Privacy by Design in practice

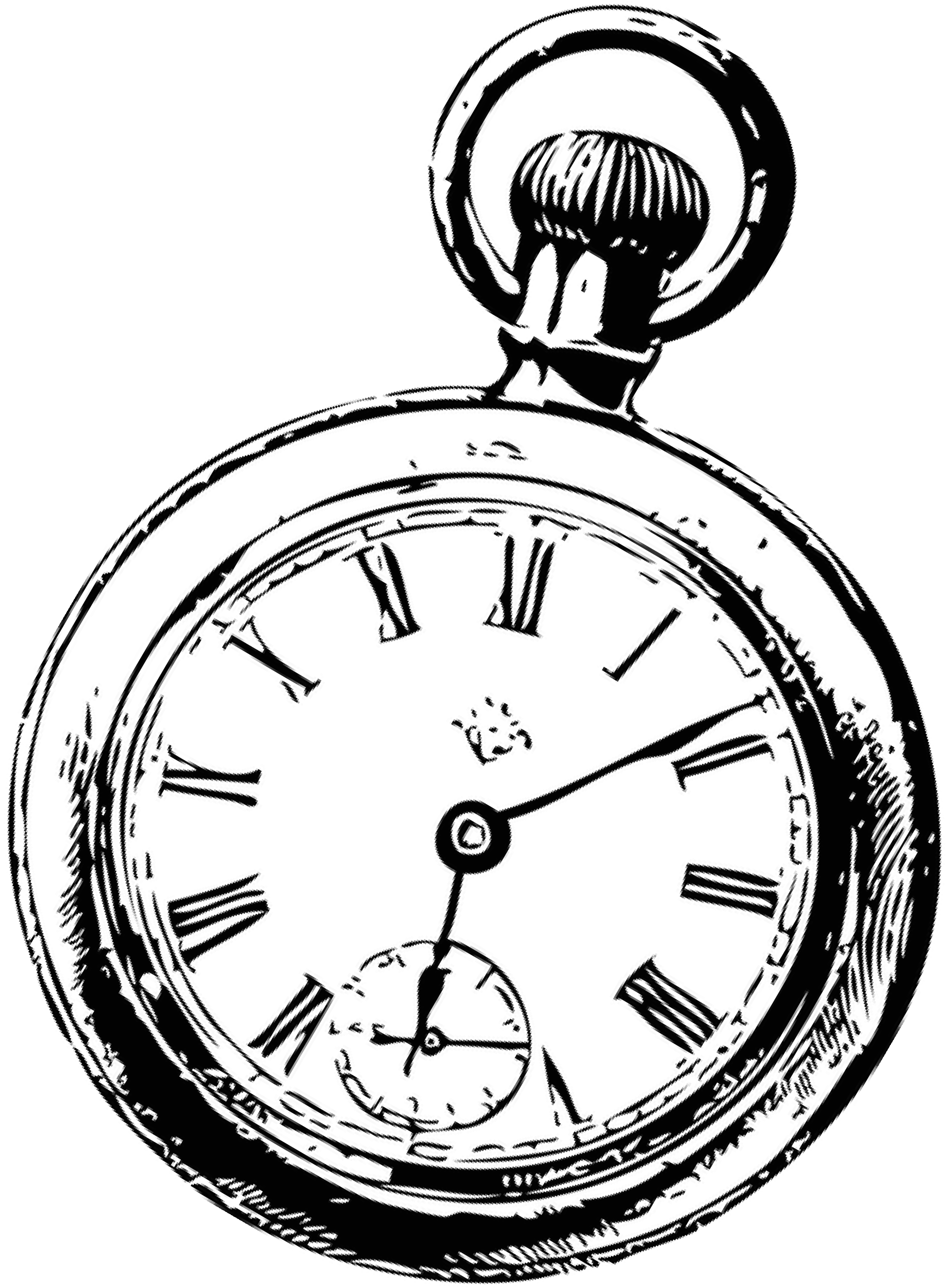
using polymorhic encryption and pseudonymisation in medical research

Jean Popma
Digital Security
Institute of Computing and Information Sciences

J.Popma@ru.nl 31-05-2017

Today’s Menu

- Privacy & Security : where am I?
- Privacy by Design in theory
- And in practice…
- Personalized Parkinson Project – a pilot, or maybe not.
- Polymorphic Encryption en Pseudonymisation
- The proof of the pudding…..



Medical research & big data … What’s the big deal?

- Medical data are the most sensitive personal data. Data breach: reputation damage for researchers
  - harmful to participants/patients and
  - undermining patient’s willingness to particpate in future research

- Legal requirements are strict
  - More so after GDPR becomes fully operational (May 2018)
  - high fines / repercussions

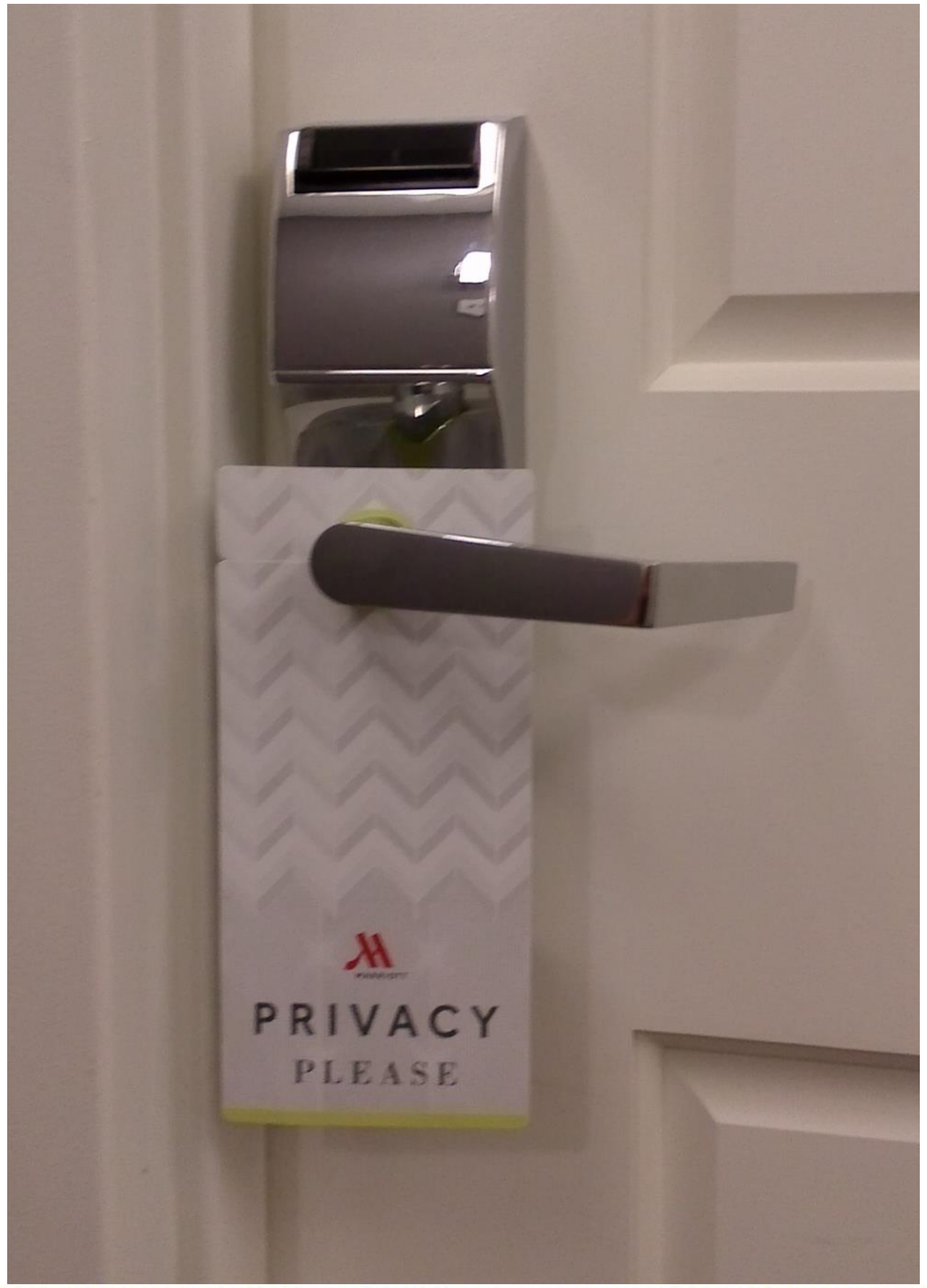
- Loss of confidence is a show-stopper for medical research

- Professional cooperation between computer scientists and medical scientists is essential; amateuristic approaches are no longer acceptable

… Because the law requires it ….

- Well Defined Purpose
- Proportionality: data minimalisation and data retention
- Transparancy
- Security
- Privacy & Security *by design* en *by default*

.. But also because it is really important ….



Intermezzo: Privacy has significance in context (after Helen Nissenbaum)

- Privacy is not an absolute concept , is not the same as secrecy, not exclusively related to data protection, and is not just a personal thing.

- We live in a natural way in different contexts

- We keep information in its context

- Breaking contextual integrity is a shock

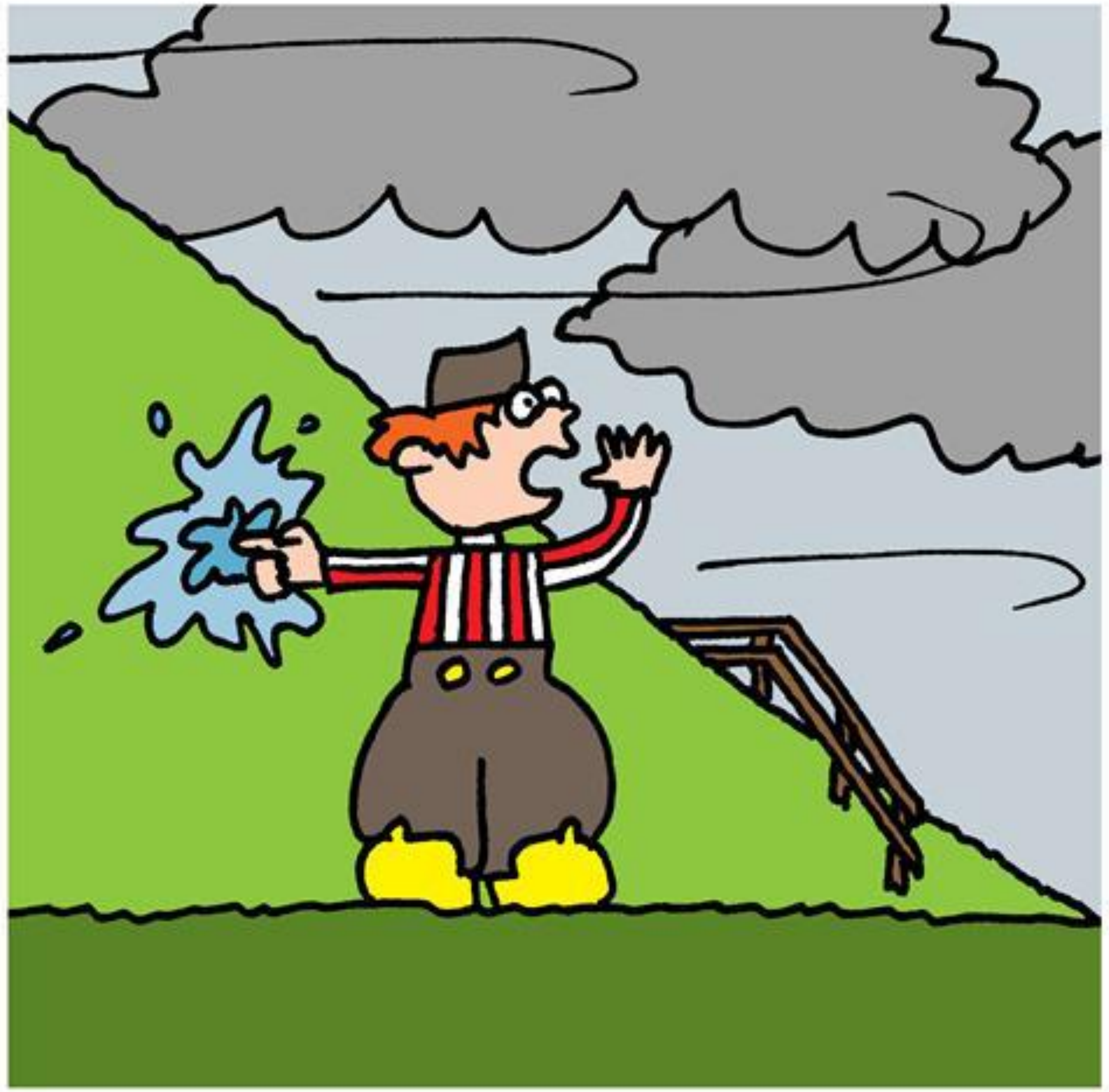
- The Googles and Facebooks in this world want to be able to trace us wherever we go, and whatever we do. They make money from breaking our contextual integrity

Mark Zuckerberg: *Having two identities for yourself is a lack of integrity* 👎👎👎

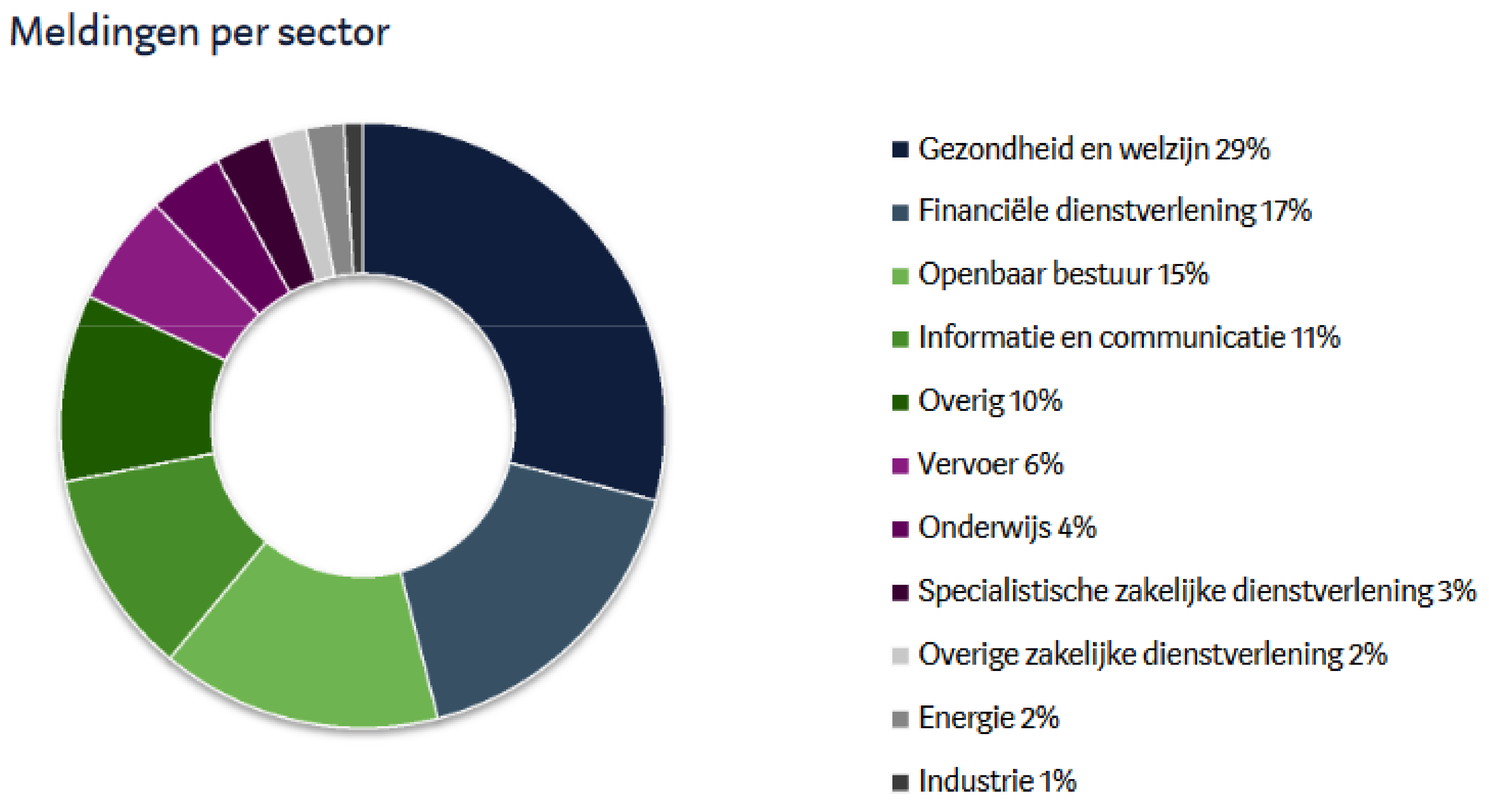
Privacy: legislation

- Constitution (art 10 t/m 13)
- 1989-2001: Wpr, Personal Registrations Act
- 1995 European Data Protection Directive
- 2001-nu: Wbp Personal Data Protection Act
- 2016 : Wbp completed with a mandatory data breach notification (meldplicht datalekken) and regulatory fines (up to 800 k€)
- 2018 (2016): GDPR, General Data Protection Regulation

Data Breach



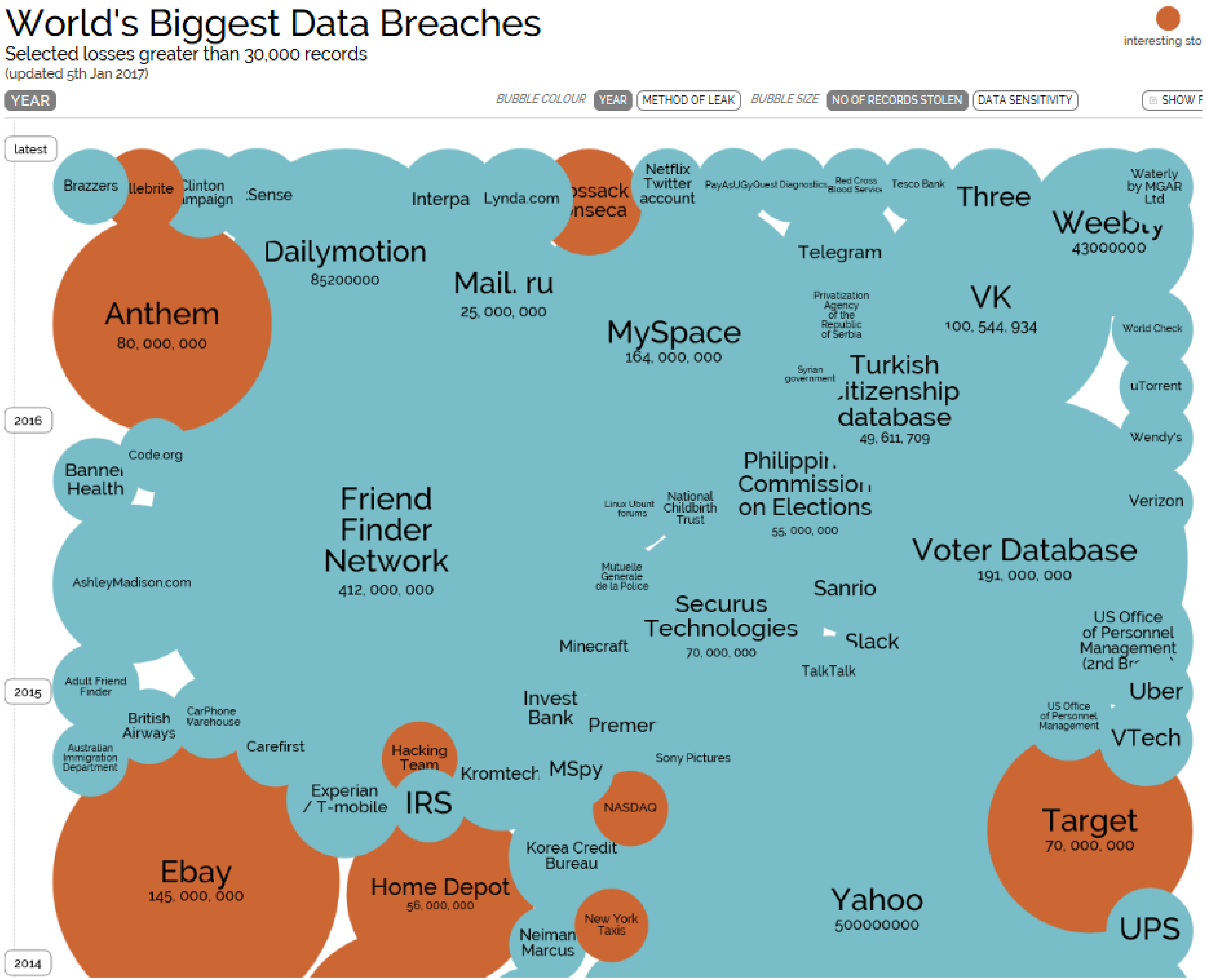
2016: 5500 Notifications to Personal Data Authority (PDA)



- about 1/3 of data breach incidents reported internally reported to PDA
- Effectiveness of detection of data breaches (internally)
- Tip of the Iceberg?

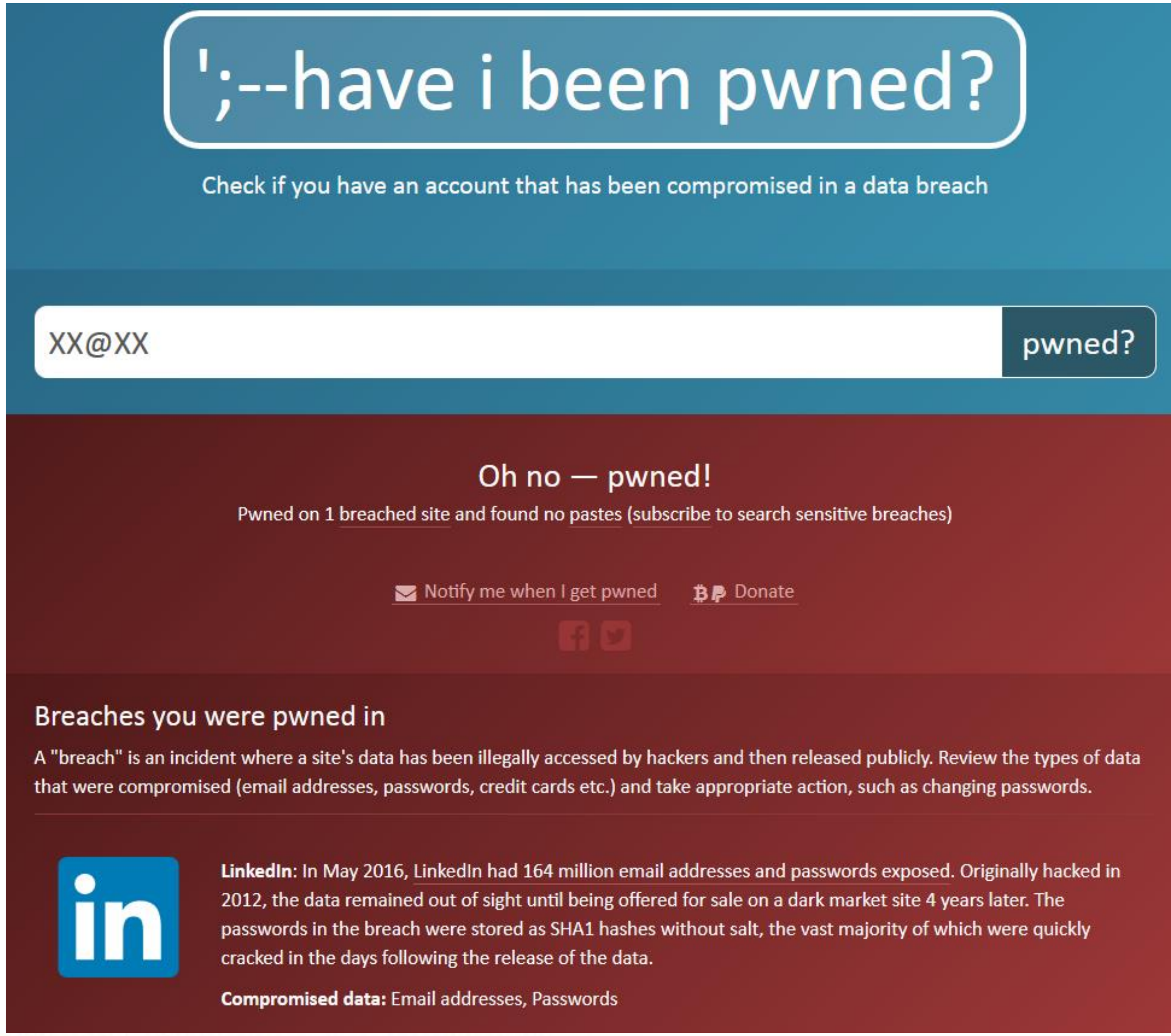
Source: Autoriteit Persoonsgegevens

… all quiet on the western front …



http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

… all quiet on the western front …



';--have i been pwned?

Check if you have an account that has been compromised in a data breach

XX@XX | pwned?

Oh no — pwned!

Pwned on 1 breached site and found no pastes (subscribe to search sensitive breaches)

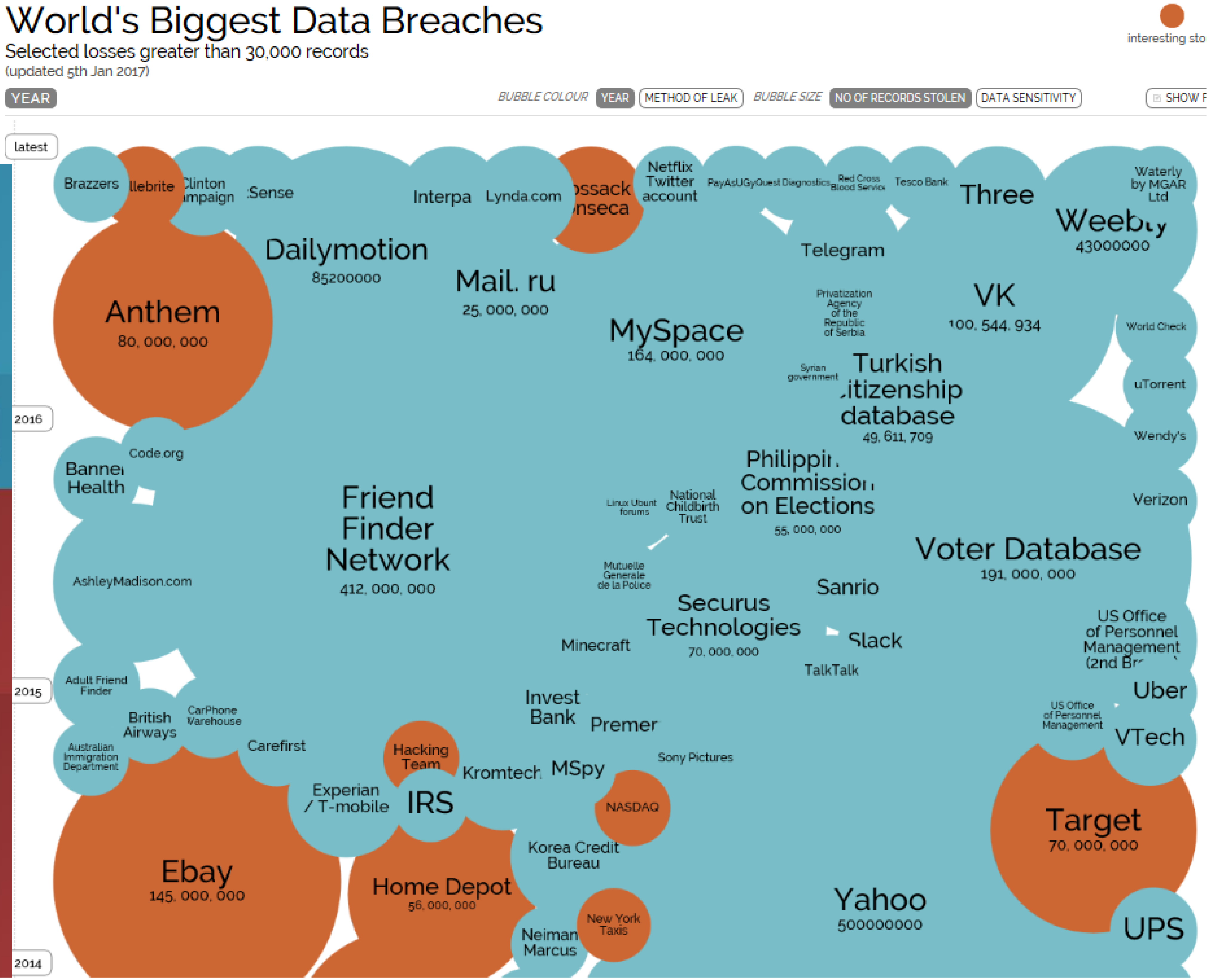
✉ Notify me when I get pwned   ₿ ₽ Donate

Breaches you were pwned in

A "breach" is an incident where a site's data has been illegally accessed by hackers and then released publicly. Review the types of data that were compromised (email addresses, passwords, credit cards etc.) and take appropriate action, such as changing passwords.

LinkedIn: In May 2016, LinkedIn had 164 million email addresses and passwords exposed. Originally hacked in 2012, the data remained out of sight until being offered for sale on a dark market site 4 years later. The passwords in the breach were stored as SHA1 hashes without salt, the vast majority of which were quickly cracked in the days following the release of the data.

Compromised data: Email addresses, Passwords



https://haveibeenpwned.com/

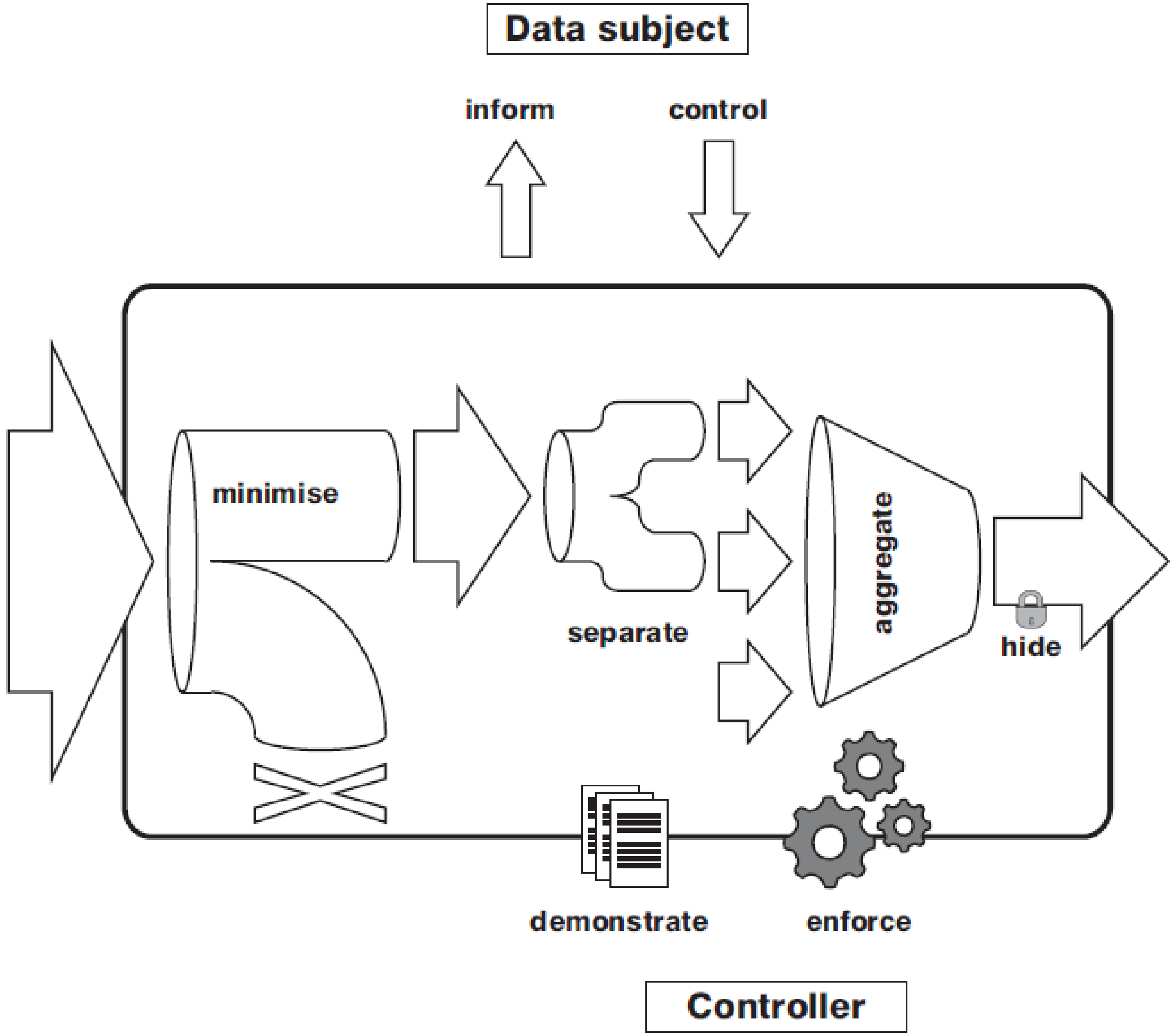
http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

Data Breach – As-A-Symptom

- Authorizations to broad
- Orphaned data
- Stolen or lost devices
- Copies traveling around
- Combination of different sources
- Identity theft
- Ransomware
- Lack of maintenance
- Hidden tracking/profiling
- Too many data collected
- Data reused for other purposes
- Fraud
- Revenge
- Spionage
- Wrongly addressed mails….
- Forgotten print output
- Active hackers
- Stupidity

Privacy By Design?

- Data Design Strategies
  - Minimise
  - Separate
  - Aggregate
  - Hide
- Proces Design Strategies
  - Inform
  - Control
  - Enforce
  - Demonstrate



Bron: Jaap-Henk Hoepman @ https://arxiv.org/pdf/1210.6621.pdf

Personalized Parkinson's Project aims at profiling patients at an early stage of the disease in order to provide more effective personalized care and treatment.

Personalised Parkinson's Project

- 650 participants suffering from Parkinson's
- 2 year study
- Data will be available to scientist at dutch UMC's and other research institutes (worldwide)

  - Clinical data (questionnaires, tests
  - Biochemical data (from blood, plasma, CSF)
  - Wearable data (2 years)
  - ECG
  - fMRI
  - Genome
  - Biome

Study Total: > 1 PB of data

Radboudumc
university medical center

parkinson op maat

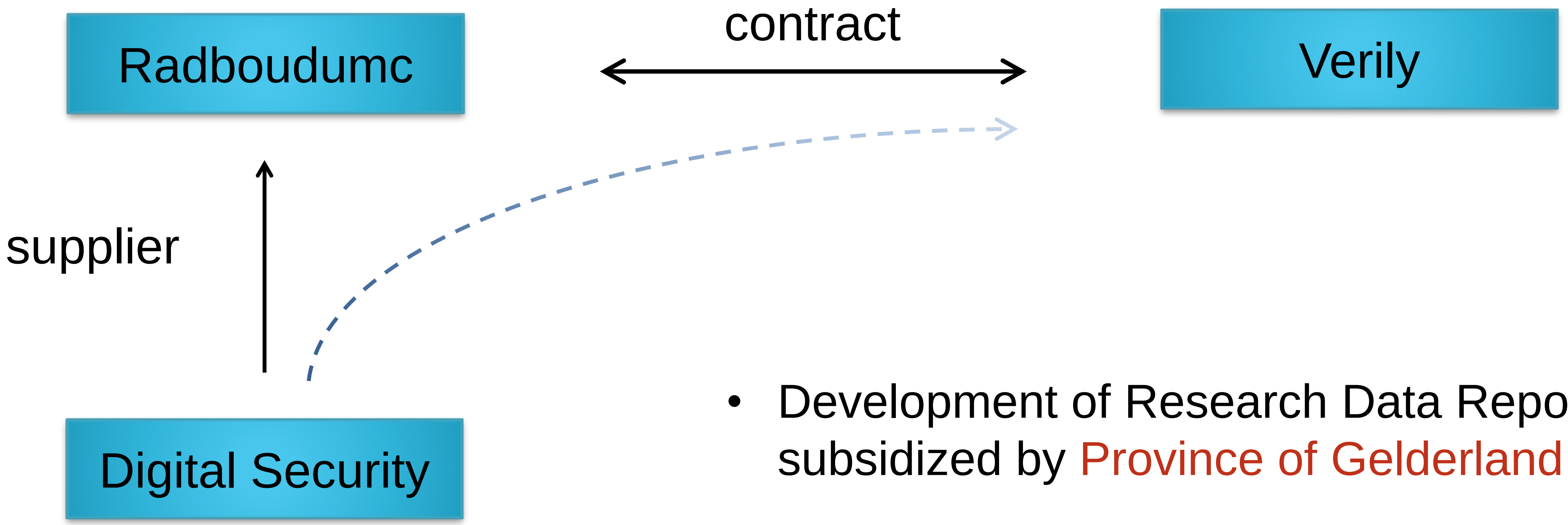
Personalized Parkinson's Project Radboudumc

Radboudumc
university medical center

Research project of Radboudumc, largely sponsored by Verily (also actively contributing to analysis and research . Verily is a medical research company, a subsidiary of Alphabet).

Research data are stored in a Research Data Repository and are protected and managed by PEP technology, developed by the Digital Security department of Radboud University.

Position Digital Security group



Radboudumc

contract

Verily

supplier

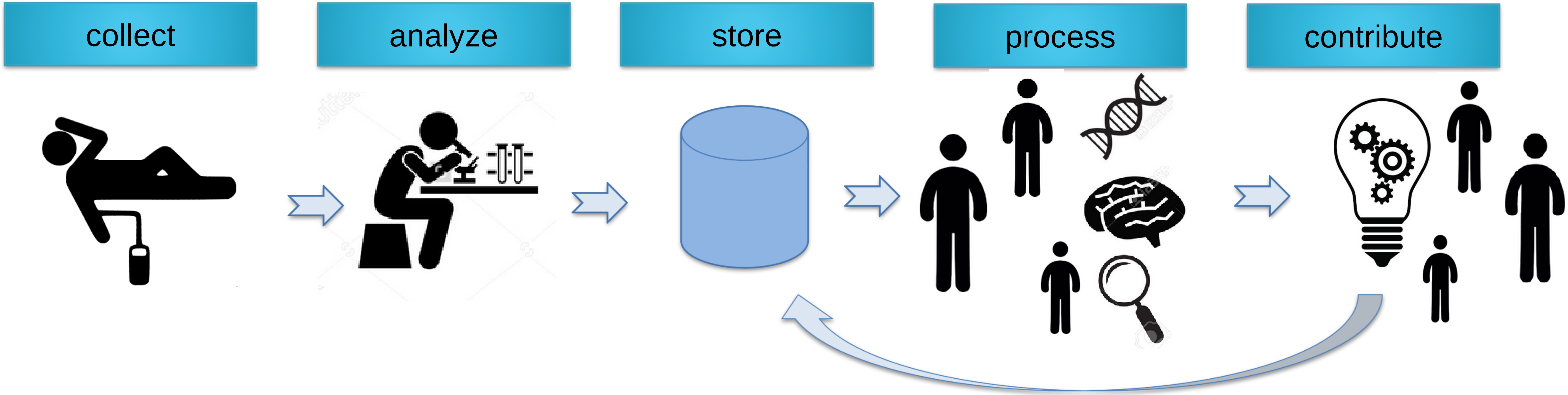
Digital Security

- Development of Research Data Repository based on PEP subsidized by Province of Gelderland: 3/4M€.

- Enabling autonomy and independence from Alphabet/Verily

Research Data Repository

2 main functions : **share** data, and ensuring **scientific integrity**



collect

analyze

store

process

contribute

Challenges

- Medical research data are very sensitive personal data

  Challenge 1:
  Protect the privacy and personal data of participants

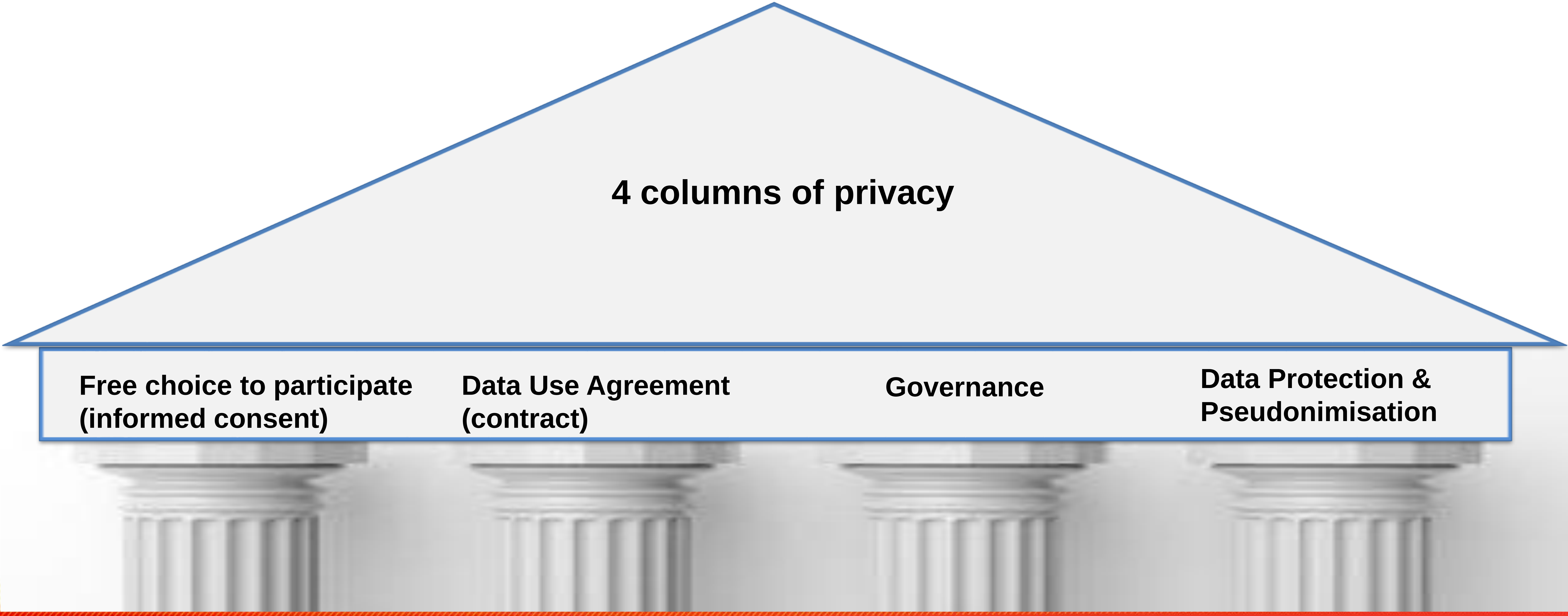
  Challenge 2:
  Enable the use of data collected for legitimate research (effectiveness)

  Challenge 3:
  All data sources have different architectures, procedures and use restrictions

- Legal requirements are clear

  Purpose is to keep data in a predefined context, in a transparent
  Purpose is to prevent illegitimate use of the data.

Protection starts with organisation, policies and agreements



4 columns of privacy

Free choice to participate (informed consent)

Data Use Agreement (contract)

Governance

Data Protection & Pseudonimisation

PEP: Data Protection & Pseudonymisation

- PEP = Polymorphic Encryption en Pseudonymisation (http://pep.cs.ru.nl)

- Innovative encryption method developed at Radboud University by prof. Eric Verheul, prof. Bart Jacobs and the PEP team.

- Very suitable for use in large scale medical research
  - Data from may different sources: the patient, hospital labs, partner labs etc.
  - All data are stored encrypted in the data-repository
  - Encryption takes place as close as possible to the source of the data.
  - Encryption and Pseudonymisation can be applied in a very flexible way in a research.

- PEP infrastructure takes care of the data management

PEP use

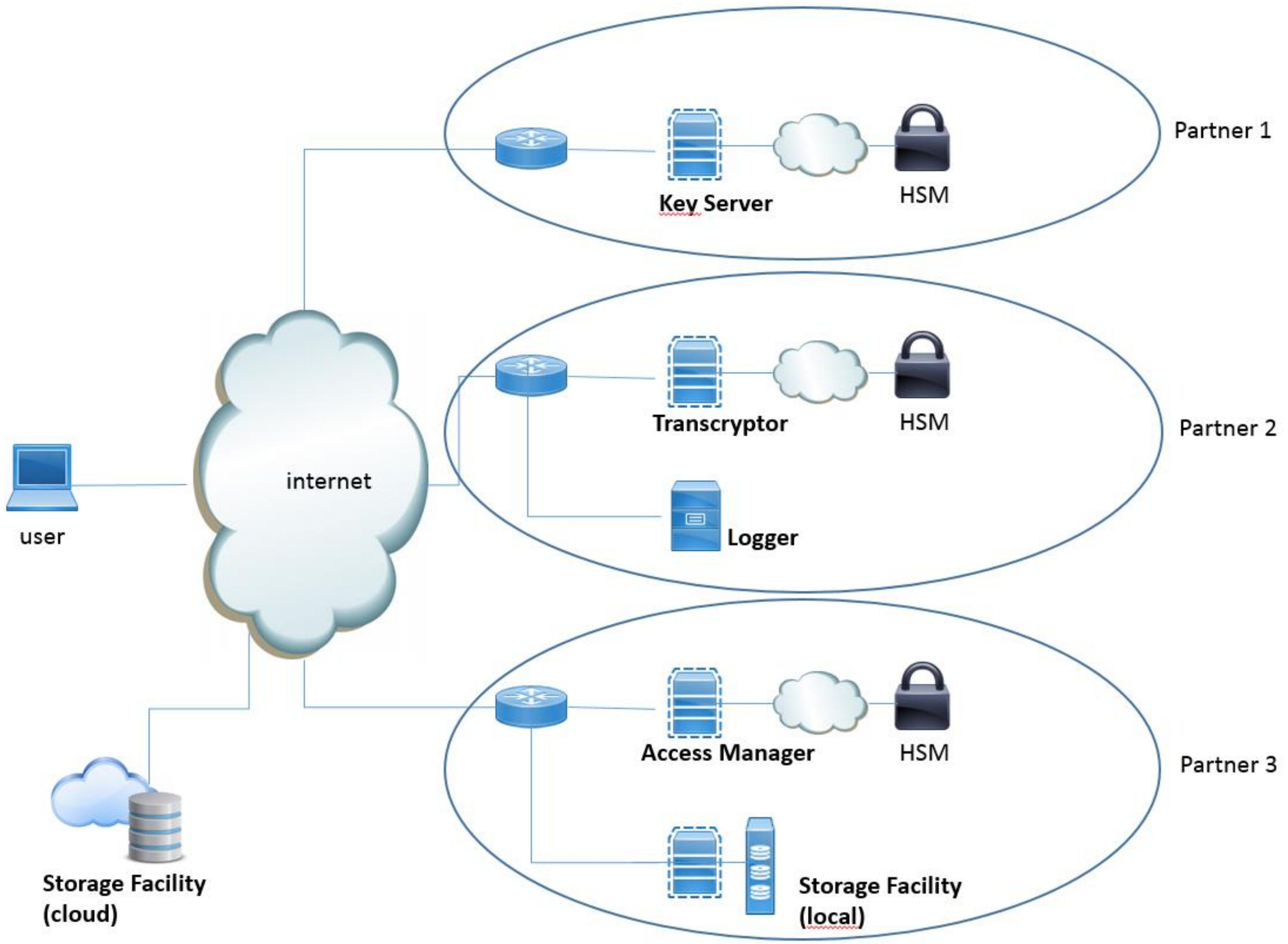
- Encryption close to the source (future: within the source)
- Upload of data to the repository in encrypted form only
- Download from repository in encrypted form only
- Use of repository data bound to restrictions (security and data-handling must comply with legal and technical requirements)
- After data analysis data can be deleted in the processing environment (historical queries will be possible during the entire project including its archival phase))
- Multiple parties involved (no single systems- or data management party can decrypt data)
- Independent logging and audit possible.

Waar does PEP stand for?

- Traditional Encryption:
  - shared secret (one key)
  - or public/private key (at the time of encryption you must know who is allowed to decrypt)

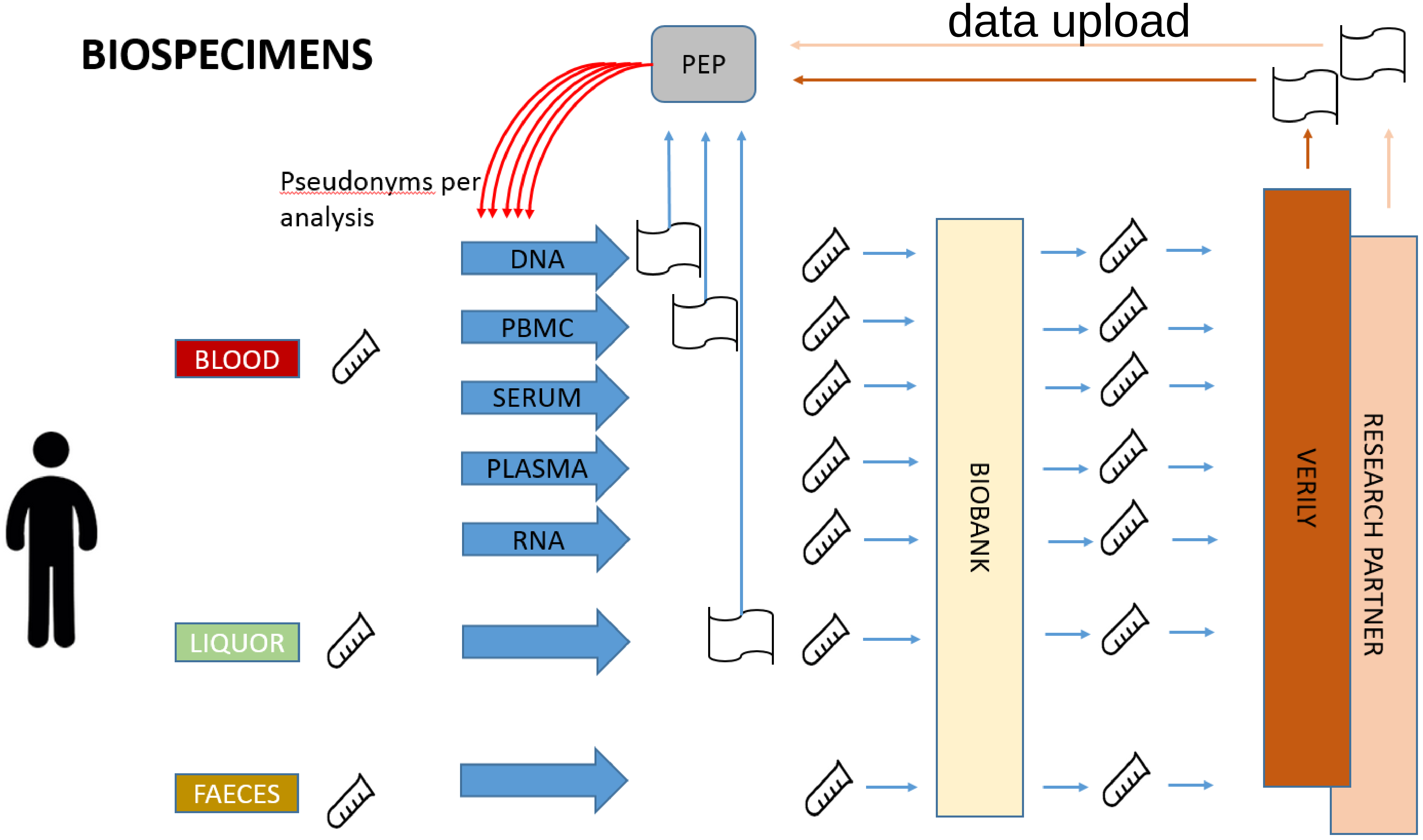
- PEP:
  - Polymorphic encryption close to the data source
  - Unique keys for different data and different users can be generated a posteriori, at the time access is granted.
  - All data streams are based on unique pseudonyms
  - Unique Pseudonyms are generated for each user (user group) obtaining data from the repository.

Distributed management of cryptographic keys



- User has client Software (encryption/decryption, up/download)
- 3 trusted parties take care of key management. Developers have no knowledge of keys.
- Keys in Hardware Security Modules (tamperproof)
- Storage Facility can be anything (public cloud)

Pseudonymisation during data collection phase



- All datastreams based on unique pseudonyms
- At the time of upload these pseudonyms are translated to a polymorphic pseudonym, so data from different sources can be linked.
- Same idea for data from other sources (wearables, MRI etc.)

Pseudonymisation for data use

- Research project requests data of certain types regarding subjects that meet certain criteria
- Request is evaluated by a committee
- In case of positive decision:
  - Data type I, III, IX of subjects PPseu3 PPseu9 PPseu125 …. PPseuN are authorized to User
  - User requests data from PEP
- User gets:
  - Access (download) to data where
  - Pseudonyms are personalized to user,
  - A unique key is constructed to decrypt these data
- In the user's authorized processing environment
  - Data are decrypted for further processing and analysis
  - Derived data can be encrypted and uploaded back to the repository

Privacy By Design revisited

- Data Design Strategies
  - Minimise √
  - Separate √
  - Aggregate X
  - Hide √
- Proces Design Strategies
  - Inform √
  - Control √
  - Enforce √
  - Demonstrate √

Bron: Jaap-Henk Hoepman @ https://arxiv.org/pdf/1210.621.pdf

Risks Mitigated?

- Authorizations to broad
- Orphaned data
- Stolen or lost devices
- Copies traveling around
- Combination of different sources
- Identity theft
- Ransomware
- Lack of maintenance
- Hidden tracking/profiling

- Too many data collected
- Data reused for other purposes
- Fraud
- Revenge
- Spionage
- Wrongly addressed mails….
- Forgotten print output
- Active hackers
- Stupidity

Risks Mitigated?

- Authorizations to broad
- Orphaned data
- Stolen or lost devices
- Copies traveling around
- Combination of different sources
- Identity theft
- Ransomware
- Lack of maintenance
- Hidden tracking/profiling

- Too many data collected
- Data re-used for other purposes
- Fraud
- Revenge
- Spionage
- Wrongly addressed mails….
- Forgotten print output
- Active hackers
- Stupidity

Risks Mitigated?

- Authorizations to broad
- Orphaned data
- Stolen or lost devices
- Copies traveling around
- Combination of different sources
- Identity theft
- Ransomware
- Lack of maintenance
- Hidden tracking/profiling

- Too many data collected
- Data re-used for other purposes
- Fraud
- Revenge
- Spionage
- Wrongly addressed mails….
- Forgotten print output
- Active hackers
- Stupidity

Risks Mitigated?

- Authorizations to broad
- Orphaned data
- Stolen or lost devices
- Copies traveling around
- Combination of different sources
- Identity theft
- Ransomware
- Lack of maintenance
- Hidden tracking/profiling
- Too many data collected
- Data re-used for other purposes
- Fraud
- Revenge
- Spionage
- Wrongly addressed mails….
- Forgotten print output
- Active hackers
- Stupidity

Summary/Results

- PEP is conditional for privacy-friendly data processing
  - PEP approach is **state-of-the-art** cryptography
  - based on: **security-by-design** en **privacy-by-design**
  - Design and software will be published as **open source**

- **integration** in medical context is a big challenge

- But risks are not reduced to 0 …
  - Data can be self-identifying (video/photo material, genome data etc.)
  - Conspiring scientists can link data based on content.

- First production version expected for July 2017.   In Q1 2018 full functionality available

Conclusions

- Strong security & privacy protection are a *licence to operate* in medical (big data) research

- This Parkinson’s study aims at a breakthrough, both on medical science and computer science challenges.
  - Unique cooperation
  - High impact

- Big Data, Cloud & Privacy can be combined if Privacy & Security are part of the design.

- With this pilot the technology can be put to the test and new areas of use can be explored.

More? http://pep.cs.ru.nl