Natural Language Processing for SE 66/2
NLP Overview (2)

Asst. Prof. Pree Thiengburanathum

Where we are now

1.	NLP Overview	3
2.	Data science methodology	3
3.	Word Tokenization, Text preprocessing	3
3.	Text extraction methods	6
4.	Machine-learning models in NLP	6
5.	Deep-learning models in NLP	6
6.	Transformers	3
7.	Evaluation metrics and explain-ability	3
8.	NLP-based Systems	3
9.	Case studies and Project	9

Announcement

- Job possible 3-6 months
  - 1*Full stack (AI-based)
  - 1*UX/UI
  - 1*QA
  - 1*Annotator
- Guest Speaker
- 21th December 14:30-15:30 (Simya lingapp)

Submission guide line

- For source code
  - link to your notebook file (all the cells with output)
  - Github URL (run all the cells with output)
- For Workshop that has only document
  - Please submit using PDF

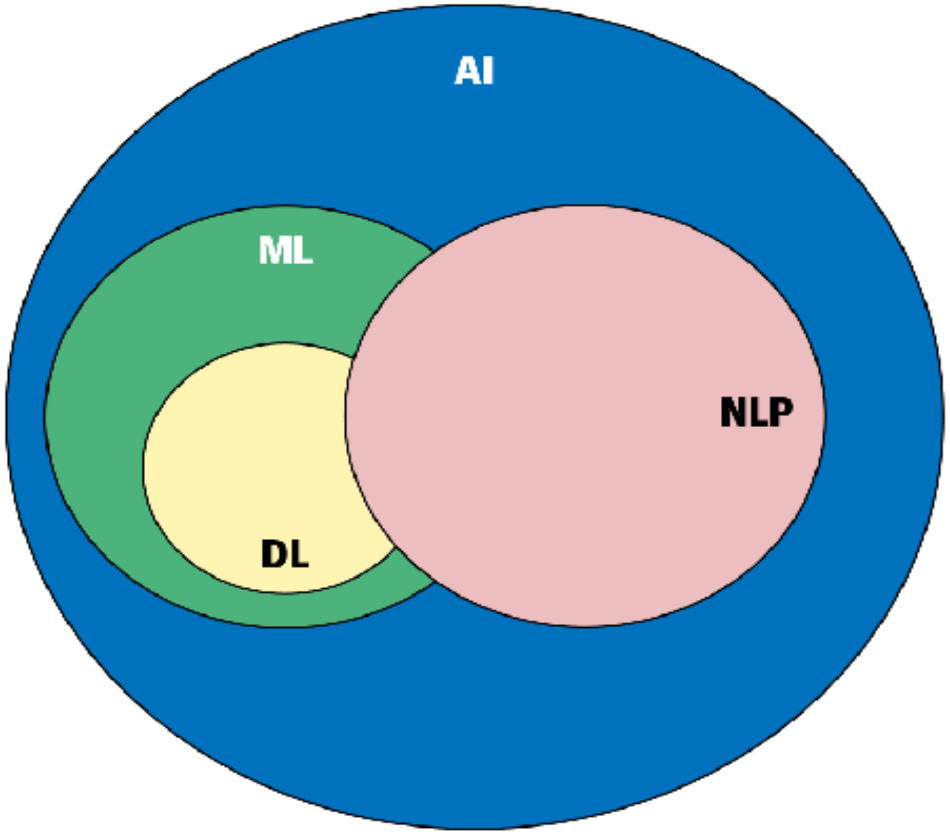
Updated on course outline

- Midterm Examination *25%*
- Final Examination *25%*
- Workshops 25% (6 works)
- Term project *25% 1 group (max 3 people) + present + docs*
- **I have right to adjust grading system based on the student performance.*
- CMU-based (i.e., a grade A cut at 80%)

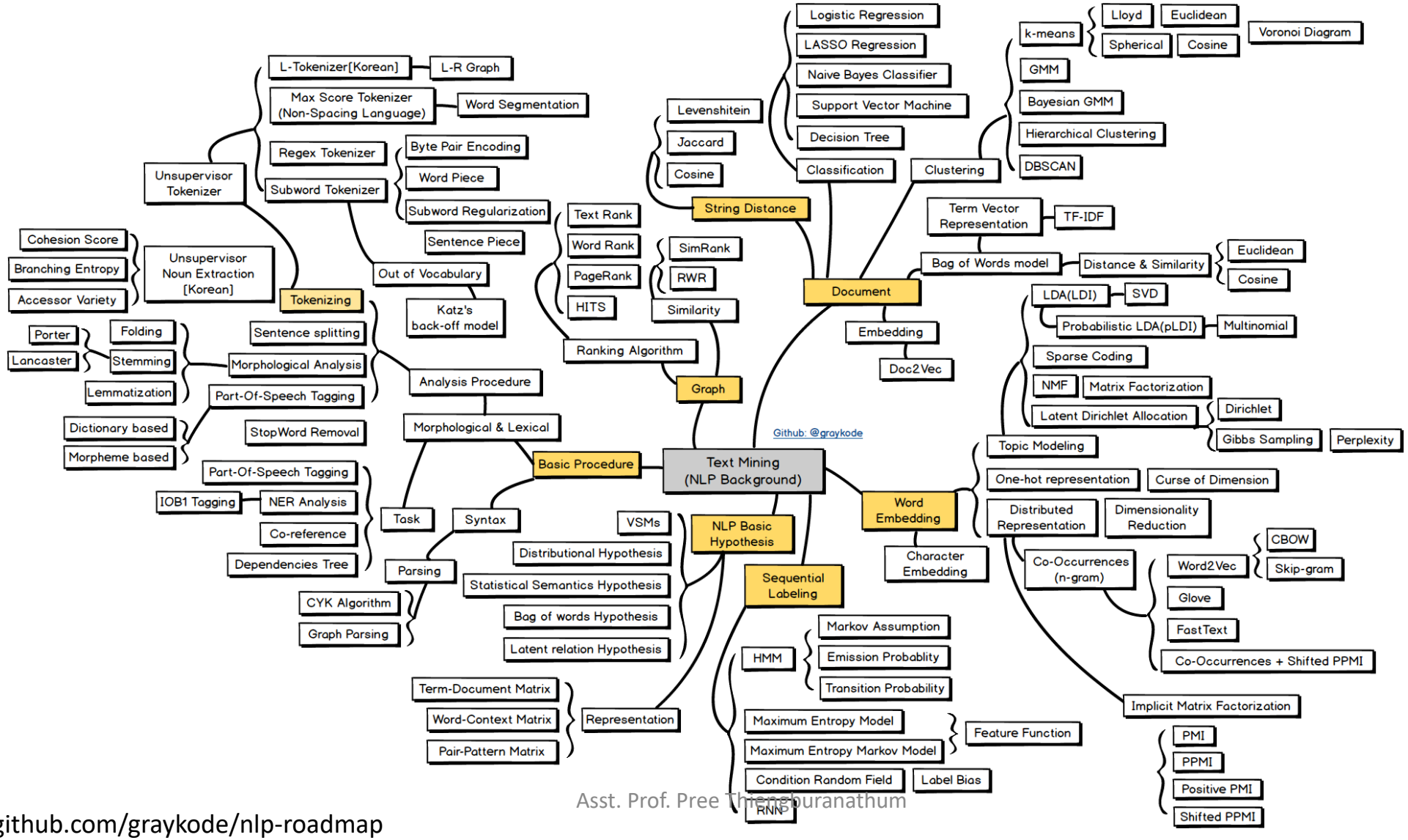
Today’s Agenda

- Use of Information Technology in Business
- DS methodology
- NLP-based project proposal

NLP related fields of studies.

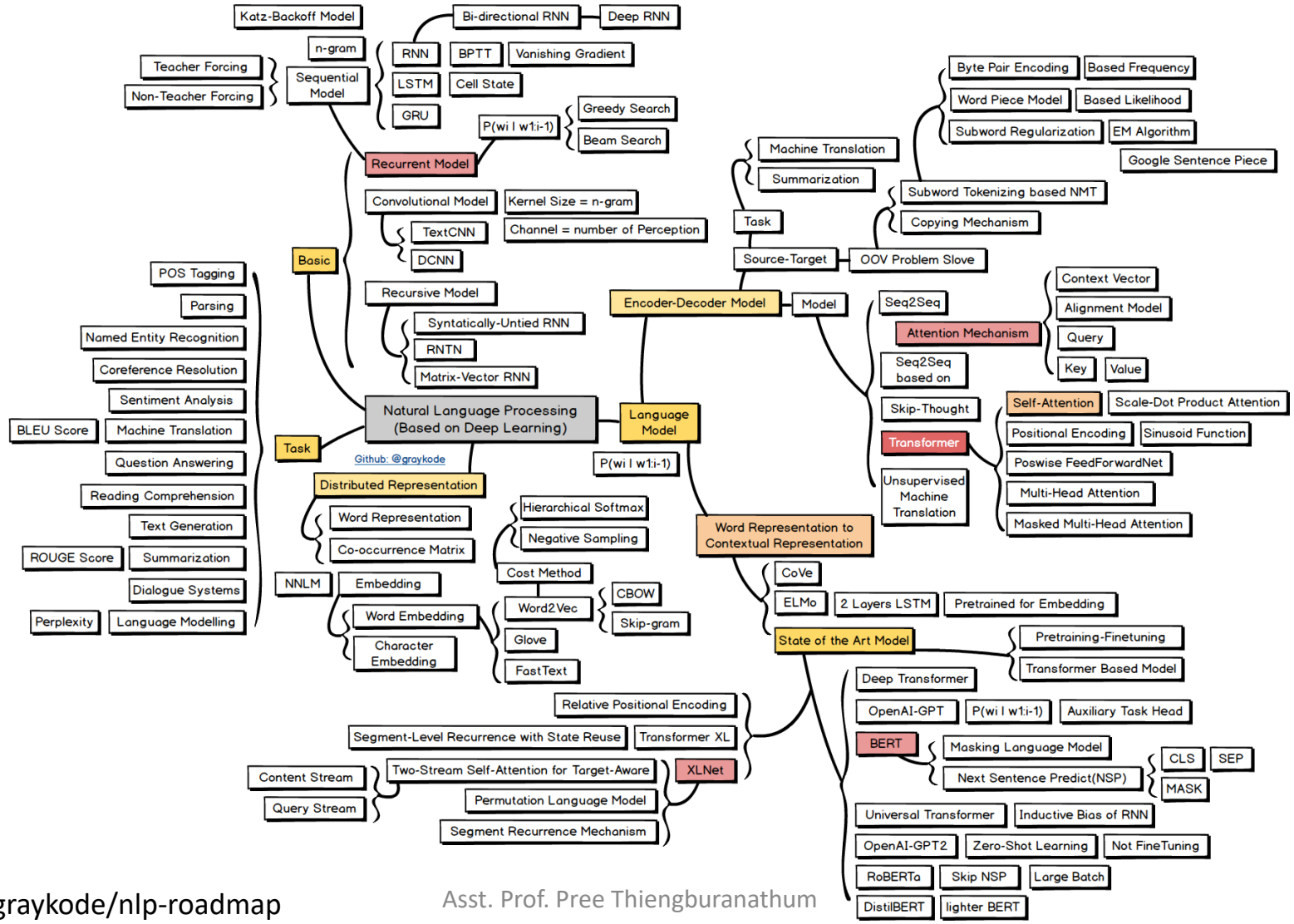


NLP road map



Source:https://github.com/graykode/nlp-roadmap

NLP road map (cont.)



Source: https://github.com/graykode/nlp-roadmap

Why NLP is challenges?

- The **ambiguity** and **creativity** of human language
- Ambiguity - uncertainty of meaning. Most human languages are inherently ambiguous
  - “I made her duck.”
  - “Call me a taxi.”
  - “The teacher said the test would be difficult tomorrow.”
- Creativity – language is not a rule-based driven.
  - Various styles, dialects
  - Poem is a great example.

Challenge in NLP (non-practical)

- **Part of speech tagging (POS-tagging**)- identify Adverb verb, noun in the sentence.
- **Text segmentation** - Chinese/Thai/Other languages.
- **Word sense disambiguation** – a word may has more than one meaning.
- **Syntactic ambiguity** – grammar is ambiguous
- **Imperfect or irregular input** – typos , grammar errors

Developer vs Data scientist

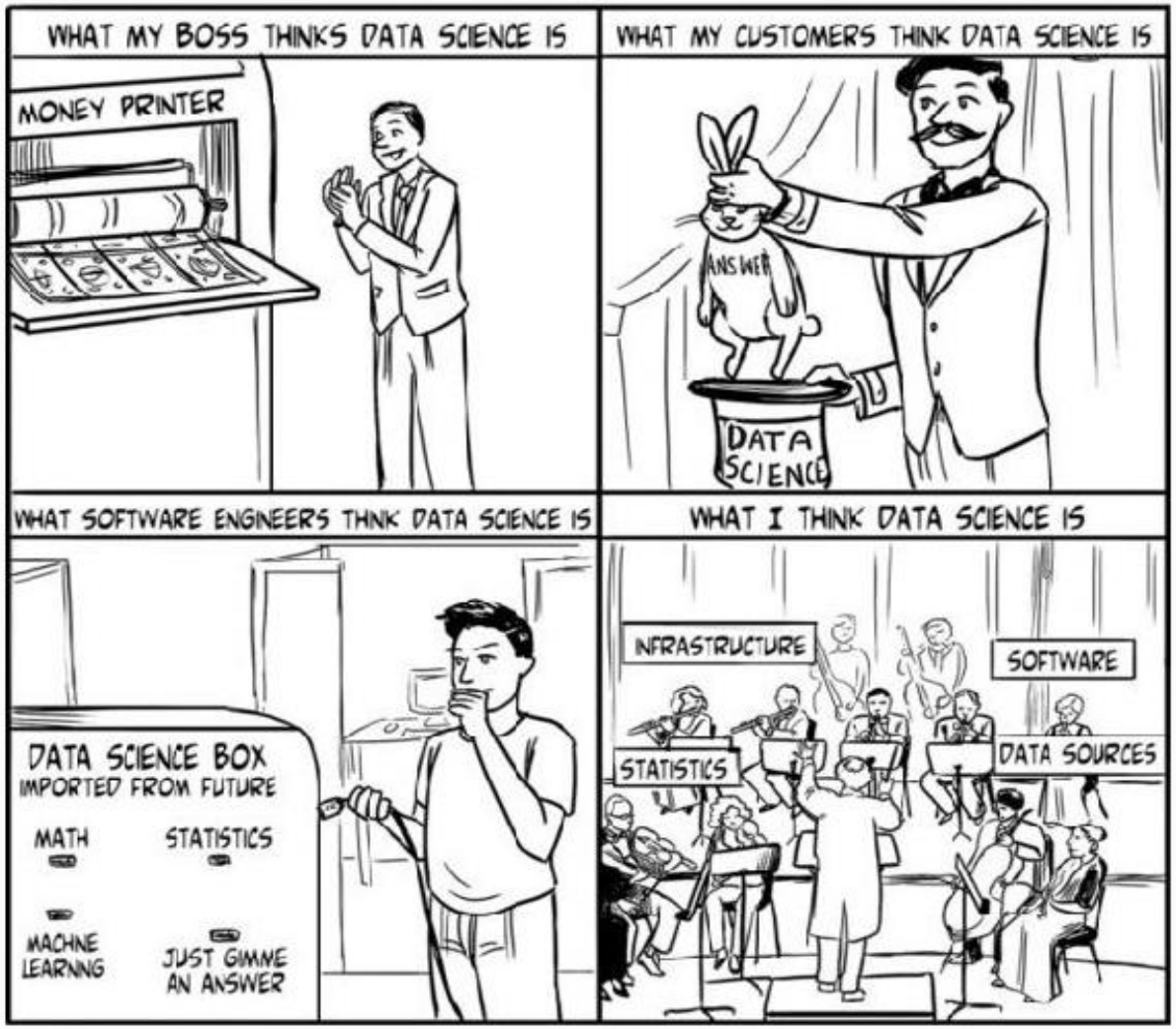
- Somewhat common (good at designing and building complex system, with tools and frameworks)
- Software dev. -> well-defined components
- Data Science -> work on component isn’t well defined (i.e., data pre-processing, analysis)
- Data Science -> create system that rely on statically results

Developer vs Data scientist (cont.)

Dealing with uncertainty is often what separates the role of a data scientist from that of a software developer.

The role of data scientist

Figure 1.1. Some stereotypical perspectives on data science



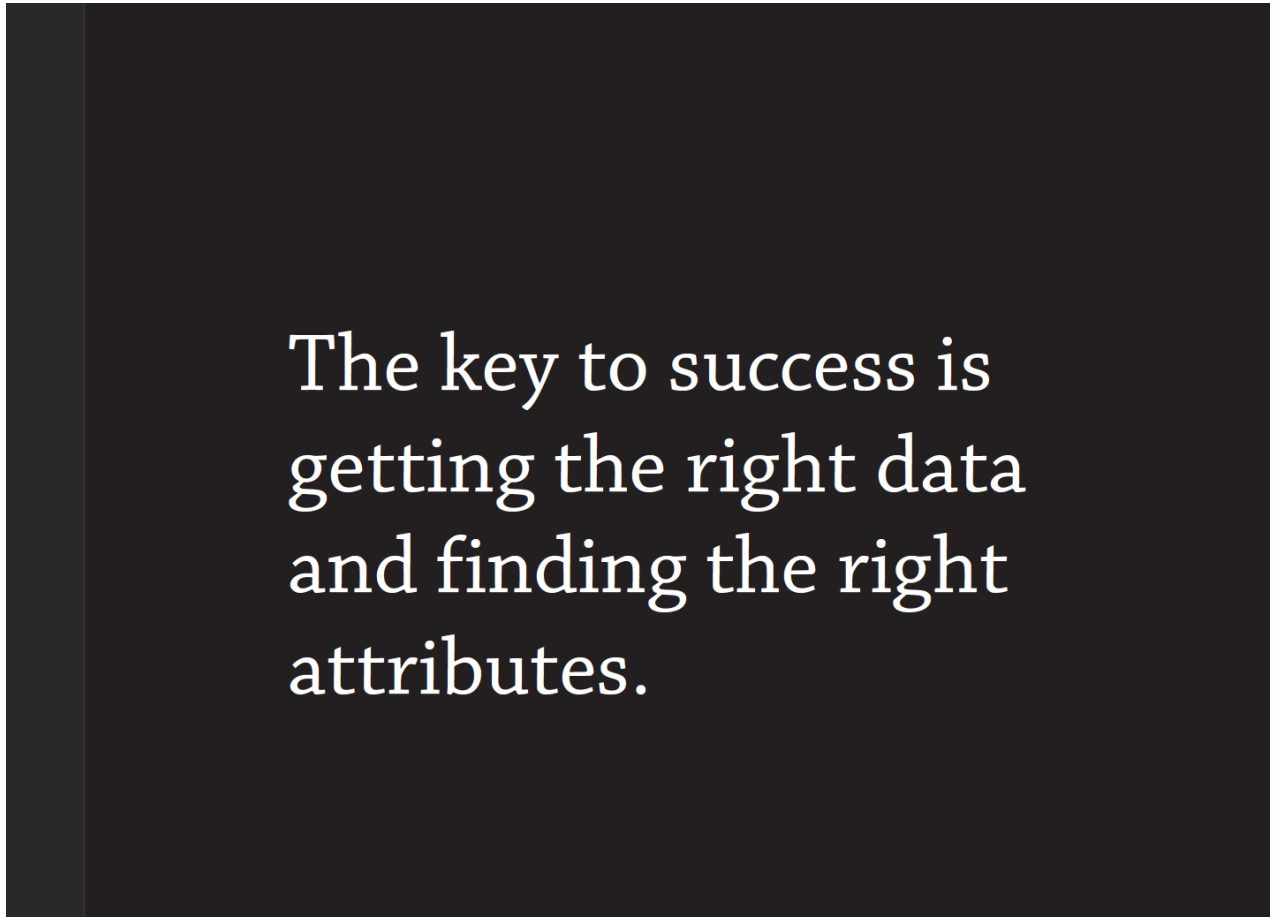
Goal of Data science

- “Find a patterns” Kenny Cheung
- “Turn **data** to data product”

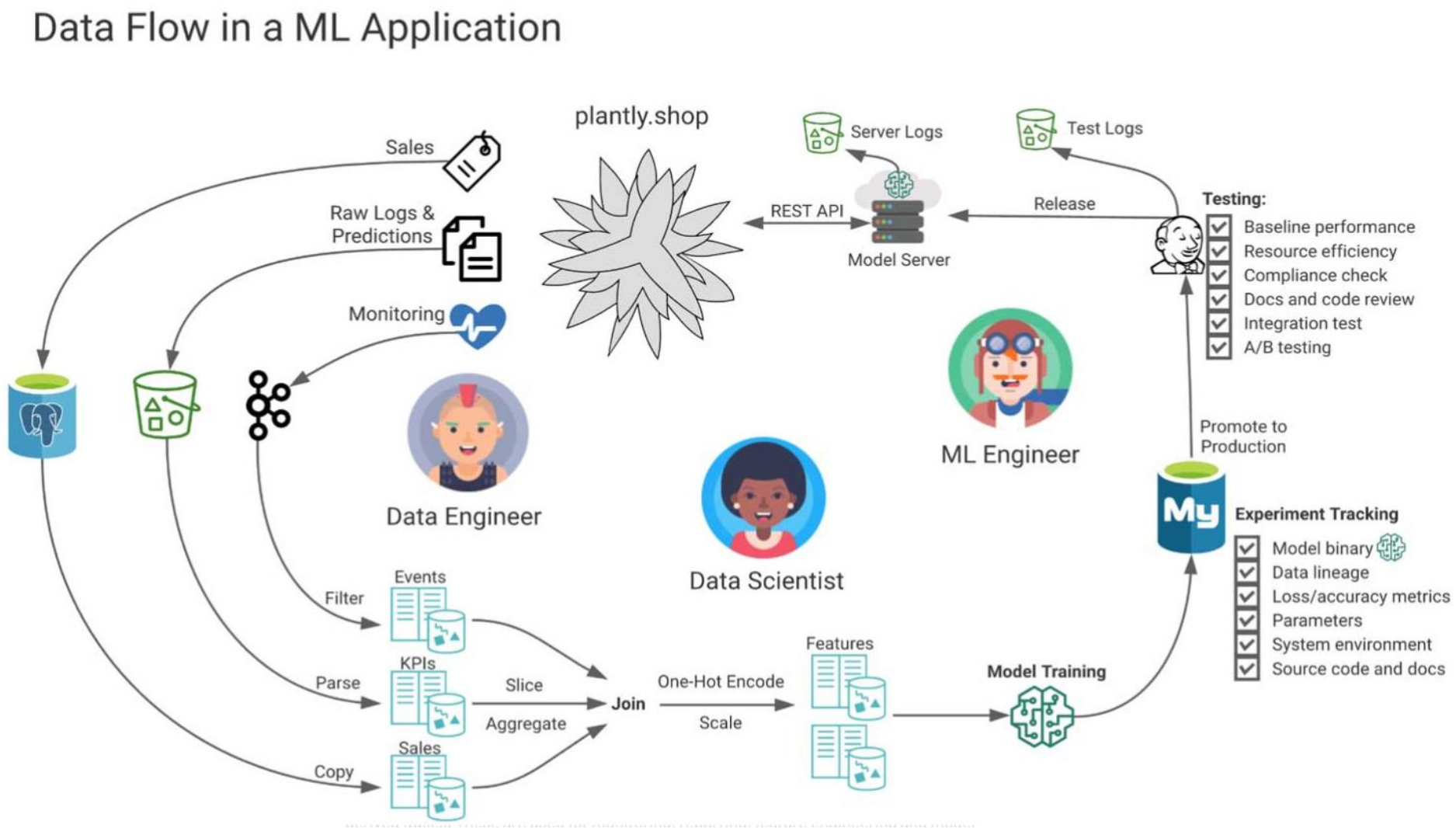
Think like data science

If a human expert can easily create a pattern in his or her own mind, it is generally not worth the time and effort of using data science to “discover” it.

Think like data science

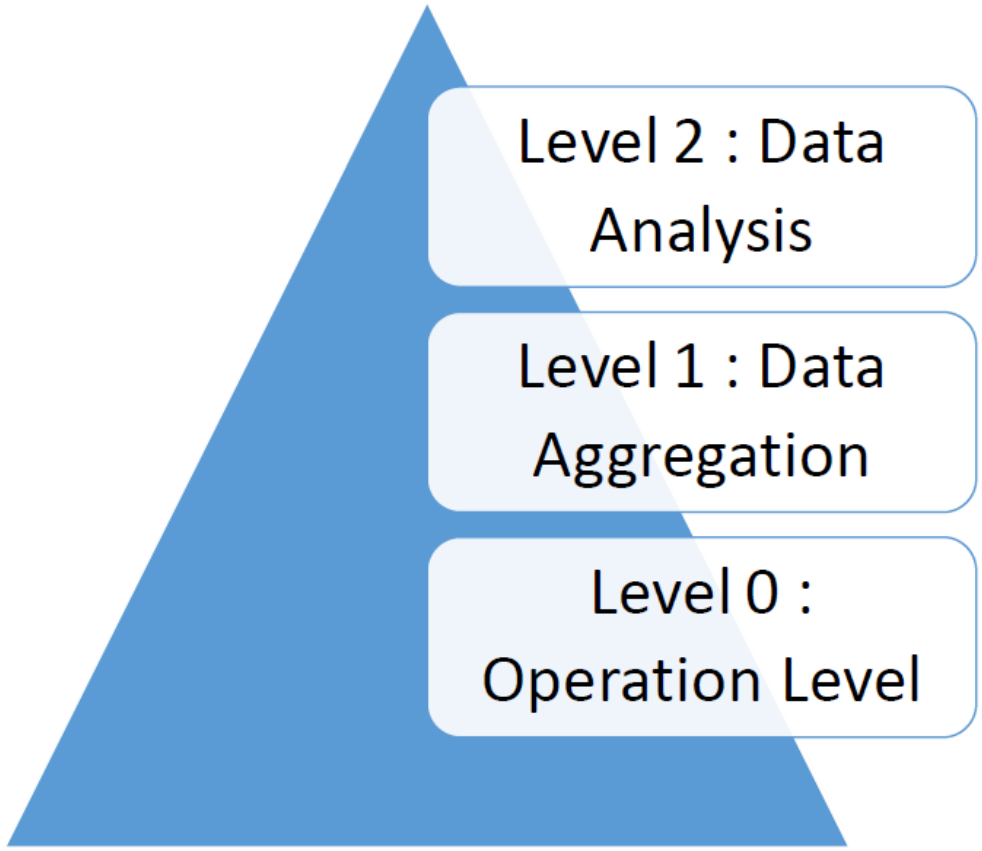


Example of Data Flow in process (back-end)



Data Science Usage and Application

Level of Information technology usage



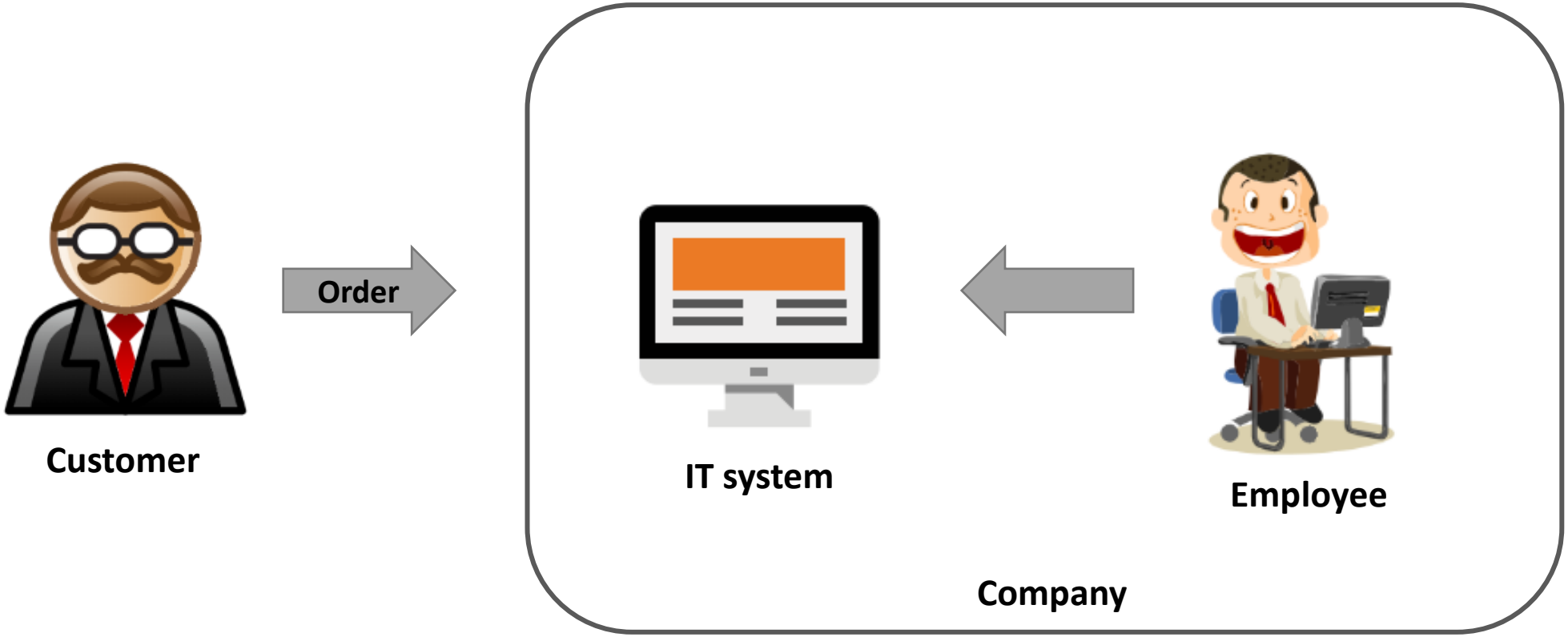
Help designing the direction

Help decision making process

Streamline inefficiency

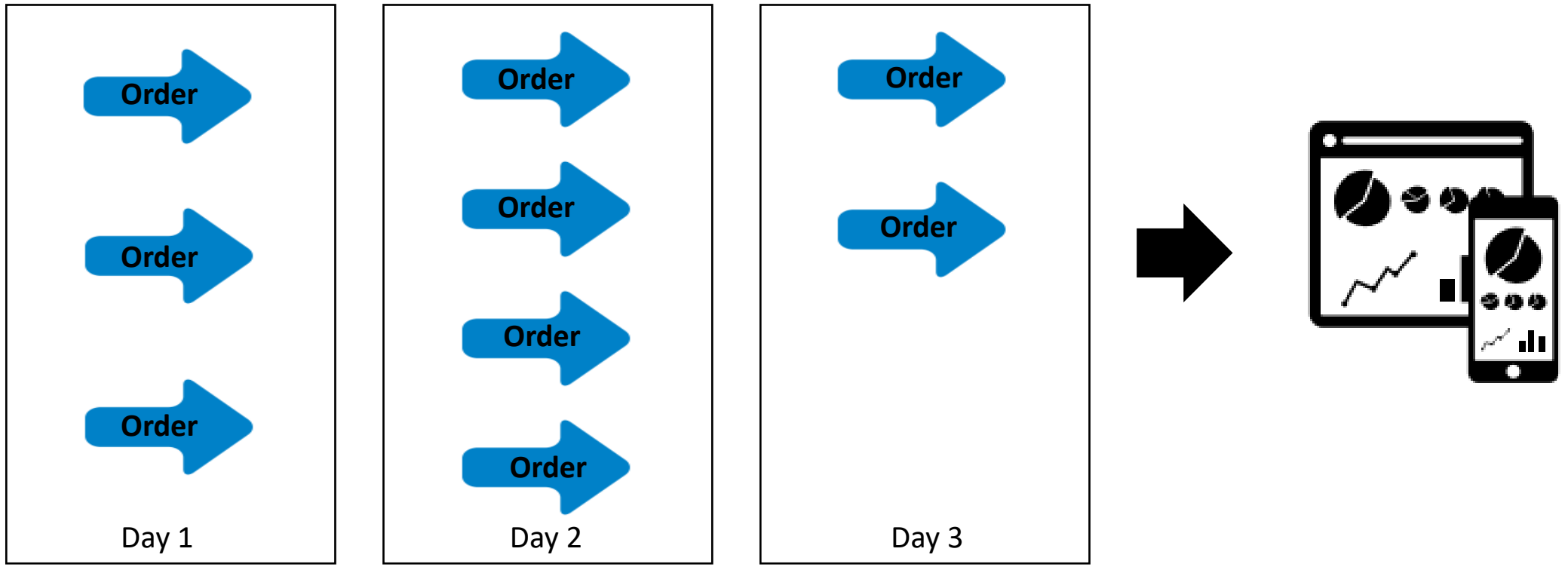
Level of Information Technology Usage (Where most SE project are)

Level 0: Operation level



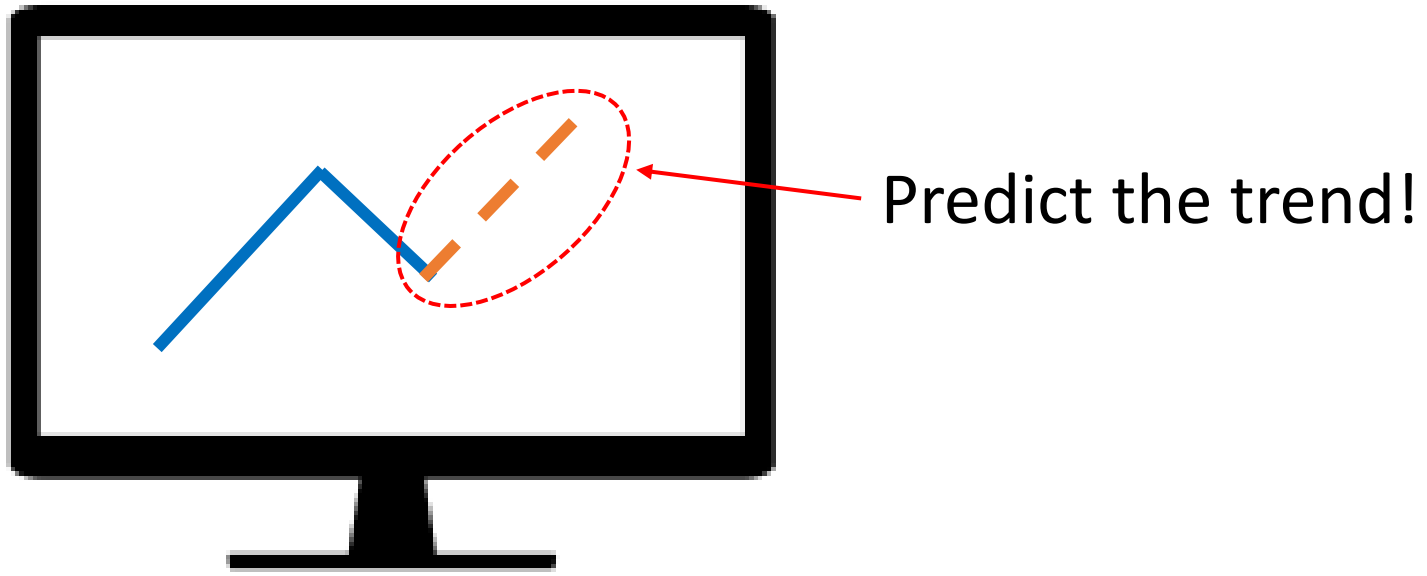
Level of Information Technology Usage

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

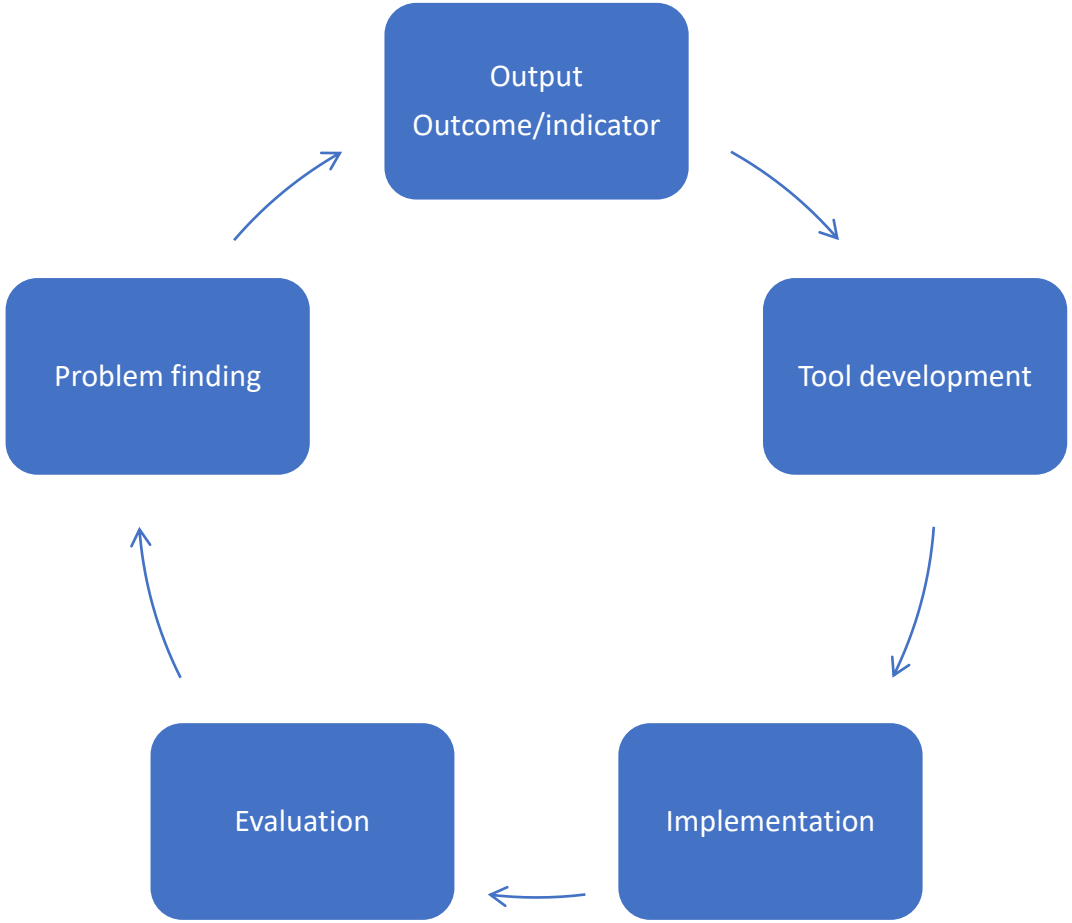
Understand the value of customers across all channels

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

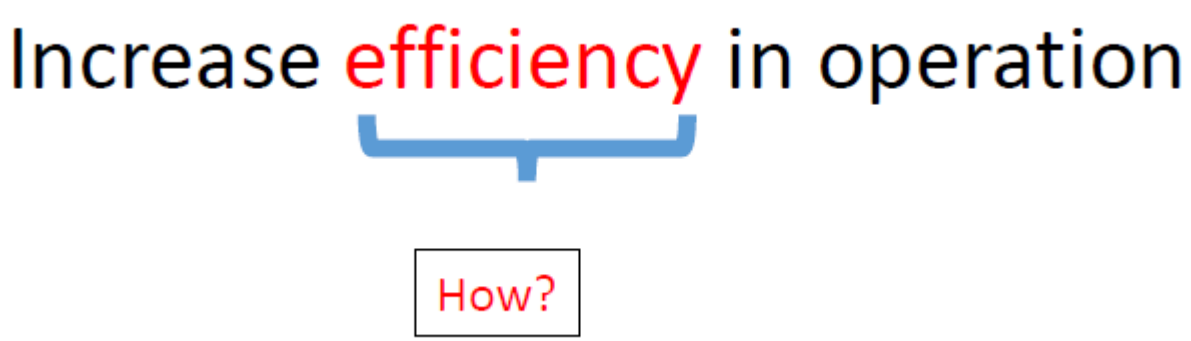
Example of outcomes:

- 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

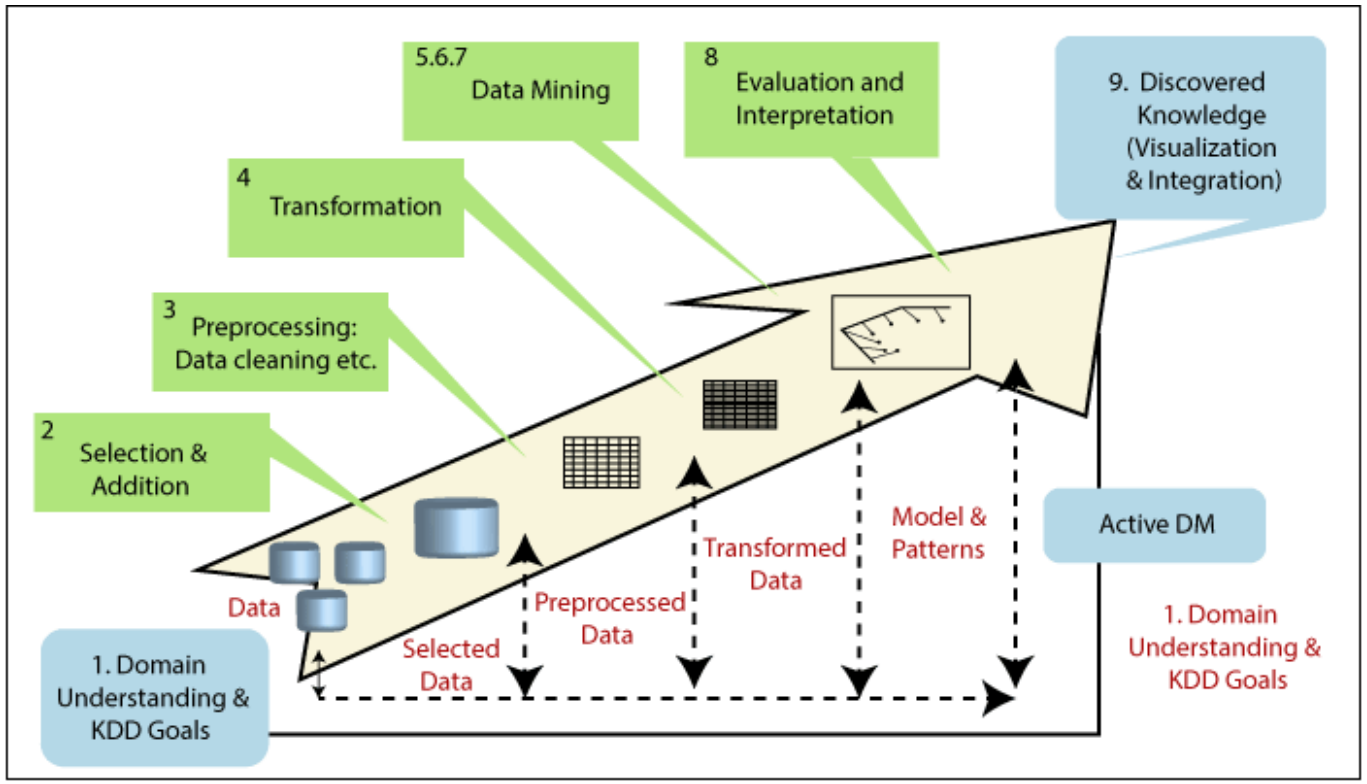
Outcome	Indicator
Increased infant breastfeeding	Number & percentage of mothers who are exclusively breastfeeding up to six months of age.
Improved work attendance by District Officials	Number of work days attended per year by District Officials
Less grade repetition	Pass rate
Beneficiaries access financial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

Solution Development

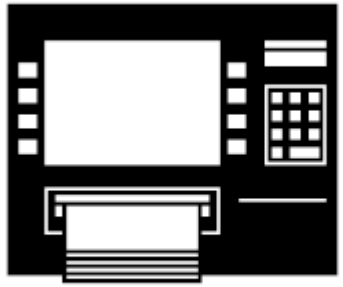
- The objective of this step is to develop a tool to solve the problem.
- The first step is to develop the **strategy**.
- The problem solving strategy is a conceptual framework to solve the problem.
- This step does not include the specification of the solution.
- The second step is to develop the **solution**.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

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



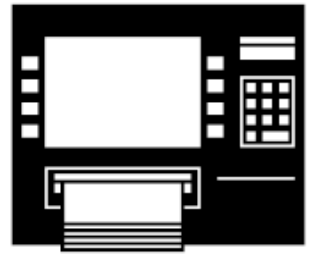
Problem identification



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

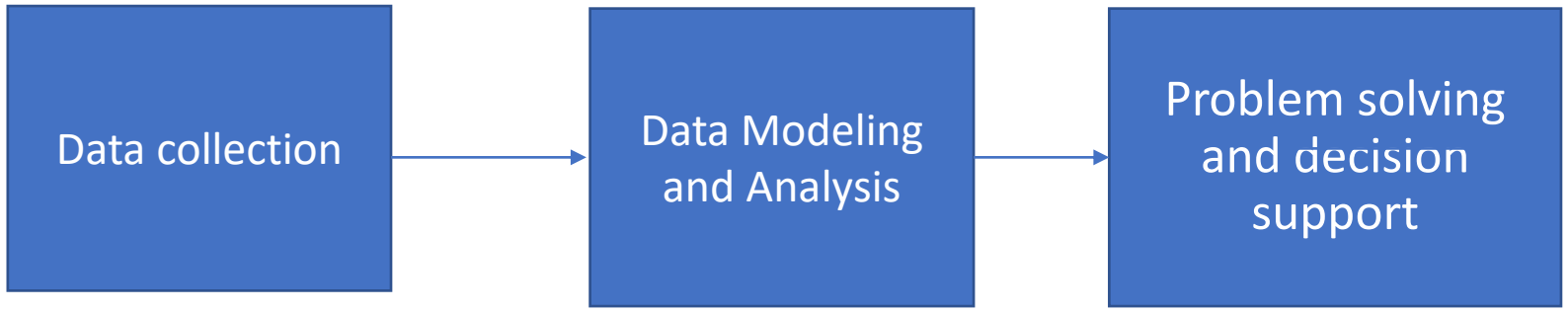
Problem

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

Supervise learning

- **human intervene (help labeling)**

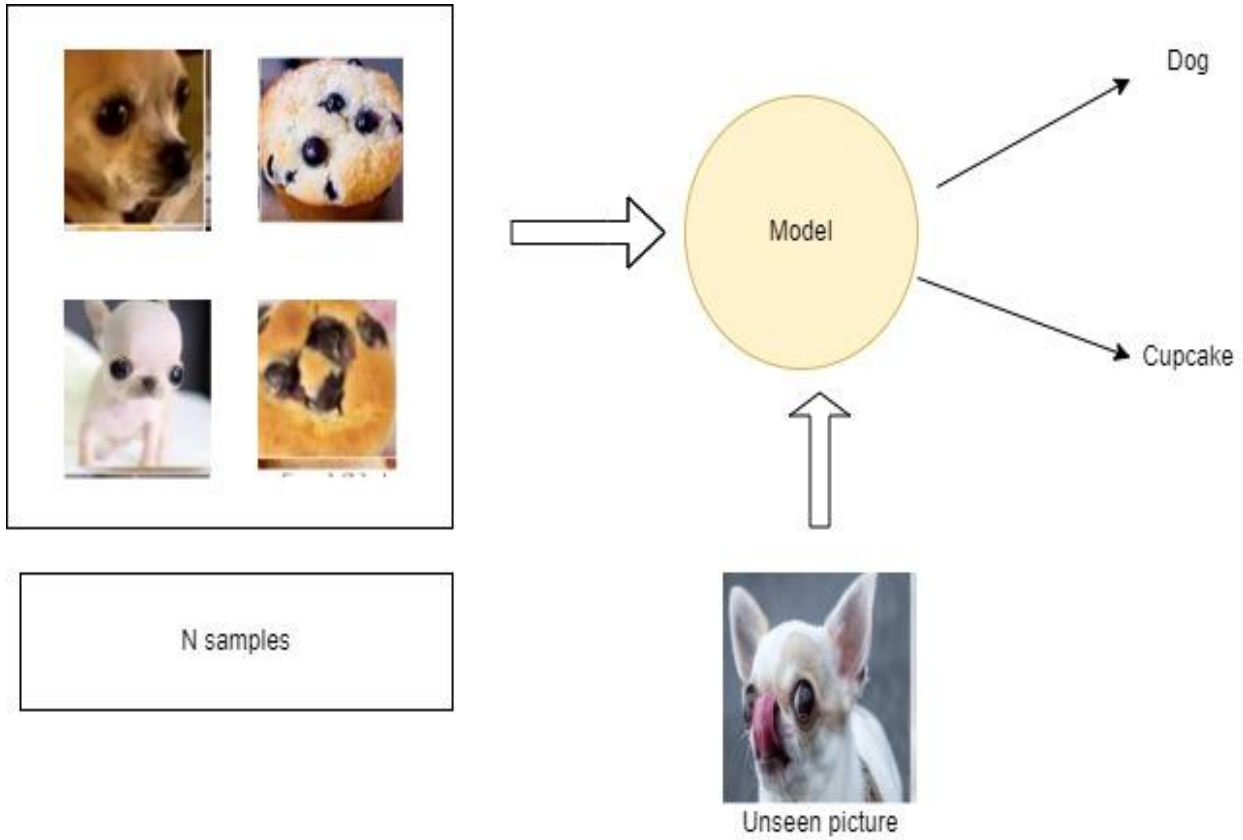
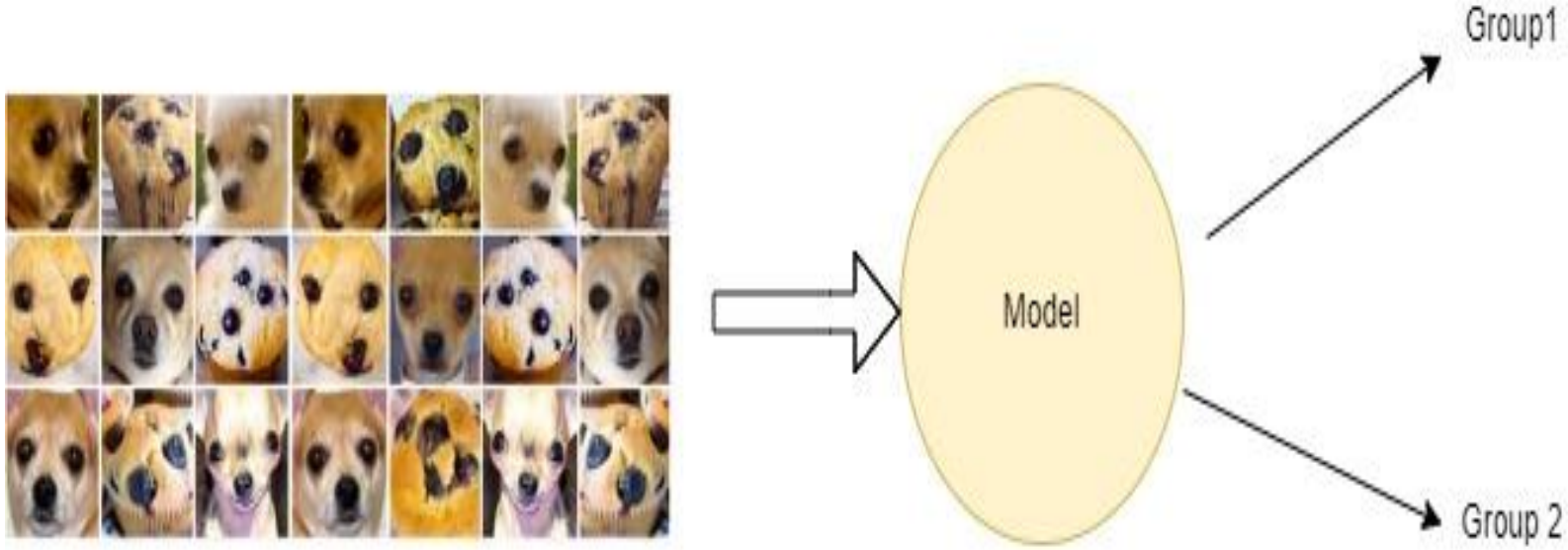


Figure 1. (a) Chihuahua and muffin, (b) Labradoodle and fried chicken

Togootogtokh, E., & Amartuvshin, A. (2018).

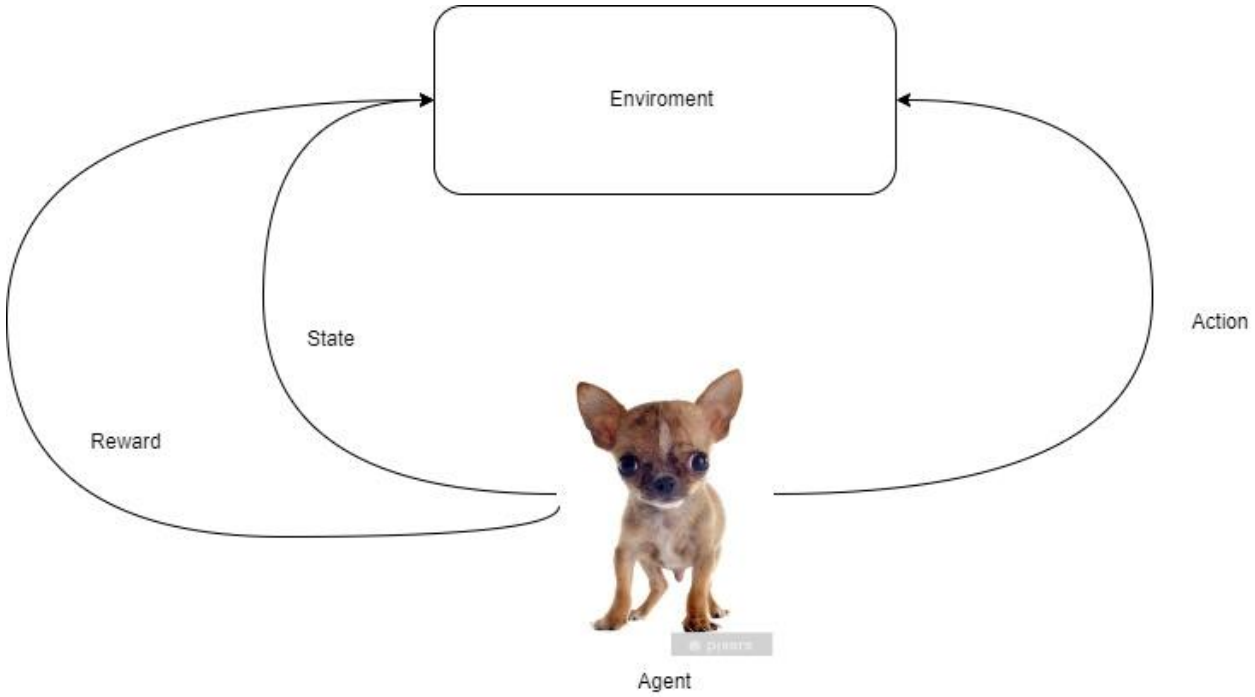
Unsupervised learning

- let the model work on its own (deal with un-labelled data)
- find similarities and differences between data points



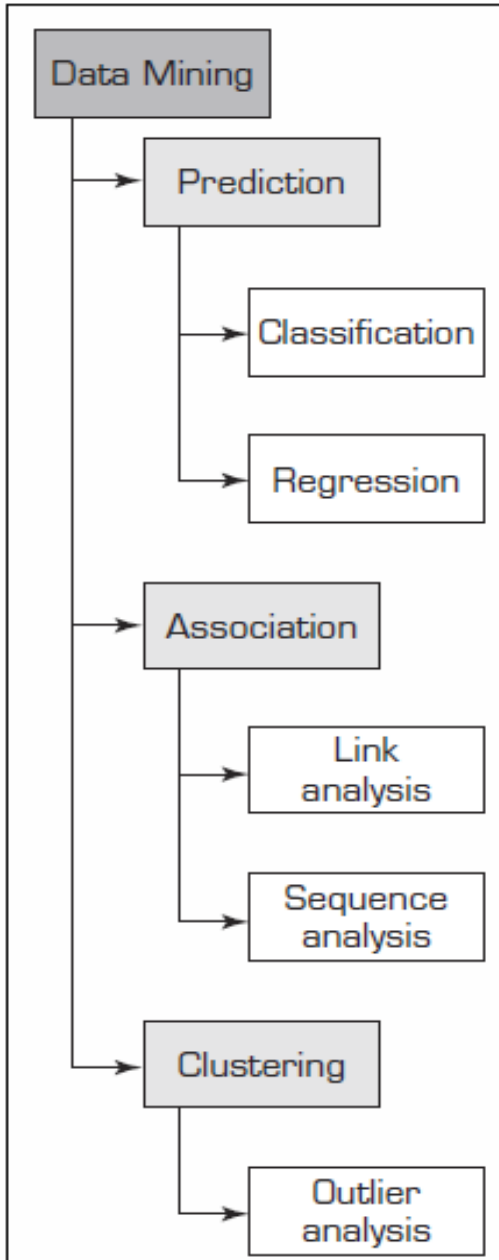
Re-enforcement learning

- Learn from mistakes
- Rewards and punishments, max(total reward of the agent)



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)



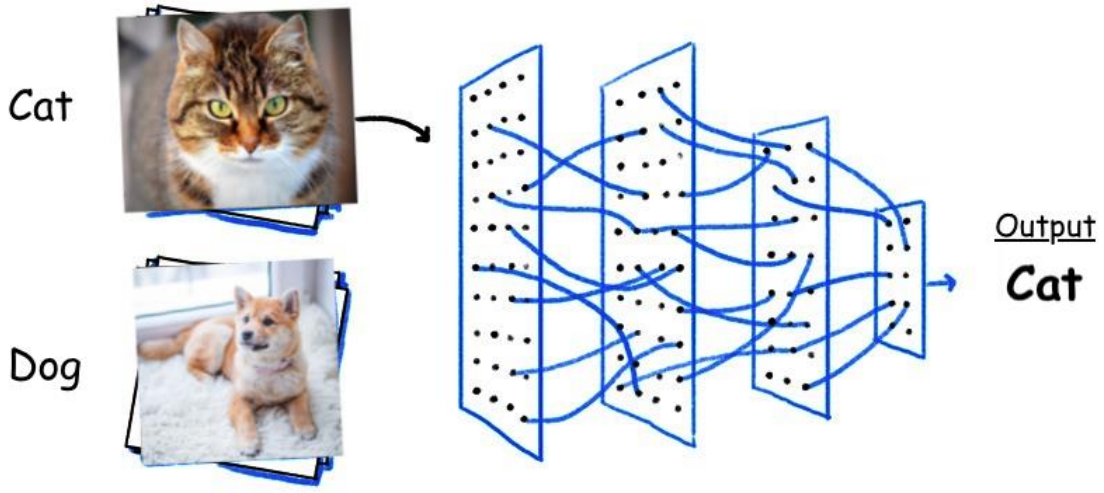
Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision Trees, ANN/MLP, SVM, Rough Sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression Trees, ANN/MLP, SVM
Unsupervised	Apriori, OneR, ZeroR, Eclat
Unsupervised	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.

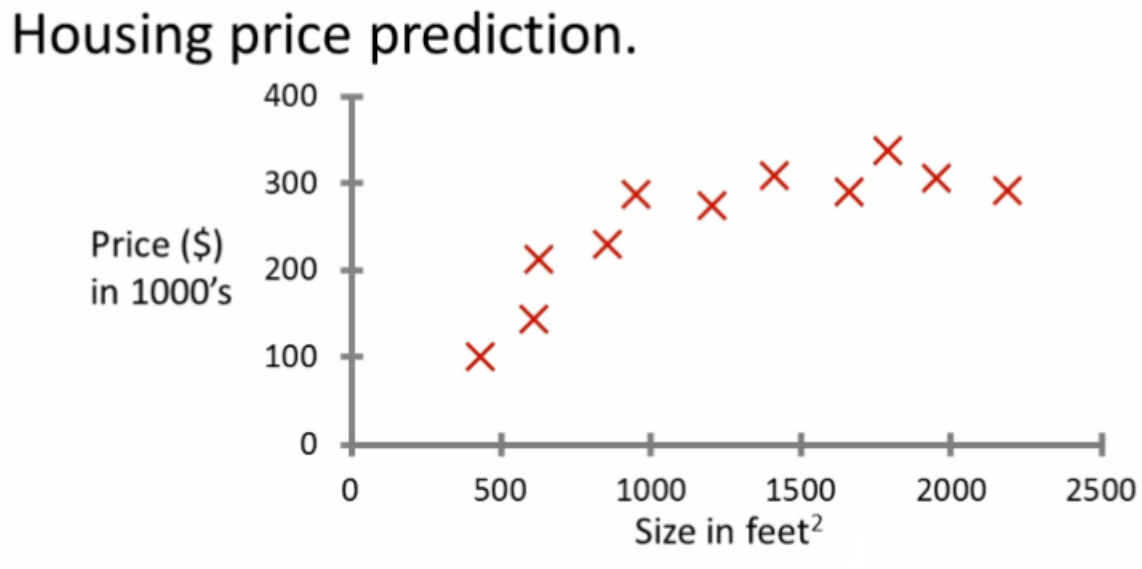
Classification

- 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.

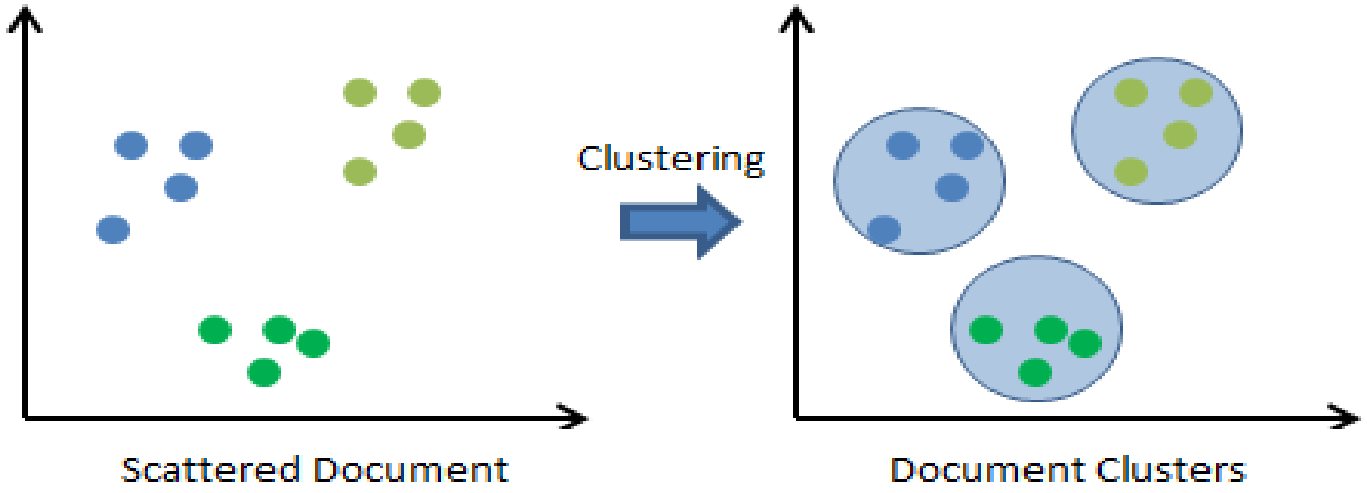


Clusters

- 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

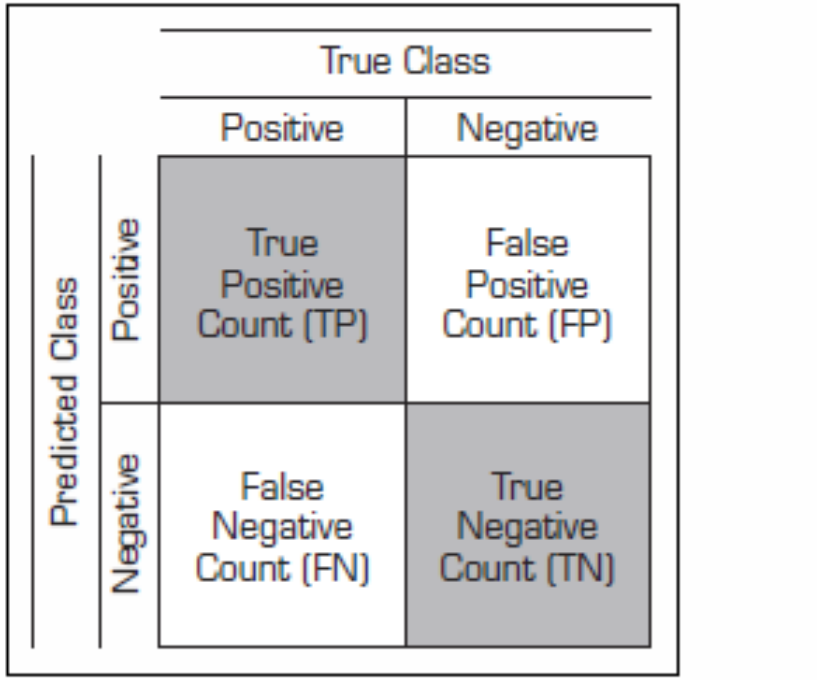
- **Predictive accuracy** – with unseen data how well the model perform in terms of %
- **Speed** – computation cost when constructing and using the model
- **Robustness** – Giving noisy data, can the model still make reasonable prediction
- **Scalability** – how’s about with larger data?
- **Interpretability**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

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

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

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

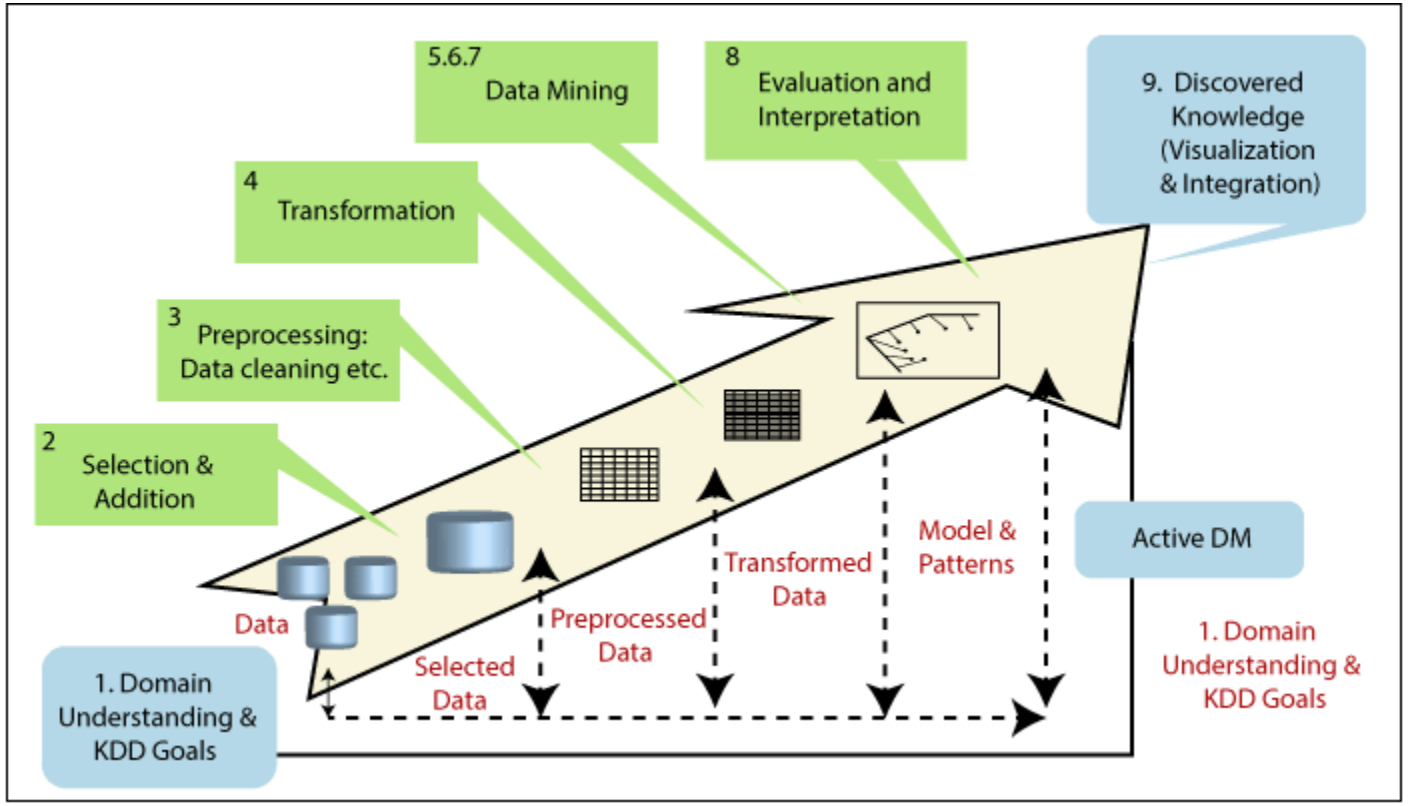
$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

Your toy dataset (SMS dataset)



Your toy dataset (SMS dataset)

In [2]:

```
df= pd.read_csv("/kaggle/input/sms-spam-collection-dataset/spam.csv",encoding='ISO-8859-1')
df
```

Out[2]:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN
...	...	...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	NaN	NaN	NaN
5568	ham	Will Ì_ b going to esplanade fr home?	NaN	NaN	NaN
5569	ham	Pity, * was in mood for that. So...any other s...	NaN	NaN	NaN
5570	ham	The guy did some bitching but I acted like i'd...	NaN	NaN	NaN
5571	ham	Rofl. Its true to its name	NaN	NaN	NaN

5572 rows × 5 columns

Your toy dataset (SMS dataset)

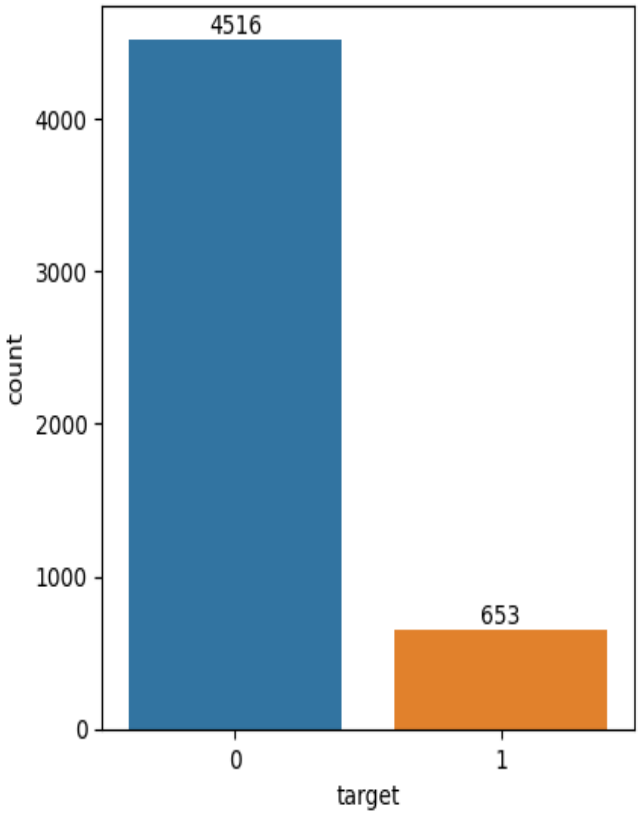
In [3]:
```
# Drop unnecessary columns from the DataFrame

columns_to_drop = ["Unnamed: 2", "Unnamed: 3", "Unnamed: 4"]
df.drop(columns=columns_to_drop, inplace=True)
```

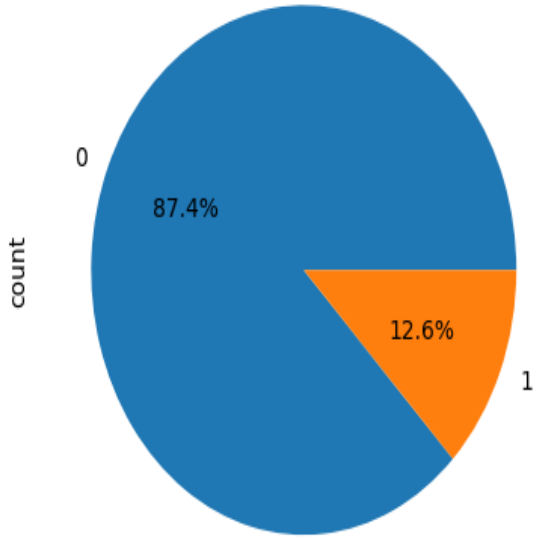
In [4]:
```
# Rename the columns
df.columns = ['label', 'message']
```

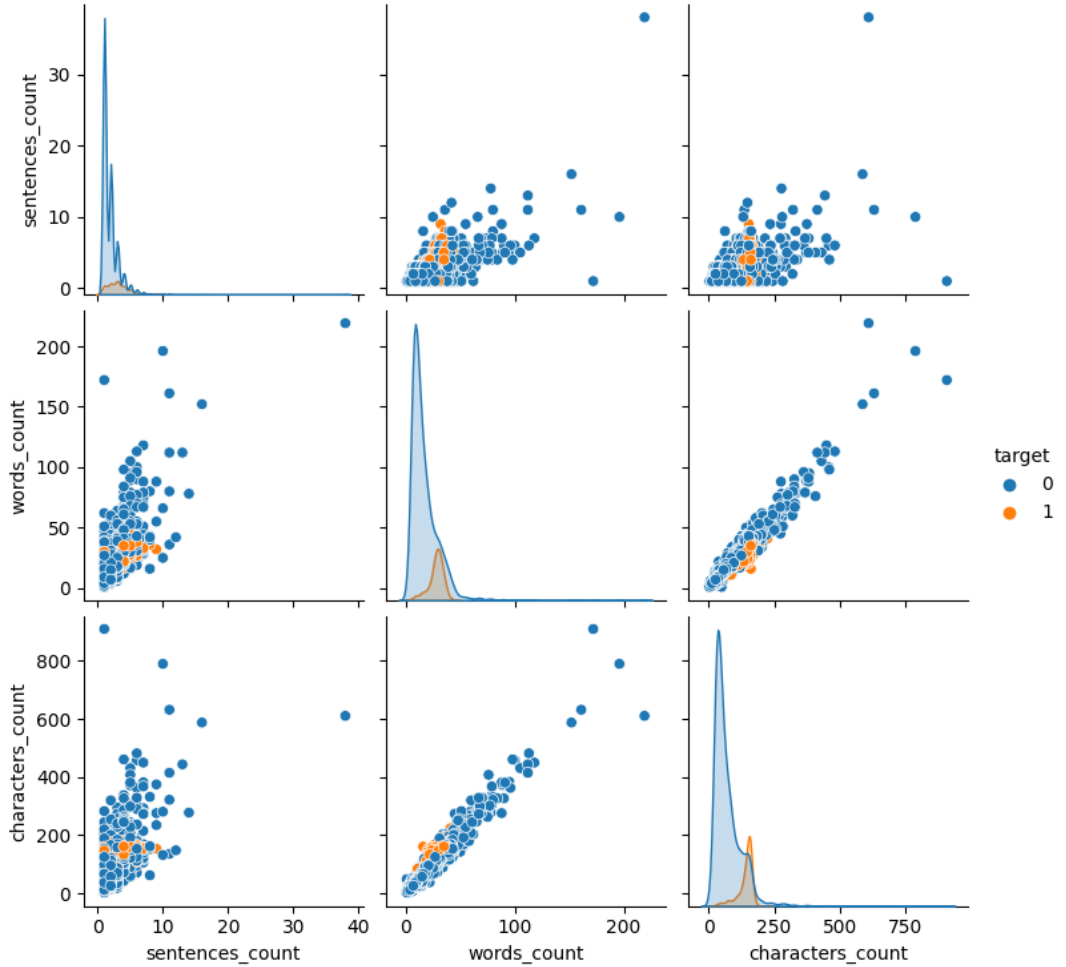
In [5]:
```
df.shape
```

Out[5]:
```
(5572, 2)
```

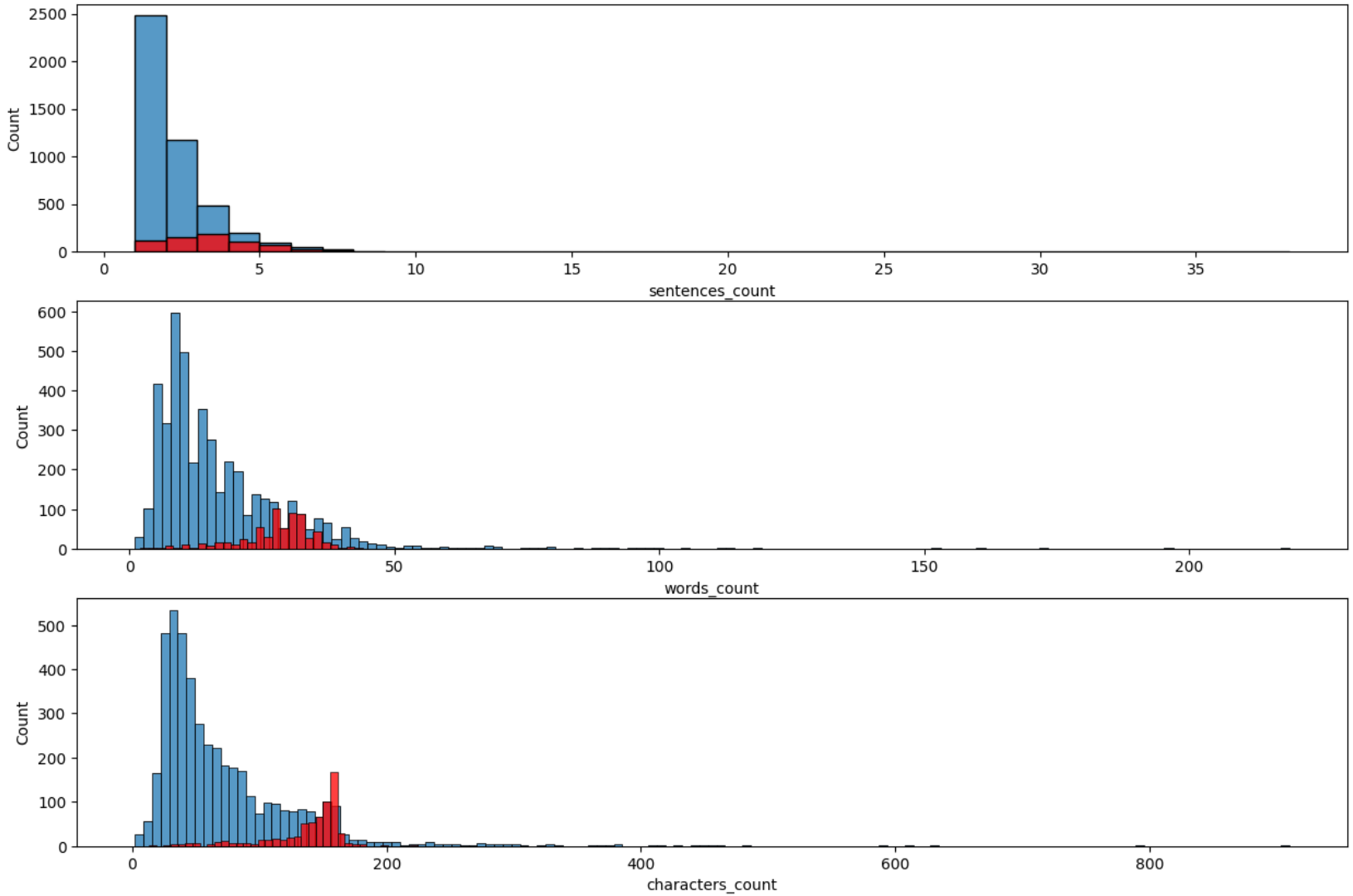
Your toy dataset (SMS dataset)EDA







Your toy dataset (SMS dataset)EDA



Domain/Data Understanding

- How many features?
- How many sample?
- What are they? How are they related? correlate?
- 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 1

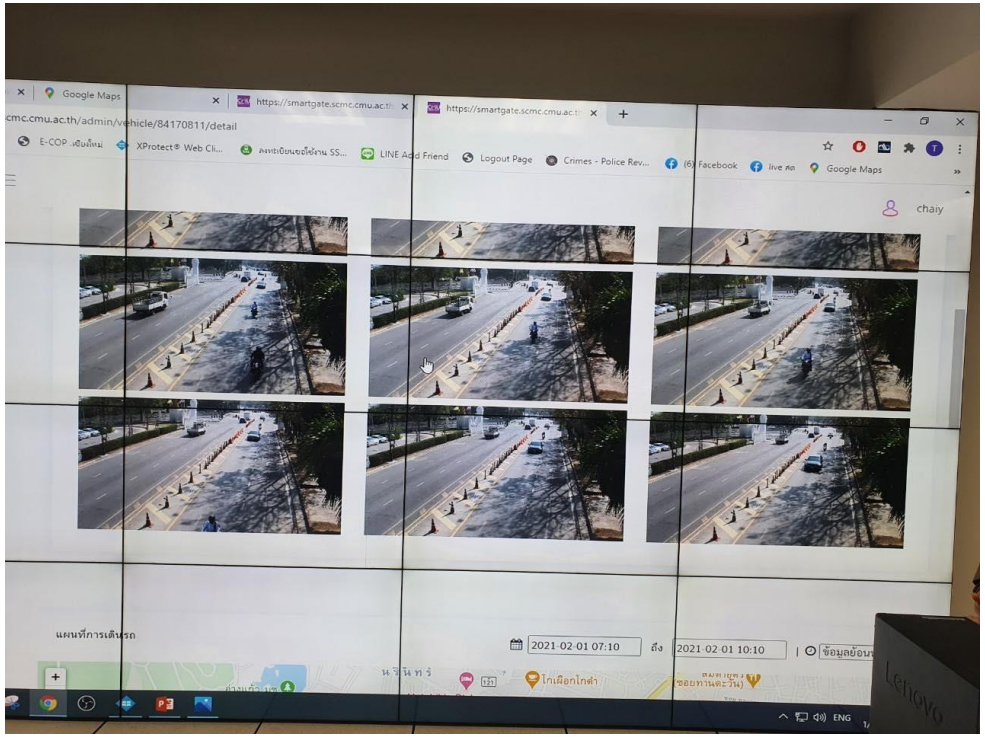
- **Business objectives, define problem (no more than one problem)**
- **Identify activities, outputs, outcomes, and indicators identification**
- **Identify Stakeholders** (who involves)
- **Identify Data source** (where can you get?, How does it look like?)
- **Identify level of IT usage (level0, 1 or 2)**
- **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 1 (continue at home)

- 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 MS team under workshop

Case study 1 : Road regulation in university campus (Beginner)

Identify who can enter or helmet detection for those who ride a motorcycle



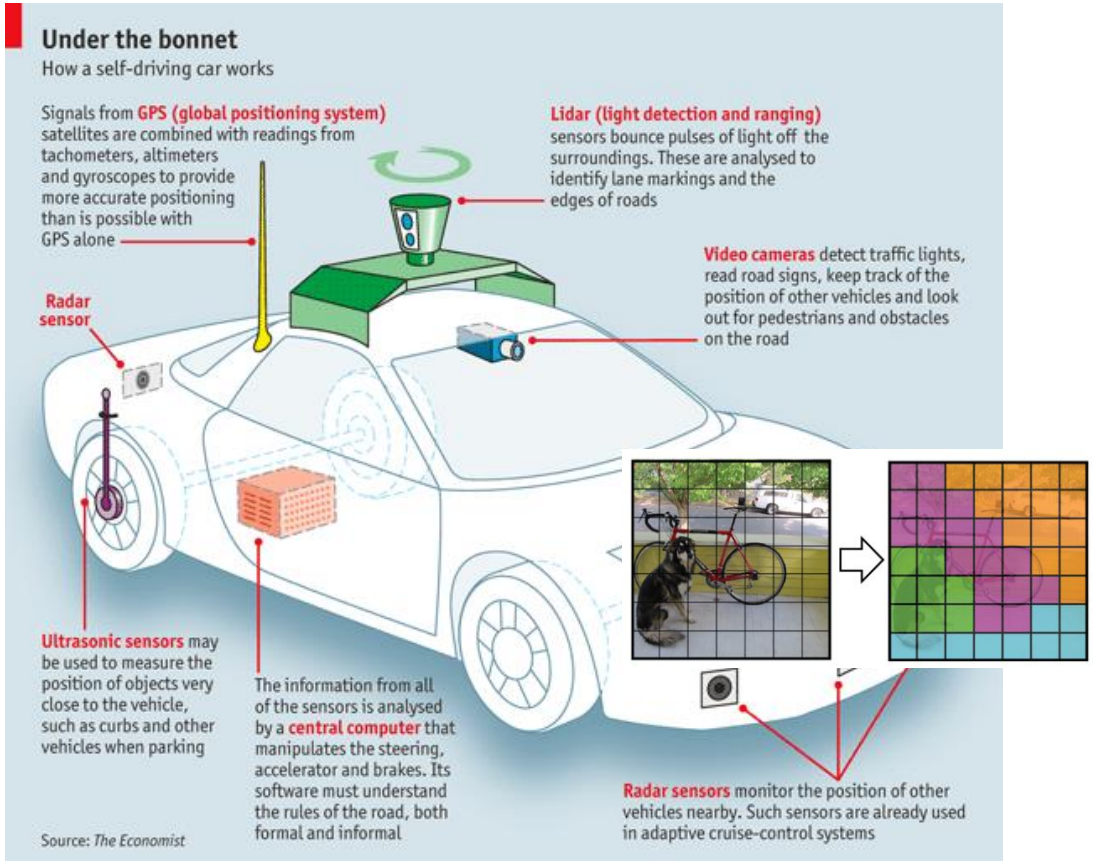
Case study 2: A university campus public transport (Intermediate, further away)

Electric bus vs Mobus

Case Study 3:

The 2020 United States presidential election (Advanced) or Any problem you think that it should belong to the group.

Sarcasm or Irony sentence detection?
How did you get here? Did someone leave your cage open?



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.
- Togootogtokh, E., & Amartuvshin, A. (2018). *Deep Learning Approach for Very Similar Objects Recognition Application on Chihuahua and Muffin Problem*. https://doi.org/10.48550/arXiv.1801.09573