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]

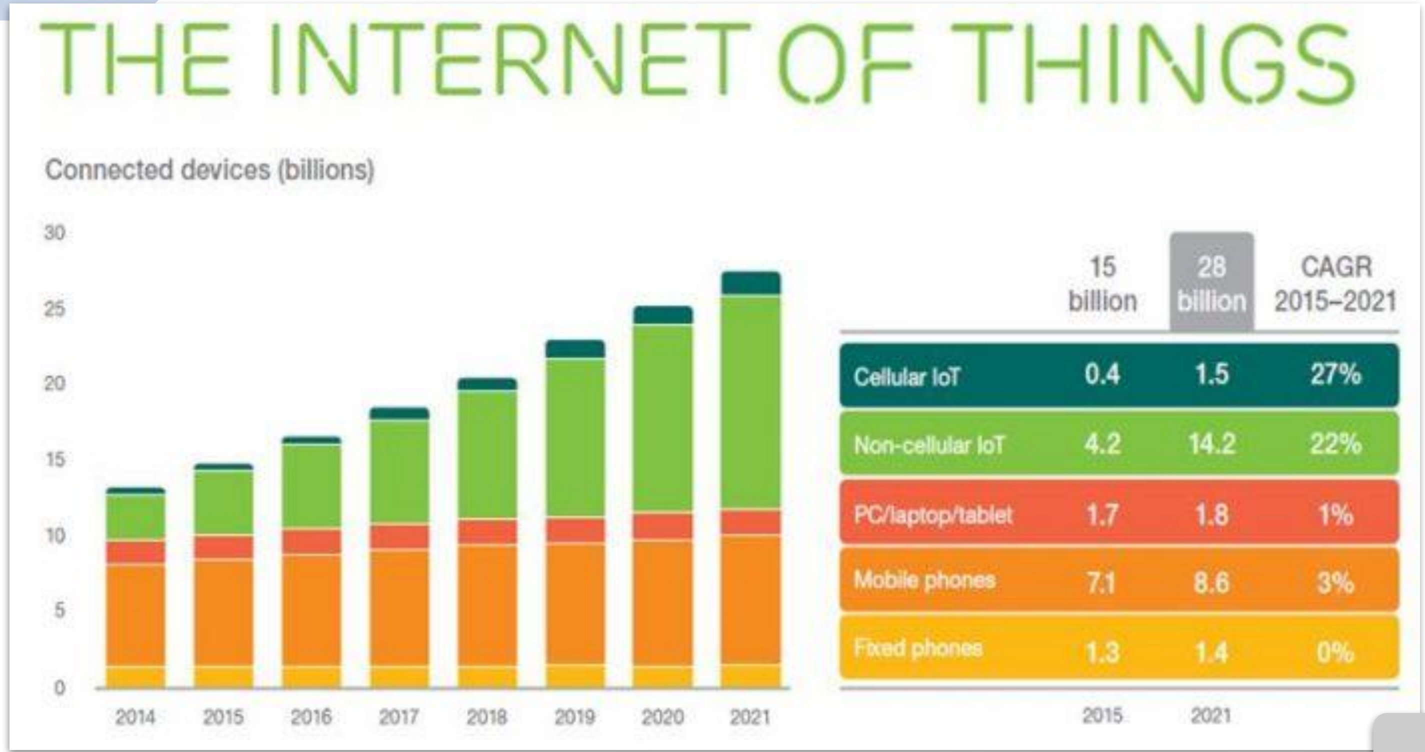
[1] statista.com

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

[1] University of Rochester

[1]Spark Summit 2017

Social Media for UN Project Fortis for humanitarian aid plan



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	

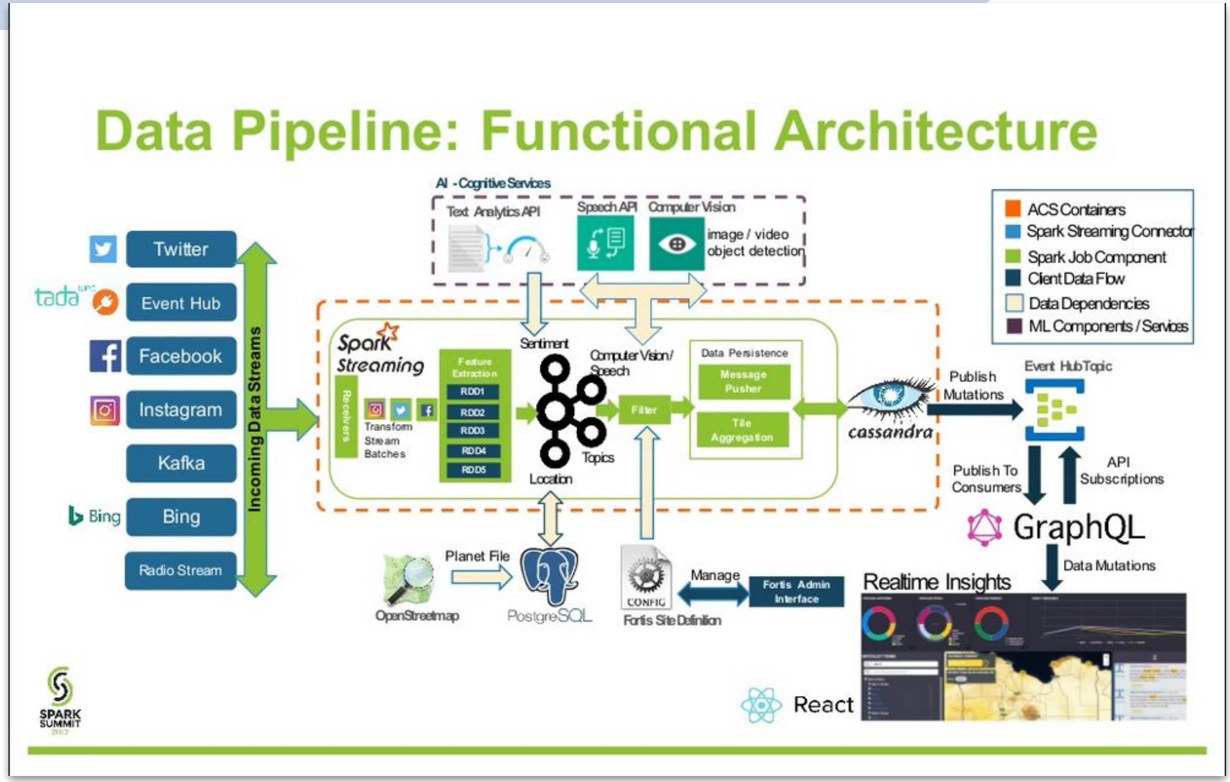
[1]ericsson.com

End Users IoT dashboards

Data per IoT Device[1]

[1] Thingsboard

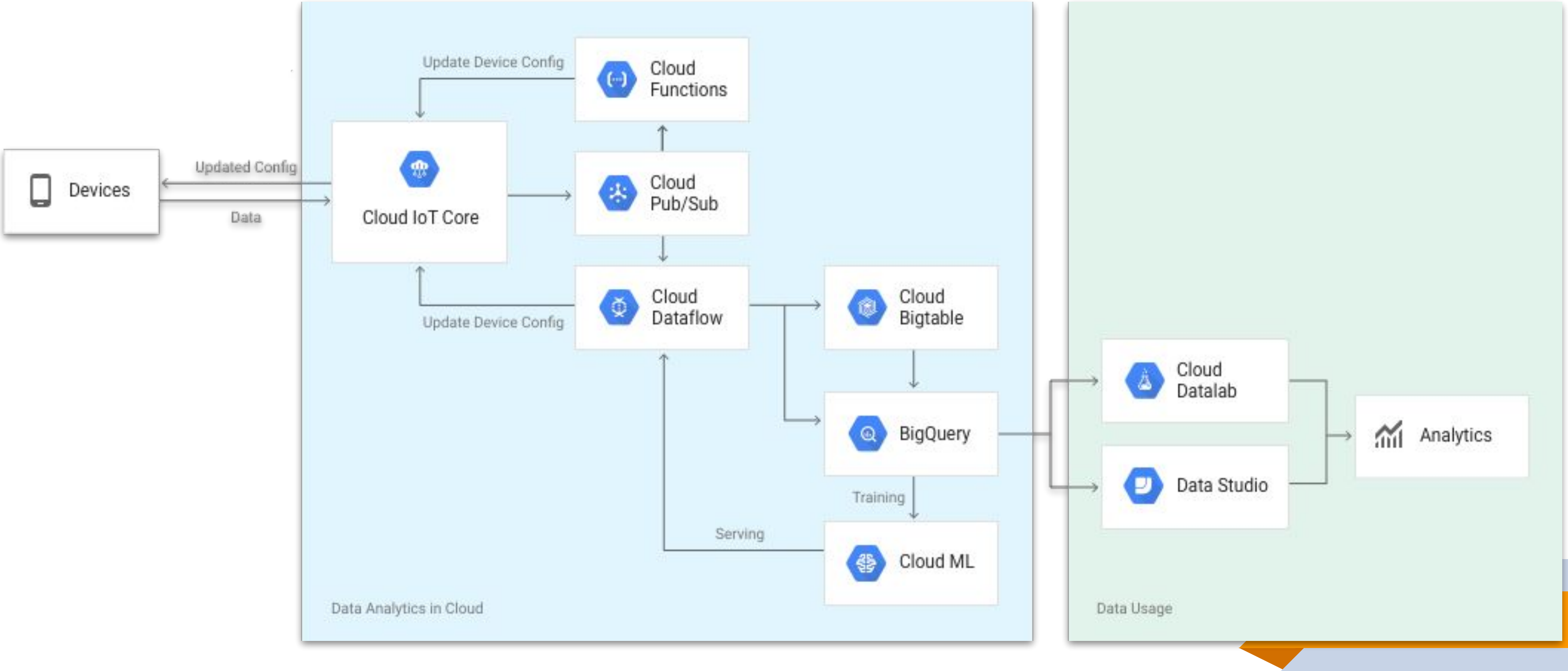
Architecture for (Social) Data Analysis$^{\text{fortis}}$



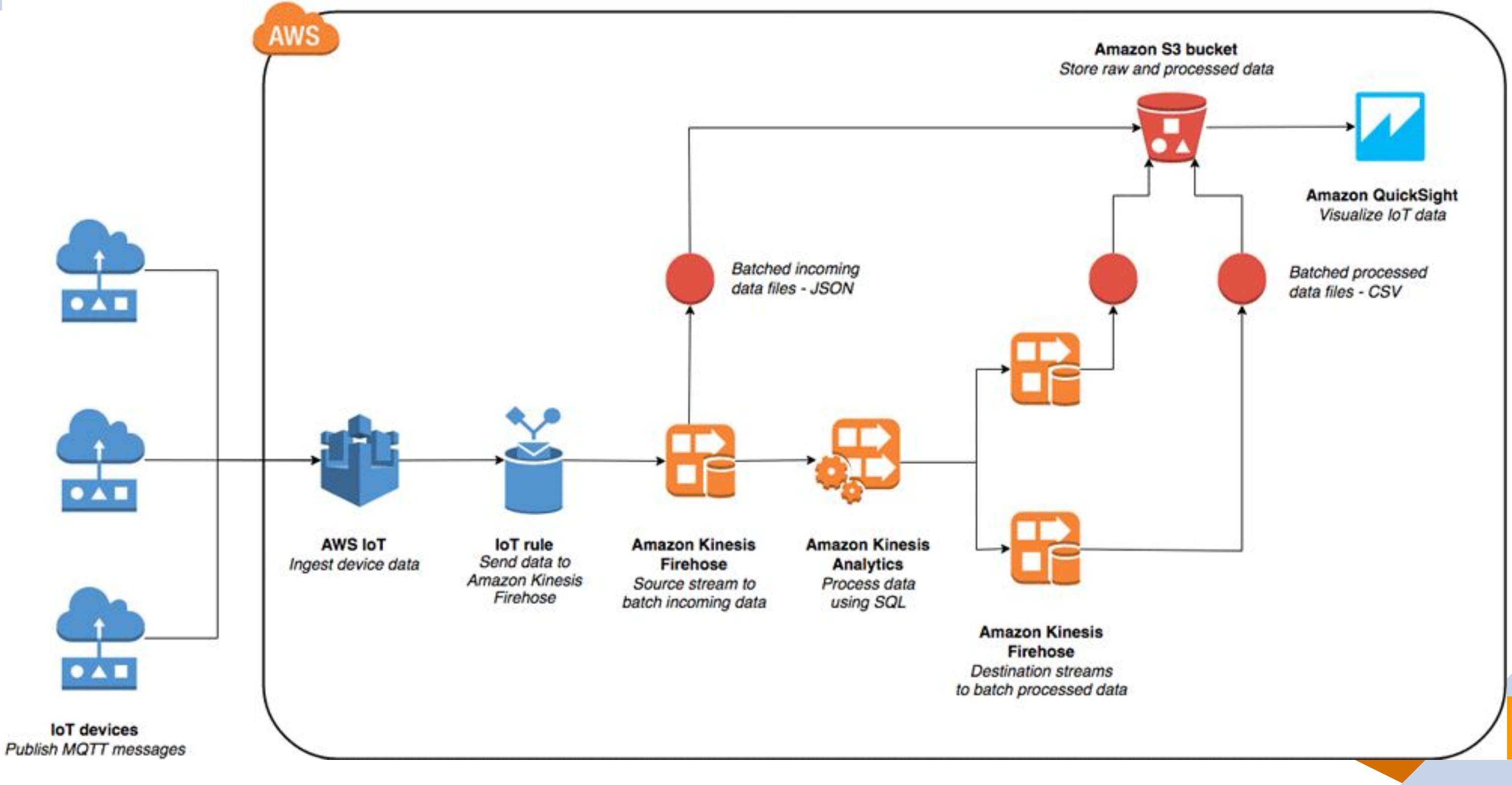
Data Pipeline: Functional Architecture

[1]Spark Summit 2017

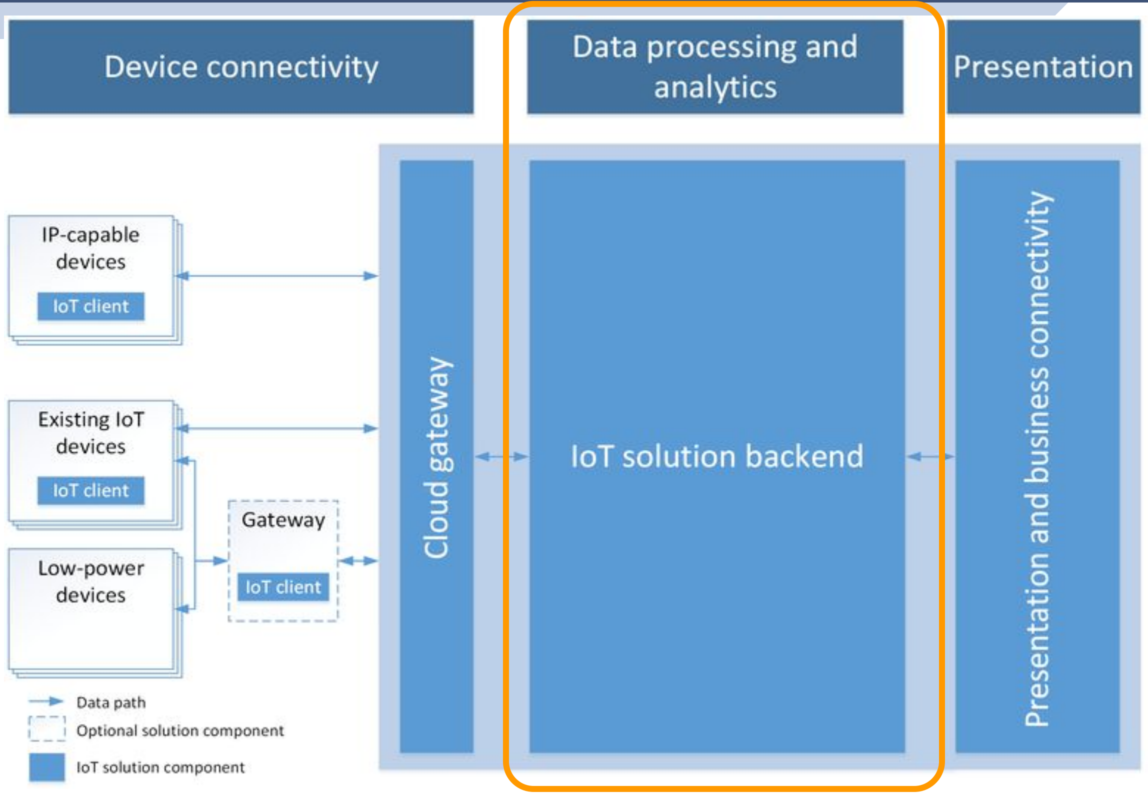
Architecture for Data Analysis^{Google Cloud}



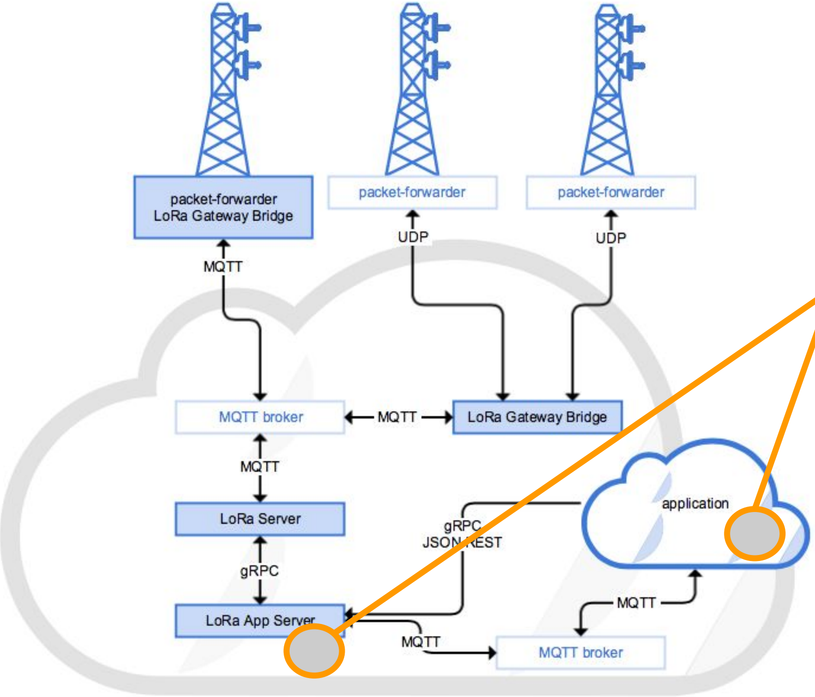
Architecture for Data Analysis$^{\text{Amazon}}$



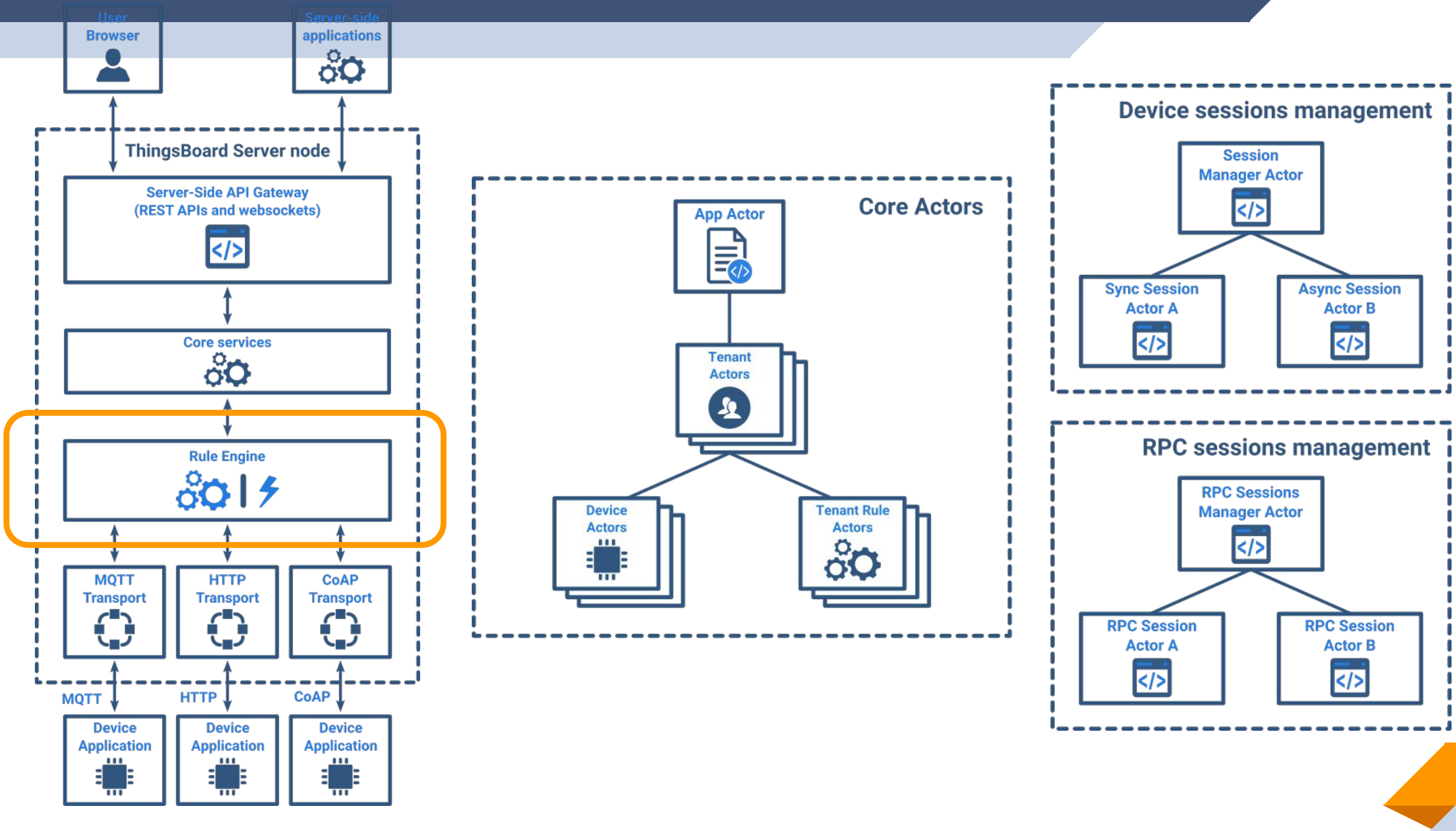
Architecture for Data Analysis$^{\text{Ms Azure}}$



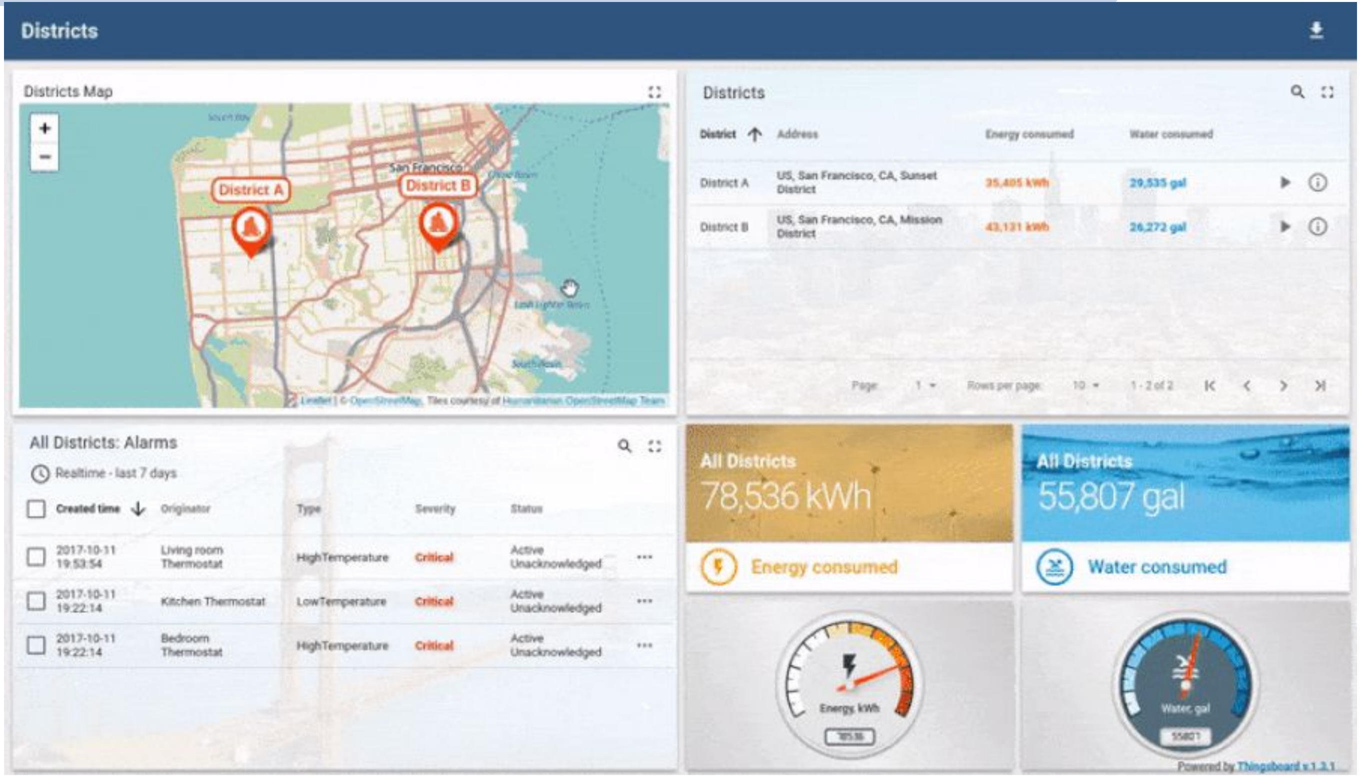
IoT Architecture for ~~Data Analysis~~Loraserver



Architecture for Data Analysis$^{\text{thingsboard}}$



Architecture for Data Analysis$^{\text{thingsboard}}$



Districts

Districts Map

District A
District B
San Francisco
Leaflet | © OpenStreetMap, Tiles courtesy of Humanitarian OpenStreetMap Team

Districts

District ↑	Address	Energy consumed	Water consumed
District A	US, San Francisco, CA, Sunset District	35,405 kWh	29,535 gal
District B	US, San Francisco, CA, Mission District	43,131 kWh	26,272 gal

Page: 1 Rows per page: 10 1 - 2 of 2

All Districts: Alarms

Realtime - last 7 days

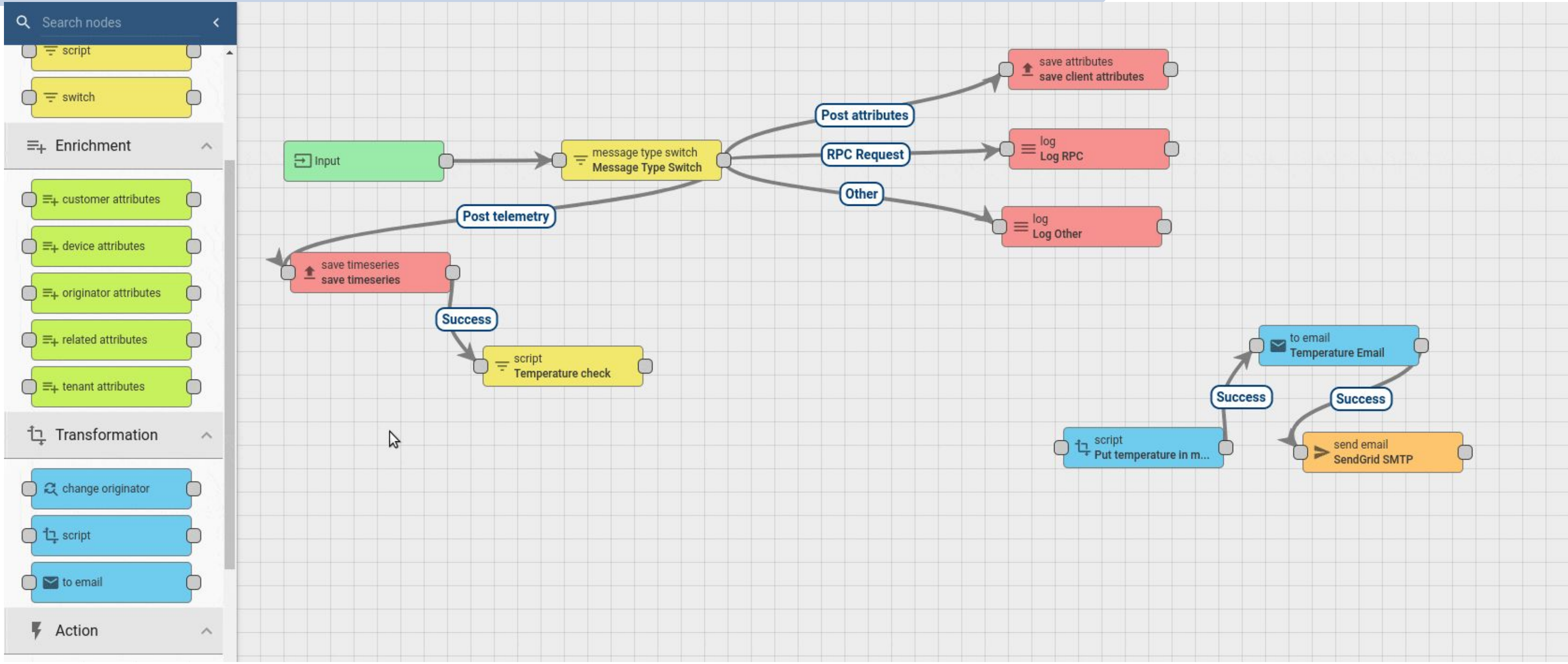
□	Created time ↓	Originator	Type	Severity	Status
□	2017-10-11 19:53:54	Living room Thermostat	HighTemperature	Critical	Active Unacknowledged
□	2017-10-11 19:22:14	Kitchen Thermostat	LowTemperature	Critical	Active Unacknowledged
□	2017-10-11 19:22:14	Bedroom Thermostat	HighTemperature	Critical	Active Unacknowledged

All Districts
78,536 kWh
Energy consumed

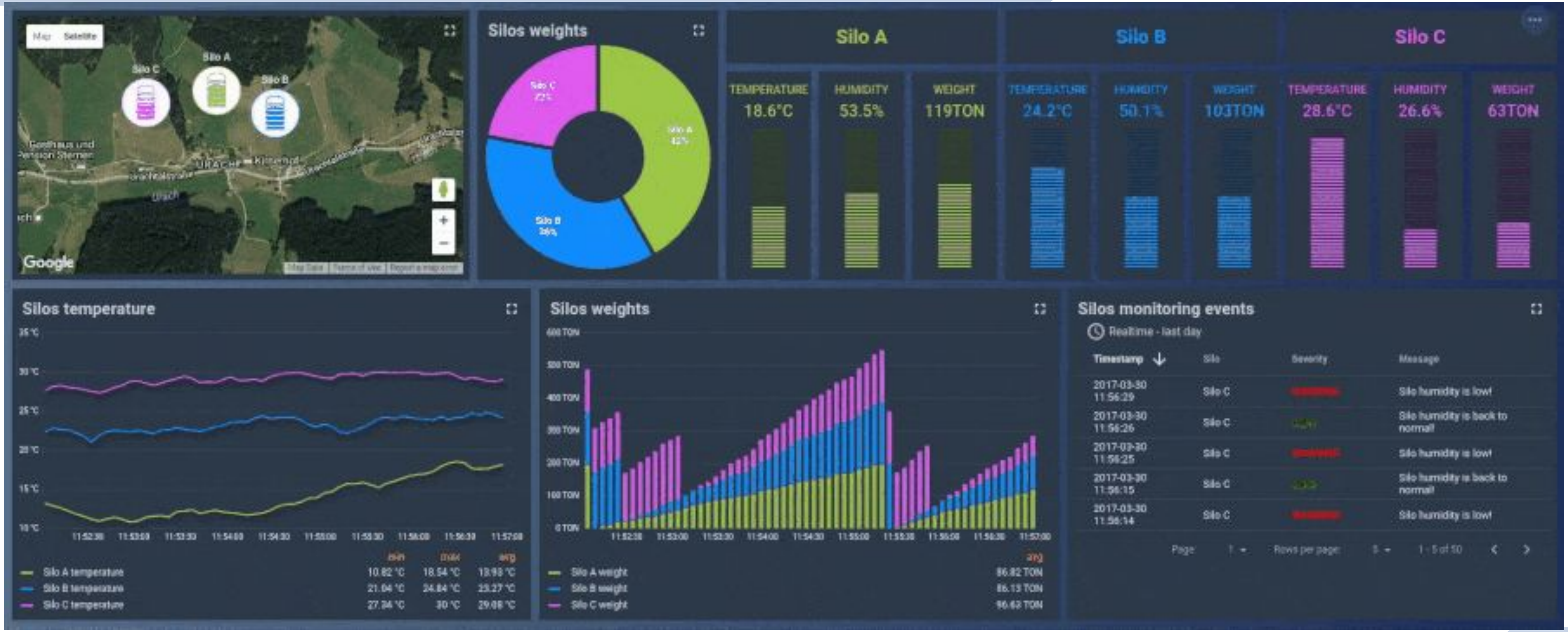
All Districts
55,807 gal
Water consumed

Powered by Thingsboard v.1.3.1

Architecture for Data Analysis thingsboard



Architecture for Data Analysis^{thingsboard}



Silo A

TEMPERATURE	HUMIDITY	WEIGHT
18.6°C	53.5%	119TON

Silo B

TEMPERATURE	HUMIDITY	WEIGHT
24.2°C	50.1%	103TON

Silo C

TEMPERATURE	HUMIDITY	WEIGHT
28.6°C	26.6%	63TON

Silos temperature

	min	max	avg
Silo A temperature	10.82 °C	18.54 °C	13.93 °C
Silo B temperature	21.04 °C	24.84 °C	23.27 °C
Silo C temperature	27.34 °C	30 °C	29.08 °C

Silos weights

	avg
Silo A weight	86.82 TON
Silo B weight	86.13 TON
Silo C weight	96.63 TON

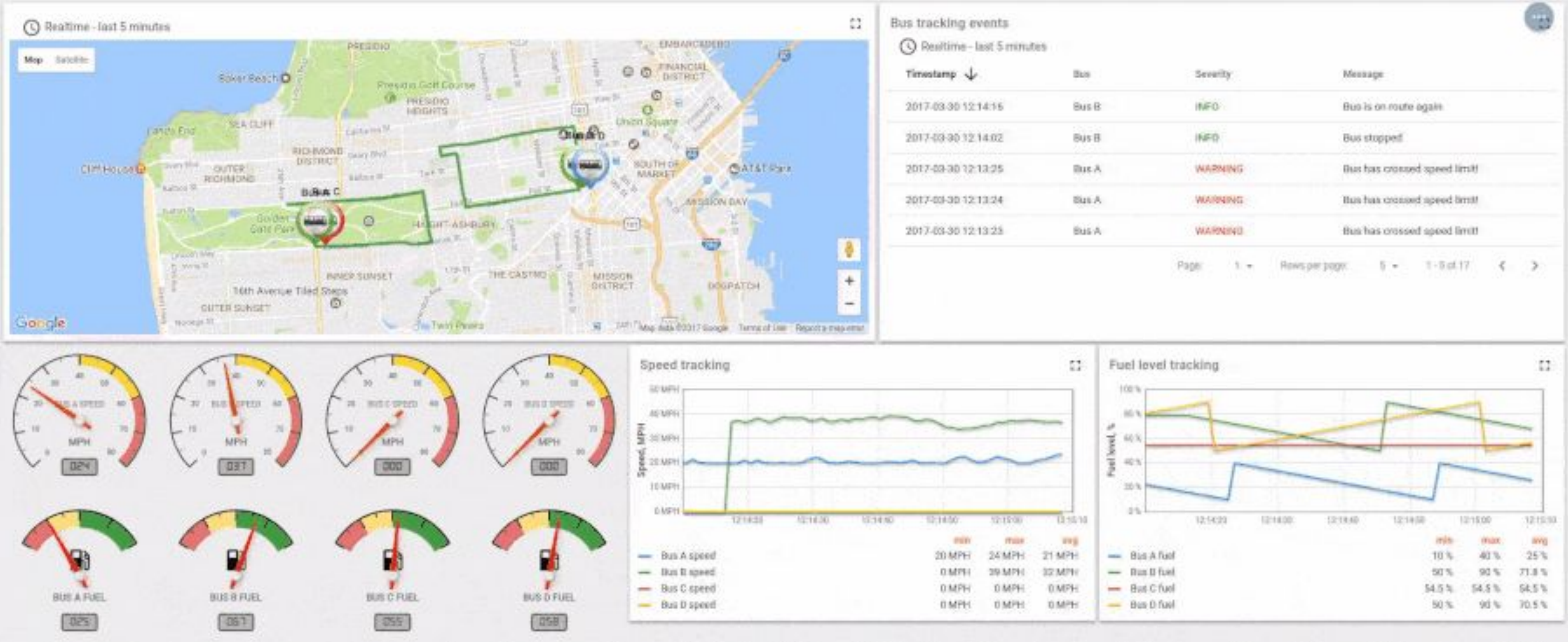
Silos monitoring events

Realtime - last day

Timestamp ↓	Silo	Severity	Message
2017-03-30 11:56:29	Silo C	[illegible]	Silo humidity is low!
2017-03-30 11:56:26	Silo C	[illegible]	Silo humidity is back to normal!
2017-03-30 11:56:25	Silo C	[illegible]	Silo humidity is low!
2017-03-30 11:56:15	Silo C	[illegible]	Silo humidity is back to normal!
2017-03-30 11:56:14	Silo C	[illegible]	Silo humidity is low!

Page: 1 ▾ Rows per page: 5 ▾ 1 - 5 of 50 ‹ ›

Architecture for Data Analysis$^{\text{thingsboard}}$



Realtime - last 5 minutes

Bus tracking events

Realtime - last 5 minutes

Timestamp ↓	Bus	Severity	Message
2017-03-30 12:14:16	Bus B	INFO	Bus is on route again
2017-03-30 12:14:02	Bus B	INFO	Bus stopped
2017-03-30 12:13:25	Bus A	WARNING	Bus has crossed speed limit!
2017-03-30 12:13:24	Bus A	WARNING	Bus has crossed speed limit!
2017-03-30 12:13:23	Bus A	WARNING	Bus has crossed speed limit!

Page: 1 ▾ Rows per page: 5 ▾ 1 - 5 of 17 ‹ ›

Speed tracking

	min	max	avg
Bus A speed	20 MPH	24 MPH	21 MPH
Bus B speed	0 MPH	39 MPH	32 MPH
Bus C speed	0 MPH	0 MPH	0 MPH
Bus D speed	0 MPH	0 MPH	0 MPH

Fuel level tracking

	min	max	avg
Bus A fuel	10 %	40 %	25 %
Bus B fuel	50 %	90 %	71.8 %
Bus C fuel	54.5 %	54.5 %	54.5 %
Bus D fuel	50 %	90 %	70.5 %

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

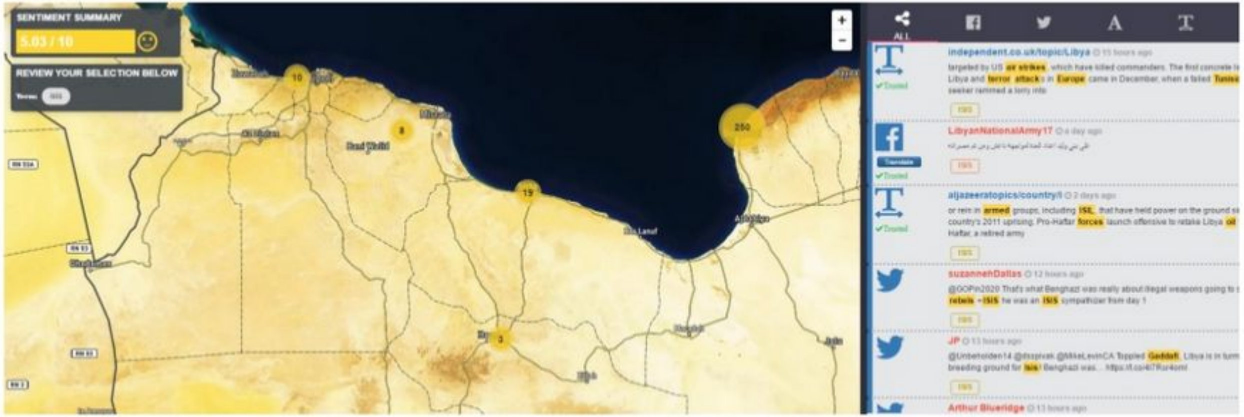
Fortis UN project goals

- **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



independent.co.uk/topic/Libya 15 hours ago
targeted by US air strikes, which have killed commanders. The first concrete l... Libya and terror attacks in Europe came in December, when a failed Tunis... seeker rammed a lorry into
ISIS

LibyanNationalArmy17 a day ago
[illegible]
ISIS

aljazeeratopics/country/l 2 days ago
or rein in armed groups, including ISIL, that have held power on the ground si... country's 2011 uprising. Pro-Haftar forces launch offensive to retake Libya oil... Haftar, a retired army
ISIS

suzannehDallas 12 hours ago
@GOPin2020 That's what Benghazi was really about illegal weapons going to ... rebels ~ISIS he was an ISIS sympathizer from day 1
ISIS

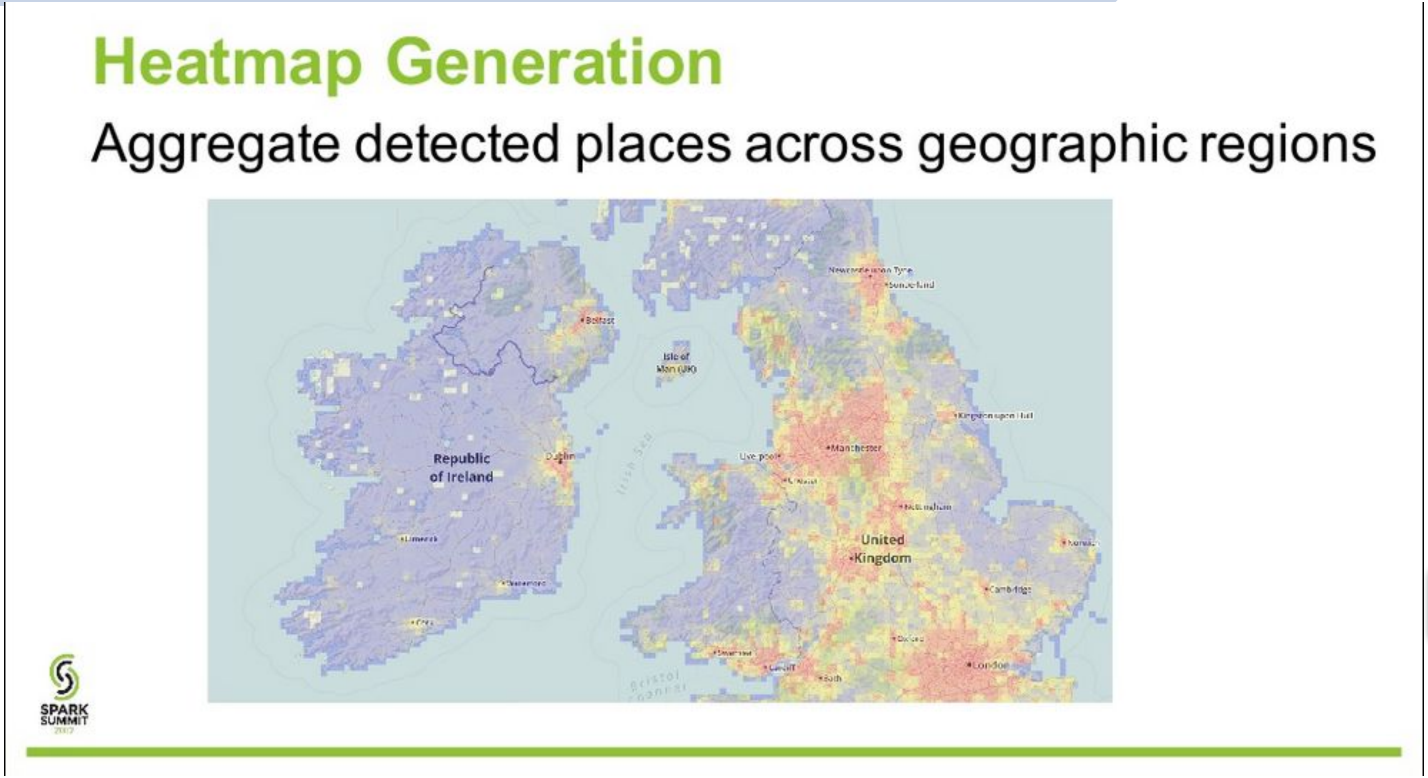
JP 13 hours ago
@Unbeholden14 @despivak @MikeLevinCA Toppled Gaddafi, Libya is in turm... breeding ground for Isis! Benghazi was... https://t.co/4i7Rur4oml
ISIS

Arthur Blueridge 13 hours ago

- 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

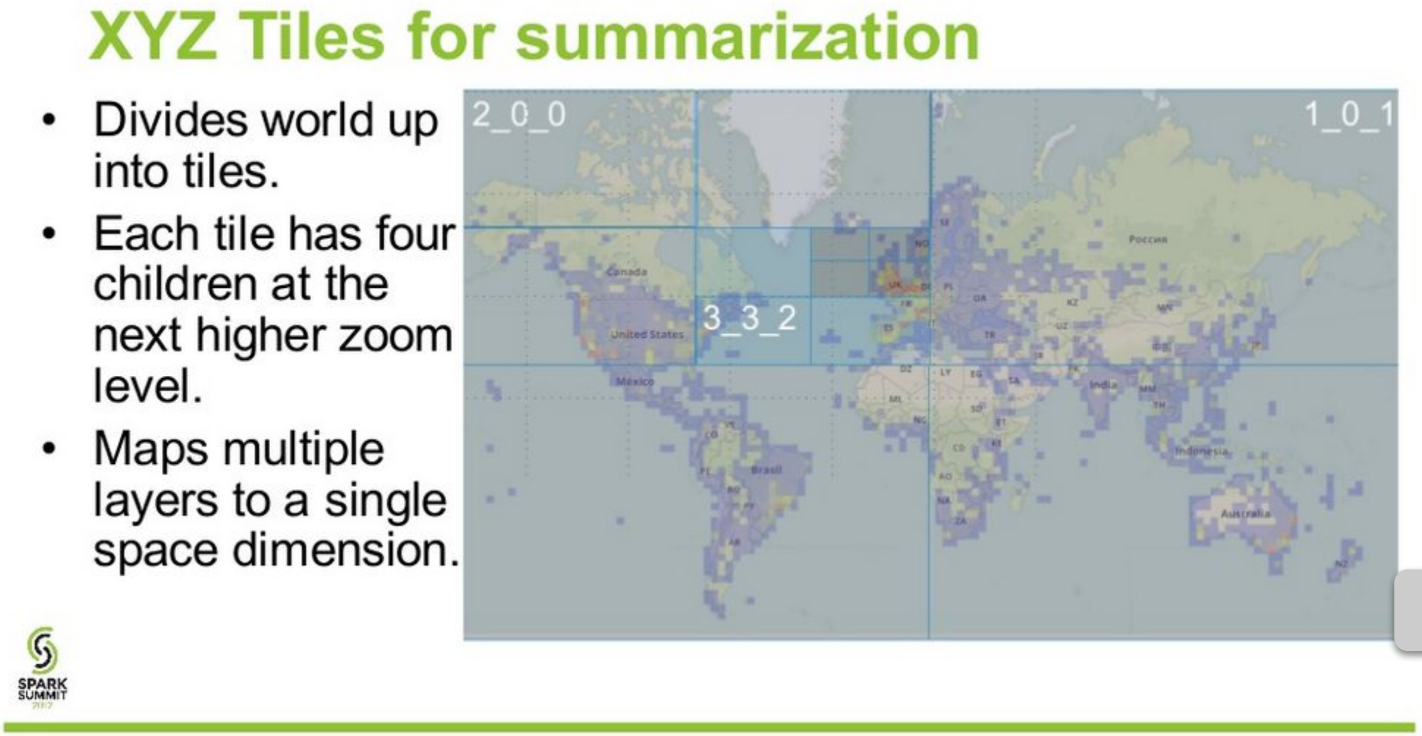


Heatmap Generation

Aggregate detected places across geographic regions

[1]Spark Summit 2017

Fortis UN project **tasks**



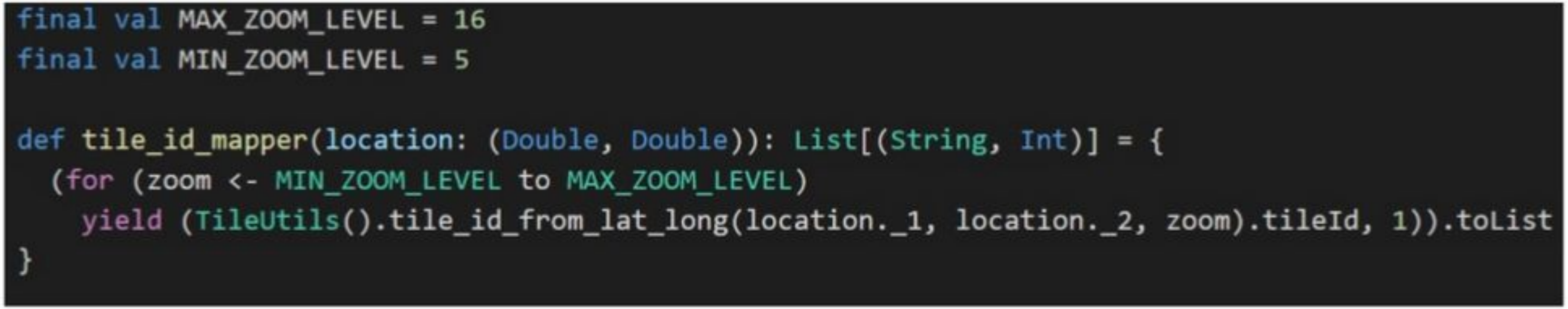
XYZ Tiles for summarization

- 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

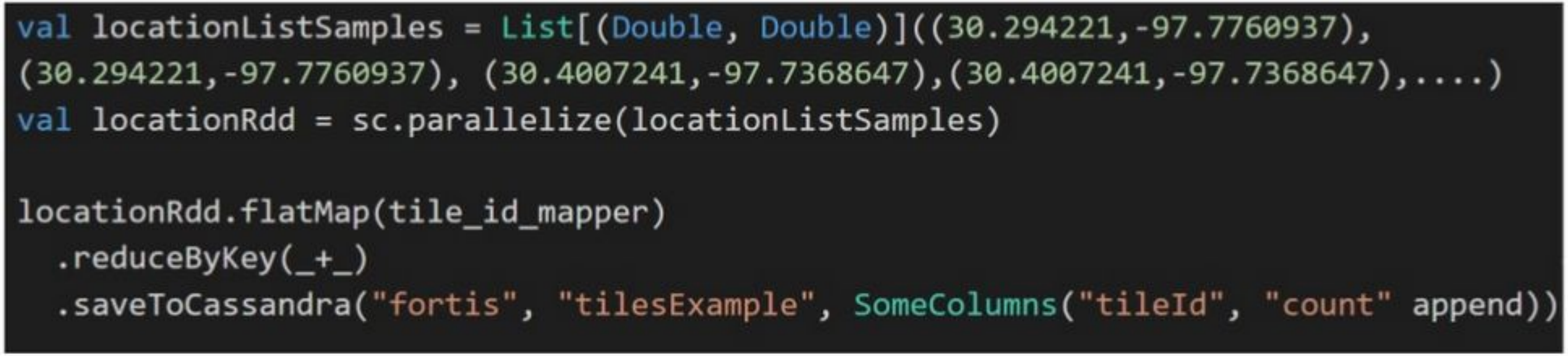
Fortis UN project tasks - Heatmap Spark

[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
}
```

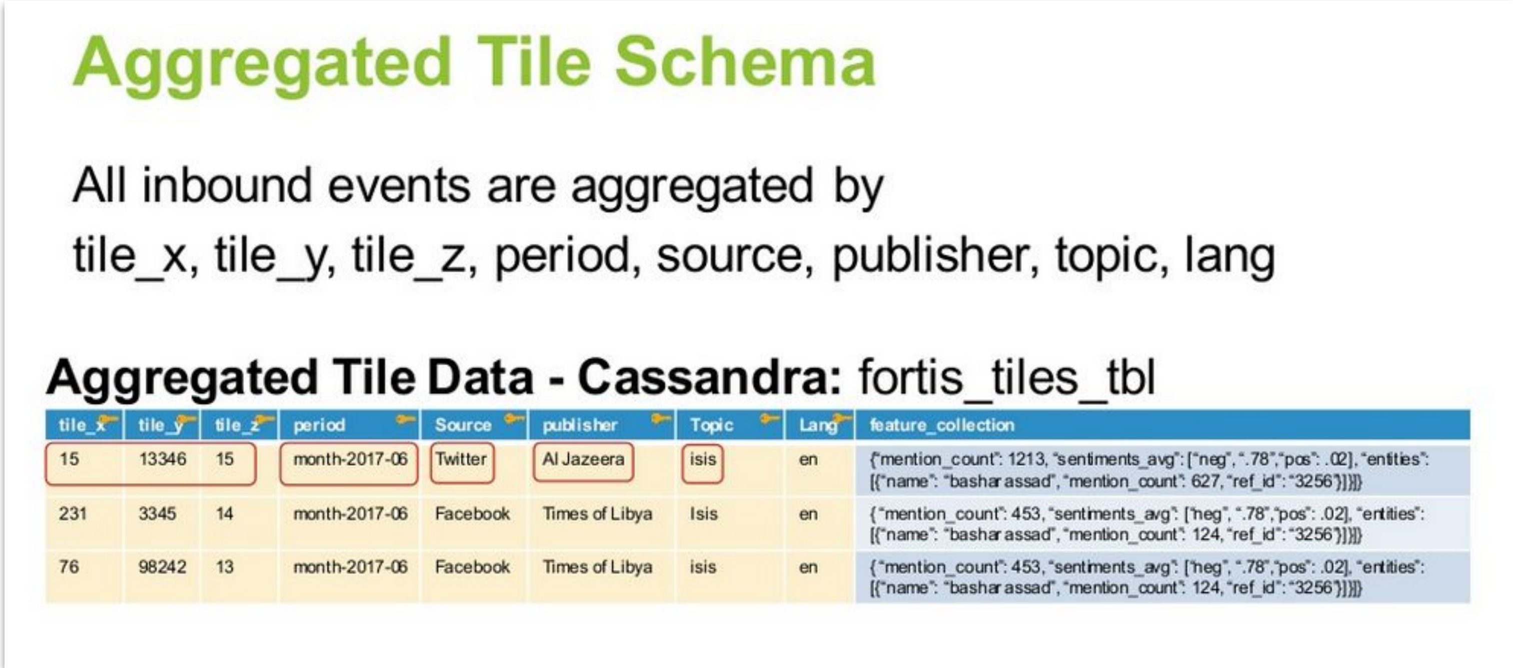


```
val locationListSamples = List[(Double, Double)]((30.294221,-97.7760937),
(30.294221,-97.7760937), (30.4007241,-97.7368647),(30.4007241,-97.7368647),....)
val locationRdd = sc.parallelize(locationListSamples)

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

Fortis UN project tasks

[1]Spark Summit 2017



Aggregated Tile Schema

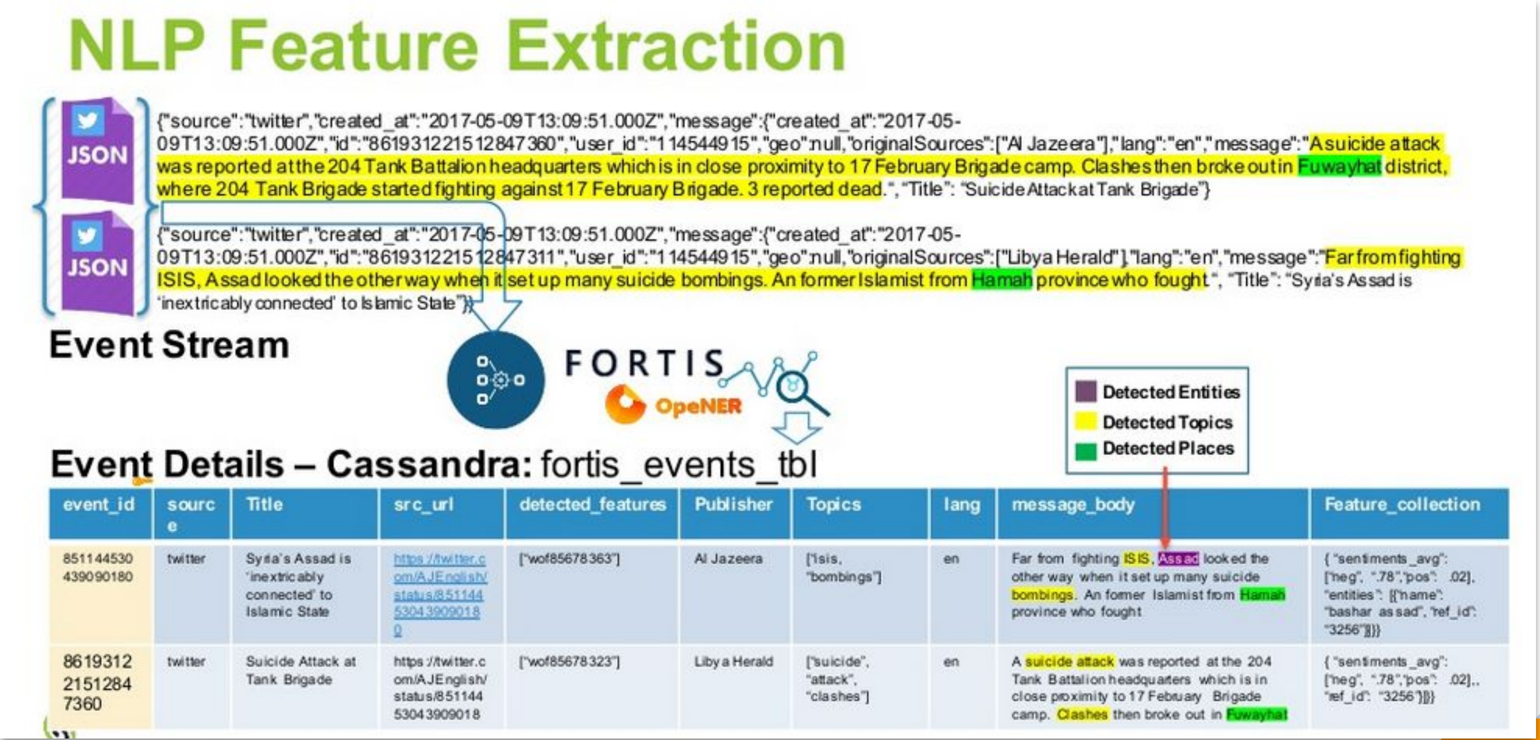
All inbound events are aggregated by
tile_x, tile_y, tile_z, period, source, publisher, topic, lang

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": "bashar assad", "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": "bashar assad", "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": "bashar assad", "mention_count": 124, "ref_id": "3256"}]}}

Fortis UN project tasks

[1]Spark Summit 2017



NLP Feature Extraction

{"source":"twitter","created_at":"2017-05-09T13:09:51.000Z","message":{"created_at":"2017-05-09T13:09:51.000Z","id":"861931221512847360","user_id":"114544915","geo":null,"originalSources":["Al Jazeera"],"lang":"en","message":"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 district, where 204 Tank Brigade started fighting against 17 February Brigade. 3 reported dead.", "Title": "Suicide Attack at Tank Brigade"}

{"source":"twitter","created_at":"2017-05-09T13:09:51.000Z","message":{"created_at":"2017-05-09T13:09:51.000Z","id":"861931221512847311","user_id":"114544915","geo":null,"originalSources":["Libya Herald"],"lang":"en","message":"Far from fighting ISIS, Assad looked the other way when it set up many suicide bombings. An former Islamist from Hamah province who fought", "Title": "Syria's Assad is 'inextricably connected' to Islamic State"}}

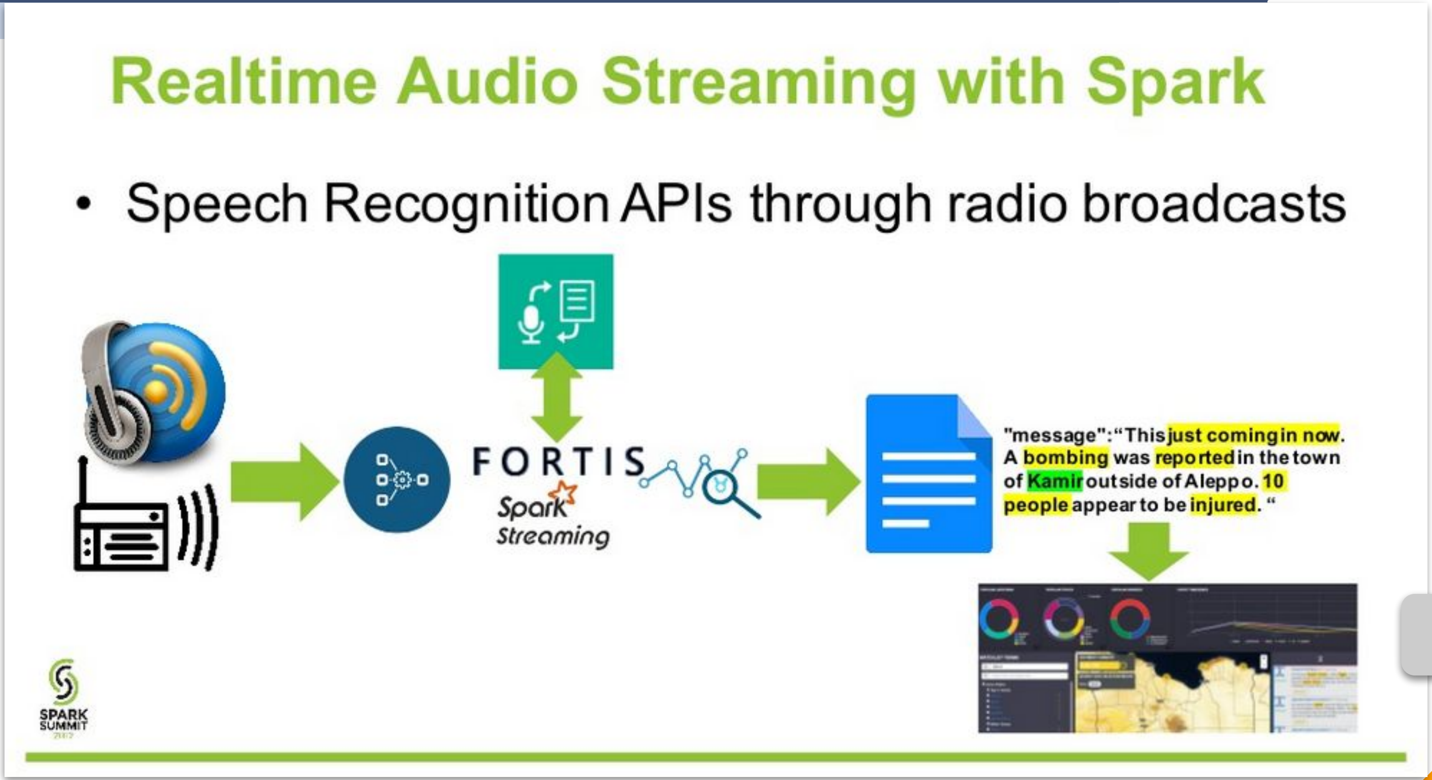
Event Stream

- 🟪 Detected Entities
- 🟨 Detected Topics
- 🟩 Detected Places

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/8511445304390901 80	["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/851144530439090 18	["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

"message":"This just coming in now. A bombing was reported in the town of Kamir outside of Aleppo. 10 people appear to be injured. "

[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.

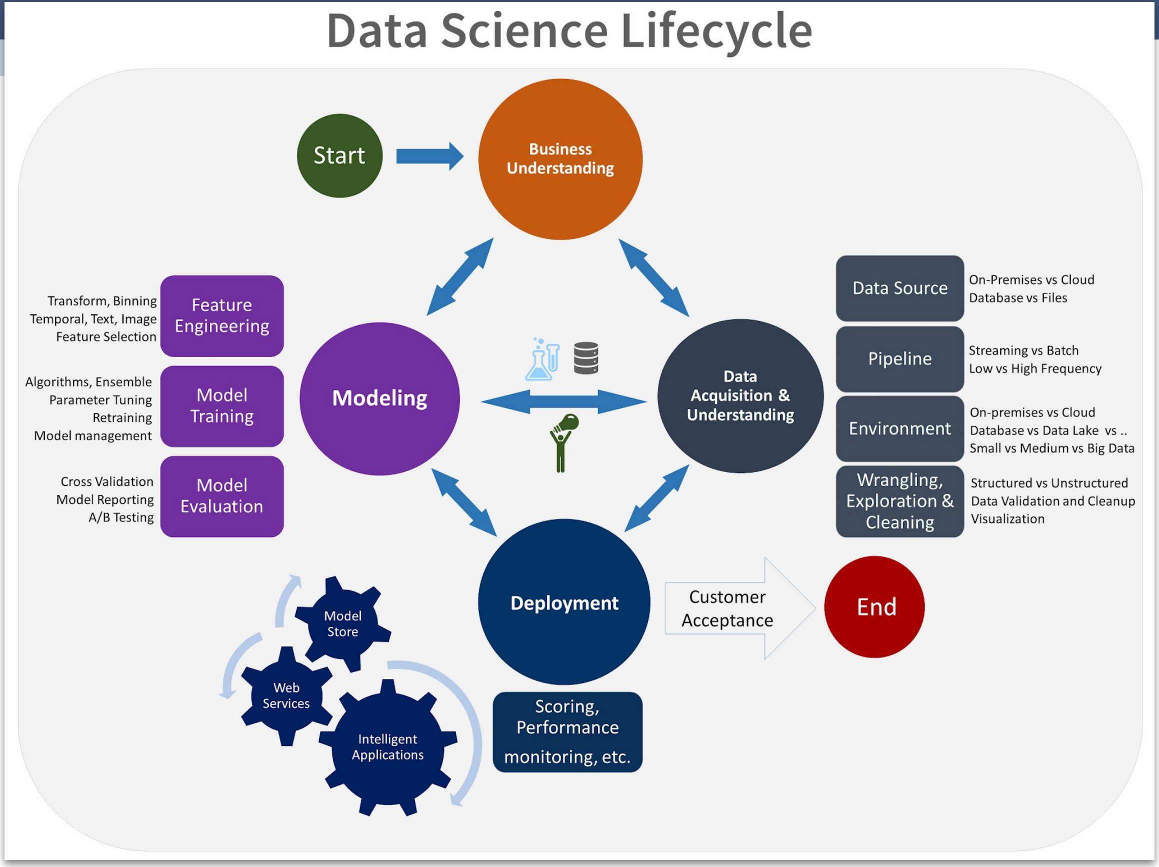
[1] Spark Summit 2017

1.3

Data Science Process

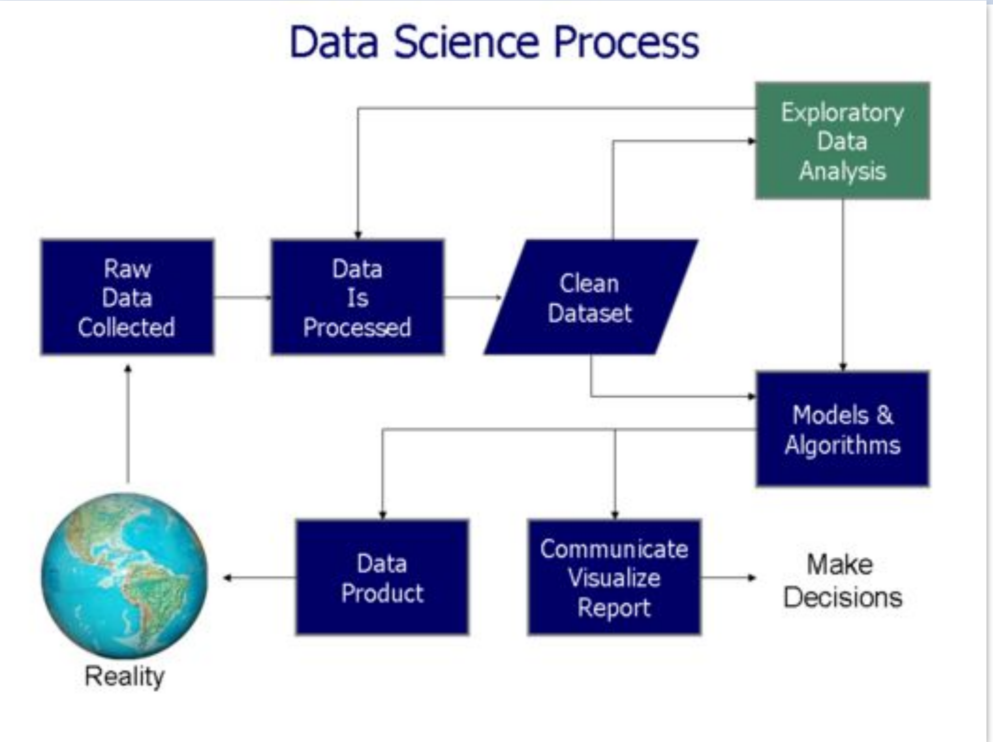
ขั้นตอนของวิทยาการข้อมูล

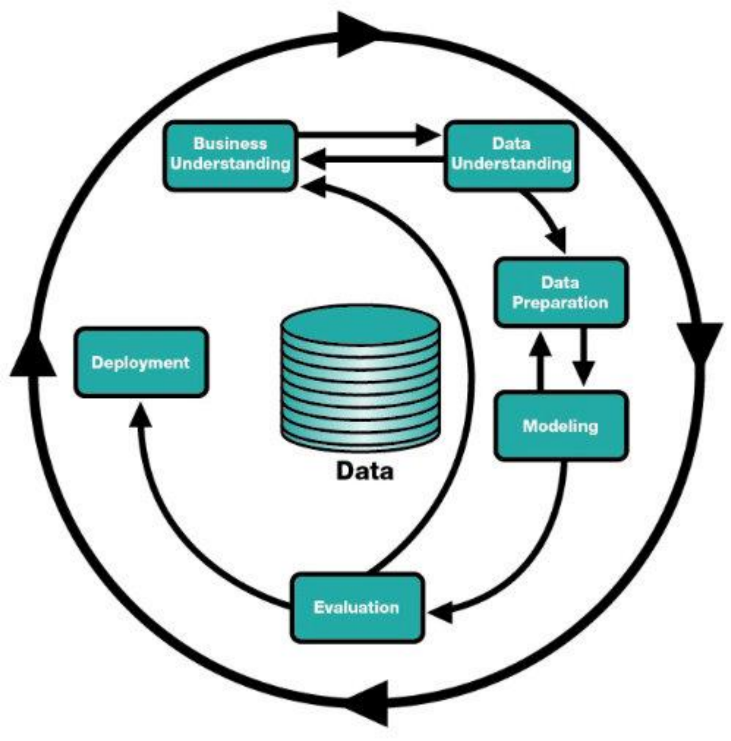
Data Science Lifecycle[1]



[1] Microsoft

Data Science Process[1,2]



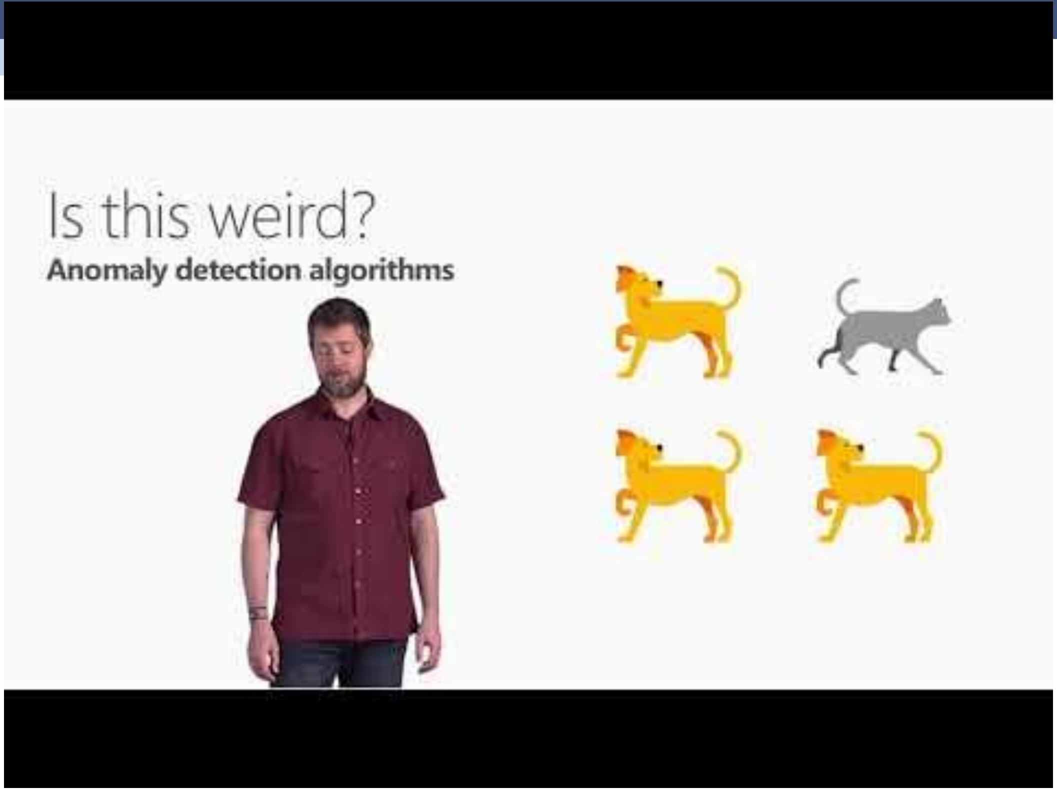


[1] kdnugget

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?



Is this weird?

Anomaly detection algorithms

[1]Daniel Yarmoluk 2018

วิทยาการข้อมูล ตอบคำถามอะไรได้บ้าง?

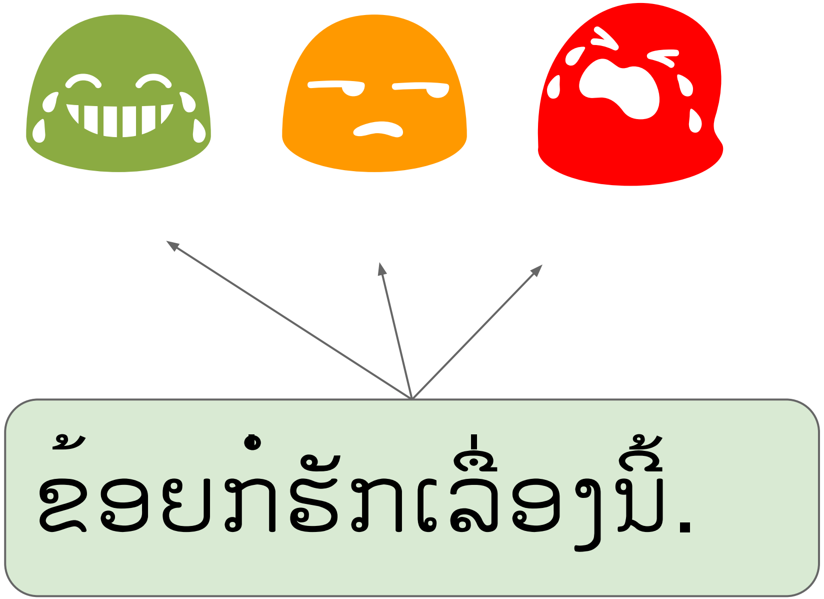
1. Is this **A** or **B**? (classification)
2. Is this **weird**? (anomaly detection)
3. How much or **how many**? (regression)
4. How is this **organized**? (clustering)
5. **What** should I do **next**? (reinforcement learning)

1.4

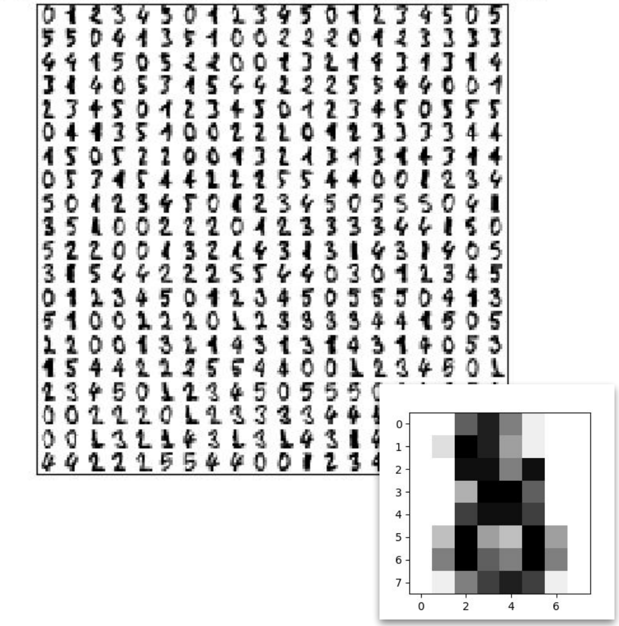
Machine Learning

หลักการเรียนรู้ของเครื่อง

Classifications

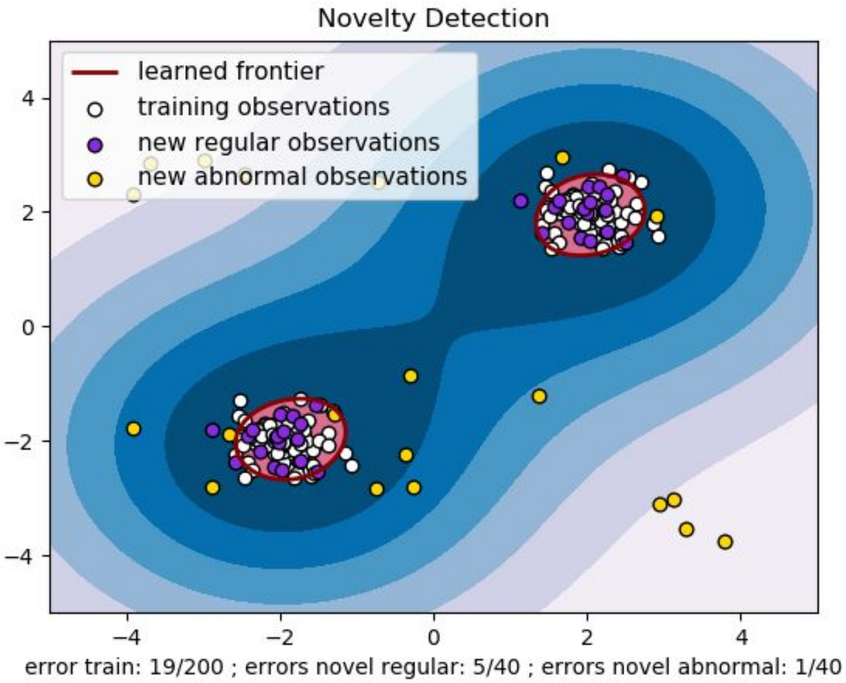


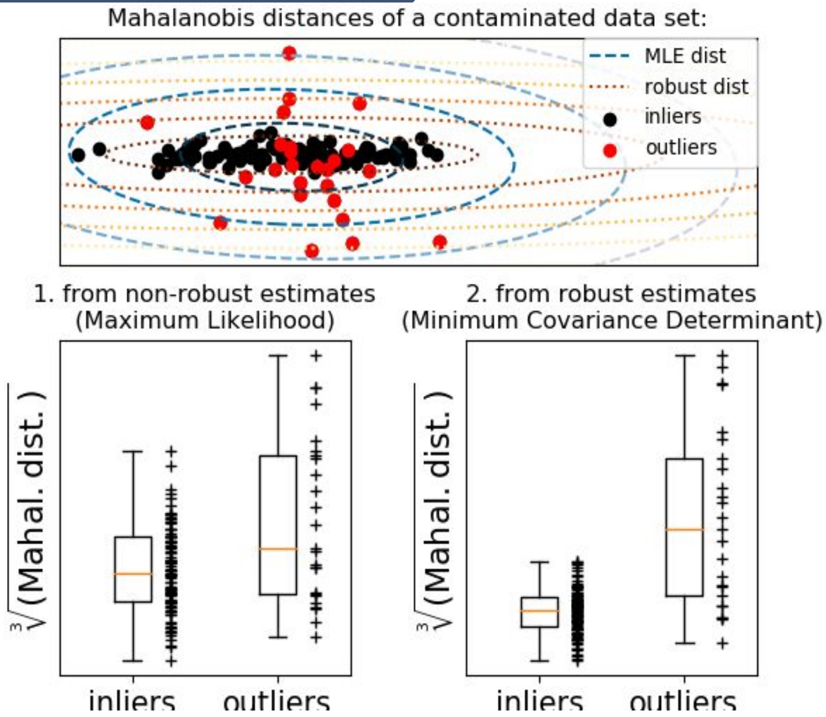
ຂ້ອຍກໍຮັກເລື່ອງນີ້.



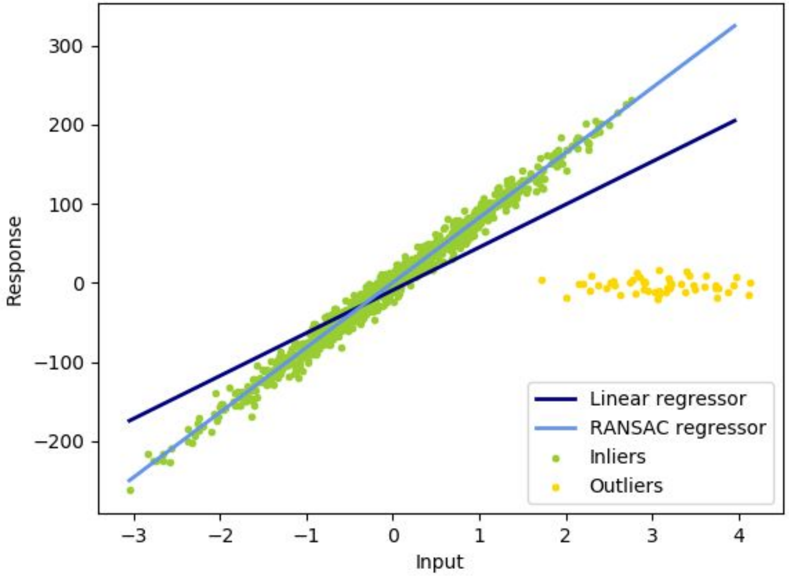
A selection from the 64-dimensional digits dataset

Anomaly Detection

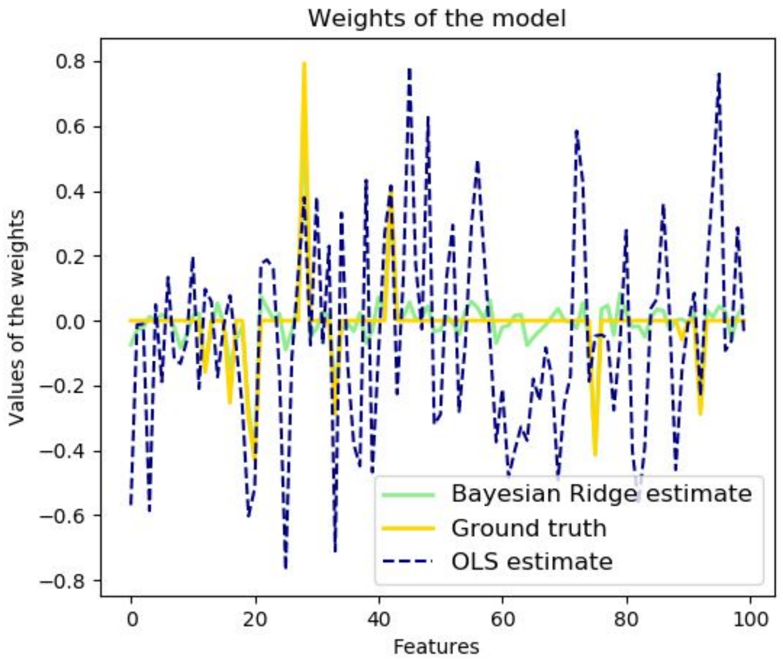




Regression

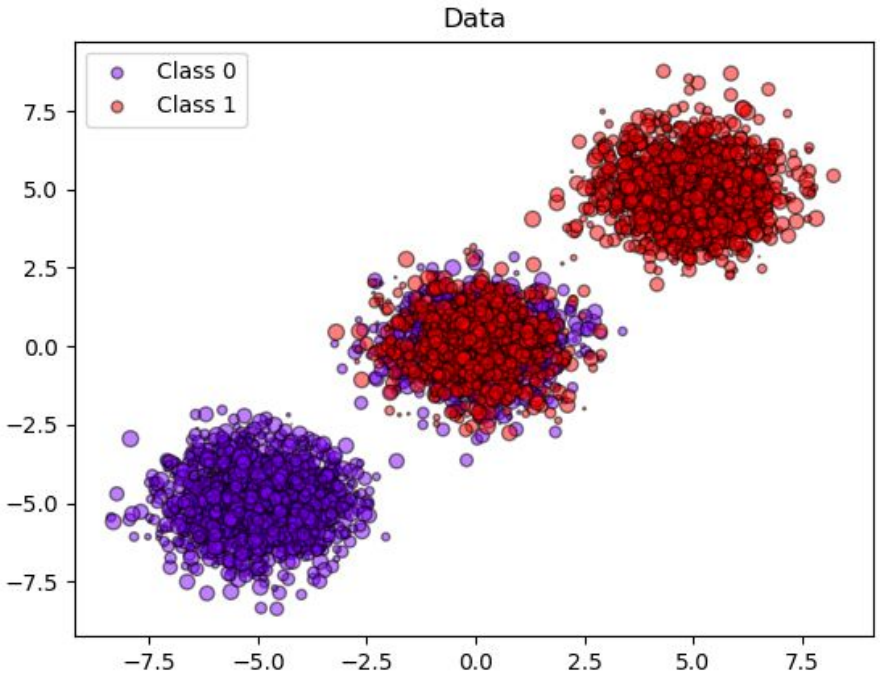


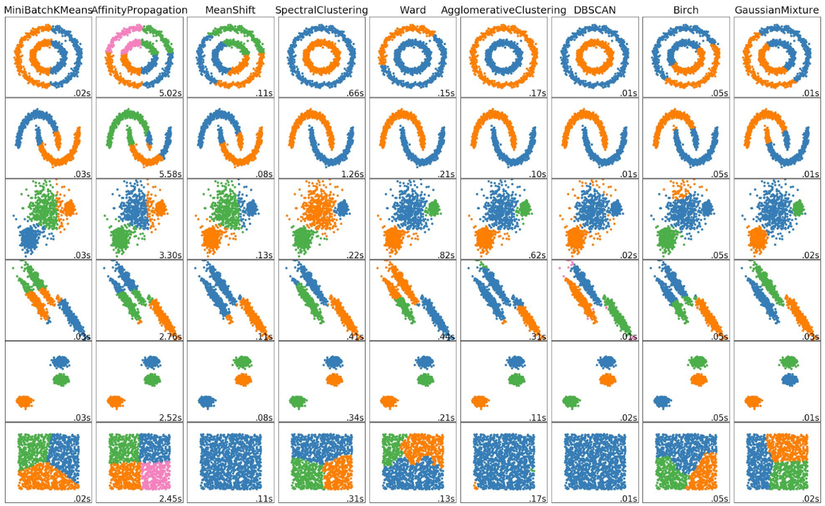
300
200
100
0
−100
−200
Response
−3
−2
−1
0
1
2
3
4
Input
Linear regressor
RANSAC regressor
Inliers
Outliers



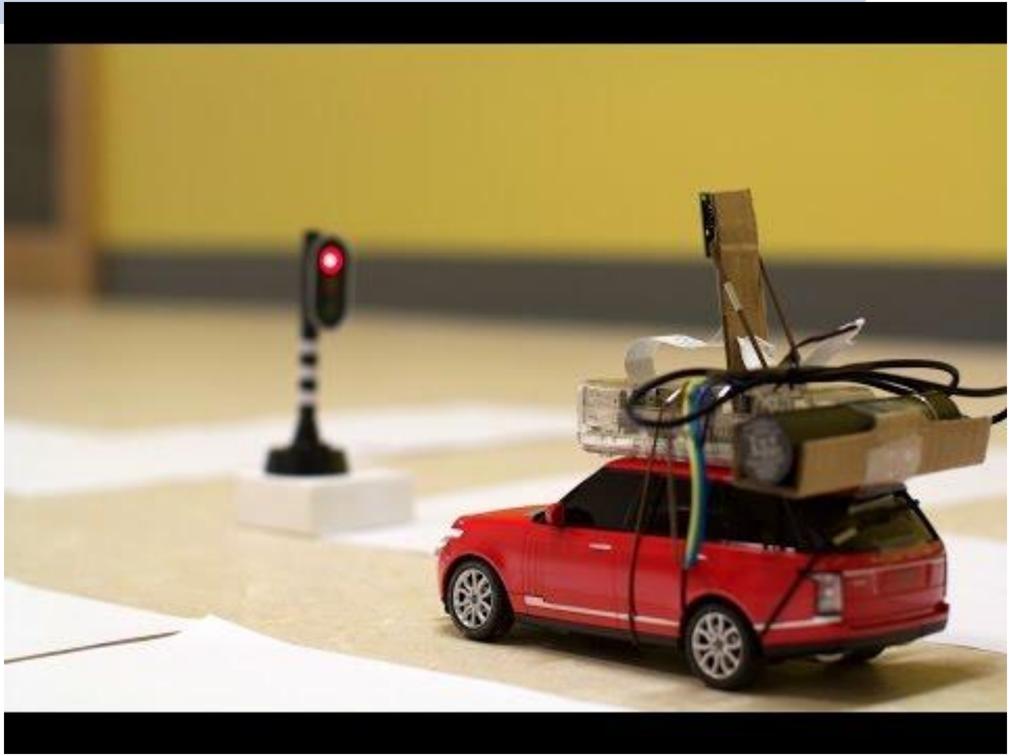
Weights of the model
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Values of the weights
0
20
40
60
80
100
Features
Bayesian Ridge estimate
Ground truth
OLS estimate

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
.02s
5.02s
.11s
.66s
.15s
.17s
.01s
.05s
.01s
.03s
5.58s
.08s
1.26s
.21s
.10s
.01s
.05s
.01s
.03s
3.30s
.13s
.22s
.82s
.62s
.02s
.05s
.02s
.03s
2.76s
.11s
.41s
.44s
.31s
.01s
.05s
.03s
.03s
2.52s
.08s
.34s
.21s
.11s
.02s
.05s
.01s
.02s
2.45s
.11s
.31s
.13s
.17s
.01s
.05s
.02s

Reinforcement Learning

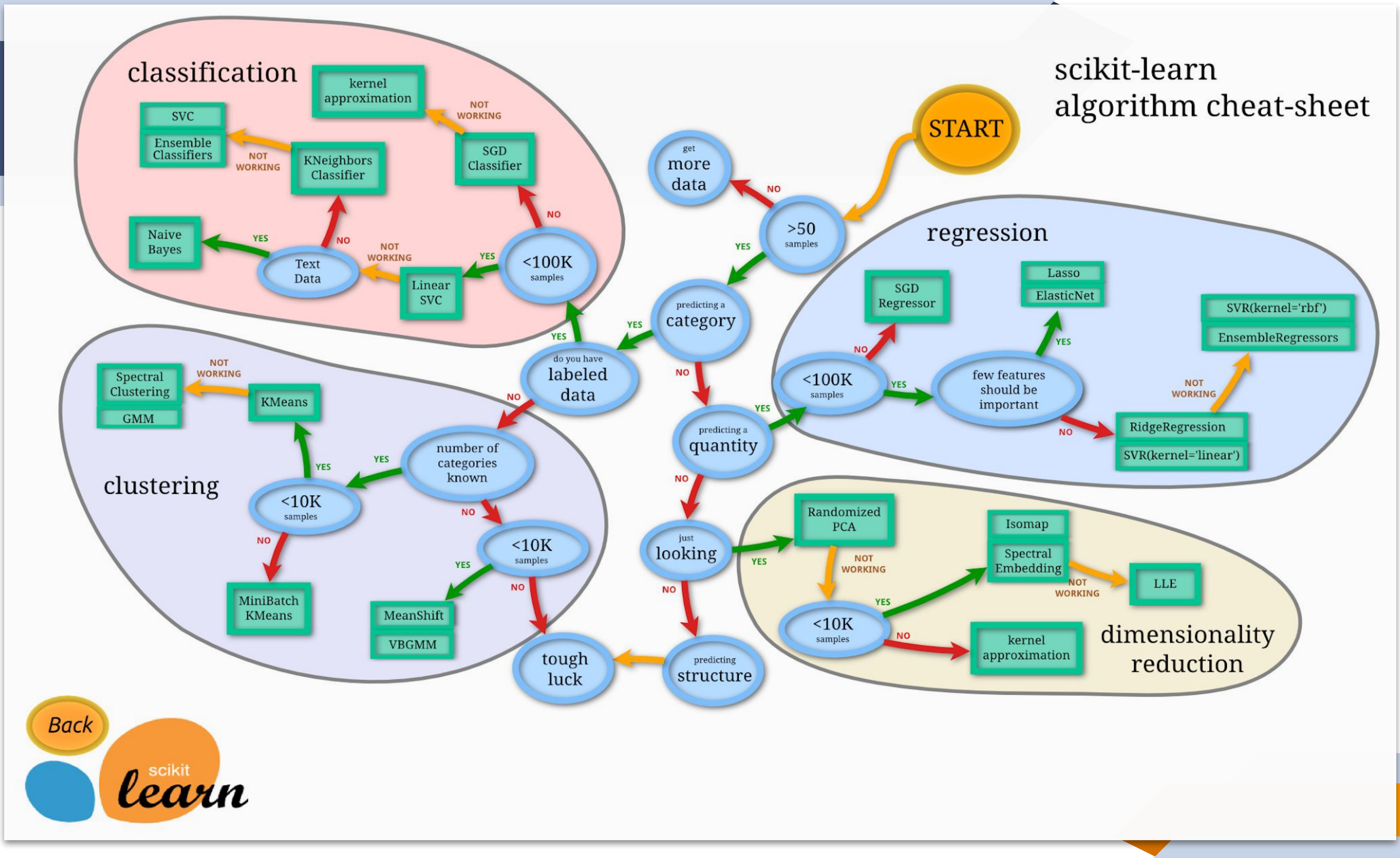


Reinforcement Learning

1.4

Algorithms

ขั้นตอนวิธีของการเรียนรู้ของเครื่อง


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

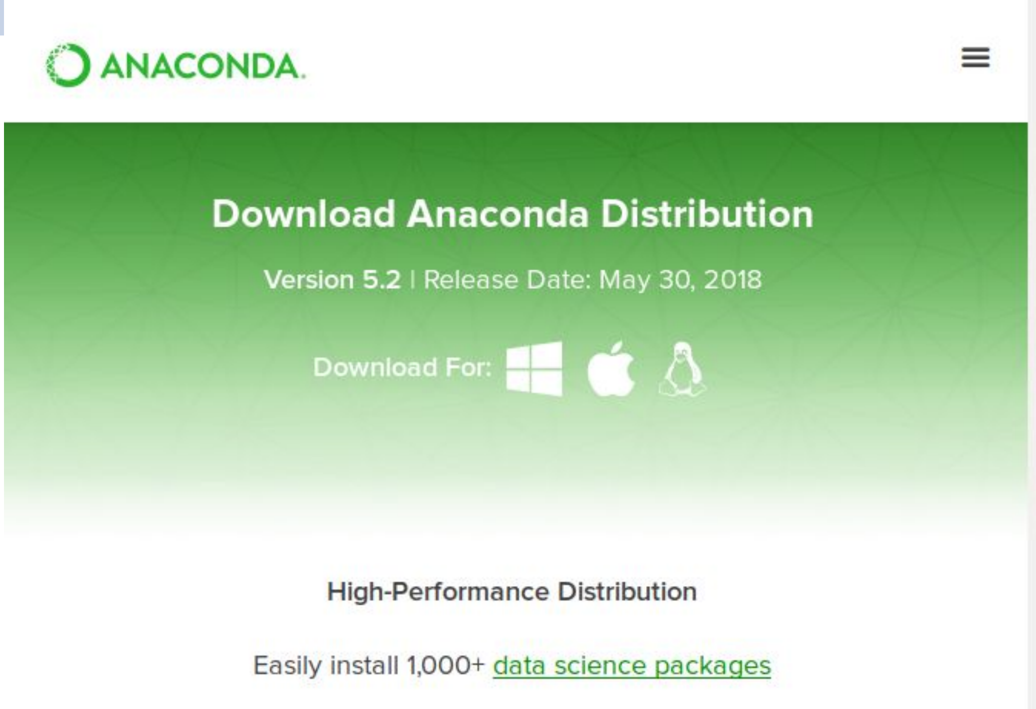
Visualization

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

Machine Learning

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

Software & Downloads



ANACONDA.

Download Anaconda Distribution

Version 5.2 | Release Date: May 30, 2018

Download For:

High-Performance Distribution

Easily install 1,000+ data science packages

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**