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…

ERSTE Group

mostly AI

…then let's talk side effects

ERSTE Group

mostly AI

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**.

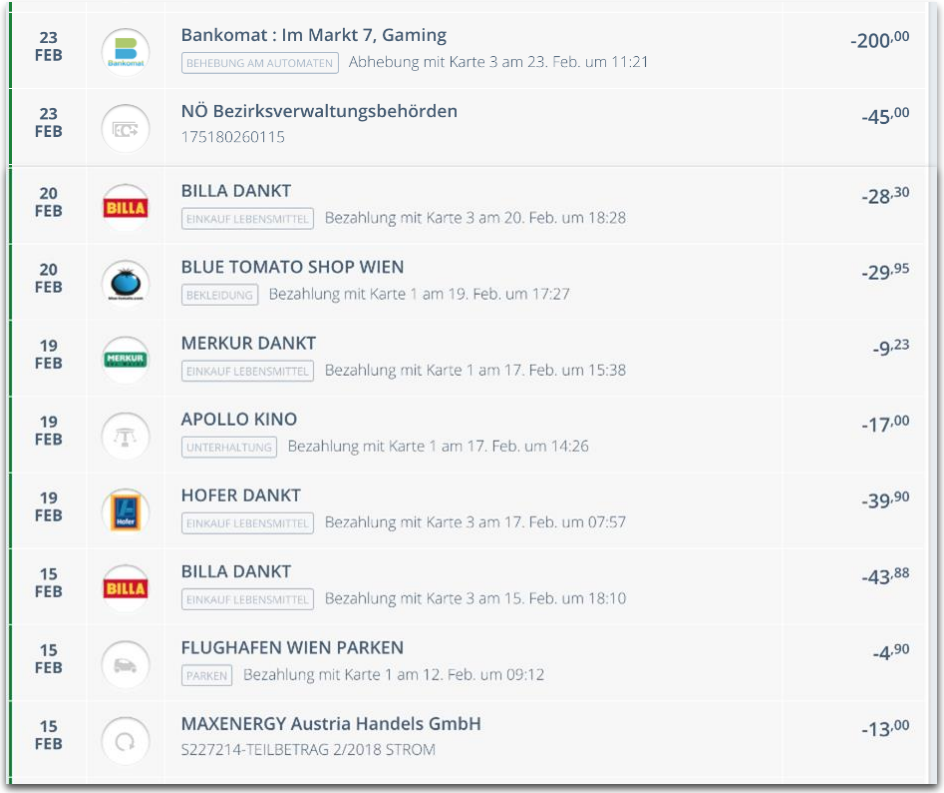
The **Privacy vs. Innovation** Clash in Finance

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



23 FEB	Bankomat : Im Markt 7, Gaming BEHEBUNG AM AUTOMATEN Abhebung mit Karte 3 am 23. Feb. um 11:21	-200,00
23 FEB	NÖ Bezirksverwaltungsbehörden 175180260115	-45,00
20 FEB	BILLA DANKT EINKAUF LEBENSMITTEL Bezahlung mit Karte 3 am 20. Feb. um 18:28	-28,30
20 FEB	BLUE TOMATO SHOP WIEN BEKLEIDUNG Bezahlung mit Karte 1 am 19. Feb. um 17:27	-29,95
19 FEB	MERKUR DANKT EINKAUF LEBENSMITTEL Bezahlung mit Karte 1 am 17. Feb. um 15:38	-9,23
19 FEB	APOLLO KINO UNTERHALTUNG Bezahlung mit Karte 1 am 17. Feb. um 14:26	-17,00
19 FEB	HOFER DANKT EINKAUF LEBENSMITTEL Bezahlung mit Karte 3 am 17. Feb. um 07:57	-39,90
15 FEB	BILLA DANKT EINKAUF LEBENSMITTEL Bezahlung mit Karte 3 am 15. Feb. um 18:10	-43,88
15 FEB	FLUGHAFEN WIEN PARKEN PARKEN Bezahlung mit Karte 1 am 12. Feb. um 09:12	-4,90
15 FEB	MAXENERGY Austria Handels GmbH S227214-TEILBETRAG 2/2018 STROM	-13,00

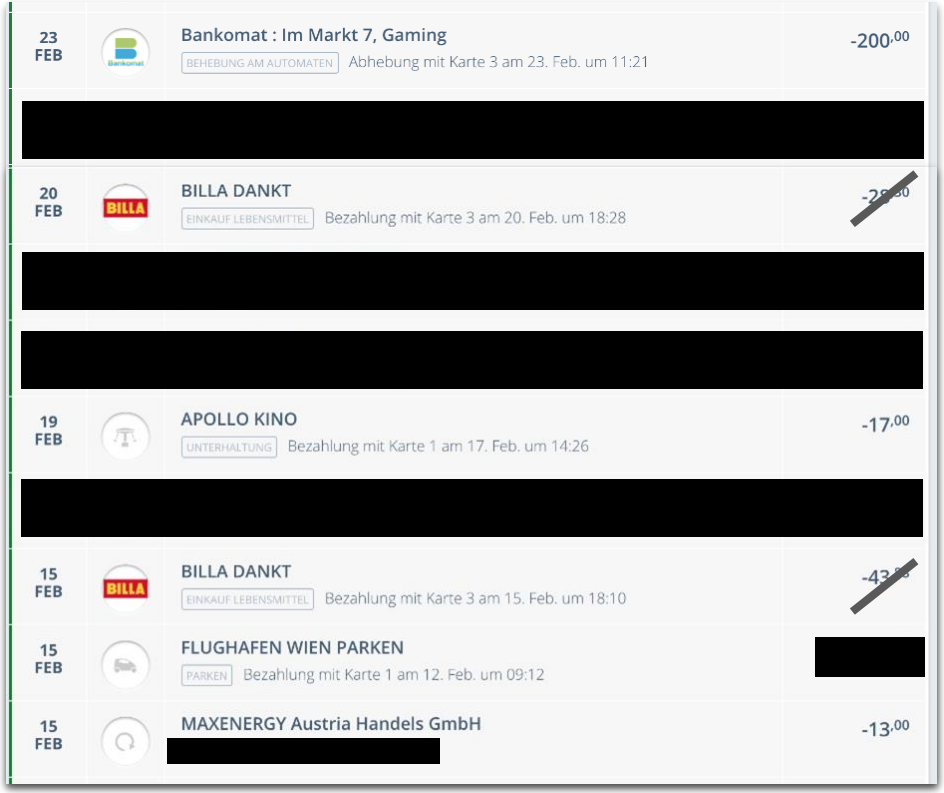
Not So Anonymous User #3289384

ERSTE Group

mostly AI

Classic Anonymization…

Still Not So Anonymous

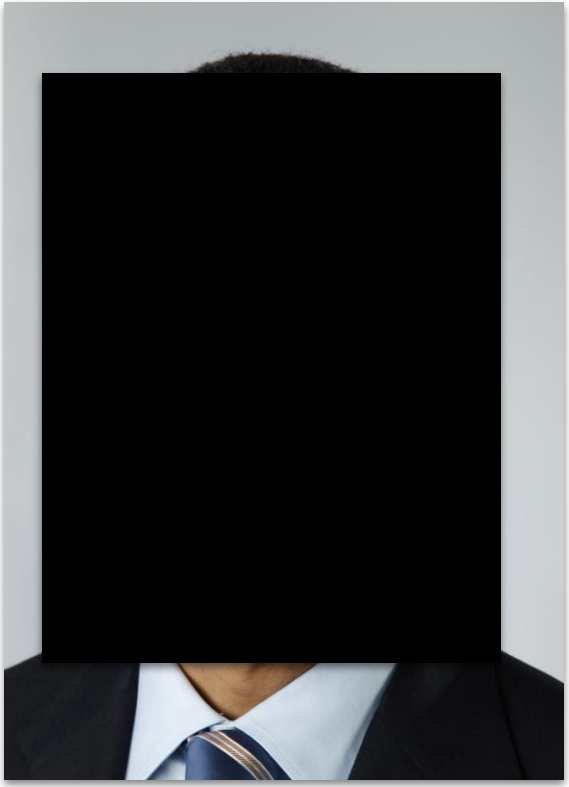


20-30€

40-50€

Still Not So Anonymous

Classic Anonymization Falls Short for Big Data



Anonymous

0-30€

0-30€

30-60€

Anonymous

ERSTE Group

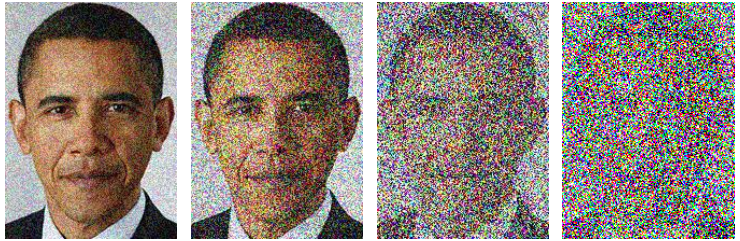
mostly AI

Classic Anonymization Falls Short for Big Data

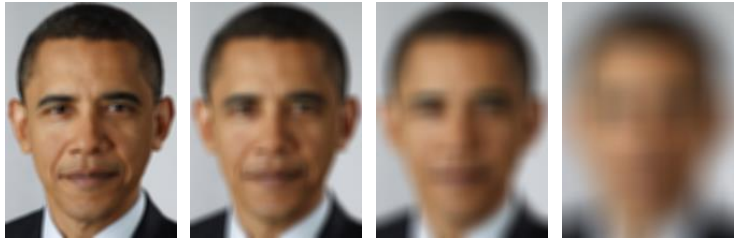
i.e. for High-Dimensional, Highly Correlated Data

masking - obfuscation

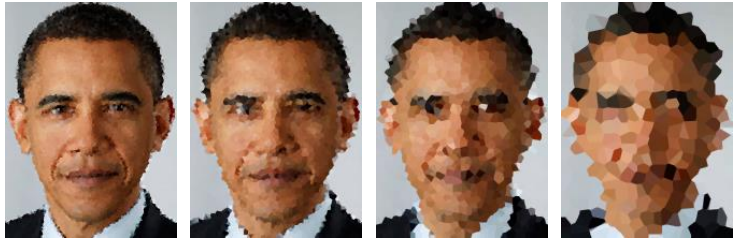
adding noise - perturbations



generalization



generalization



ERSTE Group

mostly AI

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

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
→ *"even coarse datasets provide little anonymity"*

Stalking Celebrities in NYC Taxi Dataset (2014; viz)

→ Everyone's Digital Trail is Highly Unique

ERSTE Group mostly AI

The Solution

Synthetic Data

ERSTE Group

mostly AI

Synthetic Data ?

hand-crafted: simplistic and biased

John Doe

Jane Doe

Synthetic Data ?

hand-engineered: simplistic and biased

John Doe

Jane Doe

ERSTE Group

mostly AI

AI-Generated **Synthetic Data** !

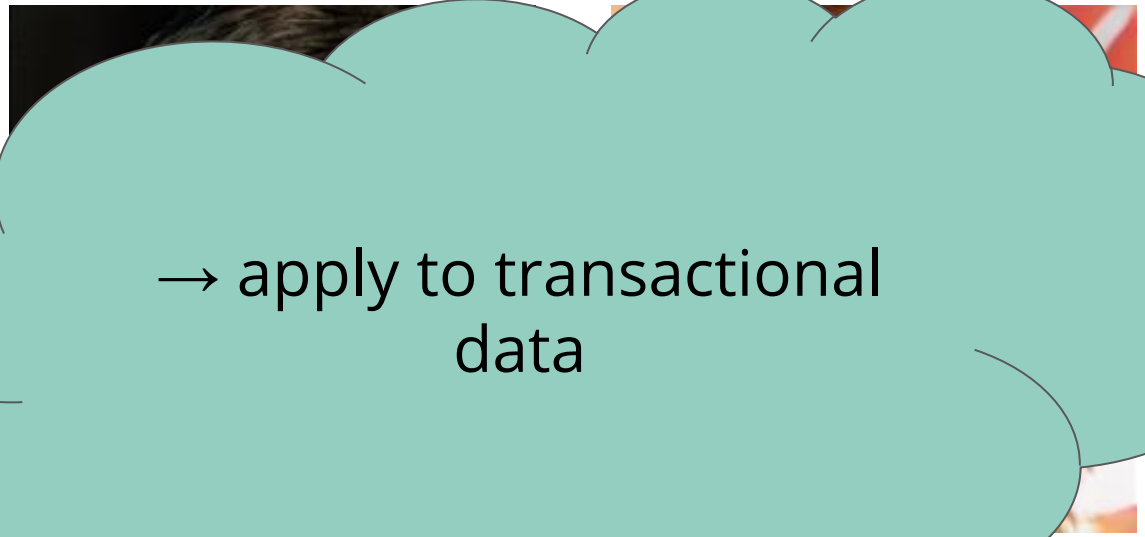
AI generated = realistic and representative

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

AI-Generated **Synthetic Data** !

AI generated = realistic and representative



→ apply to transactional data

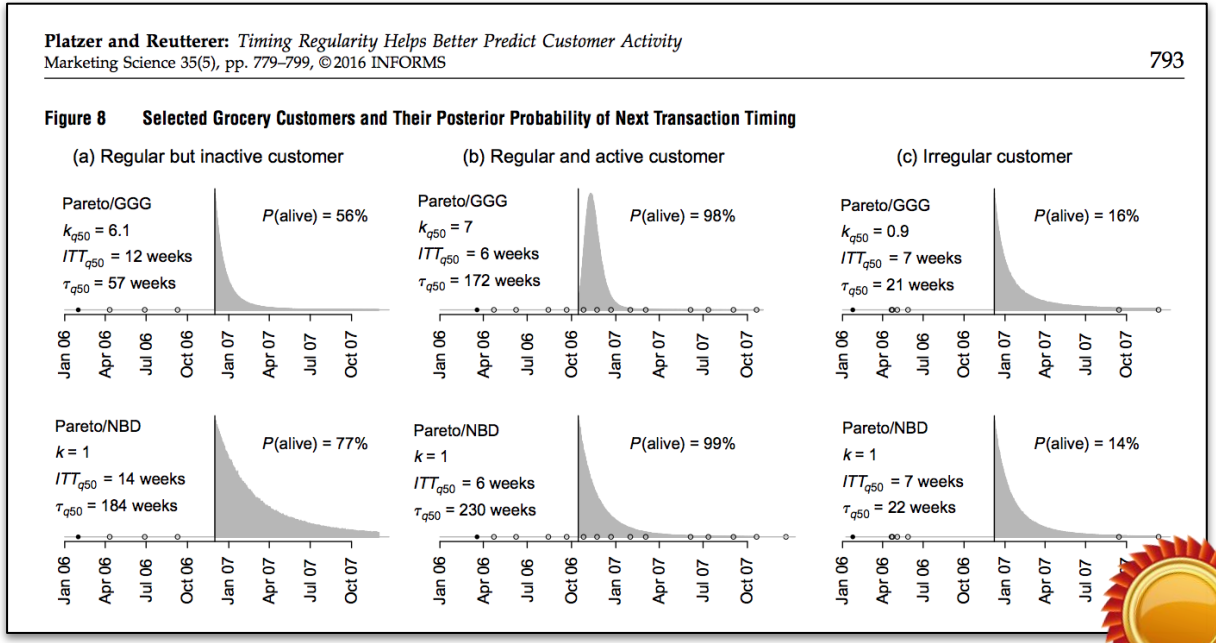
state-of-the- A on 30'000 celebrity photos

the generated new images r al persons, but they aren't

ERSTE Group

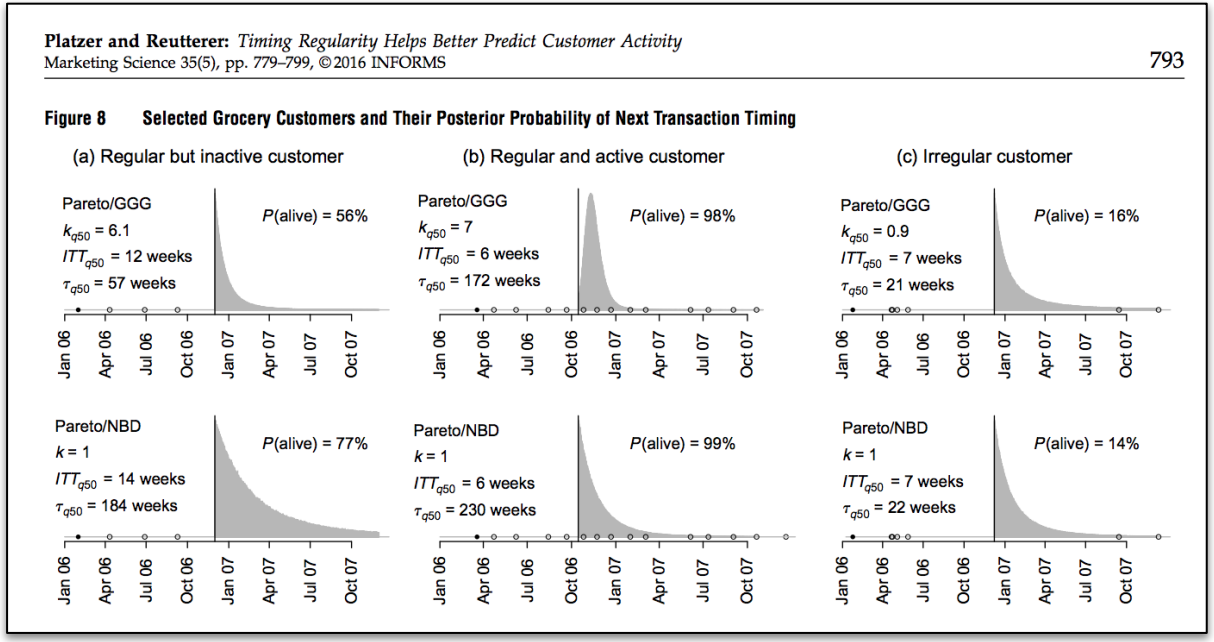
mostly AI

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)



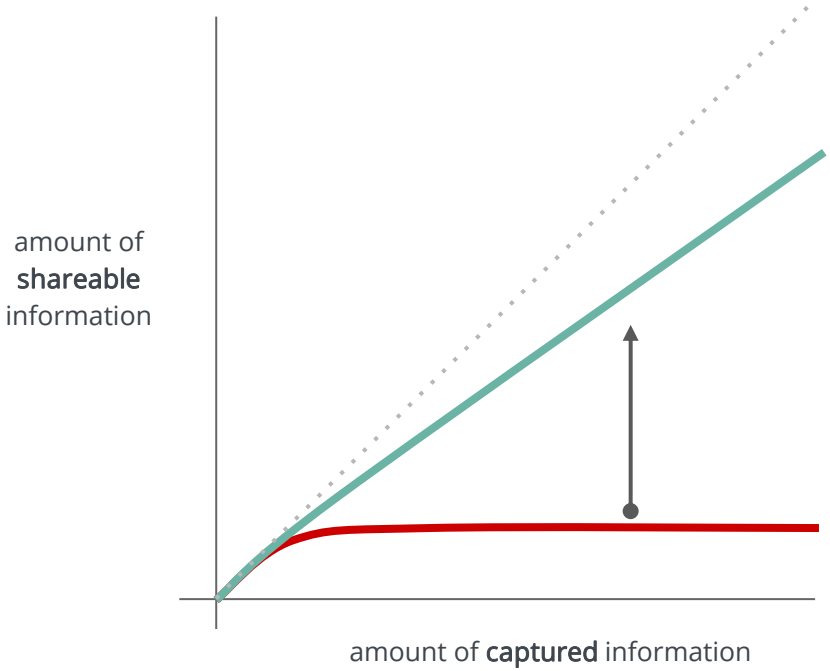
Platzer and Reutterer: *Timing Regularity Helps Better Predict Customer Activity*
Marketing Science 35(5), pp. 779–799, © 2016 INFORMS 793

Figure 8 **Selected Grocery Customers and Their Posterior Probability of Next Transaction Timing**

But computation & model capacity did not scale to big data

Scales with Data Growth

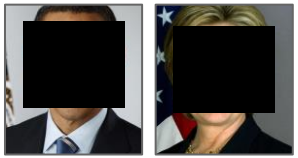
while fully preserving privacy of actual customers



...

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

Machine Learning

Learning by Being Taught

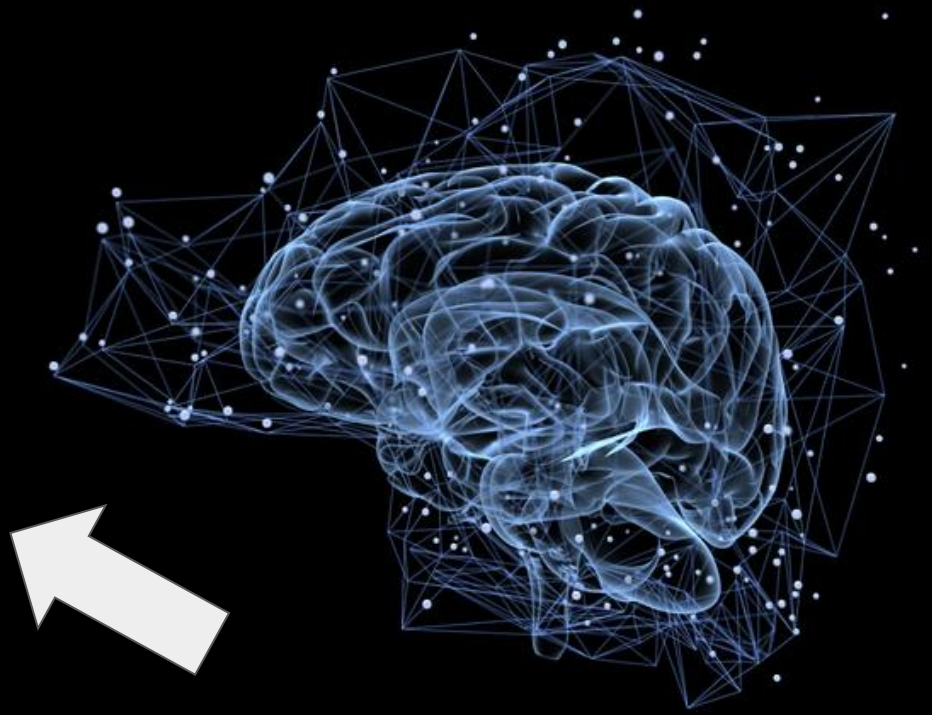
→ **Supervised Learning**

Learning by Observation

→ **Unsupervised learning**

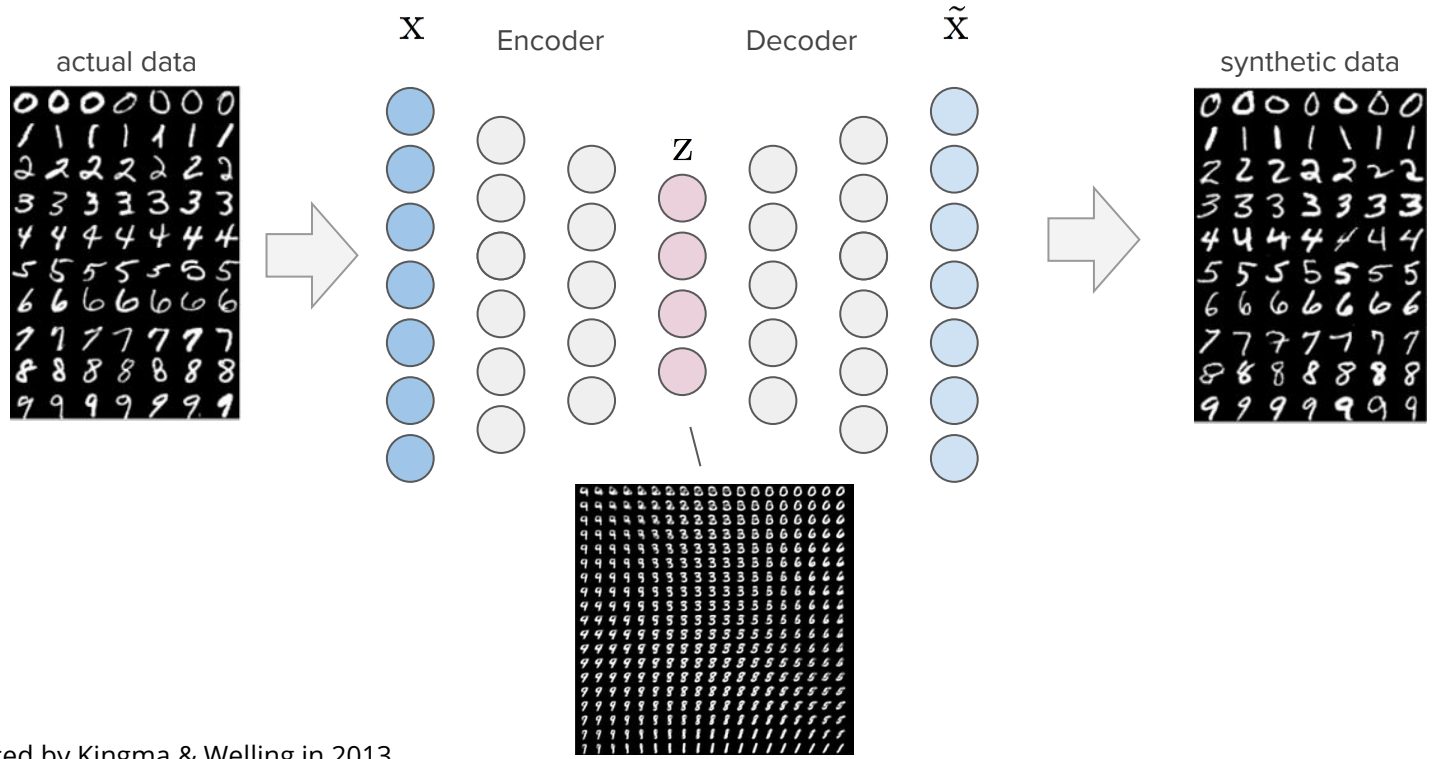
Learning by Exploration

→ **Reinforcement Learning**



Generative Deep Models – VAEs

Variational Autoencoders



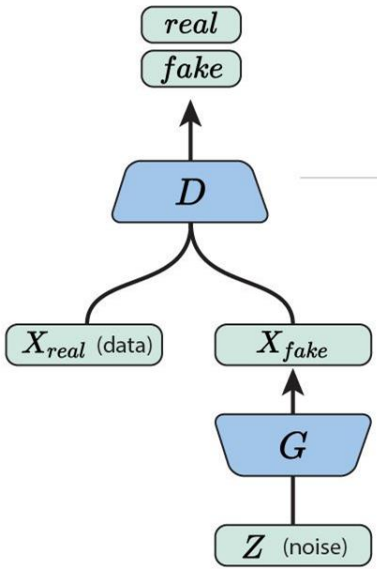
- VAE introduced by Kingma & Welling in 2013
- 600+ papers published on VAE in 2017

ERSTE Group

mostly AI

Generative Deep Models – GANs

Generative Adversarial Networks



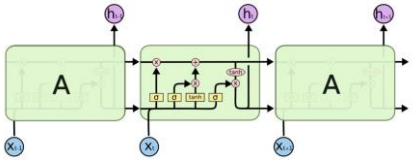
The **discriminator** tries to distinguish genuine data from forgeries created by the generator.

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

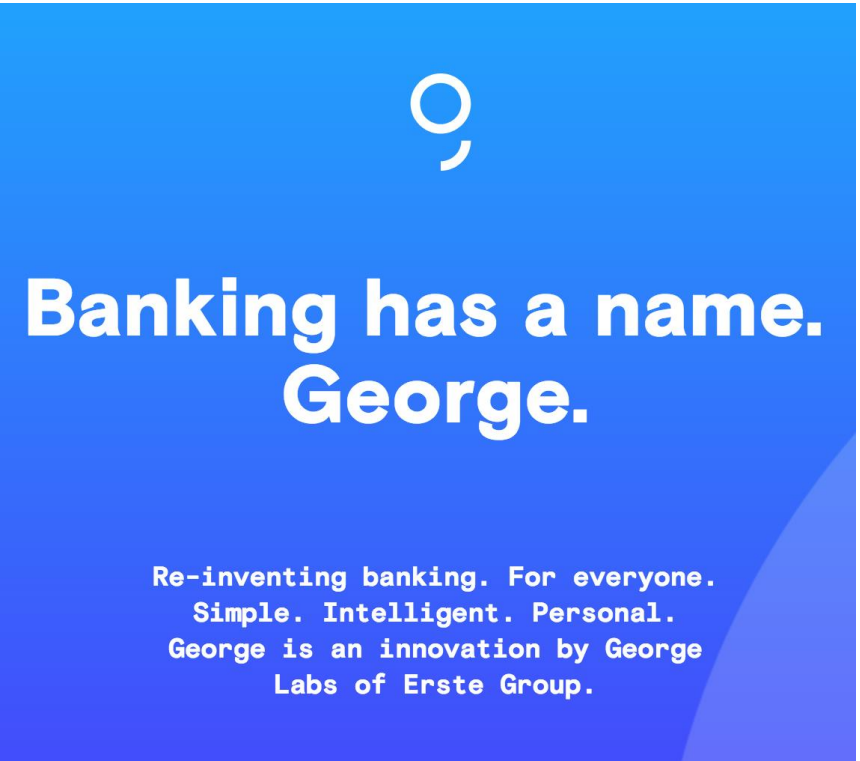
```
/*
 * If this error is set, we will need anything right after that BSD.
 */
static void action_new_function(struct s_stat_info *wb)
{
  unsigned long flags;
  int lel_idx_bit = e->edd, *sys & ~((unsigned long) *FIRST_COMPAT);
  buf[0] = 0xFFFFFFFF & (bit << 4);
  min(inc, slist->bytes);
  printk(KERN_WARNING "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

ERSTE Group mostly AI

The Customer Story

Product Development in Finance Industry

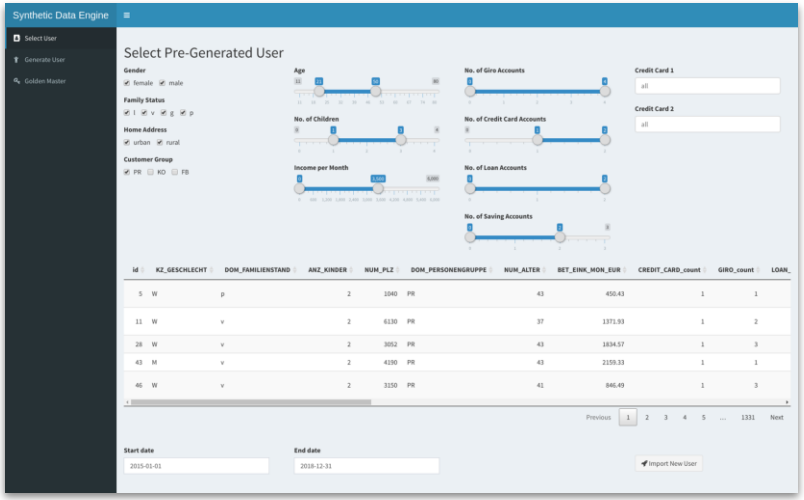


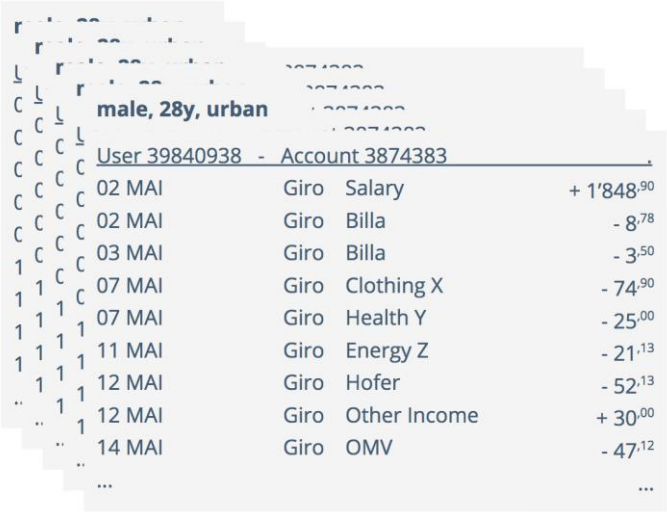
Customer Story The Business Problem

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

ERSTE Group mostly AI

Customer Story The Solution



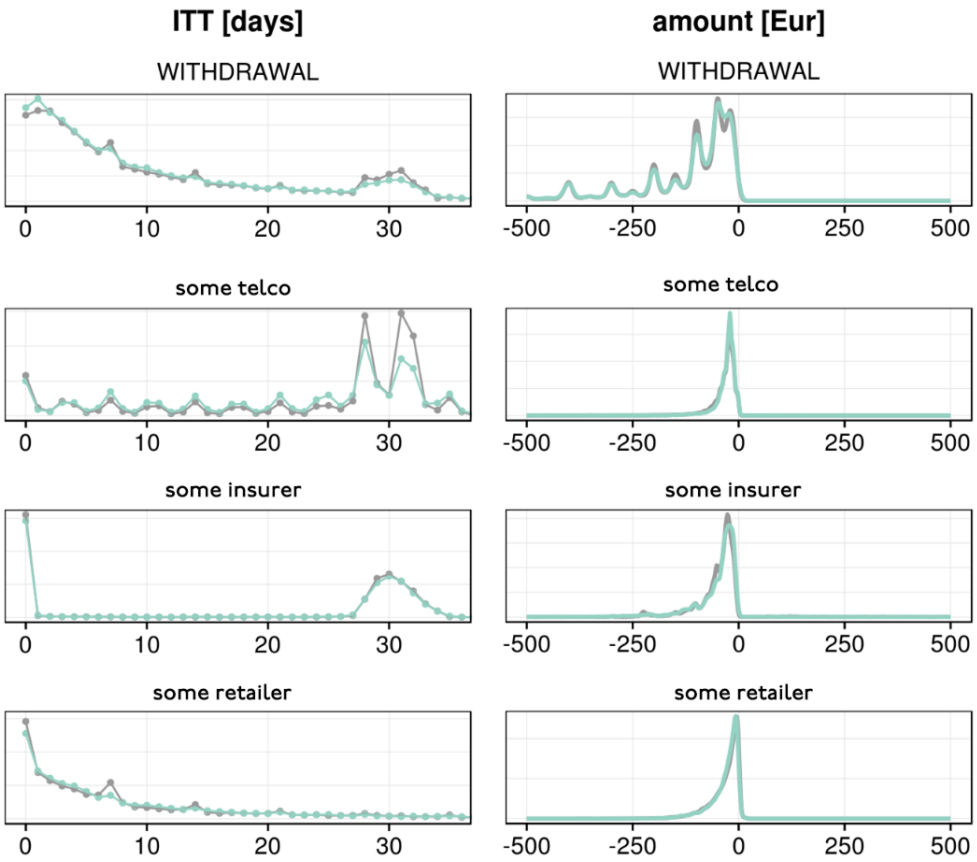


- **deep generative model** trained on 100k+ customers w/ 100m+ financial transactions
- ability to simulate an unlimited number of synthetic profiles, accounts and transactions
- results are highly realistic and representative; retain detail, structure and variation
- independent audit by bank's analytics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

ERSTE Group

mostly AI

Customer Story Synthetic Bank Transactions



Current Accounts, Credit Card,

Loans, Savings, etc.

100+ merchants & category

ERSTE Group

mostly AI

Customer Story Synthetic Bank Customers

User 39840938 - Account 3874383

02 MAI		Giro	Salary	+ 1’848,90
		+ 1’848,90		- 8,78
02 MAI		Giro	Billa	- 3,50
		- 8,78		- 74,90
03 MAI		Giro	Billa	- 25,00
		- 3,50		- 21,13
07 MAI		Giro	Clothir	- 52,13
	- 74,90			+ 30,00
07 MAI		Giro	Health	- 47,12
	- 25,00			...
11 MAI		Giro	Energy Z	
	- 21,13			
12 MAI		Giro	Hofer	

Customer Story Synthetic Bank Customers

User 39840938 - Account 3874383

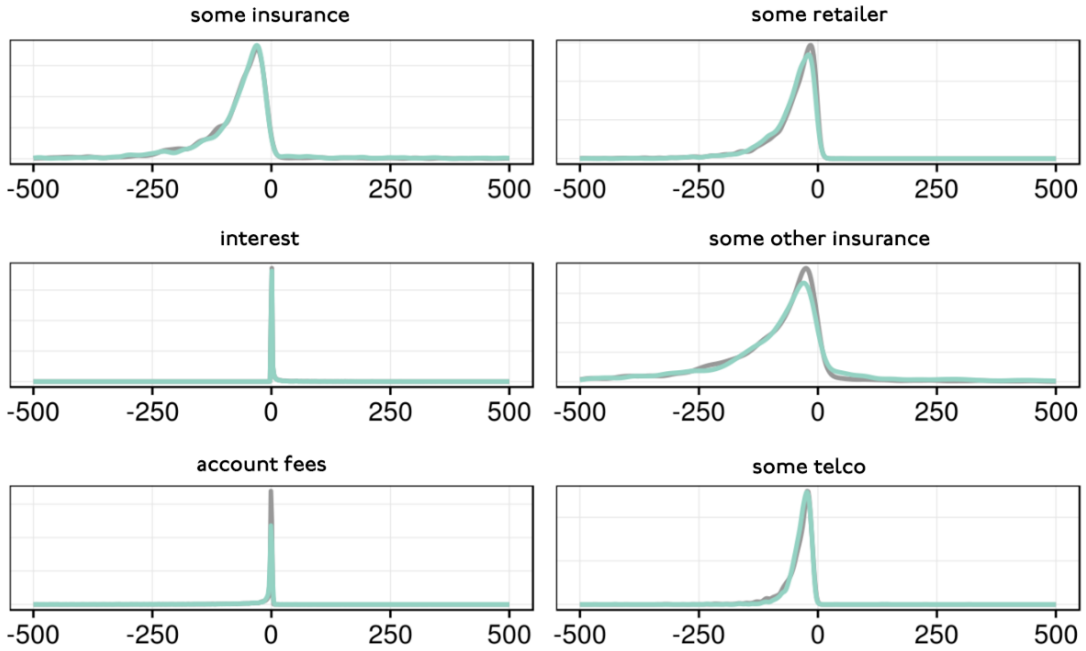
02 MAI	Giro	Salary	+ 1’848,90
	+ 1’848,90		- 8,78
02 MAI	Giro	Billa	- 3,50
	- 8,78		- 74,90
03 MAI	Giro	Billa	- 25,00
	- 3,50		- 21,13
07 MAI	Giro	Clothi	- 52,13
	- 74,90		+ 30,00
07 MAI	Giro	Health	- 47,12
	- 25,00		...
11 MAI	Giro	Hofer	Energy
	- 21,13 - 52,13		
12 MAI	Giro	Uncat.	Income
	- 21,13 - 52,13		
12 MAI	Giro	Uncat.	Hofer
	+ 30,00 - 52,13		
12 MAI	Giro	Uncat.	Hofer
	+ 30,00 - 52,13		

ERSTE Group

mostly AI

Customer Story Synthetic Bank Customers

€ / month



correlations

ERSTE Group

mostly AI

Customer Story conditional Synthetic Bank Customers

male, 28y, urban

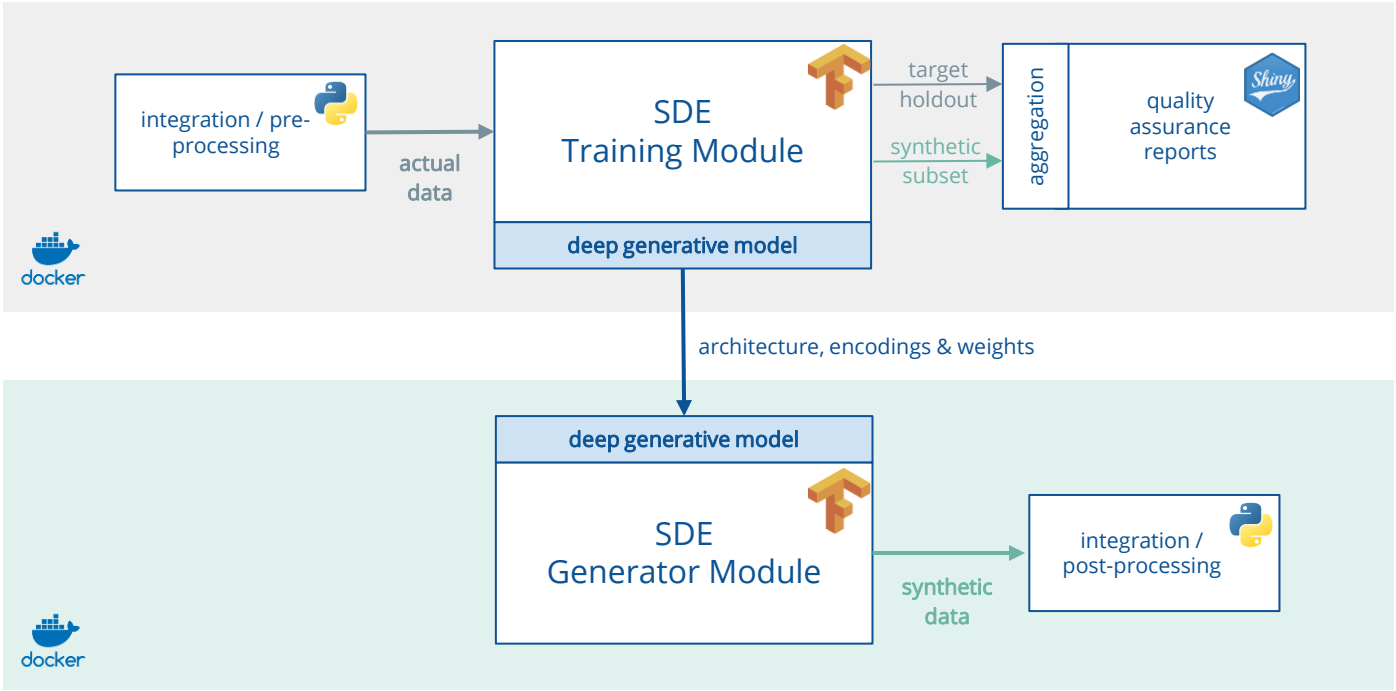
User 39840938 - Account 3874383

02 MAI	Giro	Salary	+ 1’848,90
	+ 1’848,90		- 8,78
02 MAI	Giro	Billa	- 3,50
	- 8,78		- 74,90
03 MAI	Giro	Billa	- 25,00
	- 3,50		- 21,13
07 MAI	Giro	Clothir	- 52,13
	- 74,90		+ 30,00
07 MAI	Giro	Health	- 47,12
	- 25,00		...
11 MAI	Giro	Hofer	Energy Z
	- 21,13 - 52,13		
	- 21,13 - 52,13		
	Giro - 52,13	Uncat. Income	
	Giro	Uncat. Income	

ERSTE Group

mostly AI

Customer Story System Setup



- on-premise, secure
- GPU environment

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

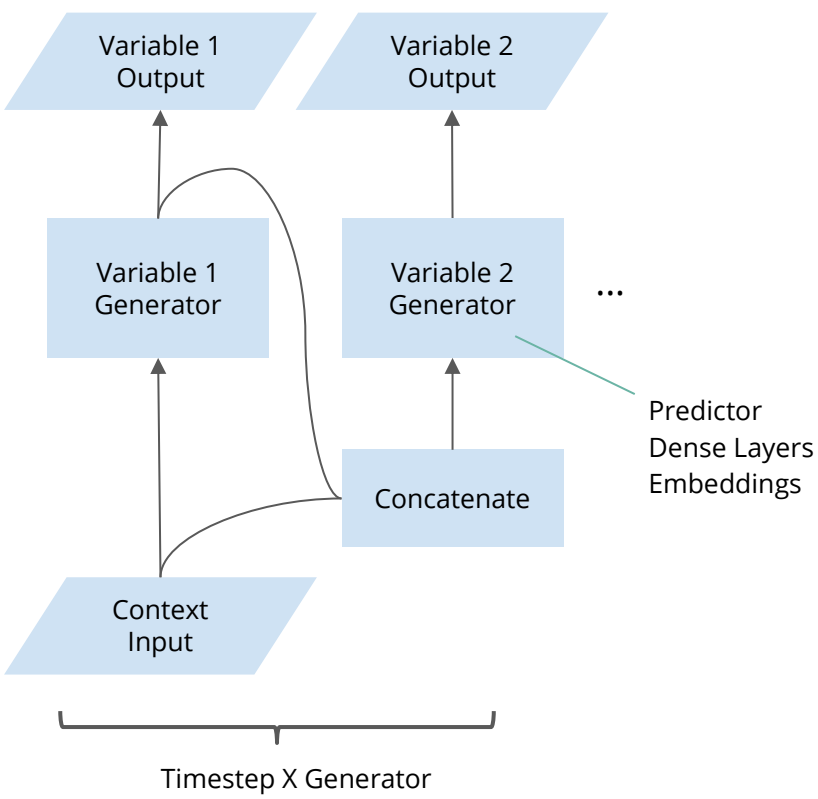
- anywhere
- anytime
- CLI / REST API

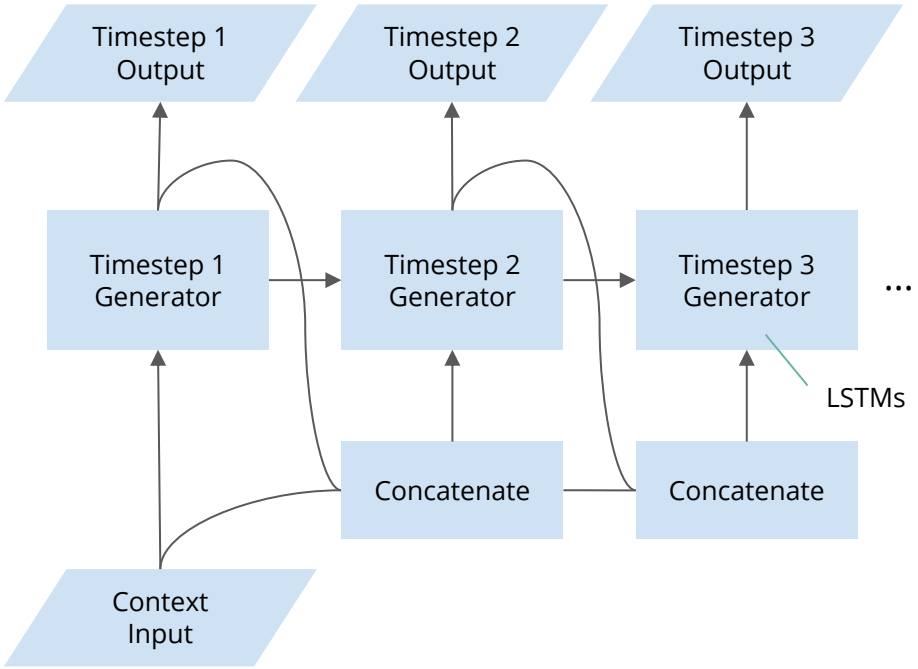
ERSTE Group

mostly AI

Customer Story Model Architecture

Fully Autoregressive Neural Network





ERSTE Group

mostly AI

Key Takeaway

PROBLEM

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.

SOLUTION

1. **Synthetic data is anonymous.**

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

ERSTE Group

mostly AI

Contact Details

Michael Platzer

Founder & CEO

michael.platzer@mostly.ai
https://mostly.ai/

Christoph Töglhofer

Data Scientist

christoph.toeglhofer@erstegroup.com
Erste Group IT International GmbH

george@erstegroup.com
https://george-labs.com/

Contact Details

Michael Platzer

Founder & CEO

michael.platzer@mostly.ai
https://mostly.ai/

Christoph Töglhofer

Data Scientist

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

mostly AI

YES, WE'RE HIRING

ERSTE Group