Privacy & Security in Data Science & Cloud -
Jon@Cognition X
Who am I?

• Professor of CS in Cambridge since 2001
  • Cloud – from Xen to Docker
  • IoT & Kids – Raspberry Pi to Computing at Schools
  • Also into community networks & social media

• Previously at UCL since 1980, building the internet
• Previously in Cambridge in the 1970s....

• Currently also 50% at the Alan Turing Institute for Data Science
  • The national data science research institute
  • Partners include Lloyds, HSBC, GCHQ, Intel, etc etc

• Next up – some projects I like to dabble in...
High Throughput & Low Latency inside pet data centers (even just rack) not *all* solved

• Layered composition is a bad idea...
  • Ousterhout (stanford)
  • 100x speedups hand crafted today

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

• Need better approaches, simply too slow
  • Specialisation – unikernels
  • Pass thru/offload fpga/gpu
  • In network processing
  • Cross layer – remove cruft:-
    • Hadoop or SparkR or graphx->linux->GPU/NIC/Switch->fabric....

See [https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=16281](https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=16281)
Decentralised – IoT/Smart-X pet warning...

• 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....
  • Relevance
    • cyberphysical data becomes exponentially irrelevant with distance&age
    • Think inverse square laws (or path loss coefficients😊)
• But there’s still plenty of centralised stuff
  • that is inherently gathered together in a cloud (and grooving with a pict😊)
• Community mesh networks with data in developing countries (GAIA)
  • 100 bucks gets you long range wifi & a terabyte...
Jon’s own pet nets are data science too

• Measure Nets
  • Traffic, topology, dynamics
  • Lots of kinds of nets (tech, social, transport, eco, neurological etc etc)

• Data sets scale
  • Log every packet, need net back to retrieve/process! 😊

• Privacy, etc
  • Traffic is confidential, traffic matrix is confidential
  • Traffic analysis can infer identity even if data de-identified
  • Anonymizing graphs is not really solved problem....

Examples:- [http://conferences.sigcomm.org/imc/2016/program.html](http://conferences.sigcomm.org/imc/2016/program.html)
Jon’s pet(small) project ideas….

• Zika – two population epidemic – infer model with partial data 😊
  • Zipfian multi-graphs? Parsimonious model?

• Highly distributed analytics (databox/hat)
  • Privacy/ by aggregation (diffpriv structurally enforced)

• UK industrial trading graph resilience
  • We design resilience into utilities – why not commerce too?
  • Risk/Expected loss in transaction if ID-theft or privacy invasion

• Is it human?
  • There’s increasing machine traffic on the net- twitterbots etc…how to tell?
Why are we here?

• Cloud/analytics ecosystem -> Big Data Hype
  • Big Data (storage/processing) affordable
  • ML tools pretty reliable (but care with reproduceable!)
  • E.g. Netflix prize

• Accidentally discovered by Google =>
  • Had to build big data center to index web
    • Store pages from Spiders&Robots
    • Run Pagerank (and 200{ special sauce heuristics) fast

• Light bulb moment – click through value.....
  • Best market research engine since Nielsson
  • Landgrab on entire advertising business
  • => Gold Rush!!!
Hyperscale is cheap

• “Quantity has a quality all of its own” -- Iosif Vissarionovich Dzhugashvili

• Cloud/data center v. HPC
  • Cloud is affordable/scale out – hadoop/spark/graphx EC2 Azure etc etc
  • HPC specialised capability – specialised stacks/libs mpi etc – talk to your provider
Hyperscale is Easy Peasy Programmable

• Python&SQL v. SparkR v. Hadoop, etc etc
  • Democratised data science

• Domain Specific Languages
  • even spreadsheet&visual
  • Integrate with map/reduce, stream, query
  • Apply/cross compile to exotic hardware
Confidentiality & Integrity – Use Cases & Law

• FCA & Farr use
• Currently caught between two forces
  • GDPR – General Data Protection Law
  • IPB – Lawful intercept++
• Add two economies of scale
  • Scale out data centers – sub-linear cost in number of cores & memory
  • Storage prices falling (1 petabyte of flash for 1M USD)
• Currently, Farr & FCA own own data centers
  • As do commercial equivalents (pharmas, banks)
  • Use strict (RBAC) access control & audit trails
  • Penalties for abuse (lose job, fine, go to prison etc)

“Privacy: It’s the law. Get Over It”
Confidentiality & Integrity - Revelation

• Queries on federated data in Farr (and FCA) can reveal personal info
  • NHS Scotland & Wales linked up all the separate data bases (federated)
  • At the Farr, you can run queries across them all
    • Who’s in this city block who is over 2 meters tall and has an STD

• Lots of more complex examples with joins
  • tuple generating queries reveal sensitive stuff not clear from simple analysis

• Require analysis of schemas & queries to prevent former

• May need Differential privacy to prevent latter
  • Differential Privacy comes out of Microsoft Silicon Valley and
  • Does clever stats to limit what level of detail is revealed by queries
  • Three approaches (all involve knowing database stats – range/max/min)
    • Don’t answer if query response too specific
    • Add chaff to raw data
    • Fuzz responses.

• What about known unknowns and unknown unknown 3rd party data
  • E.g. of re-identifying public figures in Massachusetts healthcare
  • And stars in Uber/Yellow cab ride data
Confidentiality & Integrity- Outsource Limits

• If we want to reduce costs, move out to cloud
  • But still meet GDPR requirements
  • Need to solve various problems with isolation

• Problem: Cloud operator normally has privileges
  • Access to h/w, OS, NAS, etc
  • Honest but curious (aka mission creep, shareholder value)
  • Or just exploited - even a hypervisor haz bugz 😊
    • Lets operator, bad guys outside or in other tenants access data/computation

• So rules/regulations/law don’t let you run on bare cloud platform
  • Need new tech to fix this....
Confidentiality & Integrity making safe havens

- Use of intel’s SGX with Containers or Hypervisor (virtualisation technologies)
  - Run part of OS, or container or hypervisor in SGX domain
  - TCB, with keys managed elsewise
- Can be used for enforcing isolation (if you trust intel)
  - See Imperial Scone work recently
  - Equivalent to Apple Enclave on iOS9/ARM (trust zone)
  - note possible IPB conflict (witness FBI frustration)
- Can be used for integrity checking too....
  - c.f. Microsoft VC3, Hadoop on SGX
- However, law may not comprehend this yet
  - “storage” = processing in legal terms
  - Crypted storage doesn’t get you off the hook (yet) even with keys managed by user
- Last step is add a blockchain/distributed ledger for tamper proof audit trail.....
  - May allow re-identification too....needs care

Privacy: It’s hard, but we’re working on it
Confidentiality & Transparency

• GDPR *also* requires explicable ML
• If decision/output might discriminate -1
  • Race, gender, age etc....
  • E.g. ML determining what hotel/travel/insurance to offer customer....
• Transparency may require ML include trace/audit of training set -2
  • Contradiction- training data might include ground truth
  • so allows re-identification of customers
• Hard to fix in some ML for 1&2
  • Especially trickier in deep learning
  • Less so for ML classic (radom forrest, bayesian inference)
  • E.g. If infer rule that is equivalent to a gender bias, can supress it explicitly
  • E.g. Pink cars used for school run might be correlated with women driver
  • So don’t allow a priori discount....😊
What more could I possibly say?

• Questions?
  • Now with added brexit?
  • Or we could talk about Zero Knowledge Systems (harder😊 )

“Privacy: It’s complicated, but Real Soon Now”