SoDA

Generating synthetic data

On the basis of metadata

Erik-Jan van Kesteren

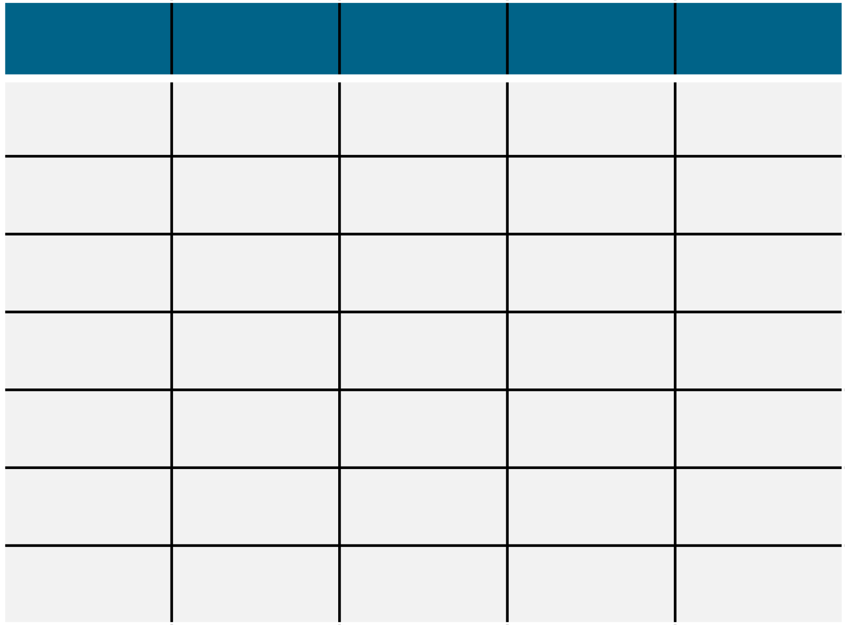
Outline

- What is synthetic data
- Why do we need synthetic data?
- The precision – privacy continuum of synthetic data
- Generating synthetic data from metadata
- Ddi-synth app
- Conclusions
- Questions

Preface: tidy data

- In this talk, I will focus on **tidy data**
- Rectangular data $\boldsymbol{X}$

- Every column is a variable.
- Every row is an observation.
- Every cell is a single value.



https://tidyr.tidyverse.org/articles/tidy-data.html

What is synthetic data?

Synthetic data

Synthetic data / fake data / generated data / simulated data

As opposed to real, natural, collected data

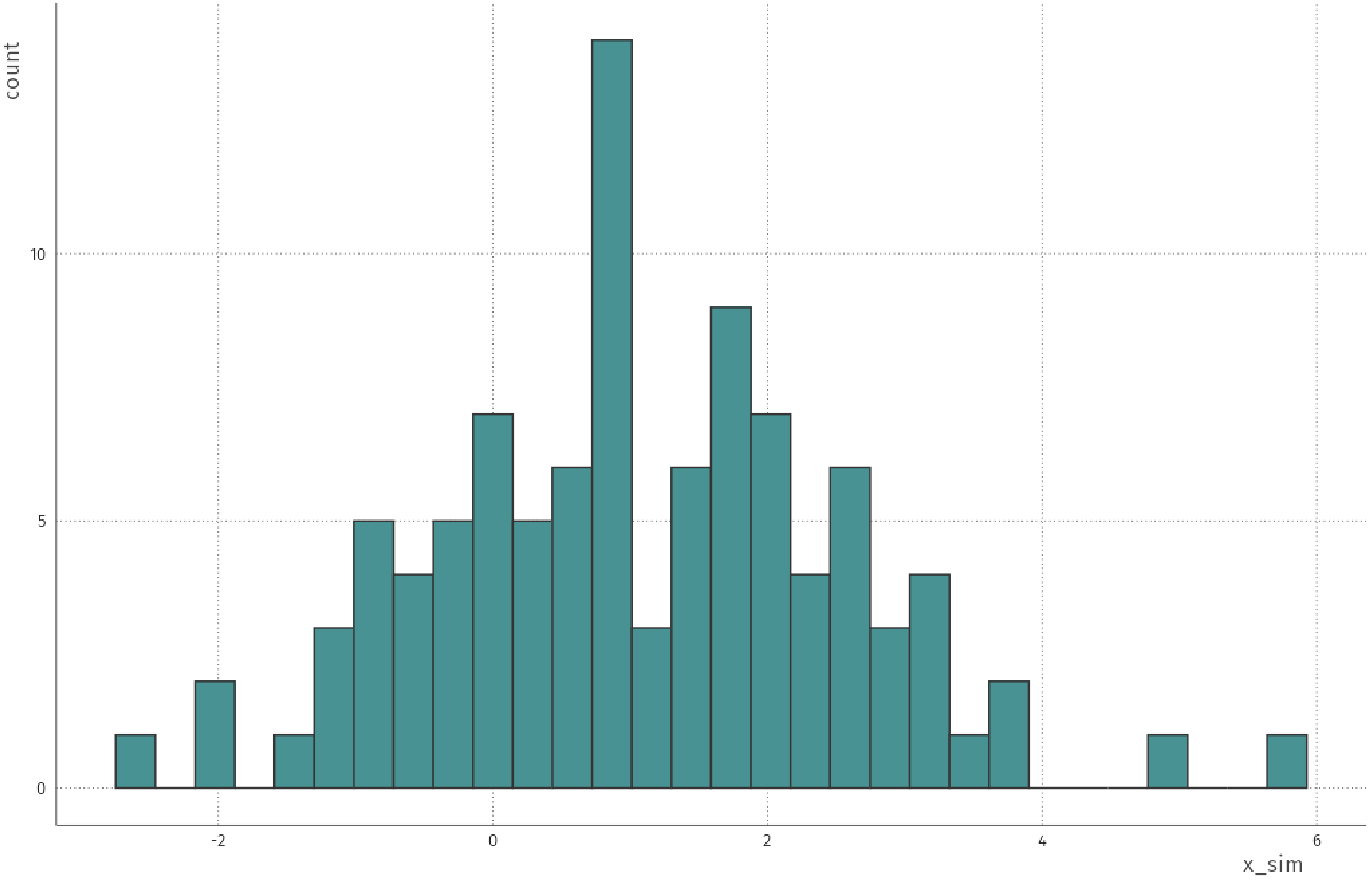
Data generated from some probability distribution

Samples from distribution $p(\boldsymbol{X}|\ \theta)$

In R code:

```
# parameters
mu <- 1.0
sigma <- 1.5

# generate data
x_sim <- rnorm(100, mean = mu, sd = sigma)
```


count
10
5
0
-2
0
2
4
6
x_sim

To generate synthetic data, you need a probability distribution and parameter values

(this is your first take-home message)

Why do we need synthetic data?

Why synthetic data?

Sometimes we cannot (easily) access the real data

CBS microdata (access control, privacy concerns)

Sometimes we have access but we want to be open about our methods

Sharing code which runs on synthetic data?

Sometimes we want to look at the data only once after the analysis is set!

Avoids p-hacking & questionable research practices

Sometimes we really do not want to see the data / have it on a server

Privacy-friendly computing

Some uses of synthetic data

Getting to know the data
Being able to inspect the variables and their values
Knowing the size of the dataset (N, P)

Using it as an toy example
For teaching purposes
To try out analyses / different methods
Statisticians love simulation

To write an analysis script
If your code runs on synthetic data, it will likely also work on the real data

To reproduce analyses and come to the same conclusions
This will be more difficult!

Know your goal. What are you going to do with the synthetic data?

(second take-home message)

Getting to know the data?

You can use data explorer from scholarsportal.info

Synthetic data not *really* necessary

https://scholarsportal.github.io/dataverse-data-explorer-v2/?siteUrl=https://dataverse.scholarsportal.info&fileId=8988

The precision – privacy continuum

Precision vs. privacy

Precision *(slight abuse of terminology for the sake of alliteration)*

When I run my analysis on synthetic data, how close are my

- Parameter estimates
- Statistical models
- Conclusions

To the real thing?

Privacy

When I have the synthetic data generated by $p(\boldsymbol{X}|\theta)$, how well can I

- Reproduce the original data? (model inversion attack)
- Determine whether a person was part of the original data? (differential privacy)

Precision vs. privacy

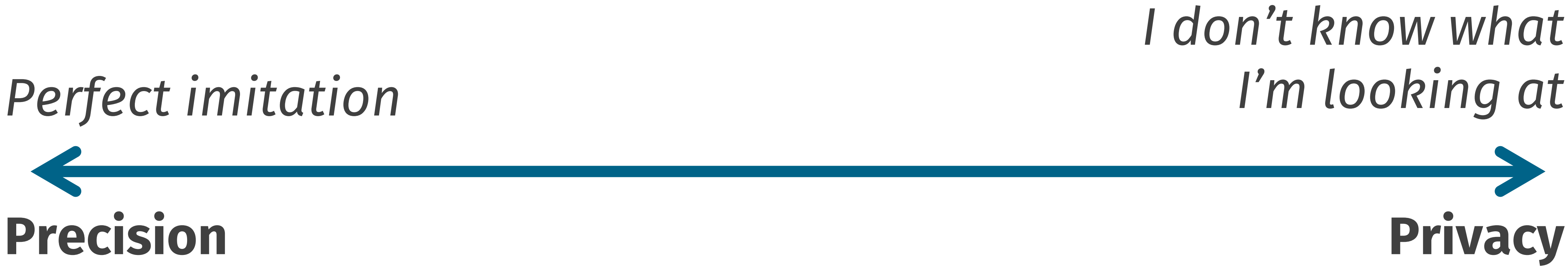
- Every parameter in the data-generating model contains information about the observations in the real data
- The more parameters (information) you use to generate synthetic data, the more precise it will be
- When the information in the parameters equals the information in the real data, we have just recreated the real data
- At that point, there is no more privacy / disclosure control

Precision and **privacy** are opposites

If you increase the precision of synthetic data, you decrease its privacy

(third take-home message)

How much does the synthetic data look like the real data?



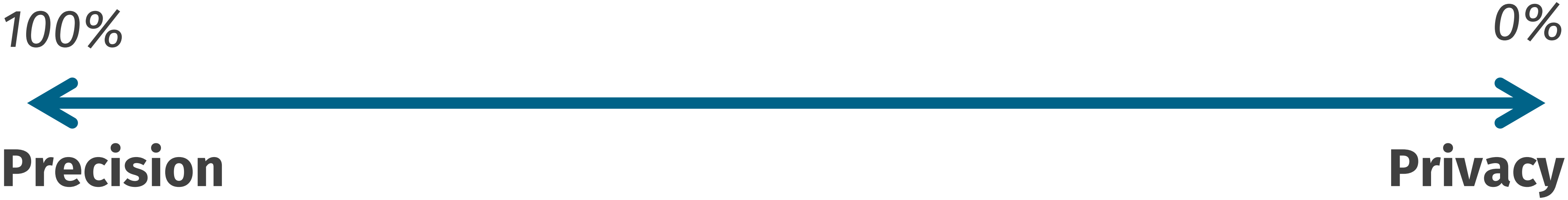
Perfect imitation

I don't know what I'm looking at

Precision

Privacy

How well do analyses performed on the synthetic data reproduce those on the real data?



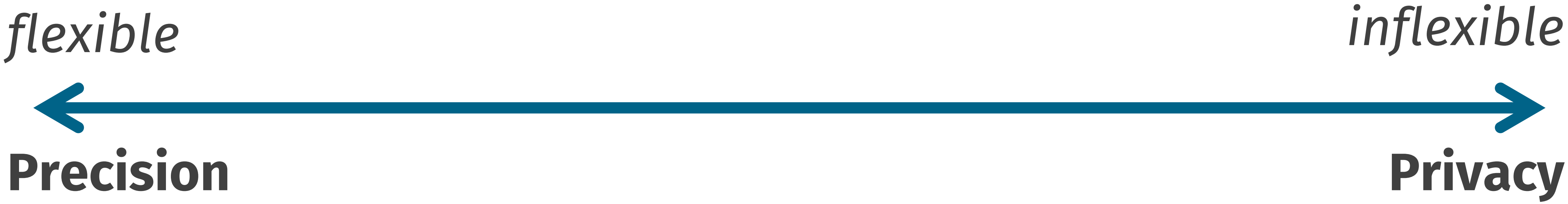
100%

0%

Precision

Privacy

How flexible does my data-generating model $p(\boldsymbol{X}|\theta)$ need to be?



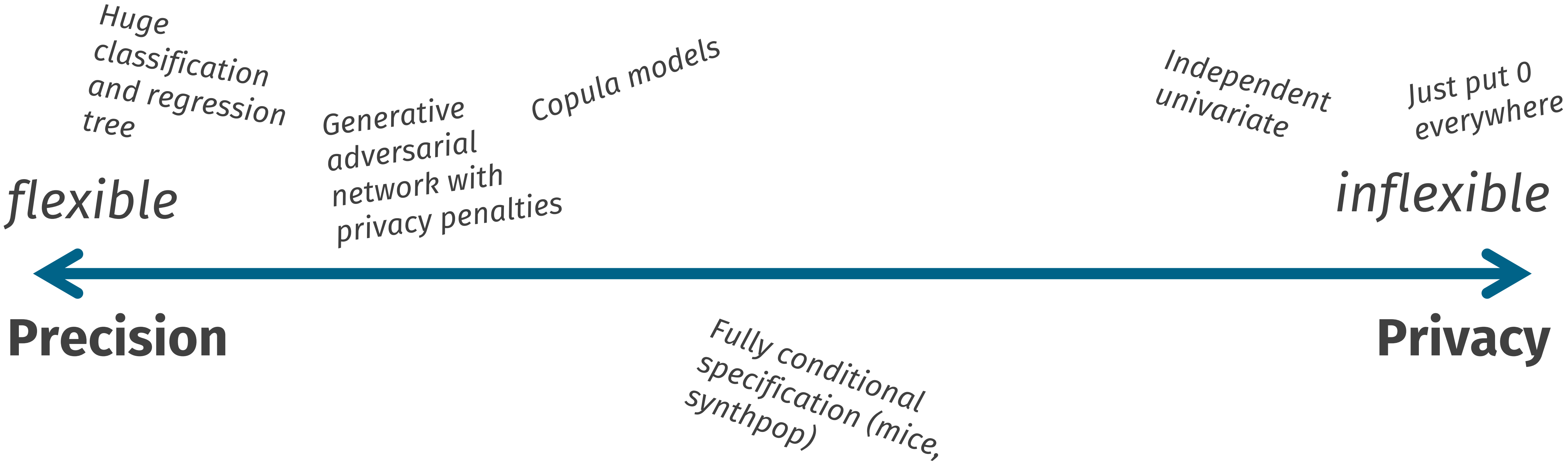
flexible

inflexible

Precision

Privacy

How flexible does my data-generating model $p(\boldsymbol{X}|\theta)$ need to be?



Huge classification and regression tree

Generative adversarial network with privacy penalties

Copula models

Independent univariate

Just put 0 everywhere

flexible

inflexible

Precision

Privacy

Fully conditional specification (mice, synthpop)

Learning $p(X|\theta)$ from data

Synthpop

https://synthpop.org.uk/

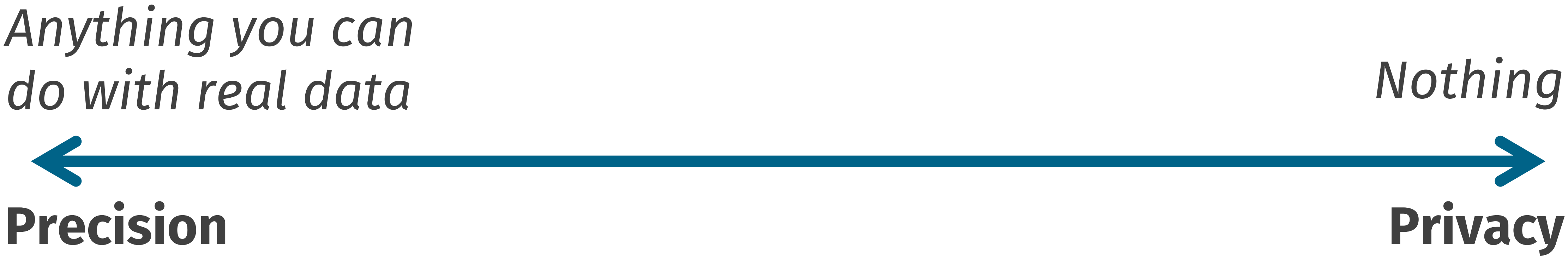
(chained equations, trees)

Synthetic data vault

https://sdv.dev/

(copula, GAN, VAE)

What can we do with the synthetic data?



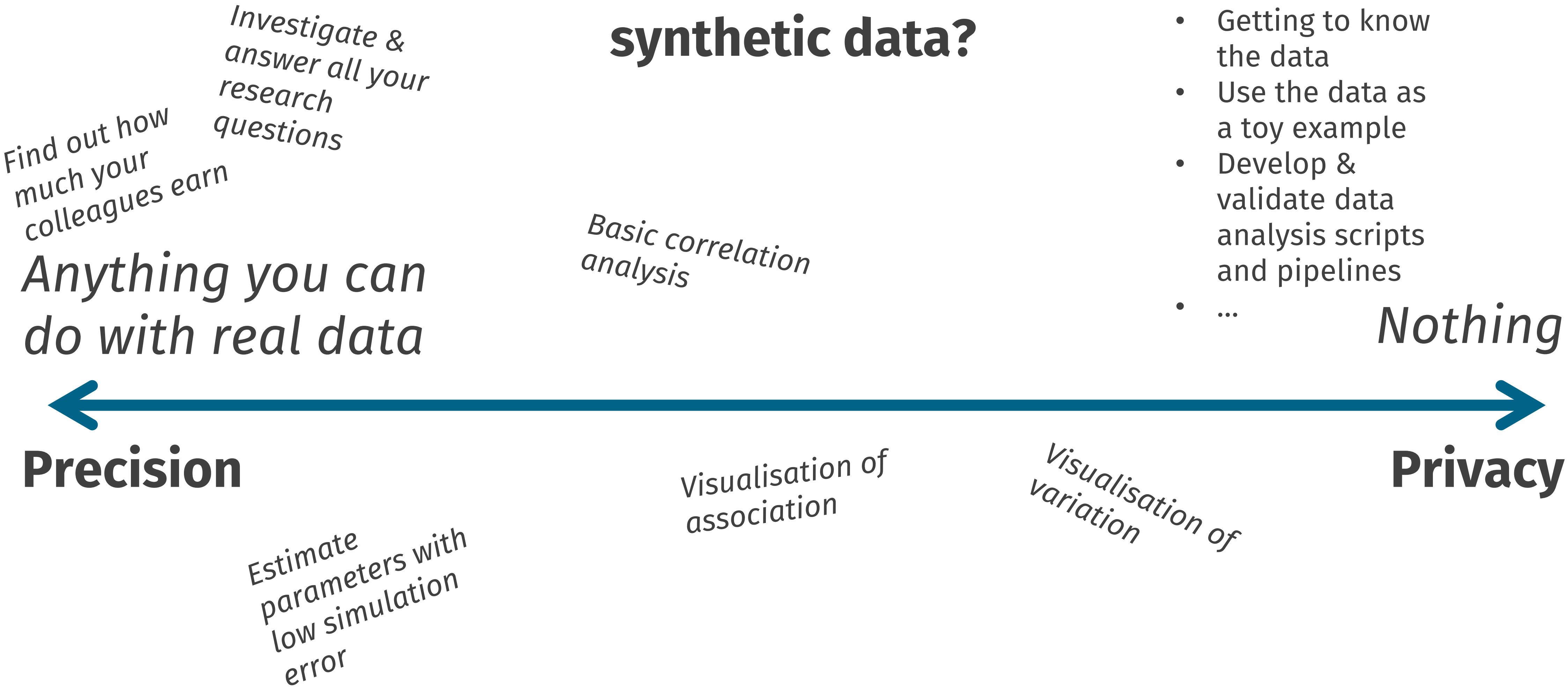
Anything you can do with real data

Nothing

Precision

Privacy

What can we do with the synthetic data?



Anything you can do with real data

Find out how much your colleagues earn

Investigate & answer all your research questions

Estimate parameters with low simulation error

Basic correlation analysis

Visualisation of association

Visualisation of variation

- Getting to know the data
- Use the data as a toy example
- Develop & validate data analysis scripts and pipelines
- …

Nothing

Precision

Privacy

What can we do with the synthetic data?

Find out how much your colleagues earn

Investigate & answer all your research questions

Anything you can do with real data

Basic correlation analysis

- Getting to know the data
- Use the data as a toy example
- Develop & validate data analysis scripts and pipelines
- …

Nothing

Precision

Privacy

Visualisation of association

Visualisation of variation

Estimate parameters with low simulation error

This privacy stuff is complicated

The trick:
generate data from metadata

Generating data from metadata

There is no more privacy concern

The information in the metadata has already been released to public

Metadata is ideally in a machine-readable format

Not just pdfs (!)

Contains variable-level information

Data-generating models using metadata are necessarily simple

Per-variable information, no association: $p(\boldsymbol{X}|\theta) = \prod_{i \in P} p(x_i|\theta)$

Similar to “naïve Bayes” model

Generating data from metadata

1. Get the metadata
2. For each variable:
   1. Determine type of outcome
   2. Determine amount of missingness
   3. Find distribution that fits: normal, truncated normal, Bernoulli, multinomial
   4. Set parameters of this distribution from metadata: mean, sd, min, max, proportion, category probabilities, category labels
   5. Generate data
3. Put it all in a nice table for the user

ddi-synth

Conclusions

Conclusions

- Synthetic data:
  - Is generated from probability distribution with parameters
  - Has different goals (know your goal!)
  - Lies on a precision-privacy continuum

- Metadata is privacy-friendly & contains parameter values
- Automatic generation of data from metadata is viable (easy?)
- ddi-synth as a particular proof-of-concept implementation

Questions?

Thank you!

SoDa