CUCS Welcomes Three New Faculty

Michael Collins named Vikram S. Pandit Professor of Computer Science

“Michael Collins revolutionized the natural language processing (NLP) field with his work on parsing. He was the first to build a robust, statistical parser, which is now the most downloaded software in NLP and is widely used in applications around the world.”

– KATHLEEN MCKEOWN

MICHAEL COLLINS received a Ph.D. in computer science from the University of Pennsylvania in 1998. From 2003 to 2010 he was a professor in computer science at MIT; prior to that he was a researcher at AT&T Labs-Research. Professor Collins’s research has focused on topics including statistical parsing, structured prediction problems in machine learning, and NLP applications including machine translation, dialog systems, and speech recognition. His awards include a Sloan fellowship, an NSF career award, and best paper awards at several conferences: EMNLP (2002, 2004, 2010), UAI (2004, 2005), and CoNLL 2008.

CUCS: Tell us about your field of research.

My research is in the area of machine learning for natural language processing (NLP). I work on applications such as machine translation, speech recognition, dialog systems and information extraction. I got into this research area for a few reasons. I first learned about natural language processing while doing my masters, and quickly became fascinated by the idea of formal or computational models of human languages. These models go back to Chomsky's work in linguistics, which were later developed further by computer scientists. In recent years there has been a great deal of interest in statistical models for NLP, the idea being to learn models from large amounts of linguistics data. Statistical NLP poses many interesting challenges, both in terms of the development of (continued on next page)

From left to right: Augustin Chaintreau, Michael Collins and Xi Chen
What are your ongoing research projects and what motivated you to focus on them? One research project I've been working on is models for statistical machine translation that make greater use of linguistic structures. In recent years, great progress has been made in machine translation through statistical models—Google translate being the obvious example of such a system. However these systems make relatively little use of linguistic structure, and because of this sometimes have trouble producing coherent output, or in modeling systematic differences in word order between languages. We've been developing statistical models that make explicit use of linguistic structure, in an effort to address these problems. A second recent research project is more on the algorithmic side, and concerns efficient inference algorithms for several statistical models found in NLP. In many applications, inference (e.g., finding the most likely translation for a sentence) is very costly, because of this, researchers have generally focused on approximate search methods, which have quite weak guarantees. We've been developing exact inference methods, based on ideas from linear programming relaxations, and combinatorial optimization. I'm excited about this work because we've been able to exactly solve several inference problems that people previously thought were intractable, and this has led to improvements on several problems.

What is your approach to research? I like working on real-world, practical problems, but I also enjoy developing formal models and algorithms. Ideally I try to develop approaches that give real improvements on problems, but that have a clear theoretical underpinning.

What are some of your ongoing research projects? Recently I have been working on the computation of market equilibria. In a classical market, there are buyers and sellers exchanging money and goods. An equilibrium is a state in which everybody feels satisfied about the outcome of the exchange. The model has been studied in economics for a long time, and has been recognized by many as the crown jewel of mathematical economics. While an equilibrium always exists under mild conditions, in general we do not know how to find one efficiently. Computational problems related to this model have been studied intensively in the computer science community in the past decade, from either the algorithmic or the complexity-theoretic perspective.

What do you consider the most challenging aspects of your current research? One of the greatest challenges I see is to bridge the gap between theory and practice. This is by no means easy and coming up with a practical research problem that is interesting in both theory and practice is probably the hardest part in many situations. I believe that working in AGTE would offer me a lot of opportunities to collaborate with researchers in other areas of CS. Therefore, the diversity and excellence of the Columbia CS faculty is definitely the factor that attracted me the most for coming here.

What attracted you most to a faculty position at Columbia? Where you work is incredibly important! I have always looked for academic rigor and energetic support from my colleagues. Visiting Columbia last year and seeing how people work here just blew me away. Another important fact is that Columbia’s “footprint” spans the social sciences, economics, business and journalism. One reason I’m moving to academia after five years in industry is that research in mobile and social networks today impacts all these fields.

What are some of your ongoing research directions? As mobile networks are now dealing with an abundance of local information, we are working on a distributed recommendation engine. We try to predict what information would be relevant to you the same way Netflix recommends movies.

Do you have a method or approach that you like to use in your research? I learned at least one important thing from my mentors: theory and system building go hand in hand. This is interesting because sometimes they might seem to compete with each other. I believe that working in AGTE would offer me a lot of opportunities to collaborate with researchers in other areas of CS. Therefore, the diversity and excellence of the Columbia CS faculty is definitely the factor that attracted me the most for coming here.

What are the most interesting aspects of your research? Certainly one thing that keeps me awake is how we will exploit the vast wealth of social and mobile information that we all produce everyday. We are starting to understand which tasks could be solved more efficiently using social networking tools, as well as the limits of these tools, and how they can be misused. This in turn leads to a series of computational challenges, and I think solving any of them could have a large impact.
The new program is the first of its kind, as it aims at bridging the gap between the evolving discipline of journalism and new information technologies, while leveraging the unique environment of New York City. Home to the three traditional major American television networks, four of the ten largest newspapers, several large cable television channels and radio stations, and most of the book and music publishing industry, New York has been called the “Media Capital of the World.” And along with traditional media, the city has seen a continuous growth of its media-oriented information-technology industry, which includes software and systems development for interactive media, networking, and Internet services. In the fall of 2011, the first class of students will start the five-semester graduate program. By taking courses in both the School of Journalism and the School of Engineering and Applied Science, as well as working on interdisciplinary projects under the joint supervision of professors from both schools, the students will simultaneously earn M.S. degrees in Computer Science and Journalism.

To better prepare students for careers in the digital-media world, the new program will teach them about the impact of digital techniques on journalism, the emerging role of citizens in the news process, the influence of social media, and the changing business models that will support news gathering, reporting, and delivery. The following is a sample of the program course offering from the two disciplines:

**Journalism**
- Reporting and Writing
- Journalism essentials: law, business, ethics, and history
- Skills classes in photo, audio, video, and other digitally based media
- Digital Media Workshop
- Business and Economic Reporting
- Science Reporting
- Interactive Media Storytelling

**Computer Science**
- Introduction to Databases
- Advanced Software Engineering
- User Interface Design
- Analysis of Algorithms
- Network Security
- Natural Language Processing
- Web-Enhanced Information Management
- Intrusion Detection
- Seminar on anonymity and privacy
- Machine Learning
- Search Engine Technology

At the same time, students will have the opportunity to apply some of the new media-related technologies that have been developed by professors and researchers in the Department of Computer Science, including: speech understanding and speech synthesis, automatic text summarization, 3D user interface design and augmented reality, computer vision and digital imaging, statistical machine learning for the analysis of natural languages, and machine learning for the analysis of massive data sets generated by social networks.

**Possible Career Paths**

**Online Editor/Manager of Information Technology at a large news organization.**
SKILL SET: knowledge of software creation, databases, user interfaces, security, large-scale text processing to engage readers on the web and mobile devices while ensuring content and design meet the standards for quality journalism.

**Data-mining expert for journalistic applications and investigative journalism.**
SKILL SET: knowledge of statistics, machine learning, and advanced language processing techniques, Web search for the purpose of gathering and synthesizing data.

**Entrepreneur/founder of media startup.**
SKILL SET: strong editorial background and knowledge about the business of journalism; basic knowledge of computing with specialization in social media, data mining, and other application areas.

For more information: [http://www.cs.columbia.edu/education/ms/journalism](http://www.cs.columbia.edu/education/ms/journalism)

**WHAT PEOPLE ARE SAYING**

The concept makes sense, the problem it addresses is real, and Columbia is capable of taking on the challenge. But we were most fascinated by the technologies these professors hope their graduates will contribute. – Wired Magazine

One goal of the Columbia program, according to Bill Grueskin, the dean of academic affairs, is to produce journalists who will “take it several steps beyond—to where they’re creating a lot of their own new tools.” That means learning enough computer science and software engineering to be able to design tools for information gathering, synthesis, analysis and circulation—or enough, at least, to see what technology can do for journalism. – The New York Times

Many of today’s programmer-journalists got to where they are thanks to self-directed education and hacking together courses and other educational opportunities to build their skills. But the new Columbia program, along with other initiatives, suggests that the next wave of programmer-journalists could be trained in specialized education programs that combine a traditional engineering/computer science degree with a traditional journalism education. Universities are working to either alter existing journalism programs or create new joint degrees to formalize the training of these workers. – PBS Mediashift
Yannakakis, the Percy K. and Vida L. W. Hudson Professor of Computer Science, was honored for his contributions to algorithms and computational complexity. He is one of 77 new members elected this year and joins a total membership of nearly 2,500 worldwide.

NAE membership honors those who have made outstanding contributions to engineering research, practice, or education, including, where appropriate, significant contributions to the engineering literature, and to the pioneering of new and developing fields of technology, making major advancements in traditional fields of engineering, or developing/implementing innovative approaches to engineering education.

“It is a great honor for me to join such a distinguished body,” Yannakakis says. “I am very happy to be elected.”

The honor has garnered him praise from Columbia Engineering leaders.

“As one of the leading academics working to define the laws that govern the computational world, Mihalis is both an exceptional educator and a ground-breaking scholar,” - Feniosky Peña-Mora

**Featured Lectures**

**2010-11 Distinguished Lecture Series**

The Computer Science had an exciting lineup of speakers during the 2010-11 academic year. The speakers include leading figures in science in both academia and industry with a particular emphasis on the internet’s origins, internet advertising, legal aspects of the digital world and user interaction.

The lectures were open to all members of the Columbia University campus as well as the general public. Details on the talks are available via the department’s website (http://www.cs.columbia.edu/lectures). Visitors also had opportunities to meet with students and faculty in one-on-one settings. Such visits frequently inspire new research ideas at the department and help forge collaborations between our faculty and leading colleagues. We hope you will attend future lectures in our subsequent series.
ABSTRACT: In this presentation we discuss the birth and future of the Internet. The early work on packet switching is traced and then a brief description of the critical events in the growth of the Internet is given. We then present a vision of where the Internet is heading with a focus on the edge where user participation, flexible applications and services, and innovation are appearing. We foresee a network with extreme mobility, ubiquity, personalization, adaptivity, and services, and innovation are appearing. The Internet is heading with a focus on the edge where user participation, flexible applications and services, and innovation are appearing. We foresee a network with extreme mobility, ubiquity, personalization, adaptivity, with extreme mobility, ubiquity, personalization, adaptivity,
Salman Abdul Basset and Professor Henning Schulzrinne won best paper award at IFIP COMM 2010 for their paper titled “Frequency and Relay Selection in Peer-to-Peer Communication Systems.”

About Face: NPR interviews Professor Peter Bellmeur on Face Recognition. He discusses his work on face recognition on National Public Radio. The full interview is available at http://www.npr.org/2010/12/17

Mattei Gio caricato has been named the winner of the 2010 Robot-Daten Scientific Award for his Ph.D. Thesis work on “Low-dimensional Robotic Grasping: Eigensnap Subspaces and Optimized (Un)contactarction.”

Scientific American features research of Ph.D. student Oliver Cossairt and Professor Shree Nayar. “Draw the Curtailed: Gigapixel Pixel Cameras Create High-bandwidth Gigapixel Snapshots.” Today’s high-resolution camera capture images with pixel counts in the tens of billions. In the future, digital cameras may be able to capture images with billions of pixels. Columbia’s CAVE lab has shown that by using a large ball lens, an array of planar sensors, and deconvolution as a post-processing step, one can capture gigapixel images with a very compact camera.

Professor Dr. Shree Nayar received an IBM Outstanding Innovation Fellowship, which will fund Nayar for up to two years at a postdoctoral position “to advance the field of computing and its positive impact on society.”

Professor David Harmon received a CAREER Award from the National Science Foundation, an award for “CSC@CU: CCC Computing Innovation Fellowship, which will fund Allan for up to two years at a postdoctoral position “to advance the field of computing and its positive impact on society.”

Professor David Harmon received an NSF Grant for “Detection and Performance of Event Content in Social Media.”

More information can be found at http://www.nsf.gov/awardsearch/showAward.do;sfAwardNum=1073288

Professors B. MacIntyre, M. Haupt, and W. Dally were a part of the winning team for the 2010 ACM SIGMOD Conference for “Making Multithreaded Programs Deterministic.”

Professor Julia Hirschberg was interviewed on The Leonard Lopate Show on WNYC discussing speech technology and emotional speech in particular. The interview is available at http://www.studio10 Today’s high-definition computing and its positive impact on society.”

Peter Johnstone, an IBM Ph.D. Student with co-author David Elson “is the eighth year of Lasting Impact Awards.”

Applications such as traffic monitoring, mobile user management, and sensor networks need to process large volumes of updates while supporting on-line analytic queries. With large amounts of RAM, single machines are usually modeled as managing hundreds of millions of items. With multiple hardware threads, as many as 64 on modern commodity multicore chips, many operations can be processed concurrently. Processing queries and updates concurrently can cause interference. Queries need to see a consistent database state, meaning that at least some of the time, updates will need to wait for queries to complete. To address this problem, the IBM team developed a RAM-based snapshot of the database to track various points in time. Analytic queries operate over the snapshot, eliminating interference, but allowing answers to be slightly out of date. Different scheduling methods are being developed and studied, with the goal of being able to create snapshot rays quickly (e.g., in a fraction of a second) while minimizing the overhead on update processing. These problems are studied both for traditional server machines, as well as for multicore mobile devices. By keeping personalized, up-to-date data on a user’s mobile device, a wide range of potential new applications could be supported while avoiding the privacy concerns of widely distributing one’s location. The research focus is on how to efficiently utilize the many processing cores available on modern machines, both traditional and mobile devices. A primary goal is to allow performance to scale as additional cores become available in newer generations of hardware.

Malek Ben Salem and Professor Salvatore J. Stolfo won best paper award at The 2nd International Workshop on Managing the STart of Cybersecurity (MIST 2010). The paper was entitled “Detecting Masqueraders: A Comparison of One-Class Bag of Words User Behavior Modeling Techniques.”

The workshop was conducted in conjunction with the 4th IFIP International Conference on Management of Intrusion, Malware, Japan, June 2010.

Professor Henning Schulzrinne is the recipient of the 2010 William Terry Lifetime Distinguished Service Award for his contribution and service to the IEEE and Region 1. This award is intended to recognize those whose personal efforts have provided leadership, creativity, guidance, and inspiration in a wide range of IEEE activities over a long period of time.

Professor Rocco Servedio is the 2010 recipient of the Distinguished Teaching Award of the Department of Computer Science. This award is in recognition of his dedication to teaching and his efforts to make Computer Science accessible to all students.

Professor Junfeng Yang was chosen for a NSF Career Award for “Making Multithreaded Programs Deterministic.” Multithreaded programs are becoming increasingly critical driven by the rise of multicore hardware and the coming storm of cloud computing. Unfortunately, these programs remain difficult to write, test, debug. A key reason for this difficulty is nondeterminism: different runs of a multithreaded program may show different behaviors depending on how the thread interface. Nondeterminism complicates almost every development step of multithreaded programs. For instance, it weakens testing because the schedule may not be the same run of the field; it complicates debugging because representing a buggy schedule is hard.

In the past three decades, researchers have developed many techniques to address nondeterminism. Despite these efforts, the impact of nondeterminism continues to grow, especially as the programming models being implemented in parallel systems, and makes a program repeat inputs. Based on this insight, it takes an approach called “make Computer Science accessible to all students.

Professor Junfeng Yang and his students got two papers accepted to the 2010 Symposium on Operating Systems Design and Implementation (OSDI’10), which is one of the top two operating systems conferences. Out of a few hundred submissions, only 32 papers were accepted! The two Columbia papers will be presented next month in Vancouver.

More information can be found at http://www.nsf.gov/awardsearch/showAward.do;sfAwardNum=1049698
Kasparov and Thiel Visit CU CS

Chess grandmaster Garry Kasparov and PayPal co-founder Peter Thiel visited the Columbia Robotics lab on November 4. Kasparov and Thiel were filming an episode of the award winning German-French television series “Into the Night with...” In each episode, two artists meet and spend one night exploring a city. Kasparov and Thiel are both very interested in technology, and began their exploration of New York City at night with a tour of the Robotics Lab where they discussed man vs. machines and Artificial Intelligence.