SYNTHO

PROVIDING TRUST IN DATA-DRIVEN INNOVATION

Google Is Fined $57 Million Under GDPR

The New York Times

Facebook Faces Massive Damages i

Privacy hinders data-driven innovation

nalytica Suit

Bloomberg

94% of the Dutch population has privacy concerns

AUTORITEIT PERSOONSGEGEVENS

Data innovation

Privacy protection

SYNTHO

Classic ‘anonymization’ is **no solution**

Pseudonymization

Wiping / data deletion

Generalization

Row and column shuffling

Classic ‘anonymization’ is **no solution**

Original data					
Name	**Age**	**Gender**	**Item**	**Price**	**Data**
Olivia	26	Female	Shoes	€125	4 March
John	75	Male	Laptop	€695	5 March
George	41	Male	Beer	€4	7 March
…	…	…	…	…	…
George	41	Male	Shirt	€25	9 March

Classic ‘anonymization’ is **no solution**

Example 1: removing attributes

Original data					
Name	**Age**	**Gender**	**Item**	**Price**	**Data**
Olivia	26	Female	Shoes	€125	4 March
John	75	Male	Laptop	€695	5 March
George	41	Male	Beer	€4	7 March
…	…	…	…	…	…
George	41	Male	Shirt	€25	9 March

Original data with applied classic anonymization					
Name	**Age**	**Gender**	**Item**	**Price**	**Data**
xxx	26	Female	Shoes	€125	4 March
xxx	75	Male	Laptop	€695	5 March
xxx	41	Male	Beer	€4	7 March
…	…	…	…	…	…
xxx	41	Male	Shirt	€25	9 March

Classic ‘anonymization’ is **no solution**

Example 2: generalization

Original data					
Name	**Age**	**Gender**	**Item**	**Price**	**Data**
Olivia	26	Female	Shoes	€125	4 March
John	75	Male	Laptop	€695	5 March
George	41	Male	Beer	€4	7 March
…	…	…	…	…	…
George	41	Male	Shirt	€25	9 March

Original data with applied classic anonymization					
Name	**Age**	**Gender**	**Item**	**Price**	**Data**
xxx	**25-30**	Female	Shoes	€125	4 March
xxx	**70-75**	Male	Laptop	€695	5 March
xxx	**40-45**	Male	Beer	€4	7 March
…	**…**	…	…	…	…
xxx	**40-45**	Male	Shirt	€25	9 March

Classic ‘anonymization’ is **no solution**

And one continues to destroy data…

Original data					
Name	**Age**	**Gender**	**Item**	**Price**	**Data**
Olivia	26	Female	Shoes	€125	4 March
John	75	Male	Laptop	€695	5 March
George	41	Male	Beer	€4	7 March
…	…	…	…	…	…
George	41	Male	Shirt	€25	9 March

Original data with applied classic anonymization					
Name	**Age**	**Gender**	**Item**	**Price**	**Data**
xxx	25-30	Female	Shoes	€100 - €200	March
xxx	70-75	Male	Laptop	€600 - €700	March
xxx	40-45	Male	Beer	<€5	March
…	…	…	…	…	…
xxx	40-45	Male	Shirt	€20 - €30	March

Classic ‘anonymization’ is **no solution**

Original data

Name	Age	Gender	Item	Price	Data
Olivia	26	Female	Shoes	€125	4 March
John	75	Male	Laptop	€695	5 March
George	41	Male	Beer	€4	7 March
…	…	…	…	…	…
George	41	Male	Shirt	€25	9 March

Original data with applied classic anonymization

Name	Age	Gender	Item	Price	Data
xxx	25-30	Female	Shoes	€100 - €200	March
xxx	70-75	Male	Laptop	€600 - €700	March
xxx	40-45	Male	Beer	<€5	March
…	…	…	…	…	…
xxx	40-45	Male	Shirt	€20 - €30	March

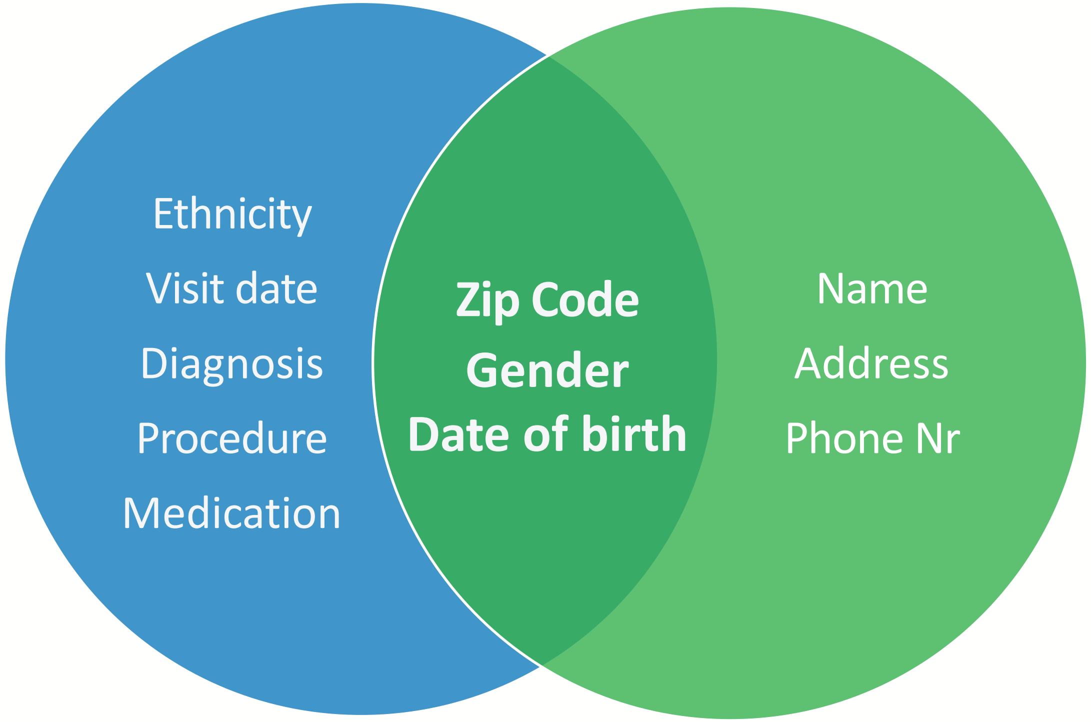
Data value destroyed

Always a privacy risk due to 1:1 relationship

How big is the risk of a linkage attack?

A ‘linkage attack’ in practice

Hospital visits **Voter list**



Quasi-identifiers	% uniquely identified of US population
▪ **5-digit ZIP** ▪ Gender ▪ Date of birth	87%
▪ **Place** ▪ Gender ▪ Date of birth	53%
▪ **Country** ▪ Gender ▪ Date of birth	18%

* L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002: 557-570
** P. Samarati. Protecting Respondents' Identities in Microdata Release. IEEE Transactions on Knowledge and Data Engineering, 13 (6), 2001: 1010-1027

Anonymization does not result in anonymous data

Which of these images is fake?

A **B** **C**

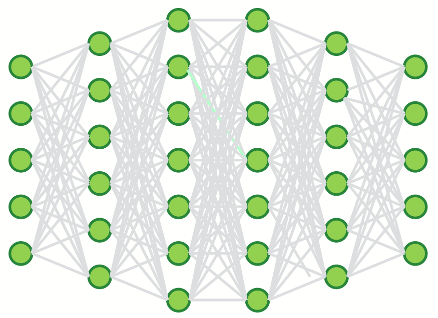
AI and privacy as allies instead of rivals

Syntho believes AI is the way forward in solving the data privacy dilemma

AI-generated synthetic data

Name	Age	Gender	Item	Price	Date	Time
Mary	26	Female	Shoes	€125	17-1-20	11:31
Olivia	65	Female	Laptop	€695	17-1-20	20:47
...						
George	41	Male	Shirt	€25	29-3-20	18:04
George	41	Male	Beer	€4	29-3-20	18:41

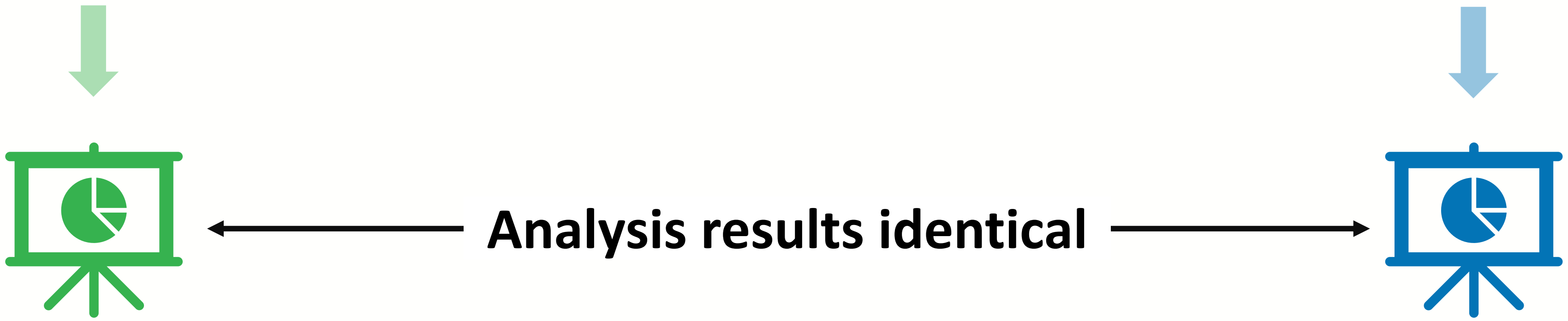
Original data



SYNTHO ENGINE

Name	Age	Gender	Item	Price	Date	Time
Emily	23	Female	Sofa	€790	29-3-20	09:12
Emily	23	Female	Scarf	€40	29-3-20	10:24
...						
Carlos	52	Male	Razor	€5	17-1-20	15:53
Sophie	35	Female	Wine	€7	17-1-20	17:28

Synthetic data



How is **Synthetic Data** different?

Original data

Name	Age	Gender	Item	Price	Data
Olivia	26	Female	Shoes	€125	4 March
John	75	Male	Laptop	€695	5 March
George	41	Male	Beer	€4	7 March
…	…	…	…	…	…
George	41	Male	Shirt	€25	9 March

Synthetic data

Name	Age	Gender	Item	Price	Data
Emily	23	Female	Sofa	€790	1 March
Emily	23	Female	Scarf	€40	3 March
Carlos	52	Male	Razor	€5	7 March
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
Sophie	35	Female	Wine	€7	9 March

No 1:1 mapping with the original data

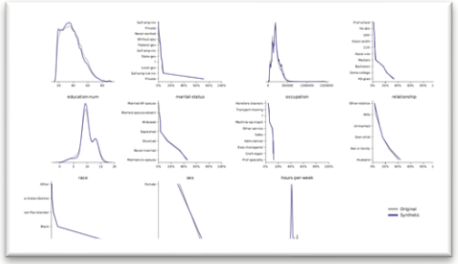
Preserved data quality

Synthetic data quality

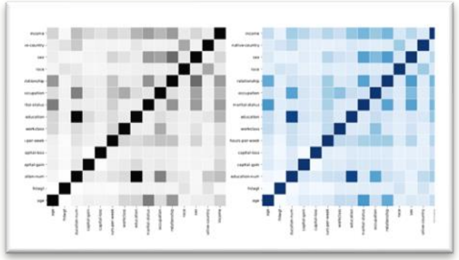
For demonstrating the quality of the synthetic data, we provide a detailed quality report and offer joint evaluation

Statistical quality report

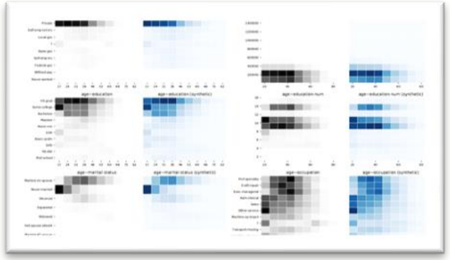
Univariate distributions



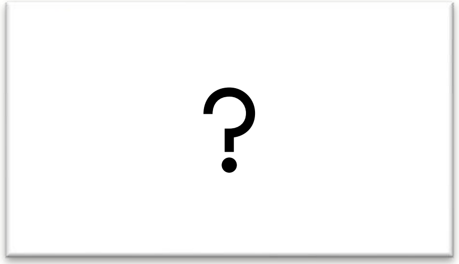
Correlations



Multivariate distributions



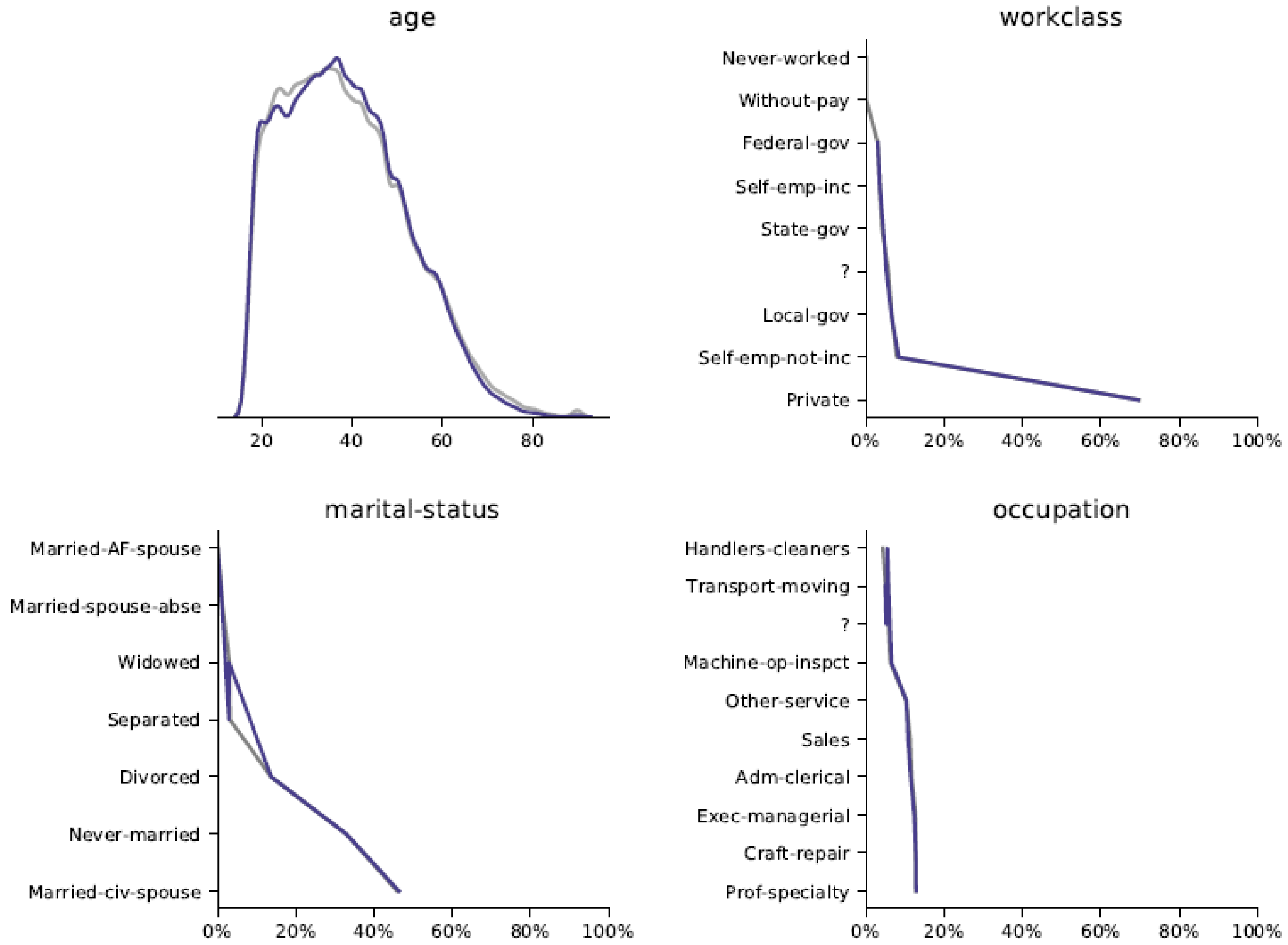
Additional measures upon request



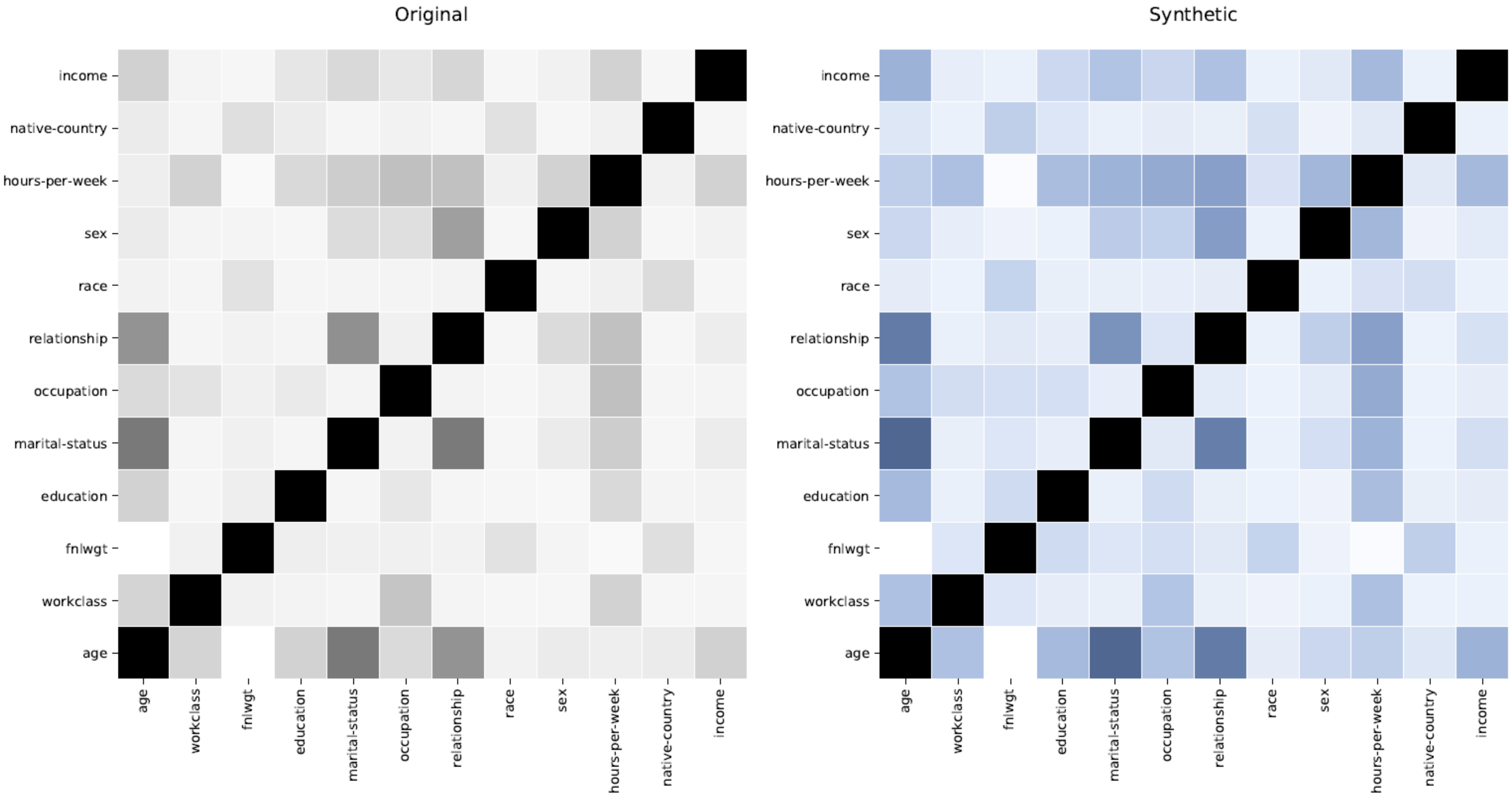
Joint evaluation

- By definition, data utility (or 'usability') can only be understood in relation to the target domain, where the data will be used, shared and / or stored.
- This is why we propose to evaluate the synthetic data with a domain expert in order **to demonstrate the synthetic data 'makes sense'**.

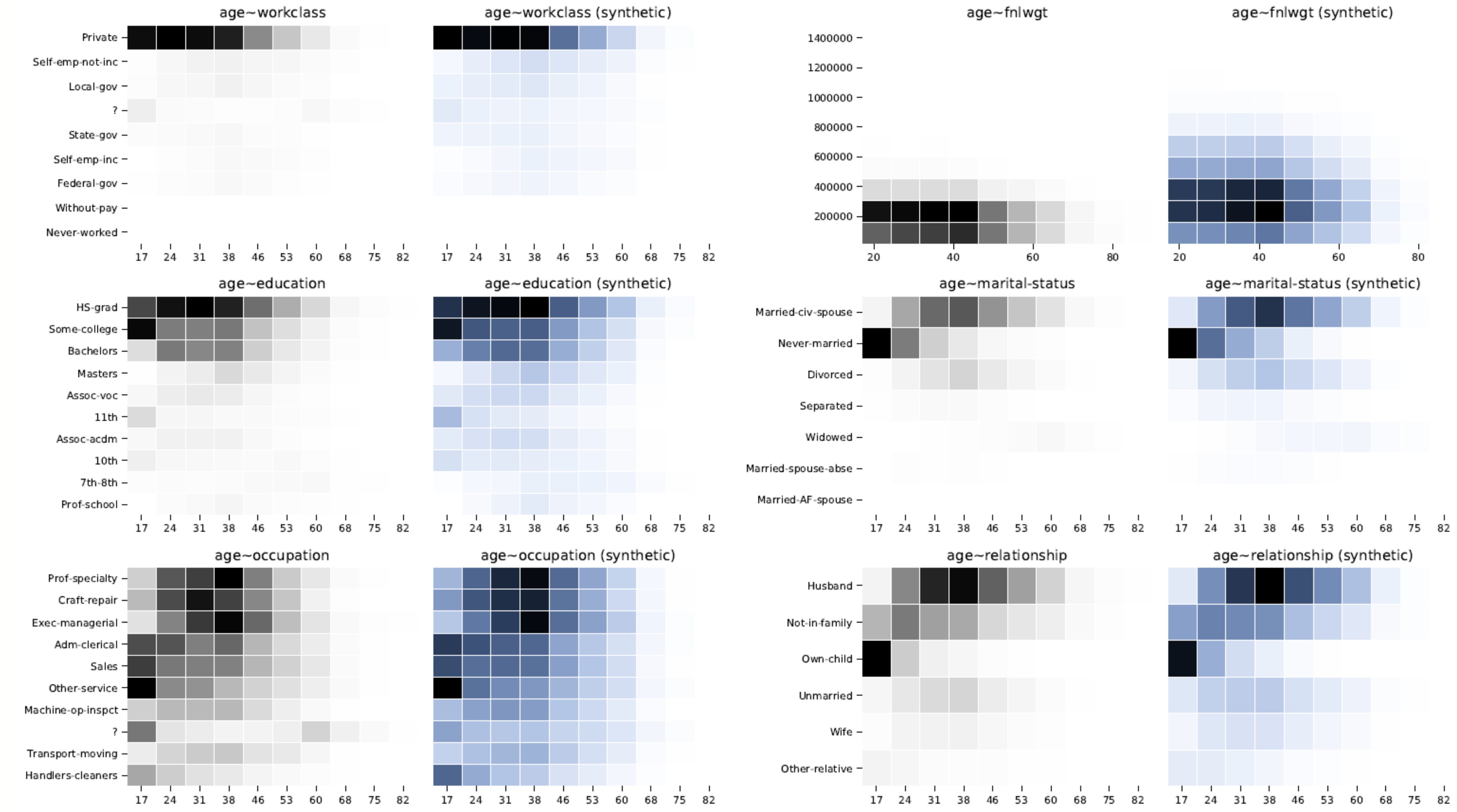
Synthetic data quality



Synthetic data quality



Synthetic data quality



Synthetic Data - an unlimited amount of records can be generated

N=100k

N = 800k?

Original data

Name	Age	Gender	Item	Price	Data
Olivia	26	Female	Shoes	€125	4 March
John	75	Male	Laptop	€695	5 March
George	41	Male	Beer	€4	7 March
…	…	…	…	…	…
George	41	Male	Shirt	€25	9 March

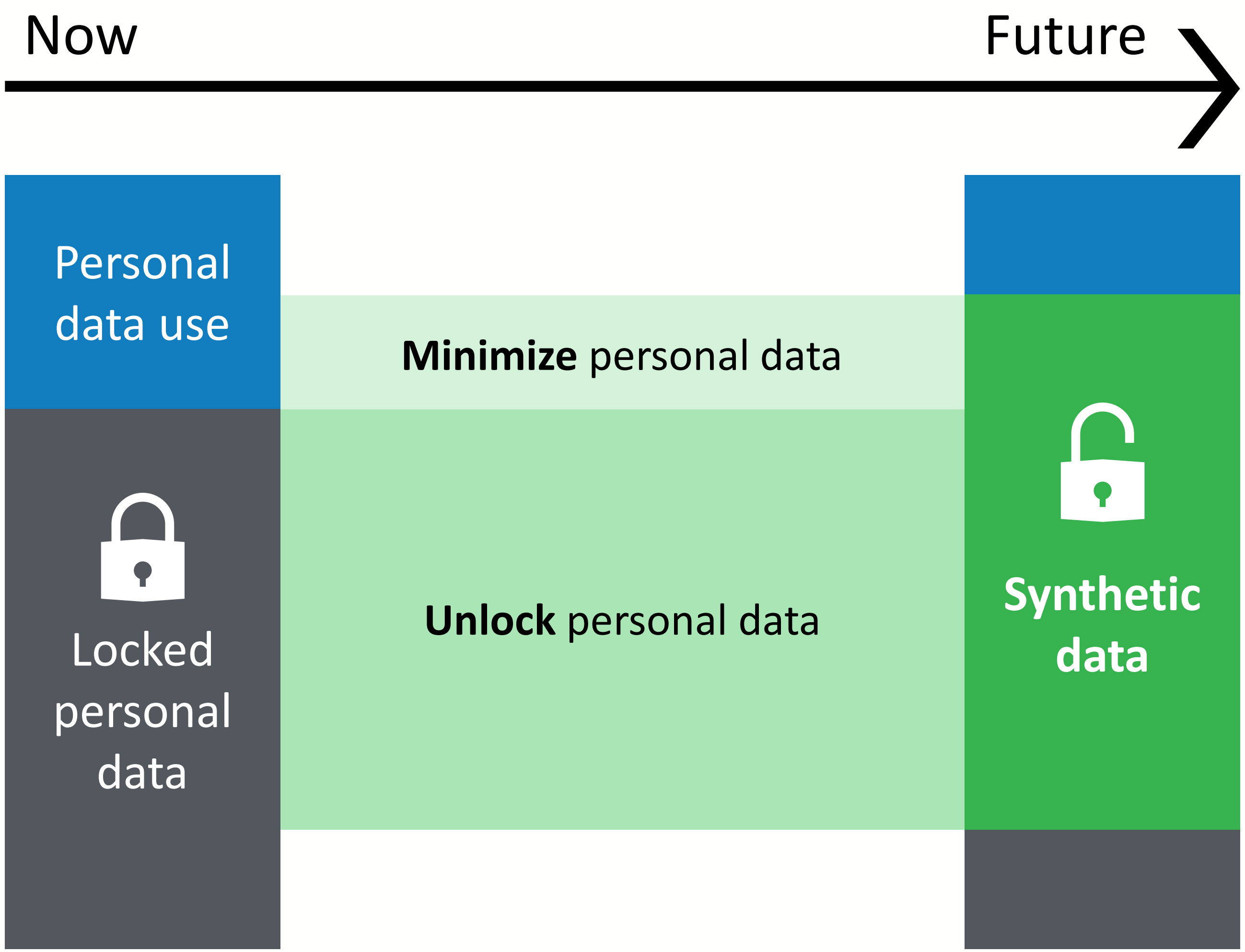
Synthetic data

Name	Age	Gender	Item	Price	Data
Emily	23	Female	Sofa	€790	1 March
Emily	23	Female	Scarf	€40	3 March
Carlos	52	Male	Razor	€5	7 March
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
Sophie	35	Female	Wine	€7	9 March

No 1:1 mapping with the original data

Preserved data quality

Why use real data when you can use synthetic data?



Key benefits

1. **Less risk**
2. **More data**
3. **Faster data access**

Boost Innovation

Synthetic data boosts data-driven innovation

Agile analytics

Eliminate time-consuming governance blocking data access and innovation

Data sharing

Privacy-preserving public and third party data sharing

Data retention

Overcome legal retention periods

Testing and development

GDPR compliant test environments

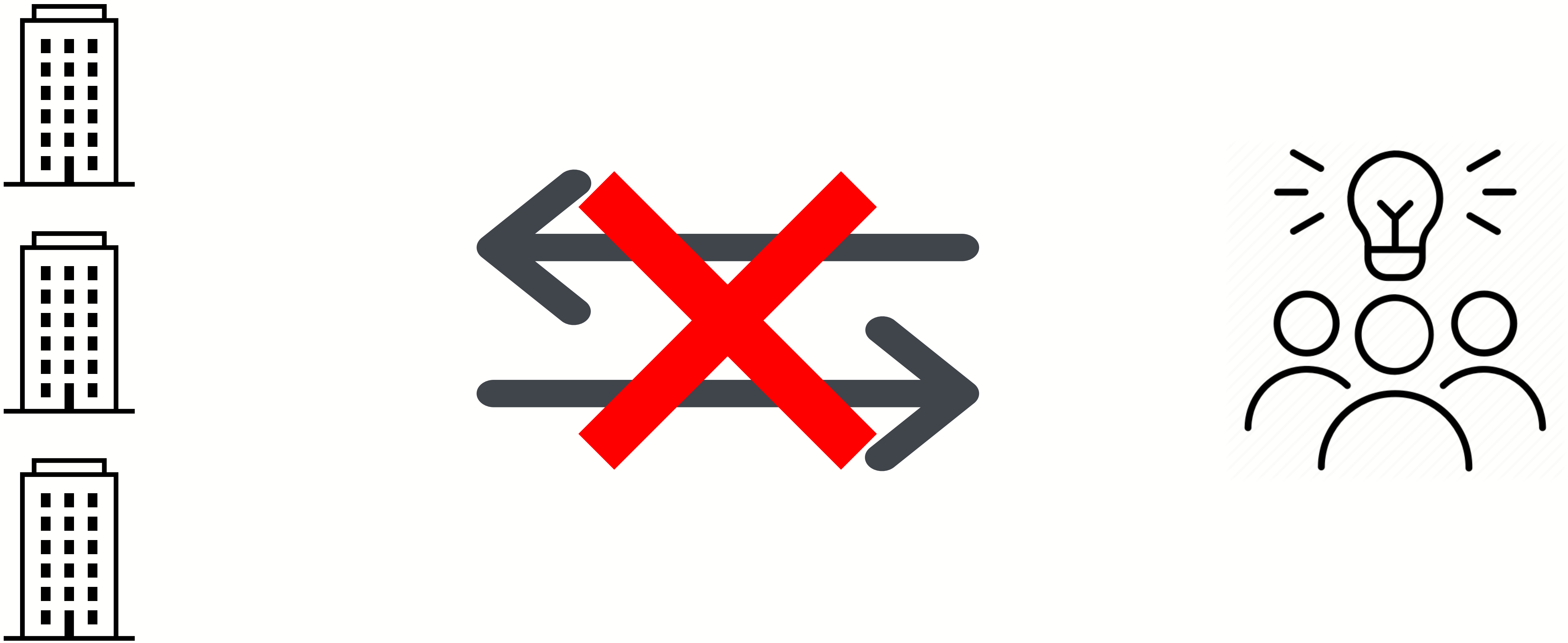
Data commerce

Responsibly monetize your data assets

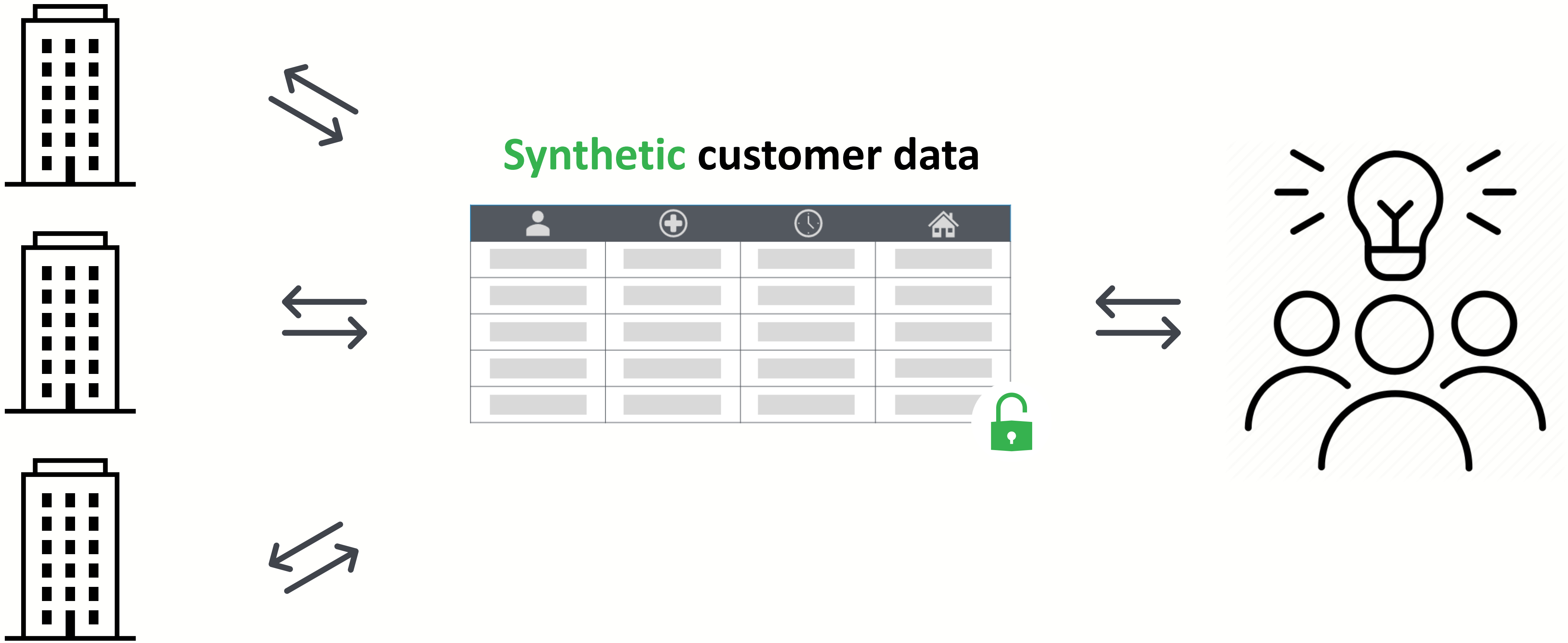
Data augmentation

Intelligent data augmentation to reduce bias and to balance datasets

Example: freely use and share anonymous synthetic data



Example: freely use and share anonymous synthetic data



SYNTHO

PROVIDING TRUST IN DATA-DRIVEN INNOVATION

info@syntho.ai

www.syntho.ai

syntho.ai/scan