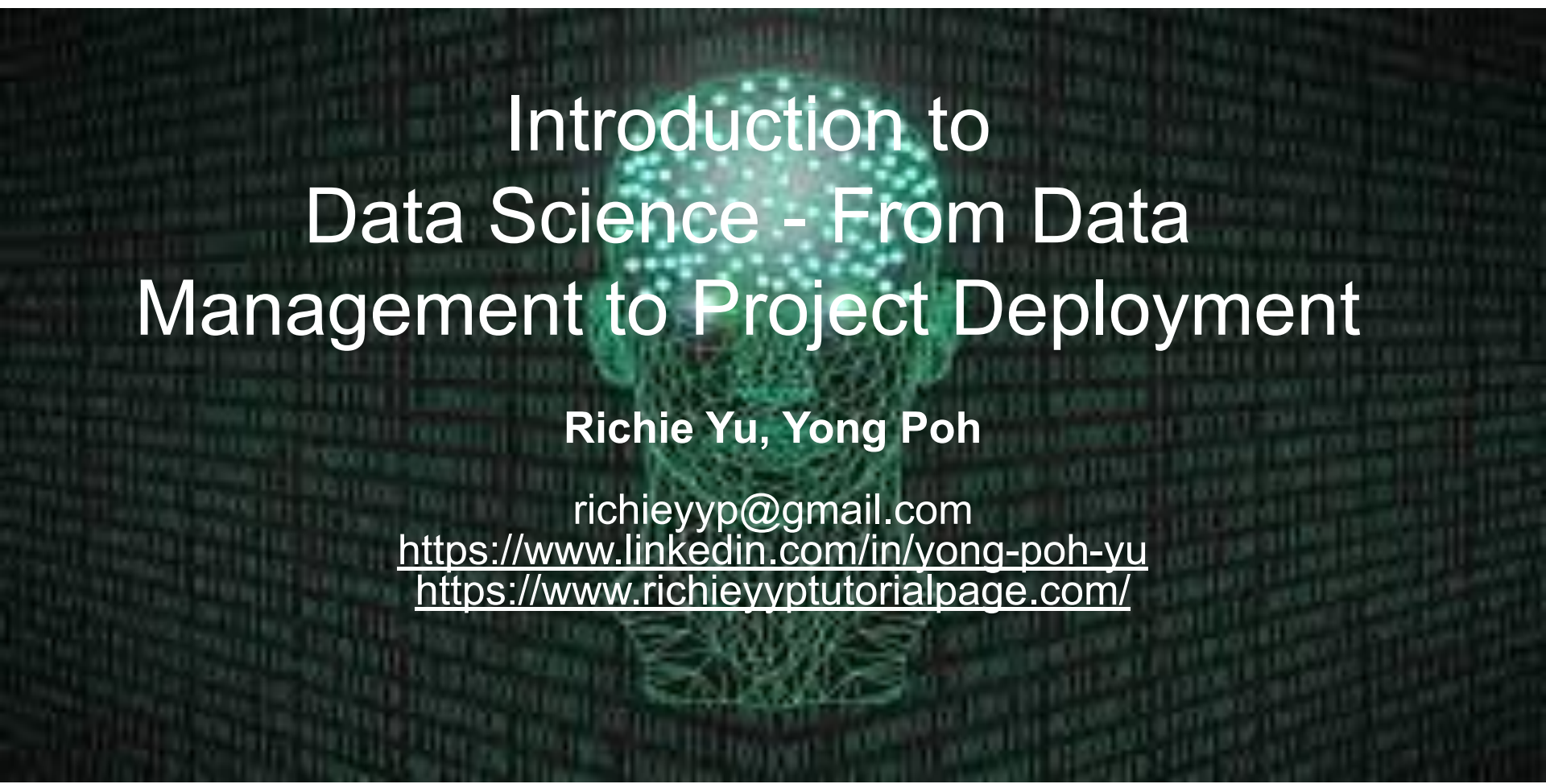


Introduction to Data Science - From Data Management to Project Deployment

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/

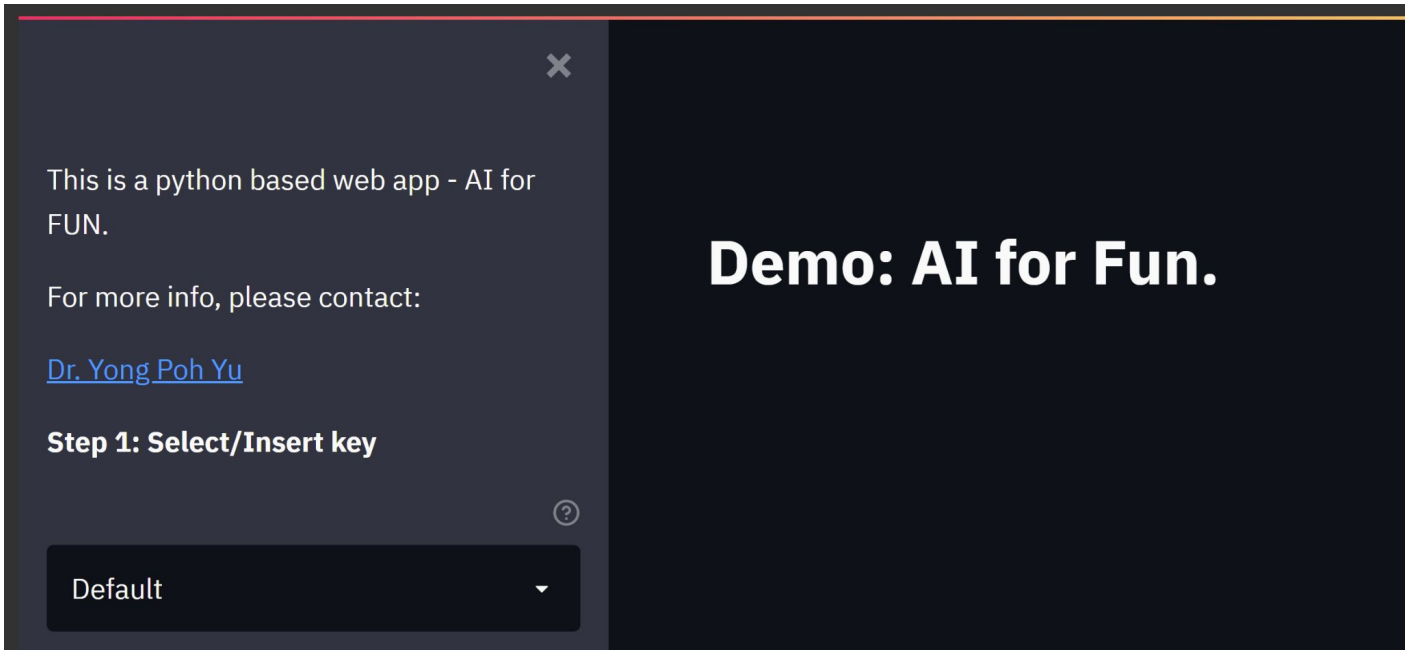
*** Open Source and Not-For-Profit Sharing / Demo**

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

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

A: Introduction to Big Data Analytics



Data science is a field for mathematical geeks

Learning a tool is the equivalent of learning data science

Data scientists will be replaced by artificial intelligence soon

What is/are …

Data Science?
Artificial Intelligence?
Machine Learning?

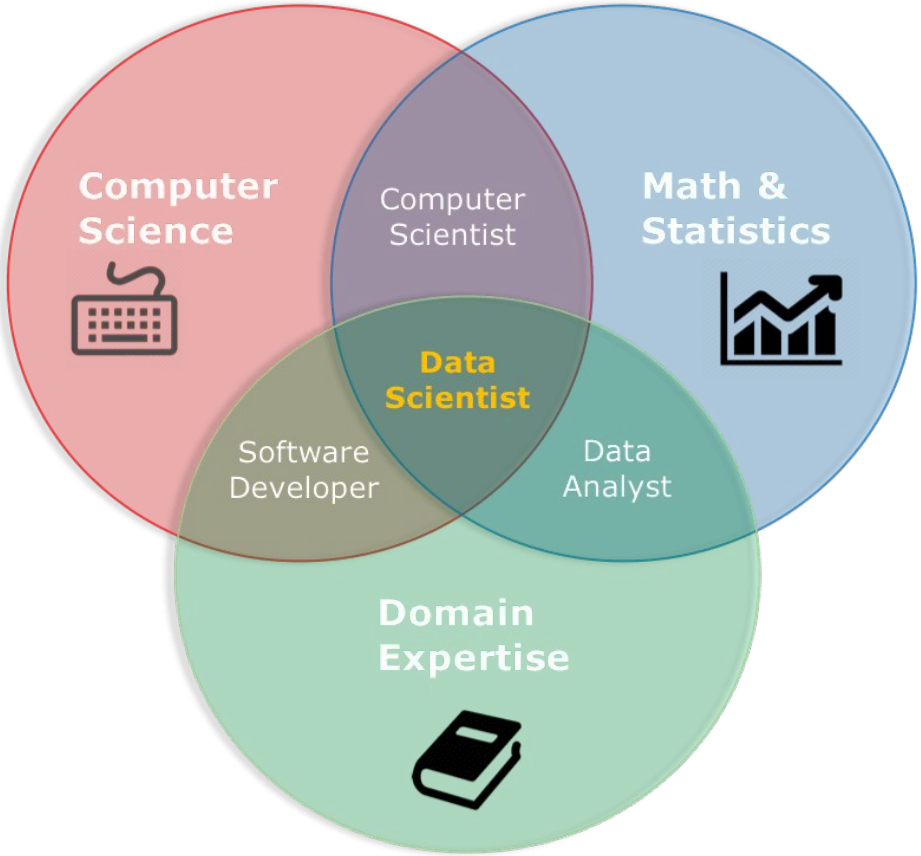
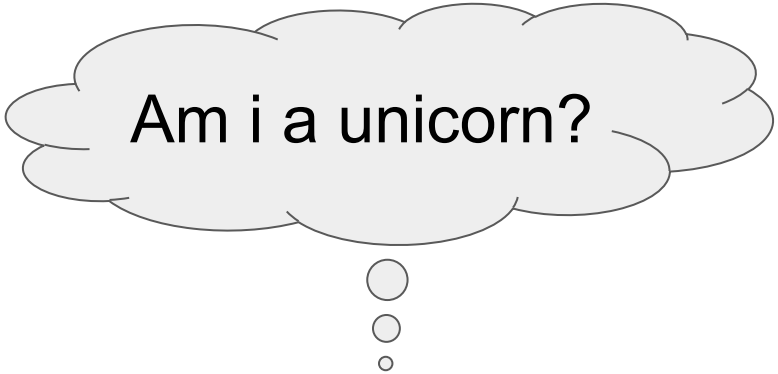


Image Credit: HERE



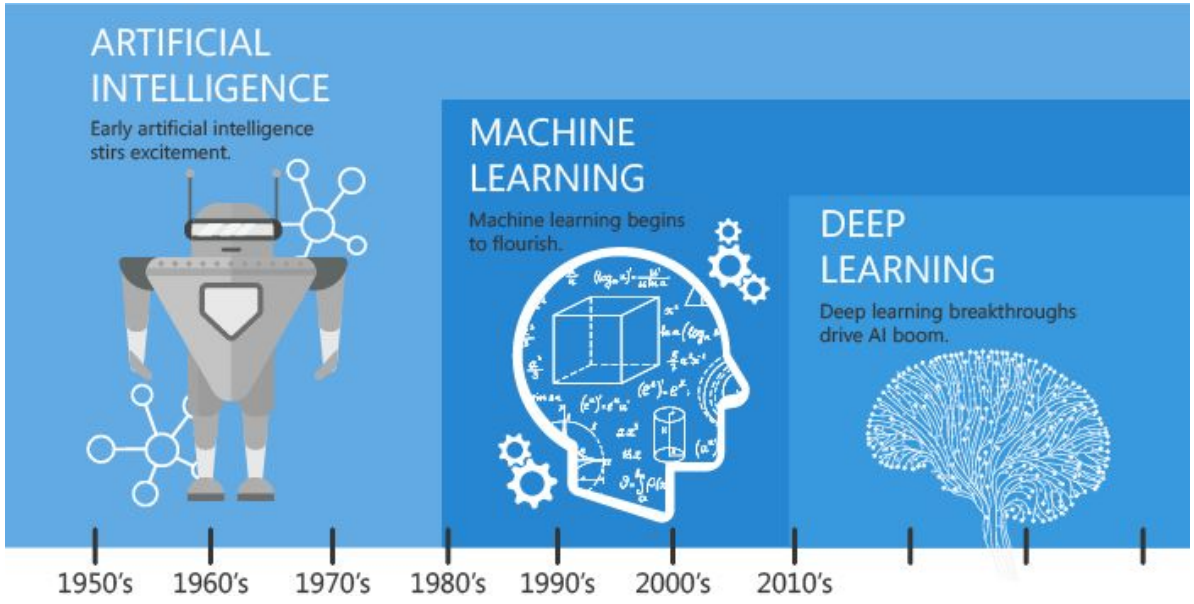


Image Credit: HERE

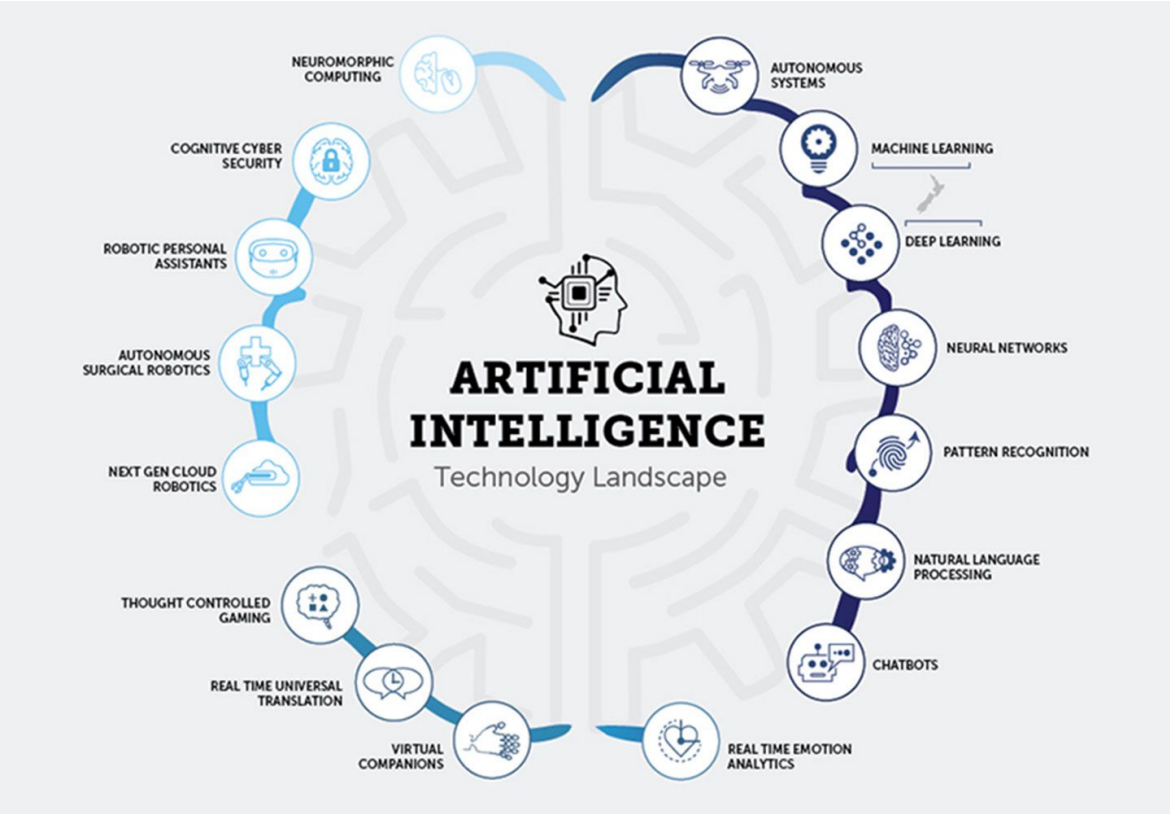


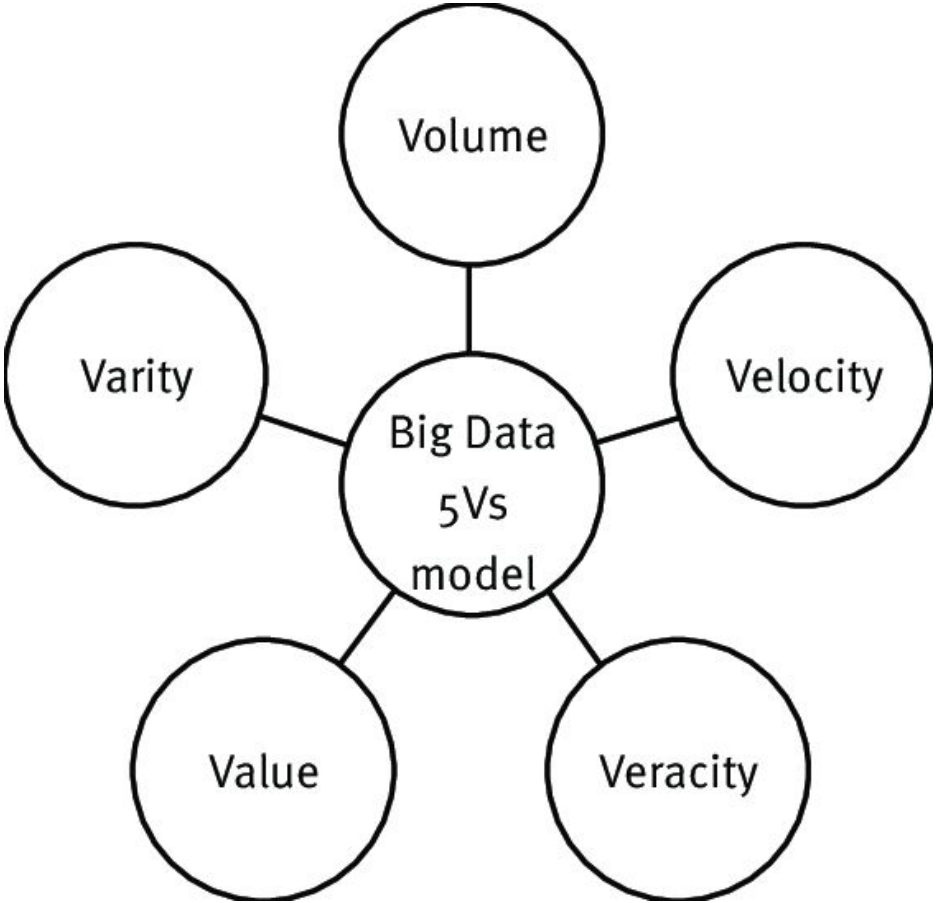
Image Credit: HERE

Data, AI, ML

Berpisah tiada

Then, what is …

Big Data?



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

- **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

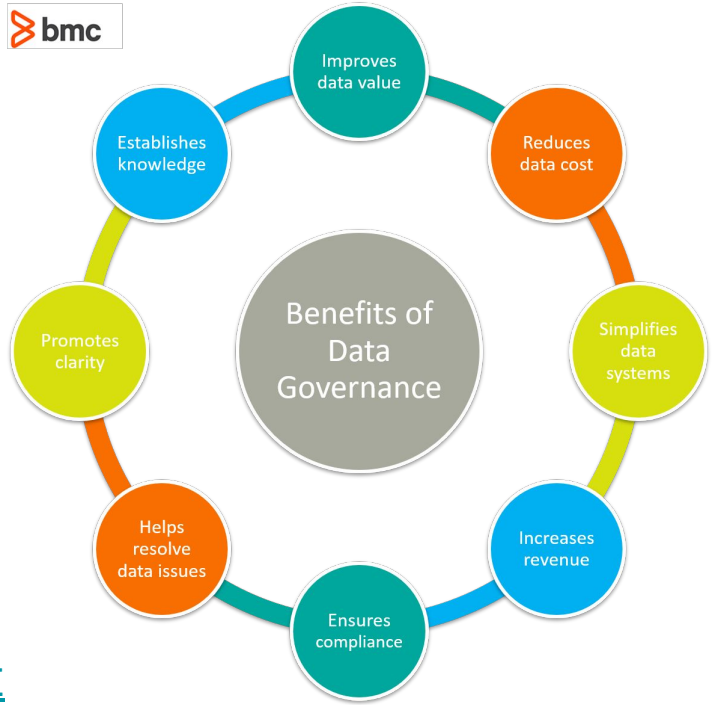


Image Credit: HERE

Data Management

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

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

Roadmap…
roadmap…

The road not taken

Developing the Right Data Strategy

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/richieyuyongpoh/creating-value-through-data-transformation

Developing the Right Data Strategy

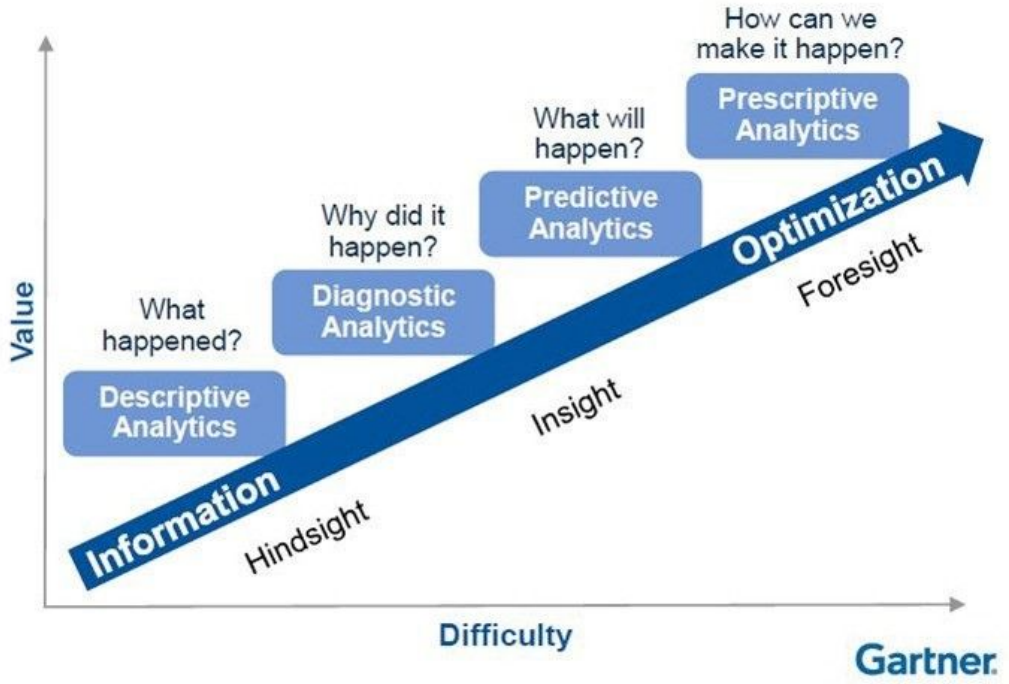


Image Credit: HERE

Developing the Right Data Strategy

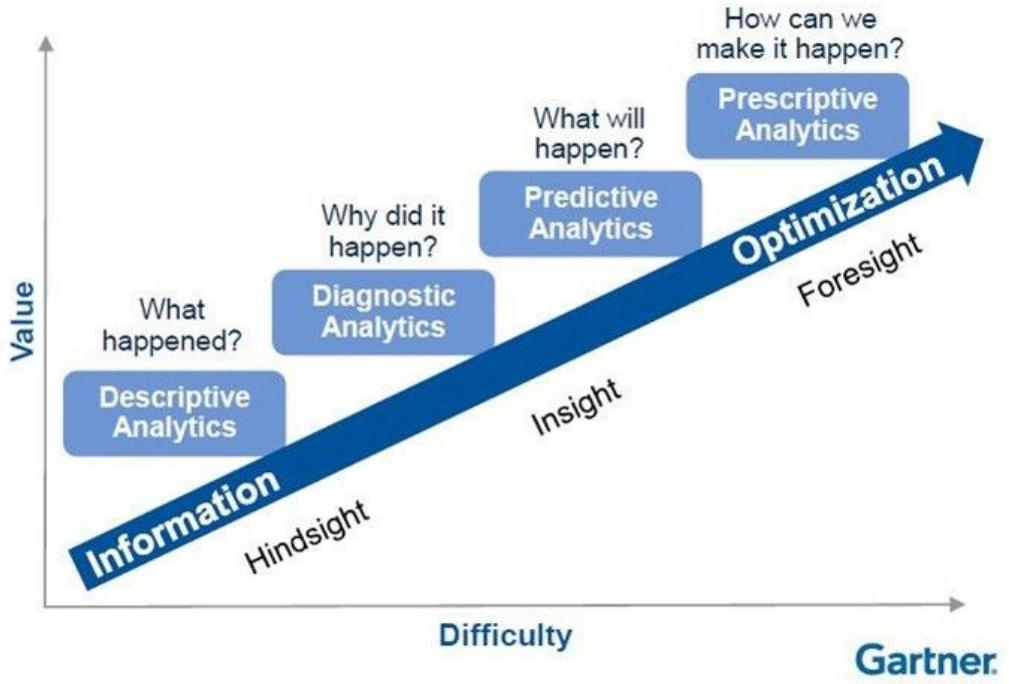


Image Credit: HERE

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)

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

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

Big Data Challenges & Solutions

- Scalability
- Proper BDA ecosystem

Big Data Challenges

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

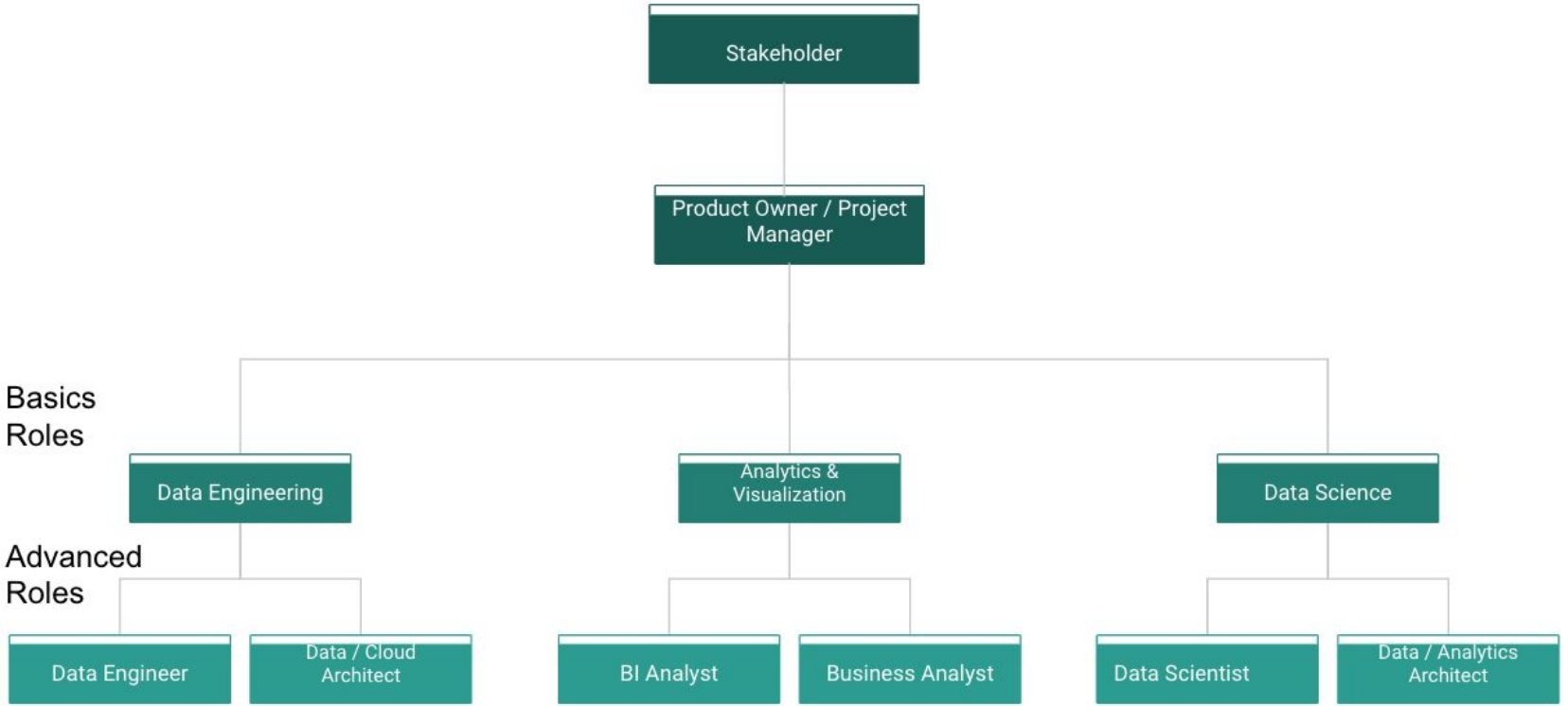


Image Credit: HERE

Data Engineer vs Data Analyst vs Data Scientist

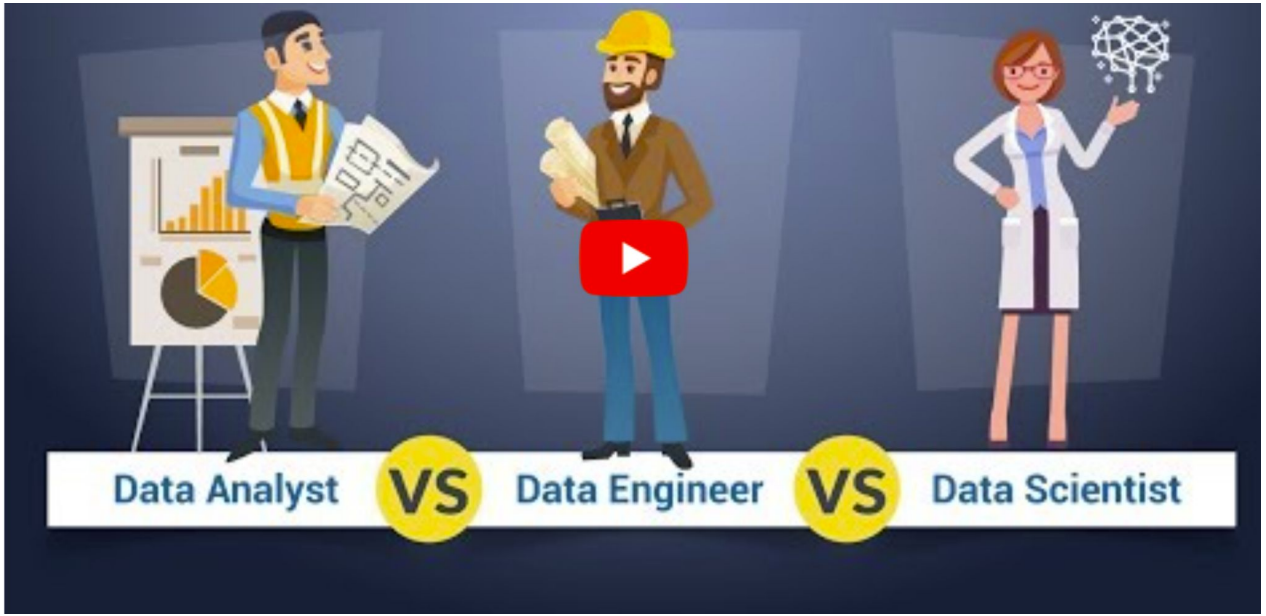


Image Credit: HERE

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

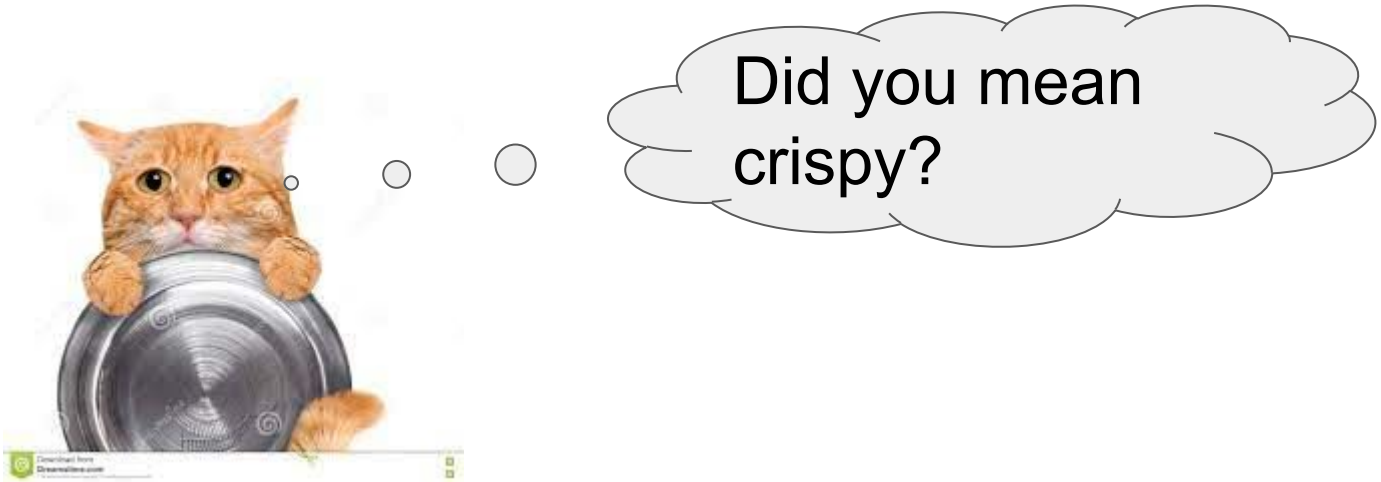
Data Analyst	Data Engineer	Data Scientist
Data Warehousing	Data Warehousing & ETL	Statistical & Analytical skills
Adobe & Google Analytics	Advanced programming knowledge	Data Mining
Programming knowledge	Hadoop-based Analytics	Machine Learning & Deep learning principles
Scripting & Statistical skills	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

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

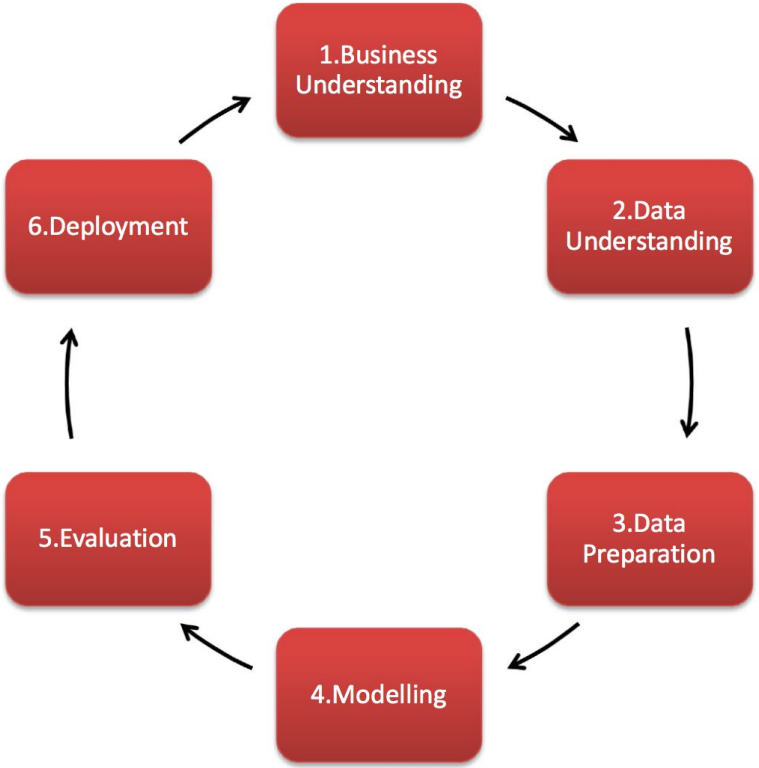
Image Credit: HERE

D: Data Science - CRISP-DM Approach

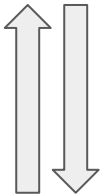


1. **Measure the right things**
2. Ask the right questions to stakeholders
3. Use segmentation to drive action
4. Use clear visualizations to convey your message
5. Discover the context of your data set
6. **Build a solid optimization plan**
7. Construct a great hypothesis
8. Integrate data sources
9.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

Source: HERE



Business



Data

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/richieyuyongpoh/creating-value-through-data-transformation

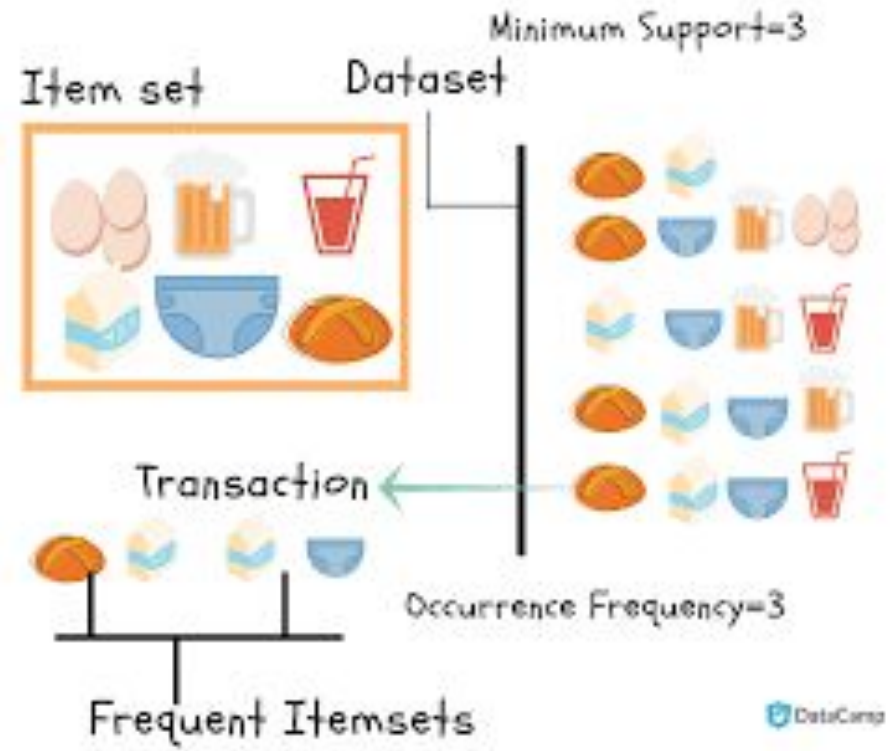
1) Business Understanding

What are the leading factors?

Automate a data-drive solution

Use Case: Stock Price Forecasting

Use Case: Market Basket Analysis

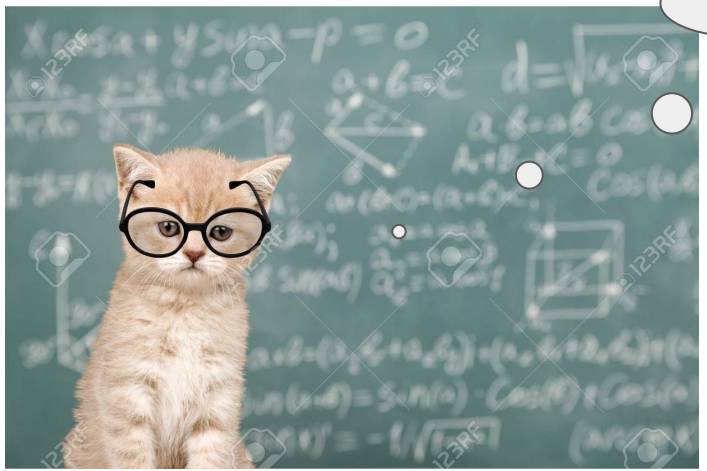


Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
https://github.com/richieyuyongpoh/creating-value-through-data-transformation

Use Case : Customer Segmentation

https://www.richieyyptutorialpage.com/demo-r-series/customer-segmention-using-k-means

<Did you know?>



Data Storytelling is important

Use Case: Household Income and Food Expenditure

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

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

<Did you know?>

When data is extraordinary ...

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/richieyuyongpoh/creating-value-through-data-transformation

What are the leading factors?

Can you give a list of potential churners?

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					

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

2) Data Understanding

What are the X?

What is the y?

Missing value?

Outliers?

3) Data Preparation

Missing value?

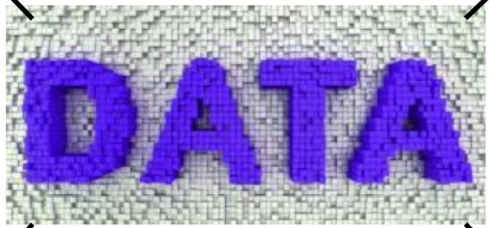
Outliers?

New Column/Variable/Feature?

I heard that ...

Garbage in, garbage out (GIGO)

Key Ingredient



Messy

The more the better

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

I heard that ...

DATA SCIENCE = DATA CLEANING

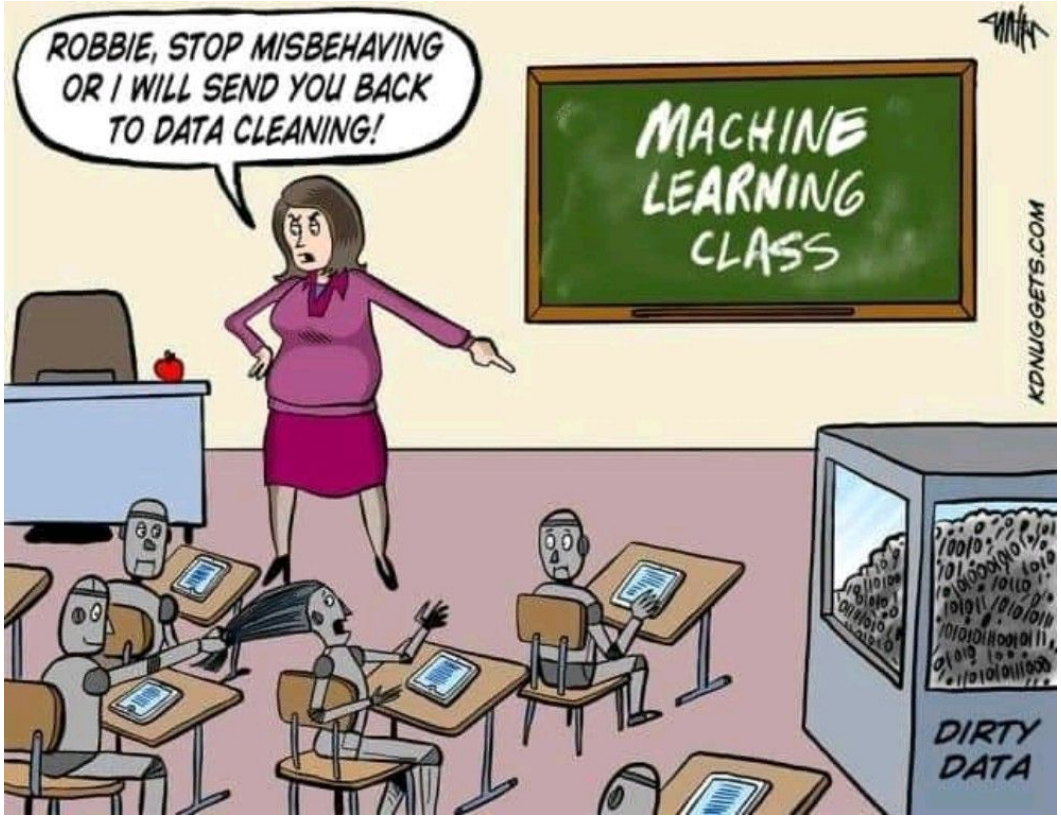


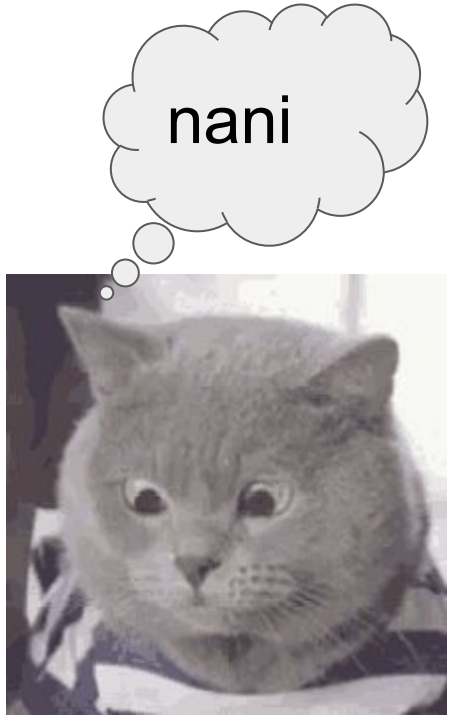
Image Credit: HERE

Data, AI, ML

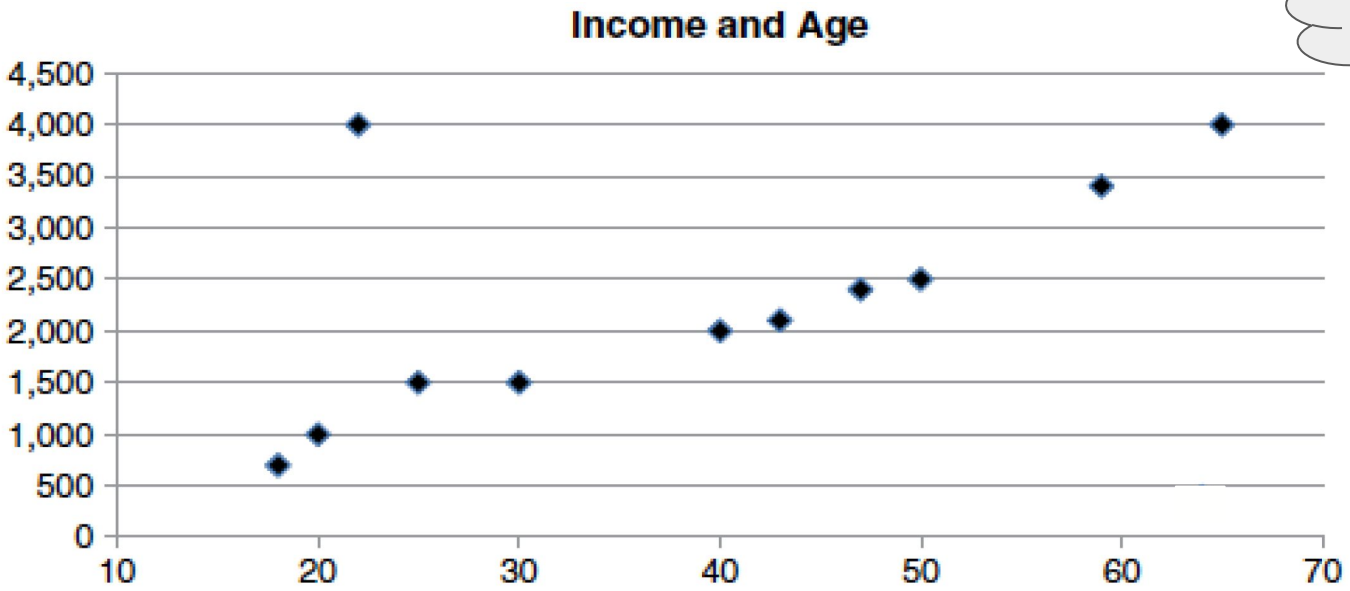
Berpisah tiada

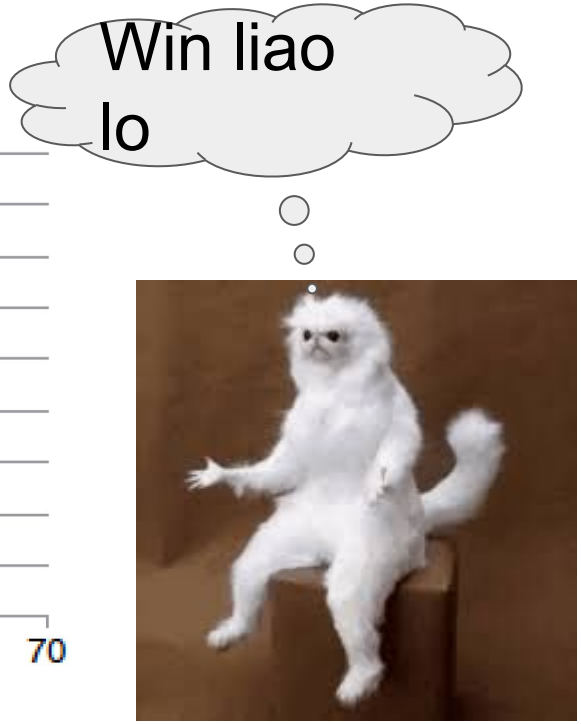
I heard that data is incomplete...

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...





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

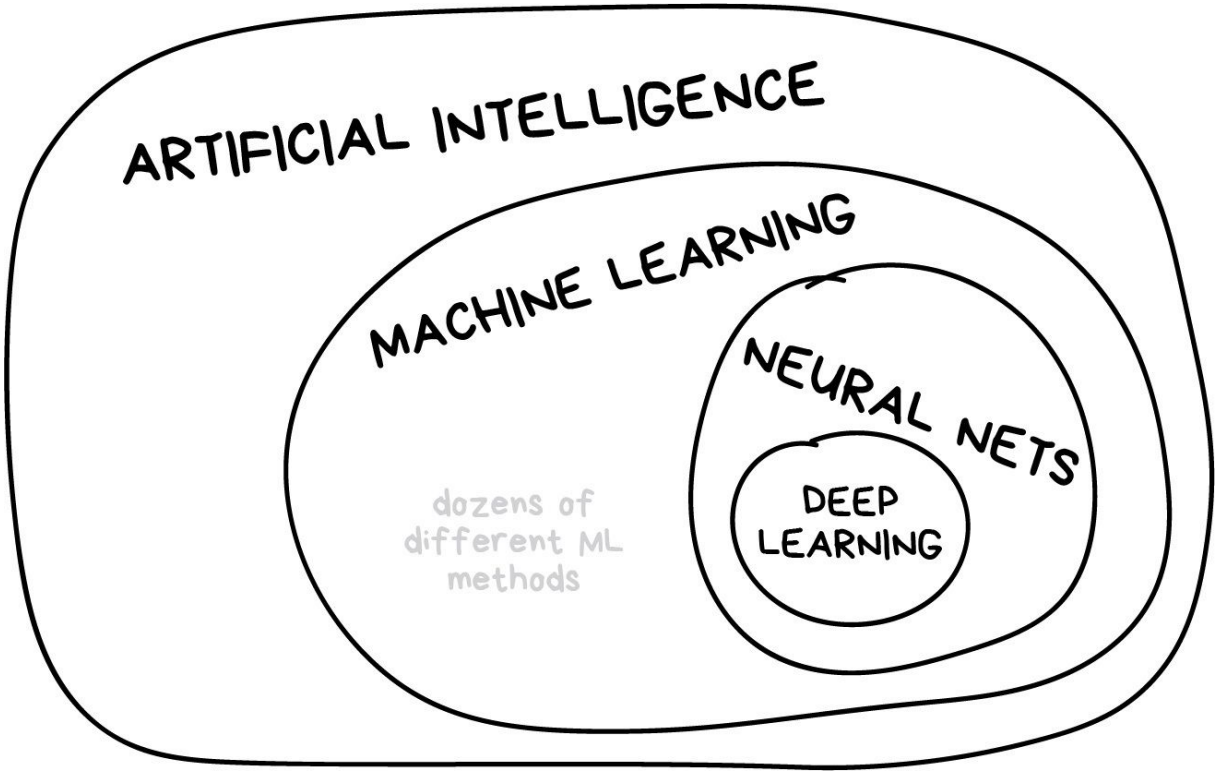
I heard that data quality is important...

Cite this wc
https://githu

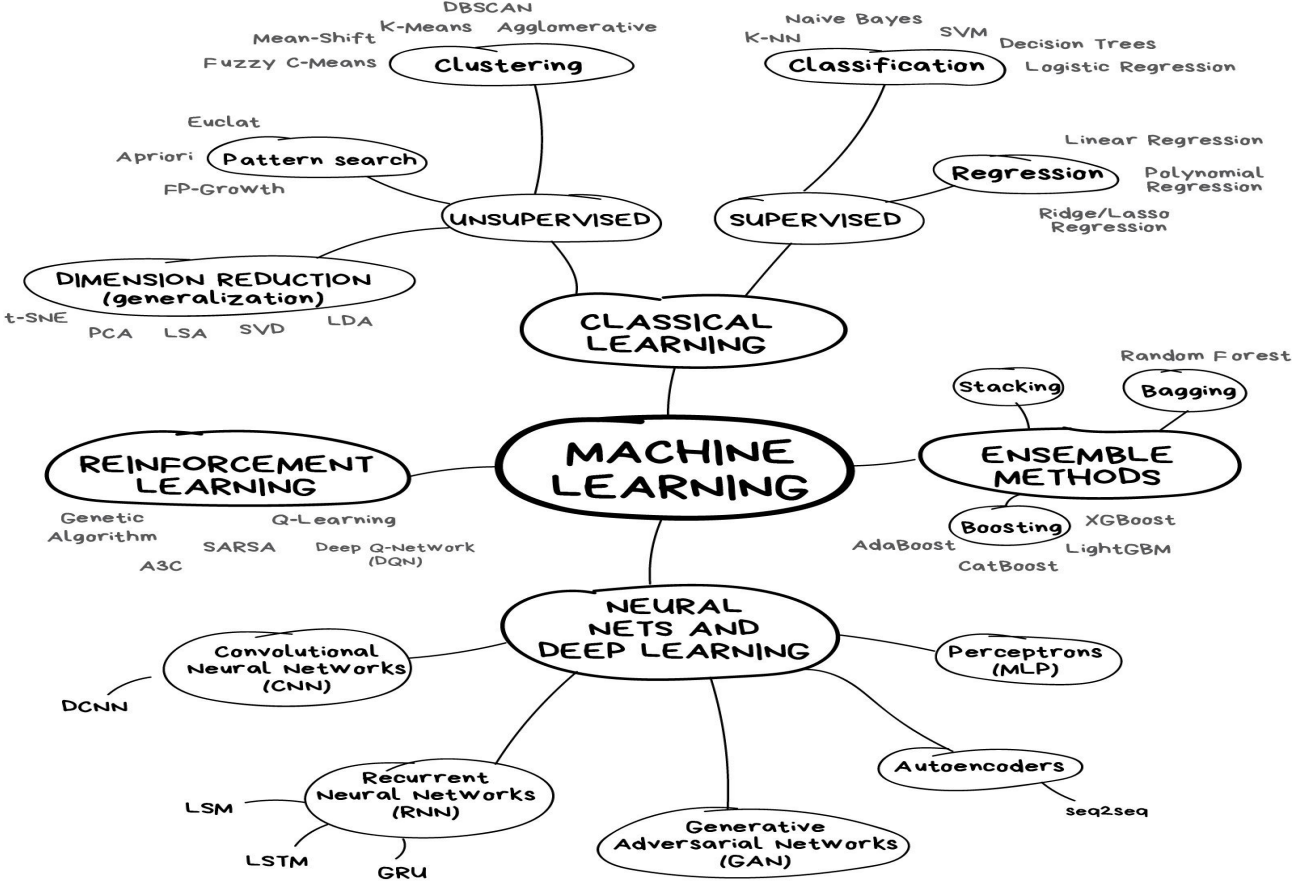
s - From Data Management to Project Deployment..

4) Modelling

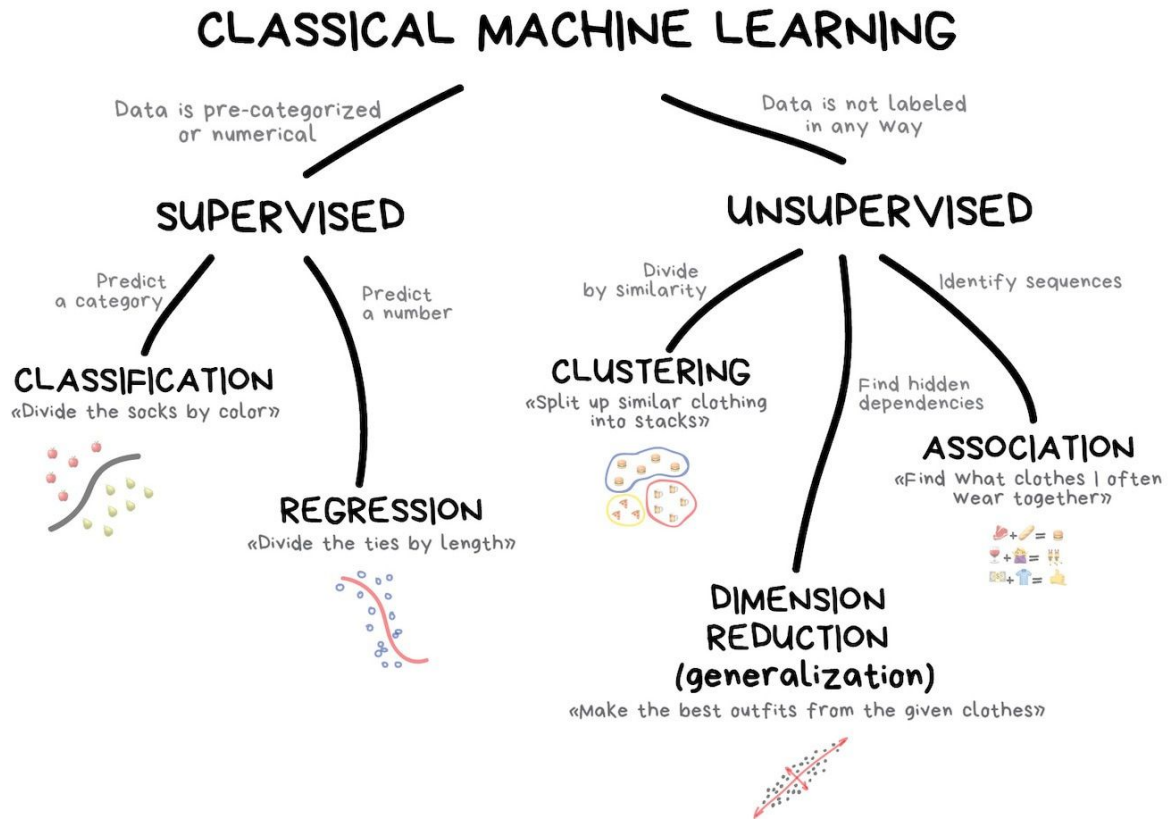
Are you looking for a model?



Reference: here

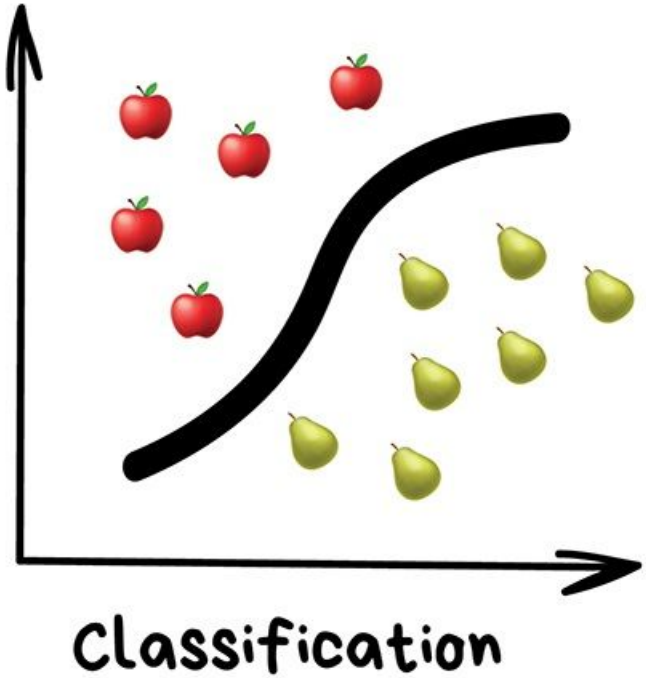


Reference: here

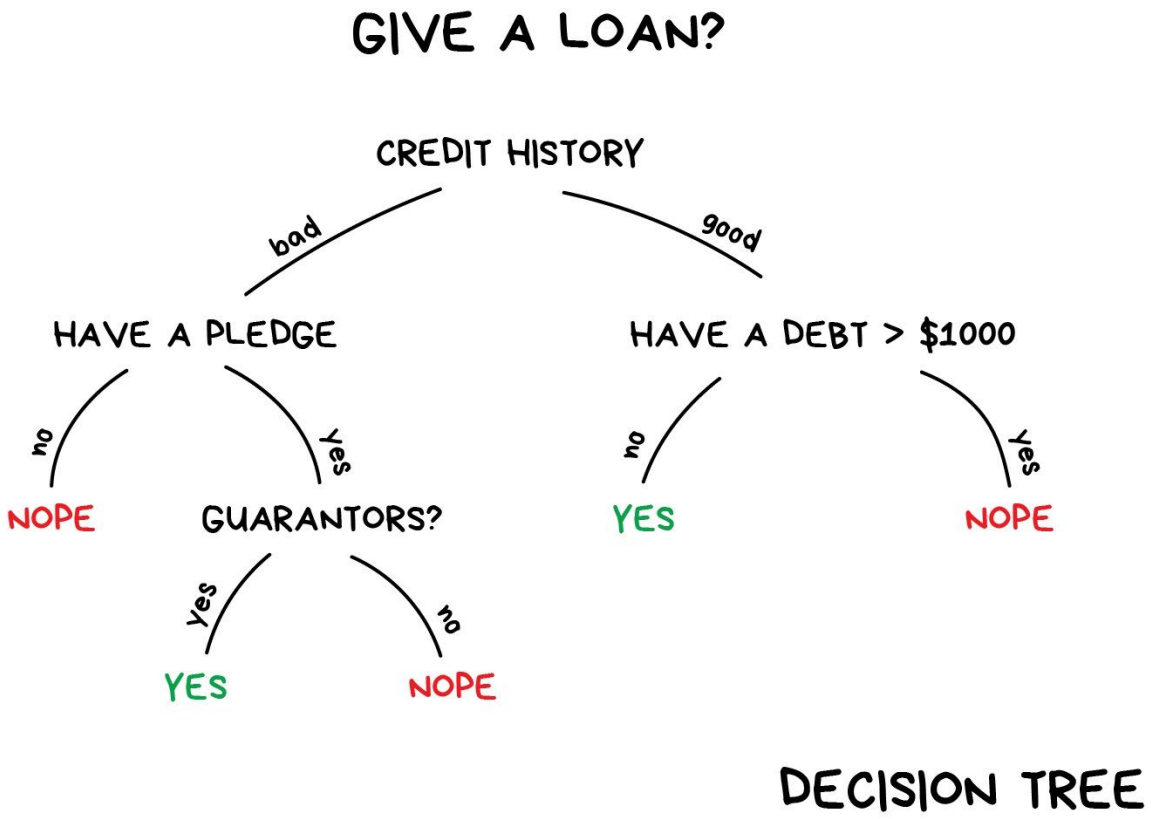


Reference: here

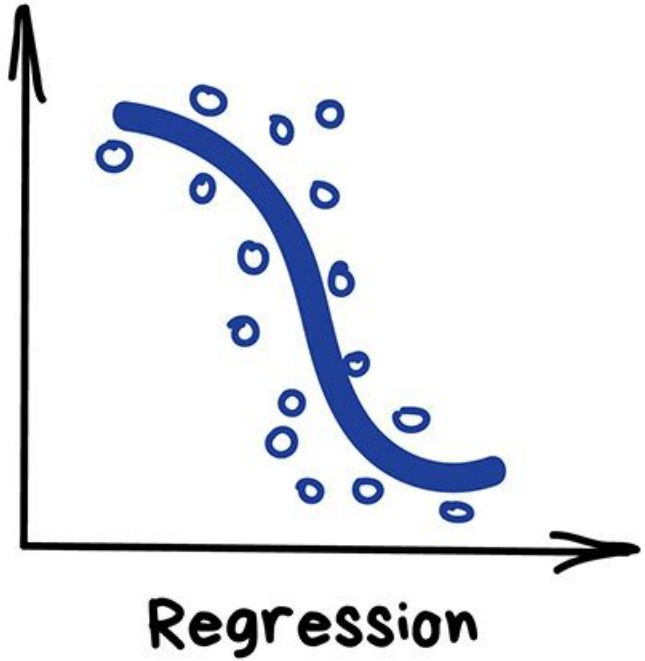
Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment.. https://github.com/richieyuyongpoh/creating-value-through-data-transformation



Reference: here

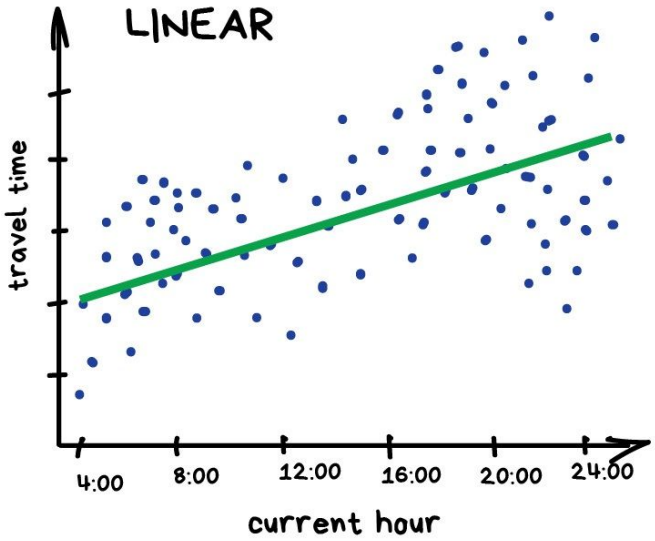


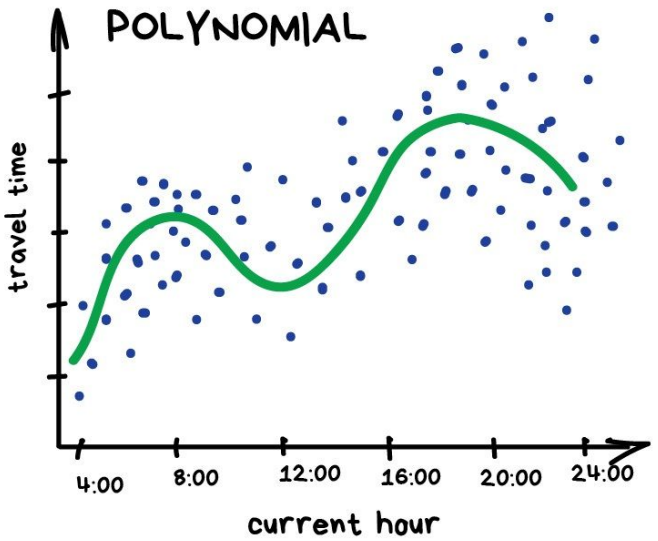
Reference: here



Reference: here

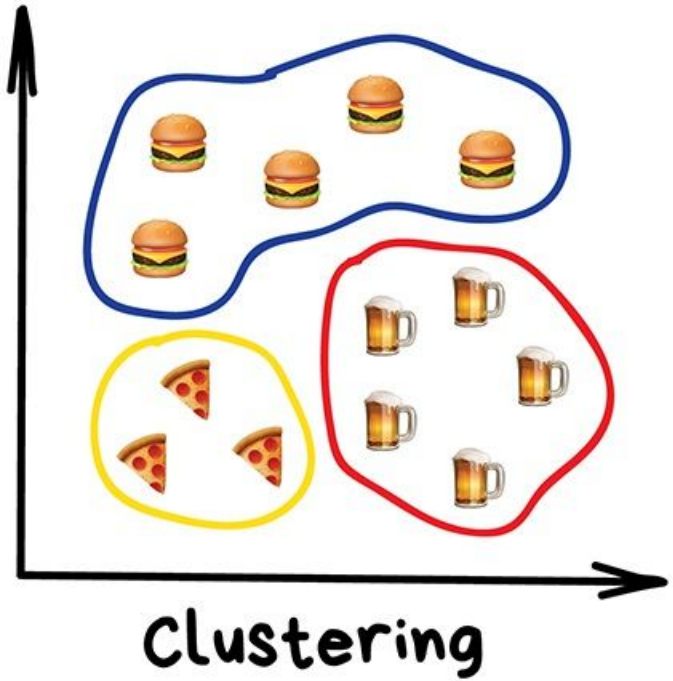
PREDICT TRAFFIC JAMS





REGRESSION

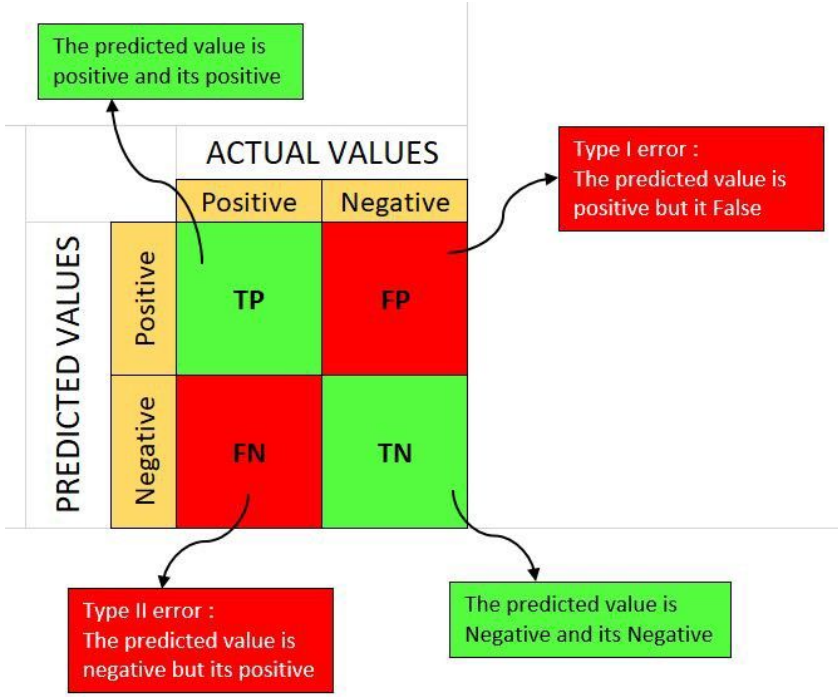
Reference: here



Reference: here

5) Evaluation

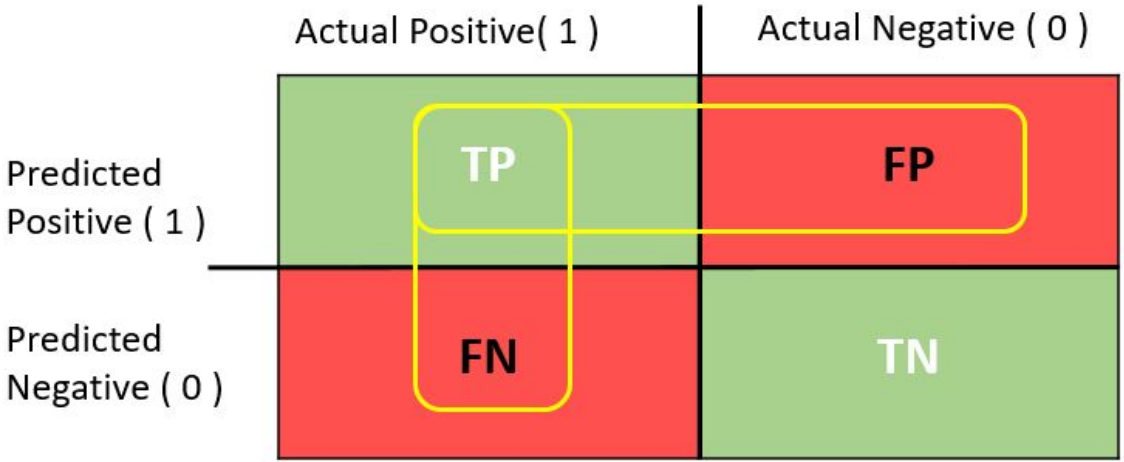
Classification: Confusion Matrix



Reference: 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/richieyuyongpoh/creating-value-through-data-transformation

Classification: Precision & Recall



Reference: here

Regression: RMSE & Rsquare

What…
maths equation…

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

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

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

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

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

Reference: here

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

6) Cloud

E: Data Engineering

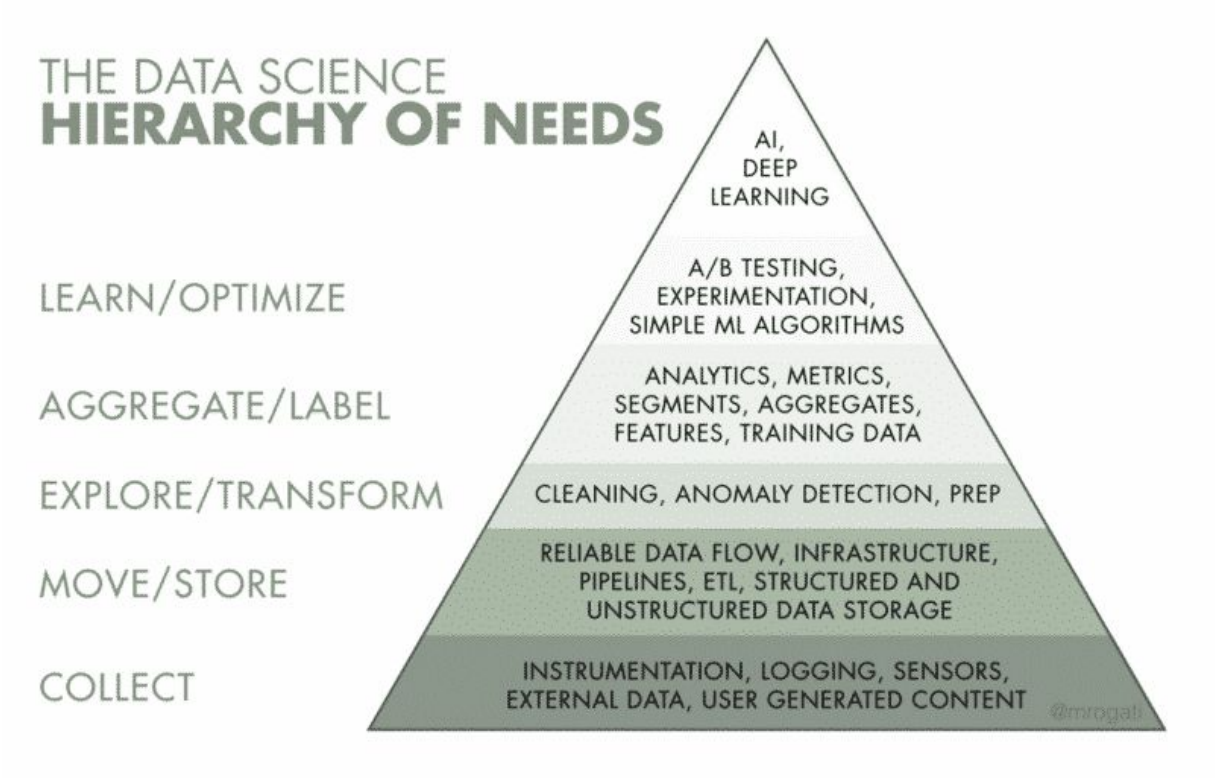


Image Credit: HERE

Data Engineering

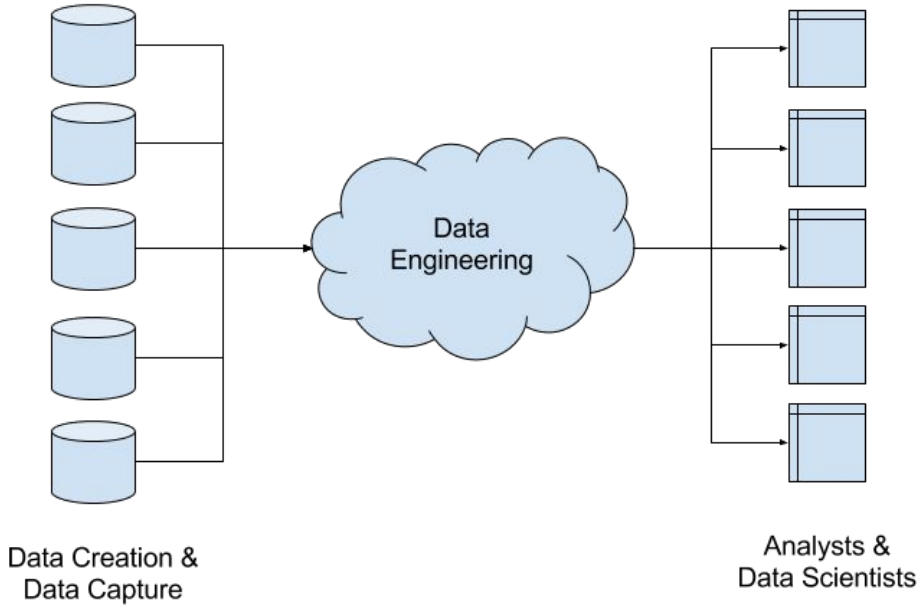


Image Credit: HERE

Data Engineering

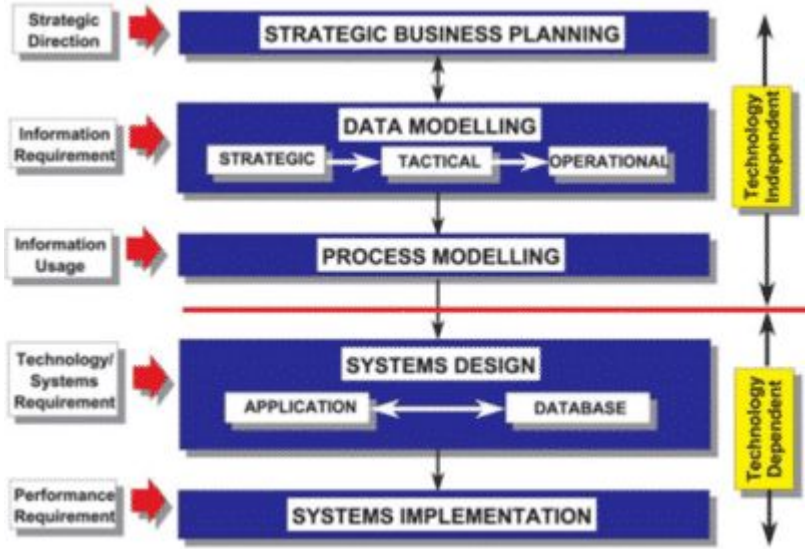


Image Credit: HERE

Data Engineering

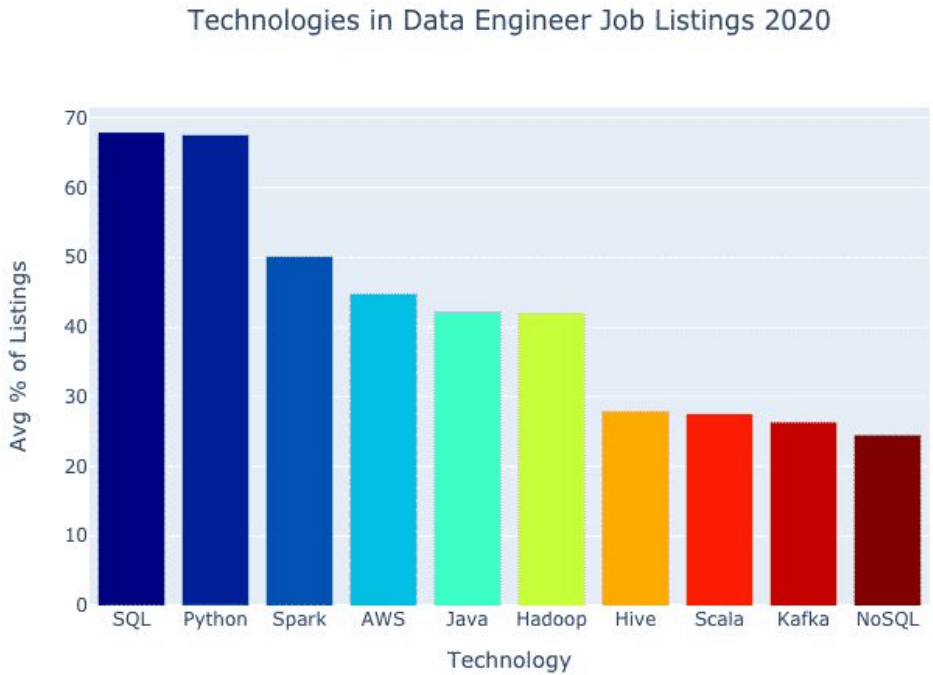


Image Credit: HERE

Data Engineering

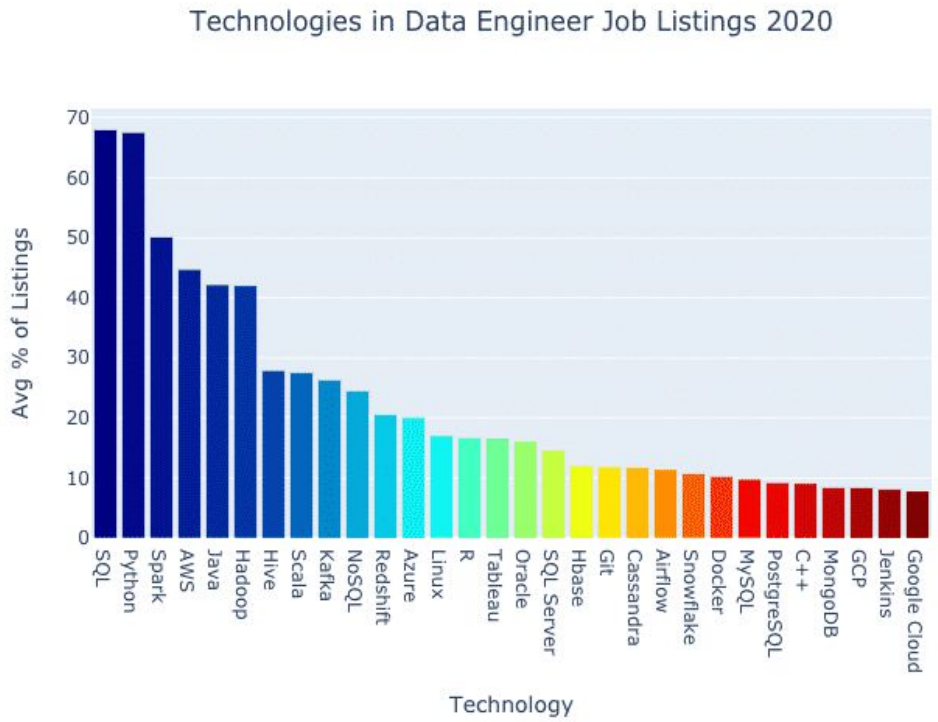


Image Credit: HERE

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/richieyuyongpoh/creating-value-through-data-transformation

Data Engineering

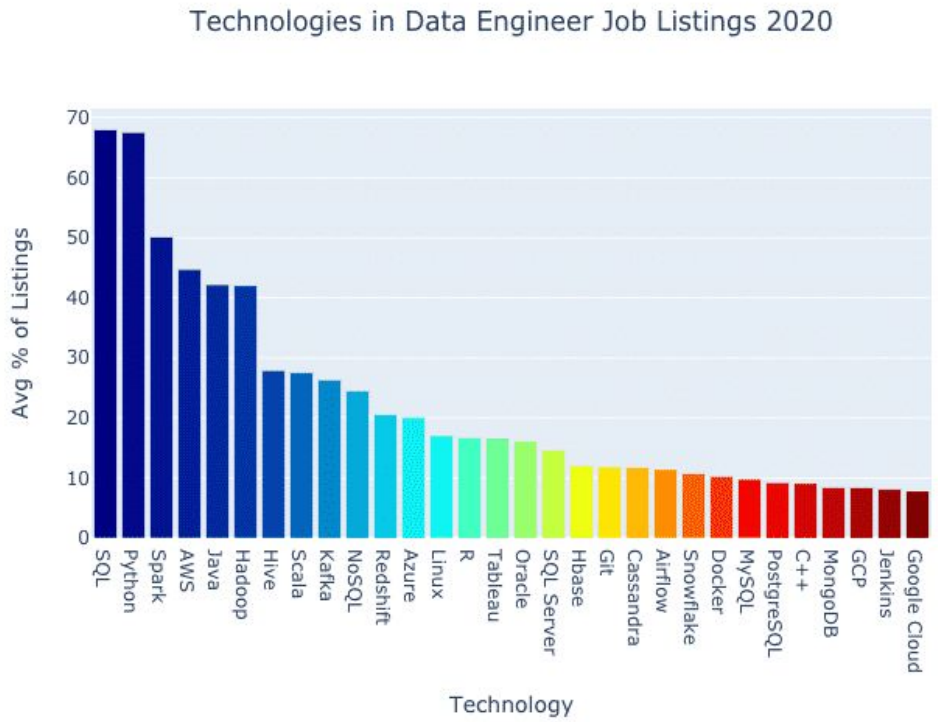


Image Credit: HERE

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/richieyuyongpoh/creating-value-through-data-transformation

MLOps - Level 0

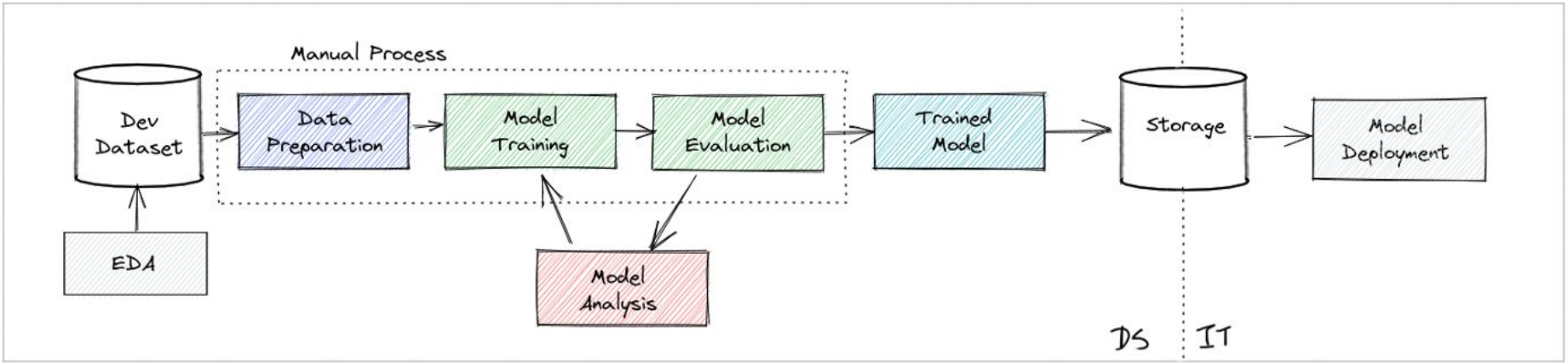


Image Credit: HERE

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/richieyuyongpoh/creating-value-through-data-transformation

MLOps - Level 1

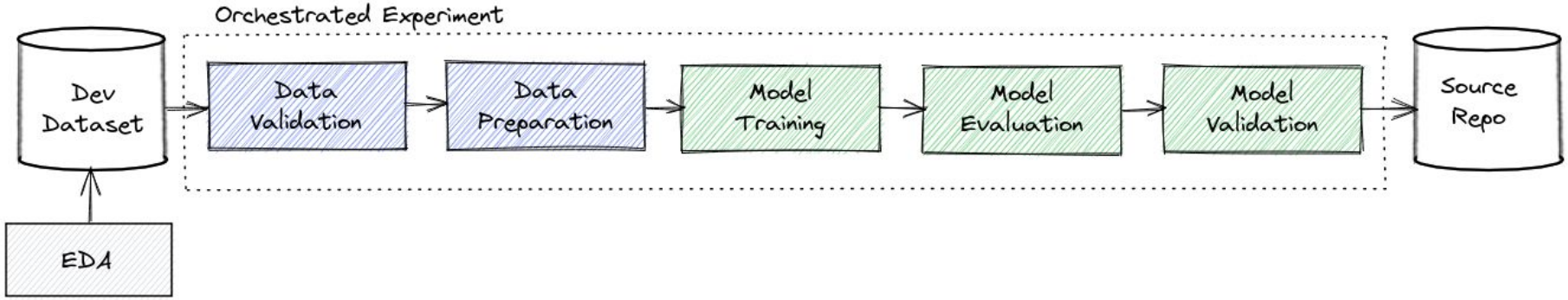


Image Credit: HERE

MLOps - Level 2

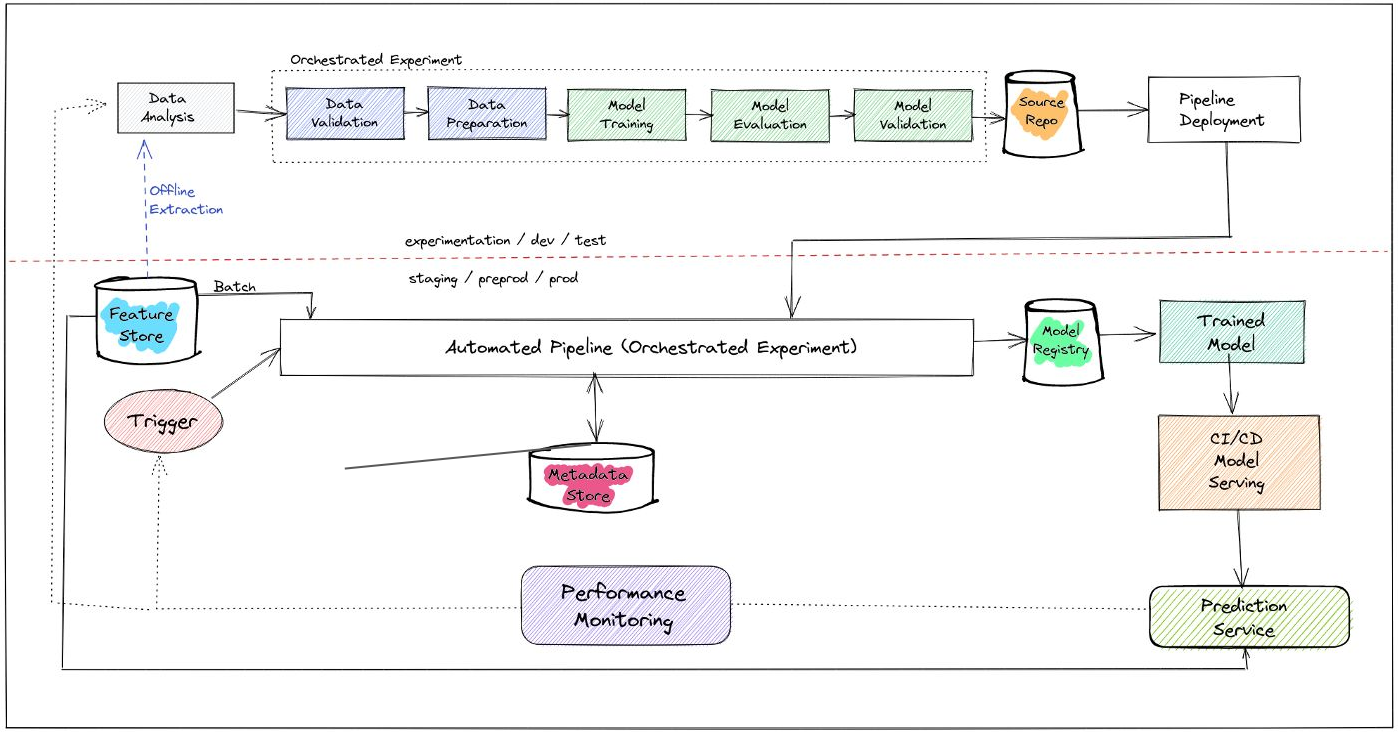


Image Credit: HERE

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/richieyuyongpoh/creating-value-through-data-transformation

F: Business Intelligence & Data Storytelling

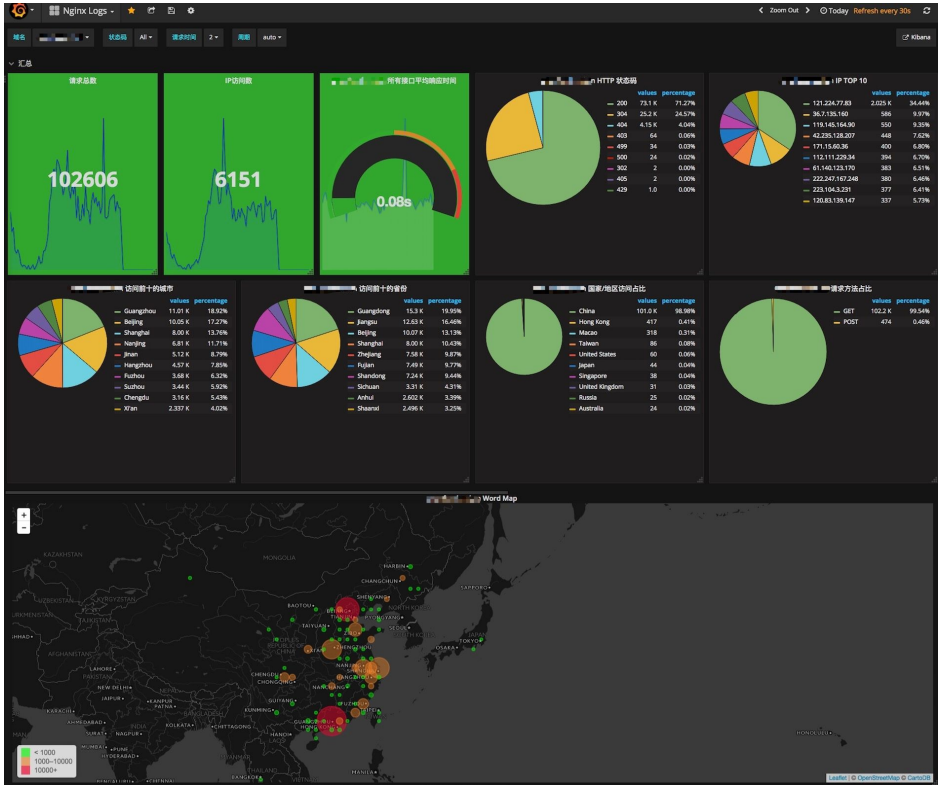
I am listening now…

"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

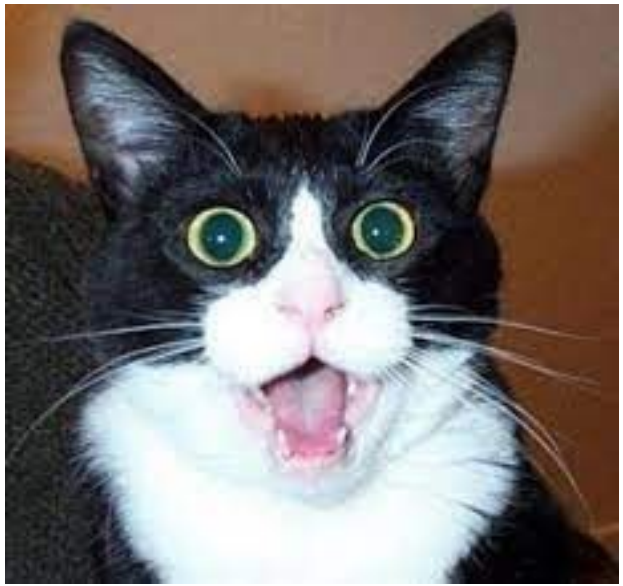


Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
https://github.com/richieyuyongpoh/creating-value-through-data-transformation

G: Unstructured Data

Well… How about unstructured data?

80% of global data will be unstructured by 2025



Source: HERE

Unstructured Data: Image

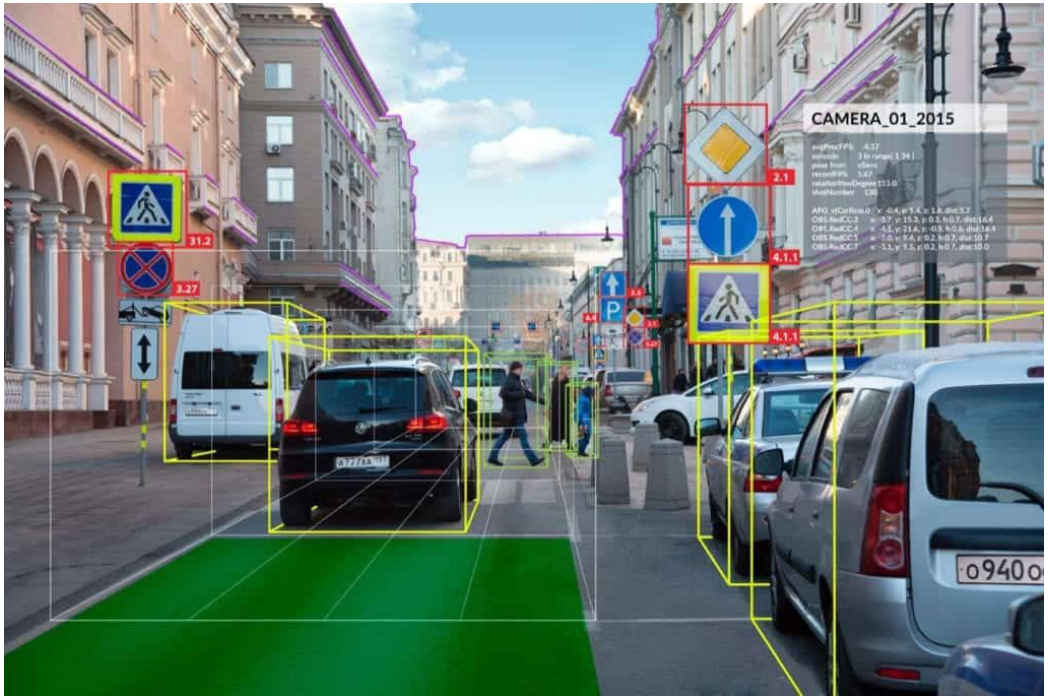
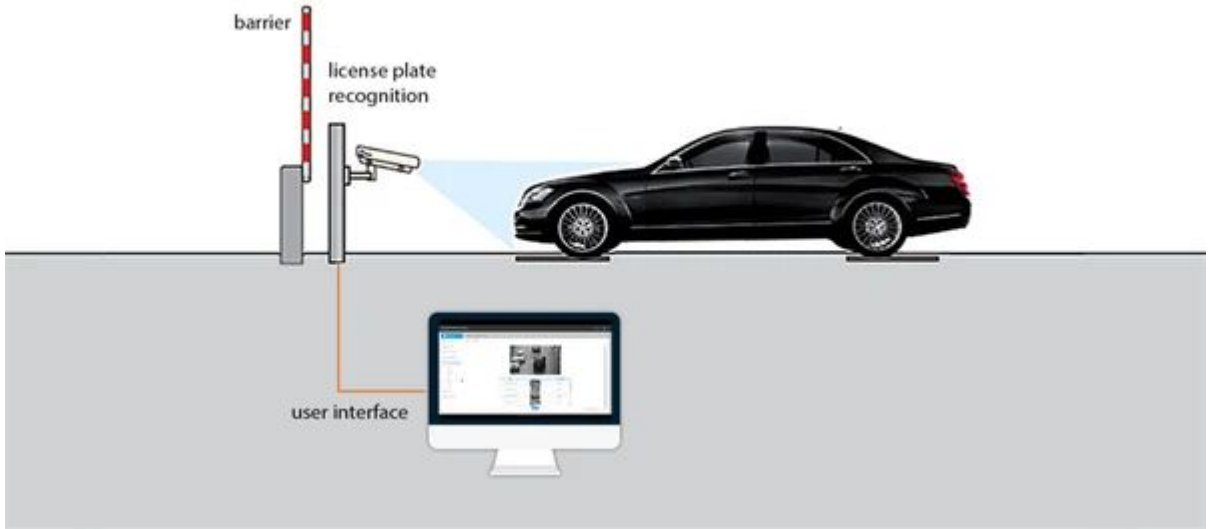


Image Credit: HERE

Image - Background

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

Example 1: Smart Parking System



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

Example 2: Manufacturing Defect Detection

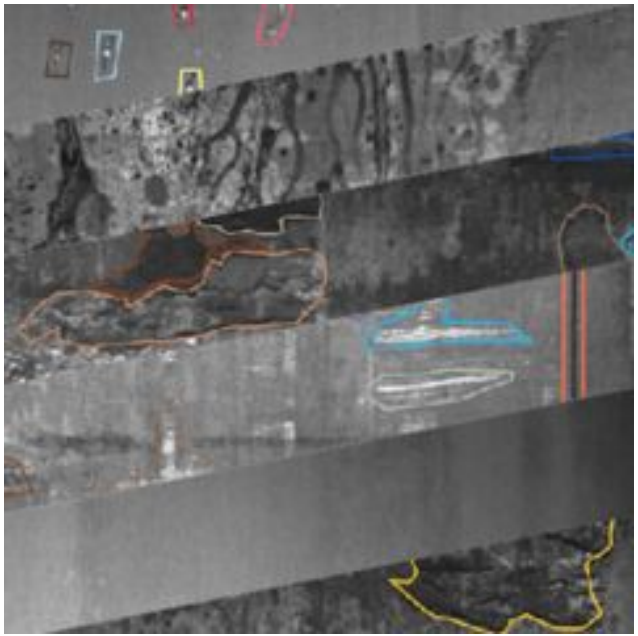


Image Credit: HERE

Example 3: Face Recognition Attendance System

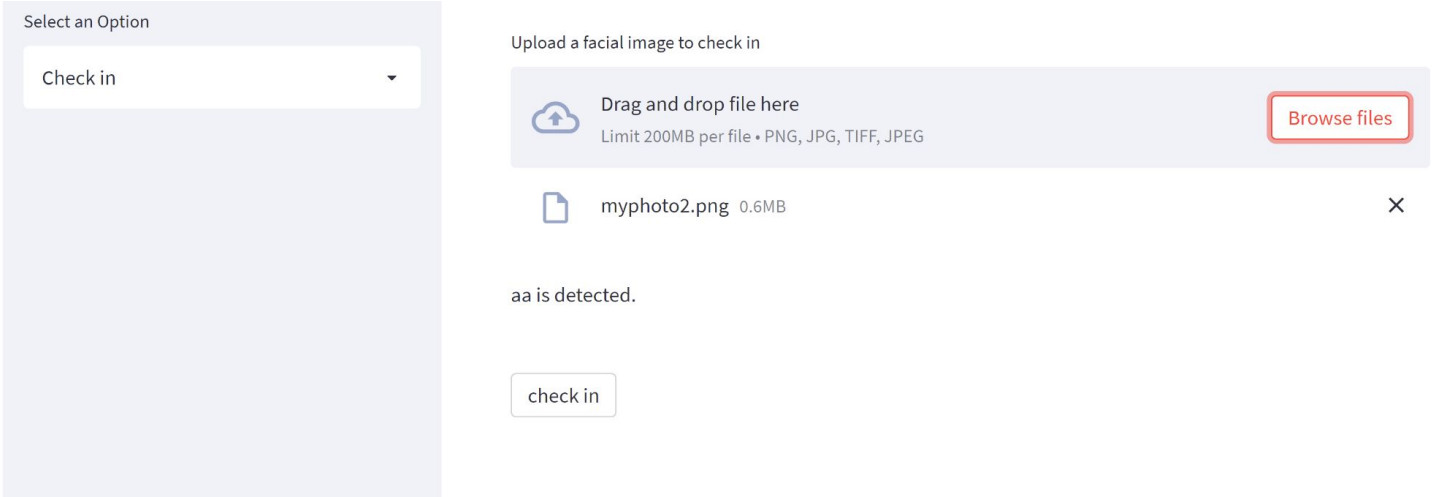


Image Credit: HERE

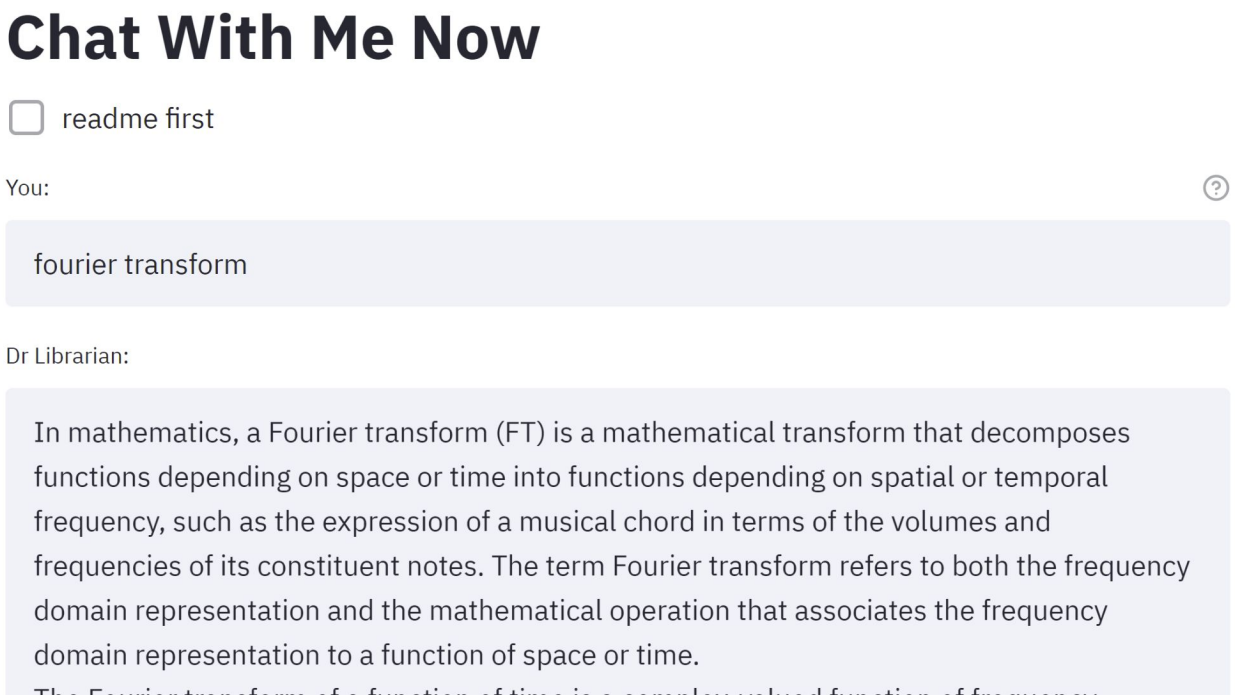
Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment.. https://github.com/richieyuyongpoh/creating-value-through-data-transformation

Demo: Face Recognition Attendance System

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

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

Unstructured Data: Text

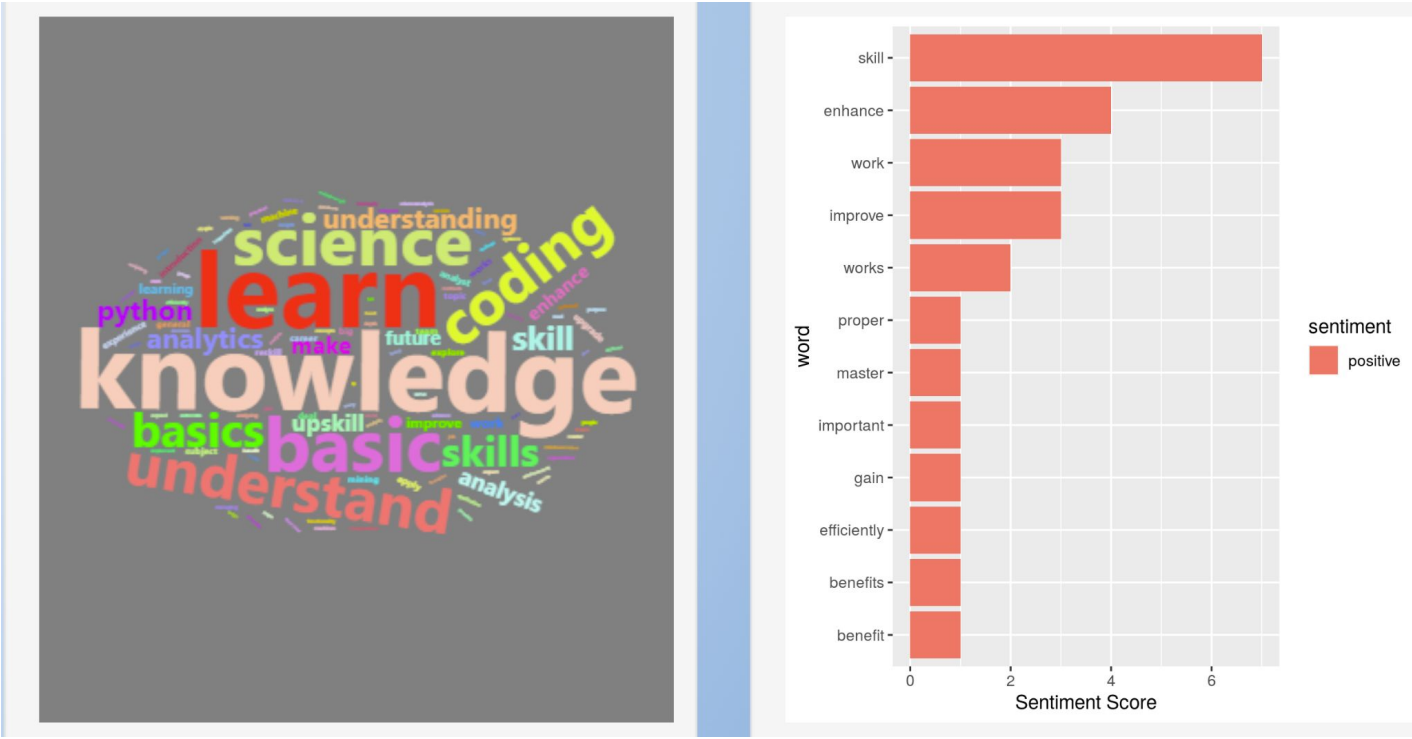


Text - Background

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

Example 1: Chatbot

Example 2: Sentimental Analytics

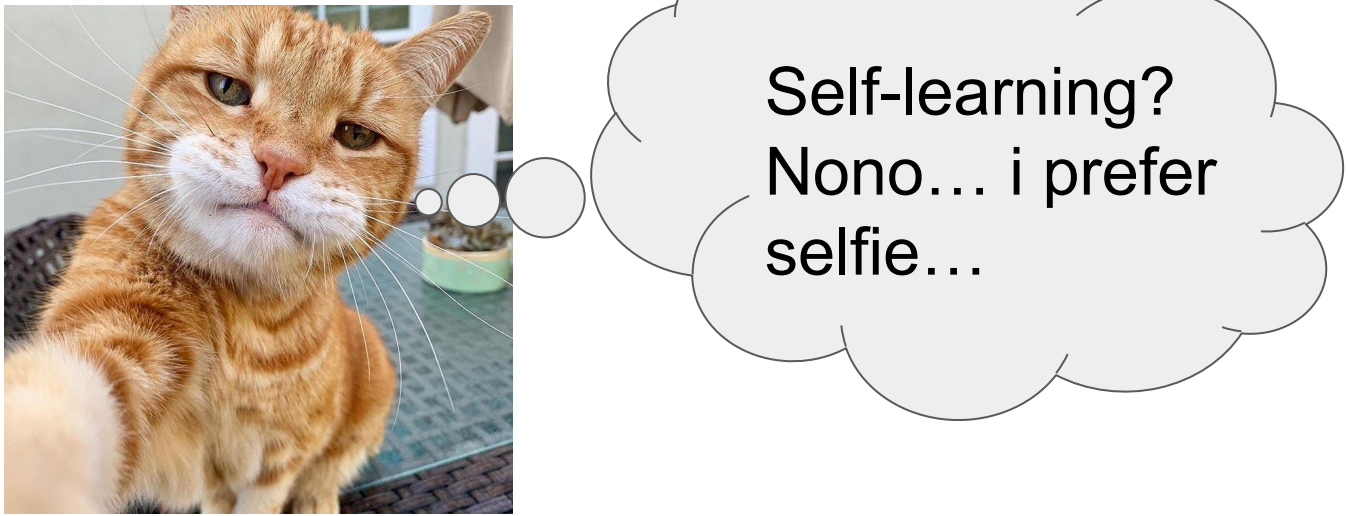


Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
https://github.com/richieyuyongpoh/creating-value-through-data-transformation

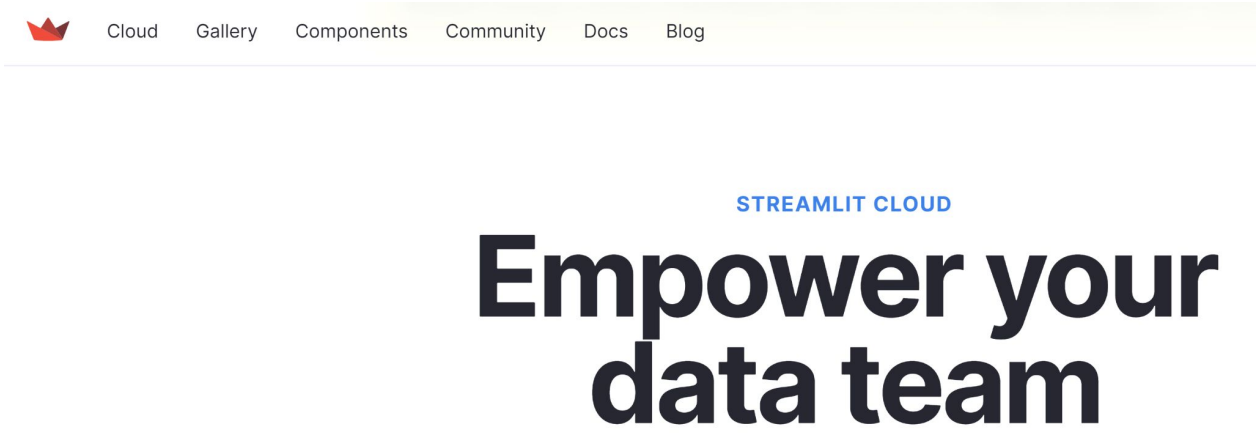
Demo: Chatting with Dr. Librarian

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

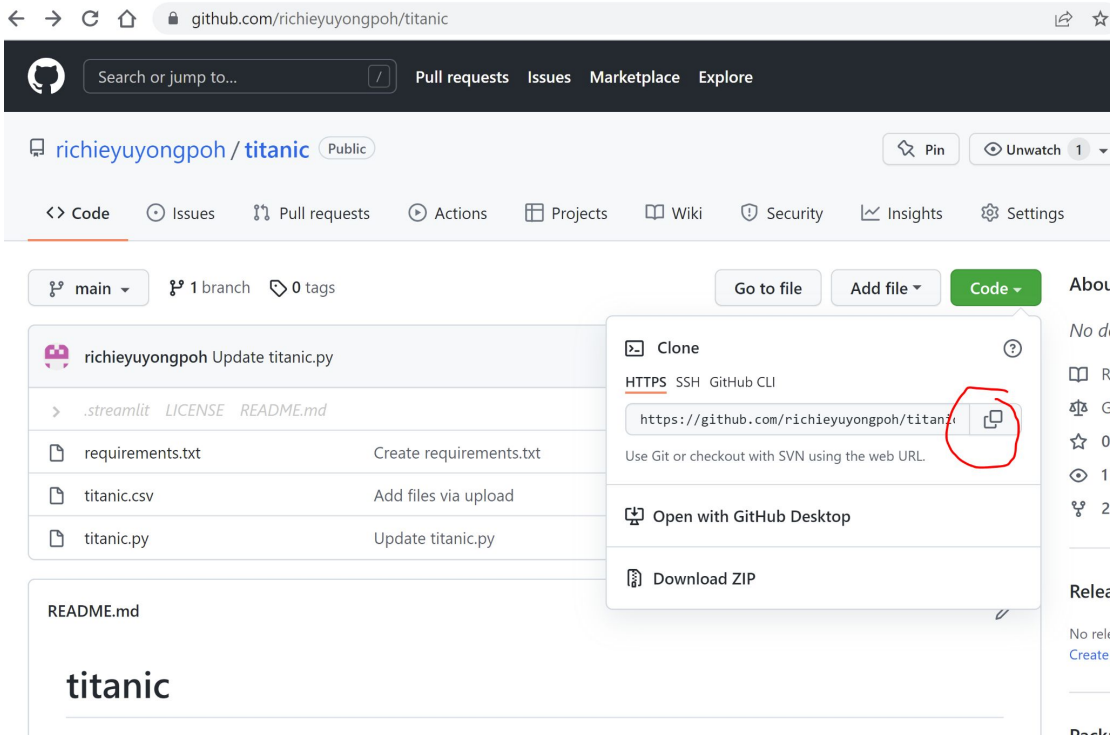
Bonus: Self-learning Activity - Project Deployment



Step1: Create a github & streamlit cloud

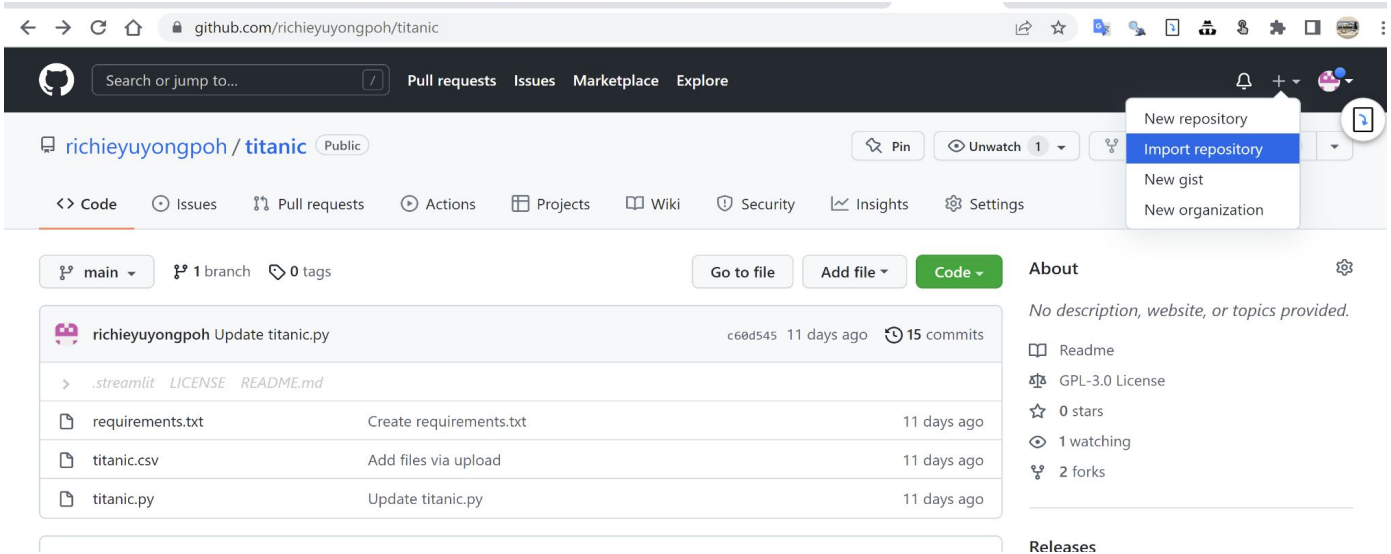


Step2: Clone the TITANIC repository

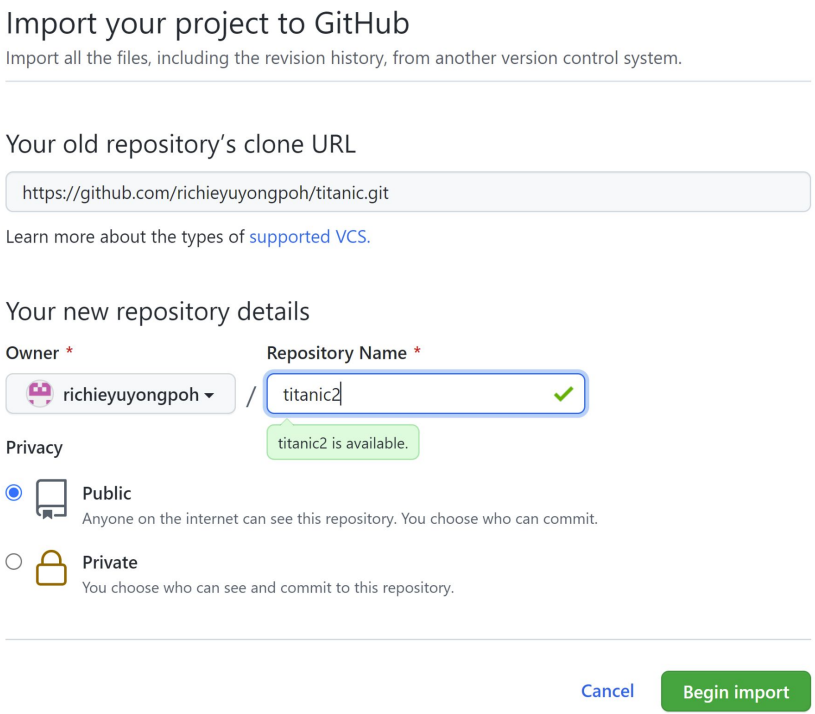


Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
https://github.com/richieyuyongpoh/creating-value-through-data-transformation

Step3: Import it to your new repository

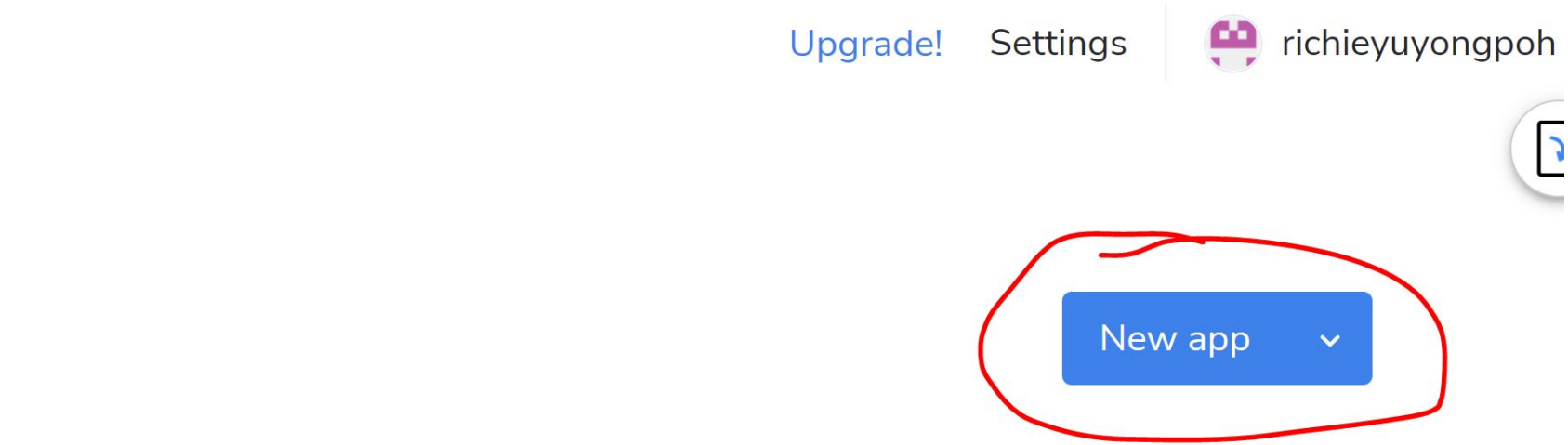


Step3: Import it to your new repository

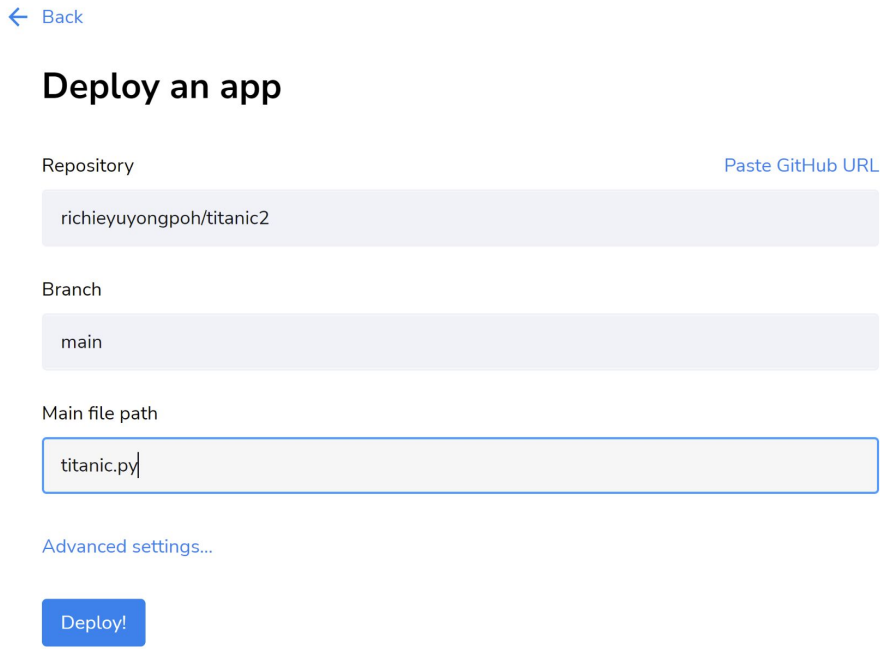


Cite this work if you are using it: Y.P. Yu (2021). Introduction to Big Data Analytics - From Data Management to Project Deployment..
https://github.com/richieyuyongpoh/creating-value-through-data-transformation

Step4: Create a new app using streamlit cloud



Step5: Deploy the app



Step6: Congratulations

Your app is in the oven


Q
A


THANK YOU...
You Rock!