INTRO TO DATA SCIENCE

Lecture 1

Introduction to Data Science

Data Science and Engineering Department
Faculty of Informatics
ELTE University

Zakarya Farou

zakaryafarou@inf.elte.hu

SEMESTER SCHEDULE AND GRADING

THE PLAN…

- Lectures
  1. Introduction
  2. Exploratory Data Analysis
  3. Clustering: K-means and DBSCAN
  4. Clustering: Hierarchical clustering
  5. Regression: linear and polynomial regressors
  6. Classification: logistic regression, regularization
  7. **Spring break – no classes**
  8. Frequent Pattern Mining
  9. **Mid-term exam**
  10. Data Types – Time series, Text, Images
  11. Data Pre-processing
  12. Model selection and validation
  13. **Mid-term retake**

- Practicals
  - Python
  - Tutored by PhD students
- Weekly quizzes
- Semester project
  - Group work
  - Select a data type and a dataset.
  - At most 5 students/group, Team leader
  - Presentation at the end of the semester.

QUIZZES, PROJECTS, MIDTERMS, AND FINAL EXAMS

- **For practical:**
  - Presence is mandatory, and you may miss at most 4 classes; if you miss more, you automatically fail the course!
  - Ten quizzes, one or two small questions about the previous class at the beginning of each practical class (5-10 mins max) (**10 x 1 points = 10 points**)
  - Semestrial project. (**30 points**)

- **For lecture:**
  - Presence is not mandatory, but I encourage you to come :)
  - Midterm- exam (**20 points**) : ***you may come to retake to improve your grade!***
  - Final Exam (**40 points**) : you have to score at least 20 points from the practical part and 10 points in the Midterm exam to be eligible to take the exam!

FINAL GRADE

$$S = \sum Q + P + E_M + E_F$$

$$G = \begin{cases} 1, & S < 50 \\ 2, & 50 \leq S < 65 \\ 3, & 65 \leq S < 80 \\ 4, & 80 \leq S < 90 \\ 5, & S \geq 90 \end{cases}$$

BRIEF INTRODUCTION

STEVE DEBENPORT VIA GETTY IMAGES

WHAT IS DATA SCIENCE?

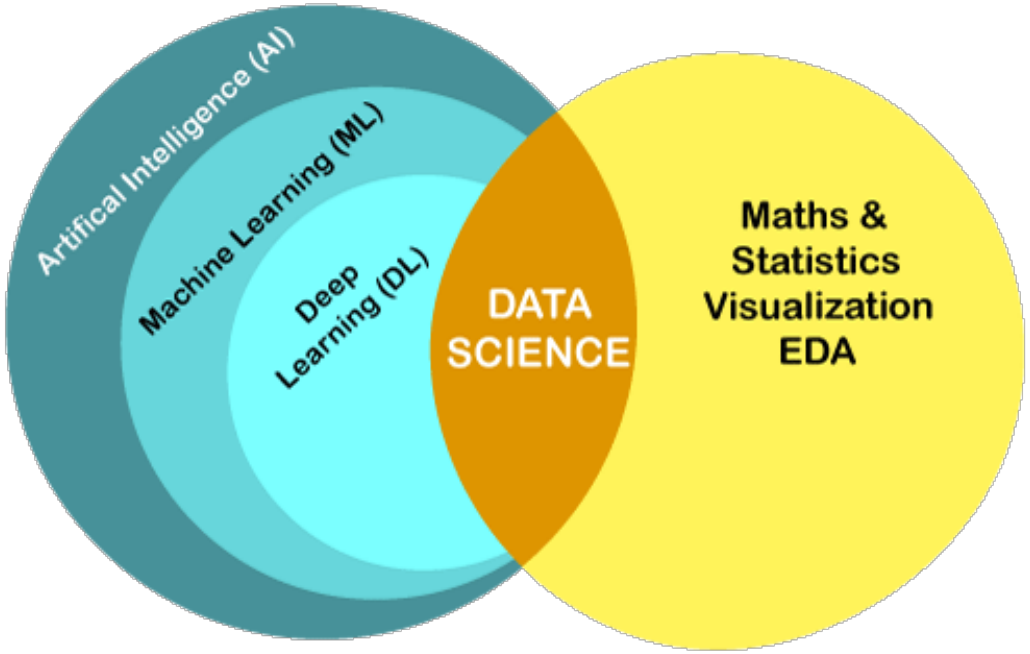
Data science (DS) is the field of study that combines:

- domain expertise,
- programming skills,
- and knowledge of mathematics and statistics.

Objective: **Extract meaningful insights from data.**

Data science practitioners apply machine learning (ML) algorithms to *numbers, text, images, video, audio, and more* to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which *analysts* and *business users* can translate into tangible business value.

DATA SCIENCE AND MACHINE LEARNING



- DS and ML are closely related to each other but have *different functionalities* and *different goals*.
- Indeed, *DS is a field to study the approaches to find insights from the raw data. Whereas ML is a technique used by the group of data scientists to enable the machines to learn automatically from the past data.*

WHERE IS MACHINE LEARNING USED IN DATA SCIENCE?

The use of machine learning in data science can be understood by the development process or life cycle of Data Science. The different steps that occur in Data science lifecycle are as follows:

Business Requirements:

Understand the requirement for the business problem for which we want to use it.

Data Acquisition:

The data is acquired to solve the given problem.

Data Processing:

Acquired raw data is transformed into a suitable format, so that it can be easily used by the further steps.

Data Exploration:

It is a step where we understand the patterns of the data and try to find out the useful insights from the data.

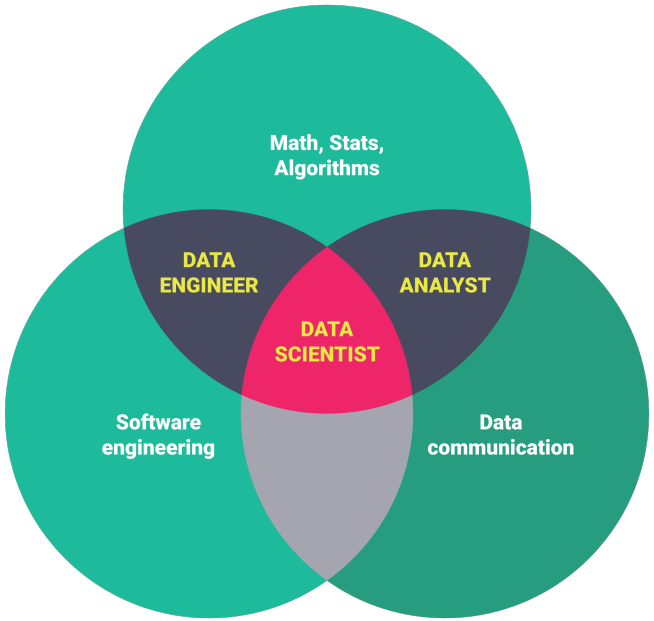
Modeling:

The usage of ML algorithms

Deployment & Optimization:

Deploying the model on an actual project and check its performance.

DATA SCIENTIST VS DATA ENGINEER



- Data science is now one of the most influential topics all around.
- Companies and enterprises are focusing a lot on gathering *data science talent* further creating more viable roles in the data science industry.
- It has also been stated that *data* engineers and *data scientists* are the two most popular career tracks as of now.

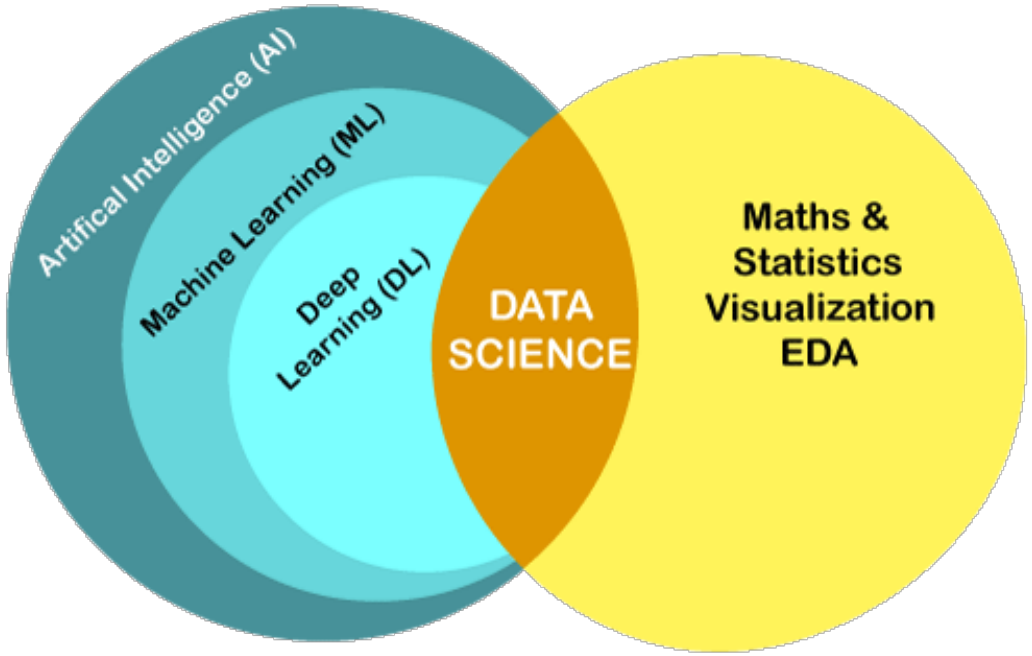
- Since the advent of big data industry, the roles were very blurred since the main objective was to get the insights. But due to a recent change in perspectives, the difference between different data science roles became more clearer than before.

DS & AI & DL

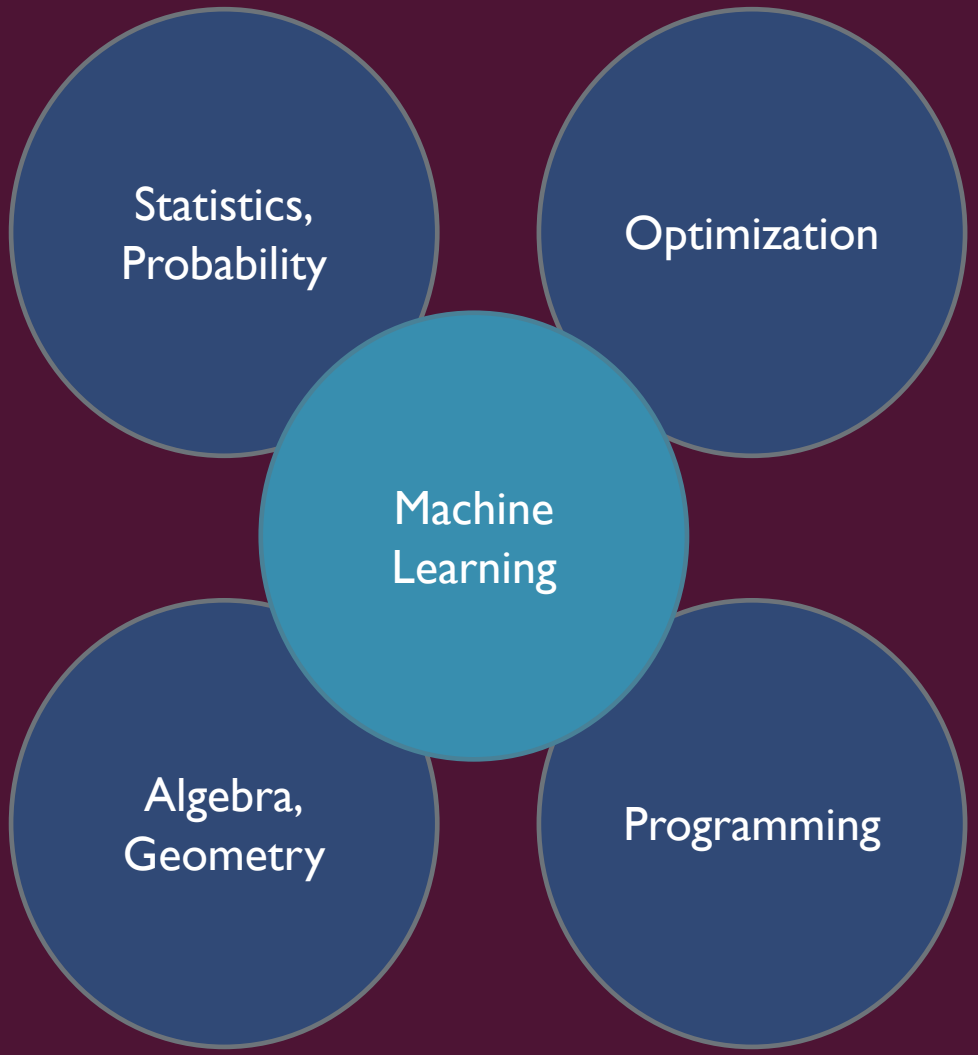
DS and ML are closely related to each other but have *different functionalities* and *different goals*.

DS is a field to study the approaches to find insights from the raw data. Whereas AI's objective is to maximize the chances of success by using a guiding ideology i.e., make machines being conscious just like humans

DL, is a subset of ML which make the computation of multi-layer neural networks feasible

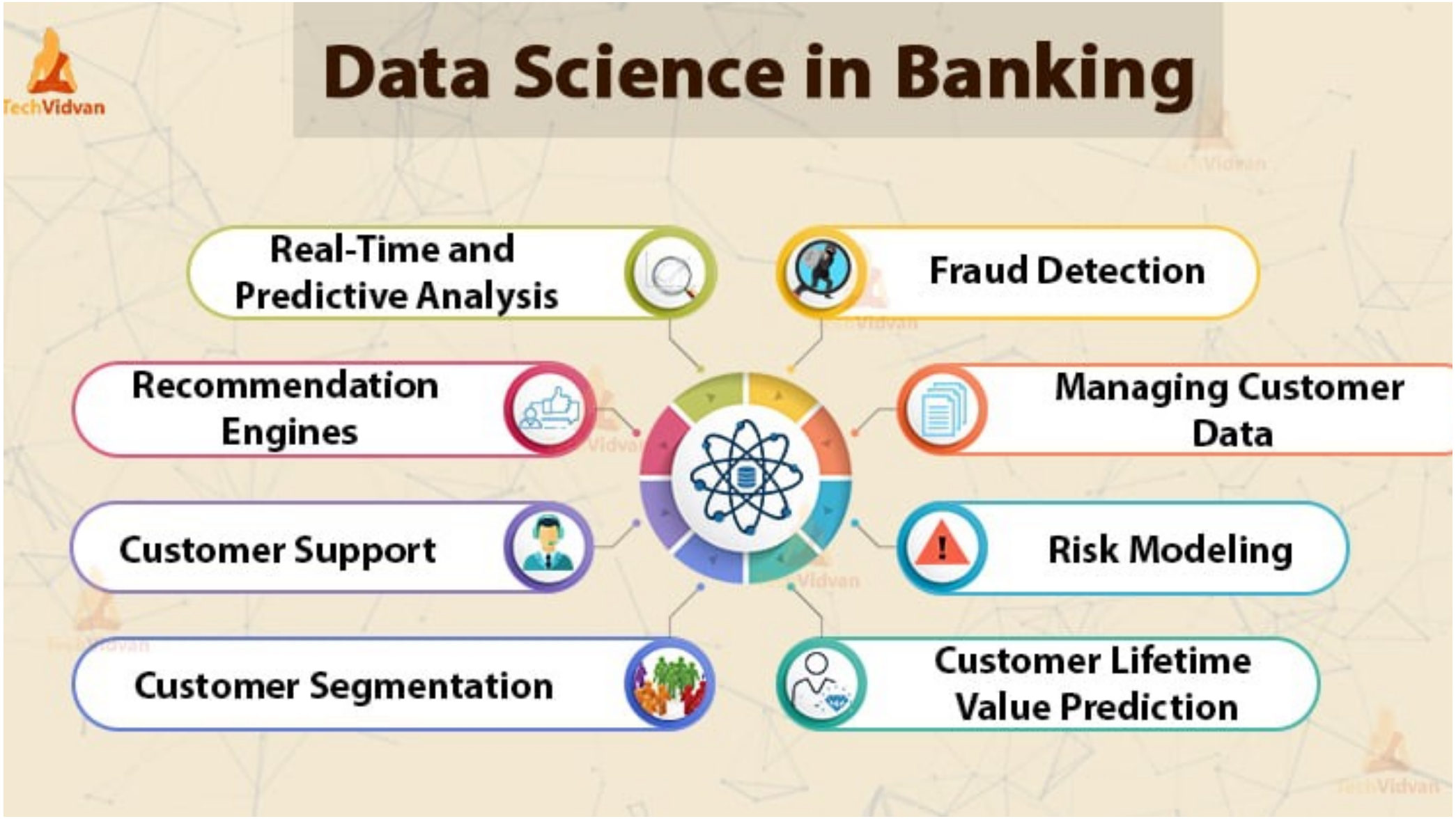


REQUIRED SKILLS TO HAVE



SOME APPLICATIONS

Data Science in Banking

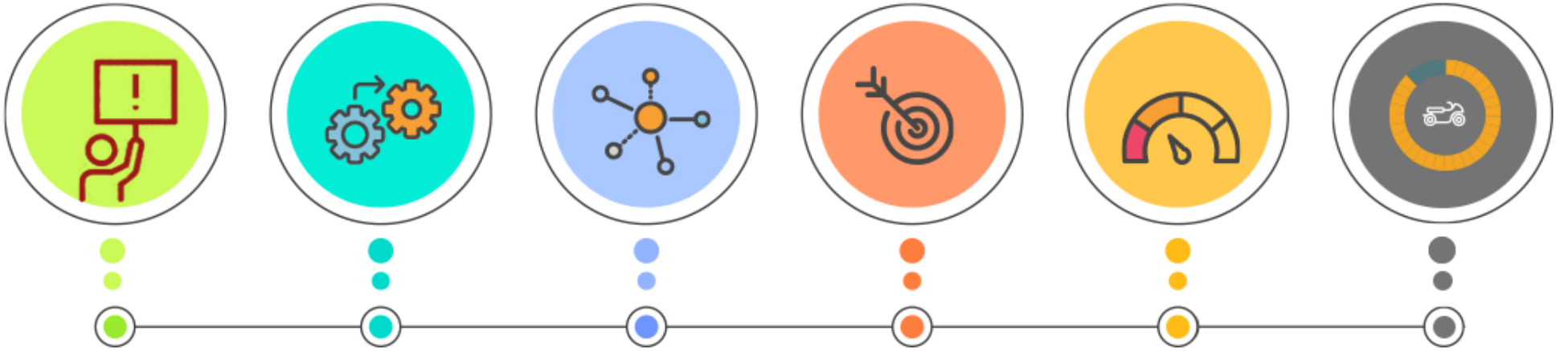


6 Data Science Use Cases in Healthcare

- Data Science for Medical Imaging
- Data Science for Genomics
- Data Science for Drug Discovery
- Predictive Analytics
- Tracking and Preventing Diseases
- Data Science for Wearables

360 DigiTMG™
Digital Transformation | Management | Governance

Applications of **Data Science** in Supply Chain



Demand Analytics

Finished Inventory Optimization

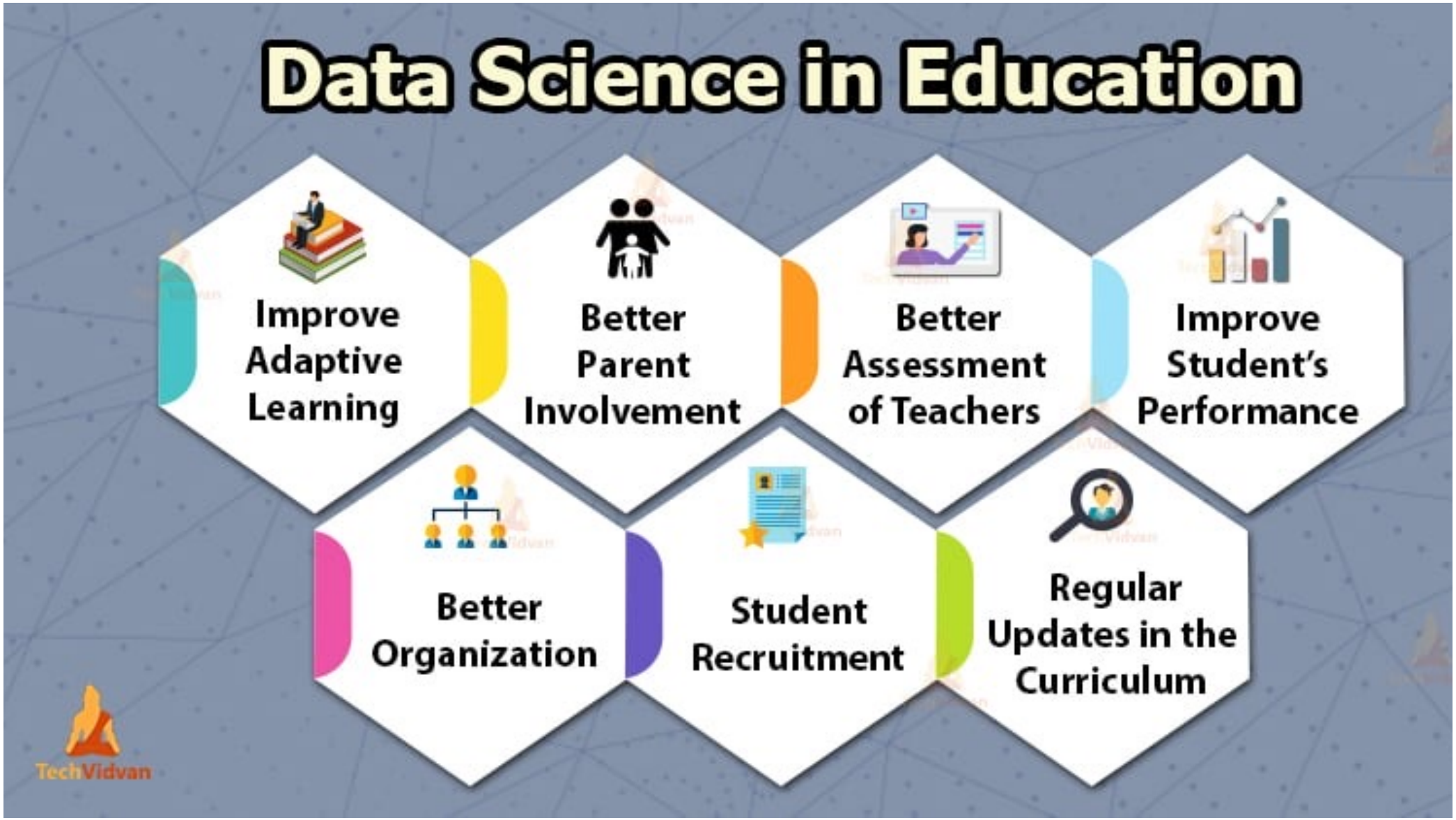
Network Planning

Replenishment Planning Analytics

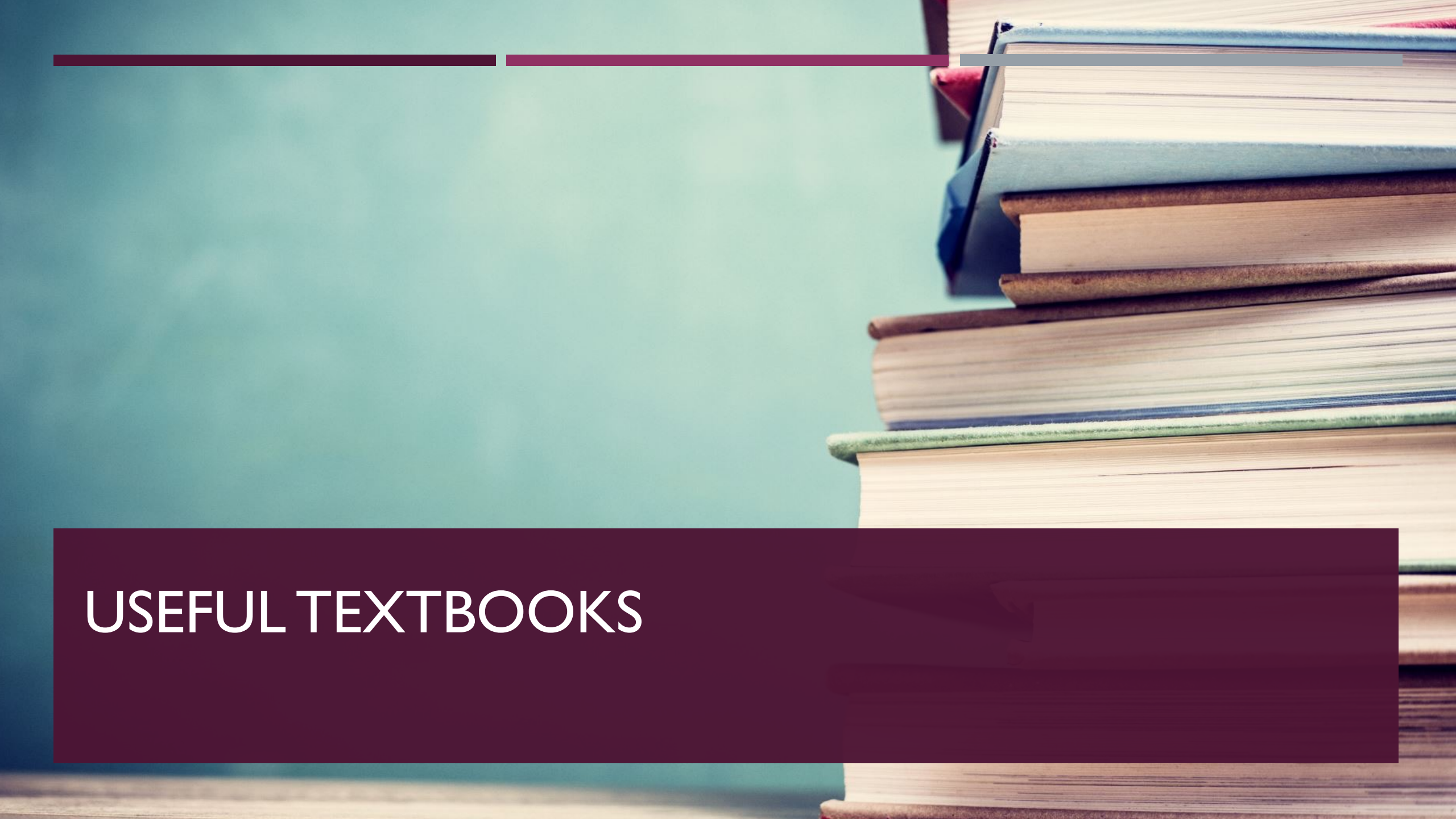
Procurement Analytics

Transportation Analytics

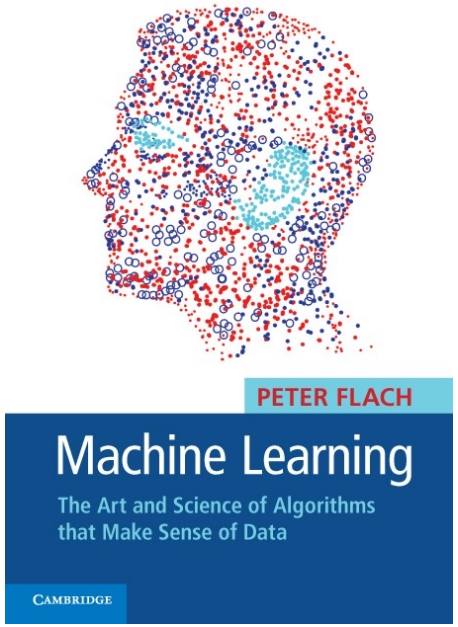
Data Science in Education

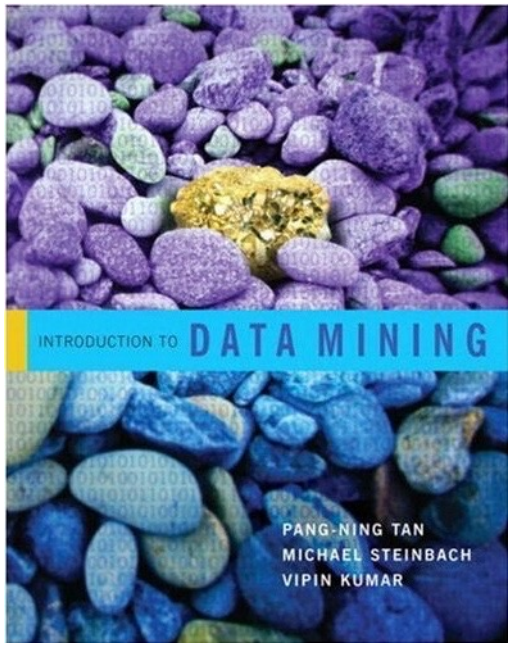


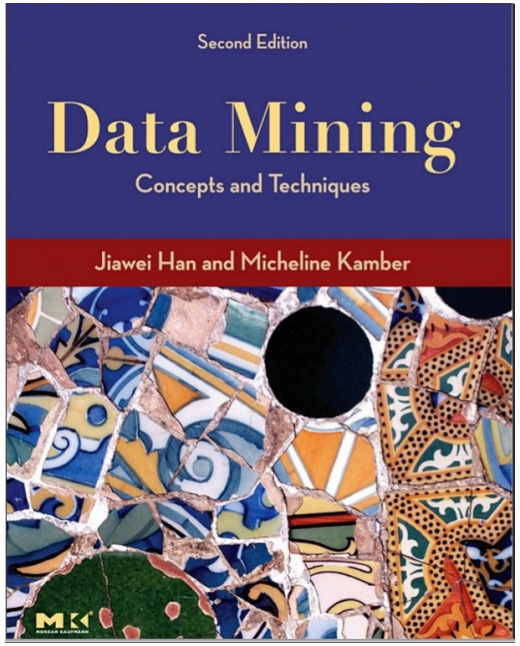

APPLICATIONS OF DATA SECIENCE IN FINANCE
Consumer Analytics
Real-Time Analytics
Customer Data Management
Risk Analytics
Fraud Detection
Providing Personalized Services
Algorithmic Trading



USEFUL TEXTBOOKS

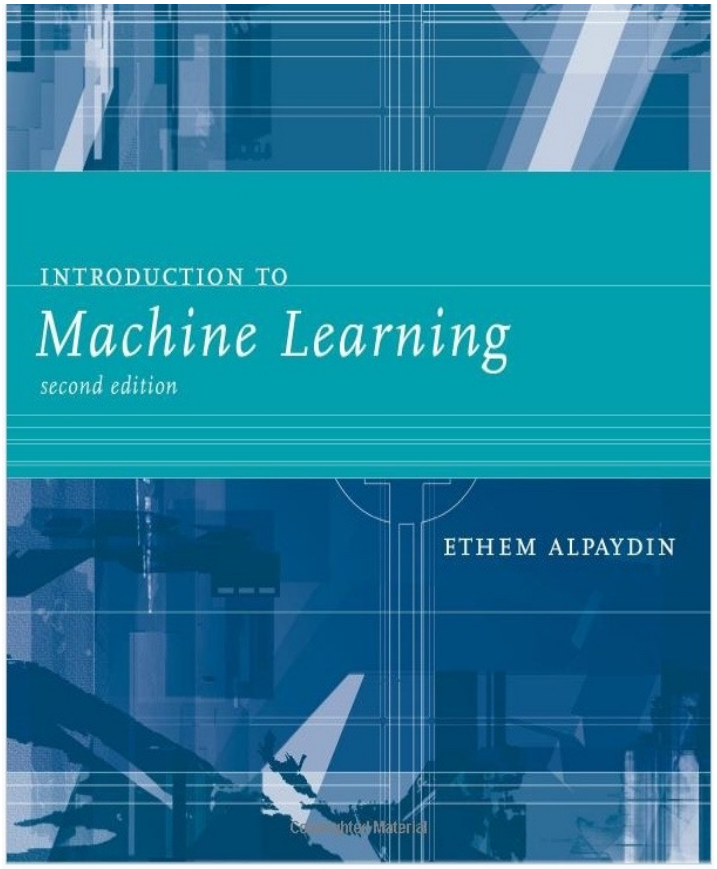

PETER FLACH
Machine Learning
The Art and Science of Algorithms that Make Sense of Data
CAMBRIDGE

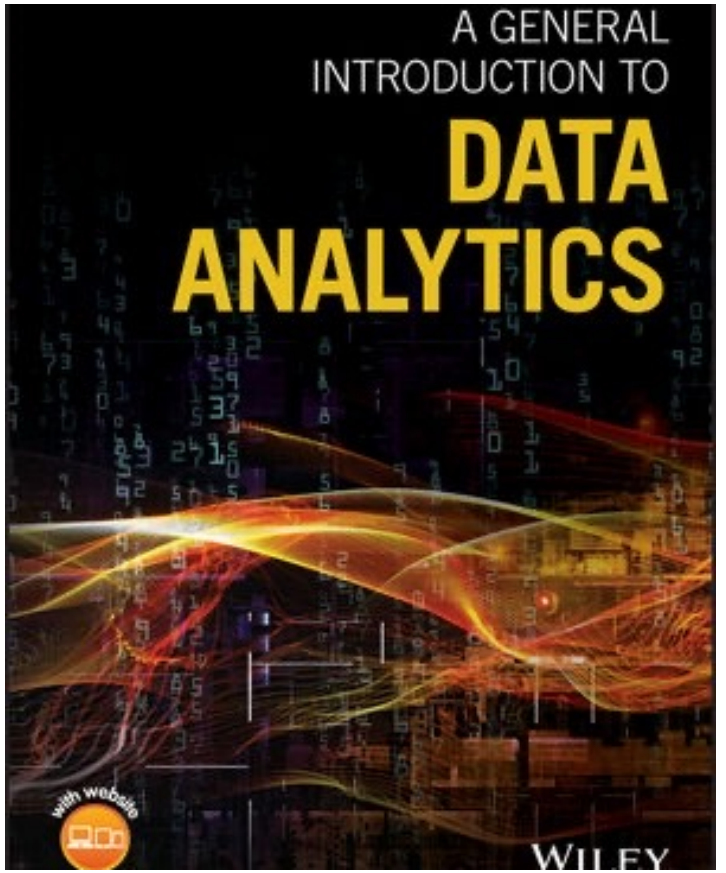

INTRODUCTION TO DATA MINING
PANG-NING TAN
MICHAEL STEINBACH
VIPIN KUMAR


Second Edition
Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber

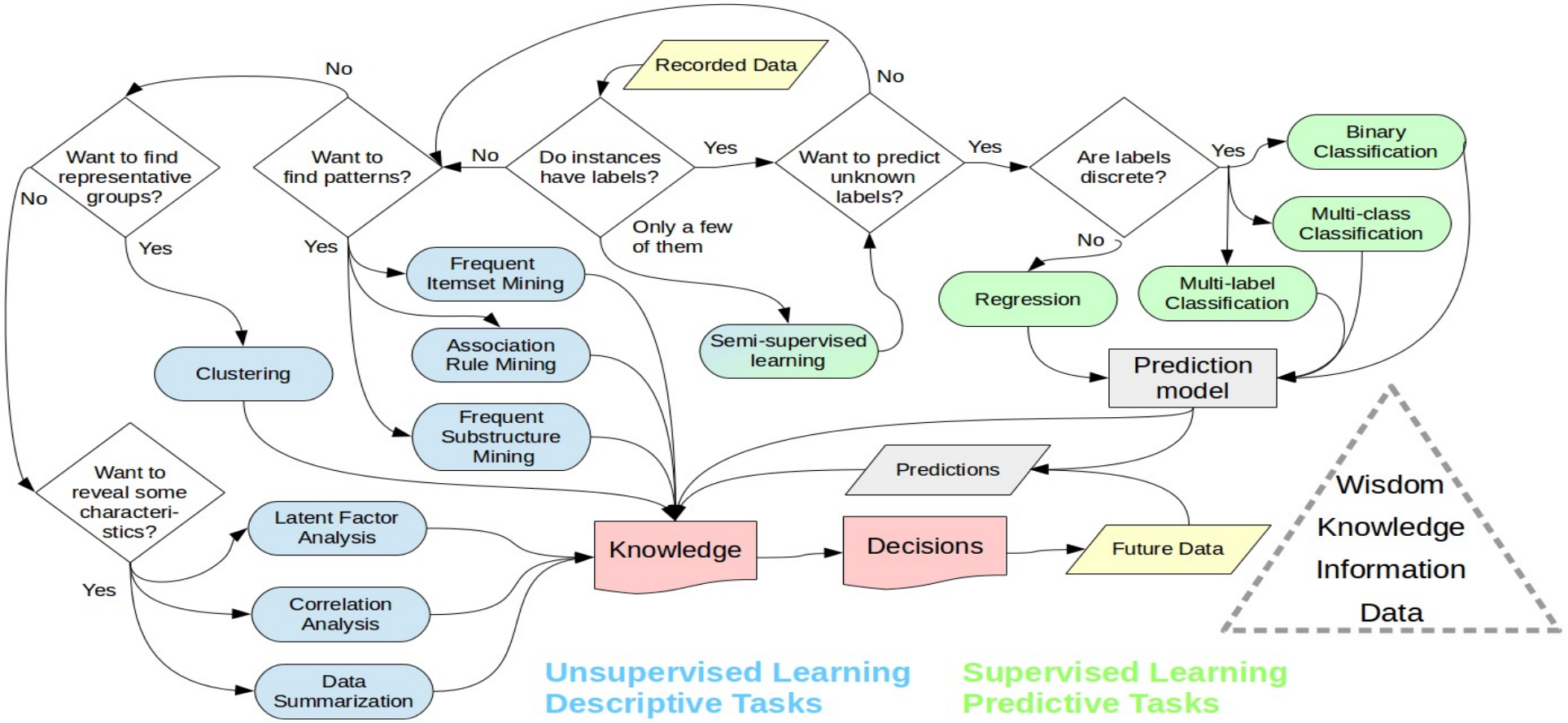
TEXTBOOKS

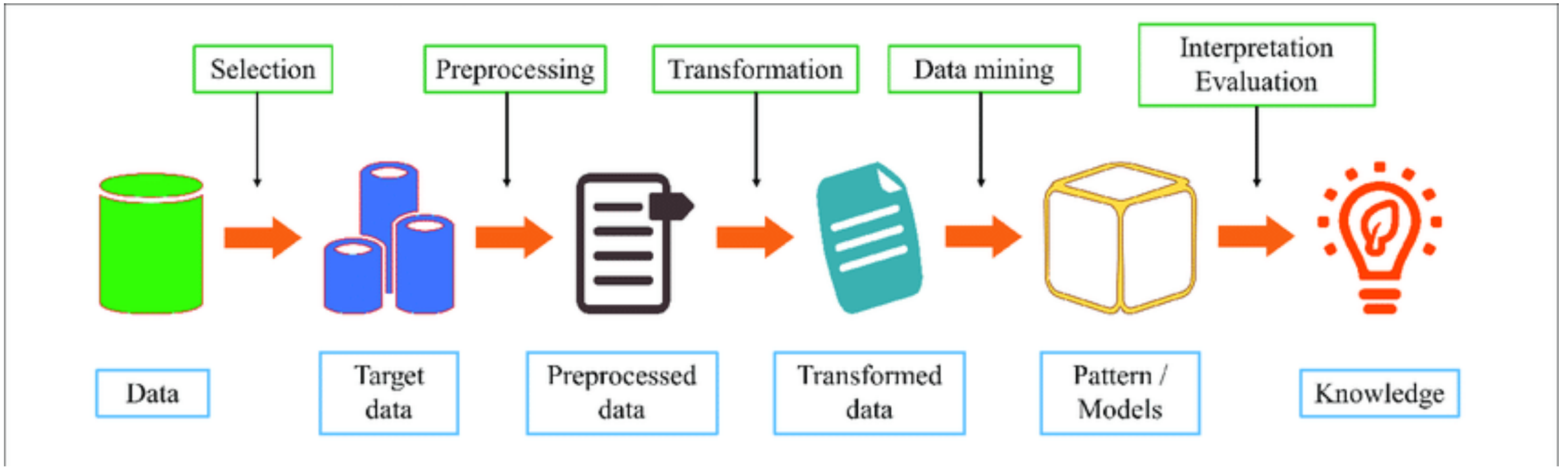
TEXTBOOKS


INTRODUCTION TO
Machine Learning
second edition
ETHEM ALPAYDIN


A GENERAL
INTRODUCTION TO
DATA
ANALYTICS
WILEY

A ROUGH OVERALL PICTURE





A DATA MINING PROJECT

DATA MINING: BIRD'S EYE VIEW

1) Collect data.
2) Data mining!
3) Profit?

Unfortunately, it's often more complicated…

DATA MINING: SOME TYPICAL STEPS

1) Learn about the application.
2) Identify data mining tasks.
3) **Collect data.**
4) Clean and preprocess the data.
5) Transform data or select valuable subsets.
6) Choose a data mining algorithm.
7) **Data mining!**
8) Evaluate, visualize, and interpret results.
9) **Use results for profit or other goals.**

(often, you'll **go through cycles** of the above)

HOW DO YOU APPROACH A DATA SCIENCE TASK?

Business Under-standing → Data Understanding → Data Preparation → Modeling → Evaluation → Deploy-meny

time consuming
(domain experts + data scientists)

boring
(data scientists)

thrilling
(domain experts + data scientists)

DATA MINING: SOME TYPICAL STEPS

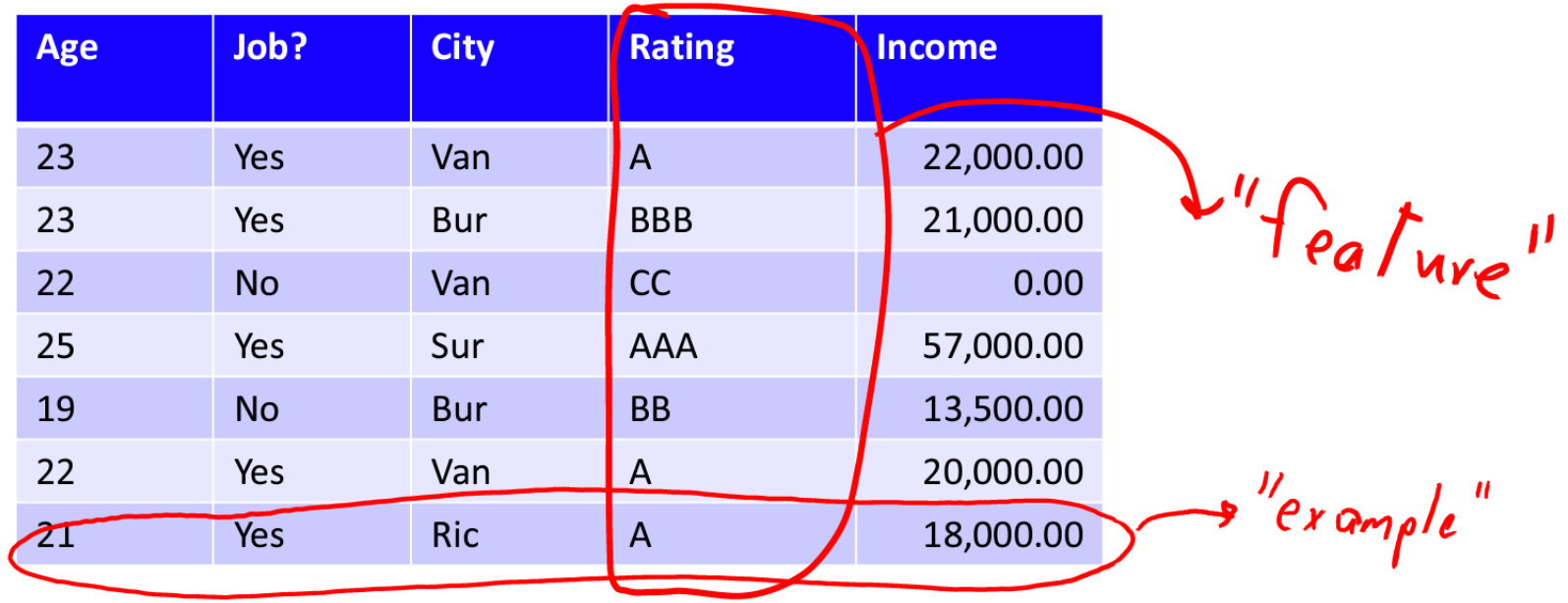
1) Learn about the application.
2) Identify data mining tasks.
3) **Collect data.**
4) **Clean and preprocess the data.**
5) **Transform data or select valuable subsets.**
6) Choose a data mining algorithm.
7) **Data mining!**
8) Evaluate, visualize, and interpret results.
9) **Use results for profit or other goals.**

(often, you'll **go through cycles** of the above)

WHAT IS DATA?

- We'll define data as a collection of examples, and their features.

Age	Job?	City	Rating	Income
23	Yes	Van	A	22,000.00
23	Yes	Bur	BBB	21,000.00
22	No	Van	CC	0.00
25	Yes	Sur	AAA	57,000.00
19	No	Bur	BB	13,500.00
22	Yes	Van	A	20,000.00
21	Yes	Ric	A	18,000.00



- Each row is an "example", each column is a "feature".
  - Examples are also sometimes called "samples", "instances".

TYPES OF DATA

- Categorical features come from an unordered set:
  - ✓ Binary: job? {yes, no} or {1,0}
  - ✓ Nominal: city. {Vancouver, Burnaby, Surrey}
- Numerical features come from ordered sets:
  - ✓ Counts like age in {0, 1, 2, 3,…}
  - ✓ Ordinal like ratings in {best (1), good (2), neutral (3), bad (4), worst (5)}
  - ✓ Continuous/real-valued like height in {173.5, 162.4, 190.2,…}

How could we convert categorical into numerical features?

CONVERTING TO NUMERICAL FEATURES

- Often want a real-valued example representation:

Age	City	Income
23	Van	22,000.00
23	Bur	21,000.00
22	Van	0.00
25	Sur	57,000.00
19	Bur	13,500.00
22	Van	20,000.00

→

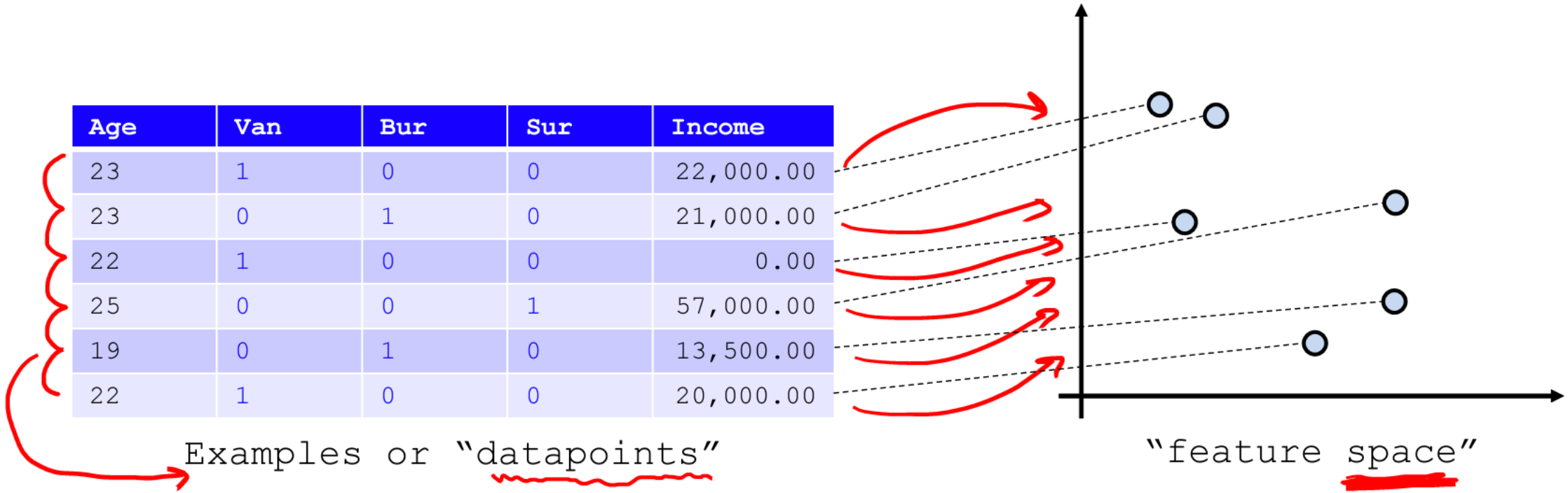
Age	Van	Bur	Sur	Income
23	1	0	0	22,000.00
23	0	1	0	21,000.00
22	1	0	0	0.00
25	0	0	1	57,000.00
19	0	1	0	13,500.00
22	1	0	0	20,000.00

- This is called a "1 of k" encoding (or "one hot" encoding).
- We can now interpret examples as points in space:

  –E.g., first example is at (23,1,0,0,22000).

DATA "SPACE"

Age	Van	Bur	Sur	Income
23	1	0	0	22,000.00
23	0	1	0	21,000.00
22	1	0	0	0.00
25	0	0	1	57,000.00
19	0	1	0	13,500.00
22	1	0	0	20,000.00



- You can compute a "distance" between examples in feature space.

Are these examples close to each other?

APPROXIMATING TEXT WITH NUMERICAL FEATURES

- Bag of words replaces document by word counts:

The **International Conference on Machine Learning** (ICML) is the leading international academic conference in machine learning

ICML	International	Conference	Machine	Learning	Leading	Academic
1	2	2	2	2	1	1

- Ignores order, but often captures general theme.
- You can compute a "distance" between documents.
  - To find similar documents, or decide if two documents are similar.

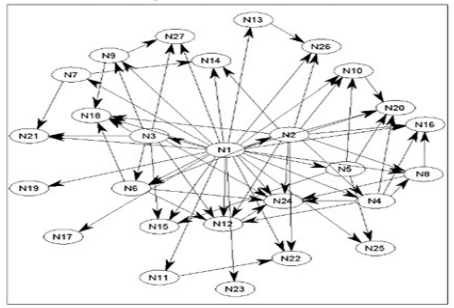
APPROXIMATING IMAGES AND GRAPHS

- We can think of other data types in this way:
  - Images:

graycale intensity →

(1,1)	(2,1)	(3,1)	…	(m,1)	…	(m,n)
45	44	43	…	12	…	35

  - Graphs:



adjacency matrix →

N1	N2	N3	N4	N5	N6	N7
0	1	1	1	1	1	1
0	0	0	1	0	1	0
0	0	0	0	0	1	0
0	0	0	0	0	0	0

DATA CLEANING

- ML typically assumes 'clean' data.
- Ways that data might not be 'clean':
  - Noise
  - Outliers
  - Missing values
  - Duplicated data
- Any of these can lead to problems in analyses.
  - Want to fix these issues, if possible.
  - Some ML methods are robust to these.
  - Often, ML is the best way to detect/fix these.

HOW MUCH DATA DO WE NEED?

- It is a difficult, if not impossible, question to answer.
- My usual answer: "More is better."
  - With the warning: "as long as the quality doesn't suffer ."

- Another popular answer: "Ten times the number of features."

DATA QUALITY

- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

A mistake or a millionaire?

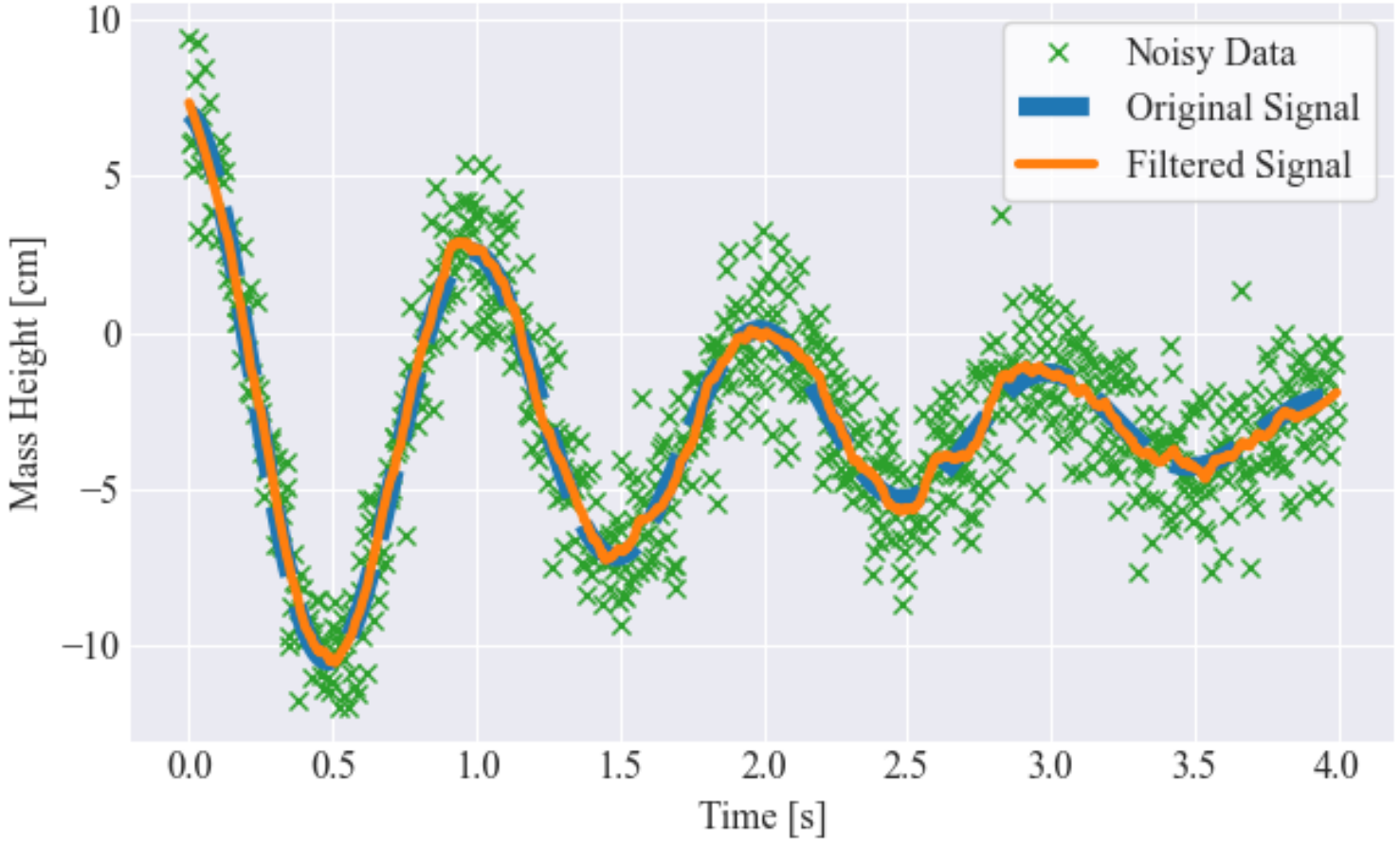
Missing values

Inconsistent duplicate entries

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

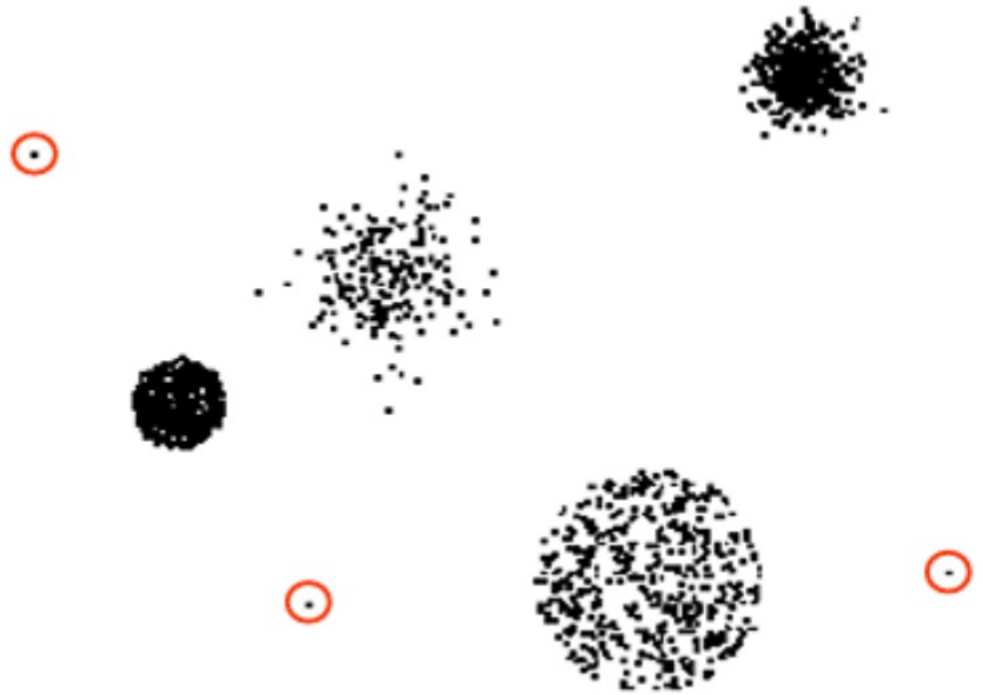
DATA QUALITY: NOISE

- Noise refers to the modification of original values.



DATA QUALITY: OUTLIERS

- Outliers are data objects with characteristics considerably different than most data objects in the data set.



DATA QUALITY: MISSING VALUES

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not apply to all cases (e.g., annual income of children)
- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

DATA QUALITY: DUPLICATE DATA

- Data sets may include data objects that are duplicates or almost duplicates of one another
  - A significant issue when merging data from heterogeneous sources.
- Examples:
  - The same person with a different ID.

id	first_name	last_name	email
1	Carine	Schmitt	carine.schmitt@verizon.net
4	Janine	Labrune	janine.labrune@aol.com
6	Janine	Labrune	janine.labrune@aol.com
2	Jean	King	jean.king@me.com
12	Jean	King	jean.king@me.com
5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
10	Julie	Murphy	julie.murphy@yahoo.com
11	Kwai	Lee	kwai.lee@google.com
3	Peter	Ferguson	peter.ferguson@google.com
9	Roland	Keitel	roland.keitel@yahoo.com
14	Roland	Keitel	roland.keitel@yahoo.com
7	Susan	Nelson	susan.nelson@comcast.net
13	Susan	Nelson	susan.nelson@comcast.net
8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

Thank you for your attention

QUESTIONS?