Privacy-Preserving Sharing of Financial Transaction Data with Deep Generative Models

Michael Platzer
Founder & CEO
Mostly AI

Christoph Töglhofer
George Labs
Erste Group
Big Data
➔ drives scientific progress
➔ fosters innovation
➔ improves services, and
➔ helps optimize processes
But if Data is the new Oil...
...then let’s talk side effects
Data Protection is here for a reason
My position is not that there should be no regulation.

Mark Zuckerberg in 2018
The **Privacy vs. Innovation** Clash in Finance

1. **Data privacy** restricts sharing of data and thus **hampers digital innovation**.
The **Privacy vs. Innovation** Clash in Finance

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

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

Günther Baumgartner

Not So Anonymous User #3289384
Classic Anonymization **falls short** for Big Data
Classic Anonymization **Falls Short** for Big Data

i.e. for High-Dimensional, Highly Correlated Data
The Underestimated De-Anonymization Risk

Simple Demographics Often Identify People Uniquely (Sweeney, 2000)
→ 87% of US citizens identified by date-of-birth, gender and ZIP

AOL search data leak (2006)

Robust De-anonymization of Large Sparse Datasets (Narayanan, 2008)
→ Partial re-identification of the Netflix dataset
“Sanitization techniques from the k-anonymity literature [..] do not provide meaningful privacy guarantees, and in any case fail on high-dimensional data.”

The privacy bounds of human mobility (Montjoye, 2013)
→ 2 spatio-temporal points are enough to uniquely identify 55% of 1.5m people

Unique in the shopping mall: On the reidentifiability of credit card metadata (Montjoye, 2015)
→ 4 spatio-temporal points are enough to uniquely identify 90% of 1.1m people
→ “even coarse datasets provide little anonymity”

Stalking Celebrities in NYC Taxi Dataset (2014; viz)

→ Everyone’s Digital Trail is Highly Unique
The Problem Untapped Big Data Potential

Growing Gap

amount of **shareable** information

amount of **captured** information

Classic Anonymization can only share very limited amount of information per person
The Solution

Synthetic Data
Synthetic Data?

hand-crafted: simplistic and biased

John Doe
Jane Doe
Synthetic Data?

hand-engineered: simplistic and biased

John Doe

Jane Doe
AI-Generated Synthetic Data!

AI generated = realistic and representative

state-of-the-art deep neural networks, trained by NVIDIA on 30'000 celebrity photos

the generated new images look like real persons, but they aren't
Scales with Data Growth
while fully preserving privacy of actual customers

- **Classic Anonymization** can only share very limited amount of information per person.

- **AI-generated Synthetic Data** retains structure and variation, but no 1:1 relationship to actual persons.
The **Synthetic Data Engine** by Mostly AI

AI-generated **rich synthetic worlds** of customers and their behavior

### Table: Actual

<table>
<thead>
<tr>
<th>NAME</th>
<th>ZIP</th>
<th>AGE</th>
<th>GENDER</th>
<th>ITEM</th>
<th>EUR</th>
<th>LAT</th>
<th>LONG</th>
<th>DAY</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tina</td>
<td>1180</td>
<td>23y</td>
<td>female</td>
<td>Car2Go</td>
<td>22€</td>
<td>48.11N</td>
<td>16.37E</td>
<td>Mon</td>
<td>10:00</td>
</tr>
<tr>
<td>Tina</td>
<td>1180</td>
<td>23y</td>
<td>female</td>
<td>Shoes</td>
<td>78€</td>
<td>48.21N</td>
<td>16.38E</td>
<td>Tue</td>
<td>12:00</td>
</tr>
<tr>
<td>Kurt</td>
<td>1170</td>
<td>39y</td>
<td>male</td>
<td>Beer</td>
<td>3€</td>
<td>47.32N</td>
<td>16.02E</td>
<td>Sat</td>
<td>19:30</td>
</tr>
<tr>
<td>Kurt</td>
<td>1170</td>
<td>39y</td>
<td>male</td>
<td>Beer</td>
<td>3€</td>
<td>47.35N</td>
<td>16.14E</td>
<td>Sat</td>
<td>20:30</td>
</tr>
</tbody>
</table>

### Table: Synthetic

<table>
<thead>
<tr>
<th>NAME</th>
<th>ZIP</th>
<th>AGE</th>
<th>GENDER</th>
<th>ITEM</th>
<th>EUR</th>
<th>LAT</th>
<th>LONG</th>
<th>DAY</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>1220</td>
<td>25y</td>
<td>female</td>
<td>Book</td>
<td>12€</td>
<td>48.20N</td>
<td>16.78E</td>
<td>Fri</td>
<td>10:00</td>
</tr>
<tr>
<td>Mary</td>
<td>1220</td>
<td>25y</td>
<td>female</td>
<td>Swim</td>
<td>8€</td>
<td>48.32N</td>
<td>16.82E</td>
<td>Fri</td>
<td>15:10</td>
</tr>
<tr>
<td>John</td>
<td>1210</td>
<td>39y</td>
<td>male</td>
<td>Lunch</td>
<td>13€</td>
<td>47.32N</td>
<td>16.02E</td>
<td>Mon</td>
<td>9:30</td>
</tr>
<tr>
<td>John</td>
<td>1210</td>
<td>39y</td>
<td>male</td>
<td>Food</td>
<td>8€</td>
<td>47.32N</td>
<td>16.02E</td>
<td>Mon</td>
<td>10:30</td>
</tr>
</tbody>
</table>
Bayesian Model of Customer Transactions (Platzer and Reutterer, 2016)

flexible MCMC estimation of parametric customer model!
Bayesian Model of Customer Transactions (Platzer and Reutterer, 2016)

But computation & model capacity **did not scale** to big data
Machine Learning

Learning by Being Taught

→ **Supervised Learning**

Learning by Observation

→ **Unsupervised 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

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

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

The generator turns random noise into imitations of the data, in an attempt to fool the discriminator.
Generative Deep Models – ARNs
Autoregressive Neural Networks

VIOLA:
Why, Salisbury must find his flesh and thought
That which I am not apt, 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.

/*
 * 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, alist->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 Shakespeare

Synthetic Linux Source Code
The Customer Story
Product Development in Finance Industry

Banking has a name. George.
Re-inventing banking. For everyone.
Simple. Intelligent. Personal.
George is an innovation by George Labs of Erste Group.
Customer Story  The Business Problem

- UX Development & Testing w/ realistic data
- platform for 3rd party development
- feature dev: forecasting account balances
- open research collaboration with university
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

actual & synthetic patterns near identical for 100+ merchants

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  ...

Customer Story  Data Quality

Customer Level

→ actual & synthetic patterns near identical for 100+ merchants
**Customer Story**  
**System Setup**

- **on-premise, secure**  
- **GPU environment**

2x Quadro P4000 (8GB)  
(training runs ~18h)

- **anywhere**  
- **anytime**  
- **CLI / REST API**
Customer Story  Model Architecture
Fully Autoregressive Neural Network
Key Takeaway

Data privacy restricts sharing of data and thus hampers digital innovation.

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

Synthetic data is anonymous.

Generative AI allows highly accurate synthetic data to be generated at scale.
Contact Details

Michael Platzer  
Founder & CEO  
michael.platzer@mostly.ai  
https://mostly.ai/  

Christoph Töglhofer  
Data Scientist  
george@erstegroup.com  
https://george-labs.com/  
christoph.toeglhofer@erstegroup.com  
Erste Group IT International GmbH  

YES, WE'RE HIRING  
mostly AI  

YES, WE'RE HIRING  
Erste Group