

Workshop: Data Analytics for IoT

Kevin Kam Fung Yuen, PhD
Senior Lecturer
School of Business
Singapore University of Social Sciences
kfyuen@suss.edu.sg,
kevinkf.yuen@gmail.com

Outline

- ◆ What is IoT?
- ◆ What is Data Analytics?
 - ◆ AI
 - ◆ Machine Learning
 - ◆ Data Analytics
- ◆ Applications and Demo
- ◆ Labs

Download: <https://github.com/kkfyuen/DA4IoT>

Definitions

- ◆ The term “Internet of Things” was coined by entrepreneur Kevin Ashton, one of the founders of the Auto-ID Center at MIT. Ashton was part of a team that discovered how to link objects to the internet through an RFID tag. He first used the phrase “Internet of Things” in a 1999 presentation – and it has stuck around ever since.

https://www.sas.com/en_us/insights/big-data/internet-of-things.html

- ◆ “The Internet of Things (IoT) refers to a vast number of “things” that are connected to the internet so they can **share** data with other things – IoT applications, connected devices, industrial machines and more. Internet-connected devices use built-in **sensors** to collect data and, in some cases, act on it. “

https://www.sas.com/en_us/insights/big-data/internet-of-things.html

IoT: Oracle

The Internet of Things (IoT) describes the network of **physical objects**—“things”—that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet.

<https://www.oracle.com/internet-of-things/what-is-iot.html>

IoT: IBM

In a nutshell, the Internet of Things is the concept of **connecting** any device (so long as it has an on/off switch) to the Internet and to other **connected** devices.

The IoT is a giant network of **connected things and people** – all of which collect and share data about the way they are used and about the environment around them.

<https://www.ibm.com/blogs/internet-of-things/what-is-the-iot/>

IoT: wikipedia

- ◆ The **Internet of Things (IoT)** is a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers (UIDs) and **the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.**

https://en.wikipedia.org/wiki/Internet_of_things

IoT: Nokia

- ◆ “Our IoT innovations enable service providers, enterprises and governments to make sense of the massive volumes of data produced by connected sensors, devices and systems. They can help you **capture** the data that matters, **manage it securely** and use it to **generate insights** that create value and improve business outcomes.”

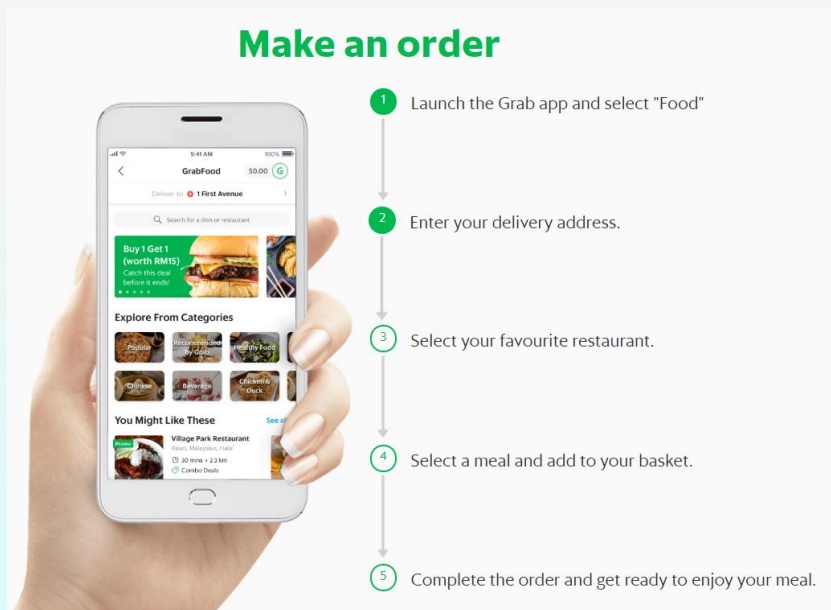
<https://www.nokia.com/networks/portfolio/internet-of-things-iot/>

IoT: Connected World

- ◆ IoT is to connect Physical World and Information World.

Discussion

- ❖ What **things** are involved for from making food order to delivering the food in IoT?
- ❖ For example, Grab food or others.



HOW IOT IS RELATED TO DATA ANALYTICS?

IoT leading to Big Data Analytics

- ◆ The rapid growth of data we experience today has led to the urgent need to develop effective and efficient techniques for big data analytics, which are much required by industries and academic communities in order to be able to discover useful information, or knowledge in big data.
- ◆ Big data analytics concerns the use of modern statistical and other machine learning techniques to analyze huge amounts of data.

(Yuen et al, 2018)

Challenging issues in Big Data Analytics

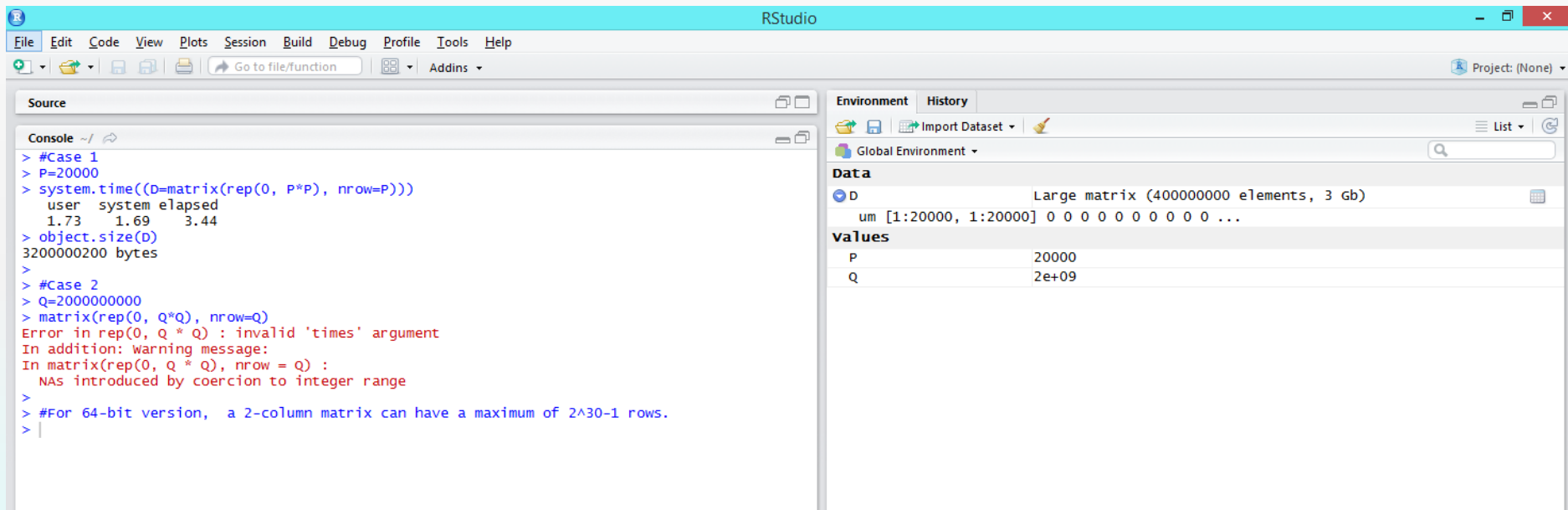
- ◆ High dimensionality of data
- ◆ Multiple objectives of the problems under study
- ◆ Conventional 5Vs,
 - ◆ large scale of data (Volume),
 - ◆ multiple sources of data (Variety),
 - ◆ rapid growth of data (Velocity),
 - ◆ quality of data (Veracity),
 - ◆ usefulness of data (Value).

(Yuen et al, 2018)

Simple example for Volume

How big is big?

problem of Matrix in R



The screenshot shows the RStudio interface. The console on the left displays the following R code and its output:

```
> #Case 1
> P=20000
> system.time((D=matrix(rep(0, P*P), nrow=P)))
  user system elapsed
  1.73   1.69    3.44
> object.size(D)
3200000200 bytes
>
> #Case 2
> Q=2000000000
> matrix(rep(0, Q*Q), nrow=Q)
Error in rep(0, Q * Q) : invalid 'times' argument
In addition: warning message:
In matrix(rep(0, Q * Q), nrow = Q) :
  NAS introduced by coercion to integer range
>
> #For 64-bit version, a 2-column matrix can have a maximum of 2^30-1 rows.
> |
```

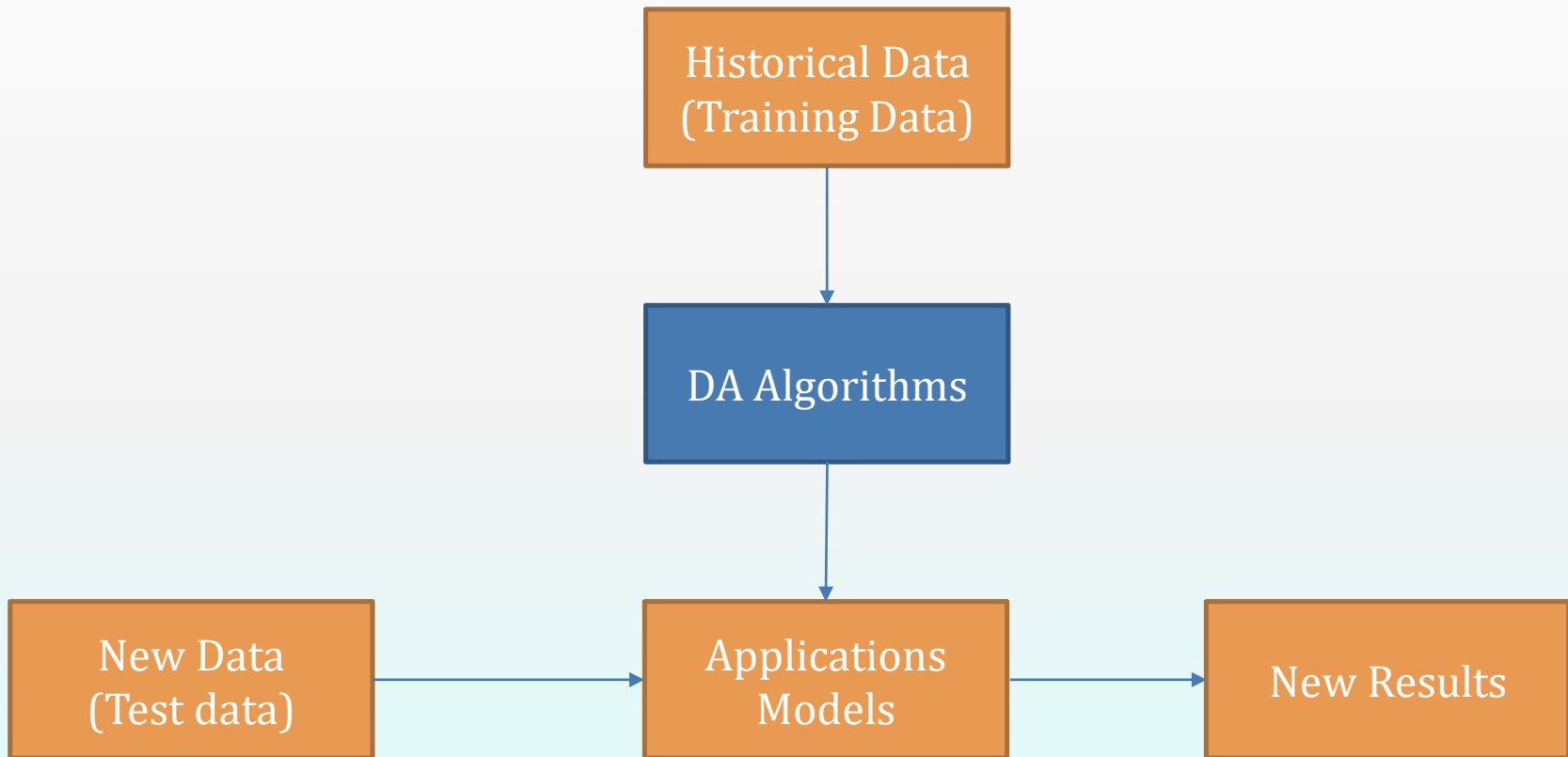
The Environment pane on the right shows the following data:

Object	Value
D	Large matrix (400000000 elements, 3 Gb)
um	[1:20000, 1:20000] 0 0 0 0 0 0 0 0 ...
P	20000
Q	2e+09

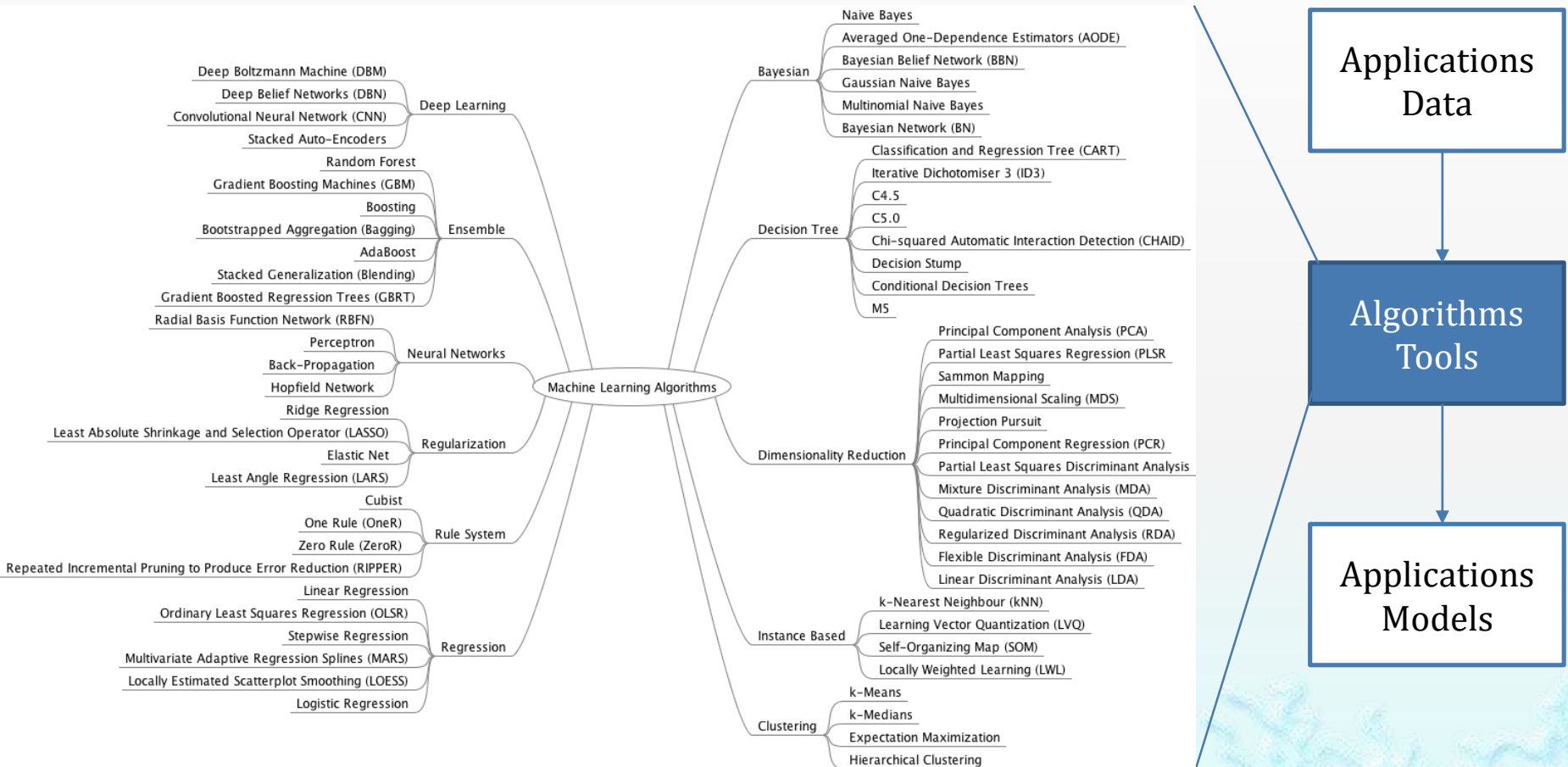
Discussion

- ◆ Search internet, what is the largest size can MS Excel handle?

Data Analytics



Data Analytics with Machine Learning



Source:

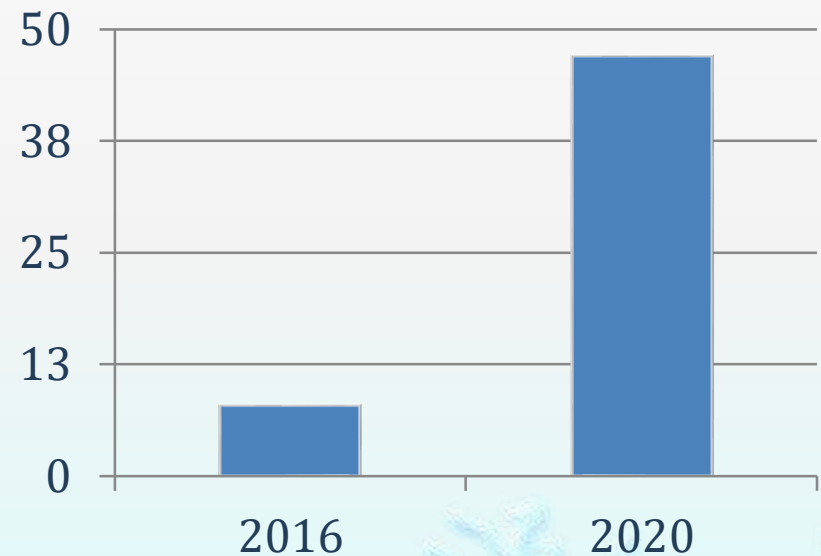
<http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>

AI Is The New Electricity

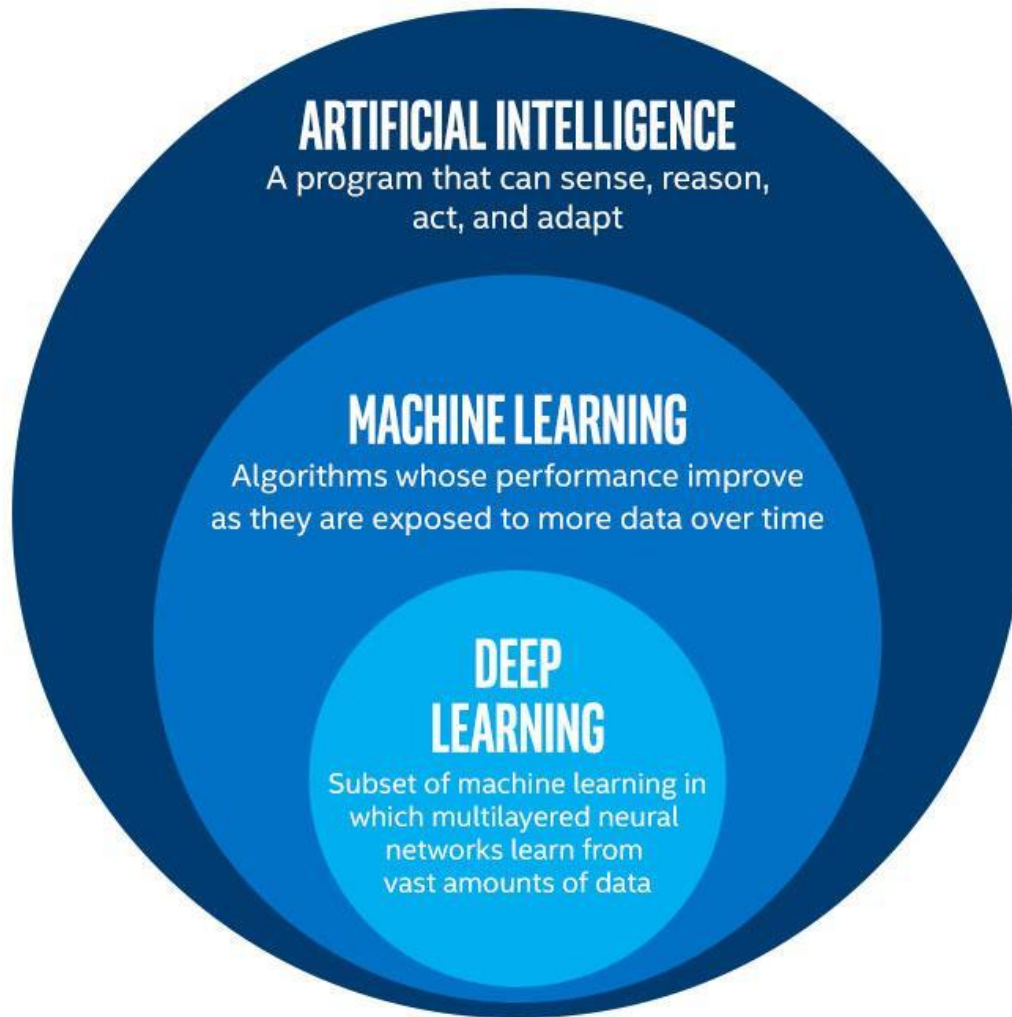
“About 100 years ago, electricity transformed every major industry. AI has advanced to the point where it has the power to transform...every major sector in coming years.”

-Andrew Ng, Stanford University

Projected Revenue (in billions USD) Generated from AI, 2016-2020 (IDC)



AI, ML and DL



Artificial Intelligence

“A branch of computer science dealing with the simulation of intelligent behavior in computers.” (Merriam-Webster)

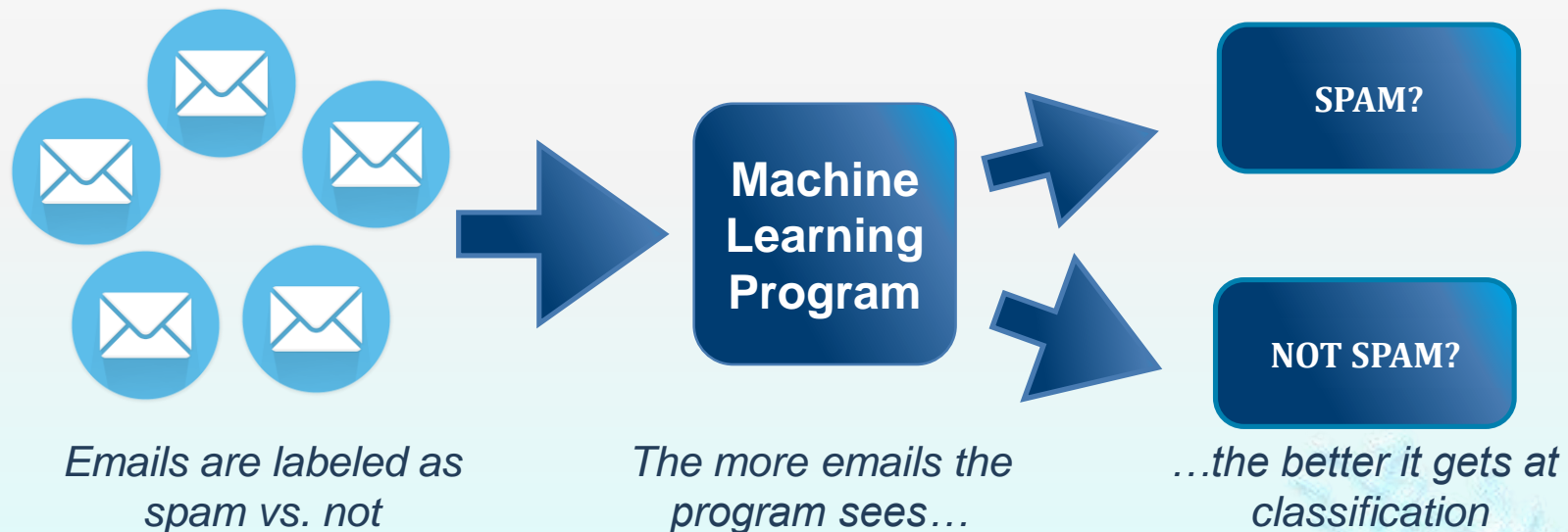
“A program that can sense, reason, act, and adapt.” (Intel)

“Colloquially, the term ‘artificial intelligence’ is applied when a machine mimics ‘cognitive’ functions that humans associate with other human minds, such as ‘learning’ and ‘problem solving’.” (Wikipedia)

Machine Learning

“The study and construction of programs that are *not explicitly programmed*, but learn patterns as they are exposed to more data over time.” (Intel)

These programs learn from repeatedly seeing data, rather than being explicitly programmed by humans.



Machine Learning Terminology

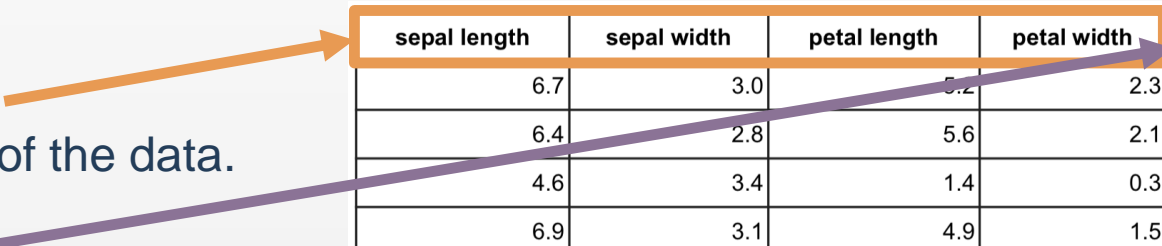
This example is learning to classify a species from a set of measurement features.

Features:

Attributes of the data.

Target:

Column to be predicted.



sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

Two Main Types of Machine Learning

	Dataset	Goal	Example
Supervised Learning	Has a target column	Make predictions	Fraud detection
Unsupervised Learning	Does not have a target column	Find structure in the data	Customer segmentation

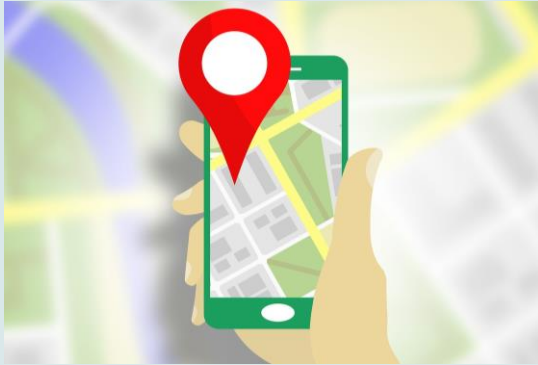
Machine Learning Example

- Suppose you wanted to identify fraudulent credit card transactions.
- You could define features to be:
 - Transaction time
 - Transaction amount
 - Transaction location
 - Category of purchase
- The algorithm could learn what feature combinations suggest unusual activity.



Applications

Navigation



Google & Waze find the fastest route, by processing traffic data.

Image classification



"Dog"

"Cat"

As of 2015, computers can be trained to perform better on this task than humans.

Content

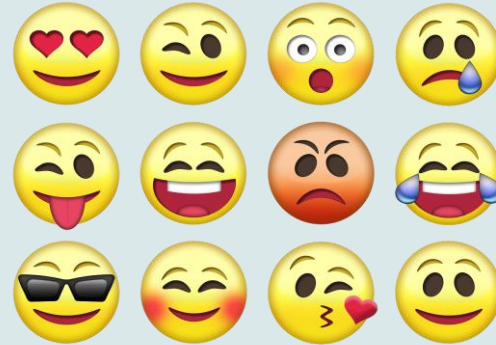
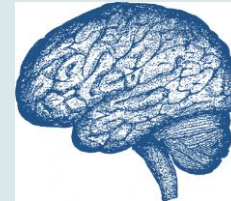


Image recognition and sentiment analysis to ensure that content of the appropriate "mood" is being served.

Machine translation

"I am a student"



"Je suis étudiant"

As of 2016, we have achieved near-human performance using the latest AI techniques.

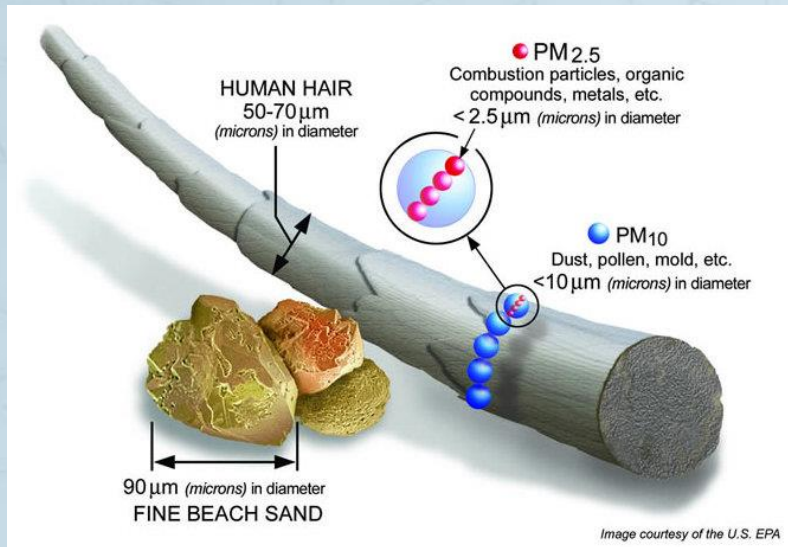
LABS



Setup R environment

- ◆ To set up the R integrated development environment, we need to install R and then install RStudio.
- ◆ Installing R According to the user's operation system (OS)
 - ◆ Microsoft Windows: <http://cran.r-project.org/bin/windows/base/>
 - ◆ MacOS: <http://cran.r-project.org/bin/macosx/>
 - ◆ Linux: <http://cran.r-project.org/bin/linux/>
- ◆ Installing RStudio
<https://www.rstudio.com/products/rstudio/download>
- ◆ There are several versions of R studio. For this foundation course, we use “RStudio Desktop Open Source License”, which is FREE.

EXAMPLE OF DATA ANALYTICS WITH REGRESSION: ANALYSIS OF AIR QUALITY



1. Yuen, K.K.F. (2017), Multiple Regression Analyses for Air Quality and Weather in Hong Kong, *Journal of Advances in Information Technology*, Vol. 8, No. 2, pp. 135-140. doi: 10.12720/jait.8.2.135-140 (open access)
2. R fundamentals, 2019, <https://github.com/kkfyuen/Rbootcamp1> (open access)



The data was collected by device and stored in the internet

MULTIPLE REGRESSION MODEL

To find \hat{W} , residual error sum of squares is minimized.

$$\text{Min}\left(\sum(\varepsilon^2)\right) = \text{Min}\left(\sum(Y - \hat{W} \cdot X)^2\right) \quad (6)$$

For an ideal regression model, ε_i is zero and \hat{W} is used to estimate W , i.e. $W \cong \hat{W}$, and the closed form solution of \hat{W} is presented as below.

$$Y = \hat{W} \cdot X \quad (7)$$

$$\hat{W} = \frac{Y}{X} = \frac{X^T Y}{X^T X} = (X^T X)^{-1} X^T Y \quad \leftarrow \quad \boxed{\text{Can we use this close form in Big Data?}} \quad (8)$$

If matrix rank of X , i.e. $\text{Rank}(X)$, is equal to $m+1$ (or $\det(X^T X) \neq 0$), $X^T X$ is invertible and W has a unique solution. The QR decomposition of one of the common approaches to find the inversion matrix. \hat{W} is used to calculate the model values as below.

$$\hat{Y} = \hat{W} \cdot X \quad (9)$$

Therefore \hat{Y} is used to estimate Y with some errors for the estimation.

1. What are core features? Do we use all??
2. How to find the optimal solution?

Date	PM _{2.5}	CO	NO ₂	NO _x	O ₃	PM ₁₀	SO ₂	Temperature	Humidity	Cloud	Wind Degree	Wind Speed
Unit	µg/m ³	10µg/m ³	µg/m ³	µg/m ³	µg/m ³	µg/m ³	µg/m ³	deg. C	%	%	Degree	Km/h
1	11	63	35	63	25	16	9	27.9	88	86	230	23.5
2	11	62	37	81	22	17	12	28.7	86	88	230	19.7
3	25	90	56	88	34	32	12	28.1	87	88	230	10.9
4	24	79	31	36	67	35	7	28.2	84	88	70	28.1
5	21	73	47	66	34	32	8	27.1	89	86	50	30
6	11	72	48	92	14	17	10	26.7	90	86	50	7.3
7	13	68	45	77	22	16	10	26.5	90	91	230	16.1
8	10	69	43	92	12	16	9	27.1	90	84	230	8.8
9	11	71	45	104	11	16	9	27	88	87	230	11.3
10	11	81	43	132	5	18	10	26.3	93	87	50	11
11	13	79	31	84	15	19	8	28.1	87	72	30	7.1
12	16	73	43	66	36	27	9	28.7	83	49	30	15.2
13	17	69	51	80	29	26	10	28.2	84	61	80	18.7
14	28	83	58	83	57	46	10	29.6	69	59	340	17.3
15	27	79	50	71	50	42	13	29.4	68	63	280	15
16	16	71	38	52	50	26	10	29	70	44	10	15.3
17	22	72	48	65	48	39	11	29.3	66	50	10	19.4
18	24	59	45	51	85	42	11	28.6	66	47	10	12.7
19	27	64	53	64	73	45	10	28.6	73	51	10	21.6
20	36	73	54	65	70	52	8	25.5	87	85	350	32
21	32	68	56	66	96	55	10	27.1	77	52	70	35.7
22	24	52	42	48	94	45	9	27.2	76	87	70	34.9
23	21	50	38	45	80	40	8	27.7	78	88	80	30.6
24	21	51	41	50	68	39	9	27.9	78	66	80	27.6
25	32	62	47	55	91	51	12	28.1	80	52	80	20.1
26	49	78	57	69	105	78	14	28.5	81	69	230	15
27	56	94	74	102	77	84	15	31.1	68	36	290	11.3
28	47	88	66	81	70	82	14	30.4	58	79	300	25.5
29	20	83	58	82	37	42	11	26.5	70	88	10	18.2
30	31	103	75	130	14	65	12	25.1	78	86	10	7.5

APPENDIX A: AIR POLLUTANTS (YUEN LONG STATION) AND WEATHER DATA IN SEP 2016 IN HONG KONG

The air pollutant data were obtained from [4] and the weather data were obtained from [5]. For the pilot test and the availability of the data, the air quality data from YUEN LONG station is chosen for the analysis in Sep 2016 in Hong Kong.

Research questions:
Use other factors to
Predict PM2.5

TABLE I: REGRESSION RESULTS AND T STATISTICS FOR REGRESSION MODEL OF 11 FACTORS

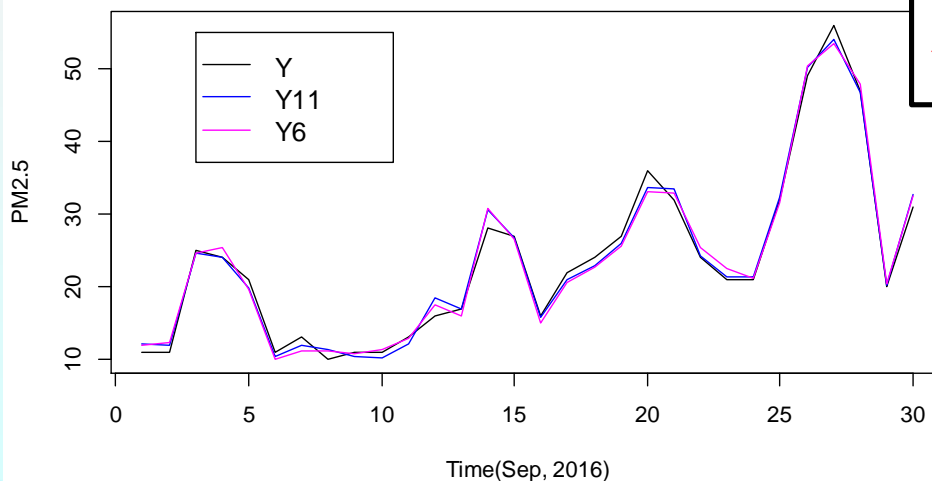
Index	Factors	\hat{W}	SE	t_w	p	t_w	Reject?
X_0	Intercept	-54.80	17.61	-3.11	0.006		
X_1	CO	0.15	0.05	3.08	0.006		
X_2	NO ₂	0.06	0.08	0.78	0.444		Yes
X_3	NO _x	-0.04	0.03	-1.22	0.239		Yes
X_4	O ₃	0.07	0.04	2.01	0.060		Margin
X_5	PM ₁₀	0.47	0.06	8.07	0.000		
X_6	SO ₂	0.16	0.30	0.52	0.607		Yes
X_7	Temperature	0.84	0.47	1.79	0.091		
X_8	Humidity	0.28	0.06	4.93	0.000		
X_9	Cloud	-0.05	0.03	-1.46	0.161		Yes
X_{10}	Wind Degree	0.01	0.00	3.29	0.004		
X_{11}	Wind Speed	0.00	0.06	0.07	0.947		Yes

Table II: Regression results and t statistics for regression model of 6 factors

Index	Factors	\hat{W}	SE	t value	$p(> t)$
X_0	(Intercept)	-64.57	11.63	-5.55	0.000
X_1	CO	0.16	0.05	3.44	0.002
X_2	O3	0.10	0.02	3.89	0.001
X_3	PM10	0.48	0.04	11.64	0.000
X_4	Temperature	1.18	0.30	3.86	0.001
X_5	Humidity	0.25	0.05	4.59	0.000
X_6	Wind Degree	0.01	0.00	3.34	0.003

Table III: R related values, F tests and ANOVA results for both full and reduced models

	Regression model of 11 factors	Regression model of 6 factors
R^2	0.9894	0.9862
R^2_{adjust}	0.983	0.9826
F	153.5	274.7
$P(F)$	0.000	0.000
SSE	42.577	55.532
df	18	23
$F^\pm = 1.0954, p_{F^\pm} = 0.3967$		



Statistics also indicates that
5 factors are completely not related.
We just need 6 factors instead of 11 factors

Figure 1: PM2.5 Values and its predicted values with 10 and 6 factors in Sep 2016

R Code

```
# import the data from internet
data <- read.csv("https://raw.githubusercontent.com/kkfuyen/Rbootcamp1/master/airquality.csv");

# browse the data
head(data,5);

#column 1 is date, and should be removed. we update the data.
data=data[,-1];
head(data,5)
|
# full model
model.full=lm(PM2.5 ~ .,data = data)

#statistics for full model
(summary.full=summary(model.full))

# reduced model
model.reduced=lm(PM2.5 ~ CO +O3 +PM10 +Temperature+Humidity+WindDegree,data = data)

#statistics for reduced model
(summary.model.reduced=summary(model.reduced))

#Call anova. the anova function only take the lm() objects as input parameters.
anova(model.full,model.reduced)
```

Example of Classification for object Detection

```
#### Install the package ####
devtools::install_github("bnosac/image",
                        subdir = "image.darknet", build_vignettes = TRUE)

#### Exploration ####
# define the jpg file in the internet
url<- 'https://raw.githubusercontent.com/kkfuyen/DA4IoT/master/NTUST.jpg'

#download locally and assign into temp variable
temp <- tempfile()
download.file(url,temp,mode="wb")

# display the jpeg file by using jpeg library
library(jpeg)
jpg = readJPEG(temp,native=TRUE)
plot(0:1,0:1,type="n",ann=FALSE,axes=FALSE)
rasterImage(jpg,0,0,1,1)

#### train by yolo ####
library(image.darknet)
weights = system.file(package="image.darknet","models","tiny-yolo-voc.weights")
labels = system.file(package="image.darknet","include","darknet","data","voc.names")
# which objects can be detected
readLines(labels)

yolo_tiny_voc =image_darknet_model(type="detect", model = "tiny-yolo-voc.cfg",
                                weights= weights,
                                labels= labels )
```

Detect object

```
#### detect objects ####  
#file =file.path(getwd(), temp) # if temp is local file name  
image_darknet_detect(file =temp,object = yolo_tiny_voc)  
  
#### display the prediction file ####  
library(imager)  
im<-load.image("predictions.png")  
plot(im)  
  
#### clean the temp file ####  
file.remove(temp)
```

K-means

- ◆ The Basic Ideas of K-means
 - ◆ <https://github.com/kkfyuen/Rbootcamp2/blob/master/kmeans.pdf>
- ◆ Develop k-means and use k-means function in code
 - ◆ <https://github.com/kkfyuen/Rbootcamp2/blob/master/Chapter2Kmeans.pdf>

Use k-means function in R code

```
# K-means Clustering

# explore the data
head(iris)
str(iris)
levels(iris$Species)
pairs(iris[,5])

# perform clustering
kiris=kmeans(iris[,5],3)

# confusion matrix
table(kiris$cluster,iris[,5])

# externalCriteria as we know the ground true
library(clusterCrit)
extCriteria(as.integer(kiris$cluster),
            as.integer(iris[,5]),
            c("Jaccard","Rand","Recall","Precision"))|
```


Apply kmeans to Network Optimization

- ◆ The background
- ◆ <https://github.com/kkfyuen/SMAH2018Guide/blob/master/SMAH2018Guide.pdf>
- ◆ The guideline
- ◆ <https://github.com/kkfyuen/Rbootcamp2/blob/master/Chapter3NCO.pdf>

Shortest path

```
library("TSP")
?ETSP
example(ETSP)
|
riderArea=rbind(ridersPoints[ci,],clusterMembers) #rider is the starting point
areaNames=rownames(riderArea)

etsp <- ETSP(riderArea,labels =areaNames) # Euclidean , use tsp to find the Havensine distance
tour <- solve_TSP(etsp,method="nn", start=1L)
stPath=as.integer(tour)
areaNames[stPath]

resultPath=riderArea[areaNames[stPath],]
```

References

- ◆ Intel AI Academy, <https://software.intel.com/en-us/ai/academy>
- ◆ Wikipedia, Internet of things
https://en.wikipedia.org/wiki/Internet_of_things
- ◆ Yuen K.K.F., Guan S. S.-U., Chan K. Y., Palade V., (2018) “Editorial for Special Issue on “Hybrid Evolutionary and Swarm Techniques for Big Data Analytics and Applications”, Big Data Research,14, pp.55-56,
<https://doi.org/10.1016/j.bdr.2018.11.002> .
- ◆ K.K.F. Yuen, Introduction and Tutorial Guide on SUSS-Microsoft Analytics Hackathon 2018, 12 pages report plus dataset and case writing, and codes. (Delivery system)
<https://github.com/kkfyuen/SMAH2018Guide>

References

- ◆ K.K.F. Yuen, 2019, R fundamentals,
<https://github.com/kkfyuen/Rbootcamp1>
- ◆ K.K.F. Yuen, 2019, Data Analytics with R,
<https://github.com/kkfyuen/Rbootcamp2>

THANK YOU!

