Introduction to Data Science

2110446 Data Science and Data Engineering (2023/2)

CHULA ENGINEERING
Foundation toward Innovation
COMPUTER

* Part of this slide is modified from a slide of Prof.Natawut

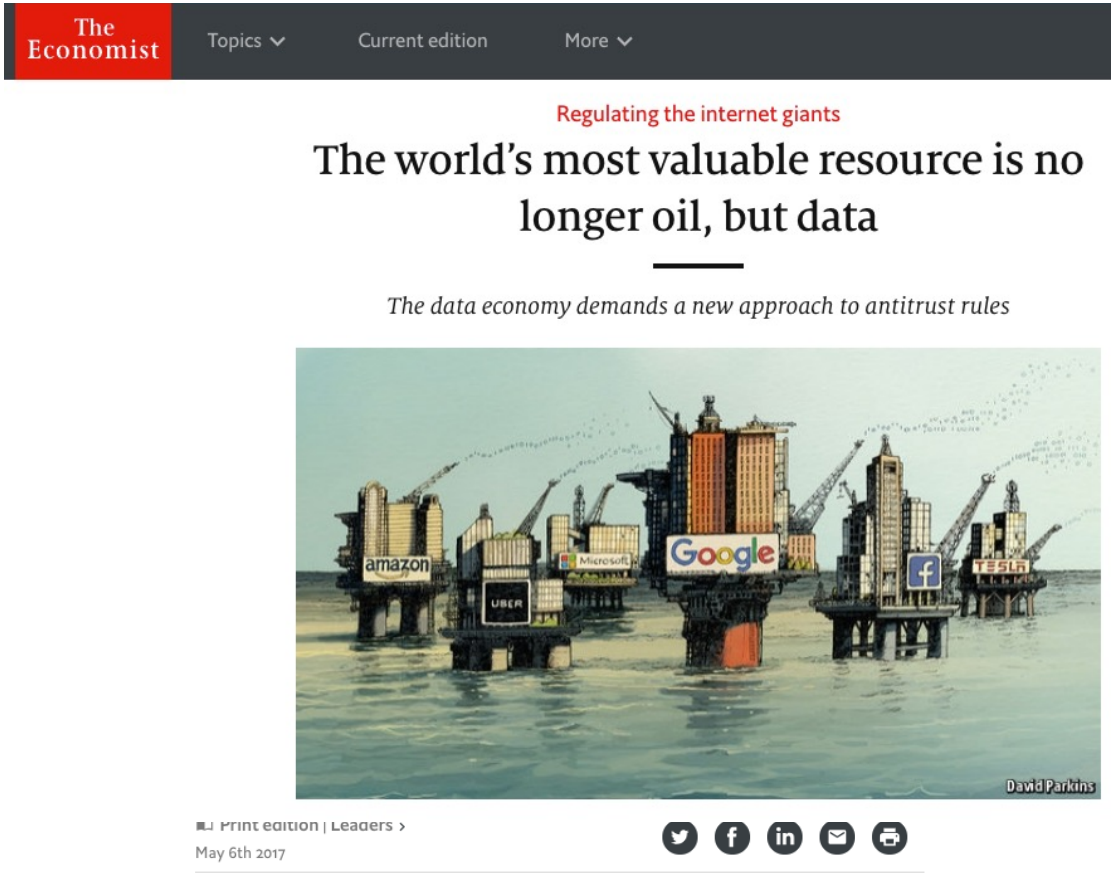
Assoc. Prof. Peerapon Vateekul, Ph.D.
Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University
Peerapon.v@chula.ac.th
www.cp.eng.chula.ac.th/~peerapon/

Outline

- Introduction
  - Data is important
  - Data Science Definition by Dr.Virote
  - Data Science Definition by Aj.Natawut
- Big Data
- Data Science Process & Data Science Trend

Introduction

Data is important (in 2017)



- Alphabet (Google's parent company), Amazon, Apple, Facebook and Microsoft
- $25bn in net profit in the first quarter of 2017
- Amazon captures half of all dollars spent online in America.
- Google and Facebook accounted for almost all the revenue growth in digital advertising in America last year

https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Data is important (in 2018)! (cont.)

The New Oil

Jennifer Presley Executive Editor, E&P Magazine Hart Energy Thursday, November 1, 2018 - 6:40am

With a number of successful projects under its collective belt, the oil and gas industry is proving Big Data is more than just a buzzword. (Source: Makhnach_S/*Shutterstock.com*; Design by Felicia Hammons)

https://www.epmag.com/new-oil-1720651

Data Science (AI,ML,DM) + Big Data

Who analyzes these data!

Harvard Business Review

ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

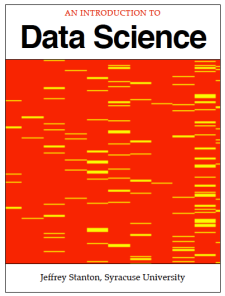
by Thomas H. Davenport and D.J. Patil

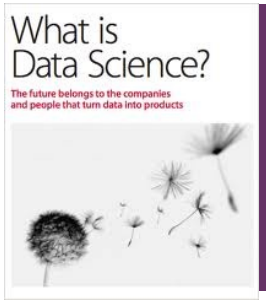
FROM THE OCTOBER 2012 ISSUE

WHAT TO READ NEXT

Competing on Analytics

What is Data Science?

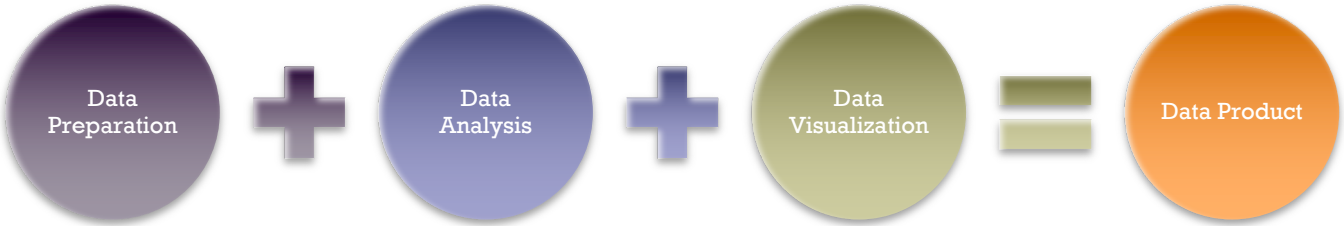




- Data
  - Facts and statistics collected for reference or analysis

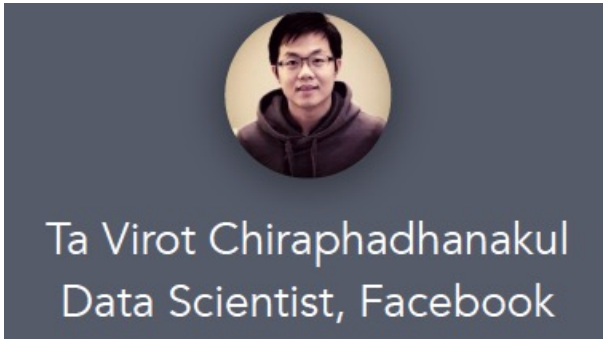
- Science
  - A systematic study through observation and experiment

- 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



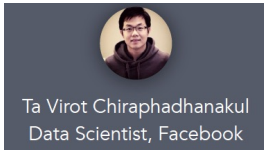
Code Mania 2 (01), Jan-2015

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.)



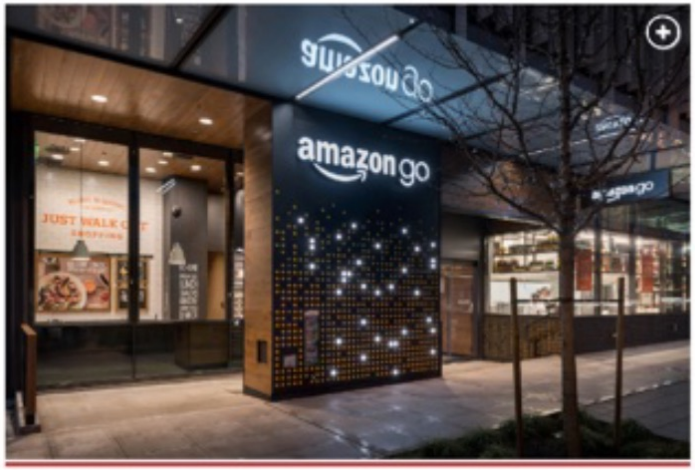
Ta Virot Chiraphadhanakul
Data Scientist, Facebook
Code Mania 2, Jan-2015

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



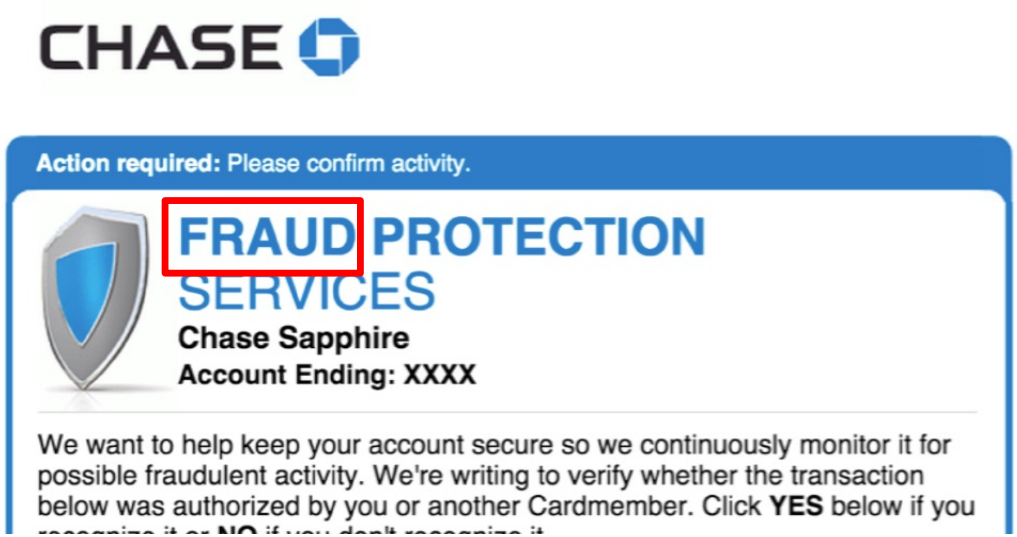
The Amazon Go storefront

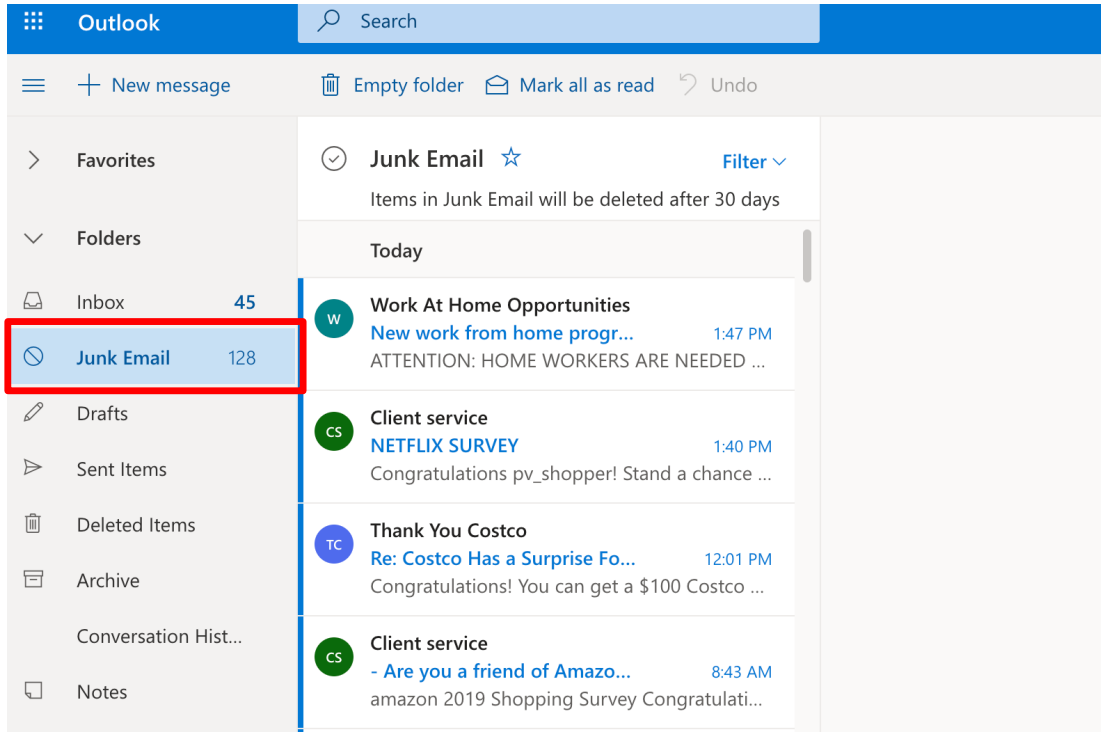
NEW YORK POST POETICS

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

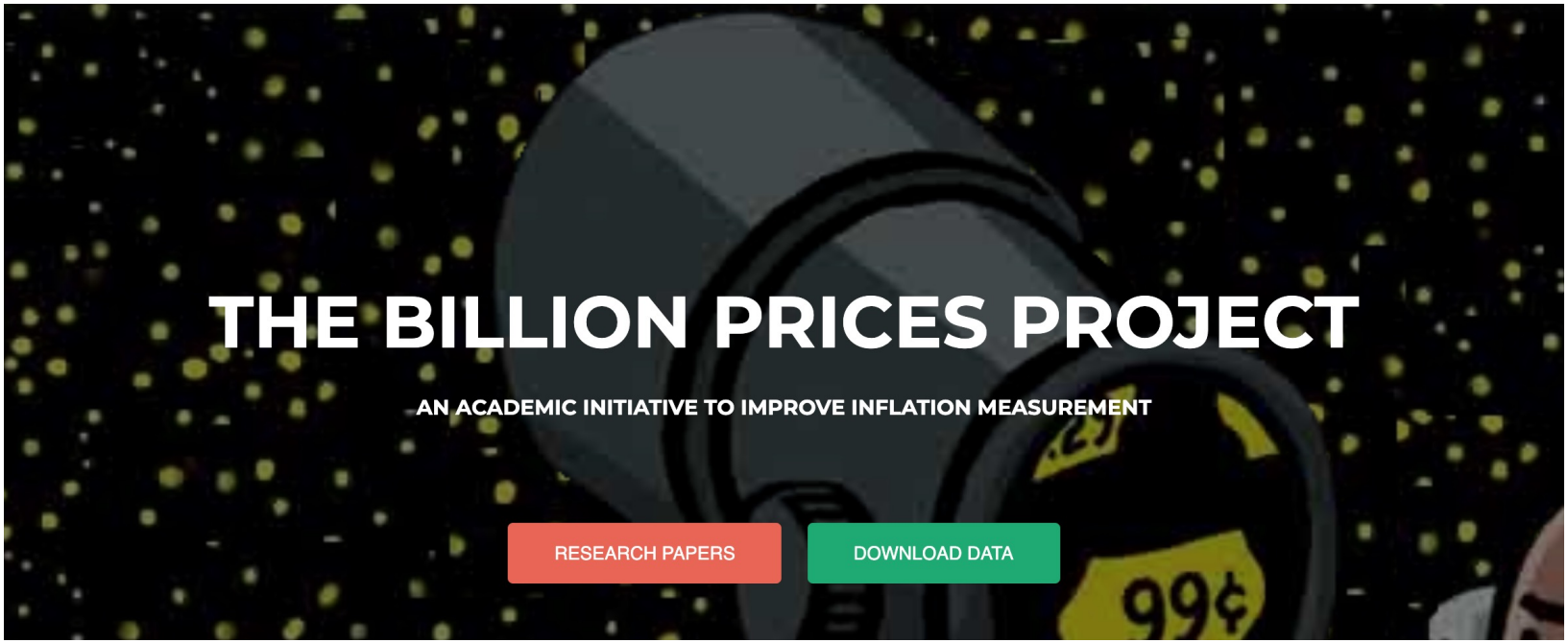
http://nypost.com/2016/12/05/amazon-introduces-next-major-job-killer-to-face-americans/

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 ก.พ. 61
แลกซื้อ 7 บาท
ถูกสุดสุดไปเลย
เติมเงินฟินเว่อร์ ใช้คูปอง ลดทันที
10 บาท
สะดวก ใกล้บ้านคุณ


TESCO Lotus express
ลดทั้งเดือน
วันที่ 28 ก.ย. - 29 ต.ค. 60


ท็อปส์ ได้รับการยกย่องเป็น Trusted Brand
ซึ่งเป็นสุดยอดแบรนด์ที่ผู้บริโภคให้ความเชื่อมั่นมากที่สุด
ต่อเนื่องกันถึง 5 ปี สาขาซุปเปอร์มาร์เก็ต 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

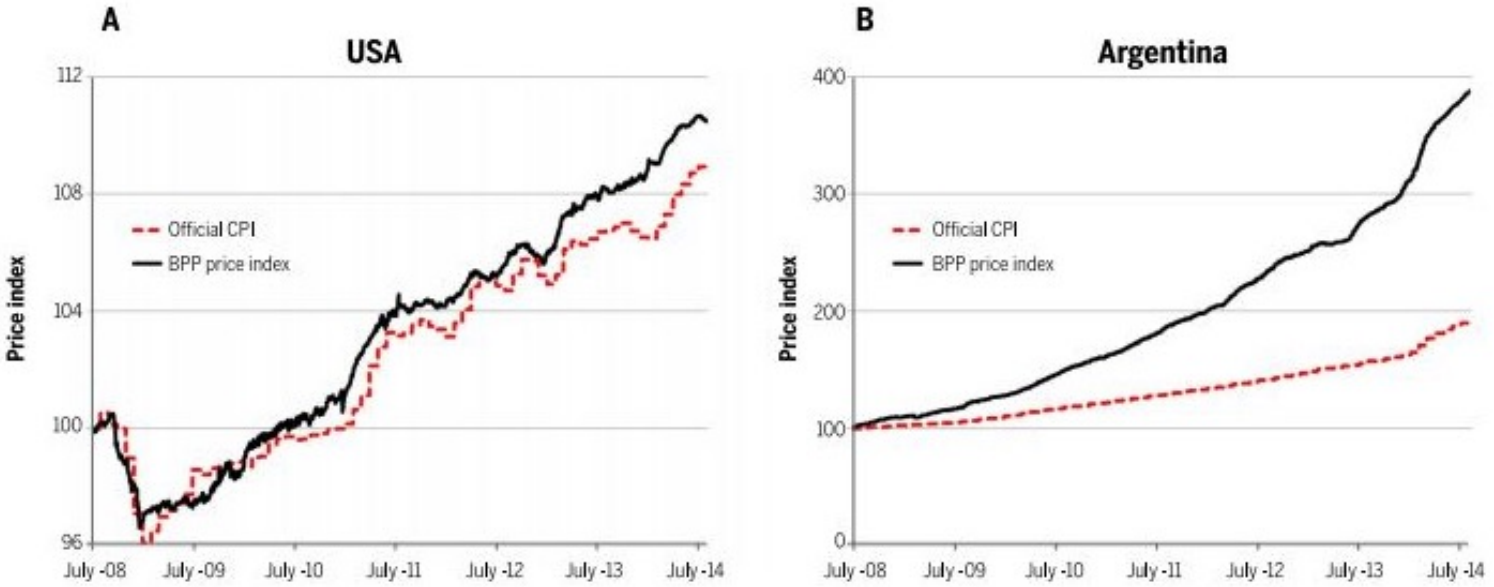
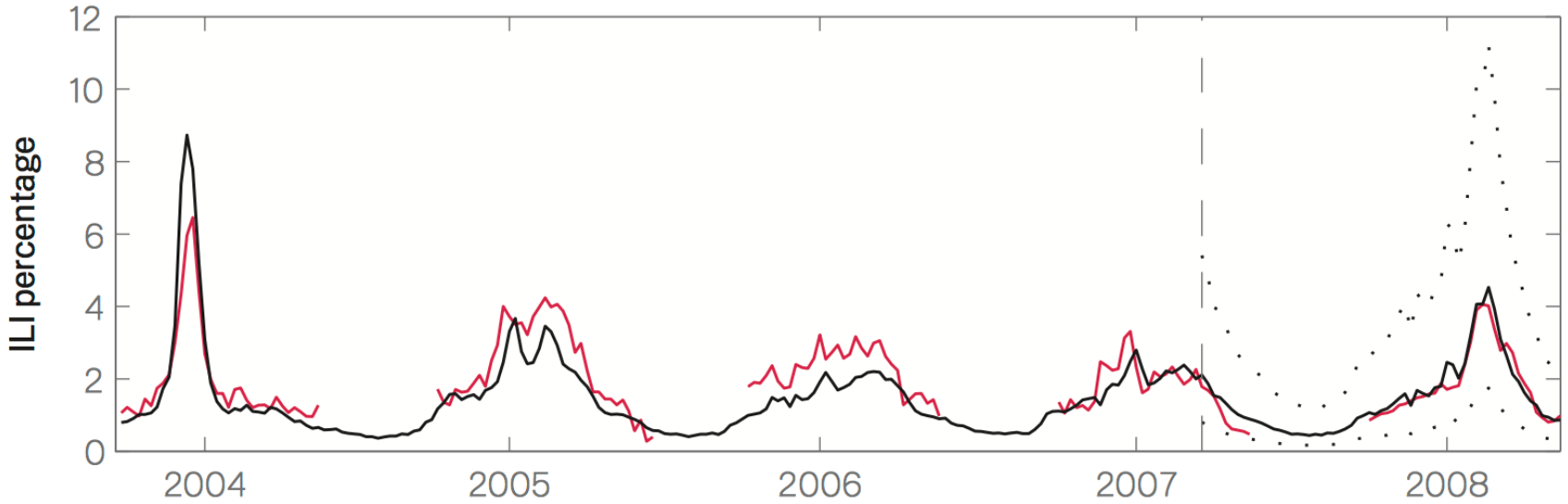


Fig. 2. BPP price index. Dashed red lines show the monthly series for the CPI in the United States (**A**) and Argentina (**B**), as published by the formal government statistics agencies. Solid black lines show the daily price index series, the "State Street's PriceStats Series" produced by the BPP, 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
  - Usually called metrics or indicators
- Direct and indirect measurement

Why do we need to measure?

- Comparison between alternatives (make a selection)
  - Choosing which notebook to buy
- Comparison after improvement or tuning
  - Should I add memory to my notebook?
- A/B Testing (split testing)
  - Let the actual users decide their preferences
  - Very popular for UI design

A/B Testing

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

Example: SimCity

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

Control

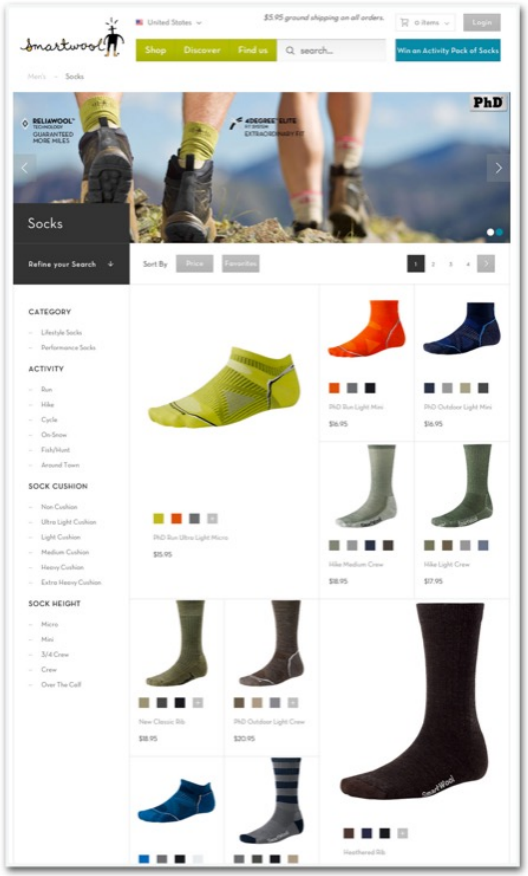
Variation

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

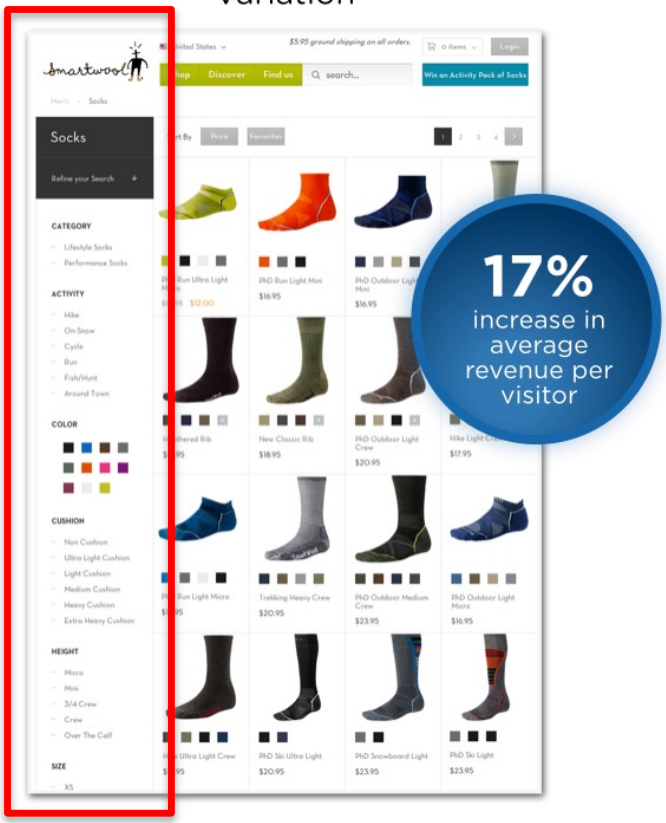
Example: SmartWool

3. 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/

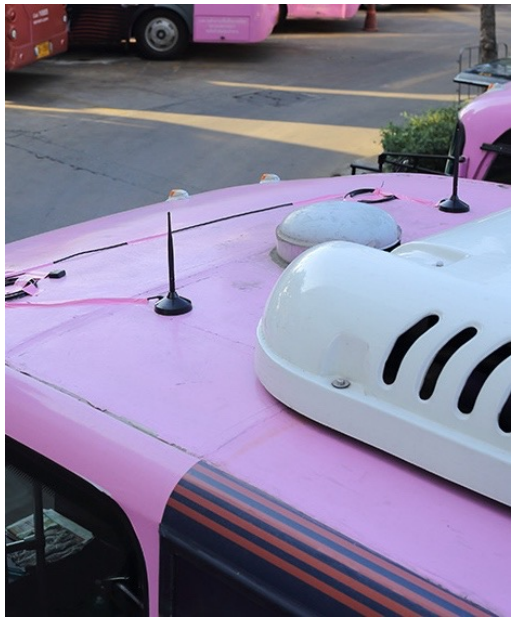
2) Insights

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

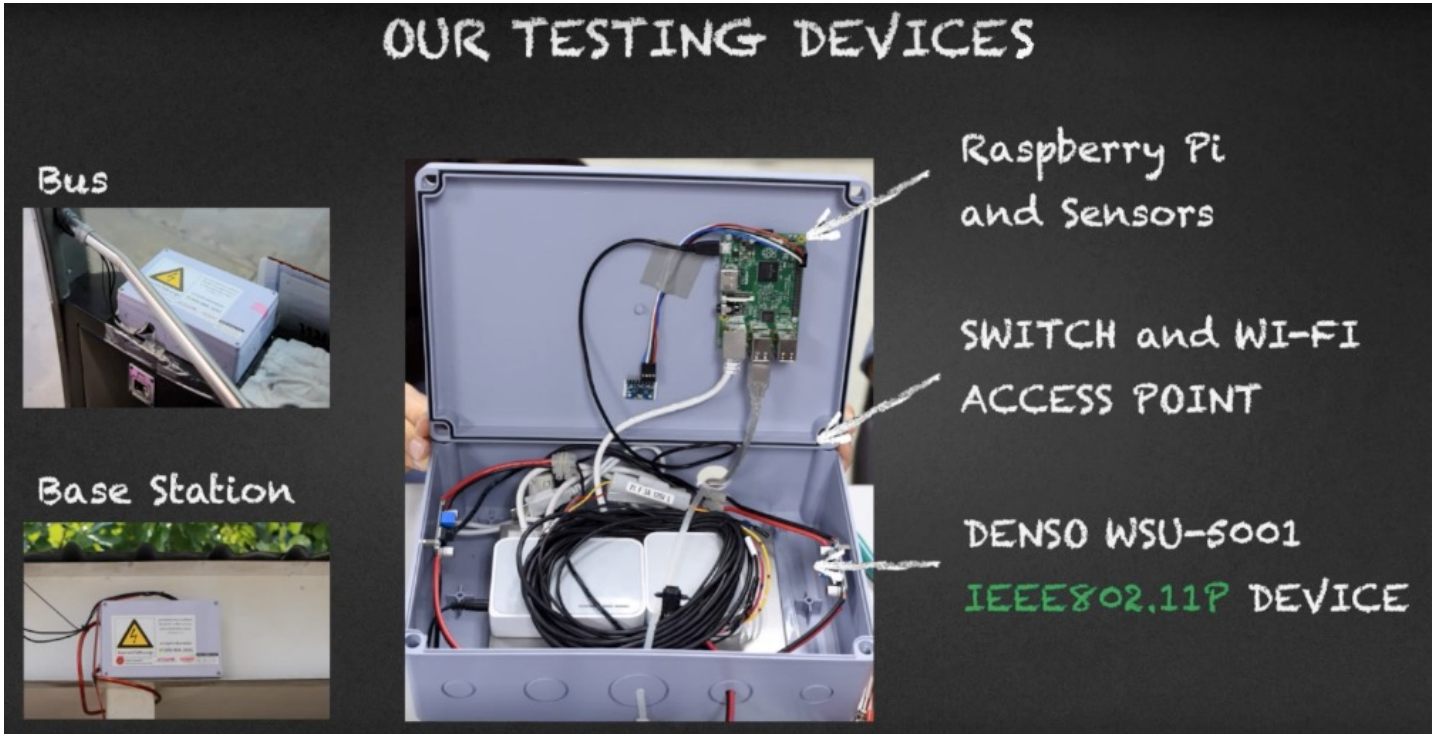
- Good understanding of user behavior can lead to new product development or improvements of the existing products

- Walmart -- Pop-Tarts before a hurricane
  - Prehurricane, Strawberry Pop- Tart sales increased about sevenfold
- 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

Example: Tracing Traffic

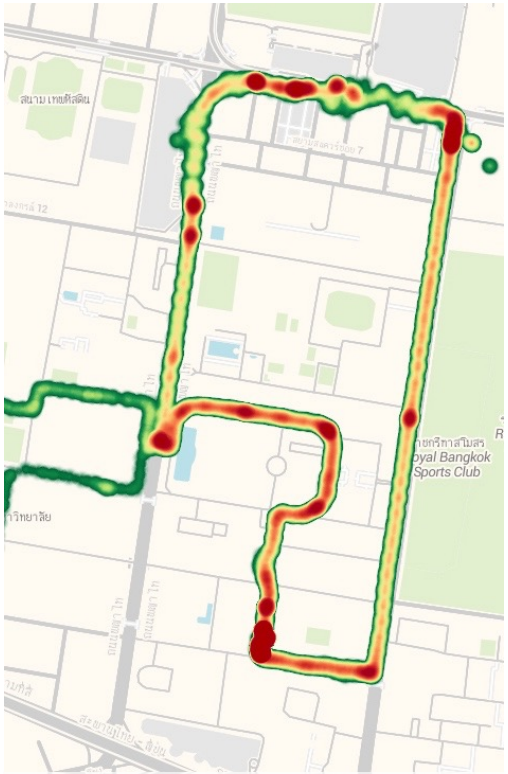


Example: Tracing Traffic

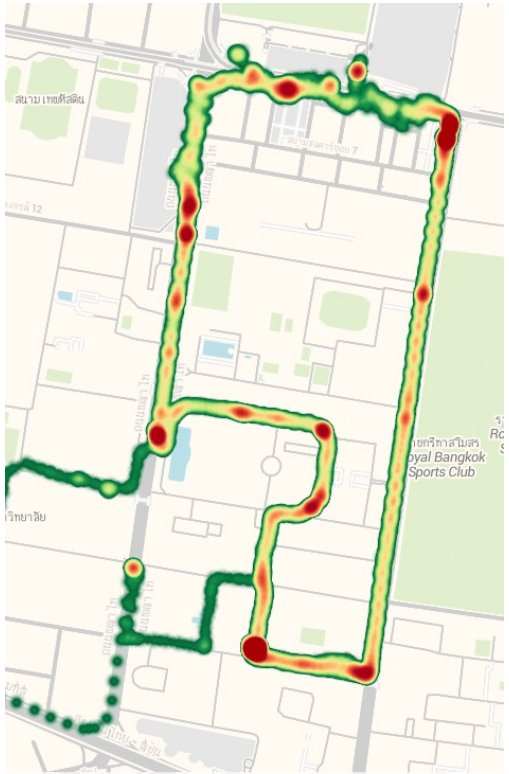


GPS Average Speed

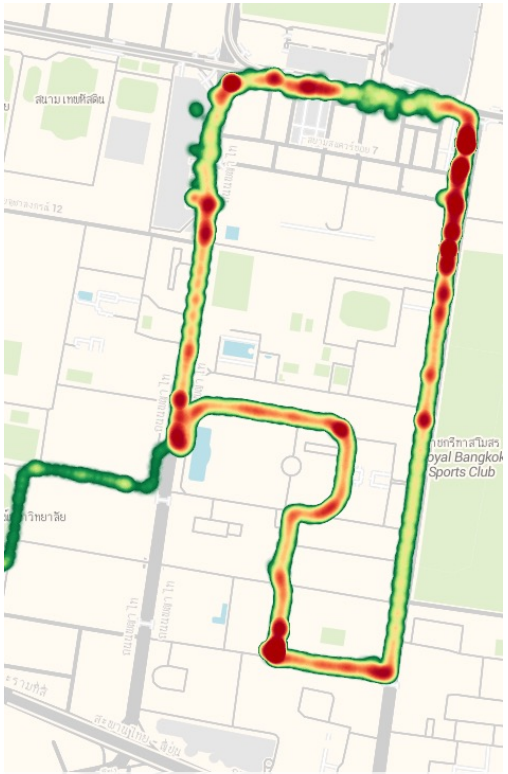
6:00-10:00



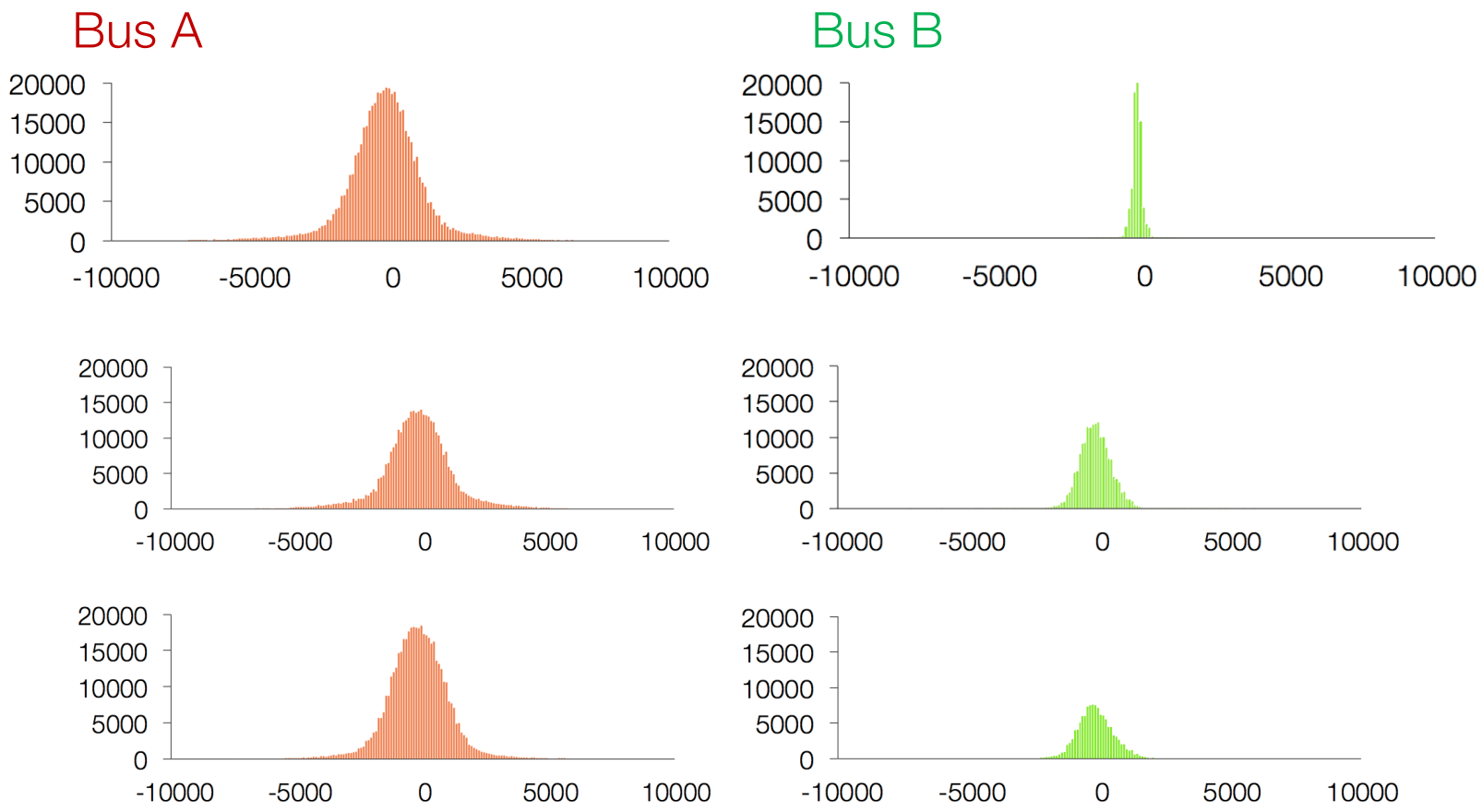
10:00-15:00



15:00-18:00



Bus Drivers' Behaviors

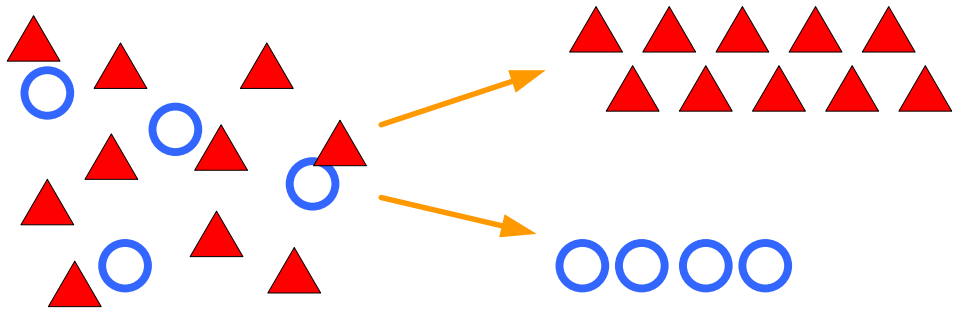


3) Data Products

- 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

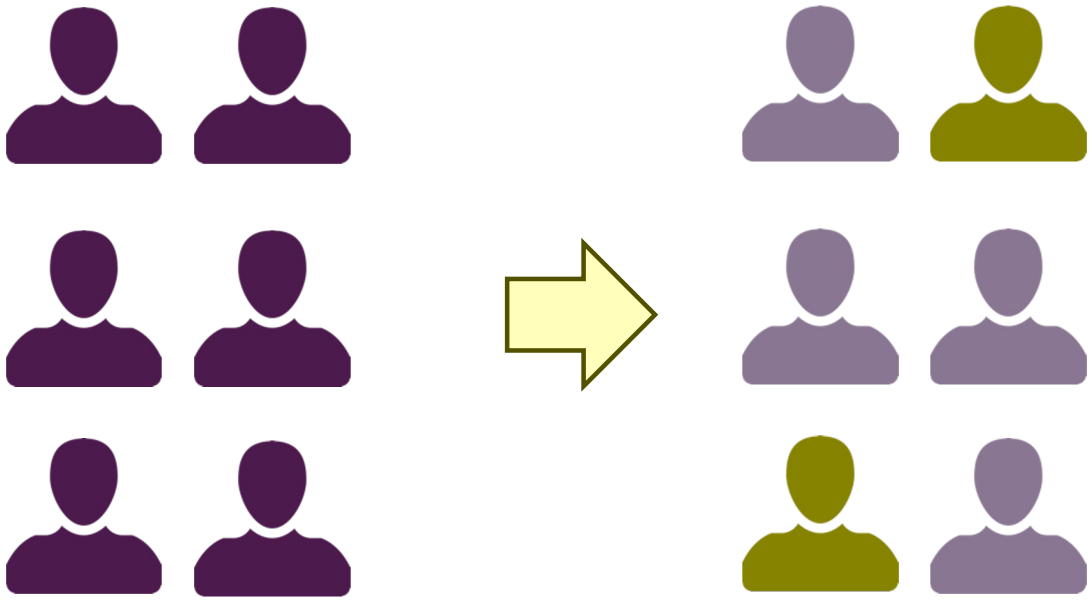
Machine Learning Classification

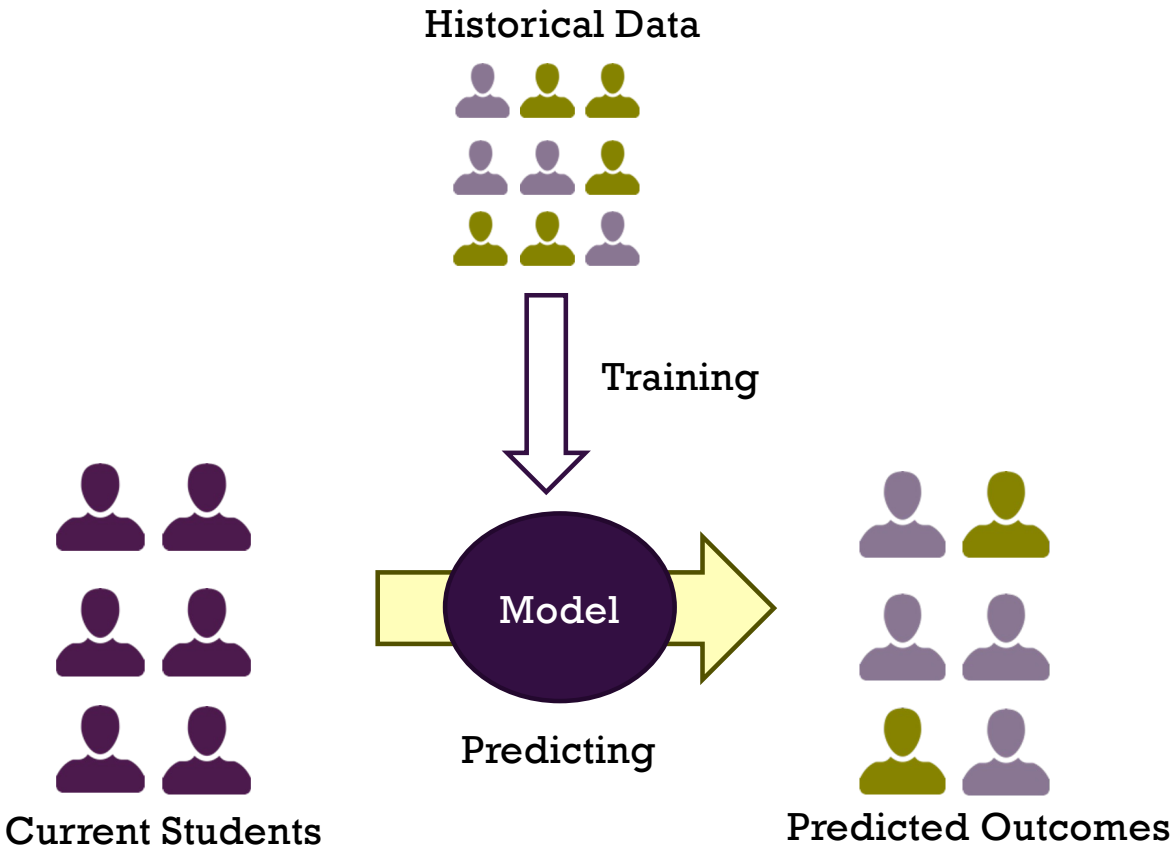
- Identify to which set of categories a new observation belong



- Example: spam filtering, customer churn prediction, complaint classification

Example: Students Grade Prediction

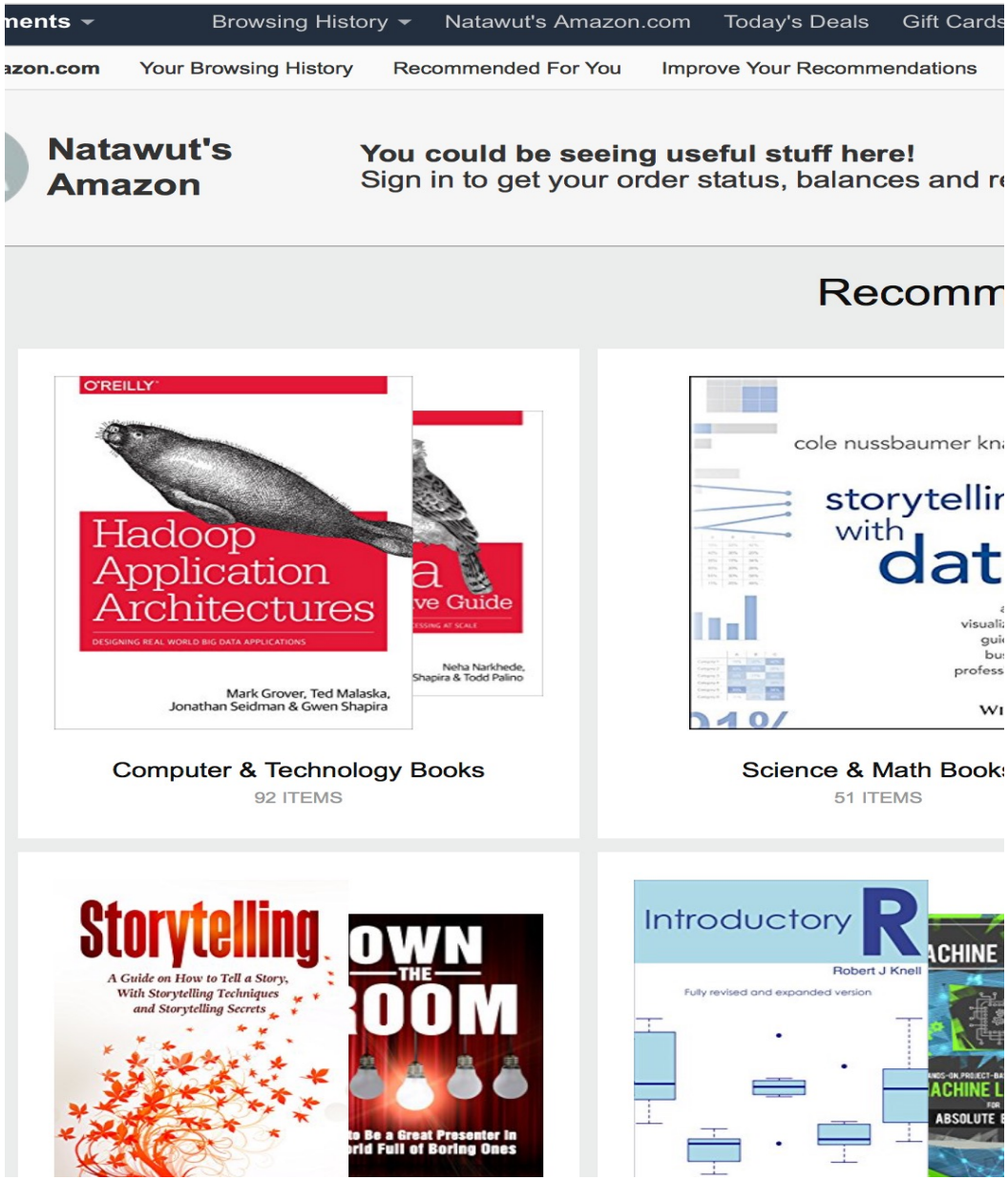




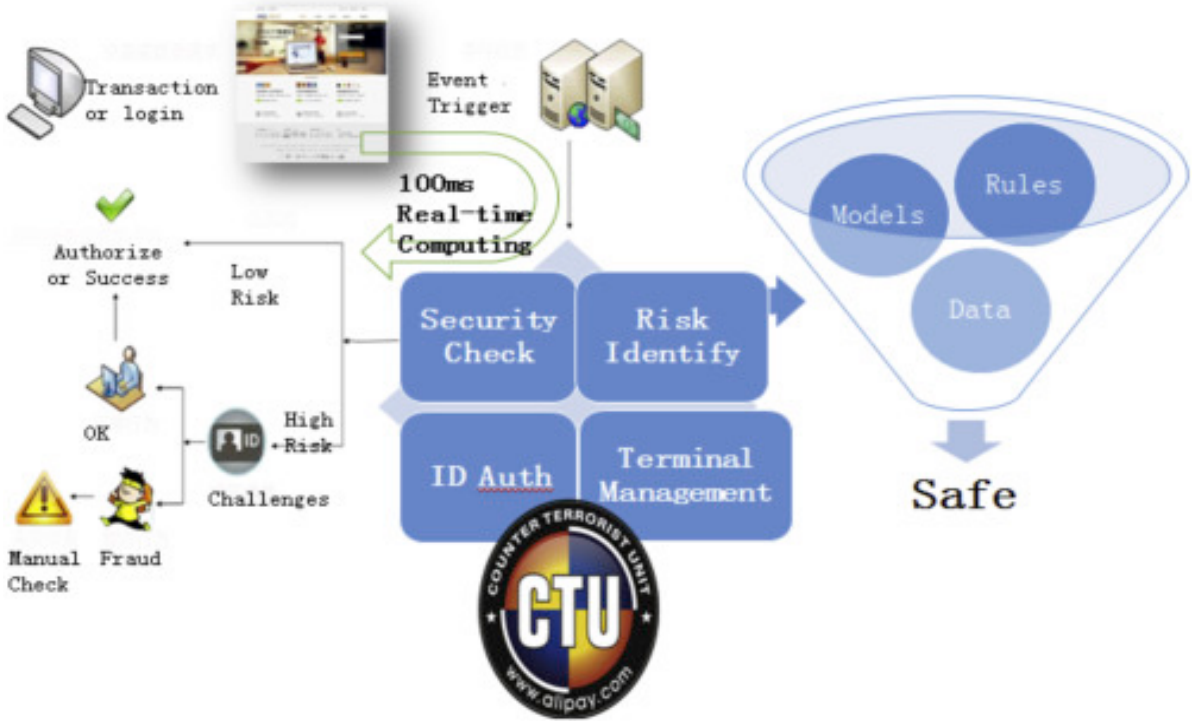
$$\frac{OS \times Data\ Struct \times Prog}{9} > 7$$

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

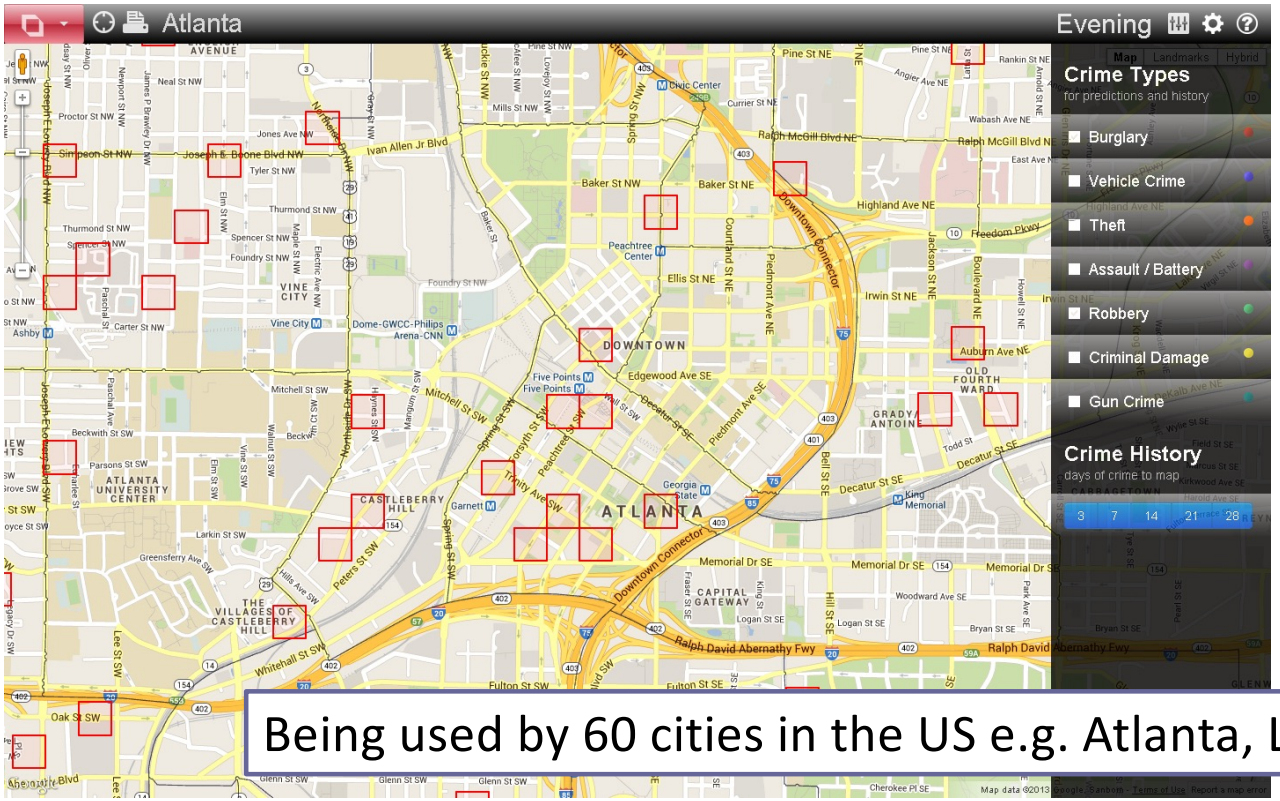


Case study: Alibaba Fraud Detection



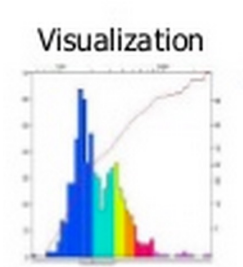
Source: http://www.sciencedirect.com/science/article/pii/S2405918815000021

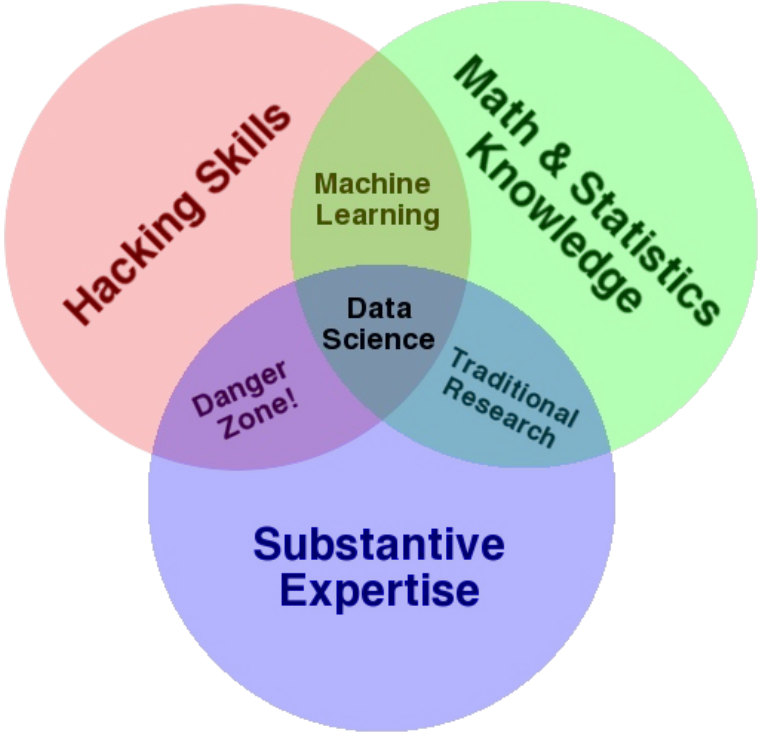
Case study: Predictive Policing

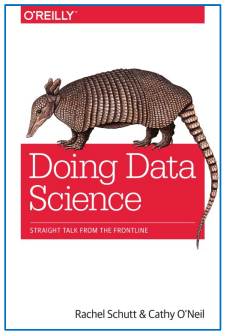


Source: http://www.forbes.com/sites/ellenhuet/2015/02/11/predpol-predictive-policing

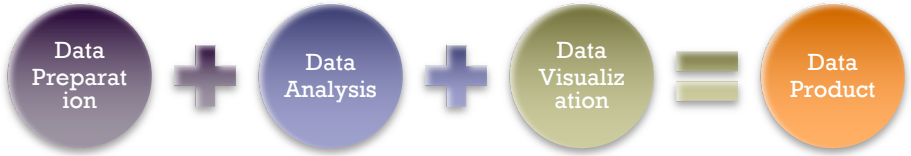
Drew Conway's Data Science Venn diagram (Skills)







Drew Conway's Venn diagram of data science, 2010

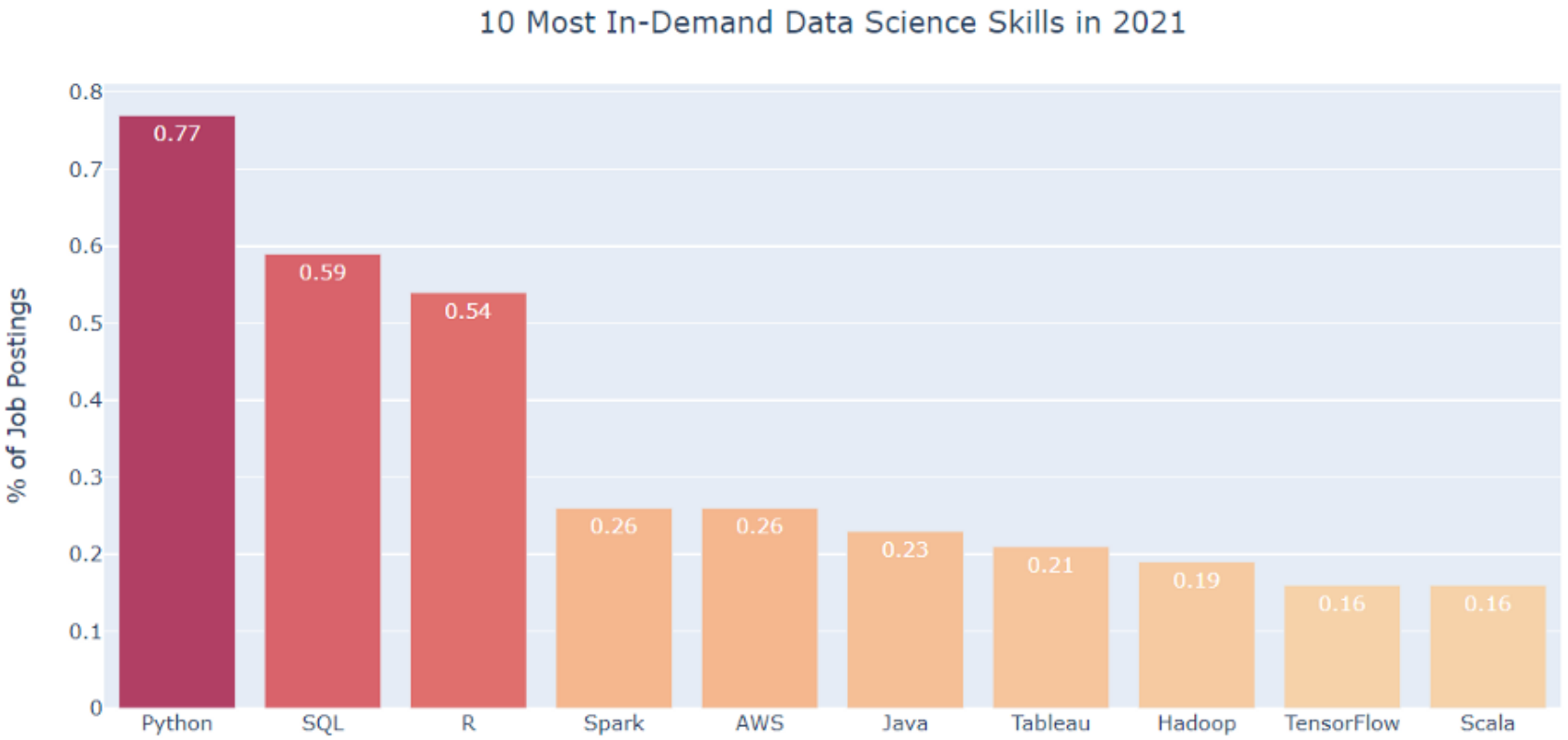


Chula Data Science

The Most In-Demand Skills for Data Scientists in 2021

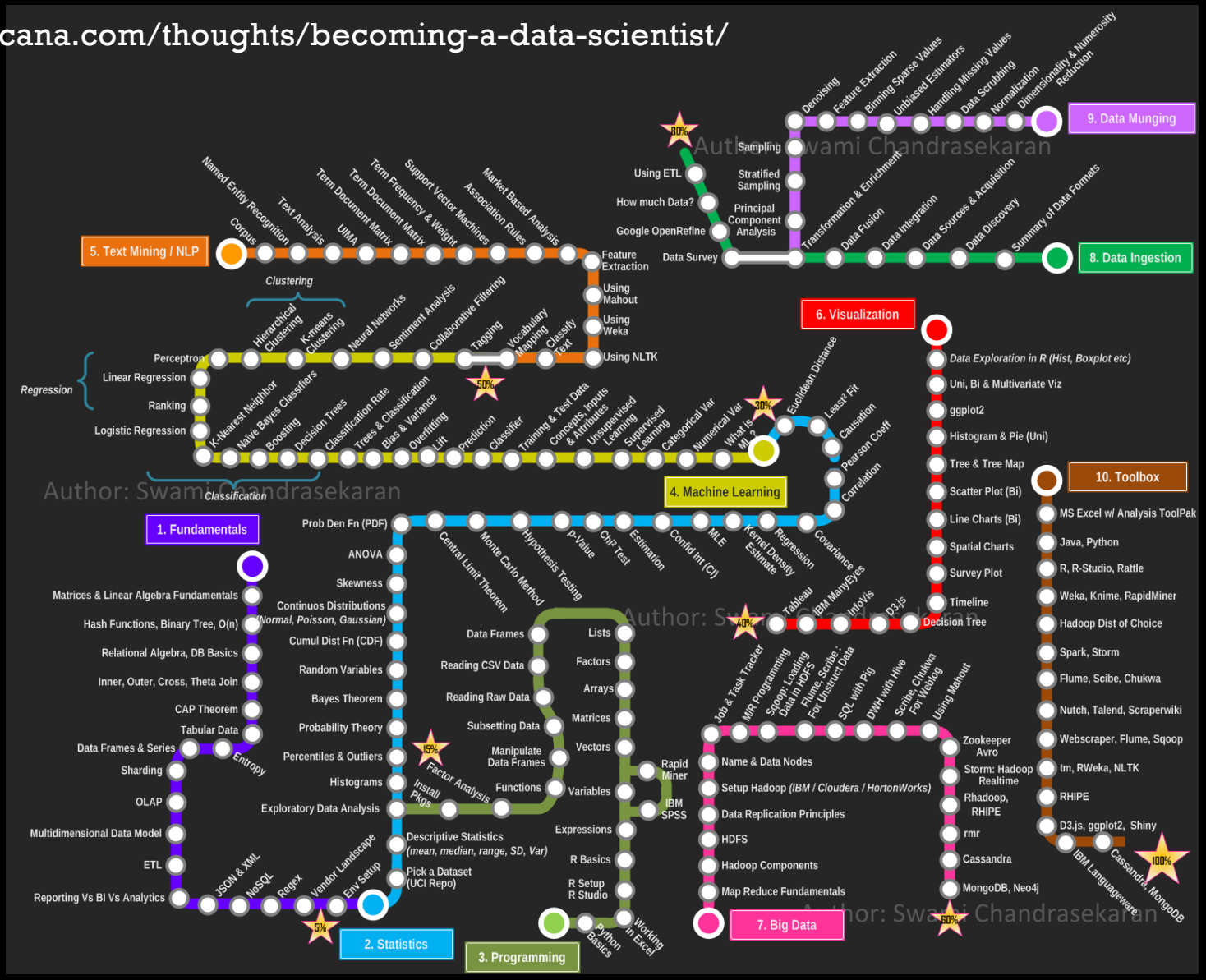
Results from webscraping over 15,000 Data Scientist job postings

Terence Shin Mar 22, 2021 · 5 min read ★

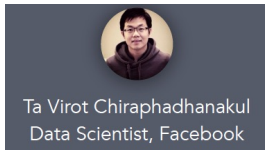


https://towardsdatascience.com/the-most-in-demand-skills-for-data-scientists-in-2021-4b2a808f4005

http://nirvacana.com/thoughts/becoming-a-data-scientist/



Needed Skills (cont.)



Harvard Business Review

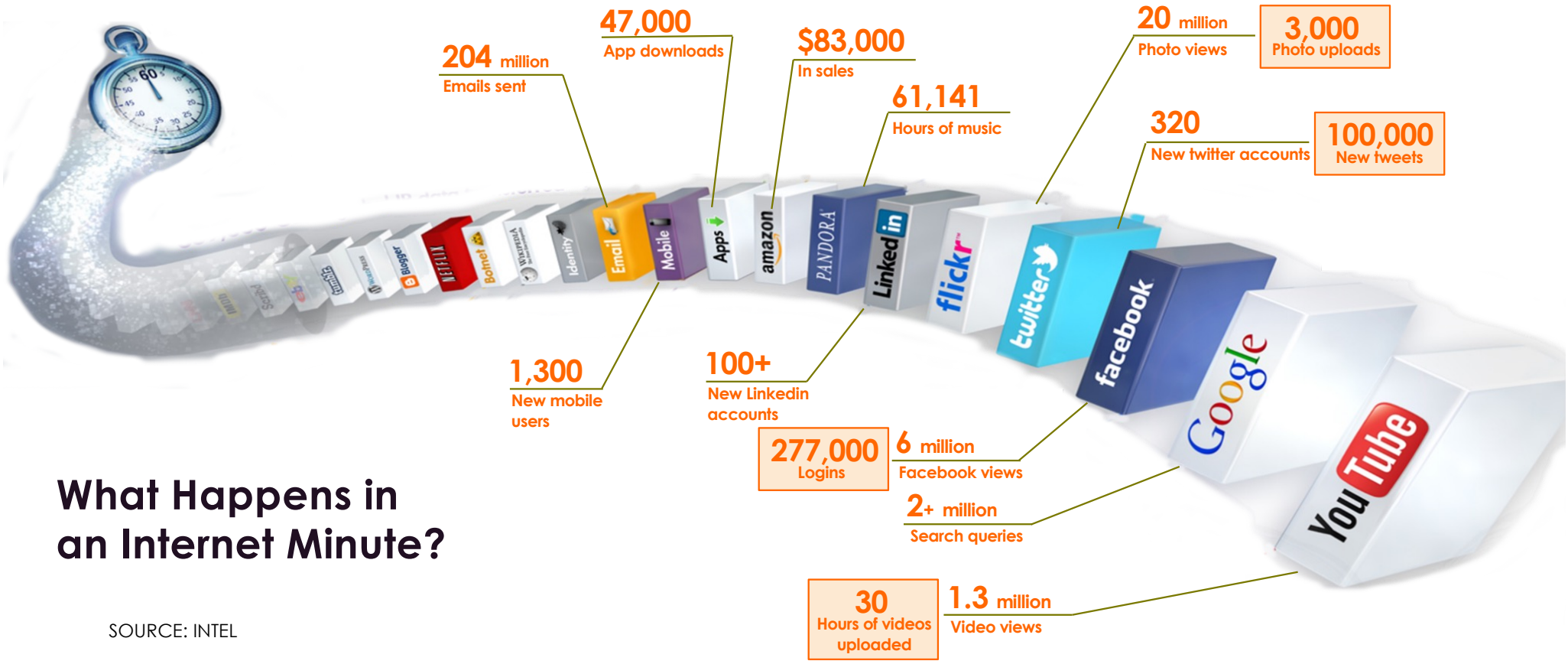
Data Scientist: The Sexiest Job of 21st Century is Tedious, and the Needs to Change

By Sean Kandel
From the April 2014 issue

+

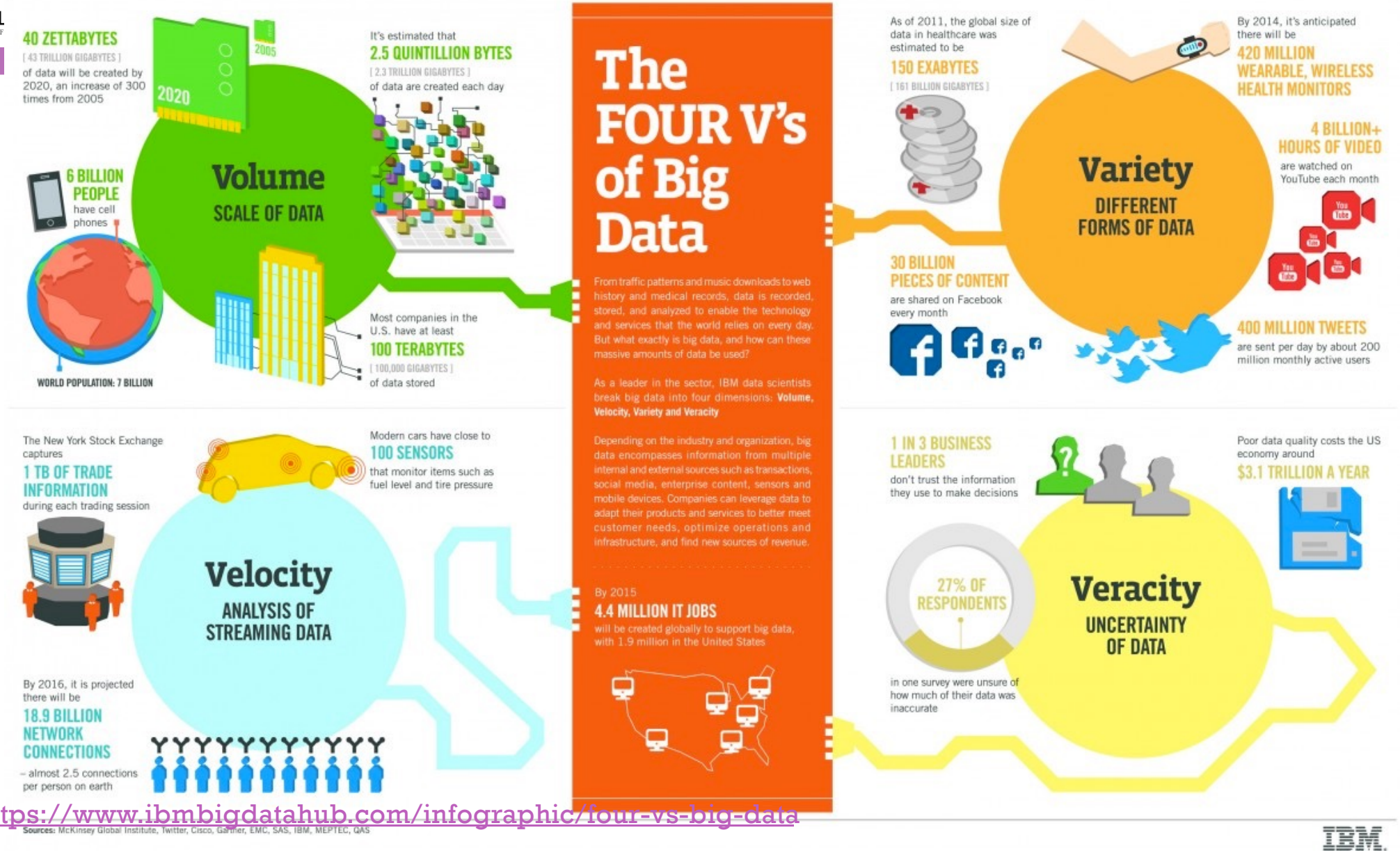
Big Data

Big Data Explosion



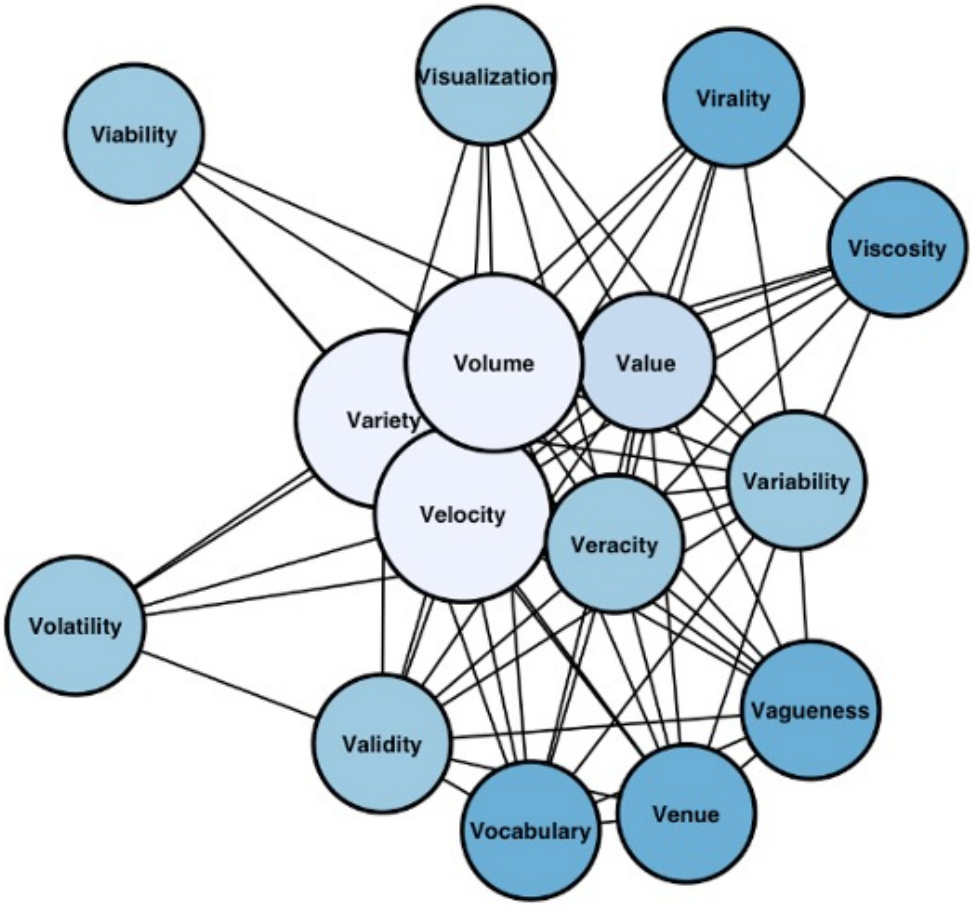
What Happens in an Internet Minute?

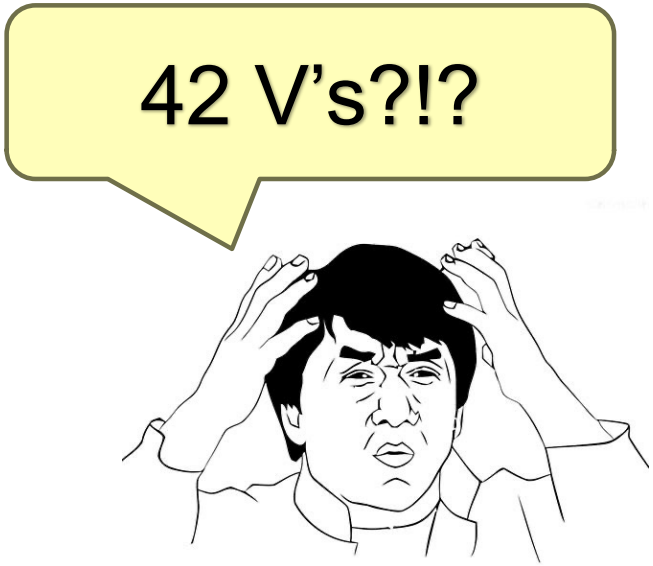
SOURCE: INTEL



https://www.ibmbigdatahub.com/infographic/four-vs-big-data

Now 42 V of Big Data





Big Data Driver: Internal + External Data

Owlet

https://owletcare.com/

KNOWING IS BEST

A Better Way to Know

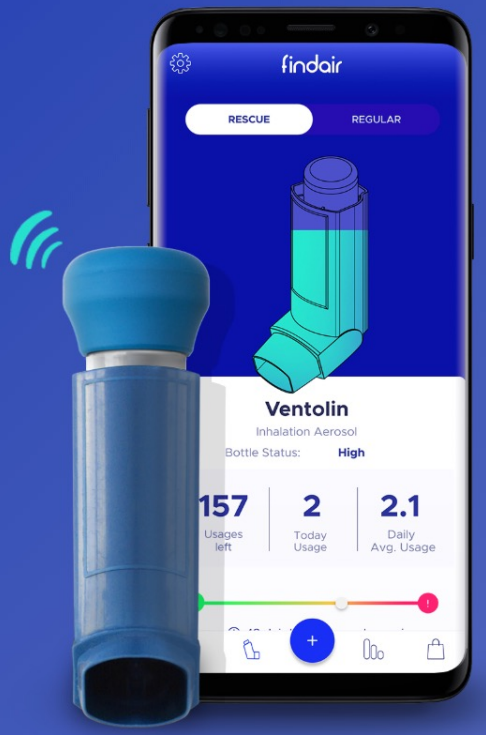
We understand what it's like to hover over the crib at night. That's why we invented the Owlet Smart Sock. It's a better way to check on your baby and smarter way to know they're okay.

findair

Home Product Contact

Control your asthma with FindAir smart inhaler

Full Control. Less asthma attacks. Better life.



Receive notifications when you forget your inhaler

Reduce the number of unexpected asthma attacks

Find out what causes your asthma attacks

2nd prize EIT HEALTH INNOSTARS AWARD 2018

1st prize UPC DIGITAL IMAGINATION CHALLANGE 2017

Best Pitch SMART HEALTH BUDAPEST 2017

1st prize INNOLABS DIGITAL HEALTH HACKHATHON 2017

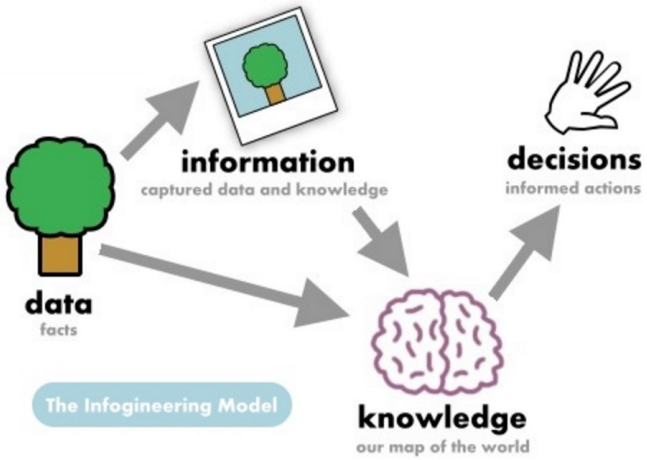
1st prize INNOLABS DIGITAL HEALTH HACKHATHON 2017

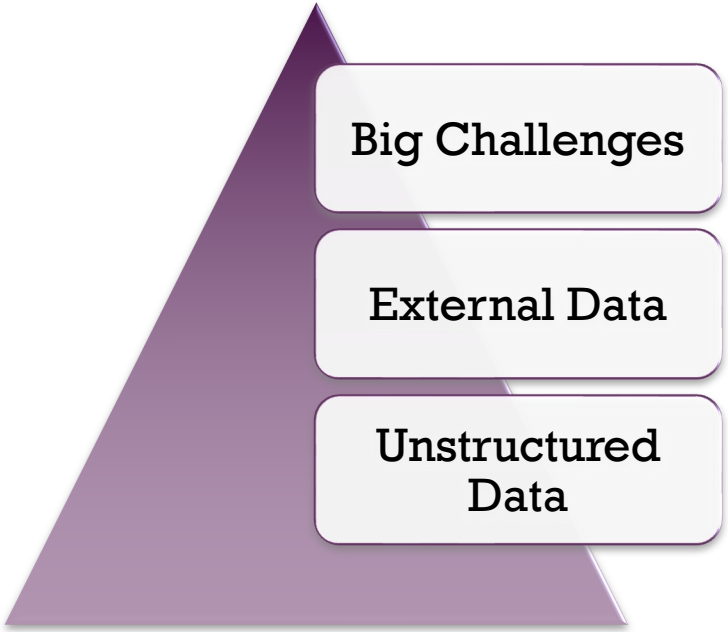
1st prize MEDTRENDS TOP TRENDS 2018

https://findair.eu/#Produkt

Big Data Analytics

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

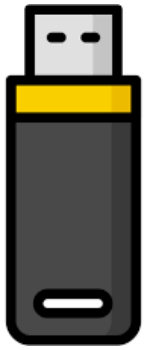




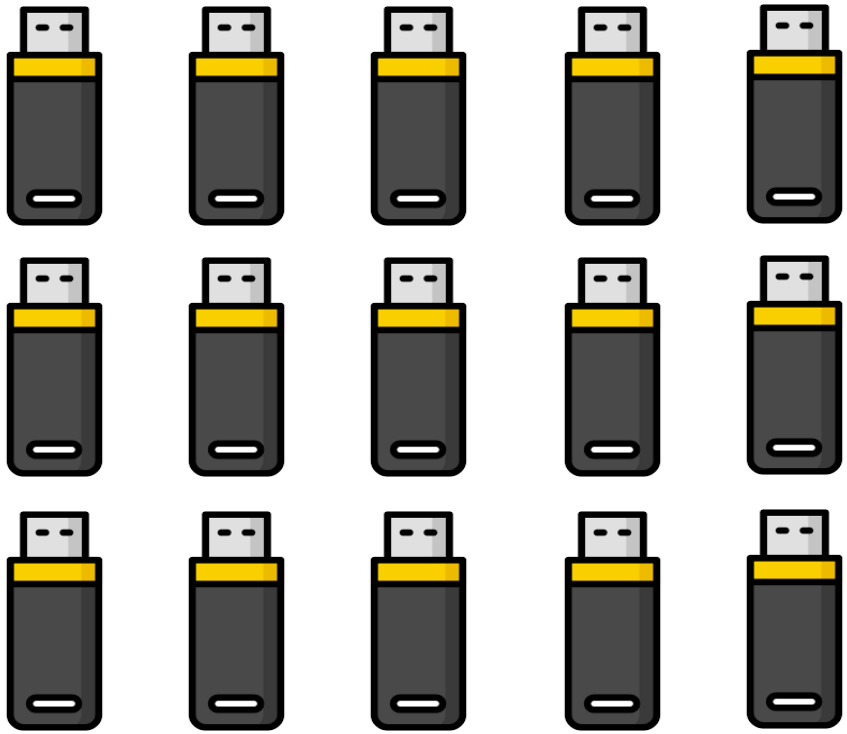
Big Data Challenges

Same tasks, but much more difficult!

2MB



200TB



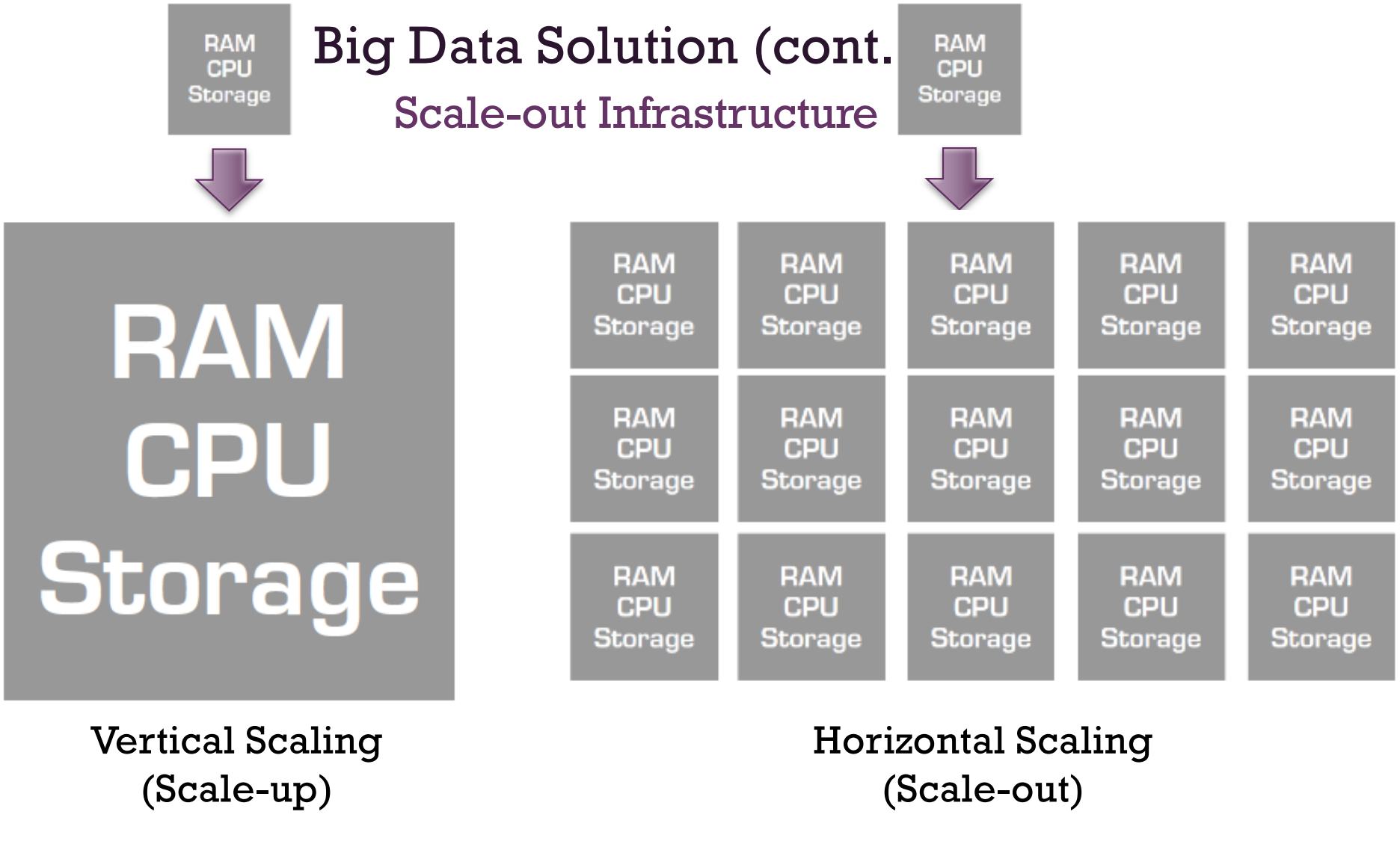
Big Data Solution

INFRASTRUCTURE

ALGORITHM

Big Data Solution (cont.

Scale-out Infrastructure

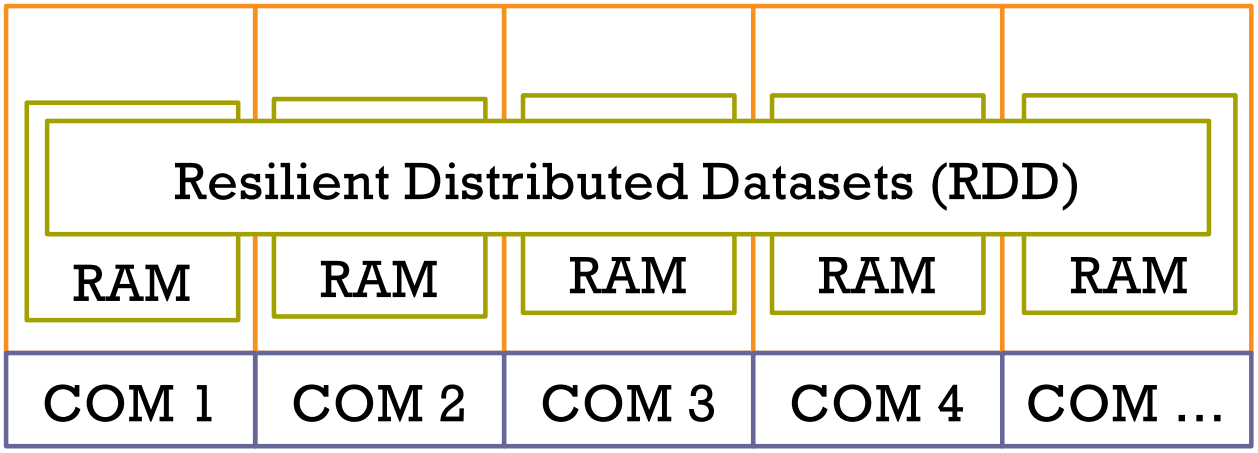


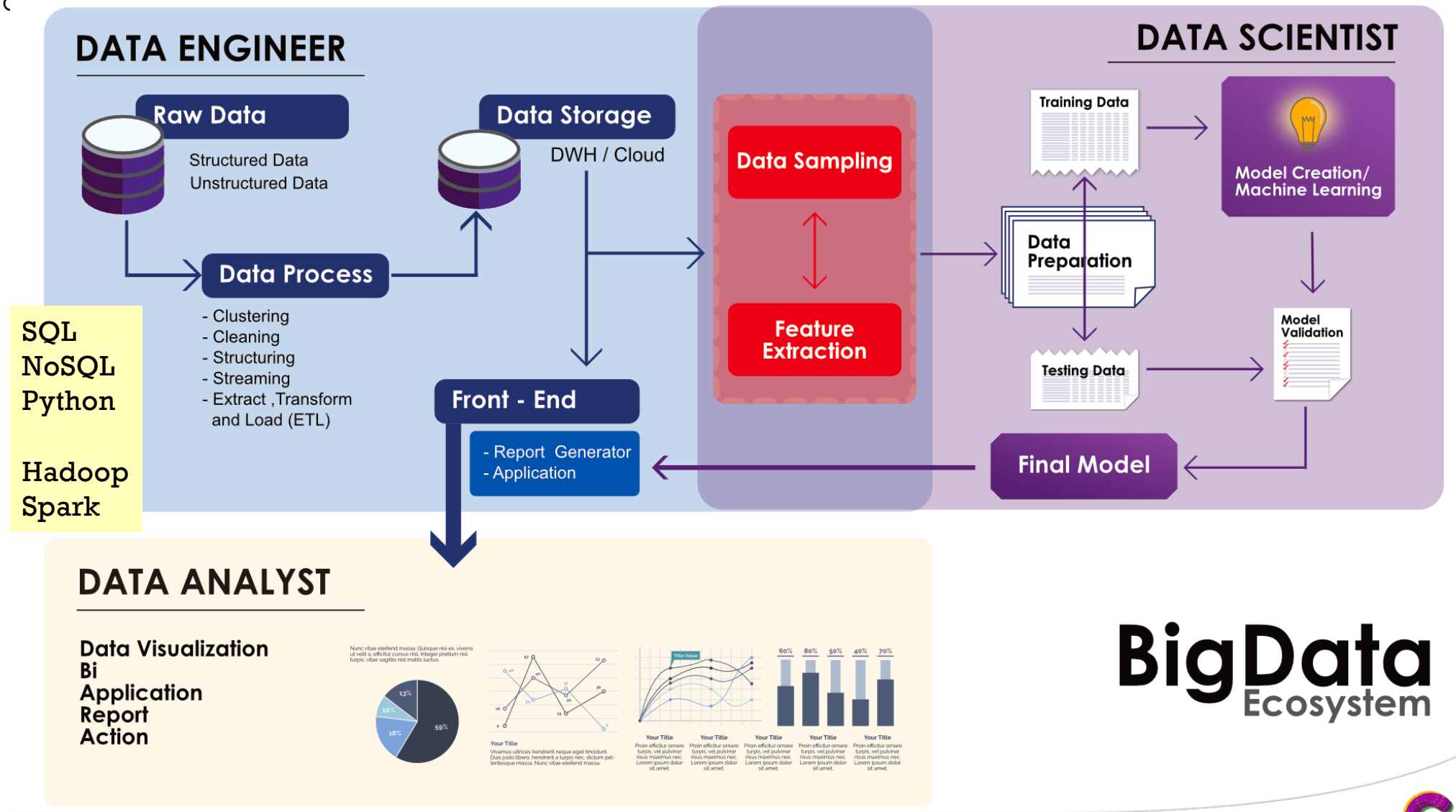
Vertical Scaling
(Scale-up)

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

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)

Data Science Process

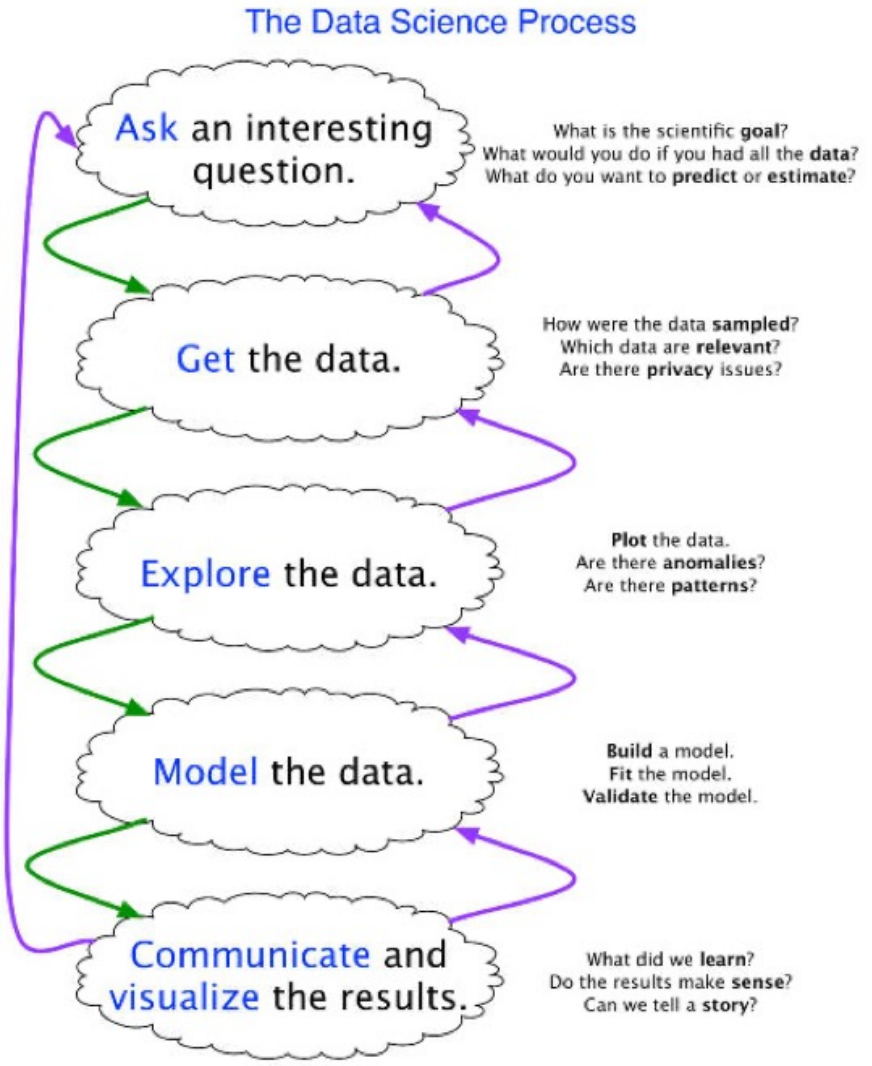
Data Science Process

Dr.Virote

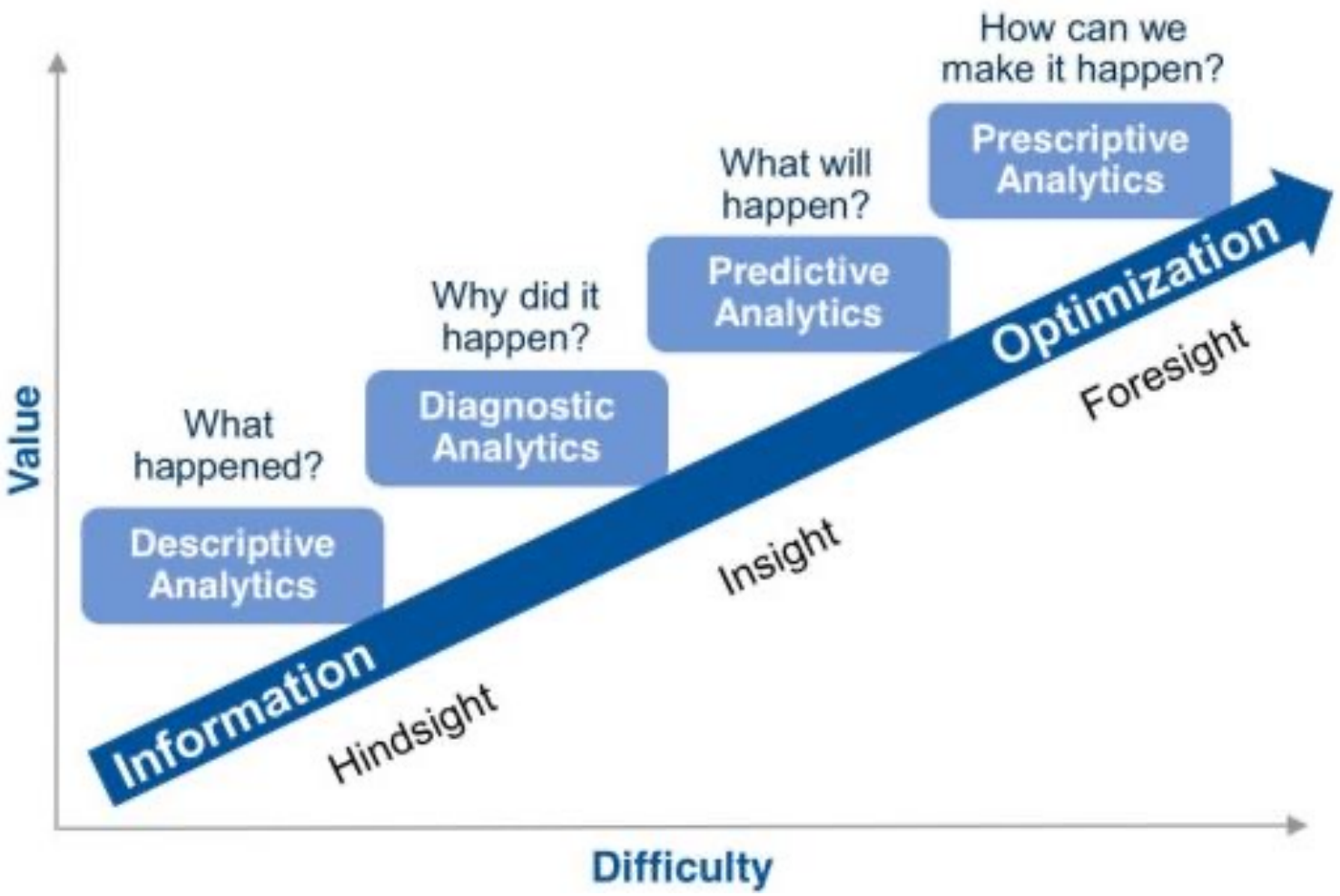
1. Transform data into valuable insights
2. Transform data into data products
3. Transform data into interesting stories

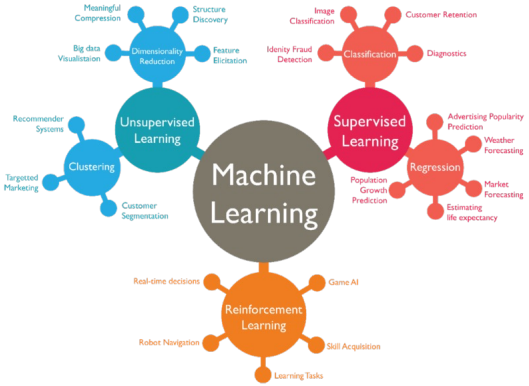
Aj.Natawut

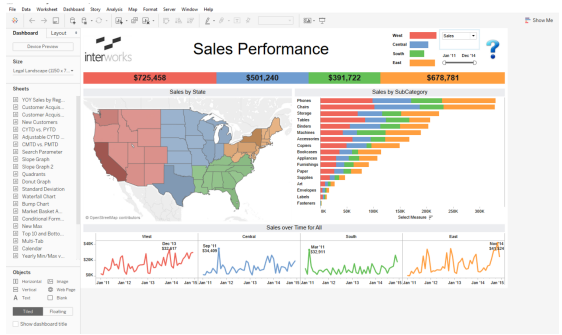
1. Measurement (decision)
2. Insights (knowledge)
3. Data Products (Innovation, Intelligent)



Data Analytics







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

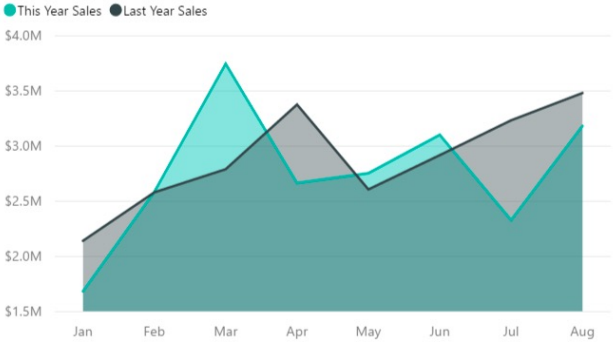
Total Stores
NEW & EXISTING STORES
104

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

This Year's Sales
BY CHAIN


New Stores, New Stores Target
YEAR TO DATE

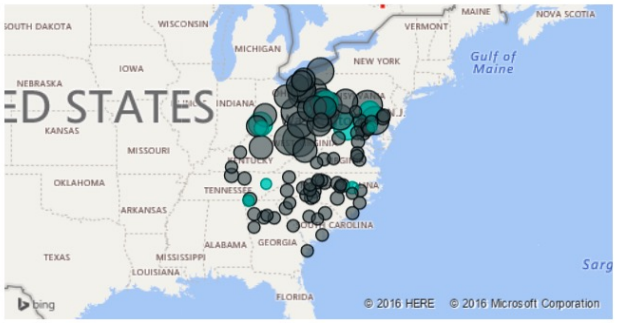

This Year's Sales
NEW STORES ONLY
$2M

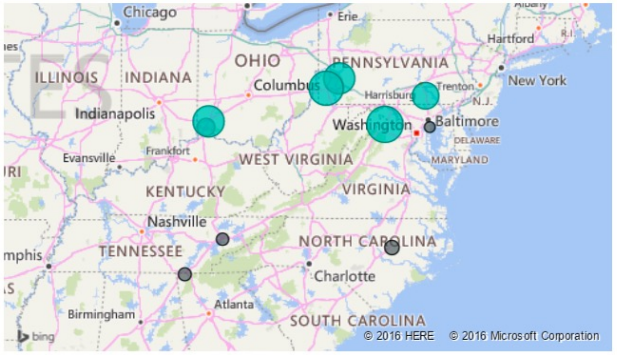
This Year's Sales, Last Year's Sales
BY FISCAL MONTH


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

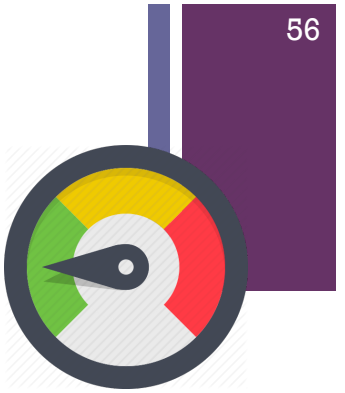

New Stores
NEW STORES ONLY
10

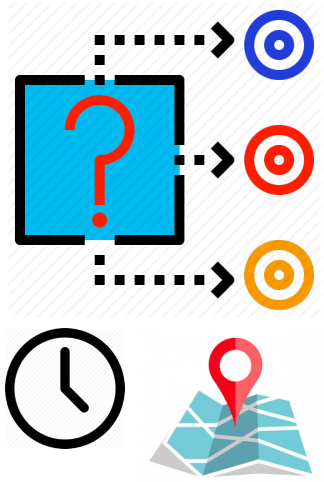
Stores Opened This Year
BY OPEN MONTH, CHAIN


This Year's Sales
BY POSTAL CODE, STORE TYPE


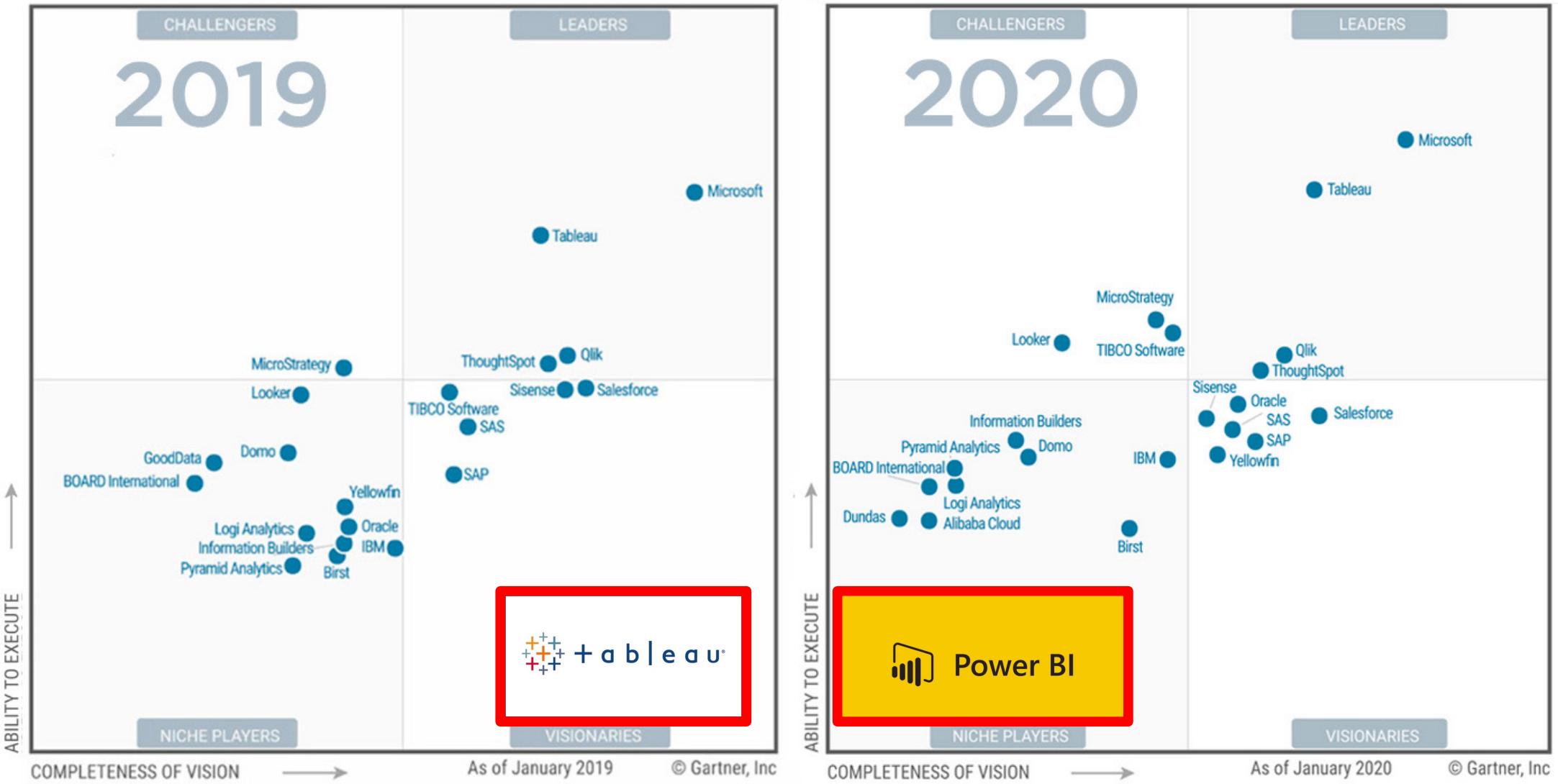
This Year's Sales
BY CITY, CHAIN


Sales Per Sq Ft
BY NAME

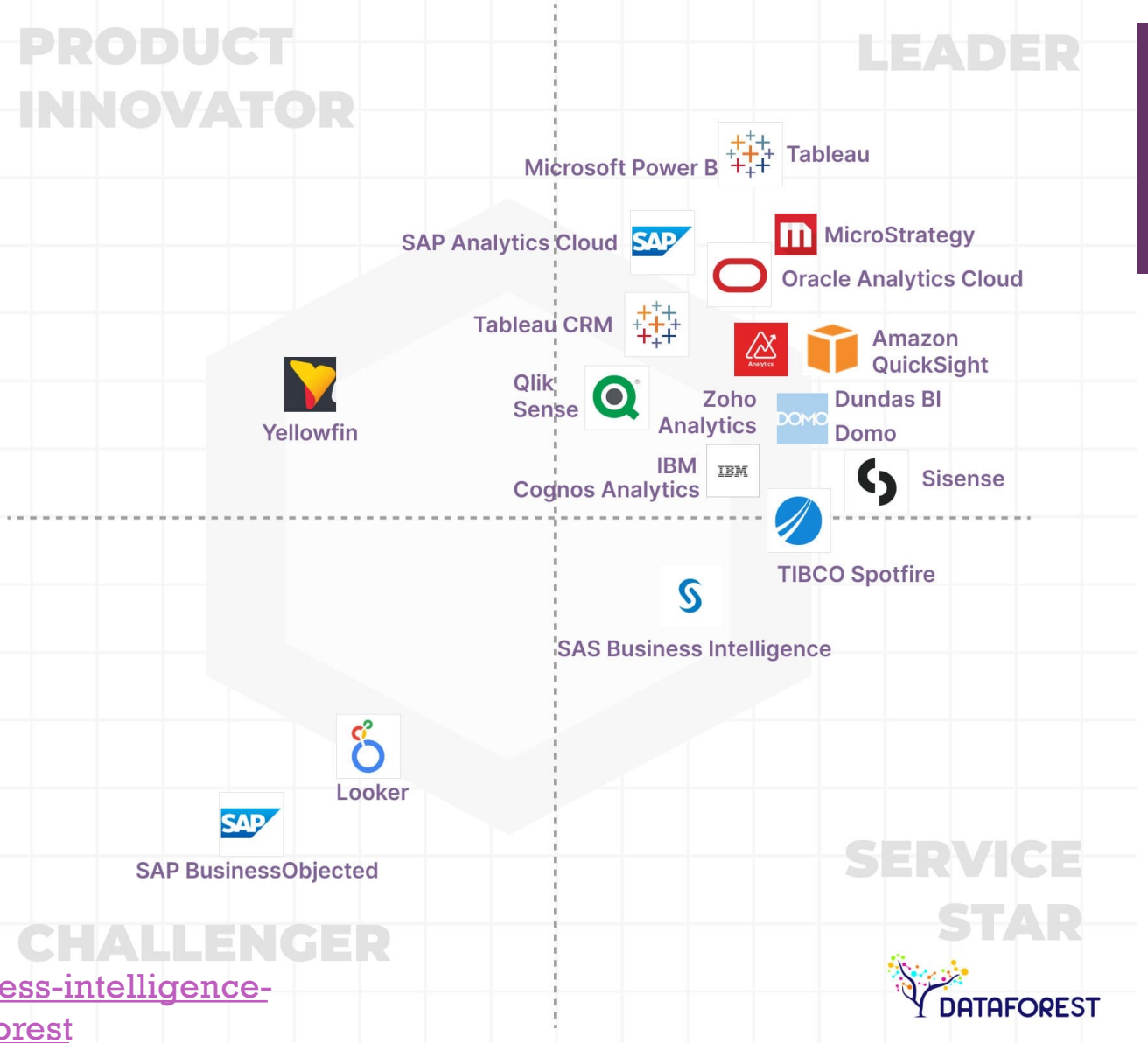


Magic Quadrant for Analytics and Business Intelligence Platforms.



Source: Gartner (Feb 2019 and 2020)

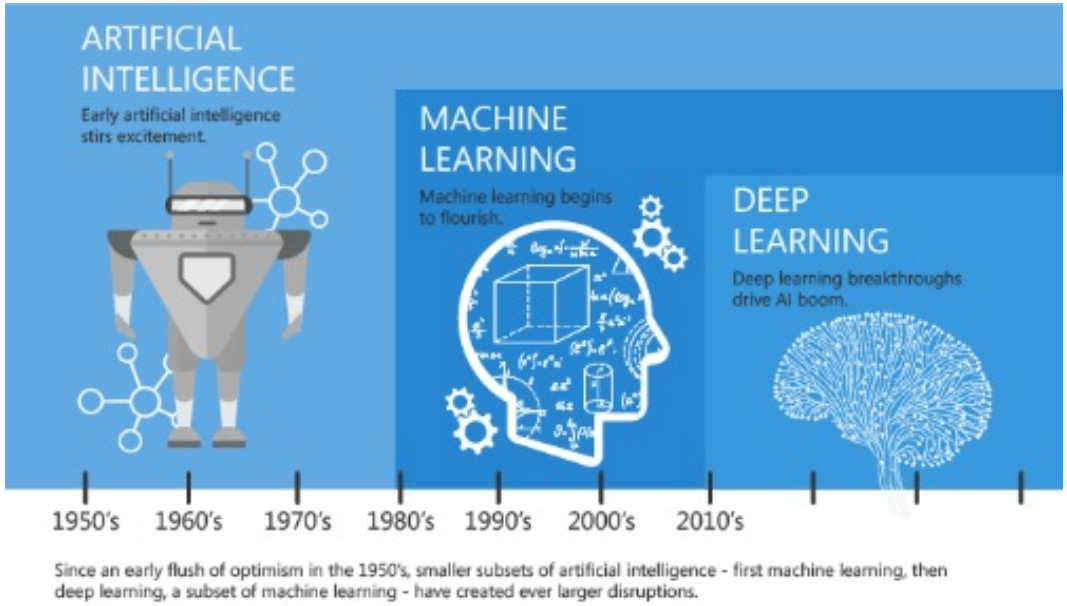
Power BI



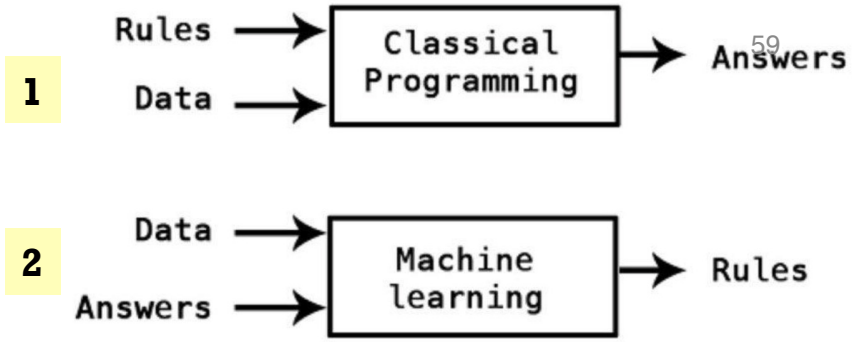
https://dataforest.ai/blog/best-business-intelligence-tool-of-2023-top-16-bi-tools-by-dataforest

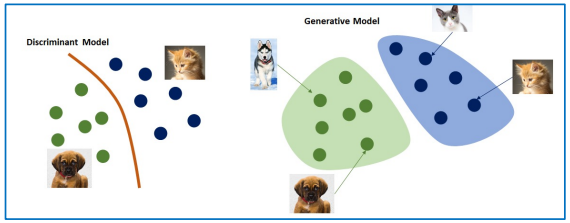
AI = Automation

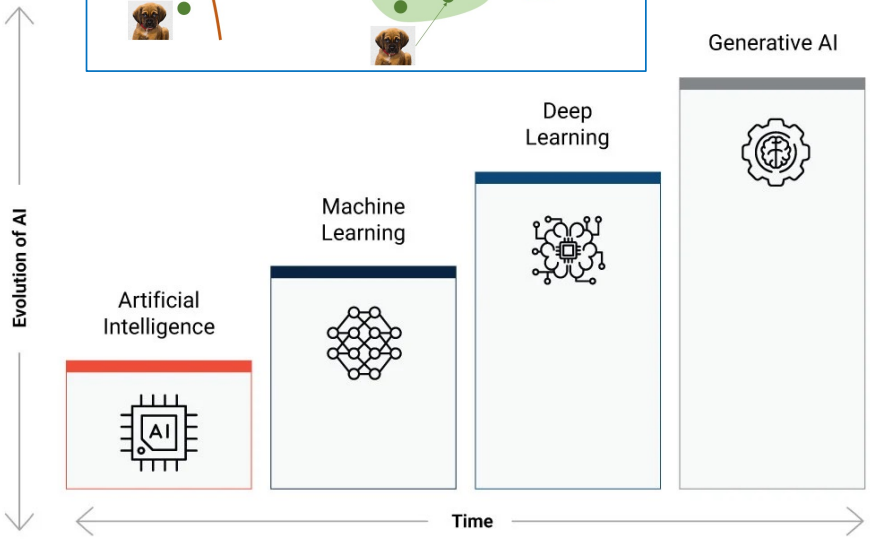
- 1) Rule-based AI
- 2) Machine Learning (ML)



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.

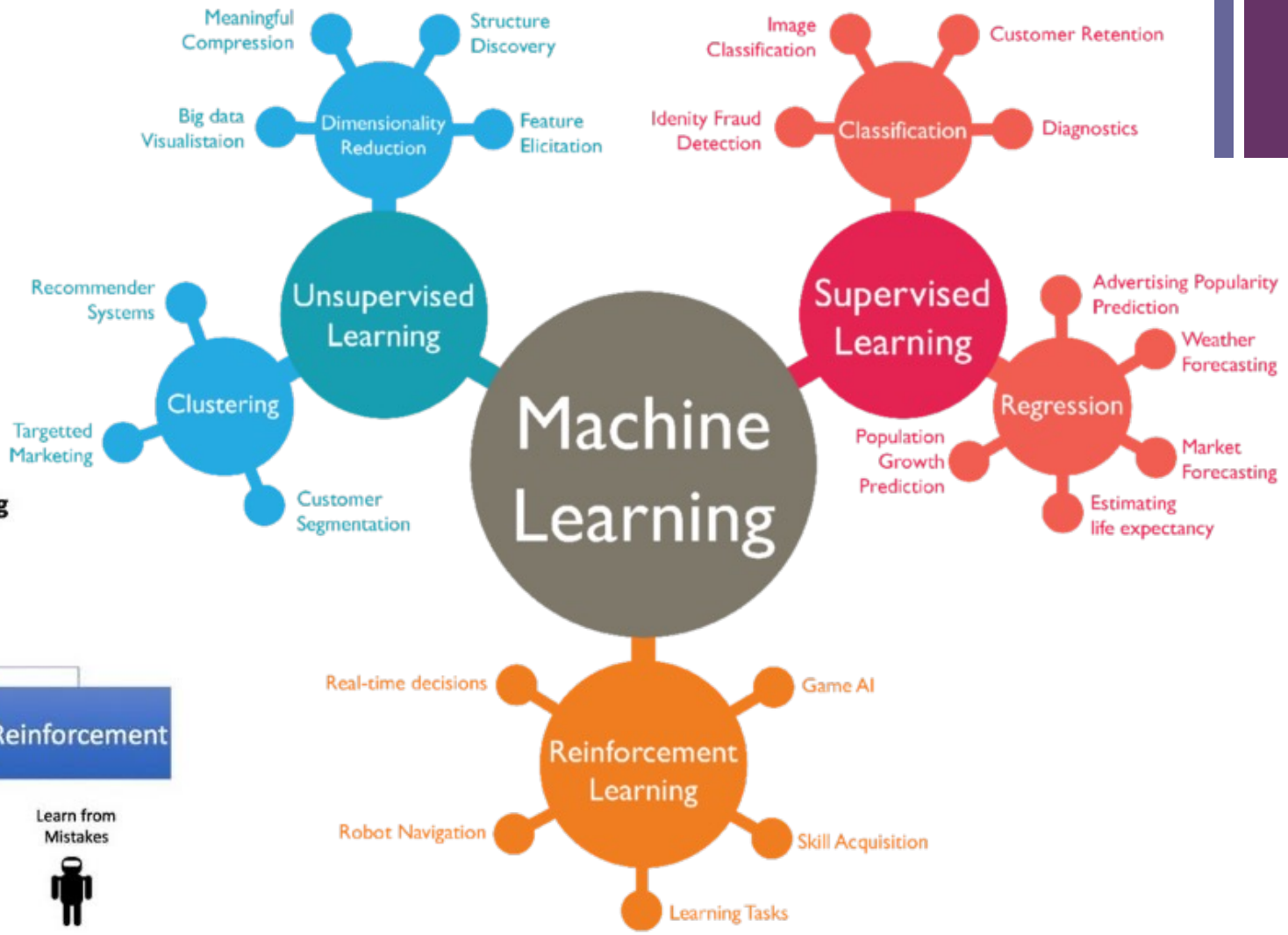


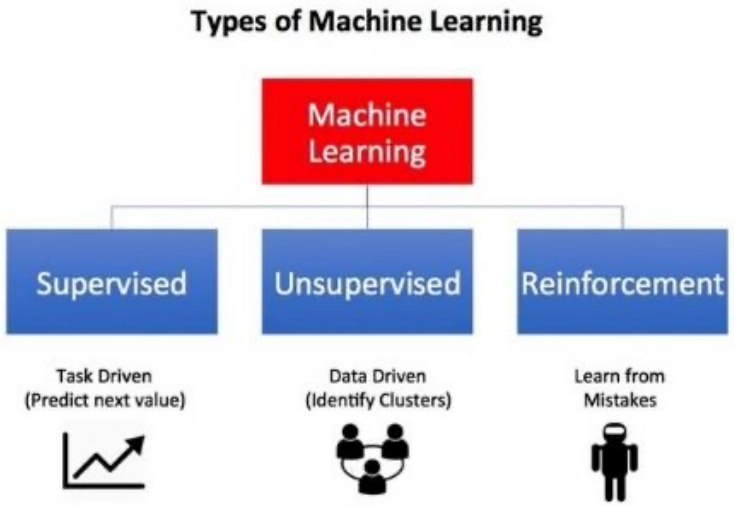


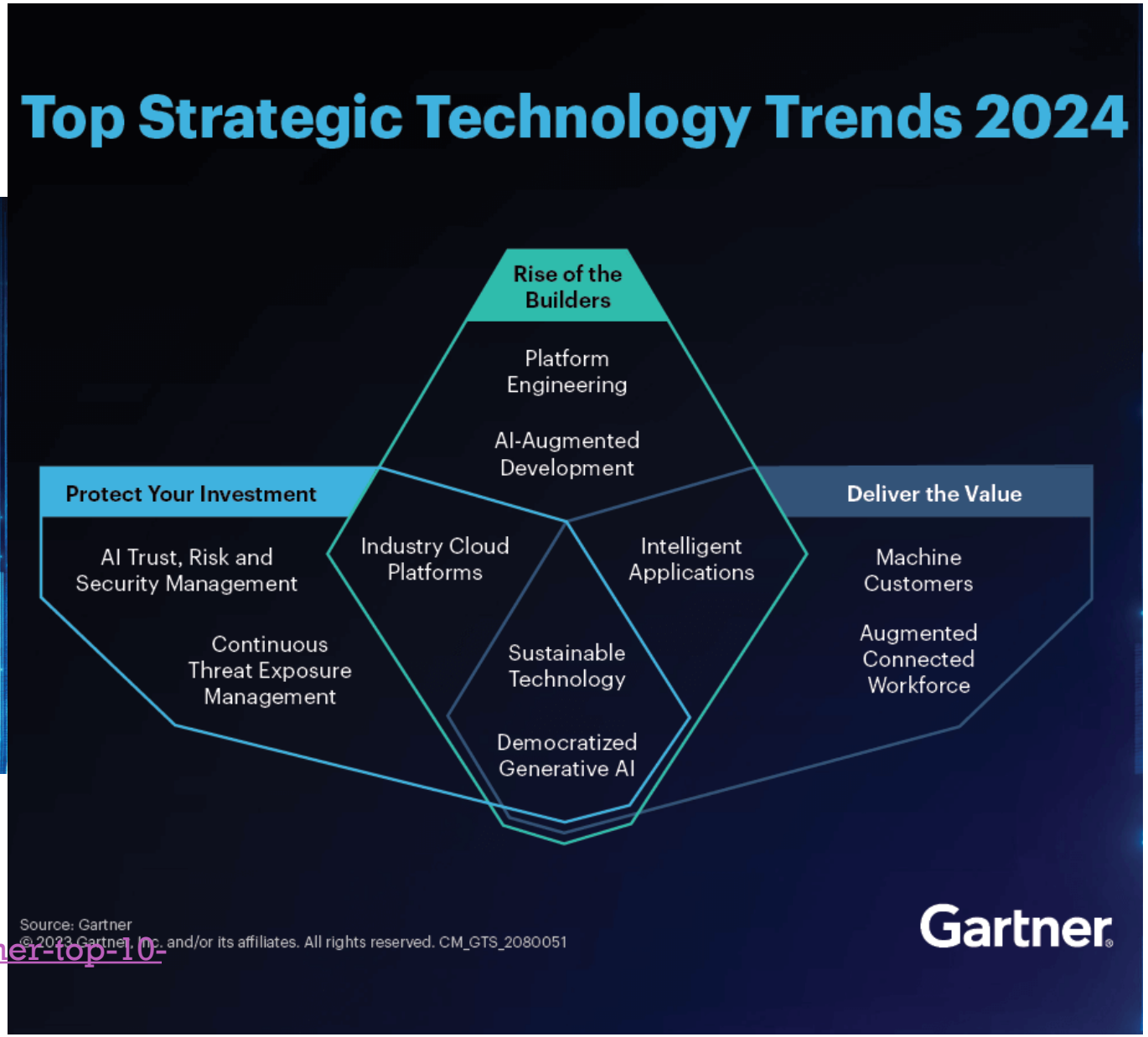


https://mc.ai/machine-learning-basics-artificial-intelligence-machine-learning-and-deep-learning/

Machine Learning (ML)



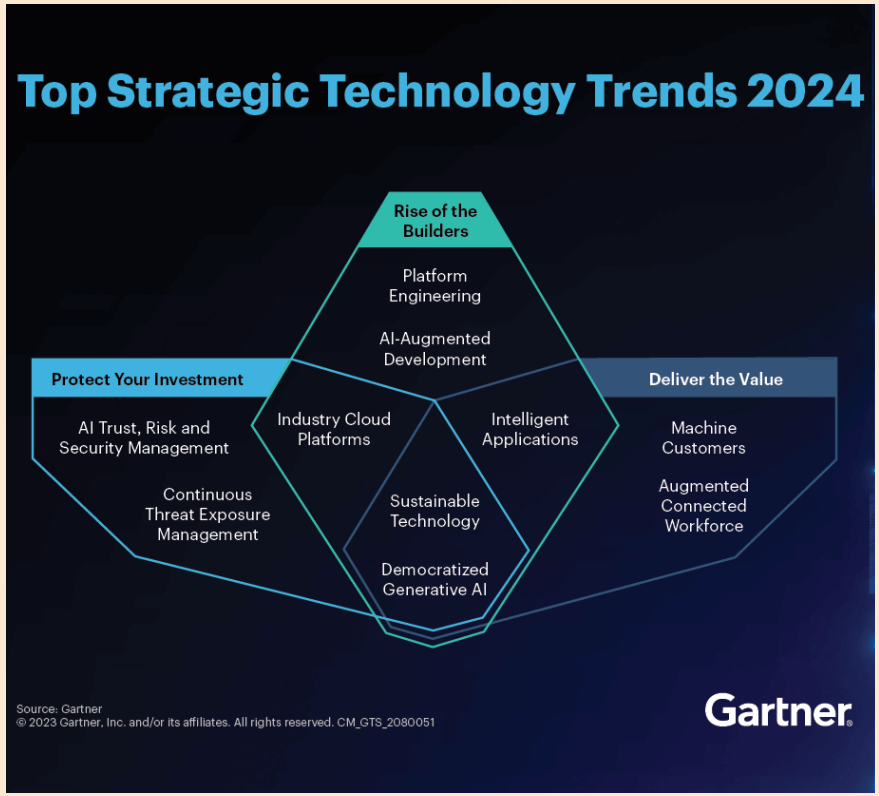




https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-tre nds-for-2024

Data Trend in 2024 (cont.)

- AI (AI everywhere & Gen AI) is the key component.
- Knowledge without action (Platform Engineering) is meaningless.
- Cloud technology is a modern infrastructure.



Gartner Magic Quardrant

Vit Niennattrakul, Ph.D.

Figure 1: Magic Quadrant for Cloud Infrastructure and Platform Services

Figure 1: Magic Quadrant for Cloud AI Developer Services

Categories of AWS services

Vit Niennattrakul, Ph.D.

- Analytics
- Application Integration
- AR and VR
- Blockchain
- Business Applications
- Compute
- Cost Management
- Customer Engagement
- Database
- Developer Tools
- End User Computing
- Game Tech
- Internet of Things
- Machine Learning
- Management and Governance
- Media Services
- Migration and Transfer
- Mobile
- Networking and Content Delivery
- Robotics
- Satellite
- Security, Identity, and Compliance
- Storage

Data preparation for AI / ML and data science

Vit Niennattrakul, Ph.D.

Logs, app data

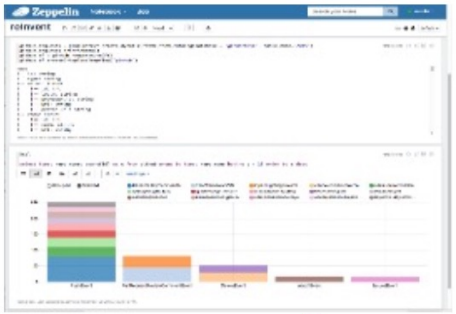
Amazon RDS

Other databases

On-premises data

Streaming data

Notebooks:
data exploration,
experimentation



AWS Glue
ingest

raw data

AWS Glue
transform

cleaned and
enriched data

AWS Glue
transform

extracted
features

AWS Glue
transform

training
data

Amazon Forecast

Amazon Comprehend

Amazon SageMaker

Amazon Rekognition

Amazon Lex

AWS Academy Service

AWS Academy Learner Lab

- Amazon API Gateway
- AWS App Mesh
- Application Auto Scaling
- AWS AppSync
- Amazon Athena
- Amazon Aurora
- AWS Backup
- AWS Certificate Manager (ACM)
- AWS Batch
- AWS Cloud9
- AWS CloudFormation
- Amazon CloudFront
- Amazon CloudSearch
- AWS CloudShell
- AWS CloudTrail
- Amazon CloudWatch
- AWS CodeCommit
- AWS CodeDeploy
- Amazon CodeWhisperer
- AWS Config
- AWS Systems Manager (SSM)
- Amazon Textract
- AWS Cost and Usage Report
- AWS Cost Explorer
- AWS Data Pipeline
- AWS DeepComposer
- AWS DeepLens
- AWS DeepRacer
- AWS Directory Service
- Amazon EC2 Auto Scaling
- AWS Elastic Beanstalk
- Amazon Elastic Block Store (EBS)
- Amazon Elastic Container Registry (ECR)
- Amazon Elastic Container Service (ECS)
- Amazon Elastic File System (EFS)
- Amazon Elastic Inference
- Amazon Elastic Kubernetes Service (EKS)
- Elastic Load Balancing (ELB)
- Amazon Elastic MapReduce (EMR)
- Amazon ElastiCache
- Amazon EventBridge
- AWS Fargate
- Amazon Timestream
- AWS Trusted Advisor
- Amazon Forecast
- AWS Glue
- AWS Glue DataBrew
- Amazon GuardDuty
- AWS Health
- AWS Identity and Access Management (IAM)
- AWS IAM Access Analyzer
- Amazon Inspector
- AWS IoT 1-Click
- AWS IoT Analytics
- AWS IoT Core
- AWS IoT Greengrass
- Amazon Kendra
- AWS Key Management Service (KMS)
- Amazon Kinesis
- Amazon Lex
- Amazon Machine Learning (Amazon ML)
- AWS Marketplace Subscriptions
- AWS Mobile Hub
- Amazon Neptune
- Amazon Virtual Private Cloud (Amazon VPC)
- AWS WAF - Web Application Firewall
- AWS OpsWorks
- Amazon Personalize
- Amazon QuickSight
- Amazon Redshift
- Amazon Relational Database Service (RDS)
- AWS Resource Groups & Tag Editor
- AWS RoboMaker
- Amazon Route 53
- AWS Secrets Manager
- AWS Security Hub
- AWS Security Token Service (STS)
- AWS Serverless Application Repository (SAR)
- AWS Service Catalog
- Amazon Simple Notification Service (SNS)
- Amazon Simple Queue Service (SQS)
- Amazon Simple Storage Service (S3)
- Amazon Simple Storage Service Glacier (S3 Glacier)
- Amazon Simple Workflow Service (SWF)
- AWS Step Functions
- AWS Storage Gateway
- AWS Well-Architected Tool
- AWS X-Ray

AWS Academy Lab Project - Cloud Data Pipeline Builder

- Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Both Learner Lab & Lab Project - Cloud Data Pipeline Builder

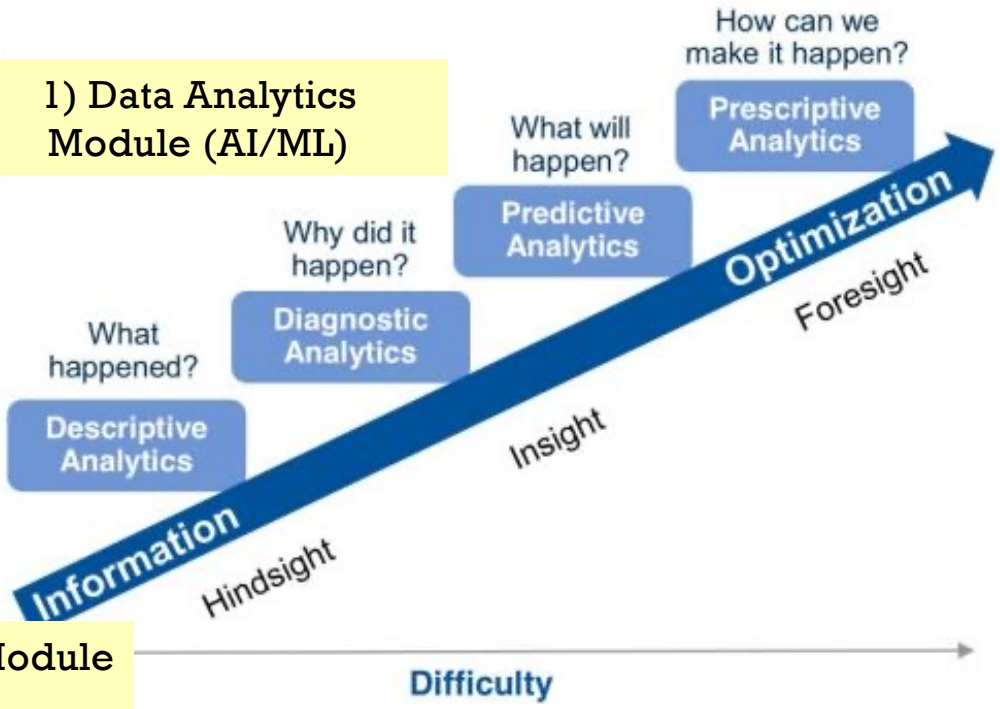
- Amazon SageMaker
- Amazon Elastic Compute Cloud (EC2)
- Amazon DynamoDB
- AWS Lambda
- Amazon Kinesis Video Streams
- Amazon Rekognition

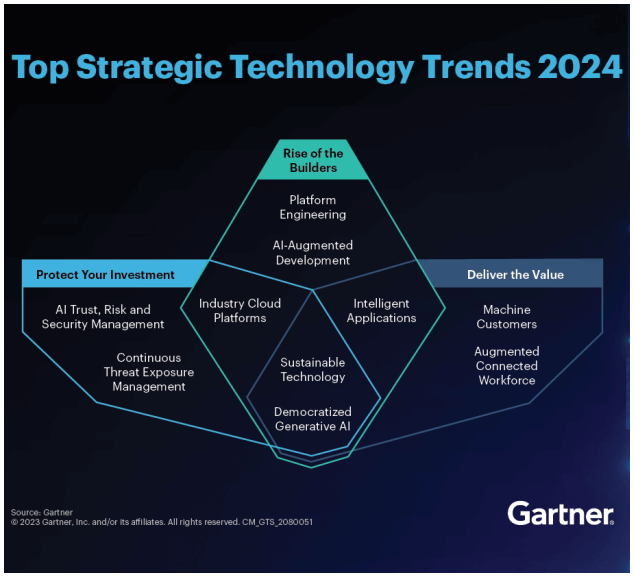
Conclusion

1) Data Analytics Module (AI/ML)

3) Data Vizualiation Module

2) Data Engineering Module (Data Pipeline)





+

Any questions? ☺