Graph-Based Anomaly Detection Using MDL Principle

Aluna Wang
HEC Paris
Hi! Paris Center

The Learning Objectives

Gain some understanding of the role of an **AI architect**

- Transform business problems into data science problems
- Envision, build, deploy and operationalize the AI solutions to business problems

Anomaly detection AI for mitigating corporate misconduct risk

- Corporate Risk Management
- Anomaly detection in metadata (multi-dimensional data points) and graph data

HEC PARIS

The Roadmap

Anomaly Detection

- What is it?
- What are some of the high-impact applications?
- What are the challenges and opportunities?

Risk Management

- What is the COSO enterprise risk management framework?
- How is anomaly detection relevant to risk management?

Accounting process and data

- What do we mean by financial transactions in the accounting information system?
- How to feed accounting data into machine learning algorithms?

A Real Case

Two Methods

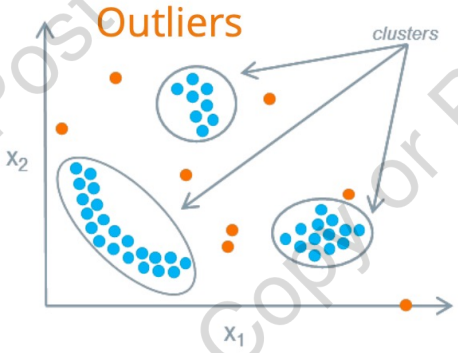
Let's dive in!

HEC PARIS

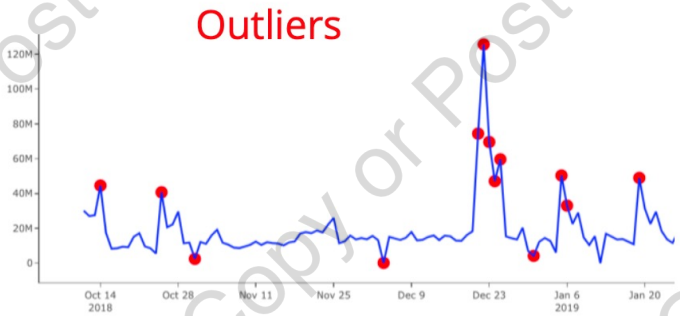
Anomaly Detection

What is anomaly detection?

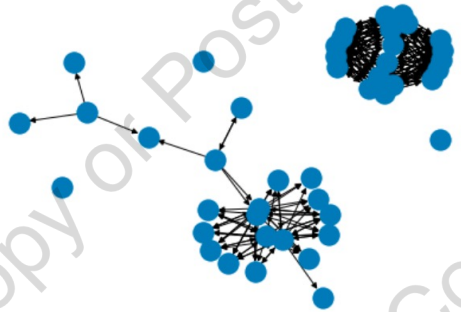
Simply put, anomaly detection aims to identify the data samples that are deviant from the general distribution.



In tabular data



In time-series data



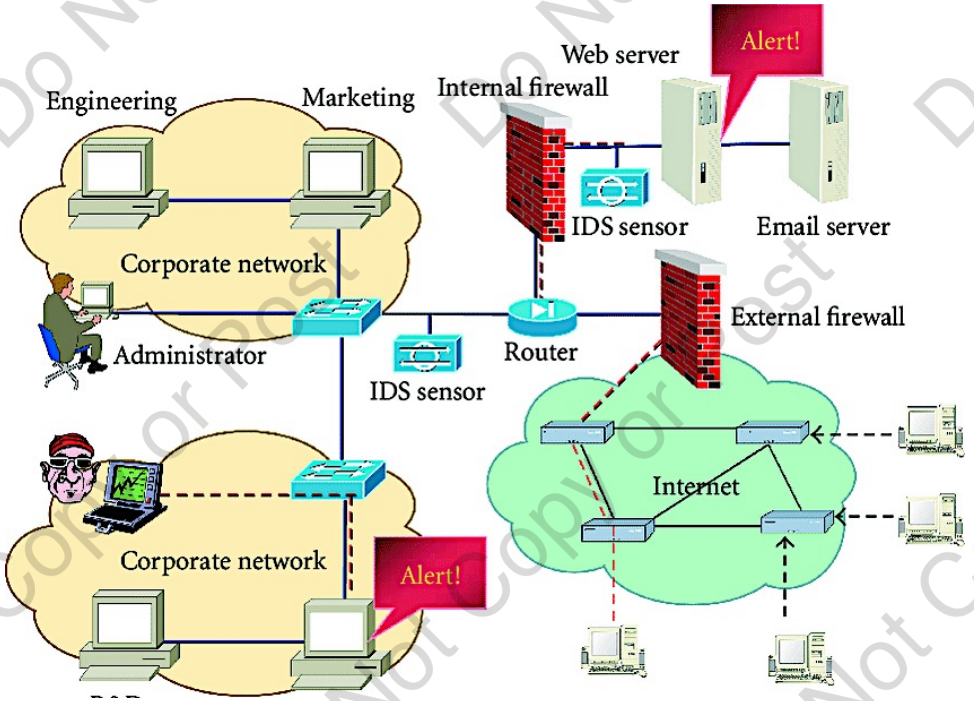
In graph data

HEC PARIS

High-Impact Application I

In information security (cybersecurity)

Intrusion detection systems are tools used to monitor network traffic and evaluate the components of the traffic to detect threats to the network.



HEC PARIS

High-Impact Application II

In medical & healthcare

The scarcity of available data chiefly determines an intricate scenario even for experts and specialized clinicians, which in turn leads to the so-called “diagnostic odyssey” for the patient. This situation calls for innovative solutions to support the decision process *via* quantitative and automated tools.

Affecting fewer than one in 2,000 people, a rare disease can include rare forms of neurological diseases, metabolic diseases, intellectual disabilities, certain cancers, rheumatologic disorders, complex epilepsies, immune deficiencies and autoinflammatory diseases. © Gorodenkoff

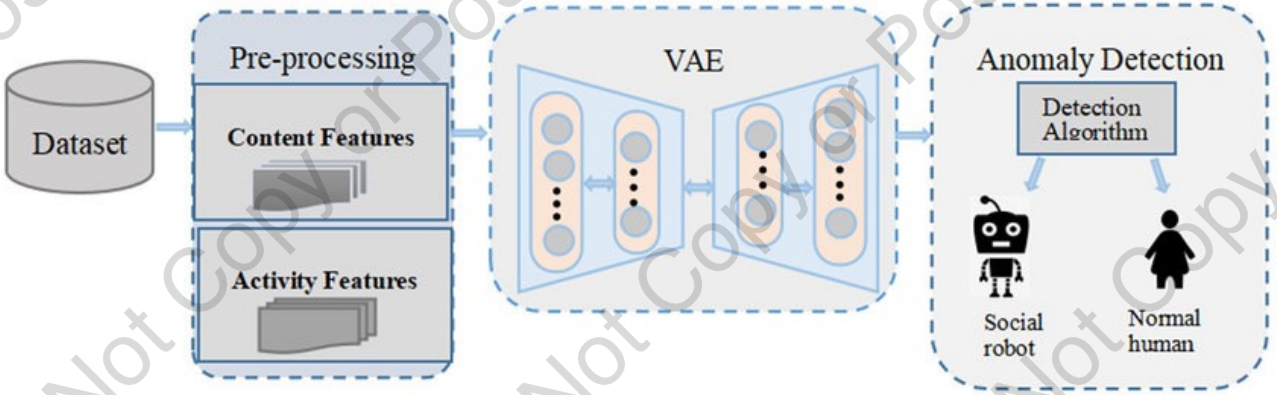
Decherchi, S., Pedrini, E., Mordenti, M., Cavalli, A. and Sangiorgi, L., 2021. Opportunities and challenges for machine learning in rare diseases. *Frontiers in Medicine*, *8*.

HEC PARIS

High-Impact Application III

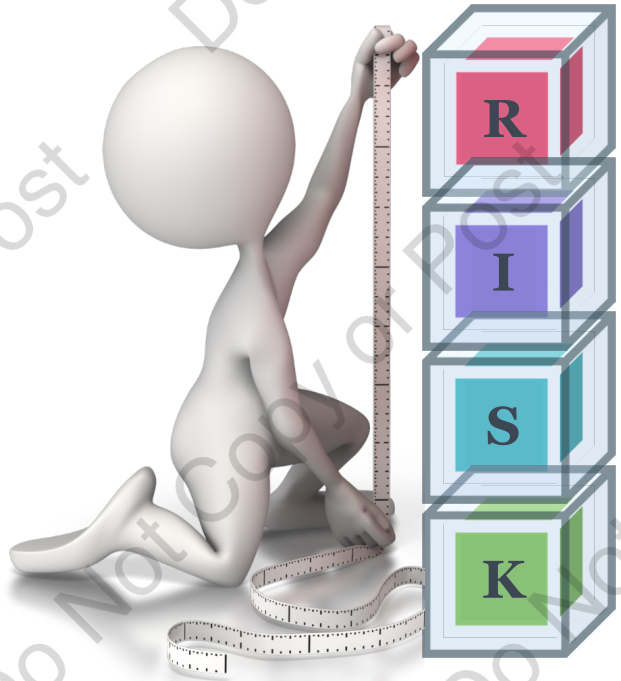
In social media monitoring

Social media bots (automated accounts) attacks are organized crimes that pose potential threats to public opinion, democracy, public health, the stock market and other disciplines.



Wang, X., Zheng, Q., Zheng, K., Sui, Y., Cao, S. and Shi, Y., 2021. Detecting social media bots with variational autoencoder and k-nearest neighbor. *Applied Sciences*, *11*(12), p.5482.

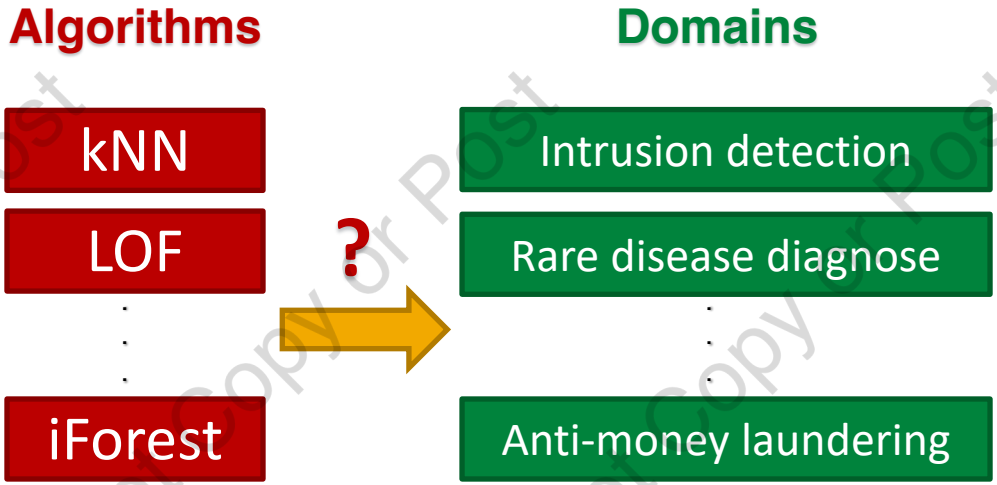
High-Impact Application in Finance and Risk Management



HEC PARIS

There is NO SINGLE BEST Anomaly Detection Algorithm!

No free lunch theorem: there is no universal learning algorithm that performs well on all problems. Simply put, **no single anomaly detection algorithm can always outperform.**



Even for the same application domain, different natures of problems or different properties of data call for different algorithms.

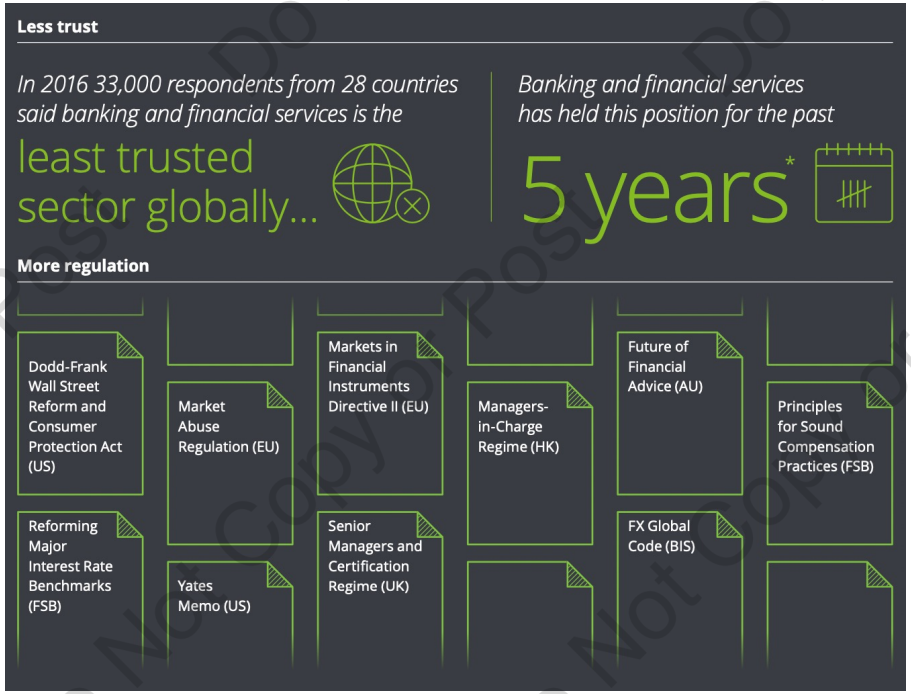
➔ **Understanding the nature of the problem and the properties of data is critical in designing the algorithms**

Source: Yue Zhao's presentation on Systems, Algorithms, and model selection

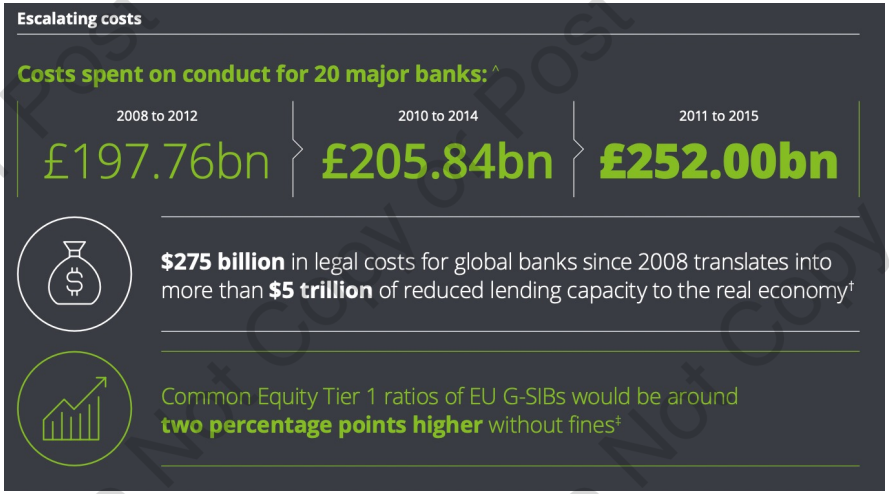
Corporate Misconduct Risk

HEC PARIS

Corporate Misconduct Risk
The Trust Issue



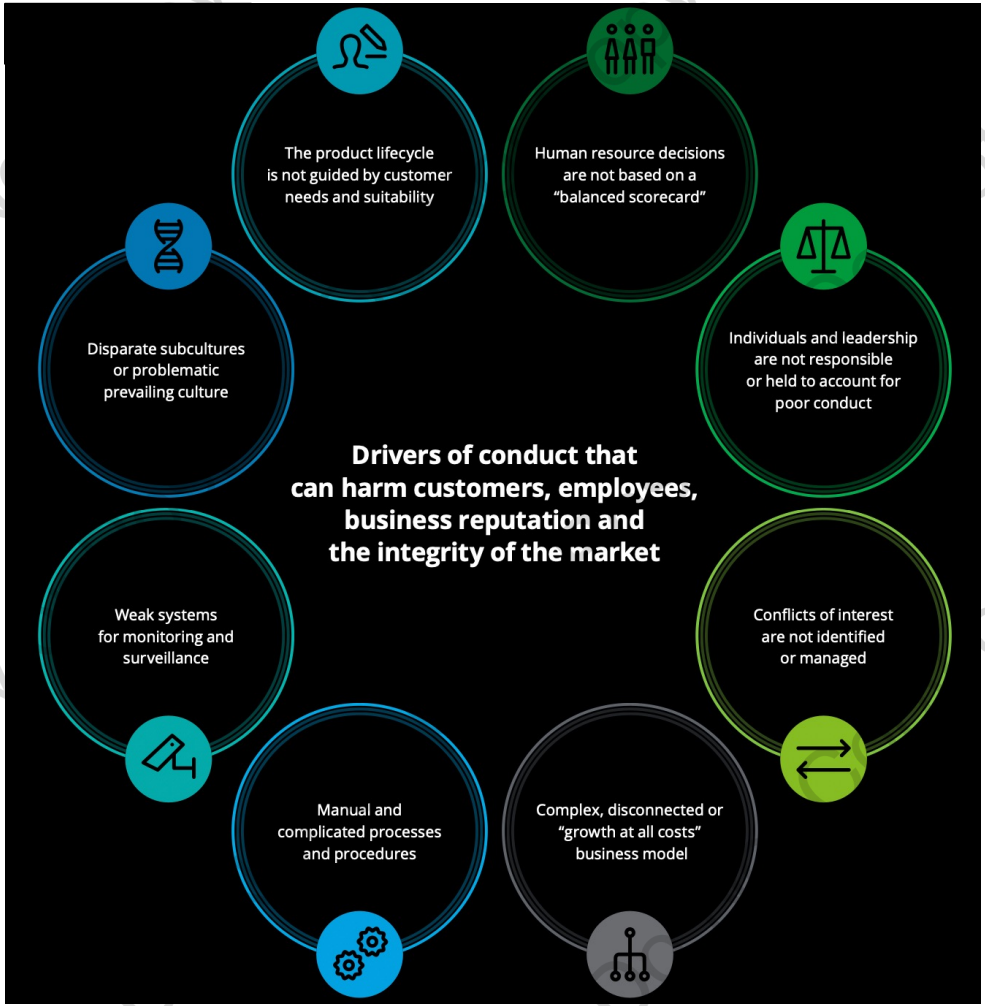
There has been no shortage of well-publicized and highly damaging misconduct scandals within the financial services industry over the past decade.



Source: Deloitte Center for Regulatory Strategy

HEC PARIS

Root Causes



Source: Deloitte Center for Regulatory Strategy

Responses

Source: Deloitte Center for Regulatory Strategy

Problems

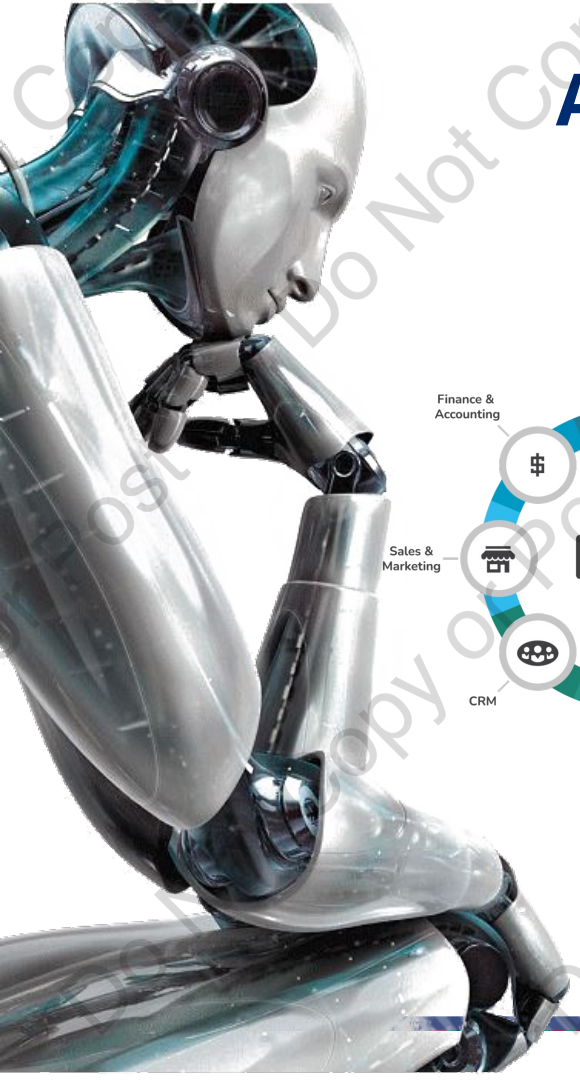
Identification and Assessment

Solutions

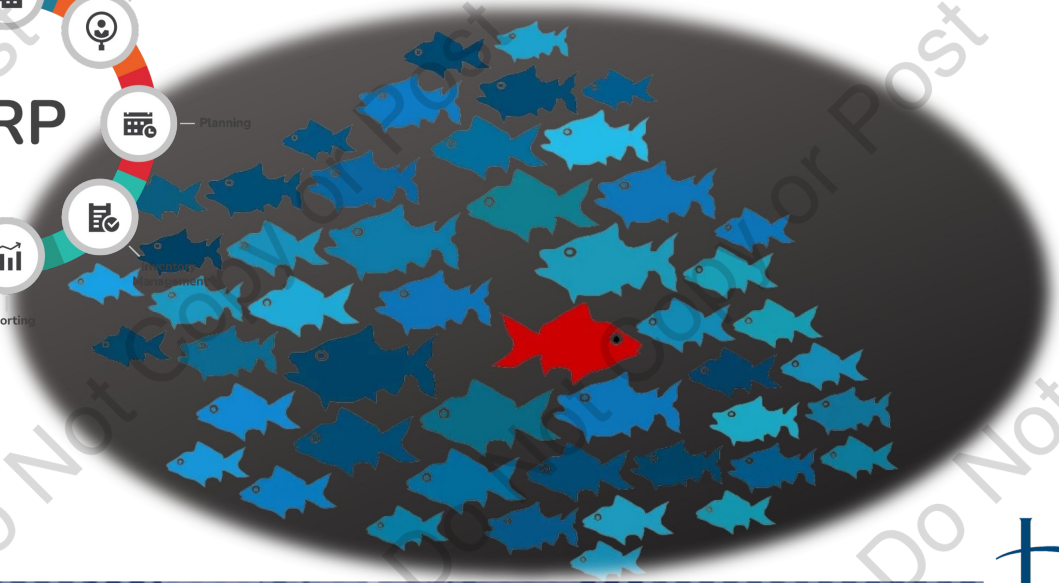
HEC PARIS

Innovative Solutions

Anomaly Detection in Financial Transaction Data



General, scalable, and explainable model for detecting anomalous transaction

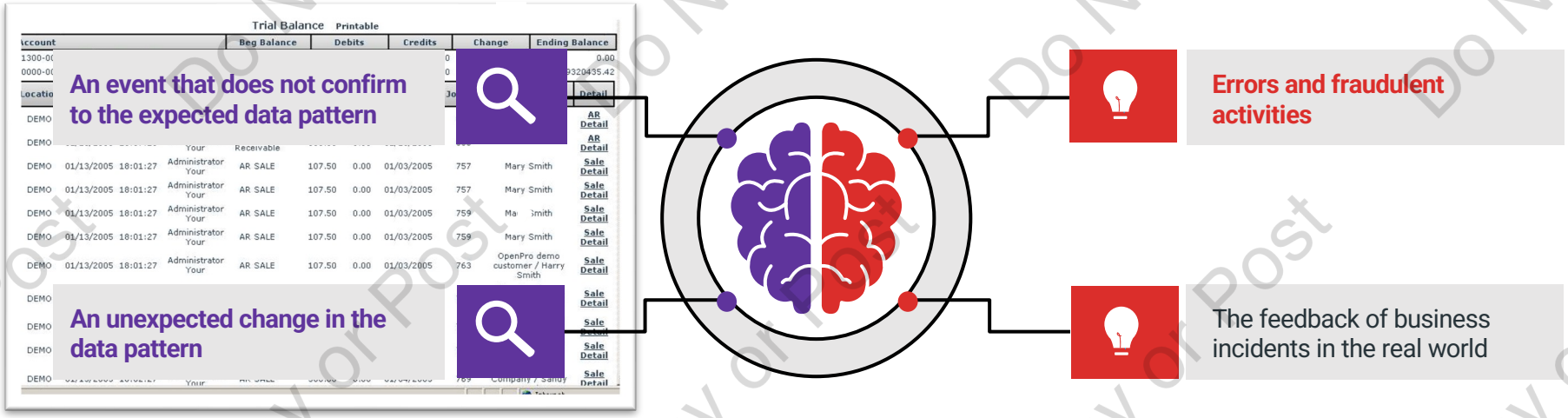


Anomalous Transaction Ranking

HEC PARIS

Background

Anomaly Detection in Corporate General Ledger



One of the most significant analytical procedures performed by auditors and risk management professionals is reading or scanning the general ledger account activity.

→ **Detect unusual transactions**

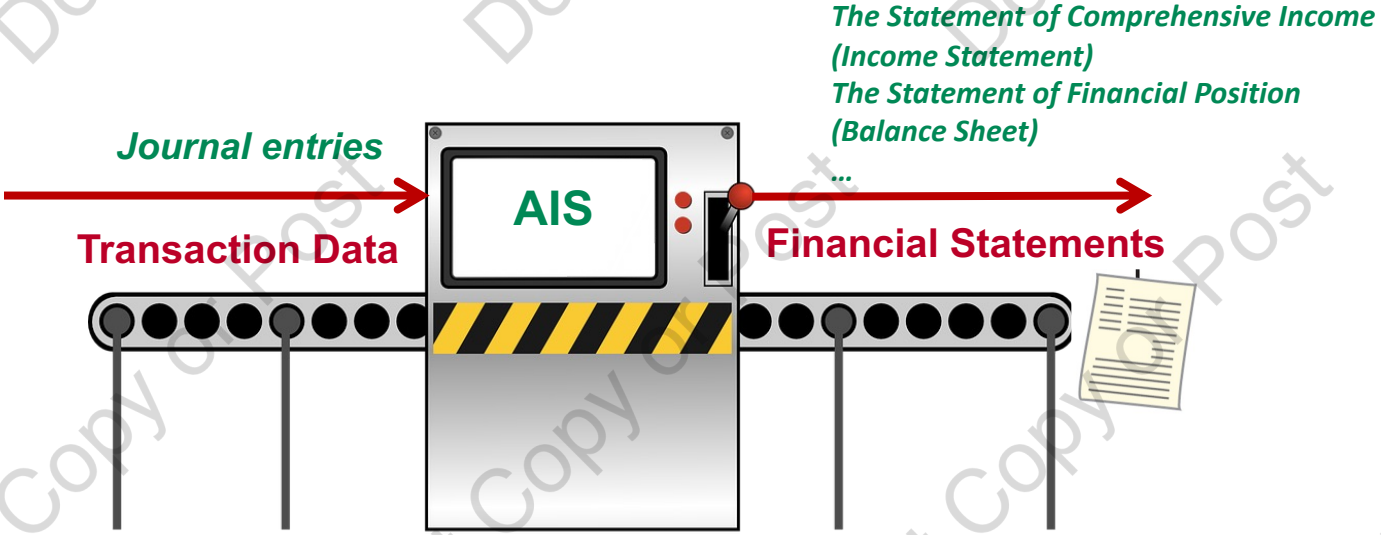
→ Obtain substantive evidence that relevant assertions for many account balances are reasonable

→ Assess the appropriateness and reasonableness of an entity's applicable reporting framework

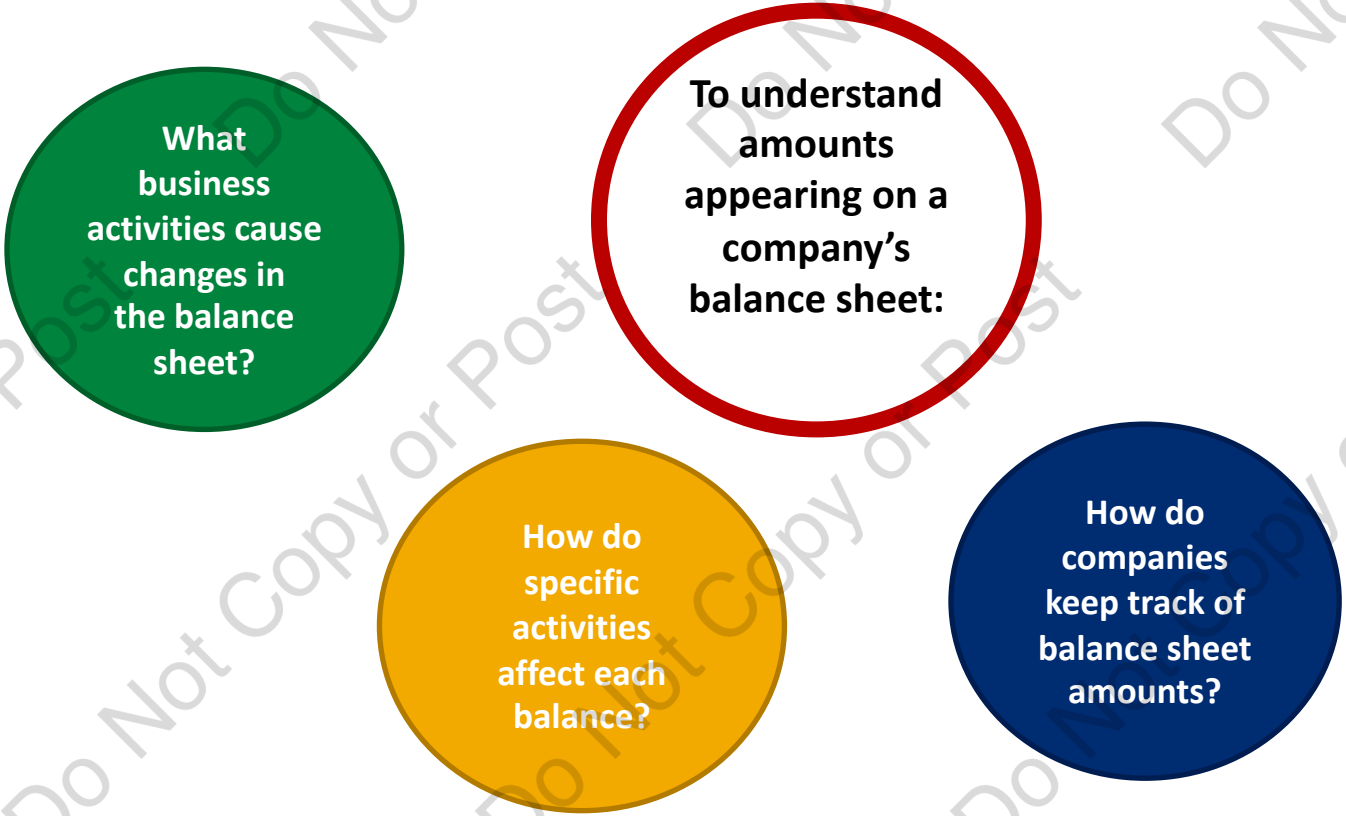
The Financial Transaction Data

HEC PARIS

The Accounting Information System



Understanding how a business measures performance



HEC PARIS

Elements of a balance sheet

$$A = L + SE$$

Assets

Economic resources with probable future benefits owned or controlled by the entity.

Liabilities

Debts or obligations (claims to a company's resources) that result from a company's past transactions and will be paid with assets or services. Entities that a company owes money to are called creditors.

Stockholders' Equity

The financing provided by the owners and business operations.

Example - Chipotle Mexican Grill, Inc. Balance Sheet

**The information has been adapted from actual statements and simplified for this chapter.*

CHIPOTLE MEXICAN GRILL, INC.

Consolidated Balance Sheet*
December 31, 2014
(in thousands of dollars, except per share data)

ASSETS

Current Assets:		
Cash		419,500
Short-term investments		338,600
Accounts receivable		34,800
Supplies		15,300
Prepaid expenses	70,300	
Total current assets		878,500
Property and equipment:		
Land		11,100
Buildings	1,267,100	
Equipment	442,500	
Total cost		1,720,700
Accumulated depreciation		(613,700)
Net property and equipment		1,107,000
Long-term investments	496,100	
Intangible assets	64,700	
Total assets		$2,546,300

Current assets

Noncurrent assets

EXPLANATIONS

"Consolidated" means all subsidiaries are combined
Point in time for which the balance sheet was prepared

Ownership of other companies' stocks and bonds
Amounts due from customers and others
Food, beverage, and packaging supplies on hand
Rent, advertising, and insurance paid in advance

Includes furniture and fixtures
Cost of property and equipment at date of acquisition
Amount of cost used in past operations

Ownership of other companies' stocks and bonds
Rights, such as patents, trademarks, and licenses

HEC PARIS

Example - Chipotle Mexican Grill, Inc. Balance Sheet

**The information has been adapted from actual statements and simplified for this chapter.*

CHIPOTLE MEXICAN GRILL, INC.

Consolidated Balance Sheet*
December 31, 2014
(in thousands of dollars, except per share data)

LIABILITIES AND STOCKHOLDERS' EQUITY

Current liabilities

Current Liabilities:

Accounts payable	$	69,600
Unearned revenue		16,800
Accrued expenses payable:		
Wages payable		73,900
Utilities payable		85,400
Total current liabilities		245,700

Noncurrent liabilities

Other liabilities	288,200
Total liabilities	533,900

Stockholders' equity

Stockholders' Equity:

Common stock ($0.01 par value)	400
Additional paid-in capital	290,200
Retained earnings	1,721,800
Total stockholders' equity	2,012,400
Total liabilities and stockholders' equity	$2,546,300

EXPLANATIONS

"Consolidated" means all subsidiaries are combined
Point in time for which the balance sheet was prepared

Amount due to suppliers
Unredeemed gift cards

Amount due to employees
Amount due for electric, gas, and telephone usage

Summary of liabilities due beyond one year

Total par value of stock issued by company to investors
Excess of amount received from investors over par

Undistributed earnings reinvested in the company

HEC PARIS

What Business Activities Cause Changes in the Financial Statement Amounts?

External Events: Exchanges between the entity and one or more parties.

Ex: Purchase of a machine from a supplier.

Internal Events: Events that are not exchanges between parties but that have a direct and measurable effect on the entity.

Ex: Using up insurance paid in advance.

HEC PARIS

Understanding how accounts play a role in capturing amounts spent/received in a transaction

Accounts are used to accumulate the dollar effect of transactions

HEC PARIS

Typical Account Titles

Assets	Liabilities	Stockholder's Equity	Revenues	Expenses
Cash	Accounts Payable	Common Stock	Sales Revenue	Cost of Goods Sold
Short-Term Investments	Accrued Expenses Payable	Additional Paid-in Capital	Fee Revenue	Wages Expense
Accounts Receivable	Notes Payable	Retained Earnings	Interest Revenue	Rent Expense
Notes Receivable	Taxes Payable		Rent Revenue	Interest Expense
Inventory (to be sold)	Unearned Revenue		Service Revenue	Depreciation Expense
Supplies	Bonds Payable			Advertising Expense
Prepaid Expenses				Insurance Expense
Long-Term Investments				Repair Expense
Equipment				Income Tax Expense
Buildings				
Land				
Intangibles				

Principles of Transaction Analysis

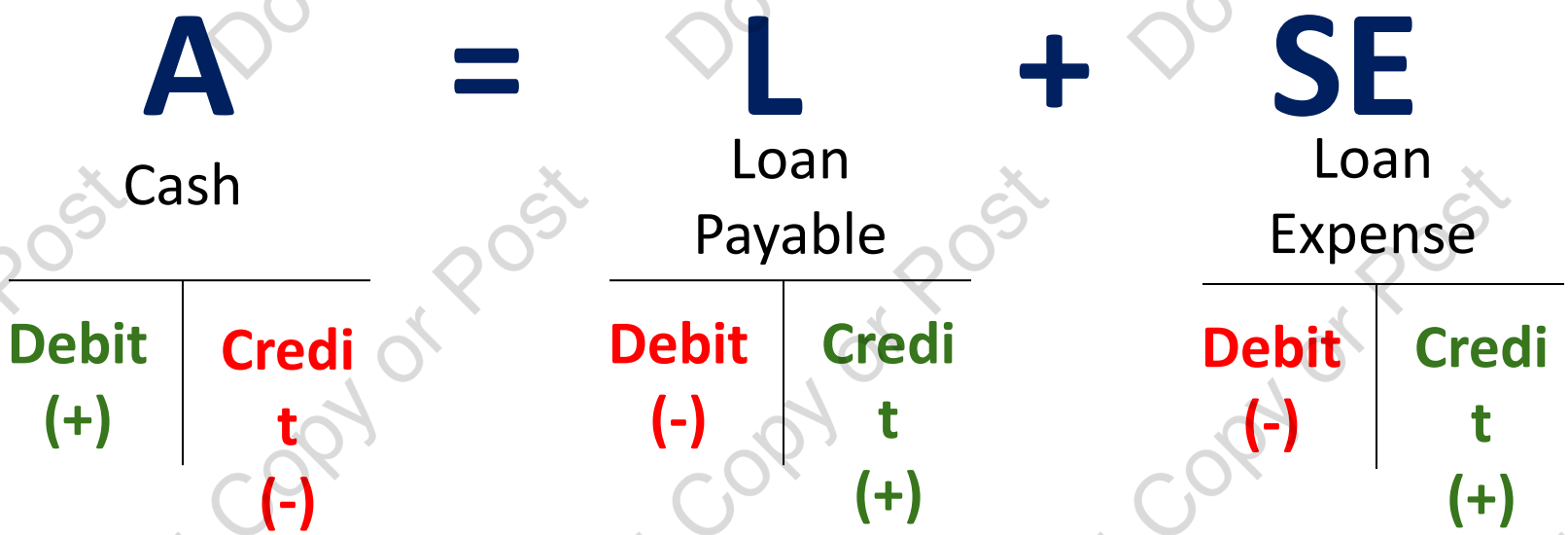
1. Double-entry - Every transaction affects at least two accounts (duality of effects)

2. Balance of equation - The accounting equation must remain in balance after each transaction.

A = L + SE

Assets | Liabilities | Stockholders' Equity

Using T-accounts to balance equations



Example - Visualizing transactions

Scenario: Hi! Paris started operations on 1/1/2021. On that date, Hi! Paris borrowed $2,000 from its local bank, signing a note for the principal to be paid in three years.

Assets	=	Liabilities	+	Stockholders' Equity

Cash

(Debit)	
(a) $2,000	

Loan Payable

	(Credit)
	(a) $2,000

		Debit	Credit
(a)	Cash (+A)	2,000	
	Loan Payable (+L)		2000

HEC PARIS

Understanding the operating cycle in the transaction

Objective

A General, Scalable, and Explainable Anomaly Detection Model

Dataset	A	B	C	D
Time Period	2016 GL 1-12			
# Transactions (Journal Entry ID)	39,092	90,628	157,002	9,941
Valid Lines (Rows)	762,402	336,162	1,433,649	896,028
Valid Variables (Columns)	26	18	26	31
Financial Flows ($Billion)	2.94	36.90	81.39	76.9
True anomalies	15	18	26	21

Fast and Reliable Detection

Whether the detection model can catch true anomalies using very few resources

Explainable AI

Whether audit professionals would be able to tell a story based on detection results

HEC PARIS

Meta-Data

Demo GL Journal Entries from Dataset C

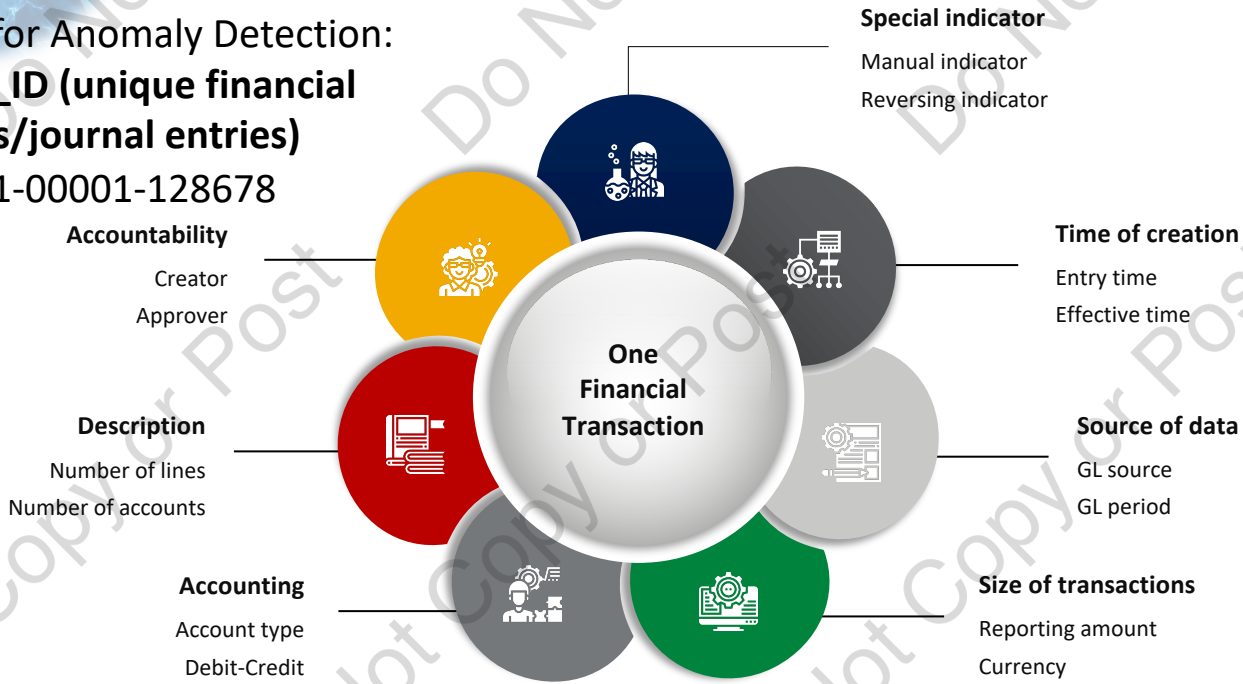
Two random journal entries from Dataset C

GL_Business_Unit_Code	GL_Account_Number	GL_Journal_ID	GL_Journal_Header_Description	GL_Entry_Date	GL_Effective_Date	GL_Fiscal_Year	GL_Period	GL_User_ID	GL_Segment01	GL_Segment02	GL_Segment03	GL_Amount_Cr_Dr_Indicator	GL_Reporting_Amount
00001	1.1235.3	2016/01-00001-128678	HOSTESS BRANDS, LLC	20150925	20160119	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20150925	GLDOC: 54434	D	250.00
00001	1.2010.1	2016/01-00001-128678	Offset By Batch V 128678	20150925	20160119	2016	01	SAVARESE	GLDCT: AE	GLDICJ: 20150925	GLDOC: 128678	C	(250.00)
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 60954	D	32.00
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 60955	D	172.80
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61234	D	2,300.00
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61236	D	212.00
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61335	D	2,300.00
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61336	D	1,100.00
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61337	D	1,100.00
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61340	D	500.00
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160111	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61341	D	1,071.62
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61361	D	1,100.00
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61387	D	1,186.13
00001	1.1235.3	2016/01-00001-136794	KRAFT FOODS INC.	20151202	20160104	2016	01	BILLYR	GLDCT: P4	GLDICJ: 20151202	GLDOC: 61413	D	687.70
00001	1.2010.1	2016/01-00001-136794	Offset By Batch V 136794	20151202	20160104	2016	01	SAVARESE	GLDCT: AE	GLDICJ: 20151202	GLDOC: 136794	C	(204.80)
00001	1.2010.1	2016/01-00001-136794	Offset By Batch V 136794	20151202	20160111	2016	01	SAVARESE	GLDCT: AE	GLDICJ: 20151202	GLDOC: 136794	C	(1,071.62)
00001	1.2010.1	2016/01-00001-136794	Offset By Batch V 136794	20151202	20160104	2016	01	SAVARESE	GLDCT: AE	GLDICJ: 20151202	GLDOC: 136794	C	(10,485.83)

HEC PARIS

Meta-Data

Datapoints for Anomaly Detection:
GL_Journal_ID (unique financial transactions/journal entries)
e.g. 2016/01-00001-128678



Challenge and Solution

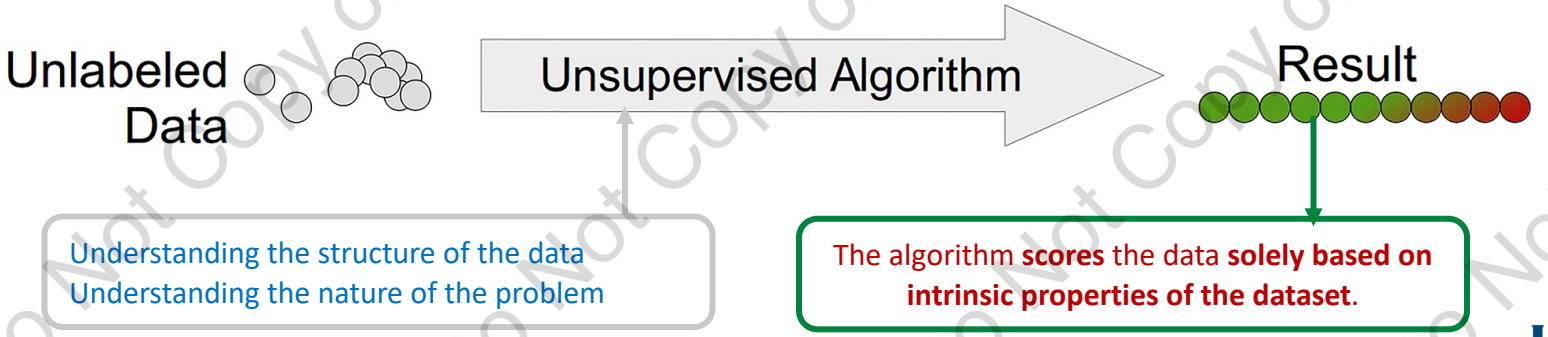
HEC PARIS

Solution: Supervised vs Unsupervised?

Major limitations of supervised learning:

- Assumption: anomalies are known and labeled correctly.
  - Reality: anomalies are often not known in advance and may occur spontaneously
- Imbalanced data: most of the journal entries are faithful records of normal financial transactions, and only a tiny portion of them are considered anomalies that interest industry professionals.

Here's the setup of our unsupervised anomaly detection.



Algorithmic Framework

The Data Science Problem Description

Given

A multi-dimensional database with categorical features

Goals:

→ Build a model for the **"norm"**

→ Find **anomalous** (rare) tuples

Features

Tuples

Financial Transaction	$$$ Amount	Account Type	Creator-Approver	Entry-Effective Time
1	a	b	c	x
2	a	b	c	x
3	a	b	c	x
4	a	b	d	y
5	a	b	d	x
6	a	b	d	y
7	a	b	d	z

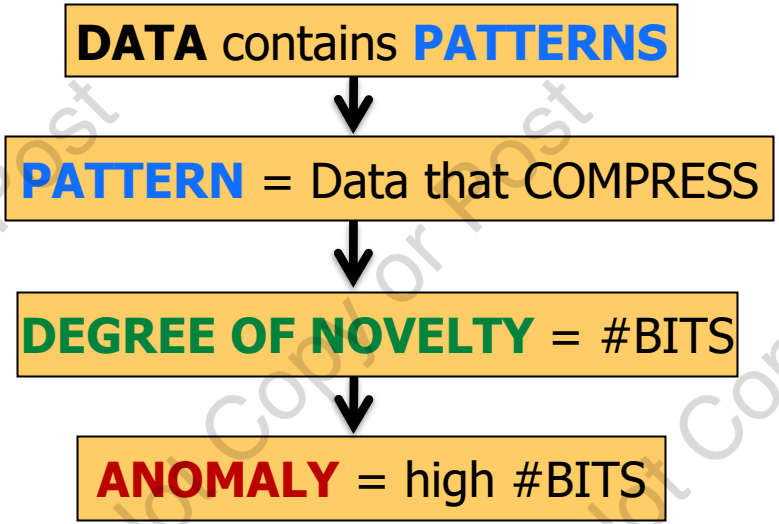
- A database D is a bag of n tuples over a set of m categorical features $\mathcal{F} = \{f_1, \dots, f_m\}$.
- Each feature $f \in \mathcal{F}$ has a domain $dom(f)$ of possible values $\{v_1, v_2, \dots\}$.
- An item is a feature-value pair $(f = v)$, with $f \subseteq F$, and $v \in dom(f)$
- An **itemset** (a pattern) is then a pair $(F = v)$, for a set of features $F \subseteq \mathcal{F}$, and $v \in dom(F)$ is a vector of length $|F|$.
- A tuple t is said to contain a **pattern** $(F = v)$, denoted as $p(F = v) \subseteq t$, if all features $f \in F, t_f = v_f$ holds.

HEC PARIS

Algorithmic Framework

The Information Theory-Based Approach

- The **more often** a pattern occurs in the database, the **shorter** its code lengths.
- Some groups of features may be highly **correlated**, and hence may be compressed together as feature groups.



Dictionary based compression

Data encoding

An illustrative database D and an example code table CT for a set of three features, $F = \{f_1, f_2, f_3\}$

Data	*Code Table*			
$f_1 f_2 f_3$	$p(F = v)$	$code(p)$	$usage(p)$	$L(code(p))$
a b x	a b x	0	4	1 bit
a b x	a c	10	2	2 bits
a b x	x	110	1	3 bits
a b x	y	111	1	3 bits
a c x				
a c y				

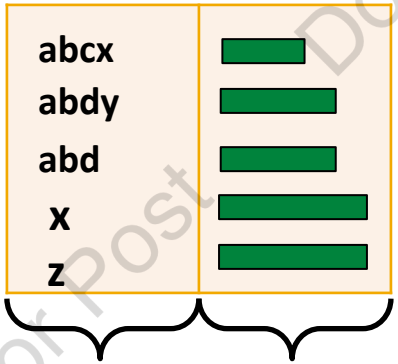
Algorithmic Framework

The Information Theory-Based Approach

Example database

Financial Transaction	$$$ Amount	Account Type	Creator-Approver	Entry-Effective Time
1	a	b	c	x
2	a	b	c	x
3	a	b	c	x
4	a	b	d	y
5	a	b	d	x
6	a	b	d	y
7	a	b	d	z

Code Table



Freq: Pr(p)

3/9
2/9
2/9
1/9
1/9

Optimal #bits (p) = $-\log_2 Pr(p)$
(theoretic prefix code length)

Frequent itemsets **imply** Good Compression

Algorithmic Framework

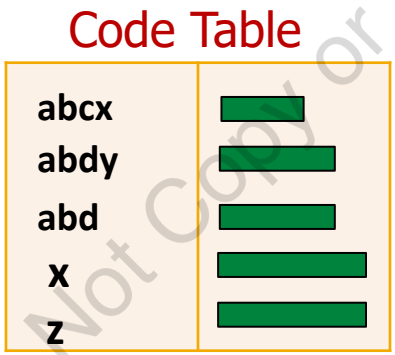
The Information Theory-Based Approach

1. How many code tables?
2. Which patterns to include in the code tables?

Main idea: Minimum Description Length Principle

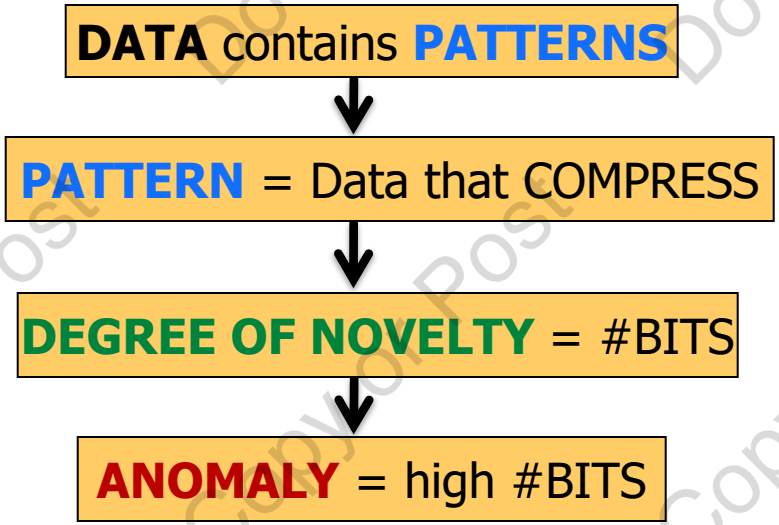
$$L(CTs) + L(D|CTs)$$

- $L(CTs)$: encoding length of code tables — Model complexity
- $L(D|CTs)$: encoding length of data with patterns — Data complexity



HEC PARIS

A Detour to Minimum Description Length (MDL)



- The **minimum description length (MDL) principle** in machine learning says that the **best description of the data is given by the model which compresses it the best.**
- Put another way, learning a model for the data or predicting it is about capturing the regularities in the data and any regularity in the data can be used to compress it.
- Thus, the more we can compress data, the more we have learnt about it and the better we can predict it.

HEC PARIS

Objective formulation

Given a **database D** with m features in F

Find

a **partitioning P**: $\{F_1, \ldots, F_k\}$ of F

a set of associated **code tables CT**: $\{CT_1, \ldots, CT_k\}$

such that total MDL-encoding cost **L(P, CT, D)** is minimized

Financial Transaction	$$$ Amount	Account Type	Creator-Approver	Entry effective time
1	a	b	c	x
2	a	b	c	x
3	a	b	c	x
4	a	b	d	y
5	a	b	d	x
6	a	b	d	y
7	a	b	d	z

Code table #1

3	abc
4	abd

Code table #2

4	x
2	y
1	z

Objective formulation

Given a **database D** with m features in F
Find
a **partitioning P**: $\{F_1, \ldots, F_k\}$ of F
a set of associated **code tables CT**: $\{CT_1, \ldots, CT_k\}$
such that total encoding cost is minimized

$$L(\mathcal{P}, \mathcal{CT}, D) = \underbrace{L(\mathcal{P})}_{\text{length in bits: P}} + \underbrace{\sum_{F \in \mathcal{P}} L(\pi_F(D) | CT_F)}_{\text{length in bits: data, encoded by CT}} + \underbrace{\sum_{F \in \mathcal{P}} L(CT_F)}_{\text{length in bits: code tables CT}}$$

$$\sum_{k=0}^{2^{|F|}} \binom{2^{|F|}}{k} \times k!$$

A Detour to Minimum Description Length (MDL)

Ockham chooses a razor

- MDL is connected to **Occam's Razor**, stating that **"other things being equal, a simpler explanation is better than a more complex one."**
- In MDL, the simplicity of a model is interpreted as **the length of the code** obtained when that model is used to **compress** the data.
- The **ideal version** of MDL is given by the Kolmogorov Complexity, which is defined as the length of the shortest computer program that prints the sequence of observed data and halts.
- **Uncomputable!!!** (It can be shown that there exists no computer program that, for every set of data D, when given D as input, returns the shortest program that prints D). Moreover, for finite length sequences, the best program may depend on the sequence itself.

HEC PARIS

The Practical (Crude) MDL

The crude MDL is a two-stage code approach based on the notion that we can specify the descriptive properties of a model for data in two stages:

1. encode the **model** with some codelength $L(\boldsymbol{q})$
2. encode the **data** using the **model** with codelength $\boldsymbol{L_q}(\boldsymbol{x^n})$. Now pick the model which minimizes the total codelength of the two-stage code:

$$q_\gamma(\boldsymbol{x^n}) = \arg\min_{\boldsymbol{q}\in\boldsymbol{Q_\gamma}}\{L(\boldsymbol{q}) + L(\boldsymbol{x^n})\}$$

HEC PARIS

Algorithmic Framework 1

The Information Theory-Based Approach

Given a **database D** with m features in F

→ Find

a **partitioning P**: $\{F_1, \ldots, F_k\}$ of F

a set of associated **code tables CT**: $\{CT_1, \ldots, CT_k\}$

such that total MDL-encoding cost $\boldsymbol{L(P, CT, D)}$ is minimized

1	a	b	c	x
2	a	b	c	x
3	a	b	c	x
4	a	b	d	y
5	a	b	d	x
6	a	b	d	y
7	a	b	d	z

Code table #1

3	abc	
4	abd	

Code table #2

4	x	
2	y	
1	z	

HEC PARIS

Algorithmic Framework 1

The Information Theory-Based Approach

1	a	b	c	x
2	a	b	c	x
3	a	b	c	x
4	a	b	d	y
5	a	b	d	x
6	a	b	d	y
7	a	b	d	z

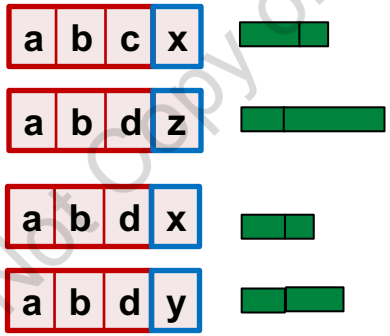
Code table #1

3	abc	
4	abd	

Code table #2

4	x	
2	y	
1	z	

Compression cost:



high cost!

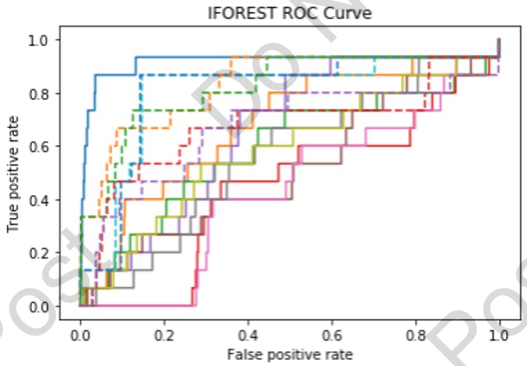
- top-k
- statistical test

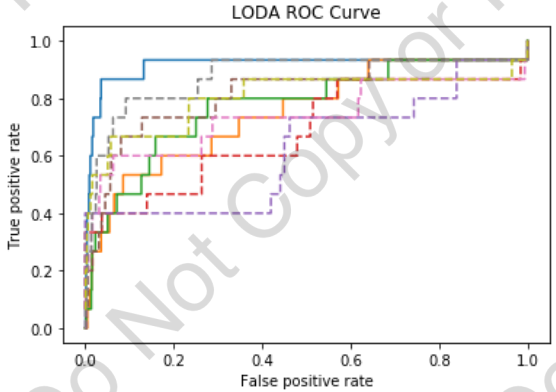
Anomalous!!!

HEC PARIS

Algorithmic Framework 1 (Blue ROC Curve)

Anomaly Detection Performance (Dataset C)





AUC: 0.9633

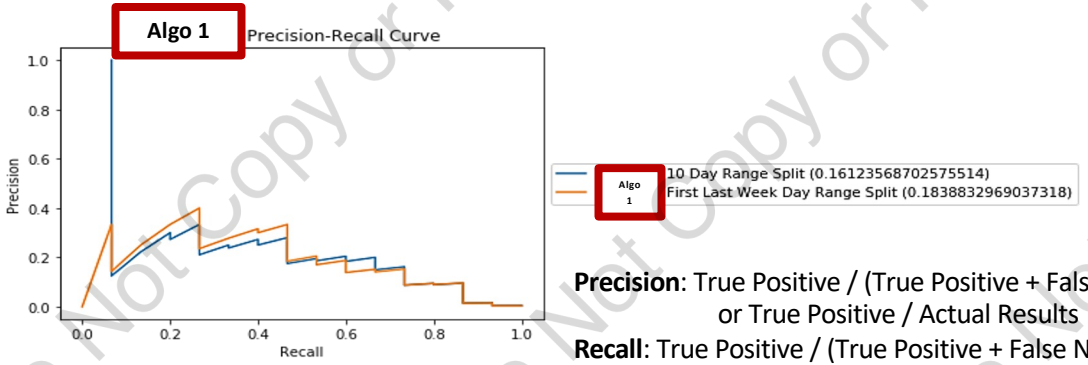
26/157,002 ≅ 0.017%

Algo 1 Ranking	# True Anomalies	
Top 5	3 of 26	
Top 10	3 of 26	
Top 20	3 of 26	
Top 50	7 of 26	
Top 100	8 of 26	
Top 200	16 of 26	
Top 1000	22 of 26	0.6%
Top 10000	23 of 26	6%

HEC PARIS

Why Does Algo 1 Perform So Well on Accounting Data?

- **Information theory-based approach**
  - Complex multi-dimensional database, unbalanced data
- **Parameter-free**
  - Builds dictionaries directly from data
  - No specified parameters such as distance function or similarity measures
- **Flexible in feature engineering**



Precision: True Positive / (True Positive + False Positive)
or True Positive / Actual Results

Recall: True Positive / (True Positive + False Negative)
or True Positive / Predicted Results (# anomalies)

Algorithmic Framework 1

Highly correlated features are grouped together (Dataset C)

1	Other				
2	GL_Reversal_Indicator				
3	GL_Reporting_Amount				
4	GL_Entry_Minute	GL_Entry_Hour	GL_Entry_Total		
5	GL_Entry_IsWeekday	GL_Effective_IsWeekday	GL_Effective_Entry_Diff	GL_Entry_Weekday	GL_Effective_Weekday
6	GL_Manual_Indicator	GL_User_ID	GL_Approver_ID		
7	Assets	Liabilities	Expenses	Num_Accounts	GL_Source
8	GL_Entry_Month	GL_Effective_Month	GL_Period	GL_Entry_Day_Range	GL_Effective_Day_Range
9	Num_Lines	Revenue			

HEC PARIS

Algorithmic Framework 1

Designing an explainable AI (Top 1 anomaly in Dataset C)

Feature Group	1	2	3	4	
Feature	Other	GL_Reversal_Indicator	GL_Reporting_Amount	GL_Entry_Hour	GL_Entry_Total
Value	1	1	1,424	17	1064
P| Code Length	99	99	17	84	

Feature Group	5	6				
Feature	GL_Entry_Minute	GL_Entry_Is Weekday	GL_Effective_Is Weekday	GL_Effective_Entry_ Diff	GL_Entry_Wee kday	GL_Effective_We ekday
Value	44	True	True	0	2	2
P| Code Length	98	7				

HEC PARIS

Algorithmic Framework 1

Designing an explainable AI (Top 1 anomaly in Dataset C)

Feature Group	7			8				
Feature	GL_Manual_Indicator	GL_User_ID	GL_Approver_ID	Assets	Liabilities	Expenses	Num_Accounts	GL_Source
Value	1 (M)	2 (Ezra)	9 (Randy)	0	1	0	2	3 (G J6)
P| Code Length	99			99				

Feature Group	9			10		11	
Feature	GL_Entry_Month	GL_Effective_Month	GL_Period	GL_Entry_Day_Range	GL_Effective_Day_Range	Num_Lines	Revenue
Value	10	10	2	2	2	2	0
P| Code Length	99			97		0	

HEC PARIS

Why Graph?

Algorithmic Framework 2

Graph-based technique for more adversarially-robust detection tools

Representation of financial transactions:

- ✓ The Journal Entry
- ✓ The Linear Algebra
- ✓ The Graph

GL_Account_Number	GL_Reporting_Amount	CA_FS_Caption	…
40060000 (Revenue)	-1575 (Credit)	Gross Sales (GSL)	…
10415000 (Assets)	1575 (Debit)	Accounts Receivable (ARV)	…

$$A \cdot y = x = \begin{bmatrix} -1575 \\ 1575 \end{bmatrix}$$

Gross Sales

Accounts Receivable

Gross Sales

Accounts Receivable

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix} = A$$

$$[1575] = y$$

Ijiri (1975)
Arya et al. (2000)
Arya et al. (2004)

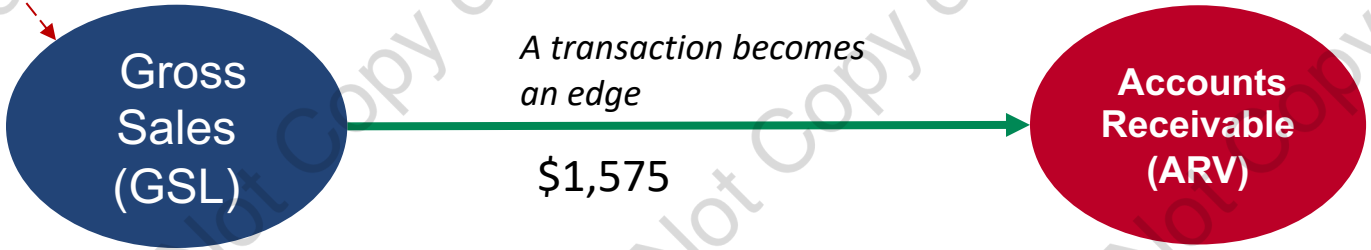
HEC PARIS

Graph Data

Every simple journal entry can be represented as an edge between two nodes

- Account → Node
- Entry → Edge
- Debit → Inflow
- Credit → Outflow

GL_Account_Number	GL_Reporting_Amount	CA_FS_Caption	…
40060000 (Revenue)	-1575 (Credit)	Gross Sales (GSL)	…
10415000 (Assets)	1575 (Debit)	Accounts Receivable (ARV)	…

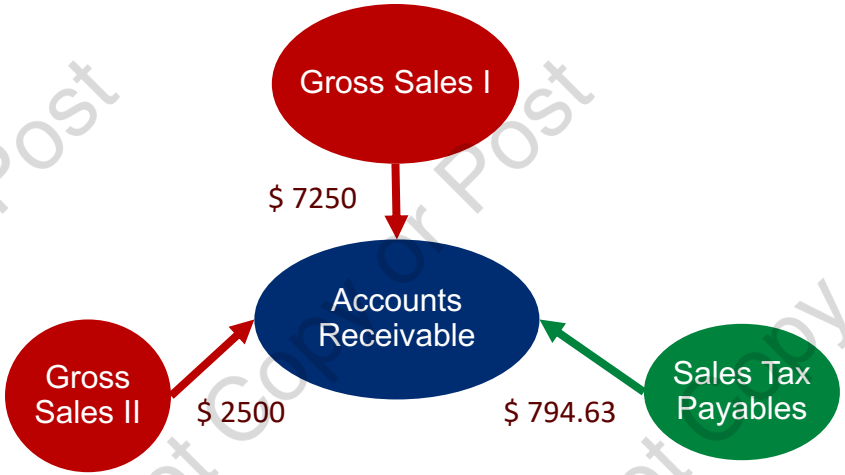


Graph Data

- One debit and many credits
- One credit and many debits
- Total amount offset

GL_Account _Number	CA_FS_Caption	Cr/Dr	GL_Reporting _Amount
40020000 (Revenue)	Gross Sales (GSL)	C	-7250
40020000 (Revenue)	Gross Sales (GSL)	C	-2500
20830000 (Liabilities)	Sales Tax Payables (STP)	C	-794.63
10390000 (Assets)	Accounts Receivable (ARV)	D	10544.63

Source/Sink Stars

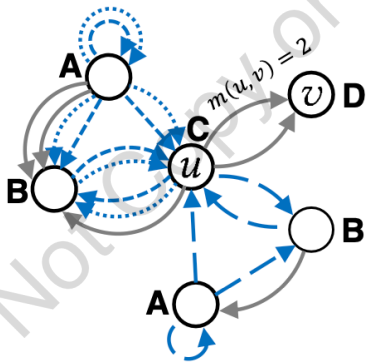


The Data Science Problem (The Graph Version)

A large set of J graphs $G = \{G_1, \ldots G_J\}$ is given. Each graph $G_j = (V_j, E_j, \tau)$ is directed, node-labeled, multi-graph which may contain multiple edges that have the same end nodes.

$\tau: V_j \rightarrow \Gamma$ is a function that assigns label from an alphabet Γ to nodes in each graph.

The number of realization of an edge $(u, v) \in E_j$ is called its multiplicity, denoted $m(u, v)$.



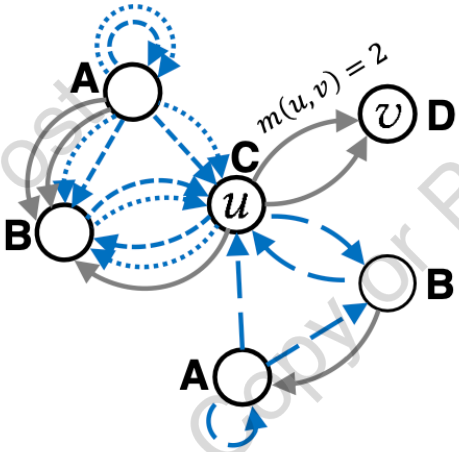
To find anomalous graphs in datasets G

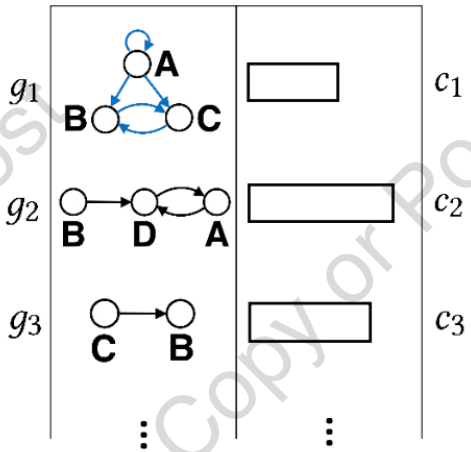
→ To identify key characteristics patterns of the data that "explain" or compress the data well

→ To flag those graphs that do not exhibit such patterns as expected.

HEC PARIS

Graphs vs Motifs





DATA contains **PATTERNS**

↓

PATTERN = Data that COMPRESS

↓

DEGREE OF NOVELTY = #BITS

↓

ANOMALY = high #BITS

HEC PARIS

Algorithmic Framework

MDL Principle!!!

1. How many motif tables?
2. Which patterns to include in the motif tables?

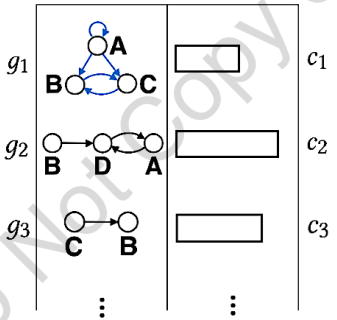
Main idea: Minimum Description Length Principle

$$\underbrace{L(MTs)}_{\text{encoding length of Motif tables}} + \underbrace{L(G|MTs)}_{\text{encoding length of data with patterns}}$$

Model complexity

Data complexity

Motif Table



HEC PARIS

Algorithm 1. Motif Encoding

Input: Motif $g_i = (V_i, E_i)$
Output: Encoding of g_i ▶ Note: the values after symbol ▷ summed over the course of the algorithm provides the total encoding length $L(g_i)$

1: Encode $n_i = |V_i|$; # of nodes in g_i ▷ $L_{\mathbb{N}}(n_i)^2$
2: **repeat**
3: Pick an unmarked node $v \in V_i$ at random where $indeg(v) = 0$ (if none, pick any unmarked node), and mark v
4: **procedure** RecurseNode($mid(v)$)
5: Encode $mid(v)$; v's motif-node-ID ▷ $\log_2(n_i)$
6: Encode v's node label ▷ $\log_2(T)$
7: Encode # of v's out-neighbors $\mathcal{N}_{\text{out}}(v)$ ▷ $L_{\mathbb{N}}(\text{outdeg}(v))$
8: Encode motif-node-IDs of $\mathcal{N}_{\text{out}}(v)$ ▷ $\log_2 \binom{n_i}{\text{outdeg}(v)}$
9: **for each** unmarked node $u \in \mathcal{N}_{\text{out}}(v)$ **do**
10: Mark u
11: RecurseNode($mid(u)$)
12: **until** all nodes in V_i are marked

→ Encode the motif table
→ Encode a graph given the motif table

HEC PARIS

Algorithm 2. Graph Encoding

Input: Graph G_j, Motif table MT, $\mathcal{CS}(G_j, MT)$
Output: Encoding of G_j such that it can be decoded losslessly

1: **do**
2: Pick any occurrence $g_{kj} \in \mathcal{CS}(G_j, MT)$ and communicate $code_{MT}(g_k)$ ▷ $|code_{MT}(g_k)| = L(code_k)$ Eq. (2)
3: Encode matching graph-node-IDs ▷ $L_{\text{perm}}(|V_j|, n_k)$ Eq. (6)
4: Encode g_k's multiplicity on V_{kj} ▷ $L_{\mathbb{N}}(m(g_k, G_j, V_{kj}))$
5: $\mathcal{CS}(G_j, MT) \leftarrow \mathcal{CS}(G_j, MT) \setminus \{g_{lj} \mid V_{lj} = V_{kj}\}$
6: **while** $\mathcal{CS}(G_j, MT) \neq \emptyset$

HEC PARIS

Algorithm 3. Search (Greedy Algorithm for MIS)

To compress as a large portion of the input graphs as possible using motifs

→ To find a large set of non-overlapping motif occurrences that cover these graphs

→ An instance of the **Maximum Independent Set (MIS)** problem on the occurrence graph G_o.

- → In G_o, the nodes represent motif occurrences and edges connect two occurrence that shares a common edge.
- → Maximum Independent Set ensures that occurrences in the solution are non-overlapping (thanks to independence, no two are incident to the same edge).
- → Moreover, MIS helps us identify motifs that have large usages, i.e. number of non-overlapping occurrences, which is associated with shorter code length and hence better compression.

Input: Motif occurrence graph $G_O = (V_O, E_O)$ for graph G_j
Output: Set $O \subset V_O$ of independent nodes (non-overlapping occurrences)

1: Solution set $O = \emptyset$
2: **while** $V_O \neq \emptyset$ **do**
3: $v_{\min} = \arg\min_{v \in V_O} deg_{G_O}(v)$
4: $O := O \cup \{v_{\min}\}$
5: Remove $v_{\min}$ and $\mathcal{N}(v_{\min})$ along with incident edges from V_O and E_O accordingly

HEC PARIS

Algorithm 4. Memory-efficient Greedy MIS

Input: Simple occurrences $sg_{1j}, \ldots, sg_{tj}$, edge multiplicities in G_j
Output: A list $\mathcal{S}$ of simple occurrences

$\mathcal{S} = \emptyset$
Calculate degrees of all simple occurrences $deg_{G_O}(sg_{ij}), \forall i = 1 \ldots t$ using Eq. (15)
$deg_{\max} = \max_{i=1\ldots t} deg_{G_O}(sg_{ij})$
while $\exists i \in \{1 \ldots t\}$ s.t. $deg_{G_O}(sg_{ij}) < deg_{\max} + 1$ **do**
    $i^* = \arg\min_{i \in \{1\ldots t\}} deg_{G_O}(sg_{ij})$
    $\mathcal{S} = \mathcal{S} \cup \{sg_{i^*j}\}$
    $m((u,v)) = m((u,v)) - 1\ ,\ \forall (u,v) \in E_{i^*j}$
    for $l = 1 \ldots t, E_{lj} \cap E_{i^*j} \neq \emptyset$ **do**
        if $m((u,v)) > 0\ \forall (u,v) \in E_{lj}$ **then**
            Recalculate $deg_{G_O}(sg_{lj})$ using Eq. (15)
        else
            $deg_{G_O}(sg_{lj}) = deg_{\max} + 1$

HEC PARIS

Algorithm 5. Motif Table Search

Input: Database $\mathcal{G} = \{G_1, \ldots, G_J\}$, candidate motifs C, node labels $\mathcal{T}$, $(W)MIS$ solution $\mathcal{S}_j$ $\forall G_j \in \mathcal{G}$
Output: MT containing set of motifs $\mathcal{M}$

$\text{Cnt}_j(s,d) = 0$, for $j = 1 \ldots J$, and $\forall s, d \in \mathcal{T}$
for $G_j = (V_j, E_j) \in \mathcal{G}$ **do**
  for each $(u,v) \in E_j$ **do**
    $\text{Cnt}_j(t(u), t(v)) := \text{Cnt}_j(t(u), t(v)) + m(u,v)$
$\mathcal{M} = \{(s,d) | s, d \in \mathcal{T}$ and $\exists j \in [1, J]$ where $\text{Cnt}_j(s,d) > 0\}$
while $C \neq \emptyset$ **do**
  for $g = (V, E) \in C$ **do**
    for $G_j \in \mathcal{G}$ **do**
      $\mathcal{O}(g, \mathcal{CS}_j) = \{sg \simeq g | sg \in \mathcal{S}_j\}$
      $\text{Cnt}_{jg} := \text{Cnt}_j$
      for each $(u,v) \in E$ **do**
        $\text{Cnt}_{jg}(t(u), t(v)) := \text{Cnt}_{jg}(t(u), t(v)) - |\mathcal{O}(g, \mathcal{CS}_j)|$
    $\mathcal{M}_g = \mathcal{M} \cup \{g\} \backslash \{(s,d) | s, d \in \mathcal{T}, \text{Cnt}_{jg}(s,d) = 0 \; \forall j\}$
    Get $L(\mathcal{M}_g, \mathcal{G})$ using $\mathcal{O}(g_k, \mathcal{CS}_j)$ $\forall g_k \in \mathcal{M}_g$ s.t. $n_k \geq 3$ and $usage_{G_j}(s,d) := \text{Cnt}_{jg}(s,d)$ $\forall (s,d) \in \mathcal{M}_g; \forall j$
  $g_{\min} = \arg\min_{g \in C} L(\mathcal{M}_g, \mathcal{G})$
  if $L(\mathcal{M}_{g_{\min}}, \mathcal{G}) \leq L(\mathcal{M}, \mathcal{G})$ **then**
    $\mathcal{M} := \mathcal{M}_{g_{\min}}$, $C := C \backslash \{g_{\min}\}$, $\text{Cnt}_j := \text{Cnt}_{jg_{\min}}$
  else
    break

HEC PARIS

The Graph-Based Detection!

Open your Jupyter notebook and try the algorithms!

HEC PARIS

The Complementarity of the Two Algorithms

GL_Journal_ID	GL_Journal_ID_original	Anom_Type	Algo 1	Algo 2
20036-ACTUAL	Anom-local-9	random	4	24
20027-ACTUAL	Anom-manual-5	<Matt>	6	1450
20581-ACTUAL	Anom-local-6	random	7	1
20662-ACTUAL	Anom-local-0	random	12	185
20114-ACTUAL	Anom-local-5	random	24	3
20020-ACTUAL	Anom-local-4	random	30	21
20581-ACTUAL	Anom-local-7	random	37	33
20036-ACTUAL	Anom-global-2	unusual times long backdate	39	2095
20668-ACTUAL	Anom-local-8	random	41	8
20027-ACTUAL	Anom-local-2	random	42	119
20027-ACTUAL	Anom-global-0	high amnt period end	43	1477
20664-ACTUAL	Anom-local-3	random	74	41
20033-ACTUAL	Anom-local-1	random	89	3038
20667-ACTUAL	Anom-manual-1	<Matt>	90	61
20661-ACTUAL	Anom-manual-4	<Matt>	161	7
20520-ACTUAL	Anom-manual-3	<Matt>	333	9
20027-ACTUAL	Anom-global-5	new user	752	1480
20033-ACTUAL	Anom-manual-2	<Matt>	774	116
20033-ACTUAL	Anom-global-4	create approve	915	1696
20036-ACTUAL	Anom-global-1	new account combo	952	2
20036-ACTUAL	Anom-manual-6	<Chris>	2800	1117

- Recording financing arrangements as revenue
- Wrong tax code that is entered on a purchase transaction resulting in redundant Value Added Tax (VAT) payments
- A clerk that switches two bank account numbers causing an amount being transferred to a wrong supplier

HEC PARIS

Solving the Ensemble Learning Problem

Even the simplest fusion of the two algorithms we created show promise of a roust GL anomaly detection machinery, industry leaders call for ensemble learning and software engineering.

Algorithm 1

Feature engineering incorporating domain expertise

Algorithm 2

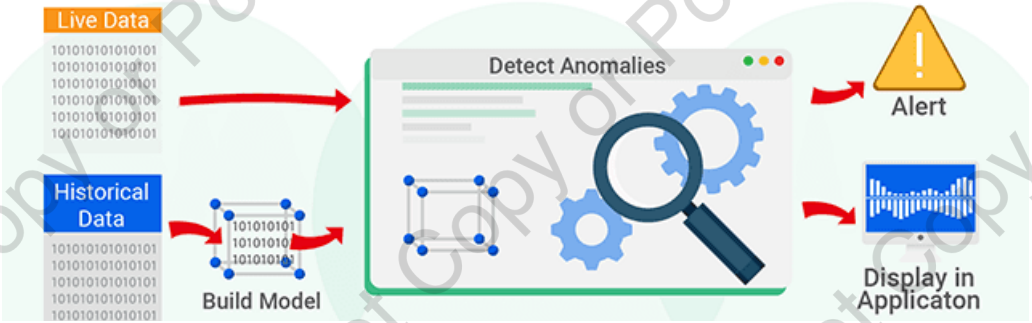
Graph-based analysis focusing on structural anomalies

Top10	19 journal entries	**<0.2%** of the financial transactions	**8/21** More than one-third of the anomalies!!!
Top100	191 journal entries	**<2%** of financial transactions	**17/21** More than 80% of the anomalies!!!

HEC PARIS

Integrate detection tools based on our algorithms into the internal audit and risk management space

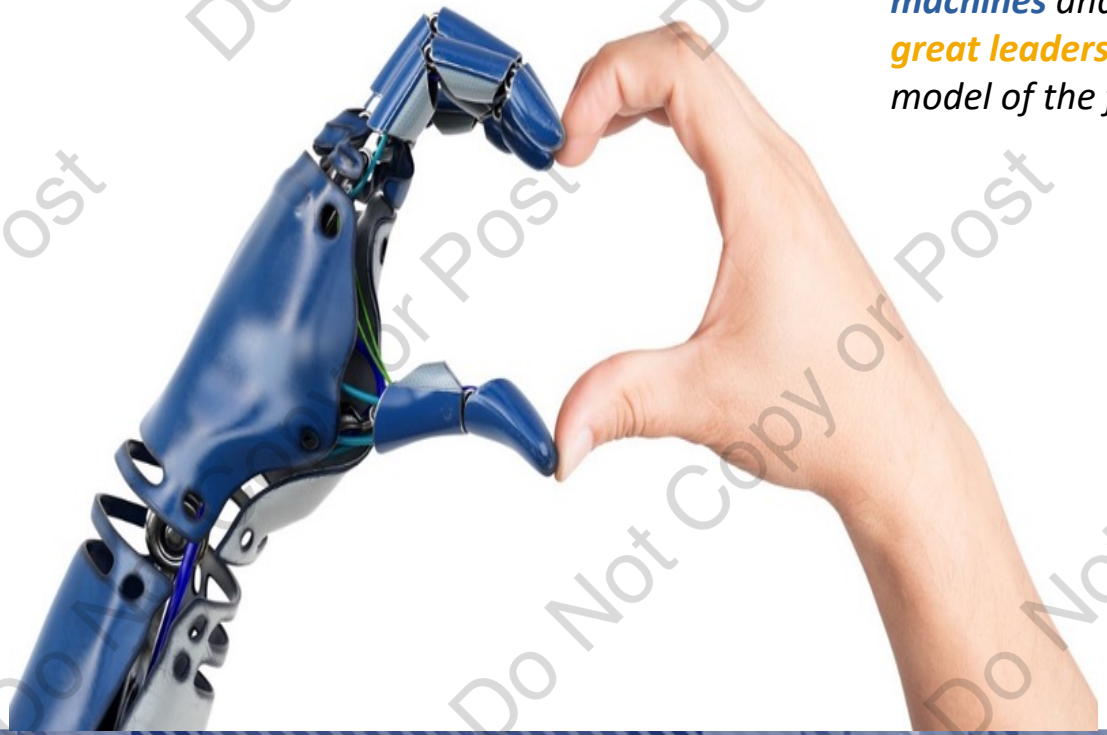
Continuous Monitoring and Continuous auditing—Building Risk Intelligent Businesses



- Honest mistakes and suspicious activities
- Process deviation and other business problems

Content credit: PwC

Thank you very much!



*"The most powerful weapon today is the alliance between the **mathematical smarts of machines** and **the imaginative human intellect of great leaders.** Together they make the business model of the future."*

-The Mathematical Corporation

Aluna Wang
Contact: *wanga@hec.fr*