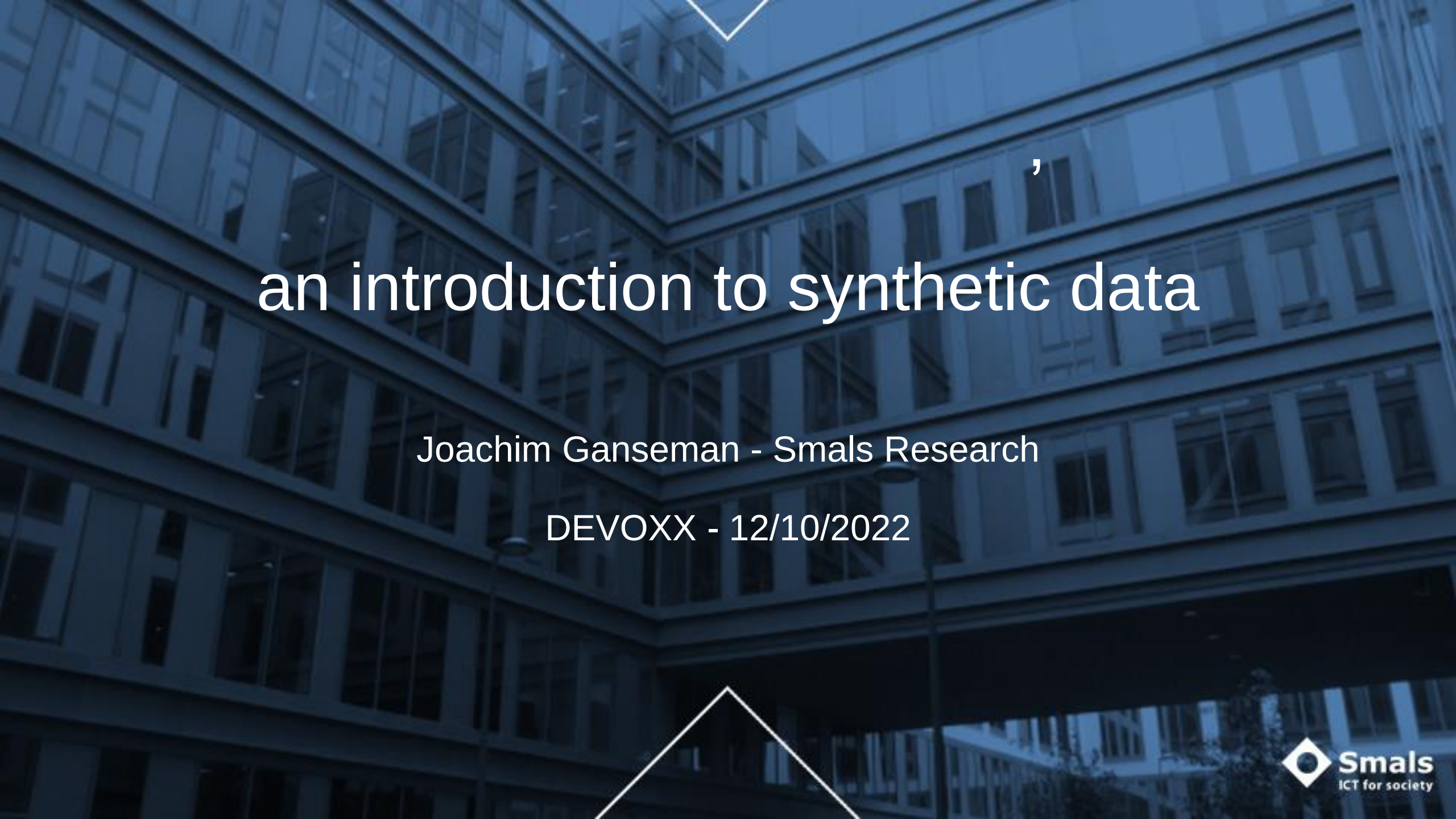


,

an introduction to synthetic data

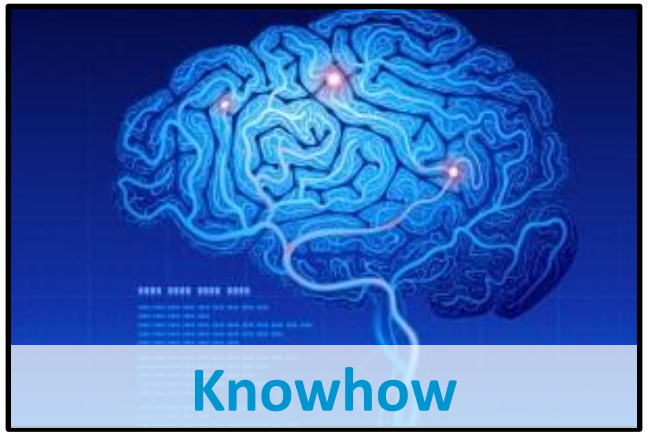
Joachim Ganseman - Smals Research

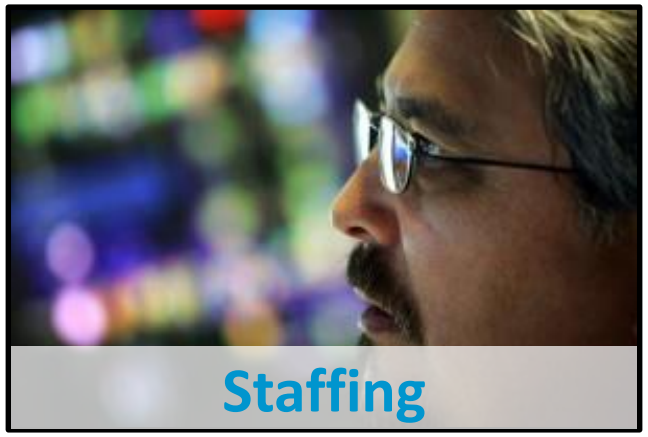
DEVOXX - 12/10/2022

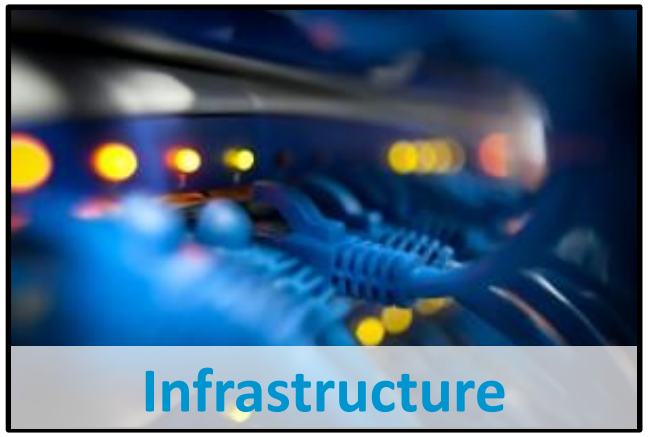
Smals
ICT for society

Smals
ICT for society

SUPPORT FOR E-GOVERNMENT







g-cloud

top EMPLOYER BELGIË BELGIQUE
CERTIFIED EXCELLENCE IN EMPLOYEE CONDITIONS

WWW.SMALS.BE

Smals Research 2022

Innovation with new technologies

Consultancy & expertise

Internal & external knowledge transfer

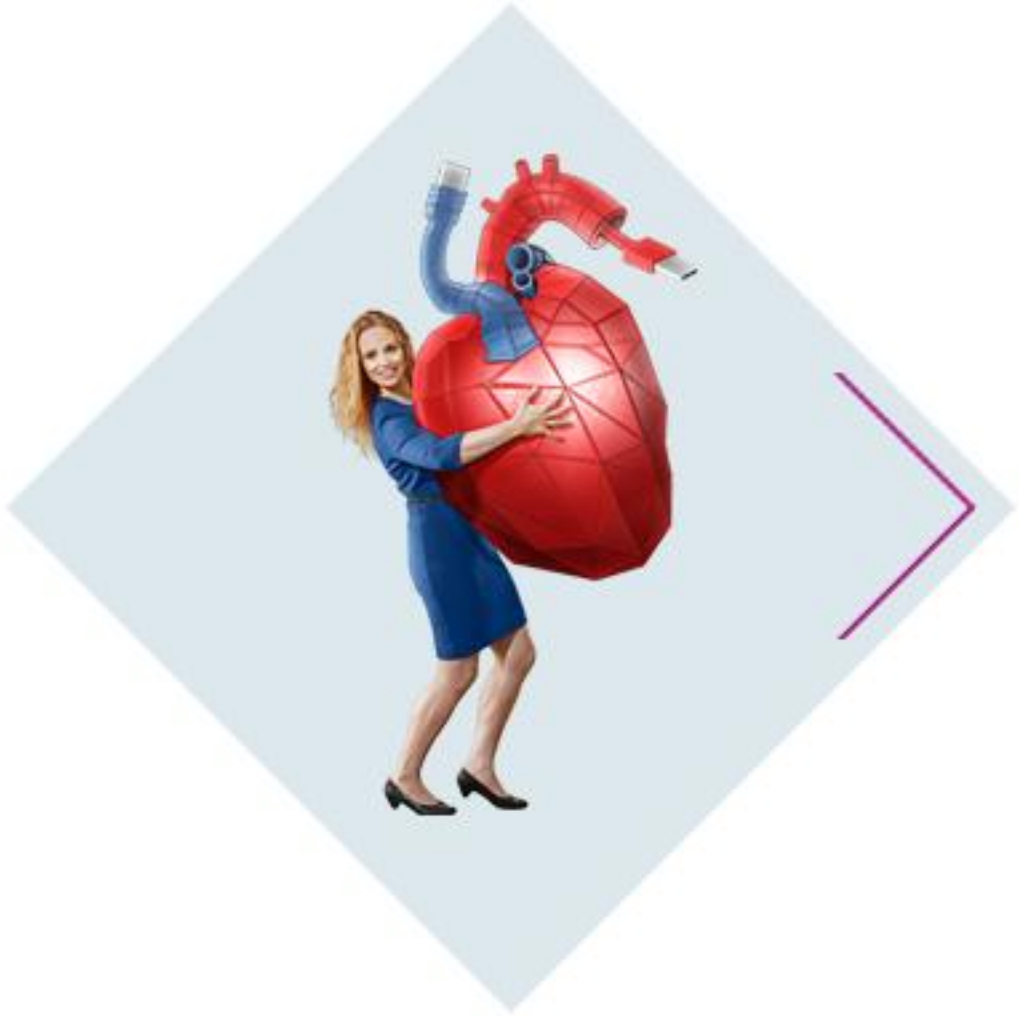
Support for going live

2022

- Low Code Development
- AI-Augmented Software Development
- Resilient Application Architectures
- On-premise Cloud: Anthos
- GIS for Analytics
- Near real time Translation
- Passwordless Authentication using FIDO
- Flexible Authorization Management
- Recommender Systems
- Named Entity Recognition Service
- Approach for detecting AI Cases
- Pseudonymization Service
- Format-Preserving Encryption Service
- Synthetic Data

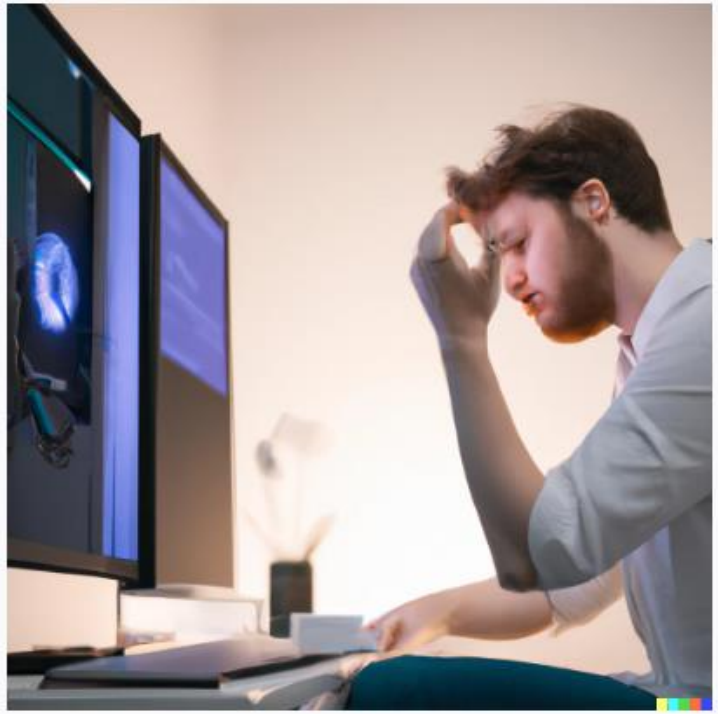
Smals ICT for society

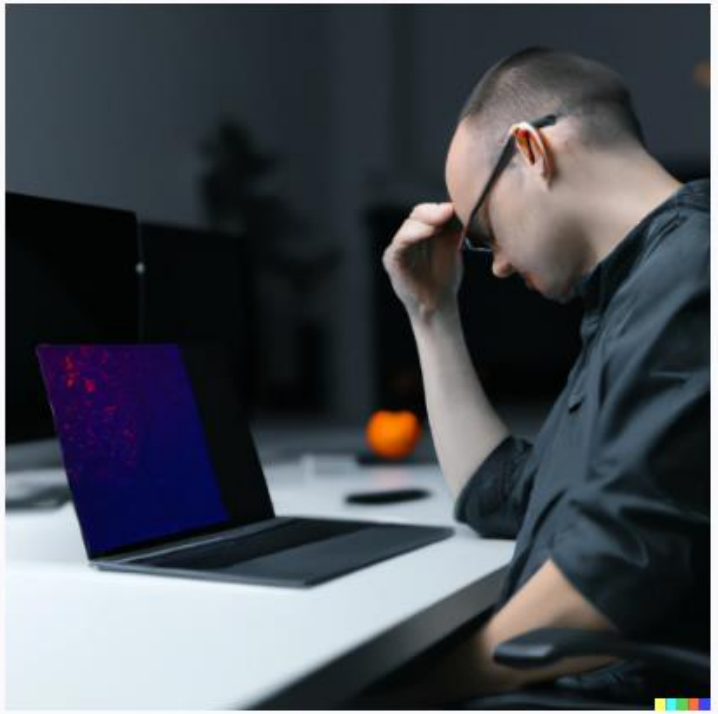
www.smalsresearch.be



Introduction

“A developer upgraded to Ubuntu 22, which broke his computer setup, one week before an important presentation” [DALL-E 2]





Synthetic data

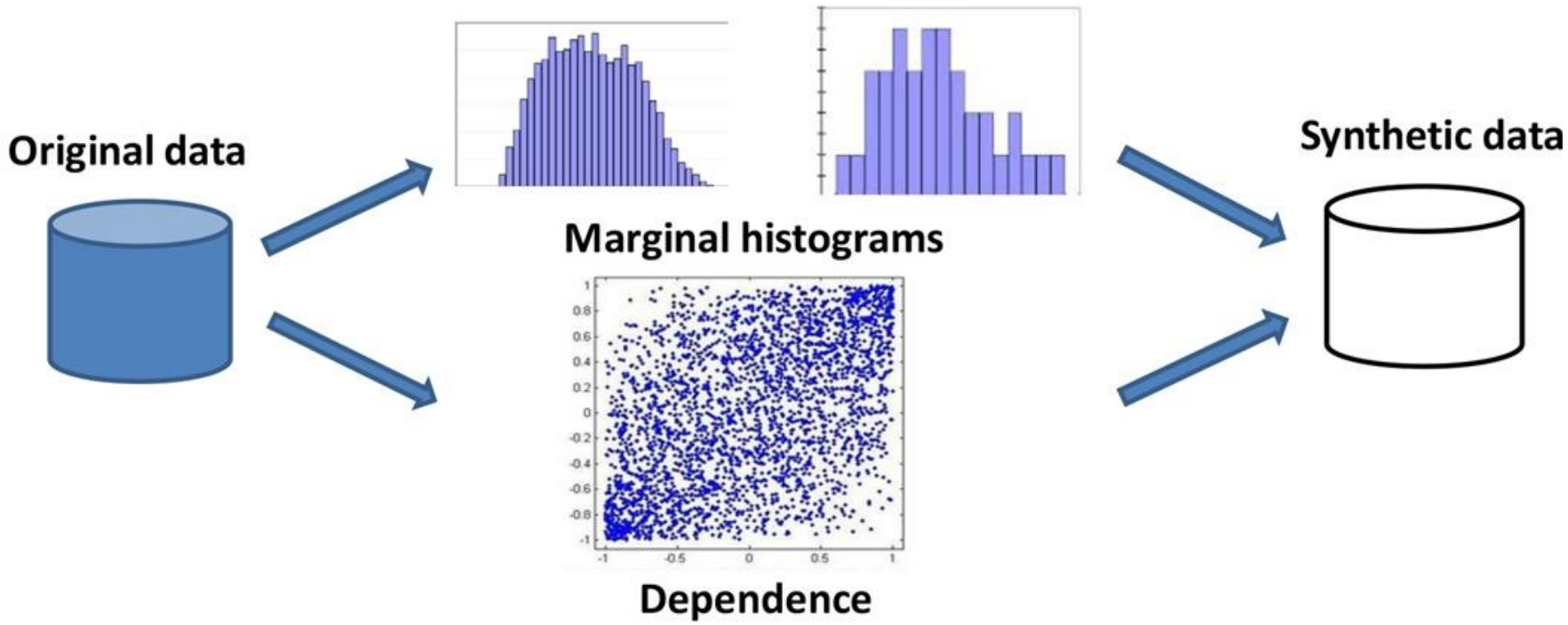
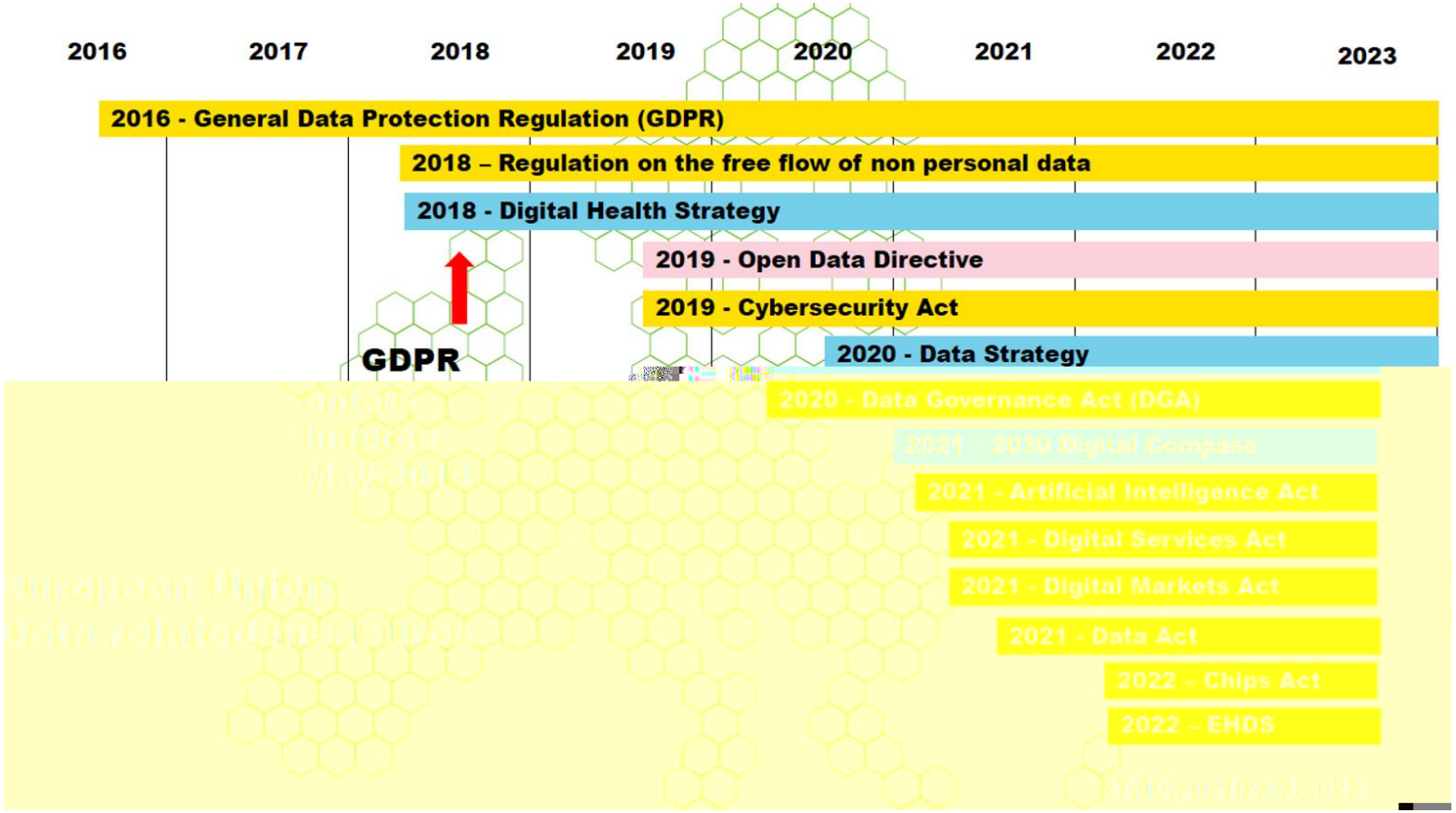


Image © Haoran Li, Li Xiong, Lifan Zhang, and Xiaoqian Jiang,
"DPSynthesizer: Differentially Private Data Synthesizer for Privacy Preserving Data Sharing"

Create a fictitious dataset that mimics an actual dataset
by learning its structure and generating plausible datapoints

Access to data is not always as simple as “sign this NDA and it’s fine”



Source: "Towards the European Health Data Space", Markus Kalliola, TEHDAS

Why?

Regulatory requirements for re-use of sensitive data, not limited to:

- “Sufficient / adequate” technical and organizational measures
- Explicit permission from data subjects
- Anonymization / aggregation
- Obligations to delete data
- Writing impact assessments, keeping registries, …

Real data can be

- Expensive to collect
- Unbalanced
- Incomplete

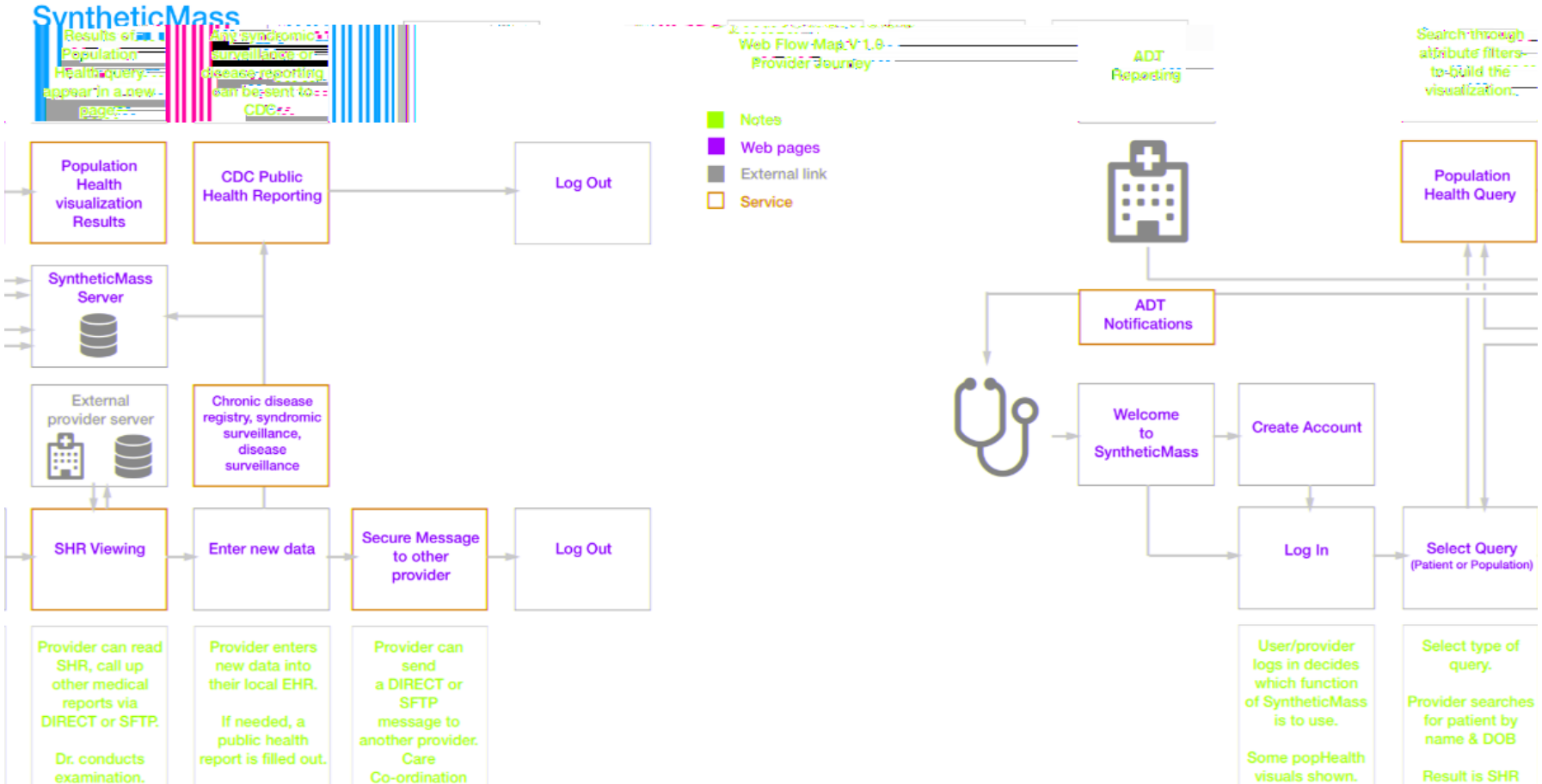
→ Synthetic data can avoid more than one headache

Smals ICT for society

Example: SyntheticMass

"SyntheticMass is a model of synthetic residents of the state of Massachusetts, with [statistically accurate] artificial health records for the fictional residents."

Tests most aspects of an eHealth-system (patient, provider, analyst)



Focus on statistical modeling for tabular data in textual form

Dive into the practicalities with an open-source approach in Python
(yeah I know)

Usecases

- Making a realistic alternative to (sensitive) data available, e.g.
  - As a data controller, to universities for research
  - As a university researcher, to the outside world for reproducibility
  - As a company, to the architects, developers and testers that build your software
- Data augmentation for ML applications
- Realistic simulations / generate test data
- …

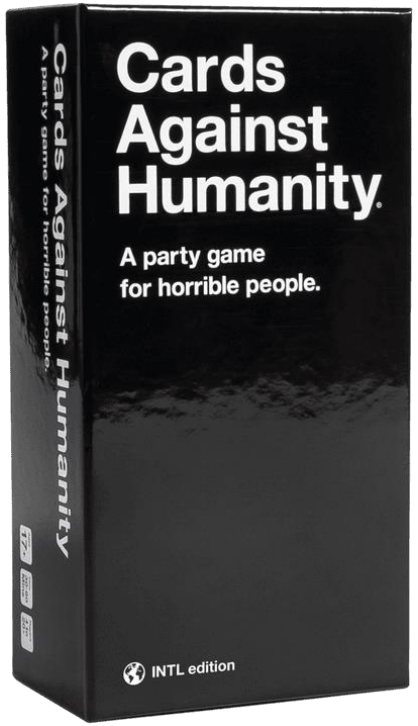
Approaches

Lorem ipsum …

Predefined structure/schema in which the gaps need to be filled

- Random number/text generators
- Lists of names, addresses, cities, locales…
- Shuffling existing data
- Inverse of regex matching
- …

Libraries that do this will offer several generators for common data types, formats and locales



Faker [https://faker.readthedocs.io/]

Also for PHP, Perl, Ruby, Java, …

Mimesis [https://mimesis.name/]

```
from mimesis import Generic
from mimesis.locales import Locale
g = Generic(locale=Locale.ES)

g.datetime.month()
# Output: 'Agosto'

g.code.imei()
# Output: '353918052107063'

g.food.fruit()
# Output: 'Limón'
```

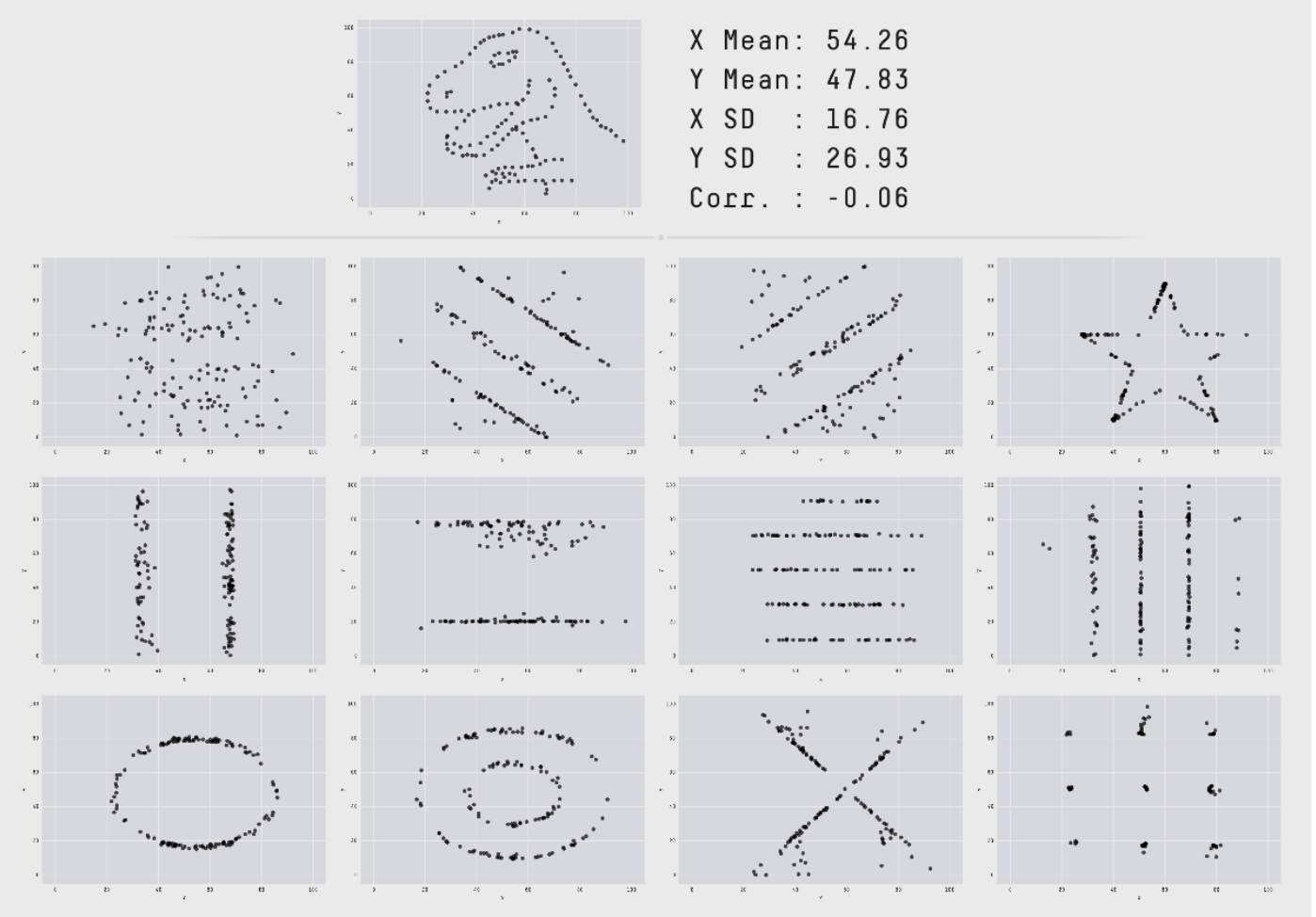
```
from faker import Faker
fake = Faker('it_IT')
for _ in range(10):
    print(fake.name())

# 'Elda Palumbo'
# 'Pacifico Giordano'
# 'Sig. Avide Guerra'
# 'Yago Amato'
# 'Eustachio Messina'
# 'Dott. Violante Lombardo'
# 'Sig. Alighieri Monti'
# 'Costanzo Costa'
# 'Nazzareno Barbieri'
# 'Max Coppola'
```

```
>>> Faker.seed(0)
>>> for _ in range(5):
...     fake.vat_id()
...
'BE6048764759'
'BE8242194892'
'BE1157815659'
'BE8778408016'
'BE9753513933'
```

Extensible with custom generation routines and schemas for your own datatypes

Smals ICT for society

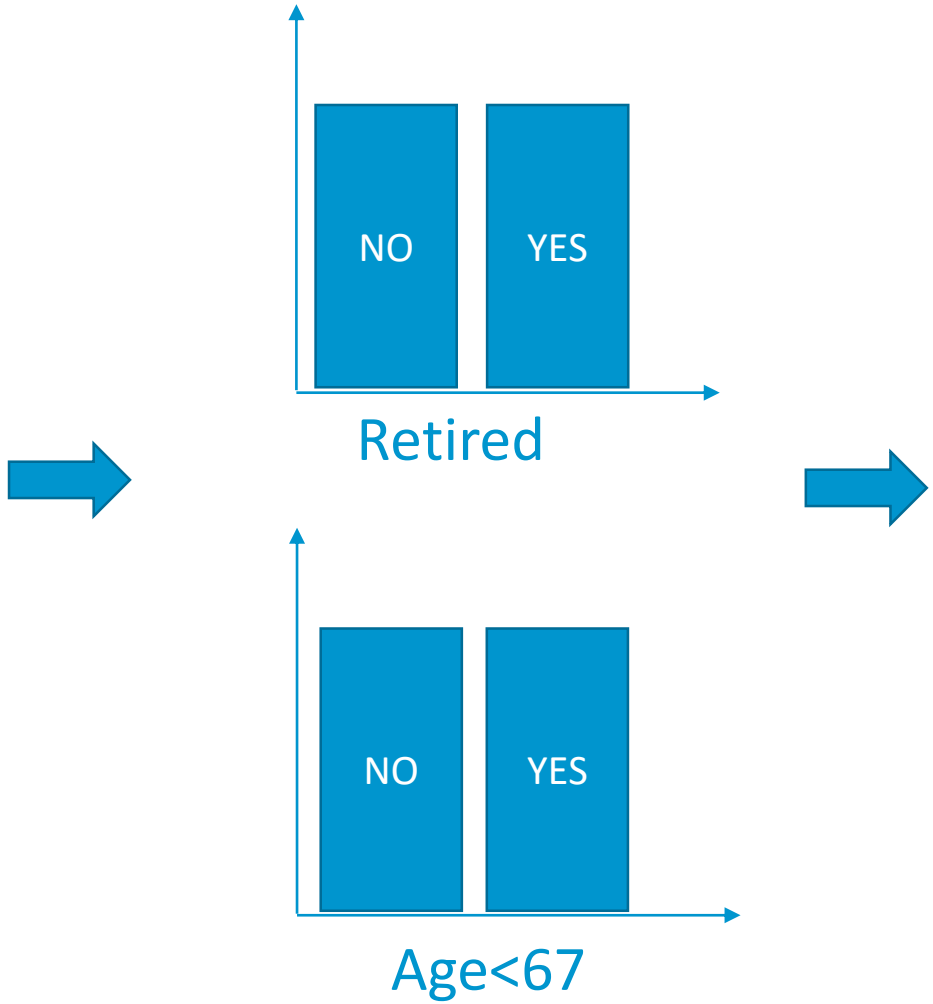
Similar “summary statistics” ≠ good mimicking of original data



Smals
ICT for society

Conservation of distributions ≠ conservation of correlations

Age	Retired
15	FALSE
24	FALSE
50	FALSE
68	TRUE
72	TRUE
88	TRUE



Age	Retired
88	FALSE
68	TRUE
50	FALSE
15	TRUE
72	FALSE
24	TRUE

Smals
ICT for society

1. Learn (joint) distributions from original data → model
2. Repeatedly “sample” this model

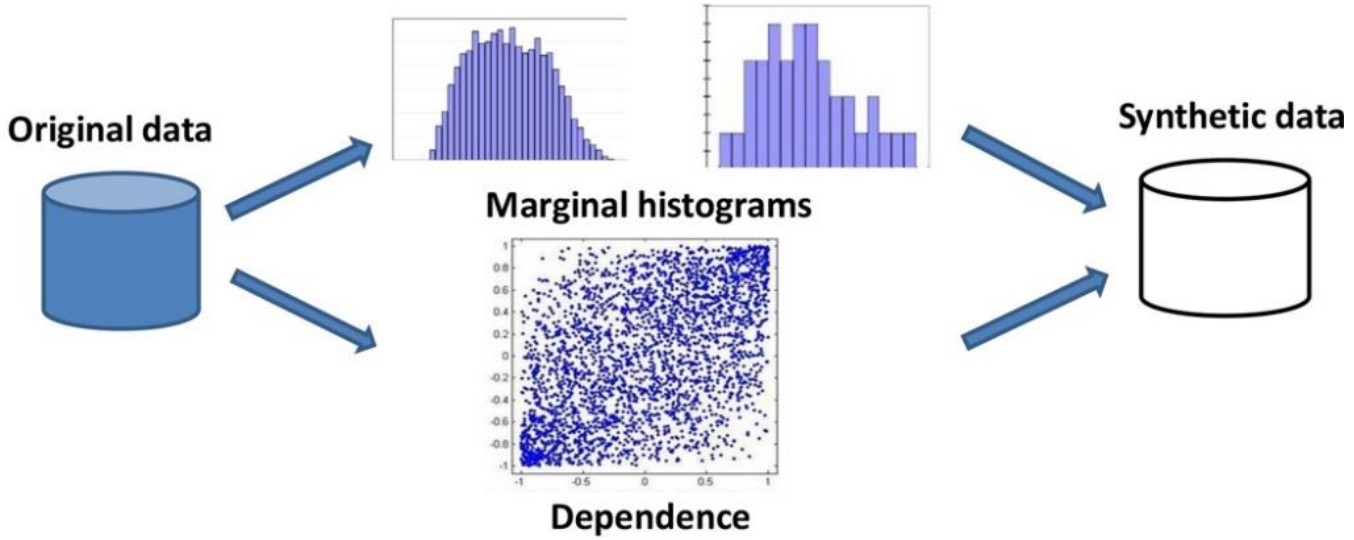


Image © Haoran Li, Li Xiong, Lifan Zhang, and Xiaoqian Jiang,
“DPSynthesizer: Differentially Private Data Synthesizer for Privacy Preserving Data Sharing”

Conditional sampling: keep some values fixed to sample a subset

Smals
ICT for society

Generate data for rare or expensive events

Create annotated datasets for machine learning

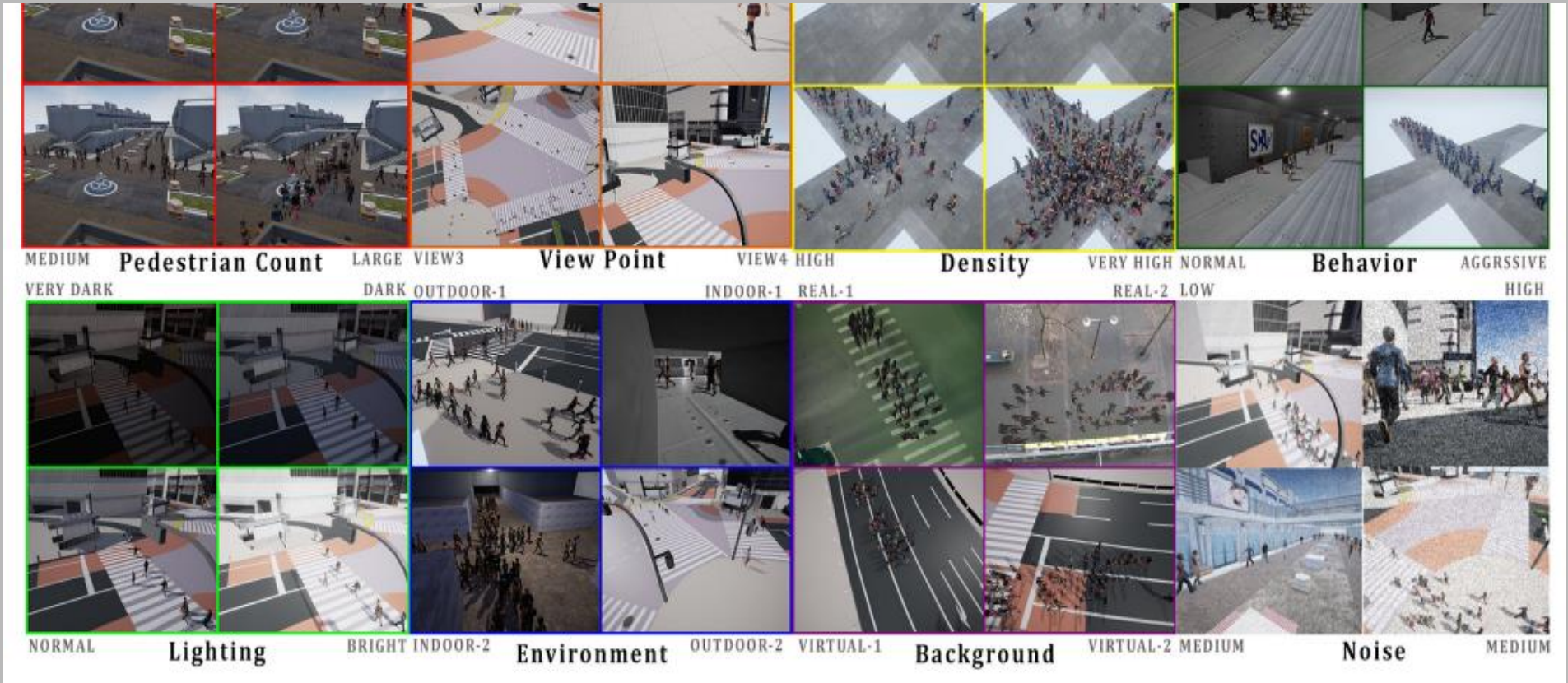


Image © Ernest Cheung, T.K. Wong, Aniket Bera, Xiaogang Wang, Dinesh Manocha
"LCrowdV: Generating Labeled Videos for Simulation-based Crowd Behavior Learning"

Smals ICT for society

Agent-based Modeling

Complex dynamic systems (e.g. physics/biology simulations)
Generate interaction data
Tools: specialized frameworks: Repast (C++), MASON (Java), Mesa (Python), …

Virtual Environments

Robotics, self-driving, VR
Generate large amounts of different scenarios
Tools: 3D engines: Unity3D, GTA, X-Plane, …

Synthesizers

Audio, speech, generative artwork
Generate multimedia from (textual) annotations
Tools: text-to-speech systems, MIDI, WaveNet, Processing, …

In practice

Let’s take a dataset and pick a software library:

	age	workclass	fnlwgt	education	marital-status	occupation	relationship	race	sex	hours-per-week	native-country	capital	income
0	39	State-gov	77516	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	2174	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	13	United-States	0	<=50K
2	38	Private	215646	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	40	United-States	0	<=50K
3	53	Private	234721	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	40	United-States	0	<=50K
4	28	Private	338409	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	40	Cuba	0	<=50K

[Source: Kaggle, “Adult Census Income” dataset: https://www.kaggle.com/datasets/uciml/adult-census-income]

SYNTHEA

SyntheticMass

MITRE

synthia

SDV
The Synthetic Data Vault

data RESPONSIBLY

BENERATOR
THE SMART WAY
TO HANDLE DATA

https://sdv.dev/

One of the larger open source Synthetic Data libraries (in Python)

Supported methods:

- Statistical: “copula” multivariate distributions
- Deep learning: CTGAN (GAN for tabular data) / TVAE (Tabular Var. AutoEncoder)
- Time series (in development) : PAR (Probabilistic AutoRegression)
- Relational data (linked tables) : hierarchical model with underlying copula models

Encode your own constraints

Some evaluation and benchmarking options (limited)

Commercial support by Datacebo Inc.

Getting started

```python
from sdv import load_demo, SDV

# Use pre-loaded demo tables
metadata, tables = load_demo(metadata=True)

sdv = SDV()
sdv.fit(metadata, tables)

synthetic_data = sdv.sample()
print(synthetic_data)
```

```
1 # Display basic statistics about the dataset
2 print("Data description - categoricals:")
3 actual_data.describe(include='object', datetime_is_numeric=True)
```

Data description - categoricals:

	workclass	education	marital-status	occupation	relationship	race	sex	native-country	income
count	48842	48842	48842	48842	48842	48842	48842	48842	48842
unique	7	16	7	15	6	5	2	42	2
top	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States	<=50K
freq	33906	15784	22379	6172	19716	41762	32650	43832	37155

```python
# display count of every categorical value
for var in OPTS['categorical_vars']:
    if var in actual_data:
        actual_data[var].value_counts()

```

```
Male      32650
Female    16192
Name: sex, dtype: int64
```

```
United-States                 43832
Mexico                          951
?                               857
Philippines                     295
Germany                         206
Puerto-Rico                     184
Canada                          182
El-Salvador                     155
                              ⋮
Outlying-US(Guam-USVI-etc)       23
Yugoslavia                       23
Scotland                         21
Honduras                         20
Hungary                          19
Holand-Netherlands                1
Name: native-country, dtype: int64
```

```python
# Display basic statistics about the dataset
print("Data description - integers:")
actual_data.describe(datetime_is_numeric=True)
```

Data description - integers:

	fnlwgt	hours-per-week	capital		age
0	4.884200e+04	48842.000000	48842.000000	count	48842.000000
5	1.896641e+05	40.422382	991.565313	mean	38.643585
0	1.056040e+05	12.391444	7475.549906	std	13.710510
0	1.228500e+04	1.000000	-4356.000000	min	17.000000
0	1.175505e+05	40.000000	0.000000	25%	28.000000
0	1.781445e+05	40.000000	0.000000	50%	37.000000
0	2.376420e+05	45.000000	0.000000	75%	48.000000
0	1.490400e+06	99.000000	99999.000000	max	90.000000

Smals
ICT for society

Results out-of-the-box (statistical Copula model)

	age	workclass	fnlwgt	education	marital-status	occupation	relationship	race	sex	hours-per-week	native-country	capital	income
0	39	State-gov	77516	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	2174	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	13	United-States	0	<=50K
2	38	Private	215646	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	40	United-States	0	<=50K
3	53	Private	234721	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	40	United-States	0	<=50K
4	28	Private	338409	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	40	Cuba	0	<=50K

Generated 48842 synthetic samples. Displaying the first few rows:

	age	workclass	fnlwgt	education	marital-status	occupation	relationship	race	sex	hours-per-week	native-country	capital	income
0	46	Private	129352	Some-college	Married-civ-spouse	Farming-fishing	Not-in-family	Black	Male	52	South	1775	<=50K
1	21	Private	466882	5th-6th	Never-married	Prof-specialty	Not-in-family	White	Male	43	United-States	7510	<=50K
2	52	Local-gov	129500	Some-college	Divorced	Prof-specialty	Husband	White	Male	59	United-States	41618	<=50K
3	37	Self-emp-inc	124908	Some-college	Married-civ-spouse	Tech-support	Not-in-family	White	Female	43	United-States	7586	<=50K
4	38	Federal-gov	149033	Some-college	Married-civ-spouse	Adm-clerical	Wife	White	Male	42	South	1889	<=50K

Smals
ICT for society

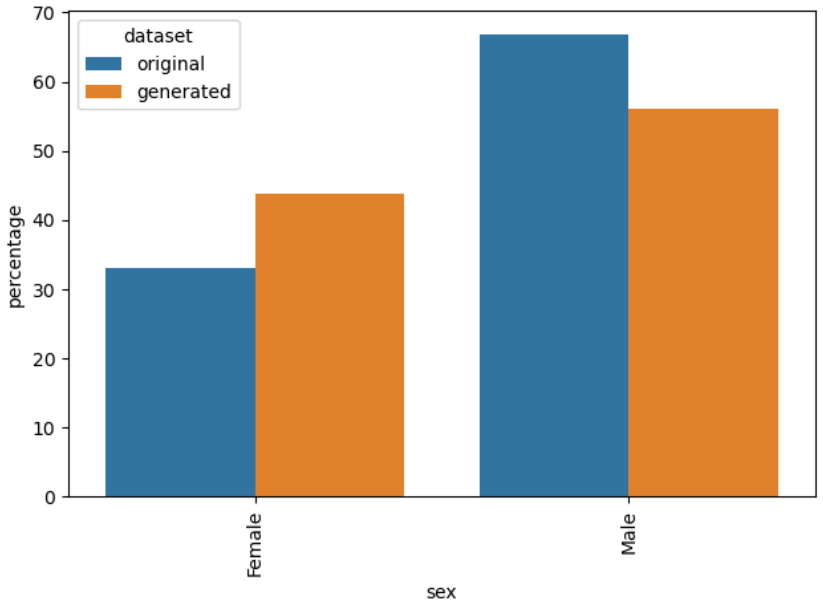
```python
# Display basic statistics about the dataset
print("Data description - categoricals:")
actual_data.describe(include='object', datetime_is_numeric=True)
```

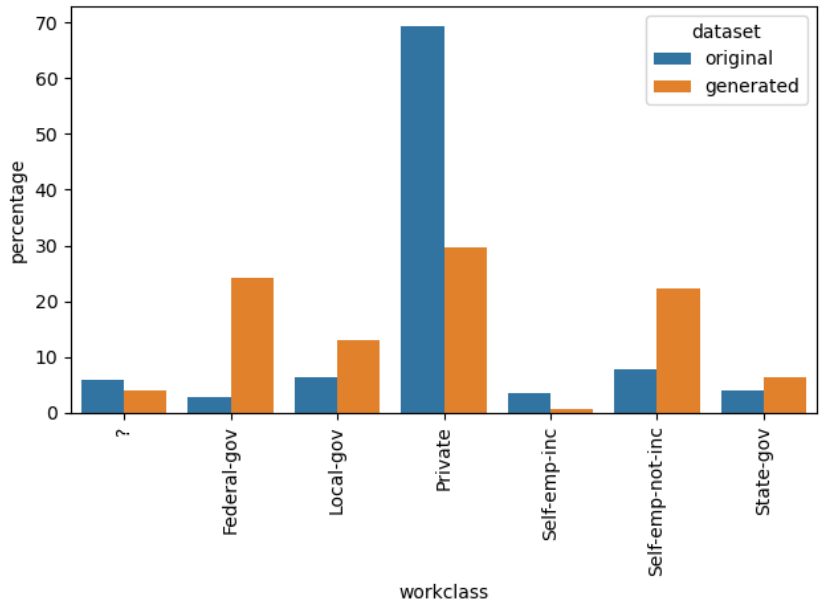
```
Data description - categoricals:
```

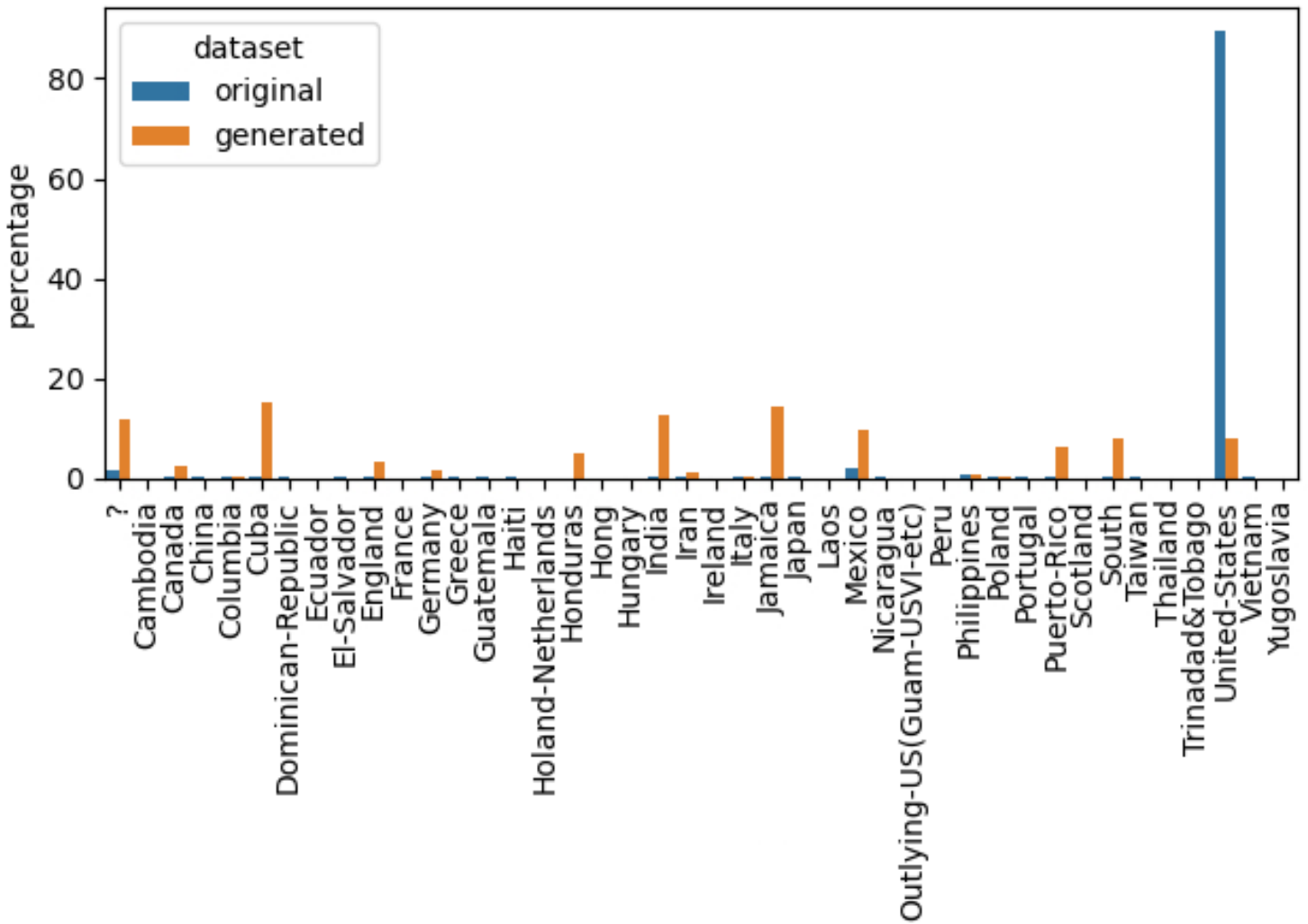
	workclass	education	marital-status	occupation	relationship	race	sex	native-country	income
count	48842	48842	48842	48842	48842	48842	48842	48842	48842
unique	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
top	United-States	<=50K	[illegible]	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male
freq	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

	workclass	education	marital-status	occupation	relationship	race	sex	native-country	income
count	48842	48842	48842	48842	48842	48842	48842	48842	48842
unique	7	16	7	15	6	4	2	24	2
top	Private	Some-college	Married-civ-spouse	Exec-managerial	Husband	White	Male	Cuba	<=50K
freq	14538	15604	19432	6293	15757	26779	27479	7253	31275

Smals
ICT for society


dataset
original
generated
percentage
sex
Female
Male


dataset
original
generated
percentage
workclass
?
Federal-gov
Local-gov
Private
Self-emp-inc
Self-emp-not-inc
State-gov


dataset
original
generated
percentage
?
Cambodia
Canada
China
Columbia
Cuba
Dominican-Republic
Ecuador
El-Salvador
England
France
Germany
Greece
Guatemala
Haiti
Holand-Netherlands
Honduras
Hong
Hungary
India
Iran
Ireland
Italy
Jamaica
Japan
Laos
Mexico
Nicaragua
Outlying-US(Guam-USVI-etc)
Peru
Philippines
Poland
Portugal
Puerto-Rico
Scotland
South
Taiwan
Thailand
Trinadad&Tobago
United-States
Vietnam
Yugoslavia

SDV's default built-in models deal particularly badly with:

- Highly skewed or irregular distributions
- Distributions with long tails
- Rare or unique values (tend to be ignored)

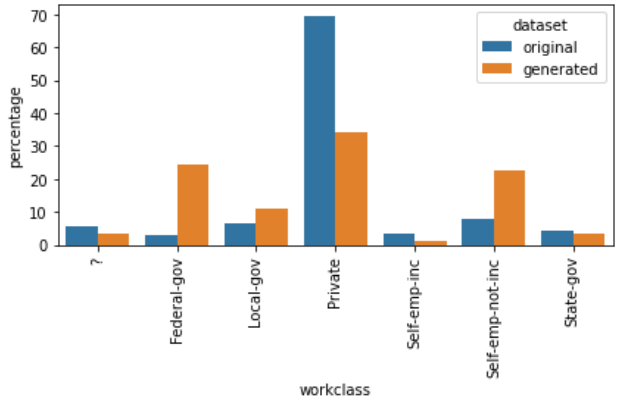
→ but this is all very common in real life datasets!

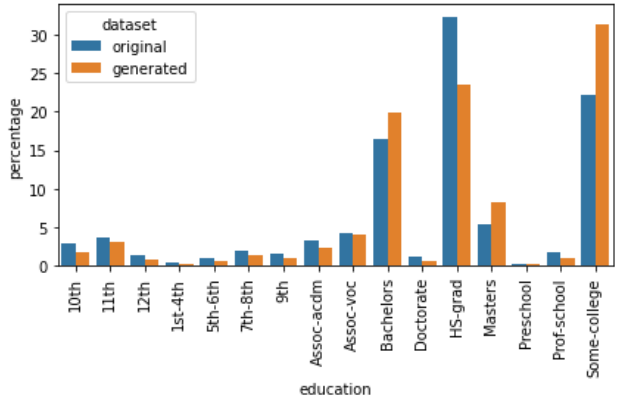
There is a structural limit:
for rare values, there are not enough datapoints to be able to learn conditional distributions, nor correlations with other variables

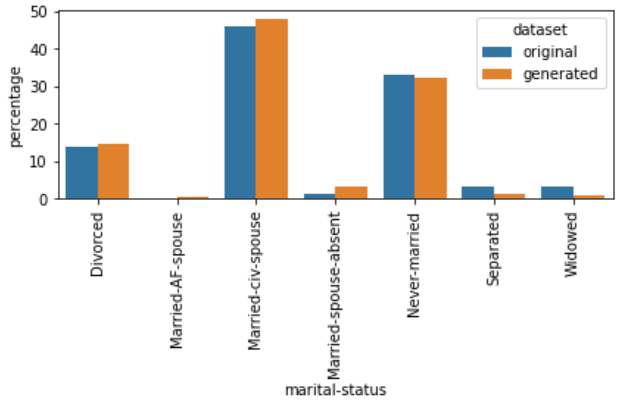
Tweaking SDV's parameters can help but doesn't do miracles

Adding CTGAN to the mix

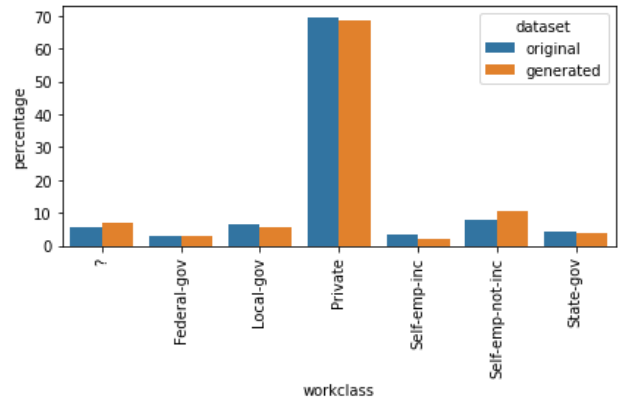
Copula (stat.)

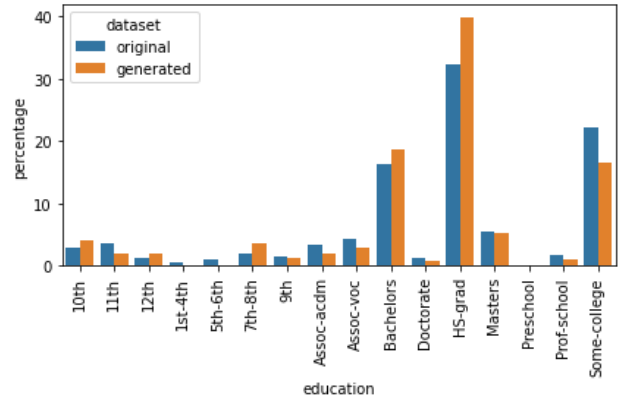


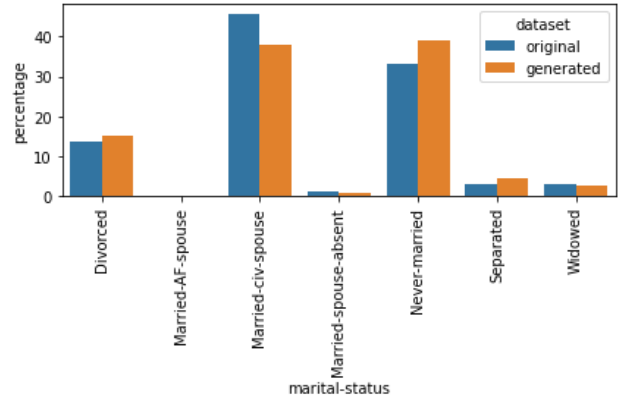




Copula+CTGAN

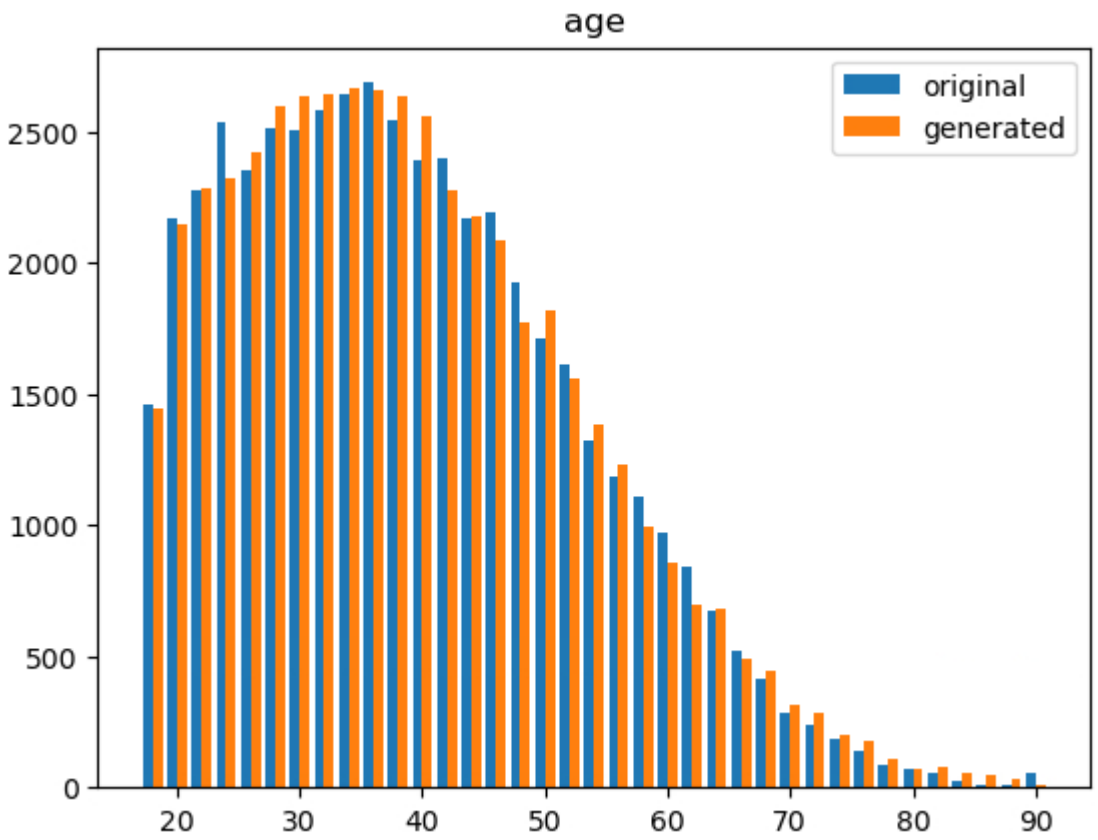


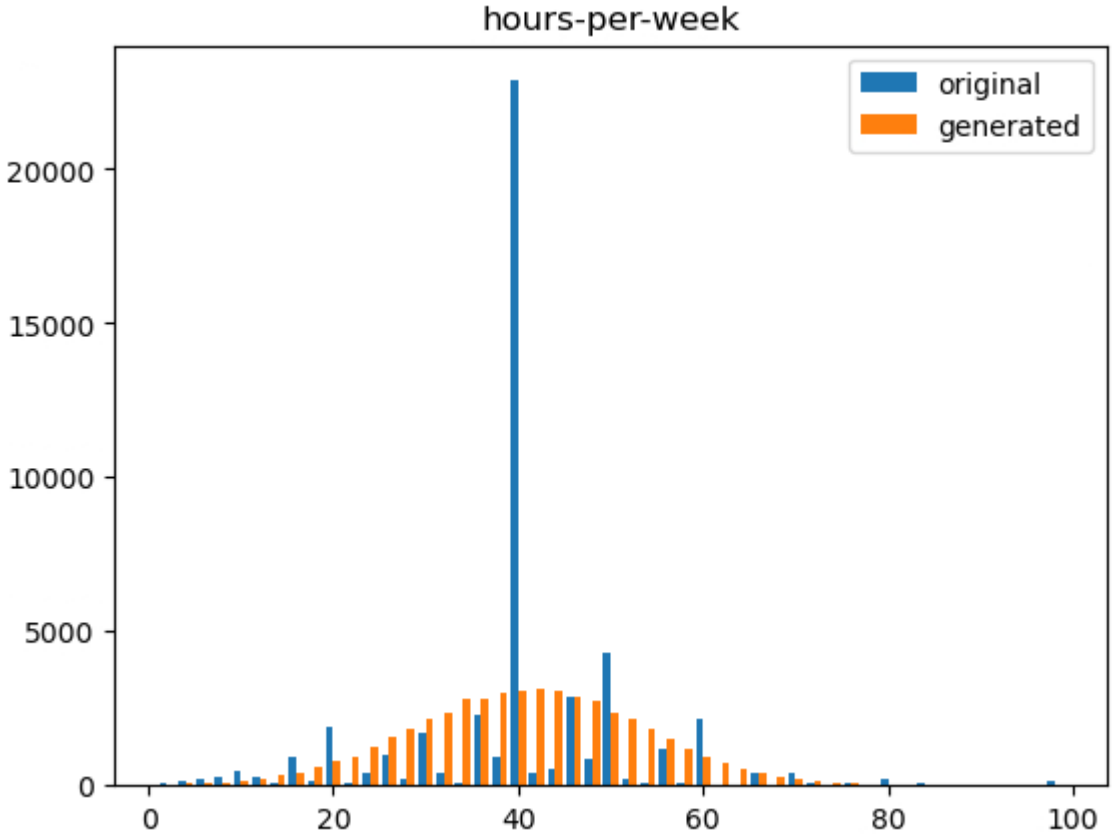




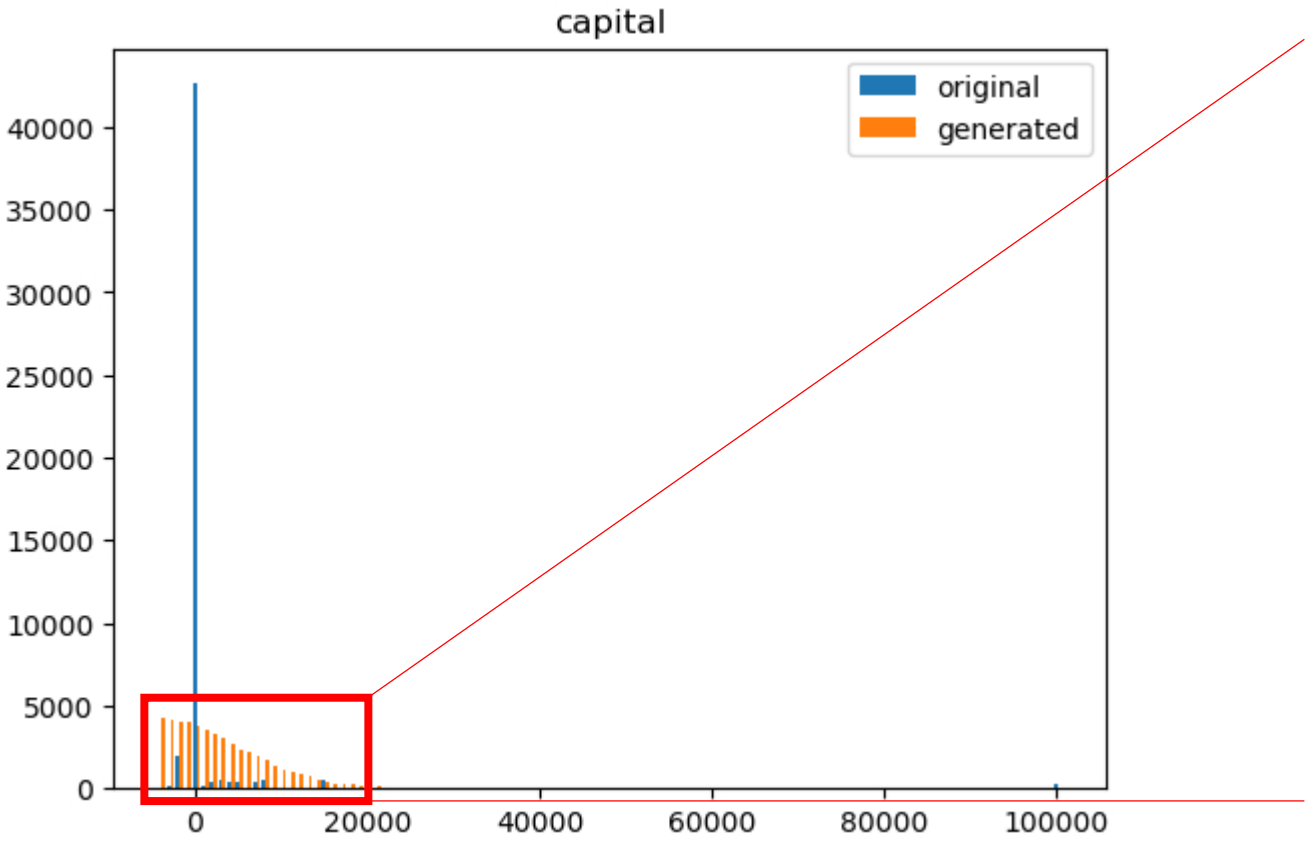
Smals
ICT for society

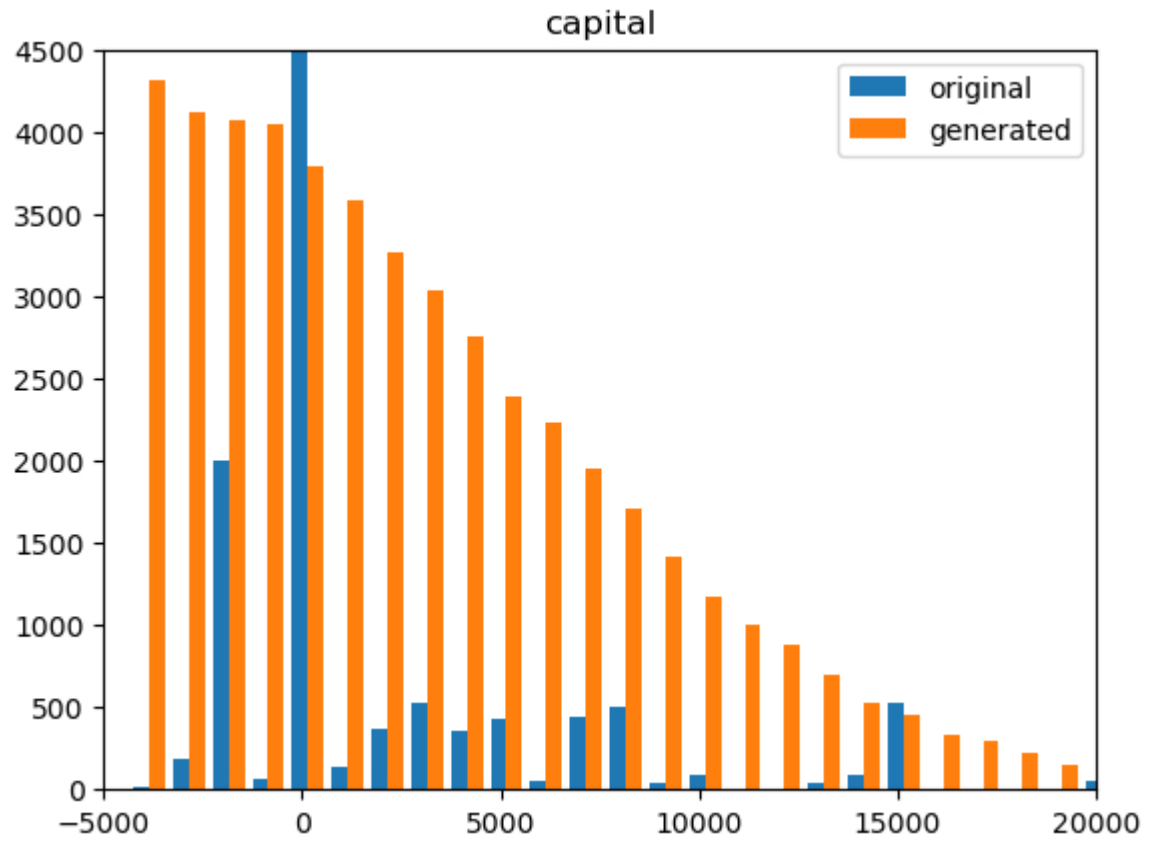
What happens with numbers?





What happens with numbers?






```
1 actual_data['capital'].nunique()
```

```
221
```

```
1 new_data['capital'].nunique()
```

```
17428
```

Smals
ICT for society

SDV does not know the meaning of any number

Assumes smooth distributions -> interpolates when sampling

In this dataset

- “hours-per-week” behaves more like a categorical variable (fulltime, parttime, …)
- “capital” has plenty of 0, which might mean “no info / null” instead of “capital = $0”

It’s not as easy as
“Load data & press start”…

Missing data

What does a missing data point mean?

- Numericals: can appear as NaN, but sometimes also 0, -1, -2^{32}, …
- A missing year of birth doesn't mean a person was born in the year 0
- Booleans: missing data = FALSE or missing data = third category?

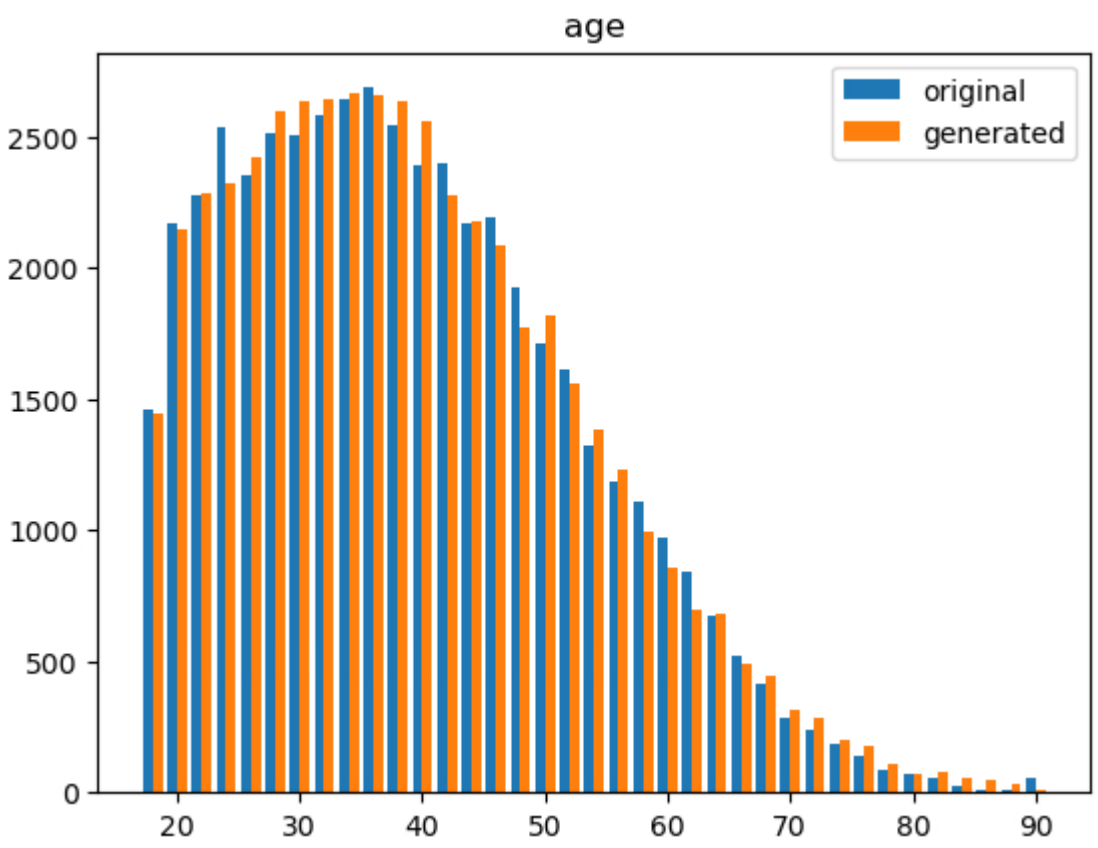
SDV deals in a peculiar way with missing data:

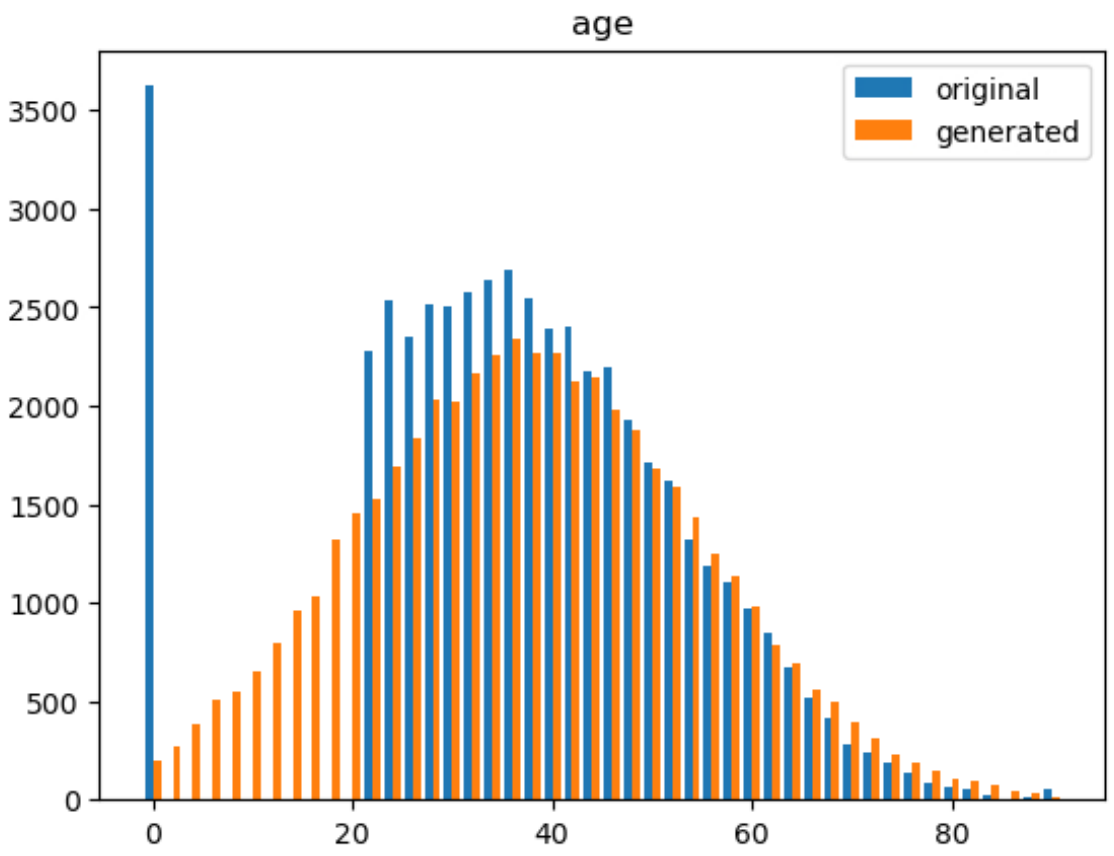
- Categorical variable: considered just another category within the variable
- Numerical variables are split in 2
  - 1 Boolean variable to decide "is it null"?
  - 1 numerical variable trained on all non-null values
- Boolean variable: null values not supported

Not possible (yet) to generate data conditional on the absence of other data.

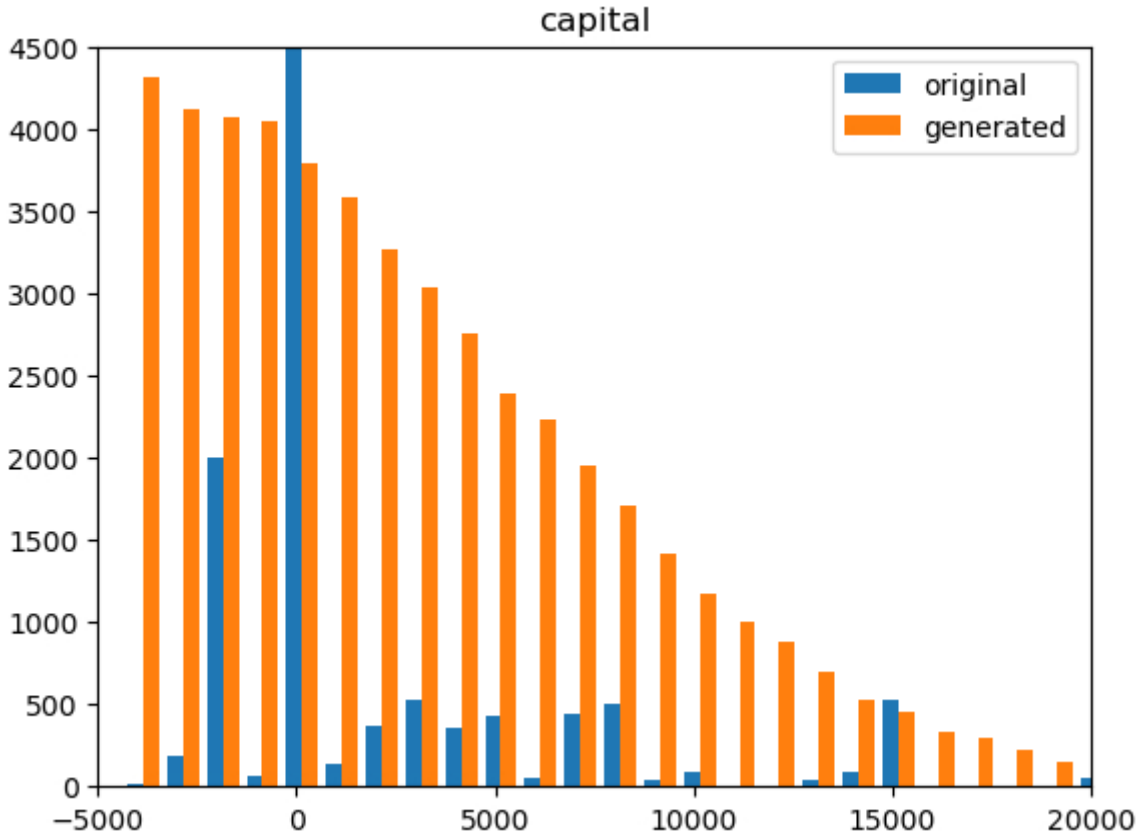
Smals
ICT for society

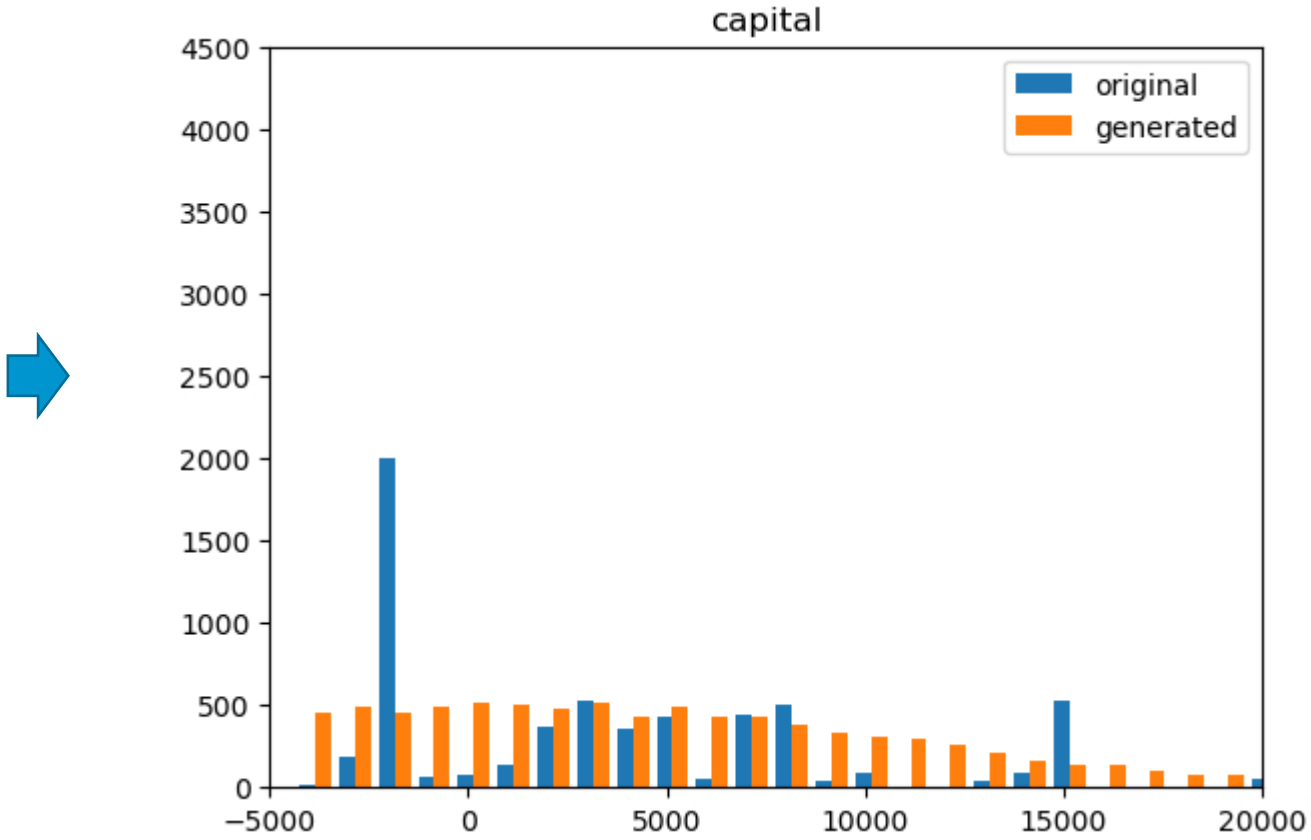
Suppose our dataset masked the ages of all people <21 by setting it to -1





Suppose that ‘capital==0’ actually means: no data (NaN)





Some datasets contain count data:

Frost	Rain	Sun	# days
No	No	Yes	52
No	Yes	No	43
Yes	No	Yes	1
No	No	No	187
No	Yes	Yes	10

(obviously this table does not say that frost is present in 20% of the datapoints)

For a correct data model:

1. “unroll” the data (= undo the counting, expand) and delete count variable
2. train the model and generate synthetic data
3. recount / regroup the synthetic data

Forcing dataset coverage

There are no guarantees that a particular value will be drawn from the distribution, especially when those values are rare:

	race	sex	native-country
count	48842	48842	48842
unique	5	2	42
top	White	Male	United-States
freq	41762	32650	43832

	race	sex	native-country
count	48842	48842	48842
unique	4	2	24
top	White	Male	Cuba
freq	26779	27479	7253

Conditional generation allows to force the desired quantity of a value

Conditioning on rare values may give repetitive results
(because not enough data to properly learn conditional distributions)

Dependent columns

Columns can be fully dependent on others:

X	Y	X+Y	2Y-X
2	4	6	6
8	7	15	6
0	1	1	2
1	0	1	-1

SDV cannot detect dependencies, only approximately learns correlations

For a correct data model:

- Remove dependent columns
- Learn model and generate data
- Re-calculate and re-add the dependent columns

Smals ICT for society

Constraints

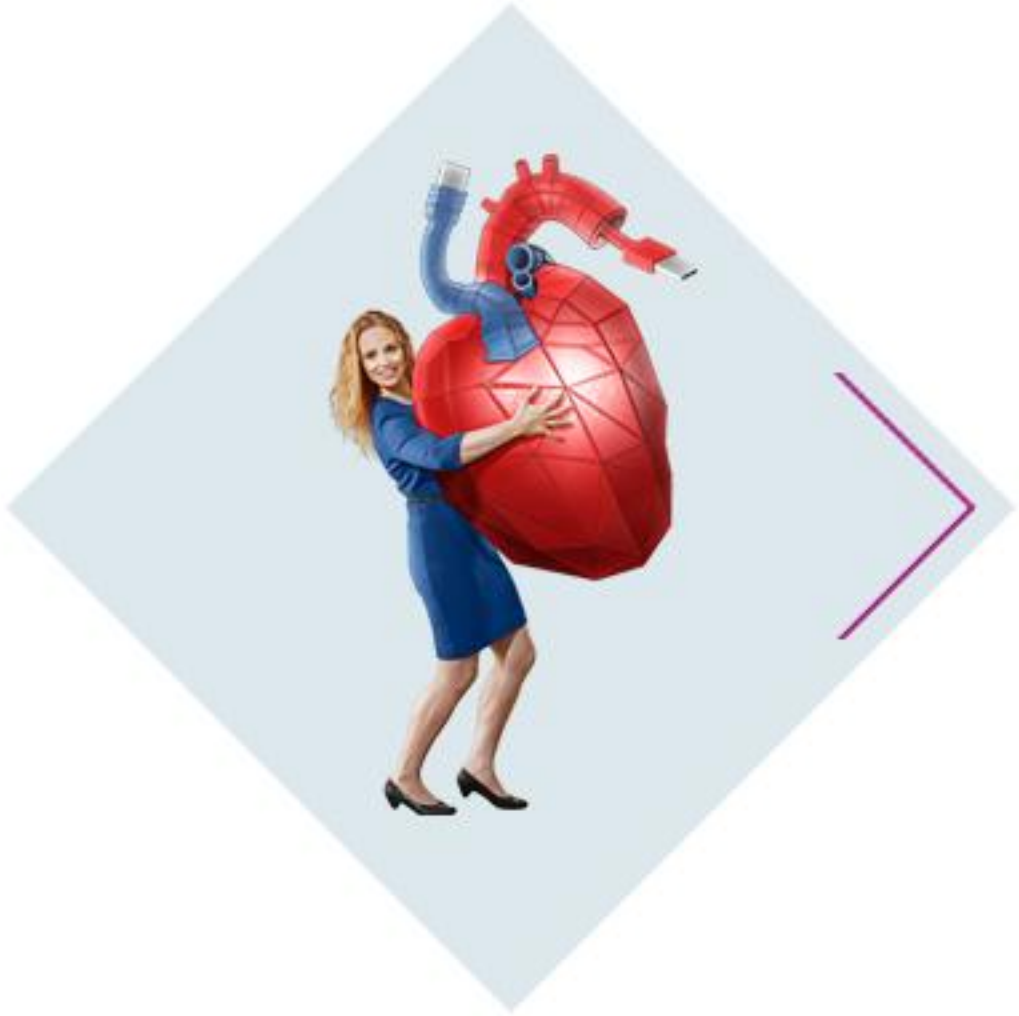
The meaning of the data may imply other dependencies

- Date of birth < date of death
- City = Antwerp → Province = Antwerp → postcode.startsWith('2')
- Age < 18 → child_benefits = true
- Distance > 0
- $ORCL = Oracle
- A 25-year-old in year X, cannot be 36 in year X+1

Encode these in constraints

- Some can be incorporated in the model
- Others can be enforced by fusing columns
- Others can be enforced through rejection sampling

Smals ICT for society



Wrapping up

Takeaways

Synthetic data tools rarely work out of the box for datasets “in the wild”

- Your model likely needs finetuning -> iterative process
- You’ll often want to add custom preprocessing for your data

Long-tailed or irregular distributions complicate things
and can give rise to statistical instability

- Various strategies can mitigate the worst side-effects, but there is no silver bullet

Know your data

Smals
ICT for society

Minimize the number of columns

- To anonymize a dataset with *address* and *gender* : only synthesize new addresses
- Discard columns that don’t need to be resynthesized

Exploit knowledge about the data

- Fuse columns that are strongly correlated (e.g. city and its province)
- Use constraints to prevent generating nonsensical datapoints
- Decide what to do with outliers and missing data and why
- Merging the least-used categories into an “other” category, reduces the “long tail”

Work with a well-selected sub-dataset to speed up finetuning

Analytics?

Possibilities for analytics on synthetic data are limited!

Structure of the data is approximately mimicked

- 1 variabele statistics (min, max, avg, etc) are mostly preserved,
- Links between 2 variables (correlation, …) are somewhat preserved,
- Links between more variables (regressions, …) are poorly or not preserved,

→ The suitability of synthetic data as drop-in-replacement for real data depends on usecase and data properties…

We obtained the best results on datasets with

- Few variables (columns)
- Many datapoints for each value of each variable

Smals
ICT for society

Modify personal data such that the original can no longer be derived

Related concepts: *differential privacy, k-anonymity*

Privacy is improved by

Removing identifying attributes

Generalizing (e.g. only keep the year from a birthdate)

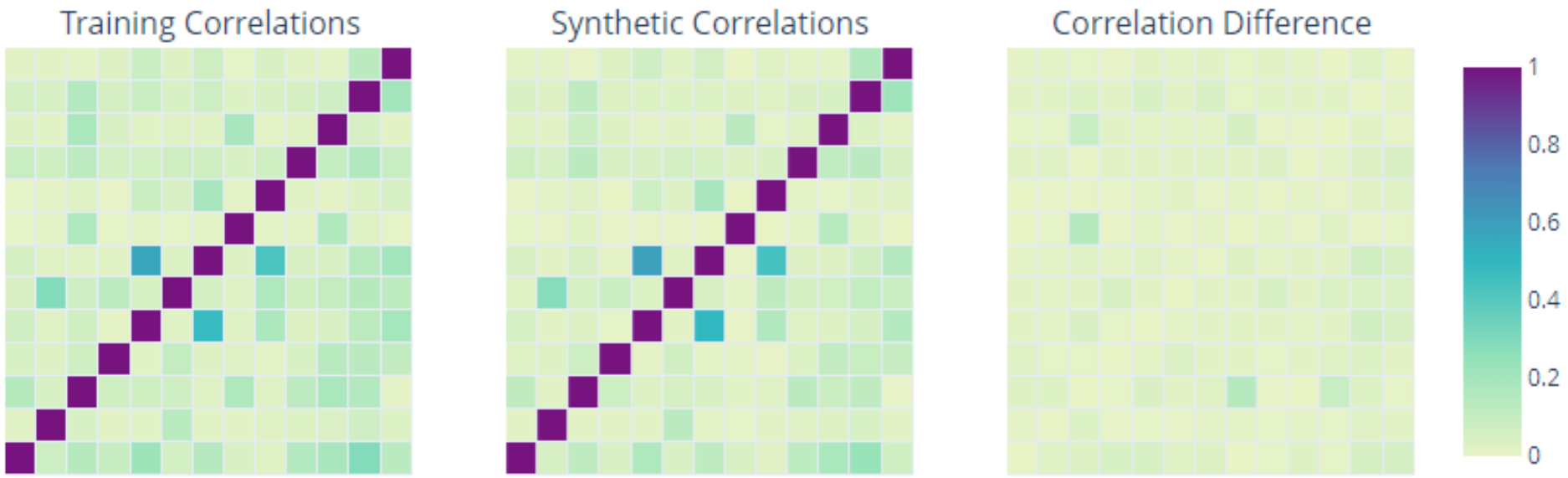
→ inevitable loss of information and/or utility (which synthetic data tries to mitigate)

On the model side: prevent overfitting / memorization of training data

Toolkit for evaluation: ARX

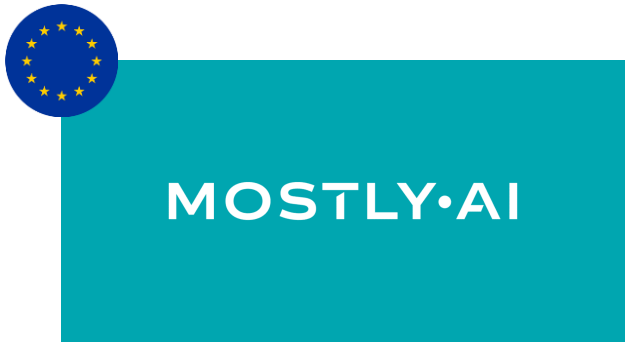
SDMetrics library (under development) provides some toolkit-agnostic evaluation routines

Commercial providers often provide well-illustrated analysis reports - e.g. these cross-correlation graphs from gretel.ai :



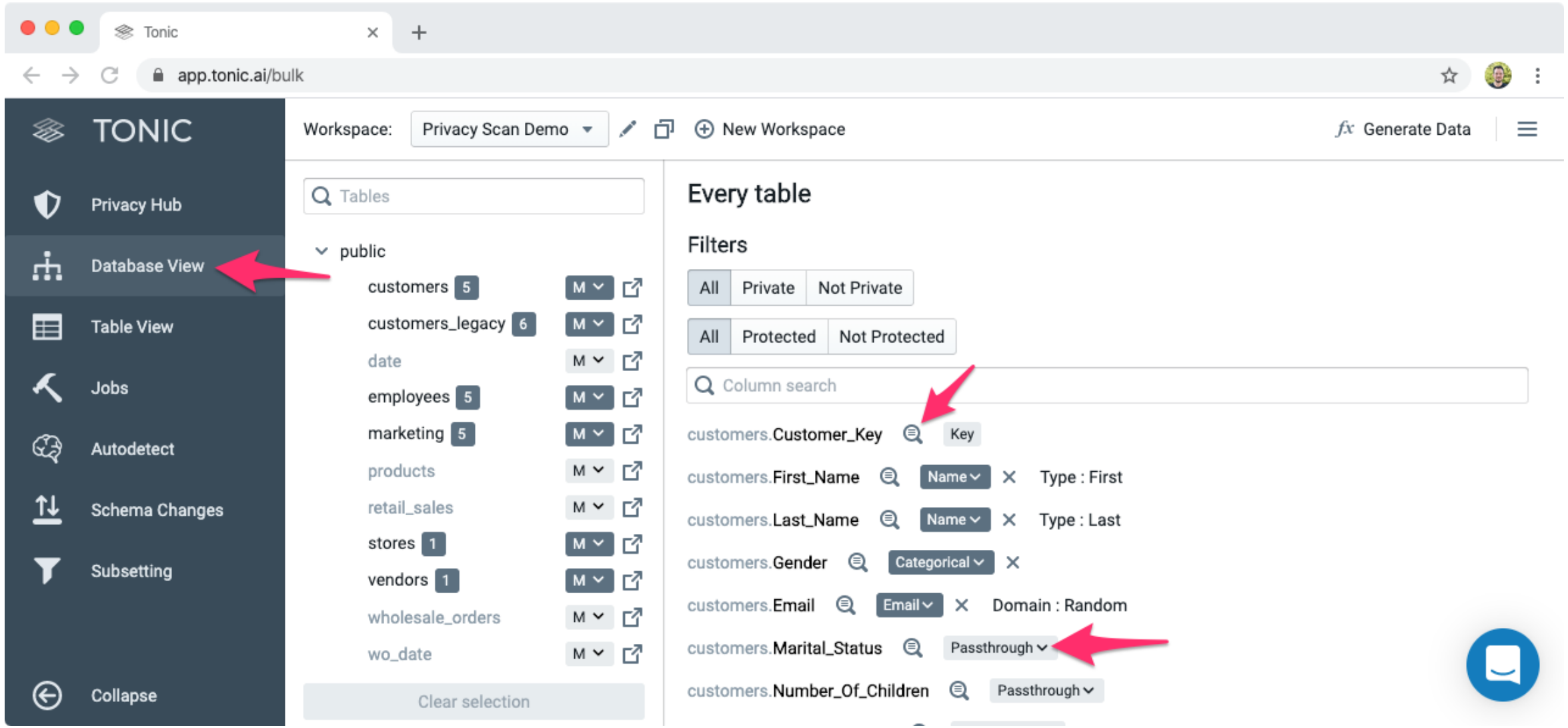
the market

Commercial market for synthetic data tools is booming



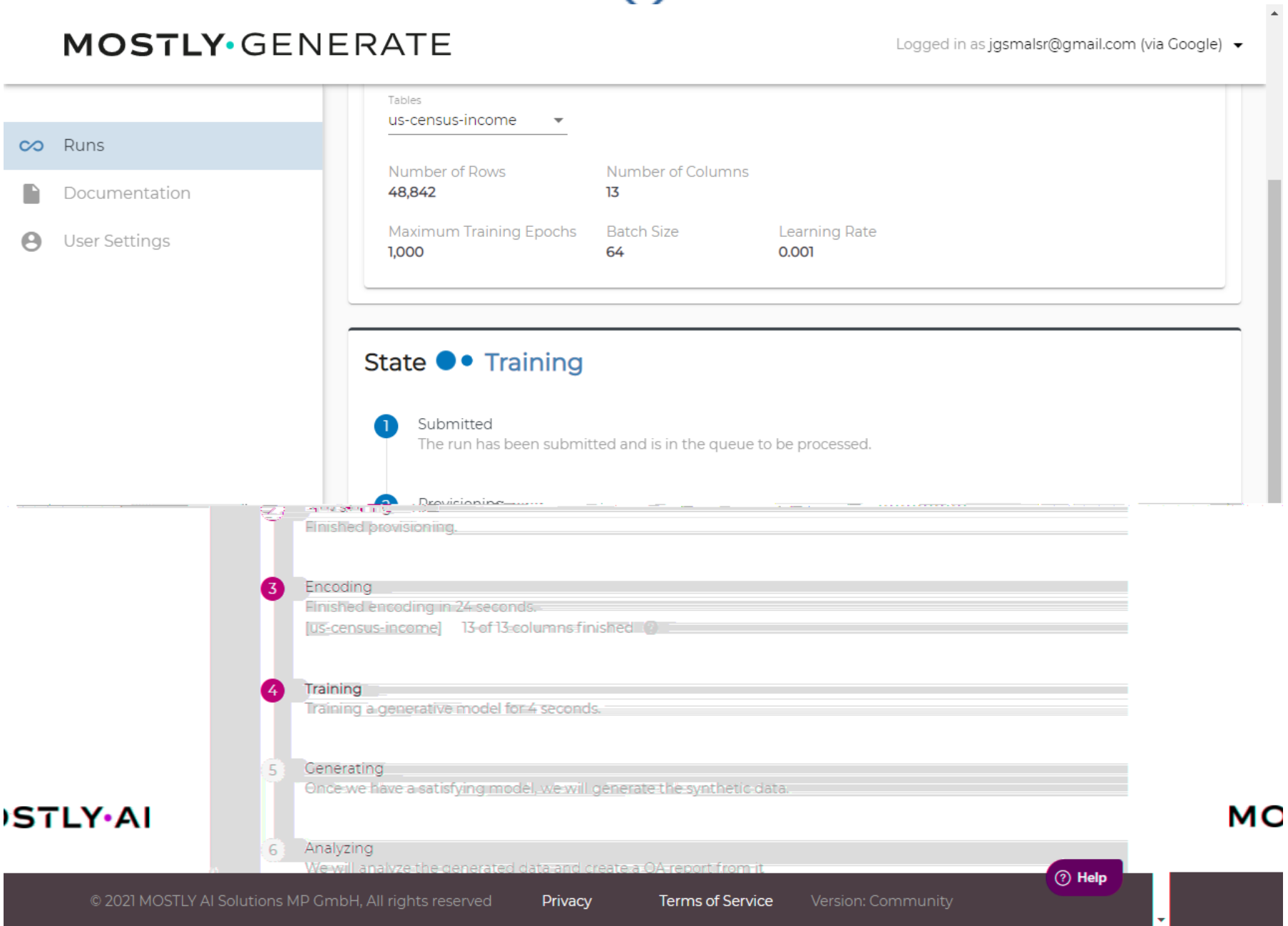
Smals
ICT for society

Configuration: example tonic.ai



Smals
ICT for society

Pipeline: example mostly.ai



Reports: example Gretel.ai

Review

Results

Generated 5,000 records

```
1   17:24:32  Preparing privacy filters
2   17:24:35  Loaded 2 privacy filters
3   17:24:35  Starting privacy filtering
4   17:24:36  Privacy filtering removed 399 records, generating replacement records — filtered_outliers 17, filtered_si
5   17:25:54  Privacy filtering removed 38 records, generating replacement records — filtered_outliers 0, filtered_simi
6   17:26:04  Privacy filtering removed 1 records, generating replacement records — filtered_outliers 0, filtered_simil
7   17:26:06  Privacy filtering complete
8   17:26:06  Saving model archive
9   17:26:09  Creating synthetic quality report
10  17:26:21  Uploading artifacts to Gretel Cloud
11  17:26:22  Model creation complete!
```

Synthetic Quality Score

Privacy Protection Level

Data summary statistics

Field Correlation Stability	Deep Structure Stability	Field Distribution Stability
88	76	82

Download Synthetic Report

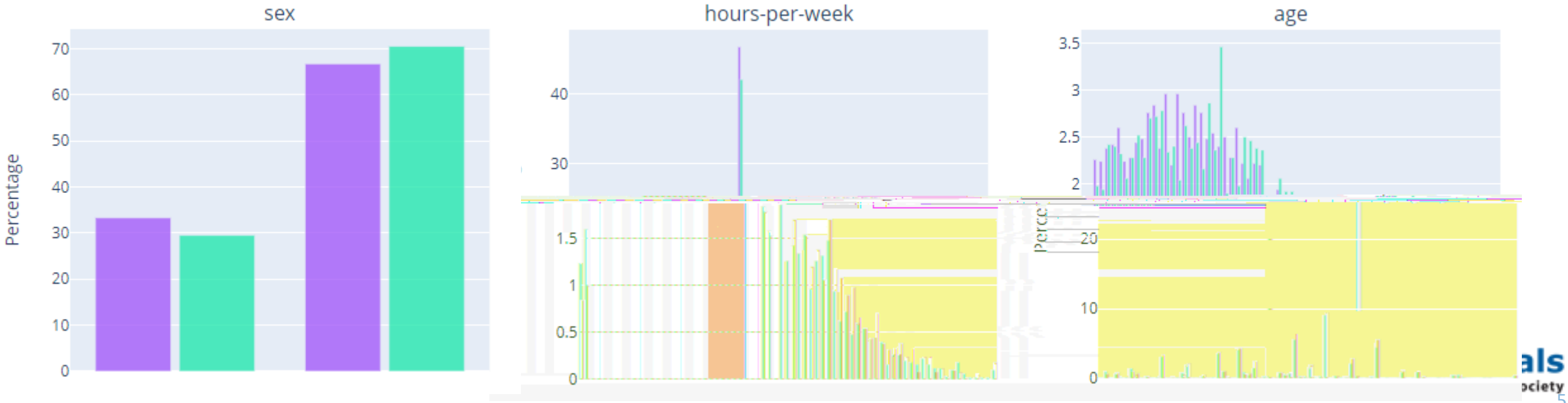
Smals
ICT for society

Better results out-of-the-box

- Better estimation of data properties and subsequent setting of parameters
- Seems more up to speed with developments in deep learning

User-friendly interfaces

Built-in reports with clean graphics



[DOI 10.2760/50072) July 2022]

using open source tools are relatively powerful but only for flat tables, with limited number of constraints, low cardinality categorical variables and continuous, without hard breaks ,

available research and open source solutions by a huge margin at the time of writing ,

may expect competitive open source

,

And now what?

The upsides

Synthetic data, when quality-checked and carefully crafted, is free of privacy and other regulatory issues.

The elimination of bureaucracy associated with sensitive data access, enables more flexibility: put synthetic data in the cloud, make it available as Open Data (democratization), …

Synthetic data can supply digital twins or test environments with plenty of data, and facilitate prototyping.

The field is fast evolving while also steadily maturing. Multiple vendors already offer qualitative solutions.

Inspired from Hradec et. al. : “Multipurpose synthetic population for policy applications” , JRC, 2022, DOI 10.2760/50072

Smals ICT for society

Current challenges

Inflated expectations: a synthetic dataset still differs from the original, and is therefore not always useful for every usecase.

Synthetic data should not be taken at face value. User discretion is advised when interpreting results based on a synthetic dataset.

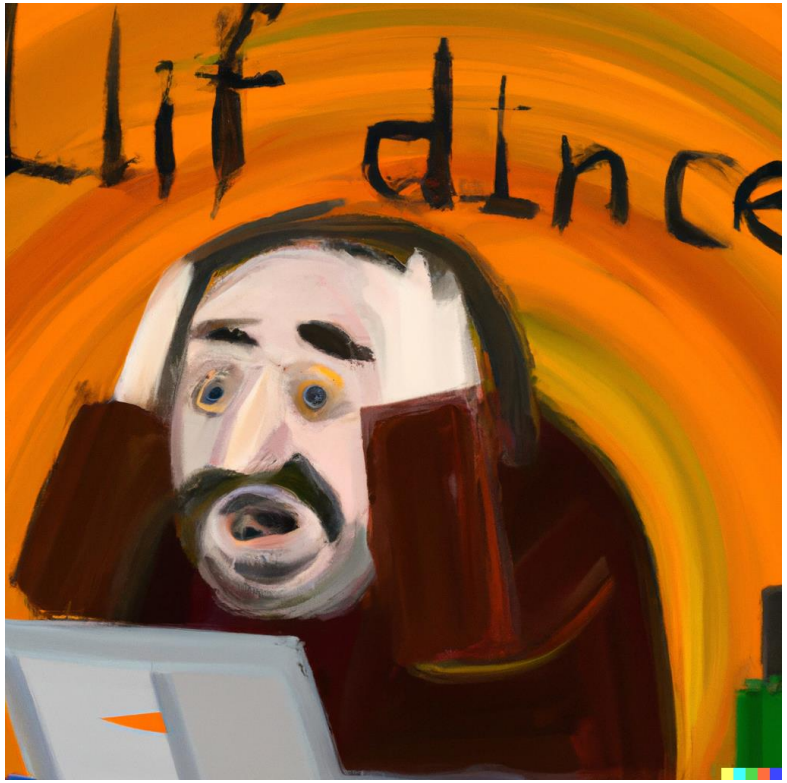
Qualitative synthesis remains challenging in some common cases:

- For hierarchical or very complex data
- For small datasets, datasets with many columns, or with many unique values

Creating good synthetic data still requires expertise, domain knowledge, careful verification and validation, and a good grasp of statistics.

Inspired from Hradec et. al. : “Multipurpose synthetic population for policy applications” , JRC, 2022, DOI 10.2760/50072

Smals ICT for society



Papers on diffusion models for tabular text are starting to appear:

TabDDPM: Modelling Tabular Data with Diffusion Models

30 Sep 2022 · Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, Artem Babenko · Edit social preview

[Source: paperswithcode.com]

Smals
ICT for society

Questions? Shoot!

(or drop by the Smals booth in the centre aisle)

Joachim Ganseman

Smals Research

www.smalsresearch.be

Joachim Ganseman

joachim.ganseman@smals.be
www.smalsresearch.be

Smals, ICT for society
02 787 57 11
Fonsnylaan 20 / Avenue Fonsny 20
1060 Brussel / 1060 Bruxelles