INTRODUCTION TO DATA ANALYTICS

Introduction

to the Data Analytics Course

Asst. Prof. Dr. Rathachai Chawuthai

Department of Computer Engineering
Faculty of Engineering
King Mongkut's Institute of Technology Ladkrabang

Agenda

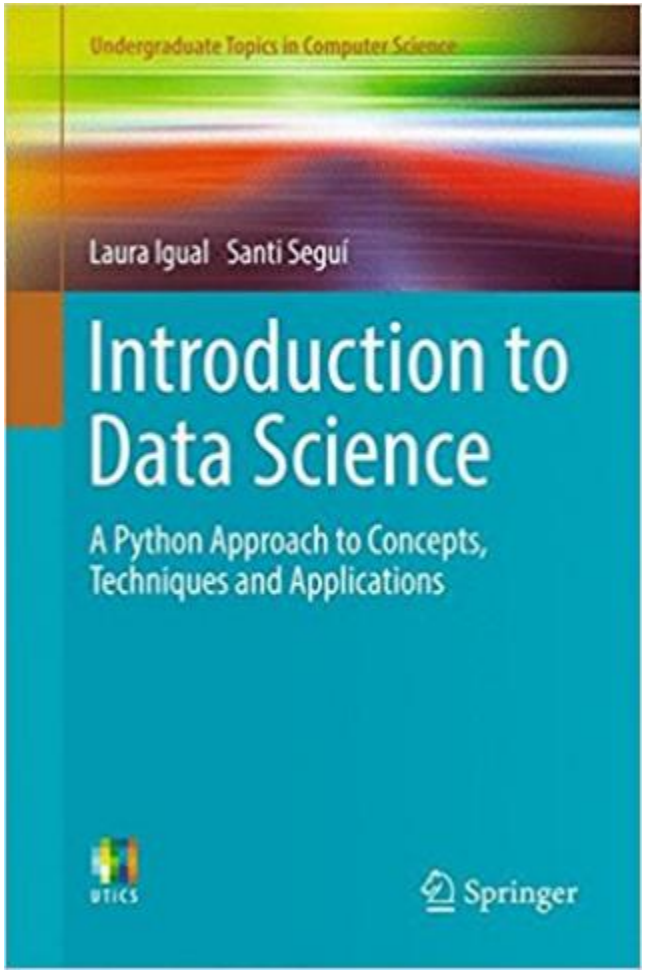
- About this Course
- Data Analytics

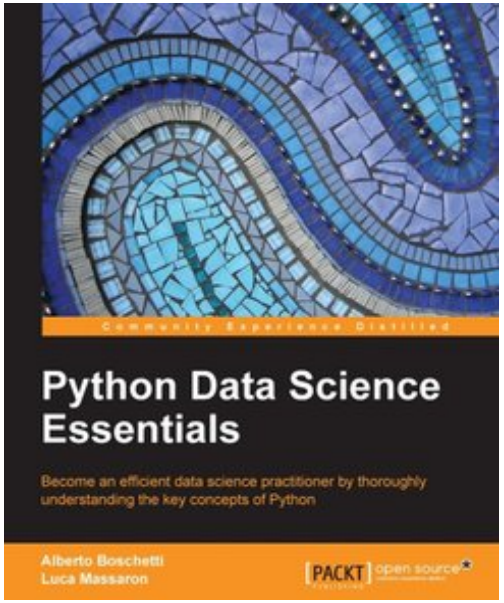
About this Course

Course Outline

- Introduction
- Python for DA
- Data Exploration
- Data Processing
- Regression Analysis
- Classification Analysis
- Cluster Analysis
- Recommender System
- Visualization

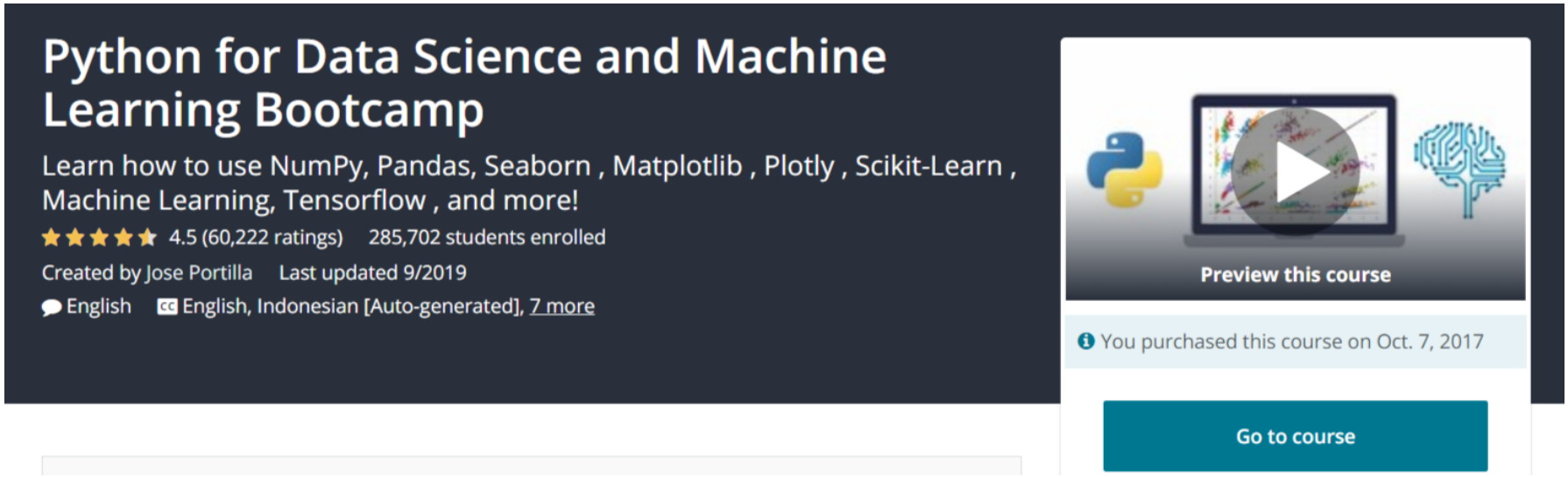
Books





http://www.springer.com/gp/book/9783319500164

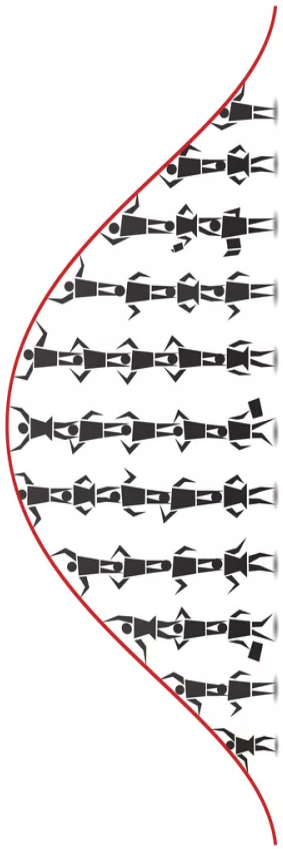
E-Learning



https://www.udemy.com/course/
python-for-data-science-and-machine-learning-bootcamp/

Grading (Plan)

- Midterm 30%
- Final 30%
- Assignments 10%
- Project 30%



Score	Grade
≥ X	A
	B+
	B
	C+
	C
	D+
	D
< Y	F

Tools

Anaconda

- Language: Python
- Package: NumPy, SciPy, Pandas, Scikit-learn, ... , etc.
- IDE: Spyder, Jupyter
- Download:

https://www.anaconda.com/download/

Data Analytics

Questions ?

ทุกการเรียงสับข้อมูล เป็นงานในการค้นหา

How much is the price of 1 acre of land in this area?

How to commute from here to Icon Siam at 8:00?

Will this be a flu or not?

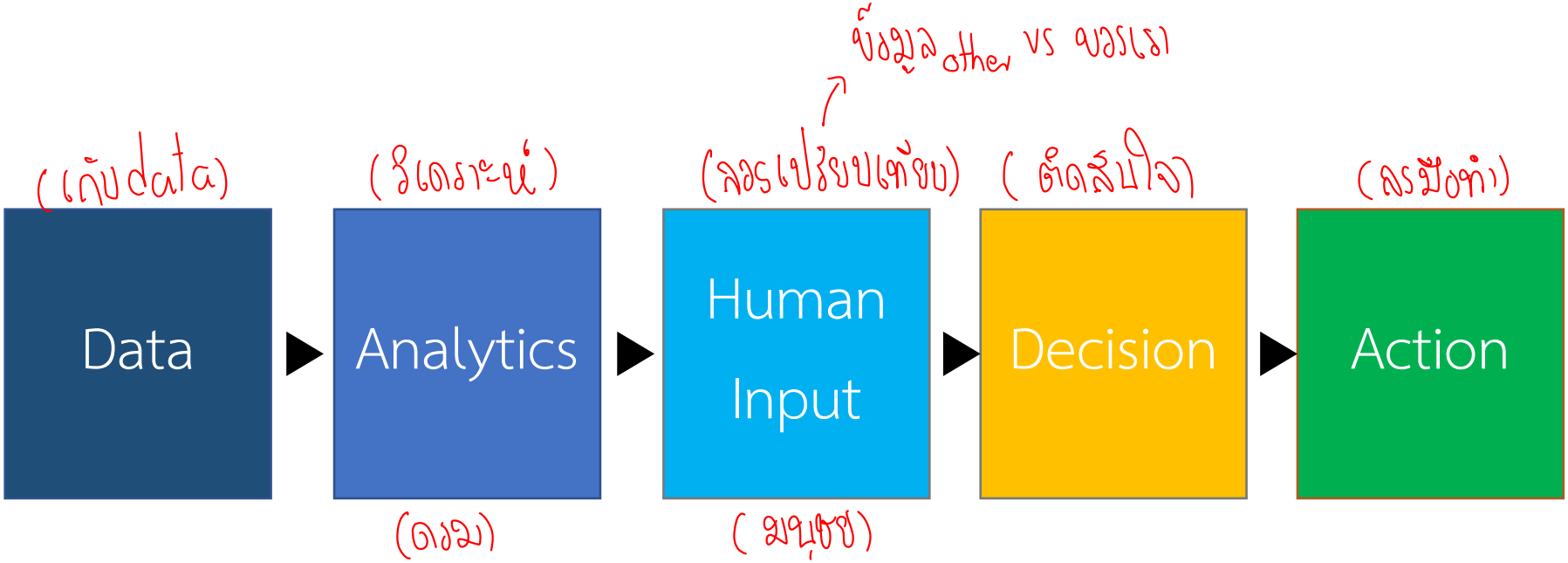
If the customer has already bought this item,

which products should be recommended next?

Where should the gas station be opened?

Data Analytics

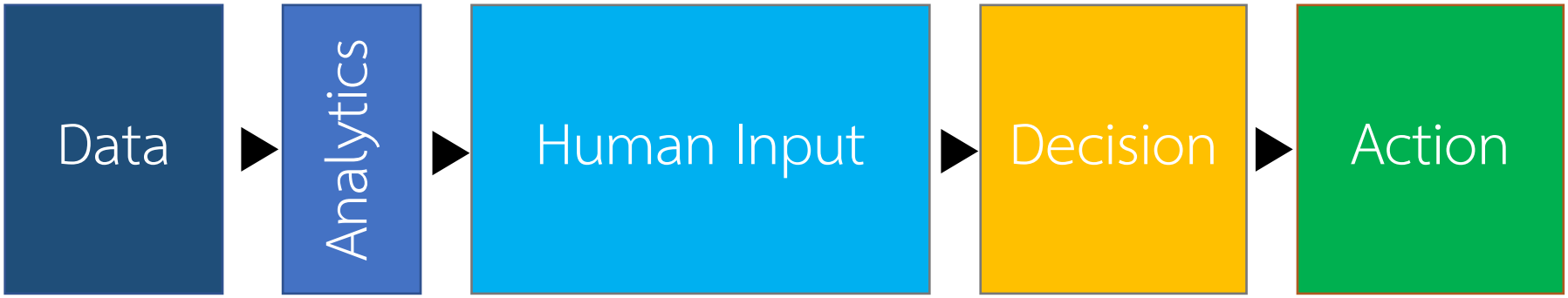
กระบวนการของ data ana



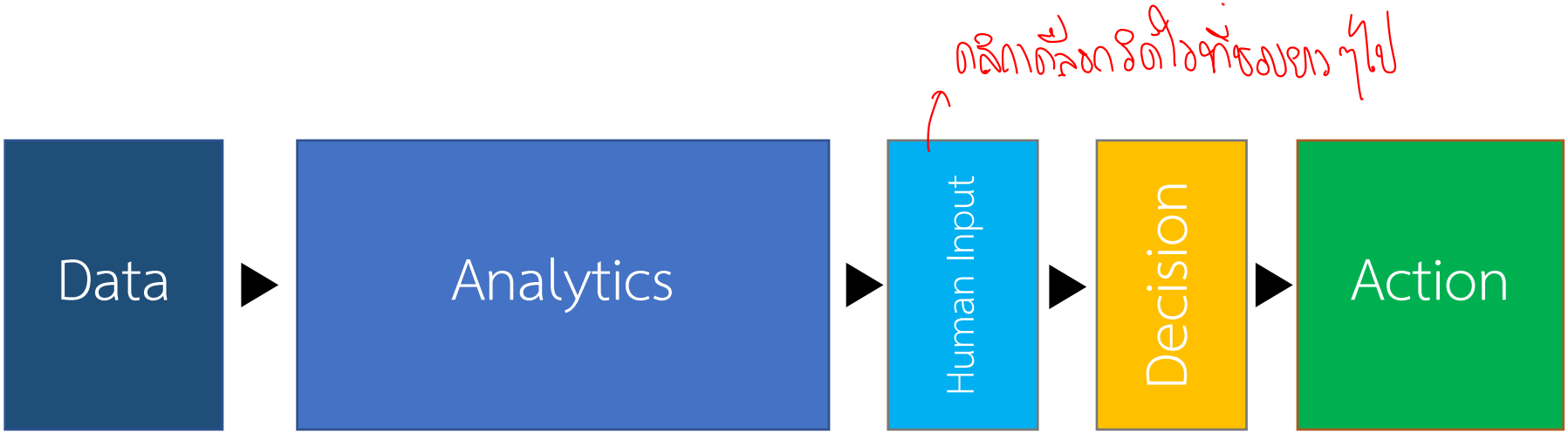
ถ้า analytics น้อย → คนก็ต้องตัดเยอะมากขึ้น

ท้ายสุดเสียอยากให้ มนุษย์ไม่ต้องออกแรงตัดเลย

Data Analytics



Data Analytics



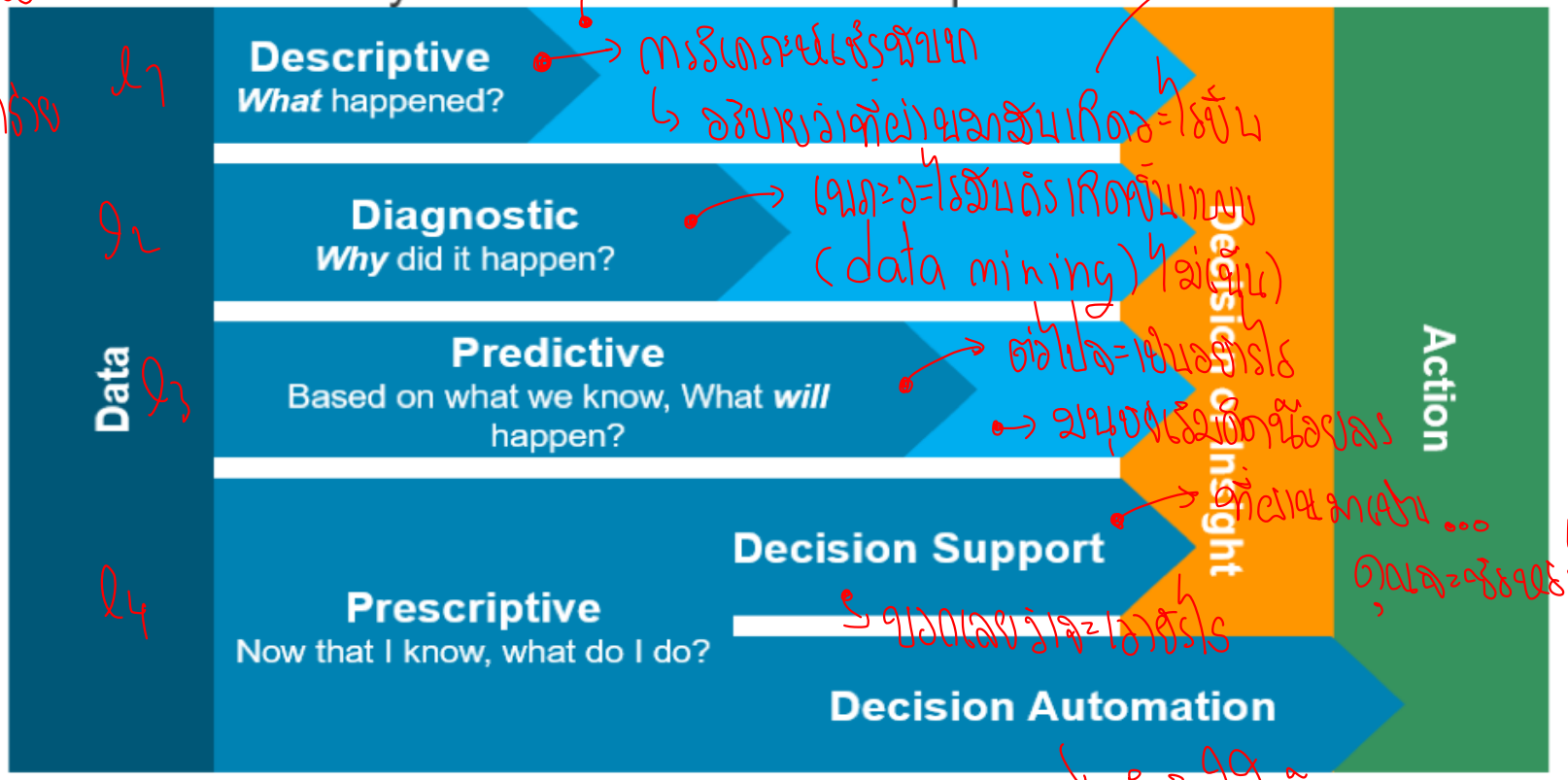
Data Analytics

Data Analytics

- Mathematics
- Business
- Programming

Ref: • (image) https://sensanalytics.com/executive-master-classes/introduction-data-science/

Data Analytics



Ref:
- Four types of analytics capability (Gartner, 2014)
- (image) https://www.healthcatalyst.com/closed-loop-analytics-method-healthcare-data-insights

Descriptive Analytics → ดึงจากข้อมูลในอดีต → สินค้าอะไรขายดีบ้าง, ช่วงเวลาไหนมากสุด

Descriptive analytics is the **interpretation of historical data** to better understand changes that have occurred in a business. Descriptive analytics describes the use of a range of historic data to draw comparisons.

Ref:
- (content) https://www.investopedia.com/terms/d/descriptive-analytics.asp
- (image) https://www.freepik.com/free-photos-vectors/chart

Descriptive Analytics

Questions:

- Which are the best-seller products?
- Which are the most or least revenue-generating products?
- Which are the most successful promotional campaigns?
- Who are the most paying customers?
- What are revenue trends for each Strategic Business Unit (SBU) of last N years, last N months?

Ref: • https://www.chulamoocachieve.com/pathway/data-science

Descriptive Analytics

Techniques:

ดูคุณภาพของข้อมูล

- Exploratory Data Analysis
- Measure of the Shape of the Distribution
- Measure of Data Summary
- Measure of Variability or Dispersion
  - Standard Deviation, Interquartile Range, Range
- Measure of Central Tendency
  - Mean, Median, Mode, Min, Max

Ref: • https://www.chulamoocachieve.com/pathway/data-science

Predictive Analytics → ดูกราฟแนวโน้ม → ใช้ทำนาย

Predictive analytics is the practice of extracting information from existing data sets in order to **determine patterns and predict future outcomes and trends**. Predictive analytics does not tell you what will happen in the future.

Ref:
- (content) https://www.webopedia.com/TERM/P/predictive_analytics.html
- (image) https://socioboard.org/blog/prediction-instagram-mobile-ad-sales-to-surpass-2-81-billion-in-2017/

Predictive Analytics

Questions:

- What is going to be likely revenue for each SBU in coming year?
- What is going to be likely attrition rate for the common year?
- Who all customers are likely to churn-out?
- Which promotional campaigns are likely to do well?
- Which products are likely to sell most in the next 6 months?

Ref: • https://www.chulamoocachieve.com/pathway/data-science

Predictive Analytics

Techniques:

- Decision Support System
  - Linear Regression
- Classification
  - Decision Tree
  - Logistic Regression
  - Support Vector Machine
  - Artificial Neural Network
  - etc.

Prescriptive Analytics

ช่วยให้เลือกสิ่งที่ดีที่สุด

Prescriptive analytics is a type of data analytics—the use of technology to help businesses make **better decisions** through the analysis of raw data. Specifically, prescriptive analytics factors information about possible situations or scenarios, available resources, past performance, and current performance, and suggests a course of action or strategy. It can be used to make decisions on any time horizon, from immediate to long term.

Ref:
- https://www.investopedia.com/terms/d/descriptive-analytics.asp
- (image) https://www.diytradenews.co.za/digital-dynamism-a-new-dawn-for-retailers/

Prescriptive Analytics

Questions:

- What world be the best channel to sell this product?
- Which of the supplier suggested promotions of adopt?
- What new or replacement items to introduce, and when?
- Ho to modify the overall product assortment for each category?
- What's the next promotion that I can offer to this customer segment?
- What is the best route from the point A to B?

Ref: • https://www.chulamoocachieve.com/pathway/data-science

Prescriptive Analytics

Techniques:

- Decision Support System
- Recommender System
- Search Engine
- Route and Direction Recommendation
- Chatbot

Ref: • https://www.chulamoocachieve.com/pathway/data-science

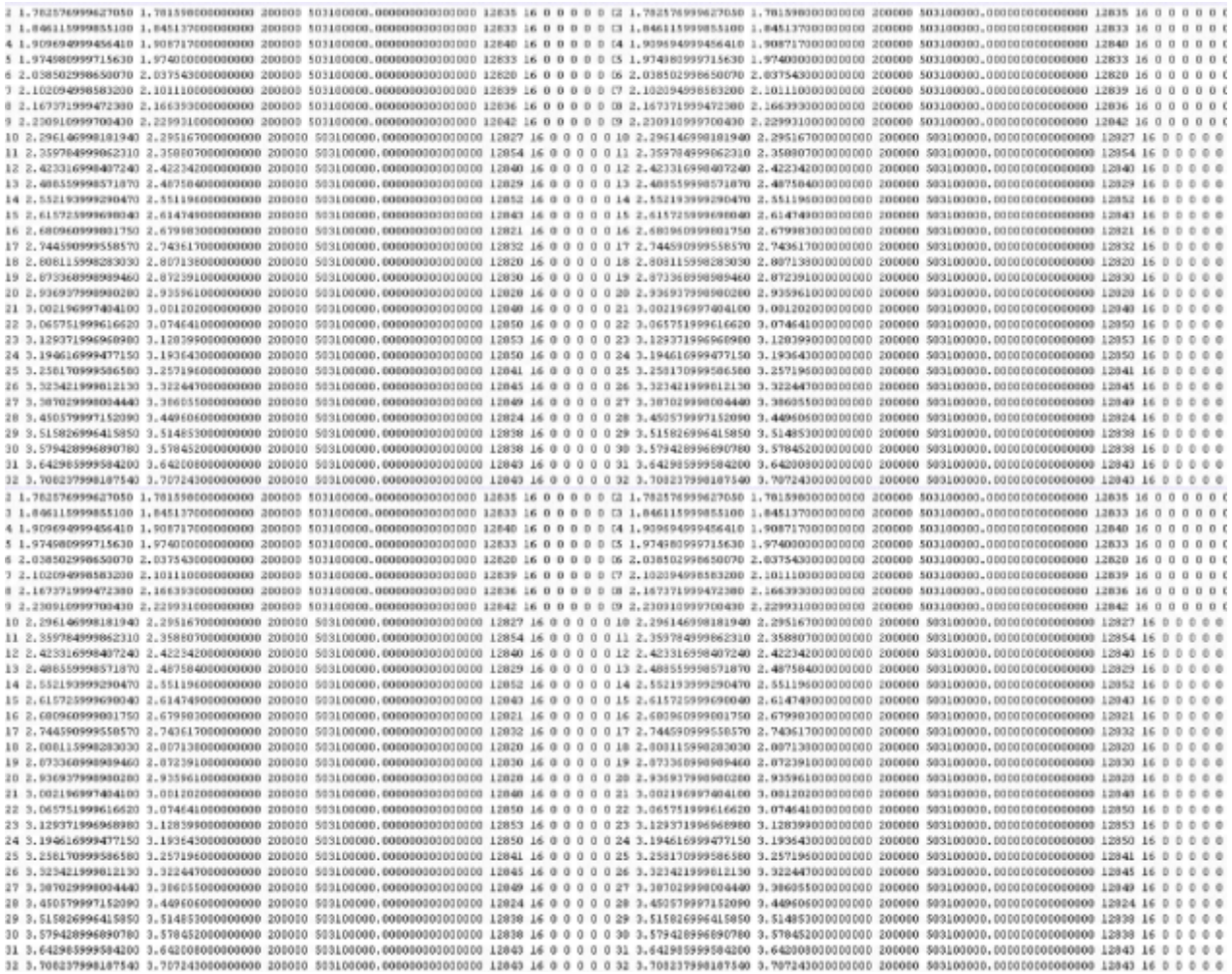
??? Analytics ???

- What was the popular product last month? → descriptive
- What is the average revenue of this product? → descriptive
- What will be the revenue of the next quarter? → predictive
- Which products should be promoted next month? → prescriptive
- Which products should be stopped selling? → prescriptive
- Which place should we promote this product? → prescriptive
- Which products will we recommended to our customers? → prescriptiv
- Will the customers cancel their orders? → predictiv

Ref: • https://www.chulamoocachieve.com/pathway/data-science

When you have a question!

OK?



Ref: • (image) URL

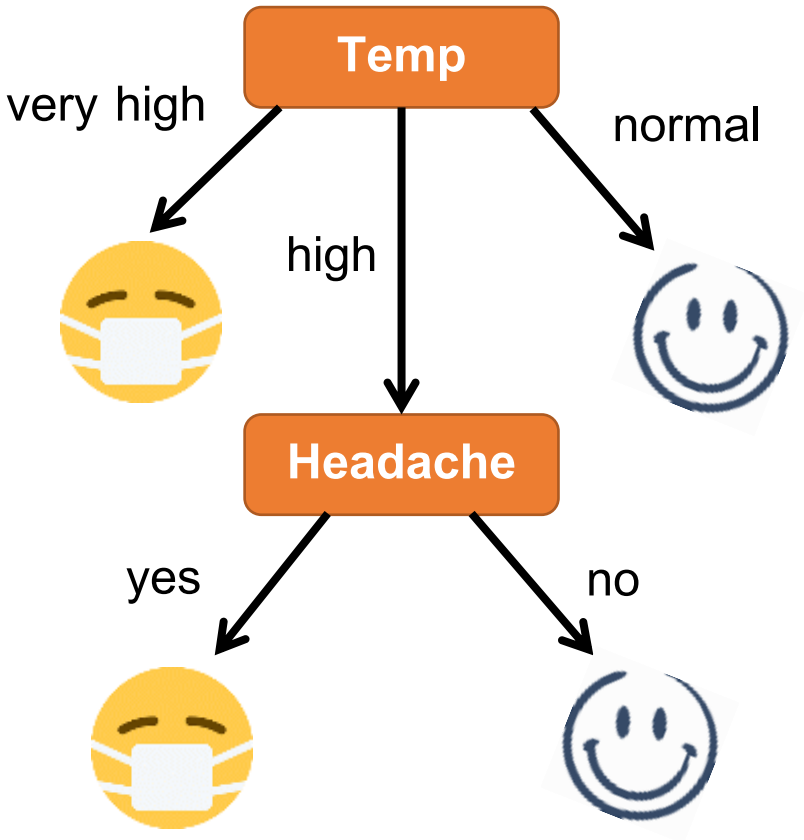
Example: a small dataset

#	Body Temperature	Headache	Nausea	FLU?
1	high	yes	-	**yes**
2	very high	yes	yes	**yes**
3	normal	-	-	**-**
4	high	yes	yes	**yes**
5	high	-	yes	**-**
6	normal	yes	-	**-**
7	normal	-	yes	**-**
	Normal	YES	YES	?

Example: a small dataset

#	Body Temp	Headache	Nausea	FLU?
1	high	yes	-	yes
2	very high	yes	yes	yes
3	normal	-	-	-
4	high	yes	yes	yes
5	high	-	yes	-
6	normal	yes	-	-
7	normal	-	yes	-

	Normal	YES	YES	?



In this age

Much Data → High Accuracy

Much Data → Headache

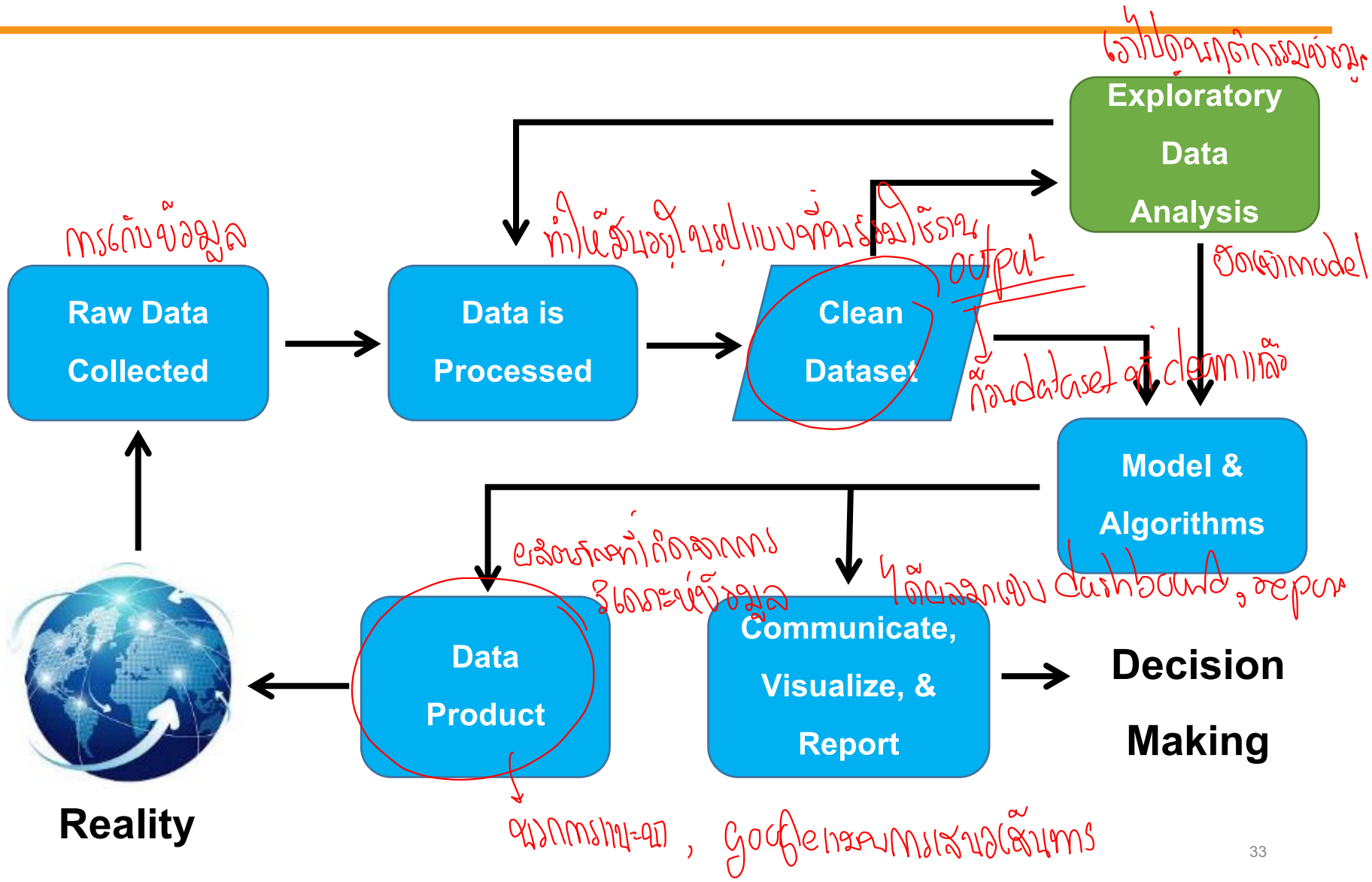
(Then, let's use Computers)

Data Science

Data science, also known as data-driven science, is an interdisciplinary field about **scientific methods, processes and systems to extract knowledge or insights from data** in various forms, either structured or unstructured, similar to Knowledge Discovery in Databases.

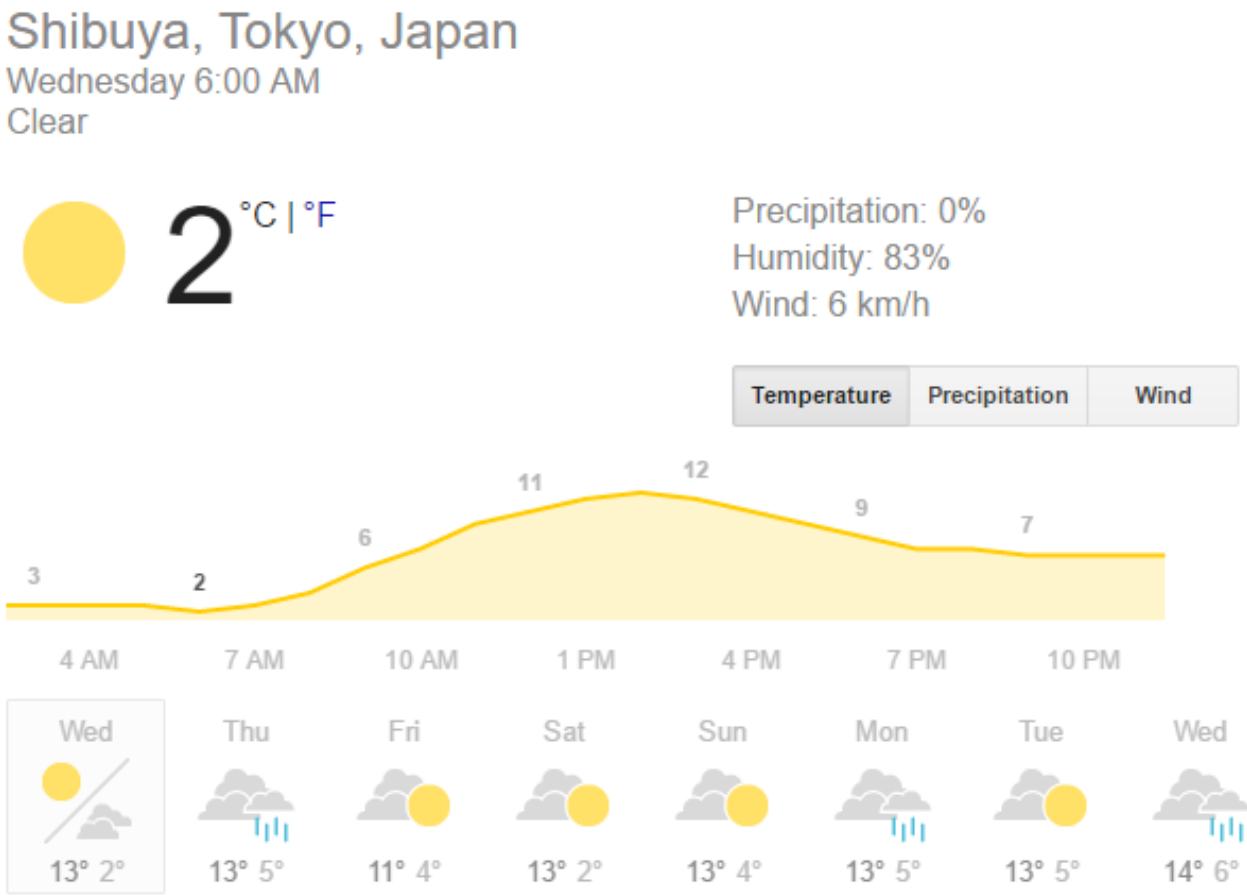
Ref: • (content) https://en.wikipedia.org/wiki/Data_science

Data Science Process



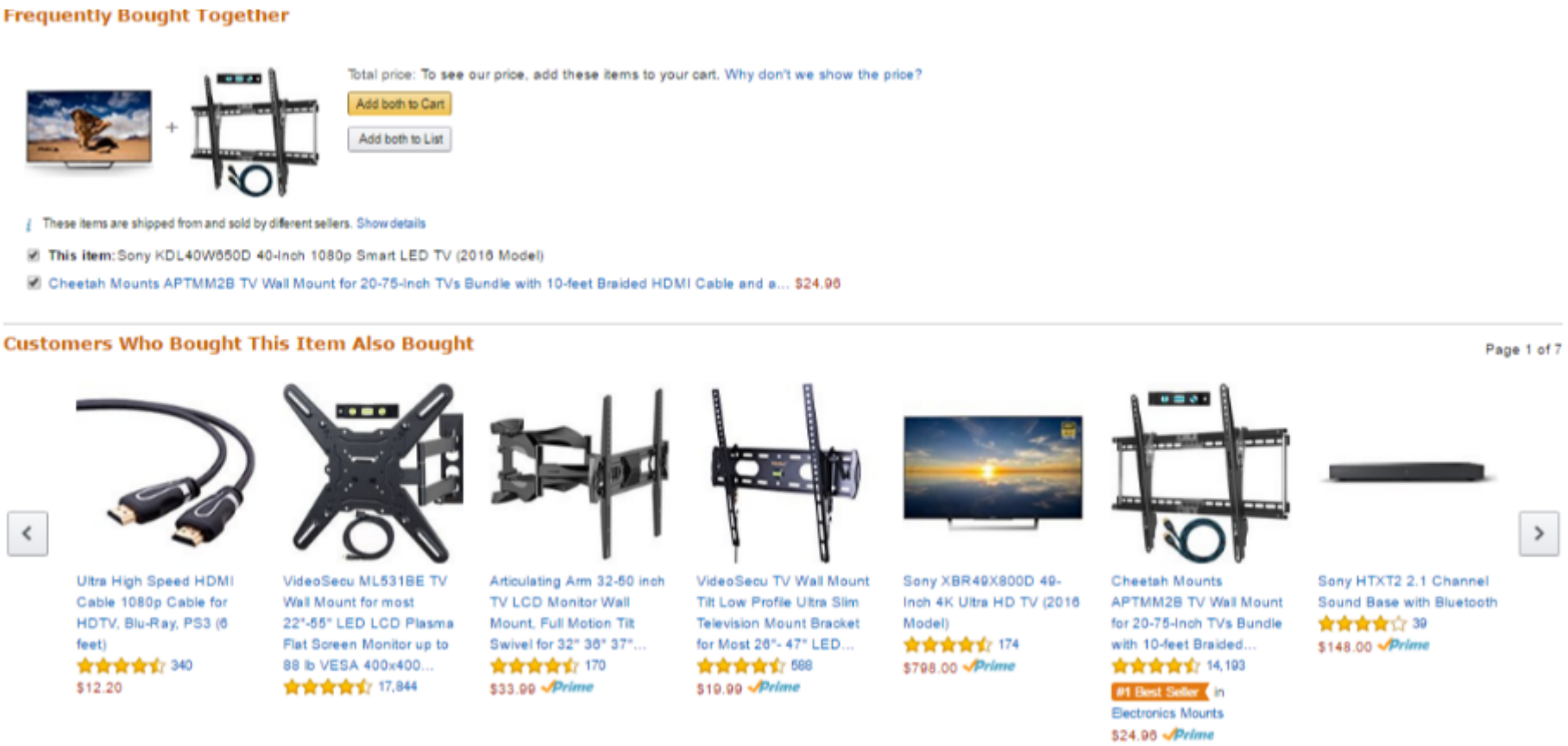
Weather Forecast

ผลิตภัณฑ์ข้อมูล (data product)



Ref: • (image) Google.com

amazon



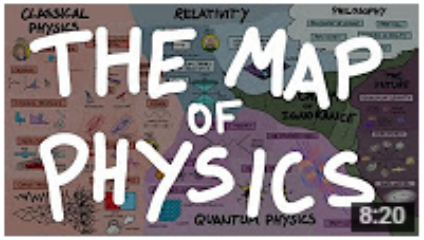
Ref: • (image) Amazon.com

YouTube

Recommended

How the blockchain will radically transform the...
TED
126,953 views • 2 months ago

How to gain control of your free time | Laura Vanderkam
TED
404,362 views • 3 weeks ago



The Map of Physics
DominicWalliman
604,613 views • 3 months ago

สามก๊ก 2010 ตอนที่ 22 (28 กุมภาพันธ์ 2560)
สามก๊ก 2010 TH
165 views • 2 hours ago

The Future of Data Science - Data Science @ Stanford
Stanford
48,516 views • 1 year ago

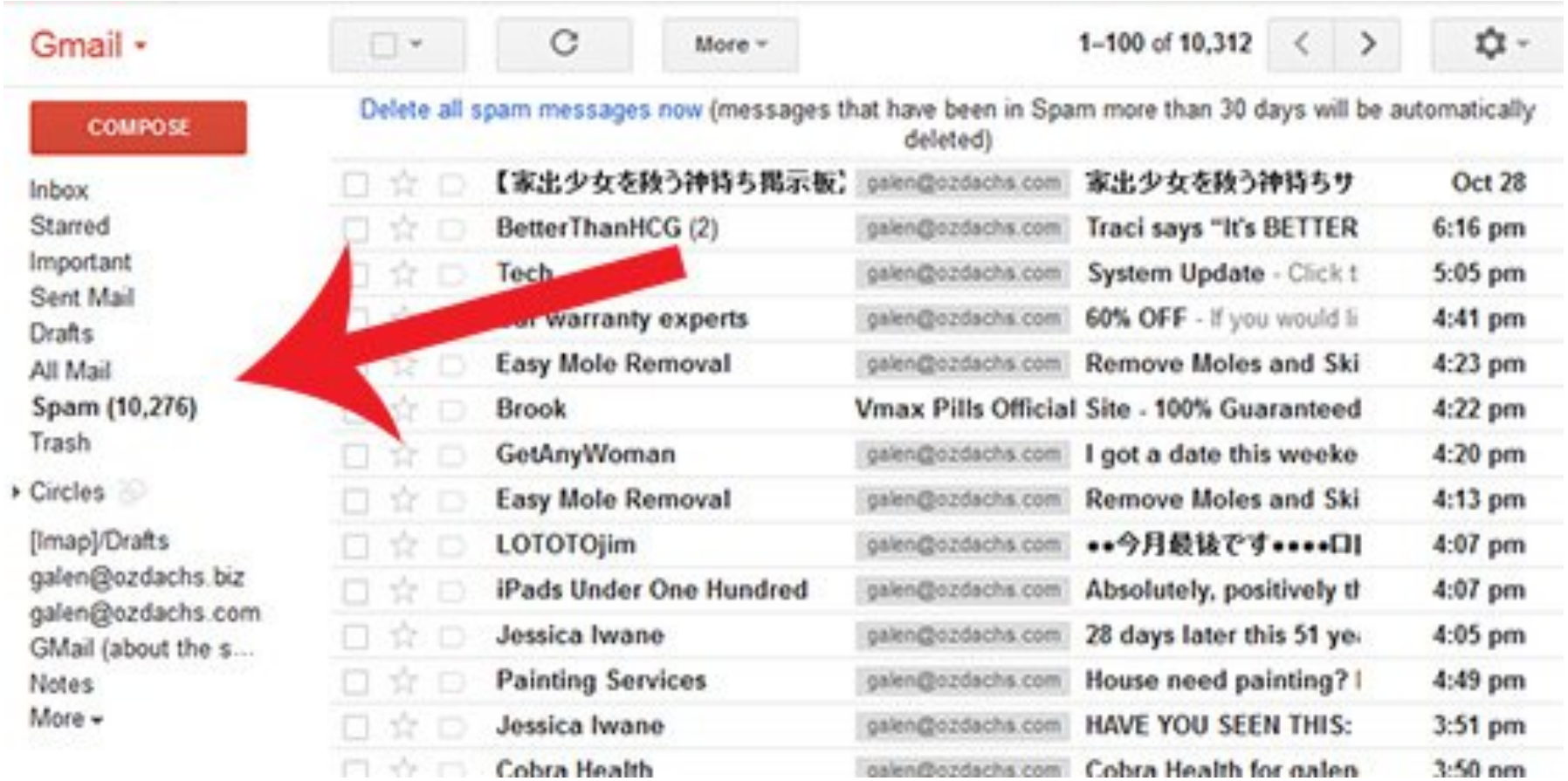
จักรวรรดิมองโกลผู้พิชิตโลก จากยุคเริ่มต้นถึงล่มสลาย
SFG Unicorn
38,487 views • 7 months ago

10 อันดับสุดยอดขุนศึกในยุคสามก๊ก by CHERRYMAN
CHERRYMAN
96,747 views • 3 weeks ago

เขียนโปรแกรม แอนดรอยด์ ด้วย Visual Studio (Develop...
Suppakit Thongdee
5,859 views • 1 year ago

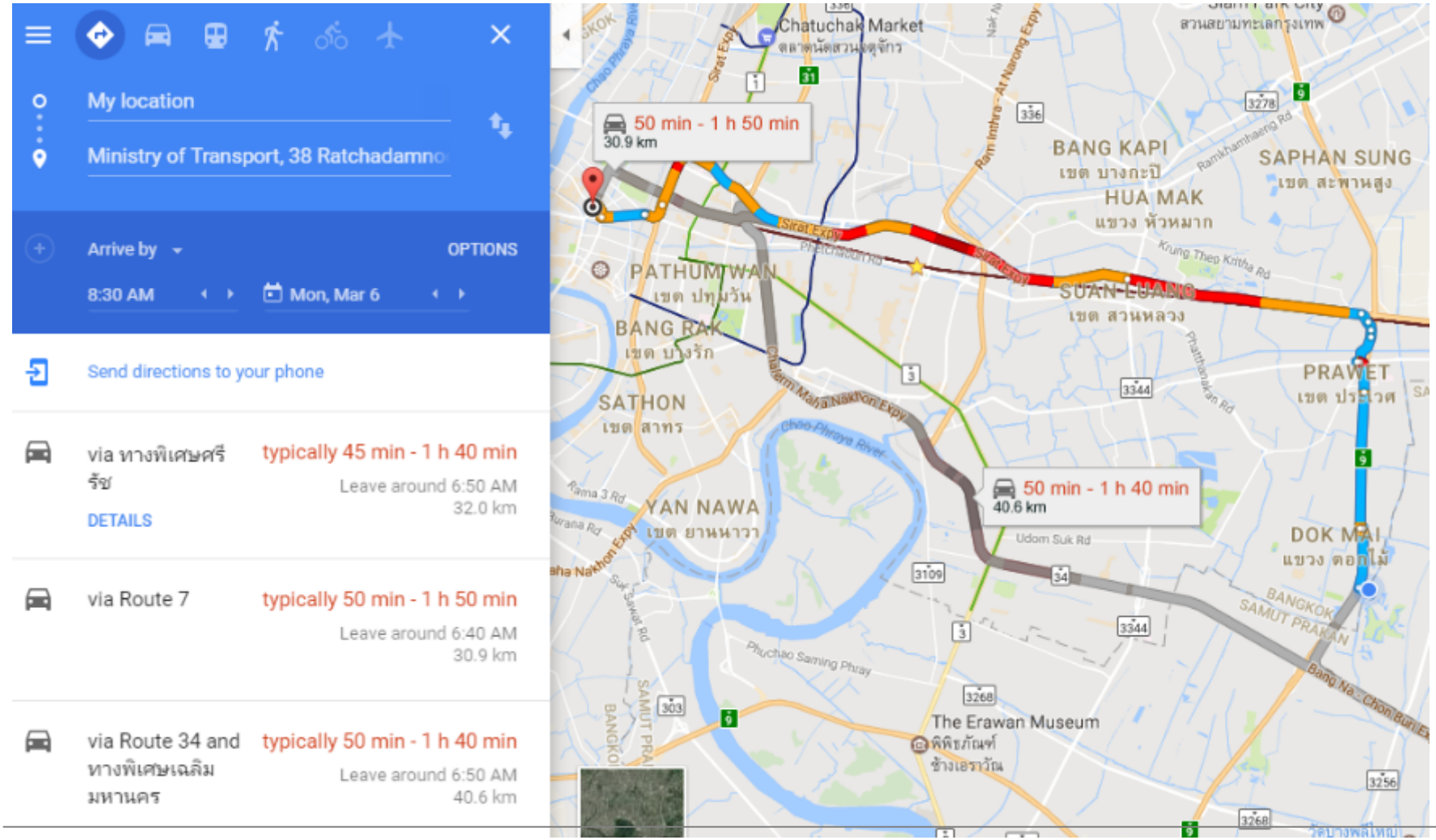
Ref: • (image) Youtube.com

Gmail



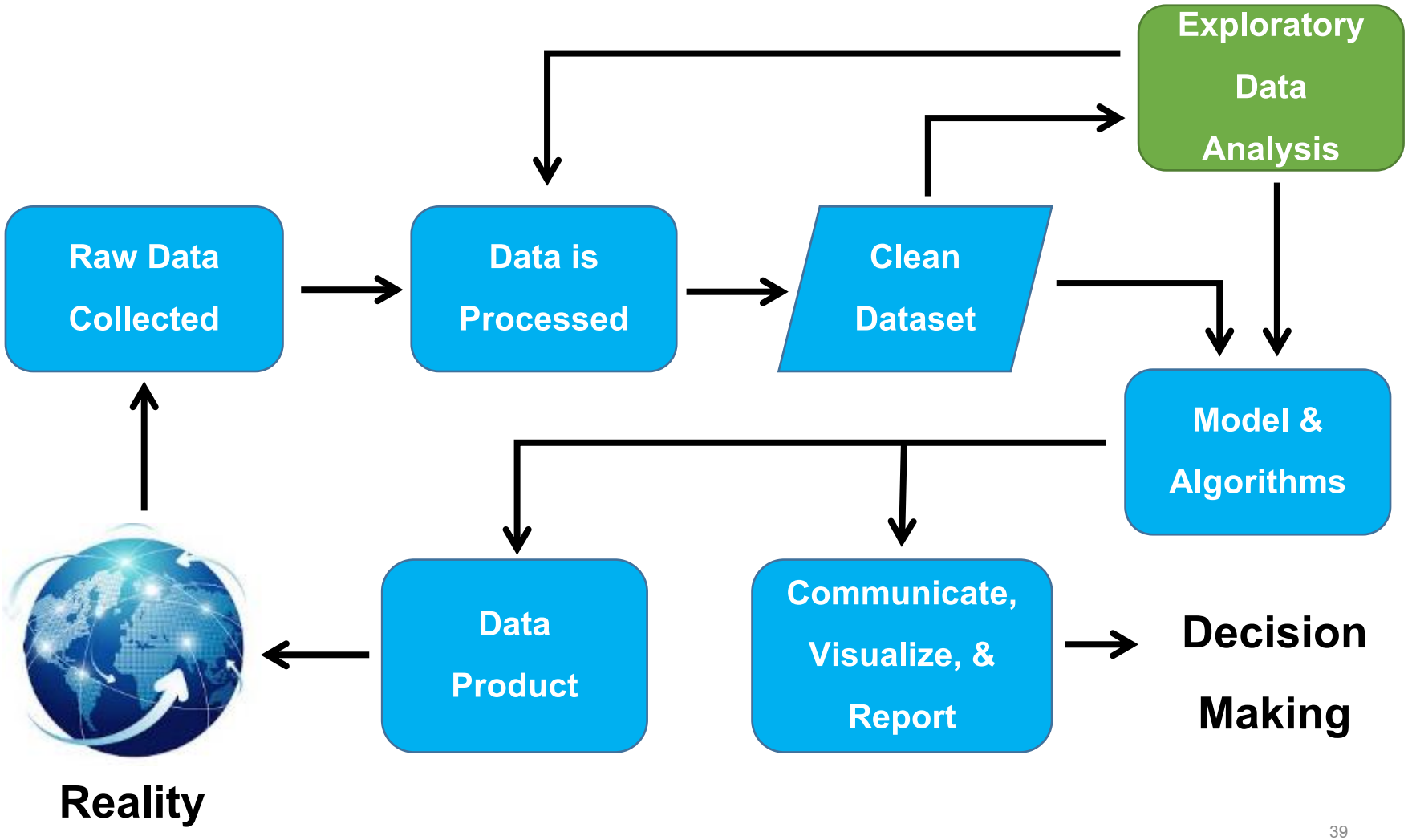
Ref: • (image) mail.google.com

Google Map

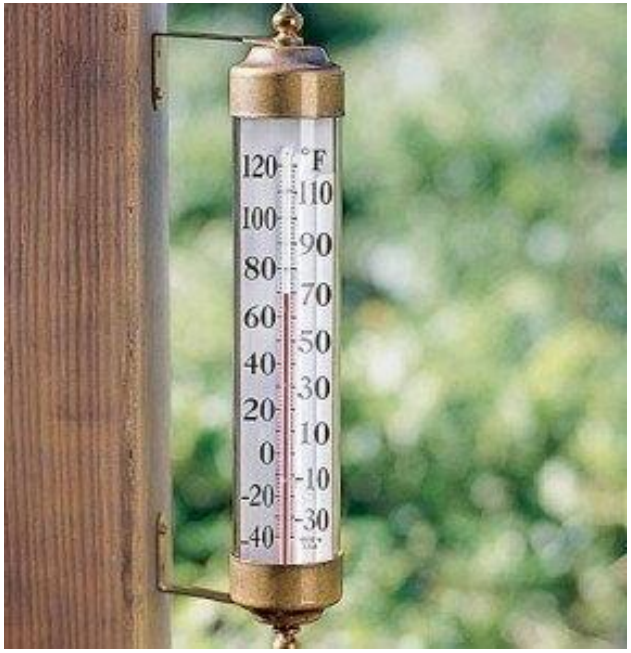


Ref: • (image) map.google.com

Data Science Process



Reality



Raw Data Collected

การเก็บข้อมูล

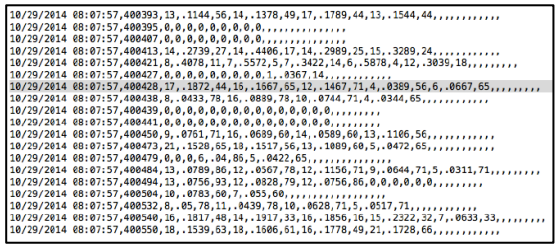
```
10/29/2014 08:07:57,400393,13,.1144,56,14,.1378,49,17,.1789,44,13,.1544,44,,,,,,,,,,
10/29/2014 08:07:57,400395,0,0,0,0,0,0,0,0,0,,,,,,,,,,,,,,
10/29/2014 08:07:57,400407,0,0,0,0,0,0,0,0,0,,,,,,,,,,,,,,
10/29/2014 08:07:57,400413,14,.2739,27,14,.4406,17,14,.2989,25,15,.3289,24,,,,,,,,,,
10/29/2014 08:07:57,400421,8,.4078,11,7,.5572,5,7,.3422,14,6,.5878,4,12,.3039,18,,,,,,,,
10/29/2014 08:07:57,400427,0,0,0,0,0,0,0,0,1,.0367,14,,,,,,,,,,
10/29/2014 08:07:57,400428,17,.1872,44,16,.1667,65,12,.1467,71,4,.0389,56,6,.0667,65,,,,,,,,
10/29/2014 08:07:57,400438,8,.0433,78,16,.0889,78,10,.0744,71,4,.0344,65,,,,,,,,,,
10/29/2014 08:07:57,400439,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,,,,,,,,
10/29/2014 08:07:57,400441,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,,,,,,,,
10/29/2014 08:07:57,400450,9,.0761,71,16,.0689,60,14,.0589,60,13,.1106,56,,,,,,,,,,
10/29/2014 08:07:57,400473,21,.1528,65,18,.1517,56,13,.1089,60,5,.0472,65,,,,,,,,,,
10/29/2014 08:07:57,400479,0,0,0,6,.04,86,5,.0422,65,,,,,,,,,,,,,,
10/29/2014 08:07:57,400484,13,.0789,86,12,.0567,78,12,.1156,71,9,.0644,71,5,.0311,71,,,,,,,,
10/29/2014 08:07:57,400494,13,.0756,93,12,.0828,79,12,.0756,86,0,0,0,0,0,0,,,,,,,,
10/29/2014 08:07:57,400504,10,.0783,60,7,.055,60,,,,,,,,,,,,,,,,
10/29/2014 08:07:57,400532,8,.05,78,11,.0439,78,10,.0628,71,5,.0517,71,,,,,,,,,,
10/29/2014 08:07:57,400540,16,.1817,48,14,.1917,33,16,.1856,16,15,.2322,32,7,.0633,33,,,,,,,,
10/29/2014 08:07:57,400550,18,.1539,63,18,.1606,61,16,.1778,49,21,.1728,66,,,,,,,,,,
```

#	Time	App / Site (1 to 10)	Category	Tags	bulk edit
1	8h 40m	Dreamweaver	Dev Tools	webdev work	[edit]
2	4h 37m	mail.google.com/a/	Comm (Email)	all-comm google-apps work	[edit]
3	3h 31m	Photoshop	Design/Presentation	design webdev work	[edit]
4	2h 36m	mail.google.com	Comm (Email)	all-comm google-apps work	[edit]
5	1h 30m	news.ycombinator.com	News/Blogs	personal	[edit]
6	1h 23m	twitter.com	Social Networking	personal social	[edit]
7	1h 10m	localhost:3000	Dev Tools	webdev work	[edit]
8	45m 20s	rescuetime.com	Personal Productivity	webdev work	[edit]
9	36m 3s	google.com/reader	News/Blogs	google-apps work	[edit]
10	34m 58s	deck.rescuetime.com	Dev Tools	design webdev work	[edit]

1 to 10 of 212 - 10 per page << prev 1 2 3 4 5 ... 21 22 next >> Export to CSV

Data is Processed

การที่ข้อมูลอยู่ใน format ที่ computer อ่านออกได้



Machine-Readable

Format

Ref:
- (content) URL
- (image) URL

Data is Processed

- Merge Data Sets — into the same format
- Rebuild Missing Data — with appropriate values
- Standardize — e.g. same column name
- Normalize — e.g. same date format
- De-Duplicated
- Verify & Enrich — e.g. update the salary values

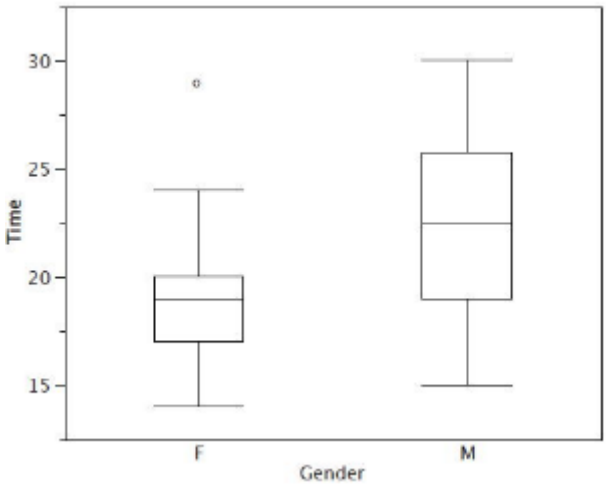
Clean Dataset

ข้อมูลแบบ clean dataset

- Good Format
- Good Shape
- Ready for Data Analysis

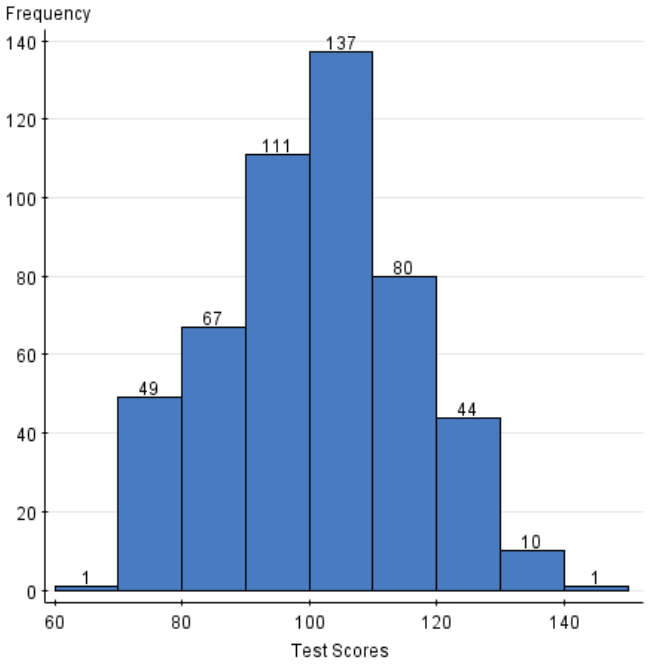
Exploratory Data Analysis สำรวจข้อมูล

เลือกใช้เทคนิคที่เหมาะสมได้ตามประเภทข้อมูล



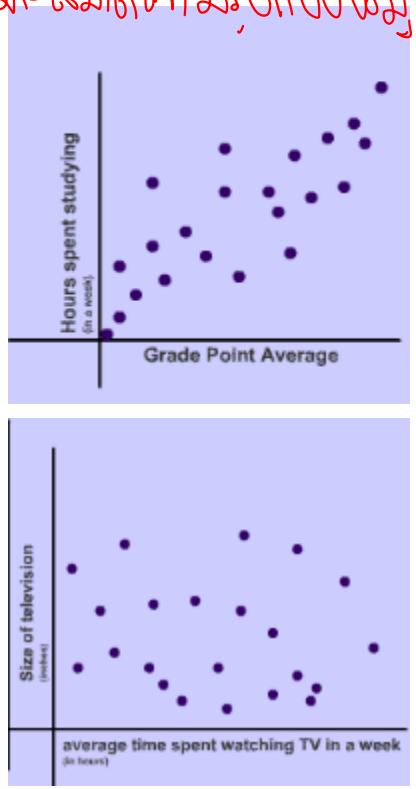
Box Plot

ความเร็วในการบอกชื่อสีที่เห็น
ระหว่างผู้หญิงกับผู้ชาย



Histogram

คะแนนสอบ
กับจำนวนผู้ที่ได้คะแนน



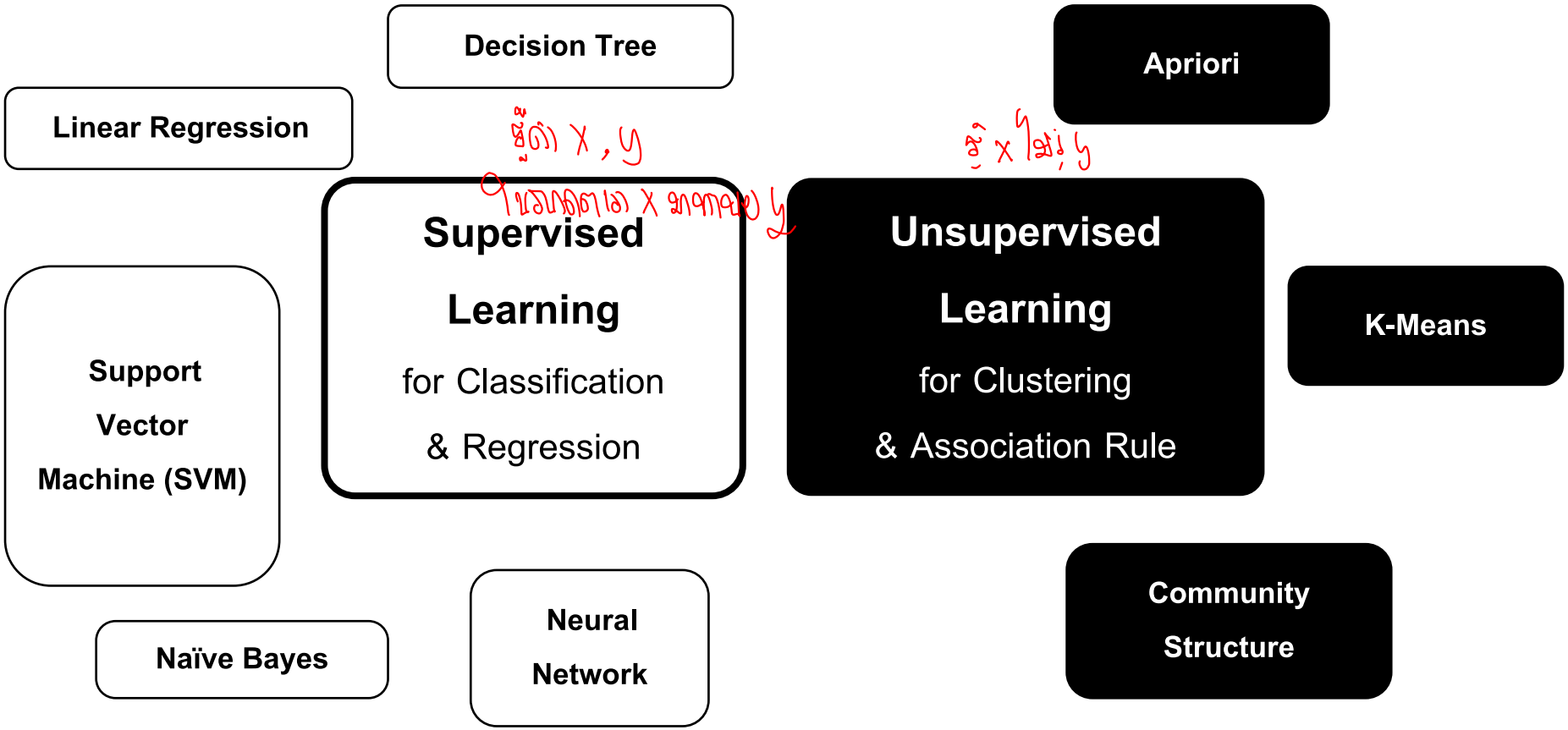
Scatter plot

ดูความสัมพันธ์

Model & Algorithms

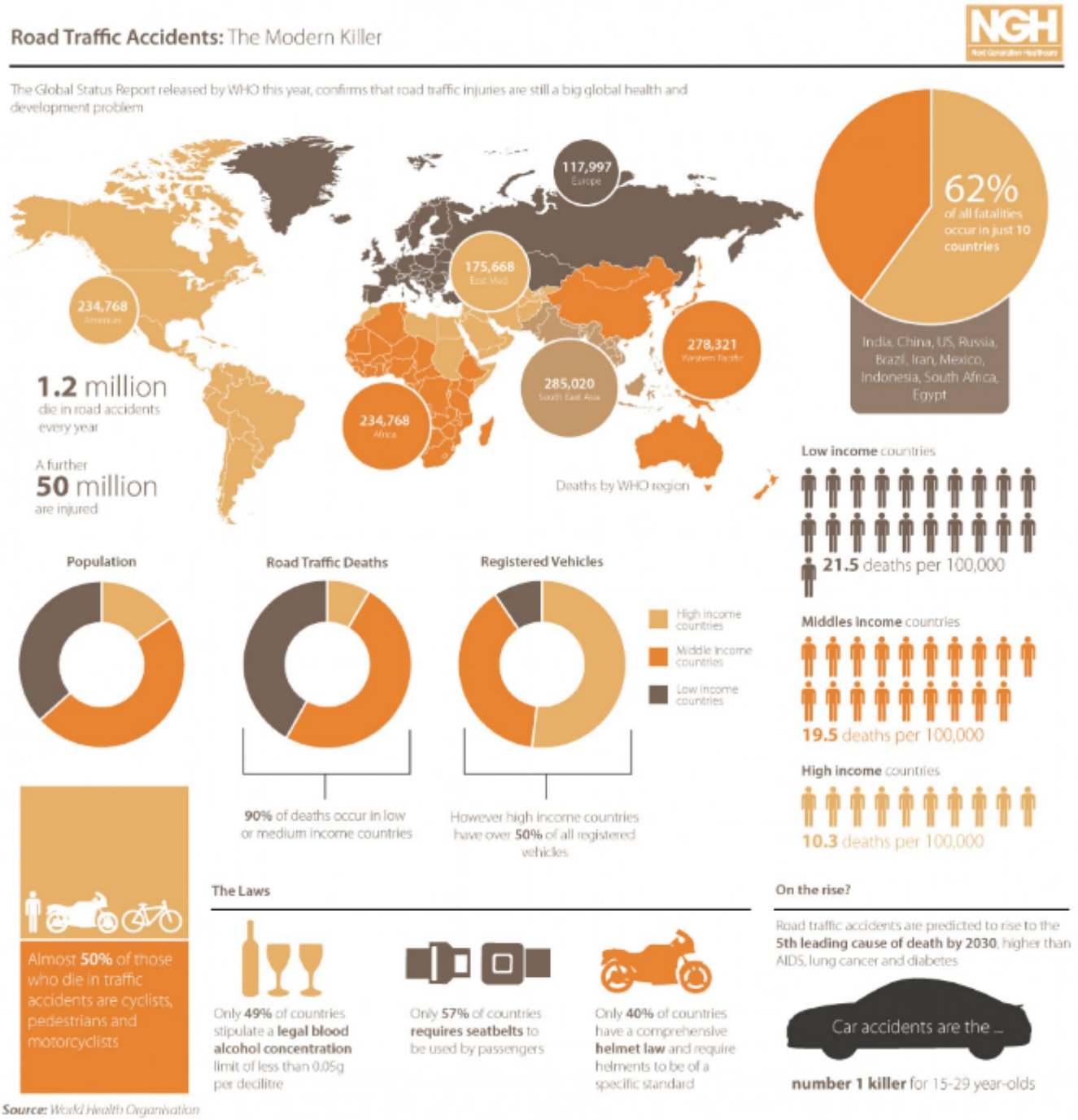
- **Algorithm** → กระบวนการที่จะสร้าง model → tree class ไหน ID Tree ?
  - a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.
- **Model** → สูตรสมการ/แผนภาพที่ได้จากการวิเคราะห์มาแล้ว → หาข้อมูลใหม่ได้
  - A model is a computation or a formula formed as a result of an algorithm that takes some values as input and produces some value as output.
- Example
  - Model: Decision Tree with structure
  - Algorithm: A process to build an accurate decision tree

Model & Algorithms (Example)



(คลิ๊กนา)

Communicate, Visualize, & Report

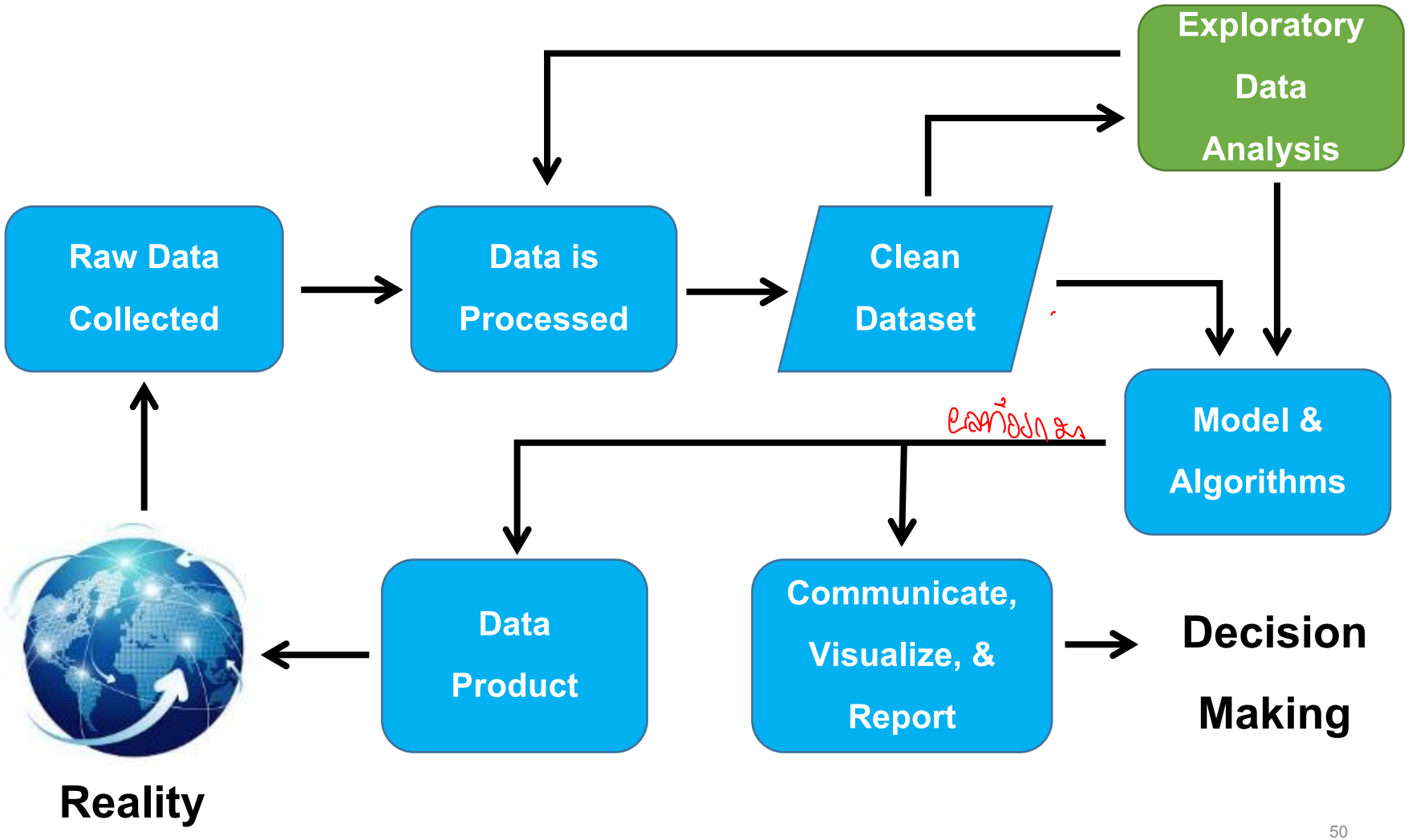


(ยกตัวอย่าง)

Data Product

- A Product that is mainly produced by data analytics
  - Amazon : Recommendation Engine for recommending next products
  - Netflix : Recommendation Engine for recommending next movies
    ↳ การแนะนำสินค้าที่ชอบต่อไป
  - Gmail : Spam filter for identifying junk emails
  - Self-Driving Car: to drive to the destination by understanding traffic signs and traffic conditions in realtime.

Data Science Process



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

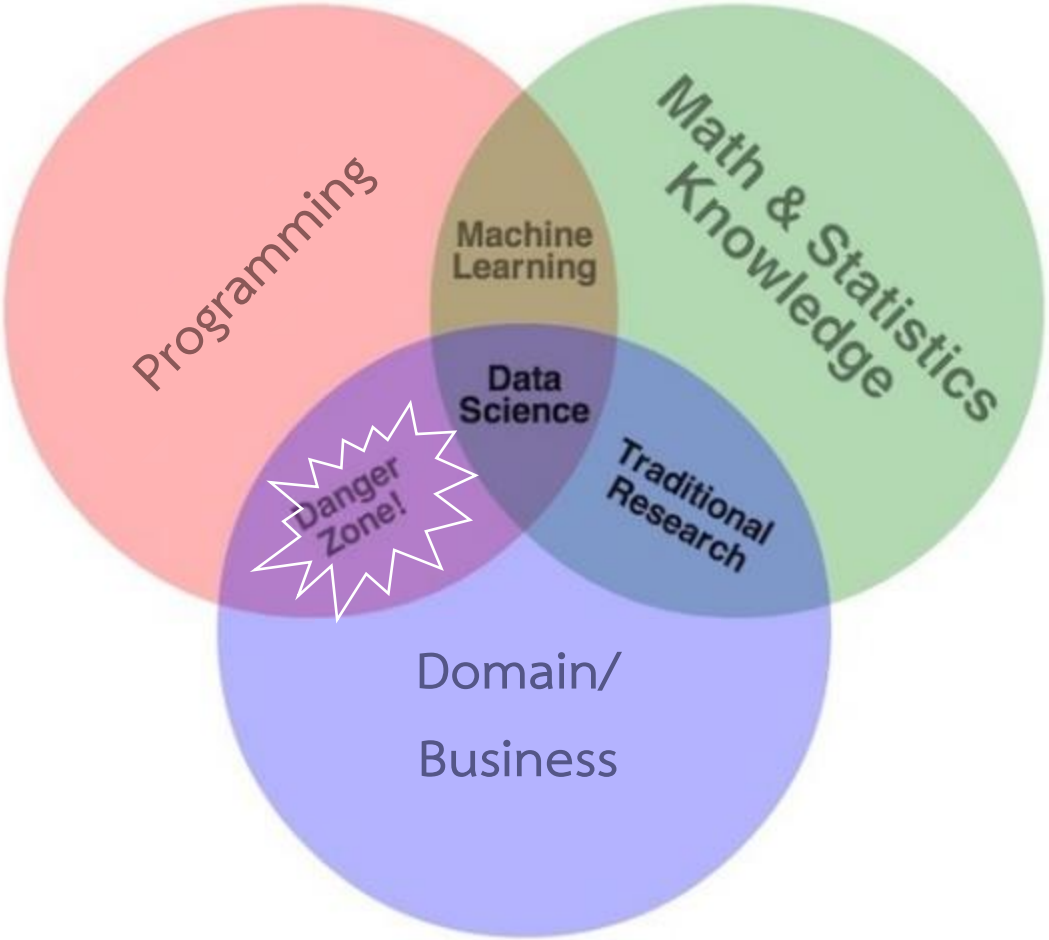
DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

One Last Thing



“

The goal is
to turn data into information,
and information into insight.

”

Carly Fiorina

つづく