



INTRODUCTION TO DATA ANALYTICS

Introduction

to the Data Analytics Course

Dr. Rathachai Chawuthai

Department of Computer Engineering

Faculty of Engineering

King Mongkut's Institute of Technology Ladkrabang

Agenda

- About this Course
- Data Analytics

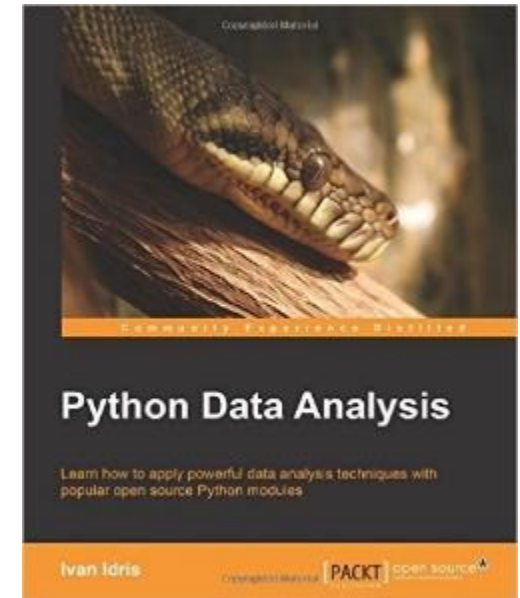
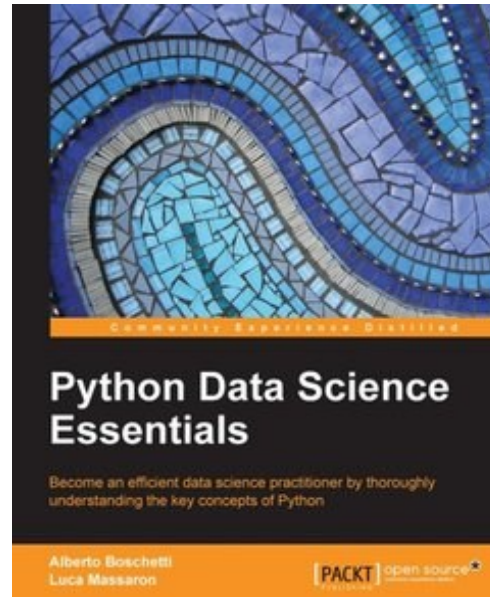
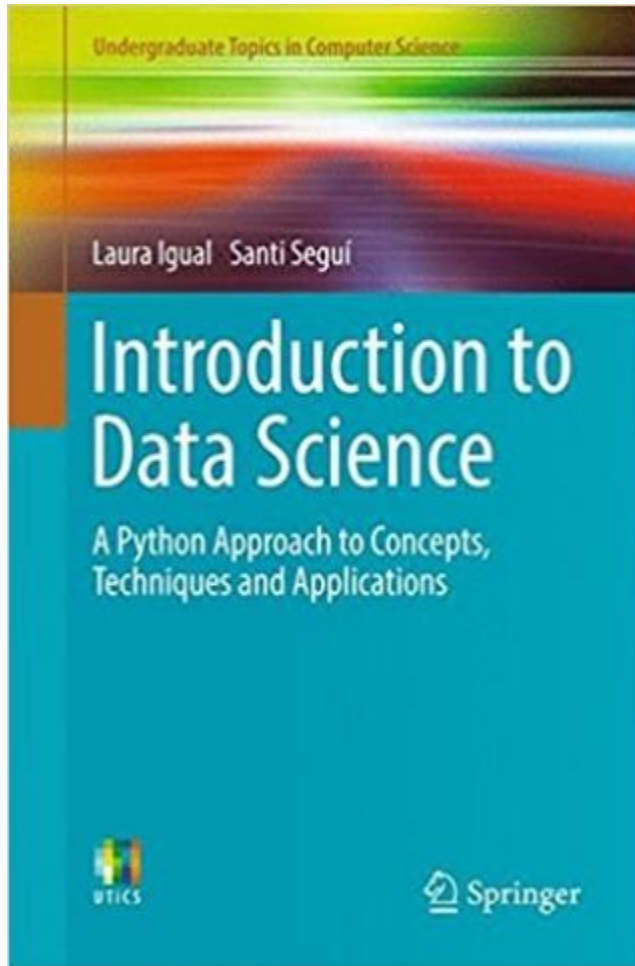
About this Course



Course Outline

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

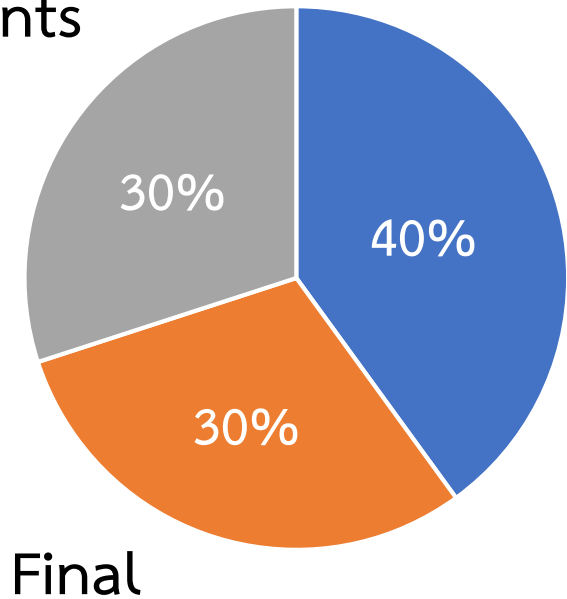
Books



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

Grading (Plan)

Assignments
& Project



Midterm

Final

Score	Grade
$\geq X$	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 ?

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

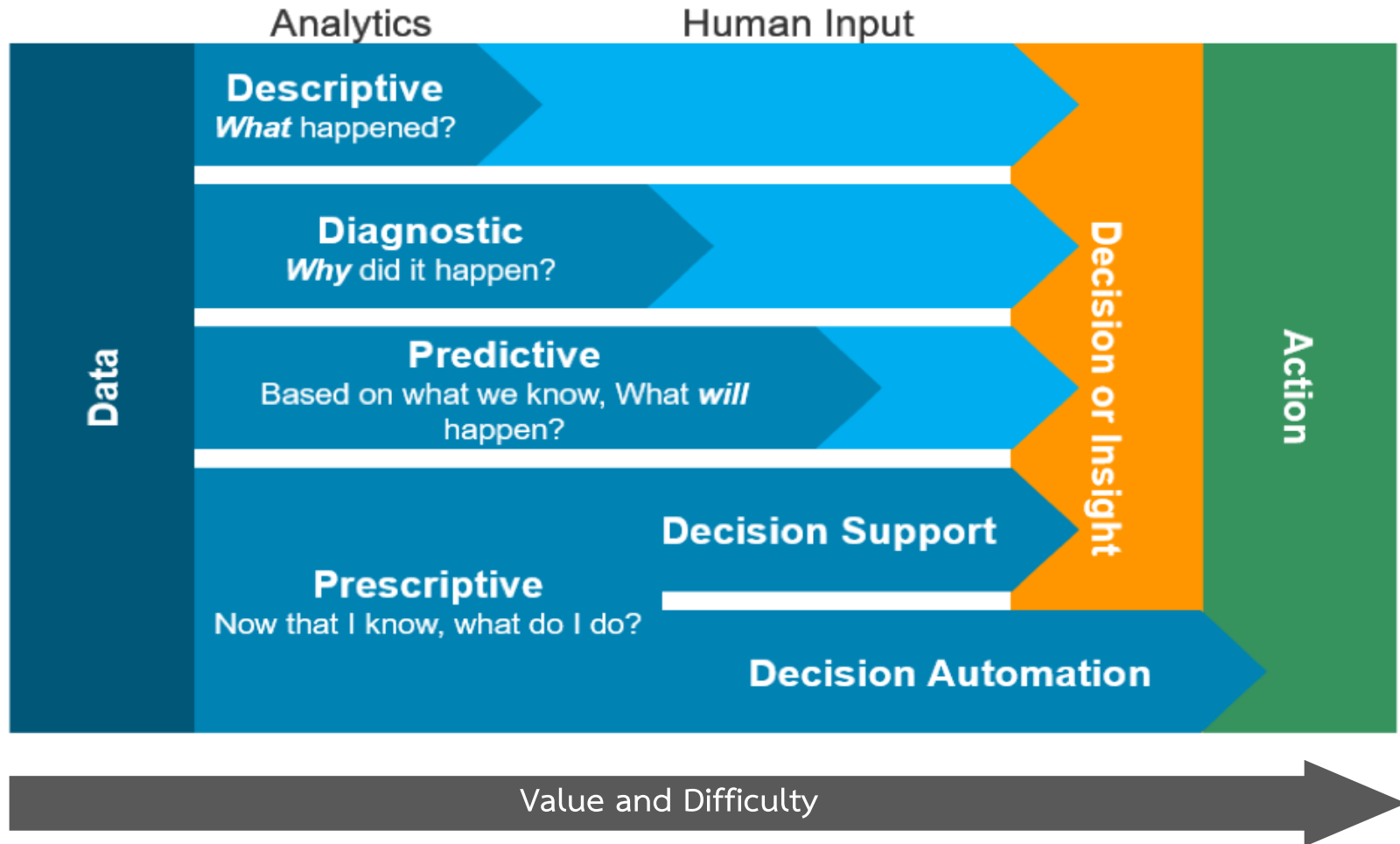
How to commute from here to Icon Siam at 8:00?

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



Descriptive Analytics

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.



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?

Descriptive Analytics

Techniques:

- 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

Predictive Analytics

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.



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?

Predictive Analytics

Techniques:

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

Prescriptive Analytics

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.



Prescriptive Analytics

Questions:

- What would 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?
- How 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?

Prescriptive Analytics

Techniques:

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

??? Analytics ???

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

When you have a question!

Finding Data

Analyzing

Finding Answer

Example: a small dataset

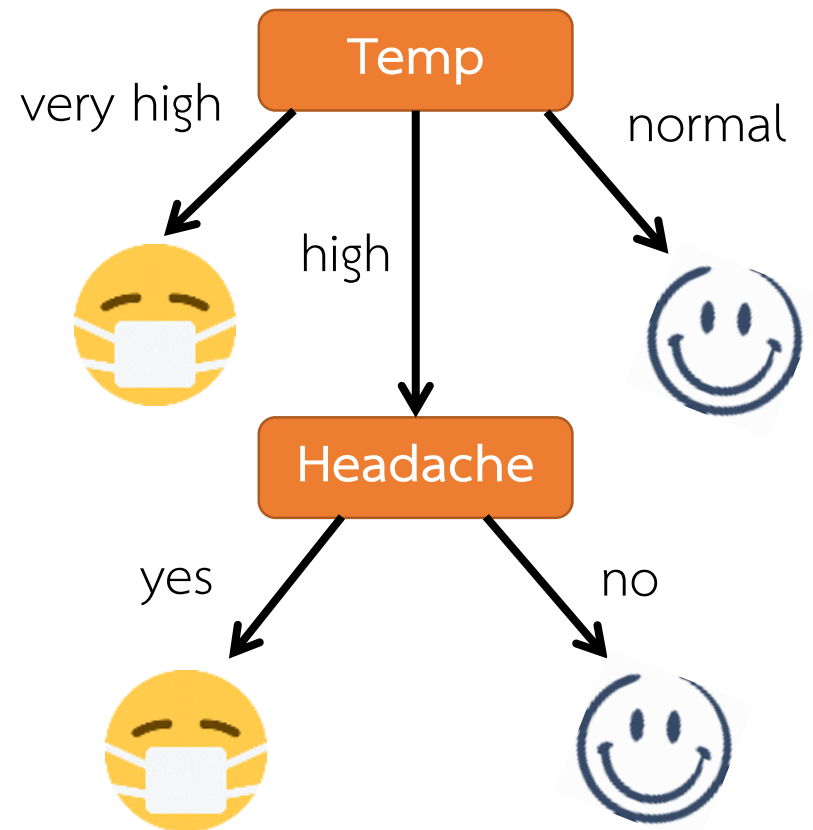
#	Body Temperature	Headache	Nausea	FLU?
1	high	yes	-	yes
2	very high	yes	yes	yes
3	normal	-	-	-
4	high	yes	yes	yes
5	high	-	yes	-
6	normal	yes	-	-
7	normal	-	yes	-

	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

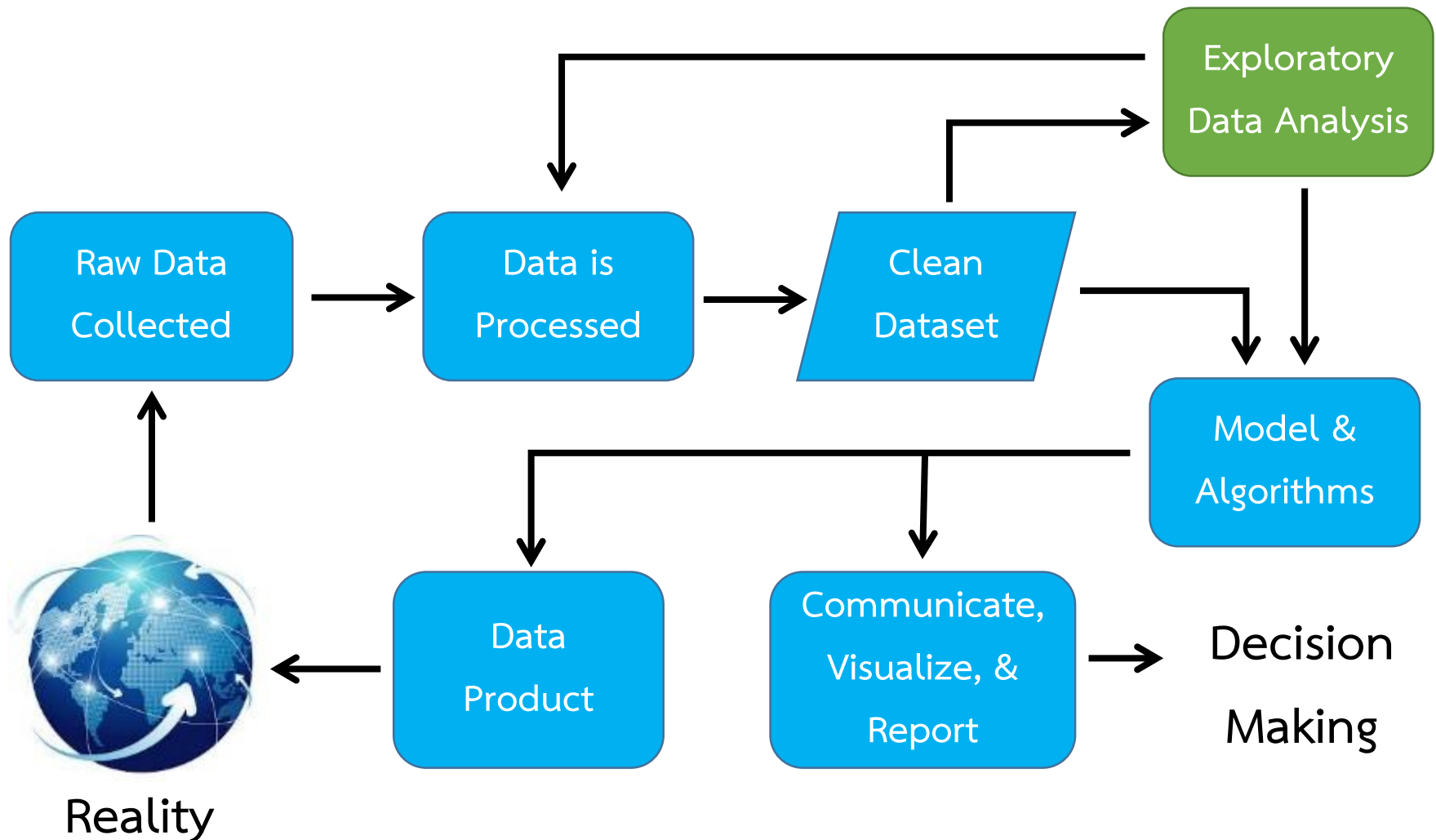
Much Data → Headache

(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.

Data Science Process



Weather Forecast

Shibuya, Tokyo, Japan

Wednesday 6:00 AM

Clear

 2 °C | °F

Precipitation: 0%

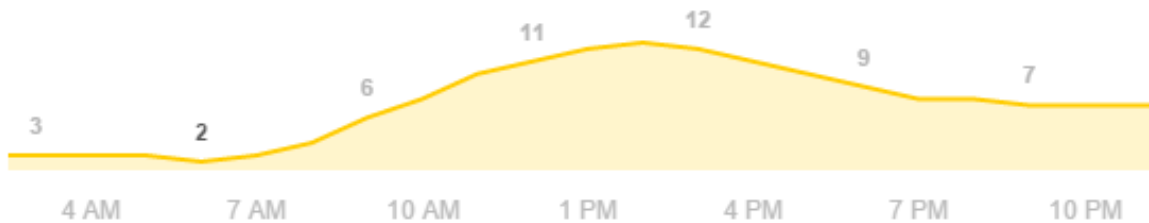
Humidity: 83%

Wind: 6 km/h

Temperature

Precipitation

Wind



Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed
							
13° 2°	13° 5°	11° 4°	13° 2°	13° 4°	13° 5°	13° 5°	14° 6°

Frequently Bought Together



Total price: To see our price, add these items to your cart. Why don't we show the price?

Add both to Cart

Add both to List

These items are shipped from and sold by different sellers. [Show details](#)

✓ **This item:** Sony KDL40W850D 40-Inch 1080p Smart LED TV (2010 Model)

✓ Cheetah Mounts APTMM2B TV Wall Mount for 20-75-Inch TVs Bundle with 10-foot Braided HDMI Cable and a... **\$24.95**

Customers Who Bought This Item Also Bought

Page 1 of 7

Ultra High Speed HDMI Cable 1080p Cable for HDTV, Blu-Ray, PS3 (6 feet)
★★★★☆ 340
\$12.20

VideoSecu ML531BE TV Wall Mount for most 22"-55" LED LCD Plasma Flat Screen Monitor up to 88 lb VESA 400x400...
★★★★☆ 17,844

Articulating Arm 32-50 inch TV LCD Monitor Wall Mount. Full Motion Tilt Swivel for 32" 36" 37"..."
★★★★☆ 170
\$33.99 **Prime**

VideoSecu TV Wall Mount Tilt Low Profile Ultra Slim Television Mount Bracket for Most 26"- 47" LED...
★★★★☆ 588
\$19.99 **Prime**

Sony XBR49X800D 49-Inch 4K Ultra HD TV (2016 Model)
★★★★☆ 174
\$798.00 **Prime**

Cheetah Mounts APTMM2B TV Wall Mount for 20-75-Inch TVs Bundle with 10-foot Braided...
★★★★☆ 14,193
#1 Best Seller in Electronics Mounts
\$24.95 **Prime**

Sony HTXT2 2.1 Channel Sound Base with Bluetooth
★★★★☆ 39
\$148.00 **Prime**

YouTube

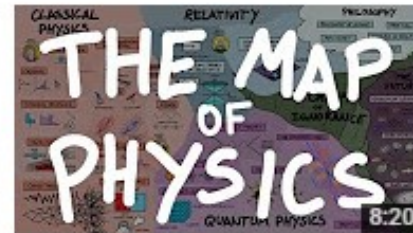
Recommended



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 จาก
ยุคเริ่มต้นถึงล่มสลาย
SFG Unicorn
38,487 views • 7 months ago



10 อันดับสุดยอดขุนศึกในยุคสาม
ก๊ก by CHERRYMAN
CHERRYMAN
96,747 views • 3 weeks ago



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

Gmail

Gmail - ☐ - 1-100 of 10,312

COMPOSE

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

Inbox	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	【家出少女を救う神待ち掲示板】	galen@ozdachs.com	家出少女を救う神待ちサ	Oct 28
Starred	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	BetterThanHCG (2)	galen@ozdachs.com	Traci says "It's BETTER	6:16 pm
Important	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	Tech	galen@ozdachs.com	System Update - Click t	5:05 pm
Sent Mail	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	for warranty experts	galen@ozdachs.com	60% OFF - If you would li	4:41 pm
Drafts	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	Easy Mole Removal	galen@ozdachs.com	Remove Moles and Ski	4:23 pm
All Mail	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	Brook	Vmax Pills Official	Site - 100% Guaranteed	4:22 pm
Spam (10,276)	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	GetAnyWoman	galen@ozdachs.com	I got a date this weeke	4:20 pm
Trash	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	Easy Mole Removal	galen@ozdachs.com	Remove Moles and Ski	4:13 pm
⌵ Circles <input type="button" value="Add"/>	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	LOTOTOjim	galen@ozdachs.com	..今月最後です.....口	4:07 pm
[imap]/Drafts	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	iPads Under One Hundred	galen@ozdachs.com	Absolutely, positively t	4:07 pm
galen@ozdachs.biz	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	Jessica Iwane	galen@ozdachs.com	28 days later this 51 ye	4:05 pm
galen@ozdachs.com	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	Painting Services	galen@ozdachs.com	House need painting? I	4:49 pm
GMail (about the s...	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	Jessica Iwane	galen@ozdachs.com	HAVE YOU SEEN THIS:	3:51 pm
Notes	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>	Cobra Health	galen@ozdachs.com	Cobra Health for nalen	3:50 pm
More -	<input type="checkbox"/> <input type="star"/> <input type="checkbox"/>				

Google Map

Google Maps interface showing route options from 'My location' to 'Ministry of Transport, 38 Ratchadamno'. The interface includes a blue header with navigation icons, a search bar, and a list of route options with estimated travel times and distances.

My location

Ministry of Transport, 38 Ratchadamno

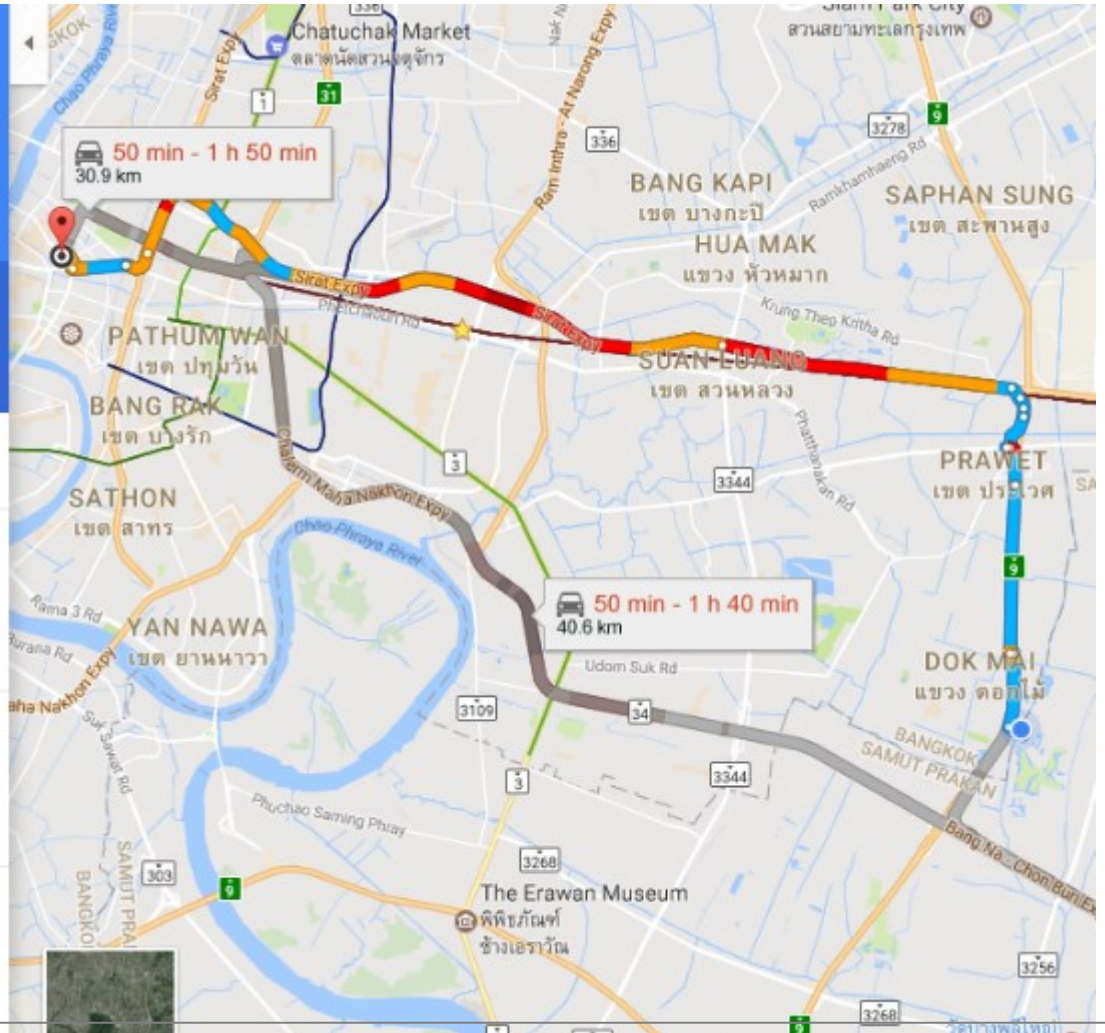
Arrive by 8:30 AM Mon, Mar 6

Send directions to your phone

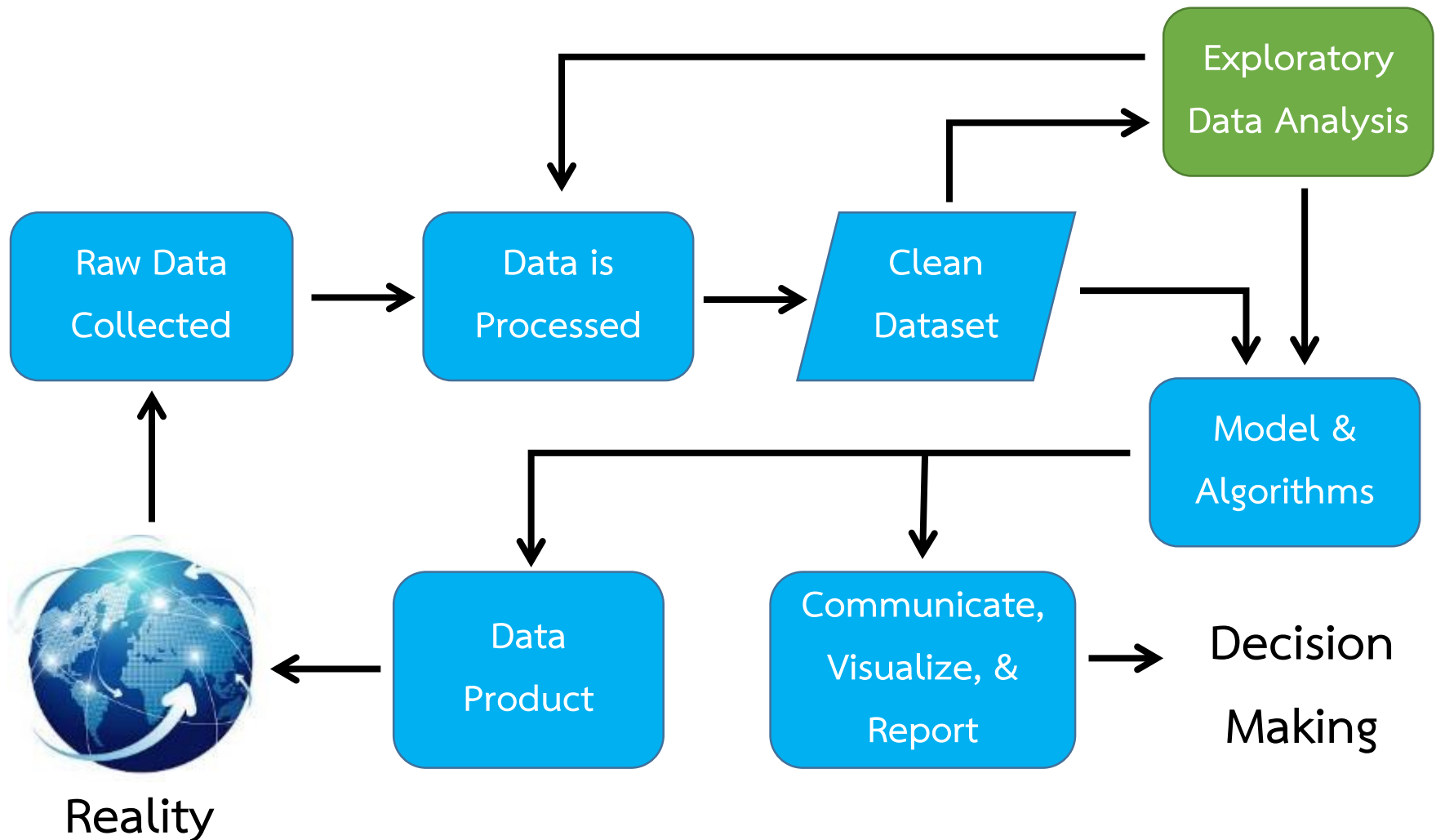
via ทางพิเศษศรีรัช typically 45 min - 1 h 40 min
Leave around 6:50 AM
32.0 km
DETAILS

via Route 7 typically 50 min - 1 h 50 min
Leave around 6:40 AM
30.9 km

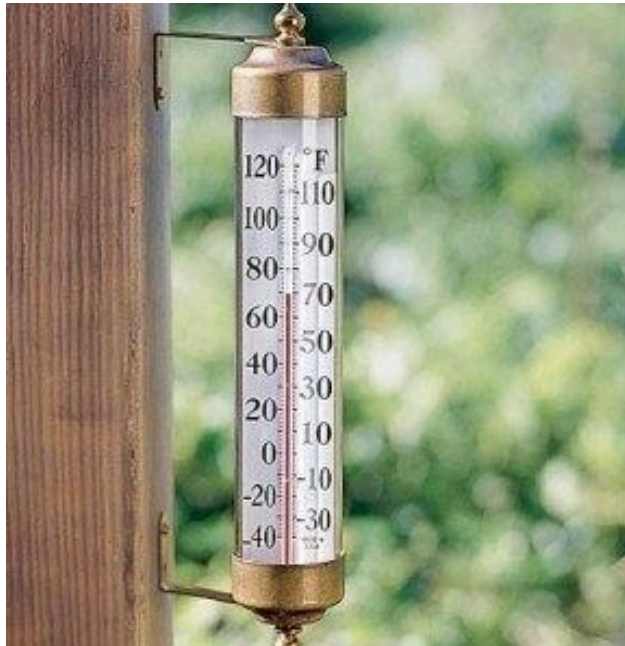
via Route 34 and ทางพิเศษเฉลิมมหานคร typically 50 min - 1 h 40 min
Leave around 6:50 AM
40.6 km



Data Science Process



Reality



A close-up photograph of a document with a repeating pattern of text, likely a form or a page from a book, showing multiple overlapping sheets. The text is printed in a small, uniform font and is arranged in a grid-like pattern across the pages. The pages are slightly offset, creating a sense of depth and repetition. The overall tone is monochromatic, with shades of gray and white.


#	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

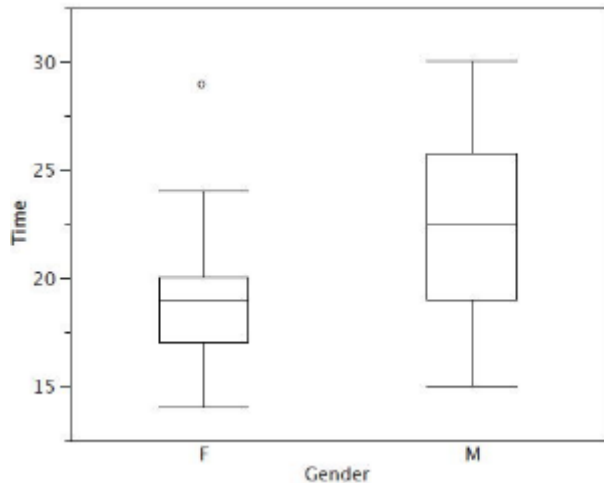
Data is Processed

- 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

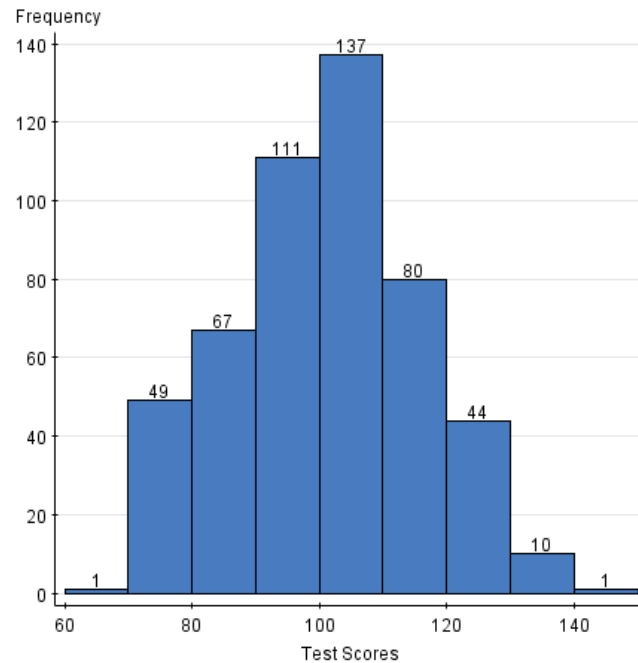
- Good Format
- Good Shape
- Ready for Data Analysis

Exploratory Data Analysis



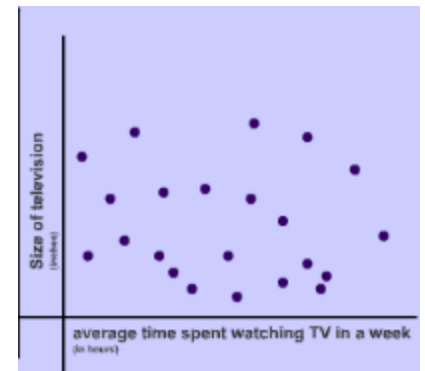
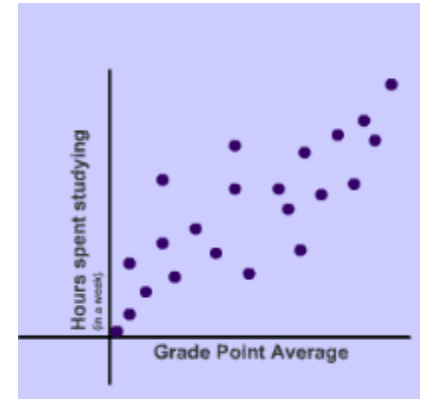
Box Plot

ความเร็วในการบอกข้อดีที่เห็น
ระหว่างผู้หญิงกับผู้ชาย



Histogram

คะแนนสอบ
กับจำนวนผู้ที่ได้คะแนน



Scatter plot

ดูความสัมพันธ์

Model & Algorithms

■ Algorithm

- a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

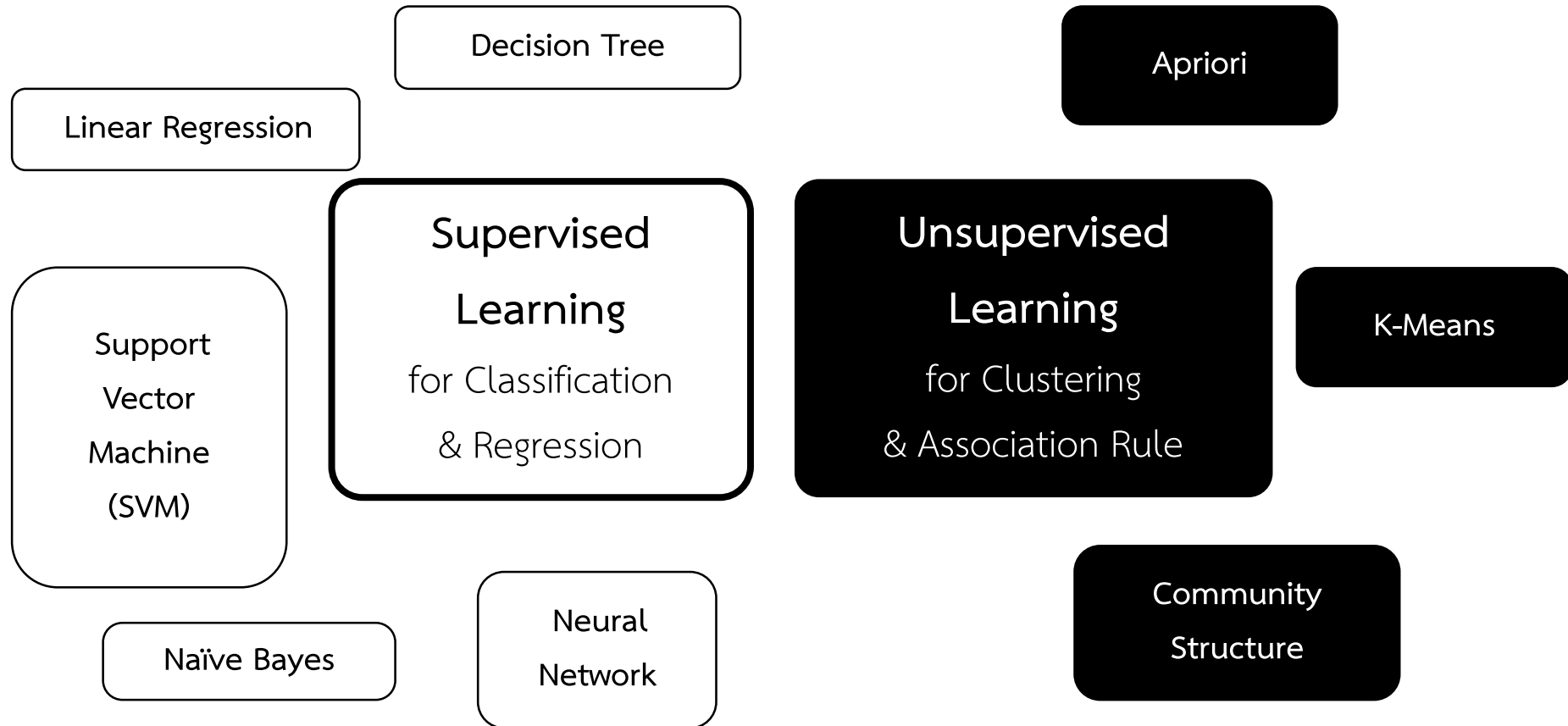
■ Model

- 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

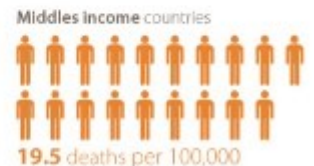
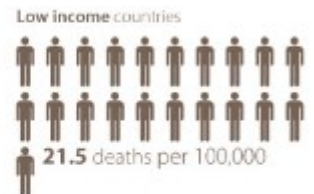
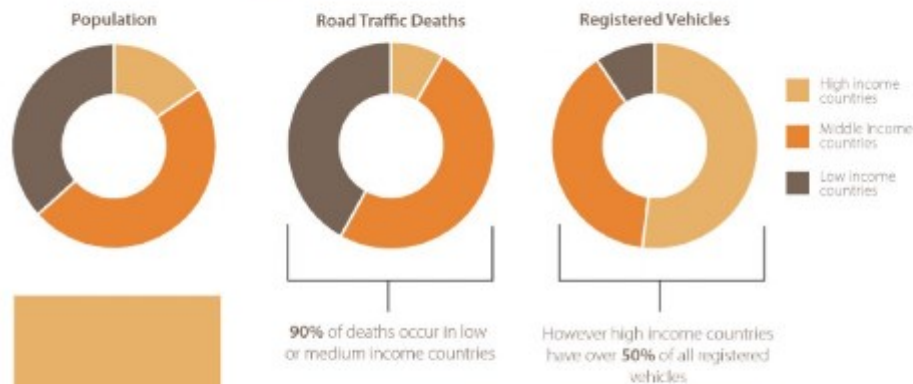
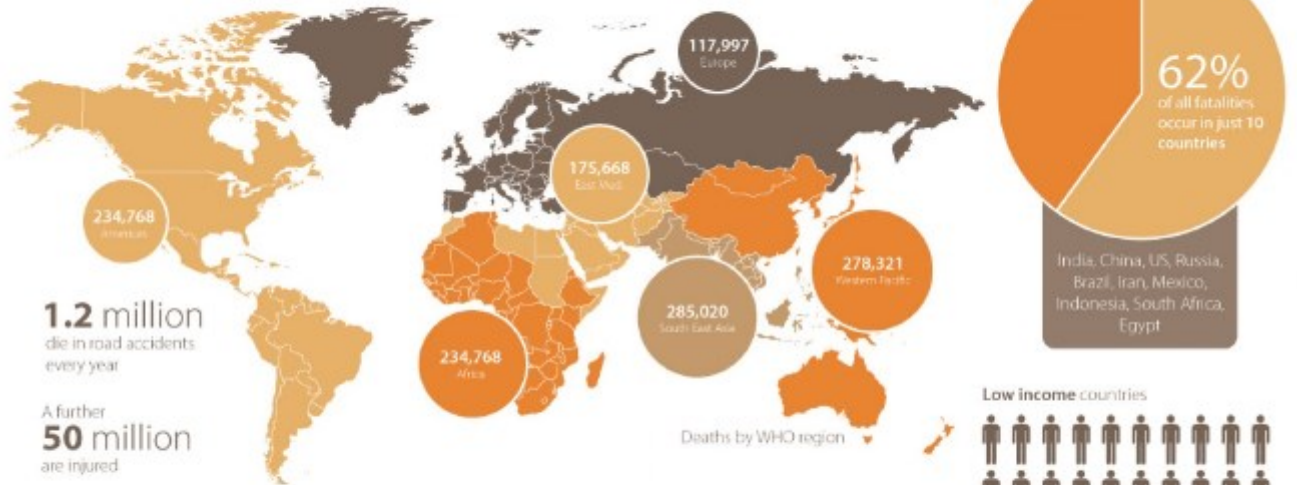
Model & Algorithms (Example)



Communicate, Visualize, & Report

Road Traffic Accidents: The Modern Killer

The Global Status Report released by WHO this year, confirms that road traffic injuries are still a big global health and development problem



Almost **50%** of those who die in traffic accidents are cyclists, pedestrians and motorcyclists

The Laws



Only **49%** of countries stipulate a **legal blood alcohol concentration** limit of less than 0.05g per decilitre



Only **57%** of countries **requires seatbelts** to be used by passengers



Only **40%** of countries have a comprehensive **helmet law** and require helmets to be of a specific standard

On the rise?

Road traffic accidents are predicted to rise to the **5th leading cause of death by 2030**, higher than AIDS, lung cancer and diabetes

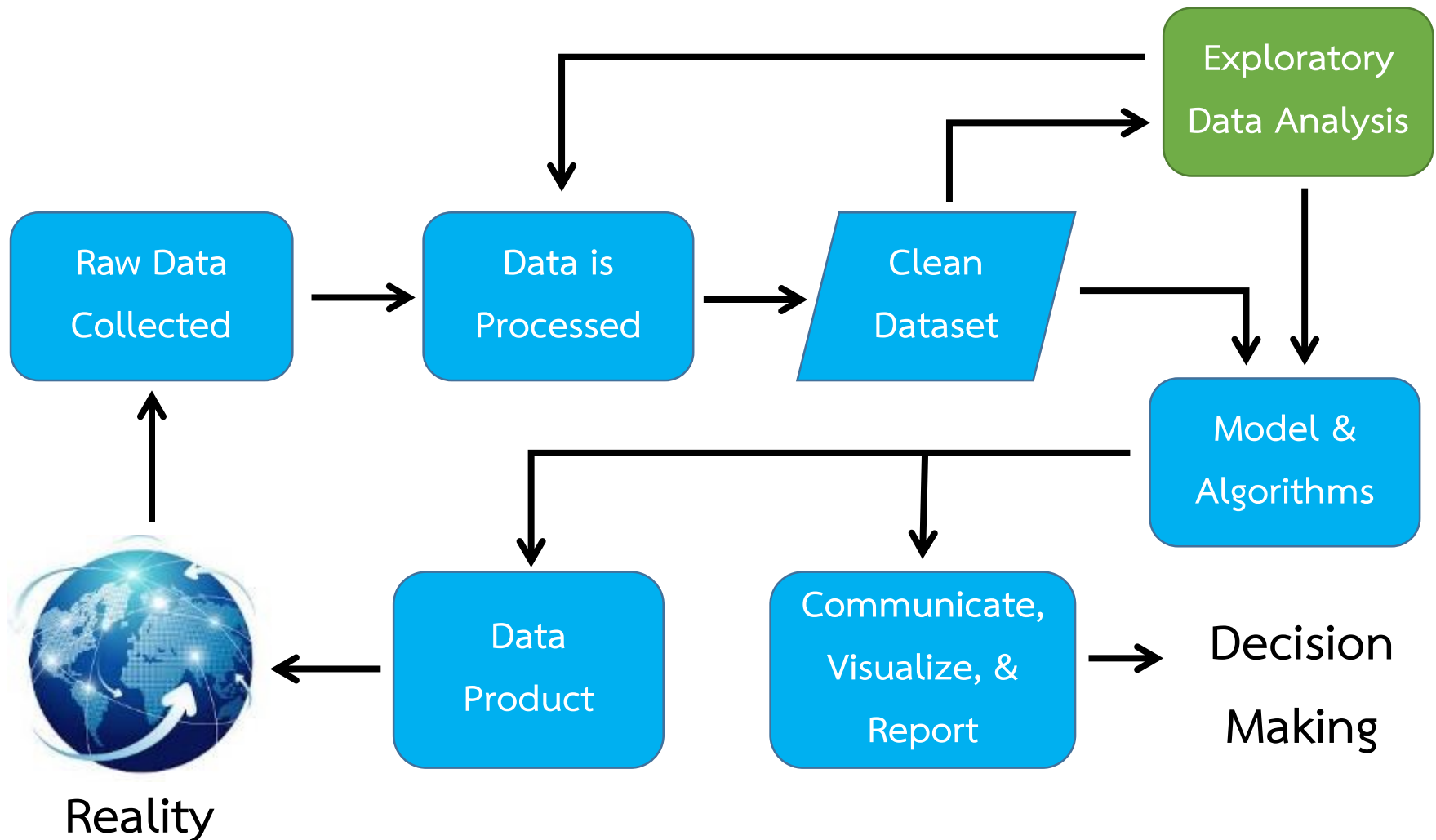


Car accidents are the **number 1 killer** for 15-29 year-olds

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
 - 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 21st 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

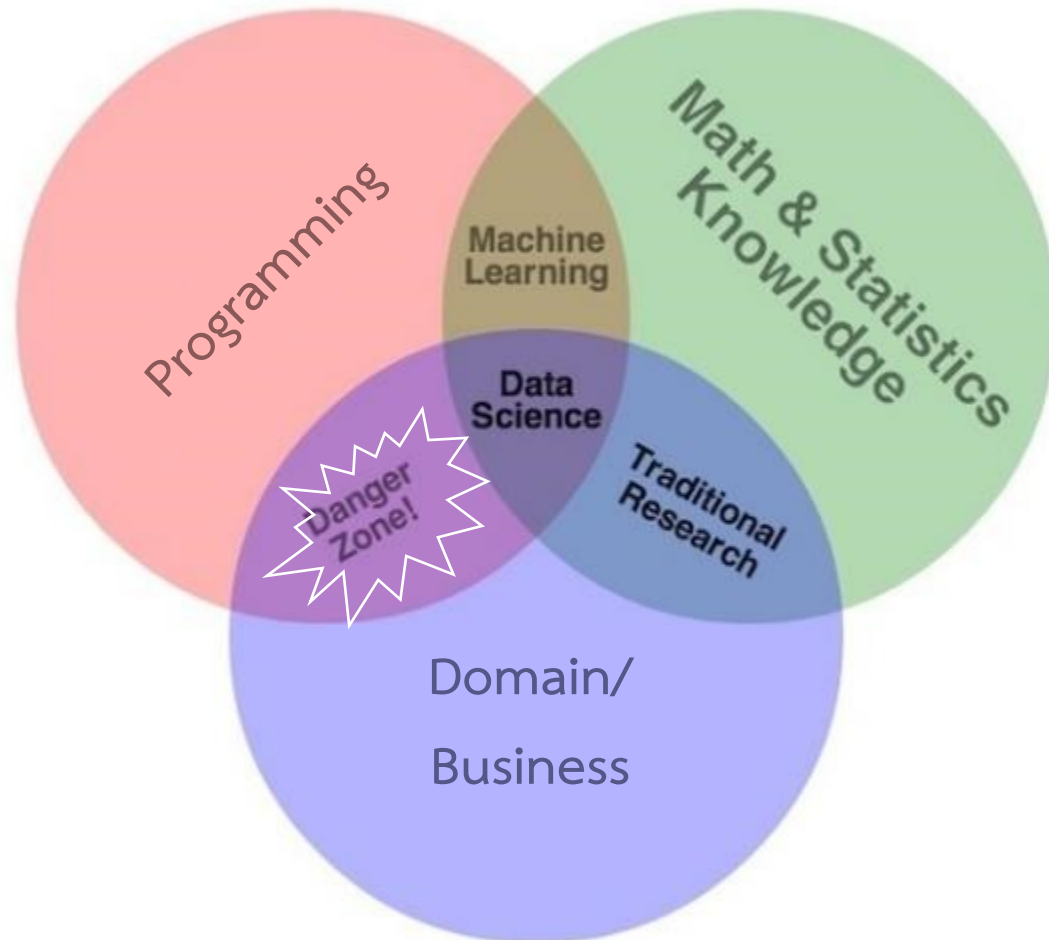
- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

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