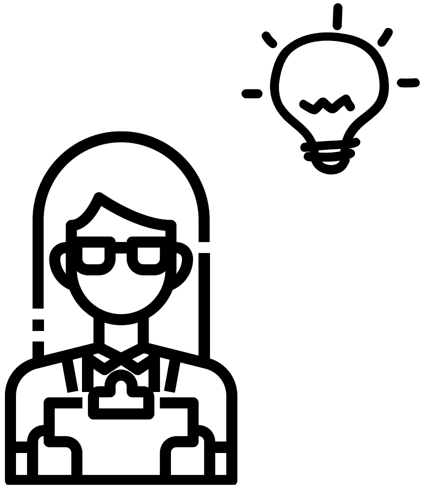
How to create synthetic data

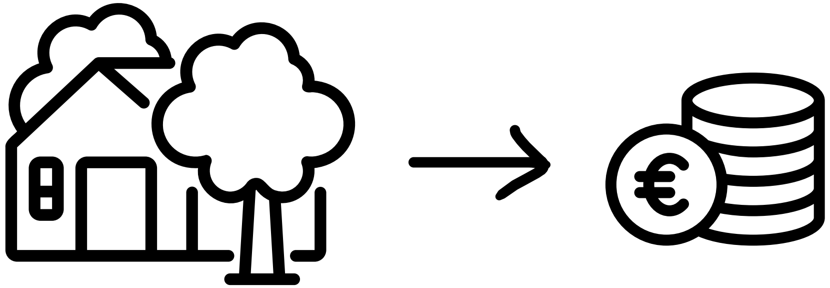
A tool for open science

Erik-Jan van Kesteren
Raoul Schram
Thom Volker

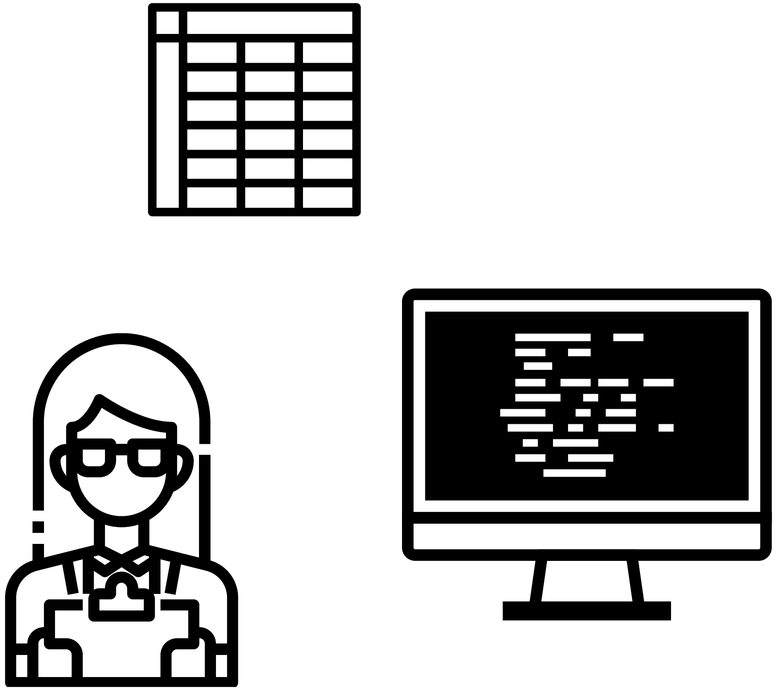
Utrecht University
ODISSEI Social Data Science team

Imagine..





- Where do you live?
- How long have you lived there?
- What do you earn?
- How much do you spend on gifts for your friends?



“More generous gifting behaviour in greener neighbourhoods”

OPEN MATERIALS

```
my_data <- read_csv("super_private_data_file.csv")
```

Open data not allowed, options:

- Data just not available, good luck
- “Data available upon reasonable request”
- Data is part of a large project with data access procedures

I just want to check out the script to learn from the cool analysis!

Solution: publish open synthetic data with your open materials

What will we do this morning?

- A primer on synthetic data
- Creating privacy-friendly synthetic data based on metadata
    - Pair programming in python
- Creating & assessing high-utility synthetic data
    - Pair programming in R
- Closing

A primer on synthetic data

Synthetic data (EJ's definition)

Synthetic data is generated from a model
As opposed to real, natural, collected data

fake data
generated data
simulated data
digital twin
public use file

To create synthetic data, you need a **generative model**

Generative model

$$p(\boldsymbol{X}|\theta)$$

- A model for data $\boldsymbol{X}$
- Has parameters (θ)
- You can fit / estimate / learn θ based on real data

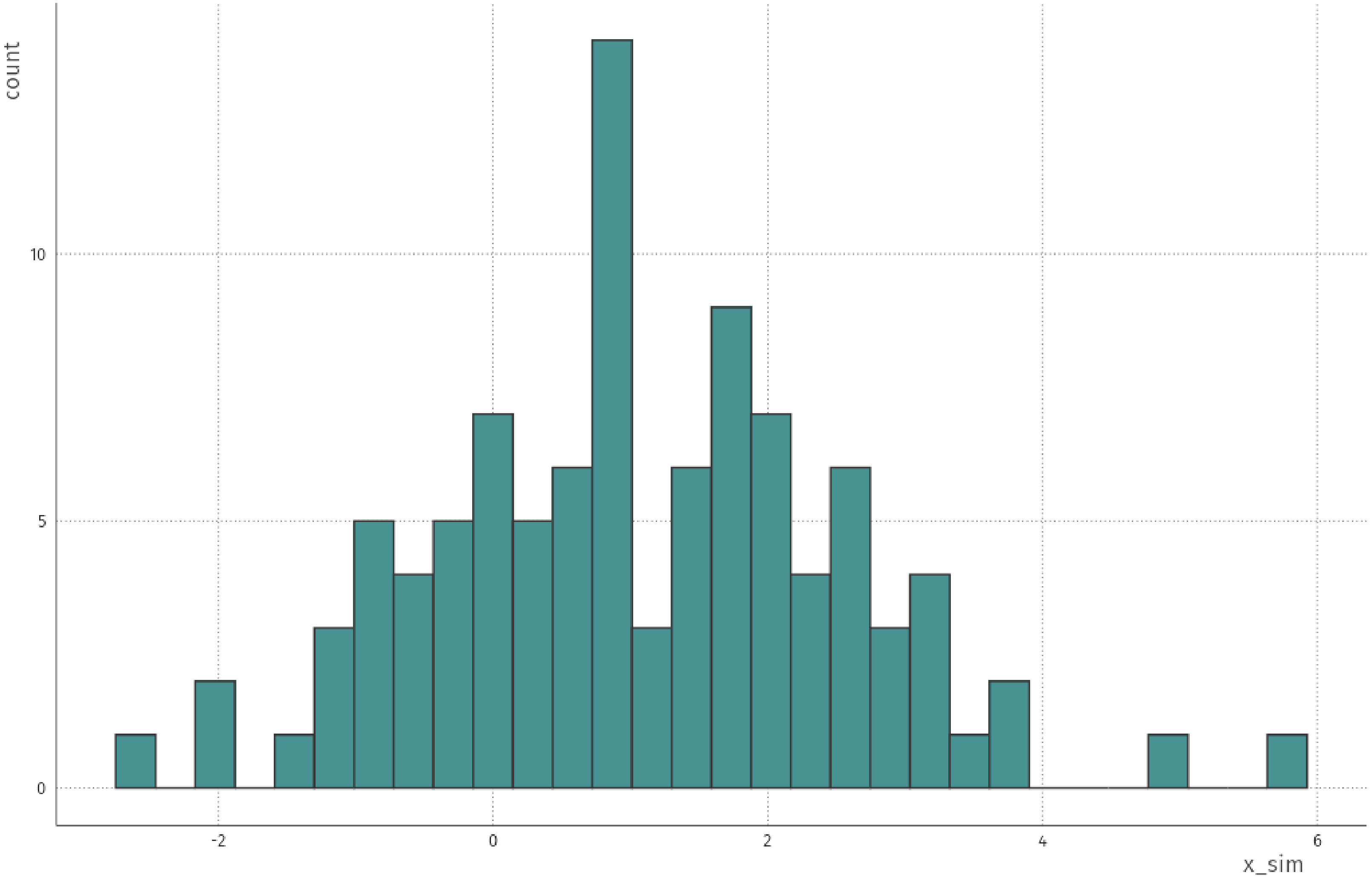
- Examples:
  - A normal distribution with parameters $\theta = \mu, \sigma$
  - A histogram with bins and proportions
  - A generative adversarial network with a million parameters

Generative model

In R code:

```r
# 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

Generative model

- Today we will fit two types of generative models:

- **Metasynth:** automatically selected univariate parametric distributions for each variable in your data
- **Synthpop:** Fully conditional nonparametric classification and regression trees to model the whole dataset

- There are infinitely many more generative models. This is an active field of research

How to make sense of all of these models for creating synthetic data in the real world?

The privacy-utility tradeoff

Utility vs. privacy

Utility

- How close is my synthetic data to my real data? Can I distinguish synthetic and real samples?
- Thom will tell you more about this

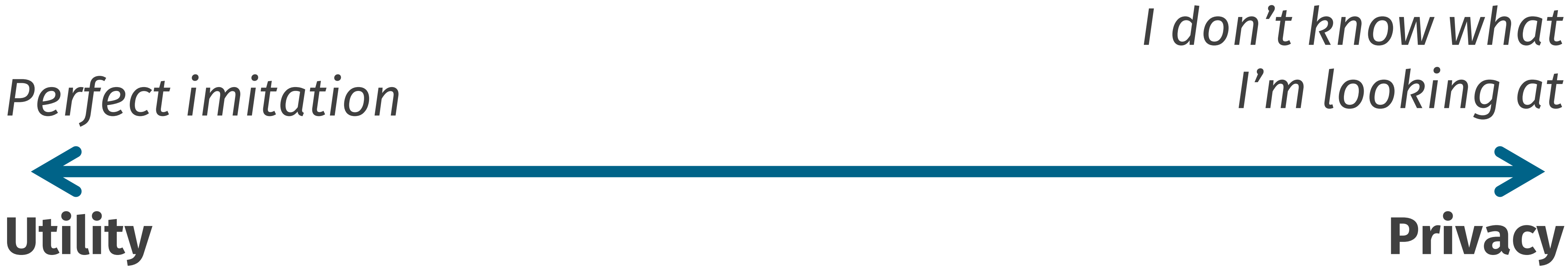
Privacy

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)
- Estimate a specific person’s income within certain bounds?
- …

Utility and **privacy** are opposites

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



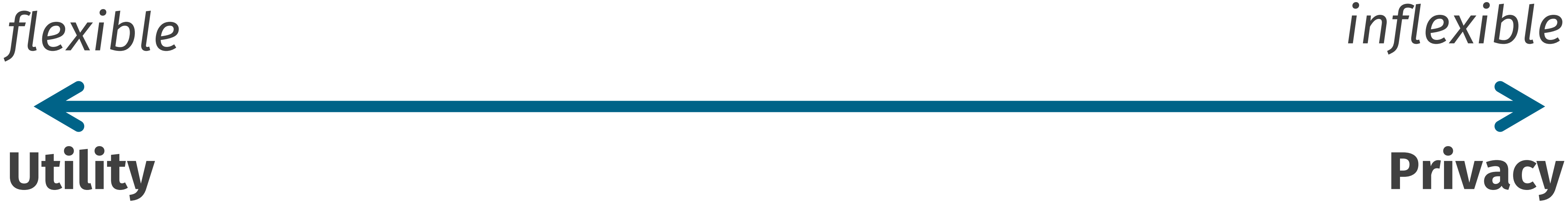
Perfect imitation

Utility

I don’t know what I’m looking at

Privacy

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



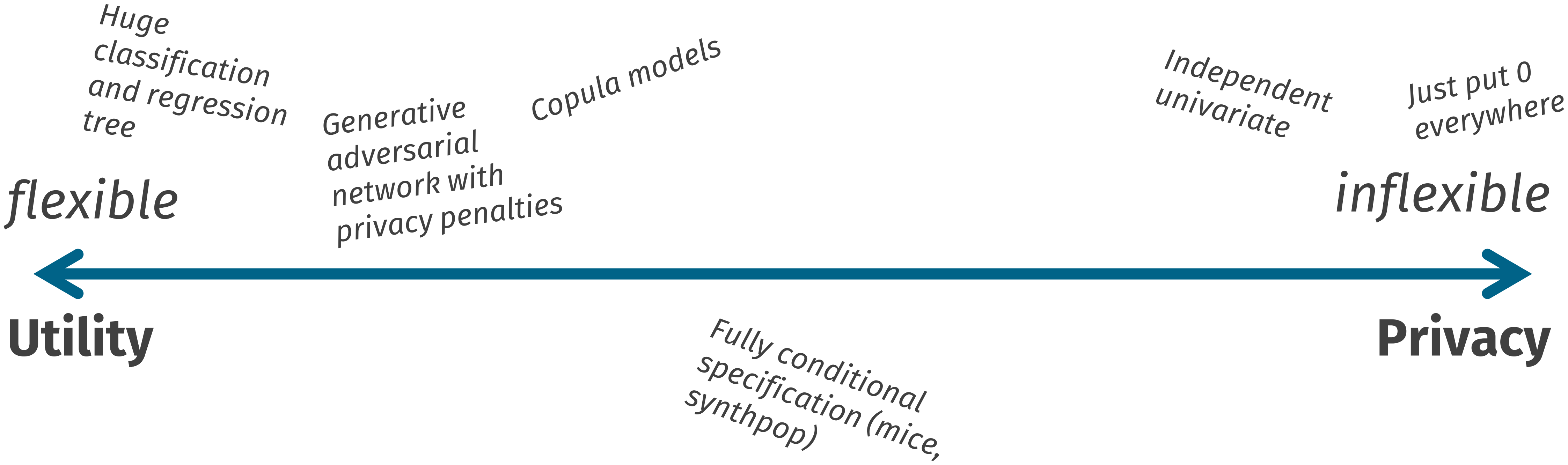
flexible

inflexible

Utility

Privacy

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



Huge classification and regression tree

Generative adversarial network with privacy penalties

Copula models

Independent univariate

Just put 0 everywhere

flexible

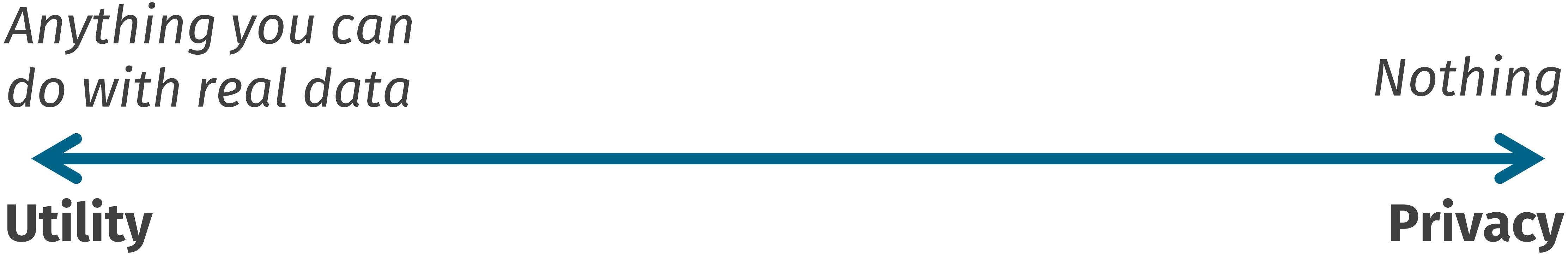
inflexible

Utility

Privacy

Fully conditional specification (mice, synthpop)

What can we do with the synthetic data?



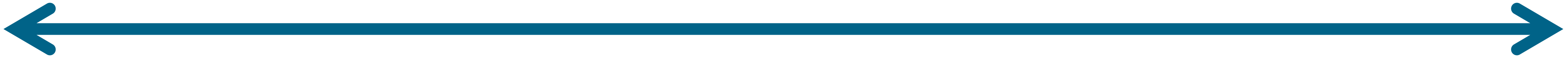
Anything you can do with real data

Nothing

Utility

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
- …

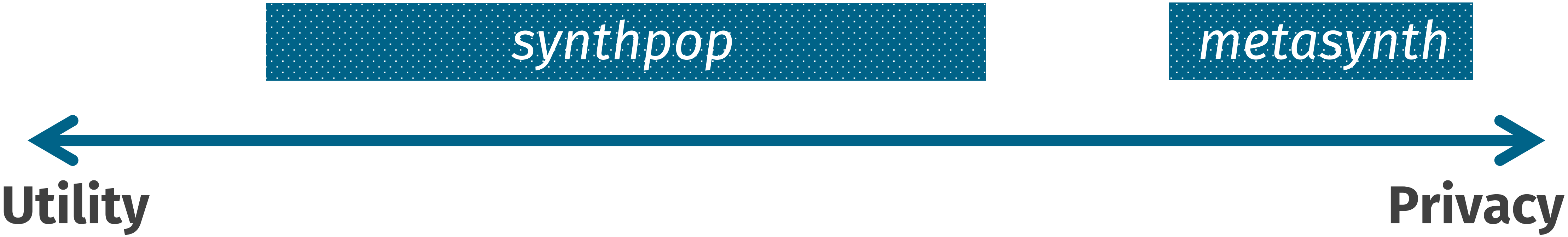
Nothing

Utility

Privacy

Different methods for different use-cases

Let's get started!



synthpop

metasynth

Utility

Privacy

Icons from the noun project

Scientist by Justicon
Idea by Icon
house tree by LUTFI GANI AL ACHMAD
Euro by Larea
people by Alice Design
Table by Alex Burte
Hacking by Alfredo
Paper by Egi Maulana
Scientist 2 by Justicon
Question by Anggara Putra

https://thenounproject.com/