Synthetic Customers

for Big Data Privacy

mostly AI

The Privacy vs. Innovation Clash

PROBLEM

1 **Data Privacy 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 **Classic Anonymization Fails 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.

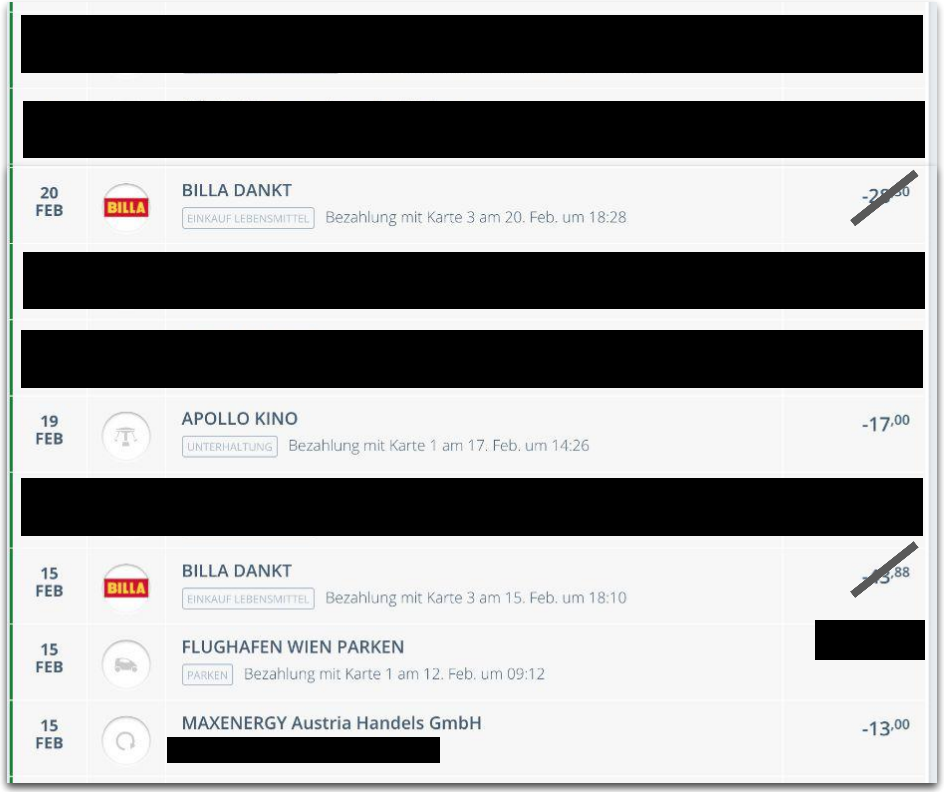
2 **Generative AI → As-Good-As-Real Synthetic Data generated at scale**

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

mostly AI

The Problem Anonymization Fails for Big Data

User #3dcf29717a9f9b39



[redacted]

[redacted]

20 FEB	BILLA	BILLA DANKT EINKAUF LEBENSMITTEL Bezahlung mit Karte 3 am 20. Feb. um 18:28	~~-2[illegible],30~~
[redacted]			
[redacted]			
19 FEB		APOLLO KINO UNTERHALTUNG Bezahlung mit Karte 1 am 17. Feb. um 14:26	-17,00
[redacted]			
15 FEB	BILLA	BILLA DANKT EINKAUF LEBENSMITTEL Bezahlung mit Karte 3 am 15. Feb. um 18:10	~~-[illegible]3,88~~
15 FEB		FLUGHAFEN WIEN PARKEN PARKEN Bezahlung mit Karte 1 am 12. Feb. um 09:12	[redacted]
15 FEB		MAXENERGY Austria Handels GmbH [redacted]	-13,00

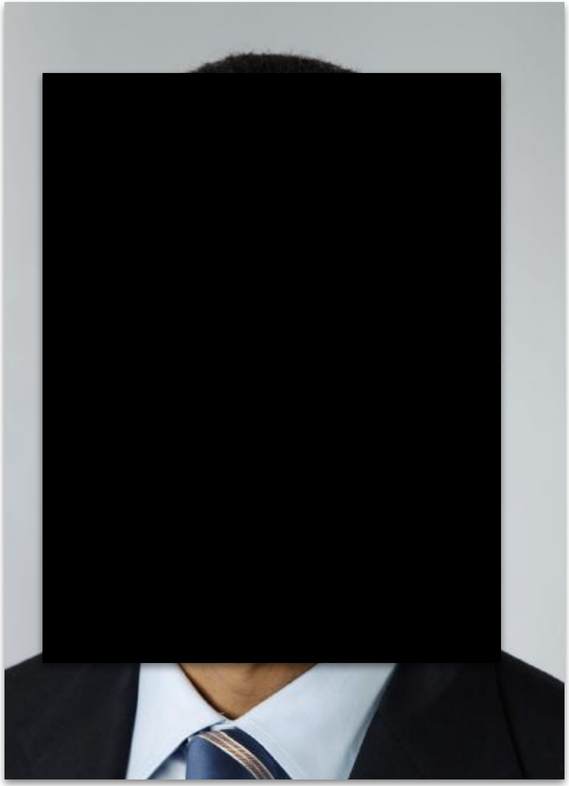
20-30€

40-50€

User #71f7c3014d2ced27

mostly AI

The Problem Anonymization Fails for Big Data



User #3dcf29717a9f9b39

0-30€

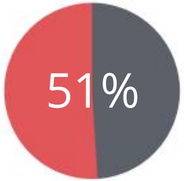
0-30€

30-60€

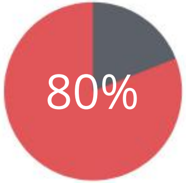
User #71f7c3014d2ced27

mostly AI

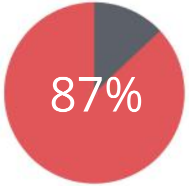
The Problem Classic Anonymization Fails for Big Data



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

WIRED

AOL: "This was a screw up"

FAST COMPANY

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

Sticky data: Why even 'anonymized' information can still identify you

The New York Times

Researchers spotlight the lie of 'anonymous' data

SCIENCE & TECHNOLOGY

You're not so anonymous

Sorry, your data can still be identified even if it's anonymized

Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data

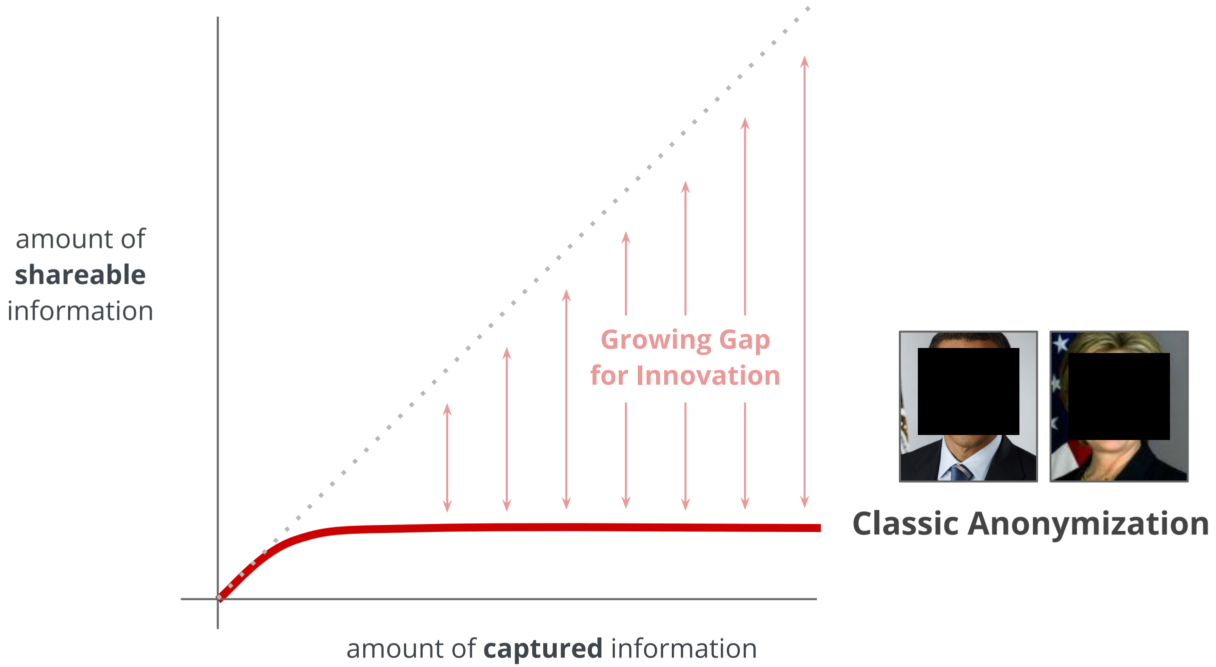
REGULATION

There's No Such Thing as Anonymous Data

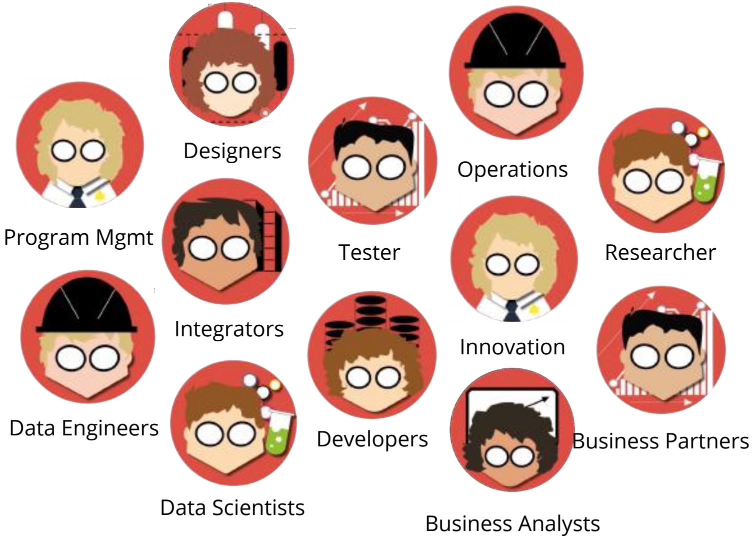
Harvard Business Review

Sources: [1], [2], [3], [4], [5], [6]

No Solution for Big Data Anonymization Exists



The Consequence



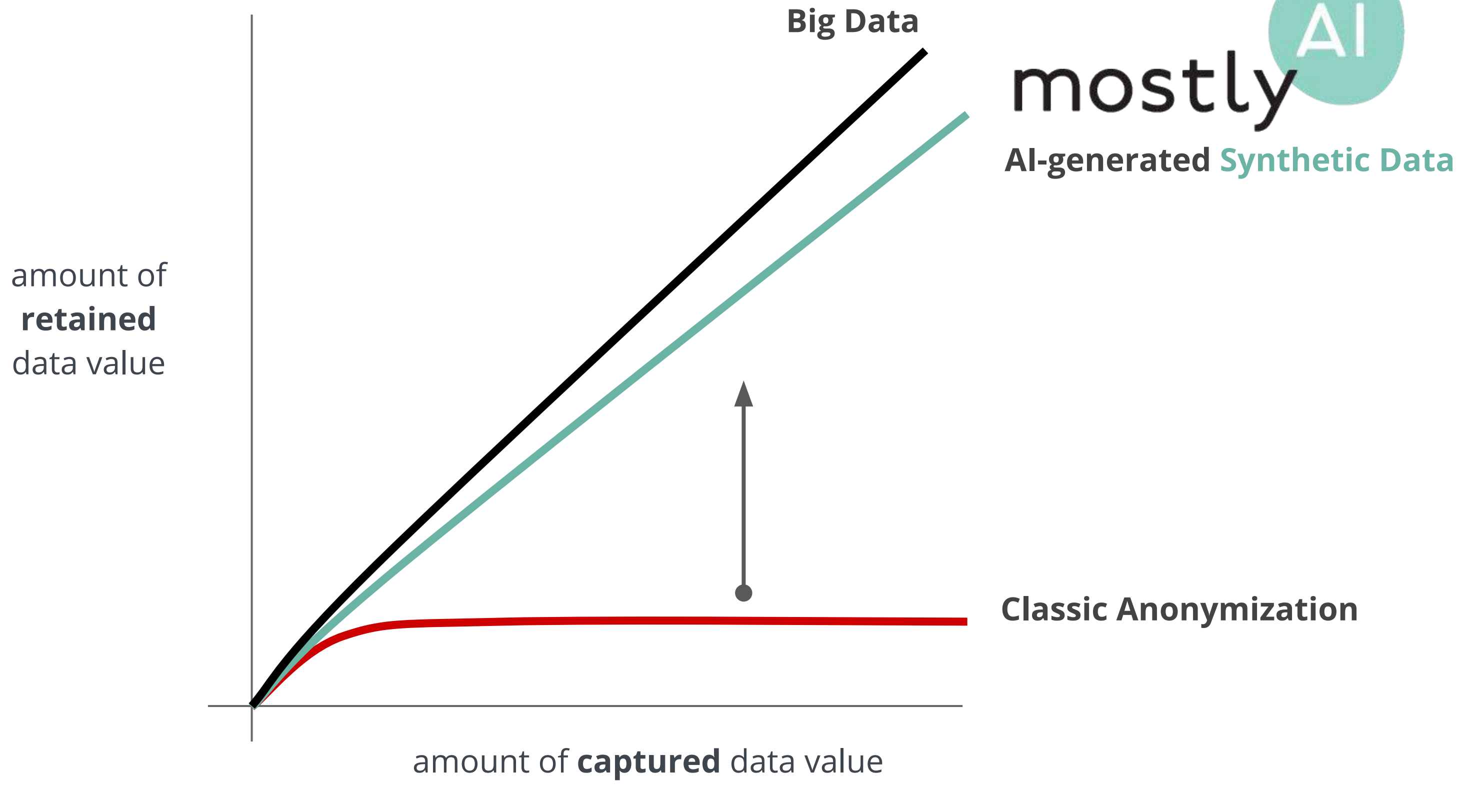
→ How to be **data-driven & customer-centric** in the era of data privacy?

The Solution AI-Generated Synthetic Data

AI-generated synthetic faces

(as demonstrated by Nvidia)

Game Changer for Big Data Anonymization



Synthetic Data allows you to <u>do both:</u>

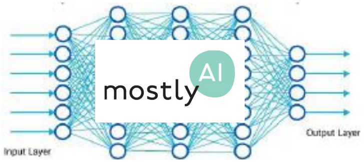
1. Retain Big Data's **Value & Information**
2. **Full** Anonymization

mostly AI

Our Solution The Synthetic Data Engine by Mostly AI

NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Mary	25y	female	Book	12€	4/2/19	8:12
John	72y	male	Pizza	34€	4/2/19	18:12
...						
Bill	18y	male	Swim	6€	4/4/19	10:02
Bill	18y	male	Shoes	123€	4/4/19	12:32

actual, privacy-sensitive data



NAME	AGE	GENDER	ITEM	EUR	DATE	TIME
Kim	29y	female	Amazon	236€	4/4/19	12:32
Kim	29y	female	Zalando	36€	4/4/19	18:58
...						
Brian	82y	male	Beer	6€	4/2/19	21:32
Sue	24y	female	Sushi	12€	4/2/19	21:32

synthetic, statistical representative data

Finance Retail Insurance Health Public

anonymous granular-level data

retains statistical value

unrestricted big data utilization

Our Solution Flexible, Scalable & Easy-To-Use

Command Line Interface

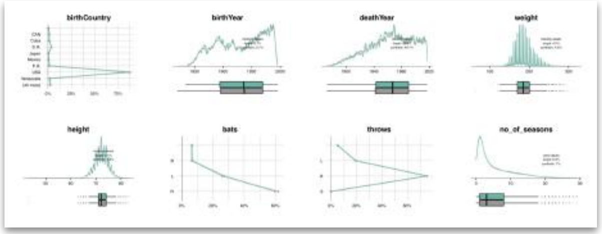
```
> mostly config /path/to/data

> mostly train /path/to/data

> mostly generate -n 1000000
```

Graphical User Interface

Quality Assurance Reports



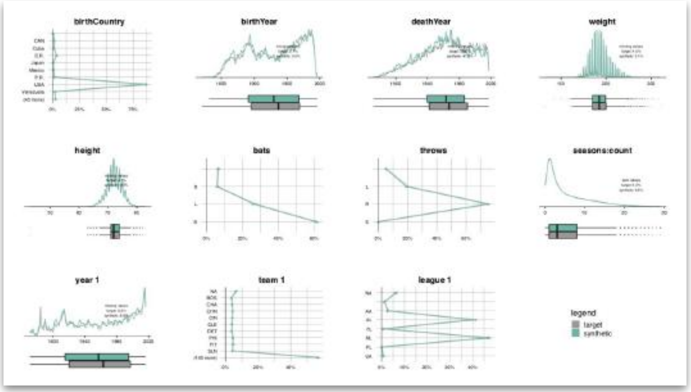
easy setup on-premise or cloud

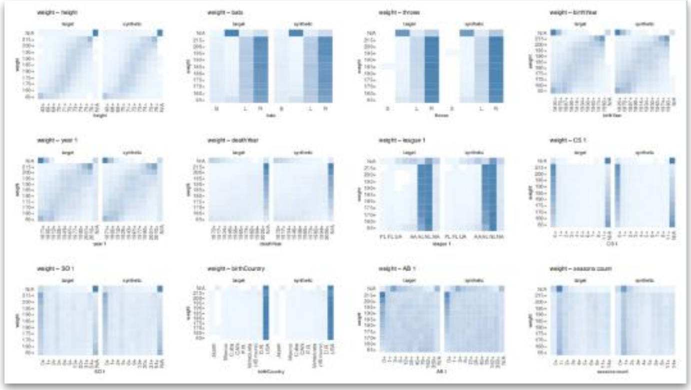
scales to millions of users

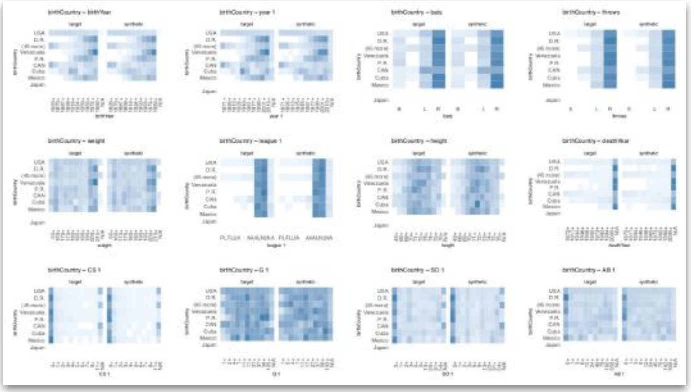
unlimited amount of synthetic data

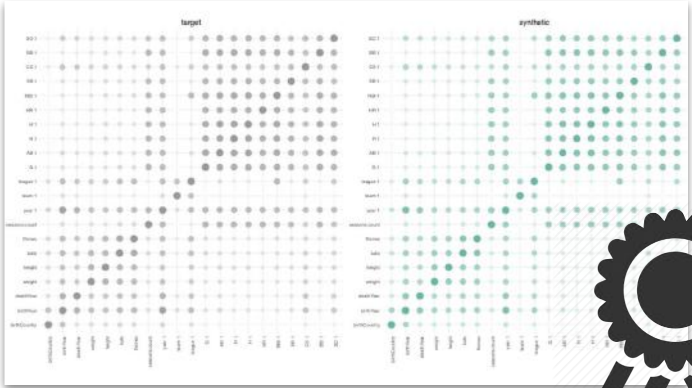
TRUSTED PRIVACY
EU data protection
ePrivacyseal

Our Solution Unparalleled Accuracy & Quality









Use Cases for Synthetic Data

for **Internal Data Sharing**

- AI Training & Analytics
- Testing & Development
- UX & Customer Centricity
- Cloud Migration
- Breaking Down Data Silos
- Advanced Predictive Analytics

for **External Data Sharing**

- Open Innovation
- Startup Collaborations
- Research Collaborations
- Vendor Validation
- Sandboxes

for **External Data Monetization**

- Strategic Partnerships
- Data Marketplaces
- Data Resellers
- Market Research Intel

Mostly AI's Synthetic Data Engine...

1.

Faster

2.

Cheaper

3.

Less Risk

...AI & Big Data Innovation!

mostly AI

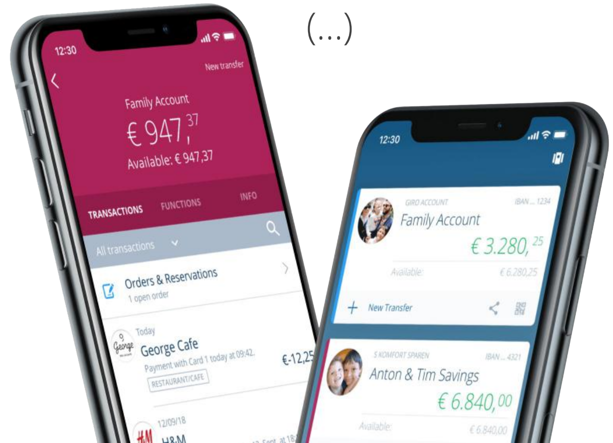
Customer Success Story

Product Development in Finance Industry

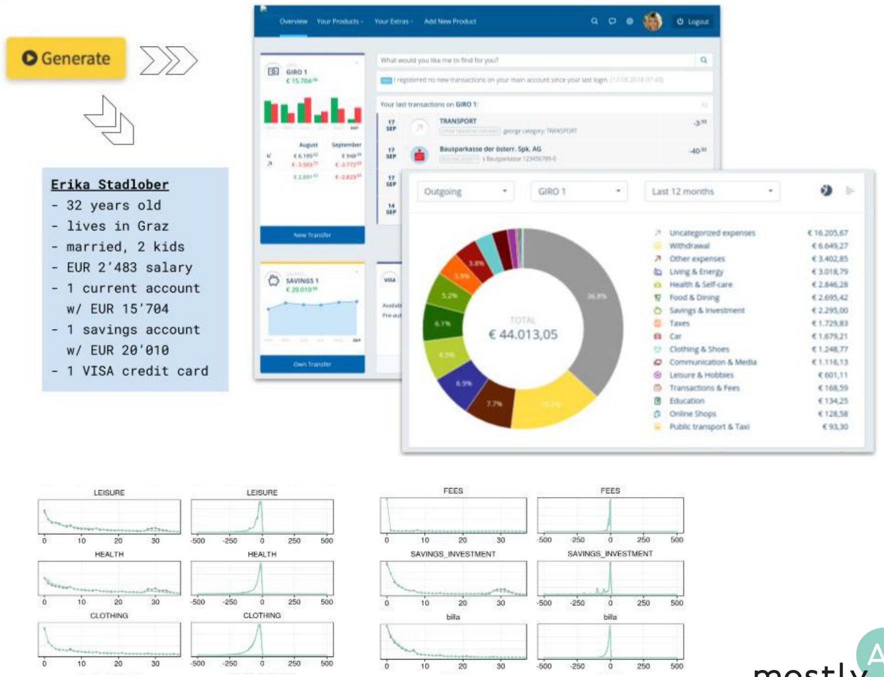
Business Needs

1. Testing with realistic data
2. UX optimization with realistic data
3. 3rd party developer ecosystem
4. Open research collaborations

(…)



Solution - Deployed & Validated



mostly AI

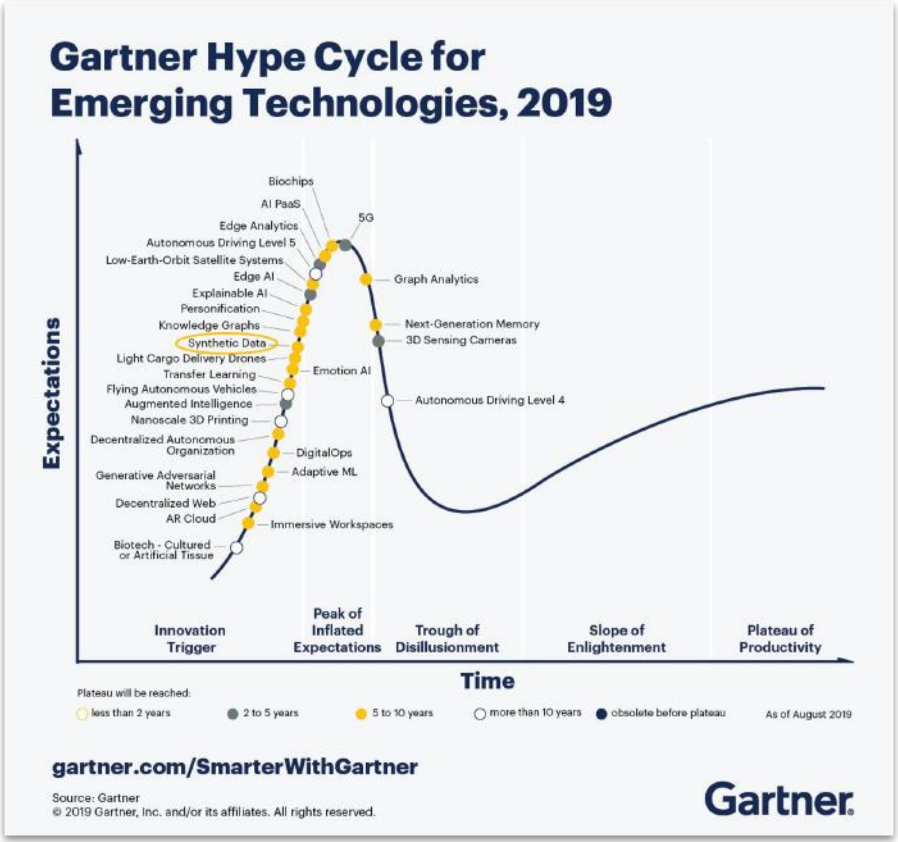
Synthetic Data is THE way forward for **Privacy-Preserving Big Data**

Forbes on Power of Synthetic

"Privacy and Security are Converging" HBR article Jan 2019

US Census goes Synthetic

Nadella: "Privacy is a Human Right"



Join the Synthetic Data Revolution Today!

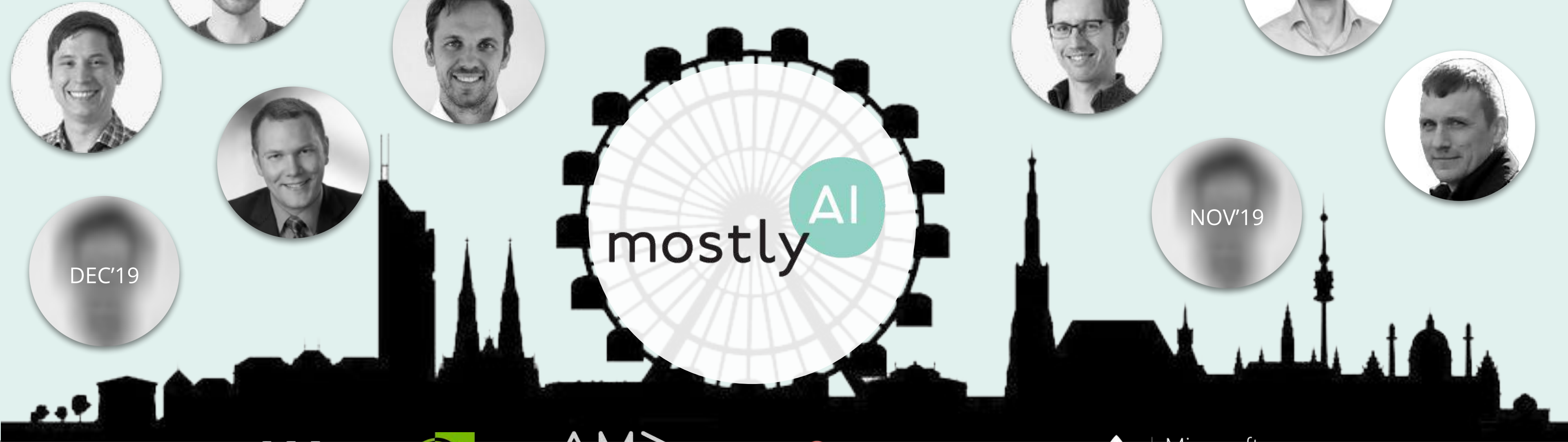



We believe in the power of data.

We believe in the right for privacy.

We are here to make it possible!





PLUGANDPLAY wayra NVIDIA AMA AMERICAN MARKETING ASSOCIATION Google Cloud for Startups Microsoft For Startups weXelerate

Innovative Company?
Big Data Assets Untapped?
Go Synthetic Today!

Alexandra Ebert

Client Relations & External Affairs

alexandra.ebert@mostly.ai

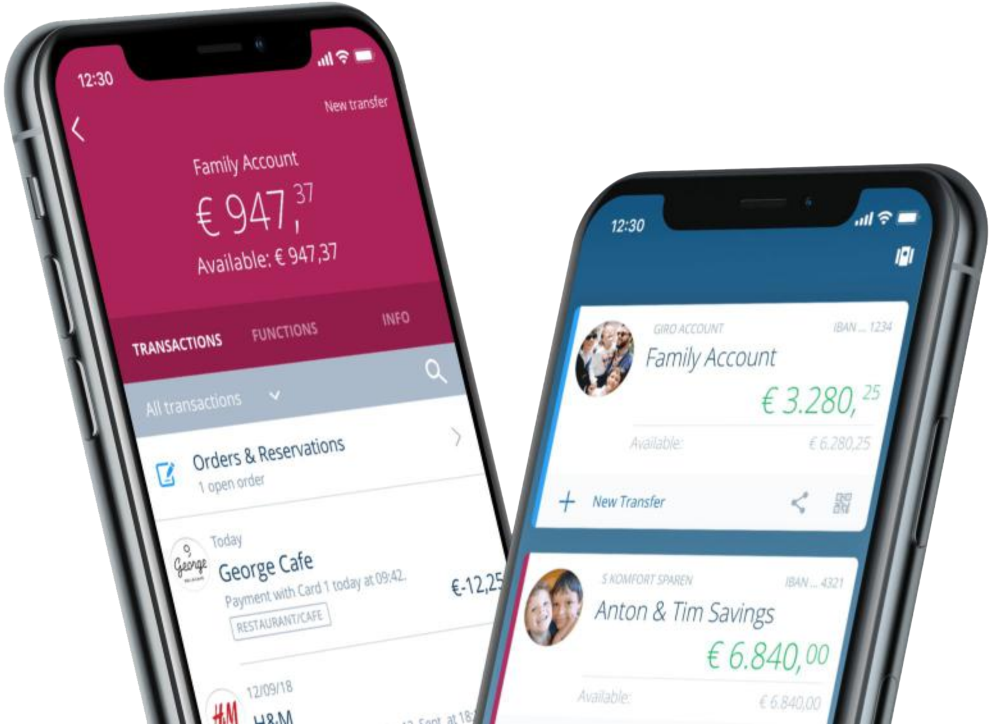
+43 664 884 711 52

www.mostly.ai

Customer Success Story

Customer Success Story

Product Development in Finance Industry



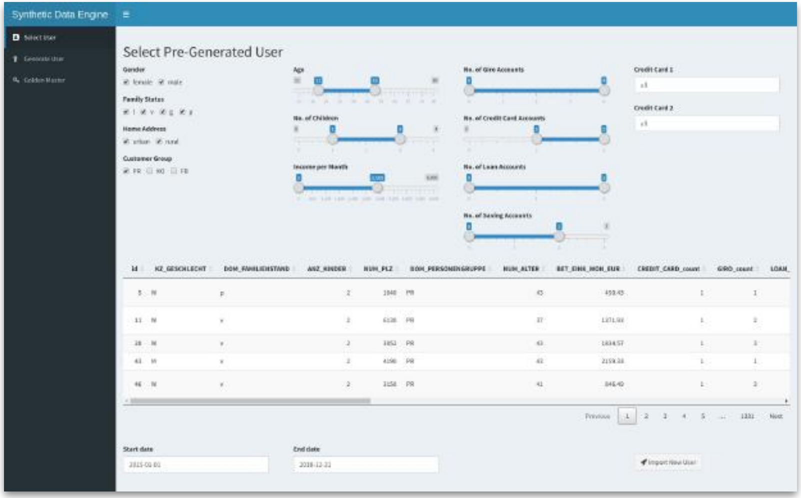
Business Needs

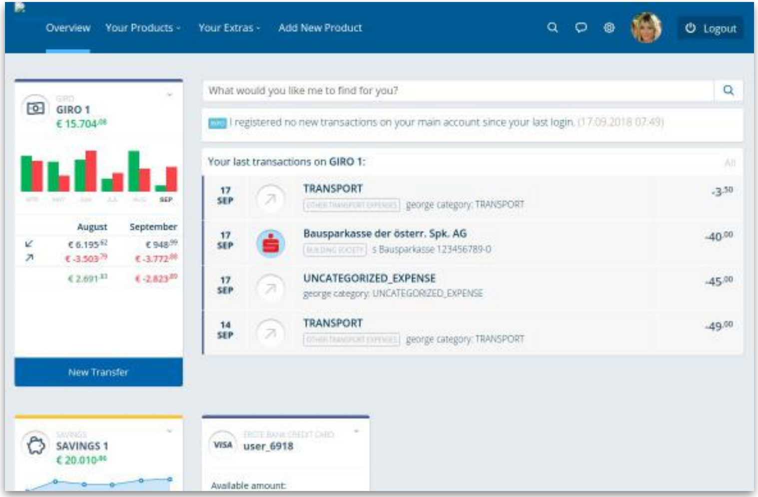
1. Testing with realistic data
2. UX optimization with realistic data
3. Development of smart features (balance forecasting)
4. 3rd party developer ecosystem
5. Open research collaborations with universities

(and more)

Customer Story

Product Development in Finance Industry





The Solution

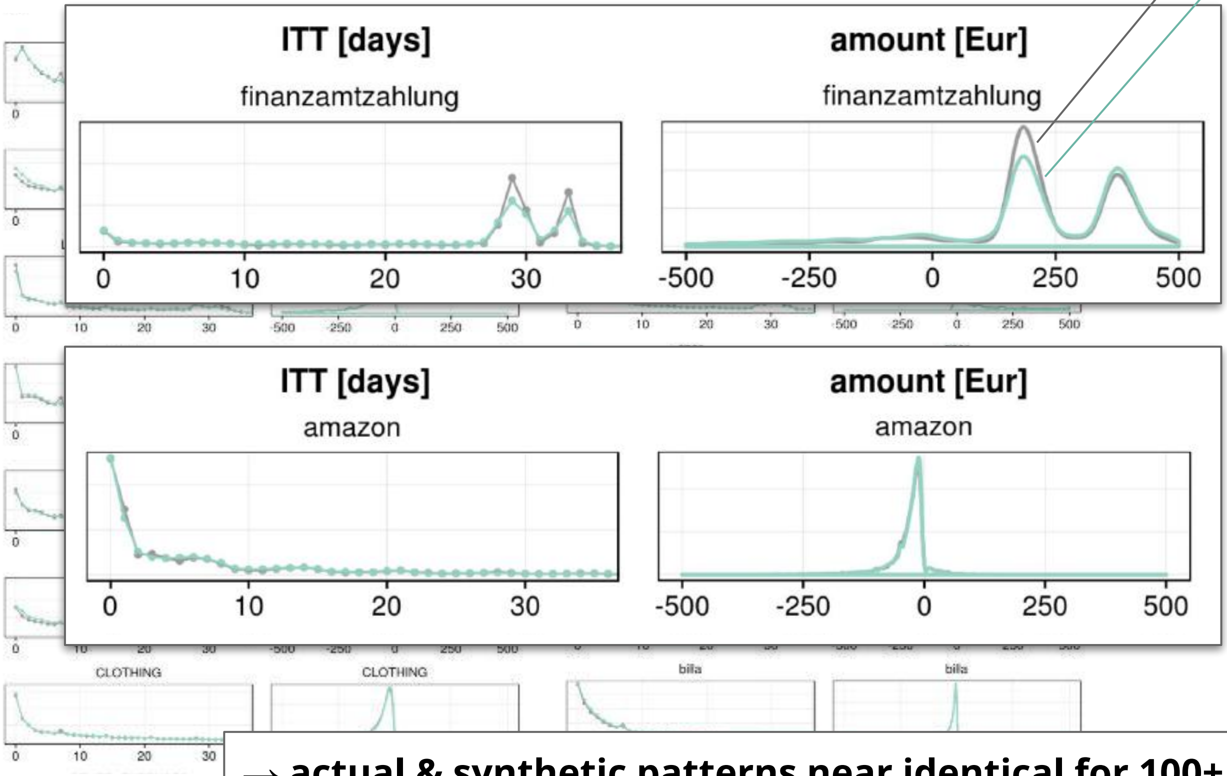
- deep generative model trained on 100k+ customers w/ 100m+ financial transactions
- ability to simulate an unlimited number of synthetic profiles, accounts and transactions
- results are highly realistic and representative; retain detail, structure and variation
- independent audit by bank's analytics team: "***over-achieved***"

Customer Story Data Quality

Transaction Level

actuals

synthetic



100+ merchants

Billa
Spar
Kirchenbeitrag
Ikea
Libro
Kik
Drei
UPC
ORF GIS
Uniqa
ÖBB
WGKK Zahlung
Wiener Netze
Radatz
…

Hofer
Merkur
Starbucks
Mediamarkt
Thalia
Fressnapf
AirBnB
EVN
OMV
A1
Bauhaus
Santander
Generali
Amazon
Finanzamt
…

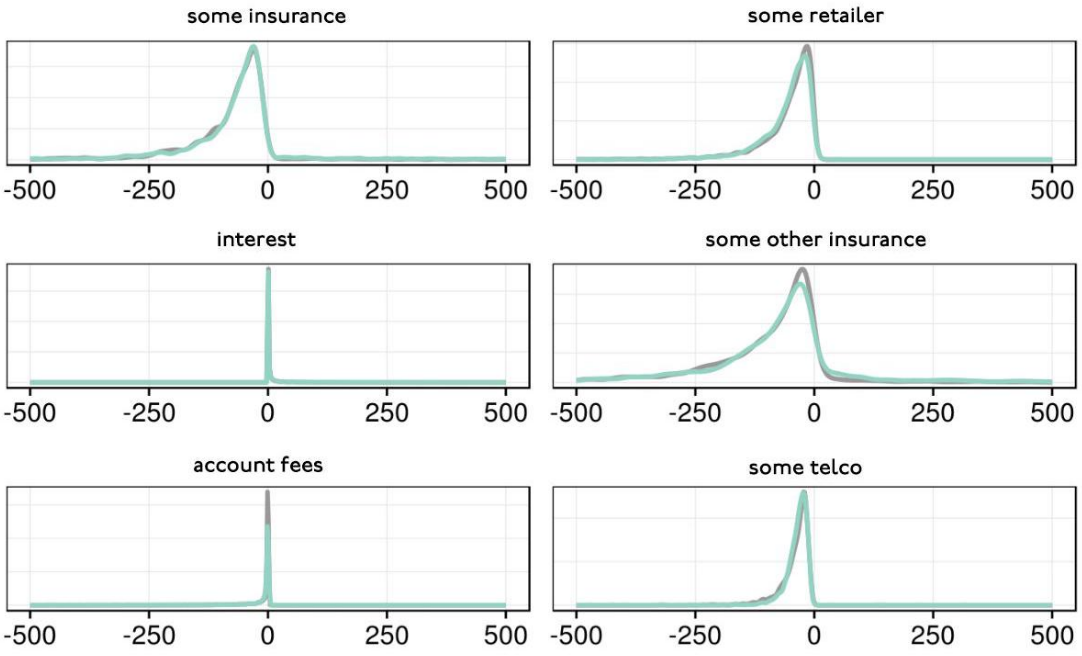
→ **actual & synthetic patterns near identical for 100+ merchants**

mostly AI

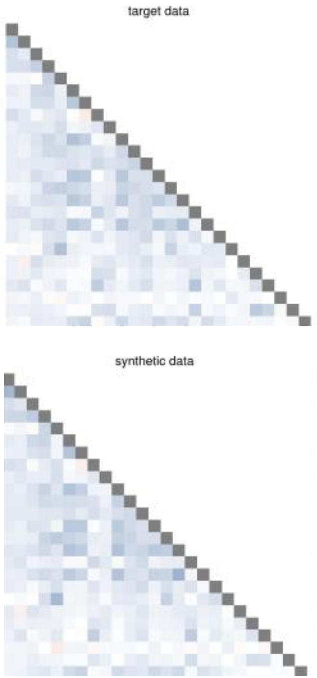
Customer Story Data Quality

Customer Level

€ / month



correlations



→ **actual & synthetic patterns near identical for 100+ merchants**

Demos

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.

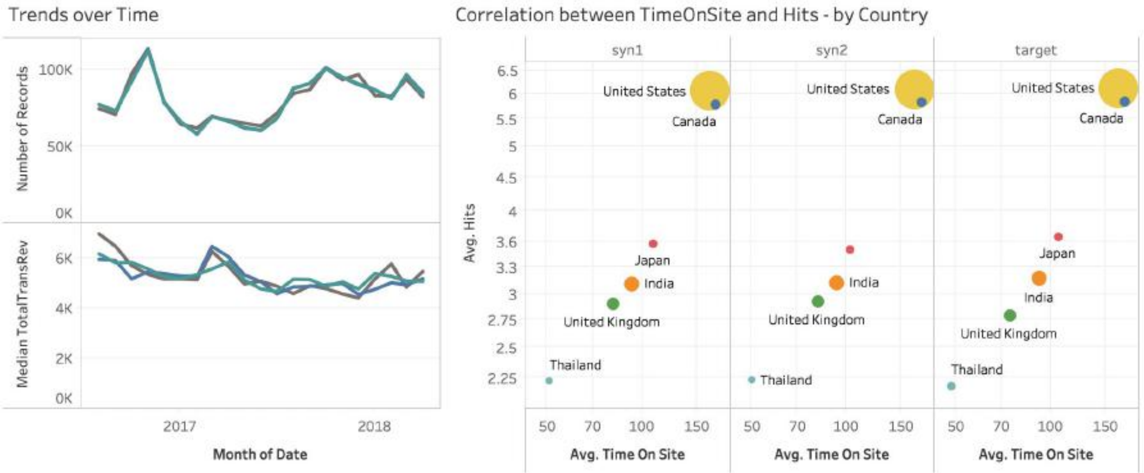
→ **synthetic data can be used for advanced ML algos**
→ **synthetic data also retains weak signals in the data**

Synthetic eCommerce Visitors

Colors: syn1 · syn2 · target

Revenue Distribution by Device Category

	desktop			mobile			tablet		
	syn1	syn2	target	syn1	syn2	target	syn1	syn2	target
Number of Records	1,178,680	1,179,133	1,171,579	461,761	462,215	471,336	64,241	64,174	65,422
Visits w/ Transactions	18,986	19,076	16,905	1,567	1,573	1,364	253	267	245
Pctg Visits w/ Transactions	1.61%	1.62%	1.44%	0.34%	0.34%	0.29%	0.39%	0.42%	0.37%
Percentile (25) of TransRev	2,496	2,418	2,499	1,599	1,696	1,595	1,899	1,699	1,899
Percentile (50) of TransRev	4,959	4,790	4,896	2,598	2,599	2,498	3,044	3,299	2,928
Percentile (75) of TransRev	10,460	10,081	10,583	5,378	5,396	5,249	5,997	5,696	6,592
Median Total Trans Rev	5,578	5,458	5,549	3,099	3,199	2,999	3,493	3,898	3,500



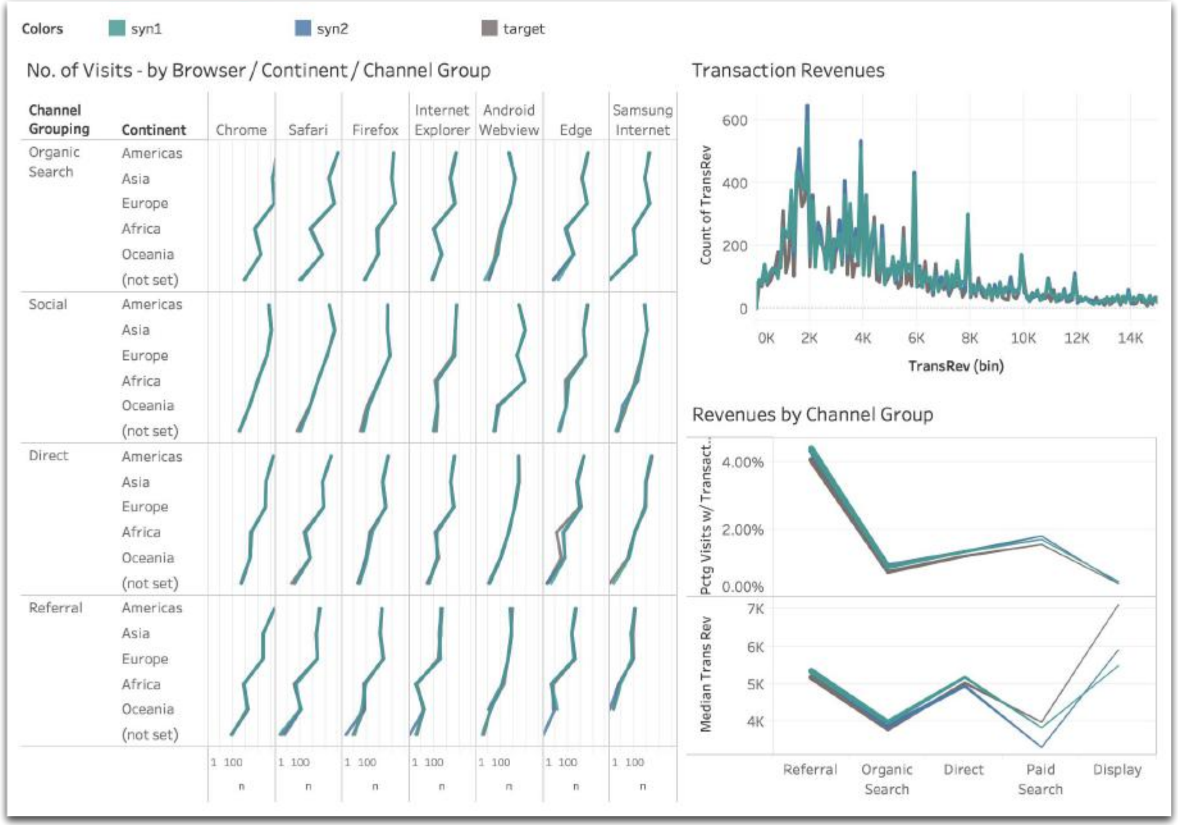
https://www.kaggle.com/c/ga-customer-revenue-prediction

- 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 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