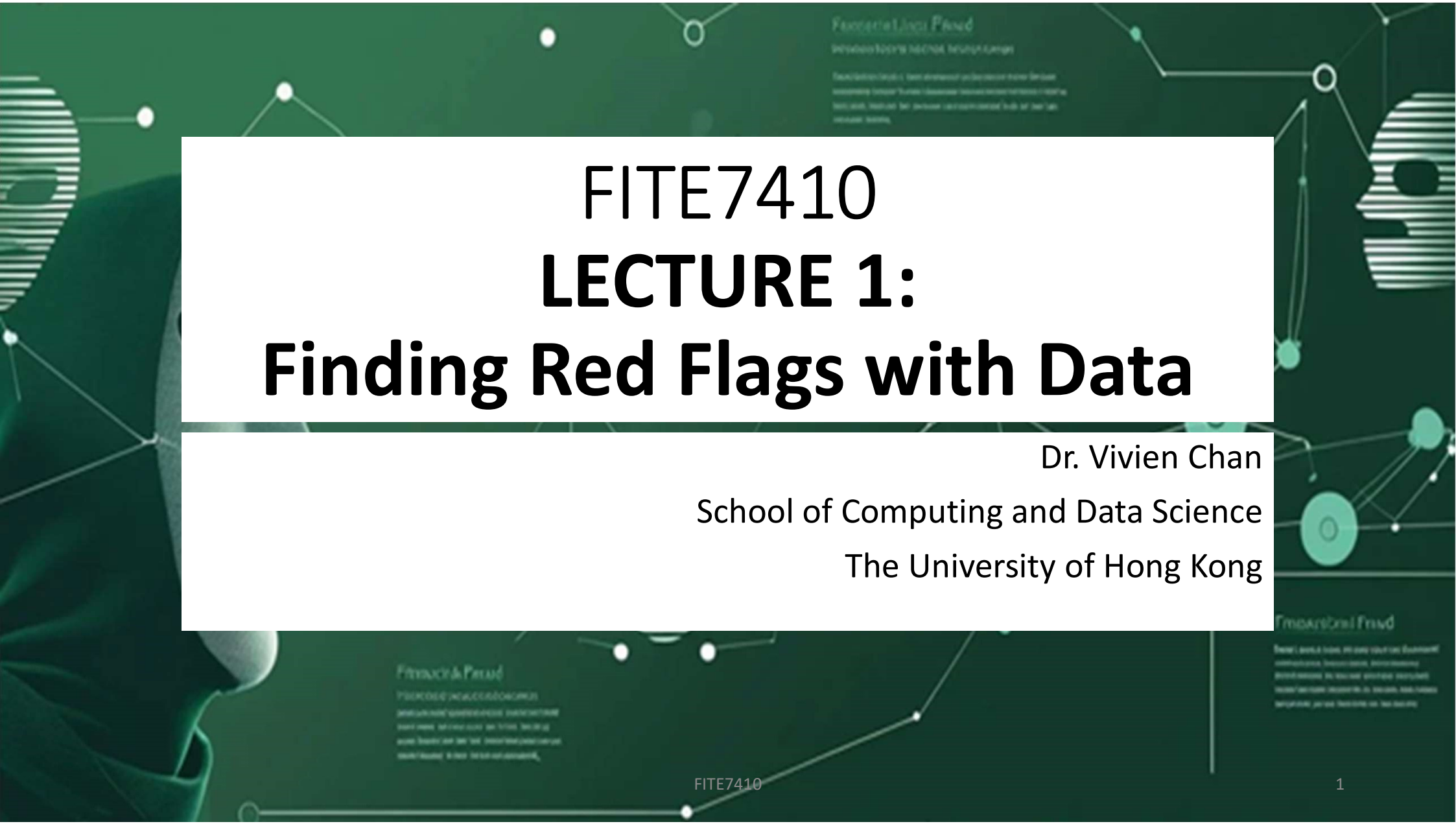


FITE7410
LECTURE 1:
Finding Red Flags with Data

Dr. Vivien Chan

School of Computing and Data Science

The University of Hong Kong

Agenda

Today's question: "**How do we use data to quickly surface suspicious activity in a sea of transactions?"**

Lecture goals:

1) Understand fraud basics and red flags
2) Learn EDA as an investigative process

Tutorial (50-60 mins)

- Introduction to using R


What is Financial Fraud Analytics?
FITE7410


What is “Financial Fraud”?
Any examples?
FITE7410

Definition of “Fraud”

English Dictionary definition

“the crime of getting money by deceiving people”

(Ref: “Cambridge Dictionary: fraud”. https://dictionary.cambridge.org/)

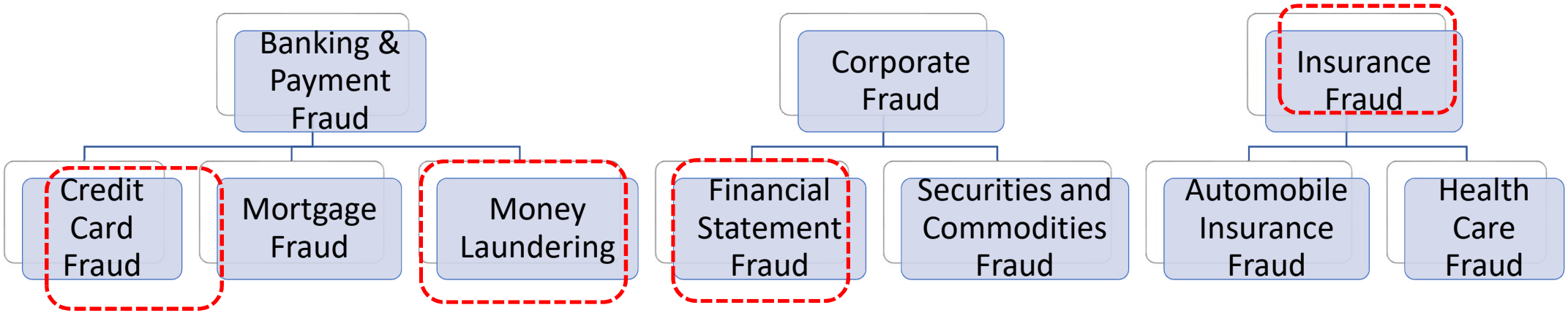
Legal Dictionary definition

“the intentional use of deceit, a trick or some dishonest means to deprive another of his/her/its money, property or a legal right. ”

(Ref: "Legal Dictionary: fraud". Law.com)

Types of Financial Frauds

- There is no single definition of financial fraud type.
- One of the categorization by FBI is as follows (selected from the list provided in the Federal Bureau of Investigation, Financial Crimes Report (2010–2011), United States):



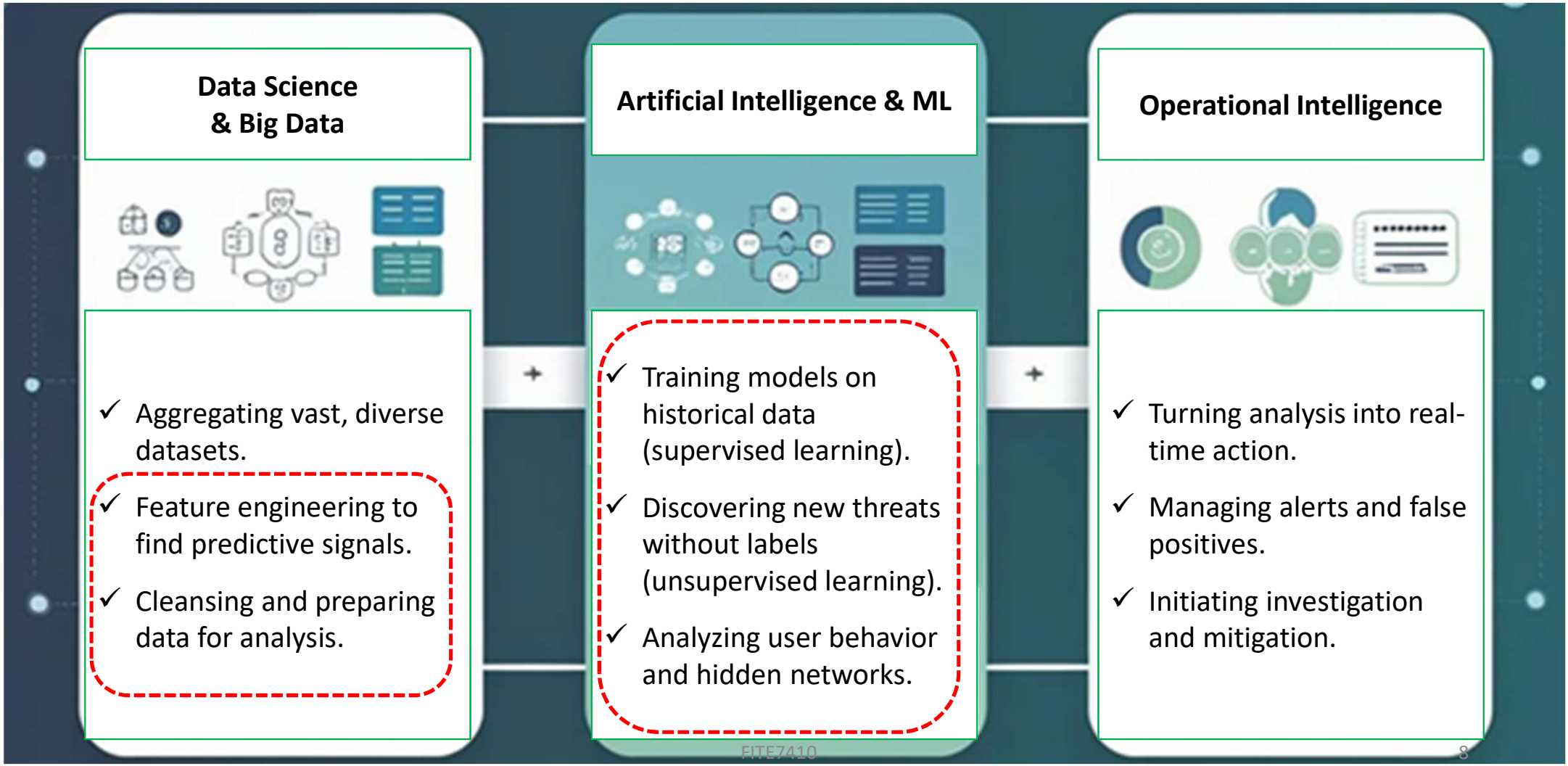
Fraud Analytics of these financial frauds to be discussed in upcoming lectures

What is Financial Fraud Analytics?

- In this course, Financial Fraud Analytics is:
  - The process of using machine learning/data mining techniques to analyse financial fraud data for red flags that correlate to financial frauds
  - The end product of fraud analytics is a sample of transactions that have a higher probability of containing one fraudulent transaction versus a random sample of transactions used to test control effectiveness
  - To answer the question:
    - Is there credible evidence that a fraud scenario is occurring?
    - Should we perform an investigation?

- In this course, Financial Fraud Analytics is NOT:
  - about forensic fraud investigation

What is Financial Fraud Analytics? A Modern Perspective



What is a Red Flag?

- A red flag is an observable condition that relates to the act of a specific fraud scenario
- Fraudsters may be detected by deviant behaviours from non-fraudsters. These deviations from normality are red flags of fraud, i.e. anomality or variation from predictable patterns of normal behaviours
- A red flag exists in:
  - Data
  - Documents
  - Internal controls
  - Behaviours
  - Public records
  - Digital footprints
  - … etc.

What are some examples of 'anomalities'?

Examples of Red Flags

- Examples of anomality:
  - Outliers
  - Inliers where they are not expected
  - Too many or too few transactions
  - Unexplained items
  - Unusual relationships between items
  - Unexpected timing of transactions or events
  - Inconsistencies
  - Gaps or duplicates of item numbers
  - Unexpected payment methods
  - Unreasonable items

Examples of Red Flags

Tax evasion fraud red flags:

- An identical financial statement, since fraudulent companies copy financial statements of nonfraudulent companies to look less suspicious
- Name of an accountant is unique, since this might concern a nonexisting accountant

Credit card fraud red flags:

- A small payment followed by a large payment immediately after, since a fraudster might first check whether the card is still active before placing a bet
- Regular rather small payments, which is a technique to avoid getting noticed

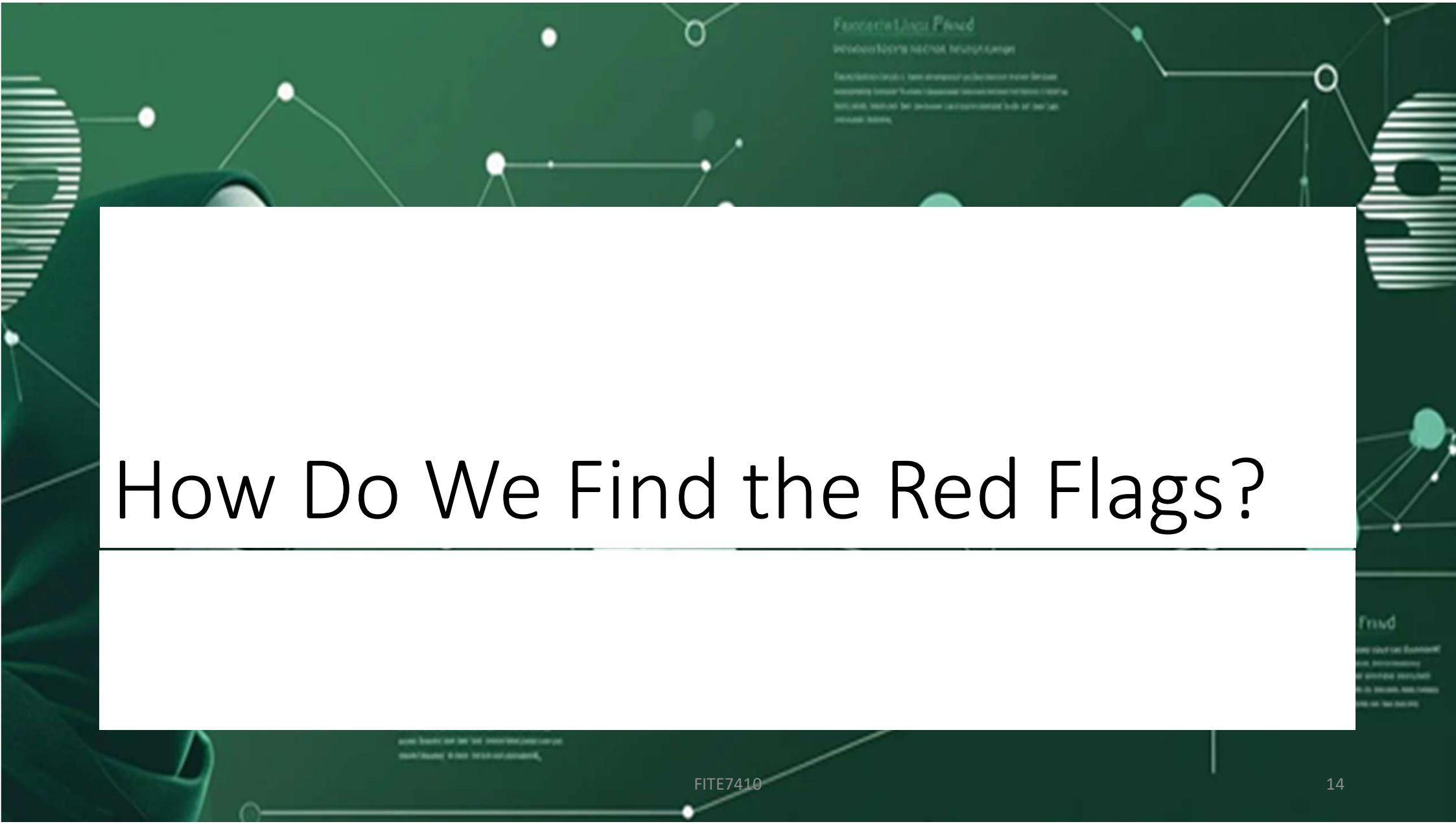
False Positive vs False Negative

	False Positive	***False Negative***
Definition	Red flags identified in the fraud data profile but the transaction is **NOT** a fraudulent transaction.	Red flags **NOT** identified in the fraud data profile but the transaction is a fraudulent transaction.
Effects	Too many false alarms might render the fraud analytics results not useful.	Missing out too many actual fraudulent transactions means the fraud data analytics is useless.
Reasons	Threshold of red flags too sensitive	• Not understanding well about the fraud scenarios • Data integrity issues, lack of data, etc.
Strategies to resolve	1. To reduce the number of false positives through fraud data analytics plan; 2. To allow user to resolve the false positives through manual procedures.	1. To review the fraud data analytics plan so as to have a better understanding of the fraud scenarios; 2. To review the issues related to data, from data collection, pre-processing, to data analysis.

The Investigator's Dilemma

If you are too aggressive, you accuse innocent people (False Positives).
If you are too cautious, you miss the criminals (False Negatives).

False Positives vs False Negatives – which one is more problematic?



How Do We Find the Red Flags?

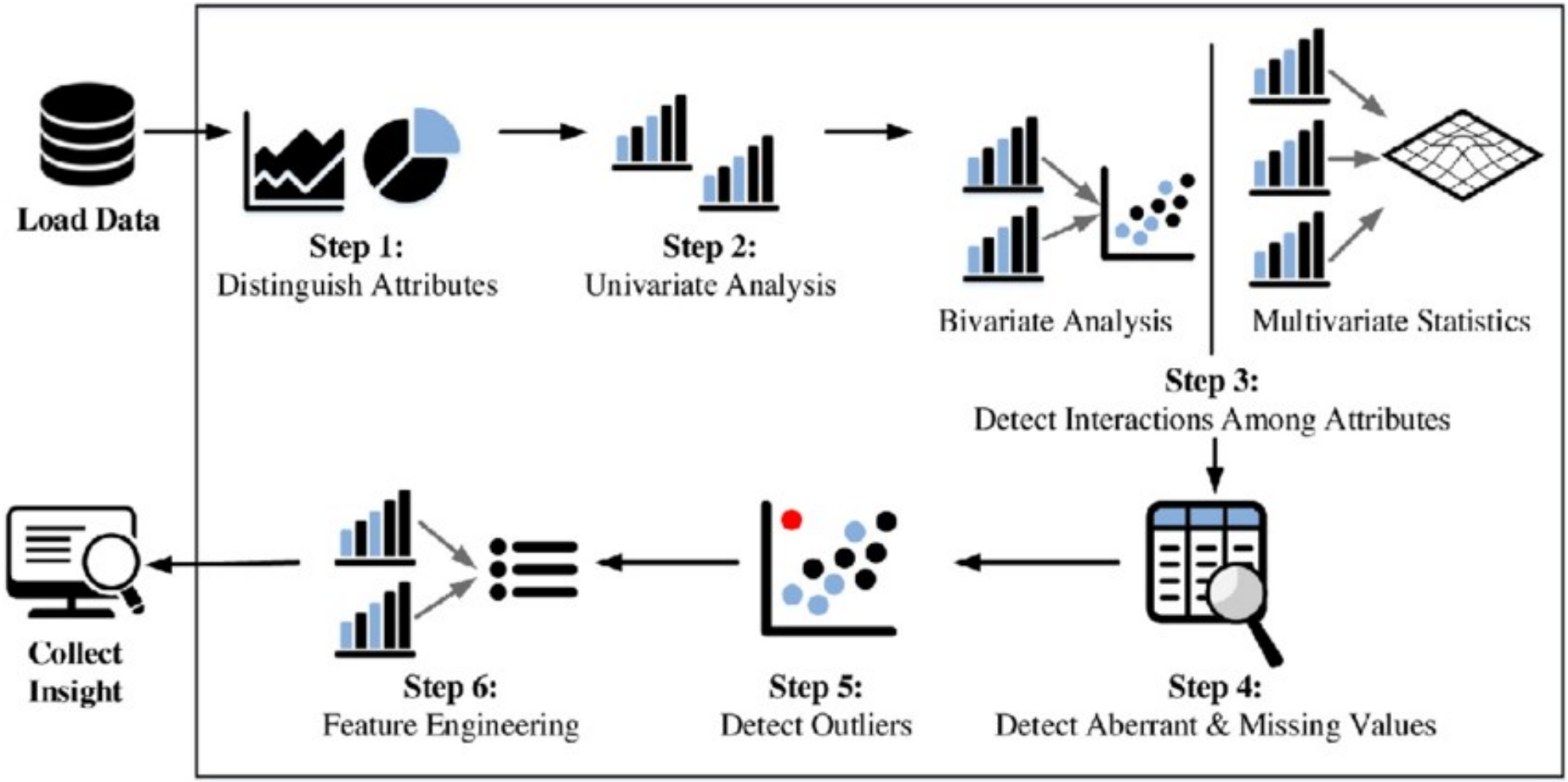
Exploratory Data Analysis (EDA)

- Exploratory Data Analysis (EDA, Tukey, 1977) is the process of using quick and simple methods for the visualization and examination of small data sets, e.g. boxplots, histograms, etc.

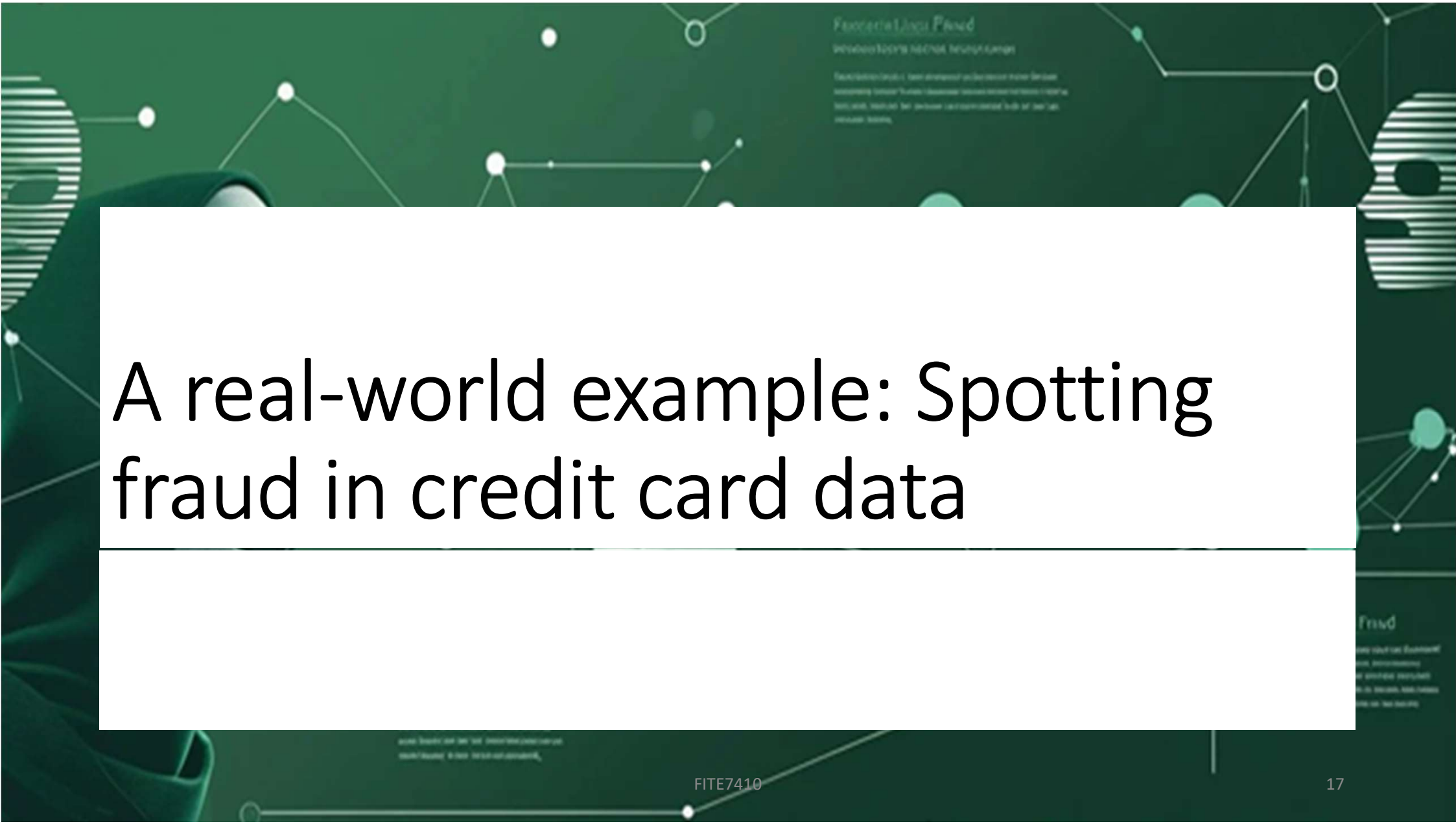
- Purpose of EDA:
  - to have a better understanding of the data before building the fraud detection model
  - to detect problems in data

- When performing EDA, fraud analyst need to keep the following in mind:
  - What is the purpose of this fraud analysis?
  - What are the insights or in-depth knowledge about the fraud that can be derived?
  - Are the objectives of EDA aligned with the problem on hand?

Step-by-step EDA (Data Pre-processing)



Ref: Ghosh et al. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets



A real-world example: Spotting fraud in credit card data

A credit card company has millions of transactions. How do we find the fraudulent ones?

The Anatomy of a Bust-Out

A bust-out fraud follows a predictable pattern that can be difficult to detect until it's too late.

1

Month 1-6: The Build-Up

A fraudster uses a synthetic identity to open a new credit card. For months, they make small, regular purchases and pay on time, building a perfect credit score. The bank's system sees a model customer.

2

Month 7: The Hit

The bank raises the credit limit.

3

Month 8: The Bust-Out

The fraudster maxes out the limit with untraceable purchases and vanishes.

The Analyst's Question:

How could we have detected this when every action looked legitimate?

Credit Card Frauds – some example scenarios

THE FBI FEDERAL BUREAU OF INVESTIGATION

CONTACT US | ABOUT US | MOST WANTED | NEWS | STATS

Newark Division

Home • Newark • Press Releases • 2013 • Eighteen People Charged in International $200 Million Credit Car

Info This is archived material from the Federal Bureau of Investigation (FBI) website. It may contain outdated information and links may no longer function.

Twitter Facebook Share

Eighteen People Charged in International $200 Million Credit Card Fraud Scam

Crime Ring Invented 7,000 Fake Identities to Obtain Tens of Thousands of Credit Cards

U.S. Attorney's Office **District of New Jersey**
February 05, 2013 (973) 645-2888

NEWARK—Federal agents in four states arrested 13 people today for allegedly creating thousands of phony identities to steal at least $200 million in one of the largest credit card fraud schemes ever charged by the Department of Justice, U.S. Attorney Paul J. Fishman announced.

- Credit Card fraud, 2 subtypes (Bolton and Hand, 2002)
  - Application fraud
    - involving individuals obtaining new credit cards from issuing companies by using false personal information
    - spending as much as possible in a short space of time
  - Behavioral fraud
    - where details of legitimate cards are obtained fraudulently and sales are made on a "Cardholder Not Present" basis
    - does not required stealing physical card
    - behavioral fraud concerns most credit card fraud

A Typology of Credit Card Fraud

Understanding the different types of credit card fraud is essential for detection and prevention.

Application Fraud

Using stolen/synthetic identities for "bust-out" schemes.

Card-Not-Present (CNP) Fraud

Stolen details for online/phone purchases (most common).

Skimming & Counterfeit Fraud

Stealing data to clone cards.

Collusive Merchant Fraud

Merchants process fake transactions, splitting profits with fraud rings.

Checklist for Credit Card Fraud

Use this checklist to identify potential fraud schemes through key indicators.

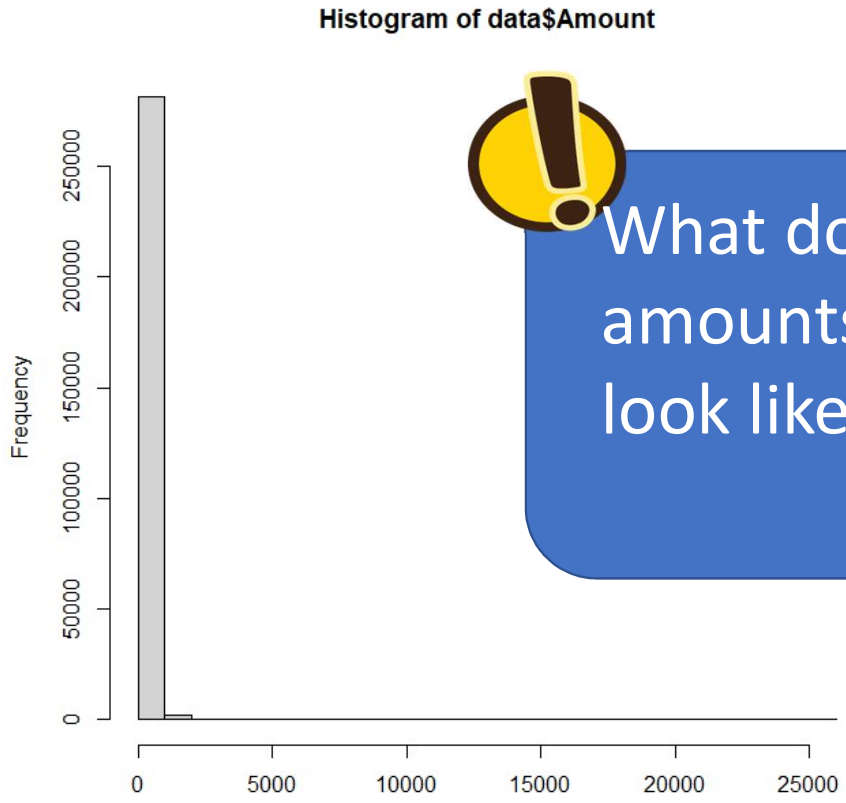
Investigative Question	Fraud Scheme Indicated	Red Flag
Is the activity too fast?	CNP Fraud, Skimming	
Is the location possible?	CNP Fraud, Skimming	
Are they testing the card?	CNP Fraud	
Does this fit the user?	Account Takeover	
Is the card new?	Application Fraud	

Example – Credit Card Fraud : “Amount”

```
help(hist)
```

Use “help” to check the R documentation

```
#Histogram
hist(data$Amount)
```



What do transaction amounts typically look like?

An Advanced Technique: Do the Numbers *Feel* Real?

Have you ever noticed what chance has a given number to start with the digit 1?

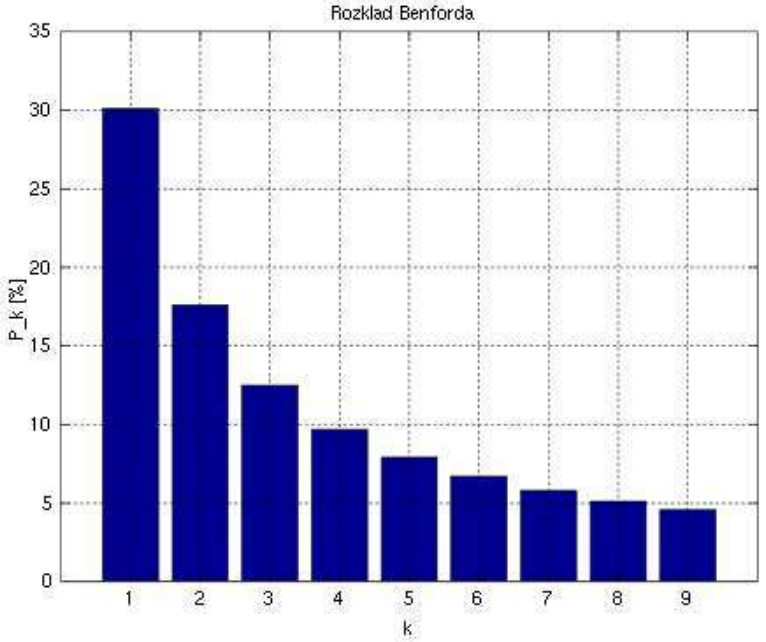
Does the digit 1 has the same probability to be a leading digit as 9?

Well, the answer is NO …

(Source: adapted from Dr. KP Chow's note)

Benford's Law (The first digit law)

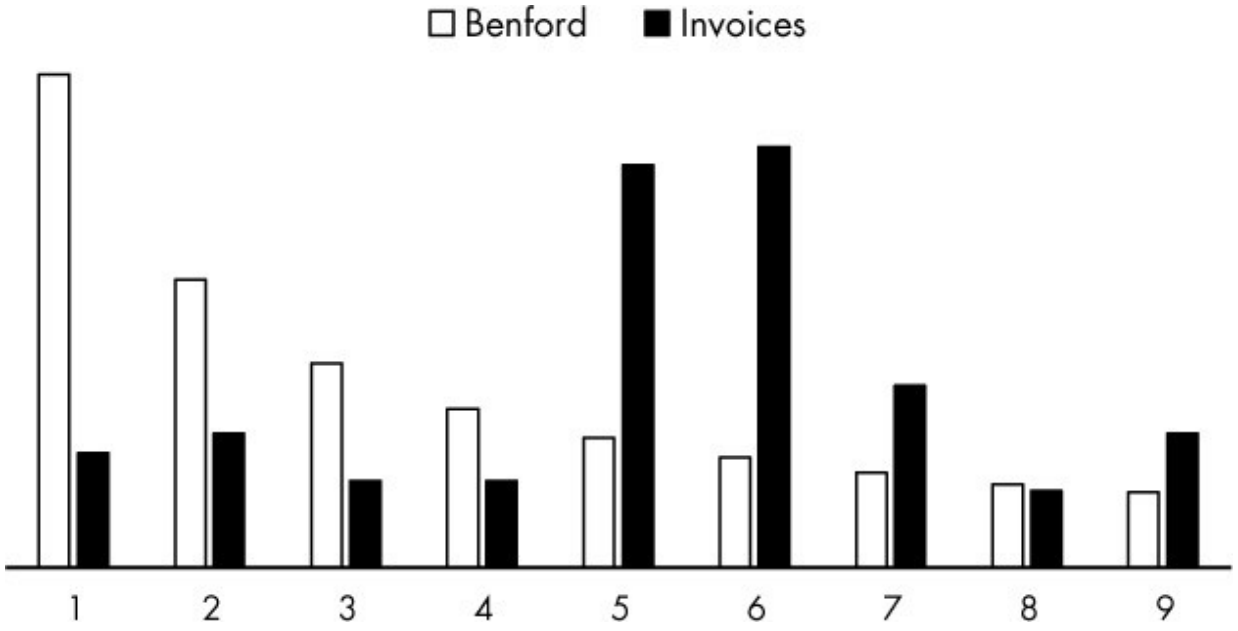
- The frequency of occurrence of the leading digits in "naturally occurring" numerical distributions is predictable and nonuniform, but closer to a power-law distribution
- A given number is six times more likely to start with a 1 than a 9!



(Source: adapted from Dr. KP Chow's note)

Example: Splitting the invoice

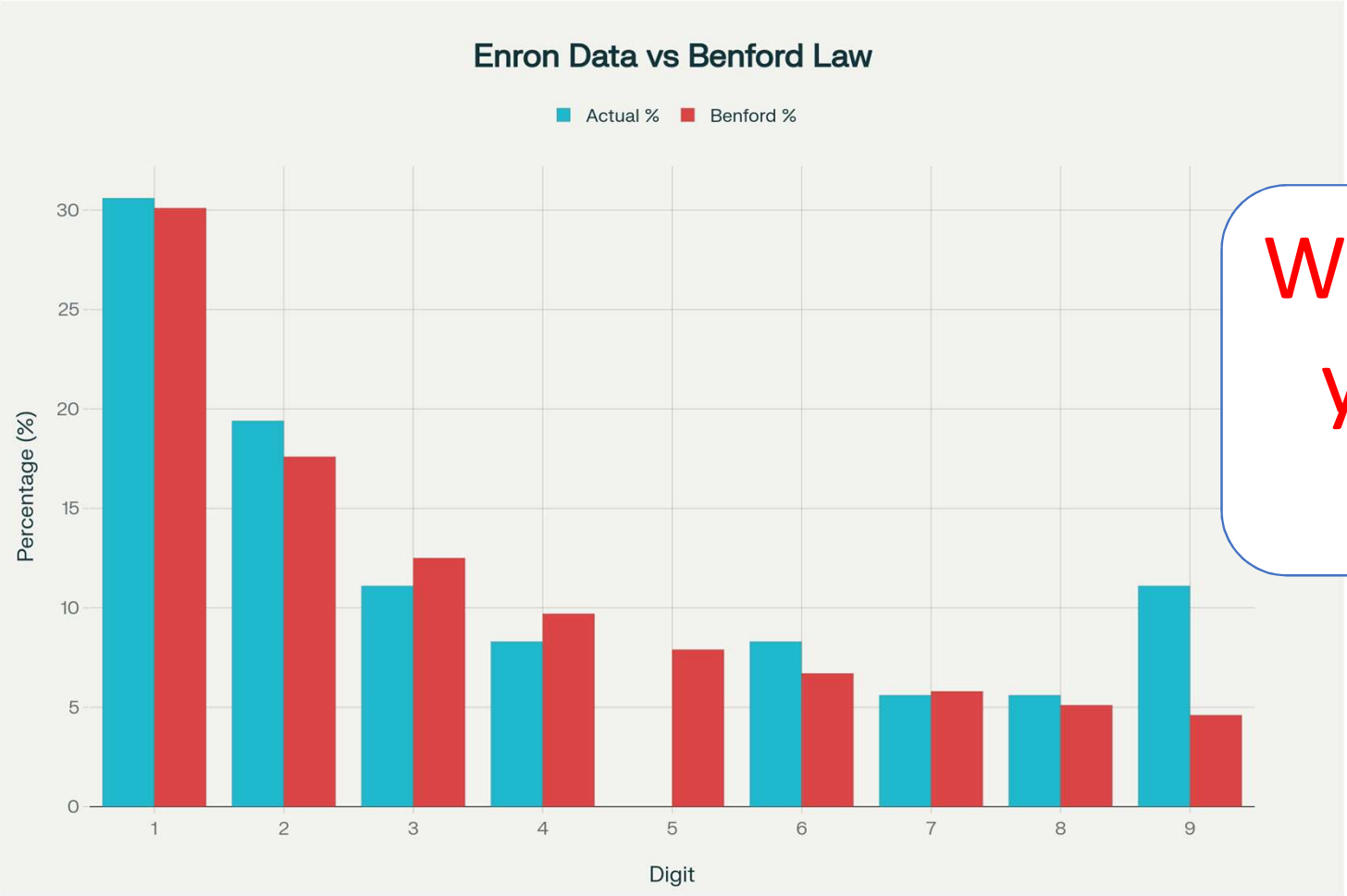
- Consider a company where any travel and entertainment expenses over $10,000 must be approved by the vice president
- Employees will split invoice with amount over $10,000 to avoid the “approval”



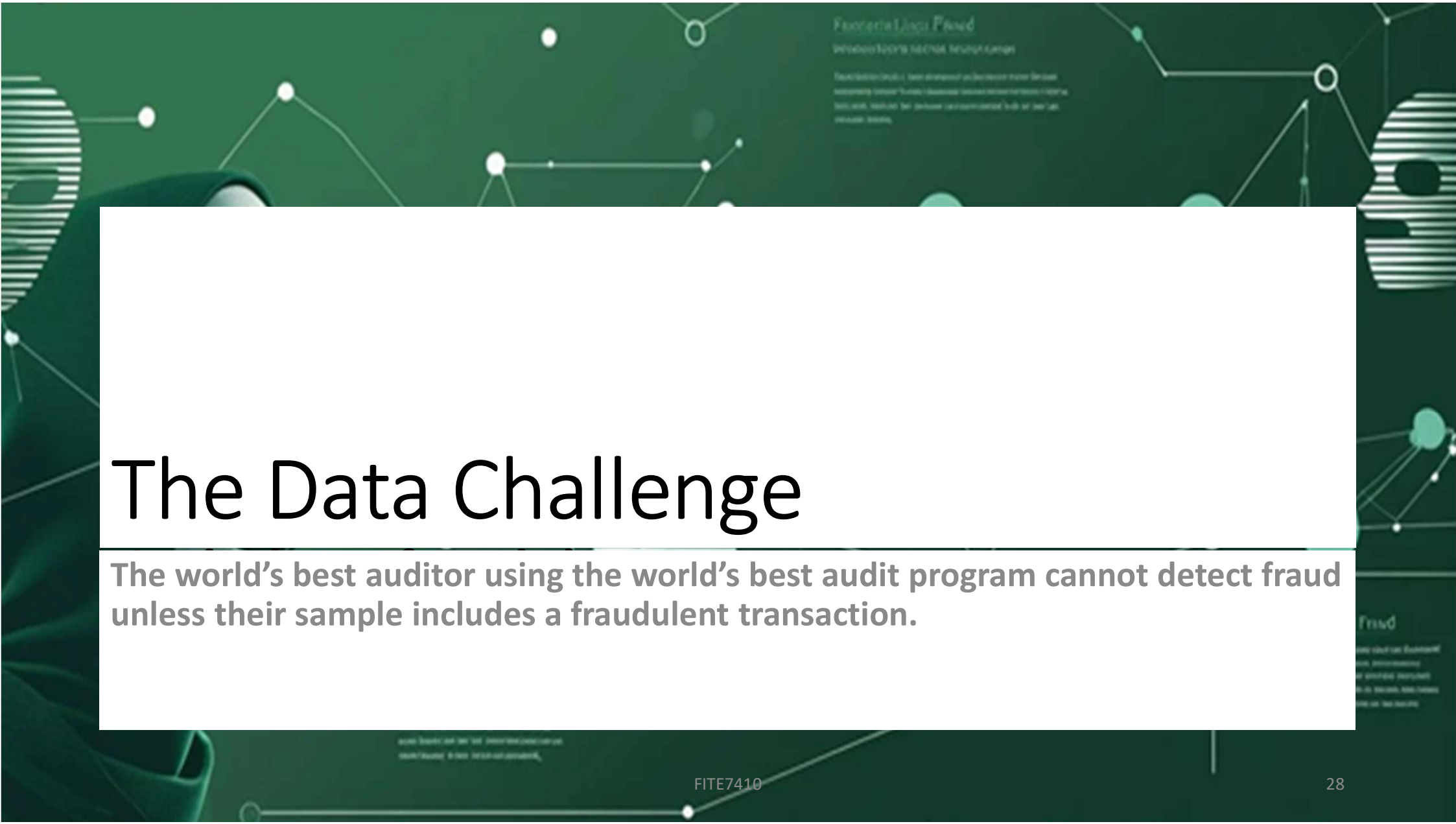
(Source: adapted from Dr. KP Chow’s note)

A spike in first-digit frequencies around 5 and 6, clear violation of Bendford’s Law

Benford's Law Analysis: Enron Financials (1997-2000)



What insight do you get from this chart?



The Data Challenge

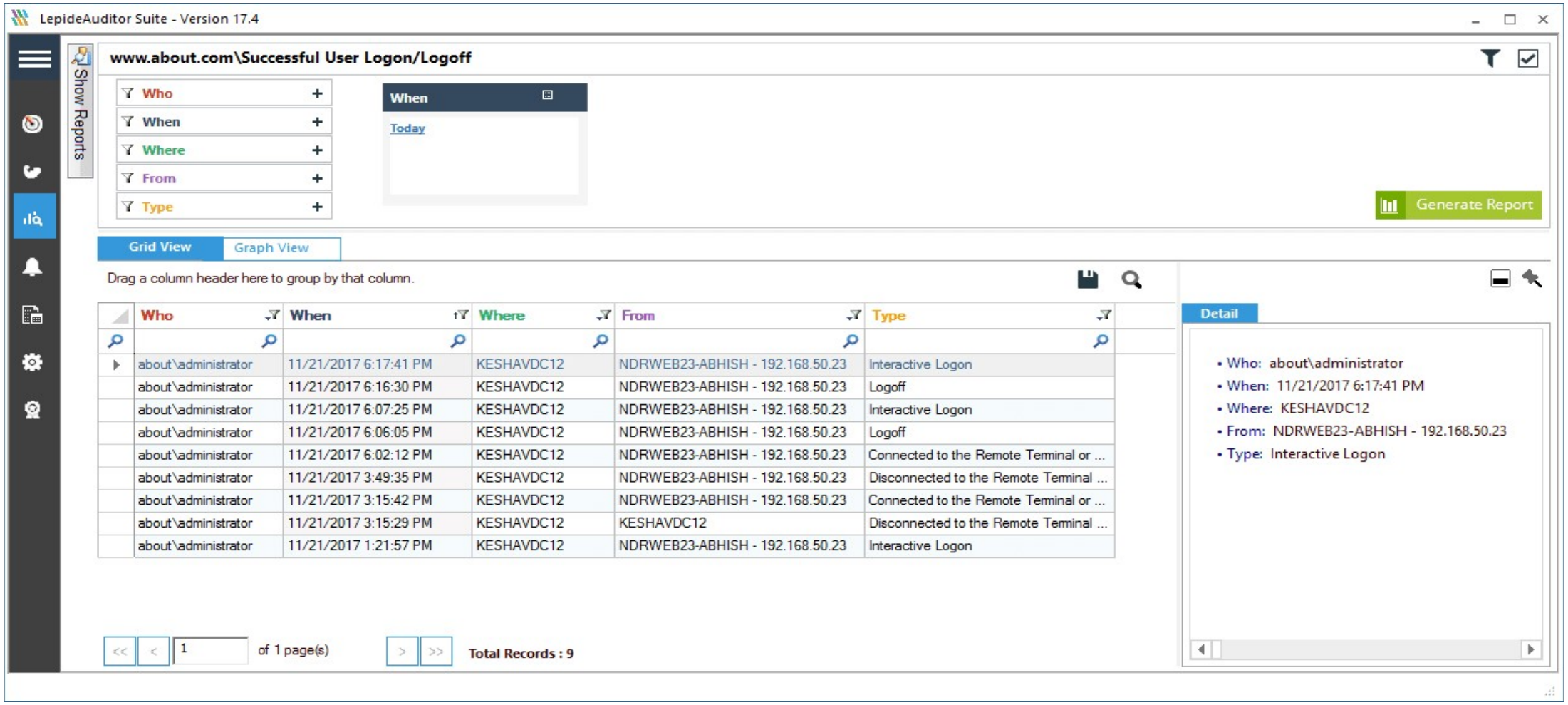
The world's best auditor using the world's best audit program cannot detect fraud unless their sample includes a fraudulent transaction.

What are “Fraud Data”?

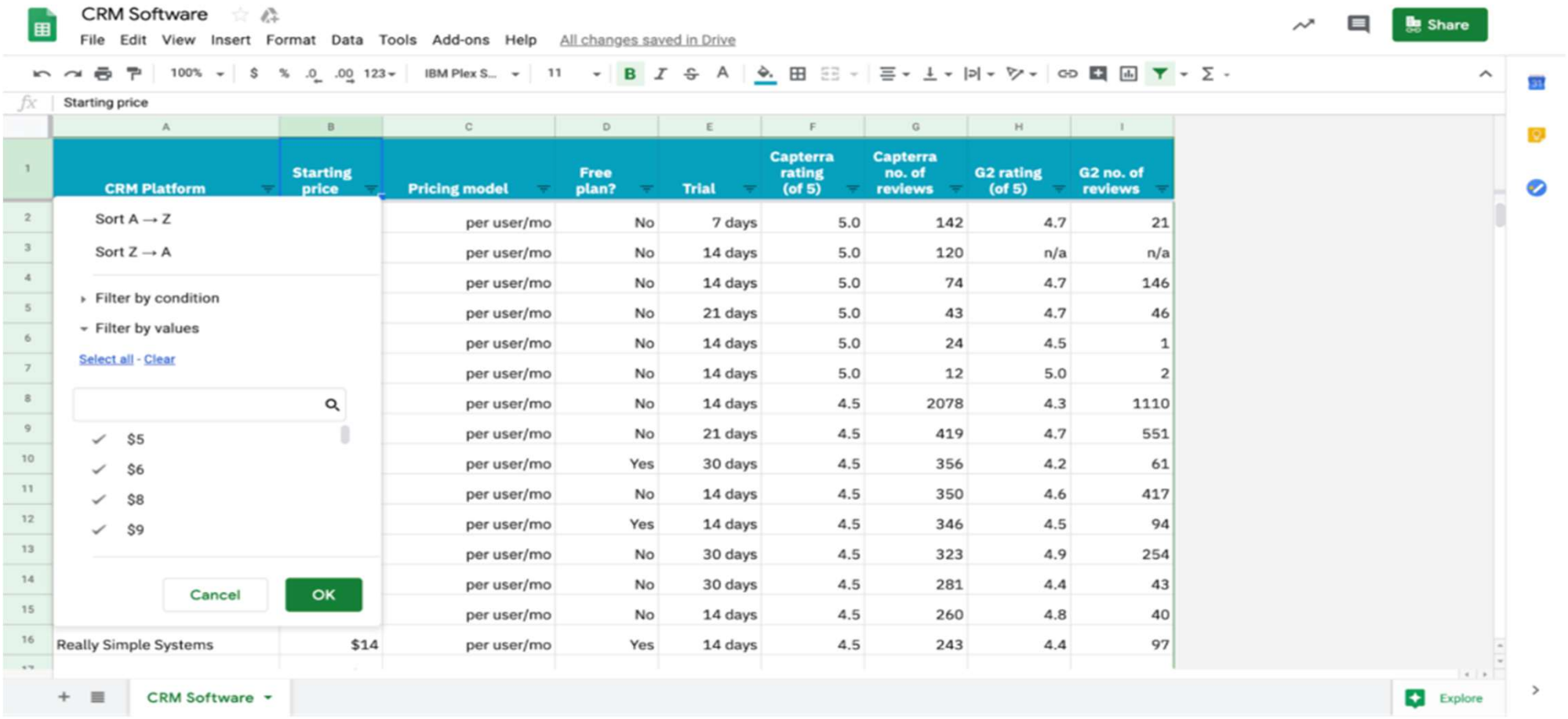
- Examples of real-world raw data:
  - Bank Statements?
  - Network logs?
  - Customer data?
  - Firewall logs?
  - Email?
  - Whatsapp/Wechat/LINE or any other IM messages?
  - FB/IG/MeWe or any other social media pages?
  - Public information, e.g. news?
  - Documentation, e.g. user guides?
  - Server room check-in and check-out logbook?
  - CCTV?
  - …

Which of these are “Fraud Data”?

Example of raw data – user login access log



Example of raw data – CRM records



CRM Platform	Starting price	Pricing model	Free plan?	Trial	Capterra rating (of 5)	Capterra no. of reviews	G2 rating (of 5)	G2 no. of reviews
		per user/mo	No	7 days	5.0	142	4.7	21
		per user/mo	No	14 days	5.0	120	n/a	n/a
		per user/mo	No	14 days	5.0	74	4.7	146
		per user/mo	No	21 days	5.0	43	4.7	46
		per user/mo	No	14 days	5.0	24	4.5	1
		per user/mo	No	14 days	5.0	12	5.0	2
		per user/mo	No	14 days	4.5	2078	4.3	1110
		per user/mo	No	21 days	4.5	419	4.7	551
		per user/mo	Yes	30 days	4.5	356	4.2	61
		per user/mo	No	14 days	4.5	350	4.6	417
		per user/mo	Yes	14 days	4.5	346	4.5	94
		per user/mo	No	30 days	4.5	323	4.9	254
		per user/mo	No	30 days	4.5	281	4.4	43
		per user/mo	No	14 days	4.5	260	4.8	40
Really Simple Systems	$14	per user/mo	Yes	14 days	4.5	243	4.4	97

Example of raw data - email

Here's what Jeffrey Skilling's folders look like:

```
localhost:CMU$ cd skilling-j

localhost:skilling-j$ ls

_sent_mail      deleted_items
all_documents   discussion_threac
calendar        inbox
contacts        mark
```

The emails are in individual text files:

```
..evans@thyme>

com>

lders\Sent

eat to meet you.  I look forward to
seeing more of you in the future.

Regards,
Jeff
```

What are the possible problems with the raw data collected?

Sample raw data files extracted from Enron case

“Data” used in Fraud Detection

- What are the problems with real-life data?
  - Inconsistencies
  - Incompleteness
  - Duplication
  - Merging
  - Data size too big to handle
  - … and many more

- Thus, it is critical to pre-process the raw data before proceeding to conduct the analysis steps

“Data” used in Fraud Detection

- What are the types of data sources?
  - Structured
    - Transactional data
    - Contractual, subscription, or account data
    - Surveys
    - Behavioural information
  - Unstructured
    - Text documents, e.g. emails, web pages, claim forms or multimedia contents
    - Contextual or network information
    - Qualitative expert-based data
    - Publicly available data
  - Semi-structured

Some Basic Concepts - Data Table

Columns = variables, fields, characteristics, attributes, features, etc.

Rows = instances, observations, lines, records, tuples, etc.

Executive Role	First Name	Last Name	Primary Email
CEO	David	Delainey	david.w.delainey@enron.com
CEO	Kenneth	Lay	kenneth.lay@enron.com
CEO	Jeffrey	Skilling	jeff.skilling@enron.com
Energy Trader	Timothy	Belden	tim.belden@enron.com
Sr. VP Broadband	Scott	Yeager	scott.yeager@enron.com
Sr. VP Eng.Ops Broadband	Rex	Shelby	rex.shelby@enron.com
CEO Broadband	Kenneth	Rice	kenneth.rice@enron.com
CEO Broadband	Joseph	Hirko	joe.hirko@enron.com
COO Broadband	Kevin	Hannon	kevin.hannon@enron.com
CFO	Andrew	Fastow	andrew.fastow@enron.com
Dir Global Finance	Michael	Kopper	michael.kopper@enron.com
Dir Investor Relations	Mark	Koenig	mark.koenig@enron.com
Treasurer CFO	Raymond	Bowen	ray.bowen@enron.com
Vice President	Christopher	Calger	chris.calger@enron.com
Chief Accounting Officer	Richard	Causey	richard.causey@enron.com
Chief Accounting Officer	Wesley	Colwell	wes.colwell@enron.com
Treasurer	Ben	Gilsan	ben.gilsan@enron.com
CFO Global Power Pipelines	Paula	Rieker	paula.rieker@enron.com

Sample data records extracted from Enron case

Basic Concepts – Data types

Continuous data

Defined on an interval, with limited or unlimited value	With or without a natural zero value	Example: • amount of transactions • balance on savings account • similarity index

Basic Concepts – Data types

Categorical data

Nominal

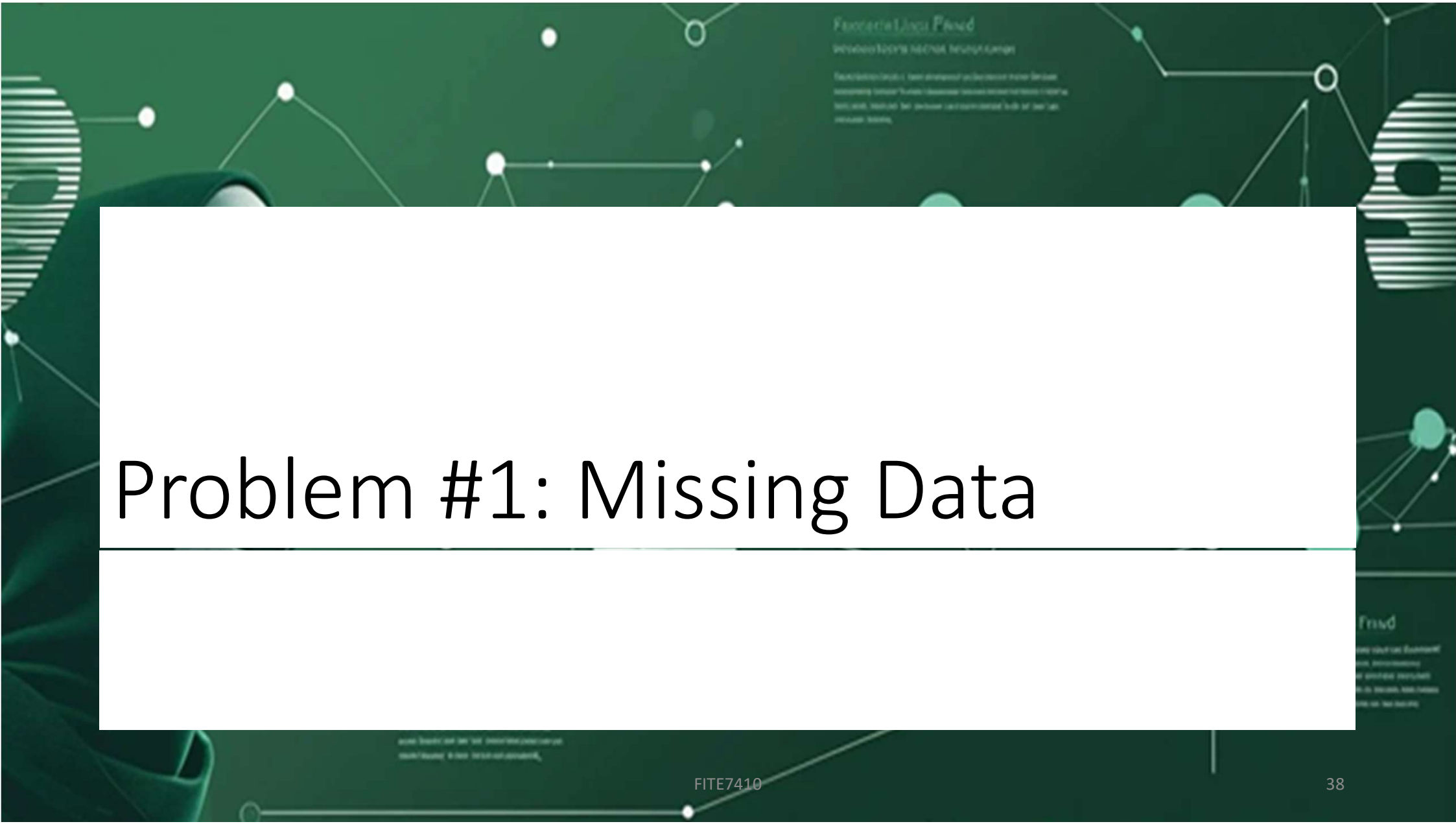
- limited set of values with no meaningful ordering in between
- e.g. marital status; payment type; country of origin

Ordinal

- take on a limited set of values with a meaningful ordering in between
- e.g. age coded as young, middle-age, and old

Binary

- can only take on two values
- e.g. yes/no



Problem #1: Missing Data

Erroneous and Missing Data

- Reasons of erroneous and missing data
  - Human input error
  - Intentionally hiding some information
  - Not applicable values, e.g. if there are records without visa card, visa card transactions will be not applicable
  - Not matching search or filter criteria, e.g. if transaction > 1 billion

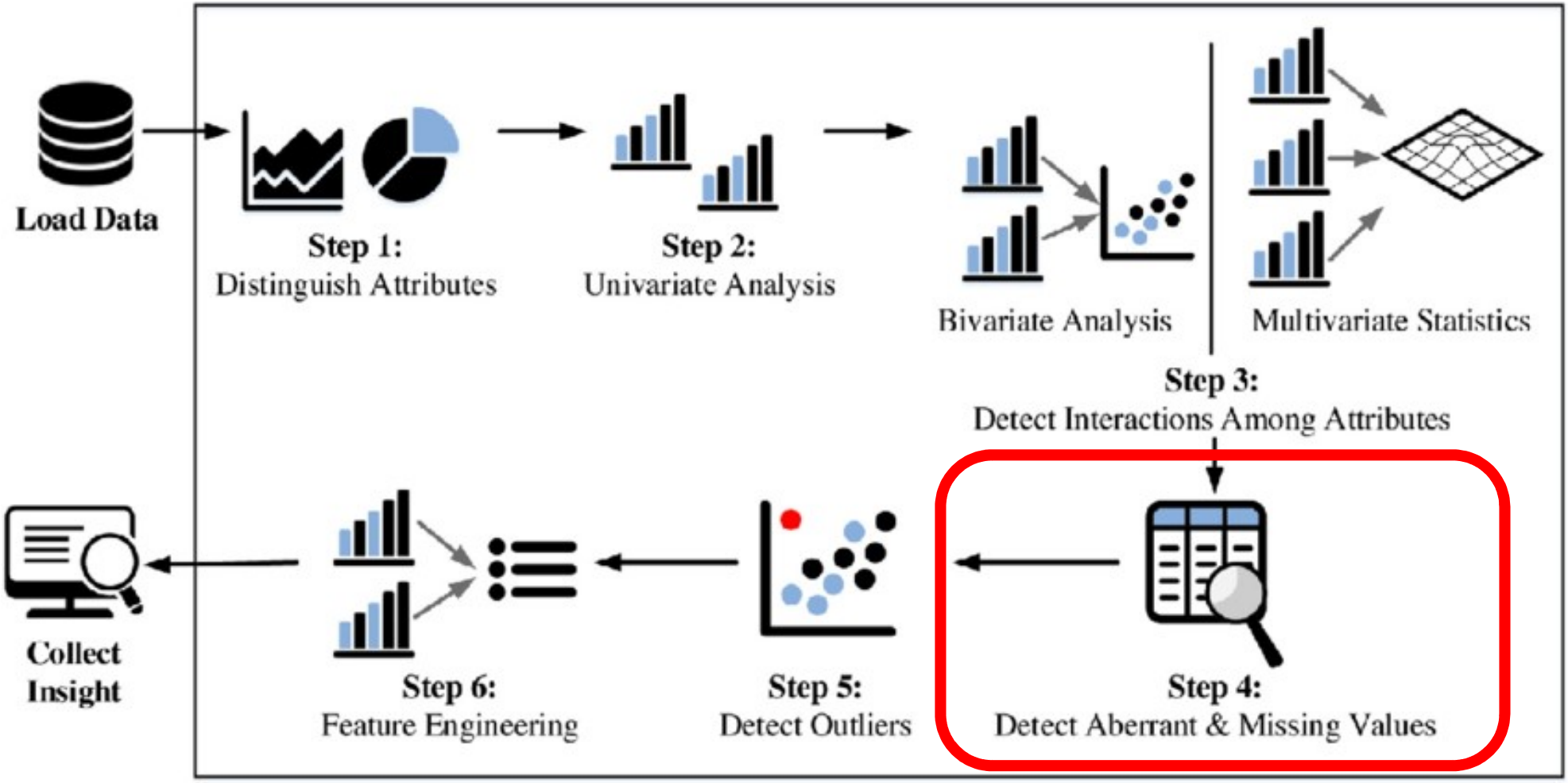
Table 2.1 Dealing with Missing Values

ID	Age	Income	Marital Status	Credit Bureau Score	Fraud
1	34	1,800	?	620	Yes
2	28	1,200	Single	?	No
3	22	1,000	Single	?	No
4	60	2,200	Widowed	700	Yes
5	58	2,000	Married	?	No
6	44	?	?	?	No
7	22	1,200	Single	?	No
8	26	1,500	Married	350	No
9	34	?	Single	?	Yes
10	50	2,100	Divorced	?	No

Data Cleaning

- Raw data contains noise, inconsistencies and incompleteness.
- “Dirty” data can cause confusion for the data analytics procedure.
  - Incomplete/Missing data
  - Noisy data (outliers or errors)
  - Data inconsistencies (similar to data conflict in data integration)
  - Duplicate records (similar to data duplication in data integration)
- Thus, data cleaning is an essential process in data pre-processing.

Step-by-step EDA (Data Pre-processing)



Ref: Ghosh et al. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets

EDA – Step 4: Detect erroneous & missing values

Objective

- Aberrant and missing values may result in biased analysis of data. Thus, need to identify and detect any outliers and missing values in the dataset

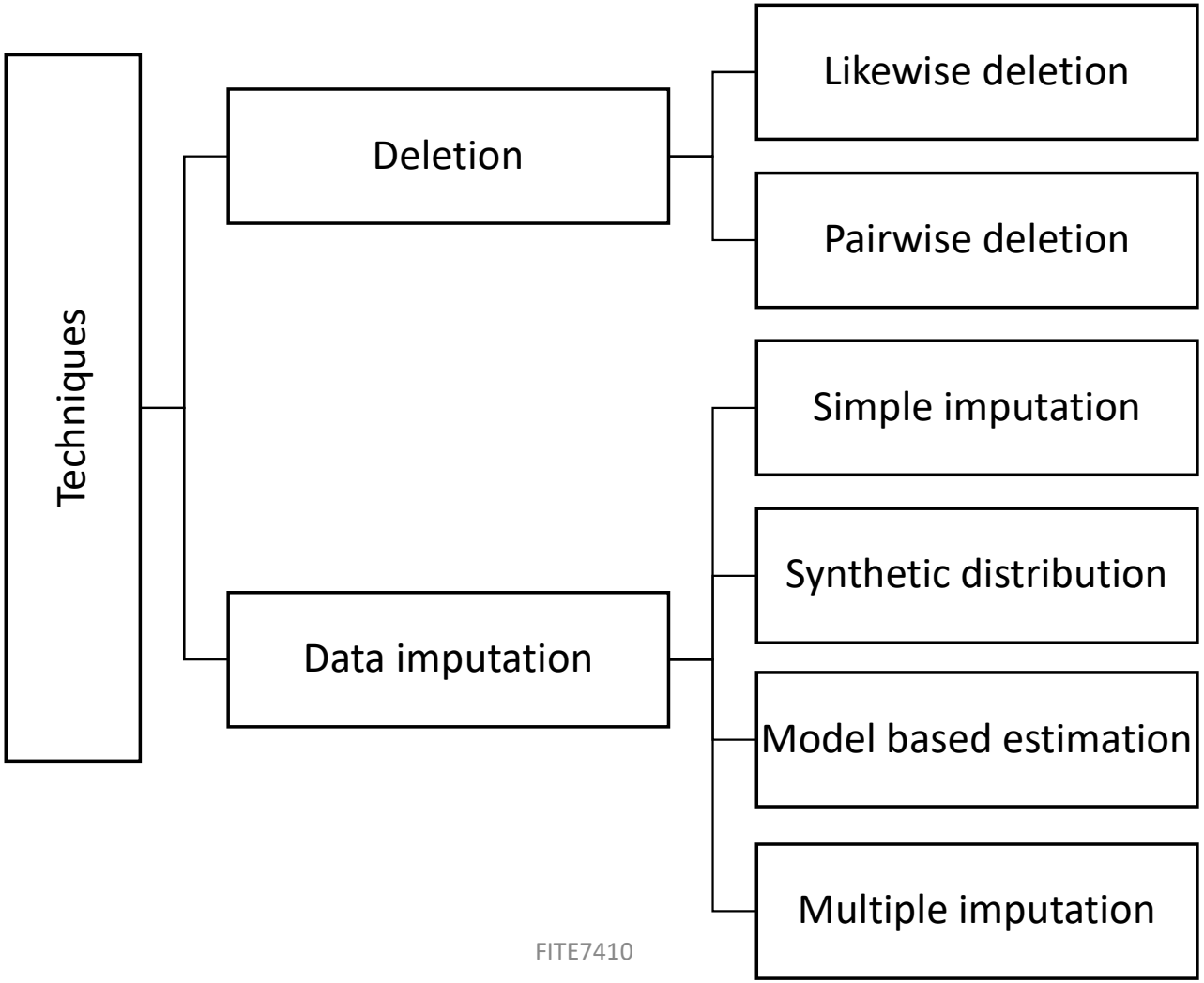
Explore what?

- Erroneous values : Erroneous values which occur as a result of incorrect user inputs or calculation errors
- Missing values : Occur in a dataset during data extraction and/or data collection.

Techniques

- Performs AFTER multivariate analysis when you have a clearer idea about the attributes
- Detection of abnormalities in univariate, bivariate and multivariate visualizations

How to handle missing data?



Handling Missing Data - deletion

Likewise/Listwise Deletion

Mobile ID	Mobile Package	Download Speed	Data Limit Usage	
1	Fast+	157	80%	
2	Lite	99	70%	
3	Fast+	167	10%	
4	Fast+	N/A	80%	← Delete
5	Lite	76	70%	
6	Fast+	155	10%	
7	N/A	N/A	95%	← Delete
8	Lite	76	77%	
9	Fast+	180	N/A	← Delete

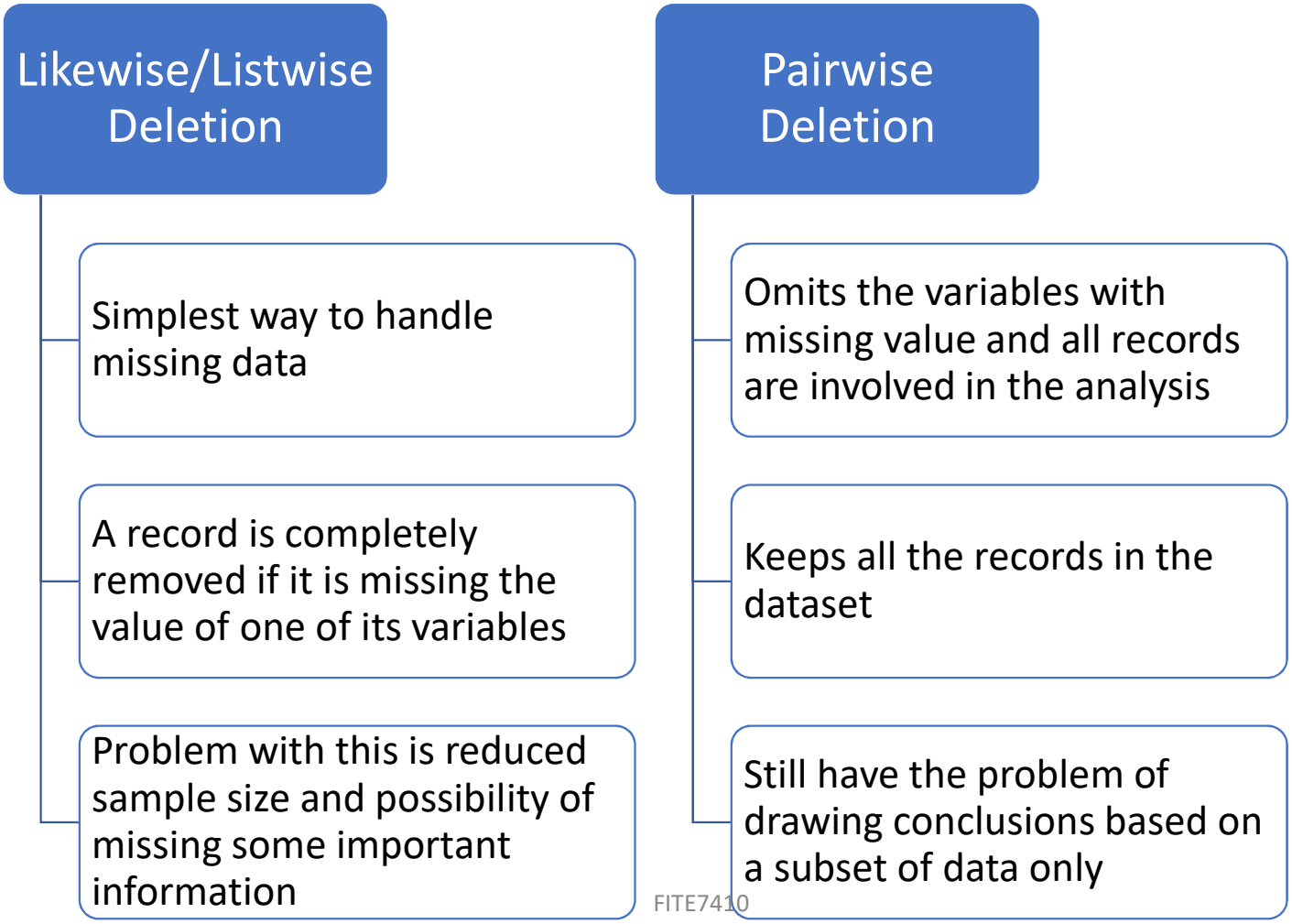
Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
5	Lite	76	70%
6	Fast+	155	10%
8	Lite	76	77%

Pairwise Deletion

Mobile ID	Mobile Package	Download Speed	Data Limit Usage	
1	Fast+	157	80%	
2	Lite	99	70%	
3	Fast+	167	10%	
4	Fast+	N/A	80%	← Delete
5	Lite	76	70%	
6	Fast+	155	10%	
7	N/A	N/A	95%	← Delete
8	Lite	76	77%	
9	Fast+	180	N/A	← Delete

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+		80%
5	Lite	76	70%
6	Fast+	155	10%
7			95%
8	Lite	76	77%
9	Fast+	180	

Handling Missing Data - deletion



Handling Missing Data-Imputation

- Data Imputation : Fill in the missing data

MICE – a package in R

Simple imputation	Synthetic distribution	Model-based estimation	Multiple imputation
Filling in missing values with a single value, such as the mean, median, mode, or a constant.	Fill in the missing data based on a synthetic distribution curve	Use a predictive or statistic model to impute the missing data based on observed data, e.g. regression, random forest, knn	Use the distribution of the observed data to estimate a set of plausible values for the missing data. Random components are incorporated into these estimated values to show their uncertainty.
• Easy to implement • Can keep the sample mean value	• Easy to understand and communicate • Preserves the distribution of the data. • Captures the relationships between variables.	• Keeps all the records and avoids altering the mean or distribution • Can provide more accurate imputations compared to simple methods.	• Keeps the variability and uncertainty of missing data, resulting in more valid statistical inference. • Accounts for uncertainty in imputed values. • Provides more accurate estimates of parameters and standard errors.
• May lead to bias if the sample is not completely random • Does not take into account relationships between variables.	Assumes outcome remains unchanged by missing data, but in real-world this usually is not the case	Missing value is created by prediction based on values from other variables, that means no actual novel information is included	• Need to build separate models for data imputation, in addition to the analysis model. • More complex and computationally intensive

Example –Step 4: Detect erroneous & missing values

```
# Load csv dataset
data <- read.csv('../input/creditcardfraud/creditcard.csv')
```

Path of the source file

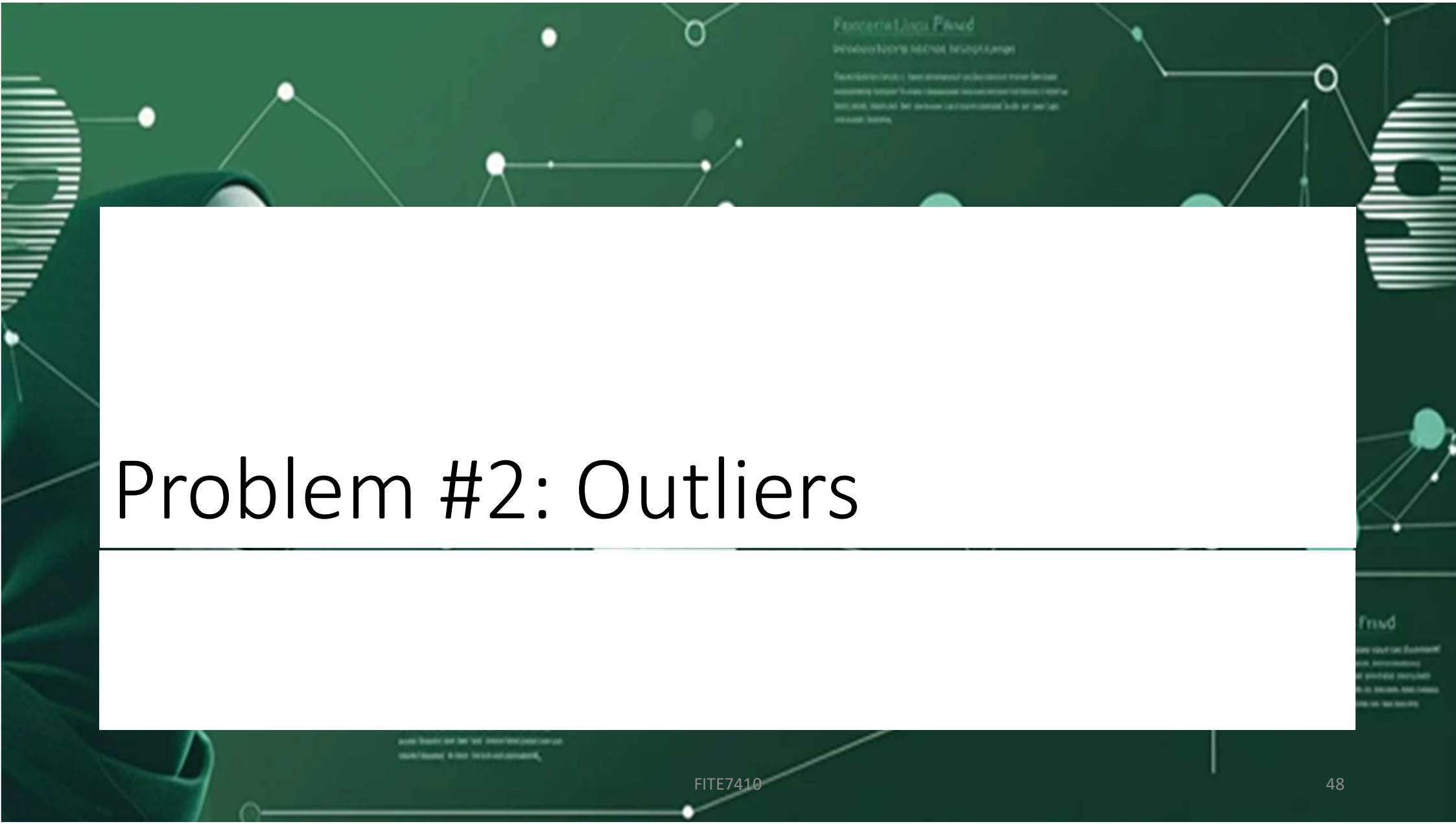
```
# Total missing values in the data set
print('Total missing values')
sum(is.na(data))

# Missing values per column
colSums(is.na(data))
```

"is.na" is the function to check missing data

```
 [1] "Total missing values"
0
```

Time: 0 **V1:** 0 **V2:** 0 **V3:** 0 **V4:** 0 **V5:** 0 **V6:** 0 **V7:** 0 **V8:** 0 **V9:** 0 **V10:** 0 **V11:** 0 **V12:** 0 **V13:** 0 **V14:** 0 **V15:** 0 **V16:** 0 **V17:** 0 **V18:** 0 **V19:** 0 **V20:** 0 **V21:** 0 **V22:** 0 **V23:** 0
V24: 0 **V25:** 0 **V26:** 0 **V27:** 0 **V28:** 0 **Amount:** 0 **Class:** 0



Problem #2: Outliers

Outliers

- Outliers – extreme observations that are dissimilar to the rest of the population
- Reasons of outliers
  - Valid observations: e.g. salary of senior management > US$1M
  - Invalid observations: e.g. age > 300

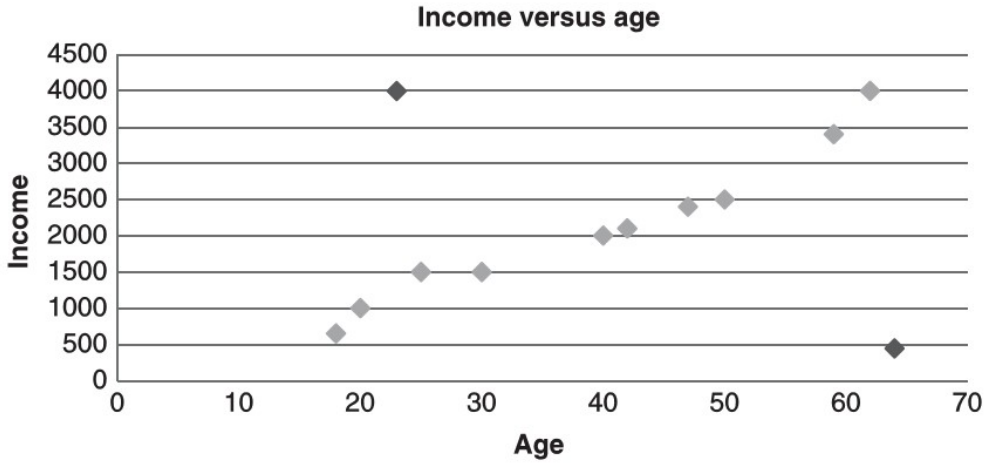
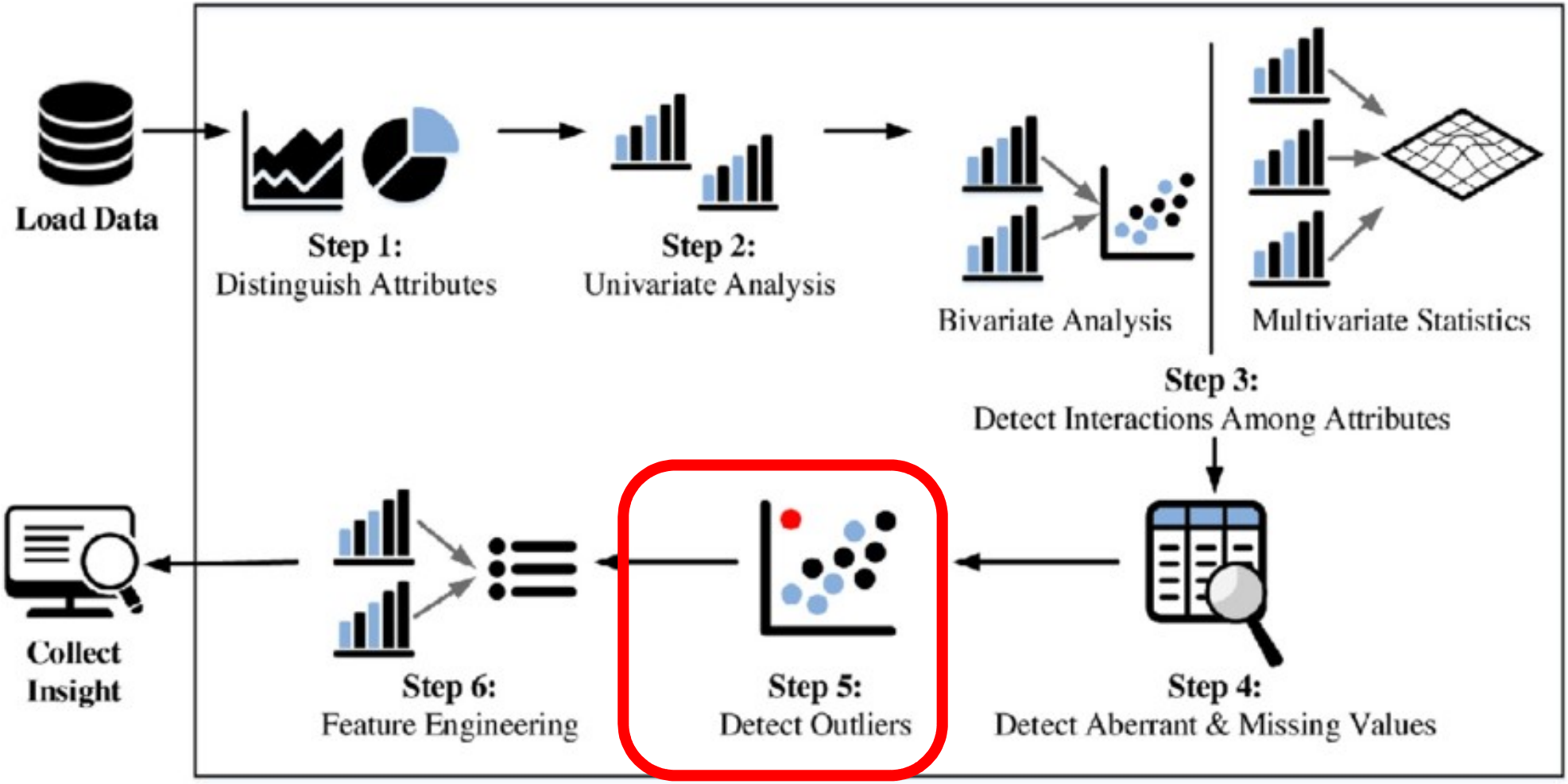


Figure 2.4 Multivariate Outliers

Step-by-step EDA (Data Pre-processing)



Ref: Ghosh et al. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets

EDA – Step 5: Detect outliers

Objective

- Outliers might result in biased analysis of data

Explore what?

- Outliers in a dataset can be primarily of three types namely: univariate,
  bivariate, and multivariate outliers

Techniques

- Univariate outliers : can be detected by calculation of the Inter-Quartile
  Range (IQR)
- Bivariate and multivariate outliers : need to inspect correlations among different attributes

Detection of outliers

- Calculate minimum, maximum, z-score values for each attribute. Define outliers when absolute value of z-score is larger than 3.
- Use visualization, e.g. histogram, box plots

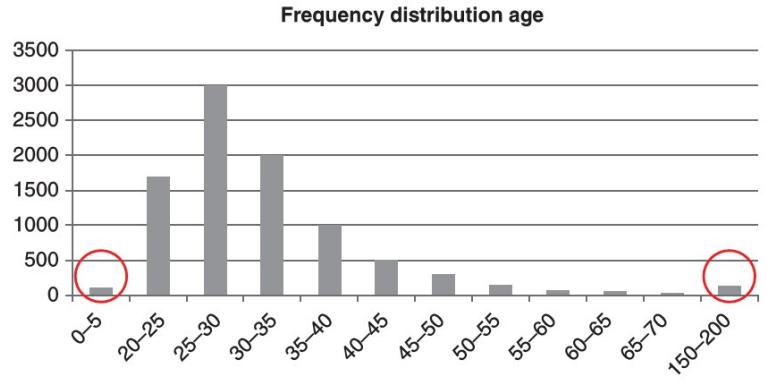


Figure 2.5 Histogram for Outlier Detection

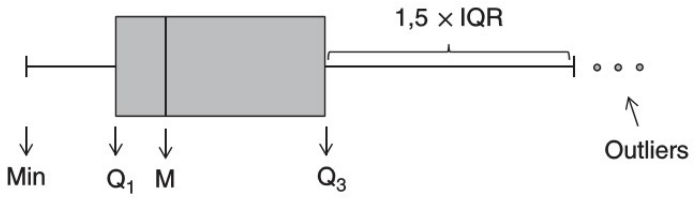


Figure 2.6 Box Plots for Outlier Detection

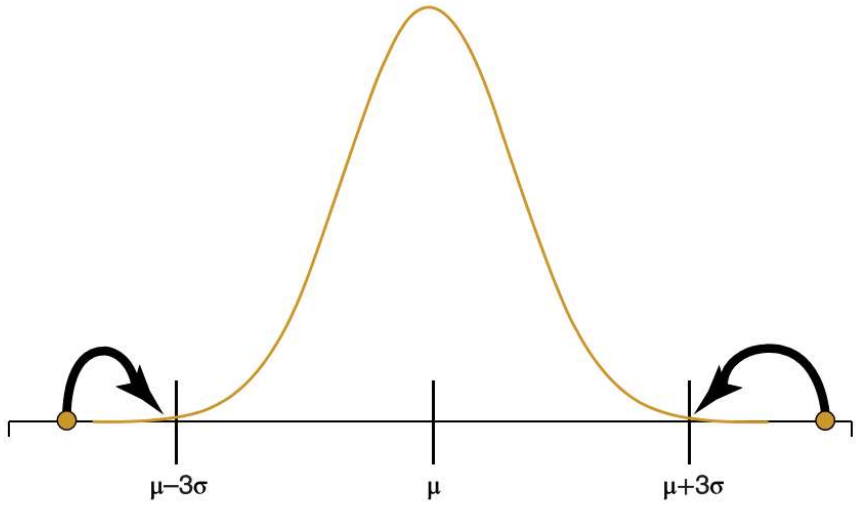
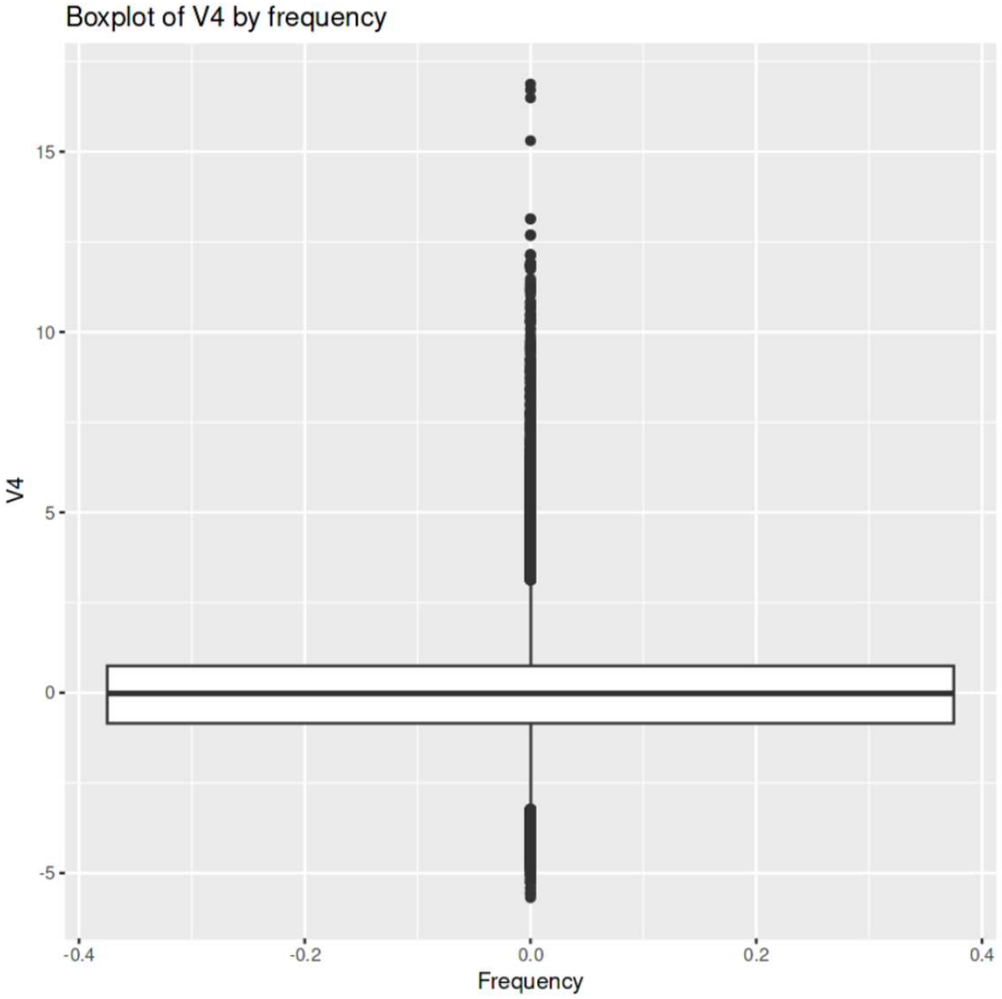


Figure 2.7 Using the *z*-Scores for Truncation

Example – Step 5: Detect outliers




```r
# Boxplot of variable V4
ggplot(data, aes(y=V4)) +
        geom_boxplot() +
        labs(x = 'Frequency', y = 'V4') +
        ggtitle('Boxplot of V4 by frequency')
```

The Critical Question: Outlier or Red Flag?

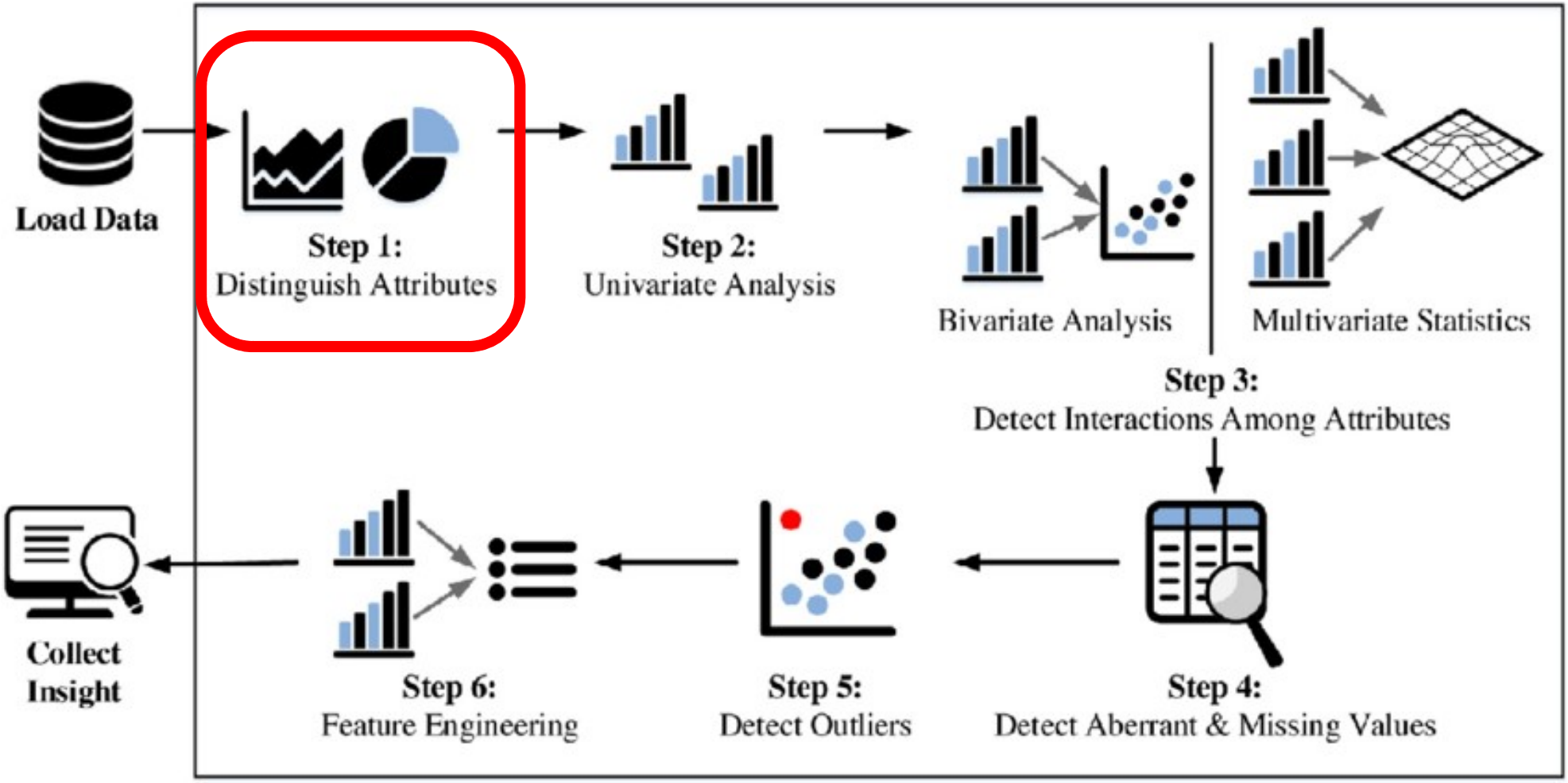
- In cases of fraud analysis, outliers may sometimes be actually the fraud cases that concern us
- Behaviours of fraudsters usually deviant from normal non-fraudsters.
- These deviations from normal patterns are Red Flags of fraud, e.g.
  - Identical financial statements may be Red Flag for tax evasion
  - Small payment followed by a large payment immediately after may be Red Flag for credit card fraud
- Caution should be taken when treating outliers, or marked these observations as outliers for further analysis

Handling outliers

- Treatment
  - For invalid observations, treat the outlier as missing value and can use one of the techniques for handling missing value to deal with outlier value
  - For valid observations, truncation/capping/winsorizing, i.e. set upper and lower limits on a variable.
    - Z-socre (stand deviation), e.g. upper/lower limit = M ± 3 x z-score
    - IQR (interquartile range), e.g. upper/lower limit = M ± 3 x IQR/(2 x 0.6745) (Van Gestel and Baesens, 2009)

Using EDA: The Initial Investigation

Step-by-step EDA (Data Pre-processing)



Ref: Ghosh et al. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets

EDA – Step 1: Distinguish Attributes

Objective

- To identify the attributes in a dataset in order to formulate a clearer goal of the data analytics process
- To understanding the meaning of each attribute before analyzing the data

Explore what?

- Attribute names, datatypes, number of attributes, etc.
- Continuous vs Categorical data types

Techniques

- Descriptive summary of the dataset

EDA – Statistics Examples

- Make use of descriptive statistics, including, mean, median, standard deviations, percentile distribution, etc.
  - Mean value and median value are basic descriptive statistics for continuous variables; while mode (i.e. most frequently occurring value) is used for categorical variables
  - Standard deviation can provide insight with respect to how much data is spread around the mean.
  - Percentile value like 10, 25, 75, 90 percentile provide information about the distribution of the data

- Descriptive statistics can be used as a preview to check the symmetry or asymmetry of the distribution, i.e. skewness of the data.

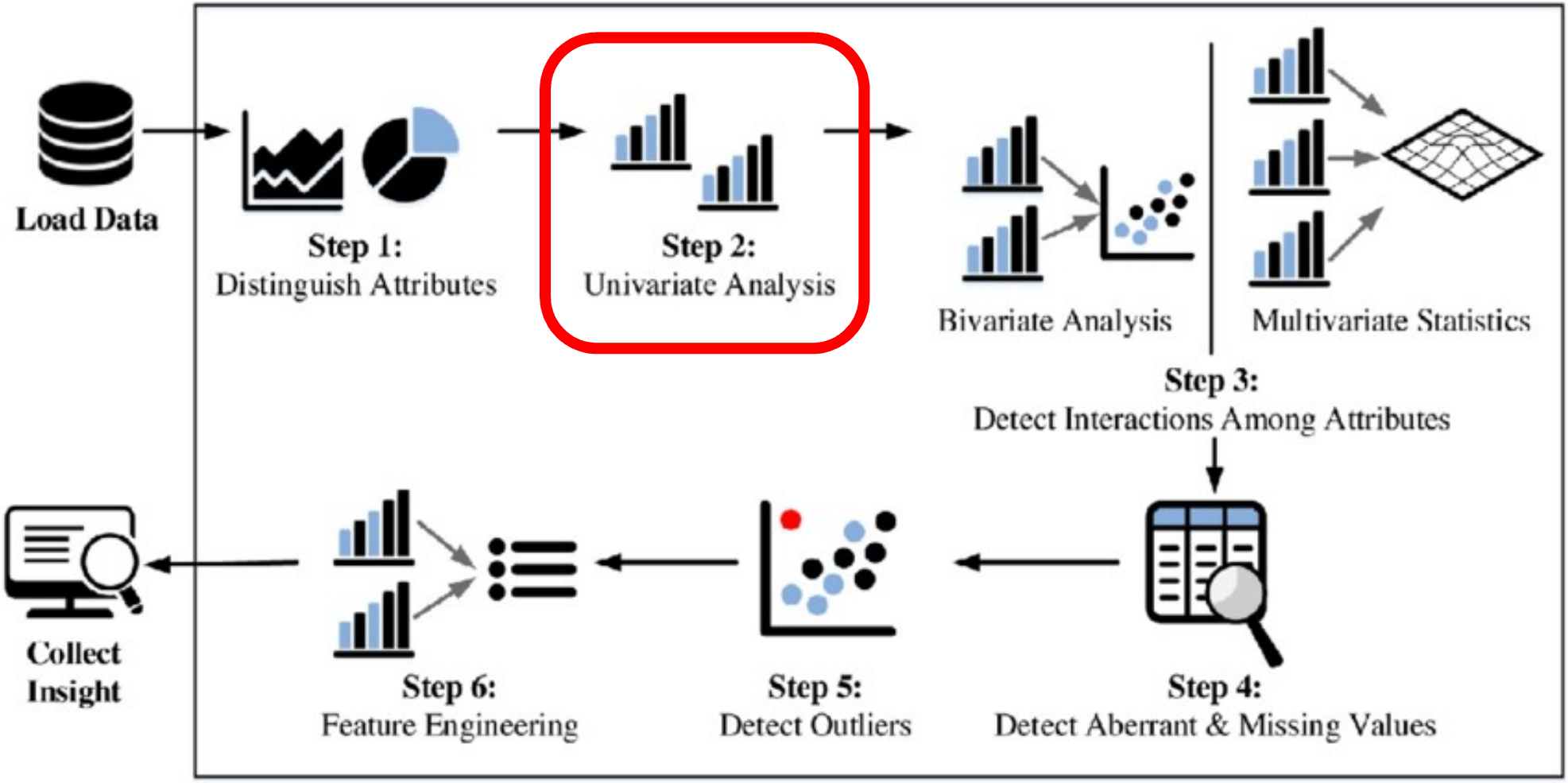
EDA – statistics examples

- Basic statistics measurement

	Time	V1	V2	V3	V4
count	1282.000000	1282.000000	1282.000000	1282.000000	1282.000000
mean	95362.820593	-0.005273	-0.031155	-0.040710	0.049703
std	47314.564520	2.121696	1.935197	1.517273	1.490202
min	172.000000	-40.470142	-37.520432	-17.474421	-3.559353
25%	55779.250000	-0.921363	-0.622638	-0.886693	-0.831241
50%	87777.500000	0.015252	0.076608	0.171572	0.033485
75%	138647.000000	1.296679	0.843327	0.969347	0.793492
max	172704.000000	2.312894	6.762856	3.428548	11.427809

- In additional to calculating the above statistics for the whole sample set, can also calculating for each target set (i.e. fraud and non-fraud cases) to observe the differences between the target sets (e.g. different average age)

Step-by-step EDA (Data Pre-processing)



Ref: Ghosh et al. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets

EDA – Step 2: Univariate Analysis

Objective

- Gain a better or deeper understanding of each attribute
- Identify combinations of attributes for subsequent analysis
- Identify “noise” in the dataset, e.g. missing values, outliers
- Discretize continuous variables, i.e. convert continuous variables into categorical variables

Explore what?

- Centrality: i.e. mean, median, mode
- Dispersion: i.e. range, variance, standard deviation, skewness and kurtosis

Techniques

- Visualization tools, e.g. histogram, boxplot, scatter-plot

Visualization

- Descriptive statistics sometimes may be difficult to interpret, visual plots of the distributions of the involved variables will give insights about the data and problem
- Missing data, outliers, and distribution can be more easily identified using visualization of the data
- Visualization might include plotting of histograms, box plots, etc.

EDA – visualization example

- Pie Charts – represents a variable's distribution
- Histogram – visualize the central tendency and variability of the data
- Scatter plot – visualize the correlation patterns in data
- …

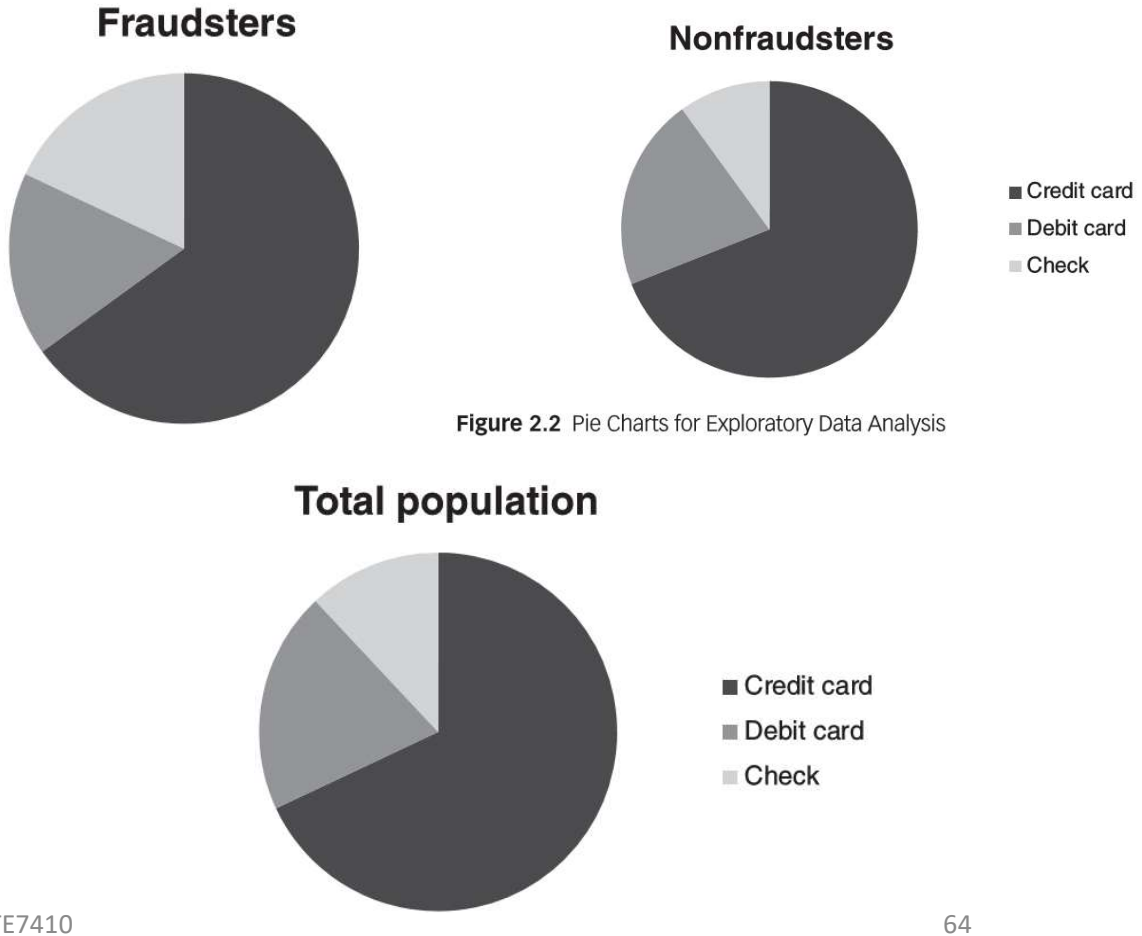


Figure 2.2 Pie Charts for Exploratory Data Analysis

EDA – visualization example

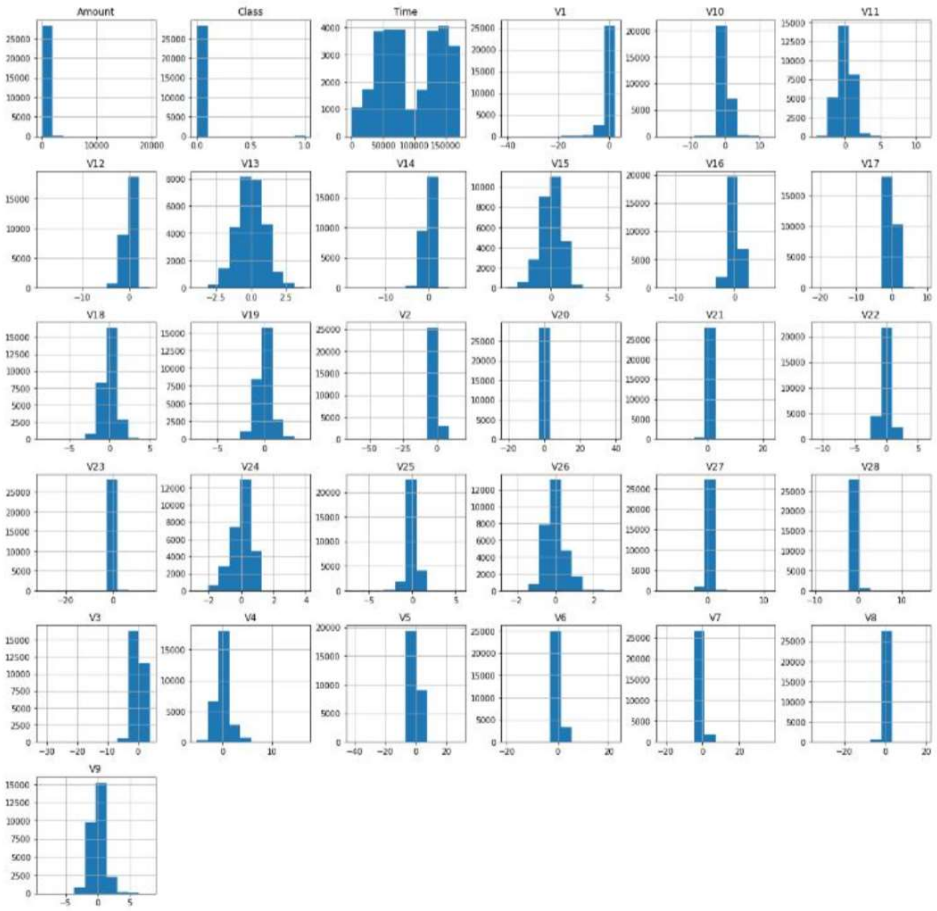
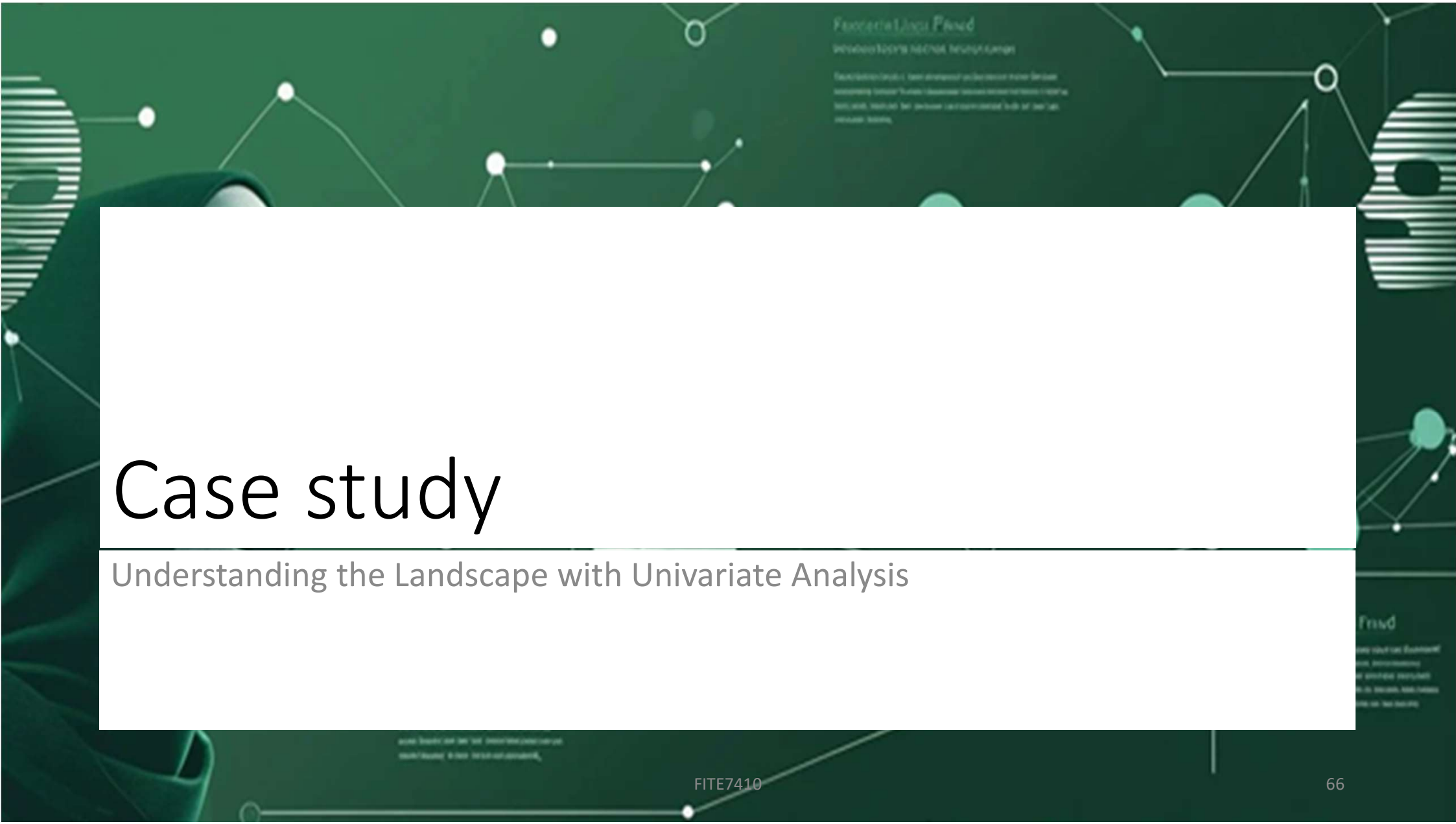


Figure 1: Histogram Showing Dataset Features



Case study

Understanding the Landscape with Univariate Analysis

Case Background

- Background:
  - Credit card division of a large international bank in Brazil
  - Credit card holders who don't want to pay the annual fees may call the bank asking for cancelation or a fee reduction
  - Bank representatives negotiate with the clients about the fees. During the discount negotiation process, bank representatives should follow the bank policy; they cannot offer discounts higher than their authority. And within their jurisdiction, they should also give top priority to the benefit of the bank. In other words, they should offer the lowest discounts acceptable to the clients.

QUESTION: A bank finds that it's losing revenue on credit card renewals because representatives are giving out discounts. Is this just good customer service, or could it be fraud? Let's use our framework to break it down.

Source: Liu, Qi. (2019). An Application of Exploratory Data Analysis in Auditing – Credit Card Retention Case. 10.1108/978-1-78743-085-320191001.

Define scope of Fraud Data Analytics

- What is the type of credit card fraud scenario?
  - Example: Credit card renewal discount offer

- What are the objectives of the fraud data analytics?
  - To identify any non-compliance behaviours of credit card representatives that would case revenue loss of the credit card company

Source: Liu, Qi. (2019). An Application of Exploratory Data Analysis in Auditing – Credit Card Retention Case. 10.1108/978-1-78743-085-320191001.

The dataset

- The account master data is a large dataset with 60,309,524 records and 504 fields.
- Description of 8 selected attributes in this credit card case

Attribute Name (Source Database)	Description
Call Length (Retention)	The duration of each call in seconds
Call Location (Retention)	The location of the customer service center
Agent Number (Retention)	ID of the bank representative answering the call
Supervisor Number (Retention)	ID of the representative's supervisor
Sequential Number (Retention & Account Master)	Sequence Number of an account
Annual Fee (Retention)	Original annual fees of a credit card
Output Annual Fee (Retention)	Actual annual fees paid by clients
Number of Cards (Account Master)	Number of cards associated with each account

Source: Liu, Qi. (2019). An Application of Exploratory Data Analysis in Auditing – Credit Card Retention Case.

Data pre-processing

- Data pre-processing
  - E.g. Data transformation
    - achieved by the logarithm function.
  - E.g. Feature re-engineering: Creation of new attributes – 'Discount'
  - 2 attributes related to 'Discount'
    - Original fee = original annual fee
    - Actual fee = actual annual fee paid

$$Discount = \frac{(Original\ fee - Actual\ fee)}{Original\ fee} \times 100\%$$

Q: What does the following values mean?

Case 1: Discount = 0%

Case 2: Discount = 100%

Case 3: Discount = -ve value

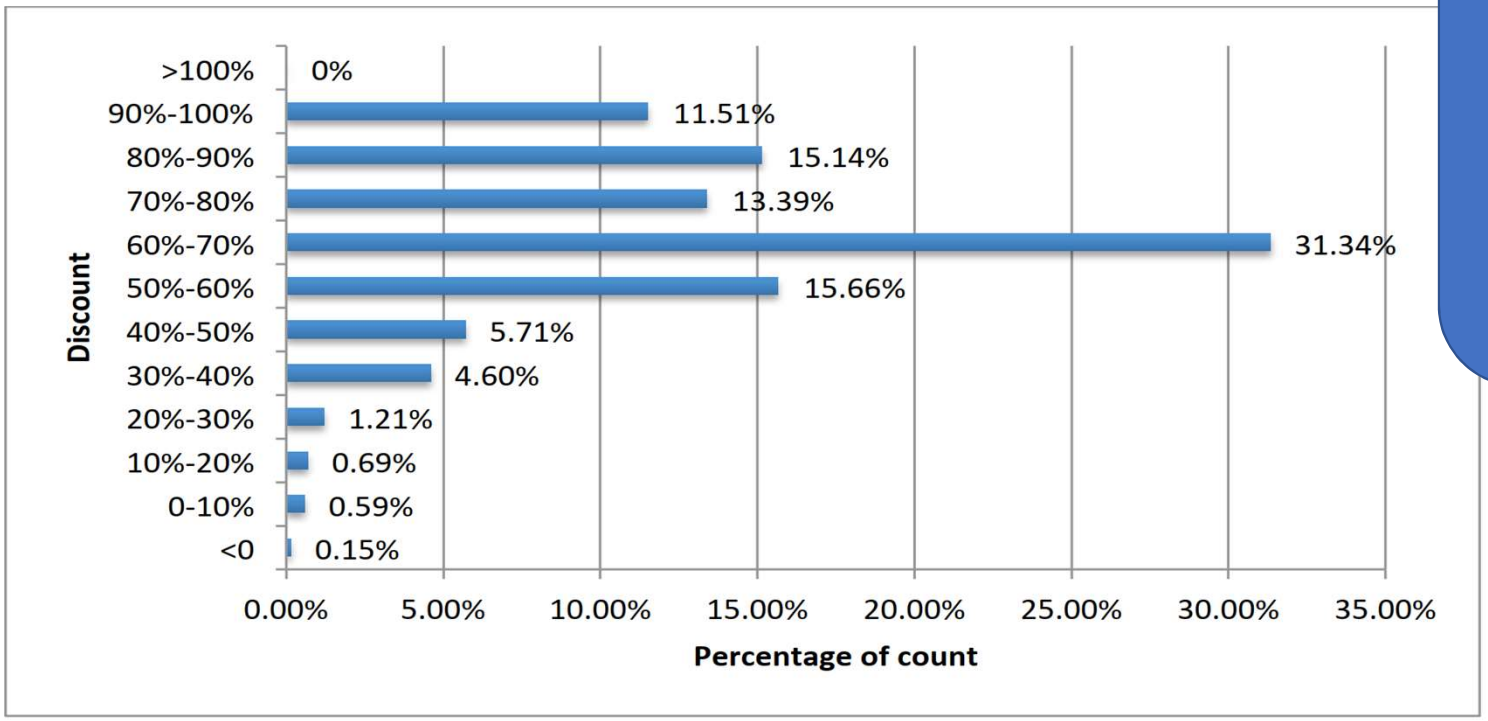
Source: Liu, Qi. (2019). An Application of Exploratory Data Analysis in Auditing – Credit Card Retention Case. 10.1108/978-1-78743-085-320191001.

EDA - Univariant Analysis

- Descriptive statistics used in this study include frequency distribution, summary statistics (mean and standard deviation), and categorical summarization.
- Looking at the following one by one:
  - ‘Discount’ offer patterns and frequencies
  - ‘Discount’ by agent number (i.e. credit card representatives)
  - ‘Discount’ by call lengths
  - ‘Discount’ by call locations
  - ….

EDA - Univariate Analysis : ‘Discount’

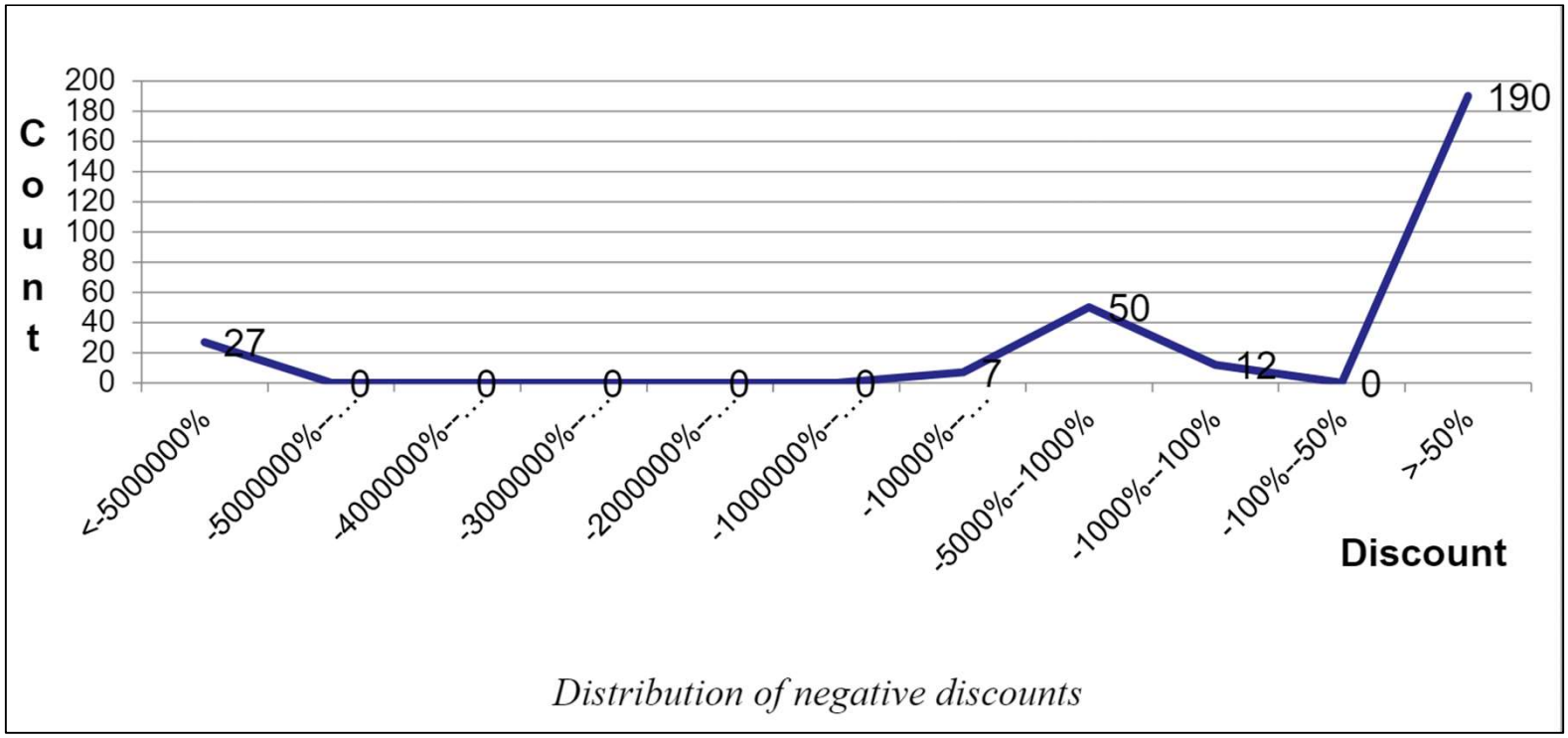
Field Name	Mean	Median	Minimum	Maximum	Standard deviation
Discount	-2.326.04%	60%	-27,944,522.22%	100.00%	219933.88%



Any insights you get from this initial univariate analysis?

EDA - Univariate Analysis : ‘Discount’

Field Name	Mean	Median	Minimum	Maximum	Standard deviation
Discount	-2.326.04%	60%	-27,944,522.22%	100.00%	219933.88%

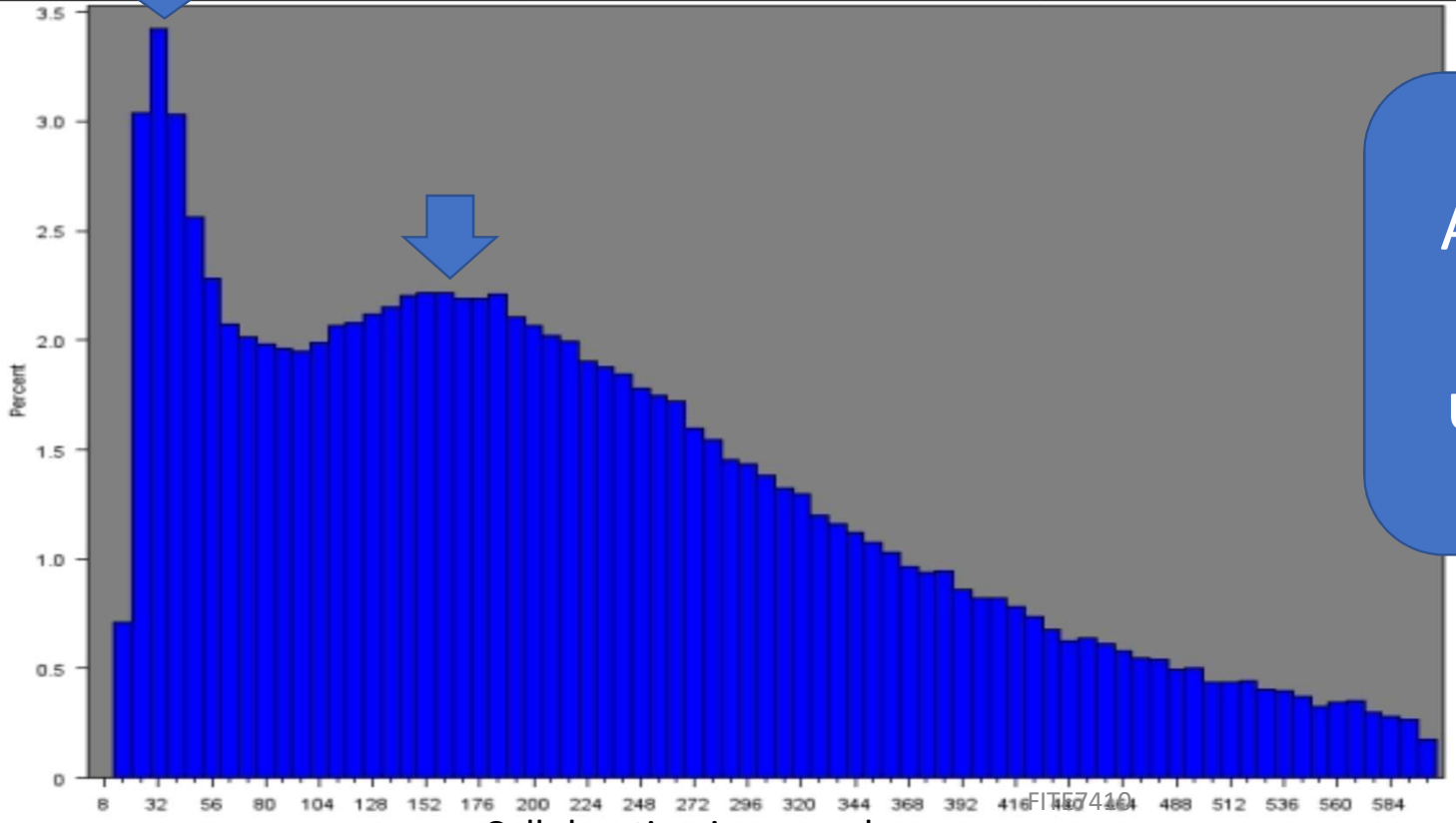


Distribution of negative discounts

Why is –ve discount recorded?

EDA - Univariate Analysis : 'Call duration'

Minimum	Maximum	Mean	Median	90th Percent	Count
10	6561	255	206	514	195694



Call duration in seconds

Any insights you get from this initial univariate analysis?

Key Takeaways

Fraud detection starts with understanding patterns and red flags.

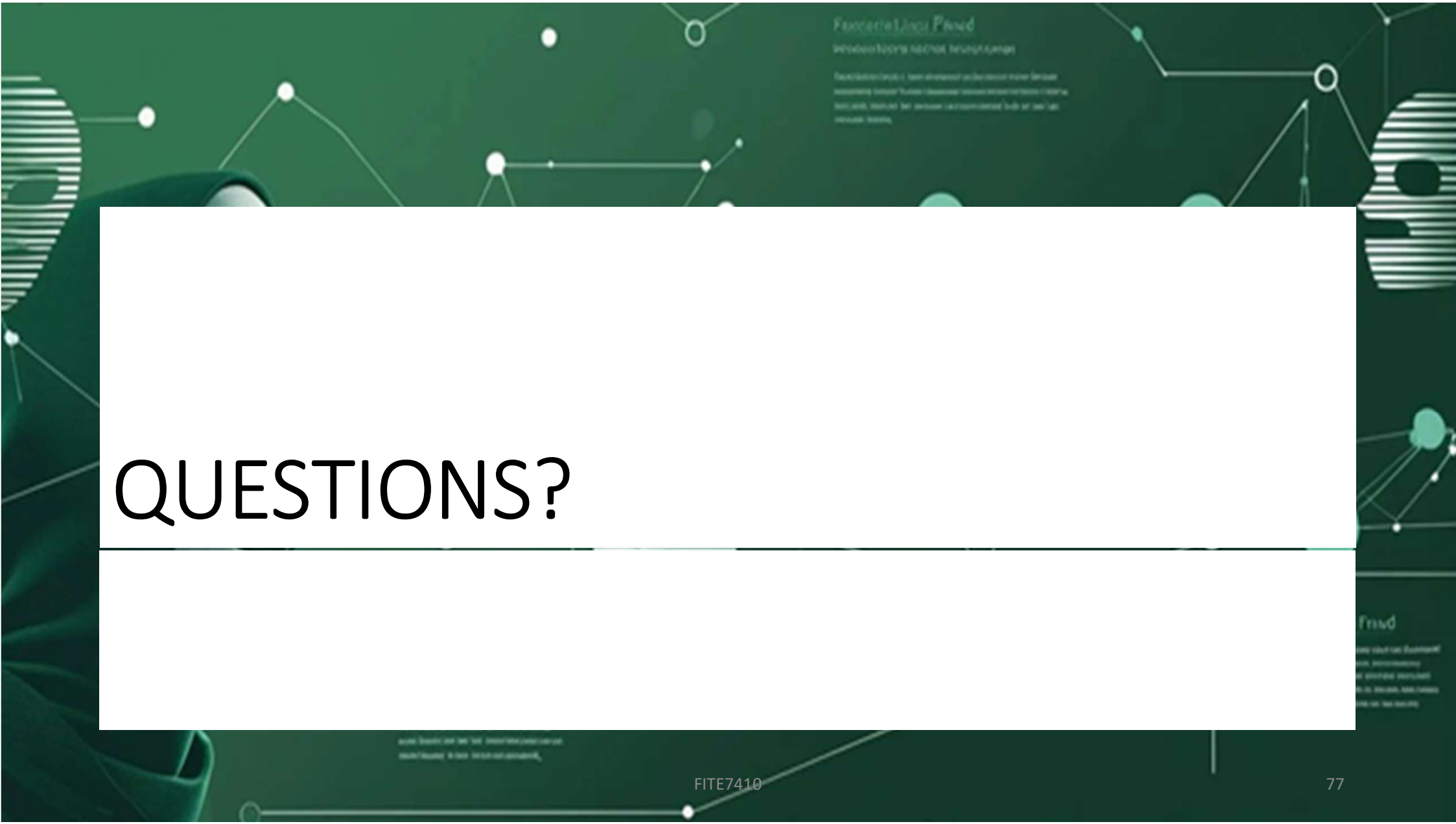
Recognising the typical patterns of fraudulent behaviour is the first step in prevention.

EDA is your first investigative step—ask questions of the data.

Exploratory Data Analysis helps uncover anomalies that warrant further investigation.

References

- Bart Baesens, Veronique Van Vlasselaer, Wouter Verbeke (2015). Fraud Analytics using Descriptive, Predictive, and Social Network Techniques, 1st ed, John Wiley & Sons Inc.
- FICO. (2025). What Are Fraud Analytics? FICO.
- Ghosh et al. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets.
- IBM. (2025). AI fraud detection in banking. IBM. https://www.ibm.com/think/topics/ai-fraud-detection-in-banking
- Kadam, P. (2024). Enhancing Financial Fraud Detection with Human-in-the-Loop Feedback and Feedback Propagation. arXiv. https://doi.org/10.48550/arXiv.2411.05859 Leonard W. Vona (2017). Fraud Data Analytics Methodology: The Fraud Scenario Approach to Uncovering Fraud in Core Business Systems, John Wiley & Sons, Inc.
- Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of the literature. Decision Support Systems, 50(3), 559-569.



QUESTIONS?