SYNTHETIC DATA AT SCALE

for Big Data Privacy

mostly AI

Using Apache Spark's GraphX library for the analysis of functional neuroimaging data

Roland Boubela

Center for Medical Physics and Biomedical Engineering
Medical University of Vienna

8. Mar. 2016

Curriculum Vitae

- 2007 - BSc in Data Engineering & Statistics at **TU Wien**
- 2010 - MSc in Statistics, Technical Mathematics at **TU Wien**
- 2015 - PhD in Medical Physics at **MedUni Wien**
- 2016 - 2017 Data Scientist at **Hutchison Drei Austria**
- 2017 - Founder & CTO of **Mostly AI**

  - AI Academies

  - Enabling Big Data Privacy via AI-Generated Synthetic Data

mostly AI

Privacy in the Era of Big Data

The Privacy vs. Innovation Clash

PROBLEM

1. **Data privacy** restricts sharing of data and thus **hampers innovation**.

   We demand highest standards for data protection, but also need to collaborate broadly on data in order to develop next-gen digital services and processes.

2. **Pseudonymization** offers **no safety**, while **Full Anonymization** falls short for big data.

   Classic anonymization techniques need to destroy most of the available information to prevent re-identification of individuals (see appendix).

SOLUTION

1. **Synthetic data is anonymous**.

   Synthetic data is not restricted in its usage, and is free to store, to use, to explore, to experiment, to modify, and to share, within and outside of the organization.

2. **Generative AI** allows **highly accurate** synthetic data to be generated at scale.

   Academic advances on deep generative neural networks have resulted in highly realistic synthetic images, near indistinguishable from real ones.

Pseudonymization

Pseudonymization offers **no safety**!

Pseudonymization

Pseudonymization offers **no safety**, while
Classic Anonymization destroys data **value**!

Pseudonymization

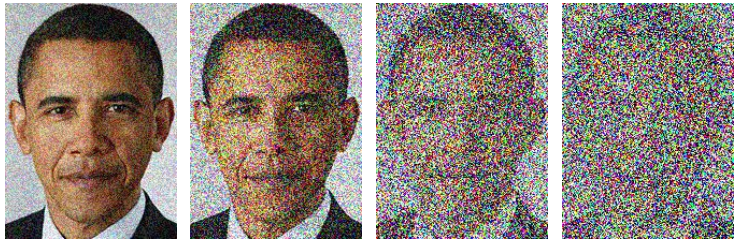
Pseudonymization offers **no safety**, while
Classic Anonymization destroys data **value**!

Classic Anonymization Fails

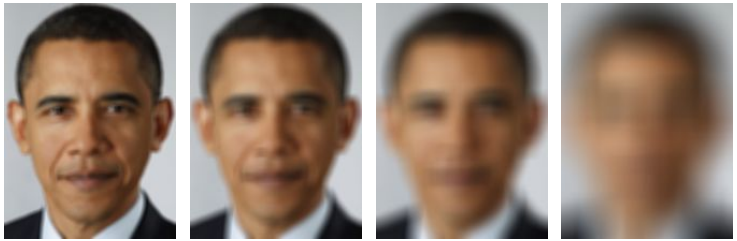
for High-Dimensional, Highly Correlated Data

masking - obfuscation

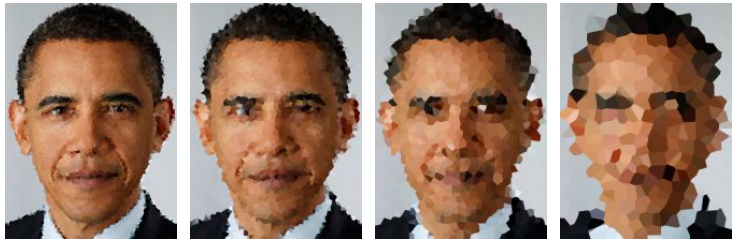
adding noise - perturbations



generalization



generalization

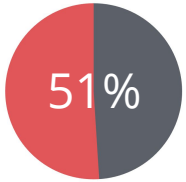


"We conjecture that the amount of perturbation that must be applied to the data to defeat our algorithm will ***completely destroy their utility*** *[..] Sanitization techniques from the k-anonymity literature such as generalization and suppression do not provide meaningful privacy guarantees, and in any case fail on high-dimensional data."*

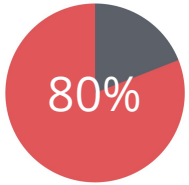
"Robust De-anonymization of Large Sparse Datasets" (Narayanan, 2008)

Classic Anonymization Fails for Big Data

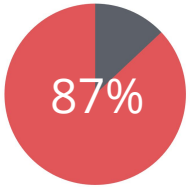
i.e. for High-Dimensional, Highly Correlated Data



of **mobile phone owners** are re-identified simply by 2 antenna signals, even when coarsened to the hour of the day



of **credit card owners** are re-identified by 3 transactions, even when only merchant and the date of transaction is revealed



of **all people** are re-identified, merely by their date-of-birth, their gender and their ZIP code of residence

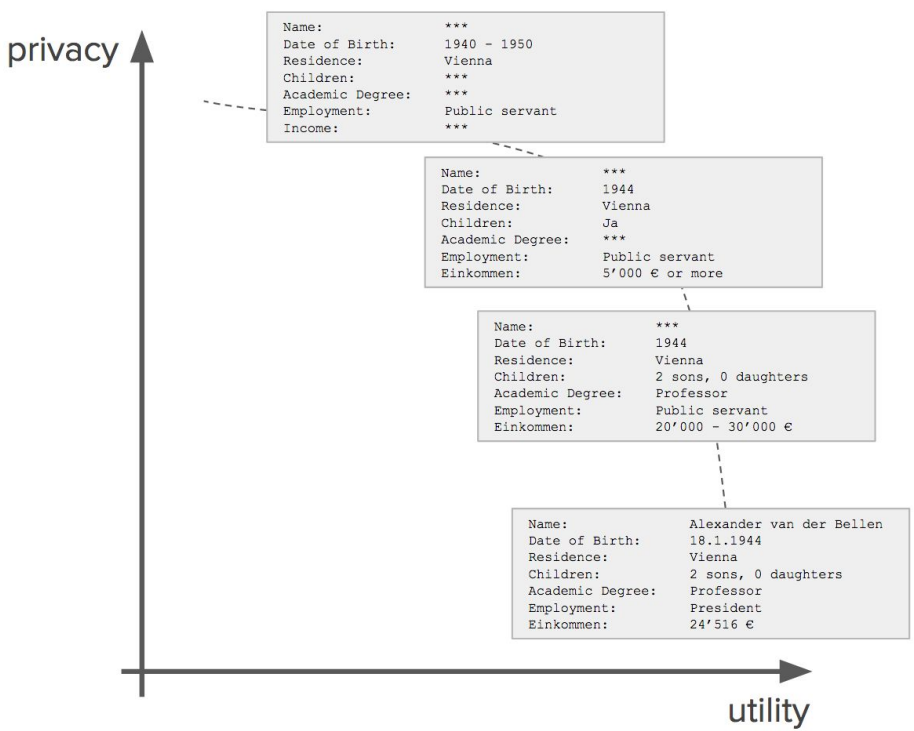
Anonymize Me!

Vorname:	Roswitha
Nachname:	Mayrhofer
Geb. Datum:	13.4.1963
Wohnort:	Hochberg 12, 5532 Filzmoos
Familienstand:	Verheiratet
Kinder:	0 Söhne, 1 Tochter
Akadem. Grad:	-
Beruf:	Handelsangestellte
Einkommen:	1’853 €
Religion:	röm.kath.

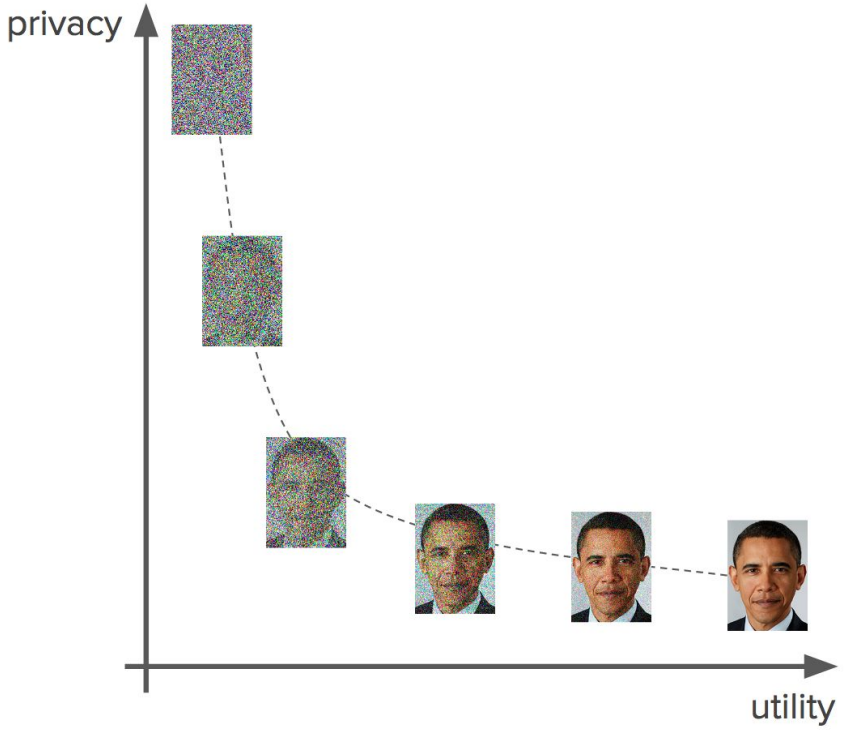
Anonymize Me!

Vorname:	Alexander
Nachname:	van der Bellen
Geb. Datum:	18.1.1944
Wohnort:	Millergasse 2, 1060 Wien
Familienstand:	Verheiratet
Kinder:	2 Söhne, 0 Töchter
Akadem. Grad:	Professor
Beruf:	Bundespräsident
Einkommen:	24’516 €
Religion:	o.B.

The Privacy vs. Utility Trade-Off



small data



big data

So, How to Utilize Big Data in the Era of Privacy?

a data-driven, customer-centric, innovative organization?

Dummy Data?

overly **simplistic and biased**

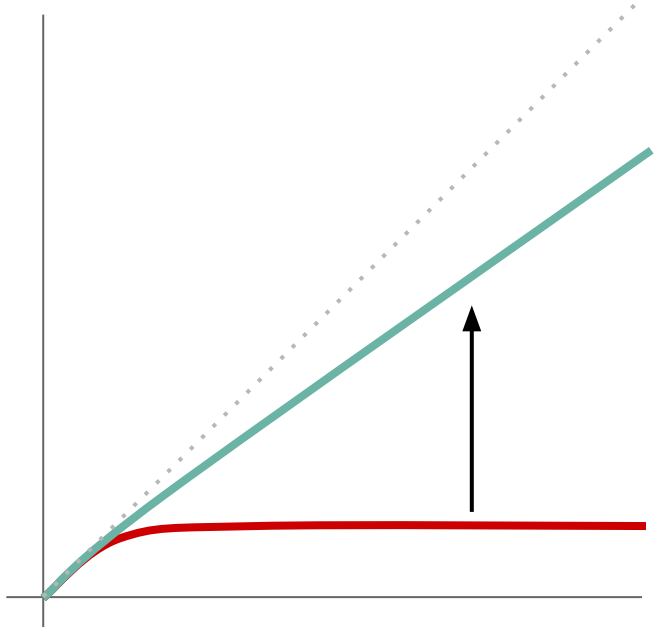
Model-Driven **Synthetic Data**

overly **simplistic and biased**

AI-Generated **Synthetic Data!**

highly realistic & representative, **as-good-as-real**, captures full richness of your actual data

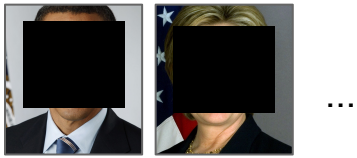
Game Changer for Big Data Anonymization



...

AI-generated Synthetic Data

retains structure and variation, but no 1:1 relationship to actual persons



...

Classic Anonymization

can only share very limited amount of information per person

mostly AI

Generative AI

Machine Learning

Learning by Being Taught

→ **Supervised Learning**

Learning by Observation

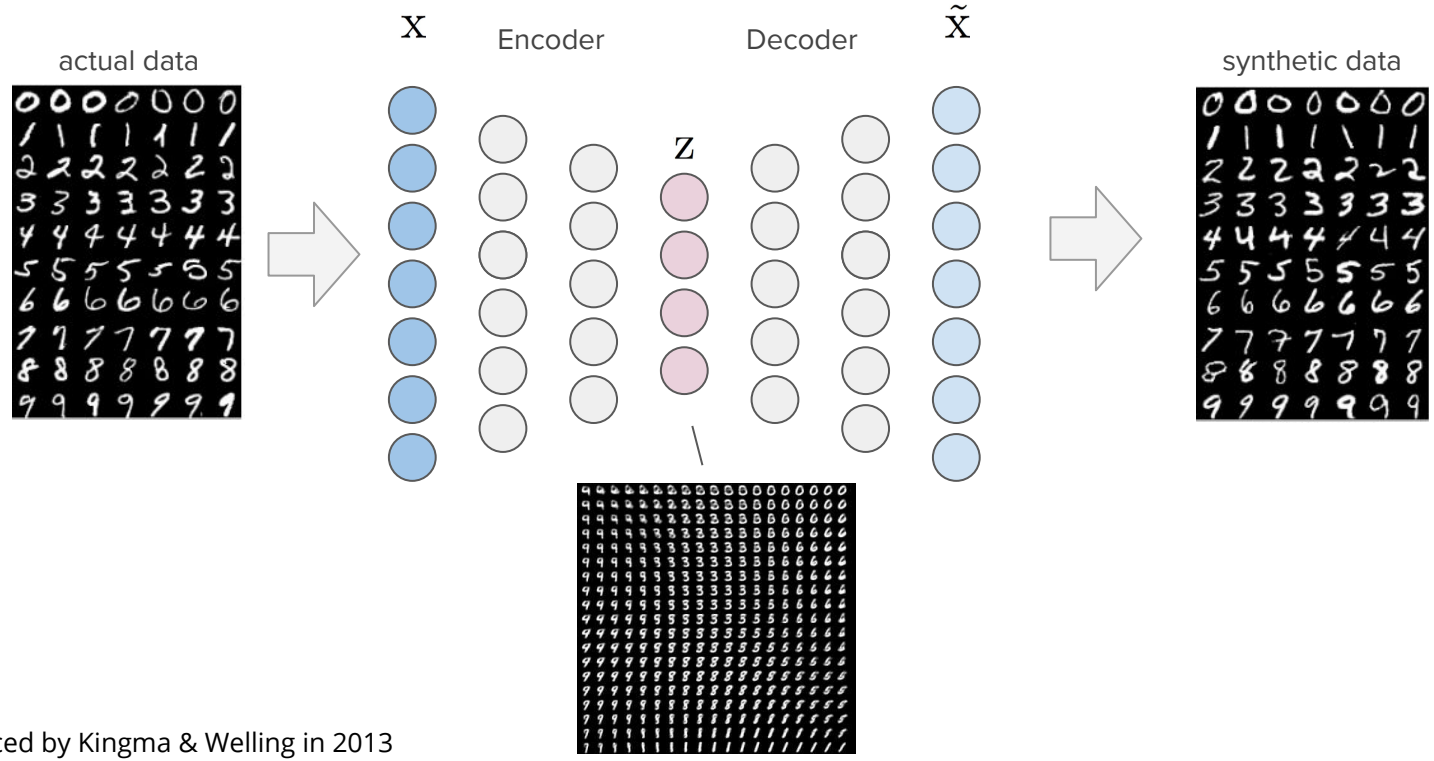
→ **Self-Supervised learning**

Learning by Exploration

→ **Reinforcement Learning**

Generative Deep Models – VAEs

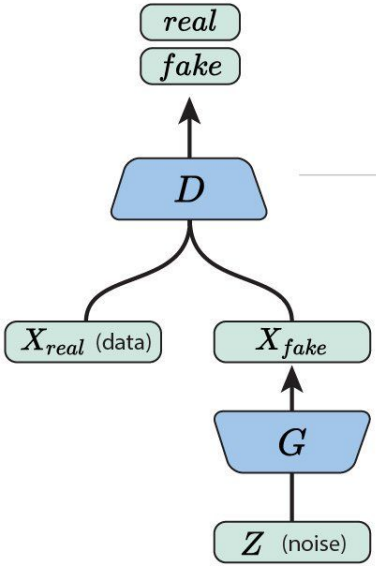
Variational Autoencoders



- VAE introduced by Kingma & Welling in 2013
- 600+ papers published on VAE in 2017

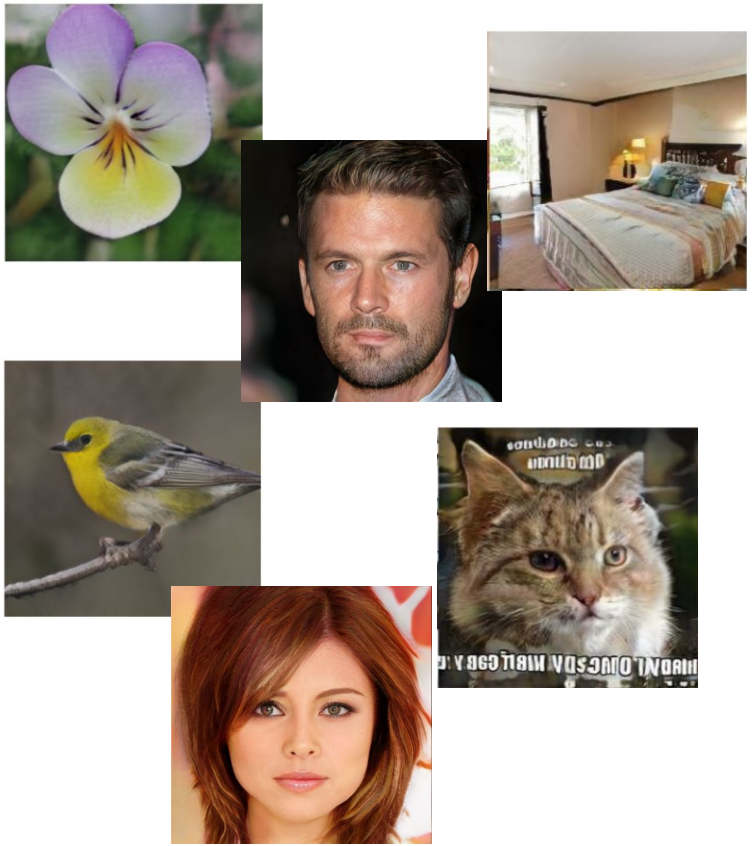
Generative Deep Models – GANs

Generative Adversarial Networks



The **discriminator** tries to distinguish genuine data from forgeries created by the generator.

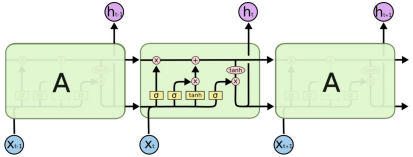
The **generator** turns random noise into immitations of the data, in an attempt to fool the discriminator.



- Introduced by Goodfellow et al. in 2014
- 1500+ papers published on GANs in 2017

Generative Deep Models – ARNs

Autoregressive Neural Networks




```
VIOLA:
Why, Salisbury must find his flesh and thought
That which I am not aps, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair are hand,
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.

KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.
```

Synthetic Shakespeare

```c
/*
 * If this error is set, we will need anything right after that BSD.
 */
static void action_new_function(struct s_stat_info *wb)
{
  unsigned long flags;
  int lel_idx_bit = e->edd, *sys & ~((unsigned long) *FIRST_COMPAT);
  buf[0] = 0xFFFFFFFF & (bit << 4);
  min(inc, slist->bytes);
  printk(KERN_WARNING "Memory allocated %02x/%02x, "
    "original MLL instead\n"),
    min(min(multi_run - s->len, max) * num_data_in),
    frame_pos, sz + first_seg);
  div_u64_w(val, inb_p);
  spin_unlock(&disk->queue_lock);
  mutex_unlock(&s->sock->mutex);
  mutex_unlock(&func->mutex);
  return disassemble(info->pending_bh);
}
```

Synthetic Linux Source Code

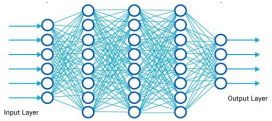
The **Synthetic Data Engine** by Mostly AI

AI-generated **synthetic populations** of customers and their behavior

NAME	ZIP	AGE	GENDER	ITEM	EUR	DATE	TIME
Mary	1220	25y	female	Book	12€	4/2/19	8:12
John	2320	72y	male	Pizza	34€	4/2/19	18:12
...							
Kevin	8329	18y	male	Swim	6€	4/4/19	10:02

actual

mostly AI

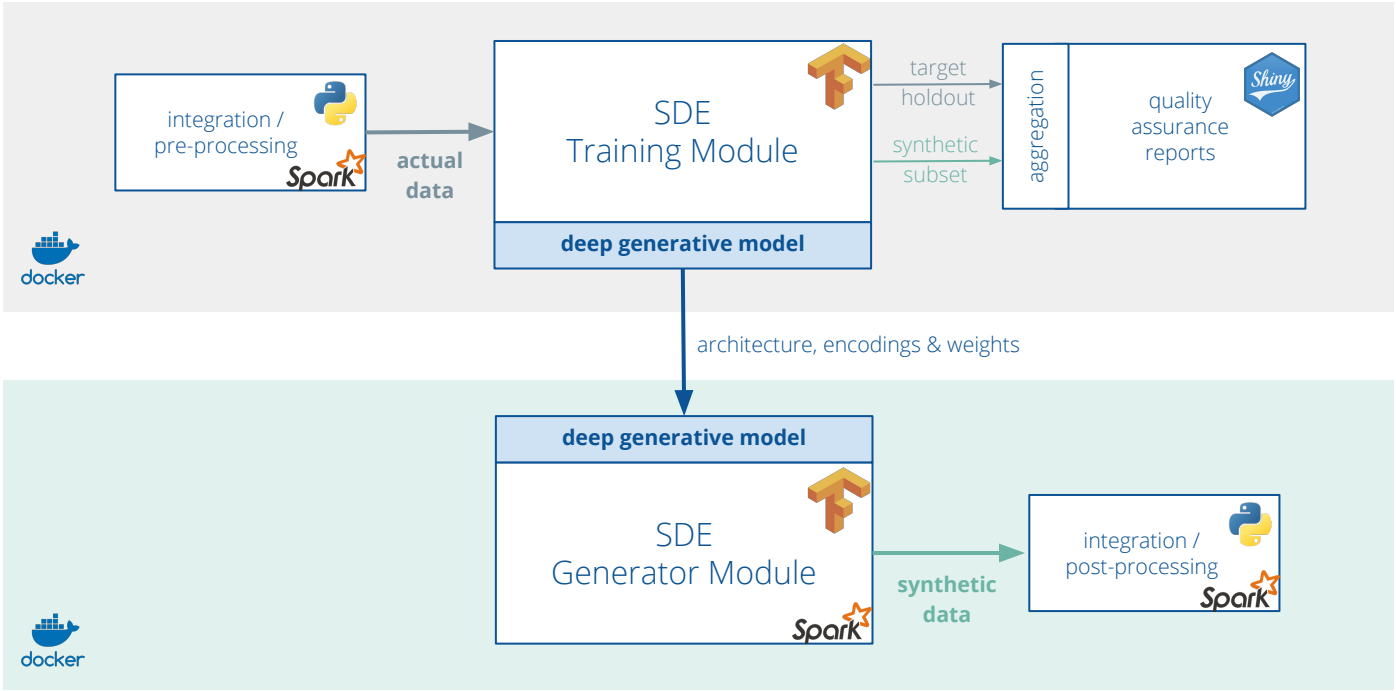


NAME	ZIP	AGE	GENDER	ITEM	EUR	DATE	TIME
Bob	3729	82y	male	Beer	6€	4/2/19	15:32
Sue	8022	24y	female	Sushi	12€	4/2/19	21:32
...							
Kim	3923	29y	female	Amazon	36€	4/4/19	12:32

synthetic

Under the Hood

The Synthetic Data Engine by Mostly AI



- on-premise or cloud
- GPU / cluster
- CLI / REST API

- anywhere
- anytime
- CLI / REST API

Demos

User Interface

mostly train

```
$ mostly train

usage: mostly train [-h] [-c [config]] [-t [tag]] [input_path]

positional arguments:
  input_path   path to directory with csv files

optional arguments:
  -h, --help   show this help message and exit
  -c [config]  path to data configuration file
  -t [tag]     name of the training run
```

User Interface

mostly generate

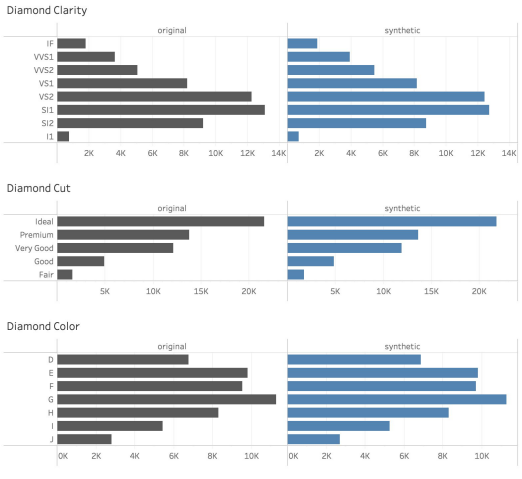
```
$ mostly generate

usage: mostly generate [-h] [-n [count]] [-d [context_data] [[context_data]
                       ...]] [-t [tag]] [-o [output_path]]

optional arguments:
  -h, --help            show this help message and exit
  -n [count]            number of records to generate
  -d [context_data] [[context_data] ...]
                        list of context data table paths
```

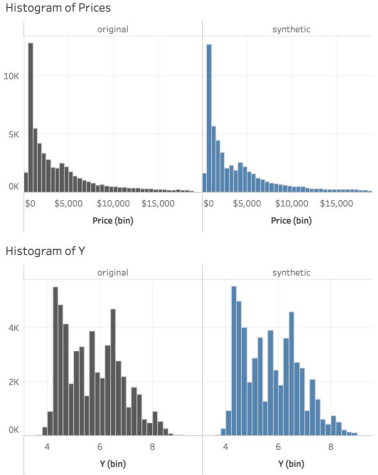
mostly AI

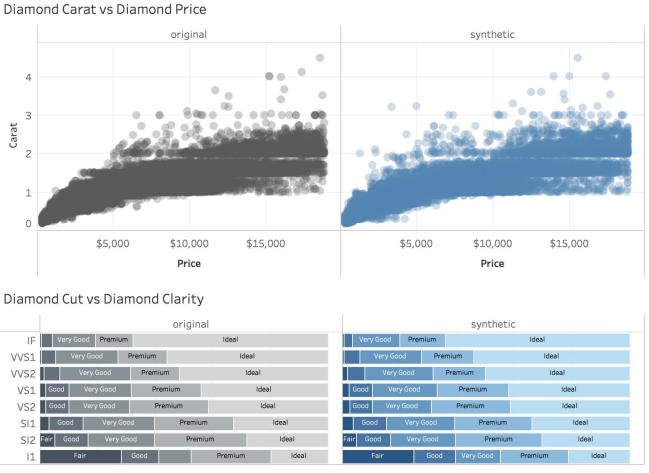
Synthetic Data Diamonds

	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55.0	326
2	0.21	Premium	E	SI1	59.8	61.0	326
3	0.23	Good	E	VS1	56.9	65.0	327
4	0.29	Premium	I	VS2	62.4	58.0	334
5	0.31	Good	J	SI2	63.3	58.0	335
6	0.24	Very Good	J	VVS2	62.8	57.0	336
7	0.24	Very Good	I	VVS1	62.3	57.0	336
8	0.26	Very Good	H	SI1	61.9	55.0	337

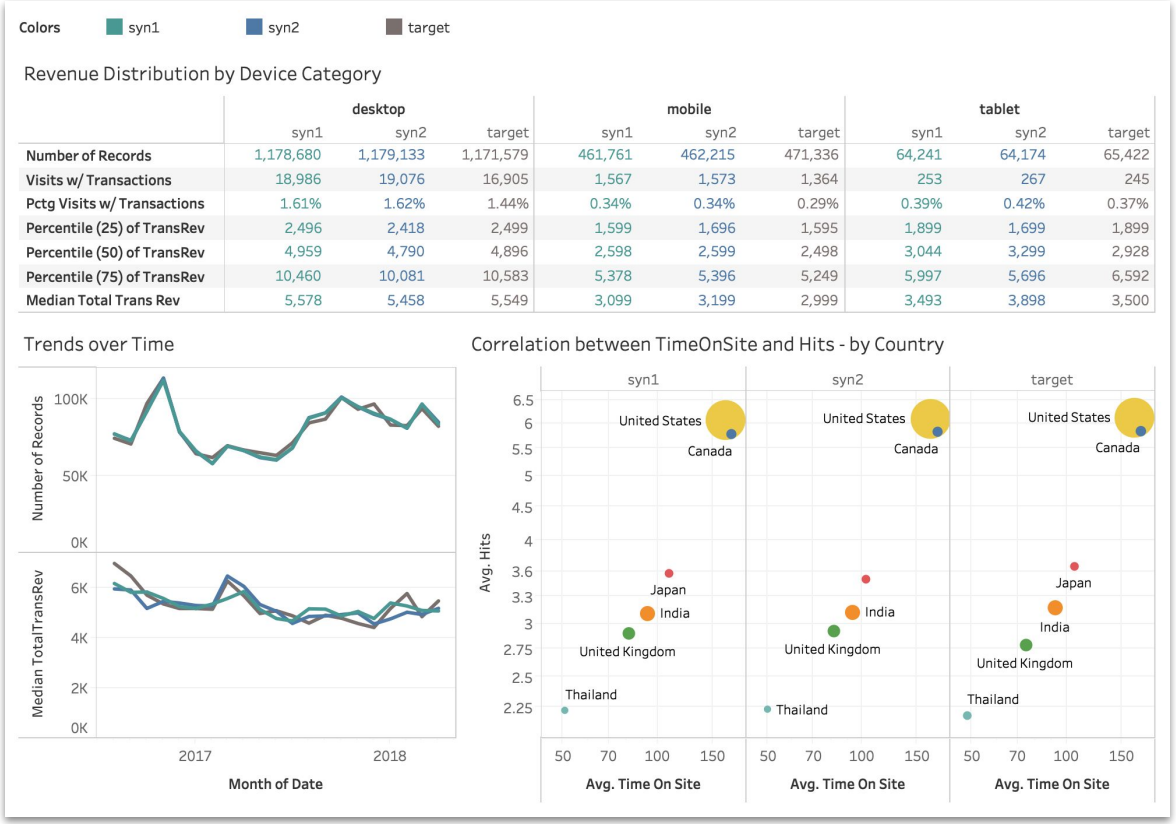
	carat	cut	color	clarity	depth	table	price
1	0.32	Premium	I	SI1	61.5	58.0	508
2	2.07	Ideal	H	SI2	60.8	56.0	12920
3	0.31	Good	E	SI1	63.8	58.0	537
4	1.05	Very Good	G	VVS2	62.9	57.0	8173
5	0.45	Premium	J	VS1	60.7	60.0	898
6	0.90	Premium	H	VVS1	61.0	58.0	4931
7	1.10	Ideal	E	IF	62.9	55.0	12508
8	0.75	Ideal	E	VS2	61.1	56.0	3169







Synthetic eCommerce Visitors



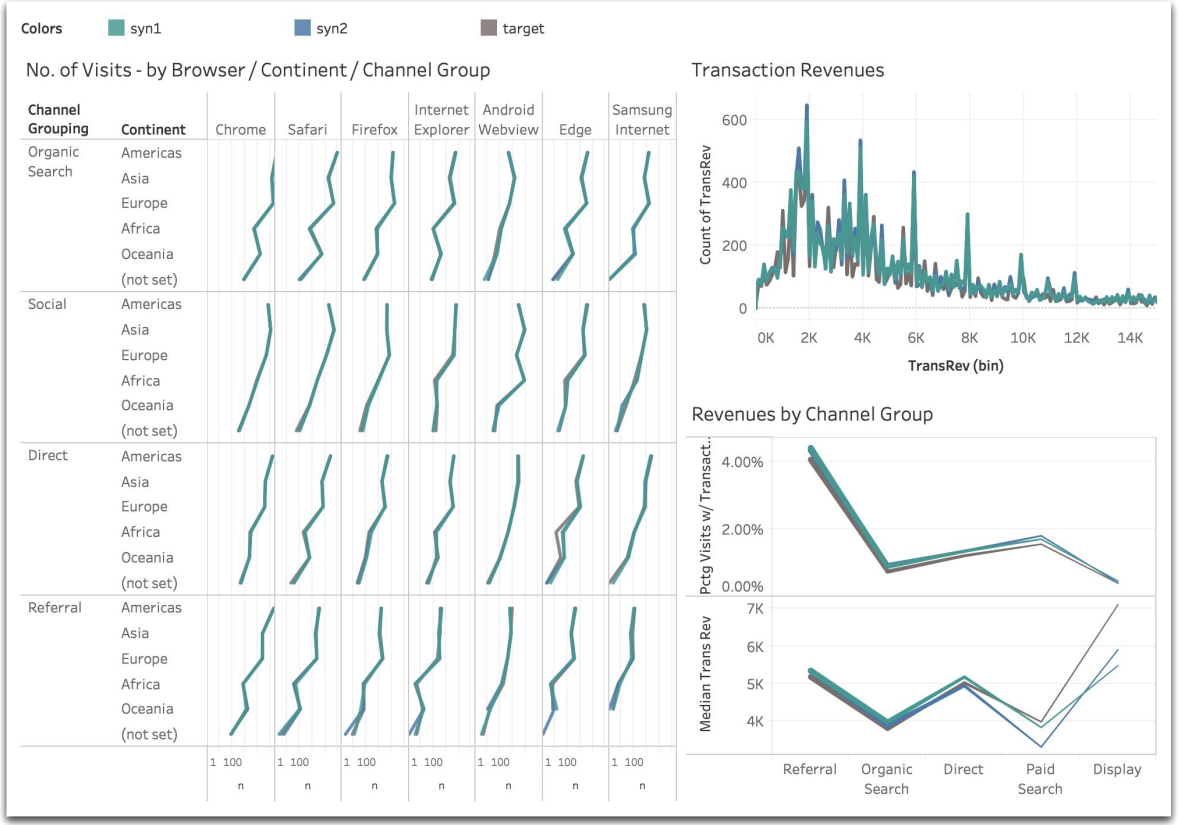
https://www.kaggle.com/c/ga-customer-revenue-prediction

- 1.3m visitors with 1.7m visits
- 40 attributes captured per visit
  - date, time
  - geography
  - browser info
  - traffic source
  - ...
- only 1.1% of visits have transactions
- transaction revenues are strongly right-skewed (~31)

2 synthetic versions of the target data are being generated via the Synthetic Data Engine, and then compared to each other.

→ **statistics match perfectly**

Synthetic eCommerce Visitors



https://www.kaggle.com/c/ga-customer-revenue-prediction

- 1.3m visitors with 1.7m visits
- 40 attributes captured per visit
  - date, time
  - geography
  - browser info
  - traffic source
  - ...
- only 1.1% of visits have transactions
- transaction revenues are strongly right-skewed (~31)

2 synthetic versions of the target data are being generated via the Synthetic Data Engine, and then compared to each other.

→ **statistics match perfectly**

Synthetic Credit Card Fraud

```r
library(data.table)
library(ranger)
library(pROC)

val <- fread('kaggle-fraud/data/cc-test.csv')
tgt <- fread('kaggle-fraud/data/cc-train.csv')
syn <- fread('kaggle-fraud/data/fraud-gen.csv')

dim(tgt)
# [1] 142403     31
mean(tgt$Class)
# [1] 0.001727492

# train random forest
m_tgt <- ranger(Class~., data = tgt)
m_syn <- ranger(Class~., data = syn)

auc(roc(as.factor(val[, Class]), predict(m_tgt, val)$predictions))
# 0.9562
auc(roc(as.factor(val[, Class]), predict(m_syn, val)$predictions))
# 0.9486

tgt[, .N, by = Class] # 246
syn[, .N, by = Class] # 265
```

https://www.kaggle.com/mlg-ulb/creditcardfraud

- 142k records
- 30 attributes
- 0,17% of cases are labelled fraud

Synthetic data of same size and structure as the original dataset is being generated via the Synthetic Data Engine. Subsequently a sophisticated machine learning algorithm (Random Forest) is trained on the original as well as on the synthetic version, and then evaluated on an actual holdout dataset in terms of accuracy. As can be seen, the accuracy of the two model is nearly the same.

→ **synthetic data can be used for advanced ML algos**
→ **synthetic data also retains weak signals in the data**