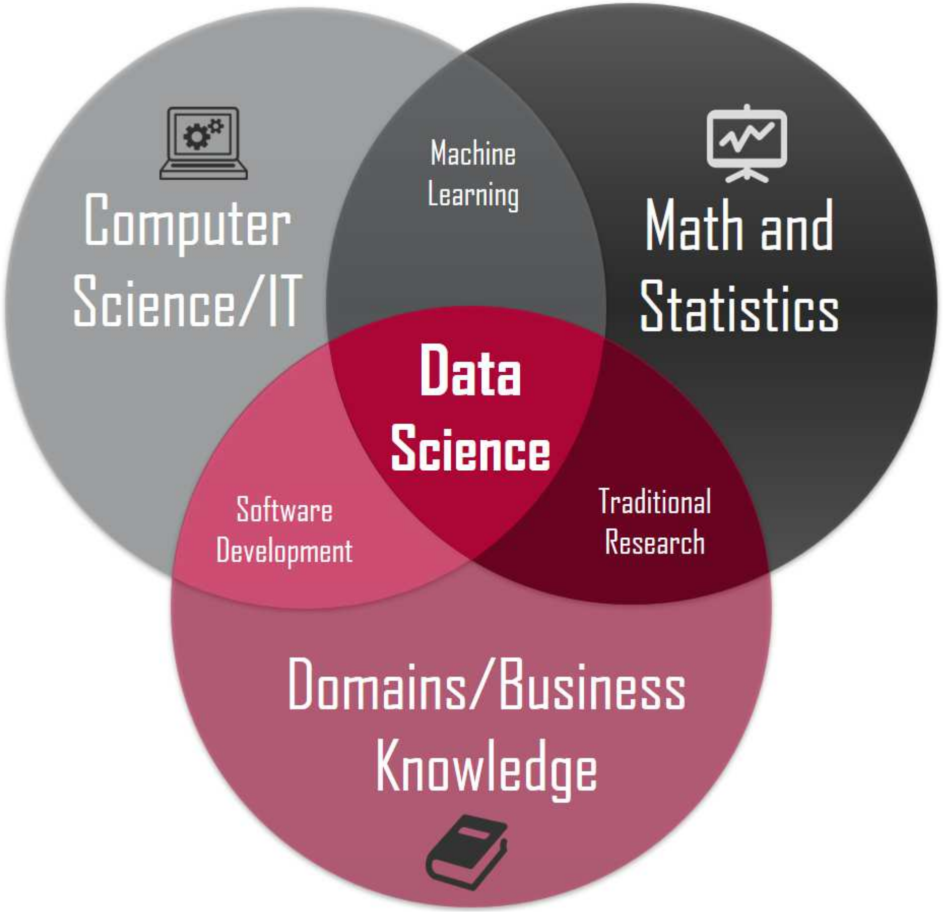
Data Science

What is it?

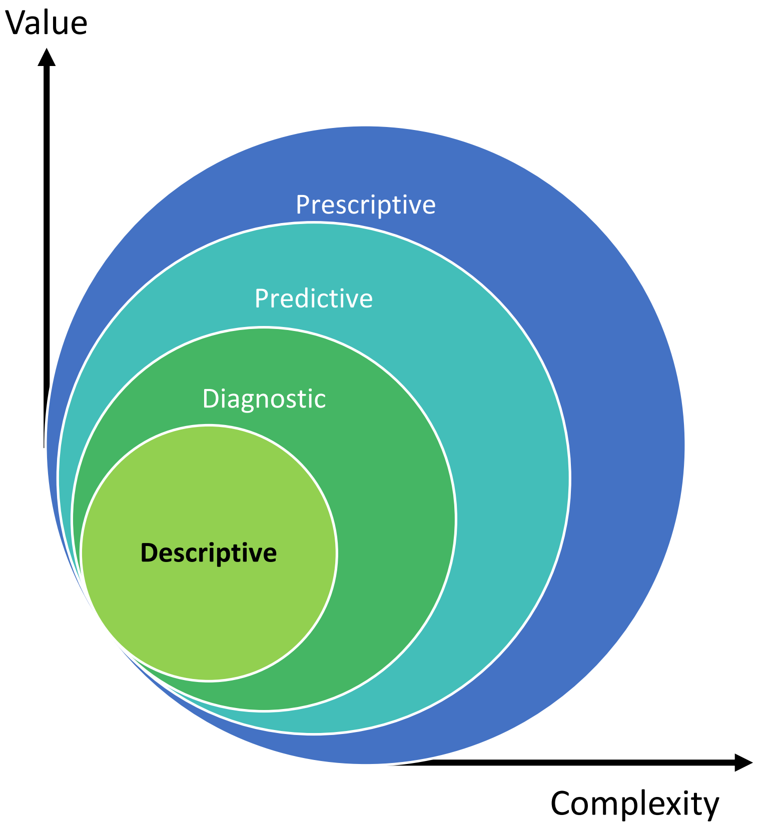
Defined by fields of study



Source: Fox school of business, Temple university

Data Analytics

'Descriptive'



- **Ask:**

  What is happening?

- **Goal:**

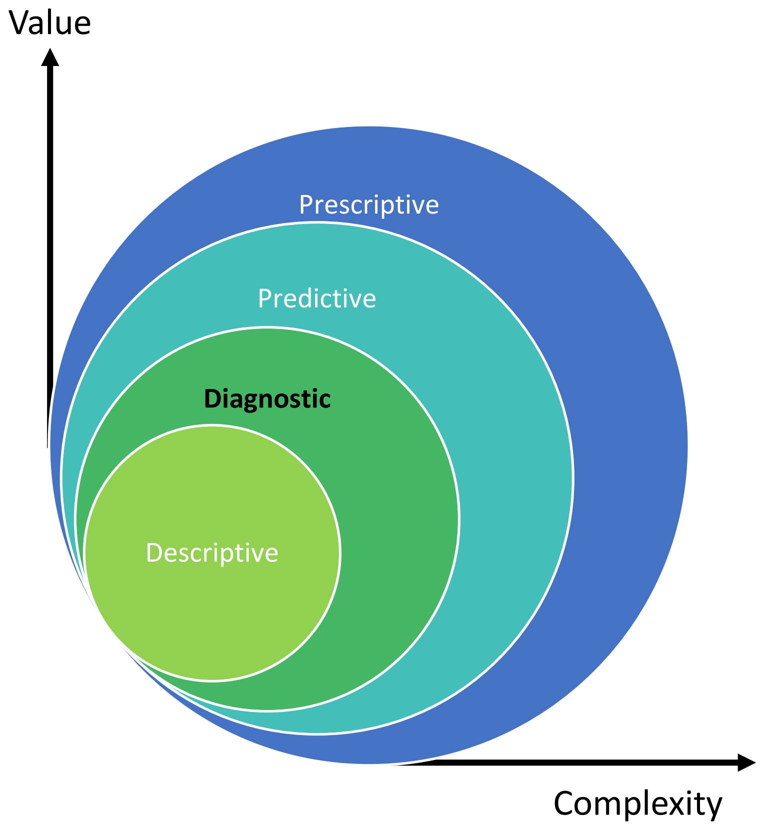
  To understand the situation

- **Example:**

  Trend of heart failure admission to a hospital is increasing during winter

Data Analytics

'Diagnostic'



- **Ask:**

  Why is it happening?

- **Goal:**

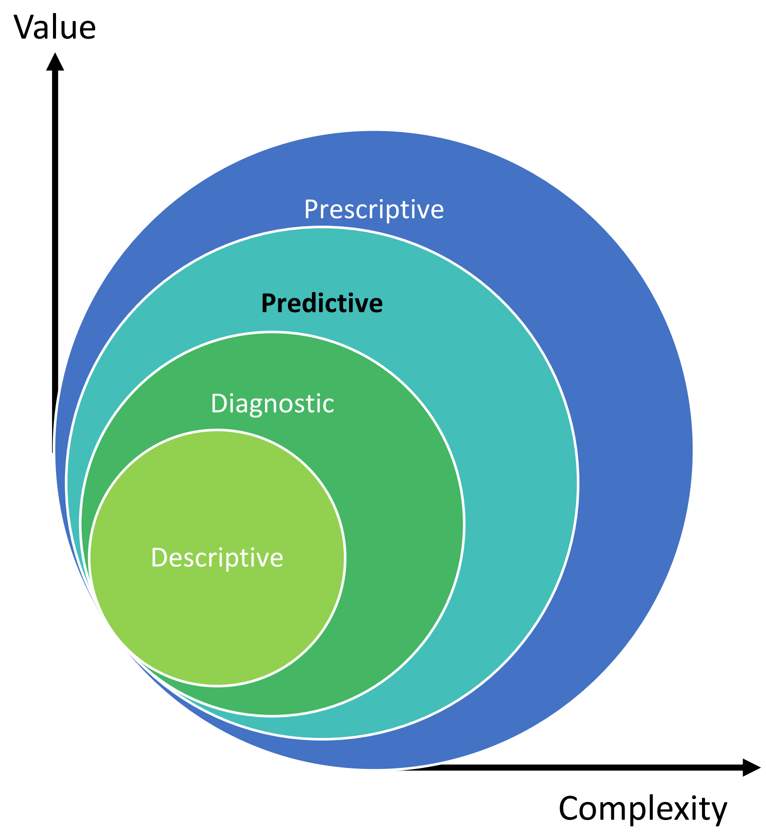
  Root cause identification

- **Example:**

  Asthma and low temperature are root cause because they stimulate airway hyper-responsiveness that makes patients cannot breathe properly resulting heart failure

Data Analytics

‘Predictive’



- **Ask:**

  What is likely to happen?

- **Goal:**

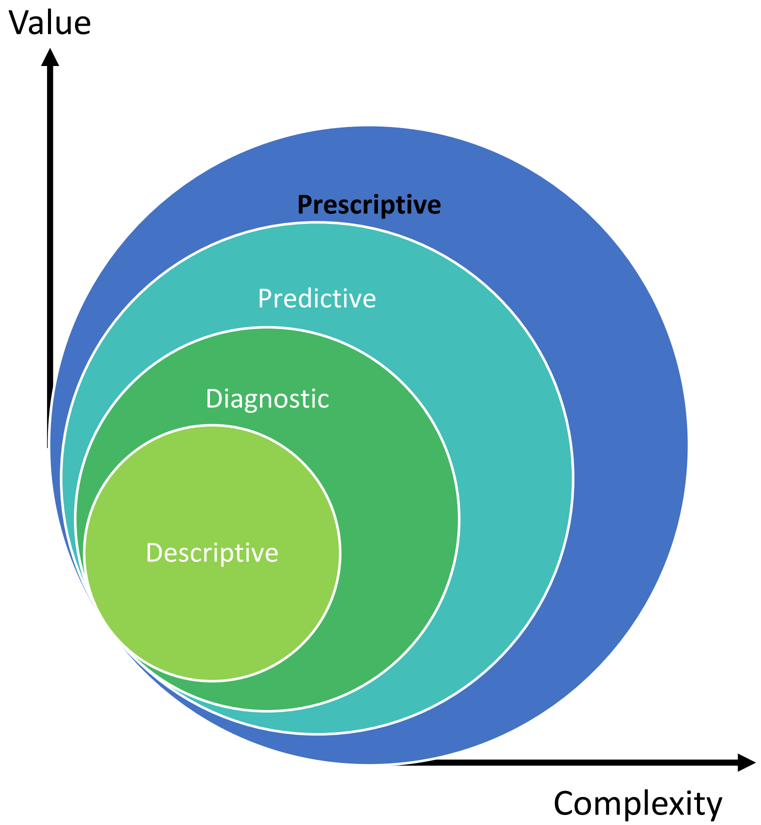
  Foreseeing the future

- **Example:**

  Heart failure predictive modeling from patient records and external data, e.g. weather.

Data Analytics

'Prescriptive'



- **Ask:**

  What are suggestions to organization?

- **Goal:**

  Actionable plan

- **Example:**

  Patient monitoring system using IoT sensor

Data Science

Why do organizations need data science?

Growth of Data

- How much data were generated by human?

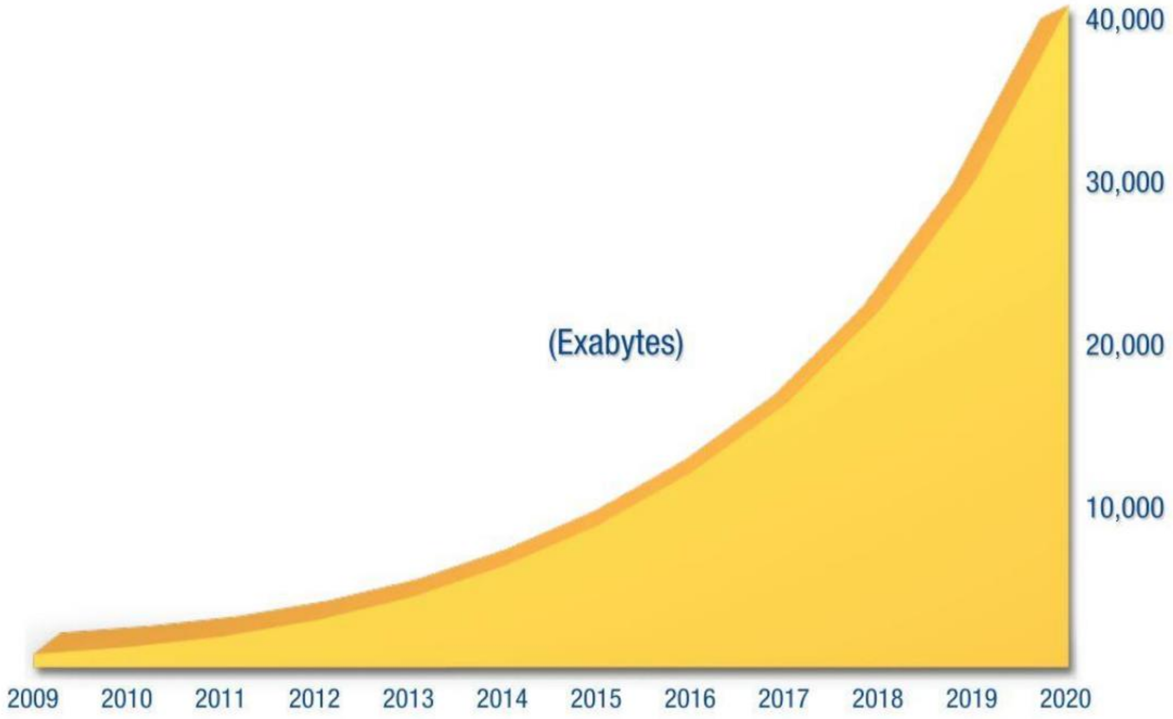
2005: 130 EB

2010: 1,200 EB

2015: 7,900 EB

2020: 40,900 EB

Growth of Data



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Types of Question

What types of question can data science answer?

Problems in Data Science

Comparing 1 sample

Is the population mean different from a given value?

Comparing 2 samples

Are two population means different?

Correlation

Direction and strength of relationship between variables

Causation

Relationship between cause and effect

Regression

Estimating outcomes given input variables

Classification

Identifying the target class of given observations

Clustering

Reveal hidden structure in data

More …

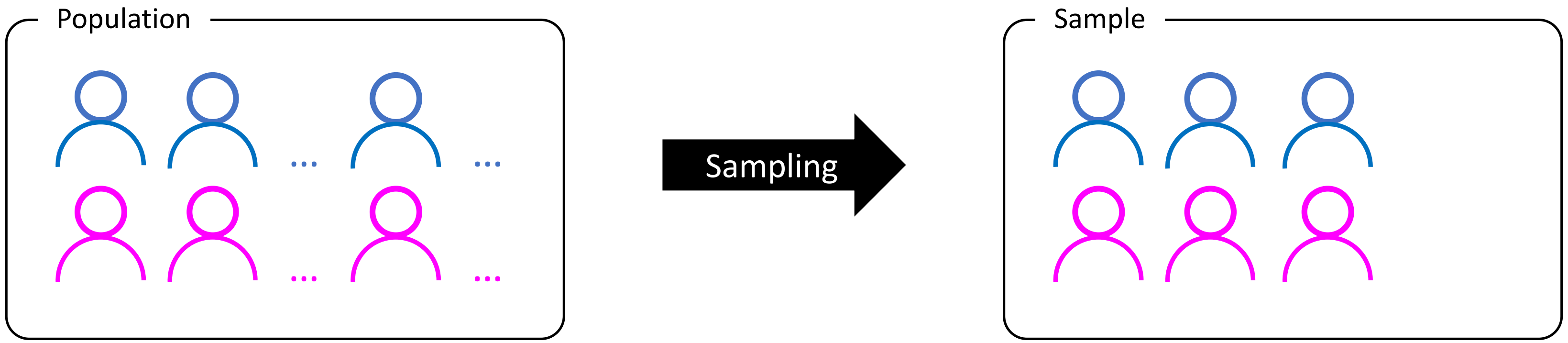
Multiple Sample Comparison, Raking, Optimization, etc.

Comparing 1 Sample

Is population mean different from a given value x?

Some Concepts

- Population and Samples



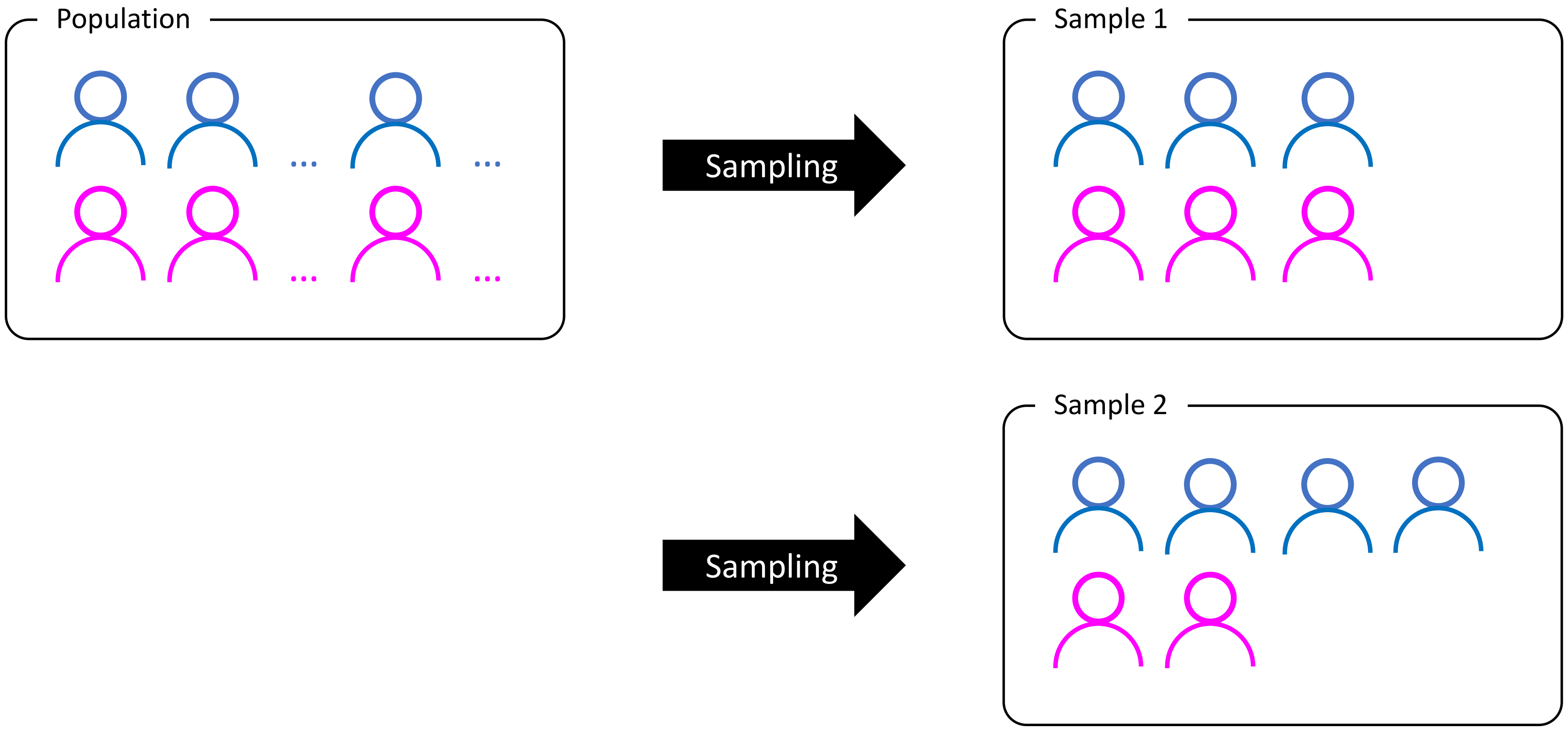
Impractical to observe due to

- Costly
- Time consuming

Practical to observe

Concept

- Variation in samples



Examples in various domains

Business

- Does our customer satisfaction achieve the target satisfaction level (4 out of 5)?

Healthcare

- Are Thai too fat (BMI > 27.5)?

Manufacturing

- Does this production line achieve the target yield (99% yield)?

Agriculture

- Do Thai rice farms have too low productivity (3.1 tons per hectare)?

Ref: https://data.oecd.org/agroutput/crop-production.htm

Examples in various domains

Education

- Do our students spend time on self-study hard enough (more than 8 hours per week)?

$

Financial

- Does our house mortgage service process fast enough (target within 3 working days)?

Tourism

- Does Europe trip expensive (ratio between spending amount to income)?

More …

- Biology
- Economy
- Military
- Medicine
- Etc.

Discussions

- From the following domains or your working area
  - Business
  - Healthcare
  - Manufacturing
  - Agriculture
  - Education
  - Financial
  - Tourism
  - More …

Give an example related to 1 sample comparison

Discussions

- Given BMI of Thai participants as follows

Samples	BMI
1	25
2	30
3	28
4	33
5	35

Average BMI is

$$\frac{25 + 30 + 28 + 33 + 35}{5} = 30.2$$

If BMI > 27.5 is consider as too fat, Can we conclude that Thai is too fat? Why?

Answer
No. Thai are population not samples

Later, we will learn how to infer from samples to population using t-statistics

2 Samples Comparison

Are two population means different?

Examples in various domains

Business

- Does our customer satisfy our product compared to our top competitor?

Healthcare

- Are Thai fatter than Vietnamese?

Manufacturing

- Do the automobile manufacturer in Thai have better quality than automobile manufacturer in Malaysia?

Agriculture

- Do Thai rice farms have better productivity compared to India?

Examples in various domains

Education

- Do our students in Bangkok spend time on self-study harder than students in Chiang Mai?

Financial

- Does our house mortgage service process faster than the top competitor?

Tourism

- Does Europe trip expensive compared to America trip?

More …

- Biology
- Economy
- Military
- Medicine
- Etc.

Discussions

- From the following domains or your working area
  - Business
  - Healthcare
  - Manufacturing
  - Agriculture
  - Education
  - Financial
  - Tourism
  - More …

Give an example related to 2 samples comparison

Discussions

- Given BMI of Thai and Vietnamese participants as follows

Samples	BMI (THA)	BMI (VNM)
1	25	23
2	30	25
3	28	30
4	33	28
5	35	32

Average BMI (THA) is 30.2
Average BMI (VNM) is 27.6

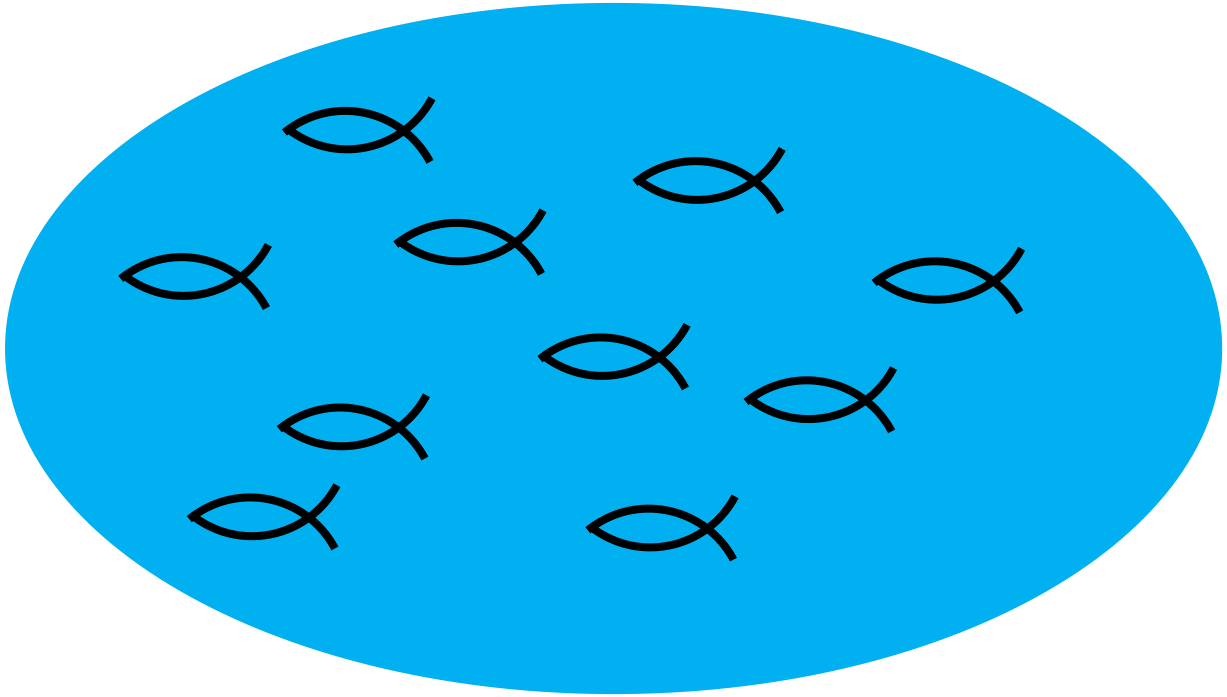
Can we conclude that Thai is fatter than Vietnamese? Why?

Answer
No. Thai and Vietnamese are population not samples

Later, we will learn how to infer from samples to population using t-statistics

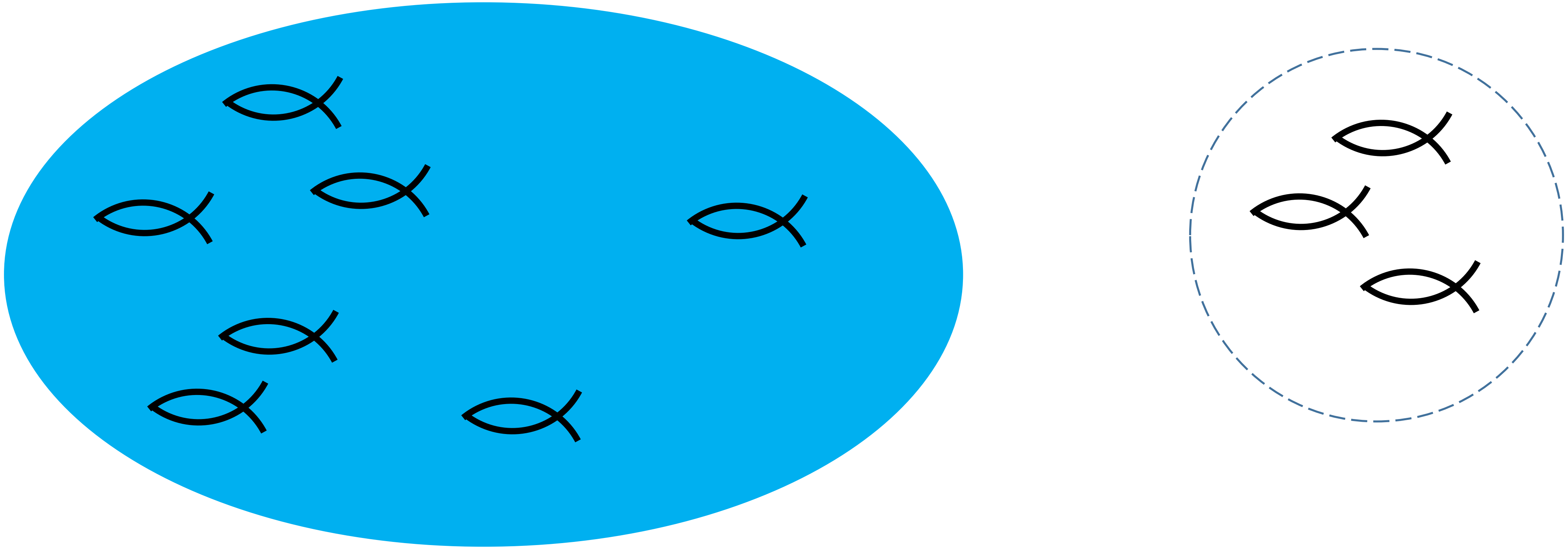
Discussions

- How to **estimate** the number of fish in this pond?



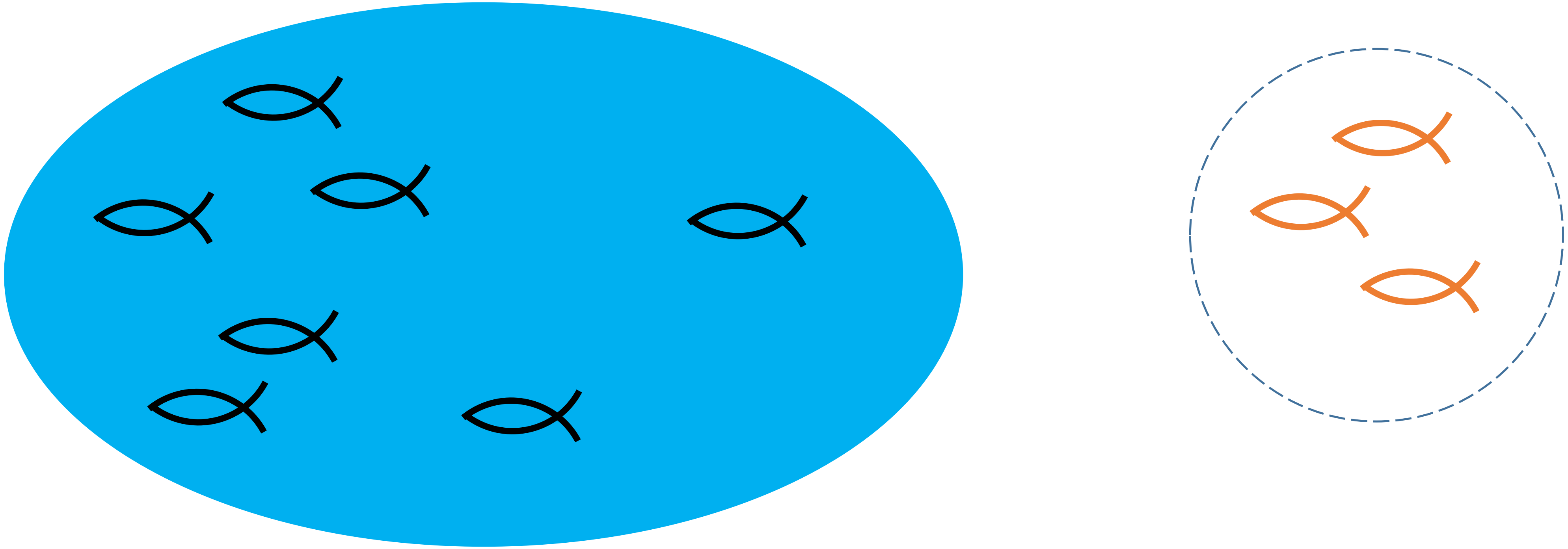
Discussions

- Let the number of fish in the pond equal to X
- Use a container to take some fish out



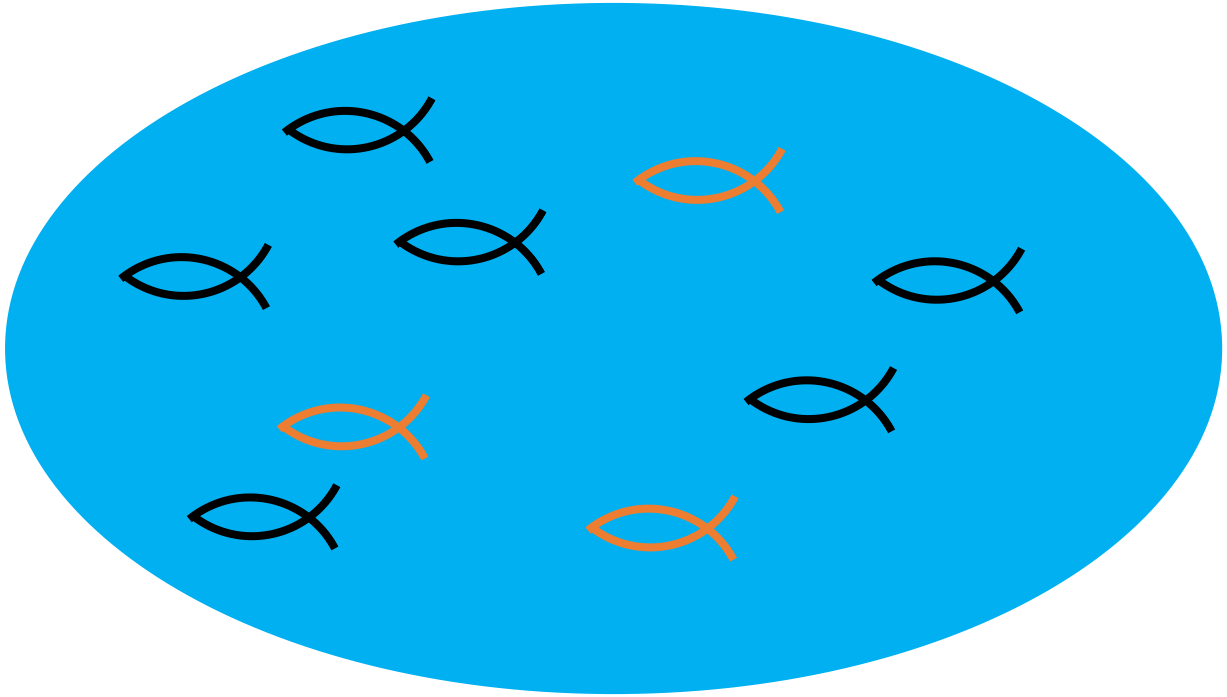
Discussions

- Suppose the number of fish in container is equal to 20
- Mark on them



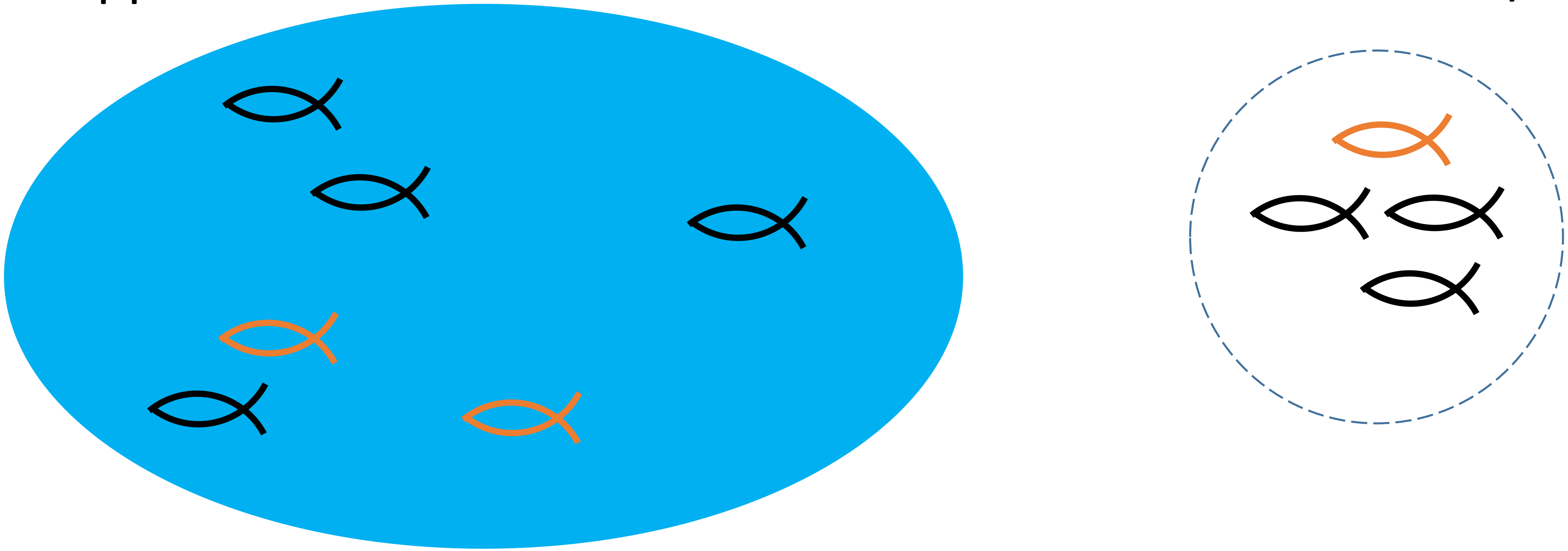
Discussions

- Put the marked fish back to the pond
- We shall assume randomization



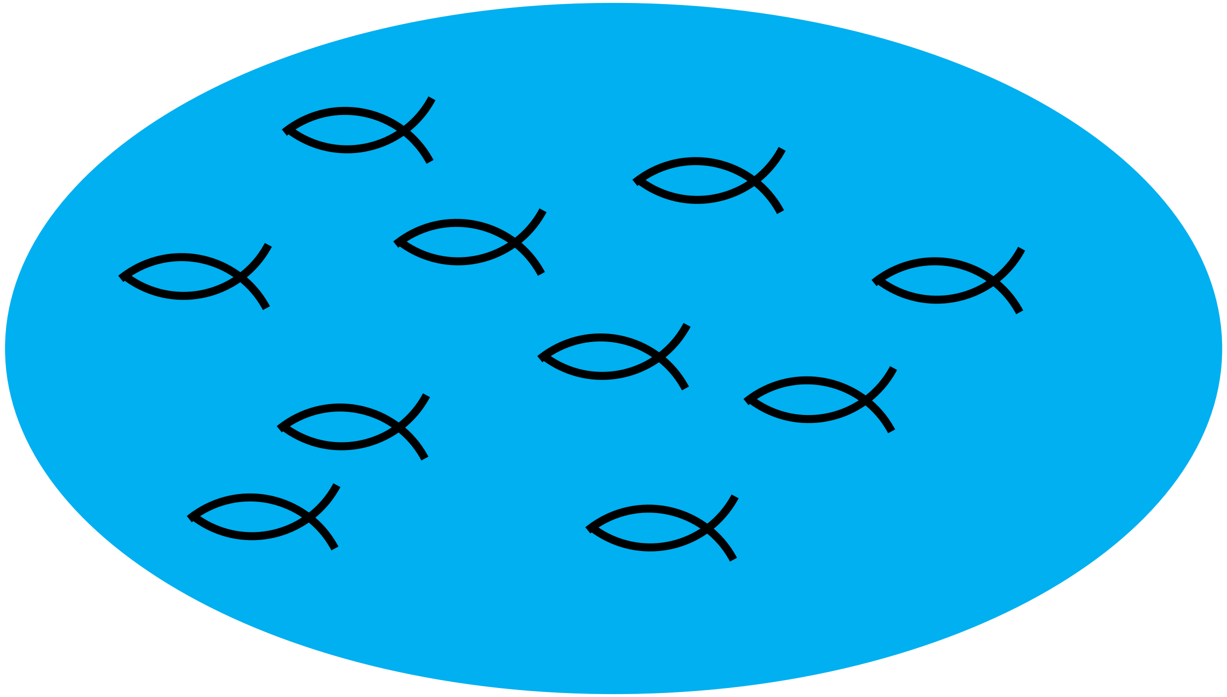
Discussions

- Use a container to take some fish out again
- Suppose we observe that the ratio of marked and unmark is 1/3



Discussions

- Then we can infer to the number of fish in the pond as follows



$$\frac{marked\ fish}{unmarked\ fish} = \frac{1}{3}$$

If marked fish = 20 then unmarked fish = 60

Hence, total fish = 80

Correlation

What is the direction of relationship between two variables?

How much strength of such relationship?

Examples in various domains

Business

- Does customer spending correlate to the number of visits to our shop?

Healthcare

- Does heart failure correlate to asthma in winter?

Manufacturing

- Does visual inspection outcomes of trainee operator correlated to trainer?

Agriculture

- Does plant growth rate correlate to the ratio of Nitrogen in soil?

Examples in various domains

Education

- Do student scores in Mathematics correlate to scores in Science?

Financial

- Does SET50 correlate to S&P500?

Tourism

- Does the number of tourist correlate to the number of nights in hotel?

More …

- Biology
- Economy
- Military
- Medicine
- Etc.

Discussions

- From the following domains or your working area
  - Business
  - Healthcare
  - Manufacturing
  - Agriculture
  - Education
  - Financial
  - Tourism
  - More …

Give an example related to correlation

Causation

Is A depend on B?

Examples in various domains

Business

- Does responseiveness cause customer churn?

Healthcare

- Does stress cause heart failure

Manufacturing

- Does humidity in air cause defect in production line?

Agriculture

- Does sunlight intensity cause seed cultivation?

Examples in various domains

Education

- Does school reputation cause student admission?

Financial

- Does customer income cause house location?

Tourism

- Does cost of living cause the yearly average number of tourists?

More …

- Biology
- Economy
- Military
- Medicine
- Etc.

Discussions

- From the following domains or your working area
    - Business
    - Healthcare
    - Manufacturing
    - Agriculture
    - Education
    - Financial
    - Tourism
    - More …

Give an example related to causation

Regression

How much or how many?

Examples in various domains

Business

- How much customer demand in the next month?

Healthcare

- How many patients to be admitted in the next month?

Manufacturing

- How many raw materials to be ordered?

Agriculture

- How much rice to be produced for the next shipment?

Examples in various domains

Education

- How much student GPA given IQ?

Financial

- How much stock price on tomorrow?

Tourism

- How many tourist coming to Thailand in this summer?

More …

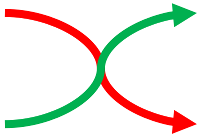
- Biology
- Economy
- Military
- Medicine
- Etc.

Discussions

- From the following domains or your working area
  - Business
  - Healthcare
  - Manufacturing
  - Agriculture
  - Education
  - Financial
  - Tourism
  - More …

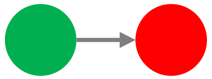
Give an example related to regression

FAQ



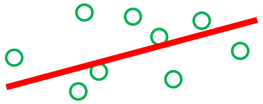
Correlation

Direction and strength of relationship between variables



Causation

Relationship between cause and effect



Simple Linear Regression

Estimating linear relationship between two variables

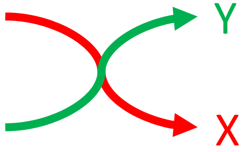
What are the difference between them?

FAQ

Correlation VS Dependency

- What are the differences between correlation and dependency?

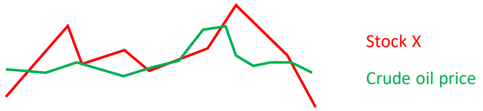
Correlation



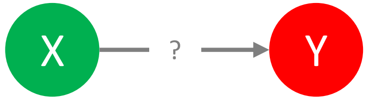
Correlation quantifies **linear** dependence between two variables

Correlation does not capture **non-linear** relationship between two variables

Ex. Does the stock price X correlate to crude oil price?



Dependency



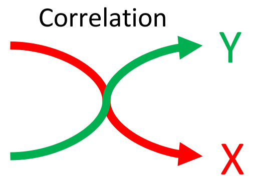
Dependency is the association whether causal or not between two variables

Ex. Is ischemic heart disease the cause of death?

FAQ

Correlation VS Causation

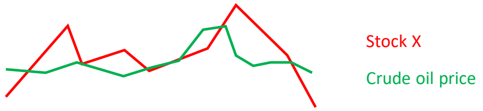
- What are differences between correlation and causation

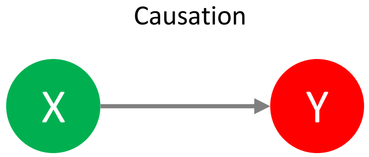


Correlation quantifies **linear** dependence between two variables

Correlation does not capture **non-linear** relationship between two variables

Ex. Does the stock price X correlate to crude oil price?





Causation indicates that one event is the result of the occurrence of the other event

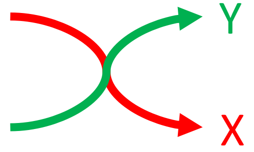
Ex. Ischemic heart disease is the top cause of death globally in 2015 reported by WHO

FAQ

Correlation VS Simple Linear Regression

- What are the differences between correlation and simple linear regression?

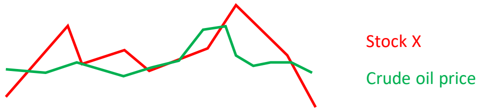
Correlation



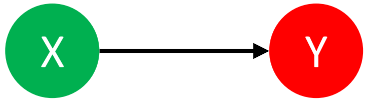
Correlation quantifies **linear** dependence between two variables

Correlation does not capture **non-linear** relationship between two variables

Ex. Does the stock price X correlate to crude oil price?

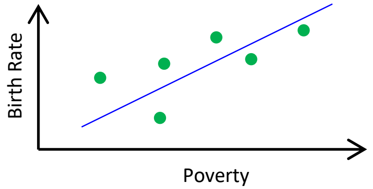


Simple Linear Regression $y = \beta_0 + \beta_1 x$



Simple linear regression estimates value of Y given value of X by assuming linear relationship

Ex. Predict birthrate from poverty rate



FAQ

Can we perform causation test?

- None of experimental design proves causal relationship
- Observation study to demonstrate correlation can only
  - Show strong or weak **evidence** of causality
  - Cannot infer causation
- No causation means zero correlation but not vise versa

Classification

What is its class?

Examples in various domains

Business

- Fraud detection

Healthcare

- Detecting diabetes from image of patient eye

Manufacturing

- Given a conveyer carrying finish goods, detecting defects

Agriculture

- Which fruit are ready to be havested?

Examples in various domains

Education

- Emotional classification in the classroom

Financial

- Loan application

Tourism

- Classify tourists from profile whether he/she is likely to have over-stay or not?

More …

- Biology
- Economy
- Military
- Medicine
- Etc.

Discussions

- From the following domains or your working area
  - Business
  - Healthcare
  - Manufacturing
  - Agriculture
  - Education
  - Financial
  - Tourism
  - More …

Give an example related to classification

Clustering

How data is organized?

Examples in various domains

Business

- Customer segmentation

Healthcare

- Patient obese from several factors, e.g. demographic factor (age, gender, deprivation), risk factors (diabetes, cardiovascular, stroke, osteoarthritis) and individual factor (metabolic)

Manufacturing

- Supplier Quality Evaluation

Agriculture

- Precision farm

Examples in various domains

Education

- Student categorization by academic performance

Financial

- Stock segmentation

Tourism

- Hotel segmentation

More …

- Biology
- Economy
- Military
- Medicine
- Etc.

Discussions

- From the following domains or your working area
    - Business
    - Healthcare
    - Manufacturing
    - Agriculture
    - Education
    - Financial
    - Tourism
    - More …

Give an example related to clustering

Summary

What we have learned

Summary

- What we have learned
  - Opportunity of Data Science
  - Data Science Problems

- What Next
  - Data Science Project Life Cycle
  - Data interactive visualization tool