Everything Starts with Data
Week 2 64/1

Pree Thiengburanathum, PhD.

Announcement

- PreTest and workshop1 have been graded?
- Hybrid classroom? Vs Covid 19

Review last week

- History Industrials to data-driven
  - The rise of powerful algorithms and large-volume of data
- Roles in Data Science world
  - Contemporize data scientist come from different background such as engineering, statistics, and even psychology.
  - Average 114345$/ year (Sexiest job in 21st century)

- Workshop 2 deadline today 11:59 pm

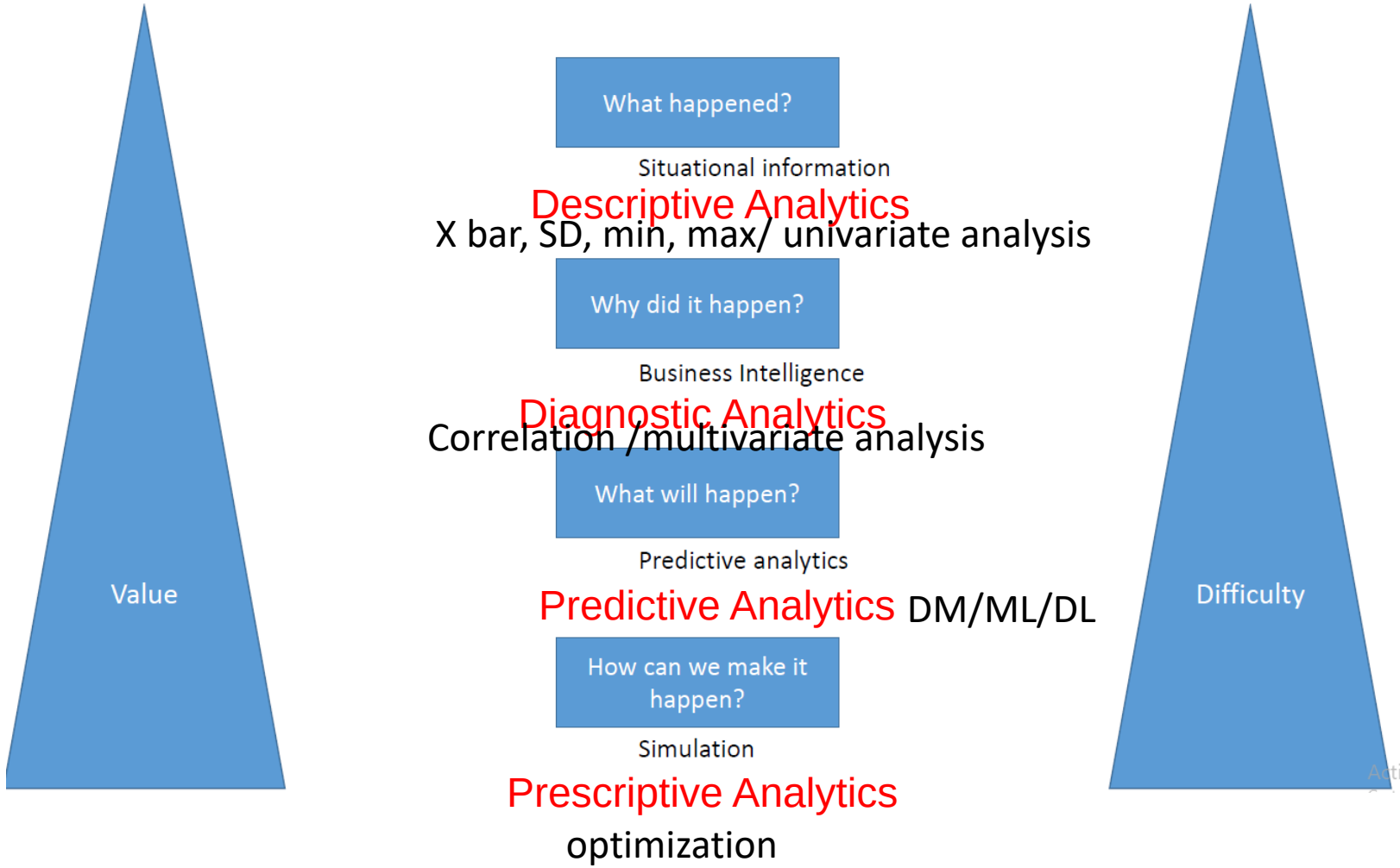
Where are we now?

Week	Topics
1	Data Scientist Foundation
2	Basic data analytics: KDD
3	Basic data analytics: Data to Data Product
4	What is Data (Str - eg. nomi, unstr - img, text)
5	Dataset (Basic manipulation)
6	Data quality (e.g., outlier, inconsistency, duplication, etc.)
7	Processes - history, e.g., turn kdd to crisp-dm -> modern
8	Processes - in action
Midterm	

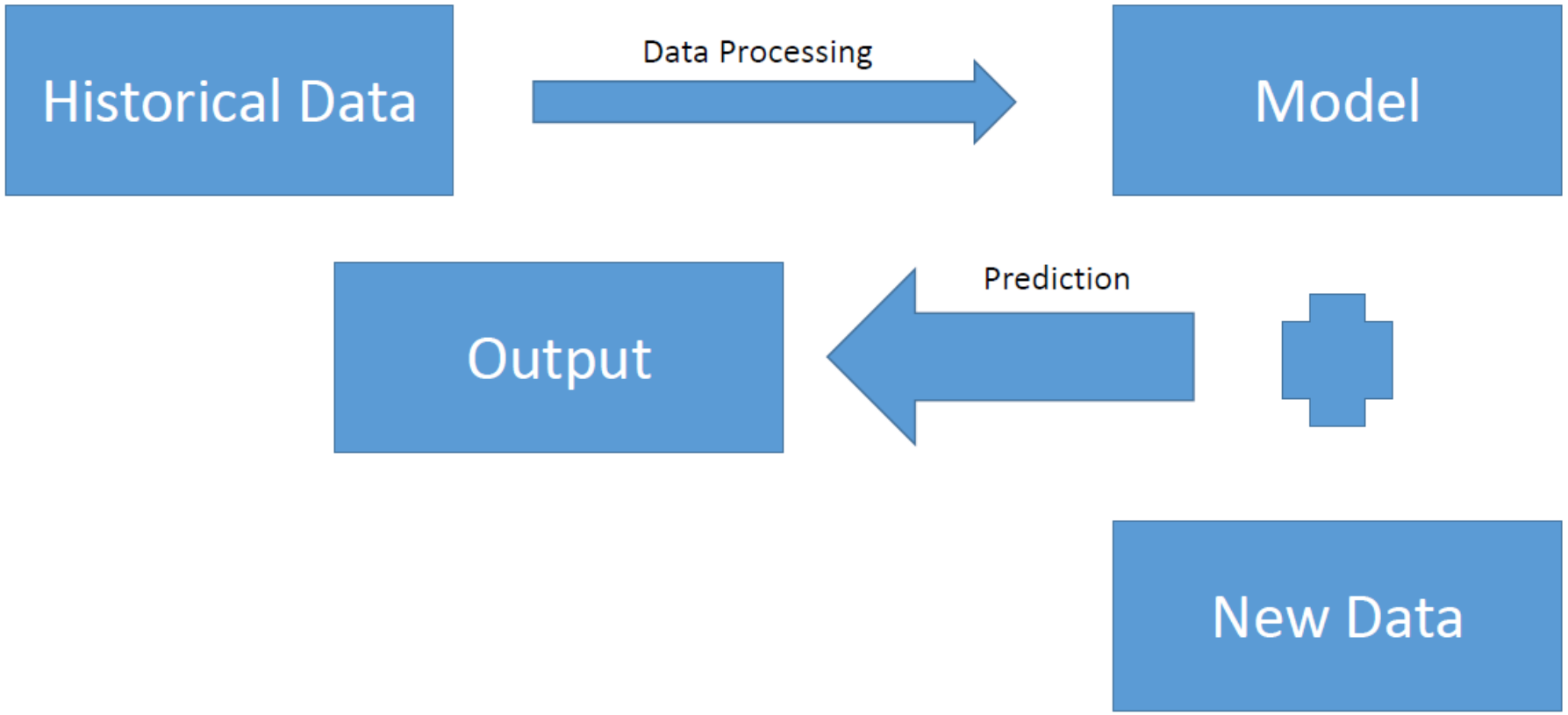
Agenda

- Introduction to Data science Part II
- Data Science Process from a simple to KDD
- KDD process – Data Mining
- Workshop 2

The data analytic



Data Science operation (naive)



Think like data science

1

Knowledge first— Get to know your problem, your data, your approach, and your goal before you do anything else, and keep those at the forefront of your mind.

2

Technology second— Software is a tool that serves you. It both enables and constrains you. It shouldn't dictate your approach to the problem except in extenuating circumstances.

3

Opinions third— Opinions, intuition, and wishful thinking are to be used only as guides toward theories that can be proven correct and not as the focus of any project.

Data Science vs. Statistics

- Intersection of three areas (math/stat, computation and a domain)
- Have a lot in common
- Both look for relationship within data
- More often, statistic starts with hypothesis, use **primary data**
- Data science usually use **secondary data** to discover novel patterns and relationships
- Different in size of data (i.e. 100 samples is enough for statistician)
- Read more at (Carmichael and Marron 2018)

Data Science vs Machine-learning

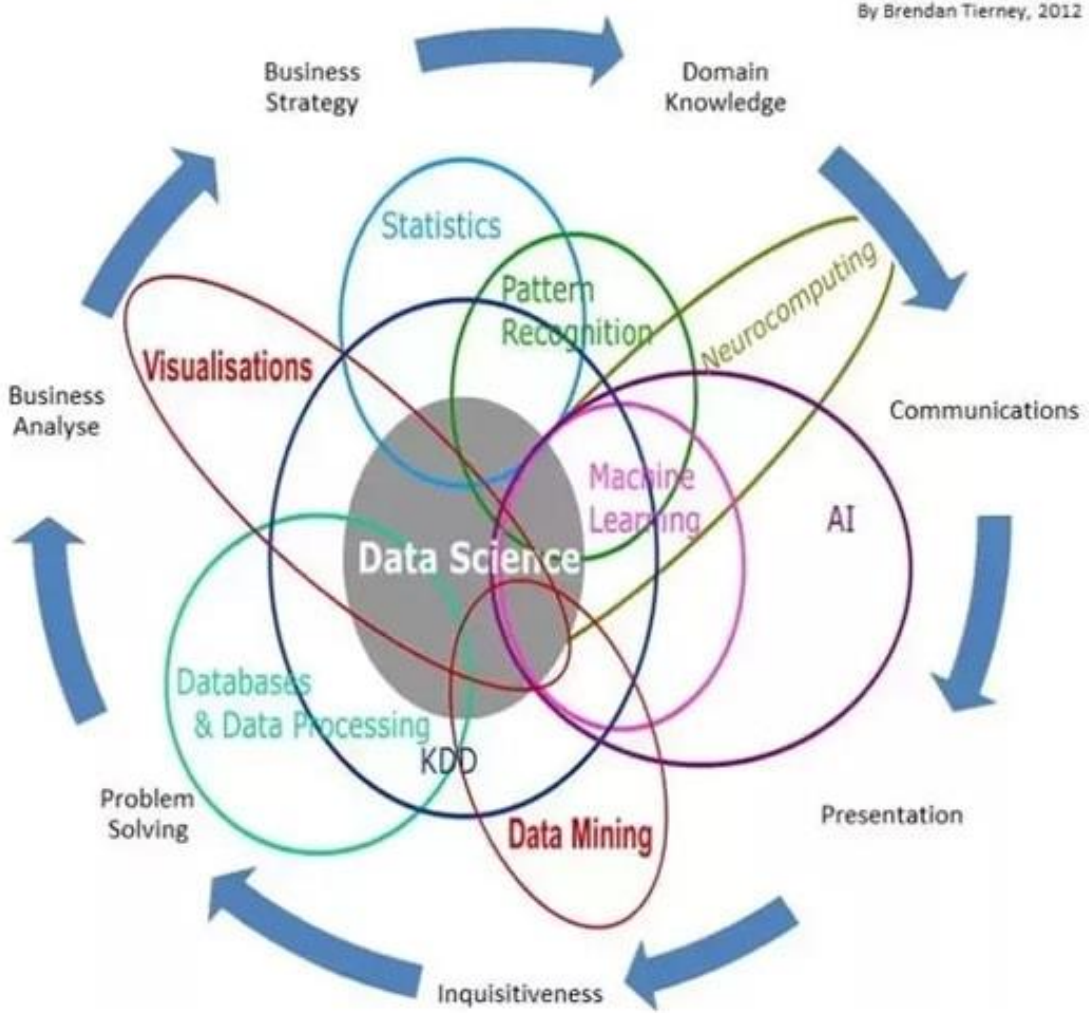
- *"A computer program is said to learn from experience with respect to some class of tasks and performance measure if its performance at tasks in T, as measured by P, improves with experience E."*
  - Mitchell, T. M. (1997). Machine learning. 1997.*Burr Ridge, IL: McGraw Hill*,*45*(37), 870-877.

Data science	Machine learning/ ML Engineer
Skilled in **Math/stat and modeling**	Utilize the proposed models
Prove the properties of the learning model/interpret results	Optimize the model/automation/Parallelizing
Domain expert	Skilled in **programming/technical**
Academic propose	Technical/practical propose

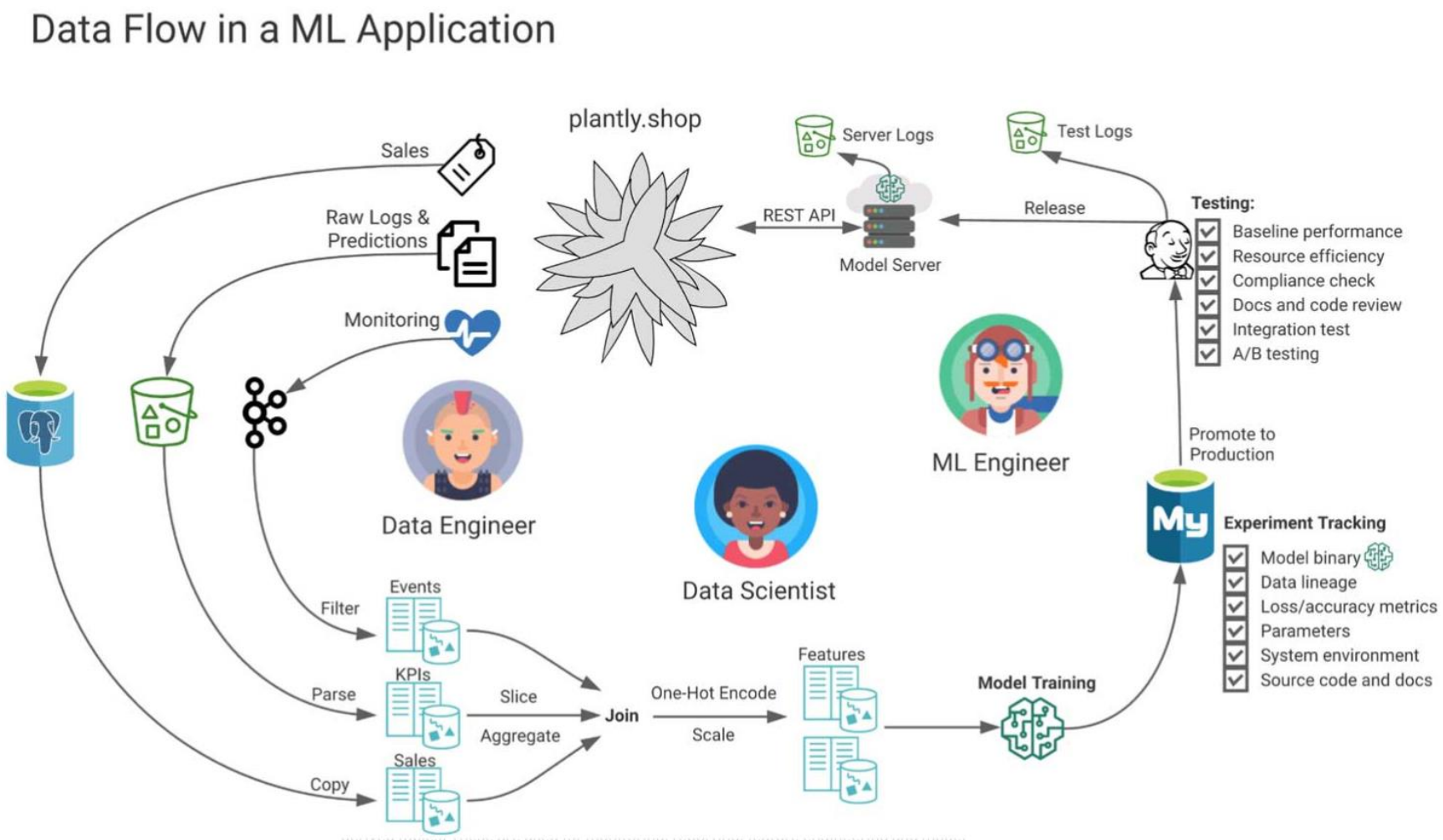
Data science vs Data Mining

Data Mining	Data Science
Business process	Scientific study
A technique to find the trends in a data set and using these trends to identify future patterns	A field study
Make data more usable	Building data-centric product
Subset of data science	Multidiscipline of Big Data Analytics, Data Mining, Predictive Modeling, Machine learning, Data Visualization, Mathematics, and Statistics

Data Science Is Multidisciplinary



Example of Data Flow in process (back-end)

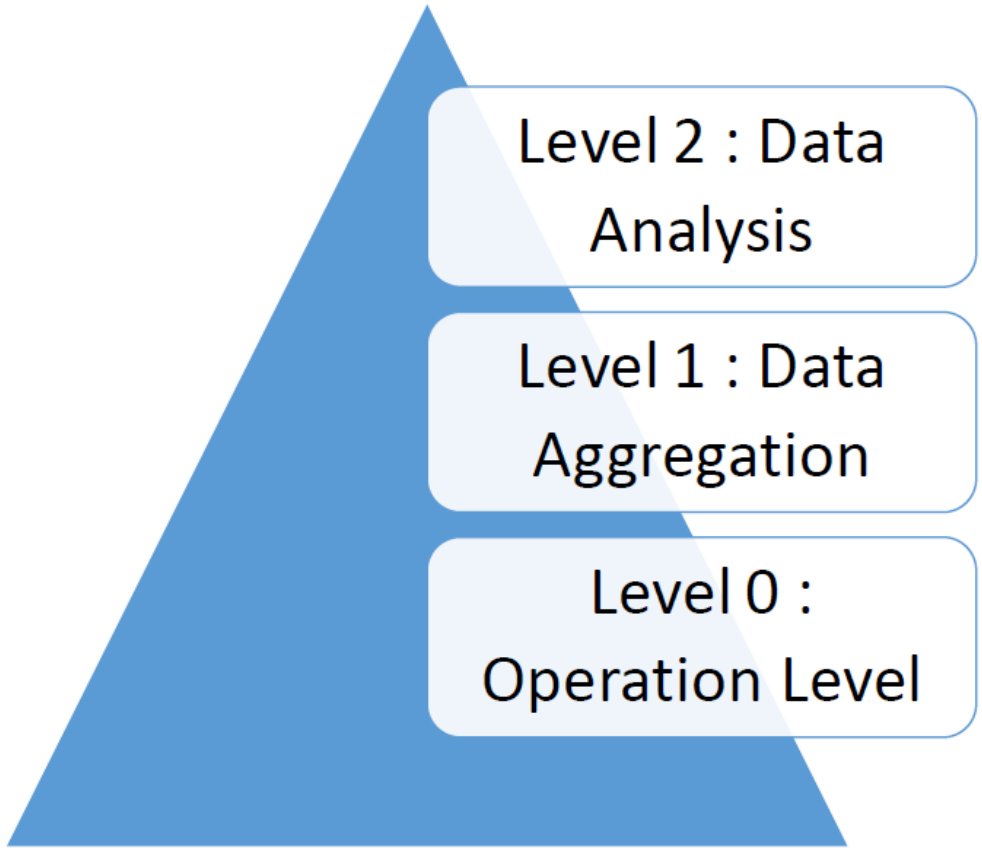


Historical preceptive

- ‘Data Science’ has been around the 1960s but back then it was used as an alternative to ‘Computer Science’. Presently, it carries a completely different meaning.
- In 2008, D. J. Patil and Jeff Hammerbacher became the first individuals to call themselves ‘Data Scientists’ in order to describe their role at LinkedIn and Facebook respectively.
- In 2012, Harvard Business Review article cited Data Scientist as the ‘Sexiest Job of the 21st Century’.
- The term Data Mining has evolved parallelly. It became prevalent amongst the database communities in the 1990s.
- Data Mining owes its origin to KDD (Knowledge Discovery in Databases). KDD is a process of finding Knowledge from information present in databases. And Data Mining is a major subprocess in KDD.
- Data Mining is often used interchangeably along with KDD.

Data Science Usage and Application

Level of Information technology usage



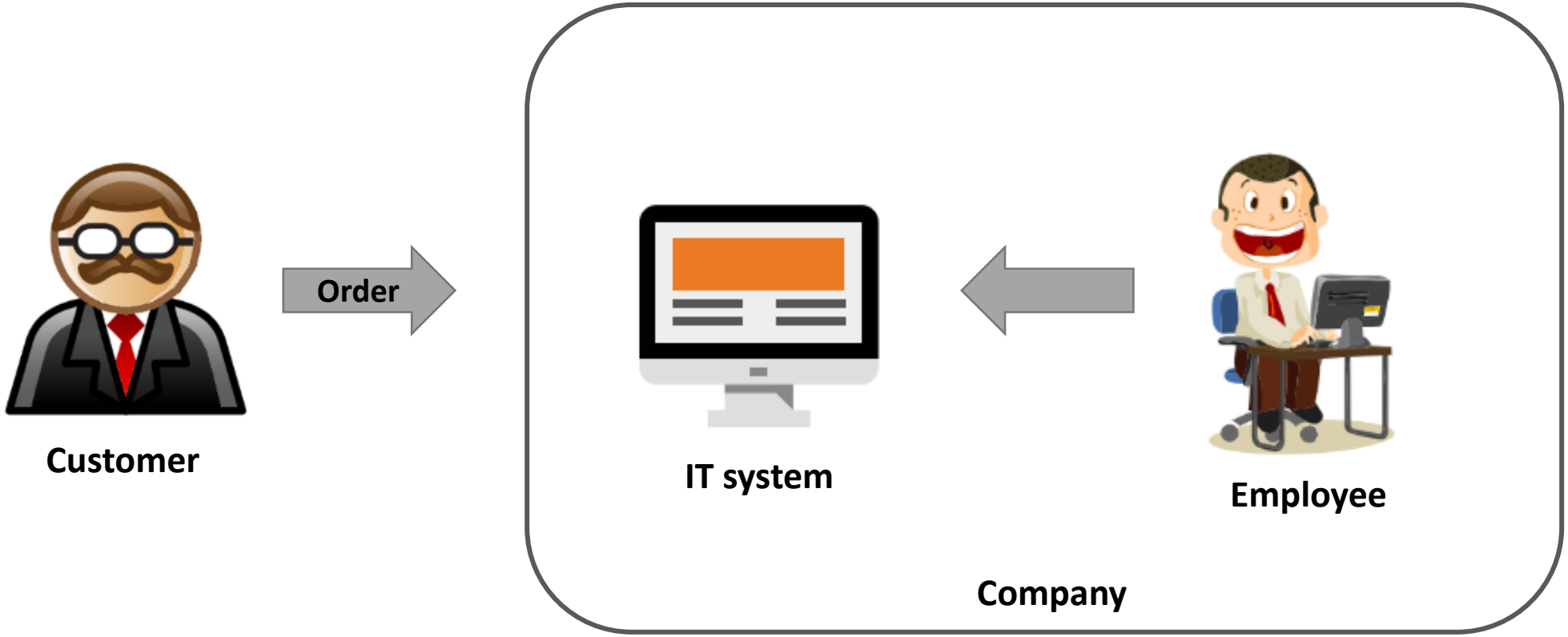
Help designing the direction

Help decision making process

Streamline inefficiency

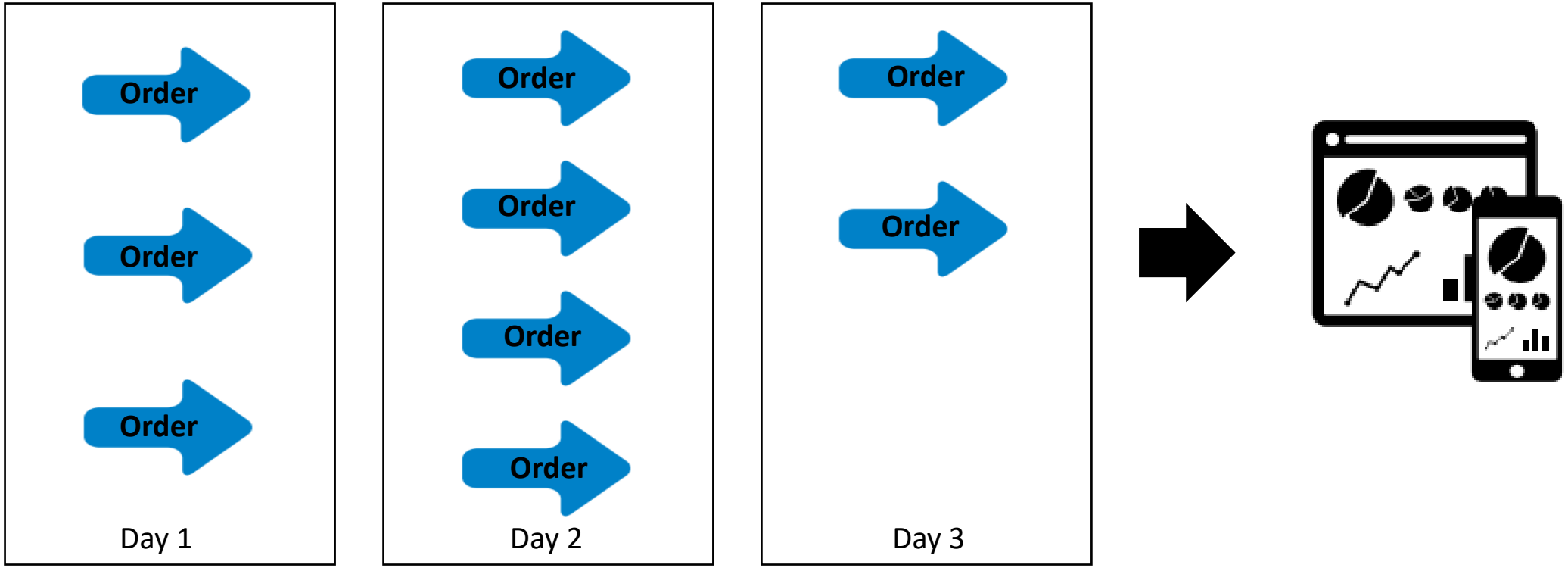
Level of Information Technology Usage

Level 0: Operation level



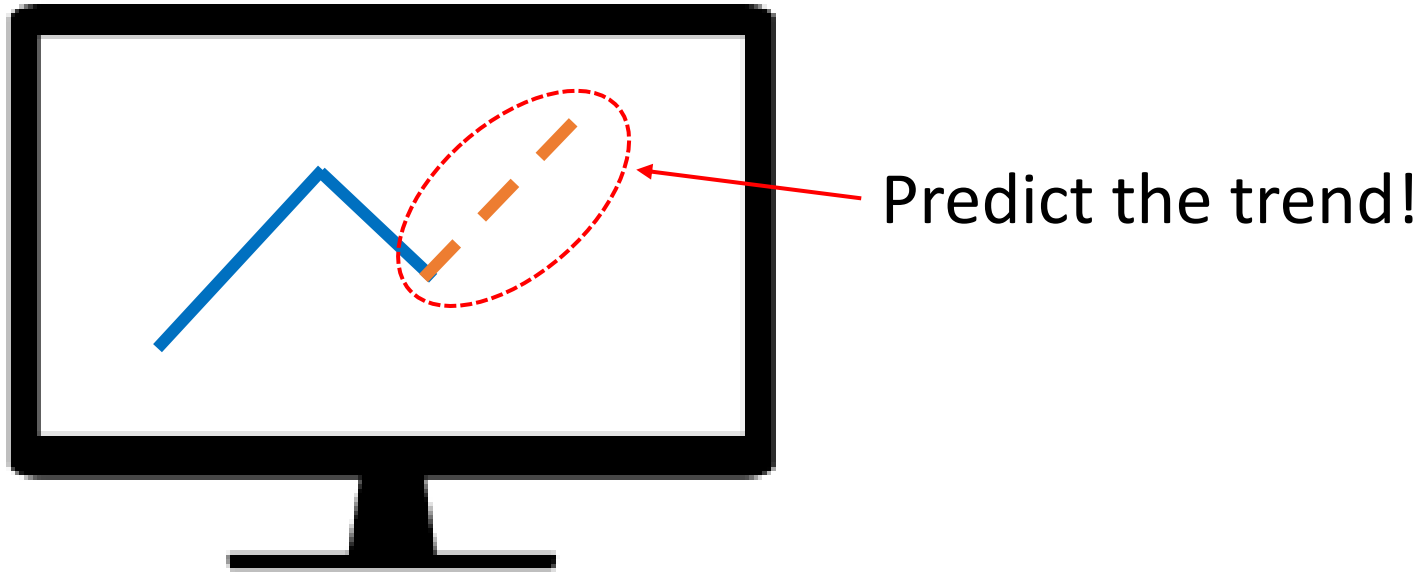
Level of Information Technology Usage

Level 1: Data aggregation



Level of Information Technology Usage

Level 2: Data analysis



Benefit of the DS tools for **Business** tools

Improve the return on its direct marketing investment

Select **optimal** site locations

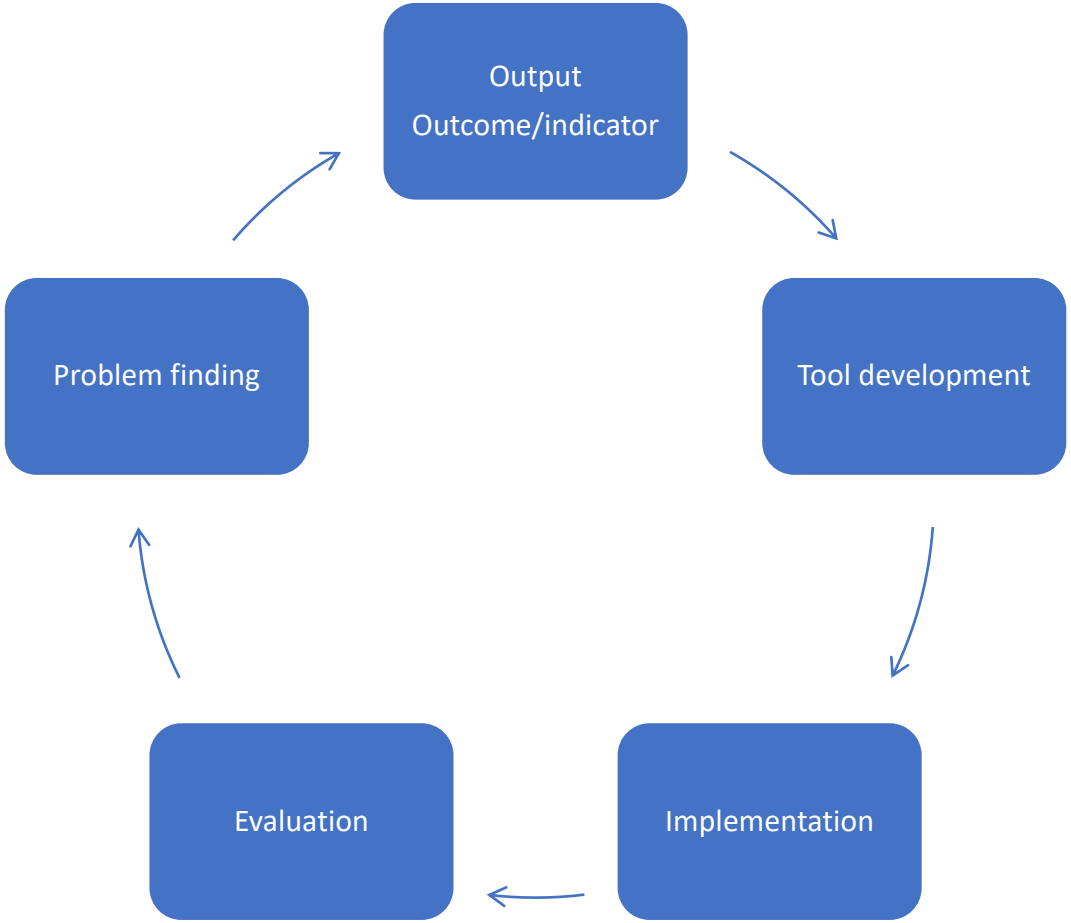
Understand the value of customers across all channels

Design promotional offers that best **enhance** sales and profitability

Tailor direct marketing offers to **customer preferences.**

Use of Information Technology in Business (Data Science)

Usage of Information Technology in Business



Problem identification

- Review the environment or contexts of the problem
- Significant academic vs. practical impact
- Talk to your stakeholders

Outputs, Outcomes and Indicators identification

Outputs

- Outputs are a quantitative summary of an activity
- I.e., the activity is ‘we provide training’ and the output is ‘we trained 50 people to NVQ level 3’. An output tells you an activity has taken place.

Activity	Output
CV checking drop ins	Number of people getting support with their CV
Parenting skills classes	Number of people attending parenting skills classes
Cardio vascular health checks	Number of health checks conducted

Source: https://www.kent.gov.uk/__data/assets/pdf_file/0009/41499/Community-Mental-Health-and-Wellbeing-Service-Market-Engagement-event-Julia-Slay-presentation.pdf

Outcome

- The **change that occurs** as a result of an activity (e.g., improved well-being of training participants)
- Outcome : change direction + target component
- Need to be cleared
- Sometimes it takes years for outcomes to take place

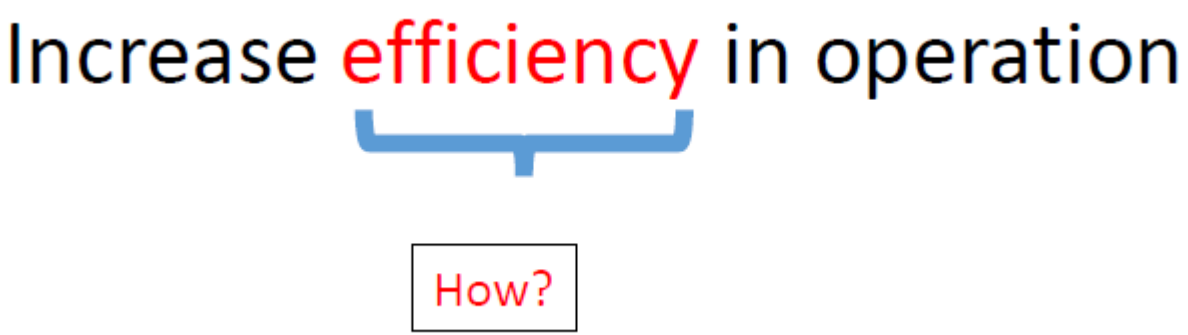
Example of outcomes:

- Reduce labor cost in organization
- Reduce computation time during training model
- Increase predictive accuracy power of the model.
- Increase usability and user experience of the recommendation system

Outcome (cont.)

- Good outcome

- Poor outcome



Indicators identification

- To identify the desirable outcome in term of processes or results (i.e., to measure something)
- Usually present in in number of percentage (ratio of, percentage of)
- Indicators can be shared: reduced school drop-out rates = graduation rate
- Good indicators must be simple, reliable and valid.
- Stakeholders are often the best people to help you identify indicators, so ask them how they know that change has happened for them

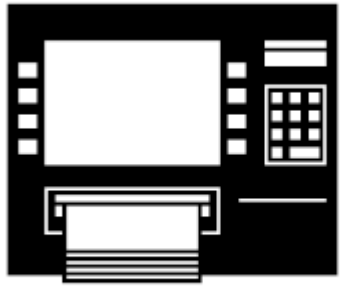
Example of indicators

Outcome	Indicator
Increased infant breastfeeding	Number & percentage of mothers who are exclusively breastfeeding up to six months of age.
Improved work attendance by District Officials	Number of work days attended per year by District Officials
Less grade repetition	Pass rate
Beneficiaries access financial support for tertiary education	Number and percentage of beneficiaries that have bursaries and student loans

Solution Development

- The objective of this step is to develop a tool to solve the problem.
- The first step is to develop the **strategy**.
- The problem solving strategy is a conceptual framework to solve the problem.
- This step does not include the specification of the solution.
- The second step is to develop the **solution**.
- 2 types of solution : develop by yourself or use the existing solution.

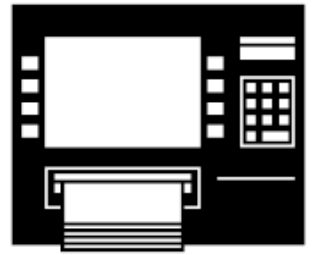
Problem identification



Problem: The institute rents the building and the labor cost is the second highest cost.

Activities	Outputs
Deployed ATM across the region.	Number of ATM machines being deployed Number of people have used
Online banking	Number of transaction

Outcomes and indicators



Outcomes	Indicators
Reduce the cost of labors	Percentage of cost of labors / months
Reduce the cost of renting the building	Percentage of cost of renting spending / months

Solution Development

Problem

The institute rents the building and the labor cost is the second highest cost.

Strategy

Develop a novel approach which does not need to rent and use less employee.

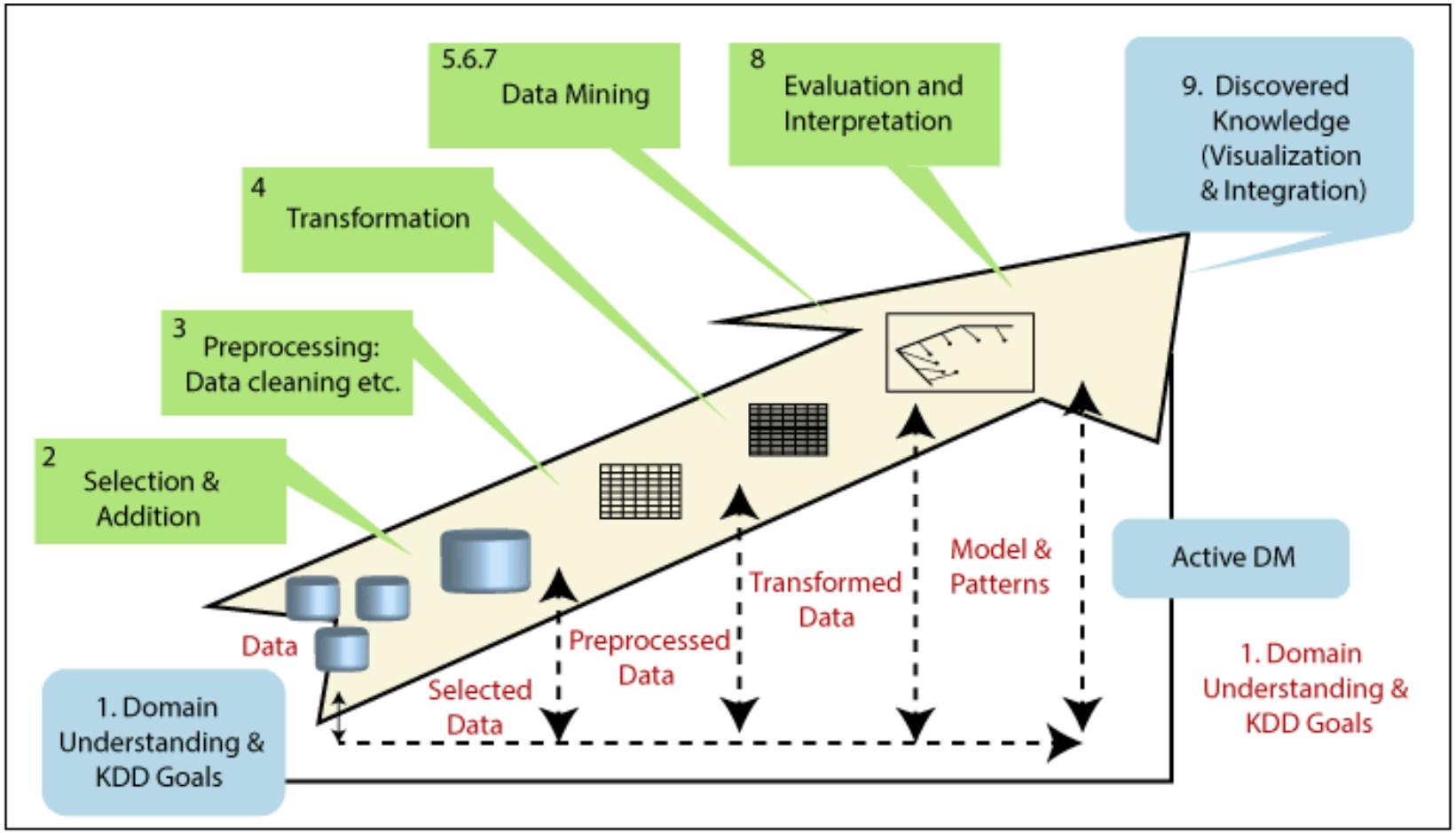
Data science simple process in 1997

In 1997, University of Michigan statistics professor C.F. Jeff Wu

Knowledge Discovery in Databases Process (KDD)

- is the process of finding valid, novel, useful and understandable patterns in data, to verify hypothesis of the user or to describe/predict the future behavior of some event

Knowledge Discovery in Databases Process

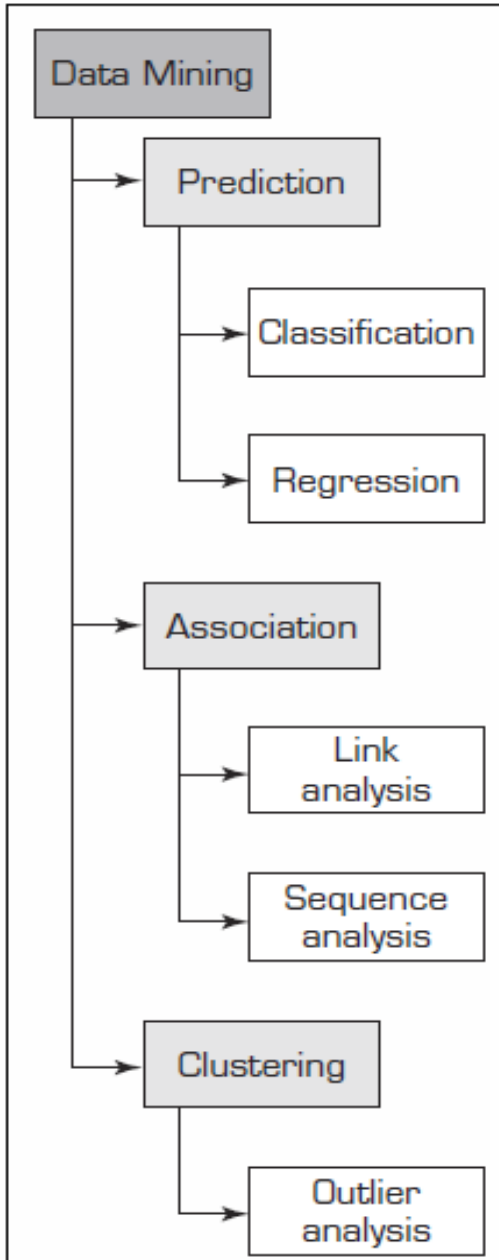


Data Science/Data Mining - Method

- Large amount of data
- Faster computation power.
- Data Mining method can be classified either **supervised** and **unsupervised**
- **Supervised learning – human intervene (help labeling)**
- **Unsupervised learning – let the model work on its own (deal with un-labellled data)**

Data science -The methods

- **Predictions**- predict the winner of football match
- **Associations** – find the commonly co-occurring group of things (beer and chips in shelf)
- **Clusters** – identify natural grouping of things based their own attributes.
- **Sequential relationships** – discover time-order event. (banking customer has c-account will open open i-account with in a period)



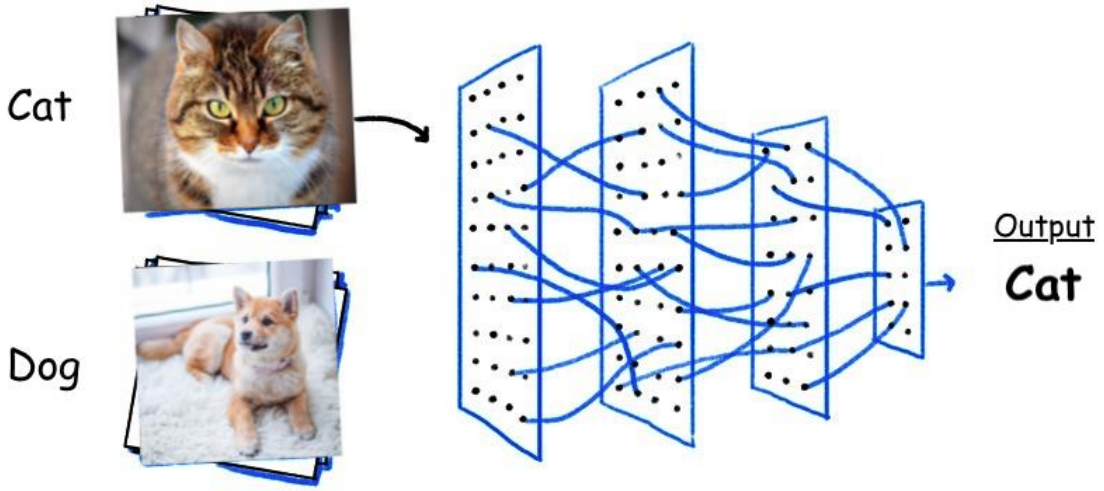
Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision Trees, ANN/MLP, SVM, Rough Sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression Trees, ANN/MLP, SVM
Unsupervised	Apriori, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization Apriori Algorithm, Graph-Based Matching
Unsupervised	Apriori Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

Predictions

- Guessing, predicting, forecasting, and recommending
- Tell the nature of future occurrences of certain events based on what has happened in the past
- I.e. forecasting the absolute temperature of a day.

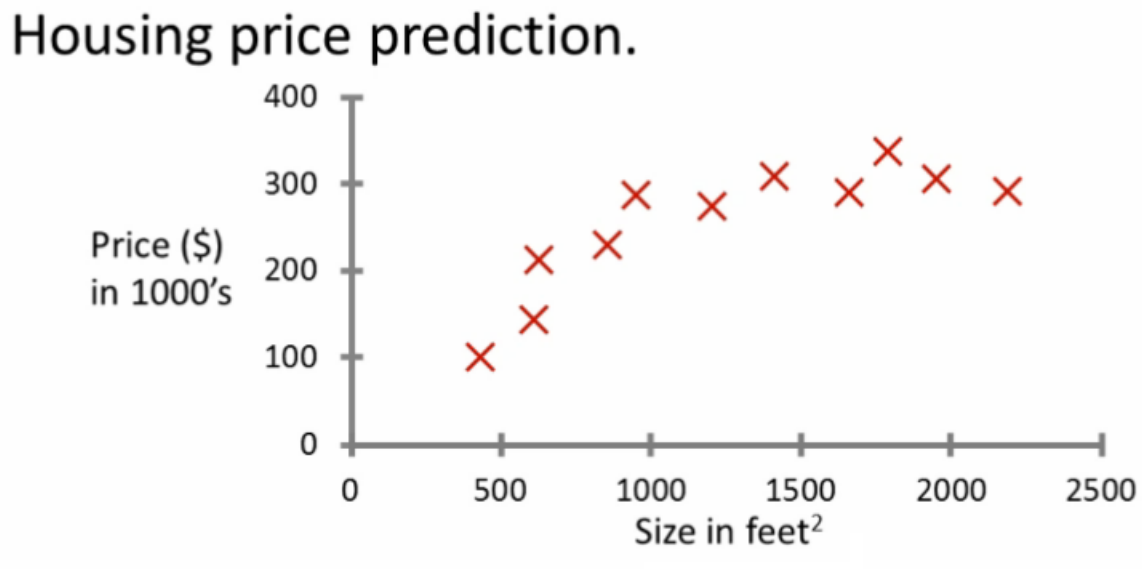
Classification

- Most **common** of all data mining tasks
- A **forced choices** or **known choices**.
- Analyze historical data and generate a **predictive model**.
- Hope that the model can be used to predict the future unclassified records
- Common classification algorithms – NN, DT, Logistic regression



Regression

- To predict value of dependent variable, based on its relationship with values of at least one independent variable.
- Explain the impact of changes in an independent variable on the dependent variable.

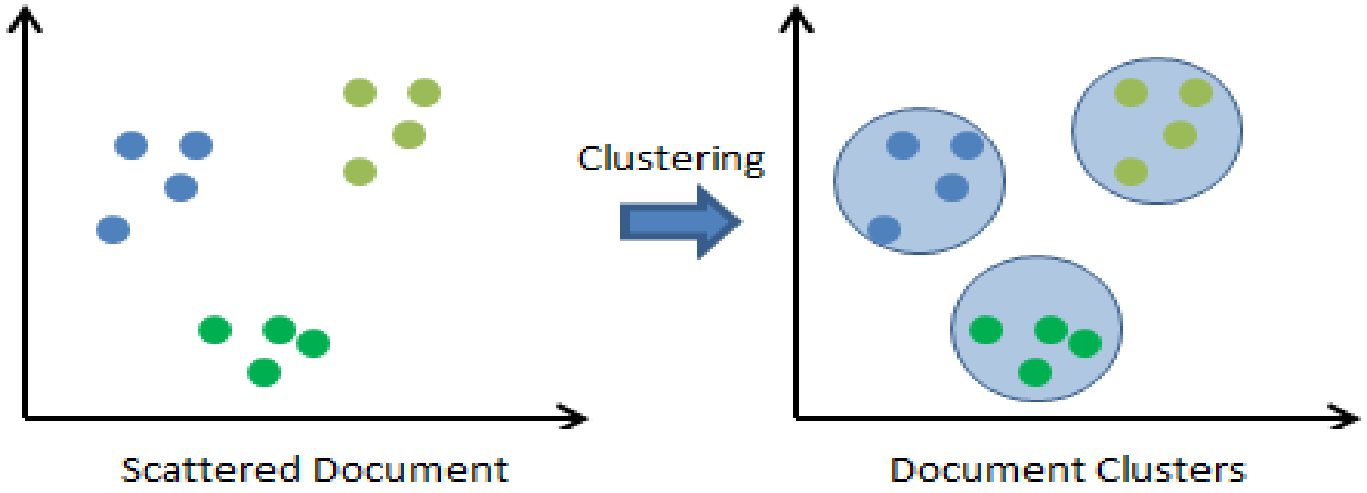


Clusters

- Identify natural groupings of things based on their known characteristics,
- I.e. assigning customers in different segments based on their demographics and past purchase behaviors.

Clustering

- Partitions a collection of things
- E.g. objects, events, etc.
- Class label are unknown



Associations

- Find association among your problem attributes or variables
- E.g. Find relations such as a patient with high-blood-pressure is more likely to have heart-attack disease.
- E.g. Find a products that customers usually purchased together.

Association Rules

- Also known as **market basket analysis**
- Association rules helps uncover relationship between items from large databases
- C1 – {Milk, Eggs, Sugar, Bread}
- C2 – {Milk, Eggs, Cereal, Bread{
- C3 – {Eggs, Sugar}
- Find associations/correlation between the different items that customers place in their basket? Which product are bought together?
- *Apriori* algorithm method
  - Frequent itemset
    - Itemset construction
    - Support count
  - Associate rules

Sequence analysis

- Discover time-ordered events.
- i.e. predicting that an existing banking customer who already has a checking account will open a savings account followed by an investment account within a year.
- Gene prediction
- Protein structure prediction
- Heath Informatics

Assessment

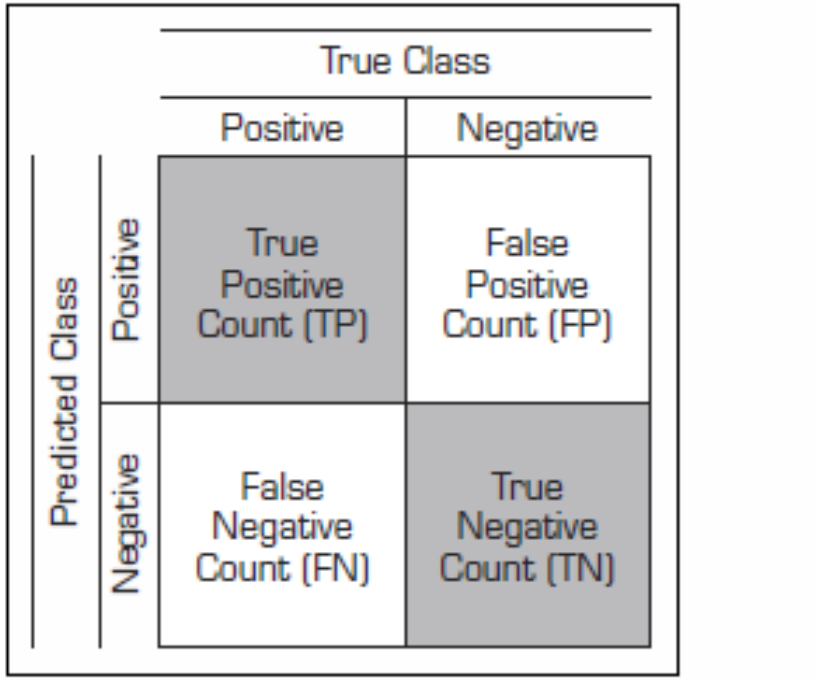
- **Predictive accuracy** – with unseen data how well the model perform in terms of %
- **Speed** – computation cost when constructing and using the model
- **Robustness** – Giving noisy data, can the model still make reasonable prediction
- **Scalability** – how’s about with larger data?
- **Interpretability** – level of understanding

Estimating the true accuracy of models

$$(True\ Classification\ Rate)_i = \frac{(True\ Classification)_i}{\sum_{i=1}^{n}(False\ Classification)_i}$$

$$(Overall\ Classifier\ Accuracy)_i = \frac{\sum_{i=1}^{n}(True\ Classification)_i}{Total\ Number\ of\ Cases}$$

Confusion matrix (getting more insight)



Estimating the error of regression models

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

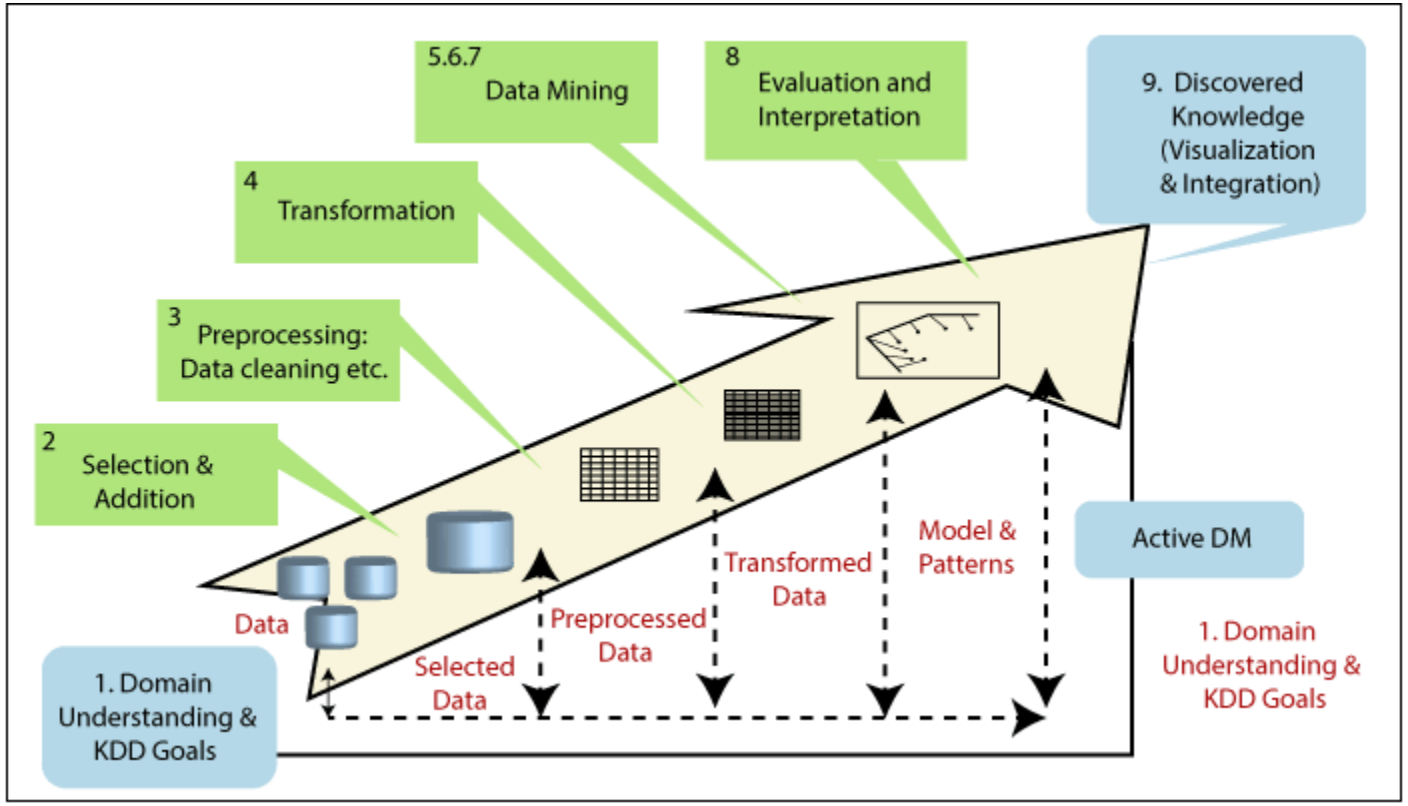
$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

Your second toy dataset (Iris dataset)



Iris Versicolor **Iris Setosa** **Iris Virginica**

- https://archive.ics.uci.edu/ml/datasets/iris

Your second toy dataset

	A	B	C	D	E
1	Sepal Length	Sepal Width	Petal Length	Petal Width	Class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3	1.4	0.1	Iris-setosa
15	4.3	3	1.1	0.1	Iris-setosa
16	5.8	4	1.2	0.2	Iris-setosa
17	5.7	4.4	1.5	0.4	Iris-setosa
18	5.4	3.9	1.3	0.4	Iris-setosa
19	5.1	3.5	1.4	0.3	Iris-setosa
20	5.7	3.8	1.7	0.3	Iris-setosa
21	5.1	3.8	1.5	0.3	Iris-setosa
22	5.4	3.4	1.7	0.2	Iris-setosa
23	5.1	3.7	1.5	0.4	Iris-setosa
24	4.6	3.6	1	0.2	Iris-setosa
25	5.1	3.3	1.7	0.5	Iris-setosa

Domain/Data Understanding

- How many features?
- How many sample?
- What are they?
- What DS task shall we perform?
- How do we do it?

Workshop

- Write 1 page essay in English for one of the three case of your choice.
- You may discuss with your classmates, but you need to write on your own.
- Your essay must include the follow topics

Workshop 2

- **Business objectives, define problem (no more than one problem)**
- **Identify activities, outputs, outcomes, and indicators identification**
- **Identify Stakeholders** (who involves)
- **Identify Data source** (where can you get?, How does it look like?)
- **Identify DM technique** (which Data science/ Datamining technique you might use?)
- **Data as data product (solution)** (what should be your output product to the users or stakeholders?)

Workshop 2

- Work as a group (maximum of 3 people)
- Write 1-2 proposal pages including topics I just mentioned
- You may insert figures, tables
- Submit to the google classroom under the workshop 2

Case study 1 : Entrance problem in university campus (easy)

Case study 2: A university campus public transport (medium)

Case Study 3: The 2020 United States presidential election (difficult)

References:

- Carmichael, Iain, and J. S. Marron. 2018. “Data Science vs. Statistics: Two Cultures?” *Japanese Journal of Statistics and Data Science* 1(1):117–38.
- http://www.dgmt-growingconfidence.co.za/content/how-selectidentify-and-write-indicators
- https://www.kent.gov.uk/__data/assets/pdf_file/0009/41499/Community-Mental-Health-and-Wellbeing-Service-Market-Engagement-event-Julia-Slay-presentation.pdf