Big data and Data Science

Chantana Chantrapornchai
Dept. of Computer Engineering
Kasetsart University

KU KASETSART UNIVERSITY

NVIDIA GPU EDUCATION CENTER

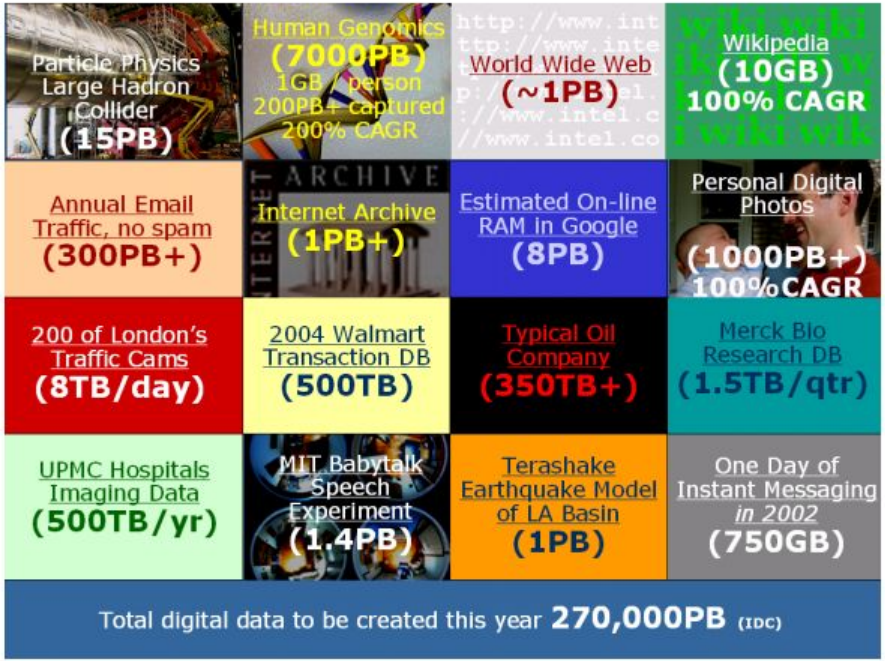
HPCNC

What is Big Data? - 3 V's

- Big Volume
  - With simple (SQL) analytics
  - With complex (non-SQL) analytics
- Big Velocity
  - How fast data is processed
- Big Variety
  - Large number of diverse data sources to integrate

http://whatis.techtarget.com/definition/3Vs

http://www.dummies.com/how-to/content/defining-big-data-volume-velocity-and-variety.html



Compound Annual Growth Rate: CAGR
http://www.investopedia.com/terms/c/cagr.asp

12 big facts about big data

1. If you stacked a pile of CD-ROMs on top of one another until you'd reached the current global storage capacity for digital information – about 295 exabytes – if would stretch 80,000 km beyond the moon. จำนวน storage ที่ใช้

2 Every hour, enough information is consumed by internet traffic to fill 7 million DVDs. Side by side, they'd scale Mount Everest 95 times. จำนวนข้อมูลที่ไหลบน Internet

3247 billion e-mail messages are sent each day… up to 80% of them are spam.

4 By 2020, IT departments will be looking after 10 x more servers, 50 x more data and 75 x more files. Meanwhile, the number of IT administrators keeping track of all that data growth with increase by 1.5 times. จำนวน server ที่ใช้

5 We can expect a 40-60 per cent projected annual growth in the volume of data generated, while media intensive sectors, including financial services, will see year on year data growth rates of over 120 per cent. ปริมาณที่เพิ่มขึ้นทุกปี

6The world's 500,000+ data centres are large enough to fill 5,955 football fields. ขนาด data center

775% of digital information is generated by individuals, whilst enterprises have liability for 80% of digital data at some point in its life.

8 There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

9 Investment in digital enterprises has increased 50% since 2005. การลงทุนในหน่วยงานเพิ่มขึ้น 50%

10 There are 30 billion pieces of content shared on Facebook every day. จำนวนข้อมูลที่ shared บน facebook ทุกวัน

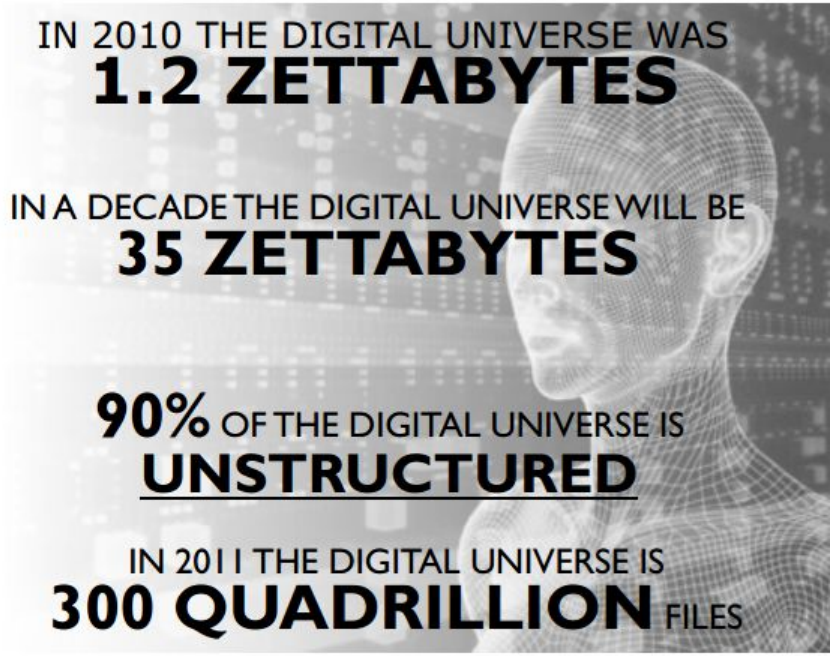
11 In 2010, 28% of the digital universe required some level of security… not all of it had the level of security it required….

12 People wishing each other Happy New Year drove a 500% surge in smartphone data within just one year, according to 3UK whose customers used a whopping 80 terabytes (TB) on the 31st December 2011, compared to just 14 TBs on the same day in 2010. ความต้องการ security

จำนวน ข้อมูลที่ส่งบน mobile ในวันปีใหม่

http://blog.kurtosys.com/12-big-facts-about-big-data/#.VBszDBY4S-J

The Economist, Feb 25, 2010



WIRED The New York Times Bloomberg Businessweek Forbes WALL STREET JOURNAL

http://www.snia.org/sites/default/files2/ABDS2012/Tutorials/RobPeglar_Introduction_Analytics%20_Big%20Data_Hadoop.pdf

How big?

- 1 ZB = 1000^7bytes = 10^{21} bytes = 1000000000000000000000 bytes = 1000 exabytes = 1 billion tera bytes.

Multiples of bytes v·t·e

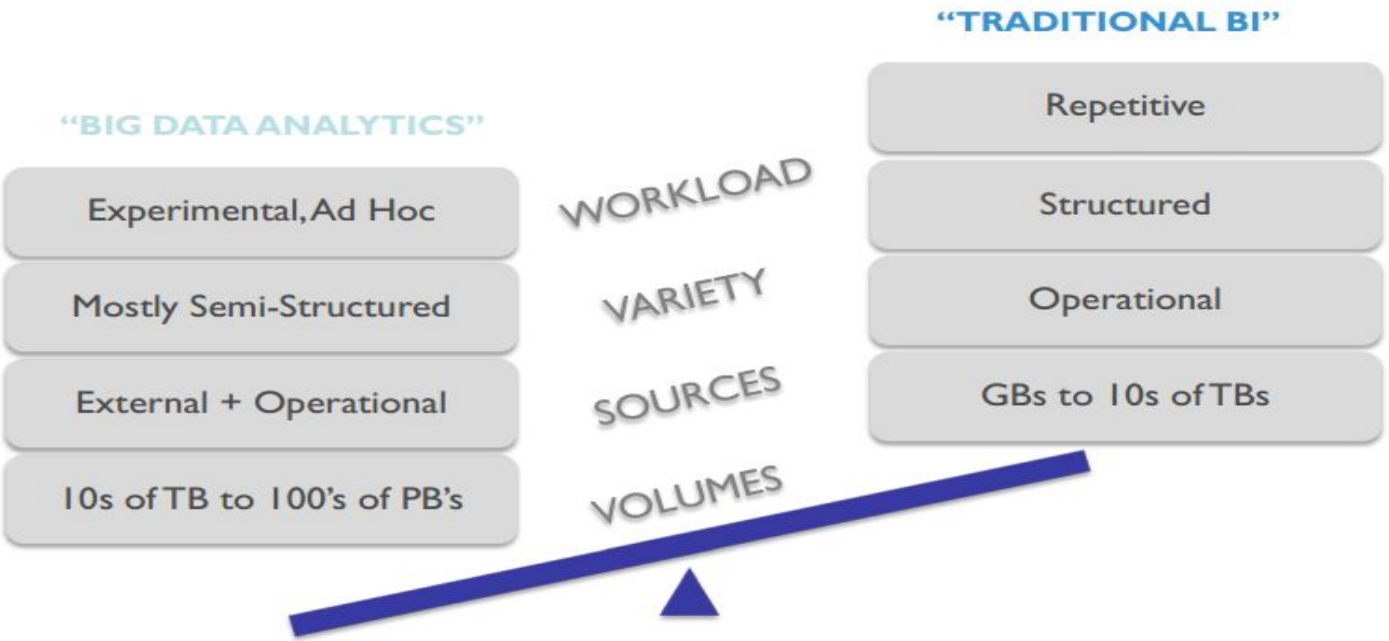
Decimal			Binary				
Value		Metric	Value		JEDEC		IEC
1000	kB	kilobyte	1024	KB	kilobyte	KiB	kibibyte
1000^2	MB	megabyte	1024^2	MB	megabyte	MiB	mebibyte
1000^3	GB	gigabyte	1024^3	GB	gigabyte	GiB	gibibyte
1000^4	TB	terabyte	1024^4	-	-	TiB	tebibyte
1000^5	PB	petabyte	1024^5	-	-	PiB	pebibyte
1000^6	EB	exabyte	1024^6	-	-	EiB	exbibyte
1000^7	ZB	**zettabyte**	1024^7	-	-	ZiB	zebibyte
1000^8	YB	yottabyte	1024^8	-	-	YiB	yobibyte

Orders of magnitude of data

http://en.wikipedia.org/wiki/Zettabyte


639,800 GB of global IP data transferred
20
New victims of identity theft
204 million
Emails sent
47,000
App downloads
$83,000
In sales
61,141
Hours of music
20 million
Photo views
3,000
Photo uploads
320+
New Twitter accounts
100,000
New tweets
135
Botnet infections
1,300
New mobile users
100+
New Linkedin accounts
6
New Wikipedia articles published
277,000
Logins
6 million
Facebook views
2+ million
Search queries
30
Hours of video uploaded
1.3 million
Video views
Email
Mobile
Apps
amazon
PANDORA
LinkedIn
flickr
twitter
facebook
Google
YouTube
intel
And Future Growth is Staggering
Today, the number of networked devices
=
the global
By 2015, the number of networked devices
=
2x
the global
In 2015, it would take you 5 years
IP
to view all video crossing IP networks each second

Processing big data: Differences from traditional BI



Ten commons applications

1. Modeling true risk
2. Customer churn analysis
3. Recommendation engine
4. Ad targeting
5. PoS transaction analysis
6. Analyzing network data to predict failure
7. Threat analysis
8. Trade surveillance
9. Search quality
10. Data "sandbox"

Financial Services 	Healthcare 
Retail 	Web/Social/Mobile 
Manufacturing 	Government 

Retail

- CRM – Customer Scoring
- Store Siting and Layout
- Fraud Detection / Prevention
- Supply Chain Optimization

Advertising & Public Relations

- Demand Signaling
- Ad Targeting
- Sentiment Analysis
- Customer Acquisition

Financial Services

- Algorithmic Trading
- Risk Analysis
- Fraud Detection
- Portfolio Analysis

Media & Telecommunications

- Network Optimization
- Customer Scoring
- Churn Prevention
- Fraud Prevention

Manufacturing

- Product Research
- Engineering Analytics
- Process & Quality Analysis
- Distribution Optimization

Energy

- Smart Grid
- Exploration

Government

- Market Governance
- Counter-Terrorism
- Econometrics
- Health Informatics

Healthcare & Life Sciences

- Pharmaco-Genomics
- Bio-Informatics
- Pharmaceutical Research
- Clinical Outcomes Research

Categories of Big Volumes

- With no analytics
- With analytics

Big volumes – no analytics

- Well addressed by data warehouse crowd
- Who are pretty good at SQL analytics on

— Hundreds of nodes

— Petabytes of data

- Column stores may be a good solution.
  - Factor of 50 or so faster than row stores

Big data- big analytics

- Complex math operations (machine learning, clustering, trend detection, ….)

— the world of the “quants” (Quantitative analyst)

— Mostly specified as linear algebra on array data

- A dozen or so common ‘inner loops’

— Matrix multiply

— QR decomposition

— SVD decomposition (Singular value decomposition)

— Linear regression

http://www.forbes.com/sites/investopedia/2013/06/07/quants-the-rocket-scientists-of-wall-street/

Big Velocity

การซื้อขายในตลาดหุ้น

- Trading volumes going through the roof on Wall Street

– breaking infrastructure

- Sensor tagging of {cars, people, …} creates a firehose to ingest
- The web empowers end users to submit transactions

– sending volume through the roof

- PDAs lets them submit transactions from anywhere….

Big Variety

Typical enterprise has 5,000 operational systems

— Only a few get into the data warehouse

— What about the rest?

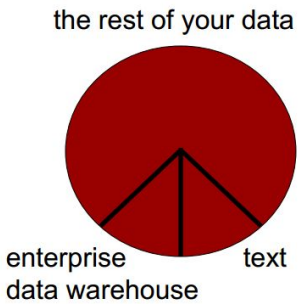
And what about all the rest of your data?

— Spreadsheets

— Access databases

— Web pages

• And public data from the web?



All data integrations

Information History

Business Intelligence.
Great variety of visual resources
to analyze data

Beginning of the use of DBs
and basic reports

Data was not stored

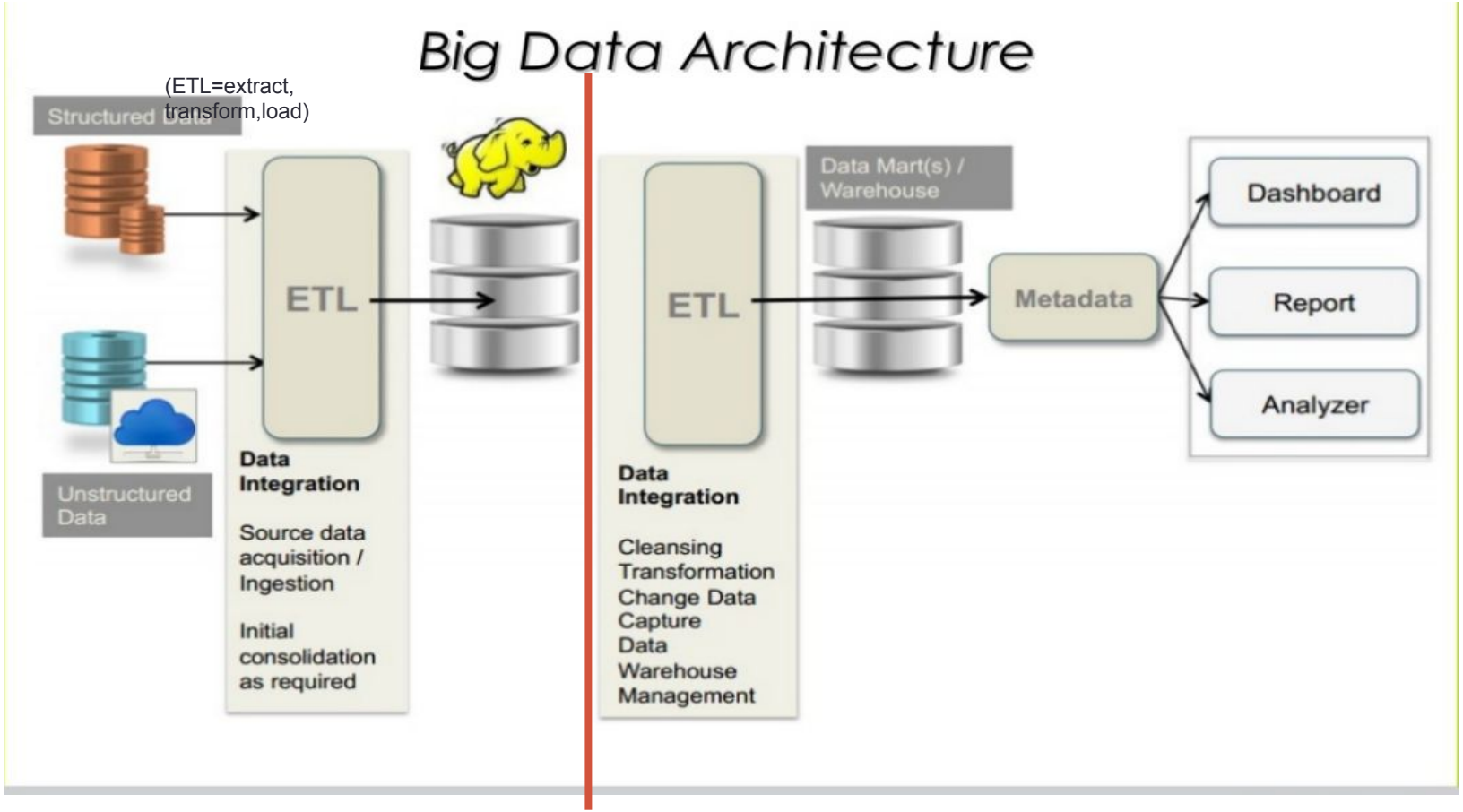
http://www.slideshare.net/zanorte/big-data-analytics-2013

Data analysis profits:

- *Competitive advantages*
- *Customer satisfaction evaluation*
- *Business process improvement*
- *Increase sales*
- …

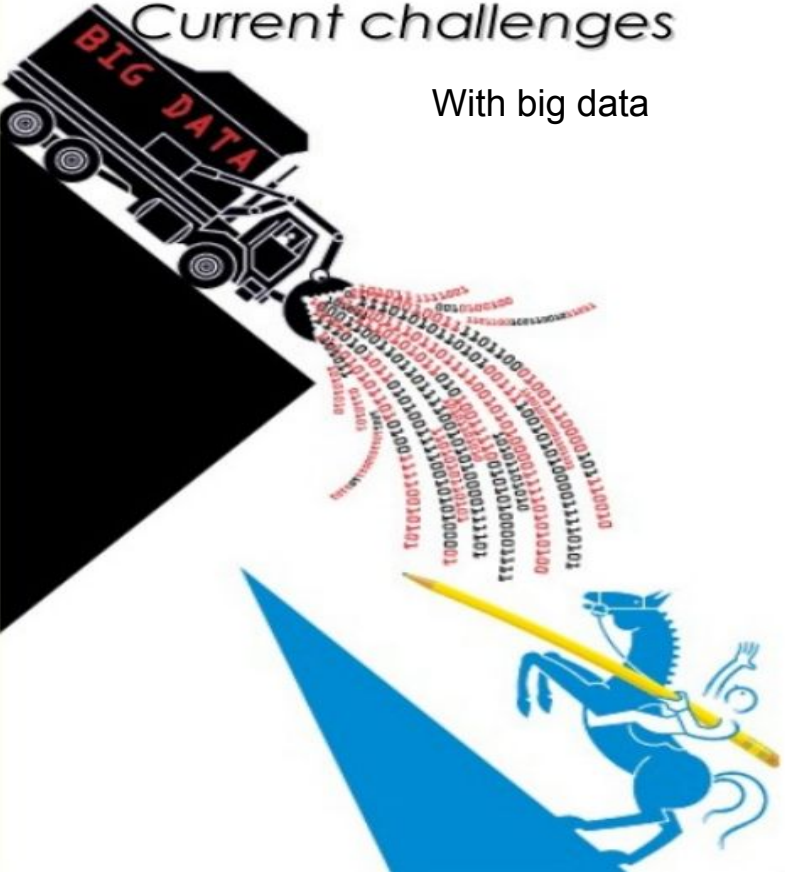
New data analysis techniques and processes

- New BI solutions
- New visual resources
- New data sources
- Cloud solutions
- Latest trends
  - Social Intelligence
  - Mailing intelligence
- …

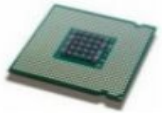

Big Data Architecture
(ETL=extract, transform,load)
Structured Data
Unstructured Data
ETL
Data Integration
Source data acquisition / Ingestion
Initial consolidation as required
ETL
Data Integration
Cleansing
Transformation
Change Data Capture
Data Warehouse Management
Data Mart(s) / Warehouse
Metadata
Dashboard
Report
Analyzer

Current challenges

With big data



- Scalability
  - Vertical
    - + CPU
    - + RAM



  - Horizontal
    - More nodes

- Data types
  - Structured
  - Unstructured

Data types

- Structured

A data structure is a particular way of storing and organizing data in a computer so that it can be used efficiently.
List: http://en.wikipedia.org/wiki/List_of_data_structures
Primitive data types: Boolean, chart, float, double …

- Unstructured

Unstructured information refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner.

semi-structure….

Types of Big Data DBs. Not Only SQL (NoSQL)

- *In response to these problems a NoSQL paradigm appeared.*
- *NoSQL* ***is not a substitute*** *for relational databases*
  - *Instead it is used in other* ***specific scenarios***
- *Not all problems can be solved using a RDBMS*
- *Developer has a* ***range of possibilities*** *and can select the best to deal with a specific problem*
- *There are several NoSQL systems focusing on typical issues (scaling, increasing performance…) in a different way*

Example of NoSQL

Document-oriented *databases*

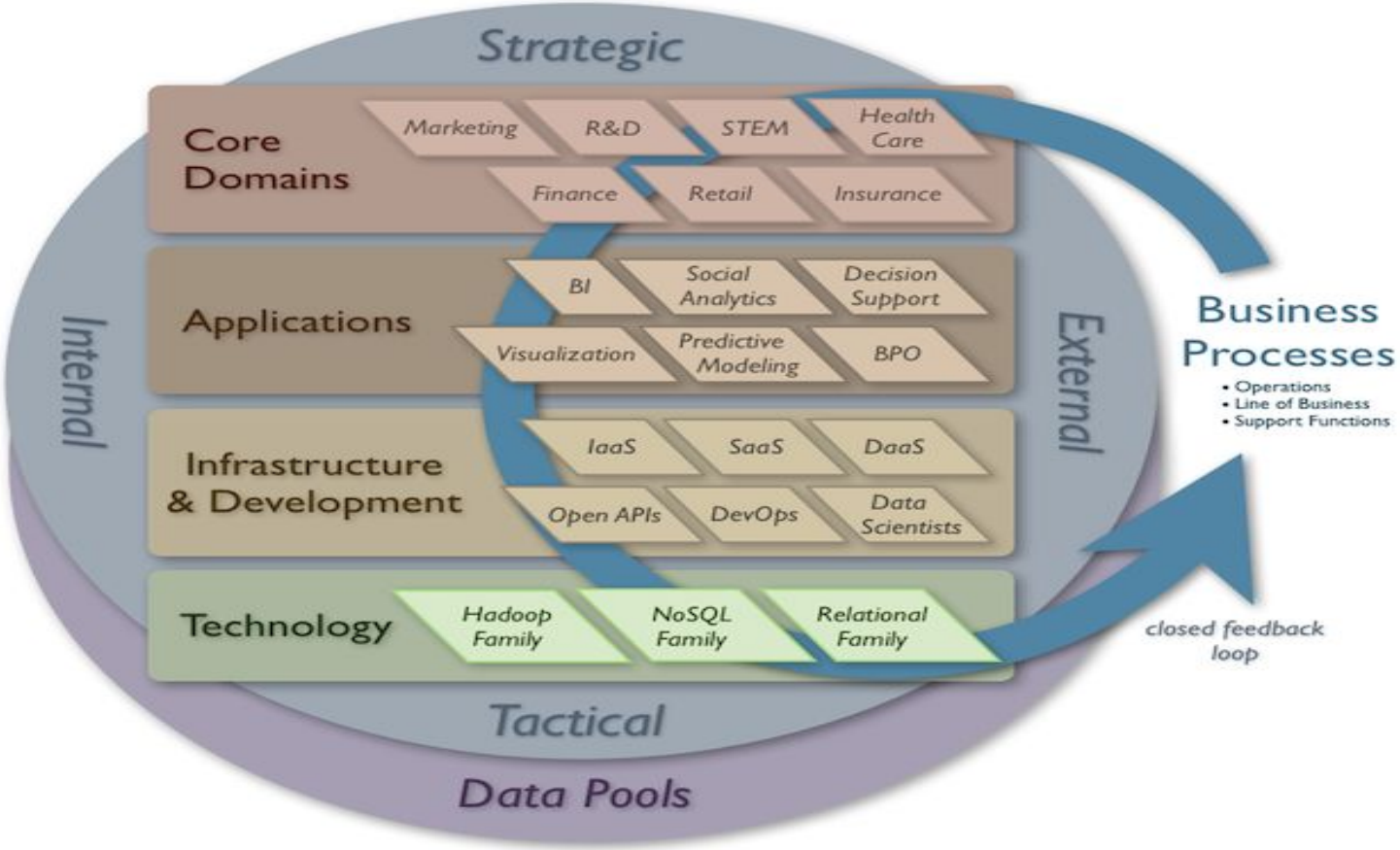
Columnar *databases*

Object oriented *databases*

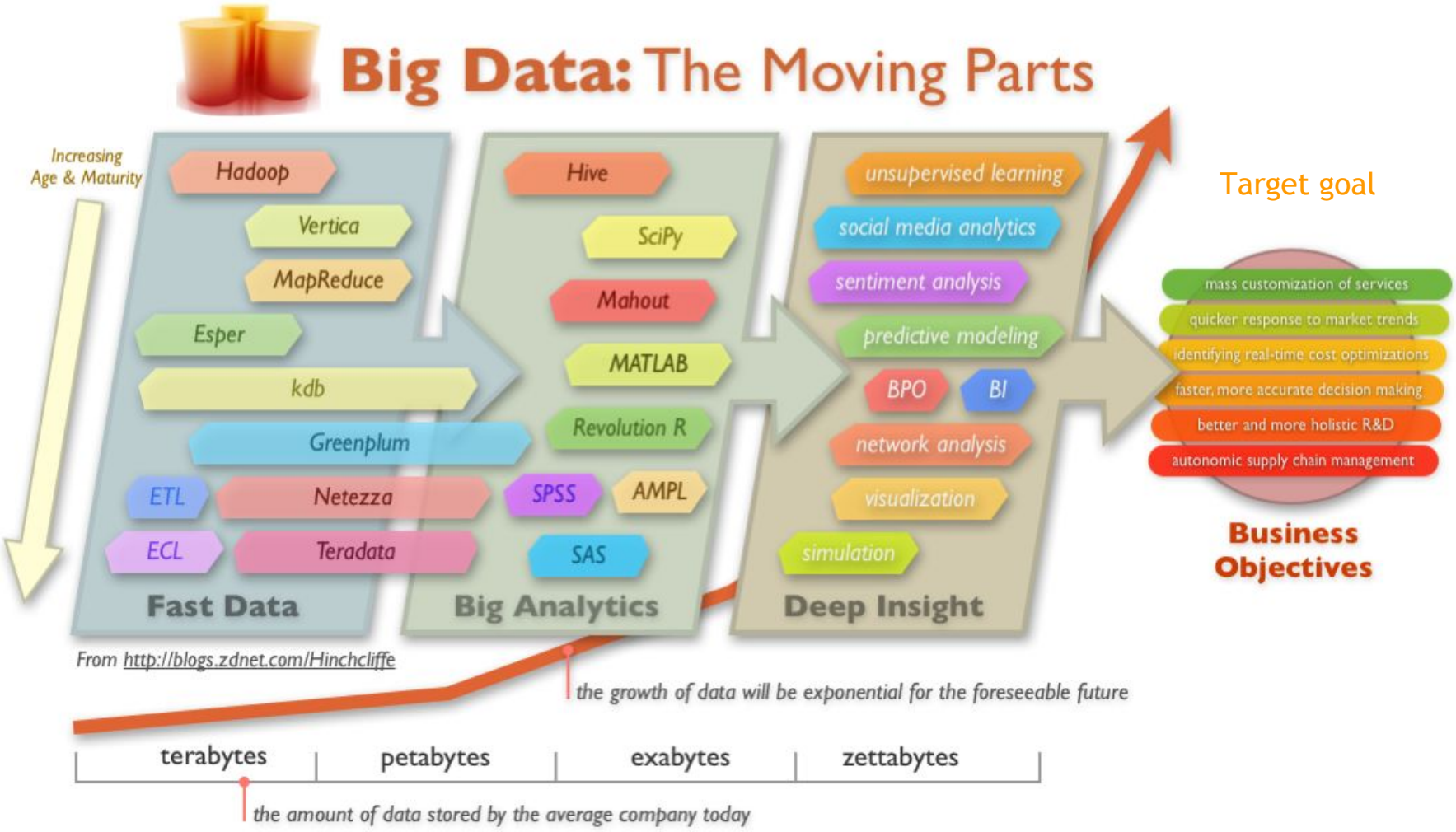
Graph *databases*

Do not replace relational model. Specific scenarios.

Big Data in the Enterprise



From http://zdnet.com/blog/hinchcliffe on ZDNet.


Big Data: The Moving Parts
Increasing Age & Maturity
Hadoop
Vertica
MapReduce
Esper
kdb
Greenplum
ETL
Netezza
ECL
Teradata
Fast Data
Hive
SciPy
Mahout
MATLAB
Revolution R
SPSS
AMPL
SAS
Big Analytics
unsupervised learning
social media analytics
sentiment analysis
predictive modeling
BPO
BI
network analysis
visualization
simulation
Deep Insight
Target goal
mass customization of services
quicker response to market trends
identifying real-time cost optimizations
faster, more accurate decision making
better and more holistic R&D
autonomic supply chain management
Business Objectives
From http://blogs.zdnet.com/Hinchcliffe
the growth of data will be exponential for the foreseeable future
terabytes
petabytes
exabytes
zettabytes
the amount of data stored by the average company today

Data Science

What is data science?

-Data Science is a field that comprises of everything that related to data cleansing, preparation, and analysis.

-Dealing with unstructured and structured data

Big Data: Big Data is humongous volumes of data that cannot be processed effectively with the traditional applications that exist.

Data Analytics: The science of examining raw data with the purpose of drawing conclusions.

It involves applying an algorithmic or mechanical process to derive insights.

Applications of data science

1. Internet search
2. Digital advertisement
3. Recommendation

Application of Big data

Big data in communication: Gaining new subscribers, retaining customers, and expanding within current subscriber bases

Big Data for financial services: Credit card companies, retail banks, private wealth management advisories, insurance firms, venture funds, and institutional investment banks use big data for their financial services.

Applications of Data Analysis:

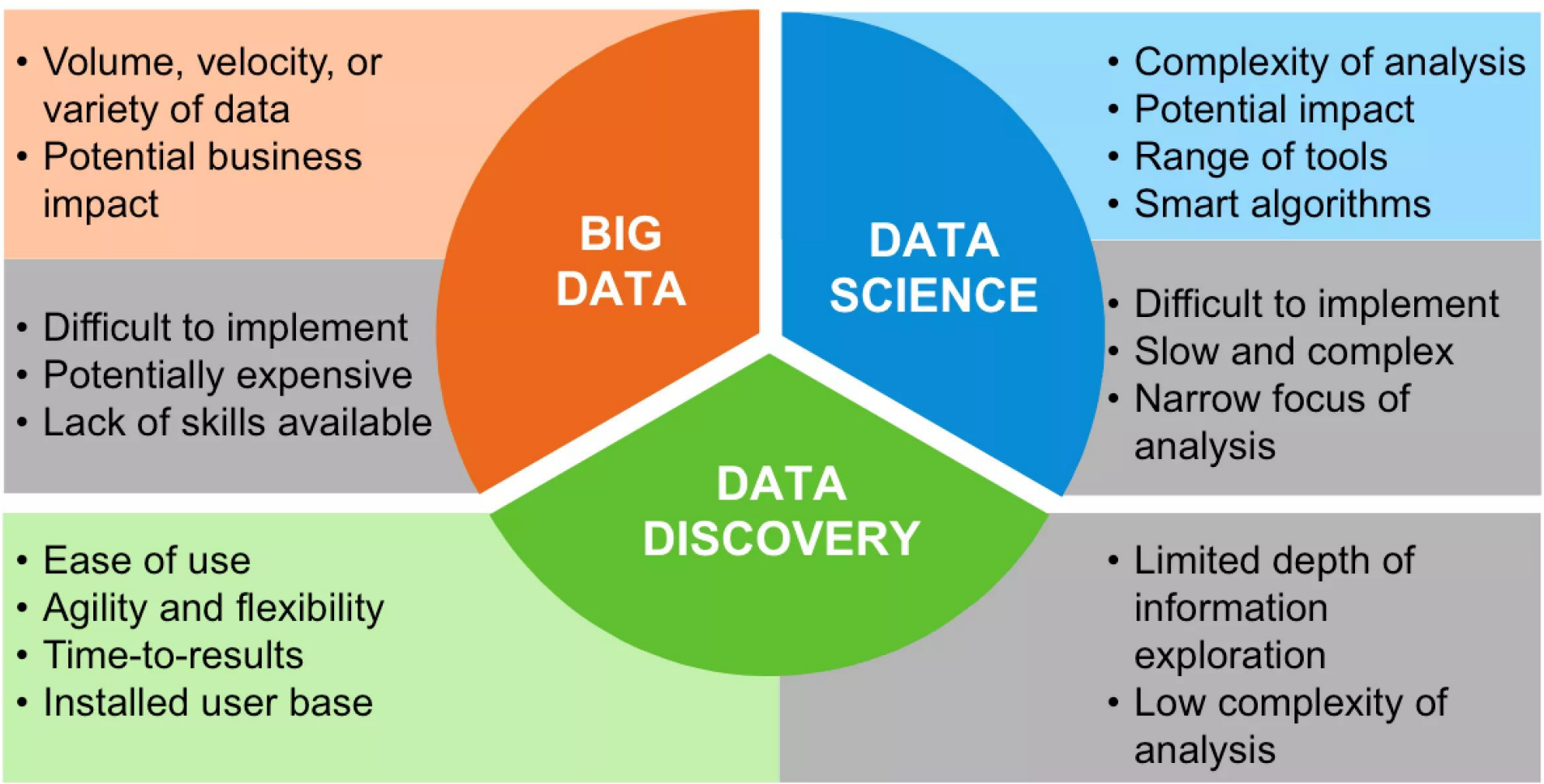
-Healthcare

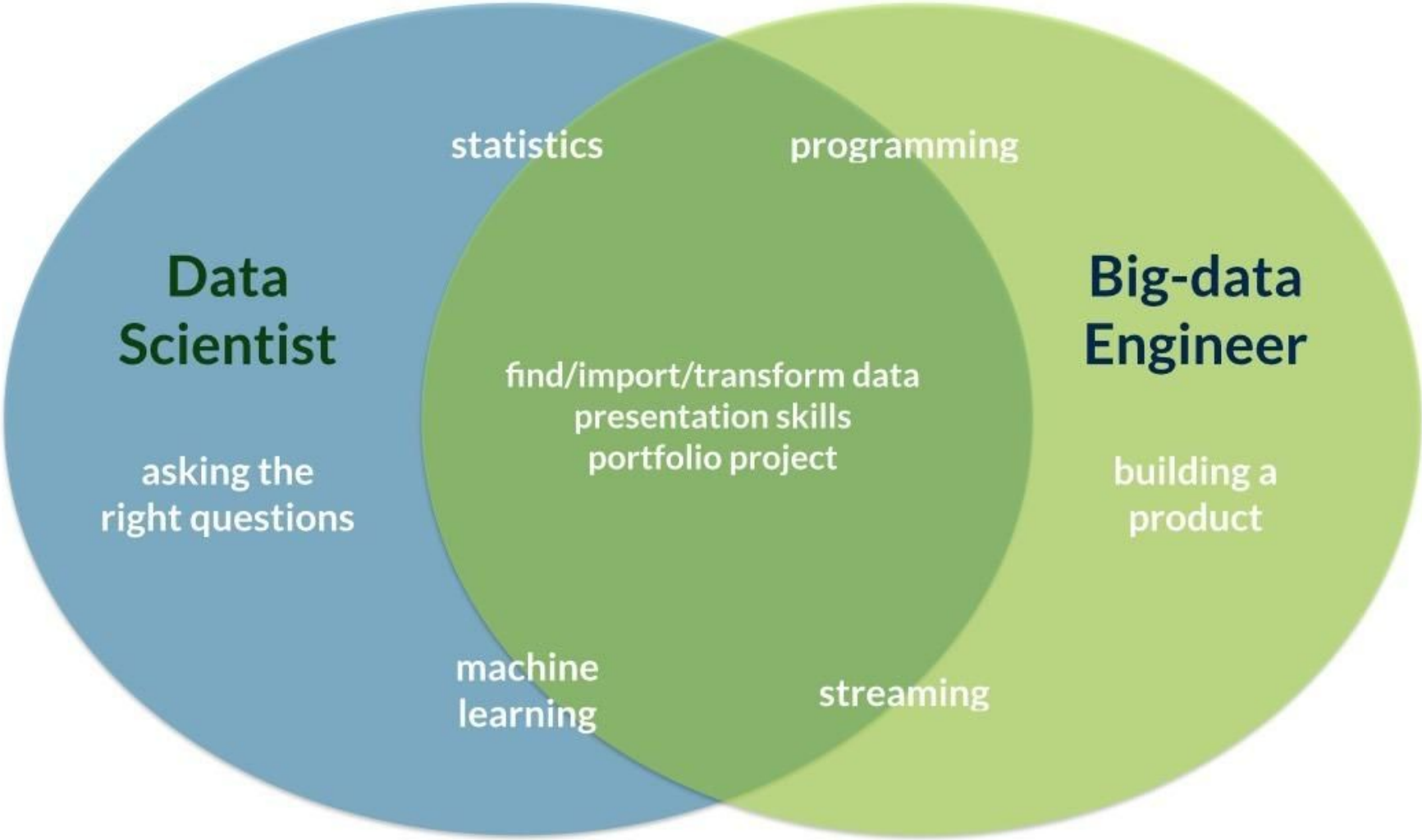
-Tourism

-Gaming

-Energy management

https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article


• Volume, velocity, or variety of data
• Potential business impact
BIG DATA
• Difficult to implement
• Potentially expensive
• Lack of skills available
• Complexity of analysis
• Potential impact
• Range of tools
• Smart algorithms
DATA SCIENCE
• Difficult to implement
• Slow and complex
• Narrow focus of analysis
DATA DISCOVERY
• Ease of use
• Agility and flexibility
• Time-to-results
• Installed user base
• Limited depth of information exploration
• Low complexity of analysis


Data Scientist
asking the right questions
statistics
machine learning
programming
streaming
find/import/transform data
presentation skills
portfolio project
Big-data Engineer
building a product


Parallel Data Analytics
Data Mining Methods
Machine Learning Algorithms
Data Science
Scientific Community Applications
Data Analysis Tools
Generic Data Methods

Our case study: IoT data

Monitoring inefficiency of energy consumption in data center.

Chantana Chantrapornchai

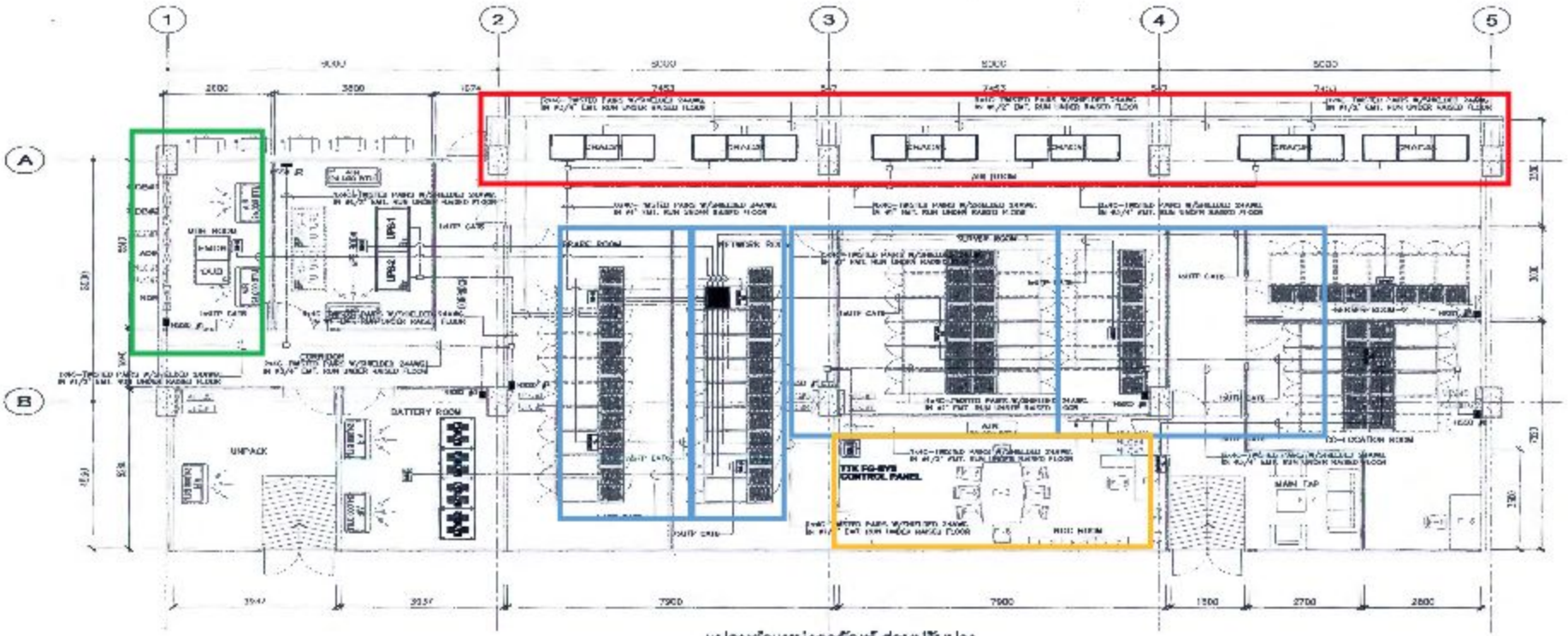
HPCNC Laboratory, Dept. of Computer Engineering

Kasetsart University

Withit Chatlatanakulchai, et.al.

CRVLAB , Dept. of Mechanical Engineering

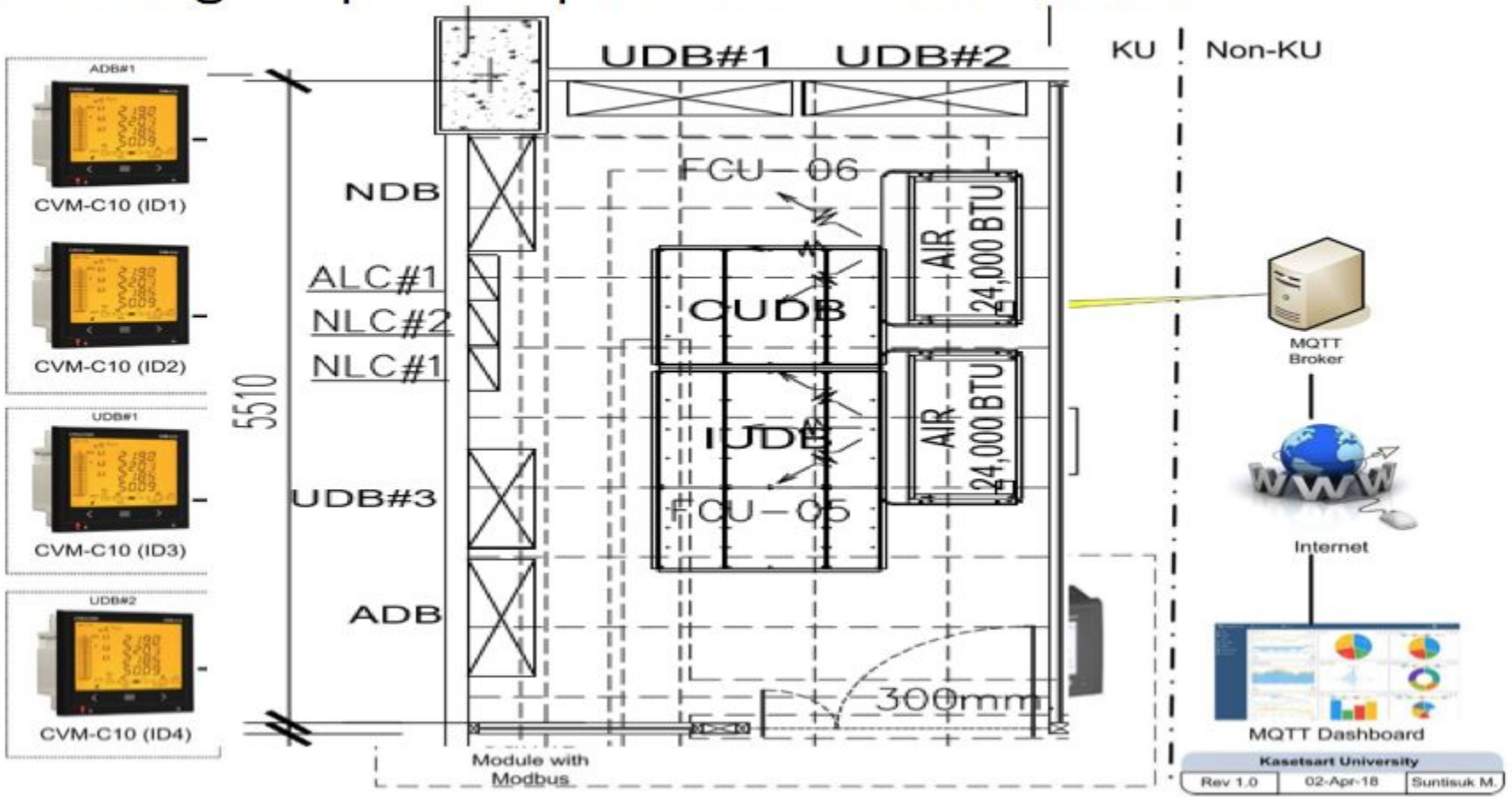
Kasetsart University



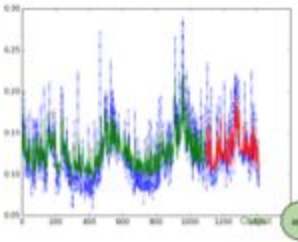
แปลนตำแหน่งครุภัณฑ์ ส่วนปรับปรุง

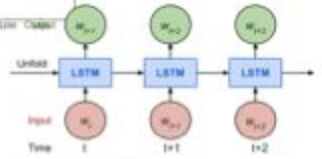
PROJECT: ปรับปรุงห้องข้อมูลอากาศและรังสี ชั้น 6 ที่ทำการบริการวิชาการสิ่งแวดล้อม มหาวิทยาลัยเกษตรศาสตร์	OWNER: สำนักบริการวิชาการสิ่งแวดล้อม มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน	TCS	AS BUILT DRAWING	REV.	DATE	DESCRIPTION	TITLE: ENVIRONMENT MONITORING SYSTEM LAYOUT DRAWN BY: S.Sompech ENGINEER: L.Wan DATE: 3 Feb. 2014	DWG. NO. PS-0

Measuring scope and power meter installation.



Relevant data integration





json

2) Temperature data outside (Webservice)

ERP 4) Power bill database

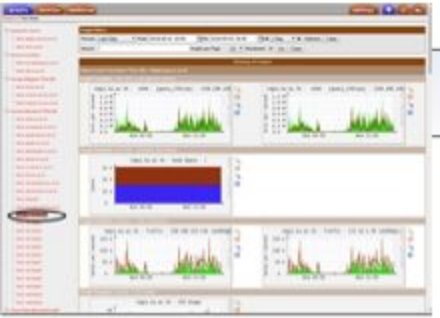
Analytic/ Prediction

1) Power data CSV. json

Database (key,value)

Visualization

Cloud storage

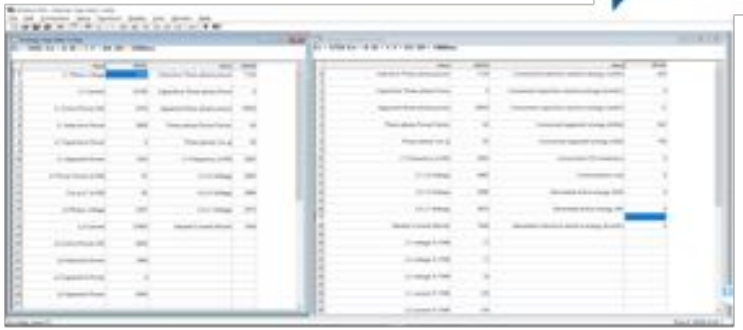


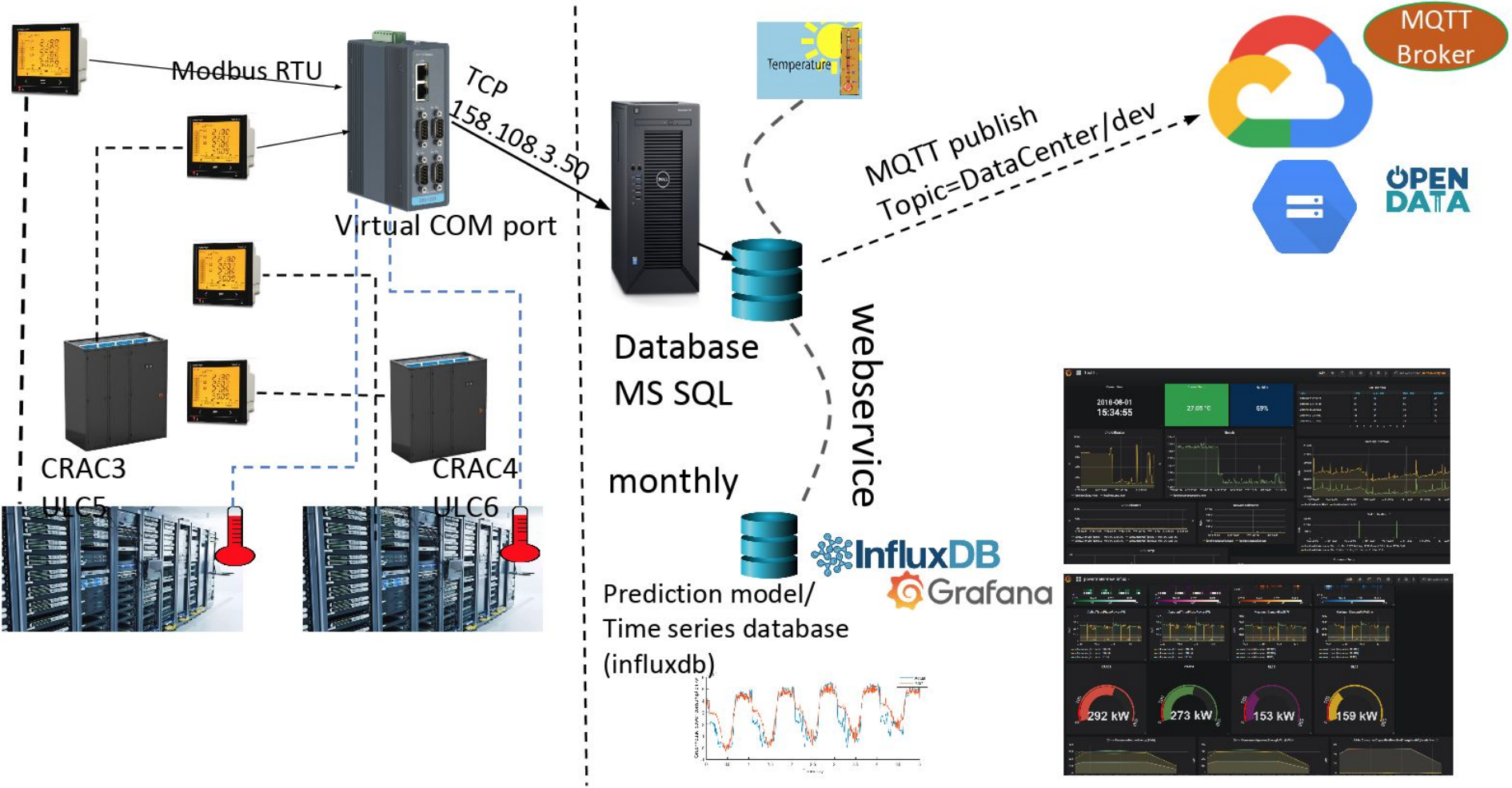
3) Load server data database

Streaming

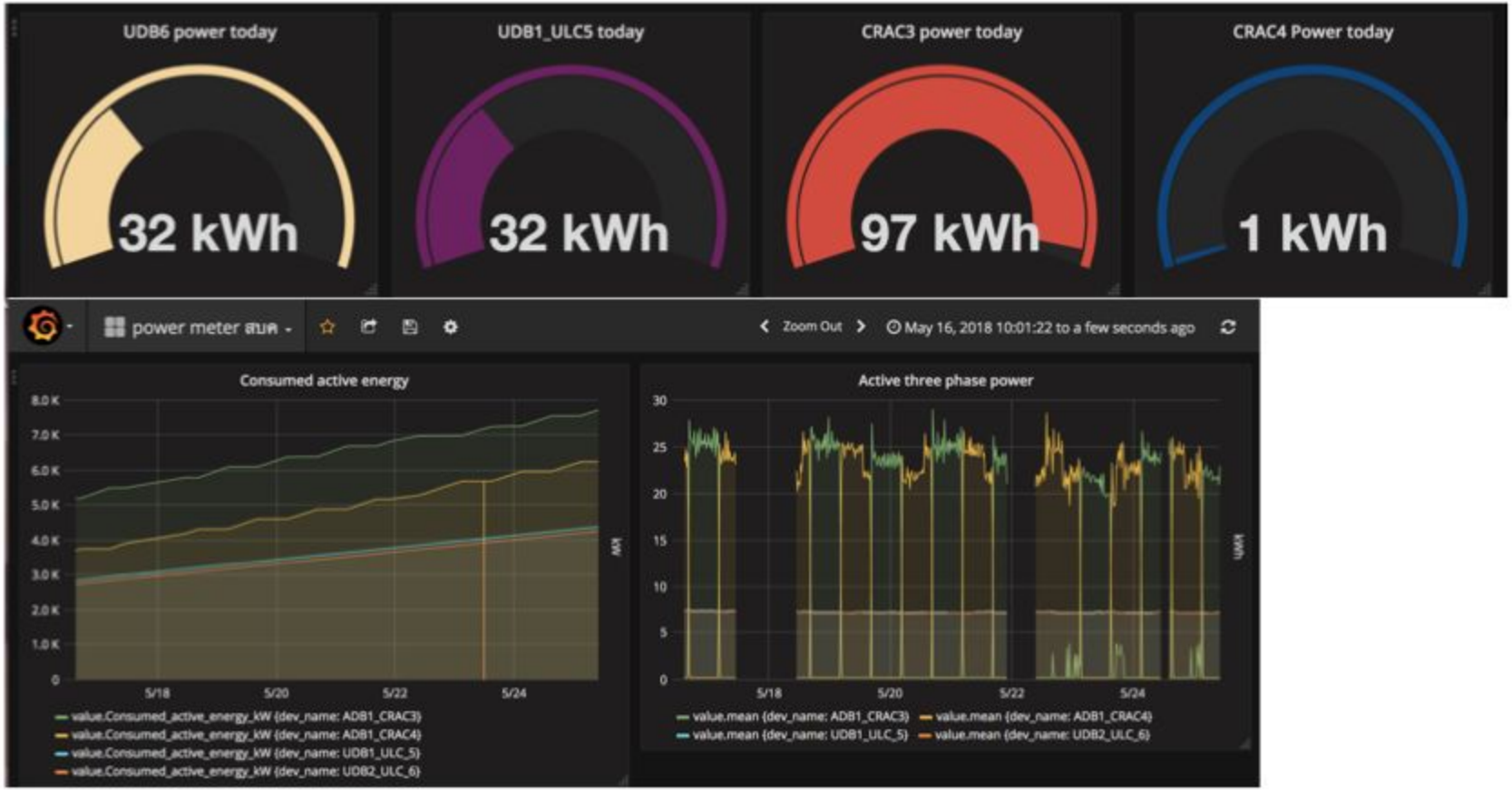
Manual keyin/Database retreival

Meter installation



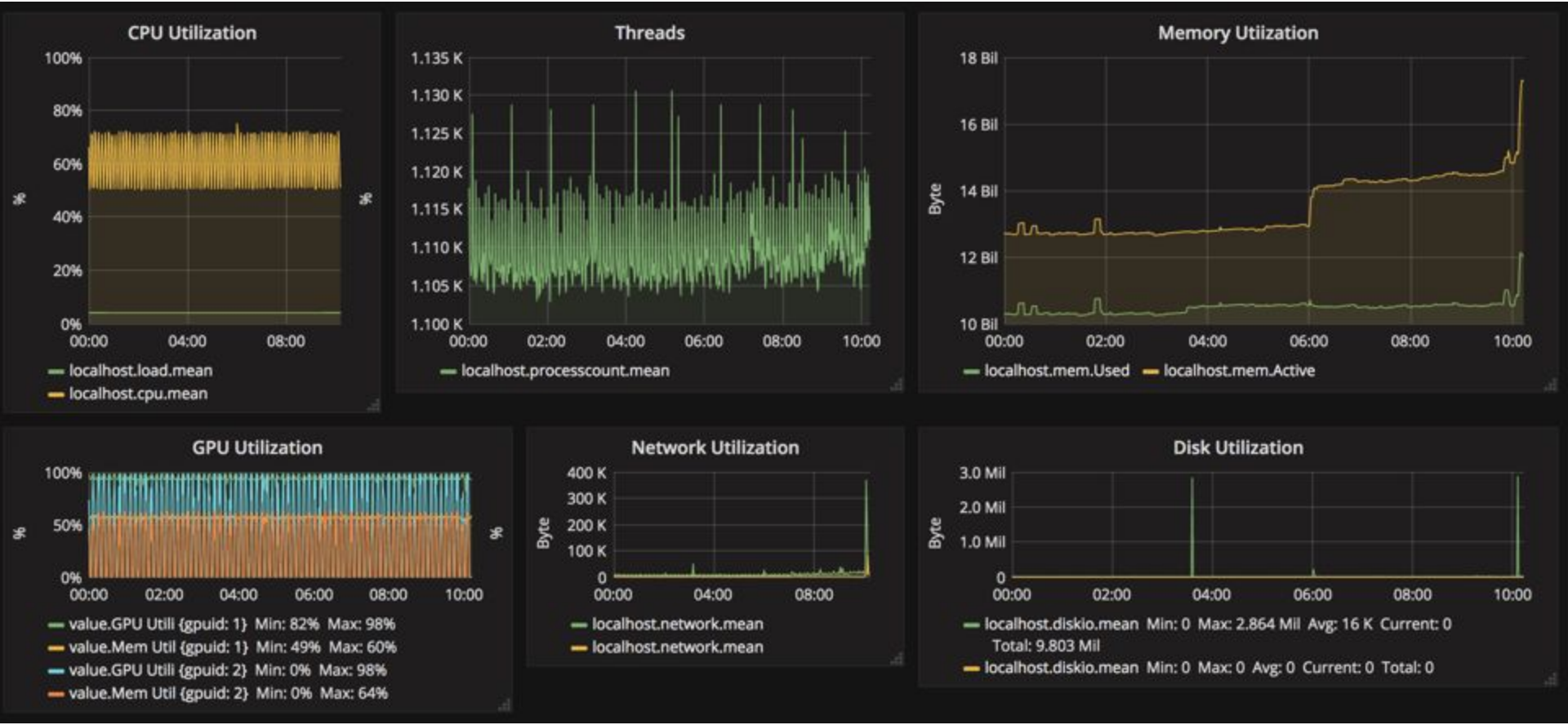

Modbus RTU
TCP
158.108.3.50
Virtual COM port
Temperature
MQTT publish
Topic=DataCenter/dev
MQTT Broker
OPEN DATA
Database
MS SQL
webservice
monthly
CRAC3
ULC5
CRAC4
ULC6
InfluxDB
Grafana
Prediction model/
Time series database
(influxdb)
292 kW
273 kW
153 kW
159 kW

Visualization


UDB6 power today
32 kWh
UDB1_ULC5 today
32 kWh
CRAC3 power today
97 kWh
CRAC4 Power today
1 kWh
power meter สบค
Zoom Out
May 16, 2018 10:01:22 to a few seconds ago
Consumed active energy
8.0 K
7.0 K
6.0 K
5.0 K
4.0 K
3.0 K
2.0 K
1.0 K
0
5/18
5/20
5/22
5/24
kW
value.Consumed_active_energy_kW {dev_name: ADB1_CRAC3}
value.Consumed_active_energy_kW {dev_name: ADB1_CRAC4}
value.Consumed_active_energy_kW {dev_name: UDB1_ULC_5}
value.Consumed_active_energy_kW {dev_name: UDB2_ULC_6}
Active three phase power
30
25
20
15
10
5
0
5/18
5/20
5/22
5/24
kWh
value.mean {dev_name: ADB1_CRAC3}
value.mean {dev_name: ADB1_CRAC4}
value.mean {dev_name: UDB1_ULC_5}
value.mean {dev_name: UDB2_ULC_6}


Power Meter New
(ConsumedActiveEnergykW_kWh+ConsumedActiveEnergyW_Wh)
30 K
25 K
20 K
15 K
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
Watt
CRAC3 Max: 28.03 K
CRAC4 Max: 25.39 K
ULC5 Max: 15.84 K
ULC6 Max: 15.77 K
ActiveThreePhasePower_W
50 K
40 K
30 K
20 K
10 K
0
Watt
CRAC3
CRAC4
ULC5
ULC6
CRAC4Avg15min
CRAC3Avg15min
ULC5Avg15min
ULC6Avg15min
ConsumedCapacitiveReactiveEnergyvarhC_varh
3.0 K
2.5 K
2.0 K
1.5 K
1.0 K
500
0
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
CRAC3
CRAC4
ULC5
ULC6
MaximumDemandkWIII_W
30 K
25 K
20 K
15 K
10 K
5 K
0
MaximumDemandkVAIII
30 K
25 K
20 K
15 K
10 K
5 K
ThreePhasePowerFactorx100
150
100
50
0
-50
-100


Test0
Zoom Out
Today so far
Refresh every 30s
Current time
2018-05-25
10:11:18
Current Temp
26.00 °C
Humidity
79%
Past Temp Value
Time
Temp
Min Temp
Max Temp
humidity
2018-05-25 09:51:42 31 30 31 79
2018-05-25 09:27:13 29 29 29 70
2018-05-25 08:51:42 27 27 28 83
2018-05-25 08:27:13 27 27 27 88
2018-05-25 07:51:42 27 27 27 88
1 2 3 4 5 6 7 8 9
CPU Utilization
100%
80%
60%
40%
20%
0%
%
00:00
04:00
08:00
Threads
1.135 K
1.130 K
1.125 K
1.120 K
1.115 K
1.110 K
1.105 K
1.100 K
00:00
02:00
04:00
06:00
08:00
10:00
Memory Utilization
18 Bil
16 Bil
14 Bil
12 Bil
10 Bil
Byte
00:00
02:00
04:00
06:00
08:00
10:00


CPU Utilization
100%
80%
60%
40%
20%
0%
%
%
00:00
04:00
08:00
localhost.load.mean
localhost.cpu.mean
Threads
1.135 K
1.130 K
1.125 K
1.120 K
1.115 K
1.110 K
1.105 K
1.100 K
00:00
02:00
04:00
06:00
08:00
10:00
localhost.processcount.mean
Memory Utiization
18 Bil
16 Bil
14 Bil
12 Bil
10 Bil
Byte
00:00
02:00
04:00
06:00
08:00
10:00
localhost.mem.Used
localhost.mem.Active
GPU Utilization
100%
50%
0%
%
%
00:00
02:00
04:00
06:00
08:00
10:00
value.GPU Utili {gpuid: 1} Min: 82% Max: 98%
value.Mem Util {gpuid: 1} Min: 49% Max: 60%
value.GPU Utili {gpuid: 2} Min: 0% Max: 98%
value.Mem Util {gpuid: 2} Min: 0% Max: 64%
Network Utilization
400 K
300 K
200 K
100 K
0
Byte
00:00
04:00
08:00
localhost.network.mean
localhost.network.mean
Disk Utilization
3.0 Mil
2.0 Mil
1.0 Mil
0
Byte
00:00
02:00
04:00
06:00
08:00
10:00
localhost.diskio.mean Min: 0 Max: 2.864 Mil Avg: 16 K Current: 0
Total: 9.803 Mil
localhost.diskio.mean Min: 0 Max: 0 Avg: 0 Current: 0 Total: 0

Let’s try out the data

Data Science Process

Data cleansing

Finding relationship

Modeling

Checking accuracy

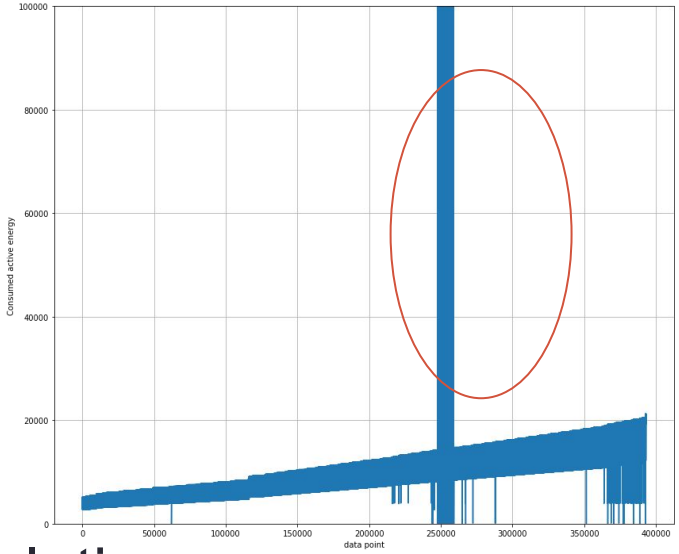
Visualization

Cleansing

Remove noises

-Constraints (Max,Min)

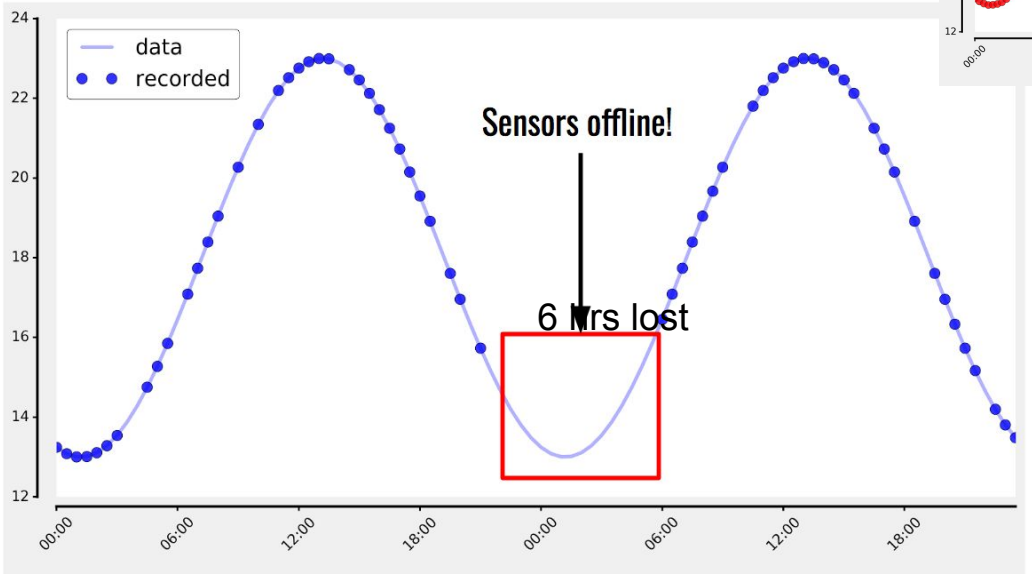
-monotonically increasing

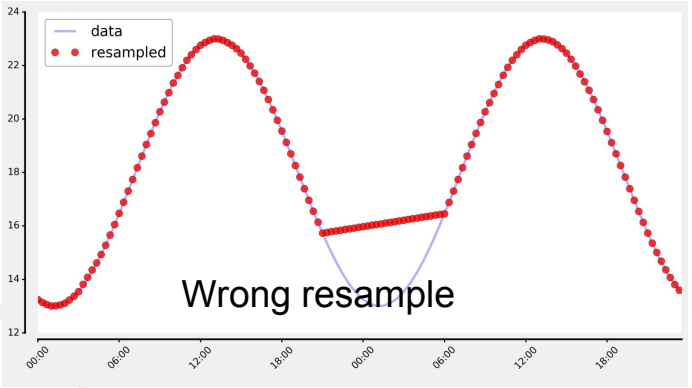


Missing data with resampling/interpolation

-Upsampling : Min -> Second

-Downsampling: Min -> Year


data
recorded
Sensors offline!
6 Hrs lost
24
22
20
18
16
14
12
00:00
06:00
12:00
18:00
00:00
06:00
12:00
18:00


data
resampled
Wrong resample
24
22
20
18
16
14
12
00:00
06:00
12:00
18:00
00:00
06:00
12:00
18:00

original

Timestamp	Consumed_active_energy_kW	Consumed_apparent_energy_kVAh \
2018-05-16 14:41:16	5180.0	5413.0
2018-05-16 14:41:46	5180.0	5413.0
2018-05-16 14:42:16	5180.0	5413.0
2018-05-16 14:42:46	5180.0	5413.0

Every 2 m

2018-07-06 05:48:00	5254.000000
2018-07-06 05:50:00	5254.000000
2018-07-06 05:52:00	5254.000000
2018-07-06 05:54:00	5255.000000

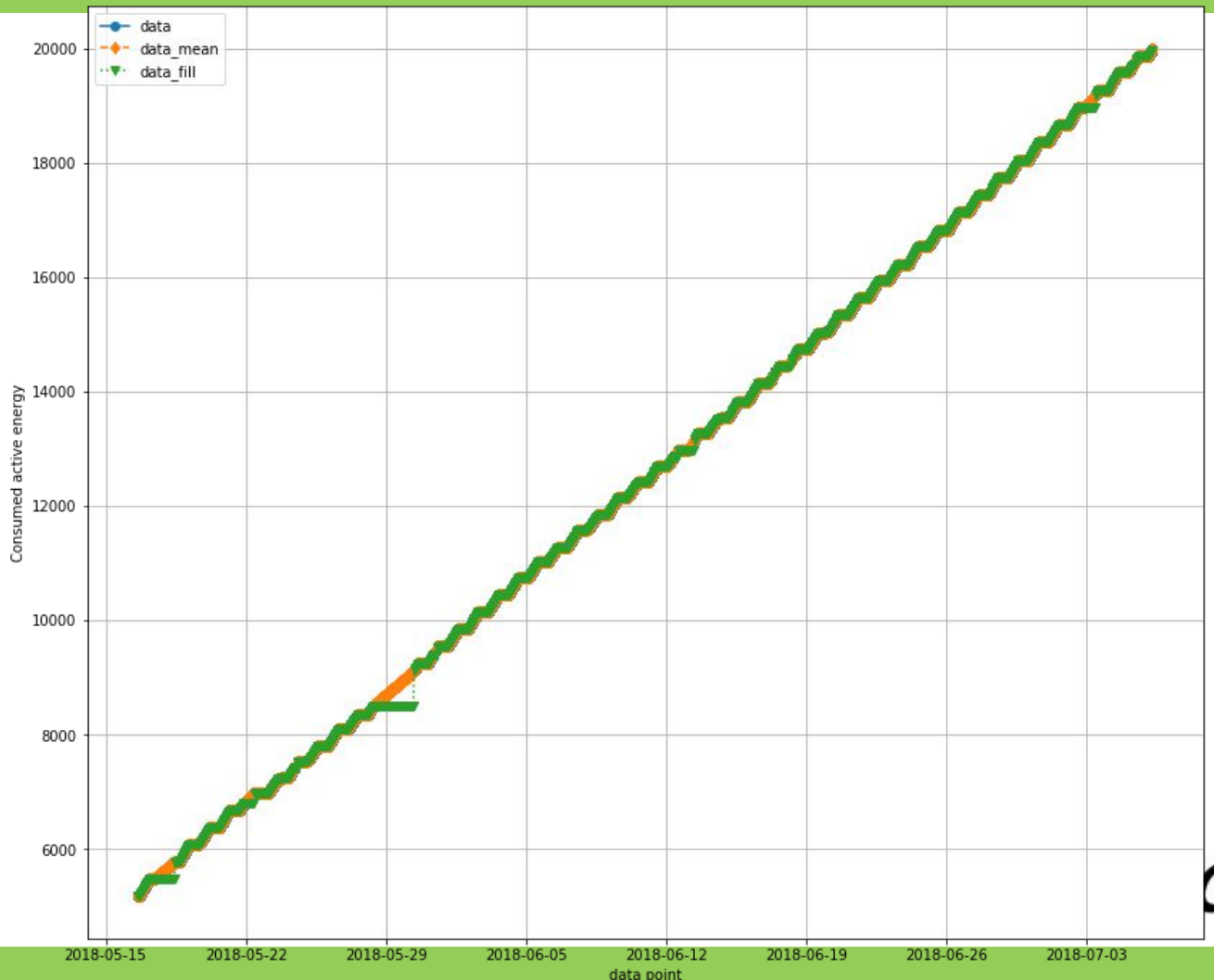
https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.resample.html

Every 50min

2018-05-16 17:30:00	5205.0	5439.0
2018-05-16 18:20:00	5227.0	5461.0
2018-05-16 19:10:00	5248.0	5483.0
2018-05-16 20:00:00	5269.0	5504.0
2018-05-16 20:50:00	5290.0	5526.0
2018-05-16 21:40:00	5312.0	5549.0

KASETSART UNIVERSITY

HPCNC



KU
KASETSART
UNIVERSITY

Finding relationships with existing attributes

Some analysis...

1. Multivariable linear regression

KU KASETSART UNIVERSITY

HPCNC

Finding correlation

It is a mutual relationship between quantities.

-Sales increases when increase marketing budget?

-Energy usages increase when temperature is high?

-Customer purchases more if there are packages promotions?

etc.

https://www.datascience.com/learn-data-science/fundamentals/introduction-to-correlation-python-data-science

KASETSART UNIVERSITY

HPCNC

Intro to correlation

It is a mutual relationship between quantities.

-Sales increases when increase marketing budget?

-Energy usages increase when temperature is high?

-Customer purchases more if there are packages promotions?

etc.

https://www.datascience.com/learn-data-science/fundamentals/introduction-to-correlation-python-data-science

KASETSART UNIVERSITY

HPCNC

Why we need correlation?

Help predict one quantity from another.

Indicate casual relationship.

Basic foundation to other modeling techniques.

KU KASETSART UNIVERSITY

HPCNC

Type of correlations

Covariances:

Covariance is a statistical measure of association between two variables X and Y.

Indicate the increase in the same direction.

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

if both variables tend to move in the same direction, we expect the "average" rectangle connecting each point (X_i, Y_i) to the means (X_bar Y_bar) to have a large and positive diagonal vector, corresponding to a larger positive product in the equation above.

$$E(X) = \sum X * P(X)$$

KU KASETSART UNIVERSITY

HPCNC

Pearson Correlation Coefficient

Pearson correlation measures the linear association between continuous variables.

This coefficient quantifies the degree to which a relationship between two variables can be described by a line.

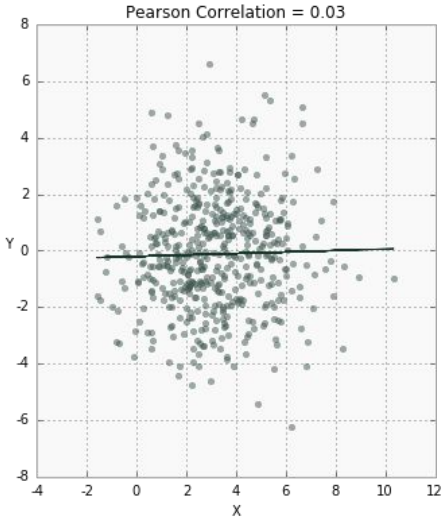
Developed by Karl Pearson over 120 years ago.

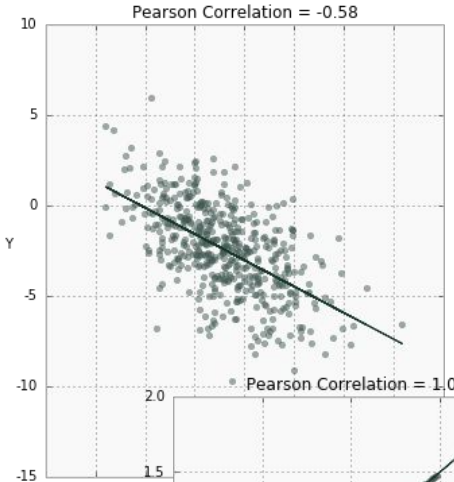
$$\rho_{X,\,Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

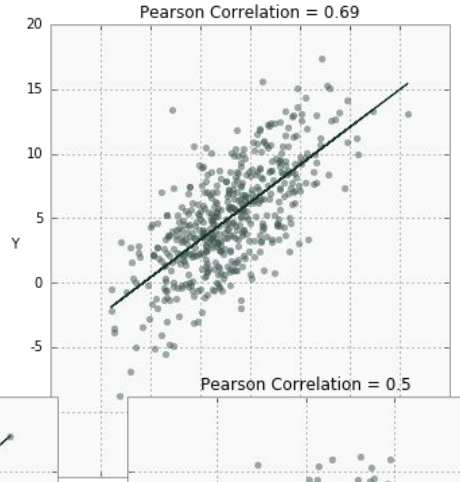
subtracting their means and re-scaled by a measure of standard deviations.

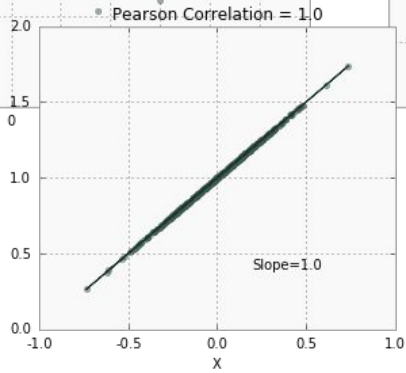
$$\rho_{X,\,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

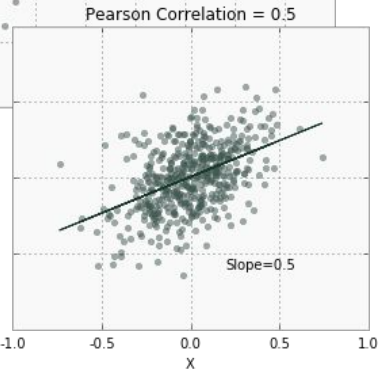
-Note that if X and Y are independent, then ρ is close to 0, but not vice versa!
-correlation does not equal slope

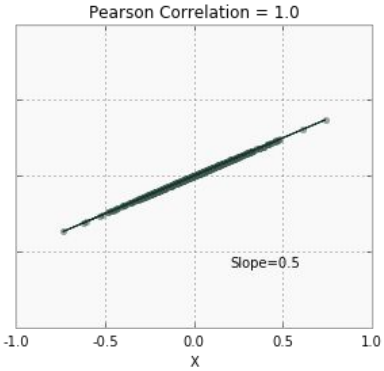












The closer ρ is to 1, the more an increase in one variable associates with an increase in the other.

Spearman's Correlation Coefficient

-Spearman's correlation is not restricted to linear relationships.

-it measures monotonic association (only strictly increasing or decreasing, but not mixed) between two variables and relies on the rank order of values.

Spearman's coefficient looks at the relative order of values for each variable.

$$\rho_{rank_X,\ rank_Y} = \frac{cov(rank_X,\ rank_Y)}{\sigma_{rank_X}\sigma_{rank_Y}}$$

Spearman's rank correlation coefficient can be defined as a special case of Pearson ρ applied to ranked (sorted) variables.

KU KASETSART UNIVERSITY

HPCNC

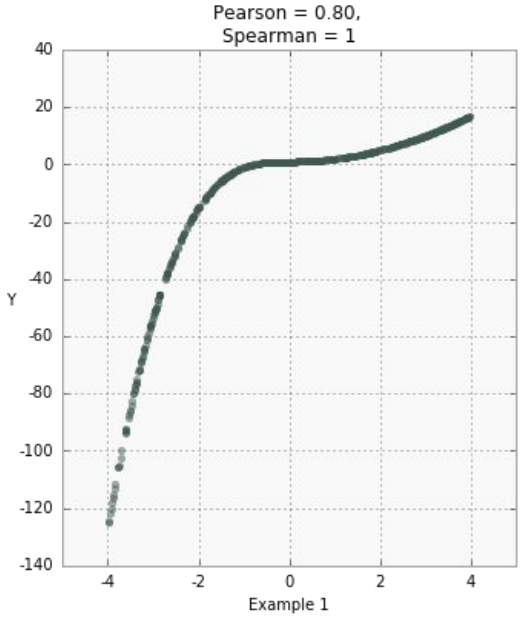
If all ranks are unique (i.e. there are no ties in ranks),

The simplified version: $$\rho_s = 1 - \frac{6\sum d_i^2}{N(N^2-1)}$$

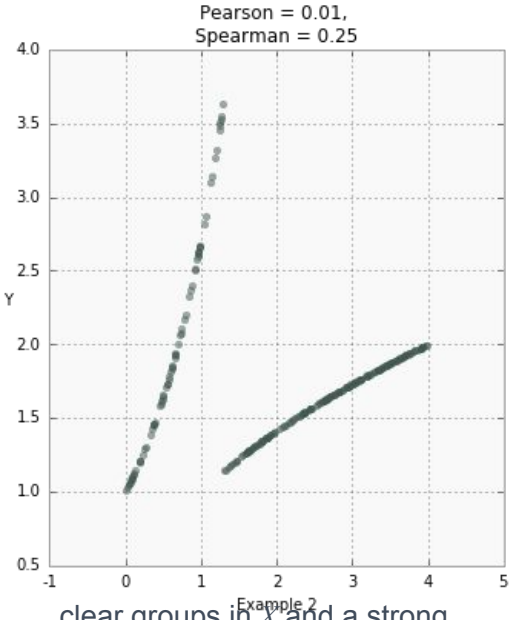
where $di = rank(Xi)$ - $rank(Yi)$,is the difference between the two ranks of each observation and N is the number of observations.

KU KASETSART UNIVERSITY

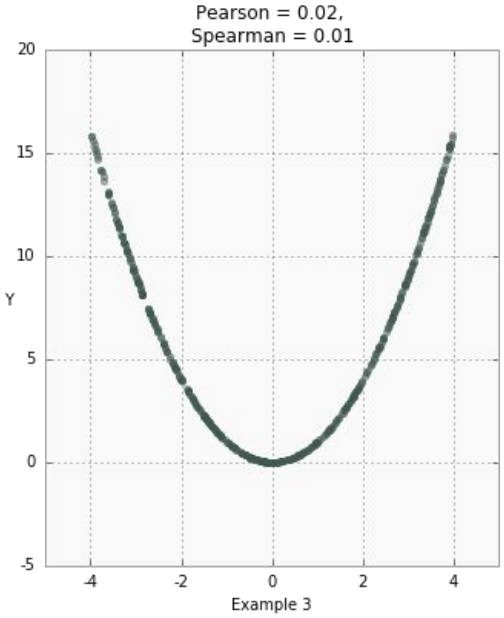
HPCNC



a clear monotonic (always increasing) and non-linear relationship.



clear groups in X and a strong, although non-monotonic, association for both groups with Y. Pearson correlation is almost 0 since the data is very non-linear. Spearman rank correlation shows a weak association since the data is non-monotonic.



a nearly perfect quadratic relationship centered around 0. However, both correlation coefficients are almost 0 due to the non-monotonic, non-linear, and symmetric nature of the data.

KASETSART UNIVERSITY

HPCNC

Kendall's Tau Coefficient

Kendall's τ does not take into account the difference between ranks — only directional agreement. Therefore, this coefficient is more appropriate for discrete data.

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{N(N-1)/2}$$

KU KASETSART UNIVERSITY

HPCNC

Concordant pairs $(x1, y1), (x2, y2)$ are pairs of values in which ranks coincide: $x1 < x2$ and $y1 < y2$ or $x1 > x2$ and $y1 > y2$.

	X	Y
a	1	7
b	2	5
c	3	1
d	4	6
e	5	9

Here (4,6) and (5,9) in rows d and e is a concordant pair.

To calculate the numerator of τ, we compare all possible pairs in the dataset and count number of concordant pairs; 6 in this case:

- (1,7) and (5,9)
- (2,5) and (4,6)
- (2,5) and (5,9)
- (3,1) and (4,6)
- (3,1) and (5,9)
- (4,6) and (5,9)

Discordant pair

- (1,7) and (2,5)
- (1,7) and (3,1)
- (1,7) and (4,6)
- (2,5) and (3,1)

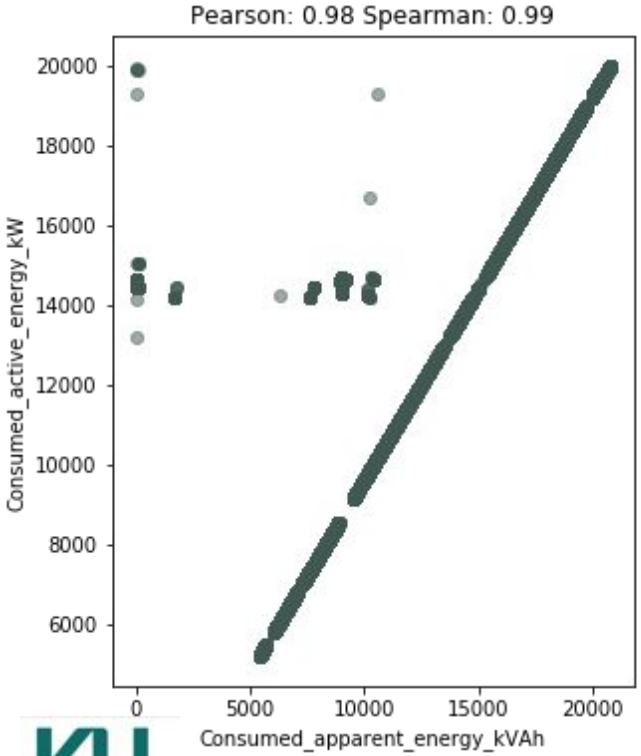
KU KASETSART UNIVERSITY

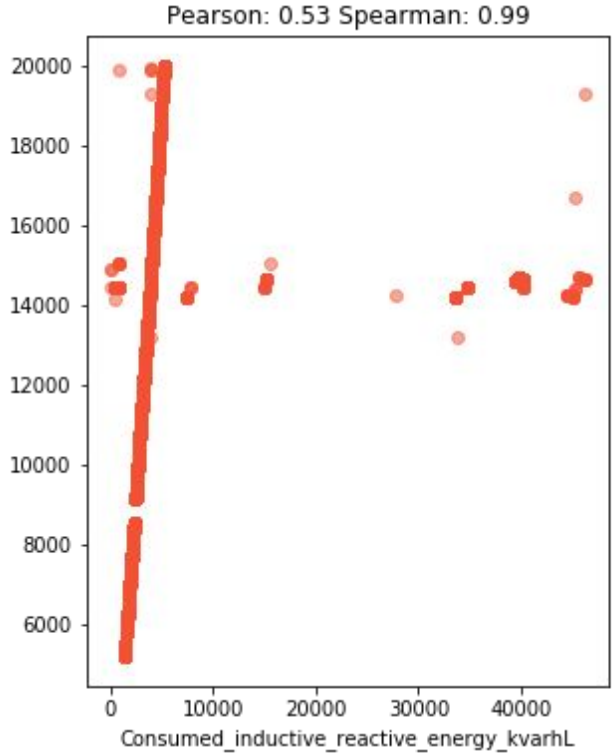
HPCNC

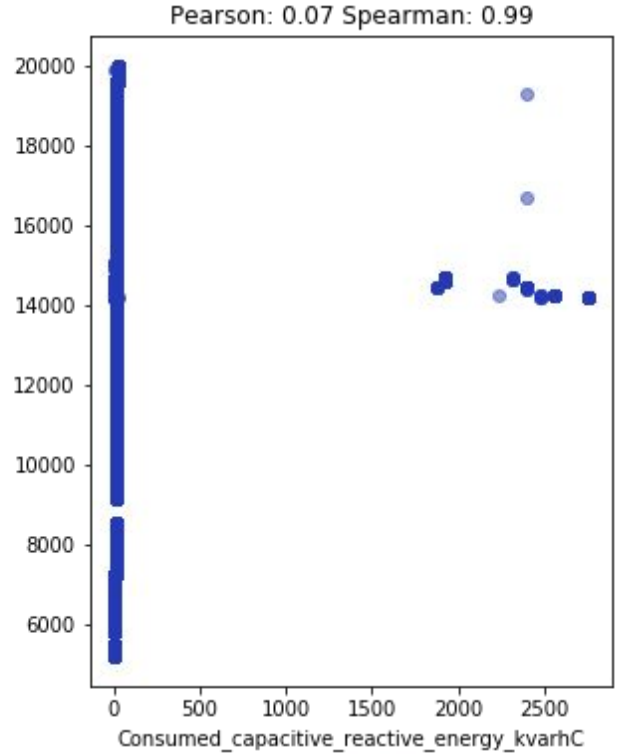
Visualization of relation

KU KASETSART UNIVERSITY

HPCNC







KU KASETSART UNIVERSITY

HPCNC

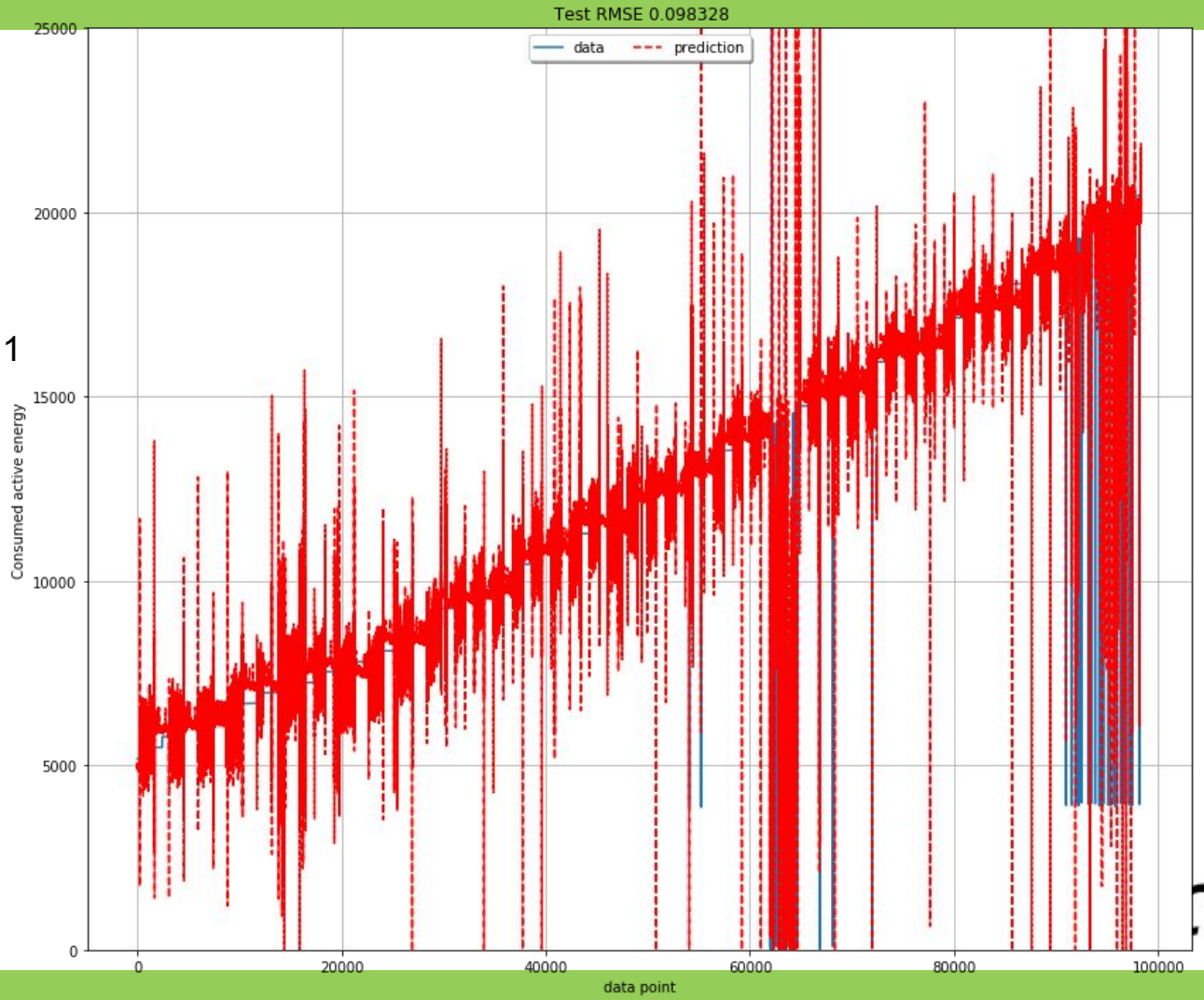
Top 20 features from correlation

Consumed_apparent_energy_kVAh	0.981747
Consumed_inductive_reactive_energy_kvarhL	0.528266
Consumed_capacitive_reactive_energy_kvarhC	0.068560
L2_voltage__THD	0.067717
Cos_L1	0.052542
L2_current__THD	0.050298
L1_current__THD	0.047905
Consumed_CO2_emissions	0.039708
Consumed_capacitive_reactive_energy_varhC	0.039587
L1_Power_Factor	0.035377
Threephase_Cos	0.034699
L1_Capacitive_Power	0.031062
L3_Phase_voltage	0.027613
Capacitive_Threephase_Powe	0.026782
Cos_L2	0.026488
Cos_L3	0.024519
L3Power_Factor	0.024410
L3_Apparent_Power	0.023013
Maximum_demand_I_L3	0.021808
L2_Power_Factor	0.021571

KASETSART UNIVERSITY

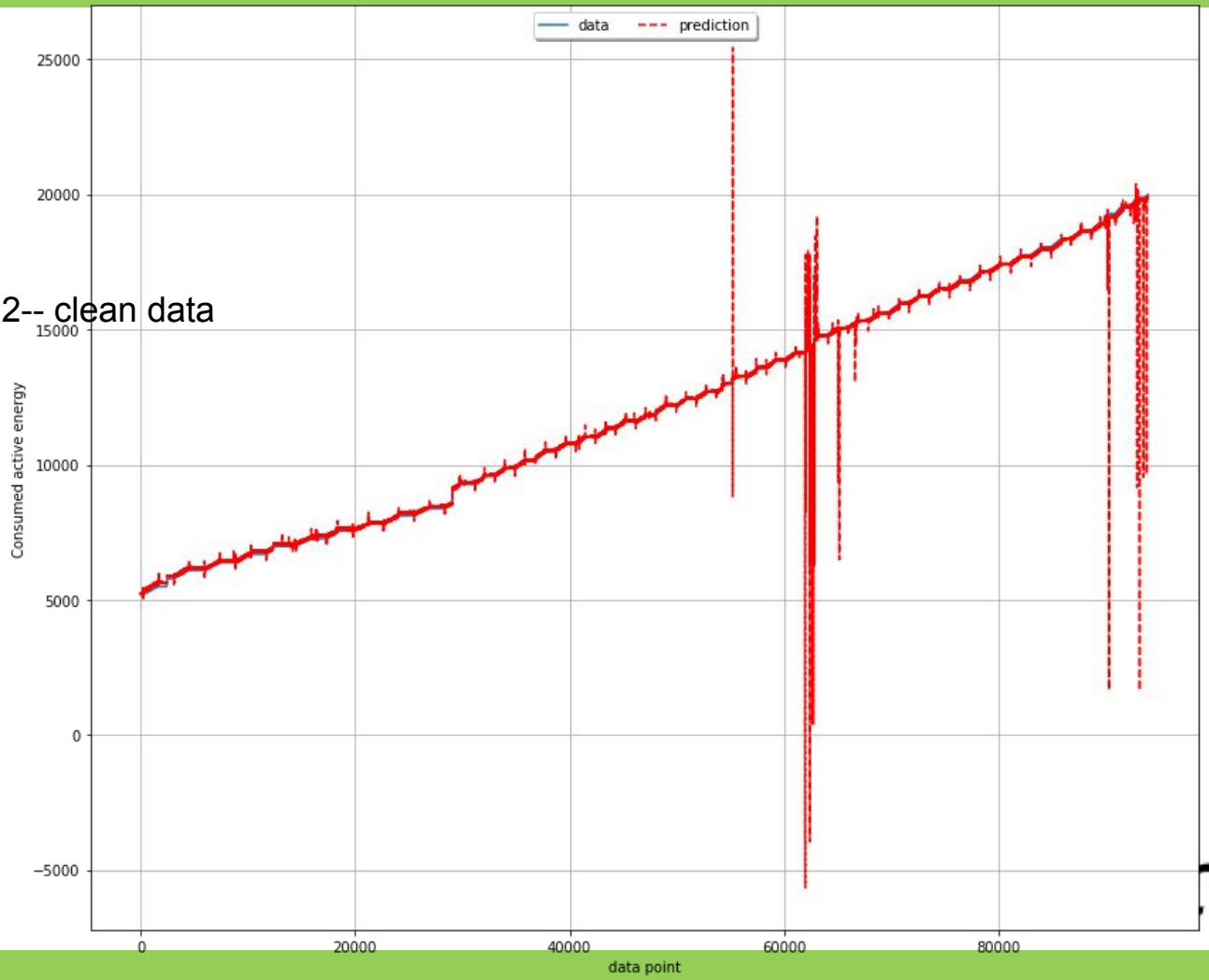
HPCNC

Linear regression1



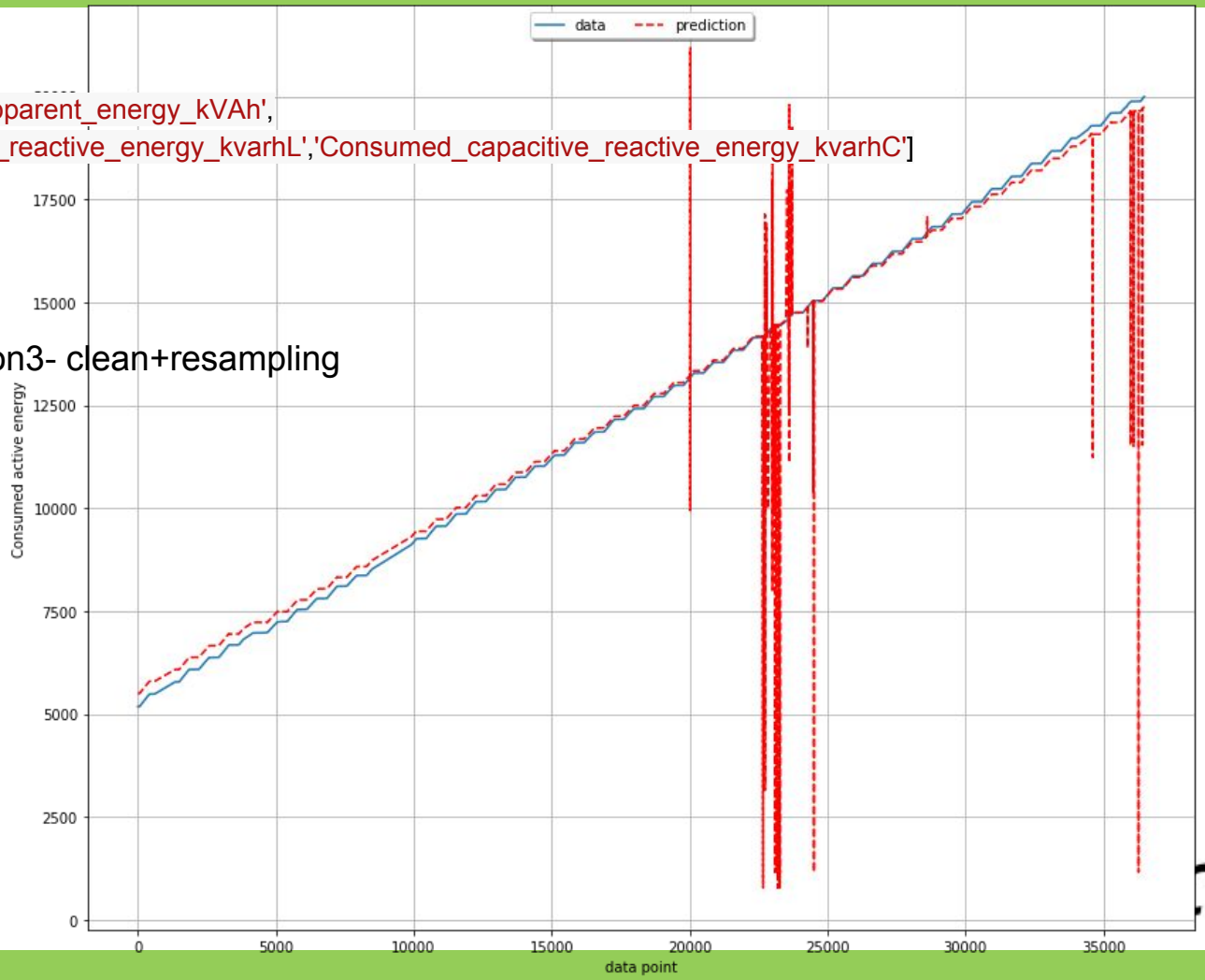
KU
KASETSART
UNIVERSITY

Linear regression2-- clean data



```
cols = ['Consumed_apparent_energy_kVAh',
'Consumed_inductive_reactive_energy_kvarhL','Consumed_capacitive_reactive_energy_kvarhC']
```

Linear regression3- clean+resampling



Resampling with multicolumn

2. Autoregression

Autoregression

Time series data

Predict x(t+1) from x(t-2), x(t-1), x(t)

--Finding autocorrelation

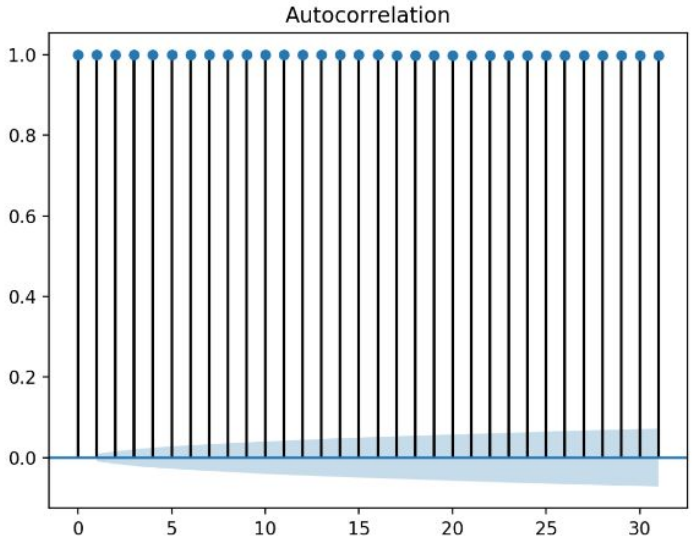
KU KASETSART UNIVERSITY

HPCNC

Autocorrelation

	X(t-1)	X(t+1)
X(t-1)	1.0	1.0
X(t+1)	1.0	1.0

good!



Lag=31

KU KASETSART UNIVERSITY

HPCNC

Stationary analysis

-Testing whether the data is stationary. It has the constant average/moving average

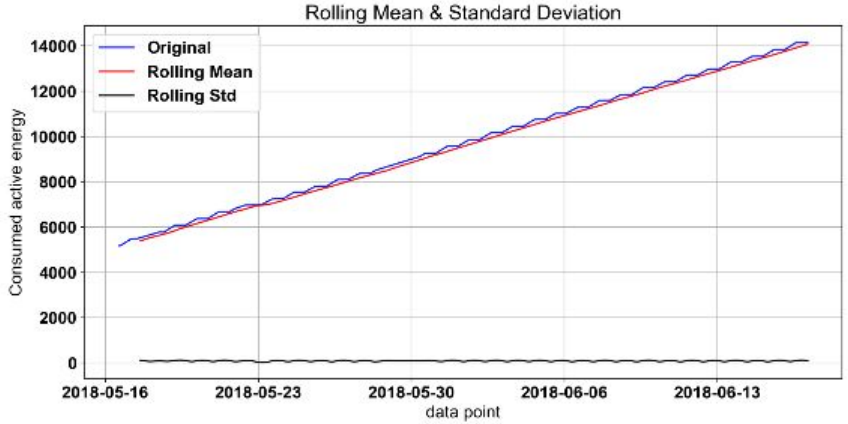
-Use Dickey Fuller test

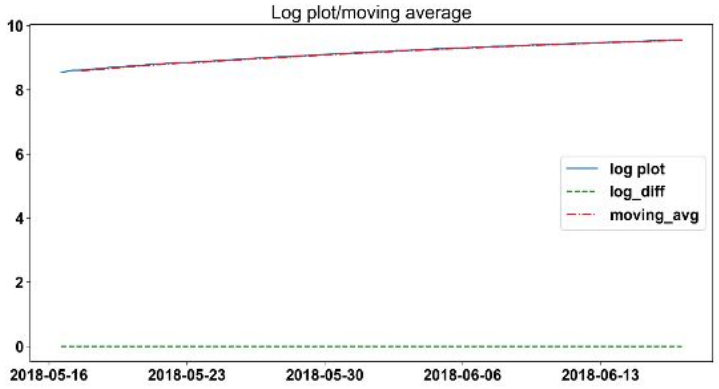
Results of Dickey-Fuller Test:	
Test Statistic	**-0.061172**
p-value	0.953171
#Lags Used	55.000000
Number of Observations Used	45332.000000
Critical Value (5%)	**-2.861604**
Critical Value (1%)	**-3.430494**
Critical Value (10%)	-2.566804

Test statistics -0.061172 should be less than Critical Value (1%)

KU KASETSART UNIVERSITY

HPCNC



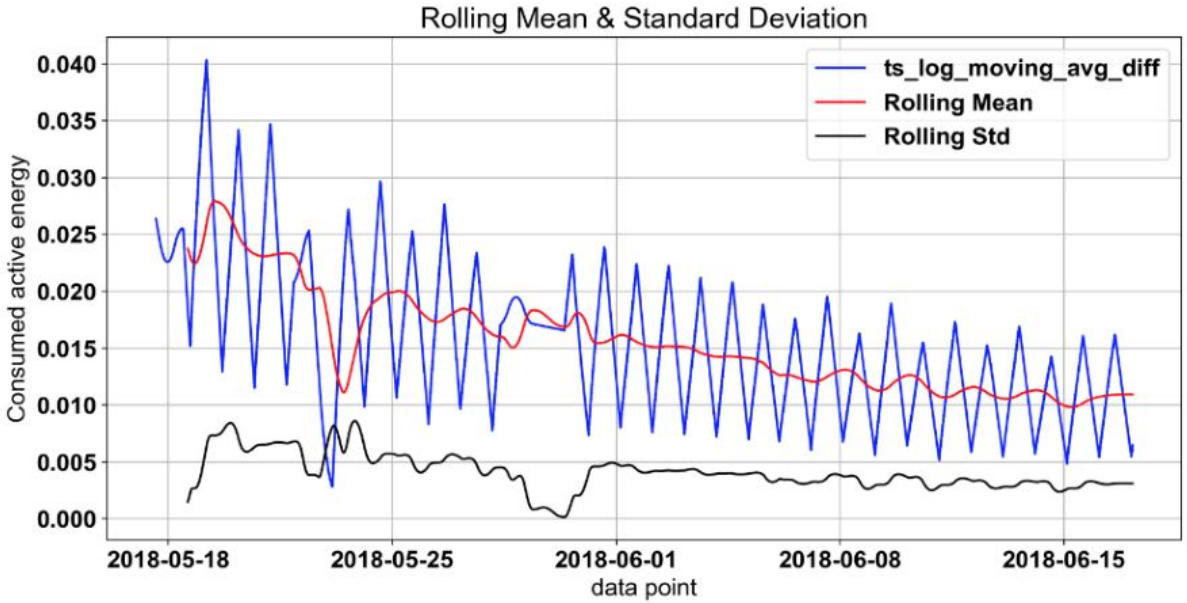


KU
KASETSART
UNIVERSITY

HPCNC

Solving non-stationary

-Take difference (first order diff : with prev value)



KU KASETSART UNIVERSITY

HPCNC

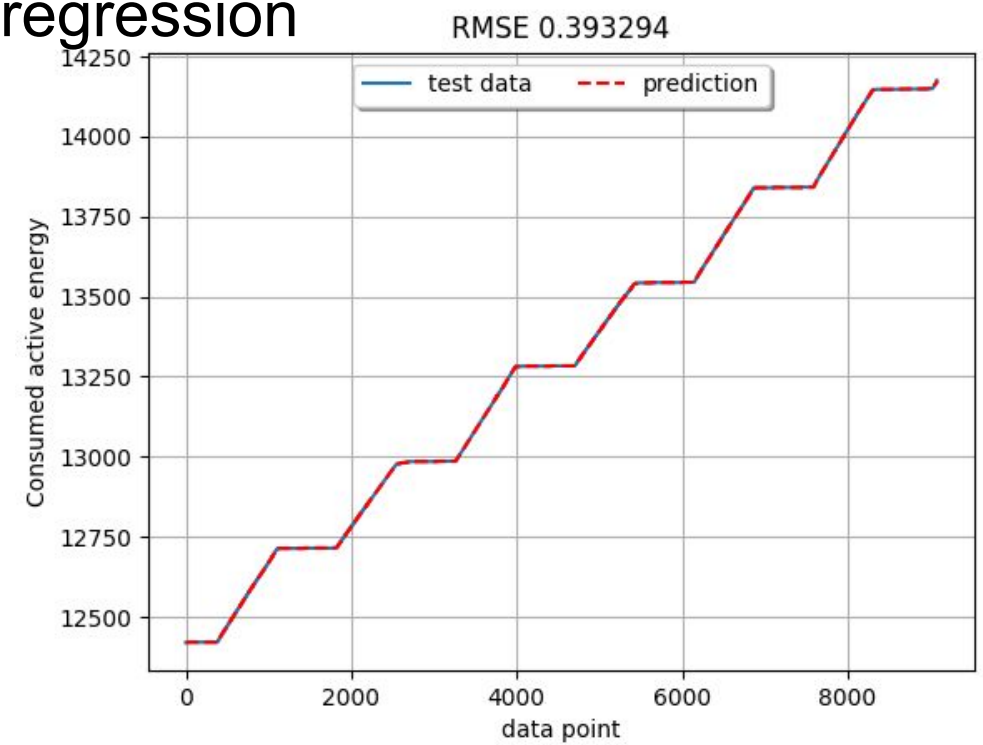
Test Statistic	-8.922295e+00
p-value	1.033276e-14
#Lags Used	5.500000e+01
Number of Observations Used	4.389300e+04
Critical Value (5%)	-2.861606e+00
Critical Value (1%)	-3.430499e+00
Critical Value (10%)	-2.566805e+00

Better!!

KU KASETSART UNIVERSITY

HPCNC

Prediction with autoregression



https://cocalc.com/projects/

Free Jupyter project: jupyter lab

https://github.com/cchantra/datascience_tutorial.git

KU Data Science Status

KU Data Science Bootcamp

KU Data Science Forum

KU Data Science Education

KU Open data

http://datascience.ku.ac.th

KU KASETSART UNIVERSITY

HPCNC

RESEARCH UNIVERSITY NETWORK THAILAND

การบูรณาการ การนำเสนอ และการวิเคราะห์ข้อมูลขนาดใหญ่ทางเกษตร

KU KASETSART UNIVERSITY

วัตถุประสงค์

1.สร้างระบบโครงสร้างพื้นฐานด้านสารสนเทศเพื่อการบูรณาการข้อมูลทางการเกษตรเพื่อสร้างข้อมูลเปิดให้ประเทศ
2.นำข้อมูลด้านการเกษตรที่สำคัญบางส่วนมา รวมไว้ที่เดียว เพื่อให้ง่ายต่อการเข้าถึง เชื่อมโยง และนำไปใช้งานต่อยอด
3.สร้างรูปแบบการเข้าถึงข้อมูลเปิด (Open data) ทางการเกษตร
4. สร้างบริการใช้ข้อมูลและวิเคราะห์ข้อมูลโดยใช้ เทคโนโลยีขั้นสูงเช่น Machine learning และ Deep learning ได้

ประกอบด้วยกิจกรรมย่อย

กิจกรรม 1 แพลตฟอร์ม และการบูรณาการข้อมูลทางการเกษตร สำหรับการสร้างข้อมูลเปิด

กิจกรรมที่ 2 การตรวจวัดคุณภาพอากาศ และอุตุนิยมวิทยาเพื่อการบูรณาการด้านการเกษตร

กิจกรรมที่ 3 การทำนายผลผลิตและสุขภาพพืชด้วยภาพถ่ายทางอากาศ

กิจกรรมที่ 4 การวางแผนการเงินเพื่อเกษตรกร โครงการนาแปลงใหญ่

กิจกรรมที่ 5 การวิเคราะห์โรคข้าวจากภาพถ่าย

กิจกรรมที่ 6 การบริการข้อมูลคุณภาพอากาศทางเว็บไซต์เพื่อการเกษตร

Visualization framework
Analytic tools
Integration Platform
Physical Data

Core Technology
-Web service
-Deep Learning/Machine Learning
-Visualization
-Analytics

Output

แพลตฟอร์มรวมข้อมูลด้านการเกษตร สร้างชุดข้อมูลเปิด โครงสร้างสำหรับบริการเว็บเข้าถึงข้อมูลเปิดและวิเคราะห์

บริการเว็บสำหรับตรวจวัดคุณภาพอากาศ และอุตุนิยมวิทยา

บริการทำนายสุขภาพพืชและทำนายผลผลิต และแอปพลิเคชัน

ระบบข้อมูลเพื่อช่วยสนับสนุนในการตัดสินใจทางการเงินแก่เกษตรกร โครงการนาแปลงใหญ่ บริการแนะนำแนวทางในการวางแผนการเงินให้แก่เกษตรกร โครงการนาแปลงใหญ่

แอปพลิเคชันให้บริการระบุชนิดโรคข้าวและแนะนำวิธีการรักษา

บริการข้อมูลคุณภาพอากาศโดยรวม ระบบแสดงผลผ่านทางเว็บไซต์ โดยสามารถนำข้อมูลไปพิจารณาถึงความเสี่ยงที่อาจส่งผลต่อการเจริญเติบโตของพืช หรือแนวโน้มการเกิดโรคของพืช

Outcome/Impact

GLOBAL OPEN DATA INDEX

จำนวนบริการข้อมูล 15 บริการ
จำนวนชุดข้อมูลเปิด 6 เรื่อง

เพิ่มจำนวน Data scientist

STARTUP 4.0 THAILAND

คาดคะเนผลผลิตที่ได้จากพื้นที่เพาะปลูก

วางแผนแก้ปัญหาผลผลิตไม่ทอเพื่อง/ล้นตลาด

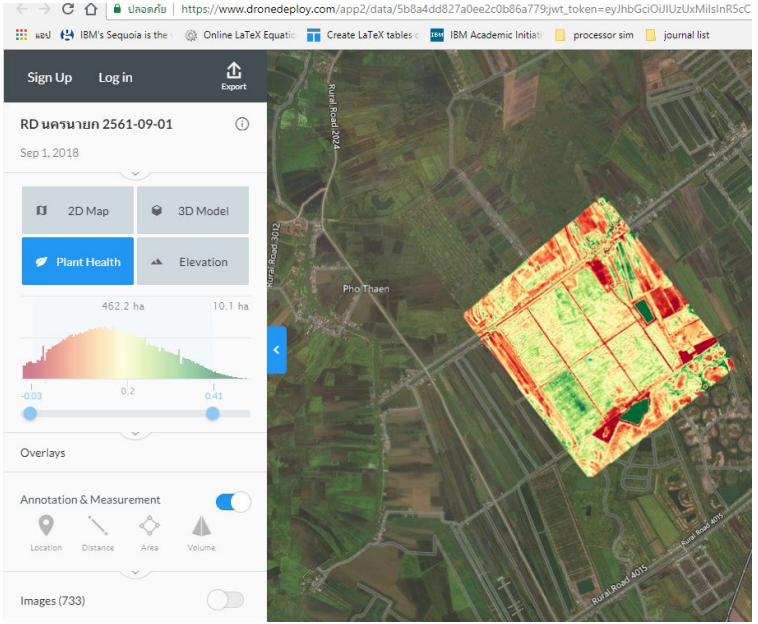
ความเสี่ยงหายที่อาจเกิดจากสภาพอากาศต่อการเพาะปลูกล่วงหน้า

เกษตรกรลดลงใช้ระบบช่วยตัดสินใจทางการเงิน 30 ครัวเรือน

ตระหนักถึงความสำคัญของการวางแผนการเงิน

เกษตรกรลดลงใช้แอปพลิเคชันการรักษาโรคข้าว 20 ราย

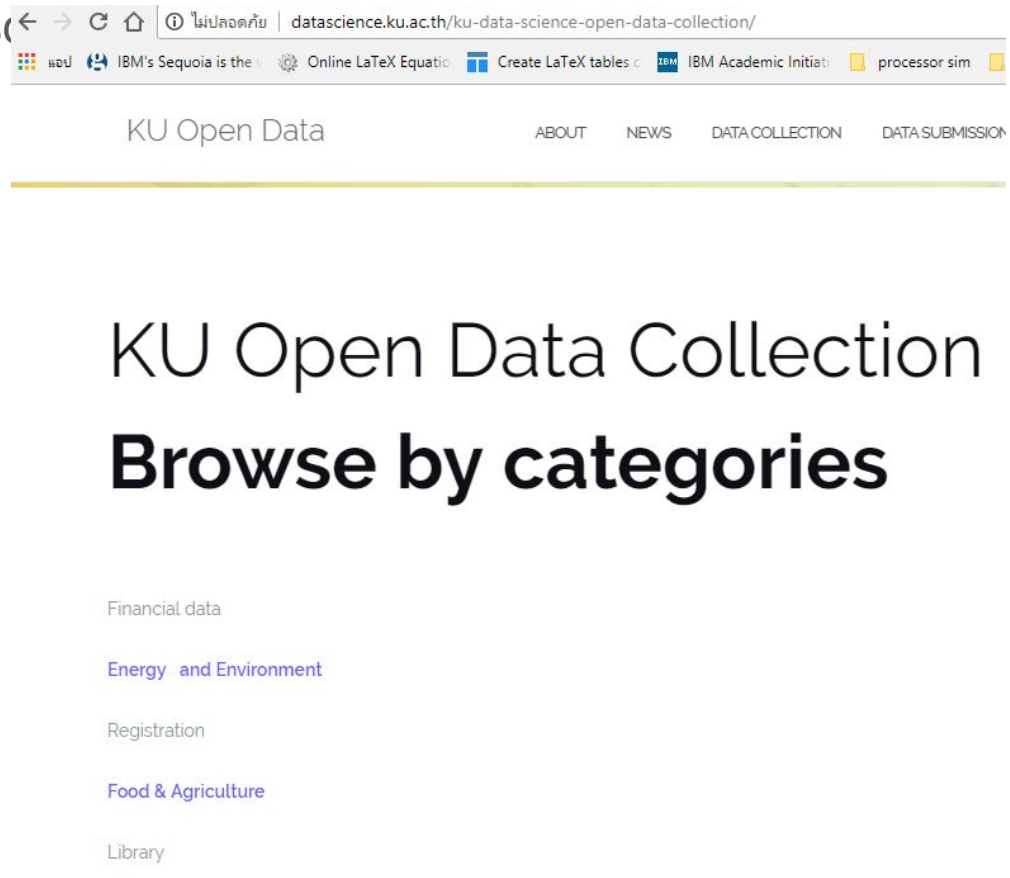
การพบปัญหาและแก้ปัญหาได้ง่ายขึ้น ลดความสูญเสียของผลผลิต



KU KASETSART UNIVERSITY

HPCNC

http://datasc



KU KASETSART UNIVERSITY

HPCNC

https://www.facebook.com/KU-Data-science-boot-camp-2018-356005521545466/?ref=bookmarks