SAARLAND UNIVERSITY
COMPUTER SCIENCE

CISPA
Center for IT-Security, Privacy and Accountability

On Epigenomic Privacy: Tracking Personal MicroRNA Expression Profiles over Time

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

21st February 2016

On Epigenomic Privacy: Tracking Personal MicroRNA Expression Profiles over Time

Epigenetics

"epi": above, over (greek)
"genetics": origin (greek)

Definition: study of cellular and phenotypic trait variations stemming from **other causes than changes in the genotype**

external factors such as:
in-utero and childhood development, environmental chemicals, aging, diet.

MicroRNA (miRNA)

discovered in the early 1990s

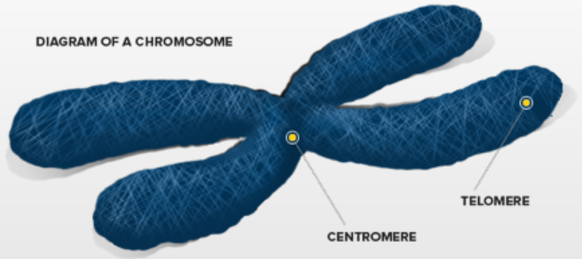
Definition: small non-coding RNA molecules that regulate gene expression in plants/animals
60% of genes coding human proteins are regulated by miRNAs

MicroRNA Expression Profiles

Real-valued number quantifying whether and how much miRNAs are active in a given set of cells/tissue.

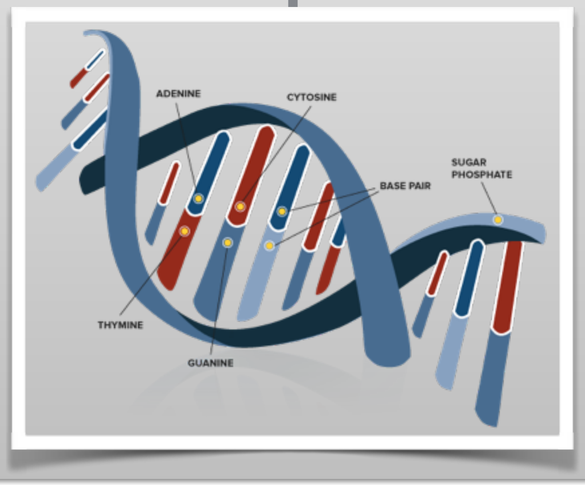
What is the purpose of MicroRNAs?

SAARLAND UNIVERSITY
COMPUTER SCIENCE
CISPA Center for IT-Security, Privacy and Accountability



Chromosomes: carry hereditary information in long strings of **DNA** called **genes**

(a region of DNA)



But all cells have the same **genes**!

What makes the cells different:

gene expression

(which genes are active in a cell)

Graphics: genographic.nationalgeographic.com

What is the purpose of MicroRNAs?

What makes the cells different:

gene expression
(which genes are active in a cell)

miRNAs regulate most of human genes!

↳ important for normal and **disease** cells

neurodegenerative diseases (e.g., Alzheimer's)
heart diseases, diabetes, majority of cancers

More on DNA and MicroRNAs!

SAARLAND UNIVERSITY
COMPUTER SCIENCE
CISPA Center for IT-Security, Privacy and Accountability

DNA

- contains receipts what a cell potentially can do,
- is (mostly) fixed over time,
- can hint on risks of getting a disease,
- has been researched a lot.

miRNAs

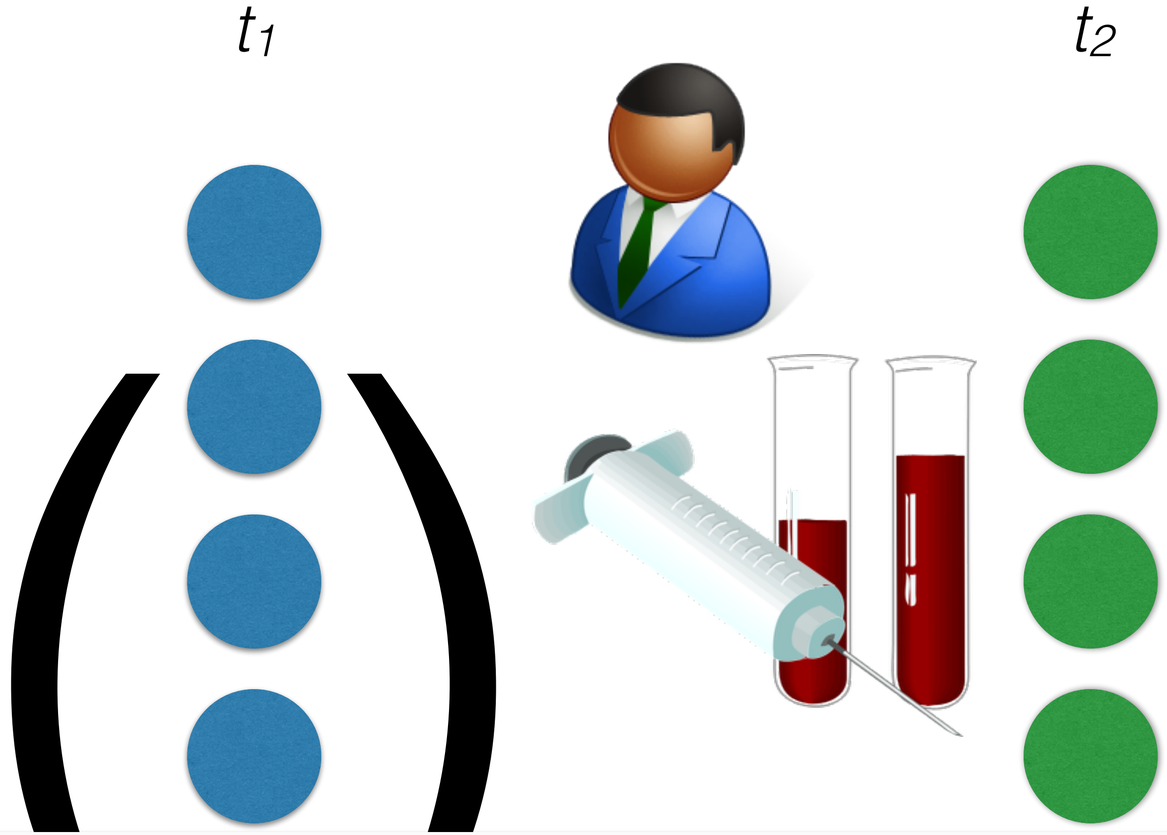
- expression regulates what a cell really does,
- expression changes over time,
- can tell whether you carry a disease,
- so far, have been largely overlooked (in privacy)!

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

identification

matching

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



hospital server

black market

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

Athletes' dataset

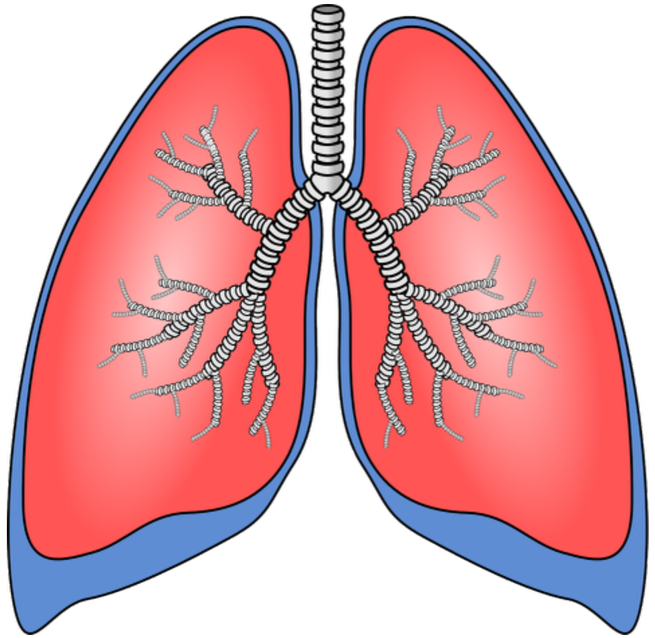
Participants: **29**
Points in time: **2** (before and after exercising)
Time shift: **1 week**
Disease: **none**

blood-based

plasma-based

1,189 miRNAs per sample

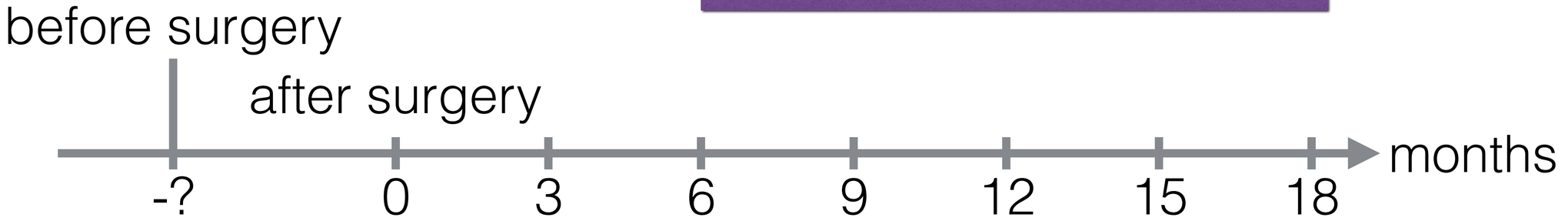
Lung cancer dataset

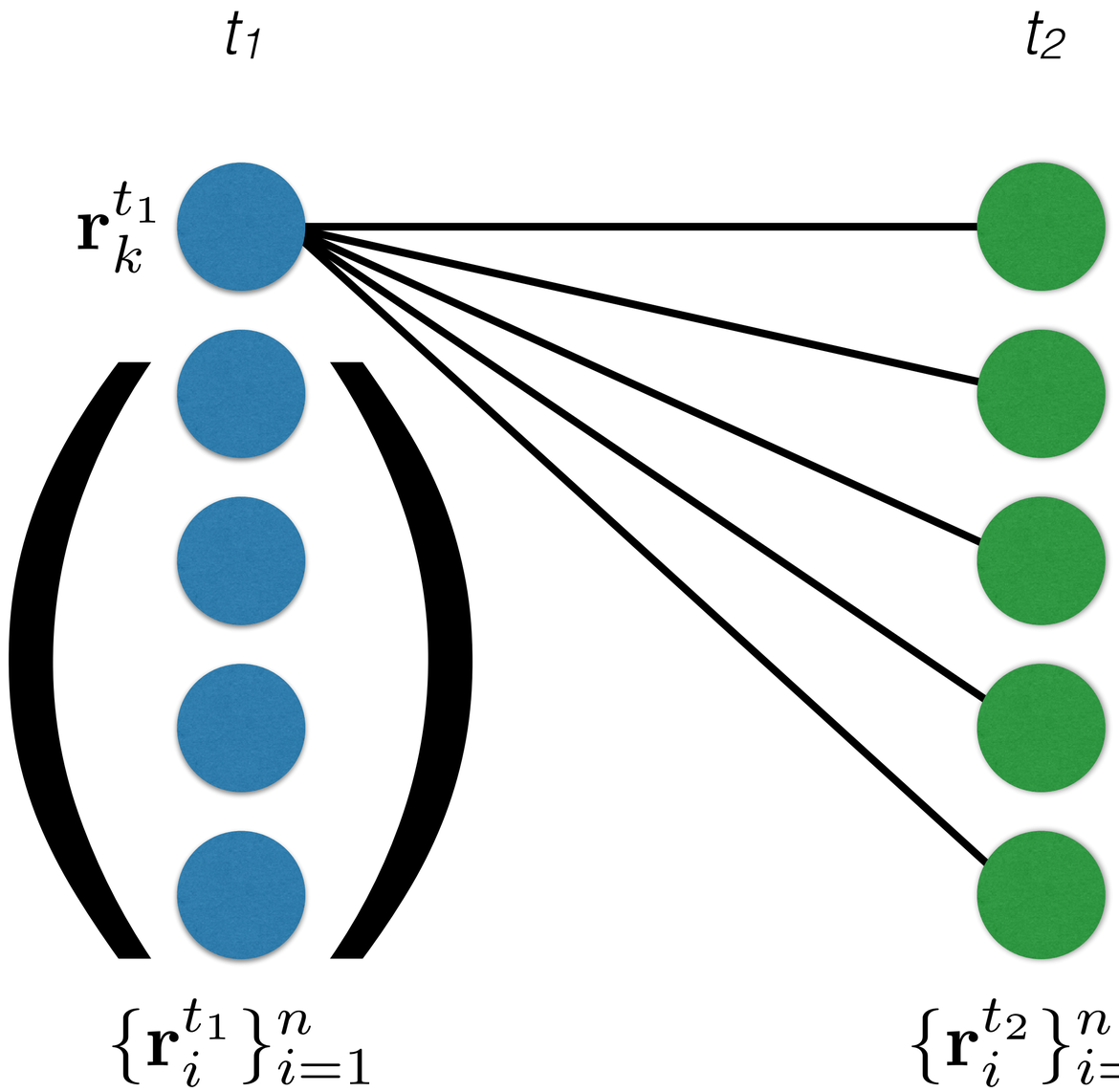


Participants: **26**
Points in time: **8**
Time shift: mostly **3 months**
Disease: **lung cancer**

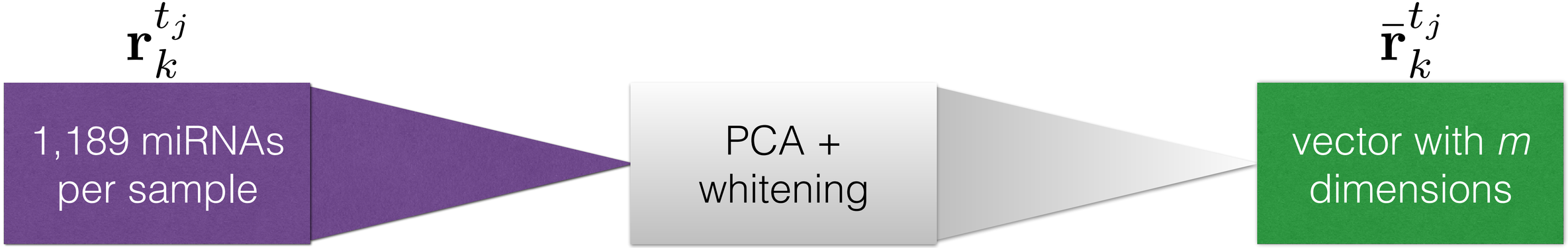
plasma-based

1,189 miRNAs per sample



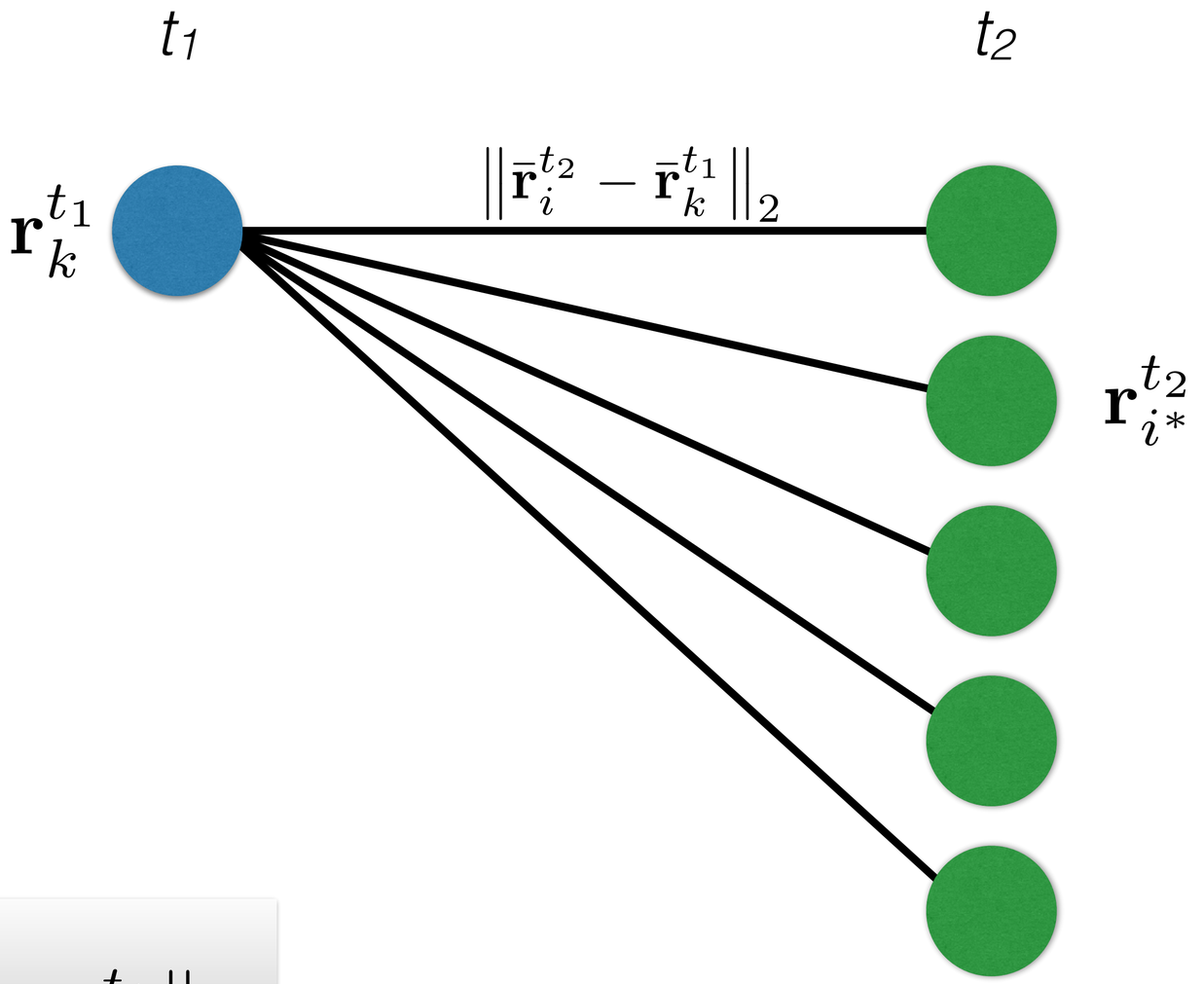


1,189 miRNAs per sample



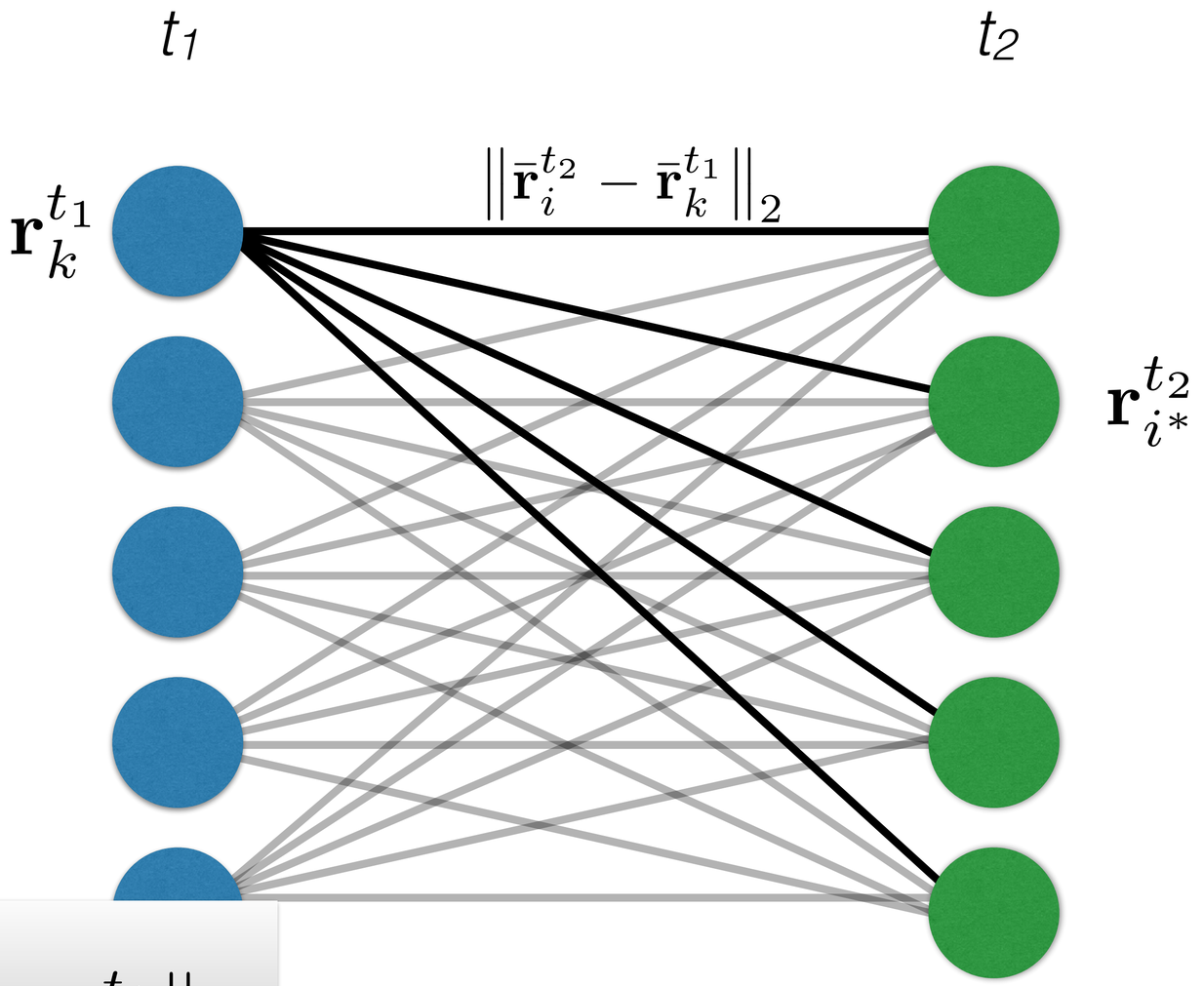
whitening: unit variance
PCA: smaller dimensionality *m*
+ uncorrelated components

Identification Attack



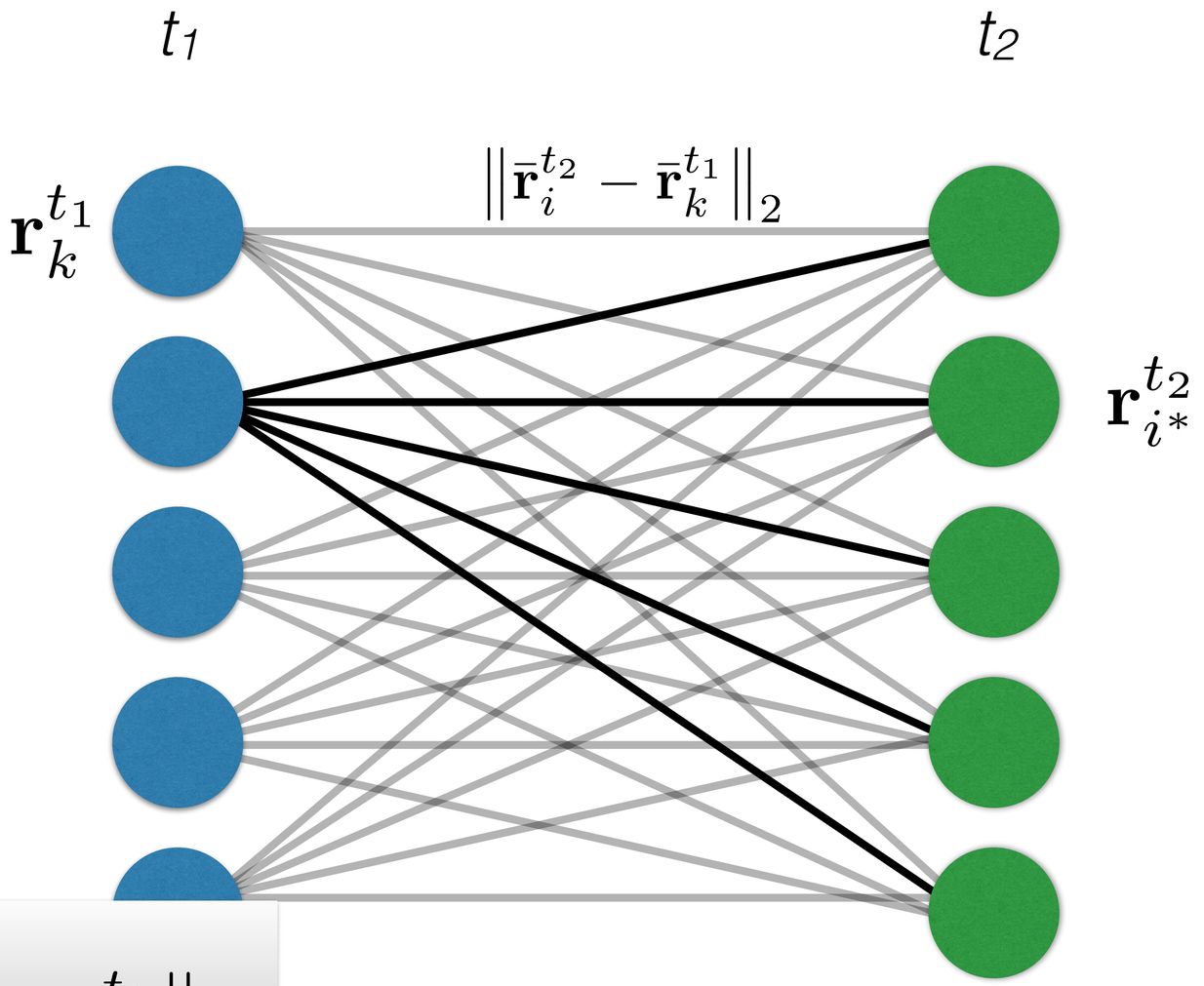
$$i^* = \arg\min_i \left\|\bar{\mathbf{r}}_i^{t_2} - \bar{\mathbf{r}}_k^{t_1}\right\|_2$$

Identification Attack



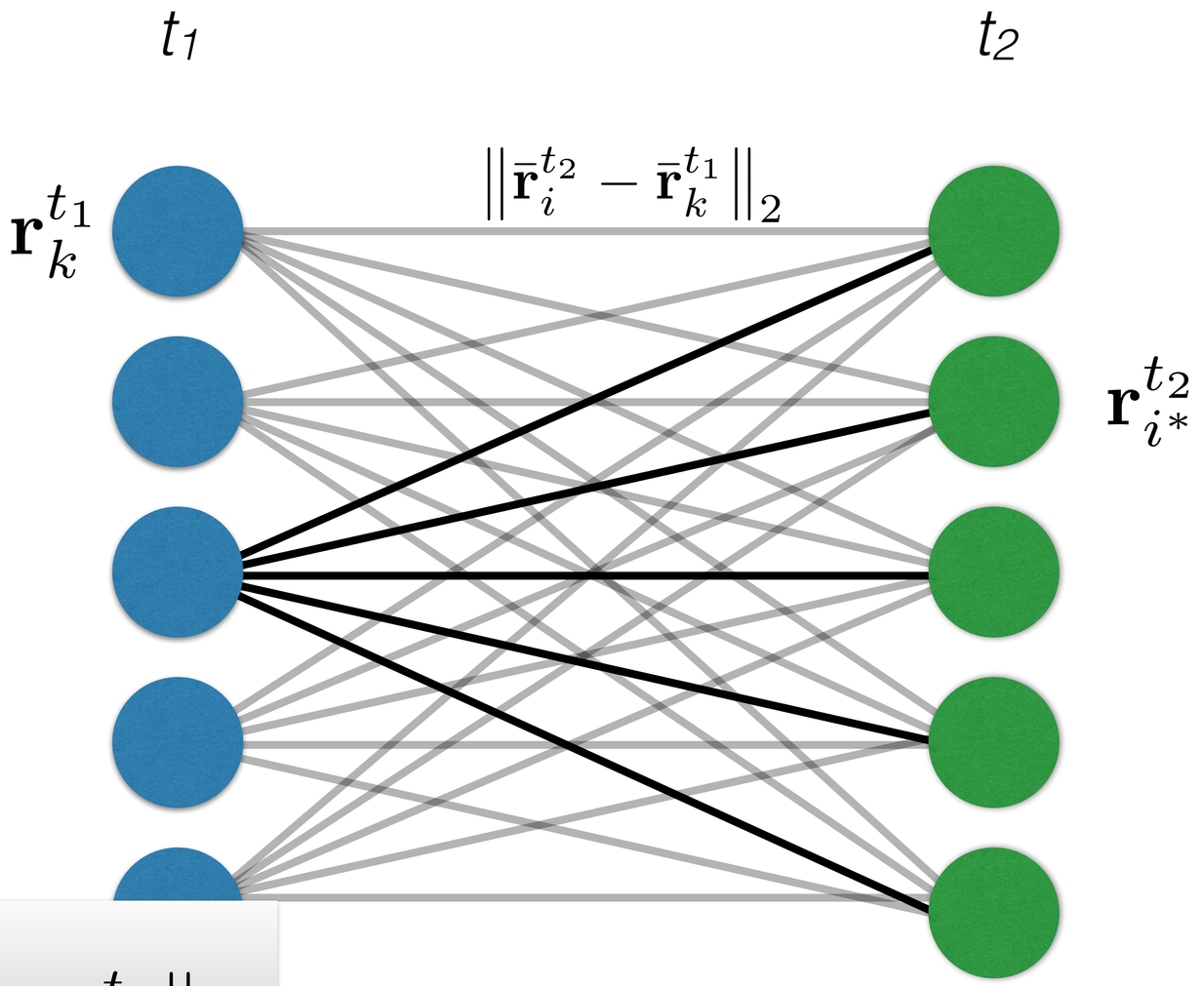
$$i^* = \arg\min_i \left\|\bar{\mathbf{r}}_i^{t_2} - \bar{\mathbf{r}}_k^{t_1}\right\|_2$$

Identification Attack



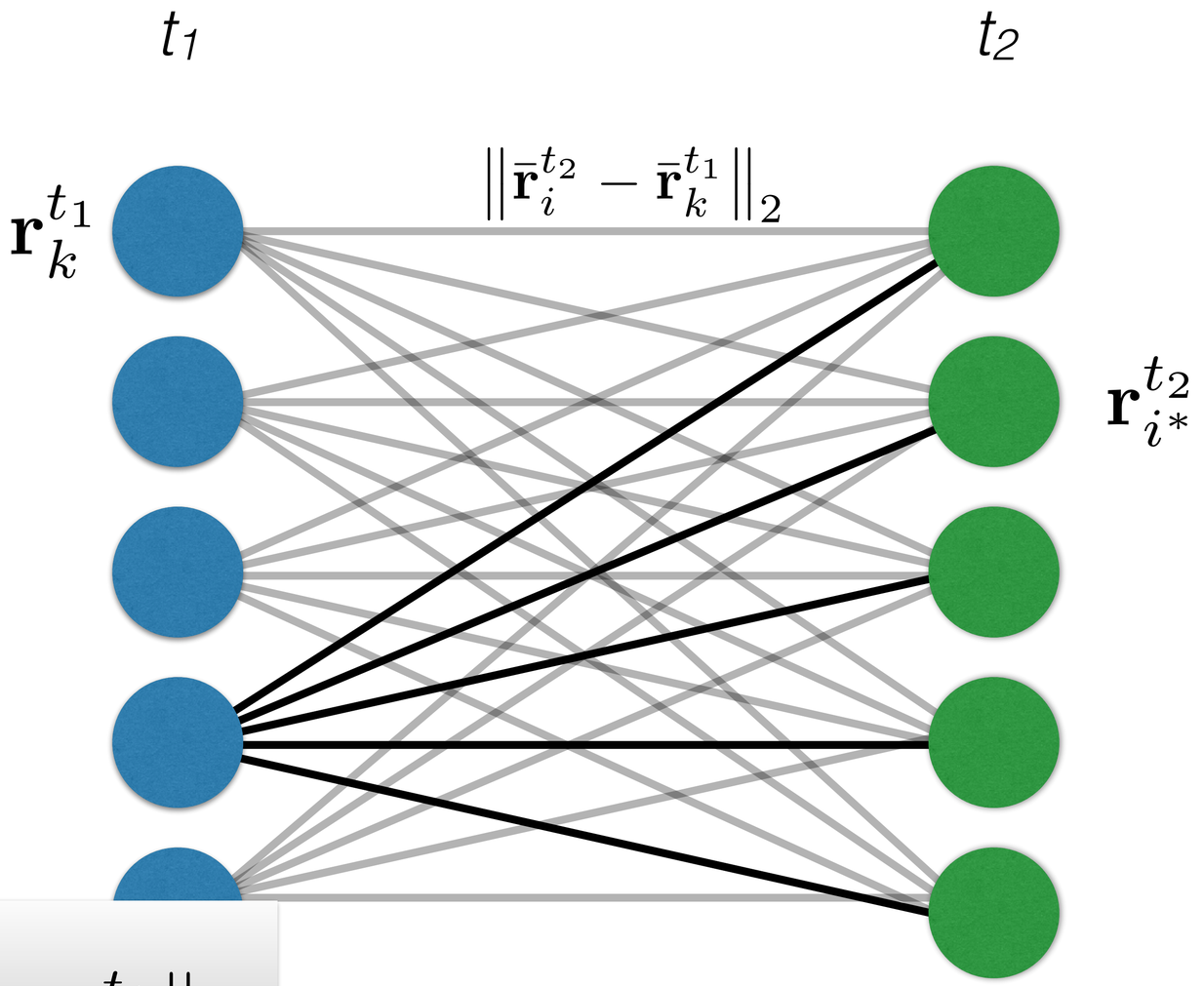
$$i^* = \arg\min_i \left\|\bar{\mathbf{r}}_i^{t_2} - \bar{\mathbf{r}}_k^{t_1}\right\|_2$$

Identification Attack



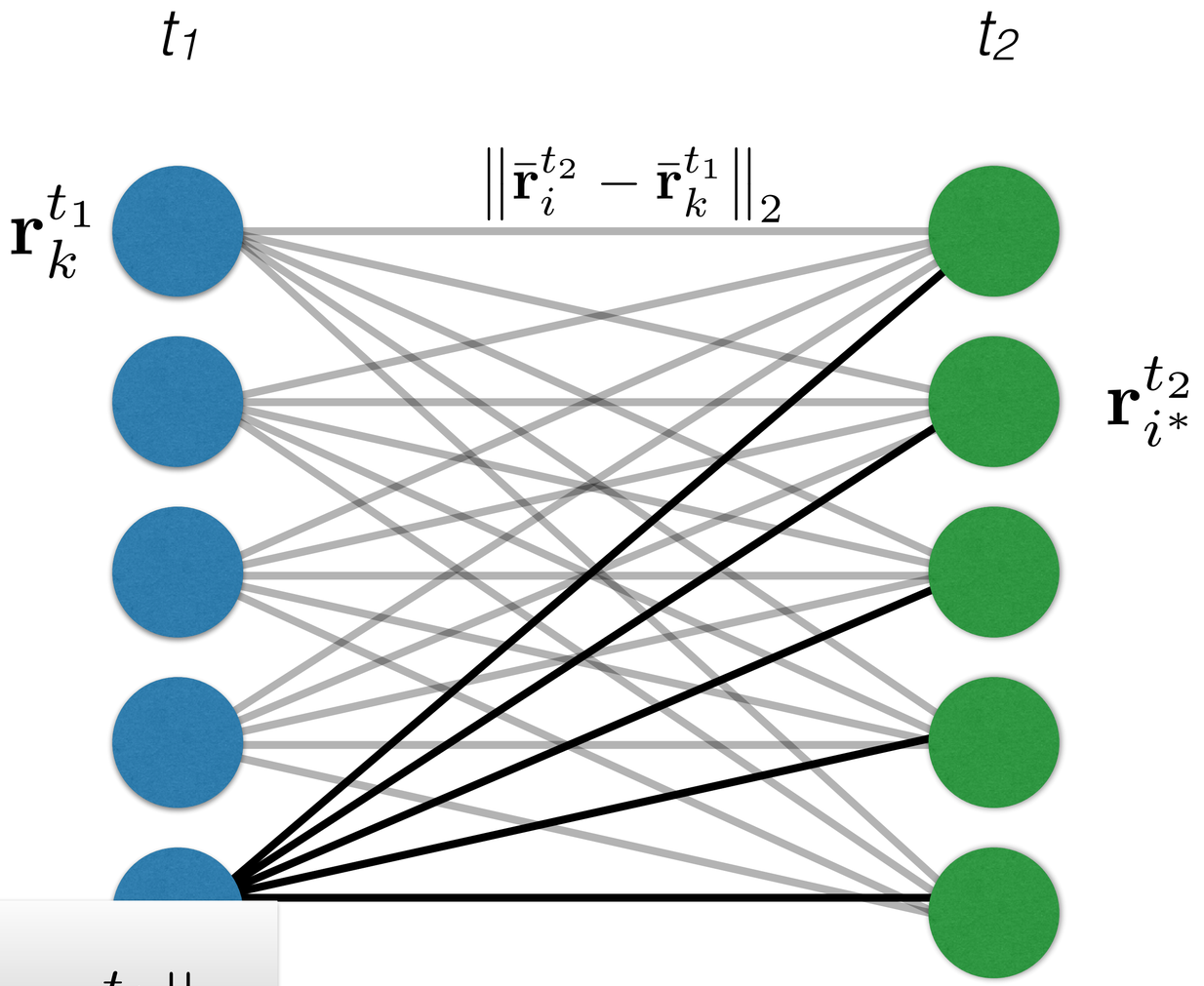
$$i^* = \arg\min_i \left\|\bar{\mathbf{r}}_i^{t_2} - \bar{\mathbf{r}}_k^{t_1}\right\|_2$$

Identification Attack



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

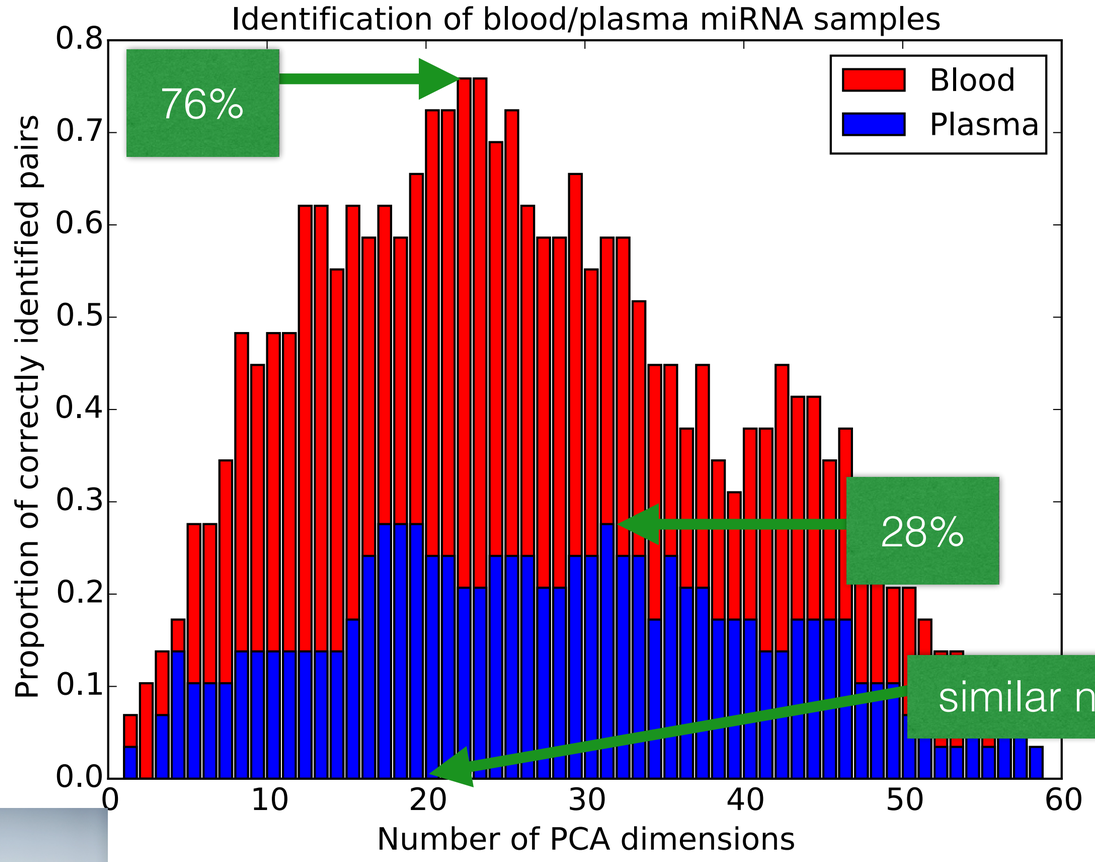
Identification Attack

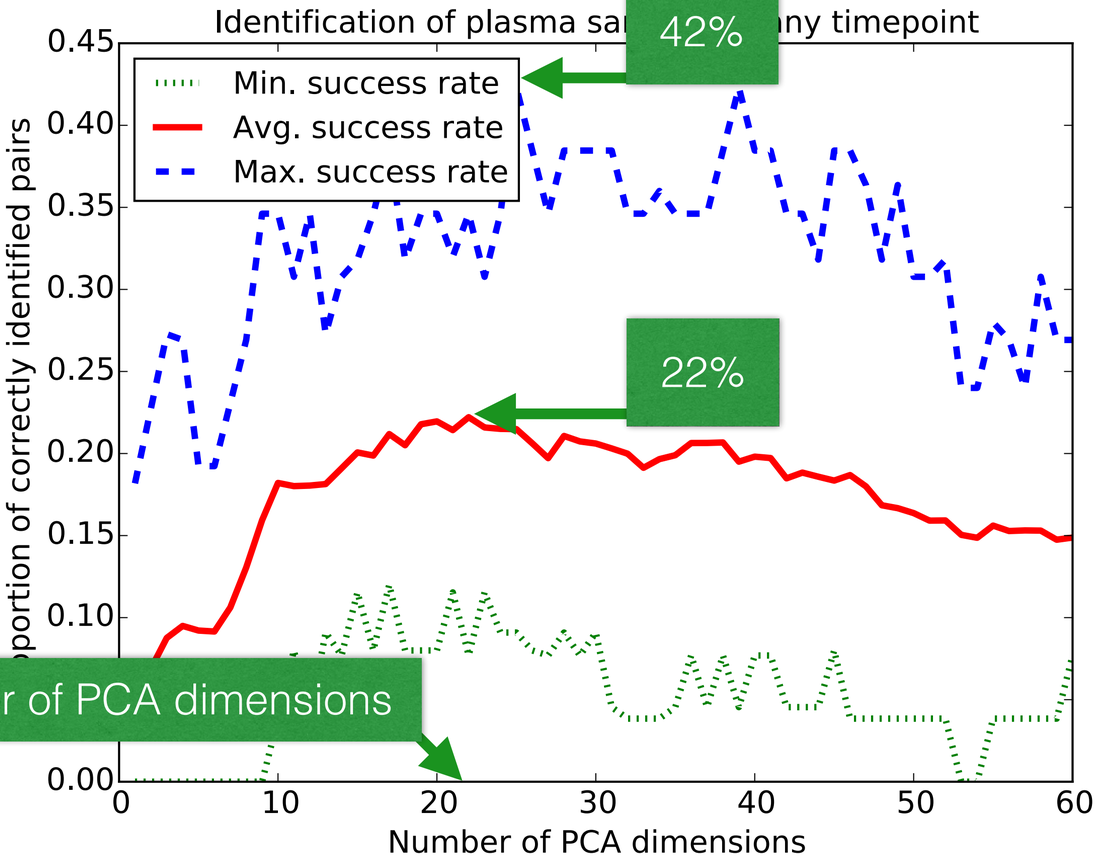


$$i^* = \arg\min_i \left\|\bar{\mathbf{r}}_i^{t_2} - \bar{\mathbf{r}}_k^{t_1}\right\|_2$$

Identification Attack

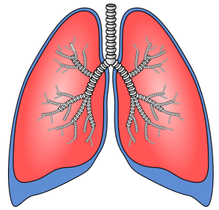
SAARLAND UNIVERSITY
COMPUTER SCIENCE
CISPA Center for IT-Security, Privacy and Accountability



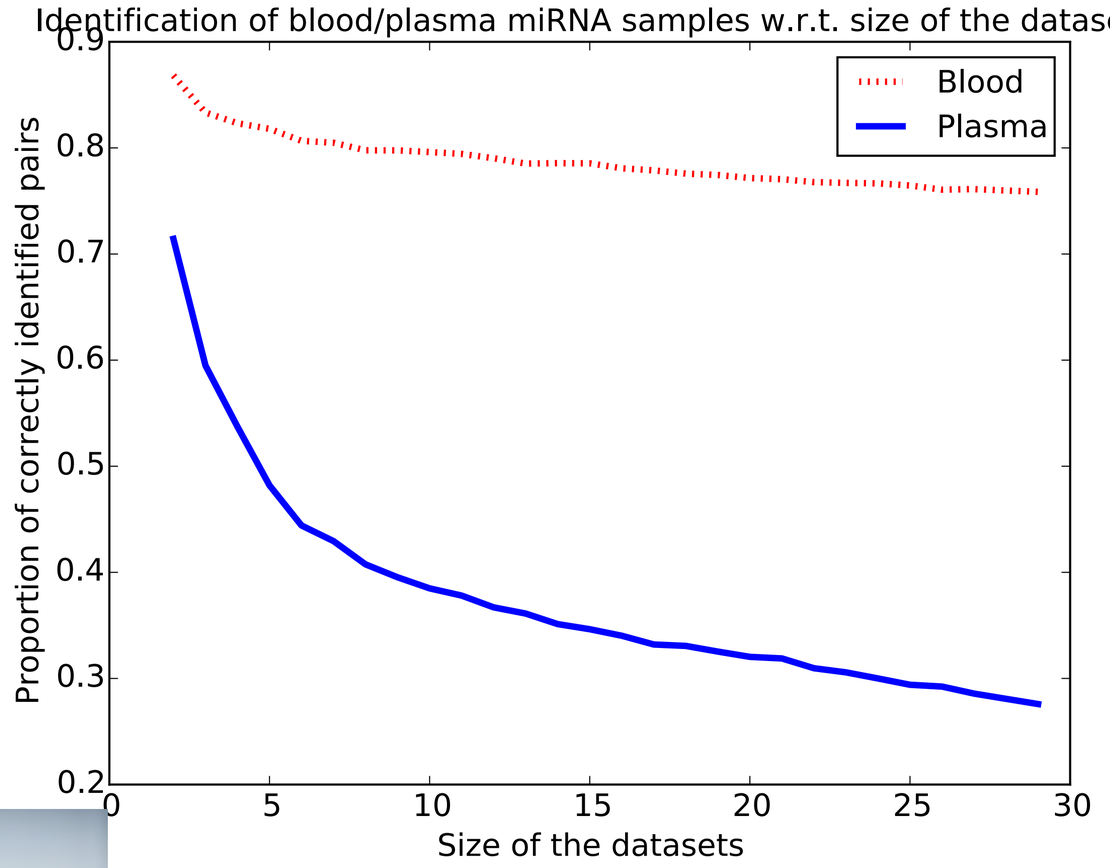


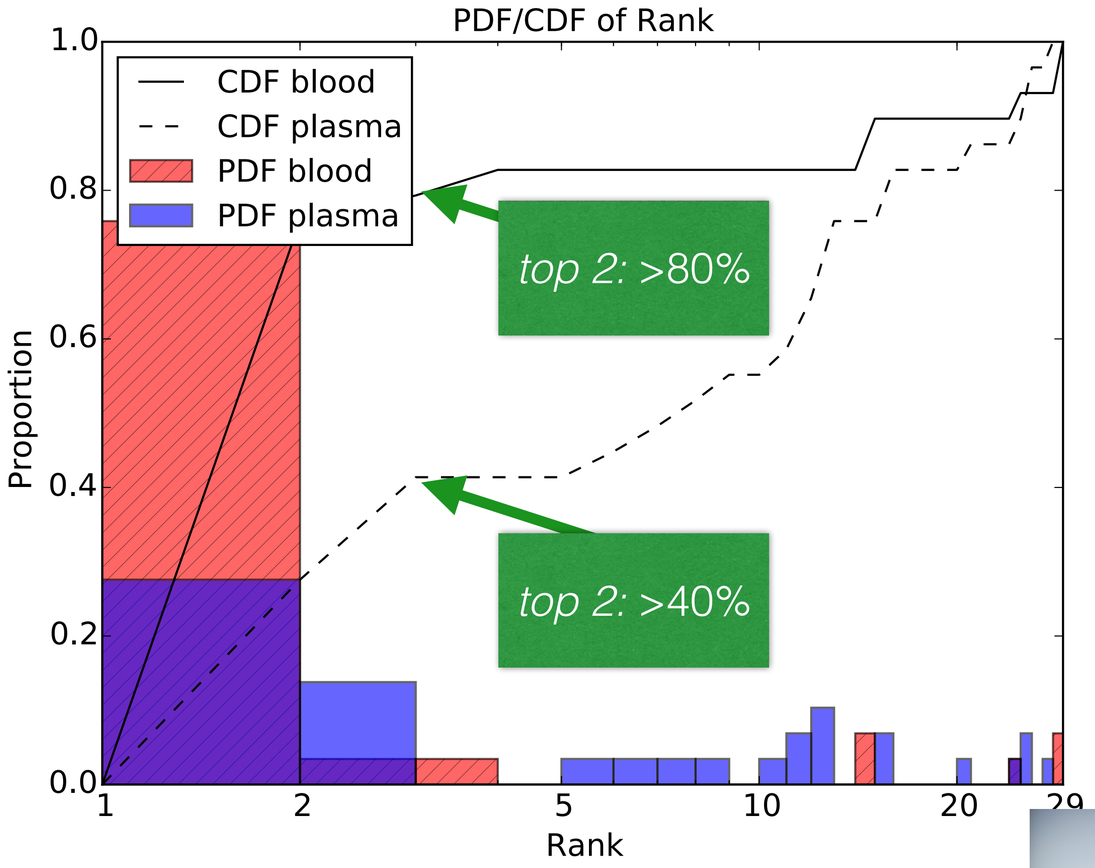
similar number of PCA dimensions

80% overlap in top10 miRNAs of first PCA component

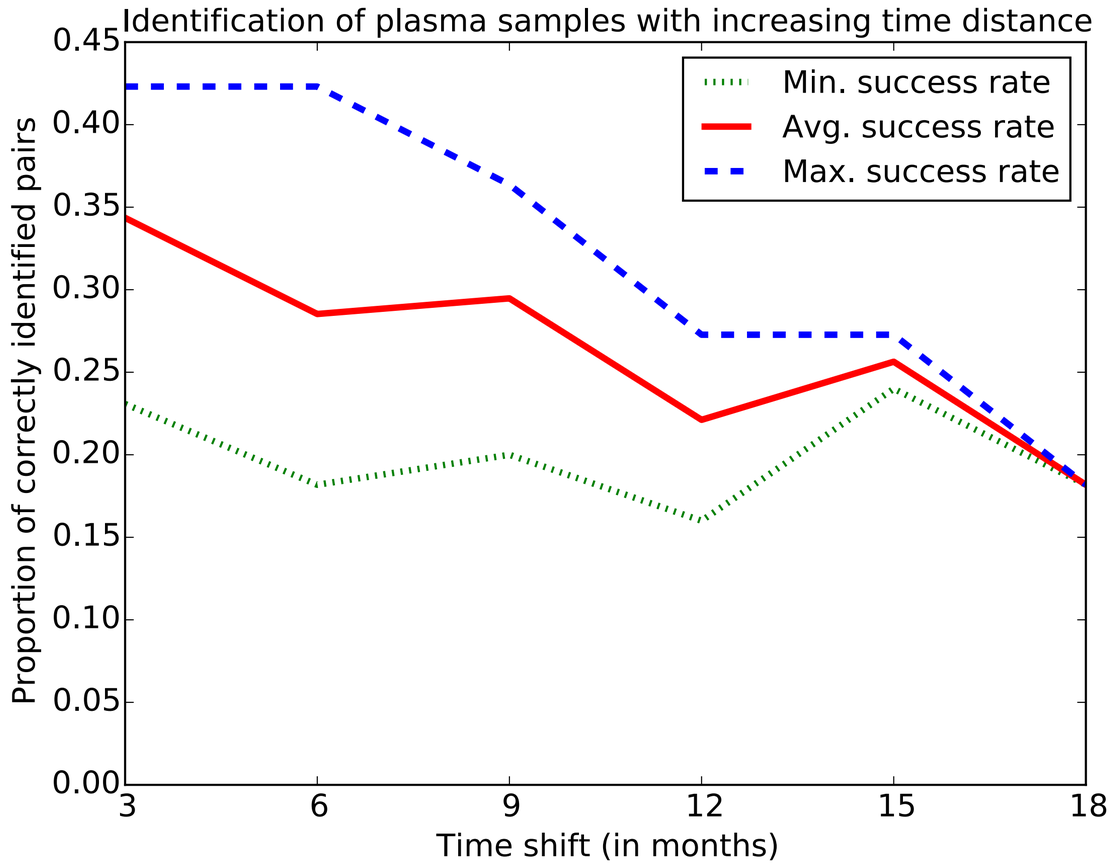


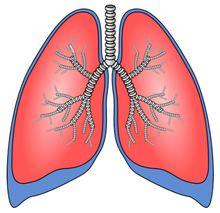
Identification Attack



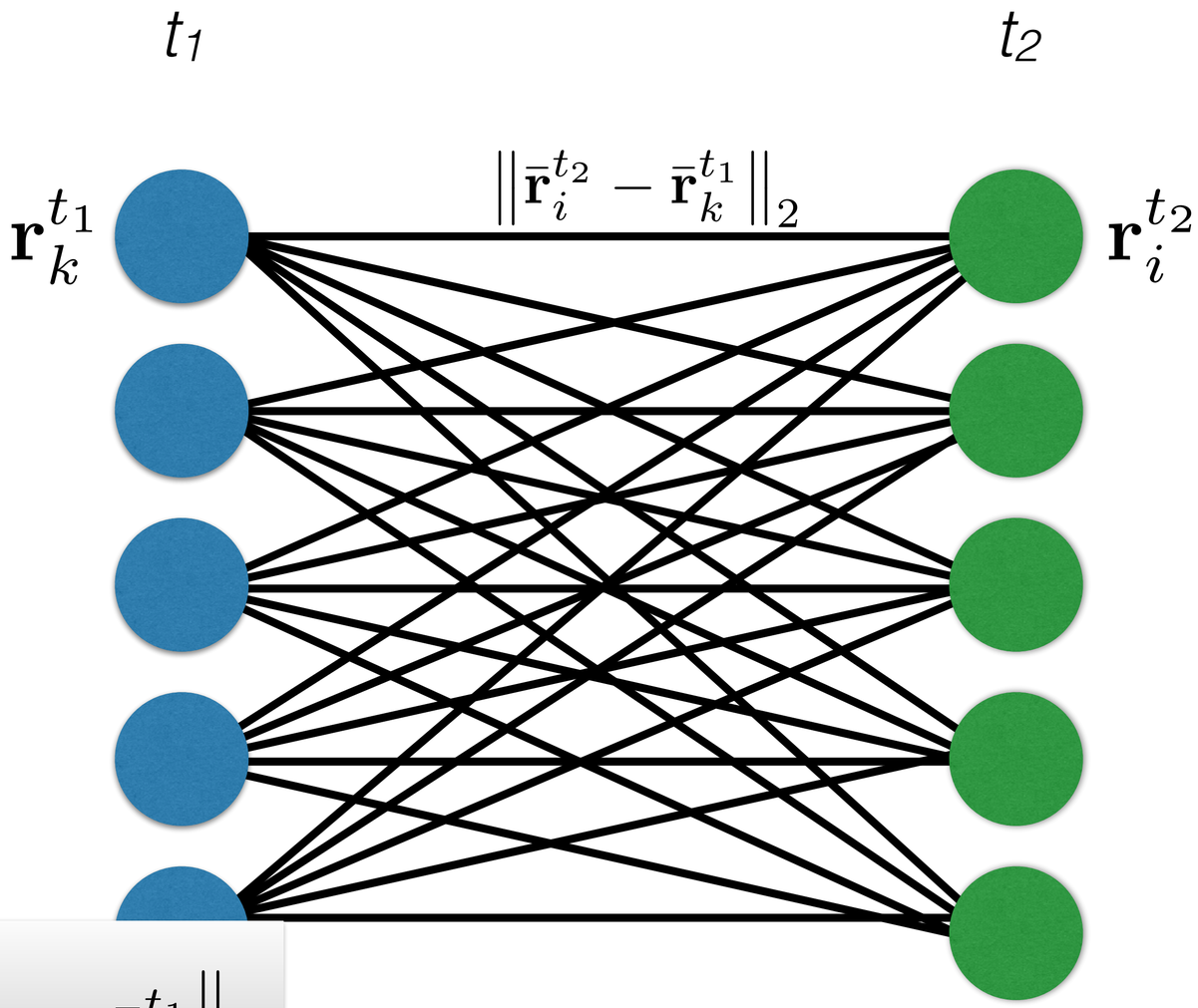


Identification Attack



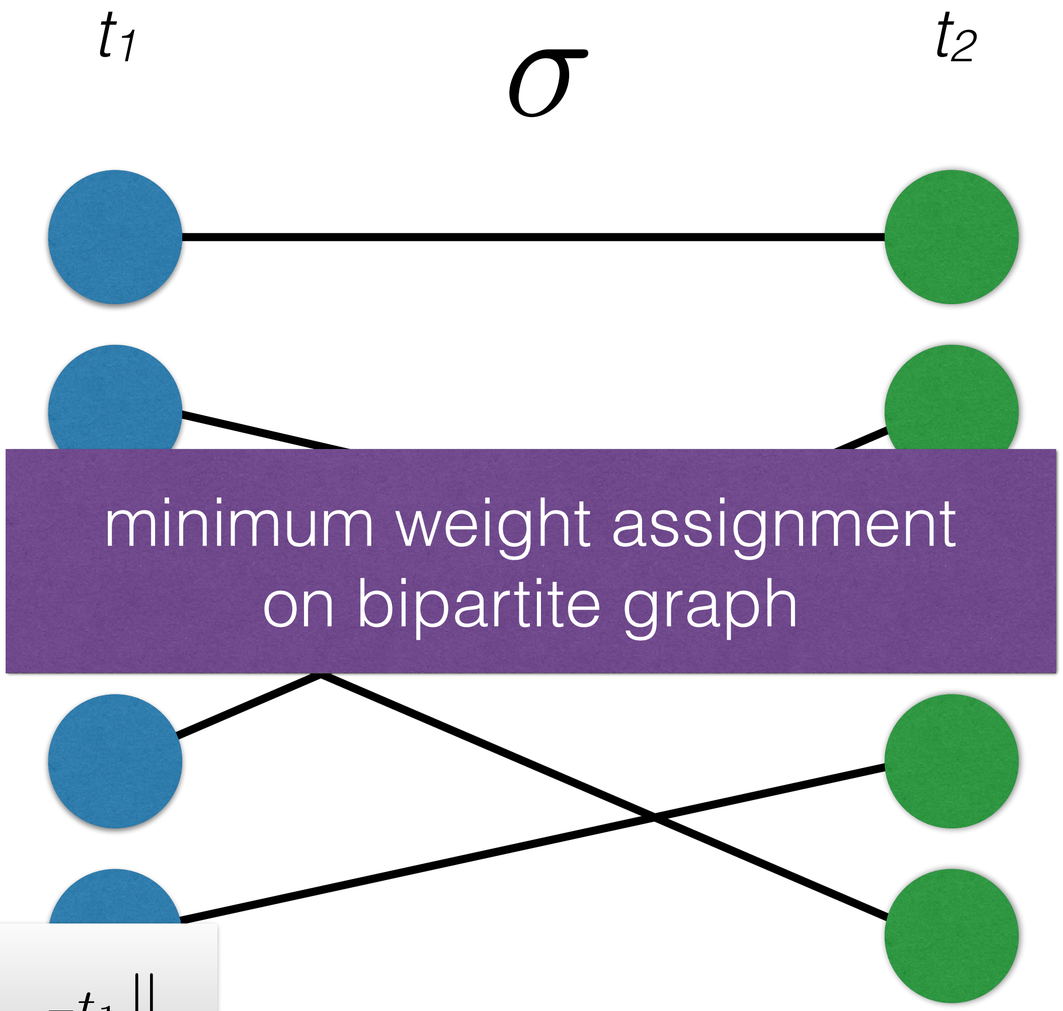


Matching Attack



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

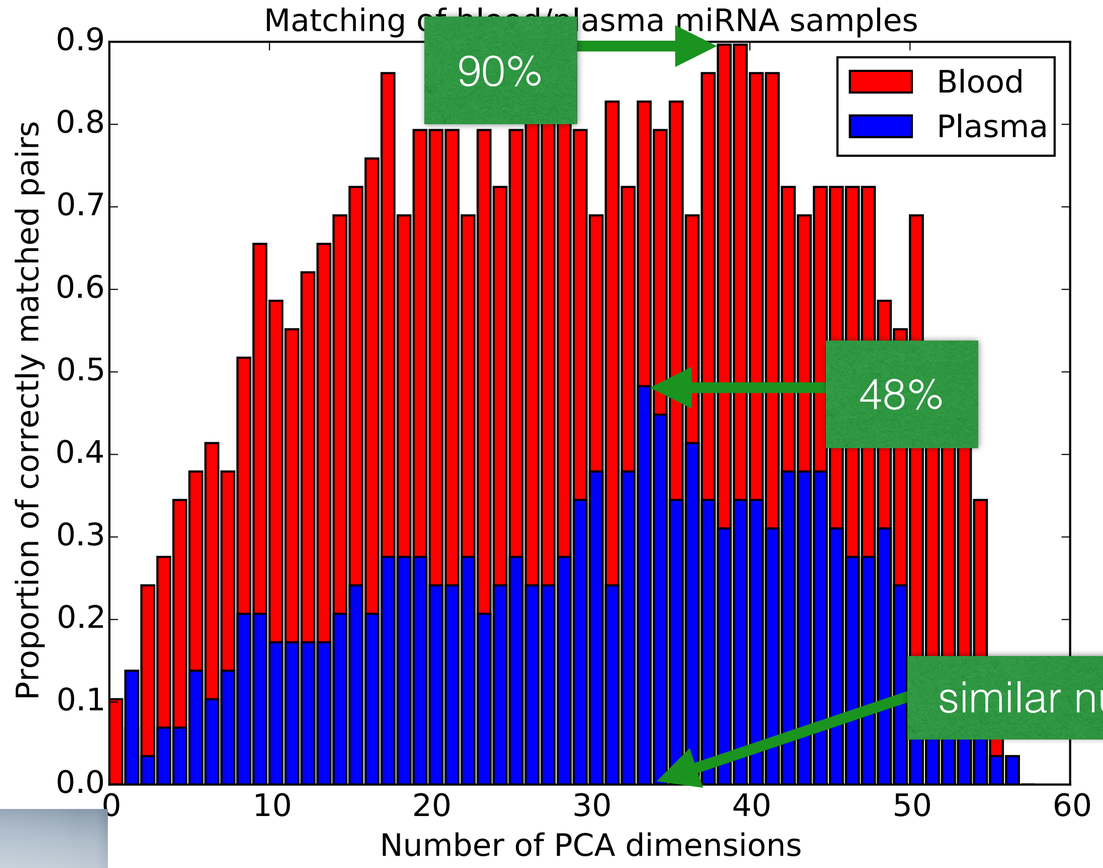
Matching Attack

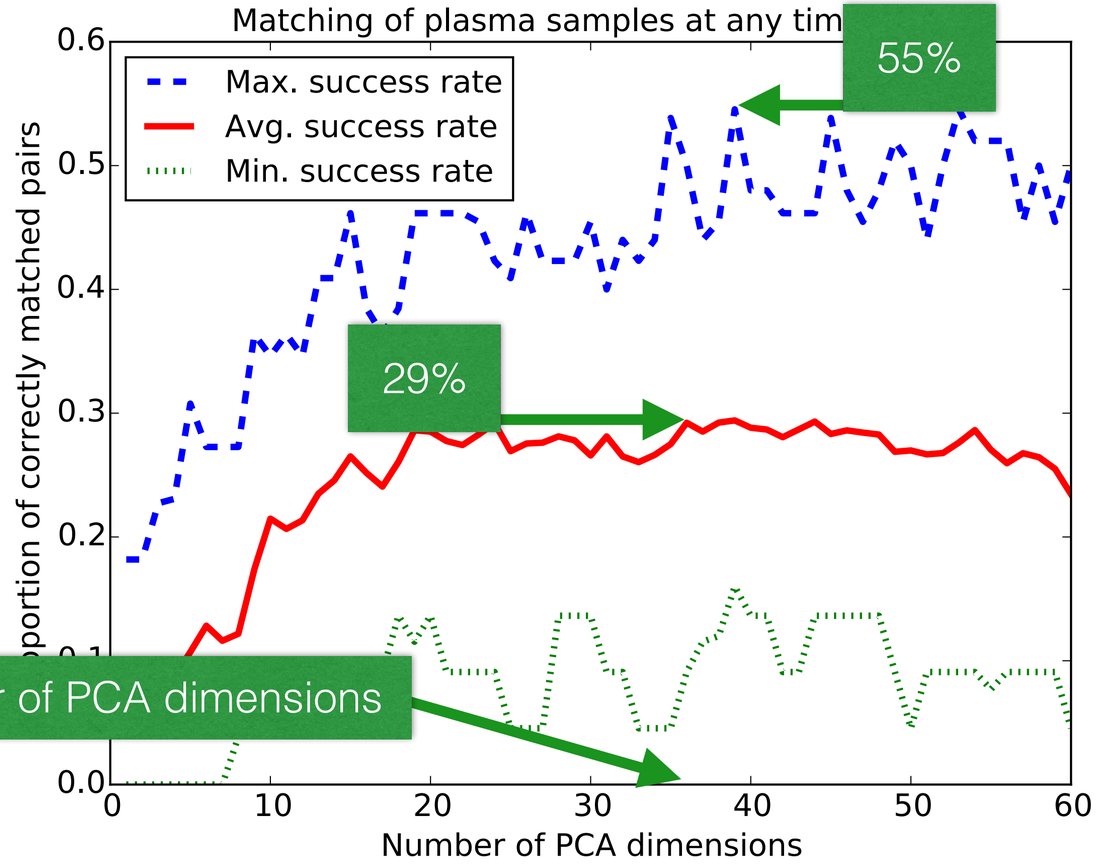


$$\sigma^* = \arg\min_{\sigma} \sum_{i=1}^{n} \left\| \bar{\mathbf{r}}_{\sigma(i)}^{t_2} - \bar{\mathbf{r}}_i^{t_1} \right\|_2$$

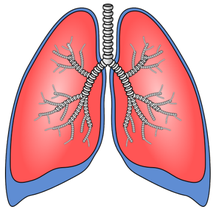
Matching Attack

SAARLAND UNIVERSITY
COMPUTER SCIENCE
CISPA Center for IT-Security, Privacy and Accountability

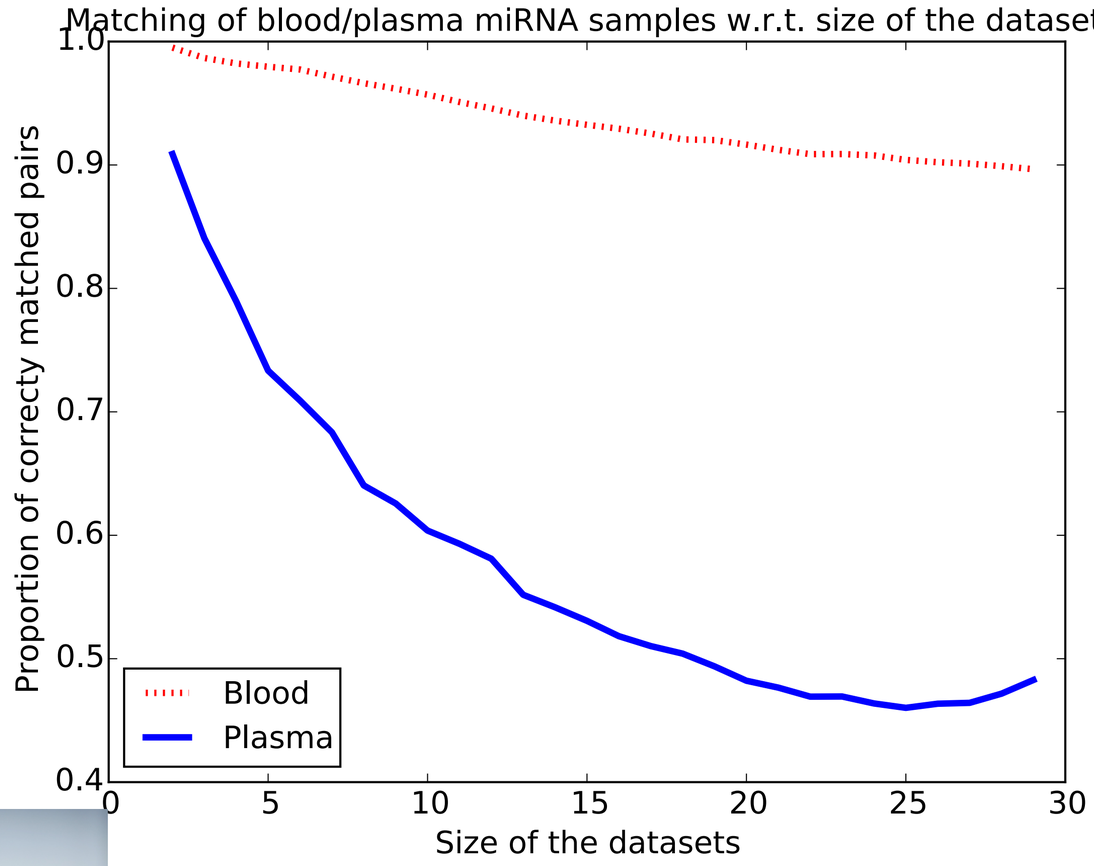


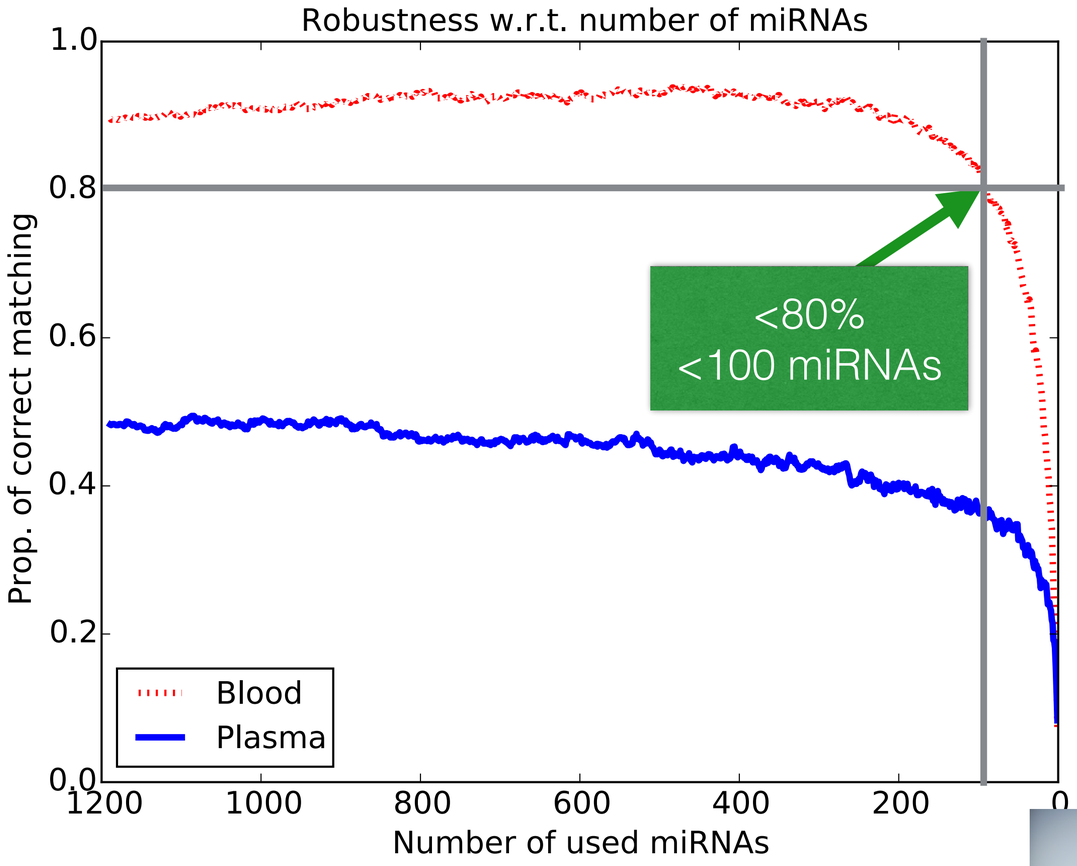


similar number of PCA dimensions

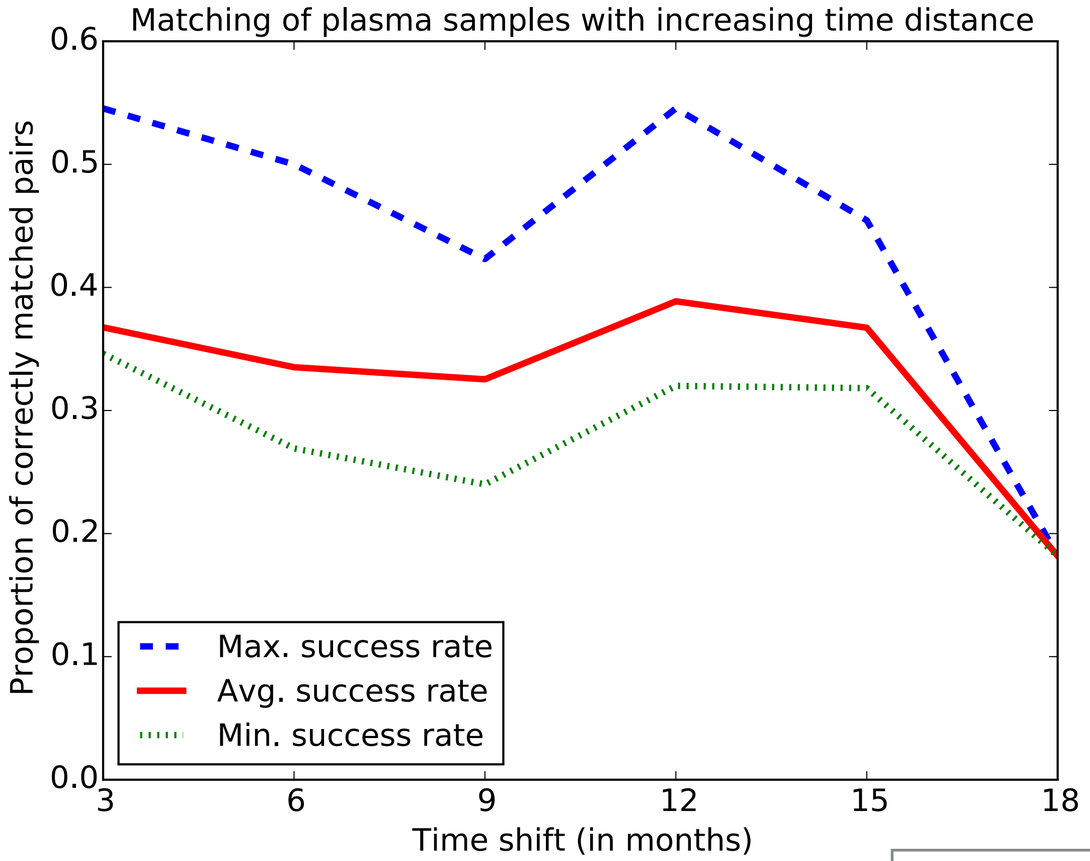


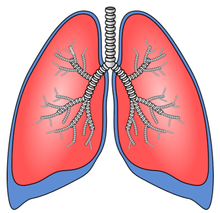
Matching Attack





Matching Attack



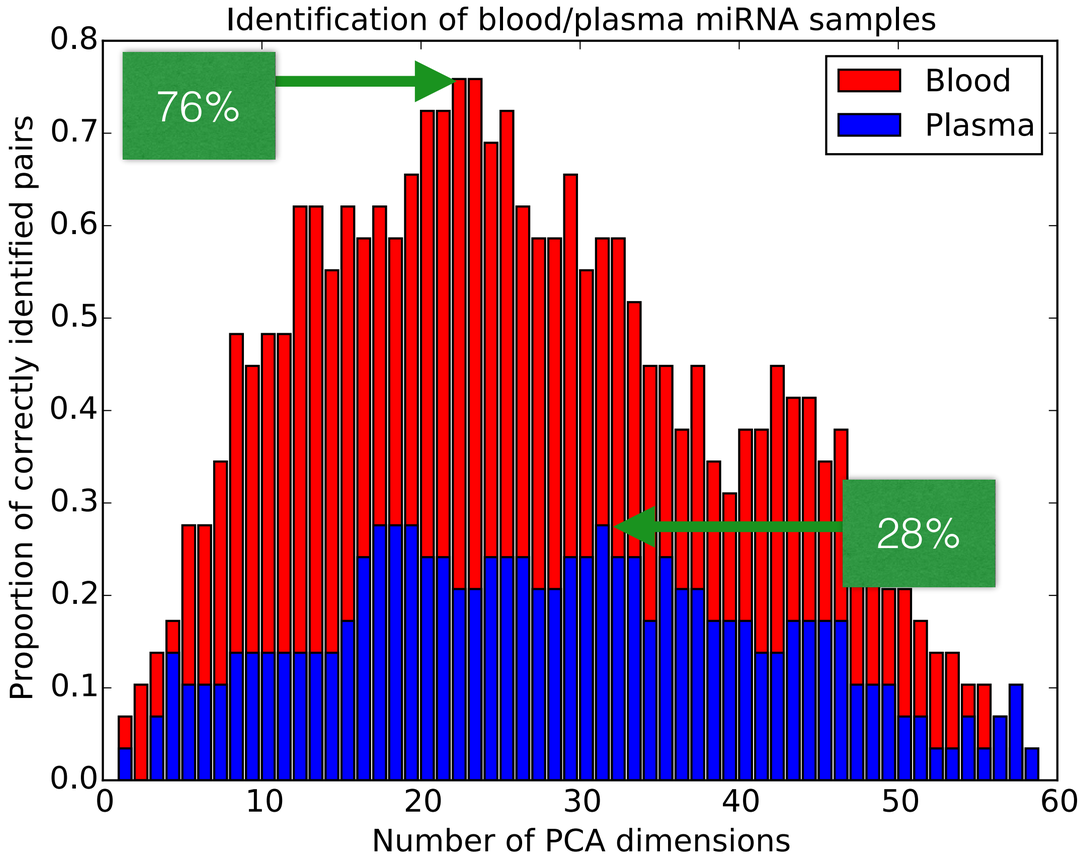


success rate remains more or less **constant in the first year**

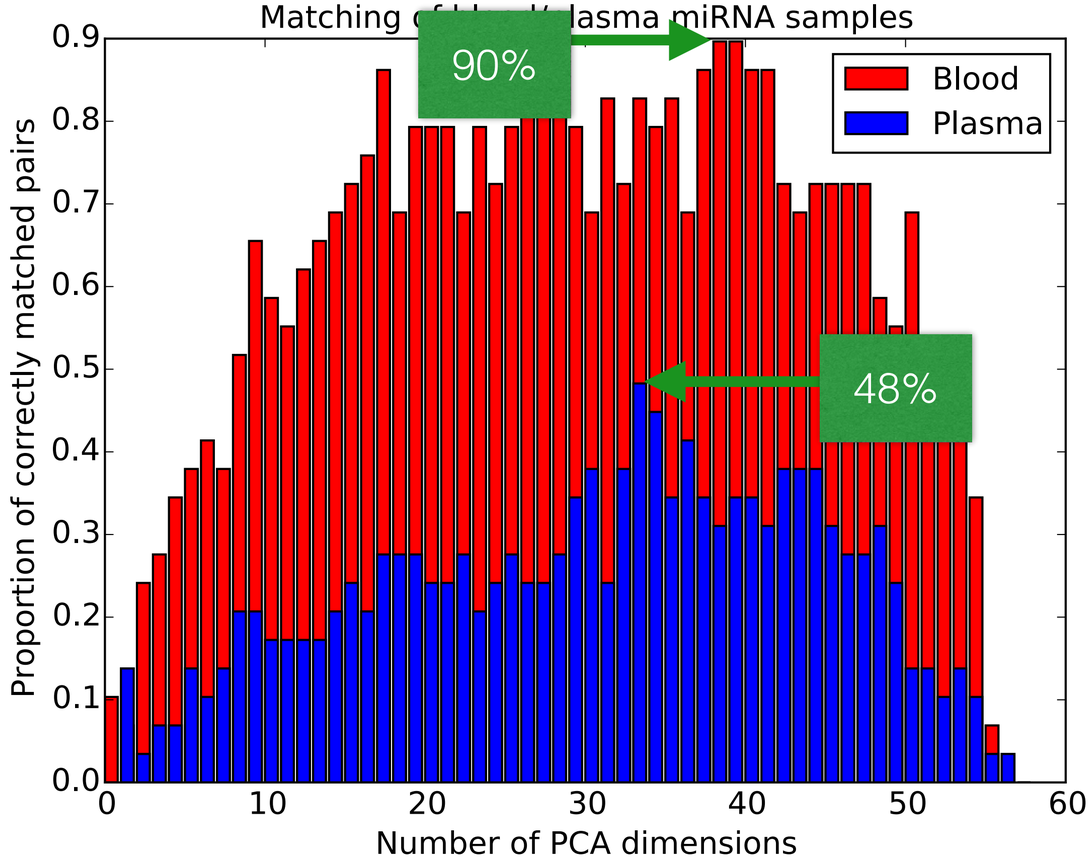
SAARLAND UNIVERSITY
COMPUTER SCIENCE

CISPA
Center for IT-Security, Privacy and Accountability

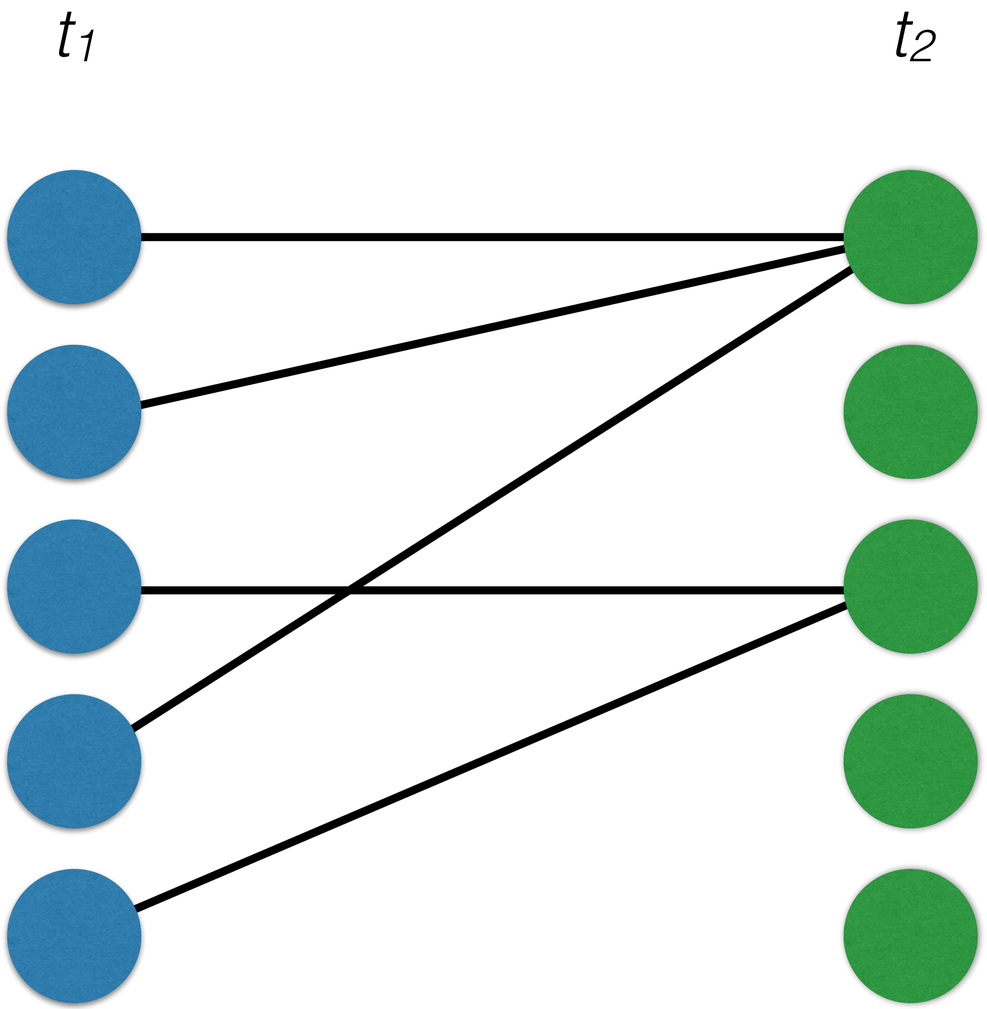
Identification Attack

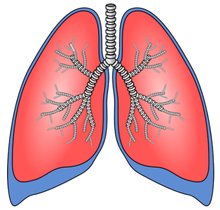


Matching Attack



Downside of Identification Attack

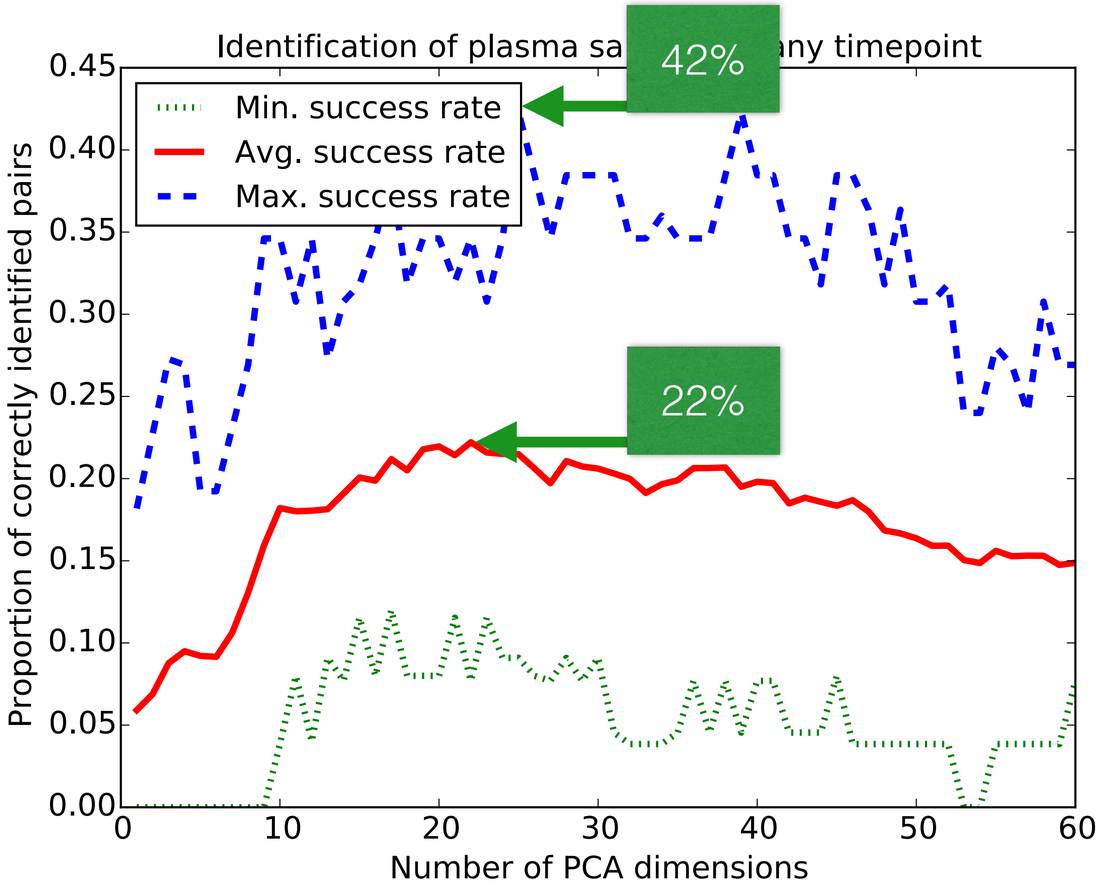




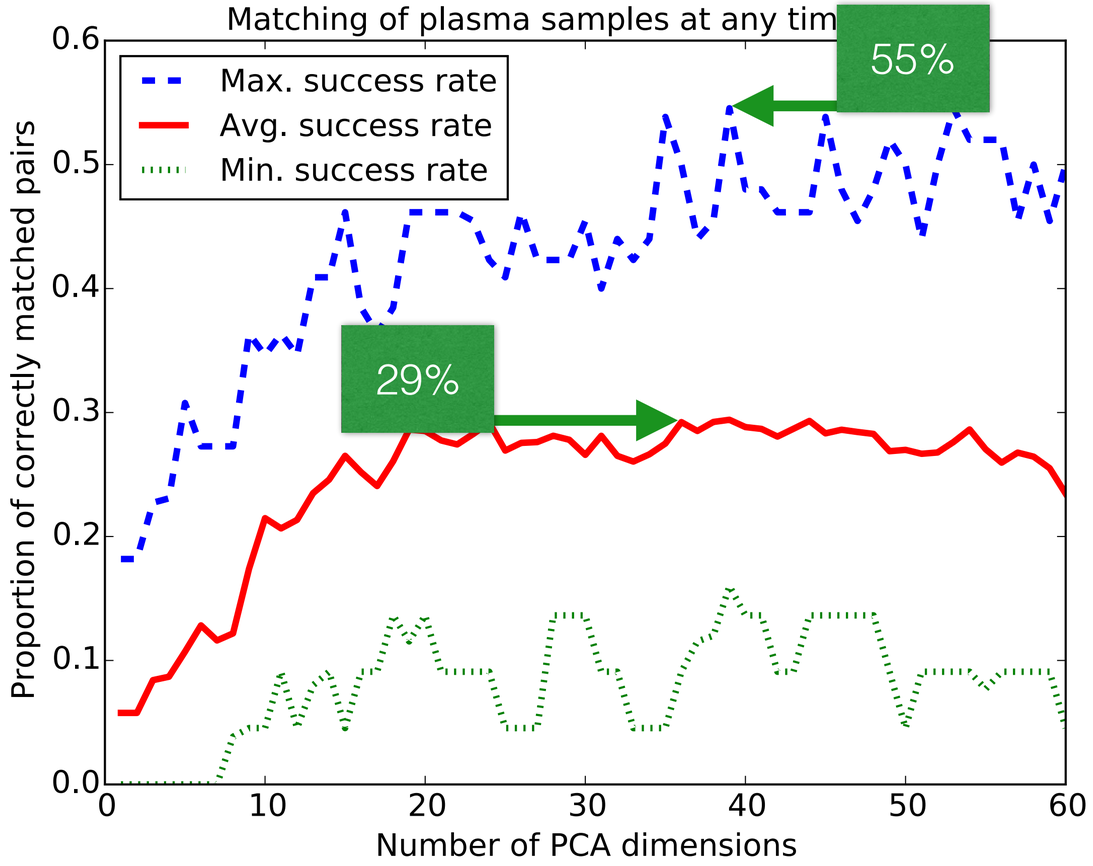
SAARLAND UNIVERSITY
COMPUTER SCIENCE

CISPA
Center for IT-Security, Privacy and Accountability

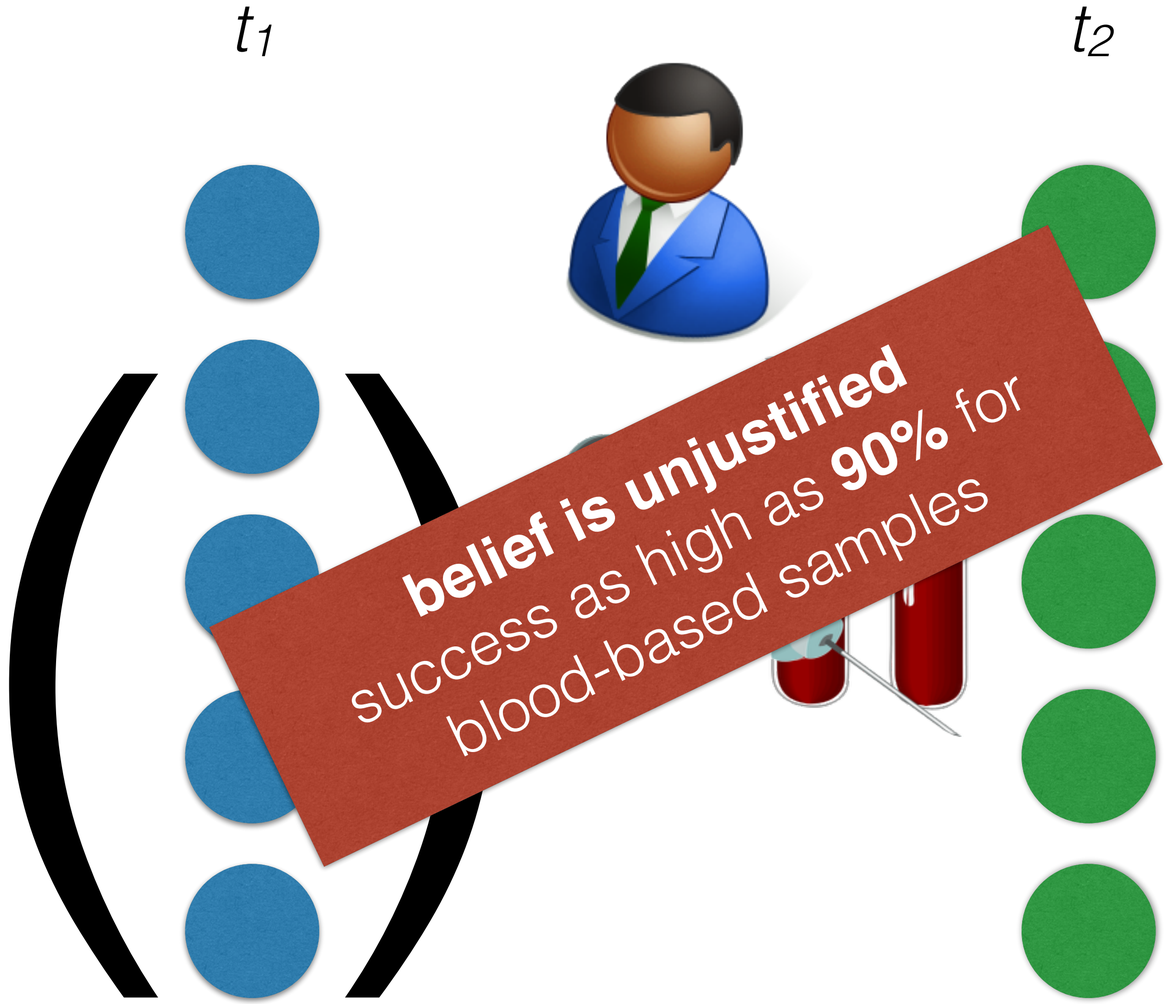
Identification Attack



Matching Attack



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



Thank you!

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

Questions?

there in fact are **privacy threats** inherent to epigenetic data

blood is easier to link than plasma

matching is more successful than identification

success rate remains more or less **constant in the first year**