DATA SCIENCE FOR NON-PROGRAMMER

PUSILKOM UI

November 2021

Course Outline

- Introduction to Data Science
- Managing a Data Science Project
- Classification and Hands on
- Exploratory Data Analysis
- Clustering
- Data Preprocessing Methods
- Regression
- Case Study

INTRODUCTION TO DATA SCIENCE

Denny, Ph.D.

PUSILKOM UI

November 2021

a.k.a. data analytics, data mining, Knowledge Discovery in Databases (KDD), knowledge extraction, business intelligence

fenomena
proses

menghasilkan
data

data

belajar
dari data

memahami

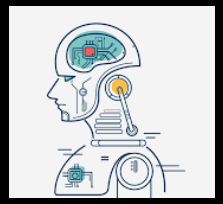
membuat
keputusan

fenomena
proses

menghasilkan
data

data

belajar
dari data



machine

model

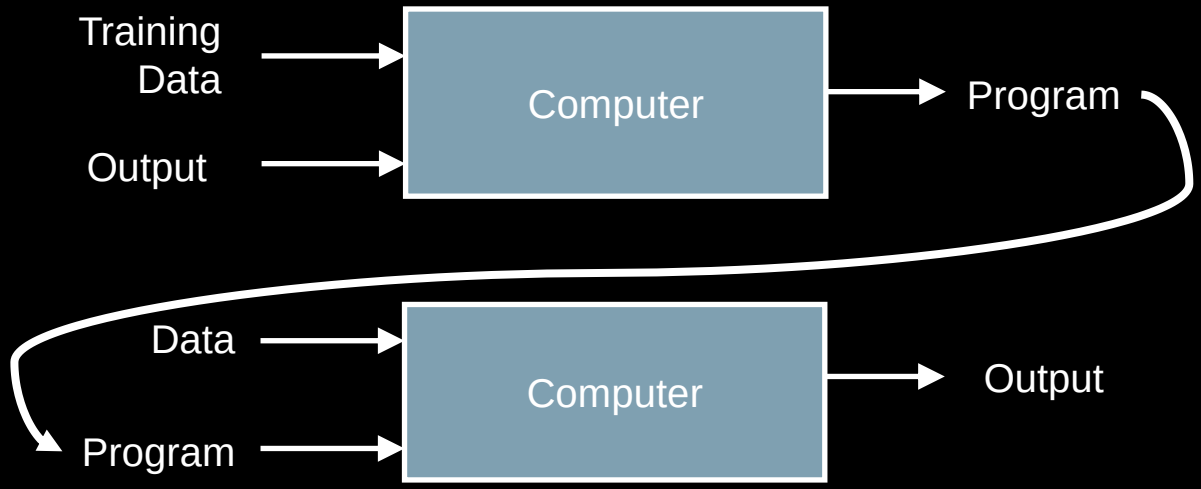
memahami

Mengapa Menggunakan Machine Learning?

- **mengapa kita tidak membuat program dengan aturan yang kita buat?**
- **data terlalu besar dan kompleks**
  - data astronomi, arsip medis, data genetic, transaksi keuangan, social media, transaksi e-commerce
- **kemampuan adaptif**
  - program yang dibuat, tidak adaptif jika ada perubahan kondisi

Traditional Programming

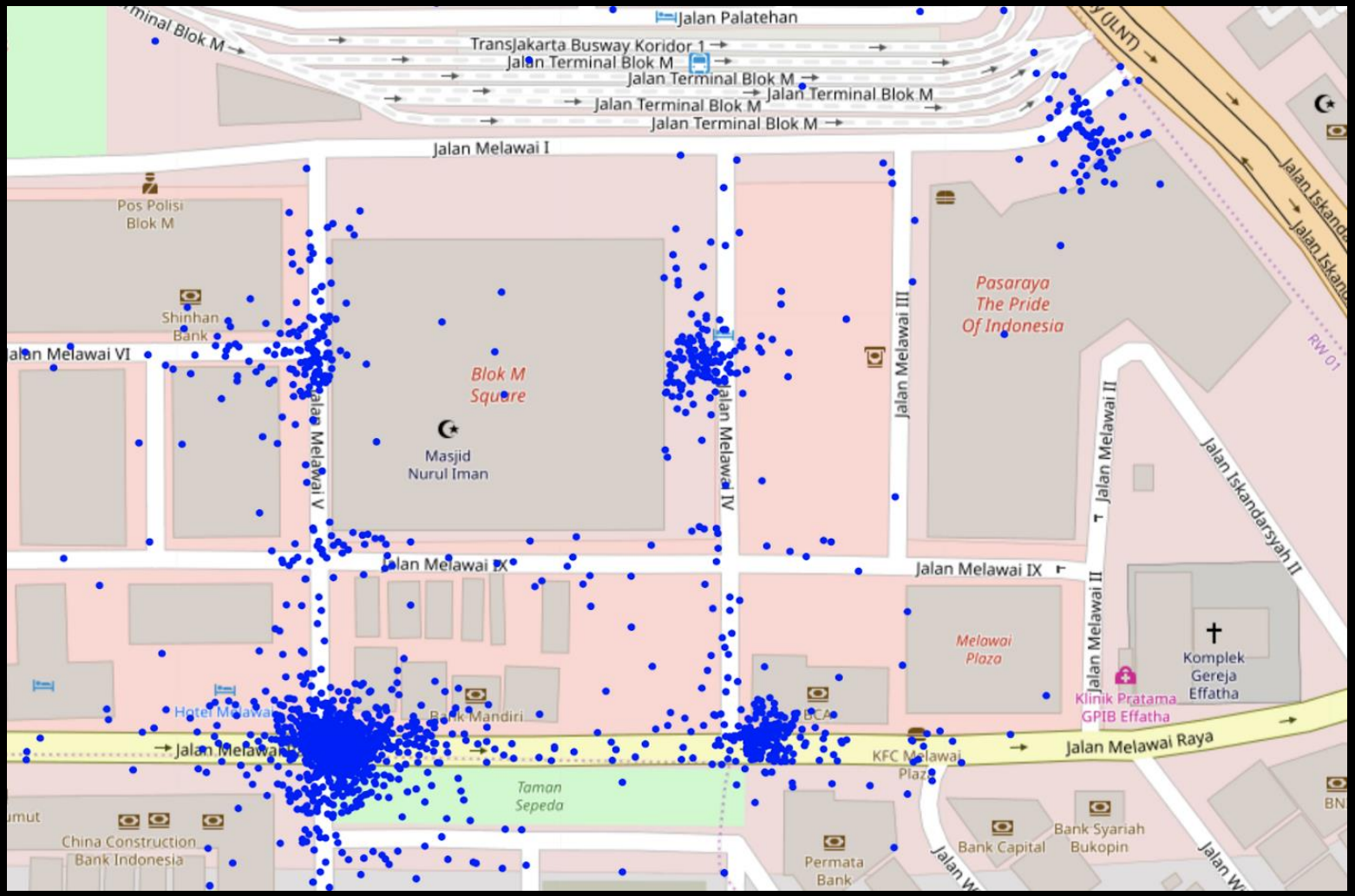
Machine Learning



MOTIVATION

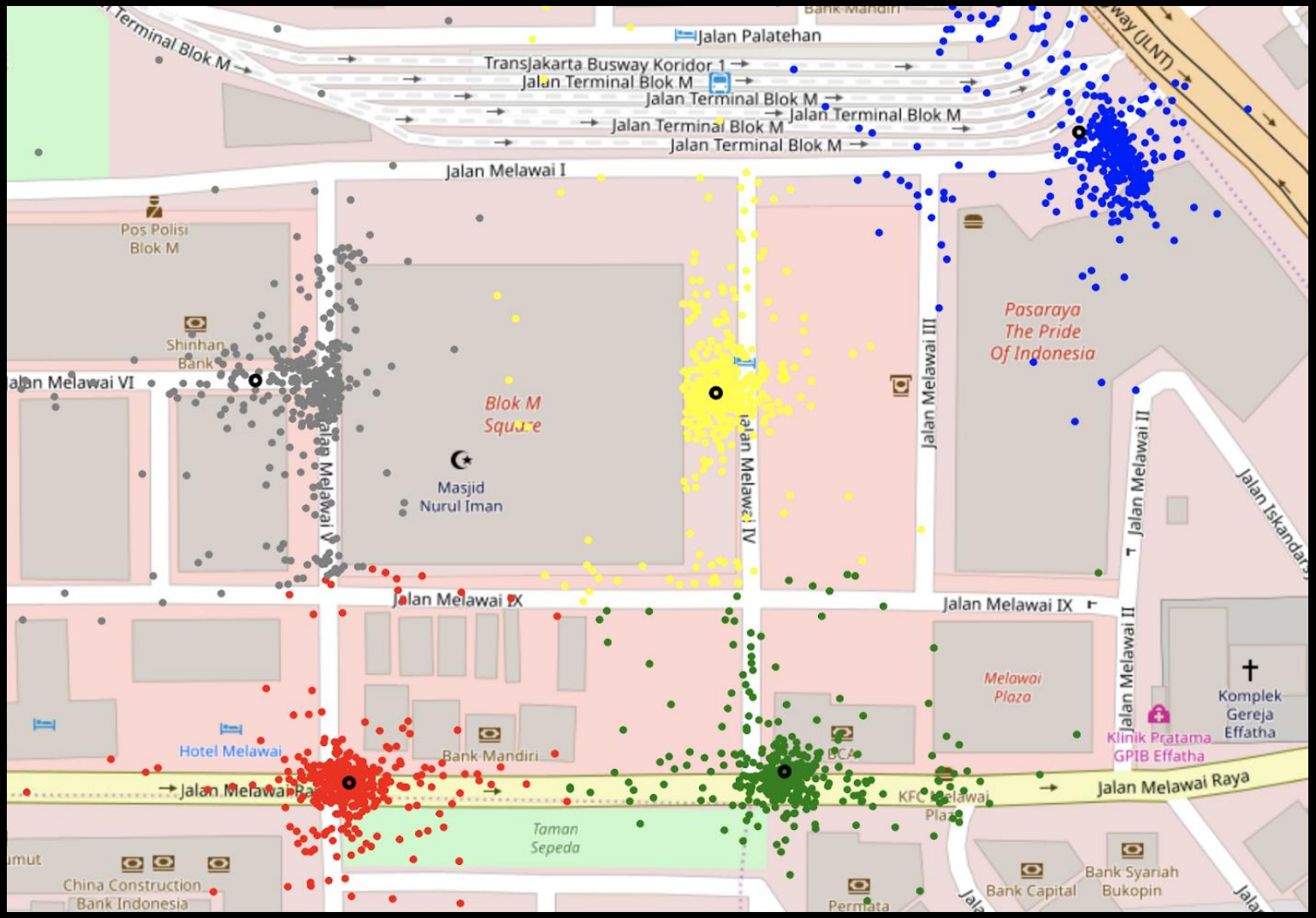
Why Data Science?

Gojek: Pickup Points

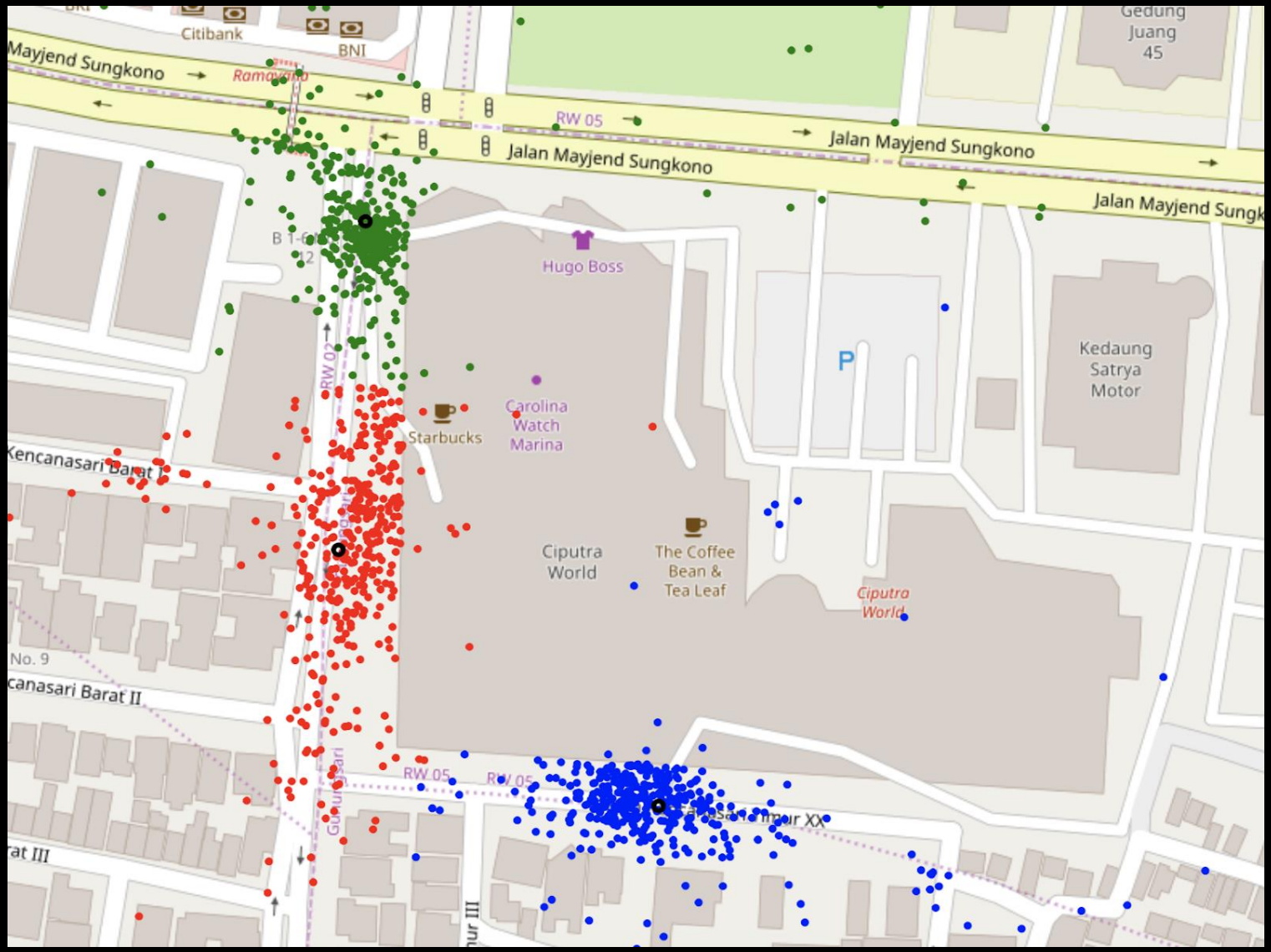


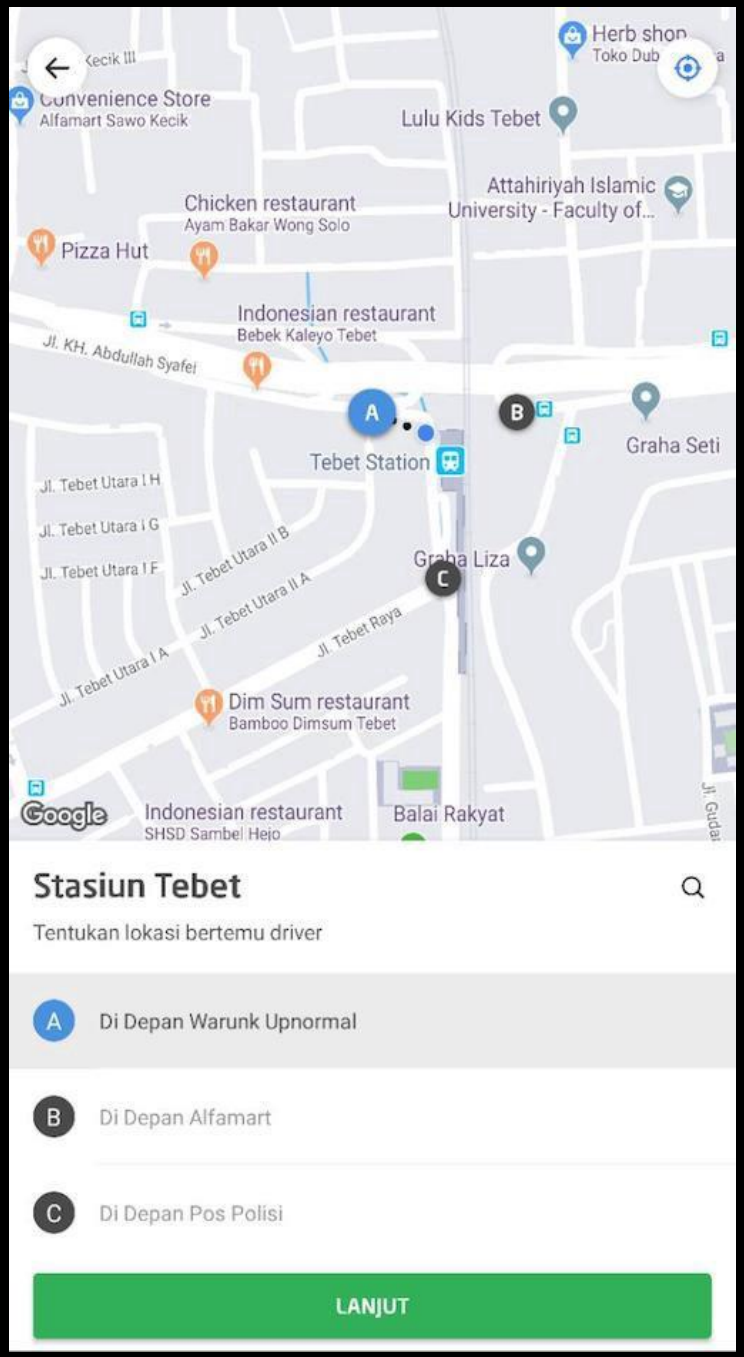
https://www.gojek.io/blog/fantastic-drivers-and-how-to-find-them

Gojek: Clustering Pickup Points



Gojek: Clustering Pickup Points





Naming Frequent Place of Interest?

Automatic naming from booking-text data

Case Study: GOJEK

- **Value:**
  - provide the best experience to the customers
- **Objective:**
  - Help customer to meet his/her GOJEK driver — without a single call

- **The use of non-transactional data**
  - GPS
  - booking text / chat

Android Keyboard

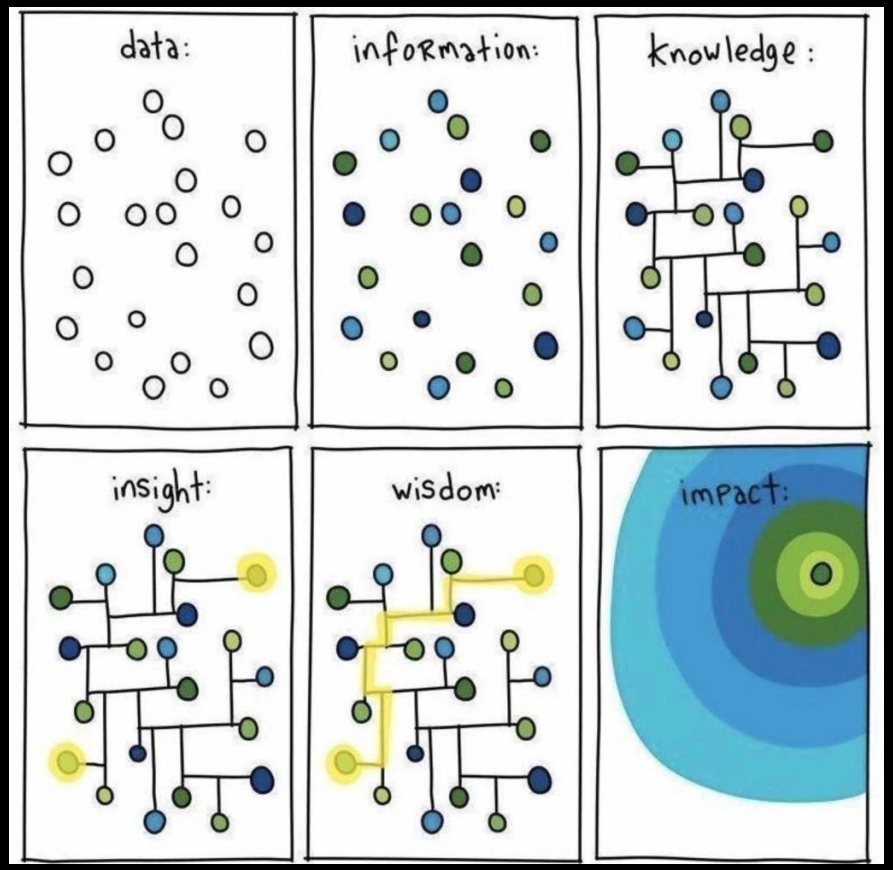
Data Science

methods to capture, process, and organize data to uncover actionable insights for current problems, and establishing the best way to present this data

translate data into a **story** to generate **insights**

then, make **decisions** for a company or institution

producing results that can lead to **immediate improvements**



Data Science Value

- **help companies improve operations and make faster, more intelligent decisions.**
- **can help a company to gain useful insight to increase revenues, get or retain customers, and improve operations.**
- **making decision based on knowledge / patterns / insights discovered from data**
  - **actionable knowledge** that creates values

Business Value of Data Science

- **Examples:**
  - fraud detection to reduce company losses
    - Q: action/decision?
  - identifying offshore tax evaders to increase tax revenue (ATO)
  - identifying items bought together and creating package to increase sales
  - Customer segmentation: personalize customer relationships for higher satisfaction and retention
  - Customer attrition analysis: prevent loss of high-value customer and perhaps let go of lower-value customers

Business Value of Data Science: ATO

- **Fraud detection**
  - Identify High Risk Refund
    - Previous practice simple business rules based on experience:
      - Total claimed investment deductions > $N
      - Ratio of self education deductions to total income > N
      - Total international transfers > N times taxable income
      - Luxury vehicle purchase $M > N times taxable income
    - Use modelling
      - regression, decision trees, random forests
      - increase Tax revenue

- **Identify Aggressive Tax Planning**
- **Assessing levels of debt: propensity dan capacity to pay**

Analytics

- **Descriptive**
  - the tasks on providing insight into the past, and then answer ”what was happened?”
- **Inferential/Diagnostic**
  - Determining the cause of phenomenon that occured in the past: ”what reason behind the event?”
- **Predictive**
  - determine the outcomes of an event that might occur in the future the tasks on providing the (statistical) model: ”what will happen?”
- **Prescriptive**
  - focus not only on prescribing the best option to follow, but also: ”how do you make the best happen?”

Business Intelligence

DATA SCIENCE / DATA MINING

a.k.a. analytics, Knowledge Discovery in Databases (KDD), knowledge extraction, business intelligence

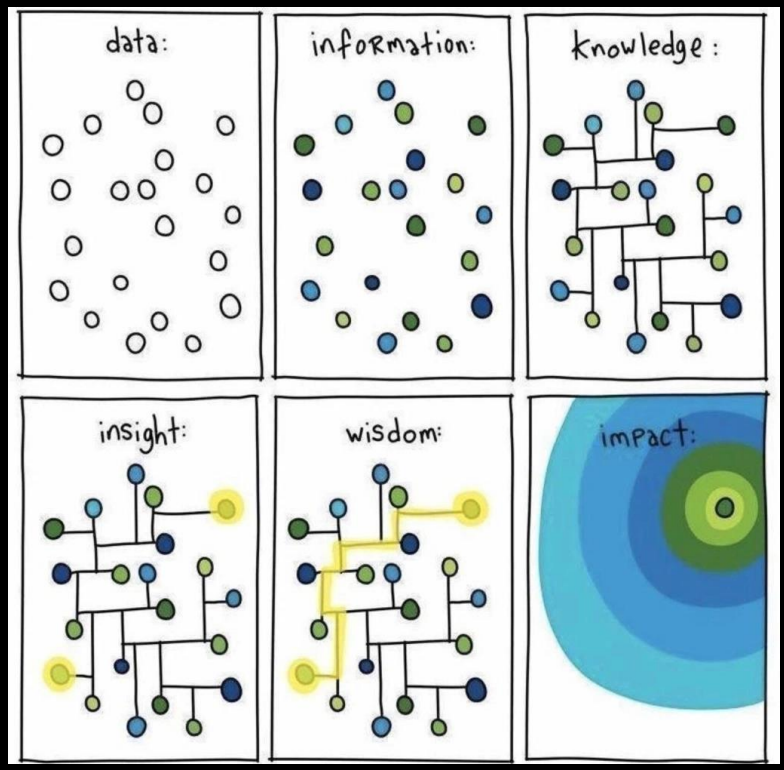
Data Science

methods to capture, process, and organize data to uncover actionable insights for current problems, and establishing the best way to present this data

translate data into a **story** to generate **insights**

then, make **decisions** for a company or institution

producing results that can lead to **immediate improvements**



Data Science, Data Mining, Analytics

the analysis of data sets to find **unsuspected relationships** and to summarize the data in novel ways that are both **understandable** and **useful** to the data owner

Approach

How can we make it happen?

What will happen?

Why did it happen?

What happened?

Prescriptive Analytics

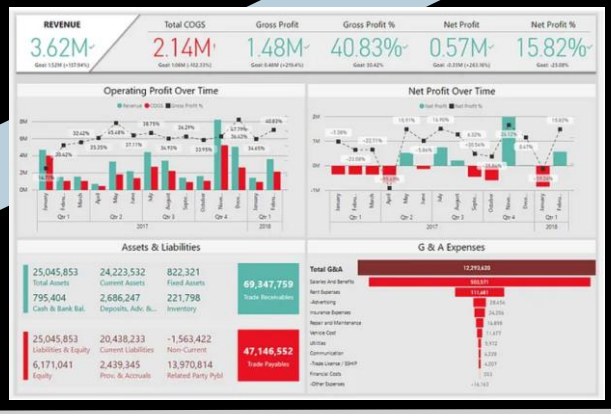
Predictive Analytics

Diagnostic Analytics

Descriptive Analytics

Optimization

Information



VALUE

DIFFICULTY

Analytics Approach

Descriptive

the tasks on providing insight into the past, and then answer ”what was happened?”

Inferential/Diagnostic

Determining the cause of phenomenon that occurred in the past: ”what reason behind the event?”

Business Intelligence

Predictive

determine the outcomes of an event that might occur in the future the tasks on providing the (statistical) model: ”what will happen?”

predict customer churn

Prescriptive

focus not only on prescribing the best option to follow, but also: ”how do you make the best happen?”

offer lower premium/other incentives for high value customer with low risk

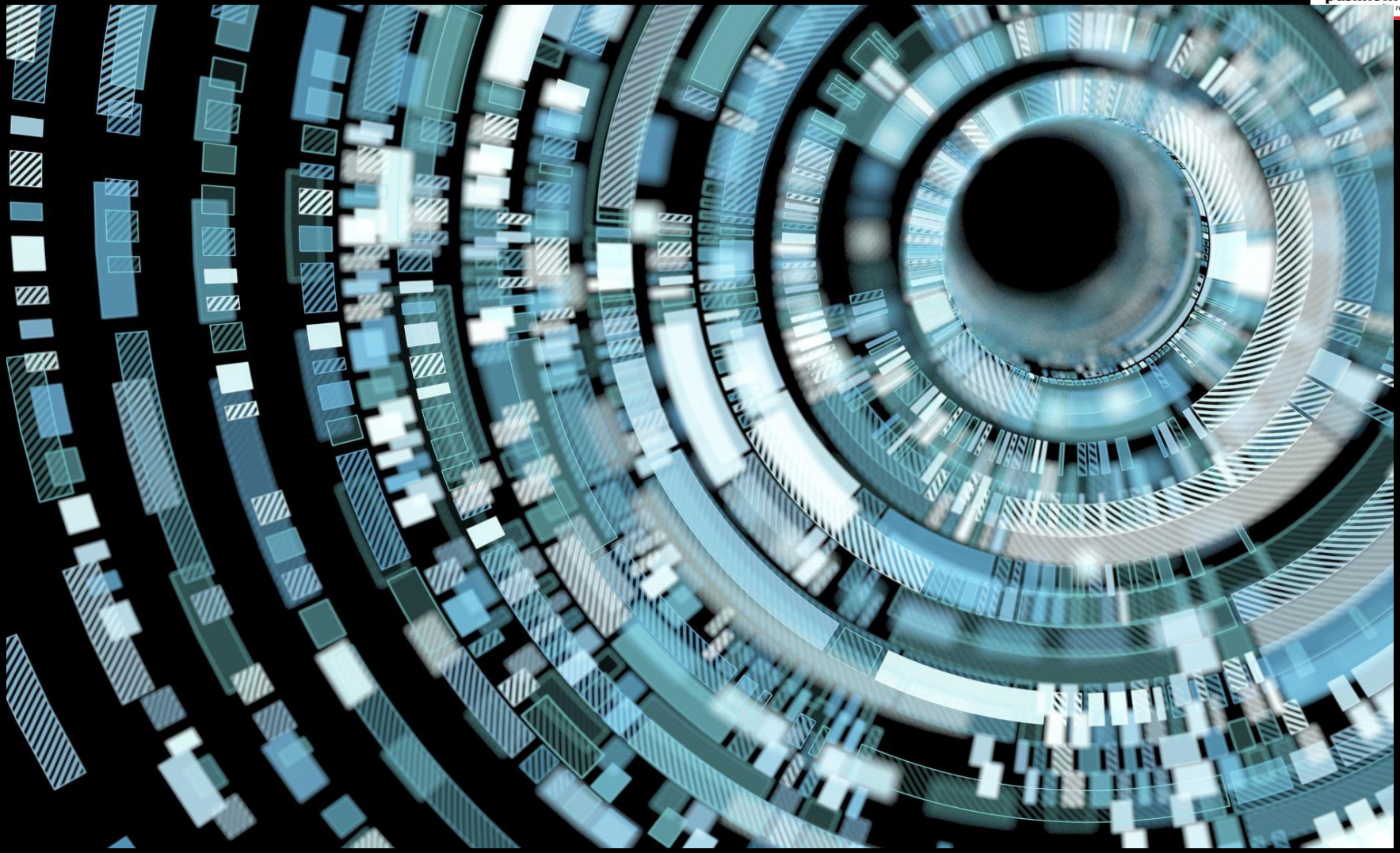
Analytics Approach

- **descriptive, diagnostic, predictive, prescriptive?**

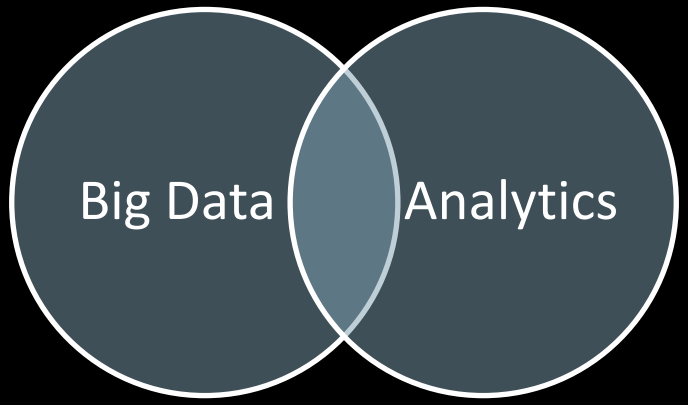
Kelompok WP	Total SPT Disampaikan	WP Wajib SPT	Kepatuhan
WP Badan	891.976	1.482.500	60,17%
WP Orang Pribadi Karyawan	12.105.924	14.172.999	85,42%
WP Orang Pribadi Non Karyawan	1.757.596	3.351.295	52,45%
	14.755.496	**19.006.794**	**77,63%**

- **other examples?**

Big Data Analytics



Big Data Analytics



Big Data

capacity to capture, gathering, store, accessing, and analyze high volume, high velocity, high variety data

structured and unstructured

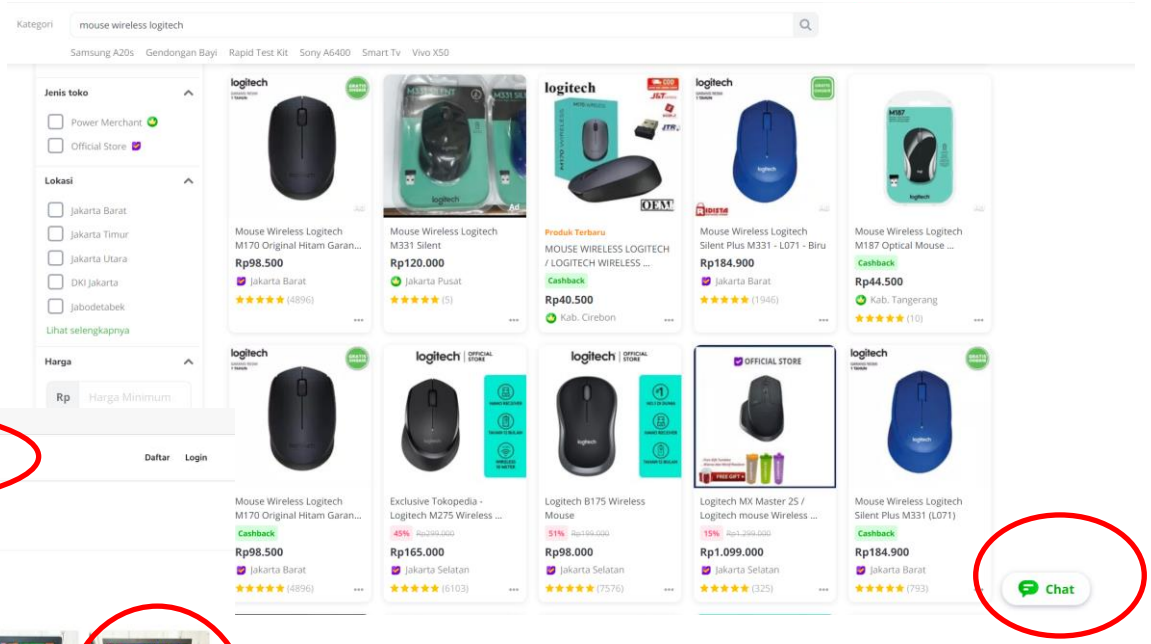
Analytics (Data Science)

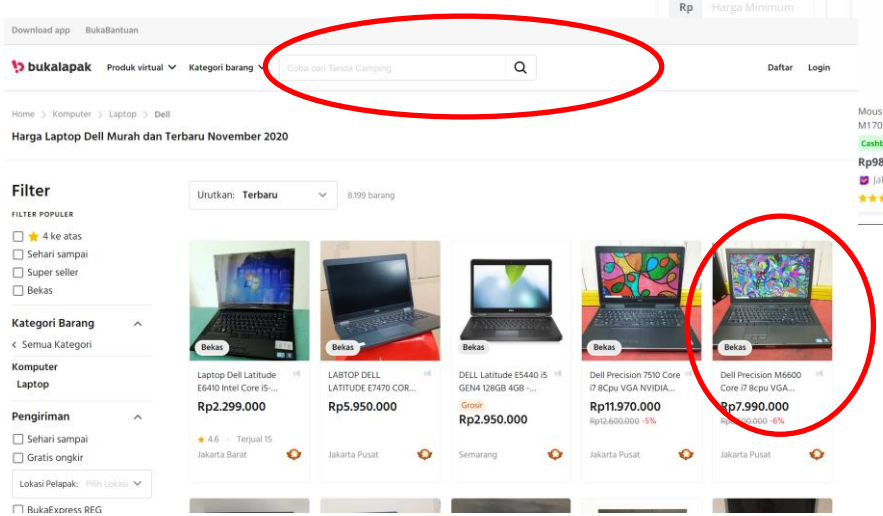
capacity to extract valuable knowledge / pattern / insight / model from data

Why Big Data?

Transactions + Interactions

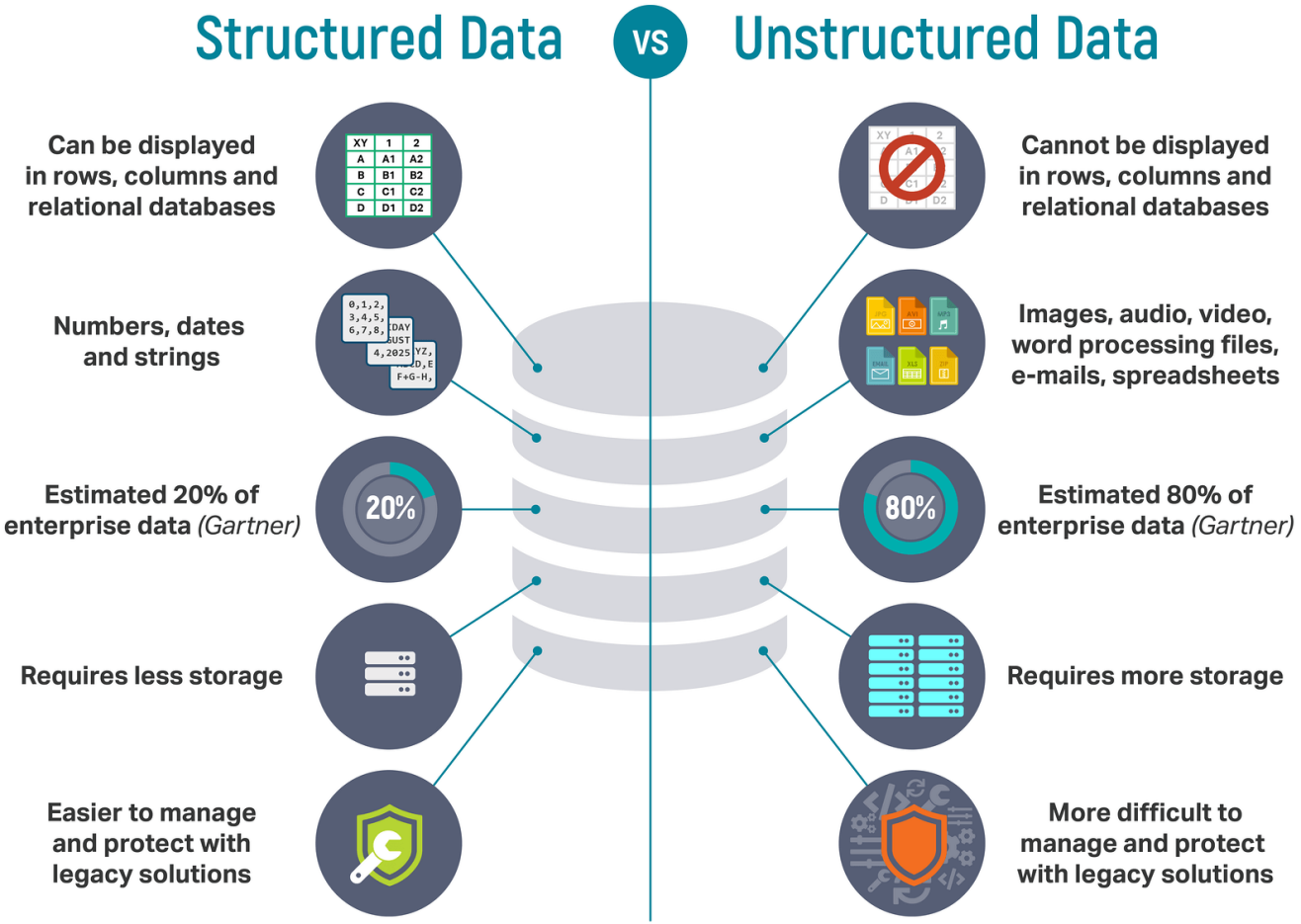
+ Observations + External Data (Social Media)





Which source of data represents the most immediate opportunity?

Structured vs Unstructured Data



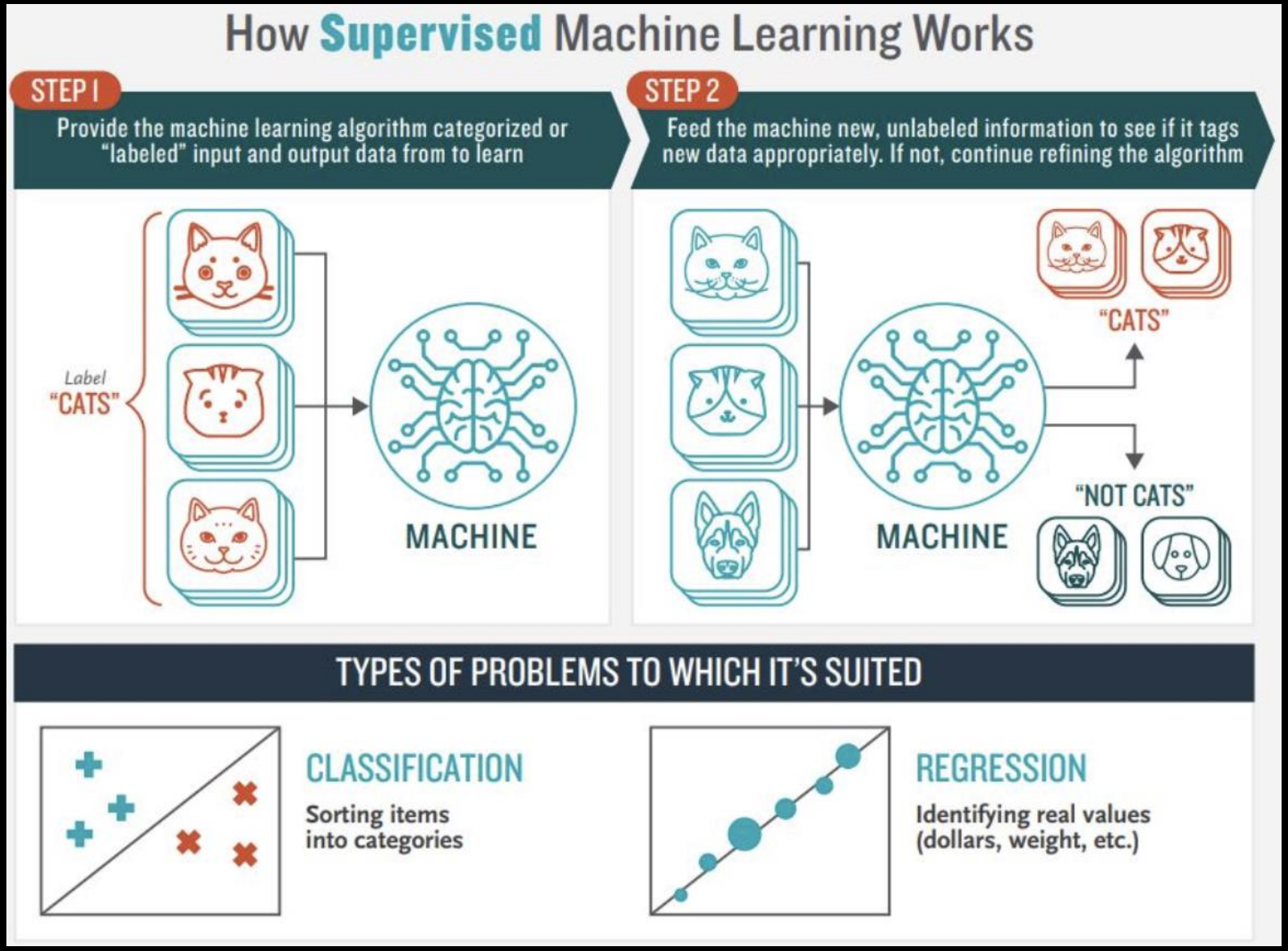
Big Data

Factors	Data Science / Analytics	Big Data
Concepts	Analysing data	Handling large data
Responsibility	Understand pattern within data and make decisions	Process huge volumes of data and generate insights
Industry	Sales, advertisement	Telecommunication, e-commerce

Big Data Analytics = Big Data + Data Science

Type of Learning

Supervised Learning

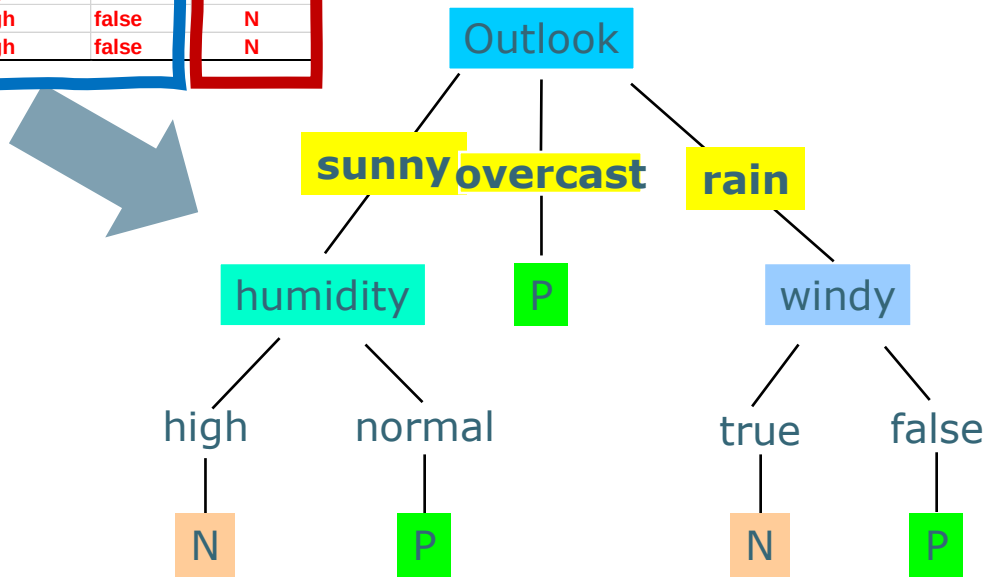


https://medium.com/@jorgesleonel/supervised-learning-c16823b00c13

predictor

Outlook	Temperature	Humidity	Windy	Class
overcast	mild	high	true	P
overcast	cool	normal	true	P
overcast	hot	high	false	P
overcast	hot	normal	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
sunny	cool	normal	false	P
rain	mild	high	true	N
rain	cool	normal	true	N
sunny	hot	high	true	N
sunny	hot	high	false	N
sunny	mild	high	false	N

response



ML Framework – Supervised Learning

Training data

		x	y	
x_1		[color = … , shape = …, texture = …]	orange	y_1
x_2		[color = … , shape = …, texture = …]	banana	y_2
x_3		[color = … , shape = …, texture = …]	apple	y_3
x_4		[color = … , shape = …, texture = …]	banana	y_4
x_5		[color = … , shape = …, texture = …]	apple	y_5

feature vector representation

Classification

Sebuah pemetaan fungsi *f* dari input data *x* ke sebuah label *y*

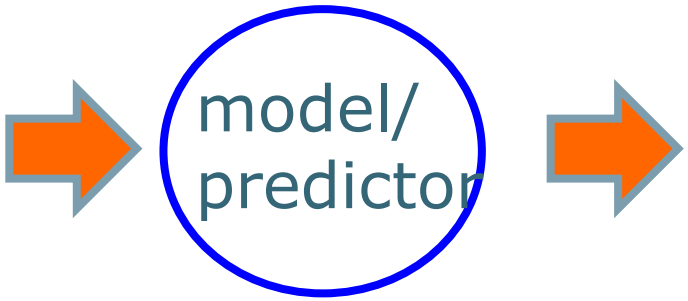
X = fitur/attribute buah

Y = {jeruk, apel, pisang}

$$f\left[\ \text{  }\ \right] = \text{banana}$$

x y

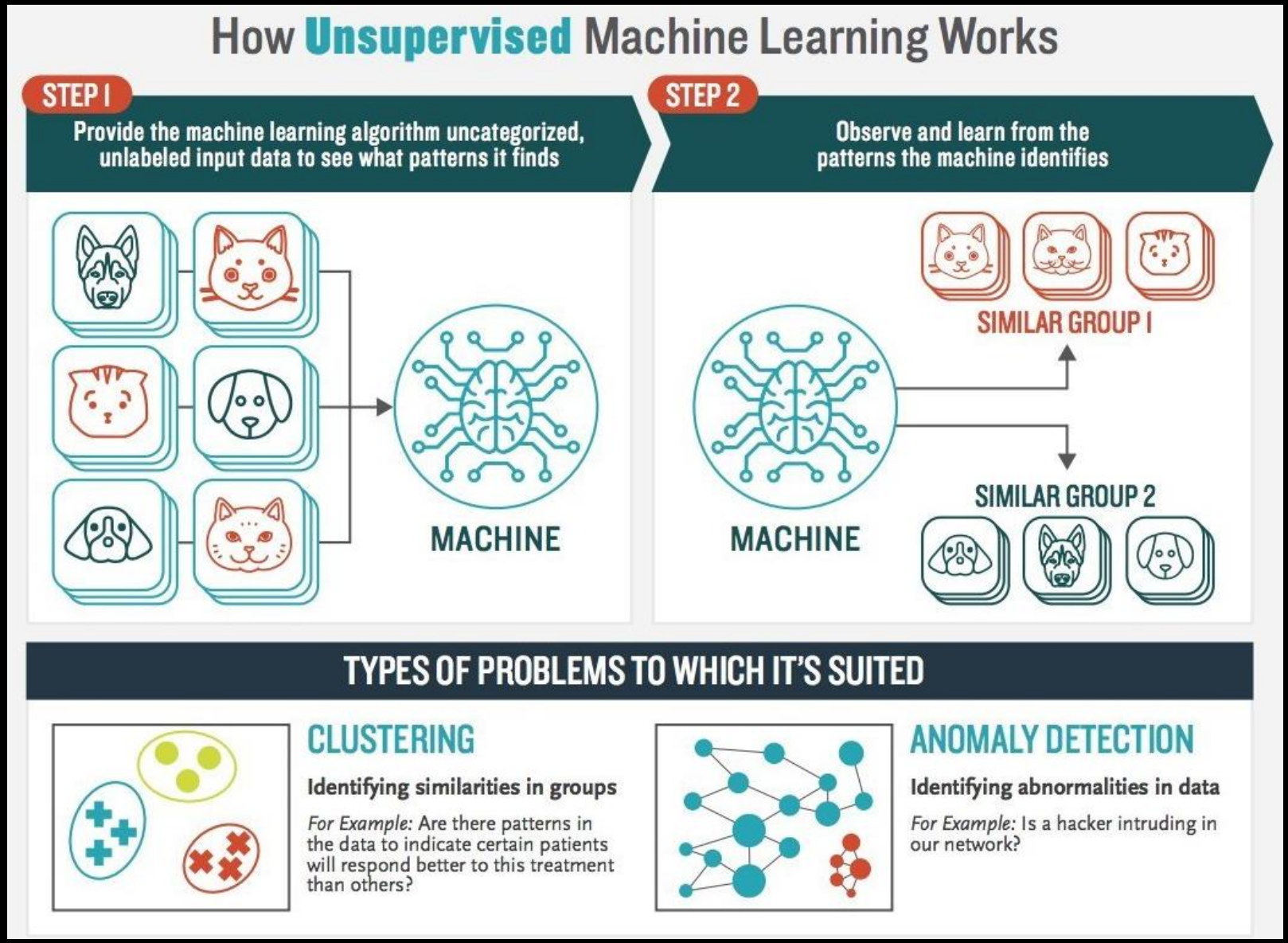
Supervised learning



predicted label

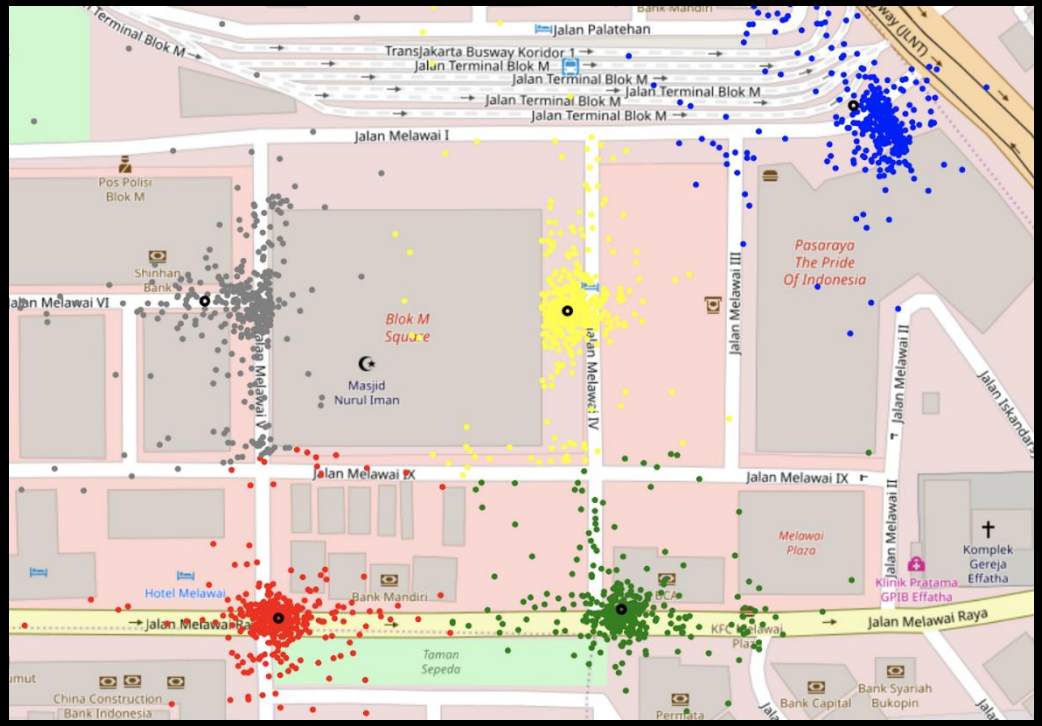
Supervised learning: learn to predict new example

Unsupervised Learning

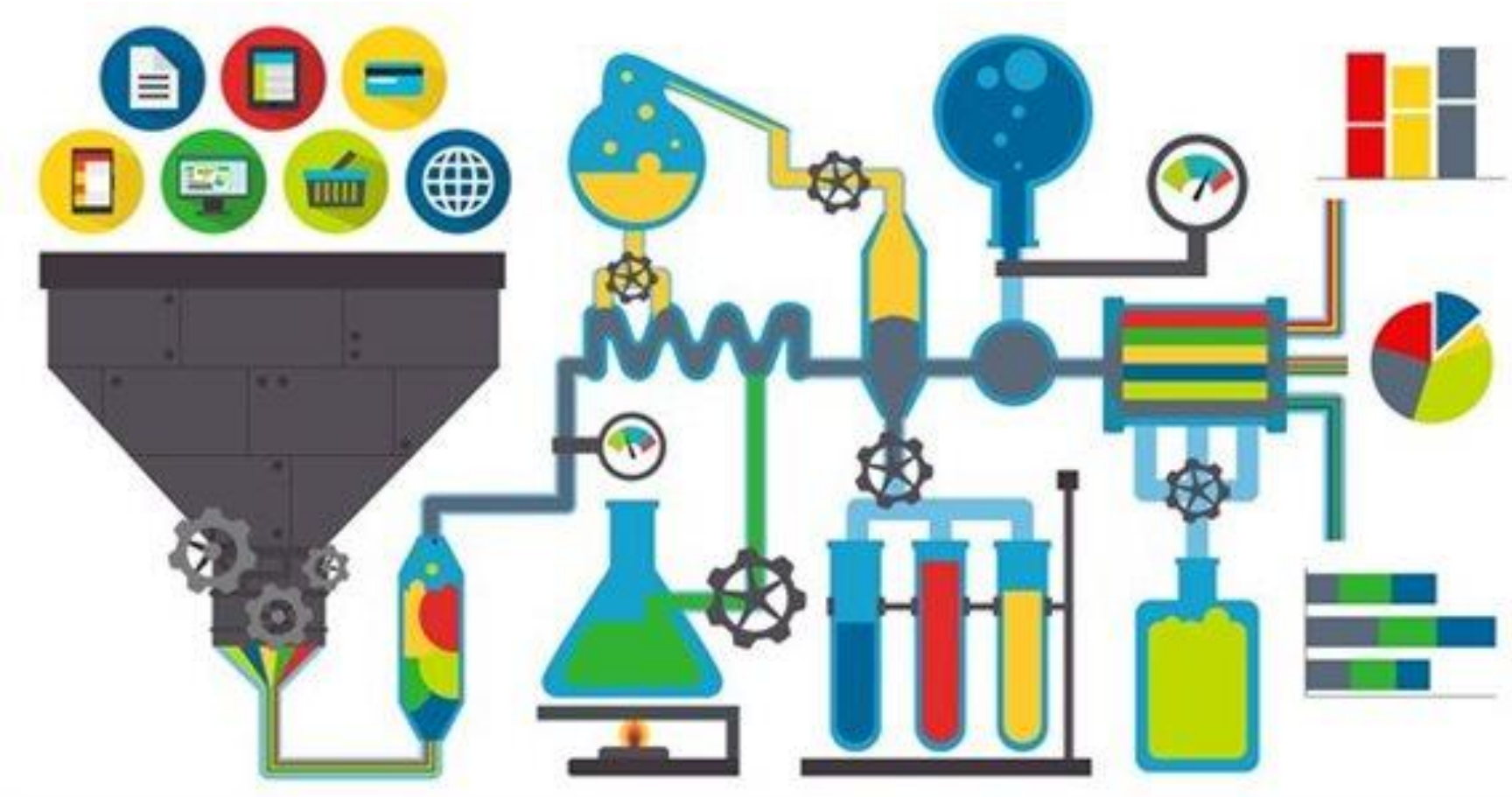


https://www.newtechdojo.com/list-machine-learning-algorithms/

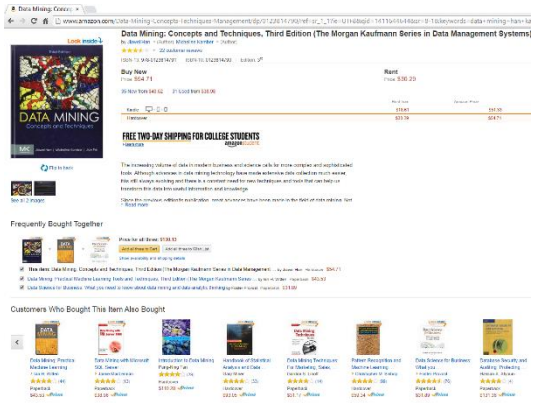
Unsupervised: Clustering Gojek Pickup Points



Data Science Functionality

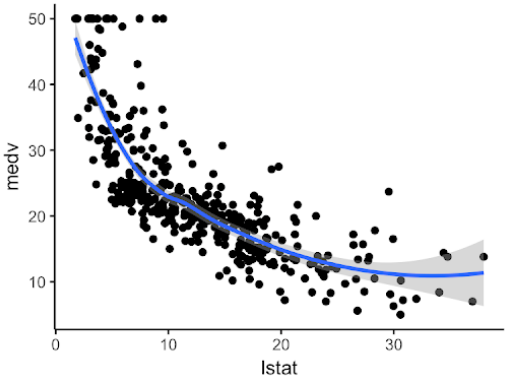


Data Science Functionality

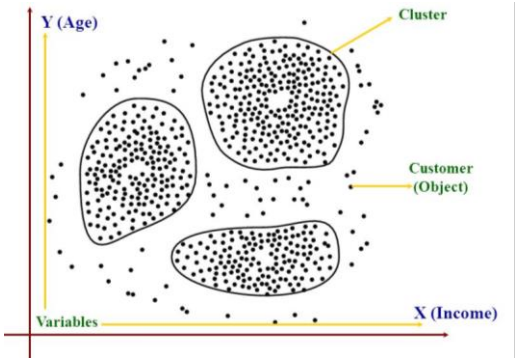


Association Rule Mining and Market Basket Analysis

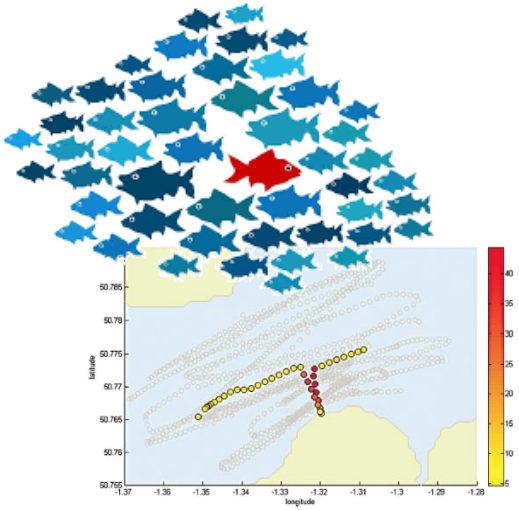
Classification



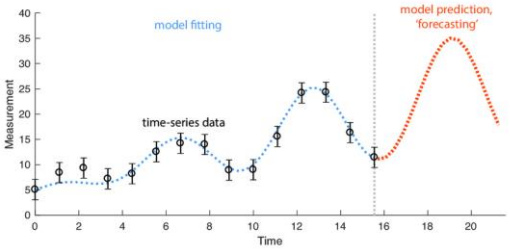
Regression



Cluster Analysis

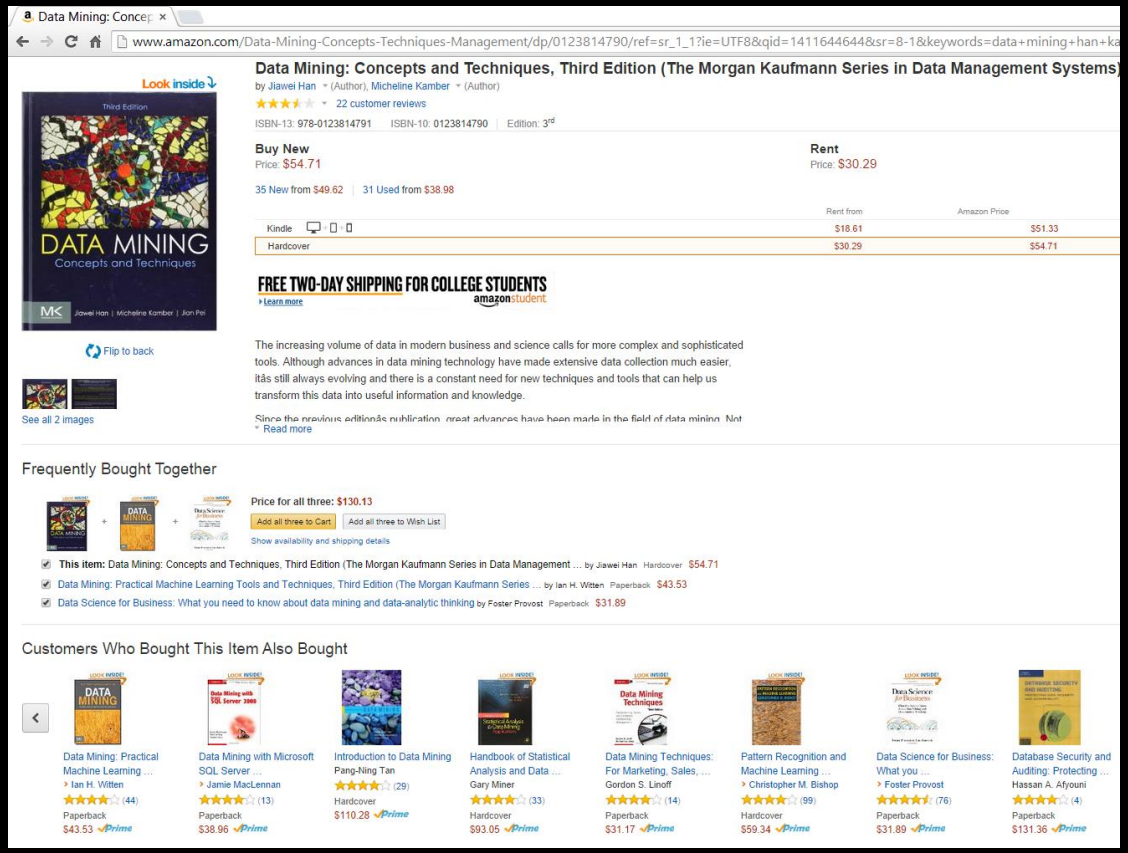


Outlier and exception data analysis

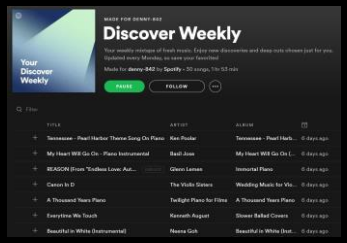


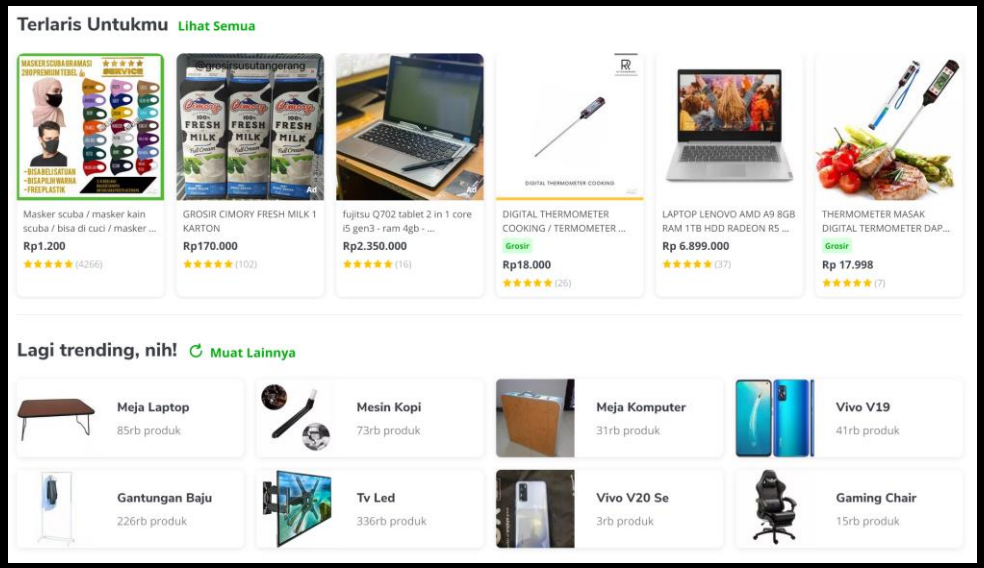
Time series analysis

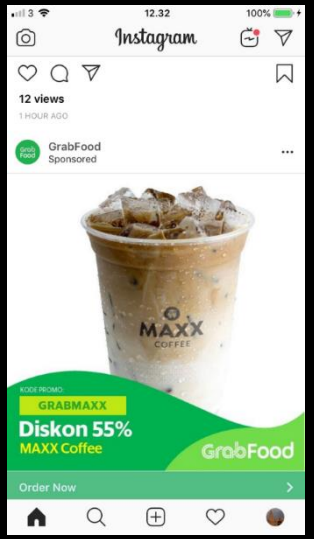
Market Basket Analysis, Association Rule Mining
Item Recommender




NETFLIX
Top Picks for DD
OBLIVION
THE RUTHLESS
BLACK SUMMER
LIFE
VIKINGS
TOM CRUISE
JACK REACHER
THE SILENCE

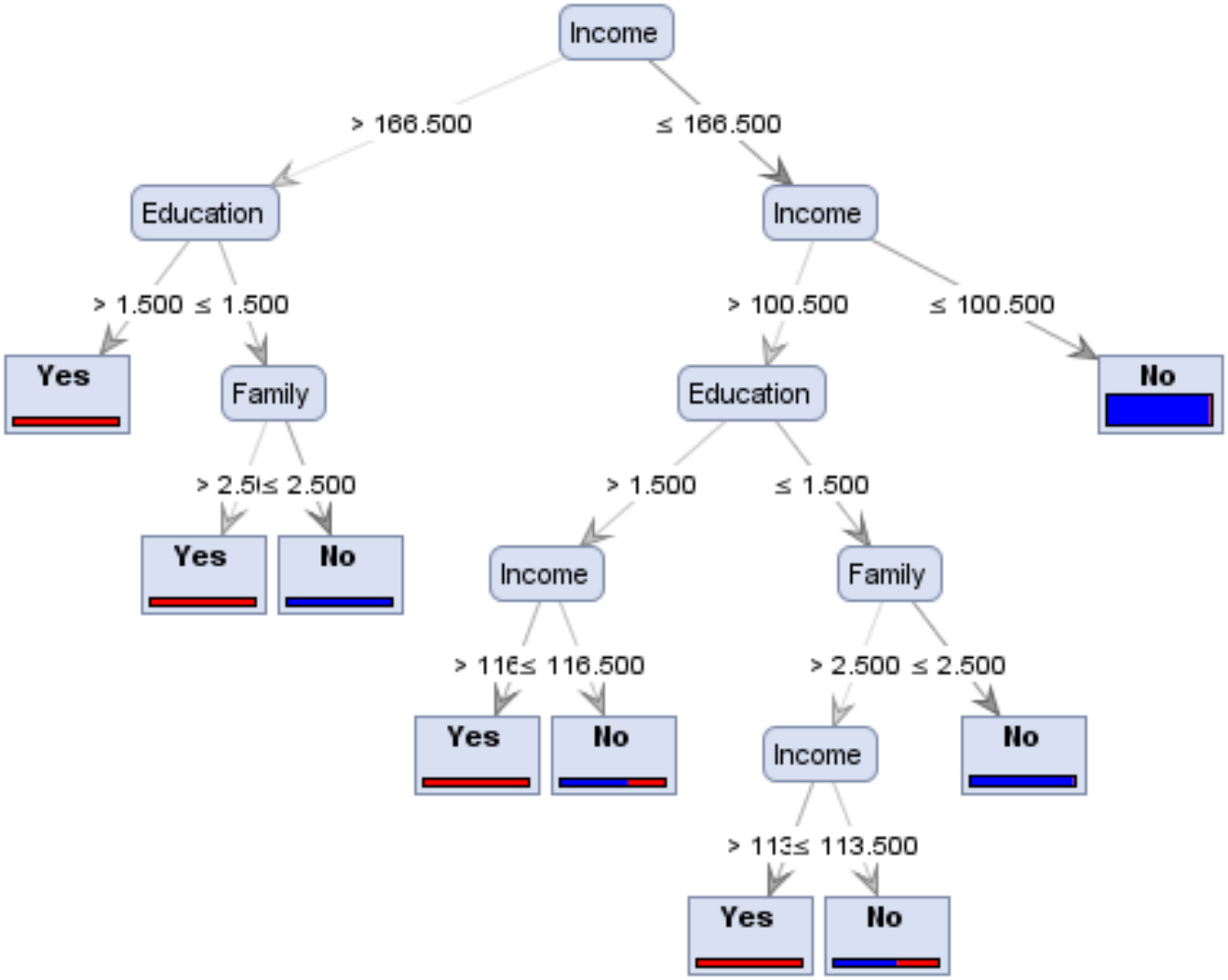

Discover Weekly
Your Discover Weekly


Terlaris Untukmu
Lihat Semua
Masker scuba / masker kain scuba / bisa di cuci / masker ...
Rp1.200
GROSIR CIMORY FRESH MILK 1 KARTON
Rp170.000
fujitsu Q702 tablet 2 in 1 core i5 gen3 - ram 4gb - ...
Rp2.350.000
DIGITAL THERMOMETER COOKING / TERMOMETER ...
Grosir
Rp18.000
LAPTOP LENOVO AMD A9 8GB RAM 1TB HDD RADEON R5 ...
Rp 6.899.000
THERMOMETER MASAK DIGITAL TERMOMETER DAP...
Grosir
Rp 17.998
Lagi trending, nih!
Muat Lainnya
Meja Laptop
85rb produk
Mesin Kopi
73rb produk
Meja Komputer
31rb produk
Vivo V19
41rb produk
Gantungan Baju
226rb produk
Tv Led
336rb produk
Vivo V20 Se
3rb produk
Gaming Chair
15rb produk

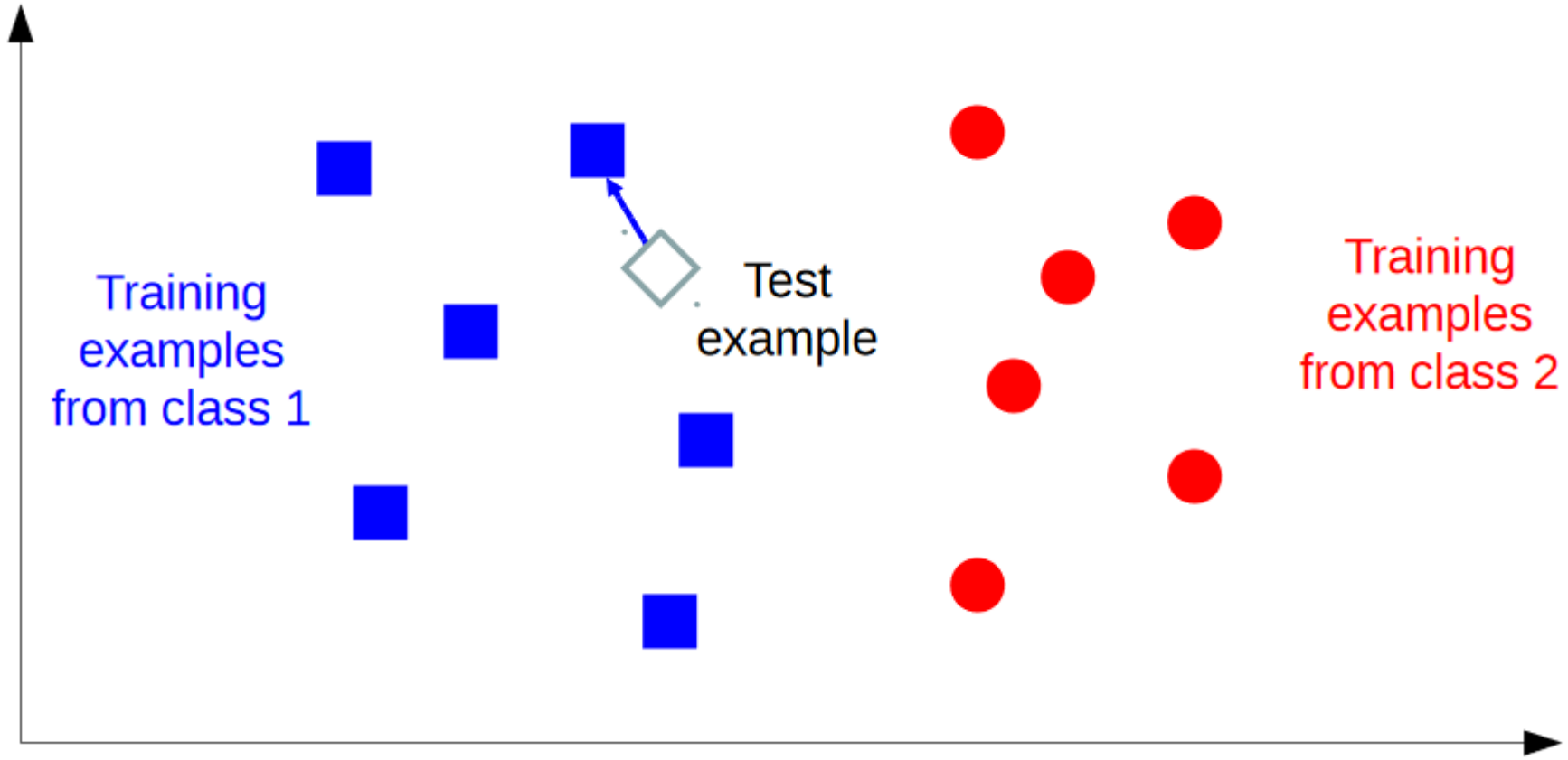

Instagram
12 views
1 HOUR AGO
GrabFood
Sponsored
MAXX COFFEE
GRABMAXX
Diskon 55%
MAXX Coffee
GrabFood
Order Now

CLASSIFICATION AND REGRESSION

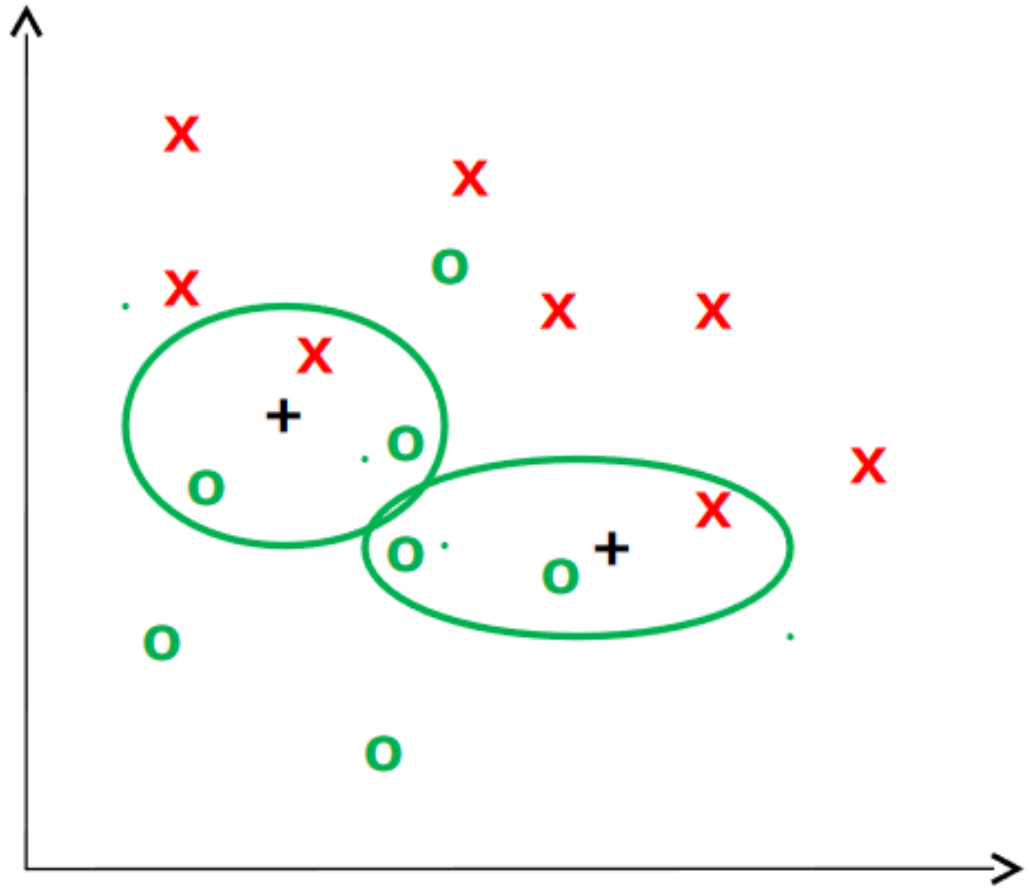
Classification: Decision Tree



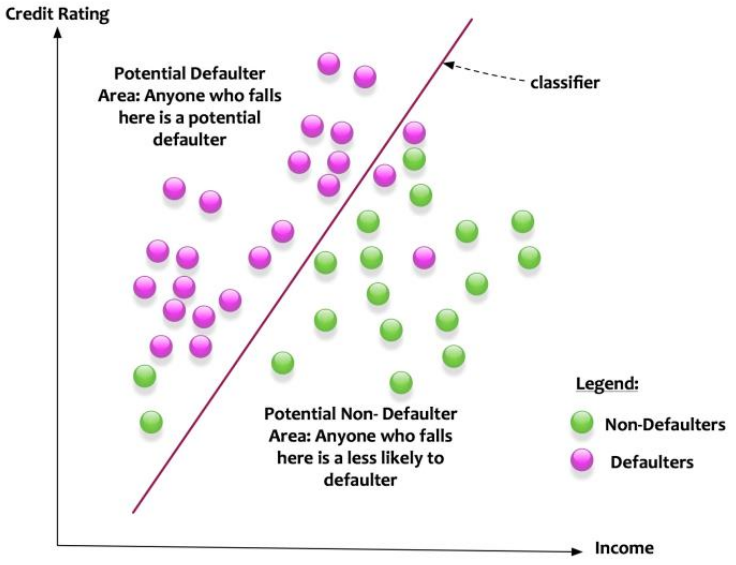
k-nearest neighbour



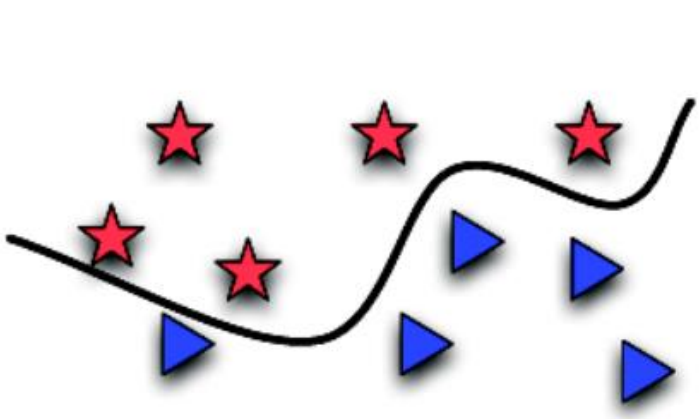
3-nearest neighbour

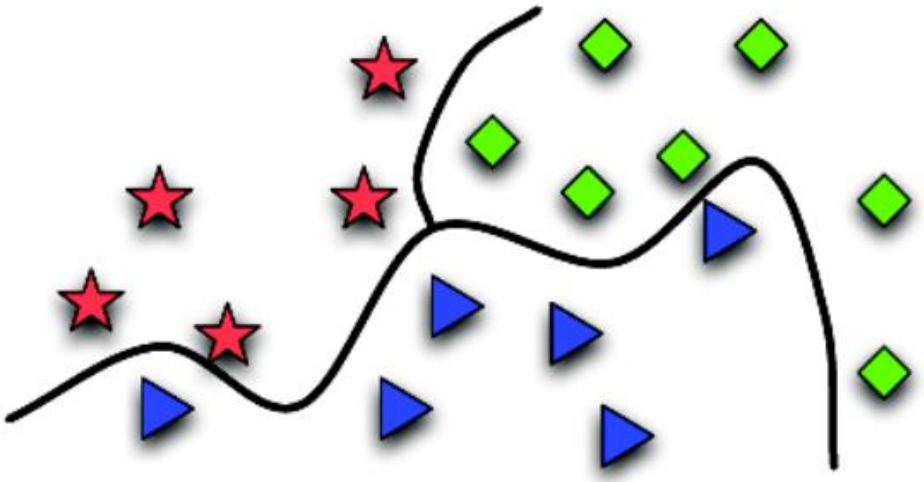


Classification: Decision Boundary

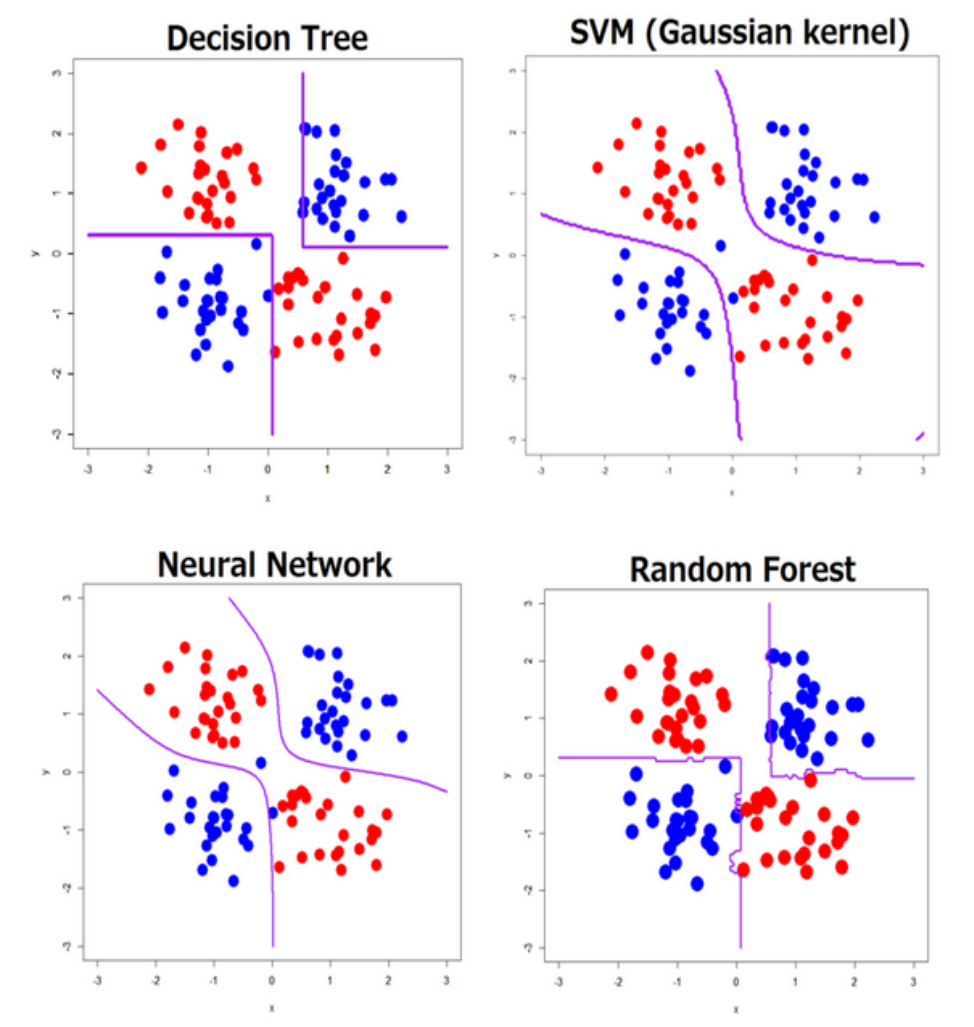


Classification: Decision Boundary

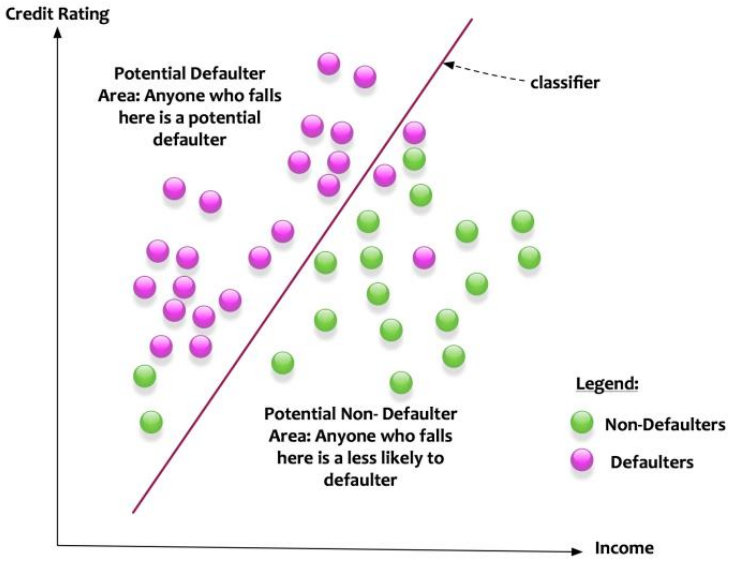




Classification

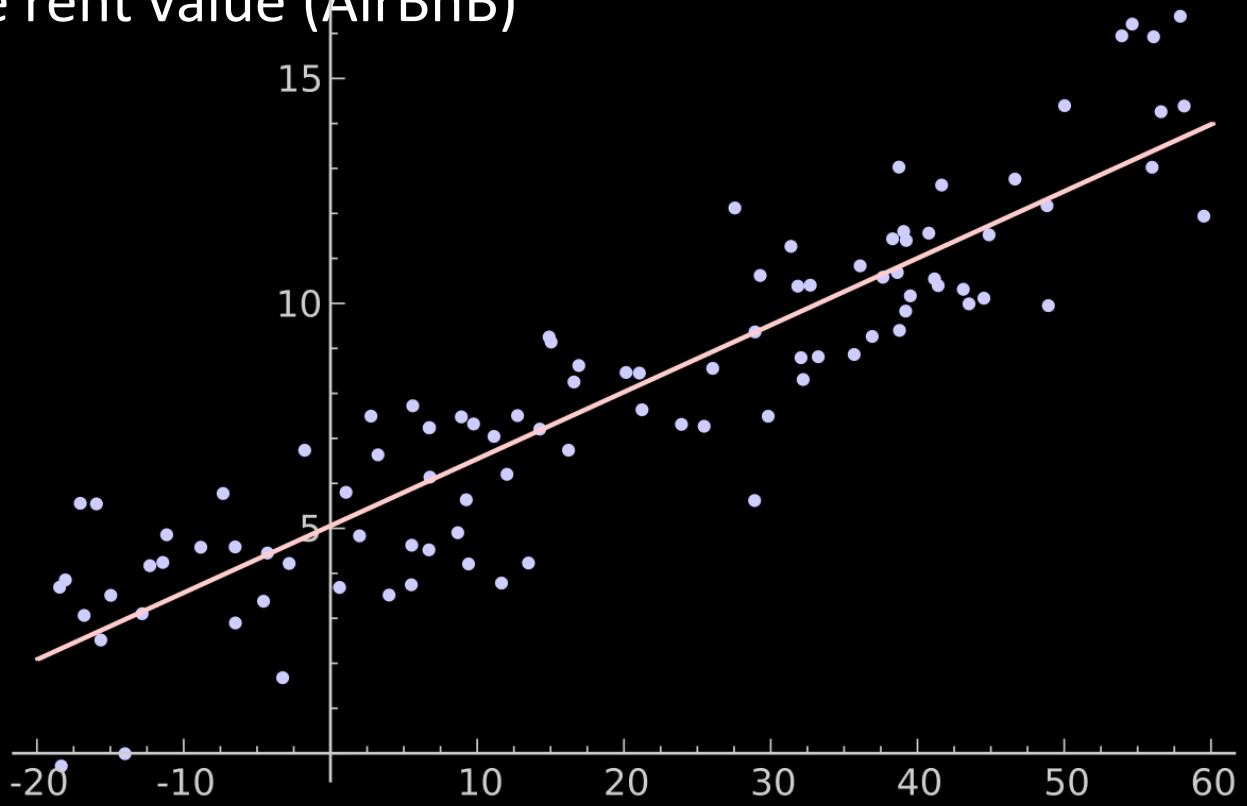


Classification: Decision Boundary

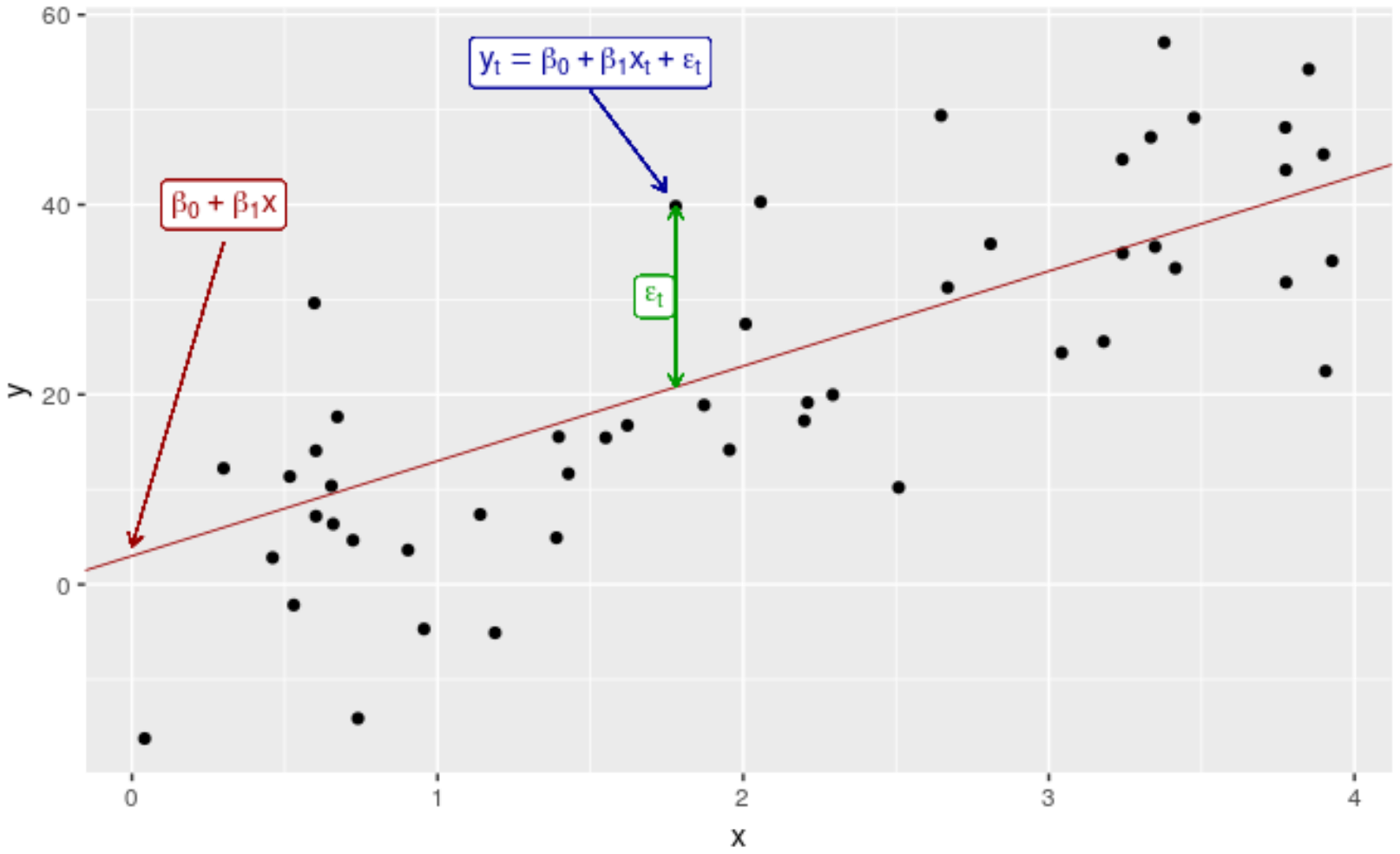


Regression

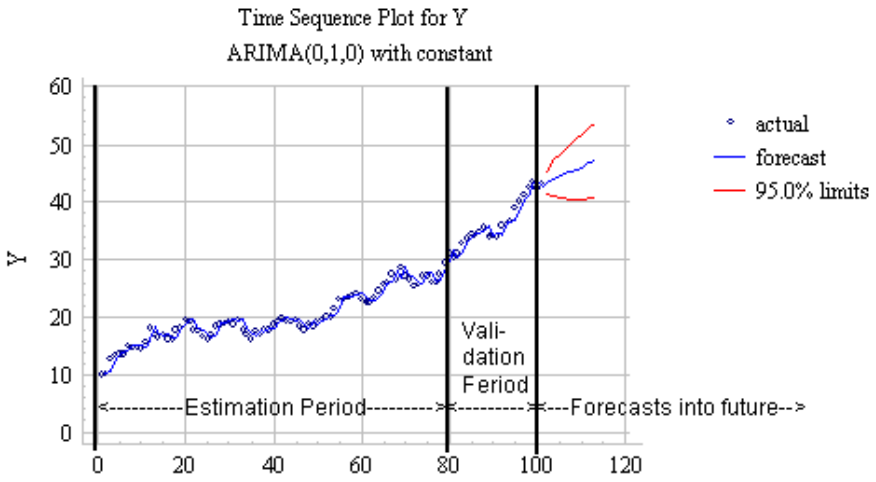
- **predict continuous value for each case in the data**
- **example:**
  - estimate value of tax adjustments
  - estimate rent value (AirBnB)

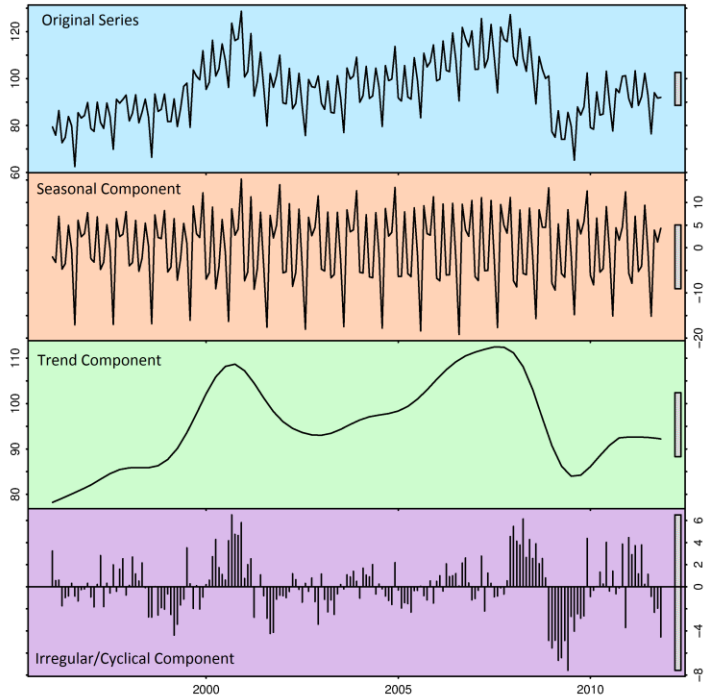


Regression


$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$
$\beta_0 + \beta_1 x$
ε_t
60
40
20
0
y
0
1
2
3
4
x

Time Series Analysis





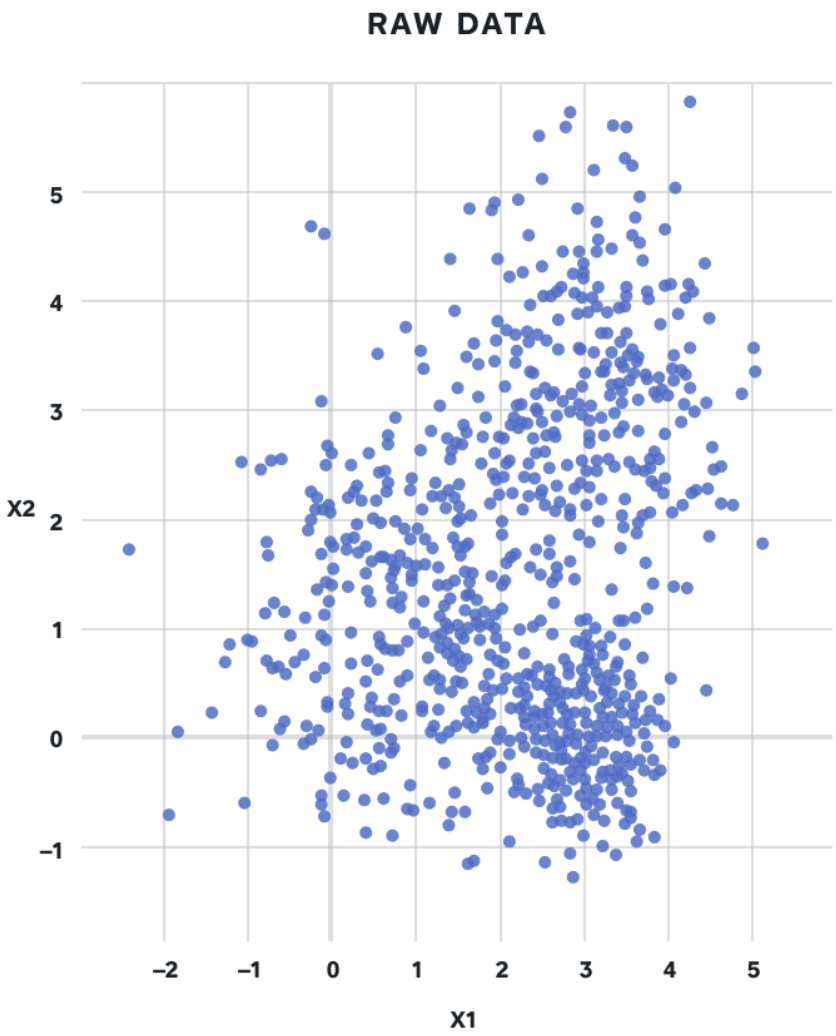
CLUSTER ANALYSIS

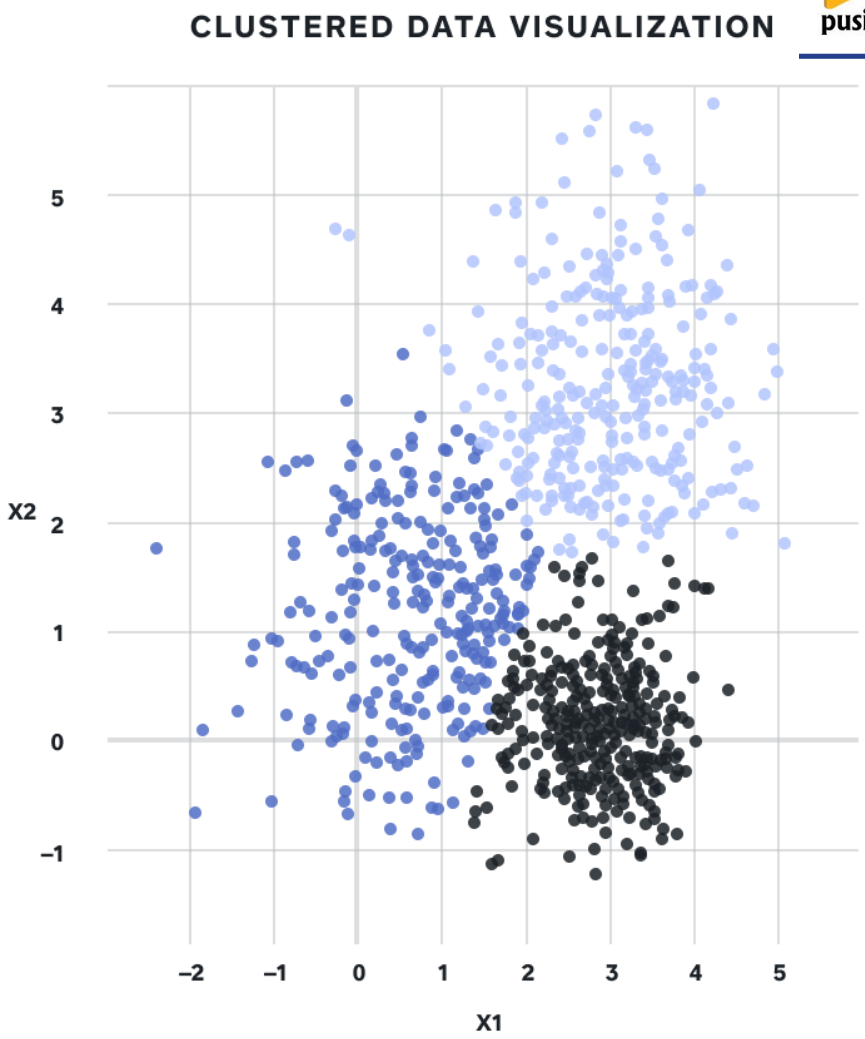
What is Cluster Analysis?

- **Cluster: a collection of data objects**
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- **Cluster analysis:**
  - Grouping a set of data objects into clusters
- **Clustering is unsupervised classification.**
  - no predefined classes

- **Useful for data understanding phase – data exploration**

Cluster Analysis

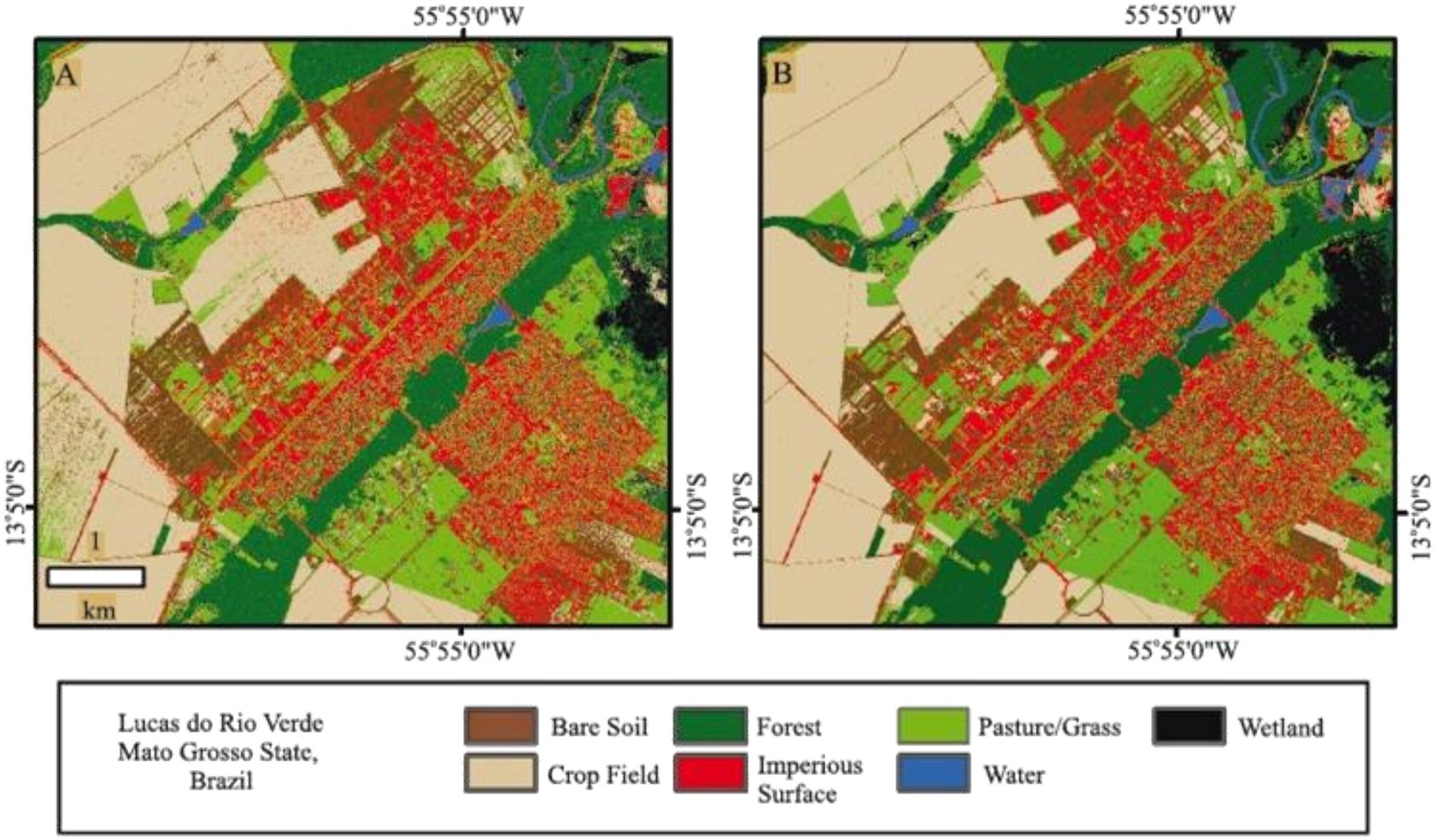




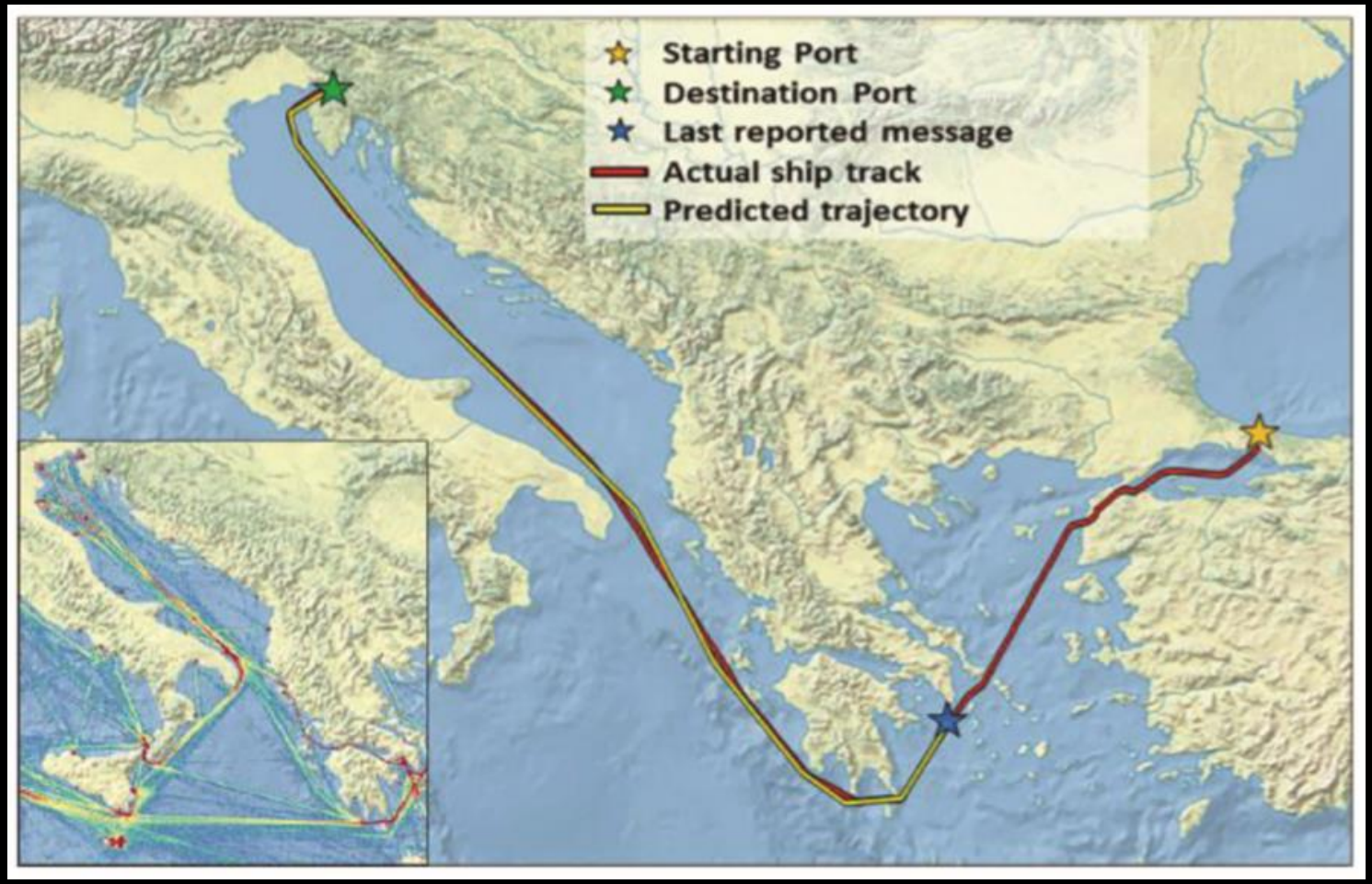
Clustering Applications

- **Marketing**: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Land use**: Identification of areas of similar land use in an earth observation database
- **Insurance**: Identifying groups of motor insurance policy holders with a high average claim cost
- **City-planning**: Identifying groups of houses according to their house type, value, and geographical location
- **Economic**: clustering countries/regions with similar development indicators
- **Maritime**: clustering ships routes to find the energy efficient route

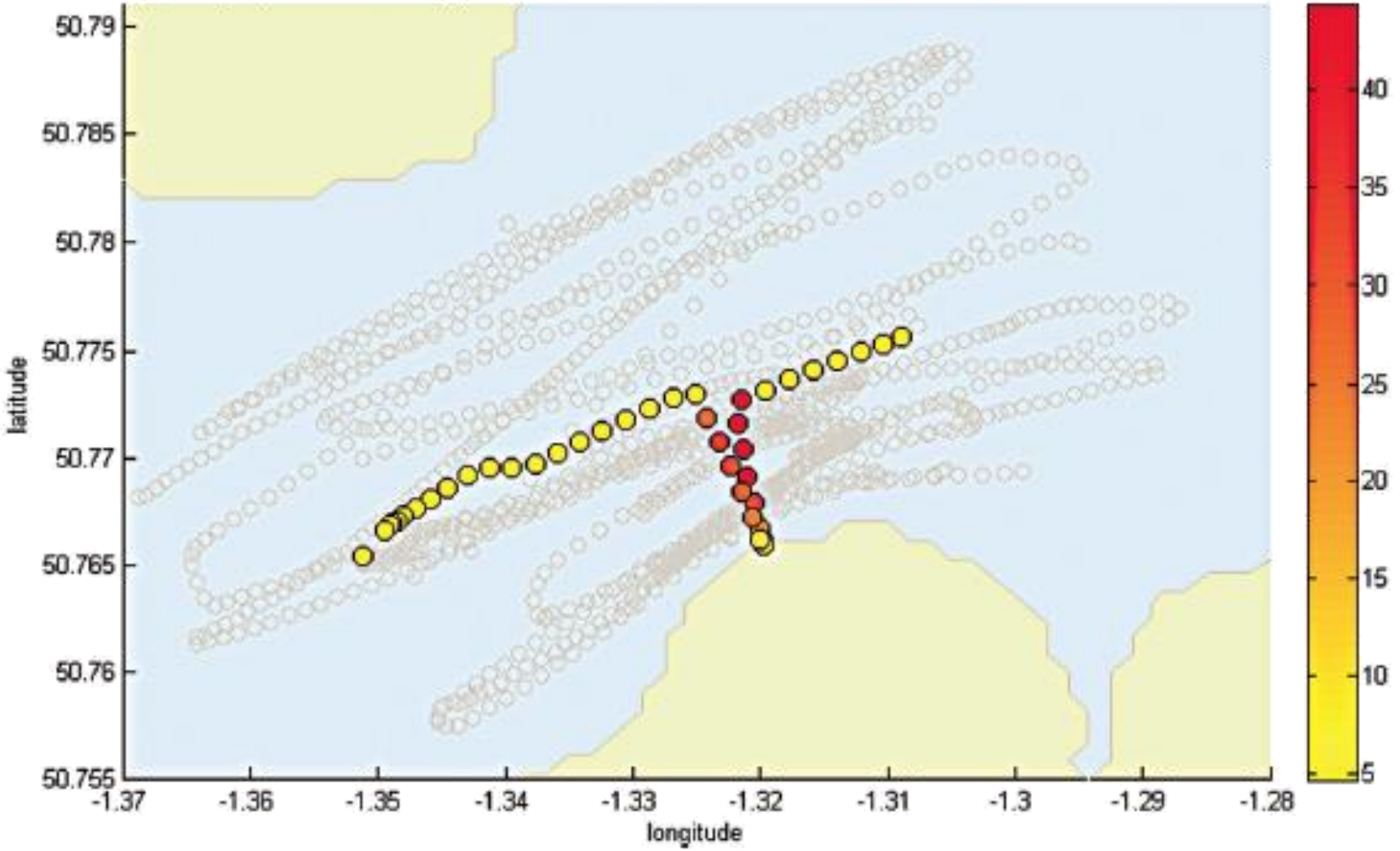
Land use/cover classification in the Brazilian Amazon



Logistic Optimization



Anomaly Detection



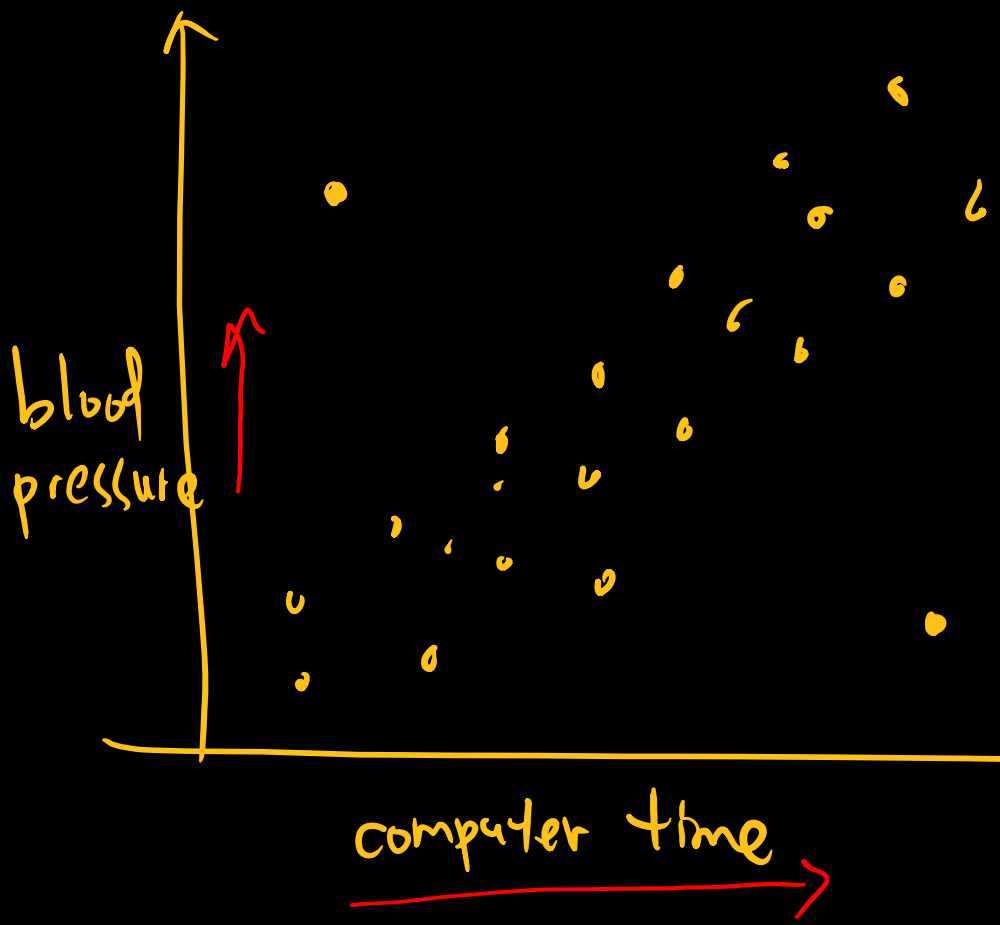
Caution

- **Data mining find regularities from history, but history is not the same as the future.**
  - Concept drift – e.g. pandemic
  - Population drift
- **Association does not dictate trend nor causality.**
  - Observational vs Experimental data
- **Some abnormal data could be caused by human.**
  - Noise, Bias

Observational versus Experimental

- **Observational study**
  - observe only, no "treatment" assigned
  - generally, a control group is not needed
  - reports an association
  - may (or not) use random sample
  - may (or not) generalize to population
- **Experiment**
  - "treatment" assigned
  - uses control group for comparison
  - reports a cause and effect
  - randomization of sample group
  - generalize to population

Observational Study



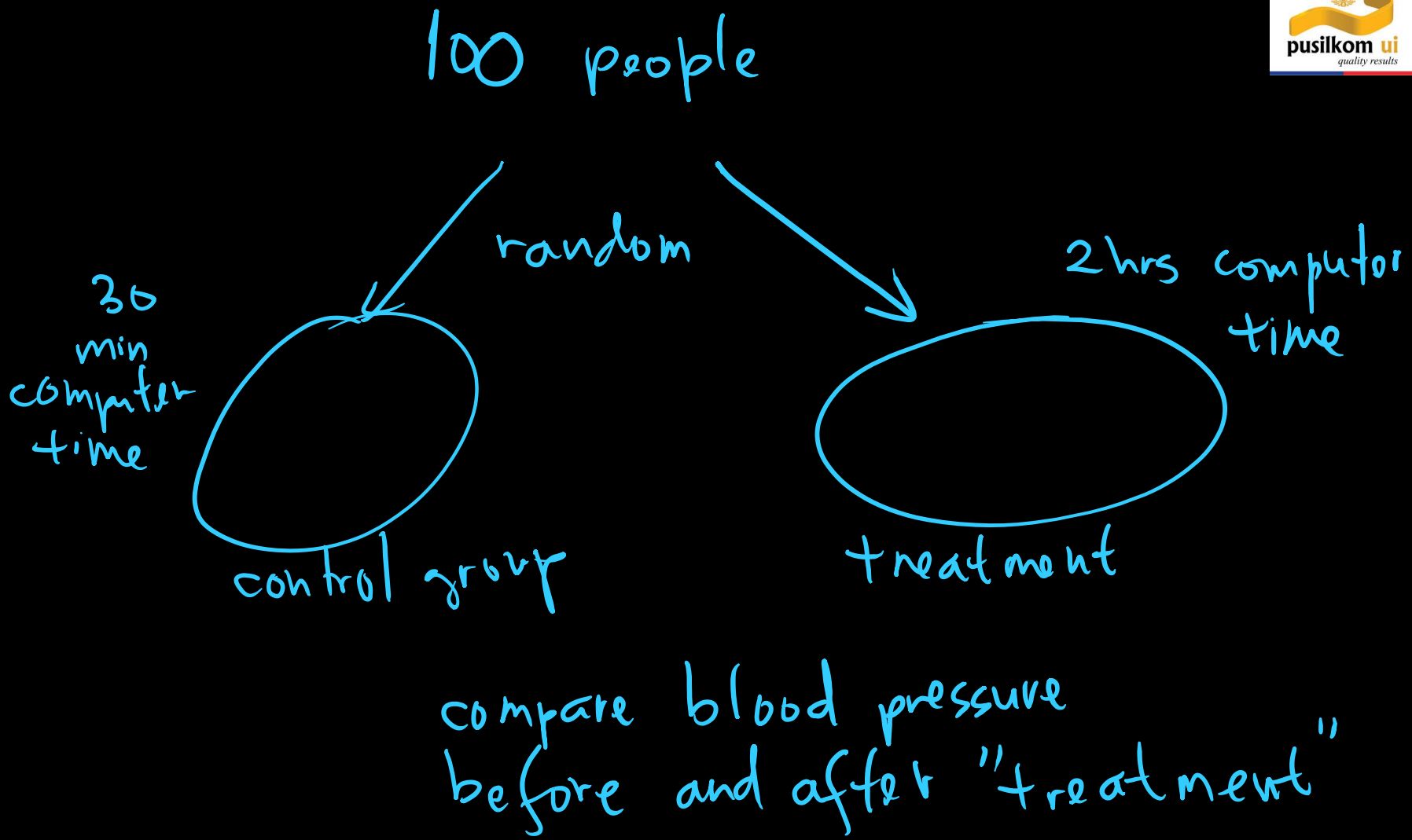
lurking / confounding variable

lack of physical activity

computer time

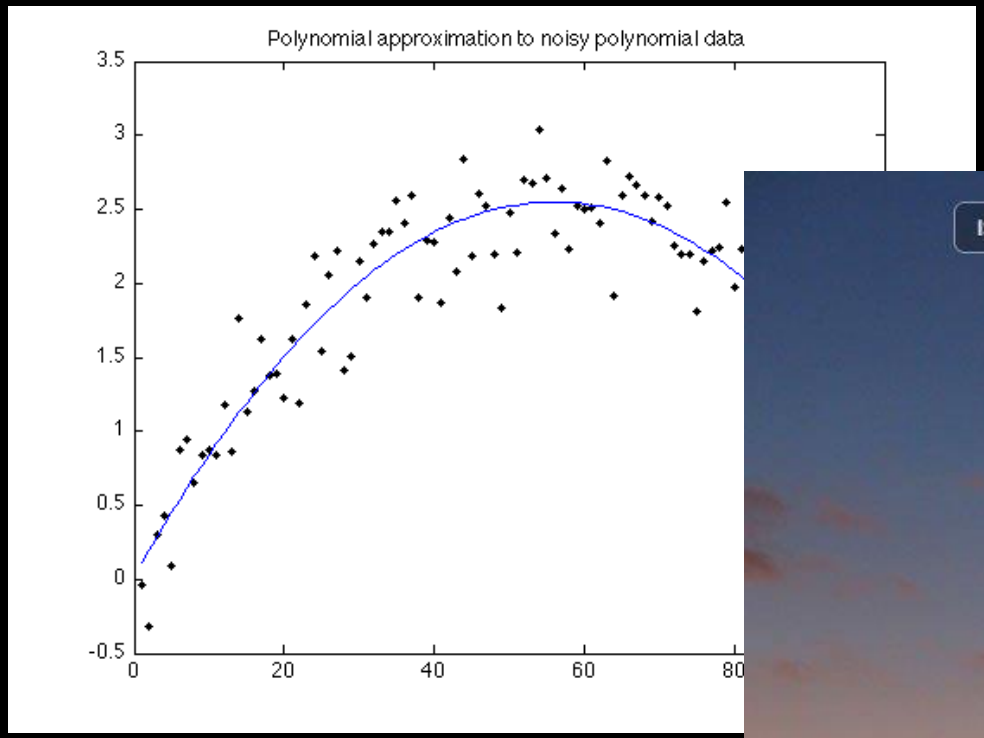
blood pressure

Experiment


100 people
random
30 min computer time
2 hrs computer time
control group
treatment
compare blood pressure before and after "treatment"

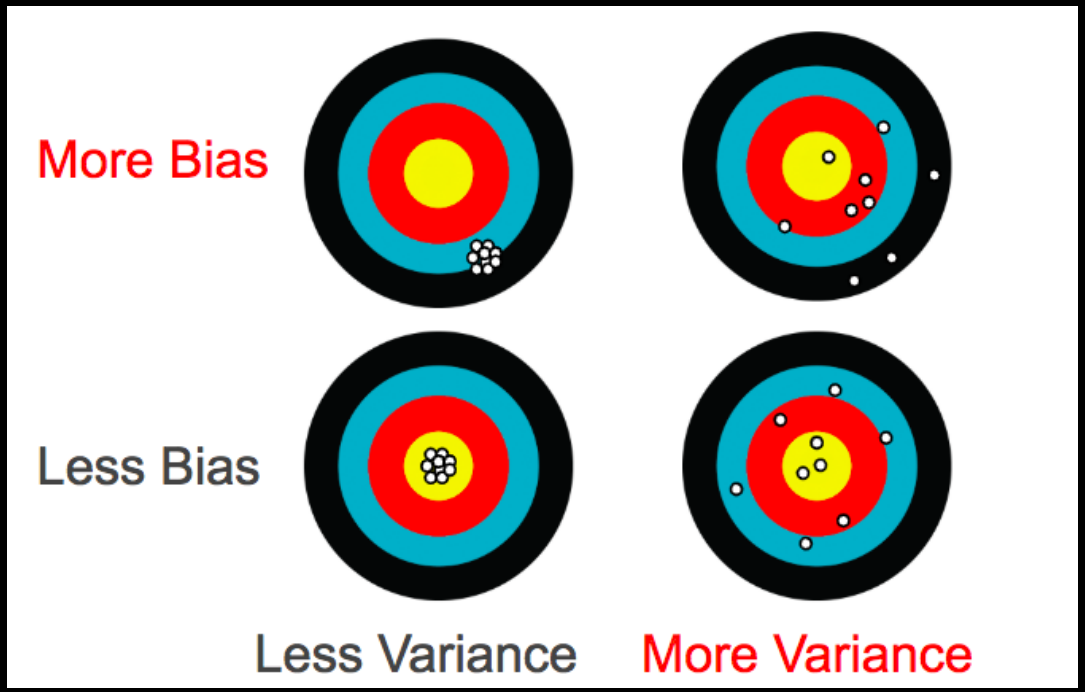
Noise

data that is corrupted, or distorted, or has a low Signal-to-Noise Ratio



Bias

- **Bias refers to *results that are systematically off the mark.***



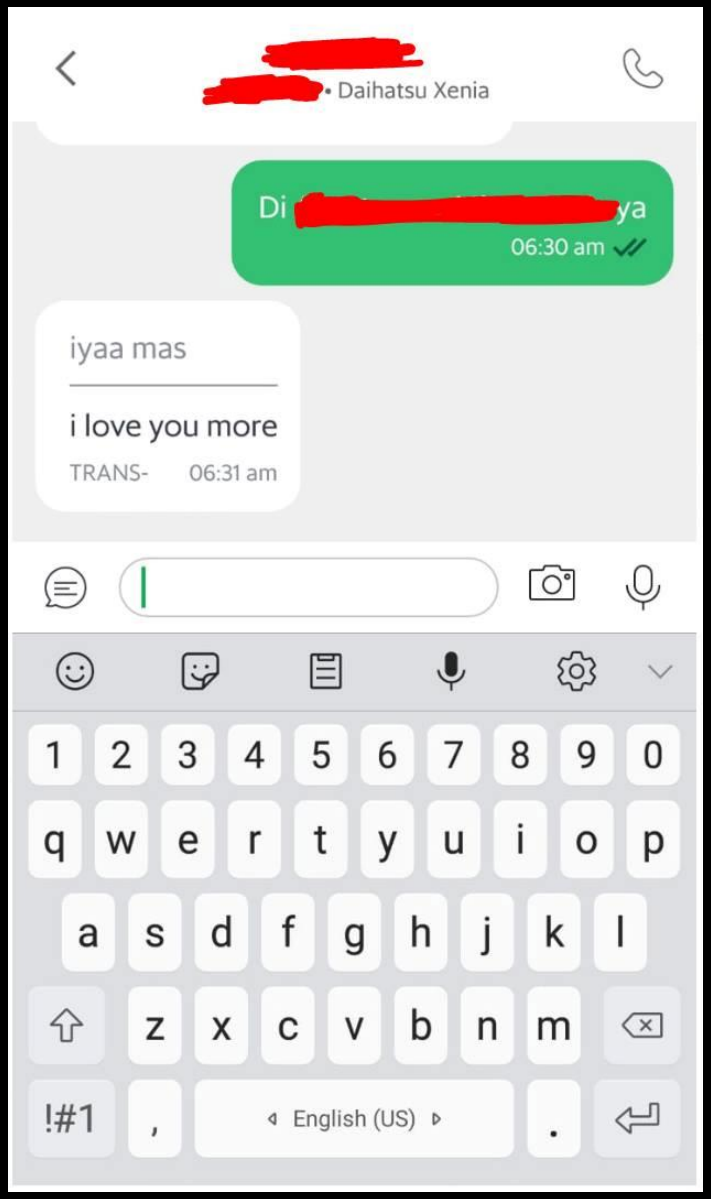
Question:

Temperature gun measure body temp at 32-34 centigrade: Bias or Noise?

Bias

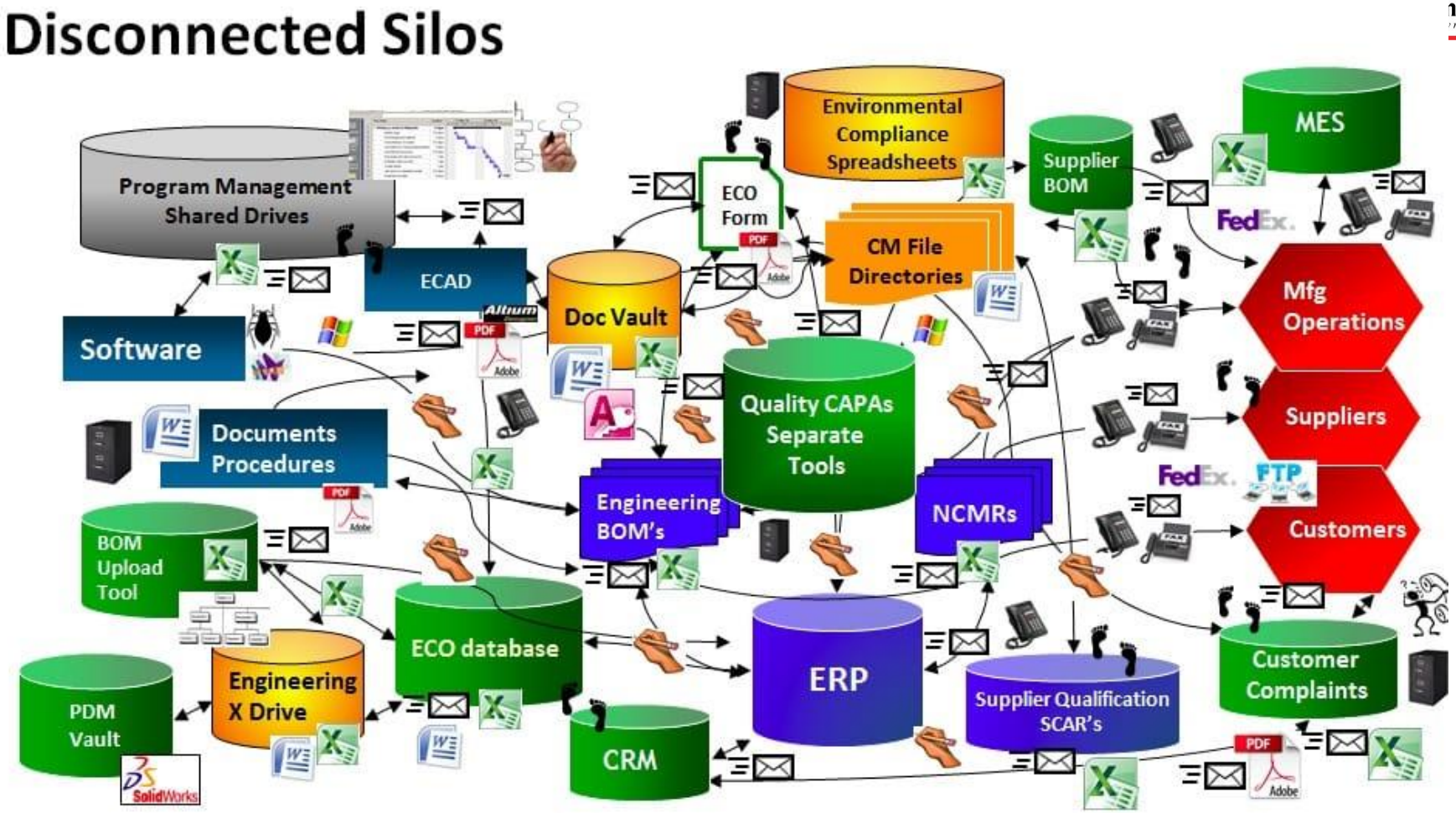
- **Sample Bias**
  - collected data doesn't accurately represent the environment the program is expected to run into
- **Exclusion bias**
  - Happens as a result of excluding some feature(s) from our dataset usually under the umbrella of cleaning our data
- **Observer bias (aka experimenter bias)**
  - The tendency to see what we expect to see, or what we want to see
- **Prejudice bias**
  - Happens as a result of cultural influences or stereotypes.
- **Measurement bias**
  - Systematic value distortion happens when there's an issue with the device used to observe or measure. This kind of bias tends to skew the data in a particular direction.

True story: In a Transportation App

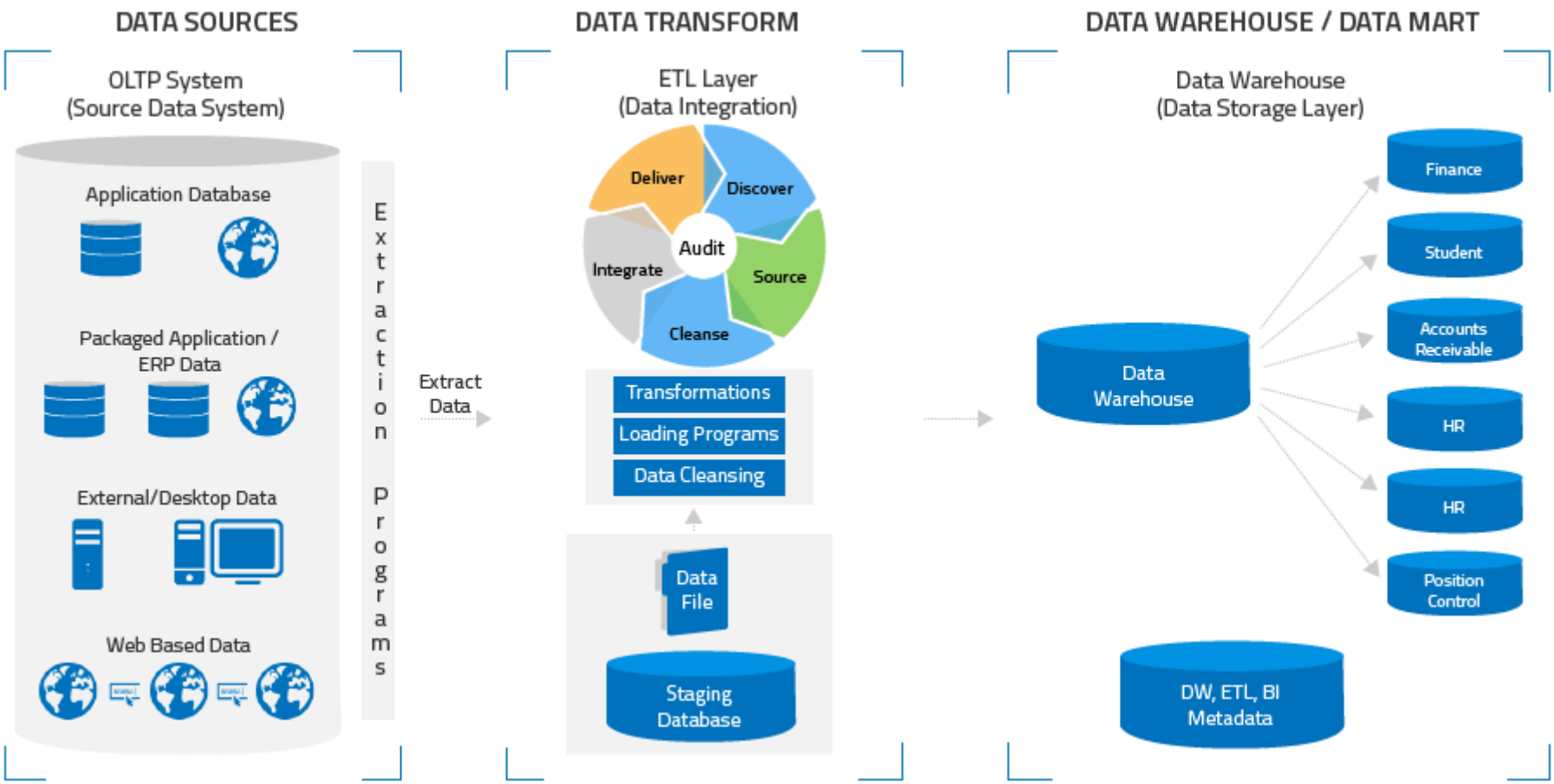


CHALLENGES

Challenges: need Data Management



Data Warehouse might help



Challenges

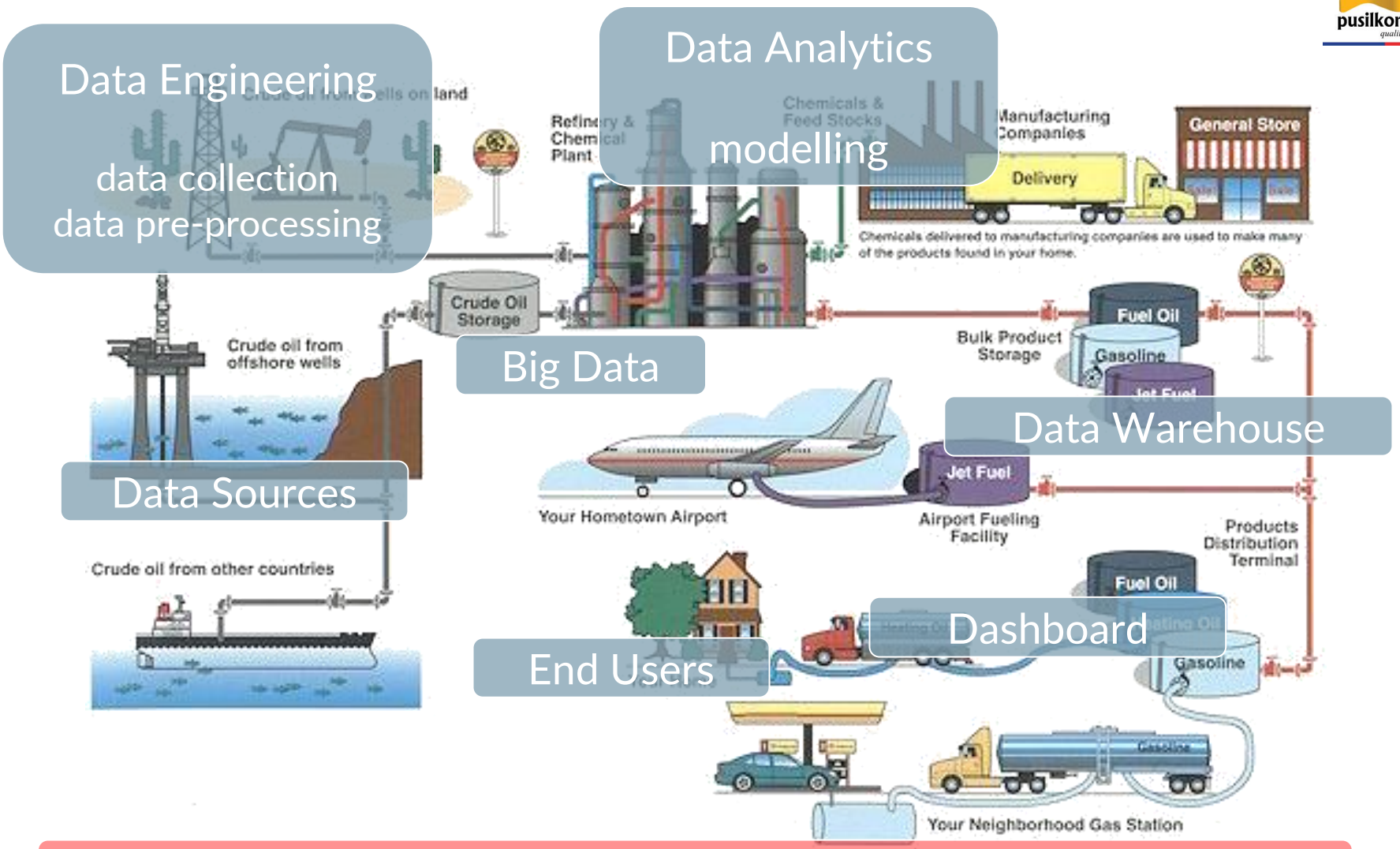
- **Data science needs skilled people, not off the shelf solutions (yet)**
- **Plus staffs who are competent in statistics, econometrics**

- **Shortage of skilled data scientist**
  - not only know how to use the tools, but understand the underlying mechanism
  - dedicated assignment?
- **Data matching / Record linkage – link datasets**

More Challenges

- **Volume**
- **External sources**
- **Unstructured data**
  - Text
  - Images
  - Audio and/or Video Streaming
  - Geospatial

Data is the new OIL!



Data Science Stages

https://ajgoldstein.com/2017/11/12/deconstructing-data-science/

Skills Required

01 FRAME THE PROBLEM

- **Domain Knowledge** (needs)
- **Product Intuition** (metrics)
- **Business Strategy** (priorities)
- **Teamwork** (people & resources)

02 COLLECT RAW DATA

- **Database Management**
  - Systems: mySQL, postgreSQL, Oracle, MongoDB
- **Querying Structured Databases**
  - SQL
- **Retrieving Unstructured Info**
  - Informational Retrieval / Text Mining
- **Distributed Storage**
  - Hadoop HDFS, Spark, Flink

03 PROCESS THE DATA

- **Scripting Language**
  - Python or R
- **Data Wrangling & Cleaning**
  - Python "Pandas" library
- **Distributed Processing**
  - Hadoop MapReduce / Spark

04 EXPLORE THE DATA

- **Scientific Computing**
  - Python: numpy, matplotlib, scipy, pandas
- **Inferential Statistics**
  - hypothesis testing
  - correlation vs. causation
- **Experimental Design**
  - A/B tests, controlled trials

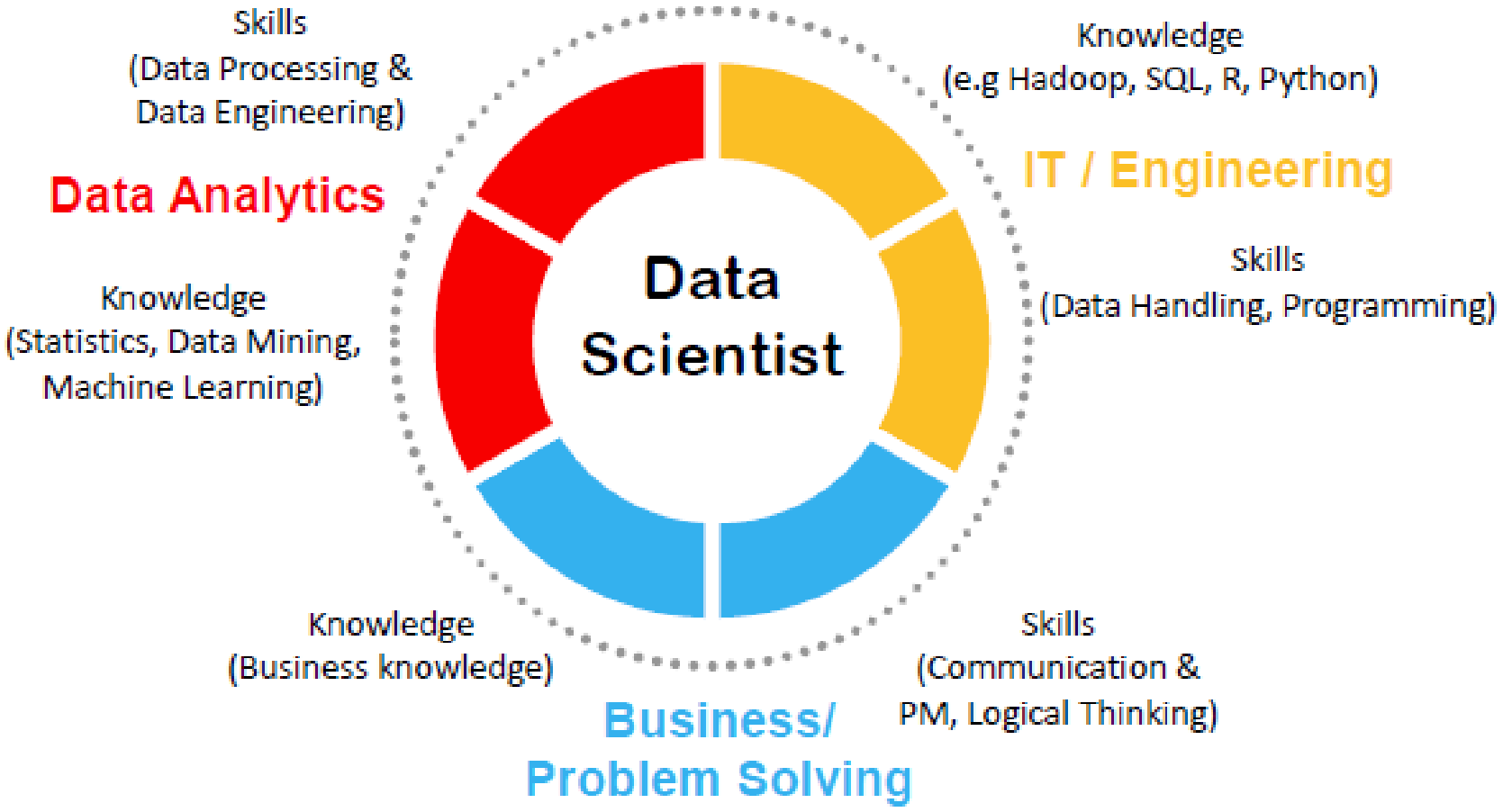
05 PERFORM IN-DEPTH ANALYSIS

- **Machine Learning**
  - Supervised / Unsupervised algorithms
  - Contextual pros/cons)
- **ML Tools Library**
  - Python: scikit-learn
- **Advanced Math**
  - Linear Algebra & Multivariate Calculus

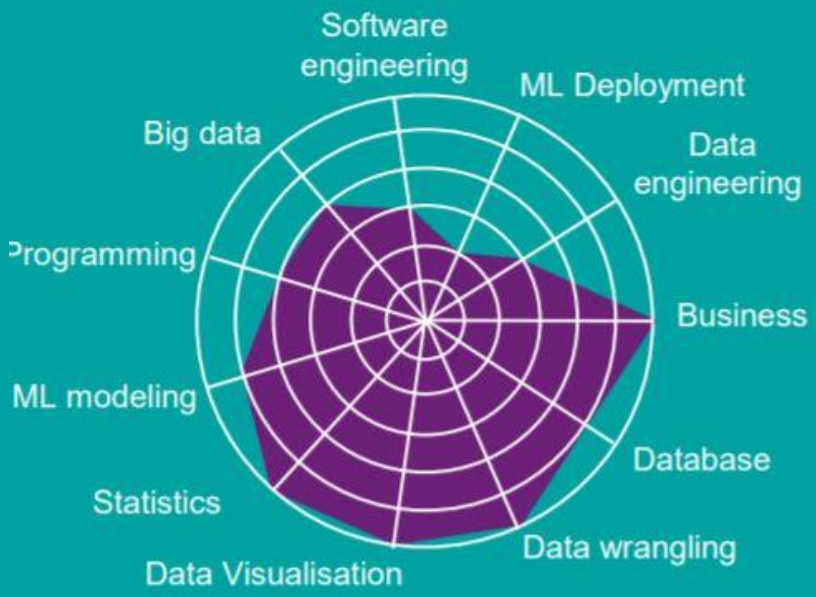
06 COMMUNICATE RESULTS

- **Business Acumen**
  - Non-technical terminology
- **Data Visualization Tool(s)**
  - Tableau, D3.js, Google visualize, matplotlib, ggplot, seaborn
- **Data Storytelling**
  - presenting & speaking
  - reporting & writing

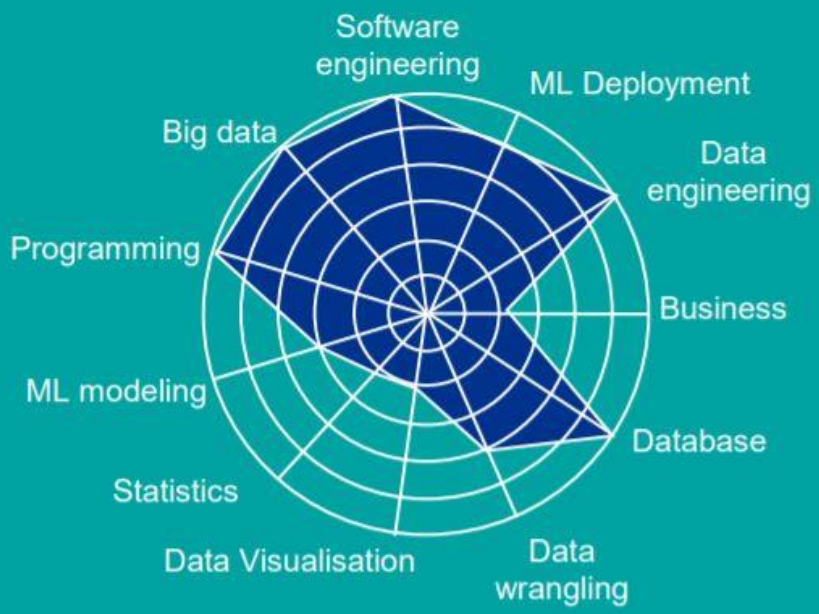
Data Scientist



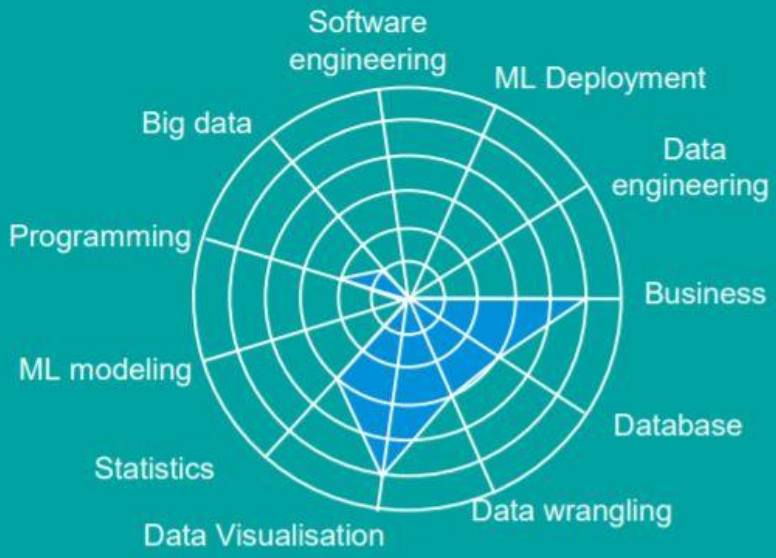
Data scientist



Data engineer



Data analyst



Big Data Analytics Training Series

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia

pusilkom ui
quality results

Data Science for Non-Programmers	Python for Data Science	Data Mining for Big Data	NLP & Text Mining
Python Programming	Big Data Engineering	Sosial Media Analysis	Advanced Data Sciences
	Knowledge Graphs	AI for Business Executives	

Summary

- **think of business value of data science, first**
- **analytics:**
  - classification, regression, clustering, time series analysis, outlier analysis
- **caution:**
  - hidden context, concept drift, noise, bias

Analytics should start from business problem rather than technology.

Clear business goals that the organization aims to achieve using Data Analytics

Don’t start without understanding the **value**!

What is my ROI?

Discussion

- **Describe the problem (more specific is better).**
  Explain the value proposition, if we can solve the problem.
- **How do you measure the value?**
- **Decision / action that can help to solve the problem**
- **Describe a solution that help to make decision or take action**
- **What kinds of data are needed? Where can we get those data? Which data are readily available?**
- **Are there any risks in implementing the proposed solution? How to mitigate the risks?**