ICRI-CI Retreat 13-14 May 2014 – Agenda+Abstracts

Start	Dur.	End	Theme	Project	Speaker
Day1	13-May				
8:00	1:00	9:00	Gathering / Breakfast		
9:00	0:30	9:30	Opening / welcome		**Ronny Ronen, Shalom Goldenberg**
9:30	3:00	12:30	**Advanced Machine Learning**		
				Deep Learning: Why and How One of the most significant recent developments in machine learning has been the resurgence of “deep learning”, usually in the form of artificial neural networks. I'll discuss the importance of deep learning and will describe recent work on novel techniques for training such networks.	**Shai Shalev Shwarts**
				Silence is Golden: Distributed Learning with Minimal Communication Learning in distributed environments is an important challenge in big-data applications, but can be hampered by slow communication between machines. I'll discuss the challenge of learning with minimal communication, and present a new Newton-type method for stochastic optimization and learning, whose communication requirements provably *shrink* with the data size.	**Ohad Shamir**
				Distributed learning in networks One of the challenges in learning over networks is scalable and distributed parameter estimation. A main building block in such problems is graphical models that characterize the natural statistical conditional independence structure within the data across the network. In this talk, we will consider algorithms that exploit this information in an efficient manner based on modified marginal distributions. Typical applications include monitoring in large scale sensor networks. Joint work with Z. Meng, D. Wei, A. O. Hero and I. Soloveychik.	**Ami Wiesel**
				Learning many tasks with a single teacher We introduce a new multi-problem framework, in which K online learners are sharing a single teacher whom is limited in her bandwidth. On each round, each of the K learners receives an input, and makes a prediction about the label of that input. Our algorithm then decides which of the K inputs will be annotated. The learner that receives the feedback (label) may update its prediction rule, and we proceed to the next round. We develop few algorithms and analyze their performance in the worst-case setting. Additionally, we show that our algorithm can be used to solve two bandits problems: contextual bandits, and dueling bandits with context, both allowed to decouple exploration and exploitation. Empirical study shows that our algorithm outperforms algorithms that use uniform allocation, and essentially makes more (accuracy) for the same labor of the annotator.	**Koby Crammer**
10:50	0:30	11:20	Break		
				Holistic NLP Parsing via Sampling One of the key goals of machine learning is to build classifiers from labeled data. Earlier work in the field focused on cases where labels were binary or confined to a small set of possible classes. However, in recent years, there has been considerable interest in tasks where labels have a more complex structure. Example tasks are semantic segmentation of images, machine translation, and syntactic parsing. The challenge in these is that there are exponentially many possible labels and that these need to satisfy complex global	**Amir Globerson**

				constraints (e.g., the translation should be grammatical, the segmentation should make sense physically etc). Learning and enforcing such "holistic" constraints is an ongoing algorithmic and empirical challenge. In this talk I will focus on the problem of syntactic dependency parsing, a key problem in natural language processing, which lies at the core of many NLP applications such as translation, entity extraction and semantic analysis. I will describe an approach to parsing that is based on sampling and simplified learning rules. The resulting system achieves state of the art results, and is currently the best performing system across as set of 14 languages	
				The intriguing relations between power, time, and information in planning and learning - a lesson for computer architecture? One of the recent developments in machine learning is understanding learning and planning under information constraints, which can include communication bandwidth, memory capacity, and computational complexity. We consider the first two using an information theoretic approach that can provide lower bounds on information processing in interactive systems as well in computers which interact with their environment or with other computers. I will first review our general framework for studying information constrained interactive systems. Then I will present a new quantitative tradeoff between computation time, power and information processing, which has interesting implications to computer design. Finally, I will argue that optimizing predictive information can suggest new type of computer architectures, where both inputs (sensing) and operations (actions) should scale with the available information about the future, suggesting a different type of hierarchal designs of computers.	**Naftali Tishby**
				Intel Machine Learning View	**Shai Fine**
12:30	1:00	13:30	**Imaging**		
				Edge Detection on a Computational Budget: A sublinear approach Consider the following problem. Your camera or other imaging hardware acquired a noisy and very large image with n^2 pixels, and your goal is to detect and localize all sufficiently long straight edges in it. However, you are operating under severe computational constraints (such as very low power computer to do so, real-time requirements, etc). Can this task be done extremely efficiently, even without processing all n^2 pixels ? If yes, what are the tradeoffs between statistical accuracy and computational run-time, and are there lower bounds on edge detection given such computational constraints? In this talk we'll show that the above is indeed possible, provide a novel framework and mathematical analysis of this problem, and present (possibly the first) sublinear algorithm for edge detection.	**Boaz Nadler**
				Inverse Volume Rendering with Material Dictionaries We present a reflectance display: a dynamic digital display capable of showing images and videos with spatially-varying, user-defined reflectance functions. Our display is passive- it operates by modulation of reflected light. As such, it does not rely on any illumination recording sensors, nor does it require expensive on-the-fly rendering. It reacts to lighting changes instantaneously and consumes only a minimal amount of energy. Our work builds on a wave optics approach to BRDF design combined with a programable liquid crystal spatial light modulator, retaining high resolution of approximately 160 dpi. Our approach enables the display of a wide family of angular reflectances, and it allows the display of dynamic content with time varying reflectance properties---"reflectance videos". We demonstrate the utility of our display with a diverse set of experiments including display of custom reflectance images and videos, interactive reflectance	**Anat Levin**

				editing, display of 3D content reproducing lighting and depth variation, and simultaneous display of two independent channels on one screen.	
				The 4th Channel - Image Statistics beyond RGB We use techniques that were successful in modeling natural image statistics to model the statistics of local patches of optical flow. Some aspects of what our models learn reinforces the common practice in optical flow estimation, but other aspects suggest novel algorithms that could lead to much better estimation.	**Yair Weiss**
13:30	1:00	14:30	Lunch		
14:30	2:00	16:30	**Visual Understanding**		
				On the development of better saliency detection algorithms The goal of salient object detection is to detect the foreground pixels in an image. Accurate detection is useful as input for many applications, including recognition, retrieval, editing, segmentation, compression, and more. In the first part of this talk we will describe state-of-the-art in saliency estimation solutions developed in our lab. The second part of the talk will focus on the evaluation of saliency detection algorithms. Clearly, the development of better algorithms relies heavily on the availability of tools for measuring the accuracy of the algorithms' output. We will show that the most commonly-used measures for assessing saliency maps do not provide a reliable evaluation. We will identify the causes and will propose a new measure that amends them.	**Lihi Zelnik-Manor**
				3D cameras in the service of psychiatry and neurology In recent years we see the emergence of technology based on 3D cameras (e.g. Microsoft Kinect and Intel Creative), which facilitates the automatic tracking of body movement and facial expressions with resolution and ease previously unavailable. This opens the doors to many applications, including medical applications that could offer standardized evaluation of psychiatric and neurological conditions. In the last year we had investigated two such applications: (i) the quantitative characterization of non-verbal behavior in schizophrenic patients, which will allow us to develop automatic tools for describing and analyzing clinically relevant measures of this illness. (ii) The development of an objective and automatic procedure to assess the severity of dyskinesia, which is a debilitating complication of chronic levodopa therapy of Parkinson’s disease. In this talk I will describe some of the technical challenges we face when addressing these problems. I will outline our preliminary results, suggesting that this line of research may lead to real progress in areas which currently rely mostly on human expertise and very little automation.	**Daphna Weinshall**
				What can we do with egocentric video? If wearable cameras will become popular, and "life logging" videos will be recorded by many people, we will have a problem what to do with all recorded video, and how to access and use it. Unlike regular video, the camera is not aimed at the location of interest and the record button is not pressed at the time of interest, and a lot if irrelevant video is recorded. This talk will present two cases to process such egocentric video. (a) In case of a single camera recording life logging video. We present a way to segment the video into several classes such as "walking", "driving", "staying in place", etc. This will allow a user, at a stage of viewing such video, to skip very quickly to the activity of interest. (b) When multiple cameras are watching the same event, such as a concert, a lecture, etc - we propose to automatically curate a single video of the event making use of all cameras, selecting the most relevant video having the highest quality.	**Shmuel Peleg**

				Compressed Sensing for Natural Images Compressed sensing (CS) refers to a branch of applied mathematics which is based on the surprising result whereby signals that are exactly “k-sparse” (i.e. can be represented by at most k nonzero coefficients in some basis) can be exactly reconstructed using a small number of random measurements. Since natural images tend to be sparse in the wavelet basis, one of the motivating examples of CS has always been to reconstruct high resolution images from a small number of random measurements. Unfortunately, there are some significant deviations between the way that natural images behave and the assumptions of the dramatic theorems, and in fact random projections perform quite poorly when applied to real images. I will describe an alternative theory, which we call “Informative Sensing”, that seeks a small number of projections that are maximally informative given a known distribution over signals. I will show experimental results demonstrating that the informative projections indeed outperform random projections, but that the savings relative to more standard imaging methods are altogether rather modest. Joint work with Hyun Sung Chang and Bill Freeman.	**Yair Weiss**
				A new type of ConvNet Architecture for Visual Classification A large variety of challenging visual tasks are being successfully applied in the recent years through Deep Convolutional Nets. These includes categorization of pictures, detection of objects in pictures, localization of objects, face recognition, alignment, and more. However, DCNs have not changed much since their inception during the 80s despite the great strides in statistical machine learning since then. Most of the recent success of DCN is attributed to the great increase of raw computing power of the past two decades and not to any clever new theoretical revelation. For example, in the 80s it took weeks to train a DCN on the MNIST digit training set and today with off-the-helf DCN packages it can take merely few minutes. We will describe a new architecture of layers of successive convolutions and pooling that are borne out of kernel large margin principle from classical statistical learning. From the theory emerge three types of pooling layers and a natural initialization principle from unlabeled data. Experiments with the new architecture demonstrate competitive results to DCN with much fewer units and easier training. This work was jointly done with Nadav Cohen	**Amnon Shashua**
				Deep Leaning and its usage for Visual Understanding An emerging method in the utilization of big data in computer vision is through the set of techniques called Deep Learning. In these techniques, multilayer neural networks are often trained using millions of training samples. The top layers of the hierarchy are used as powerful high level representations of the input signal. In my talk, I will discuss recent advances. Specifically, I will discuss a face recognition system that achieves state of the art results and an innovative video analysis system that is able to count video events.	**Lior Wolf**
16:30	0:30	17:00		**Intel Visual Understanding View**	**Ofri Wechsler**
17:00	0:30	17:30	Free time		
17:30	2:30	20:00	**Poster Session + Dinner**		

Start	Dur.	End	Theme	Project	Speaker
Day2	14-May				
8:00	0:30	8:30	Gathering /Breakfast		
8:30	2:00	10:30	**Novel Heterogeneous Computing Platforms**		
				Research project overview	**Uri Weiser**
				H-EARtH (Hetero Energy Efficient Race-to-Halt) – from research to product	**Efi Rotem**
				Accelerators based on Associative Processing We investigate alternative architectures for high-end accelerators of machine-learning (ML) applications, focusing on in-memory computing. Sparse matrix-vector and matrix-matrix multiplications, as well as the inversion of sparse matrices, form the backbone of ML computations. Associative processing (AP) architectures demonstrate better suitability than GPU-like SIMD arrays for these low-arithmetic-intensity problems. AP eliminate hot spots and operate at lower power density, being more suitable for 3D stacking of multiple AP layers with multiple DRAM chips. GP-SIMD architectures combine massive and uniform SIMD arrays with scalar processors, sharing the same memory. They appear to be faster than AP for about the same power, but more comparative study is needed in the context of ML computations. We have also found that compressing sparse matrices and adding the burden of decompression to normal processing still saves time and energy and helps balance I/O and computations. Ultra low power architectures, critical for future high-end massively parallel servers, may be enabled by sub-threshold circuits, requiring careful matching of circuits, technology, architectures and design styles.	**Ran Ginosar**
				Semantic Locality and its usage in Memory Prefetching Data locality, or the correlation between memory addresses accessed by a program, is an artifact of program semantics that dictate the patterns in which it accesses memory. In this project, we establish the notion of "semantic locality" as a generalization of spatio-temporal locality and argue that semantic locality is orthogonal to the layout of data in memory. Semantic locality can thus assist in predicting accesses to irregular data structures. Consequently, we demonstrate a context-based locality predictor, a reinforcement learning approach to predicting future memory accesses based on program context, which comprise the machine state and compiler injected cues. The context-based predictor is used to construct a context-based memory prefetcher that improves the cache behavior of irregular memory workloads.	**Yoav Etsion**
				Memory Intensive Architecture Over the past years, new memory technologies such as RRAM, STT-MRAM, PCM etc., have emerged. These technologies, located in the metal layers of the chip, are relatively fast, dense, and power-efficient, and can be considered as memristors. Usually, the use of these devices has been limited to flash, DRAM, and SRAM replacement. This talk is focused on different uses of memristors. For example, new memory structures, different than the conventional memory hierarchy, opening opportunity to a new era in computer architecture – the era of Memory Intensive Computing. Memristors can also be integrated with CMOS in logic circuits. Alternatively, they can be used as a stand-alone logic, suitable to perform logic within the memory and provide opportunity for new computer architectures, different than classical von Neumann.	**Uri Weiser (Shahar Kvatinsky)**
10:30	0:30	11:00		**Intel Accelerators View**	**Debbie Marr**
11:00	0:30	11:30	Break		

11:30	1:00	12:30	**Cognition**		
				Identifying Authorships of very short texts using flexible patterns We present a system for identifying authors of very short texts, demonstrating its success on single Twitter tweet (Schwartz et al., EMNLP 2013). Our algorithm uses the recently developed "flexible patterns" technique, which is shown to be more robust compared to previous approaches while requiring less human annotation.	**Ari Rappoport (Roy Schwartz)**
				Providing Arguments in Argumentative Discussions Over the last fifteen years, argumentation has come to be increasingly central as a core study within Artificial Intelligence, in general, and multi-agent research in particular. However, only very few attempts have been made to study models of human argumentations. Such models can be used for supporting people or representing them in argumentation. Formal argumentation theories that are grounded in computational logic, consider arguments as abstract entities and study their interaction as introduced by Dung are not useful when modelling human argumentation. People do not adhere to the optimal, monolithic strategies that can be derived analytically. Their argumentation behavior is affected by a multitude of social and psychological factors. In some contexts, especially in law and medicine, where the goal of the argumentation is to reveal the truth, computational tools are available to help people reach the true conclusions. In deliberation dialogues where people express opinions and no objective truth exists, helping people or arguing with people is more difficult. We propose that the first step in the study of argumentation in human-computer interaction is to try to predict human argumentation choice, and we will discuss an extensive study toward this goal. Based on extensive empirical experiments, we will show that in addition to justification concepts that are commonly used in argumentation theory, the relevance of the argument to the deliberation and the psychological aspects are important features in the prediction process. We will conclude by providing directions for building automated agents that can support people in deliberations.	**Sarit Kraus**
				Pythia 2.0: A mobile service for social relationship enhancement Mobile services using location and social network graphs are not new. In Pythia 2.0 we aim to build on the capabilities of Pythia 1.0 and enhance a person's social life by merging their social graph with their location graph across multiple life dimensions. Our service will first recognize who your close friends are, and then proceeds to identify opportunities to get you and your friends to meet in the Real World (RW). The service takes into account friend closeness, time since last meeting, mean time between meetings with a friend, spatial distance between friends, future planned locations of friends, mean time to navigate to a meeting, current estimated travel time to a potential meeting with a friend, and location and time of the meeting after it. Once a potential meeting is identified, the service will notify both friends about the opportunity, remind them when they last met and if both agree, a calendar entry will be created and a reminder will be added to their meeting cue. When a meeting comes up, a notification will navigate the friends towards their meeting. The application depends upon a novel separation of function between mobile and the cloud, and incorporates novel calendar information extraction technology.	**Amnon Dekel**
12:30	0:30	13:00		**Intel Cognitive Computing View**	**Gadi Singer**
13:00	0:30	13:30	**Concluding remarks**		
13:30	1:00	14:30	Lunch		