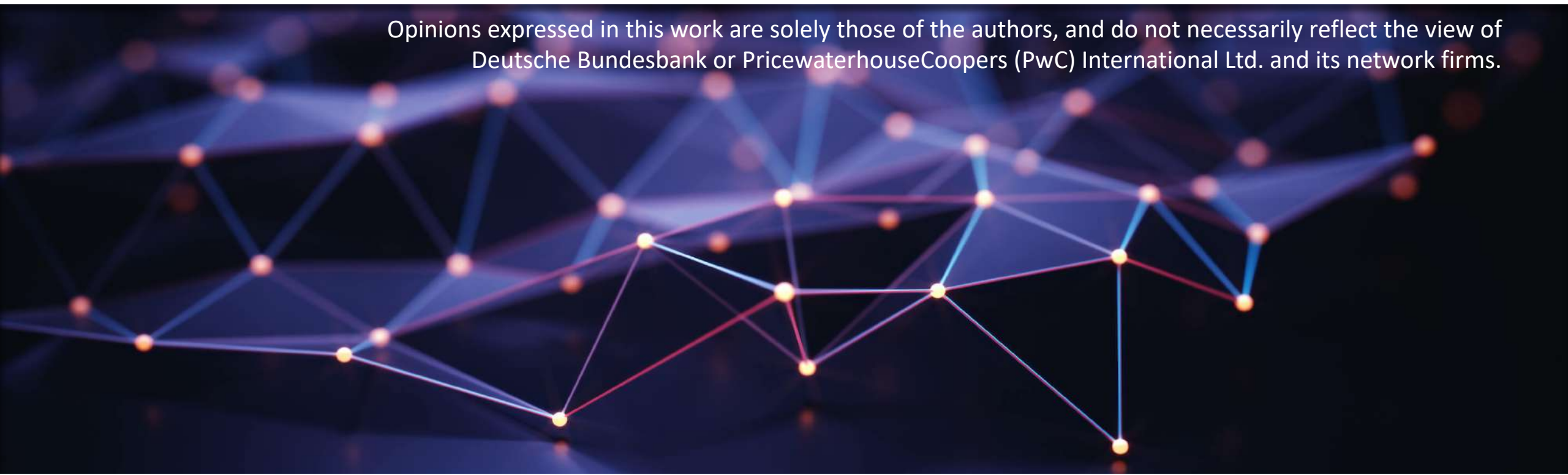University of St.Gallen

Opinions expressed in this work are solely those of the authors, and do not necessarily reflect the view of Deutsche Bundesbank or PricewaterhouseCoopers (PwC) International Ltd. and its network firms.

DEUTSCHE BUNDESBANK EUROSYSTEM

pwc

ACM International Conference on AI in Finance (ICAIF) 2020

Learning Sampling in Financial Statement Audits using Vector Quantised Autoencoder Neural Networks
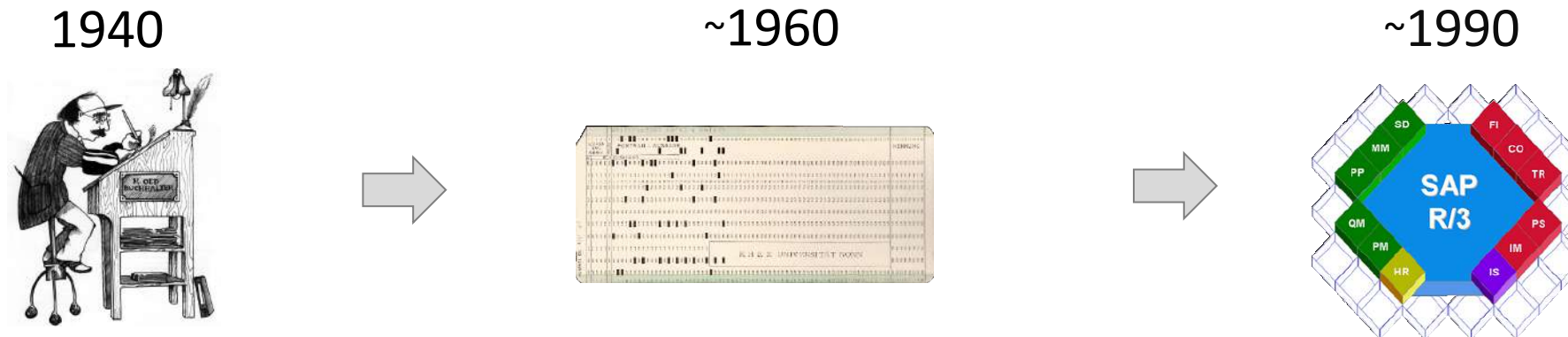
M. Schreyer[1], T. Sattarov[2], A. Gierbl[1,3], B. Reimer[3] & D. Borth[1]
[1]University of St.Gallen, [2]Deutsche Bundesbank, [3]PricewaterhouseCoopers GmbH

October 16th, 2020

Motivation & Background

Evolution of Recording and Processing Accounting Data

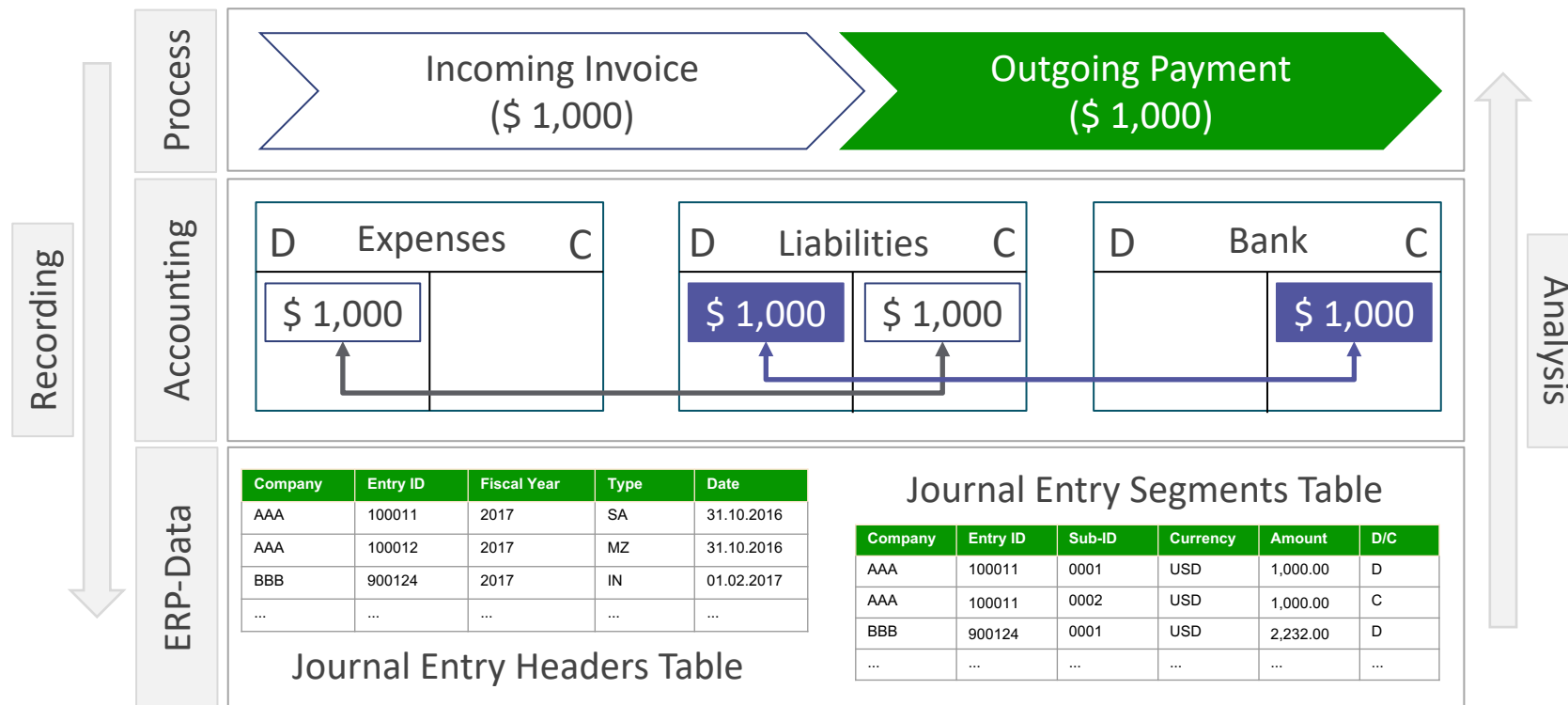| 1940	~1960	~1990

Data Volume

Towards the "Data Driven Economy"

- Displacement of the non-digital processing of organizational activities
- Accumulation of exhaustive volumes of transactional and accounting data
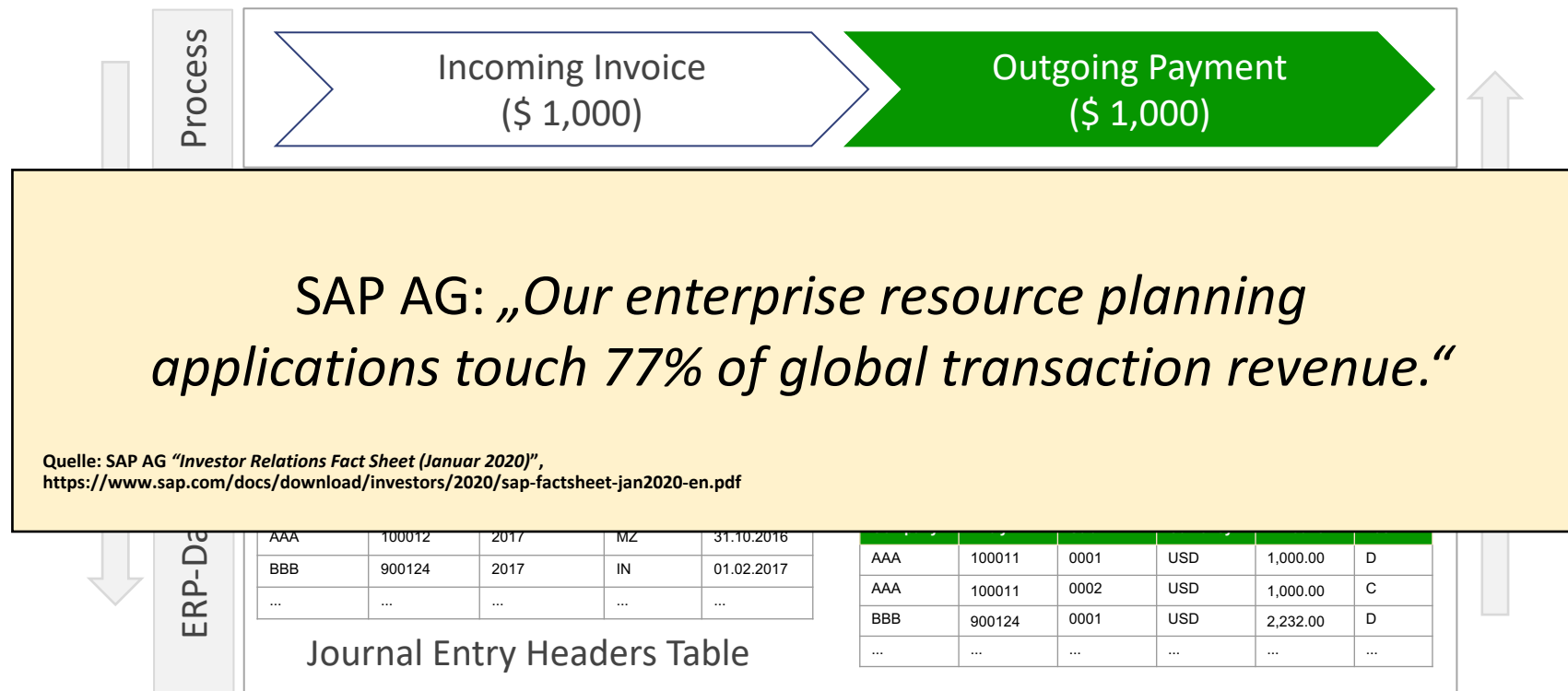- Almost every activity within an organization leaves **a digital trace** … !

Motivation & Background

Evolution of Recording and Processing Accounting Data

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
...	...	...	...	...	...

Motivation & Background

Evolution of Recording and Processing Accounting Data

Incoming Invoice ($ 1,000)		Outgoing Payment ($ 1,000)		

SAP AG: *„Our enterprise resource planning applications touch 77% of global transaction revenue."*

Quelle: SAP AG *"Investor Relations Fact Sheet (Januar 2020)"*,
https://www.sap.com/docs/download/investors/2020/sap-factsheet-jan2020-en.pdf

AAA	100012	2017	MZ	31.10.2016
BBB	900124	2017	IN	01.02.2017
...	...	...	...	...

Journal Entry Headers Table

AAA	100011	0001	USD	1,000.00	D
AAA	100011	0002	USD	1,000.00	C
BBB	900124	0001	USD	2,232.00	D
...	...	...	...	...	...

Motivation & Background

Responses of the Audit Practice

Challenge

"True and Fair Presentation"   "Big Data"

Audit sampling: "The application of audit procedures **to less than 100% of items within a population** of audit relevance such that all sampling units have a chance of selection in order to provide the auditor with a reasonable basis on which to draw conclusions about the entire population."

Source: International Auditing and Assurance Standards Board, International Standard on Auditing 530 - "Audit Sampling", Effective December 2009

Sampling Risk: "The risk that the auditor's conclusion based on a sample **may be different** from the conclusion if the entire population were subjected to the same audit procedure."

Source: American Institute of Certified Professional Accountants, AU-C Section 530 "Audit Sampling", Effective December 2012

Related Work (Current Practice)

Statistical Audit Sampling - "Attribute Sampling Techniques"

Objective: Estimate the failing of internal controls via the test of controls.

Random Sampling

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100011	2017	SA	31.10.2016	1009000	2,685.00	...	FB01
AAA	100012	2017	MZ	31.10.2016	1008010	250,000.00	...	FB01
BBB	900124	2017	IN	01.11.2016	6019000	68,488.45	...	SA03
AAA	100013	2017	SA	14.11.2016	1008010	4,000.00	...	FB01
AAA	100014	2017	MZ	25.11.2016	5000110	4,395.00	...	FB01
BBB	900125	2017	IN	01.12.2016	6019000	5,995.00	...	SA03
AAA	100015	2017	MZ	12.12.2016	1009000	180.00	...	FB01
BBB	900126	2017	IN	22.11.2016	9400000	18,500.00	...	SA03
BBB	900125	2017	IN	20.12.2016	5000110	697.15	...	SA03
BBB	900126	2017	IN	21.01.2016	6019000	95,400.00	...	SA03
AAA	100015	2017	MZ	01.01.2017	1009000	1,017,862.23	...	FB01
BBB	900126	2017	IN	02.01.2017	5000110	234.50	...	SA03
...	...	...	...	...	...	...	...	...

Population

Stratified Sampling

Population A

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100011	2017	SA	31.10.2016	1009000	2,685.00	...	FB01
AAA	100012	2017	MZ	31.10.2016	1008010	250,000.00	...	FB01
AAA	100013	2017	SA	14.11.2016	1008010	4,000.00	...	FB01
AAA	100014	2017	MZ	25.11.2016	5000110	4,395.00	...	FB01
AAA	100015	2017	MZ	12.12.2016	1009000	180.00	...	FB01
...	...	...	...	...	...			

Population B

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
BBB	900124	2017	IN	01.11.2016	6019000	68,488.45	...	SA03
BBB	900125	2017	IN	01.12.2016	6019000	5,995.00	...	SA03
BBB	900126	2017	IN	22.11.2016	9400000	18,500.00	...	SA03
BBB	900125	2017	IN	20.12.2016	5000110	697.15	...	SA03
BBB	900126	2017	IN	21.01.2016	6019000	95,400.00	...	SA03
...	...	...	...	...	...			

Systematic Sampling

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100011	2017	SA	31.10.2016	1009000	2,685.00	...	FB01
AAA	100012	2017	MZ	31.10.2016	1008010	250,000.00	...	FB01
BBB	900124	2017	IN	01.11.2016	6019000	68,488.45	...	SA03
AAA	100013	2017	SA	14.11.2016	1008010	4,000.00	...	FB01
AAA	100014	2017	MZ	25.11.2016	5000110	4,395.00	...	FB01
BBB	900125	2017	IN	01.12.2016	6019000	5,995.00	...	SA03
AAA	100015	2017	MZ	12.12.2016	1009000	180.00	...	FB01
BBB	900126	2017	IN	22.11.2016	9400000	18,500.00	...	SA03
BBB	900125	2017	IN	20.12.2016	5000110	697.15	...	SA03
BBB	900126	2017	IN	21.01.2016	6019000	95,400.00	...	SA03
AAA	100015	2017	MZ	01.01.2017	1009000	1,017,862.23	...	FB01
BBB	900126	2017	IN	02.01.2017	5000110	234.50	...	SA03
...	...	...	...	...	...	...	...	...

Population

Haphazard Sampling

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100011	2017	SA	31.10.2016	1009000	2,685.00	...	FB01
AAA	100012	2017	MZ	31.10.2016	1008010	250,000.00	...	FB01
BBB	900124	2017	IN	01.11.2016	6019000	68,488.45	...	SA03
AAA	100013	2017	SA	14.11.2016	1008010	4,000.00	...	FB01
AAA	100014	2017	MZ	25.11.2016	5000110	4,395.00	...	FB01
BBB	900125	2017	IN	01.12.2016	6019000	5,995.00	...	SA03
AAA	100015	2017	MZ	12.12.2016	1009000	180.00	...	FB01
BBB	900126	2017	IN	22.11.2016	9400000	18,500.00	...	SA03
BBB	900125	2017	IN	20.12.2016	5000110	697.15	...	SA03
BBB	900126	2017	IN	21.01.2016	6019000	95,400.00	...	SA03
AAA	100015	2017	MZ	01.01.2017	1009000	1,017,862.23	...	FB01
BBB	900126	2017	IN	02.01.2017	5000110	234.50	...	SA03
...	...	...	...	...	...	...	...	...

Population

Sources: *D.M. Guy, D.R. Carmichael, and R. Whittington. Audit Sampling: In Introduction (5th Edition), John Wiley & Sons Inc., 2002.*
T.W. Hall, J.E. Hunton, and B.J. Pierce. Sampling Practices of Auditors in Public Accounting, Industry, and Government. Accounting Horizons, 16(2):125-136, 2002.

Related Work (Current Practice)

Statistical Audit Sampling - "Variable Sampling Techniques"

Objective: Estimate misstatement via conducting substantive audit procedures.

Difference Estimation

Sample

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100011	2017	SA	31.10.2016	1009000	2,685.00	...	FB01
AAA	100012	2017	MZ	31.10.2016	1008010	250,000.00	...	FB01
BBB	900124	2017	IN	01.11.2016	6019000	68,488.45	...	SA03

$$AuditValue_{Population} = \frac{BookValue_{Sample} - AuditValue_{Sample}}{SampleSize} * PopulationSize$$

Population

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100013	2017	SA	14.11.2016	1008010	4,000.00	...	FB01
AAA	100014	2017	MZ	25.11.2016	5000110	4,395.00	...	FB01
BBB	900125	2017	IN	01.12.2016	6019000	5,995.00	...	SA03
AAA	100015	2017	MZ	12.12.2016	1009000	180.00	...	FB01
BBB	900126	2017	IN	22.11.2016	9400000	18,500.00	...	SA03
BBB	900125	2017	IN	20.12.2016	5000110	697.15	...	SA03
...	...	...	...	...	...	...	...	...

Mean per Unit Estimation

Sample

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100011	2017	SA	31.10.2016	1009000	2,685.00	...	FB01
AAA	100012	2017	MZ	31.10.2016	1008010	250,000.00	...	FB01
BBB	900124	2017	IN	01.11.2016	6019000	68,488.45	...	SA03

$$AuditValue_{Population} = \frac{AuditValue_{Sample}}{SampleSize} * PopulationSize$$

Population

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100013	2017	SA	14.11.2016	1008010	4,000.00	...	FB01
AAA	100014	2017	MZ	25.11.2016	5000110	4,395.00	...	FB01
BBB	900125	2017	IN	01.12.2016	6019000	5,995.00	...	SA03
AAA	100015	2017	MZ	12.12.2016	1009000	180.00	...	FB01
BBB	900126	2017	IN	22.11.2016	9400000	18,500.00	...	SA03
BBB	900125	2017	IN	20.12.2016	5000110	697.15	...	SA03
...	...	...	...	...	...	...	...	...

Ratio Estimation

Sample

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100011	2017	SA	31.10.2016	1009000	2,685.00	...	FB01
AAA	100012	2017	MZ	31.10.2016	1008010	250,000.00	...	FB01
BBB	900124	2017	IN	01.11.2016	6019000	68,488.45	...	SA03

$$AuditValue_{Population} = \frac{BookValue_{Sample}}{AuditValue_{Sample}} * BookValue_{Population}$$

Population

| Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100013	2017	SA	14.11.2016	1008010	4,000.00	...	FB01
AAA	100014	2017	MZ	25.11.2016	5000110	4,395.00	...	FB01
BBB	900125	2017	IN	01.12.2016	6019000	5,995.00	...	SA03
AAA	100015	2017	MZ	12.12.2016	1009000	180.00	...	FB01
BBB	900126	2017	IN	22.11.2016	9400000	18,500.00	...	SA03
BBB	900125	2017	IN	20.12.2016	5000110	697.15	...	SA03
...	...	...	...	...	...	...	...	...

Monetary Unit Sampling

Population

Company	Entry ID	Fiscal Year	Type	Date	GL Account	Amount	...	T-Code
AAA	100015	2017	MZ	01.01.17	1009000	1,017,862.23	...	FB01
AAA	100012	2017	MZ	31.10.16	1008010	250,000.00	...	FB01
BBB	900126	2017	IN	21.01.16	6019000	95,400.00	...	SA03
BBB	900124	2017	IN	01.11.16	6019000	68,488.45	...	SA03
BBB	900126	2017	IN	22.11.16	9400000	18,500.00	...	SA03
BBB	900125	2017	IN	01.12.16	6019000	5,995.00	...	SA03
AAA	100014	2017	MZ	25.11.16	5000110	4,395.00	...	FB01
AAA	100013	2017	SA	14.11.16	1008010	4,000.00	...	FB01
AAA	100011	2017	SA	31.10.16	1009000	2,685.00	...	FB01
BBB	900125	2017	IN	20.12.16	5000110	697.15	...	SA03
BBB	900126	2017	IN	02.01.17	5000110	234.5	...	SA03
AAA	100015	2017	MZ	12.12.16	1009000	180	...	FB01
...	...	...	...	...	...	...	...	...

Sources: *D.M. Guy, D.R. Carmichael, and R. Whittington. Audit Sampling: In Introduction (5th Edition), John Wiley & Sons Inc., 2002.*
   T.W. Hall, J.E. Hunton, and B.J. Pierce. Sampling Practices of Auditors in Public Accounting, Industry, and Government. Accounting Horizons, 16(2):125-136, 2002.
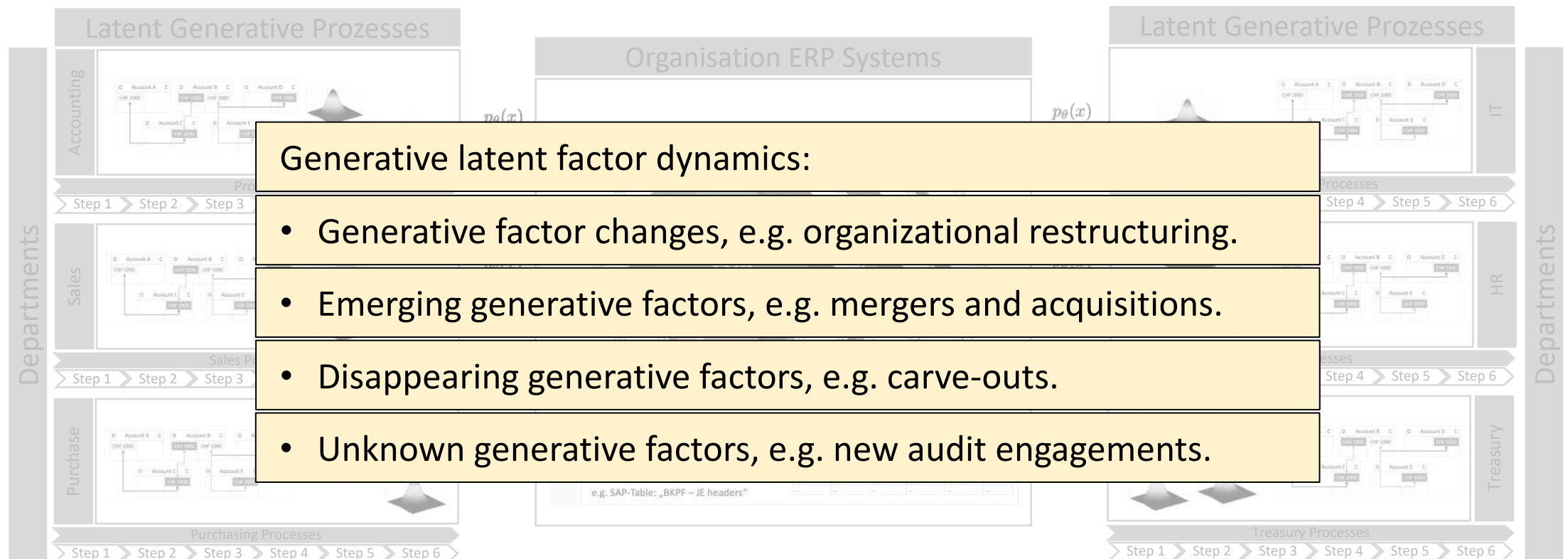
Motivation & Background
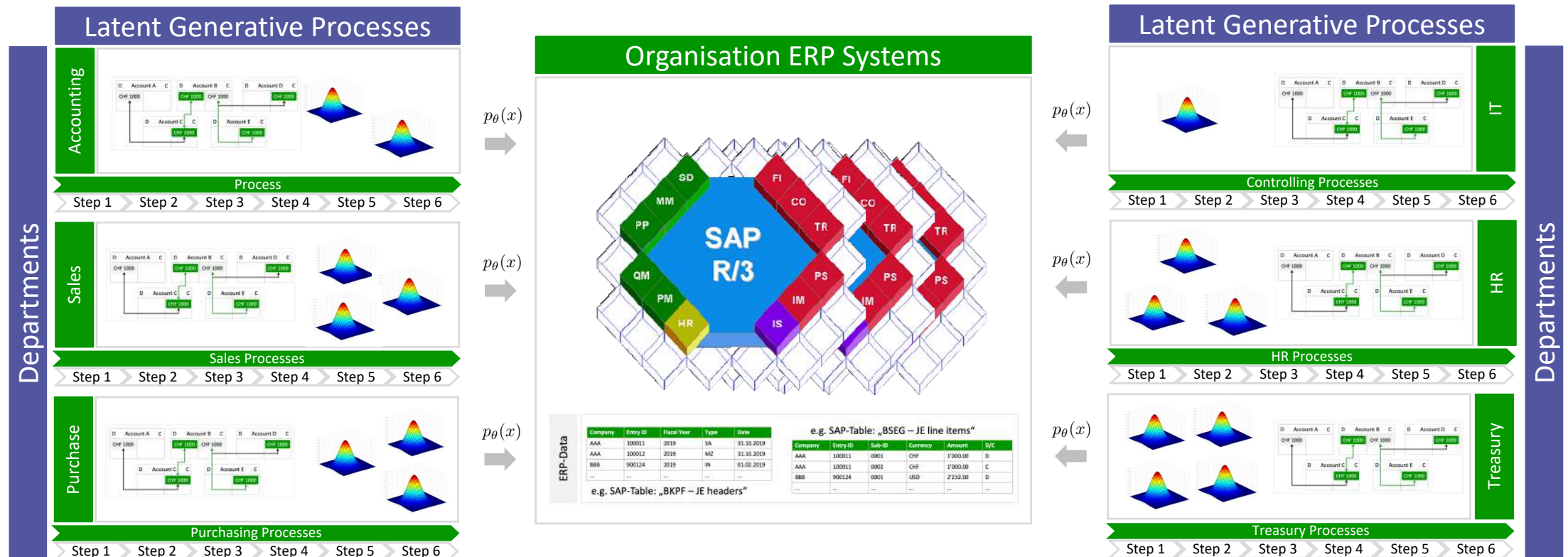
Latent Generative Business and Accounting Processes

Motivation & Background

Latent Generative Business and Accounting Processes

Generative latent factor dynamics:

- Generative factor changes, e.g. organizational restructuring.

- Emerging generative factors, e.g. mergers and acquisitions.

- Disappearing generative factors, e.g. carve-outs.

- Unknown generative factors, e.g. new audit engagements.

Motivation & Background

Questions of the Audit Practice
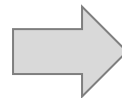

Challenge

"True and Fair Presentation"

"Big Data"

New Audit Engagement

Carve-Outs

Mergers & Acquisitions

Restructuring

…
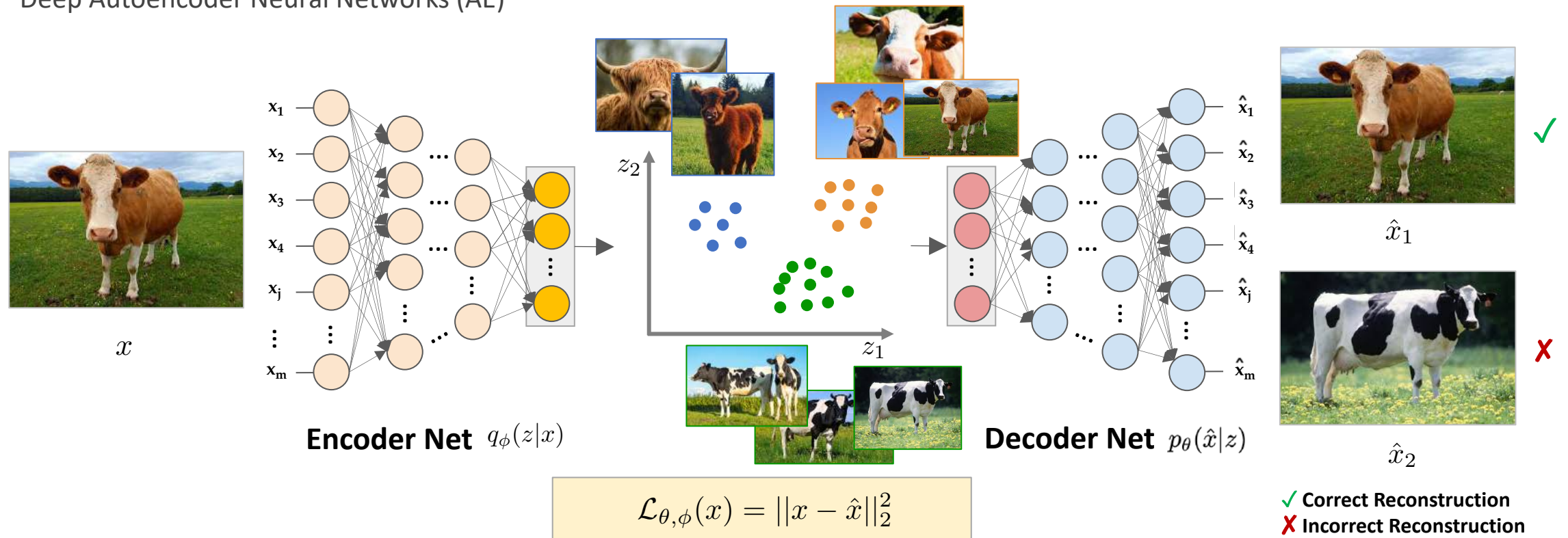
1. "Can we learn to uncover the generative latent factors?"

2. "Can we learn to disentangle the generative factors?"

3. "Can we learn representative audit sampling?"

Sources: R.J. Elder, A.D. Akresh, S.M. Glover, J.L. Higgs, and J. Liljegren. Audit Sampling Research: A Synthesis and Implications for Future Research. Auditing: A Journal of Practice & Theory, 32(1):99-129, 2013.
M. Reichert and B. Weber. Ad hoc Changes of Process Instances. Enabling Flexibility in Process-Aware Information Systems, pages 153-217. Springer, 2012.

Methodology

1. Can we learn to uncover the generative latent factors?

Deep Autoencoder Neural Networks (AE)

Encoder Net $q_\phi(z|x)$

Decoder Net $p_\theta(\hat{x}|z)$

$$\mathcal{L}_{\theta,\phi}(x) = ||x - \hat{x}||_2^2$$

✓ **Correct Reconstruction**
✗ **Incorrect Reconstruction**

Sources: G.E. Hinton and R. R. Salakhutinov. Reducing the Dimensionality of Data with Neural Networks. Science 313, no. 5786, pages 504-507, 2006.
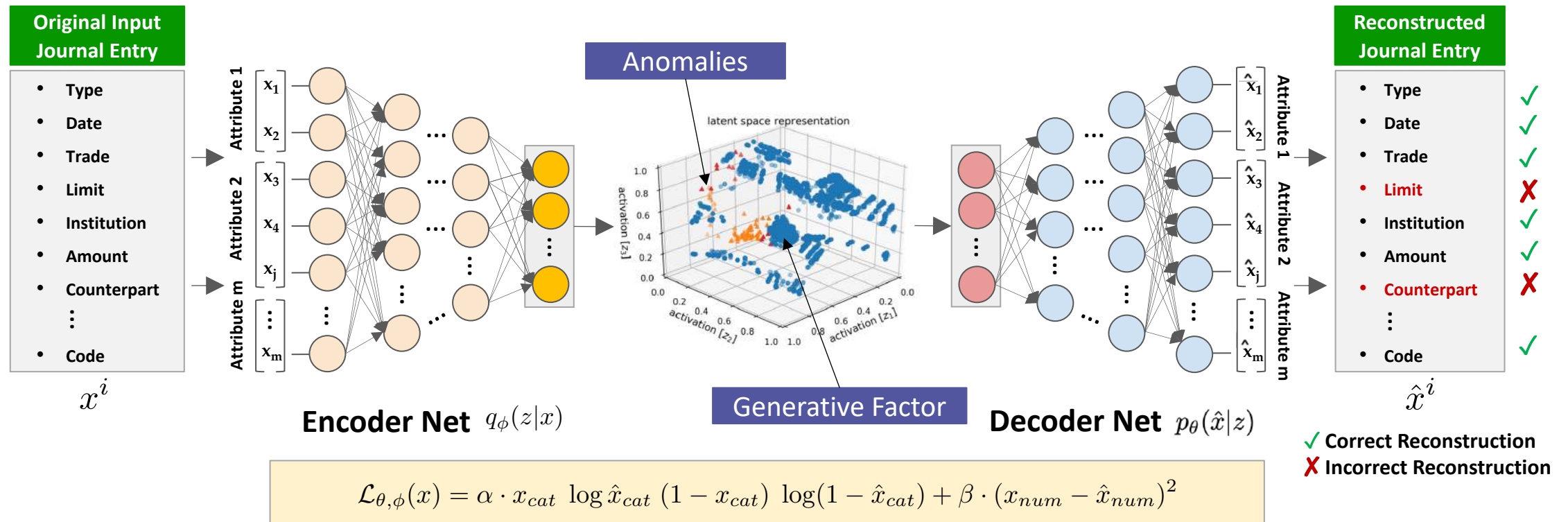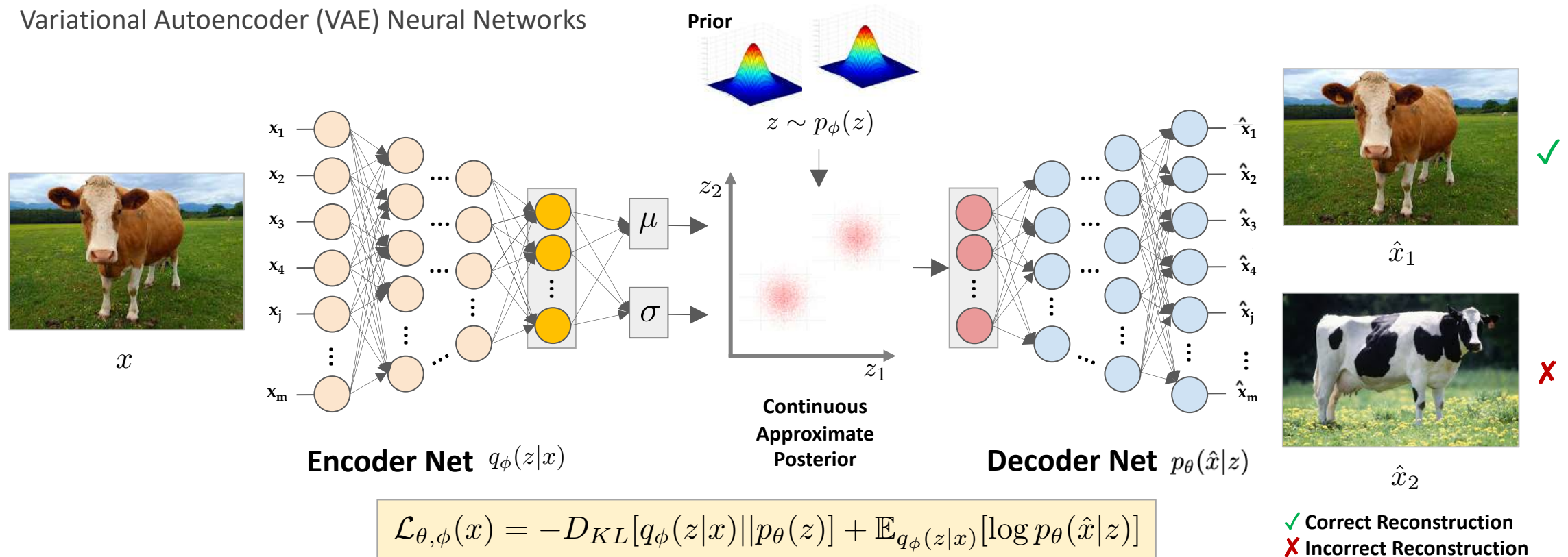   Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspective. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, no. 8, pages 1798-1828, 2013.

Methodology

1. Can we learn to uncover the generative latent factors?   Schreyer & Sattarov *et al.*, 2017

Deep Autoencoder Neural Networks (AE)

Original Input Journal Entry
- Type
- Date
- Trade
- Limit
- Institution
- Amount
- Counterpart
- ⋮
- Code

x^i

Encoder Net $q_\phi(z|x)$

Anomalies

latent space representation

Generative Factor

Decoder Net $p_\theta(\hat{x}|z)$

Reconstructed Journal Entry
- Type ✓
- Date ✓
- Trade ✓
- Limit ✗
- Institution ✓
- Amount ✓
- Counterpart ✗
- ⋮
- Code ✓

$\hat{x}^i$

✓ **Correct Reconstruction**
✗ **Incorrect Reconstruction**

$$\mathcal{L}_{\theta,\phi}(x) = \alpha \cdot x_{cat} \, \log \hat{x}_{cat} \, (1 - x_{cat}) \, \log(1 - \hat{x}_{cat}) + \beta \cdot (x_{num} - \hat{x}_{num})^2$$

Sources: G.E. Hinton and R. R. Salakhutinov. *Reducing the Dimensionality of Data with Neural Networks. Science 313, no. 5786, pages 504-507, 2006.*
    M. Schreyer, T. Sattarov, D. Borth, B. Reimer, A. Dengel. *Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks. arXiv preprint. arXiv: 1709.05254, 2017*

Methodology

2. Can we learn to disentangle the generative factors?

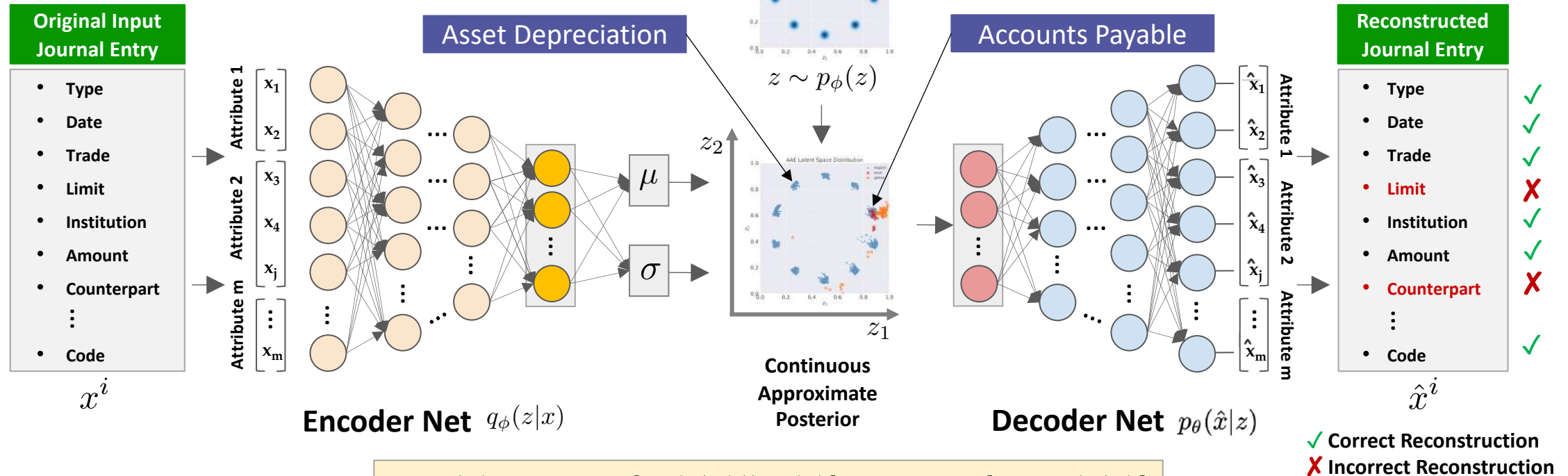Variational Autoencoder (VAE) Neural Networks

Prior

$z \sim p_\phi(z)$

z_2

z_1

**Continuous
Approximate
Posterior**

Encoder Net $q_\phi(z|x)$

x

Decoder Net $p_\theta(\hat{x}|z)$

μ

σ

$\hat{x}_1$
$\hat{x}_2$
$\hat{x}_3$
$\hat{x}_4$
$\hat{x}_j$
$\hat{x}_m$

x_1
x_2
x_3
x_4
x_j
x_m

$\hat{x}_1$ ✓

$\hat{x}_2$ ✗

✓ **Correct Reconstruction**
✗ **Incorrect Reconstruction**

$$\mathcal{L}_{\theta,\phi}(x) = -D_{KL}[q_\phi(z|x)||p_\theta(z)] + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\hat{x}|z)]$$

Source: D.P. Kingma and M. Welling. Auto-encoding Variational Bayes. arXiv preprint. arXiv:1312.6114, 2013.

Methodology

2. Can we learn to disentangle the generative factors?

Schreyer & Sattarov *et al.*, 2019

Variational Autoencoder (VAE) Neural Networks

Prior

$z \sim p_\phi(z)$

Original Input Journal Entry
• **Type**
• **Date**
• **Trade**
• **Limit**
• **Institution**
• **Amount**
• **Counterpart**
⋮
• **Code**

Asset Depreciation

Accounts Payable

z_2

z_1

Continuous Approximate Posterior

Reconstructed Journal Entry
• **Type** ✓
• **Date** ✓
• **Trade** ✓
• **Limit** ✗
• **Institution** ✓
• **Amount** ✓
• **Counterpart** ✗
⋮
• **Code** ✓

x^i

$\hat{x}^i$

Encoder Net $q_\phi(z|x)$

μ

σ

Decoder Net $p_\theta(\hat{x}|z)$

✓ **Correct Reconstruction**
✗ **Incorrect Reconstruction**

$$\mathcal{L}_{\theta,\phi}(x) = -D_{KL}[q_\phi(z|x)||p_\theta(z)] + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\hat{x}|z)]$$

Source: D.P. Kingma and M. Welling. Auto-encoding Variational Bayes. arXiv preprint. arXiv:1312.6114, 2013.

M. Schreyer, T. Sattarov, C. Schulze, B. Reimer, D. Borth. Detection of Accounting Anomalies in the Latent Space using Adversarial Autoencoder Neural Networks. KDD'19 Workshop on Anomaly Detection in Finance, 2019.

M. Schreyer, T. Sattarov, B. Reimer, D. Borth. Adversarial Learning of Deepfakes in Accounting. NeurIPS'19 Workshop on Robust AI in Financial Services, 2019.

Methodology

3. Can we learn representative Audit Sampling?

Learning Objective

x

Encoder Net $q_\phi(z|x)$

z_2

z_1

**Discrete/Quantised
Approximate Posterior**

Decoder Net $p_\theta(\hat{x}|z)$

$\hat{x}_1$

$\hat{x}_2$

✓ **Correct Reconstruction**
✗ **Incorrect Reconstruction**
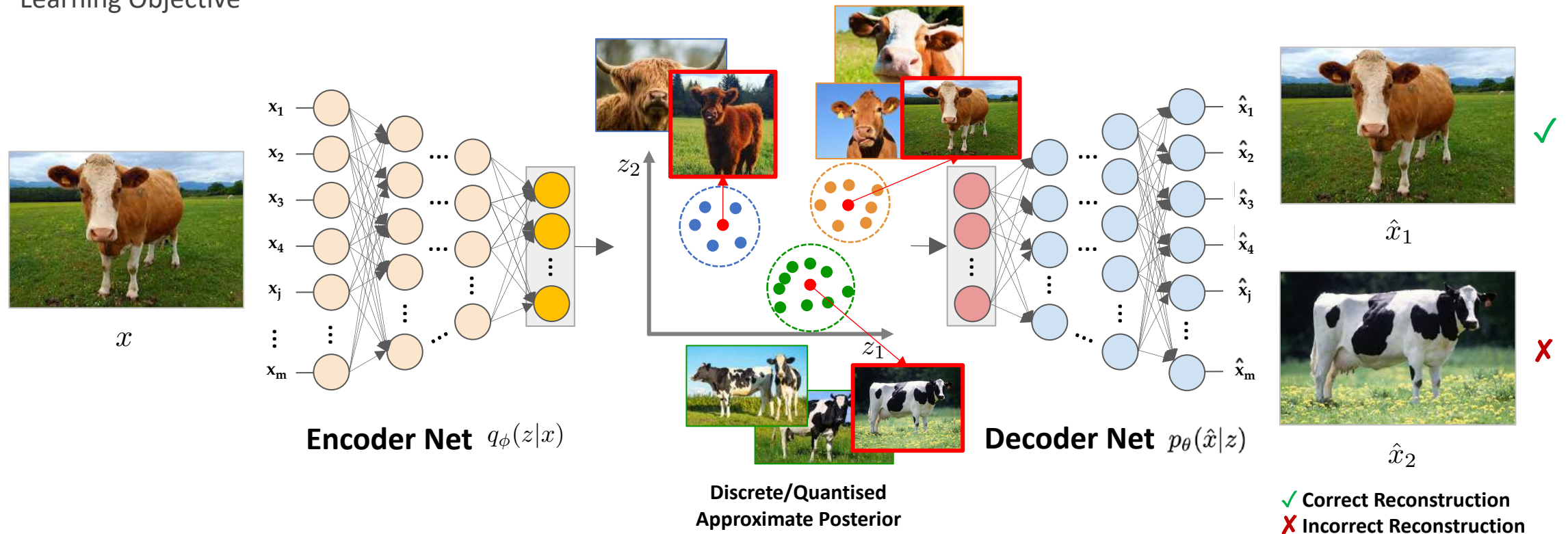
Methodology

3. Can we learn representative Audit Sampling?   This work

Learning Objective

Methodology

3. Can we learn representative Audit Sampling?

Embedding Space

Vector Quantized Variational Autoencoder
(VQ-VAE) Neural Networks

D

$e_1 \quad e_2 \qquad e_i \qquad e_k$

$\nabla_z \mathcal{L}$

**Quantised
Approximate
Posterior**

$z_e(x)$

$q(z|x)$

$z_q(x)$

x

$x_1, x_2, x_3, x_4, x_j, x_m$

$\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4, \hat{x}_j, \hat{x}_m$

Encoder Net q_θ

Decoder Net p_ϕ

$\hat{x}_1$ ✓

$\hat{x}_2$ ✗

✓ **Correct Reconstruction**
✗ **Incorrect Reconstruction**

$$\mathcal{L}_{\theta,\phi}(x) = \log p(x|z_q(x)) \; + \; \alpha||sg[z_e(x)] - e||_2^2 + \; \beta||z_e(x) - sg[e]||_2^2$$

Sources: A. van den Oord, O. Vinyals. Neural Discrete Representation Learning. Advances in Neural Information Processing Systems. Pages: 6306-6315, 2017.

Methodology

3. Can we learn representative Audit Sampling?

Vector Quantized Variational Autoencoder
(VQ-VAE) Neural Networks

Embedding Space
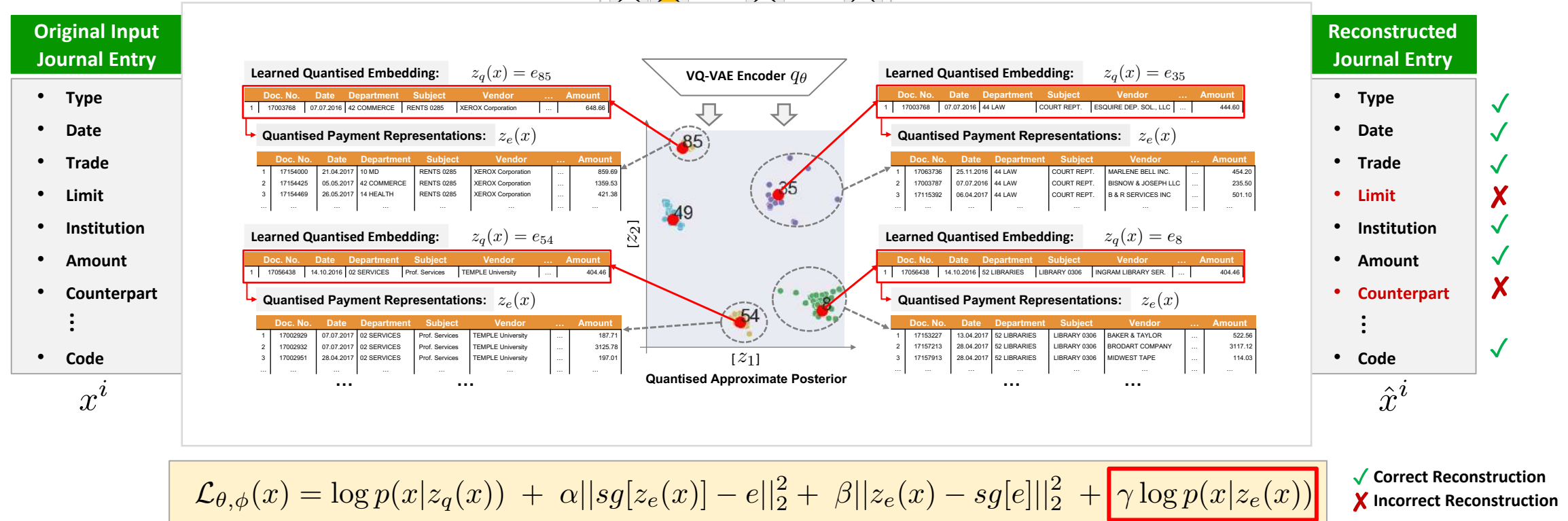
$$\mathcal{L}_{\theta,\phi}(x) = \log p(x|z_q(x)) \;+\; \alpha||sg[z_e(x)] - e||_2^2 \;+\; \beta||z_e(x) - sg[e]||_2^2 \;+\; \boxed{\gamma \log p(x|z_e(x))}$$

✓ **Correct Reconstruction**
✗ **Incorrect Reconstruction**

Sources: A. van den Oord, O. Vinyals. Neural Discrete Representation Learning. Advances in Neural Information Processing Systems. Pages: 6306-6315, 2017.
            V. Fortuin, M. Hüser, F. Locatello, H. Strathmann, G. Rätsch. SOM-VAE: Interpretable Discrete Representation Learning on Time Series. arXiv preprint, preprint arXiv:1806.02199, 2018.

Methodology

3. Can we learn representative Audit Sampling?

Training objective: $\mathcal{L}(x) = \log p(x|z_q(x)) + \alpha||sg[z_e(x)] - e||_2^2 + \beta||z_e(x) - sg[e]||_2^2 + \boxed{\gamma \log p(x|z_e(x))}$
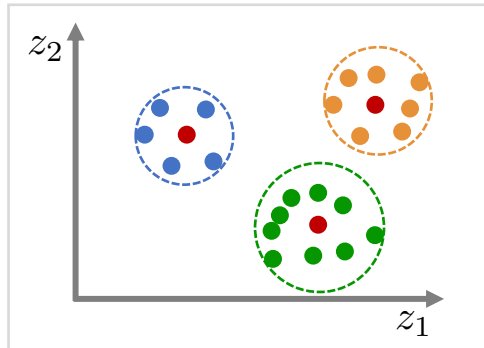
q-Reconstruction Loss:

Quantised embeddings z_q are informative representations of the encoder output:

$$\mathcal{L}_a(x) = \log p(x|z_q(x))$$
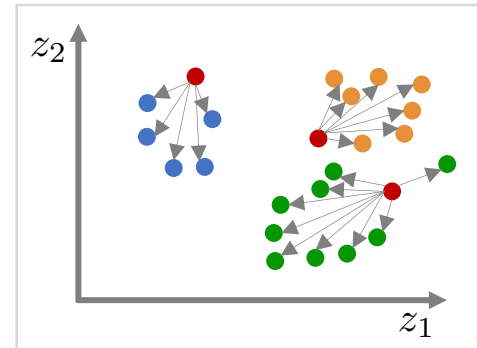
Embedding Loss:

Guarantees that the $e = z_q$ quantised embeddings z_e commit to the encoder output:

$$\mathcal{L}_b(x) = \alpha||sg[z_e(x)] - e||_2^2$$
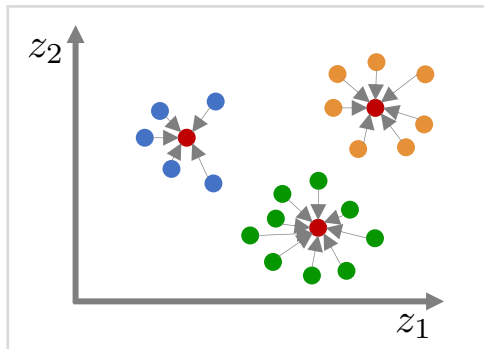
Commitment Loss:

Guarantees that the encoder output z_e commits to one of the embeddings $e = z_q$:

$$\mathcal{L}_c(x) = \beta||z_e(x) - sg[e]||_2^2$$
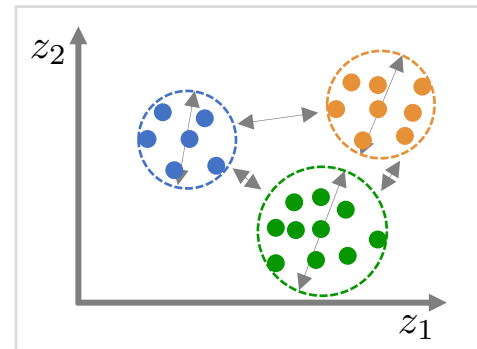
e-Reconstruction Loss:

Encoder embeddings z_e are informative representations of the data semantics:

$$\mathcal{L}_d(x) = \gamma \log p(x|z_e(x))$$

Experimental Setup

3. Can we learn representative Audit Sampling?

Datasets and Data Preprocessing:

Dataset A: Philadelphia "City Payments"

- $4.2 billion city payments of the fiscal year 2017
- 60 offices, departments, boards and committees
- N=238,894 payments: 10 categorical, 1 numerical attribute
- 8,565 'one-hot' encoded dimensions: $x^i \in \mathcal{R}^{8,565}$

City of **Philadelphia**

check_date	document_no	dept	department_title	char_	character_title	sub_obj	sub_obj_title
2016-11-25T00:00:00Z	CHEK17063379	1	01 CITY COUNCIL	3	03 MATERIALS AND SUPPLIES	325	PRINTING 0325
2017-01-26T00:00:00Z	CHEK17087297	1	01 CITY COUNCIL	2	02 PURCHASE OF SERVICES	240	ADVERTISING/PROMOTIONAL ACTIVITIES
2016-08-04T00:00:00Z	CHEK17012548	1	01 CITY COUNCIL	2	02 PURCHASE OF SERVICES	240	ADVERTISING/PROMOTIONAL ACTIVITIES
2016-08-26T00:00:00Z	ACHD17028318	1	01 CITY COUNCIL	2	02 PURCHASE OF SERVICES	210	POSTAGE 0210
2017-06-30T00:00:00Z	ACHD17192218	1	01 CITY COUNCIL	3	03 MATERIALS AND SUPPLIES	309	CORDAGE AND FIBERS 0309
2017-06-08T00:00:00Z	CHEK17141080	1	01 CITY COUNCIL	3	03 MATERIALS AND SUPPLIES	325	PRINTING 0325
2017-04-12T00:00:00Z	CHEK17117562	1	01 CITY COUNCIL	2	02 PURCHASE OF SERVICES	255	DUES 0255
2017-04-13T00:00:00Z	ACHD17151640	1	01 CITY COUNCIL	3	03 MATERIALS AND SUPPLIES	304	BOOKS AND OTHER PUBLICATIONS 0304

Dataset B: Chicago "Vendor Payments"

- $5.3 billion vendor payments of the fiscal years 1996 -2020
- 30 departments, 928 contracts
- N=72,814 payments: 7 categorical, 1 numerical attribute
- 2,354 'one-hot' encoded dimensions: $x^i \in \mathcal{R}^{2,354}$

CHICAGO DATA PORTAL

VOUCHER NUMBER	AMOUNT	CHECK DATE	DEPARTMENT NAME	CONTRACT NUMBER	VENDOR NAME	CASHED
CV54195400322	94.80	07/01/2019	DEPT OF COMMUNITY DEV...	50236	18TH STREET. DEVELOPME...	Yes
CVIP174102790	-6.56	04/30/2019	DEPARTMENT OF HEALTH	64612	RESPIRATORY HEALTH ASS...	No
CVIP175004687	1,380.00	01/02/2019	DEPT OF FAMILY AND SUPP...	33322	FEATHERFIST	Yes
CVIP182500015	2,686.10	01/08/2019		82037	PHALANX FAMILY SERVICES	Yes
CVIP182500016	12,831.19	01/09/2019		82037	PHALANX FAMILY SERVICES	Yes
CVIP182500017	48,589.00	01/25/2019		87489	CHICAGO CITYWIDE LITERA...	Yes
CVIP182500021	115.47	04/29/2019		89573	CATHOLIC CHARITIES OF T...	Yes
CVIP182500022	963.83	04/29/2019		89573	CATHOLIC CHARITIES OF T...	Yes

Experimental Setup

3. Can we learn representative Audit Sampling?

Architectural Details and Training Setup:

Architectural details:

- Encoder & Decoder networks: 10 fully connected layers
- Input-Hidden: Leaky ReLUs (α = 0.4), Output: Sigmoid
- Discrete latent vectors: $e_j \sim \mathcal{U}(-1, 1)$, $e_j \in \mathcal{R}^2$
- Codebook sizes: $K \in \{2^3, 2^4, 2^5, 2^6, 2^7\}$

Neurons per Layer:

	Net	Dataset	ℓ = 1	2	3	4	...	10	11
Encoder:	$q_\theta(z\|x)$	A	5,096	2,048	1,024	512	...	4	2
Decoder:	$p_\phi(\hat{x}\|z)$	A	2	4	8	16	...	2,048	5,096
Encoder:	$q_\theta(z\|x)$	B	2048	1024	512	256	...	4	2
Decoder:	$p_\phi(\hat{x}\|z)$	B	2	4	8	16	...	1024	2048

Training setup:

- SGD max. 4,000 training epochs (early stopping)
- Mini-batch size: 128 journal entries
- Optimisation: ADAM, $\beta_1 = 0.9$, $\beta_2 = 0.999$
- Learning rate: 1e-4 (cosine update schedule)

Hyperparameter sweeps:

- $\alpha, \beta, \gamma \in [0.0, 0.5, 1.0, 1.5, 2.0]$
- Training epoch "warm-up" : 0, 20, 50, 100 epochs
- Discrete vector initialisations: uniform, circle, zeros, normal
- 5 random network parameter initialisations
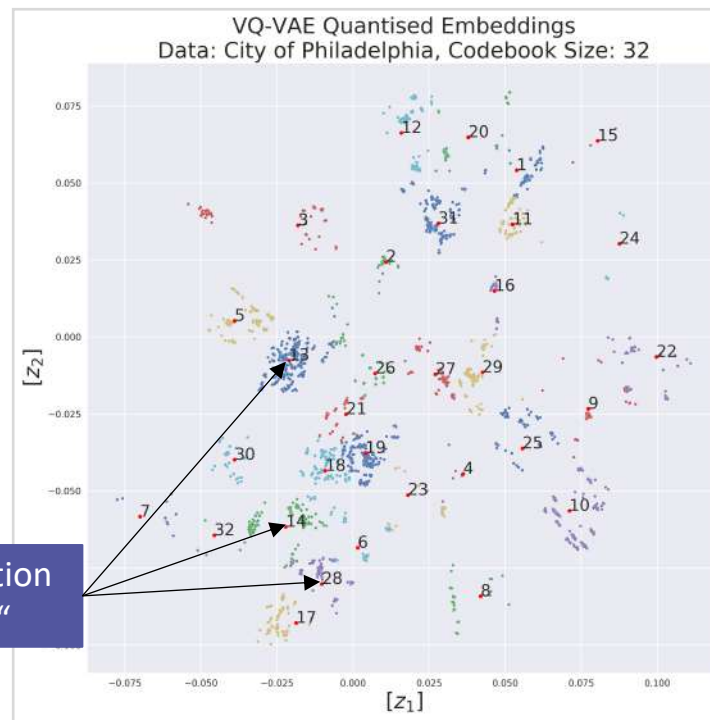- Discrete vector update: direct vs. EMA

Experimental Results

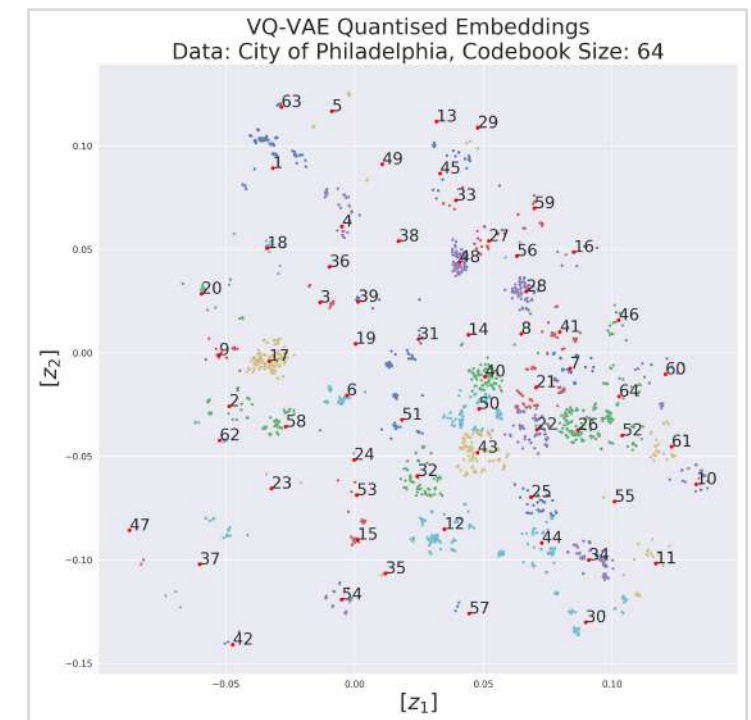**3. Can we learn representative Audit Sampling?**

Exemplary Qualitative Results – Philadelphia City Payments:
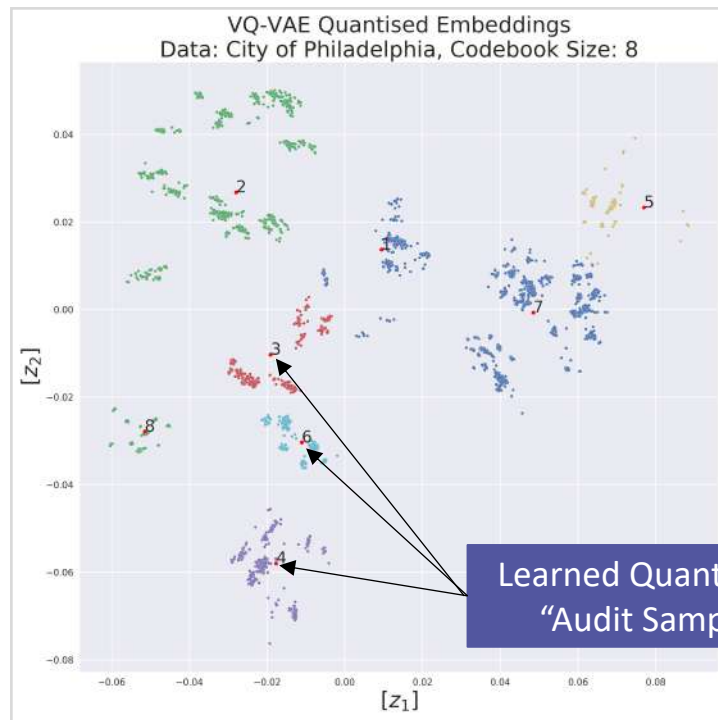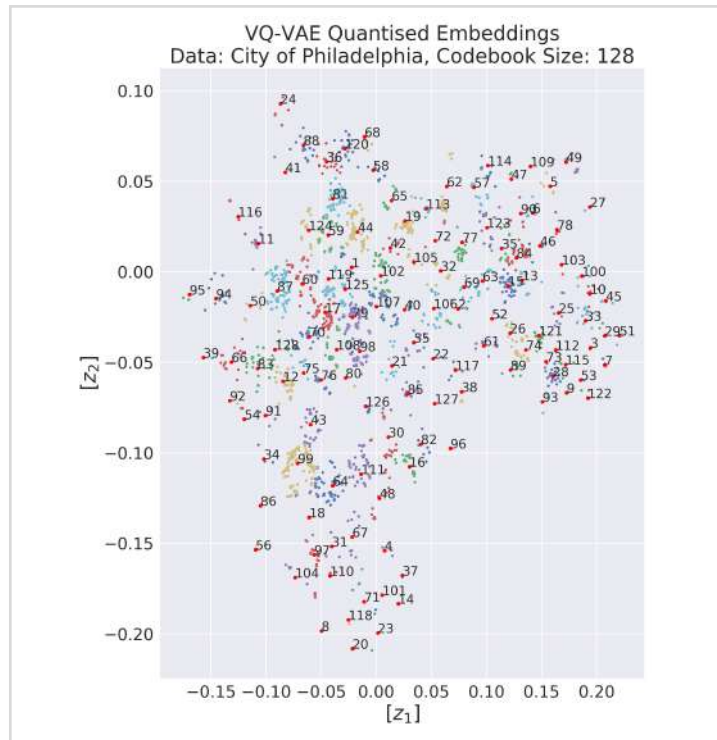

Audit Sample Size N = 8

Audit Sample Size N = 32
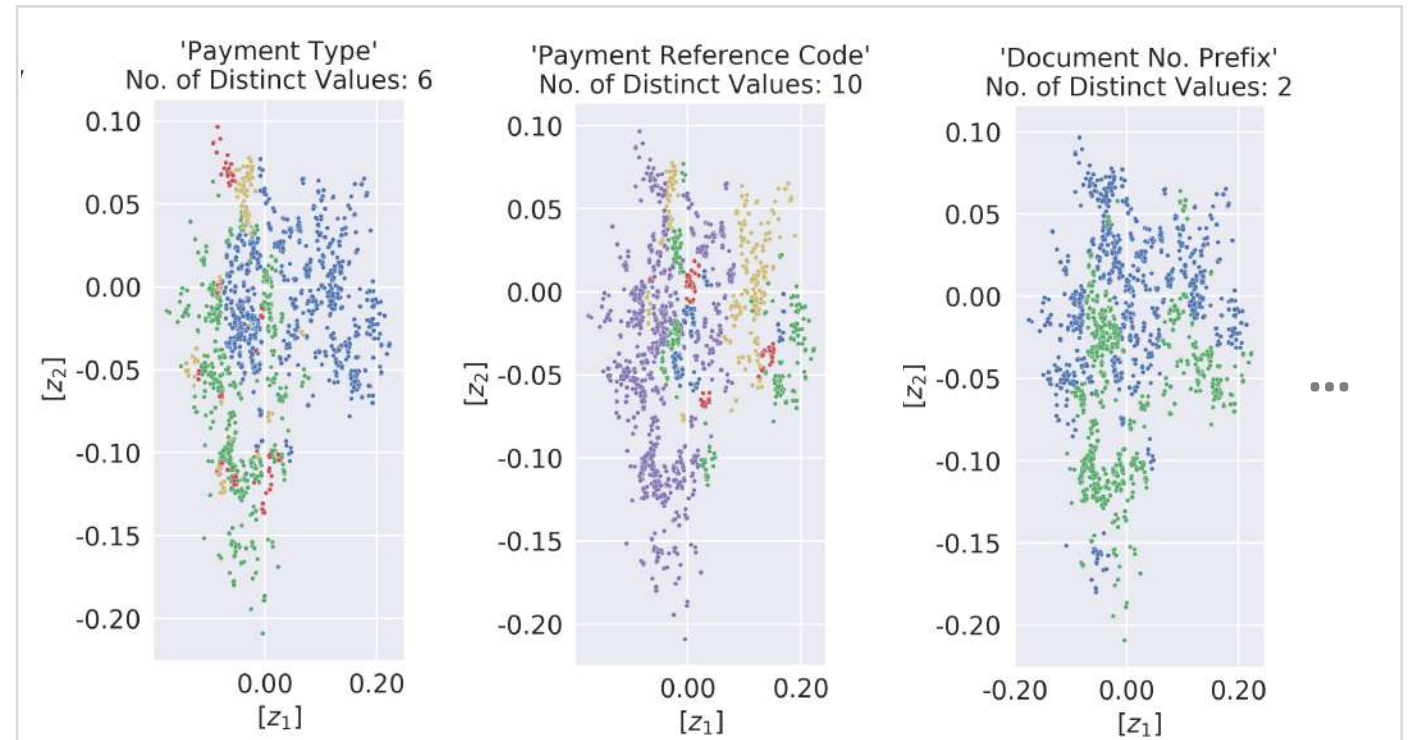
Audit Sample Size N = 64

Experimental Results

3. Can we learn representative Audit Sampling?

Exemplary Qualitative Results – Philadelphia City Payments:
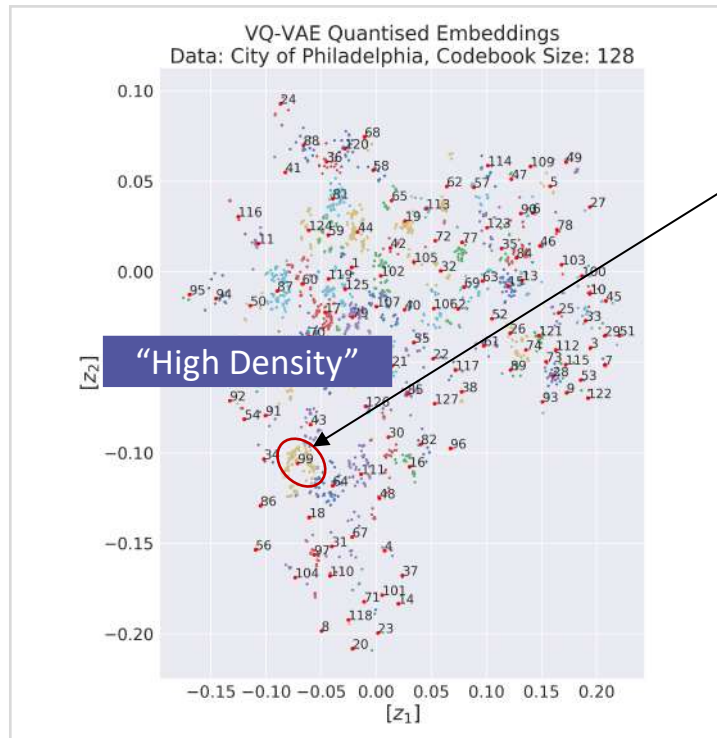

Sample Size N = 128

Disentangled City Payment Attributes

Experimental Results

3. Can we learn representative Audit Sampling?

Exemplary Qualitative Results – Philadelphia City Payments:

VQ-VAE Quantised Embeddings
Data: City of Philadelphia, Codebook Size: 128

"High Density"

Sample Size N = 128

Quantisation - "Audit Sample" (99)

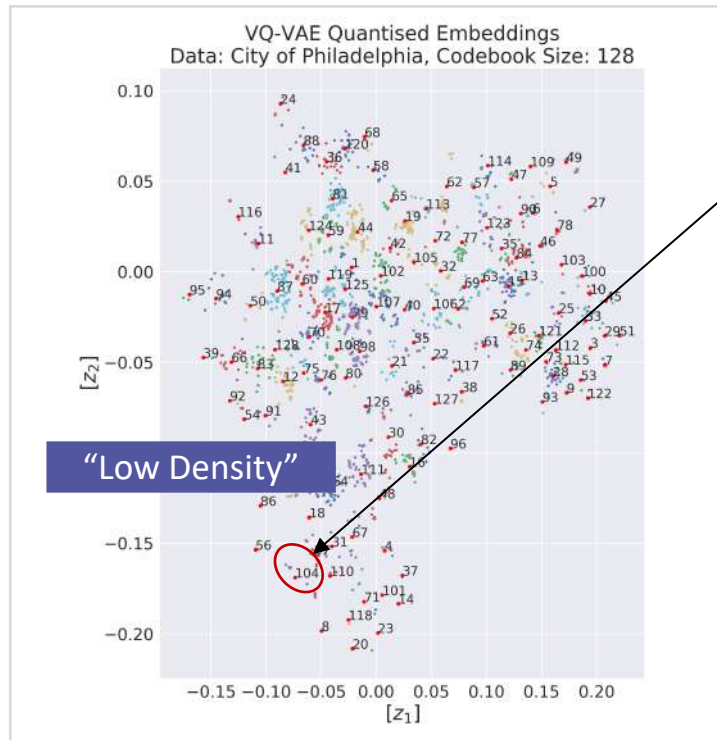| FY	Date	Payment	Department	Type	Object	Vendor	Details	Amount
2017	20.04.17	CHEK17121800	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	G.L. SAYRE INCORPORATED	auto parts	132.59

Quantised Representation - City Payments

FY	Date	Payment	Department	Type	Object	Vendor	Details	Amount
2017	14.07.16	CHEK17005079	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	TRANSTECK INC	auto parts	-32.17
2017	20.04.17	CHEK17121800	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	G.L. SAYRE INCORPORATED	auto parts	132.59
2017	04.05.17	CHEK17125451	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	TRANSTECK INC	auto parts	37.60
2017	01.06.17	CHEK17136892	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	CHAPMAN FORD SALES	auto parts	-293.71
2017	01.06.17	CHEK17138737	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	JACK DOHENY COMPANIES INC	auto parts	539.70
2017	08.06.17	CHEK17139311	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	TRANSTECK INC	auto parts	11.46
2017	29.06.17	CHEK17149083	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	TRANSTECK INC	auto parts	15.21
2017	29.06.17	CHEK17150281	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	G.L. SAYRE INCORPORATED	auto parts	78.84
2017	11.08.16	CHEK17014102	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	TRANSTECK INC	auto parts	136.76
2017	25.11.16	CHEK17063268	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	GILES AND RANSOME INCORPORATED	auto parts	570.22
2017	12.01.17	CHEK17082148	25 FLEET MANAGEMENT	03 MATERIALS AND SUPPLIES	MOTOR VEHICLE PARTS AND ACCESSORIES 0328	CHAPMAN CHEVROLET LLC	auto parts	2.62
…	…	…	…	…	…	…	…	…

Experimental Results

3. Can we learn representative Audit Sampling?

Exemplary Qualitative Results – Philadelphia City Payments:

Quantisation - "Audit Sample" (104)

| FY	Date	Payment	Department	Type	Object	Vendor	Details	Amount
2017	23.06.17	ACHD17189704	52 FREE LIBRARY	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	363.2

Quantised Representation - City Payments

| FY	Date	Payment	Department	Type	Object	Vendor	Details	Amount
2017	22.07.16	ACHD17011732	13 FIRE	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	228.68
2017	23.06.17	ACHD17189704	52 FREE LIBRARY	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	113.00
2017	19.08.16	ACHD17026108	52 FREE LIBRARY	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	354.20
2017	28.10.16	ACHD17059715	22 HUMAN SERVICES	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	539.20
2017	23.12.16	ACHD17090986	42 COMMERCE	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	213.00
2017	13.01.17	ACHD17102628	25 FLEET MANAGEMENT	02 PURCHASE OF SERVICES	REPAIR AND MAINTENANCE CHARGES 0260	INTERNETWORK SERVICES INC	auto parts	5953.50
2017	20.01.17	ACHD17105752	22 HUMAN SERVICES	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	479.08
2017	17.02.17	ACHD17121338	22 HUMAN SERVICES	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	320.60
2017	23.06.17	ACHD17189446	22 HUMAN SERVICES	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	GREYHOUND LINES, INC.	payment voucher	530.80
2017	23.06.17	ACHD17189703	28 WATER	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	91.00
2017	23.06.17	ACHD17189763	28 WATER	02 PURCHASE OF SERVICES	TRANSPORTATION 0211	AMERICAN EXPRESS	American Express travel vouchers	519.10
...	...	...	...	...	...	...	...	...

Sample Size N = 128

Experimental Results

3. Can we learn representative Audit Sampling?

Exemplary Qualitative Results – Chicago City Vendor Payments:
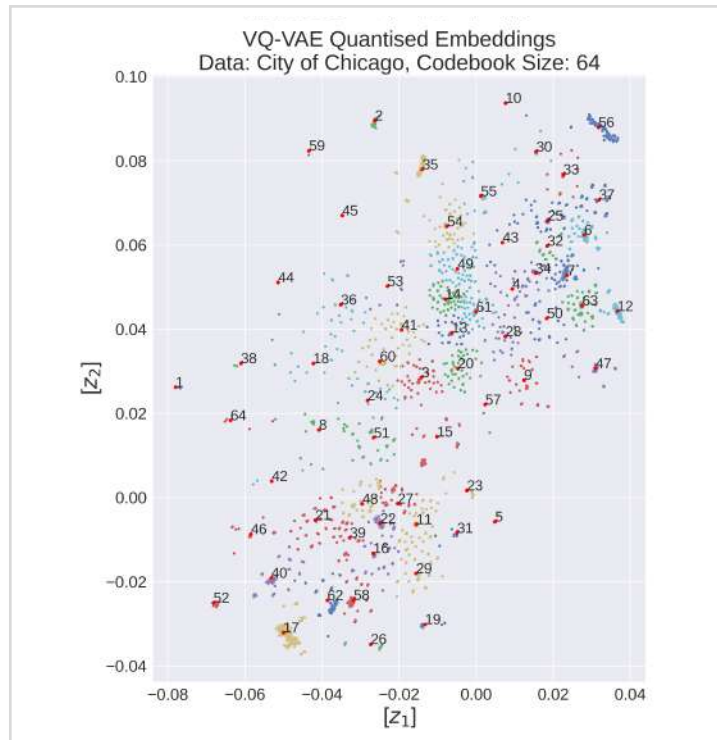
Sample Size N = 64
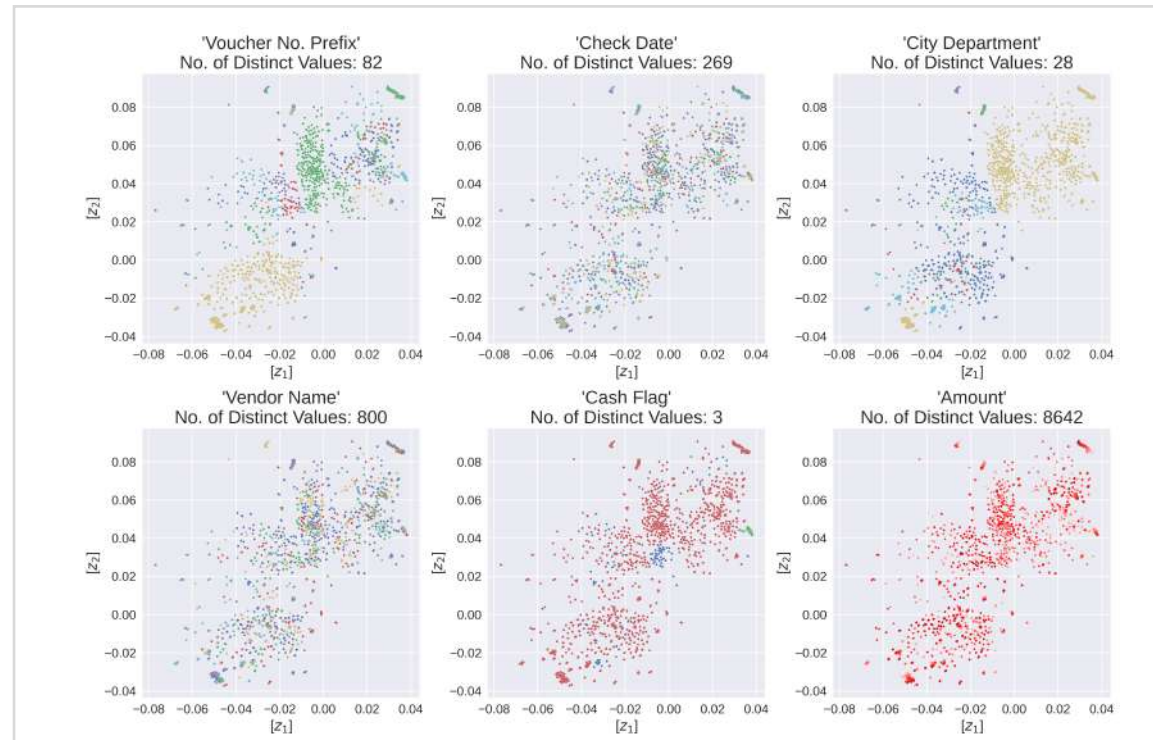
Disentangled City Vendor Payment Attributes

Experimental Results

3. Can we learn representative Audit Sampling?

Exemplary Quantitative Results – Both City Payment Datasets:

<div style="background:green;color:white">Quantisation Results</div>

Data	K	$\mathcal{L}^{zq}_{MSE}$	$\mathcal{L}^{ze}_{MSE}$	$\mathcal{P}_{erp}$	$\mathcal{P}_{urity}$
A	2^3	0.577 ± 0.13	0.453 ± 0.24	5.394 ± 3.81	0.864 ± 0.09
A	2^4	0.454 ± 0.06	0.312 ± 0.34	12.668 ± 0.62	0.845 ± 0.01
A	2^5	0.417 ± 0.10	0.281 ± 0.33	22.024 ± 1.29	0.853 ± 0.02
A	2^6	0.382 ± 0.05	0.232 ± 0.13	37.677 ± 1.72	0.872 ± 0.01
A	2^7	0.345 ± 0.17	0.208 ± 0.19	59.755 ± 2.83	0.890 ± 0.01
B	2^3	1.675 ± 0.03	1.535 ± 0.04	6.082 ± 0.17	0.440 ± 0.03
B	2^4	1.622 ± 0.06	1.424 ± 0.10	9.788 ± 1.12	0.416 ± 0.01
B	2^5	1.587 ± 0.07	1.351 ± 0.15	17.941 ± 2.87	0.377 ± 0.06
B	2^6	1.467 ± 0.02	1.121 ± 0.04	31.217 ± 5.04	0.373 ± 0.01
B	2^7	1.407 ± 0.05	1.071 ± 0.07	42.171 ± 9.39	0.321 ± 0.02

Philadelphia: (rows A), Chicago: (rows B)

Variances originate from parameter initialization using five distinct random seeds.

<div style="background:green;color:white">Disentanglement Results</div>

Data	K	β-VAE [24]	Fac-VAE [28]	MIG [11]	DCI [14]
A	2^3	0.160 ± 0.02	0.110 ± 0.06	0.025 ± 0.02	0.039 ± 0.01
A	2^4	0.166 ± 0.01	0.108 ± 0.06	0.029 ± 0.01	0.038 ± 0.01
A	2^5	0.166 ± 0.03	0.119 ± 0.01	0.031 ± 0.02	0.046 ± 0.01
A	2^6	0.182 ± 0.02	0.134 ± 0.01	0.068 ± 0.08	0.061 ± 0.04
A	2^7	0.193 ± 0.01	0.149 ± 0.03	0.081 ± 0.06	0.139 ± 0.07
B	2^3	0.244 ± 0.01	0.142 ± 0.02	0.051 ± 0.03	0.690 ± 0.03
B	2^4	0.240 ± 0.01	0.145 ± 0.03	0.057 ± 0.03	0.703 ± 0.02
B	2^5	0.277 ± 0.04	0.144 ± 0.02	0.053 ± 0.04	0.709 ± 0.01
B	2^6	0.290 ± 0.02	0.144 ± 0.01	0.067 ± 0.01	0.715 ± 0.01
B	2^7	0.324 ± 0.01	0.146 ± 0.01	0.080 ± 0.03	0.717 ± 0.01

Philadelphia: (rows A), Chicago: (rows B)

Variances originate from parameter initialization using five distinct random seeds.

Increased codebook size K:
➡ Decrease in quantisation error and perplexity. Convergence of embedding error.

Increased codebook size K:
➡ Increase in payment attribute disentanglement across all disentanglement metrics.

➩ The qualitative and quantitative results indicate that the VQ-VAE provides the ability to learn embeddings that quantise the latent generative factors of accounting data and therefore constitute a representative audit sample.

Happy Sampling! 😎

University of St.Gallen

Contacts: **Marco Schreyer**
Artificial Intelligence & Machine Learning
University of St.Gallen (HSG)
marco.schreyer@unisg.ch

Timur Sattarov
Securities and Money Market Statistics (S5)
Deutsche Bundesbank
timur.sattarov@bundesbank.de