SODA - Scalable Oblivious Data Analytics

Meilof Veeningen, Philips Research

The project SODA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731583.

The problem – Privacy in joint data analytics

- Unlocking the huge potential of Big Data requires combining datasets:
  - from mutually distrusting parties, e.g., hospitals and insurance companies
  - containing sensitive information for which the consent of data subjects is required
- With huge legal/financial/reputation consequences if something goes wrong
- On the other hand, potentially huge gains…
  - In healthcare, “extracting knowledge is the fastest, least costly, and most effective path to improving people’s health” (Prof. Butte, Stanford)
  - Potential healthcare savings of the use of big data: 300B€ in US alone

NHS to carry on selling patients' medical data to insurance firms despite history of blunders over illegal use of the information

- NHS will continue to sell patients' medical records to insurance firms
- Data includes personal details of diagnoses, dates of birth and postcodes
- Expert said proper checks were now in place to prevent misuse of records

THE DAILY MAIL

UPDATED: 10:23 GMT, 27 November 2014

By BEN
PUBLISH

Dan Munro
Contributor

I write about the intersection of healthcare innovation and policy.

Opinions expressed by Forbes Contributors are their own.

PHARMA & HEALTHCARE 12/31/2015 @ 9:11PM | 27,034 views

Data Breaches In Healthcare Tot
112 Million Records In 2015

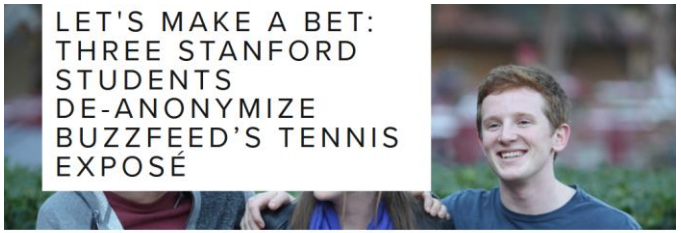
Disclosure: *Our family is one of the “tens of millions” of Americans potentially affected by the Anthem breach.*

Healthcare’s “wall of shame” for 2015 officially ends tonight at midnight. It’s not really a “wall,” it’s just a website, but it’s the online mechanism for the Office of

Anonymization and the Privacy/Utility Trade-Off

- In some combined data scenarios, anonymization is a solution, but it has **inherent limitations**

- The "**privacy-utility trade-off**":
  - Hard to properly anonymize data while preserving anonymity
  - In famous examples, supposedly anonymized datasets were de-anonymized
  - No way back when data is out in the open



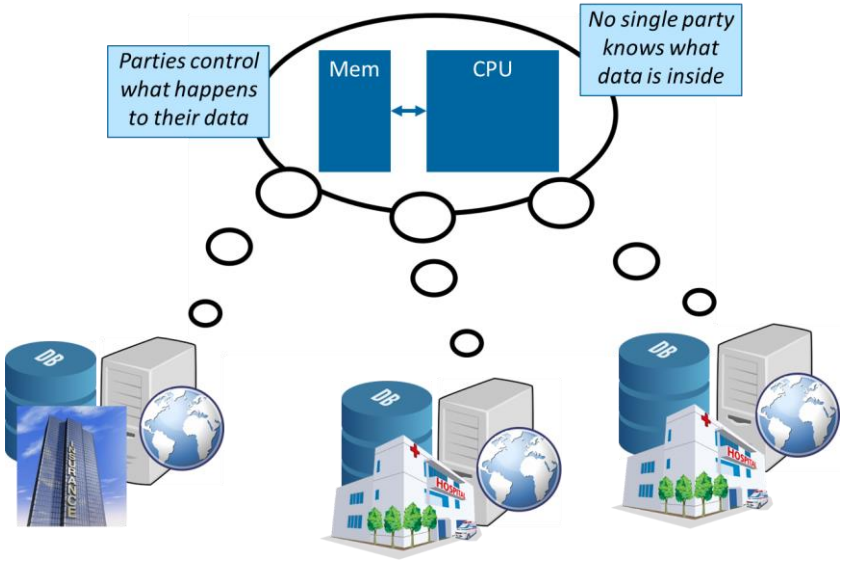
THE SCANDAL FIRST SURFACED when Buzzfeed released an article containing allegations against 16 unnamed professional tennis players for accepting money to manipulate their match outcomes. According to the story, more than half of the the players, whose names they refused to release, would be competing in the Australian Open in just a few days.

- Some types of analysis **cannot take place** on anonymized data, and some data simply **cannot be anonymized**
  - E.g., correlation mining between many sensitive attributes cannot be anonymous
  - E.g., any meaningful set of genomic data is already identifying

Multi-Party Computation: Beyond the Trade-Off

- Cryptographic techniques enable joint data analytics **without the need to share the underlying data**

- Main techniques:
  - Multi-Party Computation (MPC)
  - Fully Homomorphic Encryption (FHE)



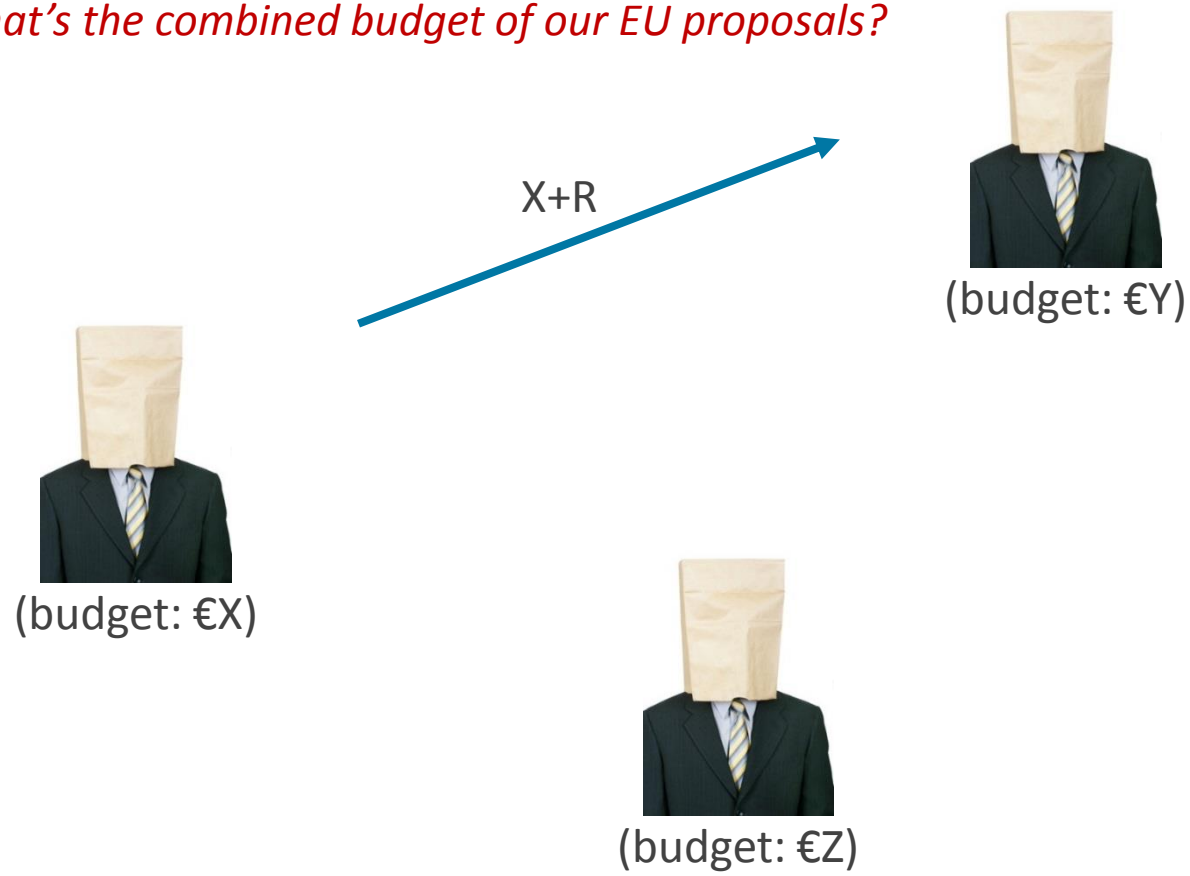
- Privacy guaranteed by encryption, but **without any loss in utility**

- MPC (picture right): each individual data item is stored and processed in a distributed way, data owners retain control of what happens

Multi-Party Computation: A Simple Demo

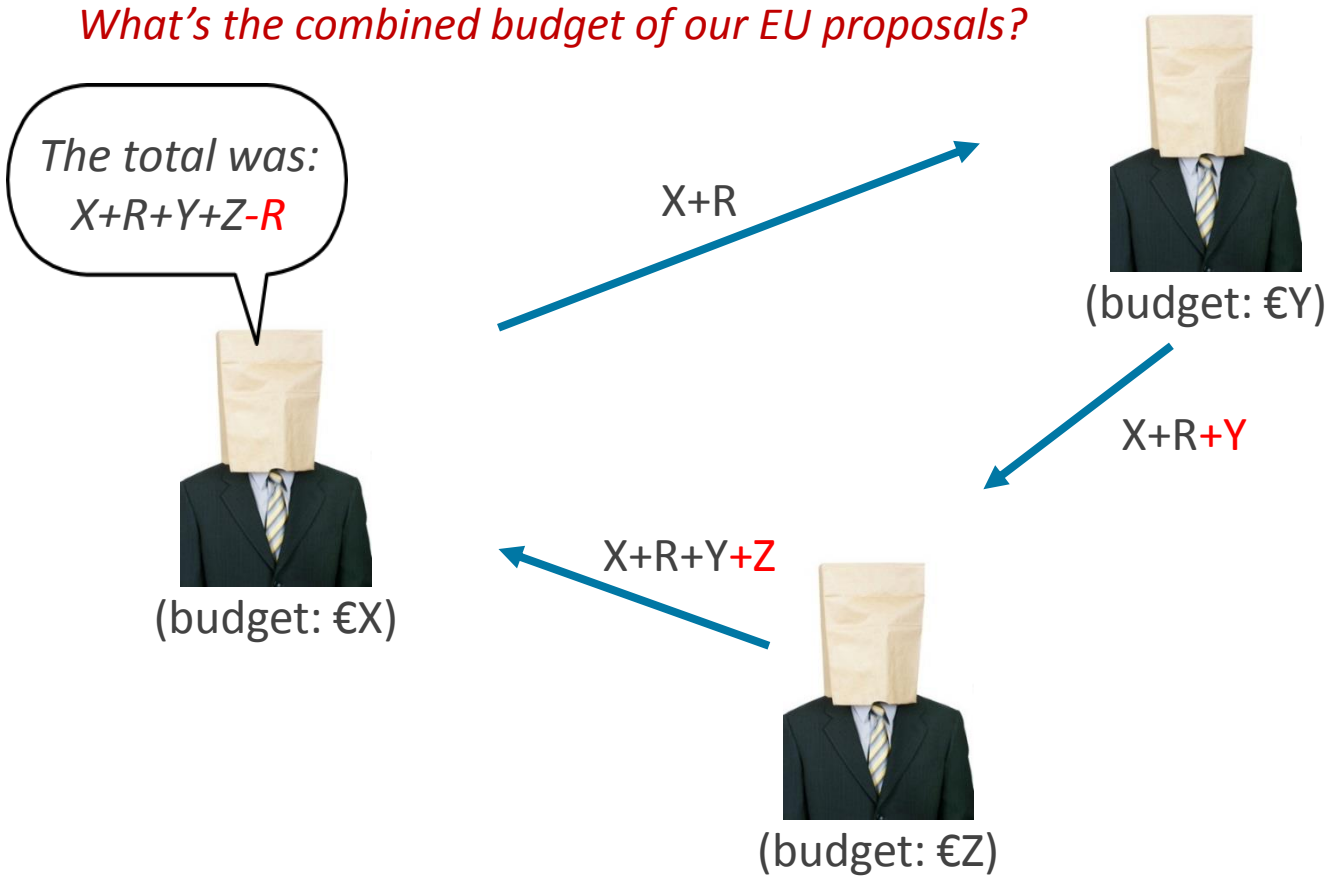
(Demo: Sebastiaan de Hoogh)

What's the combined budget of our EU proposals?



Multi-Party Computation: A Simple Demo

(Demo: Sebastiaan de Hoogh)



SODA: From Crypto Theory to Big Data Practice

- With MPC, we can perform any data mining algorithm on distributed data, while only revealing mining results to the intended parties

- However, challenges remain before MPC can be widely used:

1. **Performance** – MPC does not scale well to Big Data needs
   - Example: gradient descent matrix factorization: 11 days on 1M database (2015)
   - Example: statistics on two joined databases: 5 hours on 10M+600K records (2016)

2. **Disclosure control** – MPC controls *how* data is mined, not *what* is mined
   - Comprehensive solutions should combine MPC with, e.g., differential privacy

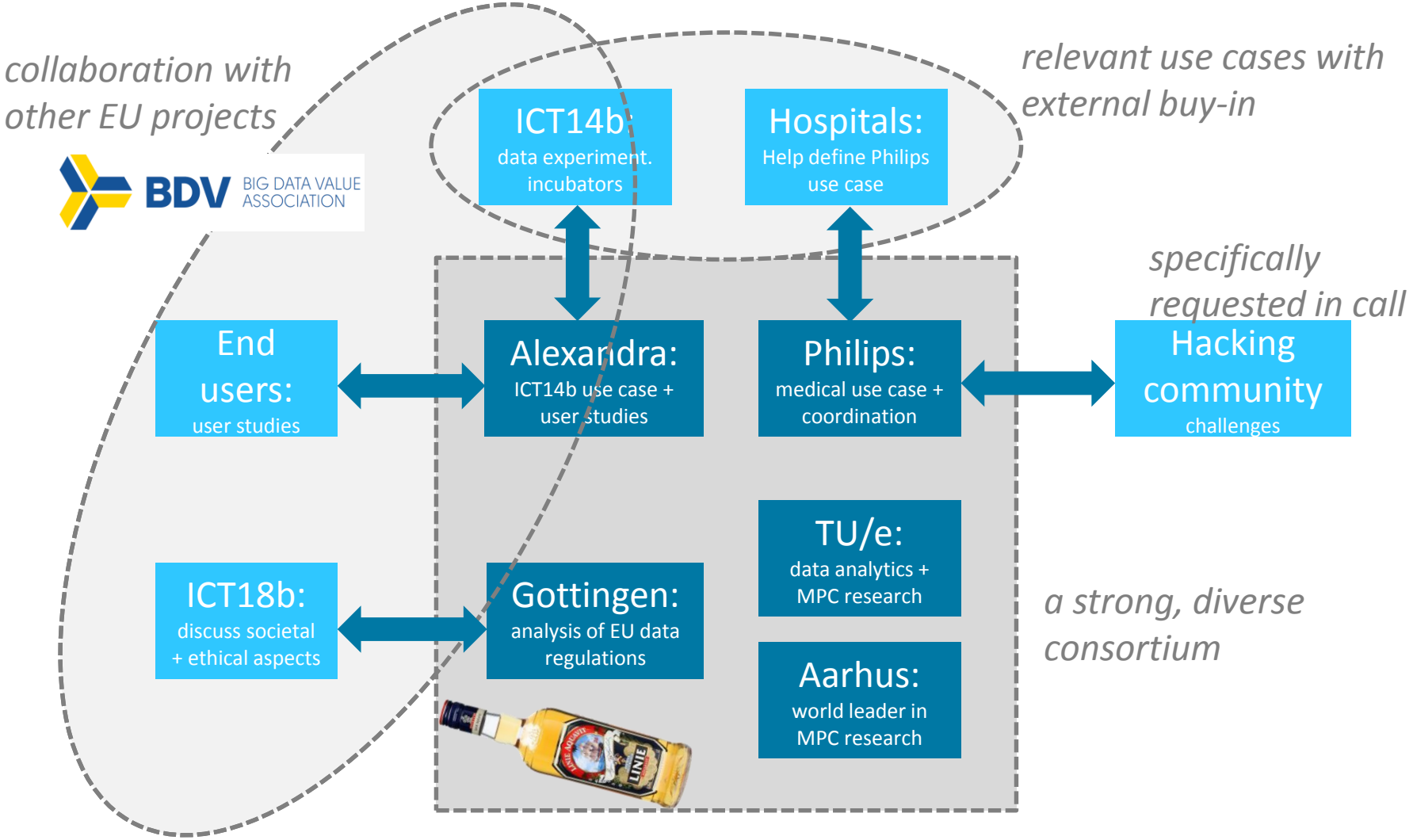
3. **Legal and social** – In a big system, unclear how data subject/owner is helped
   - End goals: increased willingness of data subject to enable analysis; increased compliance of data owner with EU data protection regulation

The SODA Mission

We will **enable practical privacy-preserving analytics on Big Data** by:

1. Making the performance of multi-party computation techniques for privacy-preserving shared Big Data processing **practical for real-world use cases**, moving **beyond the traditional privacy-utility trade-off**;
2. Ensuring that the performance improvements of our technical tools lead to real privacy improvements for relevant stakeholders, by:
   – Combining privacy-preserving processing with rigorous **bounds on the privacy leakage** of aggregated data analytics results;
   – Showing that our overall approach leads to **improved compliance with EU privacy law**;
   – Developing ways of explaining our techniques accessibly, hence **making data subjects more confident to allow processing** with our techniques.
3. Validating our approach in terms of functionality and security with **demonstrators** and **hacking challenges.**

SODA: A Collaborative Effort



9

Conclusions

- SODA is a collaborative effort to **enable MPC-based practical privacy-preserving analytics on Big Data,** moving **beyond the privacy/utility trade-off**
- The project runs from January 2017 to December 2019
- We are looking forward to collaborate with other projects (e.g., ICT14b, ICT18b) projects to place our technical contributions in a broader context
- Project leader: Meilof Veeningen (meilof.veeningen@philips.com)

any questions?


LINIE AQUAVIT
LINIE