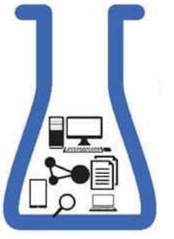


Buzzwords surrounding Data Science

Eiko Yoneki
eiko.yoneki@cl.cam.ac.uk
http://www.cl.cam.ac.uk/~ey204

Systems Research Group
University of Cambridge Dept. Computer Science and Technology
Computer Laboratory

Outline

- Data Science Community
- Landscape of Research in Alan Turing Institute
  - Probabilistic Programming
  - Optimisation
  - Ethics
  - Frameworks
- Becoming Data Scientist?
- Fill the Gap between Research and Practice

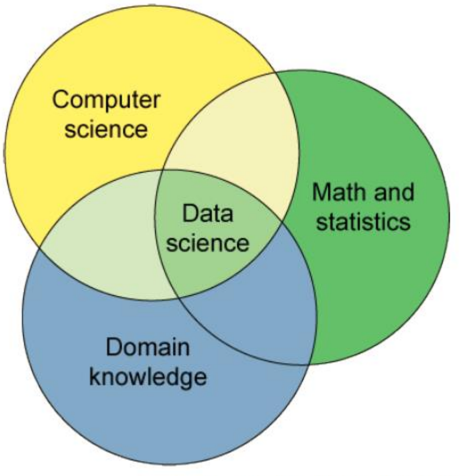
Data Science: Any new intellectual content?

- What does it mean to Computer Science?

- 1970's: EE + Math → Computer Science
- 2010's: CS + Stats + ?? → Data Science

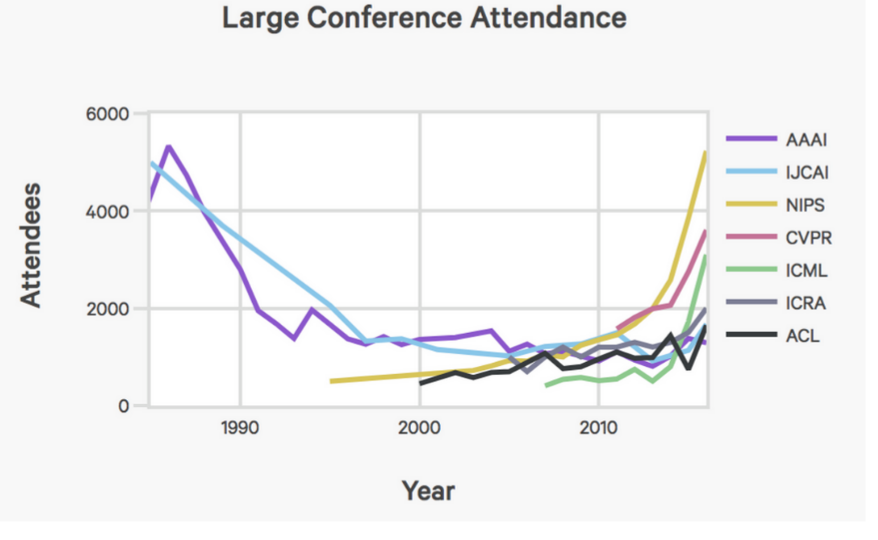
- Is something fundamental emerging here?

- Data Science is a very broad discipline
- Data Science PhD?
  - PhD normally with a narrow field with depth…



based on Drew Conway, NYU

Scale of Community Size in ML/AI



NIPS: 8000 Attendees in 2017

- Randomness of Paper acceptance?
- In 2016, 2,406 papers were submitted and 568 were accepted for a 24% acceptance rate. In 2017, 679 papers out of 3,240 submitted were accepted for a 21% acceptance rate.
- In 2014, Corinna Cortes and Neil Lawrence ran the NIPS experiment where 1/10th of papers submitted to NIPS went through the NIPS review process twice, and then the accept/reject decision was compared.
- The 26% disagreement: Given 22% acceptance rate, The immediate implication is that between 1/2 and 2/3 of papers accepted at NIPS would have been rejected if reviewed a second time.
- http://blog.mrtz.org/2014/12/15/the-nips-experiment.html
- In particular, about 57% of the papers accepted by the first committee were rejected by the second one and vice versa. In other words, most papers at NIPS would be rejected if one reran the conference review process (with a 95% confidence interval of 40-75%).

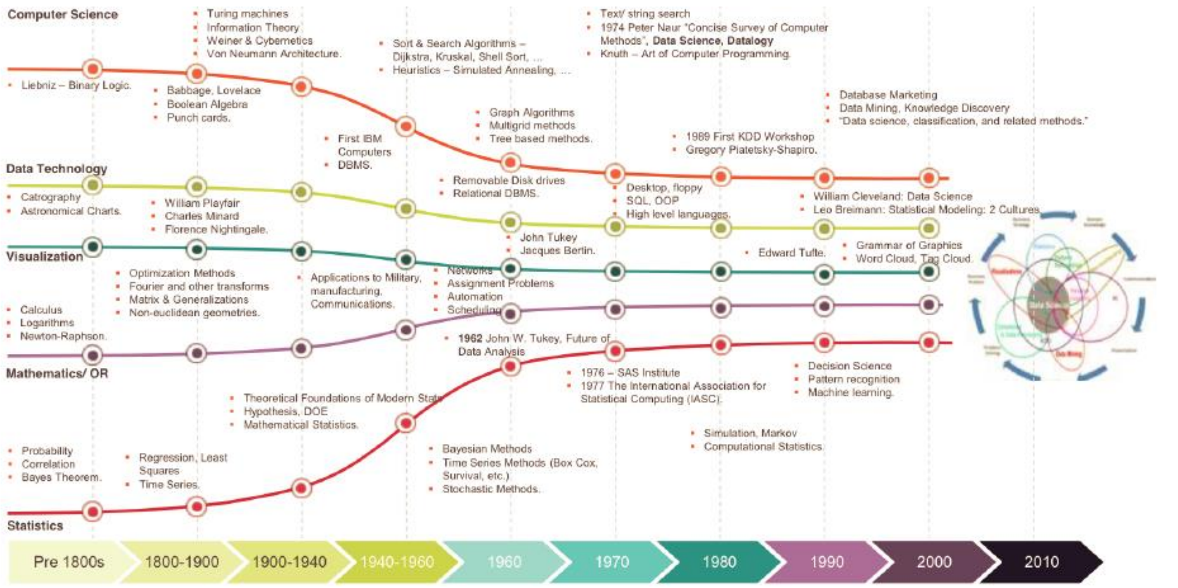
SysML Conference spawn in 2018-2019

- SysML is a conference targeting research at the intersection of systems and machine learning
- Aims to elicit new connections amongst these fields, including identifying best practices and design principles for learning systems, as well as developing novel learning methods and theory tailored to practical machine learning workflows

Steering Committee

Jennifer Chayes
Bill Dally
Jeff Dean
Michael I. Jordan
Yann LeCun
Fei-Fei Li
Alex Smola
Dawn Song
Eric Xing

Trajectory



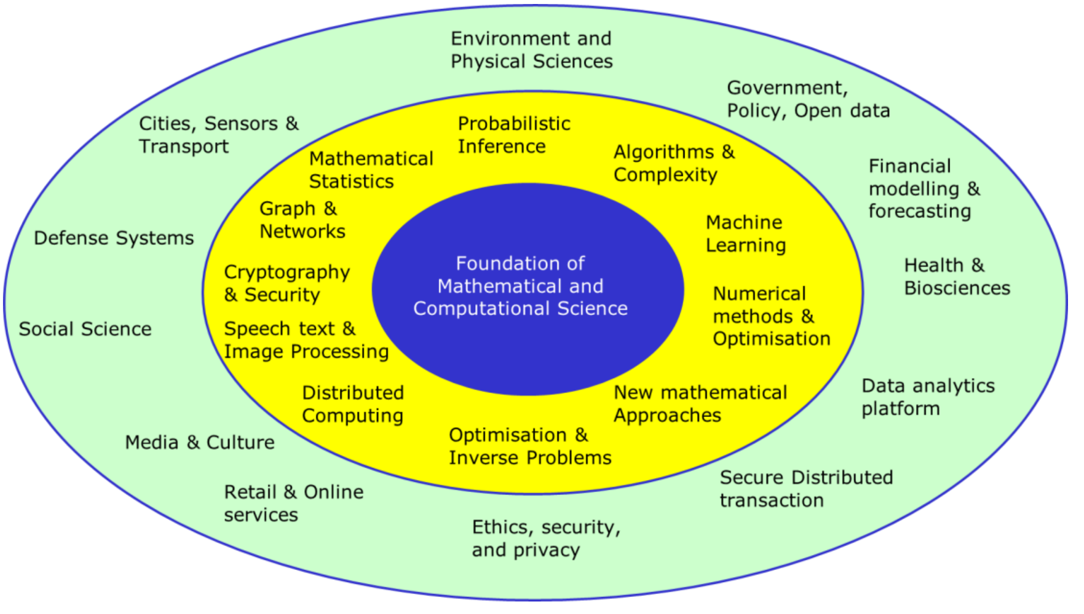
UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Alan Turing Institute (ATI)

- Established in 2015 in London as a National Institute for Data Science
- >£20M Capital Investment from Government
- Originally 5 Universities formed core body (UCL, Warwick, Edinburg, Oxford and Cambridge) and now expanded to 13 universities
- Goal: Data Science and after 2018 Artificial Intelligence
- Translating output into practice

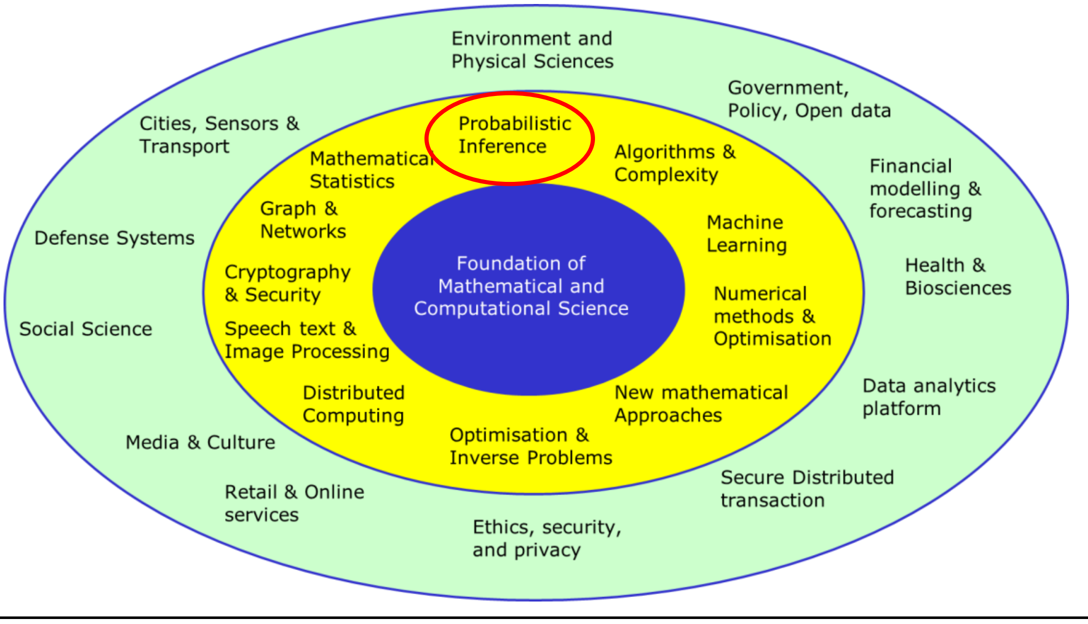
UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Broad Landscape of Research in ATI in 2015



UNIVERSITY OF CAMBRIDGE
Computer Laboratory

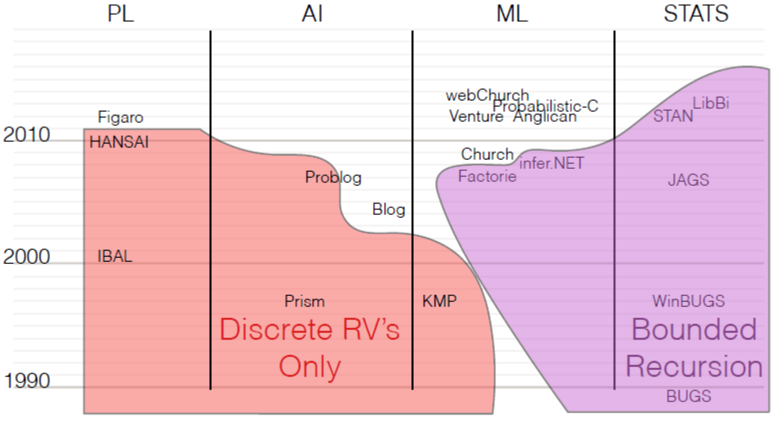
Broad Landscape of Research in ATI in 2015



Probabilistic Model

- Probabilistic models incorporate random variables and probability distributions into the model
  - Deterministic model gives a single possible outcome
  - Probabilistic model gives a probability distribution
- Used for various probabilistic logic inference (e.g. MCMC-based inference, Bayesian inference…)

Probabilistic Programming



Edward based on Python

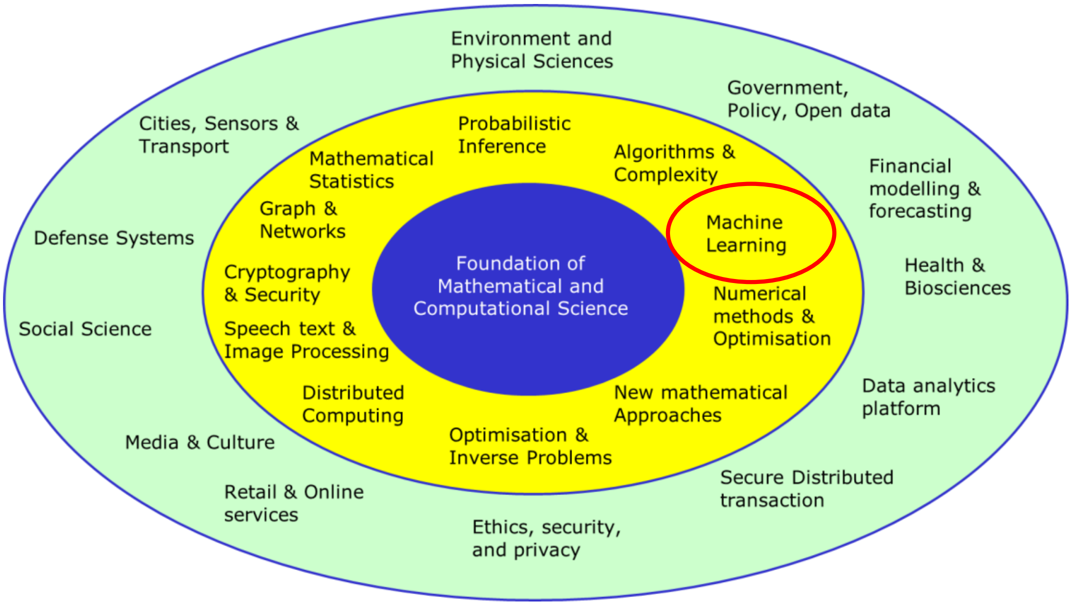
Probabilistic C++

Improbable – Java version

Simula

Prolog

Broad Landscape of Research in ATI in 2015

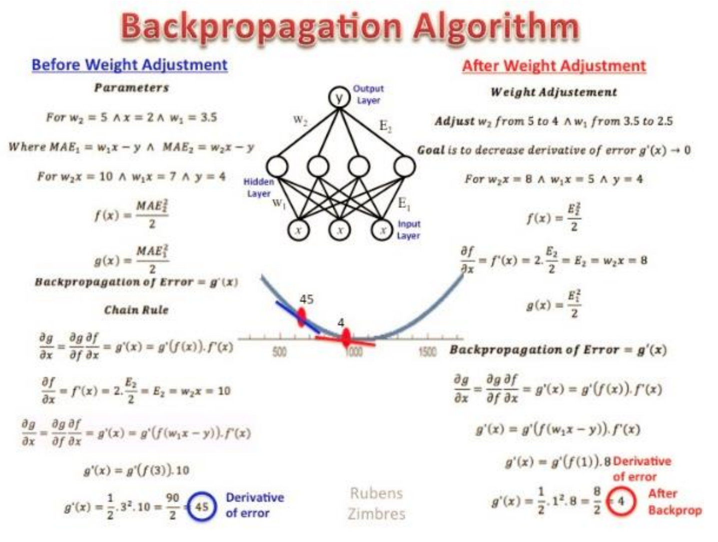


4 Great Pictures Illustrating Machine Learning Concepts

- Neural Networks: The Backpropagation algorithm
- Cheat Sheet on Probability
- 24 Neural Network Adjustments
- Matrix Multiplication in Neural Networks

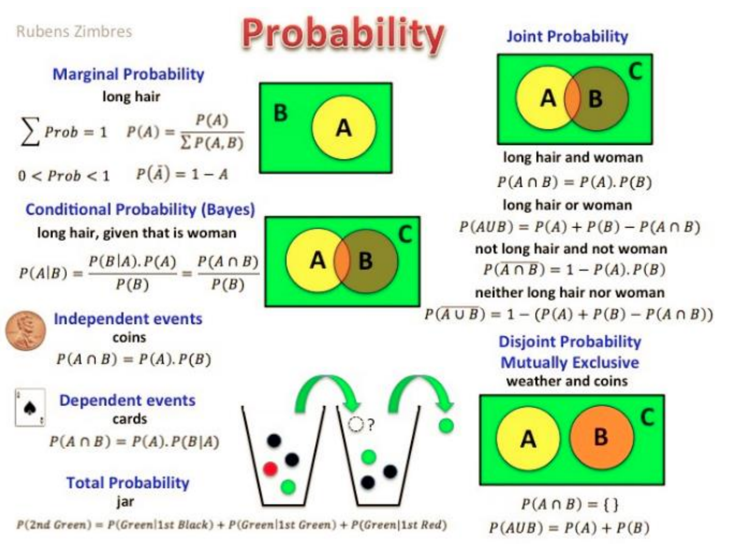
See https://www.datasciencecentral.com.

Neural Networks: The Backpropagation algorithm



UNIVERSITY OF CAMBRIDGE Computer Laboratory

Cheat Sheet on Probability



24 Neural Network Adjustments

ARCHITECTURE

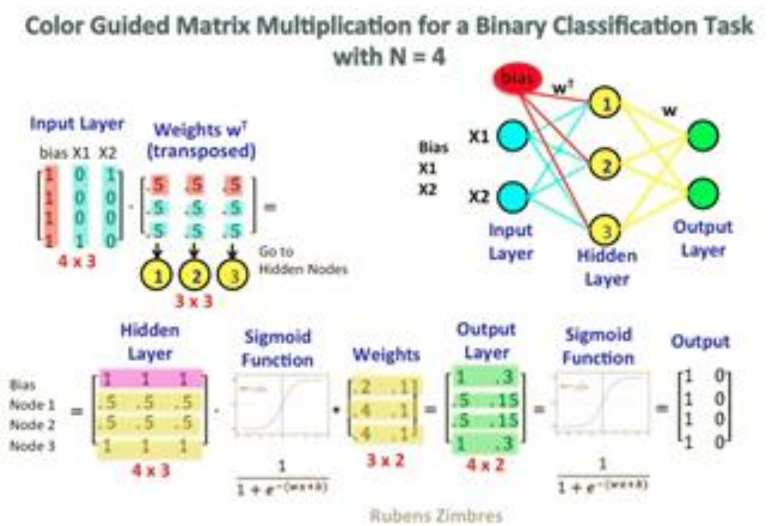
- Variables type
- Variable scaling
- Cost function
- Neural Network type:
  - RBM,FFN,CNN,RNN…
- Number of layers
- Number of hidden Layers
- Number of nodes
- Type of layers:
  - LSTM, Dense, Highway
  - Convolutional, Pooling…
- Type of weight initialization
- Type of activation function
  - Linear, sigmoid, relu…
- Dropout rate (or not)
- Threshold

HYPERPARAMETER TUNING

- Type of optimizer
- Learning rate (fixed or not)
- Regularization rate (or not)
- Regularization type: L1, L2, ElasticNet
- Type of search for local minima:
  - Gradient descent, simulated
  - annealing, evolutionary…
- Batch size
- Nesterov momentum (or not)
- Decay rate (or not)
- Momentum (fixed or not)
- Type of fitness measurement:
  - MSE, accuracy, MAE, cross-entropy,
  - precision, recall
- Epochs
- Stop criteria

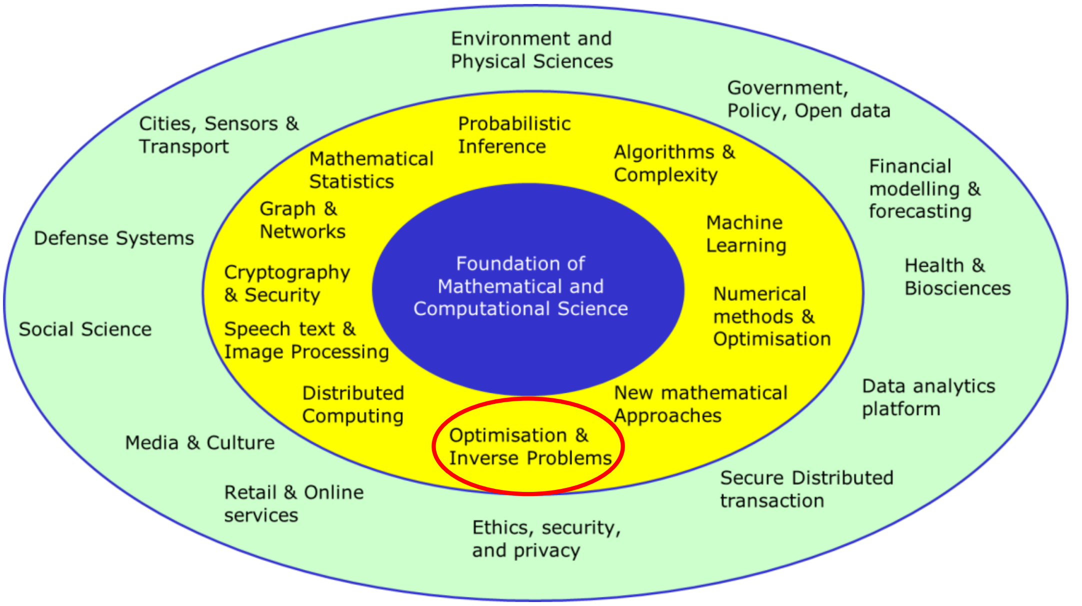
UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Matrix Multiplication in Neural Networks



UNIVERSITY OF CAMBRIDGE Computer Laboratory

Broad Landscape of Research in ATI in 2015

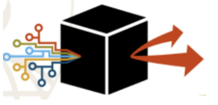


UNIVERSITY OF CAMBRIDGE Computer Laboratory

Tuning Computer Systems

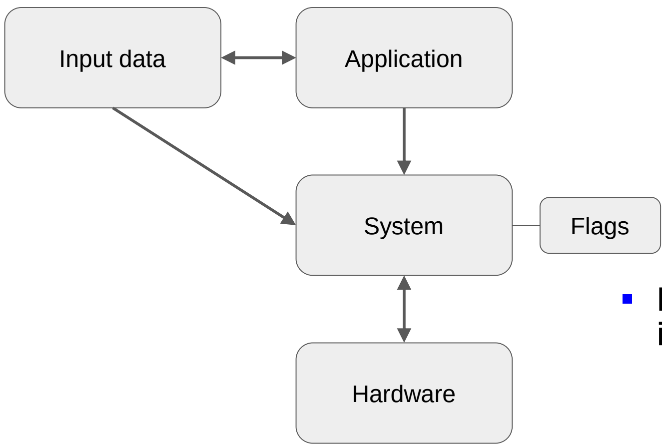
- Complex configuration parameter space and increasing number of parameters
- Configurations need tuning to optimise resource utilisation, minimise cost
- Containerisation means software is frequently restarted and deployed in new contexts
- Hand-crafted solutions impractical, often left static or configured through extensive offline analysis
- Auto-tuning: automatically determine parameters, e.g. resources for Hadoop job

Optimisation of Complex Data Processing in Computer Systems



- Auto-tuning to deal with complex parameter space using machine-learning
  - Structured Bayesian Optimisation, Reinforcement Learning
  - Build a solid auto-tuning platform in a complex and large parameter space
- e.g. Tuning Cluster task scheduling, ML framework, JVM garbage collector, NN model, Compiler, DB indexing…

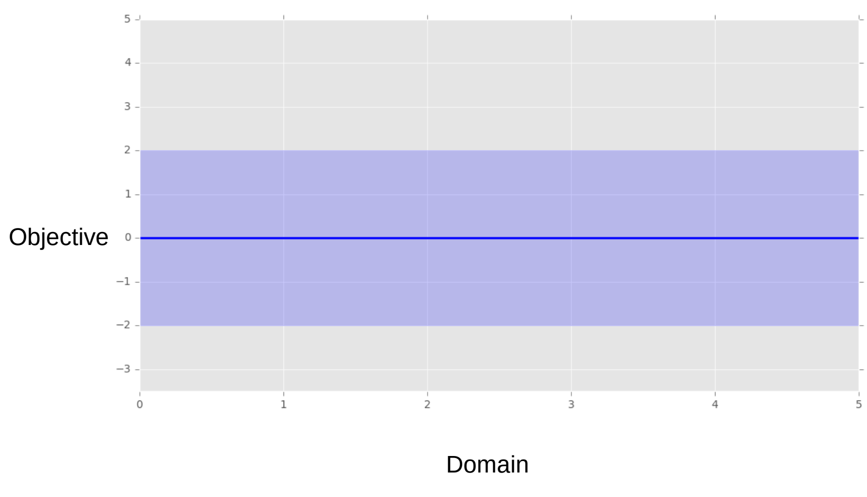
Auto-tuning systems



- Properties:
  - Many dimensions (30+)
  - Expensive objective function
  - Understanding of the underlying behaviour
- Bayesian Optimisation is popular tool

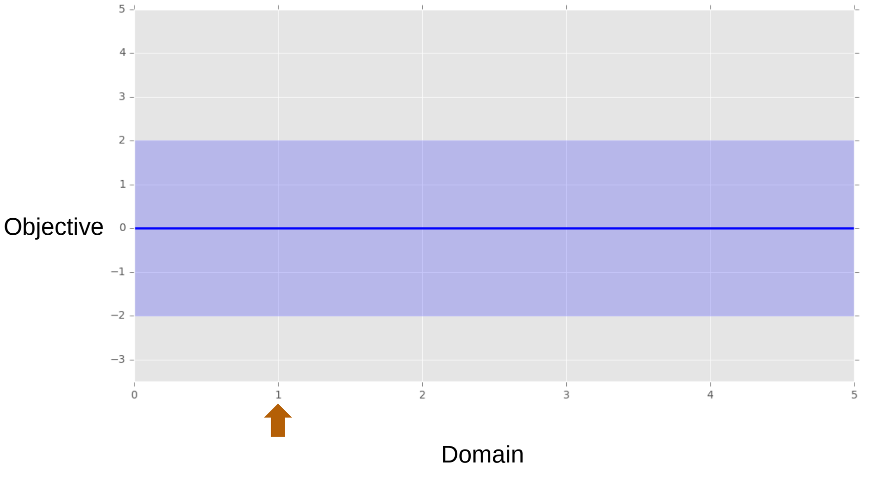
UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Bayesian optimisation



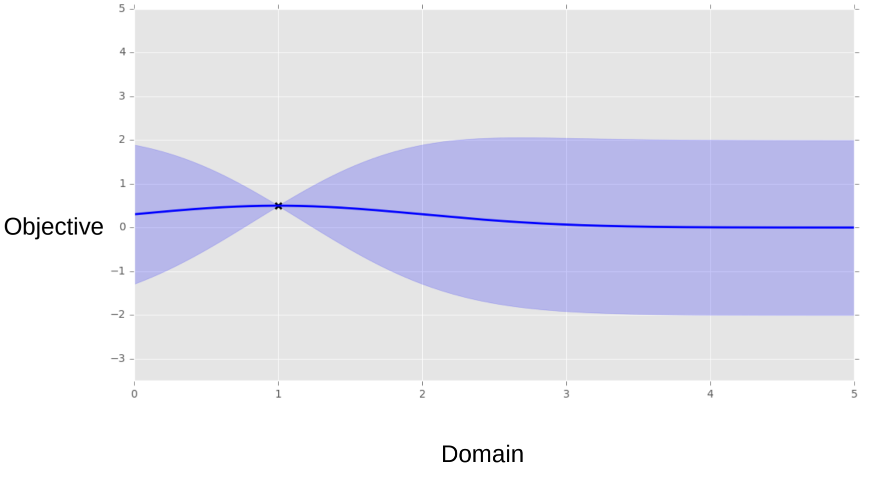
UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Bayesian optimisation

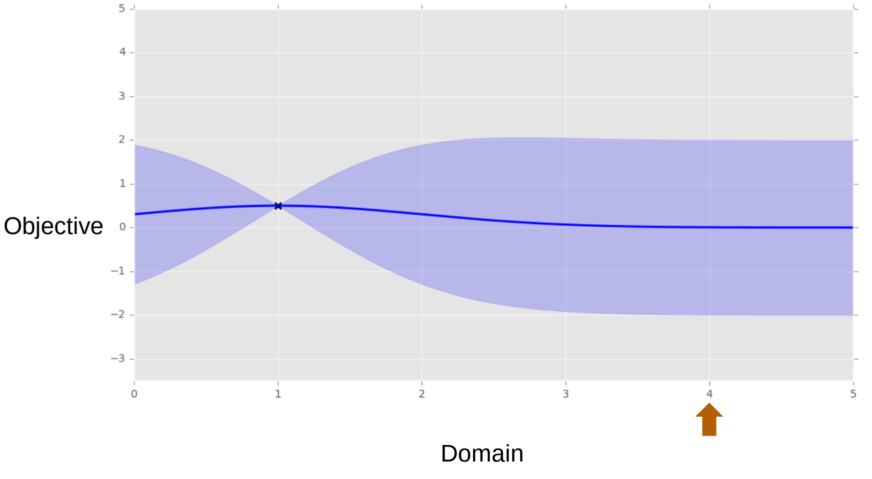


UNIVERSITY OF CAMBRIDGE
Computer Laboratory

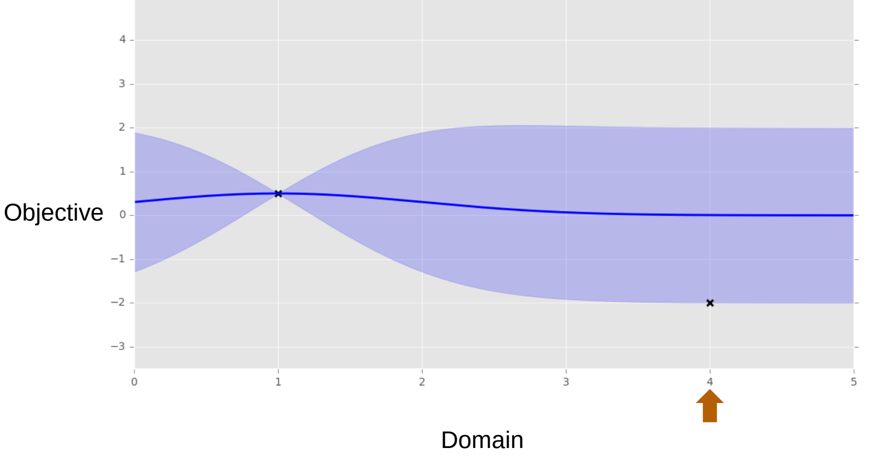
Bayesian optimisation



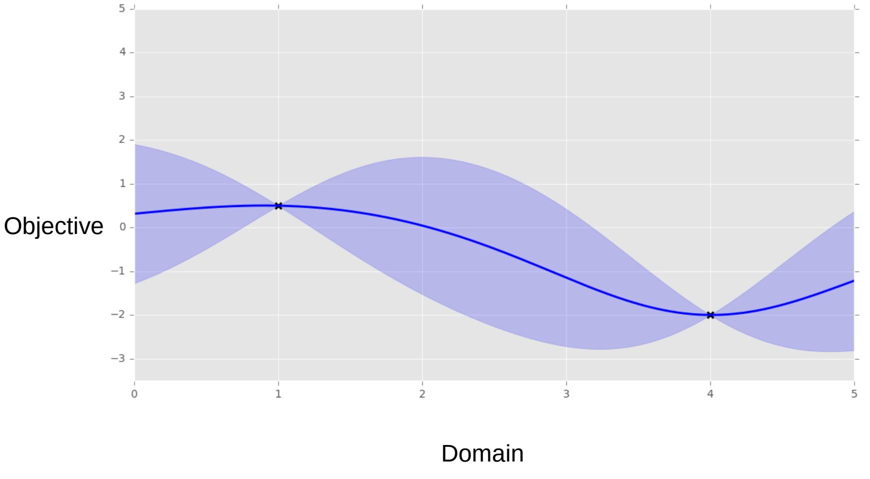
Bayesian optimisation



Bayesian optimisation

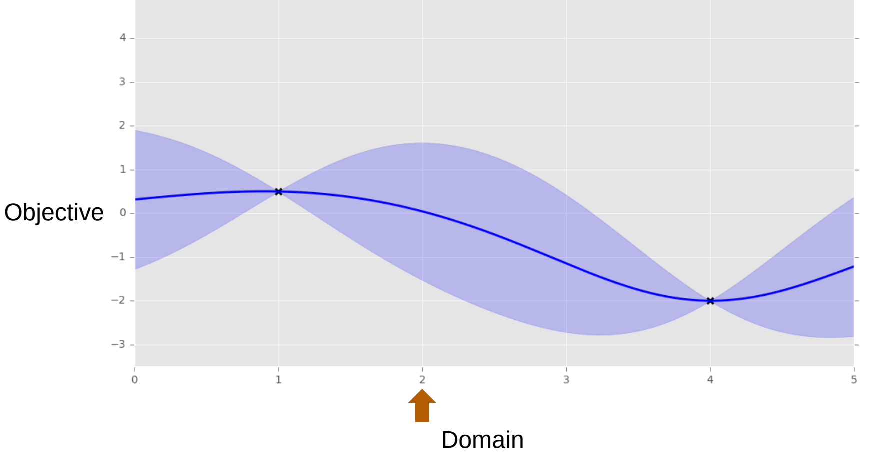


Bayesian optimisation

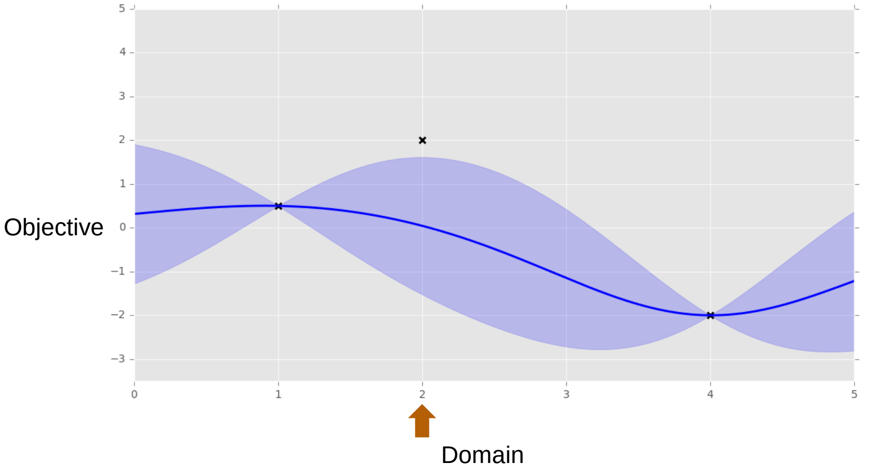


UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Bayesian optimisation

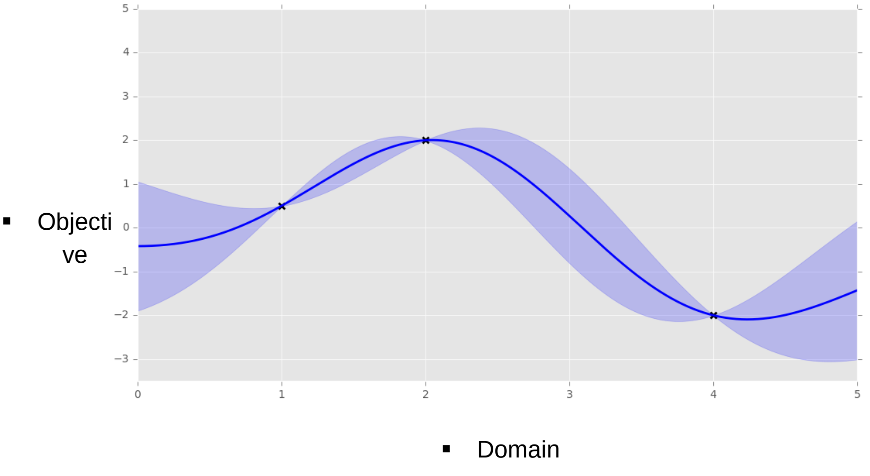


Bayesian optimisation

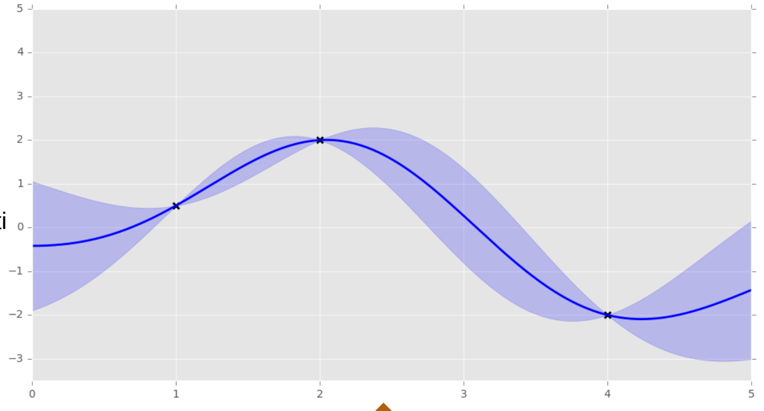


UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Bayesian optimisation



Bayesian optimisation

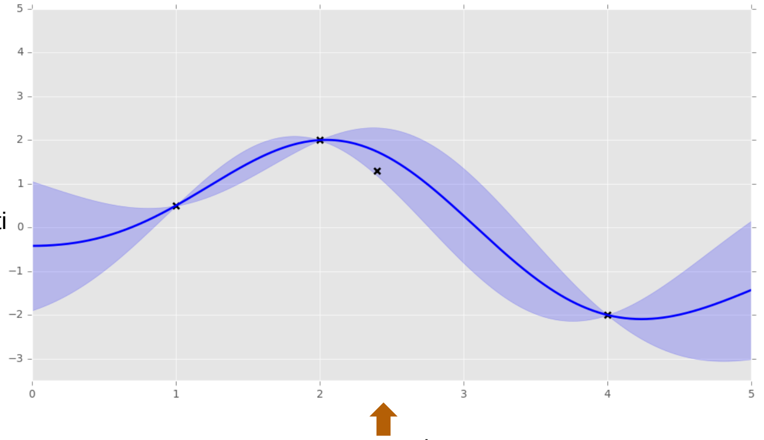


- Objective

- Domain

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

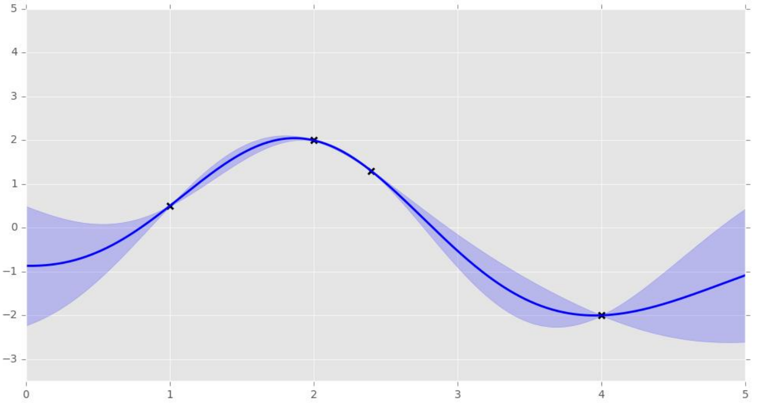
Bayesian optimisation



- Objective

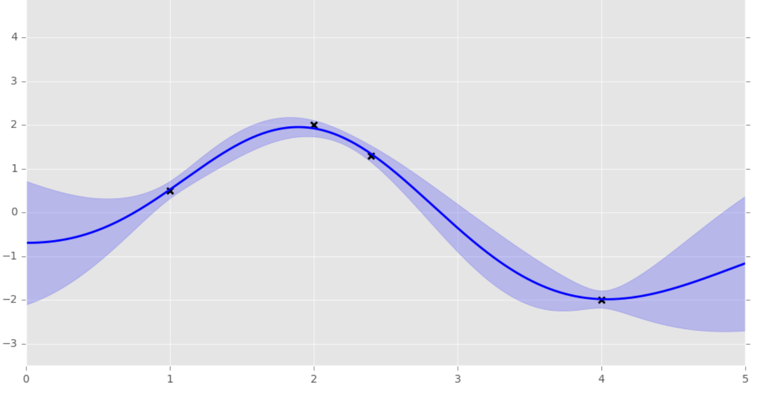
- Domain

Bayesian optimisation



- Objective
- Domain

Bayesian optimisation

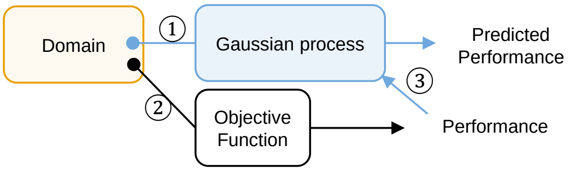


- Objective
- Domain

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Bayesian optimisation

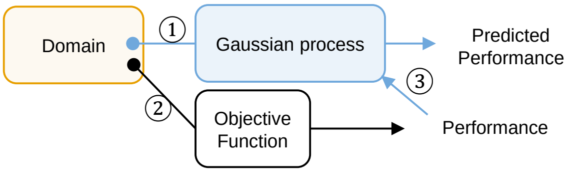
- Iteratively build a model of the objective function



1. Find promising point
2. Evaluate it
3. Perform inference

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Bayesian optimisation



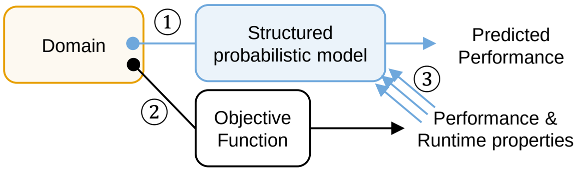
1. Find promising point
2. Evaluate it
3. Perform inference

Pros:
- ✓ Data efficient: converges in few iterations
- ✓ Able to deal with noisy observations

Cons:
- ✗ In many dimensions, model does not converge to the objective function

Solution: Use the known structure of the optimisation problem

Structured Bayesian optimisation



Three desirable properties:
- Able to use many measurements
- Understand the trend of the objective function
- High precision in the region of the optimum

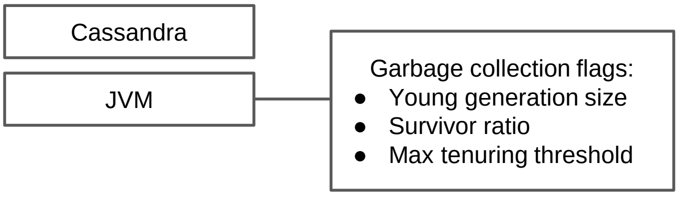
✓ Better convergence
✓ Use all measurements

- **BOAT:** a framework to build **B**esp**O**ke **A**uto-**T**uners
- It includes a probabilistic library to express these models

- V. Dalibard, M. Schaarschmidt, and E. Yoneki: BOAT: Building Auto-Tuners with Structured Bayesian Optimization, WWW 2017. (Morning Paper on May 18, 2017)

Example:

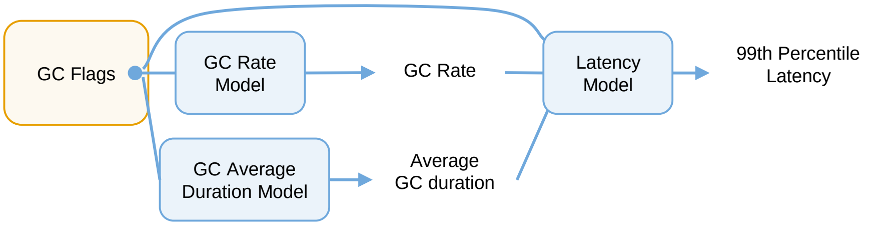
- Cassandra's garbage collection



- Minimise 99th percentile latency of Cassandra

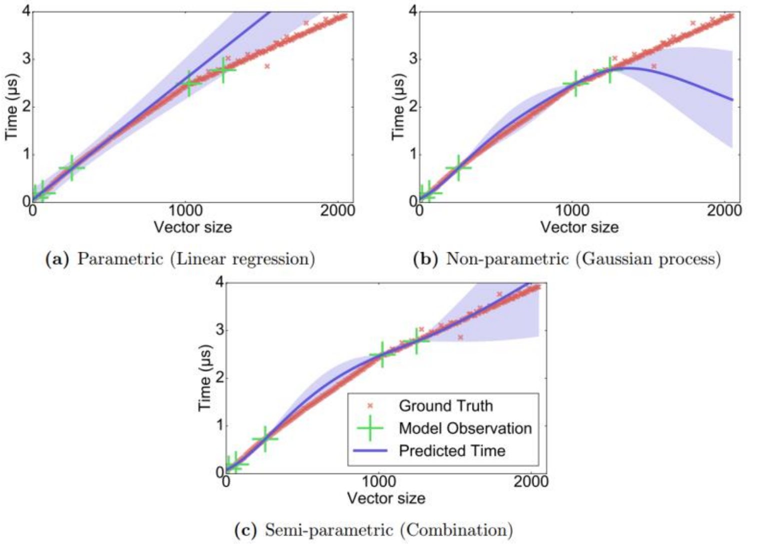
Using many measurements

- Define a directed acyclic graph (DAG) of models



Tune three JVM parameters of a database (Cassandra) to minimise latency

Semi-parametric models



(a) Parametric (Linear regression)

(b) Non-parametric (Gaussian process)

(c) Semi-parametric (Combination)

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

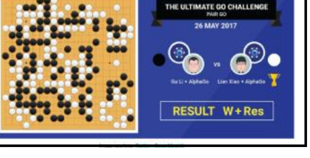
Computer Systems Optimisation Models

- Long-term planning: requires model of how actions affect future states. Only a few system optimisations fall into this category, e.g. network routing optimisation.
- Short-term dynamic control: major system components are under dynamic load, such as resource allocation and stream processing, where the future load is not statistically dependent on the current load. Bayesian optimisation is sufficient to optimise distinct workloads. For dynamic workload, Reinforcement Learning would perform better.
- Combinatorial optimisation: a set of options must be selected from a large set under potential rules of combination. For this situation, one can either learn online if the task is cheap via random sampling, or via RL and pre-training if the task is expensive, or massively parallel online training given sufficient resources.

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Towards Reinforcement Learning

- Given a set of actions with some unknown reward distributions, maximise the cumulative reward by taking the actions sequentially, one action at each time step and obtaining a reward immediately.
- To find the optimal action, one needs to explore all the actions but not too much. At the same time, one needs to exploit the best action found so-far by exploring.
- What makes reinforcement learning different from other machine learning paradigms?
  - There is no supervisor, only a reward signal
  - Feedback is delayed, not instantaneous
  - Time really matters (sequential)
  - Agent’s actions affect the subsequent data it receives



AlphaGo defeating the Go World Champion

UNIVERSITY OF CAMBRIDGE Computer Laboratory

Practical Issues: Scalable action spaces

Many systems problems are combinatorial in nature:

- Mapping tasks to resources
- Selecting a subset of different configuration options
- Large number of options, e.g. thousands of devices in a cluster

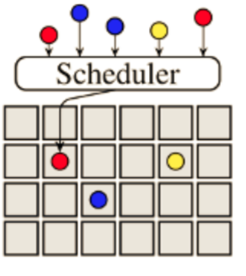
'Standard RL' cannot deal with this – training data requirements grow exponentially

UNIVERSITY OF CAMBRIDGE Computer Laboratory

Problem: Controlling dynamic behaviour

Assume workload dynamic, e.g. seasonality, load spikes, shared resources, failures..

- Algorithm: workload → behavior **distribution**
- Involves approximations to NP-complete problems, e.g. bin packing, sub-graph isomorphism, ..



Source: firmament.io

UNIVERSITY OF CAMBRIDGE Computer Laboratory

Trade-offs in iterative optimisation

Grid search: Discretised sweep

Random search: No risk of 'getting stuck', potentially many samples required

Evolution strategies: Evaluate permutations against fitness function

Bayes Opt: Sample efficient, requires continuous function, some configuration

More requirements

Computation more expensive

Fewer samples

UNIVERSITY OF CAMBRIDGE Computer Laboratory

Trade-offs in dynamic control

Single static configuration/rule: E.g. FIFO scheduler

Online estimate of distributions: E.g. join-order in query planning

Workload clustering: Identify distinct classes, e.g. write-heavy, read-heavy workloads, per-class decisions

Fully adaptive: Optimal per-task behavior, unrealistic in practice

Robust, predictable, low deployment cost

Analytical overhead

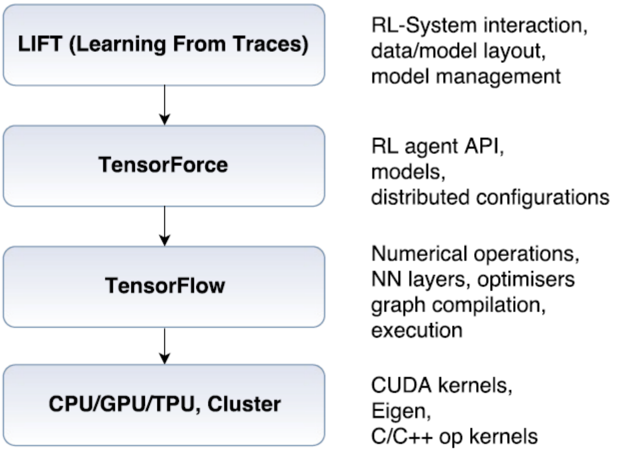
Training/deployment cost factor

Practical Issues continued…

- Many deep learning tools, **no standard library** for modern RL (~2014-2018)
- **Exploration** in production system not a good idea
  - Unstable, unpredictable
- **Simulations** can oversimplify problem
  - Expensive to build, not justified versus gain
- **Online steps take too long**

UNIVERSITY OF CAMBRIDGE Computer Laboratory

LIFT: A software stack for deep RL

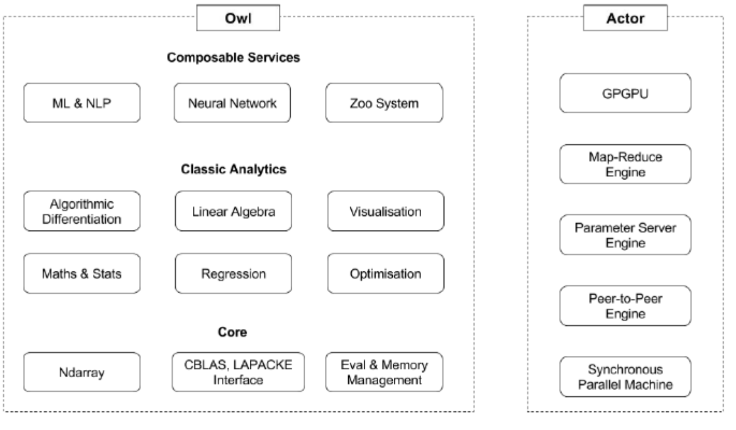


Dynamic Control Flow in Current Frameworks

- There are static computation frameworks WITHOUT dynamic control flow (mxnet, cntk) -> dynamic control flow is in the out of graph host program.
- There are dynamic computation graph frameworks WITH dynamic control flow (PyTorch, DyNet) -> graph is only implicitly defined via imperative operations, cannot do static graph optimisations, typically slower but easier to use.
- There is static computation with dynamic differentiable control flow in the graph -> only TensorFlow offers this among modern deep learning frameworks.

UNIVERSITY OF CAMBRIDGE Computer Laboratory

OWL Architecture for OCaml



Owl + Actor = Distributed & Parallel Analytics

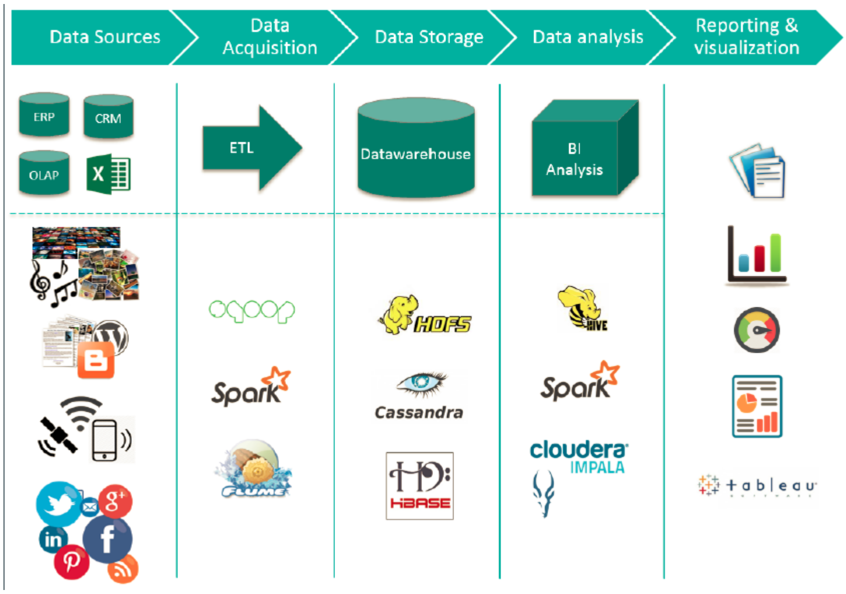
Owl provides numerical backend; whereas Actor implements the mechanisms of distributed and parallel computing. Two parts are connected with functors.

Various system backends allows us to write code once, then run it from cloud to edge devices, even in browsers.

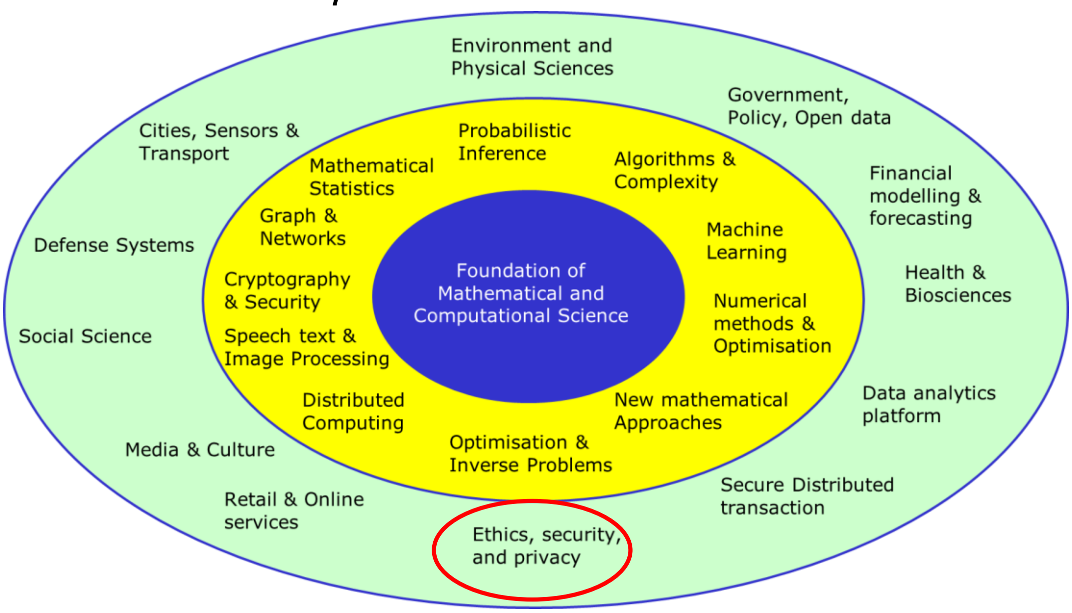
Same code can run in both sequential and parallel mode with Actor engine.

By Liang Wang in 2018

Pipeline of Data Processing…



Broad Landscape of Research in ATI in 2015



Network Structures

- Eight Friends Are Enough: Social Graph Approximation via Public Listings by Joseph Bonneau et al. 2009.
- Many interesting properties can be accurately approximated. This has disturbing implications for online privacy, since leaking graph information enables transitive privacy loss.

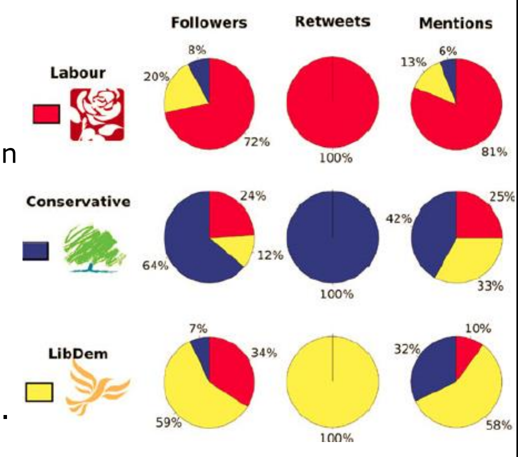
- Cambridge Analytica was caught tampering with elections by exploiting Facebook, but chances are that this is the tip of the iceberg, and that many others.

Cambridge Analytica

Social Network Analysis

- What's in Twitter: I Know What Parties are Popular and Who You are Supporting Now!
- 1,150,000 messages from the main stream of Twitter related to the 2010 UK General Election between the 5th and the 12th of May from about 220,000 users in Twitter.

- With party characteristics, classification algorithm based on Bayesian framework to compute political preferences of users is developed. Accuracy of 86 % without any training and the network topology information.



Modern Data Scientist: The sexiest job of 21th century

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

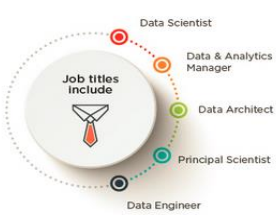
Many Courses offered: e.g. Master Certificate

- Data scientist is the pinnacle rank in an analytics organisation. You will be required to understand the business problem, design the analysis, collect and format the required data, apply algorithms or techniques using the correct tools, and finally make recommendations backed by data.

£ 1799

ENROLL NOW

* VAT Included



Course 1
Data Science with SAS Training

Course 2
Data Science Certification Training - R Programming

Course 3
Big Data Hadoop and Spark Developer

Course 4
Data Science with Python

Course 5
Business Analytics with Excel

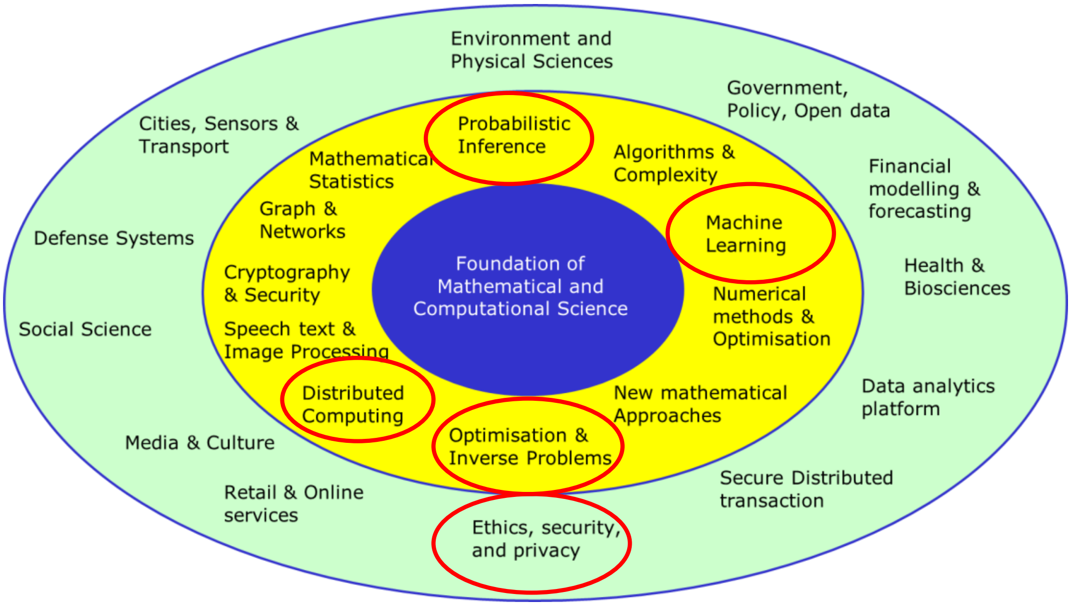
Course 6
Machine Learning

Course 7
Deep Learning with TensorFlow

Masters Certificate
*You will get individual certificates for each course.

UNIVERSITY OF CAMBRIDGE Computer Laboratory

Broad Landscape of Research in ATI in 2015



UNIVERSITY OF CAMBRIDGE Computer Laboratory

Modern Theory of Deep Learning

Why does it work so well?

- On the Expressive Power of Deep Neural Networks PMLR 2017: understanding of how and why neural network architectures achieve their empirical successes is still lacking.
- Ali Rahimi's talk at NIPS(NIPS 2017 Test-of-time award presentation)
- Deep Learning works in Practice. But Does it work in Theory? By L. Hoang and R. Guerraoui. (https://arxiv.org/pdf/1801.10437.pdf)

- Understanding deep learning requires rethinking generalisation
- Fundamental theory behind the paradoxical effectiveness of deep learning. Still open research problem…

Gap between Research and Practice

Device Placement Optimization with Reinforcement Learning

20H with 80GPUs!

Azalia Mirhoseini[*1,2] Hieu Pham[*1,2] Quoc V. Le[1] Benoit Steiner[1] Rasmus Larsen[1] Yuefeng Zhou[1]
Naveen Kumar[3] Mohammad Norouzi[1] Samy Bengio[1] Jeff Dean[1]

Research opportunities ahead!

http://www.cl.cam.ac.uk/~ey204/

Acknowledgements: Michael Schaarschmidt (RL) and Valentin Dalibard (BOAT).