CUCS Resources

You can subscribe to the CUCS news mailing list at **http://lists.cs.columbia.edu/mailman/listinfo/cucs-news/**

CUCS colloquium information is available at **http://www.cs.columbia.edu/lectures**

Visit the CUCS alumni portal at **http://alum.cs.columbia.edu**

Columbia University
in the City of New York
Department of Computer Science
1214 Amsterdam Avenue
Mailcode: 0401
New York, NY 10027-7003

ADDRESS SERVICE REQUESTED

NON-PROFIT ORG.
U.S. POSTAGE
PAID
NEW YORK, NY
PERMIT 4332

CS@CU

NEWSLETTER OF THE DEPARTMENT OF COMPUTER SCIENCE AT COLUMBIA UNIVERSITY VOL.2 NO.1 WINTER 2006

Bill Gates Visits Columbia

Bill Gates shares his vision of the future with a capacity crowd at Columbia on October 13.

On October 13, Columbia University played host to a distinguished guest when Bill Gates, the Chairman and Chief Software Architect of Microsoft, visited campus. Gates visited Columbia as part of a six-university tour undertaken with the goal of fostering interest in computer science and related fields; this follows a similar tour of five other campuses last spring. Computer Science department chair Henning Schulzrinne stated that "It validates Columbia as one of the very small number of places in the U.S. that Gates put on his list of five or six places that he's visiting... I think that's a very good sign that Columbia is one of those places that somebody of his stature would want to visit."

The Faculty Roundtable

Before speaking to a packed audience of students in Lerner Hall, Gates participated in a roundtable discussion with Computer Science department faculty. Department chair Henning Schulzrinne started things off with a brief presentation on "Columbia's Vision for Tomorrow's Global Intelligent Systems." Henning gave a bird's-eye view of the many different active research areas in computer science at Columbia, ranging across topics such as theory, digital systems design, computer systems and networks, interacting with humans, graphics/ robotics/ vision, and making sense of data. He highlighted several innovative research projects underway at CUCS, including the Newsblaster system for automatic summarization that has been developed by Professor McKeown and her team and the work on distributed channel allocation that is being carried out by Professors Misra and Rubenstein.

Following Henning's presentation, Professor Shree Nayar gave an overview of his lab's research on computational cameras and displays. Shree started off by talking about his research on wide angle imaging (which has led to commercial development of 360-degree cameras for videoconferencing and other applications) and on radial stereoscopic imaging, which enables a "depth camera" to generates three-dimensional images.

Shree also described some exciting recent work using randomized point clouds which may soon lead to inexpensive three-dimensional displays constructed of laser-damaged glass. At the end of Shree's presentation, he gave Bill Gates the actual glass cube that was used to implement a three-dimensional version of Pac-Man. (As Shree noted, this was one of the few things that Bill would have a hard time acquiring for himself.)

Perhaps the high point of the faculty roundtable was a question-and-answer session between Gates and Computer Science faculty members. To start things off, Professor Kathy McKeown noted that

(continued on next page)

Cover Story (continued)

Bill Gates greets an enthusiastic crowd.

there are major obstacles to increasing the number of women and minorities in computer science, and that the undergraduate computer science curriculum at Columbia has recently been broadened to address this. (See page 3 for a related article on the Women in Computer Science organization at Columbia.) She asked Gates how Microsoft and universities might systematically work together to address these issues. In response, Gates stated that there is no easy solution, and that the dropoff in women and minorities occurs at every stage of the pipeline—from early interest in math and science, through introductory computer science courses, selecting a major, and even into one's professional career. Just as there is dropoff at all stages, he advocated working to increase the participation of women and minorities in computer science at all stages—from early outreach at summer camps and high schools all the way through college and industry efforts. (As Gates briefly noted, Microsoft Senior Vice President of Research Rick Rashid is a trustee of the Anita Borg Institute for Women and Technology.)

A second question from Professor Al Aho concerned the future of computer science. Al had previously canvassed the Computer Science faculty for their "top five" most important unsolved questions in computer science, and they came up with the following list:

1. How do we prove that certain problems are hard to solve?
2. How can we make truly reliable software? Is there something analagous to Shannon's work on error correction and Von Neumann's work on reliability through redundancy?
3. How should we architect systems so that they can be more easily maintained and evolved?
4. How can we make computers understandable and usable for people of all backgrounds, ages, and abilities, in the many different situations we encounter in life?
5. How can we program computers to have human qualities such as consciousness, intelligence and emotion?

Al asked Gates how these questions fit into Microsoft's view of the future of computer science. In response, Gates noted that questions 2 and 3 (reliability and maintability) are largely the same thing, and that this is a tremendous issue for Microsoft; evolving existing systems is a tremendous challenge because of deep compatibility constraints. He pointed out that tools for establishing program correctness are now becoming useful in practical settings, so there is reason for hope on these questions.

Gates also added his own "big question" how to design concurrent systems. Since processor clock rates are not likely to increase in the next decade the way they have in the past, there is a tremendous challenge in sustaining the exponential growth in computing ability to which users have grown accustomed to. Given constraints on clock speeds, personal computers will only continue to get faster and better if we are able to make strides in writing concurrent programs.

Speaking to Students

After the faculty roundtable, Gates addressed and fielded questions from a standing-room-only crowd of roughly 1200 students in Arledge Auditorium. His talk started off with a brief video presentation highlighting some familiar faces: Columbia CS students who are working or interning at Microsoft.

The main message of Gates's address was that while the past 30 years of computer and software development have radically changed the way we live and work, the coming years will be even more exciting and revolutionary. By many measures such as memory, bandwidth and display abilities, today's computers are more than a million times better than the machines of three decades ago; these tremendous capabilities can only be exploited to their fullest if the best minds continue to be interested in software development. (Gates highlighted the importance of recruiting top talent for Microsoft with a humorous video in which he co-starred with Jon Heder of "Napoleon Dynamite.")

Gates outlined his vision of a near future in which every computer will be connected to a single network; we will interface with this network through a range of different modalities (wristwatches, phone-size devices, tablets, desktop machines and immersive home environments) but much of the computation will be done remotely. As an illustrative example, when you are travelling in a foreign country you might take a photo of a street sign with your phone and have it instantly translated using image processing and natural language technology running on a remote server. He anticipated widespread use of very large displays and a shift away from paper-based activities; as an example of how such a shift can take place, he pointed out how thoroughly online encyclopedias have displaced traditional paper encyclopedias.

Another theme of Gates's address was that computer technology is a powerful equalizing force; he observed that any student who is connected to the Internet today has far better tools at their disposal to satisfy intellectual curiosity than even the most privileged students did a decade ago. Gates also demonstrated some upcoming Microsoft products and technologies, ranging from software tools to "smart" desktops to the Xbox 360.

SEAS undergraduate Daniel Medrano asks Gates a question after his talk.

After his talk Gates fielded a wide range of questions from student audience members. In response to the question "with Microsoft so large, how do you create an environment

2005 Technical Report Series

The following technical reports were published through October 2005.

All reports are available at *http://www.cs.columbia.edu/research/publications*

CUCS-001-05
The Virtual Device: Expanding Wireless Communication Services Through Service Discovery and Session Mobility
Ron Shacham, Henning Schulzrinne, Srisakul Thakolsri, and Wolfgang Kellerer

CUCS-002-05
A Uniform Programming Abstraction for Effecting Autonomic Adaptations onto Software Systems
Giuseppe Valetto and Gail Kaiser

CUCS-003-05
Dynamic Adaptation of Rules for Temporal Event Correlation in Distributed Systems
Rean Griffith, Joseph L. Hellerstein, Yixin Diao and Gail Kaiser

CUCS-004-05
Genre Classification of Websites Using Search Engine Snippets
Suhit Gupta, Gail Kaiser, Salvatore Stolfo and Hila Becker

CUCS-005-05
Adding Self-healing capabilities to the Common Language Runtime
Rean Griffith and Gail Kaiser

CUCS-006-05
Manipulating Managed Execution Runtimes to Support Self-Healing Systems
Rean Griffith and Gail Kaiser

CUCS-007-05
Internet Routing Dynamics and NSIS Related Considerations
Charles Shen, Henning Schulzrinne, Sung-Hyuck Lee and Jong Ho Bang

CUCS-008-05
P2P Video Synchronization in a Collaborative Virtual Environment
Suhit Gupta and Gail Kaiser

CUCS-009-05
Adaptive Interactive Internet Team Video
Dan Phung and Giuseppe Valetto and Gail Kaiser

CUCS-011-05
A Lower Bound for Quantum Phase Estimation
Arvid J. Bessen

CUCS-012-05
The Power of Various Real-Valued Quantum Queries
Arvid J. Bessen

CUCS-013-05
802.11b Throughput with Link Interference
Hoon Chang and Vishal Misra

CUCS-014-05
Similarity-based Multilingual Multi-Document Summarization
David Kirk Evans, Kathleen McKeown, Judith L. Klavans

CUCS-015-05
Multi-Language Edit-and-Continue for the Masses
Marc Eaddy and Steven Feiner

CUCS-016-05
A Hybrid Hierarchical and Peer-to-Peer Ontology-based Global Service Discovery System
Knarig Arabshian and Henning Schulzrinne

CUCS-017-05
Improving Database Performance on Simultaneous Multithreading Processors
Jingren Zhou, John Cieslewicz, Kenneth A. Ross, and Mihir Shah

CUCS-018-05
Incremental Algorithms for Inter-procedural Automaton-based Program Analysis
Christopher L. Conway, Kedar Namjoshi, Dennis Dams, and Stephen A. Edwards

CUCS-019-05
Classical and Quantum Complexity of the Sturm-Liouville Eigenvalue Problem
A. Papageorgiou and H. Wozniakowski

CUCS-020-05
A Hybrid Approach to Topological Mobile Robot Localization
Paul Blaer and Peter K. Allen

CUCS-021-05
Optimal State-Free, Size-aware Dispatching for Heterogeneous M/G/-type systems
Hanhua Feng, Vishal Misra, and Dan Rubenstein

CUCS-022-05
Merging Globally Rigid Formations of Mobile Autonomous Agents
Tolga Eren, Brian D.O. Anderson, Walter Whiteley, A. Stephen Morse and Peter N. Belhumeur

CUCS-023-05
Time-Varying Textures
Sebastian Enrique, Melissa Koudelka, Peter Belhumeur, Julie Dorsey, Shree Nayar and Ravi Ramamoorthi

CUCS-024-05
The Appearance of Human Skin
Takanori Igarashi, Ko Nishino, and Shree K. Nayar

CUCS-025-05
Tractability of Quasilinear Problems I: General Results
Arthur Werschulz and Henryk Wozniakowski

CUCS-026-05
Quantum algorithms and complexity for certain continuous and related discrete problems
Marek Kwas

CUCS-027-05
Lexicalized Well-Founded Grammars: Learnability and Merging
Smaranda Muresan, Tudor Muresan, and Judith Klavans

CUCS-028-05
Pointer Analysis for C Programs Through AST Traversal
Marcio Buss, Stephen Edwards, Bin Yao, and Daniel Waddington

CUCS-029-05
Learning mixtures of product distributions over discrete domains
Jon Feldman, Ryan O'Donnell, and Rocco A. Servedio

CUCS-030-05
Agnostically Learning Halfspaces
Adam Kalai, Adam Klivans, Yishay Mansour, and Rocco A. Servedio

CUCS-031-05
Generic Models for Mobility Management in Next Generation Networks
Maria Luisa Cristofano, Andrea G. Forte, and Henning Schulzrinne

CUCS-032-05
Parsing Preserving Techniques in Grammar Induction
Smaranda Muresan

CUCS-033-05
PachyRand: SQL Randomization for the PostgreSQL JDBC Driver
Michael Locasto and Angelos D. Keromytis

CUCS-034-05
Tractability of quasilinear problems. II: Second-order elliptic problems
A. G. Werschulz and H. Wozniakowski

CUCS-035-05
DotSlash: Providing Dynamic Scalability to Web Applications with On-demand Distributed Query Result Caching
Weibin Zhao and Henning Schulzrinne

CUCS-037-05
The Pseudorandomness of Elastic Block Ciphers
Debra Cook, Moti Yung, and Angelos Keromytis

CUCS-038-05
A General Analysis of the Security of Elastic Block Ciphers
Debra Cook, Moti Yung, and Angelos Keromytis

CUCS-039-05
On Elastic Block Ciphers and Their Differential and Linear Cryptanalyses
Debra Cook, Moti Yung, and Angelos Keromytis

CUCS-040-05
Square Root Propagation
Andrew Howard and Tony Jebara

CUCS-041-05
Performance and Usability Analysis of Varying Web Service Architectures
Michael Lenner and Henning Schulzrinne

CUCS-042-05
Approximating the Reflection Integral as a Summation: Where did the delta go?
Aner Ben-Artzi

CUCS-043-05
TCP-Friendly Rate Control with Token Bucket for VoIP Congestion Control
Miguel Maldonado, Salman Abdul Baset, and Henning Schulzrinne

"What I did with my summer vacation"

After a busy year in classrooms and labs at Columbia, many CUCS students left the Columbia campus—but not their love of computer science—behind over the summer. Here are brief reports on what some computer science students did with their summer vacations.

Marcio Buss spent 12 weeks this summer as an intern at Bell Laboratories Research in Holmdel, NJ. Bell Labs is the lab that invented Unix, C, C++, the laser, digital signal processors, cellular telephone technology, the transistor, and many other innovations that changed the world. Indeed, the water tower on the park-like property of Bell Labs is actually modeled after the design of an early transistor, created in 1947. (Their first experimental transistor, a misshapen clump of ceramic and wire, would make for a poor water tower.) Marcio worked with his mentor, Bin Yao, in source-code static analysis aimed at dead-code and feature removal. Part of this work is being presented at the 5th IEEE Source Code Analysis and Manipulation Workshop in Budapest, Hungary.

The Bell Labs water tower.

Wisam Dakka writes "Over the summer, I joined a hand-picked group of 150 researchers and engineers known as 'the underdogs.' A year ago, they were assembled to build a large-scale, Windows-based search engine, known to the outside world as MSN Search. In contrast to Google, they reduced to nearly zero the number of people/robots needed to replace hard drives daily. Before I went to Redmond, I had not imagined I would meet people able to cause a revolution inside a company as established as Microsoft. Up against the odds, they did. The group's current focus is improving the quality of the search results and making them as compatible as possible with results from other engines. As an intern, I worked on large-scale problem related to query analysis and ranking. Working on these problems made me realize that having a huge amount of data markedly changes the way we think about problem-solving. Before returning to the East Coast, I hiked and camped my way from Seattle to Washington's Olympia National Park to the Lost Coast in Northern California. The views of the Pacific Ocean were astonishing and inspire me still."

Ariel Elbaz, **Homin K. Lee**, and **Andrew Wan** attended the Fifth Canadian Summer School on Quantum Information. The session was held in Montreal, Quebec in August. Experts in quantum information gave lectures on both the fundamentals and important research developments. The summer school was a great opportunity to learn about advanced topics in quantum information and quantum computation. These included the meaning of negative information, quantum interactive proofs, quantum query complexity, quantum communication complexity, quantum lower bounds, and many other topics that start with the word "quantum." It also gave Ariel, Homin and Andrew a chance to explore Montreal, a city which has a young international population, a vibrant nightlife, and tasty inexpensive food like smoked meat and poutine (similar to disco fries but with cheese curds).

Ariel Elbaz spent 12 weeks this summer as an intern at Mitsubishi Electric Research Labs in Boston (no, they do not make cars—Mitsubishi Motors has a similar name but is an independent company). MERL has ~70 researchers and accepts about the same number of interns each summer. Mr. Elbaz worked with his host, Shai Avidan, in cooperation with his advisors at Columbia, Tal Malkin and Rocco Servedio, on cryptographically secure protocols for private computation and machine learning applications. In addition to a paid research internship and making new connections, he toured Boston, sailed on the Charles river, and missed all the Red Sox games he could.

Joshua Reich spent his summer in Albuquerque, New Mexico working as a summer intern in Sandia National Laboratories' Center for Cyber Defenders.Sandia serves as one of the primary engineering labs in the national laboratories system and was founded as an offshoot of Los Alamos' Z-division (for history buffs). Sandia/NM has about 8,000 employees, with about 30 working in the Center for Cyber Defenders. Josh worked on projects involving Software Signature Generation and Mobile Sensor Simulation Systems. In his spare time, Josh and his wife managed to see quite a bit of New Mexico and its local flavor—including llama hiking, balloon rides, rock climbing, and of course, eating lots and lots of green chiles.

Julia Stoyanovich spent the summer at the IBM Almaden Research Center, where she worked in the Advanced Database Optimization group under the guidance of Jun Rao and Volker Markl. This internship gave Julia an opportunity to work with some of the top people in the field, as well as to trade the heat and humidity of New York City for the perfect Californian climate.

Krysta Svore spent her summer as a summer research associate at IBM Research in Yorktown Heights, NY. She worked in the Physics of Information Group, where she developed, with her mentors David DiVincenzo and Barbara Terhal, a lattice layout for quantum computers. She simulated quantum gates on this layout to determine failure rates of components of quantum circuits. In addition to the internship, she also co-organized the "IBM Workshop on Fault-tolerant Quantum Computing", which took place in August at IBM Research. This fall, she is continuing to work with her IBM colleagues on other projects in quantum computation.

Alejandro Troccoli spent his summer as an intern at Microsoft Research (MSR) in Redmond . Alejandro worked with Sing Bing Kang, a member of the Interactive Visual Media group, on image based rendering from multiple images acquired with varying exposure. This was a great chance to escape from New York City (and its hot summer), walk the mountains, swim in the lakes and watch the whales.

that fosters innovation the same way startups do?" Gates pointed out that many of the remaining "grand challenges" of technology innovation (such as speech recognition) require a level of skill and a investment in long term research that are only possible at a major organization like Microsoft. Several of the students who asked questions were Gates Millenium Scholars who took the opportunity to publicly thank Bill Gates for his philanthropy. Other student questions dealt with topics such as identity theft and the role of open source software, as well as some lighthearted questions including "can you get me one of those Xbox 360s?" and "will you help me with my computer science homework?" (the answer: "I'd love to but it might be cheating.")

The afternoon closed with Dean Zvi Galil presenting a panoply of gifts to Bill Gates, including an official pair of SEAS cufflinks. All in all, it was a memorable day for computer science at Columbia.

Feature Article

Women In Computer Science (WICS) at Columbia

In our field a woman's face is as hard to find as a fix for a buggy program that is due in an hour. Women in Computer Science (WICS) is a group that is changing this situation by bringing together women involved in computer science in the Columbia University community. WICS hopes to encourage prospective computer science majors and provide direction and assistance to current computer science students. As a network for women undergraduates, graduate students, faculty and staff, the group promotes interaction on social, academic and professional levels.

WICS hosts social events, corporate information sessions, and educational functions primarily catered to women but also open and often very relevant to men as well. Among WICS' most popular events are faculty panels that address topics relevant to undergraduate and graduate students such as "Giving Talks," "Chosing Research Topics," and "Job Options After Graduate School." Naturally, one of the most important aspects of WICS is the forum it provides to address women's concerns, questions, ideas and goals. To this end WICS organizes book discussions on topics specific to women (such as "Women Don't Ask: Negotiation and the Gender Divide," by Linda Babcock and Sara Laschever and "Women in Science: Meeting Career Challenges," by Angela M. Pattatucci, ed.), and recently participated in a university-wide panel on Sexual Harrassment. WICS also facilitates access to mentorship opportunities and other educational and professional resources. For example, two WICS student members participated in the CRA-W 2005 Grad Cohort Workshop held in San Francisco, and some students were partly funded to attend the Grace Murray Hopper Conference. WICS also designed and distributed a survey in an effort to identify the main difficulties faced by women in our department. Other recent activities include a visit to Google's office in Times Square, Welcome and End-of-Semester parties, and joint WICS/CS movie nights.

WICS Executive Board. Standing, from left to right: Elena Filatova, Knarig Arabshian, Julia Hirschberg, Lauren Wilcox. Front, from left to right: Dan Phung, Ani Nenkova, Marta Arias.

WICS is also very active outside the scope of the computer science department. It is currently part of a SEAS-wide group, WICSE, which has funding to support activities of all women students in SEAS. In a joint effort with Carnegie Mellon University, WICS organized a dinner with distinguished CMU visitors and graduate students and faculty from NYU, Stevens, Hunter, and NJIT, funded by CMU. The dinner was followed by a CRA-W Distinguished Lecture given by Dr. Lenore Blum of CMU and lunch sponsored by CRA-W with a panel on graduate research and careers (hosted by Dr. Joann Ordille, Avaya Labs). The activities finished with a presentation by CMU students "Exploring Graduate School: The Women @SCS Outreach Roadshow".

WICS is currently planning a Fall welcome party, a student/faculty panel to address hidden barriers to success in academia, and a distinguished lecture by a local woman scientist.

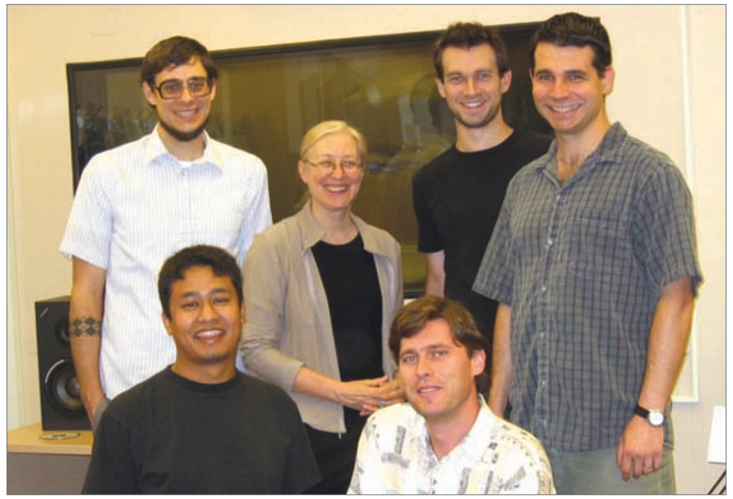
WICS' activities are funded by the Computer Science Department, Google, and by the Vice-Provost for Diversity and by the SEAS Dean Zvi Galil.

WICS' Executive Board includes faculty, research staff, graduate and undergraduate students.

In summary, WICS is an active and visible voice in the computer science community that provides both women and men with the opportunity to organize and participate in activities seeking to improve the university experience as a whole. For more information you can visit the WICS website at http://www.cs.columbia.edu/wics or contact WICS by email at cu-wics@cs.columbia.edu.

WICS Executive Board. From left to right: Sarah Friedman, Emily Stolfo, Stephanie Lee.

The **Natural Language** and **Speech Processing** Groups



Front row, from left to right: Sameer Maskey, Dr. Stefan Benus. Back row: Jackson Liscombe, Professor Julia Hirschberg, Andrew Rosenberg, Frank Enos. Not pictured: Augustin Gravano, Noemie Elhadad, Dr. Mona Diab, Dr. Martin Jansche.

Natural language and speech processing deal with the theory and practice of enabling computers to understand and produce language, both written and spoken.

Human-computer interfaces based on language understanding and generation have the potential to make online information universally accessible to the general public.

At Columbia, the Natural Language and Speech Processing Groups are addressing problems in analyzing and providing access to large amounts of information, whether spoken or written. We are developing browsing and summarization systems for news (both published written sources and broadcast news), email, journal articles and meetings. Our research in spoken language processing also aims at identifying deceptive speech and emotions. We are also developing techniques for multilingual analysis, whether spoken or written.

The group includes Prof. Julia Hirschberg and Prof. Kathleen McKeown, research scientists Martin Jansche and Owen Rambow, post-docs Stefan Benus, Mona Diab, Nizar Habash, and Advaith Siddharthan, and consultant Rebecca Passonneau. 14 PhD students and two programmers are also part of the group.

Summarization

Our research addresses automatic summarization of a wide range of genres, from formal written language (e.g., medical journal articles) to informal, conversational speech (e.g., meetings).

Newsblaster

Current technology in summarization and topic detection and tracking is mature enough to be used reliably in a live, online environment. A number of people in our group have developed Newsblaster, a system that crawls news sites, filters out news from non-news (e.g., ads), groups news into stories on the same event, and generates a summary of each event. Newsblaster generates summaries by extracting important sentences and by generating new sentences that fuse information from different documents. News is automatically categorized into five main categories (US, World, Entertainment, Finance and Sports) and within each category, news stories are automatically clustered into stories on the same event. It tracks events across days and can generate an update, identifying what's new on a given day in comparison to previous days. We have also developed a multilingual version of Newsblaster, which crawls foreign language news sites and generates English summaries of documents on the same event drawn from many languages. Newsblaster runs on a daily basis, crawling 25 news sites (see http://newsblaster.cs.columbia.edu).

Broadcast News

In our work at Columbia summarizing Broadcast News, we have pursued a **two-level** approach to the problem of summarizing errorful spoken material: First, we identify domain-specific aspects of newscasts to provide an **outline** of the newscast, which users can navigate in a GUI interface, following links from e.g. headlines to stories and speakers to the speech they contribute. In domains like Broadcast News, the material to be summarized exhibits fairly regular patterns from one speech document to another: news broadcasts generally open with a news anchor's introduction of the major news stories to be presented in the broadcast, followed by the actual presentation of those stories by anchor, reporters, and possibly interviewees, and are usually concluded in a fairly conventionalized manner as well. We are identifying key elements that appear in any broadcast, using a combination of acoustic, prosodic, lexical, and structural features obtained from the news transcript and from the original speech. Second, we use similar features to extract portions of news stories to serve as summaries. Thus a newscast can be searched or browsed, to locate stories of interest, and these stories can subsequently be summarized for the user.

Email

We are working on a system to summarize threads of email conversations, i.e., coherent exchanges of email messages among several participants. Our email summarization interface is a system for on-the-fly processing of email conversations that can be seamlessly integrated into a user's existing email client such as Microsoft Outlook. Various functionalities have been identified and implemented, including summarization of individual emails and email threads, and categorization of emails. The summarization system finds key questions posed within the thread and their corresponding answers and integrates these with sentences that it determines contain important information. To find these sentences, it uses a machine learning approach based on features related to email structure (e.g., a sentence that occurs immediately following quoted sentences). We have recently begun developing techniques that allow summarization of an entire mailbox, in addition to thread summarization.

Bruce Abramson (BA '83, MS '85, Ph.D. '87) writes: "My first book intended for a general audience, Digital Phoenix: Why the Information Economy Collapsed and How it Will Rise Again (MIT Press, 2005), is now available through fine booksellers everywhere. Digital Phoenix is my attempt to make sense of the major information economy events of the past decade—in particular, the Internet investment bubble, the Microsoft trial, the rise of open source, and the P2P wars. In doing so, I draw upon some results that those of us well-versed in artificial intelligence, software engineering, industrial organization, network economics, antitrust, and intellectual property law may know well, but that remain news to most of the rest of the world. I also spend a chapter discussing the relationship between the information economy and the broader social and political arenas that are beginning to feel the transition from industrial age to information age. Despite that potentially intimidating profile, I worked hard to make the material accessible to a general audience, and above all to make the book an entertaining read. I've collected early reviews on my website, at http://www.theinformationist.com/index/bruce/comments/reviews/."

Andrew Arnold (CC '03) is in his second year of Carnegie Mellon's Ph.D.program in machine learning. He reports that he is "enjoying Pittsburgh, but still pining for the city."

Regina Barzilay (Ph.D. '02) was selected as one of this year's top 35 technology innovators under the age of 35 by Technology Review. She is currently an assistant professor at the Computer Science and Artificial Intelligence Laboratory at MIT. Regina is being recognized for her work in teaching computers to read and write: as the Technology Review article reports, "For her doctoral dissertation at Columbia University, computer scientist Regina Barzilay led the development of Newsblaster, which does what no computer program could do before: recognize stories from different news services as being about the same basic subject, and then paraphrase elements from all of the stories to create a summary."

Jeannie Fromer (Barnard '96) is moving down to Washington, DC, where she will be clerk for Justice David H. Souter of the U.S. Supreme Court.

Stuart Haber (Ph.D. '88) is the General Chair of the 25th International Cryptology Conference (CRYPTO 2005), which is the main conference on cryptology held annually in the United States. Stuart was quoted in an August 17 article in the New York Times on cryptologists from China who were not granted visas in time to present their work at CRYPTO 2005: "It's not a question of them stealing our jobs... We need to learn from them, but we are shooting ourselves in the foot."

Kai-Fu Lee (BA '83) was the subject of a New York Times Week in Review article "The Key to Google's $10 Million Man." Lee, who graduated with highest honors from Columbia and holds a Ph.D. in computer science from Carnegie Mellon, is a former corporate vice president at Microsoft who is now heading Google's operations in China.

A student under the supervision of **Tom O'Donnell** of Siemens Corporate Research (entered Ph.D. Program 1990) has won the Best Student Paper in the Area of Segmentation and Processing in this year's MICCAI (Medical Image Computing and Computer Aided Intervention) conference. "Quantification of Delayed Enhancement Images" By Engin Dikici, Thomas O'Donnell, Randolph Setser, and Richard D. White was awarded a cash prize of 500 Euros.

Jonathan Rosenberg (Ph.D. '01), was profiled in Computer Reseller News as part of their annual list of the 10 most innovative technologists. Jonathan is one of the co-authors of the Session Initiation Protocol (SIP), a widely used telephone technology.

Montek Singh (Ph.D. '02) writes "I am now an Assistant Professor at UNC Chapel Hill, working on VLSI CAD, asynchronous circuits and systems, and graphics hardware. I was awarded an IBM Faculty Award last year, and very recently won a DARPA grant (headed by Boeing) to develop an industrial-strength CAD tool for automating the design of high-speed asynchronous digital systems. The DARPA work will be done jointly with Philips Semiconductors and Steve Nowick."

CUCSers and their families enjoy a barbecue on a beautiful fall day.

conference every eighteen months. The group is chaired by Prof. Don Towsley, UMass.

Professor **Steve Nowick** has been brought onto the $11M DARPA CLASS program, a major government initiative to make asynchronous digital design viable for the commercial and military sectors. There were 20 large-scale proposals submitted, and only one funded, headed by Boeing, with participation of Philips Semiconductors, two asynchronous startups and two smaller academic efforts. The two goals of the project are to build a large-scale asynchronous demonstration chip (for Boeing) and design an asynchronous CAD tool for use future asynchronous designs. Prof. Nowick and his former PhD student **Montek Singh** (currently an assistant professor at UNC), will play a key role in transferring their high-speed asynchronous pipeline style, MOUSETRAP, to the Philips commercial asynchronous tool flow, and providing optimizations for several of the other CAD tools.

PhD student **Rafi Pelossof** and senior research scientist **Yoav Freund**, both at the Center for Computational Learning Systems (CCLS), presented their work on their state-of-the-art "particle filtering tracking" system for traffic and pedestrian scenes in a full-day tutorial at the CVPR (Computer Vision and Pattern Recognition) conference, held in San Diego in June.

Ph.D. student **Dan Phung** won the Best Student Paper Award for "Adaptive Internet Interactive Team Video", co-authored by Dan Phung, **Giuseppe Valetto** and **Gail Kaiser**, at the 4th International Conference on Web-based Learning (ICWL 2005), August 2005.

Professor **Henning Schulzrinne** was interviewed about the i2hub file sharing system by ABC Nightline and by Fox News.

Professor **Henning Schulzrinne** won the 2005 Sputnik Innovator Award for his contributions to VoIP. Prof. Schulzrinne received the award at the 2005 forward2business conference in Halle, Germany.

Professor Angelos Keromytis and his wife Elizabeth Doran confront an applied security problem in the Bahamas.

Professor **Henning Schulzrinne**, along with a team consisting of researchers from Pennsylvania State University, University of California-Santa Barbara and Lucent Technologies, won a National Science Foundation grant titled WORKIT: A Universal Wireless Open Research KIT. The WORKIT project addresses the need for wireless network tools and platforms as recommended in the 2003 NSF Wireless Network Workshop report. The project will build on the IOTA (Integration of Two Access Technologies) project at Bell Labs. The PI's will enhance and develop IOTA for a software and systems package in a distributable form called the Wireless Open Research Kit (WORKIT). WORKIT will include source code and documentation and also be embodied in low-cost off the shelf hardware. WORKIT will be an enabler for research in mobility management, interlayer awareness, software algorithms for optimal network selection, reconfiguration, security, accounting, authentication, policy download and enforcement, and hybrid wireless networking. Broader impacts of this project include use of WORKIT in education and enabling stronger university/industry collaborations in this area of emerging importance at colleges and universities.

Professor **Rocco Servedio** was elected as a board member of the Association for Computational Learning; the members of the board comprise the steering committee for the Conference on Computational Learning Theory (COLT).

Professor **Rocco Servedio** was awarded a grant from the NSF program on Emerging Models and Technologies for Computation (EMT). The EMT cluster seeks to advance the fundamental capabilities of computer and information sciences and engineering by capitalizing on advances and insights from areas such as biological systems, quantum phenomena, nanoscale science and engineering, and other novel computing concepts. The award will support Rocco's research on connections between quantum computation and computational learning theory.

Two DHS grants have been awarded to CounterStorm and Columbia's IDS Lab, headed by Professor **Sal Stolfo**. This project aims to research and develop a new generation of collaborative, cross-domain security technologies to detect and prevent the exploitation of network-based computer systems. The core concept is to deploy a number of strategically placed sensors across a number of participating networks that collaborate by sharing information in real-time to defend the entire network and each of its members. A novel content-based anomaly detector, PAYL, identifies likely new exploits targeting vulnerable systems. The Worminator project has developed a new generation of scalable, collaborative, cross-domain security systems that exchange alert information including profiled behaviors of attacks and privacy-preserving anomalous content alerts to detect severe zero-day security events. The work is a joint collaboration with CounterStorm, a New York City based company spun out from the DHS and DARPA-sponsored Columbia IDS lab.

Professor **Joseph Traub** was named as the new chair of the Computer Science and Telecommunications Board (CSTB) of the National Academies. The CSTB deals with critical issues facing the nation in the area of computer science and telecomuniations. Projects include cybersecurity research, biometrics, IT to enhance disaster management, and building certifiably dependable systems. For more information, visit www.cstb.org. Prof. Traub's appointment marks his return to the CSTB, as he was also its founding chair. "In 1986, along with Marjory Blumenthal, Joe's vision and dedication established the model that has made CSTB one of the strongest boards at the Academies. At this particular point in CSTB's history, I could not think of another person better suited to assume the chair and to guide CSTB to new heights," said Bill Wulf, President of the National Academy of Engineering.

Medical Journal Articles and Textbooks

We have developed a system to summarize documents (i.e., journal articles, textbooks and online consumer health) as part of a medical digital library being developed at Columbia University, called PERSIVAL (PErsonalized Retrieval and Summarization of Images, Video and Language), which aims to provide tailored presentation of the relevant medical literature for both physicians and lay consumers. The PERSIVAL summarization component takes as input documents relevant to a user's query, retrieved by a search component. It generates an English summary of one or more paragraphs of the group of documents, highlighting facts that are common to all documents and pointing out differences between them. A key feature of the PERSIVAL summarizer is the ability to personalize its content given information about patient status available in the online patient record; the result is a summary that highlights information that is more likely to be relevant to the user, whether it is the patient or a physician treating the patient.

Meetings

Meetings involve multi-party conversation with overlapping speakers; the language is informal and utterances tend to be partial, fragmentary, ungrammatical and include many ellipses and pronouns. At Columbia, we are working on meeting summarization as part of a larger project entitled *Mapping Meetings* where the goal is to create methods for effectively recognizing, browsing and visualizing meetings. Our work to date has focused on methods for identifying important content and for generating the sentences of a summary. We have developed techniques for segmenting the meeting into different topics, for recognizing agreement and disagreement in dialog, and we are currently working on methods for generating summary sentences that can remove disfluencies and extraneous material.

Analyzing Spoken Language

Deceptive Speech

We are currently examining the feasibility of automatic detection of deception in speech, using lexical, prosodic, and acoustic cues. Our current study creates a context in which the subject is positively motivated to deceive an interviewer, where deceptive and truthful statements are marked by the subject. The interview is recorded and features of the subject's speech extracted for statistical analyses and machine learning experiments that attempt to find those features best able to distinguish truth from lie.

Emotional Speech

This research seeks to identify acoustic and prosodic cues to human emotions, based on subjective judgments of emotion as well as more 'objective' eye-tracking studies of subjects asked to match audio data with human faces which display different emotional states. The goals of these studies are first, to identify those acoustic and prosodic features of an utterance which are most characteristic of utterances judged to represent different emotions using machine learning techniques, and second, to gain additional insight into how humans identify emotion in speech and which emotions appear most closely related to one another.

Charismatic Speech

People at an instinctual level are drawn to certain public speakers. What about it makes their speech charismatic? Our research examines acoustic and lexical features of politicians' speech in a variety of contexts to identify the acoustic, prosodic and lexical indicators that best correlate with subjects' judgments of charisma. Though our work to date has been on American English, parallel work on Palestinian Arabic is being conducted to identify cultural differences in the perception of charisma.

Arabic Dialect Processing

Three members of the Natural Language Processing Group are part of the Center for Computational Learning Systems where they are concentrating on issues relating to Arabic and its dialects. The Arabic language is actually a collection of dialects with phonological, morphological, lexical, and syntactic differences comparable to those among the Romance languages. However, throughout the Arab world, the standard written language is the same, Modern Standard Arabic (MSA), which is also used in some official spoken communication (newscasts, parliamentary debates). MSA is based on Classical Arabic and is itself not a native spoken language. Other forms of Arabic (generally referred to as "dialects" of MSA) are what people use for daily spoken communication. In unofficial written communication, in particular in the now growing electronic media, often ad-hoc transcriptions of dialects are used. The boundaries between MSA and a dialect are not well defined and borrowing between the two forms occurs often.

Within the Arabic-dialect family, MSA is the variant with most available resources in terms of corpora (monolingual text, English-Arabic parallel texts, and Treebanks), morphological analyzers, etc. There are far fewer resources available for most other dialects; since they are mainly spoken, even unannotated corpora are not common.

The focus of our current research on Arabic Dialects is to create NLP tools for Arabic dialects, including resource-poor dialects. Our approach is to create frameworks that exploit the commonalities among the dialects. One resource we are creating is the Pan-Arab Morphological Analyzer (PAMA). Initially, we will populate it with morphological knowledge for MSA and Egyptian, but hope to be able to extend this work to other dialects (Levantine, Iraqi, Moroccan, Yemeni, and so on). We plan to extend initial work we have done on automatic morphological disambiguation for MSA to the dialects, using PAMA as the core component. Moreover, PAMA is being used as a core component in a machine translation (MT) component which approximates dialect-to-MSA translation by intentionally overgenerating MSA hypotheses given dialect input.

Front row, from left to right: Ani Nenkova, Dr. Becky Passonneau, Lokesh Shrestha, Professor Kathy McKeown, Smara Muresan, Jen-Yuan Yeh, Elena Filatova, Dr. Barry Schiffman. Back row: Aaron Harnly, Sasha Blair-Goldensohn, David Elson, Michel Galley, Dr. Nizar Habash, Dr. Owen Rambow.

Recent & Upcoming **PhD Defenses**

David Evans

Advisor:
Kathy McKeown

Identifying Similarity in Text: Multilingual Analysis for Summarization

Abstract: Early work in the computational treatment of natural language focused on summarization and machine translation. In my research I have concentrated on the area of summarization of documents in different languages. This thesis presents my work on multilingual text similarity. This work enables the identification of short units of text (usually sentences) that contain similar information even though they are written in different languages. I present my work on SimFinderML, a framework for multilingual text similarity computation that makes it easy to experiment with parameters or similarity computation and add support for other languages. An in-depth examination and evaluation of the system is performed using Arabic and English data. I also apply the concept of multilingual text similarity to summarization in two different systems. The first improves readability of English summaries of Arabic text by replacing machine translated Arabic sentences with highly similar English sentences when possible. The second is a novel summarization system that supports comparative analysis of Arabic and English documents in two ways. First, given Arabic and English documents that describe the same event, SimFinderML clusters sentences to present information that is supported by both the Arabic and English documents. Second, the system provides an analysis of how the Arabic and English documents differ by presenting information that is supported exclusively by documents in only one language. This novel form of summarization is a first step at analyzing the difference in perspectives from news reported in different languages.

Shlomo Hershkop

Advisor:
Sal Stolfo

Behavior-based Email Analysis with Application to Spam Detection

Abstract: Email is the "killer network application". Email is ubiquitous and pervasive. In a relatively short timeframe, the Internet has become irrevocably and deeply entrenched in our modern society primarily due to the power of its communication substrate linking people and organizations around the globe. Much work on email technology has focused on making email easy to use, permitting a wide variety of information and information types to be conveniently, reliably, and efficiently sent throughout the Internet. However, the analysis of the vast storehouse of email content accumulated or produced by individual users has received relatively little attention other than for specific tasks such as spam and virus filtering. As one paper in the literature puts it, "the state of the art is still a messy desktop".

The Problem: Email clients provide only partial information—users have to manage much on their own making it hard to search or prioritize large amounts of email. Our thesis is that advanced data mining can provide new opportunities for applications to increase email productivity and extract new information from email archives.

This thesis presents an implemented framework for data mining behavior models from email data. The Email Mining Toolkit (EMT) is a data mining toolkit developed for the purposes of mining email data. EMT was designed to analyze offline email corpora, including the entire set of email sent and received by an individual user revealing much information about individual users as well as the behavior of groups of users in an organization. A number of machine learning and anomaly detection algorithms are embedded in the system to model the user's email behavior in order to classify email for a variety of tasks. The work has been successfully applied to the task of classification, spam detection, and forensic analysis to reveal information about user's behavior.

We organize the core functionality of EMT into a package called the Profiling Email Toolkit (PET). A novel contribution in PET is the focus on analyzing real time email flow information from both an individual and an organization in a standard framework. PET includes new algorithms that combine multiple models using a variety of features extracted from email to achieve higher accuracy and lower false positive than any one individual model for a variety of analytical tasks.

Gaurav KC

Advisor: Al Aho

Defending Software Against Process-Subversion Attacks

Abstract: We address the significant problem of protecting computing resources against attack mechanisms commonly utilised by malware such as Internet worms and Internet cracking tools. These attacks typically exploit security vulnerabilities in applications, middleware, system libraries, and even the underlying operating system, to break in and compromise the targeted software. We analysed the anatomy of common process-subversion attacks, and derived from it a model for describing and evaluating defence techniques that thwart such attacks. We have incorporated this model into a software tool called esBench, which can be used to perform security evaluations of defence techniques in implemented systems.

We illustrate the comprehensiveness of our model and the power of esBench by evaluating three defence techniques, Casper, RiSA, and e-NeXSh. We designed these techniques to create a secure software-execution environment to counter the problem of process-subversion attacks. These techniques build defensive support within application code or the underlying operating system, and are incorporated into specific components of the x86/Linux run-time environment, including the compiler, system libraries, the operating system, and the underlying hardware. These techniques extend the run-time environment in different manners with capabilities to automatically detect and thwart a variety of attack mechanisms. Our techniques are general and practical, and can be applied to other systems.

Simon Lok

Advisor:
Steve Feiner

Automated Layout of Information Presentations

Abstract: *Layout* refers to the process of determining the sizes and positions of the visual objects that are part of an information presentation. *Automated layout* refers to the use of a computer program to automate either all or part of the layout process. This field of research lies at the crossroads of artificial intelligence and human computer interaction. Automated layout of presentations is becoming increasingly important as the amount of data that we need to present rapidly overtakes our ability to present it manually.

We present a set of novel techniques that can assist an automated layout system to produce an effective presentation of information, given a set of components to display and metadata describing relationships between those components. Unlike the vast majority of previous approaches to layout and numerous related problems, our techniques attempt to mimic the workflow used by a human graphic designer when manually creating a presentation.

Our techniques include a method to control the length of text being

research staff involved include Professor **Gail Kaiser**, Dr. **Christina Leslie**, Professor **Ken Ross**, Dr. **David Waltz** and Professor **Yechiam Yemini**. Funded through the National Centers for Biomedical Computing (NCBC) program, a component of the National Institutes of Health Roadmap for Medical Research, the National Center for the Multiscale Analysis of Genomic and Cellular Networks (MAGNet) will address this challenge through the application of both knowledge-based and physics-based methods. The Center will provide an integrative computational framework to organize molecular interactions in the cell into manageable context dependent components. Furthermore, it will develop a variety of interoperable computational models and tools that can leverage such a map of cellular interactions to elucidate important biological processes and to address a variety of biomedical applications.

Professor **Angelos Keromytis** chaired the Workshop On Rapid Malcode (WORM) held in Washington, DC, in November.

Research scientist **Christina Leslie** (of the Center for Computational Learning Systems) was awarded a five-year NIH R01 grant of $1.4M entitled "Recognizing protein folds with discriminative machine learning". The project is to build a system for predicting the structural class of a protein from its sequence, using string kernel support vector machine technology developed by the Leslie lab.

Spring PhD recipient **Simon Lok** and his company Lok Technology were featured in a May Entrepreneur magazine article "The Kid's Got It".

Professor **Tal Malkin** and her group were prominently mentioned in congressional testimony on identity theft. On September 22, Bruce E. Bernstein, President of the New York Software Industry Association (NYSIA), testified in writing to the U.S. Senate Committee on Banking, Housing and Urban Affairs during a Hearing on "Examining the Financial Services Industry's Responsibilities and Role in Preventing Identity Theft and Protecting Sensitive Financial Information", mentioning Prof. Malkin's project analyzing the security configuration of TLS-protected servers. Part of the testimony read: "Dr. Malkin and her team made a systematic study of the cryptographic strength of thousands of "secure" servers on the Internet... Dr. Malkin's study probed 25,000 secure Web servers to determine if SSL was being properly configured and whether it was employed in the most secure way. Improper configuration can lead to attacks on servers, stolen data identity theft, break-ins, etc. Dr. Malkin's project is the most extensive study of actually existing server security on the Internet. The team's findings, relevant to these hearings, included some serious weaknesses in how Web servers, including eCommerce servers employed by financial service companies, are currently being configured... These security shortcomings are quite serious, and pose risks both to the consumers and the providers in the financial services industry. Financial server security can be increased both by popularizing the correct configurations and, possibly, by greater government oversight in this area."

Amelie Marian (Ph.D. '05) started as a tenure-track assistant professor at Rutgers University, in the Computer Science Department, in Fall 2005.

The Henry and Gertrude Rothschild Chair in Computer Science was donated by Henry Rothschild and Gertrude Neumark Rothschild. Gertrude is the Howe Professor Emerita of Materials Science in the Fu Foundation School of Engineering. Pictured are Professor Rothschild and Professor Kathleen McKeown, the recipient of the chair.

The Natural Language Processing (NLP) research group led by Professor **Kathy McKeown** and Professor **Julia Hirschberg**, together with Professor Ellis (EE) and researchers in the Center for Computational Learning Systems (CCLS), has won a large DARPA grant for their "Novel Information Gathering and Harvesting Techniques in a Global Autonomous Environment (NIGHTINGALE)" project. The grant was awarded to a team lead by SRI and consisting of researchers at Columbia University, University of Massachusetts Amherst, University of California San Diego, University of California Berkeley, University of Washington, Technical University Aachen (Germany), and Systran. The research to be conducted at the Center for Computational Learning Systems (CCLS) will center on building natural language processing tools for Arabic and its dialects, concentrating on leveraging linguistic knowledge when few resources (annotated corpora or even unannotated corpora) are available. **Mona Diab, Nizar Habash,** and **Owen Rambow** will build on work accomlished under an existing NSF grant. In addition, Nizar Habash will continue his work on generation-heavy hybrid machine translation.

Professors **Vishal Misra** and **Dan Rubenstein** were recently elected to the IFIP Working Group 7.3 on Computer Performance Modeling and Analysis. The WG accepts new members every two years and includes the most recognized experts in that field. The work of the Group is directed toward improving the art of analyzing and optimizing performance and costs of data processing systems through the use of analytical models. The group plays an important and active role in fostering education and research in these areas. It organizes or coorganizes a

On October 17, a reception was held in honor of the recent establishment of several endowed chairs in the Computer Science Department. Seated from left are Kathleen McKeown, Henry and Gertrude Rothschild Professor in Computer Science; Shree Nayar, T.C. Chang Professor in Computer Science; Mihalis Yannakakis, Percy K. and Vida L. W. Hudson Professor in Computer Science; and Alfred Aho, Lawrence Gussman Professor in Computer Science.

CS concentrator **Scott Brinker** (GS '05) was the 2005 Valedictorian for the School of General Studies.

Bogdan Caprita (MS '05) was the Valedictorian of the Columbia Class of 2005. He completed a BS in Applied Mathematics and Computer Engineering and an MS in Computer Science in four years.

The database research group hosted the first DB/IR Day at Columbia University on April 15, 2005 to bring together researchers in database and information retrieval. More than 120 researchers and students from academic and research institutions across the greater New York area attended this inaugural workshop, making it a very successful event. The program consisted of three technical keynote lectures from Alon Halevy (University of Washington), Craig Nevill-Manning (Google Inc.) and Michael Stonebraker (MIT), and a poster session for graduate students to present their latest research. The event was sponsored by IBM research, with additional funding from Columbia's Graduate Student Advisory Council.

Sebastian Enrique, **Alpa Shah**, **Mark Threshock**, **Eugene Ie**, **William Beaver**, **Abhinav Kamra**, and **Joshua Weinberg** were named as “extraordinary TAs” for the 2004 fall semester, based on the evaluation of students in their classes.

Professor **Steve Feiner**'s lab was featured in a March Technology Review article on augmented reality, and an August Knight-Ridder syndicated news story on applications of GPS. Professor Feiner received a grant from the Air Force Research Lab to fund research on augmented reality for maintenance assistance, and during the spring, gave invited talks in the MIT CSAIL HCI Seminar Series and at INRIA Futurs in Orsay, and a keynote talk at the GIS Planet Conference in Estoril, Portugal.

Dean **Zvi Galil**, Julian Clarence Levi Professor of Mathematical Methods and Computer Science and Dean of the School of Engineering and Applied Science, was elected as a Fellow of the American Academy of Arts and Sciences.

The paper “Modeling and Managing Content Changes in Text Databases,” by **Panos Ipeirotis** (a Fall 2004 Columbia PhD graduate, now an assistant professor at NYU), Alexandros Ntoulas (a PhD student at UCLA), Junghoo Cho (an assistant professor at UCLA), and Professor **Luis Gravano**, won the Best Paper Award at the 21st IEEE International Conference on Data Engineering (ICDE) 2005 conference held April 2005 in Tokyo. ICDE is a highly selective and prestigious database conference.

Professor **Eitan Grinspun** has won a large National Science Foundation grant to investigate simulating physical systems at multiple resolutions. The work will be performed by researchers at Columbia University, led by Eitan, and Caltech and was funded jointly by the NSF math and computer science directorates. The proposal is titled “Computational and Mathematical Foundations for the Synthesis of Multi-resolution Representations with Variational Integrators and Discrete Geometry”.

PhD student **Suhit Gupta** received the Best Student Poster Award at WWW 2005 (together with Professor **Gail Kaiser** and Professor **Sal Stolfo**) for his poster “Extracting Context To Improve Accuracy For HTML Content Extraction”.

Professor **Julia Hirschberg** was elected as president of the International Speech Communication Association (ISCA). The International Speech Communication Association (ISCA) is the major international organization devoted to speech science and technology, with approimately 1500 members. ISCA runs annual conferences which draw 1000-1300 participants. The main goal of the Association is “to promote Speech Communication Science and Technology, both in the industrial and Academic areas”, covering all the aspects of Speech Communication (Acoustics, Phonetics, Phonology, Linguistics, Natural Language Processing, Artificial Intelligence, Cognitive Science, Signal Processing, Pattern Recognition, etc.).

Professors **Julia Hirschberg** and **Henning Schulzrinne** both received the IBM Faculty Award.

The Machine Learning Laboratory, led by Professor **Tony Jebara**, recently received two one-year awards from the KDD Program (a joint effort between the Intelligence Community and the NSF). The first was a KDD Program Award, made in February, for a proposal on “Correspondence in Learning via Permutation Algorithms.” The second was a KDD Challenge Award, made in August, for a proposal on “Text and Author Identity as a Permutation Learning Problem.”

The National Institutes of Health (NIH) has announced that Columbia will host one of the three NIH Roadmap National Centers for Biomedical Computing awarded in 2005. The National Center for Multi-Scale Study of Cellular Networks, MAGNet, at Columbia will be focusing on the comprehensive mapping and analysis of molecular cellular interactions. Columbia Computer Science faculty and

displayed that is inspired by the capability of a designer to ask a copy editor for a change in the content length to suit a particular layout. We have also developed a method for automatically evaluating and improving the visual balance of a layout that closely mirrors the process used by graphic designers. Finally, we have developed a method to deliver layouts produced by our techniques using a thin-client approach.

Our techniques are uniquely capable of capturing the irregular and sometimes counterintuitive tactics used by human designers to create layouts. When combined with a model of the workflow used by human graphic designers, our techniques allow systems to automatically generate layouts that are highly effective, superior in both form and function to those generated by investigations into this field to date.

Amelie Marian

Advisor: Luis Gravano

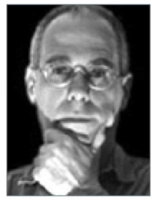
Evaluation of Top-k Queries over Structured and Semi-structured Data

Abstract: Traditionally, query processing strategies for structured (e.g., relational) and semi-structured (e.g., XML) data identify the “exact matches” for the queries. This exact-match query model is not appropriate for many database applications and scenarios where queries are inherently fuzzy—often expressing user preferences and not hard Boolean constraints—and are best answered with a ranked, or “top-k,” list of the best matching data objects.

In this thesis, we present a variety of top-k query processing algorithms. For efficiency, these algorithms focus on the objects that are most likely to be in the top-k query answers, and discard —as early as possible—objects that are guaranteed not to qualify for the answers. We center the discussion around a number of important scenarios, each presenting fundamental challenges for top-k query definition and processing. Specifically, we first focus on a common web-application scenario where the data is structured and only available through autonomous web services with heterogeneous access interfaces and constraints. By taking into account the peculiarities of the sources and potentially choosing a different query execution plan for each individual candidate object, our adaptive algorithms significantly reduce the query processing time over previously existing algorithms, which select coarser “global” query execution plans.

We also discuss an important XML data integration scenario where XML data comes from heterogeneous sources, and therefore may not share the same schema. In this scenario, exact query matches are too rigid, so XML query answers should be ranked based on their “similarity” to the queries in terms of both content and structure. Processing top-k queries efficiently in such a scenario is challenging, as the number of candidate answers increases dramatically with the query size. By pruning irrelevant data fragments as early as possible, our algorithms minimize the number of candidate answers considered during query evaluation.

In summary, this thesis discusses the general problem of ranking query answers over structured and semi-structured data, and returning the “best” objects, according to user preferences, in a time-efficient manner.



Barry Schiffman

Advisor: Kathy McKeown

Learning to Identify New Information

Abstract: This thesis is an investigation into a new problem in natural language processing: new-information detection. It is a similar task to first-story detection, but with a very large difference. First-story detection operates on the document level, while new-information detection is on the statement level. In its fundamental guise, new-information detection is the ability for a machine to be able to compare two textual statements and decide whether they say the same thing or not. But the task is complicated by the fact that each new statement must be tested against all previous statements.

In this thesis, I show that the sentence is a poor choice of syntactic unit for this task since sentences are arbitrarily composed of one or more structures. Thus, the system must do a deeper syntactic analysis of the inputs than recognizing sentence boundaries. At the same time, I found that context is important, and I developed a mechanism to look beyond sentence boundaries for evidence of novelty. Thus, the system I developed considers a mixture of features, from a micro perspective, looking within sentences, and from a macro perspective, looking beyond sentence boundaries. I apply machine learning techniques to combine the features coherently into a unified hypothesis for the problem, using rule induction.

The system is designed to function in a multi-document summarization system, like Columbia's Newsblaster, for which it produces update summaries focusing on the developments of the day in an event that has interested the public over a period of several days. The new-information system provides all the novel statements to the DEMS summarizer, which I had previously built for Newsblaster, for the final selection of material.

The system also includes a semantic unit that improves performance a bit. The system uses a plugin lexicon that is largely taken from the WordNet data at present.

Xiaotao Wu

Advisor: Henning Schulzrinne

Ubiquitous Programmable Internet Telephony End System Services

Abstract: In Internet telephony, end systems usually have CPU and memory, so they are programmable and can perform services. This is very different from traditional communication networks, which usually assume dumb endpoints. The enhanced capabilities of end systems make services more distributed, and introduce many new services. However, most of the existing service research focus only on services in the network, which are not well-suited for end systems. Therefore, it is important to conduct research specifically on end system services, including the definition of end system services, how to develop an efficient and friendly end system service creation environment, how to handle feature interactions, and how to integrate telephony services with other Internet services, such as web, email, location-based services, and networked appliance control.

In this dissertation, I analyzed the differences between end system services and network services. The analysis motivated me to define a new scripting language called the Language for End System Services (LESS) specifically for end system service creation. Compared with the existing service creation techniques, LESS is simple, safe, portable, extensible, and can directly interact with users and directly control media streams.

LESS has a tree-like structure, which makes it easy to handle feature interactions among LESS scripts. I proposed a LESS-based feature interaction handling method by using action conflict tables and a tree merging algorithm. Once a potential feature conflict is detected, my solution can also help to resolve the detected conflict.

The tree-like structure also suggests a way to automatically generate services by performing decision tree induction. Since users may not be aware of what services are available and not know how to customize or create their own services, automatically generating services based on

Recent & Upcoming **PhD Defenses** (continued)

users' communication behaviors can greatly help users to create their desired services. I developed a service learning system that uses the Incremental Tree Induction (ITI) algorithm to generate decision trees for call handling. The decision trees can then be translated to LESS scripts. Auto-generated services may sometimes perform unwanted actions that may cause users to lose calls, money, or privacy. I name these kinds of problems service risks. In this dissertation, I also did an investigation on how to avoid or reduce service risks in the service learning system.

Since location-based services are critical to Internet telephony, especially for emergency call handling, I also did a comprehensive analysis on location-based services in Internet telephony. I will present the location-based services implementation in our lab environment. In addition, I will introduce our prototype implementation on emergency call handling in SIP-based Internet telephony systems. Besides, I will also discuss our SIP-based global-scale ubiquitous computing architecture. The architecture uses Service Location Protocol (SLP) to find available resources based on the location information, and uses SIP third-party call control architecture to control the available resources.

I have integrated my research on end system services into our SIP user agent—SIPc. In addition to basic SIP functions and an end system service execution environment, SIPc also supports many other Internet functions, such as service discovery, event notification, networked appliance control, instant messaging, and session announcement handling. Multiple functions in SIPc can interact with each other to provide new services. In this dissertation, I will discuss different ways for multi-function integration and interactions in SIPc.

New **Faces**

Elli Androulaki graduated from National Technical University of Athens with a Diploma in Electrical and Computer Engineering in June 2005. She worked at Delta Cingular Corporation for the construction of the Athens 2004 Olympic Games webpage, and started as a Ph.D. student in Fall 2005. Elli's main research interests are network security and peer to peer database systems. She is currently working with Profesor Keromytis's and Professor Bellovin's group at the network security lab.

Marta Arias has an undergraduate degree from the Polytechnic University of Catalunya (Barcelona, Spain) and a Ph.D. from Tufts University (Massachusetts, USA) both in Computer Science. She moved to the Center for Computational Learning Systems at Columbia in September 2004. Her interests are in Machine Learning and Computational Biology. She is involved in several projects in applied machine learning ranging from predicting genetic regulation in the fruitfly to predicting failures in New York City's electric grid.

Omer Boyaci graduated from Bilkent University with a B.S. in Computer Science in June 2003. He did a M.S. in computer science at Columbia from 2003-2004. He worked at Avaya Labs Research in summer 2005 and started as a Ph.D. student in Fall 2005. Omer's main research interests are Multimedia Networking, Wireless Networking and Software Engineering. He is currently working with Professor Schulzrinne's group studying Multimedia Networking and VoIP.

Seung Geol Choi graduated from Seoul National University with a B.S. in Computer Engineering in February 1999. He worked at Bluebird Soft Co. in Korea from 1999-2002, did an M.S. in Computer Science and Engineering at Seoul National University from 2003-2005, and started as a Ph.D. student in Fall 2005. Seung Geol's main research interests are cryptographic protocols and applying cryptographic techniques (e.g., bilinear mapping based cryptography) to problems arising in privacy issues. He is also interested in other aspects of theoretical cryptography and the interaction of cryptography and secure systems. He is currently working with Professor Moti Yung.

Rebecca L. Collins graduated from the University of Tennessee at Knoxville with an M.S. in Computer Science in December 2004. She began Columbia's Ph.D. program in Fall 2005. Her research interests include design and scheduling of distributed systems, and erasure codes. She is currently working with Professor Nieh's group studying client to client communications in thin client environments and with Professor Carloni exploring the cycle balancing problem in latency insensitive systems. She is currently supported by a three-year NDSEG Fellowship.

Mona Diab received her PhD in Computational Linguistics in 2003, from the University of Maryland College Park. Her PhD work focused on solving problems of ambiguity in natural language semantics exploiting multilingual evidence. She did her postdoctoral work at Stanford University working on issues of Arabic Natural Language processing. She then joined Columbia University in Feb 2004. Her research interests include computational lexical semantics, machine learning, computational psycholinguistics and comparative linguistics. As an Associate Research Scientist at the Center for Computational Learning Systems (CCLS), Dr. Diab's current work addresses problems of processing and modeling Arabic dialects specifically and resource poor languages generally. For more info check http://www.cs.columbia.edu/~mdiab.

David K. Elson graduated from Columbia University with a B.S. in Computer Science in May 2002. He remained on campus and worked in Prof. McKeown's Natural Language Processing Group as a programmer/analyst from 2002-2005. He was also a part-time student during this time, receiving his M.S. in Computer Science in May 2005. David's work is in computational narratology, which synthesizes his interests in AI and creative writing. He is working on his project with Prof. McKeown as well as professors and students in the School of the Arts.

Jinwei Gu graduated from Tsinghua University in China with a B.S. in June 2002 and a M.S. in June 2005, both in Department of Automation. He was a visiting student in Microsoft Research Asia from 2003 to 2004, and started as a Ph.D. student in the Computer Science Department in Columbia University in Fall 2005. His main research interests are Model-based Graphics and Vision, Machine Learning, Image Analysis, and Biometrics. He is currently working with Profesor Shree Nayar, Professor Peter Belhumeur and Prof. Ravi Ramamoorthi studying Bidirectional Texture Function and its application to realistic rendering.

Nizar Habash received his PhD in 2003 from the Computer Science Department, University of Maryland College Park. He is an Associate Research Scientist at the Center for Computational Learning Systems in Columbia University. His research interests include machine translation, natural language generation, lexical semantics, and combining rule-based and corpus-based resources for natural language processing. He is currently working on computational modeling of Arabic dialects and Arabic-English Machine Translation. For more information, visit his website at http://www.NizarHabash.com.

Neeraj Kumar graduated from Georgia Tech in May 2005 with a double degree in Computer Science and Aeronautical Engineering, and he started as a Ph.D. student in computer science at Columbia in Fall 2005. He is in the vision and graphics group working with Professor Shree Nayar on topics related to computer vision and image processing. He has interned in past summers at Microsoft Research and NVIDIA corporation.

Kevin Matulef visited the computer science department for the summer of 2005. While at Columbia he worked with Ronitt Rubinfeld and Rocco Servedio on algorithms and lower bounds for property testing of Boolean functions. Kevin is a third year Ph.D. student in theoretical computer science at MIT.

Ronitt Rubinfeld visited the computer science department for the summer of 2005 (her second summer in a row at Columbia). While at Columbia she worked with Kevin Matulef and Rocco Servedio on algorithms and lower bounds for property testing of Boolean functions. Ronitt is a Professor of Electrical Engineering and Computer Science at MIT.

Pannagadatta K. Shivaswamy graduated from University of Mysore with a B.E. in Computer Science in 2001. He worked at Cisco Systems and Yahoo! India before joining Indian Institute of Science in 2003 for an M.E. in Systems Science and Automation. Pannagadatta's main research interests are in Machine Learning and its applications. He is currently working with Prof Tony Jebara and Dr. Martin Jansche. His current work focuses on applications to power and permutation invariance in Support Vector Machines.