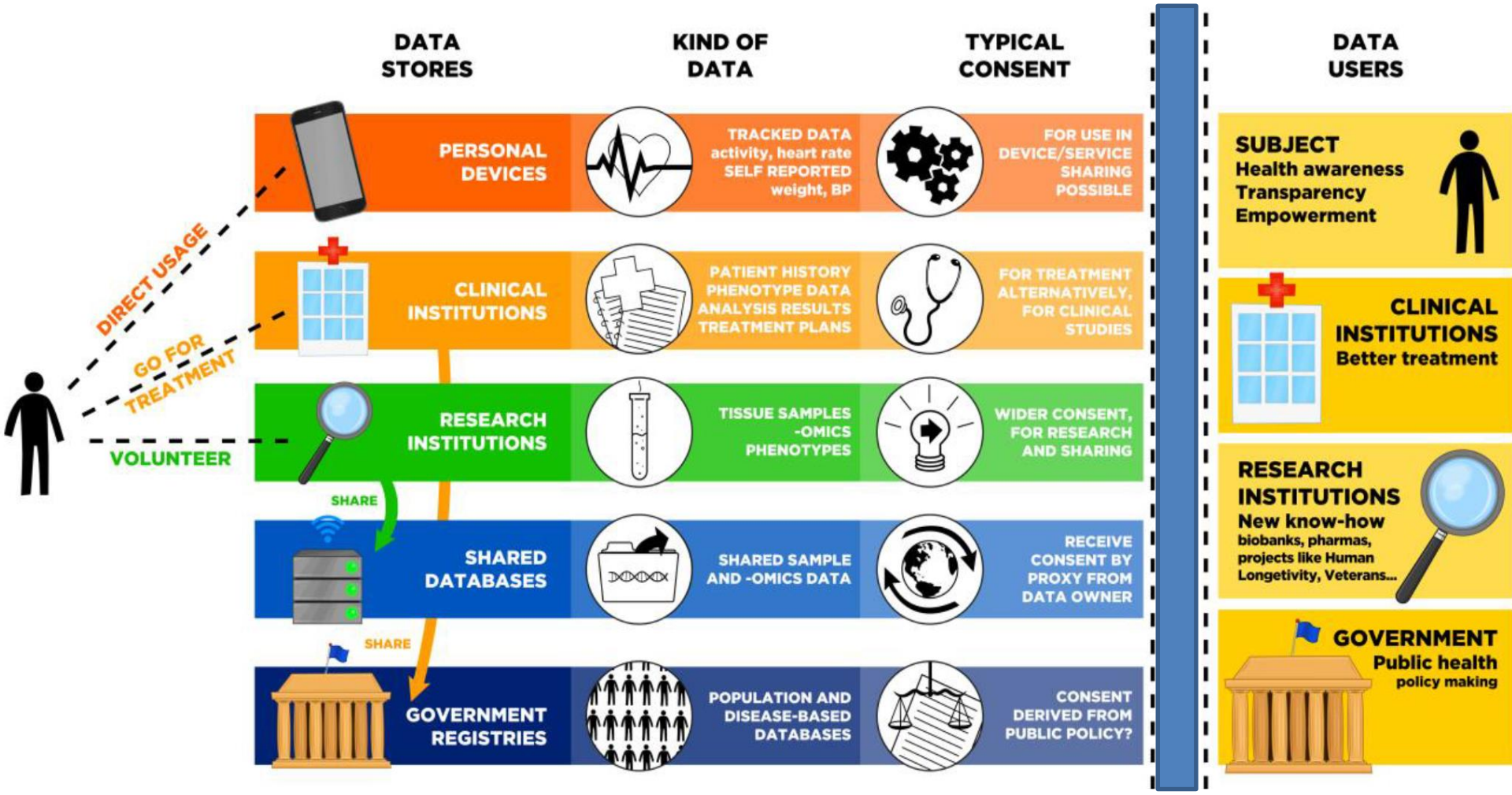
Privacy Challenges of Genomic Data Sharing

Erman Ayday

January 2018

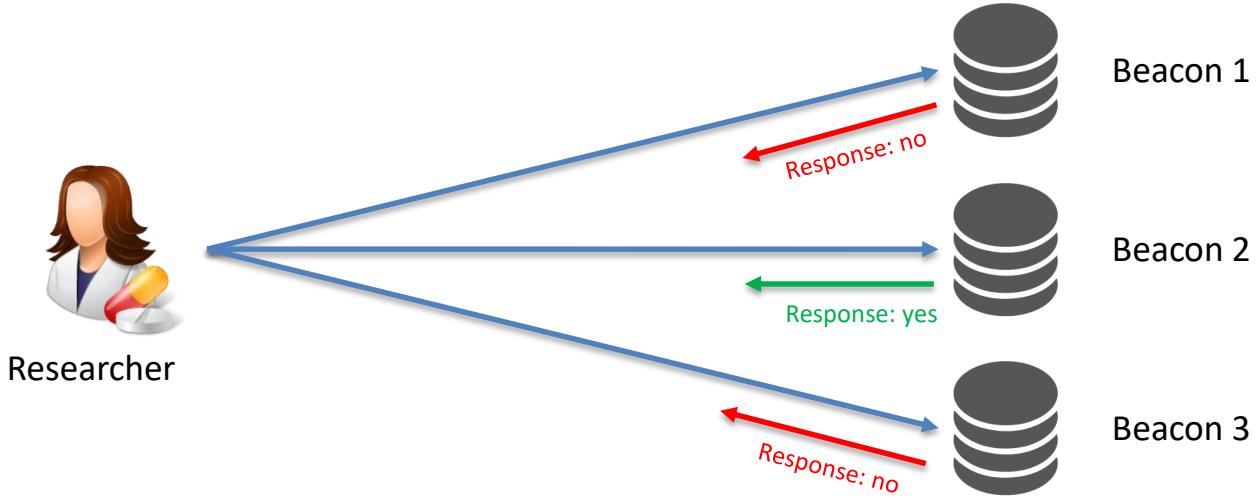
Biomedical Data Sharing



Outline

- Privacy risks of genomic data sharing beacons
  - Membership inference and more
- Privacy-preserving genome sharing
  - Optimization
  - Differential privacy
- Liability of genomic data
  - Watermarking
- Open research directions

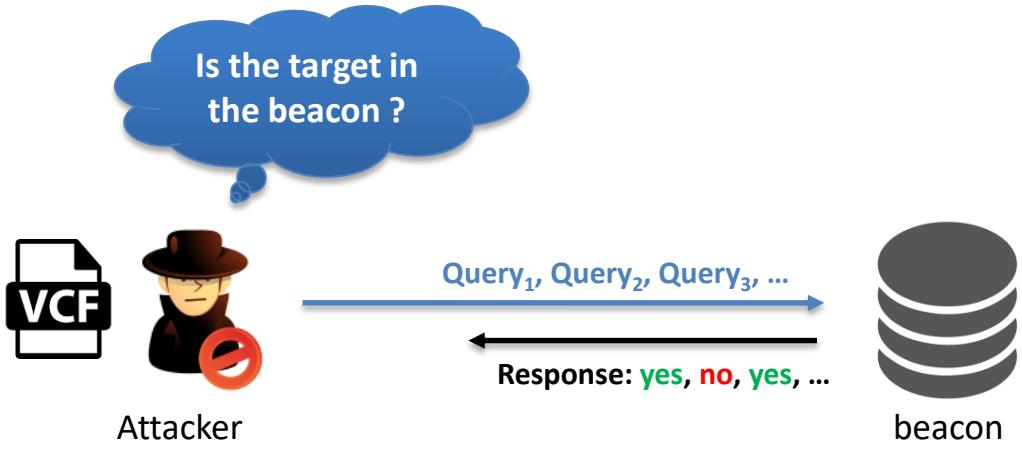
GA4GH Beacon Project

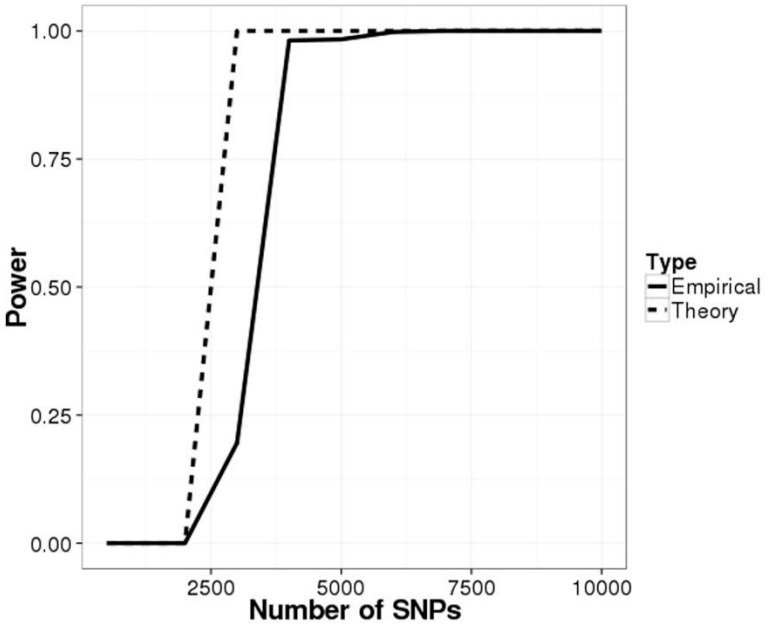


Main features:

- Allows researchers to quickly query multiple database to find the sample they need
- Encourages cross-borders collaboration among researchers
- Only provides minimal responses back in order to mitigate privacy concerns

Beacon used as an oracle: the SB attack





Shringarpure SS, Bustamante CD. Privacy risks from genomic data-sharing beacons. The American Journal of Human Genetics. 2015 Nov 5;97(5):631-46.

- The attack relies on the assumption that the adversary knows:
  - **The set of variants (VCF file) of the target individual**
  - The size of the beacon
- The attack is based on a likelihood ratio test where the adversary repeatedly queries the beacon in order to re-identify the individual
- The attack can be **extremely dangerous** if the beacon is associated with a sensitive phenotype (e.g., cancer)

Can the Attacker do Better?

Can we infer the beacon answers without actually asking them?

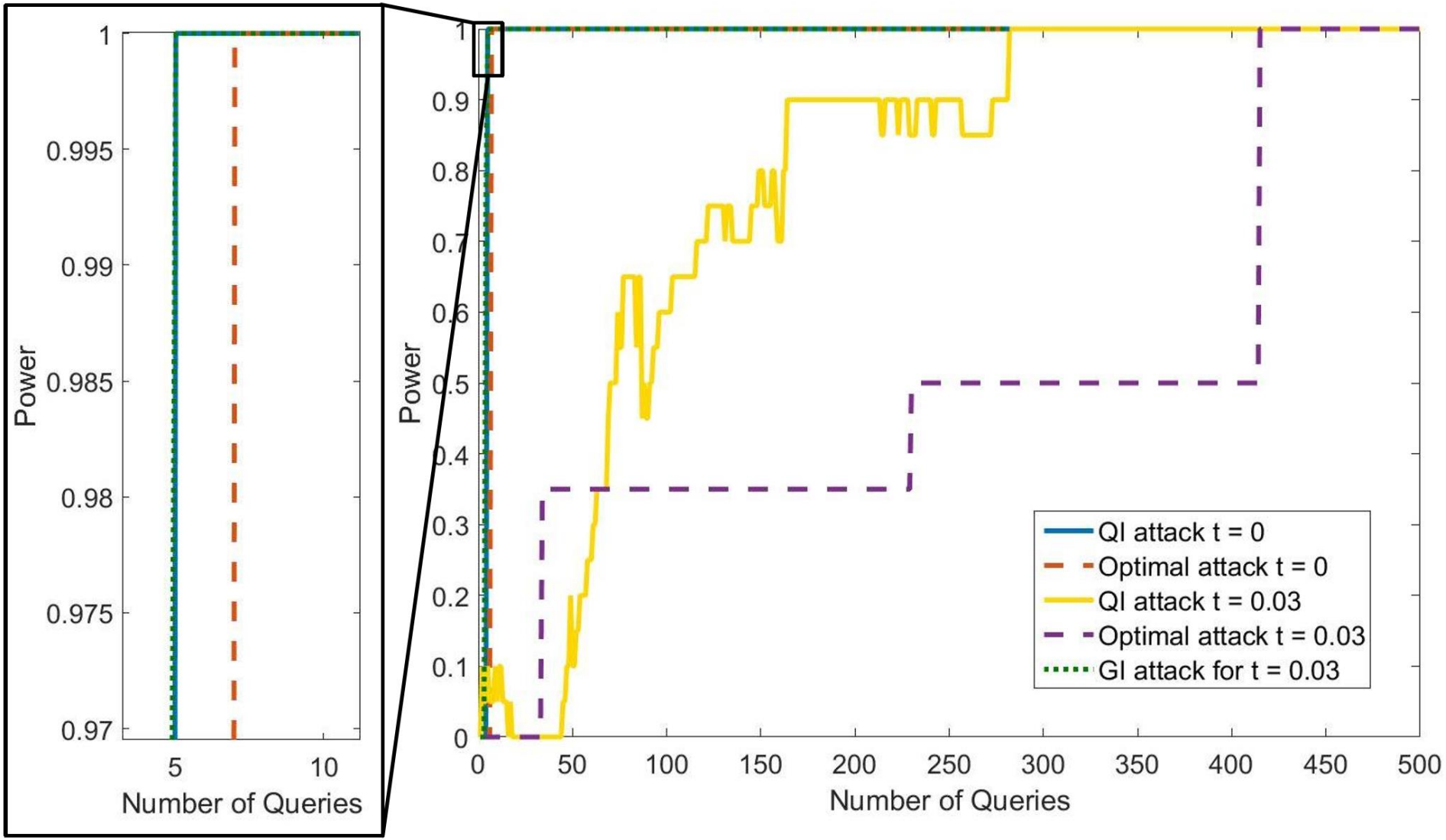
→ Yes, using linkage disequilibrium

Can we infer alleles when parts of the genome is hidden?

→Yes, using high-order Markov chains

N. V. Thenen, A. E. Cicek, and E. Ayday. “Re-Identification of Individuals in Genomic Data-Sharing Beacons via Allele Inference”, bioRxiv 200147; doi: https://doi.org/10.1101/200147, 2017.

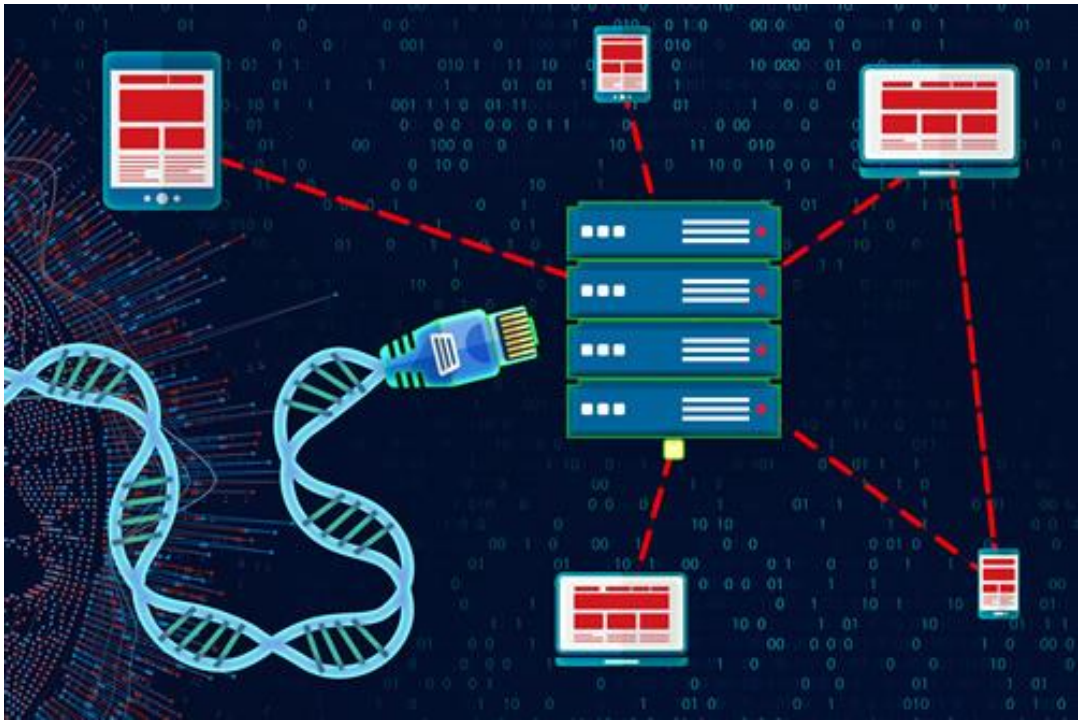
Results – Power Curves



Simulated beacon with 65 CEU individuals from HapMap

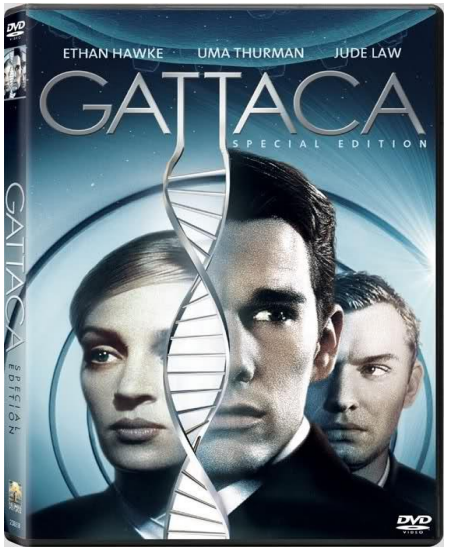
Similar Threats

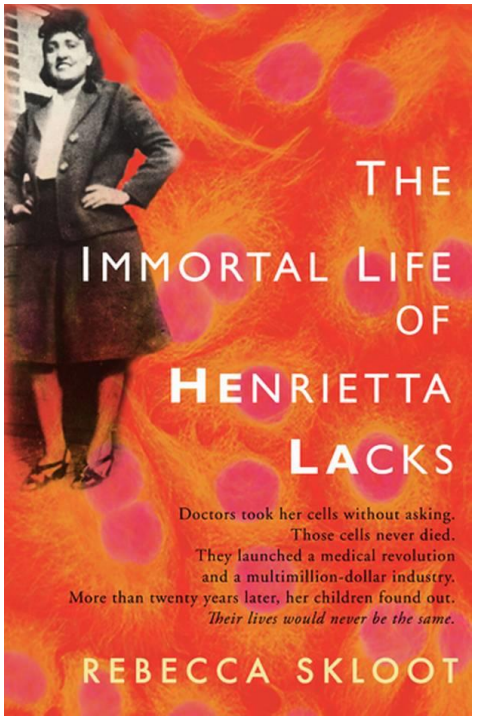
- Membership inference using auxiliary info
  - Family members
  - Phenotype
- Genome inference
- Dynamic beacons
  - Add/Remove
- Multiple beacons



Consequences of Open Data

- Genetic discrimination
- Kinship inference
- Surname inference



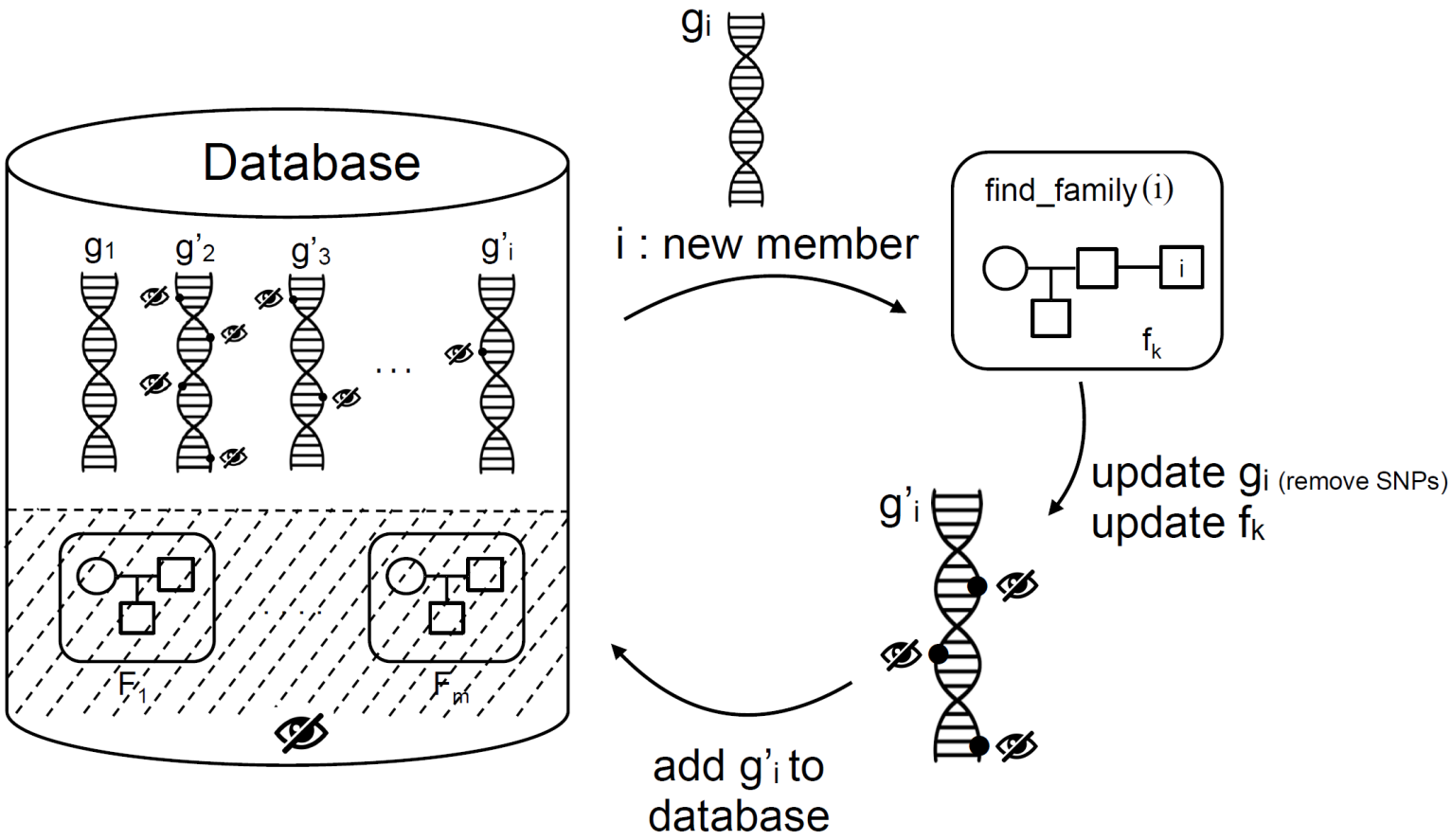


Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

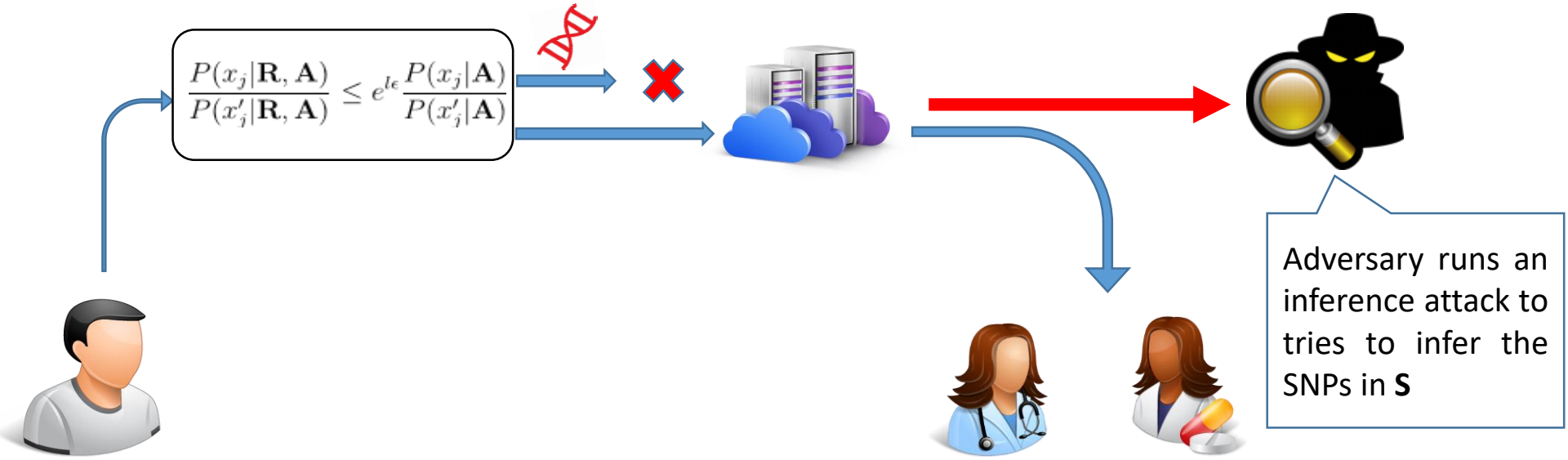
Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Protecting Kinship Inference



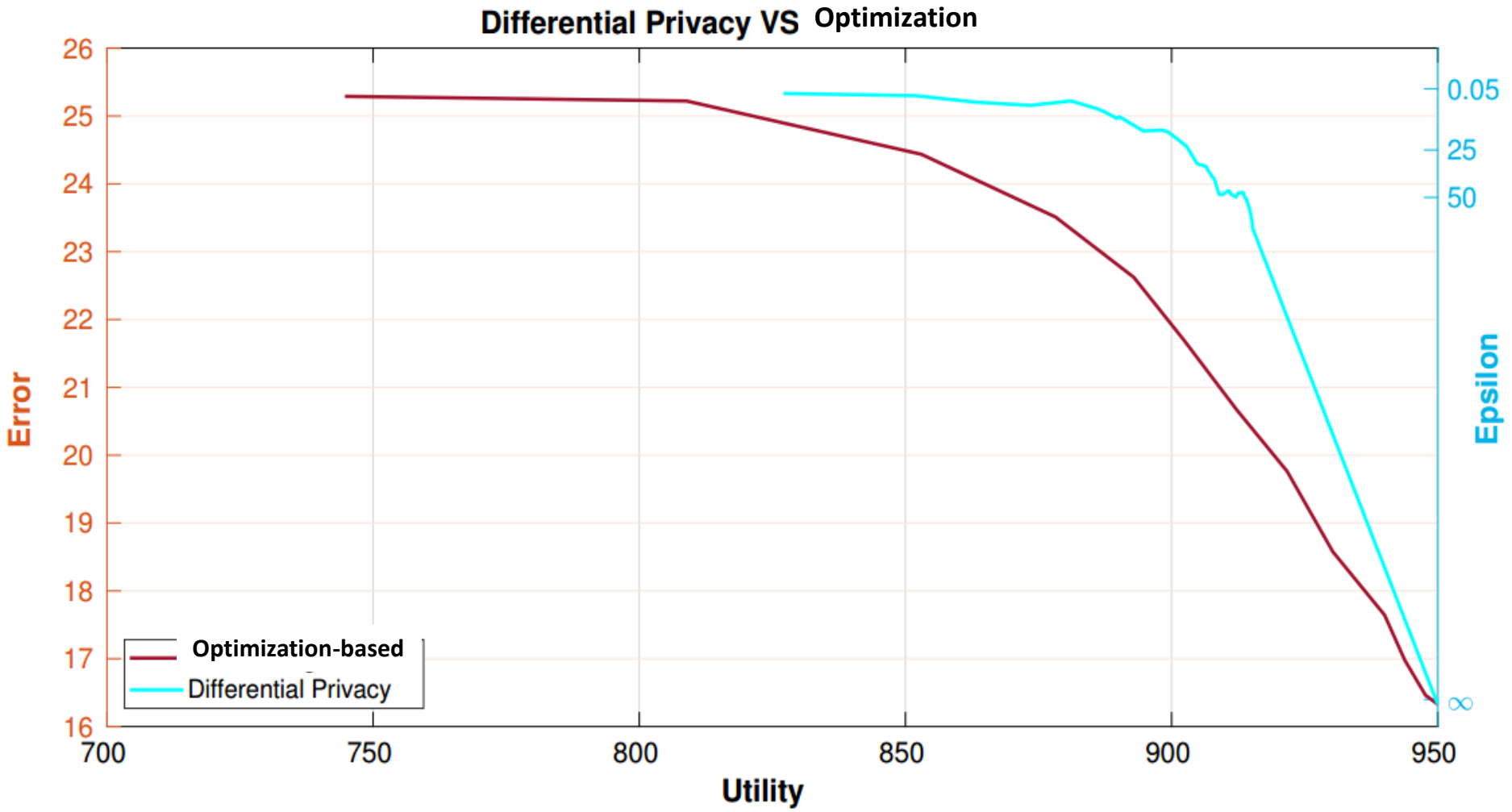
G. Kale, E. Ayday, and O. Tastan. "A Utility Maximizing and Privacy Preserving Approach for Protecting Kinship in Genomic Databases", Bioinformatics, 2017.

Differential Privacy for Individual Data Release



M. Mobayen and E. Ayday. "Tradeoff between Utility and Privacy of Genomic Data,"

Differential Privacy for Individual Data Release

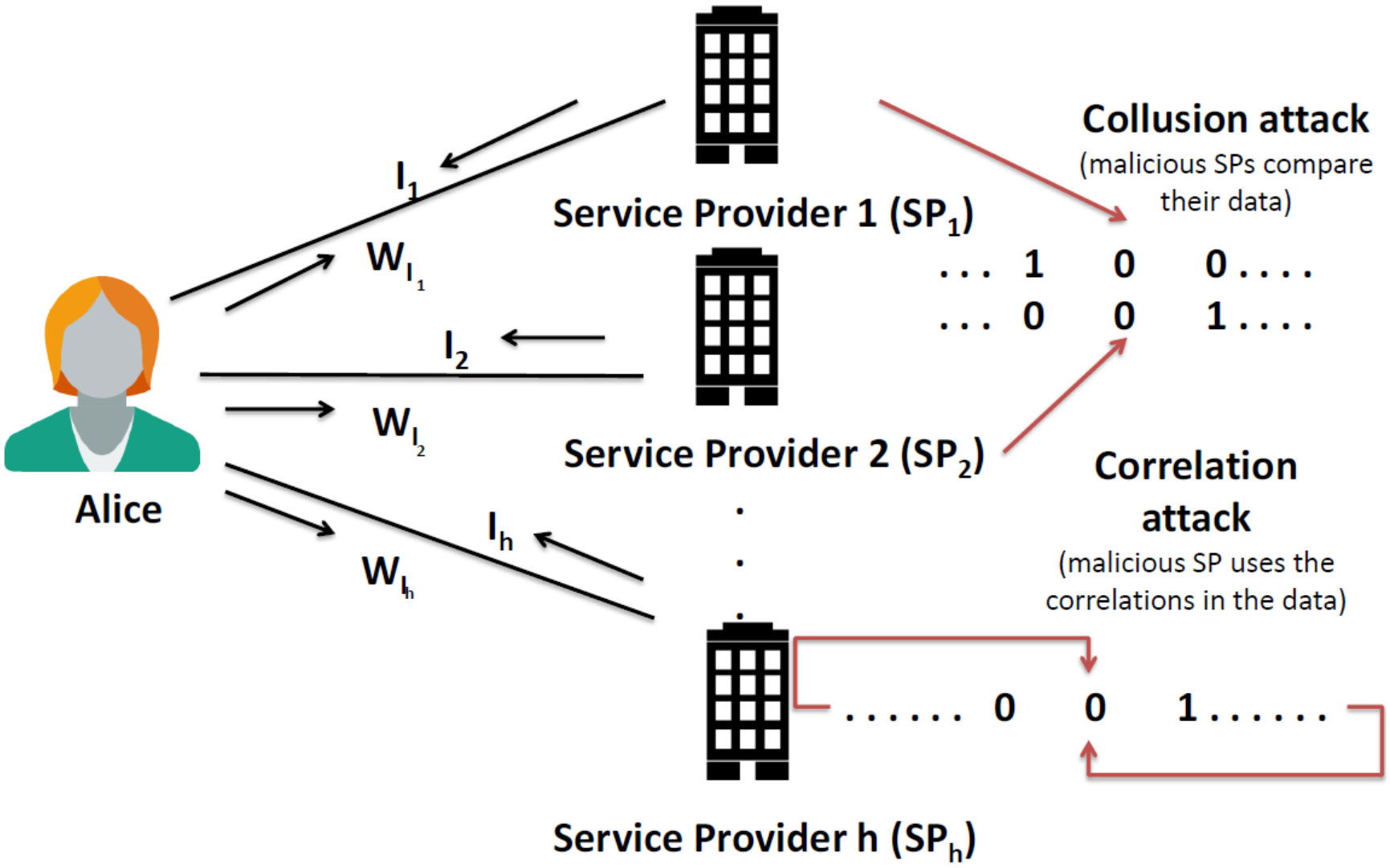


Liability of Genomic Data

- Find the source of unauthorized sharing by checking a watermark
- Goals:
  - SP that receives the data cannot understand the watermark
  - When more than one SPs aggregate their data, they still cannot determine the watermark
  - Watermark should be robust against intentional noise and partial sharing
  - Added watermark should be compliant with the nature of the corresponding data
  - Maximize the utility

A. Yilmaz and E. Ayday. “Collusion-Secure Watermarking for Sequential Data”, arXiv preprint arXiv:1708.01023, 2017.

System Model



Open Research Directions

- Genomic data sharing between different entities
- Credibility and privacy
- Privacy vs. utility of genomic data
- Interoperability
- One-time programming

Advertisement

- Looking for Ph.D. students and postdocs to work on security and privacy
- Starting as early as January 2018

Conclusion

- 35 Zettabytes (billion terabytes) of data will be generated annually by 2020 (source:IBM)
  - Most of it is will be biomedical data
- It is crucial to
  - Come up with techniques to quantify the risk on this data
  - Develop techniques to preserve its security and privacy

erman.ayday@case.edu erman@cs.bilkent.edu.tr

http://cs.bilkent.edu.tr/~erman/

@eayday