# Data Science - IoT Data

**Provincial Electricity Authority co.,**

# Data Science for IoT - Agenda

- 09.00 - 10.30 **Introduction to Data Science**
- **Coffee Break**
- 10.45 - 12.00 **Data Modeling**
- **Lunch**
- 13.00 - 14.30 **Model Validation**
- **Coffee Break**
- 14.45 - 16.30 **Case Study**

# Paul

Wichit Sombat, PhD
wichit.s@ubu.ac.th

wichit2s

# K

Kriengsak Treeprapin, PhD
kriengsak.t@ubu.ac.th

# Tom

Phaichayon Kongchai, PhD
phaichayon.k@ubu.ac.th

# 1

# Intro to Data Science

แนะนำวิทยาการข้อมูล

> *Unify **statistics**[1], **data analysis**[2], **machine learning**[3] to scientifically understand data.*
>
> Wikipedia, 2018

Maths + Stats + Info Sci + Com Sci

# 1.1

## Introduction

ที่มา หลักการและเหตุผล

# Why???
# Understanding Data

ทำไมต้องเอาข้อมูลมาศึกษา?


เรียนรู้สาเหตุและป้องกันภัย

คาดการและปรับปรุงระบบการทำงานตามสถานะการ

# Data from <40,000[1] planes

ข้อมูลการบินจาก เครื่องบิน สี่หมื่น ลำ(ทั่วโลก)

[1]telegraph.co.uk

Air Traffic from social network posts (twitter)
ข้อมูลมาได้จากหลากหลายแหล่ง  twitter ~ 336M[1]

[1]statista.com

Tracking Influenza Via Twitter - 2013
โปรแกรม GermTracker[1] จากข้อมูล twitter

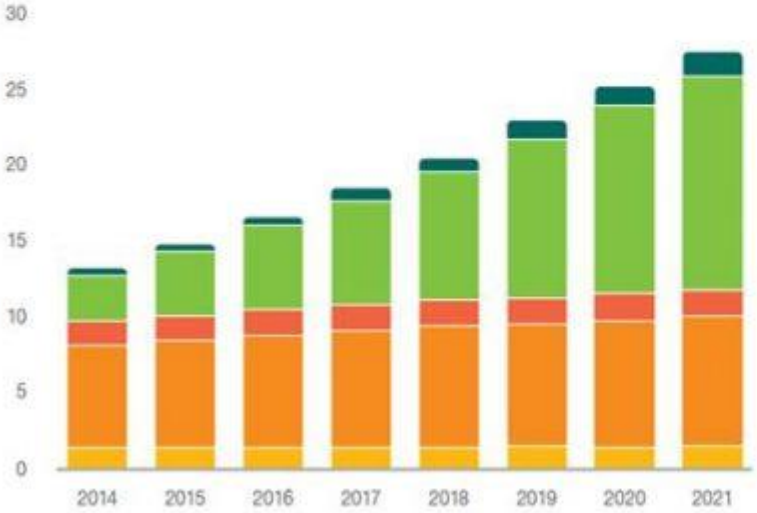
[1]University of Rochester


Social Media for UN Project Fortis for humanitarian aid plan

[1]Spark Summit 2017

# THE INTERNET OF THINGS

Connected devices (billions)



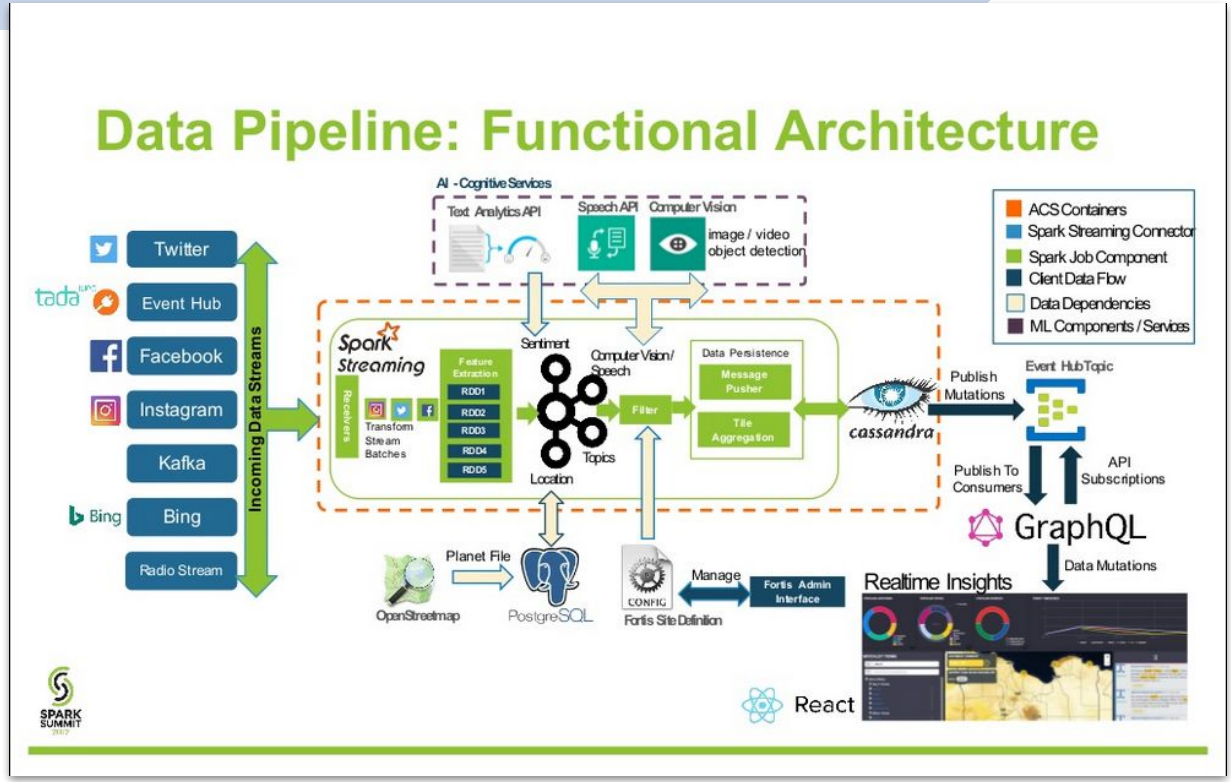

| | 15 billion | 28 billion | CAGR 2015–2021 |
|---|---|---|---|
| Cellular IoT | 0.4 | 1.5 | 27% |
| Non-cellular IoT | 4.2 | 14.2 | 22% |
| PC/laptop/tablet | 1.7 | 1.8 | 1% |
| Mobile phones | 7.1 | 8.6 | 3% |
| Fixed phones | 1.3 | 1.4 | 0% |
| | 2015 | 2021 | |

[1]ericsson.com


Data per IoT Device[1]

[1]Thingsboard

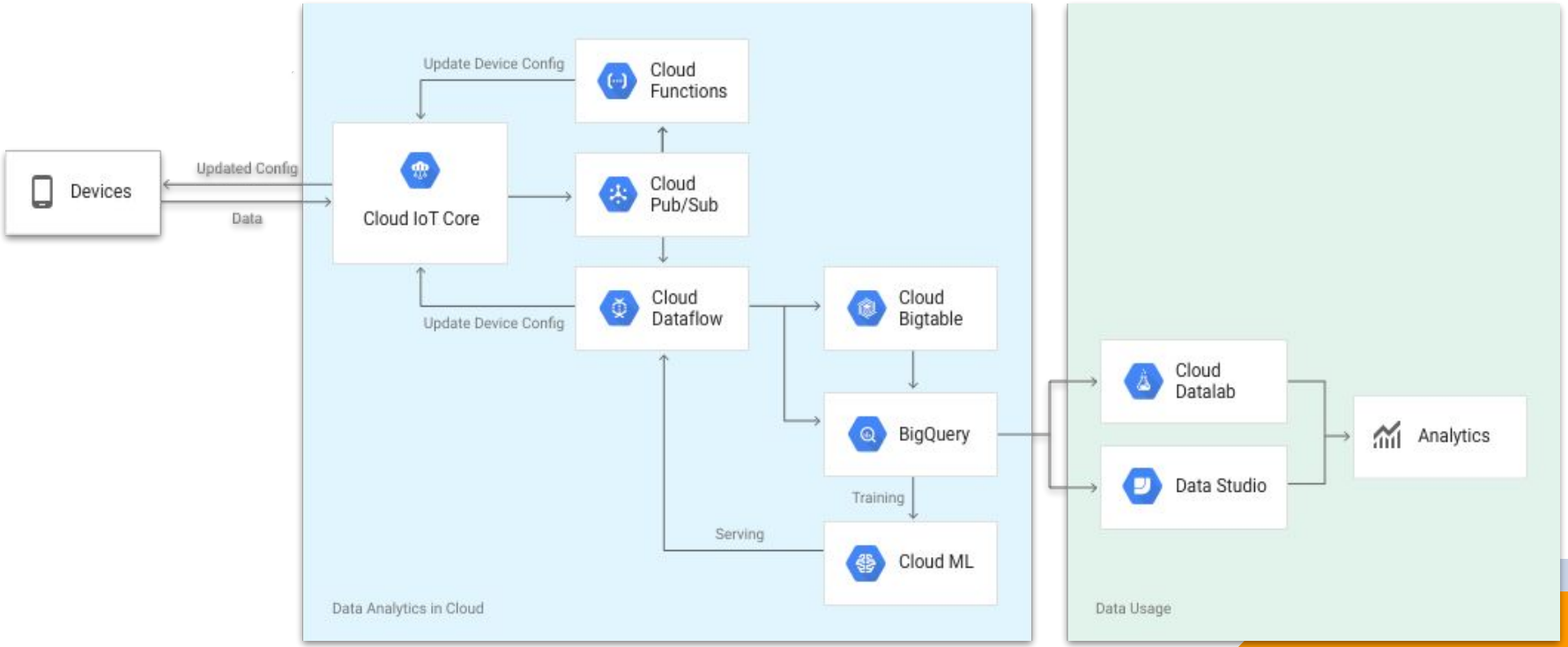
# Architecture for (Social) Data Analysis[fortis]



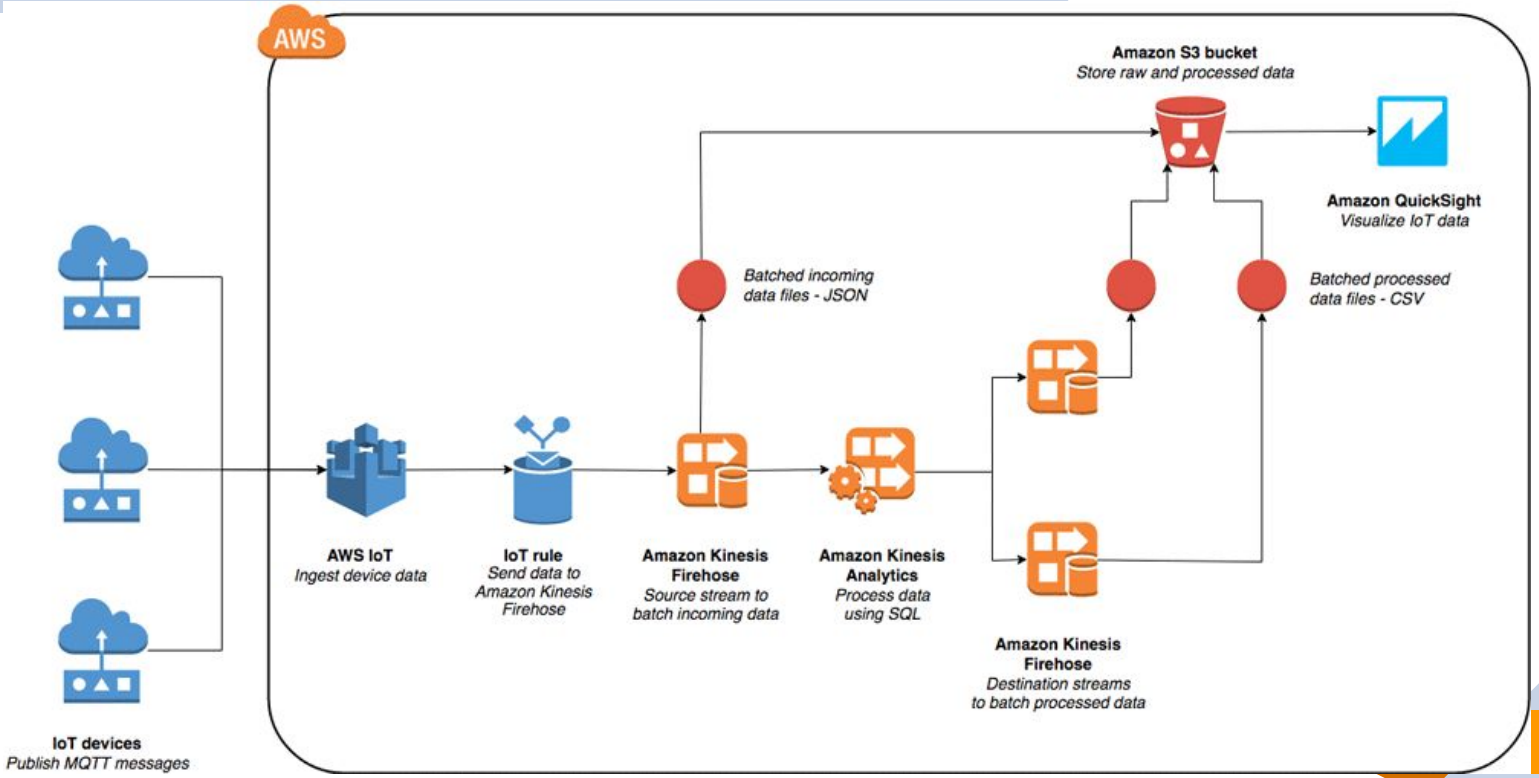

[1]Spark Summit 2017

# Architecture for Data Analysis

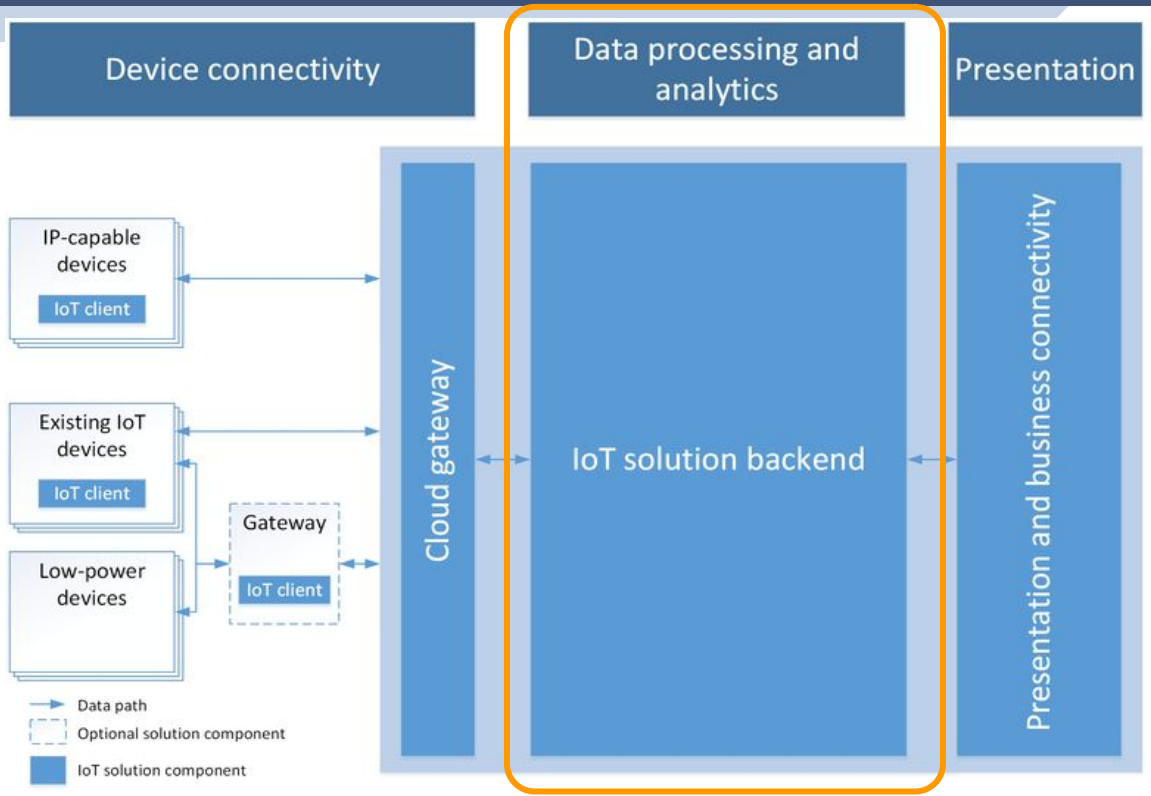
Google Cloud



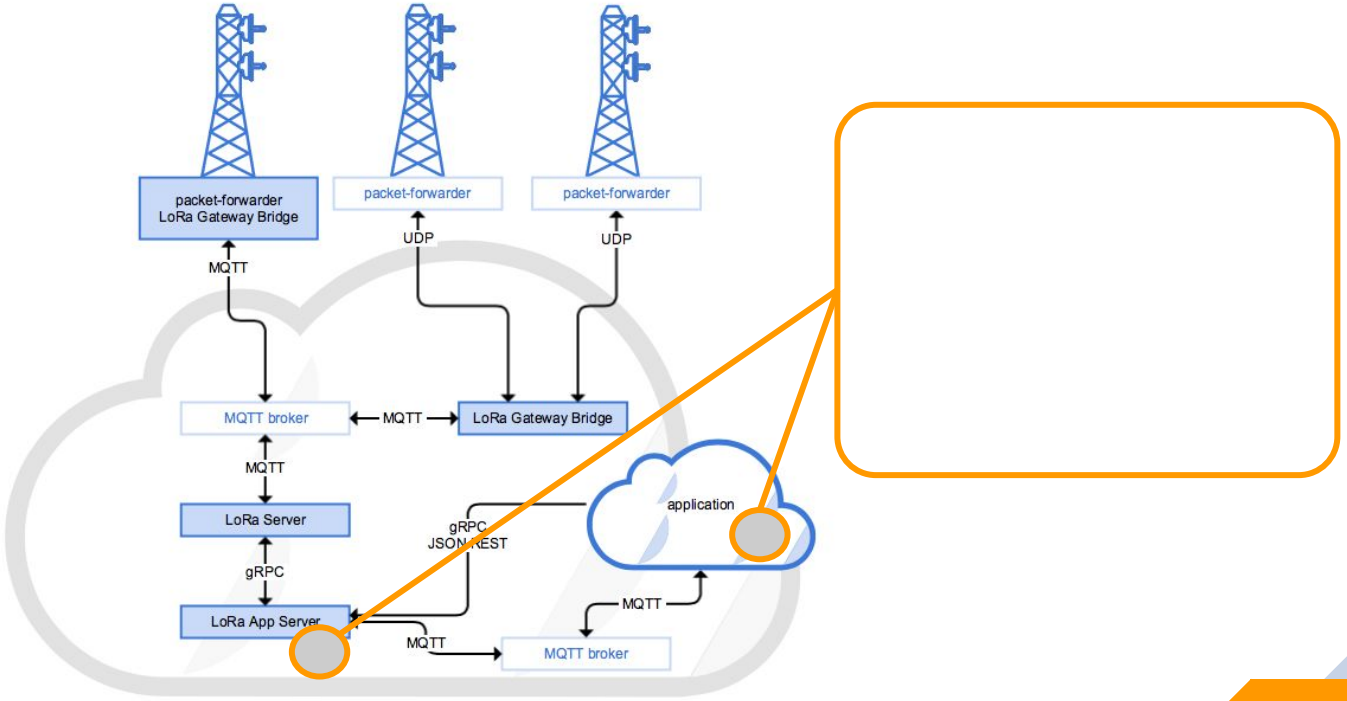
# Architecture for Data Analysis[Amazon]



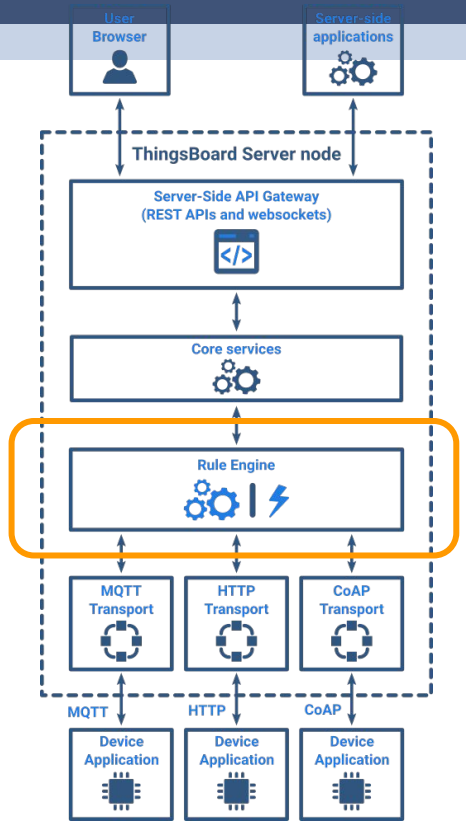
# Architecture for Data Analysis[Ms Azure]

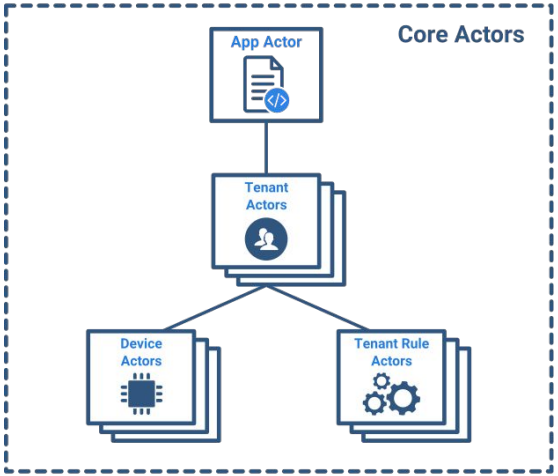


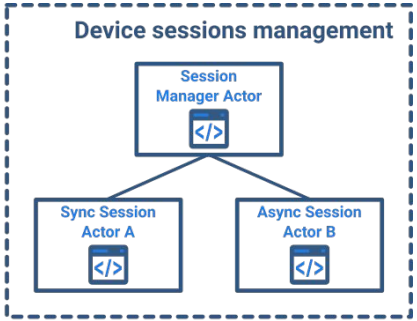
# IoT Architecture for ~~Data Analysis~~ Loraserver

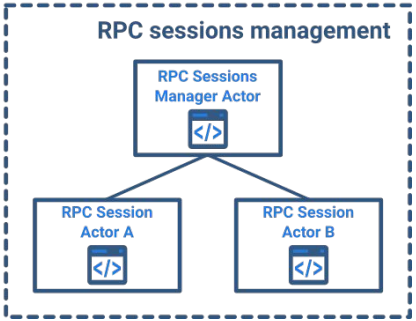


# Architecture for Data Analysis thingsboard

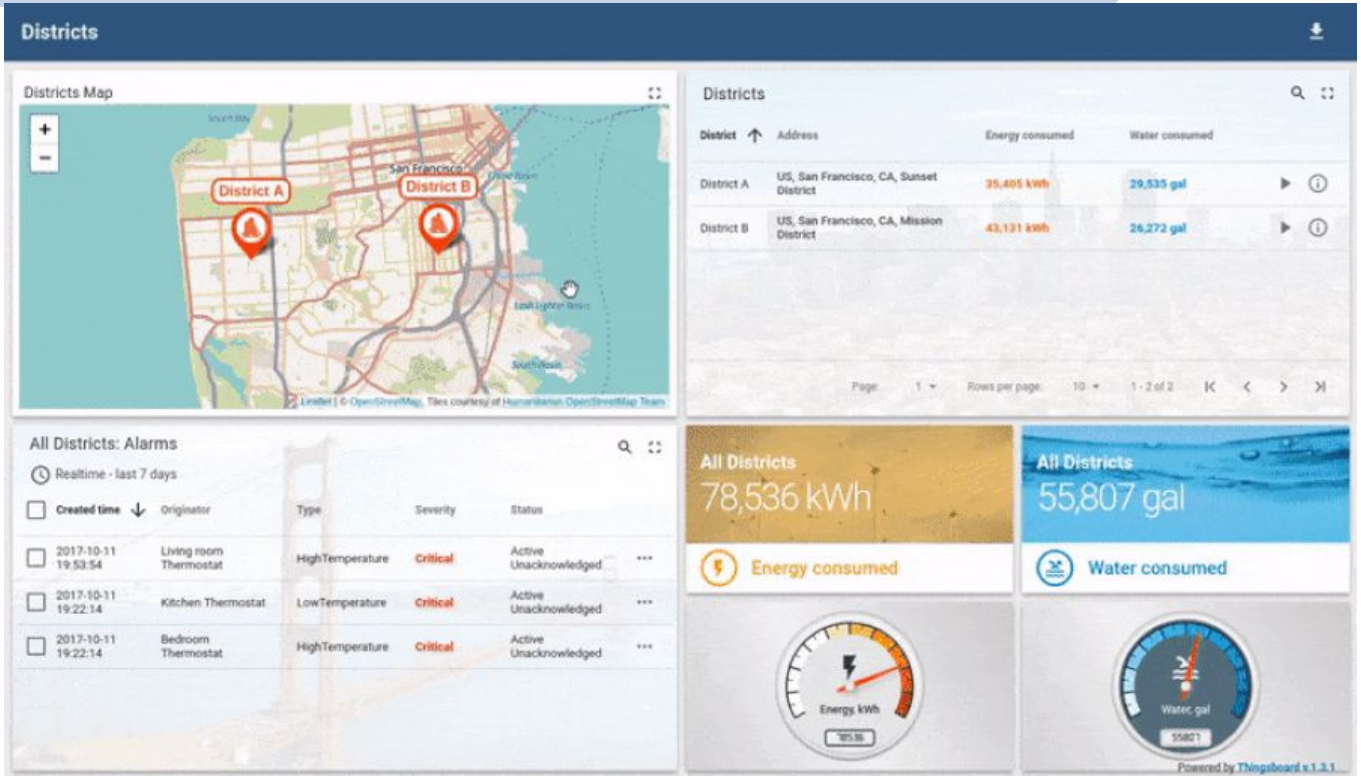


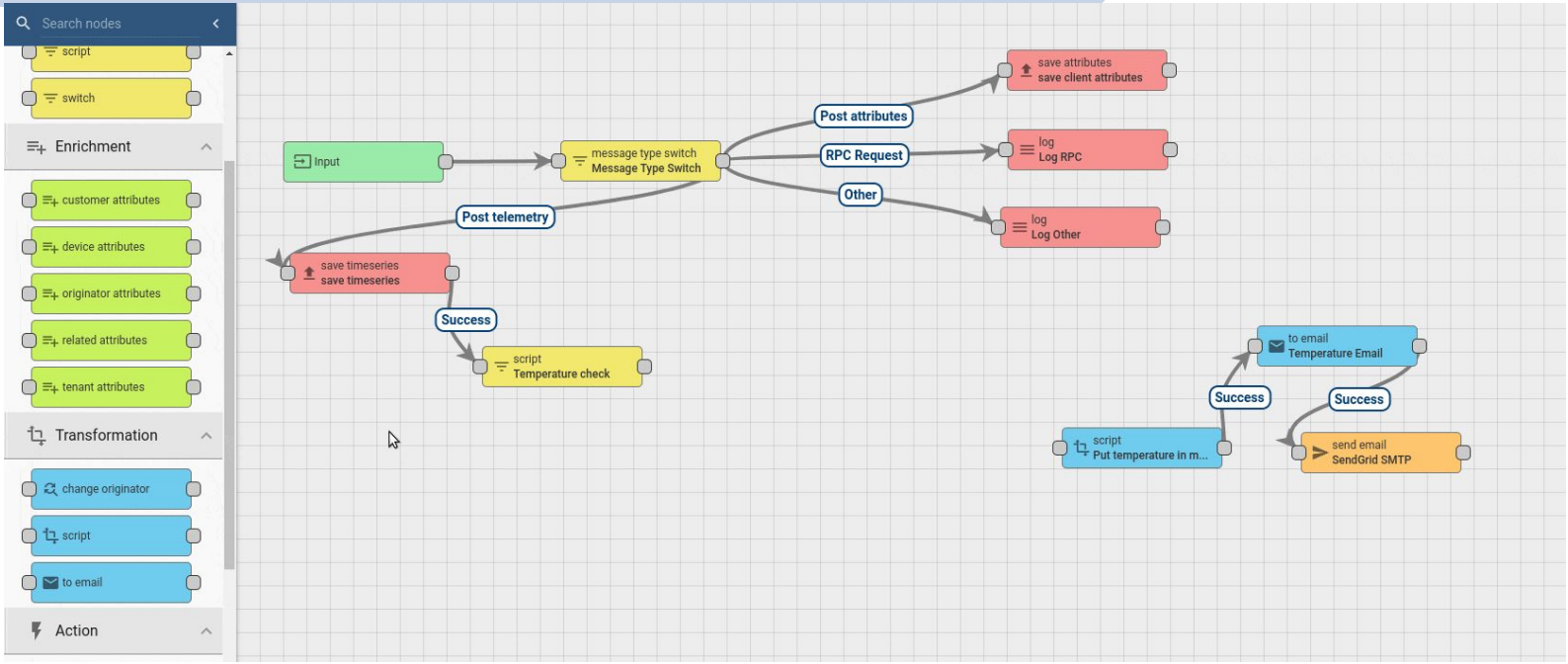



# Architecture for Data Analysis[thingsboard]

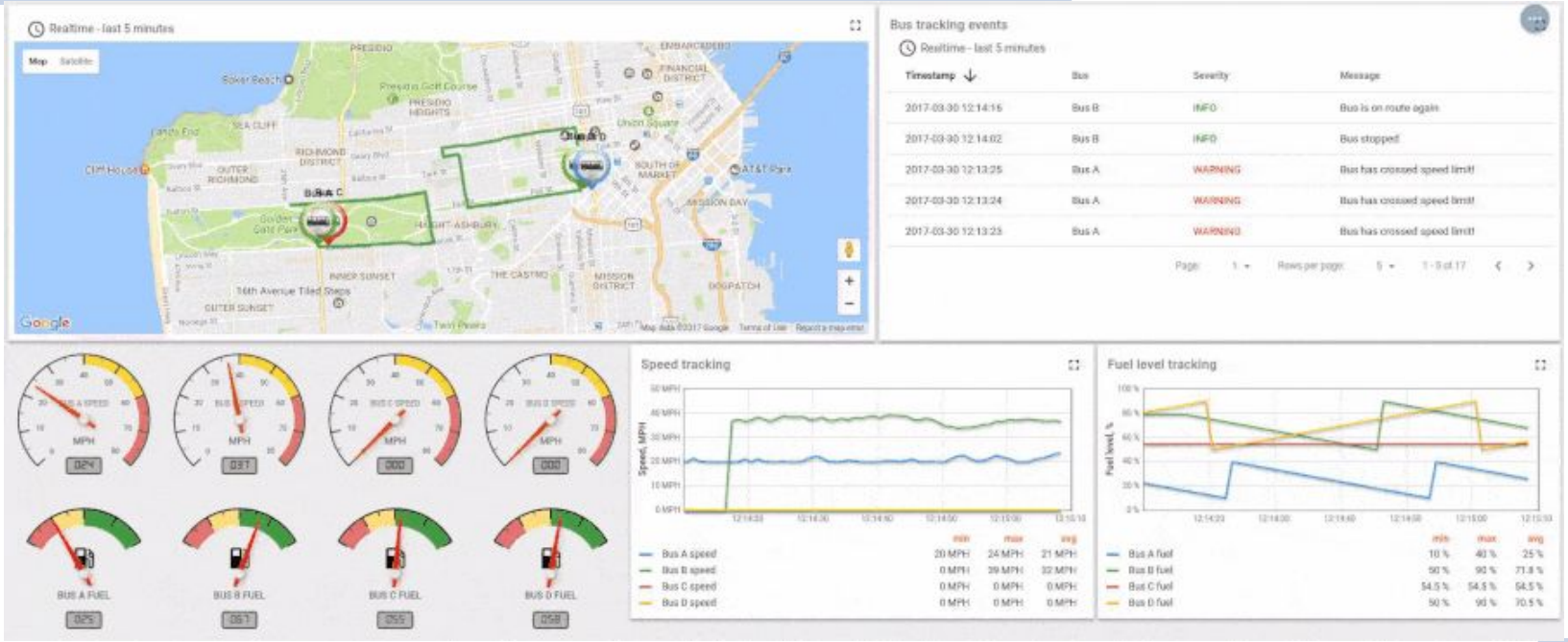


# Architecture for Data Analysis[thingsboard]



# Architecture for Data Analysis[thingsboard]

# Architecture for Data Analysis[thingsboard]



# 1.2

## Data Science Process

ขั้นตอนหลักของวิทยาการข้อมูล

# How Fortis UN generate heatmap?


Social Media for UN Project Fortis for humanitarian aid plan

[1]Spark Summit 2017

# Humanitarian aid plans are manually composed

[1]Spark Summit 2017


# Fortis UN project goals

- **Accelerate** the construction of aid **planning**
- **Improve** its data **accuracy**
- Provide **deeper insights** and **trends**
- **Real-time** analytics
- More intelligence and insight to enable better **forecasting**

[1]Spark Summit 2017

# Fortis UN project tasks

## Geospatial Requirements

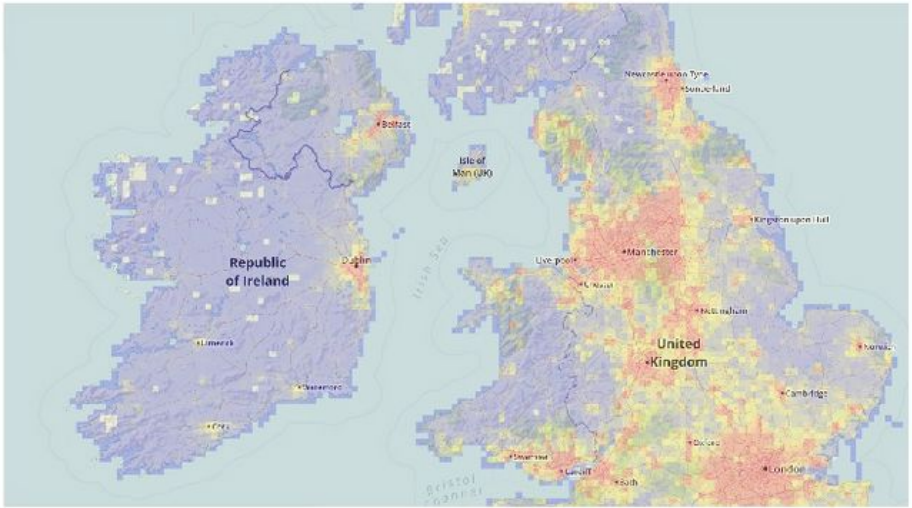

- Geotag mentioned places from conversations
- To query activities within a bounding box
- Filter activities based on topic(s)

[1]Spark Summit 2017

# Fortis UN project tasks

## Heatmap Generation

Aggregate detected places across geographic regions

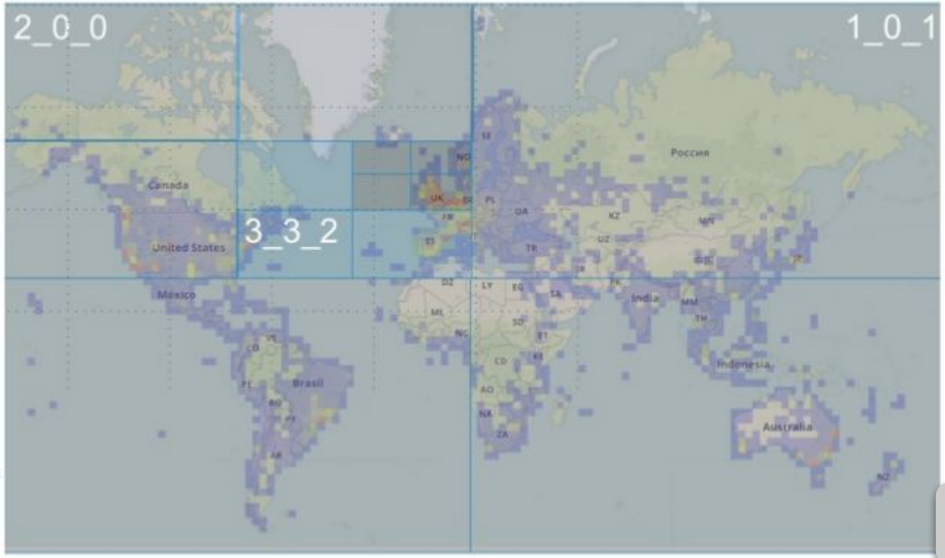



[1]Spark Summit 2017

# Fortis UN project tasks

## XYZ Tiles for summarization

- Divides world up into tiles.
- Each tile has four children at the next higher zoom level.
- Maps multiple layers to a single space dimension.




[1]Spark Summit 2017

# Fortis UN project tasks - Heatmap Spark

[1]Spark Summit 2017

```
final val MAX_ZOOM_LEVEL = 16
final val MIN_ZOOM_LEVEL = 5

def tile_id_mapper(location: (Double, Double)): List[(String, Int)] = {
  (for (zoom <- MIN_ZOOM_LEVEL to MAX_ZOOM_LEVEL)
    yield (TileUtils().tile_id_from_lat_long(location._1, location._2, zoom).tileId, 1)).toList
}
```

```
val locationListSamples = List[(Double, Double)]((30.294221,-97.7760937),
(30.294221,-97.7760937), (30.4007241,-97.7368647),(30.4007241,-97.7368647),....)
val locationRdd = sc.parallelize(locationListSamples)

locationRdd.flatMap(tile_id_mapper)
  .reduceByKey(_+_)
  .saveToCassandra("fortis", "tilesExample", SomeColumns("tileId", "count" append))
```

# Fortis UN project tasks

[1]Spark Summit 2017

## Aggregated Tile Schema

All inbound events are aggregated by
tile_x, tile_y, tile_z, period, source, publisher, topic, lang

**Aggregated Tile Data - Cassandra:** fortis_tiles_tbl

| tile_x | tile_y | tile_z | period | Source | publisher | Topic | Lang | feature_collection |
|---|---|---|---|---|---|---|---|---|
| 15 | 13346 | 15 | month-2017-06 | Twitter | Al Jazeera | isis | en | {"mention_count": 1213, "sentiments_avg": ["neg", ".78","pos": .02], "entities": [{"name": "basharassad", "mention_count": 627, "ref_id": "3256"}]}} |
| 231 | 3345 | 14 | month-2017-06 | Facebook | Times of Libya | Isis | en | {"mention_count": 453, "sentiments_avg": ["neg", ".78","pos": .02], "entities": [{"name": "basharassad", "mention_count": 124, "ref_id": "3256"}]}} |
| 76 | 98242 | 13 | month-2017-06 | Facebook | Times of Libya | isis | en | {"mention_count": 453, "sentiments_avg": ["neg", ".78","pos": .02], "entities": [{"name": "basharassad", "mention_count": 124, "ref_id": "3256"}]}} |

# Fortis UN project tasks

[1]Spark Summit 2017

## NLP Feature Extraction


**Event Details – Cassandra:** fortis_events_tbl

| event_id | source | Title | src_url | detected_features | Publisher | Topics | lang | message_body | Feature_collection |
|---|---|---|---|---|---|---|---|---|---|
| 851144530439090180 | twitter | Syria's Assad is 'inextricably connected' to Islamic State | https://twitter.com/AJEnglish/status/851144530439090180 | ["wof85678363"] | Al Jazeera | ["isis, "bombings"] | en | Far from fighting ISIS, Assad looked the other way when it set up many suicide bombings. An former Islamist from Hamah province who fought | { "sentiments_avg": ["neg", ".78","pos": .02], "entities": [{"name": "bashar assad", "ref_id": "3256"}]} |
| 861931221512847360 | twitter | Suicide Attack at Tank Brigade | https://twitter.com/AJEnglish/status/851144530439090180 | ["wof85678323"] | Libya Herald | ["suicide", "attack", "clashes"] | en | A suicide attack was reported at the 204 Tank Battalion headquarters which is in close proximity to 17 February Brigade camp. Clashes then broke out in Fuwayhat | { "sentiments_avg": ["neg", ".78","pos": .02]., "ref_id": "3256"}]} |

# Fortis UN project tasks

## Realtime Audio Streaming with Spark

- Speech Recognition APIs through radio broadcasts


[1]Spark Summit 2017

# 1.2

## Data Science Skills

ทักษะที่สำคัญของวิทยาการข้อมูล

# ทักษะที่จำเป็น ของวิทยาการข้อมูล

- **การวิเคราะห์ข้อมูล**
  - Data Mining, Machine Learning, NLP, NN, etc.
- **การเขียนคำสั่งโปรแกรม**
  - Python, R, Scala, JavaScript, SAS, etc.
- **การจัดการข้อมูลและนำเสนอ**
  - Data Visualization, Infographics, Plots, etc.

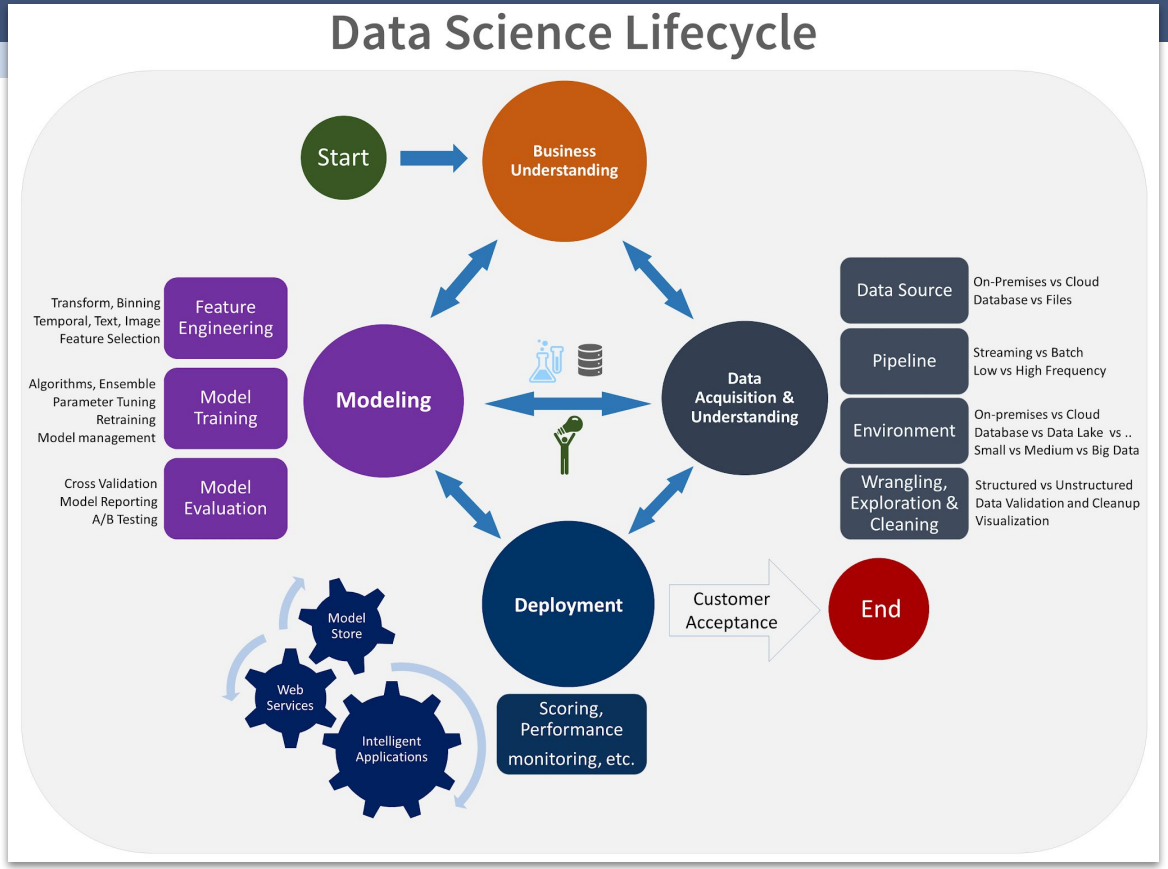
[1]Spark Summit 2017

# 1.3

## Data Science Process

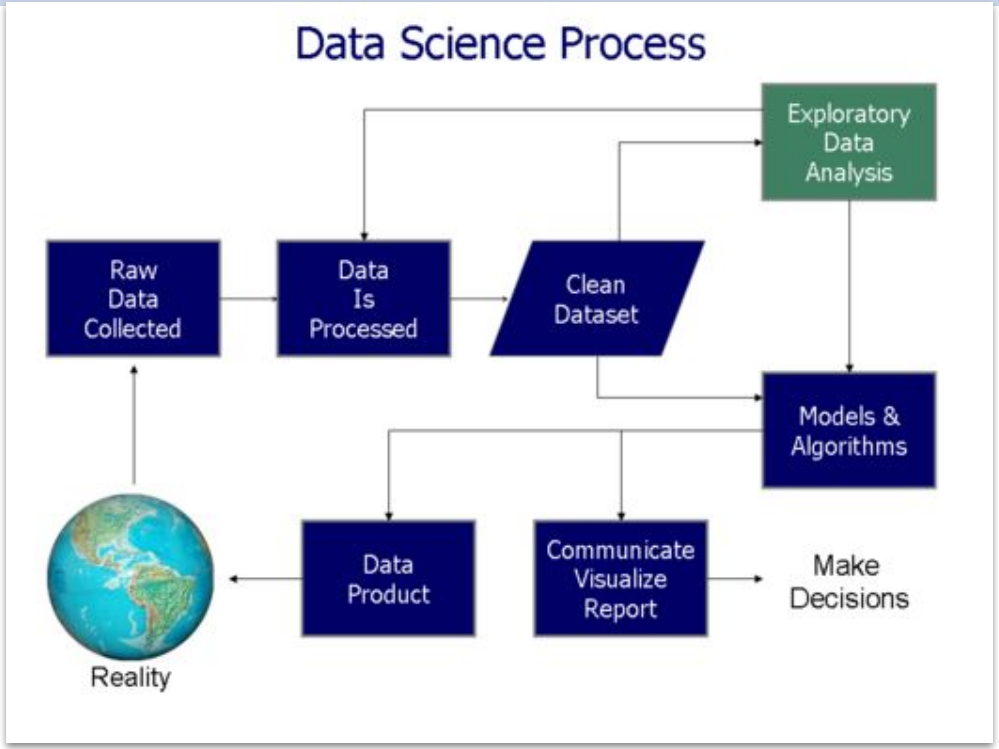
ขั้นตอนของวิทยาการข้อมูล

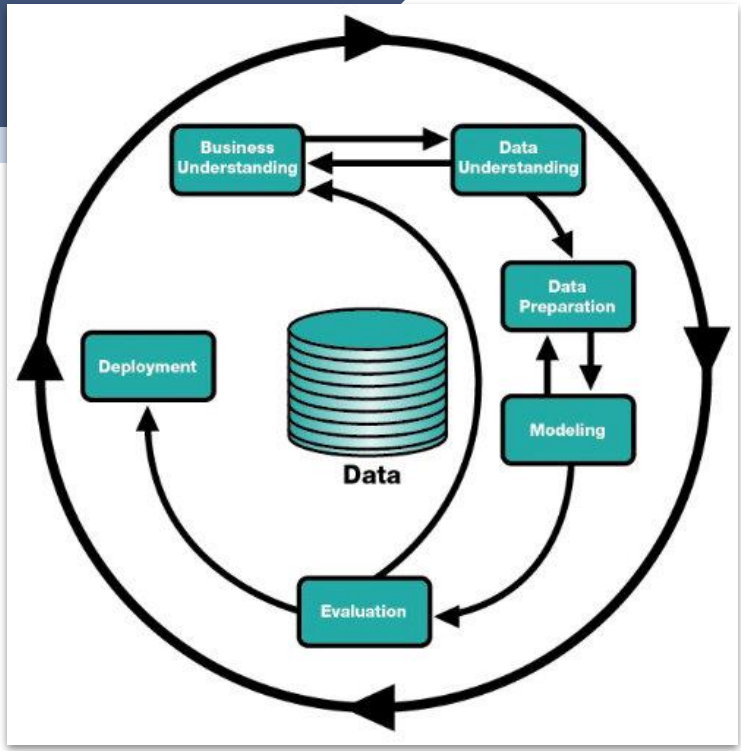
# Data Science Lifecycle[1]




[1]Microsoft

# Data Science Process[1,2]



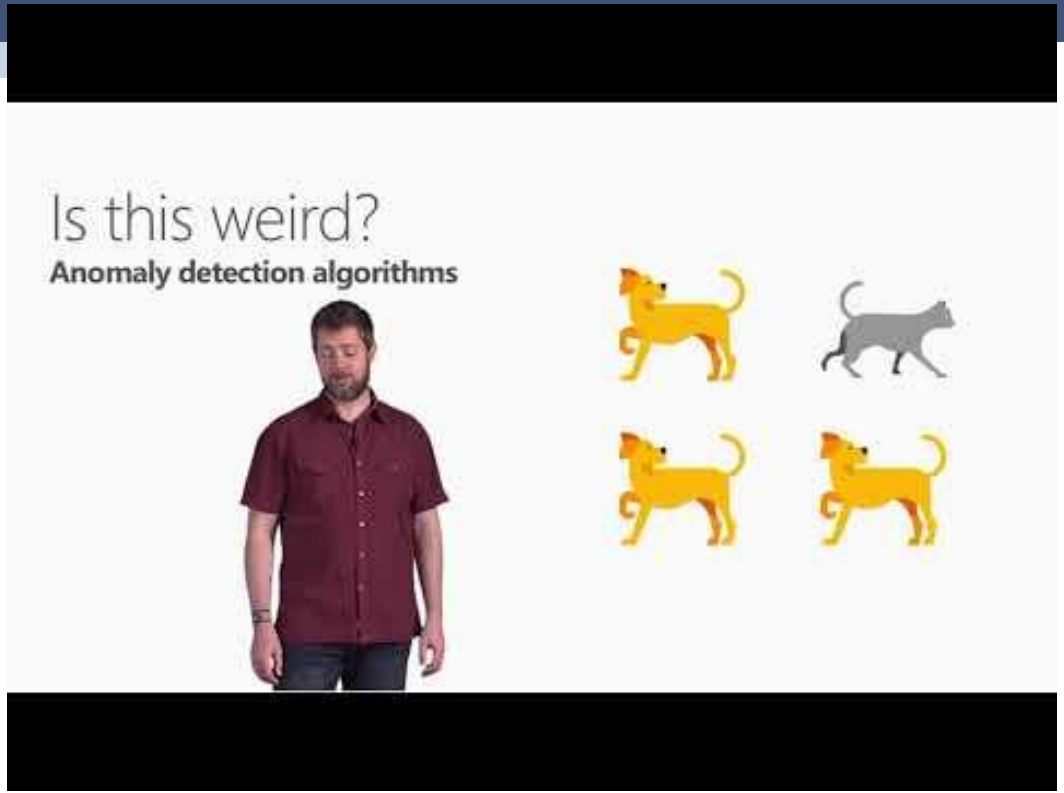




[1]kdnugget

# ขั้นตอนหลักของวิทยาการข้อมูล

1. **ตั้งคำถาม** - goal?  What to predict or estimate?
2. **รวบรวมข้อมูล** - นำเข้าข้อมูล,  ทำความสะอาดข้อมูล
3. **วิเคราะห์ข้อมูล** - plot, anomalies, patterns
4. **พัฒนาต้นแบบ**  สร้าง, ทดสอบ - build, fit, validate
5. **นำเสนอผลลัพธ์**  นำต้นแบบไปใช้งาน

# วิทยาการข้อมูล ตอบคำถามอะไรได้บ้าง?




[1]Daniel Yarmoluk 2018

# วิทยาการข้อมูล ตอบคำถามอะไรได้บ้าง?

1. Is this **A** or **B**? (classification)
2. Is this **weird**? (anomaly detection)
3. How much or **how many**? (regression)
4. How is this **organized**? (clustering)
5. **What** should I do **next**? (reinforcement learning)

# 1.4

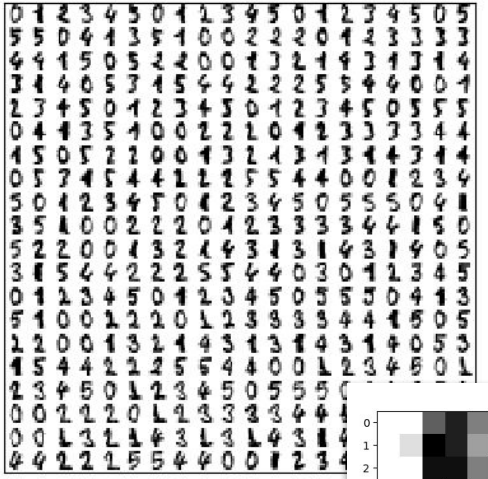
## Machine Learning

หลักการเรียนรู้ของเครื่อง

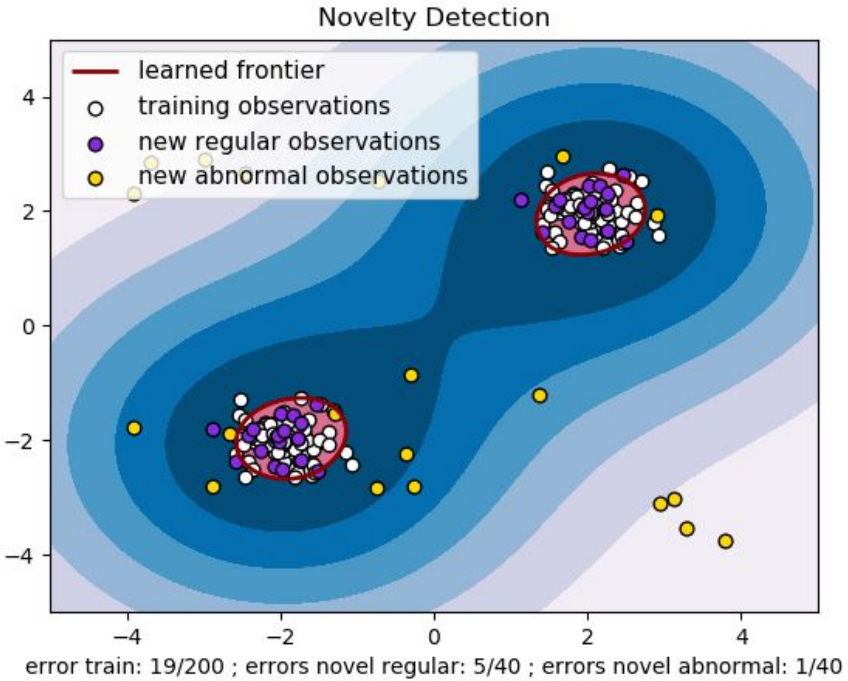
# Classifications

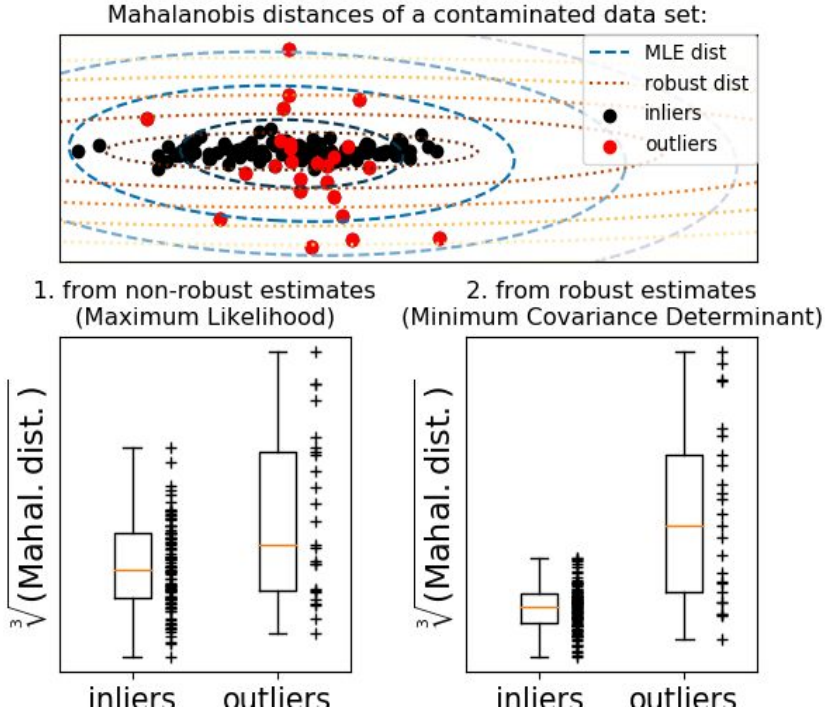
ຂ້ອຍກໍຮັກເລື່ອງນີ້.

A selection from the 64-dimensional digits dataset

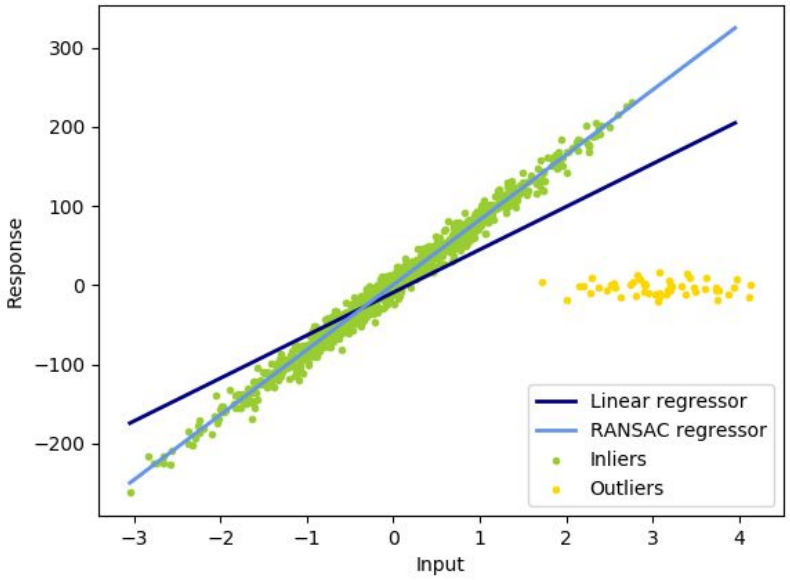


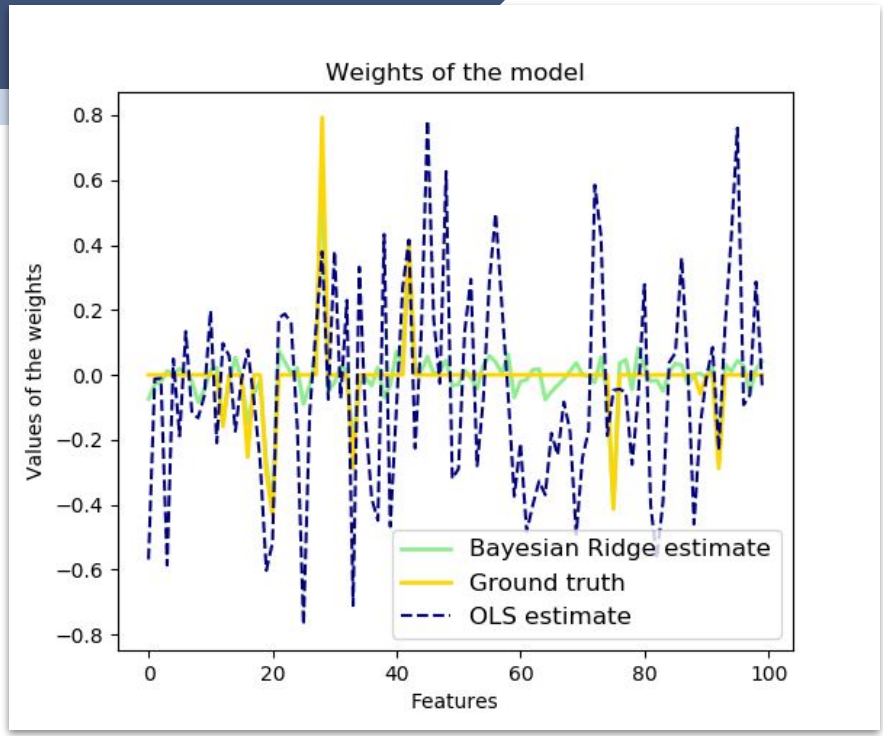
# Anomaly Detection



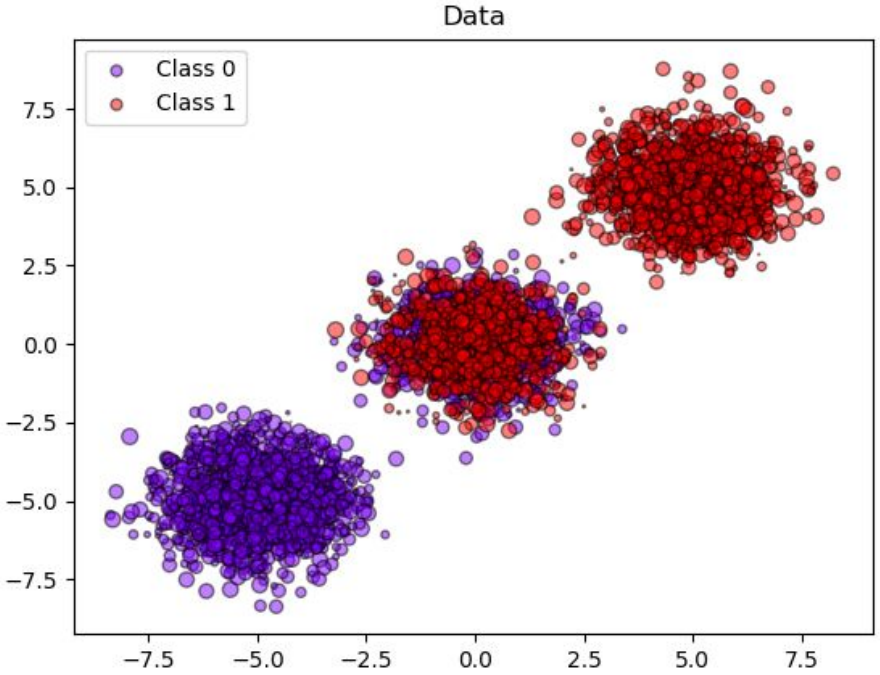



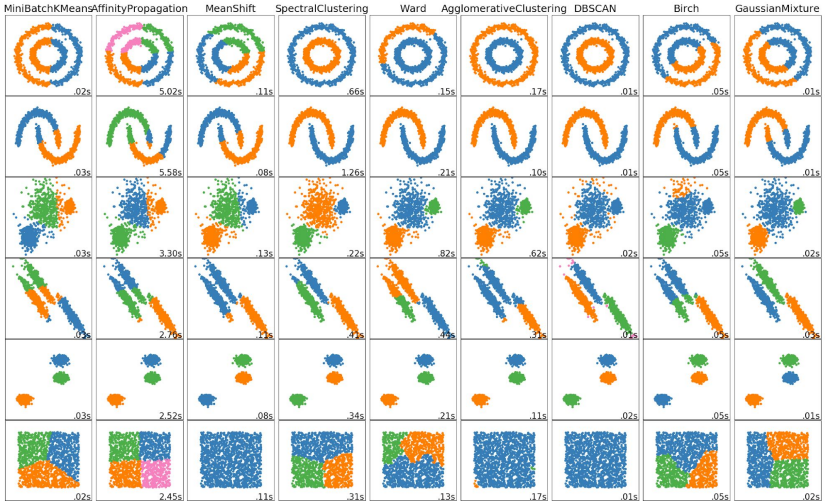
# Regression


Response
Input
Linear regressor
RANSAC regressor
Inliers
Outliers

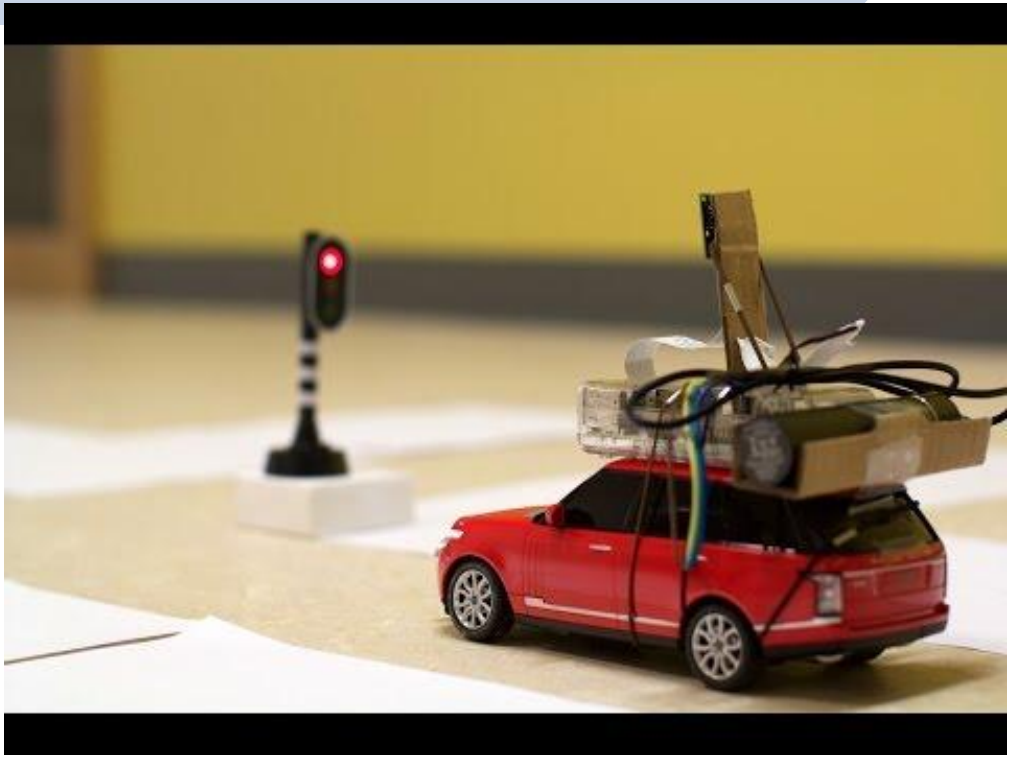

Weights of the model
Values of the weights
Features
Bayesian Ridge estimate
Ground truth
OLS estimate

# Clustering


Data
Class 0
Class 1
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture

# Reinforcement Learning

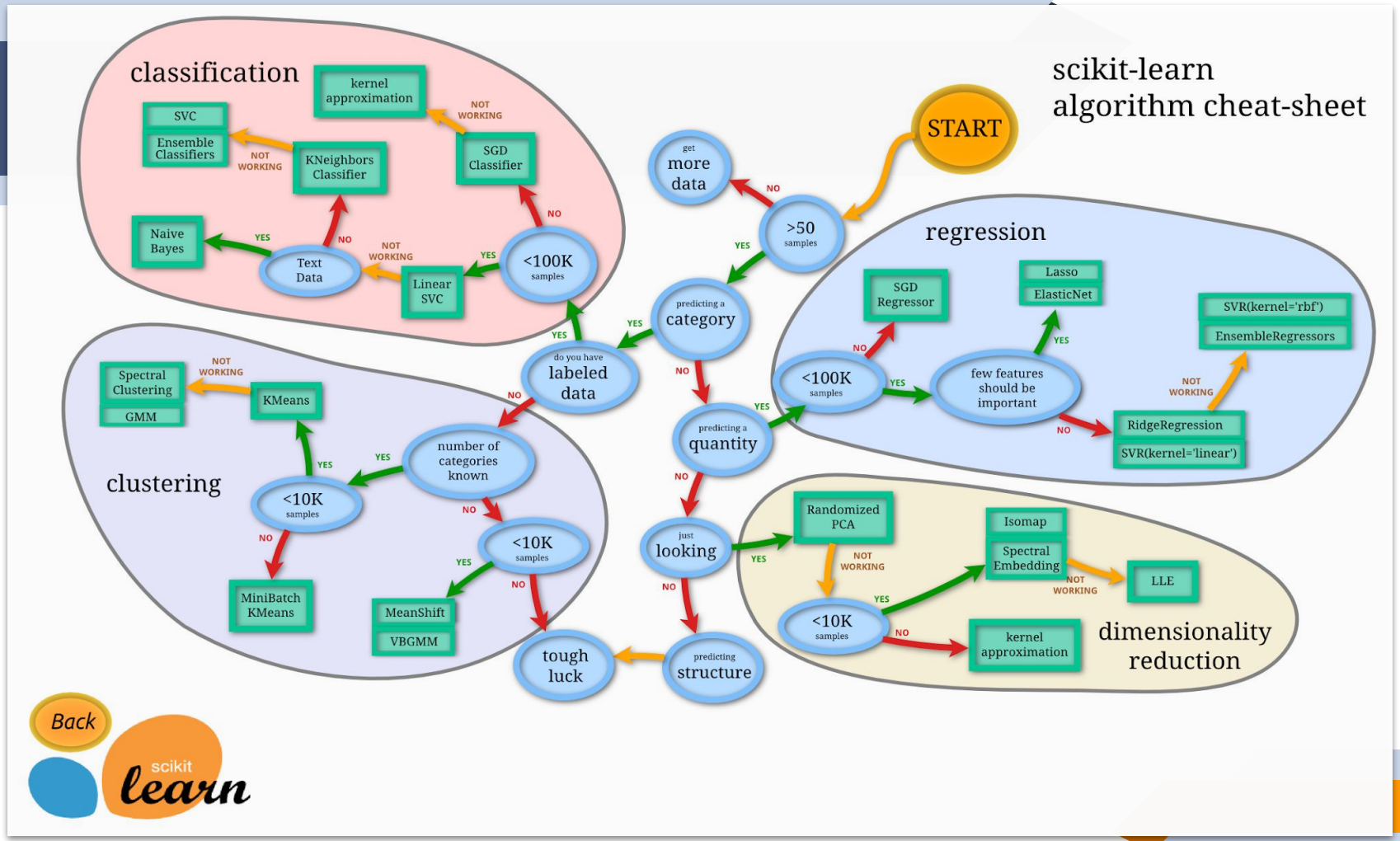


# Reinforcement Learning

# 1.4

## Algorithms

ขั้นตอนวิธีของการเรียนรู้ของเครื่อง


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
NO
RidgeRegression
SVR(kernel='linear')
predicting a category
YES
NO
do you have labeled data
YES
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
NO
MeanShift
VBGMM
just looking
YES
NO
dimensionality reduction
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
tough luck
predicting structure
Back
scikit learn

# Machine Learning Frameworks

TensorFlow™

Caffe

Microsoft Cognitive Toolkit
aka.ms/CognitiveToolkit
Microsoft

scikit learn

amazon web services
**Amazon** Machine Learning
aws.amazon.com/machine-learning

**Apache SINGA**
A General Distributed
Deep Learning Platform

Torch
Accord.NET
Apache Mahout
Theano
Brainstorm
etc,.

# Python Libraries for Data Science - ชุดคำสั่ง

## Core Libraries

- **Numpy**
- Scipy
- **Pandas**
- Statsmodels

## Data Acquisition

- **Scrapy**
- **Quadl**
- Requests
- Selenium
- lxml
- Beautiful Soup

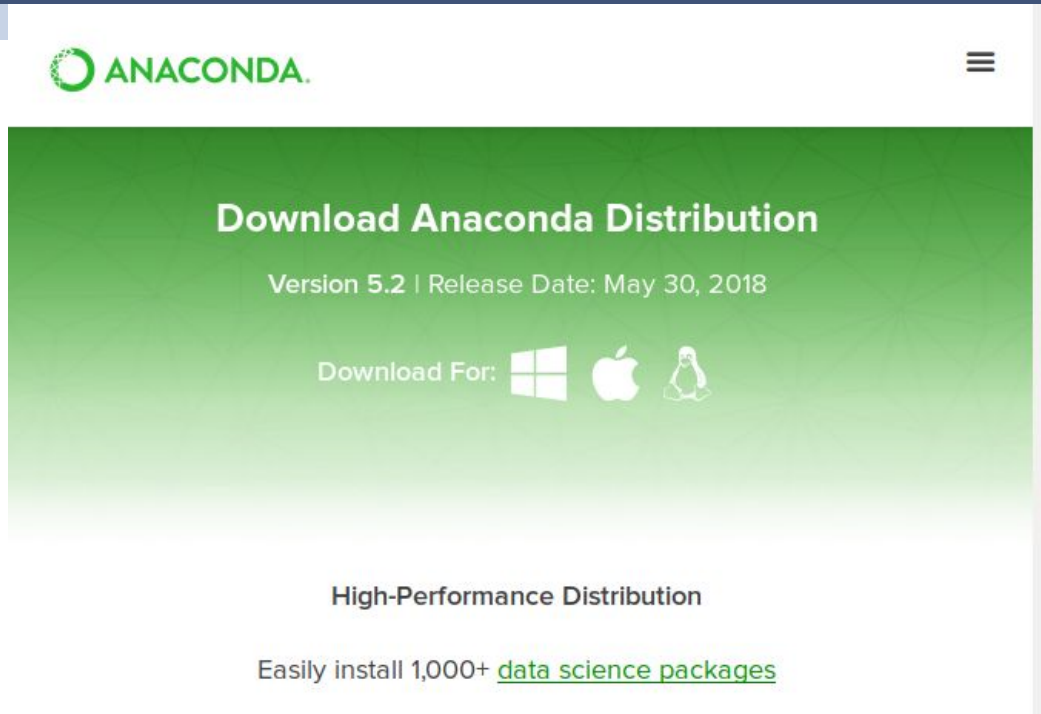
## Visualization

- Matplotlib
- **Seaborn**
- Bokeh
- Plotly

## Machine Learning

- **SciKit-Learn**
- Theano
- **Tensorflow**
- **TFLearn**
- Keras
- NLTK
- Gensim

# Software & Downloads




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

**Includes...**

- Numpy
- Pandas
- Tensorflow
- Scrapy
- Seaborn
- ...

## Hand-on Lab http://gg.gg/b1jmx

https://mybinder.org/v2/gh/baldwint/PythonDataScienceHandbook/py35

- 10.45 - 12.00 **Data Modeling**
- **Lunch**
- 13.00 - 14.30 **Model Validation**
- **Coffee Break**
- 14.45 - 16.30 **Case Study**