INTERNATIONAL
COMPUTER SCIENCE
INSTITUTE

International Computer Science Institute Activity Report 2008

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 2008 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)
Google
IM2 National Centre of Competence in Research, Switzerland
Microsoft
National Science Foundation (NSF)
Qualcomm
Spanish Ministry of Science and Innovation (MICINN) (formerly the Ministry of Science (MEC))

Affiliated 2008 Sponsors

Appscio
Intel
National Institutes of Health (NIH)
SAP
Volkswagen
XORP, Inc.

Corporate Officers

Prof. Nelson Morgan (President and Institute Director)
Prof. Scott Shenker (Vice President)
Maria Eugenia Quintana (Secretary)
Prof. Richard Karp (Treasurer)

Board of Trustees, January 2009

Prof. Javier Aracil, MICINN and Universidad Autónoma de Madrid
Prof. Hervé Bourlard, IDIAP and EPFL
Vice Chancellor Beth Burnside, UC Berkeley
Dr. Adele Goldberg, Agile Mind, Inc. and Pharmaceutrix, Inc.
Dr. Greg Heinzinger, Qualcomm
Mr. Clifford Higgerson, Walden International
Prof. Richard Karp, ICSI and UC Berkeley
Prof. Nelson Morgan, ICSI (Director) and UC Berkeley
Dr. David Nagel, Ascona Group
Prof. Prabhakar Raghavan, Stanford and Yahoo! Research
Prof. Stuart Russell, UC Berkeley EECS Department Chair
Prof. Shankar Sastry, UC Berkeley, Dean of the College of Engineering (Chairman)
Prof. Scott Shenker, ICSI and UC Berkeley
Dr. Eero Silvennoinen, TEKES
Dr. David Tennenhouse, New Venture Partners
Prof. Wolfgang Wahlster, DFKI GmbH

2008 International Visitor Program

Name	Country	Group	Sponsor
Bin Dai	China	Networking	CSC
Po-Ching Lin	China	Networking	CSC
Li Tang	China	Networking	CSC
Bo Xu	China	Networking	CSC
Youquan Zheng	China	Networking	CSC
Joan Isaac Biel	EU	Speech	AMIDA
Rosemary Orr	EU	Speech	AMIDA
Korbinian Riedhammer	EU	Speech	AMIDA
David Van Leeuwen	EU	Speech	AMIDA
Oriol Vinyals	EU	Speech	AMIDA
Yoshia Hirase	Finland	Other	TEKES
Jyri Kivinen	Finland	Algorithms	TEKES
Kimmo Kuusilinna	Finland	Other	TEKES
Teemu Koponen	Finland	Networking	TEKES
Tommi Lampikoski	Finland	Other	TEKES
Annukka Näyhä	Finland	Other	TEKES
Boris Nechaev	Finland	Networking	TEKES
Teemu Roos	Finland	Other	TEKES
Ville-Pekka Seppä	Finland	Other	TEKES
Jan Baumbach	Germany	Algorithms	DAAD
Bernd Bohnet	Germany	Speech	DAAD
Gerald Friedland	Germany	Speech	DAAD
Tobias Friedrich	Germany	Algorithms	DAAD
Martin Gairing	Germany	Algorithms	DAAD
Martin Hilpert	Germany	AI / FrameNet	DAAD
Thomas Kleinbauer	Germany	AI / FrameNet	DAAD
Christian Kreibich	Germany	Networking	DAAD
Birte Lönneker-Rodman	Germany	AI / FrameNet	DAAD
Gregor Maier	Germany	Networking	DAAD
Andreas Maletti	Germany	AI	DAAD
Christian Müller	Germany	Speech	DAAD
Andreas Raabe	Germany	Architecture	DAAD
Ulrich Rueckert	Germany	Algorithms	DAAD
Felix Salfner	Germany	Other	DAAD
Thomas Sauerwald	Germany	Algorithms	DAAD
Guido Schryen	Germany	Networking	DAAD

Alberto Amengual	Spain	AI	MICINN
Lucia Conde	Spain	Algorithms	MICINN
Oscar Ferrandez	Spain	AI	MICINN
Eduardo Lopez	Spain	Speech	MICINN
Carlos Subirats	Spain	AI / FrameNet	MICINN
Enrique Torres	Spain	Architecture	MICINN
Neha Garg	Switzerland	Speech	IM2
Nikhil Garg	Switzerland	Speech	IM2
Bao-Lan Huynh	Switzerland	Speech	IM2
David Imseng	Switzerland	Speech	IM2
Adish Singla	Switzerland	Speech	IM2
Leo Juan	Taiwan (China)	Networking	iCAST
Chih-Hung Lin	Taiwan (China)	Networking	iCAST
Jamon Liu	Taiwan (China)	Networking	iCAST

AMIDA: Augmented Multi-party Interaction with Distance Access
CSC: China Scholarship Council
DAAD: Deutscher Akademischer Austausch Dienst
iCAST: International Collaboration for Advancing Security Technology
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 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). 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.

The particular areas of 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 areas: Internet Research, including Internet architecture, related theoretical questions, and network security; and Perceptual and Cognitive Systems, including speech, 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, etc. There are approximately one hundred scientists in residence at ICSI including permanent staff, postdoctoral Fellows, visitors, affiliated faculty, and students. Senior investigators are listed at the end of this overview, along with their current interests. The current Director of the Institute is Professor Nelson Morgan of the UC Berkeley Electrical Engineering faculty.

1 Institute Sponsorship for 2008

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.

US Federal funding in 2008 came from a range of grants to support research Institute-wide. Most of this funding comes from the National Science Foundation, DARPA, and IARPA. International support in 2008 came from the Ministry of Education and Research in Germany, the Ministry of Education and Science in Spain, the National Technology Agency of Finland, and the Swiss National Science Foundation (through the Swiss Re-

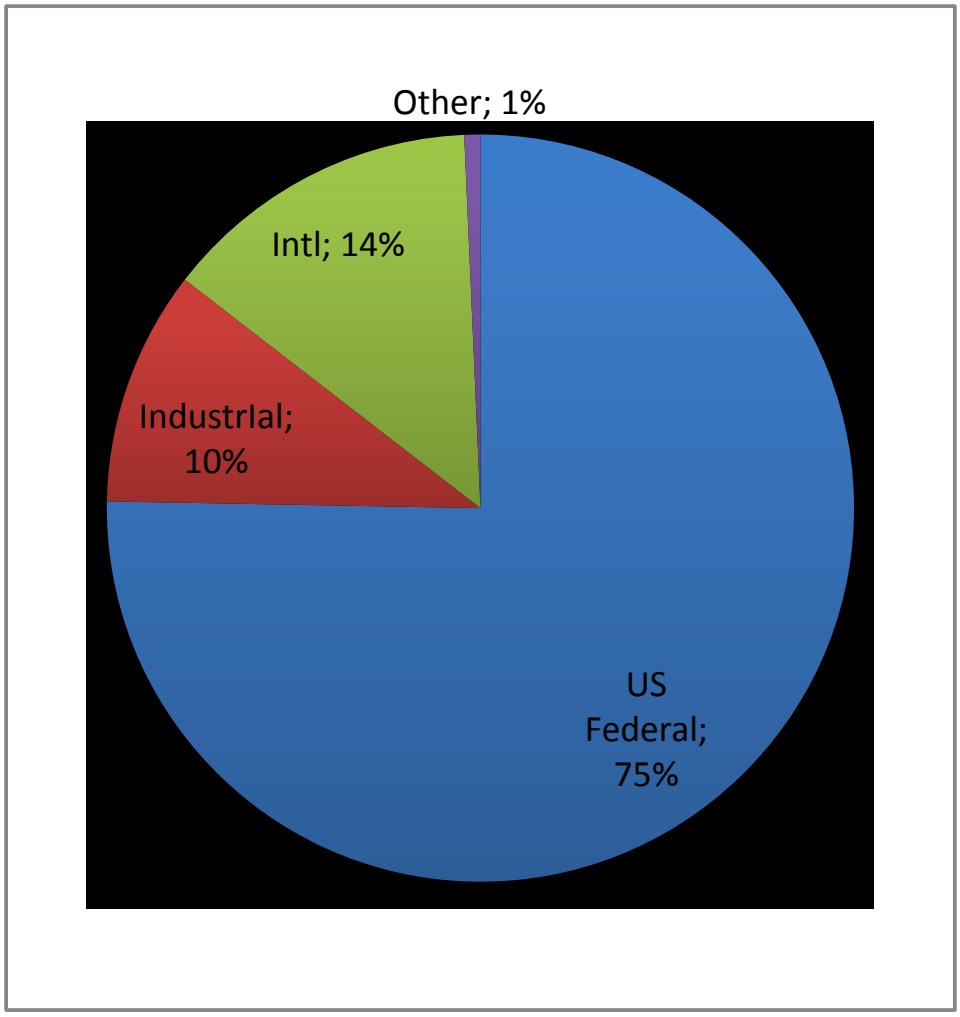

Figure 1: Distribution of sources of ICSI revenue for 2008.

search Network IM2). Additional support came from the European Union (as a partner in the Integrated Project, AMIDA). Industrial support in 2008 was provided by Qualcomm, Google, SAP, Lockheed, Microsoft, Intel, Volkswagen, AppScio, and Xorp, Inc.. Total ICSI revenue was $8.5M in 2008.

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: the Administrative and Research organizations.

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 Chair of that organization is Professor Shankar Sastry of the EECS Department. Ongoing operation of the Institute is the responsibility of Corporation Officers, namely the President, Vice President, the Secretary and the Treasurer. 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 Controller. All other support activities come under the general heading of Administrative Operations, and are supervised by the Operations Manager; these activities include the visitor program, office assignments, 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. For some years we have had four groups: the Networking Group (internet research), the Algorithms Group, the AI Group, and the Speech Group. In 2008 we expanded to two other groups: 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 four groups.

Across these activities, there is a theme: scientific studies based on the growing ubiquity of connected computational devices. In the case of Networking studies, 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, though, 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 2009) senior research staff members at ICSI, along with a brief description of their current interests and the Research Group that the researcher is most closely associated with. 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): neural 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;

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 and Networking): mathematics of computer networking, computational molecular biology, computational complexity, combinatorial optimization;

Paul Kay (AI): analyzing 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;

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

John Moody (Algorithms): machine learning, multi-agent systems, statistical computing, time series analysis, computational finance;

Srini Narayanan (AI): probabilistic models of language interpretation, graphical models of linguistic aspect, graphical models of stochastic grammars, semantics of linguistic

aspect, on-line metaphor interpretation, and embodied rationality; more recently models of the role of sub-cortical structures (like basal ganglia-cortex loops) in attentional control;

Vern Paxson (Networking): intrusion detection; internet measurement; measurement infrastructure; packet dynamics; self-similarity;

Scott Shenker (Networking): congestion control, internet topology, game theory and mechanism design, scalable content distribution architectures, and quality of service;

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);

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 the year 2007, 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 assists underrepresented students in computer science and engineering in getting college acceptances and scholarship dollars.

1.1 Networking

- In a novel study, ICSI researchers and colleagues measured the efficacy of several spam campaigns from an "inside" perspective. By interceding on the control mechanisms used by a spammer to send spamming instructions to compromised systems, the researchers were able to introduce perturbations to the spammer's directives such that they could match up emails sent with subsequent visits to the web sites advertised in the spam, including which visitors then attempted to make purchases. The measurements led to the remarkable estimate that the spammer had to send more than 12 million spams to yield a single $100 purchase of Viagra.

- ICSI spun out its eXtensible Open Router Platform (XORP) technology to XORP, Inc. (http://xorp.net), a startup founded by the leaders of the XORP.org project (http://xorp.org). This coincided with the debut of the XORP 1.5 Release, the first

community release of XORP to be made under the auspices of XORP, Inc. In a message to the XORP.org community, the company stated its commitment both to supporting the XORP.org community and code base and to building a successful open source business based on XORP.

- Recent joint work with Stanford University on secure enterprise networks has led to the development (led by the startup Nicira Networks) of a Network Operating System, called NOX, that is now freely available (under a GPL license) at noxrepo.org. This system is currently being used by several research groups to explore different approaches to managing enterprise networks.

1.2 Algorithms

- Kyoto Prize: Richard Karp became received the Kyoto Prize in Information Science, which is granted every four years. The citation states that Karp's work NP-completeness "streamlined algoritm design for problem solving, accelerated algorithm engineering and brought computational complexity within the scope of scientific research."

- Statistical Genetics: Eran Halperin has led our diverse efforts in statistical genetics. One of a number of related highlights is the development of algorithms and software (LAMP, SWITCH and WINPOP) for the inference of ancestry in recently admixed populations such as African-Americans or Latinos in the United States. This work has many applications, particularly in medicine, since disease association studies that are performed on such populations directly depend on such methods. Participants in these and other investigations in this area include Lucia Conde, Michael Jordan (UCB), Richard Karp, Bonnie Kirkpatrick, Bogdan Pasaniuc, Sriram Sankaramanan and Meromit Schuster.

- Genetic Regulation: A fundamental tool in genomics is the BLAST program for matching a genomic query sequence against a target database. Over the past few years Richard Karp and colleagues at UC San Diego and Tel-Aviv University have pioneered the extension of this methodology from sequences to genetic regulatory networks representing interactions among proteins.In recent work we present efficient algorithms to identify conserved regulatory structures by searching a target network for compact subnetworks whose proteins match the proteins in a known pathway.

- Load Balancing: Tobias Friedrich, Thomas Sauerwald and colleagues from UCB have worked on load-balancing on distributed networks. They proved that a new randomization technique significantly speeds up the convergence of the load balancing process and measured precisely how much randomness is needed to do so.

1.3 Artificial Intelligence

- Five graduate students from the AI group completed their PhD dissertations in 2008. Three students (John Bryant, Nancy Chang, and Eva Mok) worked on language acquisition and use using the Embodied Construction Grammar (ECG) framework

developed by the NTL group. Steve Sinha combined the NTL event modeling and inference algorithms with FrameNet to build a semantically oriented question answering system. Joseph Makin completed the first extant computational model and comprehensive analysis of the mammalian coagulation pathway.

- In 2008, Srini Narayanan was appointed as a Fellow at the Institute for Advanced Study in Berlin to write a monograph on his work in the NTL project. This appointment is for the 2008-2009 academic year.

- The ICSI AI group started a multi-year collaboration with Hesperian Press to bring primary healthcare access to the emerging world using speech and local language interfaces coupled with semantic search technologies. The initial pilot project is being funded by Google, the Bill and Melinda Gates foundation, and the Rockefeller foundation.

- The FrameNet (FN) group expanded its collaboration and guidance of parallel FrameNets in multiple languages. The group now has active collaborations with developers of Spanish FN (U A Barcelona), the SALSA Project (German–U Saarlandes), Japanese FN (Keio U and U Tokyo), Chinese FN (Shanxi U, Taiyuan), and others.

1.4 Speech

- Group members significantly improved sentence segmentation, extending methods that mainly rely on local features by the use of syntax.

- The summarization of meetings and multiple documents have been greatly improved by the use of keyphrase based methods and integer linear programming.

- Diarization efforts moved towards online systems by achieving comparable performance for much shorter speech segments than were previously possible.

- Using multi-stream speech recognition with up to 28 streams reduced the number of errors for noisy speech by nearly half.

1.5 Computer Architecture

- The architecture group began at ICSI in summer of 2007, and in 2008 welcomed the first two international postdoc visitors in this area.

- The architecture group received a DARPA award to study the application of monolithic silicon photonics to processor-memory interconnect.

- Architecture group leader Asanovic was a key player in the successful initiation of the Microsoft/Intel-funed Parallel Lab.

1.6 Computer Vision

- The new ICSI Vision group moved from MIT in 2008, with ICSI research beginning in January (based on significant funding from DARPA, NSF, and industry) and the group physically arriving in July.
- The group already has major publications at CVPR, ICML, and NIPS.

2 Networking

2.1 Measurements and Modeling

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

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 [46]. 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 *ISPprobe* project [33], we are developing a Web-driven measurement apparatus that informs Internet users about the extent of ISP-level interference with their Internet connectivity, allowing us to collect this information on a broad scale for later analysis. Among the analyses that our framework currently supports are port blocking, network address translation, generic and HTTP-specific proxying, inline content modification, in-modem packet buffering, DNS-level resolver security, and DNS query redirection.

A third effort assess 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 [53]. 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

Reactive measurement: Reactive measurement (REM) is a measurement technique in which one measurement's results are used to decide what (if any) additional measurements are required to further understand some observed phenomenon. While reactive measurement has been used on occasion in measurement studies, what has been lacking is *(i)* an examination of its general power, and *(ii)* a generic framework for facilitating fluid use of this approach. We believe that by enabling the coupling of disparate measurement tools, REM holds great promise for assisting researchers and operators in determining the root causes of network problems and enabling measurement targeted for specific conditions. We are currently exploring reactive measurement using two approaches. First, we have a prototype REM system that a researcher can use to connect measurement tools together and we have been using this to explore both HTTP and DNS failures. In addition, we have built a plug-in for the Firefox web browser that takes ancillary measurements when an error in loading a web page is found. These measurements are then sent back to a central database for analysis. While these efforts are in fairly early stages, the overall architecture is more mature [9].

An open measurement platform: In this project we pursue the development of a measurement framework that researchers can use to abstract away the mundane logistical details that tend to dog every measurement project [8]. The measurement community has outlined the need for better ways to gather assessments from a multitude of vantage points, and our system is designed to be an open community-oriented response to this desire. While many previous efforts have approached this problem with heavyweight systems that ultimately fizzle due to logistical issues (e.g., hosts breaking and no money to replace them), we take the opposite approach and attempt to use the lightest-weight possible framework that allows researchers to get their work done. In particular, we take the approach of designing a system without any sort of central "core" component, and therefore the system has no single point of failure. In addition, our proposed system is community-oriented in that there is no central control; we build just enough mechanism for the community to get their work done and police the infrastructure. In addition, our proposed system works in an open fashion, such that results from the community's infrastructure are immediately provided to the community through publicly available "live feeds".

Measuring OpenDHT: A number of applications have proposed using distributed hash tables (DHTs) for a variety of database tasks (e.g., new naming schemes or aggregating RSS feeds). DHTs present logistical challenges to set up and deploy. Therefore, OpenDHT was developed as a general-purpose DHT service that provides a simple *get()/put()* interface to developers and is advertised as useful across many applications. While in principle various proposed databases fit nicely within the DHT abstraction, and OpenDHT is attractive for its ease of use, each application will ultimately have its own performance requirements which the DHT system may or may not be able to meet. Therefore, in practice the suitability of using a DHT data structure may not be readily apparent. In this project

we have captured an initial set of simple measurements to assess the performance of the OpenDHT service. In addition, we are assessing the reliability and responsiveness for various operations from an application's viewpoint.

2.2 Security, Malware, and Intrusion Detection

Exploiting multi-core processors to parallelize network intrusion prevention: It is becoming increasingly difficult to implement effective systems for preventing network attacks, due to the combination of (1) the rising sophistication of attacks requiring more complex analysis to detect, (2) the relentless growth in the volume of network traffic that we must analyze, and, critically, (3) the failure in recent years for uniprocessor performance to sustain the exponential gains that for so many years CPUs enjoyed ("Moore's Law"). For commodity hardware, tomorrow's performance gains will instead come from *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 [45].

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 [47], and evaluated the implementation using stress-testing [51]. We are now working with LBNL to expand the prototype deployment there to one suitable for 24x7 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.

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 two studies using infiltration of the well-known *Storm* botnet [39, 38].

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.

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.

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 specialized products and services. This

project, joint with UC San Diego, aims to explore these marketplaces in an attempt to characterize their constituencies, impact, and sundry elements, in the hope that such an analysis might shed light on bottlenecks/weakspots present in the underground economy that can then be targeted to provide maximal benefit for defenders [29]. 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 [7]. 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 in addition 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 this latter (the most common), we then attempt to refine the analysis to assess whether the scanning *targeted* the site in particular, or was merely part of a much broader, indiscriminate scan. Our preliminary results indicate our analysis using *purely local information* generally yields estimates of global targeting scope quite close to those obtained more directly from the global *DShield* repository of Internet scanning activity [40].

Building a network "time machine": Insight into past network traffic can have enormous value, both for forensics when analyzing a problem detected belatedly, and to augment real-time decision-making, both to inform *reactive measurement* (see above) and to give additional pinpoint context to a network intrusion detection system (NIDS). This project aims to develop a network *time machine*, which works by passively bulk-recording as much network traffic as possible. The time machine maintains a ring buffer of recent network traffic that matches a given criteria. This criteria needn't be a simple static filter—the decision of what to capture and for how long could be much richer and incorporate more context. This buffer resides in RAM for fast access, with the decision of what traffic to record in the buffer, and how to filter it (e.g., retaining the first N bytes of each connection), being driven off of a collection of policies describing retention for different types of activity. In addition, recorded traffic migrates from RAM to a given allocation of disk space, which is also managed per a collection of policies that again determine which traffic to migrate, how to filter it, and how to expire it as the disk allocation fills up.

Recent efforts on this project have focused on integrating our Time Machine implementation with a real-time NIDS by providing an API by which a NIDS (in our case, the Bro system) can query activity seen in the recent past for given connections or hosts. This coupling has the potential to greatly offload the NIDS, allowing it to process only lighter-weight request streams and not response streams, unless it sees a problematic request, in which case it can at that point ask the time machine for a copy of the reply to that particular request [41]. We also are pursuing development of a clusterized version of the Time Machine to enhance its scalability for monitoring very high volume environments.

Robust TCP stream normalization: One of our previous efforts investigated algorithms by which hardware devices can reassemble TCP bytestreams even in the presence of adversaries who will attempt to subvert the hardware's operation by overwhelming its state management. This follow-on project looks at the next step: how to *normalize* the byte stream in order to assure that we can remove any ambiguities in terms of inconsistent TCP retransmissions. Hardware vendors have asserted that simply hashing the contents of previously seen packets suffices to provide such normalization. We have found that by itself, this approach renders a great deal of "collateral damage" in terms of retransmitted traffic that must be discarded because it does not align with previously recorded hashes, or in *evasion* opportunities, if such traffic is allowed to proceed in the absence of a previous hash against which we can check it. The approach we have developed in this regard, however, is robust to such variations, as well as to attackers who deliberately target the state we must manage to provide such normalization [48].

Rate-based scan detection: This project explores developing light-weight worm detection algorithms that offer significant advantages over fixed-threshold methods. The first algorithm, RBS (rate-based sequential hypothesis testing), aimed at the large class of worms that attempt to quickly propagate, thus exhibiting abnormal levels of the rate at which hosts initiate connections to new destinations. The foundation of RBS derives from the theory of sequential hypothesis testing, the use of which for detecting randomly scanning hosts was first introduced by our previous work with the *Threshold Random Walk*

algorithm [34]. The sequential hypothesis testing methodology enables engineering the detectors to meet false positives and false negatives targets, rather than triggering when fixed thresholds are crossed. In this sense, the detectors that we introduce are truly adaptive.

We then developed RBS+TRW, an algorithm that combines fan-out rate (RBS) and probability of failure (TRW) of connections to new destinations. RBS+TRW provides a unified framework that at one end acts as a pure RBS and at the other end as pure TRW, and extends RBS's power in detecting worms that scan randomly selected IP addresses. Using four traces from three qualitatively different sites, we evaluated RBS and RBS+TRW in terms of false positives, false negatives, and detection speed, finding that RBS+TRW provides good detection of high-profile worms, internal Web crawlers, and a network monitoring tool that we used as proxies for targeting worms. In doing so, RBS+TRW generates fewer than 1 false alarm per hour for wide range of parameter choices [36, 37].

Predicting resource consumption of network intrusion detection systems: When installing network intrusion detection systems (NIDSs), operators are faced with a large number of parameters and analysis options for tuning trade-offs between detection accuracy versus resource requirements. In this effort we set out to assist this process by understanding and predicting the CPU and memory consumption of such systems. We started towards this goal by devising a general NIDS resource model to capture the ways in which CPU and memory usage scale with changes in network traffic. We then used this model to predict the resource demands of different configurations for specific environments, leading to an approach to derive site-specific NIDS configurations that maximize the depth of analysis given predefined resource constraints.

We have validated our approach by applying it to the open-source Bro NIDS, testing the methodology using real network data, and developing a corresponding tool that automatically derives a set of configurations suitable for a given environment based on a *sample* of the site's traffic. While no automatically generated configuration can ever be optimal, these configurations provide sound starting points, with promise to significantly reduce the traditional trial-and-error NIDS installation cycle [23].

Testing evasion resilience of network intrusion detection systems: Network intrusion detection systems (NIDS) face a difficult, fundamental problem in the degree to which attackers can exploit ambiguities present when monitoring network traffic in order to undermine the correctness of the NIDS's analysis to evade detection. However, many of today's NIDSs lack the additional mechanisms required to resist different forms of evasion, because the underlying problems are subtle and—critically—*not visible* to the customers who purchase these systems.

Remedying this common shortcoming of modern NIDS functionality requires the widespread availability of *test suites* oriented towards probing the degree to which a NIDS exhibits evasion vulnerabilities. In this project we undertake the creation of a framework to facilitate the development of such test suites. Our prototype system takes as input a packet trace and from it constructs a configurable set of variant traces that introduce different forms of ambiguities that can lead to evasions. Our test harness then uses these variant traces in either an *offline* configuration, in which the NIDS under test reads traffic from the traces

directly, or a *live* setup, in which we employ replay technology to feed traffic over a physical network past a NIDS reading directly from a network interface, and to potentially live victim machines. Summary reports of the differences in NIDS output tell the analyst to what degree the NIDS's results vary, reflecting sensitivities to (and possible detections of) different evasions. We have used it to test the open-source *Snort* and *Bro* systems [35].

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 [49]. 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 [4].

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 [52]. We in addition use network traces to evaluate potential defenses.

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 [12]. This effort focuses on revising this previous work to clarify issues and ambiguities that have been identified since its publication [11].

Reacting to spurious retransmissions: TCP and SCTP both provide reliability by retransmitting lost data. In addition, losses are taken as an indication of network congestion and used to trigger congestion control in the data sender (i.e., a reduction in the

sending rate). Spurious retransmissions are generally caused by a transport's lack of ability to cope with the dynamics of a network path (e.g., a widely varying round-trip time). Several schemes have been devised to detect spurious retransmissions. In this effort, we investigate a response to spurious timeouts whereby we alter the calculation of the retransmission timer in an attempt to account for more variation in the round-trip time and prevent further spurious retransmissions [20].

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 [5] allows the transport to reduce (in certain special circumstances) the number of duplicate acknowledgments required to trigger a fast retransmission. Doing so allows the transport to use Fast Retransmit to recover packet losses that would otherwise require a lengthy retransmission timeout.

Equation-Based Congestion Control: A few years ago ICSI researchers, along with others, developed an approach to congestion control called *equation-based* congestion control. Rather than imitating TCP's window adjustment algorithm, equation-based congestion control seeks to match TCP's bandwidth usage equation in the longer run (the equation describing TCP's bandwidth usage resulting from a given packet-drop rate), while avoiding TCP's halving of the sending rate in response to a single packet drop. A protocol specification for this congestion-control mechanism, called TFRC (TCP-Friendly Rate Control), was standardized in early 2003 in RFC 3448. In 2008 the authors of RFC 3448 produced a revised document, RFC 5348 [28], updating the TFRC standard. The main change to the TFRC standard is the addition of mechanisms to address the case of a data-limited sender.

Transport Modeling: We believe that research in congestion control mechanisms has been seriously hampered by the lack of good models underpinning analysis, simulation, and testbed experiments, and we are participating in a larger effort towards improving the models that we use in the evaluation of transport protocols. One part of this work is a document discussing the metrics to be considered in an evaluation of new or modified congestion control mechanisms for the Internet [26]. This includes metrics for the evaluation of new transport protocols, of proposed modifications to TCP, of application-level congestion control, and of Active Queue Management (AQM) mechanisms in the router. A second part of this work is an initial proposal for a test suite for the evaluation of proposed TCP modifications [19]. Several Internet-Drafts are in progress to further develop test suites for the evaluation of congestion control mechanisms.

2.4 Novel Internet Architectures

Architectural support for network trouble-shooting: Troubleshooting is an inherent part of network operation: no matter how well networks are designed, something eventually fails, and in large networks, failures are ever-present. In the past, troubleshooting has mostly relied on *ad hoc* techniques cobbled together as afterthoughts. However,

both the importance and difficulty of troubleshooting has intensified as networks have become crucial, ubiquitous components of modern life, while at the same time their size and complexity continues to grow. These twin pressures highlight the urgent need to integrate troubleshooting as a first-class citizen when developing a network architecture.

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

The Usefulness of Best-Effort Traffic: In this effort we develop a number of observations on the capabilities and limitations of "simple best-effort" traffic, defined loosely as Internet traffic that is not covered by Quality of Service mechanisms, congestion-based pricing, cost-based fairness, admissions control, or the like [27]. One core observation is that simple best-effort traffic serves a useful role in the Internet, and is worth keeping. While differential treatment of traffic can clearly be useful, we believe such mechanisms have utility primarily as adjuncts to simple best-effort traffic, not as replacements. A second observation is that for simple best-effort traffic, some form of rough "flow rate fairness" is a useful goal for resource allocation, by which we mean attaining equal rates for different flows over the same path.

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 [15]. 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.

Pathlet Routing: We have developed a new multipath routing protocol, pathlet routing, in which networks advertise fragments of paths (pathlets) over virtual nodes (see [30]). 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 routers 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 New Communications API: We have developed NetAPI (see [54]), a flexible communications interface. Although the ubiquitous Sockets API lets applications select among a number of mechanisms to accomplish networking tasks, it binds them tightly to their chosen mechanisms. Consequently, the network stack has little freedom in selecting the best protocols and mechanisms for each application, and innovating below the API is extremely difficult. NetAPI allows applications to specify their communication intentions against an abstract interface that hides implementation mechanisms, encouraging innovation below the API. Application intents are combined with user policies and environmental conditions to let the network meet application goals in varied ways. We have also implemented a prototype of NetAPI called PANTS for the iPhone platform.

Rethinking Packet Forwarding Hardware: We have developed a new approach to building network forwarding devices (e.g., switches and routers) that provides a good compromise between low cost/performance and high flexibility (see [44, 43]). In this approach, all forwarding decisions are made by software, and thus are fully flexible, but after the initial software-based decision all subsequent forwarding actions are implemented by hardware, using some form of a classification engine. We have built a full realization of this design, using NetFPGA hardware, and have shown how to use this approach for a variety of traditional (e.g., L2 and L3), and recently proposed (e.g., SEATTLE) forwarding algorithms, and have analyzed the performance of this approach over real traces.

NOX: Towards an Operating System for Networks: As anyone who has operated a large network can attest, enterprise networks are difficult to manage. That they have remained so despite significant commercial and academic efforts suggests the need for a different network management paradigm. We can turn to operating systems as an instructive example in taming management complexity.

In the early days of computing, programs were written in machine languages that had no common abstractions for the underlying physical resources. This made programs hard to write, port, reason about, and debug. Modern operating systems facilitate program development by providing controlled access to high-level abstractions for resources (e.g., memory, storage, communication) and information (e.g., files, directories). These abstractions enable programs to carry out complicated tasks safely and efficiently on a wide variety of computing hardware. In contrast, networks are managed through low-level configuration

of individual components. Moreover, these configurations often depend on the underlying network; for example, blocking a users access with an ACL entry requires knowing the users current IP address. In this way, an enterprise network resembles a computer without an operating system, with network-dependent component configuration playing the role of hardware-dependent machine-language programming.

What we clearly need is an "operating system" for networks, one that provides a uniform and centralized programmatic interface to the entire network. Analogous to the read and write access to various resources provided by computer operating systems, a network operating system provides the ability to observe and control a network. In particular, a network operating system allows management applications to be written as centralized programs over high-level names as opposed to the distributed algorithms over low-level addresses we are forced to use today. While clearly a desirable goal, achieving this transformation from distributed algorithms to centralized programming presents significant technical challenges. In [32] we describe a scalable and robust Network Operating System called NOX. It is now available at noxrepo.org.

Packet Caches on Routers: Many past systems have explored how to eliminate redundant transfers from network links and improve network efficiency. Several of these systems operate at the application layer, while the more recent systems operate on individual packets. A common aspect of these systems is that they apply to localized settings, e.g. at stub network access links. In this project, we explored the benefits of deploying packet-level redundant content elimination as a universal primitive on all Internet routers. Such a universal deployment would immediately reduce link loads everywhere. However, we argue that far more significant network-wide benefits can be derived by redesigning network routing protocols to leverage the universal deployment. We develop "redundancy-aware" intra- and inter-domain routing algorithms and show that they enable better traffic engineering, reduce link usage costs, and enhance ISPs' responsiveness to traffic variations. In particular, employing redundancy elimination approaches across redundancy-aware routes can lower intra and inter-domain link loads by 10-50%. We also address key challenges that may hinder implementation of redundancy elimination on fast routers. Our current software router implementation can run at OC48 speeds.

OpenFlow: A group of researchers have proposed OpenFlow as a way for researchers to run experimental protocols in the networks they use every day (see [42]). OpenFlow is based on an Ethernet switch, with an internal flow-table, and a standardized interface to add and remove flow entries. Our goal is to encourage networking vendors to add OpenFlow to their switch products for deployment in college campus backbones and wiring closets. We believe that OpenFlow is a pragmatic compromise: on one hand, it allows researchers to run experiments on heterogeneous switches in a uniform way at line-rate and with high port-density; while on the other hand, vendors do not need to expose the internal workings of their switches. In addition to allowing researchers to evaluate their ideas in real-world traffic settings, OpenFlow could serve as a useful campus component in proposed large-scale testbeds like GENI.

Accountable Internet Protocol: We have developed AIP (Accountable Internet Protocol), a network architecture that provides accountability as a first-order property. AIP uses a hierarchy of self-certifying addresses, in which each component is derived from the public key of the corresponding entity. In [18] we discuss how AIP enables simple solutions to source spoofing, denial-of-service, route hijacking, and route forgery. We also discuss how AIPs design meets the challenges of scaling, key management, and traffic engineering.

Anomaly-Cognizant Forwarding: It is well known that BGP convergence can cause widespread temporary losses of connectivity resulting from inconsistent routing state. In this project (see [25]) we developed Anomaly-Cognizant Forwarding (ACF) - a novel technique for protecting end-to-end packet delivery during periods of convergence. Our preliminary evaluation demonstrates that ACF succeeds in eliminating nearly all transient disconnection after a link failure without the use of precomputed backup routes or altering the dynamics of BGP.

2.5 Distributed Systems

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 hosted 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 [16].

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 (see [24]) 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 (see [21]), 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 (see [22]), we reallocate the "fault-tolerance budget" of a long-term service. We split the service into service 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.

2.6 Research Community Activities

We have contributed to the community's discussion of the overall review process for papers [2], as well as the specific question of how program committees and editorial boards should treat methodologically questionable papers [1].

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

Mark Allman chaired the IRTF's Internet Measurement Research Group (IMRG), co-chaired the IETF's TCP Maintenance and Minor Extensions (TCPM) Working Group and is a member of the IETF's Transport Area Directorate and General Area Review Team.

Vern Paxson co-founded and serves on the steering committee of the *USENIX Workshop on Large-scale Exploits and Emergent Threats*. He taught a tutorial *Understanding and Addressing the Threat of Internet Worms* for the Summer Undergraduate Program in Engineering Research at Berkeley (SUPERB) and gave invited briefings for the US Dept. of State / Office of Naval Research / National Science Foundation *Workshop on International Strategy and Policy for Cyber Security* and for the European Aeronautic Defence and Space Company. His work on Internet measurement was recognized by his receipt of the 2008 ACM Grace Murray Hopper Award. He chaired the ACM SIGCOMM *Test-of-Time*

Award committee and served on the steering committee of the first *Workshop on GENI and Security*. He presented a Distinguished Lecture at the University of Michigan entitled *Coming To Grips With Live Attackers*.

Robin Sommer co-taught a *Summer School on Internet Security*, organized by the German National Academic Foundation and held in La Villa, Italy.

References

[1] M. Allman (2008) "What Ought a Program Committee to Do?, Proceedings of USENIX Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems (WOWCS) San Francisco, California, April 2008.

[2] M. Allman (2008) "Thoughts on Reviewing, *ACM Computer Communication Review*, Vol. 38, Issue 2, pp. 47-50 April 2008.

[3] M. Allman (2008) "TCP Slow Start Survey: Standards and Issues, IETF Internet-Draft Draft-ietf-tcpm-early-rexmt-00.txt, August 2008, in progress.

[4] M. Allman (2009) Comments on Selecting Ephemeral Ports, *ACM Computer Communication Review* April 2009, under submission. THIS INFORMATION IS FROM ALLMANS WEB SITE

[5] M. Allman, K. Avrachenkov, U. Ayesta, J. Blanton, and P. Hurtig (2008) "Early Retransmit for TCP and SCTP, IETF Internet-Draft Draft-ietf-tcpm-early-rexmt-00.txt, August 2008, in progress.

[6] M. Allman, K. Christensen, B. Nordman, and V. Paxson (2007) "Enabling an Energy-Efficient Future Internet Through Selectively Connected End Systems, Proceedings of ACM Special Interest Group on Data Communications Workshop on Hot Topics in Networks (ACM SIGGCOMM HotNets-VI) Atlanta, Georgia, November 2007.

[7] 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.

[8] M. Allman, L. Martin, M. Rabinovichz, and K. Atchinson (2008) "On Community-Oriented Internet Measurement, Proceedings of Passive and Active Measurement Conference, pp. 112-121 Cleveland, Ohio, April 2008.

[9] M. Allman and V. Paxson (2008) "A Reactive Measurement Framkework, Proceedings of Passive and Active Measurement Conference, pp. 92-101 Cleveland, Ohio, April 2008.

[10] 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.

[11] M. Allman, V. Paxson, and E. Blanton (2008) "TCP Congestion Control," IETF Internet-Draft Draft-ietf-tcpm-rfc2581bis-04.txt, April 2008, work in progress.

[12] M. Allman, V. Paxson, and W. Stevens (1999) "TCP Congestion Control," IETF RFC 2581, April 1999.

[13] M. Allman and V. Paxson (2007) "Issues and Etiquette Concerning Use of Shared Measurement Data, Proceedings of ACM SIGCOMM Conference on Internet Measurement, pp. 135-140 San Diego, California, October 2007.

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

[15] M. Allman, M. Rabinovich and N. Weaver (2008) "Relationship-Oriented Networking, In preparation.

[16] Z. Al-Qudah, H. Alzoubi, M. Allman, M. Rabinovich, and V. Liberatore (2008) "Efficient Application Placement in a Dynamic Hosting Platform, Proceedings of the International World Wide Web Conference Madrid, Spain, April 2009, to appear.

[17] A. Anand, A. Gupta, A. Akella, S. Seshan, and S. Shenker (2008) "Packet Caches on Routers: The Implications of Universal Redundant Traffic Elimination, Proceedings of ACM Special Interest Group on Data Communications Conference (SIGCOMM 2008), pp. 219-230 Seattle, Washington, August 2008.

[18] D. Andersen, H. Balakrishnan, N. Feamster, T. Koponen, D. Moon, and S. Shenker (2008) "Accountable Internet Protocol (AIP), Proceedings of ACM Special Interest Group on Data Communications Conference (SIGCOMM 2008), pp. 339-350, Seattle, Washington, August 2008.

[19] 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.

[20] J. Blanton, E. Blanton, and M. Allman (2008) "Using Spurious Retransmissions to Adapt the Retransmission Timeout, International Computer Science Technical Report 08-005 August 2008

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

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

[23] 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), pp. 135-154 Cambridge, Massachusetts, September 2008.

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

[25] A. Ermolinskiy and S. Shenker (2008) "Reducing Transient Disconnectivity using Anomaly-Cognizant Forwarding, Proceedings of ACM Special Interest Group on Data Communications Workshop on Hot Topics in Networks (ACM SIGGCOMM HotNets-VII) Calgary, Canada, October 2008.

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

[27] S. Floyd and M. Allman (2008) "Comments on the Usefulness of Simple Best-Effort Traffic, Informational RFC 5290, July 2008.

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

[29] 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), pp. 375-388 Alexandria, Virginia, October 2007.

[30] P. Godfrey, I. Ganichev, S. Shenker, and I. Stoica (2008) "Pathlet Routing, Proceedings of ACM Special Interest Group on Data Communications Workshop on Hot Topics in Networks (ACM SIGGCOMM HotNets-VII) Calgary, Canada, October 2008.

[31] 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), pp. 139-149 Alexandria, Virginia, October 2007.

[32] N. Gude, T. Koponen, J. Pettit, B. Pfaff, M. Casado, N. McKeown, and S. Shenker (2009) "NOX: Towards an Operating System for Networks," *ACM SIGCOMM Computer Communications Review*, Vol. 38, Issue 3, pp. 105-110 July 2008.

[33] *The ICSI ISPprobe*, `nettest.icir.org/`.

[34] 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, pp. 211-225 Oakland, California, May 2004.

[35] 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 (Extended Abstract)," Proceedings of Fifth GI International Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA), pp. 267-278 Paris, France, July 2008.

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

[37] 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, Number 1, pp. 197-210 February 2008.

[38] 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 ACM Computer and Communication Security Conference (ACM CCS), pp. 3-14 Alexandria, Virginia, October 2008.

[39] 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 San Francisco, California, April 2008.

[40] Z. Li, A. Goyal, Y. Chen and V. Paxson (2009) "Automating Analysis of Large-Scale Botnet Probing Events," Proceedings of ACM Symposium on InformAtion, Computer, and Communications Security (ASIACCS09) Sydney, Australia, March 2009, to appear.

[41] 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), pp. 183-194 Seattle, Washington, August 2008.

[42] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner (2008) "OpenFlow: Enabling Innovation in Campus Networks," *ACM Computer Communication Review*, Vol. 38, Issue 2, pp. 69-74 April 2008.

[43] D. Moon, M. Casado, T. Koponen, and S. Shenker (2009) "Bridging the Software/Hardware Forwarding Divide, in submission, 2009.

[44] D. Moon, M. Casado, T. Koponen, and S. Shenker (2008) "Rethinking Packet Forwarding Hardware, Proceedings of ACM Special Interest Group on Data Communications Workshop on Hot Topics in Networks (ACM SIGGCOMM HotNets-VII) Calgary, Canada, October 2008.

[45] 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, pp. 1-7 Princeton, New Jersey, April 2007.

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

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

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

[49] N. Weaver and M. Allman (2008) "On Constructing a Trusted Path to the User," November 2008, under submission.

[50] 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), pp. 199-206 Monterey, California, February 2007.

[51] 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.

[52] N. Weaver(2009) "Comprehensive DNS Resolver Defenses Against Cache Poisoning, work in progress.

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

[54] M. Zaharia, G. Ananthanarayanan, K. Heimerl, M. Demmer, T. Koponen, A. Tavakoli, K. Fall, I. Stoica, S. Shenker (2009) "A New Communications API, 2009, in submission.

3 Algorithms

3.1 Highlights of 2008

Kyoto Prize

Richard Karp received the Kyoto Prize in Information Science, which is granted every four years. The citation states that Karp's work on NP-completeness "streamlined algorithm design for problem solving, accelerated algorithm engineering and brought computational complexity within the scope of scientific research."

Statistical Genetics

Eran Halperin has led our diverse efforts in statistical genetics. One highlight is the development of algorithms and software (LAMP, SWITCH and WINPOP) for the inference of ancestry in recently admixed populations such as African-Americans or Latinos in the United States. This work has many applications, particularly in medicine, since disease association studies that are performed on such populations directly depend on such methods.

Other work in statistical genetics includes: full-genome association studies of Non-Hodgkin's Lymphoma, one of the most common types of cancer in the United States; methods that use the coalescent model to enhance association studies; methods for haplotype inference in complex pedigrees; and studies of the limitations of statistical methods that try to jeopardize the privacy of individuals when summary DNA data is exposed.

Gene Regulation

A fundamental tool in genomics is the BLAST program for matching a genomic query sequence against a target database. Over the past few years Richard Karp and colleagues at UC San Diego and Tel-Aviv University have pioneered the extension of this methodology from sequences to genetic regulatory networks representing interactions among proteins. In recent work we present efficient algorithms to identify conserved regulatory structures by searching a target network for compact subnetworks whose proteins match the proteins in a known pathway.

3.2 Introduction

Computational biology is the major focus of activity for the Algorithms Group. Our work in statistical genetics has centered around the study of associations between genetic variation and disease. We have conducted a genome-wide association study of Non-Hodgkin's Lymphoma, developed methods for the inference of local ancestry in recently admixed populations, developed techniques for correcting genotype measurement errors and im-

puting values to unobserved components of genotypes, and explored the extent to which the privacy of individual participants can be compromised by pooled genotype data. In the areas of gene regulation and functional genomics we have explored specific regulatory pathways in microbes and yeast, explored the nucleosomal structure of DNA, developed query-based methods for identifying conserved pathways of protein-protein interaction and studied how interactions between quantitative trait loci can predict interactions between genes. We have also conducted research on diverse topics in the theory of computation and the design and analysis of algorithms, including computational game theory, computational learning theory and specific algorithmic problems such as implicit set covering problems, randomized load balancing and the testing problem for digital microfluidic chips.

3.3 Statistical Genetics

This work was led by Eran Halperin. Participants include Lucia Conde, Michael Jordan (UCB), Richard Karp, Gad Kimmel, Bonnie Kirkpatrick, Bogdan Pasaniuc, Javier Rosa, Sriram Sankararaman, Christine Skibolla(UCB)and the Buck Institute for Aging.

Inference of Local Ancestry in Recently Admixed Populations: We have developed methods for the inference of ancestry in recently admixed populations. Recently admixed populations are populations that originated in two or more ancestral populations that were separated for many generations, and then started mixing in the last few generations. Examples of such populations are the African American or Latino populations in the US. 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.

Our goal is to infer the ancestry of each individual in each position in the genome. 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 extended previous methods for the inference of locus-specific ancestry by the incorporation of a refined model of recombination events. We introduced an efficient dynamic programming algorithm for inferring the locus-specific ancestries in this model, resulting in a method that obtains improved accuracy; the improvement is most significant when the ancestral populations are closely related. We evaluated our approach on a wide range of scenarios showing that indeed, locus-specific ancestry can be accurately inferred. As a potential utility for our method in subsequent analysies, 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.

Our software package LAMP (Local Ancestry in adMixed Populations), was shown to accurately identify ancestries even when the ancestral populations are unknown. LAMP has been downloaded from the ICSI server more than 150 times since its publication.

In further work we developed a hidden Markov model that is augmented with recombination events. Parameter estimation in this model is done by an EM algorithm initialized

by the LAMP solution. The advantage of using such a model is that a number of biologically interesting parameters can now be estimated. For instance, we have used the model to predict historical recombination events and to estimate the allele frequencies in an ancestral population given those in the other population.

Imputation Methods in Association Studies: Recent advances in Single Nucleotide Polymorphism(SNP) genotyping technologies have made possible large scale genome-wide association studies that promise to uncover the genetic basis of complex human diseases. The validity of associations uncovered by these studies critically depends on the accuracy of the genotype data. Despite recent progress in genoype calling programs, SNP genotyping errors remain present at levels that can invalidate statistical tests for disease association, particularly for methods based on haplotype analysis. Furthermore, since causal SNPs are unlikely to be typed directly due to the limited coverage of current genotyping platforms, imputation of genotypes at untyped SNP loci has recently emerged as a powerful technique for increasing the power of association studies. We have developed several methods for performing these imputations, and comprehensive benchmarking of all the existing state of the art methods for genotype imputation is ongoing.

One such method is based on the haplotype information from repositories of human variation such as the HapMap project. At the core of this method is a left-to-right hidden Markov model used to represent the Linkage Disequilibrium (LD) patterns observed in the genotype population under study. A second method, WINPOP, uses locus-specific ancestry to improve imputation results in admixed populations. Two additional approaches are described below.

Association Mapping and Significance Estimation Via the Coalescent

The central questions asked in whole-genome association studies are how to locate associated regions in the genome and how to estimate the significance of these findings. Researchers usually do this by testing each SNP separately for association and then applying a suitable correction for multiple-hypothesis testing. However, SNPs are correlated by the unobserved genealogy of the population, and a more powerful statistical methodology would attempt to take this genealogy into account. Leveraging the genealogy in association studies is challenging, however, because the inference of the genealogy from the genotypes is a computationally intensive task, especially when recombination is modeled, as in ancestral recombination graphs. Furthermore, if large numbers of genealogies are imputed from the genotypes, the power of the study might decrease if these imputed genealogies create an additional multiple-hypothesis testing burden. In this work we showed that several existing methods that aim to address this problem suffer either from low power or from a very high false-positive rate; their performance is generally not better than the standard approach of separate testing of SNPs. We therefore suggested a new genealogy-based approach, CAMP (coalescent-based association mapping), that takes into account the trade-off between the complexity of the genealogy and the power lost due to the additional multiple hypotheses. Our experiments show that CAMP yields a significant increase in power relative to that of previous methods and that it can more accurately locate the associated region.

Haplotype Inference in Complex Pedigrees

Despite the desirable information contained in complex pedigree data sets, analysis

methods struggle to efficiently process these data sets. The attractiveness of pedigree data sets is their power for detecting rare variants, particularly in comparison with studies of unrelated individuals. In addition, rather than assuming individuals in a study are unrelated, knowledge of their relationships can avoid spurious results due to confounding population structure effects. However, a major challenge for the applicability of pedigree methods is the ability to handle complex pedigrees having multiple founding lineages, inbreeding and half-sibling relationships.

A key ingredient in association studies is imputation and inference of haplotpes from genotype data. Existing haplotype inference methods either do not efficiently scale to complex pedigrees or are limited in their accuracy. In this work, we present algorithms for efficient haplotype inference and imputation in complex pedigrees. Our method, PhyloPed, leverages the perfect phylogeny model, resulting in an efficient method with high accuracy. In addition, PhyloPed effectively combines the founder haplotype information from different lineages and is immune to inaccuracies in prior information about the founders.

Whole-Genome Association Studies of Non-Hodgkin's Lymphoma: We are conducting whole-genome-scan association studies of Non-Hodgkin's Lymphoma (NHL). NHL is the fifth most common cancer in the U.S. and its causes remain largely unexplained. We aimed to identify novel risk factors and genes that may be associated with development of the disease using high-throughput genotyping technologies and bioinformatics techniques. The characterization of these genetic factors involved in the etiology of lymphoma will help to identify which individuals within susceptible populations are most at risk and will lead to better treatment strategies. In [17] we describe some previously unknown associations between SNPs and NHL. A second manuscript is under review [16].

Genomic Privacy: Limits of Individual Detection in a Pool: Large-scale data from genome-wide association studies has presented the scientific community with the problem of how best this data can be shared while protecting the privacy of the participants. Many studies, until recently, pooled the participants and only made the allele frequencies at each SNP public. However, it was shown in a recent paper by Homer et al that whole-genome SNP arrays could be used to detect the presence of an individual, even in fairly large pools. As a result, such summary data have been removed from the public domain.

For many applications, it is useful to have summary data for a subset of the SNPs publicly available, provided an acceptable level of privacy can be ensured. We analyze the upper bound on the power achievable by any privacy-preserving method as a function of the number of SNPs m, the pool size n, the maximum power β and the false positive rate α. We can show that the allowed number of exposed SNPs m is linear in n (with the constant being a function of α and β), provided the minor allele frequencies are large and the SNPs are chosen to be independent. Importantly, our emprical results closely match the analysis. To this end, we plan to release a program that might serve as a practical guide in allowing researchers to share their data while preserving privacy.

3.4 Gene Regulation and Functional Genomics

Integrated Analysis, Visualization and Reconstruction of Microbial Gene Regulatory Networks: Appropriate handling of changing environmental conditions is crucial for any living microbial organism and is triggered by complex molecular strategies. These are coordinated by gene regulatory networks, which are stored and analyzed in reference databases, such as our CoryneRegNet platform. The rapidly growing number of sequenced organisms requires the necessity to transfer knowledge of regulatory networks from model organisms to others. Within the ICSI Algorithms Group we developed reliable computational models that attack several sub-problems for this monumental task [2]. (Jan Baumbach)

Topology-Free Querying of Protein Interaction Networks: In the network querying problem, one is given the protein-protein interaction network of species B and a query subnetwork from species A. The goal is to identify subnetworks of B that are similar to the query subnetwork. Existing approaches mostly depend on knowledge of the interaction topology of the query subnetwork; however, in practice, this topology is often not known. To combat this problem, we develop a topology-free querying algorithm, which we call TORQUE. Given a query, represented as a set of proteins, TORQUE seeks a matching set of proteins that are sequence-similar to the query proteins and span a connected region of the network, while allowing a limited number of mismatches consisting of insertions and deletions of proteins. The algorithm uses alternatively dynamic programming and integer linear programming for the search task.We test TORQUE with queries from yeast, fly and human, where we compare it to the QNet topology-based approach, and with queries from less studied species, where only topology-free algorithms apply. TORQUE detects many more matches than QNet, while in both cases giving results that are highly functionally coherent.(Richard Karp, in collaboration with the laboratories of Ron Shamir and Roded Sharan at Tel-Aviv University)

Nucleosomal Mapping Through The Use of Next-Generation Sequencing: Non-protein-coding sequences in mammalian genomes contain large amounts of regulatory information used to program the complexity of a mammalian gene environment. There is growing evidence that these sequences are associated with the nucleosomal structure of the DNA (i.e., the manner in which the chromosomal DNA coils around histone proteins to form chromatin). In this work we study the nucleosomal structure of the chromatin through the use of high-throughput next-generation sequencing. The next-generation sequencing technologies have been proposed recently as an alternative to the classical Sanger sequencing. They rely on carrying out many parallel reactions, with each reaction sequencing a short DNA fragment. The vast number of reactions that are run in parallel yield an enormous overall output (e.g., the Applied Biosystems SOLiD next-generation sequencing system produces around 40 million reads of 30-40 base pairs each, totaling 1 billion bases for each run). The huge amount of data generated by such platforms raises complex computational and statistical challenges both in mapping the short reads to the reference human genome and in interpreting the mapping results. Currently we are in the process of analyzing such datasets, with promising preliminary results showing the relation

between the epigenetic factors such as nucleosomal location and the development of stem cells. (Eran Halperin and Bogdan Pasaniuc)

Analysis of the Responses of the Yeast Molecular Network to Oxidative Stress: In response to environmental challenges, biological systems respond with dynamic adaptive changes in order to maintain the functionality of the system. Such adaptations may lead to cumulative stress over time, possibly leading to global failure of the system. When studying such system responses, it is therefore important to understand them in a system-wide and dynamic context. We hypothesized that dynamic changes in the topology of functional modules of integrated biological networks reflect their activity under specific environmental challenges. We introduced Topological Enrichment Analysis of Functional Subnetworks (TEAFS), a method for the analysis of integrated molecular profile and interactome data, which we validated by comprehensive metabolomic analysis of dynamic yeast response under oxidative stress. TEAFS identified activation of multiple stress response related mechanisms, such as lipid metabolism and phospholipid biosynthesis. We identified, among others, a fatty acid elongase IFA38 as a hub protein which was absent at all time points under oxidative stress conditions. The deletion mutant of the IFA38 encoding gene is known for the accumulation of ceramides. By applying a comprehensive metabolomic analysis, we confirmed the increased concentrations over time of ceramides and palmitic acid, a precursor of de novo ceramide biosynthesis. Our results imply that the connectivity of the system is being dynamically modulated in response to oxidative stress, progressively leading to the accumulation of (lipo)toxic lipids such as ceramides. (Eran Halperin, Peddinti V. Gopalacharyulu, Vidya R. Velagapudi, Erno Lindfors, and Matej Oresic).

Genetic Interactions in Natural Populations: Screens for genetic interaction, in which pairs of genes are systematically mutated and scored for their effects on phenotype, are a powerful resource for understanding molecular function. In contrast, gene linkage and association studies characterize the myriad natural mutations in a population to link genotype to phenotype. A significant open question is how to uncover genetic interactions in data from natural populations, and whether these interactions coincide with genetic interactions from systematic screens. Here we exploit structure within a recent gene linkage study in yeast, integrated with information on known protein complexes, to map a natural genetic network containing 2,163 interactions between distinct genomic regions.We observe significant overlap between this natural network and networks derived through systematic genetic analysis, even though the phenotypes used to generate these networks are different. Both networks also enrich for functional interactions within a core set of 32 complexes, although the interactions among the complexes are divergent. This study demonstrates how integrative analysis can elucidate combinations of factors underlying complex traits.(Richard Karp, in collaboration with the Trey Ideker lab at UCSD)

3.5 Design and Analysis of Algorithms

Implicit Set Covering Problems: Let U be a finite set and S a family of subsets of U. Define a *hitting set* as a subset of U that intersects every set in S. The optimal hitting set

problem is: given a positive weight for each element of U, find a hitting set of minimum total weight. This problem is equivalent to the classic weighted set cover problem. We consider the optimal hitting set problem in the case where the set system S is not explicitly given, but there is an oracle that will supply members of S satisfying certain conditions; for example, we might ask the oracle for a minimum-cardinality set in S that is disjoint from a given set Q. The problems of finding a minimum feedback arc set or minimum feedback vertex set in a digraph are examples of implicit hitting set problems. Our interest is in the number of oracle queries required to find an optimal hitting set. After presenting some generic algorithms for this problem we focus on our computational experience with an implicit hitting set problem related to multiple genome alignment (Richard Karp and Erick Montero Centero(UCB)).

Congestion Games and Covering Games: In [5], we extend Bayesian congestion games by including players who might act in a malicious way. In such a *malicious Bayesian congestion game* each player is of on of two *types*. Either the player is a rational player seeking to minimize her own delay, or – with a certain probability – the player is *malicious* in which case her only goal is to disturb the other players as much as possible.

We show that such games do in general not possess a Bayesian Nash equilibrium in pure strategies (i.e. a *pure Bayesian Nash equilibrium*). Moreover, given a game, we show that it is NP-complete to decide whether it admits a pure Bayesian Nash equilibrium. This result even holds when resource latency functions are linear, each player is malicious with the same probability, and all strategy sets consist of singleton sets of resources. For a slightly more restricted class of malicious Bayesian congestion games, we provide easily checkable properties that are necessary and sufficient for the existence of a pure Bayesian Nash equilibrium.

Moreover, we study the impact of the malicious types on the overall performance of the system (i.e. the *social cost*). To measure this impact, we use the *Price of Malice*. We provide (tight) bounds on the Price of Malice for an interesting class of malicious Bayesian congestion games. We show that for certain congestion games the advent of malicious types can also be beneficial to the system in the sense that the social cost of the worst case equilibrium decreases. We provide a tight bound on the maximum factor by which this happens.

In [6], we consider a covering problem from a game theoretic prospective. In the *general covering problem*, we are given a universal set of weighted elements E and n collections of subsets of the elements. The task is to choose one subset from each collection such that the total weight of their union is as large as possible. This is a generalization of the well studied *max n-cover* problem, where n subsets have to be selected from a common collection of subsets.

Given a general covering problem, we define *covering games* with n rational players. Each player has her own strategy set, which is a set of subsets of the elements E. For covering an element, the players receive a payoff defined by a non-increasing *utility sharing function*. This function defines the fraction that each covering player receives from the weight of the elements.

We show how to construct a utility sharing function f, such that every Nash equilibrium

covers at least a $(1 - \frac{1}{e})$-fraction of the weight of the optimum solution, i.e. the price of anarchy for f is $(1 - \frac{1}{e})$. We also show that this is best possible, even for very restricted settings. For an important subclass of the covering games, we provide a family of utility sharing functions – with price of anarchy arbitrarily close to $(1 - \frac{1}{e})$ – where any sequence of unilateral improving steps of the players is polynomially bounded. So, a pure Nash equilibrium can be computed in polynomial time. This gives rise to a family of polynomial-time approximation algorithms with approximation ratio arbitrarily close to $(1 - \frac{1}{e})$. this approximation ratio is essentially the best possible (Martin Gairing).

Randomized Load Balancing: In the standard abstraction of load balancing in networks, processors are modeled as the vertices of a graph and connections between them as edges. Each process has an intial collection of unit-size jobs which we call tokens.. Tokens are routed through the network by transmitting them along the edges according to some local rule. The quality of such a network is measured by the discrepancy, which is the maximum difference between the numbers of tokens at any two vertices after the balancing operations have ended.

We consider and analyze a new algorithm for balancing indivisible tokens on a distributed network with n processors. In every time step paired processors balance their load as evenly as possible. In case the number of involved tokens is odd, the location of the excess token is chosen randomly.

We have proven that in comparison to the corresponding model of Rabani, Sinclair and Wanka [13] with arbitrary roundings, the randomization yields an improvement of roughly a square root of the achieved discrepancy in $O(\max(\log n, (\log \log n)^3))$ rounds. This is optimal up to $\log \log n$ factors while the best previous algorithms in this setting either require $\Omega(\log^2 n)$ time or can only achieve a logarithmic discrepancy. Our new result also demonstrates that with randomized rounding the difference between discrete and comntinuous load balancing vanishes. (Tobias Friedrich and Thomas Sauerwald).

Computational Learning Theory: Traditional approaches to concept drift detection in data streams are often based on statistics that capture a fixed set of characteristics of the underlying data stream. This makes those approaches perform worse than necessary, if the change in the distribution affects the statistic only to a small degree. To overcome this problem, we extended results from computational learning theory to design adaptive statistics that automatically adjust to the relevant properties of the data at hand. Based on these results we designed three novel drift detection methods, which were shown to outperform the multivariate Wald-Wolfowitz test in experiments. A paper about this research was accepted at the 2009 SIAM International Conference on Data Mining [4].

Asecond topic deals with capacity control for partially ordered feature sets with linear classifiers. Such feature sets appear naturally in many classification settings with structured input instances, for example, when the data instances are graphs and a feature tests whether a specific substructure occurs in the instance. Since such features are partially ordered according to a "is substructure of" relation, the information in such datasets is stored in an intrinsically redundant form. We were able to prove that the capacity of the hypothesis class of linear classifiers does not decrease in general for partially ordered feature

sets. However, if the data generating distribution assigns lower probabilities to instances in the tail of the hierarchy induced by the partial order, the capacity of the hypothesis class can be bounded by a smaller upper bound. This theoretical result indicates that underfitting rather than overfitting is the more prominent capacity control problem in this setting. A paper about this work was submitted to the 2009 International Conference on Machine Learning. (Ulrich Rueckert)

The Testing Problem for Digital Microfluidics Biochips: Microfluidic biochips enable the precise manipulation of nanoliter volumes of biological fluids and chemical reagents, due to their high speed and sensitivity and very low manufacturing and operating costs. Microfluidic biochips are expected to significantly impact the standard laboratory procedures by performing a broad range of biochemical analyses in genomics, proteomics, clinical diagnostics, environmental monitoring and bio-defense directly on the chip. While early generations of microfluidic biochips rely on continuous fluid flow through permanently etched capillaries, recently, a new generation of biochips has emerged that employ a digitalization of the flow; the fluids are manipulated in discrete droplets using a regular array of electrodes by controlling the voltages applied to two neighbor electrodes. This new architecture provides increased flexibility since many different biochemical analyses can be performed on the same biochip by simply programming the pattern in which electrode voltages are applied. Our work focuses on detecting defective electrodes in such digital biochips; this is done by routing droplets along paths covering all electrodes on the chip and checking their arrival at a sink. Using multiple test droplets can reduce test time, but requires coordination to avoid droplet interference. We introduce new formulations for the problem and give exact and heuristic algorithms together with theoretical bounds showing the efficiency of our algorithm.(Bogdan Pasaniuc)

References

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

[2] 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.

[3] C. Daskalakis, A. G. Dimakis, R. M. Karp, and M. J. Wainwright (2008) "Probabilistic Analysis of Linear Programming Decoding," *IEEE Transactions on Information Theory*, Vol. 54, Issue 8, pp. 3565-3578 August 2008.

[4] A. Dries and U. Rückert (2009) "Adaptive Concept Drift Detection," Proceedings of SIAM International Conference on Data Mining (SDM) Sparks, Nevada, April 2009, to appear.

[5] M. Gairing (2008) "Malicious Bayesian Congestion Games," Proceedings of the 6th Workshop on Approximation and Online Algorithms (WAOA'08), pp. 119-132 Universität Karlsruhe, Germany, September 2008.

[6] M. Gairing (2009) "Covering Games: Approximation Through Non-Cooperation," submitted.

[7] M. Gairing, T. Lücking, M. Mavronicolas, B. Monien, and M. Rode (2008) "Nash Equilibria in Discrete Routing Games with Convex Latency Functions," *Journal of Computer and System Sciences*, Vol. 74, Issue 7, pp. 1199-1225 November 2008.

[8] M. Gairing, B. Monien, and K. Tiemann (2008) "Selfish Routing with Incomplete Information," *Theory of Computing Systems*, Vol. 42, Issue 1, pp. 91-130 January 2008.

[9] V. P. Gopalacharyulu, V. R. Velagapudi, E. Lindfors, E. Halperin, and M. Oresic (2009) "Dynamic Network Topology Changes in Functional Modules Predict Responses to Oxidative Stress in Yeast," *Molecular BioSystems*, Vol. 5, Issue 3, pp. 276-287 2009.

[10] R. M. Karp (2008) "George Dantzig's Impact on the Theory of Computation," *Discrete Optimization*, Vol. 5, Issue 2, pp. 174-185 May 2008.

[11] G. Kimmel, R. M. Karp, M. I. Jordan, and E. Halperin (2008) "Association Mapping and Significance Estimation via the Coalescent," *The American Journal of Human Genetics*, Vol . 83, Issue 6, pp. 675-683 November 2008.

[12] H. Lin, C. Amanatidis, R. M. Karp, M. Sideri, and C. H. Papadimitriou (2008) "Linked Decomposition of Networks and the Power of Choice in Polya Urns," Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 993-1002 San Francisco, California, January 2008.

[13] Y. Rabani, A. Sinclair, and R. Wanka (1998) "Local Divergence of Markov Chains and the Analysis of Iterative Load Balancing Schemes," Proceedings of 39th IEEE Symposium on Foundations of Computer Science (FOCS '98), pp. 694-705 Palo Alto, California, November 1998.

[14] S. Sankararaman, G. Kimmel, E. Halperin, and M. I. Jordan (2008) "On the Inference of Ancestries in Admixed Populations," *Genome Research*, Vol 18, Issue 4, pp. 668-675, April 2008. Also presented at 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2008) Singapore, March 2008.

[15] S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin (2008) "Estimating Local Ancestry in Admixed Populations," *The American Journal of Human Genetics*, Vol. 8, Issue 2, pp. 292-303 February 2008.

[16] C. F. Skibola, P. M. Bracci, E. Halperin, L. Conde, D. W. Craig, L. Agana, K. Iyadurai, N. Becker, A. BrookesWilson, J. D. Curry, J. Spinelli, E. A. Holly, J. Riby,

L. Zhang, A. Nieters, M. T. Smith, and K. M. Brown "Genetic Variants at the PSORS1 Locus on 6p21.33 Influence Lymphoma Risk," submitted.

[17] C. F. Skibola, P. M. Bracci, E. Halperin, A. Nieters, A. Hubbard, R. A. Paynter, D. R. Skibola, L. Agana, N. Becker, P. Tressler, M. S. Forrest, S. Sankararaman, L. Conde, E. A. Holly, and M. T. Smith (2008) "Polymorphisms in the Estrogen Receptor 1 and Vitamin C and Matrix Metalloproteinase Gene Families Are Associated with Susceptibility to Lymphoma," *Proceedings of the Library of Science (PLoS)*, Vol. 3, Issue 7 July 2008.

[18] S. Suthram, A. Beyer, R. M. Karp, Y. Eldar, and T. Ideker (2008) "E-QED: An Efficient Method for Interpreting eQTL Associations Using Protein Networks," *Molecular Systems Biology*, Vol. 4 March 2008.

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

4 Artificial Intelligence and its Applications

In 2008, the Artificial Intelligence group made significant progress in both basic and applied projects. One indication of the progress is the completion of five Berkeley EECS PhD dissertations [12, 33, 30, 44, 13]. The basic research of the group continues to be language learning, computational biology, and neural modeling. In 2008, the applied role was expanded to the areas of natural language Processing (NLP) for developing regions, predictive analysis, and multilingual semantic resources.

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

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

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

3. Applications of the groups' research included the following efforts in 2008.

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

   (b) Completion of NLP based smart search project (sponsored by Ask.com).

   (c) A new collaborative project with Hesperian Press (http://www.hesperian.org) on Multilingual Semantic Resource Development for Emerging regions funded by Google and the Rockefeller foundation.

   (d) A new exploratory effort on metaphoric inference funded by an NSF SGER grant.

   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 addition, in 2008 the group continued work on hybrid state models of biological processes. Joseph Makin completed his PhD working with Srini Narayanan on the hybrid State model of the mammalian coagulation pathway Previous work had resulted in the first comprehensive computational simulation of the mammalian coagulation pathway. In 2008, we completed our analysis using methods from nonlinear control theory with hybrid system modeling and verification to perform sensitivity analysis and design controllers for therapeutic interventions in the presence of specific clotting disorders. The resulting model has generated a fair bit of interest from the medical community and part of the work in 2009 would be to directly investigate the clinical potential of this work.

In 2008, the AI group started a new effort on investigating NLP 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 lead efforts in over 100 countries around the world. In 2008, graduate student Matt Gedigian, working with Srini Narayanan and with Hesperian Press developed a novel approach that uses Semantic Wikis 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 or the ICSI AI group in 2009. The project was awarded a Rockefeller Foundation planning grant for 2009.

Detailed accounts of progress in specific projects in 2008 follows.

4.1 The Neural Theory of Language

The NTL project of the AI group works in collaboration with other units on the UCB campus and elsewhere. It combines basic research in several disciplines with applications to natural language processing systems. Basic efforts include studies in the computational, linguistic, neurobiological and cognitive bases for language and thought and continues to yield a variety of theoretical and practical findings. In 2008, we have 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 book is being used in a number of courses at 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 will also serve 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 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 interpretation could be used to compute subtle semantic distinctions in English, interpret Mandarin child-parent dialogs, and predict processing time difficulty in a manner consistent with experimental data. The resulting program pays an central role in the linguistics doctoral thesis of Ellen Dodge, which is essentially complete and the completed CS doctoral work of Eva Mok. A new invited handbook chapter by Feldman, Dodge, and Bryant presents a coherent picture of this development for linguists. Bryant's system is also now in use

at other labs internationally. A major 2008 addition to handle morphology was an undergraduate honors project by Nate Schneider, who is now in graduate school at CMU. A second undergraduate honors project, by Luca Gilardi, adds 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.

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 late in 2008.

Eva Mok completed her dissertation research on a computational model of context-driven early grammar acquisition. 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, Eva's research represents a first step towards a unified computational model in which both the grammatical units and usage statistics are learned simultaneously from naturalistic, contextually-grounded Mandarin Chinese input.

The 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 2008. 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.

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 [39] that the group has been developing.

Ben Bergen (U. Hawaii) continues to cooperate with the group after completing a UCB 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 and Feldman have completed a major invited book chapter showing how NTL helps explain the ancient mystery of how people can learn new concepts. This is also being used in various classes.

In 2008, 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 to be 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. A companion planning grant proposal to the CRI program was submitted to focus on a second (complimentary) set of issues regarding corpus selection, the design on metaphor annotation schemas and community wide participation, feedback, and evaluation of the annotation effort.

In 2008 there was a very significant increase in the use of the group's results in UCB 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. J. Feldman ran an interdisciplinary class in Spring 2008 and several of the research efforts from that class are being incorporated into the project. George Lakoff is incorporating much of the new NTL book into his undergraduate course and it is being widely used elsewhere.

There were also several invited talks and seminars presented by NTL members. Jerry Feldman presented the NTL project at the Cognitive Science conference and the NTL project was the main subject of the UCB Cognitive Science Faculty retreat in 2008. German post doctoral fellow Birte Loenneker-Rodman and Srini Narayanan contributed an invited article on computational models of figurative language [29] to the Cambridge Encylopedia of Psycholinguistics. Srini Narayanan was invited to be a Fellow at the Institute for Advanced Study in Berlin for the academic year 2008-2009. His project at the Institute for Advanced Study is to write a monograph on NTL related research.

4.2 Reinforcement Learning with Multiple Goals

In Reinforcement Learning (RL), an agent interacts with the environment to learn optimal behavior. Most RL techniques are based on a scalar reward, i.e., they aim to optimize an objective that is expressed as a function of a scalar reinforcement. A natural extension to traditional RL techniques is thus the case where there are multiple rewards. In many realistic domains, actions depend on satisfying multiple objectives simultaneously (such as achieving performance while keeping costs low, a robot moving efficiently toward a goal while being close to a recharging station, or a government funding both military and social programs). Learning optimal policies in many real-world domains thus depends on the ability to learn in the presence of multiple rewards. However, the resulting policies depend heavily on the preferences over these rewards, and they may change swiftly as preferences vary. In 2008, working with Srini Narayanan, graduate student Leon Barrett developed an algorithm for the general case of learning all optimal policies under all assignments of

linear priorities for the reward components, and a proof showing the correctness of the algorithm [8].

Our method learns the set of optimal policies for *all* $\overrightarrow{w}$ at the same time. Once the agent has learned all these policies, it can change reward weights at runtime to get a new optimal behavior, without having to do any relearning. For a fixed priority scheme (fixed weight vector $\overrightarrow{w}$) over the multiple reward components, our algorithm results in the standard recurrence for Q-values that is analogous to the equation for the average weighted reward case as in:

$$Q^*_{\overrightarrow{w}}(s,a) = E\left[\overrightarrow{w} \cdot \overrightarrow{r}(s,a) + \gamma \max_{a'} Q^*_{\overrightarrow{w}}(s',a')|s,a\right]$$

In the general case, where we do not know the relative priorities over the reward components, our algorithm exploits the fact that the extrema of the set of Q-values vectors (Q vectors that are maximal for some weight setting) is the same as the *convex hull* of the Q-value vectors. (The convex hull is defined as the smallest convex set that contains all of a set of points. In this case, we mean the points that lie on the boundary of this convex set, which are of course the extreme points–the ones that are maximal in some direction.

Our algorithm extends the single-$\overrightarrow{w}$ case (which is the standard expected discounted reward framework into the following recurrence:

$$\mathring{Q}(s,a) = E\left[\overrightarrow{r}(s,a) + \gamma \texttt{Hull} \bigcup_{a'} \mathring{Q}(s',a')|s,a\right] \quad (1)$$

$\mathring{Q}(s,a)$ is the convex hull of possible Q-value vectors for taking action a at state s. That is, instead of repeatedly backing up maximal expected rewards, we back up the set of expected rewards that are maximal for some $\overrightarrow{w}$. While the expectation over hulls looks awkward, it is the natural equivalent of an expectation of maxima, and it arises for the same reason. We must take an expectation over s', but once in s', we can choose the best action, no matter what our $\overrightarrow{w}$.

4.3 Smart Search

In 2008, the ICSI AI group completed the smart search project with Ask.com. The aim of the project was to explore the use of machine learning and probabilistic modeling techniques to improve the Ask smart search results. This came as a follow-up to a small 2007 pilot project that explored the use a combination of Natural Language Processing (NLP) and machine learning techniques, many of which are already developed and in use at ICSI for general purpose NLP processing, to enhance people search and the presentation of results through the automatic extraction of key information on the person.

In 2008, the ICSI group used machine learning techniques to induce topics and profiles from social network pages. John Bryant joined the ICSI-ASK collaboration team headed by Srini Narayanan as a post doctoral researcher along with graduate students Matt Gedigian and Srini Ramaswami. We worked on a set of problems related to smart search and query disambiguation which has the potential to make a qualitative difference in web search in cases where there are structured answer sources.

In one of the problems, the inputs comprised of an unfilled registration form (that you need to fill to register), and the profile page (in html). The output was the filled-in registration form induced from the profile (what the person would have entered to result in the profile page for the specific social network site). The data (including training data) was provided by ASK for a variety of social network sites. The training data was typically around 30 examples, and the testing data was several orders of magnitude higher. The ICSI group extracted features from the html including content and meta-information (such as html tags) and used a variety of classification techniques (including SVM, Decision trees, Bayesian models) to compute the profile information. We evaluated our results against a gold-standard (built manually) for the different sites. We compared item by item the profiles extracted with the gold standard using a variety of sampling techniques and with specific guidance on extreme and important cases (provided by the sponsor).

Our results were very encouraging ($> 93\%$ accuracy) and the Ask team felt they could extend the general approach to other sorts of wrapper induction problems within their company. We have given them the training and testing code for the classifier and they are extending and evaluating the use of the technique for other problems.

4.4 Probabilistic Models for Pathway Analysis

Our pathway inference research was conducted within the context of a larger project (PAINT) funded by IARPA. Graduate student 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 a 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 2008, The PAINT program built 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 produces 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, monetary expense, 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), we can use the information value metric to choose the optimal probe to help us answer Hypothesis Disambiguation questions.

We evaluated this system in three multi-university demonstrations. 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 2009. The hypothesis disambiguation problem forms a core test case for the ICSI probabilistic modeling framework and is described more in [44].

4.5 Hybrid System Models of Human Blood Clotting

In 2008, Joe Makin and Srini Narayanan continued work on the blood clotting model, and this work has resulted in the first complete model of the mammalian coagulation pathway. Previous work has modeled specific segments and components. The coagulation pathway is complex, and while the structure is mostly known, there is quantitative data only for part of the pathway. The model uses techniques from hybrid systems to capture both quantitative and qualitative knowledge of the pathway. Using hybrid Petri nets as the underlying graphical model, Joe built and simulated both normal and pathological clotting situations. His output variable (Thrombin) matched existing data (from the literature and from the few real time in-vivo studies) for both the cases. Since the data in the literature is sparse, we performed a monte-carlo sampling of the model to do a sensitivity analysis and showed that the most sensitive variables in the model were consistent with the literature and with a previous experimental analysis.

As part of his PhD thesis, Joe Makin then turned to analysis and controller design. Techniques in hybrid system analysis (such as reachability) cannot deal tractably with the complexity of the model. Solving the Hamilton-Jacobi equations for more than five variables is intractable; his model had over 100. Joe came up with clever way to decompose the model into an initial and final hybrid part and a central (thrombin production) part which is described by around 100 nonlinear (multi-affine) ODEs. He then applied tools from non-linear control theory to design controllers for controlling thrombin production.

Joe's work on applying analysis techniques to control this complex biological system exploits insights and techniques from nonlinear control. First, he showed that the complete system is not linearizable using state feedback. This was quite tricky because you have to compute two Jacobians for every Lie derivative for over 100 variables; which generates far too many symbolic variables (due to the chain rule and proliferation of symbolic derivatives over polynomials) to be computationally feasible. Using recent results in multi-affine control, Joe demonstrated that the full state was not controllable using feedback linearization. A side effect of this proof was a new result that asserts that the locally accessible manifold has as an upper bound the number of reactions of the chemical system. This is applicable to all SISO chemical systems of mass-action kinetics in that a necessary (though insufficient) condition for full-state linearization is that there be at least as many reactions as state variables. This result should be useful to show feedback linearizability (or not) of fair variety of systems of chemical kinetics.

While the full state is not controllable, for coagulation, controlling thrombin is sufficient

to control clotting. Using fairly standard techniques from non-linear control, Joe showed that thrombin can indeed be controlled by feedback linearization. He then designed two types of controllers based on heparin (the standard therapeutic input for controlling hypercoagulation) as input. One required no sampling, but was sensitive to the model correctness while the other is more robust but requires sensing the state. A key result of designing the controller was the realization that the system was non-regular in that the relative degree changed with time. Joe made a crucial observation that the singularity occurred only in the initial condition (with thrombin value 0) and so sliding the controller to start later (thrombin > 0) regularized the system kept the relative degree of the system to a constant (2). It turns out that turning on the control input (heparin) upto 60 seconds after the clotting initiation can still be effective in controlling the thrombin output. This could have clinical implications.

In this case, and in most serious computational biology efforts, the model building process involves a great deal of careful culling of knowledge from experimental and clinical results in multiple scientific fields (in this case from blood chemistry, proteomics, protein engineering, bio-sensing, synthetic biology, and from animal model studies). There is the additional complication of parameters being culled from both in-vitro and in-vivo studies where often we don't have a good mapping function to translate one to the other. All this leads to the fundamental question of model validation and clinical relevance. This is an especially hard problem because the clinical data are likely to be quite sparse. While, there is no clear general answer to this question that emerges from this (or any other) effort, our work points to a possible trajectory to address this vexing issue. The basic idea is to combine the model based predictions with a data driven approach in a machine learning framework where model outputs (in this case thrombin) under parameter perturbations (monte-carlo simulations) are clustered into different classes. Existing gold-standard clinical data points provide a strong prior constraint on the number of classes, their centroid, and boundaries between classes. The clusters correspond to the different classes of interest informed by the different types of diseases (in this case hyper and hypo coagulatory) in the clinical data. Once the data are clustered, a classifier (such as a decision tree) can be trained to learn the general mapping from the space of parameters to the clusters (classes). This then allows for calculating lowest cost moves from one class (diseased) to another (normal) which is of clinical relevance. While the results here are preliminary, they show promise and the approach is general to cover all cases where you want to combine information from an imperfect but good model with sparse gold-standard clinical data. However, moving from initial promise to robust results with this approach remains future work.

The model has attracted considerable interest from the medical community and we now have now in our team clinical experts from UCSF (Dr. Mark Shuman, Dr. Tracy Manchiello, and Dr. Patrick Fogarty) examining the model for accuracy and robustness and testing the model predictions on known quantitative data on disease pathologies. We have also been contacted by different Pharmaceutical companies to use the model as part of their development toolkit. We are currently testing and refining the model, including using the model to predict clinical data on various tests (such as PTtime and APTtime). Ultimately, we hope to be able to package and make available the model as a routine diagnostic aid for hematologists.

4.6 A Multiply Annotated American National Corpus

Under a subcontract on NSF grant #0708952 "CRI: CRD: A Richly Annotated Resource for Language Processing and Linguistic Research" (Prof. Nancy Ide of Vassar, PI), the FrameNet group is annotating texts from the American National Corpus (ANC) ([24], `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 is to do the manual annotation on a small portion of the corpus, training automatic semantic role labeling (ASRL) software to do the rest.

In order to make effective use of ASRL, we need the ability to run SRL on a text, view the output, and decide, sentence by sentence, whether to accept the automatic label or to correct (or delete) it. We have implemented software for this purpose, as an enhancement to the FNDesktop annotation tool. It is now possible to read in an XML file (in the same format as the desktop exports it) containing annotations and view it without importing it into the database. The user can then decide whether to import the annotations or not, correcting them when needed. The corrections to the automatic annotation will give a clear indication of where the system needs more training.

We have manually annotated in 42 texts during 2008; 2,052 annotation sets were created this process. Small amounts of additional annotation was done in previously annotated texts, as new LUs were added to the lexicon, but most of the new annotation was on new texts, which fall into two groups: (1) text from travel guides contributed by Berlitz to the ANC (13,418 words); (2) a number of business letters from charities, soliciting contributions (2,308). The annotated texts, along with others from 2007, have been sent to the ANC for inclusion in the first release of the ANC "Multiply annotated subcorpus (MASC), expected in January, 2009. Dr. Baker went to New York City, Aug. 13-14, for one meeting with our partners in the MASC project, and since then has participated by phone and video connection in two other meetings in which the other partners met in New York.

4.7 WordNet-FrameNet Alignment

We have continued work funded under NSF #0705155 on aligning FrameNet with WordNet (WN), the largest machine-readable English lexicon. There are large areas of the lexicon in which the frames evoked by the LUs are not particularly "interesting" or useful; most common nouns are of this type. In these areas, there is no intention of creating a detailed structure of frames; instead we are creating a small number of general frames that will "cover" large numbers of LUs in less detail, and will depend on the WordNet definitions and lexical hierarchy to provide most of the semantics. We have identified NN such domains, and created the appropriate general frames for them.

We also have worked on detailed comparisons of WN and FN sense divisions for NN common words, in part for the WN-FN alignment study and also as a contribution to the ANC MASC project; these words will be annotated by the WN annotators at Vassar

throughout the MASC. 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.

4.8 Development of Word Sketch Engine for Rapid Vanguarding

We are continuing the collaboration with Adam Kilgarriff on a new system for the vanguarding portion of the FrameNet work, as described in our last annual report. This consists of an implementation of a GUI based on Dr. Kilgarriff's Word Sketch Engine [26, 27], funded under NSF #0535297 "Rapid Development of a Frame Semantic Lexicon". We have completed the software development needed to integrate it with the FNDesktop annotation system. We still need to encode the full ANC and preprocess it for producing word sketches; we have already done so for the BNC. Then we plan to measure the time required per LU for the vanguarding step and compare the old and new vanguarding systems. The original grant has ended, but we were able to obtain a one-year, no cost extension, which should be enough to complete the planned comparisons.

4.9 Changes in data release files and formats

The last formal release of the FrameNet data, in 2006, contains a large number of different file formats. There are files for the lexical annotation in a different format than files for the full-text annotation, and each of these is provided in both XML and HTML format. One of our goals in preparing for the next data release is to reduce the number of different file formats. We are doing this by creating a new XML format that contains all of the information needed for display, and packaging with the XML files an XSL file containing Javascript. The combination of the XML and the XSL files means that standard web browsers can view the data as if it were HTML, while having it in XML keeps it easy for programs to read. The new format and the XSL file have to be developed and tested on various browser/ operating system combinations, but the result will be a much simpler set of files to distribute. To date, we have the XML specification and the report generator, and an XSL file that provides most of the desired functionality. We hope to finish work on this and be ready for a new data release early in 2009.

4.10 Pending Proposals

We submitted four grant proposals to NSF between September and December, 2008, all of which are currently pending:

1. In collaboration with Prof. Christiane Fellbaum of Princeton University, we requested a Community Resource Development planning grant to hold a conference to present the results from work on FN-WN alignment so far and to determine the level of support for a full-scale alignment of WordNet and FrameNet.

2. We requested another CRD grant to maintain and improve the FN database, including producing a fully Unicode compliant version of the data, a full set of APIs for various programming languages, and elimination of certain inconsistencies in the

annotated data. Included would be support for a student researcher to monitor and respond to users' questions,

3. We asked for money under the "Robust Intelligence" area of Information and Intelligent Systems to develop a set of web-based games to be used to gather annotation data; the basic idea is that the player would decide whether annotation of a sentence proposed by and ASRL system was correct. By making this a game, with points awarded for accuracy and agreement with other players, we hope to be able to gather a large amount of manually annotated text more rapidly than we could in-house. Similar project for collection of other kinds of data, including labeling of images from the web [1] and gathering sets of associated words [2] have succeeded in gathering large amounts of data very quickly.

4. We also requested another Robust Intelligence grant to set up a website to be used for collaborative development of frame definitions, including FE definitions and lists of LUs. The proposed mechanism is a Semantic Media wiki, in which relations between frames would be mapped to relations between frames, FEs, and LUs, allowing transfer to and from the FN database.

4.11 Development of the FrameNet Database

During 2008, the number of frames increased by 65 to 949; the number of lexical units (LUs) in the FrameNet database increased by 446 to 11,548. Of these, 277 were new LUs in existing frames; examples of these are shown in Table 1. The other 170 LUs were in the 65 new frames; examples of these new frames and LUs are shown in Table 2.

Architectural part	*foundation.n*
Attack	*infiltrate.v, infiltration.n*
Biological area	*rainforest.n*
Buildings	*airport.n, pub.n*
Building subparts	*laundry room.a*
Calendric unit	*century.n*
Cause motion	*jerk.v, knock.v, launch.v, propel.v, punt.v, roll.v, stick.v*
Collaboration	*in league.a, jointly.adv, together.adv, work together.v*
Compliance	*disobey.v, in accordance.a*
Desirability	*idyllic.a, miserable.a, popular.a, unfortunate.a*
Emotion directed	*crushed.a, demolished.a, outrage.n*
Execute plan	*force ((into force)).n, institute.v*
Experiencer obj	*appeal.v, arouse.v, crush.v, demolish.v, destroy.v, devastate.v, engage.v, fulfill.v, grate.v, harass.v, stagger.v, worry.v*
Fame	*epic.a, notoriety.n, notorious.a, reputation.n*
Fields	*history.n, industrial.a*
Getting vehicle underway	*lift (off).v, take (off).v*
Giving	*advance.v, charity.n, leave.v, volunteer.v, will.v*
Hiding objects	*camouflage.v*
Leadership	*administer.v, administration.n, crown prince.n, executive.n, lawmaker.n, legislator.n, major general.a, representative.n, secretary.n, spearhead.v*
Locale by use	*campus.n, downtown.n, zoo.n*
Locative relation	*mainland.n, northeast.prep, southeast.prep, to.prep*
Organization	*committee.n, delegation.n, group.n*
Origin	*Assyrian.a, Colombian.a, Cuban.a, Egyptian.a, Jordanian.a, ottoman.a, roman.a, Saudi.a, swiss.a, Turkish.a*
People by vocation	*agent.n, architect.n, businessperson.n, journalist.n, judge.n, mechanic.n, mole.n, oilman.n, reporter.n, scholar.n, speculator.n, spy.n, trader.n, veterinarian.n*
Quantity	*a little.n, both.a, smattering.n*
Relational natural features	*shoreline.n*
Relational quantity	*little.art*
Removing	*cut.v, rip.v, scrape.v, tear.v*
Render nonfunctional	*incapacitate.v*
Self motion	*advance.v, file.v, take to the air.v*
Sending	*barge.v, express.v, wire.v*
Sign	*mark.v, usher in.v*
Speed	*quickly.adv, rapid.a, rapidly.adv, speedily.adv*
Temporal collocation	*ancient.a, around.prep, at present.adv, at the time.adv, during.prep, early.a, newly.adv, prehistoric.a*
Text	*epic.a, issue.n, paper (article).n, paper (newspaper).n*

Table 1: Examples of New Lexical Units in Existing Frames Created in 2008

Abandonment	*abandoned.a, abandonment.n, abandon.v, leave.v*
Agriculture	*cultivate.v, farming.n, farm.v*
Annoyance	*annoyed.a, frustrated.a, irritated.a*
Attending	*attendance.n, attend.v, go (to).v*
Beat opponent	*beat.v, defeat.n, demolish.v, prevail.v, rout.v, trounce.v, ...*
Becoming dry	*dehydrate.v, dry up.v, dry.v, desiccate.v*
Being strong	*impregnable.a, strong.a*
Breadth of coverage	*broad.a, narrow.a, specific.a, sweeping.a, targeted.a*
Cause emotion	*affront.n, concern.v, insult.n, insult.v, offend.v, ...*
Ceasing to be	*disappear.v, dissolve.v, fade.v, go away.v, vanish.v, ...*
Confronting problem	*confront.v, face.v*
Conquering	*capture.v, conquer.v, fall.v, takeover.n, take.v*
Degree	*a little.adv, far.adv, heavily.adv, so.adv, very.adv, ...*
Duration relation	*last.v, persist.v*
Dying	*dying.a, moribund.a*
Emotions of mental activity	*delight (in).v, delight.n, drudgery.n, enjoyment.n, enjoy.v, luxuriate.v, pleasure.n, relish.v, savor.v*
Emotions success or failure	*dissatisfied.a, fulfilled.a, satisfied.a, unfulfilled.a*
Fear	*afraid.a, dread.n, fear.n, freaked.a, scared.a, terror.n, ...*
Food gathering	*bring in.v, gather.v, harvest.v, pick.v*
Growing food	*grow.v, raise.v*
Hunting	*fish.v, hunt.v, seal.v*
Invading	*invade.v, invasion.n, overrun.v*
Just found out	*shocked.a, shock.n, surprised.a, surprise.n*
Make possible to do	*allow.v, let.v, permit.v*
Mental stimulus exp focus	*absorbed.a, engrossed.a, enthralled.a, fascinated.a, infatuated.a, interested.a, smitten.a, wrapped up (in).a, ...*
Mental stimulus stimulus focus	*absorbing.a, captivating.a, engrossing.a, enthralling.a, fascinating.a, interesting.a*
Others situation as stimulus	*compassion.n, empathize.v, feel (for).v, pity.v, sympathize.v*
Out of existence	*a thing of the past.n, gone.a, history.n, toast.n*
Planting	*plant.v, sow.v*
Prevent from having	*deny.v, deprive.v, starve.v*
Punctual perception	*get a(n) eyeful.v, glimpse.n, glimpse.v*
Relating concepts	*associate.v, connect.v, link.v, relate.v, tie.v*
Repel	*repel.v, resist.v*
Ruling legally	*rule.v*
Soaking up	*absorb.v, soak up.v, sponge.v*
Supporting	*bolster.v, buttress.v, support.v*

Table 2: Examples of New Frames (and Lexical Units) Created in 2008

4.12 FN-related Publications by FN Staff and Alumni

- C. Fellbaum and C. Baker (2008) "Can WordNet and FrameNet Be Made 'Interoperable'?" Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGLR 2008), pp. 67-74 Hong Kong, China, January 2008.

- C. J. Fillmore (2008) "Border Conflicts: FrameNet Meets Construction Grammar in Bernal," Proceedings of the XIII EURALEX International Congress, pp. 49-68 Barcelona, Spain, July 2008.

- C. J. Fillmore (2008) Review Article: *A Valency Dictionary of English*, T. Herbst et al, eds., *International Journal of Lexicography* October 2008.

- S. Kuboya, K. Ohara, and H. Saito (2008) "The Annotation Environment for Japanese FrameNet," Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing (ANLP), pp. 639-642 Tokyo, Japan, March 2008.

- B. Lönneker-Rodman, C. Baker and J. Hong (2008) "The New FrameNet Desktop: A Usage Scenario for Slovenian in Webster," Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGLR 2008), pp. 147-154 Hong Kong, China, January 2008.

- K. H. Ohara (2008) "Semantics of Lexical Items and Constructions in Japanese FrameNet: A Contrastive Analysis of a Parallel Corpus," (in Japanese), Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing (ANLP), pp. 857-860 Tokyo, Japan, March 2008.

- K. H. Ohara (2008) "Lexicon, Grammar, and Multilinguality in the Japanese FrameNet," Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pp. 3264-3268 Marrakech, Morroco, May 2008.

- H. Saito, S. Kuboya, T. Sone, H. Tagami, and K. Ohara (2008) "The Japanese FrameNet Software Tools," Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pp. 1835-1838 Marrakech, Morroco, May 2008.

- H. Sato (2008) "New Functions of FrameSQL for Multilingual FrameNets," Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), pp. 758-762 Marrakech, Morroco, May 2008.

- T. Schmidt (2008) "The Kicktionary: Combining Corpus Linguistics and Lexical Semantics for a Multilingual Football Dictionary," *The Linguistics of Football*, E. Lavric, G. Pisek, A. Skinner, and W. Stadler, eds., pp. 11-23 Gunter Narr, 2008.

- T. Sone, K. Ohara, and H. Saito (2008) "Automatic Semantic Role Labeling in Japanese FrameNet Using Hierarchical Information of Lexicon," (in Japanese), Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing (ANLP), pp. 635-638 Tokyo, Japan, March 2008.

References

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

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

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

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

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

[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. Mac Dermed (2008) "A (Somewhat) New Solution to the Variable Binding Problem," *Neural Computation*, Vol. 20, Issue 9, pp. 2361-2378 September 2008.

[8] L. Barrett and S. Narayanan (2008) "Reinforcement Learning with Multiple Criteria," Proceedings of International Conference in Machine Learning (ICML) Helsinki, Finland, July 2008.

[9] B. Bergen and J. Feldman (2008) "It's the Body, Stupid: Concept Learning According to Cognitive Science," /em Handbook of Embodied Cognitive Science, P. Calvo and R. Gomila, eds. Elsevier, 2008.

[10] 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, No. 5, pp. 733-764 September 2007.

[11] 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.

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

[13] N. Chang (2008) "Constructing grammar: A Computational Model of the Emergence of Early Constructions," UC Berkeley EECS dissertation 2008.

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

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

[16] Richard S. Cook, Paul Kay, and Terry Regier (2005) "The World Color Survey Database: History and Use," *Handbook of Categorisation in the Cognitive Sciences*, H. Cohen and C. Lefebvre, eds. Elsevier, 2005.

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

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

[19] 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 Press, May 2004.

[20] C. J. Fillmore (1976) "Frame Semantics and the Nature of Language," *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Vol. 280, Issue 1, pp. 20-32 October 1976.

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

[22] C. Fillmore, J. Ruppenhofer, and C. Baker (2004) "FrameNet and Representing the Link between Semantic and Syntactic Relations," *Computational Linguistics and Beyond*, C.-R. Huang and W. Lenders, eds., pp. 19-62 Academia Sinica Press, 2004.

[23] J. Hobbs and S. Narayanan (2003) "Spatial Representation and Reasoning," *Encyclopedia of Cognitive Science*, L. Nadel, ed. Nature Publishing Group, 2003.

[24] N. Ide, R. Reppen, and K. Suderma (2002) "The American National Corpus: More Than the Web Can Provide," Proceedings of the Third Language Resources and Evaluation Conference (LREC), pp. 839-844 Canary Islands, Spain, May 2002.

[25] P. Kay (2005) "Color Categories Are Not Arbitrary," *Cross Cultural Research*, Vol. 39, No. 1, pp. 72-78 February 2005.

[26] A. Kilgarriff and D. Tugwell (2001) "Word Sketch: Extraction and Display of Significant Collocations for Lexicography," Proceedings of COLLOCATION: Computational Extraction, Analysis and Exploitation Workshop at 39th Annual Meeting of the ACL, pp. 32–38 Toulouse, France, July 2001.

[27] A. Kilgarriff and D. Tugwell (2001) WASP-Bench: An MT Lexicographers'Wworkstation Supporting State-of-the-Art Lexical Disambiguation," Proceedings of Machine Translation Summit VII (MT Summit VII), pp. 187-190 Santiago de Compostela, Spain, September 2001.

[28] B. Lönneker-Rodman "Beyond Syntactic Valence: FrameNet Markup of Example Sentences in a Slovenian-German Online Dictionary," Proceedings of Computer Treatment of Slavic and East European Languages: The Fourth International Seminar (Slovko 2007), pp. 152-164 Bratislava, Slovakia, October 2007.

[29] B. Lönneker-Rodman and S. Narayanan (2008) "Computational Models of Figurative Language," /em Cambridge Encyclopedia of Psycholinguistics (to appear).

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

[31] J. Makin and S. Narayanan (2008) "Real Time Control of Human Blood Clotting," *Proceedings of the Library of Science (PLoS), Computational Biology* 2008, submitted.

[32] J. Makin and S. Narayanan (2008) "A Hybrid System Model of Human Blood Clotting," *Proceedings of the Library of Science (PLoS), Computational Biology* 2009, to appear. Also ICSI Technical Report TR-08-002, February 2008.

[33] E. Mok (2008) "Contextual Bootstrapping for Grammar Learning," UC Berkeley EECS dissertation 2008.

[34] 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.

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

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

[37] S. Narayanan (2008) "The Thermal Qualities of Substance," Proceedings of the International Cognitive Science Conference (CogSci), pp. 2290-2295 Washington, D.C., July 2008.

[38] S. Narayanan and S. Harabagiu (2004) "Question Answering based on Semantic Structures," Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Article 693, Geneva, Switzerland, August 2004.

[39] S. Narayanan, K. Sievers, and S. Maiorano (2007) "OCCAM: Ontology-Based Computational Contextual Analysis and Modeling," Lecture Notes in Computer Science, Vol. 4635, /em Modeling and Using Context, pp. 356-368 Springer Berlin: Heidelberg, 2007.

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

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

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

[43] J. Scheffczyk, C. F. Baker, and S. Narayanan (2008) "Ontology-based Reasoning about Lexical Resources," *Ontology and Lexical Resources in Natural Language Processing* Cambridge University Press, 2008.

[44] S. Sinha (2008) "Answering Questions about Complex Events," UC Berkeley EECS dissertation 2008.

[45] S. Sinha and S. Narayanan (2005) "Model-Based Answer Selection," Proceedings of Workshop on Textual Inference for Question Answering, pp. 24-31 Pittsburgh, Pennsylvania, July 2005.

5 Speech Processing

2008's Speech efforts were headed by research staff members Dilek Hakkani-Tür, Adam Janin, Gerald Friedland, 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 working with Hynek Hermansky and Hervé Bourlard of IDIAP, and Mari Ostendorf of the University of Washington. We benefited greatly by our close collaboration with SRI, not only via the efforts of Shriberg and Stolcke, about 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. As always, major contributions were also made by our team of students, research associates, postdoctoral Fellows, and international visitors. (See http://www.icsi.berkeley.edu/groups/speech/members.html for 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 described is by no means exhaustive, but it should provide a useful overview of the major speech activities for the year.

5.1 Speech Recognition

ASR for GALE: In 2008, we completed the attempt to apply the Gabor filter approaches that had successfully been used for small vocabulary tasks (see below) to the GALE Mandarin broadcast data. As we had found with other novel discriminative features, we were able to significantly reduce error rates (from 25.5% to 22.1%) by incorporating such a stream. We are still working on ways to usefully include such features in an evaluation system that already includes a number of other streams. We also worked on adapting MLP-based features to reduce word error rate on Mandarin Broadcast Conversation data, which was quite high in comparison with the error rate on Mandarin Broadcast News. Thus far we have been unable to improve performance (either using retrained or adapted MLPs or by retraining the Gaussian mixtures for the HMM) beyond what one could achieve by having an equivalent amount of training data for the two genre. Since these tests our colleagues at SRI have shown that that they could achieve improvements from such an approach for the Gaussian mixture components using a finer granularity of genre classification; we had used the "show" as the classified element. We hope use an utterance-based genre classification for MLP adaptation in the coming year. In the meanwhile, we applied our multistream nonlinear discriminant transformations to the new DARPA evaluation data for 2008.

Cortically-inspired features for speech recognition: We continued our work in the use of multiple streams comprising Gabor-filter-based spectro-temporal receptive fields, transformed via MLPs into discriminative feature streams. As noted above, some experiments were run on a difficulty large vocabulary task, GALE Mandarin broadcast data that was used for a machine translation evaluation. However, most of our experiments continue to be run on the OGI Numbers 95 corpus, which is more useful for running a

large number of tests. Experiments were conducted with both the original Numbers 95 corpus and with a noise-added version. The noises used for the latter were selected from the RSG-10 collection; they include speech babble, factory floor noise, car interior noise, F-16 cockpit noise, Destroyer operations room noise, and Leopard vehicle noise [20]. Five signal-to-noise ratios, ranging from 0dB to 20dB, along with the no-noise-added condition were used. Experiments were conducted with a four-stream system, an 8 stream system, a 16 stream system, and a 28 stream system that was composed of all of the streams in the other 3 systems. In each case, the streams incorporated a range of temporal and spectral modulations and were calculated for each critical band; also, the resulting features were always appended to an MFCC-based feature vector (MFCCs plus first and second time derivatives) for improved performance. For the noise-added cases, the best performance was achieved using the 28 stream system. In this case, the stream weights were calculated using an MLP that had been trained with one-hot labels for the best stream in each frame (for phone class per frame), and with MFCC and scaled inverse entropy as inputs. The resulting word error rate in the noisy condition was 8.1% in contrast to the MFCC baseline which only achieved a 15.3% word error rate. This was nearly a factor of two improvement. Results were more modest for the "clean" case, and were actually the best for the four-stream system, reducing the word error rate from 3.0% to 2.0%. We are still exploring ways to teach a many-stream system to choose the smaller number of streams for cases that do not benefit from the larger number being used.

Ensemble feature selection: In 2008 we explored automated feature selection for multi-stream ASR, using a hill-climbing algorithm that changes one feature at a time if the change improves a performance score. This is a break from prior published work, which has selected features for multi-stream systems in blocks (e.g., adding in a PLP stream given MFCCs for the first stream). For most of the experiments, there were three streams, which drew from a pool incorporating MFCC, PLP, and MSG features (where the latter are features developed at ICSI by Brian Kingsbury in the late 1990's). In some experiments the three streams were initialized to be these three feature types, while in others a random initialization was used. For both clean and noisy data sets (using the OGI Numbers corpus and noise additions as described above), hill climbing almost always improved performance on held out data compared to multi-stream baselines, though it did not improve over the best single-stream baseline that incorporated all the features. For noisy data, the improvements were achieved even for noise types that were not seen during the process. In tests involving mismatch between clean and noisy data, however, ensemble feature selection outperformed both singlestream and multi-stream baselines when a scoring formula due to Opitz was used. We found this scoring formula, which blends single-classifier accuracy and ensemble diversity, outperformed ensemble accuracy as a performance score for guiding the hill climbing. Our noisy version of the Numbers corpus, our multi-layerperceptron-based Numbers ASR system, and our hill climbing scripts are available online. A complete description of the research can be found in [25].

Parallelization: Late 2007 saw the founding of ParLab, a project led by the University of California Berkeley and funded in part by Microsoft and Intel with the ambitious goal

of making it easy to program parallel computers for the 90% of real-world programmers who are not trained in parallel programming. Research in ParLab runs the gamut from hardware and language design all the way up to applications. ICSI is a member of ParLab with responsibility for the Speech Recognition application (see also the Architecture Group's section of this report for other ICSI involvement). Initial experiments on parallelizing Speech Recognition have been performed on an NVIDIA cards. Although NVIDIA is primarily known for it's graphics cards, they provide an environment for general purpose programming as well. On an NVIDIA GTX 280 (containing 240 simple cores), neural network training is 2.5 times faster than our current highest-end training system consisting of an 8 core 2.2 GHz AMD Opteron. Currently in progress is porting the "decoder" component of a speech recognition engine to the NVIDIA platform using two different approaches. The first approach involves parallelizing only over the Gaussian mixture computation. The second approach runs the entire search on the parallel hardware. Although the latter method has potential for much higher performance, it suffers from a shortage of memory, leading to the use of simpler, less accurate models.

5.2 Speaker Diarization

After working on efficient speaker diarization in 2007, this year's efforts focused on reducing the average latency in diarization (where latency means the maximum amount of input data the algorithm can process before emitting the final output for that sample) and to find ways to incorporate video data to improve diarization.

Especially in the projects AMIDA and IM2, where the diarization is used as an upstream application for remote collaboration, a speaker diarization method that can work during the meeting is desirable. Therefore we continued our work on low-latency or online diarization approaches. We developed a two-step online approach where speaker models are trained in advance, either by manually providing a recorded segment of a speaker or by using the output of an offline diarization system. The obtained models are then used for online diarization. Having a completly unsupervised system has several advantages, e.g. no training data is required a-priori. This means, speakers can be separated even if they are not known to the system. However, because there is no training involved, there is no way that the output of such a system can contain real names. Traditionally, the segments are labeled "speaker_0", "speaker_1", etc... This is not typically helpful for applications.

Among other things, we experimented with pre-trained speaker models in order to be able to map speaker clusters to real names. We found that with only 50 seconds of speech per speaker the system is able to perform the diarization task on a subset of the AMI meetings, a total length of more than 9.5 hours, with a Diarization Error Rate better than the offline system. When the pre-trained models are also labeled with real names, the assignment of speaker clusters to real names becomes trivial. The system was presented at the AMI Knowledge and Know How day at MLMI 2008 and demonstrated at ACM Multimedia 2008.

Under the assumption that a person that is speaking has more visual activity than a person who is not speaking, we experimented with automatically associating a video of the corresponding speaker rather than labeling the speaker regions with numbers. However, extracting faces and correlating mouth movements is often difficult in natural conversation

settings, especially in a single-camera view. In a series of experiments, we tested the feasibility of using more gross body motions as a basis for audio-visual synchrony. Rather than explicitly identifying body parts or even kinematic models, we opted for a simpler approach and tried to observe approximations of local body movement. We investigated 20 different methods of audio-visual synchrony based on findings in psychology literature on human body movements in discourse. Using a mixed modality set-up with a single audio source and 2–4 cameras on a 4.5 hours subset of the AMI meeting corpus, the task was to find the current speaker in a video given the observed activity of all speakers and correlate the results with an online version of the speaker diarization engine. A total of 720 experiments using different error metrics was performed. Initial findings of the study have been published at the European Conference on Computer Vision (ECCV2008) [31] and the final, 39-page study has been invited for submission to Elsevier's Computer Vision and Image Understanding journal (CVIU) [33].

In summary, the results have provided evidence that body motion (which means head and hand activity) is strongly correlated with speech and can be used to locate the current speaker in a meeting. The best algorithm/feature combination resulted in an average accuracy (over 12 real-world meetings containing 4 speakers) of 41.89 % DER in the 4-camera case and 42.38 % DER in the 2-camera case. This includes a speech/non-speech error of 12.20 % (overlapping speech regions where tagged as non-speech as well). We found that audio-visual correlation using body motions can help improve the noise-robustness of diarization: When varying the SNR of the input audio signal, most correlation methods showed no degradation in the score. Also, it was encouraging for us to see that the number of cameras used had little effect.

After studying the audio-visual correlation we started creating a system to actually improve the diarization accuracy both on the four close-up views and on the single camera view. Using a combination of a skin-color detector and MPEG-4 compressed-domain motion vectors, we obtain the overall activity in one cell of an evenly partitioned video image.

Adding these video features from the single-camera view resulted in a 14 % relative improvement over the audio-only baseline, and adding the video features from the 4-camera view resulted in a 25 % relative improvement. The results were submitted and accepted at IEEE ICASSP. The experiment showed that it is possible to treat audio and video data as part of the same optimization problem to help improve the diarization task. However, the combined training of audio and video models allows for more than just improved accuracy. In a second pass over the video, we obtain the location of the current speaker's skin patches by "inverting" the visual models. Given the current speaker (as output in the previous step), we assume activity for each individual skin patch in a frame to calculate the likelihoods of belonging to a speaker using the visual Gaussian Mixture Models for each of the patches. Those skin patches that belong to the camera (four-camera case) or the part of the frame (single-camera) that is most likely to be active given the speaker determined by the diarization in the first step are tagged (for visualization or further processing). This enables a completely unsupervised diarization and tracking of the speakers in meetings.

The results and their explanation will be refined in the next year.

Automatic Detection of Overlapping Speech: We continued our work on overlap detection. We found that the combination of MFCCs, RMS energy, LPC residual energy, and posterior entropy yields a detection precision that is able to improve Diarization Error Rate. In addition, we found that feature warping techniques improve overlap detection.

Having identified regions of overlapped speech, this information can then be used to modify segment and label information output by the diarization system. The procedure is as follows. In an overlapped segment, the frame-level speaker posteriors are summed over the frames of the segment to obtain a single score for each speaker. Typically the diarization system will have assigned the segment to the speaker with the highest score, in which case the speaker with the second highest score is chosen as the other speaker. In the event that the system has chosen another speaker, then this highest scoring speaker is selected as the additional speaker. Note that this procedure limits the number of possible overlapping speakers to two, but that two-speaker overlap typically comprises 80 % or more of the instances of overlapped speech.

In the end, we obtained a relative improvement of 6.8 %. The results was presented at Interspeech 2008 [4] and resulted in a PhD thesis [2].

Automatic Segmentation of Laughter: We continued our work on laughter detection. Specifically, we developed and then used a median filtered MLP (MF MLP) laughter detection system to compute the posterior probability of laughter for each frame (10 ms). This posterior was then used as a feature for the overlap detection system in combination with MFCC features. The overlap detection system was then used in the ICSI Diarization engine. We trained the laughter detection system using only the original training set used for laughter before and only the training set used by the overlap system. We also trained using two classes (laughter and other vocalized sounds), where we did not train on non-speech segments, and using three classes (laughter, other vocalized sounds, and silence). After including the posterior probability of laughter as a feature in the overlap detection system (which was then used in the ICSI diarization system), the baseline diarization system improved by 6 % relative. After looking further at the overlap detection dataset, we found that of all the overlap time 25 % contained laughter. Furthermore, 75 % of laughter time was overlapped.

The work resulted in a master thesis [34]

5.3 Multimodal Analysis Framework

Under the sponsorship of a Silicon-Valley-based startup company, Appscio, the speech group was involved 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 framweork aims to provide a unified approach that standardizes the entire process of development, deployment, and integration of components. Multimedia is often regarded a synonym for audio and video. In the philosophy of the Appscio framework, however, multimedia includes not only video, audio, and related meta data, but any other source of sensory information that can help accomplish a certain task.

A first beta version of the framework was released and described in [9].

5.4 Punctuation Insertion

Efforts on punctuation insertion to speech recognition output at ICSI involve two tasks: *sentence segmentation* and *comma detection*. *Sentence segmentation* from speech is part of a process that aims at enriching the unstructured stream of words output by standard speech recognizers. Its role is to find the sentence units in this stream of words. It is of particular importance for speech related applications, as most of the further processing steps, such as parsing, machine translation, information extraction, assume and benefit from the presence of sentence boundaries [6, 41]. In 2008, we continued working on multilingual (English, Mandarin and Turkish – which is new this year) sentence segmentation in the framework of the DARPA GALE and CALO projects. Our technical contributions include using long distance models that rely on syntax and check grammaticality of the formed sentences [7] and use of morphological features for the sentence segmentation of morphologically rich languages, in addition to lexical and prosodic information [28].

For using syntax in sentence segmentation, we leverage global syntactic information from a syntactic parser, which is better able to capture long distance dependencies. While other previous work has also included syntactic features, ours is the first to do so in a tractable, lattice-based way, which is crucial for scaling up to long-sentence contexts. Specifically, an initial hypothesis lattice is constructed using a local, word boundary classification model. Candidate sentences are then assigned syntactic language model scores. These global syntactic scores are combined with local low-level scores in a log-linear model. The resulting system significantly outperforms the most popular long-span model for sentence segmentation (the hidden event language model) on both reference text and automatic speech recognizer output from news broadcasts.

In order to use morphological information in sentence segmentation, we introduced a new set of morphological features, extracted from words and their morphological analysis. We also extended the established method of hidden event language modeling to factored hidden event language modeling and combined it with discriminative classification techniques, boosting and conditional random fields (CRFs). We experimented with Turkish broadcast news data and use of morphological features resulted in significant improvements in classification accuracy of all methods used.

Comma detection aims to include commas and similar punctuation marks (such as caesuras in Mandarin) into the raw word sequence output from a speech recognizer. We studied comma detection for English and Mandarin, in the framework of the DARPA GALE project. Restoring commas have been shown to help part of speech tagging [30] and information extraction [6, 40]. It can also help in finding appositions, for instance, that would greatly enhance coreference resolution and more generally further language processing tasks, such as question answering and summarization. For comma detection, similar to sentence segmentation, we exploit several lexical and prosodic features. Furthermore, we derive syntactic features at the boundary between consecutive words. We then apply conditional random fields and factored hidden event language models in order to restore commas in news broadcasts. The new syntactic features bring 7% improvement over lexical, part-of-speech tag and prosodic features [8].

5.5 Information Distillation and Summarization

Information distillation aims to extract the most useful pieces of information related to a given query from massive, possibly multilingual, audio and textual document sources. This year, the distillation task was very focused and included extraction of answers to 5 W-questions (who, what, when, where, and why) that can be asked to sentences from the multilingual, audio and textual documents.

Our distillation approach is based on syntactic parses annotated with function tags, such as subject and object, as well as semantic role labels annotated according to PropBank. Syntactic and semantic parsers trained according to each genre are used to parse sentences, and these parses are then used to extract answers to the 5 W-questions. The output from syntactic and semantic parsers may differ. In order to benefit from both sources, our distillation approach combines answers from different sources using majority voting and hidden Markov models trained to recognize answers to these questions.

In addition to answering the 5 W-questions, our team also worked on multi-document summarization, and participated in the NIST Text Analysis Conference (TAC) update summarization task evaluations for the first time this year. The ICSI multi-document system is an extractive summarization system that tries to pick sentences that have the highest information content from the original documents [26]. It relies on a general framework that casts summarization as a global optimization problem with an integer linear programming solution. Our primary submission to the NIST TAC evaluations, a simple sentence extractor with an n-gram document frequency heuristic, gives results at least as good as any reported on the first part of the update summarization task. Our secondary submission also considers compressed sentence alternatives, and achieves high ROUGE [39] scores but lower manual evaluation scores.

5.6 Spoken Language Processing in Meetings

Our work on spoken language processing in meetings was focused on two research problems: *meeting summarization* and *argument diagramming of meetings*. Previous work on meeting summarization mainly focused on applying unsupervised methods, such as maximum marginal relevance (MMR) [5], and studied term weighting methods for meeting summarization [43, 54, among others]. However, mainly because of the lack of a shared task evaluation for meeting summarization, the different research used different baselines, and constraints, resulting in a difficulty for comparing these approaches. Our first work on meeting summarization mainly focused on creating baselines and proposed methods for computing oracle accuracies for the widely used ROUGE and weighted precision metrics, for meeting summarization [45].

A major problem in extractive meeting summarization using MMR is finding a proper query: the centroid based query which is commonly used in the absence of a manually specified query cannot significantly outperform a simple baseline system that includes longest sentences in the summary. We introduced a simple yet robust algorithm to automatically extract keyphrases from a meeting which can then be used as a query in the MMR algorithm. We showed that the keyphrase based system significantly outperforms both baseline and centroid based systems and allows for human interaction via keyphrase refinement. As

human refined keyphrases showed even better summarization performance, we formed a graphical user interface that allows for interactive summarization to match the user's needs in terms of summary length and topic focus [44]. Later on, instead of using the greedy MMR algorithm, we extended our summarization approach with integer linear programming to find the subset of sentences that cover as many high-weight keyphrases as possible, while satisfying the summary length constraint [27].

Argument diagramming aims at tagging the utterances and their relationships to represent the flow and structure of reasoning in conversations, especially in discussions and arguments. Argument diagrams extracted from meetings can be useful for meeting participants, to help them in following discussions and catch up with arguments, if the maps can be extracted during the meeting [21]. There is a wide body of work that focuses on visualization of argument maps, as entered by the conversation participants [21, 1]. Argument diagrams can also help users in browsing past meetings, tracking progress across several meetings and can be useful in meeting summarization. [47] describes experiments with human subjects, and their results indicated that argumentation information from meetings can be useful in question answering.

In our initial work, we tackled the problem of automatically assigning node types to user utterances using several lexical and prosodic features. We performed experiments using the AMI Meeting Corpus annotated according to the the Twente Argumentation Schema (TAS) [46]. In TAS, argument diagrams are tree-structured; the nodes of the tree contain speech act units (usually parts of or complete speaker turns) and the edges show the relations between the nodes, the edges emanate from parents and end at children nodes, where the children nodes follow parent nodes in time. There are five types of nodes, that are categorized into to higher level types: *issues and statements*. The *issue* nodes mainly open up an issue and request a response and are further categorized into three depending on the form of the response they expect: *open issue* (OIS), *A/B issue* (AIS) and Yes/No issue (YIS). More information about TAS can be found in [46]. Our results indicate that while lexical and prosodic features both provide orthogonal information for this task, using a cascaded approach that first detects and eliminates backchannel utterances and then categories the remaining non-backchannel utterances improves the node type detection performance [29]. With this final approach, when all features are used, we achieved about 9% relatively better error rates than a simpler classifier based on only lexical features.

5.7 Paraphrasing

Given a sentence as input, the ICSI "Mutaphraser" outputs a huge number of semantically similar sentences. For example, given the input "I like eating cheese", the system outputs the syntactically distant "Eating cheese is liked by me", the semantically distant "I fear sipping juice", and thousands of other sentences. Much of the work on the Mutaphrase algorithm in 2008 was under the hood in components that are not directly visible to the user. The major added component was a generalized unification routine that combines two input objects into a composite object that is consistent with both sources of information according to a flexible specification. This was necessary in a number of places in the algorithm, but most importantly this enables the algorithm to correctly deal with optional sentential elements, producing a much larger number of potential paraphrases and to more

robustly produce sentences that follow the rules of grammar (such as verbs having the right form to go with their subjects). Another added component, traversing the frame semantic hierarchy, allows the Mutaphraser to produce a broader set of paraphrases based on more distantly related semantic material.

We also made some progress on integration of automatic frame parsers, but encountered a major obstacle. The tags that the automatic frame parsers use for Grammatical Function and Phrase Types are inconsistent with those used in FrameNet, and there is no simple mapping from one to the other. We are in the process of working with the authors of the automatic frame parsers to provide a compatible tag set.

5.8 Dealing with real-world sound mixtures

The majority of real-world sound analysis problems involve challenges of interfering sources and imperfect acoustics (reverberation, etc.). We have continued our work on overcoming these issues through detailed models of the circumstances, and the automatic identification of adverse interference.

For the scenario in which speech is received in the presence of competing talkers and noise, and in a room with significant reverberant reflections, we have developed the MESSL-SP model (for Model-based Expectation Maximization Source Separation and Localization using Source Priors). This model identifies in time-frequency the energy that constitutes the "direct path" of a particular target voice by estimating the pertinent between-channel level and phase differences from a pair of microphones, and further refines its estimate of the target speech by using an adapted model of speech acoustics (the "source prior") (Weiss et al. 2008). The speech source model is based on a classical hidden Markov model (HMM) from speech recognition, but employs a high-resolution spectrum (320 frequency samples) to be able to lock in to the individual harmonics in a mixture. To allow the model to match individual speakers closely, while enforcing sufficient constraints to enable the matching to be done even on mixed or corrupted speakers, we use an eigen-voice technique, where the space of thousands of HMM parameters is spanned by a small number of basis functions, obtained from principal component analysis of a small set of speaker-dependent models [53]. Variations in overall speaker level and gross channel characteristics can be accommodated with a few additional basis functions, similarly estimated from the mixture signal alone.

Sometimes, interference encountered can have a great deal of structure in its own right. Music is a common example, particularly in audio recorded "in the field" (such as the soundtrack of consumer-recorded videos). Identifying this kind of interference can be useful to compensate (or suppress) speech recognition, and can be an interesting indexable attribute in its own right. We have developed mechanisms to detect music in environmental recordings that rely on two key features of musical sound: notes and rhythm. By looking for stable, persistent peaks in the long-period lags of autocorrelation, we can distinguish between notes (whose basic period, indicating the musical note, is often stable for hundreds of milliseconds) and the more fleeting periodicities associated for instance with speech. By looking for peaks in the autocorrelation of the overall signal envelope at even longer periods, we can effectively recognize the pulse of rhythmic music. While certain music may have only one of these properties (notes and pulse), it is hard to imagine music that would have neither, and our evaluations show that this detector is highly effective [36]. This

work is a special case of our broader project to be able to effectively classify audio-visual content based on features of the soundtrack, for instance to aid in search and retrieval of large personal video archives [37]. Of several techniques we have investigated to solve this general classification problem, the most effective has been using probabilistic Latent Semantic Analysis (pLSA), in which a discrete probability distribution is modeled as arising from the combination of a set of distinct, automatically-learned "topics". We use this to describe histograms of component use a large Gaussian Mixture model trained to fit the per-frame MFCC features of the soundtrack, and then learn how best to associate the inferred topic weights with explicit, human-generated labels, such as the tags applied to YouTube videos.

5.9 Speaker Recognition

Structured approaches to data selection for speaker recognition: As noted in previous studies (and in last year's annual report), 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) to improve speaker recognition performance. We are now interested in developing a systematic approach to determining sound units that could be employed in this way. Towards 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. Once suitable measures are obtained, they can be applied in a unit-selection procedure, where almost all conceivable units are examined for their utility for speaker recognition.

Many experiments have been done, but the most interesting result currently is that a set of features oriented towards the detection of nasalized sounds is one of the most effective ways of selecting useful data that we have found. The regression of the mean and variances of a set of 12 features for the detection of nasalized sounds currently give us a 90% correlation with the equal error rates using a set of 30 phones as speech units. We have also found it to be useful to get an estimate of the mutual information between unit-constrained feature vectors (i.e., those feature vectors generated for speech segments aligned to labels for the chosen units) and the speaker identity, as well as the overall kurtosis of the unit-constrained feature vectors. According to [38], the mutual information and kurtosis have roughly -83.5% and 71.5% correlations with the equal error rates.

In terms of applying the measures to the actual selection of speech units to incorporate into a overall speaker recognition system, we have attempted to select units that have high speaker discriminative capability according to the mutual information measure (high-relevance), and low-redundancy of its feature vectors with those of other selected units (according to a set of correlations involving the relative speaker recognition improvement of the score-level combination of unit pairs) [38]. We have found that selecting units via our relevance-redundancy technique allows us to select units that have a slightly lower overall equal error rate in combination than simply selecting the unit with the top individual equal error rates [38].

Studies of intrinsic speaker variation re value analysis for speaker recognition: Speaker signals can vary greatly because of extrinsic sources of variability, such as noise or

channel characteristic. In the work reported here, we focused on the sources of variability due to intrinsic factors. To do so, we employed a basic GMM speaker recognition system, using 1024 Gaussians with CMS and T-norm applied as channel compensation techniques. We obtained scores for every possible target-test conversation pair in Switchboard-1, and used the scores along with corresponding metadata (e.g., sex, birth year, dialect area, etc) to look for effects of different speaker characteristics on the system scores. Speaker sex and age difference between speakers (more than 5 years versus less) both had an evident impact, while education level and dialect area did not show any clear effect.

Then, making use of system submissions from NIST's 2008 Speaker Recognition Evaluation, we found that the target score distributions have positive skew, even after mean subtraction of the average target score for each speaker. Informal listening experiments suggested that the cause for target score outliers was often a match (or mismatch) in some aspect of the training and test data. The match (or mismatch) may be in speaker mode (e.g., enthusiastic, higher pitched, laughing vs. serious, lower pitched), general speaking style and vocabulary (e.g., varying slang usage and pronunciation depending on the listener), language (even in primarily English conversations), or channel.

MLP-based Speaker ID: We have continued our work on using neural networks for speaker recognition, where the inputs to the networks are short-duration acoustic features from a test speaker and a target speaker and the output is 1.0 if the test speaker is the same person as the target speaker and 0.0 otherwise. A single neural network trained on all combinations of phones (and silence) yielded a 22% equal error rate, but ran prohibitively slowly. Experiments using super-vector systems yielded significantly inferior results (enough so that it deserves additional investigation). Another system consisting of 48 different networks, one for each phone, was also implemented. This yielded a system more than 50x faster than the original. The best single phones included /DH/, /NG/, and /IY/, each with equal error rate of about 23%. The best combination of 10 phones had equal error rate of 17.91%. Interestingly, the best combination of 4 phones (/AW/, /ER/, /AA/ and /NG/) had a nearly identical equal error rate of 17.98%. These rates are comparable with simple gaussian mixture systems. This suggested that the new systems might combine well with more standard systems, but we have not yet demonstrated this.

5.10 Speech/Audio Hybrid Coding

2008 marked the third year of our collaboration with IDIAP and Qualcomm on the development of a hybrid speech/audio codec based on Frequency Domain Linear Prediction (FDLP). In this phase of the project, we sought better compression efficiency while reducing the algorithmic latency of last year's codec. Our colleagues at IDIAP concentrated on modifying the codec to operate under 200 ms latency instead of the original 1 sec. To achieve this reduction, a lower delay (32 ms) non-uniform Quadrature Mirror Filterbank (QMF) was implemented. The 200 ms codec achieved the same perceptual quality while paying a small 5% penalty in bitrate. Moreover, a speech-only codec operating at low bit-rates (12 kbps) and under various low latencies (25 ms to 200 ms) was proposed which achieved better perceptual performance than AAC but not as good as AMR.

At ICSI we focused on the second stage filterbank applied on the FDLP residual. We showed improved performance by replacing last year's DFT with a Modified Discrete Cosine Transform (MDCT). The phase of the DFT was very hard to code while the MDCT coefficients proved to vector-quantize (VQ) well. We used the MPEG1 psychoacoustic model in order to scalar-quantize (SQ) the spectral coefficients. We applied SQ on the output of the MDCT and we experimented using FDLP on low delay-MDCT. We implemented the noiseless coding stage of AAC which uses a set of static Huffman codes that are dynamically searched at runtime to adapt to the local statistics of the signal. We collected statistics for 2- and 4-tuple-coefficient Huffman tables in order to compress SQ coefficients. The tables achieve compression of around 1.5x which is close to the published AAC performance.

References

[1] M. S. Bachler, S. J. Buckingham Shum, D. C. De Roure, D. T. Michaelides, and K. R. Page (2003) "Meeting Support by Visualizing Discussion Structure and Semantics," Proceedings of the first International Workshop on Hypermedia and the Semantic Web (HTSW2003) Nottingham, United Kingdom, August 2003.

[2] K. Boakye (2008) "Audio Segmentation for Meetings Speech Processing," UC Berkeley Ph.D. Thesis 2008.

[3] K.A. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland (2008) "Overlapped Speech Detection for Improved Speaker Diarization in Multiparty Meetings," Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4353-4356 Las Vegas, NV, March 2008.

[4] K. Boakye, O. Vinyals, and G. Friedland (2008) "Two's a Crowd: Improving Speaker Diarization by Automatically Identifying and Excluding Overlapped Speech", Proceedings of the 9th International Conference of the ISCA (Interspeech 2008), pp. 32-35 Brisbane, Australia, September 2008.

[5] J. Carbonell and J. Goldstein (1998) "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Inforrmation Retrieval, pp. 335-336 Melbourne, Australia, August 1998.

[6] B. Favre, R. Grishman, D. Hillard, H. Ji, D. Hakkani-Tür, and M. Ostendorf (2008) "Punctuating Speech for Information Extraction," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5013-5016 Las Vegas, Nevada, March 2008.

[7] B. Favre, D. Hakkani-Tür, S. Petrov, and D. Klein (2008) "Efficient Sentence Segmentation using Syntactic Features, Proceedings of IEEE Workshop on Spoken Language Technologies (SLT2008), pp. 77-80 Goa, India, December 2008.

[8] B. Favre, D. Hakkani-Tür, and E. Shriberg (2009) "Syntactically Informed Models for Comma Prediction," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) Taipei, Taiwan, April 2009, to appear.

[9] G. Friedland, E. Hensley, J. Schumacher, and R. Jain (2008) "Appscio: A Software Environment for Semantic Multimedia Analysis," Proceedings of IEEE International Conference on Semantic Computing, pp. 456-459 Santa Clara, California, August 2008.

[10] G. Friedland, H. Hung, and C. Yeo (2008) "Multi-Modal Speaker Diarization of Real-World Meetings Using Compressed-Domain Video Features," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) Taipei, Taiwan, April 2009, to appear, also International Computer Science Institute Technical Report 08-007 October 2008.

[11] G. Friedland, W. Hürst, and L. Knipping (2008) "Multimedia Education – Can we find Unity in Diversity?" Proceedings of 16th ACM International Conference on Multimedia, pp. 1115-1116 Vancouver, Canada, October 2008.

[12] G. Friedland, L. Knipping, and W. Hürst (2008) "Multimedia Education in Computer Science – A Little Bit of Everything Is Not Enough," *IEEE Multimedia Magazine*, Vol. 15, No. 2, pp. 78-82 April-June 2008.

[13] G. Friedland, L. Knipping, and W. Hürst (2008) "Automated Lecture Recording," in *Encyclopedia of Multimedia*, Burko Furht, ed., 2nd edition Springer, October 2008.

[14] G. Friedland, L. Knipping, and W. Hürst (guest editors) (2008) "Educational Multimedia," Special Section in *IEEE Multimedia Magazine*, pp. 54-74 July-September 2008.

[15] G. Friedland and R. Rojas (2008) "Anthropocentric Video Segmentation for Lecture Webcasts," *Journal on Image and Video Processing*, Vol. 8, Issue 2, Article 9 January 2008.

[16] G. Friedland and O. Vinyals (2008) "Live Speaker Identification in Conversations," Proceedings of 16th ACM International Conference on Multimedia, pp. 1017-1018 Vancouver, Canada, October 2008.

[17] 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* 2009, to appear.

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

[19] G. Friedland and D. van Leeuwen (2009) "Speaker Diarization and Identification," chapter in *Semantic Computing*, P. Sheu et al., eds. IEEE Press/Wiley, 2009, to appear.

[20] D. Gelbart (2007) "Noisy Numbers data and Numbers testbeds," International Computer Science Institute, Berkeley, CA. Online: http://www.icsi.berkeley.edu/speech/papers/gelbart-ms/numbers/.

[21] K. Fujita and K. Nishimoto and Y. Sumi and S. Kunifuji and K. Mase (1998) "Meeting Support by Visualizing Discussion Structure and Semantics," Proceedings of the 2nd International Conference on Knowledge-Based Intelligent Electronic Systems, Vol. 1, pp. 417-422 Adelaide, Australia, April 1998.

[22] S. Ganapathy, P. Motlicek, H. Hermansky, and H. Garudadri (2008) "Temporal masking for Bit-rate Reduction in Audio Codec based on Frequency Domain Linear Prediction," Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4781-4784 Las Vegas, Nevada, April 2008.

[23] S. Ganapathy, P. Motlicek, H. Hermansky, and H. Garudadri (2008) "Autoregressive Modeling of Hilbert Envelopes for Wide-Band Audio Coding," Proceedings of 124th Convention of Audio Engineering Society (AES), paper 7481 Amsterdam, the Netherlands, May 2008.

[24] S. Ganapathy, P. Motlicek, H. Hermansky, and H. Garudadri (2008) "Spectral Noise Shaping: Improvements in Speech/Audio Codec Based on Linear Prediction in Spectral Domain," Proceedings of the 9th International Conference of the ISCA (Interspeech 2008) Brisbane, Australia, September 2008.

[25] D. Gelbart (2008) "Ensemble Feature Selection for Multi-stream Automatic Speech Recognition," Ph.D. Thesis, University of California at Berkeley

[26] D. Gillick, B. Favre, and D. Hakkani-Tür (2008) "The ICSI Summarization System," Proceedings of Text Analysis Conference (TAC2008) Gaithersburg, Maryland, November 2008.

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

[28] U. Guz, B. Favre, D. Hakkani-Tür, and G. Tur (2009) "Generative and Discriminative Methods using Morphological Information for Sentence Segmentation of Turkish," *IEEE Transactions on Speech, Audio and Language Processing, Special Issue on Processing Morphologically Rich Languages* May 2009, to appear.

[29] D. Hakkani-Tür (2009) "Towards Automatic Argument Diagramming of Multiparty Meetings," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) Taipei, Taiwan, April 2009, to appear.

[30] D. Hillard, Z. Huang, H. Ji, R. Grishman, D. Hakkani-Tür, M. Harper, M. Ostendorf, and W. Wang (2006) "Impact of Automatic Comma Prediction on POS/Name Tagging of Speech," Proceedings of IEEE Workshop on Spoken Language Technologies (SLT2006), pp. 58-61 Palm Beach, Aruba, December 2006.

[31] H. Hung and G. Friedland (2008) "Towards Audio-Visual On-Line Diarization Of Participants In Group Meetings," Proceedings of European Conference on Computer Vision (ECCV) 2008 Marseille, France, October 2008.

[32] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez (2008) "Estimating the Dominant Person in Multi-Party Conversations Using Speaker Diarization Strategies," Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2197-2200 Las Vegas, NV, March 2008.

[33] H. Hung, C. Yeo, and G. Friedland (2009) "Approaching On-Line Speaker Diarization and Audio-Visual Localization Through Aspects of Human Discourse," invited to *Computer Vision and Image Understanding* Elsevier, 2009, to appear.

[34] M. Knox (2008) "Automatic Laughter Segmentation," UC Berkeley Masters Thesis 2008.

[35] M. Knox, N. Morgan, and N. Mirghafori (2008) " Getting the Last Laugh: Automatic Laughter Segmentation in Meetings," Proceedings of the 9th International Conference of the ISCA (Interspeech 2008), pp. 797-800 Brisbane, Australia, September 2008.

[36] K. Lee and D. Ellis (2008) "Detecting Music in Ambient Audio by Long-Window Autocorrelation," Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 9-12 Las Vegas, Nevada, April 2008.

[37] K. Lee and D. Ellis (2009) "Audio-Based Semantic Concept Classification for Consumer Video," submitted to *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[38] H. Lei (2009) "Towards Structured Approaches to Arbitrary Data Selection and Performance Prediction for Speaker Recognition", accepted to 3rd International Conference on Biometrics, 2009.

[39] C. Lin (2004) "ROUGE: a Package for Automatic Evaluation of Summaries," Proceedings of ACL Text Summarization Workshop, pp. 74-81 Barcelona, Spain, July 2004.

[40] J. Makhoul, A. Baron, I. Bulyko, L. Nguyen, L. Ramshaw, D. Stallard, R. Schwartz, and B. Xiang (2005) "The Effects of Speech Recognition and Punctuation on Information Extraction Performance," Proceedings of 9th European Conference on Speech Communication and Technology (Eurospeech), pp. 57-60 Lisbon, Portugal, September 2005.

[41] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney (2007) "Improving Speech Translation with Automatic Boundary Prediction," Proceedings of International Conference on Spoken Language Processing (Interspeech), pp. 2449-2452 Antwerp, Belgium, August 2007.

[42] P. Motlicek, S. Ganapathy, H. Hermansky, H. Garudadri, and M. Athineos (2008) "Perceptually Motivated Sub-Band Decomposition for FDLP Audio Coding," Proceedings of 11th International Conference on Text, Speech, and Dialogue (TSD 2008), pp. 435-442 Brno, Czech Republic, September 2008.

[43] G. Murray and S. Renals (2007) "Term-Weighting for Summarization of Multi-Party Spoken Dialogues," *Machine Learning for Multimodal Interaction IV*, pp. 155-166 *Lecture Notes in Computer Science*, Vol. 4892, Springer, 2007.

[44] K. Riedhammer, B. Favre, and D. Hakkani-Tür (2008) "A Keyphrase Based Approach to Interactive Meeting Summarization," Proceedings of IEEE Workshop on Spoken Language Technologies (SLT2008), pp. 153-156 Goa, India, December 2008.

[45] K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tür (2008) "Packing the Meeting Summarization Knapsack," Proceedings of the 9th International Conference of the ISCA (Interspeech 2008), pp. 2434-2437 Brisbane, Australia, September 2008.

[46] R. Rienks, D. Heylen, and E. van der Weijden (2005) "Argument Diagramming of Meeting Conversations," Proceedings of Workshop on Multimodal Multiparty Meeting Processing at the 7th International Conference on Multimodal Interfaces (ICMI), pp. 85-92 Trento, Italy, October 2005.

[47] R. Rienks and D. Verbree (2006) "About the Usefulness and Learnability of Argument-Diagrams from Real Discussions," Proceedings of 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms Washington, D.C., May 2006.

[48] O. Vinyals and G. Friedland (2008) "A Hardware-Independent Fast Logarithm Approximation with Adjustable Accuracy," Proceedings of the 10th IEEE International Symposium on Multimedia, pp. 61-65 Berkeley, California, December 2008.

[49] O. Vinyals and G. Friedland (2008) "Modulation Spectrogram Features for Speaker Diarization," Proceedings of the 9th International Conference of the ISCA (Interspeech 2008), pp. 630-633 Brisbane, Australia, September 2008.

[50] O. Vinyals and G. Friedland (2008) "Live Speaker Identification in Meetings: 'Who is Speaking Now?'," International Computer Science Institute Technical Report TR-08-001 January 2008.

[51] O. Vinyals and G. Friedland (2008) "Towards Semantic Analysis of Conversations: A System for the Live Identification of Speakers in Meetings," Proceedings of IEEE International Conference on Semantic Computing, pp. 426-431 Santa Clara, California, August 2008.

[52] R. Weiss and D. Ellis (2009) "A Variational EM Algorithm for Learning Eigenvoice Parameters in Mixed Signals," Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP) Taipei, Taiwan, April 2009.

[53] R. Weiss, M. Mandel, and D. Ellis (2008) "Source Separation Based on Binaural Cues and Source Model Constraints," Proceedings of the 9th International Conference of the ISCA (Interspeech 2008), pp. 419-422 Brisbane, Australia, September 2008.

[54] S. Xie and Y. Liu (2008) "Using Corpus and Knowledge-Based Similarity Measure in Maximum Marginal Relevance for Meeting Summarization," Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4985-4988 Las Vegas, Nevada, March 2008.

6 Architecture

The architecture group began at ICSI in the summer of 2007, with the major focus being 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 this year were in silicon photonics [3] and the vector-thread architecture [2].

6.1 Monolithic Silicon Photonics

In a collaboration with the MIT Center for Integrated Photonic Systems, ICSI architecture researchers are exploring the use of silicon photonics for processor-to-memory interconnect. Projected advances in electrical signaling seem unlikely to fulfill the memory bandwidth demands of future manycore processor chips. Monolithic silicon photonics, which integrates optical components with electrical transistors in a conventional CMOS process, is a promising new technology that could provide large improvements in achievable interconnect bandwidth. In a DARPA-funded effort, the ICSI architecture group is exploring possible memory interconnect schemes to exploit the new photonic device and circuit technology being developed at MIT.

The MIT-developed monolithic silicon photonics technology is well-suited for integration with standard bulk CMOS processes, which reduces costs and improves opto-electrical coupling compared to previous approaches. The technology supports dense wavelength-division multiplexing with dozens of wavelengths per waveguide. The MIT group have fabricated several test chips and measured the device properties. Exploiting key features of this photonics technology, we have developed a processor-memory network architecture for future manycore systems based on an opto-electrical global crossbar. Based on the MIT device measurements, and through the use of both analytical models and detailed microarchitectural simulations, we have shown that for a power-constrained system with 256 cores connected to 16 DRAM modules, aggregate network throughput can be improved by $\approx 8 - -10\times$ compared to an optimized purely electrical network [3].

We are continue to explore the space of possible network designs and improvements to global flow control and routing. We are also working with our MIT collaborators to define requirements to guide future device development and fabrication.

6.2 Maven (Malleable Array of Vector-thread ENgines)

In earlier work at MIT, Asanović's team developed the Scale vector-thread architecture and processor prototype [4], 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 [2].

Our primary work over this period has been in designing an efficient microarchitecture with a considerable reduction in control logic complexity compared to the earlier Scale design. We are progressing with the development of an architectural simulator and a software compilation toolchain for Maven.

6.3 Other Collaborations

The architecture group works closely with the new Parallel Computing Laboratory (Par Lab) in the Computer Science Division at UC Berkeley, which is developing a new parallel software stack for future parallel architectures [1]. The ICSI work will be using 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 RAMP (Research Accelerator for Multi-Processors) consortium hosted at the Berkeley Wireless Research Center (BWRC), which is developing technology for rapid simulation of large-scale multiprocessors using FPGAs (Field-Programmable Gate Arrays). The RAMP infrastructure will be used as the simulation platform to evaluate new architecture and device technologies designed at ICSI.

References

[1] K. Asanović, R. Bodik, J. Demmel, T. Keaveny, K. Keutzer, J. D. Kubiatowicz, E. A. Lee, N. Morgan, G. Necula, D. A. Patterson, K. Sen, J. Wawrzynek, D. Wessel, and K. A. Yelick (2008) "The Parallel Computing Laboratory at UC Berkeley: A Research Agenda Based on the Berkeley View," University of California at Berkeley EECS Department Technical Report UCB/EECS-2008-23 March 2008.

[2] C. Batten, H. Aoki, and K. Asanović (2008) "The Case for Malleable Stream Architectures," Proceedings of the Workshop on Streaming Systems at MICRO-41 Lake Como, Italy, November 2008.

[3] C. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C. Holzwarth, M. Popović, H. Li, H. Smith, J. Hoyt, F. Kärtner, R. Ram, V. Stojanović, and K. Asanović (2008) "Building Manycore Processor-to-DRAM Networks with Monolithic Silicon Photonics," Proceedings of IEEE Symposium on High-Performance Interconnects (Hot Interconnects 2008), pp. 21-30 Stanford, California, August 2008.

[4] 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 Vision

Prof. Trevor Darrell and his group members moved from MIT to UC Berkeley and ICSI in 2008 to start ICSI's new Vision group. ICSI-funded activity started in January while group members still resided in Cambridge, MA: Research Scientist Dr. Raquel Urtasun, and MIT Doctoral Students Mario Christhoudias and Kate Saenko were funded to work on DARPAs URGENT and CALO projects. (And urgent it was, given the hectic pace of DARPA deliverables!) MIT students Tom Yeh and Arianda Quattoni remained at MIT CSAIL with MIT funding and continue to be Prof. Darrells advisees and affiliate group members. Two UCB EECS graduate students joined the group during the fall term, and a postdoctoral fellow, Dr. Mario Fritz, joined from TU-Darmstadt in October 2008, winning a prestigious Humboldt fellowship. The group continues to grow rapidly, with three additional postdocs (Brian Kulis, Mathieu Salzmann, and Carl Ek) scheduled to start by the end of February 2009 and recruiting underway for several new EECS graduate admits.

The Vision group focuses on solutions to two core computer vision problemsthe perception of human motion and expression for multimodal interfaces, and the recognition of object and scene categories–and their application to mobile phone interfaces and to interactive robotics. Awareness of the environment and of a user's presence and/or expression is critical for smart mobile interfaces, and of course for successful real-world action on mobile agents.

Funding for ICSI Vision research in 2008 was obtained from several sources, including:

- - DARPA projects on gesture recognition (ULTRA-VIZ), meeting understanding (CALO), as well as object category recognition in urban environments (URGENT).
- an NSF grant for understanding perceptually situated human-robot dialog
- a Google gift for developing efficient category-level object indexing algorithms
- a contract from DHHS and Children's Hospital Boston to develop facial image indexing interfaces
- a contract from Toyota to develop practical object recognition capabilities for domestic service robots

7.1 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, families could be separated for months. After Katrina it was 6 months until the last child was reunited with her family. We are working on a system where 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.

7.2 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.

7.3 Probabilistic Models for Multi-View Learning and Distributed Feature Selection

Many problems in machine learning contain datasets that are comprised of 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. 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.

7.4 Interactive Image Matching for Information Retrieval and Human Computer Interaction

Recent advances in content-based image retrieval have made it possible to index and search millions of images accurately and efficiently. Finding images, instead of an end itself, can be an effective mean for human users to perform a wide variety of interesting tasks in information retrieval and human-computer interactions. For example, by finding images online that resemble an object a human user is looking at, the text surrounding the online images may contain useful information about the object. Moreover, in terms of usability, the user may find it more intuitive to simply take the image of the object as the input query, compared to using keywords to describe the object. In this project, we examine various ways to exploit the vast amount of online multimedia data with both image and text. Specifically, we develop prototype systems to search online catalog for product information and screenshots for software tutorial. Also, we investigate usability issues such as learnability (how easy is it to learn to use the system to find information) and efficiency (how much quicker can the user find it).

7.5 Multi-Modal Learning and Sensing for Mobile Robotic Systems

One of the most vital competences of mobile robotic platform is to the ability to recognize and categorize entities in its environment. Tasks as reasoning about the current state of the world, assessing consequences of possible actions, as well as planning future episodes build on such a basic "understanding" of what roles objects and places may possibly play. We are driven by the goal to enable such systems with this basic capability by making best possible use of the available sensing modalities. Therefore we plan to extend our research on visual categorization towards optimal sensor fusion and dealing with heterogenous training data.

References

[1] C. M. Christoudias, R. Urtasun, and T. Darrell (2008) "Multi-View Learning in the Presence of View Disagreement," Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI) Helsinki, Finland, July 2008.

[2] C. M. Christoudias, R. Urtasun, and T. Darrell (2008) "Unsupervised Distributed Feature Selection for Multi-View Object Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Anchorage, Alaska, June 2008.

[3] M. Fritz and B. Schiele (2008) "Decomposition, Discovery, and Detection of Visual Categories Using Topic Models," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Anchorage, Alaska, June 2008.

[4] A. Quattoni, M. Collins, and T. Darrell (2008) "Transfer Learning for Image Classification with Sparse Prototype Representations," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Anchorage, Alaska, June 2008.

[5] K. Saenko and T. Darrell (2008) "Unsupervised Learning of Visual Sense Models for Polysemous Words," Proceedings of Neural Information Processing Systems Conference (NIPS) Vancouver, Canada, December 2008, to appear.

[6] K. Saenko, K. Livescu, J. Glass, and T. Darrell (2008) "Multistream Articulatory Feature-Based Models for Visual Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI) December 2008, to appear.

[7] M. Salzmann, R. Urtasun, and P. Fua (2008) "Local Deformation Models for Monocular 3D Shape Recovery," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Anchorage, Alaska, June 2008.

[8] R. Urtasun, D. J. Fleet, A. Geiger, J. Popovic, T. Darrell, and N. D. Lawrence (2008) "Topologically-Constrained Latent Variable Models," Proceedings of International Conference in Machine Learning (ICML), pp. 1080-1087 Helsinki, Finland, July 2008.

[9] R. Urtasun and T. Darrell (2008) "Local Probabilistic Regression for Activity-Independent Human Pose Inference," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Anchorage, Alaska, June 2008.

[10] T. Yeh, J. Lee, and T. Darrell (2008) "Photo-Based Question Answering," Proceedings of the ACM International Conference on Multimedia, pp. 389-398 Vancouver, Canada, December 2008.

[11] T. Yeh and T. Darrell (2008) "Dynamic Visual Category Learning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Anchorage, Alaska, June 2008.