(Message from the Chair continued from page 1)

All such changes probably need to be explained by multiple factors, but the extensive coverage in the popular press of the PC and, for the second crest, the Internet may well have motivated many students to choose this "hot" major—and parents and guidance counselors probably did their part. Now that biomedical and nano technologies dominate the newspaper science and technology section and computer science-related topics tend towards the not-so-pleasant side effects of a networked world, this external push into computer science is unlikely to return. Thus, instead of relying on others to tell our story, we need to convince students that computer science remains the hub of technology, offering opportunities to contribute in a large variety of human endeavors, from the arts to finance and computing technology.

Traditionally, with some empirical justification, undergraduate education was seen by the outside as training for writing code, except for the small percentage of students continuing on to advanced graduate study and research. In our local experience, our graduates have much more diverse career paths, often straddling multiple disciplines. I believe that, as a discipline, we need to do a better job of understanding where our students go, what they do and why students choose, or do not choose, computer science. There will be misunderstandings to correct, e.g., about job prospects, but there are also likely to be opportunities to make a computer science education more relevant, as versatile as a liberal arts education, but with better prospects. In the years to come, for example, we anticipate that the increasing emphasis on computational methods for finding new drugs will further strengthen the need for computer science talent in the New York metropolitan area, a major focal point for pharmacological research. It remains to be seen whether students perceive a better value in a CS major, or combining CS as a minor with another subject matter. Recent indications are that there is strong growth in areas that CS programs have traditionally paid little attention to, such as system and network administration.

The transition to smaller class sizes affords us the opportunity to take a closer look at what we could be doing. We intend to be part of such explorations and experiments and invite others to share their impressions and experiences with us.

You can keep up with Columbia Computer Science news as it happens, by subscribing to our mailing list at *http://lists.cs.columbia.edu/mailman/listinfo/cucs-news*. Our nascent departmental blog, with topics of general interest in computer science resides at *http://columbiacs.blogspot.com*. Alumni can find our portal at *http://alum.cs.columbia.edu*, where they can look up fellow alumni, see job listings received by the Department and share their current whereabouts and activities. We always appreciate hearing from our former students, faculty and staff.

Henning Schulzrinne
Professor and Chair
(hgs@cs.columbia.edu)

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

CS@CU

NEWSLETTER OF THE DEPARTMENT OF COMPUTER SCIENCE AT COLUMBIA UNIVERSITY VOL.1 NO.4 SPRING 2005

Students Athena Ledakis, Ron Coleman, Josef Bryks Schenker, Akshay Kumar, and Justin Titi with the Muddrover robot that they built and programmed in Stephen Edwards's CSEE 4840 "Embedded System Design" course.

See pages 2-3 for other highlights of student projects from this course.

Message from the Chair

Henning Schulzrinne, Professor & Chair

The end of the academic year is a good opportunity to reflect on the year past. We are proud of our graduating class of 2005, consisting of 55 seniors and 82 masters students, in addition to 14 PhD students who defended their theses throughout the year.

We are delighted to be able to welcome Prof. Steve Bellovin as a new member of our faculty. Prof. Bellovin strengthens our security and networking areas, adding expertise in cryptography, network security protocols and other networking topics.

Under the stewardship of our new CS@CU newsletter editor, Prof. Rocco Servedio, we have added a few new sections to the newsletter that I hope you'll find useful. We introduce new faces in our department on pgs.10-11 and summarize recent dissertations on pgs.4-5. The Department has published technical reports since before its official existence. To provide a more detailed look at some of our current research, this issue of the newsletter summarizes some of the recent reports published by our faculty, research staff and students. Interested readers can subscribe to a mailing list announcing new reports by visiting our department web site.

One of the advantages of being part of an institution that has a smaller undergraduate student body is that we can offer project-oriented courses. We profile one such course in this newsletter (pgs. 2-3), where students use an FPGA board to develop custom logic, creating a variety of (mostly) useful, but always challenging, designs.

Our Programming Systems lab, introduced on pg.6, has a tradition of producing software-related research with practical applications. Our Database group, introduced on pg. 7, is engaged in a wide range of projects including interdisciplinary work with scholars in anthropology, earth and environmental sciences, art history and other areas.

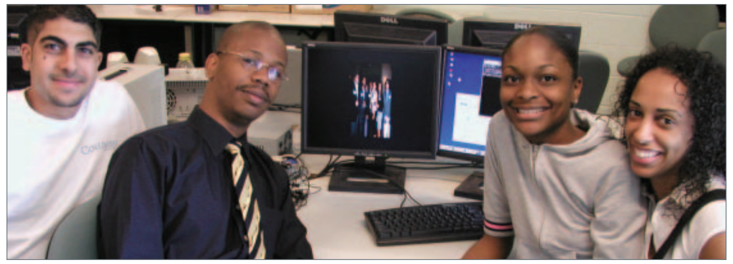
Like everywhere else, our undergraduate enrollment has decreased significantly since its peak in 2002. We draw most of our undergraduates from the School of Engineering and Applied Science, where the overall number of admitted students has remained largely constant. We saw large increases in the fraction of undergraduates choosing computer science in the mid-1980s, and then again starting around 1997.

(continued on page 16)

Highlights of CSEE 4840: Embedded System Design

In spring 2004, Professor Stephen Edwards introduced "CSEE 4840: Embedded System Design" as a replacement for ELEN E3940. **The main focus of the class is an independent group project on the design and implementation of a small system involving hardware and software.**

Scott Arfin, Gabriel Glaser, and Ron Weiss with their Terrormouse system.



Essa Farhat, Eveliza Herrera, Rhonda Jordan and Amon Wilkes with their Thing-a-ma-Flipper Video Effects Generator.

Students implemented their projects on an FPGA board (the XSB-300E) from XESS Corporation. This board contains a Xilinx Spartan IIE FPGA (an XC2S300E) capable of holding both a 32-bit RISC microprocessor core (a "Microblaze") and quite a lot of student-designed custom logic. Virtually every project incorporated a combination of C code running on the processor and custom logic written in the VHDL hardware description language.

Below is a list of selected projects from the spring 2004 class that illustrates the breadth and complexity of what the students were able to achieve. Professor Edwards stated that "I was very impressed with what these students could create in half a term."

Scott Arfin, Gabe Glaser, and Ron Weiss designed "Terrormouse," a MIDI synthesizer that exemplifies hardware/software codesign. They decoded the high-level MIDI stream, which consists mostly of note on/note off events, in software and used this information to control twelve oscillators, six based on the Karplus-Strong string synthesis algorithm, the other six based on FM synthesis. By changing the FM synthesis parameters, such as modulation frequency and depth, they implemented ten different "patches" (sounds) ranging from a pure sinewave to a metallic-sounding organ.

The large "Muddrover" group, which consisted of Ron Coleman, Akshay Kumar, Athena Ledakis, Josef Bryks Schenker, and Justin Titi, executed an ambitious project which integrated a Lego Mindstorms-based robot, a video camera, and custom hardware and software to produce a robot that could run laps around a black line on white paper. They took the raw video signal in through the Philips SAA7019 video decoder chip on the XSB-300E board, processed it, divided the screen into ninths, and used information about how much black was in each area to decide whether to advance or turn the robot. These commands were fed out a serial port to an IR tower to control the Lego robot. Prof. Sklar lent the group the video-camera and extensive advice.

Essa Farhat, Eveliza Herrera, Rhonda Jordan, and Amon Wilkes built the "Thing-a-ma-flipper" dynamic video effects generator. Their hardware was able to scale and distort a still video image stored in memory under software control. The result was a digital fun-house mirror.

Sangeeta Das, Waclaw Aleksander Godycki, Laxmikant Joshi, and Stephen Tarzia implemented a unique project involving an unusual peripheral: a formula-style racecar built with a space-frame chassis powered by a 600cc motorcycle engine. Their goal was real-time remote data acquisition from the car, letting them capture data such as engine RPM, throttle position, and manifold air pressure from a safe, quiet, still location. They bought an off-the-shelf RF

CUCS-045-04
Extracting Context To Improve Accuracy For HTML Content Extraction
Suhit Gupta, Gail Kaiser, and Salvatore Stolfo

CUCS-046-04
Design and Verification Languages
Stephen A. Edwards

CUCS-047-04
WebPod: Persistent Web Browsing Sessions with Pocketable Storage Devices
Shaya Potter and Jason Nieh

CUCS-048-04
End System Service Examples
Xiaotao Wu and Henning Schulzrinne

CUCS-049-04
Obstacle Avoidance and Path Planning Using a Sparse Array of Sonars
Matei Ciocarlie

CUCS-050-04
Remotely Keyed CryptoGraphics - Secure Remote Display Access Using (Mostly) Untrusted Hardware
Debra L. Cook, Ricardo Baratto, and Angelos D. Keromytis

CUCS-051-04
Sequential Challenges in Synthesizing Esterel
Cristian Soviani, Jia Zeng, and Stephen A. Edwards

CUCS-052-04
Determining Interfaces using Type Inference
Stephen A. Edwards and Chun Li

Alumni News

Galina Datskovsky Moerdler (PhD 1990), appeared on the cover of the March 2005 issue of KM World as the leader of one of the "100 companies that matter in knowledge management". Galina is the CEO of MDY Advanced Technologies.

Jonathan Katz (PhD 2002) is currently an assistant professor at the University of Maryland. He was recently awarded a Faculty Early Career Development Award (CAREER) from the NSF, for a project titled "Models and Cryptographic Protocols for Unstructured, Decentralized Systems".

Eleazar Eskin (PhD 2002, now an Assistant Professor in Residence in Computer Science & Engineering at UCSD) writes: "We have a paper that just came out on the cover of Science. It involves the HAP project which is the main project in my group. It is the first publication of a whole genome map of human variation (with Perlegen Sciences) and we performed some of the analysis. We have another paper that is currently under review where we performed more analysis over the same data and a third in the pipe which is analysis over an even larger dataset."

Srinivasa Narasimhan (MA 1999, PhD 2003) joined the tenure-track faculty of the Robotics Institute, School of Computer Science, CMU, in Fall 2004. His research interests are in vision and graphics.

Frank Lai (BA 2003) writes, "I graduated in 2003 from FF-SEAS. I wasn't really sure what I wanted to do for a living, but I fondly remembered my experience working at a summer camp for academically-gifted children. So, I figured at the time that the logical next step would be to go into teaching Math, given my background in Computer Science. As a result, I applied to and was successfully accepted by Rutgers University's Graduate School of Education Masters program in Math Education. I enjoyed my experience there so much, that before I finished my Masters degree, I transferred into Rutgers's doctoral program in Math Education. To bolster my research capabilities, I am simultaneously working towards a Masters in Math over there as well. I am quite grateful to Columbia's Computer Science program for providing me with the strong analytical ability needed for what I am doing now."

Andrew Arnold (BA 2003) writes: "After receiving my degree in 2003, I moved to 86th and 2nd ave and began a programming job at Bloomberg. After nine months at Bloomberg, I moved to Google in Times Square and worked on their local search feature. In September 2004 I made the difficult, but rewarding, decision to leave my job, friends, and New York City to begin a PhD in the Center for Automated Learning and Discovery (CALD) at Carnegie Mellon University in Pittsburgh. My focus is on automatic feature extraction and I plan (hope) to be done by 2009, in time to move back to New York for the 2012 Olympics."

Eric Siegel (PhD 1998; Assistant Professor 1997-2001) offers business analytics and data mining services with San Francisco-based Prediction Impact (www.predictionimpact.com). His clients range from Fortune 100 companies down to small R&D think tanks. Eric is also happily married to Maria de Fatima Callou! Eric can be reached at eric@predictionimpact.com or (415) 385-1313.

Wenke Lee (PhD 1999) has been promoted to the rank of Associate Professor with tenure in the College of Computing at Georgia Institute of Technology.

Jim Kurose (PhD 1984) was presented with an award for "Exemplary Service to the Community" at Infocom, the biggest networking conference. The award was bestowed by the Technical Committee on Computer Communications (TCCC), a part of IEEE Communications Society. The other recipient was Henning Schulzrinne.

Michelle Zhou (PhD 1999) received the IUI 2005 (International Conf. on Intelligent User Interfaces) Outstanding Paper Award for "A Graph-Matching Approach to Dynamic Media Allocation in Intelligent Multimedia Interfaces" by Michelle X. Zhou, Zhen Wen, and Vikram Aggarwal. Michelle is a research staff member and manager at IBM TJ Watson Research Center.

2004 Technical Report Series

All reports are available at *http://www1.cs.columbia.edu/~library/2004.html*

CUCS-001-04
Automating Content Extraction of HTML Documents
Suhit Gupta, Gail Kaiser, Peter Grimm, Michael Chiang, and Justin Starren

CUCS-002-04
Secret Key Cryptography Using Graphics Cards
Debra L. Cook, John Ioannidis, Angelos D. Keromytis, and Jake Luck

CUCS-003-04
The Complexity of Fredholm Equations of the Second Kind: Noisy Information About Everything
Arthur G. Werschulz

CUCS-004-04
Feature Interactions in Internet Telephony End Systems
Xiaotao Wu and Henning Schulzrinne

CUCS-005-04
Secure Isolation and Migration of Untrusted Legacy Applications
Shaya Potter, Jason Nieh, and Dinesh Subhraveti

CUCS-006-04
Asymptotic bounds for MX/G/1 processor sharing queues
Hanhua Feng and Vishal Misra

CUCS-007-04
DotSlash: A Scalable and Efficient Rescue System for Handling Web Hotspots
Weibin Zhao and Henning Schulzrinne

CUCS-008-04
A Virtual Environment for Collaborative Distance Learning with Video Synchronization
Suhit Gupta and Gail Kaiser

CUCS-009-04
Optimizing Quality for Collaborative Video Viewing
Dan Phung, Giuseppe Valetto, Gail Kaiser, and Suhit Gupta

CUCS-010-04
Elastic Block Ciphers
Debra L. Cook, Moti Yung, and Angelos Keromytis

CUCS-011-04
Failover and Load Sharing in SIP Telephony
Kundan Singh and Henning Schulzrinne

CUCS-012-04
Collaborative Distributed Intrusion Detection
Michael E. Locasto, Janak J. Parekh, Sal Stolfo, Angelos D. Keromytis, Tal Malkin, and Vishal Misra

CUCS-013-04
When one Sample is not Enough: Improving Text Database Selection Using Shrinkage
Panagiotis G. Ipeirotis and Luis Gravano

CUCS-014-04
MobiDesk: Mobile Virtual Desktop Computing
Ricardo Baratto, Shaya Potter, Gong Su, and Jason Nieh

CUCS-015-04
Improved Controller Synthesis from Esterel
Cristian Soviani, Jia Zeng, and Stephen A. Edwards

CUCS-016-04
Jitter-Camera: High Resolution Video from a Low Resolution Detector
Moshe Ben-Ezra, Assaf Zomet, and Shree K. Nayar

CUCS-017-04
Blurring of Light due to Multiple Scattering by the Medium, a Path Integral Approach
Michael Ashikhmin, Simon Premoze, Ravi Ramamoorthi, and Shree Nayar

CUCS-018-04
Host-based Anomaly Detection Using Wrapping File Systems
Shlomo Hershkop, Linh H. Bui, Ryan Ferster, and Salvatore J. Stolfo

CUCS-019-04
Exploiting the Structure in DHT Overlays for DoS Protection
Angelos Stavrou, Angelos Keromytis, and Dan Rubenstein

CUCS-021-04
Elastic Block Ciphers: The Feistel Cipher Case
Debra L. Cook, Moti Yung, and Angelos Keromytis

CUCS-022-04
Orchestrating the Dynamic Adaptation of Distributed Software with Process Technology
Giuseppe Valetto

CUCS-023-04
On decision trees, influences, and learning monotone decision trees
Ryan O'Donnell and Rocco A. Servedio

CUCS-024-04
Efficient Algorithms for the Design of Asynchronous Control Circuits
Michael Theobald

CUCS-025-04
Efficient Shadows from Sampled Environment Maps
Aner Ben-Artzi, Ravi Ramamoorthi, and Maneesh Agrawala

CUCS-026-04
The Simplicity and Safety of the Language for End System Services (LESS)
Xiaotao Wu and Henning Schulzrinne

CUCS-027-04
THINC: A Remote Display Architecture for Thin-Client Computing
Ricardo A. Baratto, Jason Nieh, and Leo Kim

CUCS-028-04
Group Ratio Round-Robin: O(1) Proportional Share Scheduling for Uniprocessor and Multiprocessor Systems
Bogdan Caprita, Wong Chun Chan, Jason Nieh, Clifford Stein, and Haoqiang Zheng

CUCS-029-04
Cross-Dimensional Gestural Interaction Techniques for Hybrid Immersive Environments
Hrvoje Benko, Edward W. Ishak, and Steven Feiner

CUCS-030-04
Modeling and Managing Content Changes in Text Databases
Panagiotis G. Ipeirotis, Alexandros Ntoulas, Junghoo Cho, and Luis Gravano

CUCS-031-04
Using Execution Transactions To Recover From Buffer Overflow Attacks
Stelios Sidiroglou and Angelos D. Keromytis

CUCS-032-04
A Theoretical Analysis of the Conditions for Unambiguous Node Localization in Sensor Networks
Tolga Eren, Walter Whiteley, and Peter N. Belhumeur

CUCS-033-04
Voice over TCP and UDP
Xiaotang Zhang and Henning Schulzrinne

CUCS-034-04
Information Structures to Secure Control of Rigid Formations with Leader-Follower Structure
Tolga Eren, Walter Whiteley, Brian D.O. Anderson, A. Stephen Morse, and Peter N. Belhumeur

CUCS-035-04
An Investigation Into the Detection of New Information
Barry Schiffman and Kathleen R. McKeown

CUCS-036-04
Machine Learning and Text Segmentation in Novelty Detection
Barry Schiffman and Kathleen R. McKeown

CUCS-037-04
Live CD Cluster Performance
Haronil Estevez and Stephen A. Edwards

CUCS-038-04
Building a Reactive Immune System for Software Services
Stelios Sidiroglou, Michael E. Locasto, Stephen W. Boyd, and Angelos D. Keromytis

CUCS-039-04
An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol
Salman A. Baset and Henning Schulzrinne

CUCS-040-04
Programmable Conference Server
Henning Schulzrinne, Kundan Singh, and Xiaotao Wu

CUCS-041-04
Microrobotic Streak Seeding For Protein Crystal Growth
Atanas Georgiev, Peter K. Allen, Ting Song, Andrew Laine, William Edstrom, and John Hunt

CUCS-042-04
Preventing Spam For SIP-based Instant Messages and Sessions
Kumar Srivastava and Henning Schulzrinne

CUCS-043-04
Service Learning in Internet Telephony
Xiaotao Wu and Henning Schulzrinne

CUCS-044-04
Peer-to-Peer Internet Telephony using SIP
Kundan Singh and Henning Schulzrinne

transmitter/receiver pair, interfaced it with a PIC microcontroller, and designed and built their own analog signal conditioning hardware.

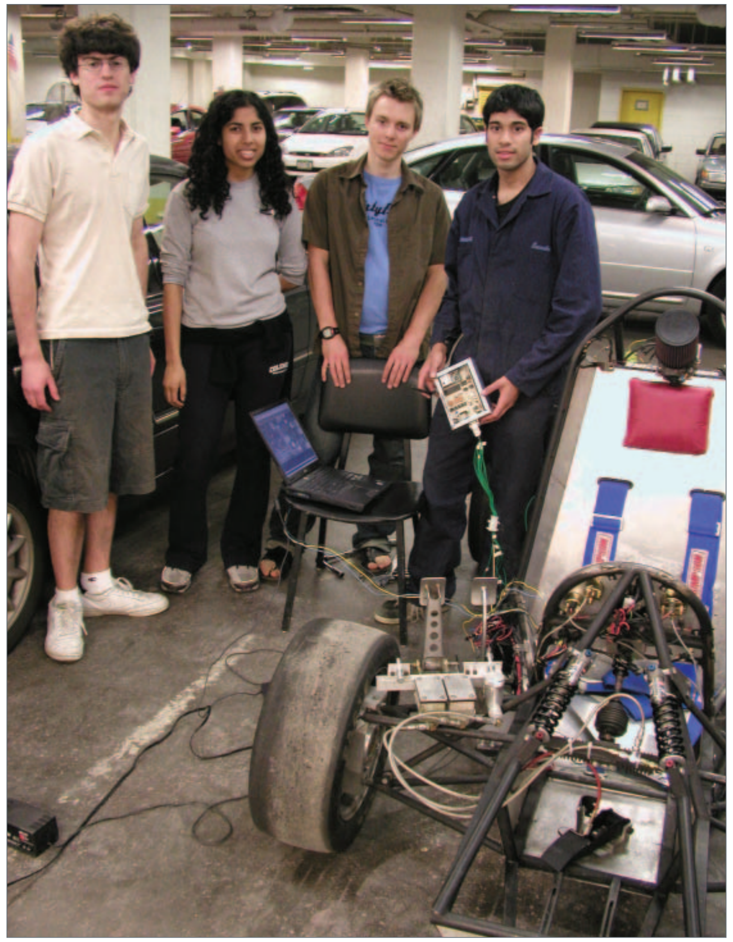
Ang Cui, Jeng-Ming Hwang, Yen Yen Ooi, Kashif Siddiqui, and Ting-Hsiang Wu implemented an idea based on work by Professor Shree Nayar: extracting depth information from a video camera pointed at a mirror. Their project took the video from the camera, identified two spots on the image that were projected by a laser pointer onto a piece of paper, and determined how far away the point was using parallax. They revised their initial mirror configuration after reading Shree's paper on the subject.

Two groups, one consisting of Charles Finkel, Dagna Harasim, and David Soofian, the other of Winston Chao, Eric Li, and Ke Xu, each implemented a video game based very loosely on Pac-Man. The display used a combination of characters for the maze and sprites (overlaid graphic objects) for the characters. The game logic was implemented in C.

Yaniv Schiller, Avrum Tilman, and Joshua Weinberg built "JAYcam," a system that took real-time black-and-white video, reduced its resolution, and transmitted it over a 100 Mbit/s Ethernet link to a simple Java program that displayed it on the screen. This project implemented several different, complicated standards: NTSC video, Ethernet, and UDP.

Philip Coakley, Tecuan Flores and Joshua Mackler pushed the computing power of the XESS board by implementing a real-time graphical spectrum analyzer able to perform a real-time 2048-point fast Fourier transform (FFT) on a pair of 48 kHz audio signals, bin the results, and display them graphically. Most of the FFT was done in software, but they implemented a very fast complex multiplier peripheral to speed up the inner loop.

Students are developing a whole new crop of projects for the spring 2005 edition of the class, and they promise to be just as interesting as these.


Sangeeta Das, Waclaw Aleksander Godycki, Laxmikant Joshi and Stephen Tarzia with their DAQ-T Student Vehicle Telemetry System.

New **Faculty**

Computer Science Department Welcomes **Professor Steven Bellovin**

Professor Steven Bellovin

The Computer Science department is delighted to welcome Steven M. Bellovin as a Professor of Computer Science.

Professor Bellovin joined the Columbia faculty in 2005 after many years at Bell Labs and AT&T Labs Research.

Bellovin received a B.A. degree from Columbia University, and an M.S. and PhD in Computer Science from the University of North Carolina at Chapel Hill. While a graduate student, he helped create USENET; for this, he and the other perpetrators were award the 1995 Usenix Lifetime Achievement Award. The Usenix Lifetime Achievement Award recognizes and celebrates singular contributions to the Unix community in both intellectual achievement and service. USENET was an experiment started in 1979 to create an electronic bulletin board to facilitate the posting and reading of news messages and notices. Today it has more than 10,000 discussion groups, known as newsgroups, on a wide variety of subjects, tens of thousands of USENET sites, and many millions of participants. Bellovin joined AT&T Bell Laboratories in 1982, where he became an AT&T Fellow. He was elected a member of the National Academy of Engineering in 2001.

Bellovin is the co-author of "Firewalls and Internet Security: Repelling the Wily Hacker", and holds several patents on cryptographic and network protocols. He has served on many National Research Council study committees, including those on information systems trustworthiness, the privacy implications of authentication technologies, and cybersecurity research needs; he was also a member of the information technology subcommittee of an NRC study group on science versus terrorism. He was a member of the Internet Architecture Board from 1996-2002; he was co-director of the Security Area of the IETF from 2002 through 2004.

Recent **PhD Defenses**

Tiberiu Chelcea

Design and Optimization of Large-Scale Asynchronous and Mixed Timing Systems

Abstract: Modern VLSI design is rapidly moving towards building large-scale and mixed-timing systems, which contain multiple synchronous domains, as well as asynchronous domains. However, two key issues are not completely addressed in these systems: the performance and scalability in communication between mixed-timing domains, and the synthesis of large-scale asynchronous systems. This thesis presents solutions to each of the two challenges in modern VLSI design. First, a complete set of mixed-timing FIFO's is introduced to interface any combination of synchronous and asynchronous systems. These FIFO's exhibit low latency, high throughput in communication, and have low area overheads. In addition, these FIFO designs have been adapted to solve the issue of long interconnect delays between mixed-timing domains. The second contribution of the thesis is an optimizing back-end, Balsa-CUBE, for a large-scale asynchronous synthesis system. The proposed back-end incorporates a set of new peephole and resynthesis optimizations. To facilitate the optimizations, a new asynchronous modeling language is also introduced. Balsa-CUBE incorporates a number of original CAD tools, as well as several existing academic and industrial CAD programs. Experimental results on several asynchronous systems show speed improvements of up to 56%.

Giuseppe Valetto

Orchestrating the Dynamic Adaptation of Distributed Software with Process Technology

Abstract: Software systems are becoming increasingly complex to develop, understand, analyze, validate, deploy, configure, manage and maintain. Much of that complexity is related to ensuring adequate quality levels to services provided by software systems after they are deployed in the field, in particular when those systems are built from and operated as a mix of proprietary and non-proprietary components. That translates to increasing costs and difficulties when trying to operate large-scale distributed software ensembles in a way that continuously guarantees satisfactory levels of service. A solution can be to exert some form of dynamic adaptation upon running software systems: dynamic adaptation can be defined as a set of automated and coordinated actions that aim at modifying the structure, behavior and performance of a target software system, at run time and without service interruption, typically in response to the occurrence of some condition(s). To achieve dynamic adaptation upon a given target software system, a set of capabilities, including monitoring, diagnostics, decision, actuation and coordination, must be put in place. This research addresses the automation of decision and coordination in the context of an end-to-end and externalized approach to dynamic adaptation, which allows to address as its targets legacy and component-based systems, as well as new systems developed from scratch. In this approach, adaptation provisions are superimposed by a separate software platform, which operates from the outside of and orthogonally to the target application as a whole; furthermore, a single adaptation possibly spans concerted interventions on a multiplicity of target components. To properly orchestrate those interventions, decentralized process technology is employed for describing, activating and coordinating the work of a cohort of software actuators, towards the intended end-to-end dynamic adaptation. The approach outlined above, has been implemented in a prototype, code-named Workflakes, within the Kinesthetics eXtreme project investigating externalized dynamic adaptation, carried out by the Programming Systems Laboratory of Columbia University, and has been employed in a set of diverse case studies. This dissertation discusses and evaluates the concept of process-based orchestration of dynamic adaptation and the Workflakes prototype on the basis of the results of those case studies.

Jingren Zhou

Architecture-Sensitive Database Query Processing

Abstract: During the last decade, microprocessors have experienced tremendous improvement. This architectural growth has not been equally distributed over all aspects of hardware performance. Recent advances in the speed of commodity CPUs have far outpaced advances in memory latency. Main memory access is therefore becoming a significant cost component of database operations. Database systems face new performance bottlenecks, such as memory access and poor utilization of sophisticated execution hardware. Research has shown that the DBMS hardware behavior is suboptimal, compared with scientific workloads. This illustrates the importance of rethinking and developing database query processing algorithms in the context of new computer architectures. This thesis focuses on studying interactions between DBMMs and modern processor architecture. We design techniques to exploit architectural innovations and to alleviate performance bottlenecks throughout the CPU and the memory hierarchy, including caches, memory, and disks. The first part of this thesis presents a novel technique to implement database operations with a high degree of intra-instruction parallelism. We describe new SIMD instructions commonly available in commodity processors and show how database operations can be accelerated using SIMD instructions. Database research has demonstrated that the dominant memory stalls are due to the data cache misses on the second-level cache and the instruction cache misses on the first-level instruction cache. We address both issues in the second part of this thesis. We propose buffering techniques to improve the data cache performance of index structures and to improve the instruction cache performance of query processing in database systems. The main benefit derives from better data and instruction reference locality. Our techniques can be easily integrated into current database systems without significant changes. The final part of this thesis describes a new storage model called MBSM (Multi-resolution Block Storage Model) for laying out tables on disks. Compared to other data storage models, MBSM has both good I/O performance and good cache utilization in main-memory.

Panagiotis Ipeirotis

Classifying and Searching Hidden-Web Databases

Abstract: Many valuable text databases on the web have non-crawlable contents that are "hidden" behind search interfaces. Hence, traditional search engines do not index this valuable information. One way to facilitate access to "hidden-web" databases is through Yahoo!-like directories, which organize these databases manually into categories that users can browse. An alternative way is through "metasearchers," which provide a unified query interface to search many databases at once. As part of my thesis, I have developed QProber, a system to automatically categorize and search autonomous, hidden-web databases. To categorize a database, QProber uses

Eitan Grinspun is organizing the first-ever course on Discrete Differential Geometry at this year's SIGGRAPH conference. SIGGRAPH is the premier conference in computer graphics, with an annual attendance between 20,000 and 40,000.

Joseph Traub participated in a site visit to the University of Waterloo for the Natural Sciences and Engineering Research Council of Canada (NSERC). The purpose of the site visit was to evaluate a major proposal on quantum computing by a alliance of seven Canadian universities, several government partners, and a group of industrial partners.

Henning Schulzrinne was presented with an award for "Exemplary Service to the Community" at Infocom, the biggest networking conference. The award was bestowed by the Technical Committee on Computer Communications (TCCC), a part of IEEE Communications Society. The other recipient was **Jim Kurose**, Henning's advisor and a Columbia Computer Science PhD alumnus.

Sebastian Enrique, **Alpa Shah**, **Mark Treshock**, **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. Congratulations to these outstanding TAs!

CCLS Researchers Applying Machine Learning at ConEd

CCLS Director
David Waltz

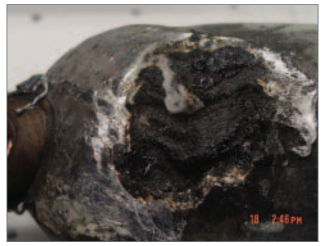
CCLS Research Scientist **Phil Long**

PhD student
Phil Gross

A team of Columbia researchers, with members from the Lamont-Doherty Earth Observatory, the Center for Computational Learning Systems, and the Computer Science Department, has received an $820K grant from Consolidated Edison to support joint research with ConEd engineers. These funds support research in 2005; plans are in the works for 2006 and beyond.

The group is working on applying and extending machine learning techniques to identify what equipment is most likely to fail next. "We're like the Department of Pre-Crime from that movie Minority Report," says Roger Anderson of LDEO, "our job is to catch the failures before they happen." Successful predictions can save money by allowing needed maintenance to be done during normal working hours. Also, by enabling technicians to clear incipient faults, they reduce the chance of cascading failures leading to outages. The team is also working on estimating the long-term survivability of cables, to inform prioritization in an ongoing cable replacement program.

In addition to Anderson, the team includes Albert Boulanger of LDEO, David Waltz and Phil Long of CCLS, and Phil Gross of the Computer Science Department. The project started when Anderson and Boulanger, working with the ConEd R&D department, identified a variety of problems where machine learning techniques might be profitably applied. In 2004 they started working on one of them, the failure prediction problem, using indications of likely failure uncovered in separate studies by ConEd researchers and consultant David Allen. Anderson and Boulanger then brought in Waltz and Long, and all collaborated on the design and evaluation of a prediction method using the variables for which data was available. The team has found that a machine learning technique called boosting, developed by CCLS member Yoav Freund and Robert Schapire of Princeton, has stood out as particularly useful for this application. Gross joined the effort this year; among other things, he is applying ideas from his PhD thesis on processing data from distributed remote sensors. These sensors will be used to generate real-time features for learning algorithms.



Failure at the lead wipe region of a overhead lead splice.

Professor **Al Aho** was named to the Lawrence Gussman Professor Chair and Professor **Kathy McKeown** was named to the Henry and Gertrude Rothschild Professor Chair. Both are recognized for their distinguished contribution to computer science, their service to the profession, the University, and the School.

Angelos Keromytis was featured in a WWORTV (Channel 9 News) piece on computer identity theft and wireless networking security.

Angelos Keromytis and **Moti Yung** are co-chairing the Third Applied Cryptography and Network Security (ACNS) Conference, to be held at Columbia June 7-10. **John Ioannidis** of the Columbia CCLS is the general co-chair. The ACNS conference brings together industry and academic researchers interested in the technical aspects of cryptology and the latest advances in the application of crypto systems; see http://acns2005.cs.columbia.edu for details.

One of **Henning Schulzrinne**'s developments, commercialized by SIPquest, was recognized as an Internet Telephony 2004 "Product of the Year". This development greatly reduces the hand-over delay in WiFi wireless networks for voice applications.

The largest German news magazine, Der Spiegel, did a photo feature about Internet telephony including **Henning Schulzrinne** and **Dorgham Sisalem** (one of his former students).

PhD student **Simon Lok** was featured in a Forbes magazine "Mavericks" profile article, "(Not So) Simple Simon." Lok heads his own company, Lok Technology, that makes secure networking devices.

Steven Bellovin has been named to the Department of Homeland Security's Science and Technology Advisory Committee.

Bhargav Bhatt, **Bogdan Caprita** and **Aaron Roth** were recognized with the Computing Research Association's (CRA) Outstanding Undergraduate Award for 2005. Aaron and Bhargav received Honorable Mention, while Bogdan was one of only ten finalists in the United States. The winners will receive their award at an upcoming CRA conference. The CRA noted: "This year's nominees were a very impressive group. A number of them were commended for making significant contributions to more than one research project, several were authors or coauthors on multiple papers, others had made presentations at major conferences, and some had produced software artifacts that were in widespread use. Many of our nominees had been involved in successful summer research or internship programs, many had been teaching assistants, tutors, or mentors, and a number had significant involvement in community volunteer efforts. It is quite an honor to be selected as one of the top members of this group."

Prof. **Steven Nowick** is general chair of the 11th IEEE Async Symposium, which was hosted at Columbia on March 13-16, 2005. Async-05 Symposium is the top symposium on advances in asynchronous (i.e., clockless) circuits and systems. The symposium typically has 100-120 attendees, and over 60 submitted papers. This year's invited speakers include Turing award-winner Ivan Sutherland with Robert Drost (Sun Microsystems Lab), Bob Colwell (the former Intel manager of several Pentium projects), and a tutorial on high-speed clocking with Prof. Ken Shepard (EE Department) and Phil Restle (IBM TJ Watson).

Prof. **Steven Nowick** was one of the invited speakers at Washington University's symposium, "Clockless Computing: Coordinating Billions of Transistors," to honor both the University's 150th Anniversary and the 30th anniversary of the completion of the seminal project on Macromodule Computer Design, work that anticipated current endeavors to go clockless, or asynchronous. Other invited speakers included Turing Award winner Ivan Sutherland, the founder of computer graphics, and Wes Clark, the designer of the world's first personal computer.

A proposal from the Columbia Robotics Lab was chosen as one of ten winners for the CanestaVision 3D sensing design competition. Columbia PhD student **Matei Ciocarlie** and Research Scientist **Andrew Miller** headed the proposal which focuses on developing an "Eye-in-Hand" range sensor for robotic grasping. Each of the winners will receive a $7,500 development kit that consists of a CanestaVision 3D sensor chip, a USB interface, and application program interface (API) software. These hardware and software development kits will be used to actually build the applications, and enter them in the "implementation" phase of the contest which boasts a $10,000 first prize for best use of the technology. Stay tuned for the Phase II winners in June!

Ravi Ramamoorthi won a National Science Foundation CAREER award to support his research into mathematical and computational fundamentals of visual appearance for computer graphics. The CAREER award is a highly-competitive early-career grant for new faculty. Ravi's research focuses on the mathematical and computational fundamentals of visual appearance, seeking to understand the intrinsic computational structure of illumination, reflection and shadowing, and develop a unified approach to many problems in graphics and vision.

MIT Technology Review's online edition features a story about Georgia Tech Assistant Prof. **Blair MacIntyre** (PhD '99) and his research on developing augmented reality systems. The article also includes commentary from Prof. **Steve Feiner**.

John R. Kender was invited to give the Keynote Address to the Sixth Annual ACM International Workshop on Mulitmedia Information Retrieval (MIR04), in October 2004. His talk was entitled, "The Psychology of MIR".

Ravi Ramamoorthi and **Rocco Servedio** received Alfred P. Sloan Research Fellowships. Prof. Ramamoorthi's research focuses on developing the mathematical representations and computational models for the visual appearance of objects, digitally recreating the complexity of natural appearance. Prof. Servedio's research interests are in computational learning theory, with strong interests in computational complexity theory, quantum computation, randomized algorithms, computational biology, cryptography, and combinatorics. The Sloan Research Fellowships are intended to enhance the careers of the very best young faculty members in specified fields of science. There were 14 of these fellowships awarded in Computer Science in 2005; the only institutions with two or more computer science fellowship recipients this year were Columbia, University of Washington, and Carnegie Mellon.

Yoav Freund of the Center for Computational Learning Systems and Robert Schapire of Princeton University won the Paris Kanellakis Theory and Practice Award for their contribution to highly accurate prediction rules used in web search engines. The Kanellakis Award honors specific theoretical accomplishments that significantly affect the practice of computing.

just the number of matches generated by a small number of query probes derived using state-of-the-art machine learning techniques. To search over "uncooperative" hidden-web databases, QProber exploits the database categorization to extract a small, topically-focused document sample from each database, from which a statistical summary of the database contents is produced. The content summaries can then be used during metasearching to select the most appropriate databases for a given query, a critical task for search scalability and effectiveness. Specifically, QProber identifies the most relevant databases for a query by exploiting both the database classification information and the extracted summaries. QProber produces high-quality database selection decisions, which in turn help return highly relevant search results. QProber also handles databases with contents that change over time. Using a "survival analysis" model, QProber predicts when content summaries need to be updated, and updates the summaries only when needed, avoiding overloading remote hidden-web databases unnecessarily.

Gong Su

MOVE: Mobility with Persistent Network Connections

Abstract: A key problem in today's mobile computing is how to preserve the ongoing network communication between two computation units when they move from one place to another. Because current network infrastructure and protocols are designed to support stationary endpoints only.

We have developed MOVE, a fine-grain end-to-end connection migration architecture, to address the problem. MOVE distinguishes itself by achieving, in a single system, several essential goals of a mobile communication architecture: (1) end system only without any infrastructure demand, transport protocol independence, and backward compatibility; (2) fine-grain connection migration and unlimited mobility scope; (3) secure migration with both handoff and suspension/resumption support; and (4) very low performance overhead both before and after migration.

We first analyze the key technical problems of end-to-end mobile communication: state inconsistency, conflict, and synchronization. We develop a simple and elegant abstraction called CELL, which provides a virtual, private, and labeled namespace for connection states so that they can be transparently migrated anywhere free of the problems mentioned above. We then develop a unique handoff protocol called H2O, which can handoff a connection securely in a single one-way end-to-end trip with minimal impact on the migrating connection. We finally integrate MOVE with a process migration mechanism to enable zero service disruption in proxy-based server clusters during a scheduled server maintenance.

We have implemented MOVE on a commodity OS without requiring any change to the OS and applications and conducted extensive performance measurements using both microbenchmarks and real world applications to validate the feasibility of MOVE.

Eugene Agichtein

Extracting Relations from Large Text Collections

Abstract: A wealth of information is hidden within unstructured text. Often, this information can be best exploited in structured or relational form, which is well suited for sophisticated query processing, for integration with relational database management systems, and for data mining. This thesis addresses two fundamental problems in extracting relations from large text collections: (1) portability—tuning extraction systems for new domains and (2) scalability—scaling up information extraction to large collections of documents. To address the first problem, we developed the Snowball information extraction system, a domain- independent system that learns to extract relations from unstructured text based on only a handful of user-provided example relation instances. Snowball can then be adapted to extract new relations with minimum human effort. Snowball improves the extraction accuracy by automatically evaluating the quality of both the acquired extraction patterns and the extracted relation instances. To address the second problem, we developed the QXtract system, which learns search engine queries that retrieve the documents that are relevant to a given information extraction system and extraction task. QXtract can dramatically improve the efficiency of the information extraction process, and provides a building block for extracting structured information and text data mining from the web at large.

Pablo Duboue

Indirect Supervised Learning of Strategic Generation Logic

Abstract: The Strategic Component in a Natural Language Generation (NLG) system is responsible for determining content (Content Selection) and structure (Document Structuring) of the generated output.

An implementation for the Strategic Component uses Content Selection rules to select the relevant knowledge and Document Structuring schemata to guide the construction of the document plan. This implementation is better suited for descriptive texts with a strong topical structure and little intentional content. In such domains, special communicative knowledge is required to structure the text, a type of knowledge referred to as Domain Communicative Knowledge.

In this thesis, we investigate the automatic acquisition of Content Selection rules and the automatic construction of Document Structuring schemata from an aligned Text-Knowledge corpus.

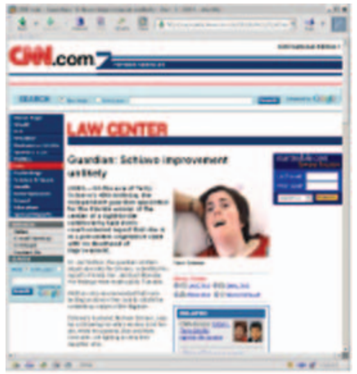
These corpora are a collection of human-produced texts together with the knowledge data a generation system is expected to use to construct texts similar to the human texts.

In two domains, medical reports and biographical descriptions, we have found aligned Text-Knowledge corpus for the learning task.

Aligned Text-Knowledge corpora only provide indirect information about the selected or omitted status of each piece of knowledge and their relative placement. Our methods involve Indirect Supervised Learning (ISL) as my proposed solution. ISL has two steps; in the first step, the Text-Knowledge corpus is transformed into a dataset for supervised learning, in the form of matched texts. In the second step, supervised learning machinery acquires the CS rules and schemata from this dataset.

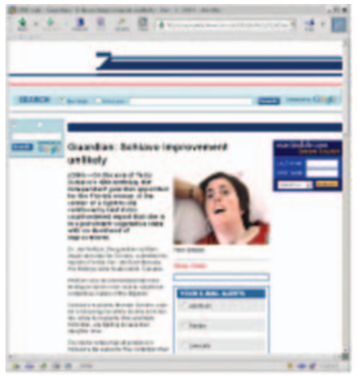
The main contribution of the thesis is to define empirical metrics over rulesets or schemata based on the training material.

Research Focus: The Programming Systems Lab



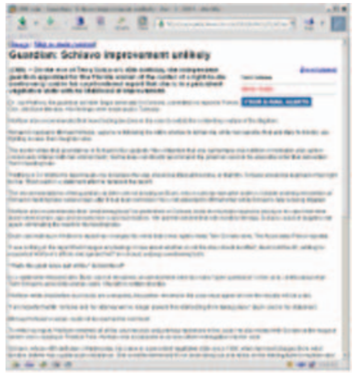
Guardian: Schiavo improvement unlikely

Original web page captured by Crunch web proxy



Guardian: Schiavo improvement unlikely

Crunch output for a PDA (whose browser will in turn reformat the elements, depending on screen orientation and scrolling mechanisms, and possibly remove the image, depending on screen resolution)



Guardian: Schiavo improvement unlikely

Crunch output for speech rendering

Professor Gail Kaiser's **Programming Systems Lab** takes a *software engineering* approach to programmed systems.

Rather than concentrate on any specific kind of system, such as operating systems or databases, PSL investigates and develops methodologies and technologies intended to apply to a wide range of software systems.

PSL's recent research emphasizes *self-managing systems*, addressing the problem that many software systems are too complex for human administrators to manage efficiently while maintaining availability, particularly in the face of increasing security threats. PSL's solution assumes models of a system's expected and required behaviors, enabling the system itself to monitor for divergences from those models, continuously analyze any evidence of deviations to dynamically construct reconfiguration plans intended to prevent, tolerate or repair the disruption, and then enact those plans—with no human involvement at run-time, but still aiming to make the process understandable and controllable by human users.

PSL has focused on externally injecting its **Kinesthetics eXtreme** (KX) self-management technology into legacy systems, and on experimental validation with real-world systems developed by third parties (e.g., ISI's GeoWorlds geographical intelligence analysis system, Telecom Italia's heterogeneous instant messaging service). KX components have also been used in collaborative intrusion detection systems across multiple institutions and for synchronizing streaming video across multiple distributed viewers.

Self-managing systems have been popularized as "autonomic computing," a term coined by IBM in 2001, in analogy to the autonomic nervous system—which does not include the higher functions of the brain. Since PSL's first results came earlier, in 1999, Prof. Kaiser has been an invited speaker at several autonomic computing forums, including the IBM Almaden Institute Symposium on Autonomic Computing in 2002, the 5th Annual International Active Middleware Workshop and the Autonomic Computing briefing of the Woodrow Wilson International Center for Scholars, both in 2003, and the Technology Transfer Institute Vanguard conference on The Challenge of Complexity in 2004.

PSL also continues research related to *software development environments*, its earliest focal area—sometimes leading to unanticipated applications, e.g., for collaborative work. Most recently, PSL's **Crunch** framework for "content extraction" from HTML web pages has proven very useful for preprocessing web pages for speech rendering and other devices for the blind and visually impaired. Crunch integrates a collection of heuristic plugins concerned with how to prune the DOM (document object model) tree representing the web page; these heuristics are often customizable, e.g., to specify the text-to-link ratio threshold for when to remove links, which tends to be different on news vs. shopping sites. Crunch incrementally clusters websites based on their search engine snippets, to determine the closest "genre" for a previously unknown website, to select these settings. The Crunch web proxy can be used with any browser and tuned to multiple applications, such as speech rendering or small PDA screens as well as the originally envisioned input to content summarization tools. The KX components and Crunch can be downloaded from *http://www.psl.cs.columbia.edu.*

Hassan Malik graduated with a Masters degree in Computer Science from North Carolina State University, Raleigh in 2003 and started as a doctoral student at Columbia in Fall 2004. He is originally from Karachi, Pakistan where he completed his undergraduate studies in Computer Science in 1999. During his high-school and undergraduate years, he founded a software company, Logicators and managed it for 7 years. Logicators developed computer game engines, multimedia applications and business applications for a variety of national and international clients and won several awards. He also worked as a Technical Leader at Extensibility Inc., a successful XML startup company and built core components of Tibco BusinessWorks, at Tibco Software Inc., in Palo Alto, CA. He is currently working as Manager of Applications Architecture at Liberty Travel in Northern NJ, where he is leading a large team to design the Next Generation Travel System. He is working with Professor John Kender and conducting research in the areas of Multimedia Information Retrieval and Web Mining.

Arezu Moghadam graduated from Amirkabir University of Technology with a BS in Electrical Engineering in February 2001. She did her MS in Electrical Engineering at University of Pennsylvania from 2002-2004, and started as a PhD student in Fall 2004. Arezu's main research interests are Network Architecture and Protocol Design for Multimedia Networking and Network Security. She is currently working with Professor Henning Schulzrinne's group on Wireless Multimedia Networking.

Remi Moss received her BA in history from UC Berkeley and her master's degree in education from UCLA. Prior to joining the Columbia community, she worked at UCLA as a study abroad advisor. In the CS department, she serves as a PhD Program Administrator, and she coordinates Black Friday and helps students resolve registration issues and the like. She enjoys reading novels and essays, and exploring New York City.

Raphael A. Pelossof graduated from Tel-Aviv Academic College with a BA in Computer Science and a BA in Economics and Management in June 2001. He worked in industry from 2000-2002 and did an MS in computer science at Columbia University from 2002-2003. He worked for a year in the Cognitive Psychology laboratory and in the Computational Learning Center at Columbia, and started as a PhD student in Spring 2004. Raphael's main research interests are Machine Learning, Vision and Robotics. He is currently working in the Learning Center with Dr. David Waltz as his advisor, studying Machine Learning and its application to Vision.

Lokesh S. Shrestha graduated from Princeton University with a BSE in Computer Science in June 2000. After graduation he joined Abridge Inc. as a Software Engineer where he was involved in the research and development of an email management software. Since June 2003, he has been with the Natural Language Processing Lab at Columbia University working with Professor McKeown and Dr. Rambow in investigating various approaches to the summarization of email threads. He completed his requirements for an MS in Computer Science in Fall 2004 and started as a PhD student in Spring 2005. His main research interests are statistical and machine learning approaches to natural language processing in general, and email thread summarization in particular.

Francois-Xavier Standaert received the Electrical Engineering degree and PhD degree from the Université catholique de Louvain, respectively in June 2001 and June 2004. He is currently a Fulbright Visiting Researcher at Columbia University (September 2004-January 2005) and MIT Medialab (February 2005-June 2005). His research interest includes digital design and FPGA's, cryptographic hardware, design of cryptographic primitives and side-channel analysis. In fall 2004, he worked with Professors Moti Yung and Tal Malkin (Department of Computer Science, Networks Security and Theory of Computation groups) about models for the physical security of cryptographic implementations.

Julia Stoyanovich graduated from UMass Amherst with a BS in Computer Science and a BS in Mathematics and Statistics in August 1998. She then moved to New York, where she worked at 2 startups and one large corporation. Julia did her MS in Computer Science at Columbia in 2003-2004, and started as a PhD student in the Fall of 2004. Julia's main research interests are in Database Systems, in particular Query Processing and Optimization. She is currently working with Professor Ross.

Olivier Tardieu got his PhD in September 2004 from École des Mines de Paris, France. He prepared his dissertation "Loops in Esterel: from operational semantics to formally specified compilers" under the supervision of Gérard Berry in the INRIA French public research institute. He started in Columbia both as a lecturer and postdoctoral research scientist in January, working with Professors Aho and Edwards. His current research interests include programming languages (design, semantics, code generation, static analysis) and digital circuits (again design, synthesis, formal verification).

Andrew Tzu-Kang Wan graduated from Columbia University with a BA in Philosophy-Economics in 2000. In the fall of 2004, Andrew started his PhD. He is studying theory of computation with Professors Tal Malkin and Rocco Servedio. His main research interests right now are computational complexity, computational learning theory, and cryptography.

Sean White graduated from Stanford University with a BS in Computer Science in 1992 and a MS in Computer Science in 1993. He worked at Interval Research, WhoWhere, Lycos, and NeoCarta Ventures until entering Columbia's ME graduate program where he earned an MS investigated carbon nanotubes for fuel cells and solar cells. He started the PhD program in CS at Columbia in Spring 2005 and is currently working for Professor Steve Feiner's group studying mobile augmented reality, visualization, and user interface.

New **Faces**

Melinda Y. Agyekum graduated from the Georgia Institute of Technology with a Bachelor of Science degree in computer engineering in December 2001. After graduating from Georgia Tech, she entered Columbia University in Fall 2002 and obtained a Master of Science degree in computer science in the spring of 2004. She is currently a first year PhD student with a research concentration in the area of computer systems and hardware design. She works under the supervision of Professor Steven Nowick in the Asynchronous Circuits and System Group studying state machine decomposition and their application to Burst Mode Controllers.

Spyridon Antonakopoulos graduated from the National Technical University of Athens in June 2004 with a Diploma in Electrical and Computer Engineering, and started as a PhD student in Computer Science at Columbia University in Fall 2004. Spyridon's research interests include approximation and online algorithms, computational complexity and computational learning theory. He is currently working with Professor Yannakakis on approximation algorithms for survivable network design problems.

Salman A. Baset graduated from Ghulam Ishaq Khan Institute of Engineering Sciences and Technology with a BS in Computer System Engineering in May 2001. He worked at Avaz Networks from 2001-2003, did an MS in computer science at Columbia from 2003-2004, and started as a PhD student in Fall 2004. Salman's main research interests are multimedia and peer-to-peer networks, ubiquitous computing and network security. He is currently working in the Internet Real-Time Lab supervised by Professor Henning Schulzrinne and is studying TCP dynamics for constant bit rate (CBR) traffic.

Stefan Benus joined the Spoken Language Processing Group on 2/1/05 as a post-doc research scientist working with Professor Julia Hirschberg. He holds a PhD in linguistics from New York University; his dissertation deals with articulatory phonology and dynamic modeling. He is primarily involved in investigating prosodic cues that correlate with deceptive speech and other pragmatic functions. He is also interested in dialectal and cross-linguistic intonational patterns, as well as the relationship between continuous and discrete aspects of prosody.

Simon Bird is the new Assistant Director of Academic Programs in the CS department advising students and helping administer the CS undergraduate and MS programs. His role includes advising students regarding program planning, course choices, and graduation clearance, administering the MS program admissions process, dealing with general student academic concerns, and acting as liaison between students and faculty. Hailing from the sunny islands of Britain, Simon has a bachelor degree in zoology, a PhD in insect ecology, and was a landscape ecologist in a previous life. Manhattan is the Bird's natural habitat, however. Outside of the CS department, Simon is a pseudo guitarist/song writer, photographer, and writer (i.e., a generalist).

Eli Brosh graduated from Tel-Aviv University, Israel, with a BS in Computer Science, Statistics, and Operations Research in July 1997. He worked in industry from 1997-2003, did an MS in Computer and Electrical engineering at Tel-Aviv University 1999-2003, and started as a PhD student at Columbia University in Fall 2004. Eli's main research interests are design and analysis of algorithms for communication networks, and performance evaluation of network systems. He is currently working with Professors Dan Rubenstein and Vishal Misra studying network security and resiliency in the context of overlay and peer-to-peer networks.

John M. Cieslewicz graduated from Stanford University with a BS in Computer Science in June 2004. A PhD student since Fall 2004, John's research interests include databases and data streams. John is currently working with his adviser, Professor Ken Ross, on architecture sensitive optimizations for in-memory database operations.

Franz Coriand is a master's student from the Augmented Reality Lab at Bauhaus University Weimar (Germany). He is doing a research internship supervised by Professor Steven Feiner (Computer Graphics & User Interfaces Lab). His main research interests include future display technologies (projected displays) and advanced human-computer interfaces.

Corey Goldfeder graduated from Yeshiva University with a BA in Computer Science and Mathematics in 2004. After receiving an NDSEG fellowship from the Department of Defense, he entered the PhD program in the fall of 2004. Corey is currently a member of Dr. Peter Allen's Robotics group, working on 3D shape matching algorithms, with applications to 3D search and robotic grasp prediction.

Maryam Kamvar graduated from Princeton University in June 2002 with a BA in Computer Science and a Certificate in French Language and Literature. She received an MS in Computer Science from Columbia University in June 2004 and is currently enrolled as a PhD student at Columbia with Betsy Sklar's lab. For the past year she has been working at Google in Mountain View, CA on the wireless team. She plans to return to Columbia in the fall and resume her research in Human Computer Interaction.

Homin K. Lee graduated from Columbia University with a BA in Music and Mathematics-Computer Science in 2000. After enjoying a delirious year riding the technology boom and a manic year of unemployment, Homin obtained his MS from Columbia in Spring 2004, and started in the PhD program in Fall 2004. Homin's main research interests are in Computational Complexity, Cryptography, and Computational Learning Theory. Homin is a proud member of Dixon's Theory Lab, and is advised by Rocco Servedio and Tal Malkin.

Featured **Group**

The **Database** Group

Professor
Kenneth Ross

Professor
Luis Gravano

Professor **Mihalis Yannakakis**

Database Systems is the branch of Computer Science that deals with the storage and management of large volumes of data.

Many kinds of applications need to deal with large databases, and the world-wide-web creates numerous interesting problems related to the storage and access of large volumes of data. Database Systems has a tradition of innovation in research, both at industrial research laboratories and at universities. Among the most notable advances is the **relational data model** that uses tables to represent information in a database. Today, commercial relational database systems constitute a multi-billion dollar market.

The database group at Columbia includes Professors Gravano, Ross and Yannakakis, and their students. Recent PhD graduates of the database group (since 2000) include Eugene Agichtein (now at Microsoft Research), Nicolas Bruno (Microsoft Research), Panos Ipeirotis (NYU), Kazi Zaman (Siebel), Jun Rao (IBM Almaden Research Center), and Jingren Zhou (Microsoft Research). Current and recent database group projects are described below. For more information about the database group, go to *http://www.cs.columbia.edu/database/*

Top-k Query Processing: 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 objects. Professor Gravano, his students, and their external collaborators have developed top-k query processing algorithms for a variety of important database applications and scenarios. 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.

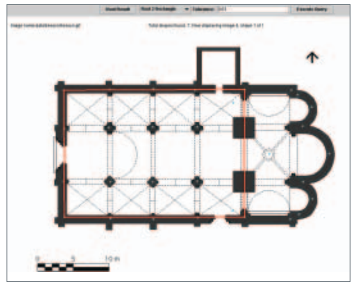
Structured Querying of Unstructured Text Documents: Text documents often contain valuable structured data that is "buried" in natural-language sentences. For example, news articles and web pages hide a wealth of information that is invaluable for business intelligence. Companies and organizations generally rely on structured, relational database management systems for data processing and analysis, which creates a gap between relatively unstructured text documents and highly structured databases. To bridge this gap, information extraction algorithms are needed for text. Professor Gravano and his students have designed critical building blocks for efficient information extraction, which help extract—with minimal to no human supervision—inherently structured information that is buried in web pages and text databases. A number of long-term challenges still need to be addressed, including adapting to the vast heterogeneity—in quality, format, scope, and intended audience—and the unprecedented volume of the information available over the web, which together yield existing information extraction techniques inadequate.

Architecture Sensitive Databases: Today's commercial database systems were designed at a time when I/O was the performance bottleneck. Many technological changes have happened since then, including much faster processors, much larger (but not much faster) memories, and much faster sequential data transfer from disk. Today, for many database workloads, cache misses have become the performance bottleneck. Professor Ross and his students are looking at ways to redesign data structures and algorithms for database operations to avoid or reduce cache misses. Other CPU-level delays, such as branch misprediction penalties, are also relevant.

A Database for Archaeology: Several Columbia University scholars from computer science, earth and environmental sciences, anthropology, historic preservation, classics, and art history and archaeology are creating new computational tools for modeling, visualizing, and analyzing historic structures and archaeological sites. Professor Ross and his group are developing new database technology to catalogue and access a site's structures, artifacts, objects, and their context.

A Database for Finding Shapes in Architectural Plans: Professor Ross is collaborating with Columbia Art History and Archeology Professor Stephen Murray on the study of a group of French churches. The historic region of the Bourbonnais flourished in the eleventh and twelfth centuries, and there are numerous churches in the region. No written records of how these churches were built remain; the only way to understand how they were built is to study the churches themselves. The geometric proportions of churches and their substructures can give strong clues about the technology used to construct them. Professor Ross and his group are building a system to query a set of points marked on church cross-sections for geometric shapes.

Multiobjective Optimization of Queries: In optimizing the evaluation of queries, a number of different criteria may come into play, e.g., minimization of resources such as time, memory, communication, maximizing quality and accuracy of the results, etc. The criteria may depend also on the underlying query processing context (e.g., centralized, distributed, streaming). It is generally impossible to optimize simultaneously all the objectives, and thus selecting an evaluation method involves making a trade-off between the objectives. Related issues arise when it is not the evaluation methods but the results of the query that are compared along different criteria, and a limited number of results has to be produced that strike the optimal balance between the criteria. Professor Yannakakis and his collaborators investigate approximation algorithms for optimizing the trade-offs between different objectives in query evaluation, and for computing a small number of solutions that represent the whole range of possible trade-offs.



A screenshot of our shape query engine searching for rectangles with a 1 by √2 side ratio. These rectangles are potentially interesting to art historians, because builders in the 13th century were known to construct √2 lengths by holding a rope along the diagonal of a unit square. In this screenshot, the section of the church at Besson highlighted in red forms an answer to the query. The blue points (which form the corners of shapes) are points that have been marked as reference points by art historians.

New Computer Science Courses **in 2004-2005**

Professor Servedio offered COMS 6998-2, **Advanced Topics in Computational Learning Theory,** in the Spring semester. The course focuses on state-of-the-art, provably correct and efficient algorithms to learn rich classes of Boolean functions such as decision trees, DNF formulas, and geometrically defined functions such as intersections of halfspaces and polynomial surfaces. A recurrent theme in the course is the close interplay between structural results in complexity theory and efficient algorithms in learning theory. A major component of the course is a personalized research project on an active research topic (of the student's choice) in learning theory. Each student presents the results of her or his research in an in-class presentation.

Jon Feldman offered COMS 4995-2, **Introduction to Coding Theory: a Computer Science Perspective** in the Fall semester. Coding theory is the study of error-correcting codes, which are used to transmit digital information in the presence of noise. Their direct applications range from deep-space communication to recovering from packet-loss on the Internet. As combinatorial objects, they also have many important applications to problems in complexity theory and cryptography. This course gives the basics of coding theory using the language of theoretical computer science. Both classical and modern topics in coding theory are covered, with an emphasis on problems concerning algorithm design and asymptotic analysis. This course also surveys some of the applications of coding theory to problems in digital communication and theoretical computer science.

A student-generated image from "Computer Animation".

Professor Grinspun introduced COMS 4167, **Computer Animation**, covering fundamental techniques of computer animation and simulation. Students focused on both theory and implementation through a combination of lectures, theory and programming assignments. In the second assignment, students designed and programmed their own "hopping robots" and then in class competed in a virtual Robot Olympics. In the second half of the semester students formed small teams, and then carried out a significant animation project of their choice. This challenging course spans all stages of animation from design and scripting through production and post-production, including traditional animation techniques, keyframing, kinematic rigging, simulation and dynamics, free-form animation, behavioral and procedural animation, production scheduling and post-production.

Professor Feiner offered COMS W4172, **3D User Interfaces,** in the Spring semester. 3D user interfaces are already essential to fields as diverse as scientific visualization and video games, and will become even more important as the major consumer "desktop" user interfaces begin to incorporate 3D technology in upcoming software releases. COMS W4172 provides an introduction to this exciting way of interacting with computers, emphasizing methods for designing and evaluating effective 3D user interfaces. Topics include interaction metaphors; selection and manipulation; traveling and wayfinding; perception, displays, and interaction devices; tangible user interfaces; and virtual and augmented reality. During the second half of the semester, students carry out team projects in conjunction with students in the Visual Studies Workshop "Architecture and Memory," taught by Prof. Rory O'Neill in the Graduate School of Architecture, Planning and Preservation.

Tracked user of a multimodal augmented reality system, shown with live overlaid graphics depicting 3D selection geometry.

Professor Stolfo offered COMS 6998-1, **Intrusion and Anomaly Detection Systems,** in the Spring semester. The course focuses on new methods and algorithms for detecting attacks against networked computer systems. The topics include an assessment of different kinds of attacks against systems, network- and host-based intrusion detection and prevention techniques and systems, and new state-of-the-art algorithms for behavior-based anomaly detection to cover the zero-day attack problem, and the insider attack problem. Methods are described to defend against worm/virus attack, as well as masquerade (or impersonation) detection. An invited speaker from industry presented a lecture on the true state of security operations in large corporations, and a view on future large-scale attacks that critical infrastructure industries should plan for. In the first project, students pair with each other to alternate roles as attacker and defender. Their task is to stealthily probe their partner's machine, while their partner's task is to identify when they are under surveillance and possibly imminent attack. A second project provides familiarity with common open source intrusion detection systems, while the final project tests the student's creativity in defending their host computer from malicious misuse.

Students in the course "Computational Aspects of Geometric Design" learn the algorithms behind such varied modeling tasks as character development (in computer animation) and massing studies (in architecture).

The course **Computational Aspects of Geometric Design** was offered by **Michael Reed** beginning in the Spring, 2004 semester. This course is an introduction to representation and manipulation of 3-dimensional shape as used in computer-aided design and manufacture, animation and special effects, and related fields. A variety of representations of 3D geometry are studied, with additional topics including curve and surface acquisition, rapid prototyping, and CAD system architecture. The material is illustrated using a combination of applets, physical models, short films, and guest lecturers from the animation and design fields. *(class website: www.cs.columbia.edu/cagd)*

Professor Carloni offered COMS E6988-03, **Distributed Concurrent Systems** in the Spring semester. The course, which was also offered on CVN, is an an inter-disciplinary graduate-level seminar on the design of distributed embedded systems. Emphasis is put on system robustness in the presence of highly variable communication delays and heterogeneous component behaviors. The course has a two-fold structure: the study of the enabling technologies (VLSI circuits, communication protocols, embedded processors, RTOSs), models of computation, and design methods is coupled with the analysis of modern domain-specific applications including on-chip micro-networks, multiprocessor systems, fault-tolerant architectures, and robust deployment of embedded software. Common research challenges include design complexity, reliability, scalability, safety, and security. The course requires substantial reading, class participation and a research project.

Professor Hirschberg offered COMS W4706, **Spoken Language Processing,** in the Spring semester. This course introduces students to computational approaches to speech generation and understanding. It focuses on speech recognition and understanding, speech analysis for computational linguistics research, and speech synthesis, with applications to spoken dialogue systems, data mining, summarization, and speech-to-speech translation. Students get hands-on experience performing speech analysis with software tools and building their own text-to-speech system, all using freeware components.

Markus Hofmann offered E6998-9, **Content Networking,** in the Spring semester. The course focuses on the challenges in distributing content on the World Wide Web (WWW), describes basic concepts and principles for improving content delivery over the Internet and outlines possibilities for tapping into the huge potential of custom-tailored content services. The lectures concentrate on explaining and evaluating underlying principles, concepts and mechanisms, and use many examples and case studies for illustration. Specific protocols are selected as examples of how concepts and mechanisms can be incorporated in real-life networks, but the main aim of lectures are to provide a systematic and architectural view of the content networking and content services field. All parts of the lectures will have a mix of research and industry flavor, addressing seminal research concepts and looking at the technology from an industry angle. More information is available at *http://lecture.mhof.com/*

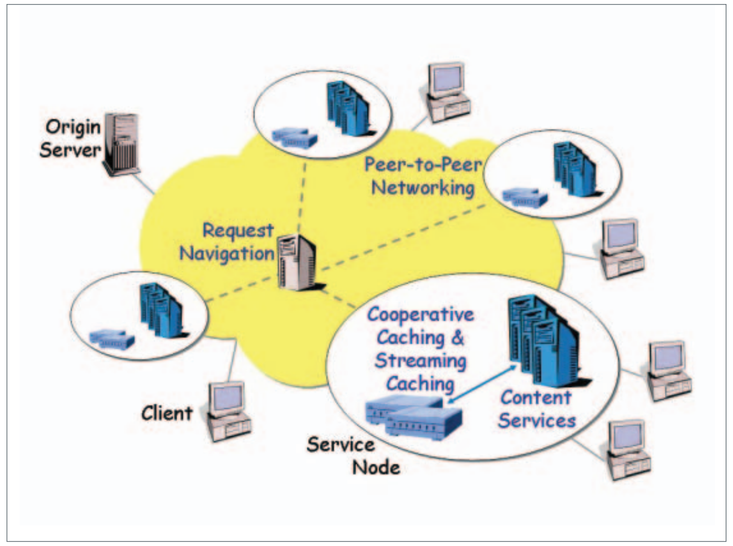


Illustration from "Content Networking".