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 ?

Ref : http://www.dbjournal.ro/archive/14/14_3.pdf

Data mining

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

Search	Discover
Data Retrieval	Data Mining

Ref : http://www.dbjournal.ro/archive/14/14_3.pdf

Data mining

- **Mining patterns from data**

- Is it statistics ?
  - Non Functional form
  - Speed are important
  - Data size

Ref : http://www.dbjournal.ro/archive/14/14_3.pdf

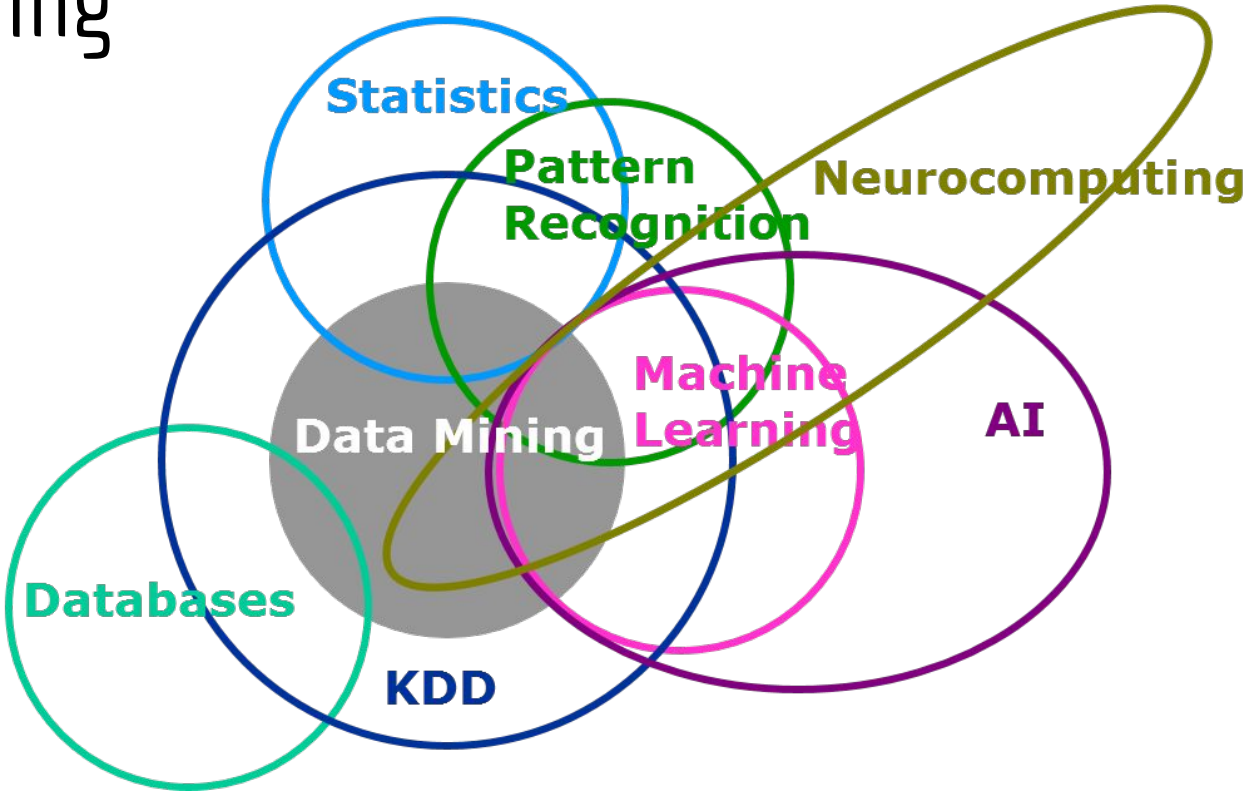
Data mining

- **Mining patterns from data**

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

Ref : http://www.dbjournal.ro/archive/14/14_3.pdf

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

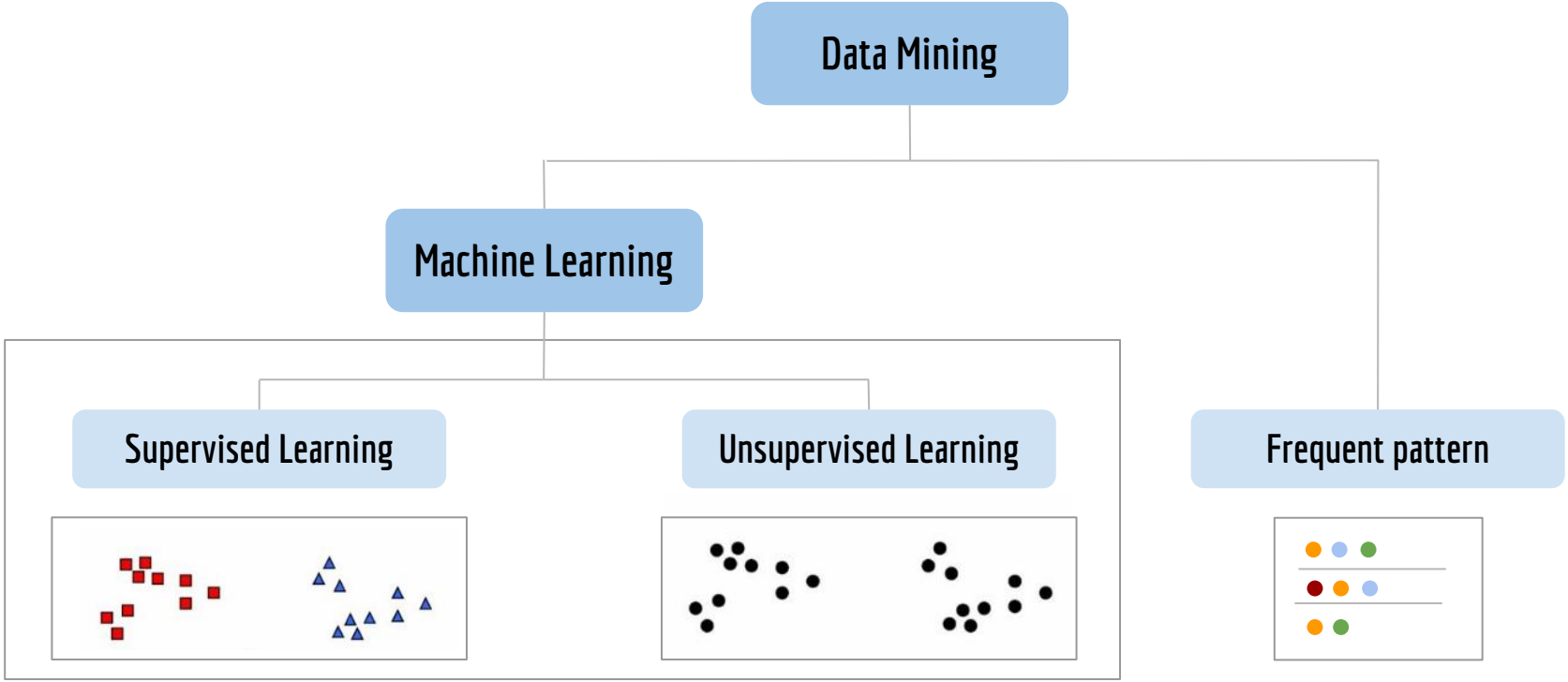
- Risk factor identification
- Fraud detection

Social

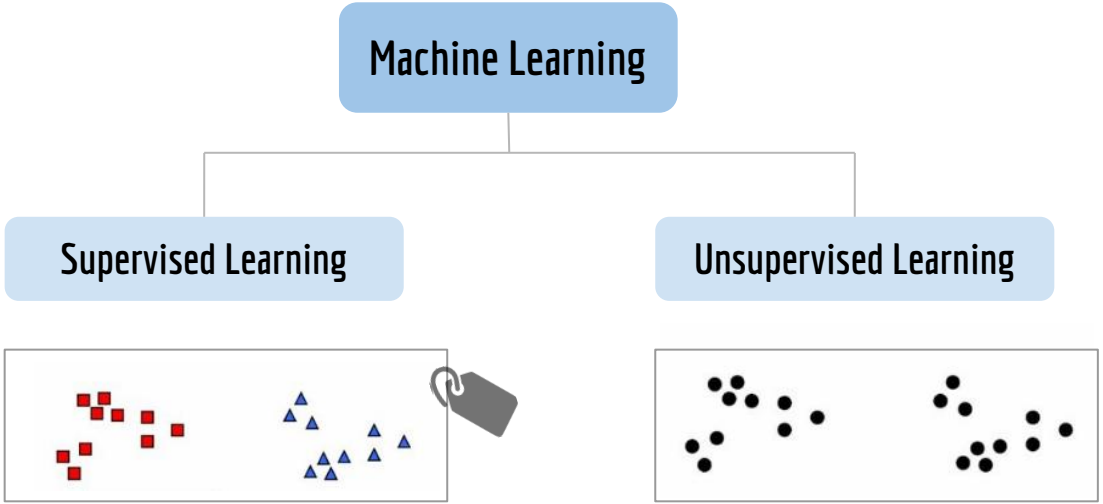
- Keyword Suggestions
- Face recognition
- Recommendation

Ref : http://www.dbjournal.ro/archive/14/14_3.pdf

Data Mining Tasks



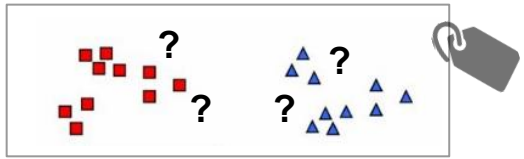
Machine Learning to Data Mining



Machine Learning to Data Mining

Machine Learning

Supervised Learning



Predictive

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

Example: Spam mail filtering

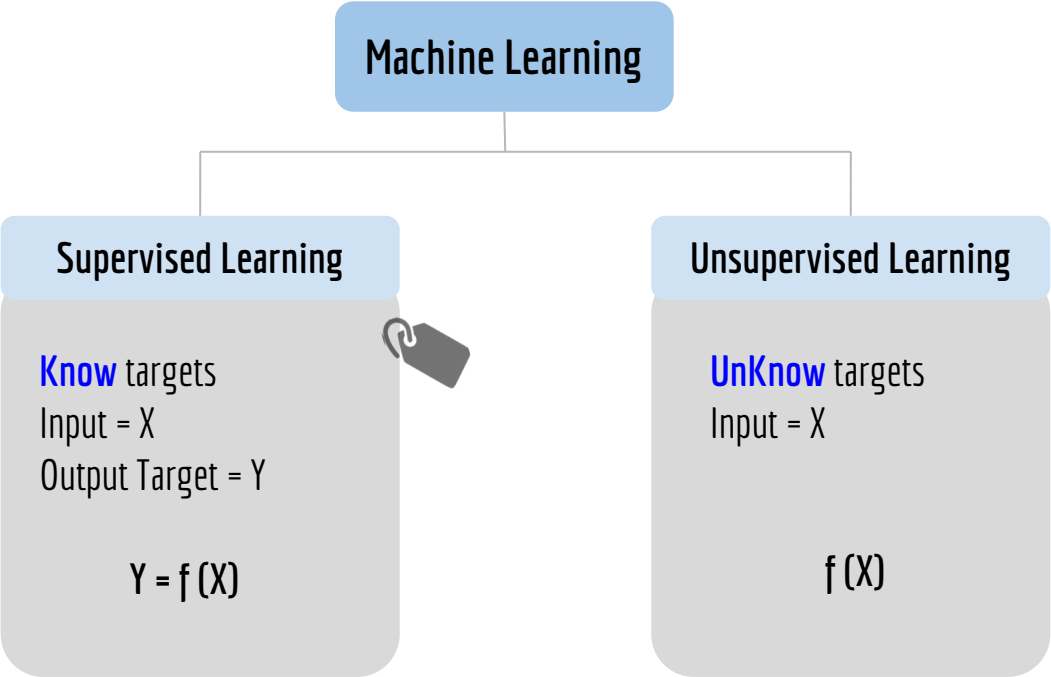
Unsupervised Learning

Descriptive

Extracting structure from data.
There is no right answer.

Example: Customer behaviors segmentation

Machine Learning to Data Mining



Machine Learning to Data Mining

Machine Learning

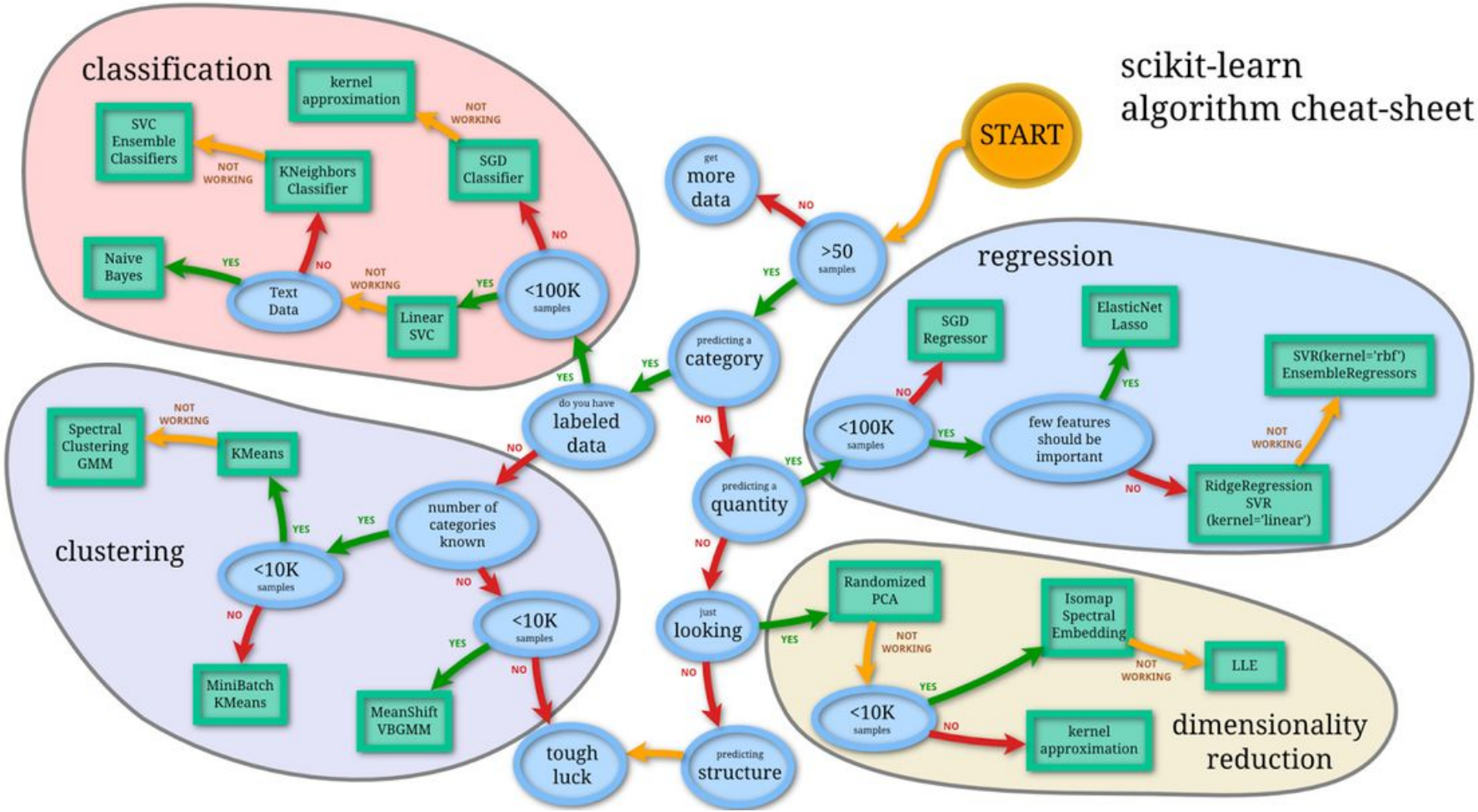
Supervised Learning

Classification
- Decision tree
- Naive Bayes
- SVM
- KNN

Regression

Unsupervised Learning

Clustering
- K-means



Ref : http://5047-presscdn.pagely.netdna-cdn.com/wp-content/uploads/2015/04/drop_shadows_background2.png

Kanda Tiwatthanont @ TNI

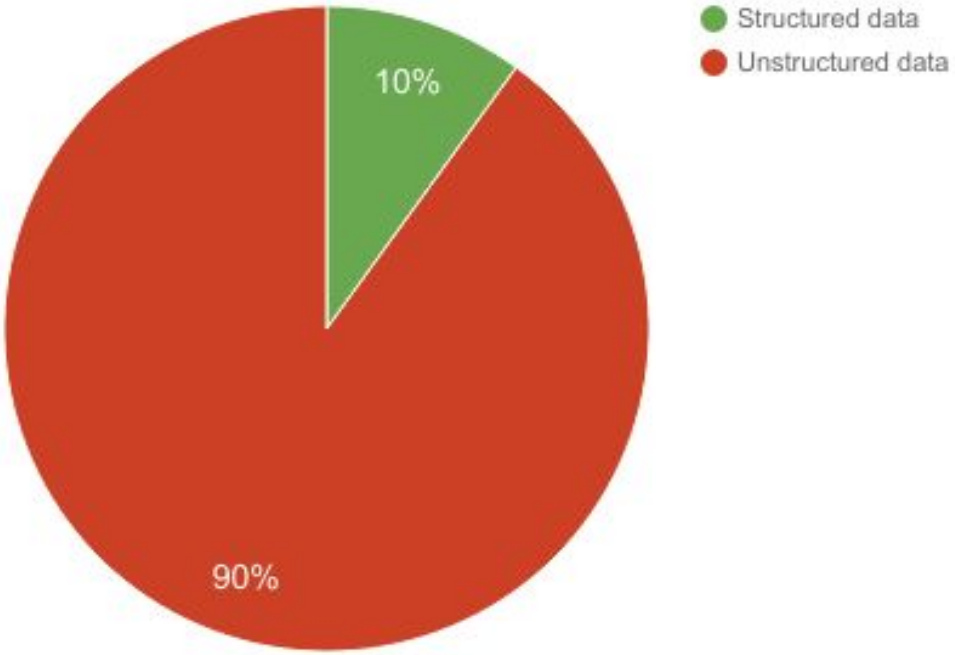
Part 1

Introduction
Text Mining

Text mining

- Mining patterns from unstructured data

Structured vs. Unstructured data

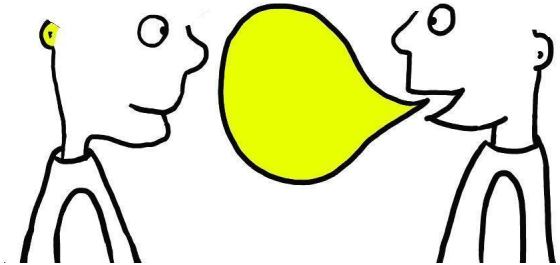


http://www.robertprimmer.com/blog/structured-vs-unstructured.html

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		Data Retrieval	Data Mining
Unstructured data		Information Retrieval	Text Mining

Text Mining Application

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

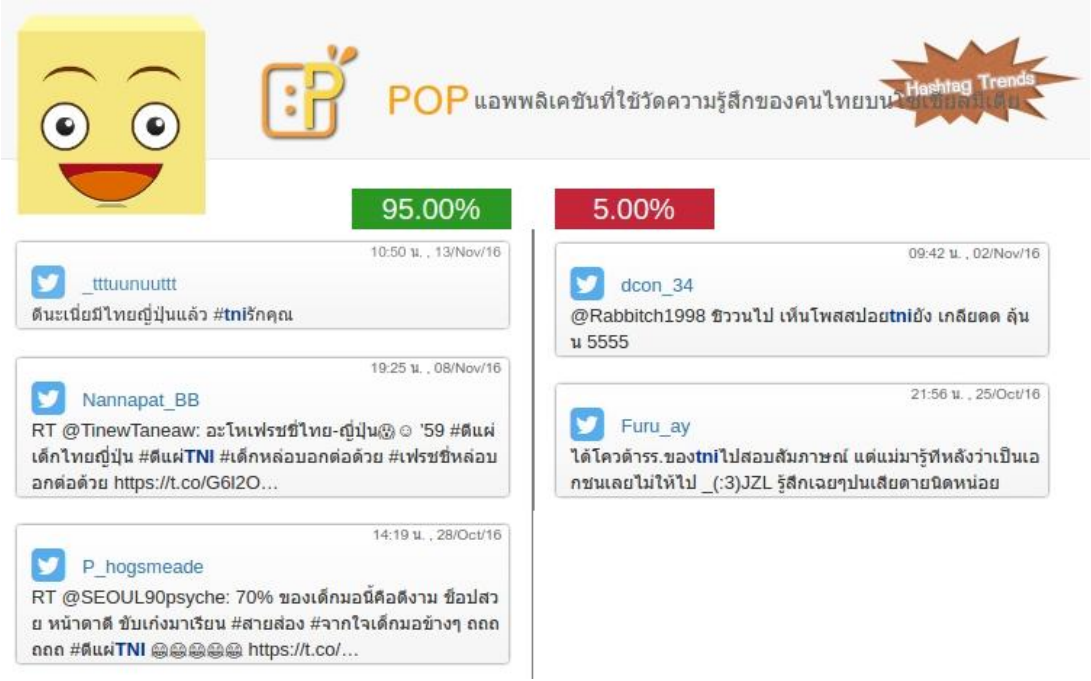
http://textcompactor.com/

S-SENSE

Text Mining Application

Text mining

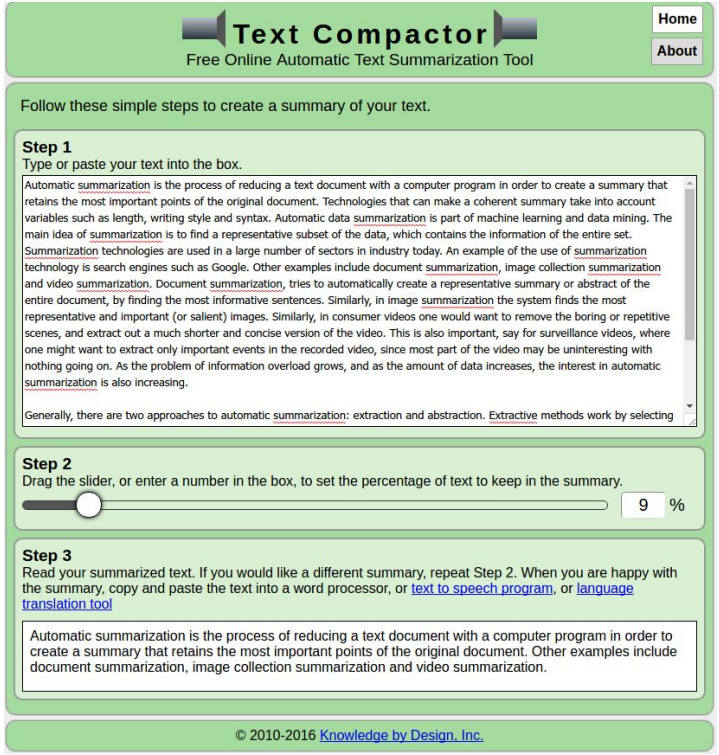
Keyword “TNI”



Text Mining Application

Text summarization

http://textcompactor.com/



Text Mining Application

- ★ Brand Monitoring
- ★ Feedback
- ★ Competitors Comparing

Data Analytics Specialization
Which tool to Choose?

Mining Tools

Business

rapidminer

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

orange
DATA MINING
FRUITFUL&FUN

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

Academic

WEKA
The University of Waikato

Very sophisticated
Free, but hard to use

R

Statistical and math computing

scikit learn

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

Basic

MySQL
- Easy to use operators
- Very popular among data scientist

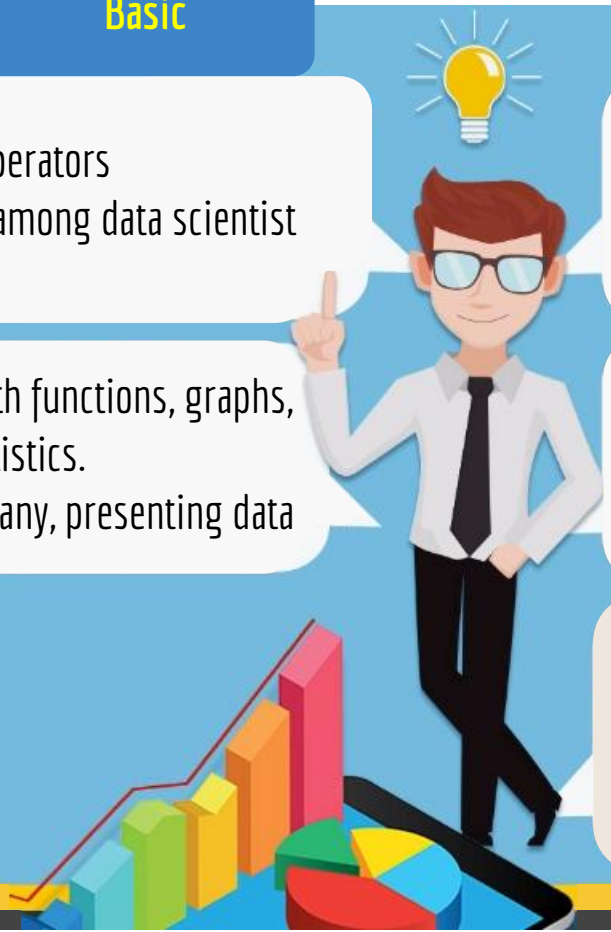
X
- Interactively with functions, graphs, and limited statistics.
- apply in a company, presenting data

Advance

sas
- Easy to learn
- Enterprises using SAS

R
- 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)



Ref : http://www.digitalvidya.com/blog/choose-data-analytics-specialization/

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%

Languages

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

Deep Learning

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

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

Programming Languages (Data Analytics)

Analysis / Data mining :

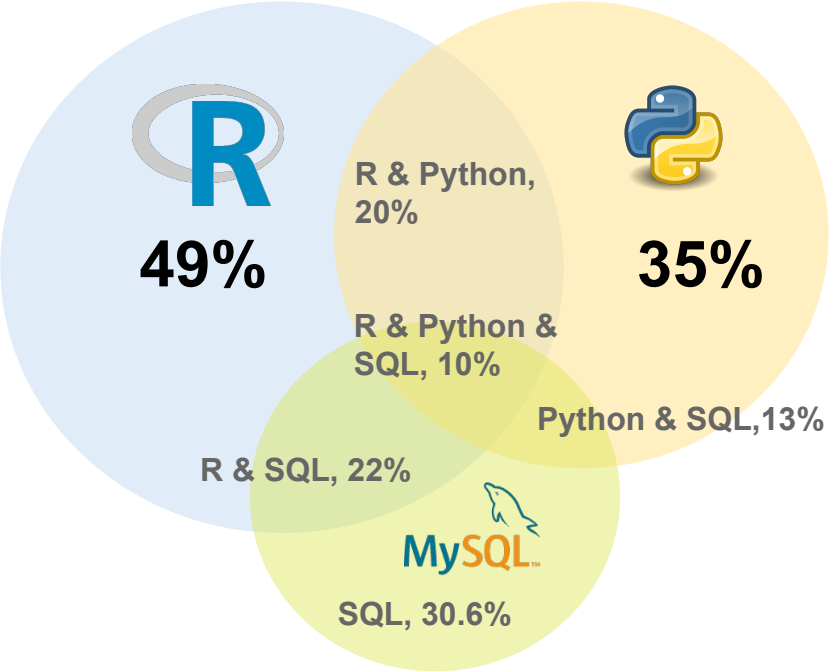
- R Language
- **Python**
- SQL

Big Data (Hadoop)

- Java
- **Python**

Visualization

- JavaScript



Next

Part 2 - Python

Kanda Tiwatthanont @ TNI