



Introduction to Data Science

2110446 Data Science and Data Engineering (2025/2)

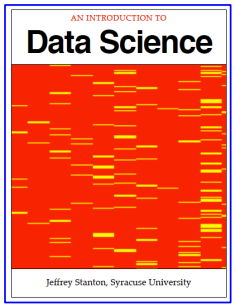
Prof. Peerapon Vateekul, Ph.D.
Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University
Peerapon.v@chula.ac.th
www.cp.eng.chula.ac.th/~peerapon/

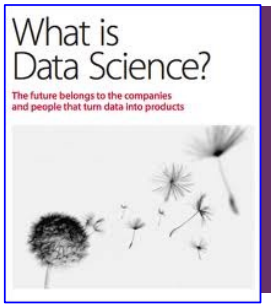
Outline: Understand definitions and terminologies

- Introduction
  - Data is important
  - Data Science Definition by Dr.Virote
  - Data Science Definition by Aj.Natawut
- Key Data Science Activities
  - Data Science Process
  - Types of Data Science Projects
  - AI/ML/DL/GenAI
  - Data Engineering + Big Data Analytics
  - MLOps
- Conclusion
- Disclaimer

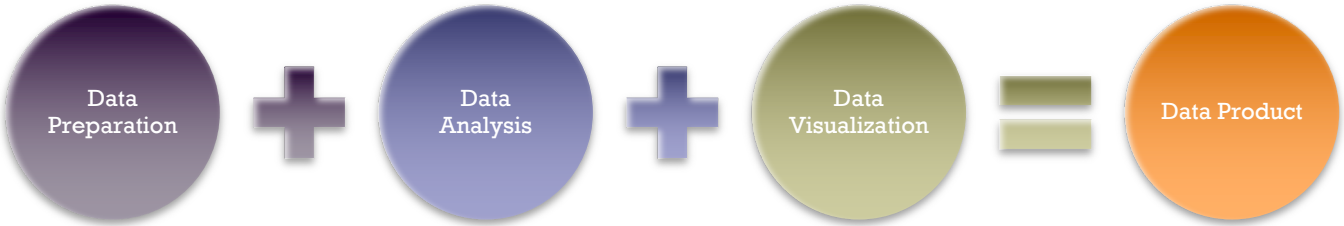
Introduction

What is data science? (aka. data analytics)



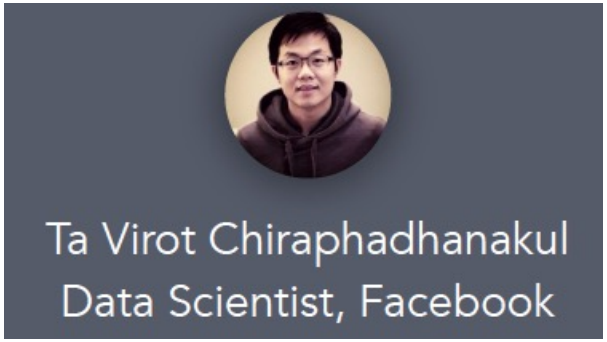


- Data
  - Facts and statistics collected for reference or analysis
- Science
  - A systematic study through observation and experiment
- Data Science
  - The scientific exploration of data to extract meaning or insight,
  - and the construction of software to utilize such insight in a business context.



What is data science? (cont.)

1. Transform data into valuable insights
2. Transform data into data products
3. Transform data into interesting stories



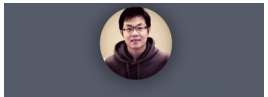
Code Mania 2 (01), Jan-2015

1) Transform data into valuable insights

How Obama Raised $60 Million by Running a Simple Experiment

By Dan Siroker
November 29, 2010
A/B Testing Case Studies, Multivariate Testing, Political Campaign

1) Transform data into valuable insights (cont.)



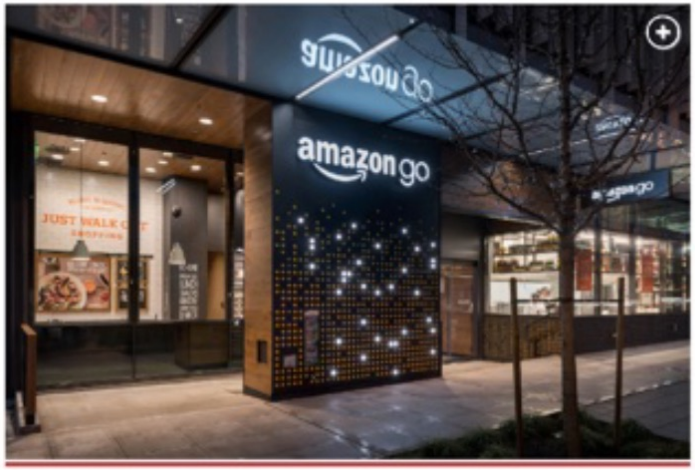
Ta Virot Chiraphadhanakul
Data Scientist, Facebook
Code Mania 2, Jan-2015

NEW YORK POST

BUSINESS

Amazon introduces next major job killer to face Americans

By James Covert, Linda Massarella and Bruce Golding December 5, 2016 | 9:59pm | Updated



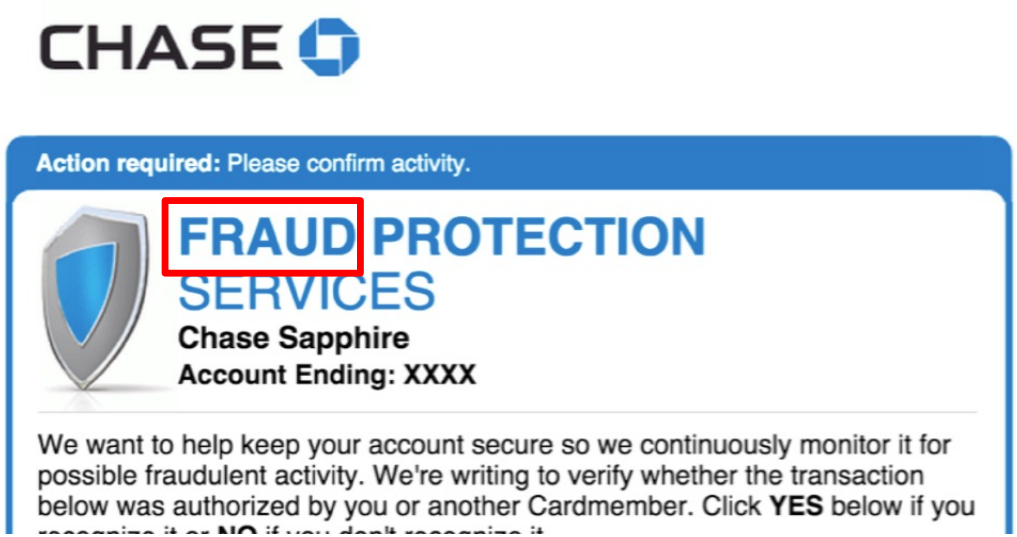
The Amazon Go storefront

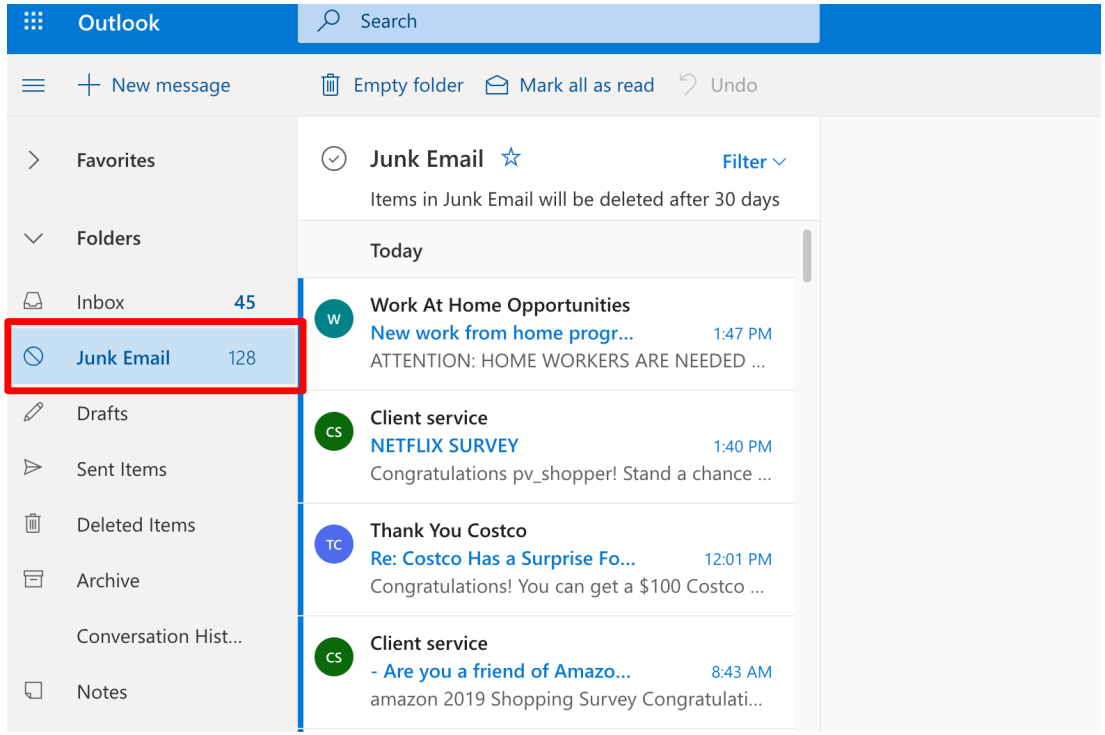
NEW YORK POST POETICS

Amazon's new supermarket will blow your mind — and cost America jobs
Why employment may be optional in the near future
Amazon's futuristic grocery store could spell employment doom
Amazon introduces next major job killer to face Americans
Amazon's latest idea could kill off jobs forever

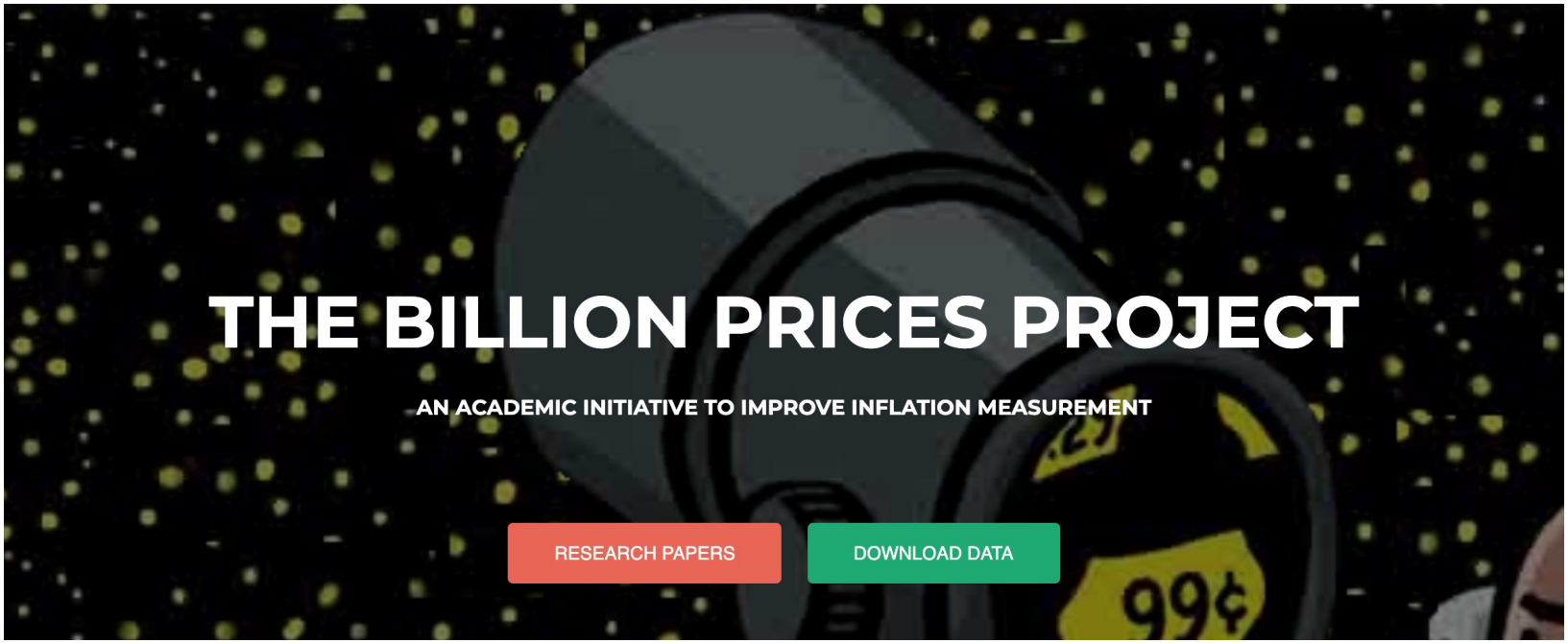
http://nypost.com/2016/12/05/amazon-introduces-next-major-job-killer-to-face-americans/

2) Transform data into data products





3) Transform data into interesting stories Consumer Price Index (CPI) - Inflation



http://www.thebillionpricesproject.com/


mini Big C
แลกซื้อสุดคุ้ม
เมื่อซื้อสินค้าที่มินิบิ๊กซี ครบ 79 บาทขึ้นไป
22-28
แลกซื้อ 10 บาท
แลกซื้อ 7 บาท
POCKY Choco Banana
MILO
แลกซื้อ 69 บาท
บรีส
แลกซื้อ 35 บาท
แลกซื้อ 19 บาท
ลดสุดสุดไปเลย
ประหยัด 2 บาท
ประหยัด 15 บาท
ประหยัด 15 บาท
เติมเงินฟินเว่อร์ ใช้คูปองลดทันที
เมื่อเติมเงิน 150 บาท ขึ้นไป
10 บาท
สิทธิพิเศษ เฉพาะสมาชิก
สะดวก ใกล้บ้านคุณ


TESCO Lotus express
ลดทั้งเดือน
วันที่ 28 ก.ย. - 29 ต.ค. 60
อิชิตัน
โค้ก
ทีพอท
โอวัลติน
โฟร์โมสต์
มอลโคน่า
โอวัลติน
เนสกาแฟ
ย่ำย่ำจัมโบ้
พมรุ้ง
ฉัตร
ลักส์
ซันซิล
คอลเกต
เคลียร์,
ไลปอน
ลอรีเอะ
เทสโก้
มามี่โพโค
ไฮยีน
ดาวน์นี่
บรีส
แอทแทค
โอโมพลัส


ท็อปส์ ได้รับการยกย่องเป็น Trusted Brand
ซึ่งเป็นสุดยอดแบรนด์ที่ผู้บริโภคให้ความเชื่อมั่นมากที่สุด
ต่อเนื่องกันถึง 5 ปี สาขาซูเปอร์มาร์เก็ต จากนิตยสาร รีดเดอร์ส ไดเจสท์
Red Hot
ซื้อ 1 แถม 1 BUY 1 GET 1 FREE
Tops super

The Billion Prices Project: Using Online Prices for Measurement and Research ǂ

Alberto Cavallo MIT and NBER
Roberto Rigobon MIT and NBER

This Version: April 8, 2016

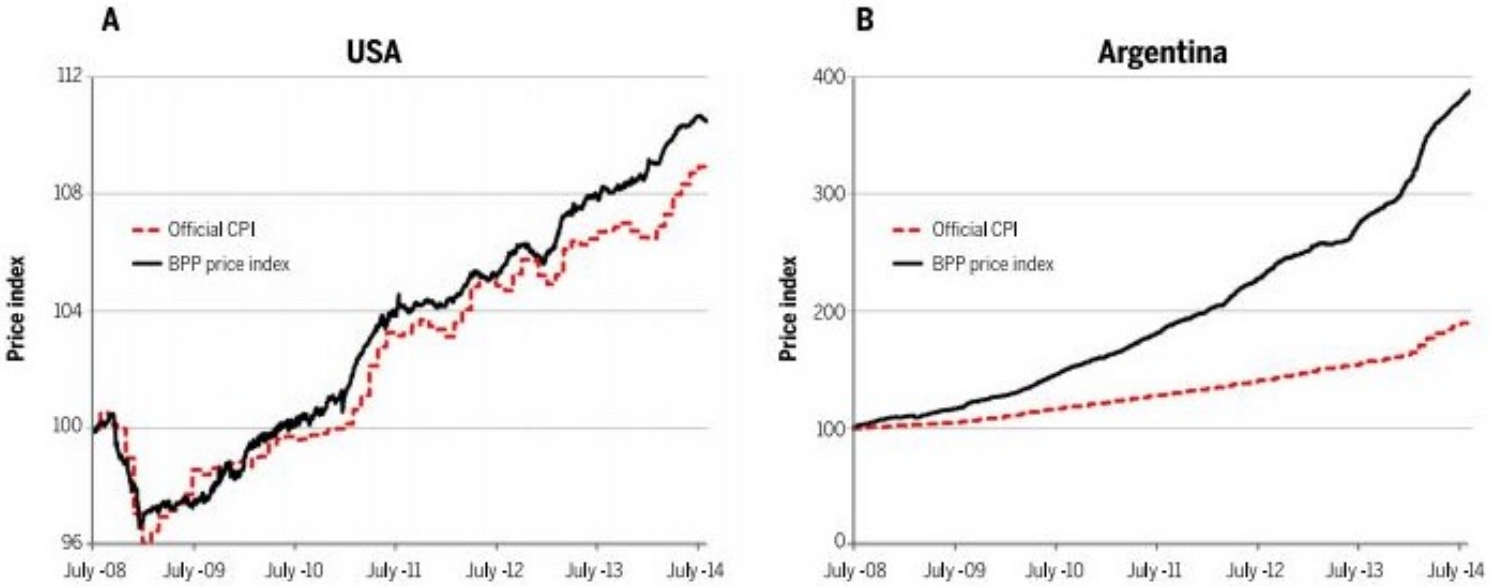
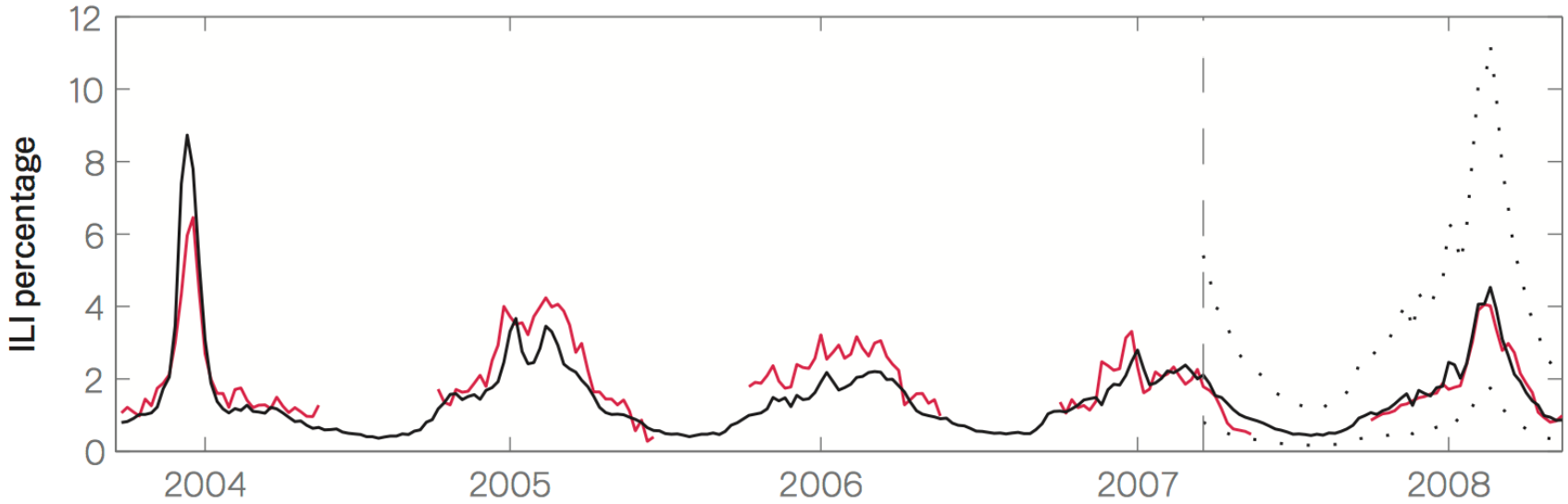


Fig. 2. BPP price index. Dashed red lines show the monthly series for the CPI in the United States (**A**) and Argentina (**B**), as published by the formal government statistics agencies. Solid black lines show the daily price index series, the "State Street's PriceStats Series" produced by the BPP, which uses scraped Internet data on thousands of retail items. All indices are normalized to 100 as of July 2008. In the U.S. context, the two series track each other quite closely, although the BPP index is available in real time and at a more granular level (daily instead of monthly). In the plot for Argentina, the indices diverge considerably, with the BPP index growing at about twice the rate of the official CPI. [Updated version of figure 5 in (*18*), provided courtesy of Alberto Cavallo and Roberto Rigobon, principal investigators of the BPP.]

https://www.hbs.edu/faculty/Publication%20Files/BPP_JEP_m_13b5e009-4162-4f2c-b507-593a9a98c082.pdf

Google Flu Trend



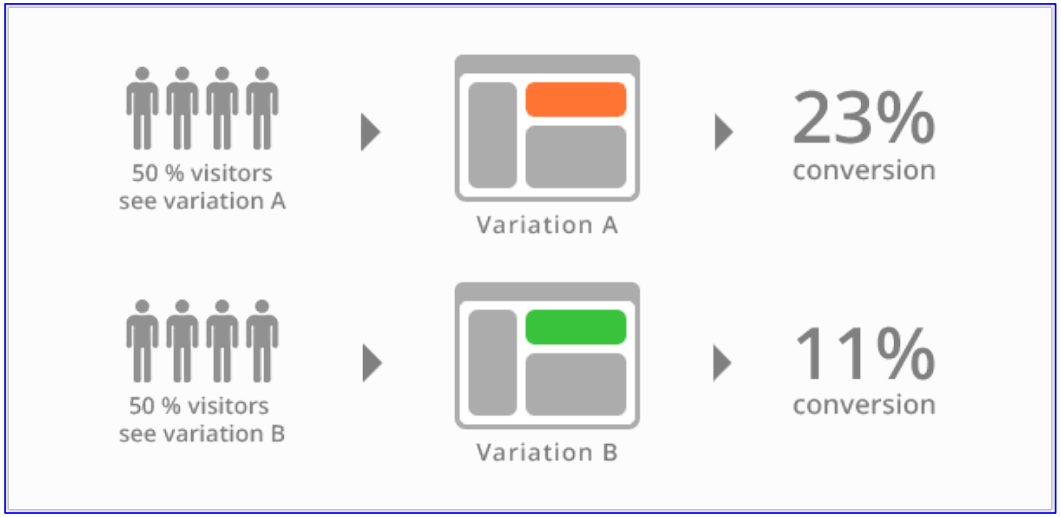
Ginsberg, Jeremy; Mohebbi, Matthew H.; Patel, Rajan S.; Brammer, Lynnette; Smolinski, Mark S.; Brilliant, Larry (19 February 2009). "Detecting influenza epidemics using search engine query data". *Nature*. **457** (7232): 1012–1014.

What are they using data science for?

1. Measurement
2. Insights
3. Data Products

1) Measurement

- To make a decision based on data (aka. Benchmarking)
- Turning qualitative information into quantitative values (metrics or indicators)
- Support comparison:
  - Between options (e.g., which notebook to buy)
  - Before vs. after (e.g., tuning, upgrades)
  - A/B Testing



Source: https://vwo.com/ab-testing/

Example: SimCity

Remove product banner: SimCity sees 43% more conversions without hero banner image

Control

Variation

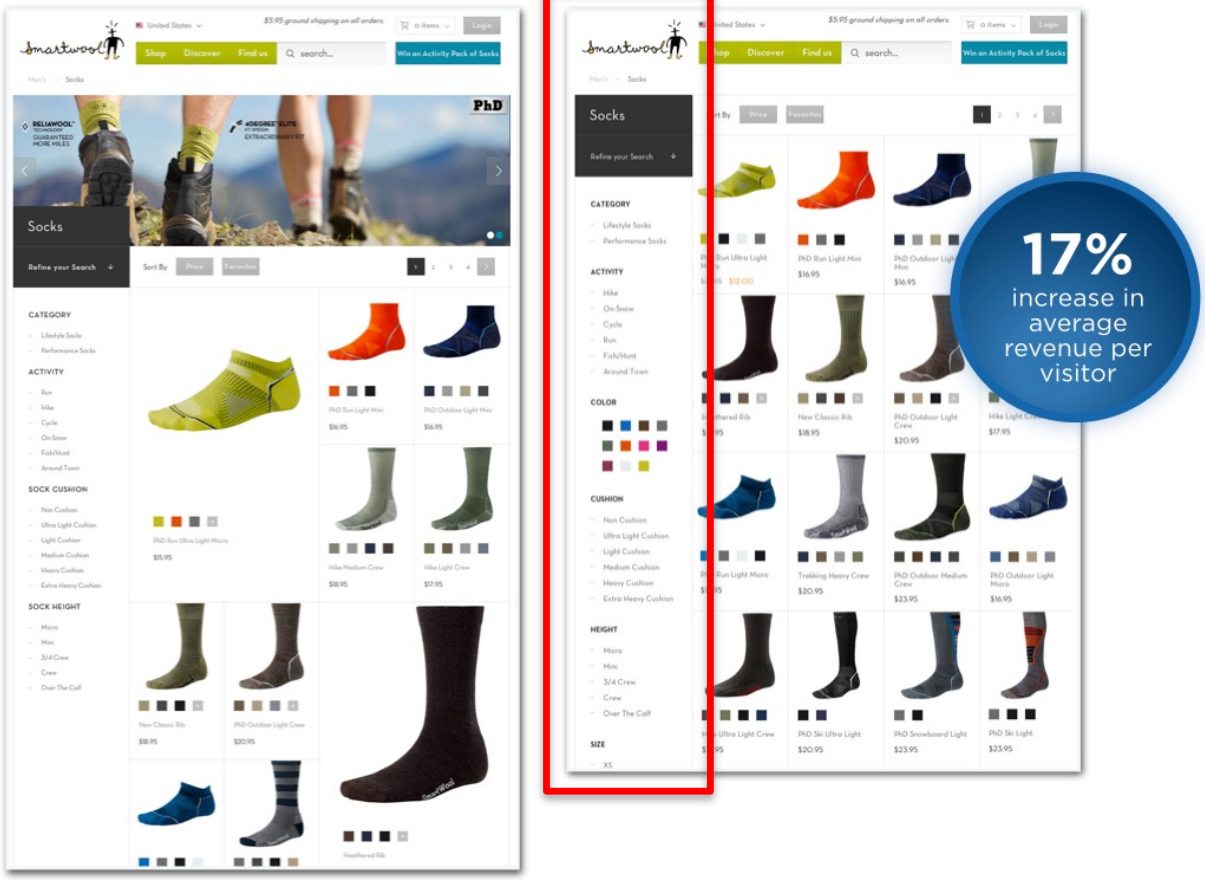
Source: https://blog.optimizely.com/2015/06/04/ecommerce-conversion-optimization-case-studies/

Example: SmartWool

Use a well-defined grid layout for your online shopping experience: Uniform product page images increase ARPV 17% for SmartWool

Control

Variation



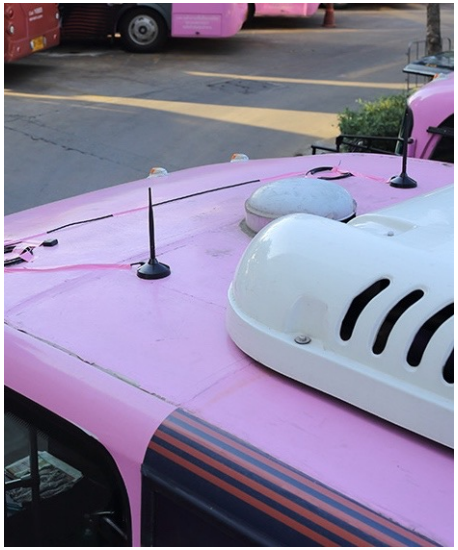
Source: https://blog.optimizely.com/2015/06/04/ecommerce-conversion-optimization-case-studies/

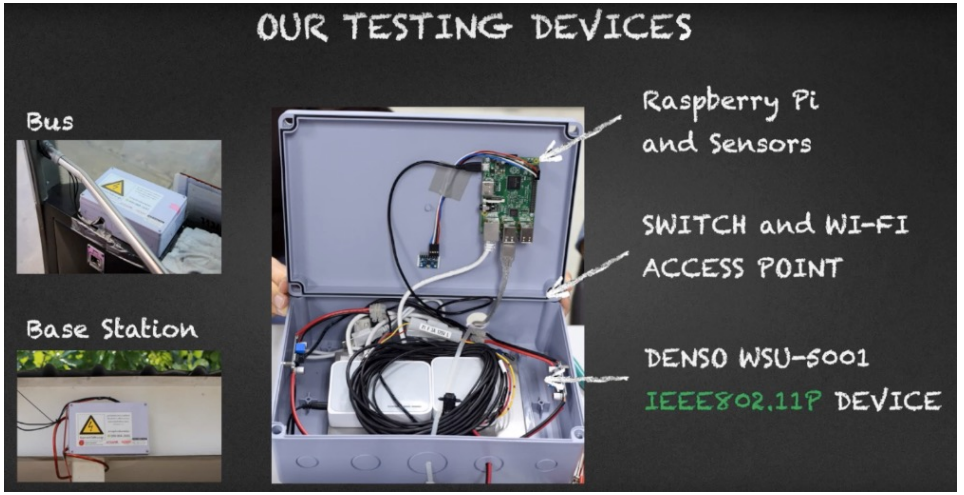
2) Insights

- Good understanding of user behavior can lead to new product development or improvements of the existing products
- Financial startup -- Typing with proper capitalization indicates creditworthiness
  - Online loan applicants who complete the application form with the correct case are more dependable debtors
- Starbucks use customer purchase information from My Starbucks Mobile Apps to figure out new products

https://blogs.scientificamerican.com/guest-blog/9-bizarre-and-surprising-insights-from-data-science/

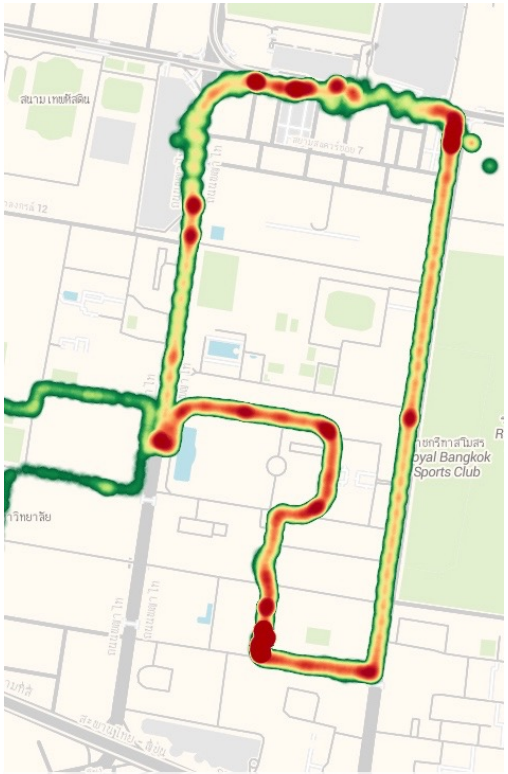
Example: Tracing Traffic



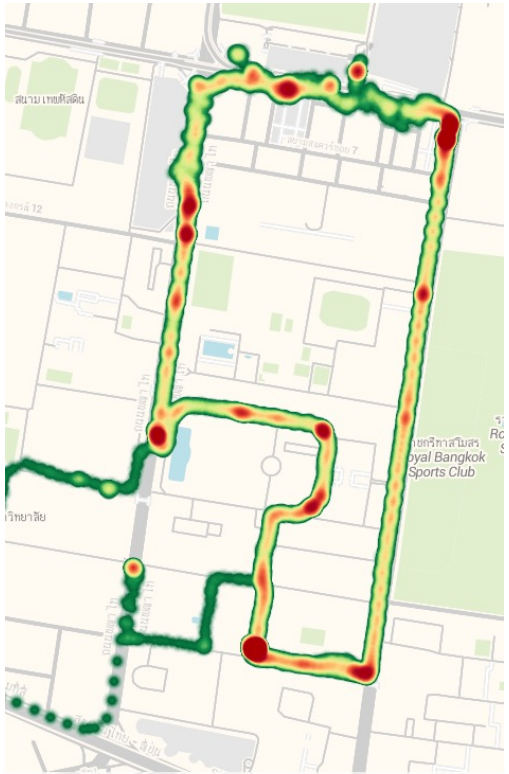


GPS Average Speed

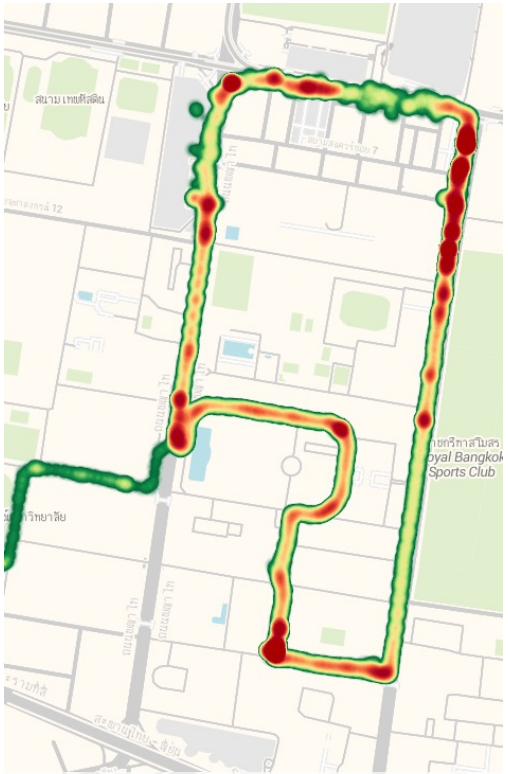
6:00-10:00



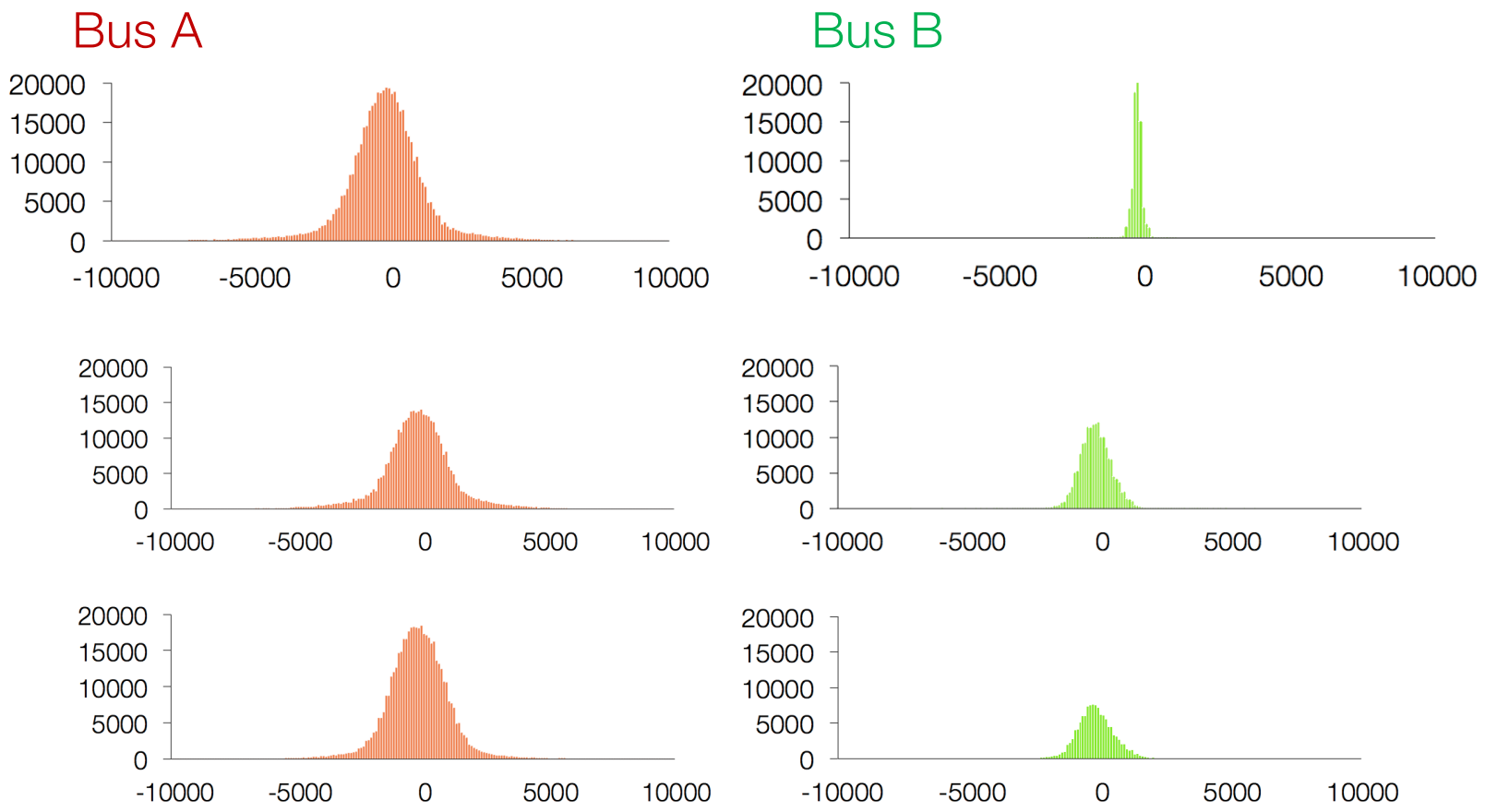
10:00-15:00



15:00-18:00



Bus Drivers' Behaviors

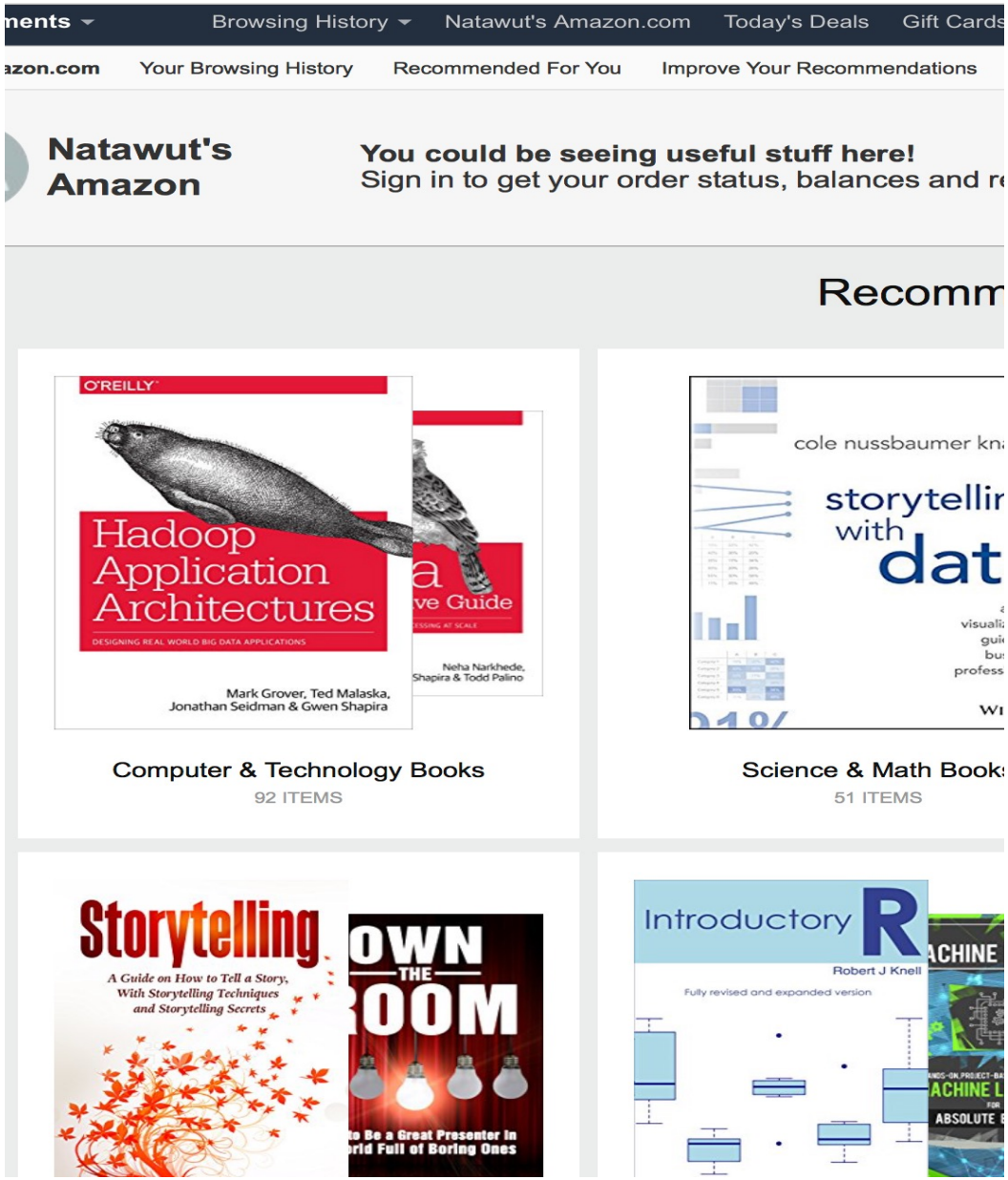


3) Data Products

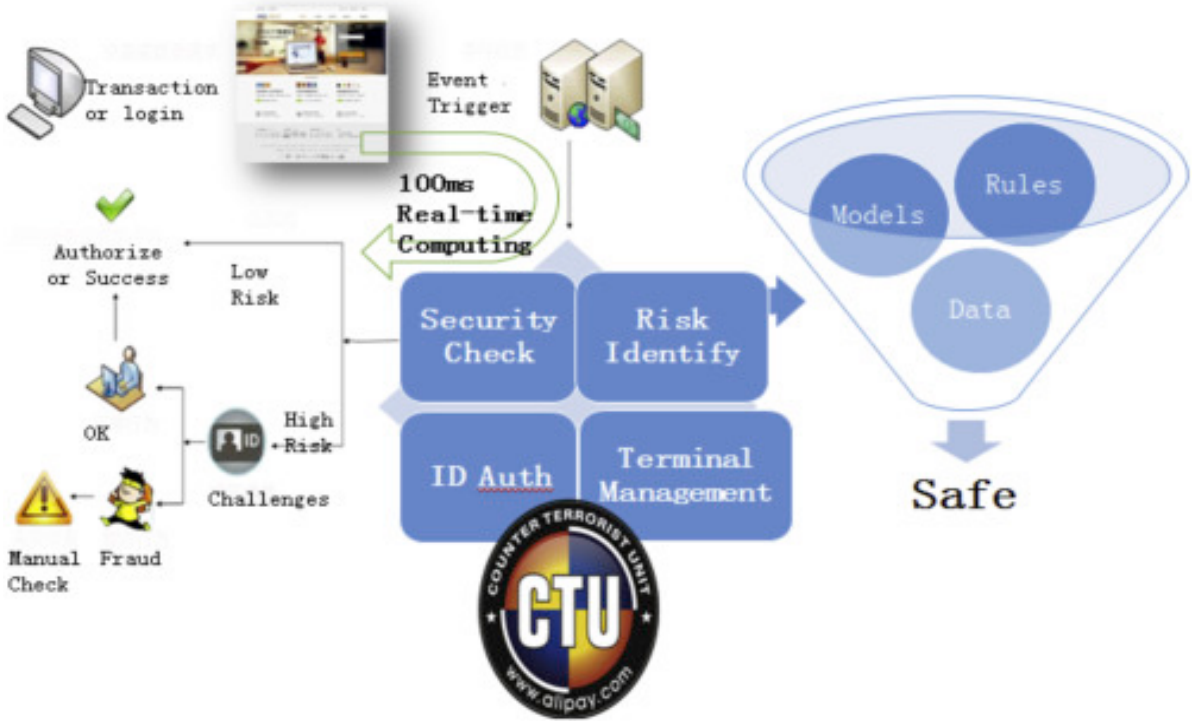
- An application or system that uses data to provide "intelligent" products or services, which create more data that can be further used
- Machine Learning plays an important role in building great data products

Example: Amazon Recommendation

- Amazon sells 480M products (485k new products per day)
- Use recommendation systems to bring products to customers
- Analyze data from 300M customers
  - Purchase history
  - Reviews / Ratings
  - Search history
  - Views

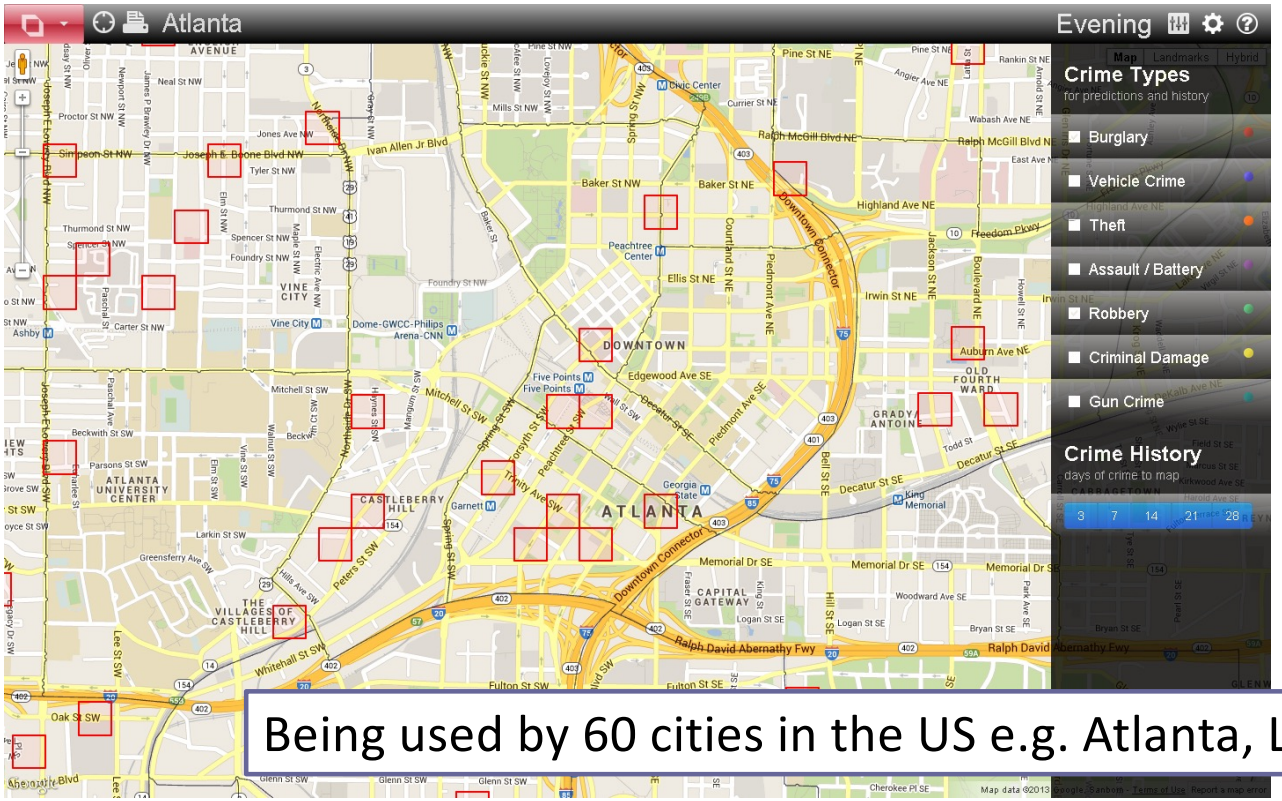


Case study: Alibaba Fraud Detection



Source: http://www.sciencedirect.com/science/article/pii/S2405918815000021

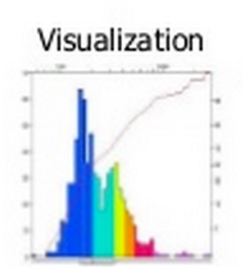
Case study: Predictive Policing

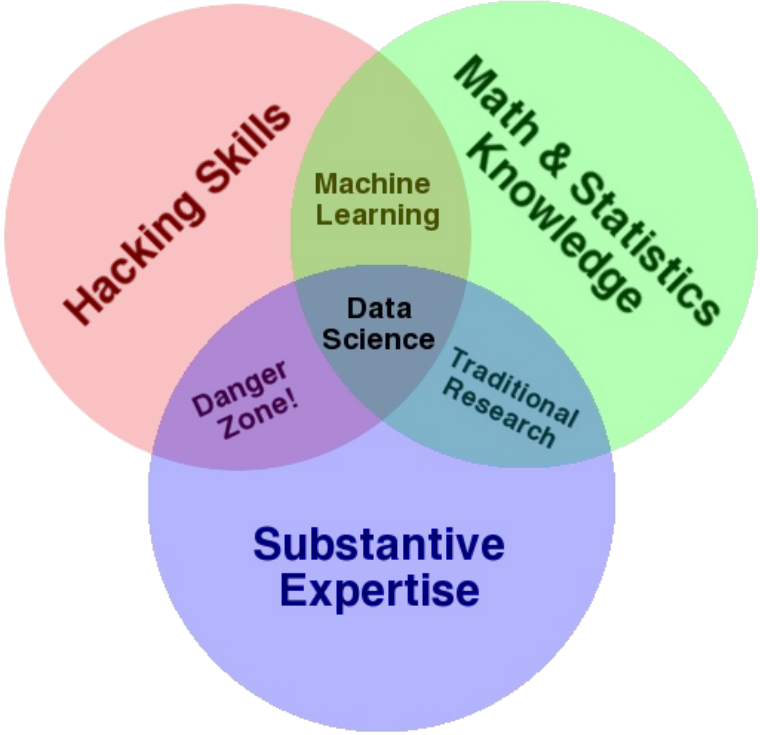


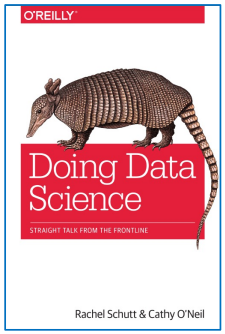
Being used by 60 cities in the US e.g. Atlanta, LA, etc.

Source: http://www.forbes.com/sites/ellenhuet/2015/02/11/predpol-predictive-policing

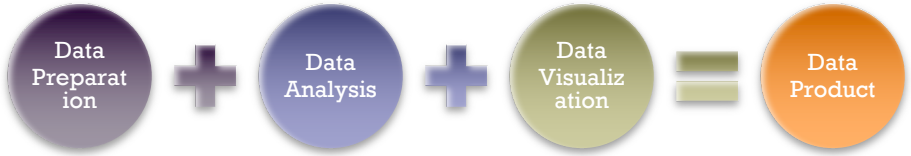
Drew Conway's Data Science Venn diagram (Skills)







Drew Conway's Venn diagram of data science, 2010



Chula Data Science

CH

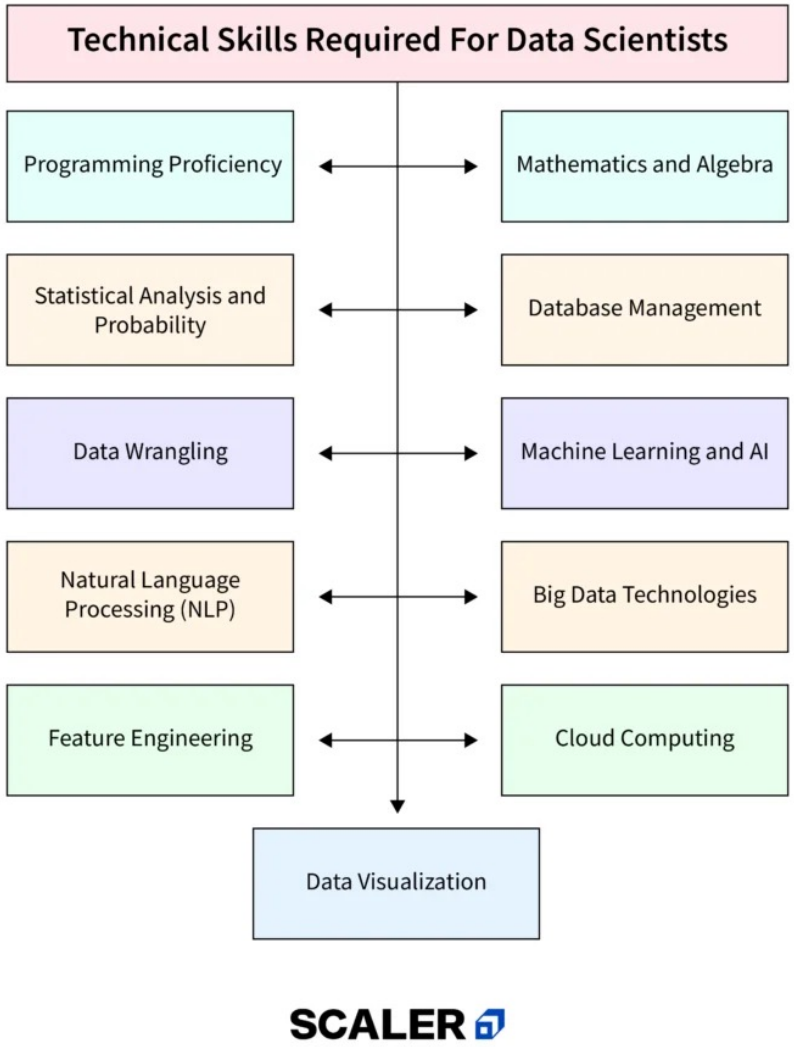
DATA SCIENCE AND BUSINESS ANALYTICS

Top 20 Data Scientist Skills You Need in 2025

24 Min Read

Written by: SCALER TEAM

Last updated: December 18, 2024 7:49 pm



https://www.scaler.com/blog/data-scientist-skills/

Key Data Science Activities

Data Science Process
Types of Data Science Projects
AI/ML/DL/GenAI
Data Engineering
MLOps
Clond Technologies

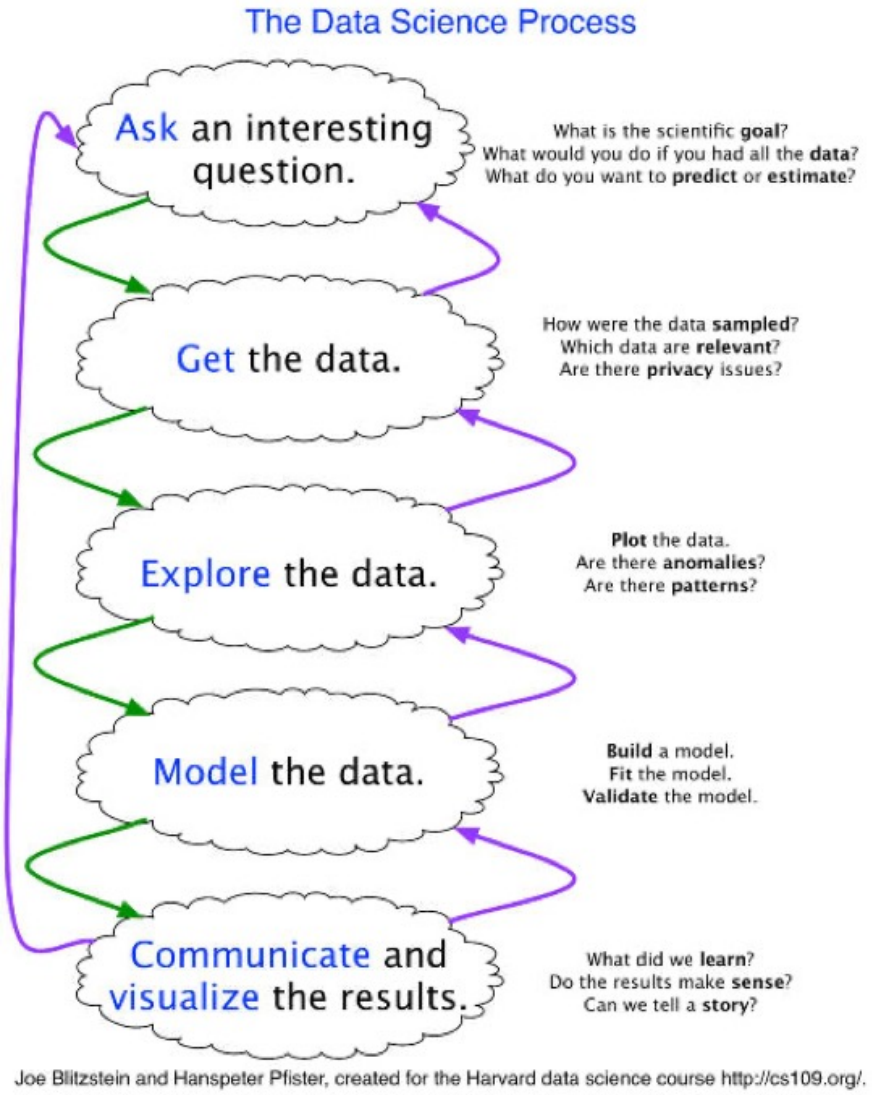
Data Science Process

Dr.Virote

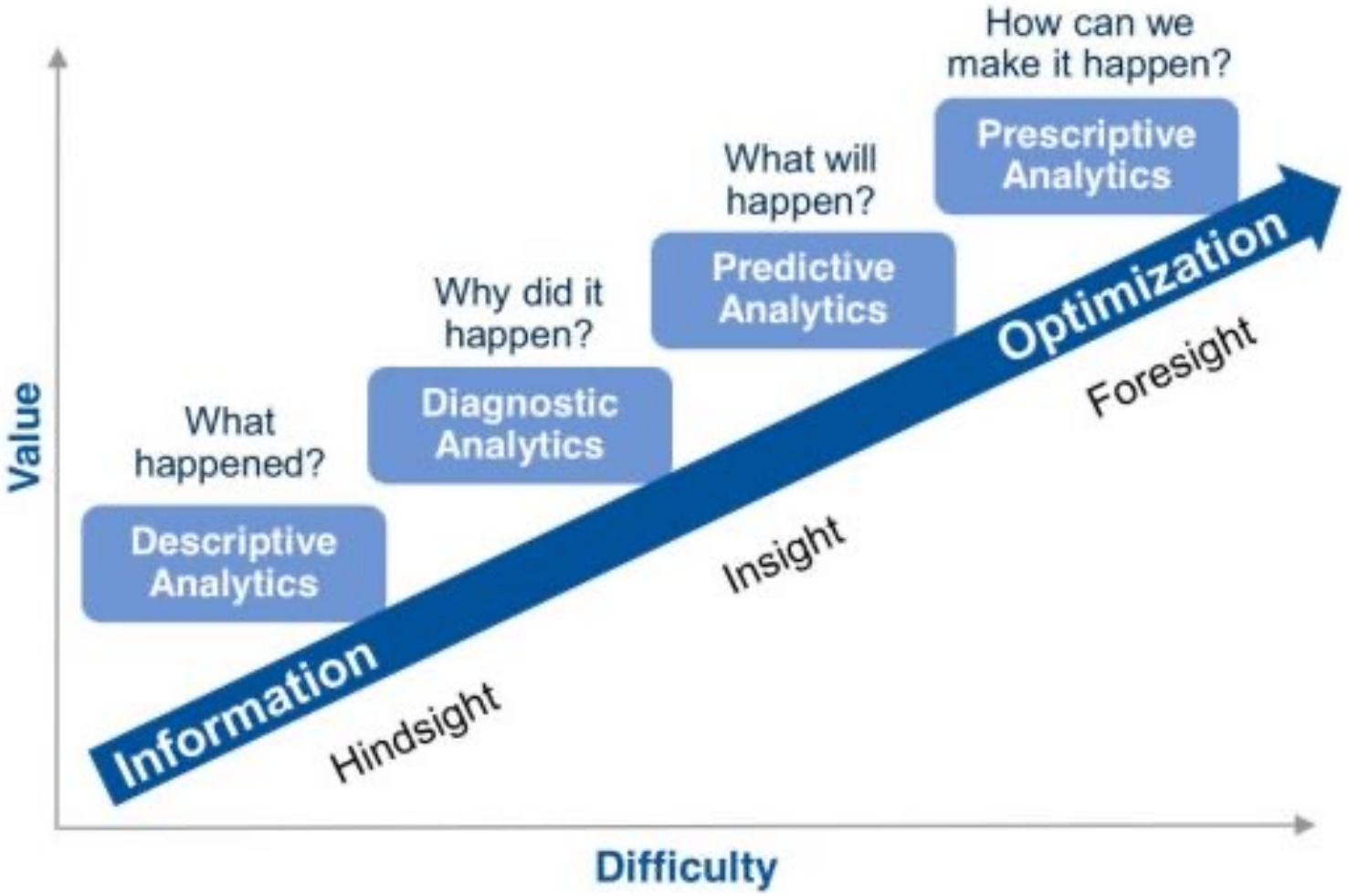
1. Transform data into valuable insights
2. Transform data into data products
3. Transform data into interesting stories

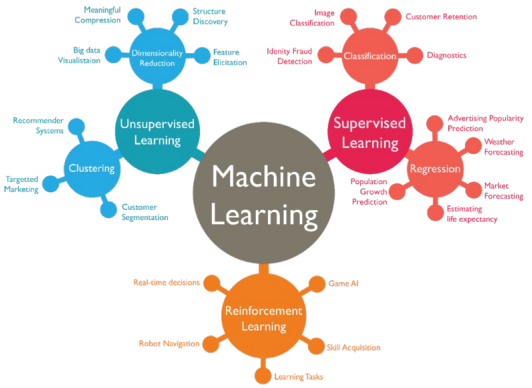
Aj.Natawut

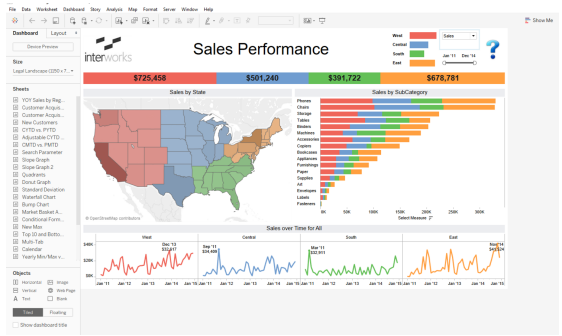
1. Measurement (decision)
2. Insights (knowledge)
3. Data Products (innovation, intelligent)



Data Analytics (Data Science)







Types of Data Science Projects

Valuable insights

- Data visualization
- Analytical skills & storytelling
  - Infographic

Advanced analytics

- AI/Machine Learning/Deep Learning
  - Prediction, Forecasting, Clustering, etc.

Retail Analysis Sample

Ask a question about your data

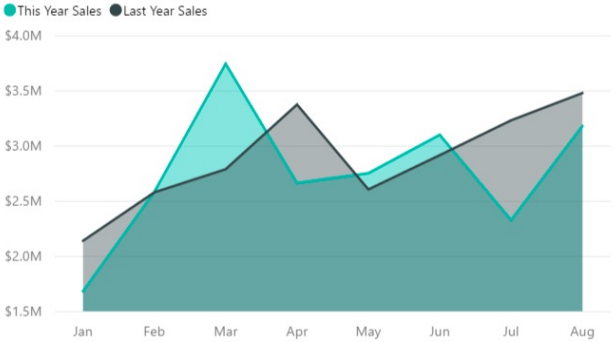
Total Stores
NEW & EXISTING STORES
104

This Year's Sales
NEW & EXISTING STORES
$22M

This Year's Sales
BY CHAIN


New Stores, New Stores Target
YEAR TO DATE

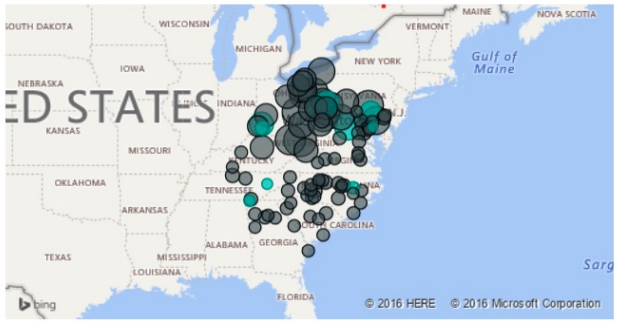

This Year's Sales
NEW STORES ONLY
$2M

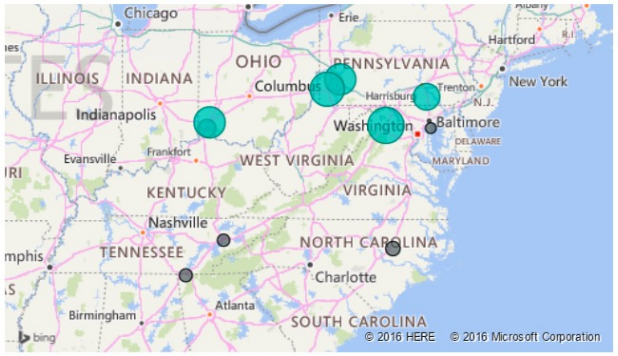
This Year's Sales, Last Year's Sales
BY FISCAL MONTH


Total Sales Variance %, Sales Per Sq Ft, This Year's Sales
BY DISTRICT

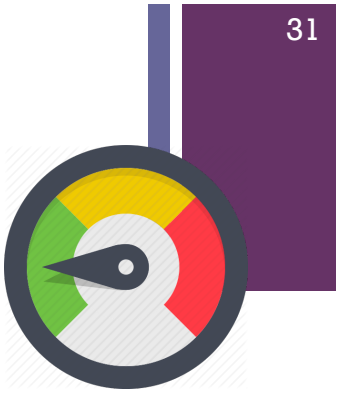

New Stores
NEW STORES ONLY
10

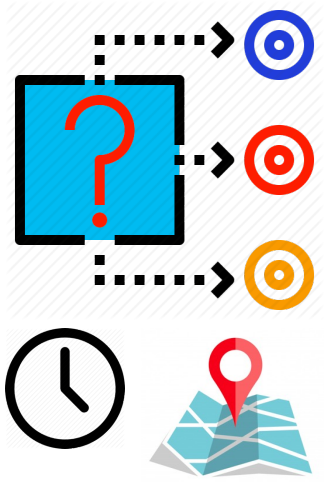
Stores Opened This Year
BY OPEN MONTH, CHAIN


This Year's Sales
BY POSTAL CODE, STORE TYPE


This Year's Sales
BY CITY, CHAIN


Sales Per Sq Ft
BY NAME



2025 Gartner® Magic Quadrant™ for Analytics and Business Intelligence Platforms

Power BI

tableau

amazon QuickSight

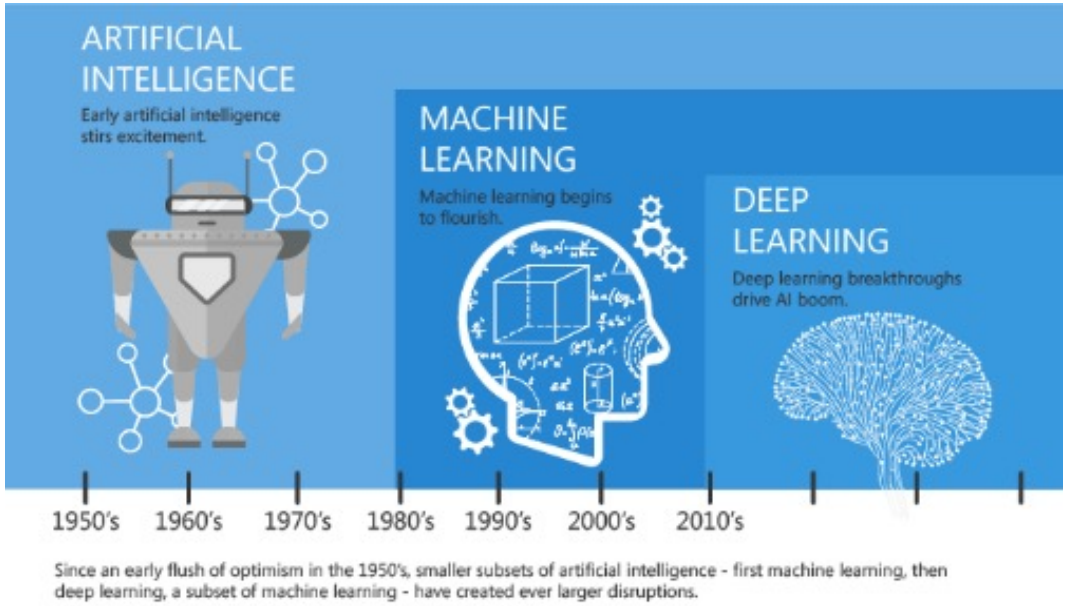
https://www.qlik.com/us/gartner-magic-quadrant-business-intelligence

Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms

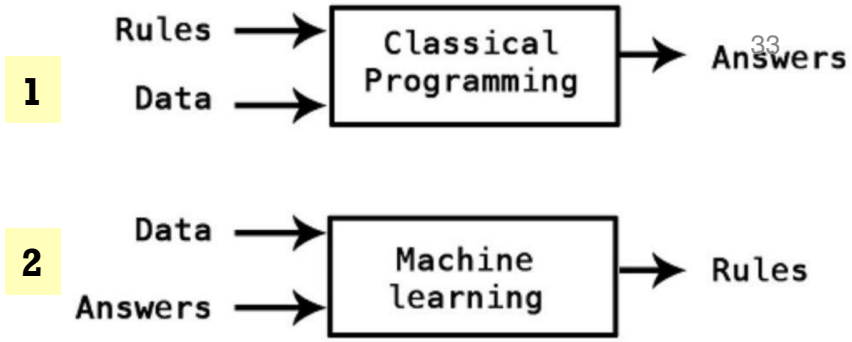
Gartner

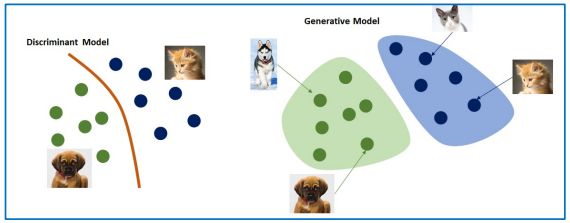
AI = Automation

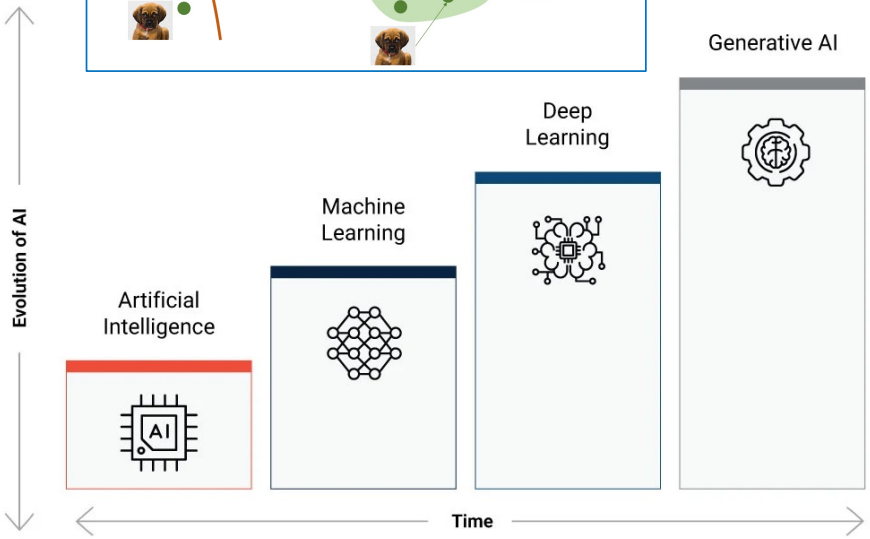
- 1) Rule-based AI
- 2) Machine Learning (ML)



Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.



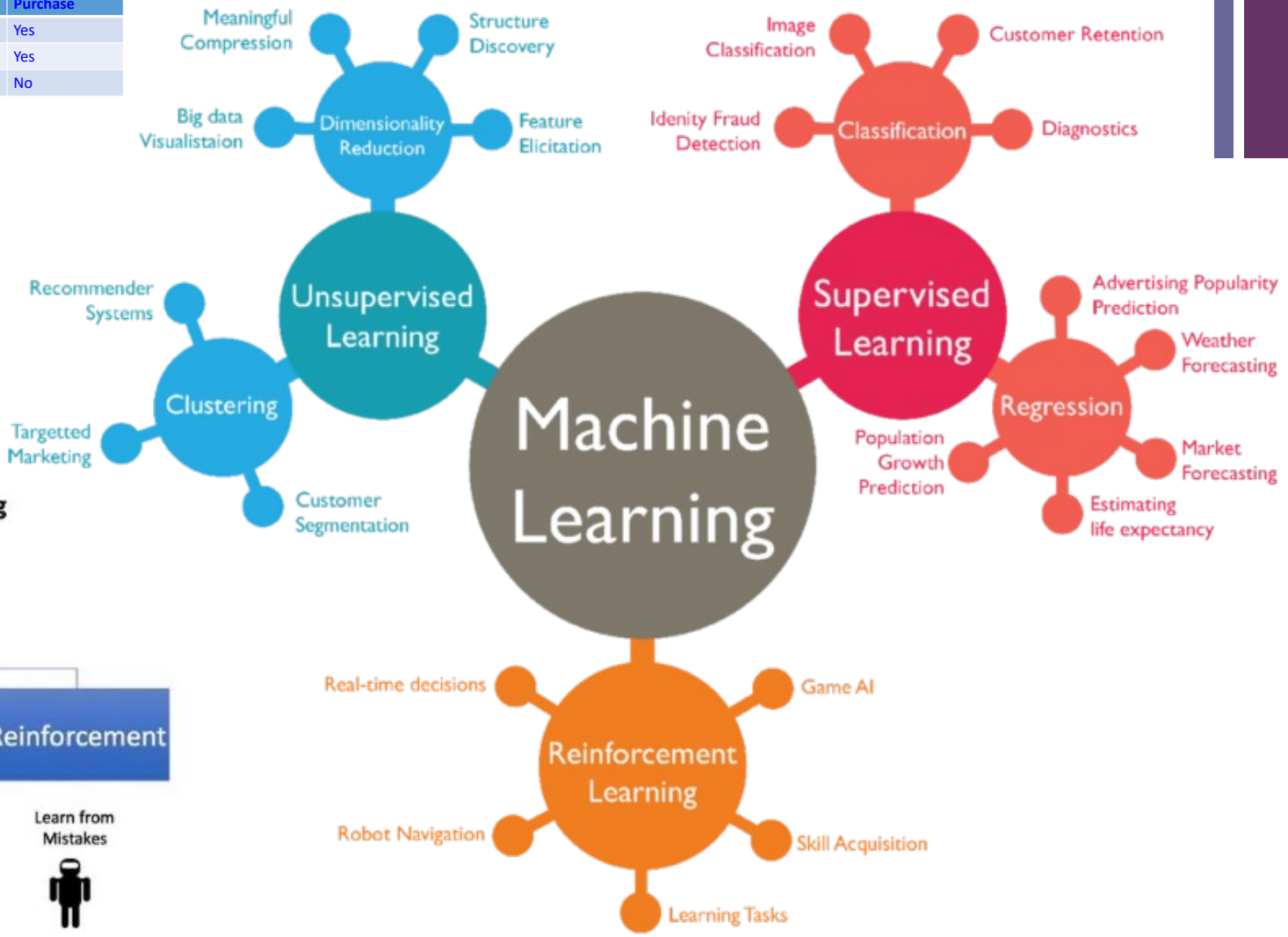


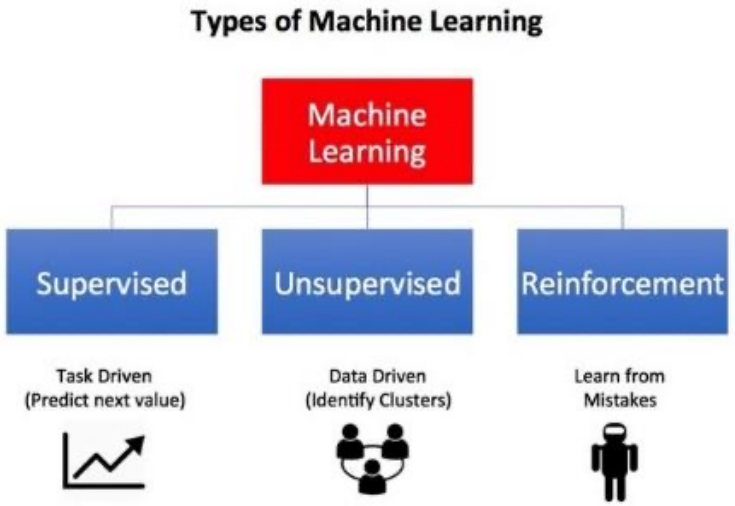


https://mc.ai/machine-learning-basics-artificial-intelligence-machine-learning-and-deep-learning/

Machine Learning (ML)

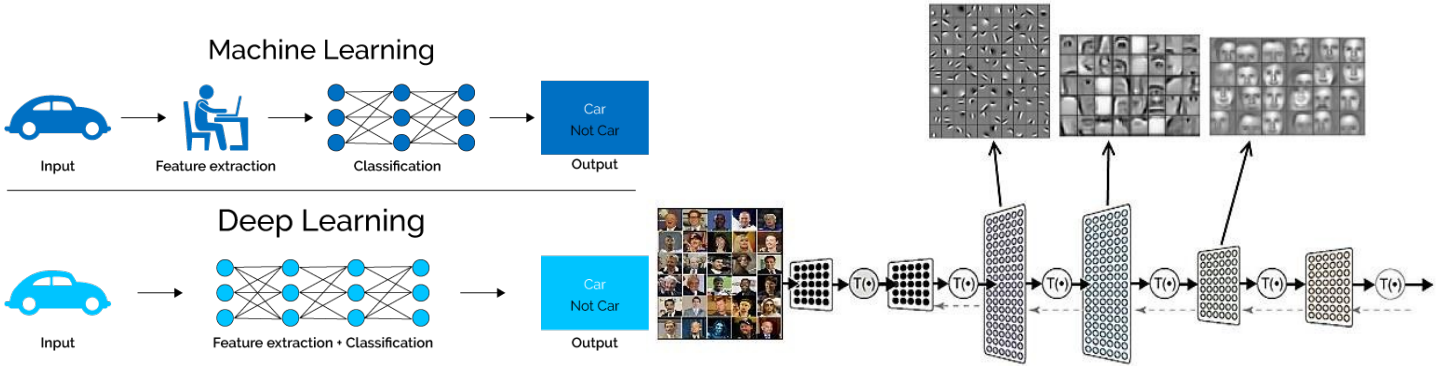
Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes
32	35,000	Male	Bangkok	No





CHULA ENGINEERING COMPUTER
Foundation toward Innovation

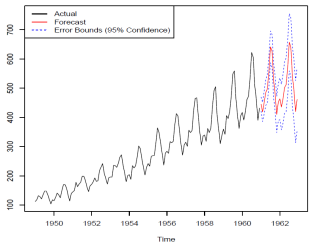
Deep Learning (DL)

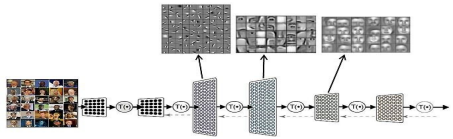


Speech Recognition

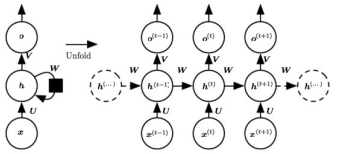
Computer Vision

Natural Language Processing





CNN



RNN (LSTM)

English | 简体中文

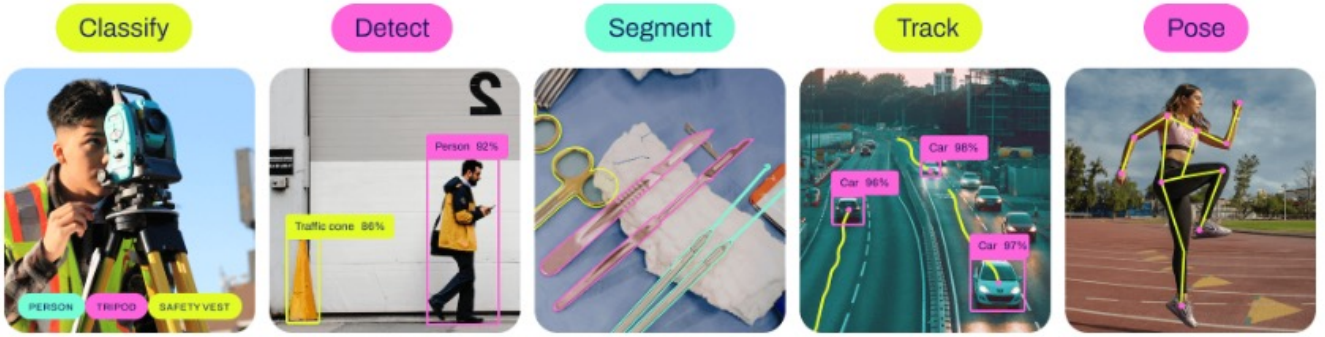
Ultralytics CI passing | codecov 87% | DOI 10.5281/zenodo.7347926 | docker pulls 22k

Run on Gradient | Open in Colab | Open in Kaggle

Ultralytics YOLOv8 is a cutting-edge, state-of-the-art (SOTA) model that builds upon the success of previous YOLO versions and introduces new features and improvements to further boost performance and flexibility. YOLOv8 is designed to be fast, accurate, and easy to use, making it an excellent choice for a wide range of object detection and tracking, instance segmentation, image classification and pose estimation tasks.

Models

YOLOv8 Detect, Segment and Pose models pretrained on the COCO dataset are available here, as well as YOLOv8 Classify models pretrained on the ImageNet dataset. Track mode is available for all Detect, Segment and Pose models.



All Models download automatically from the latest Ultralytics release on first use.

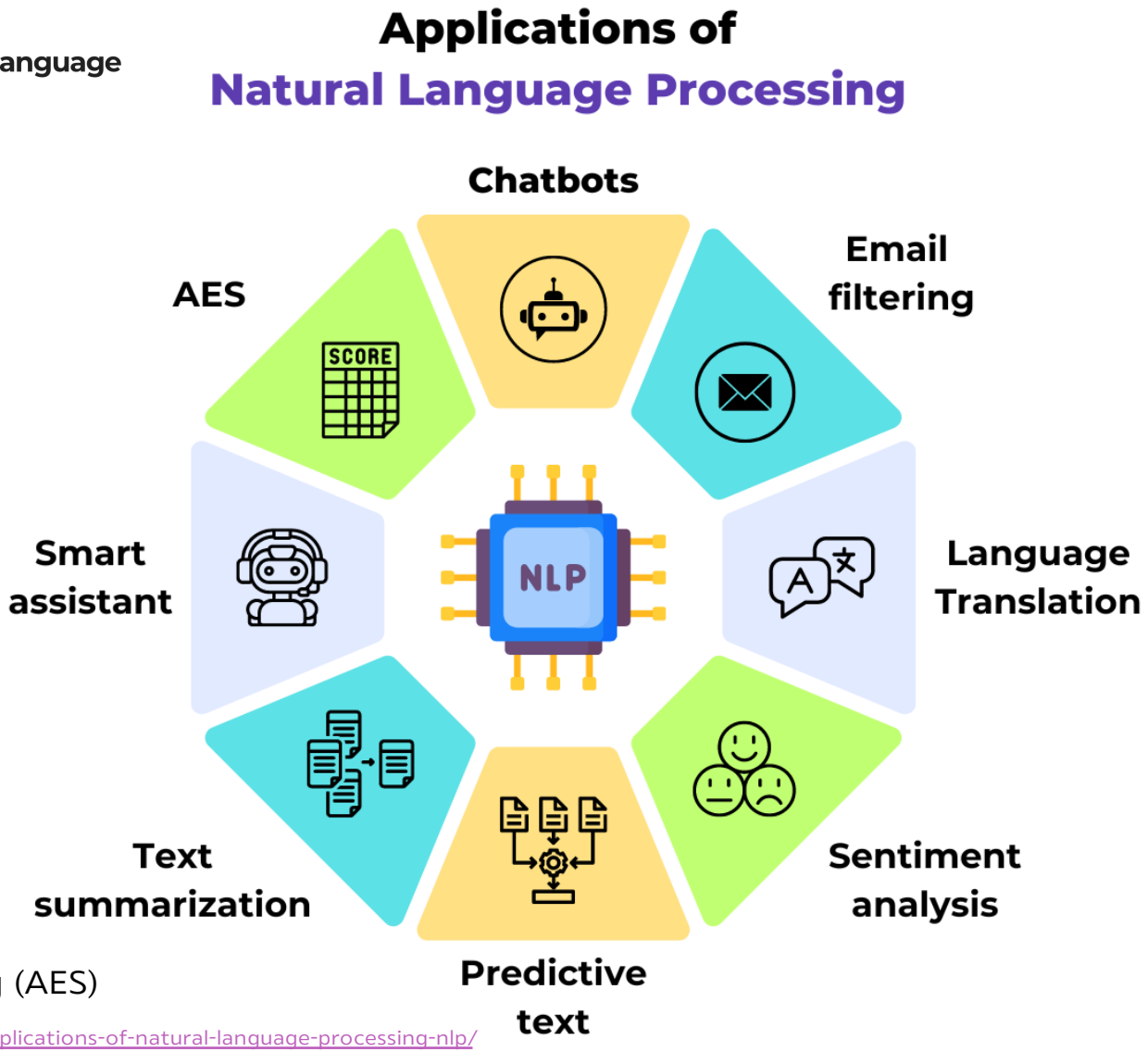
https://github.com/ultralytics/ultralytics

Home › AI

January 26, 2024

Top 8 Applications of Natural Language Processing (NLP)

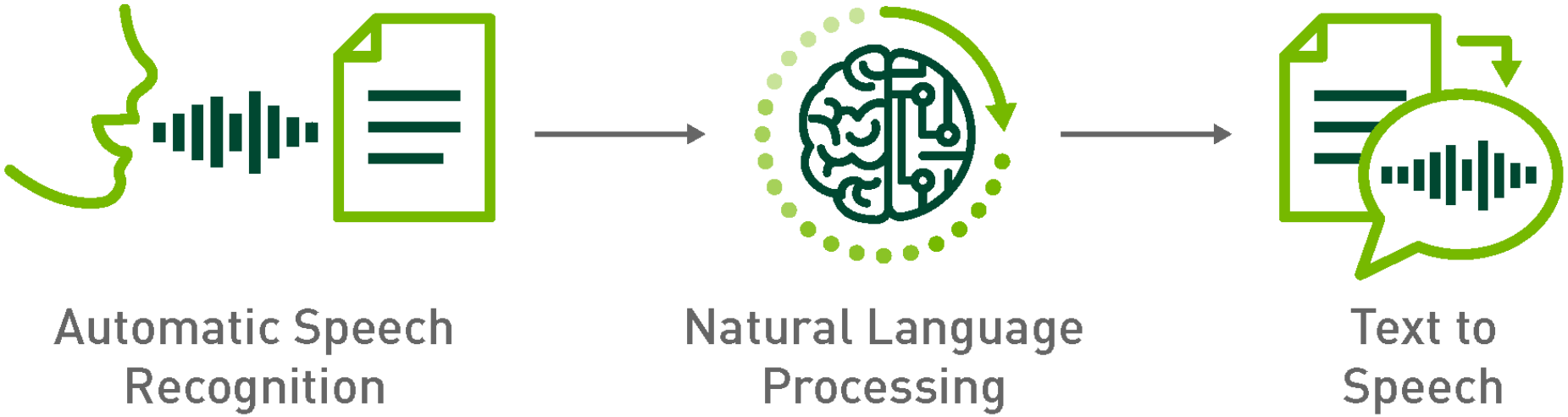
NLP



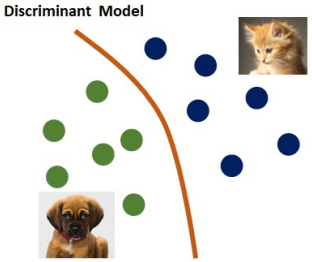
Automated Essay Scoring (AES)

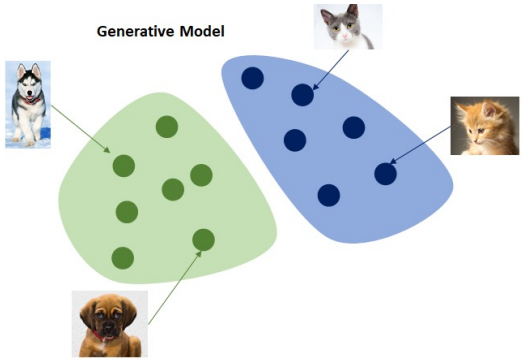
https://eastgate-software.com/top-8-applications-of-natural-language-processing-nlp/

Automatic Speech Recognition (ASR) & Text-to-Speech (TTS)

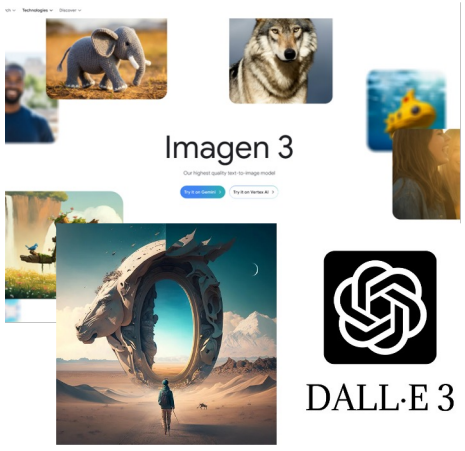


Generative AI





1) Image



2) Text

Claude

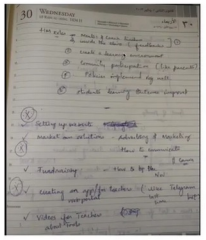
Gemini

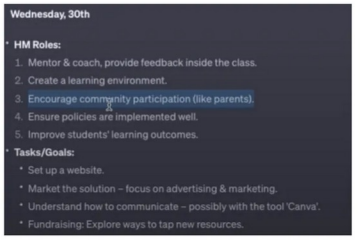
deepseek

3) Video

Sora

4) Multimodal





Types of Data Science Projects (recap)

If we don't have data, can we still perform these tasks?

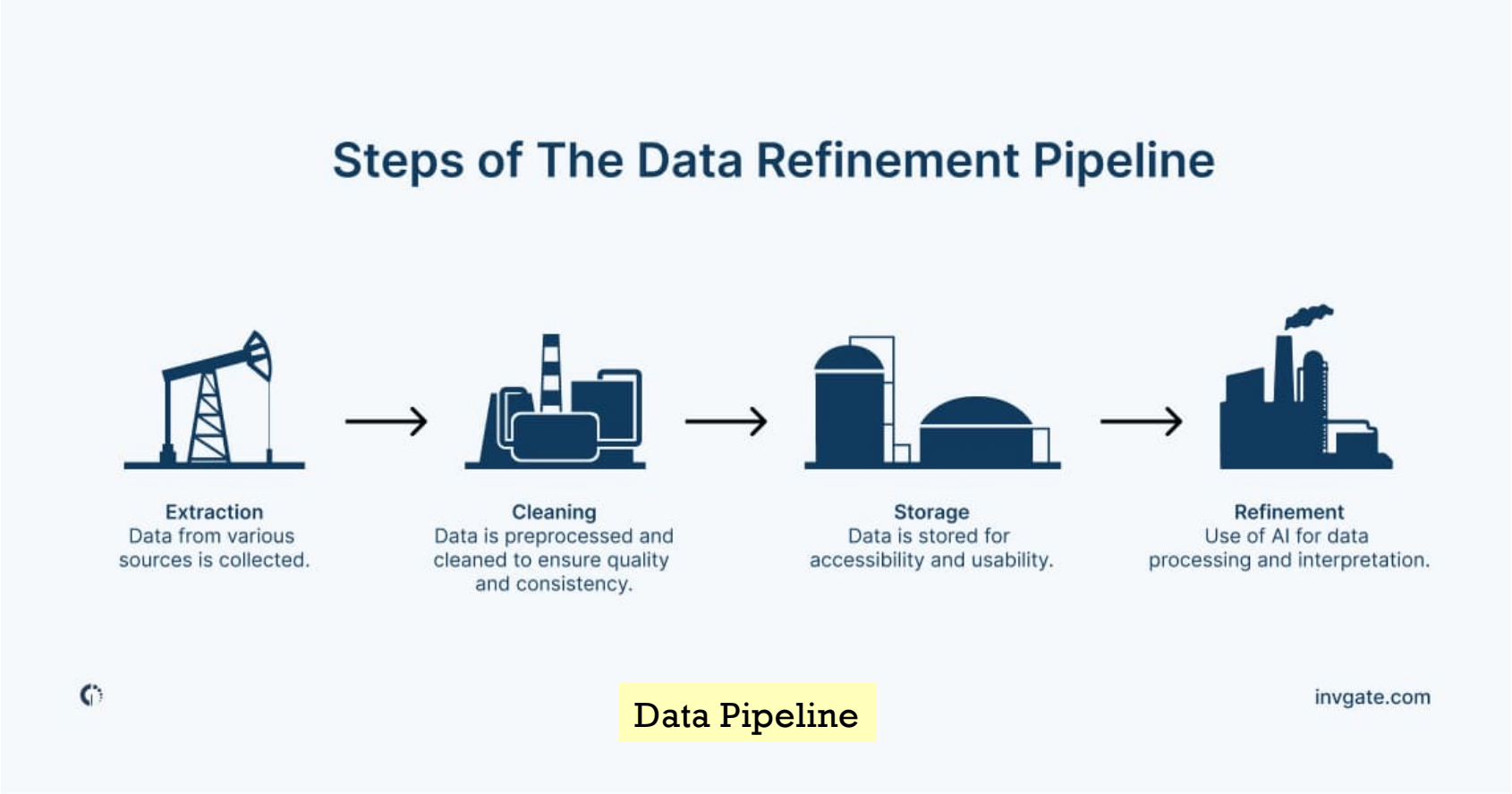
Valuable insights

- Data visualization
- Analytical skills & storytelling
  - Infographic

Advanced analytics

- AI/Machine Learning/Deep Learning
  - Prediction, Forecasting, Clustering, etc.

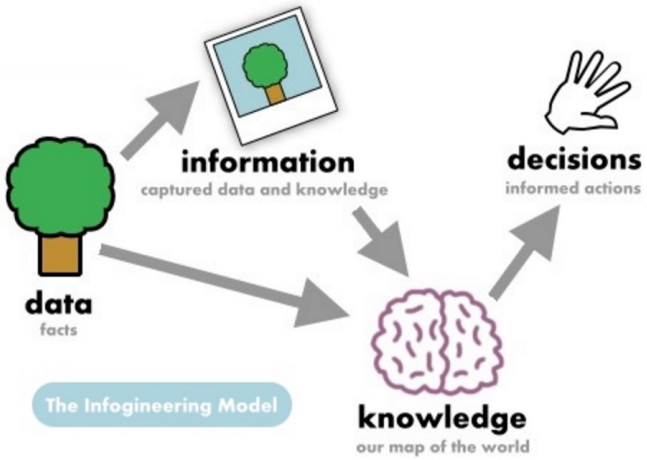
Data engineering: data acquisition (e.g., SQL, web scraping), cleansing, and storing

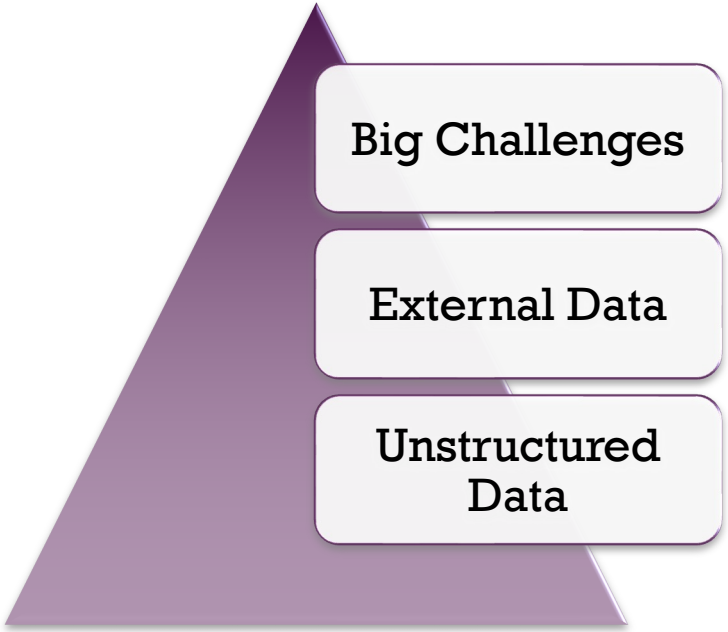


Data Pipeline

Big Data Analytics

- It is a process of examining Big Data to uncover useful information and knowledge.
- More data means better decision!

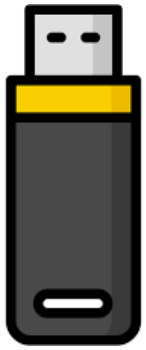




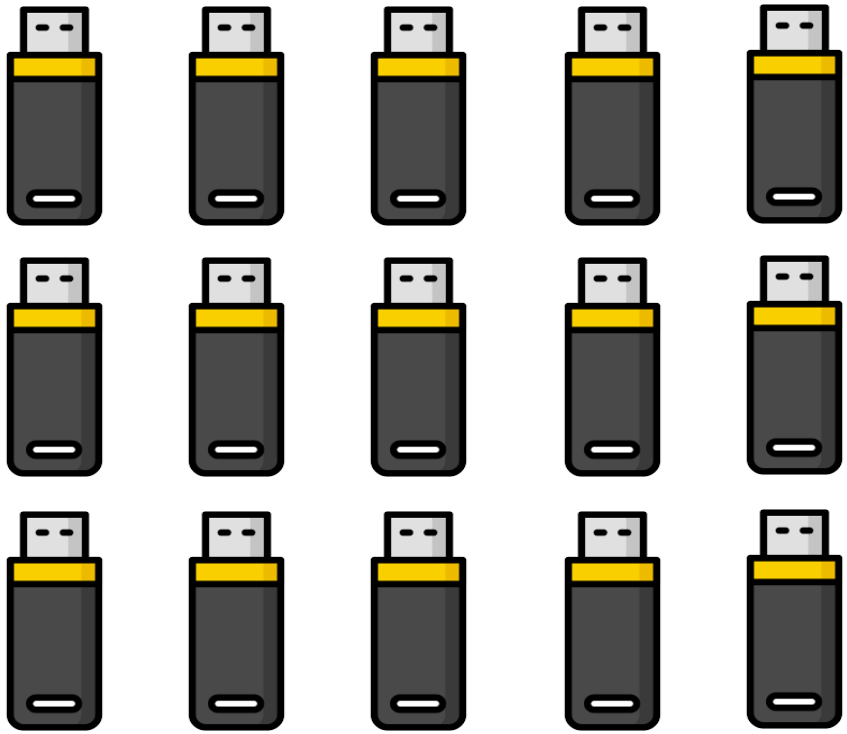
Big Data Challenges

Same tasks, but much more difficult!

2MB



200TB



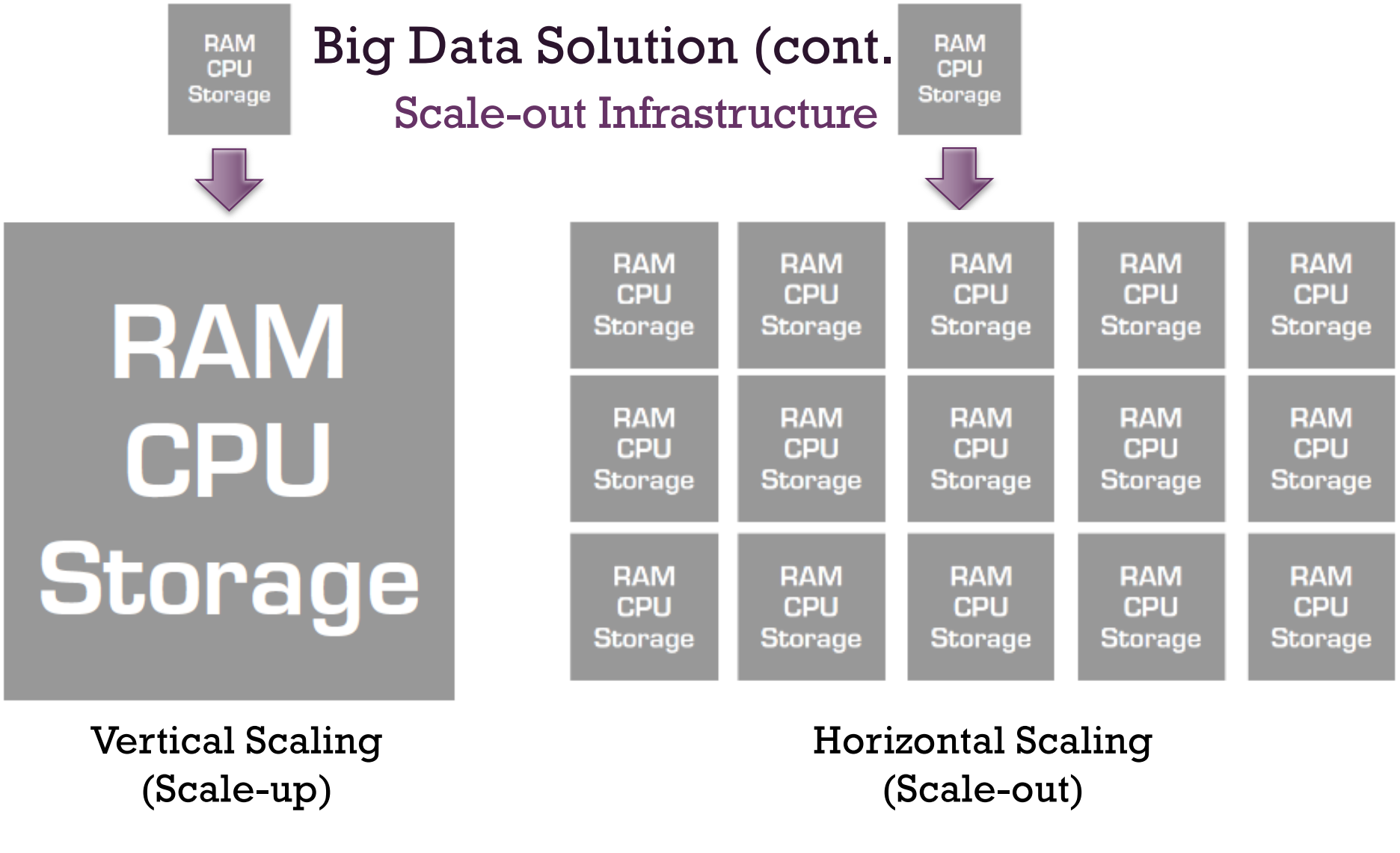
Big Data Solution

INFRASTRUCTURE

ALGORITHM

Big Data Solution (cont.

Scale-out Infrastructure

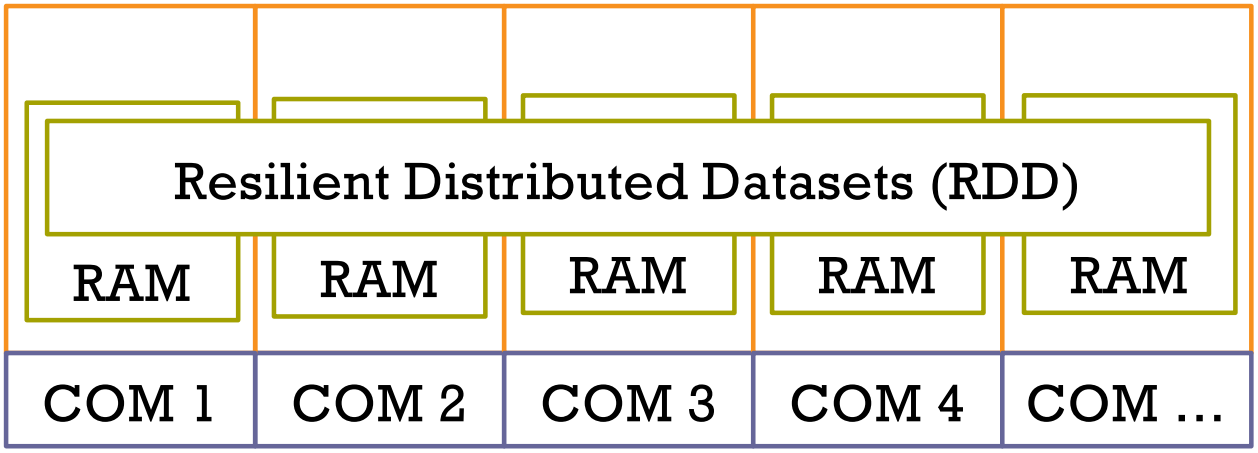


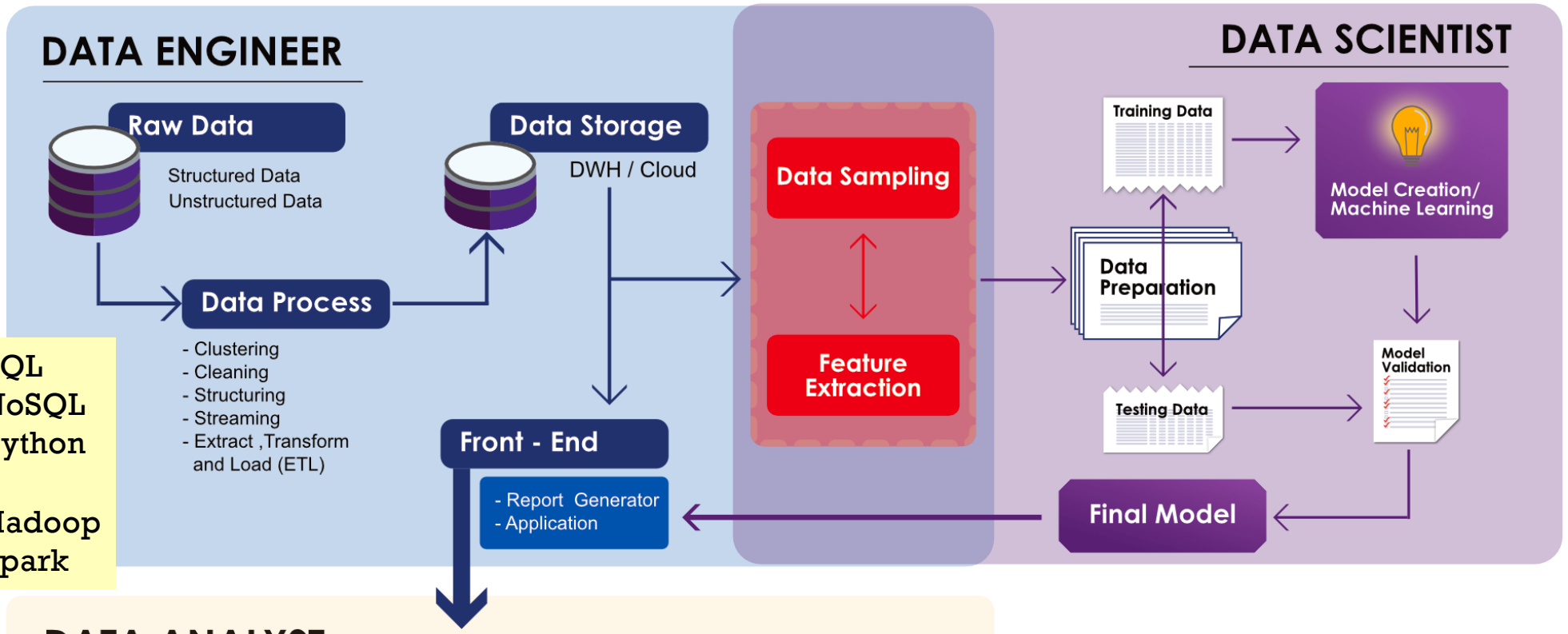
Vertical Scaling
(Scale-up)

Horizontal Scaling
(Scale-out)

Big Data Solution (cont.)

In-memory & Distributed Computing



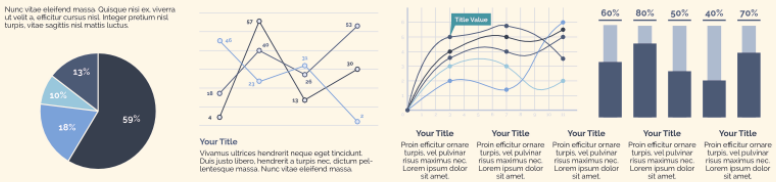


SQL
NoSQL
Python

Hadoop
Spark

DATA ANALYST

Data Visualization
Bi
Application
Report
Action



BigData Ecosystem

LINK

https://blog.datath.com/data-engineer-guide/

Top Chef Thailand ตอนสุดท้าย ที่ผู้เข้าแข่งขันต้องช่วยกันทำงานเป็นทีม – ขอบคุณรูปจาก one31

Data Engineer ก็เหมือนกับผู้ช่วยเชฟ มีหน้าที่จัดเตรียมข้อมูลจากแหล่งต่าง ๆ มารวมกันไว้ในจุดเดียว โดยต้องทำให้ข้อมูลมีความถูกต้อง และดูแลระบบว่าทำงานได้ไม่เกิดปัญหาอะไร (ในชีวิตจริงนี่ต่อให้เราวางแผนมาดีแค่ไหน เจอข้อมูลเยอะ ๆ วันดีคืนดีก็ล่มได้ครับ T_T)

We have the data and the ML model.

So, are we all set? Or is there more to consider?

Do you think that user can use this code to get the prediction result? NO!!!

2_Linear-Regression-v2.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

```
[ ]  1 coeff_df = pd.DataFrame(lm.coef_,lm.feature_names_in_,columns=['Coefficient'])
     2 coeff_df
```

Does this make sense? Probably not because I made up this data. If you want real data to repeat this sort of analysis, check out the boston dataset:

```
from sklearn.datasets import load_boston
boston = load_boston()
print(boston.DESCR)
boston_df = boston.data
```

Predictions from our Model

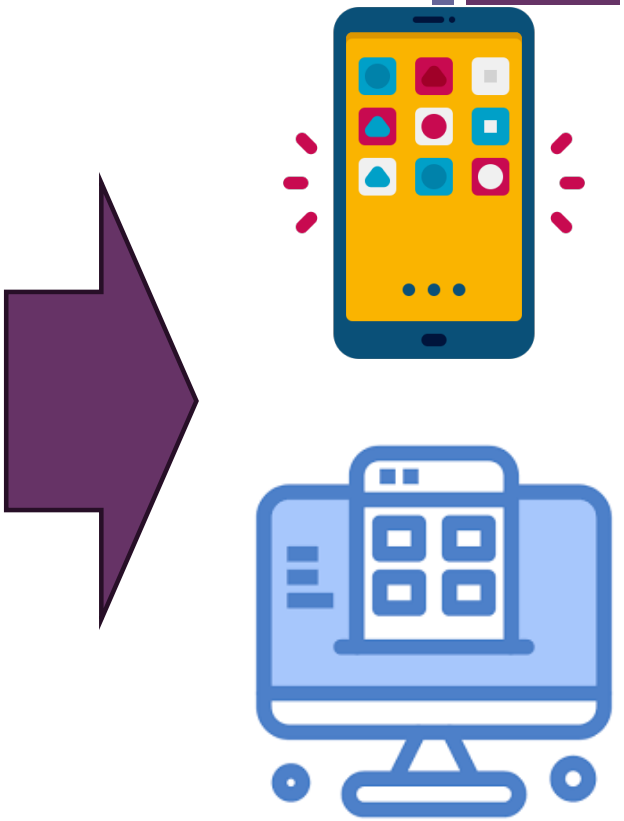
Let's grab predictions off our test set and see how well it did!

```
[ ]  1 predictions = lm.predict(X_test)
```

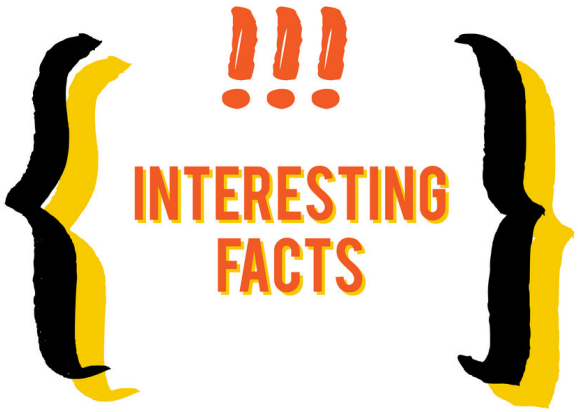
```
[ ]  1 plt.scatter(y_test,predictions)
```

Residual Histogram

```
[ ]  1 sns.distplot((y_test-predictions),bins=50);
```



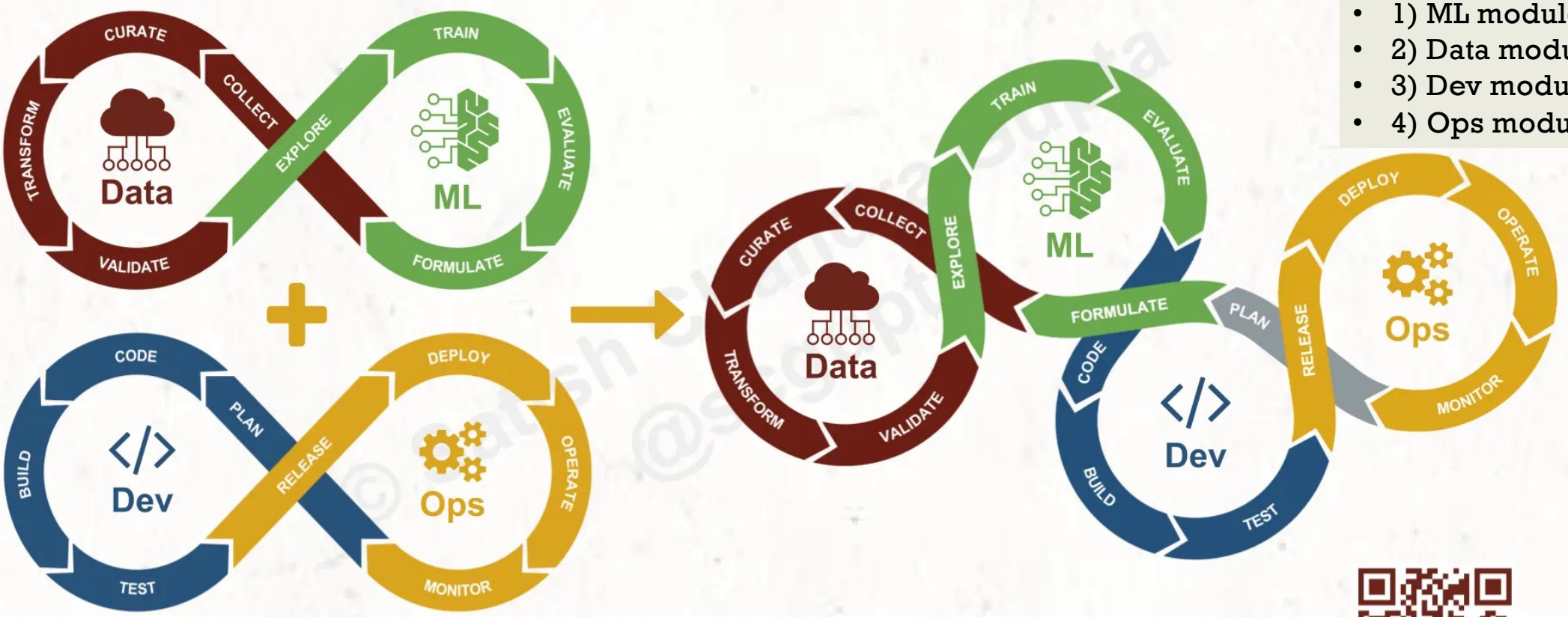
Interesting facts



- More than 50% of AI projects were FAILED since they didn't plan about the deployment.
- For the AI project, finish building the model is only 50% of the work.
- The remaining work is about the deployment as a touchpoint to the target user.

MLOps = DataML + DevOps

ml4devs.com/mlops-lifecycle



- 1) ML module
- 2) Data module
- 3) Dev module
- 4) Ops module

scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta

https://www.ml4devs.com/images/illustrations/ml-lifecycle-fusing-model-and-software-development.webp

Data Scientist + ML Engineer

Data Scientist

vs

Data Engineer

vs

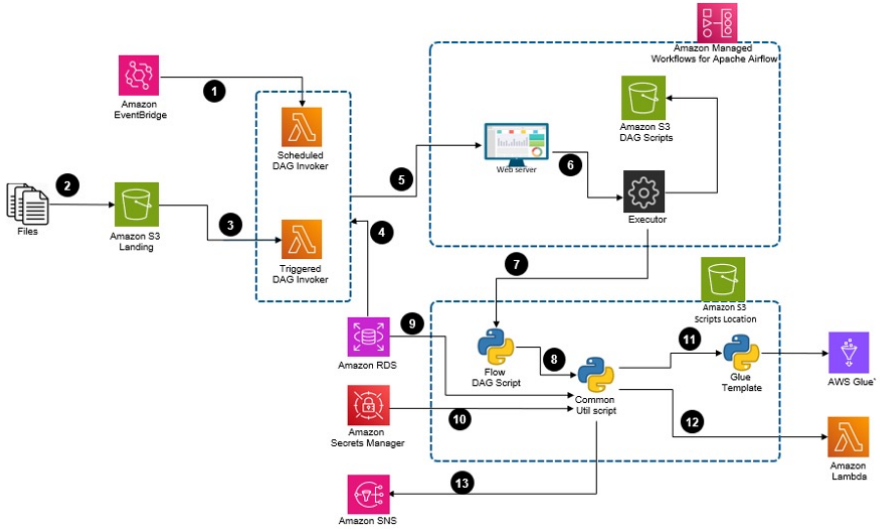
ML Engineer

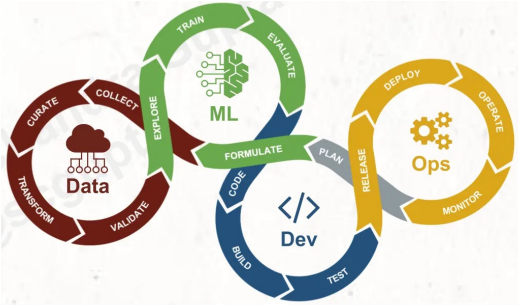
vs

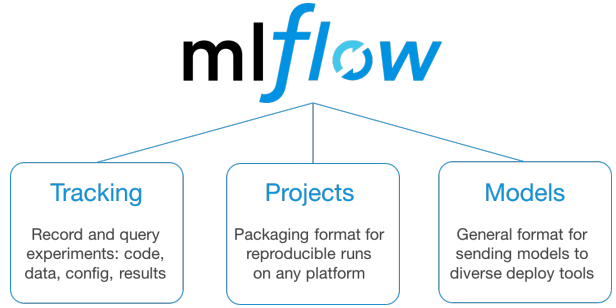
MLOps Engineer

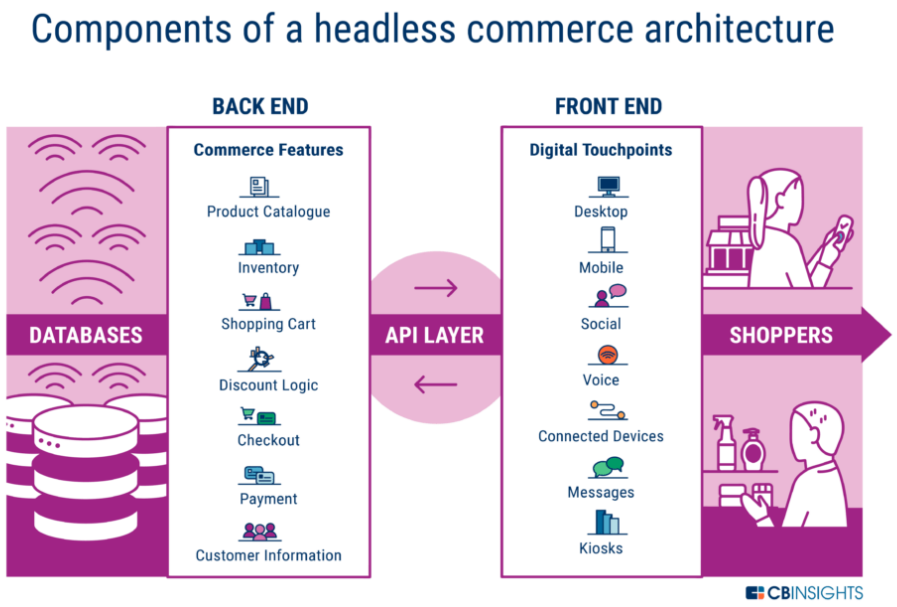
https://vocal.media/education/data-scientist-vs-data-engineer-vs-ml-engineer-vs-ml-ops-engineer

Many MLOps Tasks: Deployment, Automation, Integration, Monitoring









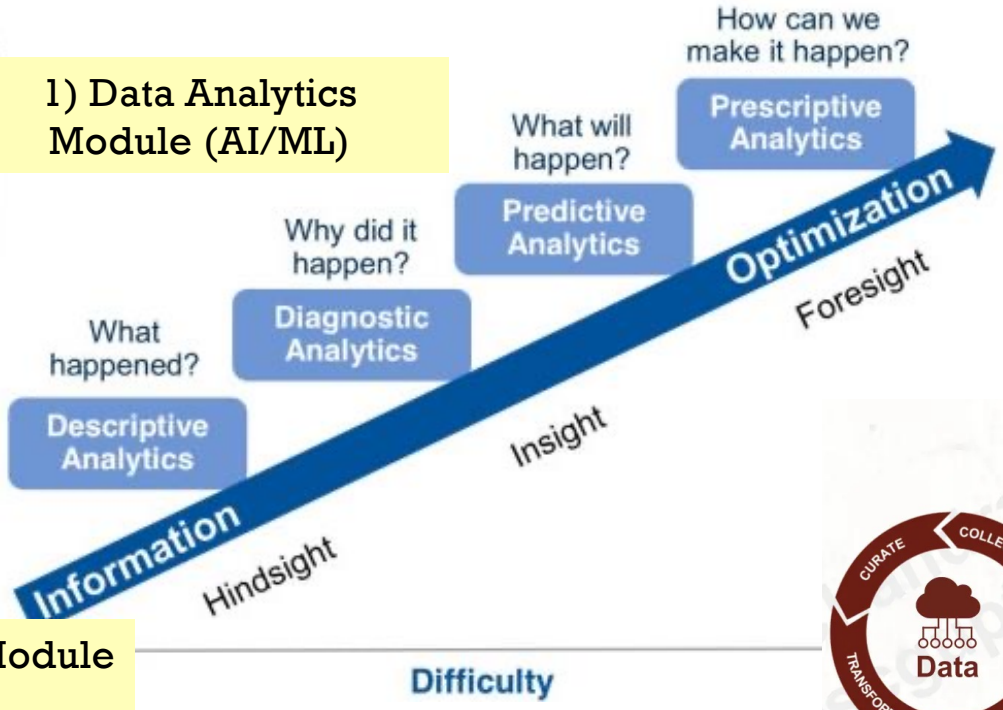
Conclusion

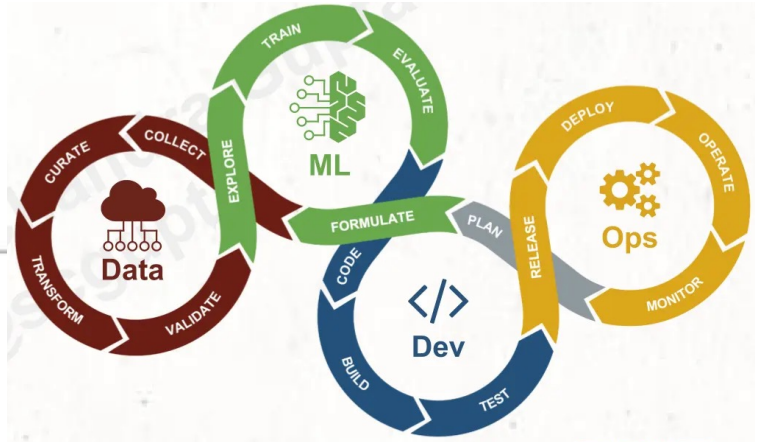
Conclusion

1) Data Analytics Module (AI/ML)

3) Data Vizualiation Module

2) Data Engineering Module (Data Pipeline)

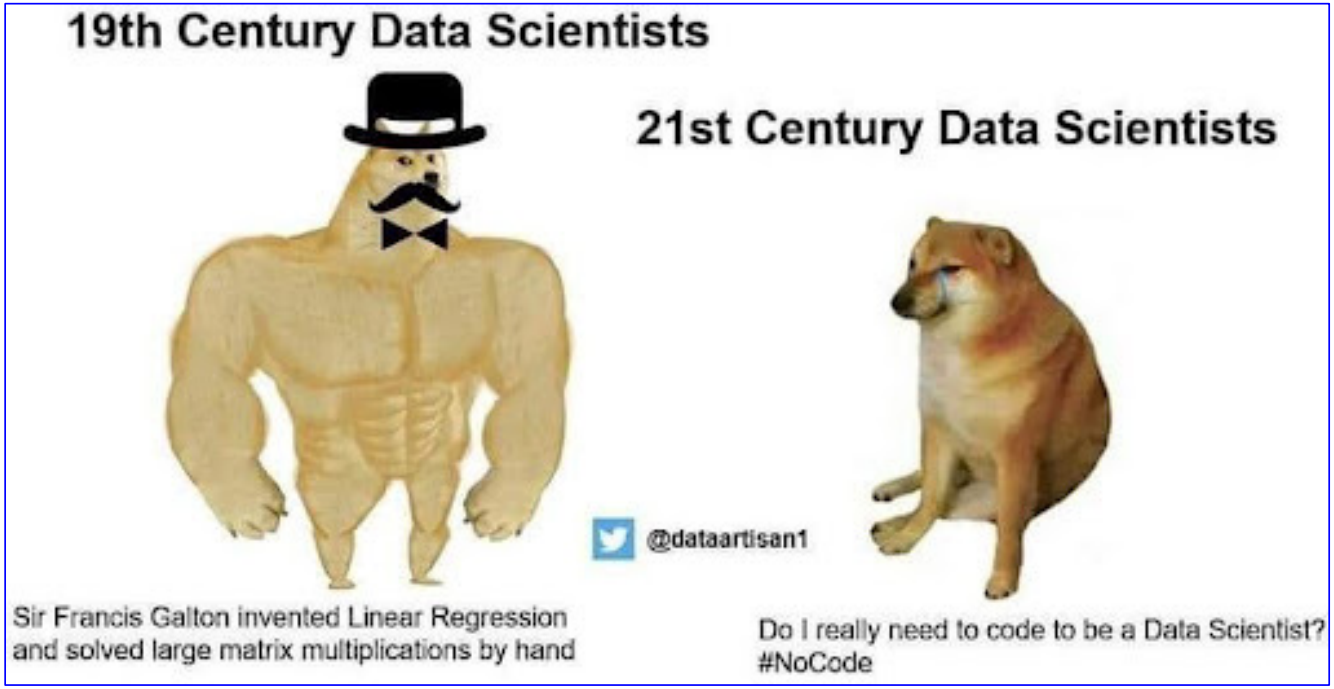




Is Data Science Still the Sexiest Job of 2025?

Muhammad Mujtaba Raza · Follow
Published in Analyst's corner · 4 min read · 4 days ago

It is more essential than ever, as data science skills have become a fundamental knowledge that everyone needs to acquire.

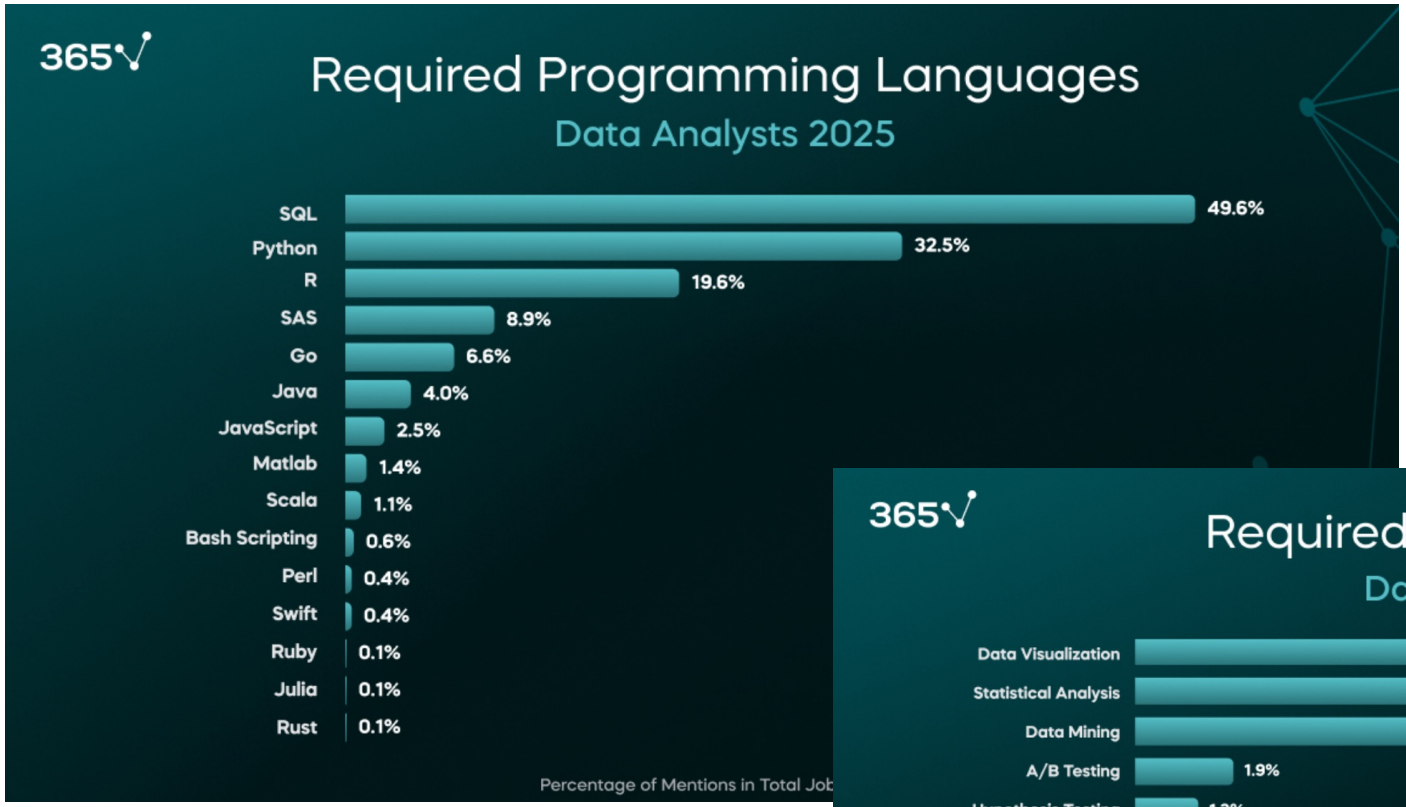


https://medium.com/analysts-corner/is-data-science-still-the-sexiest-job-of-2025-38ed9c438a34

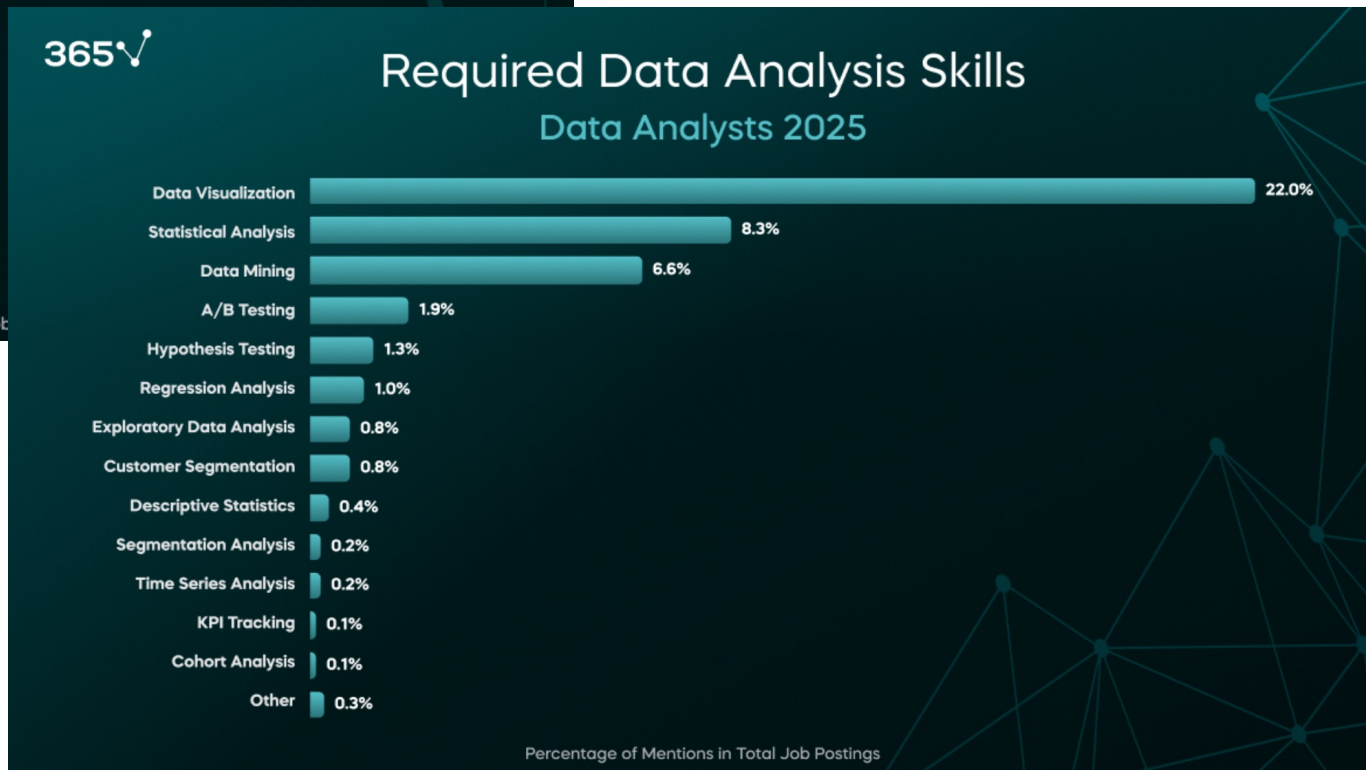
Data Analyst Job Outlook 2025 [Research on 1,000 Job Postings]

Sophie Magnet • 8 Apr 2025 • 21 min read

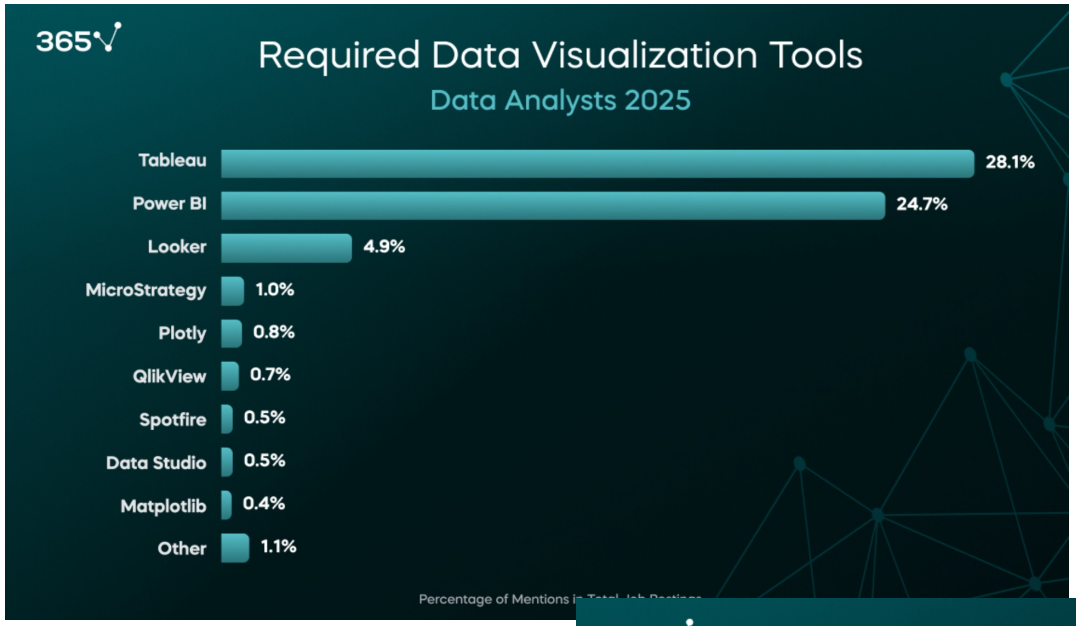
https://365datascience.com/career-advice/data-analyst-job-outlook-2025/

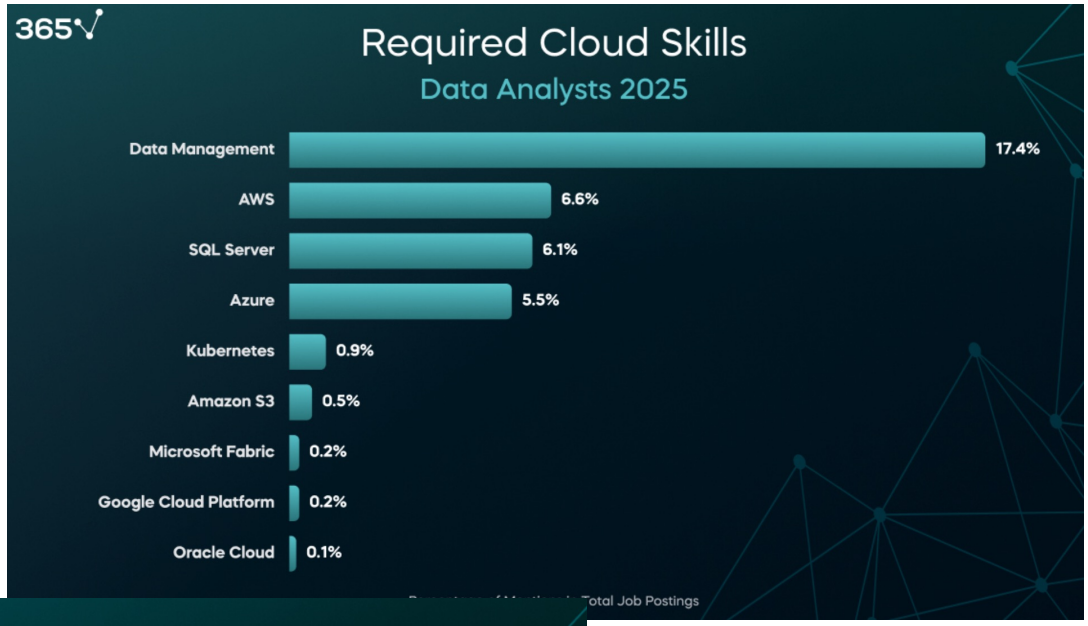


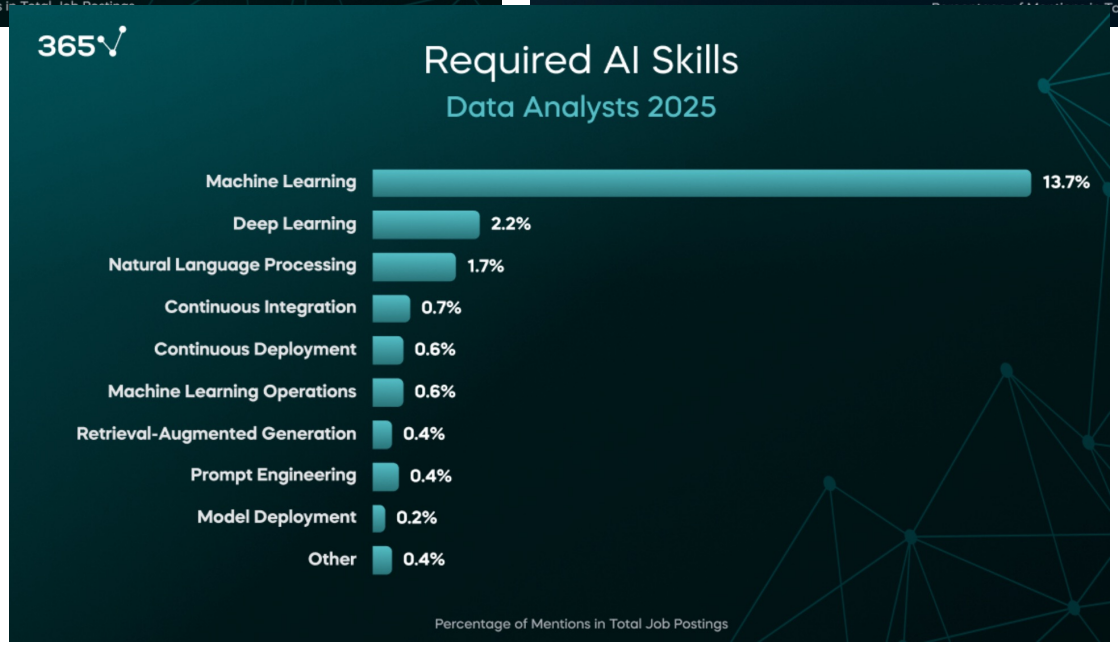
- SQL
- Python



- Data visualization
- Statistical analysis
- AI/ML
- Exploratory Data Analysis


365
Required Data Visualization Tools
Data Analysts 2025
Tableau 28.1%
Power BI 24.7%
Looker 4.9%
MicroStrategy 1.0%
Plotly 0.8%
QlikView 0.7%
Spotfire 0.5%
Data Studio 0.5%
Matplotlib 0.4%
Other 1.1%

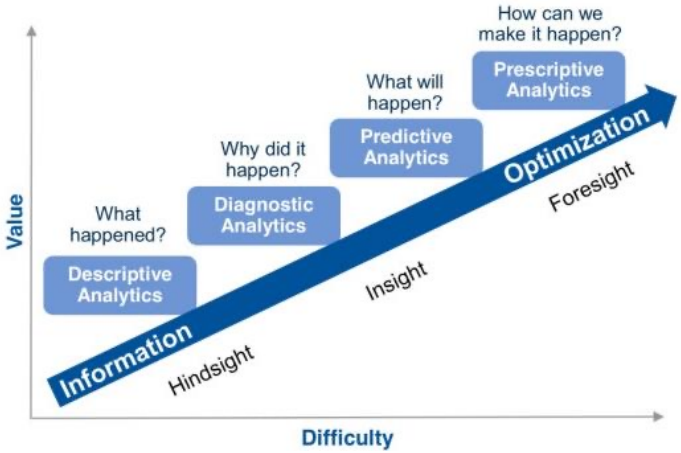

365
Required Cloud Skills
Data Analysts 2025
Data Management 17.4%
AWS 6.6%
SQL Server 6.1%
Azure 5.5%
Kubernetes 0.9%
Amazon S3 0.5%
Microsoft Fabric 0.2%
Google Cloud Platform 0.2%
Oracle Cloud 0.1%


365
Required AI Skills
Data Analysts 2025
Machine Learning 13.7%
Deep Learning 2.2%
Natural Language Processing 1.7%
Continuous Integration 0.7%
Continuous Deployment 0.6%
Machine Learning Operations 0.6%
Retrieval-Augmented Generation 0.4%
Prompt Engineering 0.4%
Model Deployment 0.2%
Other 0.4%
Percentage of Mentions in Total Job Postings

Disclaimer

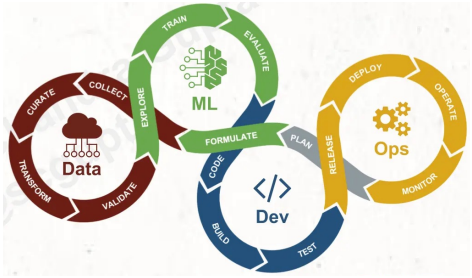
2110446: DS & DE
Practical Data Analytics & ML Pipeline

- This course focuses on:
- Fundamental data processing
- ML pipeline
- Practical process (MLOps)



- Remark1: DS != AI/ML
- We cannot cover all algorithms (NN, DL, NLP, CV, etc.) in this course!

- Remark2: DS != DE (aka. DW)
- There are a lot more DE tasks (NoSQL, ETL, etc.)



+

Any questions? ☺