BIOBANKCLOUD

Your PaaS for Biobanking

Secure Storage of DNA Data

Alysson Bessani

(joint work with Vinicius Cogo, Francisco Couto and Paulo Verissimo)

Ciências ULisboa

Faculdade de Ciências da Universidade de Lisboa

Motivation: Personalized Medicine

- Drugs and treatments specifically designed for an individual and its current condition
  - Genomic analysis will be a common practice

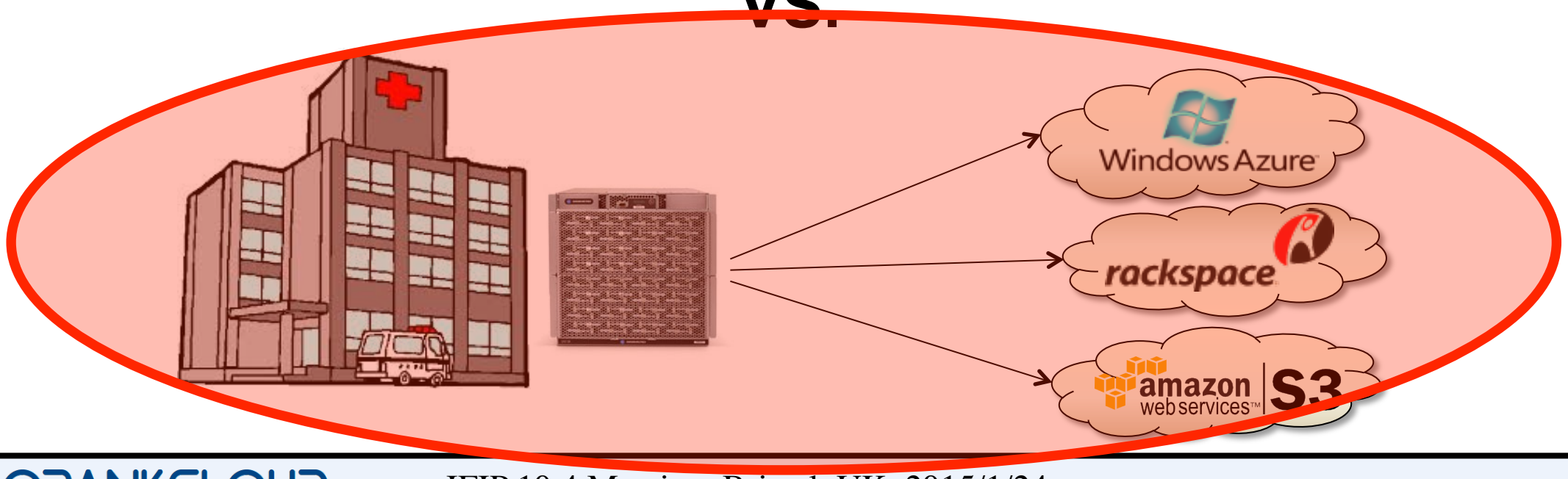
- Required infrastructure
  - Sequencing machine
  - Computing infrastructure
  - Storage infrastructure

Genomics Raw Data: FASTQ files

- Format generated by sequencing machines
- Each FASTQ file represents a number of reads
  - Each read contains a sequence and its quality score
- A genome can be represented by one or more FASTQ files -> up to 300 GB!!!

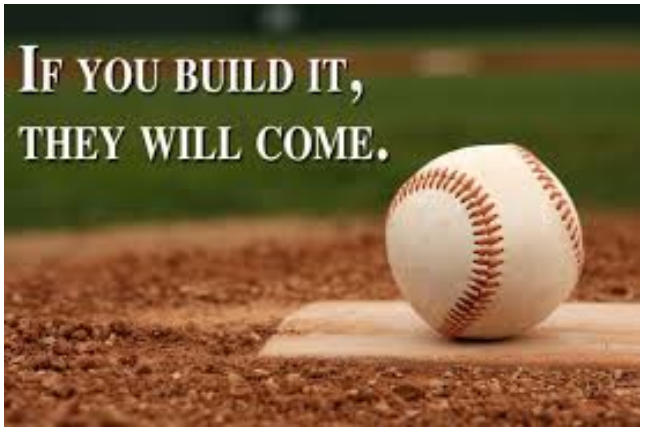
PM Infrastructure

vs.



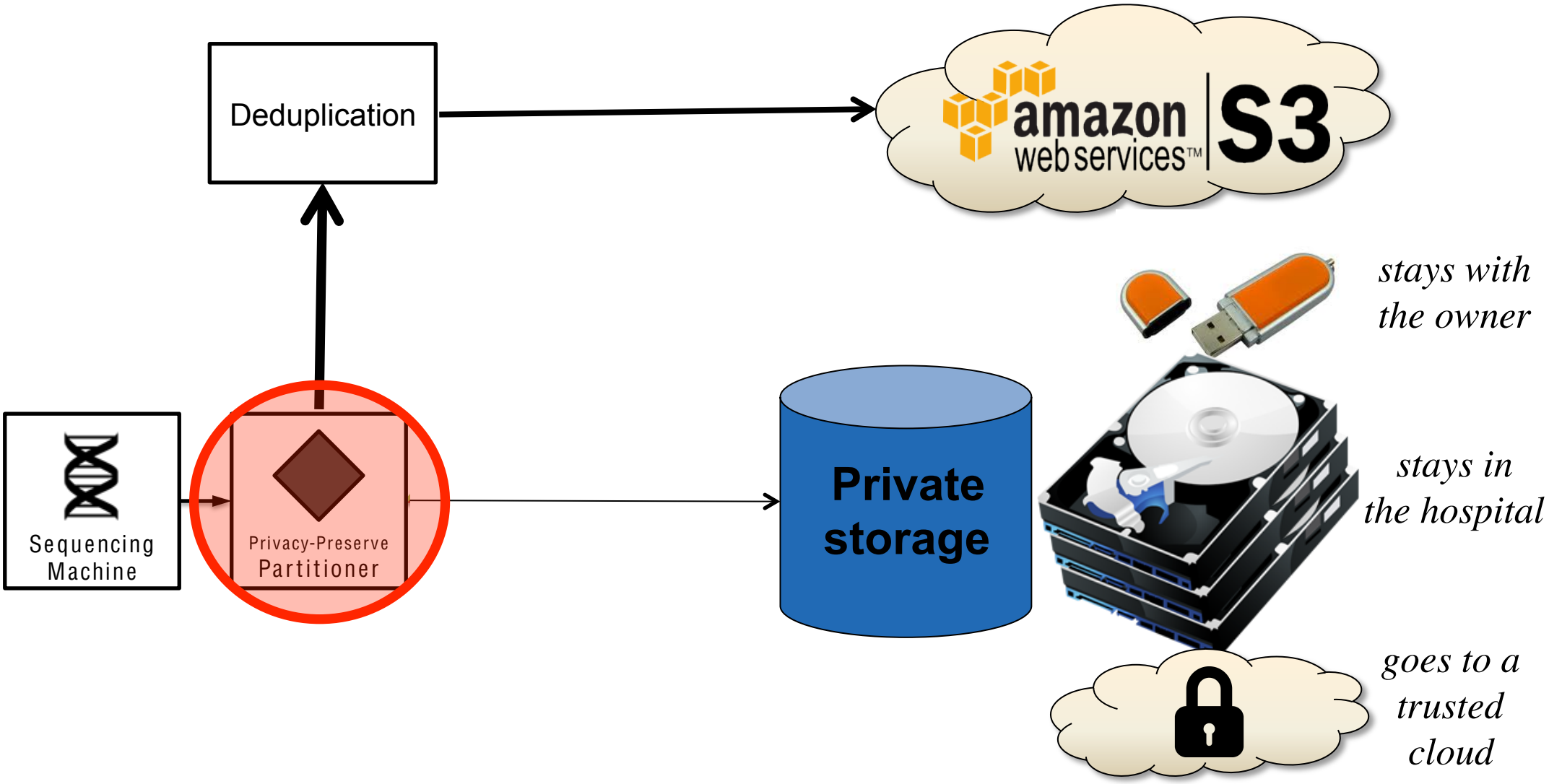
Challenges in Using Public Clouds for PM

- It's mostly illegal!
  - -Huge privacy concerns
  - -But things may change…



- How to decide what goes to the public storage and what not?
- How to minimize storage space in the cloud?

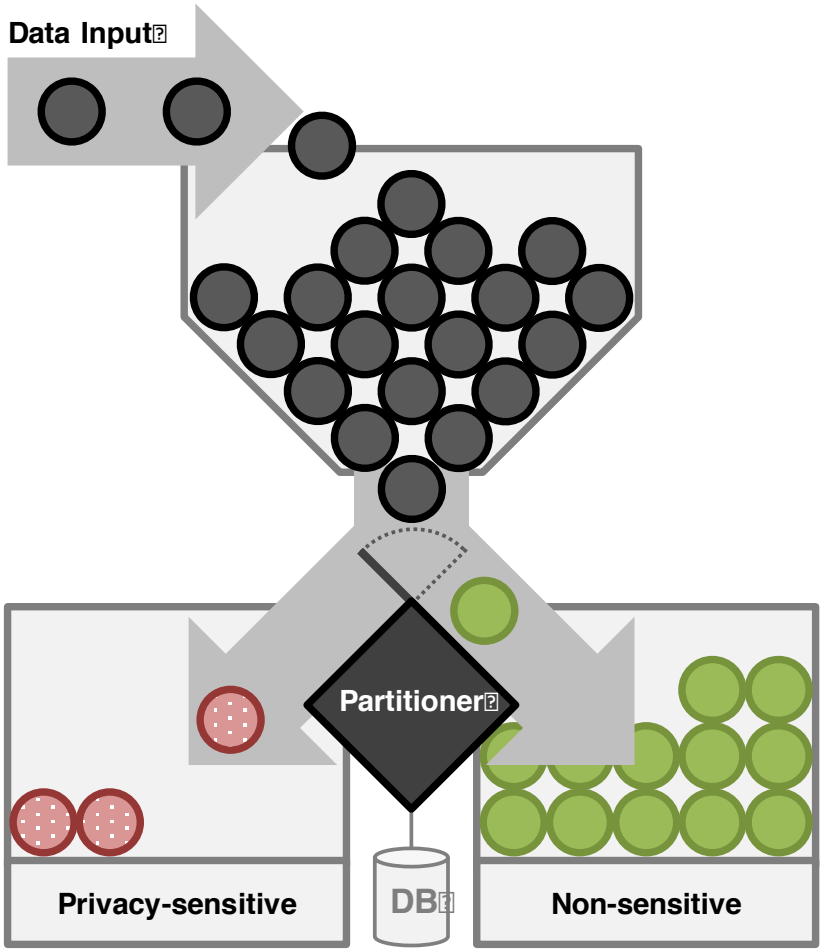
Storage of Raw Genomics Data



How to decide what can go to the cloud?

- Privacy-preserving DNA partitioner
  - Classify DNA sequences (reads) in privacy-sensitive (can leak information about its donor) and non-sensitive

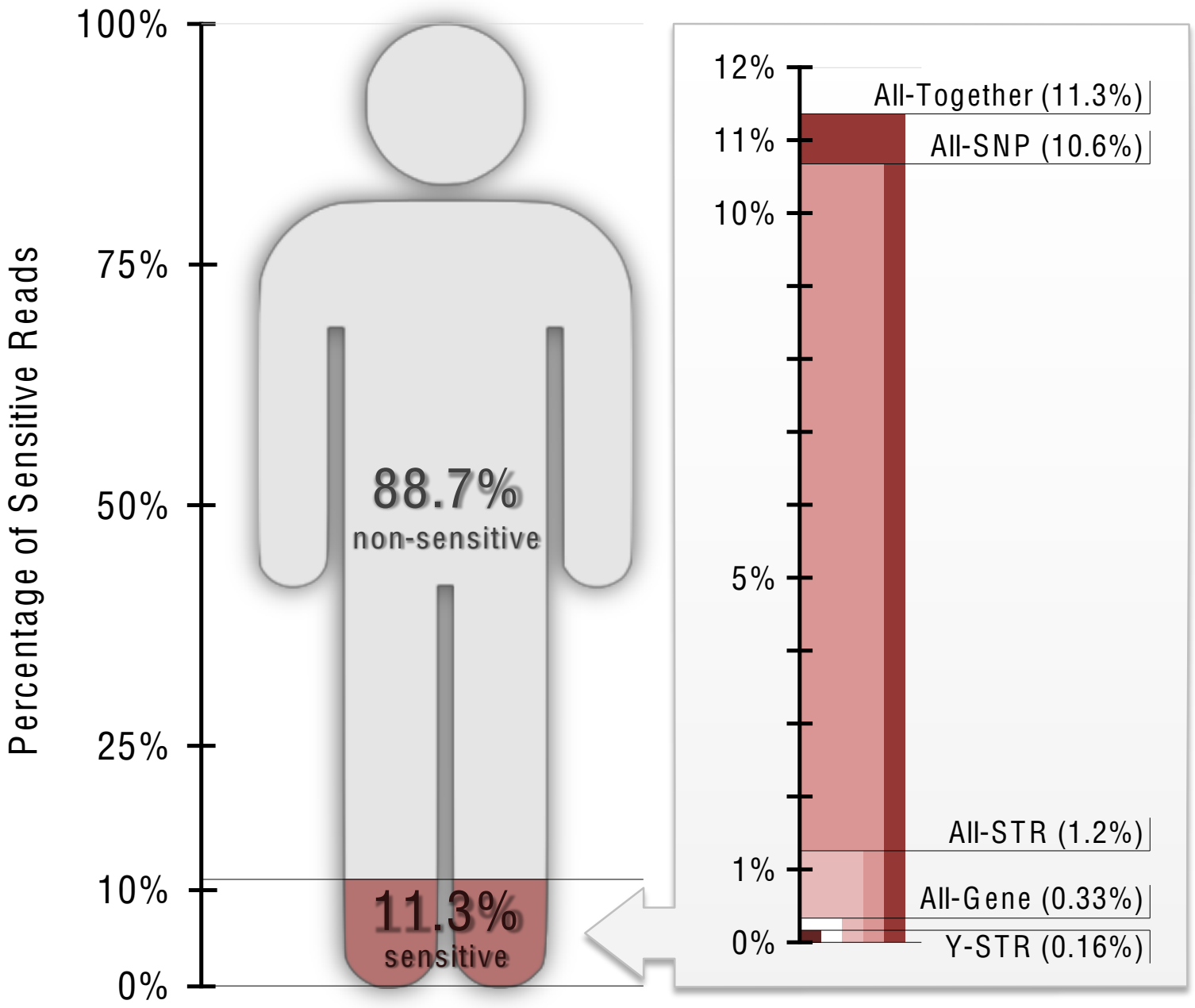
- Features
  - Accuracy
  - Performance
  - No false negatives



Privacy-preserving DNA partitioner

- The **blacklist database** of privacy-sensitive sequences can be generated using two methods
  - Genetic genealogy profiling (Y-STR)
  - Rare variants presented in individuals (Allele Frequency)
  - Disease-related Genes (using protein databases)
- These three methods are sufficient to filter sequence data and avoid all known attacks
- Prototype implemented using a big **bloom filter**
- Important: The database can be updated as new threats are identified

How much of a human DNA is privacy-sensitive?



Privacy-preserving DNA Partitioner

- Some results for a conservative partitioner
  - False positive rate: **1 in a Million**
  - Number of sequences in the blacklist: **1178 Million**
  - Size of the filter: **5.6GB** (fits in main memory)
    - 7x smaller than the size of the sequences in the blacklist
  - Throughput/core: **13.2 Million bp/sec**
    - 44x faster than the fastest sequencing machine
    - Scales linearly to up to 8 cores

Open Questions

- What to do if new attacks are discovered?
  - How often this happens?
  - We only filter what is in the black list
  - For this reason we should encrypt the data

- How effective is deduplication?
  - Within a single individual
  - Within multiple individuals
  - Informed guess: 80%