École Pour l'Informatique et les Techniques Avancées – EPITA

Masters program – 02 April 2022

Course: Data Privacy by Design

EPITA
ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

Course instructor: M Salman Nadeem
mohammad-salman.nadeem@epita.fr

Data Privacy by Design (PbD)

Course schedule (tentative)

Date & Time	No.	Topics	Duration (in hours)
04/03/2022 14:30-17:30	1	Data & its types, Information & knowledge, Introduction to Data Privacy by Design (PbD)	3 hours
18/03/2022 14:30-17:30	2	DPbd Case studies, Data privacy risks & solutions	3 hours
02/04/2022 10:00-13:00	**3**	**Privacy Enhancing Technologies (PET's)**	**3 hours**
22/04/2022 14:30-17:30	4	General Data Protection Regulation (GDPR), PbD and GDPR	3 hours
29/04/2022 14:30-17:30	5	Open session, Putting it all together, Quiz, Final project presentation	3 hours
		Total Lecture (hours)	***15***

Evaluation: 10% Class attendance + 10% Class participation + 30% Class/home exercises + 50% Final Evaluation

Lecture 3 Outline

- Privacy Enhancing Technologies (PETs)
  - Data Anonymization techniques
  - Differential privacy
  - K-anonymity
  - Tor/Panoramix
  - Systematic approaches
  - General Security controls
- Class exercise 5

Data Anonymization techniques

- It is difficult!
  - One data anonymization company, Aircloak, even acknowledges that true anonymization is extremely difficult: "as is the case with IT security, no 100% guarantee can be given, and often there is the **need for a risk assessment**"
- Gazillion Anonymization techniques:
  - Often embodied as "Privacy Enhancing Technologies" (PETs):
    - Soft: 3rd parties can be trusted for data processing (through compliance control and audit), example technologies: differential privacy, SSL, etc
    - Hard: 3rd parties cannot be trusted, example technologies: onion routing, secret ballot, etc
- When is data considered as anonymized?

Per European Data Protection Board (EDPB) guidelines, when it not possible to:

1. Single out an individual from a larger group
2. Link different records related to the same individual
3. Infer unknown information about an individual

Source: https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_20200420_contact_tracing_covid_with_annex_en.pdf

Differential privacy

- Noise addition using a single value: epsilon (ϵ), which is a measure of how private a data release (output) is
  - Higher values of ϵ gives accurate, less private answers
  - low-ϵ systems give highly random answers
- The outcome of any analysis on output dataset is essentially equally likely, independent of whether any individual joins, or refrains from joining, the input dataset
  - Used by: Apple, Microsoft, Google, Uber …

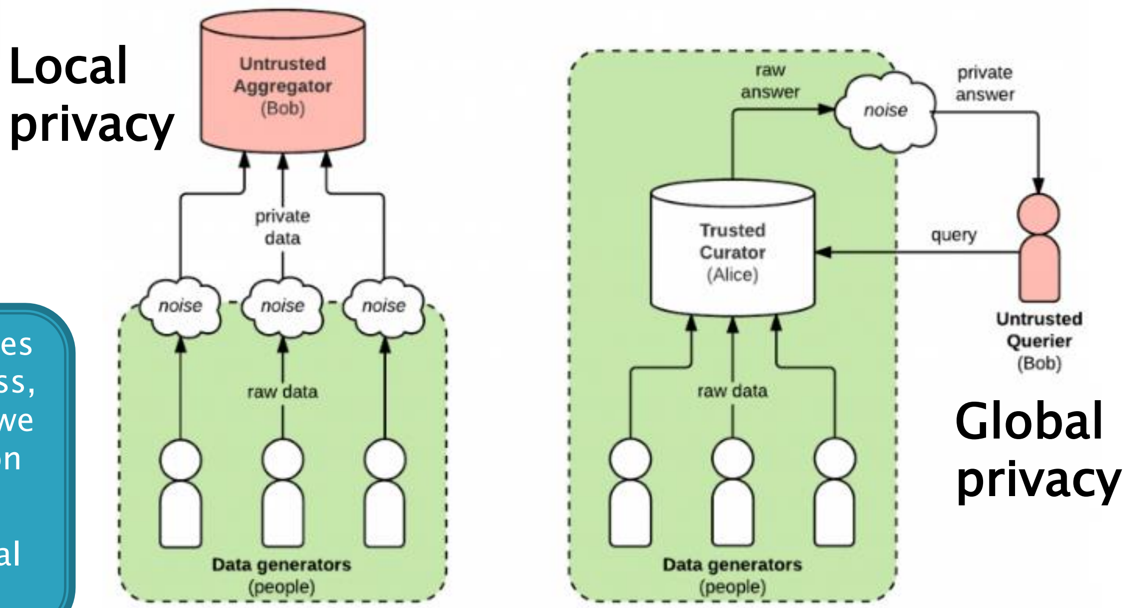
$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{A}(D_2) \in S],$$

Two data sets: $\boldsymbol{D_1}$, $\boldsymbol{D_2}$
Randomized algorithm: $\boldsymbol{A}$
All events/subsets: $\boldsymbol{S}$

The **algorithm** $\boldsymbol{A}$ is said to provide e-differential privacy, for all datasets (D_1, D_2), that differ on a single element (i.e., the data of one person)…

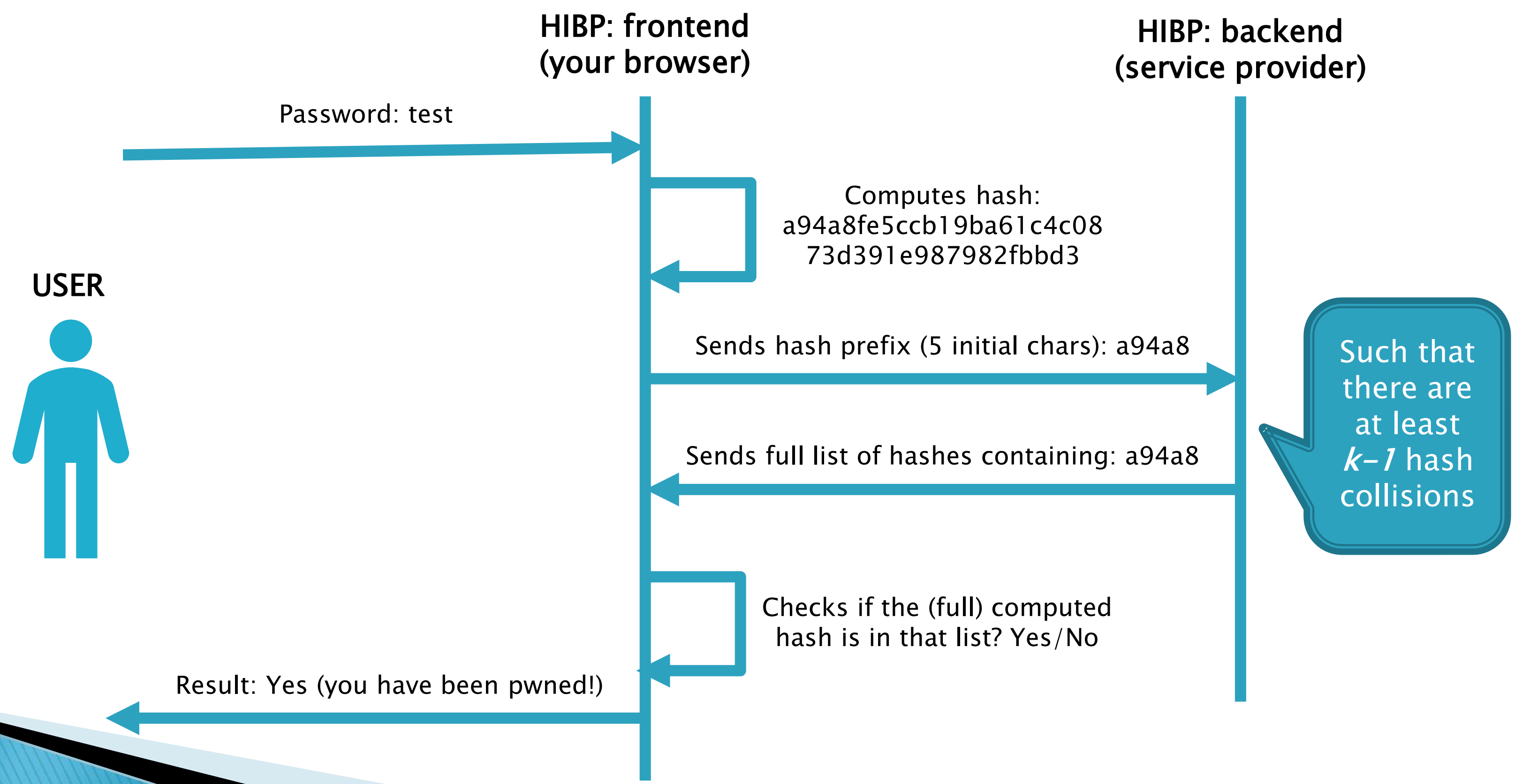
A introduces randomness, such that we get epsilon

(ϵ) differential privacy

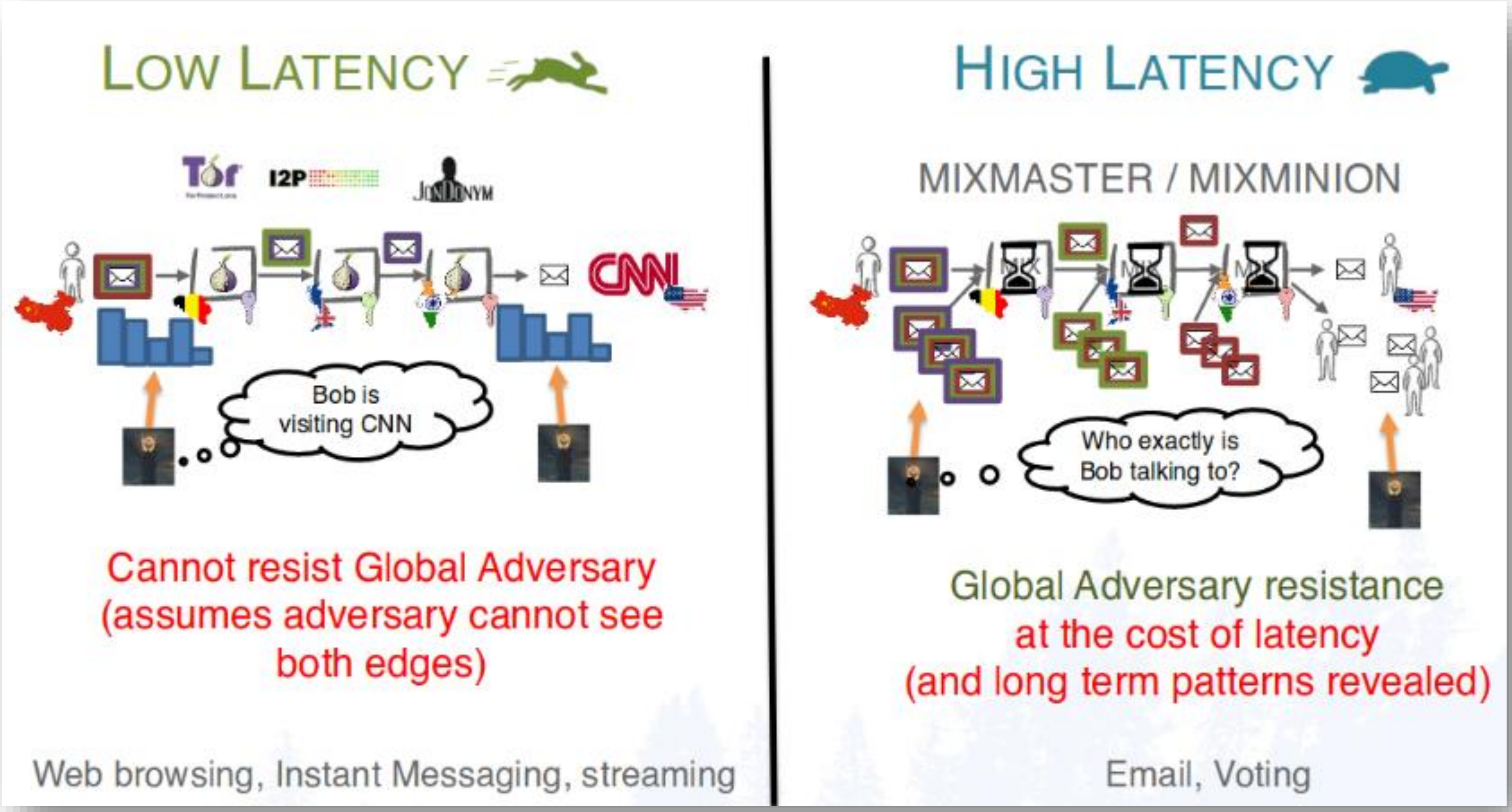


K-anonymity (range queries)

- If atleast ‘k’ individuals share same quasi-identifier(s) in the same data set, then no individual can be uniquely traced
- E.g., HIBP (https://haveibeenpwned.com/Passwords) should not know your password in order to be able to tell if it was breached



Tor/Panoramix



Other examples: I2P, freenet

Not all techniques works for all cases! (1/3)

- Netflix [Competition 'Prize' (2006)]
  - Competing teams had to create an algorithm to predict user ratings for films
  - Provided dataset included ~100M ratings, ~480k users for ~17k movies
    - Anonymization:
      - Replaced name of users with random chars
      - Replaced random ratings with fake one's

How To Break Anonymity of the Netflix Prize Dataset

Arvind Narayanan, Vitaly Shmatikov

(Submitted on 18 Oct 2006 (v1), last revised 22 Nov 2007 (this version, v2))

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.
We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Subjects: **Cryptography and Security (cs.CR)**; Databases (cs.DB)
Cite as: **arXiv:cs/0610105 [cs.CR]**
(or **arXiv:cs/0610105v2 [cs.CR]** for this version)

Bibliographic data

[Enable Bibex (What is Bibex?)]

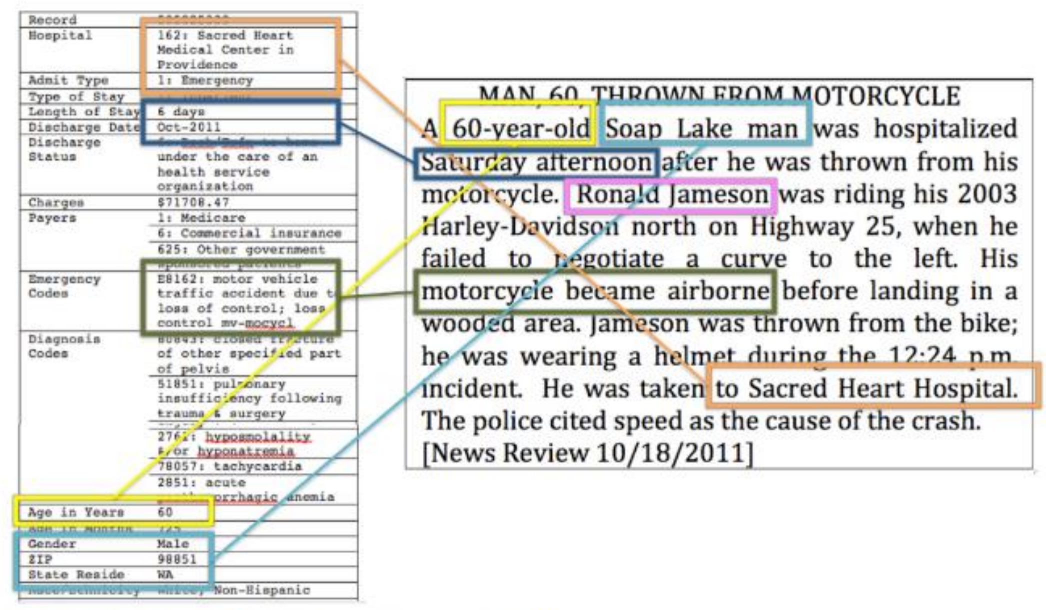
Submission history

From: Vitaly Shmatikov [view email]
[v1] Wed, 18 Oct 2006 06:03:41 UTC (128 KB)
[v2] Thu, 22 Nov 2007 05:13:06 UTC (313 KB)

2007 -> Researchers successfully denonymized the Netflix dataset by combining it with the data of IMDB (Linkage attack)

Not all techniques works for all cases! (2/3)

- Another example of re-identification from the Journal of Technology Science that
  - An “anonymous” medical record is cross-referenced with a newspaper brief about a motorcycle crash
  - Patient in question is identified



Matching public medical information to news stories to identify patients.

Ref. https://techscience.org/a/2015092903/

Not all techniques works for all cases! (3/3)

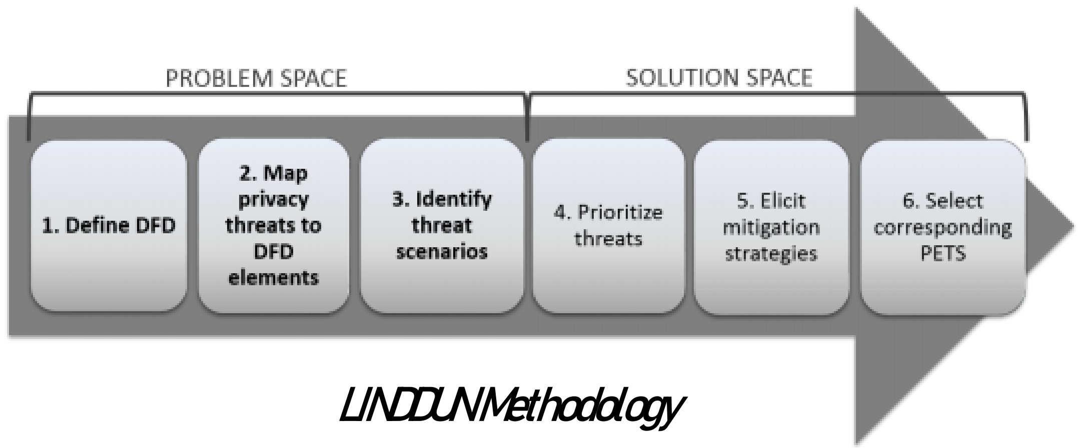
- Many possible attacks exist!
  - Background information attack
  - Unsorted matching attack
  - Complementary release attack
  - Temporal attack
  - …

Carry out independent audits/reviews to ensure that the anonymized data-set is not vulnerable to de-anonymization attacks!

Systematic approaches

STRIPED	Spoofing	Tampering	Repudiation	Information disclosure	Privacy	Elevation of privilege	Denial of service
LINDDUN	Linkability	Identifiability	Non-repudiation	Detectability	Disclosure of information	Unawareness	Non-compliance

Scientific renown
Industry acceptance:
(ISO 27550, EDPS PbD opinion, ENISA PbD)



LINDDUN Methodology

Other factors: Data lifecycle, maintenance, …
Brainstorm sessions & ad hoc basis…
Do what is feasible for your team!

General security controls

- Unique and random passwords of all administrative, and other sensitive channels!
- Use of suitable crypto. mechanisms on all appropriate levels
  - Including the use of Anonymization/Pseudonymisation techniques
- Intrusion detection & prevention systems (SIEM, …)
- User access control: ACL's, RBAC, … (both in-house, and for end-users)
- Secure data backup strategy
- Properly configured Firewalls, Real-time monitoring of systems (Log analysis & management, …)
- Regular software updates, if appropriate, by using patch management software
- Safe disposal of software and hardware
- …

Without sufficient security controls, all data privacy protections/guarantees will be **ineffective**!

Lecture 3 Outline

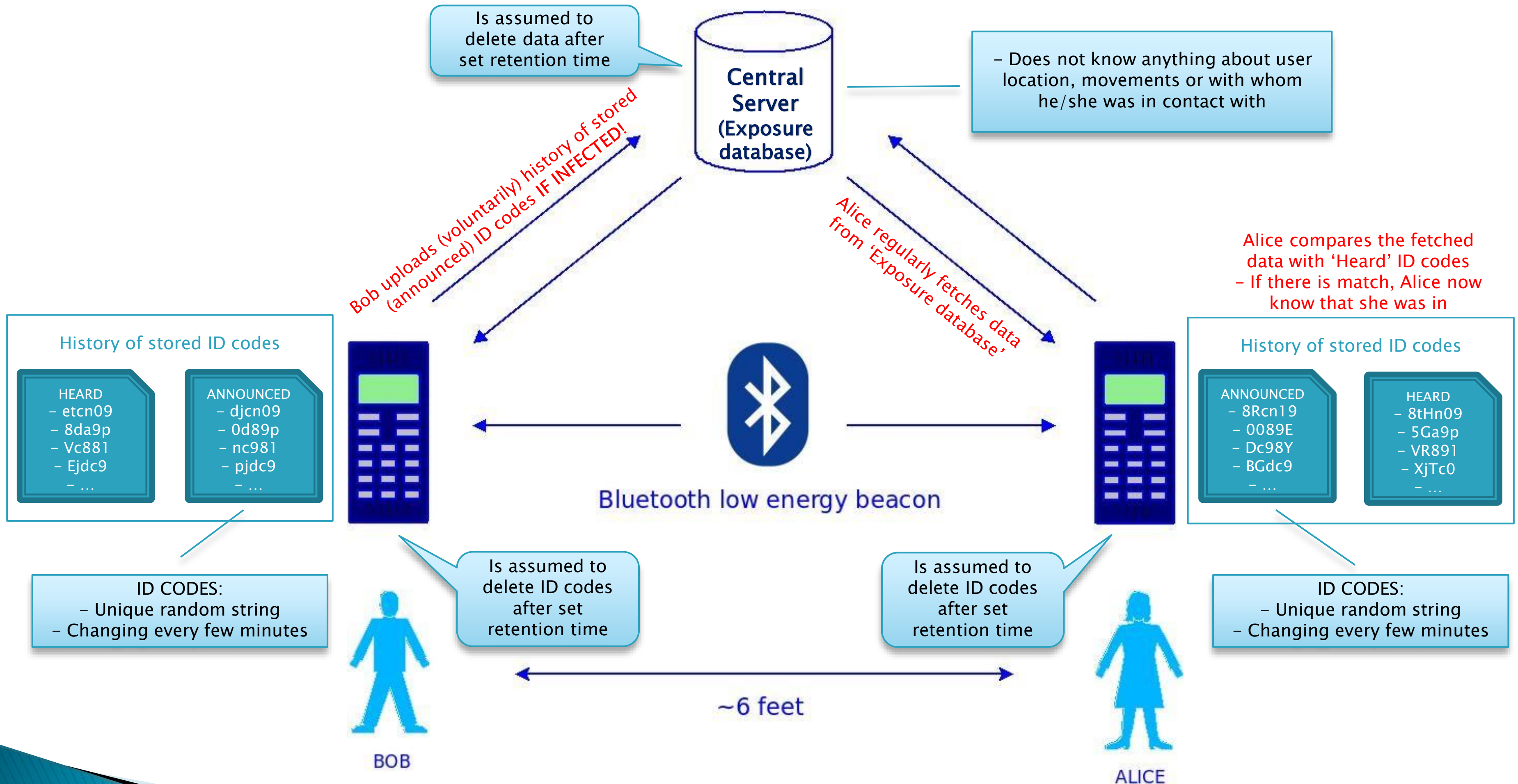
- Privacy Enhancing Technologies (PETs)
  - Data Anonymization techniques
  - Differential privacy
  - K-anonymity
  - Tor/Panoramix
  - General Security controls
  - Systematic approaches
- **Class exercise 5**

Class exercise 5 (1/2)

- **Case Study: Technology-assisted "contact tracing" (TACT) to curb the spread of COVID19**
  - System rely on location or proximity detection by mobile phones to selectively deliver alerts about potential exposures to COVID19 positive individuals
- Analyze the DP-3T proposal, and explain how it achieves DPbD goal using respective 6 strategies
  - **Use the reference model in the next slide** and **document your observation/result in a '.doc' file**
  - Other useful links:
    - https://www.aclu.org/report/aclu-white-paper-principles-technology-assisted-contact-tracing
    - https://en.wikipedia.org/wiki/Decentralized_Privacy-Preserving_Proximity_Tracing
    - https://github.com/DP-3T/

Deadline: See 'Teams' Assignment section

Class exercise 4 (2/2)



Lecture 3 ends here

- Course Slides: Go to MS Teams:
  ‘Data Privacy by Design Spring (S1) Spring 2022’
  -> Files section

- Send your questions by email:
  mohammad-salman.nadeem@epita.fr
  OR via direct message using MS Teams

- Thank You!