INTERNATIONAL
COMPUTER SCIENCE
INSTITUTE

International Computer Science Institute Activity Report 2009

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 2009 Sponsors

Defense Advanced Research Projects Agency (DARPA)
Intelligence Advanced Research Projects Activity (IARPA, formerly DTO)
European Union (via University of Edinburgh)
Finnish National Technology Agency (TEKES)
German Ministry of Education and Research (via the DAAD)
IM2 National Centre of Competence in Research, Switzerland
National Science Foundation (NSF)
Spanish Ministry of Science and Innovation (MICINN)

Affiliated 2009 Sponsors

Appscio
AT&T
Google
IBM
Intel
Lockheed
Microsoft
Panasonic
Qualcomm
Toyota

Corporate Officers

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

Board of Trustees, January 2010

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

2009 International Visitor Program

Name	Country	Group	Sponsor
Bin Dai	China	Networking	CSC
Bo Xu	China	Networking	CSC
Carl Henrick Ek	EU	Vision	AMIDA
Carolin Mende	EU	Speech	AMIDA
Shasha Xie	EU	Speech	AMIDA
Yoshia Hirase	Finland	Other	TEKES
Anniina Huttunen	Finland	Other	TEKES
Joakim Koskela	Finland	Networking	TEKES
Kimmo Kuusilinna	Finland	Other	TEKES
Tommi Lampikoski	Finland	Other	TEKES
Tiina Lindh-Knuutila	Finland	AI	TEKES
Annukka Näyhä	Finland	Other	TEKES
Boris Nechaev	Finland	Networking	TEKES
Jarno Rajahalme	Finland	Networking	TEKES
Pasi Sarolahti	Finland	Networking	TEKES
Jouni Similä	Finland	Other	TEKES
Marko Turpeinen	Finland	Other	TEKES

Jan Baumbach	Germany	Algorithms	DAAD
Bernd Böhme	Germany	Networking	DAAD
Joos-Hendrick Böse	Germany	Networking	DAAD
Nicolas Cebron	Germany	Vision	DAAD
Tobias Friedrich	Germany	Algorithms	DAAD
Martin Gairing	Germany	Algorithms	DAAD
Oliver Günther	Germany	Other	DAAD
Sascha Hunold	Germany	Architecture	DAAD
Jörg Lässig	Germany	Algorithms	DAAD
Ulrich Rückert	Germany	Algorithms	DAAD
Benjamin Satzger	Germany	Algorithms	DAAD
Thomas Sauerwald	Germany	Algorithms	DAAD
Guido Schryen	Germany	Networking	DAAD
Dirk Sudholt	Germany	Algorithms	DAAD
Holger Ziekow	Germany	Other	DAAD
Eduardo Lopez	Spain	Speech	MICINN
Carlos Subirats	Spain	AI / FrameNet	MICINN
Enrique Torres	Spain	Architecture	MICINN
Nikhil Garg	Switzerland	Speech	IM2
David Imseng	Switzerland	Speech	IM2

AMIDA: Augmented Multi-party Interaction with Distance Access
CSC: China Scholarship Council
DAAD: Deutscher Akademischer Austausch Dienst
IM2: Interactive Multimodal Information Management, National Centre of Competence in Research, Switzerland
MICINN: Ministerio de Ciencia e Innovación
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 started in 1986 and 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 "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 2009, in addition to sponsored visitor programs with Finland, Germany, Spain, and Switzerland, we had many visitors associated with specific projects, such as the EU project AMIDA, and the Berkeley TRUST center's collaboration with Taiwan on internet security . 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 in-

vestigators 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 2009

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.

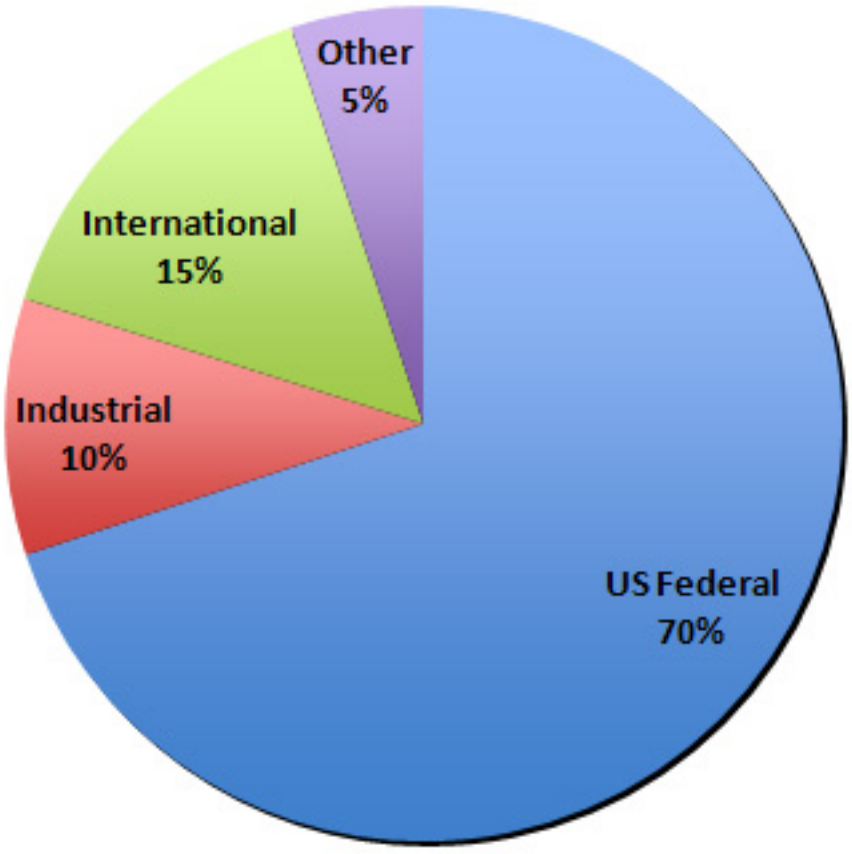

Figure 1: ICSI revenue sources for 2009.

US federal funding in 2009 came from a range of grants to support research institute-wide. Most of this funding came from the National Science Foundation, DARPA, and IARPA. International support of our visitor program came from the Ministry of Education and Research in Germany, the Ministry of Science and Innovation in Spain, 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). Additional support came from the European Union (as a partner in the Integrated Project, AMIDA). Industrial support was provided by Qualcomm, AT&T, IBM, Lockheed, Microsoft, Intel, AppScio, Toyota, and Panasonic. Total ICSI revenue was $8.3M in 2009.

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 current (March 2010) senior research staff members at ICSI 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;

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;

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;

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, InformationTechnology for Emergng 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

Part II
Research Group Reports

1 Research Group Highlights

The following are a selection of key achievements in our research groups for 2009, 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

- In collaboration with colleagues at Stanford and several Silicon Valley companies, ICSI researchers have developed a radically new approach to networking called "Software-Defined Networking" (SDN). SDN combines simple switches that expose an interface to the forwarding path with a network operating system that provides global network abstractions for management. SDN provides far greater control than traditional approaches to network management, at far lower cost. Research embodiments of SDN include the NOX network operating system (www.noxrepo.org) and the OpenFlow switch interface (openflow.org).

- The "Netalyzr" (netalyzr.icsi.berkeley.edu/netalyzr.com) provides a comprehensive and easy to use network measurement and debugging service that evaluates the functionality of people's Internet connectivity. Users access the service as a Java applet via their Web browser. The applet 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. Users have recorded over 100,000 Netalyzr measurements internationally for future extensive measurement studies of edge-network properties.

- ICSI researchers and colleagues developed a novel approach for fighting email spam by leveraging a spammer's own tools. "Botnet judo" works by running deliberately infected systems ("bots") in a highly controlled environment to observe the email spam messages they attempt to send. These measurements allow defenders to monitor new spam as it is created, and consequently infer the underlying *template* used to generate polymorphic e-mail messages. By then employing recognizers for these general templates, it is possible to automatically filter *any* spam generated by the bots precisely and with virtually no false positives.

1.2 Algorithms

- In addition to developing generic computational methods for the analysis of genome-wide association studies, the Algorithms group has been collaborating with researchers who study specific diseases, including asthma, breast cancer, and non-Hodgkin's lymphoma. Together with a group from the School of Public Health at UC Berkeley, the group published the first genome-wide association study of non-Hodgkin's lymphoma in Nature Genetics.

- Researchers have defined a broad class of combinatorial optimization problems called implicit hitting set problems and have developed a generic method for their solution. They developed a highly effective computer program for solving one problem in this

class, multi-genome sequence alignment. As a byproduct of this work they developed a general methodology, derived from machine learning theory, for constructing and validating heuristic algorithms for NP-hard combinatorial optimization problems.

- Richard Karp received the Dickson Prize in Science from Carnegie-Mellon University. He also co-authored "A New Biology for the 21st Century," a report issued by the National Research Council.

1.3 Artificial Intelligence

- A major new initiative was the production and release of a new Embodied Construction Grammar (ECG) wiki, http://ecgweb.pbwiki.com/ which contains a tutorial and other pedagogical materials, a downloadable ECG analyzer and development environment, and also serves as a coordination point for community wide grammar development and analysis.

- In 2009, as the first steps toward our vision of universal health information access, we created the first semantically searchable repository of primary health materials called the Hesperian Digital Commons (HDC) in three languages (English, Spanish and Tamil) (http://www.hesperian.net). This initial effort was funded by a Google Research grant and from a Rockefeller Foundation planning grant.

- From July 31 to August 2, a conference entitled "Frames and Constructions" was held at UC Berkeley to honor the founder of the ICSI FrameNet project, Prof. Charles J. Fillmore. More than 50 papers were presented by linguists from all over the world, roughly half of which were primarily about frame semantics. The complete program with abstracts is available at http://linguistics.berkeley.edu/ fillmorefest/program.html.

1.4 Speech

- Researchers participated in the ACM Multimedia Grand Challenge and won the first prize with a system for the navigation in sitcoms based on speaker identification and acoustic event detection.

- Researchers participated in the NIST Rich Transcription 2009 evaluation and continued to show results at the state of the art in speaker diarization and speaker-attributed speech recognition (the latter collaboratively with SRI).

- The summarization of meetings have been greatly improved by the inclusion of sentence scoring mechanisms and prosodic features.

- The summarization of multiple documents has been improved by developing methods for better sentence segmentation and sentence compression, as evidenced by performance in NIST TAC evaluations.

- The group expanded into a number of new areas, including music processing (for cover song detection, performing very well in an international competition called MIREX 2009), and statistical machine translation, where we demonstrated an efficient approach that parses and aligns simultaneously in a joint model, leading to improved translations.

1.5 Computer Vision

- The group grew in size from approximately four members to over ten, maintained sponsorship from government and corporate sources, and engaged in a serious of productive projects with significant publications.

- The group addressed core computer vision challenges, including learning good local feature representations, modeling human motion, and learning category models from web imagery,

- The group continued research on core machine learning challenges, including dimensionality reduction, metric learning, and active learning.

1.6 Computer Architecture

- The architecture group received a new DARPA award to study extending photonic interconnects into DRAM chips, to provide a fully photonic memory interconnect.

- The photonics work resulted in five publications, including an invited paper at the leading electro-optics conference.

- Christopher Batten completed his PhD on vector-thread architectures and has taken a position as an Assistant Professor in the Electrical and Computer Engineering Department at Cornell University.

2 Networking

2.1 Measurements and Modeling

Assessing Internet traffic manipulation: Internet users generally 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.

First, while Web pages sent over HTTP have no integrity guarantees, it is commonly assumed that such pages are not modified in transit. In the *Tripwires* project we developed evidence of surprisingly widespread and diverse changes made to Web pages between the server and client [43]. Over 1% of Web clients in our study received altered pages, and we showed that these changes often had undesirable consequences for Web publishers or end users. Such changes included: popup-blocking scripts inserted by client software, advertisements injected by ISPs, and even malicious code likely inserted by malware using ARP poisoning. Additionally, we found that changes introduced by client software can inadvertently cause harm, such as introducing cross-site scripting vulnerabilities into pages that a client visits. To help publishers understand and react appropriately to such changes, we developed client-side JavaScript code that can detect most in-flight modifications to a web page.

In the *Netalyzr* project [23], we are developing a 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. We began to offer the service to the public in June 2009 and have since collected 90,000 sessions from 170 countries, which we are currently analyzing.

A third effort assesses the degree to which network operators employ devices to enforce usage restrictions by actively terminating connections deemed undesirable. While the spectrum of the application of such devices is large—from ISPs limiting the usage of P2P applications to the "Great Firewall of China"—many of these systems implement the same approach to disrupt the communication: they inject artificial TCP Reset (RST) packets into the network, causing the endpoints to shut down communication upon receipt. In this project we study the characteristics of packets injected by such traffic control devices [58]. We found that by exploiting the race-conditions that out-of-band devices inevitably face, we not only can detect such interference but often also fingerprint the specific device in use. We developed an efficient injection detector and demonstrated its

effectiveness by identifying a range of disruptive activity seen in traces from four different sites, including termination of P2P connections, anti-spam and anti-virus mechanisms, and the finding that China's "Great Firewall" has multiple components, sometimes apparently operating without coordination. We also found a number of sources of idiosyncratic connection termination that do *not* reflect third-party traffic disruption, including NATs, load-balancers, and spam bots. In general, our findings highlight that *(i)* Internet traffic faces a wide range of control devices using injected RST packets, and *(ii)* to reliably detect RST injection while avoiding misidentification of other types of activity requires considerable care.

Measuring characteristics of residential broadband traffic: While residential broadband Internet access is popular in many parts of the world, only a few studies have examined the characteristics of such traffic. This project (in collaboration with researchers at TU-Berlin) draws upon monitoring of the network activity for more than 20,000 residential DSL customers in an urban area. To ensure privacy, all data is immediately anonymized. We augment anonymized packet traces with information about DSL-level sessions, IP (re-)assignments, and DSL link bandwidth. To date our analysis has revealed a number of surprises in terms of the mental models arising from the prior measurement literature. For example, we found that HTTP traffic, rather than peer-to-peer, dominates by a significant margin; that more often than not the home user's immediate ISP connectivity contributes more to the round-trip times the user experiences than the WAN portion of the path; and that the DSL lines are frequently not the bottleneck in bulk-transfer performance [33]. We are continuing to gather additional measurements, including a set from a different residential environment, and broadening our analysis to include an assessment of Internet security properties.

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.

To date we have captured many hundreds of GBs of data from the Lawrence Berkeley National Laboratory. Our initial research efforts have grappled with a number of measurement calibration issues that require careful consideration [37]. First, we must correctly account for redundant copies of packets that appear due to switch flooding, which if not accurately identified can greatly skew subsequent analysis results. We found that a simple, natural rule one might use for doing so in fact introduces systematic errors, but an altered version of the rule performs significantly better. We then employed this revised rule to aid with calibration issues concerning the fidelity of packet timestamps and the amount of measurement loss that our collection apparatus incurred. Additionally, we develop techniques to "map" the monitored network in terms of identifying key topological components, such as subnet boundaries, which hosts were directly monitored, and the

presence of "hidden" switches and hubs. The initial analyses we then performed using the calibrated data demonstrated that the magnitude and diversity of traffic at the subnet level is in fact striking, highlighting the importance of obtaining and correctly calibrating switch-level enterprise traces.

Longitudinal HTTP analysis: In this project we have begun an analysis of 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 delivery of content is structured 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 aimed at a unique long-term characterization [10].

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].

Assessing and mitigating the impact of P2P systems: Bulk data peer-to-peer file sharing exhibits unique economic properties that create incentives for traffic manipulation on the part of ISPs. Rather than cost savings, bulk-data P2P is instead a wholesale cost-shifting from the content providers to the ISPs. We developed a simple economic model of different content-delivery mechanisms and showed how the addition of caching into the infrastructure changes the cost model substantially, from a system that benefits the content providers at great detriment to the ISPs, to a system that benefits both the content providers and ISPs [52, 53].

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 detection analysis, (2) the relentless growth in the volume of network traffic that we must analyze, and, critically, (3) the failure in recent years of uniprocessor performance to sustain the exponential gains that for so many years were routine for CPUs ("Moore's

Law"). For commodity hardware, tomorrow's performance gains will instead come from *multicore* 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 towards 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 completed, and thus that it is safe to forward the pending packet—providing none of the analysis elements previously signaled that the packet should instead be discarded [46].

Bro cluster: While the task of parallelizing network intrusion analysis might at first blush seem fairly simple—split the traffic among multiple CPUs on a per-connection basis, and we're done—in reality, such a division becomes significantly more subtle when we must consider higher-level analyses that require coordination of information *across* connections or hosts. This project has developed a *clusterizable* version of the Bro intrusion detection system, with a focus on an approach whereby if one can dedicate N commodity PCs to the task of executing Bro, then the execution of Bro will be approximately N times more efficient.

To date, this effort has developed prototypes successfully operating at the Lawrence Berkeley National Laboratory and the University of California at Berkeley [50], and evaluated the implementation using stress-testing [57]. We are now working with LBNL to expand the prototype deployment there to one suitable for 24/7 operational monitoring, and seeking to expand the UCB deployment to facilitate more pervasive and in-depth monitoring than the prototype deployment there can currently sustain. We are also working on developing a more efficient architecture for distributing events between the multiple nodes than the current mesh/broadcast model, which burdens large clusters with excessive communication overhead.

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 are building a new monitoring infrastructure for the UC Berkeley campus, due to go into operation in Spring 2010. The new setup will consist of about 30 PCs forming a "Bro cluster" (see above). The new cluster will serve as a powerful research platform for

future studies, providing unprecedented capabilities for analyzing a large-scale operational network in depth. Furthermore, on a technical level it will allow 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 will also facilitate the continuation of security research tied to the operational requirements of one of the largest academic network environments in our country.

Spam campaign analysis via botnet infiltration: Over the last decade, unsolicited bulk email, or *spam*, has transitioned from a minor nuisance to a major scourge, adversely affecting virtually every Internet user. Spam is used not only to shill for cheap pharmaceuticals, but has also become the de facto delivery mechanism for a range of criminal endeavors, including phishing, securities manipulation, identity theft and malware distribution. While there is a considerable body of research focused on spam from the recipient's point of view, we understand considerably less about the *sender's* perspective: how spammers test, target, distribute and deliver a large spam campaign in practice.

In this project we pursue a new methodology—*botnet infiltration*—for measuring spam campaigns *from the inside*. By hooking into a botnet's *command-and-control* protocol, we can infiltrate a spammer's distribution platform and measure spam campaigns as they occur. To date we have conducted three studies using infiltration of the well-known Storm botnet [30, 28, 31].

In the first of these, we examined the system components used to support spam campaigns, including a work queue model for distributing load across the botnet, a modular campaign framework, a template language for introducing per-message polymorphism, delivery feedback for target list pruning, per-bot address harvesting for acquiring new targets, and special test campaigns and email accounts used to validate that new spam templates can bypass filters. We also measured the dynamics of how such campaigns unfold, analyzing the address lists to characterize the targeting of different campaigns, delivery failure rates (a metric of address list "quality"), and estimated total campaign sizes as extrapolated from a set of samples.

The second study expanded our monitoring of campaign orchestration over a longer period of time, focusing on the content of campaigns rather than their mechanics. We identified 90 different campaign types hosted on the Storm platform during the timeframe of our investigation, targeting 630 million different email addresses and harnessing 90,000 different spamming zombies. We classified individual campaigns by topic and time, and studied the evasive maneuvers employed by the spammers to stay ahead of filtering infrastructure. We also studied the spammers' campaign targeting strategies, including usage patterns of "spamvertized" URLs, harvested email addresses, target group selection, and target list maintenance.

In a follow-on effort, we undertook the first systematic assessment of the "conversion rate" of a spam campaign—the rate at which an unsolicited e-mail ultimately elicits a "sale." From the spammer's perspective, the conversion rate underlies the entire spam value proposition. Using a parasitic infiltration of the Storm botnet's infrastructure, we analyzed two spam campaigns: one designed to propagate a malware Trojan, the other marketing on-line pharmaceuticals. For nearly a half billion spam e-mails we identified

the number that are successfully delivered, the number that passed through popular anti-spam filters, the number that elicited user visits to the advertised sites, and the number of "infections" and "sales" produced. We found that achieving one infection required sending about a quarter of a million spams, and achieving one pharmaceutical sale required sending more than 12 million spams.

Automated reversing of botnet C&C protocols: Automatic protocol reverse-engineering is important for many security applications, including the analysis and defense against botnets. Understanding the *command-and-control* (C&C) protocol used by a botnet is crucial for anticipating its repertoire of nefarious activity and to enable active botnet infiltration. With regard to this latter, security analysts frequently need to rewrite messages sent and received by a bot in order to contain malicious activity and to provide the botmaster with an illusion of successful and unhampered operation. Such analysis and rewriting requires detailed information about the intent and structure of the messages in *both directions* of the communication, despite the fact that we generally only have access to the implementation of one endpoint, namely the bot binary. Current techniques cannot enable such rewriting.

In this project we investigate techniques to extract the format of protocol messages *sent* by an application that implements a protocol specification, and to infer the field semantics for messages both *sent* and *received* by the application. Our techniques enable applications such as rewriting the C&C messages for active botnet infiltration. We implemented our techniques in *Dispatcher*, a tool to extract the message format and field semantics of both received and sent messages. We use Dispatcher to analyze MegaD, a prevalent spam botnet employing a hitherto undocumented C&C protocol, and showed that the protocol information extracted by Dispatcher sufficed to rewrite the C&C messages [11].

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 [42]. 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 criminalized malware has been the development of *marketplaces* that criminals use to foster a specialized economy of buyers and sellers of particular 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 [19]. 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.

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 [4]. 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 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 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 the last (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 [32].

Developing a trusted path to the user: One of the fundamental activities within a network is authorization. Current, largely password-based, schemes fail for a number of reasons, but crucially because passwords are both easy to steal (via host compromise or phishing) and easy to use once obtained. Stronger authentication schemes (e.g., using cryptography) have failed to gain prevalence due to their complexity for the general user. We have begun designing a *trusted path to the user* as an essential building block for the Future Internet architecture [55]. The particular notion we are exploring is that of a "key fob" that readily fits on a user's physical key ring and can provide such a trusted path from Internet services to users regardless of the state of the components of that path.

Selecting ephemeral ports: Careless selection of the ephemeral port number portion of a transport protocol's connection identifier has been shown to potentially degrade security by opening the connection up to injection attacks from "blind" or "off path" attackers (attackers that cannot directly observe the connection). In this effort we empirically evaluated a number of algorithms for choosing the ephemeral port number that attempt to obscure the choice from such attackers [1].

DNS spoofing vulnerabilities: DNS resolvers are vulnerable to numerous attacks on their network communication, ranging from "blind" attacks to man-in-the-middle (MITM) interception. Although a full MITM attack can only be countered with cryptography, there are layers of defenses that apply to less powerful attackers. Of particular interest are defenses which only require changing the DNS resolvers, not the authoritative servers or the DNS protocols. This project develops a taxonomy of attacker capabilities and desires, and explores defenses against different classes of attackers, including: detecting non-disruptive attacks, entropy budgeting, detecting entropy stripping, semantics of duplication, and cache policies to eliminate "race-until-win" conditions [54]. We in addition use network traces to evaluate potential defenses.

Securing Web content: Security in the WWW architecture is based on authenticating the source server and securing the data during transport without considering the content itself. The traditional assumption is that a page is as secure as the server hosting it. However, modern web sites have often a composite structure where components of the web page are authored by different actors and one logical page contains components collected from disparate servers. Applying a single security policy to a whole page is inadequate. For instance, because a blog hosting service might be trustworthy but the content an individual blogger or reader making a comment is not. Building on previous work [3], we introduce a new model to protect users from web-based malware. We have developed a new model that uses opportunistic personae to better secure web content by adding integrity and accountability to individual elements [29]. We have both designed an overall mechanism and built a small prototype.

Using traffic characteristics to detect spam: In this project we evaluate the efficacy of using a machine-learning-based model of the 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 transport characteristics 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 [39].

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 have designed, and started to build, a 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. [47]

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. In a paper accepted for publication in 2010, we examine the differences between the network intrusion detection problem and other areas where machine learning regularly finds much more success [45]. Our main claim is that the task of finding attacks is fundamentally different 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 have begun 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.

DNS analysis: Some forms of malware change a victim's client-side DNS configuration from using a trusted DNS resolver (typically provided by the local site/ISP) to a malicious server under the attacker's control. Once achieved, the malicious server is in the position to redirect a victim's traffic arbitrarily by returning wrong DNS replies. We have observed a significant number of such malicious resolvers in the wild, and jointly with colleagues we have begun an analysis of a large set of malware samples for signs of such behavior. The goals of our study are three-fold: (1) estimate the scope of the problem by analyzing the degree to which clients are using malicious DNS resolvers, both in absolute terms as well as relative compared to the usage of benign non-local resolvers; (2) understand the strategies employed by malicious resolvers for returning fake replies, in particular in terms of which DNS names they are redirecting; and (3) evaluate the impact of potential counter-measures that network operators can deploy to avoid their clients becoming victims of this type of attack. We have developed an analysis that monitors a network's external DNS traffic for local clients not using any of the network's canonical DNS resolvers. The analysis compiles a list of all external resolvers seen in use, which we can then examine for signs of malicious replies. We are also working on an online detector that can identify clients talking to malicious resolvers in real-time.

Securing mediated trace access using black-box permutation analysis: The lack of public access to current, representative datasets significantly hinders the progress of network research as a scientific pursuit, and the difficulty of securely sanitizing traces for public release presents a daunting obstacle to seeing this situation improve. An alternative paradigm for enabling network research is *mediated trace analysis*: rather than bringing the data to the experimenter, bring the experiment to the data, i.e., researchers send their analysis programs to data providers who then run the programs on their behalf and return the output. A key technical hurdle for this paradigm is ensuring that the analysis program does not leak sensitive information from the data it processes.

Previously proposed frameworks for securing such mediation have had the significant limitation of requiring researchers to code their analysis programs in terms of pre-approved modules or a specific language. In this project we propose and explore a powerful alternative approach that can work with nearly arbitrary analysis programs while imposing only modest requirements upon researchers. The main idea behind *black-box permutation analysis* is to check for information leaks by permuting the sensitive fields in the input trace and analyzing resulting changes in the program output [35]. We have developed an

analytic framework for the approach and showed that we can detect violations of a data provider's privacy policy using only a modest number of black-box permutations. Our technique can also account for innocuous changes in program output via canonicalization using a researcher-supplied *output template* and an *audit trail* generated at run time.

2.3 Internet Protocols

Updating standard TCP congestion control: ICSI researchers have been instrumental in codifying algorithms for TCP congestion control (previously developed by V. Jacobson) as Internet standards [6]. We revised the previous work to clarify issues and ambiguities that have been identified since its publication [5].

Early Retransmit: In this effort we introduce a new mechanism for TCP and SCTP for recovering lost segments in the presence of a small congestion window. The "Early Retransmit" mechanism [2] 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.

2.4 Novel Internet Architectures

Pathlet routing: We have developed a new multipath routing protocol, pathlet routing, in which networks advertise fragments of paths (pathlets) over virtual nodes [20]. 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 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 [44] 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.

Practical declarative network management: In recent years, high-level declarative management languages have gained traction in the commercial world. Examples include XACML for declaring access-controls over middleware and webservices, and P3P for declaring privacy policies in web environments. In addition, many email readers use declarative languages for message filtering.

These and other successes provide evidence for the utility of declarative management techniques. However, the adoption of these languages has been almost entirely at the application layer and above. In contrast, enterprise networks continue to be managed through a number of disparate low-level mechanisms, including the use of VLANs and subnetting for isolation, ACLs for access control, NAT for client protection, and policy routing for source-based policies and the integration of middleboxes. As has been frequently lamented in the literature, these traditional approaches for network configuration result in networks whose connectivity is dictated by thousands of lines of brittle, low-level configuration code that grows stale as the network evolves. Thus, enterprise networks provide an ideal example of where declarative management techniques could provide substantial benefits.

In this project [22], we developed a Flow-based Management Language (FML), a high-level declarative language for expressing network-wide policies about a variety of different management tasks within a single, cohesive framework. While there have been numerous authorization languages proposed in the literature, we believe FML has a unique position in the academic design space as it was purpose-built to replace existing network configuration practices, and has been extensively tested in practice.

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. Further, we have started the initial design of the basic building blocks necessary to implement our vision.

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 [36] 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.

2.5 Distributed Systems

Reducing energy waste in networked systems: Networked end-systems such as desktops and set-top boxes are often left powered-on, but idle, leading to wasted energy consumption. An alternative would be for these idle systems to enter low-power sleep modes. Unfortunately, today, a sleeping system sees degraded functionality: first, a sleeping device loses its network "presence" which is problematic to users and applications that

expect to maintain access to a remote machine and, second, sleeping can prevent running tasks scheduled during times of low utilization (e.g., network backups). Various solutions to these problems have been proposed over the years including wake-on-lan (WoL) mechanisms that wake hosts when specific packets arrive, and the use of a proxy that handles idle-time traffic on behalf of a sleeping host. However, no in-depth evaluation of the potential for energy savings, and the effectiveness of proposed solutions, had not been carried out.

To remedy this, [38] we collected data directly from 250 enterprise users on their end-host machines capturing network traffic patterns and user presence indicators. With this data, we answer several questions such as: what is the potential value of proxying or using magic packets? which protocols and applications require proxying? how comprehensive does proxying need to be for energy benefits to be compelling?

We find that, although there is indeed much potential for energy savings, trivial approaches are not effective. We also find that achieving substantial savings requires a careful consideration of the tradeoffs between the proxy complexity and the idle-time functionality available to users, and that these tradeoffs vary with user environment. Based on our findings, we propose and evaluate a proxy architecture that exposes a minimal set of APIs to support different forms of idle-time behavior.

Rethinking concurrency control in storage area networks: Clustered applications in storage area networks (SANs), widely adopted in enterprise datacenters, have traditionally relied on distributed locking protocols to coordinate concurrent access to shared storage devices. In this project [16] we examined the semantics of traditional lock services for SAN environments and asked whether they are sufficient to guarantee data safety at the application level. We found that a traditional lock service design that enforces strict mutual exclusion via a globally-consistent view of locking state is neither sufficient nor strictly necessary to ensure application-level correctness in the presence of asynchrony and failures. We also found that in many cases, strongly-consistent locking imposes an additional and unnecessary constraint on application availability. Armed with these observations, we developed a set of novel concurrency control and recovery protocols for clustered SAN applications that achieve safety and liveness in the face of arbitrary asynchrony, crash failures, and network partitions. We then developed Minuet, a new synchronization primitive based on these protocols that can serve as a foundational building block for safe and highly-available SAN applications.

Diverse replication for single-machine Byzantine-fault tolerance: New single-machine environments are emerging from abundant computation available through multiple cores and secure virtualization. In this project [13], we explored the research challenges and opportunities around diversified replication as a method to increase the Byzantine-fault tolerance (BFT) of single-machine servers to software attacks or errors.

Tiered fault tolerance for long-term integrity: Fault-tolerant services typically make assumptions about the type and maximum number of faults that they can tolerate while providing their correctness guarantees; when such a fault threshold is violated,

correctness is lost. This project revisits the notion of fault thresholds in the context of long-term archival storage. Fault thresholds are inevitably violated in longterm services, making traditional fault tolerance inapplicable to the long-term. In this work [14], we reallocate the "fault-tolerance budget" of a long-term service. We split the service into several pieces, each of which can tolerate a different number of faults without failing (and without causing the whole service to fail): each piece can be either in a critical trusted fault tier, which must never fail, or an untrusted fault tier, which can fail massively and often, or other fault tiers in between. By carefully engineering the split of a long-term service into pieces that must obey distinct fault thresholds, we can prolong its inevitable demise. We have demonstrated this approach with Bonafide, a long-term key-value store that, unlike all similar systems proposed in the literature, maintains integrity in the face of Byzantine faults without requiring self-certified data.

Application placement in hosting platforms: Today's Web transactions involve a large variety of components that are unseen by the user. In particular, replicated application servers often do much of the heavy-lifting for large web services. These servers are increasingly hosted on shared platforms. One particularly attractive hosting service model calls for physical servers to be dynamically allocated among multiple applications, with the active application (or applications, if sharing is allowed) dependent on the current workload. These servers therefore must be able to take applications in and out of service in a dynamic fashion. While this notion has been previously developed, the solutions essentially require the overall application churn to be low due to the heavy application startup costs. In this project we investigate techniques to make these application servers more agile by (i) running all applications simultaneously and suspending those not in use and (ii) using new operating system memory management techniques to reduce the cost of both paging a process out and back in when it is to be activated. We have implemented our solution and demonstrated its effectiveness [7].

2.6 Datacenters

Extending networking into the virtualization layer: While virtualization's impact on computing is well known, its implications for networking are far less explored. In particular, virtualization imposes requirements on network mobility, scaling, and isolation that are far beyond what is required in most physical deployments. Virtualization also provides features making networking easier. For example, in virtualized environments, the virtualization layer can provide information about host arrivals and movements.

Thus, networking in a virtualized world presents its own set of challenges and opportunities. However, the typical model for internetworking in virtualized environments is standard L2 switching or IP router functionality within the hypervisor or hardware management layer. This virtual networking component manages communication between co-located virtual machines, and connectivity to the physical NIC. There have been some attempts to adapt the virtual network layer to its unique set of properties, but none of the implementations adequately handle the full range of challenges.

To rectify this situation, we have developed Open vSwitch [41], a network switch specifically built for virtual environments. Open vSwitch differs from traditional approaches in

that it exports an external interface for fine-grained control of configuration state and forwarding behavior. Open vSwitch can be used to tackle problems such as isolation in joint-tenant environments, mobility across subnets, and distributing configuration and visibility across hosts.

Applying NOX to the datacenter: Internet datacenters offer unprecedented computing power for a new generation of data-intensive computational tasks. There is a rapidly growing literature on the operating and distributed systems issues raised by these datacenters, but only recently have researchers turned their attention to the datacenter's unique set of networking challenges. Of particular interest are a series of network designs that, while varying along many design dimensions, are all specifically tailored to the datacenter environment.

In the more general networking literature, the 4D project initiated (in 2004) a renaissance in the network management literature by advocating a logically centralized view of the network. The goal of this approach was to provide a general management plane, not specialized to a particular context (such as the datacenter). A recent development in this vein is the NOX network operating system. Enterprise network management systems built on NOX have been in production use for over a year, and an early version of NOX is freely available under the GPL license at www.noxrepo.org.

The philosophical question behind the work reported on in [49] is whether the general-purpose approach in networking, which has served the Internet and enterprise so well, can be extended to specialized environments like the datacenter, or if special-case solutions will prevail. The more practical instantiation of this question is: How well does a general-purpose management system, like NOX, cope with the highly specific and stringent requirements of the datacenter? We find that not only can NOX provide reasonable management of datacenter environments, it also offers operators a choice of several points in the datacenter design spectrum, rather than locking them into one specific solution.

Delay scheduling for cluster computing: As organizations start to use data-intensive cluster computing systems like Hadoop and Dryad for more applications, there is a growing need to share clusters between users. However, there is a conflict between fairness in scheduling (giving each user his/her allocated share) and data locality (placing tasks on nodes that contain their input data). To address the conflict between locality and fairness, we propose a simple algorithm called delay scheduling: when the job that should be scheduled next according to fairness cannot launch a local task, it waits for a small amount of time, allowing other jobs to launch tasks instead. We find that delay scheduling achieves nearly-optimal data locality in a variety of workloads and can increase throughput by a factor of 2 while preserving fairness. In addition, the simplicity of delay scheduling makes it applicable under a wide variety of scheduling policies beyond fair sharing.

A common substrate for cluster computing: Cluster computing has become mainstream, resulting in the rapid creation and adoption of diverse cluster computing frameworks. We believe that no single framework will be optimal for all applications, and that organizations will instead want to run multiple frameworks in the same cluster. Further-

more, to ease development of new frameworks, it is critical to identify common abstractions and modularize their architectures. To achieve these goals, we propose Nexus, a low-level substrate that provides isolation and efficient resource sharing across frameworks running on the same cluster, while giving each framework maximum control over the scheduling and execution of its jobs. Nexus fosters innovation in the cloud by letting organizations run new frameworks alongside existing ones and by letting framework developers focus on specific applications rather than building one-size-fits-all frameworks.

2.7 Research Community Activities

Sally Floyd co-chairs the Transport Modeling Research Group (TMRG) of the Internet Research Task Force (IRTF).

Mark Allman serves on the IETF's Transport Area Directorate and IANA's ports review team.

Scott Shenker was a Cray Distinguished Speaker at the University of Minnesota and was invited to address the JASON Defense Advisory Group. He also participated in an FCC panel on the "The Future of the Internet". He continues to serve on NetSE Council (formerly the GENI Science Council).

Vern Paxson serves on the steering committee of the *USENIX Workshop on Large-scale Exploits and Emergent Threats* and as a member of the Scientific Advisory Board for Thomson (recently renamed Technicolor). He gave invited talks at the 2009 USENIX Security Symposium (*How The Pursuit of Truth Led Me To Selling Viagra*) and the 2009 iCAST/CMU/TRUST Joint Conference (*Towards Greater Depth in Network Security Monitoring*), along with three lectures for the Microsoft Research India Summer School on Networking on the topic *Network-Based Detection of Attacks*. He was an invited participant of the *MIT-Sandia Workshop on Cyber Data Gap Analysis*, sponsored by DHS, NSF, and ONR.

Vern Paxson and Robin Sommer led two 2.5-day *Bro Hands-On Workshops* at the Lawrence Berkeley National Laboratory. They also taught a tutorial on Bro at the *Annual Computer Security Applications Conference*.

References

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

[2] M. Allman, K. Avrachenkov, U. Ayesta, J. Blanton, and P. Hurtig (2009). "Early Retransmit for TCP and SCTP." IETF Internet-Draft draft-ietf-tcpm-early-rexmt-04.txt, January 2009 (work in progress). In RFC-Editor queue awaiting publication.

[3] M. Allman, C. Kreibich, V. Paxson, R. Sommer, and N. Weaver (2007). "The Strengths of Weaker Identities: Opportunistic Personas." Proceedings of USENIX Workshop on Hot Topics in Security (HotSec 07), Boston, Massachusetts, August 2007.

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

[5] M. Allman, V. Paxson, and E. Blanton (2009). "TCP Congestion Control." IETF Request for Comments 5681, September 2009.

[6] M. Allman, V. Paxson, and W. Stevens (1999). "TCP Congestion Control." IETF Request for Comments 2581, April 1999.

[7] Z. Al-Qudah, H. Alzoubi, M. Allman, M. Rabinovich, and V. Liberatore (2009). "Efficient Application Placement in a Dynamic Hosting Platform." Proceedings of the International World Wide Web Conference (WWW), Madrid, Spain, pp. 281-290,April 2009.

[8] Z. Al-Qudah, M. Rabinovich, and M. Allman (2010). "Web Timeouts and Their Implications." To appear in the proceedings of the Tenth Passive and Active Measurement Conference (PAM 2010), Zurich, Switzerland, April 2010.

[9] L. Andrew, C. Marcondes, S. Floyd, L. Dunn, R. Guillier, W. Gang, L. Eggert, S. Ha, and I. Rhee (2008). "Towards a Common TCP Evaluation Suite." Proceedings of the International Workshop on Protocols for Fast Long-Distance Networks (PFLDnet), Manchester, United Kingdom, March 2008.

[10] T. Callahan, M. Allman, and V. Paxson (2010). "A Longitudinal View of HTTP Traffic." To appear in the proceedings of the Tenth Passive and Active Measurement Conference (PAM 2010), Zurich, Switzerland, April 2010.

[11] 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 (CCCS), Chicago, Illinois, pp. 621-634,November 2009.

[12] M. Casado, M. Freedman, J. Pettit, J. Luo, N. Gude, N. McKeown, and S. Shenker (2009). "Rethinking Enterprise Network Control." *IEEE/ACM Transactions on Networking*, Vol. 17, Issue 4, August 2009.

[13] B. Chun, P. Maniatis, and S. Shenker (2008). "Diverse Replication for Single-Machine Byzantine-Fault Tolerance." Proceedings of USENIX Annual Technical Conference, Boston, Massachusetts, pp. 287-292, June 2008.

[14] B. Chun, P. Maniatis, S. Shenker, and J. Kubiatowicz (2009). "Tiered Fault Tolerance for Long-Term Integrity." Proceedings of the 7th USENIX Conference on File and Storage Technologies (FAST '09), San Francisco, California, February 2009.

[15] H. Dreger, A. Feldmann, V. Paxson, and R. Sommer (2008). "Predicting the Resource Consumption of Network Intrusion Detection Systems." Proceedings of the International Symposium on Recent Advances in Intrusion Detection (RAID), Cambridge, Massachusetts, pp. 135-154, September 2008.

[16] A. Ermolinskiy, D. Moon, B. Chun, and S. Shenker (2009). "Minuet: Rethinking Concurrency Control in Storage Area Networks." Proceedings of the 7th USENIX Conference on File and Storage Technologies (FAST '09), San Francisco, California, February 2009.

[17] S. Floyd (2008). "Metrics for the Evaluation of Congestion Control Mechanisms." Request for Comments 5166, Informational, March 2008.

[18] S. Floyd, M. Handley, J. Pahdye, and J. Widmer (2008). "TCP Friendly Rate Control (TFRC): Protocol Specification." Request for Comments 5348, Proposed Standard, September 2008.

[19] 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.

[20] P. Godfrey, I. Ganichev, S. Shenker, and I. Stoica (2009). "Pathlet Routing." Proceedings of ACM Special Interest Group on Data Communications Conference (SIGCOMM), Barcelona, Spain, pp. 111-122, August 2009.

[21] J. Gonzalez, V. Paxson, and N. Weaver (2007). "Shunting: A Hardware/Software Architecture for Flexible, High-Performance Network Intrusion Prevention." Proceedings of ACM Computer and Communication Security Conference (ACM CCS), Alexandria, Virginia, pp. 139-149,October 2007.

[22] T. Hinrichs, N. Gude, M. Casado, J. Mitchell, and S. Shenker (2009). "Practical Declarative Network Management." Proceedings of the 1st ACM workshop on Research on enterprise networking (WREN '09), Barcelona, Spain, August 2009.

[23] *The ICSI Netalyzr*, `http://netalyzr.icsi.berkeley.edu`.

[24] L. Juan, C. Kreibich, C-H. Lin, and V. Paxson (2008). "A Tool for Offline and Live Testing of Evasion Resilience in Network Intrusion Detection Systems." Proceedings of the 5th GI International Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA), Paris, France, pp. 267-278, July 2008.

[25] J. Jung, V. Paxson, A. Berger, and H. Balakrishnan (2004). "Fast Portscan Detection Using Sequential Hypothesis Testing." Proceedings of IEEE Symposium on Security and Privacy, Oakland, California, May 2004.

[26] J. Jung, R. Milito, and V. Paxson (2007). "On the Adaptive Real-Time Detection of Fast-Propagating Network Worms." Proceedings of the 4th GI International Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA), Lucerne, Switzerland, pp. 175-192, July 2007.

[27] J. Jung, R. Milito, and V. Paxson (2008). "On the Adaptive Real-Time Detection of Fast-Propagating Network Worms." *Journal on Computer Virology*, Vol. 4, Issue 1, pp.197-210, February 2008.

[28] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage (2008). "Spamalytics: An Empirical Analysis of Spam Marketing Conversion." Proceedings of the 15th ACM Conference on Computer and Communications Security (ACM CCS), Alexandria, Virginia, pp. 3-14, October 2008.

[29] J. Koskela, N. Weaver, A. Gurtov, and M. Allman (2009). "Securing Web Content." To appear in the proccedings of the ACM CoNext Workshop on Re-Architecting the Internet (ReArch 2009), Rome, Italy, December 2009.

[30] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage (2008). "On the Spam Campaign Trail." Proceedings of the First USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 2008), San Francisco, California, April 2008.

[31] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage (2009). "Spamcraft: An Inside Look At Spam Campaign Orchestration." Proceedings of the 2nd USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 09), Boston, Massachusetts, April 2009.

[32] 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.

[33] G. Maier, A. Feldmann, V. Paxson, and M. Allman (2009). "On Dominant Characteristics of Residential Broadband Internet Traffic." Proceedings of the 2009 Internet Measurement Conference (IMC 2009), Chicago, Illinois, pp. 90-102, November 2009.

[34] G. Maier, R. Sommer, H. Dreger, A. Feldmann, V. Paxson, and F. Schneider (2008). "Enriching Network Security Analysis with Time Travel." Proceedings of ACM Special Interest Group on Data Communications Conference (SIGCOMM 2008), Seattle, Washington, pp. 183-194, August 2008.

[35] P. Mittal, V.Paxson, R.Sommer, and M. Winterrowd (2009). "Securing Mediated Trace Access Using Black-Box Permutation Analysis." To appear in the proceedings of the 8th ACM Workshop on Hot Topics in Networks (HotNets-VIII), New York City, New York, October 2009.

[36] C. Muthukrishnan, V. Paxson, M. Allman, and A. Akella (2010). "Integrating Middleboxes Into Network Architecture Using Annotated Networking." January 2010. Under submission.

[37] 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.

[38] S. Nedevschi, J. Chandrashenkar, B. Nordman, S. Ratnasamy, and N. Taft (2009). "Skilled in the Art of Being Idle: Reducing Energy Waste in Networked Systems."

Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation, Boston, Massachusetts, pp. 381-394, April 2009.

[39] 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.

[40] V. Paxson, R. Sommer, and N. Weaver (2007). "An Architecture for Exploiting Multi-Core Processors to Parallelize Network Intrusion Prevention." Proceedings of IEEE Sarnoff Symposium, Princeton, New Jersey, pp. 1-7, May 2007.

[41] B. Pfaff, J. Pettit, K. Amidon, M. Casado, T. Koponen, and S. Shenker (2009). "Extending Networking Into the Virtualization Layer." Hotnets, 2009.

[42] 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." To appear in the proceedings of the 17th Annual Network and Distributed System Security Symposium (NDSS Symposiom 2010), San Diego, California, March 2010.

[43] C. Reis, S. Gribble, T. Kohno, and N. Weaver (2008). "Detecting In-Flight Page Changes with Web Tripwires." Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2008), San Francisco, California, pp. 31-44, April 2008.

[44] A. Seehra, J. Naous, M. Walfish, D. Maziéres, A. Nicolosi, and S. Shenker (2009). "A Policy Framework for the Future Internet." To appear in the proceedings of the 8th ACM Workshop on Hot Topics in Networks (HotNets-VIII), New York City, New York, October 2009.

[45] R. Sommer and V. Paxson (2010). "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection." To appear in the proceedings of the IEEE Symposium on Security and Privacy 2010, Oakland, California, May 2010.

[46] 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.

[47] R. Sommer, N. Weaver, and V. Paxson (2009). "HILTI: An Abstract Execution Environment for High-Performance Network Traffic Analysis." ICSI Technical Report 10-003, February 2010.

[48] L. Tang, J. Li, Y. Li, and S. Shenker (2009). "An Investigation of the Internet's IP-layer Connectivity." *Computer Communications*, Vol. 32, Issue 5, March 2009.

[49] A. Tavakoli, M. Casado, T. Koponen, and S. Shenker (2009). "Applying NOX to the Datacenter." To appear in the proceedings of the 8th ACM Workshop on Hot Topics in Networks (HotNets-VIII), New York City, New York, October 2009.

[50] M. Vallentin, R. Sommer, J. Lee, C. Leres, V. Paxson, and B. 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), Queensland, Australia, September 2007.

[51] M. Vutukuru, H. Balakirshnan, and V. Paxson (2008). "Efficient and Robust TCP Stream Normalization." Proceedings of IEEE Symposium on Security and Privacy, Oakland, California, pp. 96-110, May 2008.

[52] N. Weaver (2009). "Edge-Caches Should Be Free." Proceeedings of the Telecommunications Policy Research Conference, August 2009.

[53] N. Weaver (2009). "Peer to Peer Localization and Edge Caches." IETF draft *draft-weaver-alto-edge-caches-00*, 2009.

[54] N. Weaver. "Comprehensive DNS Resolver Defenses Against Cache Poisoning." work in progress.

[55] N. Weaver and M. Allman (2009). "On Constructing a Trusted Path to the User." ICSI Technical Report 09-009, December 2009.

[56] N. Weaver, V. Paxson, and J. Gonzalez (2007). "The Shunt: An FPGA-Based Accelerator for Network Intrusion Prevention." Proceedings of International Symposium on Field Programmable Gate Arrays (FPGA), Monterey, California, pp. 199-206, February 2007.

[57] N. Weaver and R. Sommer (2007). "Stress Testing Cluster Bro." Proceedings of USENiX DETER Community Workshop on Cyber Security Experimentation and Test, Boston, Massachusetts, August 2007.

[58] N. Weaver, R. Sommer, and V. Paxson (2009). "Detecting Forged TCP Reset Packets." Proceedings of the 16th Annual Network and Distributed System Security Symposium (NDSS 2009), San Diego, California, February 2009.

3 Algorithms

3.1 Introduction

The work of the Algorithms Group in 2009 focused on computational problems arising in biology and computational methods inspired by analogies with biological processes. Eran Halperin led an intensive effort in computational genetics, with active participation from Bonnie Kirkpatrick, Bogdan Pasaniuc, and Sridhar Sankararaman. Richard Karp developed a methodology for constructing and validating heuristic combinatorial algorithms and applied the methodology to the discovery of functional modules within protein-protein interaction networks and the construction of genome-wide sequence alignments. Jan Baumbach developed algorithms and software for several problems related to genetic regulation. Shuai Li achieved significant advances in protein folding algorithms. Dirk Sudholt and Jörg Lässig conducted novel investigations of combinatorial algorithms based on mimicking processes in biology such as the coordinated activity of ant colonies. Other activities included an analysis of genomic privacy by Sankararaman, Halperin, Prof. Michael Jordan of UC Berkeley, and research on planning by Benjamin Satzger.

3.2 Computational Genetics

In recent years, sequencing and genotyping technologies have advanced in an unprecedented manner, leading to exciting discoveries in human genetics, regarding the relations between genes and diseases. The large amounts of data generated by these technologies sets computational and statistical challenges; a large part of the activity of the ICSI Algorithms group consists in searching for methods that cope with these challenges.

The group has been mainly focusing on the development of computational methods for the analysis of genome-wide association studies, in which a set of cases (individuals carrying a disease) and a set of controls (healthy individuals) are genotyped, and the genetic variation of the two populations is compared. Typically, the diseases studied in this manner are complex and common diseases such as cancer, diabetes, or coronary artery disease.

The analysis of such association studies is complicated by the fact that the genome itself is not competely sequenced for each of the cases, or the controls. Instead, a subset of the positions in the genome (called SNPs - single nucleotide polymorphisms) are 'genotyped', and analyzed. Furthermore, there are other complications such as ambiguity and noise in the genotyped data, and missing data. A few major efforts in the group this year were to develop algorithms that deal with such incomplete data by imputing missing SNPs under different scenarios, including instances where the cases and/or controls include relatives, which complicates the computational analysis considerably.

In addition, the Algorithms Group developed other methods for SNP analysis, including methods to protect the privacy of genomic data in a database, and methods for ancestry inference from SNP data. For the latter, the group developed methods for the inference of ancestry on a local level in the genome; using such algorithms, it is possible to infer the ancestry of each base in the genome. Such analysis is crucial in order to detect genomic regions in which rare mutations may increase the risk for a disease.

Apart from developing computational methods for the analysis of such studies, the Algorithms Group has been collaborating with groups that study specific diseases, including asthma, breast cancer, and non-Hodgkin's lymphoma. For the latter, we have recently published together with a group from UC Berkeley, the first genome-wide association study on non-Hodgkin's lymphoma, published in the journal *Nature Genetics*.

Inferring missing genotype data: The last century saw great strides in understanding simple diseases, such as malaria susceptibility, cystic fibrosis, Huntington's, and Tay-Sachs. Complex diseases, however, like cancer, asthma, Crohn's disease, and schizophrenia, remain mysterious due to complex genetic contributions and interactions with environmental factors. We would like to find the genetic mechanisms and the causes for these diseases, in order to usher in an era of personalized medicine.

The main research paradigm is to perform large case-control association studies with the goal of finding a correlation between the presence of the disease and genetic or environmental factors. The basic statistics for correlation analysis are quite simple; however, this task is complicated by both missing and correlated data. Bonnie Kirkpatrick's focus has been on efficiently inferring the missing data and correcting for the correlations.

Roughly one in every thousand nucleotides in the genome is variable in the population, meaning that different individuals may have different nucleotide bases at those polymorphic positions in the genome. Since humans are diploid organisms, meaning that we inherit a copy of each chromosome from both parents, it is possible for a single individual to have variation within their genome at a particular position. A genotype assay allows us to discover the two particular nucleotides that an individual has at each polymorphic position. However, these data do not tell us which nucleotides were inherited from which parent (a.k.a. the haplotype). This leaves us with a missing data problem known as the phasing problem. Furthermore, relationships between individuals leave structure in the data and prevent us from treating the individuals as independent samples from some population.

When relationships between individuals are known in the form of a pedigree, it is an NP-hard problem to calculate the likelihood for the data given the inheritance constraints. We have used a Gibbs sampler which involves solving a constraint satisfaction problem before doing the probability calculation [Kirkpatrick, et al. 2009]. We later extended this work by testing how well disease associations can be detected from the haplotypes estimated by our Gibbs sampler [Kirkpatrick, et al. In press]. We have also investigated the impending problem of phasing a pedigree when given haplotype data rather than genotype data. In haplotype data the phase information is known for some individuals, but the parental origin is unknown, leaving missing individuals with unknown phase. Such data will soon be available from genome sequencing methods. Although in some situations the haplotypes provide more information, the likelihood calculation with haplotypes is at least as computationally tractable as the genotype version of the problem [Kirkpatrick, submitted]. This result is counter-intuitive and is contrary to what many colleagues have conjectured during informal conversations.

Inference of local ancestry in recently admixed populations: A characterization of the genetic variation of recently admixed populations may reveal historical population events, and is useful for the detection of SNPs associated with diseases through association studies and admixture mapping. Inference of locus-specific ancestry is crucial to our understanding of the genetic variation of such populations. While a number of methods for the inference of locus specific ancestry are accurate when the ancestral populations are quite distant (e.g. African-Americans), current methods incur a large error rate when inferring the locus-specific ancestry in admixed populations where the ancestral populations are closely related (e.g., Americans of European descent). In this work, we extend previous methods for the inference of locus-specific ancestry and evaluate our approaches on a wide range of scenarios showing that indeed, locus-specific ancestry can be accurately inferred. As a potential utility of our method in subsequent analyses, we showed that the accuracy of genotype imputation methods can be improved by the incorporation of accurate locus-specific ancestries, when applied to admixed populations.

Leveraging genetic variability across populations for the identification of causal variants: Genome-wide association studies have been performed extensively in the last few years, resulting in many new discoveries of genomic regions that are associated with complex traits. It is often the case that a SNP found to be associated with the condition is not the causal SNP, but a proxy to it as a result of linkage disequilibrium. For the identification of the actual causal SNP, fine-mapping follow-up is performed, either with the use of dense genotyping or by sequencing of the region. In either case, if the causal SNP is in high linkage disequilibrium with other SNPs, the fine-mapping procedure will require a very large sample size for the identification of the causal SNP. In this work we show that by leveraging genetic variability across populations, we significantly increase the localization success rate (LSR) for a causal SNP in a follow-up study that involves multiple populations as compared to a study that involves only one population. Thus, the average power for detection of the causal variant will be higher in a joint analysis than that in studies in which only one population is analyzed at a time. On the basis of this observation, we developed a framework to efficiently search for a follow-up study design: our framework searches for the best combination of populations from a pool of available populations to maximize the LSR for detection of a causal variant. This framework and its accompanying software can be used to considerably enhance the power of fine-mapping studies.

Accurate estimation of expression levels of homologous genes in RNA-seq experiments: Next generation high throughput sequencing (NGS) is poised to replace array-based technologies as the experiment of choice for measuring RNA expression levels. Several groups have demonstrated the power of this new approach (RNA-seq), making significant and novel contributions and simultaneously proposing methodologies for the analysis of RNA-seq data. In a typical experiment, millions of short sequences (reads) are sampled from RNA extracts and mapped back to a reference genome. The number of reads mapping to each gene is used as proxy for its corresponding RNA concentration. A significant challenge in analyzing RNA expression of homologous genes is the large frac-

tion of the reads that map to multiple locations in the reference genome. Currently, these reads are either dropped from the analysis, or a naive algorithm is used to estimate their underlying distribution. In this work, we present a rigorous alternative for handling the reads generated in an RNA-seq experiment within a probabilistic model for RNA-seq data; we develop maximum likelihood based methods for estimating the model parameters. In contrast to previous methods, our model takes into account the fact that the DNA of the sequenced individual is not a perfect copy of the reference sequence. We show with both simulated and real RNA-seq data that our new method improves the accuracy and power of RNA-seq experiments.

Improving genotype imputation accuracy by selecting a personalized reference panel for every region in the genome: An important component in the analysis of genome-wide association studies (GWAS) involves the imputation of genotypes that have not been measured directly in the studied samples. The imputation procedure uses the linkage disequilibrium structure in the population to infer the genotype of an unobserved single nucleotide polymorphism. The linkage disequilibrium 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. In this work we introduce 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. Our method is generic and can potentially be incorporated in any of the available imputation methods as an add-on.

Genomic privacy: With rapid increases in the volume of genomic data being churned out, there has been a growing realization that free sharing of this data is essential for its potential to be realized. Sharing individual genomic data brings with it concerns of privacy. As increasing volumes of genomic data are generated, the question of how we can share data while addressing privacy concerns is going to become increasingly important.

Due to the unique characteristics of genomic data, naive attempts to address these concerns, such as exposing only summary data, do not provide acceptable privacy guarantees. For instance, recent studies have shown that statistical methods can detect individuals within a study group based on summary SNP data from genome-wide association studies (GWAS). What is needed is a characterization of the level of privacy that can be achieved when summary data for a subset of SNPs are exposed. We provide analytical and empirical bounds on the statistical power of these methods. Our analysis aims to provide quantitative guidelines for researchers wishing to make a limited number of SNPs available publicly without compromising privacy. We have also implemented a tool, SecureGenome, that determines such a set of SNPs given an input genotype dataset (http://securegenome.icsi.berkeley.edu/securegenome/).

3.3 Combinatorial Optimization

Implicit hitting set problems: A *hitting set problem* is specified by a finite ground set in which each element has a positive weight, and a family of subsets of the ground set called *circuits*. The problem is to find a minimum-weight set that has a nonempty intersection with every circuit. An *implicit hitting set problem* is one in which the circuits are not explicitly listed. Instead, a polynomial-time *generator* is given which, given any subset H of the ground set, either certifies that H is a hitting set or returns a circuit that does not intersect H. Several standard NP-hard problems are shown to be implicit hitting set problems. Each of the following is an implicit hitting set problem: Fedback vertex set uin a graph or digraph, feedback edge set in a digraph, max cut, k-matroid intersection and maximum feasible set of inequalities.

Richard Karp and Erick Moreno Centeno developed a generic algorithm for solving an implicit hitting set problem by using a generator to reduce it to a nested sequence of explicit hitting set problems. They have shown how to cast a multi-genome alignment problem, as an implicit hitting set problem, defined a specialized algorithm for this problem, and presented very favorable computational experience on a thousands of problems involving the alignment of multiple worm genomes.

Methodology for constructing and validating heuristic algorithms: Building on our experience in designing heuristic algorithms for solving NP-hard problems arising in systems biology, we have proposed a general methodology for the construction and validation of effective heuristic algorithms for NP-hard optimization problems. The methodology has five components: an *algorithmic strategy* capable of being realized by a large number of concrete algorithms; a *training set* and a *verification set* consisting of representative problem instances a systematic *tuning process* for converging on the concrete realization that achieves the best performance on the training set, and an *evaluation procedure*. Our highly effective algorithm for multi-genome alignment was created using this methodology, and below we describe an application of the methodology to the *colorful subgraph problem*, a natural problem arising in the contexts of biological and social networks.

The colorful subgraph problem: A network is given together with a set of colors. Associated with each vertex is a color. The problem is to find a connected subgraph G= (V,E) with a minimum number of vertices, containing at least one vertex of each color. [6, 7, 8].

The colorful subgraph problem has previously been attacked in the Torque program [6, 7, 8] by a combination of dynamic programming and integer programming.

We consider an algorithmic strategy that constructs a sequence of feasible solutions $V_1, V_2, ..., V_t$ with fewer and fewer vertices.

Given V_i, an iteration consists of the following steps:

1. Partition the set of colors into two sets, A, the infrequent colors, and B,the frequent colors.

2. Consider the subgraph $G[W]$ induced by W, the set of vertices with infrequent colors.

3. Delete components from this subgraph that are not needed for matching all the infrequent colors.

4. Add a minimal number of additional vertices to interconnect the remaining components and enable all colors to be assigned.

5. Delete redundant vertices of degree 1.

There is considerable latitude in the choice of algorithms for each of the major steps within this algorithmic strategy, and we are using our tuning approach to find the combination of such algorithms that will give the best performance on the training set.

3.4 Analysis of Regulatory Networks

Protein modules: A *protein module* is a set of proteins that work in unison to perform a cellular function. Protein modules tend to be identifiable by the richness of the interactions among their members and by the uniformity of the functional categories within which their members lie.

Two proteins in different species are called *functional orthologs* if they perform the same or similar functions in two different species. Functional orthologs are often descended in evolution from a single ancestral protein.

Informally, protein modules in two species are said to be *conserved* if there are many functional orthologies between proteins in the two modules, and similar patterns of interaction between the proteins in one module and their counterparts in the other module. This concept can be extended to conservation of protein modules across several species.The determination of such correspondences among protein modules is sometimes referred to as the *alignment* of the modules.

The discovery of conserved protein modules rests on two types of data: protein-protein interaction (PPI) networks specifying the best available information about which pairs of proteins in a species have direct physical interactions; and data identifying the functional orthology relations among proteins in different species. Our previous work on discovery of conserved protein modules is reported in [18, 37, 27, 45]

Richard Karp's Ph.D. student Luke Hodgkinson is implementing a new linear-time algorithm called Produles for finding conserved protein modules in large protein interaction networks. The method uses the clustering algorithm Nibble [39] to identify dense subgraphs in a single PPI network, together with a method of finding connected subgraphs in other species that contain functional orthologs of the proteins in these dense subgraphs. Hodgkinson is also co-developer of EasyProt, a tool for maintaining and organizing a wide variety of data useful for understanding regulatory networks of proteins.

Analysis of genetic Interactions: Synthetic genetic anaysis is a powerful technique in which pairs of genes are systematically mutated and screened for interactions affecting phenotype. In [12] we exploit structure within a recent association study in yeast to map a "natural" genetic network containing 2,023 interactions between distinct genomic regions. By integrating this network with known protein complexes, we identify 208 complex-complex interactions and annotate these interactions with underlying gen

expression traits. This natural genetic network overlaps with networks derived through synthetic screens and provides coverage in areas not yet tested by synthetic analysis. As proof of principle, we experimentally confirm novel functional relationships between the INO80 chromatin remodeling complex and nine other protein complexes, as suggested by natural interactions. This study provides a paradigm for how genome-wide association astudy data can be used to elucidate the combinations of factors underlying human disease [12].

Data integration, visualization, and network transfer: We developed an integrated approach for the inter-species knowledge transfer of gene regulatory networks. First, we utilize the known network of well-studied model organisms as a template. Subsequently, we identify conserved regulatory sequence elements within the source organism's genome and the genomes of a group of target species.

Our computational biology pipeline assures reliability of the predicted gene regulatory interactions rather than completeness. Anyways, right now we work on extended ideas that might also provide suggestion for trustworthy regulatory interactions in the near future. So far, we evaluated our method by using the CoryneRegNet platform. Starting with the experimentally proven data stored in our reference database, we first use the developed MoRAine tool to enhance the transcription factor binding site annotations for each of the known regulators in the model organism, in our case Corynebacterium glutamicum. Afterwards, we performed TFBS predictions based on so-called Position Weight Matrices (PWMs) on the target organisms by using PoSSuMsearch. That provided us with information about potentially conserved binding sites between the source and the target organisms. The drawback of this method is the comparably high number of false positives. Hence, we combined it with Transitivity Clustering (described below) to gain further evidence for a predicted regulation by incorporating homology information based on proteins that are clustered together within one group. If applied to C. glutamicum as source organism and the taxonomically closely related targets C. efficiens, C. jeikeium, and C. diphtheriae, we have been able to extend the database content of CoryneRegNet by factor 4.2. Subsequently, we incorporated external databases and slightly improved our approach to demonstrate the potential for the organism Mycobacterium tuberculosis.

Recently, Jan Baumbach extended the CoryneRegNet platform with the genomes of two strains of the pathogen C. pseudotuberculosis. This work was done in collaboration with Vasco Azevedo from Minas Gerais, Brazil. We also included data on C. urealyticum into CoryneRegNet. For all three organisms, we have been able to successfully transfer hundreds of gene regulatory networks from C. glutamicum for further integrated analyses and visualizations with the updated CoryneRegNet platform. Additionally, CoryneRegNet now provides several functionalities that will soon allow external researchers to directly submit/upload their data temporarily or permanently to our platform for integrated analyses with the CoryneRegNet databases content. The respective release 6.0 will be published soon.

Protein homology detection with transitivity clustering: Here, the first goal was to enhance the previously introduced clustering approach FORCE, a software tool for

large-scale protein homology detection. All goals, as proposed in the first project proposal have been achieved successfully: FORCE now integrates n-dimensional graph layouts as preprocessing. It integrates methods inspired by ant colony behavior (ACC, Ant Colony Clustering). Besides these algorithmic improvements, extensions regarding the software implementation (e.g. threads for shared-memory parallel computing) have been performed. Recently we integrated the FORCE strategy into a more comprehensive clustering environment: Transitivity Clustering. The corresponding web site, data sets, an online interface, and the developed Cytoscape plug-ins are available at http://transclust.cebitec.uni-bielefeld.de.

Transcription factor binding site analysis: With MoRAine, an applied approach for the online re-assessment of questionably annotated transcription factor binding sites is available. We are now able to identify putative annotation errors in our regulatory databases. As proposed, we showed that the previous adjustment of TFBSs generally improves subsequent binding site predictions. The method is based on two clustering approaches that try to optimize the average information content of the resulting PWM. We recently integrated Transitivity Clustering into MoRAine to close the gap between running time and accuracy; MoRAine 1.0.

3.5 Protein Folding

Classical protein structure classification methodsinvolve computing and storing a pairwise distance matrix. This hinders the design of efficient algorithms. As a result, no tools are available to solve very large sets of decoys. In his current research, instead of computing pairwise distance, Shuai Cheng Li computes the centroid structure of each cluster by superimposing structures within the cluster to avoid the pairwise distance computation. The results are compared with up-to-date programs. The speed is at least four time faster, and the accuracies are comparable. In addition, our program can cluster 100K decoy instances, which is the first available tool to achieve this.

3.6 Biologically Inspired Algorithms

Dirk Sudholt's research at ICSI has focused on three topics in foundations of bio-inspired randomized search heuristics: memetic evolutionary algorithms, ant colony optimization, and parallel evolutionary algorithms.

Memetic evolutionary algorithms are popular heuristics that hybridize evolutionary algorithms with local search. After creating new offspring by genetic operators, local search may be used to refine the new solutions in the hope to speed up the search process. An important question in the design of these algorithms is how to find a proper balance between evolutionary and local search. In an earlier work [40] we presented two hierarchy results for constructed functions showing that small changes to the parametrization can decide between polynomial and superpolynomial running times [43]. In [41] we considered memetic algorithms for single instances of different combinatorial problems—Mincut, Knapsack, and Maxsat—and showed exemplarily that these algorithms can be successful where many other popular heuristics fail [40] [42].

In ant colony optimization (ACO) candidate solutions for a problem are constructed by random walks of artificial ants through a so-called construction graph. These random walks are guided by so-called pheromones on the edges of the graph. [13] deals with the performance of ACO on shortest path problems. We have chosen shortest paths because the ACO paradigm is inspired by the capability of real ant colonies of finding the shortest path between their nest and a food source. The mentioned study proves that ACO indeed can locate shortest paths efficiently and gives insights into the working principles of ACO algorithms. We also worked on an extension to stochastic shortest path problems where the edge weights are subject to noise [14].

Another investigation considered ACO variants with so-called iteration-best update. Instead of reinforcing the best solution found so far, iteration-best update reinforces the best solution from the current iteration. This update scheme has led to astonishing performance results for a simple test function if the strength of the pheromone update is small enough [28]. A surprising fact is that only two ants per iteration are sufficient to locate the global optimum efficiently. In contrast to this, there is a phase transition to exponential running times for all functions with unique optimum if the update strength is increased.

A more recent area of interest is the analysis of parallel evolutionary algorithms, where a large population is split into smaller subpopulations that evolve in parallel. In so-called island models at specific points of time copies of good individuals are exchanged between the subpopulations to coordinate the search. Besides the obvious advantages of parallelization, phases of independent evolution can lead to an increased diversity within the population. A first rigorous analysis demonstrates the benefits of this approach for a constructed problem where both phases of independent evolution and an exchange of individuals are essential [26].

Since arriving at ICSI Jörg Lässig has published a conference paper about the choice of crossover selection strategies in genetic algorithms, a short journal article about the solution of optimization problems for cooperation generation tasks in enterprise networks (in German) and a book chapter, analyzing different approaches of particle swarm optimization, different boundary conditions and the application of particle swarm optimization for the optimization of Hub and Spoke Inventory Systems.

He also generalized Sudholt's results about the influence of the choice of probability distributions in global optimization heuristics, covering also other methods besides genetic algorithms and including migration strategies in parallel hybrid optimization, which was originally part of his research plan in 2010 and he submitted a comprehensive version of this generalized result including a survey on previous results with the title "How to Choose Distribution Functions in Global Optimization Heuristics Optimally" to the Journal of Global Optimization, where it is under review at the moment.

Together with Dirk Sudholt he worked on the analysis of advantages of a limited amount of communication in parallel genetic algorithms and especially in the island model. Parallelization is becoming a more and more important issue for solving difficult optimization problems. Various implementations of parallel evolutionary algorithms (EAs) have been applied in the past decades. Island models combine phases of independent evolution with migration where genetic information is spread out to neighbored islands. Compared to panmictic models, this mechanism can lead to an increased diversity within the population. They perform a first rigorous runtime analysis for island models and construct a function

where phases of independent evolution as well as communication among the islands is essential. A simple island model with migration finds a global optimum in polynomial time, while panmictic populations as well as island models without migration need exponential time, with very high probability. Their results lead to new insights on the usefulness of migration and contribute to the theoretical foundation of parallel EAs.

The conference paper "On the Structure of a Best Possible Crossover Selection Strategy in Genetic Algorithms" considers 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. Many different schemes have been described in the literature as possible strategies for this task but so far comparisons have been predominantly empirical. It is 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 means uniform probabilities have to be assigned to a group of the best individuals of the population but probabilities equal to zero to individuals with lower fitness, assuming that the probability distribution to choose individuals from the current population can be chosen independently for each iteration and each individual. This result is then generalized also to typical practically applied performance measures, such as maximizing the expected fitness value of the best individual seen in any generation.

The book chapter "Comparative Study of Different Approaches to Particle Swarm Optimization in Theory and Practice" addresses a research area, which in particular has gained increasing attention in recent years by investigating the application of biological concepts to various optimization tasks in science and technology. Technically, global extrema of objective functions in a d-dimensional discrete or continuous space have to be determined or approximated, which is a standard problem with an ample number of applications. In this chapter the particle swarm optimization paradigm is described in di fferent variations of the underlying equations of motion and studied comparatively in theory and empirically on the basis of selected optimization problems. To facilitate this task, the different variants are described in a general scheme for optimization algorithms. Then di fferent side constraints as initial conditions, boundary conditions of the search space and velocity restrictions are investigated in detail. This is followed by an efficiency comparison of swarm optimization to a selection of other global optimization heuristics, such as various single state iterative search methods (simulated annealing, threshold accepting, great deluge algorithm, basin hopping, etc.), ensemble based algorithms (ensemble-based simulated annealing and threshold accepting), and evolutionary approaches. In speci fic, the application of these algorithms to combinatorial problems and standard benchmark functions in high-dimensional search spaces is examined. Further, we show how particle swarm optimization can be e ffectively used for the optimization of so-called single-warehouse multi-retailer systems to optimize the ordering strategy and the transportation resources. Optimization tasks of this class are very common in economy as e.g. in the manufacturing industry and for package delivery services.

Together with Sascha Hunold, Lässig is working on algorithm benchmarking and especially on the comparison of algorithms based on a histogram of the different byte-code instructions when running certain benchmark instances so that it is possible to compare

algorithms completely machine independent. This addresses the problem that in empirical papers different machines and different measures are used to compare algorithms. Especially for global optimization heuristics, the number of objective function evaluations or the number of iterations is applied very often. The goal of this project is to develop a portable running time measure and to apply statistical tests to develop a test suite where an (implemented) algorithm can be tested against other algorithms for the same problem to get automatically more information about its strengths and weaknesses. At the moment we are able to dynamically measure the byte code instructions. We are planning to develop a Web-based tool which is freely available.

Variants of genetic algorithms for the optimization of n-location inventory systems with lateral transshipments: Together with Christian Hochmuth and Stefanie Thiem, Lässig is working on a book chapter about simulation-based optimization of n-location inventory systems with lateral transshipments using genetic algorithms. Simulation-based optimization is used to optimize complex systems that cannot be handled analytically. It turns out that genetic algorithms have a few advantages compared to other methods for this task, where particle swarm optimization and ensemble-based threshold accepting have been investigated as well. The chapter is almost ready and has to be submitted until February 24. Christian Hochmuth, who is currently working on his PhD at Bosch Rexroth in Germany, will visit the institute from May through June.

Lässig continues to work with Dirk Sudholt on topics of migration strategies in parallel evolutionary algorithms, where now the influence of the different migration structures and parameters as the migration interval is investigated more deeply. It would be interesting to investigate these methods besides LOLZ (leading ones leading zeros) also in more detail for a more practically relevant problem. We plan to submit further work on the topic at PPSN (11th International Conference on Parallel Problem Solving from Nature).

Lässig and Benjamin Satzger are working on instance sensitive global optimization methods. The idea is to use machine learning techniques to accelerate optimization by choosing methods and parameters of methods based on the problem. Publications are not in the pipeline yet.

3.7 Planning

Planning is recognized as a means of controlling and coordinating autonomous agents, and can be used as a vehicle to build computer systems with so-called self-capabilities. Such applications tend to result in repeatedly occurring planning tasks within the same domain. It would be desirable to take information obtained by successfully solved problems and use it to improve planning performance over time.

Given a representation of an initial state, goal states, and actions, planning deals with finding a sequence of actions transferring the initial state into a goal state. A straightforward way to solve planning problems is to search the space of planning states. Combined with sophisticated heuristic estimates to guide the search, this approach turned out to be highly competetive. Planning heuristics are often based on the idea of problem relaxation and use the solution to a relaxed problem to produce estimations.

Benjamin Satzger has worked on the generation of search heuristics for state-space planning using machine learning algorithms. Plans resulting from successfully solved problems can be used to train heuristics. These can perform equally well than the most successful traditional heuristics but are able to generate estimates much faster. In this way the overall planning performance can be increased significantly.

References

[1] F. Bagci, F. Kluge, B. Satzger, and T. Ungerer (2009). "Towards Indoor Location Estimation and Tracking with Wireless Sensors." Proceedings of the 6th IEEE International Symposium on Intelligent Signal Processing (WISP 2009), Budapest, Hungary, pp. 235-240, August 2009.

[2] J. Baumbach, S. Rahmann, and A. Tauch (2009). "Reliable Transfer of Transcriptional Gene Regulatory Networks Between Taxonomically Related Organisms." *BMC Systems Biology*, Vol. 3, Issue 8, January 2009.

[3] J. Baumbach, S. Rahmann, and A. Tauch (2009). "Towards the Integrated Analysis, Visualization, and Reconstruction of Microbial Gene Regulatory Networks." *Briefings in Bioinformatics*, Vol. 10, Issue 1, pp. 75-83, January 2009.

[4] J. Baumbach, T. Wittkop, K. Kleindt, and A. Tauch (2009). "Integrated Analysis and Reconstruction of Microbial Transcriptional Gene Regulatory Networks Using CoryneRegNet." *Nature Protocols*, Vol. 4, Issue 6, pp. 992-1005, June 2009.

[5] J. Baumbach, T. Wittkop, J. Weile, T. Kohl, and S. Rahmann (2008). "MoRAine - A Web Server for Fast Computational Transcription Factor Binding Motif Re-Annotation." *Journal of Integrative Bioinformatics*, Vol. 5, Issue 2, pp. 91, August 2008.

[6] S. Bruckner, F. Hüffner, R. M. Karp, R. Shamir, and R. Sharan (2009). "Topology-Free Querying of Protein Interaction Networks." Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology (RECOMB2009), Tucson, Arizona, pp. 74-89, May 2009.

[7] S. Bruckner, F. Hüffner, R. M. Karp, R. Shamir, and R. Sharan (2009). "Torque: Topology-Free Querying of Protein Interaction Networks." *Nucleic Acids Research*, Vol. 37, pp. 106-108, July 2009.

[8] S. Bruckner, F. Hüffner, R. M. Karp, R. Shamir, and R. Sharan. "Torque: Topology-Free Querying of Protein Interaction Networks." To appear in the *Journal of Computational Biology*.

[9] T. Connelly, P. Sharp, D. Ausiello, M. Bronner-Fraser, I. Burke, J. Burris, J. Eisen, A. Janetos, R. M. Karp, P. Kim, D. Lauffenburger, M. Lidstrom, W. Lim, M. McFall-Ngai, E. Meyerowitz, and K. Yamamoto (2009). "A New Biology for the 21st Century."

To appear in a report of the Committee on a New Biology for the 21st Century: Ensuring the United States Leads the Coming Biology Revolution, National Academies Press, 2009.

[10] D. Emig, N. Salomonis, J. Baumbach, T. Lengauer, B. R. Conklin, and M. Albrecht (2010). "Statistical and Visual Analysis of Exon Expression Data Integrated with Molecular Networks and Pathways." Under review, 2010.

[11] T. Friedrich, P. S. Oliveto, D. Sudholt, and C. Witt (2009). "Analysis of Diversity-Preserving Mechanisms for Global Exploration." *Evolutionary Computation*, Vol. 17, Issue 4, pp. 455-476, December 2009.

[12] G. Hannum, R. Srivas, A. Guenole, H. van Attikum, N. J. Krogan, R. M. Karp, and T. Ideker (2009). "Genome-Wide Association Data Reveal a Global Map of Genetic Interactions Among Protein Complexes." *PLOS Genetics*, Vol. 5, Issue 12, e1000782, December 2009.

[13] C. Horoba and D. Sudholt (2009). "Running Time Analysis of (ACO) Systems for Shortest Path Problems." Proceedings of the Engineering Stochastic Local Search Algorithms (SLS 2009), Brussels, Belgium, pp. 76-91, September 2009.

[14] C. Horoba and D. Sudholt (2010). "Ant Colony Optimization for Stochastic Shortest Paths." Submitted to Genetic and Evolutionary Computation Conference (GECCO 2010), Portland, Oregon, July 2010.

[15] T. E. Ideker, T. Thorsson, and R. M. Karp (2000). "Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design." Proceedings of the Pacific Symposium on Biocomputing (PSB 2000), Oahu, Hawaii, pp. 302-313, January 2000.

[16] T. Jansen and D. Sudholt (2010). "Analysis of an Asymmetric Mutation Operator." *Evolutionary Computation*, Vol. 18, No. 1, pp. 1-26, January 2010.

[17] J. Keilwagen, J. Baumbach, T. Kohl, and I. Grosse (2009). "MotifAdjuster: A Tool for Computational Reassessment of Transcription Factor Binding Site Annotations." *Genome Biology*, Vol. 10, Issue 5, p. R46, May 2009.

[18] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker (2003). "Conserved Pathways Within Bacteria and Yeast as Revealed by Global Protein Network Alignment." *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 100, Issue 20, pp. 11394-11399, September 2003.

[19] B. Kirkpatrick (2010). "Haplotypes Versus Genotypes on Pedigrees." Submitted to the 18th International Conference on Intelligent Systems for Molecular Biology (ISMB 2010), Boston, Massachusetts, June 2010.

[20] B. Kirkpatrick, J. Rosa, E. Halperin, and R. M. Karp (2009). "Haplotype Inference in Complex Pedigrees." Proceedings of the 13th Annual International Conference on

Research in Computational Molecular Biology (RECOMB 2009), Tucson, Arizona, pp. 108-120, May 2009.

[21] B. Kirkpatrick, J. Rosa, E. Halperin, and R. M. Karp. "Haplotype Inference in Complex Pedigrees." To appear in the *Journal of Computational Biology*.

[22] J. Krawczyk, J. Kalinowski, A. Goesmann, and J. Baumbach (2009). "From Corynebacterium Glutamicum to Mycobacterium Tuberculosis - Towards Reliable Transfers of Gene Regulatory Networks and Integrated Data Analyses with MycoRegNet." *Nucleic Acids Research*, Vol. 37, Issue 14, e97, August 2009.

[23] J. Lässig, S. Heinrich, and U. Trommler (2009). "Aufträge, Partner, Kosten - Kooperation Statt Konkurrenz." *IT&Production*, ISSN 1439-7722, Issue 10, pp. 74-76, November 2009.

[24] J. Lässig and K. H. Hoffmann (2009). "On the Structure of a Best Possible Crossover Selection Strategy in Genetic Algorithms." Proceedings of the 29th SGAI International Conference on Artificial Intelligence (AI 2009), Cambridge, England, pp. 263-276, December 2009.

[25] J. Lässig and K. H. Hoffmann (2009). "How to Choose Distribution Functions in Global Optimization Heuristics Optimally." Submitted to *Journal of Global Optimization*, December 2009.

[26] J. Lässig and D. Sudholt (2010). "The Benefit of Migration in Parallel Evolutionary Algorithms." Submitted to Genetic and Evolutionary Computation Conference (GECCO 2010), Portland, Oregon, July 2010.

[27] 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.

[28] F. Neumann, D. Sudholt, and C. Witt (2010). "A Few Ants Are Enough: (ACO) with Iteration-Best Update." Submitted to Genetic and Evolutionary Computation Conference (GECCO 2010), Portland, Oregon, July 2010.

[29] B. Pasaniuc, R. Avinery, T. Gur, C. F. Skibola, P. M. Bracci, and E. Halperin. "A Generic Coalescent-Based Framework for the Selection of a Reference Panel for Imputation." Under review.

[30] B. Pasaniuc, R. Garfinkel, I. I. Mandoiu, and A. Zelikovsky. "Optimal Testing of Digital Microfluidic Biochips." To appear in *INFORMS Journal on Computing*.

[31] B. Pasaniuc, J. Kennedy, and I. I. Mandoiu (2009). "Imputation-Based Local Ancestry Inference in Admixed Populations." Proceedings of the Fifth International Symposium on Bioinformatics Research and Applications (ISBRA 2009), Fort Lauderdale, Florida, pp. 221-233, May 2009.

[32] B. Pasaniuc, S. Sankaraman, G. Kimmel, and E. Halperin (2009). "Inference of Locus-Specific Ancestry in Closely Related Populations." *Bioinformatics*, Vol. 25, No. 12, pp. i213-i221, June 2009. Presented at the 17th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), Stolkhom, Sweden, June 2009.

[33] B. Pasaniuc, N. Zaitlen, and E. Halperin (2010). "Accurate Estimation of Expression Levels of Homologous Genes in RNA-Seq Experiments." To appear in the proceedings of the 14th International Conference on Research in Computational Biology (RECOMB 2010), Lisbon, Portugal, April 2010.

[34] R. Röttger and J. Baumbach (2009). "Estimating the Size and Completeness of Gene Regulatory Networks." Proceedings of the 10th Annual Symposium on Biomedical Computation at Stanford (BCATS 2009), Stanford, California, November 2009.

[35] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin (2009). "Genomic Privacy and Limits of Individual Detection in a Pool." *Nature Genetics*, Vol. 41, No. 9, pp. 965-967, September 2009.

[36] T. Sauerwald and D. Sudholt (2010). "A Self-Stabilizing Algorithm for Cut Problems in Synchronous Networks." *Theoretical Computer Science*, Vol. 411, Issues 14-15, pp. 1599-1612, March 2010.

[37] J. Scott, T. Ideker, R. M. Karp, and R. Sharan (2006). "Efficient Algorithms for Detecting Signaling Pathways in Protein Interaction Networks." *Journal of Computational Biology*, Vol. 13, Issue 2, pp. 133-144, March 2006.

[38] R. Sharan and T. Ideker (2006). "Modeling Cellular Machinery Through Biological Network Comparison." *Nature Biotechnology*, Vol. 24, Issue 4, pp. 427-433, April 2006.

[39] D. A. Spielman and S.H. Teng (2008). "A Local Clustering Algorithm for Massive Graphs and Its Application to Nearly Linear-Time Graph Partitioning." *CoRRabs*, 0809.3232, 2008.

[40] D. Sudholt (2009). "The Impact of Parametrization in Memetic Evolutionary Algorithms." *Theoretical Computer Science*, Vol. 410, Issue 26, pp. 2511-2528, June 2009.

[41] D. Sudholt (2010). "Hybridizing Evolutionary Algorithms with Variable-Depth Search to Overcome Local Optima." To appear in *Algorithmica*, 2010.

[42] 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, 2010.

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

[44] D. Sudholt and J. Lässig (2010). "The Benefit of Migration in Parallel Evolutionary Algorithms." Submitted to Genetic and Evolutionary Computation Conference (GECCO 2010), Portland, Oregon, July 2010.

[45] S. Suthram, R. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, R. Sharan, and T. Ideker (2005). "Conserved Patterns of Protein Interaction in Multiple Species." *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 102, Issue 6, pp. 1974-1979, February 2005.

[46] S. Thiem and J. Lässig (2009). "Comparative Study of Different Approaches to Particle Swarm Optimization in Theory and Practice." To appear in *Particle Swarm Optimization: Theory, Techniques, and Applications*, 2009.

[47] I. Ulitzky, R. M. Karp, and R. Shamir (2008). "Detecting Disease-Specific Dysregulated Pathways via Analysis of Clinical Expression Profiles." Proceedings of the 12th Annual International Conference on Research in Computational Moleculary Biology (RECOMB2008), Singapore, pp. 347-359, March 2008.

[48] T. Wittkop, D. Emig, S. Lange, J. H. Morris, S. Rahmann, J. Stoye, M. Albrecht, and J. Baumbach (2010). "Clustering by Unraveling Hidden Transitive Substructures." Under review, 2010.

[49] T. Wittkop, S. Rahmann, and J. Baumbach (2010). "Efficient Online Transcription Factor Binding Site Adjustment by Integrating Transitive Graph Projection with MoRAine 2.0." To appear in *Journal of Integrative Bioinformatics*. Presented at the 6th International Symposium on Integrative Bioinformatics (IB2010), Cambridge, United Kingdom, March 2010.

[50] N. Zaitlen, B. Pasaniuc, T. Gur, E. Ziv, 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.

4 Artificial Intelligence and its Applications

In 2009, the Artificial Intelligence Group made significant progress in both basic and applied projects. Both the main projects of the group, FrameNet (http://framenet.icsi.berkeley.edu) and NTL (http://www.icsi.berkeley.edu/NTL) received funding from highly competitive grants from the National Science Foundation. Building on the previous year's effort, a major new thrust of the group is focusing on Natural Language Processing (NLP) for social development. The basic research of the group continues to be language learning, computational biology, and neural modeling. In 2009, the applied role was expanded to the areas of NLP for developing regions, information technology for primary health care, crowd-sourcing of semantic annotation, and semantic 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 group's research included the following efforts.

   (a) A continuing effort on Predictive Analysis (PAINT) funded by IARPA in which the role of the ICSI team is to develop the probabilistic modeling framework for modeling biological and technological development pathways.

   (b) A collaborative project with Hesperian Press (http://www.hesperian.org) on multilingual semantic resource development for emerging regions funded by Google and the Rockefeller Foundation.

   (c) Two new exploratory efforts on metaphoric inference and metaphor corpus construction funded by NSF SGER grants.

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

   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.

In 2009, the AI Group investigated novel NLP and semantic computing techniques for providing multilingual health care information services to rural populations in developing countries. The ICSI AI Group is partnering with a well known non-profit organization, Hesperian Press (http://www.hesperian.org), whose books in over 80 languages on primary health care are being used by rural health workers in community-led efforts in over 100 countries around the world. Graduate student Matt Gedigian, working with Srini

Narayanan and with Hesperian Press, built the first version of the Hesperian Digital Commons (http://www.hesperian.net), which uses semantic wiki technology to combine wiki-based distributed, collaborative editing for semantic annotation and ontology generation from Hesperian materials. This technique has the potential of transforming the primary health care information produced by Hesperian into a semantic database greatly enhancing multimedia (including cellphone) and multilingual content retrieval and search. This project is expected to be one of the major foci for the ICSI AI Group in 2010. The project was awarded a Rockefeller Foundation planning grant for 2009.

The ICSI AI work continues to receive international recognition. Srini Narayanan was awarded a resident Fellowship (2008-2009) from the Institute for Advanced Study in Berlin to work on issues related to the embodiment of language. The yearly Fellowship is awarded to forty of the most promising international scholars from different disciplines spanning the humanities, the sciences, and mathematics. Narayanan was invited to be a resident Fellow as a member of a group comprising of philosophers, anthropologists, linguists, biologists, psychologists, and computer scientists studying the emergence and use of language from multiple perspectives. As part of the Fellowship, Narayanan gave a series of seminars on the work within the NTL group at ICSI. The group also organized an invited workshop on Embodiment and Language at the Institute for Advanced Study in Berlin that was attended by scholars from Europe, North America, and Asia.

Detailed accounts of progress in specific projects in 2009 follows.

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 the group's work continues to yield a variety of theoretical and practical findings. In 2009, 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 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 is being used in a number of courses at UC Berkeley and elsewhere.

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

One core NTL computational question is finding the best match of constructions to an utterance in a 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.

John Bryant finished his dissertation on 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 interpreta-

tion 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. The resulting program plays an central role in the linguistics doctoral thesis of Ellen Dodge, which is nearly complete, and the completed computer science doctoral work of Eva Mok.

One major milestone in 2009 was the appearance of an ECG handbook chapter by Feldman, Dodge, and Bryant which presents a coherent picture of the ECG development for linguists [22]. This puts ECG on an equal footing with much older established grammar approaches. An additional step forward will be a 2010 handbook on computational approaches to construction grammar, edited by Hans Boas. There are five chapters covering various aspects of ECG, including Bryant's thesis described above.

A major recent addition to handle morphology was an undergraduate honors project by Nate Schneider, who is now in graduate school at CMU. While at CMU, Schneider extended this work to also treat English morphology. A second undergraduate honors project, by Luca Gilardi, added a beautiful graphical interface to Bryant's best-fit analyzer. This combined system has proven very valuable in our research and is also being used in courses. Gilardi continues to refine the interface and work with others in the group on their projects.

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 is incorporated into her thesis research, which focuses on the development of an algorithm that learns such constructions from a set of utterance-situation pairs. She completed her dissertation early in 2009 and is now at the SONY Paris lab. She presented work at several European labs in 2009.

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. Argument omission in pro-drop languages exacerbates the problem by making the meaning of utterances heavily dependent on context. Aspects of this problem are well-studied by psychologists: children's social-intentional abilities in service of language learning, the development of syntactic knowledge, and the implicit learning of statistical regularities in the language input. However, most accounts of grammar development underspecify the learning processes involved. Using Embodied Construction Grammar and extending the work of Nancy Chang, Mok's research represents a first step toward 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 of Chang and Mok will feature prominently in the forthcoming Hans Boas book.

One area of progress in 2009 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. 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. Both of these results will be chapters in the forthcoming book by Hans Boas.

The group continues to be active in exploring the computational and scientific foundations of the Neural Theory of Language. Leon Barrett worked primarily on three projects in 2009. A paper with Feldman and MacDermed, published in *Neural Computation*, explores a new solution to the classical problem of neural binding of variables. He and Srini Narayanan developed an idea and technique for extending standard reinforcement learning techniques to get a much more comprehensive class of result. Essentially, this method can solve not just a single RL problem, but also, in doing so, it simultaneously solves an entire class of related problems. This was published and presented in a leading annual conference, ICML. Feldman published a review article describing the mythical neural code, covering both traditional and currently fashionable utility theory approaches [21].

For his dissertation, Barrett has 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 what 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 [49] 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. Bergen has taken a new tenured position at UC San Diego, one of the leading institutions in the field. Bergen and Feldman have completed a major invited book chapter showing how NTL helps explain the ancient mystery of how people can learn new concepts [8]. This is also being used in various classes.

In 2008 and 2009, the NTL group received funding from NSF (PI: Narayanan) for a new project on metaphor inference. The funding is for preparatory work addressing the key scientific challenges posed by the creation of a computational model of metaphor. The work conducted in 2009 builds on previous modeling work within the NTL group and leverages results from cognitive linguistics to explore techniques to a) design and populate a machine readable metaphor ontology, b) analyze the metaphoric encoding of crucial discourse information, including event structure and communicative intent, and c) use machine learning algorithms for metaphor recognition from textual sources. There are prospects for a large national project building on these results.

There was a very significant increase in the use of the group's results in UC Berkeley courses and in linguistics research. Collaboration with the FrameNet project has been broadened and deepened with positive results for both efforts, some of which are described in this report. Feldman ran an interdisciplinary class in Spring 2008 and several of the research efforts from that class were incorporated into the project in 2009 [22]. George Lakoff is incorporating much of the new NTL book into his undergraduate course and it is being widely used elsewhere. An ongoing collaboration of Feldman and Lakoff with Manuel Castells of the University of Southern California resulted in NTL models being

given a prominent role in Castells's new book [13].

There were also several invited talks and seminars presented by NTL members. Feldman presented the NTL project at the Cognitive Science conference and the NTL project was the main subject of the UC Berkeley Cognitive Science Faculty retreat in December 2008. The Linguistics Summer Institute was held in Berkeley in 2009 and NTL was featured in courses by Lakoff and Feldman. Srini Narayanan has begun to introduce NTL ideas into the first course in cognitive science at UC Berkeley in Spring 2010. Bergen and Nancy Chang gave a series of invited lectures on Introduction to Construction Grammar at the Summer School on Computational Linguistics, Cortona, Italy, September.

4.2 Metaphor Inference

In 2009, the NTL group (PI: Srini Narayanan) completed preparatory work addressing the key scientific challenges posed by the creation of a metaphor inference system. This one-year effort was in response to an NSF panel recommendation (on proposal 0812401 to IIS RI) to fund a pilot project before a full-scale effort be undertaken. The work focused on issues of metaphor representation and inference. A companion planning grant from the CRI program was awarded in 2010 to focus on the attendant issues of corpus selection, preparation, creation, and annotation.

During the course of the one-year pilot effort, we performed four tasks that were geared toward a better understanding of the challenges in building a scalable metaphor inference system. Specifically, the following four tasks were performed either fully or partly funded by the NSF SGER project.

1. A survey of computational modeling and inference with figurative language performed by the PI and postdoctoral researcher Birte Loenneker-Rodman. Our efforts and results will appear in an article [36] entitled "Computational Models of Figurative Language" in *The Cambridge Encyclopedia on Psycholinguistics*.

2. Srini Narayanan and graduate student Matt Gedigian of the NTL group continued work on metaphor recognition, extraction, and analysis from text. One long standing question that was addressed this year was the linguistic representation of metaphor that could support deep semantic analysis. Our previous work involved a metaphor inference system, KARMA (described in previous reports and in [44, 20, 23]), which relied on partially hand parsed input. Graduate student John Bryant (partly funded by this effort) developed a semantic analyzer for an Embodied Construction Grammar (ECG) for his 2008 UC Berkeley PhD thesis. Luca Gilardi (an undergraduate student) working with Jerome Feldman developed a representation of metaphors in ECG that can work with the Bryant semantic analyzer. Their paper [22] describes this in detail.

3. Previous work provided evidence on the ubiquity of metaphor and was able to model the wide range of inferences that resulted from metaphoric discourse in newspaper stories on politics and economics. As part of this project, we performed a pilot investigation of the empirical predictions of the KARMA model of metaphor. The papers [46] and [45] have details on the various predictions of the model.

4. One central open question in metaphor research is the automatic acquisition of metaphors. While there were descriptive theories of the process, there has not been any computational model of metaphor acquisition that is consistent with the convergence of psychological and linguistic evidence. During this year, we came up with the first neurally plausible computational model of metaphor acquisition. The exciting aspect of this model is that it uses robust biological data on Spike Timing Dependent Plasticity (STDP) in combination with the psycholinguistic and developmental evidence on the metaphor comprehension to propose a model of metaphor acquisition that appears to fit the developmental sequence observed in the data. This is ongoing work by the PI. Current results are reported in [46]. Section IV has further details. Together, the four efforts outlined above give us confidence that we can build robust computational models that can recognize, represent, reason with, and learn metaphor mappings. Furthermore, our empirical experiments underline the ubiquity of metaphor in most discourse and language use (from newspaper articles to technical stock and market commentary) and the utility of building systems that can deal with figurative language.

4.3 Probabilistic Models for Pathway Analysis

Our pathway inference research was conducted within the context of a larger project (PAINT) funded by IARPA. In 2009, postdoctoral Fellow Steve Sinha worked on this problem with Srini Narayanan. The main goal of the ICSI work was to classify a partially-observable real world production pathway as fitting one of a given set of hypotheses about it.

Classification of dynamic system pathways is a challenging and important research problem that can be addressed using event modeling and reasoning. Pathways are evolution trajectories of complex dynamic systems that serve a particular goal. Examples include metabolic pathways which express protein products or technological/engineering processes with specific goals (such as vaccine research or production). To classify an unknown, partially-observable pathway requires comparing the behavior of the unknown pathway to the expected behavior of each pathway hypothesis. In 2009, the PAINT program demonstrated algorithms for a classification system in which ICSI created a central component, the Pathway Inference Engine.

Given the default set of inputs for an unknown pathway, observations of the evolution of the pathway may be insufficient for hypothesis disambiguation (determining which of the several hypotheses about the pathway is likely). The observations expected under each hypothesis for that input set may be insufficiently distinguishable. The question then becomes one of constructing *probes*. Probes are external events that are interventions designed to produce effects on the observable segments. They can be passive (measurements on observables, e.g., monitoring an additional resource) or active (structural or input changes to the pathway, e.g., changing the resource profile for a specific segment). The effect of a probe can manifest itself differently for each hypothesis, ideally creating better separation in the expected observations and thus greater diagnosticity. Probes inherently trade the cost of probing (how easy is it, how detectable it is, how expensive it is, etc.) with the value of the information obtained. We proposed a method for calculating the

information value of probes. When weighted by the costs of the probes (which are application specific), the information value metric can be used to choose the optimal probe to help us answer Hypothesis Disambiguation questions.

We evaluated this system in three multi-university demonstrations and at a final 2009 community wide PAINT Phase I meeting. The purpose of the demos was to test and show how active probes could increase diagnosticity (using our value of information metrics), providing greater separation of hypotheses. This is an important step before classifying a real unknown pathway. Our efforts within the PAINT program have much wider applications in identification of partially observable dynamic systems which we plan to explore in 2010. The hypothesis disambiguation problem forms a core test case for the ICSI probabilistic modeling framework and is described more in [55].

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

Bringing health information access to all communities in the world is a grand scientific challenge that cuts across the disciplinary boundaries and is an opportunity to demonstrate the value and potential of a universal human network. AI Group PI Srini Narayanan and graduate student Matt Gedigian have been collaborating 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. 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 three languages (English, Spanish, and Tamil) [47]. During 2010, the effort is being extended to cover and adapt to other languages, modalities, and materials. The effort has attracted considerable interest with other groups at ICSI, UC Berkeley, and other universities. One possible avenue of extension being explored involves close

collaboration with the Speech Group at ICSI on multilingual cell phone access to and dissemination of the HDC information. Another avenue is deeper semantic annotation of the materials using FrameNet frames. Both these ideas have led to proposal submissions to the NSF in 2010. The design and implementation of the HDC is iterative and ongoing field tests of the Digital Commons in multiple countries will inform the specific form and functionality of current and future developments of the resource.

4.5 ANC MASC Collaboration

Under a subcontract on NSF grant #0708952 "CRI: CRD: A Richly Annotated Resource for Language Processing and Linguistic Research" (PI:Nancy Ide of Vassar), the FrameNet group is annotating texts from the American National Corpus (ANC) ([30], `http://www.anc.org`). 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 plans to do the manual annotation on a small portion of the corpus, training automatic semantic role labeling (ASRL) software to do the rest.

We continued our collaboration with the ANC in 2009, shifting from annotating full texts to annotating sample sentences for a growing list of words that the ANC team at Vassar and Columbia Universities 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.

4.6 WordNet-FrameNet Alignment

We have continued work funded under NSF #0705155 on aligning FrameNet with WordNet (WN) [40, 24], the largest machine-readable English lexicon. In 2009, we worked on detailed comparisons of WN and FN sense divisions for roughly 60 common words, in part for the WN-FN alignment study and also annotating example sentences as a contribution to the ANC MASC project; these words are being annotated in the MASC by annotators at Vassar and Columbia. The method is as follows:

1. A list of roughly 10 words, including nouns, verbs, and adjectives, all displaying moderate polysemy, are chosen by agreement between the FrameNet staff and Christiane Fellbaum of Princeton University.

2. Then the annotators annotate a pilot random sample of 50 sentences containing the word (strictly speaking, one of the wordforms of the lemma) with the existing WordNet senses, adding notes of any senses which seem hard to distinguish, and any sentences which seem not to fit in any of the existing senses.

3. Next, Fellbaum and the FrameNet team study the sense divisions in WordNet and those in FrameNet in the light of the results of the pilot annotation; they decide whether changes are needed in either resource, in the light of how the other has handled the word.

In the most common case, WordNet has more senses for the word than FrameNet, and both sides agree that some of the WordNet senses can be collapsed, and that some senses need to be added to FrameNet corresponding to existing WordNet senses. Sometimes this process results in complete agreement on the number of senses, with a one-to-one alignment possible.

More generally, however, such perfect agreement is not reached. For example, WN contains many examples of extremely rare or obsolete senses of words, (e.g. *plug*.n 'an overworked horse'; *cheese*.v 'wind into a cheese; *cheese the yarn*'). FN deliberately omits these on the grounds that matches on them are much more likely to be wrong than right in most texts handled in NLP.

4. Once a canonical list of senses for WordNet is decided, the annotators at Vassar and Columbia annotate up to 1000 sentences containing each word in text from the ANC, marking the WN sense on each.

5. From those WN annotated sentences, a random sample of 100 sentences is drawn and sent to ICSI, where FN annotators label the frame of each instance of the word and the frame elements found in the sentence. These annotations are sent back to be merged into the ANC.

The results to date show that a real alignment of those LUs that appear in both WN and FN will require changes to both resources [5].

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 [25] of the University of Alicante and German Rigau (forthcoming) of Universidad Politécnica de Valencia.

4.7 Development of Word Sketch Engine for Rapid Vanguarding

We are continuing development of a new system for the vanguarding portion of the FrameNet work, consisting of an implementation of a GUI based on the Word Sketch Engine [32, 33]. The original grant, NSF #0535297 "Rapid Development of a Frame Semantic Lexicon," has ended, but we have been able to obtain a second one-year, no cost extension. In 2009, we completed the software development needed to integrate the GUI with the FNDesktop annotation system, including encoding of the full ANC so that it can be used in the new interface. An experienced annotator has begun using the new system for subcorporation (the extraction and grouping of sentences for annotation according to their syntactic structure and the occurrence of particular collocations). The new system is roughly twice as fast as the older, rule-based system, since the collocates to be found are determined automatically according to statistical criteria, and subcorpora for several senses of the word can be created simultaneously. However, because of the statistical cutoffs, some collocates which the annotator knows to exist (by virtue of being a speaker of the language) may not appear in the automatically-generated tables. In cases where such collocates are important, it has been necessary to revert to the old system of writing rules out manually.

4.8 Crowdsourcing

Our earlier proposal for developing games for FrameNet annotation was not approved, but we were able to submit and were approved for a smaller Exploratory Grant for Experimental Research (IIS-0947841) to test the waters in this new field. We have rehired Jisup Hong as a programmer and he 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 goal is to test whether or not we can collect annotations of sentences in this way, but we are beginning with a simpler task: i.e., asking workers to decide which frame a given instance of a polysemous lemma is in. 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 Element of that frame. In other words, the task is being broken down into two stages and we will soon be testing Mechanical Turk for the first stage. Previous research suggests that it is necessary (and sufficient) to gather roughly eight responses to each item and then to combine together to produce the equivalent of one response from an expert [57]. This may seem wasteful, but it is likely that the cost of gathering the non-expert responses will still be less than the cost of the single expert response. We hope to have some preliminary results on this line of research in the spring of 2010.

4.9 Conference on Upgrading FrameNet

Our CRI proposal on upgrading FrameNet was declined on the grounds that we had not demonstrated sufficient support for FrameNet in the NLP community. However, we received a grant (CNS-0855271) 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.

We plan to hold this conference in April, with two venues, the UC Berkeley campus and the Princeton University campus. We will provide good audio and video links between the East and West Coast participants so that it will be "virtually" a single conference, but with reduced costs and travel time. We are currently deciding on the the names of invitees and the topics to be discussed; we plan to set up a wiki (`http://en.wikipedia.org/wiki/Wiki`) to allow the discussion to begin before the actual meeting and to continue after it is over.

4.10 Development of the FrameNet Database

During 2009, the number of frames increased by 18 to 976; the number of lexical units (LUs) in the FrameNet database increased by 168 to 11,709. Of these, 109 were new LUs in existing frames; examples of these are shown in Table 1 on page 57. The other 59 LUs were in the new frames; examples of these new frames and LUs are shown in Table 2. The theory behind the database, including an analysis of different ways of extending FrameNet to other languages was discussed in a new article [35].

Attention_getting	*excuse me.a, hello there.a, yoohoo.intj*
Avoiding	*stay away.v*
Bearing_arms	*draw.v*
Being_obligated	*assignment.n, commission.n, contract.n, mission.n*
Being_operational	*functional.a*
Causation	*so.c*
Desiring	*loath.a, reluctant.a*
Fall_asleep	*drift off.v, faint.v, nod off.v, pass out.v, zonk out.v*
Frequency	*daily.adv*
Likelihood	*assured.a, long.a*
Locale_by_use	*port.n, work.n*
Locative_relation	*throughout.prep*
Measurable_attributes	*deep.a, high.a, long.a, tall.a, thick.a*
Medical_conditions	*AIDS.n, Alzheimer's.n, disorder.n*
Origin	*Jamaican.a, Portuguese.a*
Part_inner_outer	*central.a*
People_by_vocation	*plain-clothes man.n*
Perception_body	*be killing.v*
Range	*sight.n*
Reshaping	*curl.v*
Rite	*exercise.n*
Setting_fire	*ablaze.a, alight.a, on fire.a, set fire to.v*
Shapes	*ellipse.n, line.n, oval.n, triangle.n, wedge.n*
Sleep	*out.a, unconscious.a*
Successful_action	*go wrong.v*
Waking_up	*come back around.v*

Table 1: Examples of New Lexical Units Created in 2009 in Existing Frames

4.11 Visitors and Events

Carlos Subirats-Rüggeberg of Universidad Autónoma de Barcelona, head of the Spanish FrameNet effort, was at ICSI for most of 2009; the development of Spanish FrameNet is proceeding rapidly, closely based on the English FrameNet frames, as discussed in [59, 58, 60].

Another long-term visitor in 2009, Bernd Bohnet, a German postdoctoral Fellow, officially came to ICSI on a DAAD fellowship to work with the Speech Group, but also worked on semantic role labeling for FrameNet while here from September, 2008 to August, 2009.

Ephrem Fernández of the University of Texas at San Antonio visited for about three weeks in December; his speciality is evaluation of patient's reports of pain, and he discussed his work and the FrameNet work on vocabulary in this domain with us.

Shorter-term visitors included Patrick Svensson, Director of the Humanities Computing Lab at UmeåUniversity (Sweden); Aldo Gangemi, of ISTC-CNR in Rome, who is researching how to align FrameNet with ontologies; Buonaventura Coppola, of the University of Trento, who works on FN semantic role labeling: and Jaqueline Visconti, who is helping

Attempt_means	*try.v*
Capacity	*fit.v, seat.v, sleep.v, take.v*
Cause_to_perceive	*demonstrate.v, depict.v, exhibit.v, point.v, present.v, represent.v, show.v*
Concessive	*although.scon, but.c, despite.prep, however.adv, in spite of.prep, nevertheless.adv, though.scon, to be fair.adv, while.scon*
Event_instance	*once.adv, time.n, twice.adv*
Labor_product	*oeuvre.n, work.n*
Locale_by_ownership	*estate.n, land.n, property.n*
Location_in_time	*time.n*
Long_or_short_selling	*sell long.v, sell short.v*
Manipulate_into_shape	*coil.v, loop.v, twist.v, wind.v*
Pattern	*formation.n, pattern.n*
Representing	*be symbol.v, symbolize.v*
Sequence	*order.n, sequence.n, sequential.a, series.n, succession.n, successive.a*
Sound_level	*loud.a, quiet.a*
Timespan	*Day.n, time.n*
Tolerating	*bear.v, endure.v, stand.v, tolerant.a, tolerate.v, toleration.n*
Turning_out	*end up.v, prove.v, turn out.v*

Table 2: New Frames (and Lexical Units) Created in 2009

to develop an Italian FrameNet in the legal domain.

From July 31 to August 2, a conference entitled "Frames and Constructions" was held at UC Berkeley to honor the founder of FrameNet, Charles J. Fillmore. More than 50 papers were presented, roughly half of which were primarily about frame semantics. The complete program with abstracts is available at `http://linguistics.berkeley.edu/~fillmorefest/program.html`. Many of the papers were about FrameNets in other languages, in various stages of development; many of these projects are discussed in more detail in [11]. One of the speakers recently completed a dissertation on using FrameSemantics in the domain of molecular biology [18].

References

[1] L. von Ahn and L. Dabbish (2004). "Labeling Images with a Computer Game." Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '04), Vienna, Austria, pp. 319-326, April 2004.

[2] 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.

[3] 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, Issue 3 & 4, pp. 229-238, September 2008.

[4] C. 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.

[5] 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.

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

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

[8] B. Bergen and J. Feldman (2008). "Embodied Concept Learning." In *Handbook of Cognitive Science: An Embodied Approach*, P. Calvo and T. Gomila, eds., pp. 313-332, Elsevier Ltd., 2008.

[9] 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.

[10] H. Boas (2005). "Semantic Frames as Interlingual Representations for Multilingual Lexical Databases." *International Journal of Lexicography*, Vol. 18, Issue 4, pp. 445-478, December 2005.

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

[12] J. Bryant (2008). "Best-Fit Constructional Analysis." Ph.D. thesis, UC Berkeley EECS, 2008.

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

[14] N. Chang (2008). "Constructing Grammar: A Computational Model of the Emergence of Early Constructions." Ph.D. thesis, UC Berkeley EECS, 2008.

[15] N. Chang and E. Mok (2006). "Putting Context in Constructions." Presented at the Fourth International Conference on Construction Grammar (ICCG4), Tokyo, Japan, Spetember 2006.

[16] 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.

[17] 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*, H. Cohen and C. Lefebvre, eds., pp. 224-242, Elsevier, 2005.

[18] A. Dolbey (2009). "BioFrameNet: a FrameNet Extension to the Domain of Molecular Biology." Ph.D. thesis, UC Berkeley, 2009.

[19] 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.

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

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

[22] J. Feldman, E. Dodge, and J. Bryant (2009). "Embodied Construction Grammar." In *Oxford Handbook of Linguistic Analysis*, B. Heine and H. Narrog, eds., pp. 111-138, Oxford University Press, 2009. ¡http://www.icsi.berkeley.edu/cgi-bin/pubs/publication.pl?ID=002693¿.

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

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

[25] O. Ferrández, M. Ellsworth, R. Muñoz, and C. F. Baker (2010). "A Graph-Based Measure of FrameNet-WordNet Alignment." Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong, China, January 2010.

[26] C. Fillmore (1976). "Frame Semantics and the Nature of Language." In *Annals of the New York Academy of Sciences*, Conference on the Origin and Development of Language and Speech, Vol. 280, pp. 20-32, 1976.

[27] C. Fillmore, C. Baker, and H. Sato (2004). "Framenet as a 'Net."' Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp. 1091-1094, May 2004.

[28] C. Fillmore, J. Ruppenhofer, and C. Baker (2004). "FrameNet and Representing the Link Between Semantic and Syntactic Relations." In *Computational Linguistics and Beyond*, Language and Linguistics Monographs Series B, pp. 19-62, Academia Sinica Press, 2004.

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

[30] 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), Las Palmas, Canary Islands, Spain, pp. 839-44, May 2002. ¡http://americannationalcorpus.org/pubs.html¿.

[31] P. Kay (2005). "Color Categories are Not Arbitrary." *Cross-Cultural Research*, Vol. 39, Issue 1, pp. 72-78, February 2005.

[32] 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 Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the ACL (ACL-EACL 2001), Toulouse, France, pp. 32-38, July 2001.

[33] A. Kilgarriff and D. Tugwell (2001). "WASP-Bench: An MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation." Proceedings of the Machine Translation Summit VIII: Machine Translation in the Information Age (MT Summit VIII), Santiago de Compostela, Spain, pp. 187-190, September 2001.

[34] B. Löeker-Rodman (2007). "Beyond Syntactic Valence: FrameNet Markup of Example Sentences in a 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.

[35] 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.

[36] B. Loenneker-Rodman and S. Narayanan (2010). "Computational Models of Figurative Language." To appear in *Cambridge Encyclopedia of Psycholinguistics*, Cambridge University Press, 2010.

[37] J. Makin (2008). "A Computational Model of Human Blood Clotting: Simulation, Analysis, Control, and Validation." Ph.D. thesis, UC Berkeley EECS, 2008.

[38] J. Makin and S. Narayanan (2008). "A Hybrid System Model of Human Blood Clotting." Submitted. Also ICSI Technical Report TR-08-002, Berkeley, California, February 2008. ¡http://www.icsi.berkeley.edu/s̃narayan/clot.pdf¿.

[39] J. Makin and S. Narayanan (2010). "Real Time Control of Human Blood Clotting." Submitted to *PLoS Computational Biology*, 2010.

[40] G. 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, Issue 4, pp. 235-44, 1990.

[41] E. Mok (2008). "Contextual Bootstrapping for Grammar Learning." Ph.D. thesis, UC Berkeley EECS, 2008.

[42] 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.

[43] 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.

[44] S. Narayanan (1997). "KARMA: Knowledge-Based Active Representations For Metaphor and Aspect." Ph.D. thesis, UC Berkeley Computer Science, 1997.

[45] S. Narayanan (2009). "Testing the Predictions of a Computational Model of Metaphor." Proceedings of the International Conference on Cognitive Linguistics (ICLA 2009), January 2009.

[46] S. Narayanan (2010). "A Neurally Plausible Computational Model of Metaphor Acquisition." In preparation. 2010.

[47] S. Narayanan and M. Gedigian (2009). "The Hesperian Digital Commons: A Multilingual Primary Health Care Resource." Proceedings of the CRA-CCC Workshop on Computer Science and Global Development, Berkeley, California, August 2009.

[48] 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.

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

[50] S. Padó and M. Lapata (2009). "Cross-Lingual Annotation Projection of Semantic Roles." *Journal of Artificial Intelligence Research*, Vol. 36, pp. 307-340, September 2009.

[51] S. Petrov, L. Barrett, and D. Klein (2006). "Non-Local Modeling with a Mixture of PCFGs." Proceedings of the International Conference of Computational Natural Language Learning (CONLL), New York, New York, pp. 14-20, June 2006.

[52] S. Petrov, L. Barrett, R. Thibeaux, and D. Klein (2006). "Learning Accurate, Compact, and Interpretable Tree Annotation." Proceedings of the 21st International Conference of the Association of Computational Linguistics (ACL), Sydney, Australia, pp. 433-440, July 2006.

[53] T. Regier (1996). *The Human Semantic Potential*. MIT press, September 1996.

[54] J. Scheffczyk, C. F. Baker, and S. Narayanan (2010). "Ontology-Based Reasoning About Lexical Resources." To appear 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., Cambridge University Press, 2010.

[55] S. Sinha (2008). "Answering Questions About Complex Events." Ph.D. thesis, UC Berkeley EECS, 2008.

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

[57] 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 on Natural Language Processing (EMNLP), Honolulu, Hawaii, October 2008. ¡http://www.stanford.edu/j̃urafsky/amt.pdf¿.

[58] 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.

[59] 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.

[60] C. Subirats (2009). "La función del corpus en FrameNet Español." Proceedings of the First International Conference on Corpus Linguistics (CILC 09), Murcia, Spain, May 2009.

[61] L. von Ahn, M. Kedia, and M. Blum (2006). "Verbosity: A Game for Collecting Common-Sense Facts." Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '06), Montreal, Canada, pp. 75-78, April 2006.

5 Speech Processing

Speech efforts in 2009 were headed by research staff members Gerald Friedland, Dilek Hakkani-Tür, Adam Janin, Nelson Morgan, Elizabeth Shriberg (ICSI and SRI), and Andreas Stolcke (ICSI and SRI). Our work also continued to be bolstered by external collaborators such as Dan Ellis of Columbia University. 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 also played a critical role in our progress. Independent consultant George Doddington worked with the group to help formulate research directions for speaker recognition. We also partnered with IBM for work on speech recognition. There were also two activities within the group that were not, strictly speaking, speech processing: (1) UC Berkeley Faculty Associate Dan Klein worked on improving machine translation as part of our collaboration with DARPA contractors SRI and BBN; and (2) Suman Ravuri and Dan Ellis made significant progress on the musical application "cover song detection."

Major contributions were made by students, research associates, postdoctoral Fellows, and international visitors (http://www.icsi.berkeley.edu/groups/speech/members.html has a current list of group members, collaborators, and alumni). The sections below describe a number of the year's major activities in speech processing. 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 Speaker Diarization

We continued our work on low-latency diarization, multimodal diarization, and dialocalization (acoustic diarization and visual localization as joint optimization problem), we participated in the NIST RT'09 evaluation, and we improved the robustness of the diarization algorithms against varying recording length by developing a new initialization method for the ICSI speaker diarization engine. We also started work on parallelized diarization. In addition to improving diarization methodology, we are also exploring how speaker diarization methods can be used in various applications [11, 7].

Novel initialization methods for speaker diarization: We worked on novel initialization methods for the ICSI speaker diarization engine. We know that the current ICSI Speaker Diarization engine performs suboptimally on shorter segment durations, e.g., 300 seconds, 200 seconds, or even 100 seconds. Therefore the goal was to improve the performance on short segment durations. The reason for this was that some of the main parameters of the engine, namely the optimal number of initial Gaussians and the number of initial clusters, are highly dependent on the length of the overall meeting recording. Therefore we performed experiments to reduce the number of manually tunable parameters and get a more accurate speaker diarization result at the same time. We also experimented with a prosodic clustering approach to estimate the number of initial clusters. In the end,

the Diarization Error Rate (DER) on meetings of 100-second duration was improved by more than 50 % relative, measured on different NIST meeting sets.

The details of this work were published in papers presented at the IEEE ASRU workshop and IEEE ICASSP, as well as in an article in the *IEEE Transactions on Audio, Speech, and Language Processing* [23, 24, 25].

Participation in the NIST rich transcription 2009 evaluation: ICSI participated in the NIST Rich Transcription 2009 evaluation. The NIST Rich Transcription evaluation series promotes and gauges advances in the state-of-the-art in several automatic speech recognition technologies. The goal of the evaluation series is to create recognition technologies that will produce transcriptions which are more readable by humans and more useful for machines. As such, a set of research tasks has been defined that are broadly categorized as either Speech-to-Text Transcription (STT) tasks or Metadata Extraction (MDE) tasks. ICSI submitted eight different speaker diarization systems and collaborated with SRI in the speaker-adaptive-speech-to-text tasks. The diarization systems submitted varied in methodology and the data that they processed and are listed as follows:

1. A primary submission for the Multi-Distant-Microphone task with the prosodic feature initialization as described last quarter,
2. the same system as 1) as a secondary submission but with using prosodic features also for modeling speakers [12, 13],
3. the same two system as 2 for the single-distant microphone condition,
4. a low-latency system for the multiple-distant microphone condition,
5. a low-latency system for the single-distant-microphone condition,
6. a multi-modal system based on many cameras and a single-distant microphone,
7. the same system as 1) for the Mark-III Microphone Array Task, and
8. the same system as 1) or the all-distant-microphone task.

In addition we participated in four of the speaker-attributed speech-to-text tasks by combining the respective ICSI speaker diarization system with a speech recognizer from SRI.

The NIST RT evaluation is not a competition, as scores are only reported internally. However, we are permitted to report that we did extremely well. Unfortunately, ICSI was the only participant in any of the new tasks in 2010 (multimodal and low-latency) and the ADM and Mark-III tasks. The results were presented and discussed together with the other participants at the NIST RT workshop at the Florida Institute of Technology.

Multimodal diarization and dialocalization: We continued the experiments on multimodal diarization [10]. As reported in the previous annual report, we discovered a novel audio-visual approach for unsupervised speaker localization in both time and space by only minimally extending the ICSI multimodal speaker diarization engine. Using recordings from a single, low-resolution room overview camera and a single far-field microphone, ICSI's audio-only speaker diarization system (speaker localization in time) was extended so that both acoustic and visual models could be estimated as part of a joint unsupervised optimization problem. The speaker diarization system first automatically determines the speech regions and estimates "who spoke when," then, in a second step, the visual models are used to infer the location of the speakers in the video. We started to call this process "dialocalization." This year, we quantitatively investigated the different properties of the algorithm and also systematically investigated the differences between a multi-modal approach, a multi-sensor approach, and a multi-feature approach to this problem. Our descriptions of the experiments were accepted as a full paper at ACM Multimedia 2009 [14] and later extended for a submission to *ACM Transactions on Multimedia Computing, Communications and Applications*.

We also continued our investigations on novel multimodal features for speaker diarization. We studied different video features and audio features, as well as different feature-level integration strategies, such as mutual information.

Low-latency diarization: After the initial low-latency experiments conducted for the NIST RT'09 evaluation, we continued to work on developing a low-latency diarization system (aka online diarization system). In order to have a true speaker diarization system, no recording-specific training step was to be required a-priori (in contrast to the initial system submitted to RT'09, where an explicit per-meeting training step was still needed). Therefore, we experimented with an approach based on a Universal Background Model (UBM). The resulting system again consists of two subsystems: an offline subsystem, which runs in the background and generates speaker labels with high latency, and a low-latency subsystem, which uses the UBM to assign audio segments to speakers in real time. The models from the UBM are updated regularly using the labels generated by the offline subsystem. In this way, the output of the diarization starts with being almost random for the first few minutes and ends with offline accuracy after a certain recording time. New speakers introduced in the meeting might be classified correctly using the UBM but can usually be handled well once the offline system running in the background has updated the labels. Experiments on 4.5 hours of the AMI meeting dataset have shown that after 303 seconds of "booting," the system converges to a globally measured DER of 33.42 %. This is very close to the DER of the offline system, which was 32.09 %. We anticipate a submission to the ISCA Interspeech conference in 2010.

Parallel diarization: As part of the ParLab project, we started to investigate parallel variants of agglomerative hierarchical clustering. We started porting the core code of the ICSI speaker diarization engine to CUDA. As a side effect, we developed an algorithm that significantly eases the process by automatically classifying functions into "host" and "device" functions. Since device functions can only call certain system functions and not

call host functions, the algorithm constructs the call graph from the original (serial) source code and analyzes it to find conflicts. In order to parallelize the code using CUDA, the developer then only has to focus on the identified conflicting function calls. The work on the parallelization of the diarization will be continued in 2010 [33].

Speaker identification in cars: ICSI, together with the German Research Center for Artificial Intelligence (DFKI), organized the 1st International Workshop on Multimodal Interfaces for Automotive Applications (MIAA), at the IUI conference in Florida (sponsored by ACM SIGCHI) [32]. We also performed experiments to explore the difficulties of using state-of-the-art speaker recognition and diarization methods under realistic car noise conditions. We collected the two sets of car recordings by mounting an iPhone to the wind shield of an automatic Dodge Neon. The data was then annotated for speaker identification. We experimented with different filters and training models. The best result we could obtain on our initial experiments was 37.4 % DER and was presented at the IEEE International Conference on Semantic Computing [21].

Tele-immersion: Tele-immersion aims to enable users in geographically distributed sites to collaborate in real time in shared simulated environments ("intelligent spaces") as if they were in the same physical space. Motivated by the results of the Dialocalization system (see above), one idea was to extend the system to 3D localization. We therefore conducted preliminary experiments in collaboration with UC Berkeley's Tele-Immersion Lab where four microphones, one at each corner of the recording area, recorded the sound. We found that using time-delay-of-arrival features, as used in the NIST RT'09 evaluation, already enabled localization within about 10 cm accuracy even with very few samples.

Speaker diarization as front-end for speech recognition: While often used as the basic motivation for speaker diarization research, it also seems intuitively clear that speech recognition in a multi-person scenario could benefit from the output of speaker diarization for effective feature normalization and model adaptation. Such benefits have been elusive in the very challenging domain of meeting recognition from distant microphones. In a study conducted together with SRI we investigated the problem and showed that recognition gains are indeed possible but need careful postprocessing of the diarization output. Still, recognition accuracy may suffer when the underlying diarization system performs worse than expected, even compared to far less sophisticated speaker clustering techniques. We obtained a more accurate and robust overall system by combining recognition output obtained with multiple speaker segmentations and clusterings. We evaluated our methods on data from the 2009 NIST Rich Transcription meeting recognition evaluation using different speaker diarization engines and postprocessing strategies. The results were published in an article that will be presented at IEEE ICASSP 2010 [37].

Navigation in sitcoms: The ACM Special Interest Group on Multimedia (SIGMM) published a set of problems "geared to engage the Multimedia research community in solving relevant, interesting and challenging questions about the industry's 2-5 year horizon for multimedia" and started a competition called the ACM Multimedia Grand Challenge.

As a response to the Yahoo! Video Segmentation task, we created a narrative-theme segmentation system for sitcoms, following common artistic production rules for sitcoms. We present the user with the basic semantic elements of a sitcom such as the scenes, punchlines, and individual dialog segments on top of a standard video player interface. A per-actor filter helps to search only for elements that contain certain protagonists. The user is able to selectively skip certain parts and to directly navigate into elements he or she remembers.

The system consists of two main elements: an analysis step and an online video browser. The analysis step consists of an acoustic event detection and speaker identification step, and of a semantic element analysis step. The online video browser then uses the output of the semantic element analysis step to present the video navigation interface to the user.

The system was tested on three different sitcoms: *Seinfeld*, *I Love Lucy*, and *The Big Bang Theory*. Even though we obtained a Diarization Error Rate of 23.25 %, in a subjective "Turing Test"-like evaluation we conducted, the test subjects weren't able to differentiate between the navigation based on manual annotation versus the automatically generated annotation [9].

After being selected for and presented in the finals at ACM Multimedia 2009 in Beijing [8], the system won the first prize.

5.2 Speaker Recognition

Structured approaches to data selection for speaker recognition: As noted in last year's annual report, to improve speaker recognition performance, it was often useful to incorporate variables computed from selected parts of the speech signal (e.g., word n-grams for common words or phone n-grams). We are now interested in developing a systematic approach to determining sound units that could be employed in this way. Toward this end, we have worked on performance measures that correlate well with speaker recognition accuracy (i.e., have a strong negative correlation with equal error rate) for different types of speech units. We have developed three main categories of measures for this purpose: one based on mutual information, one based on kurtosis, and one based on nasality features.

Currently, we are working toward discovering speech segments that perform well according to our measures. A dynamic programming approach has been implemented to compute the mutual information, kurtosis, and nasality measures on many frame sequences within phone n-grams to determine how each sequence performs with respect to the measures. We examined sequences within the same phone n-gram across different conversation sides so that the sequences are phonetically matched. Sequences that performed well with respect to the kurtosis and mutual information measures (i.e., more speaker discriminative as determined by those measures) are used as the new units. Note that the nasality measures were inconsistent for this task and were hence excluded. We performed keyword spotting of the new units (5 total) on new conversation sides, using the ANN/HMM connectionist approach for automatic speech recognition [31]

Having located the new units, we ran a GMM/UBM speaker recognition system on the new conversation sides using data where only the new units exist, and where existing monophone units exist. We determined that all new units performed within the top third amongst the existing monophone units. Using the same conversation sides that were used

to dynamically compute our measures across the phone n-grams in our speaker recognition experiment, we found that 4 of the 5 new units performed amongst the top 13% of the monophone units, suggesting that the new units may not have been located properly in the new conversation sides due to errors from keyword spotting. We are currently examining other new units using different phone n-grams across the conversation sides to see if we can improve upon our results.

Preparing for the NIST speaker recognition evaluation 2010: We have experimented with various state-of-the art speaker recognition techniques for the upcoming speaker recognition evaluation. Among those include joint-factor-analysis (JFA) [27], frame discounting for speaker model training, various JFA scoring approaches [20], and various score-normalization techniques. Our goal is to submit a combination of systems giving the best overall speaker recognition performance across different channels and other sources of speaker variability. Our best system so far is a simplified factor-analysis approach based on the ALIZE open-source toolkit [5]. This approach involves training gender-dependent UBMs, speaker models, and eigenchannel matrices, and log-likelihood scoring based off of channel integration. We are also working to integrate a support vector machine (SVM)-based system with our current best system. Finally, NIST is using a simple version of our system to determine difficult vs. easy speaker pairs for use in some future work.

Assessing discriminability of speaker pairs: Using features that can characterize speakers' voices, and thus provide a measure of speaker similarity, we aimed to predict which impostor speaker pairs would be most difficult for automatic speaker recognition systems to distinguish. Features tested include pitch, jitter, shimmer, formant frequencies, energy, long-term average spectrum energy, histograms of frequencies from roots of linear predictive coding (LPC) coefficients, and spectral slope. Per-speaker features were calculated using conversations from NIST's 2008 Followup Speaker Recognition Evaluation. Then, a measure of closeness of these features was calculated for each unique speaker pair. Measures utilized include absolute and percent differences, Euclidean distance, and correlation. Based on the values of these measures, the speaker pairs with the smallest or biggest differences were selected as potentially difficult or easy speaker pairs. System performance for these selected speaker pairs was then determined for a variety of system submissions to the NIST 2008 Speaker Recognition Evaluation. Ultimately, the largest performance difference between difficult and easy speaker pairs occurred when speaker pairs were selected by correlation coefficients of histograms of frequencies from LPC roots with magnitude greater than 0.78. Furthermore, relative to all speaker pairs, a larger performance difference existed for easy speakers than for difficult ones.

Building from this work, an even more successful measure for selecting difficult speaker pairs proved to be an estimate of the Kullback-Leibler (KL) divergence of Gaussian mixture models (GMMs) adapted to each speaker (using all conversations available). The KL divergence was also used to select difficult target speakers in the following way. For each target speaker, the KL divergence was estimated for each pair of GMMs trained on conversations of that speaker, and the average and standard deviation of these KL divergences were calculated. The best measure for selecting difficult (and easy) target speakers turned

out to be the ratio of the average to the standard deviation, where smaller values of this ratio indicated more difficult target speakers. Current work is underway to explore the possibilities of using such information about the speaker, i.e., whether it is likely to be difficult or easy for the system, in order to improve overall system performance.

5.3 Speech Recognition

In 2009 we continued to work with our multi-stream features for broadcast speech for the DARPA GALE project, first for Mandarin in collaboration with our partners at SRI, and later in the year on Arabic in collaboration with new partners at IBM. In both cases, we were able to provide substantial improvements in the overall evaluation systems. However, most of the ASR effort for the year was on exploring combination schemes for many Gabor-based streams, development tasks for parallel implementations on GPUs (in collaboration with partners at the campus ParLab, particularly Jike Chong), and on an open source ASR system based around IDIAP's Juicer.

Parallelization and many-stream ASR: Speed has always been the Achilles heel of artificial neural networks, particularly for training large networks. In the past (1989-1996) we developed a number of hardware/software accelerators (in particular, the RAP and the Spert). In recent years commercial hardware has reached a level of performance through Moore's Law technology changes that has made many of our ASR research efforts possible. However, given the saturation of clock speeds due to energy dissipation considerations, the most recent performance gains have been necessarily due to parallelizing computation on an increasing number of CPU cores. In order to learn how to further increase throughput (and thus facilitate research on more effective algorithms), we teamed with our colleagues at the Berkeley ParLab. The ParLab's motivation for working with us was to be part of an applications focus, so that their goal of easier development of efficient parallel programs would have a grounding in realistic use by so-called "productivity programmers", who would prefer not to be mired in the details of parallel programming.

As reported last year, one of our interests related to ParLab was to derive algorithms that could benefit from highly parallel hardware. Gabor-filtered streams that compute spectro-temporal receptive fields comprised a key part of one such algorithm. This "many-stream" approach, as we reported last year, reduced small vocabulary errors in a noisy condition by nearly a factor of two [46]. The most computationally intensive part of this algorithm is the forward computation (and, offline, the training) of large multi-layer perceptrons (MLPs). As noted above, we have developed hardware and software for this purpose for many years. However, we wanted to see how training and recall could be parallelized on current commercial parallel hardware, in particular, on a GPU; we successfully completed this software (for an NVIDIA processor) in 2009. One key part of this process, which we worked on extensively, was handling floating point overflows and underflows in the GPU computation, particularly in the calculation of exponentials and softmax quotients. These problems became more serious as we started to greatly increase the number of frames to "bunch" together into matrix-matrix computations, a step which was necessary to reduce communication costs when parallelizing to a large number of arithmetic elements. By re-

arranging the exponentials, it was possible to eliminate this problem without otherwise altering the results.

Open source speech recognition: In 2009, we began work on MySTT (My Speech-to-Text), a set of components that implement a full speech recognition system aimed at transcription of natural, human-to-human communication in real-time. MySTT uses the multimedia streaming framework "gstreamer," which allows efficient composition of processing components using a well-defined and stable interface, and a meta-data composition framework built on top of gstreamer known as MPF. All components of the system, including the initial models, are open source. Training of new models depends on HTK, a freely available but non-open toolkit. The front-end components of MySTT, including a speech/nonspeech segmenter, are complete, and we are in the process of porting Juicer (an open source speech decoder from our Swiss colleagues at IDIAP) to the framework.

5.4 Speech Understanding

Rich transcription of speech: In 2009, we focused our punctuation detection efforts on detection of questions in meetings and conversations. Identifying questions in human dialogs is an important first step in automatically processing and understanding natural speech. In the case of human-computer dialog systems, it can be critical for the conversational agent to know that a user asked it a question so that the dialog can be directed accordingly. In more passive systems, such as those utilized in multiparty meetings, question detection is useful for meeting indexing and summarization.

In our work [1], in addition to the features that were used in the previous work, such as the word n-grams, prosody, and part-of-speech (POS) tags, we investigated the use of syntactic parse tree features. An analysis of features used for classification revealed that lexico-syntactic features are most useful for this task, with turn- and pitch-related features providing complementary information in combination. The lexico-syntactic features seem to enable the classifier to correctly identify the cues that signal a question. This is particularly the case for word n-grams, which slightly outperform parse trees and significantly outperform POS tags. Furthermore, the added syntactic information in parse trees as compared to POS tags was shown to greatly improve performance.

Evaluation of automatic semantic role labeling on speech transcriptions: Semantic role labeling (SRL) is a natural language processing task consisting of the detection of the semantic arguments associated with the predicates or verbs of a sentence and their classification into their specific roles. On the speech processing side, SRL has become an important component of spoken language understanding systems [22]. For example, ATIS dialog systems obtained flight information by spotting departure and arrival information in user utterances, a shallow, task-dependent form of SRL. The MEDIA project has annotated tourist information dialogs with shallow semantics [3]. However, all of these previous studies only provide an indirect evaluation of dependency parsing and SRL on speech, and focus on evaluating the overall understanding task. Even though extrinsic evaluation of

SRL systems is useful, it does not allow researchers to assess directly the performance of their SRL components.

In our work [6], we extended the standard evaluation metrics for joint dependency parsing and SRL of text in order to be able to handle speech recognition output with word errors and sentence segmentation errors. We proposed metrics based on word alignments and bags of relations, and compare their results on the output of several SRL systems on broadcast news and conversations of the OntoNotes corpus. We evaluated and analyzed the relation between the performance of the subtasks that lead to SRL, including ASR, part-of-speech tagging, and sentence segmentation. The tools have been made available to the community.

Information distillation: In the first half of 2009, we continued our work on information distillation, in the framework of the DARPA GALE project. The 2008-9 distillation task was very focused and included extraction of answers to 5 W-questions (who, what, when, where, and why) that can be asked for sentences from the multilingual, audio and textual documents. The first part of our work mainly focused on the combination of our two approaches to this task, one of which is based on the use of syntactic parse trees and function tags, and the other of which is based on the use of SRL. We then trained statistical language models (LM) to predict whether the answers returned by semantic or by syntactic parsers are more likely [43]. We evaluated this approach using the OntoNotes dataset. Our experimental results indicated that the proposed LM-based combination strategy was able to improve the performance of the best individual system.

In our final evaluation submission, we also combined our system output with two other information distillers from New York University and Columbia University, and used a classification-based system selection at the overall sentence level [44]. Furthermore, we merged efforts of the three teams and performed a detailed analysis of the individual and combination systems [34].

We also studied the generic question answering task which aims to answer questions using spoken documents. We proposed an anchored speech recognition method which improved the recognition of the relevant sentences from the spoken documents by re-recognizing them using a language model tuned to the information in the question [45].

In preparation for the following years, we continued working on the template-based question answering distillation task. We investigated two perspectives that use shallow language processing for answering open-ended distillation queries, such as *List me facts about [event]* [39]. The first approach is a summarization-based approach that uses the unsupervised maximum marginal relevance (MMR) technique to successfully capture relevant but not redundant information. The second approach is based on supervised classification and trains support vector machines (SVMs) to discriminate relevant snippets from irrelevant snippets using a variety of features. Furthermore, we investigated the merit of using the ROUGE metric for its ability to evaluate redundancy alongside the conventionally used F-measure for evaluating distillation systems. Our experimental results with textual data indicate that SVM and MMR perform similarly in terms of ROUGE-2 scores while SVM is better than MMR in terms of F1 measure. Moreover, when speech recognizer output is used, SVM outperforms MMR in terms of both scores.

Summarization of multiple documents: In 2009, we participated in the NIST Text Analysis Conference (TAC) evaluation's query-focused multi-document summarization, update task for the second time. The update summarization task aims to generate short, fluent, and user-focused summaries of multiple news articles. For each multi-document set about a topic, participants are given a topic statement expressing the information need of a user, and two chronologically ordered batches of articles about the topic. The task is to generate a 100-word summary for each batch of articles. The summary of the second batch of articles is formed under the assumption that the user has already read the previous group of articles and wants to find out about new information related to the topic.

The improvements in our system over the previous year's submission include:

- Improved sentence boundary detection [17],
- Pruning of sentences that are unlikely to work well in a summary,
- Leveraging of sentence position to improve update summarization, and
- A high-precision variety of sentence compression that is based on semantic role labeling and dependency parsing [2] to improve readability rather than content, and incorporation of the compressed alternatives of each sentence in the integer linear programming approach we proposed last year [18].

Similar to the previous year, the ICSI system did very well on ROUGE scores and on several other manual evaluation categories, out of over 50 submissions.

Summarization of meetings: Meetings provide an efficient way of sharing knowledge among people with different areas of expertise. Every day, especially within organizations, people meet for various reasons, such as discussing issues, assigning tasks, and planning. These get-togethers can be recorded for documenting and archiving the progress of the group. Meeting summarization aims to generate a compact, summary version of meeting discussions. These summaries can be formed by extracting original speaker utterances (extractive summarization) or by formulating new sentences for the summary (abstractive summarization).

In 2009, we analyzed and proposed several new methods for meeting summarization. First, we proposed an optimal formulation for the widely used maximum marginal relevance (MMR) algorithm based on integer linear programming (ILP) [36]. We compared this method with our previous ILP-based approach that targets optimal selection of utterances covering as many unique important concepts as possible.

Second, we proposed to leverage sentence importance weights in this concept-based model [41]. Three ways are introduced to combine the sentence weights, within the concept-based optimization framework: selecting sentences for concept extraction, pruning unlikely candidate summary sentences, and using joint optimization of sentence and concept weights. Our experimental results on the ICSI meeting corpus show that our proposed methods can signicantly improve the performance for both human transcripts and ASR output compared to the baseline of the concept-based approach, and this unsupervised approach achieves results comparable with those from supervised learning approaches presented in previous work.

Third, we proposed an unsupervised, graph-based approach for extractive summarization of meetings. Graph-based methods such as TextRank have been used for sentence extraction from news articles. These methods model text as a graph with sentences as nodes and edges based on word overlap. A sentence node is then ranked according to its similarity with other nodes. The spontaneous speech in meetings leads to incomplete, ill-formed sentences with high redundancy and calls for additional measures to extract relevant sentences. We proposed an extension of the TextRank algorithm that clusters the meeting utterances that are about the same or similar topic and uses these clusters to construct the graph. Our clustering method starts with assigning each sentence a separate cluster, and then iteratively merges adjacent clusters that have the maximum similarity. We evaluated this method on the AMI meeting corpus and showed a significant improvement over TextRank and other baseline methods.

Finally, we investigated methods aiming to effectively use acoustic/prosodic features for extractive meeting summarization, and how to integrate prosodic features with lexical and structural information for further improvement. Speech contains additional information that text doesn't, and that can be valuable for automatic speech summarization. Prosodic features aim to capture how the utterances are said in contrast to content information captured by the lexical features. To properly represent prosodic features, we proposed new normalization methods based on speaker, topic, or local context information. Our experimental results showed that using only the prosodic features we achieve better performance than using the non-prosodic information on both the human transcripts and recognition output. In addition, a decision-level combination of the prosodic and non-prosodic features yielded further gain, outperforming the individual models.

5.5 Computational Methods for Conversation Analysis

In 2009, ICSI (together with SRI International) was awarded an IARPA grant to work on automatic processing methods to detect socio-cultural content in language. The aim in this project 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 in our project is multi-party meetings and broadcast conversations in multiple languages. We aim to discover (1) what types of evidence in spoken language support inferences about social constructs, a term we use to encompass social goals, relations, identities, roles, and other social phenomena, and (2) how this evidence can be modeled automatically.

The project will first produce human annotations of linguistic phenomena and social constructs, which will then be used for training machine learning methods to be used in automatic extraction. Inter-annotator agreement of human annotators will be measured to estimate how well these categories are defined, and examples on which human annotators disagreed will be used to refine the definitions of categories. We completed annotating the LDC-distributed English broadcast conversations with person addressing and mentions, disfluencies and speaker roles. Furthermore, we created a first version of automatic annotation for speaker roles that uses lexical features and dialog act tags.

The output of these systems will be used by information analysts. Hence just outputting

categories, such as "Speaker 1 is the host," is not enough, and an automatic explanation of why the specific category was chosen is necessary. For example, in addition to assigning the role of "talk show host" to a broadcast conversation participant, we need to generate an explanation, such as "This speaker was assigned the host role, since he spoke 50% of the time, and asked 80% of all the questions," and a support statement, such as "The speakers who speak more than 40% of a broadcast show are hosts in 90% of the shows."

5.6 Human Robot Interaction

Humans benefit from both audio and visual modalities when interacting naturally with each other and, in addition to following what is being said, they take advantage of how objects are referred to with the help of visual gestures, and distinguish commands from comments and questions with the help of distinct prosody. It is straightforward for humans to merge these multi-modal information sources ranging from words with the prosody of speech to visual cues, such as head and hand gestures, to surrounding objects and people. Recent advances in vision and speech processing have made it possible to track and identify objects, recognize gestures of conversation participants, and human/computer spoken dialog applications are becoming more popular, especially in the customer care domain. However, despite these advances, these tasks are still error-prone and fusion of information from the vision and speech modalities along with contextual domain semantics for better spoken language understanding remains a significant challenge.

Currently, the ICSI Vision Group is funded to research perceptually situated, natural conversation models that robots can use to interact multi-modally with human operators or participants. We have been colaborating with the ICSI vision team, to supplement this current project with features from speech prosody as well as features from the speech of the other conversation participants, and combine information from vision and speech to improve the accuracy of the speech understanding module for human/robot interaction. More specifically, our goal is to improve the performance of the automatic speech recognition (ASR) and addressee detection models for speech understanding in natural human/robot interaction. For example, the words that were previously unseen by the ASR language models can be recognized if an utterance is accompanied with visual cues, similar to how humans acquire language. When the human asks the robot to pick up the *Elmo toy*, where the word *Elmo* is an out-of-vocabulary (OOV) word, the system can match the concept with an object in scene. This idea can be generalized to perform n-best or lattice re-scoring to get better performance even when such words are not OOV. For instance, the acoustic information is too ambiguous to make the distinction between the words "pen" and "pan," but their visual characteristics are distinct enough to allow an image-based classifier to correct the ASR errors.

On the other hand, regarding addressee detection, one can combine the gaze and contextual information with prosodic cues, assuming people talk to machines in a different way than they talk to each other. This is analogous to the natural fact that prosody of adult speech changes when they are addressing children instead of other adults [26]. Furthermore, the users can talk to the robot in a more commanding mode, and with a different prosody than to other conversation participants. Prosodic features have been shown to be

useful for dialog act tagging in previous studies [38], and hence are expected to be useful for addressee detection in human/robot interaction.

We designed a set of scenarios where one can study the impact of fusion of speech and vision information on human robot interaction. Furthermore, we investigated various set-ups for data collection, with the aim of making the collected set usable for other research.

5.7 Multimodal Analysis Framework

Under the sponsorship of a Silicon-Valley-based start-up company, Appscio, the speech group continued their involvement in the development of an open-source software framework based on GStreamer. The goal of the software platform is to ease the creation of multimedia content analysis applications that consist of components provided from multiple sources and different programming languages. The framework aims to provide a unified approach that standardizes the entire process of development, deployment, and integration of components. We contributed to the design of the framework as well as components as part of the MySTT project (see Section 5.3).

5.8 Handling Complex Sound Environments

While listeners are comfortable with sound environments consisting of multiple, overlapping sounds, the same scenario presents a serious challenge to machine recognition systems. Inspired by human listeners, we have been investigating the separation of sources and recognition of speech in noisy, reverberant environments with only one or two "ears" (microphones) available – too few to apply classical beamforming techniques. We have continued our previous work on binaural separation by investigating the theoretical and empirical bounds on performance of systems that separate only based on the statistics of between-ear differences, and showed that our MESSL system (Model-based Expectation Maximization Source Separation and Localization) comes close to this oracle bound [30]. Listeners, of course, can use more information to separate sources, and our Eigenvoice approach is meant to approximate the kind of "general knowledge" about the characteristics of voices that can be very important in helping listeners to infer fragments of speech that are otherwise obscured by noise. This year, ICSI External Fellow Dan Ellis and his Columbia colleagues developed more efficient algorithms for fitting eigenvoice models to speech mixtures based on a variational approximation to the complex likelihood function [40].

Complementary to the problems of recognizing speech in noisy contexts is the task of classifying and describing the noisy environment itself. Building on our earlier work of tagging consumer video soundtracks with labels such as "beach," "crowd," "party," etc., we have looked at refining the temporal scope of such labels to subdivide the soundtrack into segments which may individually reflect different specific characteristics. This should give a better description of such videos, and can also improve labeling at the video layer by reducing the danger of long stretches of uninteresting sound "washing out" brief but highly informative sounds [28].

A particularly significant attribute of a sound event may be that it recurs frequently, in more or less the same form, throughout a soundtrack – for instance, the sound of a

particular machine, or doorway, or other location- or activity-specific acoustic event. We have developed an efficient technique for identifying such commonly-repeating events, that can succeed in identifying near-repeats even in the context of intense background noise. This approach, inspired by music fingerprinting but much more robust to small variations in the sound, describes sound events as clusters of Match Pursuit atoms (greedily extracted) which are stored in a locality-senstive hash table. Later sounds with similar characteristics will end up in the same hash bin, and are thus immediately matched to relevant earlier events. This showed that, by focusing on the most energetic atoms in any sound, we can extract and recognize certain kinds of transient events (such as horse steps) with good immunity to background noise [4]. Such an approach is a vital foundation for improving the performance of automatic taggers for video soundtracks.

5.9 Music Processing: Cover Song Detection

UC Berkeley graduate student Suman Ravuri created a system to arbitrarily identify whether or not a pair of songs is a reference/cover with around 83% accuracy. The system used multiple features and a classifier (either SVM or MLP). A paper on the approach will appear in ICASSP 2010 [35]. We also submitted the algorithm to a music information retrieval contest (MIREX 2009) in which participants were given a reference song and 1000 test songs and asked to find the ten cover songs in the set. Even though Ravuri's system was tuned for different rules, the entry won second place.

5.10 Machine Translation

Statistical machine translation systems produce new translations by recombining fragments of old ones. This process has two sub-components: learning reusable fragments from training data and searching for translations of new input sentences that use those fragments. Professor Klein and his students work on both components.

Learning models of syntactic translation involves analyzing example translations into their component parts, then inducing a correspondence between those parts. Unfortunately, syntactic analysis and translation alignment are generally done by entirely separate processes, leading to widespread mismatch errors. UC Berkeley Professor Klein and his students developed an efficient approach that parses and aligns simultaneously in a joint model. Their approach substantially improves parse accuracy, alignment accuracy, and end translation quality.

Analysis of new sentences under a translation model requires parsing with a very large grammar that encodes many substructures of the training data. These grammars are generated in such a way that they are actually many times larger than the original training data itself, greatly limiting the amount of data that can be efficiently used. Klein's group has demonstrated a new syntactic analysis system that addresses this issue with an implicit representation that encodes all training substructures without any data expansion.

References

[1] K. Boakye, B. Favre, and D. Hakkani-Tür (2009). "Any Questions? Automatic Question Detection in Meetings." To appear in the proceedings of the IEEE Workshop in Automatic Speech Recognition and Understanding (ASRU), Merano, Italy, December 2009.

[2] B. Bohnet (2009). "Efficient Parsing of Syntactic and Semantic Dependency Structures." Presented at the 13th Conference on Computational Natural Language Learning (CoNLL-2009), Boulder, Colorado, June 2009.

[3] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostef (2005). "Semantic Annotation of the French Media Dialog Corpus." Proceedings of the European Conference on Speech Communication and Technology,Lisbon, Portugal, pp. 3457-3460, September 2005.

[4] C. Cotton and D. Ellis (2009) Finding Similar Acoustic Events using Matching Pursuit and Locality-Sensitive Hashing Proc. WASPAA-09, Mohonk NY, October 2009, pp. 125-128. http://www.ee.columbia.edu/ dpwe/pubs/CottonE09-mp-lsh.pdf

[5] B. Fauve, D. Matrouf, N. Scheffer, J.F. Bonastre, and J. Mason (2007). "State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software." IEEE Transactions on Audio, Speech, and Language Processing, September 2007.

[6] 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, April 2010.

[7] G. Friedland (2009). "Analytics for Experts." Featured article, *ACM SIGMM Records*, Vol.1, Issue 1, April 2009.

[8] G. Friedland, L. Gottlieb, and A. Janin (2009). "Joke-o-mat: Browsing Sitcoms Punchline by Punchline." Proceedings of ACM International Conference on Multimedia (ACM Multimedia 2009), Beijing, China, pp. 1115-1116, October 2009.

[9] G. Friedland, L. Gottlieb, and A. Janin (2009). "Using Artistic Markers and Speaker Identification for Narrative-Theme Navigation of Seinfeld Episodes." Workshop on Content-Based Audio/Video Analysis for Novel TV Services, Proceedings of the 11th IEEE International Symposium on Multimedia (ISM2009), San Diego, California, pp. 511-516, December 2009.

[10] G. Friedland, H. Hung, and C. Yeo (2009). "Multi-modal Speaker Diarization of Real-World Meetings Using Compressed-Domain Video Features." Proceedings ofthe IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, pp. 4069-4072, April 2009.

[11] G. Friedland and D. van Leeuwen (2010). "Speaker Diarization and Identification." To appear in /it Semantic Computing, P. Sheu et al., 2010.

[12] G. Friedland, O. Vinyals, Y. Huang, and C. Müller (2009). "Fusion of Short-Term and Long-Term Features for Improved Speaker Diarization." Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, pp. 4077-4080, April 2009.

[13] G. Friedland, O. Vinyals, Y. Huang, and C. Müller (2009). "Prosodic and Other Long-Term Features for Speaker Diarization." *IEEE Transactions on Audio, Speech, and Language Processing*, Vol 17, No 5, pp 985–993, July 2009.

[14] G. Friedland, C. Yeo. and H. Hung (2009). "Visual Speaker Localization Aided by Acoustic Models." Proceedings of ACM International Conference on Multimedia (ACM Multimedia 2009), Beijing, China, pp. 195-202, October 2009.

[15] N. Garg, B. Favre, K. Riedhammer, and D. Hakkani-Tür (2009). "ClusterRank: A Graph Based Method for Meeting Summarization." Proceedings of the 10th International Conference of the International Speech Communication Association (Interspeech 2009), Brighton, United Kingdom, pp. 1499-1502, September 2009.

[16] D. Gelbart, N. Morgan, and A. Tsymbal (2009). "Hill-Climbing Feature Selection for Multi-Stream ASR." Proceedings of the 10th International Conference of the International Speech Communication Association (Interspeech 2009), Brighton, United Kingdom, pp. 2967-2970, September 2009.

[17] D. Gillick (2009). "Sentence Boundary Detection and the Problem with the U.S." Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL HLT 2009): Short Papers, Boulder, Colorado, pp. 241-244, June 2009.

[18] D. Gillick and B. Favre (2009). "A Scalable Global Model for Summarization." Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing at the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL HLT 2009), Boulder, Colorado, pp. 10-18, June 2009.

[19] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tür (2009). "A Global Optimization Framework for Meeting Summarization." Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), Taipei, Taiwan, pp. 4769-4772, April 2009.

[20] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny (2009). "Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis." Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), Taipei, Taiwan, pp. 4057-4060, April 2009.

[21] L. Gottlieb and G. Friedland (2009). "On the Use of Artificial Conversation Data for Speaker Recognition in Cars." Proceedings of the Third IEEE International Conference on Semantic Computing (ICSC-2009), Berkeley, California, pp. 124-128, September 2009.

[22] D. Hakkani-Tür, G. Tur, and A. Chotimongkol (2005). "Using Syntactic and Semantic Graphs for Call Classification." Proceedings of the Workshop on Feature Engineering for Machine Learning in Natural Language Processing at ACL'05, Ann Arbor, Michigan, June 2005.

[23] D. Imseng and G. Friedland (2009). "Robust Speaker Diarization for Short Speech Recordings." Proceedings of the 11th Biannual IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2009), Merano, Italy, pp. 432-437, December 2009.

[24] D. Imseng and G. Friedland (2010). "Tuning-Robust Initialization Methods for Speaker Diarization." To appear in *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[25] D. Imseng and G. Friedland (2010). "An Adaptive Initialization Method for Speaker Diarization based on Prosodic Features." To appear in the proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Dallas, Texas, April 2010.

[26] V. Kempe (2009). "Child-Directed Speech Prosody in Adolescents: Relationship to 2D:4D, Empathy, and Attitudes Towards Children." *Personality and Individual Differences*, Vol. 47, Issue 6, pp. 610-615, October 2009.

[27] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel (2008). "A Study of Inter-Speaker Variability in Speaker Verification." /it IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, No. 5, pp. 980-988, July 2008.

[28] K. Lee, D. Ellis, and A. Loui (2010) Detecting Local Semantic Concepts in Environmental Sounds using Markov Model based Clustering Proc. IEEE ICASSP, Dallas, to appear, March 2010. http://www.ee.columbia.edu/ dpwe/pubs/LeeEL10-localconcepts.pdf

[29] M. Levit, D. Hakkani-Tür, G. Tur, and D. Gillick (2009). "IXIR: A Statistical Information Distillation System." *Journal of Computer Speech and Language*, Vol. 23, Issue 4, pp. 527-542, June 2009.

[30] M. Mandel and D. Ellis (2009) The Ideal Interaural Parameter Mask: A Bound on Binaural Separation Systems Proc. WASPAA-09, Mohonk NY, October 2009, pp. 85-88. http://www.ee.columbia.edu/ dpwe/pubs/MandE09-IIPM.pdf

[31] N. Morgan and H. Bourlard (1995). "Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach." *IEEE Signal Processing Magazine*, Vol. 12, Issue 3, pp. 25-42, May 1995.

[32] C. Müller and G. Friedland (2009). "Multimodal interfaces for automotive applications (MIAA)." Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI 2009), Sanibel, Florida, pp. 493-494, February 2009.

[33] C. Oei, G. Friedland, and A. Janin (2009). "Parallel Training of a Multi-Layer Perceptron on a GPU." ICSI Technical Report 09-008, Berkeley, California, October 2009.

[34] K. Parton, K. R. McKeown, R. Coyne, M. T. Diab, R. Grishman, D. Hakkani-Tür, M. Harper, H. Ji, W. Y. Ma, A. Meyers, S. Stolbach, A. Sun, G. Tur, W. Xu, and S. Yaman (2009). "Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task." To appear in the proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the Fourth International Joint Conference on Natural Lanaguage Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009), Singapore, August 2009.

[35] S. Ravuri and D. Ellis (2010). "Cover Song Detection: From High Scores to General Classification." To appear in the proceedings of IEEE ICASSP, Dallas, Texas, April 2010.

[36] K. Riedhammer, B. Favre, and D. Hakkani-Tür (2010). "Global Unsupervised Methods for Keyphrase-Based Meeting Summarization." In submission, 2010.

[37] A. Stolcke, G. Friedland, D. Imseng (2010). "Leveraging Speaker Diarization for Meeting Recognition from Distant Microphones." To appear in the proceedings of IEEE ICASSP, Dallas, Texas, April 2010.

[38] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. van Ess-Dykema,and M. Meteer (2000). "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech." *Computational Linguistics*, Vol. 26, Issue 3, pp. 339-373, September 2000.

[39] B. Toth, D. Hakkani-Tür, and S. Yaman (2010). "Summarization and Learning-Based Approaches to Information Distillation." To appear in the proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Dallas, Texas, April 2010.

[40] R. Weiss and D. Ellis (2009) A Variational EM Algorithm for Learning Eigenvoice Parameters in Mixed Signals Proc. ICASSP-09, pp. 113-116, Taiwan, April 2009. http://www.ee.columbia.edu/ dpwe/pubs/WeissE09-ev-vem.pdf

[41] S. Xie, B. Favre, D. Hakkani-Tür, and Y. Liu (2009). "Leveraging Sentence Weights in Concept-Based Optimization Framework for Extractive Meeting Summarization." Proceedings of the 10th International Conference of the International Speech Communication Association (Interspeech 2009), Brighton, United Kingdom, pp. 1503-1506, September 2009.

[42] S. Xie, D. Hakkani-Tür, B. Favre, and Y. Liu (2009). "Integrating Prosodic Features in Extractive Meeting Summarization." To appear in the proceedings of the 11th Biannual IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2009), Merano, Italy, December 2009.

[43] S. Yaman, D. Hakkani-Tür, and G. Tur (2009). "Combining Semantic and Syntactic Information Sources for 5-W Question Answering." Proceedings of the 10th International Conference of the International Speech Communication Association (Interspeech 2009), Brighton, United Kingdom, pp. 2707-2710, September 2009.

[44] S. Yaman, D. Hakkani-Tür, G. Tur, R. Grishman, M. Harper, K. R. McKeown, A. Meyers, and K. Sharma (2009). "Classification-Based Strategies for Combining Multiple 5-W Question Answering Systems." Proceedings of the 10th International Conference of the International Speech Communication Association (Interspeech 2009), Brighton, United Kingdom, pp. 2703-2706, September 2009.

[45] S. Yaman, G. Tur, D. Vergyri, D. Hakkani-Tür, M. Harper, and W. Wang (2009). "Anchored Speech Recognition for Question Answering." Proceedings of North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL HLT 2009): Short Papers, Boulder, Colorado, pp. 265-268, June 2009.

[46] S. Y. Zhao, S. Ravuri, and N. Morgan (2009). "Multi-Stream to Many-Stream: Using Spectro-Temporal Features for ASR." Proceedings of the 10th International Conference of the International Speech Communication Association (Interspeech 2009), Brighton, United Kingdom, pp. 2951-2954, September 2009.

6 Vision

In 2009, the ICSI Vision Group grew in size from approximately four members to over ten, maintained sponsorship from government and corporate sources, and engaged in a series of productive projects with significant publications. Our research focii included recognition of objects, events, and gestures, and their application to situated interaction with robots and/or perceptually guided search. We remained interested in core computer vision challenges, including learning good local feature representations, modeling human motion, and learning from web imagery, and also addressed core machine learning challenges, including dimensionality reduction, metric learning, and active learning.

In the following four subsections we review vision group personnel updates as well the sponsors, projects, and publications in 2009.

6.1 Personnel

In 2009 a postdoctoral research scientist in the group, Raquel Urtasun, moved on to a faculty position at the Toyota Technological Institute in Chicago (TTI-C), where she continues to research applied machine learning methods and collaborates actively with ICSI. Several new members joined the lab, including postdoctoral researchers Brian Kulis from University of Texas at Austin, Carl Ek from Oxford Brookes University, Dr. Mario Fritz from the Technical University at Darmstadt. Mario Fritz and Kate Saenko graduated with their PhDs from MIT and also joined as postdoctoral researchers, with Saenko working from Massachusetts. Three PhD students continued to work with the group, Ashley Eden, Alex Shyr, and Dave Golland, the latter two as co-advisees of UC Berkeley professors Michael Jordan and Dan Klein, respectively. Three new PhD students joined the group in 2009: Trevor Owens, Yangqing Jia, and Sergey Karayev. The group also was fortunate to have a visitor on a postdoctoral fellowship from the German DAAD: Nicholas Cebron.

6.2 Sponsors

The Vision Group's research was supported by an array of sponsors, governmental and corporate.

Government sponsors included NSF, which supported our work on grounded human-robot language models, including research on how visual recognition of gestures and objects in the environment can be related to language use while interacting with robots. NSF also began new support of our research on face and event recognition in online social media. DARPA supported our research on gesture recognition and visual activity recognition, and other DOD sponsors supported our research on multimodal location recognition and dimensionality reduction for event and gesture modeling. The US DHHS supported our research on facial similarity modeling, with application to a system for helping identify lost children after natural disasters being developed by our colleagues at Children's Hospital Boston. (The project was inspired by the aftermath of Hurricane Katrina, and the recent unfortunate events in Haiti provided additional motivation.)

Toyota provided support for our work on robotic visual category recognition, specifically focusing on exploitation of 3-D sensing for recognition using training data from 2-D sources.

We also initiated a industrial visitor program with Panasonic's Singapore labs, and hosted a visitor interested in consumer photo applications who gained expertise learning about our methods for visual category recognition.

6.3 Projects

The paragraphs below provide a synopsis of ongoing projects in the ICSI Vision Group. For more information on each, please visit the projects tab of the Vision Group page at http://www.icsi.berkeley.edu/projects/vision_p.html

Bayesian localized multiple kernel learning: Multiple kernel learning approaches form a set of techniques for performing classification that can easily combine information from multiple data sources, e.g., by adding or multiplying kernels. Most methods, however, are limited by their assumption of a per-view kernel weighting. For many problems, the set of features important for discriminating between examples can vary locally. As a consequence these global techniques suffer in the presence of complex noise processes, such as heteroscedastic noise, or when the discriminative properties of each feature type varies across the input space. We proposed a localized multiple kernel learning approach with Gaussian Processes that learns a local weighting over each view and can obtain accurate classification performance and deal with insufficient views corrupted by complex noise, e.g., per-sample occlusion. We demonstrate our approach on the tasks of audio-visual gesture recognition and object category classification.

Facial image indexing interfaces: During a disaster a large number of children may become separated from their families. Many of these children, especially the younger ones, may be unable or unwilling to identify themselves, making the task of reuniting them with their families especially difficult. Without a system in place for hospitals to document their unidentified children and to help parents search for them, families could be separated for months. After Hurricane Katrina it was six months until the last child was reunited with her family. We are working on a system in which each hospital takes digital photos of the childrens' faces, and the system is able to automatically extract features useful for identification. We are also hoping to extend the system to automatically refine image searches based on the identification of similar looking faces. Along those lines, we are working on determining a metric for feature importance in facial similarity.

Grounded semantics: This project explores how to define the meaning of prepositions using visual data. One potential application is to be able to command a robot to arrange objects in a room. For example, in order for a robot to be able to follow the command "Put the cup there, on the front of the table," the robot must identify the target location of the cup. The robot can only identify this location if it understands the meanings of each of the components. The project specifically focuses on defining the meanings of prepositions because they are both perceptible in images and can be composed together to form higher level meanings. To learn more about this project, contact David Golland, http://www.eecs.berkeley.edu/ dsg/.

Hashing algorithms for scalable image search: A common problem in large-scale data is that of quickly extracting nearest neighbors to a query from a large database. In computer vision, for example, this problem arises in content-based image retrieval, 3-D image reconstructions, human body pose estimation, object recognition problems, and other problems. This project focuses on developing algorithms for quickly and accurately performing large-scale image searches using hashing techniques. Contributions include incorporating hashing methods for learned metrics as well as performing locality-sensitive hashing over arbitrary kernel functions, two prominent scenarios arising in modern computer vision applications. Recent work has aimed to learn appropriate hash functions for a given image search task in order to minimize the memory overhead required for accurate searches. We have applied our algorithms to several large-scale data sets including the 80 million images of the Tiny Image data set and other large content-based image retrieval data sets [6].

Nonrigid object recognition and tracking: In our everyday life, we manipulate many nonrigid objects, such as clothes. In the context of personal robotics, it would therefore be important to correctly recognize and track these objects for a robot to interact with them. While tracking and recognizing rigid objects has received a lot of attention in the computer vision community, similar tasks for deformable ones remain largely unstudied. The main challenges that need to be addressed arise from the much larger appearance variability of such objects. Furthermore, the wide range of shapes that a piece of clothing can undergo makes 3-D reconstruction and tracking very challenging. In this project, we intend to study machine learning and computer vision methods to solve the following problems:

1. Texture-based classification of the different parts of a single objects, e.g., boundaries vs. interior parts.
2. Instance-level recognition of particular pieces of cloth.
3. Category-level / material recognition, e.g., jeans vs. t-shirts.
4. 3-D shape estimation and tracking.

Most of the above problems can be tackled in a multi-modal context, where different types of input, such as video or laser, are available. Another subject of interest is the study of a principled way of combining these inputs, in particular when they are asynchronous.

Transparent object recognition: Despite the omni-presence of transparent objects in our daily environment, little research has been conducted on how to recognize and detect such objects. The difficulties of this task lie in the complex interactions between scene geometry and illuminants that lead to changing refractory patterns. Realizing that a complete physical model of these phenomena is out of reach at the moment, we seek different machine learning solutions to approach this problem. In particular we have been investigating a latent local additive feature model [3]. In stark contrast to the previous approach, this method seeks to separate different contributions to the overall gradient statistic in an unsupervised decomposition approach (Figure 2).

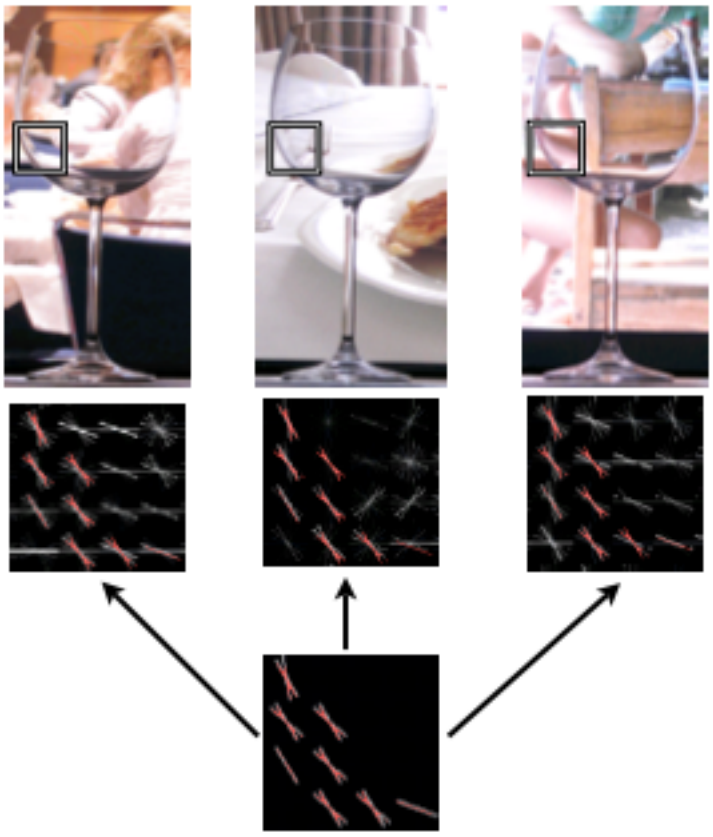

Figure 2: Example of latent local feature decomposition

Visual sense disambiguation using multiple modalities: Traditionally, object recognition requires manually labeled images of objects for training. However, there often exist additional sources of information that can be used as weak labels, reducing the need for human supervision. In this project we use different modalities and information sources to help learn visual models of object categories. The first type of information we use is the speech uttered by a user referring to an object. Such spoken utterances can occur in interaction with an assistant robot, voice-tagging a photo, etc. We propose a method that uses both the image of the object and the speech segment referring to the object to recognize the underlying category label. In preliminary experiments, we have shown that even noisy speech input helps visual recognition, and vice versa. We also explore two sources of information in the text modality: the words surrounding images on the Web, and dictionary entries for words that refer to objects. Words that co-occur with images on the Web have been used as weak object labels, but this tends to produce noisy datasets with many unrelated images. We use text and dictionary information to learn a refined model of what sense an image found on the Web is likely to belong to. We apply this model to a dataset of images of polysemous words collected via image search and show that it improves both retrieval of specific senses and the resulting object classifiers [8].

Probabilistic models for multi-view learning and distributed feature selection: Many problems in machine learning contain datasets that comprise multiple independent feature sets or views, e.g., audio and video, text and images, and multi-sensor data. In this setting, each view provides a potentially redundant sample of the class or event of interest. Techniques in multi-view learning exploit this property to learn under weak supervision by maximizing the agreement of a set of classifiers defined in each view over the training data. The ability to perform reliable inference and learning in the presence of multi-view data is a challenging problem that is complicated by many factors including view insufficiency, i.e., learning from real-world noisy observations, and coping with the potentially large amounts of information that arises when incorporating possibly many information sources for classification. In this work we propose probabilistic models built upon multi-view Gaussian Processes (GPs) for learning from noisy real-word multi-view data and for performing distributed feature selection in bandwidth constrained environments such as those typically encountered in multi-source sensor networks [1]. Initial experiments on audio-visual gesture and multi-view image datasets demonstrate that our probabilistic multi-view learning approach is able to learn under significant amounts of complex view corruption, e.g., per sample occlusions. Our work on GP-based multi-view feature selection has shown promising results for achieving compact feature descriptions from multiple sensors while preserving classification performance on a multi-view object categorization task.

In 2009 we also reported results from prior efforts on fusing speech and image cues for multimodal turntaking disambiguation [2], learning the dimensionality of latent spaces for computer vision tasks [4], efficient visual search for object recognition [9], and multi-task learning using block-normalized regression [7].

References

[1] M. Christoudias, R. Urtasun, A. Kapoor, and T. Darrell (2009). "Co-training with Noisy Perceptual Observations." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida, pp. 2844-2851, June 2009. http://people.csail.mit.edu/rurtasun/publications/christoudias_et_al_cvpr09.pdf.

[2] M. Frampton, R. Fernandez, P. Ehlen, M. Ehlen, M. Christoudias, T. Darrell, and S. Peters (2009). "Who is "You"? Combining Linguistic and Gaze Features to Resolve Second-Person References in Dialogue." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), Athens, Greece, pp. 273-281, March 2009. http://www.aclweb.org/anthology/E/E09/E09-1032.pdf.

[3] M. Fritz, M. Black, G. Bradski, and T. Darrell (2009). "An Additive Latent Feature Model for Transparent Object Recognition." Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009), Vancouver, Canada, pp. 558-566, December 2009. http://books.nips.cc/papers/files/nips22/NIPS2009_0397.pdf.

[4] A. Geiger, R. Urtasun, and T. Darrell (2009). "Rank Priors for Continuous Non-Linear Dimensionality Reduction." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida, pp. 880-887, June 2009. http://dspace.mit.edu/bitstream/handle/1721.1/42840/MIT-CSAIL-TR-2008-056.pdf?sequence=1.

[5] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell (2009). "Gaussian Processes for Object Categorization." *International Journal of Computer Vision*. Published online July 2009. http://www.springerlink.com/content/u2163rw858331135/.

[6] B. Kulis and T. Darrell (2009). "Learning to Hash with Binary Reconstructive Embeddings." Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009), Vancouver, Canada, pp. 1042-1050, December 2009. http://www.eecs.berkeley.edu/ kulis/pubs/hashing_bre_tr.pdf.

[7] A. Quattoni, X. Carreras, M. Collins, and T. Darrell (2009). "An Efficient Projection for L-1/L-Infinity Regularization." Proceedings of International Conference on Machine Learning (ICML 2009),Montreal, Quebec, pp. 857-864, June 2009. http://people.csail.mit.edu/mcollins/papers/icml09.pdf.

[8] K. Saenko and T. Darrell (2009). "Filtering Abstract Senses From Image Search Results." Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009), Vancouver, Canada, pp. 1589-1597, December 2009. http://people.csail.mit.edu/saenko/saenko_nips_2009.pdf.

[9] T. Yeh and T. Darrell (2009). "Fast Concurrent Object Localization and Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida, pp. 280-287, June 2009. http://people.csail.mit.edu/tomyeh/homepage/pubs/CVPR2009.pdf.

7 Architecture

The Architecture Group has as its major focus the realization of efficient parallel programmable architectures exploiting advances in circuit and device technologies. The group is slowly growing while leveraging extensive connections to other research groups in Berkeley and beyond.

The primary activities in the Architecture Group at ICSI continue to be in silicon photonics and the vector-thread architecture.

7.1 Monolithic Silicon Photonics

In a collaboration with the MIT Center for Integrated Photonic Systems, ICSI architecture 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, we explored the use of photonics between processors and memory controllers. With a newly awarded DARPA grant, we have been extending this work to look at how to take photonics all the way into the memory chips themselves.

The MIT-developed monolithic silicon photonics technology is well-suited for integration with standard bulk CMOS logic processes, which reduces costs and improves opto-electrical coupling compared to previous approaches. The photonics technology supports dense wavelength-division multiplexing with dozens of wavelengths per waveguide. The MIT group has fabricated several test chips and measured the device properties. Although the photonics technology is still relatively immature, initial measured results from test chips are promising despite being fabricated in a standard process with no modifications for photonic technology. In recent work, our MIT collaborators have begun exploring the monolithic integration of photonics in a standard CMOS DRAM process, with several test chips planned.

In earlier work, we explored how to use photonics to interconnect many processors to memory controllers using an opto-electrical global crossbar [7, 2]. Our results showed that aggregate throughput can be improved by ≈8–10 × compared to an optimized purely electrical network.

More recently, we have explored other applications of photonic signaling, 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 [5].

Photonics shows more promise for chip-chip interconnects, and in one line of work we have shown that the high bandwidth density of photonics enables a more efficient and scalable design style for multi-socket shared memory systems, allowing a large number of small chips to be connected with little penalty compared to using fewer, larger chips [4, 3]. Because smaller die are superlinearly cheaper to produce than larger die, simplify

system cooling, and can more easily satisfy multiple market segments with a single design, photonics should significantly reduce costs as well as improve performance and scalability.

Our newest line of research is to extend monolithic silicon photonics into memory chips themselves, providing a seamless photonic interconnect from processor cores to memory banks, with both possibly spread across scores of chips in a node. We received a new DARPA award beginning July 2009 to fund this research.

7.2 Maven (Malleable Array of Vector-thread ENgines)

In earlier work at MIT, Asanović's team developed the Scale vector-thread architecture and processor prototype [6], 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 made considerable progress on the microarchitectural design of Maven and have also completed VLSI layouts of not only the Maven prototype but also several competing design alternatives, such as traditional vector machines and GPU-style "SIMT" processors. By pushing multiple competing designs to layout we can for the first time provide quantitative evaluations of area, clock cycle time, and power. We have also developed a complete explicitly-data-parallel programming environment for Maven, based on the GNU GCC compiler, which allows users to program Maven using C++. We have modified the GCC compiler to optimize code for the vector-thread units. Our initial results are promising and show that vector-threading can provide the same efficiency as traditional vector machines on regular data-parallel code, while supporting a wider range of irregular data-parallel codes much more efficiently than GPU-style architectures. Christopher Batten recently completed his PhD in this area [1] and has taken a position as an assistant professor in the Electrical and Computer Engineering Department at Cornell University.

Over the course of the next year, we expect to continue refining the Maven architecture and to fabricate prototype test chips to help validate our simulation results.

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 is also heavily involved with the multi-university Research Accelerator for Multi-Processors (RAMP) consortium hosted at the Berkeley Wireless Research Center (BWRC), which is developing technology for rapid simulation of large-scale multiprocessors using Field-Programmable Gate Arrays (FPGAs). The Par Lab RAMP Gold manycore simulator is being extended to support both the photonics work and the vector-thread processor work.

References

[1] C. F. Batten (2010). "Simplified Vector-Thread Architectures for Flexible and Efficient Data-Parallel Accelerators." Ph.D. thesis, Massachusetts Institute of Technology, February 2010.

[2] C. F. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C. Holzwarth, M. Popović, H. Li, H. Smith, R. Ram, V. Stojanović, and K. Asanović (2009). "Building Manycore Processor-to-DRAM Networks with Monolithic CMOS Silicon Photonics." *IEEE Micro*, Vol. 29, Issue 4, pp. 8-21, July/August 2009.

[3] S. Beamer (2009). "Designing Multisocket Systems with Silicon Photonics." Master's thesis, UC Berkeley, December 2009.

[4] S. Beamer, K. Asanović, C. Batten, A. Joshi, and V. Stojanović "Designing Multisocket Systems Using Silicon Photonics." Proceedings of the 23rd International Conference on Supercomputing (ICS), Yorktown Heights, New York, pp. 521-522, June 2009.

[5] A. Joshi, C. Batten, Y. Kwon, S. Beamer, I. Shamim, K. Asanović, and V. Stojanović (2009). "Silicon-Photonic (Clos) Networks for Global On-Chip Communication." Proceedings of the Third ACM/IEEE International Symposium on Networks-on-Chip (NOCS), San Diego, California, pp. 124-133, May 2009.

[6] 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.

[7] V. Stojanović, A. Joshi, C. Batten, Y. Kwon, and K. Asanović (2009). "Manycore Processor Networks with Monolithic Integrated CMOS Photonics." Invited paper at the 29th Conference on Lasers and Electro-Optics (CLEO), Baltimore, Maryland, June 2009.