Privacy in Epigenetics:

Temporal Linkability of MicroRNA Expression Profiles

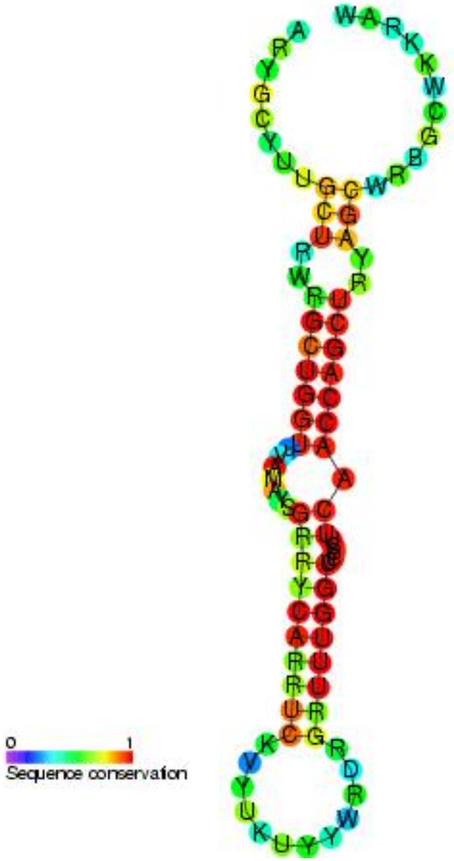
Michael Backes, **Pascal Berrang**, Anne Hecksteden, Mathias Humbert, Andreas Keller, Tim Meyer

USENIX Security Symposium
Austin, August 12th, 2016

CISPA
Center for IT-Security, Privacy and Accountability

MicroRNAs

- miRNAs are a part of epigenetics
- miRNAs are small **RNA molecules**, which **regulate** most of the **human genes**
  - i.e., they regulate what a cell does / what type of cell it is
- miRNA expression quantifies **how much** miRNAs are **active** in a given cell
  - a real-valued number for each miRNA
  - e.g., *hsa-let-7a-1: 102*



MicroRNAs



- highly influenced by **external factors**
  - e.g., ageing, diet, childhood development, environmental chemicals
  - thus, **highly variable over time**



- miRNA expression **reflects the health status**
  - due to their influence on gene expression

- associations between miRNA expression and
  - neurodegenerative diseases (e.g., Alzheimer's), heart disease, diabetes, majority of cancer types
  - and more to follow

DNA and MicroRNAs

CISPA Center for IT-Security, Privacy and Accountability

DNA

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



miRNA

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



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

see: *Routes for breaching and protecting genetic privacy*, Erich and Narayanan (2014)

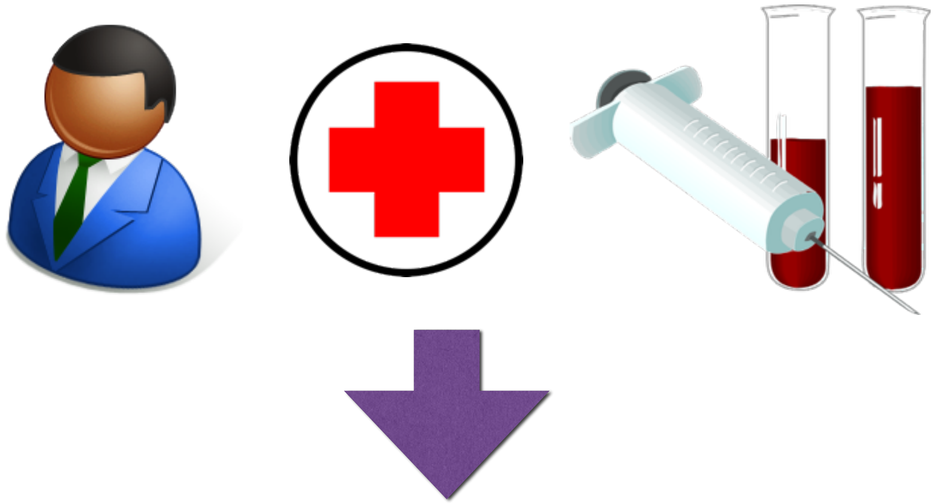
Research Questions

- MicroRNA expression-based data
  - studies, data in hospitals, etc.

- Are miRNA expression profiles vulnerable to **linkability attacks**?
  - identification attack (1 sample → database)
  - matching attack (database → database)
- If yes, how can we **protect** this kind of data?
- How does the protection affect the **utility** of the data?



How to get miRNA expression profiles?



public DB
(such as the Gene Expression Omnibus)

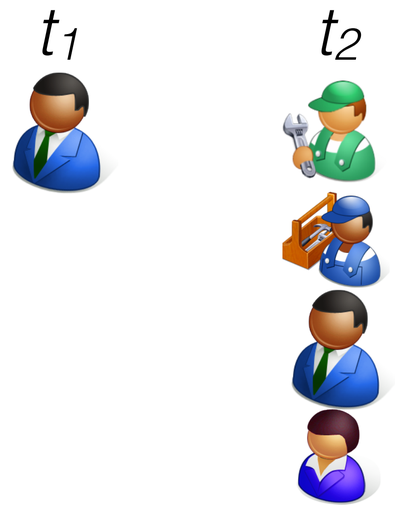
black market

cyber **attacks** against healthcare companies have **increased** by 72% within one year

since 2015:
>400 breaches, e.g., Anthem, **>78mio** records

Linkability Attacks

- Identification
  - e.g., 1 sample from breach (incl. name) and public db (excl. name)
  - **Which sample in the database corresponds to the 1 sample?**



- Matching
  - e.g., a leaked database (incl. name) and public db (excl. name)
  - **Which sample from t_1 corresponds to which sample from t_2?**



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**



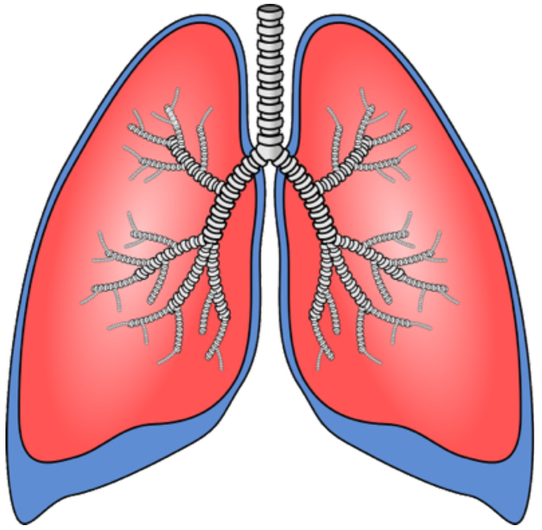
Lung Cancer Dataset

Participants: **26** (huge for a longitudinal study!)

Points in time: **8**

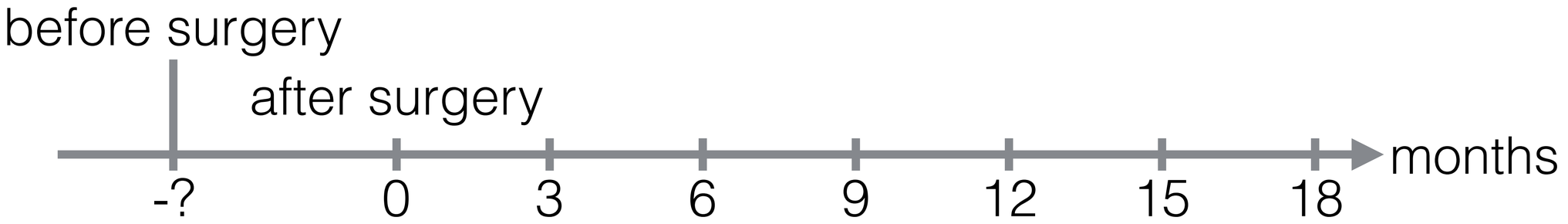
Time period: **3 months**

Disease: **lung cancer**



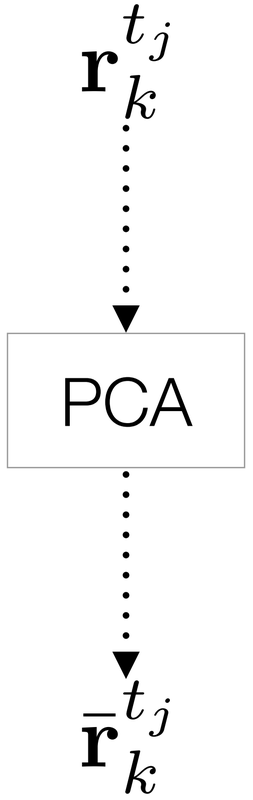
1,189 miRNAs per sample

- taken from **plasma**



Pre-processing

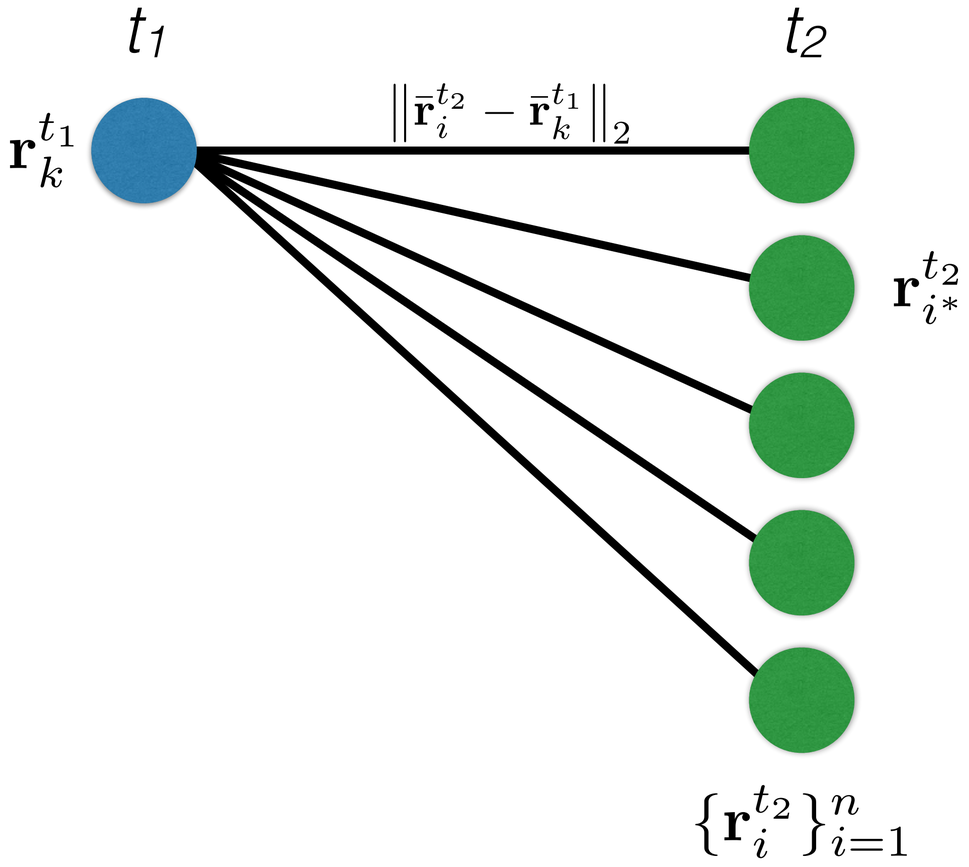
- high dimensionality: **1,189 miRNAs** per sample
  - possibly correlated and uninteresting components
- **PCA + whitening** provides
  - unit variance
  - smaller dimensionality *m*
  - uncorrelated components
- condenses data into a set of smaller dimensions with **minimal information loss**



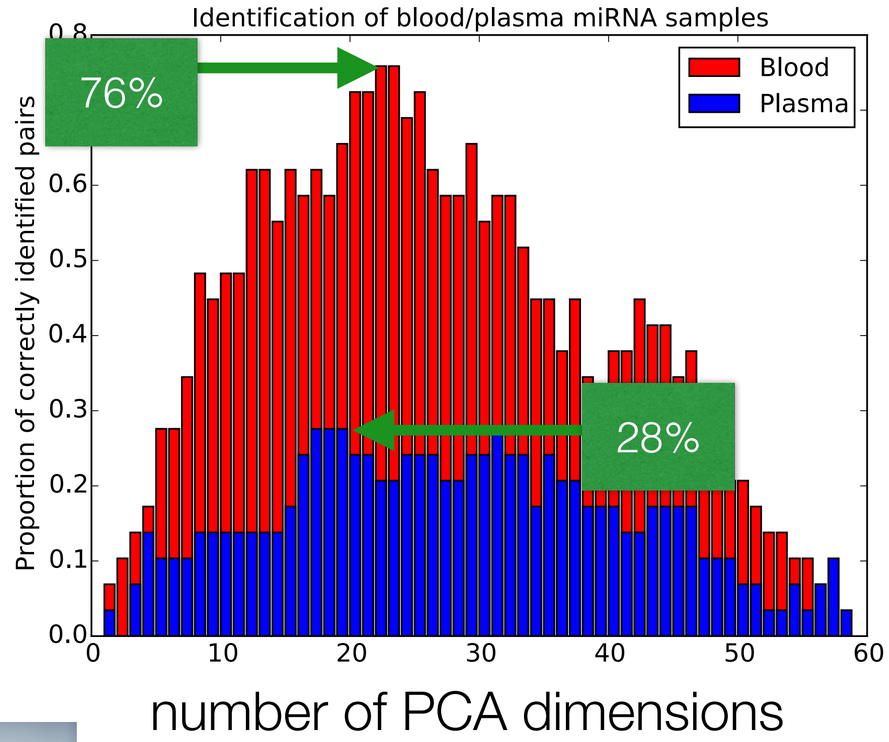
Identification Attack

Which sample in the database corresponds to the 1 sample?

$$i^* = \arg\min_i \left\| \bar{\mathbf{r}}_i^{t_2} - \bar{\mathbf{r}}_k^{t_1} \right\|_2$$



Identification Attack – Results



success rate:

percentage of successfully identified individuals

high success especially
on blood-based samples



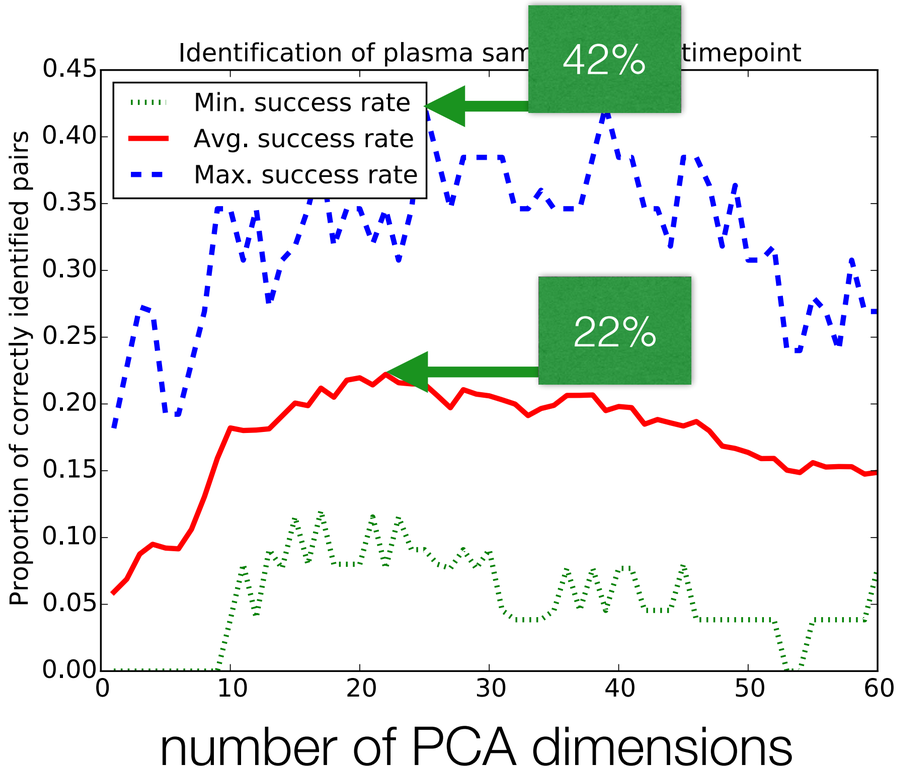
Identification Attack – Results

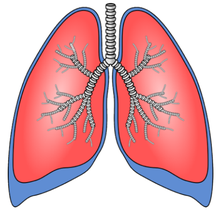
min, avg and max **success:**

over all combinations of 8 time points

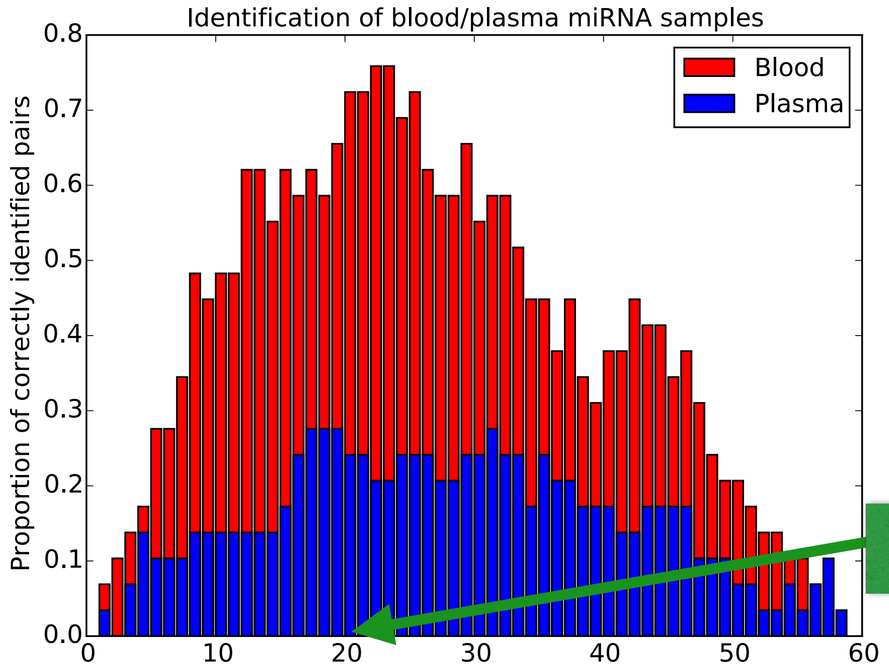
still reasonable success rates

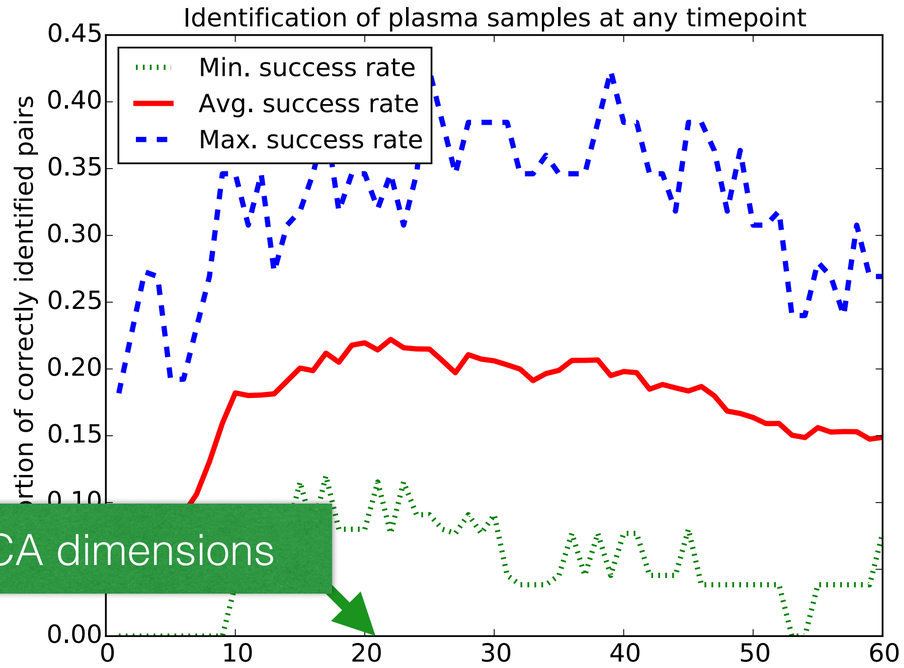
comparable to plasma-based samples in the athletes' dataset





Identification Attack – Results

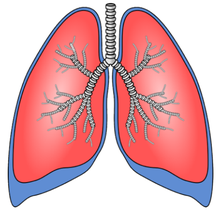




similar number of PCA dimensions



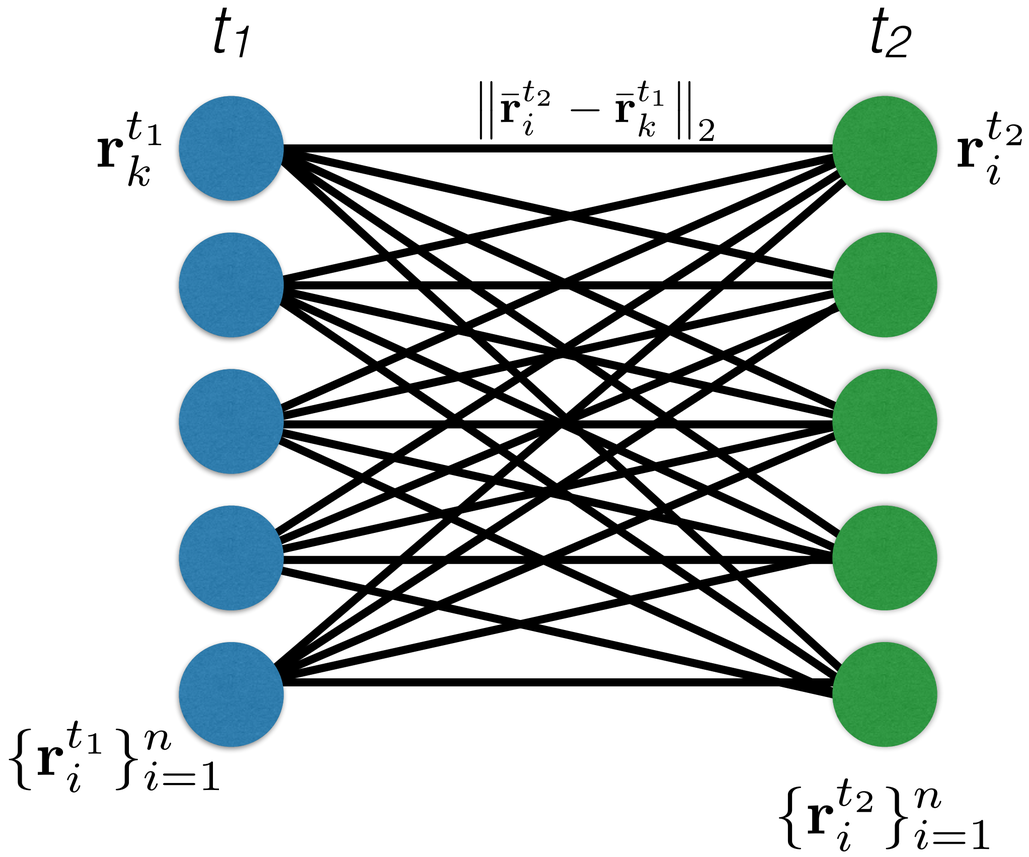
similar #dimensions and miRNAs (in 1st dimension) account for the highest success



Matching Attack

Which sample from t_1 corresponds to which sample from t_2?

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

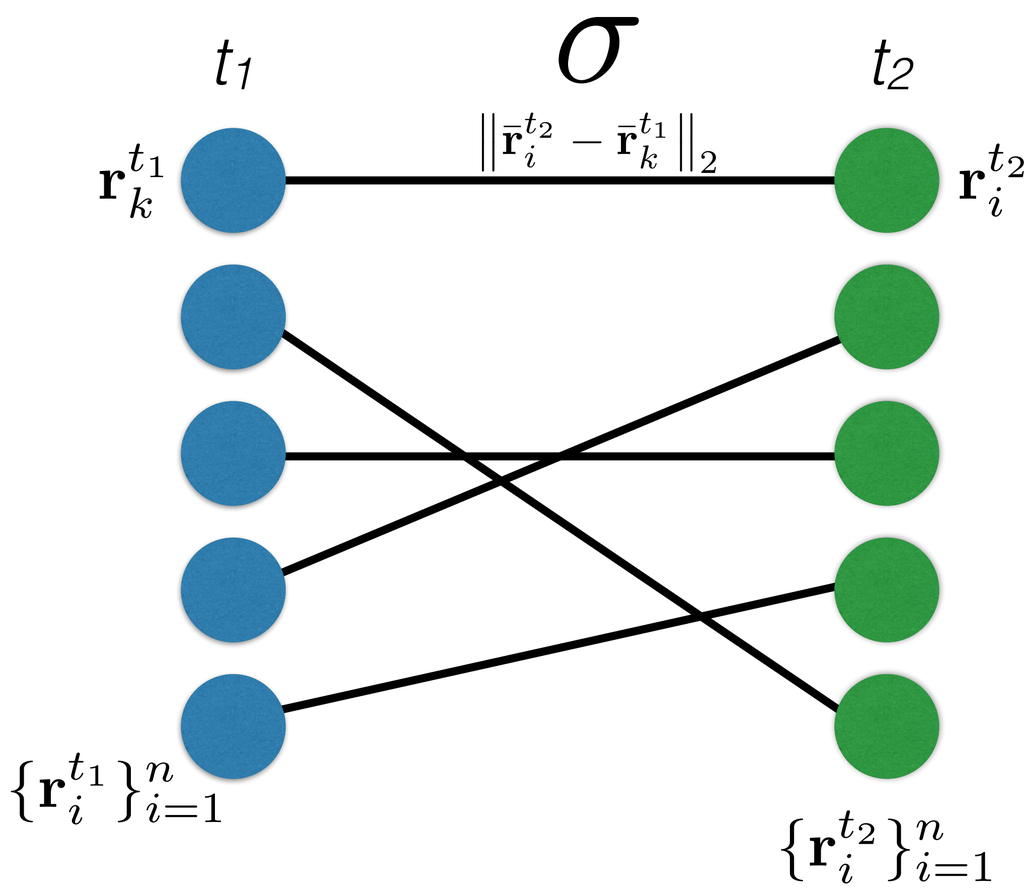


Time complexity: ***O(n³)***

Matching Attack

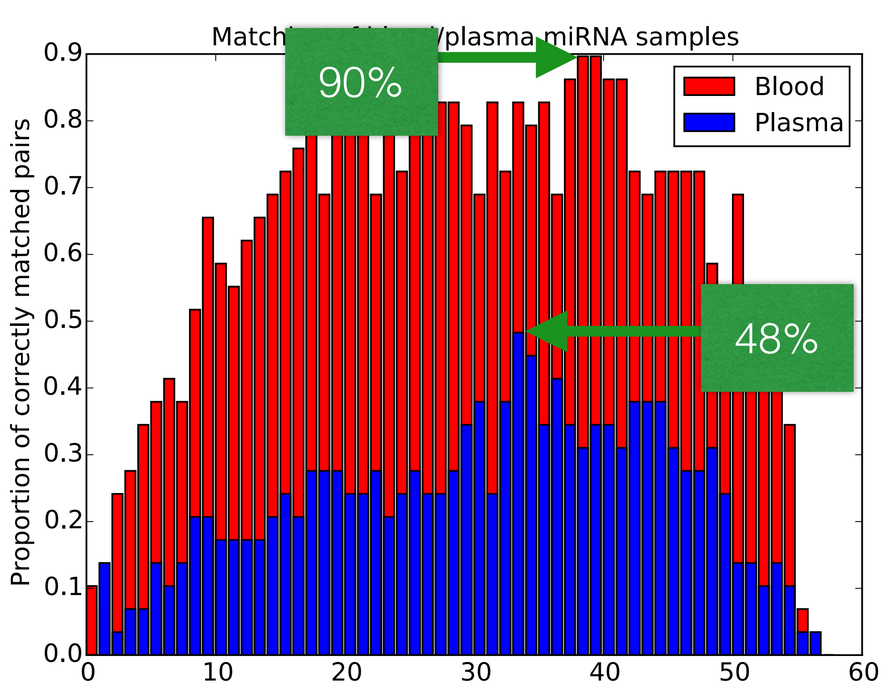
Which sample from t_1 corresponds to which sample from t_2?

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

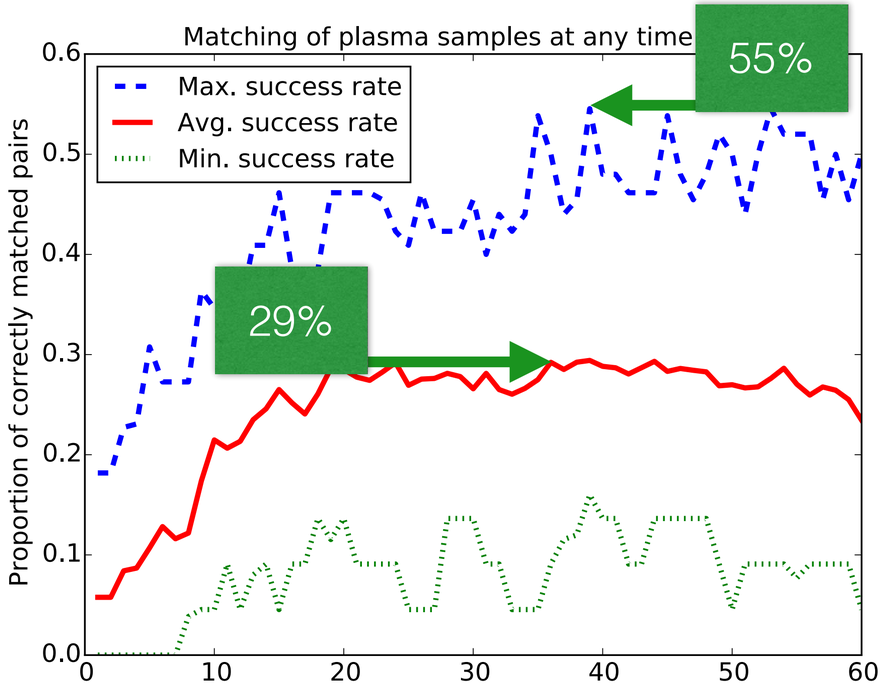


Time complexity: ***O(n³)***

Matching Attack – Results



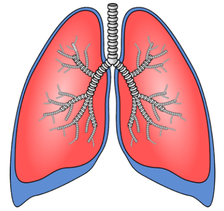
number of PCA dimensions



number of PCA dimensions



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

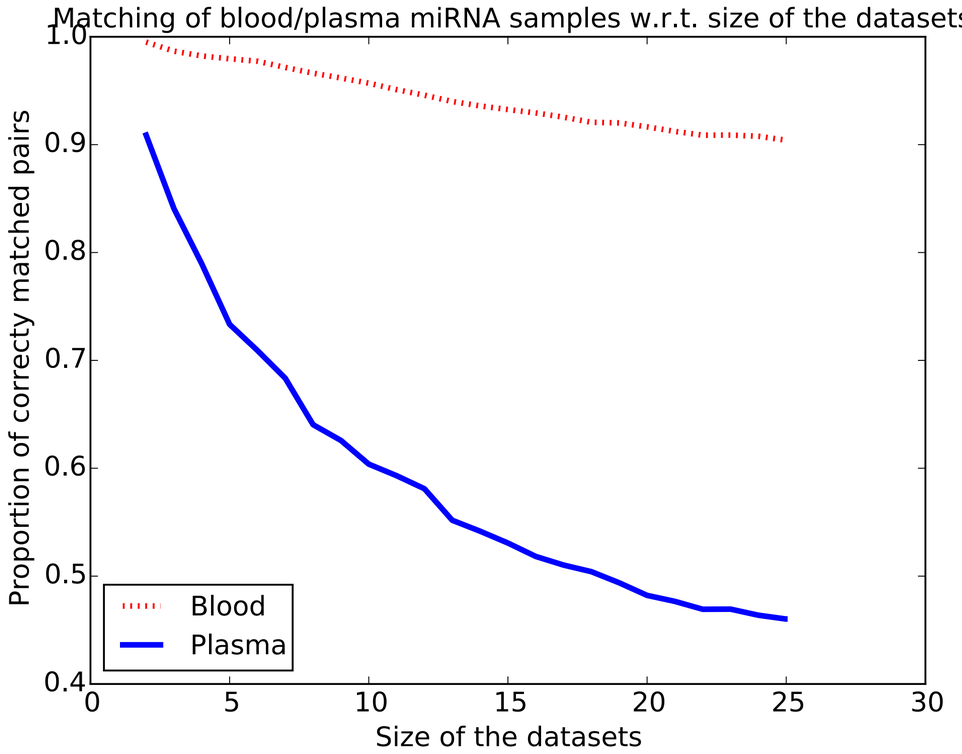


Matching Attack – Results

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

Success *decreases sharply* for plasma-based samples,

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

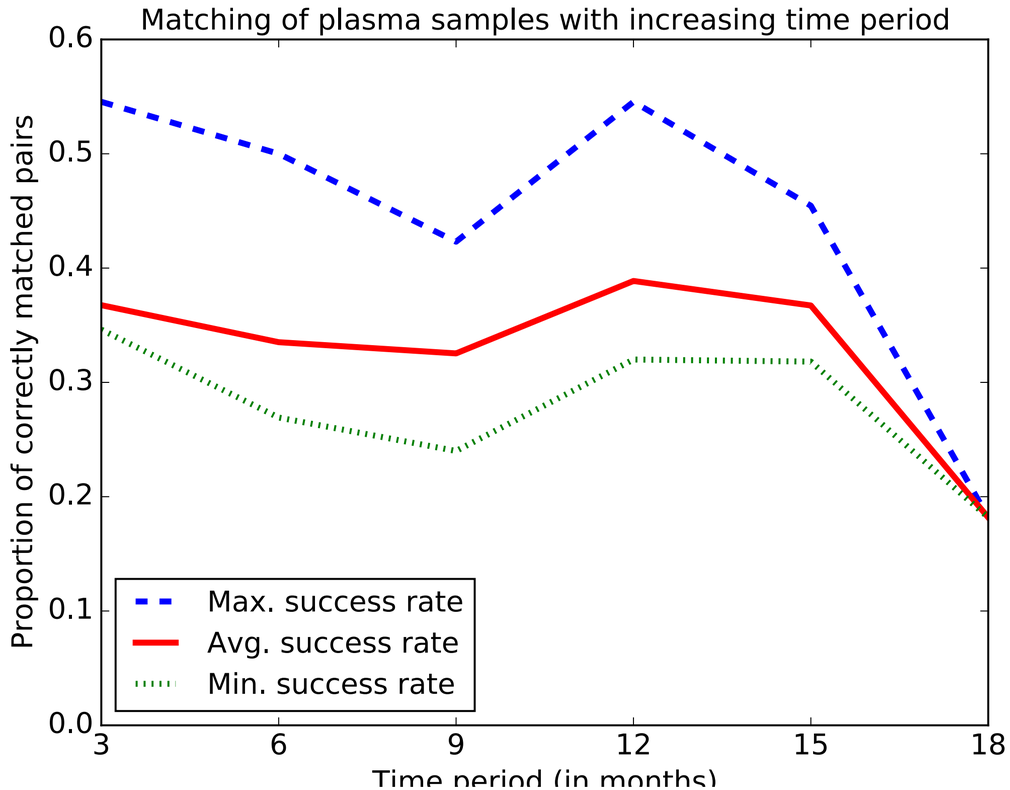


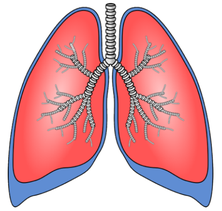


Matching Attack – Results

How does the success
evolve over **increasing time periods**?

Average success **almost constant**
within the first year.





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



belief is unjustified
success as high as **90%** for blood-based samples

Protection of miRNA Data

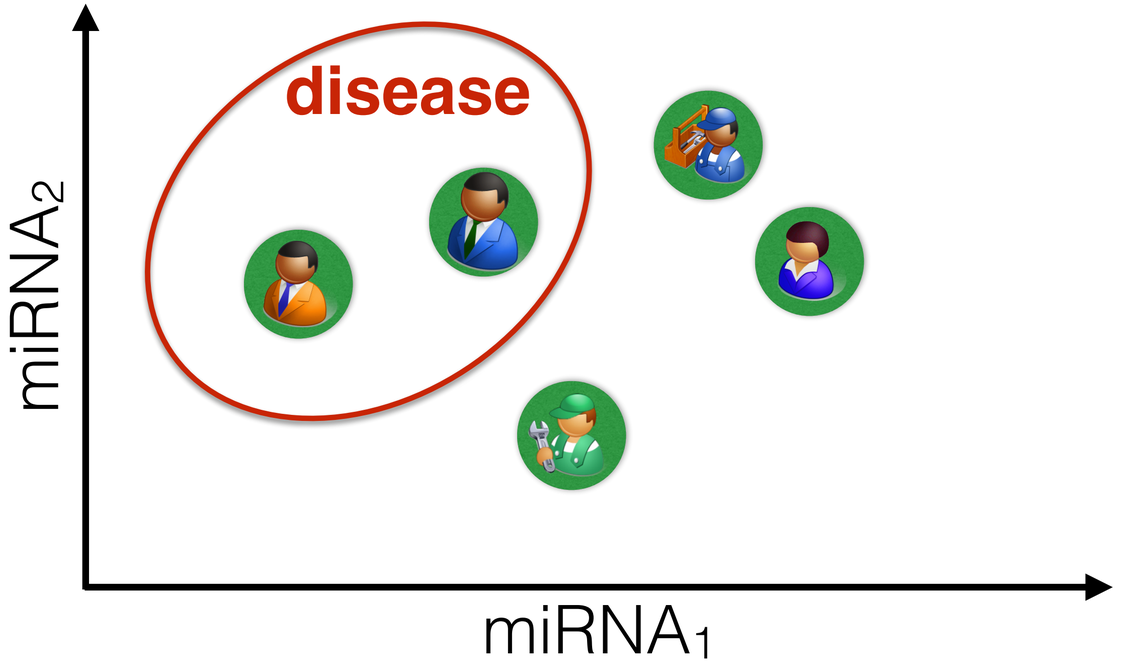
Countermeasures

- **Probabilistically sanitizing** the miRNA expression profiles
  - Adding noise in a fully distributed, differentially private manner
    → providing epigeno-indistinguishability
  - Noise drawn according to multivariate Laplacian mechanism
  - e.g., for *publishing a dataset used in a study*
- **Hiding** non-relevant miRNA expressions (not measuring/publishing)
  - sometimes, **randomization is not an option**
  - e.g., for *making a diagnosis in a hospital*
  - Downside: correlations between miRNAs

Privacy-Utility Trade-Off

Privacy: prevent **linkability** of samples

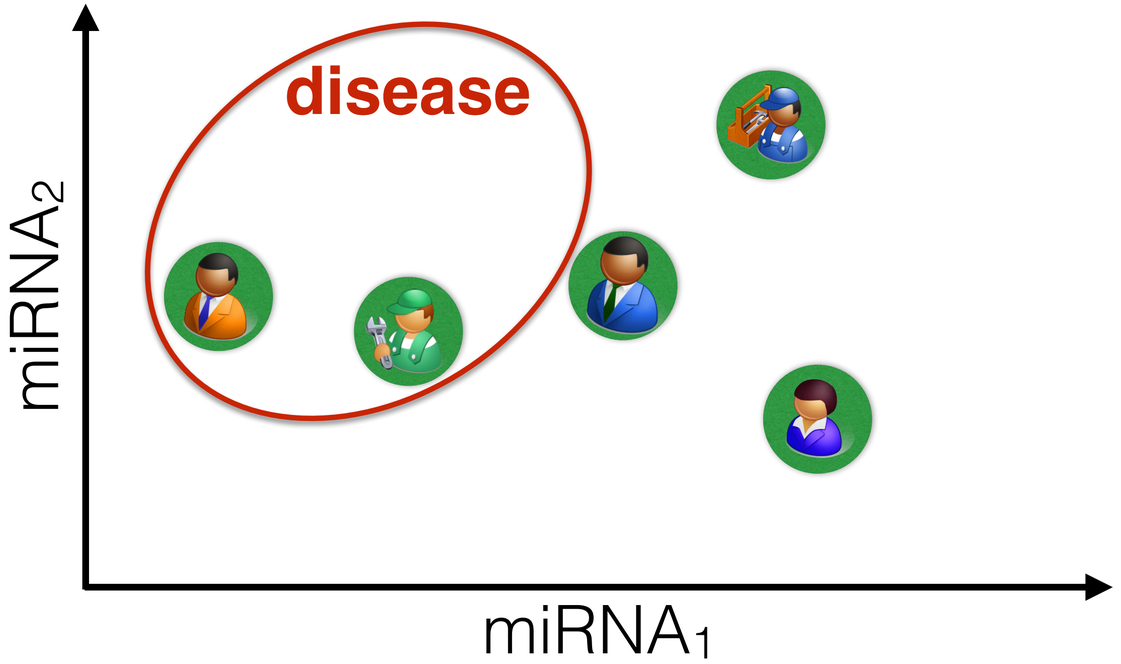
Utility: preserve accuracy of classification as **diseased / healthy**, usually using a radial SVM classifier



Privacy-Utility Trade-Off

Privacy: prevent **linkability** of samples

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



Privacy-Utility Trade-Off

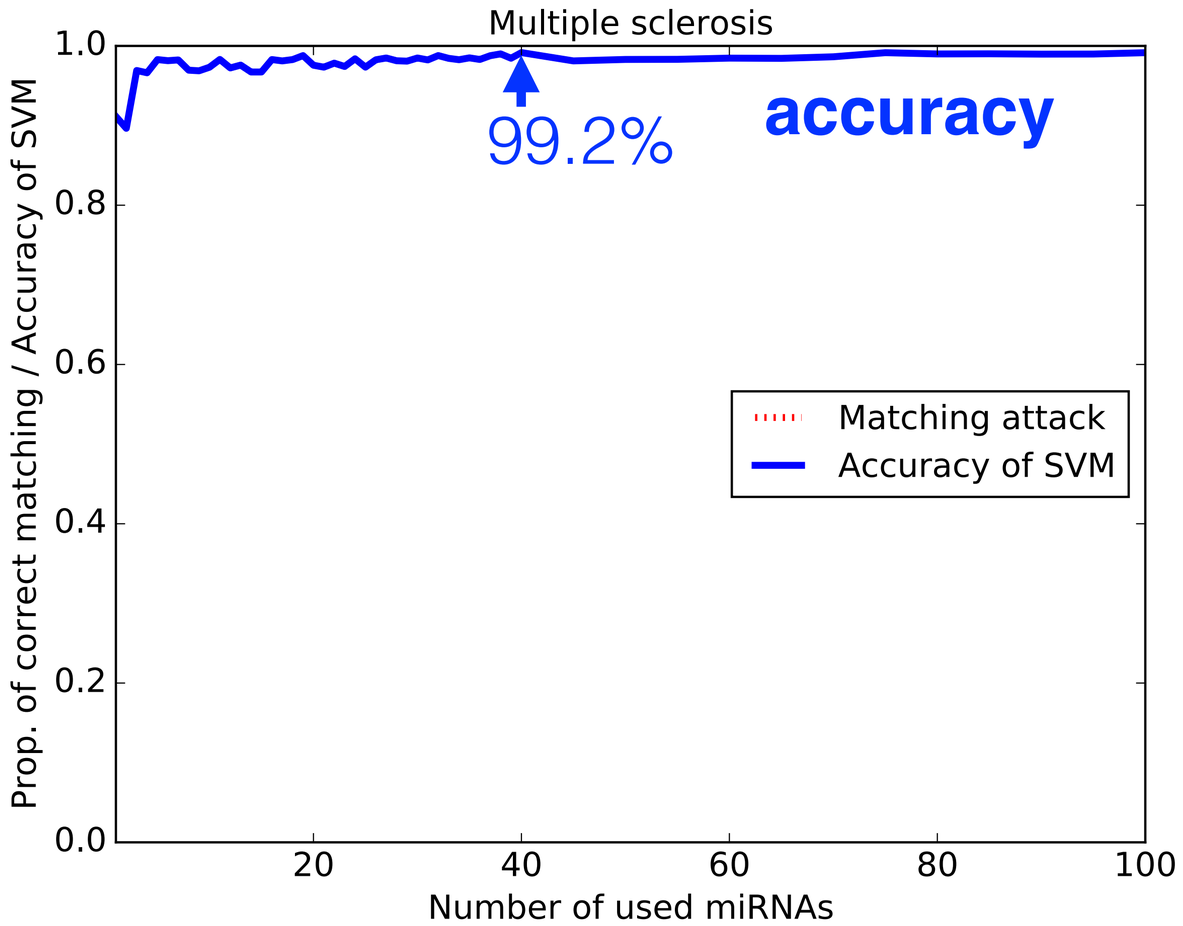
Privacy: prevent **linkability** of samples

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

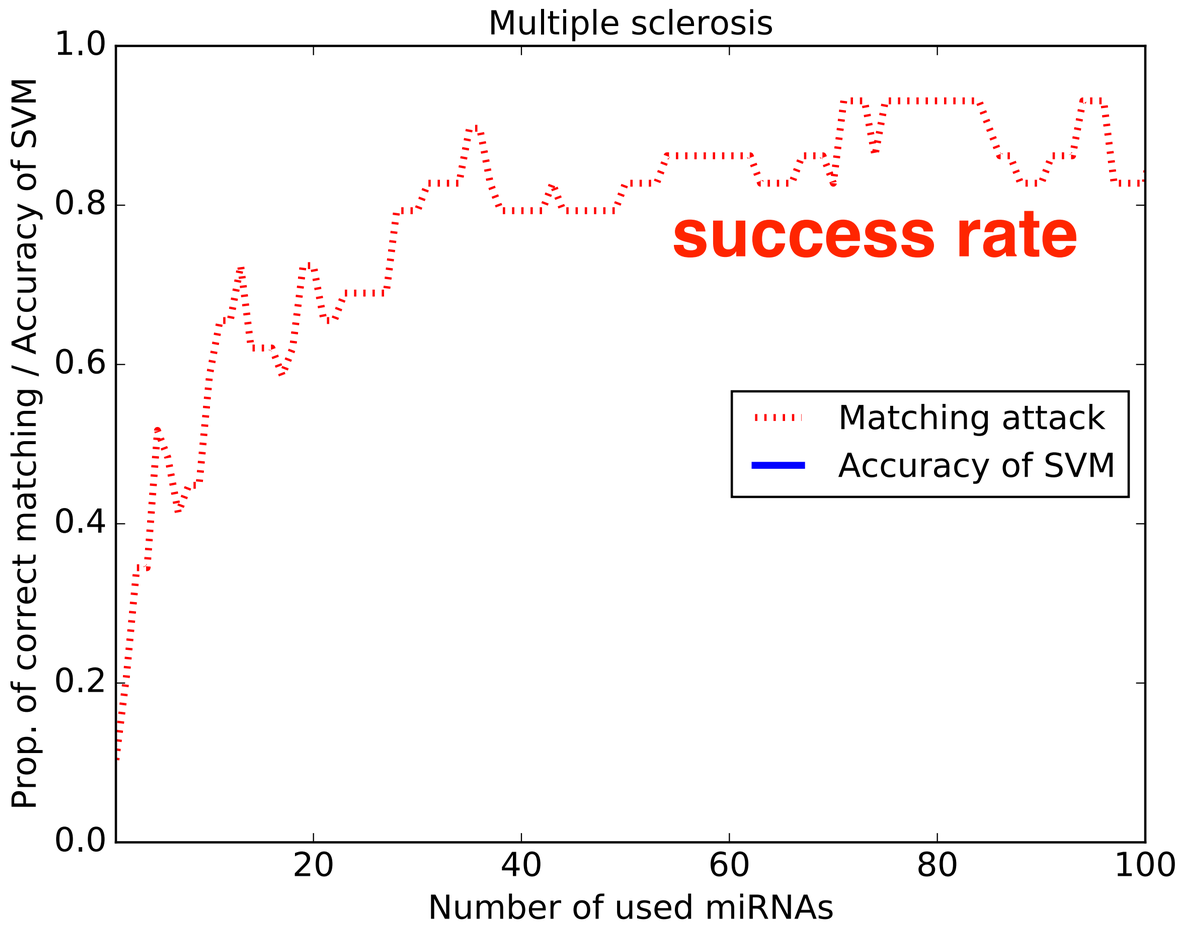
<u>another dataset for exploring utility:</u>

>1000 participants,
19 diseases,
1 time point

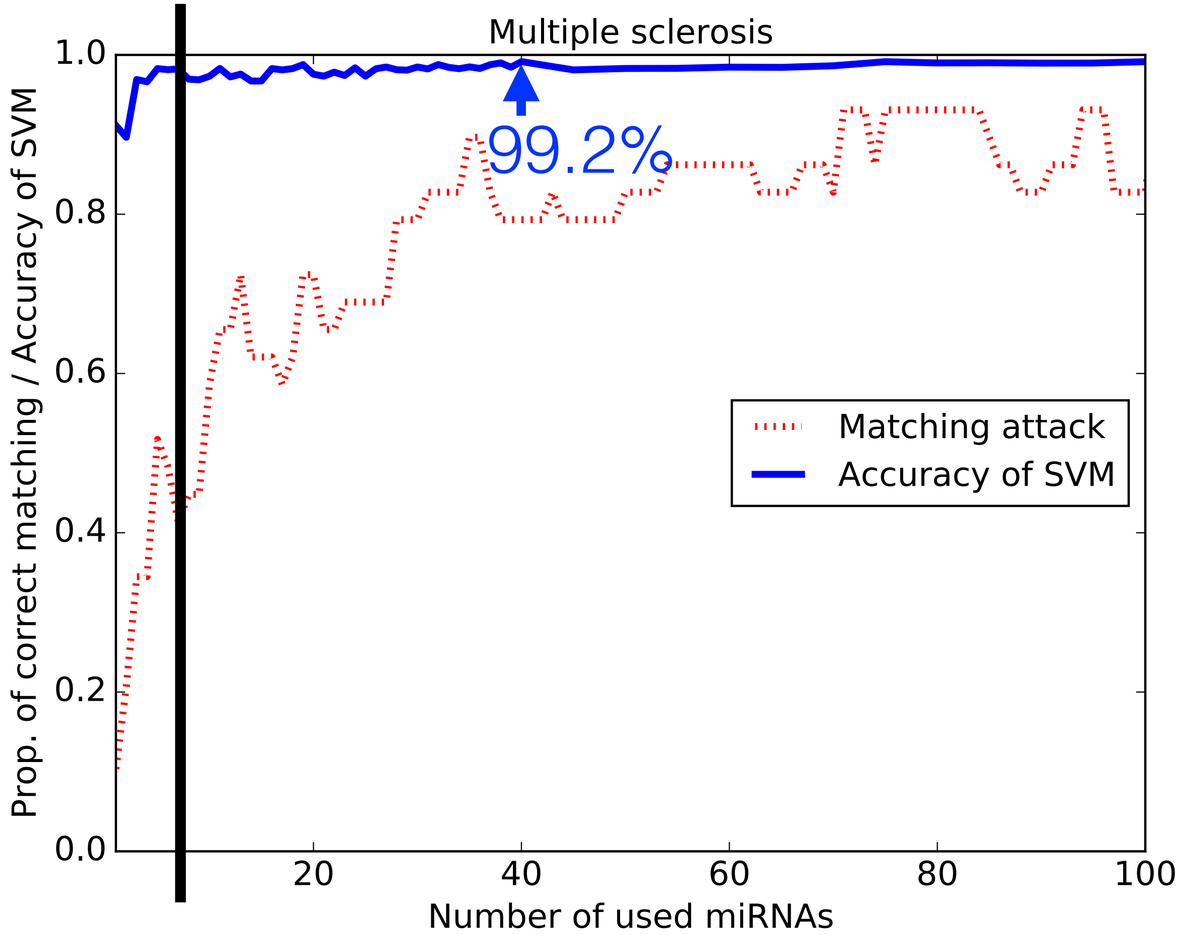
Hiding miRNAs – Results



Hiding miRNAs – Results



Hiding miRNAs – Results

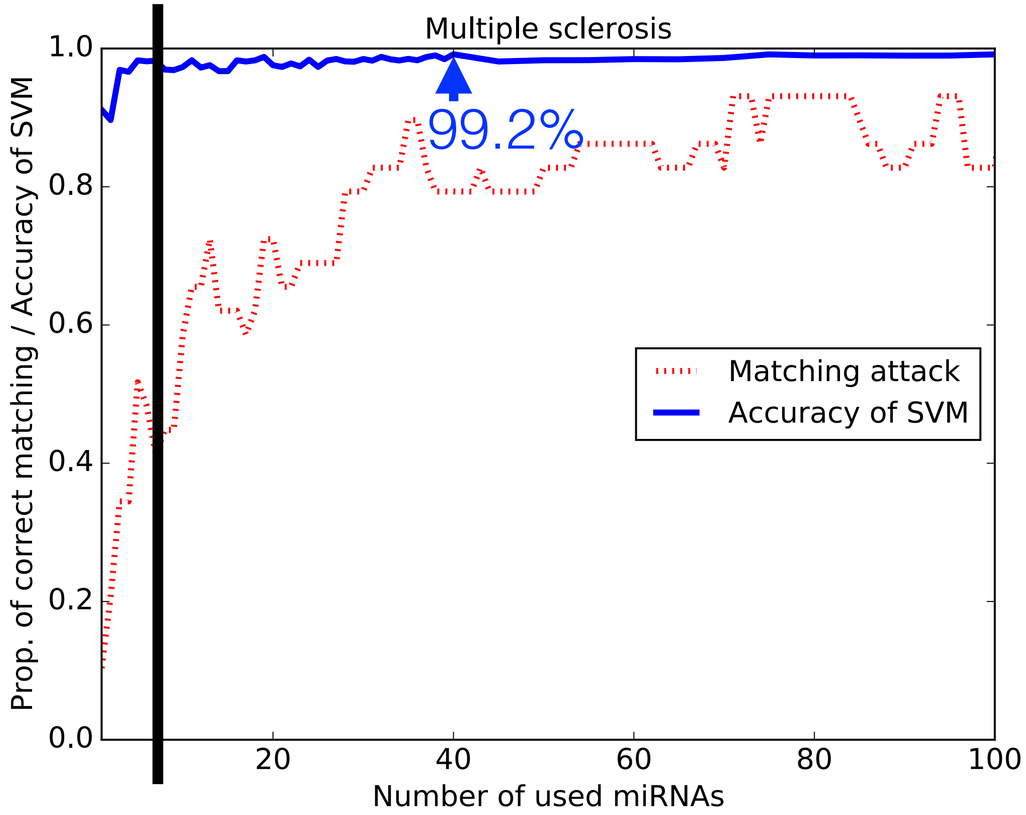


Hiding miRNAs – Results

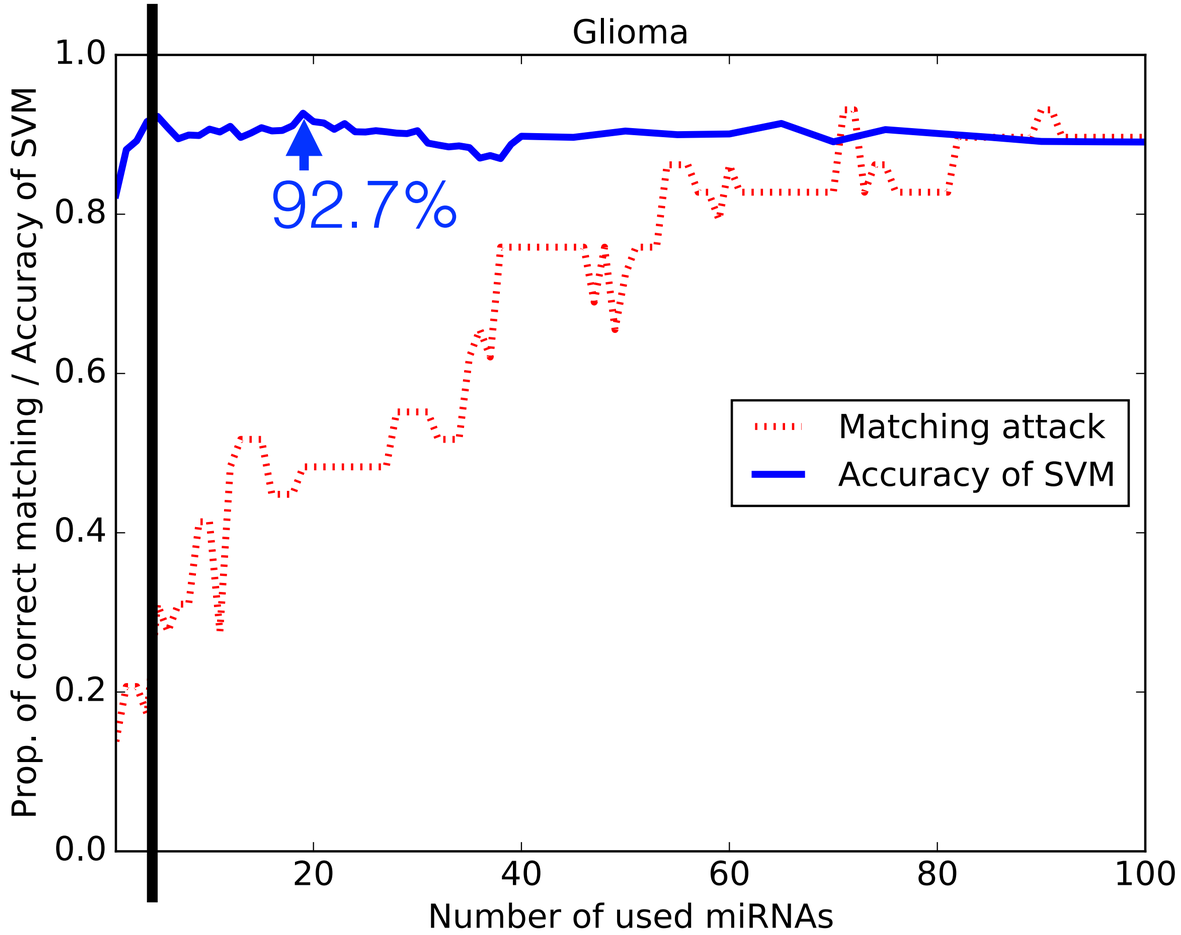
trade-off at 7 miRNAs

success decreases (relative to all)
by **54%**

accuracy decreases (relative to max)
by only **1%**



Hiding miRNAs – Results

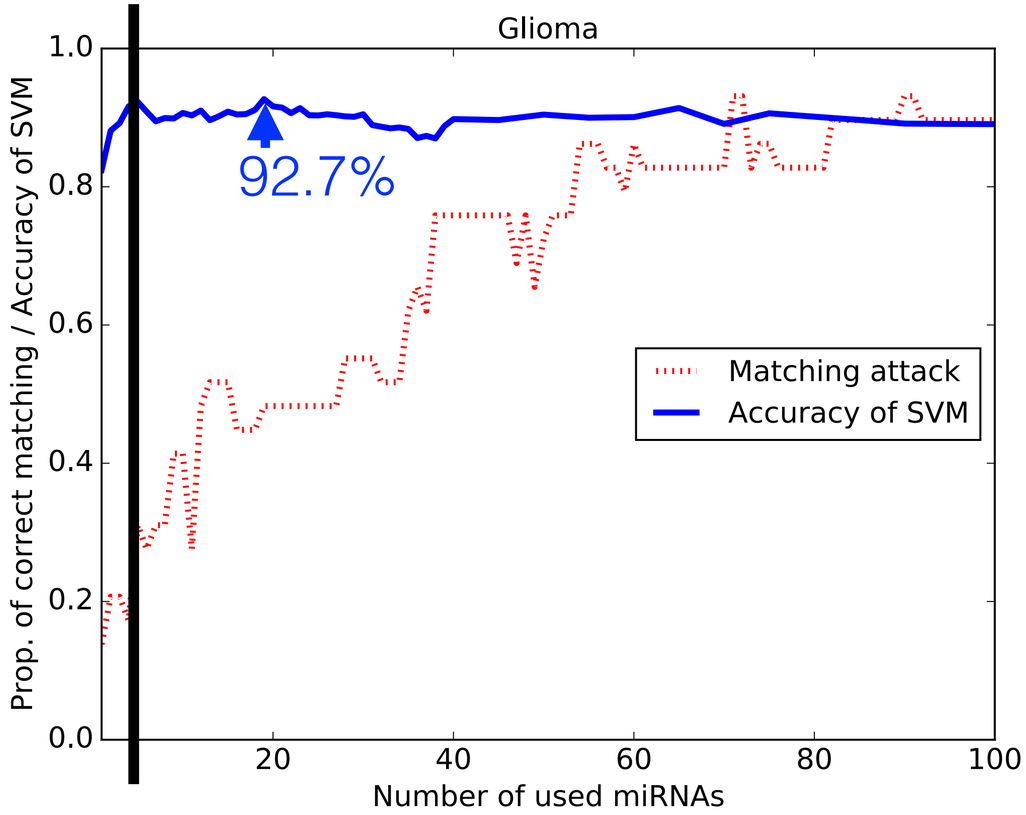


Hiding miRNAs – Results

trade-off at 4 miRNAs

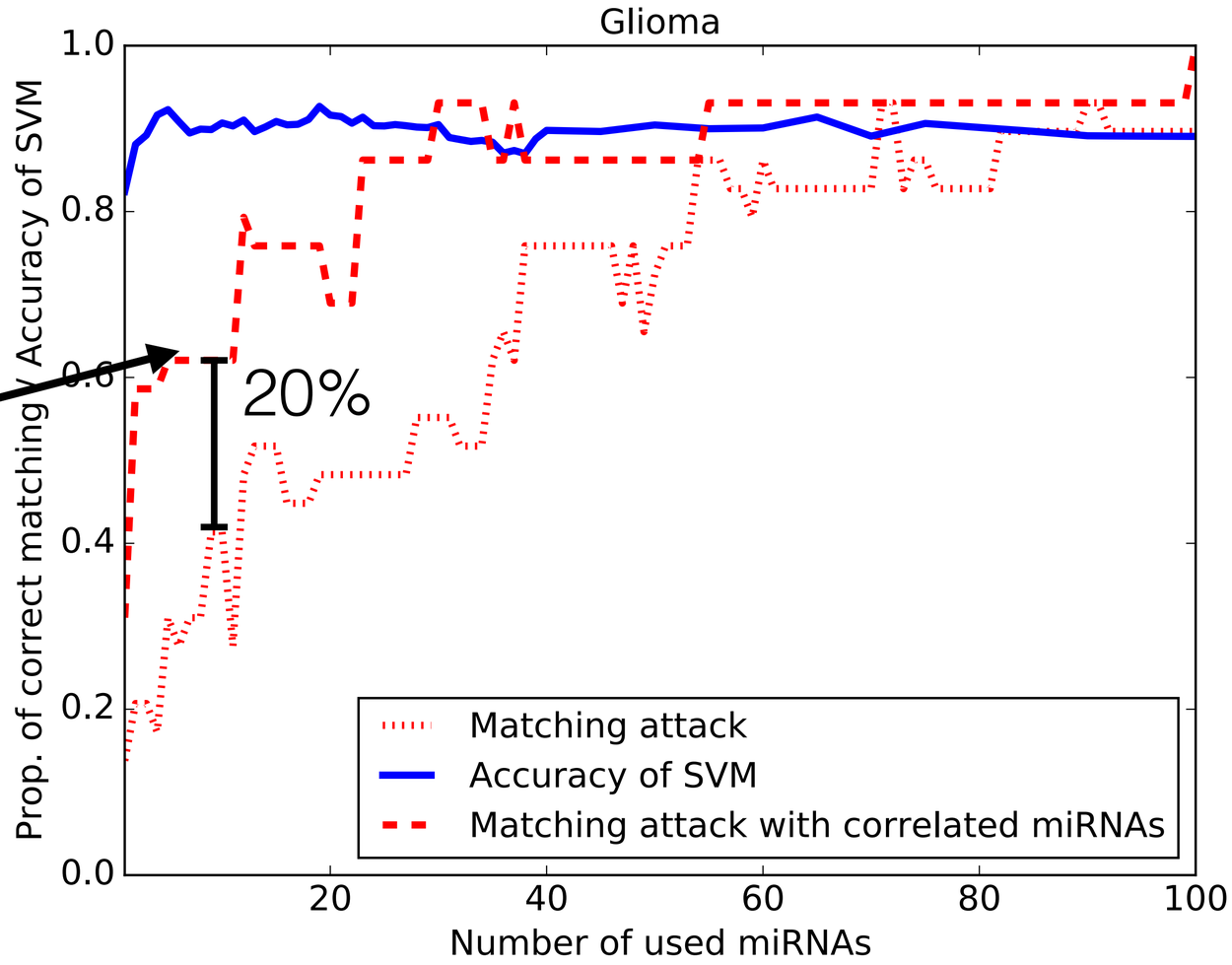
success decreases (relative to all)
by **80%**

accuracy decreases (relative to max)
by only **1%**

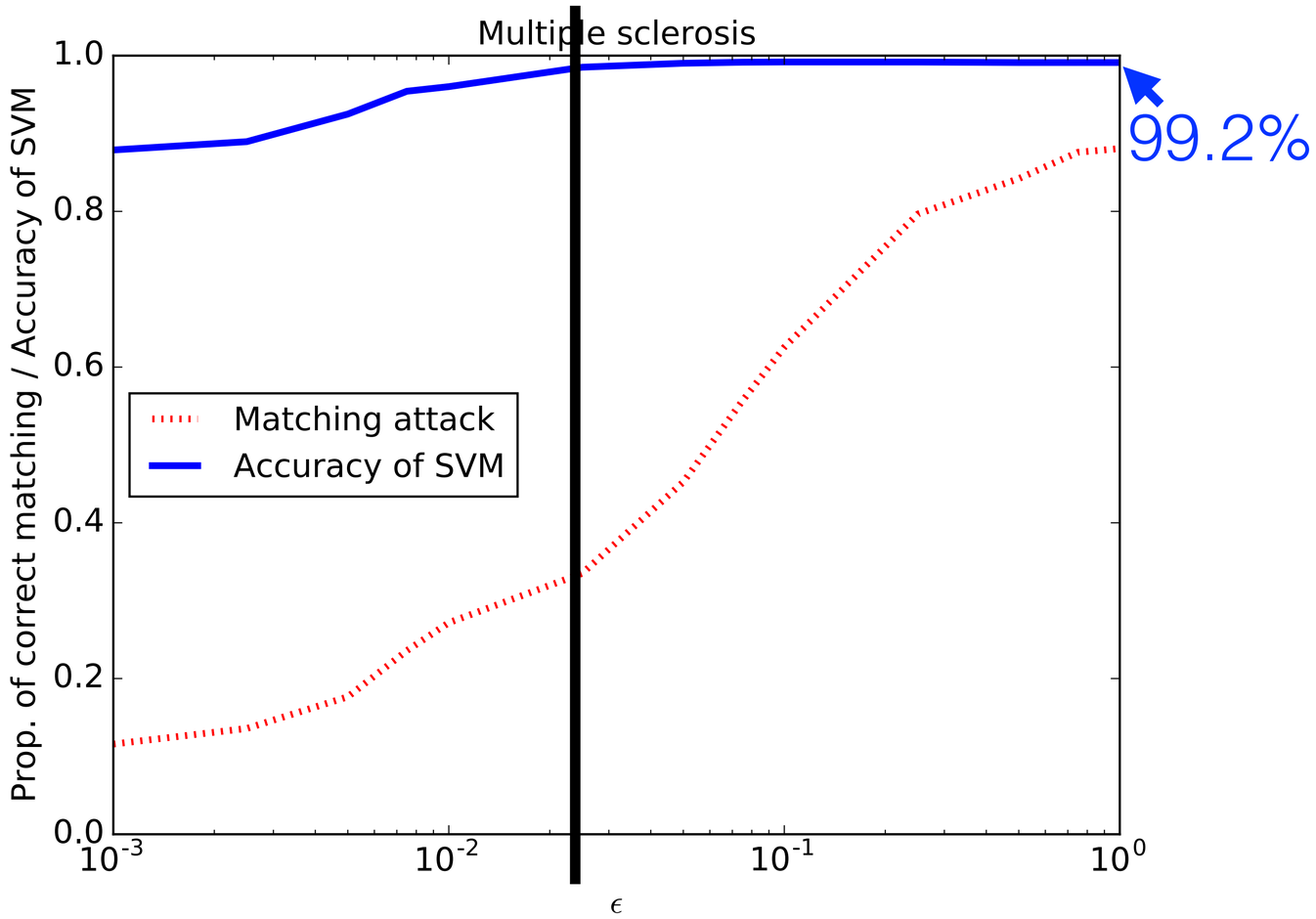


Hiding miRNAs – Results

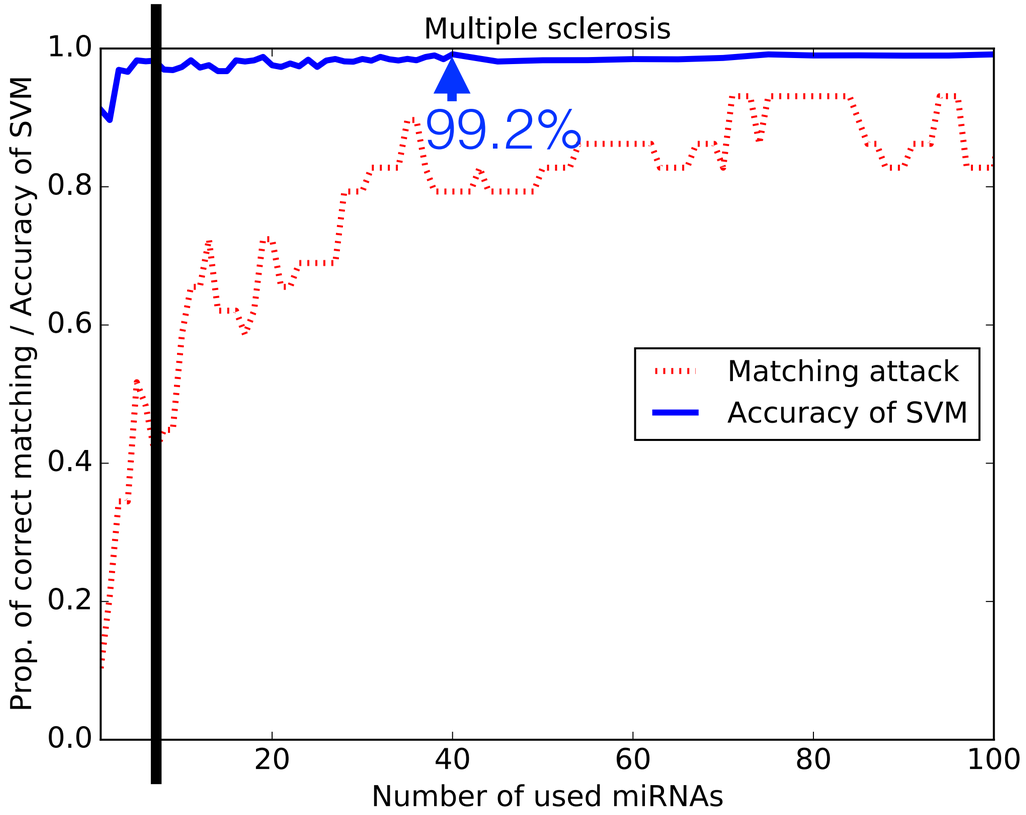
Assumption: adversary infers correlated miRNAs perfectly

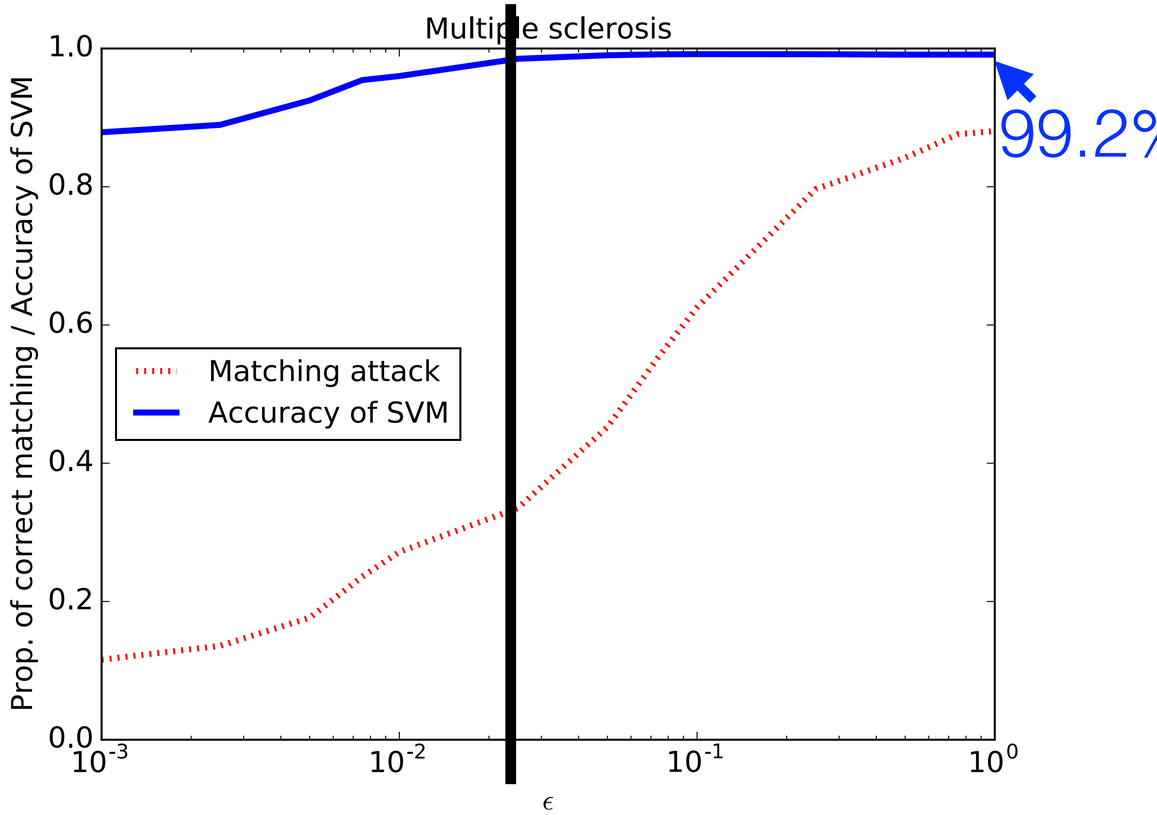


Probabilistic Sanitization – Results



Probabilistic Sanitization – Results



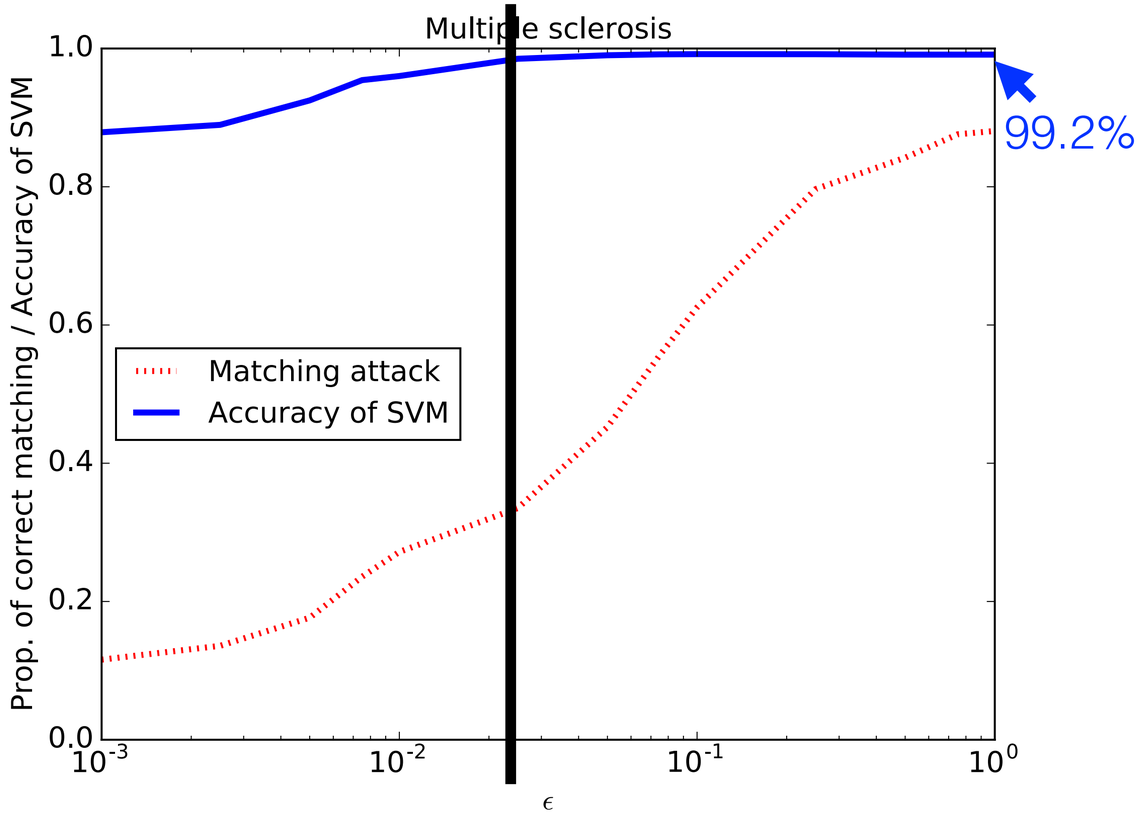


Probabilistic Sanitization – Results

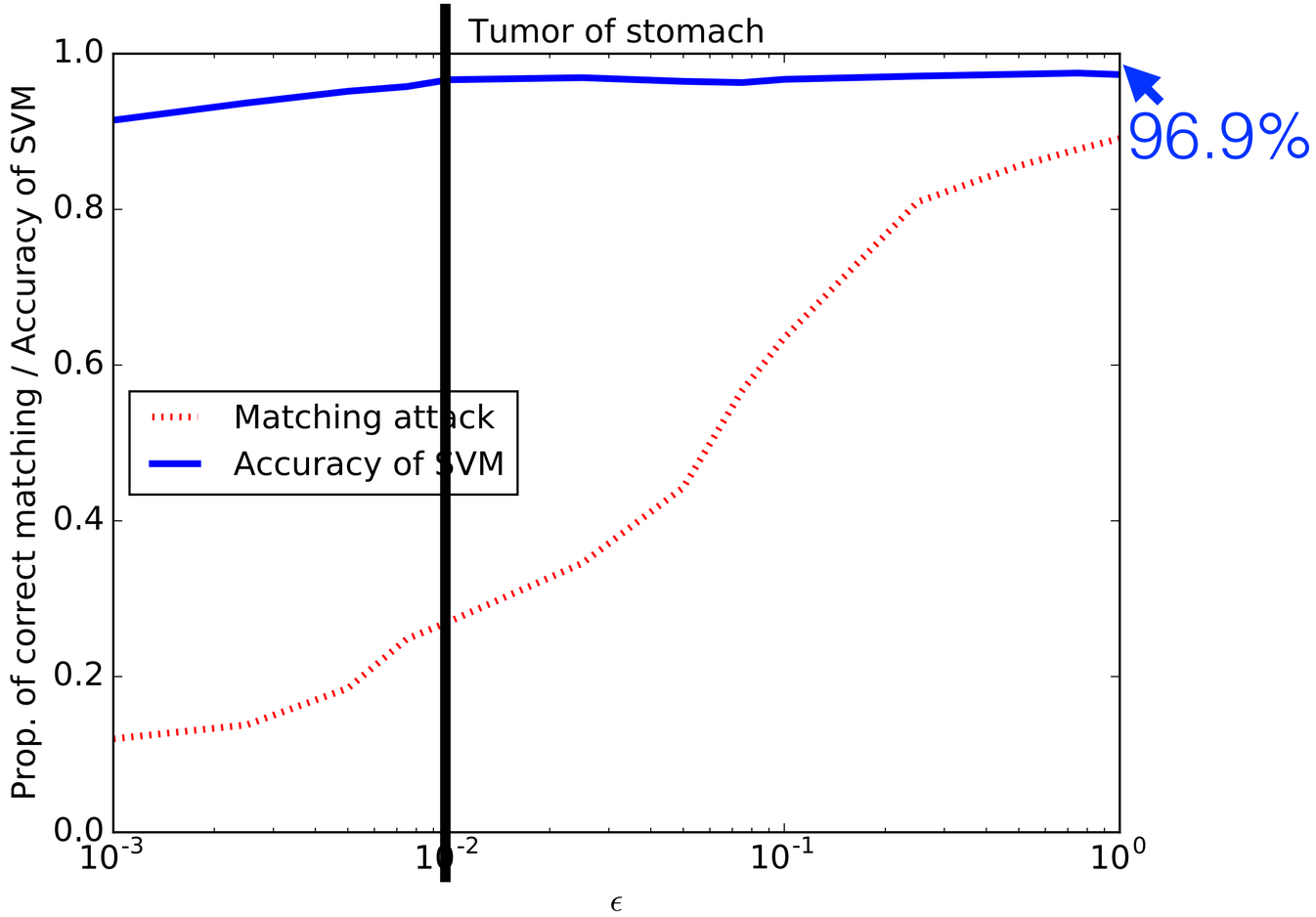
trade-off at ε=0.025

success decreases (relative to all)
by **63%** *(hiding 54%)*

accuracy decreases (relative to max)
by only **0.65%** *(hiding 1%)*



Probabilistic Sanitization – Results

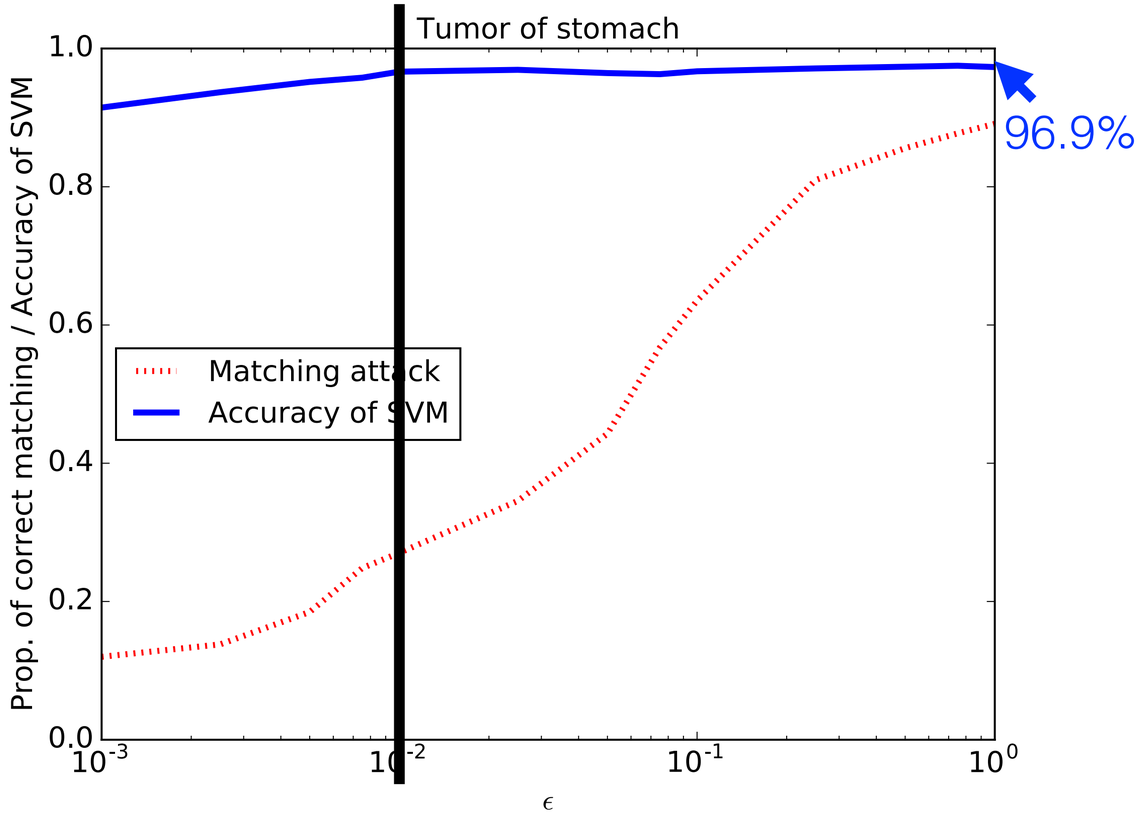


Probabilistic Sanitization – Results

trade-off at ε=0.01

success decreases (relative to all)
by **70%**

accuracy decreases (relative to max)
by only **0.2%**



Revisiting Research Questions – Conclusion

CISPA Center for IT-Security, Privacy and Accountability

Thank you!
Questions?

- Are miRNA expression profiles vulnerable to **linkability attacks**?
  - ➡ Yes, up to **90%** success rate for blood-based samples!
- If yes, how can we **protect** this kind of data?
  - ➡ Noising can **reduce the success** rate to almost random guessing.
- How does the protection affect the **utility** of the data?
  - ➡ Noising in general provides a **better** privacy-utility trade-off.
  - ➡ **Privacy doubles**, while utility cost is less than 1% for most diseases!

