ICRI-CI Retreat 5-6 May 2014 – Agenda + Abstracts (updated April 27, 2015)

Start	Dur.	End	Session
	Speaker		Title
Day 1, Tuesday, 5-May-2015			
9:00	0:30	9:30	**Opening**
	Ronny + Shalom		**Opening Notes**
9:30	2:40	12:10	**Architecture**
	Debbie Marr		**Architecture for Big Data & Machine Learning - Capstone overview**
	Uri Weiser		**Memory Traffic Reduction for Big Data & Machine Learning** The era of Big Data Computing is already here. Data centers today are reaching 1 million meter square each, while the power required to operate such center reaches close to 100 MWatts each. The electrical cost of such centers dominates operating expenses. In 2014, in the US alone, Data Centers energy consumption was at about 100 Billion KWh at a cost of close to $10B (at total of 15,000 Mega Watts). The Power Usage Efficiency – PUE – measuring total energy of data centers vs. IT energy, ranges between 1.2 to 2.0. This means that one Joule saved in computing technology (IT), saves around 1.5 Joule of energy consumed by a data center. Computing Technology energy is dominated by data movement; our past research (N. Magen 2004) showed that >60% of the energy consumed in computing chips is due to data movement. Furthermore, Stanford research (B. Dally, M. Horowitz) shows that the energy ratio between execution (e.g. 64 bit DP op – 20pJ) and moving data (DRAM access; 256bit Read/Write -16nJ) is 2-3 orders of magnitude. In this Capstones overview we will add the energy saving aspects to our previous research and describe the new paradigm in handling data movements in Big Data environment – "The Funnel"
	Ran Ginosar		**Accelerators for Big Data & Machine Learning** We address future "big data" where a single machine learning application needs to readily access one exabytes (10^18 bytes). A 100 MWatt data center would require an hour to access the entire data set once (due to power constraints). Machine learning tasks call for sparse matrix operations such as matrix-vector and matrix-matrix multiplications on either floating-point, integer or binary data. In our model, "big data machine learning" (after all funneling and data reductions have been applied) means that every item in the Exabyte data set needs to be connected with a large portion of all other data items of the same Exabyte data set (a super-linear operation). With only 100 MWatt, such computation would take days to merely access and move the data around, even before counting the required CPU power. We consider 3-D chips integrating massive storage with massively parallel accelerators in order to obtain 10-100x improvement in energy requirements for data access and processing of big data machine learning tasks. Associative Processors (AP) combine data storage and processing per each memory cell, and function simultaneously as both a massively parallel computing array and a memory. We have started with CMOS structure and more recently have investigated AP based on resistive CAM, scaling the AP from a few millions to a few hundreds of millions of processing units on a single silicon die. We also study GP-SIMD, a more conventional parallel array combined with massive memory and general processors on-die. The resistive memory technology is also applied to a resistive GP-SIMD, achieving two orders of magnitude improvement in density.
	Yoav Etsion		**Machine Learning for Architecture - Context-based Prefetching using Reinforcement Learning** Most modern memory prefetchers rely on spatio-temporal locality to predict the memory addresses likely to be accessed by a program in the near future. Emerging workloads, however, make increasing use of irregular data structures, and thus exhibit a lower degree of spatial locality. This makes them less amenable to spatio-temporal prefetchers.

			Our research introduces the concept of Semantic Locality, which uses inherent program semantics to characterize access relations. We show how, in principle, semantic locality can capture the relationship between data elements in a manner agnostic to the actual data layout, and we argue that semantic locality transcends spatio-temporal concerns. We further introduce the context-based memory prefetcher, which approximates semantic locality using reinforcement learning. The prefetcher identifies access patterns by applying reinforcement learning methods over machine and code attributes, which provide hints on memory access semantics.	
	Avinoam Kolodny		**Memory Intensive Architecture** Emerging resistive memory technologies enable novel memory structures, different than the conventional memory hierarchy. They also enable new types of high-density logic circuits, which may be provide computational capabilities within memory arrays. We view these capabilities as disruptive technologies, which may change the conventional design paradigm of computing systems. In particular, memory-intensive architectures, utilizing large volumes of dense memory structures integrated within logic and arithmetic circuits, may eliminate the memory-wall bottleneck, which currently dominates performance and power dissipation in computers. In this talk we describe the main concepts of memory-intensive architectures, and present results of a detailed evaluation of using multistate registers, based on memristive memory, for continuous flow multithreading in a processor pipeline.	
12:10	0:25	12:35	**Visual-I**	
	Ayellet Tal		**Saliency in visual data** Every visual scene tells a story. We aim to detect the essence of that story - its salient components. This talk will describe our work on saliency detection for various types of visual data: images, videos, surfaces, and point clouds. We will present some principles that guide our work and demonstrate its benefits through state-of-the-art results. We will also show some applications of our results, both in two dimensions and in three dimensions.	
12:35	1:55	14:30	**Lunch + Posters**	
14:30	3:30	18:00	**Deep Learning**	
	Boris Ginsburg		**Distributed Deep Learning Library - Capstone overview**	
	Naftali Tishby		**Optimal Deep Learning and the Information Bottleneck Method** Deep Neural Networks (DNNs) are analyzed via the theoretical framework of the information bottleneck (IB) principle. We first show that any DNN can be quantified by the mutual information between the layers and the input and output variables. Using this representation we can calculate the optimal information theoretic limits of the DNN and obtain finite sample generalization bounds. The advantage of getting closer to the theoretical limit is quantifiable both by the generalization bound and by the network's simplicity. We argue that both the optimal architecture, number of layers and features/connections at each layer, are related to the bifurcation points of the information bottleneck tradeoff, namely, relevant compression of the input layer with respect to the output layer. The hierarchical representations at the layered network naturally correspond to the structural phase transitions along the information curve. We believe that this new insight can lead to new sample complexity bounds and better deep learning algorithms.	
	Michael Zibulevsky		**Compressed sensing and computed tomography with deep learning** We show usefulness of deep feed-forward neural networks in two image restoration tasks. **Compressed sensing** is a concept of nonlinear signal / image recovery from restricted number of linear measurements. Classically this procedure is based on sparse signal representation in some signal dictionary. The goal is to create good underdetermined linear sensing matrix and afterward to recover the original signal from the under-sampled measurements. Usually classical sparse representation problem is solved at the second stage. Our idea is to train a neural network to find the optimal sensing matrix and the nonlinear restoration operator at once. The network maps an input to itself while the size of the first hidden layer is smaller than the size of the input vector, in accordance to the compression ratio. The experiments demonstrate improvement of state of the art in patch-wise mode. **Computed tomography** is a technique of image recovery from a set of projections (e.g. X-ray images), which are taken at different angles. It is widely used in medicine, science and engineering. Reconstruction methods,	

			analytic or iterative, are able to obtain several versions of reconstructed image with different bias/variance (smoothing/noise) tradeoff. Our idea is feed all these image versions as an input to neural network, and train it to estimate the original image. We improve state of the art results using this approach.	
	Lior Wolf		**Automatic Image Annotation using Deep Learning and Fisher Vectors** We present a system for tackling the holy grail of computer vision -- matching images and text and describing an image by an automatically generated text. Our system is based on combining deep learning tools for images and text, namely Convolutional Neural Networks, word2vec, and Recurrent Neural Networks, with a classical computer vision tool, the Fisher Vector. The Fisher Vector is modified to support hybrid distributions that are a much better fit for the text data. Our method proves to be extremely potent and we outperform by a significant margin all concurrent methods	
	Amnon Shashua		**SimNets: A Generalization of Convolutional Networks** The SimNet architecture consists of layers which are potentially more expressive than standard convolutional layers. The advantage is focused on allowing significantly more compact networks for the same level of accuracy of ConvNets and exploiting unsupervised data for (i) guiding network architecture, (ii) parameter initialization and (iii) requiring less labeled data for reaching same level of accuracy as ConvNets. We will show some experiments on state-of-the-art performance on CIFAR-10 while we explore the features above.	
	Shai Shalev Shwartz		**Rigorous algorithms for distributed deep learning** We describe and analyze a new approach for distributed training of deep learning. Our approach relies on an iterative process, where each iteration involves hunting of "interesting" examples followed by fast stochastic algorithm for making progress based on the newly collected examples.	
	Shie Mannor		**Outlier robust distributed learning** We propose a generic distributed learning framework for robust statistical learning on big contaminated data. The Distributed Robust Learning (DRL) framework can reduce the computational cost of traditional robust learning methods by several orders of magnitude. We provide a sharp analysis on the robustness of DRL, showing that DRL not only preserves the robustness of base robust learning methods, but also tolerates breakdowns of a constant fraction of computing nodes. Moreover, DRL can enhance the breakdown point of existing robust learning methods to be even larger than 50%, under favorable conditions. This enhanced robustness is in sharp contrast with the naive divide and fusion method where the breakdown point may be reduced by several orders. We specialize the DRL framework for two concrete cases: distributed robust PCA and distributed robust regression. We demonstrate the efficiency and the robustness advantages of DRL through comprehensive simulations.	
18:00	1:00	19:00	**Reception + Posters**	
Day 2, Wednesday, 6-May-2015				
9:00	2:40	11:40	**Conversational Understanding**	
	Moshe Wasserblat		**Conversational Speech Understanding - Capstone overview**	
	Ido Dagan		**Natural Language Knowledge Graphs** How can we capture the knowledge expressed in large amounts of text? Common knowledge representation paradigms encode knowledge in a formal language, whose vocabulary must be pre-specified and hence is inherently limited in scope. In this talk I will outline a research direction that aims to encode textual knowledge based on the available natural language vocabulary and structure, rather than requiring a Sisyphean invention of an artificial language that tries to mimic natural language expressiveness. First, we propose identifying the set of individual propositions expressed in sentences, and representing them in a canonical language-based structure. Then, we propose consolidating these propositions and inducing a global structure over them based on relevant semantic relationships. In particular, we focus first on identifying equivalent propositions and how some pieces of information elaborate over others. I will review research activities along the abovementioned goals, and provide some detail about an approach for learning targeted inference relations between words within large structured knowledge resources.	
	Ronen Feldman		**Unsupervised Extraction of Relations and Events**	

			In the talk I will describe a framework for relation learning and building of domain-specific relation extraction systems. I will demonstrate several applications of the framework in the domains of mining public medical forums and financial news and blogs. The case studies demonstrate the ability of the system to achieve high accuracy of relation identification and extraction with minimal human supervision.
	Yoav Goldberg		**Better Syntactic Parsing with Lexical-Semantic Features from Auto-parsed Data** Syn An Unsupervised Framework for Information Extractionng the meaning of a sentence. While most of the sentence structure can be inferred based purely on non-lexical information, some cases do require semantic bilexical information (relation between specific words) in order to be disambiguated correctly. In this work, we present a method to acquire such information automatically from large quantities of automatically parsed text. The method is inspired by recent advances in bilinear models of word embeddings. In contrast to previous methods, we are the first to show a significant improvement over the use of brown-clusters, the previous state-of-the-art method for semi-supervised learning for parsing.
	Roi Reichart		**Hybrid Models for Minimally Supervised Natural Language Learning With Applications to Conversation Processing** A large number of Natural Language Processing applications, including syntactic parsing, information extraction and discourse analysis, involve the prediction of a linguistic structure. It is often challenging for standard feature-based machine learning algorithms to perform well on these tasks due to modeling and computational reasons. Moreover, creating the large amounts of manually annotated data required to train supervised models for such applications is usually labor intensive and error prone. In this talk we describe a serious of works that integrate feature based methods with declarative task and domain knowledge. We address a wide variety of NLP tasks, various levels of supervision and different types of domain knowledge. We start from models trained on small amounts of manually labeled sentences that further employ manually specified declarative knowledge. We then continue to models where no manually annotated sentences are used but declarative knowledge is still specified manually. Finally, we consider unsupervised models that use no human supervision and instead induce declarative knowledge from unlabeled data. We show how these models can be applied to a variety of NLP tasks and applications including syntactic parsing and POS tagging, information extraction and discourse analysis. Our models are implemented within two different graphical model frameworks which take different approaches to the solution of the global hard optimization problem resulted from the integration of feature based models and global constraints. In experiments, these models improve over state-of-the art supervised models that require more human supervision for training. We finally discuss the potential of our framework for information extraction and summarization of human conversations, a research direction we intend to pursue under the Intel ICRI-CI-2015 research program. We discuss the unique challenges of this domain and the appropriateness of hybrid models for its modeling.
11:40	1:20	13:00	**Visual-II**
	Ronny Ronen		**Visual & Imaging - overview**
	Daphna Weinshall		**Mental phenotyping with 3D cameras** The emerging technology of 3D cameras facilitates our ability to automatically track bodily movements and facial expressions. Taking advantage of this opportunity, we are developing algorithms and learning techniques which are able to provide objective diagnostic measures of mental and neurological states. Thus for the phenotyping of Parkinson's disease, we are able to assess the severity of dyskinesia (a pathology of bodily motion) and the condition of "mask face" (a pathology of facial expressions) in a manner well correlated with trained neurologists' diagnosis. Most of our efforts, though, are focused on the mental phenotyping of Schizophrenia, involving the quantitative characterization of non-verbal behavior in schizophrenic patients. We have collected a large database of 3D videos of patients and controls participating in a battery of tasks, from interactive interviews to passive observations. Our first task is to automatically extract from these 3D videos measures which correlate with facial expressions. Based on these measures, we train classifiers to predict a variety of mental scores which are used to evaluate and diagnose Schizophrenia; for many of these measures, our predicted scores correlate well with scores given by trained psychiatrists.
	Shmuel Peleg		**Blind Video: Video Without Photographers** In ordinary videography a photographer selects the area of interest, and presses the record button at the time of interest. Surveillance cameras and wearable cameras are turned on and record continuous videos without a photographer aiming the camera or pressing the record button. The resulting video is long and

			unstructured, and mostly boring. Most such video may therefore never be watched. In this talk we will describe some approaches to make such video accessible to human viewers.	
	Yair Weiss		**Learning Optical Flow** In recent years "pure learning" methods that have very little built-in knowledge about images have achieved state-of-the-art performance in image restoration tasks (e.g. denoising, deblurring) and outperformed many hand designed approaches. I will present work-in-progress that asks whether a similar approach can also work for optical flow. Our preliminary results suggest that "pure learning" methods yield high quality energy functions (arguably better than handcrafted ones) but optimizing these energy functions still requires image-specific knowledge. Joint work with Dan Rosenbaum	
13:00			**Adjourn + Lunch**	