University of St. Gallen

Creation of Adversarial Accounting Records to Attack Financial Statement Audits

A research collaboration between the HSG, DFKI and PwC

NVIDIA's GPU Technology Conference

March, 20th 2019




M. Schreyer[1,2], T. Sattarov[3], B. Reimer[3], and D. Borth[1,2]

[1]University of St. Gallen, [2]German Research Center for Artificial Intelligence, and [3]PricewaterhouseCoopers

University of St.Gallen

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

pwc

Economic Crime and ERP-Systems

"The Footprint"

”49% respondents said that their organization have been victim of fraud or economic crime in the past 24 months”

“PwC’s Global Economic Survey 2018”,
encompassing data of 7.200 respondents in 123 countries

”The median loss of a single financial statement fraud case is $150,000... The Duration from the fraud perpetration till its detection was 18 months”

“ACFE’s 2016 Report to the Nations on Occupational Fraud and Abuse”,
encompassing 2.410 cases in 114 countries

Economic Crime

NEWS Dec.07.2017
Volkswagen manager jailed for 7 years in diesel scam

FINANCIAL POST March 15, 2018
Sino-Forest Corp. co-founder found guilty of fraud in $2.6 billion civil case

Bloomberg March 15, 2018, 4:28 PM GMT+1
Ex-Deutsche Bank Trader Bittar Pleads Guilty to Rate Rigging

August 25, 2017,
Billionaire Samsung heir convicted in mega bribery case

Cum-Ex Scandal
The Multibillion Euro Theft March 15, 2018

Business Today
PNB detects new fraud at Mumbai branch at heart of $2 billion banking scam

The Guardian Thu 14 Dec 2017
Odebrecht scandal: Ecuador vice-president given six years' jail

ZEIT 8. Juni 2017,

REUTERS BUSINESS NEWS MAY 1, 2015
BNP Paribas sentenced in $8.9 billion accord over sanctions violations

FINANCIAL TIMES NOVEMBER 28, 2018
Danske Bank charged over €200bn money-laundering scandal

Mar 19, 2018, THE TIMES OF INDIA
Three directors of Mumbai firm held for Rs 4,000 crore bank fraud

Bloomberg
London Investment Bankers Charged in German Cum-Ex Tax Case
By Karin Matussek
May 22, 2018, 2:21 PM GMT+2 *Updated on May 22, 2018, 7:21 PM GMT+2*

The Guardian edition Tue 31 Jan 2017
Deutsche Bank fined $630m over Russia money laundering claims

FORTUNE
Former HSBC Executive Mark Johnson Found Guilty of Fraud in $3.5 Billion Currency Trade

REUTERS MARCH 5, 2018
Kobe Steel admits data fraud went on nearly five decades, CEO to quit

The New York Times SEPT. 8, 2016
Wells Fargo Fined $185 Million for Fraudulently Opening Accounts

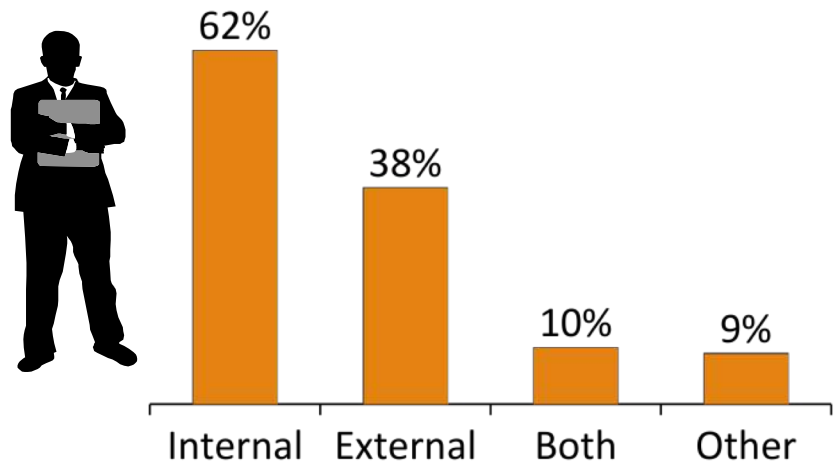
THE WALL STREET JOURNAL. March 15, 2018
Former Siemens Executive Pleads Guilty in Argentina Bribery Case

JANUARY 14, 2018 FINANCIAL TIMES
UK fraud hits 15-year high with value of £2bn

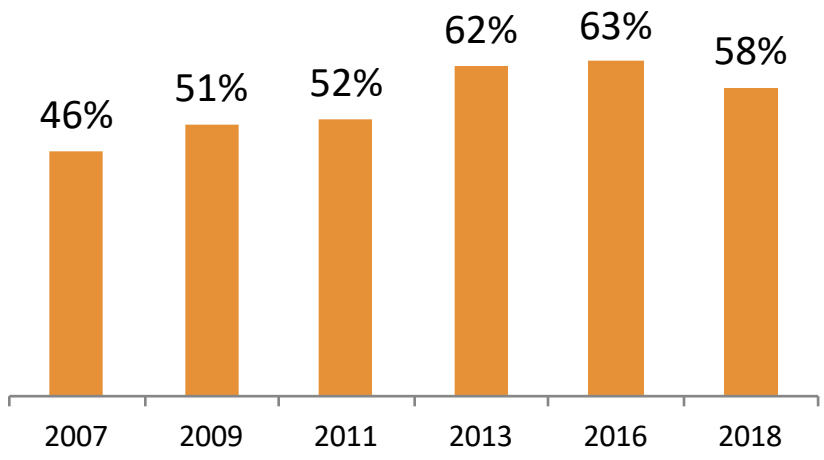
THE LOCAL fr 16 March 2018
Massive Côte du Rhône fine-wine fraud uncovered by French police

Economic Crime Committed by Internal Actors

Relationship of Actor and Victimized Organization*



Fraction of Internal Actors Conducting Economic Crime**

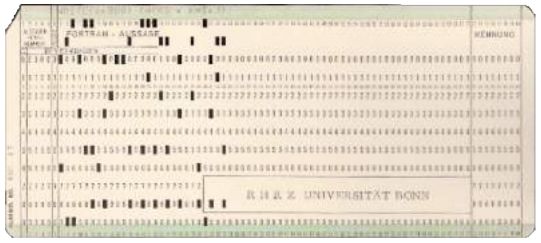


"Internal actors are the main the main perpetrators of fraud."

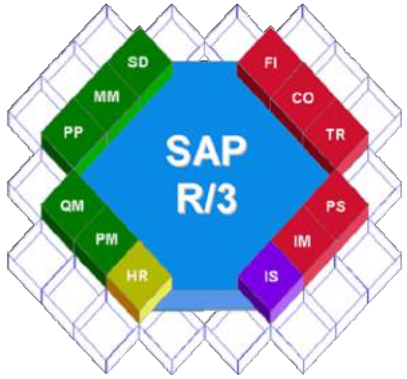
* Source: „Wirtschaftskriminalität 2018, Mehrwert von Compliance - forensische Erfahrungen", Studie der Martin-Luther-Universität Halle Wittenberg und PwC GmbH WPG
** Source: „Wirtschaftskriminalität in der analogen und digitalen Wirtschaft 2016", Studie der Martin-Luther-Universität Halle Wittenberg und PwC GmbH WPG
** Source: „Wirtschaftskriminalität und Unternehmenskultur 2013", Studie der Martin-Luther-Universität Halle Wittenberg und PwC GmbH WPG

Evolution of Recording and Processing Accounting Data

~ 1900's



~ 1950's



~ 1992's

Data Volume

- Continuous digitization of business activities and processes
- Accumulation of exhaustive transactional and business process data
- „Every" activity within an organization leaves a **digital trace** .... !

Evolution of Recording and Processing Accounting Data

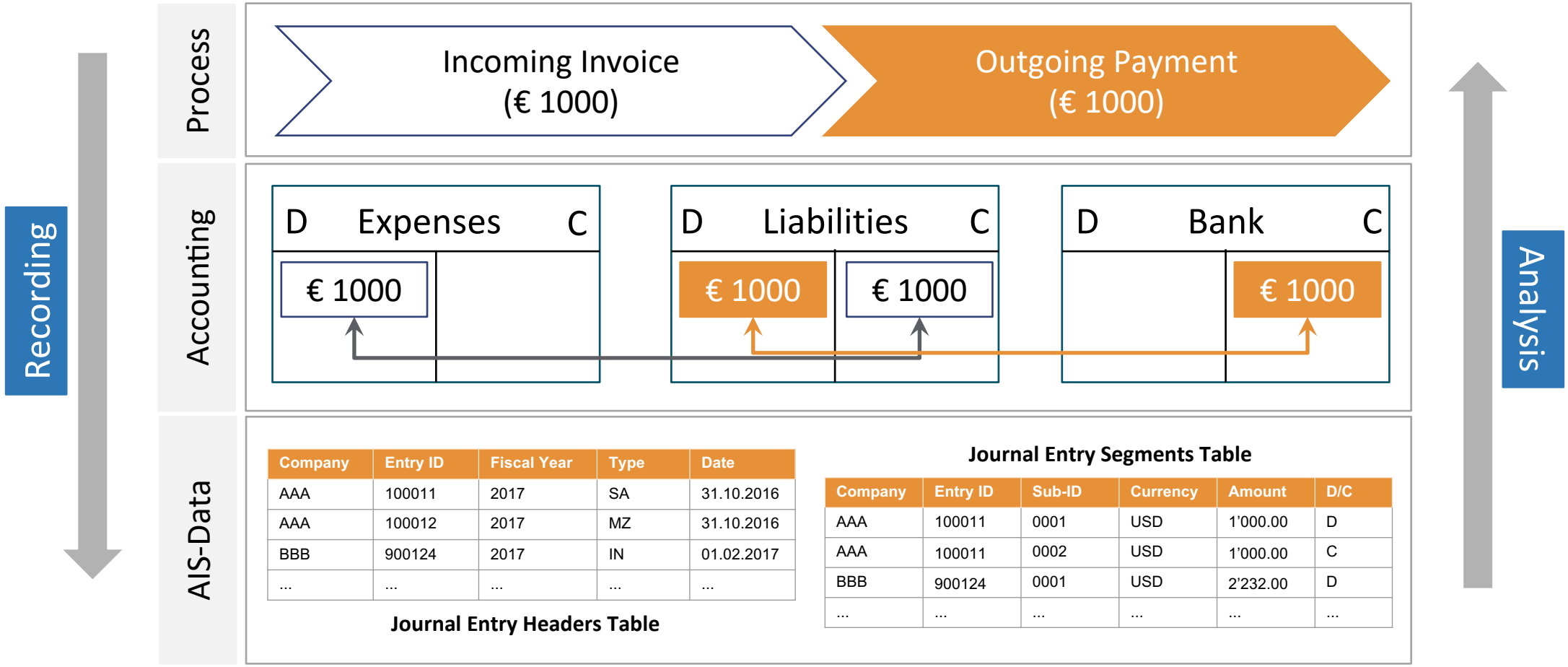
SAP AG:

”Our ERP applications touch 77% of global transaction revenue […]"

Source: “SAP at a Glance - Investor Relations Fact Sheet (October 2018)”, https://www.sap.com/docs/download/investors/2018/sap-factsheet-oct2018-en.pdf

- Continuous digitization of business activities and processes
- Accumulation of exhaustive transactional and business process data
- „Every“ activity within an organization leaves a **digital trace** …. !

Understanding the Different Layers of Abstraction



Company	Entry ID	Fiscal Year	Type	Date
AAA	100011	2017	SA	31.10.2016
AAA	100012	2017	MZ	31.10.2016
BBB	900124	2017	IN	01.02.2017
...	...	...	...	...

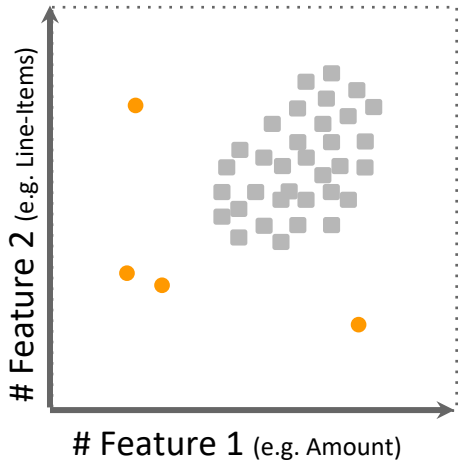
Journal Entry Headers Table

Journal Entry Segments Table

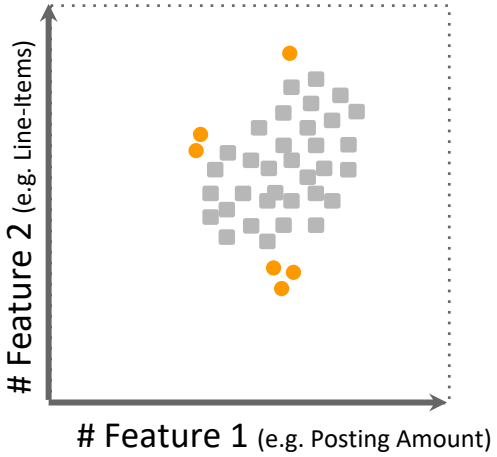
Company	Entry ID	Sub-ID	Currency	Amount	D/C
AAA	100011	0001	USD	1'000.00	D
AAA	100011	0002	USD	1'000.00	C
BBB	900124	0001	USD	2'232.00	D
...	...	...	...	...	...

Classification of Accounting Anomalies

„Global“ Accounting Anomalies



„Local“ Accounting Anomalies



Usually Rare Attribute Values

- Seldom used user accounts,
- Reverse postings, corrections

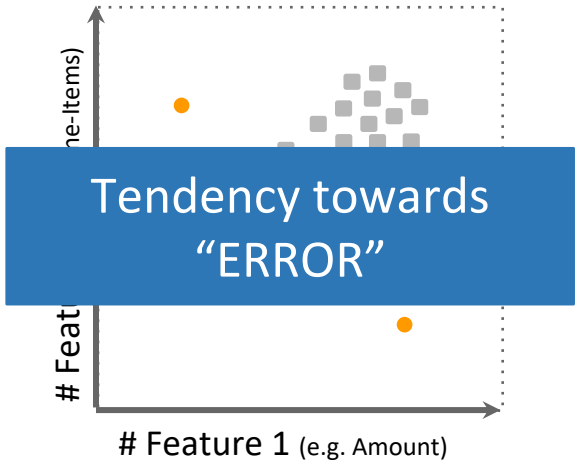
Usually Rare Attribute Combinations

- Unusual posting activities
- Deviating user behavior

[1] Kriegel et al., 2000

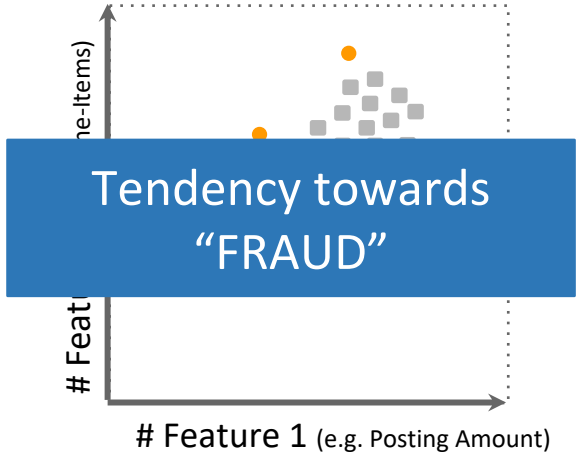
Classification of Accounting Anomalies

„Global“ Accounting Anomalies



*"Perpetrators usually don't act **completely in deviation** from the usual accounting models.”*

„Local“ Accounting Anomalies



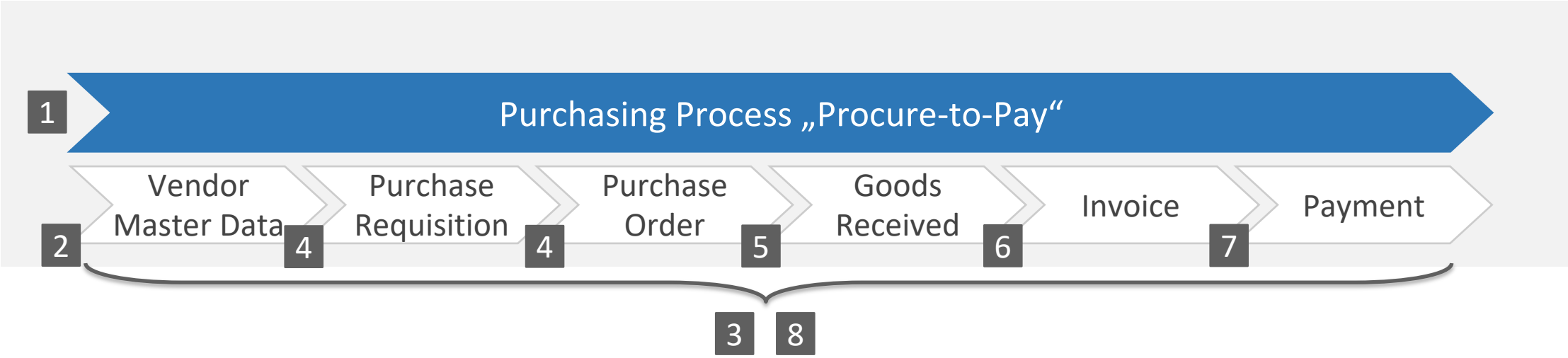
*"Perpetrators usually try to **obfuscate their behavior** to make it appear as ordinary as possible.”*

[1] Kriegel et al., 2000

Traditional “Red-Flag” Approaches

Matching Fraud Signatures

Exemplary “Red-Flags” to Detect Traces of Fraudulent Activities



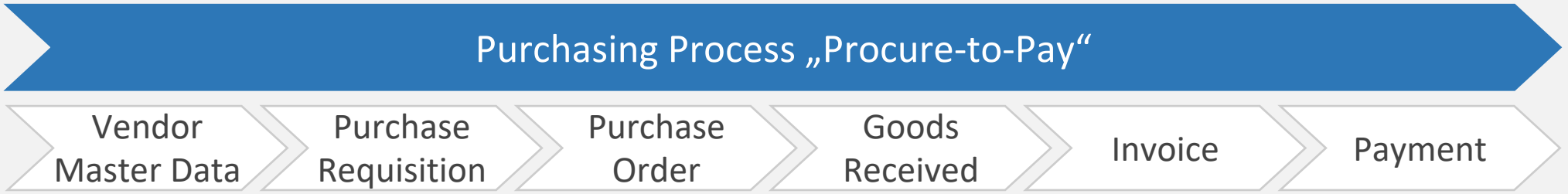
2 Vendor Master Data Analysis

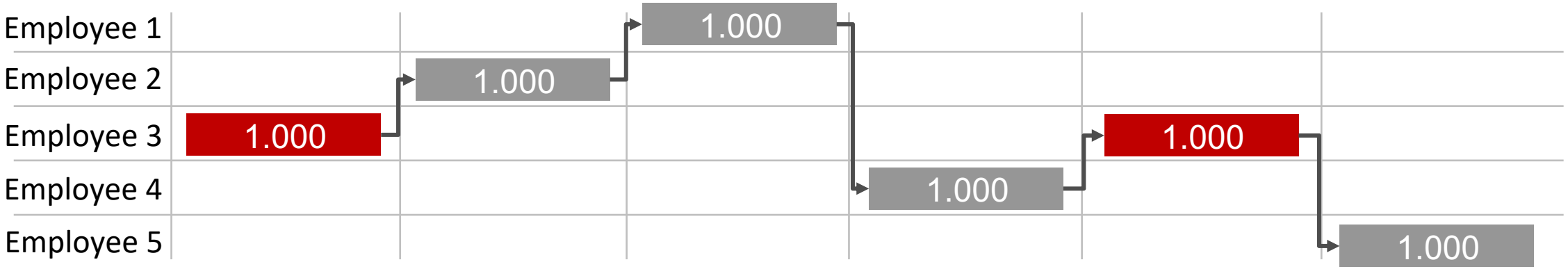
- Uncomplete vendor master data
- Short-term bank account changes
- Sanctioned or one-time vendors
- Multiple bank accounts
- …

7 Vendor Invoice Analysis

- Invoices without purchase order
- Multiple re-postings of invoices
- Short time period of invoice clearance
- Re-recorded invoice after payments
- …

Exemplary “Red-Flags” to Detect Traces of Fraudulent Activities



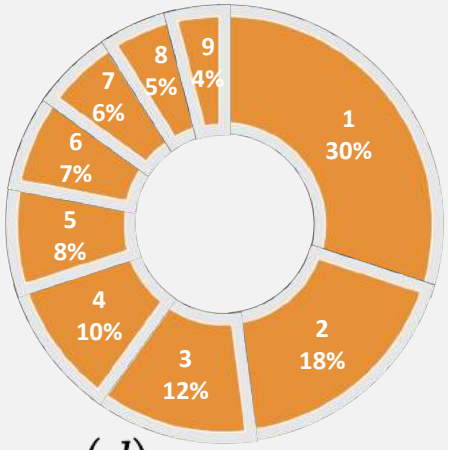


Segregation of Duties (SoD) Matrix per Process Activity

Exemplary: Distribution Analysis of Purchase Order Amounts

Benford-Newcomb Law

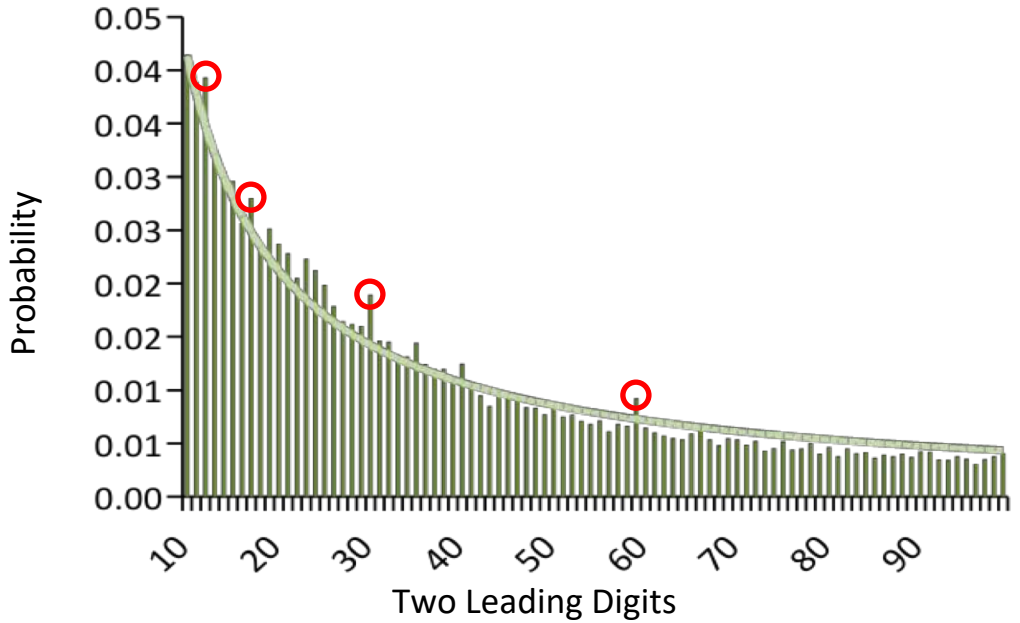
- Formalizes the uneven distribution of the leading digits in many real-life sets of numerical data



$$p(d) = \log_B\left(1 + \frac{1}{d}\right) = \log_B(d+1) - \log_B(d)$$

- Trace for the potential circumvention of financial approval limits (e.g. purchase orders)

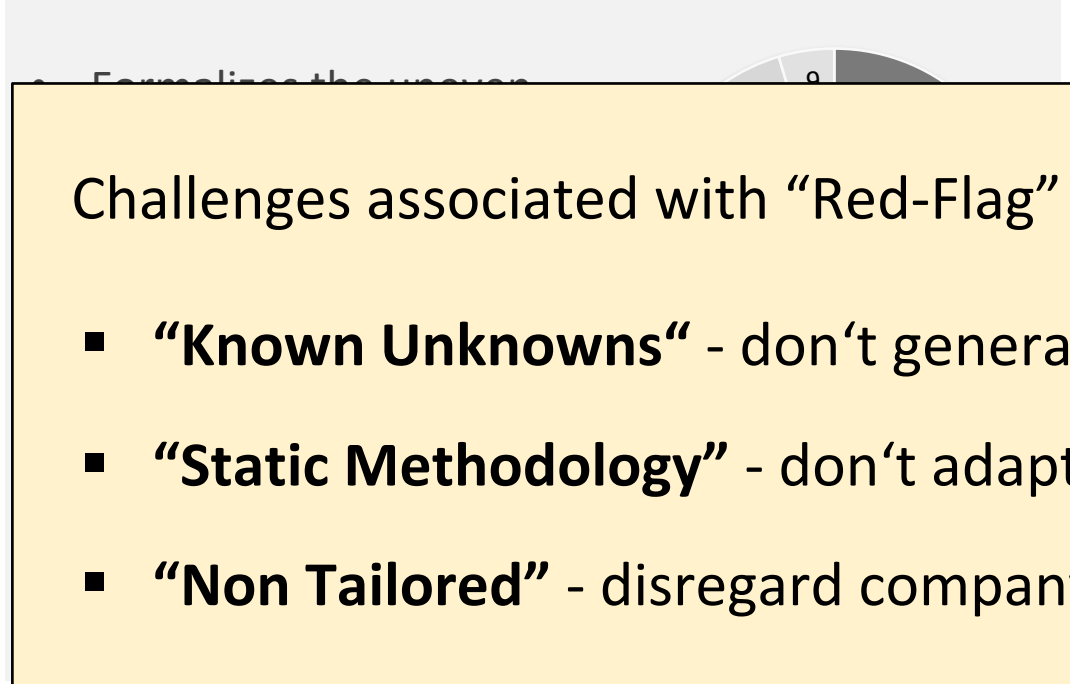
Analysis of Vendor Purchase Order Amounts



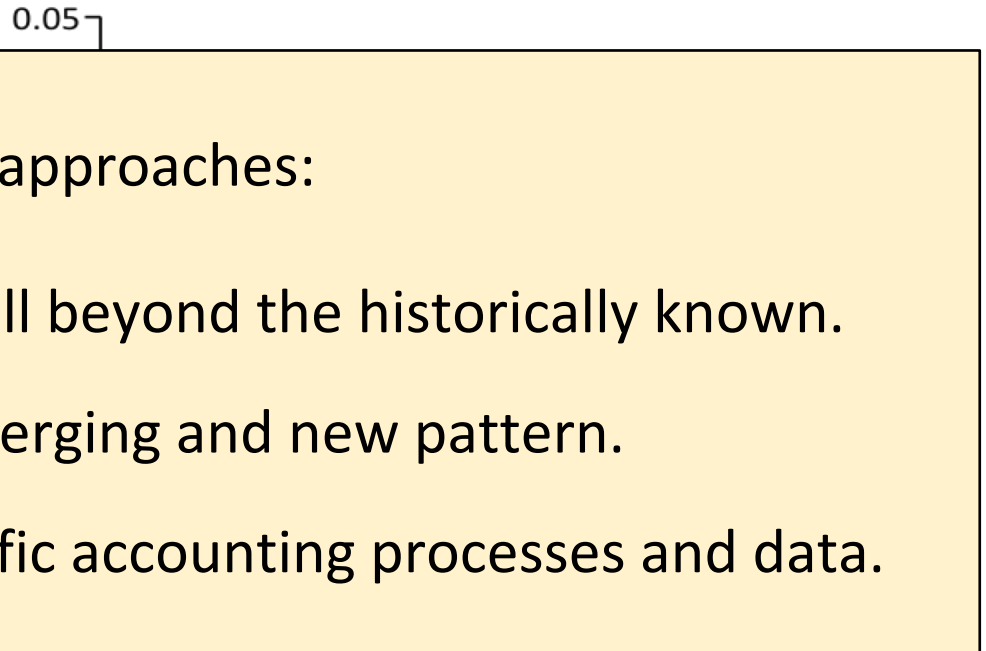
[2] Benford, Frank; 2000

Exemplary: Distribution Analysis of Purchase Order Amounts

Benford-Newcomb Law



Analysis of Vendor Purchase Order Amounts



Challenges associated with “Red-Flag” based approaches:

- **“Known Unknowns“** - don‘t generalize well beyond the historically known.
- **“Static Methodology”** - don‘t adapt to emerging and new pattern.
- **“Non Tailored”** - disregard company specific accounting processes and data.

[2] Benford, Frank; 2000

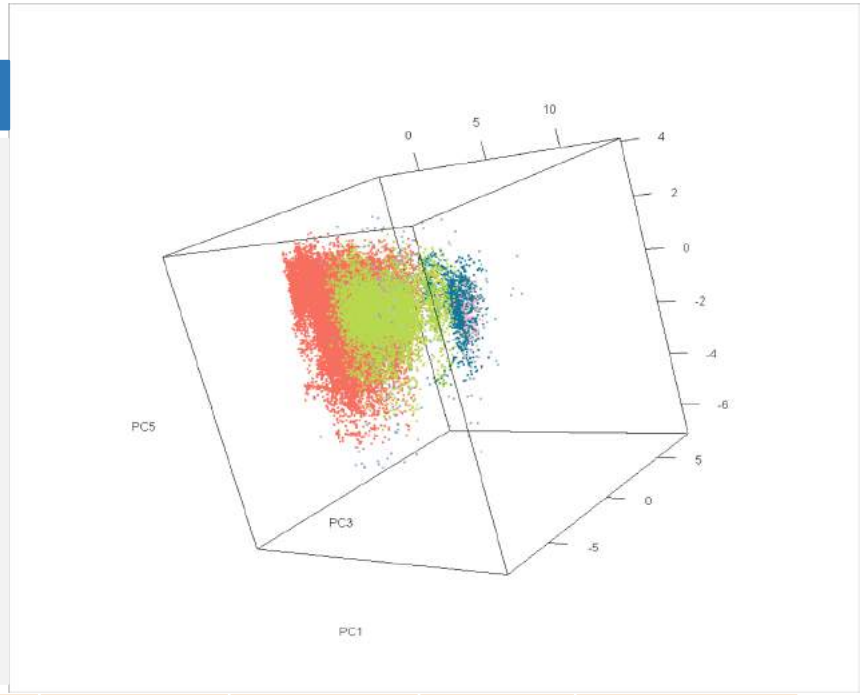
Traditional ”Data Science” Approaches

Principal Component Analysis & Clustering

Example: Multi-Dimensional Clustering of Vendor Payments

Multi-Dimensional Cluster Detection

- Exemplary analysis of SAP vendor payments:
  - Total 125.223 payment postings
  - Affecting 22 SAP-User, 3.055 Vendors
- Detected “regular” clusters:
  - Man. vendor payments („Cluster 1“)
  - Employee travel expenses („Cluster 2“)
  - Periodic payment runs („Cluster 3“)

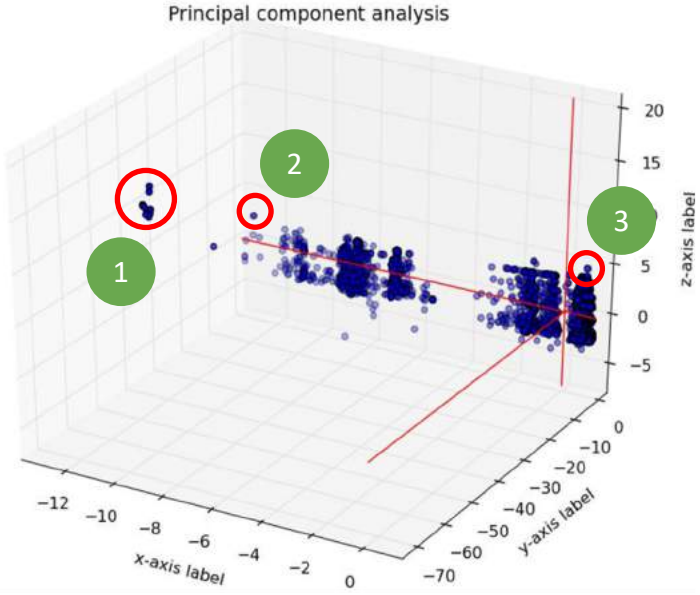


Cluster	GJAHR	BELNR	BUZEI	USNAM	BLART	TCODE	HKONT	DMBTR	LIFNR	CPUDT
1	2014	30801256	2	User A	MP	FB05	460200	2‘970.00	437970	08/18/2014
2	2014	60700394	2	User B	TR	FB1K	440000	559.68	356710	10/19/2014
3	2014	80300928	1	User C	PR	F110	440000	4‘974.2	609406	01/19/2014

Example: Multi-Dimensional Clustering of Vendor Payments

Multi-Dimensional Anomaly Detection

- Exemplary analysis of SAP vendor payments:
  - Total 125.223 payment postings
  - Affecting 22 SAP-User, 3.055 Vendors
- Detected posting anomalies:
  - Deviating man. vendor payments („Cluster 1“)
  - Late employee travel expenses („Cluster 2“)
  - Manipulated payment runs („Cluster 3“)

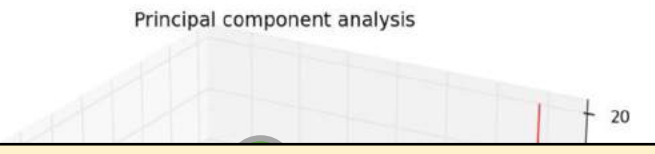


Anomaly	GJAHR	BELNR	BUZEI	USNAM	BLART	TCODE	HKONT	DMBTR	LIFNR	CPUDT
1	2014	31000007	4	User Z	MP	FBZ2	486400	14672.85	209495	01/01/2014
2	2014	60801008	2	User Y	TR	FB1K	440000	17123.98	358822	06/28/2014
3	2014	80600094	17	User C	PR	F110	440000	45376.69	364110	04/07/2014

Example: Multi-Dimensional Clustering of Vendor Payments

Multi-Dimensional Anomaly Detection

- Exemplary analysis of SAP vendor payments:



Challenges associated with traditional DS based approaches:

- **“Feature Engineering“** - difficulty to design and select relevant features.
- **“Curse of Dimensionality”** - computational complexity of the algorithms.
- **“Model Complexity”** - hurdle to model non-linear attribute relationships.

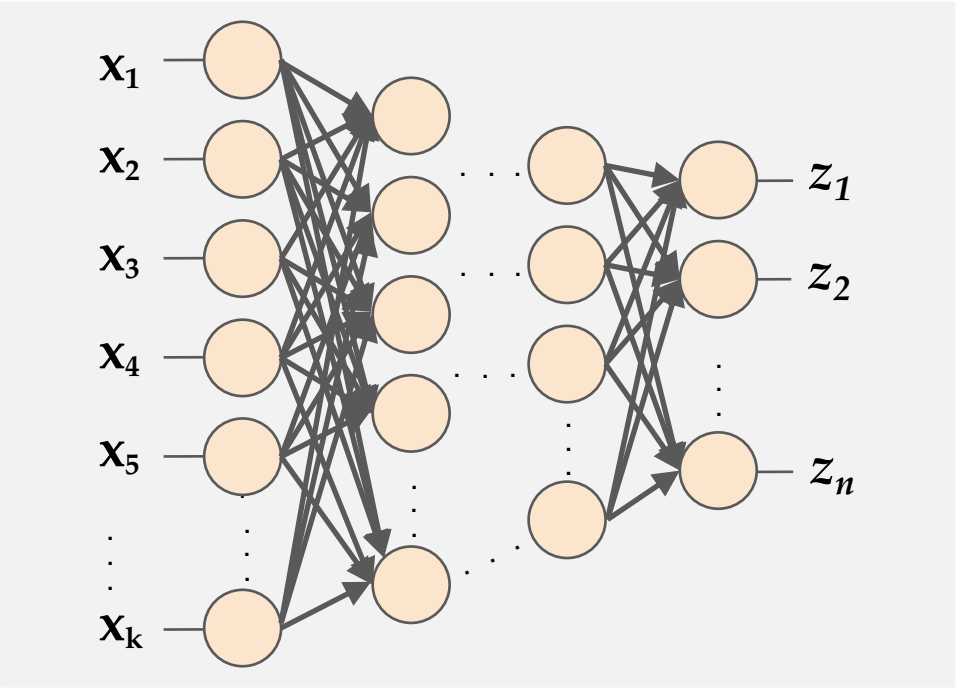
2	2014	60801008	2	User Y	TR	FB1K	440000	17123.98	358822	06/28/2014
3	2014	80600094	17	User C	PR	F110	440000	45376.69	364110	04/07/2014

“End-to-End Learning” Approaches

Autoencoder Neural Networks

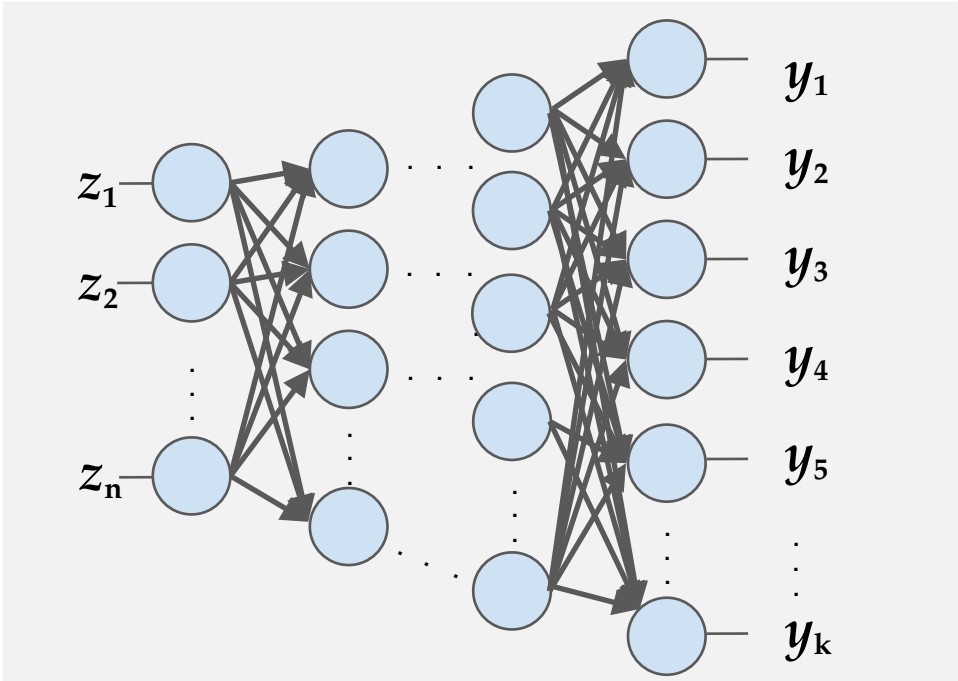
Autoencoder NNs - Network Building Blocks

Encoder-Net
Non-linear "Compression"



$$f_\Theta^l = \sigma(W f_\Theta^{l-1} + b)$$

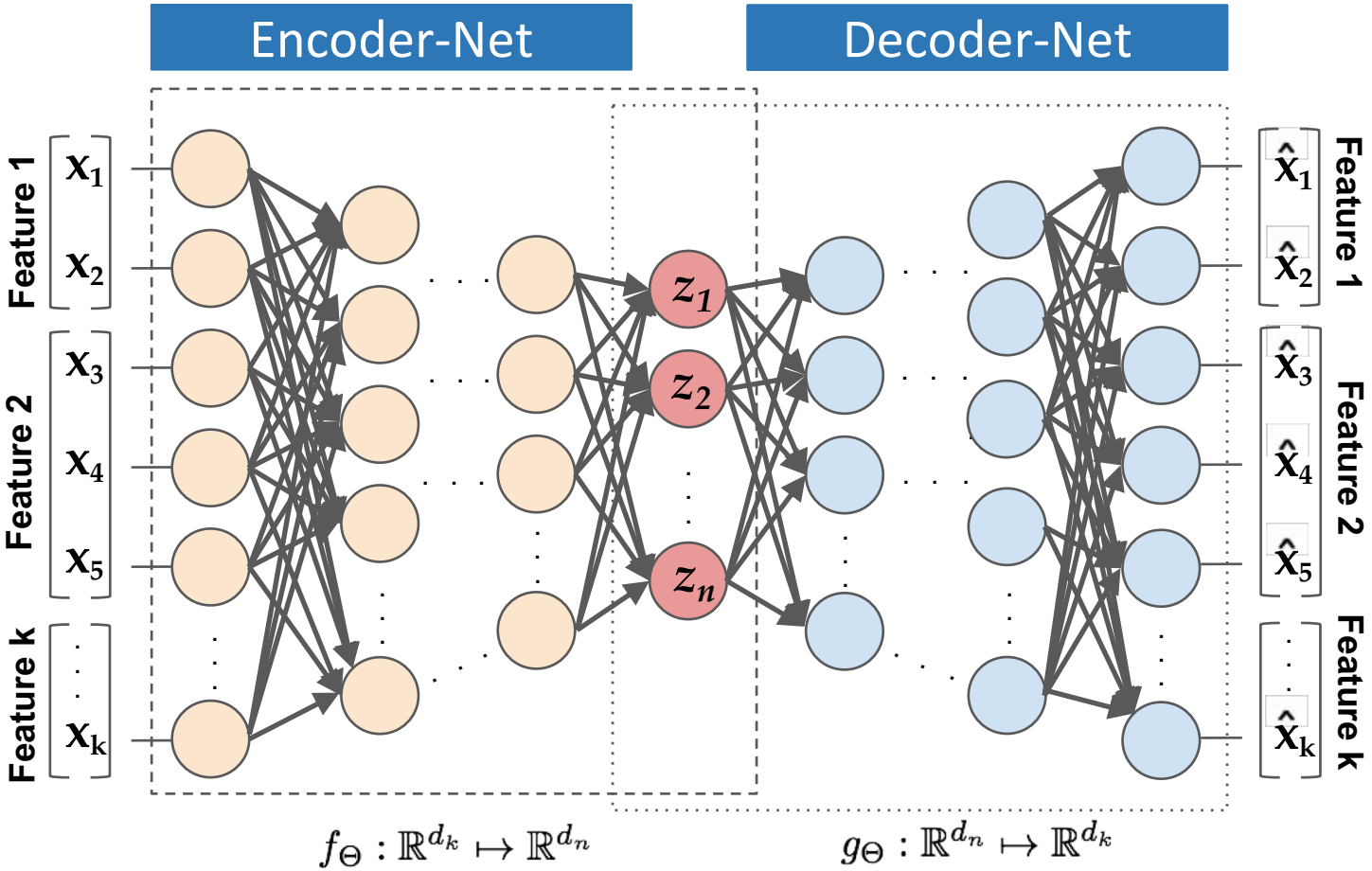
Decoder-Net
Non-linear "Reconstruction"



$$g_\Theta^l = \sigma(W' g_\Theta^{l-1} + b')$$

Autoencoder NNs - Network Building Blocks

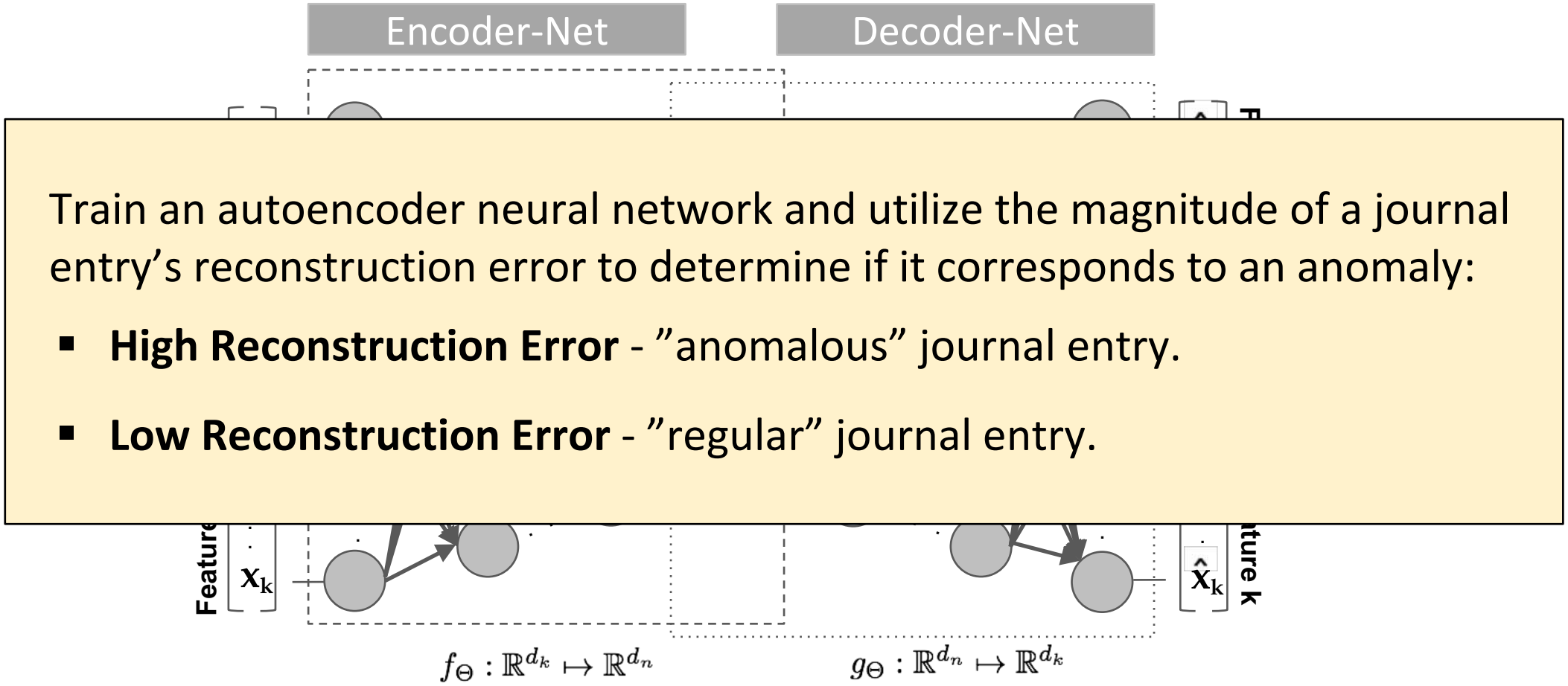
x



$\hat{x}$

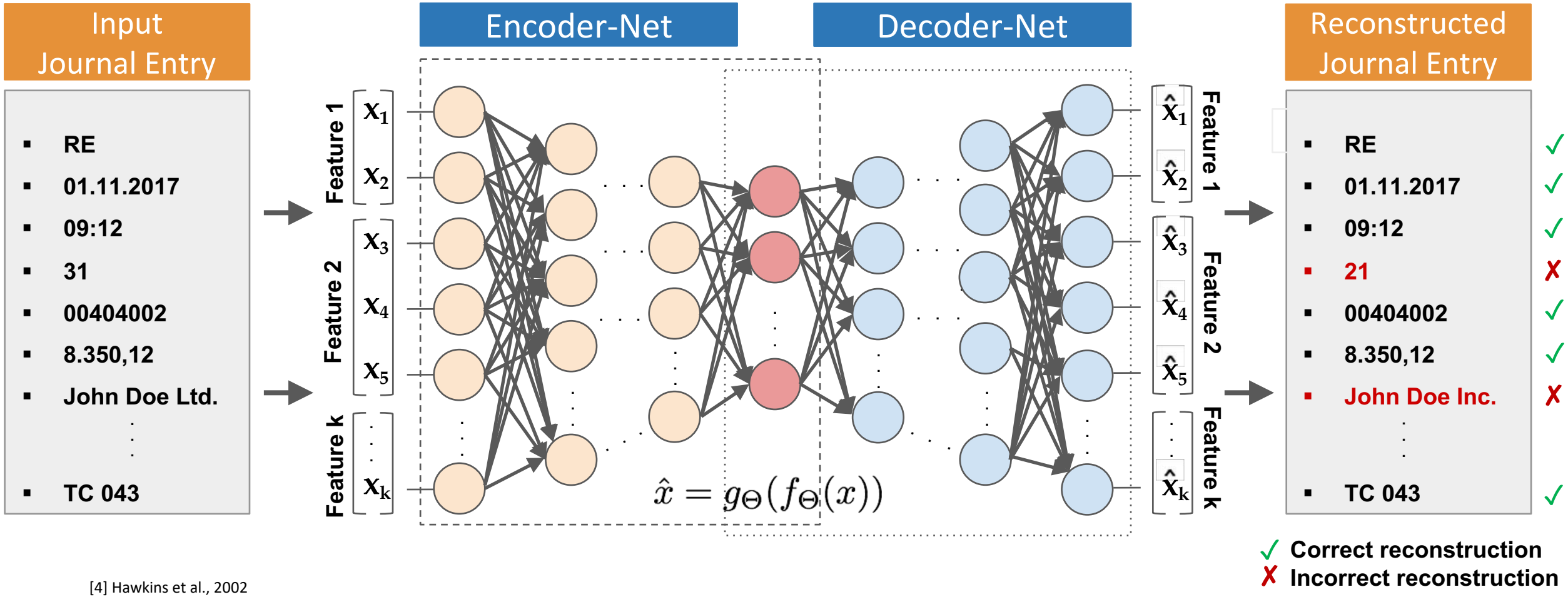
[3] Hinton, G. and Salakhutdinov, 2006

Autoencoder NNs - Network Building Blocks



[3] Hinton, G. and Salakhutdinov, 2006

Autoencoder NNs - Anomaly Detection



[4] Hawkins et al., 2002

Autoencoder NNs - Experimental Setup

Journal Entry Header

MANDT	BUKRS	BELNR	GJAHR	BLART	BLDAT	BUDAT	CPUDT	TCODE
903	1000	0100000001	2011	SA	28.02.2011	28.02.2011	02.03.2011	FB01
903	0005	0006000000	2011	KN	30.06.2011	01.07.2011	05.07.2011	FB08
903	1000	0500000003	2011	SA	04.09.2011	04.09.2011	04.09.2011	FB01
…			…				…	
…			…				…	
…			…				…	

Journal Entry Segment

MANDT	BUKRS	BELNR	GJAHR	BUZEI	SHKZG	DMBTR	HKONT
903	1000	0100000001	2011	001	S	734,45	0000100000
903	1000	0100000001	2011	002	S	100,07	0000399999
903	1000	0100000001	2011	003	H	450,40	0000113100
903	1000	0100000001	2011	004	H	384,12	0000473100

SAP FI Real World Accounting Datasets

Accounting Dataset A (anonymized):

- 307’457 journal entry line items (single FY)
- 8 attributes, 401 “one-hot” encoded dimensions
- 55 (0.016%) “global” anomalies („rare values“)
- 40 (0.015%) “local” anomalies („rare combinations“)

Accounting Dataset B (anonymized):

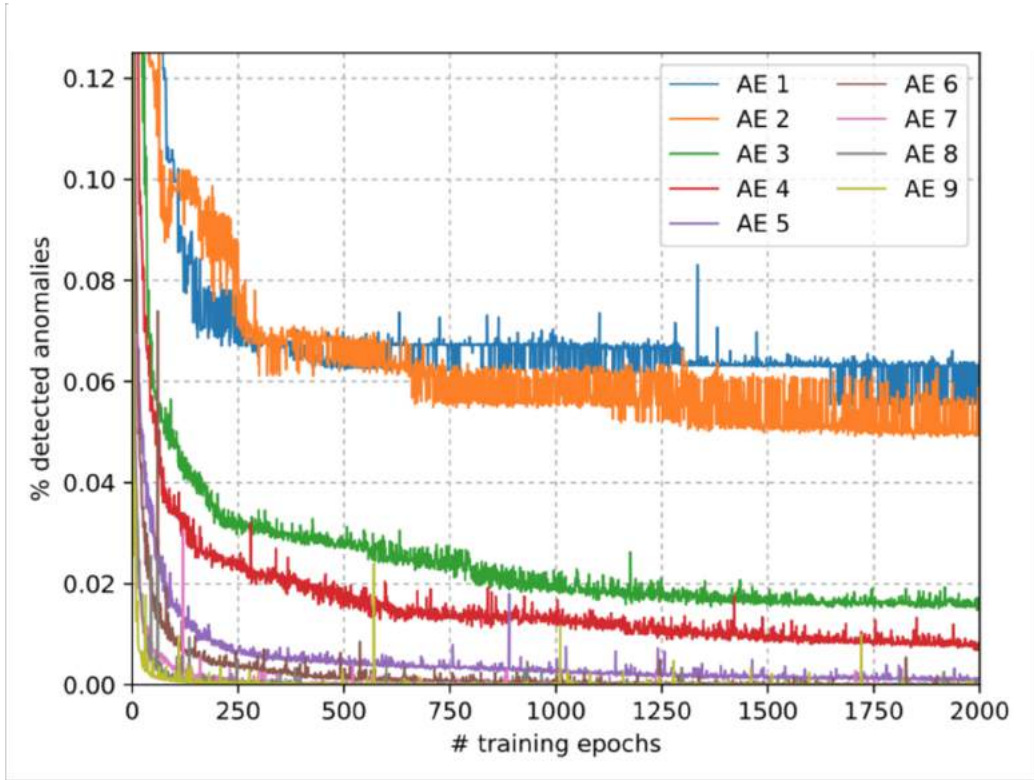
- 172’990 journal entry line items (single FY)
- 10 attributes, 576 “one-hot” encoded dimensions
- 50 (0.030%) “global” anomalies („rare values“)
- 50 (0.030%) “local” anomalies („rare combinations“)

→ Highly unbalanced class distribution!

[5] Schreyer, Sattarov et al., 2017

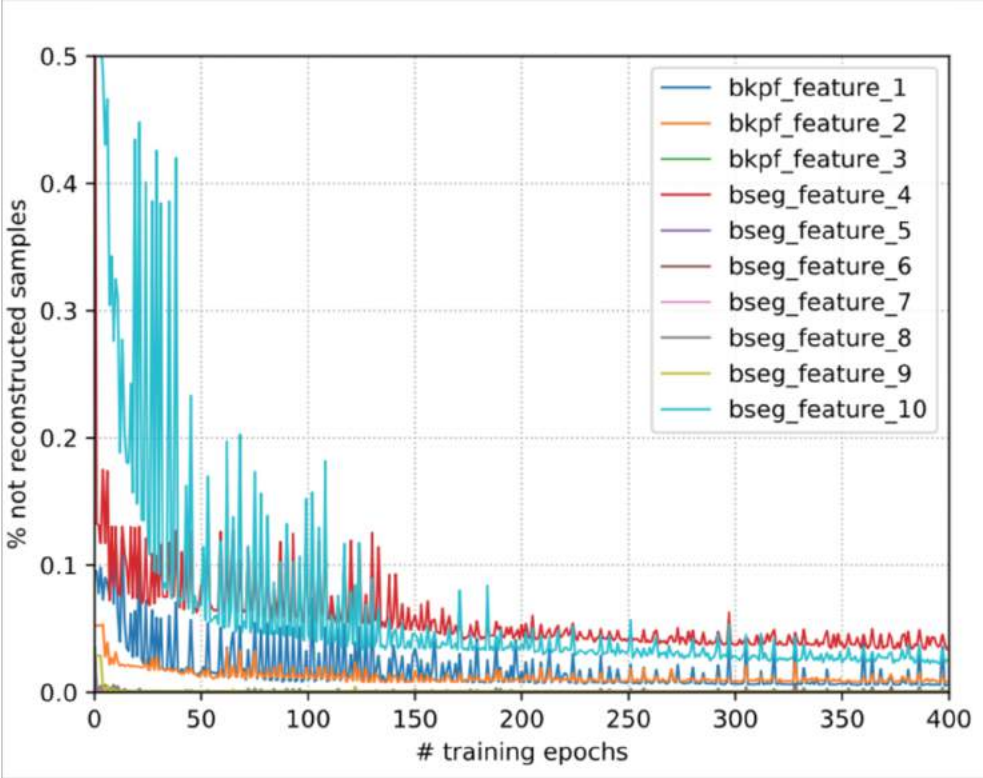
Autoencoder NNs - Training Process

AE Architecture Training Performance



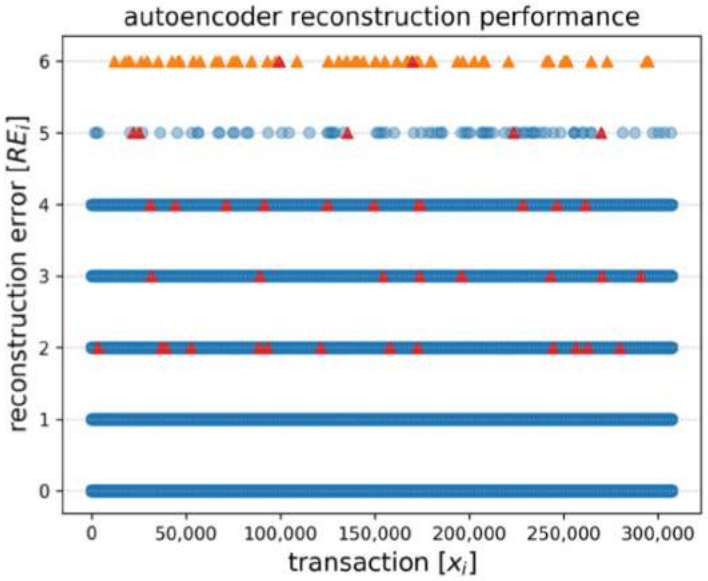
[5] Schreyer, Sattarov et al., 2017

AE Feature Learning

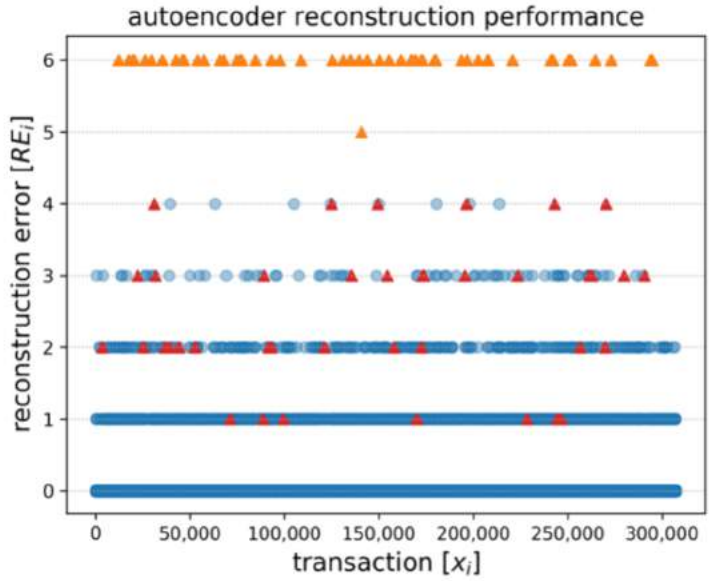


Autoencoder NNs - Training Process II

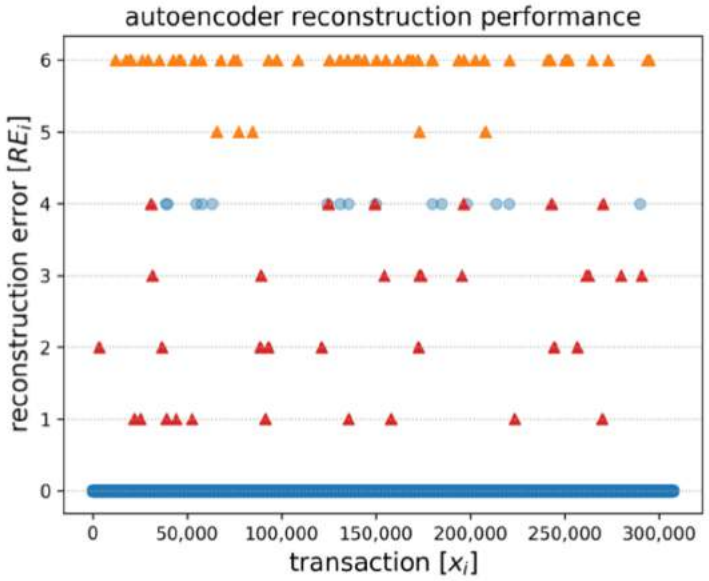
10 Training Epochs



100 Training Epochs



400 Training Epochs



Transaction class | Regular | Global Anomalies | Local Anomalies

[5] Schreyer, Sattarov et al., 2017

Autoencoder NNs - Training Results

Quantitative Evaluation – Dataset A

Model	Data	Precision	F_1-Score	Top-k	Anomalies [%]	Anomalies [#]
AE 1	A	0.0049	0.0098	0.0049	6.26	19'233
AE 2	A	0.0063	0.0126	0.0063	4.87	14'966
AE 3	A	0.0098	0.0194	0.6632	3.16	9'719
AE 4	A	0.0290	0.0564	0.7684	1.07	3'275
AE 5	A	0.0641	0.1204	0.6632	0.48	1'483
AE 6	A	0.0752	0.1398	0.5263	0.41	1'264
AE 7	A	0.0796	0.1474	0.7895	0.39	1'194
AE 8	A	0.1201	0.2144	0.5684	0.26	791
AE 9	**A**	**0.1971**	**0.3293**	**0.6947**	**0.16**	**482**

Quantitative Evaluation – Dataset B

Model	Data	Precision	F_1-Score	Top-k	Anomalies [%]	Anomalies [#]
AE 1	B	0.0020	0.0040	0.0020	0.2884	49'897
AE 2	B	0.0030	0.0059	0.0030	0.1952	33'762
AE 3	B	0.0052	0.0104	0.6200	0.1104	19'102
AE 4	B	0.0076	0.0150	0.7300	0.0765	13'238
AE 5	B	0.0087	0.0173	0.7400	0.0662	11'444
AE 6	B	0.0251	0.0489	0.6100	0.0230	3'986
AE 7	B	0.0268	0.0522	0.6400	0.0215	3'735
AE 8	B	0.0197	0.0387	0.6700	0.0293	5'070
AE 9	**B**	**0.0926**	**0.1695**	**0.4200**	**0.0062**	**1'080**

Qualitative Evaluation

- Detailed review and analysis of journal entries that result in a high reconstruction error.
- Review conducted in a joint effort with PwC's Certified Public Accountants ("Wirtschaftsprüfer").

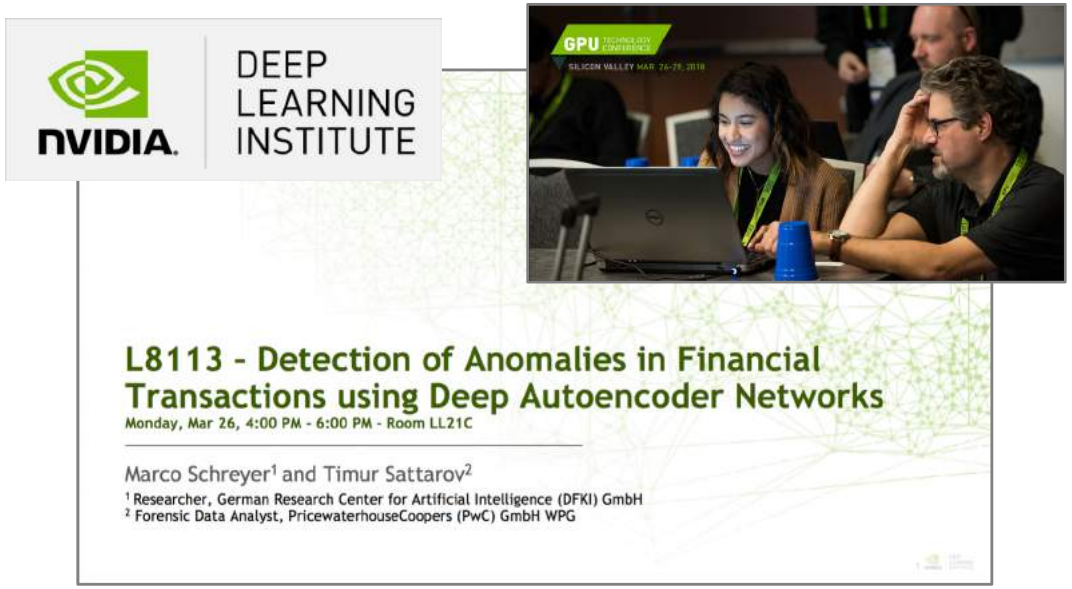
- **"Global" Anomaly Evaluation:**
  - Mostly posting errors (wrongly used GL accounts);
  - Incomplete information (tax codes, currencies).

- **"Local" Anomaly Evaluation:**
  - Cross company code shipments postings;
  - Unknown / irregular rental payments.

- **Observations revealed weak control environments!**

GTC Silicon Valley 2018

Nvidia Deep Learning Institute



Lab content jointly developed with the support of NVIDIA's DLI team Kelvin Levin, Onur Yilmaz and Patrick Hogan.

Nvidia Deep Learning Blog

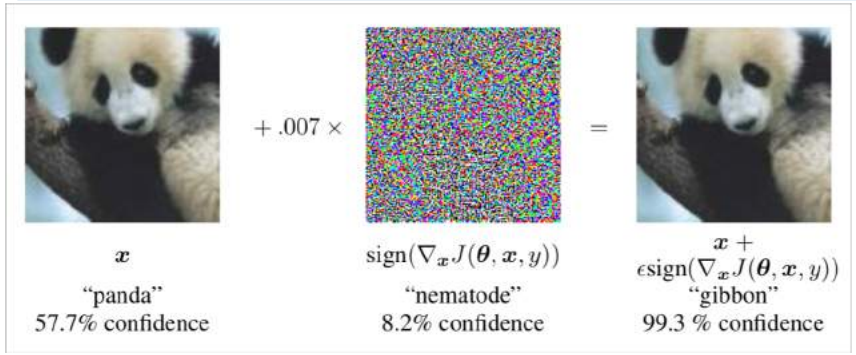
https://blogs.nvidia.com/blog/2018/04/16/finding-fraud/

Adversarial Approaches

Adversarial Autoencoder Networks

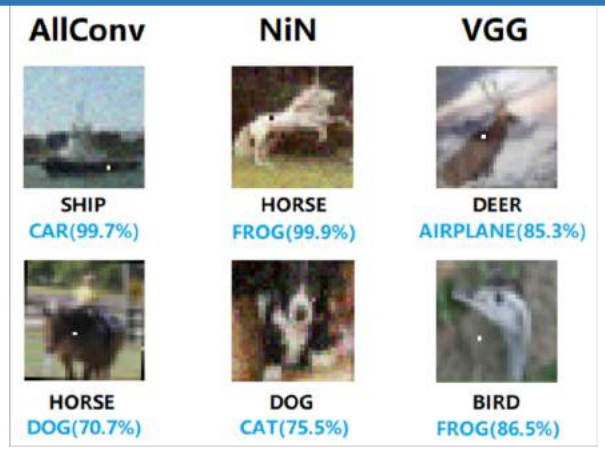
Adversarial Attacks

Goodfellow et al., 2014

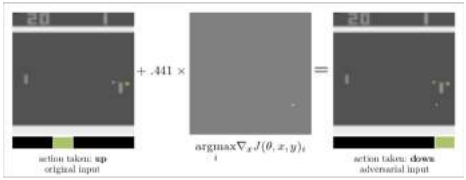


Evtimov et al., 2014

Su et al., 2017



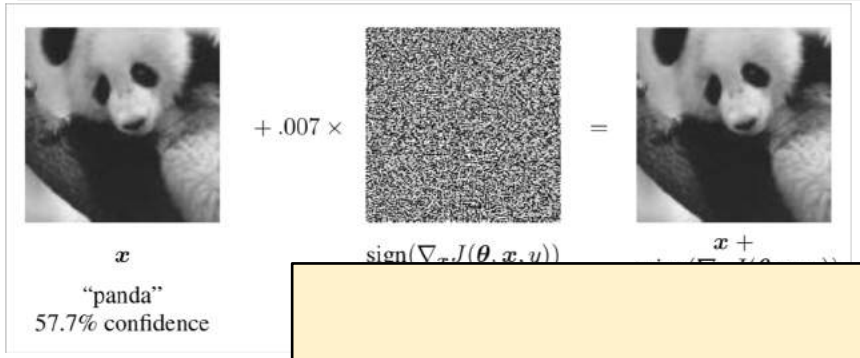
Huang et al., 2017



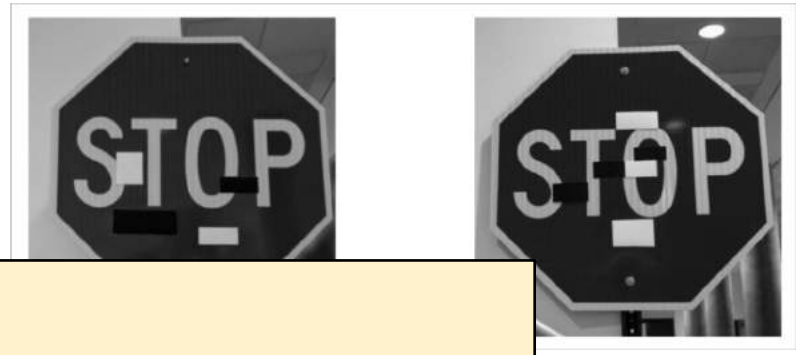
Alzantot et al., 2018

Adversarial Attacks

Goodfellow et al., 2014

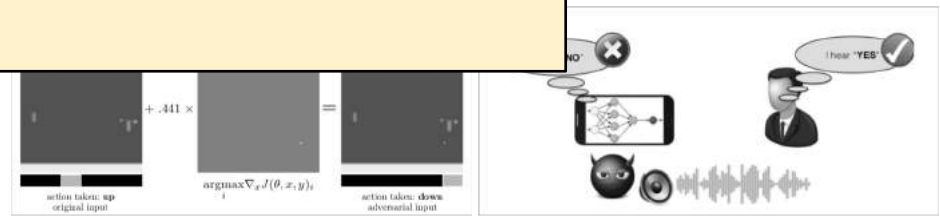


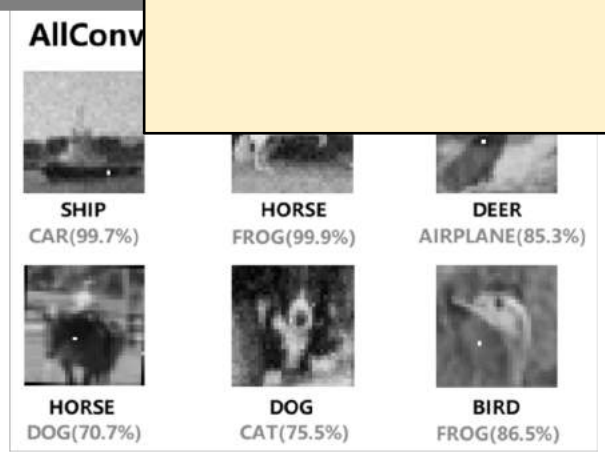
Evtimov et al., 2014



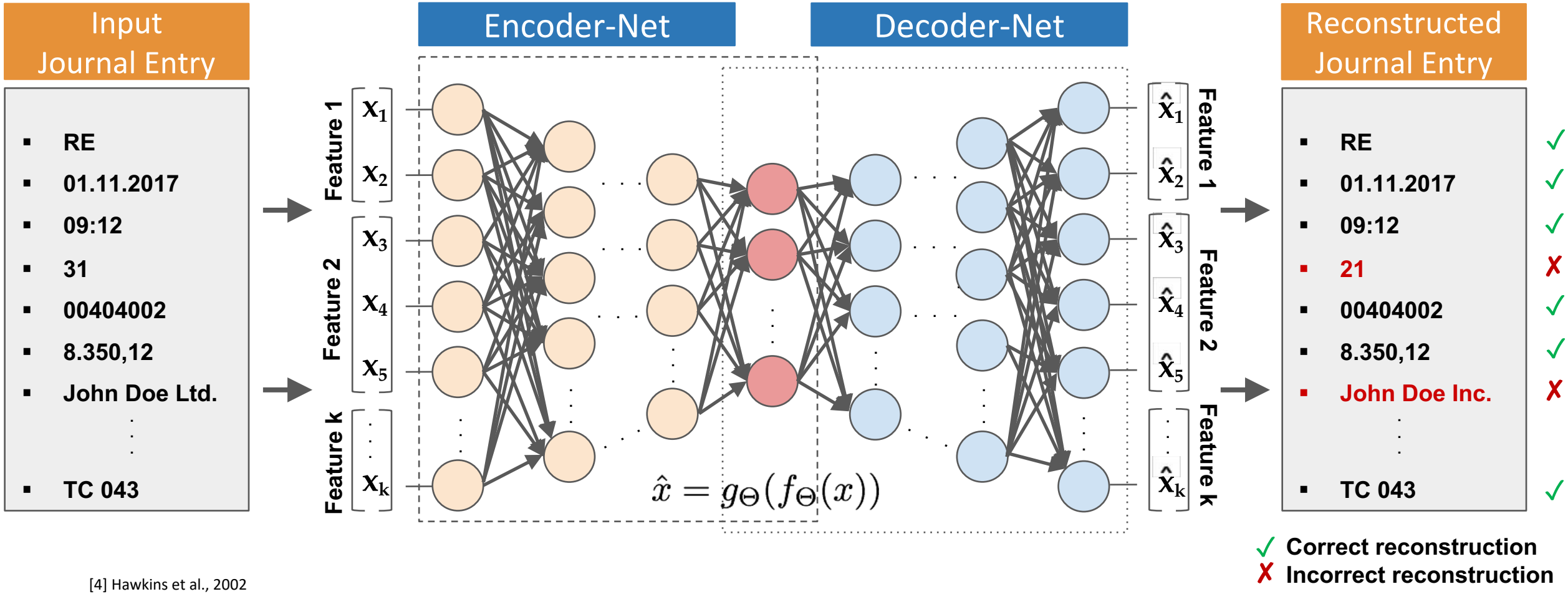
How is this relevant for the audit practice?

Ilzantot et al., 2018



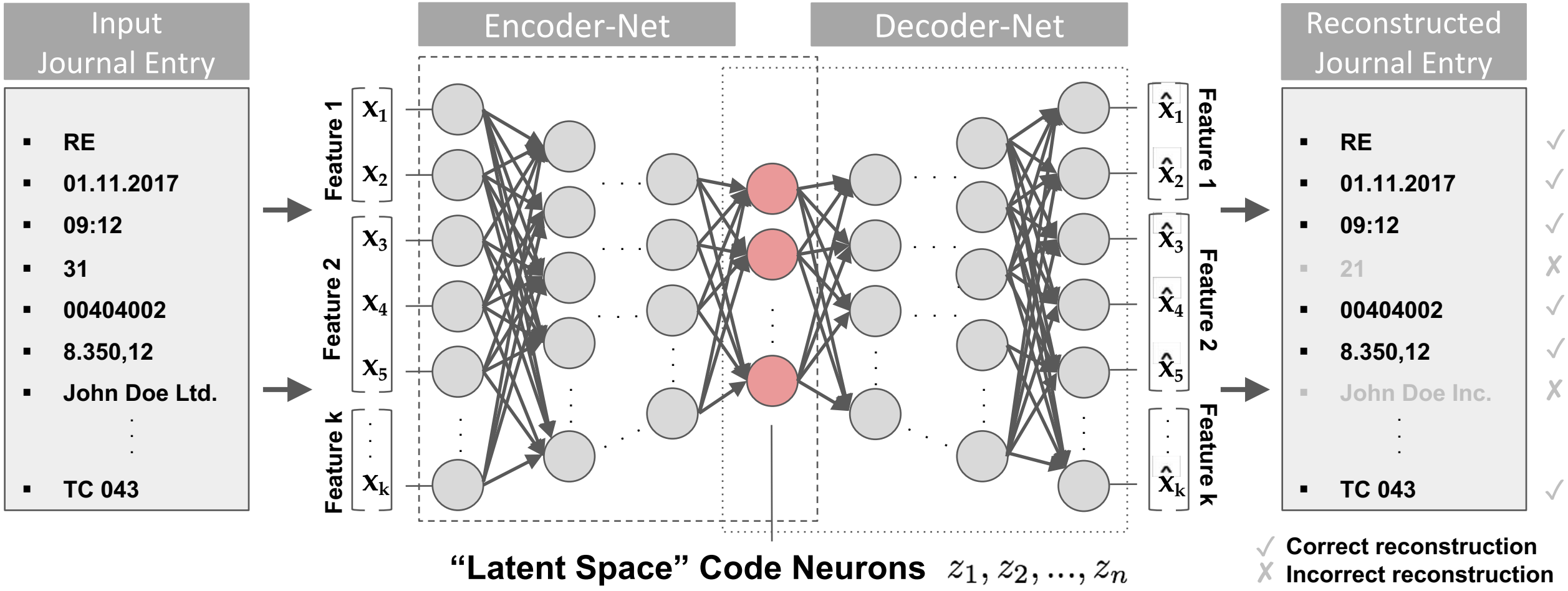


Autoencoder NNs - Latent Space Analysis



[4] Hawkins et al., 2002

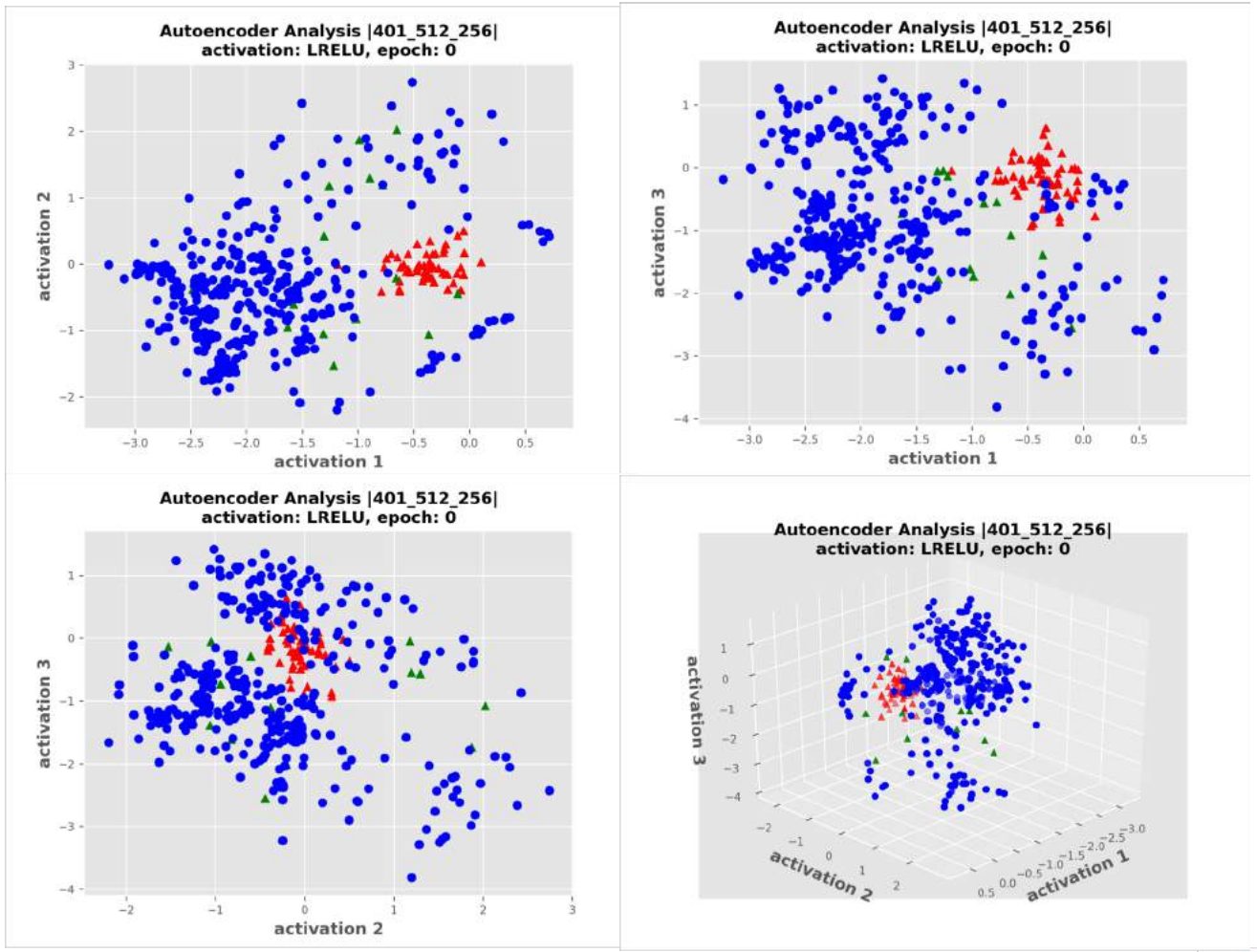
Autoencoder NNs - Latent Space Analysis



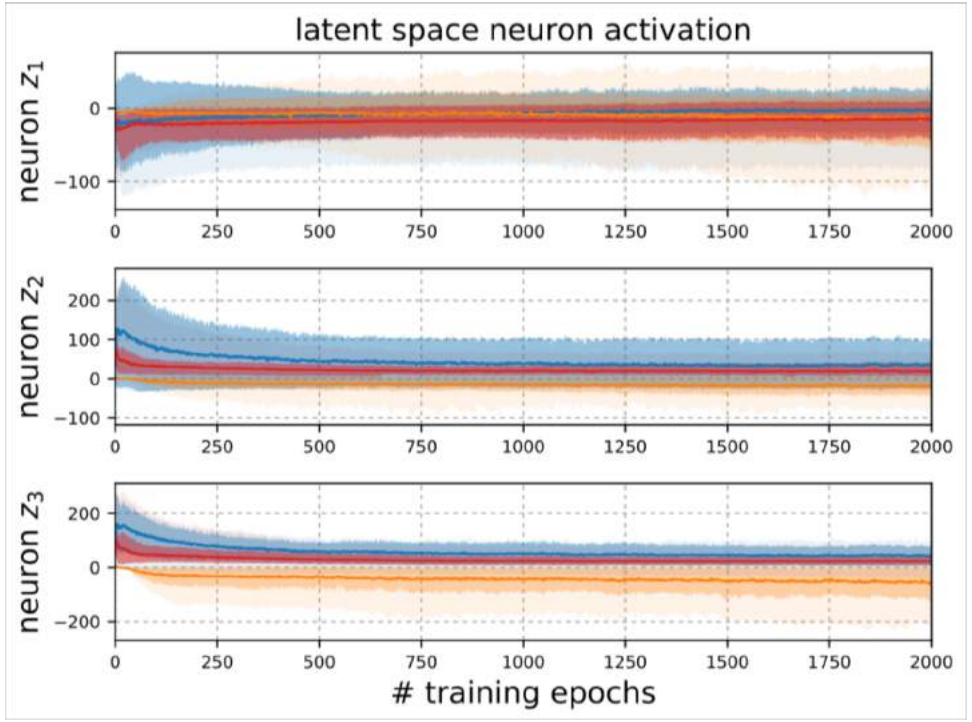
Autoencoder NNs - Latent Space Analysis I

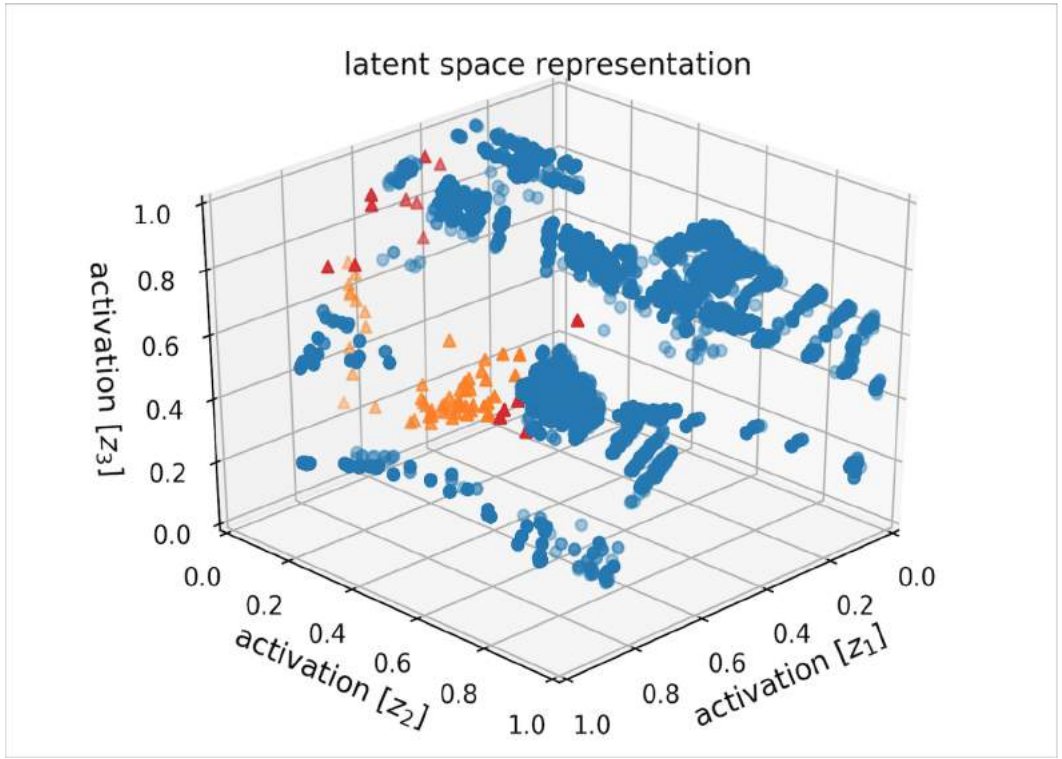
Latent Space Analysis

- Visualization of single neuron activations with progressing training.
- The Autoencoder learns a distinctive **“activation pattern” or manifold** for class of each journal entries.
- Distinct Journal Entry Classes:
  - Regular Journal Entry
  - “Global” Anomalies
  - “Local” Anomalies



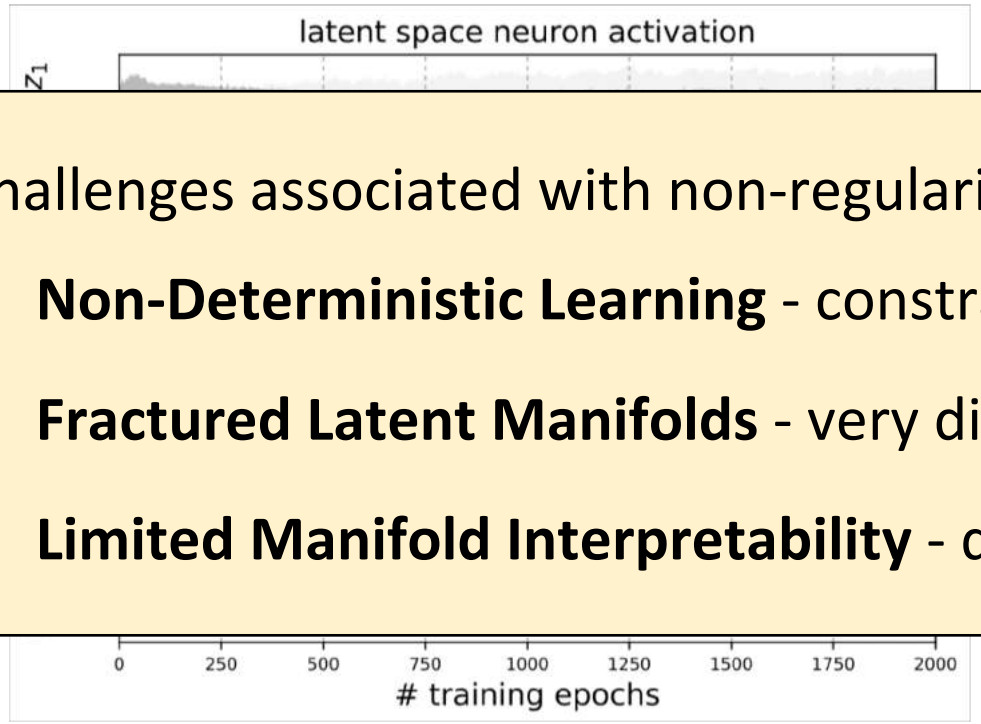
Autoencoder NNs - Latent Space Analysis II

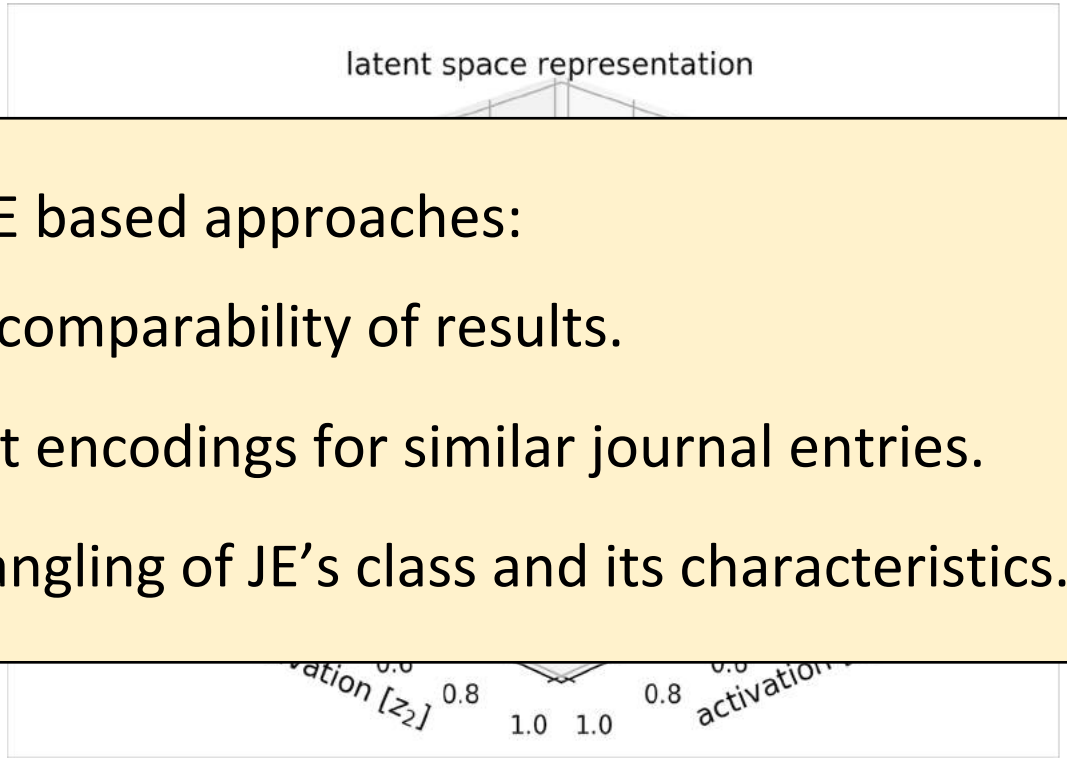




Transaction class | Regular | Global Anomalies | Local Anomalies

Autoencoder NNs - Latent Space Analysis II



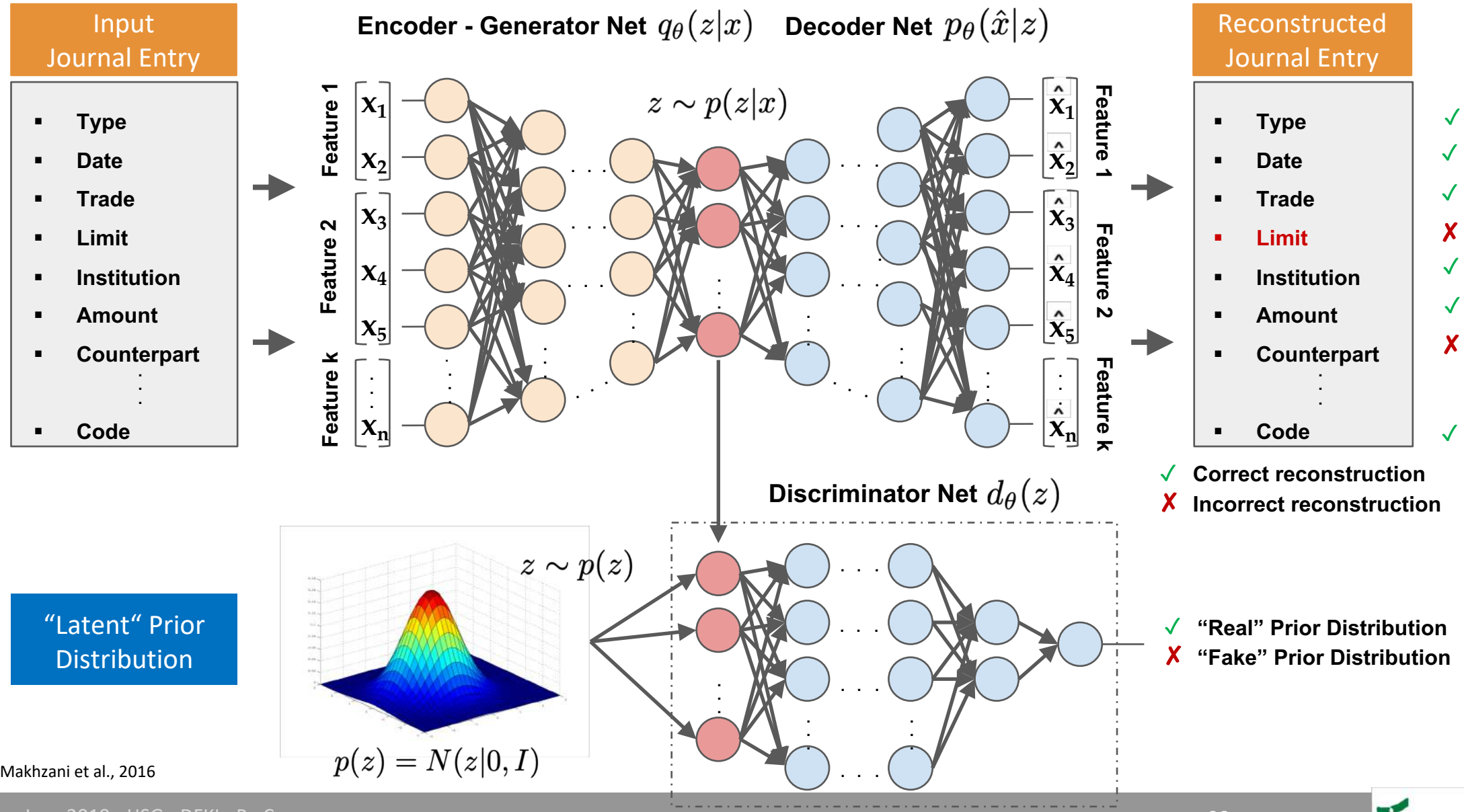


Challenges associated with non-regularized AE based approaches:

- **Non-Deterministic Learning** - constrained comparability of results.
- **Fractured Latent Manifolds** - very different encodings for similar journal entries.
- **Limited Manifold Interpretability** - disentangling of JE's class and its characteristics.

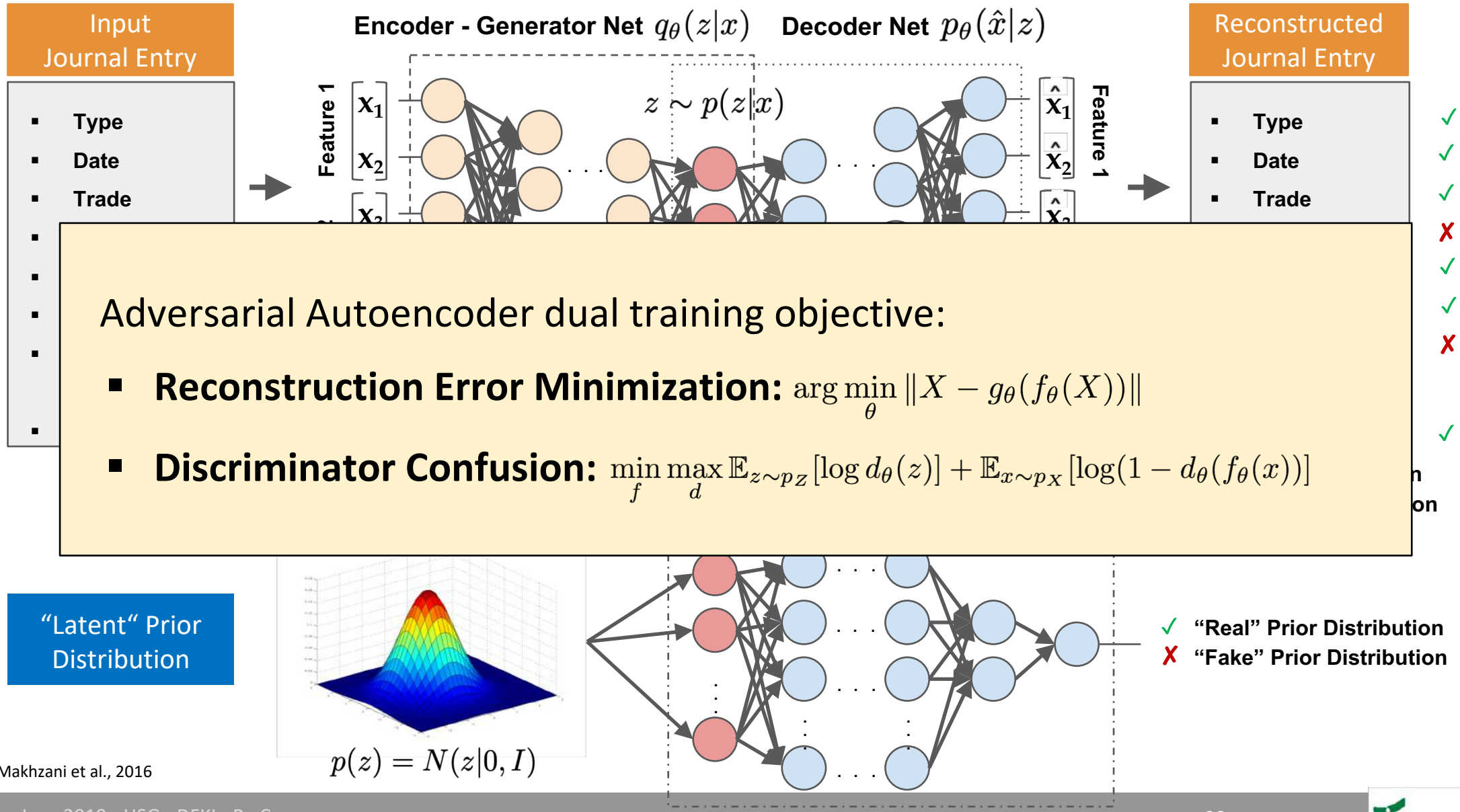
Transaction class | Regular | Global Anomalies | Local Anomalies

Adversarial Regularized Autoencoder



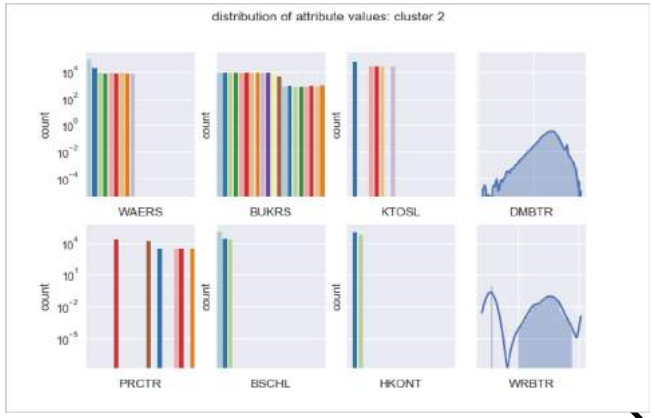
[5] Makhzani et al., 2016

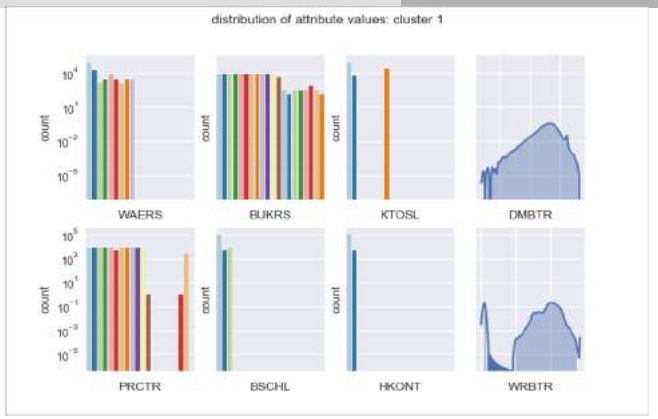
Adversarial Regularized Autoencoder

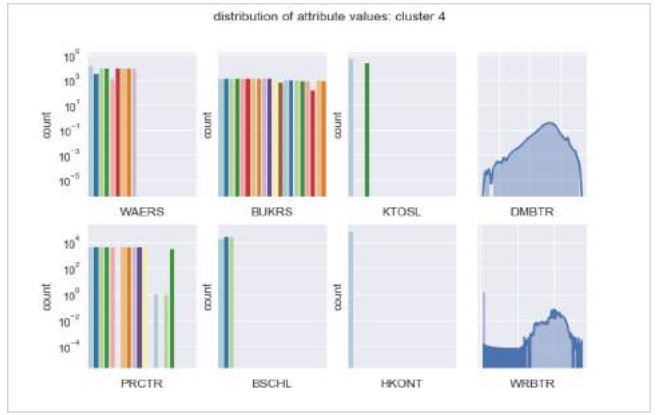


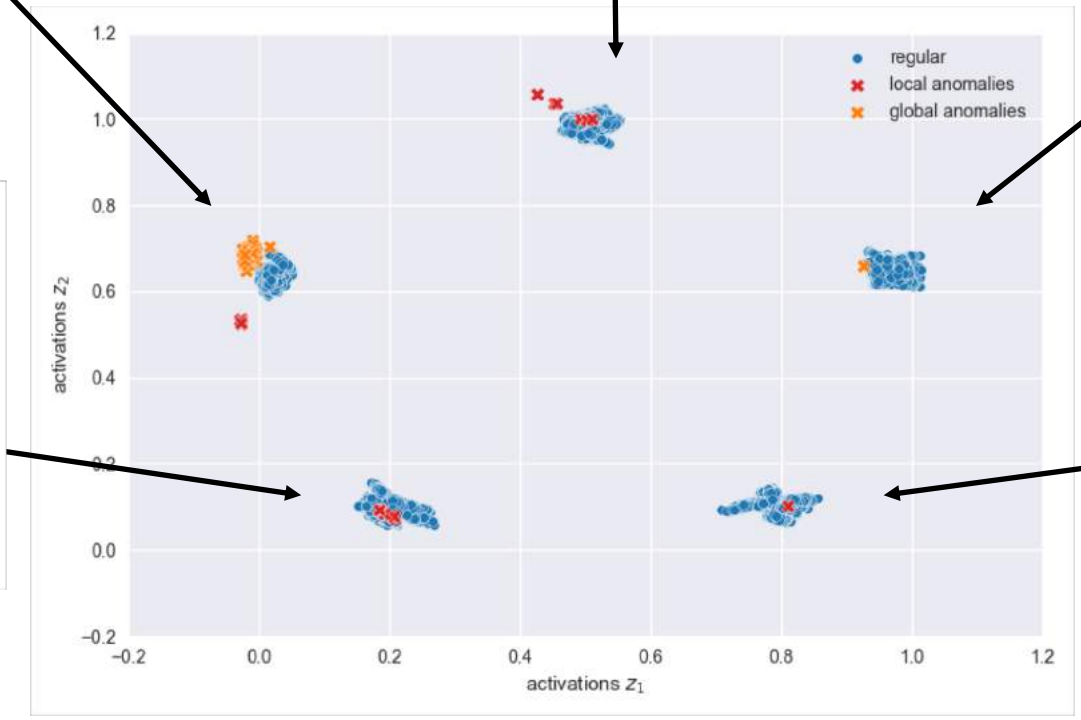
[5] Makhzani et al., 2016

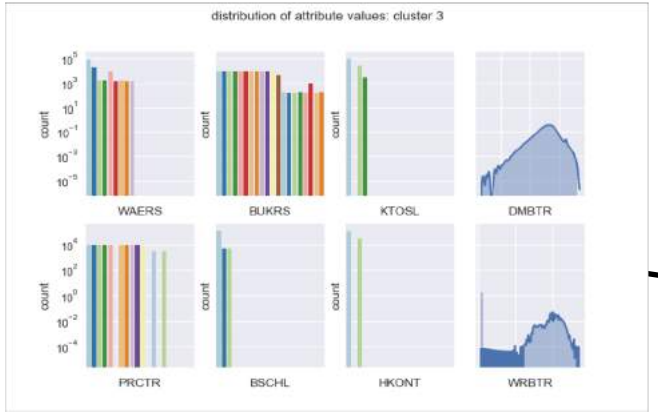
Adversarial Regularized Autoencoder

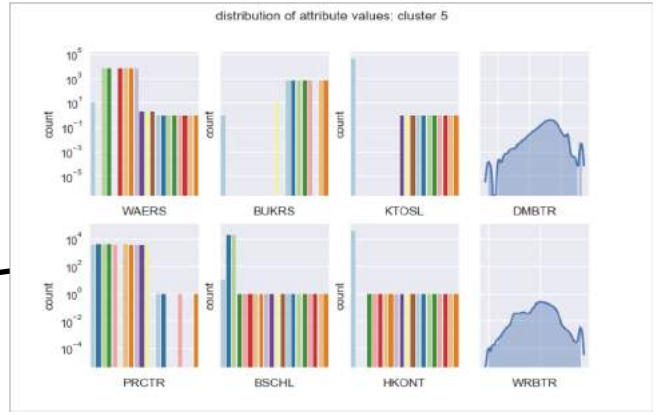




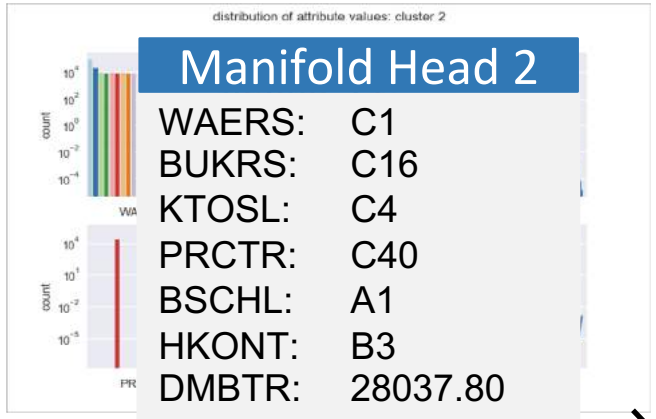






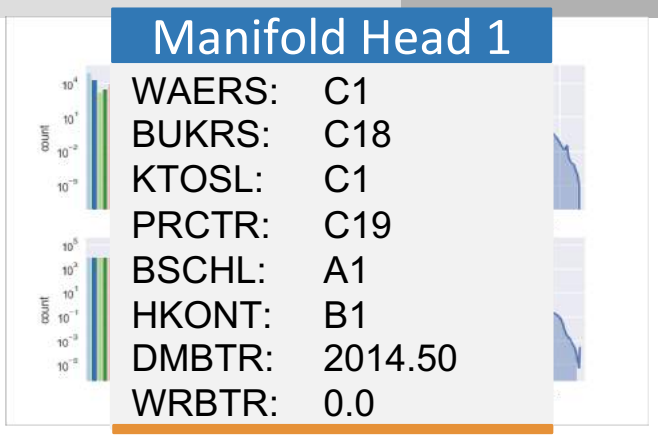


Adversarial Regularized Autoencoder



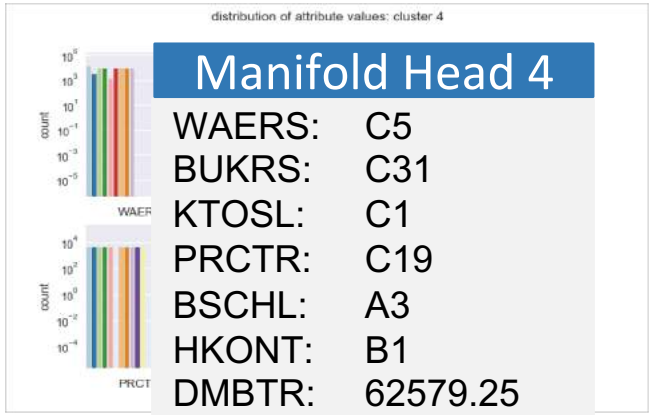
Manifold Head 2

WAERS:	C1
BUKRS:	C16
KTOSL:	C4
PRCTR:	C40
BSCHL:	A1
HKONT:	B3
DMBTR:	28037.80
WRBTR:	0.00



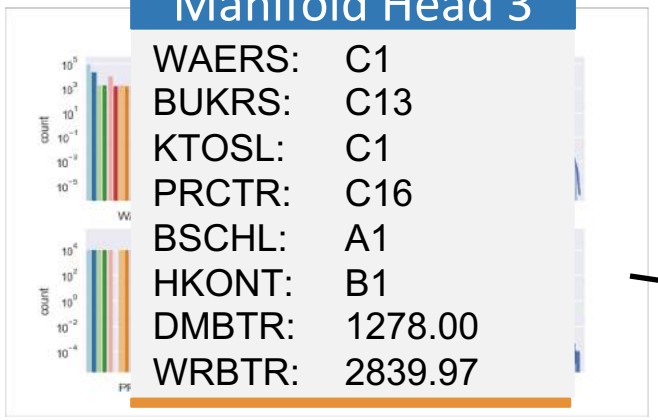
Manifold Head 1

WAERS:	C1
BUKRS:	C18
KTOSL:	C1
PRCTR:	C19
BSCHL:	A1
HKONT:	B1
DMBTR:	2014.50
WRBTR:	0.0



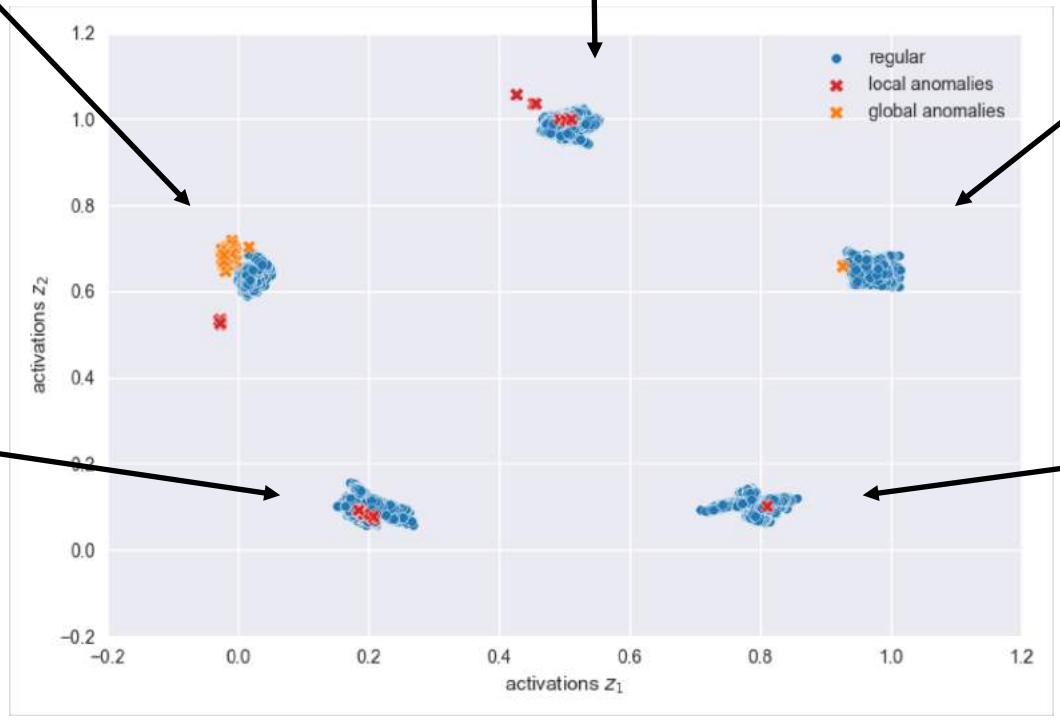
Manifold Head 4

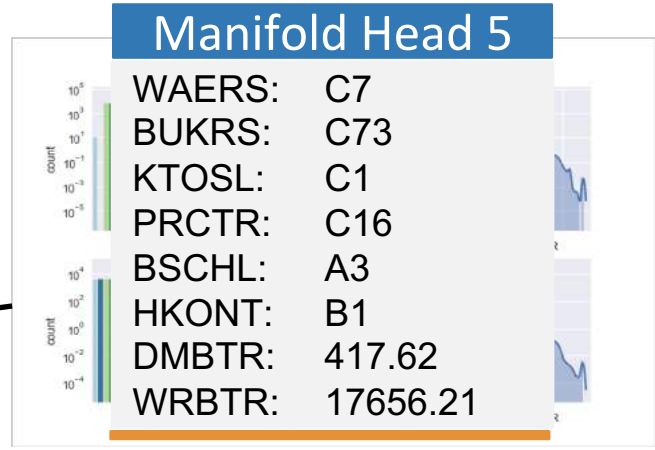
WAERS:	C5
BUKRS:	C31
KTOSL:	C1
PRCTR:	C19
BSCHL:	A3
HKONT:	B1
DMBTR:	62579.25
WRBTR:	0.0



Manifold Head 3

WAERS:	C1
BUKRS:	C13
KTOSL:	C1
PRCTR:	C16
BSCHL:	A1
HKONT:	B1
DMBTR:	1278.00
WRBTR:	2839.97

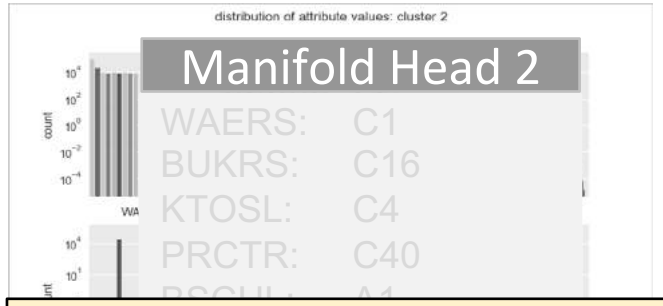


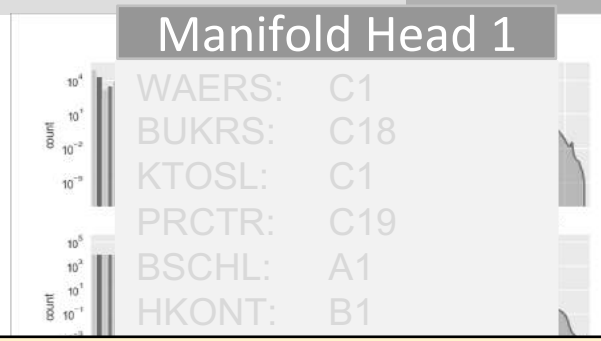


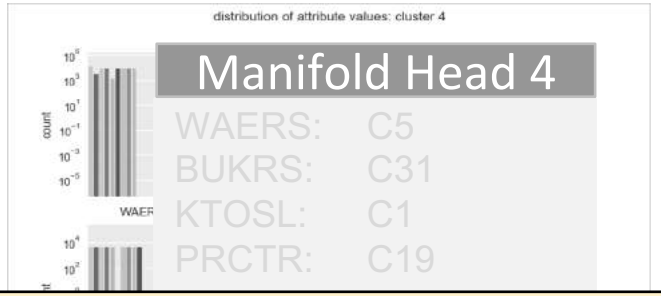
Manifold Head 5

WAERS:	C7
BUKRS:	C73
KTOSL:	C1
PRCTR:	C16
BSCHL:	A3
HKONT:	B1
DMBTR:	417.62
WRBTR:	17656.21

Adversarial Regularized Autoencoder

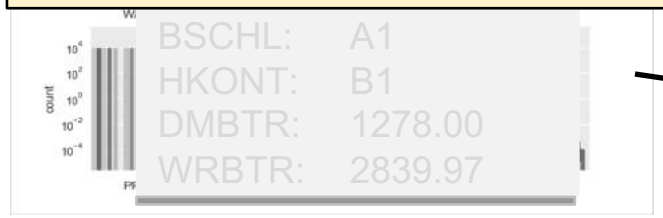


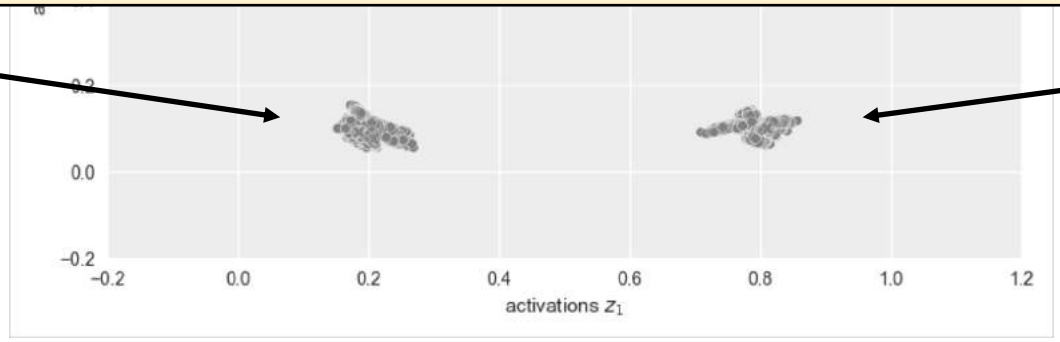


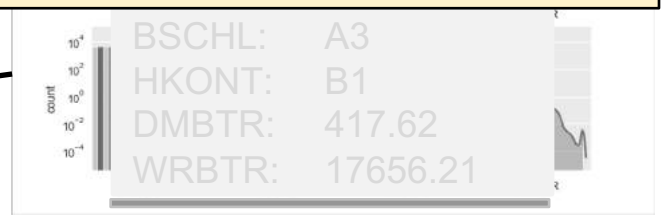


Challenges associated with learning accounting data using regularized AE:

- **Disentangle Distinct Attribute Types** - learning of categorical and numerical attributes.
- **Interpretability of the Latent Space** - determination of cluster number and distances.







Adversarial Autoencoder NNs - Training Process

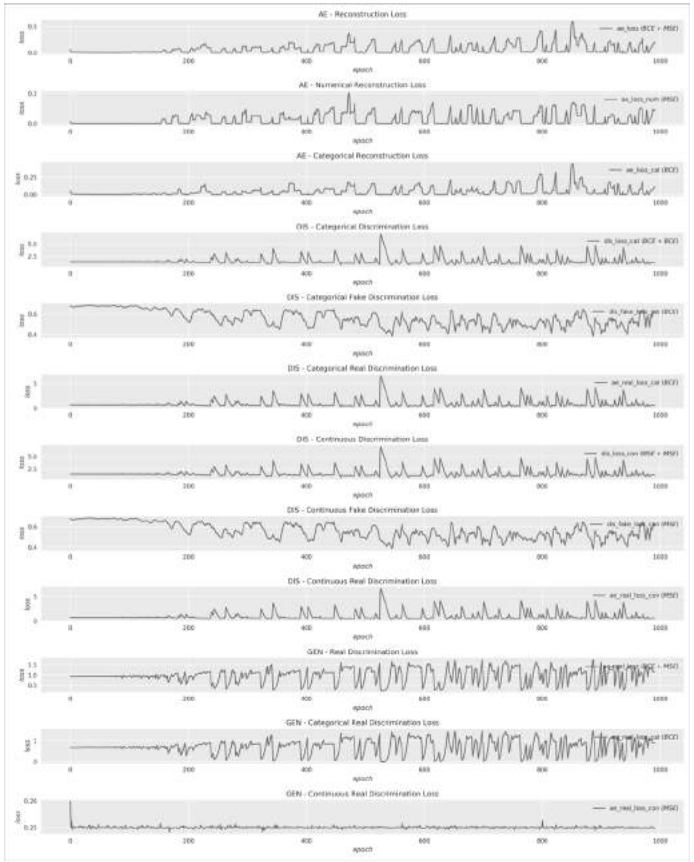
AE Losses

Cat. Dis. Losses

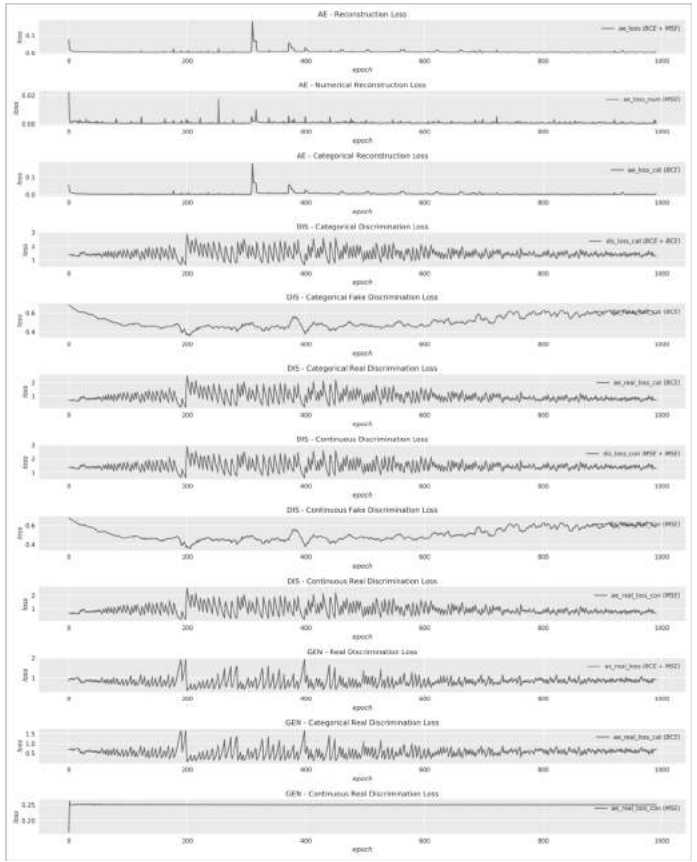
Con. Dis. Losses

Gen. Losses

AE Architecture 1 Training



AE Architecture 2 Training

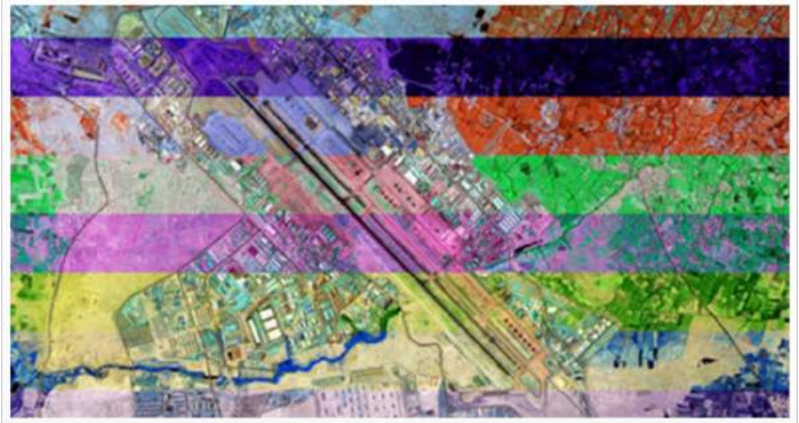


Forbes

24,584 views | Feb 19, 2019, 03:56pm

How 3 Of Europe's Universities Are Transforming AI Research

Guillaume Barat Brand Contributor
NVIDIA BRANDVOICE



- Architectures and models
- Multimedia analysis and synthesis
- Remote sensing and analyzing satellite data
- AI for financial markets

Adversarial Autoencoder NNs - Training Process

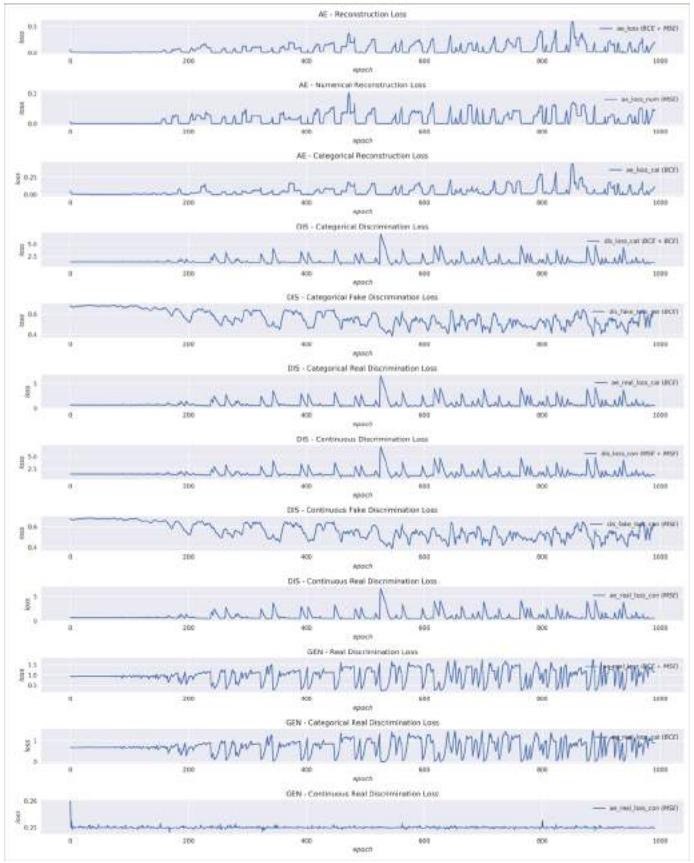
AE Losses

Cat. Dis. Losses

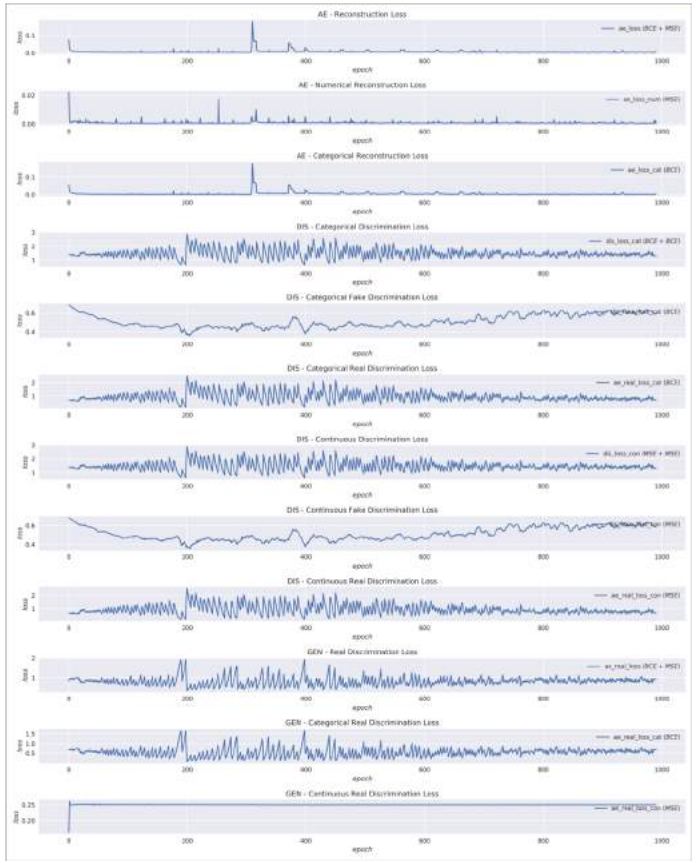
Con. Dis. Losses

Gen. Losses

AE Architecture 1 Training

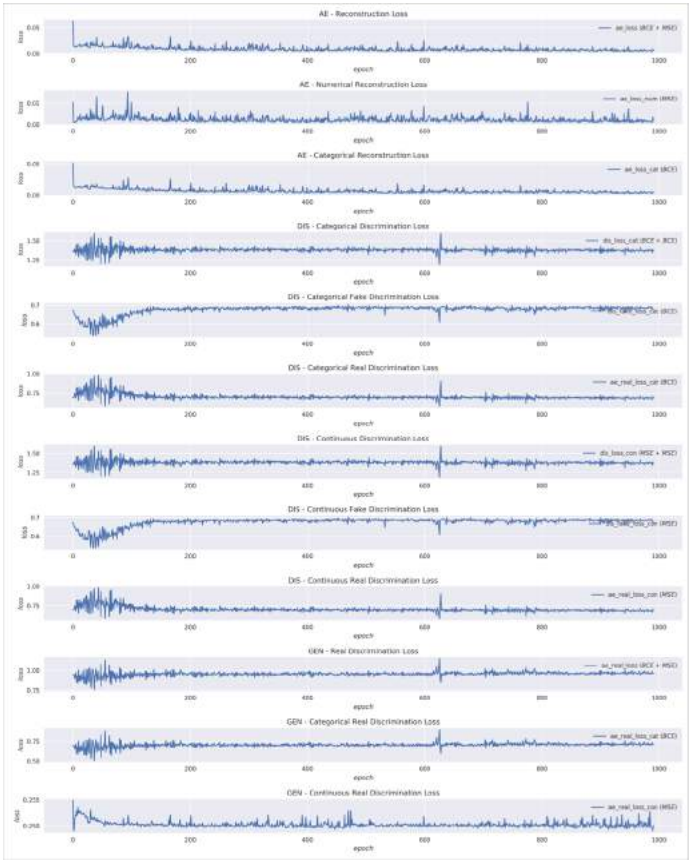


AE Architecture 2 Training



. . .

AE Architecture n Training



Adversarial Autoencoder NNs - Training Process

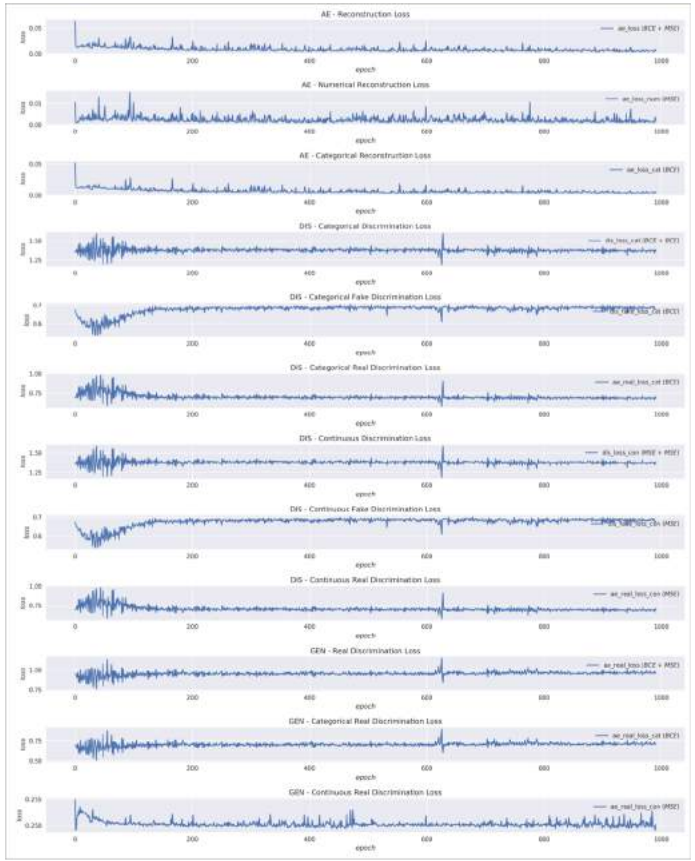
AE Losses

Cat. Dis. Losses

Con. Dis. Losses

Gen. Losses

AE Architecture Training

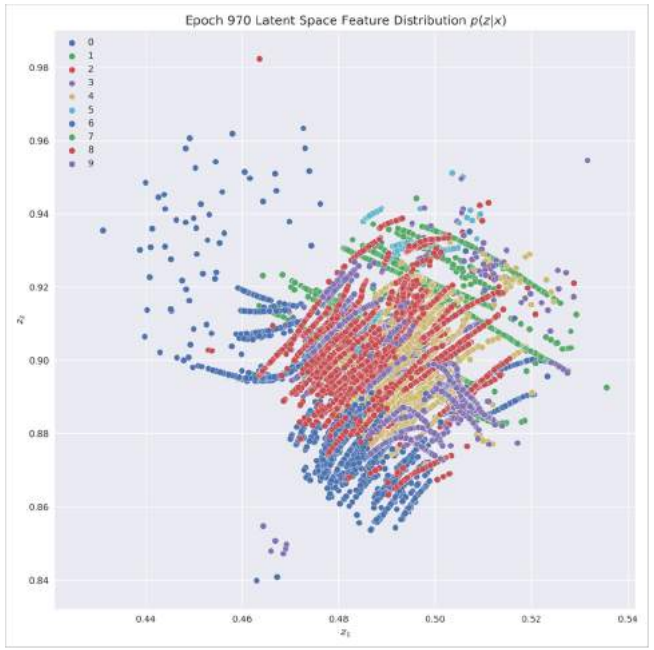


AE Feature Learning

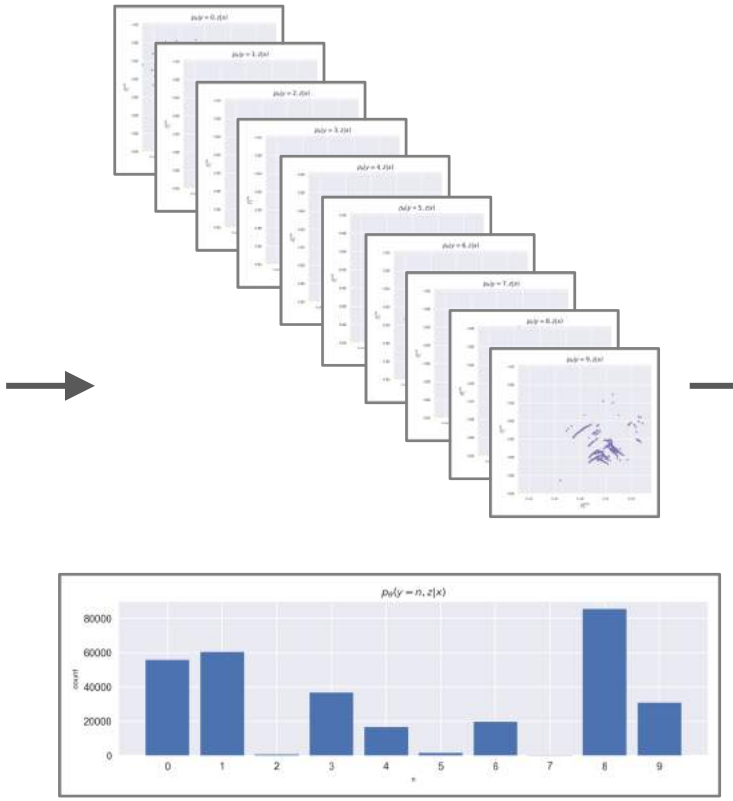
Ground Truth Labels
Anomalies vs. non-Anomalies

Learned Representations
Incl. Clusters

Latent Space Inspection



Learned Representations
incl. Clusters



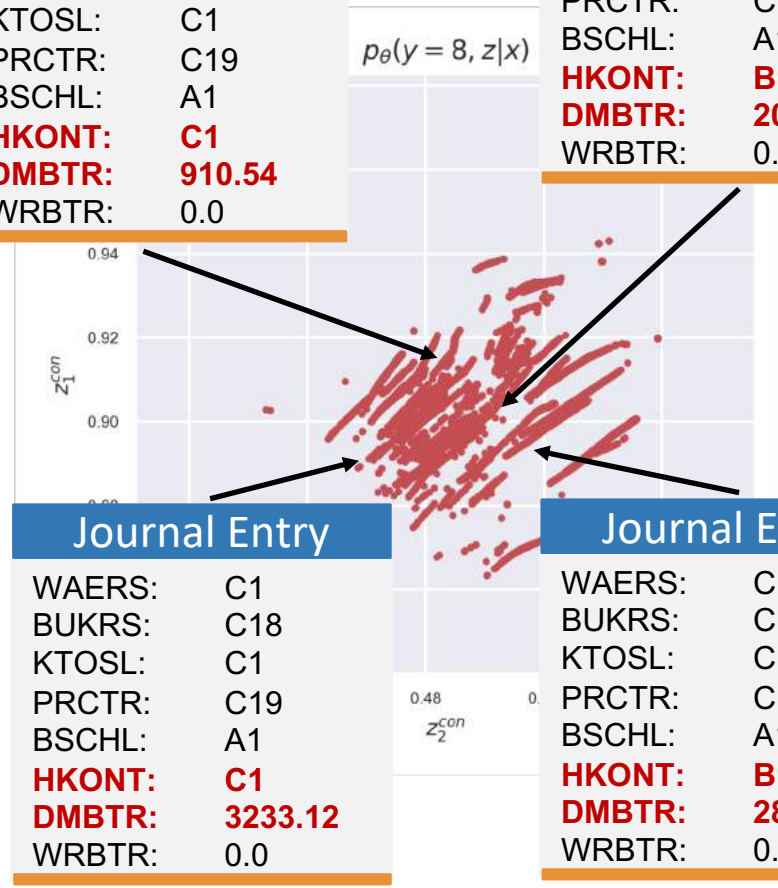
Journal Entries
Cluster Distribution

Journal Entry

WAERS:	C1
BUKRS:	C18
KTOSL:	C1
PRCTR:	C19
BSCHL:	A1
HKONT:	**C1**
DMBTR:	**910.54**
WRBTR:	0.0

Manifold Head 1

WAERS:	C1
BUKRS:	C18
KTOSL:	C1
PRCTR:	C19
BSCHL:	A1
HKONT:	**B1**
DMBTR:	**2014.50**
WRBTR:	0.0



Journal Entry

WAERS:	C1
BUKRS:	C18
KTOSL:	C1
PRCTR:	C19
BSCHL:	A1
HKONT:	**C1**
DMBTR:	**3233.12**
WRBTR:	0.0

Journal Entry

WAERS:	C1
BUKRS:	C18
KTOSL:	C1
PRCTR:	C19
BSCHL:	A1
HKONT:	**B1**
DMBTR:	**2865.97**
WRBTR:	0.0

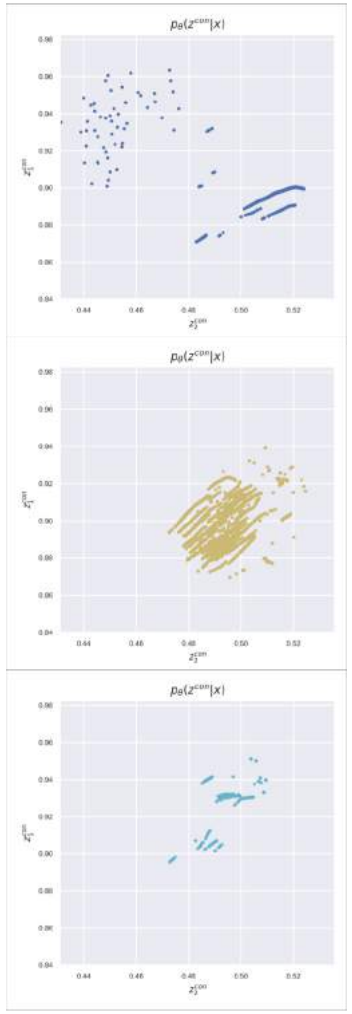
Journal Entries Within
Cluster Distribution

Latent Space Inspection

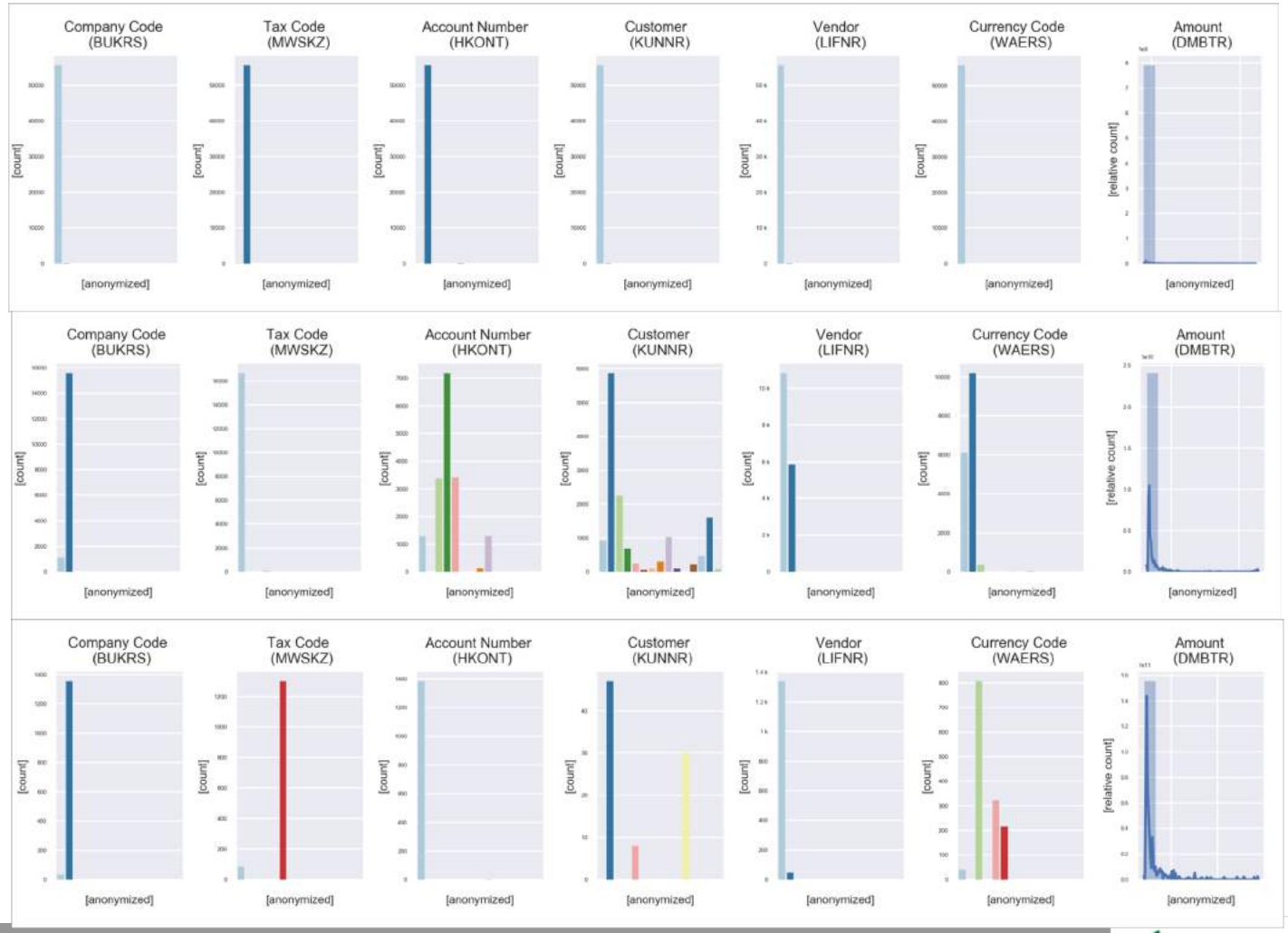
$q_\theta(y = 0, z|x)$ → "Cluster 0"

$q_\theta(y = 4, z|x)$ → "Cluster 4"

$q_\theta(y = 5, z|x)$ → "Cluster 5"



Journal Entry Characteristics - Top 15 Attribute Values

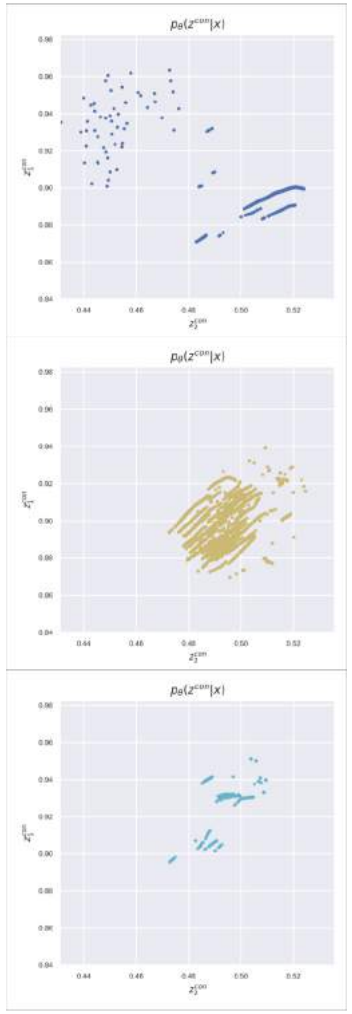


Latent Space Inspection

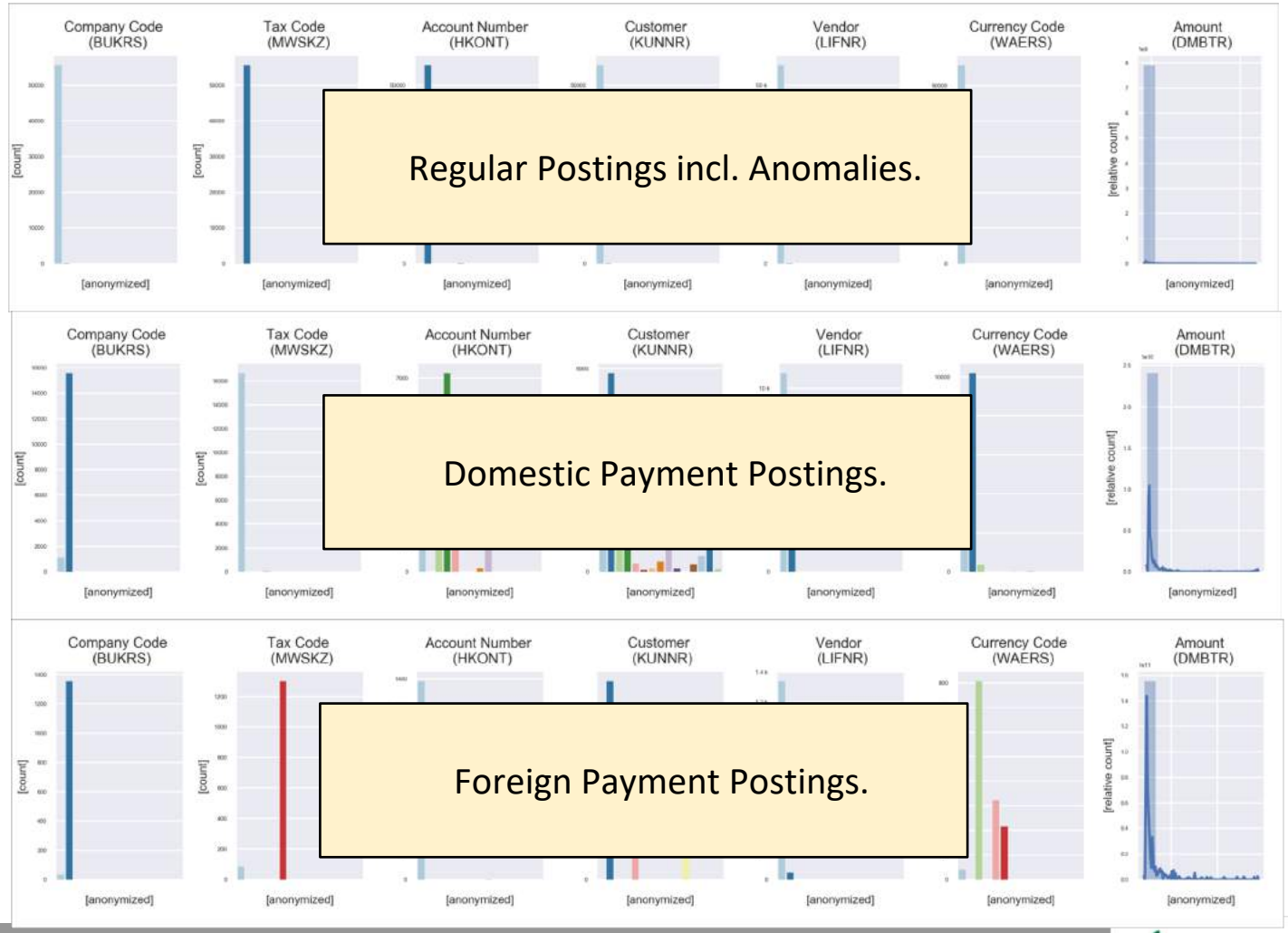
$q_\theta(y = 0, z|x)$ → "Cluster 0"

$q_\theta(y = 4, z|x)$ → "Cluster 4"

$q_\theta(y = 5, z|x)$ → "Cluster 5"



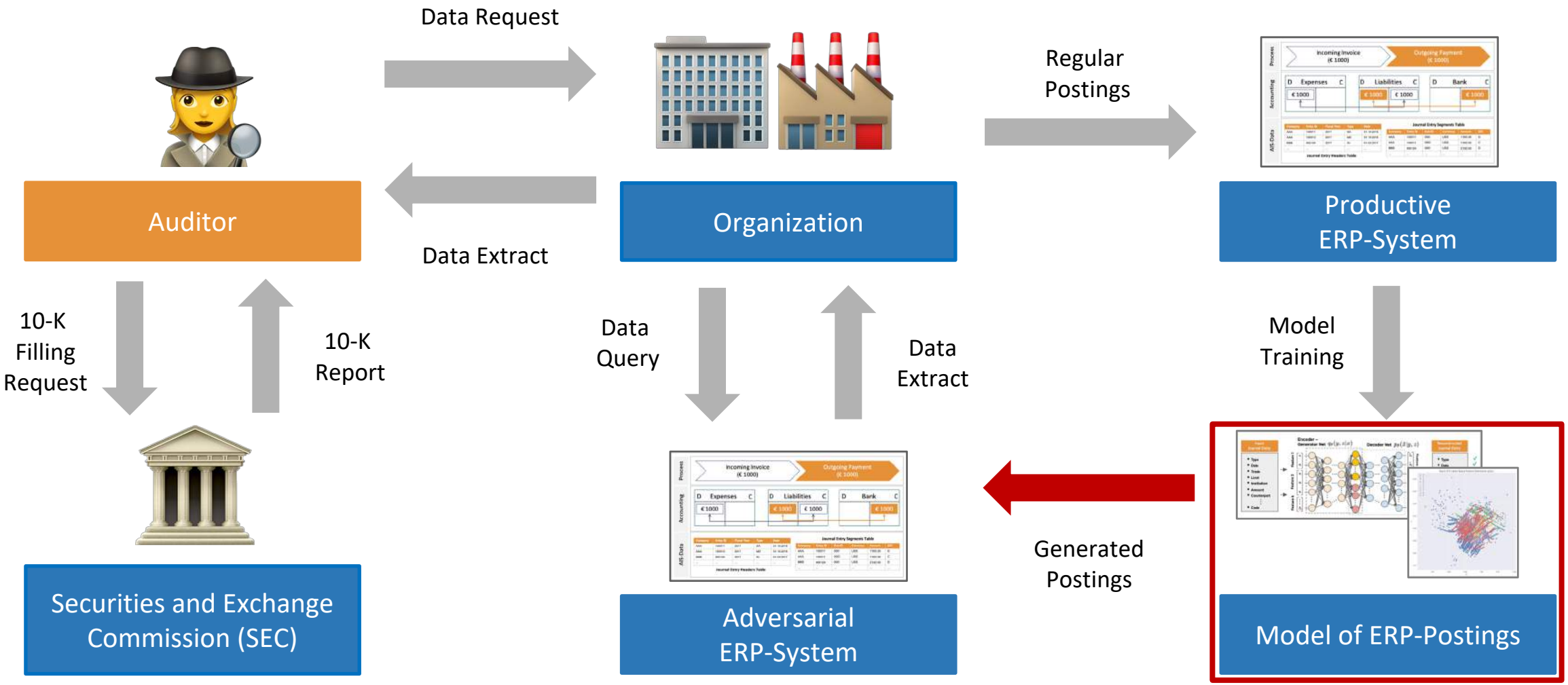
Journal Entry Characteristics - Top 15 Attribute Values



Adversarial Approaches

Generative Adversarial Autoencoder Networks

Threat Model



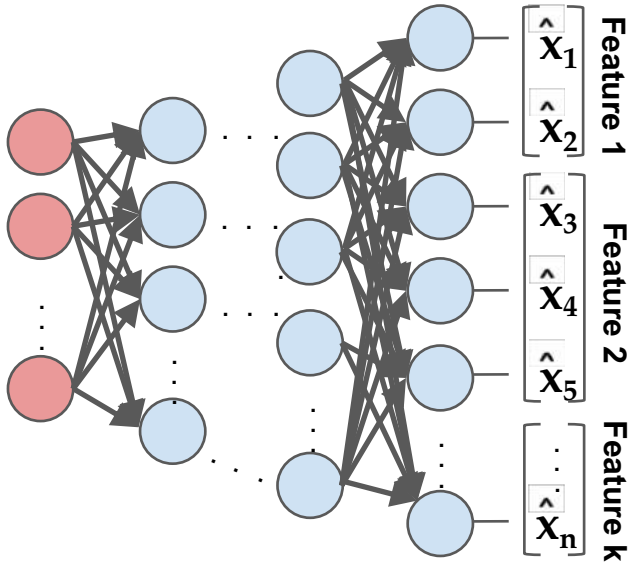
Sampling from the Latent Space

Sampling → Reconstruction → Posting

"Latent Space" Code Neurons

$z_1, z_2, \ldots, z_n$

Decoder Net g_Θ



Adversarial Journal Entry

- Type
- Date
- Trade
- Limit
- Institutio
- Amount
- Counter
- Code

"Fake" Entry

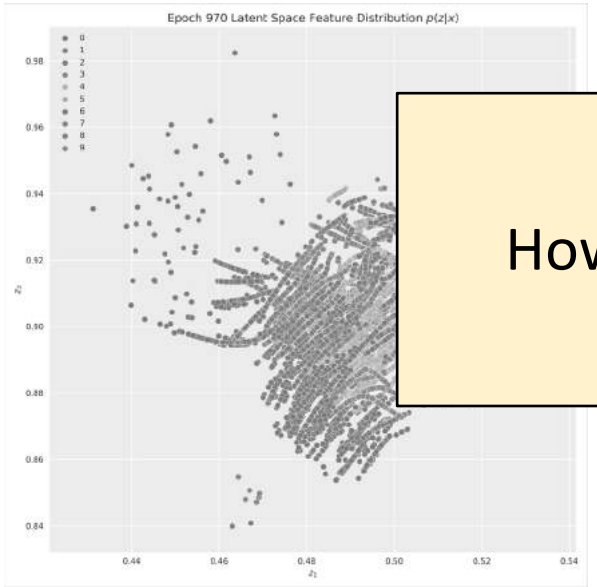
WAERS:	C1
BUKRS:	C13
KTOSL:	C1
PRCTR:	C16
BSCHL:	A1
HKONT:	B1
DMBTR:	1271488.00
WRBTR:	28394.97

Sampling from the Latent Space

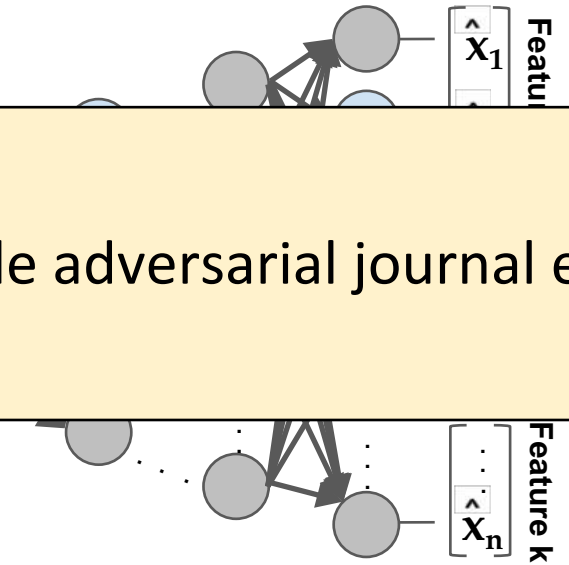
Sampling → Reconstruction → Posting

"Latent Space" Code Neurons

$z_1, z_2, \ldots, z_n$



Decoder Net g_Θ



Adversarial Journal Entry

- Type
- Date
- Trade
- Limit
- Institutio
- Amount
- Counter
- Code

How to sample adversarial journal entries?

"Fake" Entry

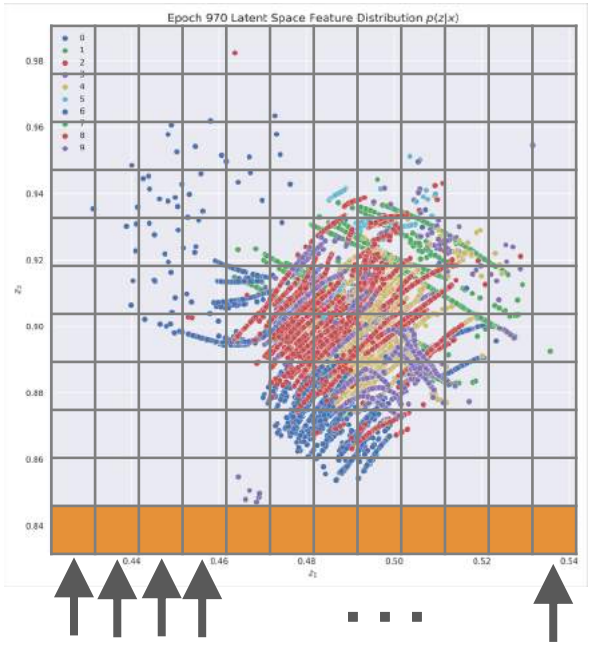
WAERS:	C1
BUKRS:	C13
KTOSL:	C1
PRCTR:	C16
BSCHL:	A1
HKONT:	B1
DMBTR:	1271488.00
WRBTR:	28394.97

Sampling from the Latent Space

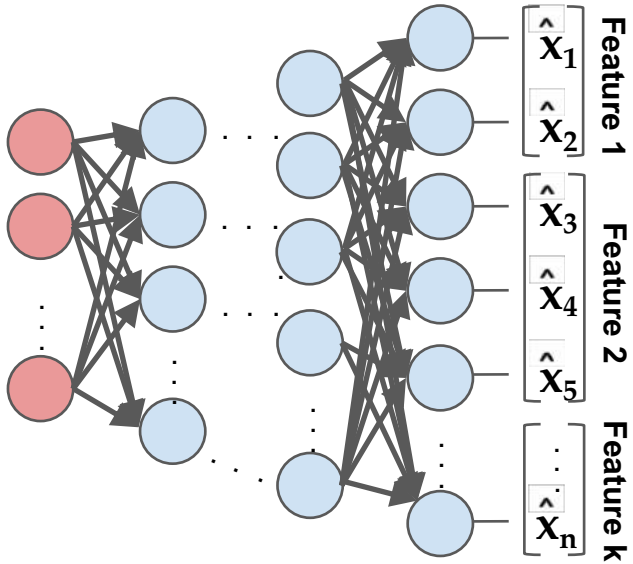
Sampling → Reconstruction → Posting

"Latent Space" Code Neurons

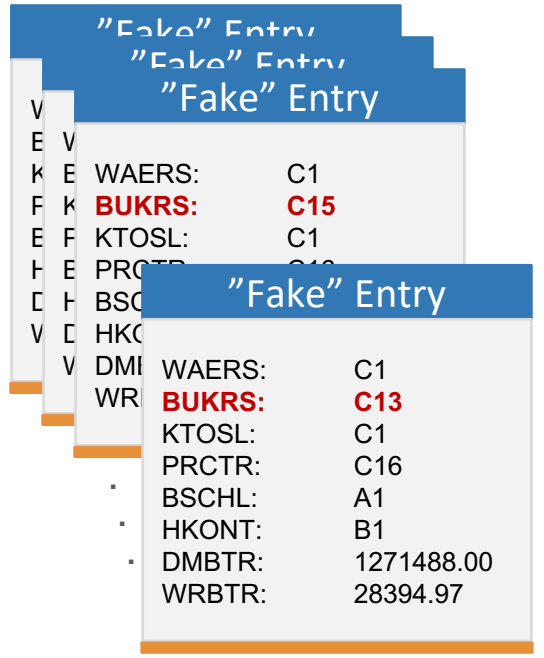
$z_1, z_2, \ldots, z_n$



Decoder Net g_Θ



Adversarial Journal Entry

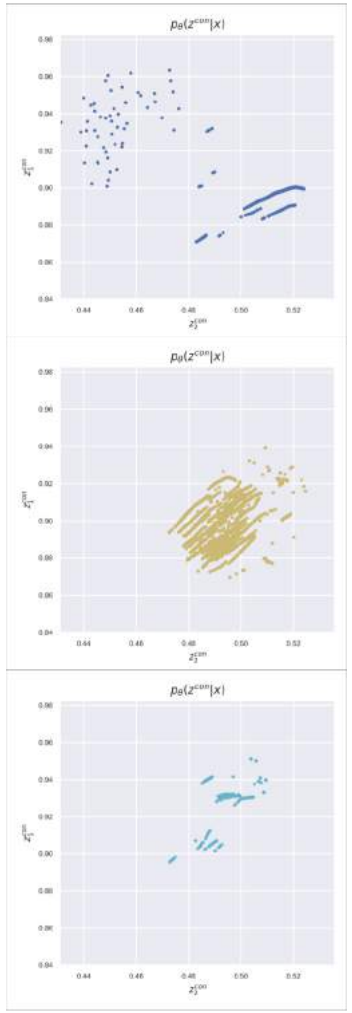


Latent Space Exploration

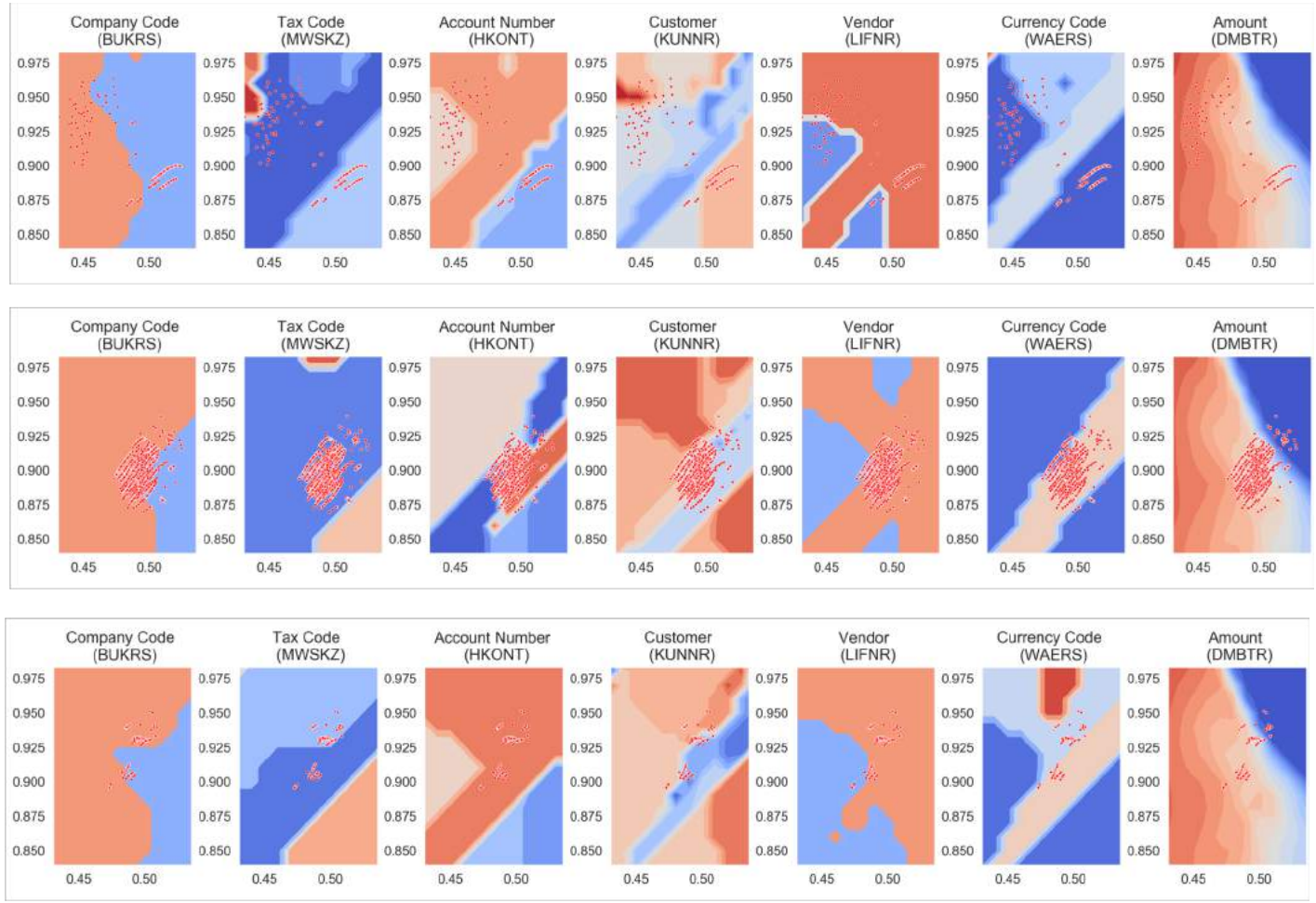
$q_\theta(y = 0, z|x)$ → "Cluster 0"

$q_\theta(y = 4, z|x)$ → "Cluster 4"

$q_\theta(y = 5, z|x)$ → "Cluster 5"



Latent Space Decision Boundaries Characteristics



Thank you

Questions?

Thank you. Questions?

References

[1] Breunig, M.M., Kriegel H.-P., Ng, R. T., and, Sander, J., ***"LOF: Identifying Density-Based Local Outliers"***, Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, 2000, USA.

[2] Benford Frank; *„The Law of Anomalous Numbers"*, Proceedings of the American Philosphical Society, Vol. 78, 1938, USA.

[3] Hinton, G. and Salakhutdinov, R., ***"Reducing the Dimensionality of Data with Neural Networks"***, Science, Vol. 313, p. 504-507, 2006.

[4] Hawkins, S., He, H., Williams, G., and, Baxter R., ***"Outlier Detection Using Replicator Neural Networks"***, Proc. International Conference on Data Warehousing and Knowledge Discovery, 2002, USA

[5] Schreyer, M., Sattarov, T., Borth, D., Dengel, A., and, Reimer, B. ***"Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks",*** arXiv preprint, arXiv: 1709.05254, 2017.

[6] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and, Frey, B., ***"Adversarial Autoencoders"***, arXiv preprint, arXiv:1511.05644, 2016.

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., ***"Generative Adversarial Nets"***, In Advances in neural information processing systems, pp. 2672-2680, 2014.

References

[8] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and, Abbeel, P., ***"InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets"***, In Advances in neural information processing systems, pp. 2172-2180, 2016.

[9] Yuan, X., He, P., Zhu, Q., and, Li, X., ***"Adversarial Examples: Attacks and Defenses for Deep Learning"***, arXiv preprint, arXiv: 1712.07107, 2017.

[10] Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., and Song, D., ***"Robust physical-world attacks on deep learning models"***, arXiv preprint arXiv:1707.08945, 2017.

[11] Su, J., Vargas, D.V. and Sakurai, K., ***"One pixel attack for fooling deep neural networks"***, IEEE Transactions on Evolutionary Computation, 2017.

[12] Huang, S., Papernot, N., Goodfellow, I., Duan, Y. and Abbeel, P., ***"Adversarial attacks on neural network policies"***, arXiv preprint arXiv:1702.02284, 2017.

[13] Alzantot, M., Balaji, B. and Srivastava, M., ***"Did you hear that? adversarial examples against automatic speech recognition"***, arXiv preprint arXiv:1801.00554, 2018.