



Workshop - Text Mining

with Python

Kanda Tiwatthanont @ TNI

Wed 17 and Mon 22 May 2017



Agenda - Day 1

- **Part 1 : Introduction** (10.00 - 11.00)
 - What is Data Mining ?
 - Text Mining -- Social Mind Extraction
- **Part 2 : Python** (11.00 - 12.00 / 13.00 - 14.00)
 - Python Introduction
 - **Anaconda** Installation (Data Science Distribution of Python)
 - **Jupyter** Introduction (Next Generation Engineering Notebook)
 - "Hello World!" in Jupyter, **and so on.**
- **Part3 : Pandas / Seaborn** (14.00 - 15.00)
 - **Pandas** (Structured Data Analysis Tool)
 - **Seaborn** (Statistical Data Visualization)

Agenda - Day 2

- **Part 4 : Data Mining Framework** (10.00 - 12.00)
 - Framework Overview
 - Scikit-learn -- Machine Learning Tool for Data Scientist
 - Data Prediction Hands-on
- **Part 5 : Sentiment Analysis** (13.00 - 15.00)
 - Introduction Text Mining
 - Unstructured to Structured Data
 - Text Classification



Workshop

Part 1 - Introduction

Data Mining & Text Mining



Kanda Tiwatthanont @ TNI

Data mining

- Mining patterns from data



Market Basket Analysis



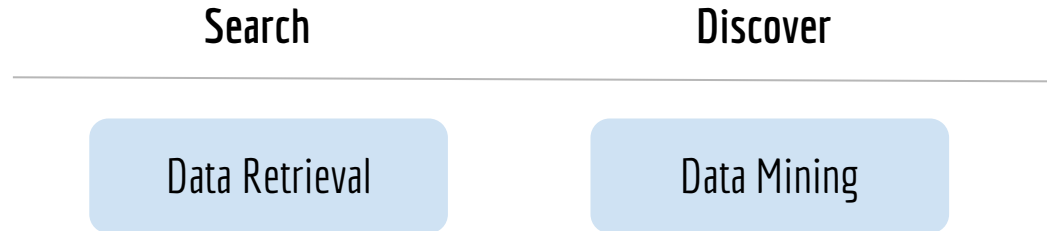
Fraud Detection

Data mining

- Mining patterns from data
- Is it database ?
- Is it statistics ?
- Is it machine learning ?

Data mining

- Mining patterns from data
- Is it **database** ?



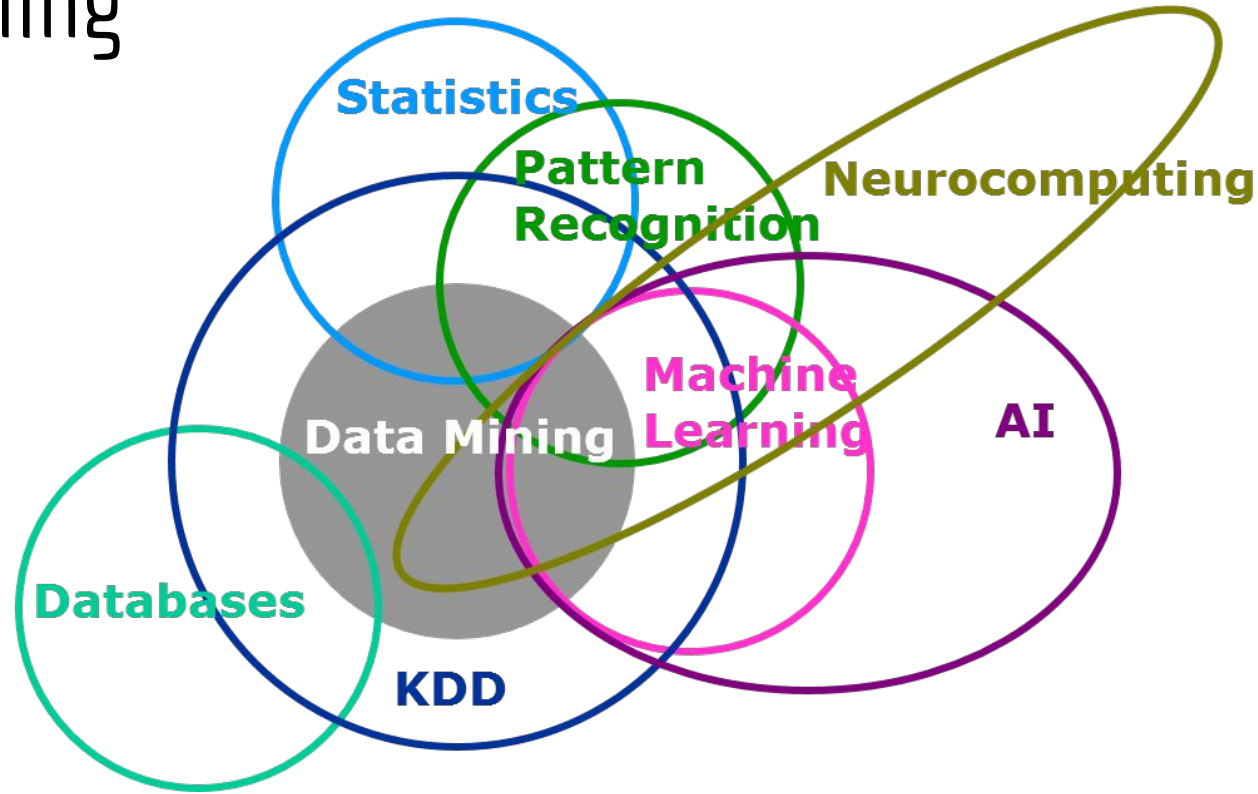
Data mining

- Mining patterns from data
- Is it **statistics** ?
 - Non Functional form
 - Speed are important
 - Data size

Data mining

- Mining patterns from data
- Is it machine learning ?
 - ML concerns speed and spaces (Algorithms)
 - Data Mining concerns data (Business side)

Data mining



Ref : SAS Institute - A Venn diagram shows how machine learning and statistic are related data mining

Data mining applications for business

Retail



- Customer shopping behaviour
- Customer segmentation
- Customer retention

Banking



- Credit score
- Customer segmentation

Insurance



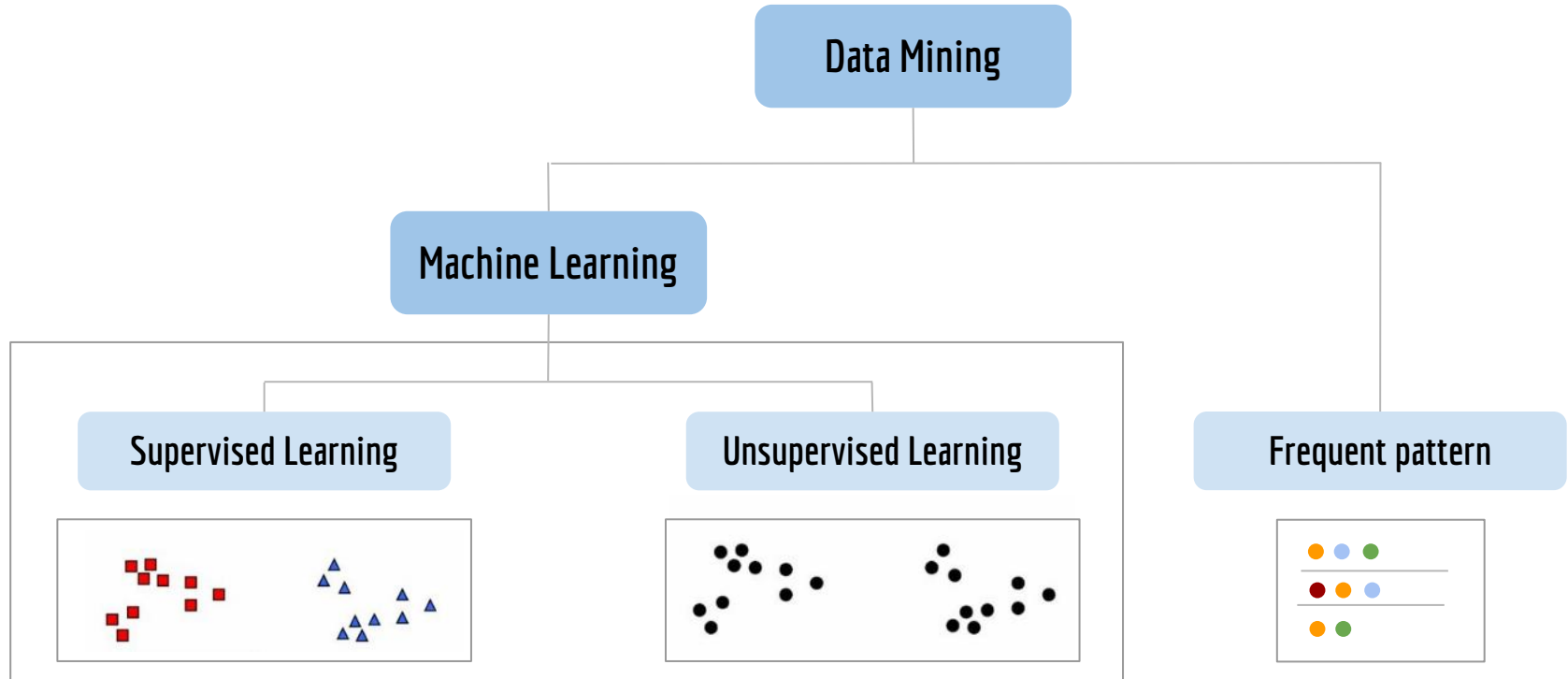
- Risk factor identification
- Fraud detection

Social

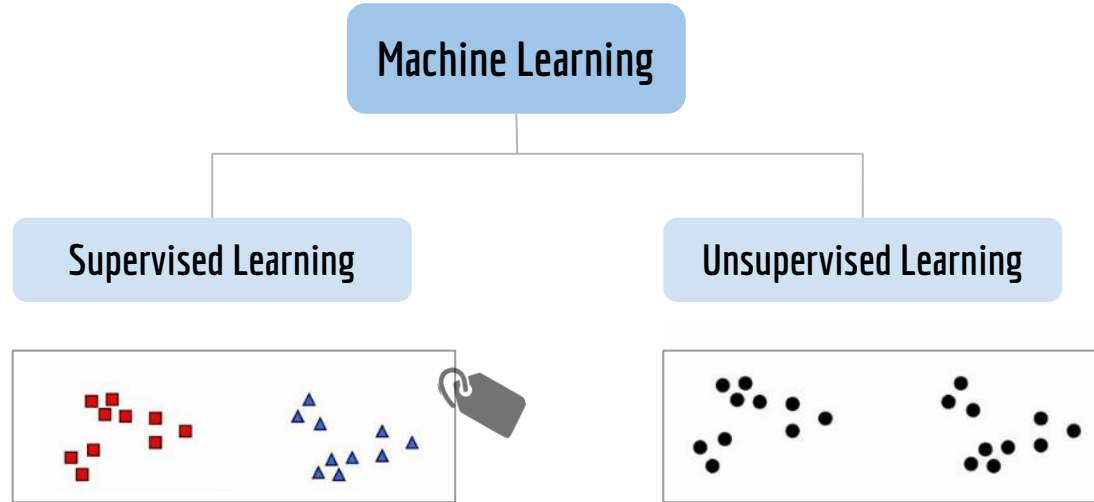


- Keyword Suggestions
- Face recognition
- Recommendation

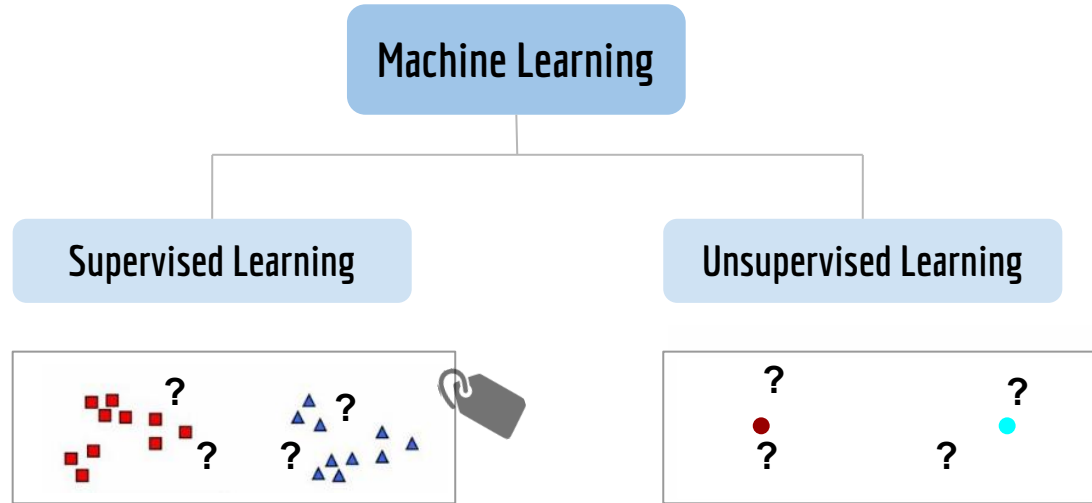
Data Mining Tasks



Machine Learning to Data Mining



Machine Learning to Data Mining



Predictive

Making predictions using data.
There is an outcome we are trying to predict.

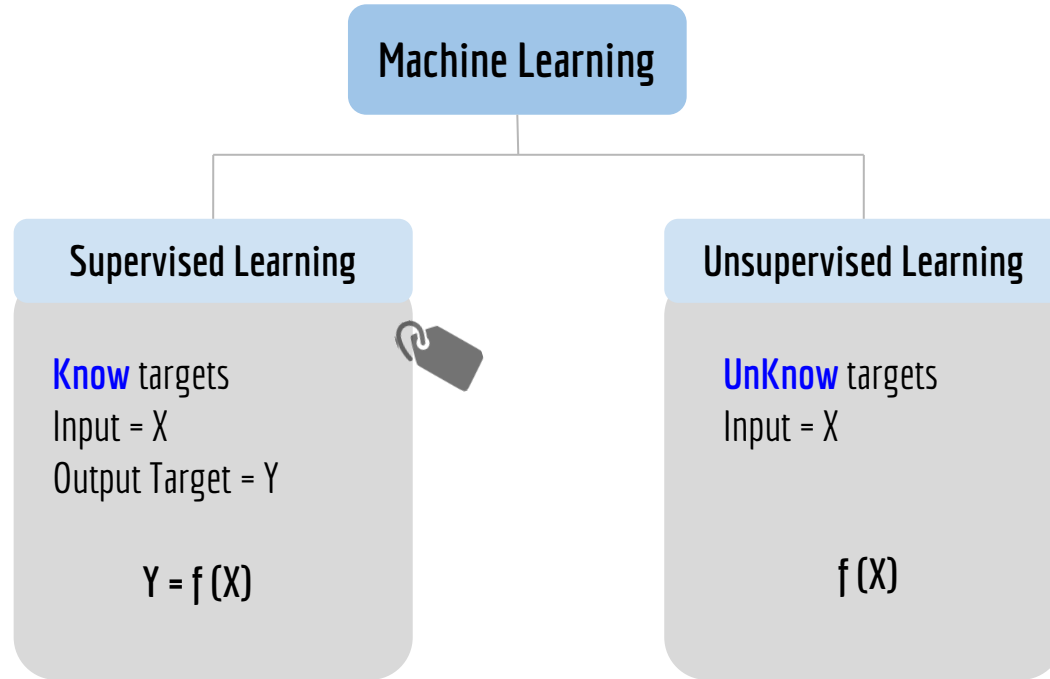
Example: Spam mail filtering

Descriptive

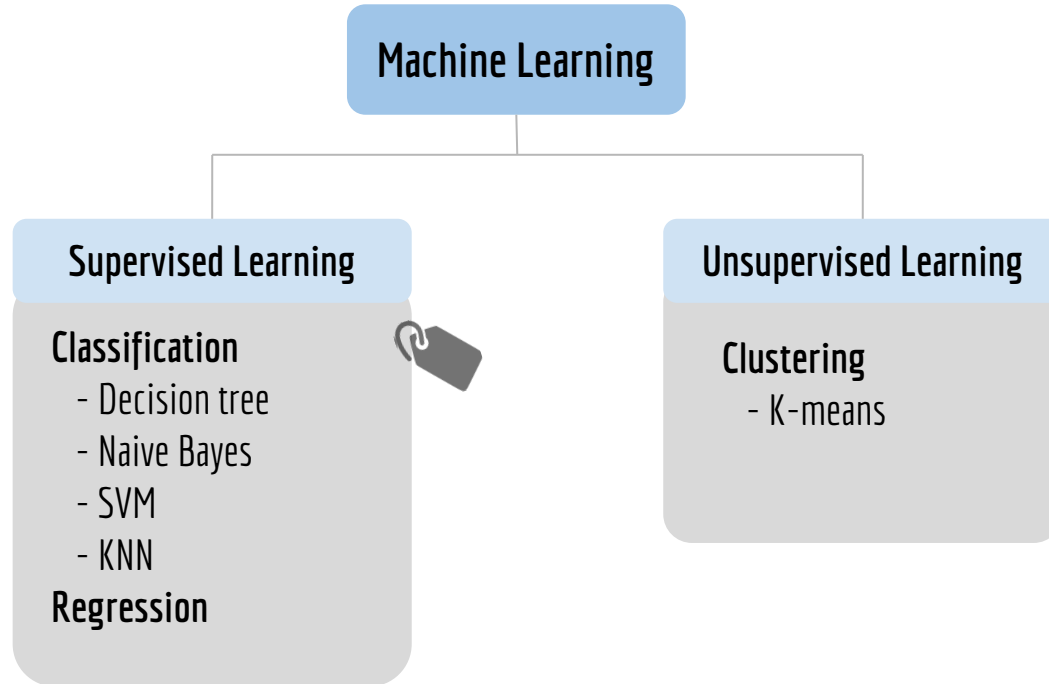
Extracting structure from data.
There is no right answer.

Example: Customer behaviors segmentation

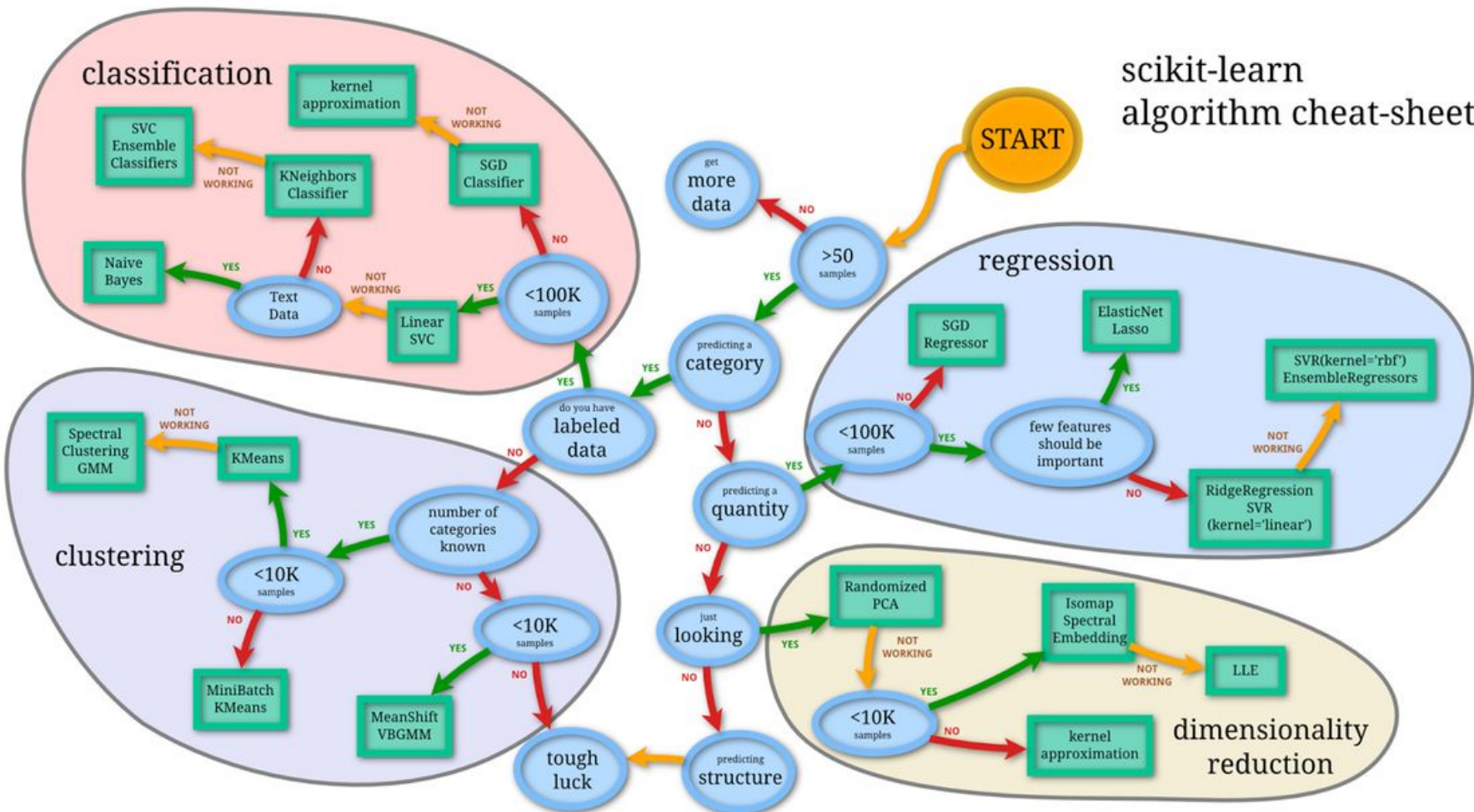
Machine Learning to Data Mining



Machine Learning to Data Mining



scikit-learn algorithm cheat-sheet





Part 1

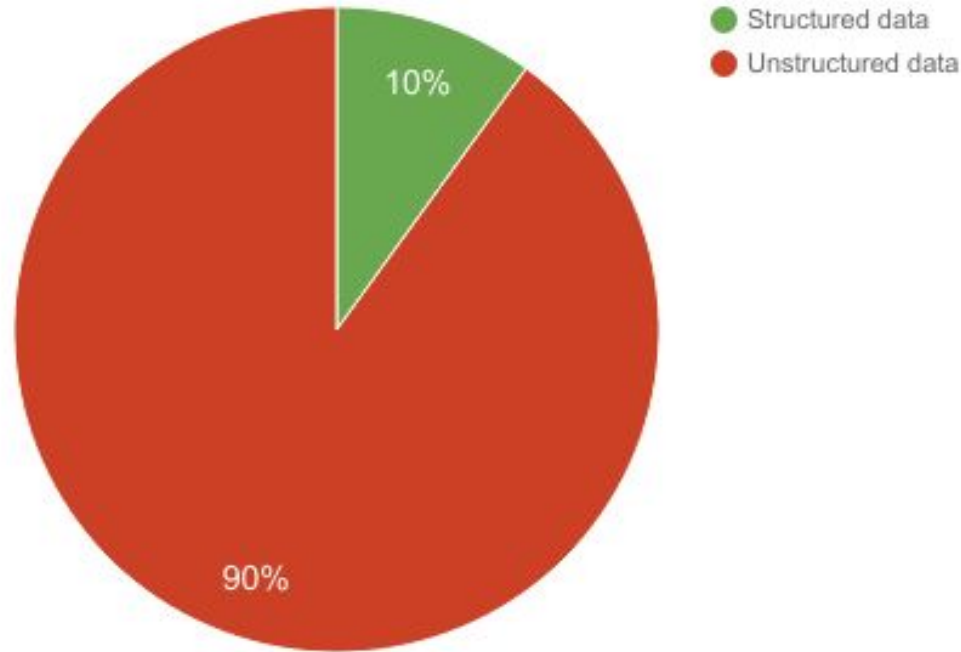
Introduction
Text Mining



Text mining

- Mining patterns from unstructured data

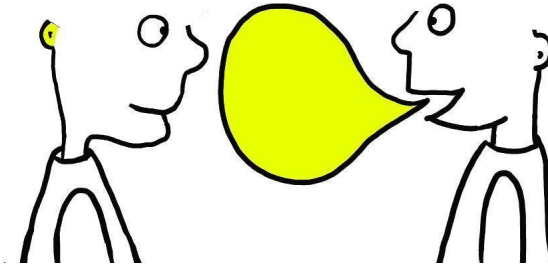
Structured vs. Unstructured data



Structured data

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
1	0	3	male	22	1	0	7.25
2	1	1	female	38	1	0	71.2833
4	1	1	female	35	1	0	53.1
5	0	3	male	35	0	0	8.05
7	0	1	male	54	0	0	51.8625
8	0	3	male	2	3	1	21.075
10	1	2	female	14	1	0	30.0708
11	1	3	female	4	1	1	16.7
13	0	3	male	20	0	0	8.05
15	0	3	female	14	0	0	7.8542
16	1	2	female	55	0	0	16
17	0	3	male	2	4	1	29.125
18	1	2	male	NA	0	0	13
19	0	3	female	31	1	0	18
20	1	3	female	NA	0	0	7.225
21	0	2	male	35	0	0	26
22	1	2	male	34	0	0	13



Unstructured data



Natural Language



Data Mining vs. Text Mining

		Search	Discover
Structured data	 <p>What you find in a DB (typically)</p>	Data Retrieval	Data Mining
Unstructured data	 <p>What you find in the 'wild' (text, images, audio, video)</p>	Information Retrieval	Text Mining

Text Mining Application

- Text summarization (การสรุปใจความสำคัญ)
- Machine translation (MT) (การแปลภาษา)
- Question answering (QA) (การถามตอบ)
- Opinion mining (การวิเคราะห์ความคิดเห็น)
- Robotic IVR



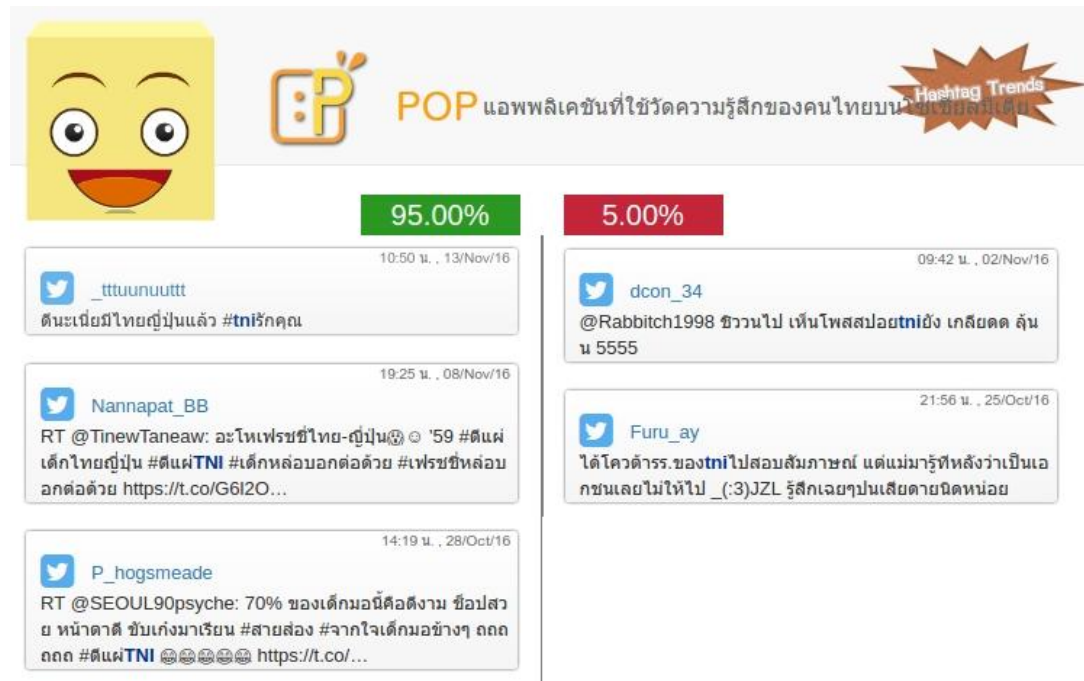
<http://textcompactor.com/>



Text Mining Application

Text mining

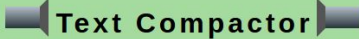
Keyword “TNI”



Text Mining Application

Text summarization

<http://textcompactor.com/>

Text Compactor

Free Online Automatic Text Summarization Tool

[Home](#)
[About](#)

Follow these simple steps to create a summary of your text.

Step 1
Type or paste your text into the box.

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. Automatic data summarization is part of machine learning and data mining. The main idea of summarization is to find a representative subset of the data, which contains the information of the entire set. Summarization technologies are used in a large number of sectors in industry today. An example of the use of summarization technology is search engines such as Google. Other examples include document summarization, image collection summarization and video summarization. Document summarization, tries to automatically create a representative summary or abstract of the entire document, by finding the most informative sentences. Similarly, in image summarization the system finds the most representative and important (or salient) images. Similarly, in consumer videos one would want to remove the boring or repetitive scenes, and extract out a much shorter and concise version of the video. This is also important, say for surveillance videos, where one might want to extract only important events in the recorded video, since most part of the video may be uninteresting with nothing going on. As the problem of information overload grows, and as the amount of data increases, the interest in automatic summarization is also increasing.

Generally, there are two approaches to automatic summarization: extraction and abstraction. Extractive methods work by selecting

Step 2
Drag the slider, or enter a number in the box, to set the percentage of text to keep in the summary.

9 %

Step 3
Read your summarized text. If you would like a different summary, repeat Step 2. When you are happy with the summary, copy and paste the text into a word processor, or [text to speech program](#), or [language translation tool](#).

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Other examples include document summarization, image collection summarization and video summarization.

© 2010-2016 [Knowledge by Design, Inc.](#)

Text Mining Application

- ★ Brand Monitoring
- ★ Feedback
- ★ Competitors Comparing



Data Analytics Specialization

Which tool to Choose?



Mining Tools

Business



written in java, drag and drop,
model form Weka, R.
Free 10K records



written in Python, drag and drop,
simple and easy to learn

Academic



Very sophisticated
Free, but hard to use



Statistical and math computing



Written in Python,
Make the learning curve easier
Fast and reliable library

Basic



- Easy to use operators
- Very popular among data scientist



- Interactively with functions, graphs, and limited statistics.
- apply in a company, presenting data

Advance



- Easy to learn
- Enterprises using SAS



- A huge number of statistical, graphical, and analytical packages
- R is not enough, many BigData use Python



- Many key advantages over R (production environments, parallel processing)
- Lacks many statistical packages (but provide pandas package)



Software you used for Data Science

What software you used for Analytics, Data Mining, Data Science, Machine Learning projects in the past 12 months ?

Data Science Platforms/Suites

- RapidMiner (554) 8%
- Excel (345) 5%
- Anaconda (258) 4%
- scikit-learn (219) 3%
- Orange (53) 1%

Deep Learning

- TensorFlow (242) 3%
- Theano (64) 1%

Languages

- R language (603) 9%
- Python (577) 8%
- SQL language (413) 6%

Ref : Vote by people who are attended rapidminer class with organized by RapidMiner (10 may 2017)

Programming Languages (Data Analytics)

Analysis / Data mining :

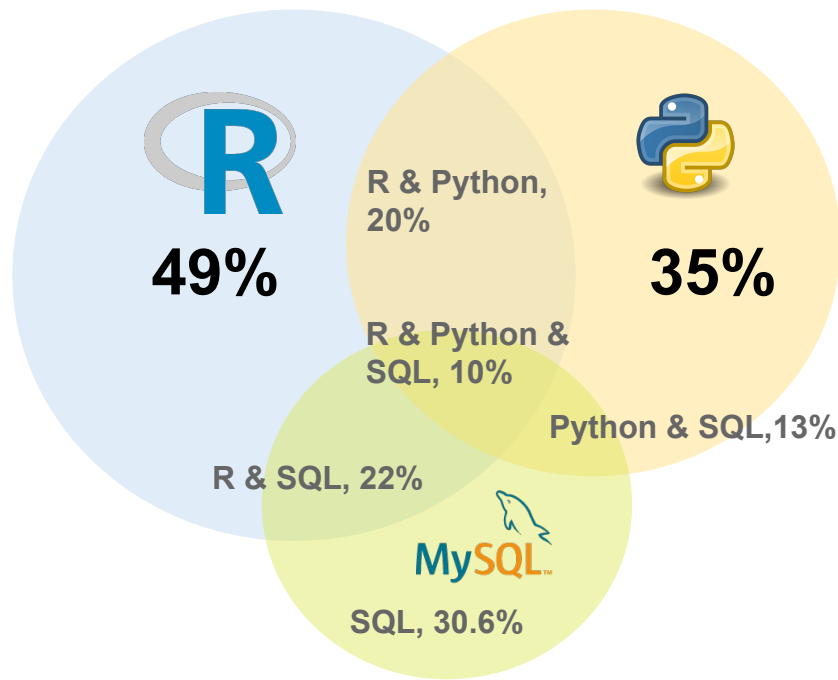
- R Language
- **Python**
- SQL


Big Data (Hadoop)

- Java
- **Python**

Visualization

- JavaScript





Next

Part 2 - Python

Kanda Tiwatthanont @ TNI

