INTERNATIONAL
COMPUTER SCIENCE
INSTITUTE

International Computer Science Institute Activity Report 2010

International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704-1198 USA
phone: (510) 666 2900 (510) fax: 666 2956 info@icsi.berkeley.edu http://www.icsi.berkeley.edu

Principal 2010 Sponsors

Cisco
Defense Advanced Research Projects Agency (DARPA)
Finnish National Technology Agency (TEKES)
German Ministry of Education and Research (via the DAAD)
Health Resources and Services Administration
IBM
Intel
Intelligence Advanced Research Projects Activity (IARPA)
Microsoft
National Geospatial-Intelligence Agency (NGA)
National Science Foundation (NSF)
Office of Naval Research (ONR)
Toyota

Affiliated 2010 Sponsors

Air Force Research Laboratory (AFRL)
Army Research Laboratory (ARL)
Brazilian Agency for Industrial Development (ABDI)
Comcast
Department of Energy (DOE)
Google
IM2 National Centre of Competence in Research, Switzerland
National Institutes of Health (NIH)
Panasonic
Qualcomm

Corporate Officers

Prof. Nelson Morgan (President and Institute Director)
Prof. Scott Shenker (Vice President)
David Johnson (Secretary)
Theresa Hilaire Hogg (Treasurer)
Maria Eugenia Quintana (Chief Administrative Officer)

Board of Trustees, January 2011

Prof. Hervé Bourlard, IDIAP and EPFL, Switzerland
Prof. David Culler, UC Berkeley
Prof. Deborah Estrin, UC Los Angeles
Prof. Graham Fleming, Vice Chancellor of Research, UC Berkeley
Mr. Cliff Higgerson, Walden International, and former Chairman, ICSI
Prof. Nelson Morgan, UC Berkeley, and Director and President, ICSI
Dr. David Nagel, Ascona Group
Dr. Peter Norvig, Google
Dr. Prabhakar Raghavan, Yahoo! Research and Stanford University, and Chairman, ICSI
Prof. Scott Shenker, UC Berkeley, and Vice President, ICSI
Dr. Eero Silvennoinen, Tekes
Dr. David Tennenhouse, New Venture Partners
Prof. Wolfgang Wahlster, DFKI GmbH

2010 International Visitor Program

Name	Country	Group	Sponsor
Dennis Brandão	Brazil	Other	ABDI
Baohua Yang	China	Networking	CSC
Kai Huotari	Finland	Other	TEKES
Dmitriy Kuptsov	Finland	Networking	TEKES
Tommi Lampikoski	Finland	Other	TEKES
Matti Mantere	Finland	Networking	TEKES
Boris Nechaev	Finland	Networking	TEKES
Jarno Rajahalme	Finland	Networking	TEKES
Pasi Sarolahti	Finland	Networking	TEKES
Petri Savolainen	Finland	Networking	TEKES
Jouni Similä	Finland	Other	TEKES
Marko Turpeinen	Finland	Other	TEKES
Rainer Böhme	Germany	Networking	DAAD
Joos-Hendrick Böse	Germany	Networking	DAAD
Nicolas Cebron	Germany	Vision	DAAD
Fabian Gieseke	Germany	Algorithms	DAAD*
Christoph Goebel	Germany	Other	DAAD
Paula Herber	Germany	Arhitecture	DAAD
Christian Hochmuth	Germany	Algorithms	DAAD*
Frank Hopfgartner	Germany	AI	DAAD
Sascha Hunold	Germany	Architecture	DAAD
Emanuel Kitzelmann	Germany	AI	DAAD
Oliver Kramer	Germany	Algorithms	DAAD
Jörg Lässig	Germany	Algorithms	DAAD
Gregor Maier	Germany	Networking	DAAD
Bernd Meyer	Germany	Speech	DAAD
Matthias Mnich	Germany	Algorithms	DAAD
Nils Peters	Germany	Other	DAAD
Benjamin Satzger	Germany	Algorithms	DAAD
Malte Schilling	Germany	AI	DAAD
Stefan Steidl	Germany	Speech	DAAD
Peer Stelldinger	Germany	Vision	DAAD
Dirk Sudholt	Germany	Algorithms	DAAD
Holger Ziekow	Germany	Other	DAAD
Yang Hua	Singapore	Vision	Panasonic
Zhong Yang	Singapore	Vision	Panasonic

*Guest students of DAAD-funded postdoctoral fellows
ABDI: Brazilian Agency for Industrial Development
CSC: China Scholarship Council
DAAD: Deutscher Akademischer Austausch Dienst
TEKES: Finnish National Technology Agency

Contents

Part I
INSTITUTE OVERVIEW

The International Computer Science Institute (ICSI) is one of the few independent, non-profit basic research institutes in the country, and is closely affiliated with the University of California campus in Berkeley, California. ICSI was founded in 1986 and officially inaugurated in 1988 as a joint project of the Electrical Engineering and Computer Science Department (and particularly of the Computer Science Division) of UC Berkeley and the GMD, the Research Center for Information Technology GmbH in Germany. Since then, Institute collaborations within the university have broadened (for instance, with the Electrical Engineering Division, as well as other departments such as Linguistics and Public Health). In addition, Institute support has expanded to include a range of international collaborations, US Federal grants, and direct industrial sponsorship. Throughout these changes, the Institute has maintained its commitment to a pre-competitive research program. The goal of the Institute continues to be the creation of synergy between world-leading researchers in computer science and engineering. This goal is best achieved by creating an open, international environment for both academic and industrial researchers.

ICSI's mission is simply defined as ''Furthering computer science research through international collaboration. Furthering international collaboration through computer science research.'' Toward these ends, our international visitor program has been an integral part of ICSI since its inception. In 2010, in addition to sponsored visitor programs with Finland, Germany, Brazil, and Switzerland, we had many visitors associated with specific projects. These visitors, who are often postdoctoral Fellows but in some cases are students or senior researchers, actively participate in publicly accessible studies and projects without regard to national or company boundaries. In addition to their access to ICSI and Berkeley campus experts in their respective fields, the synergy between the visitors from different countries is a key aspect of the ICSI environment. ICSI is particularly well suited to support the visitors administratively, helping with visas, housing, and more generally with orientation to what is often a very new cultural experience.

The particular areas of research concentration have varied over time, but are always chosen for their fundamental importance and their compatibility with the strengths of the Institute and affiliated UC Berkeley faculty. ICSI currently has a major focus on two broad areas: Internet Research, including Internet architecture, related theoretical questions, and network security; and Perceptual and Cognitive Systems, including text and visual processing. Additionally, there are efforts in theoretical computer science and algorithms for bioinformatics, a computer architecture group, and a local diversity project called the Berkeley Foundation for Opportunities in Information Technology (BFOIT).

The Institute occupies a 28,000 square foot research facility at 1947 Center Street, just off the central UC campus in downtown Berkeley. Administrative staff provide support for researchers: housing, visas, computational requirements, grants administration, accounting, etc. There are approximately one hundred scientists in residence at ICSI including permanent staff, postdoctoral Fellows, visitors, affiliated faculty, and students. Senior investigators are listed at the end of this overview, along with their current interests.

The current director of the Institute is Professor Nelson Morgan of the UC Berkeley Electrical Engineering faculty.

1 Institute Sponsorship for 2010

As noted earlier, ICSI is sponsored by a range of US Federal, international, and industrial sources. The figure below gives the relative distribution of funding among these different sponsoring mechanisms.

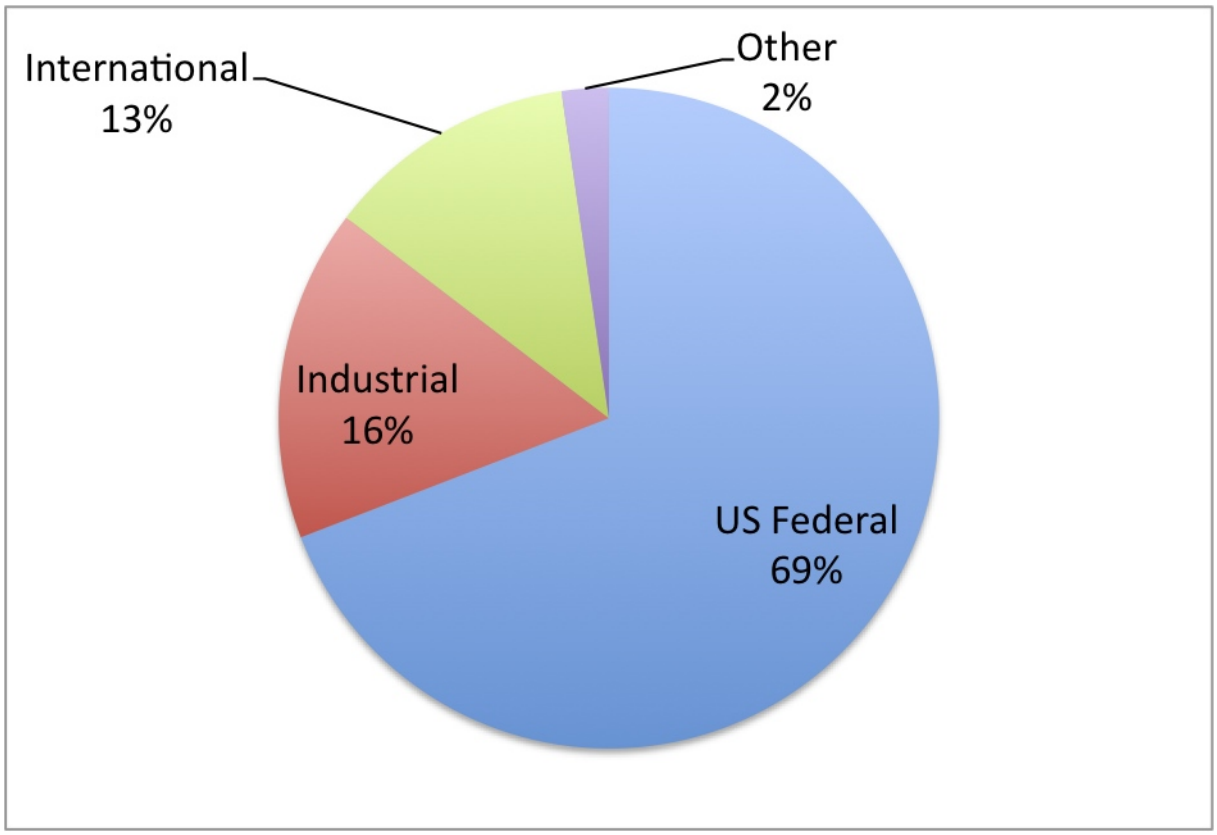

Figure 1: ICSI revenue sources for 2010.

US federal funding in 2010 came from a range of grants to support research institute-wide. Most of this funding came from the National Science Foundation, DARPA, with other Federal funds coming from AFRL, ARL, DOE, HRSA, IARPA, NGA, NIH, and ONR. International support of our visitor program came from the Ministry of Education and Research in Germany, the National Technology Agency of Finland, and the Swiss National Science Foundation (through the Swiss Research Network IM2). Brazil also recently joined the ICSI visitor program, administered through the Brazilian Competitive Movement (MBC) and the Brazilian Agency for Industrial Development (ABDI). Industrial support was provided by Cisco, Comcast, Google, IBM, Intel, Microsoft, Panasonic, Qualcomm, and Toyota. Total ICSI revenue was $7.5M in 2010.

2 Institutional Structure of ICSI

ICSI is a nonprofit California corporation with an organizational structure and bylaws consistent with that classification and with the institutional goals described in this document. In the following sections we describe the two major components of the Institute's structure: administration and research.

2.1 Management and Administration

The corporate responsibility for ICSI is ultimately vested in the person of the Board of Trustees, listed in the first part of this document. The current Chairman of the Board is Dr. Prabhakar Raghavan, the head of Yahoo! Labs. Ongoing operation of the Institute is the responsibility of Corporation Officers, namely the President, Vice President, the Secretary, the Treasurer/CFO and the Chief Administrative Officer (CAO). The President also serves as the director of the Institute, and as such, takes responsibility for day-to-day Institute operations.

Internal support functions are provided by three departments: Computer Systems, Finance, and Administrative Operations/Sponsored Projects. Computer Systems provides support for the ICSI computational infrastructure, and is led by the Systems Manager. Finance is responsible for payroll, grants administration, benefits, human resources, and generally all Institute financial matters; it is led by the CFO. All other support activities come under the general heading of Administrative Operations, and are supervised by the CAO; these activities include the visitor program, publications and communications, housing, visas, grant proposal administration, and support functions for ongoing operations and special events.

2.2 Research

Research at ICSI is overwhelmingly investigator-driven, and themes change over time as they would in an academic department. Consequently, the interests of the senior research staff are a more reliable guide to future research directions than any particular structural formalism. Nonetheless, ICSI research has been organized into groups. Our six groups currently are: Networking, Algorithms, AI, Speech, Computer Vision, and Computer Architecture. Consistent with this organization, the bulk of this report is organized along these lines, with one sub-report for each of the six groups.

Across many of these activities, there is a theme: scientific studies based on the growing ubiquity of connected computational devices. In the case of Networking, the focus is on the Internet; in the case of Speech, AI, and Vision, it is on the interfaces to the distributed computational devices. The Algorithms Group continues to develop methods that are employed in a range of computational problems, but recently has focused on problems in computational biology. The focus of the Computer Architecture Group is the realization of efficient parallel programmable architectures exploiting advances in circuit and device technologies.

Senior Research Staff: The previous paragraphs briefly described the clustering of ICSI research into major research themes and working groups. Future work could be extended to new major areas based on strategic institutional decisions and on the availability of funding to support the development of the necessary infrastructure. At any given time, ICSI research is best seen as a set of topics that are consistent with the interests of the research staff. In this section, we give the names of the senior research staff members at ICSI during 2010 and the research group that the researcher is most closely associated with, along with a brief description of their current interests. This is probably the best snapshot of research directions for potential visitors or collaborators. Not shown here are the postdoctoral Fellows, visitors, and graduate students who are also key contributors to the intellectual environment at ICSI.

Mark Allman (Networking): congestion control, network measurement, network dynamics, transport protocols and network security;

Krste Asanovic (Architecture): computer architecture, parallel programming, VLSI design, new device technologies for computing systems;

Collin Baker (AI): developing semantic frames for a large portion of the common English lexicon, and studying the extent to which these frames are applicable to other languages (including Spanish, German and Japanese) all of which have ongoing FrameNet-related projects. Also investigating the extent to which a currently manual semantic annotation process can be automated and accelerated, using automatic semantic role labeling and computer-assisted frame discovery;

Trevor Darrell (Vision): computer vision, object recognition, human motion analysis, machine learning, multimodal interfaces;

Daniel Ellis (Speech): signal processing and machine learning to extract information from sound, with applications in speech, music, and environmental recordings; source separation and acoustic scene analysis (also with Columbia University);

Jerome Feldman (AI): neurally plausible (connectionist) models of language, perception and learning and their applications;

Charles Fillmore (AI): building a lexical database for English (and the basis for multilingual expansion) which records facts about semantic and syntactic combinatorial possibilities for lexical items, capable of functioning in various applications: word sense disambiguation, computer-assisted translation, information extraction, etc.;

Sally Floyd (Networking): congestion control, transport protocols, queue management, and network simulation;

Gerald Friedland (Speech): intelligent multimedia applications, especially technology that extracts meaning (semantic) from data created for human sensory perception, including visual, acoustic, or other data, especially where the combination of modalities results in synergistic effects;

Dilek Hakkani-Tur (Speech): spoken language understanding, spoken dialog systems, active and unsupervised learning for spoken language processing;

Eran Halperin (Algorithms): computational biology, computational aspects of population genetics, combinatorial optimization, algorithm design (also with Tel Aviv University);

Jim Hieronymus (Speech): acoustic modeling for speech recognition, conversational speech recognition, spoken language understanding, spoken dialog systems, speech based

semantics, speech prosody detection and modeling and corpus based speech studies;

Adam Janin (Speech): statistical machine learning, particularly for speech recognition, speaker recognition, and language understanding; use of higher level information (e.g., semantics) in speech recognition;

Richard Karp (Algorithms/Networking): mathematics of computer networking, computational molecular biology, computational complexity, combinatorial optimization;

Paul Kay (AI): analysis of the data from the World Color Survey, which gathered color naming data in situ from 25 speakers each of 110 unwritten languages from 45 distinct language families, in order to (1) assess whether cross-language statistical universals in color naming can be observed and (2) measure the degree to which the boundaries of color categories in individual languages can be predicted from universal focal colors;

Christian Kreibich (Networking): network security; network measurement; distributed systems; botnet infiltration and containment; malware analysis;

Nikki Mirghafori (Speech): statistical machine learning and pattern recognition, specifically as applied to automatic speaker recognition, speaker diarization, speech recognition, and speech meta-data extraction.

Nelson Morgan (Speech): signal processing and pattern recognition, particularly for speech classification tasks;

Srini Narayanan (AI): cognitive computation, adaptive dynamic systems, natural language processing, structured probabilistic inference, computational semantics, computational biology, information technology for emerging regions;

Vern Paxson (Networking): large-scale Internet threats, intrusion detection/prevention, high-performance network traffic analysis, Internet measurement, forensics;

Scott Shenker (Networking): Internet architecture, network designs, datacenter computing, distributed systems;

Elizabeth Shriberg (Speech): modeling spontaneous conversation, disfluencies and repair, prosody modeling, dialog modeling, automatic speech recognition, utterance and topic segmentation, psycholinguistics, computational psycholinguistics (also with SRI International);

Robin Sommer (Networking): network security; intrusion detection and response; traffic analysis; high-performance traffic monitoring; network architectures

Andreas Stolcke (Speech): probabilistic methods for modeling and learning natural languages, in particular in connection with automatic speech recognition and understanding (also with SRI International);

Nicholas Weaver (Networking): worms and related malcode; automatic intrusion detection and response; hardware accelerated network processing;

Steven Wegmann (Speech): automatic speech recognition, acoustic modeling, diagnosing data/model mismatch, convergence properties of estimation algorithms (also with Cisco Systems Inc.).

Part II
Research Group Reports

1 Research Group Highlights

The following are a selection of key achievements in our research groups for 2010, both in group development and in research per se. Although not a complete listing and, by necessity, quite varied given the different approaches and topics of each group, it should nonetheless give the flavor of the efforts in the ICSI community for the last year. Not listed is the continuing community effort that is the Berkeley Foundation for Opportunities in Information Technology (BFOIT),which recruits high school students from underrepresented groups for potential careers in computer science and engineering, and assists in getting college acceptances and scholarship dollars.

1.1 Networking

- Mesos enables the sharing of cluster computing resources (CPU and memory) across a diverse collection of cluster computing frameworks. Using a novel "resource offer" allocation mechanism, Mesos allows a diverse set of programming frameworks (e.g., Hadoop, Pregel, etc.) to coexist on the cluster while ensuring that each framework is at least as well off in this shared cluster as they would be in a statically partitioned cluster. In addition to being used by researchers, Mesos is in production use at one large commercial datacenter.

- In collaboration with a large team including PIs from seven campuses, ICSI researchers have designed a minimal architectural ‘‘framework’’ in which comprehensive architectures can reside. The proposed Framework for Internet Innovation (FII) --- which is derived from the simple observation that network interfaces should be extensible and abstract --- allows for a diversity of Internet architectures to coexist, communicate, and evolve.

- ICSI researchers launched a new project aimed at greatly enhancing the Bro intrusion detection framework, an open-source network monitoring system that the Networking Research Group has used extensively in its past work. In collaboration with colleagues from the National Center for Supercomputing Applications, they are undertaking the establishment of a sustainable development model for the system. The National Science Foundation provides core funding for this project, supporting three full-time engineering positions dedicated to Bro development and support.

- ICSI researchers (both in the Networking and Speech Groups) demonstrated the rapidly emerging privacy threat posed by information that is published online with precise location information, such as GPS tags. While users typically realize that sharing locations has some implications for their privacy, the researchers provided evidence that many users remain unaware of the full extent of the threat that they face,

and often do not even realize that they have published such information. The work concretely explored a set of scenarios that demonstrate the ease of correlating geo-tagged data with corresponding publicly-available information in order to compromise a victim's privacy, including finding the private addresses of celebrities and the origins of anonymized Craigslist postings. This work attracted significant interest throughout the mainstream media.

1.2 Algorithms

- Eran Halperin and Lucia Conde performed the first whole-genome statistical analysis for the detection of genetic variants that may be associated with follicular lymphoma. They found two new genetic variations in the HLA region that are strongly associated with follicular lymphoma. The HLA region is a region in chromosome 6 which is related to the immune system; this discovery will lead to a better understanding of the disease mechanism and its relation to the immune system. Their work has been published in Nature Genetics.

- Halperin also collaborated with a large number of groups in Europe on association studies for cardiovascular diseases. This resulted in a paper in press in Nature Genetics, reporting on the association of 15 new genes with coronary artery disease. Halperin also received the 2010 Krill Prize for young scientists.

- Luqman Hodgkinson and Richard M. Karp designed Produles, the first algorithm with linear running time in the size of the input, for the important problem in computational biology of detecting evolutionarily conserved multi-protein modules.

- The Algorithms Group hosted four German postdoctoral scientists, Oliver Kramer, Joerg Laessig, Benjamin Satzger, and Dirk Sudholt who studied algorithms inspired by natural systems, such as evolutionary algorithms, swarm algorithms and ant colony optimization, as well as kernel-based methods from machine learning. They applied these methods to problems ranging from inventory management to adaptive control of smart energy grids.

1.3 Artificial Intelligence

- The ICSI AI group won two highly competitive basic research awards in 2010. The group was selected for a five year "Computing with Natural Language" award by the Office of Naval Research as part of their Basic Research Challenge (BRC). The group was also selected by the Templeton Foundation to research foundational issues in Mathematics and Artificial Intelligence.

- The Hesperian Digital Commons, a multilingual primary health resource developed in a collaboration between ICSI and Hesperian Foundation, is being evaluated by NGO groups around the world including Honduras (Spanish), the Philippines (Tagalog and Spanish), Lebanon (Arabic) and India (Tamil). The underlying framework was used by over 200 volunteer doctors during the aftermath of the Haiti earthquake to create a Creole-English bilingual medical term and knowledge base.

- Two new PhD dissertations were completed within the interdisciplinary NTL group. The EECS thesis by Leon Barrett resulted in a novel model of structured actions and probabilistic inference using previous research by the group on Coordinated Probabilistic Relational Models. Ellen Dodge's linguistics thesis resulted in the first compositional grammar in the Embodied Construction Grammar framework developed by the group.

- ICSI AI group member and NTL group founder Jerry Feldman was awarded the Berkeley Citation. The Berkeley Citation is one of the University's select honors, bestowed on individuals who have rendered distinguished and extraordinary service to the University. The citation noted his foundational contributions to work in neural networks and how such models can be used to explore the computational basis of complex cognition. Another ICSI NTL group member, George Lakoff, was selected to be the Richard and Rhoda Goldman Chair Professor in Linguistics and Cognitive Science (2009-2013) at the University of California, Berkeley.

1.4 Speech

- Four new part-time senior researchers were brought into the group for key roles in four new competitively awarded projects: Columbia Professor Dan Ellis for a new DARPA project on robust speech processing; Jim Hieronymus for a new DARPA project on speech analysis for emergency responders (e.g., firemen); Nikki Mirghafori for a new AFRL project on robust speaker recognition; and Steven Wegmann for a project on the diagnosis of speech recognition errors using new forms of data analysis.

- Howard Lei completed his Ph.D. dissertation on Structured Approaches to Data Selection for Speaker Recognition. This work is an essential part of the group's newest efforts in this area.

- Researchers parallelized and sped up offline speech diarization by an order of magnitude so that it now can be used in an online fashion, without hurting accuracy. This was a critical component of the group's collaboration with the Berkeley ParLab.

- Andreas Stolcke was named a Fellow of the IEEE, and Nelson Morgan was named a Fellow of the International Speech Communication Association.

- In collaboration with Robin Sommer of the Networking Group (see highlights above for more detail), Gerald Friedland exposed privacy considerations for geo-tags, for instance finding local vacationing users who might be vulnerable to burglary.

1.5 Computer Vision

- Exploitation of multiple modalities and context for effective object instance and category recognition: the group developed schemes for learning how to ''hallucinate'' one modality from another.

- Learning transformations between domains: researchers pioneered schemes for domain adaptation in vision.
- Learning rich perceptual representations for early and mid-level vision: researchers introduced novel low-level representations based on a probabilistic formulation of local histogram of gradient descriptors.
- Prof. Darrell served as Program Chair for CVPR 2010 in San Francisco, the largest meeting of the computer vision community.

1.6 Computer Architecture

- The architecture group received a third DARPA award to study extending photonic interconnects into DRAM chips. This award will pave the way for an expected DARPA-funded collaboration with Micron Technology to fabricate test chips in a true DRAM process.
- The photonics work resulted in two publications this year, including an invited paper at an optical engineering conference and a paper on photonically interconnected DRAM at the top architecture conference, ISCA.

2 Networking

2.1 Measurements and Modeling

Assessing Internet Traffic Manipulation: Internet users usually assume that their service providers enable direct, transparent, and unfettered access to the network. In reality, Internet Service Providers (ISPs) increasingly interfere with their customers' traffic, for reasons including performance optimization, security enhancements, and commercial gain. While assumptions abound regarding the prevalence of ISP-level interference with customer traffic, there has been little systematic study analyzing this issue. We have undertaken several projects to look at this practice in an empirically sound manner.

In the *Netalyzr* project [27], we developed a network measurement and debugging service that evaluates the functionality provided by people's Internet connectivity. The design aims to prove both comprehensive in terms of the properties we measure and easy to employ and understand for users with little technical background. We structure Netalyzr as a signed Java applet (which users access via their Web browser) that communicates with a suite of measurement-specific servers. Traffic between the two then probes for a diverse set of network properties, including outbound port filtering, hidden in-network HTTP caches, DNS manipulations, NAT behavior, path MTU issues, IPv6 support, and access-modem buffer capacity. In addition to reporting results to the user, Netalyzr also forms the foundation for an extensive measurement of edge-network properties. This system can be run by anyone by simply visiting http://netalyzr.icsi.berkeley.edu or http://netalyzr.com with a Java-enabled web browser.

To date, the Netalyzr project has recorded 150,000 measurement sessions from 100,000 different addresses, 7,000 organizations, and 190 countries. The measurements have found a wide range of behavior, on occasion even revealing traffic manipulation unknown to the network operators themselves. More broadly, we find chronic over-buffering of links (128KB and 256KB buffers, in particular, are common and can easily induce a second of additional latency), a significant inability to handle fragmentation (8% of senders and receivers cannot use fragmentation, while 54% of connections with path MTUs less than 1500B cannot use ICMP for signaling MTU problems), numerous HTTP caches operating incorrectly(35% of them cache strongly uncacheable content), common NXDOMAIN wildcarding (29% of our sessions are affected), impediments to DNSSEC deployment (15% of our sessions experience an effective MTU of 1500B), poor DNS performance (11% of sessions for which we can control glue records need at least 200ms to look up cached items), and deliberate manipulation of DNS results, including redirection of www.google.com and other search engines through proxy servers.

We continue to develop Netalyzr's testsuite to reflect current measurement needs (e.g., to track IPv6 adoption and correctness) and are expanding Netalyzr's coverage at additional third-party sites. We intend to operate the Netalyzr site for the foreseeable future.

Assessing Security Issues in Residential Networks: Conventional wisdom holds that residential users experience a high degree of compromise and infection, but this issue has seen little in the way of in-depth study. In this project we have undertaken a first step towards such an assessment based on monitoring the network activity (anonymized

for user privacy) of 20,000 residential DSL customers in an European urban area, roughly 1,000 users of a community network in rural India, and several thousand dormitory users at a large US university. Our assessment has focused on security issues that *overtly manifest* in network activity, such as scanning, spamming, payload signatures, and contact to botnet rendezvous points. We analyzed the relationship between overt manifestations of such activity versus the ''security hygiene'' of the user populations (antivirus and OS software updates) and potential risky behavior (accessing blacklisted URLs). We found that hygiene has little correlation with observed behavior, but risky behavior---which is quite prevalent---more than doubles the likelihood that a system will manifest security issues.

Switch-Based Measurement of Enterprise Traffic: The complexity of modern enterprise networks is ever-increasing, and our understanding of these important networks is not keeping pace. The insight into intra-subnet traffic (staying within a single LAN) available to researchers has proven particularly limited, due to the widespread use of Ethernet switches that preclude ready LAN-wide monitoring. This project builds upon an approach we have undertaken to obtain extensive intra-subnet visibility based on tapping sets of Ethernet switch ports simultaneously.

In prior work we captured more than a terabyte of data from the Lawrence Berkeley National Laboratory. The collection methodology led to a number of measurement calibration issues that require careful consideration [32]. After conducting this calibration we have pursued three parallel efforts in the past year.

- First, we conducted a preliminary analysis of performance within the enterprise, finding that the observed transmission rates often do not match the capacity of the network [31]. This initial analysis has led to the development of tools to understand more deeply the performance phenomena we observe.

- Second, we collected a new and larger dataset, taking into account the collection methodology lessons learned in our earlier calibration efforts. This new dataset has therefore allowed us to understand the accuracy of the heuristics we used to calibrate the first dataset. We find that our heuristics are reasonably sound, but, as we expected, also introduce small amounts of error. This new dataset and collection methodology, however, also raised new issues requiring a different set of calibration techniques.

- Our final effort, recently begun, focuses understanding the applications used in the enterprise context. In this new undertaking, we are beginning with an examinination of the so-called ''Windows protocols,'' which have yet to see any significant study in the literature despite their widespread use.

Longitudinal HTTP Analysis: In this project we have analyzed three and a half years of HTTP traffic observed at ICSI to characterize the evolution of various facets of web operation. While our dataset is modest in terms of user population, it is unique in its temporal breadth. We leverage the longitudinal data to study various characteristics of the

traffic, from client and server behavior to object and connection characteristics. In addition, we assess how the structure of the delivery of content across our datasets, including the use of browser caches, the efficacy of network-based proxy caches, and the use of content delivery networks. While each of the aspects we study has been investigated to some extent in prior work, our contribution is towards a unique long-term characterization [12].

Assessing Timeouts in HTTP Traffic: Timeouts play a fundamental role in network protocols, controlling numerous aspects of host behavior at different layers of the protocol stack. Previous work has documented a class of Denial of Service (DoS) attacks that leverage timeouts to force a host to preserve state with a bare minimum level of interactivity with the attacker. This effort considers the vulnerability of operational Web servers to such attacks by comparing timeouts implemented in servers with the normal Web activity that informs our understanding as to the necessary length of timeouts. We then used these two results---which generally show that the timeouts in wide use are long relative to normal Web transactions---to devise a framework to augment static timeouts with both measurements of the system and particular policy decisions in times of high load [8].

Studying Email Communication: Given the prevalence of email in everyday life, it is striking how little we know about the structure, behavior, and evolution of its transmission processes. In this work we are developing a methodology to undertake an extensive longitudinal study of the email delivery process. Before engaging the community in a large data collection effort we have been working to understand how to apply this methodology on a modest corpus of approximately 1 million email messages spanning over 20 years. While our results thus far are modest in terms of general conclusions they illustrate that this is a rich area deserving of further study and attention.

2.2 Security, Malware, and Intrusion Detection

Exploiting Multi-Core Processors to Parallelize Network Intrusion Prevention: It is becoming increasingly difficult to implement effective systems for preventing network attacks, due to the combination of (1) the rising sophistication of attacks requiring more complex analysis to detect, (2) the relentless growth in the volume of network traffic that we must analyze, and, critically, (3) the failure in recent years for uniprocessor performance to sustain the exponential gains that for so many years CPUs enjoyed (‘‘Moore’s Law’’). For commodity hardware, tomorrow’s performance gains will instead come from *multi-core* architectures in which a whole set of CPUs executes concurrently.

Taking advantage of the full power of multi-core processors for network intrusion prevention requires an in-depth approach. In this project we work toward developing an architecture customized for parallel execution of network attack analysis. At the lowest layer of the architecture is an ‘‘Active Network Interface’’ (ANI), a custom device based on an inexpensive FPGA platform. The ANI provides the in-line interface to the network, reading in packets and forwarding them after they are approved. It also serves as the front-end for dispatching copies of the packets to a set of analysis threads. The analysis itself is structured as an event-based system, which allows us to find many opportunities

for concurrent execution, since events introduce a natural, decoupled asynchrony into the flow of analysis while still maintaining good cache locality. Finally, by associating events with the packets that ultimately stimulated them, we can determine when all analysis for a given packet has been completed, and thus that it is safe to forward the pending packet---provided that none of the analysis elements has signaled that the packet should instead be discarded [40].

Based on this architecture design, we built a prototype of a multi-threaded Bro system. Turning Bro into a multi-threaded application consisted of two main parts: (1) restructuring the information flow to split the work-load across a set of threads; and (2) adapting the code base to a concurrent setting, which violates many assumptions made internally by the original Bro code about execution order. We evaluated the prototype's performance using both internal instrumentation and external profiling tools. One focus was to understand the cache effect of different thread scheduling strategies. We also have begun to port the prototype to Tilera's proprietary 64-core CPU for assessing its scaling to a larger number of cores than typical commodity systems provide today.

Large-Scale Monitoring Infrastructure for the UC Berkeley Campus: Network intrusion detection systems face major challenges in large operational networks, not only in terms of required performance but even more so in terms of the large degree of diversity exhibited in high-volume traffic. There is a world of difference between detecting attackers in a small-scale environment such as a departmental LAN (as is often used for evaluation of academic studies) and doing so at the scale of a large site. Funded by an NSF infrastructure grant, we built a new monitoring infrastructure for the UC Berkeley campus, which became operational in summer 2010. The new setup consists of about thirty PCs forming a ''Bro cluster'' [42]. The new cluster serves as a powerful research platform for network studies, providing unprecedented capabilities for analyzing a large-scale operational network in depth. Furthermore, on a technical level it allows us to systematically assess the scalability of our clustering approach to larger network loads and determine what is required to provide in-depth monitoring capabilities for other environments. The new setup also facilitates the continuation of security research tied to the operational requirements of one of the largest academic network environments in our country.

Enhancing Bro for Operational Network Security Monitoring in Scientific Environments: The popularity of our Bro system---now deployed operationally by major universities, large research labs, supercomputing centers, and open-science communities---is a two-edged sword, as it comes with demands often difficult to meet for a small research team. Funded by an NSF development grant, we have begun establishing a sustainable development model for Bro by providing explicit engineering resources outside the scope of our research efforts, and using them to advance the system to a state in which Bro's user community can take a more active role in its future development. More specifically, we are: (1) improving the perspective of Bro's end-users by providing extensive up-to-date documentation and support, and refining many of the rough edges that the system has accumulated over time; (2) unifying and modernizing Bro's current code base, which has evolved over 15 years of active development; (3) improving Bro's processing performance to

the degree required for operation in current and future large-scale scientific environments; and (4) adding new data analysis functionality in the form of an interactive graphical user interface and a transparent database interface. By specifically addressing much of the feedback the Bro team has received from users, the project enables a wide range of new sites to use Bro effectively to protect their cyberinfrastructure.

Compilation and Optimization of Protocol Analyzers for High-Speed Network Intrusion Prevention: As part of DOE-funded SBIR project, Reservoir Labs is developing a variety of hardware accelerator engines for high-speed traffic analysis for network intrusion detection. This effort includes hardware offload of regular expression matching, and parallelizing the analysis task on multi-core systems. We collaborate with them on this project, providing advice and feedback on their work, and guidance regarding interfacing their components with our Bro and BinPAC systems.

Infiltration-based Analysis of Botnet Management: Recent work, including ours, has leveraged botnet infiltration techniques to track the activities of bots over time, particularly with regard to spam campaigns. Building on our previous success in reverse-engineering botnet *command-and-control* (C&C) protocols [14], we conducted a four-month infiltration of the MegaD botnet. Our infiltration provided us with constant feeds on MegaD's complex and evolving C&C architecture, as well as its spam operations, and provided an opportunity to analyze the botmasters' operations. In particular, we collected significant evidence that the MegaD infrastructure was managed by multiple botmasters. We were also able to analyze how MegaD not only survived a ''takedown'' effort but bounced back with significantly greater vigor.

In addition, we developed new techniques for mining information about botnet C&C architecture: ''Google hacking'' to locate C&C servers scattered across the Internet, and ''milking'' C&C servers to extract not only the spectrum of commands sent to bots, but the C&C's overall structure. The resulting overall picture provided insight into MegaD's management structure, its complex and evolving C&C architecture, and its ability to withstand takedown.

Assessing Spam on Twitter: This project undertook a characterization of spam on Twitter. We found that 8% of 25 million URLs posted to the site pointed to phishing, malware, and scams listed on popular blacklists. Our analysis of the accounts that send spam found evidence that it originated from previously legitimate accounts that were compromised and thus puppeteered by spammers [22]. We based part of this analysis on a novel method we developed for determining whether a Twitter account exhibits automated behavior, based on a $\chi 2$ test to detect non-uniform posting patterns [45]. This test found that 16% of active Twitter accounts exhibit a high degree of automation, and 11% of accounts that appear to publish exclusively through the browser are in fact automated accounts that spoof the source of the updates.

Using click-through data, we analyzed spammers' use of features unique to Twitter and the degree that they affect the success of spam, finding that Twitter provides a highly successful platform for coercing users to visit spam pages; Twitter users click on embedded

URLs more than 0.1% of the time, much higher than the rates previously reported for email spam.

Given the absence of spam filtering on Twitter, we examined whether the use of URL blacklists would help to significantly stem the spread of Twitter spam. Our results indicated that blacklists are too slow at identifying new threats, allowing more than 90% of visitors to view a page before it becomes blacklisted. We also found that even if blacklist delays were reduced, the use by spammers of URL shortening services for obfuscation negates the potential gains unless tools that use blacklists develop more sophisticated spam filtering, which motivated the real-time URL classification project we discuss below.

Classifying Malicious URLs in Real-Time: On the heels of the widespread adoption of web services such as social networks and URL shorteners, scams, phishing, and malware have become regular threats. Despite extensive research, email-based spam filtering techniques generally fall short for protecting other web services. To better address this need, we developed ''Monarch,'' a real-time system based on machine-learning for rapidly identifying malicious spam URLs [23]. For a given web service, our evaluation showed that Monarch can provide accurate, real-time protection, but that the underlying characteristics of spam do not generalize across web services. In particular, spam targeting email qualitatively differs in significant ways from spam campaigns targeting Twitter, with our analysis finding that some of the basic distinctions in this regard arise from the abuse of public web hosting and redirector services. Because Monarch's decisions do not rely upon the context in which a URL appears, the system should provide sufficient generality to enable highly accurate URL filtering for a wide range of web services. Monarch's architecture emphasizes scalability, and we estimate it could protect a service such as Twitter---which needs to process 15 million URLs/day---for a bit under $800/day.

Creating High-Accuracy Spam Filters: We have traditionally viewed spam from the receiver's point of view: mail servers assaulted by a barrage of spam from which we must pick out a handful of legitimate messages. In this project we developed a system for better filtering spam by exploiting the vantage point of the spammer [37]. By instantiating and monitoring botnet hosts in a controlled environment, we are able to monitor new spam as it is created, and consequently infer the underlying *template* used to generate polymorphic e-mail messages. We have demonstrated this approach on mail traces from a range of modern botnets, and showed that we can automatically filter such spam precisely and with virtually no false positives.

Investigating the Underground Economy: One of the most disturbing recent shifts in Internet attacks has been the change from attackers motivated by glory or vanity to attackers motivated by commercial (criminal) gain. This shift threatens to greatly accelerate the ''arms race'' between defenders developing effective counters to attacks and attackers finding ways to circumvent these innovations. A major driving force behind the shift to malware has been the development of *marketplaces* that criminals use to foster a specialized economy of buyers and sellers of specialized products and services. This project, joint with UC San Diego, aims to explore these marketplaces in an attempt to characterize

their constituencies, impact, and sundry elements, in the hope that such an analysis might shed light on bottlenecks/weakspots present in the underground economy that can then be targeted to provide maximal benefit for defenders [17]. One of our current efforts in this regard concerns analyzing the use of spam campaigns conducted by cybercriminals to recruit ''mules'' for their operations; that is, essentially low-rank employees who serve to launder goods and money so that criminals can monetize the proceeds from their attacks while avoiding identification by law enforcement.

The Economic Flow of Spam: Spam is an economic phenomenon: if Spam did not produce revenue, much of it would cease to exist. In collaboration with UC San Diego, this project aims to discover the full economic and infrastructure flow required to support spam-centric industries, including online pharmacies, pirated software, and conterfeit luxury goods. The methodology involves actively probing and monitoring the DNS and HTTP infrastructure associated with spammed URLs, classifying the web sites into distinct affiliate programs, detecting the sharing of resources between affiliates, acquiring network logs of compromised systems used to host images for an illegal online pharmacy, and performing purchases. These measurements allows us to trace the full path from the origination of spam emails, through redirection infrastructure, site hosting, payment processing, and order fufilment. We can then assess each component in this path to understand how much leverage each provides for defenders to stress the spammers' value propositions. Our preliminary results indicate that disrupting payment (the merchant banks used for credit card processing, in particular) hold great promise for effectively undermining the profitability of spam.

Proactive Domain Blacklisting: In this project we explored the potential of leveraging properties inherent to domain registrations and their appearance in DNS zone files to predict the malicious use of domains proactively, using only minimal observation of domains known to be bad to drive our inference. Our analysis [16] demonstrated that 93% of the new domains that our inference procedure derives from a given known-bad domain subsequently appear suspect (based on third-party assessments), and 73% eventually appear on blacklists themselves. For these latter, proactively blocking based on our predictions provides a median head-start of about 2 days versus using a reactive blacklist, though this gain varies widely for different domains.

Visibility Into Network Activity Across Space and Time: The premise of this project is that for key operational networking tasks---in particular troubleshooting and defending against attacks---there is great utility in attaining views of network activity that are *unified across time and space*. By this we mean that procedures applied to analyzing past activity match those applied for detecting future instances, and that these procedures can seamlessly incorporate data acquired from a wide range of devices and systems. To this end, we have pursued development of *VAST* (Visibility Across Space and Time), a system that can process network activity logs comprehensively, coherently, and collaboratively [6]. The VAST system archives data from a multitude of sources and provides a query interface that can answer questions about what happened in the past, as well as notifying

operators when certain activity occurs in the future. Its policy-neutral structure allows a site to specify custom procedures to coalesce, age, sanitize, and delete data.

In addition, the VAST system can facilitate operationally viable, cross-institutional information sharing. In contrast to today's inefficient and cumbersome operational practices---phone calls, emails, manual coordination via IM---we envision a framework that enables operators to leverage each others' VAST systems. To address the important trust and privacy constraints of a such a setting, we also introduce the notion of a per-site *Clearing House* component that provides operators with fine-grained control over the flow of information, enabling them to deploy the full spectrum of information sharing, from automated sending and receiving of descriptions of activity, to holding all requests for explicit, manual approval.

Internet Situational Awareness: Effective network security administration depends to a great extent on having accurate, concise, high-quality information about malicious activity in one's network. ''Honeynets''---collections of sacrificial hosts (''honeypots'') fed traffic seen on an unused region of a network---can potentially provide such detailed information, but the volume and diversity of this data can prove overwhelming. In this project we explore ways to analyze the probes seen by honeynet data in order to assess whether a given ''event'' present in the honeynet reflects the onset of a new Internet worm, a benign misconfiguration, or a concerted effort to scan the site. For this latter (the most common), we then attempt to refine the analysis to assess whether the scanning *targeted* the site in particular, or was merely part of a much broader, indiscriminate scan. Our preliminary results indicate our analysis using *purely local information* generally yields estimates of global targeting scope quite close to those obtained more directly from the global *DShield* repository of Internet scanning activity [28, 29].

Privacy Implications of Location-Based Services: This project aims to raise awareness for a rapidly emerging privacy threat that we termed ''cybercasing'': leveraging geo-tagged information available online in order to mount real-world attacks [18]. While users typically realize that sharing locations has some implications for their privacy, we provided evidence that many users (1) are unaware of the full scope of the threat they face when doing so, and (2) often do not even realize that they *have* published such information. The threat is elevated by recent developments that make systematic search for geo-located data, and cross-inference from multiple sources, easier than ever before. We summarized the state of geo-tagging, estimated the amount of geo-information available on several major sites, including YouTube, Twitter, and Craigslist, and examined its programmatic accessibility through public APIs. We framed a set of scenarios demonstrating the ease of correlating geo-tagged data with corresponding publicly-available information for compromising a victim's privacy. For example, we were able to find private addresses of celebrities, as well as the origins of otherwise-anonymized Craigslist postings. This work attracted significant interest throughout the mainstream media, including coverage by the New York Times, New Scientist, Toronto Star, ABC News, and Fox.

The Impact of HTTP Cache Poisoning: Many web sites now include shared script content loaded from third party sites. Because of the shared nature of this code and the lack of integrity on HTTP caching, it becomes possible for an attacker who can temporarily intercept traffic (such as a user at a WiFi hotspot) to insert a malicious element into the victim's cache that can linger for a long period of time, ultimately affecting many sites. Even when not used to introduce malicious content, these analytic scripts still leak a huge amount of user behavior to third parties. We are currently evaluating both the effectiveness of this attack, and the scale of the privacy leakage due to these analytic scripts. We have found that over 1% of *all* HTTP requests in one monitored networks represent deliberate information transfers due to analytic scripts.

Using Traffic Characteristics to Detect Spam: In this project we evaluate the efficacy of using a machine learning-based model of network and transport layer characteristics of email traffic to identify spam. The underlying idea is that the manner in which spam is transmitted has an impact that is statistically observable in the traffic (e.g., in the network round-trip time or jitter between packets). Therefore, by identifying a solid set of traffic features we can construct a model that can identify spam without relying on expensive content filtering. We carried out a large scale empirical analysis of this idea with data collected over the course of one year (roughly 600K messages). With this data, we trained classifiers using machine learning methods and test several hypotheses. First, we validated prior results using similar techniques. Second, we determined which features contributed most significantly to the detection process. Third, we analyzed the behavior of our detectors over weekly and monthly intervals, and in the presence of major network events. Finally, we evaluated the behavior of our detectors in a practical setting where they are used in a filtering pipeline along with standard off-the-shelf content filtering methods, and demonstrated that they can lead to computational savings in practice [33]. Further, we reappraised results from the literature that found that a subset of such features could yield spam filtering accuracy roughly akin to traditional DNS blacklists. While the previous work leveraged data from 2,500 locations in the network, we evaluated the efficacy of a single enterprise using their own vantage point to develop models. As hypothesized, we find the earlier techniques do not work well as a blacklist replacement when observations from only a single vantage point are used [34].

An Abstract Execution Environment for High-Performance Network Traffic Analysis: When building applications that process large volumes of network traffic---such as high-performance firewalls or intrusion detection systems---one faces a striking gap between the ease with which the desired analysis can often be described in high-level terms, and the tremendous amount of low-level implementation details one must still grapple with for coming to an efficient and robust system. In a major project we are building novel environment called HILTI that provides a bridge between these two levels by offering to the application designer the high-level abstractions required for effectively describing typical network analysis tasks, while still ensuring the performance necessary for monitoring Gbps networks in operational settings. This new middle-layer consists two main pieces: (1) an abstract machine model tailored to the networking domain that directly supports the

field's common abstractions and idioms in its instruction set; and (2) a compilation strategy for turning programs written for the abstract machine into highly optimized, natively executable task-parallel code for a given target platform. The environment provides many opportunities for extensive compile-time code optimizations leveraging domain-specific context, and it holds promise for unleashing the community's potential to build libraries of efficient analysis functionality, reusable across a wide range of scenarios. [41]

A Next-Generation Parser Generator for Network Protocols: The first application to leverage the capabilities provided by the HILTI system (see above) is a successor to our previous BinPAC system (''a yacc for writing network protocol analyzers''). We conducted a poll among Bro's developer community for suggestions of what BinPAC is lacking. We received more than 50 ideas, including (1) integrating semantic constructs into the BinPAC language, rather than just syntax; (2) a better unified handling of similar protocol elements; (3) robust error checking and recovery; and (4) support for highly concurrent execution. Based on this feedback, we have begun building the next-generation BinPAC language, and we have been working on a prototype compiler translating the new specifications into the HILTI model.

Anomaly Detection for HPC Environments: In network intrusion detection research, one popular strategy for finding attacks is monitoring a network's activity for *anomalies*: deviations from profiles of normality previously learned from benign traffic, typically identified using tools borrowed from the machine learning community. However, despite extensive academic research one finds a striking gap in terms of actual deployments of such systems: compared with other intrusion detection approaches, machine learning is rarely employed in operational ''real world'' settings. To this end, we examined the differences between the network intrusion detection problem and other areas where machine learning regularly finds much more success [39]. Our main tenet is that the task of finding attacks fundamentally differs from these other applications, making it significantly harder for the intrusion detection community to employ machine learning effectively. We support this claim by identifying challenges particular to network intrusion detection, and provide a set of guidelines meant to strengthen future research on anomaly detection.

Keeping these insights in mind, we began a new DOE-funded project examining anomaly detection for high-performance computing. Jointly with colleagues from LBNL, NERSC, and UC Davis, we are investigating novel approaches for identifying malicious activity in supercomputing environments. Compared to standard Internet environments, we expect the normal workload on such systems to be rather homogeneous, making machine-learning approaches much more suitable for finding deviations from expected behavior than in more general settings.

Assessing Internet Scanning: In previous work we have used a collection of traffic logs collected at the Lawrence Berkeley National Laboratory (LBNL) to assess various aspects of scanning coming from the Internet [7]. In the previous year we have extended this work to include investigating patterns of scanning behavior. We first find that most scanners do not probe enough to draw strong conclusions about their behavior. This may

be indicative of random scanning across the entire Internet---of which we are observing only a small portion. Of the scanners that probe LBNL enough times to form an understanding of their behavior, we find that linearly marching through the address space constitutes the most prevalent form of scanning. We also find small but consistent amounts of random scanning. Early results were presented in a poster [15]. Future work will be to assess the remaining unclassified scanning behavior.

Efficient Information Flow Tracking: Byte-level information flow tracking (IFT) is a useful primitive for a number of applications, but it has not been practically feasible for the large existing base of deployed software. Prior work has demonstrated that byte-level IFT faces two significant practical limitations -- enormous performance overhead and taint explosion. In this project, we introduce several novel techniques aimed at addressing these limitations. We have developed PIFT -- a new Xen- and QEMU-based information flow tracking system that achieves a 60% performance improvement over the best previously published results and largely eliminates kernel taint explosion.

Hypervisors as a Foothold for Internet Security: The ease by which attackers can compromise end-user systems has led to enormous problems for Internet security. Today's attackers install malware on users' desktops and laptops by exploiting their large, complex software stacks, or simply by tricking users into downloading trojan horses. Once installed, the malware readily acquires full control of the machine, opening the floodgates to both information theft and further attacks on third parties via activities such as spamming or DoS.

Remedying this unhappy situation has proven extremely difficult because of fundamental limitations in the different methods currently available to deliver security solutions to end users. First, the vast complexity of consumer operating systems provides a large and inviting attack surface, so we cannot rely upon the OS's own mechanisms to protect us. Similarly, many third-party anti-malware tools run in the OS, so they are also vulnerable to subversion if the complex OS succumbs to an attack. Finally, while we can provide some security functions in the network using middleboxes, these devices suffer from fundamentally limited visibility into end-system activity and its corresponding semantics.

However, hardware trends have begun opening up a new avenue for deploying security solutions: hypervisors. As a small programmable layer between the OS, the network, and the hardware, the hypervisor is a uniquely attractive point to insert security functionality. While hypervisors have already seen considerable employment for improving security in datacenters and some enterprises, their use has largely been for isolation. Likewise, previous work by the research community has primarily (but not exclusively!) framed the isolation-based benefits of hypervisors.

There are two reasons why we think there is significantly broader potential for hypervisor-based security. First, hardware virtualization support is now coming to the home -- desktop and laptop CPUs from Intel and AMD have included it for the past few years. Thus, the near future will provide greatly expanded opportunities to leverage hypervisors. Second, we believe that a much wider variety of security functionality can be embedded in hypervisors than has been previously explored.

Put broadly, we -- the research community -- have before us a golden opportunity to have impact by producing a broadly deployable hypervisor-based security platform. The ideal outcome of such an effort would be a security-enhancing hypervisor that end users can readily install, or, even better, that OS vendors ship in an OS upgrade.

Building a Trusted Path to the User with Virtual Dedicated Clients: Establishing a trusted path between an online service and its users is a fundamental problem in Internet security. Users' interactions with services can be compromised in many ways: malware can capture user credentials or perform actions on behalf of the user, phishers may spoof websites, or attackers may simply brute-force user passwords. In this project we propose a system called Virtual Dedicated Clients (VDC) to address these problems. VDC aims to provide a level of security analogous to what would be achieved if each online service gave each of its users a locked-down, dedicated device designed only for accessing their account. VDC builds on a trusted substrate running below an unmodified OS (such as a VMM), but unlike systems based on VMs, it only adds a small amount of trusted code on top of this substrate: a generic remote UI client. VDC also introduces a novel scheme called local passwords to make it impossible for phishers to access a user's account even if they capture her password. Finally, although full VDC creates an isolated UI for accessing services, we have also designed a mechanism called VDC Lite that lets web applications in the insecure OS gain many of the benefits of VDC.

2.3 Internet Protocols

Updating TCP's Retransmission Timeout Algorithm: We have worked on updating TCP's standard retransmission timeout (RTO) scheme [35] to lower the initial RTO from 3 seconds to 1 second to provide more timely retransmits over modern networks. Our work on this involved analyzing eight packet-level datasets from four distinct vantage points and times to assess the efficacy of the change. The updated specification---including our analysis results---is under consideration by the IETF [36].

Early Retransmit: In this effort we introduced a new mechanism for TCP and SCTP for recovering lost segments in the presence of a small congestion window. The ''Early Retransmit'' mechanism [5] allows the transport to reduce (in certain special circumstances) the number of duplicate acknowledgments required to trigger a fast retransmission. Doing so allows the transport to use Fast Retransmit to recover packet losses that would otherwise require a lengthy retransmission timeout.

Updating TCP's SACK-based loss recovery: We have worked on updating TCP's selective acknowledgment-based loss recovery algorithm [9] to include a better mechanism to transitioning between steady-state transmission and loss recovery when needed [10]. The new scheme allows this transition to occur sooner and in the presence of more loss---resulting in better performance.

Re-visiting TCP: In cooperation with Google we have been re-visiting a number of aspects of TCP and its congestion control algorithms in the context of a large-scale service provider. Issues we have been working on to date include those relating to the operation of the retransmission timeout, the initial congestion window size [4], loss recovery, so-called ''dead'' connections that inexplicably end with no warning in the middle of a transaction, and a general investigation of the origins of the observed performance within the context of data gathered at Google's data centers.

2.4 Novel Internet Architectures

Architectural Support for Network Trouble-Shooting: Troubleshooting is an inherent part of network operation: no matter how well networks are designed, something eventually fails, and in large networks, failures are ever-present. In the past, troubleshooting has mostly relied on *ad hoc* techniques cobbled together as afterthoughts. However, both the importance and difficulty of troubleshooting has intensified as networks have become crucial, ubiquitous components of modern life, while at the same time their size and complexity continues to grow. These twin pressures highlight the urgent need to integrate troubleshooting as a first-class citizen when developing a network architecture.

This project pursues a key set of building blocks for developing networks that are much more amenable to troubleshooting. *Annotations* provide a means for associating meta-information with network activity. One use of annotations is to *track causality* in terms of how instances of network activity relate to previous activity. We envision much more powerful forms of *logging*, enhanced by notions of *distillation* of logged information into more abstract forms over time, and *dialog* between system components that generate log entries and the logger itself, which can call back to the component to support highly flexible distillation as well as interactive debuggers. Finally, we feed logs from multiple observation points into *repositories* that construct aggregated views of activity and mediate the ways in which sites share information for cooperative trouble-shooting.

Relationship-Oriented Networking: Humans, over centuries, have built and leveraged the notion of *relationships* in everyday actions. We have started a new project to build the notion of relationships into network architecture. Relationships can connect a variety of actors participating in a network, both users and resources, and can be woven into the network fabric to allow their usage across protocols, layers, services, components, and applications. We are exploring a number of scenarios where exposing and acting based on relationships can improve network security, trust, and usability. We have started to develop the basic building blocks necessary to implement our vision.

One of the crucial building blocks we have developed is a system for storing user-centric *meta-information*, or information required to access some piece of data and not the data itself. For instance, such meta-information might include an encoded relationship that allows one person to access another's data. As a case study, we implemented a ''personal naming'' [1] mechanism within our system, whereby users can draw upon context-sensitive and human-understandable names in their network interactions, instead of complicated URLs, disk share names, etc. In addition to our backend Meta-Information Storage System

(MISS), we have front-end plugins for Firefox and Thunderbird to allow normal users to use our naming system.

A Strongly Typed Network Architecture: Modern networks are employing a rapidly growing number of middleboxes that take complex actions based on the nature of network traffic. Unfortunately, due to lack of inherent network support, these middleboxes end-up using approximate, and often erroneous, heuristics to infer the nature of traffic, or they simply ''give up'' when inference is deemed hard. Even more crucially, actions taken by middleboxes are completely hidden from users. Users react to this by cloaking their traffic to circumventing middlebox actions, leading to an ''arms race'' and to networks making draconian enforcement decisions. Thus, we are faced with growing uncertainty in the effectiveness and correctness of middleboxes, increasing complexity in their design, growing protocol entanglement, and worsening brittleness of the network architecture.

In this effort we explore architectural approaches to resolve this increasingly untenable situation. The key ideas at the core of our current design [30] are *annotated networking* and *dialog*: we require application messages to contain trusted type information that fully describes the content being transferred, as well as trustworthy information about sending and connection properties of the application end-points. This allows a *dialog* with network elements along a path to determine whether the end system wishes to yield to the required monitoring or seek an alternative, more permissive path through the network. Our framework allows the end systems to define which portions of their communication may be inspected, modified, or kept wholly private from network elements.

Special-Purpose Social Networks for Emergency Communication: We designed a special-purpose social network that will allow people to communicate their status with friends and family when they find themselves caught up in a large disaster (e.g., sending ''I'm fine'' in the immediate aftermath of an earthquake) [3]. Since communication between a disaster zone and the non-affected world is often highly constrained due to both infrastructure failures and massive contention for the remaining resources we design the system around lightweight triggers such that people can communicate status with small messages over crude infrastructure (or even sneaker-nets).

Congestion Control Efficiency and Incentives: In [21] we studied under what conditions congestion control schemes can be both efficient, so that capacity is not wasted, and incentive compatible, so that each participant can maximize its utility by following the prescribed protocol. We show that both conditions can be achieved if routers run strict priority queueing (SPQ) or weighted fair queueing (WFQ) and end-hosts run any of a family of protocols which we call Probing Increase Educated Decrease (PIED). A natural question is whether incentive compatibility and efficiency are possible while avoiding the per-flow processing of WFQ. We partially address that question in the negative by showing that any policy satisfying a certain ''locality'' condition cannot guarantee both properties.

Building Extensible Networks with Rule-Based Forwarding: We have developed [38] a network design that provides flexible and policy-compliant forwarding. Our proposal

centers around a new architectural concept: that of packet rules. A rule is a simple if-then-else construct that describes the manner in which the network should or should not forward packets. A packet identifies the rule by which it is to be forwarded and routers forward each packet in accordance with its associated rule. Each packet rule is certified, guaranteeing that all parties involved in forwarding a packet agree with the packet's rule. Packets containing uncertified rules are simply dropped in the network. RBF supports a variety of use cases including content caching, middlebox selection and DDoS protection. Using our prototype router implementation we have verified that the overhead RBF imposes is within the capabilities of modern network equipment.

Enabling Architectural Innovation: Over the past decade there have been many attempts at clean-slate redesigns of the Internet architecture. These efforts have usually focused on addressing one or more of the current architecture's functional shortcomings, such as its poor security or its host-centric rather than content-centric design. These clean-slate activities, the most recent of which is NSF's Future Internet Architecture program, have produced, and will continue to produce, a wealth of insight into how one might build an Internet that better realizes these various features.

Our work on this project takes a radically different but wholly complementary approach, focusing not on specific functional improvements but on the broader issue of fostering architectural innovation. The reasoning behind this choice is simple: in most areas where the Internet architecture is known to be deficient, such as reliability and security, the literature is replete with proposals for how the architecture might be dramatically improved. However, almost none of these proposals have been realized in the current infrastructure because they face insurmountable deployment barriers -- barriers which are largely due to the current Internet's lack of architectural modularity -- and the literature on how to overcome this architectural impasse is comparatively barren. This suggests that the biggest intellectual challenge facing the current architecture is not a particular functional deficiency, but its inability to gracefully accommodate innovation.

Moreover, the need to support architectural innovation is of fundamental importance because any architecture, if it is unable to evolve, will eventually be found wanting in one respect or another. And yet, despite this need, we know more about how to build a more secure and reliable Internet than we do about how to build a more evolvable Internet. In short, while we know how to improve many of the Internet's features, we know far less about how to improve its overall future.

This project focuses on architecting for innovation, with the goal of allowing the Internet to evolve over time and adapt to future requirements. The key design challenge is to identify the absolutely minimal portion of the design that needs to be universally agreed upon (and therefore fixed for long periods of time) and use this core as a foundational framework that allows the rest of the architecture to evolve more freely. The proposed Framework for Internet Innovation (FII) has only three basic elements: the syntax of interdomain routing, the syntax of the network API, and a security primitive to help combat denial-of-service. There are no other globally required standards, leaving room for continuing innovation and evolution in all other aspects of Internet design. For example, there is no requirement for a uniform addressing scheme or forwarding method (as in today's IP); each domain

(i.e., an Autonomous System) can independently decide which forwarding methods and addressing schemes to deploy. Because applications are shielded from low-level network details by an abstract and extensible network API, they can operate without change over all of these architectures. This flexibility in forwarding and addressing is an example of how FII enables innovation across space (different domains) and time (different generations of design).

2.5 Datacenters

Delay Scheduling: As organizations increasingly use data-intensive cluster computing systems such as Hadoop and Dryad for a variety of applications, there is a growing need to share clusters between users. However, there is a conflict between fairness and efficiency: fairness in terms of scheduling jobs and efficiency in terms of data locality (placing tasks on nodes that contain their input data). To address this conflict, we propose a simple algorithm called delay scheduling [44]: when the job that should be scheduled next according to fairness cannot launch a local task, it waits for a small amount of time, letting other jobs launch tasks instead. We find that delay scheduling achieves nearly optimal data locality in a variety of workloads and can increase throughput by up to 2x while preserving fairness. In addition, the simplicity of delay scheduling makes it applicable under a wide variety of scheduling policies beyond fair sharing.

Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center: We have developed Mesos [25], a platform for sharing commodity clusters between multiple diverse cluster computing frameworks, such as Hadoop and MPI. Sharing improves cluster utilization and avoids per-framework data replication. Mesos shares resources in a fine-grained manner, allowing frameworks to achieve data locality by taking turns reading data stored on each machine. To support the sophisticated schedulers of today's frameworks, Mesos introduces a distributed two-level scheduling mechanism called resource offers. Mesos decides how many resources to offer each framework, while frameworks decide which resources to accept and which computations to run on them. This design enables Mesos to be small, scalable and fault tolerant. Our experimental results show that Mesos can achieve near-optimal data locality when sharing the cluster among diverse frameworks, scales up to 50,000 (emulated) nodes, and is resilient to node failures.

Fair Allocation of Multiple Resources in Datacenters: Most cluster computing frameworks, such as Hadoop and Dryad, provide fair scheduling of jobs. These schedulers fairly distribute one resource -- typically fixed-size partitions of nodes called ‘‘slots.’’ However, most data-center workloads consist of jobs with heterogeneous demands across multiple resources, such as CPUs, memory and I/O bandwidth. We have investigated what constitutes a fair division of multiple resources when users possibly prioritize resources differently. We have developed a notion of fairness called Dominant Resource Fairness (DRF) [20], which is a generalization of max-min fairness from networking to multiple resource types. We have shown that DRF, unlike other policies, satisfies a number of desirable properties. The main advantage of DRF is that it is strategy-proof, incentivizing

users to correctly report their resource demand, as they cannot increase their allocation by lying. DRF is also envy-free, which means that no user wants to trade her allocation with another's. Finally, DRF guarantees that every user's resource share is at least as high as it would be in a dedicated cluster with 1/n:th of the resources, incentivizing users to pool resources. Other schedulers that we compare to do not satisfy all these properties. We have implemented and evaluated DRF in Mesos, a cluster manager that multiplexes resources between frameworks such as Hadoop and MPI.

2.6 Software-Defined Networking

Overall Approach: The Software-Defined Networking (SDN) paradigm is a new approach that is built upon two fundamental principles:

- Software-Defined Forwarding (SDF): Forwarding functionality should be controllable by software through an open interface. This can be achieved with hardware that accepts from software a set of <header template, forwarding action> entries, where the designated forwarding actions (such as forward out a particular port, or drop) are applied to packets with headers matching the template (which can contain wildcards). OpenFlow (www.openflowswitch.org) is an example of such a hardware interface.

- Global Management Abstractions: Networks should support a basic set of global management abstractions upon which more advanced management tools can be built. These global management abstractions might include, for example: a global view of the network, triggers on network events (such as topology changes or new flows), and the ability to control network elements by inserting entries into their hardware forwarding tables. The NOX network operating system (www.noxrepo.org) is an example of a system that provides such abstractions.

The SDN approach, which was developed jointly by researchers at ICSI and Stanford University, is being used to control production networks in academia, and is now being tested in various commercial settings. The projects below are investigations of various aspects of the SDN paradigm.

Scalability of Software-Defined Forwarding: SDF involves making decisions in software and then caching entries for individual ''flows'' in hardware. The question is whether this approach can achieve reasonable performance in typical settings. We have investigated this with a variety of traces, and our early findings suggest SDF is quite scalable.

A Distributed Control Platform for Large-Scale Production Networks: Computer networks lack a general control paradigm, as traditional networks do not provide any network- wide management abstractions. As a result, each new function (such as routing) must provide its own state distribution, element discovery, and failure recovery mechanisms. We believe this lack of a common control platform has significantly hindered the development of flexible, reliable and feature-rich network control planes. To address

this, we have developed Onix [26], a platform on top of which a network control plane can be implemented as a distributed system. Control planes written within Onix operate on a global view of the network, and use basic state distribution primitives provided by the platform. Thus Onix provides a general API for control plane implementations, while allowing them to make their own trade-offs among consistency, durability, and scalability.

Ripcord: In collaboration with Nick McKeown's group at Stanford, we have developed Ripcord, a modular platform for rapidly prototyping scale-out data center networks. Ripcord enables researchers to build and evaluate new network features and topologies, using commercially available hardware and open-source software. We have used Ripcord to develop and deploy three advanced network functions, operating together on top of a 160-node cluster. The first is a routing engine that isolates classes of traffic. The second is a dynamic network manager than adjusts links and switch power states to reduce energy. The third is a statistics aggregator that supports network health monitoring and automatic alerts. See [24] for a short description of this design.

Network Virtualization Revisited: Modern datacenters have adopted the notion of ''scaling out, not up'' and have applied this paradigm successfully to computation and storage. However, attempts to scale out network functionality have only been partially successful. While recent network designs have scaled out simple packet delivery by creating a single scalable switching fabric out of a set of commodity switches, they have not addressed more sophisticated network functionality such as access control lists, monitoring, isolation, and the like. In this project [13] we have developed a new network layer, which we call a network hypervisor, that provides a clean abstraction for implementing sophisticated network-wide control. We discuss several applications of this approach, including scaling out network control stacks (such as an existing software routing suite), building distributed virtual switching layers, and creating a multi-tenant network infrastructure. This approach can be considered a way to achieve true network virtualization, finally bringing networking into parity with storage and computation in this regard.

2.7 New Routing Designs

Routing Along DAGs: Borrowing ideas from the wireless networking literature, we argue that the fundamental output of routing algorithms should be a directed acyclic graph (DAG) between each source and destination, rather than a path (or small set of paths). This ''routing along DAGs'' (RAD) approach allows a single routing framework to support both fast failover and adaptive real-time load-balancing. In particular, we show that the RAD approach guarantees connectivity without global route recomputation (as long as the graph remains connected) and provides optimal load distribution in the case of a single-destination traffic model.

Routing Reconvergence Considered Harmful: In this project we articulate the reasons for adopting a different routing paradigm [11]. Rather than recomputing paths after temporary topology changes, we argue for a separation of timescale between offline

computation of multiple diverse paths and online spreading of load over these paths. We believe decoupling failure recovery from path computation leads to networks that are inherently more efficient, more scalable, and easier to manage.

YAMR: Multipath routing is a promising technique to increase the Internet's reliability and to give users greater control over the service they receive. However, past proposals choose sets of paths that are not guaranteed to have high diversity. We propose yet another multipath routing scheme (YAMR) for the interdomain case [19]. YAMR provably constructs a set of paths that is resilient to any one inter-domain link failure, thus achieving high reliability in a systematic way. Further, even though YAMR maintains more paths than BGP, it actually requires significantly less control traffic, thus alleviating instead of worsening one of the Internet's scalability problems. This reduction in churn is achieved by a novel hiding technique that automatically localizes failures leaving the greater part of the Internet completely oblivious.

Fast Failure Resilience in Source-controlled Routing: Source-controlled routing has been proposed as a way to improve flexibility of future network architectures, as well as simplifying the data plane. However, if a packet specifies its path, this precludes fast local re-routing within the network. We propose a novel solution: specify alternate paths in the packet header itself, in the form of a compactly-encoded directed acyclic graph. We show that this can be accomplished with reasonably small packet headers for real network topologies, and results in responsiveness to failures that is competitive with past approaches that require much more state within the network. Our approach thus enables fast protection against single-link failures while preserving the benefits of source-controlled routing.

Pathlet Routing: We have developed a new multipath routing protocol, pathlet routing, in which networks advertise fragments of paths (pathlets) over virtual nodes. Sources concatenate a sequence of pathlets into an end-to-end source route. Intuitively, the pathlet is a highly flexible building block, capturing policy constraints as well as enabling an exponentially large number of path choices. In particular, we have shown that pathlet routing can emulate the policies of BGP, source routing, and several recent multipath proposals.

This flexibility allows pathlet routing to address two key challenges for interdomain routing: choice of routes for senders and scalability. When a router's routing policy has only ''local'' constraints, it can be represented using a small number of pathlets, leading to very small forwarding tables and many choices of routes for senders. Pathlet routing does not impose a global requirement on what style of policy is used, but rather allows multiple styles to coexist. Crucially, those routers that use local policies obtain the immediate benefit of small forwarding tables, regardless of what the other routers choose. Pathlet routing thus supports complex BGP-style policies while enabling and incentivizing the adoption of policies that yield small forwarding plane state and a high degree of path choice.

A Routing Policy Framework for the Future Internet: Policy is now crucially important for network design: there are many stakeholders, each with requirements that a network should support. Among many examples, senders have an interest in the paths that their packets take, providers have analogous interests based on business relationships, and receivers want to shut off traffic from flooding senders. Unfortunately, it is not clear how to balance these considerations in principle or what mechanism could uphold a large union of them in practice. To bring the policy issues into focus, in this project we (ironically) avoid predictions about which policy requirements will predominate in a future Internet and instead seek the most general policy framework we can possibly implement. To that end, we articulate a general policy principle; in condensed form, it is to empower all stakeholders by allowing communications if and only if all participants agree. Upholding this principle in the context of Internet realities, such as malicious participants, decentralized trust, and the need for high-speed forwarding, brings many technical challenges. As an existence proof that they can be surmounted, we have designed and implemented this architecture and evaluated its performance.

2.8 Research Community Activities:

Scott Shenker continues to serve on NetSE Council (formerly the GENI Science Council).

Mark Allman served as the program chair for the ACM Internet Measurement Conference (IMC), as well as joining the IMC steering committee. He continues to serve on the IETF's Transport Area Directorate. He gave invited talks at Youngstown State University, Case Western Reserve University, and Swinburne University.

Christian Kreibich gave invited talks at a meeting of the Messaging Anti-Abuse Working Group (for which the Networking Group serves as research consultants), at the NSF INCO-TRUST Workshop on International Cooperation in Security and Privacy, at a DOE Grassroots Roundtable meeting on Cybersecurity, and at the TRUST Security Seminar series at UC Berkeley.

Vern Paxson serves on the steering committee of the USENIX Workshop on Large-scale Exploits and Emergent Threats and the Network and Distributed System Security Symposium. He serves as a member of the Scientific Advisory Board for Technicolor and the Technical Advisory Board for FireEye. In 2010, he gave invited talks at Chalmers University in Sweden and the Technical University of Berlin in Germany. He also gave a briefing to the directors of the DHS National Cyber Security Division and the DHS-sponsored I3P on the topic of network situational awareness.

Robin Sommer serves on the steering committee of the Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA); as the program (co-)chair for the 2010 and 2011 Symposiums on Recent Advances in Intrusion Detection (RAID); and as Treasurer for the 2011 IEEE Symposium on Security & Privacy. He guest lectured at UC Berkeley on the difficulty of finding intrusions with anomaly detection; and gave an invited talk at Cisco on exploiting multi-core processors for parallelizing network intrusion prevention.

Nicholas Weaver has participated actively in the FCC broadband measurement workshop, and has given presentations/seminars on Netalyzr at the FCC, the Internet 2 Joint Techs workshop, NANOG, the University of Auckland, and the University of Waitomo.

At RWTH Aachen, Robin Sommer and Christian Kreibich gave a three-day block lecture on network monitoring and advanced attack detection/prevention, including an introduction to intrusion detection, the Bro IDS, challenges in conducting anomaly detection in the network context, botnets, privacy implications of geo-tagging, and edge network troubleshooting.

References

[1] M. Allman (2007). ''Personal Namespaces.'' Proceedings of ACM Special Interest Group on Data Communications Workshop on Hot Topics in Networks (HotNets-VI), Atlanta, Georgia, November 2007.

[2] M. Allman (2009). ''Comments On Selecting Ephemeral Ports.'' *ACM SIGCOMM Computer Communication Review*, Vol. 39, Issue 2, pp. 13-19, April 2009.

[3] M. Allman (2010). ''On Building Special-Purpose Social Networks for Emergency Communication.'' *ACM SIGCOMM Computer Communication Review*, Vol. 40, Issue 5, October 2010.

[4] M. Allman (2010). ''Initial Congestion Window Specification.'' Internet-Draft draft-allman-tcpm-bump-initcwnd-00.txt, work in progress, November 2010.

[5] M. Allman, K. Avrachenkov, U. Ayesta, J. Blanton, and P. Hurtig (2010). ''Early Retransmit for TCP and Stream Control Transmission Protocol (SCTP).'' Request for Comments 5827, Experimental, April 2010.

[6] M. Allman, C. Kreibich, V. Paxson, R. Sommer, and N. Weaver (2008). ''Principles for Developing Comprehensive Network Visibility.'' Proceedings of the Third USENIX Workshop on Hot Topics in Security (HotSec 08), San Jose, California, July 2008.

[7] M. Allman, V. Paxson, and J. Terrell (2007). ''A Brief History of Scanning.'' Proceedings of ACM SIGCOMM/USENIX Conference on Internet Measurement, San Diego, California, pp. 77-82, October 2007.

[8] Z. Al-Qudah, M. Rabinovich, and M. Allman (2010). ''Web Timeouts and Their Implications.'' Proceedings of the 11th Passive and Active Measurement Conference (PAM 2010), Zurich, Switzerland, April 2010

[9] E. Blanton, M. Allman, K. Fall, and L. Wang (2003). ''A Conservative Selective Acknowledgment (SACK)-Based Loss Recovery Algorithm for TCP.'' Request for Comments 3517, Standards, April 2003.

[10] E. Blanton, M. Allman, I. Jarvinen, and M. Kojo (2010). ''A Conservative Selective Acknowledgment (SACK)-based Loss Recovery Algorithm for TCP.'' Internet-Draft draft-blanton-tcpm-3517bis-00.txt, work in progress, October 2010.

[11] M. Caesar, M. Casado, T. Koponen, J. Rexford, and S. Shenker (2010). ''Dynamic Route Recomputation Considered Harmful.'' *Computer Communication Review*, Vol. 40, Issue 2, April 2010.

[12] T. Callahan, M. Allman, and V. Paxson (2010). ''A Longitudinal View of HTTP Traffic.'' Proceedings of the 11th Passive and Active Measurement Conference (PAM 2010), Zurich, Switzerland, April 2010.

[13] M. Casado, T. Koponen, R. Ramanathan, and S. Shenker (2010). ''Virtualizing the Network Forwarding Plane.'' Proceedings of the Workshop on Programmable Routers for Extensible Services of Tomorrow (PRESTO 2010), Philadelphia, Pennsylvania, December 2010.

[14] J. Caballero, P. Poosankam, C. Kreibich, and D. Song (2009). ''Dispatcher: Enabling Active Botnet Infiltration using Automatic Protocol Reverse-Engineering.'' Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS 2009), Chicago, Illinois, pp. 621-634, November 2009.

[15] T. Dooner, B. Stack, and M. Allman (2010). ''An Analysis of Internet Scanning.'' Presented at the Case Western Reserve University Intersections Undergraduate Research Symposium Poster Session, December 2010. http://www.icir.org/mallman/papers/scanning-intersections-poster-2010.pdf.

[16] M. Felegyhazi, C. Kreibich, and V. Paxson (2010). ''On the Potential of Proactive Domain Blacklisting.'' Proceedings of the Third USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 10), San Jose, California, April 2010.

[17] J. Franklin, V. Paxson, A. Perrig, and S. Savage (2007). ''An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants.'' Proceedings of ACM Computer and Communication Security Conference (ACM CCS), Alexandria, Virginia, pp. 375-388, October 2007.

[18] G. Friedland and R. Sommer (2010). ''Cybercasing the Joint: On the Privacy Implications of Geotagging.'' Proceedings of the Fifth USENIX Workshop on Hot Topics in Security (HotSec 10), Washington, D.C., August 2010. Also appeared as ICSI Technical Report TR-10-005, May 3, 2010.

[19] I. Ganichev, B. Dai, B. Godfrey, and S. Shenker (2010). ''YAMR: Yet Another Multipath Routing Protocol.'' *ACM SIGCOMM Computer Communication Review*, Vol. 40, Issue 5, October 2010.

[20] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica (2011). ''Dominant Resource Fairness: Fair Allocation of Multiple Resources in Datacenters.'' Proceedings of the 8th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2011), Boston, Massachusetts, April 2011.

[21] B. Godfrey, M. Schapira, A. Zohar, and S. Shenker (2010). ''Incentive Compatibility and Dynamics of Congestion Control.'' *ACM SIGMETRICS Performance Evaluation Review*, Vol. 36, Issue 1, pp. 95-106, June 2010.

[22] C. Grier, K. Thomas, V. Paxson, and M. Zhang (2010). ''@spam: The Underground on 140 Characters or Less.'' Proceedings of the 17th ACM Conference on Computer and Communications Security, Chicago, Illinois, pp. 27-37, October 2010.

[23] C. Grier, K. Thomas, V. Paxson, and D. Song. ''Design and Evaluation of a Real-Time URL Spam Filtering Service.'' In submission.

[24] B. Heller, D. Erickson, N. McKeown, R. Griffith, I. Ganichev, S. Whyte, K. Zarifis, D. Moon, S. Shenker, and S. Stuart (2010). ''Ripcord: A Modular Platform for Data Center Networking.'' Proceedings of ACM Special Interest Group on Data Communications Conference (SIGCOMM 2010), New Delhi, India, pp. 457-458, August 2010.

[25] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. Joseph, R. Katz, I. Stoica, and S. Shenker (2011). ''Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center.'' Proceedings of the 8th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2011), Boston, Massachusetts, April 2011.

[26] T. Koponen, M. Casado, N. Gude, J. Stribling, L. Poutievski, M. Zhu, R. Ramanathan, Y. Iwata, H. Inoue, T. Hama, and S. Shenker (2010). ''Onix: A Distributed Control Platform for Large-Scale Production Networks.'' Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation (OSDI 10), Vancouver, Canada, pp. 351-364, October 2010.

[27] C. Kreibich, N. Weaver, B. Nechaev, and V. Paxson (2010). ''Netalyzr: Illuminating The Edge Network.'' Proceedings of the Internet Measurement Conference (IMC 2010), Melbourne, Australia, November 2010.

[28] Z. Li, A. Goyal, Y. Chen, and V. Paxson (2009). ''Automating Analysis of Large-Scale Botnet Probing Events.'' Proceedings of the ACM Symposium on Information, Computer, and Communication Security (ASIACCS 2009), Sydney, Australia, pp. 11-22, March 2009.

[29] Z. Li, A. Goyal, Y. Chen, and V. Paxson (2011). ''Towards Situational Awareness of Large-scale Botnet Probing Events.'' To appear in *IEEE Transactions on Information Forensics and Security*, 2011.

[30] C. Muthukrishnan, V. Paxson, M. Allman, and A. Akella (2010). ''Using Strongly Typed Networking to Architect for Tussle.'' Proceedings of the 9th ACM Workshop on Hot Topics in Networks (HotNets-IX), Monterey, California, article no. 9, October 2010.

[31] B. Nechaev, M. Allman, V. Paxson, and A. Gurtov (2010). ''A Preliminary Analysis of TCP Performance in an Enterprise Network.'' Proceedings of the 2010 Internet Network Management Workshop/Workshop on Research and Enterprise Networking (INM/WREN '10), San Jose, California, April 2010.

[32] B. Nechaev, V. Paxson, M. Allman, and A. Gurtov (2009). ''On Calibrating Enterprise Switch Measurements.'' Proceedings of the 2009 Internet Measurement Conference (IMC 2009), Chicago, Illinois, pp. 143-155, November 2009.

[33] T. Ouyang, S. Ray, M. Allman, and M. Rabinovich (2010). ‘‘A Large-Scale Empirical Analysis of Email Spam Detection Through Transport-Level Characteristics.’’ ICSI Technical Report 10-001, January 2010.

[34] T. Ouyang, S. Ray, M. Rabinovich, and M. Allman (2011). ‘‘Can Network Characteristics Detect Spam Effectively in a Stand-Alone Enterprise?’’ Proceedings of the 12th Passive and Active Measurement Conference (PAM 2011), Atlanta, Georgia, March 2011.

[35] V. Paxson and M. Allman (2000). ‘‘Computing TCP’s Retransmission Timer.’’ Request for Comments 2988, Proposed Standard, November 2000.

[36] V. Paxson, M. Allman, H. K. Jerry Chu, and M. Sargent (2010). ‘‘Computing TCP’s Retransmission Timer.’’ Internet-Draft draft-paxson-tcpm-rfc2988bis-01.txt, work in progress, December 2010.

[37] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G.M. Voelker, V. Paxson, N. Weaver, and S. Savage (2010). ‘‘Botnet Judo: Fighting Spam with Itself.’’ Proceedings of the 17th Annual Network and Distributed System Security Symposium (NDSS Symposiom 2010), San Diego, California, March 2010.

[38] L. Popa, N. Egi, S. Ratnasamy, and I. Stoica (2010). ‘‘Building Extensible Networks with Rule-Based Forwarding.’’ Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation (OSDI 10), Vancouver, Canada, pp. 379-391, October 2010.

[39] R. Sommer and V. Paxson (2010). ‘‘Outside the Closed World: On Using Machine Learning For Network Intrusion Detection.’’ Proceedings of the IEEE Symposium on Security and Privacy 2010, Oakland, California, pp. 305-316, May 2010.

[40] R. Sommer, N. Weaver, and V. Paxson (2009). ‘‘An Architecture for Exploiting Multi-Core Processors to Parallelize Network Intrusion Prevention.’’ *Concurrency and Computation: Practice and Experience*, Vol. 21, Issue 10, pp. 1255-1279, July 2009.

[41] R. Sommer, N. Weaver, and V. Paxson (2010). ‘‘HILTI: An Abstract Execution Environment for High-Performance Network Traffic Analysis.’’ ICSI Technical Report TR-10-003, February 2010.

[42] M. Vallentin, R. Sommer, J. Lee, C. Leres, V. Paxson, and Brian Tierney (2007). ‘‘The NIDS Cluster: Scalable, Stateful Network Intrusion Detection on Commodity Hardware.’’ Proceedings of the International Symposium on Recent Advances in Intrusion Detection (RAID 2007), Queensland, Australia, September 2007.

[43] M. Walfish, M. Vutukuru, H. Balakrishnan, D. Karger, and S. Shenker (2010). ‘‘DDoS Defense by Offense.’’ *ACM Transactions on Computer Systems*, Vol. 28, Issue 1, No. 3, March 2010.

[44] M. Zaharia, D. Borthakur, J. Sarma, K. Elmeleegy, S. Shenker, and I. Stoica (2010). ‘‘Delay Scheduling: A Simple Technique for Achieving Locality and Fairness in Cluster Scheduling.’’ Proceedings of the Fifth European Conference on Computer Systems (EuroSys 2010), Paris, France, pp. 265-278, April 2010.

[45] M. Zhang and V. Paxson (2011). ‘‘Detecting and Analyzing Automated Activity on Twitter.’’ Proceedings of the 12th Passive and Active Measurement Conference (PAM 2011), Atlanta, Georgia, March 2011.

3 Algorithms

3.1 Introduction

During 2010 the three major research areas of the Algorithms group were statistical genetics, optimization, and computational molecular biology.

In statistical genetics, Bonnie Kirkpatrick worked with Eran Halperin, Richard Karp, and Shuai Li on the reconstruction of pedigree graphs by comparing the degree of Identity by Descent among extant individuals. As a byproduct of this research Shuai Li and Richard Karp developed an effecive heuristic for a novel clustering problem called the quasi-clique prediction problem. Halperin worked with Bogdan Pasaniuc on several projects including imputation using the coalescent, estimation of expression levels of homologous genes in RNA-SEQ experiments, and leveraging genetic variability across populations for the detection of causal variants. Halperin worked with Lucia Conde on a genome-wide association study of follicular lymphoma, and collaborated with a large number of groups in Europe on association studies for cardiovascular diseases.

The work in optimization covered a wide range of topics, both theoretical and practical. Jörg Lässig and Dirk Sudhold worked on evolutionary algorithms, including parallel evolutionary algorithms and hybridizations of evolutionary algorithms with local search and ant colony optimization. Lässig, Christian Hochmuth, and Stefanie Thiem worked on the application of different global optimization heuristics such as particle swarm optimization and genetic algorithms to solve real world logistics problems using simulation-based optimization (e.g., n-location inventory systems with lateral transshipments). Oliver Kramer studied evolutionary search in kernel based machine learning. Benjamin Satzger investigated control of large systems and networks and collaborated with Oliver Kramer and Jörg Lässig on approaches for adaptive control of smart energy grids. Matthias Mnich developed the theory of parametrized computational complexity

Luqman Hodgkinson and Richard M. Karp designed Produles, the first algorithm with linear running time in the size of the input, for the important problem in computational biology of detecting evolutionarily conserved multi-protein modules. Shuai Li made progress on two classical problems of protein alignment, known as the largest common point set (LCP) problem and the minimum aligned distance (MAD) problem. Richard Karp co-authored a method of using gene expression data and protein-protein interaction networks to discover pathways containing genes dysregulated in specific diseases [69]. Karp also co-authored the National Research Council Report ‘‘A New Biology for the 21st century.’’ [4]

3.2 Statistical Genetics

Analysis of Pedigree Graphs: In human genetics, recent relationships are modeled using a family tree, or pedigree graph. Traditionally, geneticists construct these graphs from genealogical records in a very tedious process of examining birth, death, and marriage records. Invariably mistakes are made due to poor record keeping or incorrect paternity information. As an alternative to manual methods, Bonnie Kirkpatrick and her coworkers addressed the problem of automatically constructing pedigree graphs from genetic data.

The genetic data of interest is single-nucleotide polymorphism (SNP)data, which are positions in the genome known to have nucleotide variation across the population. Humans are diploid individuals having two copies of each chromosome. Data for an individual can come in two forms, either haplotypes or genotypes. The haplotypes are two strings, each giving the sequence of nucleotides that appear together on the same chromosome. The genotypes, for each position in the genome, give an unordered set of nucleotides that appear. In particular the genotype is said to be ''unphased'' due to the lack of information about which nucleotide appears on which chromosome.

The most obvious way to reconstruct pedigrees from genetic data is to use a structured machine learning approach, similar to phylogenetic reconstruction. That method would involve a search over the space of pedigree graphs where the objective is to find the pedigree graph with the highest likelihood of generating the observed data. Unfortunately, this is not a good way to proceed for two reasons: the space of pedigree graphs is exponential, and the likelihood calculation has exponential running time. The likelihood calculation given genotype data was known to be NP-hard. In an attempt to address this problem for complex pedigrees, Bonnie Kirkpatrick and her collaborators used a Gibbs sampler to infer haplotypes from genotype data [18]. In a second attempt to use likelihood methods, this time for haplotype data, Kirkpatrick discovered another NP-hardness result [17].

Since likelihood-based approaches seemed completely infeasible, Kirkpatrick and her collaborators turned to a completely different approach. We focused on the problem of inferring relationships between a set of living individuals with available haplotype data. For convenience, we assumed that the inferred pedigree was monogamous without inter-generational mating. We were able to produce two heuristic and practical pedigree reconstruction methods, one for inbred pedigrees and the other for outbred pedigrees [19]. This work immediately revealed another important problem, that of evaluating the resulting inferred pedigree against a ground-truth pedigree. This can be done either by determining whether the two pedigrees are isomorphic or by finding the edit distance between the two pedigrees [20].

The *quasi-clique partition problem* is motivated by pedigree construction, where individuals (modeled as vertices) are to be classified into families (groups of sibling, referred to as quasi-cliques). We took a generative model in which the the true set of sibling groups is a union of cliques, and the observed noisy siblinghood data differs from the true siblinghood data by random perturbations consisting of edge insertions (false positives) and edge deletions (false negatives). The problem is then to reconstruct the most likely set of sibling groups from the noisy data. The problem of finding the most likely underlying set of sibling groups becomes a combinatorial clustering problem of a novel type: given a graph, find the ''closest'' graph which is a union of disjoint cliques. We devised a heuristic algorithm for this problem. For almost all of our test cases, the heuristic algorithm produced a solution with a higher objective function value than the value of the ''true'' solution, suggesting that further optimization may not lead to more accurate results. The output of the heuristic algorithm was found to have a 93/

Imputation using the Coalescent: (ICSI authors: Bogdan Pasaniuc and Eran Halperin). An important component in the analysis of genome-wide association studies involves the

imputation of genotypes that have not been measured directly in the studied samples. The imputation procedure uses the correlation structure (linkage disequilibrium) in the population to infer the genotype of an unobserved single nucleotide polymorphism (SNP). The LD structure is normally learned from a dense genotype map of a reference population that matches the studied population. In many instances there is no reference population that exactly matches the studied population, and a natural question arises as to how to choose the reference population for the imputation. We presented a Coalescent-based method that addresses this issue. In contrast to the current paradigm of imputation methods, our method assigns a different reference dataset for each sample in the studied population, and for each region in the genome. This allows the flexibility to account for the diversity within populations, as well as across populations. Furthermore, because our approach treats each region in the genome separately, our method is suitable for the imputation of recently admixed populations such as African Americans, where each position in their genome may originate from a different ancestral group. We evaluated our method across a large set of populations and found that our choice of reference data set considerably improves the accuracy of imputation, especially for regions with low LD and for populations without a reference population available as well as for admixed populations such as the Hispanic population. Our method has been published in the journal *Genetic Epidemiology* [51].

Genome-Wide Association Study of Follicular Lymphoma: (ICSI authors: Eran Halperin, Lucia Conde). We performed the first whole-genome statistical analysis for the detection of genetic variants that may be associated with follicular lymphoma. To do so, we used a three stage analysis, where in each stage we reduced the number of SNPs measured and increased the number of individuals studied; we found two new SNPs in the HLA region that are strongly associated with follicular lymphoma. The HLA region is a region in chromosome 6 which is related to the immune system; this discovery will lead to a better understanding of the disease mechanism and its relation to the immune system. Our work has been published in *Nature Genetics* [3].

Other projects and collaborations: We have collaborated with a large number of groups in Europe on association studies for cardiovascular diseases. Together with these groups we (ICSI authors: Eran Halperin) have recently published a paper in *Circulation: Cardiovascular Genetics* [53] on the design of large meta-analysis studies, and we have a paper in press in *Nature Genetics*, in which we report on the association of 15 new genes with coronary artery disease [60]. We are currently working on the association of SNPs with intermediate phenotypes that are known to be risk factors for coronary artery disease or myocardial infarction.

3.3 Optimization

Evolutionary Algorithms: Jörg Lässig considered the problem of selecting individuals in the current population in genetic algorithms for crossover to find a solution with high fitness for a given optimization problem. It has been shown that if one wishes to maximize any linear function of the final state probabilities, e.g., the fitness of the best individual

in the final population of the algorithm, then a best probability distribution for selecting an individual in each generation is a rectangular distribution over the individuals sorted in descending sequence by their fitness values. This gives new insight into the selection process of individuals of Genetic Algorithms and the dynamics of the optimization process.

Lässig and Dirk Sudholt have been working on the analysis of parallel evolutionary algorithms (EAs). They pioneered the first theoretical runtime analysis in this area. This publication won the best paper award at GECCO 2010 [39], in the Parallel Evolutionary Systems track. They presented a constructed function where parallelization was proven to be essential in a sense that a parallel evolutionary algorithm drastically outperforms comparable non-parallelized evolutionary algorithms. Experimental supplements for this function led to a second publication at PPSN 2010 [40]. In another line of research they looked at the speedup gained by parallelization, for various topologies of the underlying communication network. This has yielded another publication at PPSN 2010 that won the PPSN 2010 best paper award [41]. They also published a further paper at FOGA 2011 featuring an adaptive scheme for automatically choosing the number of processors [42].

Lässig, Christian Hochmuth, and Stefanie Thiem worked on the application of different global optimization heuristics such as particle swarm optimization and genetic algorithms to solve real world logistics problems using simulation-based optimization (e.g., n-location inventory systems with lateral transshipments). The strategy turned out to be able to optimize various large logistics systems which cannot be handled analytically.

Dirk Sudhold also worked on hybridizations of evolutionary algorithms with local search and ant colony optimization. Regarding hybrid evolutionary algorithms, he worked on an invited book chapter that deals with finding a good balance between evolutionary, global search, and local search methods. This includes a literature survey, a description of previous theoretical results on the complexity of local search, and a survey of his own previous contributions [63]. He also published a paper at ISAAC 2010 on analyzing the performance of a simple hybrid evolutionary algorithm for graph coloring in collaboration with Christine Zarges from TU Dortmund [66].

In theory of ant colony optimization he worked on a theoretical analysis for a stochastic shortest path problem, which appeared at GECCO 2010. To our knowledge this is the first running time analysis of a randomized search heuristic for a stochastic combinatorial problem. An extended version of the paper has been invited to a special issue in *Algorithmica* [12]. Three further submissions covered results for pseudo-Boolean optimization. The first one is another publication at GECCO 2010 that deals with the analysis of a new pheromone update scheme where the best solution found in one iteration is reinforced. The two publications at GECCO 2010, coauthored with researchers from Europe, were nominated for best paper awards in their respective tracks [50, 12]. The second paper on pseudo-Boolean optimization, to be presented at FOGA 2011, deals with the performance of a simple ant colony optimizer on the class of linear functions [21]. The third paper, also to be presented at FOGA 2011, is a single-authored paper where he applied techniques from Markov-Chain Monte Carlo (MCMC) for the analysis of ant colony optimization [64]. He greatly benefitted from an exciting lecture on MCMC at UC Berkeley by one of the leading experts in the field, Prof. Alistair Sinclair. We believe that these techniques will find further applications for randomized search heuristics.

In addition to these topics, Sudhold published two further papers at PPSN 2010. One is

a single-authored paper on a new method for proving lower bounds on the running time of a broad class of evolutionary algorithms [62]. The other paper deals with the performance of simple evolutionary algorithms on the class of (strictly) monotone pseudo-Boolean functions [5]. Finally, he revised and finalized a number of journal papers during his time at ICSI.

Control of Complex systems and Networks: Benjamin Satzger worked on goal-driven control of complex distributed systems. He developed an algorithm for self-configuring distributed systems using the constraint satisfaction paradigm. This allows a modeling of the desired system behavior with constraints and the use of a constraint satisfaction problem solver for system control.

A further approach for goal-driven control is automated planning. State-of-the-art planning algorithms are often based on heuristic search. Satzger worked on adaptive search heuristics based on regression. Experiments showed that the resulting heuristics provide qualitatively similar results compared to traditional ones, but the generation of a heuristic estimate is computationally more lightweight. In addition, collaborations among Oliver Kramer, Benjamin Satzger, and Jörg Lässig resulted in approaches to adaptive control of smart energy grids.

Optimizing Multi-Location Inventory Models and International Production Networks: Due to tougher competition companies must reduce cost and improve service. Thus, complex logistics systems are set up. To evaluate and to optimize systems with heterogenous locations and sophisticated demand processes, simulation-based optimization is a viable alternative to limiting, analytical models. We investigate which optimization technique is suited best for a hub-and-spoke model. In our experiments a Genetic Algorithm with elitism performs significantly competitively against Particle Swarm Optimization and ensemble-based Metropolis [37].

Furthermore, we use the Genetic Algorithm to tackle a general, extended multi-location inventory model with lateral transshipments to derive interesting insights [11, 37]. Transshipments are used to balance surplus and shortage. To optimize order and transshipment decisions, we generalize all existing transshipment rules and introduce the concept of forecasting demand. We show not only that this approach is viable. Under certain conditions, distinguished locations act as hubs redistributing to other locations. Thus, an underlying structure emerges.

In international production networks, transshipments are planned and coordinated in advance. The task is to maximize output and to optimize resource allocation. Especially in batch production, complicated alternative routings at the individual locations are defined. Switching flexibly between alternative routings is advantageous to allocate resources efficiently. Thus, complex interdependencies exist between the network layer and the location layer. The challenge is to resolve these interdependencies and to provide a transparent view of the model.

We present an intuitive, novel network flow model that reflects alternative routings at all locations and simultaneously connects these locations to a network [9, 10]. First, complex alternative routings are modeled by value streams. Second, by traversing this clear graphical representation, a linear program is set up. Using specific properties of the model,

substitution techniques reduce its complexity. The approach is suitable to support strategic decisions regarding potential locations and vertical integration by evaluating investment scenarios.

Evolutionary Search in Kernel-Based Machine Learning: Many machine learning methods use kernels. Kernels are functions with useful characteristics for the analysis of data spaces. Kernel-based techniques have grown into strong methods in solving classification, regression, and manifold learning tasks. Their optimization problems are usually solved using deterministic mathematical programming techniques. But there are good reasons to employ stochastic optimization in machine learning, of which four are (1) non-convex optimization problems, e.g., induced by non-semidefinite kernel functions, (2) noisy optimization problems, e.g., in real data observations, (3) non-differentiable loss functions, and (4) large data sets that afford approximation algorithms or parallelization. Stochastic search methods may help to overcome these problems. In particular, evolutionary computation has grown to a rich field of powerful methods for global optimization. They are embarrassingly parallelizable and thus fairly efficient search methodologies in distributed computing scenarios. Successful heuristic extensions have been proposed for special solution space conditions such as multi-modal, constrained, and multi-objective objective functions. These developments motivate the application of advanced evolutionary optimizers to kernel machines.

Evolutionary Kernel Regression: Regression has an important part to play in machine learning and data analysis. In our work we consider the intersection between evolution strategies, i.e., the $(\mu + \lambda)$-ES as well as covariance matrix adaptation (CMA) variants, and kernel regression, in particular the Nadaraya-Watson estimator. We have developed an evolutionary optimization approach for the Nadaraya-Watson estimator called evolutionary kernel regression (EKR) [25]. Kernel parameters and bandwidths are tuned with CMA evolution strategies. The approach is based on leave-one-out cross-validation and Huber's loss function. An experimental analysis on test functions known from evolutionary optimization and machine learning repositories shows that the introduced model is a robust bandwidth selection model and kernel shape optimizer. It can easily be extended to local models that adapt to local data space characteristics like varying data densities and noise. The application of the developed methods to real-world problems is another focus of our research. We have applied EKR to the prediction of energy consumption data [35].

Approximation of Equivalent Pareto-Subsets: In turn, machine learning can support evolutionary search. In many optimization problems in practice, multiple objectives have to be optimized at the same time. Some multi-objective problems are characterized by multiple connected Pareto-sets at different parts in decision space, also called equivalent Pareto-subsets, each able to cover the whole Pareto-front in objective space. We assume that the practitioner wants to approximate all Pareto-subsets to be able to choose among various solutions with different characteristics. We have proposed a clustering-based niching framework to approximate equivalent Pareto-subsets [31]. The clustering process assigns the population to niches, while the multi-objective optimization process concentrates on

each niche independently. Adaptive indicator-based extensions are introduced that allow to automatize the niching process. We presented a case-study based on rake-selection and the density-based clustering method DBSCAN. Both methods fit well into the framework, as rake-selection allows the application of self-adaptive step size control for a fast and efficient search process, and the density-based clustering method does not require an initial determination of the number of niches.

Computational Intelligence and Energy: Sustainability is very important due to increasing demands and limited resources. Many problem classes in sustainable energy systems are data mining, optimization, and control tasks. We demonstrated how techniques from computational intelligence can help to solve important task in sustainable energy systems. We have shown how statistically sound wind models can be estimated with kernel smoothing methods. Radial basis functions can be employed for wind resource visualization. Support vector machines turn out to be successful in forecasting wind energy. Monitoring of high-dimensional wind time series is possible with a self-organizing map approach. Slow driving features in wind time series can be detected with slow feature analysis. Furthermore, we have demonstrated how a learning classifier system evolves control rules for a virtual power plant with a simple demand side management model [34].

Parametrized Computational Complexity: Matthias Mnich works on exponential-time algorithms and parameterized complexity. Computational theory has shown that some problems can be too difficult to solve, and has tried to compensate for this either by finding an approximation to a problem, or by a brute force search -- that is, by checking all possible solutions to a problem. Exponential-time algorithms try to navigate these two imperfect methods by finding optimal solutions to such difficult problems. In related work, Mnich studied parameterized complexity. A problem facing theoretical computer scientists is that some problems that, in theory, are too difficult to solve are sometimes solved in practice. Parameterized complexity identifies what mechanisms are behind these theoretically impossible successes. Mnich looks at such issues in the context of computational game theory and decision science (which includes mapping voter decisions and looking at online markets).

3.4 Computational Molecular Biology

Linear-Time Algorithms to Detect Evolutionarily Conserved Multi-Protein Modules: Luqman Hodgkinson and Richard M. Karp designed Produles, the first algorithm with linear running time in the size of the input, for the important problem in computational biology of detecting evolutionarily conserved multi-protein modules. The input is a collection of protein-protein interaction networks, or interactomes, and a collection of protein homology relations. The underlying hypothesis is that conserved functional multi-protein modules exhibit high modularity in the interactomes due to modular organization of biological systems. As an interactome can be viewed as a graph with proteins as vertices and interactions as edges, a natural definition for maximizing modularity in interactomes is to maximize the fraction of interactions in which the module proteins participate that are

contained entirely within the module. This definition reduces to the problem of finding modules with low conductance in an interactome graph. General algorithms for finding sets of vertices with low conductance in a graph were modified to search only for small connected modules and to run in constant time that does not depend on the size of the input. The most important modification necessary to obtain constant time was to bound the degrees of proteins included in the modules over all intermediate steps of the algorithms.

The degree bounds were based on formulating and answering two questions. The first asks: what is the maximum degree of any protein in a module of size at most b with modularity at least d? The second asks: what is the maximum degree of any protein in a module of size at most b with modularity at least d such that the module formed by removing this protein has lower modularity? The motivation for the second question is that when searching for modules with high modularity, there may be proteins with such high degrees that it always improves the modularity to remove them from the module. Both questions have been answered precisely with two bounds, matched with corresponding families of tight examples to show that the questions have been answered optimally. Two bounds apply to the second question with a transfer between the bounds occurring at a value that is a function of b and d. The bounds derived from the second question, especially, are useful practically for Produles as they are sufficiently tight to apply in many cases of interest. The bounds and families of tight examples are easily mapped to corresponding bounds and tight examples for analogous widely applicable questions about graph conductance.

After finding reasonable module boundaries in one interactome, connected modules are found in other interactomes using protein homology to the proteins in the first interactome, refining connected components according to contribution of each protein to the modularity. This is a form of multiple validation for modules as the protein interactions are based on independent experiments in the various organisms. Luqman Hodgkinson and Richard M. Karp have defined a collection of evolutionarily-motivated graph theoretic measures of quality by which to evaluate conserved modules reported by algorithms such as Produles. Comparison with previous methods shows that Produles produces better modules by these evolutionarily-motivated measures of quality than the leading methods NetworkBlast-M [14] and Match-and-Split [49], and modules that are competitive with all previous methods. When applied with parameters to detect high-quality conserved modules in the interactomes for human and Drosophila, Produles returned a set of 29 modules with well-detected boundaries. Many of these modules correspond to known modules with known function that are known to be highly conserved, including, among many others, the TFIID general transcription factor for eukaryotic transcription and the proteosome. Produles is the first algorithm for this highly-studied problem that provably runs in linear time in the size of the input, where the input correctly includes not only the sizes of the interactomes but also the number of protein homology relationships. Because of its linear running time and efficient implementation, Produles is faster than previous methods, sometimes by orders of magnitude, and scales much better to large inputs. Moreover it can be applied with reasonable thresholds for protein similarity included in the input, allowing detection of conserved modules that would necessarily be missed by algorithms that cannot process large numbers of protein homology or protein orthology relationships due to non-linear dependence of the running time on this important component of the input.

Produles was designed in joint work by Luqman Hodgkinson and Richard M. Karp.

Protein Structure Alignment: Shuai Li has made progress on two classical problems of protein alignment, known as the largest common point set (LCP) problem and the minimum aligned distance (MAD) problem. In these problems a structure is modeled as a sequence of 3D points.

Both the LCP and MAD problems require us to identify two subsets of the same size, from two given structures, and a bijective mapping on the two subsets. The objective of the LCP problem is to maximize the subset size under the constraint of a given root-mean-square distance (RMSD) threshold θ, while the objective of the MAD problem is to find a subset of a given size ℓ which minimizes the RMSD. In some applications, a matching between the two point sets is given. That is, the bijection is to be a subset of the given matching. This problem is referred to as the model superposition problem. The complexities for these problems are unknown. We present a Fully Polynomial-Time Approximation Scheme for the MAD problem, and an $(\frac{\ell-1}{\ell})$-approximation algorithm for the LCP problem. In addition, we prove that both MAD and LCP are polynomially solvable for the model superposition problem. When the structures are proteins, we prove that both MAD and LCP are in P. The algorithm used to show this is also a pseudo-polynomial algorithm for both problems on general (non-protein) structures [44].

De Novo Discovery of Dysregulated Pathways in Human Disease: Molecular studies of the human disease transcriptome typically involve a search for genes whose expression is significantly dysregulated in sick individuals compared to healthy controls. Recent studies have found that in many human diseases, only a few genes show significant differential expression, but members of specific disease-related pathways are dysregulated in most sick individuals. These studies have defined a pathway as a set of genes known to share a specific function. However, pathway boundaries are frequently difficult to assign, and this approach cannot discover novel pathways. Protein interaction networks can potentially be used to overcome these problems. We present DEGAS (DysrEgulated Gene set Analysis via Subnetworks), a method for identifying connected genesubnetworks significantly enriched for genes that are dysregulated in sepcimens of a disease. These subnetworks provide a signature of the disease potentially useful for diagnosis, pinpoint possible pathways affected by the disease, and suggest targets for drug intervention [69].

References

[1] S. Bruckner, F. Hüffner, R. M. Karp, and R. Sharan (2010). ‘‘Topology-Free Querying of Protein Interaction Networks.’’ Journal of Computational Biology, Vol. 17, No. 3, pp. 237-252, March 2010.

[2] K. Chandrasekharan, R. M. Karp, E. Moreno Centeno, and S. Vempala (2011). ‘‘Algorithms for Implicit Hitting Set Problems.’’ Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA 2011), San Francisco, California, January 2011.

[3] L. Conde, E. Halperin, N. K. Akers, K. M. Brown, K. E. Smedby, N. Rothman, et al. (2010). ''Genome-Wide Association Study of Follicular Lymphoma Identifies a Risk Locus at 6p21.32.'' *Nature Genetics*, Vol. 42, No. 8, pp. 661-664, August 2010. DOI:10.1038/ng.626

[4] T. Connelly, P. Sharp, D. Ausiello, M. Bronner-Fraser, I. Burke, J. Burris, J. Eisen, A. Janetos, R.Karp, P. Kim, D. Lauffenburger, M. Lidstrom, W. Lim, M. McFall-Ngai, E. Meyerowitz, K. Yamamoto (2010). *A New Biology for the 21st Century*. Report of National Research Council, 2010.

[5] B. Doerr, T. Jansen, D. Sudholt, C. Winzen, and C. Zarges (2010). ''Optimizing Monotone Functions Can Be Difficult.'' Proceedings of the 11th International Conference on Parallel Problems Solving from Nature (PPSN 2010), Krakow, Poland, pp. 42-51, September 2010.

[6] B. Doerr, F. Neumann, D. Sudholt, and C. Witt (2011). ''Runtime Analysis of the 1-ANT Ant Colony Optimizer.'' To appear in *Theoretical Computer Science*, 2011. http://dx.doi.org/10.1016/j.tcs.2010.12.030.

[7] F. Gieseke, K. L. Plosterer, A. Thom, P. Zinn, D. Bomans, R.-J. Dettmar, O. Kramer, and J. Vahrenhold (2010). ''Detecting Quasars in Large-Scale Astronomical Surveys.'' Proceedings of the 9th International Conference on Machine Learning and Applications (ICMLA 2010), pp. 352-357, December 2010.

[8] T. Hein and O. Kramer (2010). ''Recognition and Visualization of Music Sequences Using Self-Organizing Feature Maps.'' Proceedings of the 33rd Annual German Conference on AI: Advances in Artificial Intelligence (KI 2010), Kralsruhe, Germany, September 2010.

[9] C. A. Hochmuth (2011). ''Modeling, Simulation, and Optimization of Complex Logistics Systems.'' Chemnitz University of Technology PhD Thesis, to be submitted.

[10] C. A. Hochmuth and J. Lässig. ''A Novel Network Flow Model for Optimizing International Production Networks.'' ICSI Technical Report, to appear.

[11] C. A. Hochmuth, J. Lässig, and S. Thiem (2010). ''Simulation-Based Evolutionary Optimization of Complex Multi-Location Inventory Models.'' Proceedings of the Third IEEE International Conference on Computer Science and Information Technology (ICCSIT), Vol. 5, Chengdu, China, pp. 703-708, July 2010.

[12] C. Horoba and D. Sudholt (2010). ''Ant Colony Optimization for Stochastic Shortest Path Problems.'' Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation (GECCO 2010), Portland, Oregon, pp. 1465-1472, July 2010.

[13] T. Jansen and D. Sudholt (2010). ''Analysis of an Asymmetric Mutation Operator.'' *Evolutionary Computation*, Vol 18, No. 1, pp. 1-26, Spring 2010.

[14] M. Kalaev, V. Bafna, and R. Sharan (2008). ''Fast and Accurate Alignment of Multiple Protein Networks.'' Proceedings of 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2008), Singapore, pp. 246-256, March 2008.

[15] R. M. Karp (2010). ''Implicit Hitting Set Problems and Multi-Genome Alignment.'' Keynote speech at the 21st Annual Symposium on Combinatorial Pattern Matching (CPM2010), New York, New York, June 2010.

[16] R. M. Karp (2011). ''Heuristic Algorithms in Computational Molecular Biology.'' Journal of Computer and System Sciences, Vol. 77, Issue 1, pp. 122-128, January 2011.

[17] B. Kirkpatrick (2010). ''Haplotypes Versus Genotypes on Pedigrees.'' Proceedings of the 10th Workshop on Algorithms in Bioinformatics (WABI 2010), Liverpool, United Kingdom, September, 2010.

[18] B. Kirkpatrick, E. Halperin, and R. M. Karp (2010). ''Haplotype Inference in Complex Pedigrees.'' *Journal of Computational Biology*, Vol. 17, No. 3, pp. 269-280, March 2010.

[19] B. Kirkpatrick, S. C. Li, R. M. Karp, and E. Halperin (2011). ''Pedigree Reconstruction Using Identity by Descent.'' To appear in the proceedings of the 15th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2011), Vancouver, Canada, March 2011.

[20] B. Kirkpatrick, Y. Reshef, H. Finucane, H. Jiang, B. Zhu, and R. M. Karp (2010). ''Algorithms for Comparing Pedigree Graphs.'' *CoRR*, abs/1009.0909, 2010.

[21] T. Kötzing, F. Neumann, D. Sudholt, and M. Wagner (2011). ''Simple Max-Min Ant Systems and the Optimization of Linear Pseudo-Boolean Functions.'' Presented at the Workshop on Foundations of Genetic Algorithms XI (FOGA 2011), Schwarzenberg, Austria, January 2011.

[22] O. Kramer. ''Analysis of Wind Energy Time Series with Neural Computing.'' Submitted to *Journal Neural Computing and Applications*, 2010.

[23] O. Kramer (2010). ''Anwendungspotenziale maschinellen Lernens in Smart Grids.'' OFFIS, Prof. Dr. Michael Sonnenschein, Carl von Ossietzky Universität Oldenburg, Germany, March 2010.

[24] O. Kramer (2010). ''Computational Intelligence and Sustainable Energy: Case Studies and Applications.'' ICSI Technical Report TR-10-010, November 2010.

[25] O. Kramer (2010). ''Covariance Matrix Self-Adaptation and Kernel Regression -- Perspectives of Evolutionary Optimization in Kernel Machines.'' *Journal Fundamenta Informaticae*, Vol. 98, No. 1, pp. 87-106, January 2010.

[26] O. Kramer (2010). ''Evolutionary Self-Adaptation -- A Survey of Operators and Strategy Parameters.'' *Journal Evolutionary Intelligence*, Vol. 3 No. 2, pp. 51-65, August 2010.

[27] O. Kramer (2010). ''Iterated Local Search with Powell's Method -- A Memetic Algorithm for Continuous Global Optimization.'' *Journal Memetic Computing*, Vol. 2, No. 1, pp. 69-83, March 2010.

[28] O. Kramer (2010). ''Multi-Objective Niching to Approximate Equivalent Pareto-Subsets.'' Invited talk at the Joint UC San Diego-Bauhaus University, Weimar Technology and Society Workshop, San Diego, California, September 2010.

[29] O. Kramer (2010). ''A Review of Constraint Handling Techniques for Evolution Strategies.'' *Journal of Applied Computational Intelligence and Soft Computing*, Volume 2010, Special Issue on Theory and Applications of Evolutionary Computation, Article ID 185063, 2010.

[30] O. Kramer (2011). ''Machine Symbol Grounding and Optimization.'' Proceedings of the Third International Conference on Agents and Artificial Intelligence (ICAART 2011), Rome, Italy, January 2011.

[31] O. Kramer and H. Danielsiek (2010). ''DBSCAN-Based Multi-Objective Niching to Approximate Equivalent Pareto-Subsets.'' Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), Portland, Oregon, pp. 503-510, July 2010.

[32] O. Kramer and F. Gieseke (2011). ''Short-Term Wind Energy Forecasting Using Support Vector Regression.'' Proceedings of the International Conference on Soft Computing Models in Industrial and Environmental Applications, Salamanca, Spain, April 2011.

[33] O. Kramer and T. Hein (2011). ''Monitoring of Multivariate Wind Resources with Self-Organizing Maps and Slow Feature Analysis.'' Submitted to the 2011 IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG), Paris, France, April 2011.

[34] O. Kramer, B. Satzger, and J Lässig (2010). ''Managing Energy in a Virtual Power Plant Using Learning Classifier Systems.'' Proceedings of the 7th International Conference on Genetic and Evolutionary Methods (GEM 2010), Las Vegas, Nevada, pp. 111--117, July 2010.

[35] O. Kramer, B. Satzger, and J. Lässig (2010). ''Power Prediction in Smart Grids with Evolutionary Local Kernel Regression.'' Proceedings of the Fifth International Conference on Hybrid Artificial Intelligence Systems (HAIS 2010), San Sebastian, Spain, pp. 262-269, June 2010.

[36] J. Lässig (2010). ''Analysis and Efficiency of Randomized Optimization Heuristics.'' *it-Information Technology*, Vol. 52, No. 6, pp. 345-349, December 2010.

[37] J. Lässig, C. A. Hochmuth, and S. Thiem (2011). ''Simulation-Based Evolutionary Optimization of Complex Multi-Location Inventory Models.'' To appear in *Variants of Evolutionary Algorithms for Real-World Applications,*'' R. Chiong, T. Weise, and Z. MichalewiczSpringer, eds., Springer, 2011.

[38] J. Lässig, B. Satzger, and O. Kramer (2011). ''Self-Stabilization in Hierarchically Structured Energy Markets.'' Proceedings of the 8th International Conference on Information Technology: New Generations (ITNG 2011), Las Vegas, Nevada, April 2011.

[39] J. Lässig and D. Sudholt (2010). ''The Benefit of Migration in Parallel Evolutionary Algorithms.'' Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation (GECCO 2010), Portland, Oregon, pp. 1105-1112, July 2010.

[40] J. Lässig and D. Sudholt (2010). ''Experimental Supplements to the Theoretical Analysis of Migration in the Island Model.'' Proceedings of the 11th International Conference on Parallel Problems Solving from Nature (PPSN 2010), Krakow, Poland, pp. 224-233, September 2010.

[41] J. Lässig and D. Sudholt (2010). ''General Scheme for Analyzing Running Times of Parallel Evolutionary Algorithms.'' Proceedings of the 11th International Conference on Parallel Problems Solving from Nature (PPSN 2010), Part 1, Krakow, Poland, pp. 234-243, September 2010.

[42] J. Lässig and D. Sudholt (2011). ''Adaptive Population Models for Offspring Populations and Parallel Evolutionary Algorithms.'' Proceedings of the Workshop on Foundations of Genetic Algorithms XI (FOGA 2011), Schwarzenberg, Austria, January 2011.

[43] J. Lässig and U. Trommler (2010). ''New Approaches to Enterprise Cooperation Generation and Management.'' Proceedings of the 12th International Conference on Enterprise Information Systems (ICEIS 2010), Madeira, Portugal, pp. 350-359, June 2010.

[44] S. Cheng Li (2010). ''Structure Alignment Under RMSD Measure.'' Submitted.

[45] S. C. Li, D. Bu, and M. Li (2011). ''Hexagon Codes Accurate Information for Side Chain Conformation.'' To appear in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2011.

[46] S. C. Li, D. Bu, J. Xu, and M. Li (2010). ''Finding Nearly Optimal GDT Scores.'' Accepted for publication in *Journal of Cell Biology*, 2010.

[47] S. C. Li and Y. K. Ng (2010). ''On Protein Structure Alignment Under Distance Constraint.'' Accepted for publication in *Theoretical Computer Science*, 2010.

[48] M. Martin, J. Maycock, F. Schmidt, and O. Kramer (2010). ''Recognition of Manual Motions with Dimension Reduction and Dynamic Time Warping.'' Proceedings of the

Fifth International Conference on Hybrid Artificial Intelligence Systems (HAIS 2010), San Sebastian, Spain, Vol. 1, pp. 221-228, June 2010.

[49] M. Narayanan and R. M. Karp (2007). ''Comparing Protein Interaction Networks via a Graph Match-and-Split Algorithm.'' *Journal of Computational Biology*, Vol. 14, Issue 7, pp. 892-907, September 2007.

[50] F. Neumann, D. Sudholt, and C. Witt (2010). ''A Few Ants are Enough: ACO with Iteration-Best Update.'' Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation (GECCO 2010), Portland, Oregon, pp. 63-70, July 2010.

[51] B. Pasaniuc, R. Avinery, T. Gur, C. F. Skibola, P. M. Bracci, and E. Halperin (2010). ''A Generic Coalescent-Based Framework for the Selection of a Reference Panel for ImputationGenetic Epidemiology.'' *Genetic Epidemiology*, Vol. 34, Issue 8, pp. 773-782, December 2010.

[52] B. Pasaniuc, N. Zaitlen, and E. Halperin (2010). ''Accurate Estimation of Expression Levels of Homologous Genes in RNA-Seq Experiments.'' Proceedings of the Fourteenth International Conference on Research in Computational Biology (RECOMB 2010), Lisbon, Portugal, pp. 397-409, April 2010.

[53] M. Preuss, I. R. Koenig, J. R. Thompson, J. Erdmann, D. Absher, T. L. Assimes, S. Blankenberg, E. Boerwinkle, L. Chen, L. A. Cupples, A. S. Hall, E. Halperin, et al. (2010). ''Design of the Coronary Artery Disease Genome-Wide Replication and Meta-Analysis (CARDIoGRAM) Study--A Genome-Wide Association Meta-Analysis Involving More than 22,000 Cases and 60,000 Controls.'' *Circulation: Cardiovascular Genetics*, Vol. 3, pp. 475-483, October 2010.

[54] J. Quadflieg, M. Preuss, O. Kramer, and G. Rudolph (2010). ''Learning the Track and Planning Ahead in a Car Racing Controller.'' Proceedings of the 2010 IEEE Symposium on Computational Intelligence and Games (CIG 2010), Copenhagen, Denmark, pp. 395-402, August 2010.

[55] R. Ronen, I. Gan, S. Modai, A. Sukacheov, G. Dror, E. Halperin, and N. Shomron (2010). ''miRNAkey: A Software for microRNA Deep Sequencing Analysis.'' *Bioinformatics*, Vol. 26, Issue 20, pp. 2615-2616, October 2010.

[56] B. Satzger, F. Bagci, and T. Ungerer (2010). ''A Novel Constraint Satisfaction Problem Solver for Self-Configuring Distributed Systems with Highly Dynamic Behavior.'' Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010), Barcelona, Spain, pp. 1-6, July 2010.

[57] B. Satzger and O. Kramer (2010). ''Learning Heuristic Functions for State-Space Planning.'' Proceedings of the Fifth IASTED International Conference on Computational Intelligence (CI 2010), Maui, Hawaii, pp. 36-43, August 2010.

[58] B. Satzger, O. Kramer, and J. Lässig (2010). ''Adaptive Heuristic Estimates for Automated Planning Using Regression.'' Proceedings of the 12th International Conference on Artificial Intelligence (ICAI 2010), Las Vegas, Nevada, pp. 576--581, July 2010.

[59] T. Sauerwald and D. Sudholt (2010). ''A Self-Stabilizing Algorithm for Cut Problems in Synchronous Networks.'' *Theoretical Computer Science*, Vol. 411, Nos. 14-15, pp. 1599-1612, March 2010.

[60] CARDIoGRAM Consortium: H. Schunkert, I. R. König, S. Kathiresan ... E. Halperin, et al. ''Large-Scale Association Analyses Identifies 13 New Susceptibility Loci for Coronary Artery Disease.'' *Nature Genetics*, in press.

[61] D. Sudholt (2010). ''Memetic Evolutionary Algorithms.'' To appear in *Theory of Randomized Search Heuristics -- Foundations and Recent Developments*, A. Auger and B. Doerr, eds., World Scientific.

[62] D. Sudholt (2010). ''General Lower Bounds for the Running Time of Evolutionary Algorithms.'' Proceedings of the 11th International Conference on Parallel Problems Solving from Nature (PPSN 2010), Krakow, Poland, pp. 124-133, September 2010.

[63] D. Sudholt (2010). ''Parametrization and Balancing Global and Local Search.'' Invited contribution to *Handbook of Memetic Algorithms*, C. Cotta, F. Neri, and P. Moscato, eds., Springer, 2011.

[64] D. Sudholt (2011). ''Using Markov-Chain Mixing Time Estimates for the Analysis of Ant Colony Optimization.'' Presented at the Workshop on Foundations of Genetic Algorithms XI (FOGA 2011), Schwarzenberg, Austria, January 2011.

[65] D. Sudholt and C. Witt (2010). ''Runtime Analysis of a Binary Particle Swarm Optimizer.'' *Theoretical Computer Science*, Vol. 411, No. 21, pp. 2084-2100, May 2010.

[66] D. Sudholt and C. Zarges (2010). ''Analysis of an Iterated Local Search Algorithm for Vertex Coloring.'' Proceedings of the 21st International Symposium on Algorithms and Computation (ISAAC 2010), Jeju Island, Korea, pp. 340-352, December 2010.

[67] S. Thiem and J. Lässig (2010). ''Particle Swarm Optimization: Theory, Techniques and Applications.'' In *Comparative Study of Different Approaches to Particle Wwarm Optimization in Theory and Practice*, A. E. Olsson, ed., Nova Science Publishers, September 2010.

[68] A. Thom and O. Kramer (2010). ''Acceleration of DBSCAN-Based Clustering with Reduced Neighborhood Evaluations.'' Proceedings of the 33rd Annual German Conference on AI: Advances in Artificial Intelligence (KI 2010), Kralsruhe, Germany, pp. 195-202, September 2010.

[69] I. Ulitzky, A. Krishnamurthy, R. M. Karp, and R. Shamir (2010). ''DEGAS: De Novo Discovery of Dysdregulated Pathways in Human Disease.'' *PLoS ONE*, Vol. 5, No. 10, e13367, 2010.

[70] N. Yosef, J. Gramm, Q. Wang, W.S. Noble, R. M. Karp, R. Sharan (2010). ''Prediction of Phenotype Information from Genotype Data.'' Communications in Information and Systems, Vol. 10, No. 2, pp. 99-114, May 2010.

[71] N. Zaitlen, B. Pasaniuc, T. Gur, E. Zic, and E. Halperin (2010). ‘‘Leveraging Genetic Variability Across Populations for the Identification of Causal Variants.’’ *The American Journal of Human Genetics*, Vol. 86, Issue 1, pp. 23-33, January 2010.

[72] Y. Zhao, B. Alipanahi, S. C. Li, and M. Li (2010). ‘‘Protein Secondary Structure Prediction Using NMR Chemical Shift Data.’’ *Journal of Bioinformatics and Computational Biology*, Vol. 8, No. 5, pp. 867-884, October 2010.

4 Artificial Intelligence and its Applications

In 2010, the Artificial Intelligence group made significant progress on both basic and applied projects linking language, cognition, and computation. One indication of the progress is the completion of two additional Berkeley PhD dissertations [8, 25] beyond several others completed recently [17, 56, 21, 54, 71]. Two other students in the group, Joshua Marker and Michael Ellsworth completed their doctoral qualifying exams.

The basic research of the group continues to be language learning and use, computational biology, and neural modeling. In 2010, the applied role was expanded to the areas of natural language processing (NLP) for developing regions, crowd-sourcing for semantic annotation, contextual inference for computing with natural language, and multilingual semantic resources.

The core scientific and technical work of the group is done within the three articulating efforts of the AI group. These are

1. The Neural Theory of Language (http://www.icsi.berkeley.edu/NTL) is a long-standing project investigating biologically plausible models of conceptual memory, language learning, and language use.

2. FrameNet (http://framenet.icsi.berkeley.edu) is an ongoing project led by Charles Fillmore that is building a semantically rich on-line lexicon based on the theory of Frame Semantics. The initial effort was an English Lexicon, but as described here, the effort has expanded to multiple languages.

3. Applications of the groups' research included the following efforts.

   (a) A new effort on ''Computing with Natural Language'' funded by a Basic Research Challenge (BRC) grant from the Office of Naval Research. ICSI is investigating the use of the NTL group's computational framework, Simulation Semantics, for natural language based interaction in a series of prototype systems in environments of increasing linguistic and situational complexity. This work will articulate with a new grant (starting in 2011) from the Templeton Foundation to research foundational issues in mathematics and artificial intelligence.

   (b) A collaborative project with Hesperian Press (http://www.hesperian.org) titled Multilingual Semantic Resource Development for Emerging Regions funded by Google, the Rockefeller Foundation, and the Bill and Melinda Gates Foundation.

   (c) Continuing exploratory efforts on metaphoric inference and metaphor corpus construction funded by NSF SGER and NSF CRI planning grants.

   (d) An effort exploring crowd sourcing techniques to scale up FrameNet funded by an NSF Exploratory Grant for Experimental Research.

   (e) A new collaboration with Professor Dan Klein on projection between languages without parallel corpora using unsupervised and semi-supervised machine learning techniques. The Klein group has been investigating the use of phylogenetic trees of language descent as an inductive bias and on joint induction of source (high resource) and target (low resource) grammars. The AI group (PI: Srini

Narayanan) will investigate semantic projection from well resourced to less resourced languages. The project is funded by a DARPA grant and is expected to be among the group's foci in 2011.

In all these cases, our main research goal is to use semantic tools and techniques developed by the the group to advance the automated analysis of information for a variety of tasks.

AI work continues to receive international recognition. Jerome Feldman gave a short course on NTL at the Autumn School in Dresden. He also gave lectures at the Max Planck Institute in Leipzig and the Medical Research Council Cognition and Brain Sciences Unit in Cambridge, England. At both of these labs, there was also discussion of ongoing collaboration on experiments on the neural basis of embodied language. This work is being actively pursued with Professor Jose Carmena at Berkeley. Feldman also participated in a Templeton Foundation Symposium on Top-Down Causality at the Royal Society in London. Srini Narayanan gave several lectures in Europe, the United States, and at IIIT Delhi in India. Narayanan was invited to be an ad-hoc reviewer for the European Science Foundation in addition to his regular panel reviews for the US National Science Foundation. Narayanan also helped organize the AAAI Fall Symposium on Computational Models of Narrative and was part of the program committee for the first conference on computing for global development held at the Imperial College in London, England.

In 2010, the AI group's work with Hesperian Foundation combining NLP and wiki-based collaboration techniques for providing multilingual health care information services was used by volunteer doctors in the aftermath of the Haiti earthquake to enter medical terms, diagnoses, treatments (with local remedies and with generic and brand name drugs), and primary health care information in Creole and provide bilingual links and translations from the Creole entries to English. This was an entirely voluntary effort and the results are continuing to be used by numerous local and international health workers who provide support and services to cope with massive devastation and disease outbreaks that followed the earthquake.

4.1 The Neural Theory of Language

The NTL project of the AI group works in collaboration with other units on the UC Berkeley campus and elsewhere. It combines basic research in several disciplines with applications to natural language processing systems. Basic efforts include studies in the computational, linguistic, neurobiological, and cognitive bases for language and thought and continues to yield a variety of theoretical and practical findings. In 2010, we made significant progress on all of these aspects.

The group has developed a formal notation for Embodied Construction Grammar (ECG), which plays a crucial role in larger, simulation based language understanding system. Jerome Feldman's book [28] on the NTL project was published by MIT Press in June 2006. The paperback version was released in 2008 and the electronic version in 2009. The book continues to be used in a number of courses at Berkeley and elsewhere.

It is been clear for some years that language understanding is best formulated as comprised of two major phases: analysis and simulation. One new finding is that the

analysis phase corresponds quite closely to public language of a community, providing a novel solution to a long standing problem [31]. Most of the research over the last few years has focused on either analysis or simulation, with the exception of the grammar learning work of Chang and Mok. A major initiative for 2010 is a unified project combining both phases, which will be further described below.

Another new initiative is the production and release of a new ECG wiki, which contains a tutorial and other pedagogical materials and also serves as a coordination point for grammar development and analysis. The wiki is at http://ecgweb.pbwiki.com/.

On the analysis side, one core NTL computational question is how to find the best match of constructions to an utterance in linguistic and conceptual context. The general computational point is that our task of finding a best-fit analysis and approximate answers that are not always correct presents a more tractable domain than exact symbolic matching. More importantly, our integrated constructions are decidedly not context-free or purely syntactic. In 2010, the crucial role of context was explored in several papers [31, 32].

One area of significant progress was the development of ECG implementations of complex constructions, mainly in English. A linguistics doctoral student, Ellen Dodge, has produced an elegant theory of conceptual compositionality and used it to illustrate basic facts about English argument structure as part of her 2010 dissertation [25]. Dodge's work will be covered as a chapter in the forthcoming book on computational approaches to construction grammar and frame semantics edited by Hans Boas [14].

A second chapter in [14] will present John Bryant's system for construction-based incremental sentence interpretation. Using psychologically plausible algorithms and a probabilistic syntax-semantics (best-fit) evaluation heuristic, he showed how a construction-based system of interpretation could be used to compute subtle semantic distinctions in English, interpret Mandarin child-parent dialogs, and predict processing time difficulty in a manner consistent with experimental data. An undergraduate honors project, by Luca Gilardi, added a beautiful graphical interface to the analyzer. This combined system has proven very valuable in our research and is also described. The resulting program plays an central role in the linguistics doctoral thesis of Ellen Dodge, and the completed computer science doctoral work of Eva Mok.

Feldman and Gilardi have extended the ECG implementation to handle Mental Spaces and Maps, which have been in the theory for some years, but not previously reduced to practice. This work forms a third chapter in [14].

A major recent addition ECG to handle morphology was an undergraduate honors project by Nate Schneider, who is now in graduate school at Carnegie Mellon University. While at CMU, Nate extended this work to also treat English morphology. This is described in another chapter of the forthcoming Boas book [14].

Nancy Chang and Eva Mok have continued developing representations and algorithms useful for an embodied approach to language acquisition and use. Chang has worked with colleagues to flesh out different aspects of a simulation-based approach to language understanding, including a formal representation for linguistic constructions. A version of the formalism was incorporated into her thesis research, which focused on the development of an algorithm that learns such constructions from a set of utterance-situation pairs. She completed her dissertation and is now at the SONY Paris lab. She presented this work at several European labs in 2010.

Eva Mok completed her dissertation research on a computational model of context-driven early grammar acquisition and is now at the University of Chicago. Grammar learning is a challenging problem for both children and machines because the target of learning - the grammatical structures - are hidden from the input. However, most accounts of grammar development underspecify the learning processes involved. Using ECG and extending the work of Nancy Chang, Eva's research represents a first step towards a unified computational model in which both the grammatical units and usage statistics are learned simultaneously from naturalistic, contextually-grounded Mandarin Chinese input. The work on grammar learning will also feature prominently in the forthcoming Boas book [14].

The group continues to be active in exploring the computational and scientific foundations of the Neural Theory of Language. Jerome Feldman published a review article describing the mythical neural code, covering both traditional and currently fashionable utility theory approaches [30]. The major effort of 2010 was an extensive review article [31] that focused on compositionality, the keystone problem in the semantics of language. This was accompanied by commentary from a wide range of experts and an author's response [32, 31].

For his dissertation [8], Leon Barrett focused on a new representation for performing action in the world. Such research is important for two reasons: first, the fundamental problem in computer science is how to represent data of any particular problem, since that determines which algorithms can be used; second, computers are particularly bad at dealing with the real world, both in perceiving and acting. His representation is specifically designed to deal with both the uncertainty and structure of actions in the real world, and he continues to develop proofs and demonstrations based on this core idea, linked to the CPRM model [65] that the group has been developing.

Ben Bergen continues to cooperate with the group after completing a UC Berkeley linguistics thesis using a statistical, corpus-based approach in combination with psycholinguistic experimentation, to explore probabilistic relations between phonology on the one hand and syntax, semantics, and social knowledge on the other. Ben has taken a new tenured position at UC San Diego, one of the leading institutions in the field. Bergen and Feldman completed a major invited book chapter showing how NTL helps explain the ancient mystery of how people can learn new concepts [10]. This is also being used in various classes.

Perhaps the most significant development of 2010 was the initiation of a new project integrating the analysis and simulation phases of ECG in a coherent system. This is being funded by a competitive grant from the Office of Naval Research and possibly additional sources. A preliminary pilot project is coupling the grammar from Dodge's thesis, the analyzer of Bryant and Gilardi, and the Robocup simulator of Barrett, all described above.

There were also several invited talks and seminars presented by NTL members. Jerome Feldman gave a short course on NTL at the Autumn School in Dresden. He also gave lectures at the Max Planck Institute in Leipzig and the Medical Research Council in Cambridge, England. At both of these labs, there was also discussion of ongoing collaboration on experiments on the neural basis of embodied language. This work is being actively pursued with Professor Jose Carmena at Berkeley.

Feldman also participated in a Templeton Foundation Symposium on Top-Down Causality at the Royal Society in London. His paper [32] will be published in a symposium

volume in 2011. Feldman has also been participating in a continuing dialog on connectionist and probabilistic models. His invited commentary [29] appeared in 2010 in *Trends in the Cognitive Sciences*. He also prepared an invited review [33] of the binding problem for the *Wiley Interdisciplinary Reviews*.

Srini Narayanan has begun to introduce NTL ideas into the first course in cognitive science at UC Berkeley. Ben Bergen and Nancy Chang gave a series of invited lectures on Introduction to Construction Grammar at the Summer School on Computational Linguistics in Cortona, Italy.

German post-doctoral fellow Birte Lönneker-Rodman and Narayanan contributed an invited article on computational models of figurative language [53] to the *Cambridge Encyclopedia of Psycholinguistics*. In 2010, the article was revised for final acceptance and is scheduled to appear in 2011. A collaborative effort between the FrameNet and NTL groups investigated the use of FrameNet for ontological inference reported in an Oxford University Press book on Lexical Resources and Ontologies in 2010 [70].

2010-2011 German post-doctoral scholar Malte Schilling started working with Narayanan on incorporating NTL models in human-robot interaction.

Srini Narayanan was invited to be a keynote speaker at a symposium on Computational Models of Narrative at Massachusetts Institute of Technology in Cambridge which brought together leading researchers in storytelling, interactive games, HCI, cognitive science, and AI. This symposium led to a AAAI Fall Symposium in 2010 that shows promise of bootstrapping a large new Government initiative on narrative and computation. Narayanan was a member of the organizing committee for the AAAI Fall Symposium and presented a paper (authored with George Lakoff) on the computational modeling of narrative [49].

4.2 Computing with Natural Language

In 2010, the AI group (PI: Srini Narayanan) won a five-year basic research challenge award from the Office of Naval Research to investigate the computational implications of simulation semantics to deal with vagueness and imprecision in natural language interaction.

Vagueness and imprecision in language manifests in multiple phenomena. Language can be uncertain, incomplete, imprecise, and ambiguous at all levels (phonological, syntactic, semantic, and pragmatic, and discourse). In addition, many languages and discourse types such as imperatives often omit arguments such as subjects, objects, instruments, and themes. To cope with this level of imprecision, the NTL group has developed algorithms that compute interpretations that are a *best-fit* [17, 56] across all the levels and in the situational and discourse context. Our approach, called simulation semantics, implements a computational model that combines best-fit construction analysis with dynamic simulation to monitor and propagate the effects of the utterance on the situational and discourse context. Simulation semantics is cognitively motivated and shows promise in handling the requisite phenomena better than any other approach that we are aware of [62, 49].

To evaluate the ability of the simulation semantics framework for contextual inference in the presence of vagueness and imprecision, we are implementing a system that can follow instructions expressed in natural language in a series of prototypes of increasing linguistic and situational complexity. This requires the system to analyze natural language and translate this language into a coordinated set of commands. In the initial phase

of the project, in 2011, a relevant command environment will be selected or designed. The evaluation will consist of naturally occurring imperative commands to be executed in this environment. The commands will be natural language instructions that must be interpreted by the system in the context of the sensed situation, previous utterances, and previous actions. The output of the system will be commands in the interface and environment API command language. The system will be judged based on the capabilities to interpret the natural language instructions, fill in missing information, and perform the right commands (in the context of the situation (right conditions, resources, goals, and outcomes)) and discourse (previous utterances, frames, and referents). We propose to evaluate both individual components as well as the overall system capability. We will report more on this exciting research in the coming years.

4.3 The Hesperian Digital Commons: A Multilingual Primary Health Care Resource

In 2010, the AI group (PI Srini Narayanan and graduate student Matt Gedigian) continued collaboration with Hesperian Foundation (http://www.hesperian.org) to investigate tools and frameworks to build a primary health digital commons that will extend the reach of health information to users with varied goals, literacy levels, multiple languages, devices, and modalities. The goal of our research is to produce information with contextually appropriate content, at the right granularity, in the appropriate language, at the right time, in the right modality, and delivered on the right device. Hesperian Foundation is a non-profit developer of of primary health materials used in over a 100 countries and adapted to over 100 languages to train health workers in violence-torn areas of Colombia, create community-based care for refugees in Thailand, provide support to children affected by HIV/AIDS in Africa, combat toxic poisoning from mining in the Philippines and support a host of other public health needs across the globe, at the community level.

As a first step toward our vision of universal health information access, we have been creating a repository or Hesperian Digital Commons (HDC) of materials that includes original Hesperian works on a variety of primary health topics. The project has been partly funded by a Google Research Grant to ICSI and by the Rockefeller Foundation through a planning grant to Hesparian Foundation. Specifically, our research has produced a set of tools and a novel collaborative editing framework with a primary health ontology and semantic annotations. This enables search and information access in multiple modalities and in multiple languages. The semantic representation also supports customized, content based *push* services such as on-demand pamphlets and book creation across materials and languages. http://www.hesperian.net has the current pilot prototype of the Hesperian Digital Commons in English, Spanish, and Tamil [63].

In 2010, the AI group's work with Hesperian Foundation combining NLP and wiki-based collaboration techniques for providing multilingual health care information services was used by volunteer doctors in the aftermath of the Haiti earthquake to enter medical terms, diagnoses, treatments (with local remedies and with generic and brand name drugs), and primary health care information in Creole and provide bilingual links and translations from the Creole entries to English. This was an entirely voluntary effort and the results are

continuing to be used by numerous local and international health workers who provide support and services to cope with massive devastation and disease outbreaks that followed the earthquake.

In addition, the HDC work was field tested by multiple groups around the world. Specifically, a) the Spanish version was evaluated in the field by health workers in Honduras, b) a health care NGO in Philippines tested the usability of the framework for development by building a Tagalog version, c) and a Lebanese group tested the customization functions in Arabic. The Haiti feedback and these field tests have given us valuable information and future directions to pursue in the next phase of our iterative design of the Hesperian Digital Commons. We hope to continue tests and report on the next design results in the coming year. Our hope is to have a fielded system in ten countries/regions covering five languages by the end of the next year.

4.4 An Annotated Metaphor Resource

In 2010, the NTL group started work on an NSF CRI program grant to focus on the issues of metaphor corpus selection, preparation, creation, and annotation. This was a follow-on grant to the NSF SGER grant (PI Srini Narayanan) in 2009 where the group completed preparatory work addressing the key scientific challenges posed by the creation of a metaphor inference system. The 2009 work focused on issues of metaphor representation and inference (please see the 2009 annual report).

The intention of the 2010 CRI planning grant is to use the theory of conceptual metaphor to plan the creation of an on-line metaphor resource that covers multiple types of language and corpora. In resolving issues related to corpus selection and preparation, our plan is to incorporate community wide evaluative feedback in the iterative design and implementation of metaphor annotations. In 2010, we designed a metaphor ontology and an algorithm for metaphor learning and annotation. In initial consultations with possible consumers of this resource, there was a need identified to perform a considerably deep analysis up-front of the conceptual relations, metaphor domains and syntactic relations in a multi-layer annotation scheme before the full value of the resource could become evident. This was the first time such a deep semantic resource was being seriously considered and members of the the user community we consulted felt that it would better enable them to assess the value and make specific recommendations on future directions. In response, rather than a first-cut corpus and annotation design evaluated in July 2010, we decided to come up with a comprehensive design and rich annotated corpus before holding the community-wide workshop. We are hoping to release the annotated corpus in the spring of 2011 and have a workshop in the summer of 2011.

4.5 Projection of Natural Language Resources

Modern statistical language processing systems require large quantities of carefully curated training data to function well. In the case of syntactic analysis tools, for example, this data comes in the form of hand-analyzed syntactic structures, which requires many dozens of person-years of effort to create and therefore exist for only a small number of economically prominent languages---indeed, many kinds of resources exist only for English.

Srini Narayanan and Dan Klein, along with UC Berkeley graduate students Taylor Berg-Kirkpatrick and David Burkett, have begun investigating methods for *language projection*, wherein linguistic resources in one language (often English) are transfered to other languages in an automatic way. In one project [12], a phylogenetic tree of language descent was used to construct a diachronic prior that ties languages' grammars together in proportion to their similarity. Using this prior, grammar induction on a collection of languages substantially improved grammar quality in each language compared to independent language modeling. In another project [18], syntactic analysis in two languages (English and Chinese) was improved by simultaneously modeling syntactic trees and language-to-language alignments. In particular, the system automatically learns which structures do align between the languages and which do not, robustly allowing for the realities of syntactic divergences. Finally, in order to capture cross-linguistic structure in a statistically efficient way, they showed a simple but general and effective method for tying together distributions using locally-parameterized features [13].

4.6 FrameNet ANC MASC Collaboration

Under a subcontract on NSF grant ''CRI: CRD: A Richly Annotated Resource for Language Processing and Linguistic Research'' (PI Professor Nancy Ide of Vassar), the FrameNet group is annotating both individual sentences and continuous texts from the American National Corpus (ANC) (`http://www.anc.org`) [45]. The ultimate goal is to combine the FrameNet annotation with other types of annotation on a large portion of the corpus, which is projected to grow to 100 million words. Obviously, manual semantic role labeling of a corpus of that size is unfeasible, so most of the annotation will have to be automatic. The current project includes the manual annotation on a small portion of the corpus, and training automatic semantic role labeling (ASRL) software to do the rest.

We continued our collaboration with the ANC in 2010, mainly annotating sample sentences for a growing list of words that the ANC team at Vassar and Columbia are annotating for WordNet senses. These are the same words that are being used in the WordNet-FrameNet Alignment pilot study, discussed in the next section. The FrameNet and WordNet annotations form part of the Manually-Annotated SubCorpus (MASC), which also includes automatically produced tokenization, part-of-speech (POS) tagging, sentence boundary detection, noun phrase chunking, and named entity recognition (NER). (cf. [43] for more details). The first release of the MASC data took place in 2010 (`http://www.anc.org/MASC`), and a second release, which will add several types of parsing, is expected soon.

Until now, the FrameNet work on MASC sentences has been performed by importing only the text and processing it using the standard FrameNet tools for tokenization, POS tagging, and NER. The FrameNet team is now completing work on a new import pipeline which will use as much of the information from the ANC as possible, so that the FrameNet annotation results will be more compatible with other annotations on the same texts.

4.7 WordNet-FrameNet Alignment

We have continued work funded by the NSF on aligning FrameNet with WordNet (WN) [55, 36], the largest machine-readable English lexicon, in collaboration with Professor Christiane Fellbaum. In 2010, we worked on detailed comparisons of WN and FN sense divisions for a total of roughly 80 common words, in part for the WN-FN alignment study and also to be used to annotate sentences for the ANC MASC project.

The results to date show that a real alignment of those lexical units that appear in both WN and FN will require changes to both resources. Since WordNet contains many more senses than FrameNet, it is likely that the majority of changes will be adding LUs to cover senses which already exist in WN; although in some cases, WN includes very rare uses of words which FN does not intend to add. Occasionally, by virtue of comparisons with FN, two or three WN senses will be collapsed, but in general, WN prefers to err on the side of finer word sense divisions [37].

Since it is obvious that only a small percentage of FN can be aligned this carefully, it will be necessary to find automatic means of aligning the bulk of the lexicon. A number of researchers are seeking algorithms for this purpose; among promising approaches are those of Oscar Ferrández [38] of University of Alicante and German Rigau [50] of UPV.

4.8 Crowd-Sourcing

We were approved for an Exploratory Grant for Experimental Research to test the waters in the new field of crowdsourcing some of the FrameNet annotation process, using different techniques, including Amazon's Mechanical Turk and online games. Our programmer Jisup Hong is developing the needed software to begin using the Amazon Mechanical Turk system (`http://en.m.wikipedia.org/wiki/Amazon_Web_Services`) to collect data for FrameNet over the Internet. Our ultimate goal is to collect full annotations of sentences in this way, but we began with a simpler task, i.e., asking workers to decide which frame a given instance of a polysemous lemma is in. For example, for the lemma *rip*.v, there are five LUs (word senses) in FrameNet, in the frames **Cause_to_fragment, Damaging, Removing, Self_motion**, and **Judgment_communication**. A sentence like *She ripped the telephone out of the wall* would be an instance of the **Removing** frame, while *You have ripped my jacket* would be in the **Damaging** frame.

Once we have a set of sentences that have been determined to be in a given frame, then we can reasonably ask people to annotate them using the Frame Elements of that frame. In other words, the task is broken down into two stages and we have been testing Mechanical Turk for the first stage. Previous research suggested that it is necessary (and sufficient) to gather roughly eight responses to each item and then to combine them to produce the equivalent of one response from an expert [73], which is still less expensive than hiring expert annotators, and a number of groups have presented results from this technique for various NLP tasks [19].

We are experimenting with different user interfaces and different ways of presenting the FN sense discrimination task, such as indicating the categories by the frame name, by the definition of the LU, or by giving a correctly categorized example, and we have tested it on words with differing degrees of polysemy. Our findings so far, however, indicate that

apparently many of the sense distinctions used in FrameNet are difficult for untrained annotators to make, so that even the result of combining annotations from eight or ten such workers is little better than chance on this task. We are also testing whether skipping the sense discrimination task and going directly to the frame semantic annotation task is feasible--it may be that seeing the frame elements available in a given frame will make it clearer whether the particular use of a lemma in a given sentence is or is not in a given frame.

4.9 Conference on Upgrading FrameNet

We received a grant for the purpose of holding a conference to discuss the future of FrameNet and issues such as aligning WordNet and FrameNet, expanding FrameNet more rapidly, and taking other steps to make it more useful to the NLP/computational linguistics community. The conference was held on May 1, 2010, with two venues, the UC Berkeley campus and the Princeton University campus, with audio and video links between them. By having half of the participants on the West Coast and the other half on the East Coast, we were able to hold what was ''virtually'' a single conference, but with reduced costs and travel time. There were eleven participants in Berkeley, seven in Princeton, and two in San Sebastian, Spain (who had heard about the conference and asked to be included); Dr. Tatiana Korelsky of NSF also phoned in to the conference for an hour or two.

Prior to the May 1 meeting, a wiki was set up to allow the discussion to begin beforehand; at the end of the meeting, there was a general discussion of conclusions and recommendations, which were posted to the wiki in real time. The participants stressed the importance of aligning WordNet and FrameNet to make them more useful in real-world NLP applications and of continuing the development of both resources, with annotated corpora. Participants further revised the wiki after the conference, and this material is being merged into a new website for FrameNet (now in alpha testing), which will include a forum for continued discussion and input from the NLP community on these topics.

4.10 Development of the FrameNet Database

During 2010, the number of frames increased by 67 to 1031; the number of lexical units (LUs) in the FrameNet database increased by 359 to 12,008. Of these, 194 were new LUs in existing frames; examples of these are shown in Table 1 on page 63. The other 165 new LUs were in the new frames; examples of these new frames and LUs are shown in Table 2 on p. 64.

Data Release 1.5: ¿From the summer through September of 2010, the FrameNet team worked intensively on preparing for a new data distribution, called release 1.5. The preparation involved a variety of efforts to test the data integrity and consistency of the FrameNet database, such as spell checking all frame, frame element (FE) FE, and LU definitions, checking to be sure that all FEs and LUs mentioned in a frame definition are in fact defined in the frame, and that definitions copied as part of the frame inheritance are modified to be appropriate to the current frame. We also checked all frames to ensure

that every FE defined in a frame is exemplified by at least one annotated sentence, that every lexical frames contains LUs, etc.

At the same time, a parallel effort was going on to develop a new XML format for the release, which would be accompanied by XSL/Javascript scripts for each XML format. This was successfully completed in September, and tested on the most popular current web browsers. This method allowed us to greatly reduce the number of different types of files we have to release, since users could simply browse the XML files, and view the different reports for each LU in the browser, with embedded links to go from one to another.

Since we completed these projects and began releasing the new version of the data in September, 2010, it has been downloaded by more than 850 users; most are NLP researchers or students in computational linguistics programs, but there is a wide range of others, including an increasing number of commercial users.

We also modified our internal wiki and the FrameNet public website to display these reports based on XML generated by the same report system. This made the process fast enough that we are now able to generate a complete new set of reports each night, so that the public website reflects new annotation within 24 hours.

4.11 Pending Proposals

- FrameNet is participating as a subcontractor on a new NSF CRI proposal to continue FrameNet annotation (similar to that described above) on the Multiply-Annotated SubCorpus (MASC) of the American National Corpus (Co-PIs: Professors Nancy Ide of Vassar and Rebecca Passonneau of Columbia),

- We have applied for a large NSF IIS grant for ''Information Integration and Informatics'' research entitled ''Talking about numbers: A common semantics for text and charts,'' with Dr. Collin Baker as PI and Dr. Gerald Friedland of ICSI and Professor Noah A. Smith of CMU as Co-PIs.

- FrameNet is also participating as a subcontractor on an SBIR grant from the US Army to Decisive Analytics Corporation (DAC) to develop frames, FEs, and LUs for frame semantic annotation of intelligence reports. The work is expected to begin in early 2011. ICSI does not handle any classified documents, but DAC will use the new frames developed at FrameNet in their in-house semantic parser.

4.12 Visitors and Events

Professor Carlos Subirats-Rüggeberg of Universidad Autónoma de Barcelona, head of the Spanish FrameNet project, was at ICSI for most of 2010; the development of Spanish FrameNet is continuing, closely based on the English FrameNet frames, cf. [76]. Subirats is also working on incorporating FrameNet frames and lexical units into the Embodied Construction Grammar parser (created by John Bryant and Eva Mok as part of their Ph.D. research, cf. [34]) in order to parse both English and Spanish.

In March, we had visits from two of our Japanese collaborators, Professor Kyoko Ohara of Keio University, PI of the Japanese FrameNet project, and Professor Hiroaki Sato of

Senshu University, developer of FrameSQL, as well as a longer visit from February to March from a student associated with Japanese FrameNet, Satoru Uchida, who demonstrated for us a sentence generator based in part on FrameNet (`http://realize.mints.ne.jp/FN/change.html`).

Professor Hinrich Schütze of the Institute for Natural Language Processing at the University of Stuttgart visited us in April to discuss the FrameNet representation of valency.

On July 28, Professor Jerry Hobbs of the Information Sciences Institute (ISI) at University of Southern California and Katja Ovchinnikova, a visiting scholar at ISI who recently presented a paper at the Language Resources Evaluation Conference on using FrameNet data for reasoning [66], came to ICSI to meet with the FrameNet staff and discuss changes in FrameNet that might make it easier to integrate FrameNet into inference systems.

On July 30, Dr. David Wible, a researcher at Academia Sinica in Taiwan, attended the FrameNet staff meeting and gave a demonstration of his online tools for discovering useful collocations in corpora, which may be of use to FrameNet in selecting examples to annotate. We hope to collaborate with Wible on the development of these tools.

Dr. Beatriz Sánchez Cardenas, a visiting scholar at UC Berkeley, attended FrameNet meetings with us from June to November and gave a presentation on her lexicographic work on numbers and counting in French and Spanish; she left to take a position at University of Granada, Spain.

Dr. Tuukka Ruotsalo, a researcher at Aalto University, Finland, arrived on September 1 for one-year stay as a visiting scholar at the iSchool at UC Berkeley and attends FrameNet staff meetings from time to time; his field is integrating cultural heritage texts into the semantic Web using XML markup.

Julia Ostanina-Olszewska, a visiting scholar at the UC Berkeley Department of Linguistics who works on translation and parallel corpora, visited FrameNet November 28 and 29 and talked with us about frame semantics across languages.

Dr. Sergio Guadarrama is a visiting scholar who arrived November 1 to begin working with Professor Lotfi Zadeh of the UC Berkeley Electrical Engineering and Computer Sciences Department on fuzzy logic/Computing with Words. He is attending FrameNet staff meetings, and will be a member of the ‘‘Talking about Numbers’’ project if it receives funding.

Dr. Malte Schilling joined the AI group in October 2010 as a DAAD-funded post-doctoral scholar. Schilling is working with Srini Narayanan on incorporating NTL models and simulation semantics in human-robot interaction.

Age	*mature.a, maturity.n, of.prep*
Aggregate	*force.n*
Ambient_temperature	*cold.n, freezing.a*
Architectural_part	*flight.n, landing.n, window.n*
Attention_getting	*hey.intj, officer.n*
Breaking_apart	*break.v, fragment.v, shatter.v, snap.v, splinter.v*
Bringing	*bus.v, shuttle.v*
Candidness	*devious.a, dishonest.a, earnest.a, earnestness.n*
Capability	*power.n, power_((statistical)).n*
Capacity	*feed.v, serve.v*
Cognitive_connection	*have to do_(with).v*
Coincidence	*chance.a, chance.n, chance.v, random.a, randomly.adv*
Destiny	*destined.a, doomed.a, fated.a, in the cards.adv, predestined.a*
Differentiation	*distinguishable.a, tell from.v*
Duration_description	*persistent.a, rapid.a*
Emotions_of_mental_activity	*groove.v*
Emptying	*expunge.v, flush.v, weed.v*
Experience_bodily_harm	*scrape.v, sunburn.v*
Filling	*dress.v, plank.v, wash.v, wax.v, yoke.v*
Frequency	*seldom.adv*
Go_into_shape	*coil.v, curl.v, twist.v*
Hiring	*contract.v, sign on.v, sign up.v, subcontract.v*
Immobilization	*straitjacket.v*
Judgment_communication	*hail.v, rip.v*
Locative_relation	*bracket.v, contact.v, meet.v, touch.v*
Manipulation	*handle.v*
Manufacturing	*assembly line.n, product.n*
Mental_property	*curious.a, suspicious.a*
Observable_body_parts	*knee.n, mustache.n*
Obviousness	*show up.v*
People_by_vocation	*magistrate.n, officer.n, police officer.n*
Physical_artworks	*image.n, poster.n*
Process_stop	*shut down.v, shutdown.n*
Quantity	*trace.n*
Resolve_problem	*handle.v*
Scrutiny	*pry.v, rifle.v*
Similarity	*image.n, take after.v, very image.n*
Storing	*cellar.v, stock.v*
Subjective_influence	*galvanize.v*
Take_place_of	*succeed.v, succession.n*
Taking_sides	*believe_(in).v*
Theft	*abstract.v, cop.v, rustle.v*
Typicality	*curious.a*
Willingness	*down.a*

Table 1: Examples of New Lexical Units Created in 2010 in Existing Frames

Appellations	*captain.n, general.n, officer.n*
Artifact_subpart	*part.n*
Assemble	*assemble.v, convene.v, meet.v*
Authority	*power.n*
Bond_maturation	*endow.v, mature.v, maturity.n*
Cause_bodily_experience	*chafe.v, itch.v, massage.v, rub.v, scratch.v, tickle.v*
Cause_to_be_included	*add.v, include.v*
Color_qualities	*dull.a, light.a, pale.a, vivid.a, warm.a*
Commutative_process	*add.v, addition.n, multiplication.n, multiply.v*
Commutative_statement	*product.n, sum.n, times.n*
Encounter	*encounter.v, stumble.v*
Exemplar	*epitome.n, exemplar.n, image.n, model.n, prototype.n*
Exercising	*exercise.n, exercise.v, work out.v*
Exhaust_resource	*devour.v, exhaust.v, expend.v, use up.v*
Heat_potential	*toasty.a, warm.a*
Impression	*image.n, persona.n*
Losing	*lose.v, misplace.v*
Losing_someone	*lose.v*
Losing_track_of_perceiver	*lose.v*
Losing_track_of_theme	*lose.v*
Meet_specifications	*fulfill.v, meet.v*
Meet_with_response	*meet.v*
Non-commutative_process	*divide.v, division.n, subtract.v, subtraction.n, take away.v*
Non-commutative_statement	*difference.n, minus.prep, quotient.n*
Opportunity	*break.n, chance.n, opportune.a, opportunity.n*
Optical_image	*image.n, shadow.n, silhouette.n*
Personal_success	*arrive.v, make it.v, succeed.v, success_((event)).n, success_((person)).n, success_((state)).n, successful.a*
Physical_strength	*force.n, power.n, strength.n*
Price_per_unit	*rate.n*
Proportion	*rate.n*
Race_descriptor	*asian.a, black.a, color.n, hispanic.a, white.a, yellow.a*
Rate_description	*fast.a, rapid.a, rapidly.adv*
Rate_quantification	*rate.n*
Relational_quantity	*rate.n*
Repayment	*pay back.v, repay.v, repayment.n*
Size	*big.a, enormous.a, huge.a, large.a, little.a, small.a, tiny.a*
Social_event_collective	*date.n*
Social_event_individuals	*go out with.v*
Subjective_temperature	*burn up.v, cold.a, cool.a, freezing.a, hot.a, warm.a*
System_complexity	*byzantine.a, complex.a, simple.a*
Time_period_of_action	*window.n*
Vehicle_subpart	*brake.n, door.n, engine.n, hood.n, part.n, seat.n, seatbelt.n, tire.n, trunk.n, wheel.n, window.n*

Table 2: New Frames (and Lexical Units) Created in 2010

References

[1] L. von Ahn and L. Dabbish (2004). ‘‘Labeling Images with a Computer Game.’’ Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2004), Vienna, Austria, April 2004.

[2] L. von Ahn, M. Kedia, and M. Blum (2006). ‘‘Verbosity: A Game for Collecting Common-Sense Facts.’’ Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2004), New York, New York, pp. 75-78, April 2004.

[3] S. Atkins, M. Rundell, and H. Sato (2003). ‘‘The Contribution of FrameNet to Practical Lexicography.’’ *International Journal of Lexicography*, vol. 16, Issue 3, pp. 333-357, September 2003.

[4] L. Aziz-Zadeh, C. Fiebach, S. Narayanan, J. Feldman, E. Dodge, and R. B. Ivry (2007). ‘‘Modulation of the FFA and PPA by Language Related to Faces and Places.’’ *Social Neuroscience*, Vol. 3, Issues 3-4, pp. 229-238, September 2008.

[5] C. F. Baker, M. Ellsworth, and K. Erk (2007). ‘‘SemEval-2007 Task 19: Frame Semantic Structure Extraction.’’ Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, pp. 99-104, June 2007.

[6] C. F. Baker and C. Fellbaum (2009). ‘‘WordNet and FrameNet as Complementary Resources for Annotation.’’ Proceedings of the Third Linguistic Annotation Workshop (LAW-III), Suntec, Singapore, pp. 125-129, August 2009.

[7] C. F. Baker, C. J. Fillmore, and B. Cronin (2003). ‘‘The Structure of the FrameNet Database.’’ *International Journal of Lexicography*, Vol. 16, Issue 3, pp. 281-296, September 2003.

[8] L. Barrett (2010). ‘‘An Architecture for Structured, Concurrent, Real-time Action.’’ UC Berkeley EECS dissertation, Berkeley, California, 2010.

[9] L. Barrett, J. Feldman, and L. Mac Dermed (2008). ‘‘A (Somewhat) New Solution to the Variable Binding Problem.’’ *Neural Computation*, Vol. 20, Issue 9, pp. 2361-2378, September 2008.

[10] B. Bergen and J. Feldman (2008). ‘‘Embodied Concept Learning.’’ In *Elsevier Handbook of Embodied Cognitive Science*, P. Calvo and T. Gomila, eds., pp. 313-332, Elsevier, 2008.

[11] B. Bergen, T. S. Lindsay, T. Matlock, and S. Narayanan (2007). ‘‘Spatial and Linguistic Aspects of Visual Imagery in Sentence Processing.’’ *Cognitive Science*, Vol. 31, Issue 5, pp. 733-764, September 2007.

[12] Taylor Berg-Kirkpatrick and Dan Klein (2010). Phylogenetic Grammar Induction. In proceedings of ACL 2010.

[13] Taylor Berg-Kirkpatrick, John DeNero, and Dan Klein (2010). Painless Unsupervised Learning with Features. In proceedings of NAACL 2010.

[14] H. C. Boas (to appear). *Computational Approaches to Construction Grammar and Frame Semantics*. Series on Constructional Approaches to Language, John Benjamins, 2011 (to appear).

[15] H. C. Boas (2005). ''Semantic Frames as Interlingual Representations for Multilingual Lexical Databases.'' *International Journal of Lexicography*, Vol. 18, No. 4, pp. 445-478, December 2005.

[16] H. C. Boas, ed. (2009). *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, Mouton de Gruyter, 2009.

[17] J. Bryant (2008). ''Best-Fit Constructional Analysis.'' UC Berkeley EECS dissertation, Berkeley, California, 2008.

[18] David Burkett, John Blitzer, and Dan Klein (2010), Learning Better Monolingual Models with Unannotated Bilingual Text. In proceedings of CoNLL 2010.

[19] C. Callison-Burch and M. Dredze, eds. (2010). Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk at the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL HLT 2010), Los Angeles, California, Los Angeles, California, June 2010.

[20] M. Castells (2009). *Communication Power*, Oxford University Press, 2009.

[21] N. Chang (2009). ''Constructing Grammar: A Computational Model of the Emergence of Early Constructions.'' UC Berkeley EECS dissertation, Berkeley, California, 2009.

[22] N. Chang and E. Mok (2006). ''Putting Context in Constructions.'' Proceedings of the Fourth International Conference on Construction Grammar (ICCG4), Tokyo, Japan, September 2006.

[23] N. Chang and E. Mok (2006). ''A Structured Context Model for Grammar Learning.'' Proceedings of the International Joint Conference on Neural Networks (IJCNN 2006), Vancouver, Canada, pp. 1604-1611, July 2006.

[24] R. S. Cook, P. Kay, and T. Regier (2005). ''The World Color Survey Database: History and Use.'' In *Handbook of Categorisation in the Cognitive Sciences*, Henri Cohen and C. Lefebvre, eds., pp. 223-242, Elsevier, 2005.

[25] E. Dodge (2010). ''Conceptual and Constructional Composition.'' UC Berkeley Linguistics dissertation, Berkeley, California, 2010.

[26] A. Dolbey (2009). ''BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology.'' UC Berkeley dissertation, Berkeley, California, 2009.

[27] M. Ellsworth and A. Janin (2007). ''Mutaphrase: Paraphrasing with FrameNet,'' Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing (TextEntail), Prague, Czech Republic, pp. 143-150, June 2007.

[28] J. Feldman (2006). *From Molecule to Metaphor: A Neural Theory of Language*, MIT Press, 2006.

[29] J. Feldman (2010). ''Cognitive Science Should Be Unified: Comment on: Griffiths et al. and McClelland et al.'' *Trends in Cognitive Sciences*, Vol. 14 No. 8, p. 341, August 2010.

[30] J. Feldman (2010). ''Ecological Expected Utility and the Mythical Neural Code.'' *Cognitive Neurodynamics*, Vol. 4, No. 1, pp. 25-35, March 2010.

[31] J. Feldman (2010). ''Embodied Language, Best-fit Analysis, and Formal Compositionality.'' *Physics of Life Reviews*, Vol. 7, Issue 4, pp. 385-410, December 2010.

[32] J. Feldman (2010). ''Simulation Semantics Can Revitalize the Formalization of Meaning: Reply to Comments on: Embodied Language, Best-fit Analysis, and Formal Compositionality.'' *Physics of Life Reviews*, Vol. 7, Issue 4, pp. 421-423, December 2010.

[33] J. Feldman. ''The Binding Problem(s).'' *Wiley Encyclopedia of Cognitive Science*, forthcoming.

[34] J. Feldman, E. Dodge, and J. Bryant (2010). ''A Neural Theory of Language and Embodied Construction Grammar.'' In *Oxford Handbook of Linguistic Analysis*, B. Heine and H. Narrog, eds., Oxford University Press, 2010.

[35] J. Feldman and S. Narayanan (2004). ''Embodied Meaning in a Neural Theory of Language.'' *Brain and Language*, Vol. 89, Issue 2, pp. 385-392, May 2004.

[36] C. Fellbaum, ed. (1998). *WordNet: An Electronic Lexical Database*.'' MIT Press, 1998.

[37] C. Fellbaum and C. Baker (2010). ''Aligning Verbs in Wordnet and Framenet.'' *Linguistics*, to appear.

[38] O. Ferrández, M. Ellsworth, R. Muñoz, and C. F. Baker (2010). ''Aligning FrameNet and WordNet Based on Semantic Neighborhoods.'' Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta, pp. 310-314, May 2010.

[39] C. J. Fillmore (1976). ''Frame Semantics and the Nature of Language.'' Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, volume 280, 20-32, 1976.

[40] C. J. Fillmore, C. F. Baker, and H. Sato (2004). ''Framenet as a 'Net.''' Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp. 1091-1094, May 2004.

[41] C. J. Fillmore, J. Ruppenhofer, and C. F. Baker (2004). ''FrameNet and Representing the Link between Semantic and Syntactic Relations.'' In *Computational Linguistics and Beyond: Language and Linguistics Monographs Series B*, C.-R. Huang and W. Lenders, eds., Institute of Linguistics, Academia Sinica Press, pp. 19-62, 2004.

[42] J. Hobbs and S. Narayanan (2003). ''Spatial Representation and Reasoning.'' In *Encyclopedia of Cognitive Science*, Nature Publishing Group, MacMillan, London, 2003.

[43] N. Ide, C. Baker, C. Fellbaum, and R. Passonneau (2010). ''The Manually Annotated Sub-Corpus: A Community Resource for and by the People.'' Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Short Papers, Uppsala, Sweden, pp. 68-73, July 2010.

[44] N. Ide, R. Reppen, and K. Suderman (2002). ''The American National Corpus: More than the Web Can Provide.'' Proceedings of the Third Language Resources and Evaluation Conference (LREC 2002), Las Palmas, Canary Islands, Spain, 839-844, May 2002.

[45] N. Ide, R. Reppen, and K. Suderman (2002). ''The American National Corpus: More Than the Web Can Provide.'' Proceedings of the Third Language Resources and Evaluation Conference (LREC 2002), Las Palmas, Canary Islands, Spain, pp. 839-844, May 2002. http://americannationalcorpus.org/pubs.html.

[46] P. Kay (2005). ''Color Categories Not Arbitrary.'' *Journal of Cross Cultural Research*, Vol. 39, pp. 39-55, February 2005.

[47] A. Kilgarriff and D. Tugwell (2001). ''WASP-Bench: An MT Lexicographers' Workstation Supporting State-of-the-Art Lexical Disambiguation.'' Proceedings of the 8th Machine Translation Summit (MT Summit VIII), Santiago de Compostela, pp. 187-190, September 2001.

[48] A. Kilgarriff and D. Tugwell (2001). ''Word Sketch: Extraction and Display of Significant Collocations for Lexicography.'' Proceedings of the workshop on COLLOCATION: Computational Extraction, Analysis, and Exploitation at the 39th Annual Meeting of the ACL and the 10th Annual Meeting of the EACL, Toulouse, France, pp. 32-38, July 2001.

[49] G. Lakoff and S. Narayanan (2010). ''Toward a Computational Model of Narrative.'' Proceedings of the AAAI Fall Symposium on Computational Models of Narrative, Arlington, Virginia, November 2010.

[50] E. Laparra, G. Rigau, and M. Cuadro (2010). ''Exploring the Integration of WordNet and FrameNet.'' Proceedings of the 5th Global WordNet Conference (GWC'10), Mumbai, India, January 2010.

[51] B. Lönneker-Rodman (2007). ''Beyond Syntactic Valence: FrameNet Markup of Example Sentences in 1a Slovenian-German Online Dictionary.'' Proceedings of Fourth

International Seminar on Computer Treatment of Slavic and East European Languages (Slovko 2007), Bratislava, Slovakia, pp. 152-164, October 2007.

[52] B. Lönneker-Rodman and C. F. Baker (2009). ''The FrameNet Model and Its Applications.'' *Natural Language Engineering*, Vol. 15, Issue 3, pp. 415-453, July 2009.

[53] B. Lönneker-Rodman and S. Narayanan (2010). ''Computational Models of Figurative Language.'' *Cambridge Encyclopedia of Psycholinguistics*, to appear.

[54] J. Makin (2008). ''A Computational Model of Human Blood Clotting: Simulation, Analysis, Control, and Validation.'' UC Berkeley EECS dissertation, Berkeley, California, 2008.

[55] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller (1990). ''Introduction to WordNet: An On-Line Lexical Database.'' *International Journal of Lexicography*, Vol. 3, No. 4, pp. 235-44, Winter 1990.

[56] E. Mok (2008). ''Contextual Bootstrapping for Grammar Learning.'' UC Berkeley EECS dissertation, Berkeley, California, 2008.

[57] E. Mok and J. Bryant (2006). ''A Best-Fit Approach to Productive Omission of Arguments.'' Proceedings of the 32nd Annual Meeting of the Berkeley Linguistics Society, Berkeley, California, February 2006.

[58] E. Mok and N. Chang (2006). ''Contextual Bootstrapping for Grammar Learning.'' Presented at the 28th Annual Conference of the Cognitive Science Society (CogSci 2006), Vancouver, Canada, July 2006.

[59] S. Narayanan (1997). ''KARMA: Knowledge-Based Active Representations For Metaphor and Aspect.'' UC Berkeley Computer Science Dissertation, Berkeley, California, 1997.

[60] S. Narayanan (2009). ''Testing the Predictions of a Computational Model of Metaphor.'' Proceedings of the International Conference on Cognitive Linguistics (ICLA 2009), Berkeley, California, July 2010.

[61] S. Narayanan (2010). ''Mind Changes: A Simulation Semantics Model of Counterfactuals.'' Accepted for publication in *Cognitive Science*.

[62] S. Narayanan. ''A Neurally Plausible Computational Model of Metaphor Acquisition.'' In preparation.

[63] S. Narayanan and M. Gedigian (2009). ''The Hesperian Digital Commons: A Multilingual Primary Health Care Resource.'' Proceedings of the Workshop on Computer Science and Global Development at the Computing Research Association Computing Community Consortium (CRA-CCC), Berkeley, California, August 2009.

[64] S. Narayanan and S. Harabagiu (2004). ''Question Answering Based on Semantic Structures.'' Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, August 2004.

[65] S. Narayanan, K. Sievers, and S. Maiorano (2007). ''OCCAM: Ontology-Based Computational Contextual Analysis and Modeling.'' *Modeling and Using Context*, pp. 356-368, Springer Berlin, 2007.

[66] E. Ovchinnikova, L. Vieu, A. Oltramari, S. Borgo, and T. Alexandrov (2010). ''Data-Driven and Ontological Analysis of FrameNet for Natural Language Reasoning.'' Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May 2010.

[67] S. Petrov, L. Barrett, and D. Klein (2006). ''Non-Local Modeling with a Mixture of PCFGs.'' Proceedings of the 10th International Conference on Conference of Computational Natural Language Learning (CoNLL-X), New York, New York, June 2006.

[68] S. Petrov, L. Barrett, R. Thibeaux, and D. Klein (2006). ''Learning Accurate, Compact, and Interpretable Tree Annotation.'' Proceedings of the 44th Annual Meeting of the Association of Computational Linguistics (ACL), Sydney, Australia, July 2006.

[69] T. Regier (1996). *The Human Semantic Potential*. MIT Press, 1996.

[70] J. Scheffczyk, C. F. Baker, S. Narayanan (2010). ''Reasoning over Natural Language Text by Means of FrameNet and Ontologies.'' In *Ontology and the Lexicon: A Natural Language Processing Perspective*, C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, and L. Prevot, eds., pp. 53-71, Cambridge University Press, 2010.

[71] S. Sinha (2008). ''Answering Questions about Complex Events.'' UC Berkeley EECS dissertation, Berkeley, California, 2008.

[72] S. Sinha and S. Narayanan (2005). ''Model-Based Answer Selection.'' Proceedings of the Workshop on Textual Inference for Question Answering at the 20th National Conference on Artificial Intelligence, Pittsburgh, Pennsylvania, July 2005.

[73] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng (2008). ''Cheap and Fast---But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks.'' Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Honolulu, Hawaii, October 2008. http://www.stanford.edu/ jurafsky/amt.pdf.

[74] C. Subirats (2009). ''FrameNet Español: Un Análisis Cognitivo del Léxico del Español.'' In *Terminología y Sociedad del Conocimiento*, A. Alcina, E. Valero, and E. Rambla, eds., pp. 309-320, Peter Lang, 2009.

[75] C. Subirats (2009). ''La Función del Corpus en FrameNet Español.'' Proceedings of the First International Conference on Corpus Linguistics (CILC 09), Murcia, Spain, pp. 1148-1155, May 2010.

[76] C. Subirats (2009). ''Spanish FrameNet: A Frame-Semantic Analysis of the Spanish Lexicon.'' In *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, H. Boas, ed., pp. 135-162, Mouton de Gruyter, 2009.

5 Speech Processing

Speech efforts in 2010 were headed by research staff members Dan Ellis (ICSI and Columbia University), Gerald Friedland, Dilek Hakkani-Tür, Jim Hieronymus, Adam Janin, Nikki Mirghafori, Nelson Morgan, Elizabeth Shriberg (ICSI and SRI), and Andreas Stolcke (ICSI and SRI). Additionally, our researchers collaborated heavily with colleagues at IDIAP in Switzerland, and at the University of Washington. We benefited greatly by our close collaboration with SRI International, not only via the efforts of Shriberg and Stolcke, but also by partnership with Gokhan Tur of SRI. Other domestic and international colleagues have played a critical role in our progress. We also partnered with IBM for work on speech recognition. There was an activity within the group that was not, strictly speaking, speech processing; namely, UC Berkeley Faculty Associate Dan Klein worked on improving machine translation (sometimes from the output of a speech recognition system) as part of our collaboration with DARPA contractor BBN.

Major contributions were made by students, research associates, postdoctoral Fellows, and international visitors. You can find a current list of group members, collaborators, and alumni at
http://www.icsi.berkeley.edu/groups/speech/members.html.

Activities in the group were primarily in the areas of speech recognition and understanding, speaker recognition and diarization, and multimodal speech processing. The sections below describe a number of the year's major activities in these areas. The list of topics is by no means exhaustive, but it should provide a useful overview of the major speech activities for the year.

5.1 Speech Recognition

Large Vocabulary ASR: In 2010 we continued to work with our multi-stream features for broadcast speech for the DARPA GALE project, focusing on Arabic in collaboration with colleagues at IBM. The recent focus in this work is experimentation with tonotopic networks of various architectures (hidden units per band in the first layer, number of hidden units in the second layer, context windows, etc.) on a 50-hour subset of the GALE Arabic data. These networks are intended as a replacement for our ''HATs'' networks, in that both approaches force some modeling power on the temporal sequence within narrow spectral bands. However, unlike ''HATs,'' which requires storage of huge files of intermediate results, the tonotopic networks can be trained and operated without significant storage requirements. Results were similar to previous experiments with other data sets - no architecture was better than 19 log critical bands with context window of 51 frames and roughly similar total number of hidden units in the two hidden layers. We are in the process of training larger nets on the full GALE training set.

Robust Small Vocabulary ASR: We have continued our efforts to understand how we can develop ASR methods that are insensitive to extrinsic sources of variability such as noise and reverberation, and intrinsic factors such as speaking rate and informal speaking style. One approach that we developed further was our many-stream Gabor filter-based methodology. In 2010 we saw that we could improve further on the ETSI standard AFE

method by using AFE to generate a spectral representation with a higher SNR, followed by the spectro-temporal filtering provided by the Gabor filters (and translated to discriminant features by MLPs) [29]. In another development, we explored the transformation of the filter outputs to MFCC-like features.

We also studied whether the use of the so-called ''Deep Belief Networks'' (DBNs) added any substantial gain over our typical single-hidden-layer MLP on the Aurora2 speech recognition task under mismatched noise conditions. Our findings suggest that DBNs outperform single layer MLPs under the clean condition, but the gains diminish as the noise level is increased [34]. Furthermore, using MFCCs in conjunction with the posteriors from DBNs outperforms merely using single DBNs in low to moderate noise conditions. MFCCs, however, do not help for the high noise settings. We are currently trying to learn whether it is the unsupervised pre-training step of the DBNs or the multiple layers that provide the gains we are seeing.

Data Collection for ASR: We are in the process of setting up the infrastructure to collect spontaneous meetings from participants' own laptops and cellphones in order to produce the ''Poor Quality'' Meeting Corpus. This corpus will represent a more accurate and natural corpus for training and testing of algorithms related to audio and video on meetings. Once the infrastructure is complete, we will collect a pilot corpus, annotate it, and hold a workshop to finalize plans with the community for a larger collection.

Exploratory Data Analysis for Speech Recognition: The central idea of this project is that if we want to improve recognition performance through acoustic modeling, then we should first understand how the current best model -- the hidden Markov model (HMM) -- fails to adequately model speech data. While the use of statistical modeling and estimation is ubiquitous in the field of speech recognition, the use of statistical diagnostics, exploratory data analysis, and maybe even hypothesis testing are not. Speech recognition researchers have been eager to import into their domain certain aspects of general statistical practice but, for some reason, not others. Interestingly, the parts of statistics that have been largely ignored have to do with diagnostics, the process of understanding how the properties of your data are at variance with your assumptions, and with inference, the process of formally drawing conclusions (is some particular technique useful or not?).

In particular, it has always been obvious that an assumption at the core of hidden Markov modeling, namely the statistical independence of frames, is false for speech data. In [36] we used resampling techniques to demonstrate that not only is the dependence in real speech data responsible for significant numbers of recognition errors, but it appears to be the single largest source of recognition errors in an HMM-based recognition system. Motivated by these results, this project is concentrating on developing a better understanding of the nature of the dependence in real speech data by studying how this data departs from an HMM, i.e., the residual of the data from the fitted HMM. This research is a form of exploratory data analysis, using a well estimated HMM to help us identify regions of model/data mismatch and to uncover structure in these regions.

We have just begun this project, so we are still in the process of setting up the necessary infrastructure. Our main accomplishment in 2010 is that we have put together an HMM-

based large vocabulary speech recognition system that uses open source software (HTK) for training and recognition. We have also identified and obtained a large corpus of transcribed audio from a single speaker which will allow us to remove between speaker variability from the model/data residual.

5.2 Speech Understanding

In 2010, the work done at ICSI on speech understanding focused on: multi-document summarization using topic models, speech-based automated cognitive status assessment, detection of socio-cultural content in language from multi-party meetings and broadcast conversations, and the study of the structure of meetings, particularly with regard to annotation methods. Additionally, we have begun an effort on situational awareness for emergency responders.

Multi-Document Summarization: Multi-document summarization (MDS) is the task of creating a short summary answering a user-formulated query, given a set of related documents. For newswire text, this task has been evaluated in the framework of Document Understanding Conferences (DUC), resulting in test beds that can be used for summarization research. We proposed two generative approaches for MDS: Summary Focused Topic Model (sLDA) [5] and hybrid hierarchical summarizer (HybHSum) [6].

The sLDA approach views sentences as mixtures of two separate sets of topics, summary and non-summary topic distributions, making it possible to formulate this task as the problem of discovering salient sentences that are likely to be included in summary text. We predict expected summary topic distributions in sentences and use them as latent output variables, along with the set of features characterizing sentences in documents to build a discriminative classifier model. The trained model is then used to infer summary latent distributions in sentences of test documents to construct a summary. Hence, the approach relies on two major steps to generate coherent and compact summaries: (1) capturing summary-focused contents from sentences in documents with a probabilistic model, and (2) building a classification-based approach for inferring content distributions in sentences of test documents, instead of building a probabilistic model for test documents. In this approach, we introduced a new parameter into the latent Dirichlet allocation (LDA) [3], that determines from which distributions the topics should be samples. Specifically, if a word exists in one of the manually formed summaries, then topics are samples from the summary topics, and otherwise from non-summary topics. In our experiments, we found that the sLDA approach is effective in comparison to the state-of-the-art methods.

HybHSum is similar to sLDA in that a generative model is used to capture summary-focused contents, and then a classification-based approach is used to characterize candidacy of a sentence for inclusion in the system-formed summary. The main differences are that a hierarchical model, hierarchical Latent Dirichlet Allocation (hLDA) [2], is used, and instead of using the new parameter, a sentence's belonging to the summary or not is determined using novel similarity measures that use the learned hierarchical topic structure. According to manual evaluations, this approach generates less redundant and more coherent summaries in comparison to other state-of-the-art methods.

In other work on multi-document summarization, we extended last year's work by adding machine learning to the problem of joint sentence selection and compression. We collected a new data set by asking annotators to delete words from 150-word summaries to create 100-word summaries, and used this data to train an end-to-end summarization system to directly maximize word overlap with the references, improving substantially on the state of the art [1]. We also developed a set of guidelines for evaluation of summaries by amateur annotators through Amazon's Mechanical Turk, showing that non-experts can reasonably reproduce expert judgements of linguistic quality, but not content, as poor grammar or organization often proves prohibitively distracting [16].

Socio-Cultural Content in Language from Speech-Enabled Technology in Multiple Languages: The aim of the IARPA-funded Socio-Cultural Content in Language Program is to discover the social goals of groups and their members by correlating these goals with automatically detected phenomena in the language used by the members. A further aim is to develop cross-cultural methodologies and software for providing insight into social dimensions of groups and their members.

The focus of our work in 2010 was on spoken broadcast conversations. We studied detection of speaker roles and agreements and disagreements in Arabic, English, and Mandarin broadcast conversations. We defined a paradigm with three levels of information: (1) low-level features, (2) linguistic phenomena, and (3) social constructs, such as speaker roles and social relations including agreements and disagreements. Low-level features are defined as individual and automatically extractable events or measurements. Examples of these features include start and end points of a talker's speech, presence of a word in automatic speech recognition output, duration of a pause, and pitch contour values. In some cases, a feature itself provides direct supporting evidence for a social construct (e.g., in ''I am a union member,'' the words directly inform social identity). But in many cases, there is good motivation for an intermediate, linguistically based category that serves as evidence for social constructs. We therefore refer to the evidence supporting social constructs as ''linguistic phenomena.'' A very simple example illustrating features and linguistic phenomena is the case of hedges (such as ''well ... I ...''), a linguistic phenomenon that tends to precede dispreferred responses. Simply the presence of the word ''well'' is ambiguous, as it can occur with other meanings. By considering the position of ''well'' in a turn (a feature with respect to speaker interaction patterns) and its prosody (features such as duration and pitch pattern), one can more accurately estimate whether it is an instance of a hedge. We manually annotated a set of linguistic phenomena that included discourse markers, prefaces, extreme case formulations, person mention, and addressing terms. Using the manual annotations as training data, we built automatic extractors, to use these as is or extract further features for detection of agreement and disagreements and social roles. We modeled the detection of agreement and disagreements as a sequence classifiction problem and used conditional random fields.

For inferring the social roles of speakers, we extended our previous work based on global speaker modeling by modeling the turn-taking behavior of the speakers with dynamic Bayesian networks (DBNs), which provide the capability of naturally formulating the dependencies between random variables. More specifically, we first explore the usefulness

of a simple DBN, namely, a hidden Markov model (HMM), for this problem. As it turns out, the knowledge of the segments that belong to the same speaker can be augmented into this HMM structure, which results in a more sophisticated DBN. This information places a constraint on two subsequent speaker roles such that the current speaker role depends not only on the previous speaker's role but also on the most recent role assigned to the same speaker. We conducted an experimental study to compare these two modeling approaches using broadcast shows. In our experiments using the English broadcast conversations data distributed by the Linguistic Data Consortium, the approach with the constraint on same speaker segments assigned 89.5% turns the correct role whereas the baseline HMM-based approach assigned 79.2% of turns their correct role. The details of this approach and experiments were presented at Interspeech 2010 [37].

Speech-Based Automated Cognitive Status Assessment: Cognition comprises mental functions such as the ability to think, reason, perceive, judge, and remember. These functions are often evaluated using various non-automated tests to assess the natural progression of cognitive decline, to study the need for care and the capacity of independent living, and to evaluate treatment [27]. Verbal interviews performed by trained clinicians are a common form of assessments to measure cognitive decline. We studied the usability of automated methods for evaluating verbal cognitive status assessment tests for the elderly. If reliable, such methods for cognitive assessment can be used for frequent, non-intrusive, low-cost screenings and provide objective and longitudinal cognitive status monitoring data that can complement regular clinical visits and would be useful for early detection of conditions associated with language and communication impairments.

Our study focused on two types of tests: a story-recall test, used for memory and language functioning assessment, and a picture description test, used to assess the information content in speech. Story retelling is part of the Logical Memory subtest of the Wechsler Memory Scale (WMS) [35] and has been useful to show degradation in memory skills related to very mild dementia of the Alzheimer type [21]. Picture description is included in standard aphasia evaluation instruments, such as the western aphasia battery (WAB) test [22].

Jointly with researchers from SRI International, we designed a data collection for this study involving recordings of about 100 people, mostly over 70 years old, performing these tests, and the data collection was funded by SRI. The speech samples were manually transcribed and annotated with semantic units in order to obtain manual evaluation scores. We explored the use of automatic speech recognition and language processing methods to derive objective, automatically extracted metrics of cognitive status that are highly correlated with the manual scores. We used recall and precision-based metrics based on semantic content units associated with the tests. Our experiments showed high correlation between manually obtained scores and the automatic metrics obtained using either manual or automatic speech transcriptions. More details on this work can be found in [17].

Meeting Structure: This past year, starting in July, Michael Ellsworth worked with Nelson Morgan and Dilek Hakkani-Tür on planning a workshop on dialog act annotation of meetings. The preparatory work included a literature review and a sample annotation of the ICSI Meeting Corpus using the Twente Argumentation Schema (TAS). TAS was

altered to allow arbitrary spans and a non-tree topology of dialog relations, since we were especially interested in whether the natural meetings of the ICSI Corpus would differ in structure from the scripted AMI Corpus meetings that TAS was designed for. A total of five hours of meetings were annotated by two annotators using the TAS annotation guide, but with special instructions to pay attention to how the annotation would affect downstream processing like summarization or action-item detection. Annotators were not forced to conform to each other when their annotation seemed reasonable, resulting in low agreement especially in chunking size (less than 50% exact span match, partially due to the fact that one annotator had 25% more annotation instances), but this was useful for promoting discussion at the workshop on whether there was a ''correct'' annotation.

As for the meeting itself, early in the process we determined a general timeframe (after the SLT conference in December of last year) and venue (ICSI) and a list of invitees, many of whom attended the workshop, which took place on December 16. There we presented a short literature review of dialog act annotation theories, proposals, and corpora, followed by a presentation of our sample annotation. Discussion was lively, and the resulting notes will be rendered into a post-workshop summary for the participants to respond and add to.

Speech-Based Situational Awareness: In late 2010, ICSI (together with UC Irvine and SRI International) was awarded a DARPA contract to work on Speech Based Situational Awareness for Firefighters. The FireTalk project will have three major parts, a situational awareness display for the fire chief and engine company chiefs, a robust speech recognition system to work in noisy firefighting environments, and a spoken language understanding system which will determine what part of the conversation involves information which would be useful for situational awareness. Another part of the system will be a series of telemetry streams which will provide additional situational awareness information, like the location, ambient temperature, and air supply remaining of every firefighter. High quality situational awareness involves the integration and maintenance of information about the dynamically changing environment of a fire.

This includes the location and severity of the fire, the location and air supply of each firefighter, a list of requests for actions and their status, the stability of the building including floors, ceilings, and roofs, and strategies for extinguishing the fire. Much of this information is communicated by speech and is not available by any other means. The difficulty for the fire chief is the large amount of information and requests which he or she has to monitor and use to plan the operation.

Since the quality of the speech on firefighting radio systems is not suitable for speech recognition, a second channel will be developed which consists of a cell phone streaming audio from the Bluetooth microphone in the breathing apparatus to the speech recognition platform. The first task is to collect speech from firefighters fighting real fires and in training exercises and have it transcribed for words and meaning. This will lead to a vocabulary and grammar for the language firefighters use which will be use in the ASR system, which will be developed by SRI. Noise canceling techniques will be developed specifically for firefighter breathing apparatus.

The meaning labels will be used to develop a semantic understanding component which will be based on machine learning techniques using bags of N-grams and unification

grammars. It will be necessary to be able to determine a request, a statement about fire conditions, equipment malfunction, dangerous conditions, etc. For example, if the firefighter reported ''the fire is out here,'' the system would know where the firefighter was and mark the fire out at that location on the situational awareness board. The Android smart phone which will be used in this project also has information on the location of the firefighter. Most phones now have GPS, but this is not useful inside buildings. It will be necessary to use inertial navigation within the building to locate the firefighter. The breathing apparatus has a digital readout of the remaining air supply, which can also be streamed through the smart phone. Having location and air supply information for each firefighter will facilitate the planning which the chief must do to keep his or her firefighters safe and effective. In the case of an emergency where the firefighter calls for help, having the location information will be extremely valuable.

The situational awareness board is a visual interface between the situational awareness system and the fire chiefs. UC Irvine has been developing situational awareness systems with displays for emergency responders and firefighters in the Center for Emergency Response Technologies, who are our partner in this project. They will add speech based information and individual firefighter locations to their Situational Awareness for Fire Fighters (SAFIRE) system. The awareness board will be redesigned to provide information derived from speech in a timely and uncluttered fashion. Several iterations will be required with feedback from the firefighters for the situation awareness board to be easy to use and optimally helpful in a fire.

The project has just begun; results will be reported in next year's document.

5.3 Speaker Recognition

In 2010 we continued our work on data selection for speaker recognition, as well as assessing the discriminability of speaker pairs. A visiting Spanish colleague, Professor Joaquin Gonzales, also began a new effort on forensic speaker recognition. Finally, we began a new project on robust speaker identification.

Open-Set Speaker Identification in Noise: In October of 2010, we started work on a new project funded by the Air Force Research Lab (AFRL) which focuses on robust open-set target/non-target speaker separation (i.e., speaker identification) systems in real environmental conditions. The challenges of the overall problem are: (1) environmental noise, (2) open-set speaker identification, (3) channel mismatch between landline and microphone channels, and, to a lesser extent, (4) accented English. The ROSSI database, specific to this task, has a variety of test and training conditions, such as noisy/matched channel/cross environment, for example, with training data from cell phone conversations in public places and testing data from cell phone conversations in moving vehicles and noisy/cross channel/cross environment, for example, with training data from roadside conversations placed on cell phones and testing data from calls made from a landline in an office.

The focus of work in this three-phase project is:

1. Benchmarking and evaluating ICSI's current SID systems for the ROSSI database,

2. Research on unit-based SID, including nasality, kurtosis, and Shannon's mutual information measures, and

3. Acoustic characterization and noise removal.

Our initial efforts so far have focused on benchmarking results on a baseline Gaussian mixture model-Universal background model (GMM-UBM) speaker identification system using verification equal error rate (EER), open set EER, and closed set accuracy measures. Additionally, we have been testing factor analysis subspace modeling, as well as applying kurtosis relevance measure to select relevant data for speaker recognition task. For three out of the five tested conditions (those involving cell-phone speech in the environmental conditions office, public place, and vehicle), the application of Kurtosis-based unit selection as well as factor analysis improved the EER. Short term future work will be on further explorations on the kurtosis measure (e.g., discarding data with kurtosis values in top 50% as opposed to top 25%), as well as benchmarking GMM-SVM SID system on all the test conditions of this database.

Structured Approaches to Data Selection for Speaker Recognition: In 2010, Howard Lei completed his PhD dissertation on structured approaches to data selection for speaker recognition [25]. The work involves the use of speaker-discriminability measures to determine temporal regions of speech (i.e., speech units) with potentially good speaker recognition performance. The measures investigated include mutual information, kurtosis, and a set of 11 nasality features. The measures are first computed using features vectors temporally constrained by a set of 30 mono-phones as the speech units, such that each phone has measure values associated with it. A speaker recognition system is run using data constrained by each of the 30 phones, and an EER is obtained for each phone. Correlations of the measure value and EER across the 30 phones indicate that mutual information and kurtosis are the best standalone measures (with correlations of 0.8 and 0.7 respectively).

Once the best standalone measures are determined, a data-selection algorithm is used to select arbitrary feature sequences as new speech units. Long phone N-gram sequences across all speech conversation sides are used for measure computation across many frame sequences within the phone N-gram, and the frame sequences with the highest measure values according to the mutual information and kurtosis measures are selected as new speech units. The ANN/HMM speech recognizer is used to train and decode the new speech units in all conversation sides, and speaker recognition systems are run on all new speech units. The EERs of the new speech units are in general better than the EERs of the existing mono-phone speech units. A relevance- and redundancy-based unit selection criterion is used to select units for MLP-based unit-combination. Results show that the use of the measures based on the nasality features allows for the selection of units that combine to produce lower EERs. The new speech units also combine to give lower EERs in general than the existing mono-phone units.

Assessing discriminability of speaker pairs In her PhD thesis work (estimated to be completed in 2011), Lara Stoll investigated the correlation of acoustic and prosodic features with various speaker attributes. In particular, she considered information collected about

the speakers in the 2008 NIST Speaker Recognition Evaluation data, namely the age, height, weight, number of years of education, and the age at which the speaker learned English (for those whose primary language is not English). Using multiple linear regression, she calculated the correlation of Mel-frequency cepstral coefficients MFCC features (coefficients 1-20) averaged over a phone or set of phones, or prosodic features such as formant frequency statistics and pitch statistics, with the aforementioned speaker attributes. The strongest correlations with MFCC features (over 0.7) were found for the speaker height, followed by the speaker weight; correlations with other attributes were very weak (less than 0.3). Of the prosodic features, the strongest correlations with speaker attributes occurred for formant frequency statistics, and these were weaker than the correlations of the attributes with MFCC features. One promising application of this work is to predict demographic information about a speaker from his speech; as an example, such information would be useful in profiling a suspect from a bomb threat phone call.

Previous work aimed to predict which impostor speaker pairs would be most difficult for automatic speaker recognition systems to distinguish, using features that can provide a measure of speaker similarity. With a goal of improving overall system performance, Stoll considered a few simple approaches for utilizing information about how difficult-to-recognize a speaker might be, with limited success. She is currently exploring more sophisticated methods that should be better able to make use of such knowledge.

Toward Robust, Transparent and Testable Forensic Speaker Recognition Through Feature-Based Likelihood Ratios: This research is focused on building bridges between the forensic speaker identification community and the automatic speaker recognition community. The current approaches for the former tend to be expert-based, performing a multi-faceted case-dependent analysis, where from a single piece of evidence (the analysis of the questioned and control recordings) one has to move from the specificities of speaker-dependent sound units to the highest linguistic levels (idiolects, sociolects, etc.). The main challenge of this expert-based approach is evaluating these expert judgments within a transparent and testable reporting scheme, thus properly validating the forensic expert's opinion. On the other hand, state-of-the-art approaches to speaker recognition, based on subspace variability analysis and compensation (including session, speaker, and channel) produce a single score, ideally a calibrated log-likelihood ratio. The main challenge of this ''automatic'' approach is to establish the correctness of the calibration of the log-likelihood ratio that is produced. The objective is to create and validate objective reporting tools that analysts can then use to provide individual strength-of-the-evidence corresponding to different sets of analysis units. This goal is strongly motivated by the National Academy of Science report entitled ''Strengthening Forensic Science in the US'' that highlights the paradigm shift in which present forensic science is involved, moving from the classical expert-based approach to evidence reporting, to a data driven statistical approach based on a transparent and testable estimation of the weight of the evidence, emulating the well grounded DNA approach which has become the new gold standard for forensic science reporting. Specific measures of individuality are the focus of the project, together with methods to best select and combine them, depending on what is possible for each case. Special attention is being paid to critical conditions appearing in real casework e.g., strong

mismatch in durations/channel/noise/reverb, speaking style, emotions, and other effects that are far from the conditions present at NIST SRE evaluations.

Several conclusions can be obtained from the work completed in the last four to five months of 2010:

- Speaker information is present in formant and bandwidths contours, our objective being to exploit it in conversational speech in a fully automatic unsupervised way
- Reasonable results in terms of calibrated likelihood ratios are obtained in realistic conditions with a very limited inventory of units
- A fast and efficient system has been developed, allowing for any phone, diphthong, or diphone to be processed and combined

5.4 Speaker Diarization

Speaker diarization is the task of partitioning an input stream into speaker homogeneous regions, or in other words, to determine ''who spoke when.'' This year, we continued our work on parallelization of speaker diarization and multimodal speaker diarization. In addition, we developed a new discriminative approach and investigated combination of different diarization systems.

Multimodal Diarization: While approaches to speaker diarization have traditionally relied entirely on the audio stream, the availability of accompanying video streams in recent diarization corpora has prompted the study of methods based on multimodal audio-visual features. This year, we worked on the use of robust video features based on oriented optical flow histograms. Using the state-of-the art ICSI diarization system, we show that, when combined with standard audio features, these features improve the diarization error rate by 14% compared to other multimodal diarization systems, including the system used for NIST RT'09 [23].

Parallelized Speaker Diarization: As part of UC Berkeley's ParLab project, we started to investigate parallel variants of agglomerative hierarchical clustering. Using GPU parallelism has the advantage of being able to use fine-grain parallel resources on many cores. However, the cores are less powerful and implementation restrictions, such as the lack of I/O operations and operating system calls, making it challenging to port code to a GPU. CPU parallelism, on the other hand, is easier to implement but there are significantly fewer cores and the current software solutions for implementing CPU parallelism do not allow for the same level of fine granularity as GPU tools. Therefore we discarded the initial CUDA implementation from last year and decided to go for a hybrid multicore/CUDA approach. The goal for parallelizing speaker diarization was to increase the speed without harming the accuracy of the system. We parallelized about 10k lines of code and brought down the runtime from 0.6 realtime to 0.07 realtime on an 8-core Intel CPU with an NVidia GTX280 card without affecting the accuracy.

A Discriminative Extension to Speaker Diarization: We developed a discriminative extension to agglomerative hierarchical clustering, a typical technique for speaker diarization, that fits seamlessly with most state-of-the art diarization algorithms. We use maximum mutual information using bootstrapping, i.e., initial predictions are used as input for retraining of models in an unsupervised fashion. We showed an absolute improvement of 4% DER with respect to the generative approach baseline[32].

Combination of Diarization Systems: System combination or fusion is a popular, successful, and sometimes straightforward means of improving performance in many fields of statistical pattern classification, including speech and speaker recognition. While there is significant work in the literature which aims to improve speaker diarization performance by combining multiple feature streams, there is little work which aims to combine the outputs of multiple systems. We worked on combining the outputs of two state-of-the-art speaker diarization systems, namely ICSI's bottom-up and LIA-EURECOM's top-down systems. We show that a cluster matching procedure reliably identifies corresponding speaker clusters in the two system outputs and that, when they are used in a new realignment and resegmentation stage, the combination leads to relative improvements of 13% and 7% DER on independent development and evaluation sets [4].

5.5 Multimodal Location Estimation

The project is a collaboration with ICSI's Computer Vision group. Location estimation is the task of estimating the geo-coordinates of the content recorded in digital media. Our project aims to leverage the GPS-tagged media available on the Web as training set for an automatic location estimator.

Ambulance Detection on Videos in the Wild: As a preliminary experiment, we considered a scenario that would be a common case for city-level location estimation: the classification of distinctive objects commonly found in cities, and, as an initial detailed case study, we focused on the classification of ambulances. Therefore, we collected 200 YouTube videos filmed in 11 cities, manually chosen to contain an ambulance. The data is inherently challenging as it derives from real users and is not recorded under controlled conditions. Our first task toward understanding location detection is thus limited to classifying which city an ambulance comes from. We used GMM/SVM classifier for the audio track and a SIFT/BOW SVM classifier for the image part of the videos. Multimodal integration was performed using early and late fusion followed by an SVM classifier. The maximum accuracy obtained on the problem was 47.72 %. The article was accepted in the highly selective brave-new-topics track at ACM Multimedia [14].

Participation in MediaEval: We developed the ICSI location estimation system, which participated in the European MediaEval 2010 Placing task [7]. The task was to automatically guess the location (latitude and longitude) of 5000 Flickr test videos using any or all of metadata, visual/audio content, and/or social information. Metadata contains user annotated title, tags, description, and the user's home location, among other information.

We submitted two different approaches that used the prior distribution of the training dataset in combination with GeoNames, a geographical gazetteer, for toponym (placename) resolution. Although currently only using metadata, both systems performed well among the top systems.

Audio/WiFi Indoor Localization: This research was performed in collaboration with Professor Ruzena Bajcsy's Tele-Immersion group at UC Berkeley. We developed a novel approach for indoor localization using multiple modalities of information that are typically available indoors: The presence of microphones in the devices that we carry in connection with various wireless signals sensed by current smartphones serve to indicate the location with about 3m accuracy [33]. Our proposed approach is computationally lightweight and, by making use of recent machine learning techniques for integrating modalities, achieves greater accuracy than current work in the area. Also the modalities truly complement each other: wireless signal localization is global as it indicates location in a specific building in the world with about room accuracy; audio localization is local with sub-room accuracy. The ACM Multimedia article [26] was selected finalist in the ACM Multimedia Grand Challenge 2011.

Privacy Implications of Geolocation: In collaboration with ICSI's Networking group, we also investigated the potential privacy issues of exact geo-location, especially as publicly available in the form as geo-tags [13]. We investigated several scenarios demonstrating the surprising power of combining publicly available geo-information resources for what we termed cybercasing: using online tools to check out details, make inferences from related data, and speculate about a location in the real world for questionable purposes. We were, e.g., able to find private addresses of celebrities as well as the origins of otherwise anonymized Craigslist postings. In one search of YouTube videos, we were able to find users with homes near downtown Berkeley that were vacationing and thus might be vulnerable to burglary given the subject matter of the videos they had posted.

The article resulted in a discussion in the international press, including *The New York Times* and ABC News.

5.6 Other Acoustic Processing

Robust Processing for Speech Activity Detection, Speaker ID, Language ID, Keyword Spotting: As part of our involvement in the DARPA RATS program, we have been looking at ways to characterize noisy and distorted speech. The program involves processing speech signals from marginal-quality radio reception, using narrow-band FM or single-sideband AM modulation. The impact of this on the resulting signal is a combination of additive noise, channel filtering (convolutive noise), and nonlinear distortion. In addition, the noise characteristics are often nonstationary, varying at both fast (impulsive) and slower (fading) time scales, and can vary greatly for different participants in a discussion. All these factors make this a very challenging style of data for conventional speech processing algorithms. In fact, it is even difficult to accurately characterize the data for the purposes of evaluating performance as a function of distortion level, and for the production of test

corpora intended to cover the full range of operational conditions. For this reason, we have investigated and developed a number of different measures of the speech quality for use with this data. These include:

- Conventional signal-to-noise ratio, based on an additive-noise assumption, and relying on voice-activity labels (e.g., from manual annotation)
- Blind SNR estimation based on energy histograms, which attempt to identify the ''noise floor'' as a low-energy mode (as realized in the NIST STNR tool).
- Blind SNR estimation based on amplitude distributions, exploiting the fact that noise generally has a Gaussian amplitude distribution, whereas speech has a distribution better characterized by a Gamma distribution (the WADA approach of Kim & Stern, Interspeech 2008).
- Objective speech quality measures based on the difference between auditory model envelopes for the clean and distorted speech, such as PESQ, a standard defined for voice codec evaluation (but which relies on having a clean reference speech signal).
- Decomposition of distorted speech into a component that is linearly dependent on the original, clean reference, and a residual composed of nonlinear distortion and additive noise. This has the advantage of describing the channel distortion as a by-product, but also relies on having a clean reference signal).

None of these approaches is entirely adequate on its own, either because its assumptions do not match our conditions, or because it relies on data beyond the distorted waveform that may not be available. We have produced a tool to calculate all of these measures, where available, which we have shared with the LDC to help them in producing simulated data for the task. By comparing the different measures against subjective ratings of speech quality for different recording and transmission conditions, it will be possible to calibrate the measures appropriately for each condition.

See http://labrosa.ee.columbia.edu/projects/snreval/ .

Feature Extraction for Sentence Segmentation: We continued research into discriminative training of pitch-related prosodic features, as compared to human-designed features prevalent in sentence segmentation systems. An attempt to implement an adaptation technique similar to fMPE from ASR did not pan out due to issues training a GMM system, which did not compare favorably with the current BoosTexter and SVM classifiers. Revisiting the LDA work performed earlier, a heteroscedastic LDA transform was tried. When combined with feature selection, the HLDA transformation produced a significant improvement over its input features and the baseline human-designed feature set, though not quite to the performance of the baseline features after feature selection.

5.7 Machine Translation

Statistical machine translation systems produce new translations by recombining fragments of old ones. These systems contain two primary sub-systems: a translation model, which

extracts reusable fragments from training data, and a language model, which scores the fluency of output sentences. UC Berkeley Professor Dan Klein and his students work on both components.

Translation models first analyze example translations into their component parts, then induce a correspondence between those parts. Standard approaches unnaturally partition examples into separate, non-overlapping fragments. However, real translation units tend to overlap, with relevant information encoded across the sentence in complex ways. Klein and his students developed a new translation modeling approach that explicitly represents overlapping features, improving final translation quality while maintaining an efficient pipeline.

Once the translation model has assembled and scored a possible translation, the language model also judges its quality. Language models score the plausibility of output sentences by comparing their parts to known word sequences. The more data that is used to train the language model, the better the overall output quality will be. However, using massive amounts of data requires slower methods, such as disk- or network-backed storage, which in turn slows down translation decoding. Klein and his students developed a fast, compact language modeling tool which can store -- in memory -- models derived from many billions of words of training data. Their system is not only many times more compact than the standard tools, but also substantially faster.

References

[1] T. Berg-Kirkpatrick, D. Gillick, and D. Klein (2011). ‘‘Learning to Jointly Extract and Compress.’’ Submitted to the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACM HLT 2011), Portland, Oregon, June 2011.

[2] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum (2003). ‘‘Hierarchical Topic Models and the Nested Chinese Restaurant Process.’’ Proceedings of the 17th Annual Conference on Neural Information Processing Systems, Vancouver, Canada, December 2003.

[3] D. M. Blei, A. Ng, and M. Jordan (2003). ‘‘Latent Dirichlet Allocation.’’ *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.

[4] S. Bozonnet, N. Evans, X. Anguera, O. Vinyals, G. Friedland, and C. Fredouille (2010). ‘‘System Output Combination for Improved Speaker Diarization.’’ Proceedings of the 11th International Conference of the International Speech Communication Association (Interspeech 2010), Makuhari, Japan, pp. 2642-2645, September 2010.

[5] A. Celikyilmaz and D. Hakkani-Tür (2010). ‘‘Extractive Summarization Using a Latent Variable Model.’’ Proceedings of the 11th International Conference of the International Speech Communication Association (Interspeech 2010), Makuhari, Japan, September 2010.

[6] A. Celikyilmaz and D. Hakkani-Tür (2010). ‘‘A Hybrid Hierarchical Model for Multi-Document Summarization.’’ Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden, pp. 1149-1154, July 2010.

[7] J. Choi, A. Janin, and G. Friedland (2010).‘‘ ‘‘The 2010 ICSI Video Location Estimation System.’’ Proceedings of the MediaEval 2010 Workshop, Pisa Italy, October 2010.

[8] J. Chong, G. Friedland, A. Janin, N. Morgan, and C. Oei (2010). ‘‘Opportunities and Challenges of Parallelizing Speech Recognition.’’ Proceedings of the Second USENIX Workshop on Hot Topics in Parallelism (HotPar ’10), Berkeley, California, June 2010.

[9] B. Favre, B. Bohnet, and D. Hakkani-Tür (2010). ‘‘Evaluation of Semantic Role Labeling and Dependency Parsing of Speech Recognition Output.’’ Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP 2010), Dallas, Texas, pp. 5342-5345, April 2010.

[10] G. Friedland, J. Chong, and A. Janin (2010). ‘‘Parallelizing Speaker-Attributed Speech Recognition for Meeting Browsing.’’ Proceedings of the 2010 IEEE International Symposium on Multimedia (ISM2010), Taiwan, pp. 121-128, December 2010.

[11] G. Friedland, J. Chong, and A. Janin (2010). ‘‘A Parallel Meeting Diarist.’’ Proceedings of the Workshop on Searching Spontaneous Conversational Speech (SSCS) at the ACM International Conference on Multimedia (ACM Multimedia 2010), Florence, Italy, 57-60, October 2010.

[12] G. Friedland and D. van Leeuwen (2010). ‘‘Speaker Recognition and Diarization.’’ In *Semantic Computing*, P. Sheu, H. Yu, C. V. Ramamamoorthy, A. K. Joshi, and L. A. Zadeh, eds., pp. 115-130, IEEE Press/Wiley, 2010.

[13] G. Friedland and R. Sommer (2010). ‘‘Cybercasing the Joint: On the Privacy Implications of Geotagging.’’ Proceedings of the Fifth USENIX Workshop on Hot Topics in Security (HotSec 10), Washington, D.C., August 2010. Also appeared as ICSI Technical Report TR-10-005, May 3, 2010.

[14] G. Friedland, O. Vinyals, and T. Darrell (2010). ‘‘Multimodal Location Estimation.’’ Proceedings of the ACM International Conference on Multimedia (ACM Multimedia 2010), Florence, Italy, pp. 1245-1251, October 2010.

[15] G. Friedland, C. Yeo., and H. Hung (2010). ‘‘Dialocalization: Acoustic Speaker Diarization and Visual Localization as Joint Optimization Problem.’’ *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 6, No. 4, Article 27, November 2010.

[16] D. Gillick and Y. Liu (2010). ‘‘Non-Expert Evaluation of Summarization Systems Is Risky,’’ Proceedings of the Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk at the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL HLT 2010), Los Angeles, California, pp. 149-151, June 2010.

[17] D. Hakkani-Tür, D. Vergyri, and G. Tur (2010). ''Speech-Based Automated Cognitive Status Assessment.'' Proceedings of the 11th International Conference of the International Speech Communication Association (Interspeech 2010), Makuhari, Japan, September 2010.

[18] D. Imseng and G. Friedland (2010). ''An Adaptive Initialization Method for Speaker Diarization based on Prosodic Features.'' Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Dallas, Texas, pp. 4946-4949, April 2010.

[19] D. Imseng and G. Friedland (2010). ''Tuning-Robust Initialization Methods for Speaker Diarization.'' *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, No. 8, pp. 2028-2037, November 2010.

[20] A. Janin, L. Gottlieb, and G. Friedland (2010). ''Joke-O-Mat HD: Browsing Sitcoms with Human Derived Transcripts.'' Proceedings of the ACM International Conference on Multimedia (ACM Multimedia 2010), Florence, Italy, pp. 1591-1594, October 2010.

[21] D. K. Johnson, M. Storandt, and D. A. Balota (2004). ''Discourse Analysis of Logical Memory Recall in Normal Aging and in Dementia of the Alzheimer Type.'' *Neuropsychology*, Vol 17, pp. 82-92, 2004.

[22] A. Kertesz (1982). *The Western Aphasia Battery*. Grune and Stratton, 1982.

[23] M. Knox and G. Friedland, ''Multimodal Speaker Diarization Using Oriented Optical Flow Histograms,'' Proceedings of the 11th International Conference of the International Speech Communication Association (Interspeech 2010), Makuhari, Japan, pp. 290-1184, September 2010.

[24] K. Lee, D. Ellis, and A. Loui (2010). ''Detecting Local Semantic Concepts in Environmental Sounds using Markov Model Based Clustering.'' Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Dallas, Texas, March 2010. http://www.ee.columbia.edu/ dpwe/pubs/LeeEL10-localconcepts.pdf

[25] H. Lei (2010). ''Structured Approaches to Data Selection for Speaker Recognition.'' UC Berkeley Dissertation, Berkeley, California, December 2010.

[26] E. Martin, O. Vinyals, G. Friedland, and R. Bajcsy (2010). ''Precise Indoor Localization Using Smart Phones.'' Proceedings of the ACM International Conference on Multimedia (ACM Multimedia 2010), Florence, Italy, pp. 787-790, October 2010.

[27] I. McDowell and C. Newell (1996). *Measuring Health: A Guide to Rating Scales and Questionnaires*. Oxford University Press, 1996.

[28] S. Ravuri and D. Ellis (2010). ''Cover Song Detection: From High Scores to General Classification.'' Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Dallas, Texas, pp. 65-68, April 2010.

[29] S. Ravuri and N. Morgan (2010). ''Using Spectro-Temporal Features to Improve AFE Feature Extraction for ASR,'' Proceedings of the 11th International Conference of the International Speech Communication Association (Interspeech 2010), Makuhari, Japan, pp. 1181-1184, September 2010.

[30] A. Stolcke, G. Friedland, and D. Imseng (2010). ''Leveraging Speaker Diarization for Meeting Recognition from Distant Microphones.'' Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Dallas, Texas, pp. 4390-4393, April 2010.

[31] C. Vaquero, O. Vinyals, and G. Friedland (2010). ''A Hybrid Approach to Online Speaker Diarization.'' Proceedings of the 11th International Conference of the International Speech Communication Association (Interspeech 2010), Makuhari, Japan, pp. 2642-2645, September 2010.

[32] O. Vinyals, G. Friedland, and N. Morgan (2010). ''Discriminative Training for Hierarchical Clustering in Speaker Diarization.'' Proceedings of the 11th International Conference of the International Speech Communication Association (Interspeech 2010), Makuhari, Japan, pp. 2326-2329, September 2010.

[33] O. Vinyals, E. Martin, and G. Friedland (2010). ''Multimodal Indoor Localization: An Audio-Wireless-based Approach.'' Proceedings of the Fourth IEEE International Conference on Semantic Computing (ICSC-2010), Pittsburgh, Pennsylvania, pp. 120-125, September 2010.

[34] O. Vinyals and S. Ravuri (2011). ''Comparing Multilayer Perceptron to Deep Belief Network Tandem Features for Robust ASR.'' Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), Prague, Czech Republic, May 2011.

[35] D. Wechsler (1954). ''Standardized Memory Scale for Clinical Use.'' *Journal of Pyschology*, Vol. 19, pp. 87-95, 1954.

[36] S. Wegmann and L. Gillick (2010). ''Why Has (Reasonably Accurate) Automatic Speech Recognition Been So Hard to Achieve?'' In arXiv.org under `arXiv:1003.0206 [cs.CL]`.

[37] S. Yaman, D. Hakkani-Tür, and G. Tur (2010). ''Social Role Discovery from Spoken Language using Dynamic Bayesian Networks.'' Proceedings of the 11th International Conference of the International Speech Communication Association (Interspeech 2010), Makuhari, Japan, September 2010.

6 Vision

The Vision group, led by Professor Trevor Darrell, studies fundamental methods for perception and recognition. In 2010, the group comprised postdoctoral researchers Mario Christoudias, Mario Fritz, Brian Kulis, Kate Saenko, and Mathieu Salzmann; DAAD postdoctoral visiting scholars Nicholas Cebron and Peer Stelldinger; and UC Berkeley EECS graduate students Ashley Eden, Allie Janoch, Yangqing Jia, Sergey Karayev, Trevor Owens, Alex Shyr, and Hyun Oh Song.

The Vision group's research focused on several key themes:

- Learning rich perceptual representations for early and mid-level vision
- Exploiting multiple modalities and context for effective object instance and category recognition
- Recognizing activities in multiple domains
- Learning transformations between domains

These themes are reflected in the following specific project efforts, listed below.

6.1 Projects

Learning Local Descriptors: One of the most successful and widely used developments in computer vision has been the rise of low-level local feature descriptors such as scale-invariant feature transform (SIFT). Such local feature descriptors allow compact yet discriminative coding of information about gradient orientations in small patches of the image. These features have successfully been used for scene and object recognition through representing densely extracted descriptors in terms of learned *visual words*, which are essentially cluster centers in descriptor space. On top of this quantized representation, more global image representations such as bags of words or pyramid-based representations are usually assembled.

Recent publications in the field have started re-evaluating the hard clustering approach of visual words in favor of ''softer'' representations that allow a single descriptor to be represented as a mixture of multiple components. The idea of such factorized representations is intuitively appealing, permitting higher robustness to noise in the data, and it has been shown that a soft factorization results in state-of-the-art performance on existing object recognition datasets, and allows good performance on novel datasets. Despite increased representative power, these local features are still input directly to global object representations. While this approach has yielded some of the best recognition performance to date, there are reasons to believe that developing intermediate visual representations could improve recognition performance by increased robustness to variance.

Mario Fritz and Sergey Karayev have been investigating in the past year a recursive Bayesian model to learn and represent visual features at different levels of complexity. Our model is based on Latent Dirichlet Allocation (LDA), a particularly interesting probabilistic factorization method that was originally formulated in the field of text information retrieval, and has recently achieved state-of-the-art performance on an object detection task [2]. The

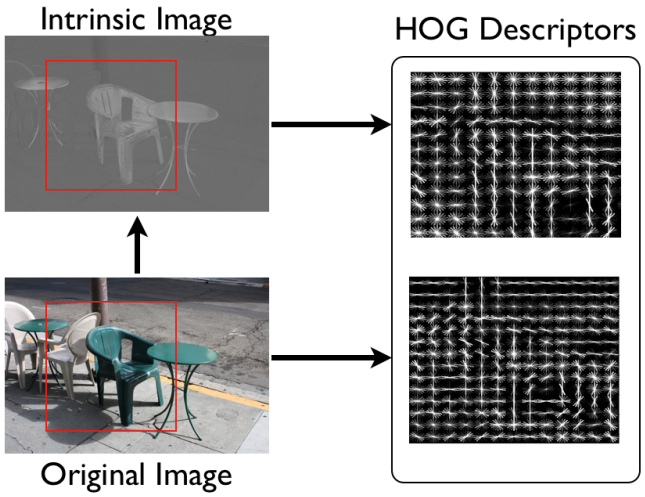

Figure 2: An overview of our main method (Original + Intrinsic). First, the original image is processed to create an intrinsic, shadow invariant image. Next, the two images are separately processed to compute two different HOG descriptors (the bounding boxes mark the area of the image shown in the descriptors). Finally, the feature vectors for each image are combined into a two-layer model.

approach is tested on standard recognition datasets showing state-of-the-art performance with respect to related models. Our results demonstrate two important properties of our proposed model: (1) adding an additional layer to the LDA representation increases performance over the flat model, and (2) a full Bayesian approach outperforms a feedforward implementation of the model.

Intrinsic Representations: Karayev, with Allison Janoch, has also investigated the idea of combining intrinsic images with traditional object detection in order to produce a more robust, lighting invariant detector. Intrinsic images and shadow detection have been the subject of much work in computer vision, yet these methods are largely unused in modern statistical object recognition methods based on feature histograms. Intrinsic images and other information about lighting and shadows may help to reduce confusion caused by variations in lighting. They may also allow detectors to take advantage of additional information conveyed by shadows. We proposed a novel descriptor based on Histograms of Oriented Gradients (HOGs) that combines feature vectors from the original image and its intrinsic light invariant image. This descriptor is tested in an object detection framework on PASCAL VOC 2009, and a novel shadow-prevalent dataset, where it shows increased average performance, especially in high-precision conditions. Figure 2 illustrates our approach.

Cross-Modal Learning: We investigated the problem of exploiting multiple sources of information for object recognition tasks when additional modalities that are not present in the labeled training set are available for inference. This scenario is common to many robotics sensing applications and is in contrast to the assumption made by existing approaches

that require at least some labeled examples for each modality. To leverage the previously unseen features, we make use of the unlabeled data to learn a mapping from the existing modalities to the new ones. This allows us to predict the missing data for the labeled examples and exploit all modalities using multiple kernel learning. We demonstrate the effectiveness of our approach on several multi-modal tasks including object recognition from multi-resolution imagery, grayscale and color images, as well as images and text. Our approach outperforms multiple kernel learning on the original modalities, as well as nearest-neighbor and bootstrapping schemes. This work was undetaken by Mario Christoudias, Mathieu Salzmann, Kate Saenko, and Mario Fritz. [1, 5]

Multimodal Latent Space Learnings: Salzmann and Yangqing Jia have focused on algorithms that reveal generative latent spaces for data with multiple modalities. One direction is to find factorized latent spaces that factorize the information into parts that are shared across all modalities, and parts that are private to each view. We proposed an efficient learning algorithm by imposing structured sparsity. Furthermore, the resulting factorized latent spaces generalize over existing approaches in that they allow having latent dimensions shared between any subset of the views. This approach outperforms state-of-the-art methods on the task of human pose estimation. The other direction is to find a shared latent topic space that connects the image and text; thus multi-modal image retrieval can be achieved from a loosely related text. We proposed a Markov random field of topic models that connects the documents from different modalities based on their similarity, and jointly learns the topics for all the modalities. Experiments on the Wikipedia data show that our model finds more semantically plausible cross-modality similarities than previous models [3].

Finding Lost Children: Christoudias and Ashley Eden developed a content-based image retrieval system based on relevance feedback and automatically extracted facial attributes for the speedy reunification of children with their parents in a large scale disaster. To our knowledge, our system is the first system for reunifying children and parents that utilizes image analysis to expedite the search. With our approach children are enrolled into a central database of lost children, their picture is taken, and attributes are automatically extracted. Parents can then efficiently search for their child by browsing the database using relevance feedback and their child's attributes. In collaboration with the Boston Children's Hospital our system was evaluated using real parents, where it resulted in relatively quick searches and significantly outperformed random search.

Multi-View Color Constancy: Color is known to be highly discriminative for many object recognition tasks, but is difficult to infer when only uncalibrated training images are available. Traditional methods for color constancy can improve surface reflectance estimates in uncalibrated images, but depend significantly on individual backgrounds. In many recognition and retrieval applications, we have access to image sets that contain multiple views of the same object in different environments; in this work Trevor Owens and Kate Saenko show that correspondences between these images provide important constraints that can improve color constancy. We introduce the multi-view color constancy

Figure 3: Examples from the DVDSBOOKS dataset, showing the illuminant corrected images, $x = M^{-1}y$, for the ground truth (GT), uniform, single image spatial correlations (SC) method and multiview spatial correlations method (MVSC). The number in brackets denotes the angular error of the illuminant estimate from ground truth. This is an example where the single image estimate of the illuminant is very poor, due to the uniform background, yet the multiple image estimate performs remarkably well.

problem, and present methods to recover estimates of underlying surface reflectance based on joint estimation of the surface properties and illuminants present in multiple images. Correspondences can be formed using a number of alignment methods; we perform matching of books and DVD covers using local region features. Our results show that multi-view constraints can significantly improve estimates of both scene illuminants and true object color (reflectance) when compared to baseline methods.

Probabilistic Segmentation Models: From conventional wisdom and empirical studies of annotated data, it has been shown that visual statistics such as object frequencies and segment sizes follow power law distributions. Using these two as prior distributions, the hierarchical Pitman-Yor process has been proposed for the scene segmentation task. In the group's line of research, Alex Shyr adds label information to the previously unsupervised Bayesian nonparametric model. This approach exploits the labeled data by adding constraints on the parameter space during the variational learning phase. We plan to extend our model to include class-dependent shape models and to perform object segmentation instead of scene segmentation. We are generalizing the models developed earlier by Erik Sudderth and colleagues. This work is in collaboration with Professor Michael Jordan of UC Berkeley and Dr. Raquel Urtasun of TTI-Chicago.

3D Descriptors: Peer Stelldinger worked to develop local 3D shape descriptors for surface meshes which can be efficiently computed on different scales. He considered the idea of an incremental descriptor generation based on the local mesh neighborhood. In order to compute such descriptors in an efficient way, we combine the local mesh data propagation with an efficient mesh simplification and remeshing algorithm. A randomized mesh simplification algorithm which runs in linear time regarding the number of mesh edges (i.e., it avoids the O(n log n) complexity of typical simplification algorithms based on priority queues) can easily be combined with the incremental descriptor generation. A non-optimized java implementation already allows to compute the complete scale space of a 1M points mesh in a few seconds. Moreover, a modification of the algorithm has been developed which allows the generation of a fast approximation of the skeleton of a 3D object based on a combination of mesh smoothing and mesh simplification.

Domain Adaptation: We have recently developed novel methods for domain adaptation in computer vision in order to address the problem that the fundamental assumption of supervised machine learning is more often than not violated, and that the distribution of test examples differs significantly from that of training data. We introduced this paradigm for the problem of object category recognition in a recent paper at ECCV2010, proposing a new approach based on metric learning; we further developed a new technique based on generic transform learning which learns an optimal projection of training data from one domain (e.g., Web images) to a target domain (e.g., an office environment where a robot may try to recognize an object). Most critically, our method works when there is no training data for a target class in a new environment: a transformation between domains is learned based on available data in both domains, and then can be applied to improve recognition of an object in an environment based solely on training data external to that environment (compared to baselines using that training data in unaltered form.) [5]

In the past year, ICSI researchers have collaborated with Professor Todd Zickler at Harvard University on the integration of machine learning and ''physics-based'' vision methods, in particular tackling the problem of color constancy from a new, multi-view perspective. The key insight our technique provides is to frame an object-centered approach to color constancy, rather than an image-centered approach. We consider object instances (and later, categories) in view in multiple uncalibrated images (such as found on the Web), and use the color correspondences (implicit or explicit) to constrain the estimate of illuminant color in each image. Informally, when one sees a single image of an object, one can only use the classic methods for color constancy: Retinex, grey world, grey edge, etc. But when one sees the same object in multiple images, the estimates of both object and illuminant color are constrained in an interesting fashion. We have shown how these constraints can be exploited efficiently and can dramatically improve the estimate of true surface color when multiple object images are available. This allows calibrated color models to be estimated from images available on the Web, even though each of those images individually has very weak color information.

Other Publications: Other publications in 2010 include a journal length version of our earlier work on active learning and probabilistic classification [4] and a earlier method for

learning shared latent spaces [6].

6.2 Transitions and Visits

The end of 2010 brought several transitions in the group. Ashley Eden completed her PhD in December 2010, and several postdocs moved to new positions in Europe and beyond: Fritz moved to start a new group at MPI-Saarbrücken; Christoudias moved to a lab at EPFL in Switzerland; Salzmann moved to the Toyota Technological Institute at Chicago.

Professor Tinne Tuytelaars from the Katholieke Univiersiteit Lueven in Belgium visited our group during summer 2010; Ryan Farrell from the University of Maryland and Sanja Fidler from Slovenia were visiting graduate students during the summer and spring, respectively.

References

[1] M. Christoudias, R. Urtasun, M. Salzmann, and T. Darrell (2010). ''Learning to Recognize Objects from Unseen Modalities.'' Proceedings of the 11th European Conference on Computer Vision (ECCV 2010), Crete, Greece, Part IV, pp. 677-691, September 2010.

[2] M. Fritz, K. Saenko, and T. Darrell (2010). ''Size Matters: Metric Visual Search Constraints from Monocular Metadata.'' Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS 2010), Vancouver, Canada, pp. 622-630, December 2010.

[3] Y. Jia, M. Salzmann, and T. Darrell (2010). ''Factorized Latent Spaces with Structured Sparsity.'' Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS 2010), Vancouver, Canada, pp. 982-990, December 2010.

[4] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell (2010). ''Gaussian Processes for Object Categorization.'' *International Journal of Computer Vision*, Vol. 88, No. 2, pp. 169-188, June 2010.

[5] K. Saenko, B. Kulis, M. Fritz, and T. Darrell (2010). ''Adapting Visual Category Models to New Domains.'' Proceedings of the 11th European Conference on Computer Vision (ECCV 2010), Crete, Greece, Part IV, pp. 213-226, September 2010.

[6] M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell (2010). ''Factorized Orthogonal Latent Spaces.'' Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010), Sardinia, Italy, pp. 701-708, May 2010.

7 Architecture

7.1 Monolithic Silicon Photonics

In a collaboration with the MIT Center for Integrated Photonic Systems, Architecture group researchers are exploring the use of silicon photonics to meet the bandwidth needs of future manycore processors. Projected advances in electrical signaling seem unlikely to fulfill memory bandwidth demands at feasible pinouts and power consumptions. Monolithic silicon photonics, which integrates optical components with electrical transistors in a conventional CMOS process, is a promising new technology that could provide both large improvements in achievable interconnect bandwidth and large reductions in power.

In earlier work [3], we explored how to use photonics to interconnect many processors to memory controllers using an opto-electrical global crossbar. Our results showed that aggregrate throughput can be improved by ≈8--10 × compared to an optimized purely electrical network. We also explored other applications of photonic signallng, including purely on-chip networks, where a photonic variant of the Clos network provides at best around a 2× improvement in performance relative to optimized electrical interconnect, limited by the short distance of on-chip connections.

In recent work, we are exploring how to use photonics to connect processors to DRAM chips. Our initial designs promise large improvements in bandwidth and energy efficiency [1], provided the required photonic devices can be fabricated successfully in a DRAM process. We are actively working with Micron Technology under a third award from DARPA, and hope to begin work on a larger DARPA-funded project in 2011 to begin fabricating test chips in a DRAM technology.

7.2 Vector-Thread Architectures

In earlier work at MIT, Asanović's team developed the Scale vector-thread architecture and processor prototype [2], which combines data-level and thread-level parallel execution models in a single unified architecture. Maven is the second-generation vector-thread architecture, designed to scale up to hundreds of execution ''lanes,'' and with the goal of providing very high throughput at low energy for a wide variety of parallel applications. Maven is based on a new compact lane design, which is replicated to yield a "sea-of-lanes" execution substrate. At run-time, lanes are ganged together to form variable-sized vector-thread engines, sized to match application needs.

Over the last year, we have designed and tuned many variants of data-parallel accelerator, including traditional vector machines and GPU-style ''SIMT'' processors along with our new vector-thread designs. By developing an advanced VLSI design flow, we are able to quickly prototype many variants of each design to support design optimization and ensure we are comparing the best possible variant of each design. A paper covering the efficiency versus programmer productivity tradeoffs of various data-parallel accelerators was submitted to the 2011 ISCA conference.

As an off-shoot of this work, we are collaborating on a new DoE-funded project ''Isis'' (Infrastructure for Synthesis with Integrated Simulation) at the Berkeley Wireless Research

Center (BWRC), to further develop our VLSI tools to provide parameterized generation of optimized custom many-core processors.

Over the course of the next year, we expect to continue refining the vector-thread architecture and to fabricate prototype test chips to help validate our simulation results. As an important link between the two Architecture group projects, we will also be using a variant of the vector-thread core in the proposed photonic demonstration system.

7.3 Other Collaborations

The Architecture group works closely with the Parallel Computing Laboratory (Par Lab) in the Computer Science Division at UC Berkeley. The ICSI work uses the software tools and parallel applications developed in Par Lab to evaluate new architectural ideas.

The Architecture group was also heavily involved with the multi-university RAMP (Research Accelerator for Multi-Processors) consortium hosted at BWRC, which officially concluded in August 2010 with a final project retreat held at Stanford. The RAMP project developed many new techniques in FPGA simulation technology [5] and this area continues to be a major research thrust at Berkeley. A second version of the Par Lab RAMP Gold manycore simulator [4], named RAMP Midas, is being developed and will help support both the photonics work and the vector-thread processor work.

References

[1] S. Beamer, C. Sun, Y.-J. Kwon, A. Joshi, C. Batten, V. Stojanović, and K. Asanović. ''Re-Architecting DRAM Memory Systems with Monolithically Integrated Silicon Photonics.'' Proceedings of the 37th International Symposium on Computer Architecture (ISCA 2010), Saint-Malo, France, pp. 129-140, June 2010.

[2] R. Krashinsky, C. Batten, and K. Asanović (2008) ''Implementing the Scale Vector-Thread Processor,'' *ACM Transactions on Design Automation of Electronic Systems* (TODAES), Vol. 13, Issue 3, pp. 41:1-41:24 July 2008.

[3] V. Stojanović, A. Joshi, C.Batten, Y.-J. Kwon, and K. Asanović. ''A Design-Space Exploration for CMOS Photonic Processor Networks.'' Invited paper at the Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC), San Diego, California, paper OWI1, March 2010.

[4] Z. Tan, A. Waterman, R. Avizienis, Y. Lee, H. Cook, D. Patterson, and K. Asanović. ''RAMP Gold: An FPGA-Based Architecture Simulator for Multiprocessors.'' Proceedings of the 47th Design Automation Conference (DAC-2010), Anaheim, California, pp. 463-468, June 2010.

[5] Z. Tan, A. Waterman, H. Cook, S. Bird, K. Asanović, and D. Patterson. ''A Case for FAME: FPGA Architecture Model Execution.'' Proceedings of the 37th International Symposium on Computer Architecture (ISCA 2010), Saint-Malo, France, pp. 290-301, June 2010.