ORGANISED BY
BDV BIG DATA VALUE ASSOCIATION

data-week.eu
#DataWeek24

DATAWEEK 24

JOIN.LEARN.SHARE.GET VALUE

Data Value UNLOCKED!

10 December/ Luxembourg

IN COLLABORATION WITH

LNDS
LUXEMBOURG NATIONAL DATA SERVICE

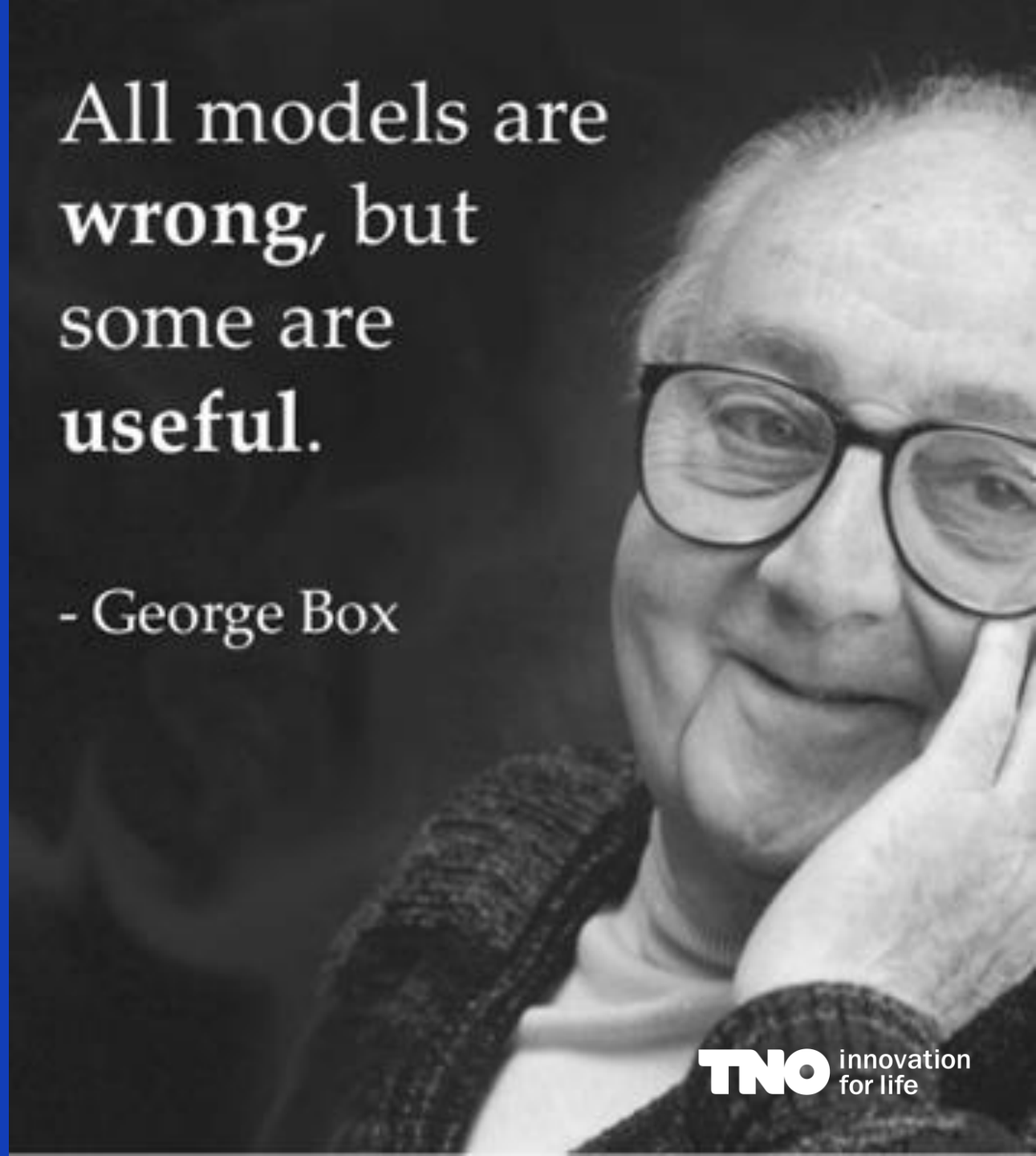
LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY | LIST

Netcompany

Synthetic data

Data week 2024, Luxembourg

Freek Bomhof | 11 December 2024



https://thisxdoesnotexist.com

Early examples

This Person Does Not Exist

The site that started it all, with the name that says it all. Created using a style-based generative adversarial network (StyleGAN), this website had the tech community buzzing with excitement and intrigue and inspired many more sites.

Created by Phillip Wang.

This Cat Does Not Exist

These purr-fect GAN-made cats will freshen your feeline-gs and make you wish you could reach through your screen and cuddle them. Once in a while the cats have visual deformities due to imperfections in the model – beware, they can cause nightmares.

Created by Ryan Hoover.

This Rental Does Not Exist

Why bother trying to look for the perfect home when you can create one instead? Just find a listing you like, buy some land, build it, and then enjoy the rest of your life.

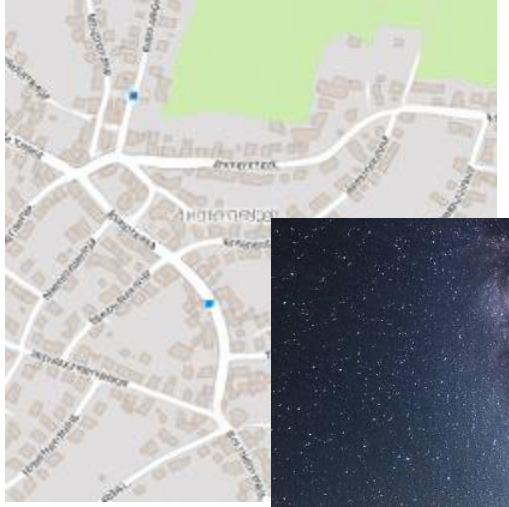
Created by Christopher Schmidt.

Synthetic data

https://thisxdoesnotexist.com

Many others

- Waifu
- Question
- Resume
- Emotion
- Vessel
- Lyrics
- Snack
- Meme
- Chair
- Foot
- Artwork
- Chemical
- Horse
- Word
- MP
- Pony
- Automobile
- Eye
- Butterfly
- City
- Food blog
- Fu***ed-up Homer
- Fursona
- Night sky
- Music video
- Beach
- Map
- Idea
- Sneaker
- Baseball player
- Flag
- Interview
- Iris
- Voice
- Pepe
- House
- Campsite
- Mountain
- Music
- Street view
- Graffiti
- Airport
- Play



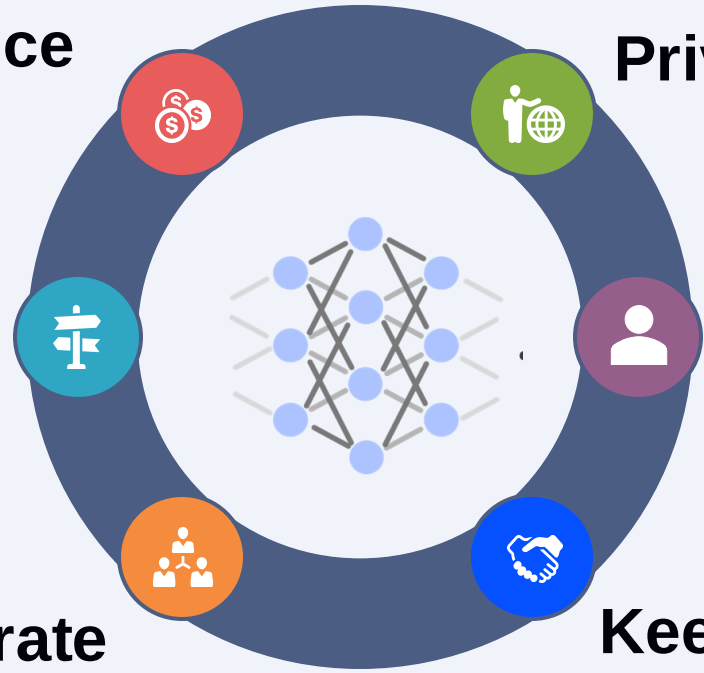
noun.

kristig

1. a synthetic black impurity, especially of iron, occurring in all samples from sheep cattle and sheep feed
   "he received a high three o'clock prickly-kristig"
2. a word that does not exist; it was invented, defined and used by a machine learning algorithm.

Link / New word / Write your own

Reasons to use synthetic data



Compliance

Privacy protection

Testsets

Bias mitigation

Collaborate

Keep data at the source

Applications of synthetic data

Testing

- Realistic data
- Create your own edge cases

Data sharing

- Training of AI models

Transparency

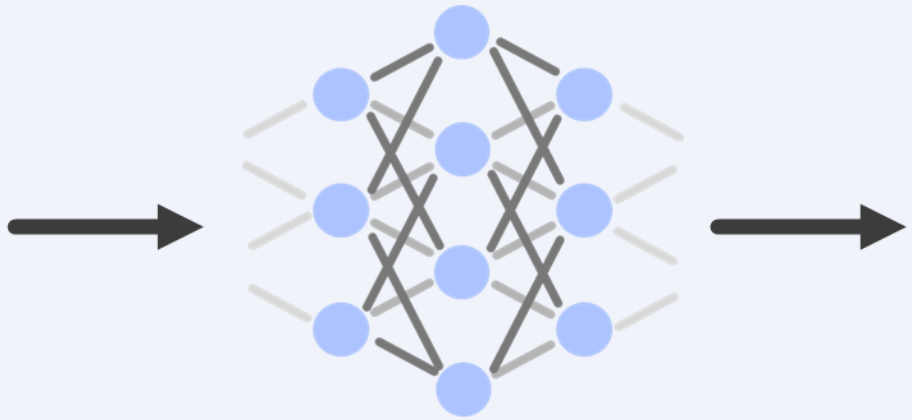
- Explainability
- Publishing results

Training data

- Augmentation
- Fairness

What is synthetic data?

ID	Age	Gender	Income
1	35	F	4300
2	27	M	5700
3	45	M	4300
...	...	...	...



ID	Age	Gender	Income
1	31	F	6600
2	37	F	4300
3	25	M	3100
...	...	...	...

- Model learns characteristics from original dataset
- Statistical (GAN, VAE, Diffusion model) or rule-based method
- Produces new samples with the same statistical properties
- No original data points are retained

Types of data

Energy usage
Time-series

HR data
Relational data

Age	Income	Dpt.
41	4200	1
55	4600	1
30	3500	3
…	…	…

Dpt.	Size
2	410
5	60
6	58
…	…

Types of data

Seized data
Text data

```
Dark Knight: you wanna pay with Monero?

Buyer2143: only BTC

Dark Knight: not gonna happen, bro

Buyer2143: other options?

Dark Knight: ....
```

Camera footage
Image data

GDPR still applies!

GDPR

- You still need a legal ground to process the data
- Synthetic data is a way to comply with data minimization requirements
- Synthetic data is not automatically anonymous data

AI Act

- Synthetic data is the preferred method to test high risk AI systems for bias (art 10.5a)

Utility

- Do the synthetic data lead to the same **conclusions** as the original data?
- Do the synthetic data mimic **patterns** that we see in the original data?

Privacy

- Can we **infer sensitive information** from the synthetic data?
- Can we make privacy guarantees about synthetic data generator?
- Can we **measure** privacy afterwards?

Privacy-utility trade-off

No traces of the original data
No utility
Perfect privacy

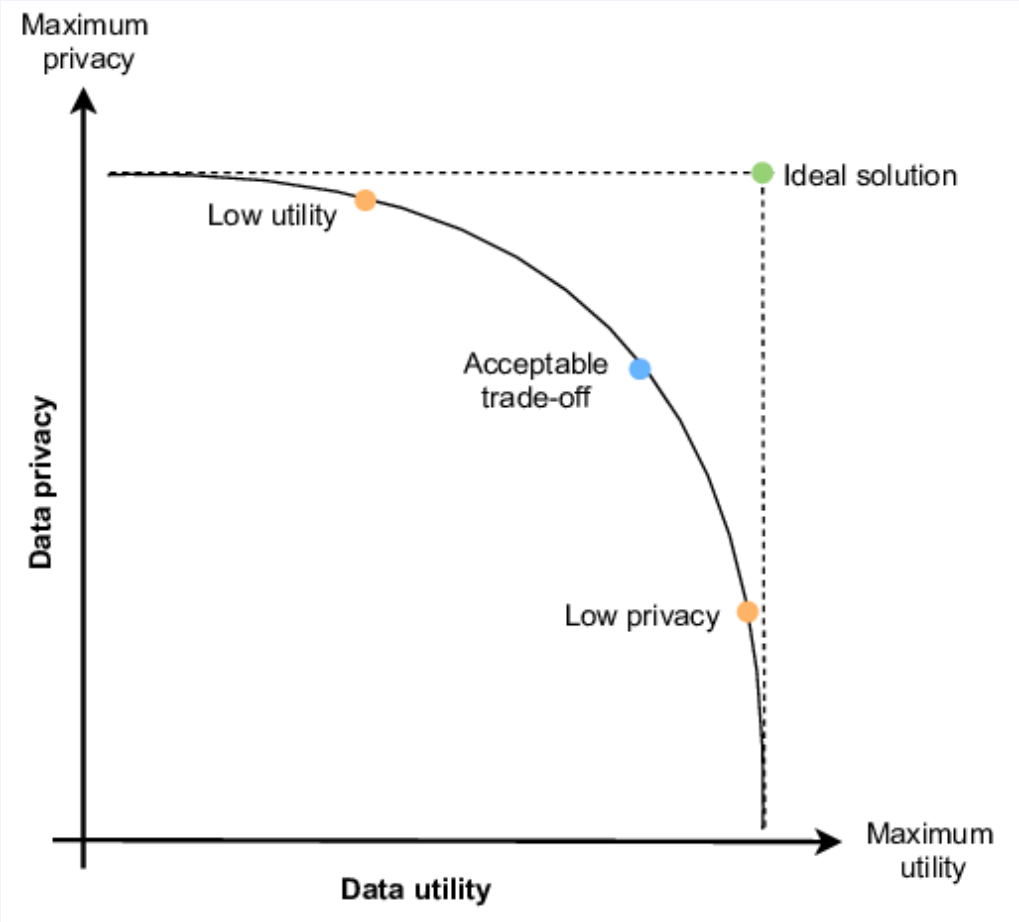
Original data
Perfect utility
No privacy

Utility vs Privacy trade-off

There is no perfect solution.

However, when privacy is a concern:

Using synthetic data is already much better than using the original sensitive data



Trade-off between privacy level and utility level of data. | Download Scientific Diagram (researchgate.net)

Ongoing work

1. Standardization of privacy and utility metrics
2. Business rules in complex relational data
3. Specific forms of textual data
4. Federated multi-tabular data

TNO innovation for life

Leaking information from synthetic data

- Despite containing only artificial records, synthetic data can leak information.
- Due to overfitting, certain individuals could be copied or closely resembled.

Original data

Age	Gender	Income
23	F	2200
62	F	12600
42	M	4500
…	…	…

Synthetic data

Age	Gender	Income
41	F	4200
62	F	12600
30	M	3500
…	…	…

Leaking information from synthetic data

Given certain **background information** can we uncover new, **sensitive information** about one or more individuals?

- Train a machine learning model to **predict** income from age and gender.
- If the model provides accurate predictions, this can indicate **low privacy**.
- Use a **holdout** set to interpret the level of accuracy.

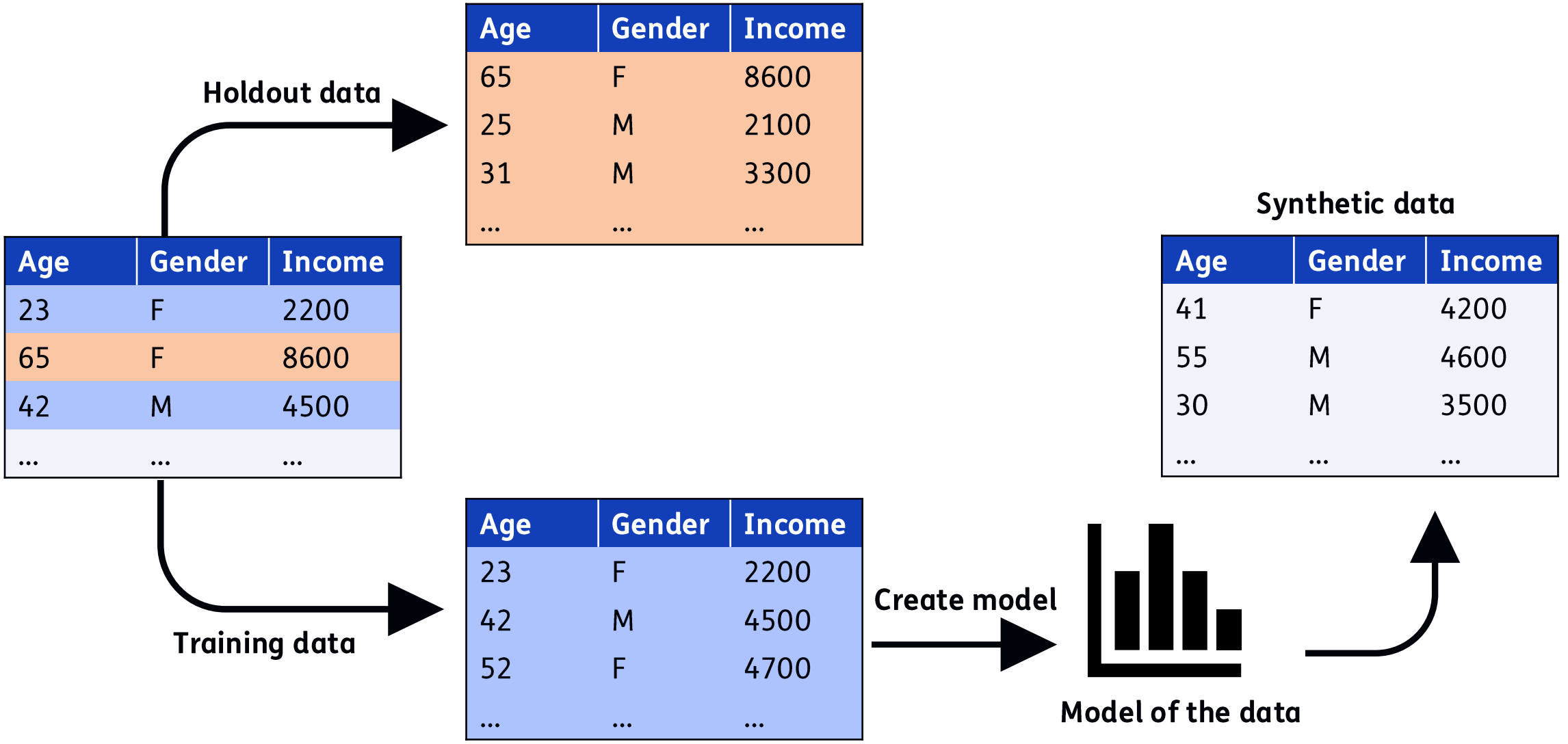
Original data

Age	Gender	Income
23	F	2200
62	F	12600
42	M	4500
…	…	…

Synthetic data

Age	Gender	Income
41	F	4200
62	F	12600
30	M	3500
…	…	…

Holdout set


Holdout data
Age Gender Income
65 F 8600
25 M 2100
31 M 3300
... ...
Age Gender Income
23 F 2200
65 F 8600
42 M 4500
... ...
Training data
Age Gender Income
23 F 2200
42 M 4500
52 F 4700
... ...
Create model
Model of the data
Synthetic data
Age Gender Income
41 F 4200
55 M 4600
30 M 3500
... ...

Holdout set

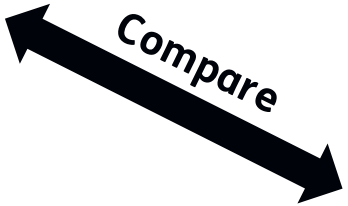
The synthetic data should be **no more similar** to the **training** data than to the **holdout** data.

People in the training data have the **same amount** of **risk** as people in the holdout data.

Holdout data set is needed to put **similarity** between synthetic and training data **in perspective**.

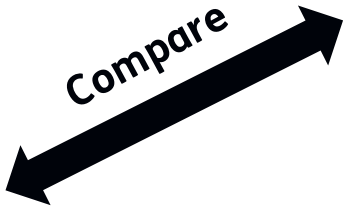
Holdout data

Age	Gender	Income
65	F	8600
25	M	2100
31	M	3300
…	…	…



Synthetic data

Age	Gender	Income
41	F	4200
55	M	4600
30	M	3500
…	…	…



Training data

Age	Gender	Income
23	F	2200
42	M	4500
52	F	4700
…	…	…