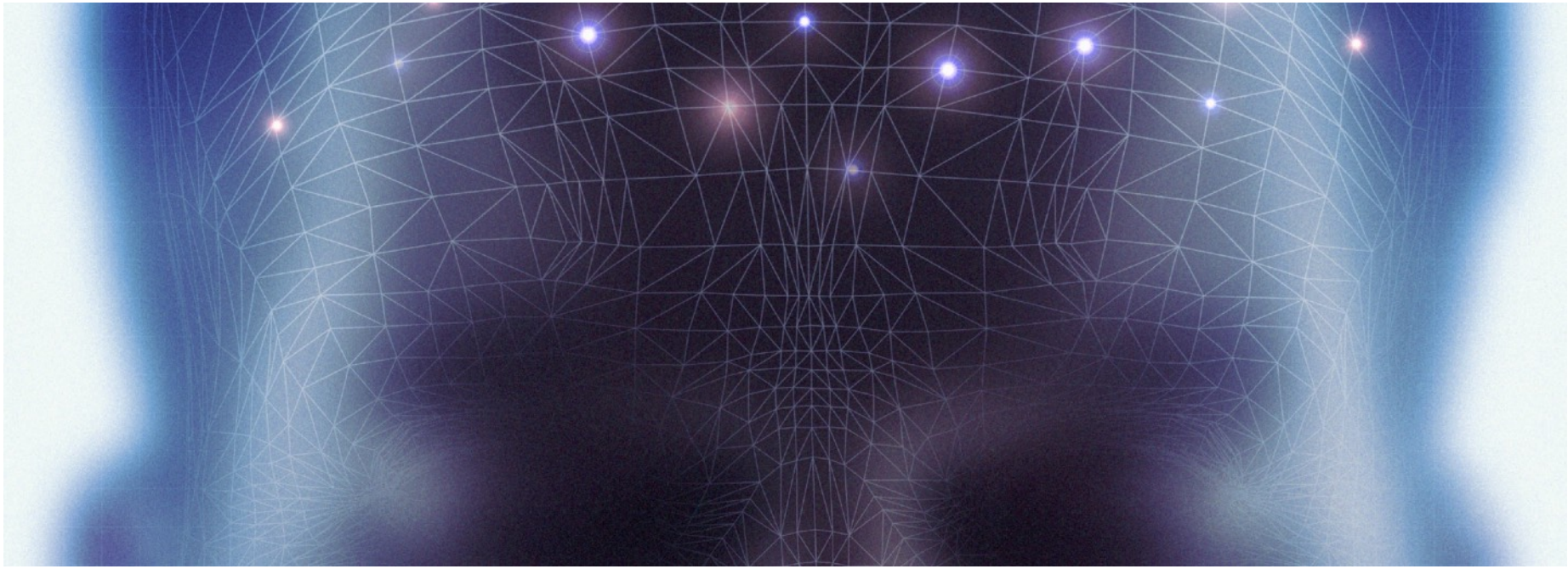


Machine Learning for Privacy
Privacy for Machine Learning

Mathias Humbert, University of Lausanne

University of Bern, November 17, 2021

Overview

Attacks

- **Machine learning for privacy**
  - How can we **evaluate privacy** in general?
    - By mimicking the adversary's behavior, by finding optimal (inference) attacks given the available data (background knowledge)
  - What is the **best (automated) approach** to infer information from data at scale?
    - Machine learning!

Tailored Defenses

- **Privacy for machine learning**
  - **Enhancing privacy** in machine-learning settings
    - No one-size-fits-all solution
    - Need to take into account the context (ML model, data, attack, …)
    - Trade-off between privacy, utility, and algorithmic efficiency

Three Key Privacy Criteria

- **Singling out**
  - Possibility to isolate some or all records which identify an individual in the dataset

- **Linkability**
  - Ability to link at least two records concerning the same individual or a group of individuals (either in the same database or in two different databases)
    → which can lead to **re-identification** if one database contains identifiers

- **Inference**
  - Possibility to deduce, with significant probability, the value of an attribute from the values of other attributes

Opinion 05/2014 on Anonymisation Techniques, **Article 29 Data Protection Working Party**, 2014

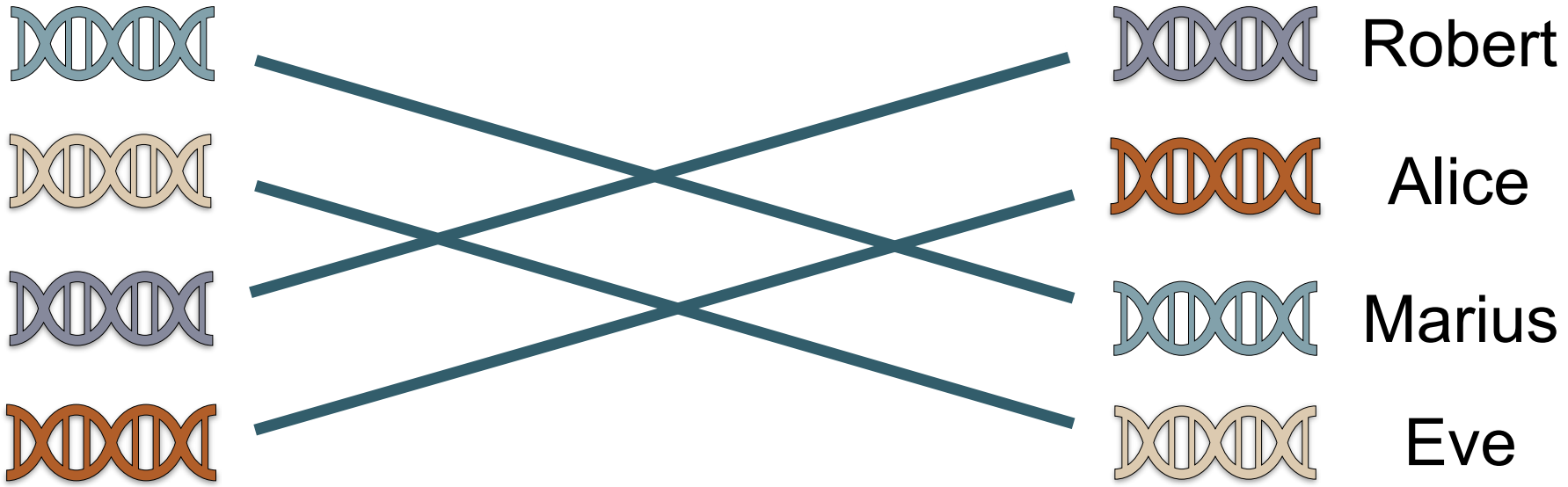
Three Key Privacy Attacks

- **Linkability/re-identification**
  - Ability to link at least two records concerning the same individual or a group of individuals (either in the same database or in two different databases)
    → which can lead to **re-identification** if one database contains identifiers

- **Attribute inference**
  - Possibility to deduce, with significant probability, the value of an attribute from the values of other attributes

- **Membership inference**
  - Possibility to deduce, with significant probability, that a specific record is part of a dataset

Linkability/Re-identification Attacks



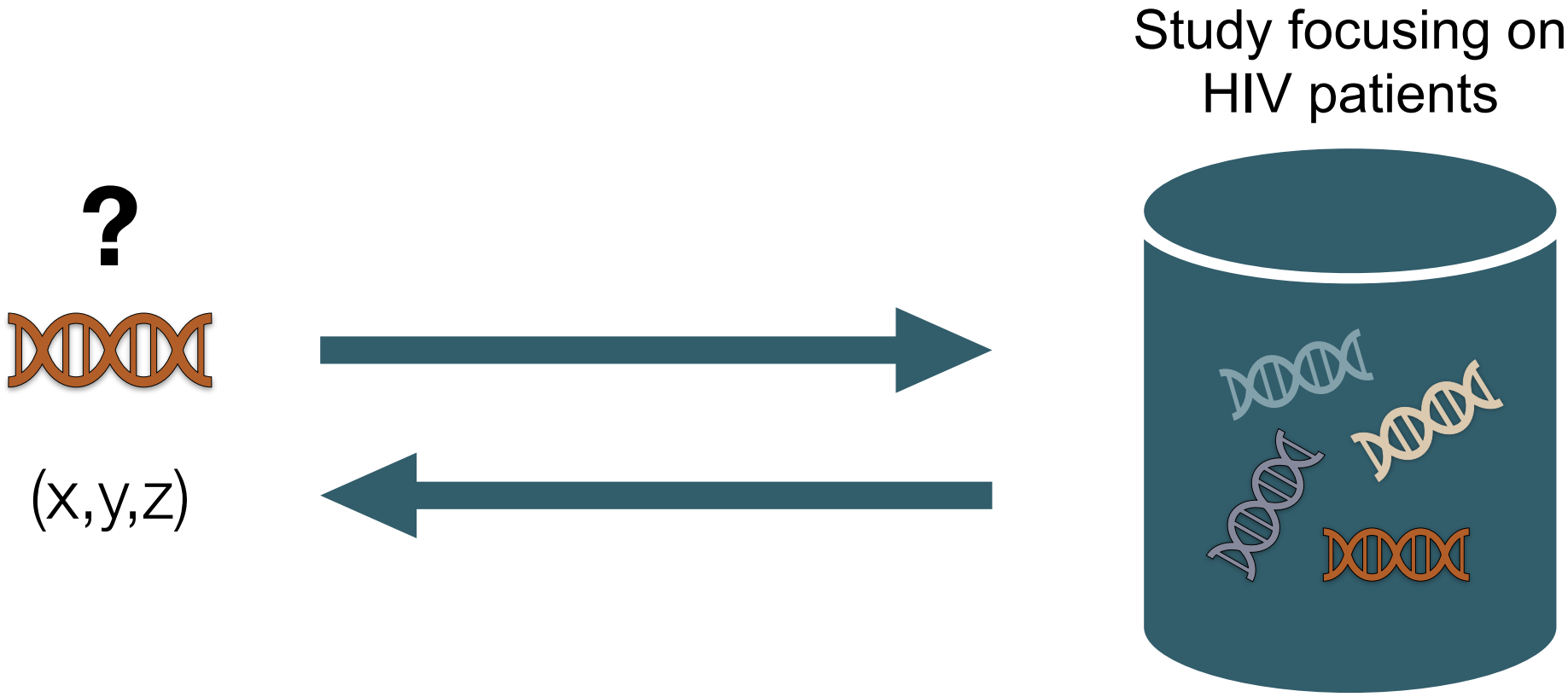
Ability to link at least two records concerning the same individual

If one dataset is not anonymized → **re-identification**

Attribute Inference Attacks

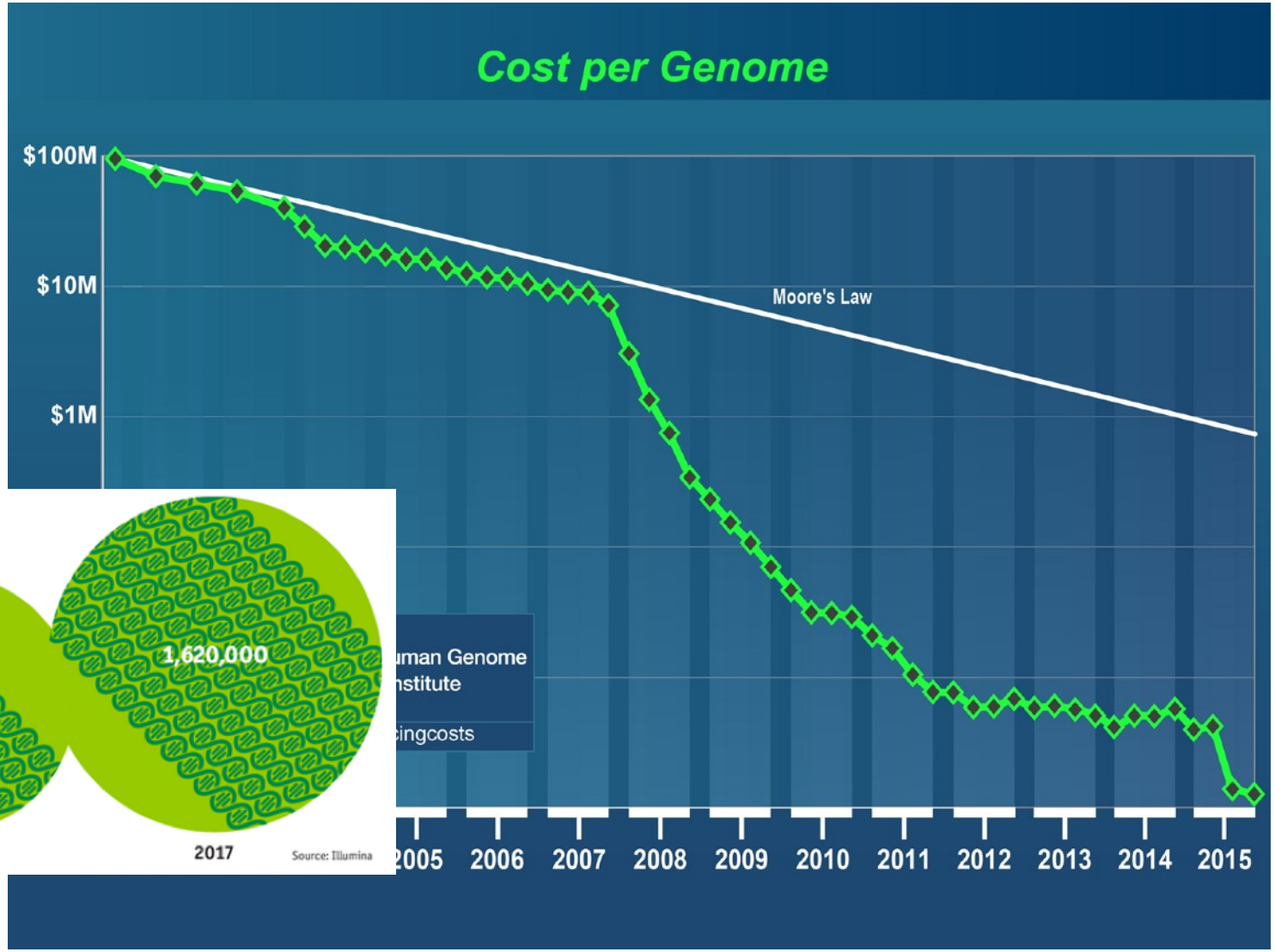
Ability to infer the value of an attribute from the values of other attributes

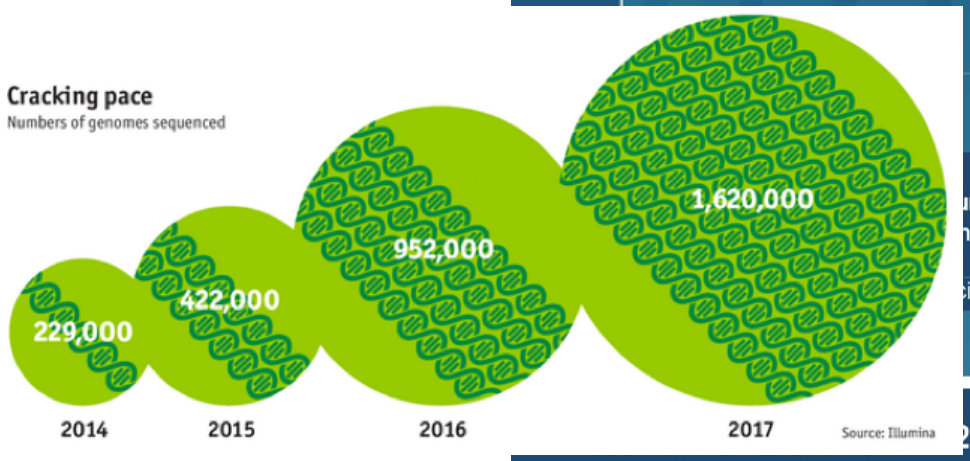
Membership Inference Attacks



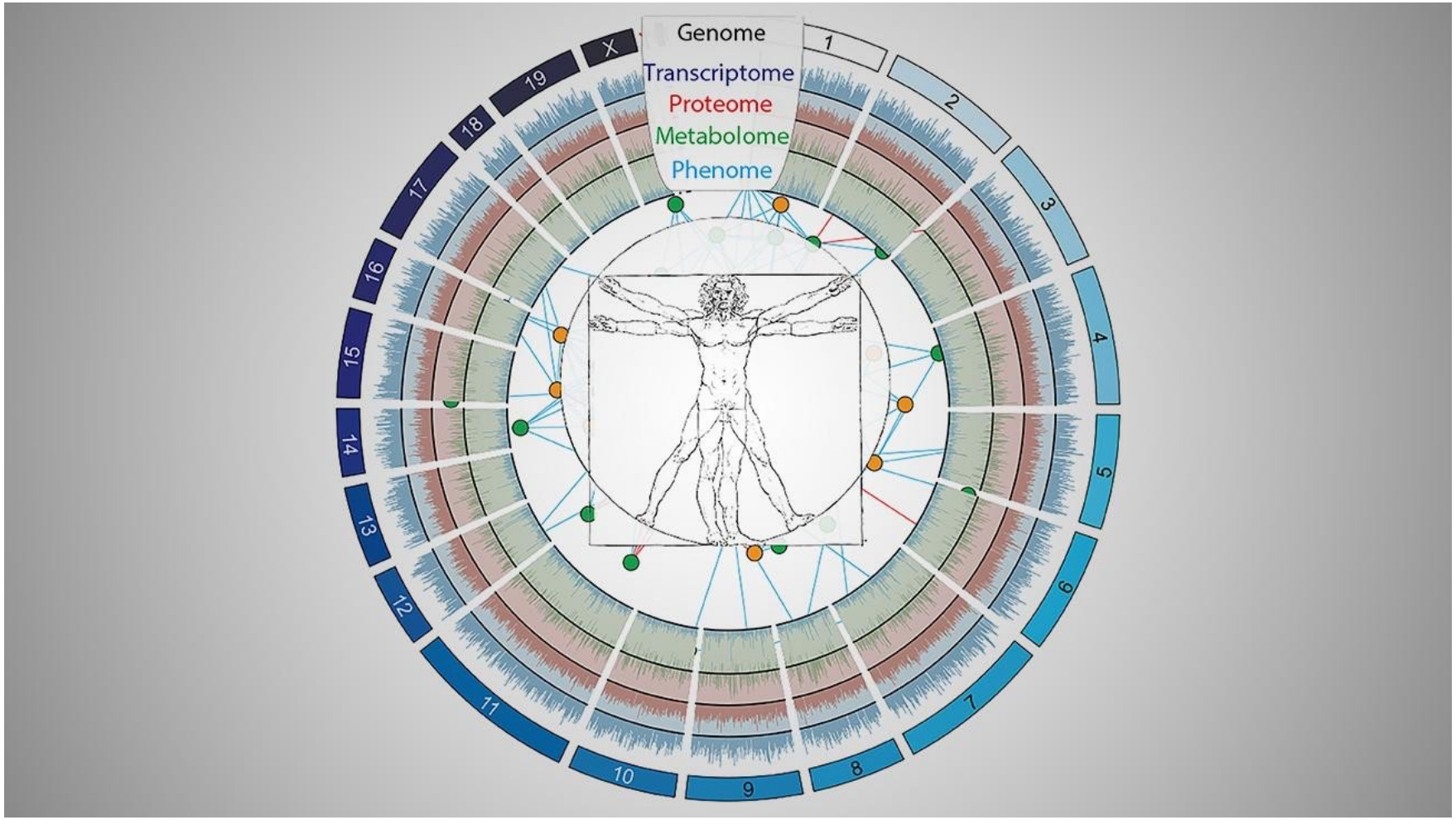
Ability to infer that a certain target is in a **specific** dataset

Deluge of Biomedical Data

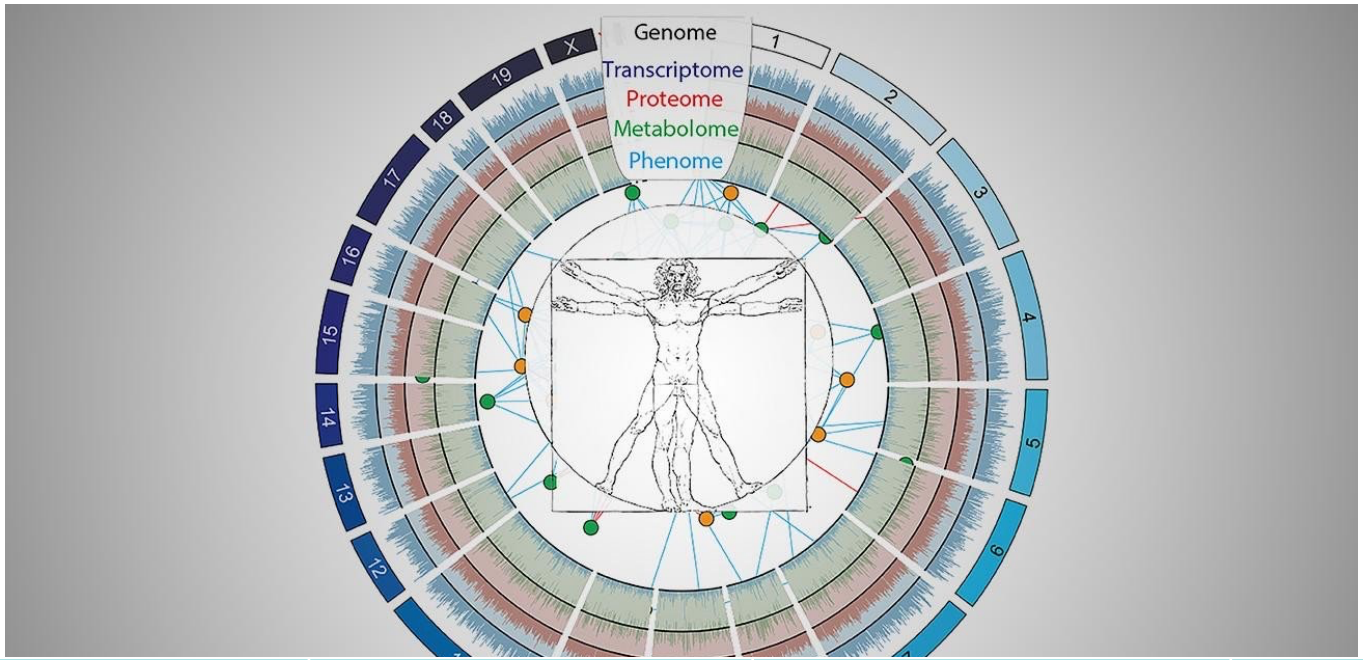




The Human OSI Stack



Biomedical Data Studied in the Research



	Data dimension	Data type	Time evolution
Genomic variants/SNPs	10^8	{0, 1, 2}	Stable
MicroRNA expression	10^3	$\mathbb{R}$	Varying+
DNA methylation	10^7	[0, 1]	Varying-

Health Privacy Risks

- **Attacks** against health companies and institutions
  - US **breach portal**: 2,000+ breaches since 2009
- Sensitive health data of **thousands of patients** ending up online due to a human mistake
- **Discriminations** based on biomedical data

Interdependent Privacy Risks

- **Interdependent** privacy [1]
  - When your privacy behavior affects **others**' privacy
  - Facebook-Cambridge Analytical scandal

- Privacy of **family members**
  - No **legal protection** framework
  - Many open questions
    - Such as **consent** of the relatives…
    - … which requires **risk assessment** to be as informed as possible

[1] Humbert et al., **A Survey on Interdependent Privacy**, ACM CSUR, 2020

Biomedicine

Do Your Family Members Have a Right to Your Genetic Code?

When a woman gets her genome sequenced, questions about privacy arise for her identical twin sister.

by Emily Mullin November 22, 2016

Twins Samantha Schilit (left) and Arielle Schilit Nitenson (right) at Nitenson's wedding in 2013.

In August 2015, Samantha Schilit went to her primary care doctor to get a blood draw. A PhD candidate at Harvard specializing in human genetics, she was itching to unlock the secrets of her genes with a test called whole-genome sequencing, which provides a full readout of a person's DNA.

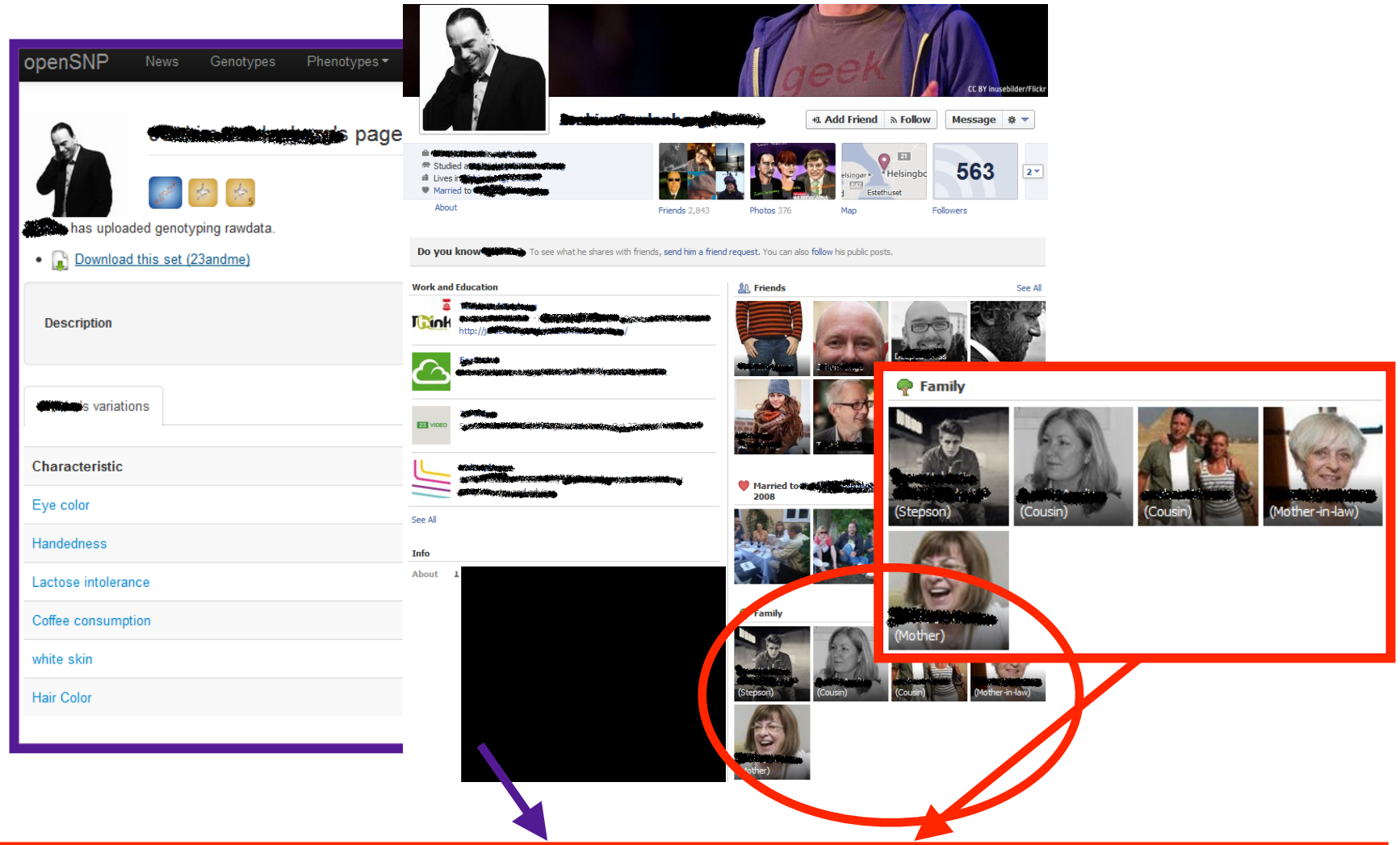
Three Key Privacy Attacks - Outline

- **Linkability and re-identification**
  - Ability to link at least two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases)
    → which can lead to **re-identification** if one database contains identifiers

- **Attribute inference**
  - Possibility to deduce, with significant probability, the value of an attribute from the values of other attributes

- **Membership inference**
  - Possibility to deduce, with significant probability, that a specific sample is part of a dataset

Attack Example via Social Network



An individual sharing his genome puts his (known) relatives' genomic privacy at risk

Quantifying Kin Genomic Privacy Risks

- Designing **efficient inference** algorithms to run the reconstruction attacks given available data

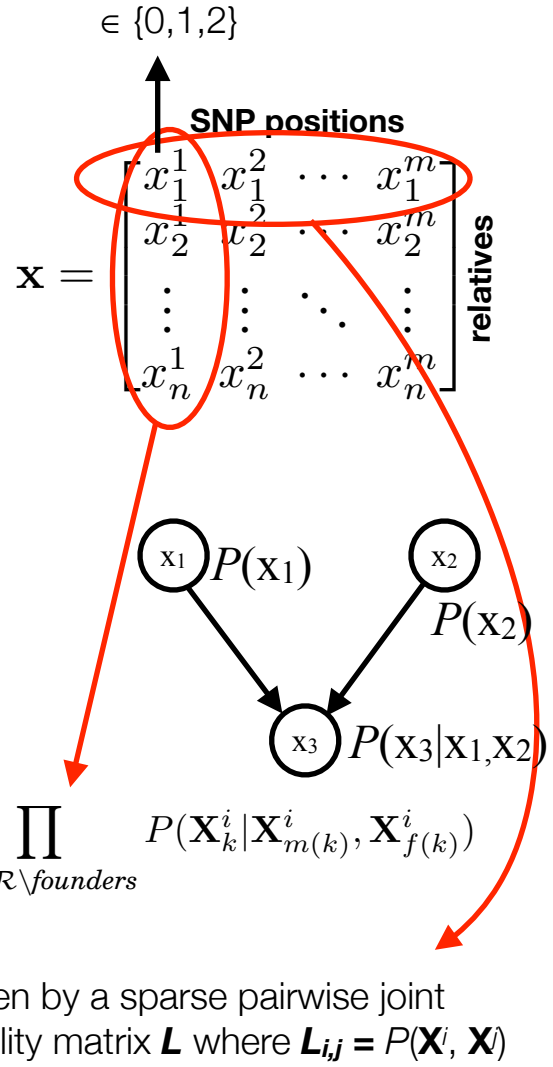
- Quantifying **privacy risks**
  - With respect to the amount of genomic data that is shared, and the relative(s) sharing them
  - Considering the background knowledge of the adversary

Inference Attack

- Adversary's **objective**:
  - Posterior marginal probabilities of the family's SNPs given:
    - Observed SNPs, priors (population statistics)
    - Inter-genome correlations (family ties), intra-genome correlations (linkage disequilibrium)

- **Naive** marginalization: computational complexity **O(3^{mn})**

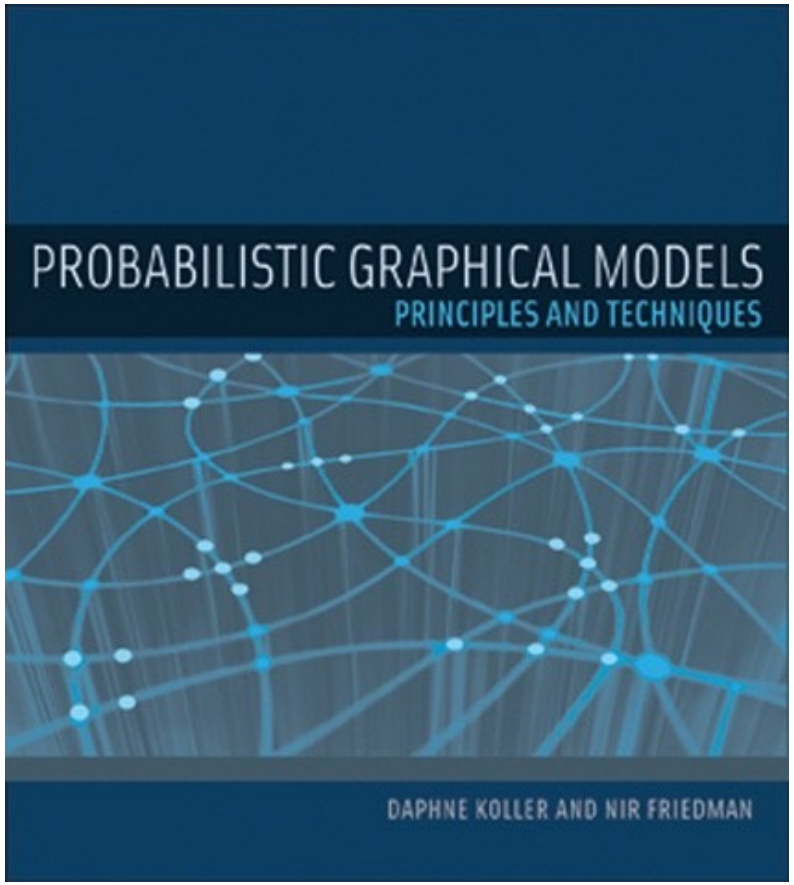
- Belief propagation on **Bayesian networks**
  - Exact inference if SNP positions are assumed to be independent
    - Junction tree algorithm
      => belief propagation on a junction tree
    - Complexity **O(mn)**



$$P(\mathbf{X}^i) = \prod_{r_j \in \textit{founders}} P(\mathbf{X}^i_j) \prod_{r_k \in \mathcal{R} \setminus \textit{founders}} P(\mathbf{X}^i_k | \mathbf{X}^i_{m(k)}, \mathbf{X}^i_{f(k)})$$

Given by a sparse pairwise joint probability matrix ***L*** where $\boldsymbol{L_{i,j}} = P(\mathbf{X}^i, \mathbf{X}^j)$

Reference Book



Quantifying Genomic Privacy

- Adversary's **expected estimation error**:

$$E_j^i = \sum_{\hat{x}_j^i \in \{0,1,2\}} P(\mathbf{X}_j^i = \hat{x}_j^i | \mathbf{X_O} = \mathbf{x_O}) \left\| x_j^i - \hat{x}_j^i \right\|$$

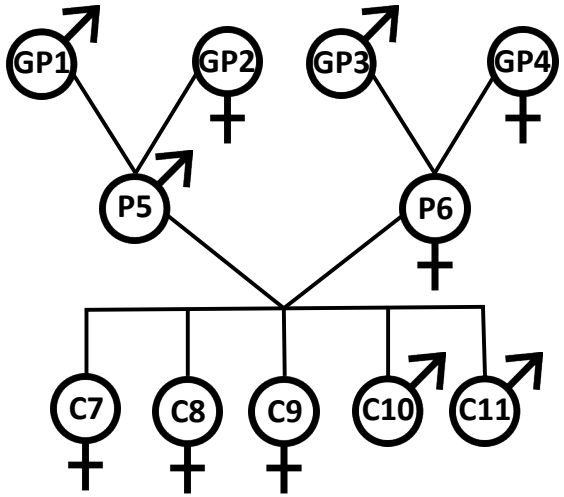
- Adversary's **uncertainty**:

$$H_j^i = \frac{-\sum_{\hat{x}_j^i \in \{0,1,2\}} P(\mathbf{X}_j^i = \hat{x}_j^i | \mathbf{X_O} = \mathbf{x_O}) \log P(\mathbf{X}_j^i = \hat{x}_j^i | \mathbf{X_O} = \mathbf{x_O})}{\log(3)} := \frac{H(\mathbf{X}_j^i | \mathbf{X_O})}{\log(3)}$$

- **Mutual information**-based metric:

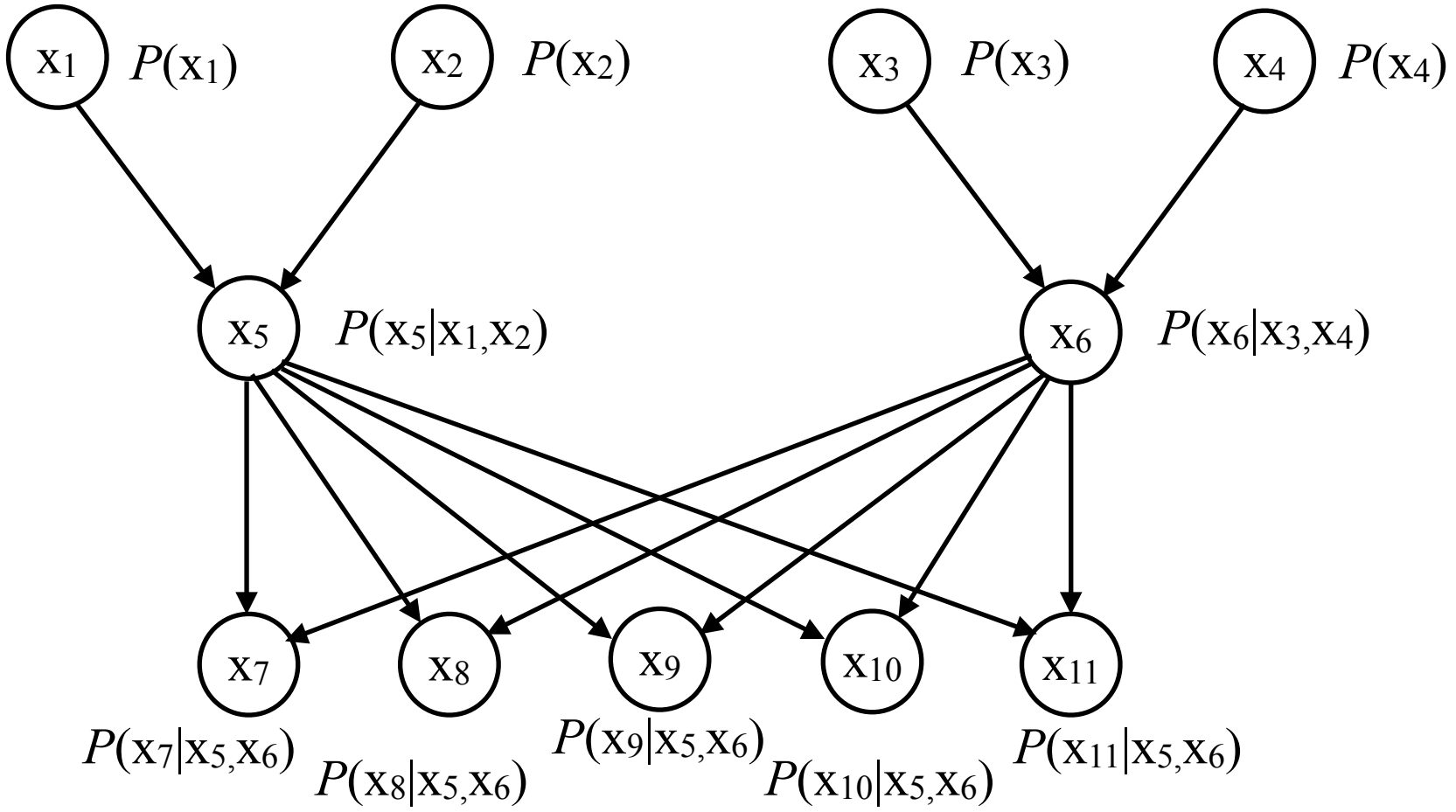
$$I_j^i = 1 - \frac{H(\mathbf{X}_j^i) - H(\mathbf{X}_j^i | \mathbf{X_O})}{H(\mathbf{X}_j^i)} = \frac{H(\mathbf{X}_j^i | \mathbf{X_O})}{H(\mathbf{X}_j^i)}$$

Framework Evaluation

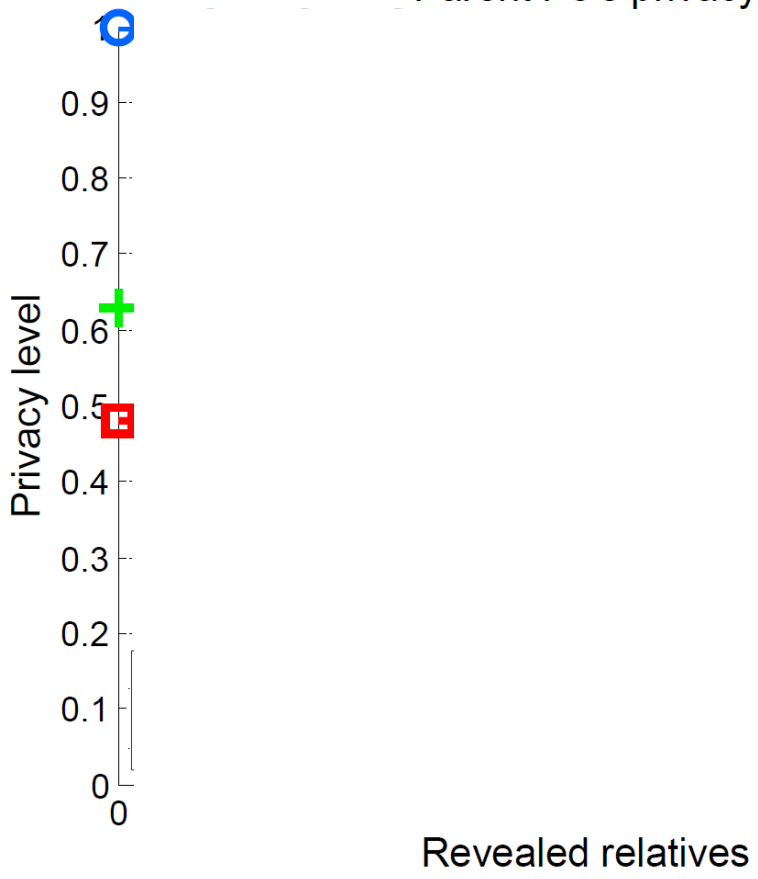


- Pedigree from Utah
  - Family containing 4 grandparents, 2 parents, and 5 children
- Focusing on **chromosome 1** (longest one)
- Relying on the three privacy metrics to quantify:
  - **Genomic privacy**: average of the privacy over all SNPs
  - **Health privacy**: weighted average over the SNPs contributing to the considered disease(s)

Bayesian Network Model



Inferring Parent P5's Genome



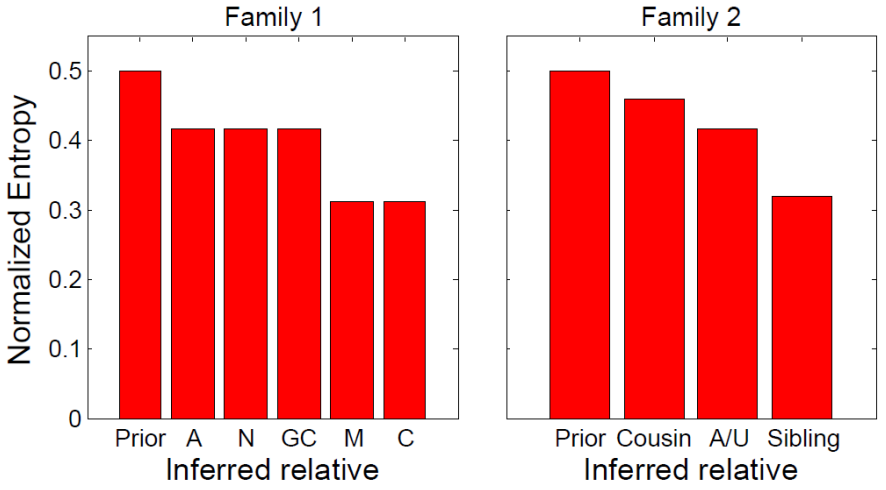
> 80,000 SNPs on chromosome 1

? P5

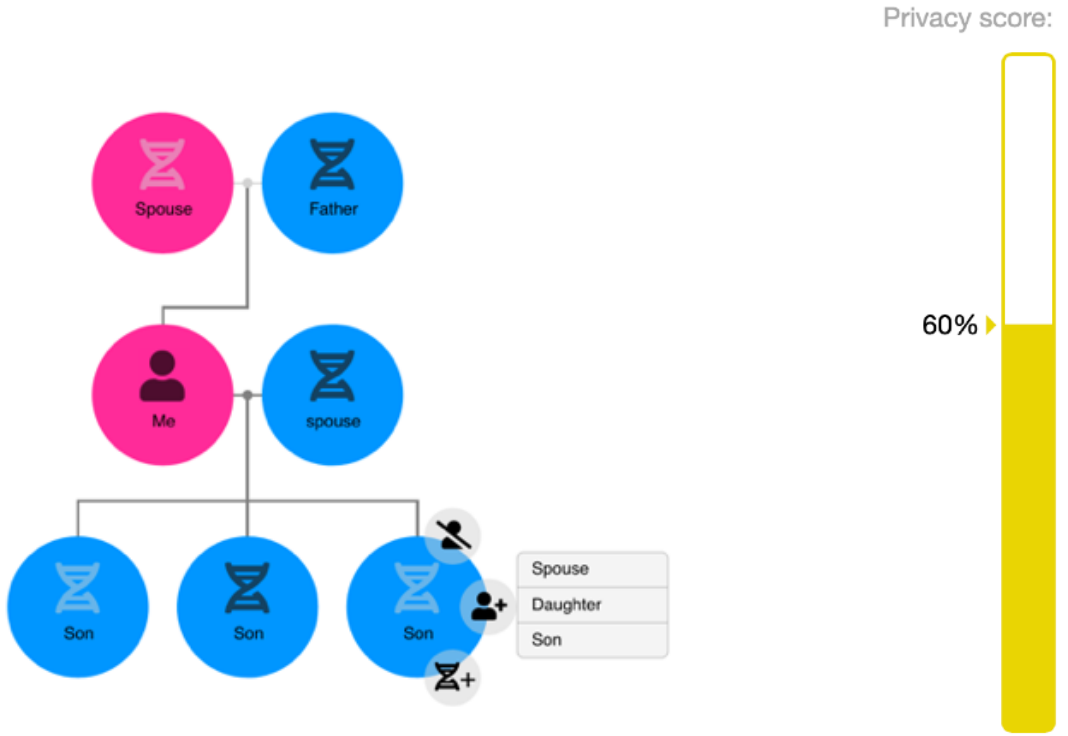
Real Attack Example

- Linking **OpenSNP** and **Facebook** profiles
  - 6 individuals sharing their genomic data on OpenSNP with a Facebook profile with relatives publicly available
  - Privacy of 29 relatives in the 6 different families at risk

- **Health-privacy** evaluation for two families
  - Alzheimer's disease
    - 2 SNPs equally contributing to the disease risk
    - 1 person/family sharing these 2 SNPs



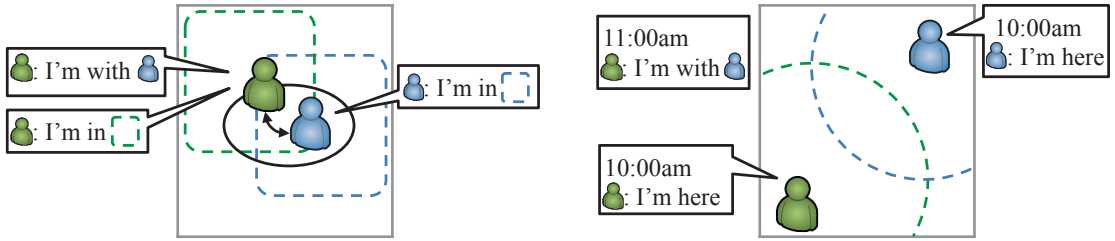
Interactive Online Tool for Estimating Genomic Privacy



Available at https://santeperso.unil.ch

Similar Model for Location Privacy

- **Co-location** information is widespread
  - E.g., on online social networks
  - Improving adversary's power to further reduce users' location privacy
  - … but increasing the complexity of the inference attack



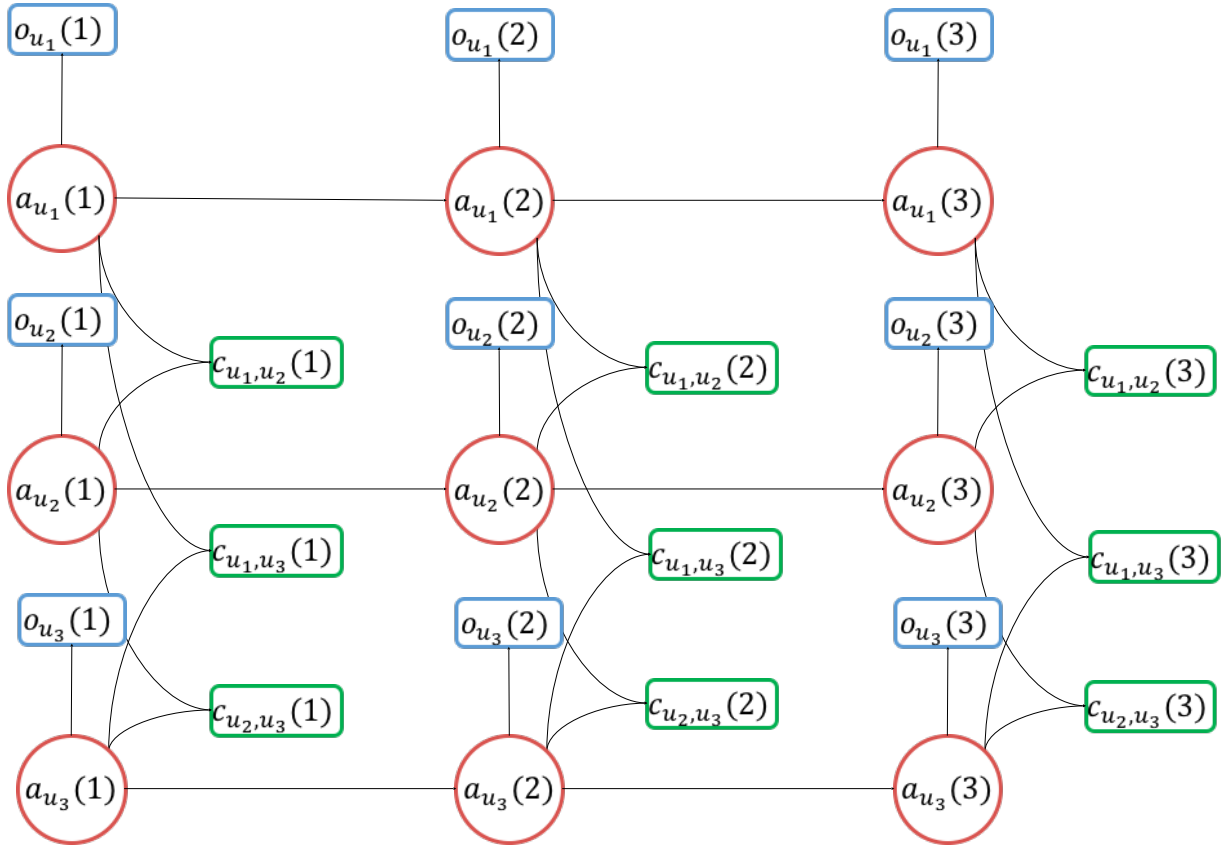
- Traditional mobility model: **hidden Markov models** [2]
  - Does not scale when users' locations are correlated
  - Our approach: Bayesian networks and belief propagation [3]

[2] Shokri et al., **Quantifying Location Privacy**, IEEE S&P, 2011

[3] Olteanu et al., **Quantifying Interdependent Privacy Risks with Location Data**, IEEE TMC, 2017

From HMMs to a Bayesian Network Model

Example: 3 mobile users, 3 time points



Three Key Privacy Attacks - Outline

- **Linkability and re-identification**
  - Ability to link at least two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases)
    → which can lead to **re-identification** if one database contains identifiers

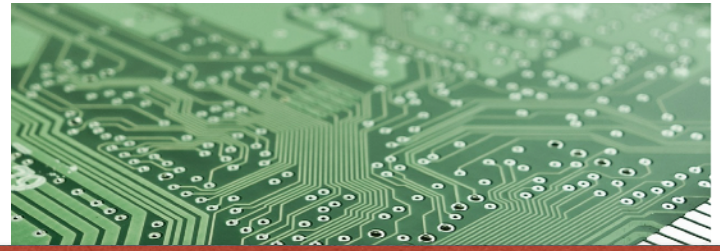
- **Attribute inference**
  - Possibility to deduce, with significant probability, the value of an attribute from the values of other attributes

- **Membership inference**
  - Possibility to deduce, with significant probability, that a specific sample is part of a dataset

Genome vs. MicroRNA Expression

The genome

- contains blueprint of what a cell potentially can do,
- is (mostly) fixed over time,
- can hint on risks of getting a disease.



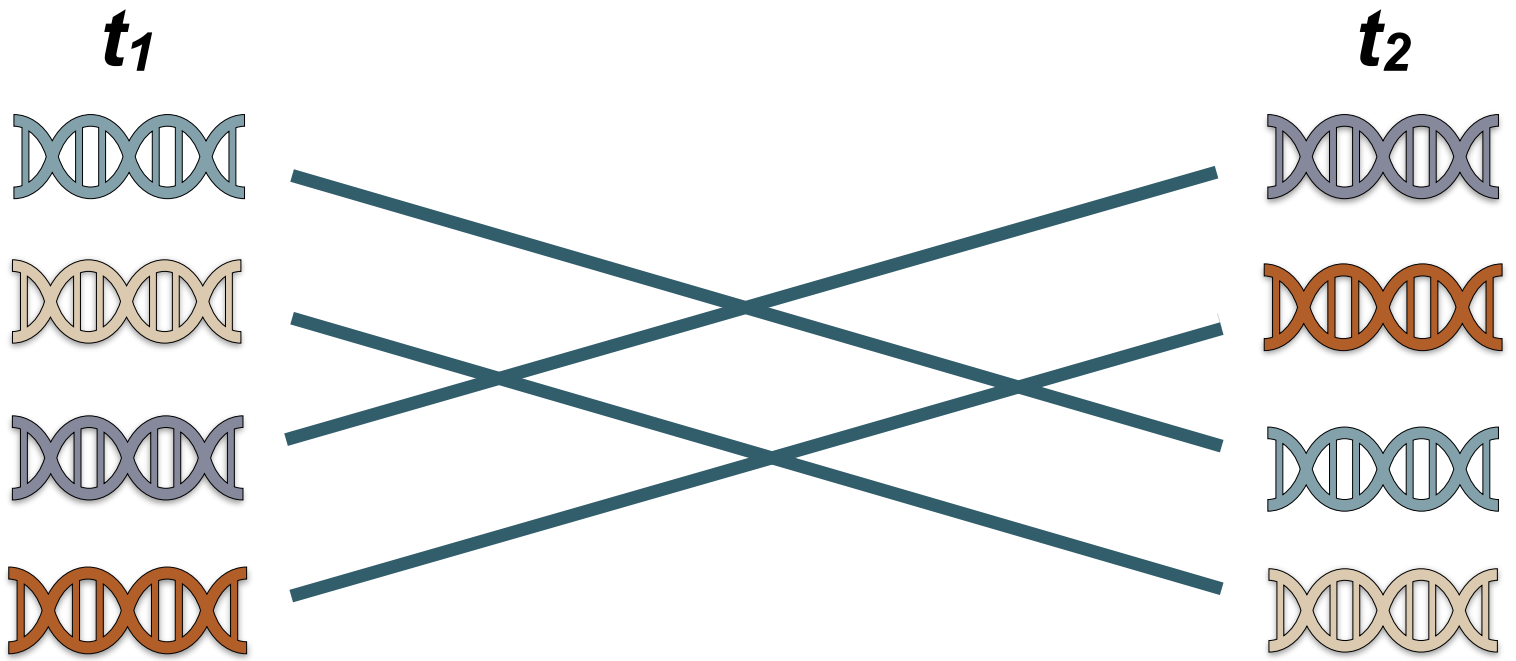
miRNA expression

- regulates what a cell really does,
- miRNA expression changes over time,
- can tell whether you carry a disease.

Common belief: **no privacy threat** from miRNAs, because of **temporal variability**

Temporal Linkability Attack

- Matching two datasets
  - E.g., a leaked database (including name) and public database (excluding name)
  - **Which sample from t_1 corresponds to which sample from t_2?**

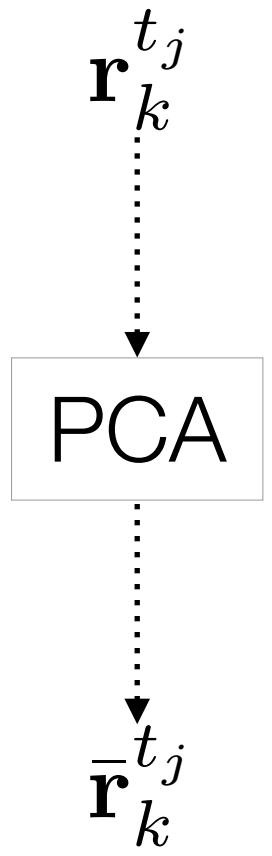


Data Pre-Processing

- High dimensionality: **1,189 miRNAs** per sample
  - Possibly correlated and uninteresting components

- **PCA + whitening** provides
  - Unit variance
  - Smaller dimensionality
  - Uncorrelated components

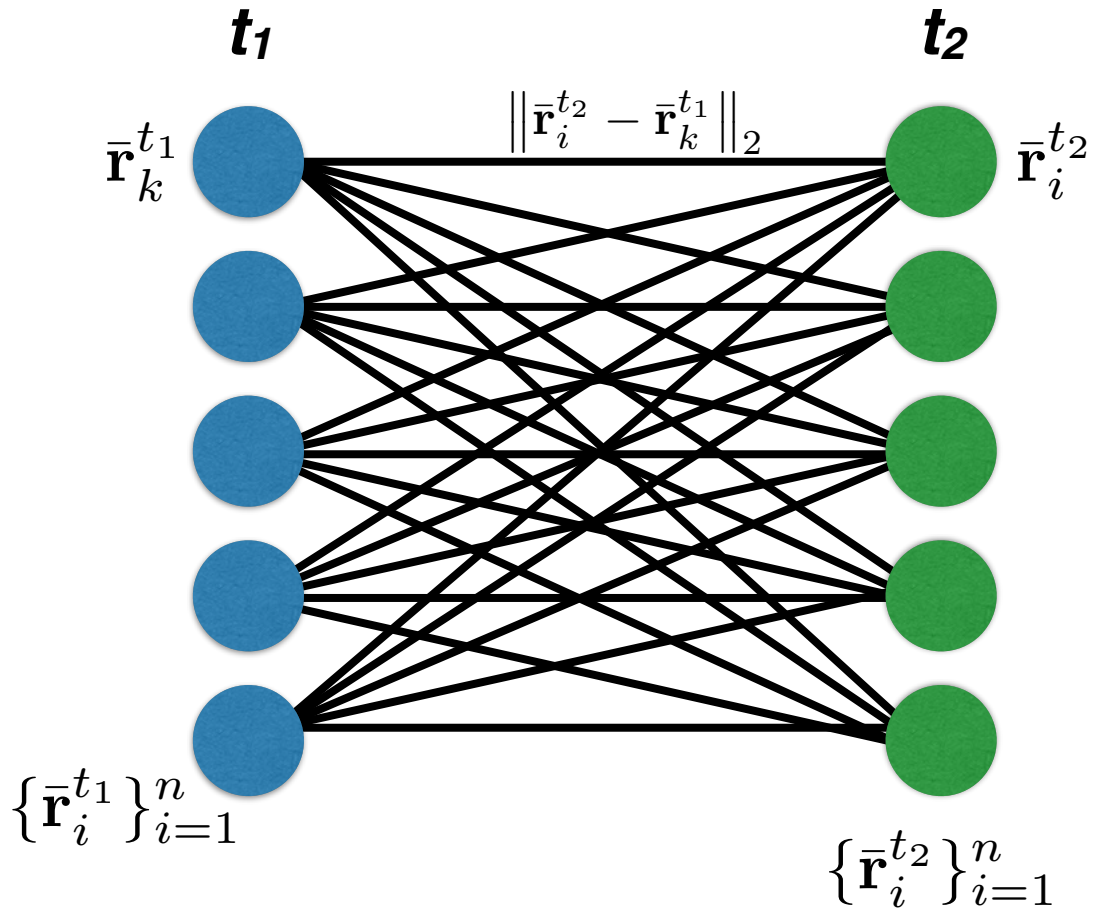
- Condenses data into a set of smaller dimensions with **minimal information loss**



Linkability Attack

Which sample from t_1 corresponds to which sample from t_2?

$$\sigma^* = \arg\min_{\sigma} \sum_{i=1}^{n} \left\| \bar{\mathbf{r}}_{\sigma(i)}^{t_2} - \bar{\mathbf{r}}_i^{t_1} \right\|_2$$

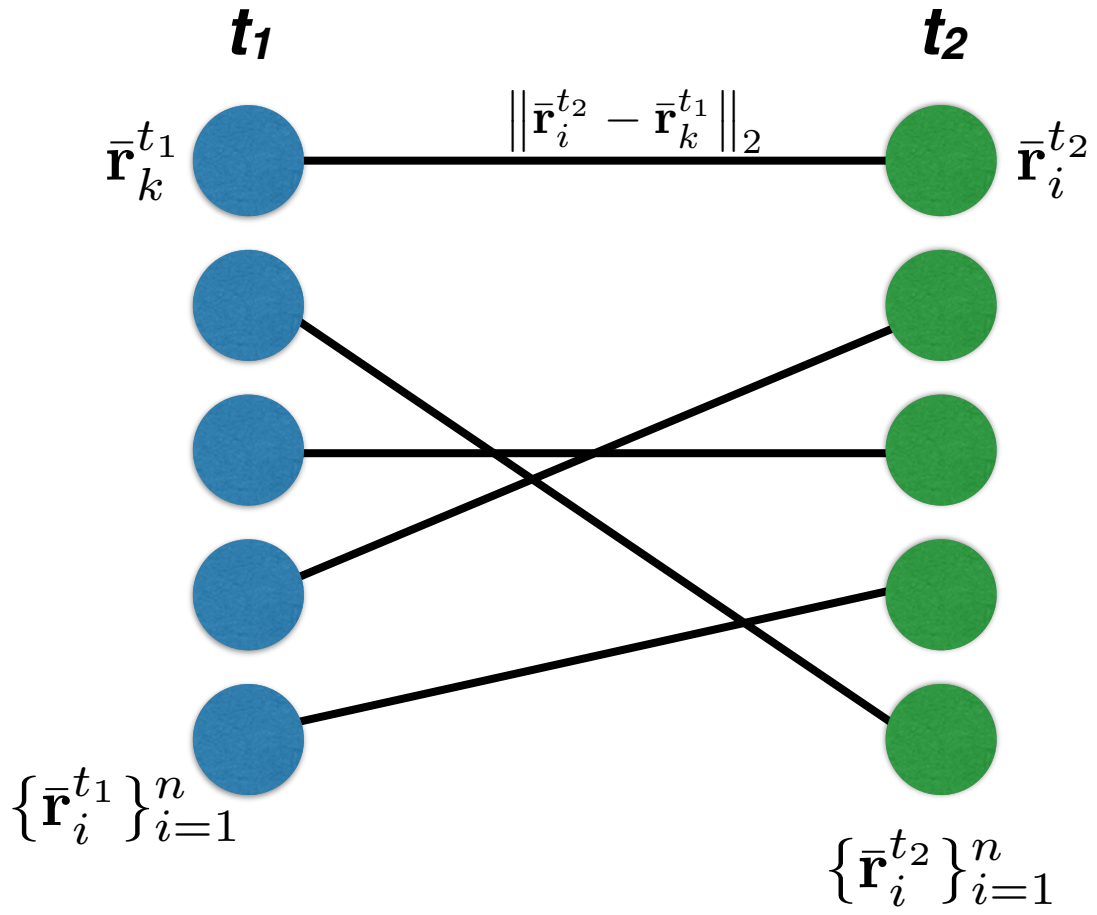


Linkability Attack

Which sample from t_1 corresponds to which sample from t_2?

$$\sigma^* = \arg\min_{\sigma} \sum_{i=1}^{n} \left\| \bar{\mathbf{r}}_{\sigma(i)}^{t_2} - \bar{\mathbf{r}}_i^{t_1} \right\|_2$$

Time complexity: ***O(n³)***



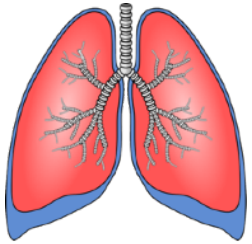
Two Longitudinal Datasets

- Athletes dataset
  - Participants: **29**
  - Points in time: **2** (before and after exercising)
  - Time period: **1 week**
  - Disease: **none**

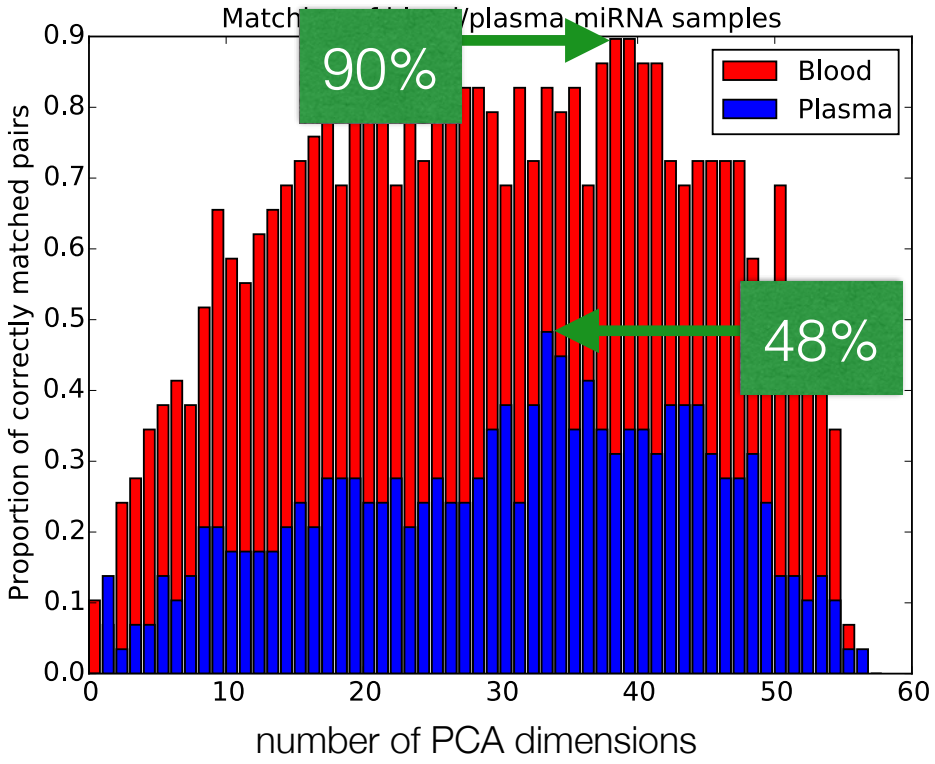
  - **1,189 miRNAs** per sample
    - taken from **blood** and **plasma**

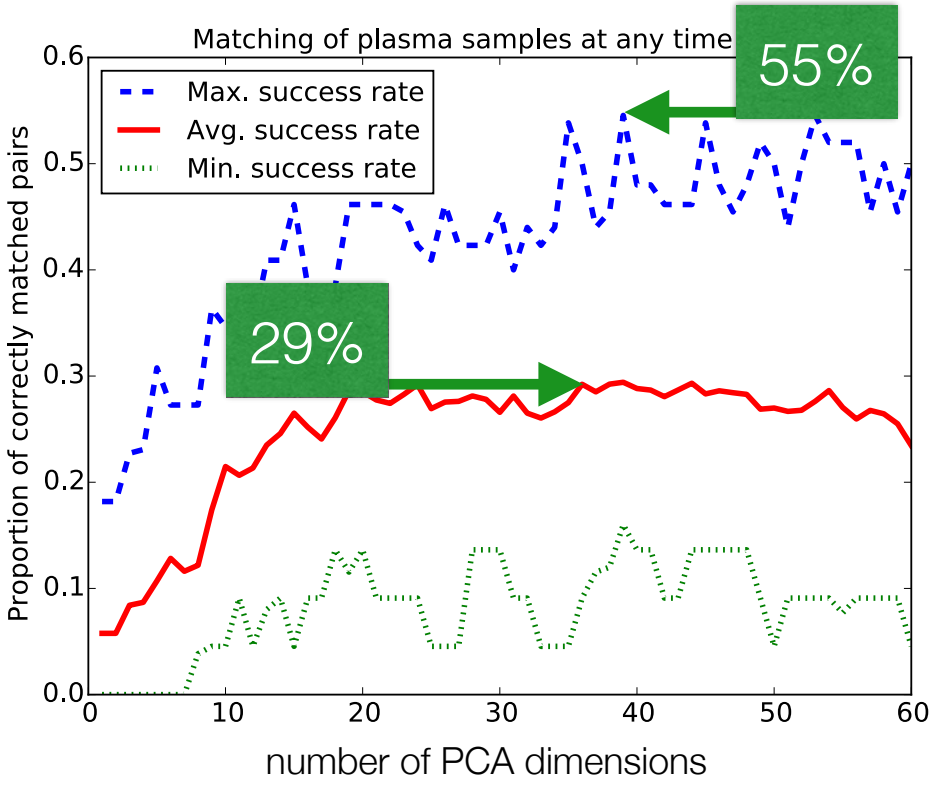
- Lung cancer dataset
  - Participants: **26**
  - Points in time: **8**
  - Time period: **18 months**
  - Disease: **lung cancer**

  - **1,189 miRNAs** per sample
    - taken from **plasma**

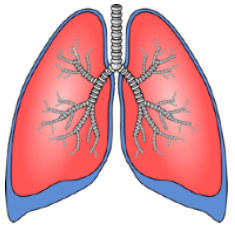


Results





success **up to 90%**
for blood-based samples

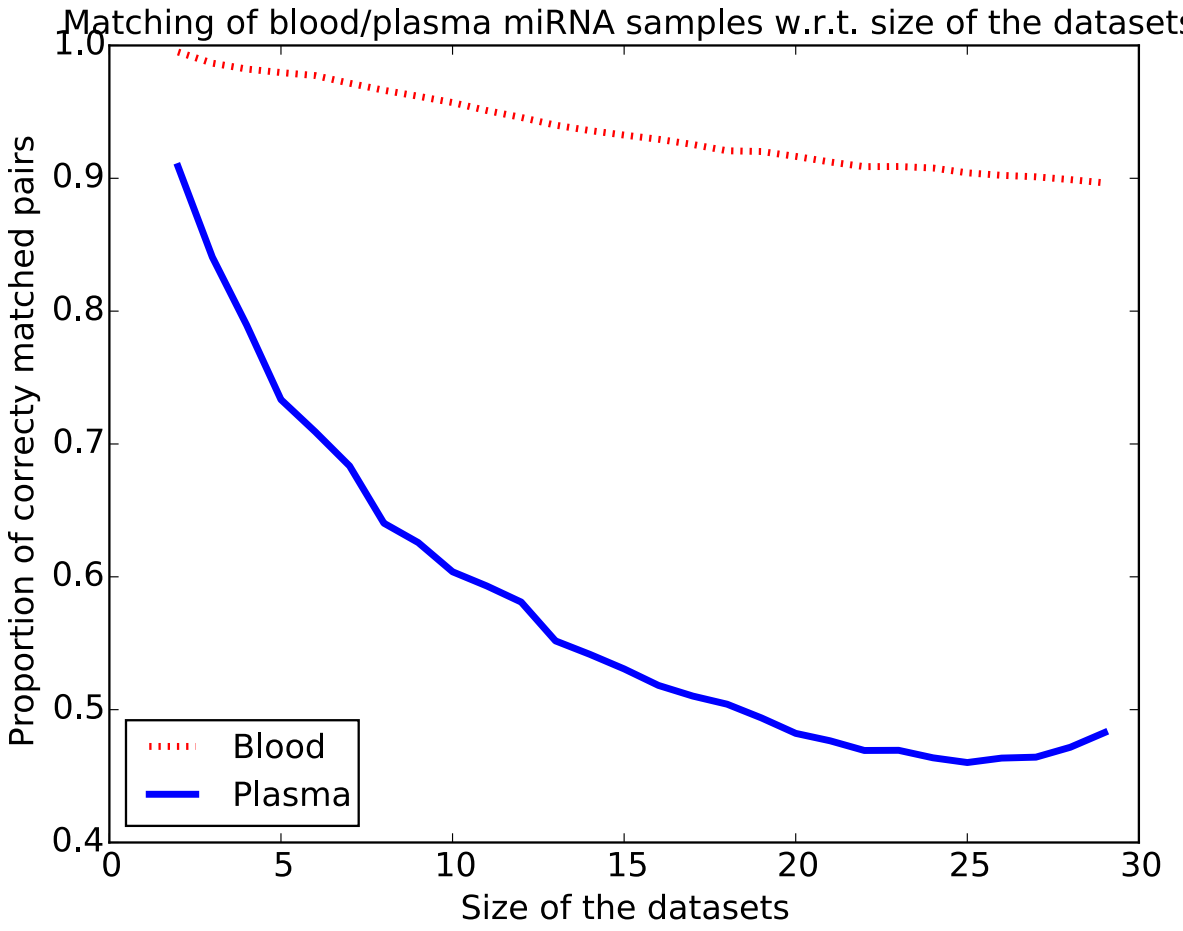


Results

How does the success change with **larger datasets**?

Success **decreases sharply** for plasma-based samples,

but **decreases linearly** for blood-based samples.



Defense Mechanisms

- **Hiding** non-relevant microRNA expressions
  - Sometimes, **randomization is not an option**
  - E.g., for *making a diagnosis in a hospital*
  - Caution: correlations between miRNAs

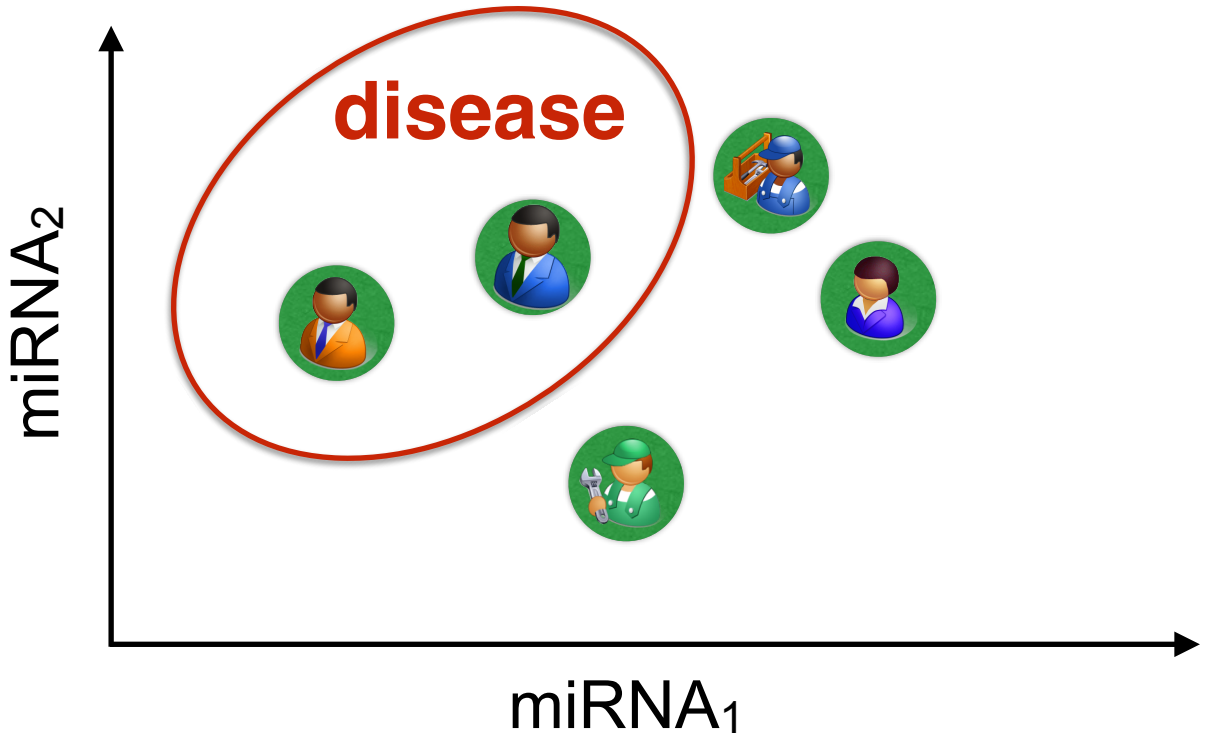
- **Randomizing** the microRNA expression profiles
  - Adding noise in a fully distributed, differentially-private manner
    → providing **epigeno-indistinguishability** (inspired by [4])
  - Noise drawn according to multivariate Laplacian mechanism
  - E.g., for *publishing a dataset used in a study*

[4] Chatzikokolakis et al., **Broadening the scope of differential privacy using metrics**, PETS, 2013

Privacy-Utility Trade-Off

Privacy: prevent **linkability** of samples

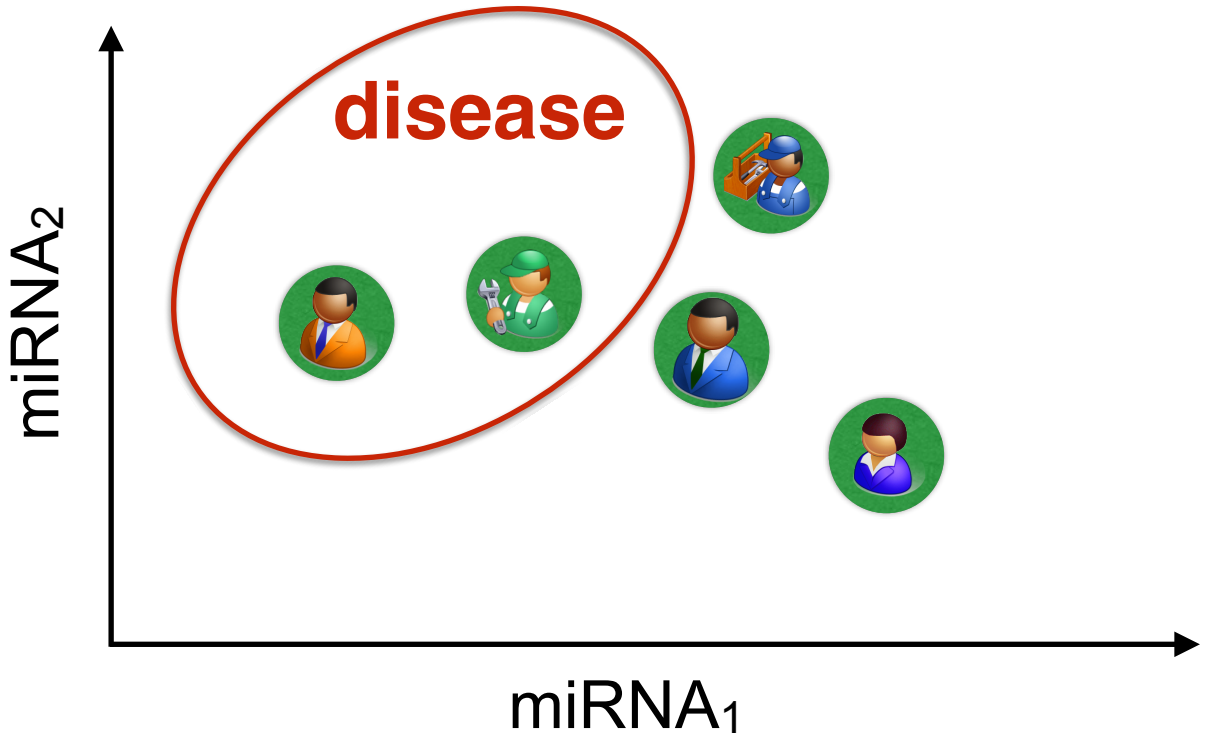
Utility: preserve accuracy of classification as **diseased / healthy**, using a radial SVM classifier



Privacy-Utility Trade-Off

Privacy: prevent **linkability** of samples

Utility: preserve accuracy of classification as **diseased / healthy**, using a radial SVM classifier



Privacy-Utility Trade-Off

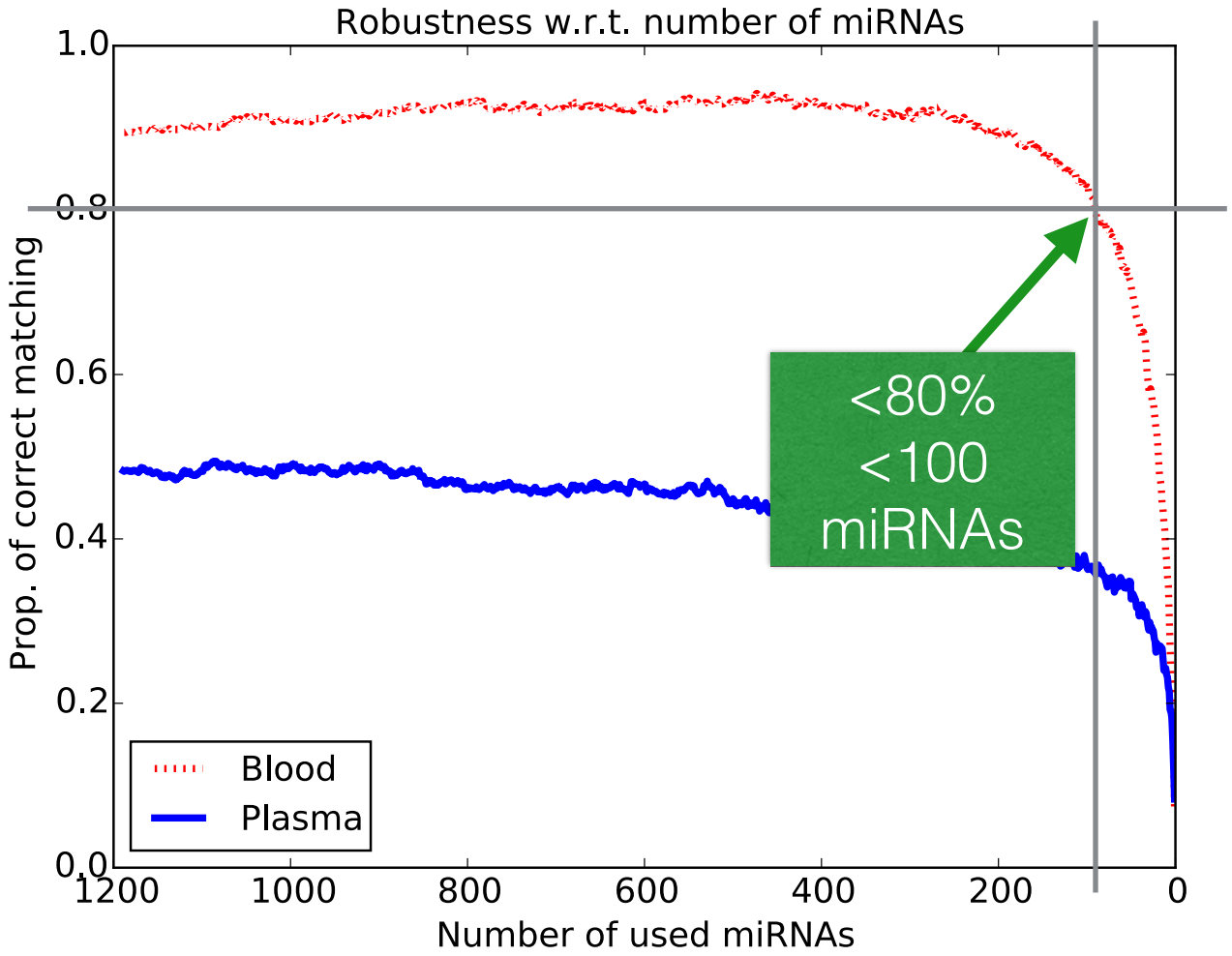
Privacy: prevent **linkability** of samples

Utility: preserve accuracy of classification as **diseased / healthy**, using a radial SVM classifier

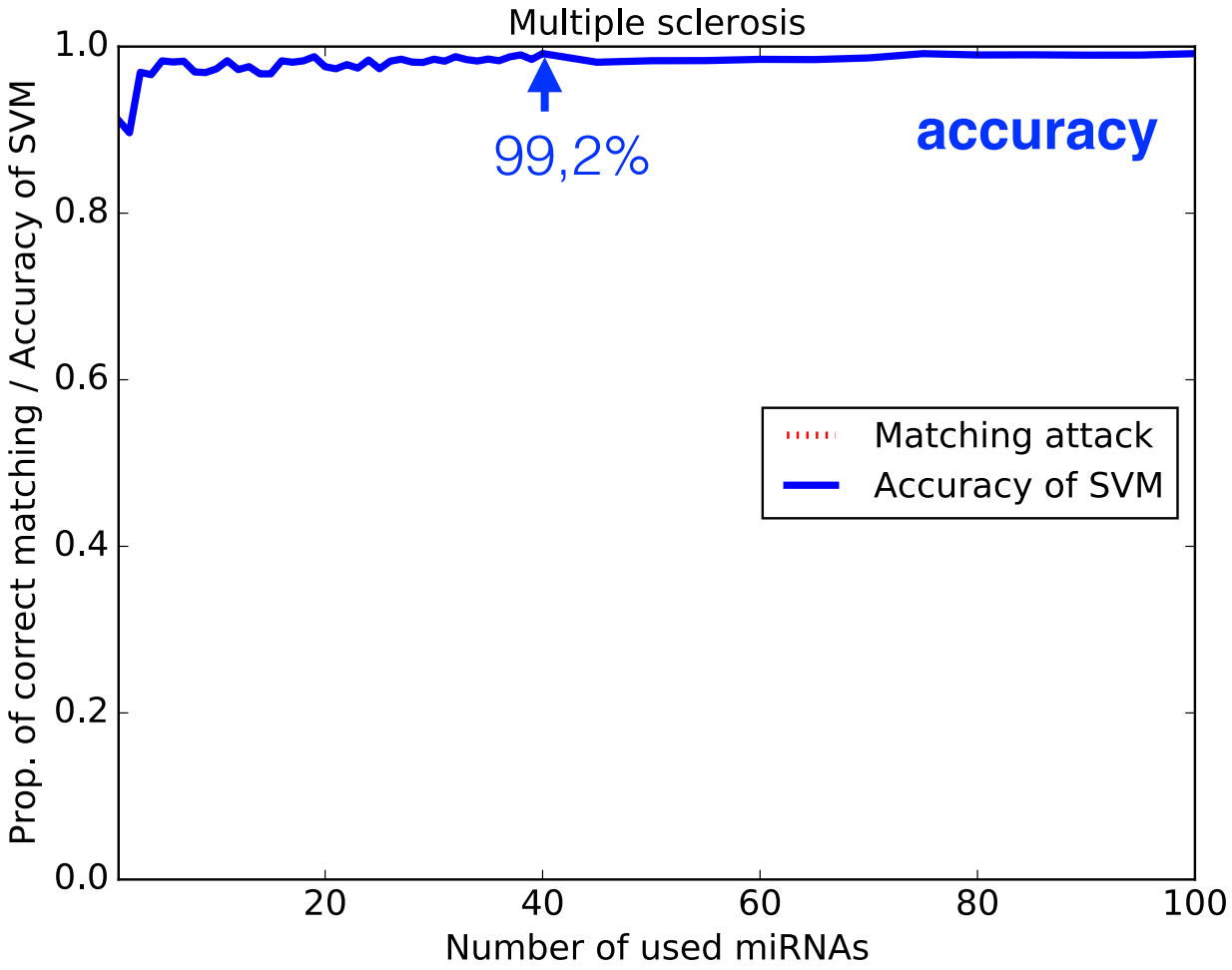
Another dataset for exploring utility:

1000+ participants,
19 diseases,
1 time point

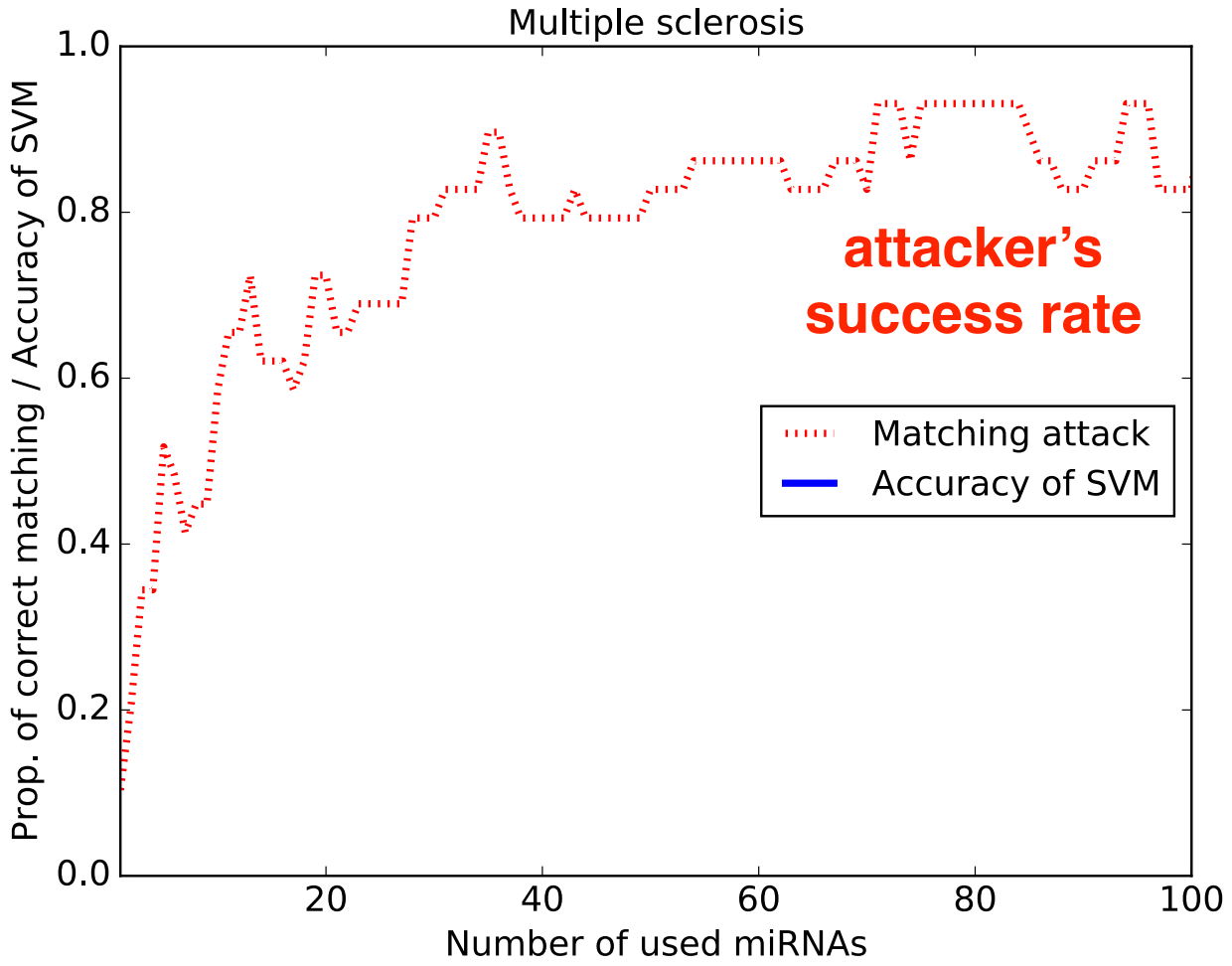
Hiding miRNAs - Results



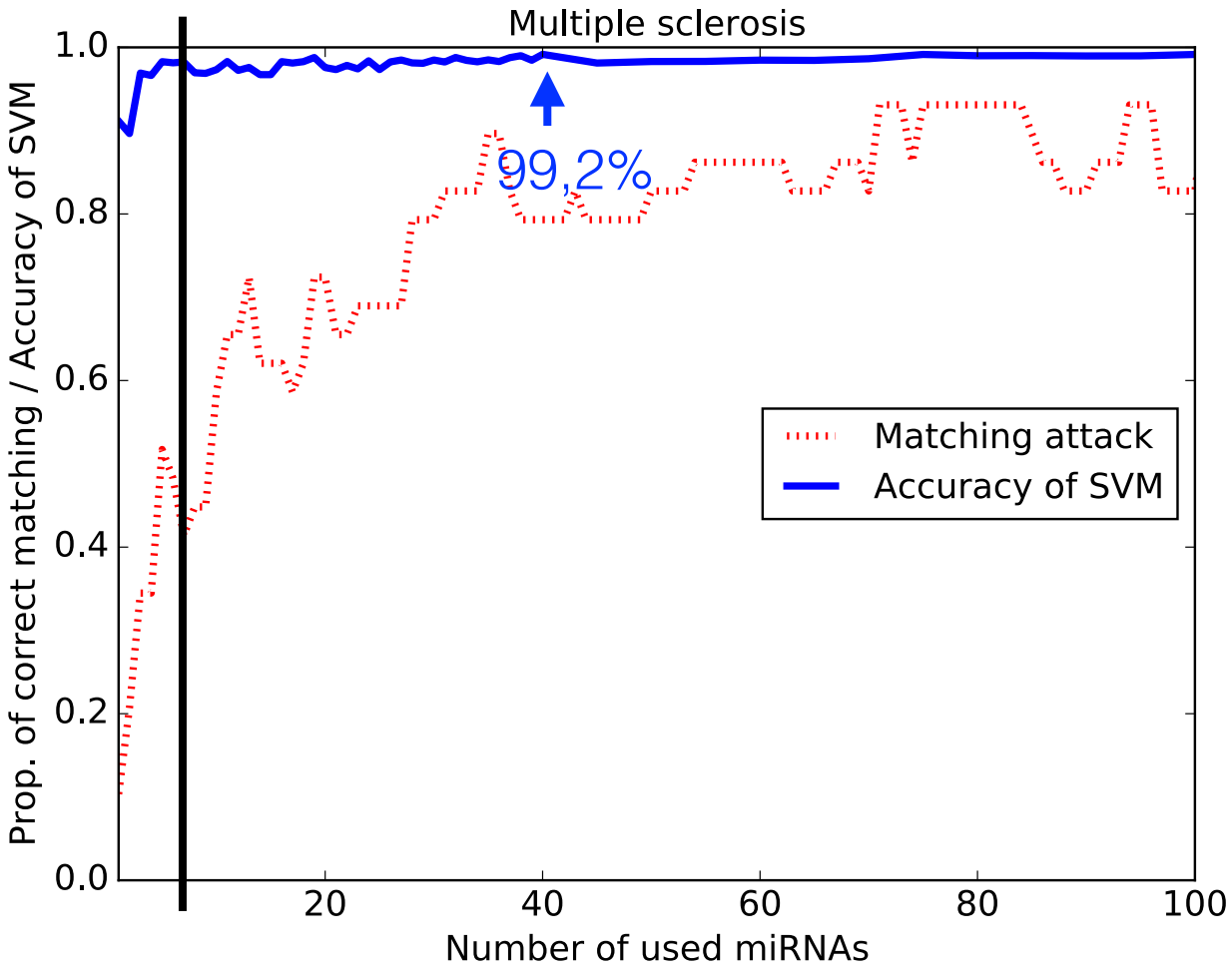
Hiding miRNAs - Results



Hiding miRNAs - Results



Hiding miRNAs - Results

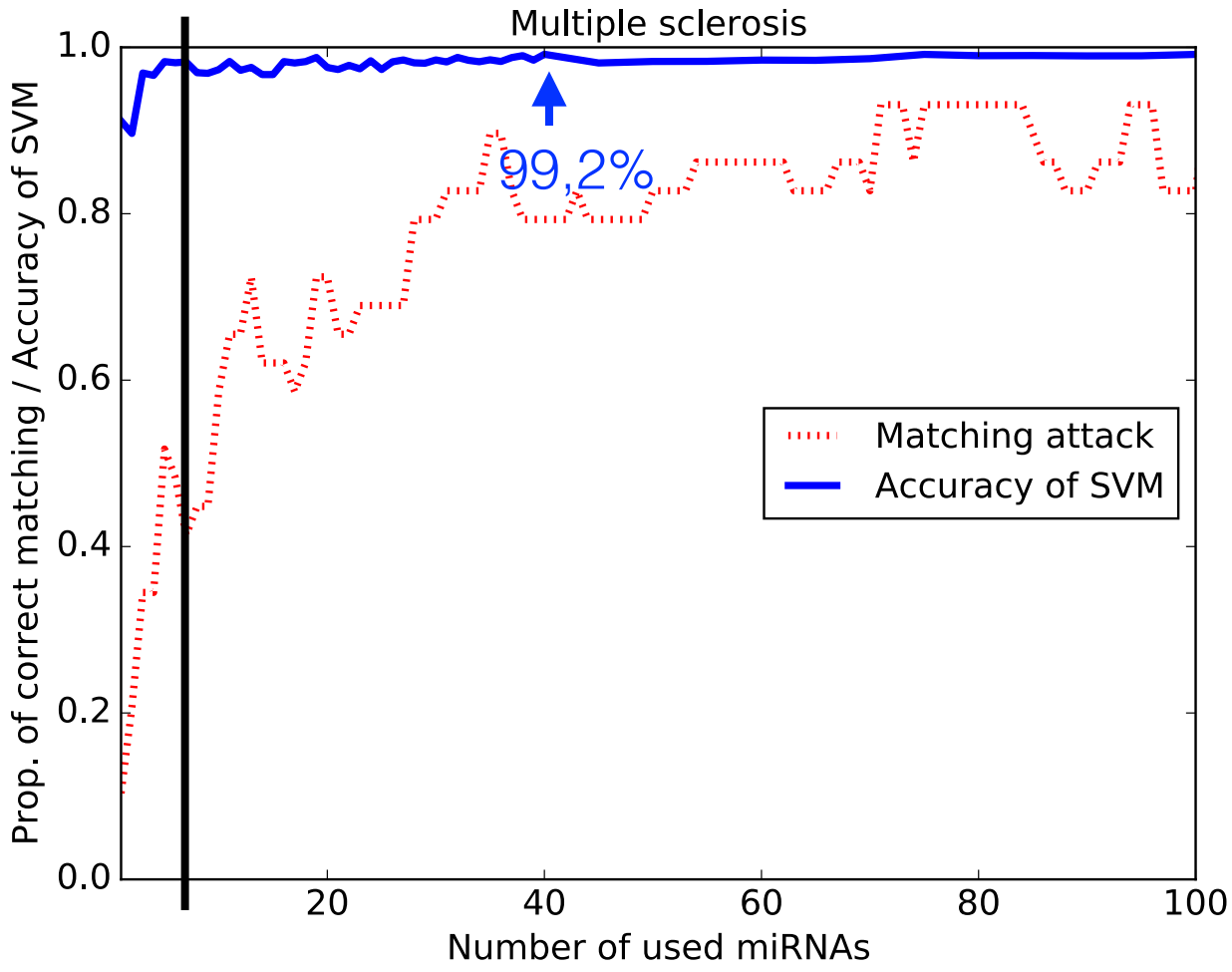


Hiding miRNAs - Results

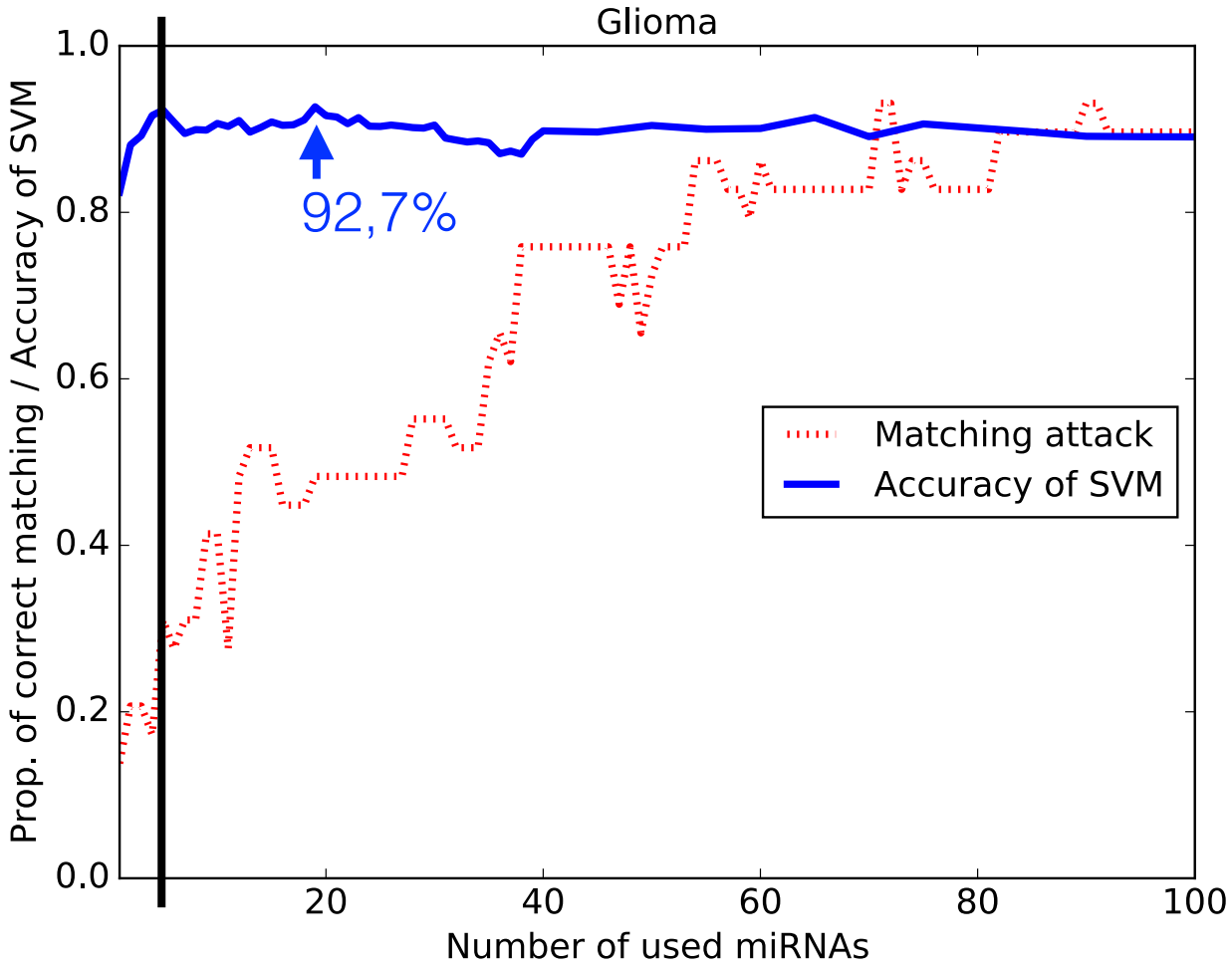
Trade-off at 7 miRNAs

Attack success decreased (relative to all) by **54%**

SVM accuracy decreased (relative to max) by only **1%**



Hiding miRNAs - Results

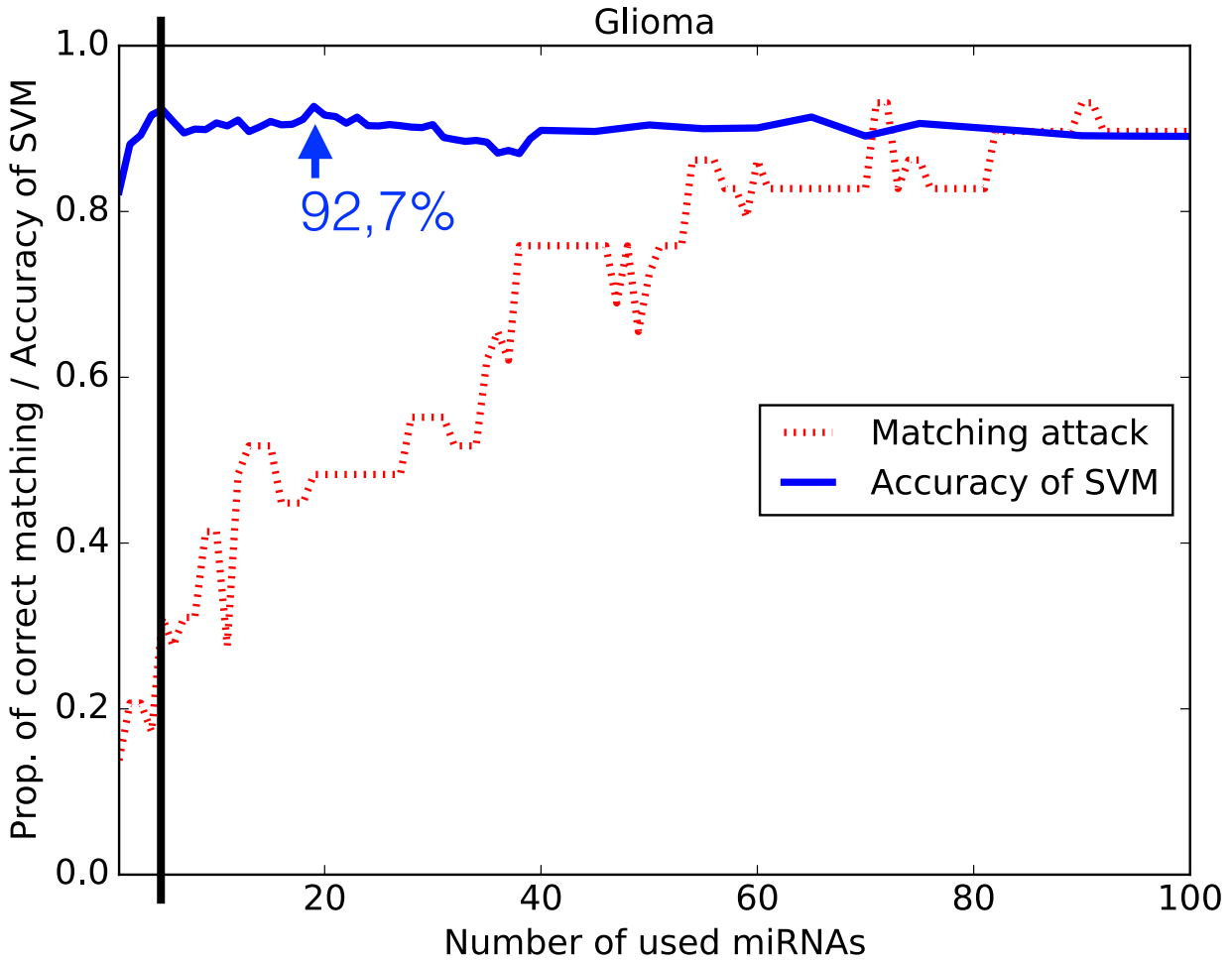


Hiding miRNAs - Results

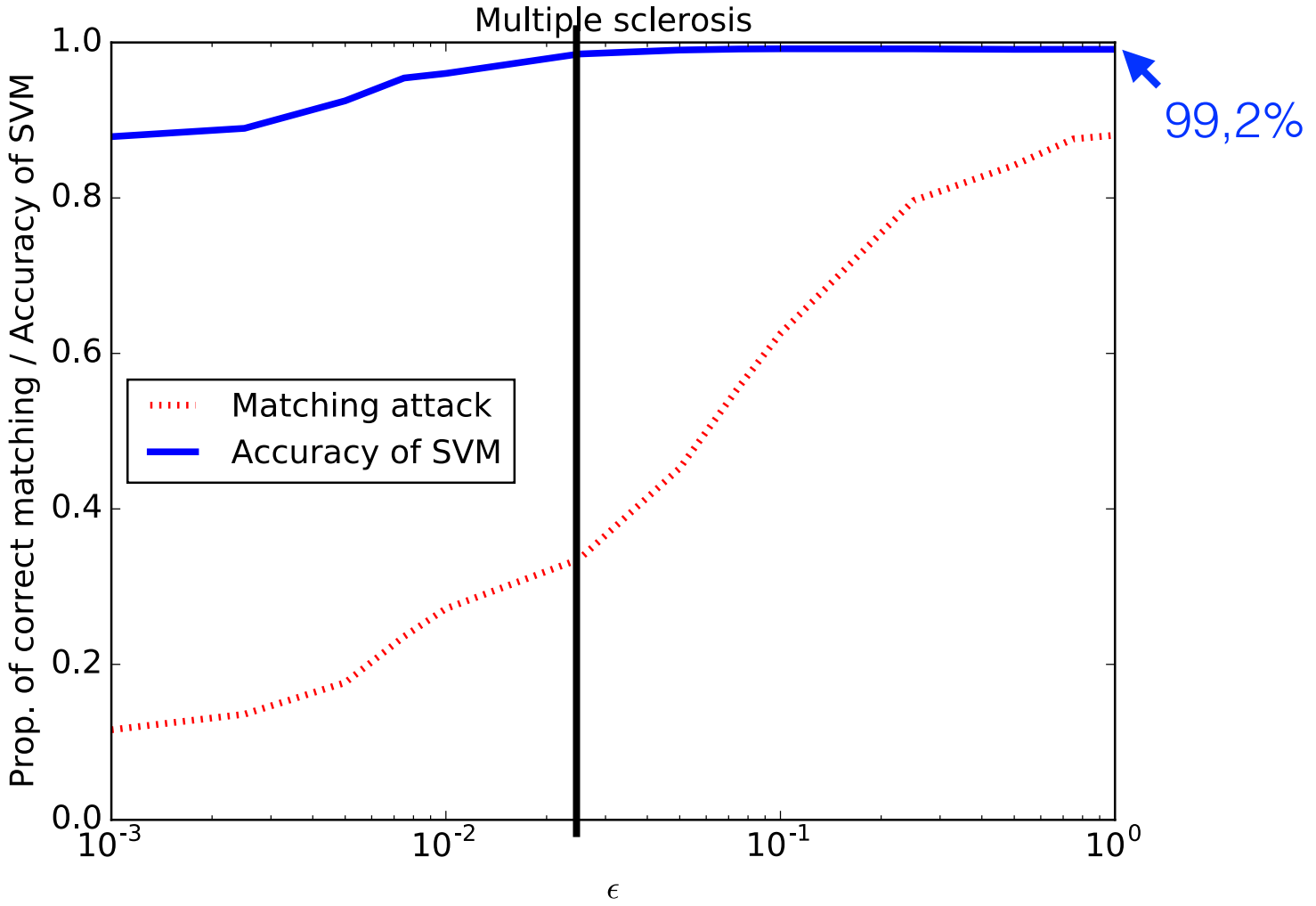
Trade-off at 4 miRNAs

Attack success decreased (relative to all) by **80%**

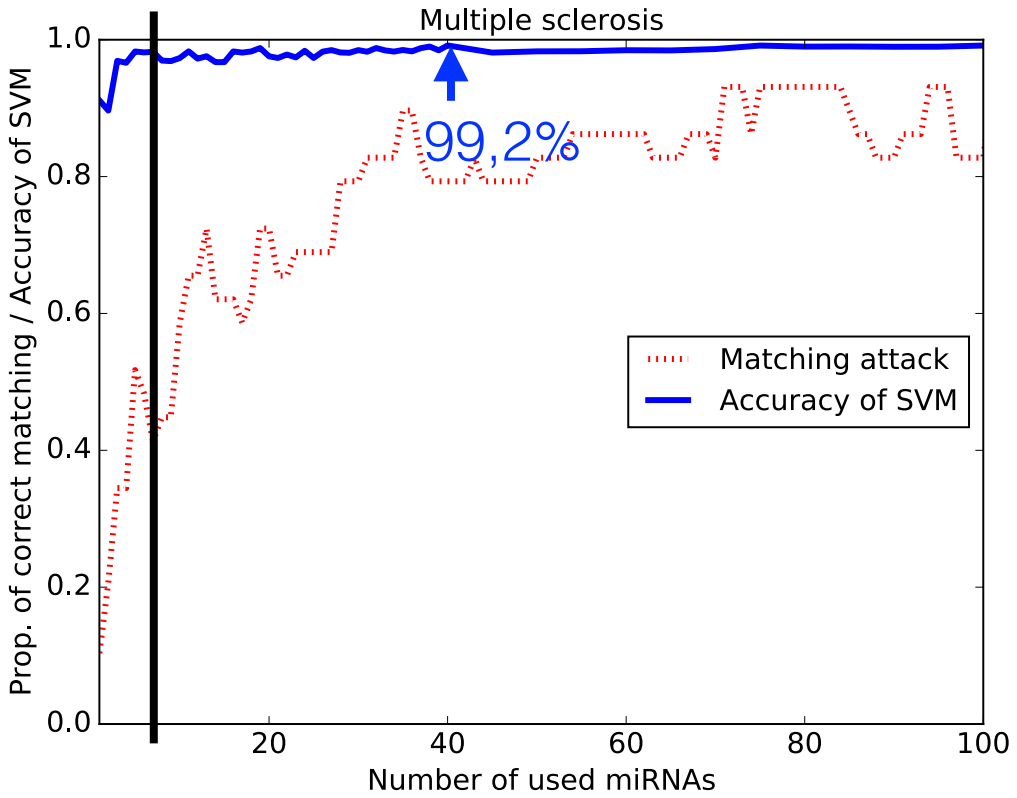
SVM accuracy decreased (relative to max) by only **1%**

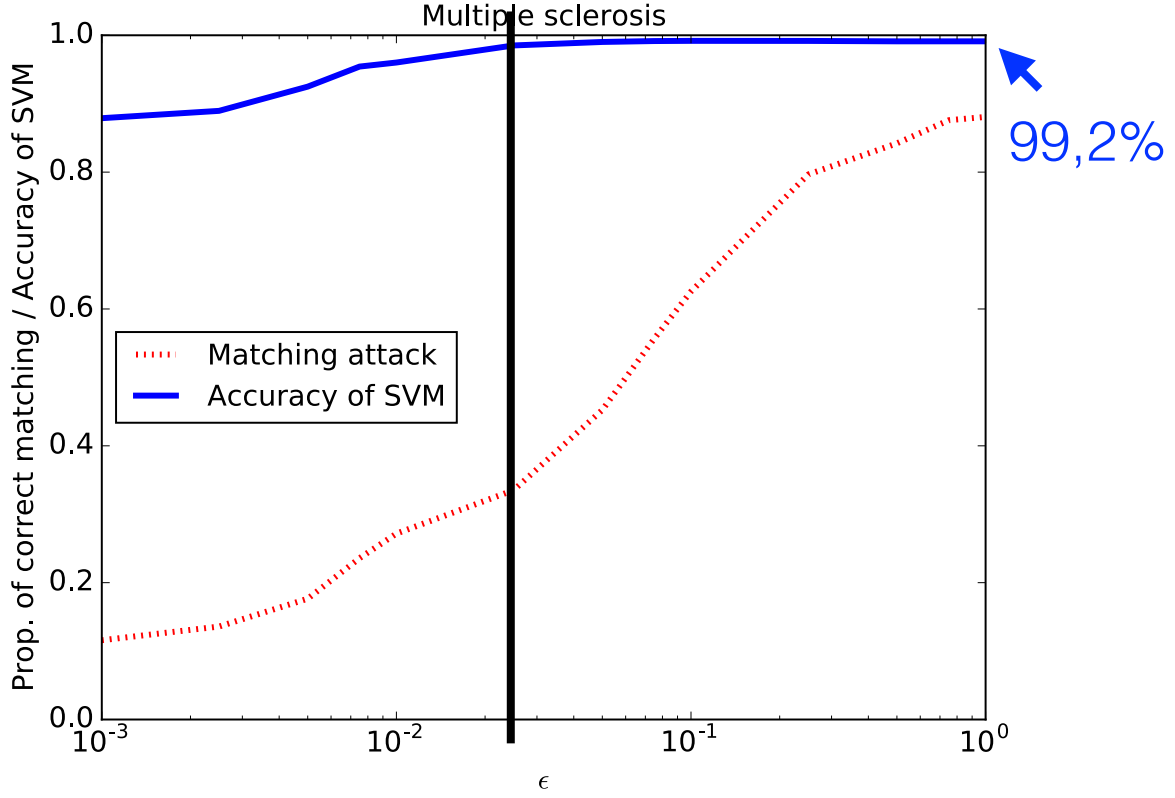


Probabilistic Sanitization - Results



Probabilistic Sanitization - Results



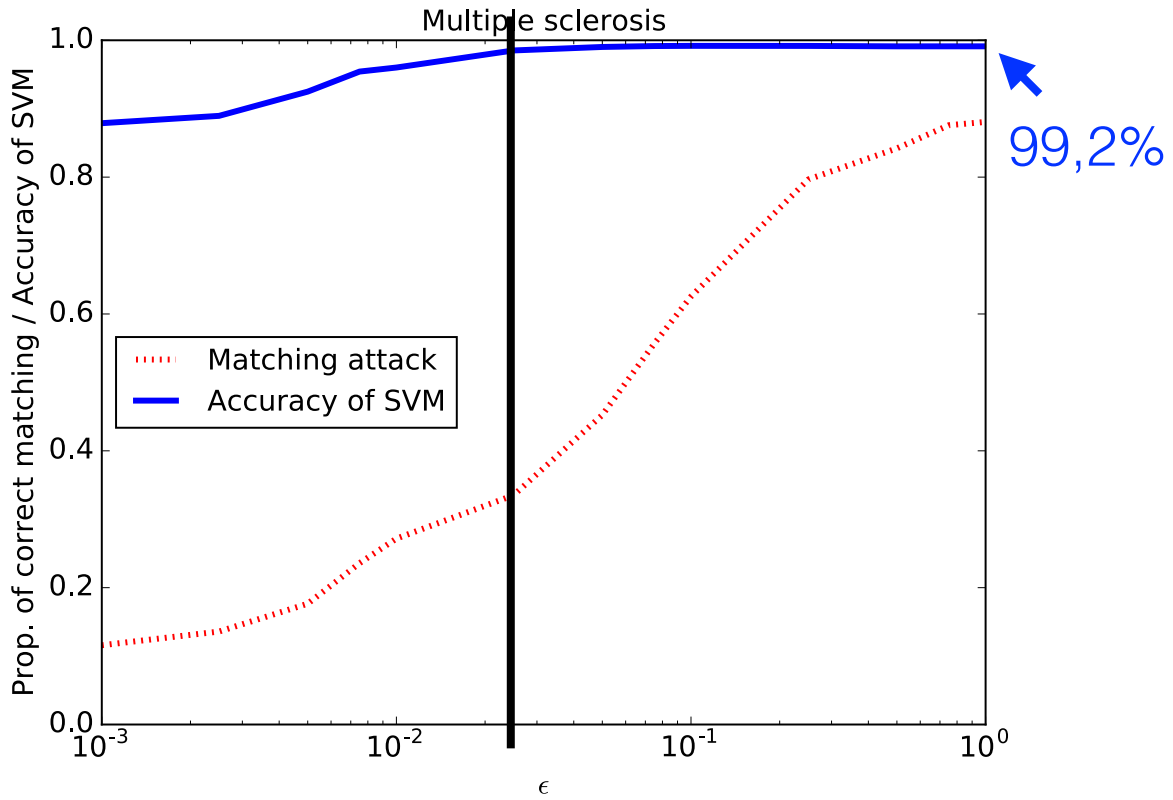


Probabilistic Sanitization - Results

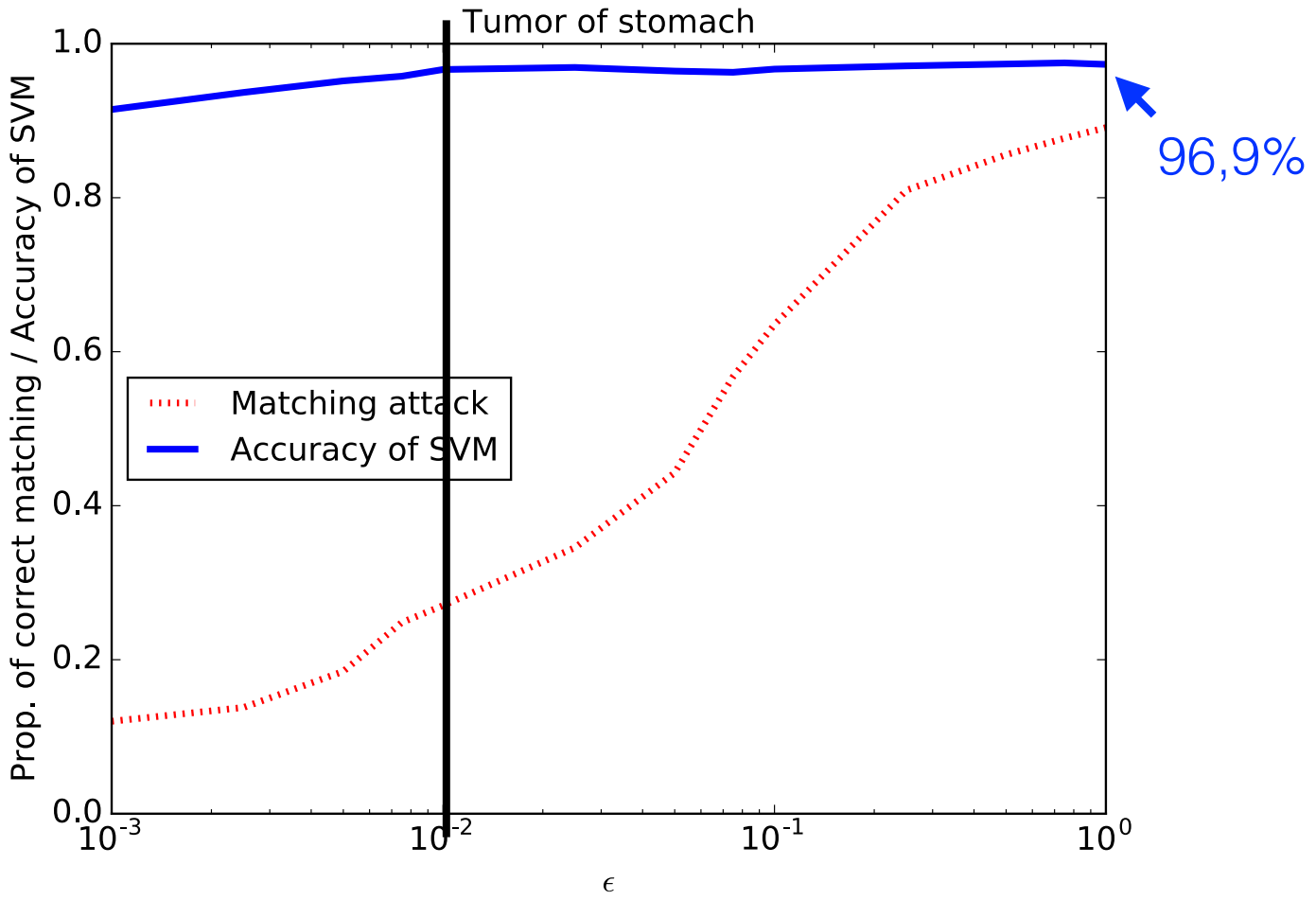
Suitable balance at ϵ = 0.025

Attack success decreased (relative to all) by **63%**

SVM accuracy decreased (relative to max) by only **0.65%**



Probabilistic Sanitization - Results

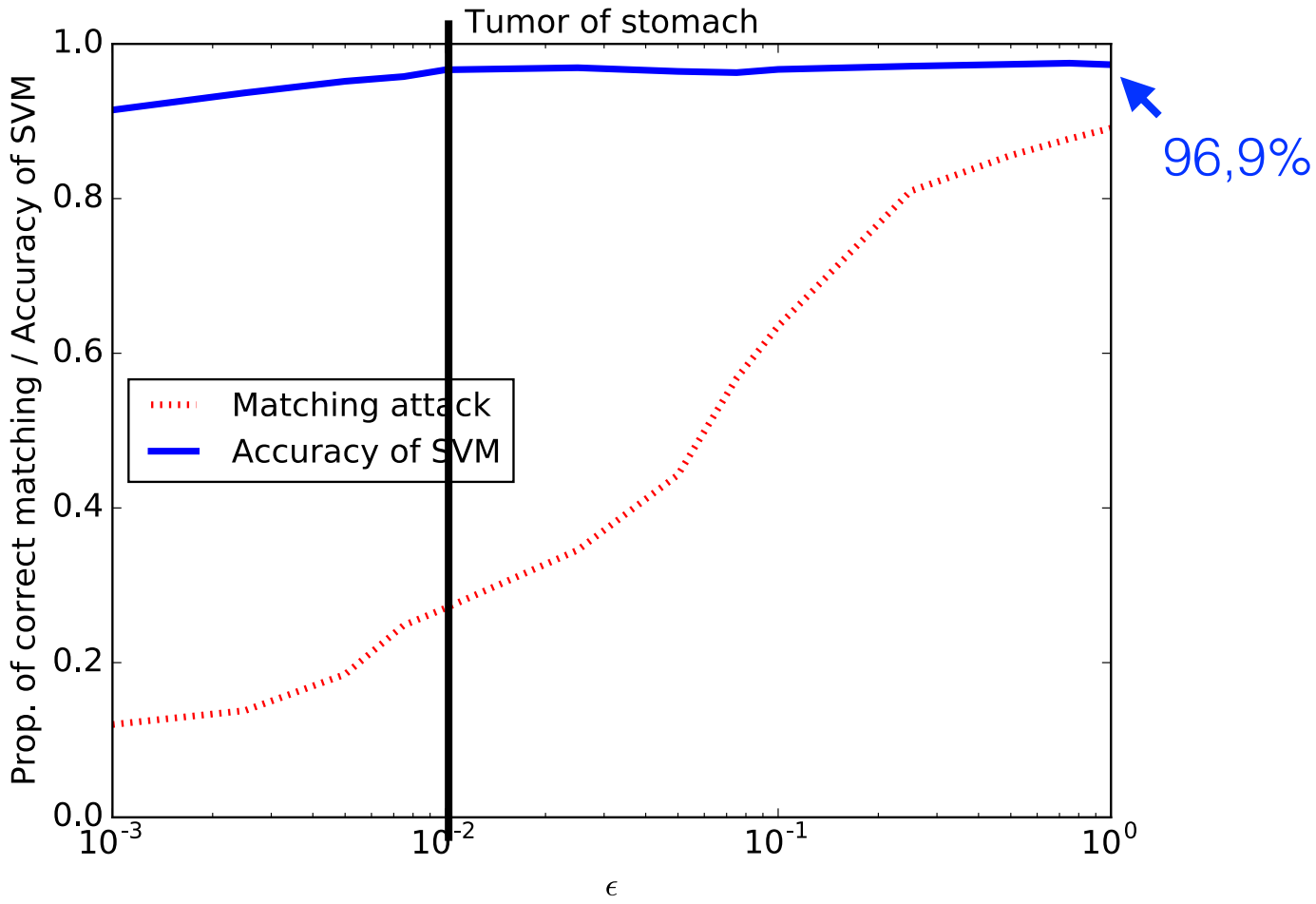


Probabilistic Sanitization - Results

Suitable balance at $\epsilon = 0.01$

Attack success decreased (relative to all) by **70%**

SVM accuracy decreased (relative to max) by only **0.2%**



Summary

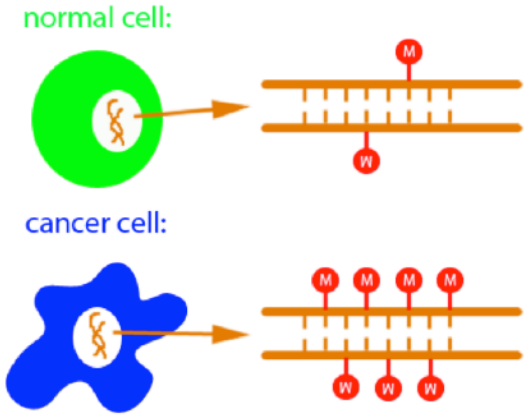
- MicroRNA expression profiles are prone to **temporal linkability attacks**
  - Up to **90% success** rate for blood-based samples

- How can we **protect** this type of biomedical data?
  - Noising can reduce the success rate to almost random guessing

- How does the protection affect the **utility** of the data?
  - **Noising** usually provides a **better** utility-privacy trade-off than **hiding**
  - **Privacy doubles** at utility cost < 1% for most diseases

DNA Methylation Data and Privacy

- **DNA methylation**
  - Very well understood epigenetic mechanism
  - Associated with human health status
    - Hyper-/hypomethylation associated with cancer
    - Smoking mother → child with asthma
  - Sensitive data → privacy must be protected



- Current **privacy practice**
  - Public release on database such as the Gene Expression Omnibus
  - Privacy precautions:
    - Anonymized samples (removal of personal identifiers)
    - Corresponding genomic data not accessible
      - Since the genome can be re-identified using various side channels [5,6]

Gene Expression Omnibus

[5] Gymrek et al., **Identifying personal genomes by surname inference**, Science, 2013
[6] Humbert et al., **De-anonymizing genomic databases using phenotypic traits**, PoPETS, 2015

Re-identifying DNA Methylation Profiles

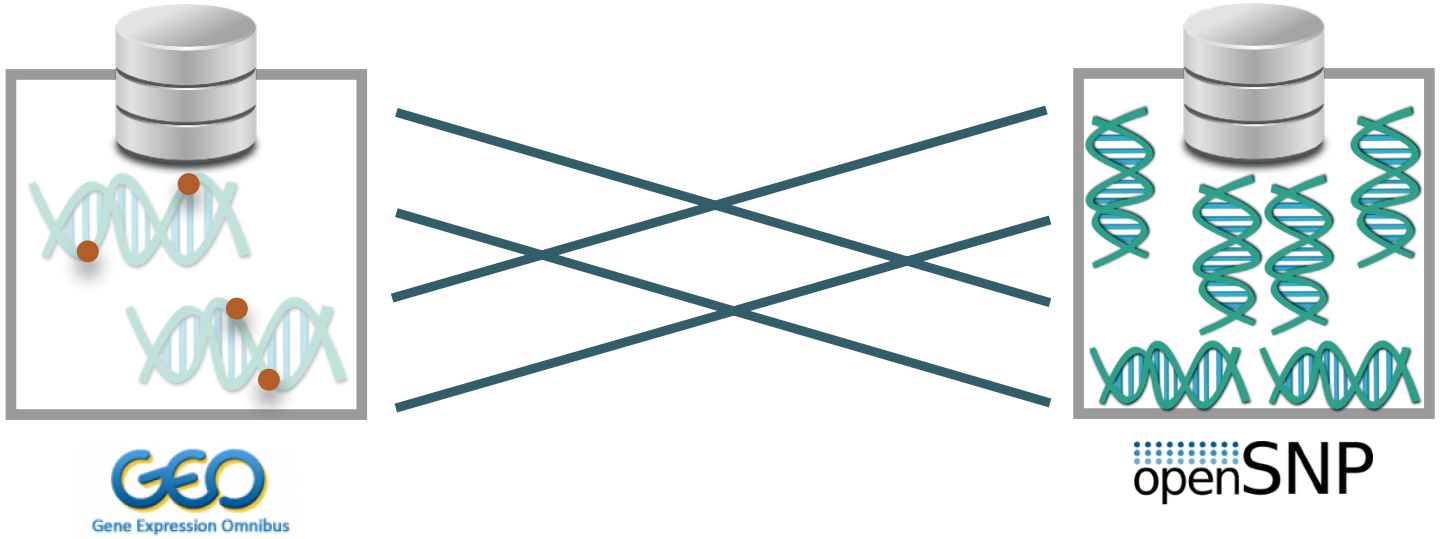
With real identifiers

DNA methylation

Genomic variants

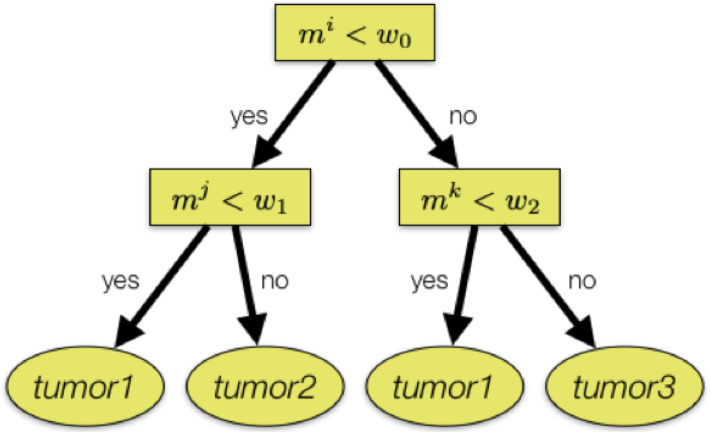
$$\Pr(G_j^i = g_j^i \mid M_j^i) = \frac{p(M_j^i \mid G_j^i = g_j^i)\Pr(G_j^i = g_j^i)}{\sum_{g_j^i} p(M_j^i \mid G_j^i = g_j^i)\Pr(G_j^i = g_j^i)}$$

Re-identifying DNA Methylation Profiles



- **Experimental results**
  - Focusing on **293** methylation regions highly correlated with genotype
  - Between **97.5%** and **100%** of matching accuracy for genotype database of size greater than **2500**
  - Wrongly matched pairs **always rejected** by our statistical test

Defense Mechanism

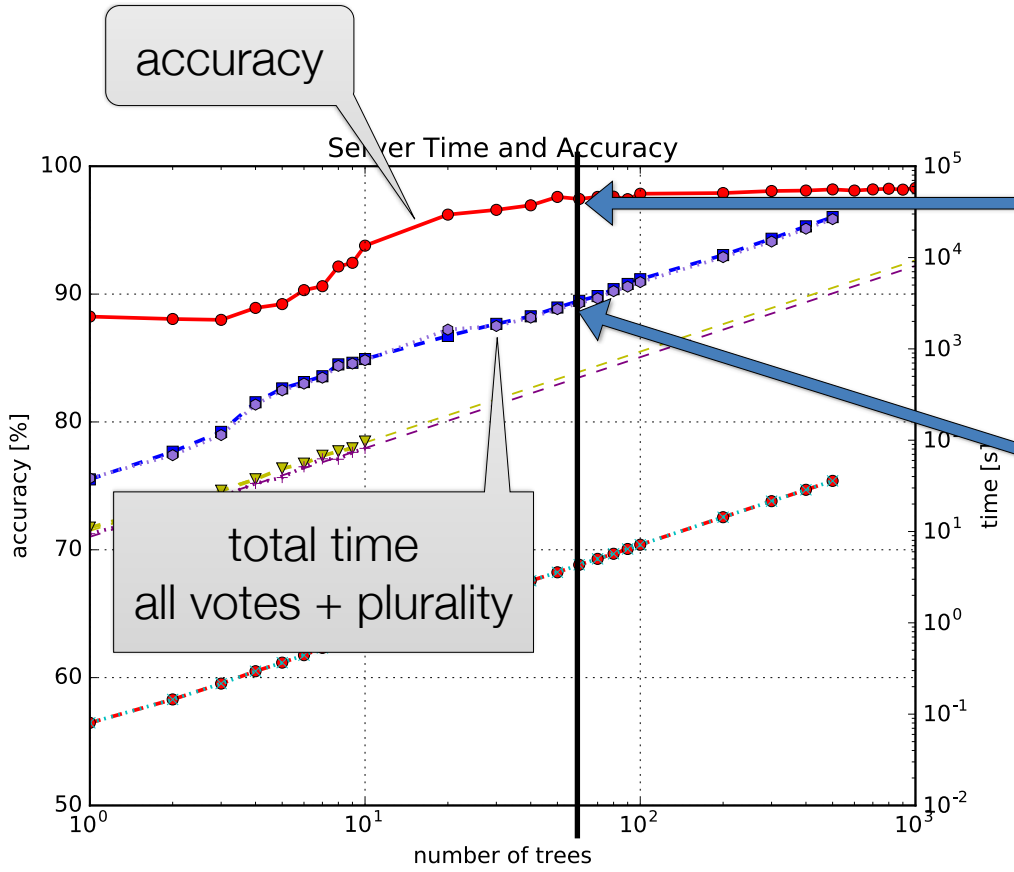


- **Private classification of brain tumors** [7]
  - Random forest classifier
  - **Cryptographic** mechanism: homomorphic encryption
  - Secure under the honest-but-curious adversary model
    - The machine-learning provider does not learn the patient's data
    - The data owner (patient) does not learn the machine-learning model
  - Typical **use case**: clinical setting, or diagnosis by third-party provider

- **Implementation** in C++
  - Classification based on 900 methylation regions
  - 9 tumor subtypes
  - Original random forest model: 1000 trees

[7] Danielsson et al. **MethPed: A DNA methylation classifier tool for the identification of pediatric brain tumor subtypes**, Clinical Epigenetics, 2015

Performance Evaluation



very good accuracy already with **< 100 trees**

100 trees take **less than 2 hours**

1000 trees take **< 12 hours**

Three Key Privacy Attacks - Outline

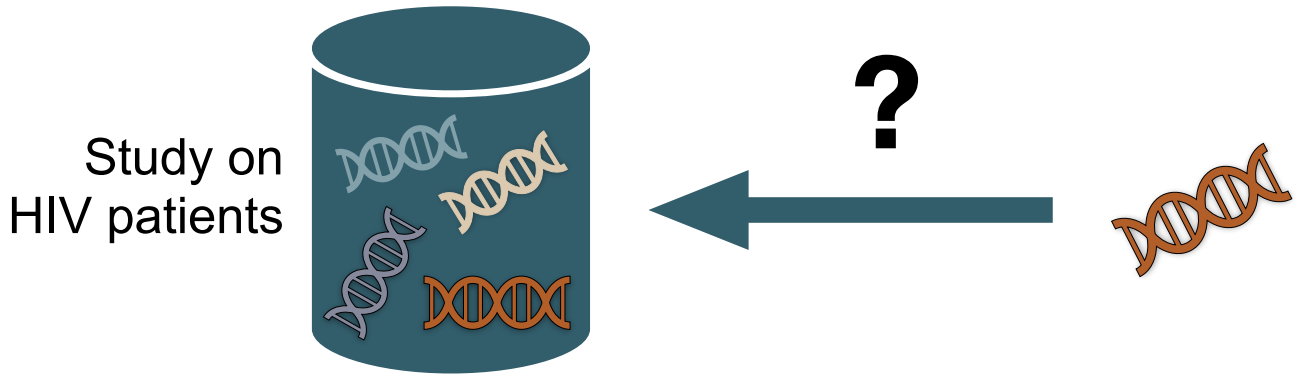
- **Linkability and re-identification**
  - Ability to link at least two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases)
    → which can lead to **re-identification** if one database contains identifiers

- **Attribute inference**
  - Possibility to deduce, with significant probability, the value of an attribute from the values of other attributes

- **Membership inference**
  - Possibility to deduce, with significant probability, that a specific sample is part of a dataset

Membership Inference Attacks

- Ability to infer that a **certain target** is in a **specific dataset**



- **Attack gaining momentum**
  - Membership inference against **location data**: **best paper** at NDSS 2018
  - Membership inference against **ML training data**: **Caspar Bowden award** for outstanding research in privacy 2018
  - Membership inference against **DNA methylation Beacons**: **best paper** at NDSS 2019

Membership Inference Against Databases

Set $\mathcal{T}$ of *m* miRNA expressions of *n* individuals

$\mathcal{T} \in \mathbb{R}^{n \times m}$

Example: study of associations between miRNAs and a disease

Summary statistics about a subset of the *m* miRNAs [e.g., **mean** values]

Curious analyst

having access to the actual miRNA expressions of a victim *v*

$\mathbf{x}^v \in \mathbb{R}^m$

Objective: determine if *v* is member of the pool $\mathcal{T}$ by using the the summary statistics and $\mathbf{x}^v$

Main differences between genomic data and miRNA expressions:

- MiRNA expressions are in $\mathbb{R}^m$ whereas genomic variants are in $\{0,1,2\}^m$
- Dimensionality of miRNA expression profiles is orders of magnitude smaller than the one of genomic data ($m \approx 10^3$ vs. $m \approx 10^8$)

Two Statistical Tests for Membership Inference

1. **L1-distances** difference *D* (Homer et al.'s idea [8]):

At miRNA *j*: $D(x_j^v) = |x_j^v - \mu_j| - |x_j^v - \hat{\mu}_j|$

mean expression in the **reference population**

mean expression in the **pool**

→0 if *v* is not in the pool
→ >0 if *v* is in the pool

⇒ **t-test**: *v* is part of the pool if the sum is strictly greater than a threshold

2. **Likelihood-ratio (LR)** test
    1. Provide the maximum achievable power (true-positive rate) for a given false-positive level
    2. Sum over all miRNAs: $LLR = \sum_{j=1}^{m} \frac{(x_j^v - \mu_j)^2}{2\sigma_j^2} - \frac{(x_j^v - \hat{\mu}_j)^2}{2\hat{\sigma}_j^2} + \log \frac{\sigma_j}{\hat{\sigma}_j}$

[8] Homer et al., **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays**. PLoS Genetics, 2008

Exact LR Test - Analytical Result

- **Theorem**:
  - The relation between the **power** β, the **false-positive rate** α, the **number of exposed miRNAs *m***, and the **number of individuals in the pool *n*** is:

$$z_\alpha + z_{1-\beta} \approx \sqrt{\frac{2m}{n^2}}$$

where z_x is the 100(1-*x*)th percentile of the standard normal distribution.

Take-home messages:

* For a successful attack, the number of exposed miRNA statistics *m* has to increase with the **square** of the size of the pool *n*
* With genomic data, it had to increase only **linearly** with *n*: $\sim\sqrt{m/n}$ [9]
* From a privacy perspective, miRNA expressions are less prone to membership inference than genomic variants

[9] Sankararaman et al., **Genomic privacy and limits of individual detection in a pool**, Nature genetics, 2009

Experimental Setup

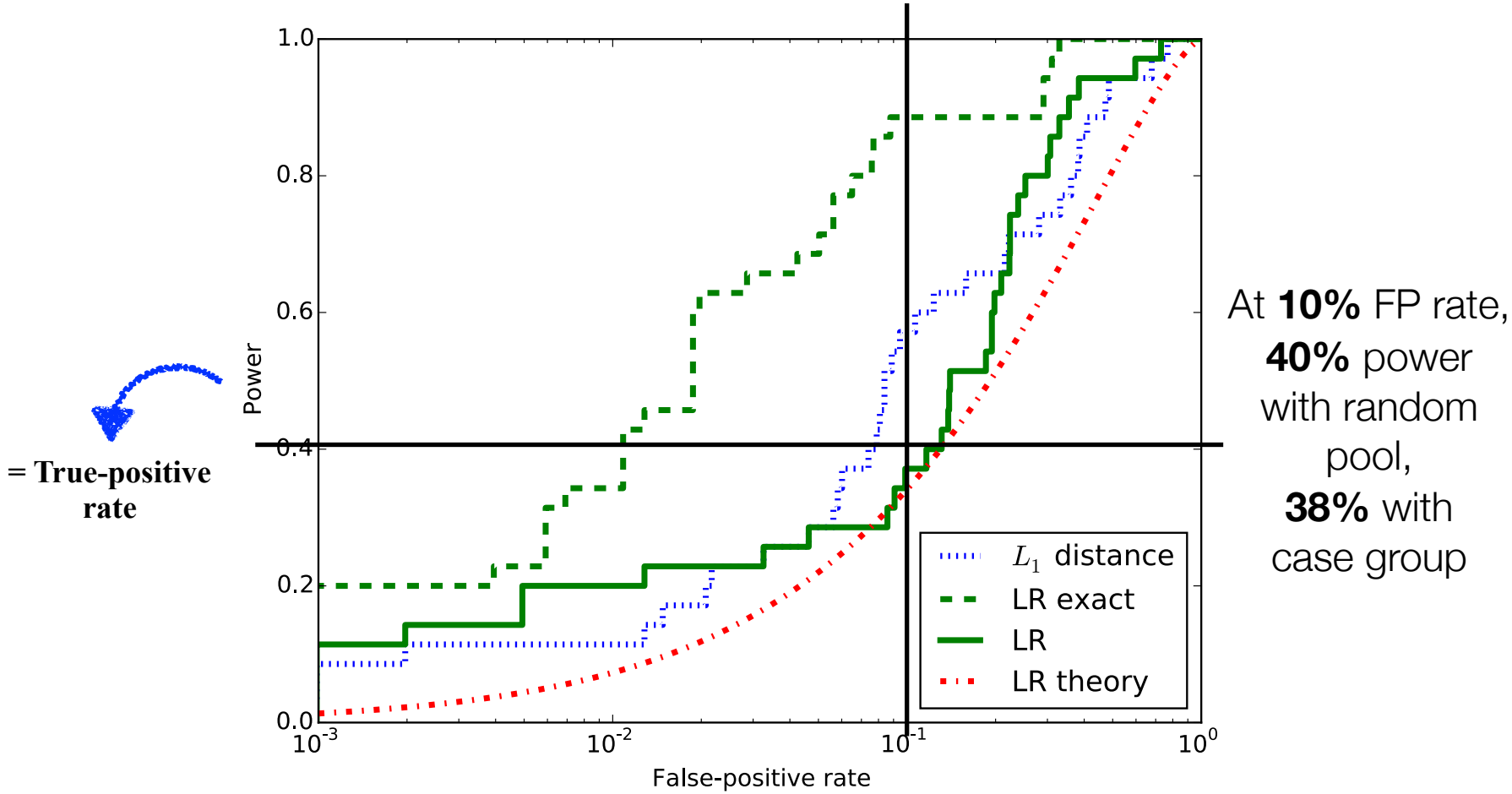
- **Dataset**:
  - **1,049 individuals** in total
    - **19 different severe diseases**
    - … from 124 individuals with Wilms tumor to 13 with stomach tumor
    - Reference population statistics computed over the whole 1,049 individuals
  - Dataset publicly available in the GEO database (ref. GSE61741)

- Two different **experimental settings**:
  - Pool = *n* randomly selected individuals among 1,049
  - Disease-specific pool (aka *case group*), from *n*=13 to *n*=124

35 Individuals in the Pool

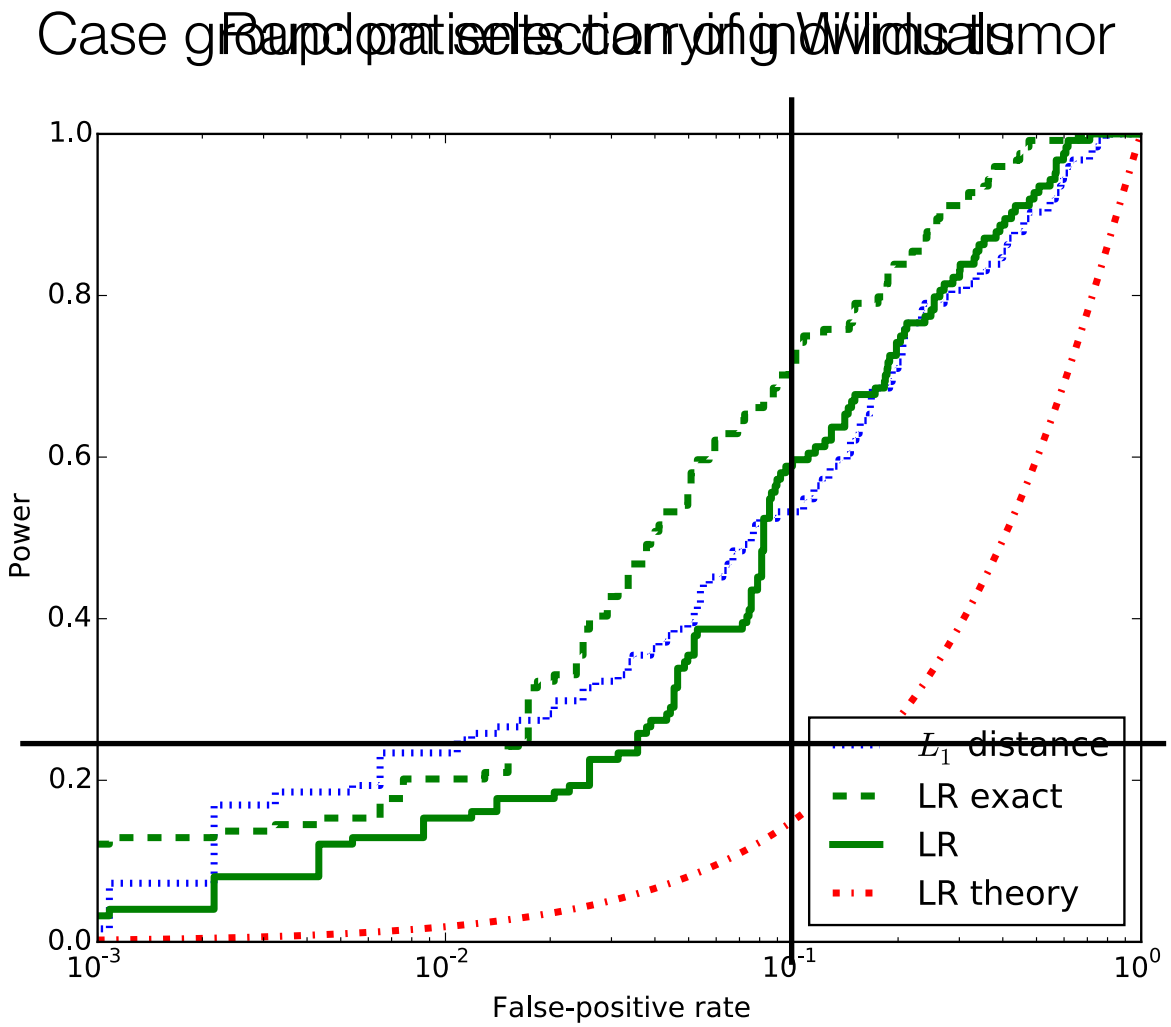
Case group: patients with benign prostatic hyperplasia

Random selection of individuals



At **10%** FP rate, **40%** power with random pool, **38%** with case group

124 Individuals in the Pool



= **True-positive rate**

At **10%** FP rate, **25%** power with random pool, **60%** with case group

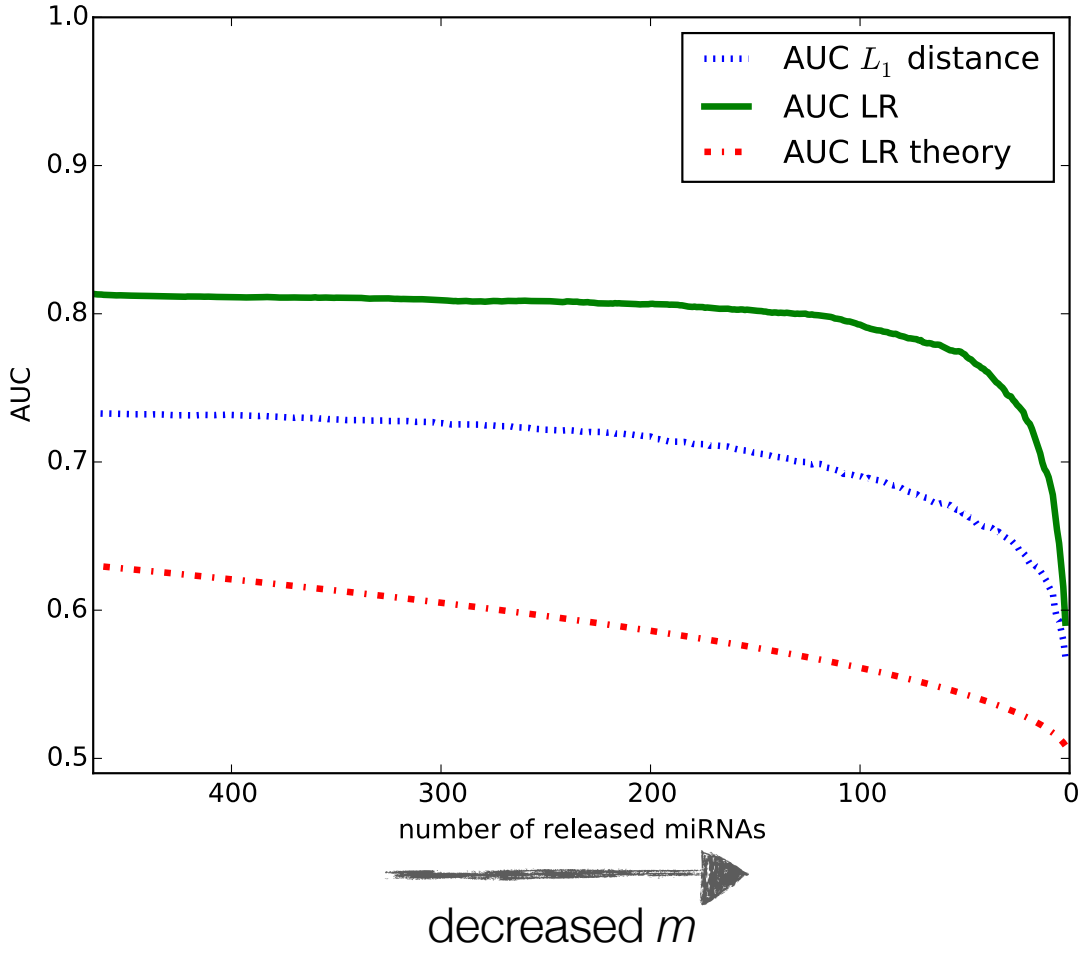
Defense Mechanisms

- **Releasing only a subset of miRNA expression statistics**
  - Publish from m=1 to m=466 miRNA expression means

- **Adding noise in a differentially private manner**
  - Laplace distribution suitably scaled to achieve ε-differential privacy
    => scale = $\Delta(f_{avg})/\varepsilon$, where $\Delta(f_{avg})$ is the sensitivity of f_{avg}
  - $\Delta(f_{avg}) = \sum \delta_i/n$, where δ_i is the range of miRNA i's expression
  - The **amount of noise** to be added **increases** proportionally to the scale, i.e., to $\sum \boldsymbol{\delta_i/\varepsilon n}$.
  - This noise is added to every miRNA expression means before release
  - **Problem**: the range δ_i of some miRNA expressions is >10,000

Hiding Mechanism

Patients carrying prostate cancer (n=65)



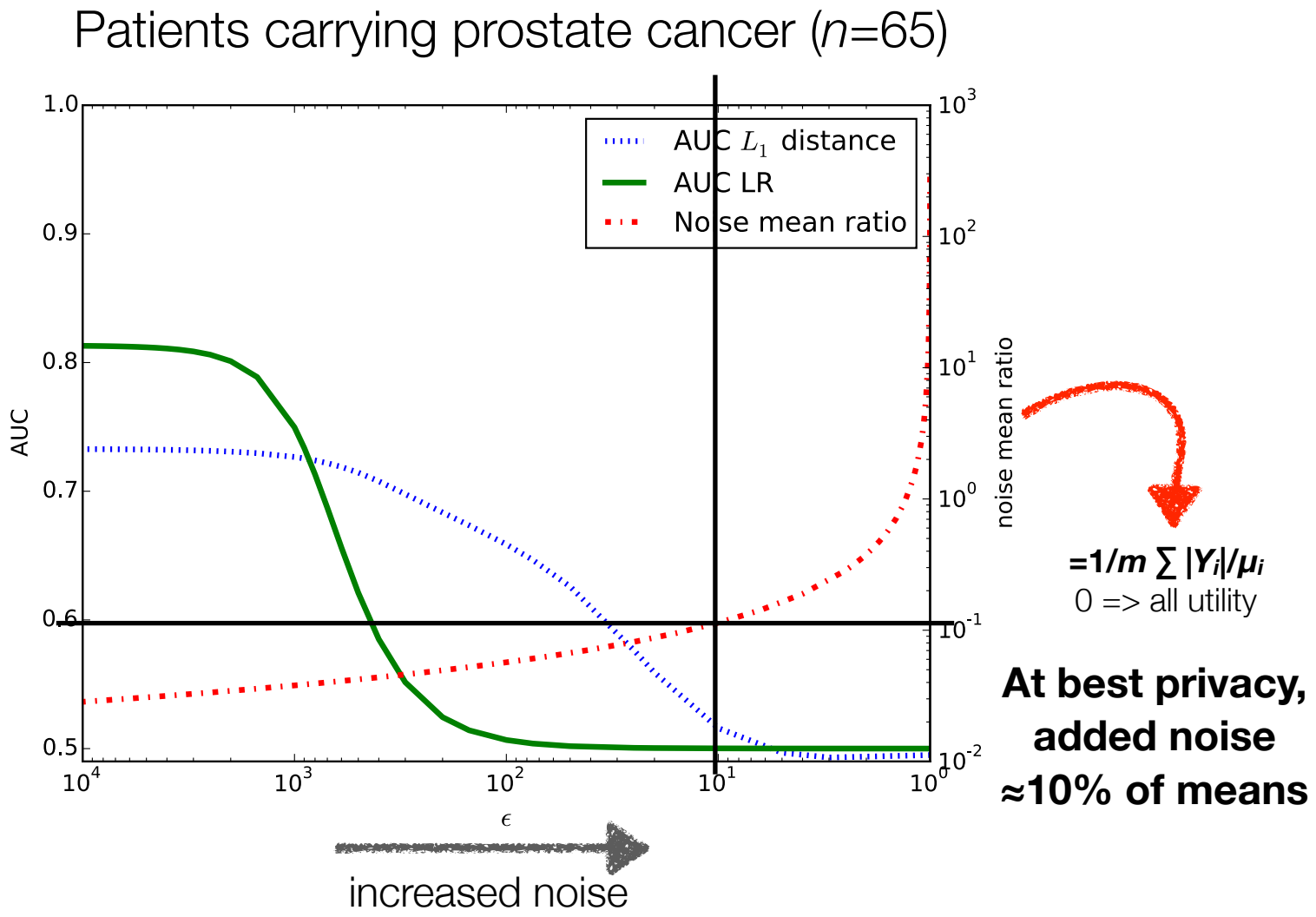
area under the ROC curve

1 => no privacy

0.5 => best privacy

decreased m

Differentially Private Mechanism



Summary

- Experimental results show that microRNA-based studies are **prone to membership inference**
  - However, the dimension *m* (number of released miRNA stats) has to scale with the **square** of the number *n* of individuals in the pool
  - *m* is much smaller than in genomic databases…
  - … but *n* as well (at least in current microRNA databases)
  - … and miRNA expressions are more affected by health status than genotype

- **Recommendation**: increase ***n*** to **>200** individuals and add a bit of noise when necessary

Membership Inference in Other Settings

- **DNA methylation data**
  - Hagestedt I. et al., **Membership Inference Against DNA Methylation Databases**, IEEE EuroS&P 2020

- **Location data**
  - Pyrgelis A. et al., **Knock Knock, Who's There? Membership Inference on Aggregate Location Data**, NDSS 2018

- **Training dataset in machine learning as a service (MLaaS)**
  - Shokri R. et al., **Membership Inference Attacks Against Machine Learning Models,** IEEE S&P 2017
  - Salem A. et al., **ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models**, NDSS 2019

Membership Inference Against Databases

$\mathcal{T} \in \mathbb{R}^{n \times m}$

Example: study of associations between miRNAs and a disease

Summary statistics about a subset of the *m* miRNAs

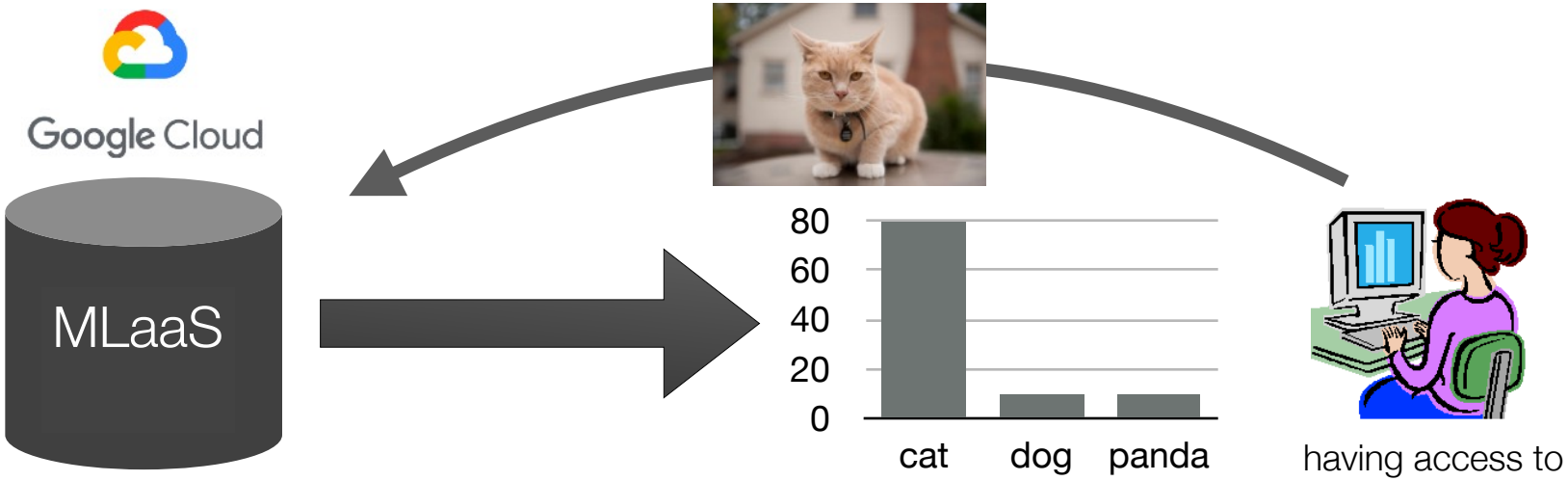
Curious analyst

having access to the actual miRNA expressions of a victim *v*

$\mathbf{x}^v \in \mathbb{R}^m$

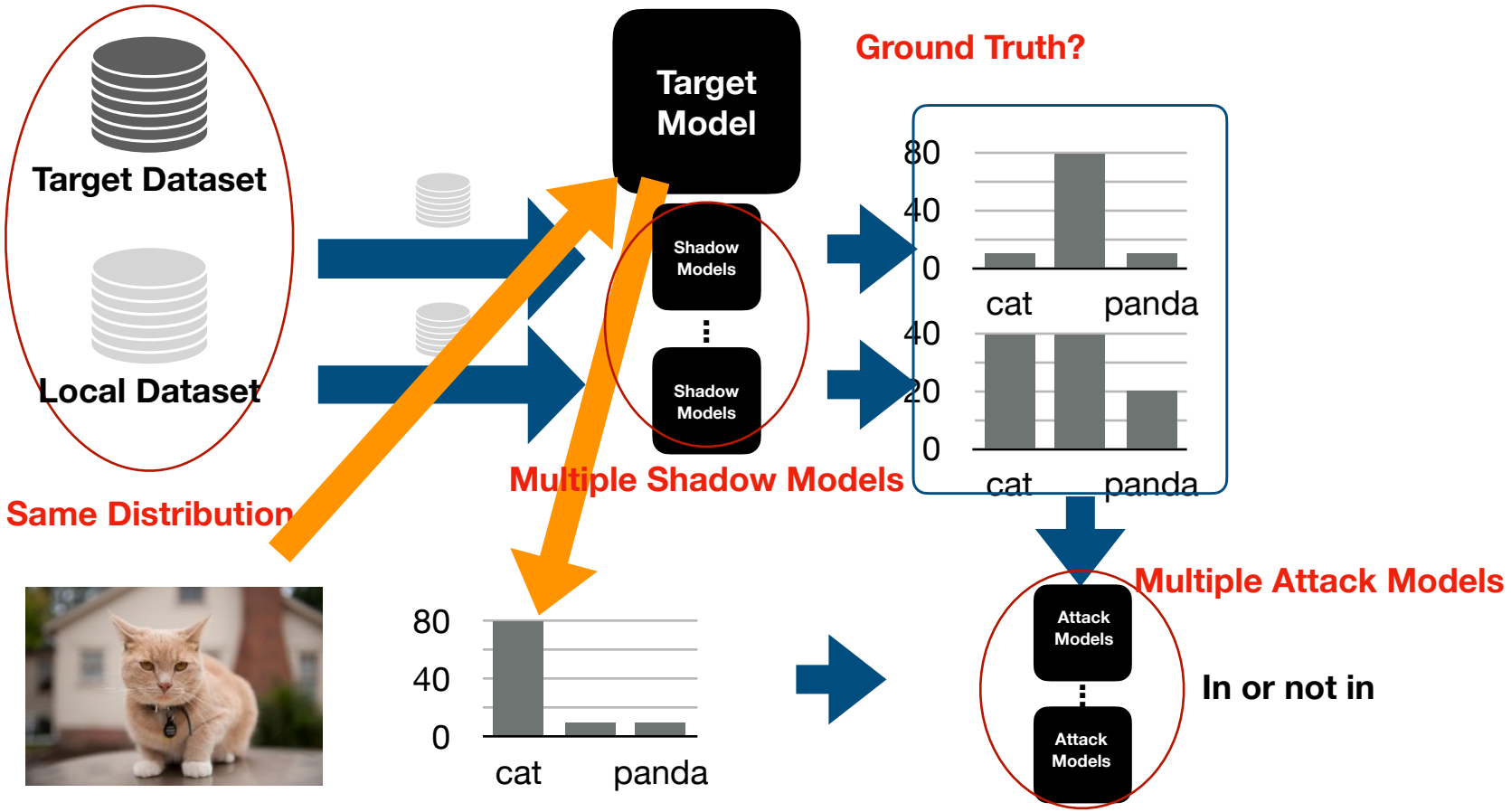
Objective: determine if *v* is member of the pool $\mathcal{T}$ by using the the summary statistics and $\mathbf{x}^v$

Membership Inference Against ML Models



Objective: determine if *v* is part of the training dataset by using the output distribution of the MLaaS system

State-of-the-Art Attack (Shokri et al.)

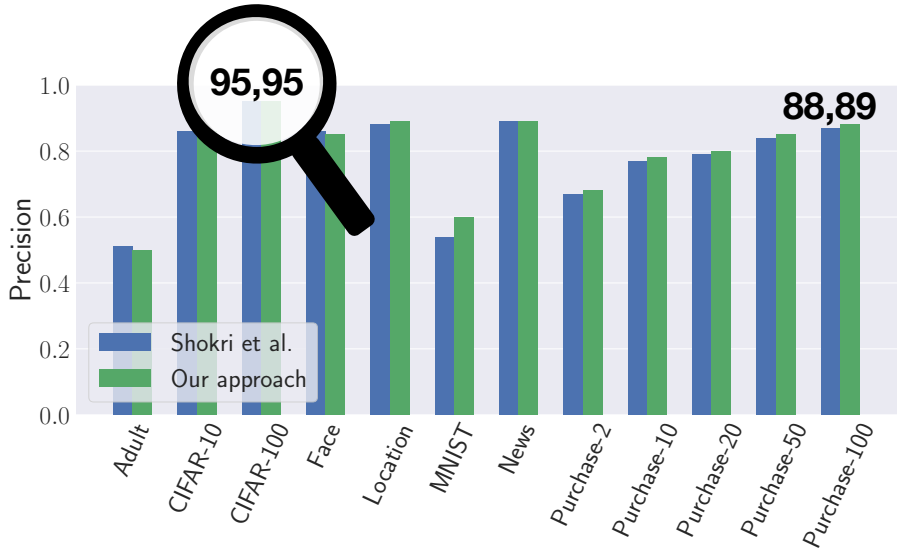


Shokri et al., **Membership Inference Attacks against Machine Learning Models**. IEEE S&P, 2017

More Realistic Attacks

- Three new adversary models:
  1. **A single** shadow model, **different structure** than the target model
     - Less costly attack

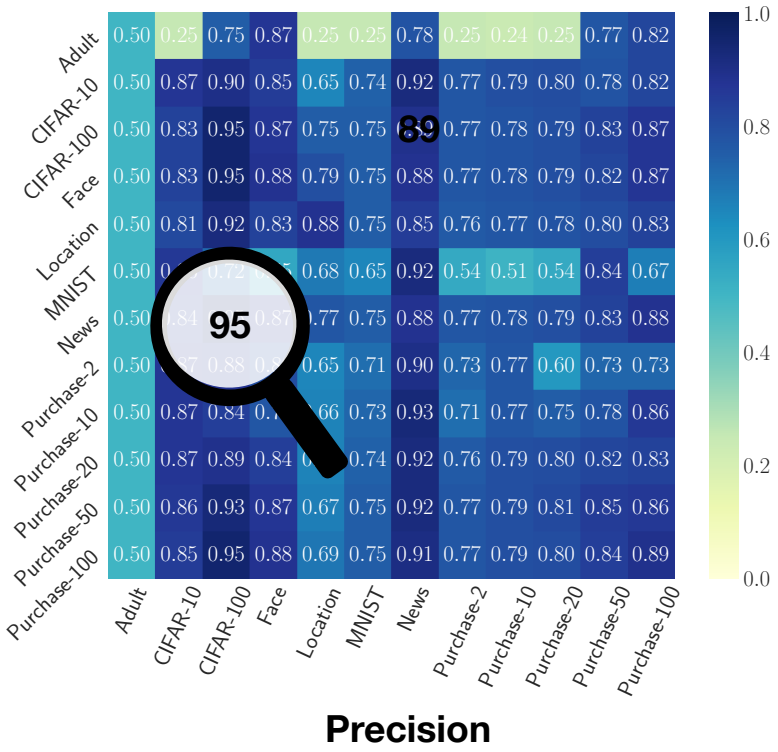
Performance of the First New Adversary



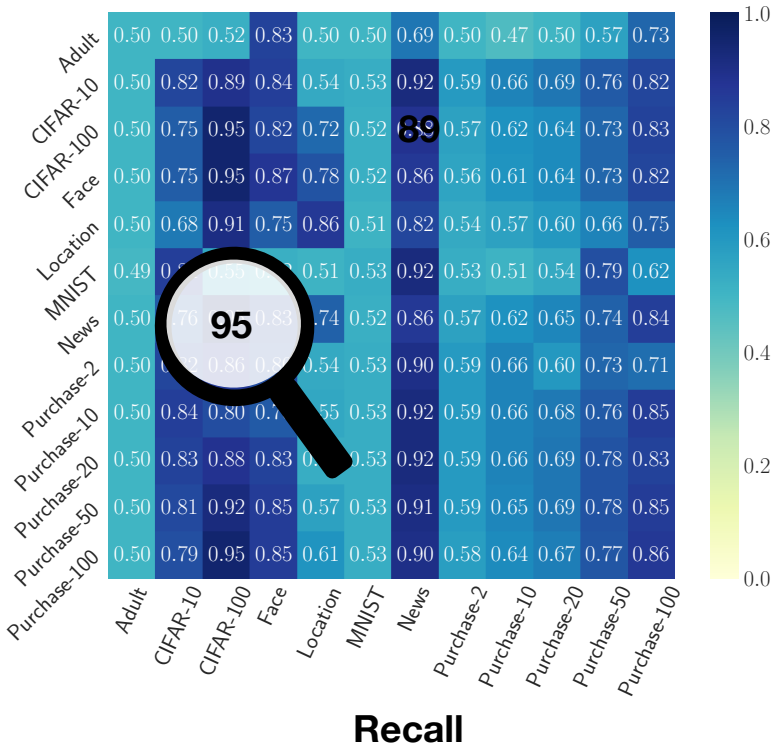
More Realistic Attacks

- Three new adversary models:
  1. **A single** shadow model, **different structure** than the target model
     - Less costly attack
     - Experimental results similar to Shokri et al.
  2. Same as 1. + **different distribution** than the original training set
     - **Transfer learning attack** -> different dataset than the training set

Performance of the Second New Adversary



Precision



Recall

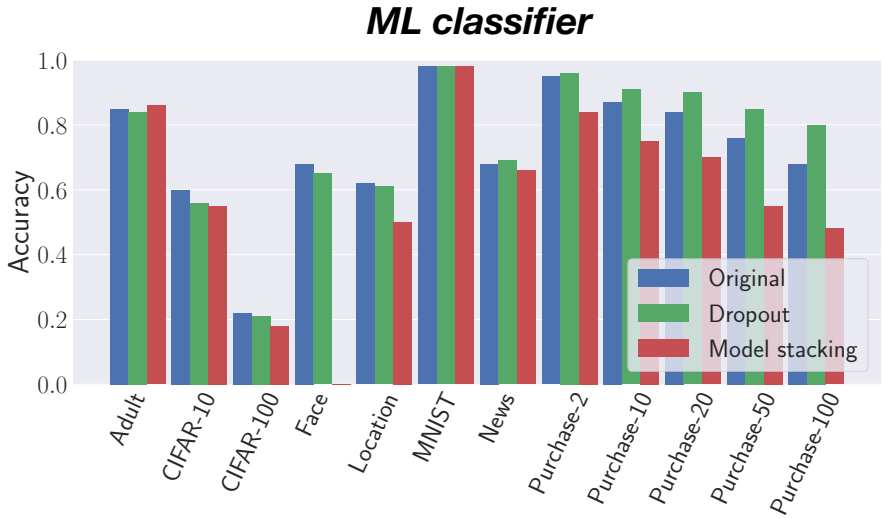
More Realistic Attacks

- Three new adversary models:
  1. **A single** shadow model, **different structure** than the target model
     - Less costly attack
     - Experimental results similar to Shokri et al.
  2. Same as 1. + **different distribution** than the original training set
     - **Transfer learning attack** -> different dataset than the training set
     - Results decreasing by a few % only
  3. **No shadow model**, different distribution than the training set
     - No training phase
     - Attack based on the **output distribution only**
     - Statistics such as ***max*** or the ***entropy*** can be sufficient
     - Good results for about half of the tested datasets

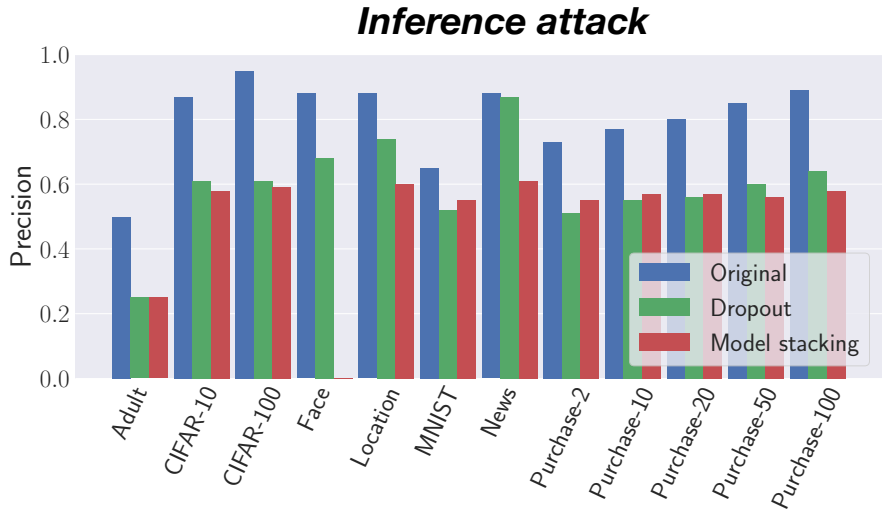
Defense Mechanisms

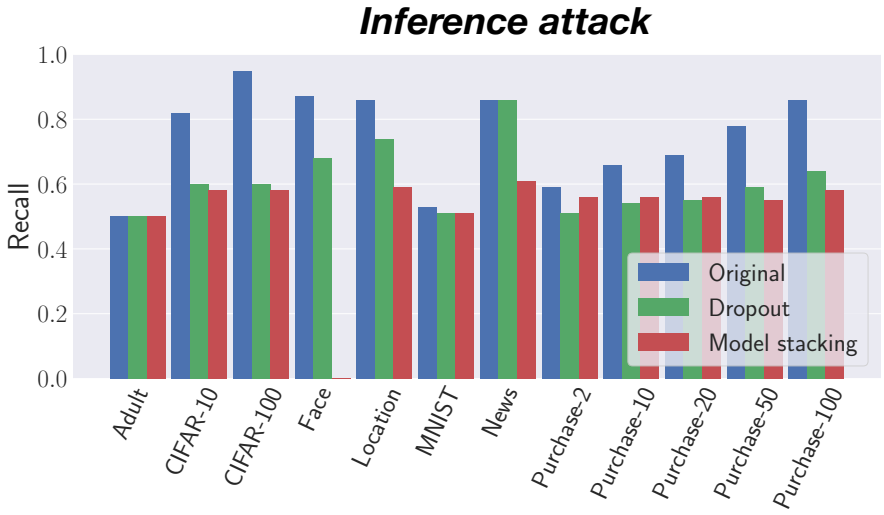
- Main reason for the success of the attack
  - **Overfitting** of training samples

- Two defense mechanisms
  - **Dropout**
  - **Model stacking**



Utility





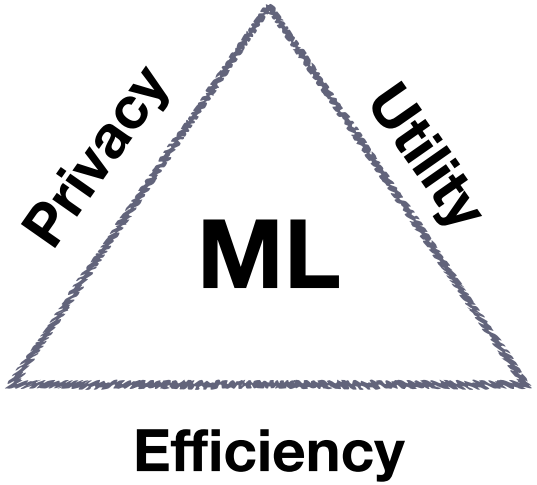
Privacy

Conclusion

- **Leveraging machine learning or statistical analysis for attacking privacy**
  - Of various biomedical data but also of other data types (location, hashtags, …)
  - For three different types of attacks (attribute inference, membership inference, linkability)

- **Improving privacy in machine-learning settings**
  - Of SVM classifiers
  - Of random forest classifiers
  - In machine learning as a service

- **There is no one size-fits-all solution**
  - Depend on the ML setting (algorithm and phase)
  - Depend on the attack
  - Depend on the data



contact: mathias.humbert@unil.ch

References

- Humbert et al., **Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy**, ACM CCS 2013
- Humbert et al., **Quantifying Interdependent Risks in Genomic Privacy**, ACM TOPS 2017
- Olteanu et al., **Quantifying Interdependent Privacy Risks with Location Data**, IEEE TMC 2017
- Backes et al., **Privacy in Epigenetics: Temporal Linkability of MicroRNA Expression Profiles**, USENIX Security 2016
- Backes et al., **Identifying Personal DNA Methylation Profiles by Genotype Inference**, IEEE S&P 2017
- Backes et al., **Membership Privacy in MicroRNA-based Studies**, ACM CCS 2016
- Salem et al., **ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models**, NDSS 2019