Future Safe Havens

Jon Crowcroft,

http://www.cl.cam.ac.uk/~jac22

Private Data Center->Public Cloud

- ATI partners e.g.
  - Farr/NHS Scotland
  - HSBC
- Motives for public cloud
  - Scale out/cost save
  - Higher Throughput analytics
  - Share "access" with more researchers
  - <Yours goes here>

Infrastructure Location

- Keep friends&enemies near:
  - Legal/Regulatory Stuff (incl GDPR)
  - Latency/Availability etc
  - Control (physical access etc)
- Need to virtualise these (better)
  - Crypt Data at rest
  - Crypt data during "processing"
  - key management etc
  - ***Enclave***... SGX,Trust Zone, AMD, CHERI

GDPR - 2018 - right to an explanaion

- MISTAKES HAPPEN
  - ERROR ON INPUT
    - E.G.AMOUT OF $, AGE, ETC
  - MISTAKE IN CODE
    - E.G. IF (C=3) {} ...
  - MISTAKE IN TRAINING
    - E.G. SELECTION BIAS - PROB OF RE-OFFENDING ONLY TRAINED ON OFFENDERS
  - MISTAKE IN ML/INFERENCE
    - E.G. ACQUIRE A LATENT VARIABLE/RULE == GENDER (OR AGE)
  - SOMETHING WE HAVNT THOUGH OF YET
    - EMERGENCE?
- RIGHT TO REDRESS, AND BALANCE ASYMMETRIC POWER

SGX opportunity

- Not the only piece, of course
  - Static/dynamic analysis etc
  - Unikernels & s/w verification
- Can use SGX on
  - Container (SCONE)
  - Platform basis, Hadoop, Flink, Spark

    https://www.microsoft.com/en-us/research/publication/vc3-trustworthy-data-analytics-in-the-cloud
  - Or application basis

MARU….@ turing.ac.uk

- ATI w/ Intel, Dstl, Docker, Microsoft
- Hiring:-

https://www.turing.ac.uk/jobs/research-associate-maru-project/

- Compare what is in SGX
  - Enter/leave cost, crypt memory o/h etc
  - Hypervisor?
- Compare w/ container on trustzone, cheri, AMD etc
  - Common APIs for keys etc
  - Virtualize?
- Pen test
  - many side channel pb
  - What if weak homomorphic crypto & diff priv?

Public Cloud->Databox (or HAT)

- Databox (and hat) take opposite view
- Re-decentralize
- Keep analytics/ML as a service
  - Mix of distributed, priv pres ML+
  - Hierachy of 3rd party aggregators, MPC
  - http://www.databoxproject.uk/
- HAT reverses direction of value…
  - Audit (distributed ledger)
  - Get paid (money (real or vurt)
  - https://www.hatdex.org/

Container – migration&replica

- Replicate (to cloud enclave)
  - for recovery (from fail,theft,loss)
- Migrate (to other personal cloud)
  - for low latency
- Most new data is append only – so use distributed ledger
  - (tamper proof logs – see datakit in docker)
- Consistency of replicas –
  - e.g. use fpaxos

Distributed Analytics

- Motives e.g.
  - Move code to data
  - Keep data close to owner/primary user
  - Guarantee can audit trail access
  - Add yours here
- Challenges
  - Depends on ML technology of choice & goal
    - PCA/Clustering, random forests
    - Curve fittign (regression etc)
    - Model Inferencing – e.g. Bayesian inference
  - Distrubuted differential privacy tricky
  - Hierarchical versus P2P?

Distributed Analytics

- Hierarchy easiest
  - Aggregation points/servers broker "model learned so far"
  - Have to be trusted by subset of leaves
  - Leaf can choose to change aggregator
- P2P just extension of this to dynamic, faster choice
- Distributed/Parallel ML
  - From data centers
  - Clustering on tuples easy If independent

Future Proof for GDPR

- Privacy by Design and by Default – HAT address all GDPR privacy requirement from its design principle to its security solution.
  - HAT ecosystem data exchange is based on fully specified privacy terms - time specific, recipient specific, minimum data points **specific** with **full intention disclosed.** Violation against any of such terms may result a ban from the Ecosystem.
- Consent by design and by default -
  - the PCST PoC mandates a "specific, informed and freely given and unambiguous" intension disclosure of data usage, for every single personal data access instances.
  - HAT technology ensures that an exchange is only authorised and kept valid by individual's case specific consent
- Rights for Individuals by design and by default – encapsulated personal data containers isolated for each individual, allows an individual is in full control of its HAT, hence inherently owns all of the following:
  - Right to Access | Right to be informed | Right to rectification | Right to restrict processing | Right to object to market
  - Right of data portability | Right to be forgotten | Right to object to automated decision making and profiling
- Accountability and governance - PCST CoP mandates every ecosystem member to higher level of accountability and governance practice.
  - Record keeping – HAT ecosystem automatically tracks data exchange, even at a much more granular level than GDPR requires – it documents the exchange parties, time of access, detailed data points, intension and T&C, for every single transaction

Things we're not covering today

- Database (Farr/ATI work now)
  - Query planning w/ privacy
  - K-anonimity
  - Weak homomorphic crypto etc
- Threat modeling
  - Assuming implicit☺
  - Suffice it to say hypervisor vulnerabilities exist
  - So need trusted stuff on untrusted platform…
  - …on new trusted stuff…

Who Am I?