PILOT SEMINAR

Organized by Madhusudan Parthasarathy and occasionally Darko Marinov

This is a seminar series for practice of academic job talks by our students and postdocs. We have several students and postdocs going on the job market each year, and this seminar aims to give them feedback from faculty who are outside their area.

What do you need to do to give a talk in this seminar?

Any graduating student or postdoc who is going to be on the academic job market can give a PILOT seminar. All faculty, postdocs, and students will be invited to attend and give comments on the talk.

These talks are meant to be the penultimate talk before the interview, where the presenter seeks feedback from faculty outside his/her area.

So we expect the presenter to have already given some practice talks earlier (we encourage at least two earlier talks, one to the advisor's group, and another to the relevant area).

To ensure there will be some faculty who attend the talks, we expect the following:

- The presenter, with the help of the advisor, must invite at least 5 faculty personally to attend their talk, and ensure that at least 3 faculty can come to their talk. These faculty should be outside the primary area of the presenter.
- The time for the seminar should be fixed based on the availability of these faculty. To find a time that doesn’t overlap with the job talks, please ask your advisor to check the available dates on the faculty Wiki that lists all upcoming job talks.

More faculty may attend the talk as it will be publicly announced, but we would like to see some effort by the presenter and advisor in ensuring at least some people come to the talk.

Once you have a list of three faculty outside your area who have promised to attend your talk at some agreed time, please (1) edit this page to enter the date/time of your talk, your name, and the talk title and abstract, and (2) email Darko (marinov@illinois.edu), Madhu (madhu@illinois.edu), Viveka (kudaliga@illinois.edu), and Samantha Smith (sdsmith3@illinois.edu) so that we can arrange a room and announce your talk (we will fill in the room number). You should ask a room to be reserved for at least 90 minutes (60 minutes to present and at least 30 minutes to get feedback).

Optional: Speakers have the option of having their seminar presentations video-recorded (if scheduled in a tech-enabled space). If you wish to have your seminar recorded, please consult with your advisor and then email your request to engrit-help@illinois.edu. Please include the time and location of your seminar and add a note that this is a PILOT seminar.

You may find it useful to read these guidelines about academic job interviews:


2018-2019 Schedule

<table>
<thead>
<tr>
<th>Date and Time</th>
<th>Room</th>
<th>Speaker</th>
<th>Title and Abstract</th>
</tr>
</thead>
</table>


Title: Pattern-Based Mining of Entity/Relation Structures from Massive Text

Abstract: Majority of information nowadays is carried by massive and unstructured text, in the form of news, articles, reports, or social media messages. This poses a major research challenge on mining entity/relation structures from unstructured text. Manual curation or labeling cannot be scalable to match the rapid growth of text. Most existing information extraction approaches rely on heavy human annotations, which can be too expensive to tune and not adaptable to new domains.

In this talk, I will present a pattern-based methodology that conducts information extraction from the massive corpora using existing resources with little human effort. The first component, WW-PIE, discovers meaningful textual patterns that contain the entities of interest. The second component, TruePIE, discovers high quality textual patterns for target relation types. I will demonstrate how semi-supervised methods can empower information extraction for broad applications and provide explainable results.

Bio: Qi Li is currently a postdoc researcher and adjunct professor at Department of Computer Science, University of Illinois at Urbana-Champaign, working with Prof. Jiawei Han. Her research interests lie in the area of data mining with a focus on the extraction and aggregation of information from multiple data sources. Qi obtained her PhD in Computer Science and Engineering from the State University of New York at Buffalo in 2017 advised by Prof. Jing Gao, and MS in Statistics from University of Illinois at Urbana-Champaign in 2012. She has received several awards including the Presidential Fellowship of University at Buffalo, the Best CSE Graduate Research Award and the CSE Best Dissertation Award at Department of Computer Science and Engineering, University at Buffalo. More information can be found at https://publish.illinois.edu/qili5/.
Abstract: The risk posed by software bugs has increased significantly as software is now essential to many areas of our daily lives. Runtime verification can help find bugs by monitoring program executions against formally specified properties. Over the last two decades, tremendous research progress has improved the performance of runtime verification. However, there has been very little focus on the benefits and challenges of using runtime verification during software testing. Yet, testing generates many executions on which properties can be monitored. In this talk, I will describe my work on studying and improving runtime verification during testing. My large-scale study was the first to show that runtime verification during testing is beneficial for finding many important bugs from tests that developers already have. However, my study also showed that runtime verification still incurs high overhead, both in machine time to monitor properties and in developer time to inspect violations of the properties. Moreover, all prior runtime verification techniques consider only one program version and would wastefully re-monitor unaffected properties and code as software evolves. To reduce the overhead across multiple program versions, I proposed the first evolution-aware runtime verification techniques. My techniques exploit the key insight that software evolves in small increments and reduce the accumulated runtime verification overhead by up to 10x, without missing new violations.

Bio: Owolabi Legunsen is a PhD candidate in Computer Science at the University of Illinois at Urbana-Champaign, where he works with Darko Marinov and Grigore Rosu. Owolabi's interests are in Software Engineering and Applied Formal Methods, with a focus on Software Testing and Runtime Verification. His research on runtime verification during software testing received an ACM SIGSOFT Distinguished Paper Award at ASE 2016. More information is available on his web page: http://mir.cs.illinois.edu/legunsen
Title: Secure Computer Hardware in the Age of Pervasive Security Attacks

Abstract:
Recent attacks such as Spectre and Meltdown have shown how vulnerable modern computer hardware is. The root cause of the problem is that computer architects have traditionally focused on performance and energy efficiency. Security has never been a first-class requirement. Moving forward, however, this has to radically change: we need to rethink computer architecture from the ground-up for security.

As an example of this vision, in this talk, I will focus on speculative execution in out-of-order processors --- a core computer architecture technology that is the target of the recent attacks. I will describe InvisiSpec, the first robust hardware defense mechanism against speculative (a.k.a transient) execution attacks. The idea is to make loads invisible in the cache hierarchy, and only reveal their presence at the point when they are safe. Once an instruction is deemed safe, our hardware is able to cheaply modify the cache coherence state in a consistent manner.

Further, to reduce the cost of InvisiSpec and increase its protection coverage, I propose Speculative Taint Tracking (STT). This is a novel form of information flow tracking that is specifically designed for speculative execution. It reduces cost by allowing tainted instructions to become safe early, and by effectively leveraging the predictor hardware that is ubiquitous in modern processors. Further improvements of InvisiSpec-STT can be attained with new compiler techniques. Finally, I will conclude my talk by describing ongoing and future directions towards designing secure processors.

BIO:
Mengjia Yan is a Ph.D. student at the University of Illinois at Urbana-Champaign (UIUC), working with Professor Josep Torrellas. Her research interest lies in the areas of computer architecture and hardware security, with a focus on defenses against transient execution attacks and cache-based side channel attacks. Her work has appeared in some of the top venues in computer architecture and security, and has sparked a large research collaboration initiative between UIUC and Intel. Mengjia received the UIUC College of Engineering Mavis Future Faculty Fellow, the Computer Science W.J. Poppelbaum Memorial Award, a MICRO TopPicks in Computer Architecture Honorable Mention, and was invited to participate in two Rising Stars workshops.
Automated Resource Management in Large-Scale Networked Systems

Abstract:
The multitude of Internet applications relies on large-scale networked environments such as the cloud for their backend support. In these multi-tenanted environments, various stakeholders have diverse goals. The objective of the infrastructure provider is to increase revenue by utilizing the resources efficiently. Applications, on the other hand, want to meet their performance requirements at minimal cost. However, estimating the exact amount of resources required to meet the application needs is a difficult task, even for expert users. Easy workarounds employed for tackling this problem, such as resource over-provisioning, negatively impact the goals of the provider, applications, or both.

In this talk, I will discuss the design of application-aware self-optimizing systems through automated resource management that helps meet the varied goals of the provider and applications in large-scale networked environments. The key steps in closed-loop resource management include learning of application resource needs, efficient scheduling of resources, and adaptation to variations in real-time. I will describe how I apply this high-level approach in two distinct environments using (a) Morpheus in enterprise clusters, and (b) Patronus in cellular provider networks with geo-distributed micro data centers. I will also touch upon my related work in application-specific context at the intersection of network scheduling and deep learning. I will conclude with my vision for self-optimizing systems including fully automated clouds and an elastic geo-distributed platform for thousands of micro data centers.

Bio:
Sangeetha Abdu Jyothi is a Ph.D. candidate at the University of Illinois at Urbana-Champaign advised by Brighten Godfrey. Her research interests lie in the areas of computer networking and systems with a focus on building application-aware self-optimizing systems through automated resource management. She is a winner of the Facebook Graduate Fellowship (2017-2019) and the Mavis Future Faculty Fellowship (2017-2018). She was invited to attend the Rising Stars in EECS workshop at MIT (2018).

Website: http://abdujyo2.web.engr.illinois.edu
<table>
<thead>
<tr>
<th>Date and Time</th>
<th>Room</th>
<th>Speaker</th>
<th>Title and Abstract</th>
</tr>
</thead>
</table>
| 2/15 Thursday 1:30PM| SC2405| Wei Yang   | **Title:** Adversarial-Resilience Assurance for Mobile Security Systems  
**Abstract:** For too long, researchers have often tackled security in an attack-driven, ad hoc, and reactionary manner with large manual efforts devoted by security analysts. In order to make substantial progress in security, I advocate to shift such manner to be automated, intelligent, and adversarial resilient. Over the course of my Ph.D. research, I have built security systems incorporating intelligent security techniques based on program analysis, natural language processing, and machine learning, and I have developed corresponding defenses and testing methodologies to guard against emerging attacks specifically adversarial to these newly-proposed security techniques. In this talk, I will first highlight two of these systems for mobile security: AppContext and WHYPER. Then I will show how to generate adversarial inputs for testing and further strengthening these systems. I will conclude by discussing how future research efforts can leverage the interplay between AI and security techniques toward a defense-driven security ecosystem. |
| 2/23 Friday 10am    | SC3403| Chao Zhang | **Title:** Knowledge Cube Construction from Massive Social Sensing Data  
**Abstract:** Social sensing data are massive and ubiquitous. Effective and scalable analytics of social sensing data can be game changing for urban science, business, healthcare, and homeland security. However, such data pose great challenges to computer science research since they are often unstructured, fragmented, noisy, and intermingled with rich contexts. In this talk, I will introduce a systematic framework, KnowCube, that addresses the above challenges by turning unstructured, noisy social sensing data into a structured, multidimensional knowledge cube. In particular, I will discuss in detail how to solve two key problems for knowledge cube construction: (1) how to extract events from noisy social sensing data; and (2) how to organize unstructured events into a multidimensional cube structure without supervision. KnowCube serves as a versatile and easy-to-use knowledge engine that can harness the power of social sensing for many applications. Finally, I will share some future research directions on better knowledge cube construction and building next-generation intelligent systems with the knowledge cube. |
| 3/9 Friday 10am     | CSL301| Izzat El Hajj | **Title:** Building Programming Systems in a World of Increasing Heterogeneity  
**Abstract:** |
The breakdown of Dennard Scaling and slowing down of Moore's Law has led to an explosion of new processor and memory technologies which is making computing systems evolve to become increasingly heterogeneous. We are seeing GPUs, FPGAs, and special purpose accelerators become central parts of systems, as well as a growing interest in persistent byte-addressable memories and near-memory acceleration. While these technologies provide massive performance gains and energy savings that are not possible on traditional systems, they tend to be very tedious to program which introduces a heavy burden on software developers and presents a significant barrier to adoption. It is therefore critical that these hardware innovations be met with software innovations that facilitate programmability.

In this talk, I will discuss my work on building programming systems (languages, compilers, runtimes, OS support) for emerging processor and memory technologies. My talk will focus on two particular systems: (1) a compiler and runtime for improving performance and programmability of irregular applications on GPUs, and (2) a novel programmable accelerator and compiler that leverage analog computing via memristive crossbars to accelerate deep learning workloads. I will also discuss my future directions in both lines of work.

Bio:

Izzat is a PhD candidate in Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, and a member of the IMPACT Research Group working with Prof. Wen-mei Hwu. Izzat's research interests are in building programming systems for emerging processor and memory technologies. He has worked on programming systems for GPUs tackling issues of performance portability (CGO'15, MICRO'16), irregular application optimization (MICRO'16), and collaborative execution via unified virtual memory (ISPASS'17). For his work on GPU programming systems, he holds the Dan Vivoli Endowed Fellowship '17-'18. He has also worked on programming systems for emerging resistive memory technologies, tackling the issue of persistent object representation (ASPLOS'16, OOPSLA'17) and the use of memristive crossbars for accelerating deep learning workloads (in submission). For the former, he received the HiPEAC paper award and has submitted multiple patent applications. Izzat received his BE in Electrical and Computer Engineering in 2011 at the American University of Beirut (AUB), where he graduated with high distinction and received the Distinguished Graduate Award.
<table>
<thead>
<tr>
<th>Date and Time</th>
<th>Room</th>
<th>Speaker</th>
<th>Title and Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>February 9 (Thursday) 330 PM</td>
<td>2405 SC</td>
<td>Matt Sinclair</td>
<td>Title: Efficient Coherence and Consistency for Specialized Memory Hierarchies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Abstract:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>As the benefits from transistor scaling slow down, specialized accelerators and heterogeneous computing are becoming increasingly important because they can significantly improve performance and energy efficiency for specific applications. An efficient and easy-to-program memory and communication architecture is critical in achieving the promise of such systems. Traditionally, accelerators in heterogeneous systems used discrete address spaces and employed specialized memories, e.g., scratchpads, for specific access patterns. These attributes make these systems difficult to program and inefficient in the presence of high data reuse and fine-grained synchronization – traits that are common in emerging applications such as graph analytics workloads. My thesis resolves these inefficiencies with cross-cutting research that rethinks the software, hardware, and hardware-software interface of heterogeneous systems. Underlying my work is the efficient support of a global address space across all accelerator memories (for easier programming), an efficient cache coherence protocol (for efficient hardware), and a familiar memory consistency model (for an appropriate hardware-software interface).</td>
</tr>
</tbody>
</table>
First, I consider heterogeneous systems that have recently started to support global address spaces. I expose the imbalance between coherence and consistency in such systems. Current systems use simple, software-based coherence protocols that require heavyweight actions at synchronization points. To deal with this, industry has moved to complex consistency models that use scoped synchronization, making consistency models for heterogeneous systems even more complex than the already complicated CPU consistency models. I introduce a low overhead cache coherence protocol, DeNovo, that adjusts the imbalance and enables heterogeneous systems to use the standard, simpler data-race-free (DRF) consistency model.

Second, I explore a further source of complexity in consistency models: relaxed atomics. These have been the Achilles heel of CPU consistency models -- they have the promise of higher performance but have no known formal semantics. Heterogeneous systems’ inefficient support for atomics makes using relaxed atomics particularly tempting. I extend the DRF consistency model to retain the efficiency benefits of relaxed atomics and provide better semantics for the common use cases of relaxed atomics in heterogeneous systems. Third, current systems continue to support specialized memories in private address spaces, which negate some of the benefits of specialization. I integrate these specialized memories into the global address space while retaining the benefits they provide. Overall, my research introduces a more efficient and easier-to-program heterogeneous memory hierarchy that significantly improves both performance and energy compared to the state-of-the-art heterogeneous systems.

Bio:
Matt Sinclair is a doctoral candidate in the Department of Computer Science at the University of Illinois at Urbana-Champaign. He is interested in computer architecture and systems, with a current focus on building efficient memory hierarchies for heterogeneous systems. His papers at the 2015 International Symposium on Computer Architecture (ISCA) and 2015 International Symposium on Microarchitecture (MICRO) were recognized as 2016 IEEE Micro Top Picks Honorable Mentions. He is the recipient of a Qualcomm Innovation Fellowship, two Mavis Future Faculty Fellowships, the Feng Chen Memorial Award, the W. J. Poppelbaum Award, and a Saburo Muroga Fellowship. He was also selected to attend and present his research at the 2016 Heidelberg Laureate Forum. He received a BS in Computer Science with Honors & Computer Engineering (2009) and a MS in Electrical Engineering (2011) from the University of Wisconsin-Madison.

February 16 (Thursday)
3.30 PM
2405 SC

Renato Mancuso

Title: Safe, Real-Time Software Reference Architectures for Cyber-Physical Systems

Abstract:

There has been an uptrend in the demand and need for complex Cyber-Physical Systems (CPS), such as self-driving cars, unmanned aerial vehicles (UAVs), and smart manufacturing systems for Industry 4.0. CPS often need to accurately sense the surrounding environment by using high-bandwidth acoustic, imaging and other types of sensors; and to take coordinated decisions and issue time critical actuation commands. Hence, temporal predictability in sensing, communication, computation, and actuation is a fundamental attribute. Additionally, CPS must operate safely even in spite of software and hardware misbehavior to avoid catastrophic failures. To satisfy the increasing demand for performance, modern computing platforms have substantially increased in complexity; for instance, multi-core systems are now mainstream, and partially re-programmable system-on-chip (SoC) have just entered production. Unfortunately, extensive and unregulated sharing of hardware resources directly undermines the ability to guarantee strong temporal determinism in modern computing platforms. Novel software architectures are needed to restore temporal correctness of complex CPS when using these platforms.
My research vision is to design and implement software architectures that can serve as a reference for the development of high-performance CPS, and that embody two main requirements: temporal predictability and robustness. In this talk, I will address the following questions concerning modern multi-core systems: Why application timing can be highly unpredictable? What techniques can be used to enforce safe temporal behaviors on multi-core platforms? I will also illustrate possible approaches for time-aware fault tolerance to maximize CPS functional safety. Finally, I will review the challenges faced by the embedded industry when trying to adopt emerging computing platforms, and I will highlight some novel directions that can be followed to accomplish my research vision.

Bio:

Renato Mancuso is a doctoral candidate in the Department of Computer Science at the University of Illinois at Urbana-Champaign. He is interested in high-performance cyber-physical systems, with a specific focus on techniques to enforce strong performance isolation and temporal predictability in multi-core systems. He has published around 20 papers in major conferences and journals. His papers were awarded a best student paper award and a best presentation award at the Real-Time and Embedded Technology and Applications Symposium (RTAS) in 2013 and 2016, respectively. He was the recipient of a Computer Science Excellence Fellowship, and a finalist for the Qualcomm Innovation Fellowship. Some of the design principles for real-time multi-core computing proposed in his research have been officially incorporated in recent certification guidelines for avionics systems. They have also been endorsed by government agencies, industries, and research institutions worldwide. He received a B.S. in Computer Engineering with honors (2009) and a M.S. in Computer Engineering with honors (2012) from the University of Rome “Tor Vergata”. 
Title: Effort-Light StructMine: Turning Massive Corpora into Structures

Abstract:
The real-world data, though massive, are hard for machines to resolve as they are largely unstructured and in the form of natural-language text. One of the grand challenges is to turn such massive corpora into machine-actionable structures. Yet, most existing systems have heavy reliance on human effort in the process of structuring various corpora, slowing down the development of downstream applications.

In this talk, I will introduce a data-driven framework, Effort-Light StructMine, that extracts structured facts from massive corpora without explicit human labeling effort. In particular, I will discuss how to solve three StructMine tasks under Effort-Light StructMine framework: from identifying typed entities in text, to fine-grained entity typing, to extracting typed relationships between entities. Together, these three solutions form a clear roadmap for turning a massive corpus into a structured network to represent its factual knowledge. Finally, I will share some directions towards mining corpus-specific structured networks for knowledge discovery.

Bio:
Xiang Ren is a Computer Science PhD candidate at University of Illinois at Urbana-Champaign, working with Jiawei Han and the Data and Information System (DAIS) Research Lab. Xiang’s research develops data-driven methods for turning unstructured text data into machine-actionable structures. More broadly, his research interests span data mining, machine learning, and natural language processing, with a focus on making sense of massive text corpora. His research has been recognized with a Google PhD Fellowship, Yahoo!-DAIS Research Excellence Award, C. W. Gear Outstanding Graduate Student Award, and has been transferred to US Army Research Lab, NIH, Microsoft, Yelp and TripAdvisor.
February 24 (Friday) 1:30 PM  
2405 SC  
Man-Ki Yoon  

Title: CPS Security: Algorithms, Analysis and Experimental Validation  

Abstract: The increased computational power and connectivity in modern Cyber-Physical Systems (CPS) inevitably introduce more security vulnerabilities. CPS poses unique security challenges, as such systems are required to meet stringent requirements such as timing constraints as well as strong safety requirements. On the other hand, it provides defenders with an opportunity to take advantage of the design and implementation constraints and the tight coupling of cyber and physical components to deter attackers. In this talk, we will discuss how such intrinsic characteristics of CPS can be used as an asymmetric advantage to detect security attacks to safety-critical CPS. We will particularly discuss (a) modeling and reasoning about the logical (temporal and spatial) and physical behaviors of CPS, (b) architectural and operating-systems supports for trusted, efficient run-time behavior monitoring, and (c) attack-resilient architectures.

Bio: Man-Ki Yoon is a PhD candidate in Computer Science at the University of Illinois at Urbana-Champaign. His research interests are in analytic tools and system design principles for secure cyber-physical and real-time embedded systems, applying computer architecture, real-time scheduling, and statistical learning techniques. He is a recipient of Qualcomm Innovation Fellowship, Qualcomm Roberto Padovani Scholarship, and IntelPhD fellowship.

March 8 (Wednesday) 1:00 PM  
2405 SC  
Snigdha Chaturvedi  

Title: Structured Approaches to Natural Language Understanding  

Abstract: Despite recent advancements in Natural Language Processing, computers today cannot understand text in the ways that humans can. My research aims at creating computational methods that not only read but also interpret and reason about text. To accomplish this, I develop machine-learning methods that incorporate different sources of social context, linguistic structure and semantic knowledge while processing text. In this talk, I exhibit this by discussing two specific applications of natural language understanding that focus on comprehension of narratives: (i) Choosing correct endings to stories, and (ii) Identifying inter-personal relationships from narratives.
Automatic narrative comprehension is a fundamental challenge in Natural Language Understanding, and can enable computers to understand social norms, human behavior and common sense by processing large corpora of such texts. In the first part of the talk, I present a model that attempts to understand a story on three semantic axes: (i) its sequence of events, (ii) its emotional trajectory, and (iii) its plot consistency. We judge the model’s understanding by inquiring if, like humans, it can develop an expectation of what will happen next in a story and predict the correct ending from possible alternatives.

In the next part of the talk I address another important aspect of Natural Language Understanding: identifying social relationships from unstructured text. Understanding such relationships is essential for developing an understanding of people's goals, actions and expected behavior in stories. We develop structured models that incorporate linguistic as well as contextual cues for capturing the evolving nature of human relationships. We automatically discover various types of relationships in a data-driven manner. I conclude with a discussion of future directions, and some real-world scenarios that would gain from such advancements in natural language understanding, including social networks, discussion fora, intelligent virtual assistants, and artificial tutors.

Bio: Snigdha Chaturvedi is a postdoctoral fellow at University of Illinois, Urbana Champaign, working with Professor Dan Roth. She specializes in the field of Natural Language Processing with emphasis on developing methods for natural language understanding. Her research has been recognized with the IBM Ph.D. Fellowship (twice), a best paper award at NAACL, and first prize at ACM student research competition held at Grace Hopper Conference. She completed her Ph.D. in Computer Science at University of Maryland, College Park in 2016 and Bachelors in Technology from Indian Institute of Technology, Kanpur in 2009. She has previously held a position as a Blue Scholar at IBM Research, India.
<table>
<thead>
<tr>
<th>March 9 (Thursday) 4:00pm-5:30pm</th>
<th>2405 SC</th>
<th>Yingyan Lin</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong> Energy-efficient Systems for Information Processing and Transfer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| **Abstract:** Machine learning (ML) algorithms are increasingly pervasive in tackling the data deluge of the 21st Century. Current ML systems adopt either a centralized cloud computing or a distributed mobile computing paradigm. In both paradigms, the challenge of energy efficiency is drawing increased attention. In cloud computing, data transfer due to inter-chip, inter-board, inter-shelf and inter-rack communications (I/O interface) within data centers is one of the dominant energy costs. This will only intensify with the growing demand for increased I/O bandwidth for high-performance computing in data centers. On the other hand, in mobile computing, energy efficiency is the primary design challenge, as mobile devices have limited energy, computation and storage resources. This challenge is being exacerbated by the need to embed ML algorithms for enabling local inference capabilities.

In this talk, I will present system-to-circuit approaches for addressing these energy efficiency challenges. First, I will describe the design of a 4-bit, 4 GS/s bit-error-rate optimal analog-to-digital converter in 90nm CMOS and its use in realizing an energy-efficient 4Gb/s serial link receiver for I/O interface in data centers. Next, I will describe two techniques that can potentially enable on-device deployment of convolutional neural networks (CNNs) by significantly reducing the energy consumption via algorithmic/architectural innovation. Finally, I will identify future research directions in the emerging area of machine learning on resource-constrained silicon platforms.

**Bio:** Yingyan Lin is a Ph.D. candidate in the Electrical and Computer Engineering Department at the University of Illinois at Urbana-Champaign under the advisement of Professor Naresh Shanbhag. She expects to receive her Ph.D. degree in June 2017. Her research includes analog and mixed signal circuits for I/O interfaces, error resiliency techniques, VLSI circuits and architectures for machine learning on resource-constrained silicon platforms. She has 11 peer-reviewed publications as the first author on the subject and designed three high-speed interface circuit IPs for large flat panel display applications that were acquired by TOSHIBA Microelectronics Corporation in Japan. She received the 2016 IEEE International Workshop on Signal Processing Systems second place Best Student Paper Award and is the recipient of the 2016-2017 Robert T. Chien Memorial Award for Excellence in Research at UIUC.
### 2015-2016 Schedule

Organized by Madhusudan Parthasarathy and Darko Marinov

<table>
<thead>
<tr>
<th>Date and Time</th>
<th>Room</th>
<th>Speaker</th>
<th>Title and Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb 8 (Monday)</td>
<td>3405SC</td>
<td>Parisa Kordjamshidi</td>
<td>Title: Declarative Learning based Programming for Structured Machine Learning in Natural Language Processing</td>
</tr>
<tr>
<td>4PM</td>
<td></td>
<td></td>
<td>Developing intelligent problem solving systems that deal with real world messy data requires addressing a range of scientific and engineering challenges. Conventional programming languages offer no help to application programmers that attempt to make use of real world data, and reason about it in a way that involves learning interdependent concepts from data, incorporating existing models, and reasoning about them. Over the last few years the research community has tried to address these problems from multiple perspectives, most notably various approaches based on Probabilistic programming, Logical programming and integrated paradigms. In this talk I present Saul, a new declarative learning based programming (DeLBP) language that aims at facilitating the design and development of intelligent real world applications that use machine learning and reasoning. Our new language addresses the following challenges: Interaction with messy data; Specifying the problem at a high level i.e. application level; Dealing with uncertainty in data and knowledge; Supporting structured learning and reasoning while considering expert knowledge that is represented declaratively. An additional advantage of such a paradigm is generating easily reusable models and code, henceforth increasing the replicability of research results. I exemplify the flexibility and the expressive power of this language using a number of applications in natural language processing domain.</td>
</tr>
</tbody>
</table>
The past decade has witnessed remarkable progress in image-based, data-driven vision and graphics. However, existing approaches often treat the images as pure 2D signals and not as a 2D projection of the physical 3D world. As a result, a lot of training examples are required to cover sufficiently diverse appearances and inevitably suffer from limited generalization capability. In this talk, I will present "inference-by-composition" approaches to overcome these limitations by modeling and interpreting visual signals in terms of physical surface, object, and scene. I will show how we can incorporate physically grounded constraints in a non-parametric optimization framework for (1) revealing the missing parts of an image due to removal of a foreground or background element, (2) recovering high spatial frequency details that are not resolvable in low-resolution observations, and (3) discovering multiple approximately linear structures in extremely noisy videos with an ecological application to bird migration monitoring at night. The resulting algorithms are simple and intuitive while achieving state-of-the-art performance without the need of training on an exhaustive set of visual examples. I will end my talk with a brief discussion of some key challenges and opportunities in visual learning with weak supervision.

Bio: Jia-Bin Huang is a Ph.D. candidate in the Department of Electrical and Computer Engineering at University of Illinois, Urbana-Champaign advised by Prof. Narendra Ahuja. His research interests include computer vision, computer graphics, and machine learning with a focus on visual analysis and synthesis with physically grounded constraints. His research received the best student paper award in IAPR International Conference on Pattern Recognition (ICPR) in 2012 for the work on computational modeling of visual saliency and the best paper award in the ACM Symposium on Eye Tracking Research and Applications (ETRA) in 2014 for work on learning-based eye gaze tracking. Huang is the recipient of the UIUC Graduate College Dissertations Completion Fellowship (2015), Thomas and Margaret Huang Award for Graduate Research (2015), Beckman Cognitive Science/Artificial Intelligence Award (2015), Sundaram Seshu Fellowship (2014), MOE Technologies Incubation Scholarship (2014), and the PURE Best Research Mentor Award (2012).

Personal website: http://www.jiabinhuang.com
# 2014-2015 Schedule

Organized by Madhusudan Parthasarathy and Darko Marinov

<table>
<thead>
<tr>
<th>Date and Time</th>
<th>Room</th>
<th>Speaker</th>
<th>Title and Abstract</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Feb 4 (Wed) 10am</th>
<th>2405</th>
<th>Milos Gligoric</th>
<th>Regression Testing: Theory and Practice</th>
</tr>
</thead>
</table>

Developers often build regression test suites that are automatically run for each code revision to check that code changes did not break any functionality. While regression testing is important, it is also expensive due to both the number of revisions and the number of tests. For example, Google recently reported that they observed a quadratic increase in daily test-suite run time (a linear increase in the number of revisions per day and a linear increase in the number of tests per revision).

In this talk, I present a technique, called Ekstazi, to substantially reduce test-suite run time. Ekstazi introduces a novel approach to regression test selection, which runs only a subset of tests whose dependencies may be affected by the latest changes; Ekstazi keeps file dependencies for each test. Ekstazi also speeds up test-suite runs for software that uses modern distributed version-control systems; by modeling different branch and merge commands directly, Ekstazi computes test sets that can be significantly smaller than the entire test suite. I developed Ekstazi for JVM languages and evaluated it on several hundred revisions of 32 open-source projects (totaling 5M lines of code). Ekstazi can reduce test-suite run time an order of magnitude, including runs for merge revisions. Finally, only a few months after the initial release, Ekstazi was adopted and used daily by many developers from several open-source projects, including Apache Camel, Commons Math, and CXF.

Bio: Milos Gligoric is a PhD candidate in Computer Science at the University of Illinois at Urbana-Champaign (UIUC). His research interests are in software engineering and formal methods, especially in designing techniques and tools that improve software quality and developers' productivity. His PhD work has explored test input generation, test quality assessment, testing concurrent code, and regression testing. He won an ACM SIGSOFT Distinguished Paper Award (ICSE 2010), and three of his papers were invited for a journal submission. He was awarded the Saburo Muroga Fellowship (2009), the C.L. and Jane W-S. Liu Award (2012), and the C. W. Gear Outstanding Graduate Award (2014) from the UIUC Department of Computer Science, and the Mavis Future Faculty Fellowship (2014) from the UIUC College of Engineering. He did internships at NASA Ames, Intel, Max Planck Institute for Software Systems, and Microsoft Research. Milos holds a BS (2007) and MS (2009) from the University of Belgrade, Serbia.
The desired output in many machine learning tasks is a structured object such as a tree, a clustering of nodes, or a sequence. Learning accurate prediction models for such problems requires training on large amounts of data, making use of expressive features and performing global inference that simultaneously assigns values to all interrelated nodes in the structure. All these contribute to significant scalability problems. We describe a collection of results that address several aspects of these problems – by carefully selecting and caching samples, structures, or latent items.

Our results lead to efficient learning algorithms for structured prediction models and for online clustering models which, in turn, support reduction in problem size, improvements in training and evaluation speed and improved performance. We have used our algorithms to learn expressive models from large amounts of annotated data and achieve state-of-the-art performance on several natural language processing tasks.

Bio: Kai-Wei Chang is a doctoral candidate advised by Prof. Dan Roth in the Department of Computer Science, University of Illinois at Urbana-Champaign. His research interests lie in designing practical machine learning techniques for large and complex data and applying them to real world applications. He has been working on various topics in Machine learning and Natural Language Processing, including large-scale learning, structured learning, coreference resolution, and relation extraction. Kai-Wei was awarded the KDD Best Paper Award in 2010 and won the Yahoo! Key Scientific Challenges Award in 2011. He was one of the main contributors of a popular linear classification library, LIBLINEAR.
Abstract: The scale of modern geometric data sets necessitates fast algorithms. In this talk I will discuss several optimal linear (or near linear) time algorithms, which work by quickly throwing out and summarizing data, creating a compact sketch of the input.

In the first part of the talk I will present a general framework called Net and Prune, which provides linear time approximation algorithms for a large class of well studied geometric optimization problems, such as k-center clustering and farthest nearest neighbor. The new approach is robust to variations in the input problem, and yet it is simple, elegant, and practical. In particular, many of these well studied problems which easily fit into our framework, either previously had no linear time approximation algorithms, or required rather involved algorithms and analysis.

In the second part of the talk I will discuss contour trees, which provide a compact description of the level set behavior of structured geometric data. These trees are used in HPC applications such as combustion, chemical and fluid mixing simulations, where they are used to both summarize and explore the significantly larger simulation data. Here I will discuss an instance optimal algorithm for their computation, which runs in linear time when the tree is balanced.

Bio: Benjamin Raichel is a PhD student in the Computer Science Department at the University of Illinois, Urbana-Champaign. His research interests are in algorithms and their applications. In particular he has developed fast and practical algorithms for a variety of geometric problems. He is currently funded by the UIUC Dissertation Completion Fellowship, and previously was awarded the Andrew and Shana Laursen Fellowship (2011-12) from the Department of Computer Science. Benjamin holds an MS degree in Computer Science (2011), as well as a BS degree with highest distinction in both Math and Physics (2009), from the University of Illinois.
Machine learning algorithms have become pervasive in multiple domains and have started to have impact in applications. Nonetheless, a key obstacle in making learning protocol realistic in applications is the need to supervise them, a costly process that often requires hiring domain experts. However, while annotated data is difficult to get, we have available large amounts of data from the Web. In this talk, I will introduce learning paradigms which use existing world knowledge to "supervise" machine learning algorithms. By "world knowledge" we refer to general-purpose knowledge collected from the Web, and that can be used to extract both common sense knowledge and diverse domain specific knowledge and thus help supervise machine learning algorithms. I will discuss two projects, demonstrating that we can perform better machine learning and text data analytics by adapting general-purpose knowledge to domain specific tasks. For the first project, I will introduce the dataless classification algorithm which requires no labeled data to perform completely unsupervised text classification. In this case, the Wikipedia knowledge is embed to represent the text documents and the category labels into the same semantic space. For the second project, I will discuss how to perform hierarchical clustering of domain-specific short texts, e.g., Web queries and tweets, using a probabilistic concept based knowledge base, Probase. In both cases, we provide realistic and scalable algorithms to address large scale and fundamental text analytics problems.

Bio: Dr. Yangqiu Song is a post-doctoral researcher at the Cognitive Computation Group at the University of Illinois at Urbana-Champaign. Before that, he was a post-doctoral fellow at Hong Kong University of Science and Technology and visiting researcher at Huawei Noah's Ark Lab, Hong Kong (2012-2013), an associate researcher at Microsoft Research Asia (2010-2012) and a staff researcher at IBM Research China (2009-2010) respectively. He received his B.E. and Ph.D. degrees from Tsinghua University, China, in July 2003 and January 2009, respectively. His current research focuses on using machine learning and data mining to extract and infer insightful knowledge from big data. The knowledge helps users better enjoy their daily living and social activities, or helps data scientists do better data analytics. He is particularly interested in working on large scale learning algorithms, on natural language understanding, text mining and visual analytics, and on knowledge engineering for domain applications.
<table>
<thead>
<tr>
<th>Feb 25 (Wed) 2:00 pm</th>
<th>3403</th>
<th>Parasara Sridhar Duggirala</th>
<th>Dynamic Analysis of Cyber-Physical Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Progress in computation and communication technologies has made it easier to integrate software in all walks of life. The social, economical, and environmental benefits of integrating software into avenues such as avionics, automobiles, power grid, and medicine lead to the rise of CPS as an important area of research. However, bugs in software systems deployed in such safety-critical scenarios can lead to loss of property and in some cases life. In this talk, I will present dynamic analysis technique for formally verifying annotated Cyber-Physical Systems and prove the absence of bugs. The annotations, called discrepancy functions, are extensions of proof certificates for analyzing convergence or divergence of systems. One of the key advantages of dynamic analysis is that it leverages the testing procedures which are the only known scalable way of ensuring the system specification. I have developed a tool C2E2 that implements this technique and verifies temporal properties of CPS. C2E2 has been applied to verify alerting mechanisms in parallel aircraft landing protocol developed by NASA and to verify specification of powertrain control system presented as a verification challenge problem by Toyota.</td>
</tr>
<tr>
<td>Bio: Parasara Sridhar Duggirala is a PhD Candidate in Computer Science Department at the University of Illinois at Urbana Champaign (UIUC). His main research interests are in Cyber-Physical Systems, Formal Methods, and Control Theory. His paper on Safety Verification of Linear Control Systems won the best paper award at the International Conference on Embedded Software (conducted as part of ESWeek) 2013. He also received Feng Chen Memorial Award in Software Engineering from the Department of Computer Science at UIUC, and has also been selected as a Young Research to attend Heidelberg Laureate Forum. He did internships at NEC Labs America and SRI International. He received his B.Tech (2009) from Indian Institute of Technology Guwahati.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Learning Invariants for Software Reliability and Security

The central problem in software verification today is in establishing invariants that prove the system reliable or secure. Current technology requires invariants to be specified manually and is the bottleneck for adoption of verification in mainstream programming. I will describe recent advances in synthesizing invariants using machine learning techniques, embracing an inductive rather than a deductive approach to this problem. In particular, I propose a new learning model called ICE (involving Implication Counter-Examples) and develop new machine learning ICE algorithms that effectively synthesize invariants. I’ll also describe ways of specifying infinite enumerated invariants using finite representations of it. These invariant generation techniques are applied to security and reliability domains including ExpressOS, a secure mobile operating system, GPU kernels, cloud systems, and the verification of the responsiveness of the full USB Windows phone driver.

Bio: Pranav Garg is a PhD candidate in the Department of Computer Science at the University of Illinois at Urbana-Champaign. His research interests span areas at the intersection of programming languages, formal methods, and software engineering. His PhD research, in particular, focuses on automating verification for building reliable and secure software systems. He received his B.Tech (2009) in Computer Science and Engineering from the Indian Institute of Technology Kanpur.