Project: A secure cloud future

PI: Jon Crowcroft (FF, 30%)

Investigators: Efi Tsamoura (RF, 50%), Brad Karp (UCL, 20%), Peter Pietzuch (Imperial, 20%), George Danezis (FF, 20%), Adria Gascon (RF, 50%), Dave Robertson (Edinburgh).

Duration: March 2017 – Feb 2019

Context

Much of today's data is stored and processed in the cloud. Cloud computing entails many, often vast data centres in heavily protected environments the size of warehouses. The data scientist would like to rent these scalable computing resources. However, the cloud may not always meet the stringent technical, legal and regulatory requirements, especially for data that is personal and subject to privacy laws, or that is commercial-in-confidence. Data “safe havens” provide strict isolation between different tenants (customers) of a data centre but this may not be sufficient for some uses -- patient records, criminal record data, financial service audit data, where there is still a residual risk from insider attacks or from software (or hardware) vulnerabilities. The next step then is to take advantage of new hardware technology for trusted computing from Intel and from ARM, who have built hardware “enclaves” to protect sensitive data. This project investigates and prototypes the secure software components that are needed to take advantage of the new hardware.

Milestones

6 months: Enclave code with logically verified security guarantees.

12 months: Build the Java runtime (JVM) into the enclave system with access as a Java library **12 months**: Electronic Health Record accelerated approvals tool with Scottish health data

18 months: Demonstrate secure isolation of enclave code from host environment

24 months: Extend the Apache Flink big data platform for ML to work with the secure enclave.

24 months: Distributed querying of federated health records with automatic checking for privacy leakage.

Project: Artificial Intelligence for Data Analytics

PI: Chris Williams (ULD, 25%), James Geddes (Turing, 30%)

Investigators: Zoubin Ghahramani (FF, 20%), Ian Horrocks (FF, 20%), Charles Sutton (FF, 10%)

Duration: Nov 2016 – Apr 2020

Context

The process of transforming a raw dataset into useful knowledge is called data analytics. It comprises many different stages and phases. The early stages involve so-called "data wrangling", starting with "messy" data and including understanding what data is available and assessing its quality, identifying and patching missing or anomalous data, integrating data from multiple, heterogeneous sources, etc. There has been little methodological research into data wrangling, even though this is often laborious and time-consuming, and accounts for up to 80% of a typical data science project. The goal is to automate, at least partially, these tasks. The Institute is uniquely positioned to bring together various research areas to tackle these problems, including machine learning and logic processing. Later stages of the data analytics process involving drawing inferences from the data may also be partially automated, using model fitting from statistics, combined with machine learning and AI to build good interfaces for human analysts. This is fundamental work with wide- ranging applications, including analysing government data and making it easily accessible to non-experts; business and financial analysis systems; and decision support systems, e.g. in healthcare.

Initial "starter" funding from LRF has been made available supporting about 1 FTE for 6 months (Nov 2016- Apr 2017), but seriously increased resource is needed in order to fulfill the ambitious agenda (see below).

Milestones

6 months: Assemble and curate exemplar datasets.

18 months: Build set of tools to address issues in data transformation, data understanding and data cleaning, combining logic and ML approaches.

30 months: Integration of tools with an interactive assistant that will step the analyst through all the issues in the current dataset.

42 months: Demonstrate full pipeline of data preparation through to automatic report writing.

Project: Fairness, Privacy & Transparency

Lead PI: Adrian Weller (FF, 30%),

Investigators: Brent Mittelstadt (RF, 50%), Josh Loftus (RF, 20%), Matt Kusner (RF, 30%), Ricardo Silva (FF, 20%), Graham Cormode (FF), Dong Nguyen (RF, 20%), David Pym (FF, 20%)

Duration: March 2017 – Feb 2019

Context

There are increasingly many examples where data collection and analysis risks oversharing personal information, enshrining biased decision making into code, and giving unwelcome decisions without explanation or recourse. A large part of these problems require a technical approach to develop:

- Robust privacy-preserving data analysis techniques that protect individuals' information
- Mathematically provable methods which ensure that appropriate non-discriminatory measures are taken to protect characteristics such as age, gender, sexuality, religion and disability status. These characteristics should not be used inappropriately as the basis of decisions, nor should attributes which are sufficiently correlated with these.
- Human-interpretable explanations of predictions or decisions emerging from machine learning systems which justify choices while still enabling high levels of accuracy

Milestones

6 Months: Initial proposals for algorithms and system design that achieve both fairness and transparency in a mathematically precise way

12 Months: Demonstrator prototypes showing how high performance machine learning techniques can be augmented to achieve a trade-off between fairness, transparency and privacy-preservation

18 Months: Application to project with partner, based on suitable data provided (example: collaboration with HSBC to show fairness in loan decisions, and transparency of reasons for rejection)

24 Months: Deliver open-source code implementation of FPT system, and accompanying journal paper describing application

Project: Computer Architecture for Data Science

PI: Peter Boyle (FF, 40%), Anthony Lee (SPD, 30%) **Investigators**: Kenneth Heafield (FF, 40%)

Duration: April 2016 – March 2018

Context

Recent advances in hardware and co-processors have had a striking effect on the efficiency of computer code for machine learning. In particular graphics coprocessors, first repurposed for machine learning have now also been re-engineered. Several companies have also built coprocessors specially to accelerate machine learning and optimised code to match. It is a compelling proposition to look radically at computer architecture, especially for data-centres, to design it specifically to accelerate codes for machine learning and data analytics. This is the focus of the Turing software-hardware co-design programme with Intel, which could have a wide-reaching effect on data analytics in the cloud. The developing architectures could be tested against a variety of workloads, such as:

- healthcare databases
- streaming analytics from urban sensor systems
- trading flows analysis

Milestones

18 months: Identification of bottlenecks in key ML software, emphasis on deep learning.

18 months: Single-node benchmarking, including FPGA-assisted computation, floating point precision studies, and application-driven algorithm parameters for convolutional neural networks and state-of-the-art recurrent neural networks used in machine translation.

18 months: Multi-node benchmarking, including distributed reductions and sparse communication.

24 months: Architectural remedies to bottlenecks proposed, both in processor architecture and in memory and interconnect design

Project: Rough paths and signatures for machine learning

PI: Terry Lyons

(FF, 40%).

Investigators: Hao

Ni (FF 40%).

Duration: Oct 2016 – Feb 2019

Context

We believe there are massive opportunities for mathematical innovation in the area of Human Computer Interfaces and that, at ATI, we are achieving state of the art results in recognising actions and expressions, particularly from movies, computer input devices etc. Although the mathematical insights also offer wider approaches to data capture, e.g. in astronomy, In this proposal we keep the focus on the capture and classification of human messages and on the supporting mathematical questions around computation, representation, scalability, and algorithms.

We see this applications of this work across a number of areas; but one in particular is mental health supporting the ATI initiative. Another might be to work safety.

Milestones

6 months: Demonstrate state of the art recognition rates on action recognition.

18 months: Create a new python package/wrapper for the signature feature set that supports the recognition of actions in movies and can be easily integrated into the lstm and other deep learning techniques. Appoint and train postdocs. Integrate them with the activity of existing ATI students and postdocs around handwriting and health. Build industry collaborations and data.

30 months: Push the scalability and dimensionality to the limits, and create a new state of the art in action recognition in collaboration with Schmid, Lianwen and his co-workers

Project: Low-dimensional structure in data

PI: Jared Tanner (ULD, 30%)

Investigators: Mihai Cucuringu (RF, 35%), Armin Eftekhari (RF, 45%), and

Hemant Tyagi (RF, 45%). **Duration**: January 2017 – December 2018

Context

Modern data sets are often well modelled as samples from a low-dimensional subspace which captures the correlation and predictive power of the data. However, data sets typically also contain outliers and are incompatible with missing values. Computationally tractable algorithms have recently been proposed which automatically decompose a data matrix into the sum of a low-rank matrix consistent with the majority of the data, and a set of outliers inconsistent with the low rank model. This research programme builds on this success by developing: scalable algorithms for low-rank plus sparse decompositions, manifold learning and low-rank partitioned models, continuous models with scalable algorithms, nonlinear dimensionality reduction, and applications to data-centric engineering.

Milestones

6 months: Develop restricted isometry constants for the low rank plus sparse model.

12 months:Design algorithms for supervised and unsupervised learning of column and/or row partitioning for matrices to model local manifold structure. Incorporate the sparse plus low rank decompositions into the Diffusion Maps framework, in order to detect anomalies and increase robustness to noise

18 months: Development of time-varying grid free model and its analysis using convex relaxations.

24 months: Apply the developed algorithms and models to real world data sets as they are completed and adapt the proposed models to better reflect the particularities of the data.

Project: Electric Vehicle Charging Networks

PI: Richard Gibbens (FF, 20%)

Investigators: L'ubos Buzna (UoZilina), Rui Carvalho (Durham), Frank Kelly

(Cambridge) **Duration**: Jan 2017 to Dec 2018

Context

Extensive use of electric vehicles (EVs) is considered a likely possibility to help reduce the environmental impact of fossil fuels and to make the transition to a low carbon transport economy. In this project we intend to develop a decentralised scheduling approach that would allow a central authority to oversee the network and resource usage but allow the vehicle owners to represent their individual preferences over charging levels and costs.

There are important mathematical connections to decentralised congestion control for the internet that has proved to be highly effective at achieving very robust and resilient network operation—characteristics that are of critical importance in electric distribution networks.

Milestones

Jun-2017: Develop essential mathematical optimisation

model **Dec-2017**: Develop core simulator platform

Jun-2018: Extend core simulator with a web-based front-end

Dec-2018: Develop demonstrator case studies to assist with technology translation

Project: Transport Network Resilience

PI: Scott Hale (FF)

Investigators: Jonathan Bright (Oxford), Graham McNeill (Oxford) **Duration**: Jan 2017 to Dec 2018

Context

This project aims to address challenges in road network resilience, developing understanding of traffic networks with impact on network theory and urban policy. It is based on an ongoing partnership with Oxford County Council, and will make use of several stores of so-called big data to which the Council has access. In particular, it will use anonymised data provided through Google's Better Cities programme, data from the Waze community traffic app, and data from the Zipabout journey planning tool. These data sources enable us to work at high spatial and temporal resolution and to overcome many of the problems associated with estimating traffic flows from lower-resolution/sparse data.

Milestones

Jun-2017: Algorithmic review, development of analysis pipeline for processing data at scale

Dec-2017: Creation of indicators for traffic anomalies; models disruptions & role of network structure **Jun-2018:** Analyse impact of shocks on real multilayer networks compared with standard metrics

Dec-2018: Research on rerouting over multilayer networks; publications; final conference/workshop

Project: Resilient Analytics for Disaster-AffectedCommunities

PI: Steve Roberts

(FF) **Investigators**:

Steve Reece (Oxford)

Duration: Jan 2017 to Dec 2018

Context

Lack of access to information and technology has a major bearing on people's ability to prepare for, survive and recover from disasters, according to the 2013 World Disasters Report. During the first critical hours after an emergency, most lives saved are actually by local people. However, following a disaster their world has changed significantly from what they knew, and probably continues to do so as disaster cascades unfold.

Rapid situation awareness is key to both survivability and recovery in the hours following the disaster. A resilient engineered network infrastructure is crucial, so that those affected can access information resources (e.g. SafetyCheck, Thrivespring, DigiDoc), social media generated maps, access to national disaster management agency (NDMA) evacuation procedures and aid centres. We intend to develop disaster-resilient data aggregation and network architectures based on movable and deployable resource units, such as the Cobham Explorer Mobile Net, that remotely pull mobile/smart phones onto a network in areas where mobile phone availability has collapsed.

Milestones

Jun-2017: Heterogeneous data fusion

Dec-2017: Aggregation and knowledge discovery

Jun-2018: Sparcity and structure for scalability; conference and journal papers

Dec-2018: Decentralised and parallelised computing; suite of machine learning algorithms; impact report

Project: Data-driven economics

PI: TBA (SPD, 50%)

Investigators:

Duration: Jan 2017 – Dec 2019

Context

The HSBC-Turing research programme in economic data science aims to stimulate the development of advanced research by economists and data scientists into new methods for harnessing and analysing financial data with ever greater detail and accuracy. The desired outcome to help economists, researchers, policymakers and businesses to better understand the UK economy, its regional constituents and their interconnectivity with global markets at a time of intense public, corporate and institutional interest.

Milestones

3 months: Data access paths tested and linked to Edinburgh secure

data centre **4 months**: SPD hired

6 months: SPD in post

Further milestones to be set by incoming SPD