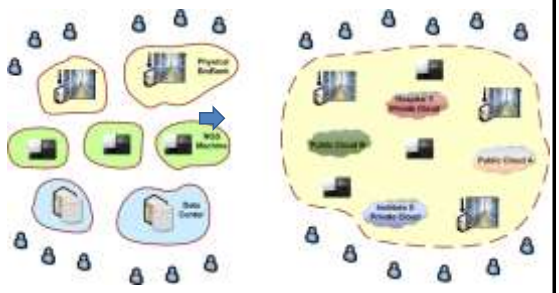
BiobankCloud @ULisboa

Francisco M. Couto

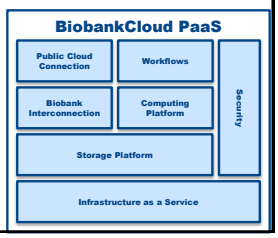
June 2, 2014
Meeting on IT/ Informatics
BBMRI-ERIC

E-biobanking

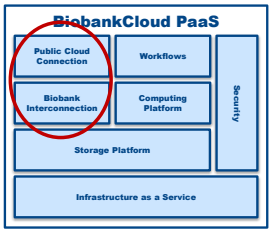


Biobank PaaS

- Distinctive characteristics to other PaaS
  - Focuses on biobanks
    - Rather than on data analysis
  - Within EU regulatory framework
  - Similar object model as LIMS



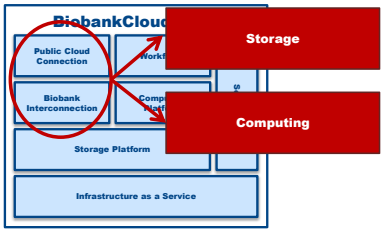
BiobankCloud @ULisboa



Overbank

Public cloud connection and biobank interconnection

Overbank Challenges

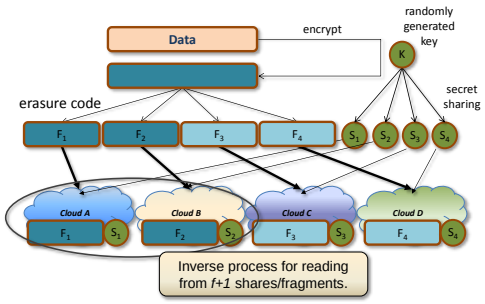


STORAGE - DATA

Storage – Charon

- Ongoing work:
  - A protocol called DepSky
    Bessani, Alysson, et al. "DepSky: dependable and secure storage in a cloud-of-clouds." ACM Transactions on Storage (TOS) 9.4 (2013): 12.
  - A FS called SCFS
    - accepted paper at USENIX ATC'14
  - Charon still needs to deal better with large files
- Specifications
  - Data encryption (for confidentiality)
  - Replication using erasure codes (for availability, avoids vendor lock-in)
  - Secret sharing (no single cloud can read the entire data)

Charon



STORAGE - METADATA

Storage – Meta-data

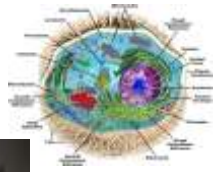
- MIABIS for sharing information about biobank collections
- Expose meta-data, while maintaining protected the data
  - Important in those countries that do not allow to externalize NGS data
- Explore data locality for running workflows where data is already located
  - Minimize the data transfers for running workflows
- HDFS+YARN already deals with data locality
  - Function-shipping instead of data-shipping

Interconnecting at the Metadata level

- MIAMES Matching
  - Unstructured text values
    - String Matching
  - Ontology mappings (e.g. Disease Ontology)
    - Semantic Similarity
  - Intermediating datasets
    - Ontology matching
    - Text Mining

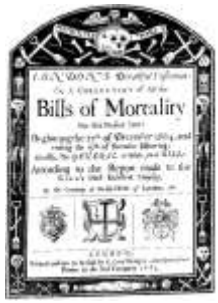
Ontology Mappings





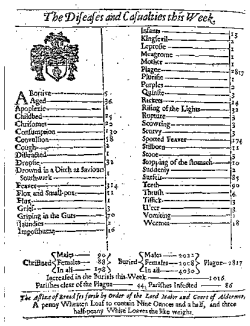
London Bills of Mortality

listed possible ways to die throughout the sixteenth, seventeenth and eighteenth centuries



Source: http://faculty.up.edu/asarnow/popular7.htm

Aggregated Stats

Similar but not the same

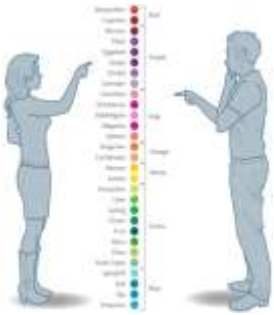
Semantic Similarity

"A semantic similarity measure is a **function** that, given two ontology terms or two sets of terms annotating two entities, returns a numerical value reflecting the closeness in **meaning** between them."

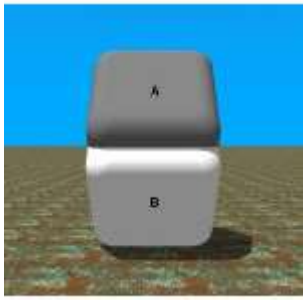
C. Pesquita, D. Faria, A. Falcão, P. Lord, and F. Couto, Semantic similarity in biomedical ontologies, PLoS Computational Biology, vol. 5, no. 7 (e1000443)

F. Couto and H. Pinto, The next generation of similarity measures that fully explore the semantics in biomedical ontologies, Journal of Bioinformatics and Computational Biology, vol. 11, no. 1371001, 2013

Different Perceptions of Similarity

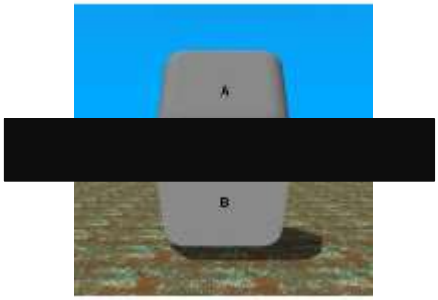


Are we good at measuring similarity?



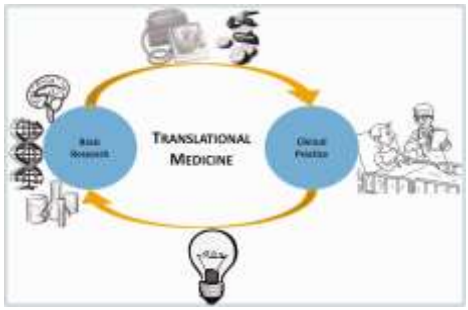
Source:
http://www.grandparents.com/food-and-leisure/did-you-know/optical-illusion-pictures

Are we good at measuring similarity?



20

Knowledge workflow in translational medicine.



Machado, Catia M., et al. "The semantic web in translational medicine: current applications and future directions." *Briefings in bioinformatics* (2013): bbt079.

SemEval-2014 Task 7

- the recognition and normalization of named entity mentions is a fundamental task
- clinical notes from MIMIC II database
  - manually annotated for disorder mentions
  - and normalized to an UMLS Concept Unique Identifier (CUI) when possible.

SemEval-2014 Tasks

TASK A
- The rhythm appears to be *atrial fibrillation*.
- The *left atrium* is moderately *dilated*.
- 53 year old man s/p *fall from ladder*.

TASK B
- atrial fibrillation
  - C0004238; UMLS preferred term *atrial fibrillation*
- left atrium...dilated
  - C0344720; UMLS preferred term *left atrial dilatation*
- fall from ladder
  - C0337212; UMLS preferred term is *accidental fall from ladder*

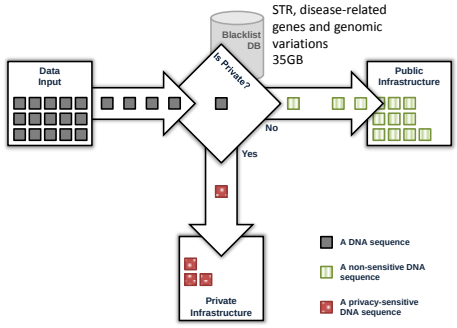
Our approach

- Stanford NER
- Lucene Disambiguation

	Task A						Task B	
	Strict			Relaxed			Strict	Relaxed
Run	Precision	Recall	F-measure	Precision	Recall	F-measure	Accuracy	Accuracy
1	0.753	0.663	0.705	0.914	0.815	0.862	0.402	0.616
2	0.752	0.660	0.703	0.909	0.806	0.855	0.404	0.612
3	0.752	0.660	0.703	0.909	0.806	0.855	**0.405**	**0.615**

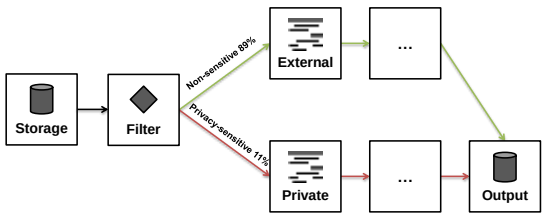
FILTER

Pipeline



Performance

- Filter throughput = 40x to 1600x faster than current NGS machines



Final Remarks

- BiobankCloud will provide a cloud-computing platform as a service (PaaS) for the storage, analysis and inter-connection of biobank data.
  - OverBank storage architecture
  - Privacy-preserving Disclosure Filter
- Meta-data Integration
  - Ontology matching, text mining, semantic similarity
  - BMSRIs principles of data management and sharing challenge:
    - "to encourage data sharing, systematic reward and recognition mechanisms are necessary".

Thanks!

www.biobankcloud.com

Francisco M. Couto:
www.di.fc.ul.pt/~fjmc