Privacy in Pangenomics: Introduction

Alexander Schönhuth

UNIVERSITÄT BIELEFELD
Faculty of Technology

Bielefeld University
October 12, 2023

Who are we?

- Research group “Genome Data Science”
  `https://gds.techfak.uni-bielefeld.de`
- Coordinates:
  Prof. Dr. Alexander Schönhuth
  email: aschoen@cebitec.uni-bielefeld.de
  office: UHG U10-128

Organization

Modules

- Lecture part of modules
  - *39-Inf-BDS Biomedical Data Science for Modern Healthcare Technology* (graded, “benotete Prüfungsleistung”)
  - *39-Inf-WP-CLS-x* (graded, Bachelor Informatik, Metamodul Computational Life Sciences, 10 LP)
  - *39-Inf-WP-DS-x* (graded, Bachelor Informatik, Metamodul Data Science, 10 LP)
  - *39-M-Inf-ABDA / _a Advanced Big Data Analytics* (ungraded/graded)
  - *39-M-Inf-INT-app / -foc Applied Interaction Technology* (graded, Metamodul Master Intelligent Systems, 5 / 10 LP)
- Look up details: `https://ekvv.uni-bielefeld.de/sinfo/publ/module`

Presentation, Reports, Papers

- Presentations:
  - Individual presentations
  - To last for approx. 30 minutes, followed by discussion
  - Present contents of scientific paper
- Reports:
  - Reports summarize contents of paper
  - Reports 8-10 pages
- Papers:
  - Papers: some already available, list will be completed
  - Papers available via Wiki:
    `https://gds.techfak.uni-bielefeld.de/teaching/2023winter/pangenomics`

SCHEDULE

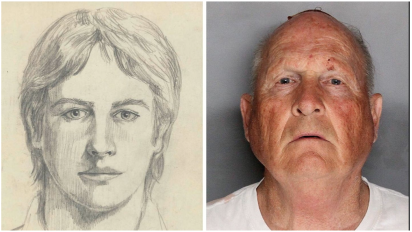
- Organization and introduction: *today*
- How to present (brief): *Oct 19* (hybrid)
- How to write (brief): *Oct 26* (hybrid)

Schedule II

- **Presentations:** *from November 30* (earlier possible if desired, but not on Nov 16 and 23)
  - Up to two presentations per week
  - Block seminar day possible as well (yet TBD)
- **Technical Report:** *after presentation:*
  - Optimally, report profits from feedback provided after presentation
  - Drafts can be submitted for discussion
  - Improving drafts based on feedback
  - *Submission deadline: February 29, 2024*

Privacy in Healthcare: Overview

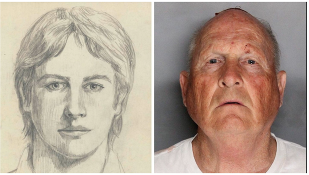
Example: Long Range Familial Searches



From www.stern.de

- Investigators uploaded crime scene sample to GEDmatch
  - GEDmatch contains 1 million DNA profiles
- GEDmatch search identified a third-degree cousin
- Genealogical search identified the perpetrator

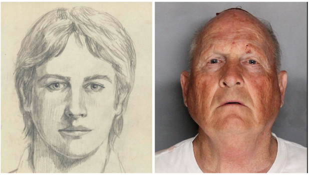
Exemplary Issues



From www.stern.de

- *Access control:*
  - Who has permission to run database searches?
  - How to organize access control?
- *Multiparty computation:*
  - Several parties share data to run computations
  - Each party’s data should stay private
  - Everyone can use data to get anonymous summaries

EXEMPLARY ISSUES

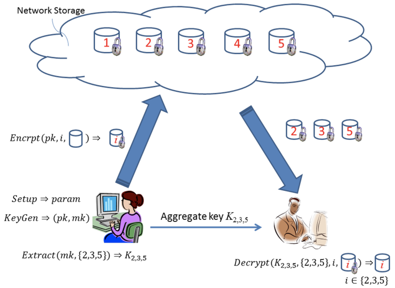


From www.stern.de

- *Homomorphic encryption:*
  - Encrypt data such that computations on encrypted data is possible
- *Differential privacy frameworks:*
  - Individual data should make no difference during analysis

Access Control

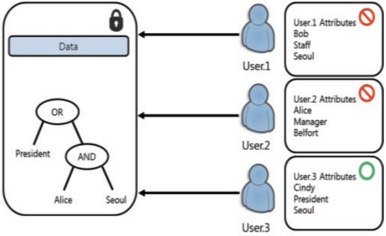
ACCESS CONTROL



From [Chu et al., 2014]

- *Key aggregate cryptography:*
  - "Master" distributes key to potential users

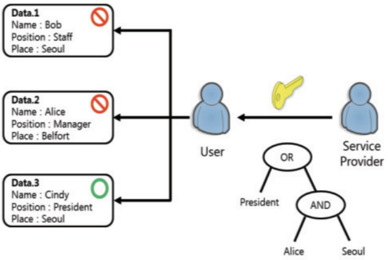
Access Control



From [Lee et al., 2015]

- *Attribute based access control:*
  - Keys depend on data characteristics

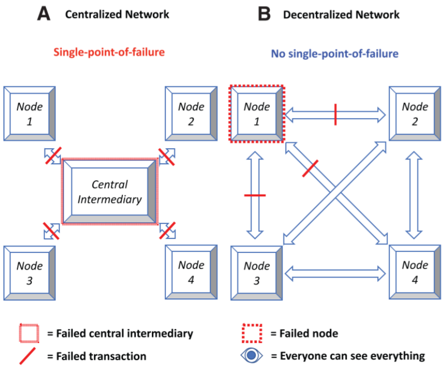
Access Control



From [Lee et al., 2015]

- *Role based access control:*
  - Keys depend on user properties

Motivation - Decentralization

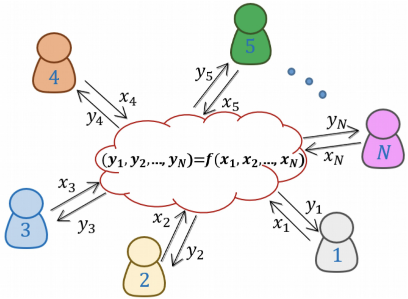


A: Central authority (e.g. running a database management system), single point of failure

B: Cluster / cloud: no single point of failure. However, no transparency, anonymity, immutability

Multiparty Computation

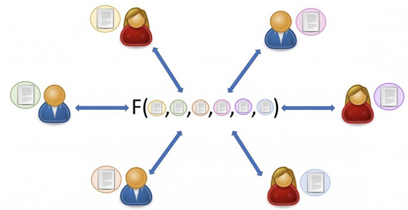
Multiparty Computation I



See www.mdpi.com

- *Multiparty computation principle:*
  - N parties provide data $x_1, \ldots, x_N$
  - Values $y_1, \ldots, y_N$ are computed
  - User providing x_i receives y_i (only)

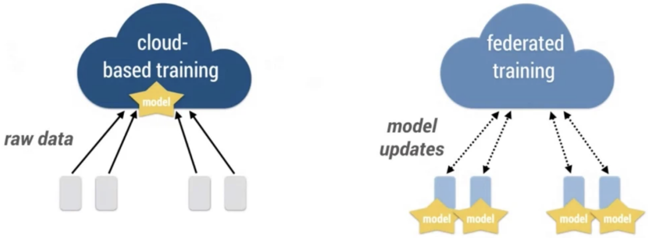
MULTIPARTY COMPUTATION II



See www.esat.kuleuven.be

- *Multiparty computation healthcare:*
  - Patients / doctors provide individual records
  - Individual analysis based on all records
  - Patients / doctors receive individual analysis results

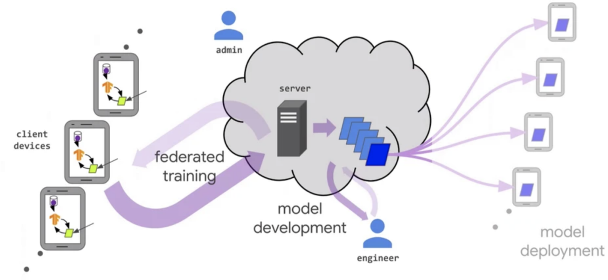
Federated Learning



See slideslive.com/38935813/federated-learning-tutorial

- *Cloud based learning:* Data transferred to cloud
- *Federated learning (FL):* Data remains stored locally
  - Reduced network strain
  - Enhanced privacy
  - Quick incorporation of new data

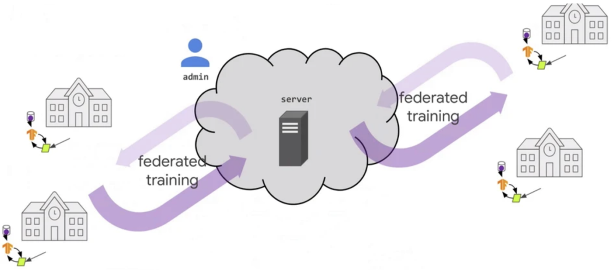
Cross-Device Federated Learning



See slideslive.com/38935813/federated-learning-tutorial

- Central engineering unit provides models to individual users
- Users train model locally with their data and return trained version
- Globally trained models used to derive individual conclusions

Cross-Silo Federated Learning



See slideslive.com/38935813/federated-learning-tutorial

- Individual institutions (clinics) store data collections
- Institutional data is used to train centrally administered models
- Institutions use globally trained models to derive conclusions

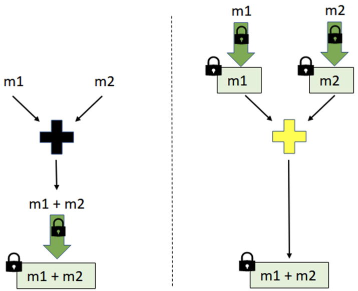
Homomorphic Encryption

Homomorphic Encryption I

See www.linksight.nl

- *Homomorphic encryption motivation:*
  - Important operations still possible after encryption
  - Decrypting data unnecessary
  - Allows users to carry out queries anonymously

Homomorphic Encryption II

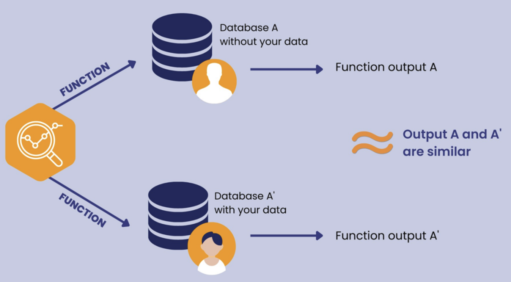


See akd13.github.io

- *Homomorphic encryption principle:*
  - Encryption and queries are mathematical operations
  - Exchanging these operations should lead to same results

Differential Privacy

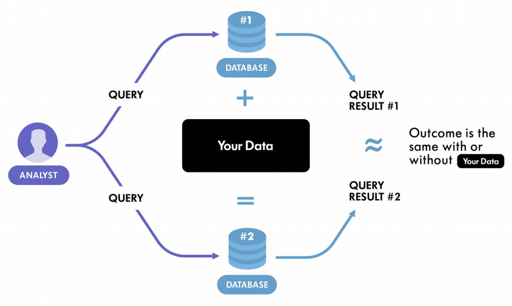
Differential Privacy I



See www.statice.ai

- *Differential privacy principle:*
  - Database A contains individual data, Database A′ does not
  - Running function returns same result on A and A′
  - *Individual data* makes no difference, so remains *unidentifiable*

Differential Privacy II



See www.winton.com

- *Differential privacy practice:*
  - Analyst runs (specially tailored) query on database with and without individual records
  - Outcomes do not differ: individual records remain anonymous

Thanks for your attention!