Data Science
Using Python

By Eng. Mohammed Marwan Shahin

Pre-requisite Courses:

- Basics of Python
- Intro to Artificial Intelligence

Outlines

- Ch1 Data Science Overview
- Ch2 Mathematical and statistical computation in python using numpy
- Ch3 Data Import / Export, Data Exploration,Data Wrangling using Pandas
- Ch4 Data Visualization using matplotlib
- Ch5 Bank Chrun Demo using Machine Learning

Data & Data Science Overview

What is Data Science?

A powerful new approach to make discoveries from data

An automated way to analyze enormous amounts of data and extract information

A new discipline that combines aspects of statistics, mathematics, programming, and visualization to turn data into information

Confused?

Business dictionary

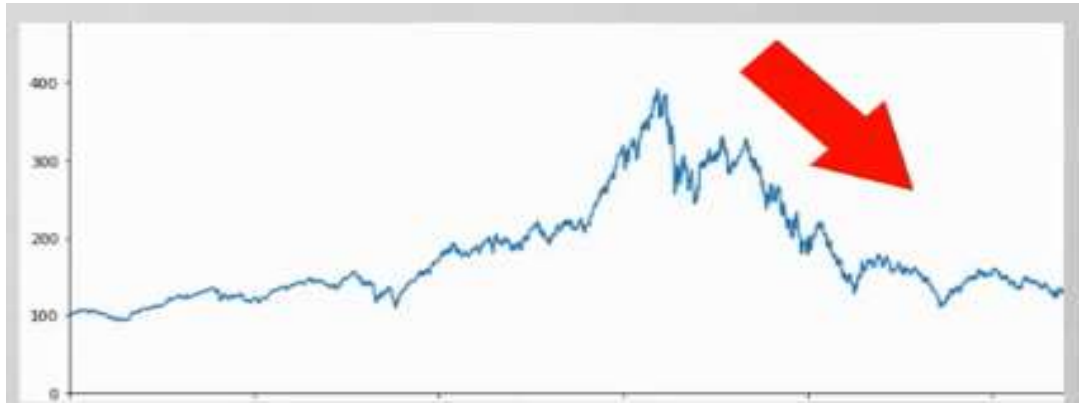
- Data
- Data team
- Big data team
- Business intelligence
- Data science
- Business analytics
- Data analytics

What is the difference?

Analysis ≟ Analytics

Analysis

Past



Explain

How? Why?

Analytics

Future

Explore potential future events

Analytics

Qualitative

=

intuition + analysis

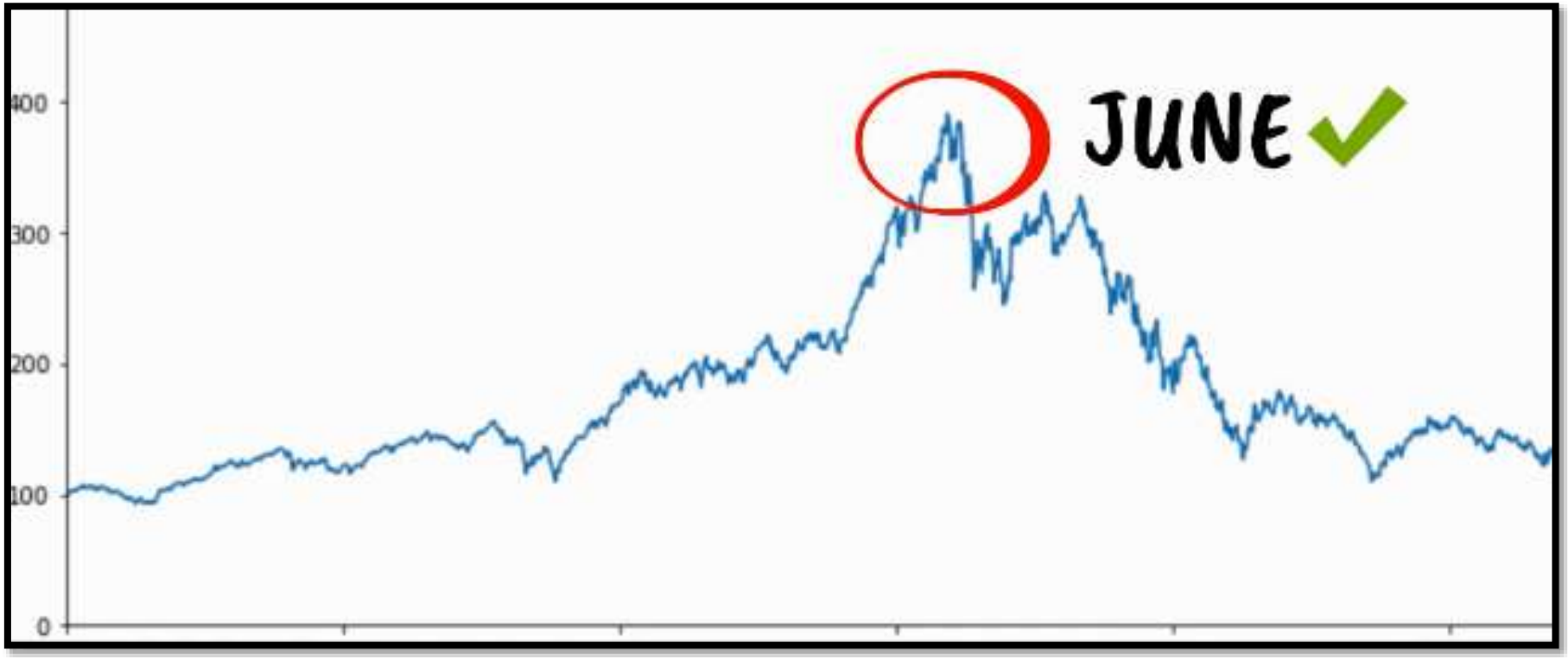
Quantitative

=

formulas + algorithms


QUALITATIVE ANALYTICS

Quantitative Analytics



Analysis

Analysis

Qualitative

Explain

How?

Why?

Quantitative

data + how sales decreased last summer

Month	($) Sales
January	9,048.12
February	9,875.55
March	10,050.11
April	10,997.93
May	11,253.26
June	11,522.56
July	12,500.35
August	11,511.08
September	10,551.10
October	9,900.65
November	8,000.50
December	7,750.25
TOTAL	122,961.46

Analysis & Analytics

Analysis ≠ Analytics

data analysis ≠ data analytics

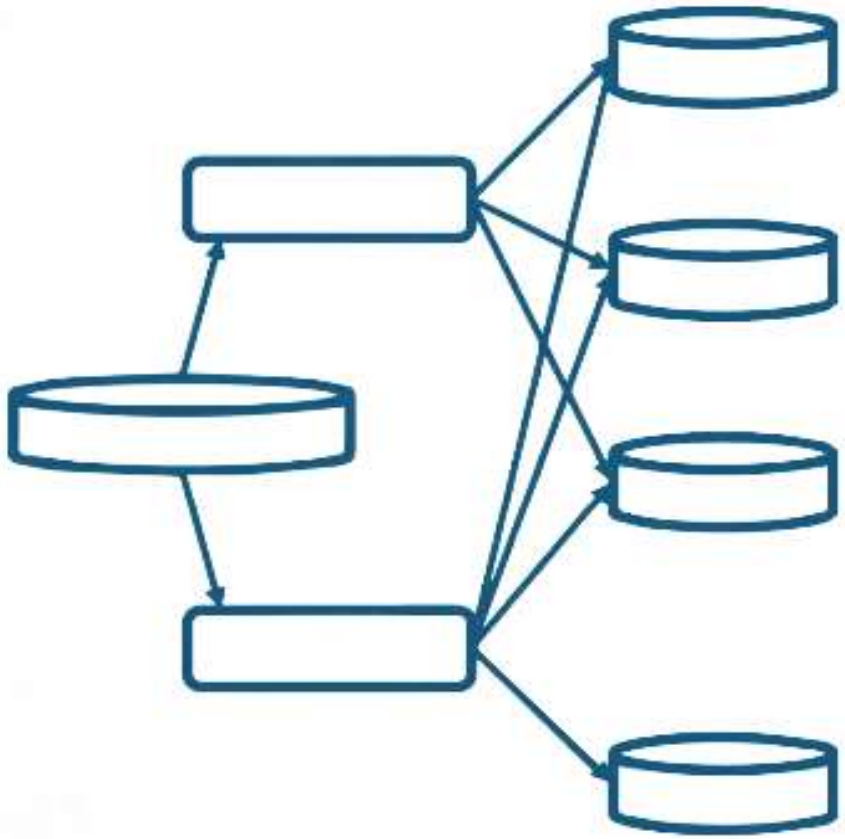
business analysis ≠ business analytics

Traditional Data

- structured
-  can be managed from 1 computer

ID	Name	Age	.....
001	John	35	.....
002	Alan	22	.....
.....	.....	.....	.....

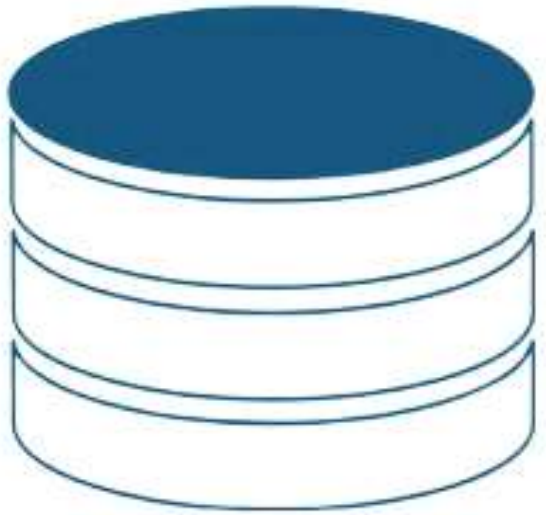
Big Data



- structured
- semi - structured
- unstructured

Traditional Data Vs. Big Data

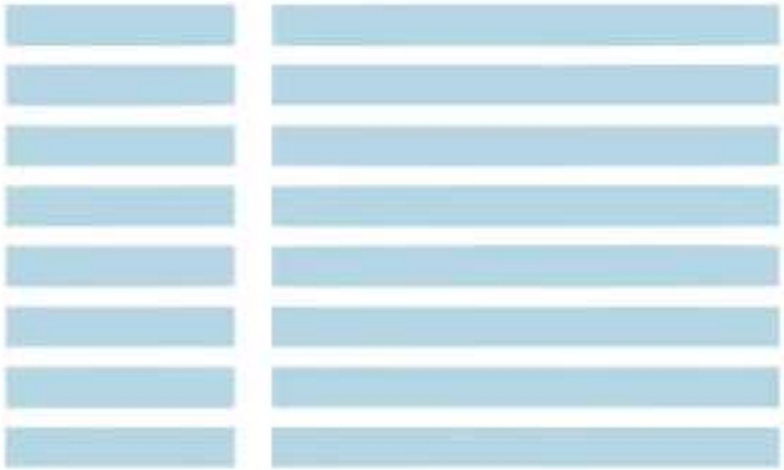
Volume:



: distributed between many computers

Traditional Data Vs. Big Data

Variety:



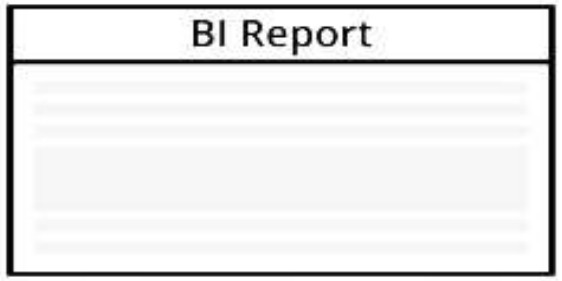
Traditional Data Vs. Big Data

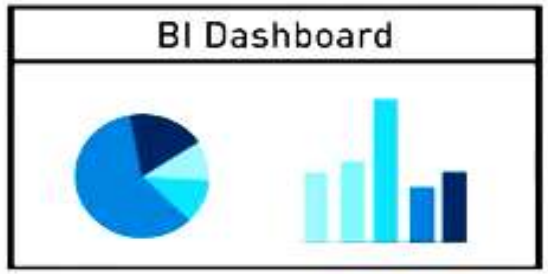
Velocity:

retrieved in real-time

Business Intelligence (BI)

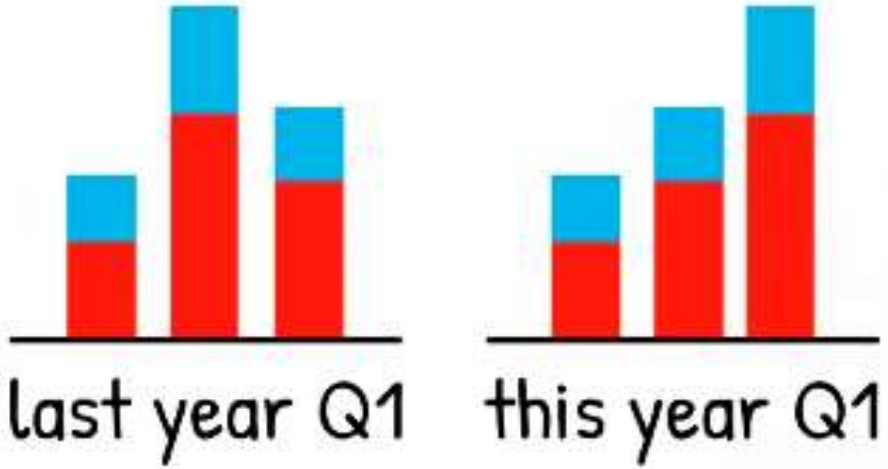
includes all technology-driven tools involved in the proess of **analyzing**, **understanding** and **reporting** available past data





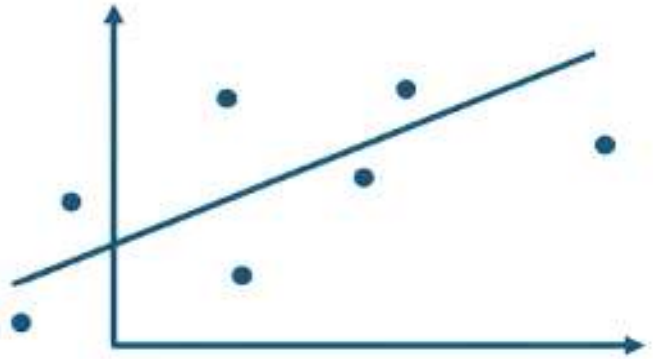
- make decisions
- extract insights
- extract ideas

Business Intelligence (BI)

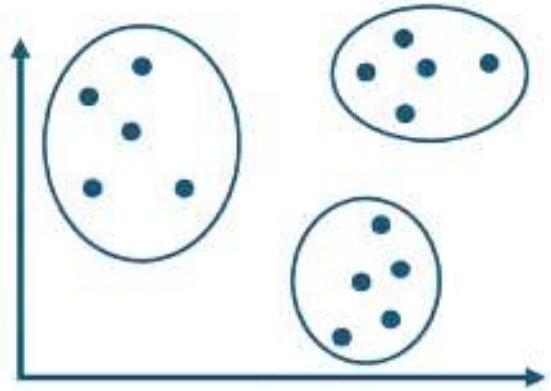


Traditional Methods

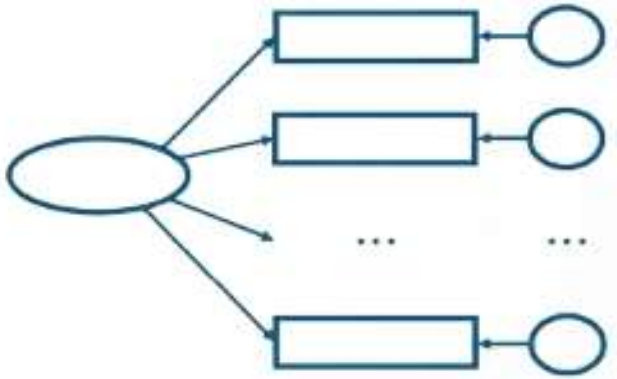
perfect for forecasting future performance with great accuracy



regression

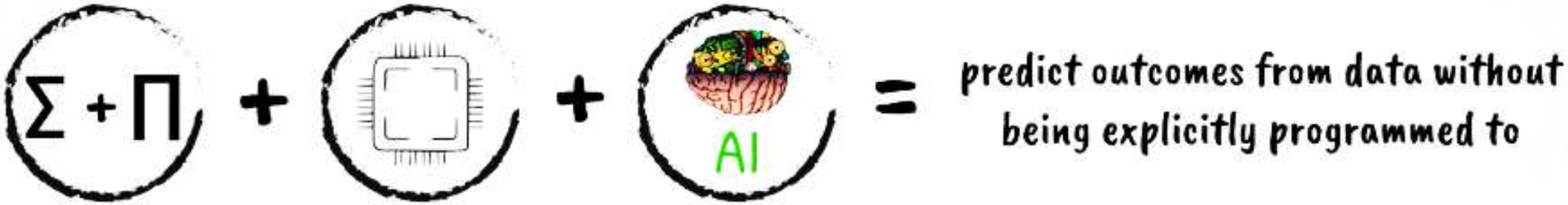


cluster

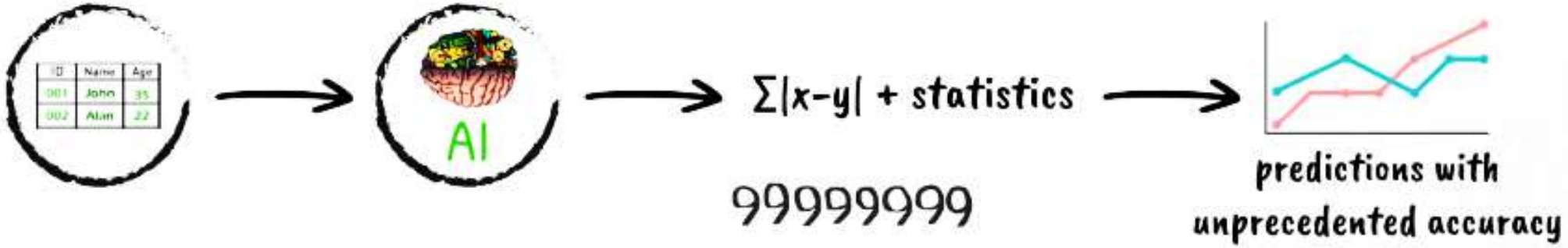


factor

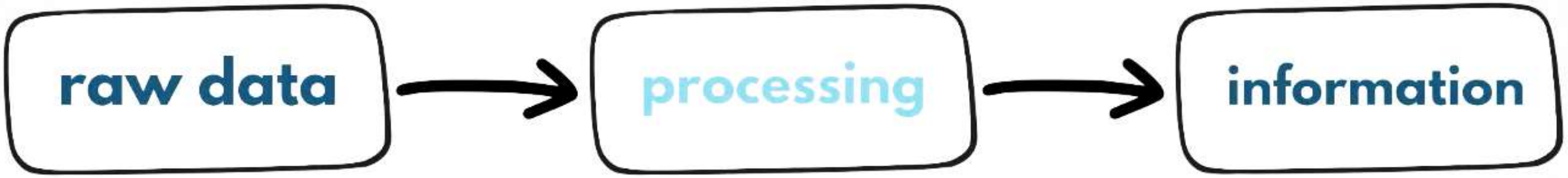
Machine Learning



Algorithm:



Traditional data


raw data
processing
information

Data = raw facts?

Data = processed data?

Data = information?

What is raw data?

raw data

- cannot be analysed straight away

= raw facts

- it is untouched data you have accumulated and stored on the server

= primary data

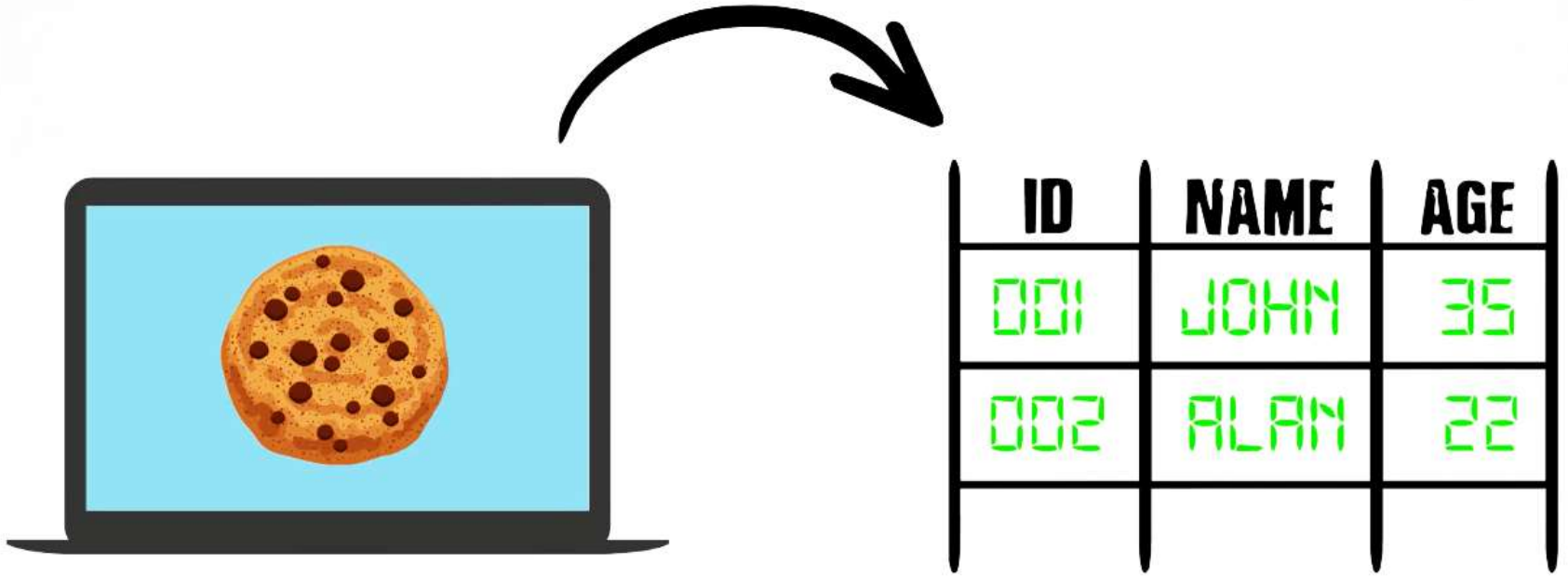
- data collection

Data Collection

On a scale of 1 to 10, how much do you like our product?

1 2 3 4 5 6 7 8 9 10

Data Collection



Data Pre-Processing

ID	NAME	AGE
001	JOHN	932
...	...	...
008	UNITED KINGDOM	24

Class Labeling

numerical

number of goods sold

9,365 units

! can be manipulated

categorical

New York, USA

! cannot be manipulated

Data Cleansing

New York

Nwe York

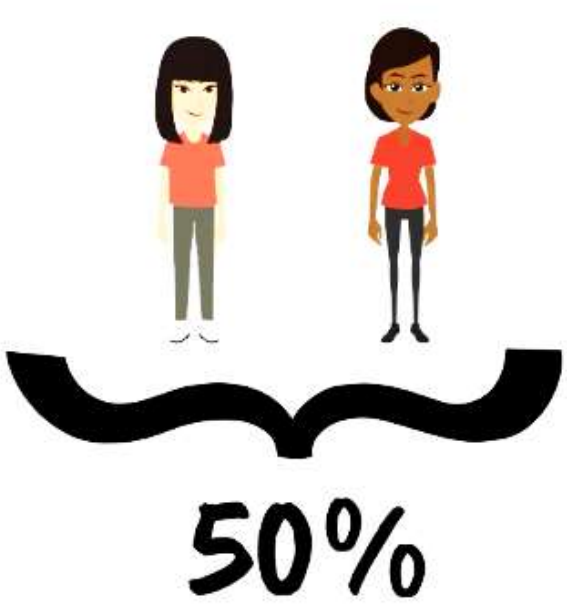
Washington DC

Washington CD

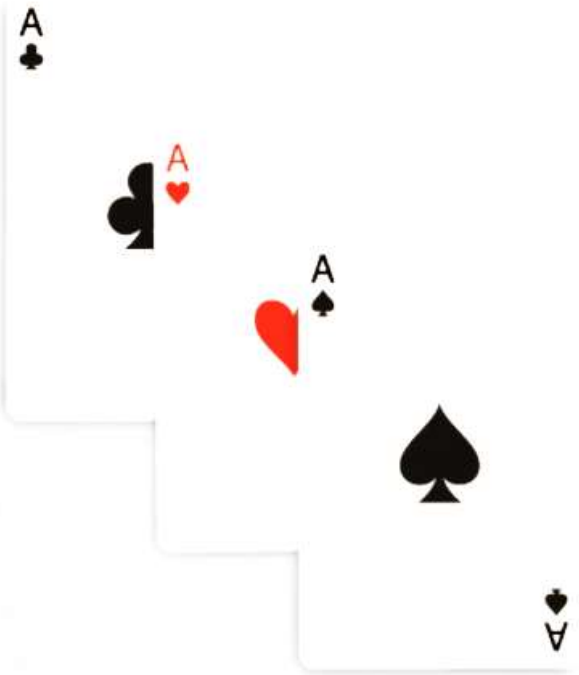
Dealing With Missing Values

ID	NAME	AGE	OCCUPATION
001	JOHN	?	DATA SCIENTIST
002	ALAN	35	ACCOUNTANT

Balancing



Data Shuffling



- prevents unwanted patterns
- improves predictive performance
- helps avoid misleading results

Basics Customer Data

numerical VS. categorical

Customers

customer_id	first_name	last_name	email_address	number_of_complaints
1	John	McKinley	john.mackinley@365careers.com	0
2	Elizabeth	McFarlane	e.mcfarlane@365careers.com	2
3	Kevin	Lawrence	kevin.lawrence@365careers.com	1
4	Catherine	Winnfield	c.winnfield@365careers.com	0

no numerical value
categorical data

TOTAL = 3
numerical data

Historical Stock Price Data

	PG
Date	
2007-01-03	46.149067
2007-01-04	45.798710
2007-01-05	45.405422
2007-01-08	45.505543
2007-01-09	45.391144
2007-01-10	45.934578
2007-01-11	46.220585
2007-01-12	46.478012
2007-01-16	46.478012
2007-01-17	46.959393
2007-01-18	47.088707

stock prices

categorical data

numerical data

Big data

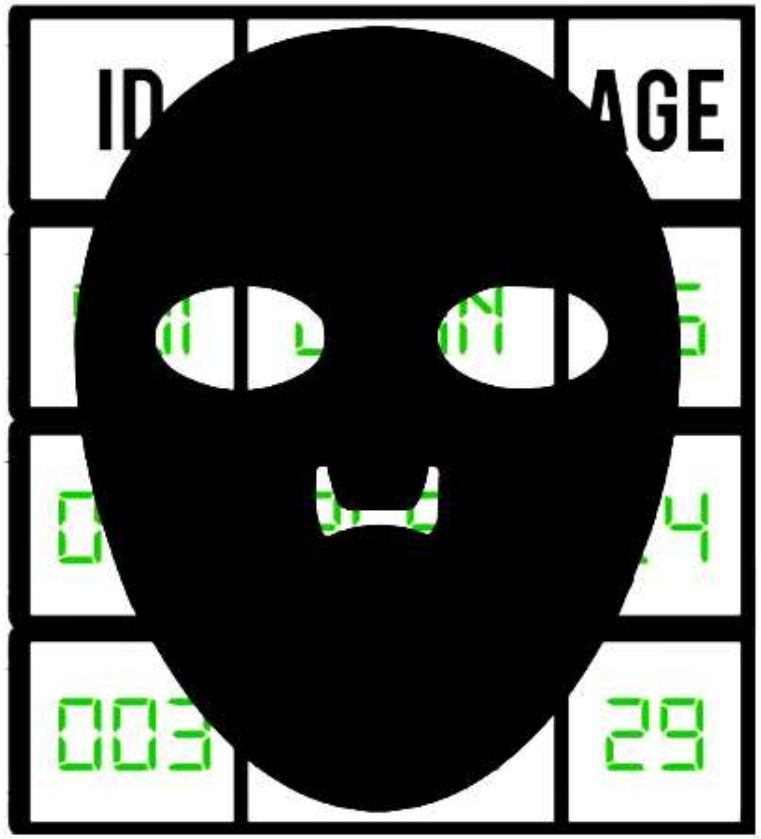
Data Cleansing

Text Data Mining

- the process of deriving valuable, unstructured data from a text

Tect Data Mining

Data Masking



Data Masking

- conceals the original data with random and false data
- conduct analysis
- keep all confidential information in a secure place

Example: confidentiality preserving data mining

Big Data – Social Media

ID	NAME	AGE
001	JOHN	35
002	ALAN	24
003	JANE	29

Big Data – Financial Trading Data

- record the stock price every second

Business Intelligence

Business Intelligence

data skills + business knowledge & intuition

explain past performance

What happened?

When did it happen?

How many units did we sell?

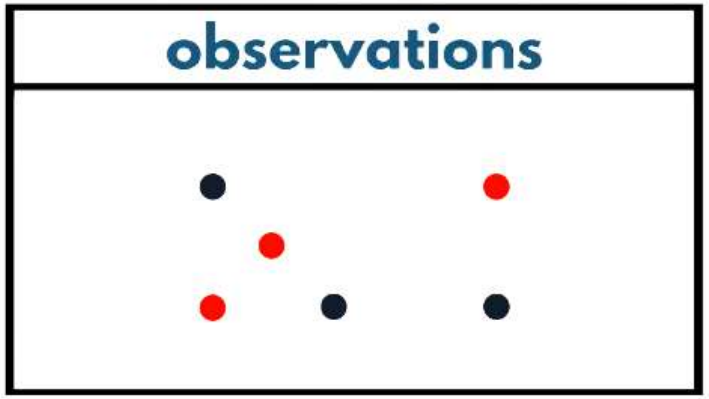
In which region did we sell the most goods?

Business Intelligence

How did our **email marketing** perform last quarter in terms of click-through rates and revenue generated ?

And how does that compare to the performance in the same quarter of last year?

Observation

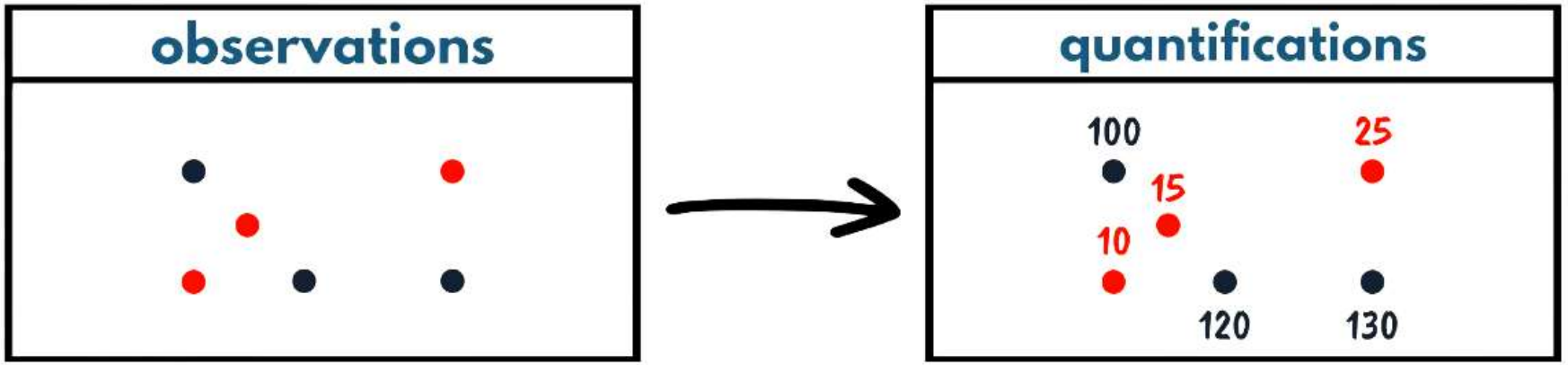


- monthly revenues
- customers

observe :

- sales volume
- new customers

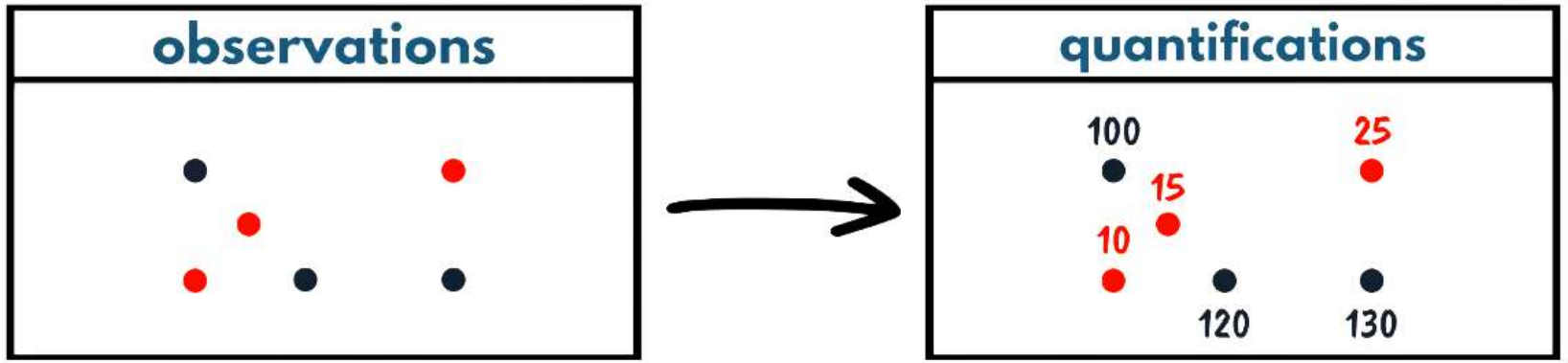
Quantification

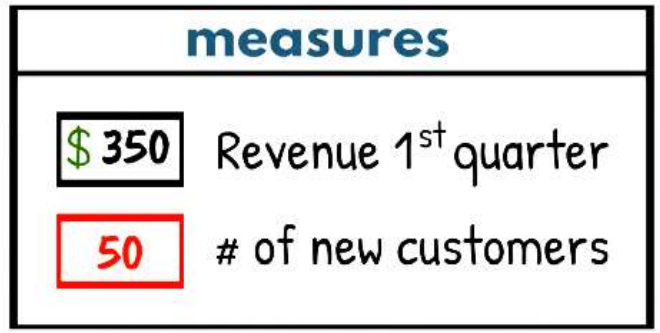


- monthly revenues
- customers

the process of representing observations as numbers

Measures





the accumulation of observations to show some information

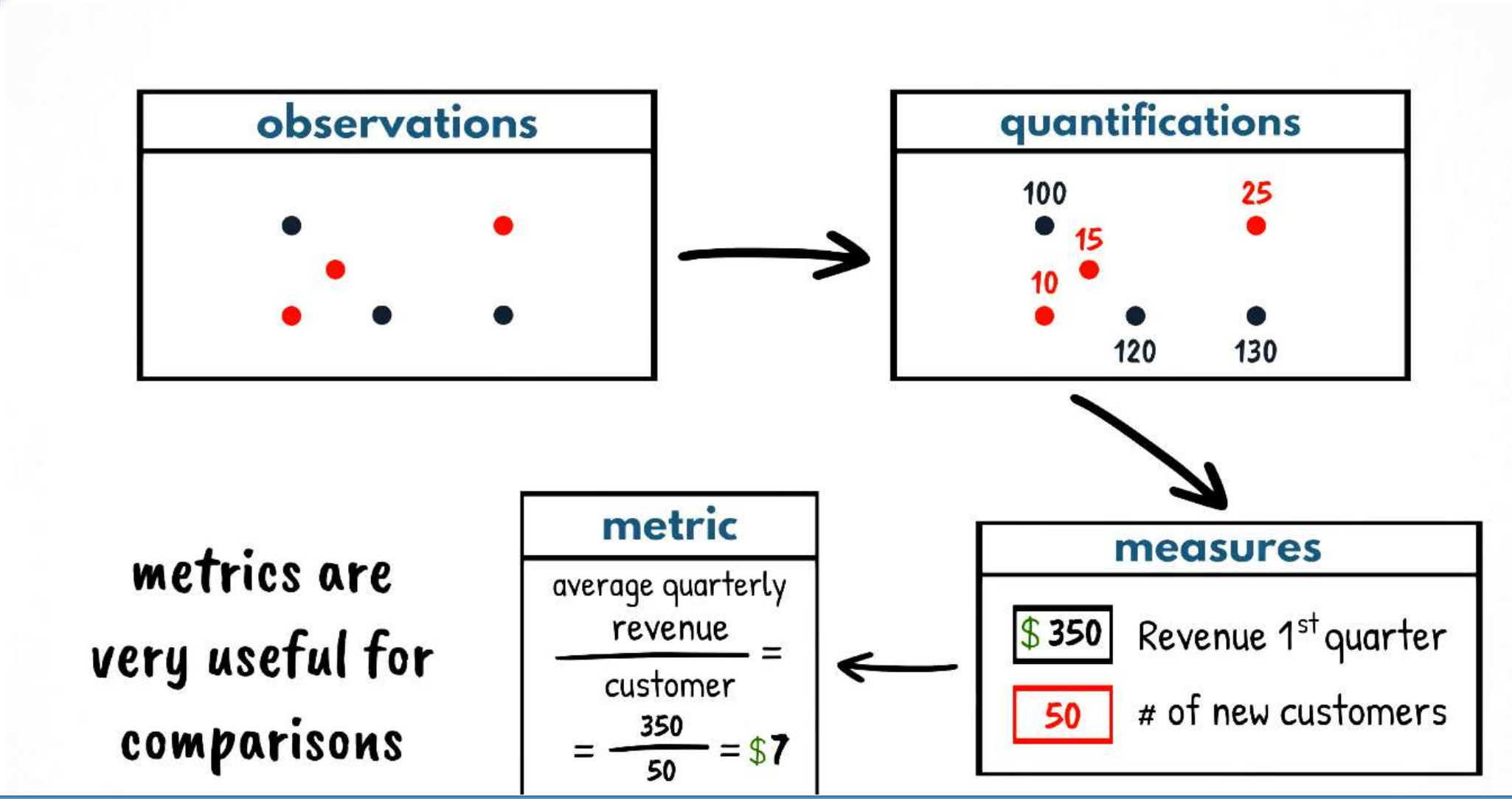
Metrics

aims at gauging business performance or progress

MEASURE - related to a simple descriptive statistics of past performance

METRIC = MEASURE + BUSINESS MEANING

Metrics


observations
quantifications
100
25
15
10
120
130
measures
$ 350
Revenue 1st quarter
50
# of new customers
metric
average quarterly revenue / customer = 350 / 50 = $7
metrics are very useful for comparisons

Key Performance Indicators (KPI)

Can we keep track of all possible metrics we can extract from a data set? ✓

Does it make sense to do that? ✗

KPIs = metrics + business objectives

= Key Performance Indicators

Key Performance Indicators (KPI)

KEY

related to your main business goals

PERFORMANCE

how successfully you have performed within a specified timeframe

INDICATORS

generated only from users who have clicked on a link provided in your ad campaign

Metric Vs. KPI

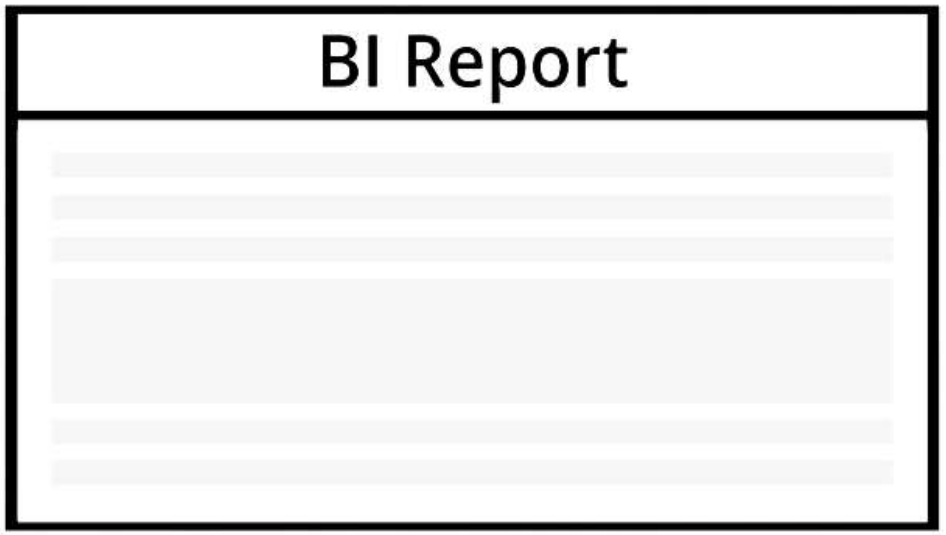
metric

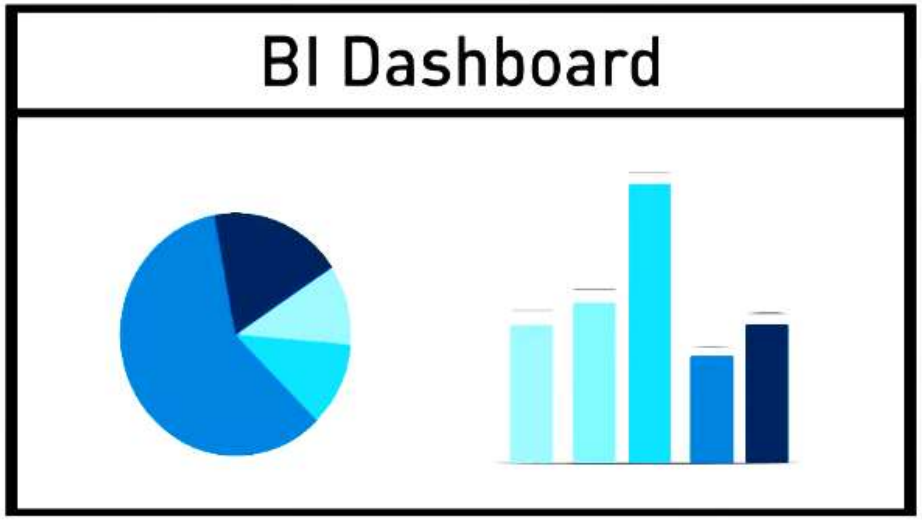
- the traffic of a page from your website that was visited by any type of user

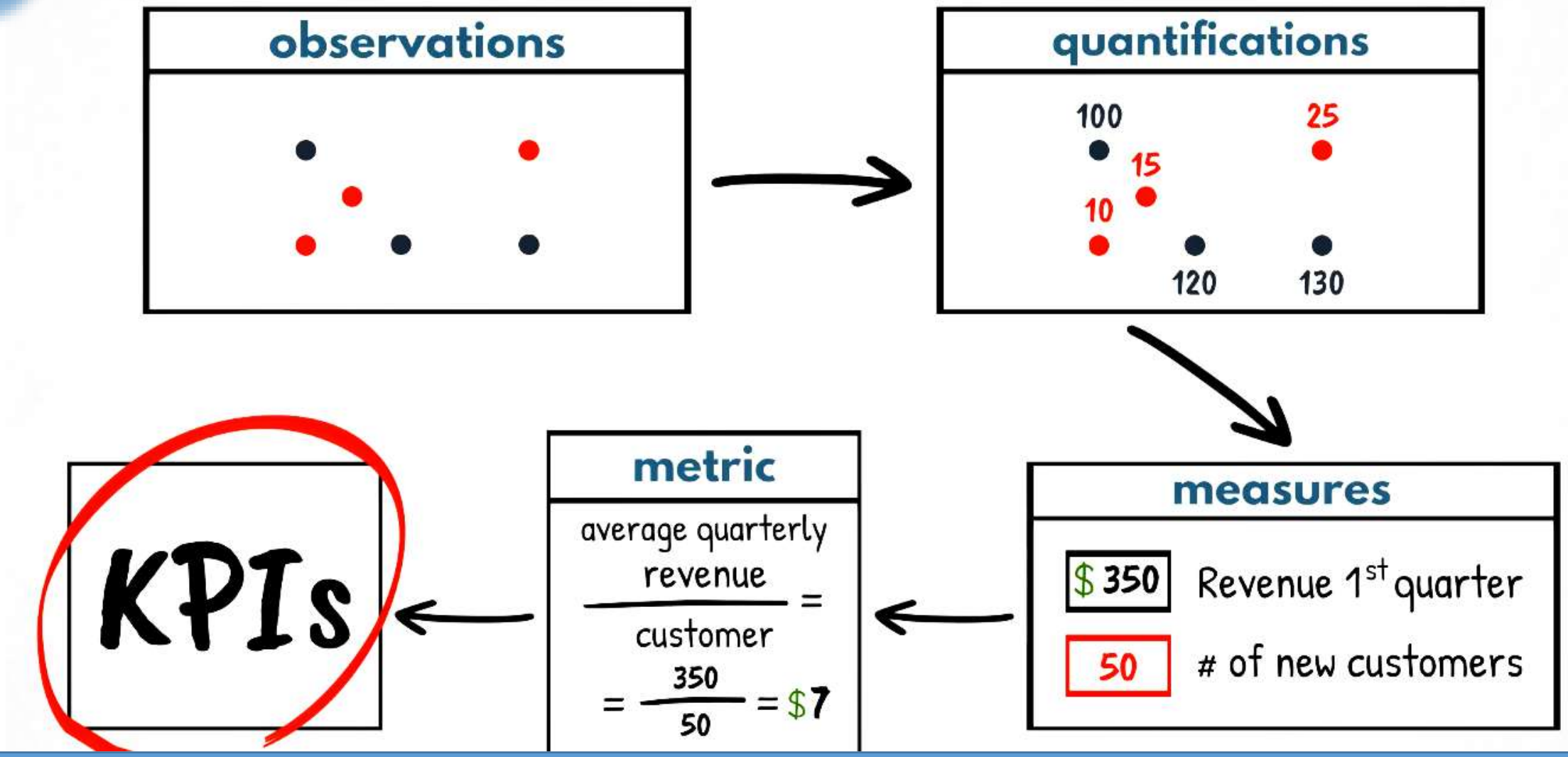
KPI

- the traffic generated only from users who have clicked on a link provided in your ad campaign

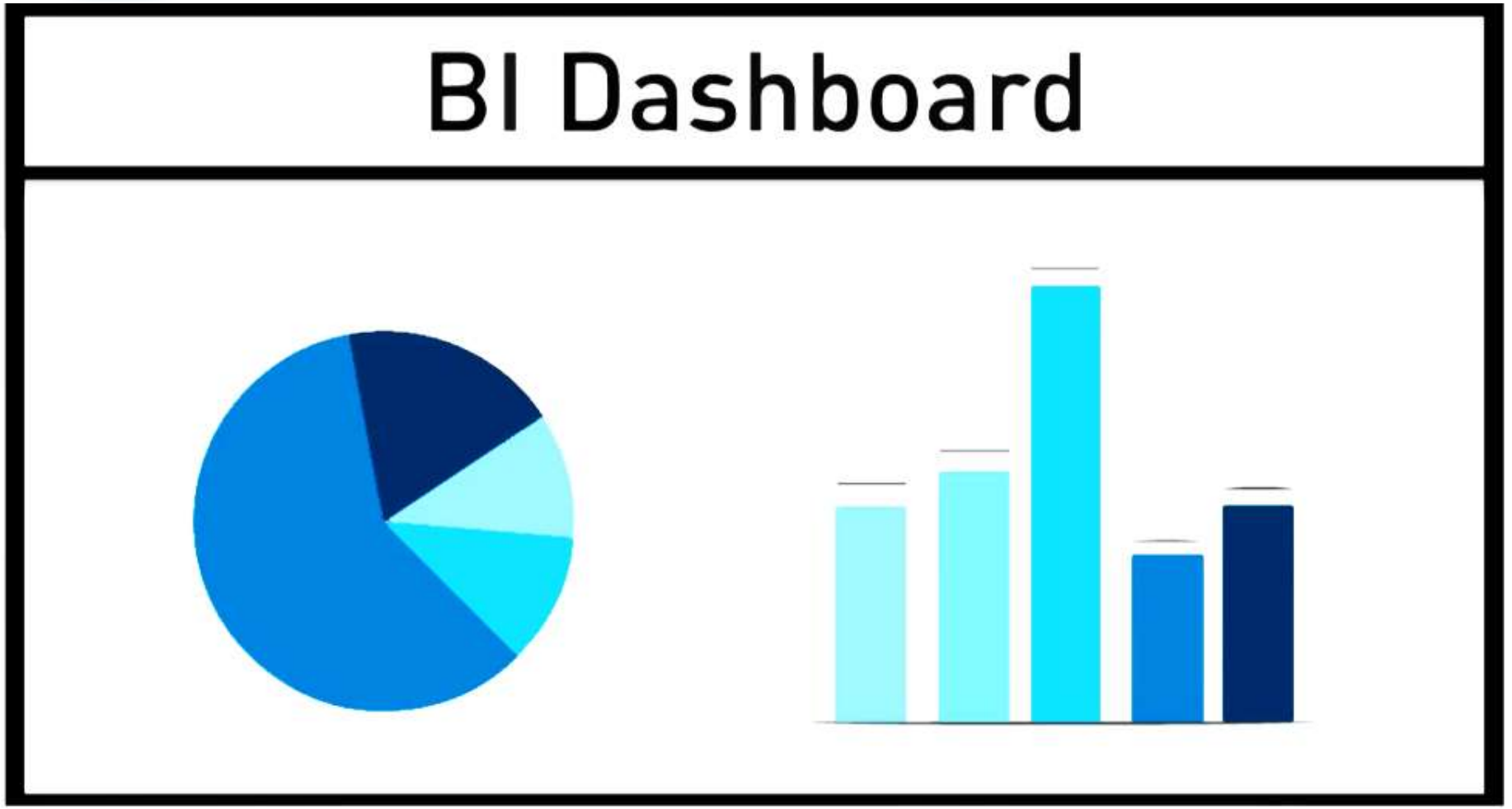
Representing the KPI in a form of Dashboard






observations
quantifications
100
25
15
10
120
130
measures
$ 350
Revenue 1st quarter
50
# of new customers
metric
average quarterly revenue / customer = 350 / 50 = $7
KPIs

Keeping the KPI for the business objective only



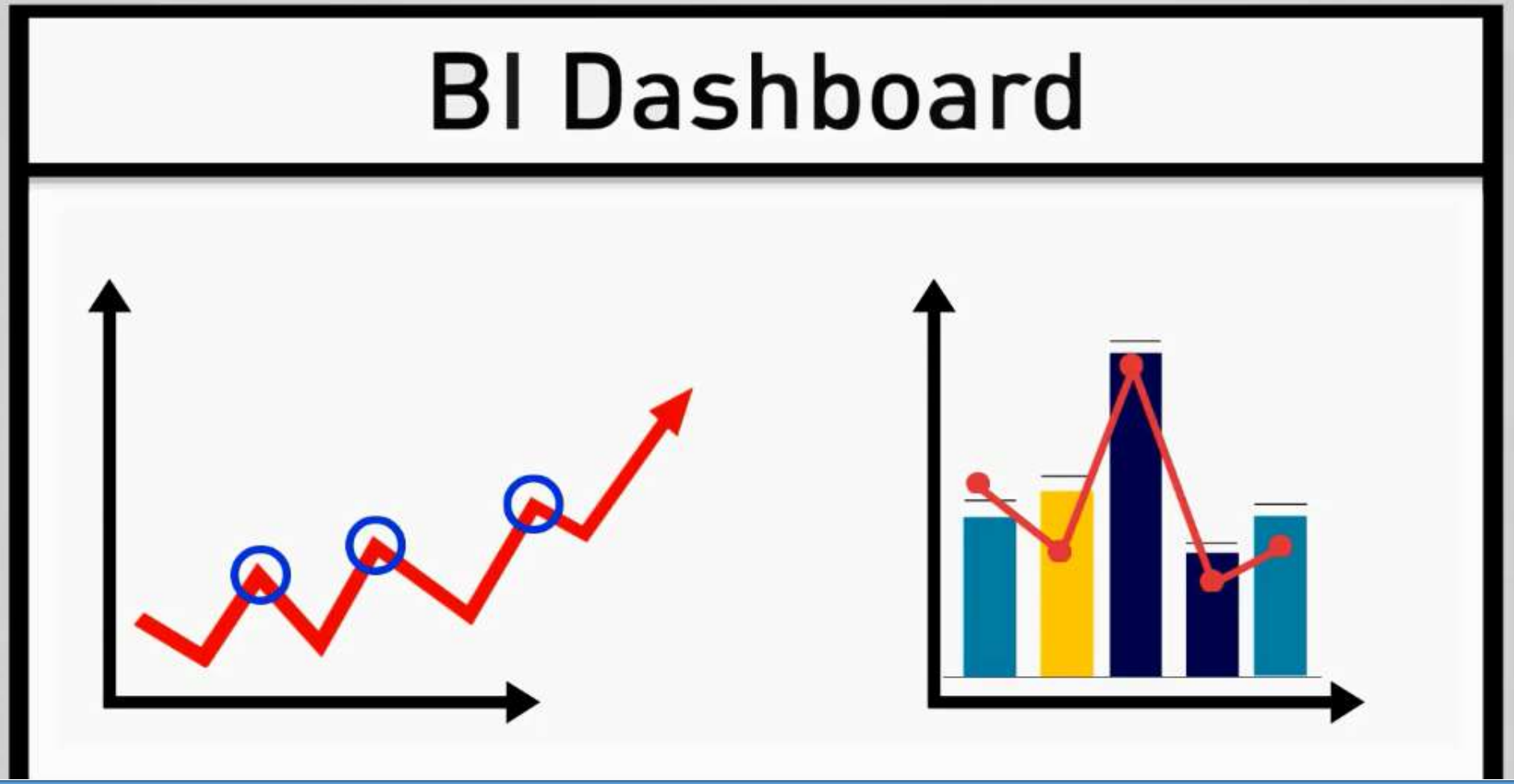
PRICE OPTIMISATION

PRICE OPTIMISATION

Technology Academy

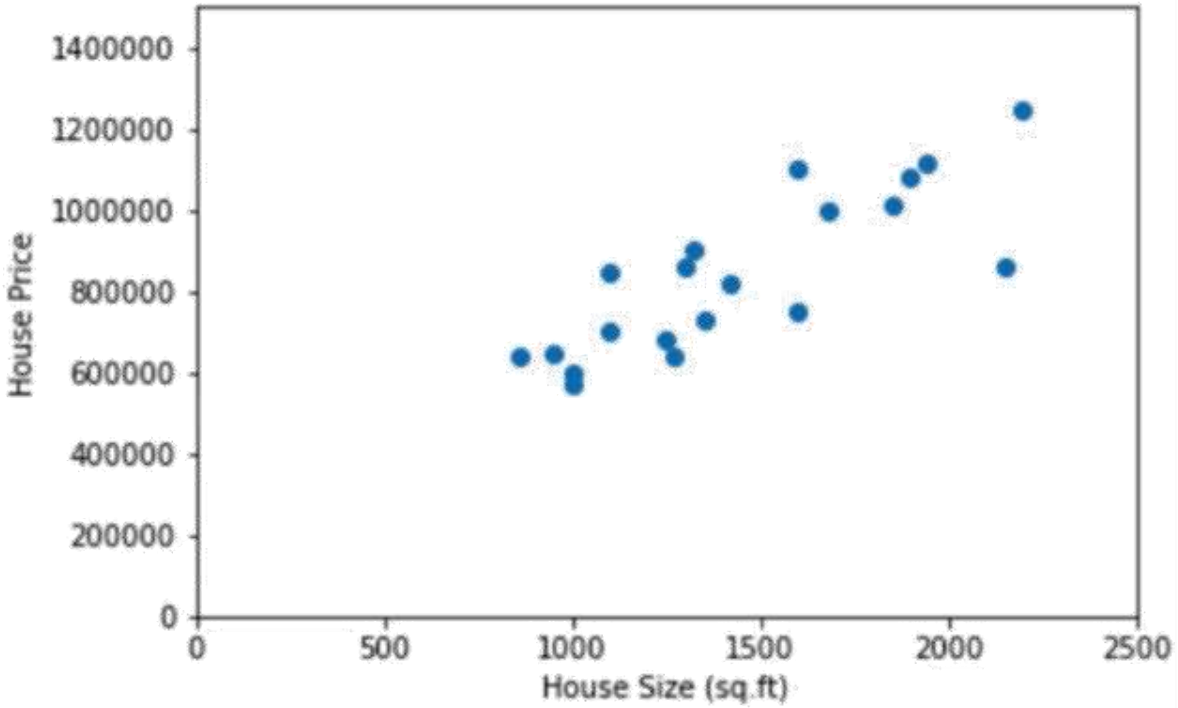
INVENTORY MANAGEMENT

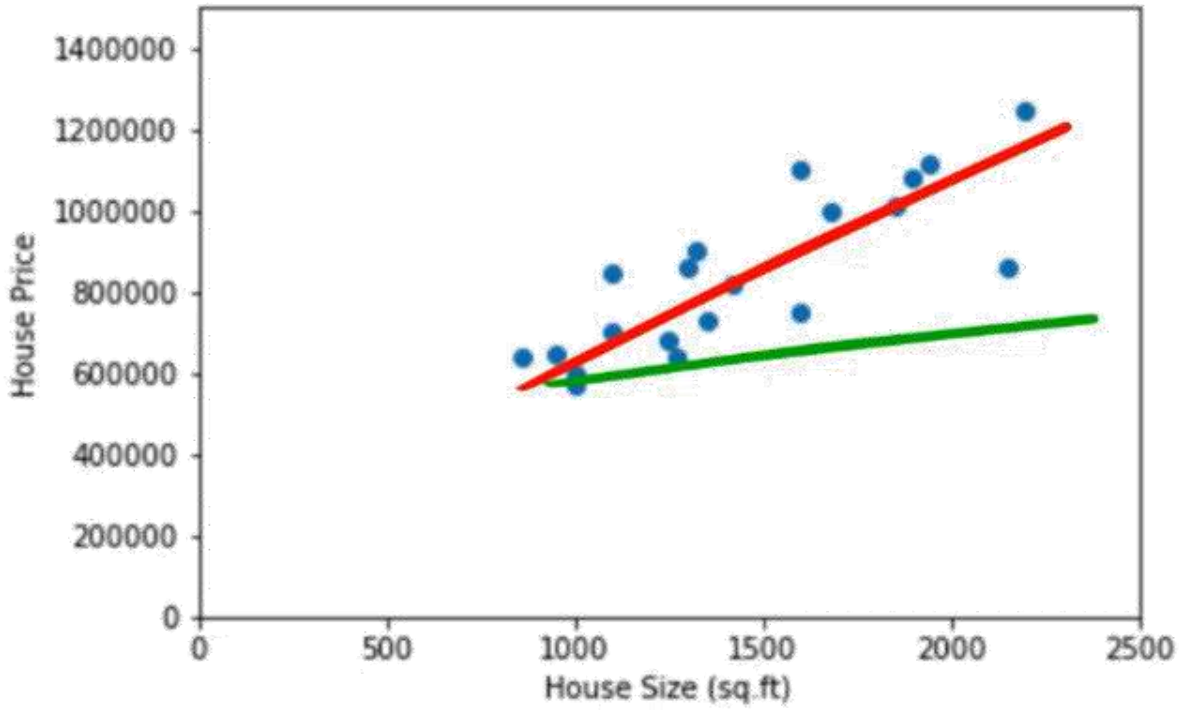
INVENTORY MANAGEMENT



Real state data

	House Price ($)	House Size (sq.ft.)
0	1116000	1940
1	860000	1300
2	818400	1420
3	1000000	1680
4	640000	1270
5	1010000	1850
6	600000	1000
7	700000	1100
8	1100000	1600
9	570000	1000
10	860000	2150
11	1085000	1900
12	1250000	2200


1400000
1200000
1000000
800000
600000
400000
200000
0
House Price
0
500
1000
1500
2000
2500
House Size (sq.ft)


1400000
1200000
1000000
800000
600000
400000
200000
0
House Price
0
500
1000
1500
2000
2500
House Size (sq.ft)

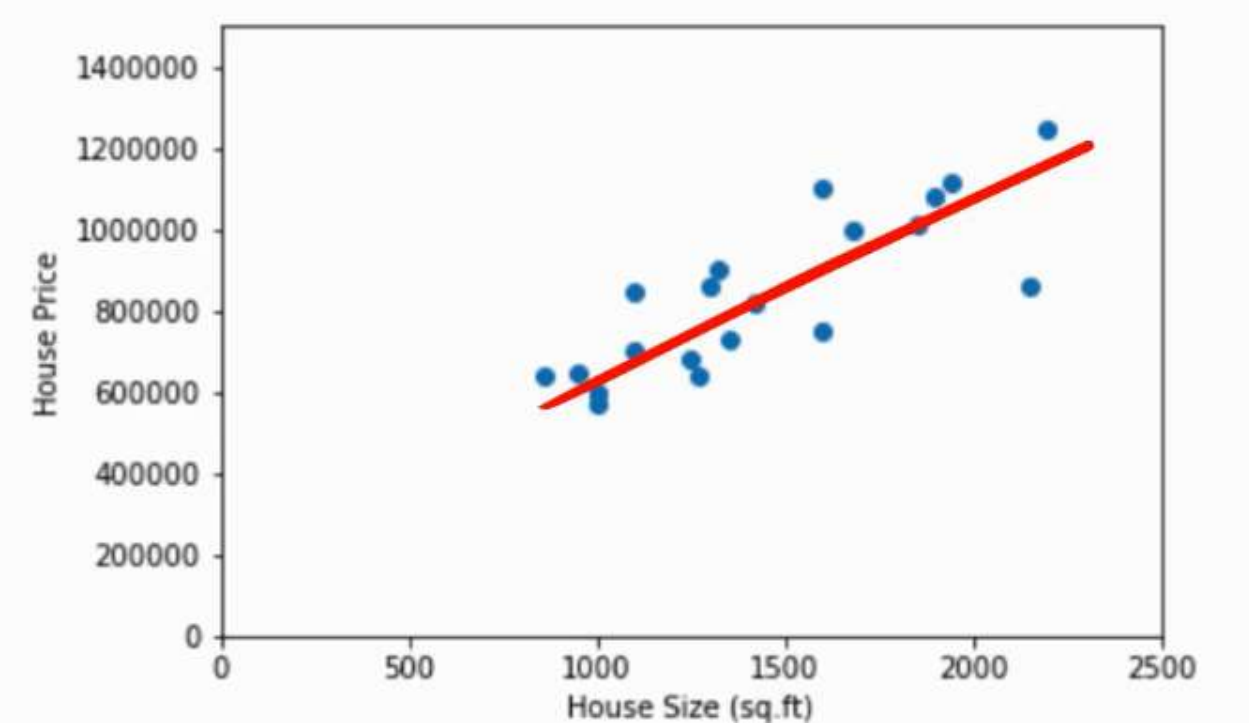
Linear Regression

$$y = bx$$

y - house price

b - coefficient

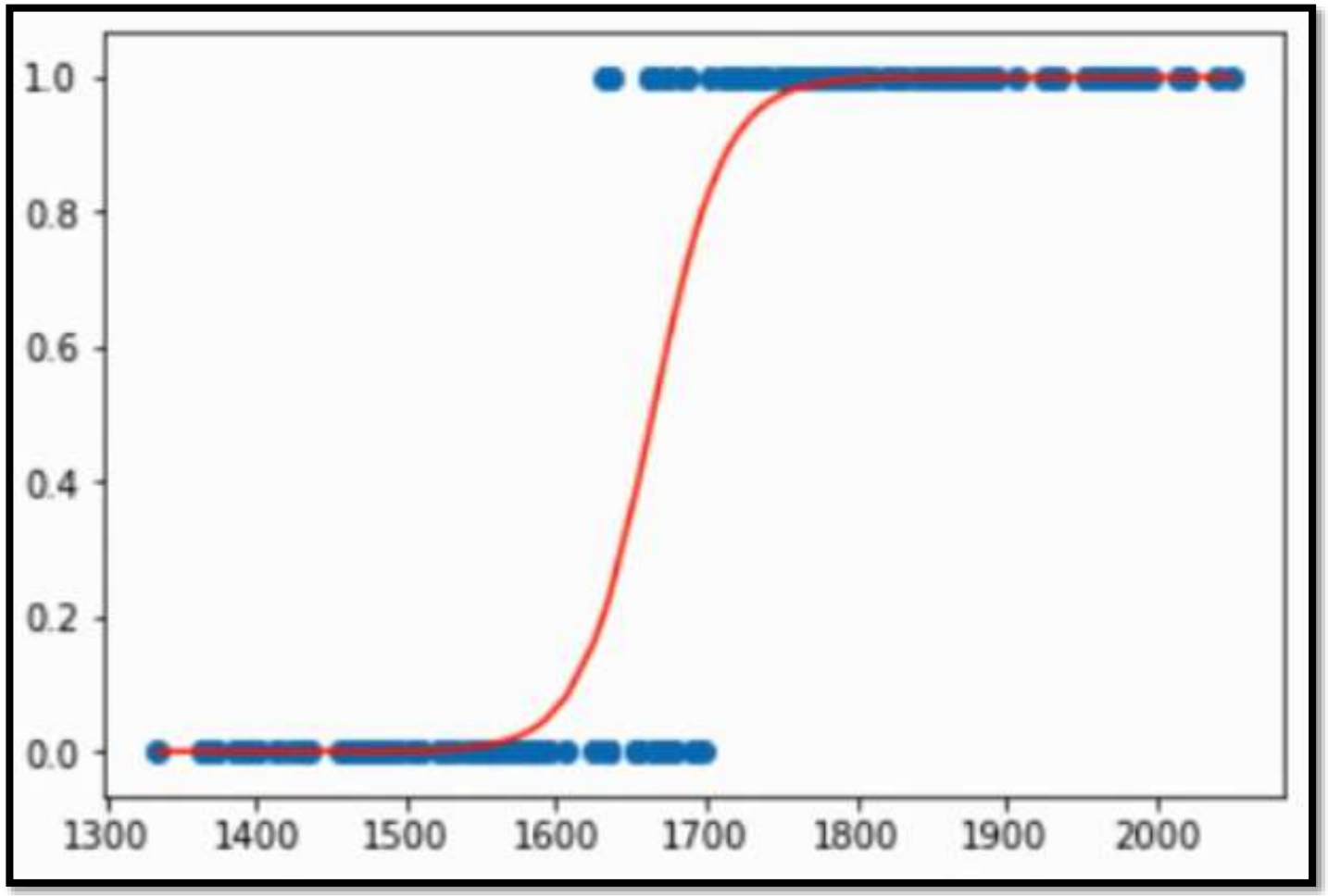
x - house size



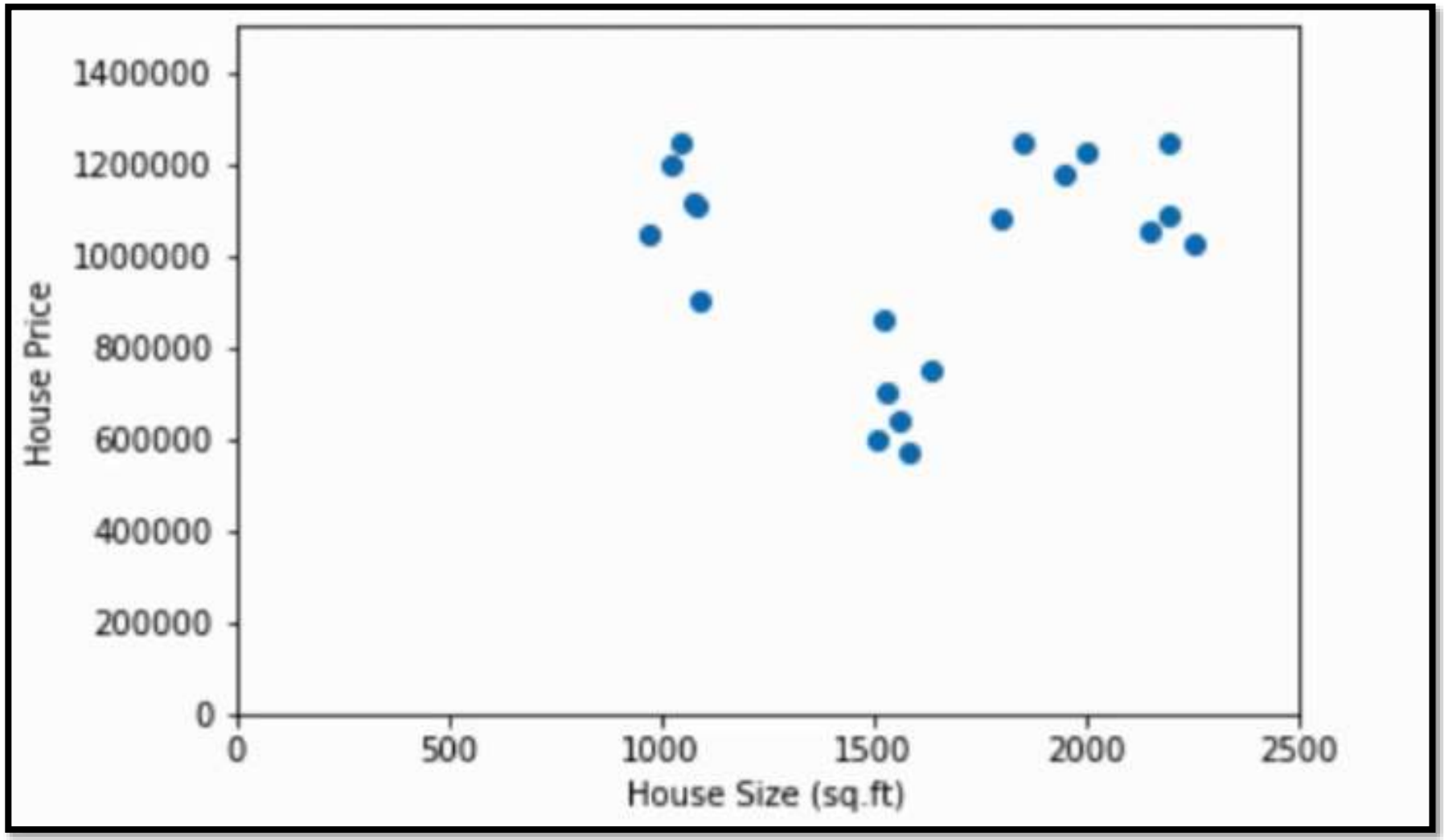
Data Science

Logistic Regression – HR Job Candidate Filtering

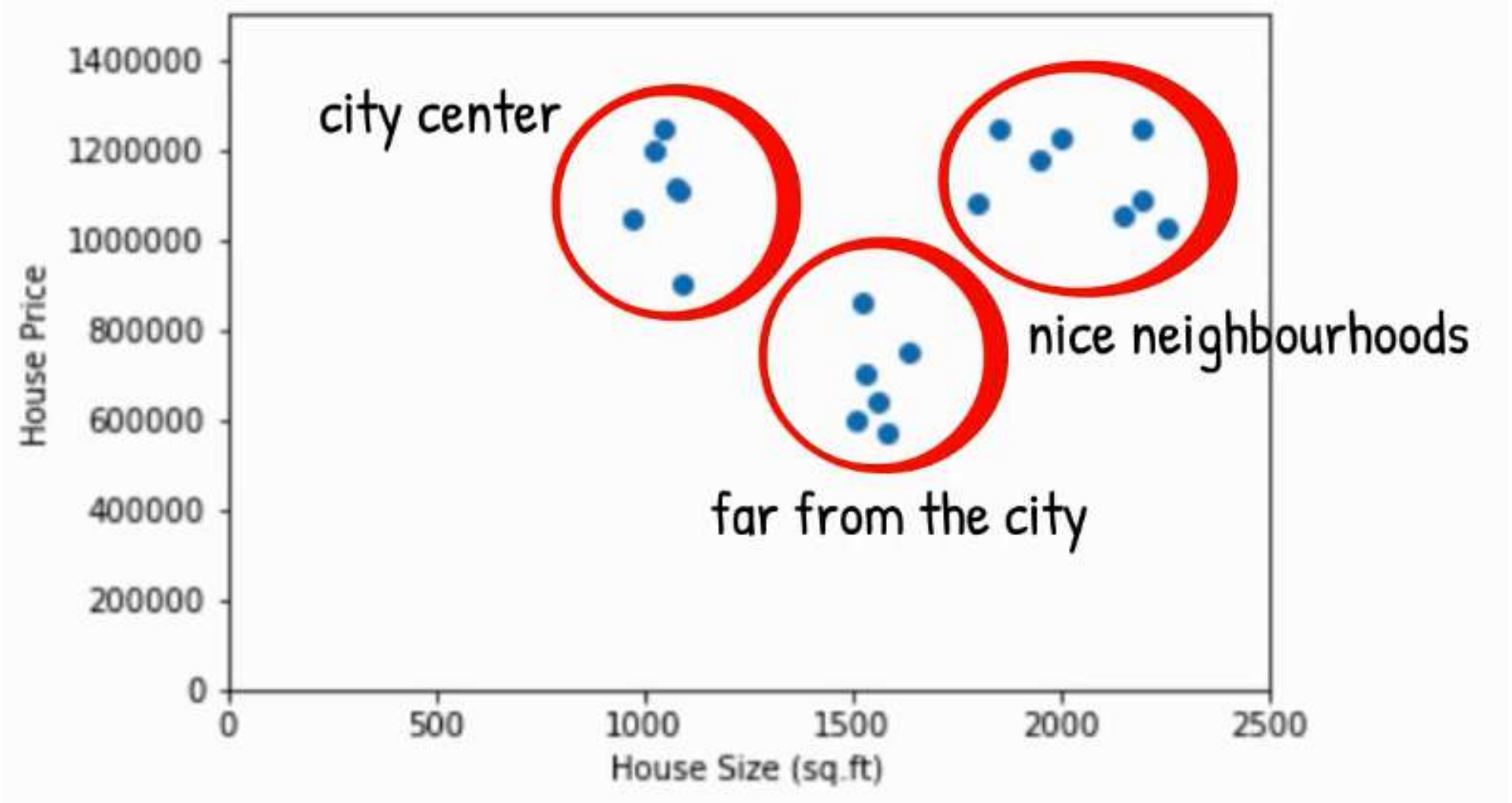
Logistic Regression Graph



Cluster Analysis



Cluster Analysis

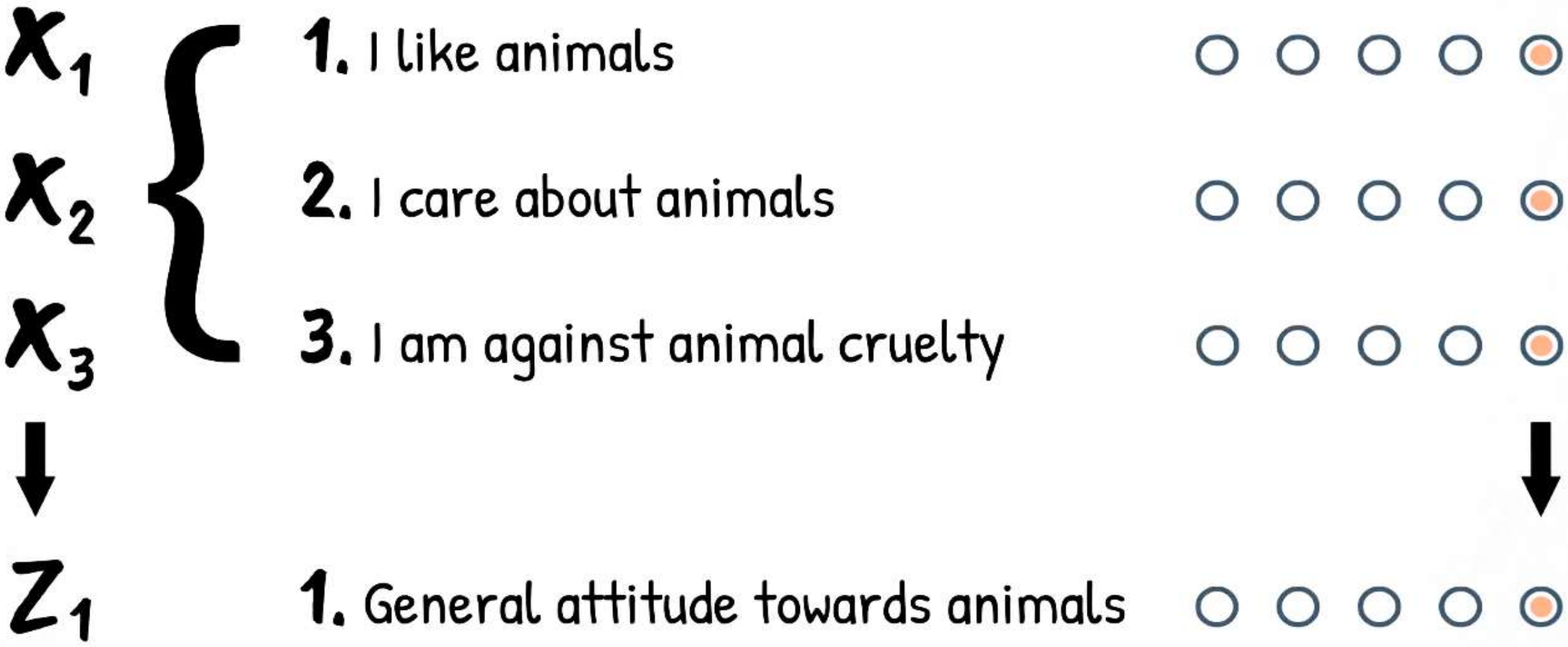


	House Price ($)	House Size (sq.ft.)	State	Number of Rooms	Year of Construction
0	1116000	1940	IN	8	2002
1	860000	1300	IN	5	1992
2	818400	1420	IN	6	1987
3	1000000	1680	IN	7	2000
4	640000	1270	IN	5	1995
5	1010000	1850	IN	7	1998
6	600000	1000	IN	4	2015
7	700000	1100	LA	4	2014
8	1100000	1600	LA	7	2017
9	570000	1000	NY	5	1997
10	860000	2150	NY	9	1997
11	1085000	1900	NY	9	2000
12	1250000	2200	NY	9	2014

Analysing a survey of 100 questions

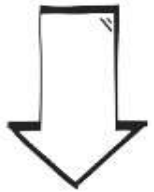
$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_{100} x_{100}$$

100 questions? → factor analysis


x_1
x_2
x_3
1. I like animals
2. I care about animals
3. I am against animal cruelty
z_1
1. General attitude towards animals

$x_1 \rightarrow x_{100}$ (100 variables):

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_{100} x_{100}$$



$z_1 \rightarrow z_{10}$ (10 factors):

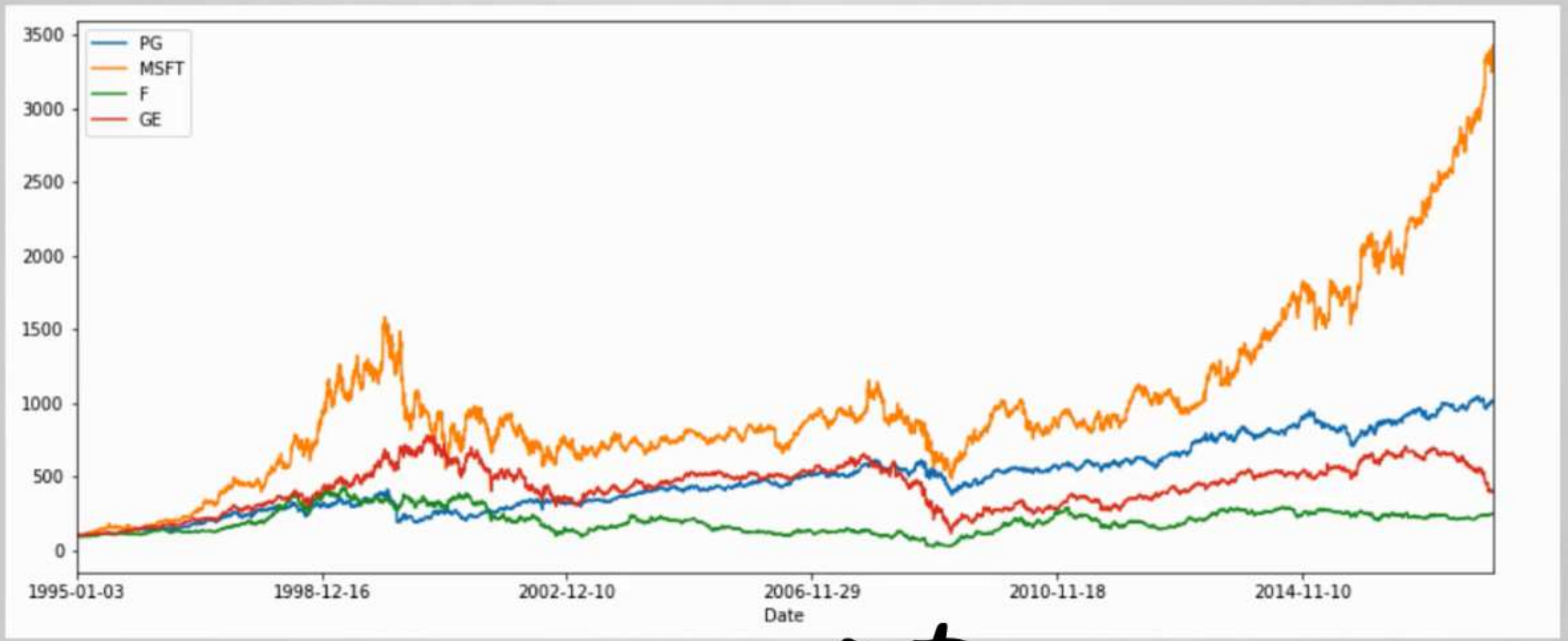
$$y = n + n_1 z_1 + n_2 z_2 + n_3 z_3 + \ldots + n_{10} z_{10}$$

Time Series

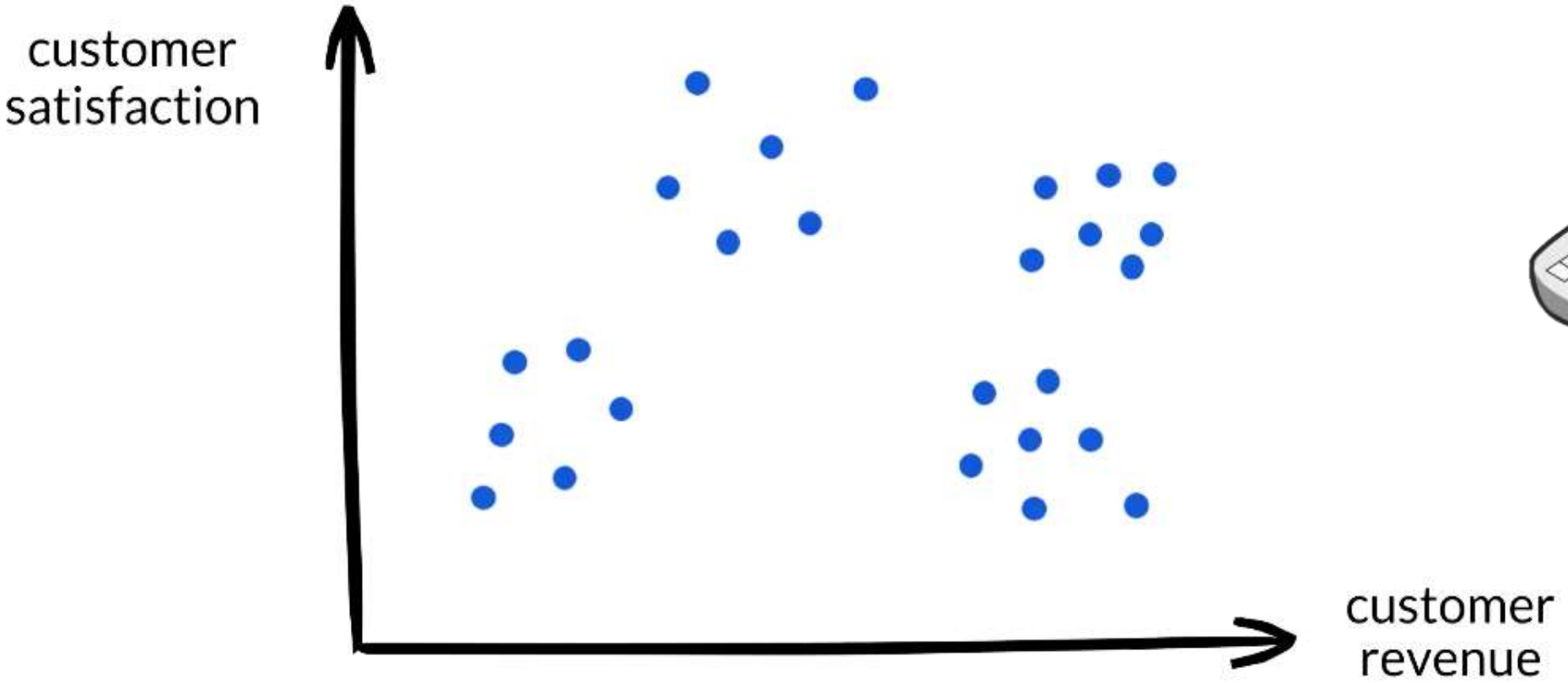
- stock price
- sales volume

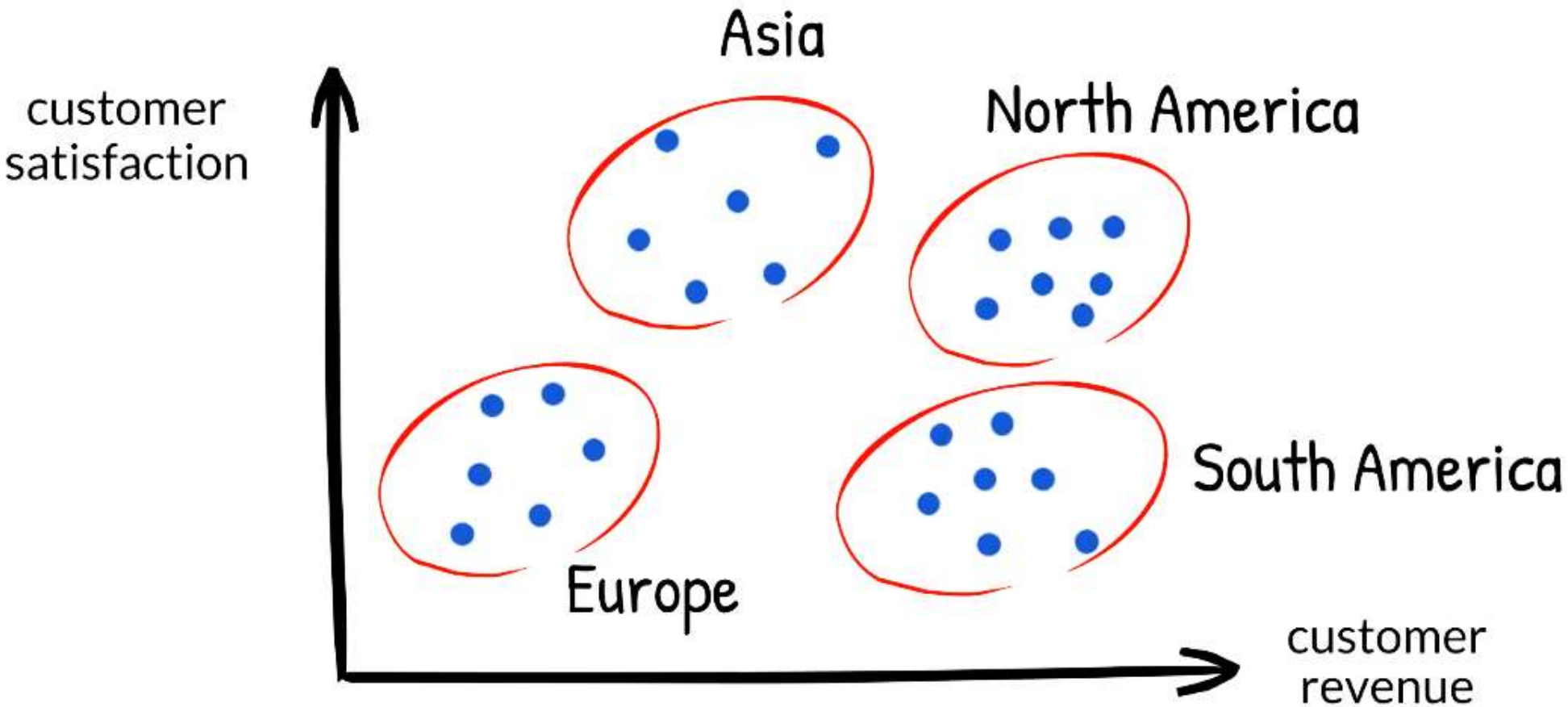
Data Science

Example : checking which stock perform better over time

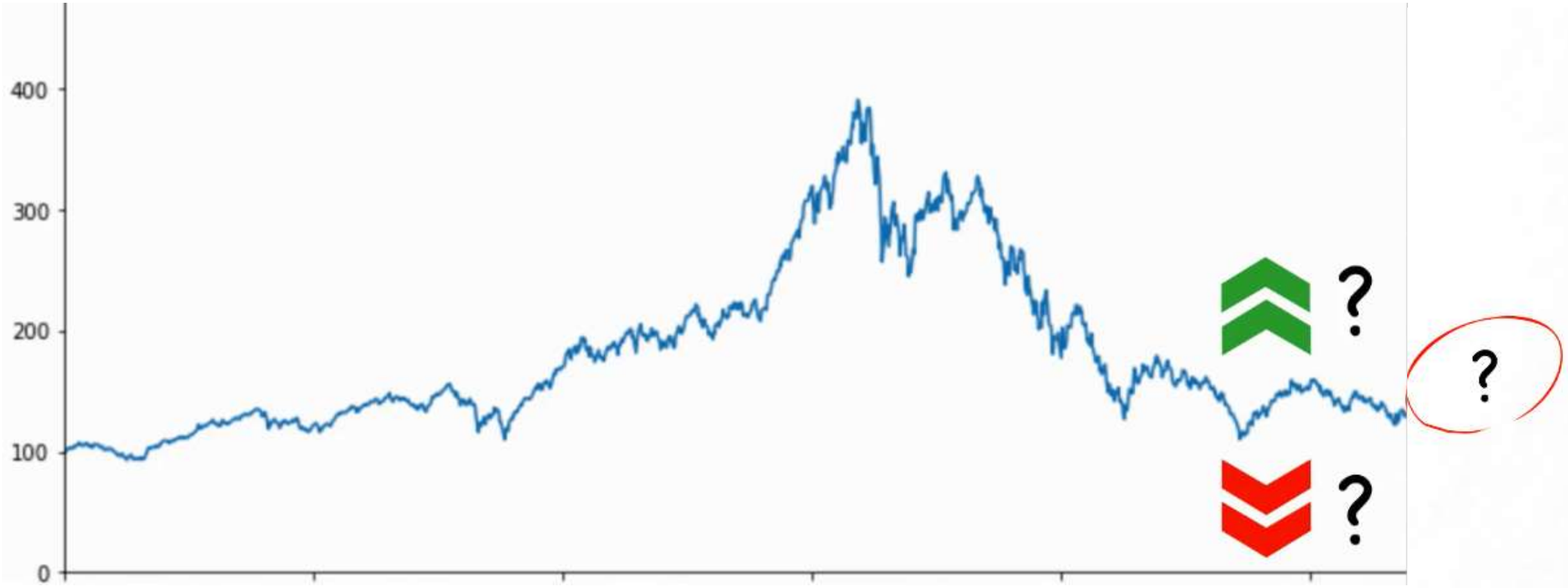


Survey : User Experience (UX)




Asia
North America
customer satisfaction
South America
Europe
customer revenue

Forecasting Sales Data



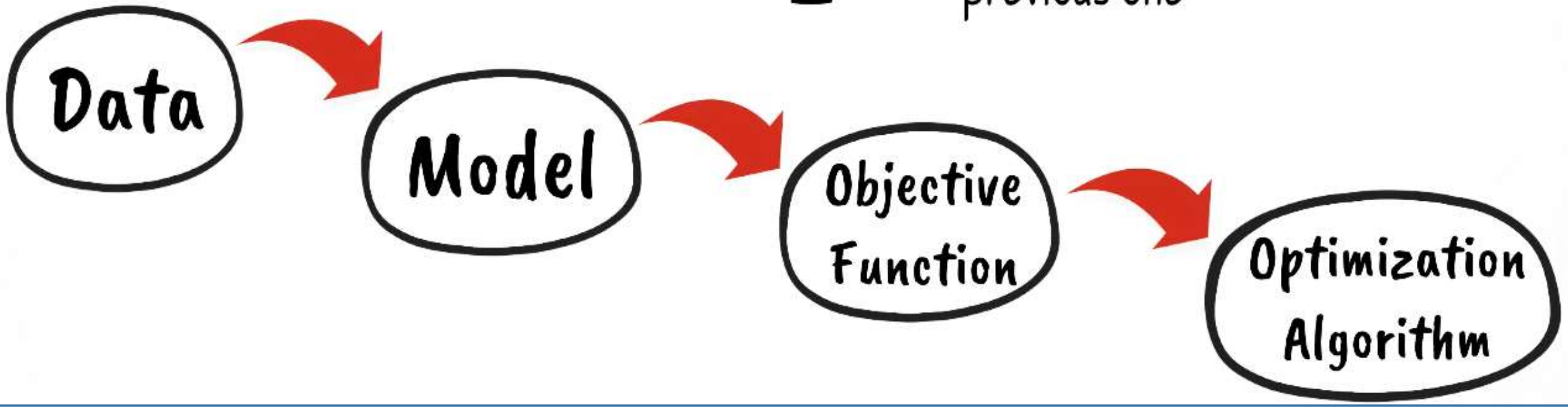
Machine Learning (ML)

Creating an algorithm, which a computer then uses to find a model that fits the data as best as possible. And makes very accurate predictions based on that

ML algorithm

a trial-and-error process

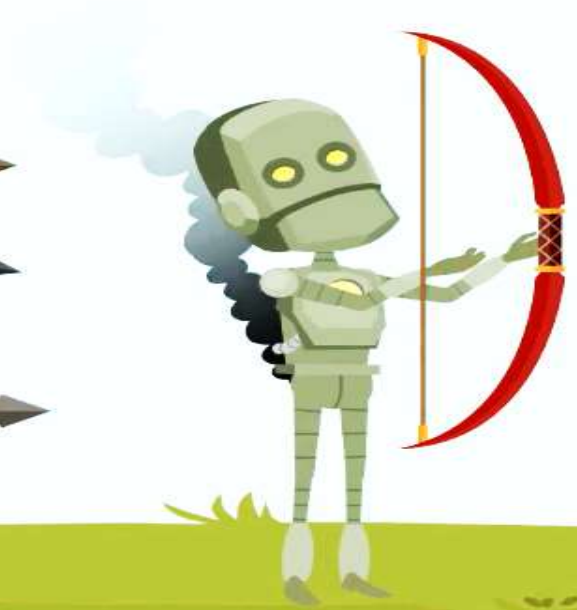
! each consecutive trial is at least as good as the previous one



Data

Model

the usage of the bow

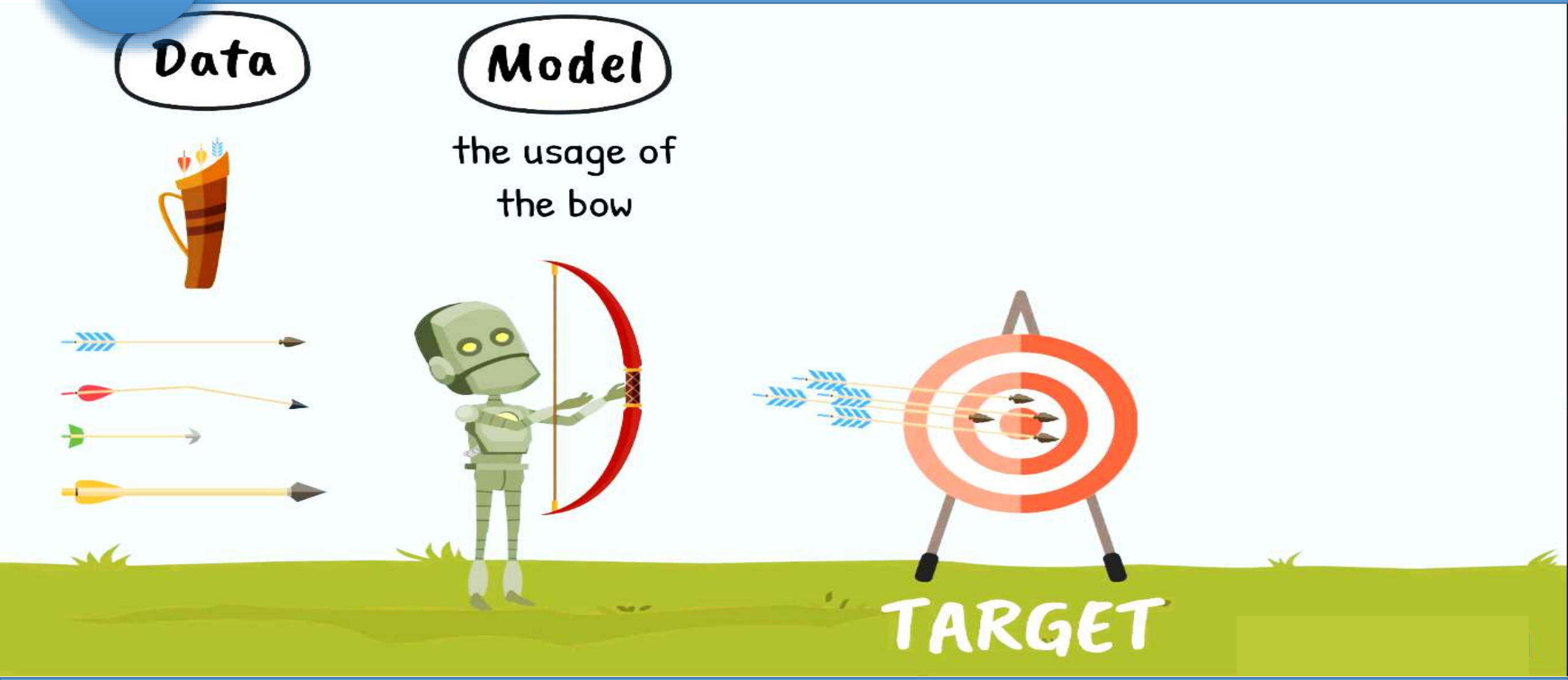


Objective Function

calculate how far from the target

Optimization algorithm

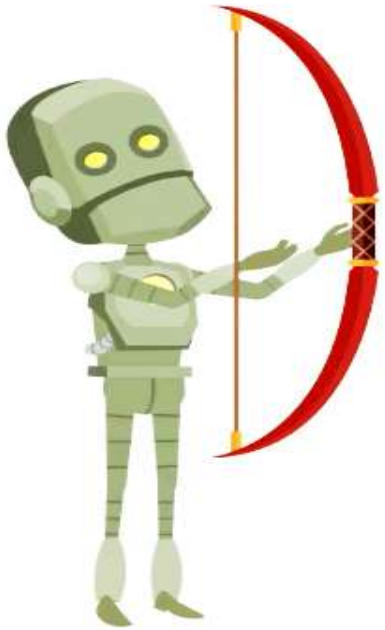
mechanics that will improve the model's performance


Data
Model
the usage of the bow
TARGET

Training your model

100,000 tries → it may have learned how to be the best archer out there

Training your model



- model: trained
- objective function: minimized
- optimization algorithm: has done its job

Machine Learning (ML)

BENEFIT: the robot can learn to fire more effectively than a human!

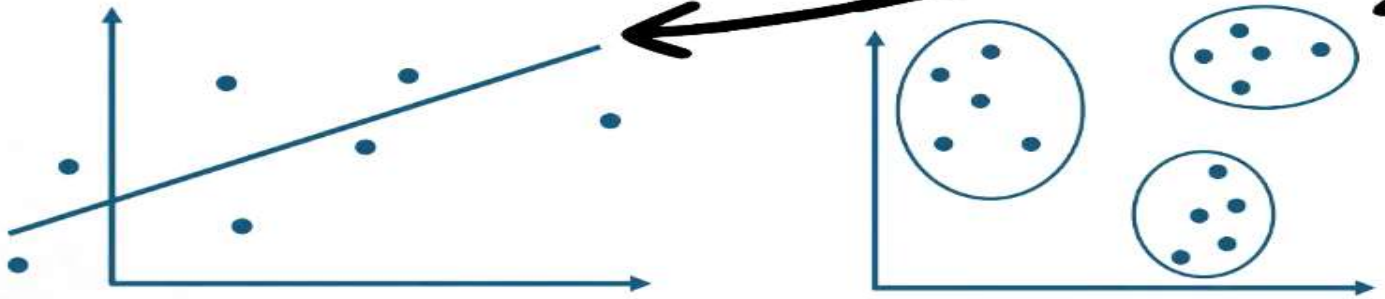
+ discover that we have been holding bows in a wrong way for centuries

Machine Learning (ML)

BENEFIT: the robot can learn to fire more effectively than a human!

USE: improve complex computational models

BUSINESS



Data Science

Fraud Detection

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Observation 1										
Observation 2										
Observation 3										
Observation 4										
Observation 5										
Observation 6										
Observation 7										
Observation 8										
Observation 9										
Observation 10										

ID	NAME	AGE
001	JOHN	35
002	ALAN	24
003	JANE	29

ID	NAME	AGE	...	...	...
001	JANE	19	...	...	...
002	JOHN	35	...	...	...
003	JESS	21	...	...	...
004	TONYA	24	...	...	...
005	IVAN	46	...	...	...
...	...	...	...	...	...

Fraud Detection

Observation 1 = good

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Observation 3										
Observation 4										
Observation 5										
Observation 6										
Observation 7										
Observation 8										
Observation 9										
Observation 10										

Observation 2 = bad

Client Retention

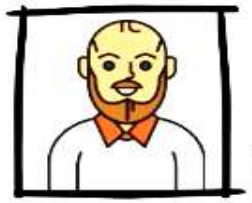
Data Science Positions – Data

DATA ARCHITECT

- designs the way data will be retrieved, processed, and consumed

DATA ENGINEER

- processes the obtained data so that it is ready for analysis



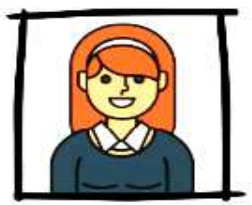
DATABASE ADMINISTRATOR

- handles this control of data
- mainly works with traditional data

Business Intelligence

BI ANALYST

- performs analyses and reporting of past historical data

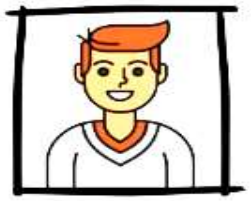


BI CONSULTANT

- 'external BI analyst'

BI DEVELOPER

Data Science & ML



DATA SCIENTIST

- employs traditional statistical methods or unconventional machine learning techniques for making predictions

DATA ANALYST

- prepares more advanced types of analyses

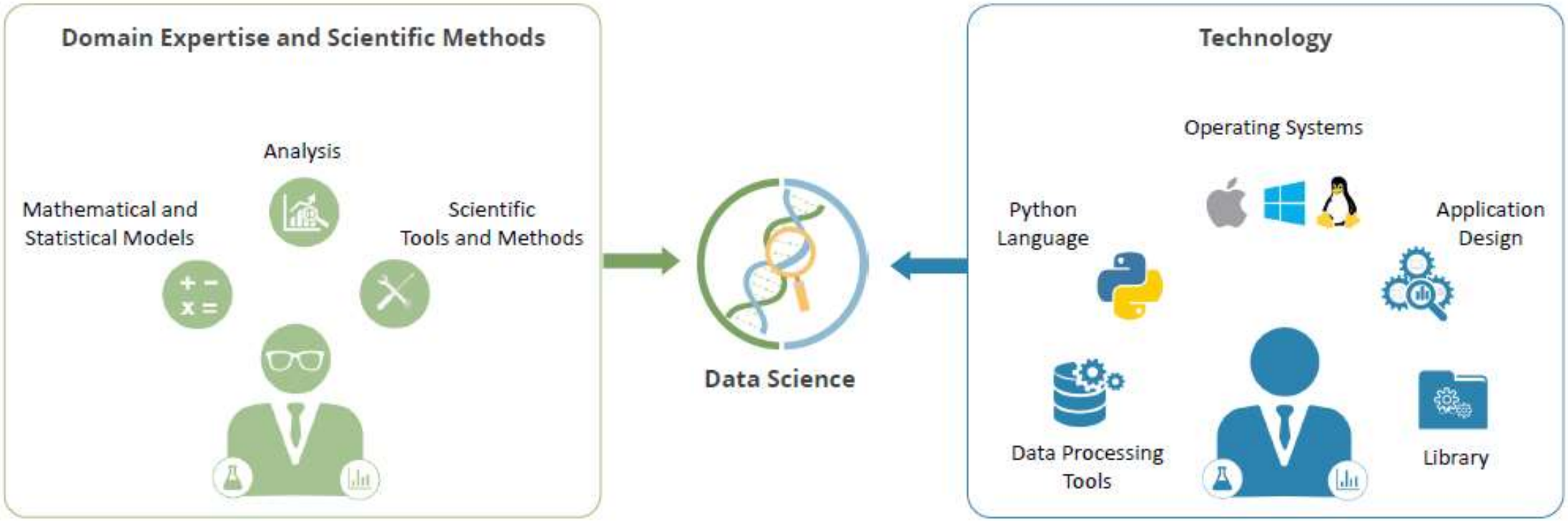
ML ENGINEER

- applies state-of-the-art computational models

Data Science

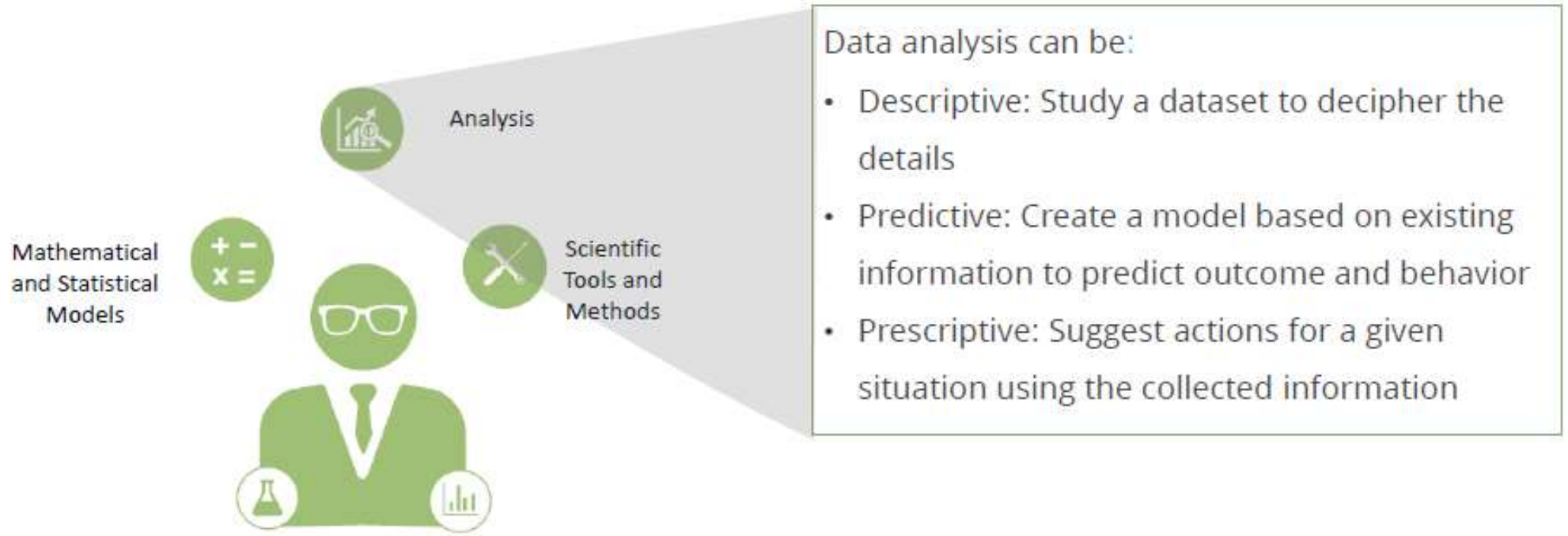
The Components of data science

When we combine domain expertise and scientific methods with technology, we get Data Science.



Domain Expertise and Scientific Methods

Data Scientists collect data and explore, analyze, and visualize it. They apply mathematical and statistical models to find patterns and solutions in the data.

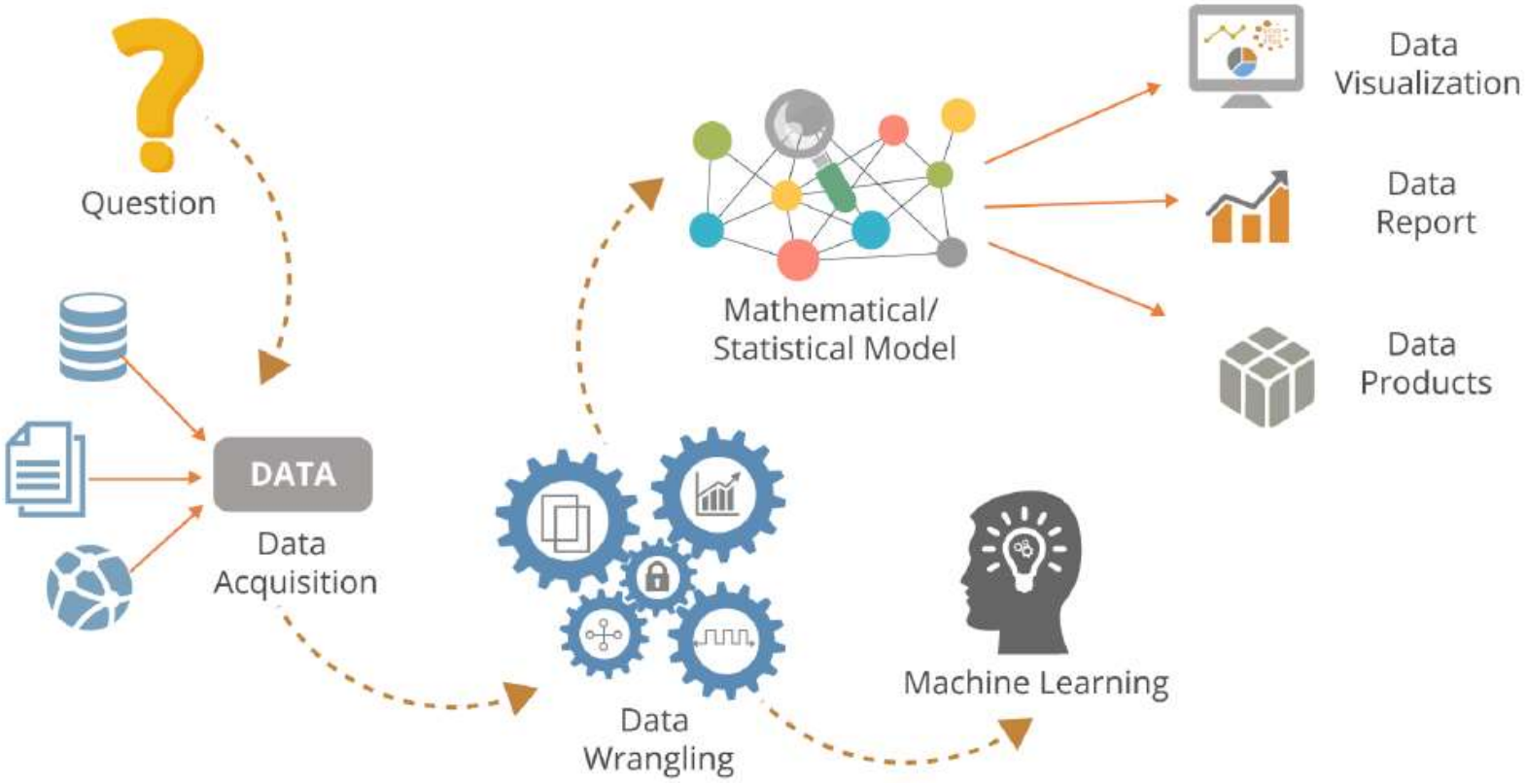


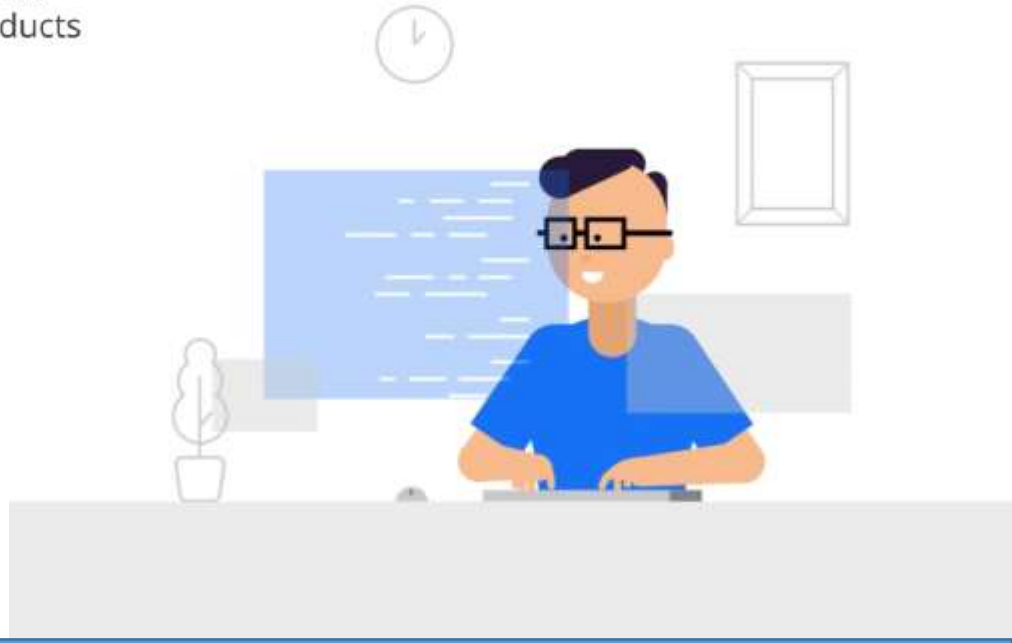
Data analysis can be:

- Descriptive: Study a dataset to decipher the details
- Predictive: Create a model based on existing information to predict outcome and behavior
- Prescriptive: Suggest actions for a given situation using the collected information

Data Science

A Day in a Data Scientist's Life





Technology Academy

Basic Skills of a Data Scientist

A Data Scientist should be able to

- Ask the right questions
- Understand data structure
- Interpret and wrangle data
- Apply statistical and mathematical methods
- Visualize data and communicate with stakeholders
- Work as a team player

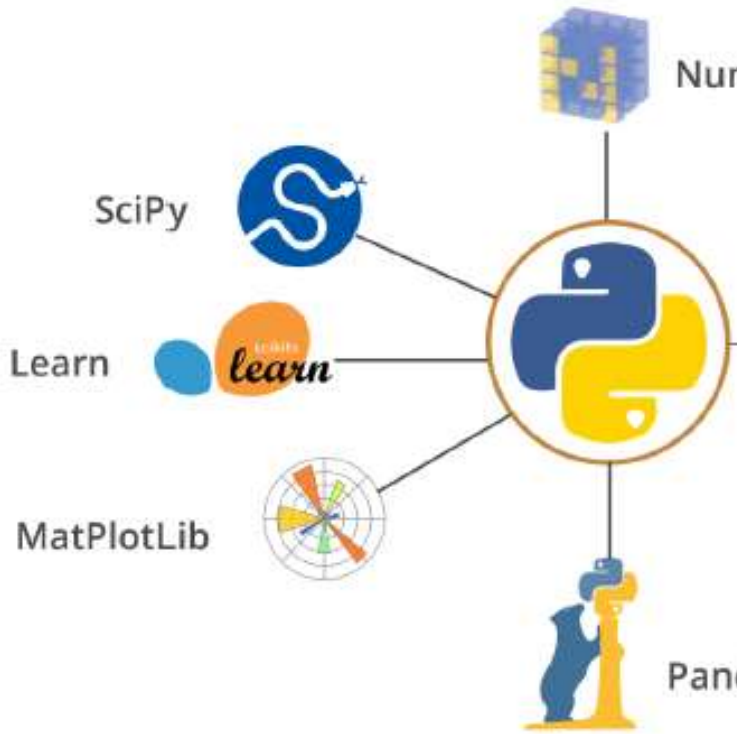
Sources of Big Data

Data Scientists work with different types of datasets for various purposes. Now that Big Data is generated every second through different media, the role of Data Science has become more important.

Python Tool and Technologies

Python is a general purpose, open source, programming language that lets you work quickly and integrate systems more effectively.

Benefits of Python



- Easy to learn
- Open source
- Efficient and multi platform support
- Huge collection of libraries, functions and modules
- Big open source community
- Integrates well with enterprise apps and systems
- Great vendor and product support

? Quiz

Quiz 1

Which of the following is not considered data science?

- **Big Data**
- **Business Intelligence**
- **Traditional Data Science methods**
- **Machine Learning**

Quiz 2

From a data scientist's perspective, the solution of every task begins:

- By suggesting a few hypothetical and theoretical solution to your boss.
- By gathering your team and deciding on which approach to follow to solve the task.
- With a proper dataset
- With a proper analysis of the past data

Quiz 3

Which of the following activities belong to the field of 'predictive analytics' and do not aim at explaining past behavior?

- **Traditional Data.**
- **Big Data**
- **Business Intelligence**
- **Traditional statistical methods**

Quiz 4

Which of the following is related to the pre-processing of a traditional data set?

- Class labeling
- Data cleaning
- Dealing with missing values
- All the above

Quiz 5

Which of the following do you encounter when working with big data?

- Text Data
- Integer
- Digital Image data
- All the above

Quiz
6

The process of representing observations as numbers is called:

- **Collection observation**
- **Quantification**
- **Measuring the accumulation of a presentation**
- **Reporting**

Quiz 7

A measure that has a business meaning attached is called:

- An observation
- A Quantification
- A Metric
- A KPI

Quiz
8

A KPI (Key Performance Indicator) can be best defined as:

- Accumulation of a observation to show some information
- A metric that is aligned to your business objectives
- A Quantification that has a business meaning attached
- An observation that can potentially be related to the business goals of a company.

Quiz 9

The job of a business intelligence analyst always involves the creation of:

- **Reports**
- **Dashboards**
- **KPI**
- **All the above**

Quiz
10

Which technique can be implemented if you want to reduce the dimensionality of a certain statistical problem?

- **Factor Analysis**
- **Cluster Analysis**
- **Time series Analysis**
- **All the above**

Quiz 11

Which technique is associated with plotting values against time, shown always on the horizontal line?

- **Factor Analysis**
- **Cluster Analysis**
- **Time series Analysis**
- **Regression Analysis**

Quiz 12

Choose the best answer. When the data is divided into a few groups, you should apply:

- Factor Analysis
- Cluster Analysis
- Time series Analysis
- None of the above

Quiz
13

Which line represents the four ingredients of any machine learning algorithm?

- Model, data , reward system , objective function
- Data , model ,objective function , optimization algorithm
- Model , labelled data, unlabeled data , optimization algorithm
- None of the above

Quiz
14

Which of the following is a typical real-life example where big data techniques are being applied?

- Basic customer data
- Social media
- Price optimization
- Inventory management

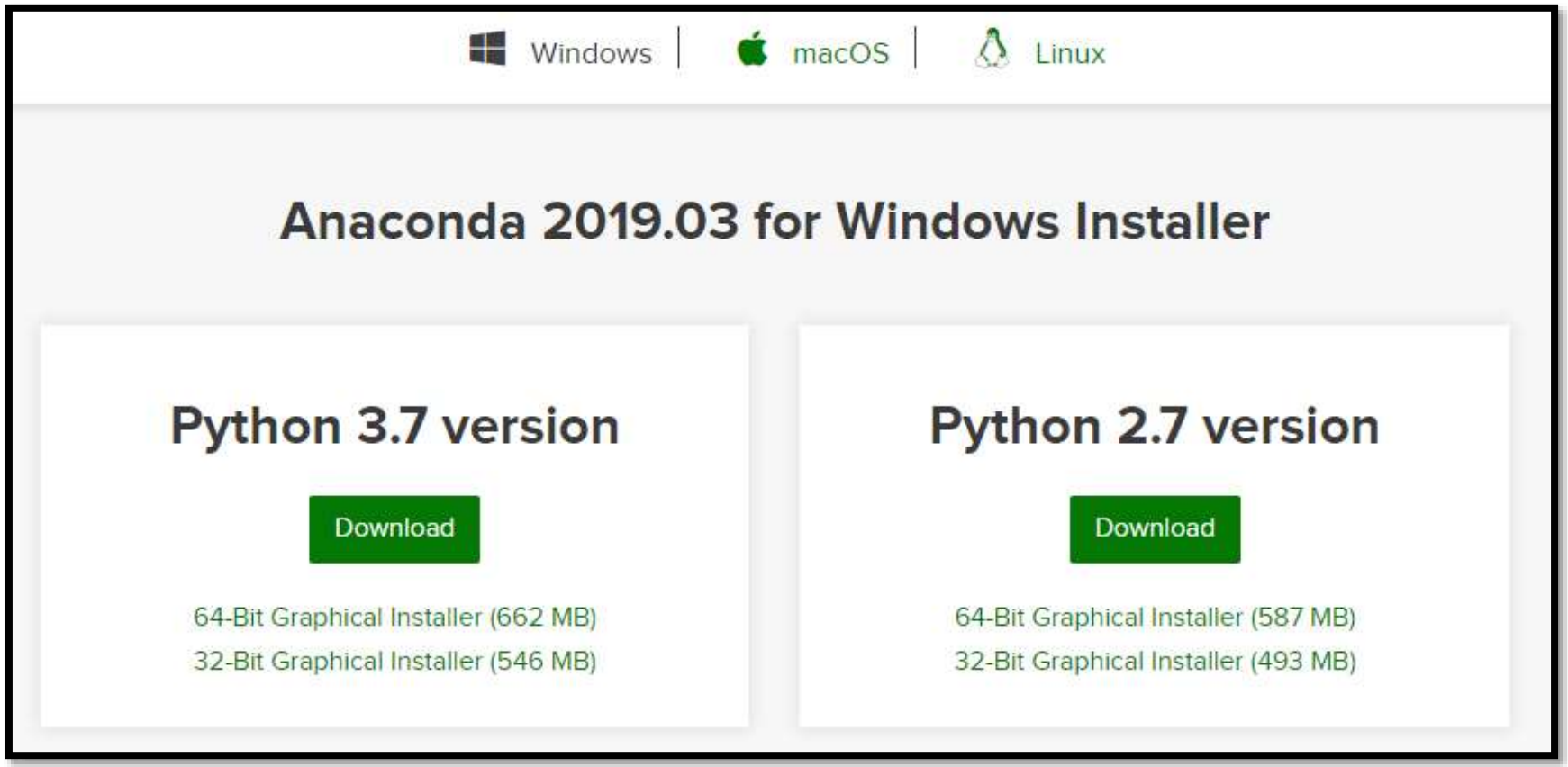
Python Environment Setup and Essentials

Why Anaconda

The open-source Anaconda Distribution is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. With over 11 million users worldwide, it is the industry standard for developing, testing, and training on a single machine, enabling *individual data scientists* to:

1. Quickly download **1,500+ Python/R data science packages**
2. Manage libraries, dependencies, and environments with **Conda**
3. Develop and train machine learning and deep learning models with **scikit-learn, TensorFlow, and Theano**
4. Analyze data with scalability and performance with **NumPy& pandas.**
5. Visualize results with **Matplotlib, Bokeh, Datashader, and Holoviews**

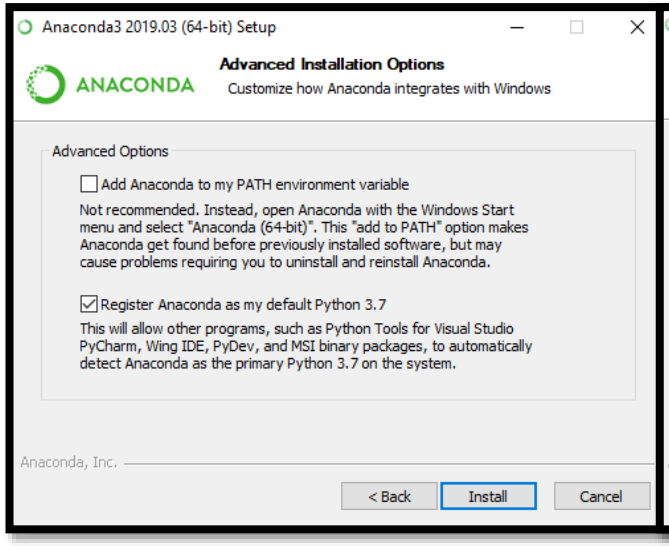
Installation of Anaconda Python Distribution

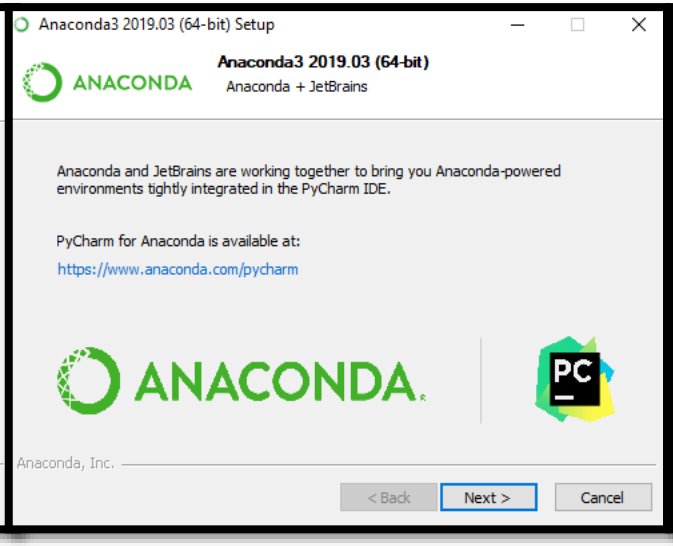


Or (https://docs.anaconda.com/anaconda/install/windows/)

Data Science

Installation of Anaconda Python Distribution

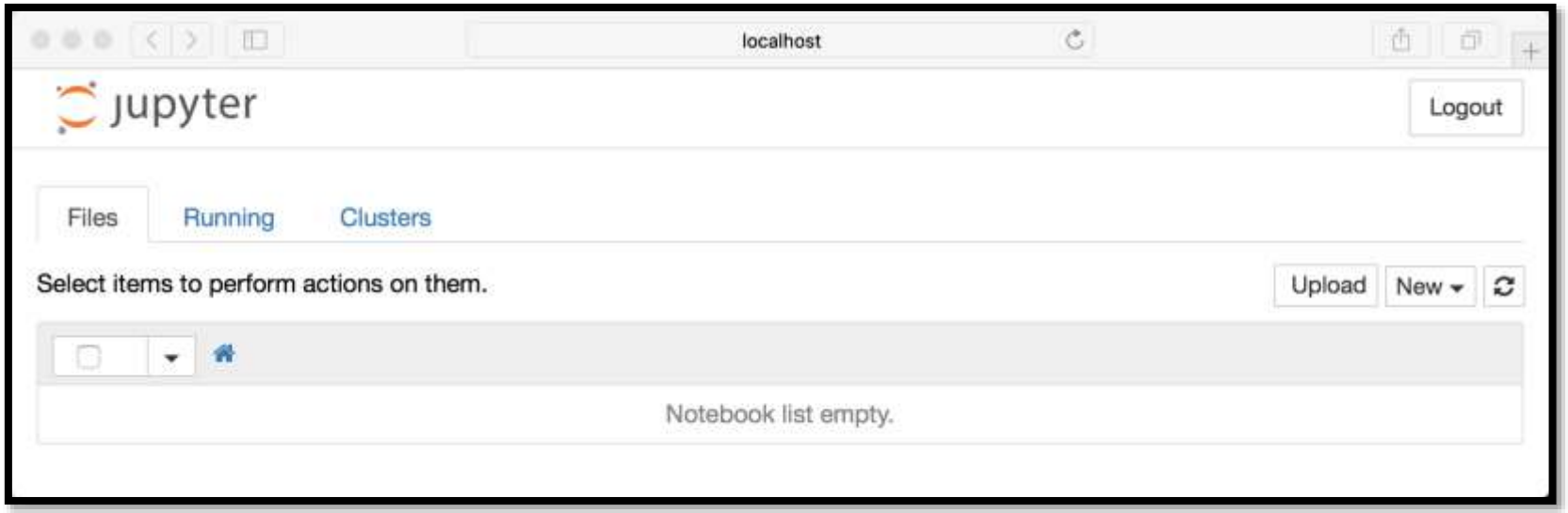




Jupyter Notebook

The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text.

The name, Jupyter, comes from the core supported programming languages that it supports: **Julia, Python, and R.**



Jupyter Notebook

To install Jupyter notebook on your system, type the command shown here on Anaconda prompt and press Enter to execute it.