

Introduction to Big Data Analytics - From Data Management to Project Deployment

Richie Yu, Yong Poh

richieyyp@gmail.com

<https://www.linkedin.com/in/yong-poh-yu>

<https://www.richieyyptutorialpage.com/>

Personal Demo Home Page:

<https://www.richieyyptutorialpage.com/>

* Open Source and Not-For-Profit Sharing / Demo

What will be covered?

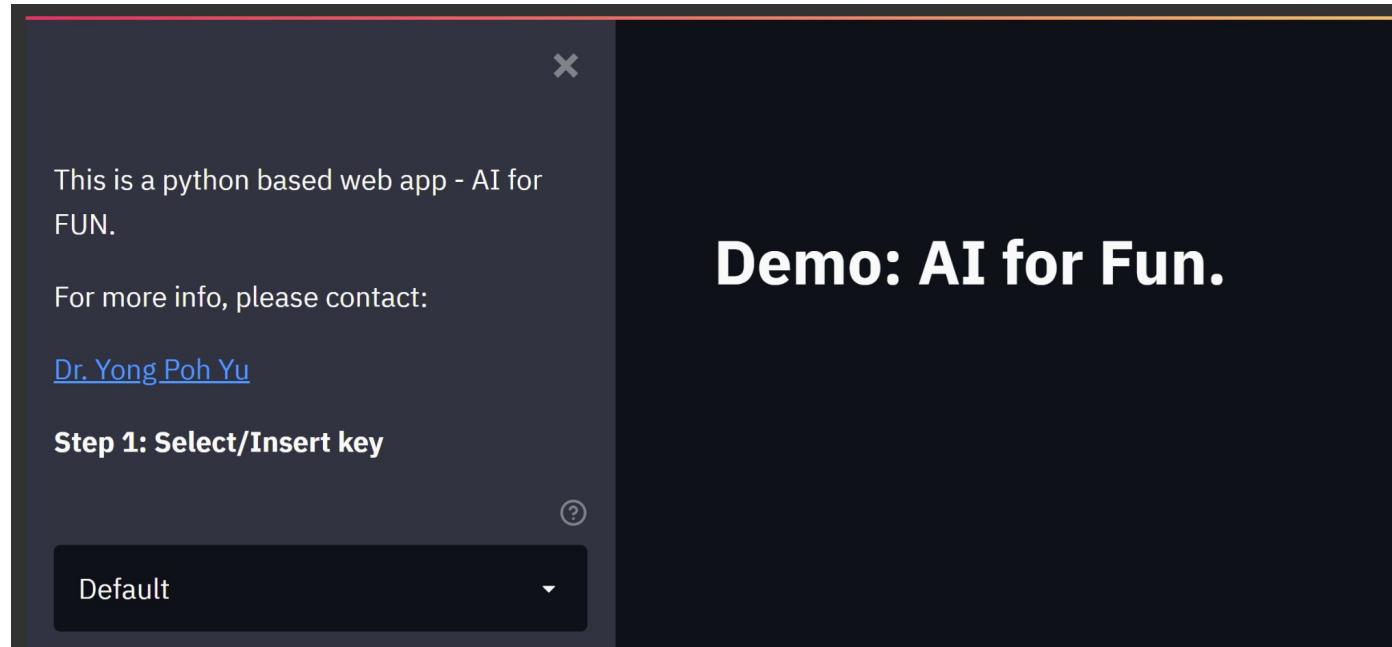
- Introduction to (Big) Data Analytics
- Data Management
- Data Governance
- Type of Analytics
- CRISP-DM
- Turning Data into Actionable Insights
- Data Engineering
- MLOps & DataOps
- Business Intelligence & Data Storytelling
- Practical Activities



Before We continue:

Please make sure you have a
Google account

A: Introduction to Big Data Analytics



Myth!

Data science is a
field for
mathematical geeks



Myth!

Learning a tool is the equivalent of learning data science



Myth!

Data scientists will be replaced by artificial intelligence soon



What is/are ...

Data Science? Artificial Intelligence? Machine Learning?



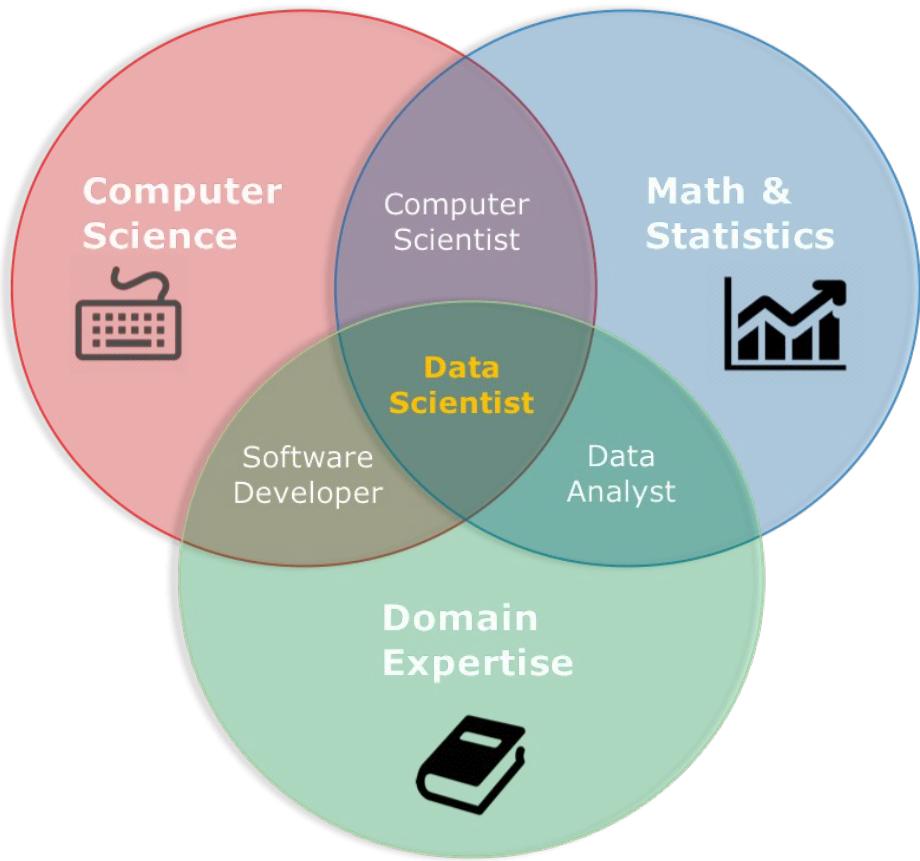


Image Credit: [HERE](#)

Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/ricchieyuyongpoh/creating-value-through-data-transformation>



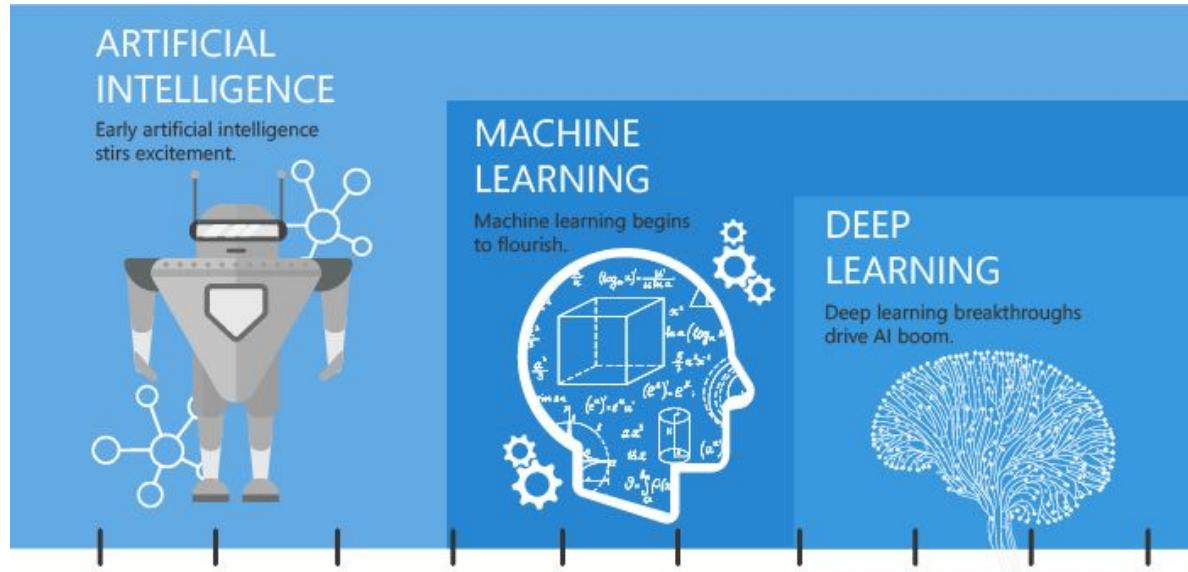


Image Credit: [HERE](#)

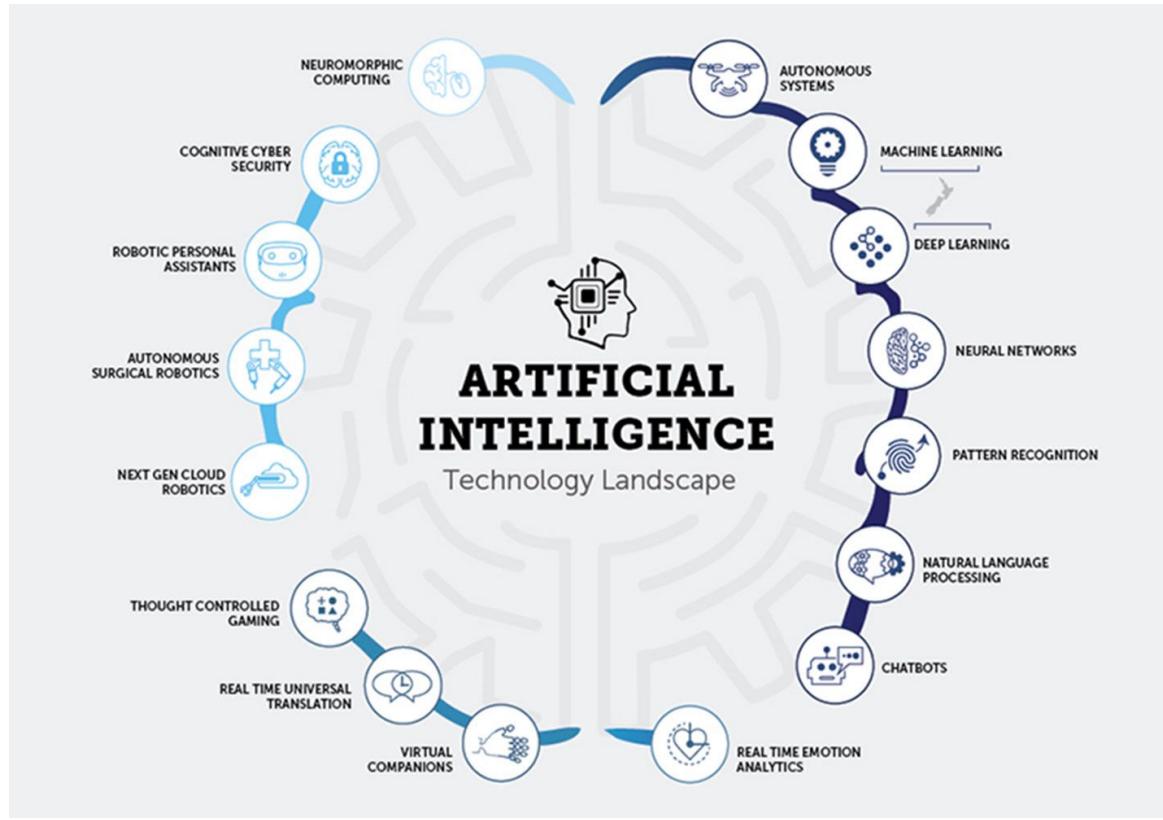


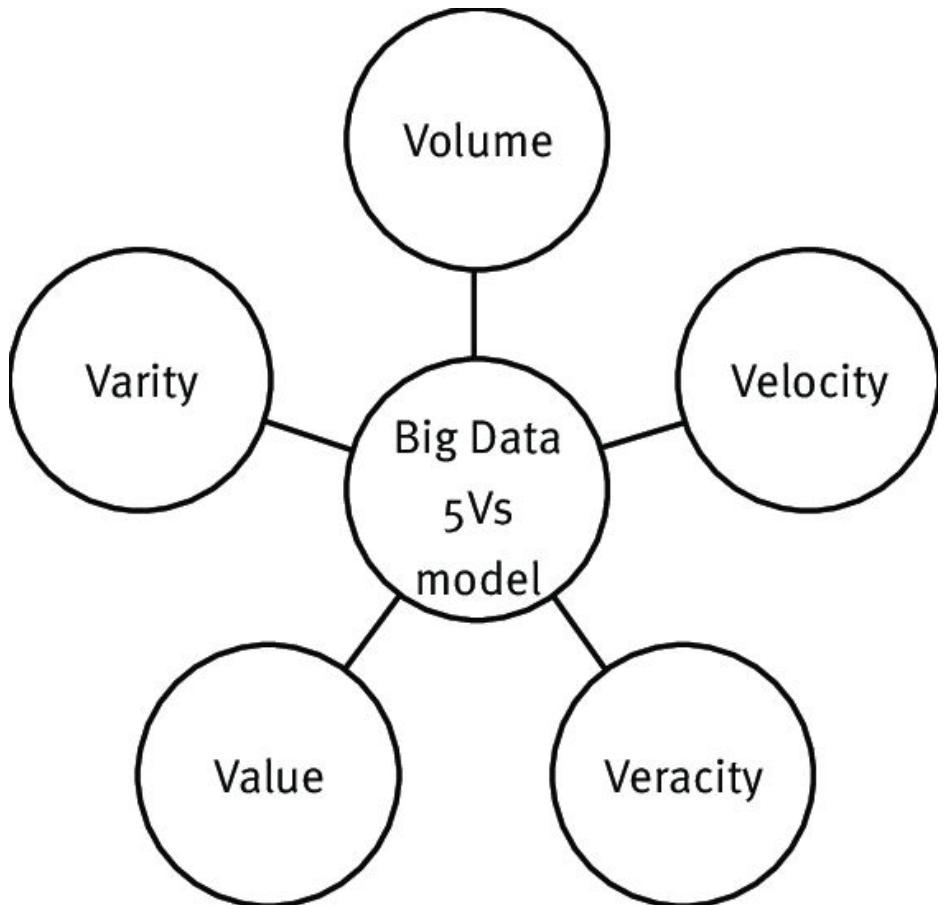
Image Credit: [HERE](#)

Data, AI, ML
Berpisah tiada



Then, what is ...

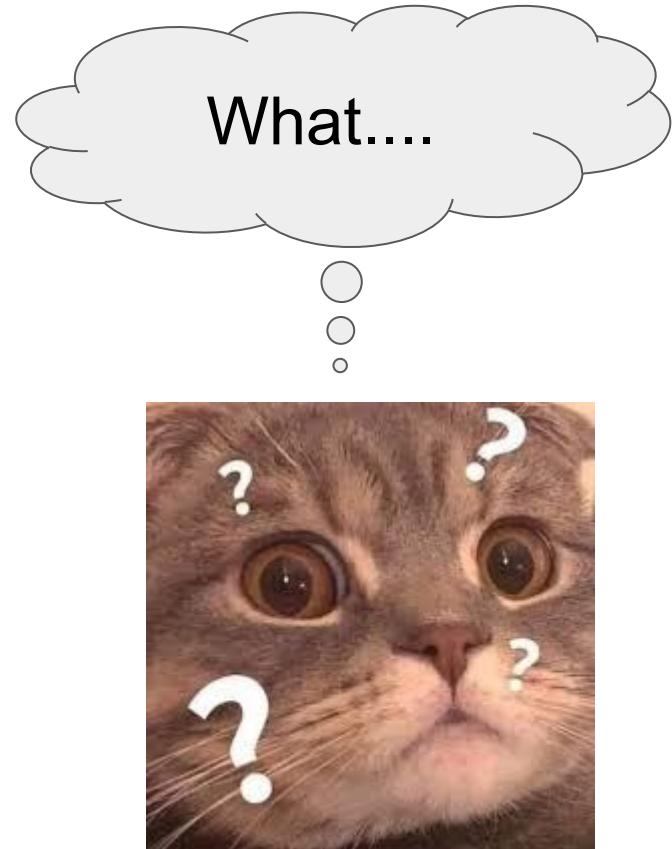
Big Data?



Am i a unicorn?



- **Volume:** very large volumes of data.
- **Variety:** different data formats and types, such as text, video, and audio.
- **Velocity:** The speed at which data is generated and stored is faster than other systems and produced more continuously.
- **Veracity:** We need to check whether the data corresponds with what we expect the data to be



Activity : Discussion



Discuss a use case/case study (based on your profession/domain/discipline) that requires big data analytics.

B: Data Management & Challenges

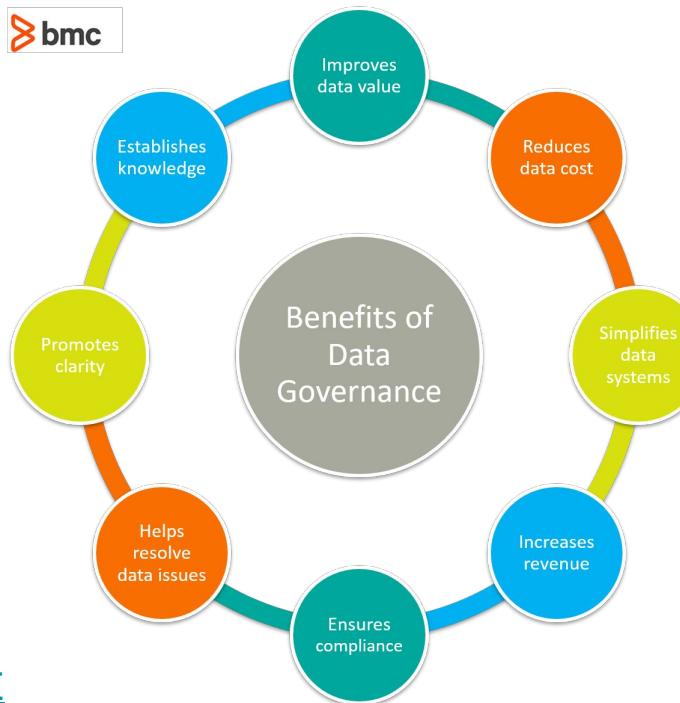


Image Credit: [HERE](#)

Data Management



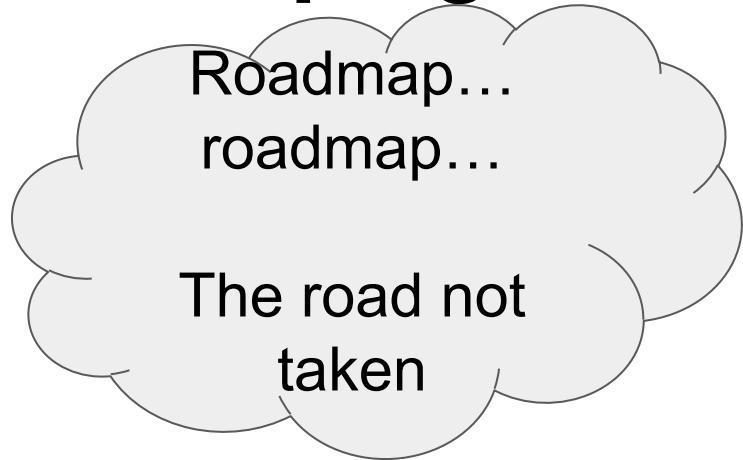
Data Management: *W* and *H*

- What problem do we try to solve? What **value** can big data bring in?
- Who holds the data, who owns the data, and who can access the data?
- What data do we need?
- Where to store the data and how long to keep them?

Data Management: *W* and *H*

- How to ensure the data quality?
- How to analyze and visualize the data?
- How to manage the complexity?

Developing the Right Data Strategy



Developing the Right Data Strategy

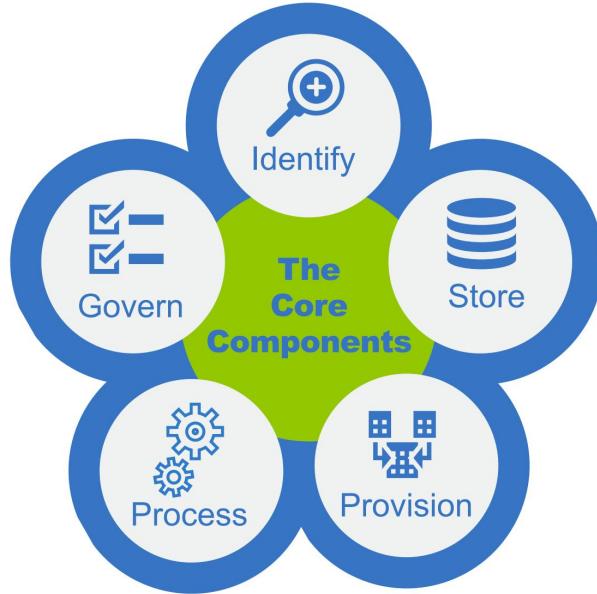


Image Credit: [HERE](#)

Developing the Right Data Strategy

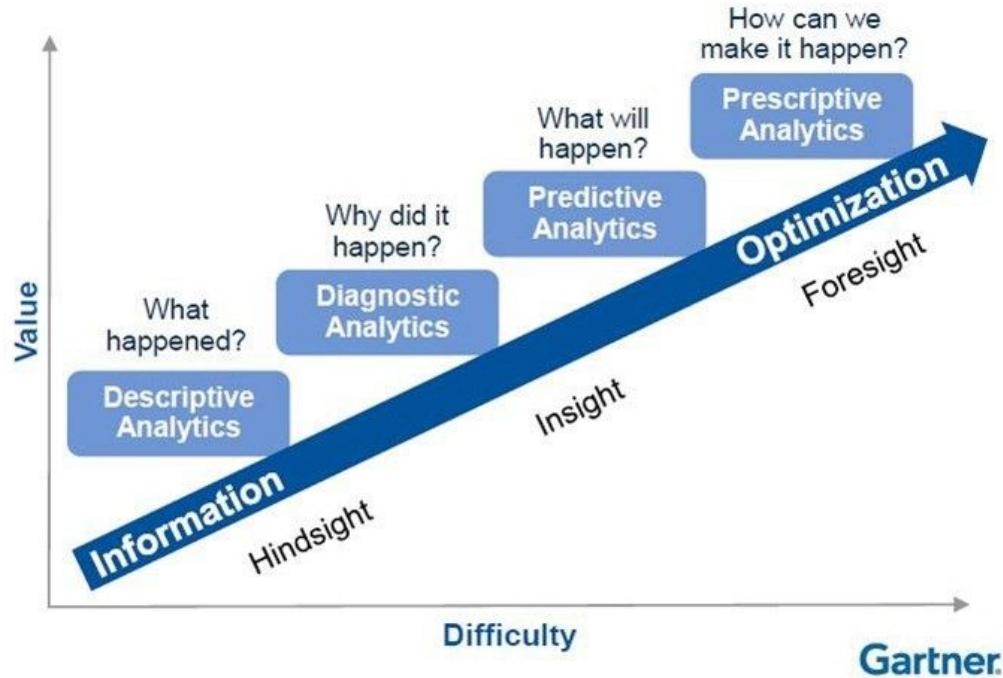


Image Credit: [HERE](#)

Developing the Right Data Strategy

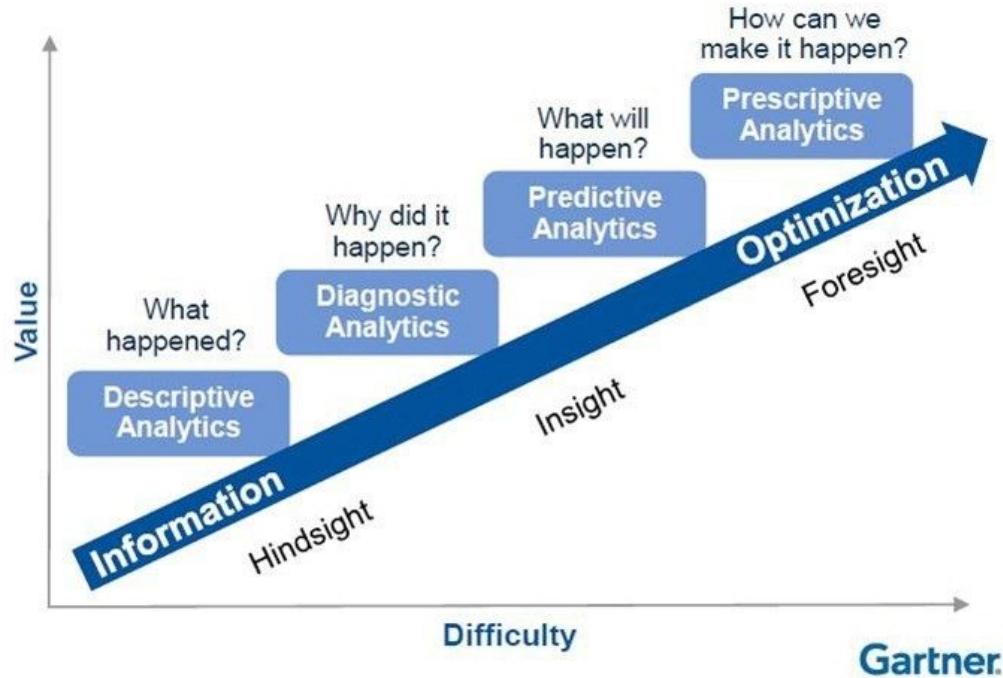


Image Credit: [HERE](#)

Big Data Challenges & Solutions

- Insufficient Knowledge
- Need top-down approach



Big Data Challenges & Solutions

- Confusing variety of Big Data Technologies
- **Talent Development, Outsourcing etc**



Big Data Challenges & Solutions

- Money, Money, Money
- Step-by-step , pay as per use (eg: cloud service)



Side Note: Job Opportunities



JobStreet by SEEK

Search Jobs

MyJobStreet

C

The screenshot shows the JobStreet search interface. At the top, there is a search bar with the query "data scientist". To the right of the search bar is a location input field with the placeholder "Area, city or town". Below the search bar are three filter buttons: "Salary", "Job type", and "Date posted". To the right of these filters is a sorting option "Sort By Relevance" with a dropdown arrow. The main search results area displays the message "1-30 of 2,406 jobs".



Data Engineer

Side Note: Job Opportunities

Employers



malaysianpaygap

Lead data science, manager,
industry data analytics

36/Malay/male

8years

Location KL

PhD in artificial intelligence

Current salary: 18,200 basic, 400 allowance

Data science is the new exciting trends. Salary pay is varied depending on qualifications and experience. MNC companies are willing to pay high if you have the right expertise and experience handling these complex projects. Many people claimed they are data scientists but come short in interview/real world applications. Ability to prove and understand your worth. You will always be in upper hands discussing salary with HR or recruiters

Employees



Side Note: Job Opportunities

Data Scientist

[REDACTED]
[REDACTED]

MYR 1,500 - MYR 2,100

Posted on 24-Feb-22



Big Data Challenges & Solutions

- Complexity of managing data quality
- Proper Data Management



Big Data Challenges & Solutions

- Data Security & Privacy
- Start from the very beginning



Big Data Challenges & Solutions

- Tricky & Tedious Processes
- Proper BDA ecosystem



Big Data Challenges & Solutions

- Scalability
- Proper BDA ecosystem



Big Data Challenges



Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Activity : Discussion



Referring to a use case/case study (based on your profession /domain /discipline) , discuss the challenges that are potentially faced by the data team.

C: Building a Data Team



Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Data Engineer vs Data Analyst vs Data Scientist

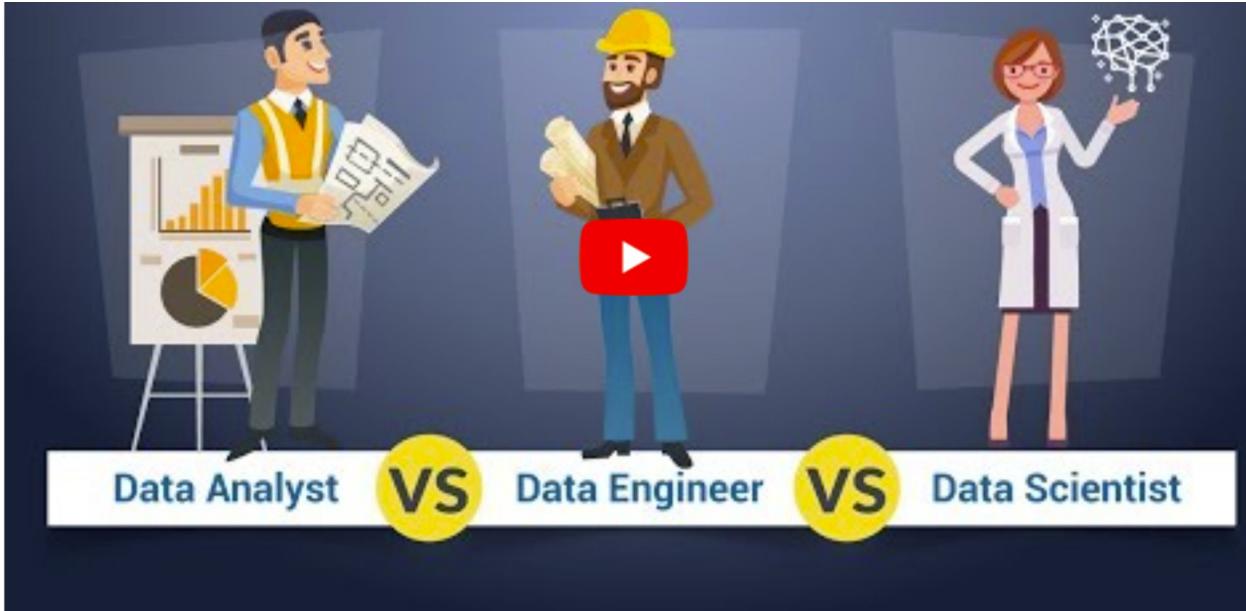


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Data Analyst	Data Engineer	Data Scientist
Data Analyst analyzes numeric data and uses it to help companies make better decisions.	Data Engineer involves in preparing data. They develop, constructs, tests & maintain complete architecture.	A data scientist analyzes and interpret complex data. They are data wranglers who organize (big) data.

Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Data Analyst	Data Engineer	Data Scientist
Data Warehousing	Data Warehousing & ETL	Statistical & Analytical skills
Adobe & Google Analytics	Advanced programming knowledge	Data Mining
Programming knowledge	Hadoop-based Analytics	Machine Learning & Deep learning principles
Scripting & Statistical skills	In-depth knowledge of SQL/ database	In-depth programming knowledge (SAS/R/ Python coding)
Reporting & data visualization	Data architecture & pipelining	Hadoop-based analytics
SQL/ database knowledge	Machine learning concept knowledge	Data optimization
Spread-Sheet knowledge	Scripting, reporting & data visualization	Decision making and soft skills

Image Credit: [HERE](#)

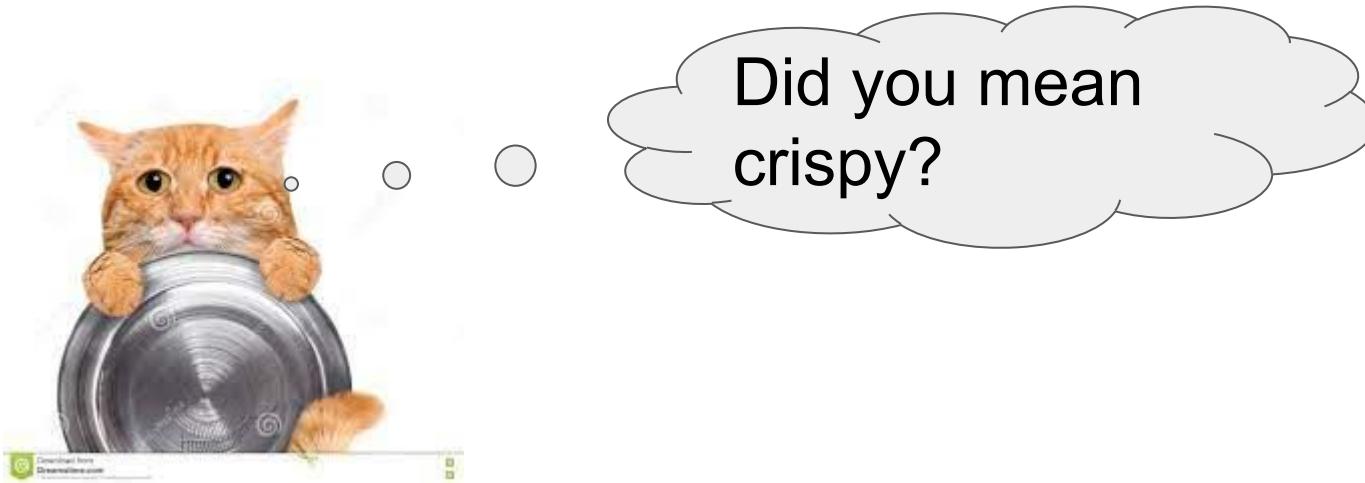
Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Data Analyst	Data Engineer	Data Scientist
Pre-processing and data gathering	Develop, test & maintain architectures	Responsible for developing Operational Models
Emphasis on representing data via reporting and visualization	Understand programming and its complexity	Carry out data analytics and optimization using machine learning & deep learning
Responsible for statistical analysis & data interpretation	Deploy ML & statistical models	Involved in strategic planning for data analytics
Ensures data acquisition & maintenance	Building pipelines for various ETL operations	Integrate data & perform ad-hoc analysis
Optimize Statistical Efficiency & Quality	Ensures data accuracy and flexibility	Fill in the gap between the stakeholders and customer

Image Credit: [HERE](#)

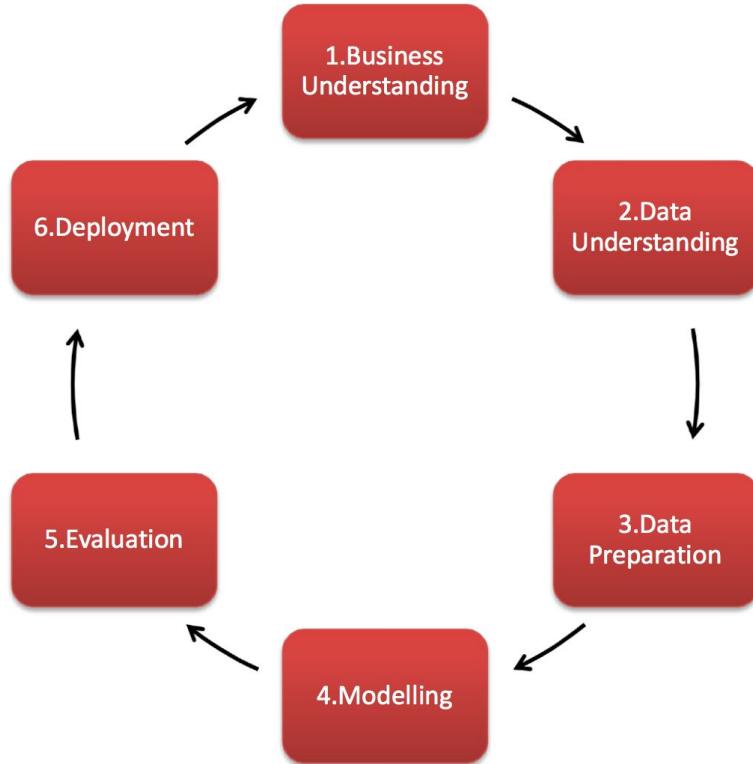
Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

D: Data Science - CRISP-DM Approach



- 1. Measure the right things**
2. Ask the right questions to stakeholders
3. Use segmentation to drive action
4. Use clear visualizations to convey your message
5. Discover the context of your data set
- 6. Build a solid optimization plan**
7. Construct a great hypothesis
8. Integrate data sources
9. Break down organizational silos
- 10. Don't forget to hire smart people**

Source: [HERE](#)



Business
Data

A vertical stack of two grey arrows, one pointing up and one pointing down, positioned between the 'Business' and 'Data' labels.

Image Credit: [HERE](#)

1) Business Understanding



What are the leading factors?

Automate a data-drive solution

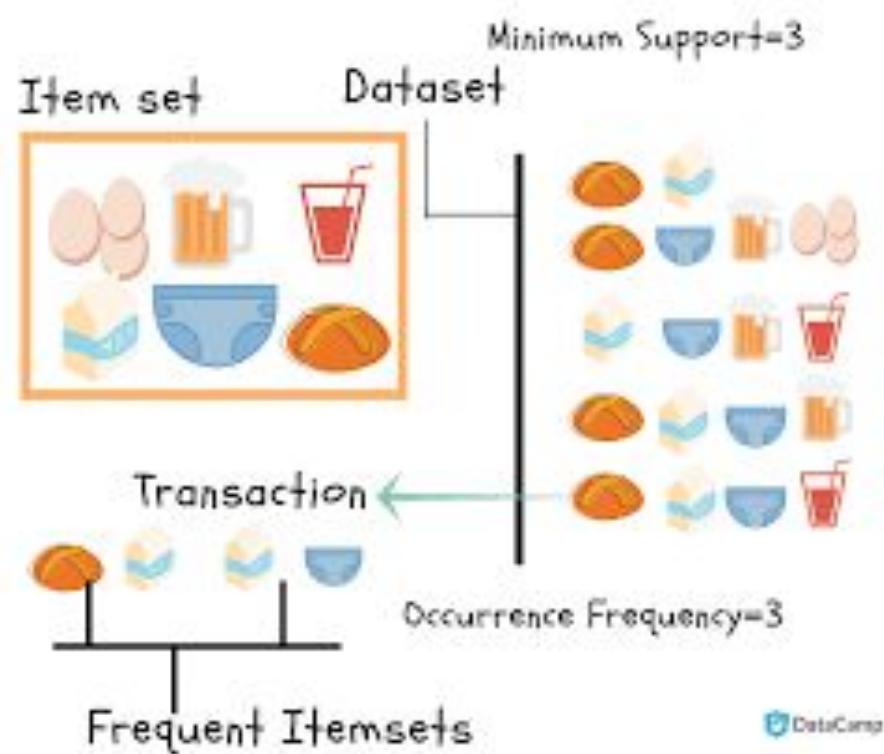
Use Case: Stock Price Forecasting



High Risk,
High Return



Use Case: Market Basket Analysis



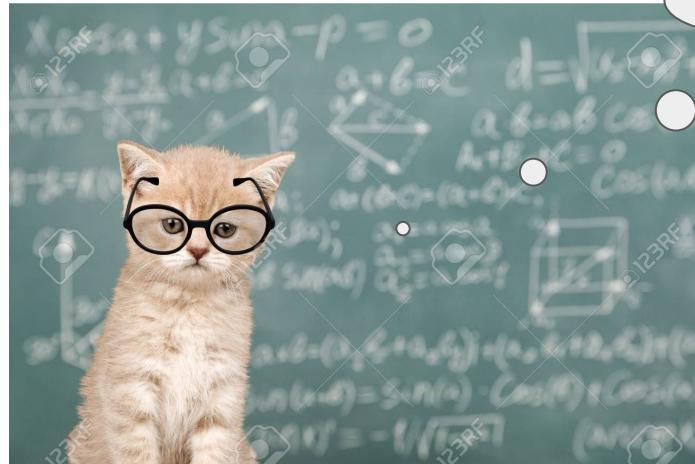
Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Use Case : Customer Segmentation



[https://www.richieyyptutorialpage.com/demo-r-series/
customer-segmentation-using-k-means](https://www.richieyyptutorialpage.com/demo-r-series/customer-segmentation-using-k-means)

What is the problem with the previous dashboard?



Data
Storytelling is
important

Customer Segmentation - Background

- Clustering
- k-means
- Spending behaviors, demographic Information etc

Use Case: Household Income and Food Expenditure



<https://www.richieyyptutorialpage.com/demo-r-series/household-income-and-food-expenditure>

What is the problem with the previous dashboard?



When data is
extraordinary ...

Background

- **Data Visualization**
- **Storytelling**
- **Domain Knowledge**

Use Case 3: Customer Churn Activity



Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

What are the leading factors?

Can you give a list of potential churners?



Background

- Classification
- Random Forest (Machine Learning)
- Garbage in, Garbage out

And MANY MANY MORE:

Marketing	Risk Management	Government	Web	Logistics	Other
Response modeling	Credit risk modeling	Tax avoidance	Web analytics	Demand forecasting	Text analytics
Net lift modeling	Market risk modeling	Social security fraud	Social media analytics	Supply chain analytics	Business process analytics
Retention modeling	Operational risk modeling	Money laundering	Multivariate testing		
Market basket analysis	Fraud detection	Terrorism detection			
Recommender systems					
Customer segmentation					

2) Data Understanding



What are the X?

What is the y?

Missing value?

Outliers?

Get Your Hands Dirty



Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Activity 1

Get the Dataset at [HERE](#). Make a copy.

How many observations (rows)
do we have? _____

How many columns (variables)
do we have? _____

A	B	C	D	E	F	G	H	I	J	K	L	M	N
PostalCode	HashCode	Age	Gender	Payment Method	rowNumber	LastTransaction	ChurnDate						
49278	BOl2gvxC	64	male	credit card	1	2012-04-17 02:05:40	2014-01-24 18:27:13						
39982	UC8cDTW	35	male	cheque	2	2011-11-25 06:58:03	2012-08-09 13:01:39						
87213	tKlbadnh	25	female	credit card	3	2012-02-15 17:29:26							
38548	RcW2Pb3w	39	female	credit card	4	2010-10-09 11:22:28	2013-11-07 10:27:31						
38794	z9twAAJ4	39	male	credit card	5	2012-06-13 10:13:08							
44573	akWNQl4e	28	female	cheque	6	2010-07-16 09:39:10	2011-06-23 07:08:53						
70936	gIhPDLzY	21	female	credit card	7	2012-03-15 22:17:03							
71302	Pn6fkbuL	48	male	credit card	8	2011-06-16 21:46:18							
49705	3rGPBX98	70	female	credit card	9	2011-03-30 14:17:44	2012-07-05 02:34:33						
36049	9Eng7yIO	36	male	credit card	10	2013-04-17 18:06:59							
26323	uP7dRmDK	22	male	credit card	11	2013-03-11 17:37:27							
42376	XlxhOIu	53	female	cash	12	2010-08-31 17:39:06	2013-08-31 17:21:18						
42215	KYmSR2vE	27	male	cash	13	2011-07-15 07:16:15							
80059	EOGj6tDH	234	männlich	credit card	14	2014-02-26 18:22:25							
57984	yTDIEBxC	22	male	cash	15	2011-05-09 06:30:35							
52245	3ANQ9shn	49	female	credit card	16	2011-02-18 10:22:00	2012-04-03 06:13:43						
56625	BDEPLkmg	24	female	cash	17	2011-01-22 03:19:03	2013-02-28 23:50:54						
50003	8vB4XorN	45	männlich	credit card	18	2013-02-11 21:16:18							
37852	1SDIHc7	45	male	credit card	19	2011-02-06 11:31:19							
66713	gQBQD28lw	66	female	cash	20	2011-03-09 23:04:48	2013-04-19 04:27:33						
70052	G1c0kZqQ	82	female	cash	21	2010-10-10 10:29:03	2013-12-17 15:21:46						
44272	UC8cDTW	35	female	credit card	22	2010-10-14 07:48:11							
71962	n5YVLU7v	17	female	credit card	23	2011-08-14 16:40:26							

Activity 1

No	Dimension	Data Type	Comment
1	PostalCode	Integer	Whole number, therefore use integer
2	HashCode	Polynomial	Different strings, so polynomial is fine.
3	Age	Integer	Whole number, therefore use integer
4	Gender	Polynomial	Should be binomial, but inconsistent data type
5	PaymentMethod	Polynomial	3 payment methods: what are they?
6	rowNumber	Integer	Integer
7	LastTransaction	Date_time	Contains date
8	ChurnDate	Date_time	Contains date

3) Data Preparation

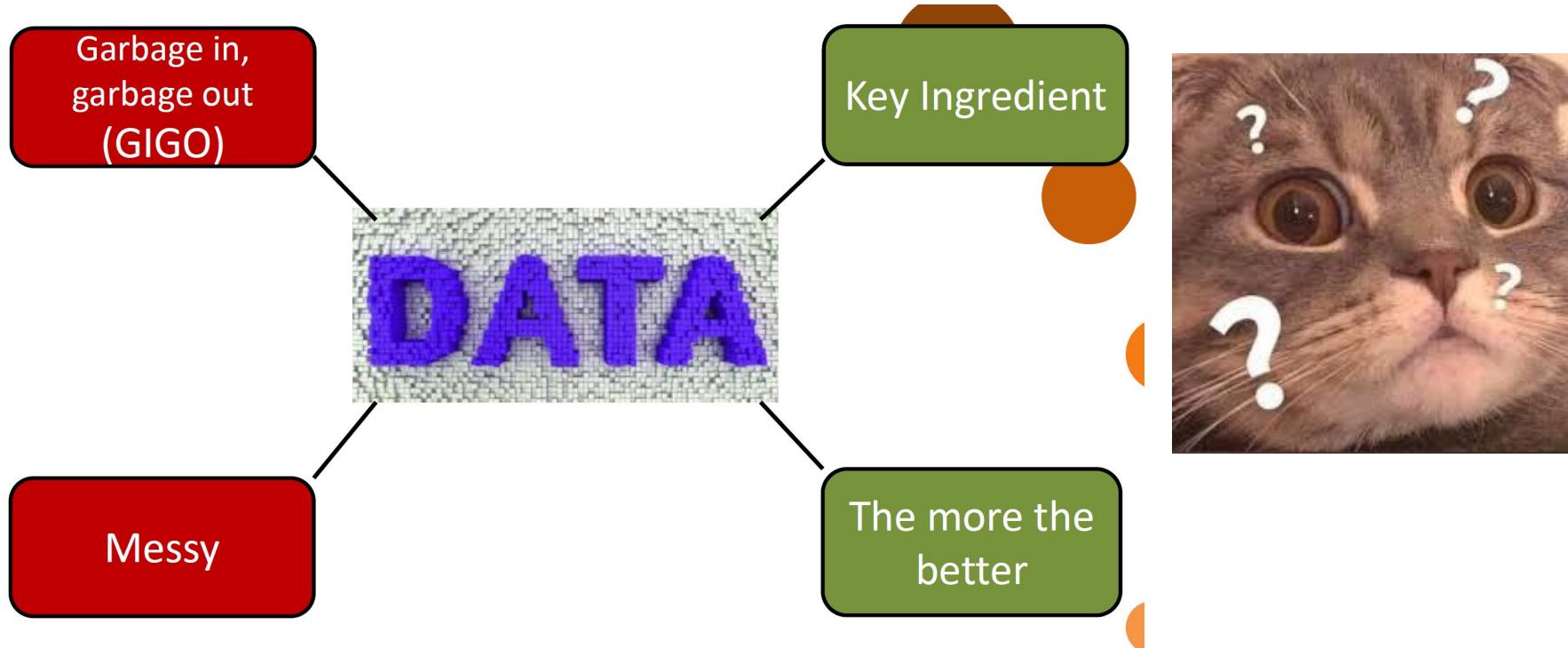


Missing value?

Outliers?

New Column/Variable/Feature?

I heard that ...



I heard that ...

DATA SCIENCE = DATA CLEANING



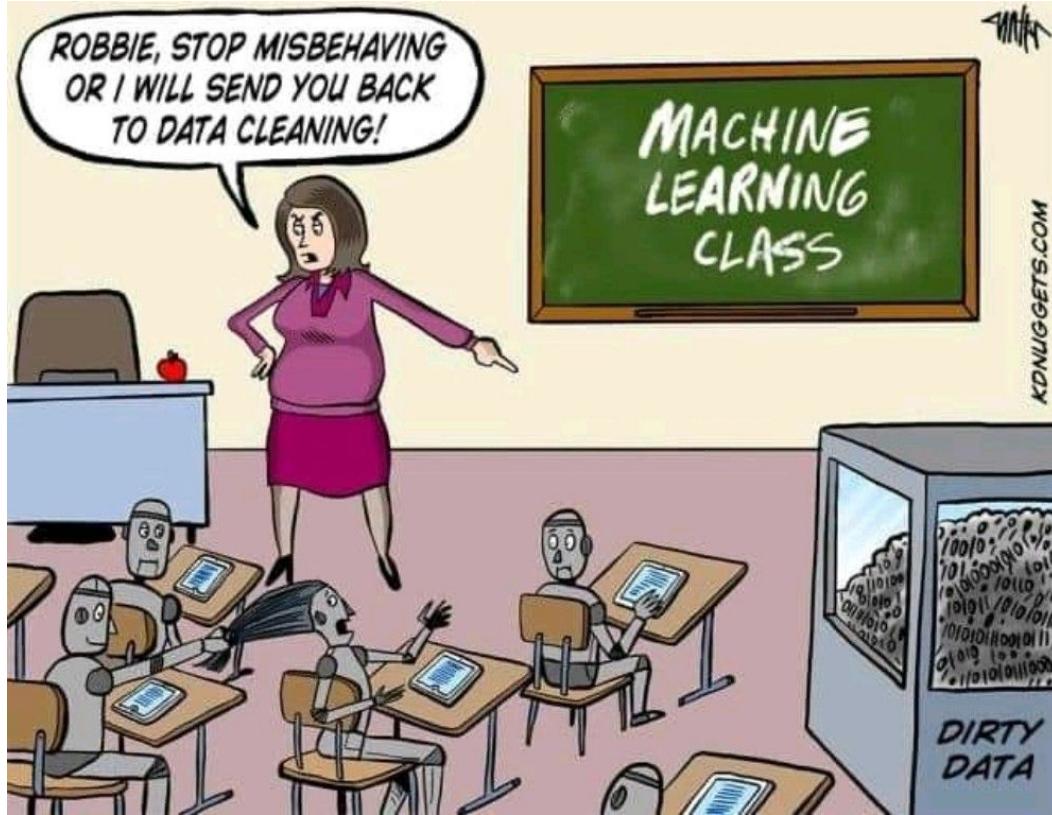


Image Credit: [HERE](#)

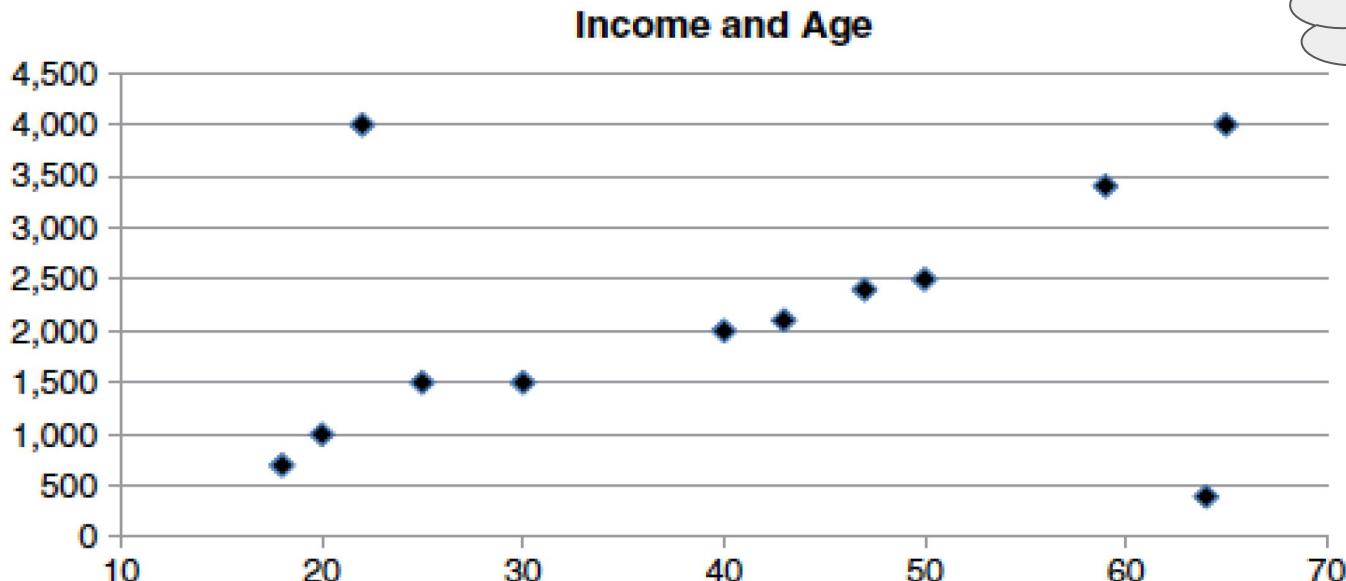
Data, AI, ML
Berpisah tiada

I heard that data is incomplete...

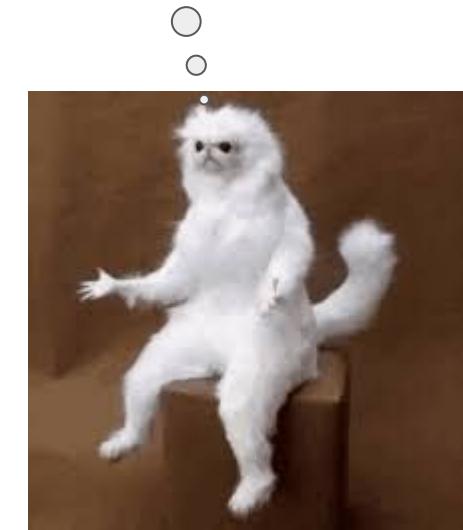
ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	1,800		620	Churner
2	28	1,200	Single		Nonchurner
3	22	1,000	Single	?	Nonchurner
4	60	2,200	Widowed	700	Churner
5	58	2,000	Married		Nonchurner
6	44				Nonchurner
7	22	1,200	Single		Nonchurner
8	26	1,500	Married	350	Nonchurner
9	34		Single		Churner
10	50	2,100	Divorced		Nonchurner



I heard that data always surprises us...



Win liao
lo



I heard that data quality is important...



Activity 2: Dealing with Missing Values

4 factors to consider when dealing with missing data:

- i. If an attribute contains too many missing data, consider discarding it.
 - ii. In a categorical (ChurnDate) attribute with missing values, consider introducing a new category eg. “*unknown*”. (“*unknown*” means something very different from missing)
 - iii. If only few rows contain missing values, consider removing those rows.
 - iv. If we cannot afford to discard data, consider imputing values into the missing values.
- Analyze your data. Which attributes possess missing values?

Activity 2: Dealing with Missing Values

For *Age* and *Gender*, only 1 missing value. Doesn't make much difference to remove (filter) them.

PostalCode	HashCode	Age	Gender	ChurnDate
49278	BOI2gvcX	64	male	05:40 2014-01-24 18:27:13
39982	IJc8cDTW	35	male	58:03 2012-08-09 13:01:39
87213	tKlbadnh	25	female	29:26
38548	RcW2Pb3w	39	female	22:28 2013-11-07 10:27:31
38794	z9twA4AJ	39	male	13:08
44573	akWNQl4e	28	female	39:10 2011-06-23 07:08:53
70936	glrPDLzY	21	female	17:03
71302	Pn6FkbuL	48	male	46:18
49705	3rGPBX98	70	female	17:44 2012-07-05 02:34:33
36049	9Eng7yl0	36	male	06:59
26323	uP7dRmDK	22	male	37:27
42376	XlxhfOlo	53	female	39:06 2013-08-31 17:21:18
42215	KYmSR2vE	27	male	16:15
80059	EOGj6tDH	234	männlich	22:25
57984	yTDIEBXc	22	male	30:35
52245	3ANQ9shn	49	female	22:00 2012-04-03 06:13:43
56625	BDEPLKmG	24	female	19:03 2013-02-28 23:50:54
50003	8vB4XorN	45	männlich	16:18
37852	1SDtHcf7	45	male	credit card

Activity 2: Dealing with Missing Values

The screenshot shows a Microsoft Excel spreadsheet titled "Copy of Customer Data" in XLSX format. The spreadsheet contains data from rows 1 to 24 across columns A through H. Column A is labeled "PostalCo", column B is "HashCode", column C is "Age", column D is "Gender", column E is "Payment Method", column F is "rowNumber", column G is "LastTransaction", and column H is "ChurnDate". The "Age" column (C) has a dropdown menu open, showing filter options: "Sort ↗ A", "Sort by color", "Filter by color", and "Filter by condition". Under "Filter by condition", the option "Is not empty" is selected. Other visible filter options include "(Blanks)", "2", "17", and "18". The "LastTransaction" column (G) shows various dates, and the "ChurnDate" column (H) shows various times.

	A	B	C	D	E	F	G	H
1	PostalCo	HashCode	Age	Gender	Payment Method	rowNumber	LastTransaction	ChurnDate
2	49278	BOI2gvcX	64				12-04-17 02:05:40	2014-01-24 18:27:13
3	39982	IJC8cDTW	35				11-11-25 06:58:03	2012-08-09 13:01:39
4	87213	tKlbadnh	25				12-02-15 17:29:26	
5	38548	RcW2Pb3w	39				10-10-09 11:22:28	2013-11-07 10:27:31
6	38794	z9twA4AJ	39				12-06-13 10:13:08	
7	44573	akWNQl4e	28				10-07-16 09:39:10	2011-06-23 07:08:53
8	70936	glrPDLzY	21				12-03-15 22:17:03	
9	71302	Pn6FkbuL	48				11-06-16 21:46:18	
10	49705	3rGPBX98	70				11-03-30 14:17:44	2012-07-05 02:34:33
11	36049	9Eng7yl0	36				13-04-17 18:06:59	
12	26323	uP7dRmDK	22				13-03-11 17:37:27	
13	42376	XlxhfOlo	53				10-08-31 17:39:06	2013-08-31 17:21:18
14	42215	KYmSR2vE	27				11-07-15 07:16:15	
15	80059	EOGj6tDH	234				14-02-26 18:22:25	
16	57984	yTDIEBXc	22				11-05-09 06:30:35	
17	52245	3ANQ9shn	49				11-02-18 10:22:00	2012-04-03 06:13:43
18	56625	BDEPLKmG	24				11-01-22 03:19:03	2013-02-28 23:50:54
19	50003	8vB4xOrN	45				13-02-11 21:16:18	
20	37852	1SDtHCF7	45				11-02-06 11:31:19	
21	66713	gQBDZ8Lw	66				11-03-09 23:04:48	2013-04-19 04:27:33
22	70052	G1c0kZqQ	82				10-10-10 10:29:03	2013-12-17 15:21:46
23	44272	IJC8cDTW	35				10-10-14 07:48:11	
24	71962	n5YVLJ7v	17				11-08-14 16:40:26	

Activity 3: Creating a new Column

1. For ChurnDate, it involves semantics and has meaning, hence removing them is NOT an option. (But we know, that if the ChurnDate is missing, the customer is still loyal)
2. So let's generate a “*Churn*” attribute that will tell us whether a customer is ‘loyal’ or ‘churn’.

Activity 3: Creating a new Column

A screenshot of a Google Sheets spreadsheet. The top menu bar shows various icons and settings like 100% zoom, currency, and font size. The formula bar contains the formula: =ArrayFormula(if(H2:H999="", "loyal", "false")). The spreadsheet has 12 columns labeled A through K. Column A is 'PostalCo', B is 'HashCode', C is 'Age', D is 'Gender', E is 'Payment Method', F is 'rowNumber', G is 'LastTransaction', H is 'ChurnDate', I is 'Churn' (which is the target column), J is blank, and K is blank. The first row has a green background. The data rows show various user IDs, their details, and their 'Churn' status. Row 11 is highlighted with a green background.

	A	B	C	D	E	F	G	H	I	J	K
	PostalCo	HashCode	Age	Gender	Payment Method	rowNumber	LastTransaction	ChurnDate	Churn		
1	49278	BOI2gvcX	64	male	credit card	1	2012-04-17 02:05:40	2014-01-24 18:27:13	false		
2	39982	IJC8cDTW	35	male	cheque	2	2011-11-25 06:58:03	2012-08-09 13:01:39	false		
3	87213	tKlbadnh	25	female	credit card	3	2012-02-15 17:29:26		loyal		
4	38548	RcW2Pb3w	39	female	credit card	4	2010-10-09 11:22:28	2013-11-07 10:27:31	false		
5	38794	z9twA4AJ	39	male	credit card	5	2012-06-13 10:13:08		loyal		
6	44573	akWNQl4e	28	female	cheque	6	2010-07-16 09:39:10	2011-06-23 07:08:53	false		
7	70936	glrPDLzY	21	female	credit card	7	2012-03-15 22:17:03		loyal		
8	71302	Pn6FkbuL	48	male	credit card	8	2011-06-16 21:46:18		loyal		
9	49705	3rGPBX98	70	female	credit card	9	2011-03-30 14:17:44	2012-07-05 02:34:33	false		
10	36049	9Eng7yl0	36	male	credit card	10	2013-04-17 18:06:59		loyal		
11	26323	uP7dRmDK	22	male	credit card	11	2013-03-11 17:37:27		loyal		

Activity 4: Removing Outliers/Extraordinary Values

Let us assume that in this case, the term and condition states that the minimum age is more than 16 years old. Hence, there are invalid age values (2 years old and 234 years old!)

Challenge: Multiple filters in one column....??

Activity 4: Removing Outliers/Extraordinary Values

Solution - Customer Data [XLSX](#)

File Edit View Insert Format Data Tools Help Last edit was seconds ago

100% \$.0 .00 123 Calibri 11 B I A fx Age

PostalCo	HashCode	Age	Gender	Payment Method	rowNumber	Last Transaction	Churn Date	Churn
49278	BOl2gvcX	64				12-04-17 02:05:40	2014-01-24 18:27:13	churn
39982	IJC8cDTW	35				11-11-25 06:58:03	2012-08-09 13:01:39	churn
87213	tKbadnh	25				12-02-15 17:29:26		loyal
38548	RcW2Pb3w	39				10-10-09 11:22:28	2013-11-07 10:27:31	churn
38794	z9twA4AJ	39				12-06-13 10:13:08		loyal
44573	akWNQl4e	28				10-07-16 09:39:10	2011-06-23 07:08:53	churn
70936	glrPDLzY	21				12-03-15 22:17:03		loyal
71302	Pn6FkbuL	48				11-06-16 21:46:18		loyal
49705	3rGPBX98	70				11-03-30 14:17:44	2012-07-05 02:34:33	churn
36049	9Eng7yI0	36				13-04-17 18:06:59		loyal
26323	uP7dRmDK	22				13-03-11 17:37:27		loyal
42376	XlxhfOlo	53				10-08-31 17:39:06	2013-08-31 17:21:18	churn
42215	KYmsR2vE	27				11-07-15 07:16:15		loyal
57984	yTDIEBxk	22				11-05-09 06:30:35		loyal
52245	3ANQ9shn	49				11-02-18 10:22:00	2012-04-03 06:13:43	churn
56625	BDEPLKmG	24				11-01-22 03:19:03	2013-02-28 23:50:54	churn
50003	8vB4xOrN	45				13-02-11 21:16:18		loyal
37852	1SDtHcf7	45				11-02-06 11:31:19		loyal
66713	gQBDZ8lw	66				11-03-09 23:04:48	2013-04-19 04:27:33	churn
70052	G1c0kZqQ	82				10-10-10 10:29:03	2013-12-17 15:21:46	churn
44272	IJC8cDTW	35				10-10-14 07:48:11		loyal
71962	n5YVLJ7v	17				11-08-14 16:40:26		loyal
30942	tldMRZn	52				10-06-23 03:58:56	2013-05-06 15:12:32	churn

Is not empty

Select all · Clear

✓ 90

✓ 91

2

234

Cancel OK

RapidMiner Data

Activity 5: Cleaning Gender Column

- Gender attribute contains male, female, *manlich* and *weiblich* (German terms for male and female).

Activity 5: Cleaning Gender Column

```
=ArrayFormula(if(D2:D1001="männlich","male",D2:D1001))
```

B	C	D	E	F	G	H	I	J
ashCode	Age	Gender	Payment Method	rowNumber	LastTransaction	ChurnDate	Churn	Gender
OI2gvcX	64	male	credit card	1	2012-04-17 02:05:40	2014-01-24 18:27:13	churn	male
C8cDTW	35	male	cheque	2	2011-11-25 06:58:03	2012-08-09 13:01:39	churn	male
Kibadnh	25	female	credit card	3	2012-02-15 17:29:26		loyal	female
W9PQZ	28	female	bank transfer	4	2012-10-20 11:22:20	2012-11-07 10:27:21	loyal	female

Activity 5: Cleaning Gender Column

The screenshot shows a Google Sheets interface with a formula bar at the top containing the formula: =ArrayFormula(if(J2:J1001="weiblich", "female", J2:J1001)). The formula is intended to replace the German word "weiblich" with "female" across the entire range J2:J1001.

HashCode	Age	Gender	Payment Method	rowNumber	LastTransaction	ChurnDate	Churn	Gender	Gender
BOl2gvcX	64	male	credit card	1	2012-04-17 02:05:40	2014-01-24 18:27:13	churn	male	male
IJC8cDTW	35	male	cheque	2	2011-11-25 06:58:03	2012-08-09 13:01:39	churn	male	male
tKlbardnh	25	female	credit card	3	2012-02-15 17:29:26		loyal	female	female
RcW2Pb3w	39	female	credit card	4	2010-10-09 11:22:28	2013-11-07 10:27:31	churn	female	female
z9twA4AJ	39	male	credit card	5	2012-06-13 10:13:08		loyal	male	male
akWNQl4e	28	female	cheque	6	2010-07-16 09:39:10	2011-06-23 07:08:53	churn	female	female
glrPDLzY	21	female	credit card	7	2012-03-15 22:17:03		loyal	female	female

Activity 6: Getting the First Character of texts in the PostalCode Column

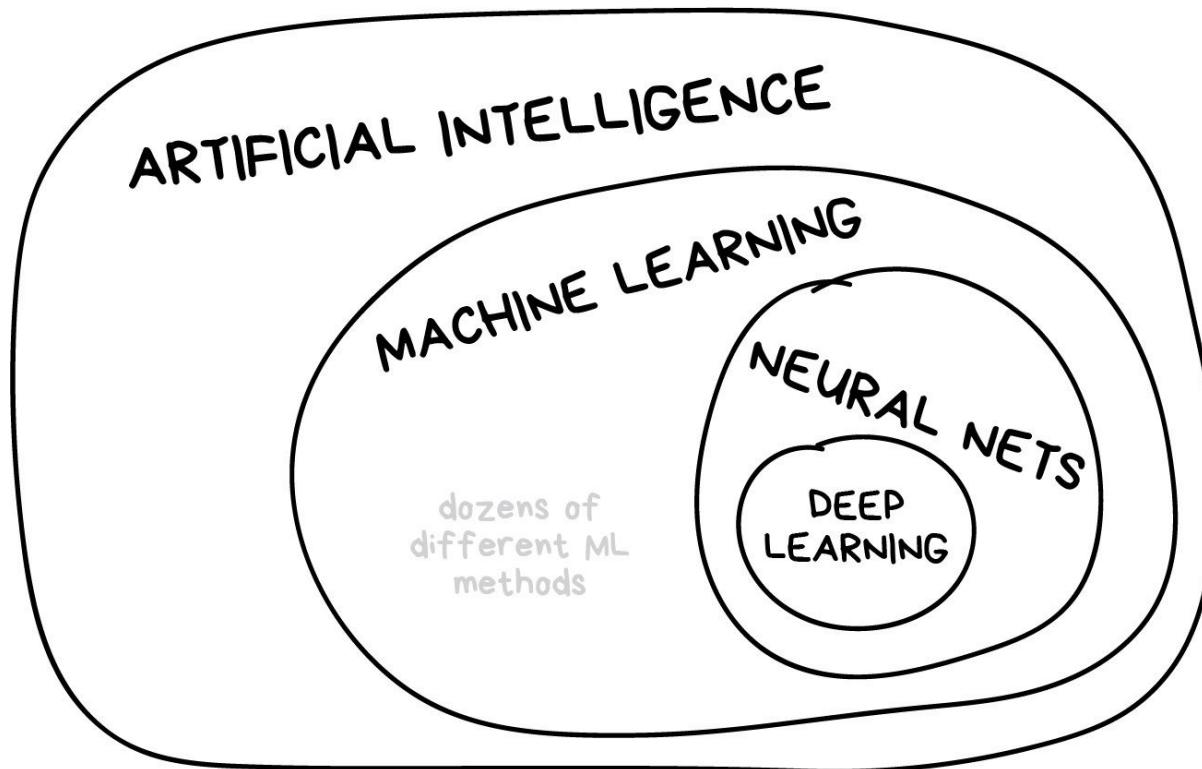
A screenshot of a Google Sheets spreadsheet. The formula `=ArrayFormula(LEFT(A2:A1001))` is entered in cell A2. The spreadsheet contains data from rows 2 to 1001 across columns B through M. The columns represent various customer attributes: ShCode, Age, Gender, Payment Method, rowNumber, LastTransaction, ChurnDate, Churn, Gender, Gender, and PostalCode. The PostalCode column shows values like '4', '3', '8', '3', '3', '4', and '7'. The formula in A2 is highlighted with a blue border.

B	C	D	E	F	G	H	I	J	K	L	M
shCode	Age	Gender	Payment Method	rowNumber	LastTransaction	ChurnDate	Churn	Gender	Gender	PostalCode	
0I2gvcX	64	male	credit card	1	2012-04-17 02:05:40	2014-01-24 18:27:13	churn	male	male	4	
8cDTW	35	male	cheque	2	2011-11-25 06:58:03	2012-08-09 13:01:39	churn	male	male	3	
badnh	25	female	credit card	3	2012-02-15 17:29:26		loyal	female	female	8	
W2Pb3w	39	female	credit card	4	2010-10-09 11:22:28	2013-11-07 10:27:31	churn	female	female	3	
twA4AJ	39	male	credit card	5	2012-06-13 10:13:08		loyal	male	male	3	
WNQI4e	28	female	cheque	6	2010-07-16 09:39:10	2011-06-23 07:08:53	churn	female	female	4	
PDLzY	21	female	credit card	7	2012-03-15 22:17:03		loyal	female	female	7	

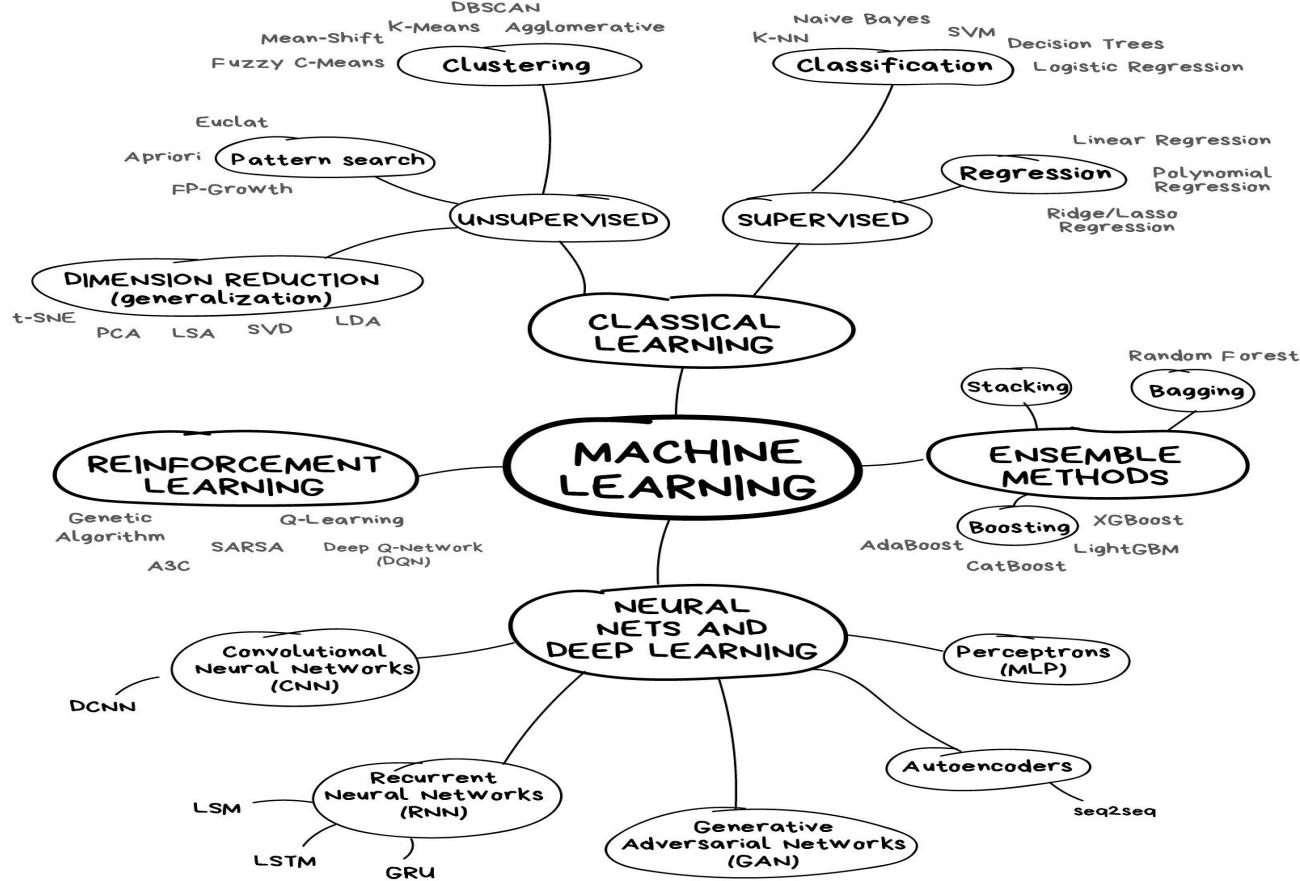
4) Modelling

Are you looking for a model?



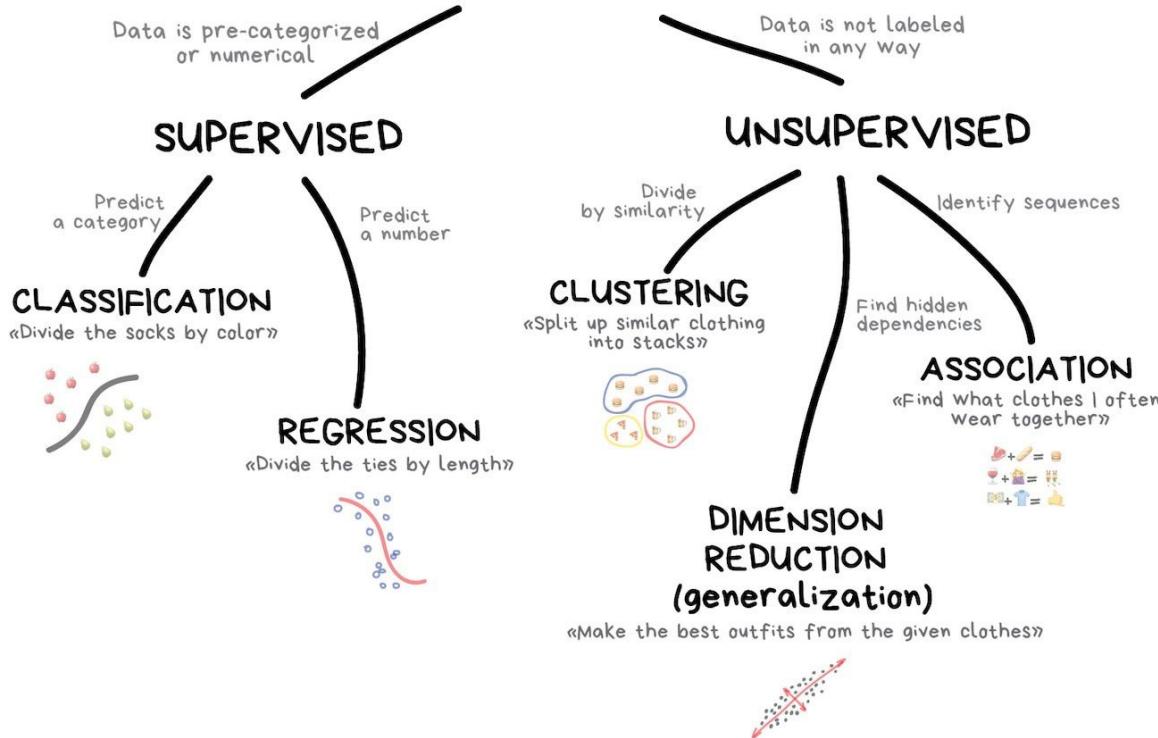


Reference: [here](#)

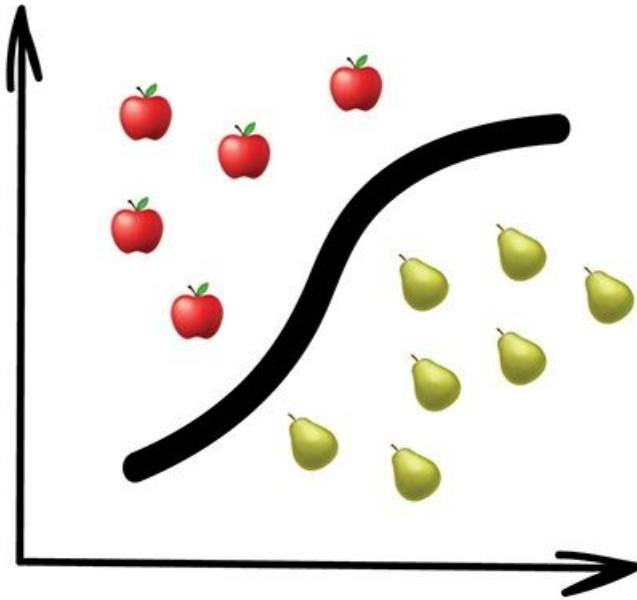


Reference: [here](#)

CLASSICAL MACHINE LEARNING



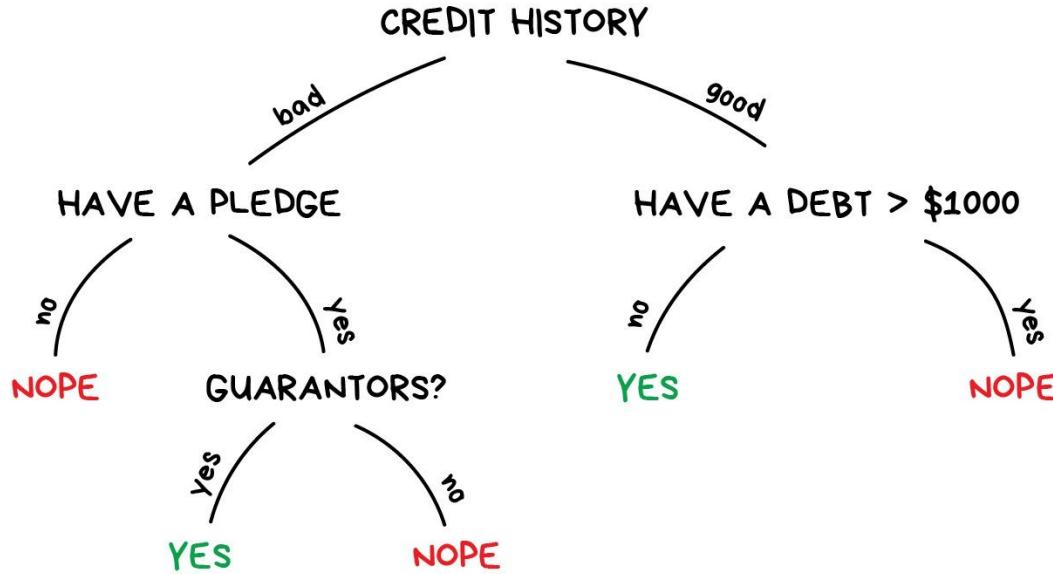
Reference: [here](#)



Classification

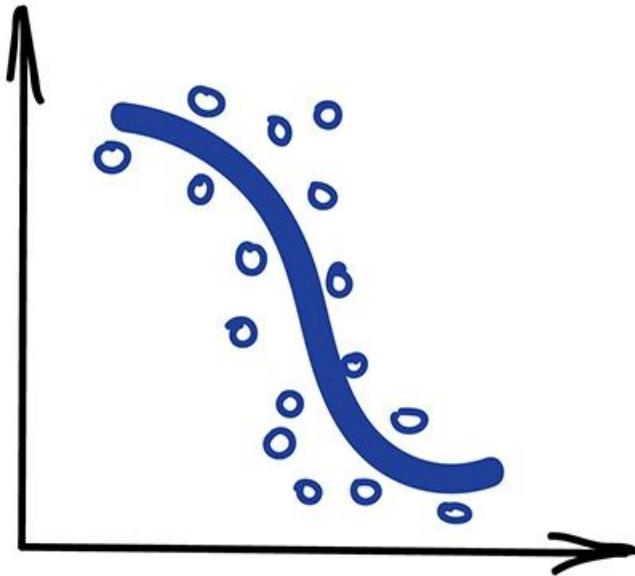
Reference: [here](#)

GIVE A LOAN?



DECISION TREE

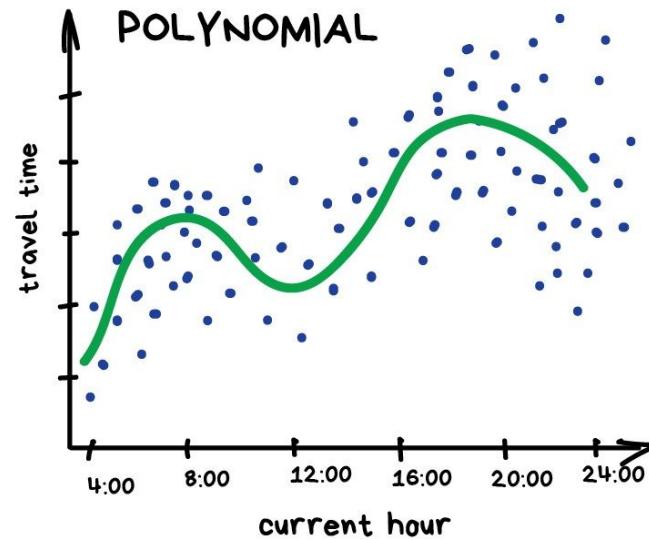
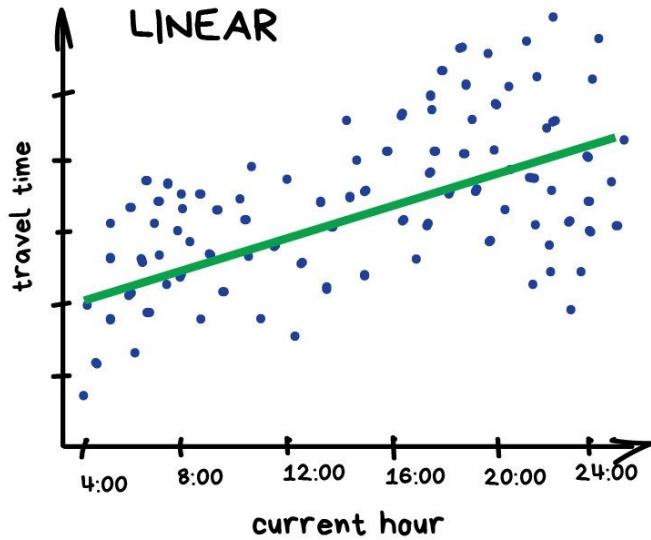
Reference: [here](#)



Regression

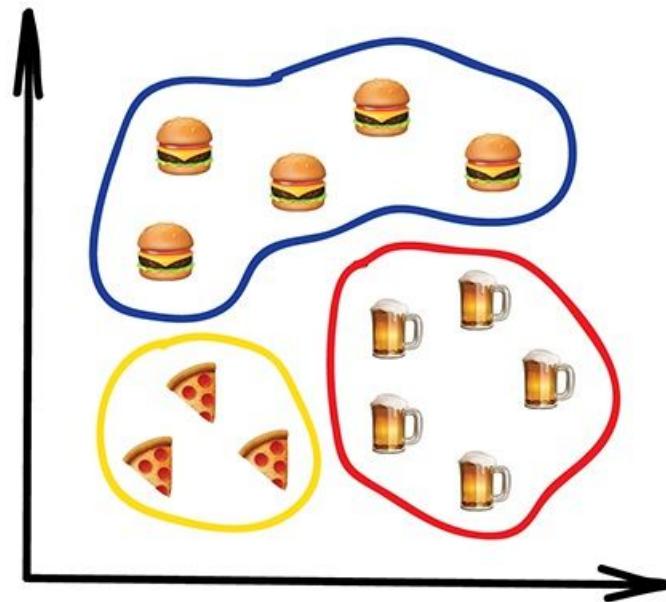
Reference: [here](#)

PREDICT TRAFFIC JAMS



REGRESSION

Reference: [here](#)

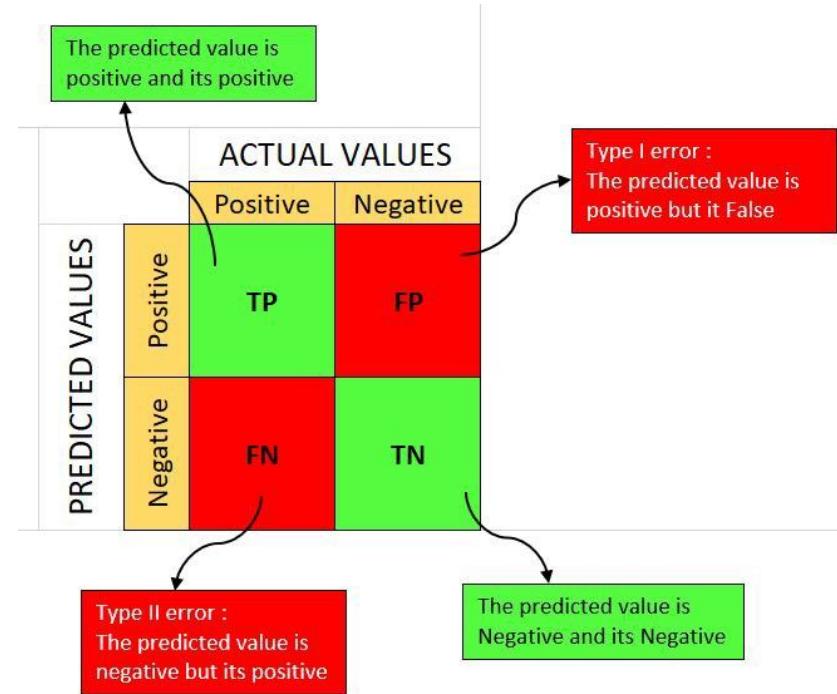


Reference: [here](#)

5) Evaluation



Classification: Confusion Matrix



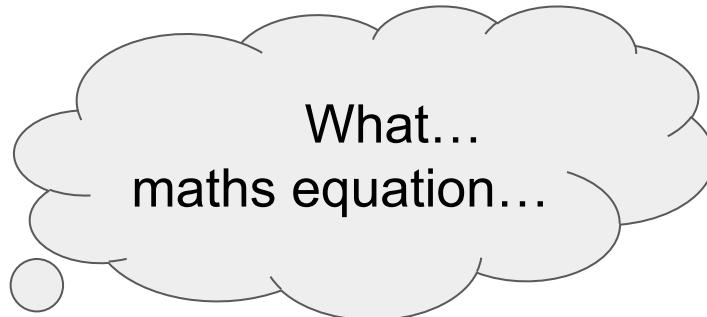
Reference: [here](#)

Classification: Precision & Recall

	Actual Positive(1)	Actual Negative (0)
Predicted Positive (1)	TP	FP
Predicted Negative (0)	FN	TN

Reference: [here](#)

Regression: RMSE & Rsquare



$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

\hat{y} – predicted value of y
 \bar{y} – mean value of y

Activity 7: Use Case - Discussion

By using a use case/case study in your profession/domain/discipline, discuss which evaluation metric is more suitable.



6) Deployment



Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

6) Cloud



E: Data Engineering

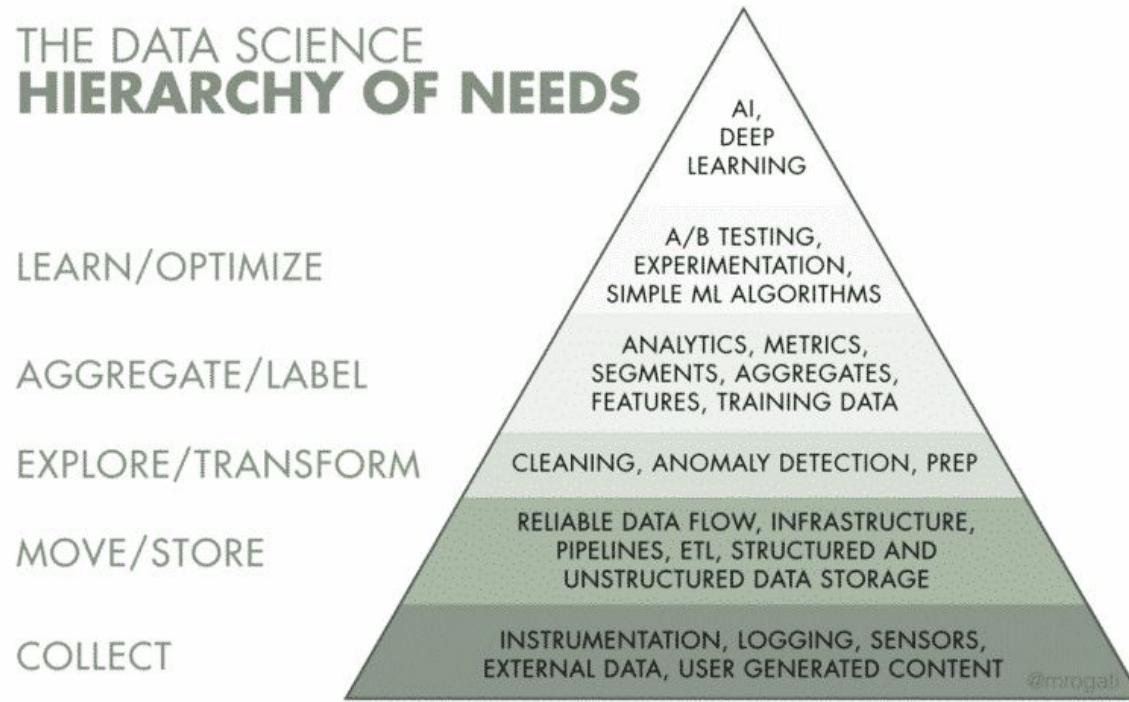


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Data Engineering

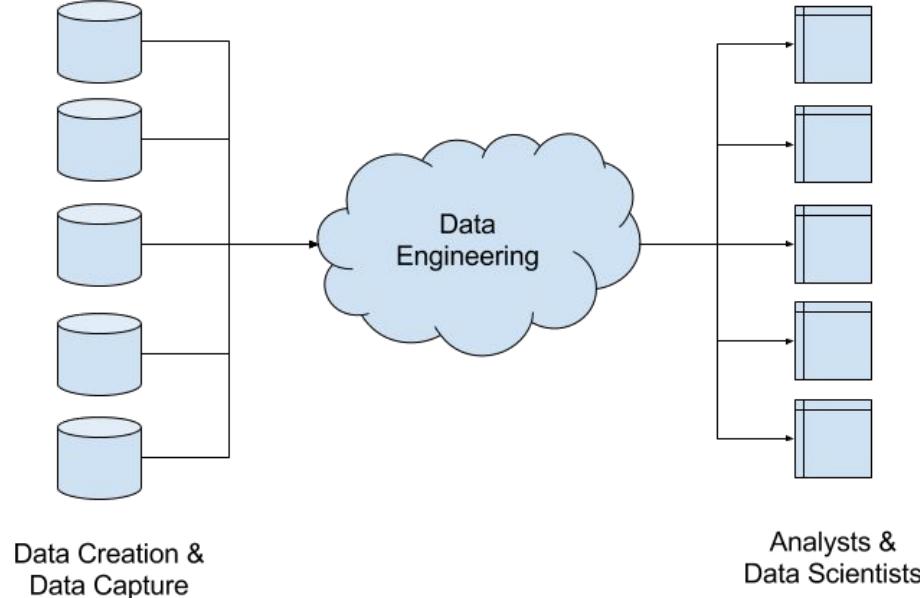


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Data Engineering

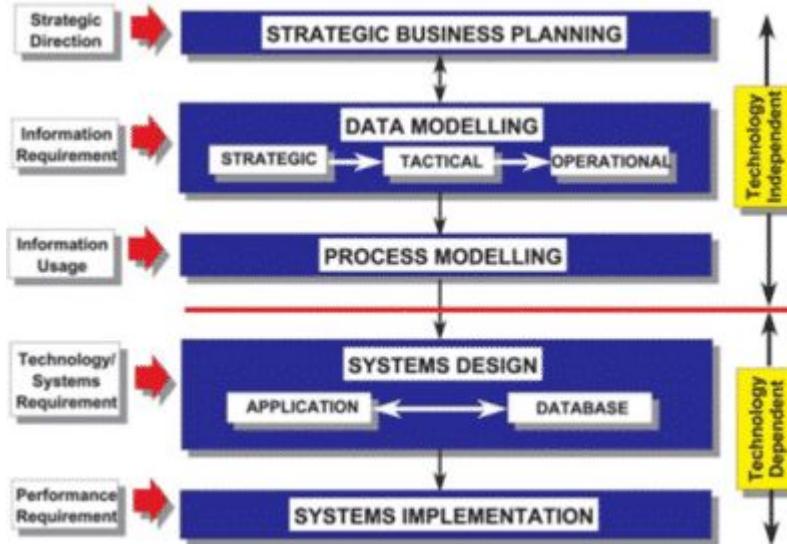


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Data Engineering

Technologies in Data Engineer Job Listings 2020

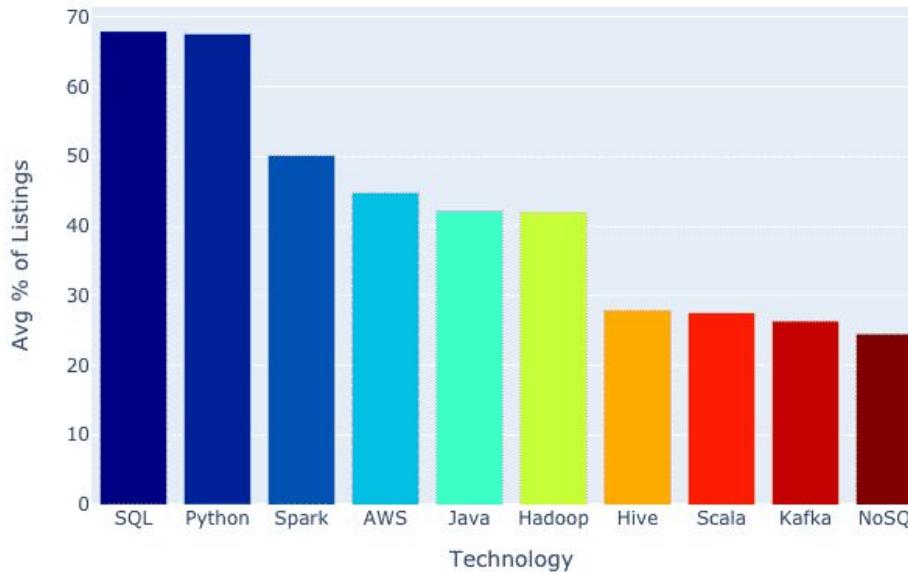


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Data Engineering

Technologies in Data Engineer Job Listings 2020

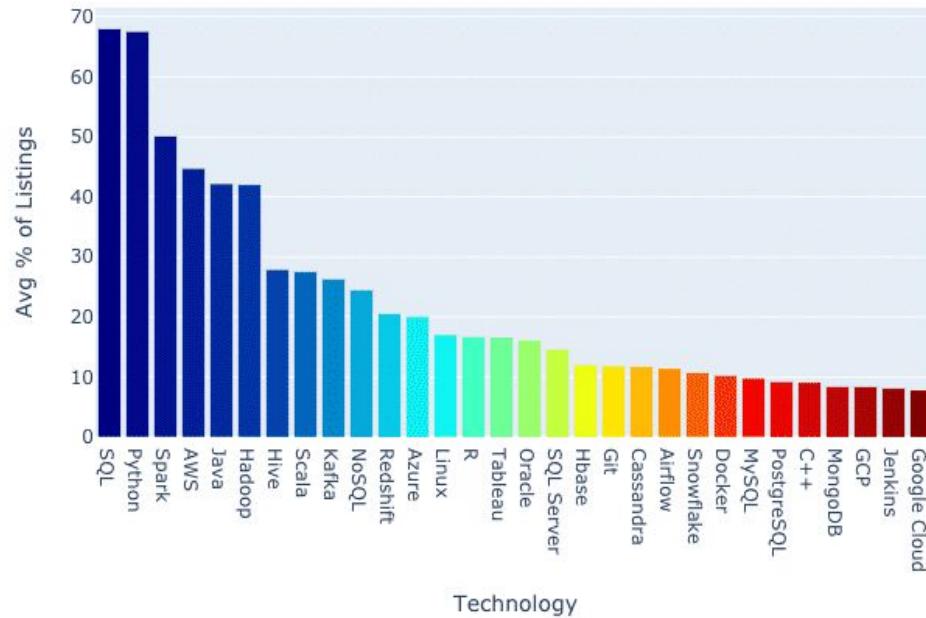


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

Data Engineering

Technologies in Data Engineer Job Listings 2020

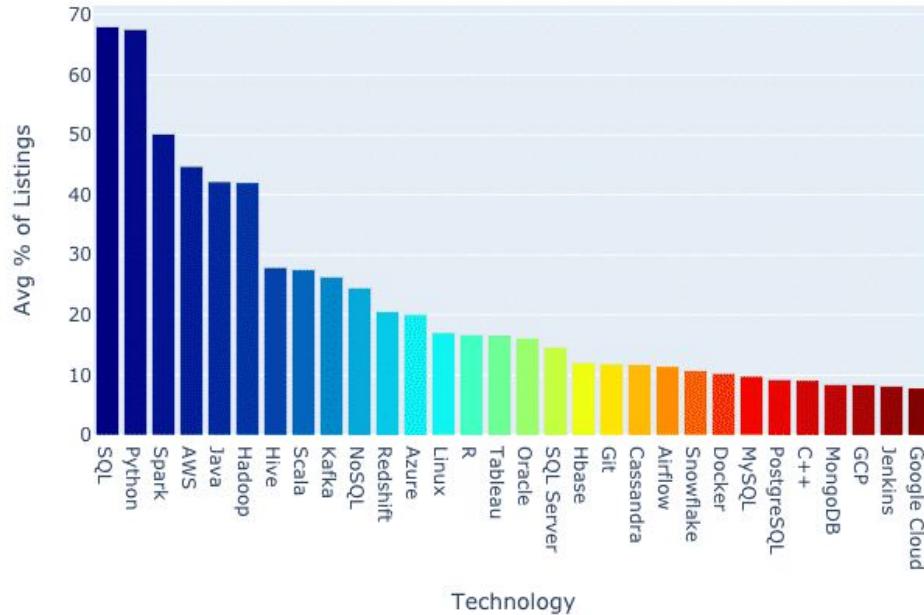


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

MLOps - Level 0

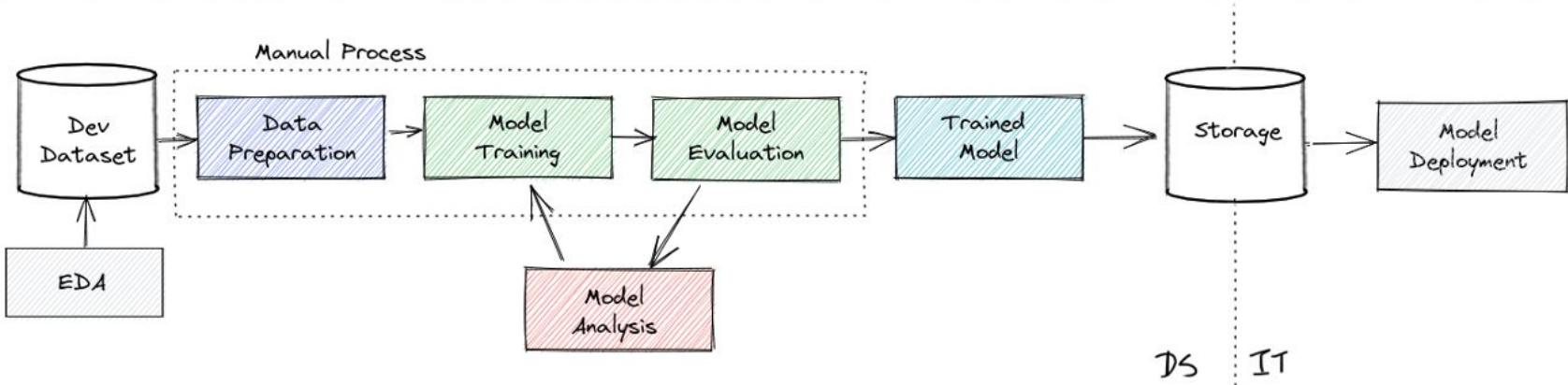


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

MLOps - Level 1

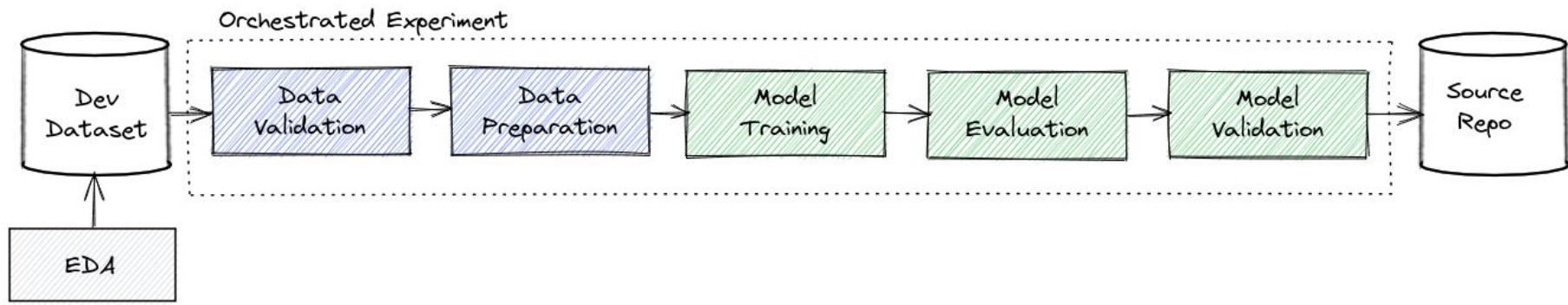


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

MLOps - Level 2

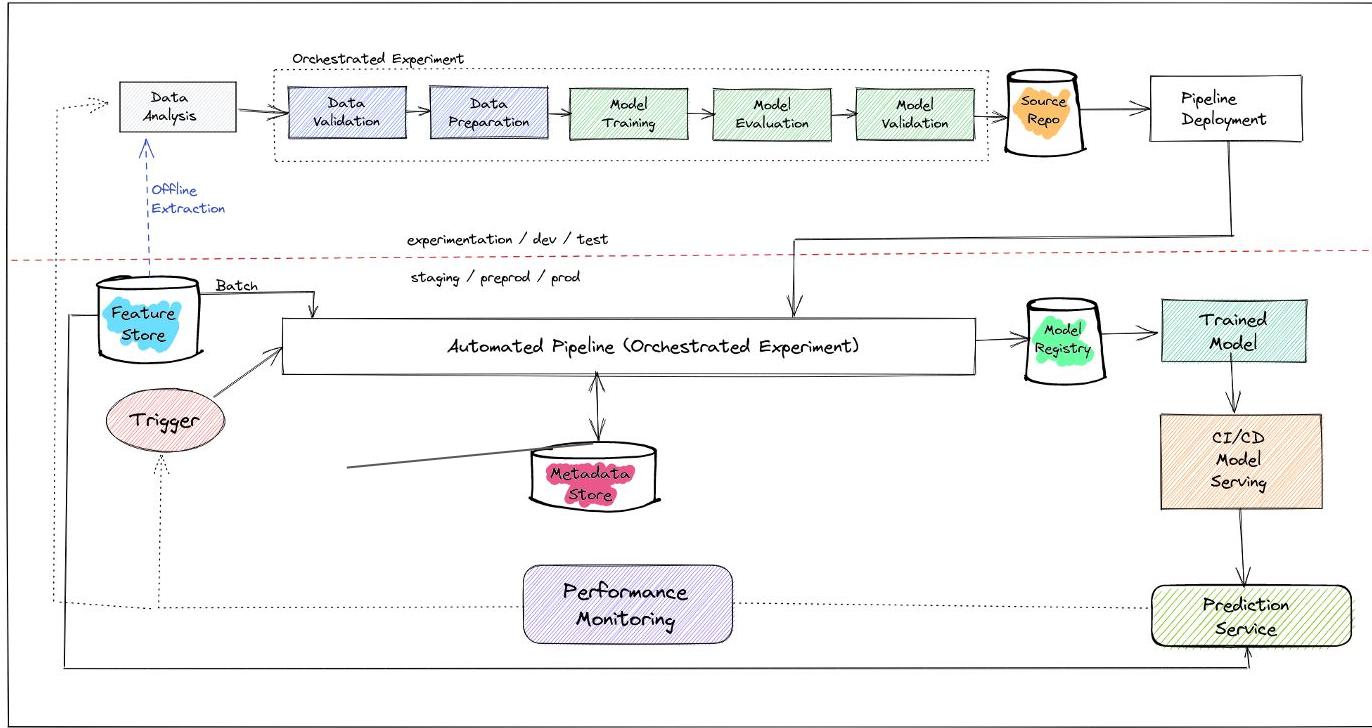


Image Credit: [HERE](#)

Cite this work if you are using it. Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richielynqph/creating-value-through-data-transformation>

F: Business Intelligence & Data Storytelling



I am listening
now...

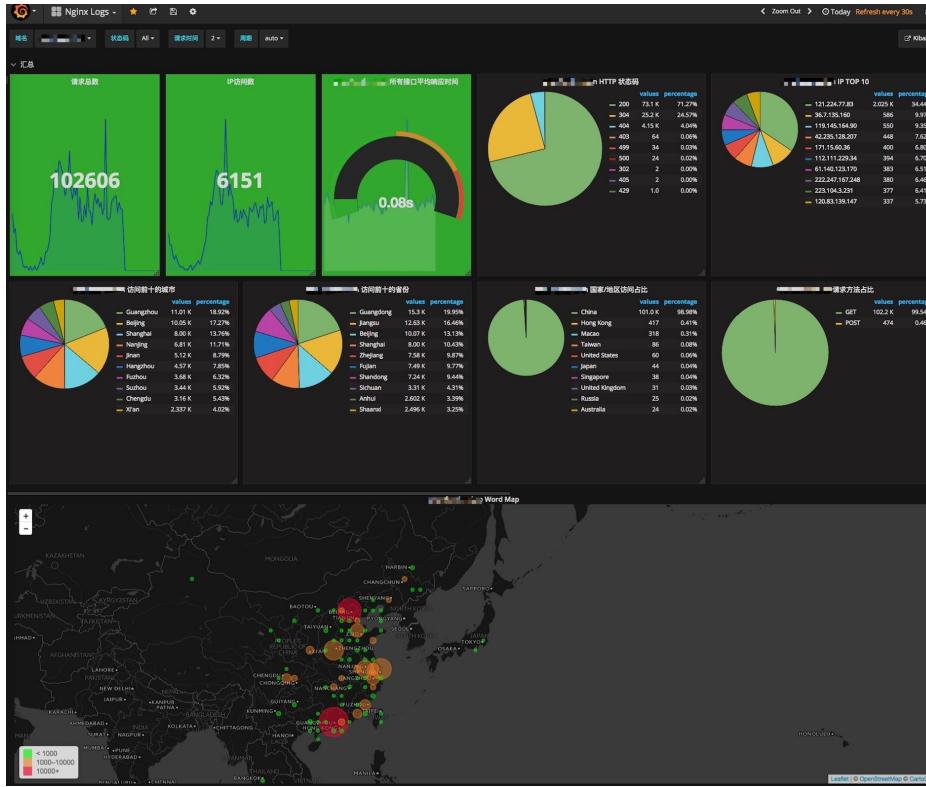


**“Let’s shrink Big Data into Small Data ...
and hope it magically becomes Great Data.”**

Image Credit: [HERE](#)

When you have mastered numbers, you will in fact no longer be reading numbers, any more than you read words when reading books. You will be reading meanings.

~ W.E.B. Du Bois



Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richeyuyongpoh/creating-value-through-data-transformation>

G: Turning Data into Actionable Insights

Well... So
many stuff
about data..
What is the
insight?





G2.com

Structured Data



Structured data is **quantitative** data in the form of numbers and values.

Unstructured Data



Unstructured data is **qualitative** data in the form of text files, audio files, video files.

Image Credit: [HERE](#)

Structured Data vs. Unstructured Data

Key Differences



Structured data is often stored in databases, while unstructured data is stored in data lakes.



Structured data is easy to access and work with, while unstructured data requires more work to process and understand.



Structured data is organized and exists in predefined formats, while unstructured data exists in different formats.



Structured data is quantitative, while unstructured data is qualitative data that cannot be processed using conventional tools.



Image Credit: [HERE](#)

Use cases for structured data

- **Customer relationship management (CRM):** CRM software runs structured data through analytical tools to create datasets that reveal customer behavior patterns and trends.
- **Online booking:** Hotel and ticket reservation data (e.g., dates, prices, destinations, etc.) fits the “rows and columns” format indicative of the pre-defined data model.
- **Accounting:** Accounting firms or departments use structured data to process and record financial transactions.

Source: [HERE](#)

Use cases for unstructured data

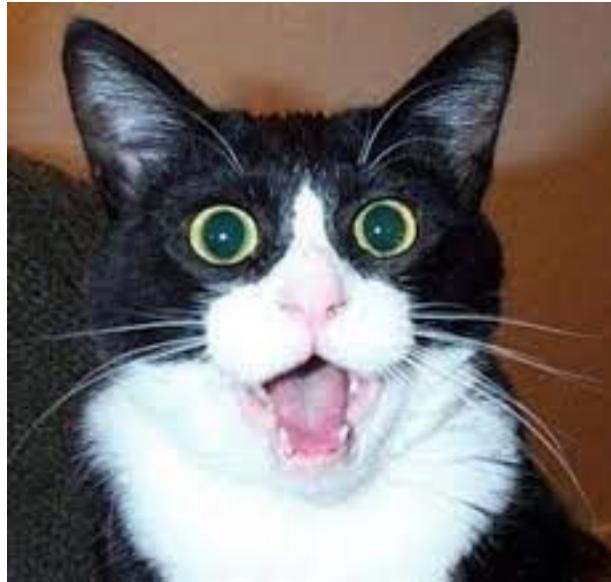
- **Data mining:** Enables businesses to use unstructured data to identify consumer behavior, product sentiment, and purchasing patterns to better accommodate their customer base.
- **Predictive data analytics:** Alert businesses of important activity ahead of time so they can properly plan and accordingly adjust to significant market shifts.
- **Chatbots:** Perform text analysis to route customer questions to the appropriate answer sources.

Source: [HERE](#)

And many
more.....



80% of global data will be unstructured by 2025



Source: [HERE](#)

Unstructured Data: Image

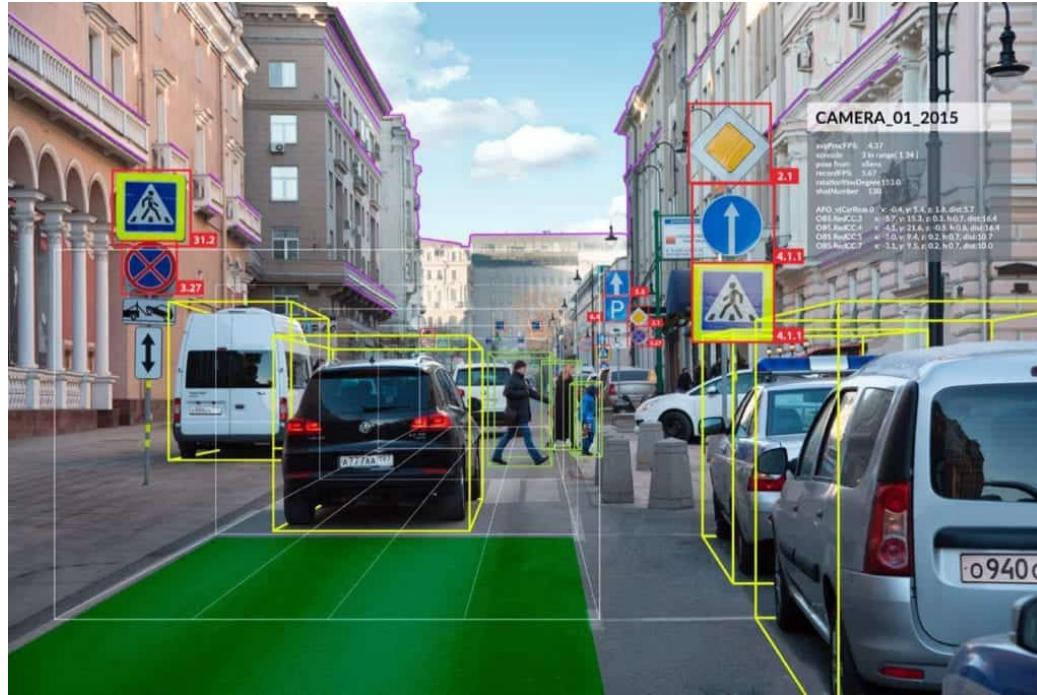
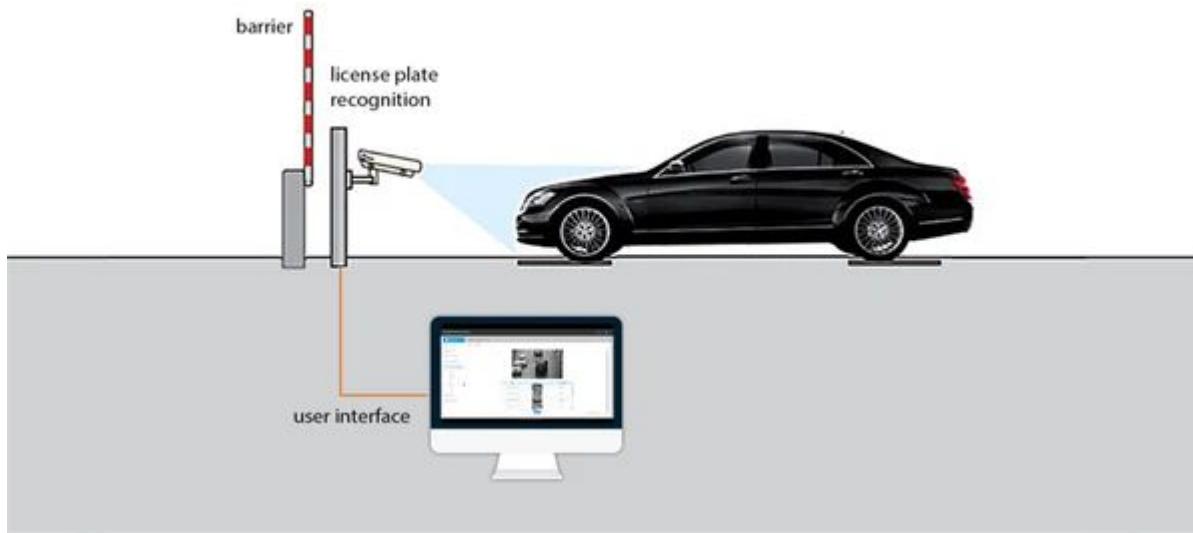


Image Credit: [HERE](#)

Image - Background

- **Image Processing & Analytics (Domain)**
- **CNN/YOLO and many more (Deep Learning)**
- **Image (Data)**

Example 1: Smart Parking System



Example 2: Manufacturing Defect Detection

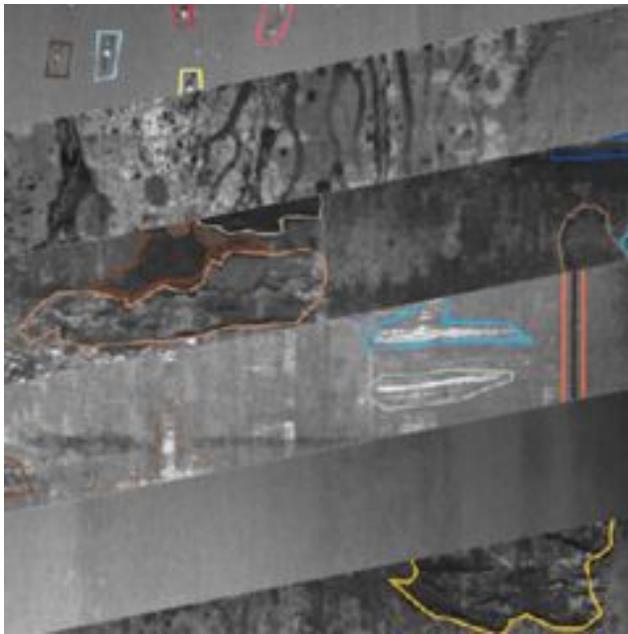


Image Credit: [HERE](#)

Example 3: Face Recognition Attendance System

Select an Option

Check in

Upload a facial image to check in

Drag and drop file here
Limit 200MB per file • PNG, JPG, TIFF, JPEG

Browse files

myphoto2.png 0.6MB X

aa is detected.

check in

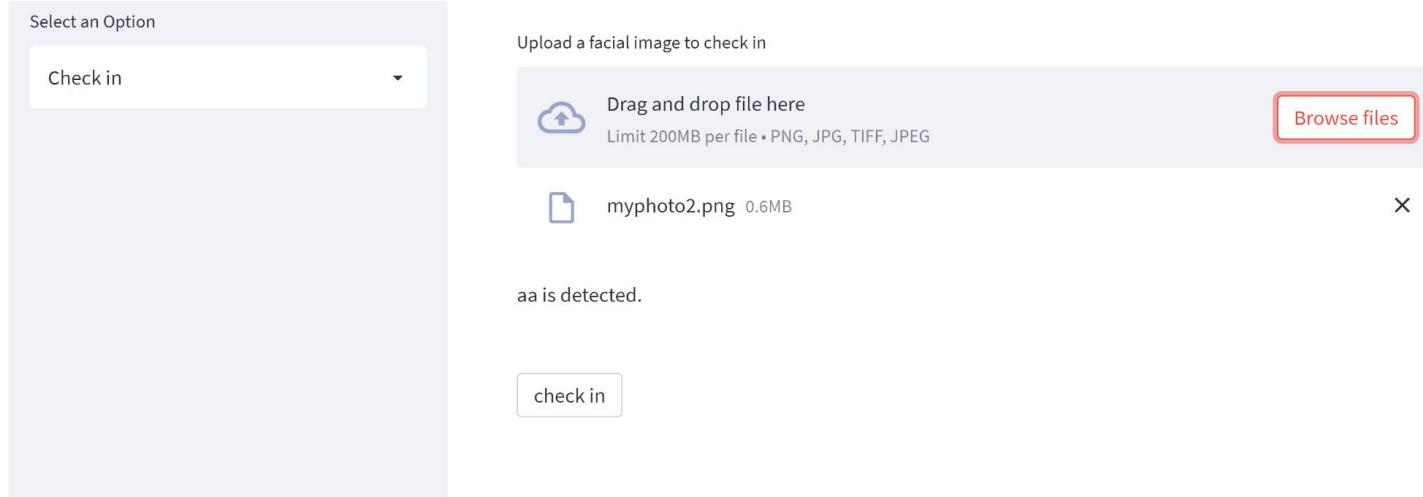


Image Credit: [HERE](#)

Demo: Face Recognition Attendance System

<https://www.richieyptutorialpage.com/demo-python-series/face-recognition-attendance-check-in-system>

Unstructured Data: Text

Chat With Me Now

readme first

You:



fourier transform

Dr Librarian:

In mathematics, a Fourier transform (FT) is a mathematical transform that decomposes functions depending on space or time into functions depending on spatial or temporal frequency, such as the expression of a musical chord in terms of the volumes and frequencies of its constituent notes. The term Fourier transform refers to both the frequency domain representation and the mathematical operation that associates the frequency domain representation to a function of space or time.

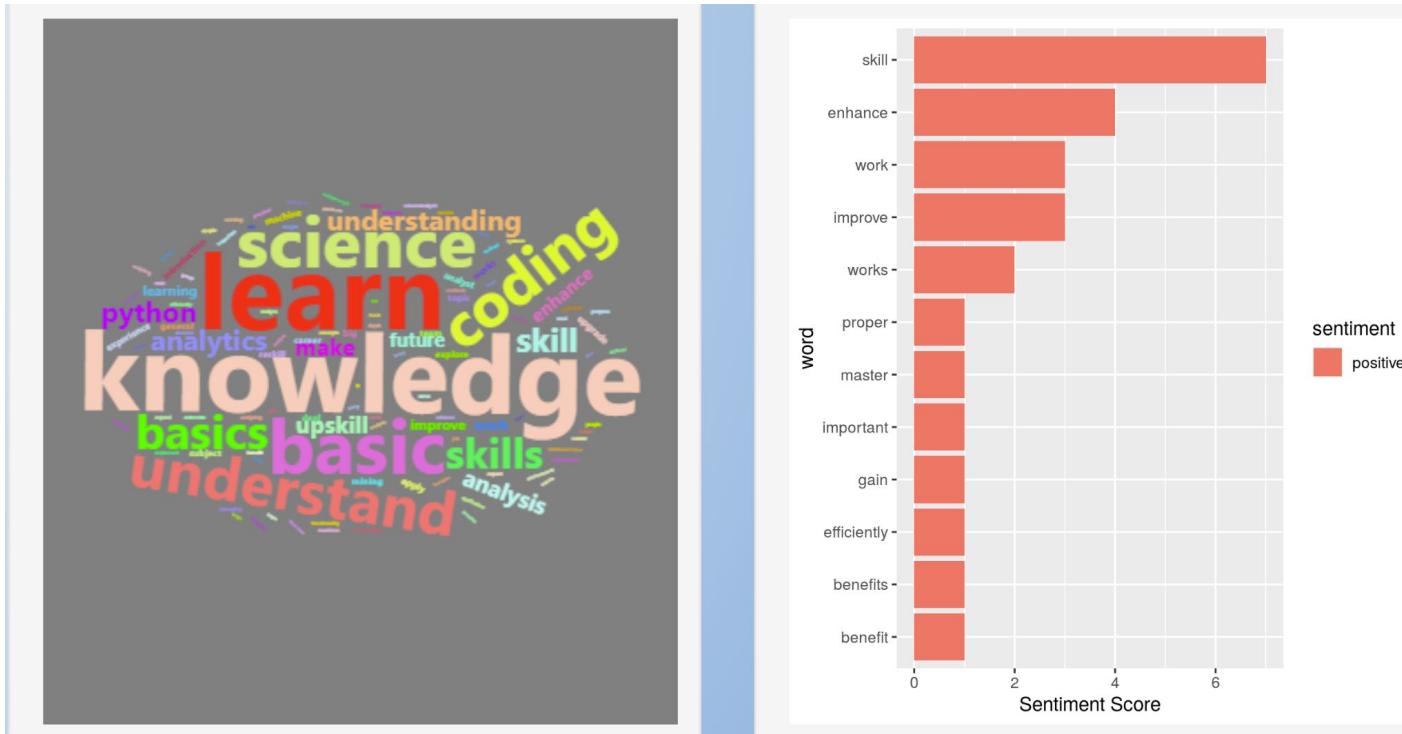
Text - Background

- **Text Analytics (Domain)**
- **Transformer/BERT/GPT3 etc (Deep Learning)**
- **Text (Data)**

Example 1: Chatbot



Example 2: Sentimental Analytics

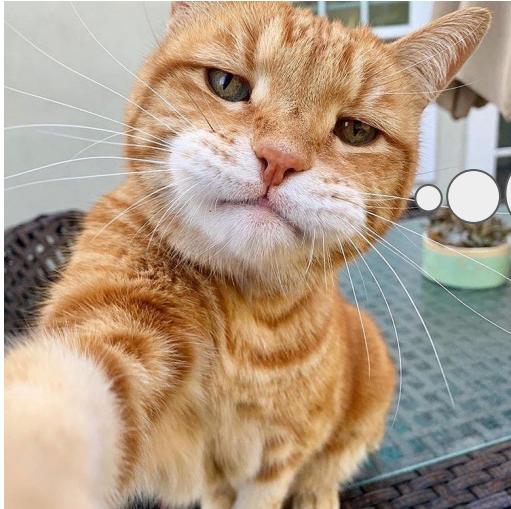


Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/ricchieyuyongpoh/creating-value-through-data-transformation>

Demo: Chatting with Dr. Librarian

<https://www.richieyyptutorialpage.com/demo-python-series/ai-chatbot-chat-with-me-now>

Bonus: Self-learning Activity - Project Deployment



Self-learning?
Nono... i prefer
selfie...

Step1: Create a github & streamlit cloud



Cloud

Gallery

Components

Community

Docs

Blog

STREAMLIT CLOUD

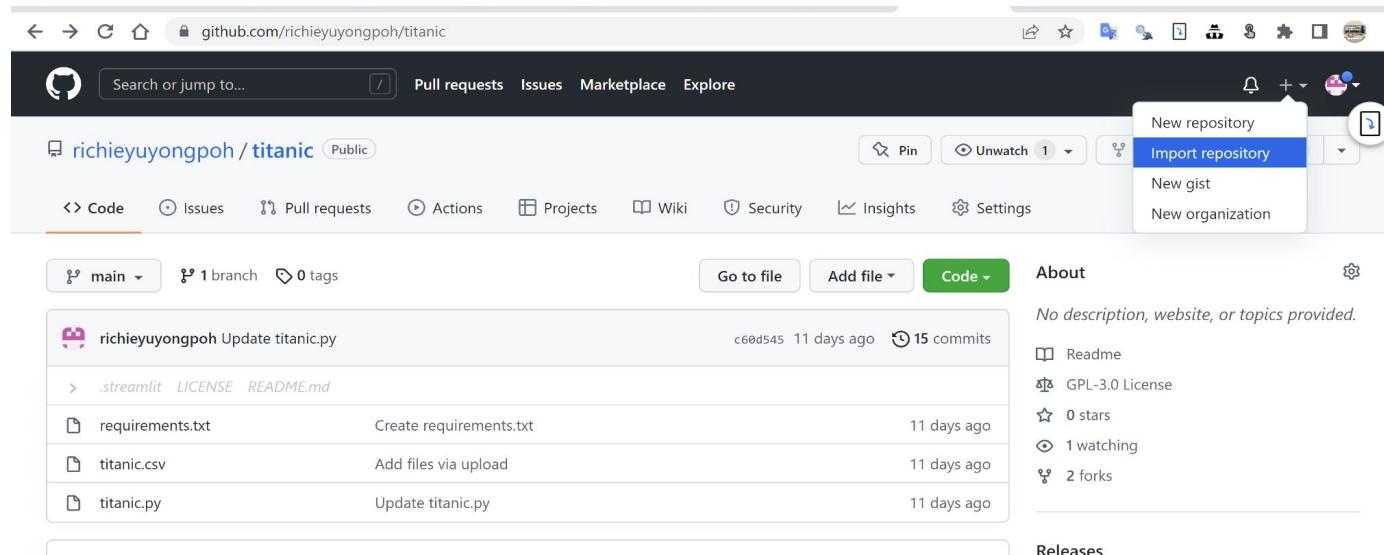
Empower your data team

Step2: Clone the TITANIC repository

The screenshot shows a GitHub repository page for the user 'richieyuyongpoh' with the repository name 'titanic'. The 'Code' tab is active. A context menu is open over the file 'titanic.py'. The 'Clone' option is highlighted and circled in red. The URL 'https://github.com/richieyuyongpoh/titanic.git' is displayed next to the 'Clone' button.

Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richieyuyongpoh/creating-value-through-data-transformation>

Step3: Import it to your new repository



Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
<https://github.com/richieyuyongpoh/creating-value-through-data-transformation>

Step3: Import it to your new repository

Import your project to GitHub

Import all the files, including the revision history, from another version control system.

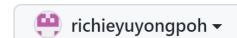
Your old repository's clone URL

`https://github.com/richieyuyongpoh/titanic.git`

Learn more about the types of [supported VCS](#).

Your new repository details

Owner *



Repository Name *

`titanic2` ✓

Privacy



Public

Anyone on the internet can see this repository. You choose who can commit.



Private

You choose who can see and commit to this repository.

Cancel

Begin import

Step4: Create a new app using streamlit cloud

Upgrade! Settings



richieyuyongpoh



New app 

Step5: Deploy the app

← Back

Deploy an app

Repository

richieyuyongpoh/titanic2

Paste GitHub URL

Branch

main

Main file path

titanic.py

Advanced settings...

Deploy!

Step6: Congratulations



Your app is in the oven

