Systems Challenges for Data Science at the ATI

scalable safe, secure, sane systems for data science.
What is the ATI
National institute for AI&Data Science/ML
What ATI does & how we know
Dog-fooding
Measure ourselves

1. Expressions of Interest analysis (by hand/eye)
   • List of 8 challenges & 10 programmes
   • https://www.turing.ac.uk/research/challenges

2. Topic models from web pages & projects (semi-automatic)
   • Crowdsourced & LDA
   • https://www.turing.ac.uk/research/research-areas
   • https://www.turing.ac.uk/people/researchers/jon-crowcroft

3. Bibliometric analysis (automagic)
   • https://arxiv.org/abs/1903.01517
Systems Challenges for Data Science at the ATI

scalable safe, secure, sane systems for data science.

Jon.crowcroft@cl.cam.ac.uk
Big Data, Science, or Market Research

• Computational Sciences == Supercomputer/HPC
  • Physics/meteo/astro
  • Genomics
  • Chem/materials

• Analytics == Marketing, Data Center
  • Facebook/Google, advertising/recommendation
  • Business optimisation (amazon)

• (BIG) Data Science in between....
  • Much Big Data is social or economic
  • Some in between (public health)
Hyperscale Challenge

• Rack scale systems in-between current DC & HPC...

• Lots of (ARM) cores 1000/socket, NUMA

• low latency interconnect

• Lots of storage – smarts included (fs, obj, blk)

• (>1 Petabyte SSD in rack, low power)
Decentralised

- Much of the data doesn’t need to go to cloud
- Stay-at-home, in office, in built environment infrastructure
- Smart home, transport, energy, even governance
- Aggregation is your friend in many ways....
Programmable

• S&Python&SQL v.Spark/R v.Hadoop/Latin?
• Or is way forward is DSLs & Functional ...
• Domain Specific Languages
  • even spreadsheet&visual
  • Integrate with map/reduce, stream, query
  • Via pure functional, clean, and specialisable...
High Throughput & Low Latency

• Layered composition is a bad idea...
  • Ousterhout (stanford)

• But one of the ways we simplify complex sys
  • Is abstraction through layering....

• Need better approaches, simply too slow
  • Specialisation – unikernels/docker
  • Pass thru/offload
  • In network processing
Confidentiality & Integrity

- FCA & Farr use cases – hard partition needed
  - Many tenants
  - Insider is a threat too, evil or incompetent

- Solution already in iOS enclave
  - But a single user device using ARM trustzone
  - With Intel SGX can do better

- So integrate hypervisor/unikernel
  - And some analytics framework with enclave
The Compliance Challenge

• Isolation & Provable Least Privileges is only part of the challenge
• Applications still must not mis-behave
  • Data should not be re-identified
  • RBAC, Information Flow Control, Provenance etc required...
• But ML/AI Based decisions will have to be justifiable/explicable
  • Harder problem – not just a systems challenge
  • Need to control input, learning and output
  • Clear how to do this in (e.g.) Bayesian inferencing or other basic tools
  • Less clear how to do this for deep learning...
Conclusions

• Ways forward with partners clear
• Have good global community
• Timely technology emerging
• Still many systems challenges
• ATI is a good UK convenor for such work
Some example other project ideas....

- Zika – two2 population epidemic – infer model with partial data 😊
  - Zipfian multi-graphs? Parsimonious model? Probabilistic programming
- Highly distributed analytics (dabox/hat)
  - Privacy/ by aggregation (diffpriv structurally enforced)
- UK industrial trading graph resilience
  - We design resilience into utilities – why not commerce too?
- Is it human?
  - There’s increasing machine traffic on the net- twitterbots etc... how to tell?