Data on the Edge

A new approach to Big Data

Rolando Martins
FCUP/CRACS, INESC-TEC

What is Big Data?

- Going behind the buzzword:
  - Set of technologies that were introduced with the appearance of cloud computing
  - Envolves the storage and processing of large data sets
  - Supported by:
    - Distributed storage
    - Fast networking (interconnects or fabrics)
    - Computational nodes
    - Computational frameworks

Current Approaches to Computation for Big Data

- Streaming
  - Processing is done in real-time, e.g., Financial - Electronic trading
- Batching
  - Processing is done off-line, e.g. Scientific computation, such as Molecular Biology
- Hybrid
  - Combines streaming and batching
    - The real-time processing has pre-processed data from past batch processing, normally historical data, e.g., recommendation systems

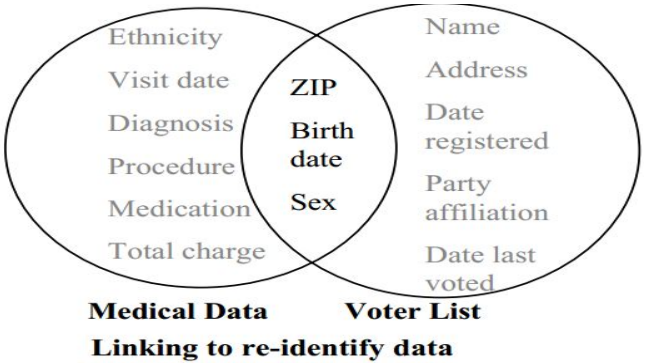
Issues with Big Data - (Lack of) Privacy

- Input

  anonymized medical database + voter list

- *Output*

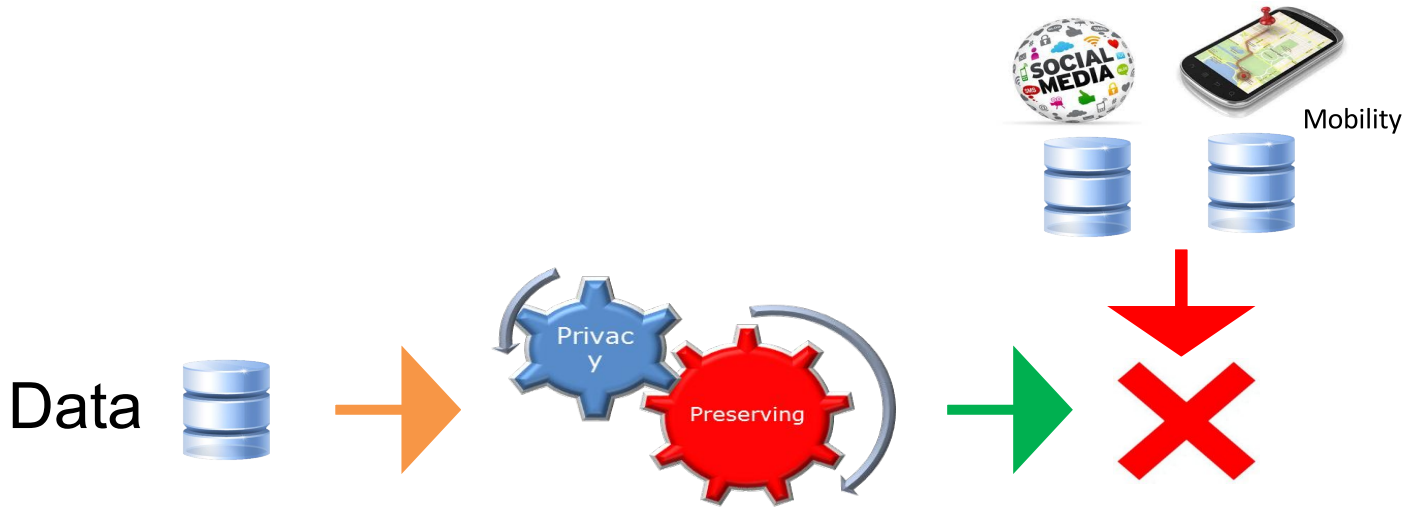
  health record of the governor of Massachu



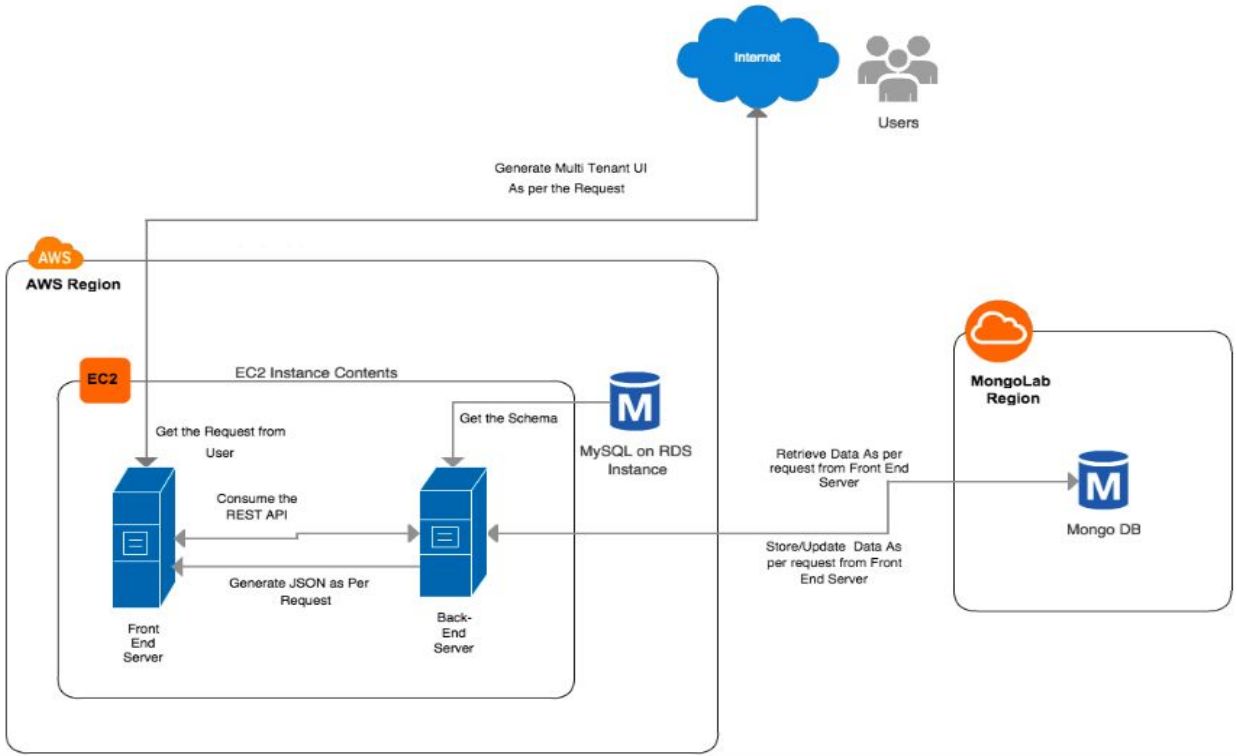
In: k-anonymity: a model for protecting privacy

k-anonymity: a model for protecting privacy
Sweeney, L. *Int. J. Uncertainty Fuzziness And Knowledge-Based Systems* **10**, 557–570 (2002).

Privacy Still An Open issue

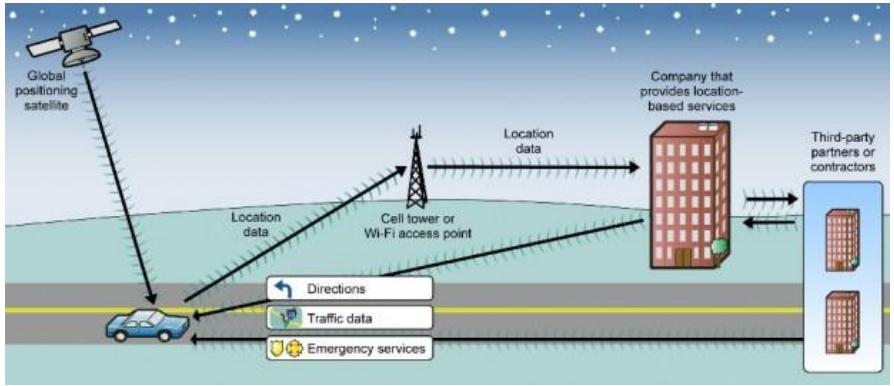


Tracking: Log collection



Lack of informed consent: Big Brother is Watching You!

- **“The Cost of Lost Privacy: How Google and Datamining Drive Economic Inequality in Our Nation”** http://www.huffingtonpost.com/nathan-newman/the-cost-of-lost-privacy-_b_891972.html
- **“Gov’t Report: Car Companies Collecting Data Through Your Car Aren’t Always Telling You What It’s Being Used For”** http://www.theblaze.com/stories/2014/01/08/govt-report-car-companies-collecting-data-through-your-car-arent-always-telling-you-what-its-being-used-for/



Big Data Outside Infrastructure Clouds?

- Infrastructure Clouds/Private Datacenters:
  - Data is privately owned and in some cases in-house generated
  - And is hosted in a single trust/authoritative domain
  - Almost infinite resources
- What if these assumptions do not apply or only are partially applicable?
  - Limited or absent connectivity to powerful backends:
    - High density scenarios, e.g., stadiums, concerts
    - Disaster scenarios
  - Privacy and security concerns:
    - Data belongs to individual users
    - Entity hosting the data must enforce privacy, data security and right-to-be-forgotten

Alternatives are here!

- Privacy Preserving Technologies:
  - TOR, GnuNet,
  - BitCoins
  - P2P Systems

- But what else can we do in limited or transient connectivity?
  - What if mobiles devices could already cooperate to offer a new cloud computing tier?
  - Could we harvest this untapped resource?
  - Could it also help us privacy-wise?

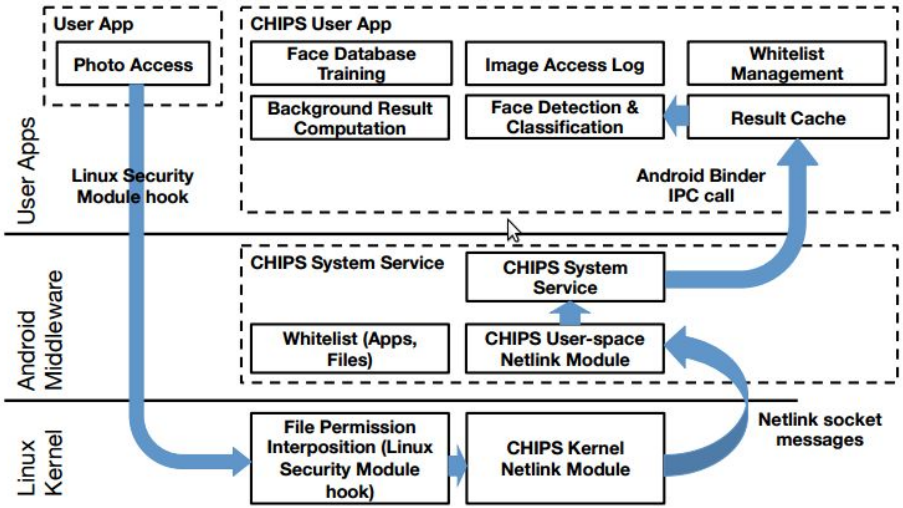
Hyrax - Exploring Edge Clouds

- A CMU-Portugal Project:
  - Partners include INESC TEC, NOVA LINCS, IT and CMU
- What is an Edge Cloud?

“a computational storage cloud comprised solely of a collection of nearby wireless edge devices, with the purpose of pooling these devices data and processing power to support a new class of proximity-aware applications that benefit the owners of these devices.”

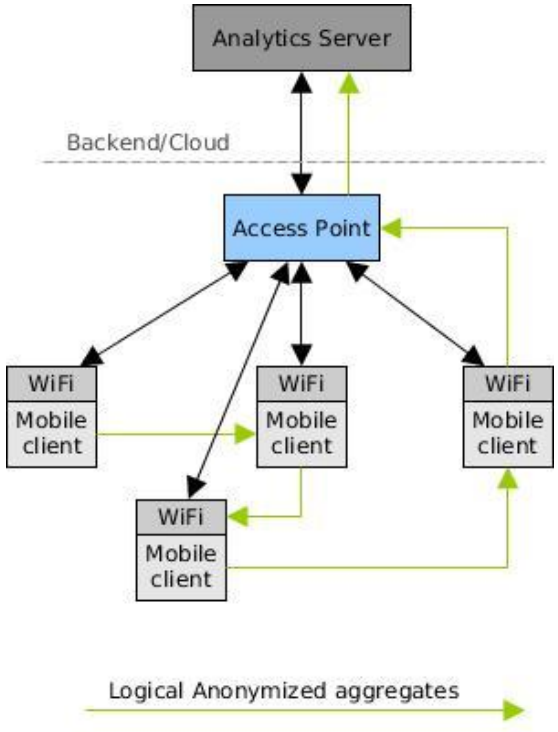
Privacy and Security in Edge Clouds - I

- CHIPS: Content-based Heuristics for Improving Photo Privacy for Smartphones (Jiaqi Tan'14)



Privacy and Security in Edge Clouds - II

- Anonymized aggregates



Conclusion

- Edge Clouds can offer support for more enhanced solutions for Big Data
  - Although with limitations, it offers a novel cloud computing tier that can:
    - Enhance privacy preserving approaches
    - Support a new class of applications that explore data locality and context awareness
    - While providing a real alternative to infrastructure clouds

Thanks!