



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

Introduction to Data Science

2110531 Data Science and Data Engineering Tools (2024/1)

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)

The Economist Topics ▾ Current edition More ▾

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules



Print edition | Leaders >
May 6th 2017

[Twitter](#) [Facebook](#) [LinkedIn](#) [Email](#) [Print](#)

- 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)

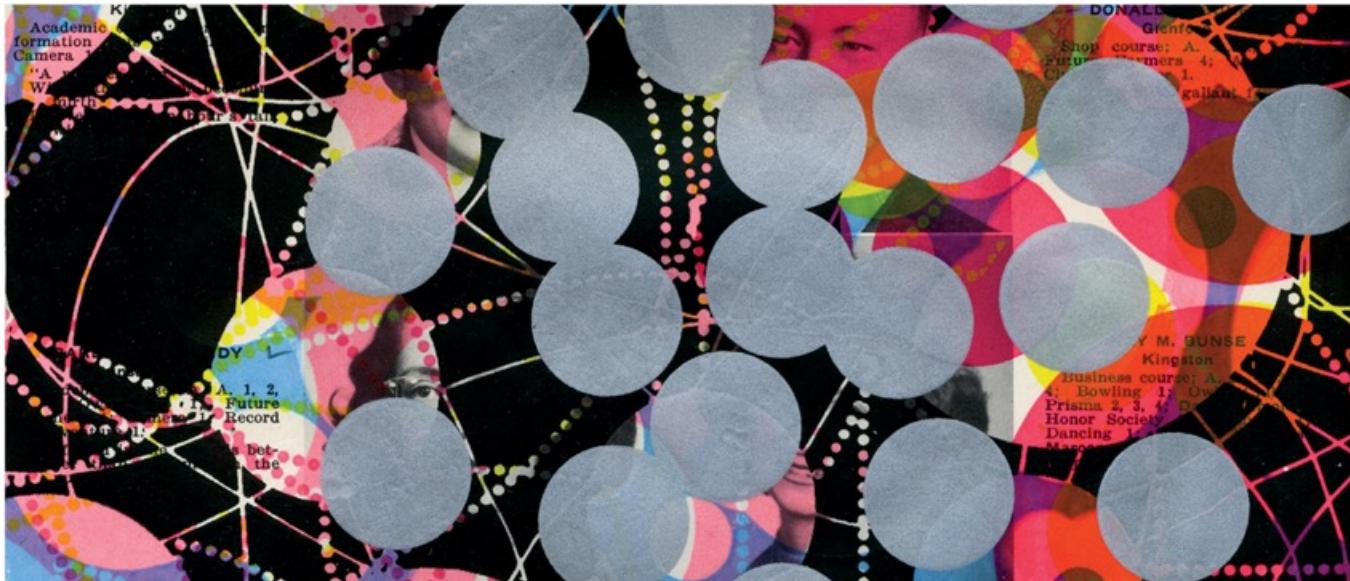
Data Science
(AI,ML,DM)
+
Big Data

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



Who analyzes these data!

Harvard
Business
Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

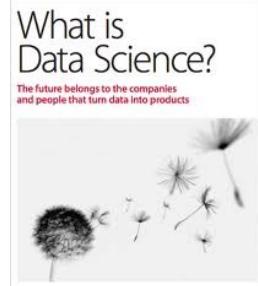
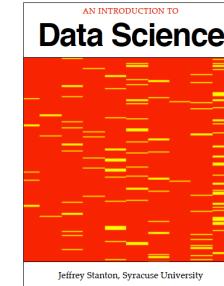
WHAT TO READ NEXT



Competing on Analytics

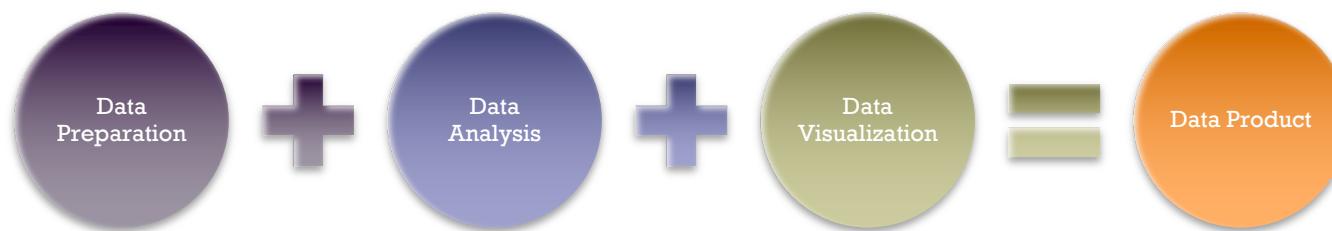


What is Data Science?



7

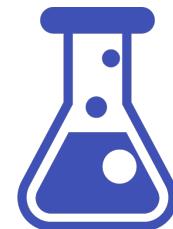
- Data
 - Facts and statistics collected for reference or analysis
- Science
 - A systematic study through observation and experiment
- Data Science
 - The scientific exploration of data to extract meaning or insight,
 - and the construction of software to utilize such insight in a business context.





What is Data Science? (cont.)

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



Ta Virot Chiraphadhanakul
Data Scientist, Facebook

Code Mania 2 (01), Jan-2015



1) Transform data into valuable insights

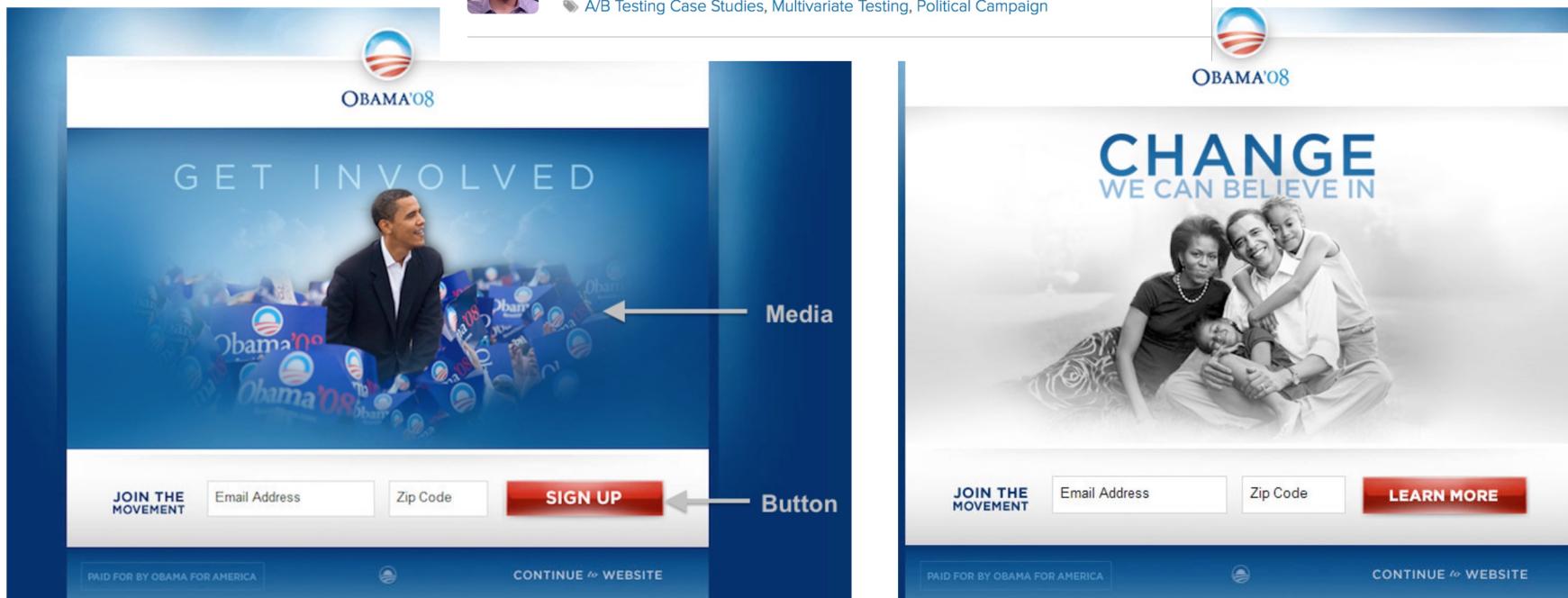
How Obama Raised \$60 Million by Running a Simple Experiment



By Dan Siroker

November 29, 2010

• A/B Testing Case Studies, Multivariate Testing, Political Campaign





1) Transform data into valuable insights (cont.)

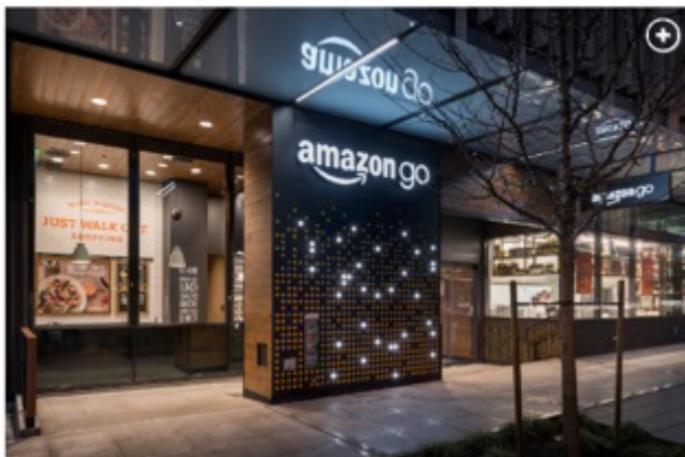


BUSINESS

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
Amazon

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





2) Transform data into data products



Action required: Please confirm activity.



FRAUD PROTECTION SERVICES

Chase Sapphire
Account Ending: XXXX

We want to help keep your account secure so we continuously monitor it for possible fraudulent activity. We're writing to verify whether the transaction below was authorized by you or another Cardmember. Click YES below if you

The screenshot shows the Microsoft Outlook interface. The top navigation bar includes 'Outlook', a search bar, and various action buttons like 'New message', 'Empty folder', 'Mark all as read', and 'Undo'. On the left, a sidebar lists 'Favorites' and 'Folders'. The 'Inbox' folder is selected, showing 45 items. A red box highlights the 'Junk Email' folder, which contains 128 items. The main pane displays several email messages from spam sources, such as 'Work At Home Opportunities', 'NETFLIX SURVEY', and 'Thank You Costco'.

From	Subject	Date
W	Work At Home Opportunities New work from home progr...	1:47 PM
CS	Client service NETFLIX SURVEY	1:40 PM
TC	Thank You Costco Re: Costco Has a Surprise Fo...	12:01 PM
CS	Client service - Are you a friend of Amazo...	8:43 AM



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

12

The Billion Prices Project

Home Our Public Data Our Research News

THE BILLION PRICES PROJECT

AN ACADEMIC INITIATIVE TO IMPROVE INFLATION MEASUREMENT

RESEARCH PAPERS DOWNLOAD DATA

<http://www.thebillionpricesproject.com/>





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

14

Alberto Cavallo

MIT and NBER

Roberto Rigobon

MIT and NBER

This Version: April 8, 2016

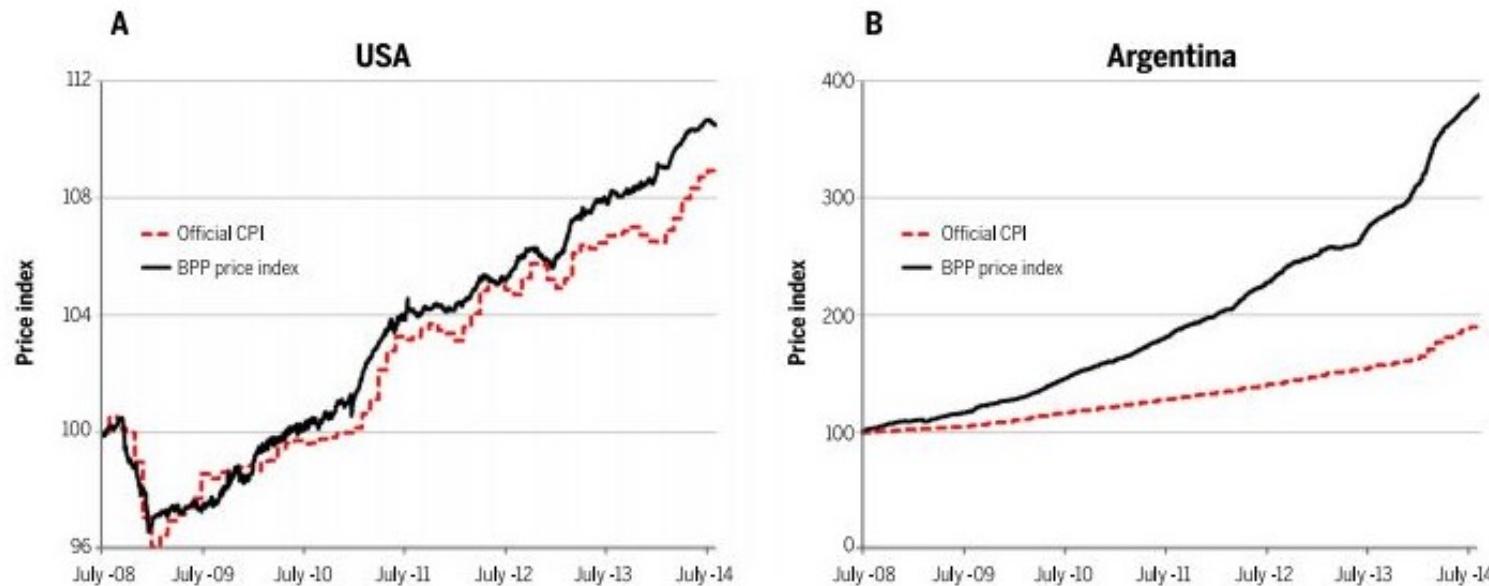


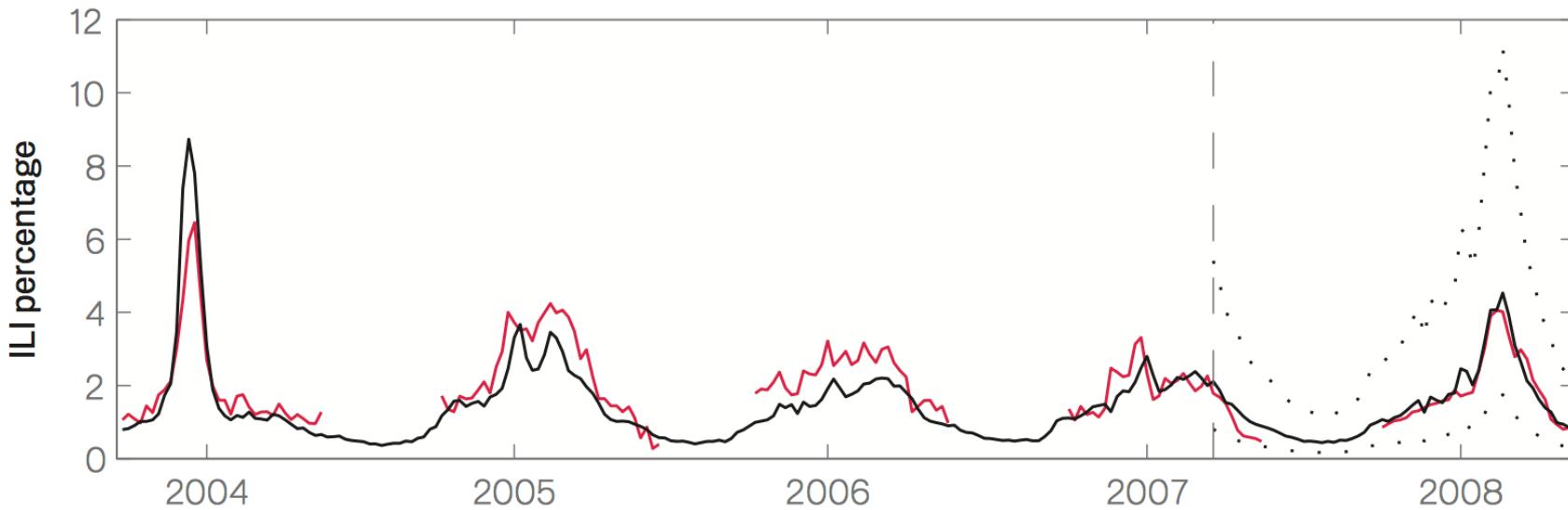
Fig. 2. BPP price index. Dashed red lines show the monthly series for the CPI in the United States (A) and Argentina (B), as published by the formal government statistics agencies. Solid black lines show the daily price index series, the "State Street's PriceStats Series" produced by the BPP, which uses 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

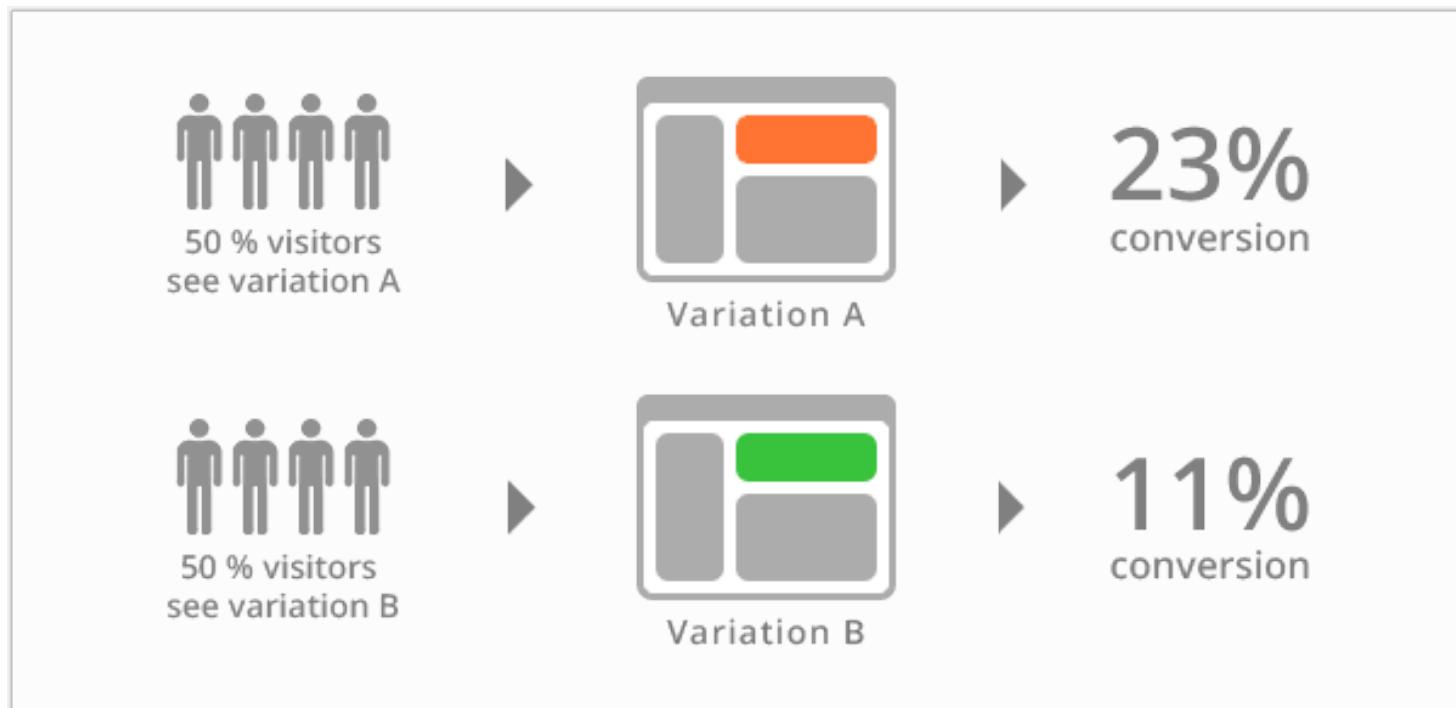


Why do we need to measure?

- 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

20

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

Control

The screenshot shows the SimCity website's homepage. A large hero banner at the top features the text "PRE-ORDER AND GET \$20 OFF YOUR NEXT PURCHASE" over a background image of the New York City skyline. Below the banner, there are two main product sections: "SIMCITY™" and "SIMCITY™ DIGITAL DELUXE EDITION". Each section includes a thumbnail image, the game title, the price (\$59.99 or \$79.99), and purchase options (PC Download or PC Physical). A red box highlights the hero banner area.

Variation

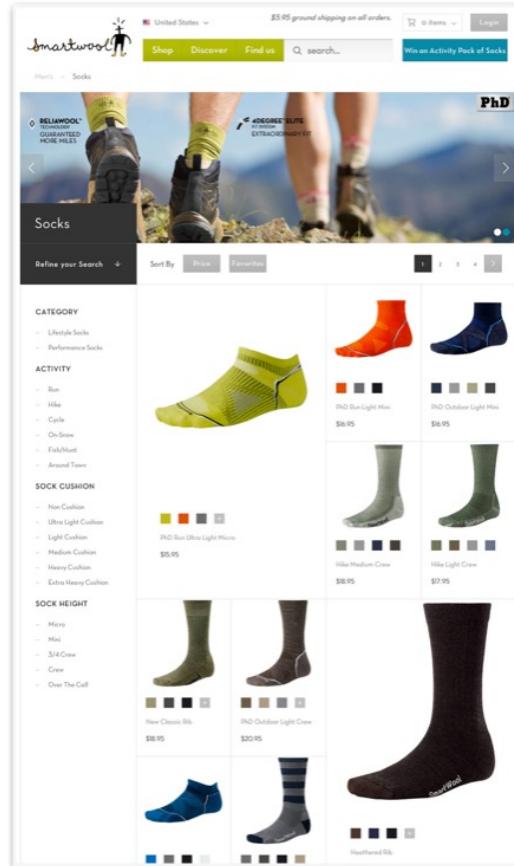
The screenshot shows the same SimCity website, but the hero banner from the control version has been removed. Instead, there is a large blue circular callout in the bottom right corner containing the text "43% increase in checkouts". The main content below remains the same, featuring the "SIMCITY™" and "SIMCITY™ DIGITAL DELUXE EDITION" sections, along with the "Key Features" and "WHAT IS SIMCITY?" sections at the bottom.

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



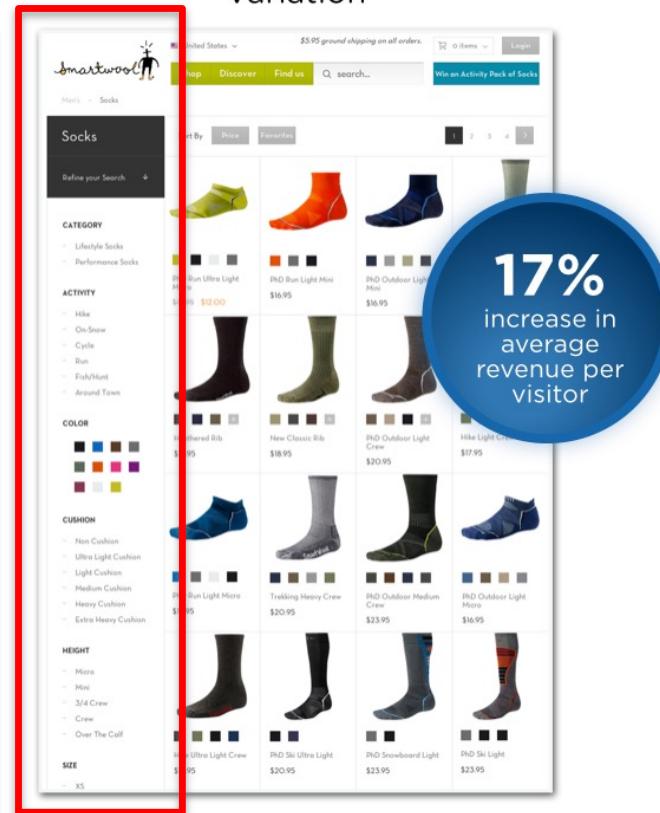
Example: SmartWool

Control



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

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

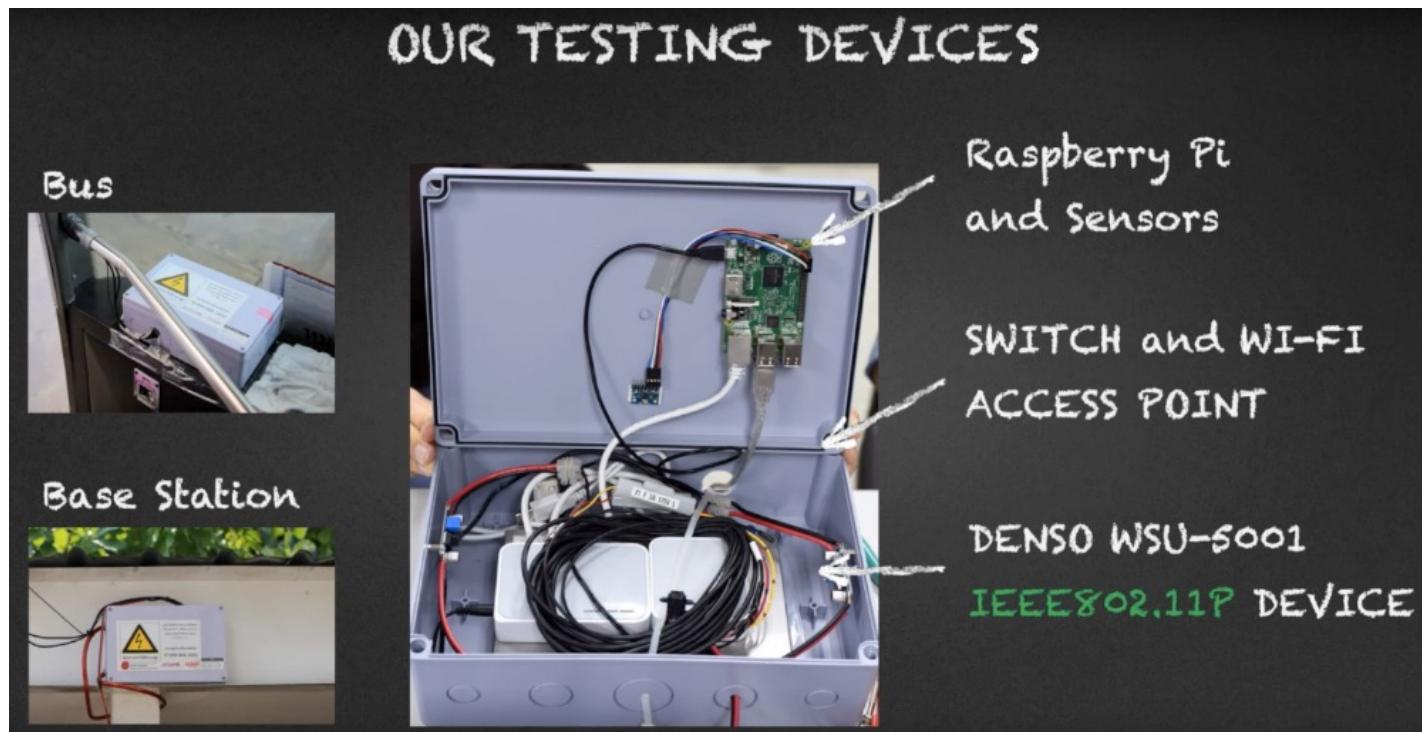


Example: Tracing Traffic





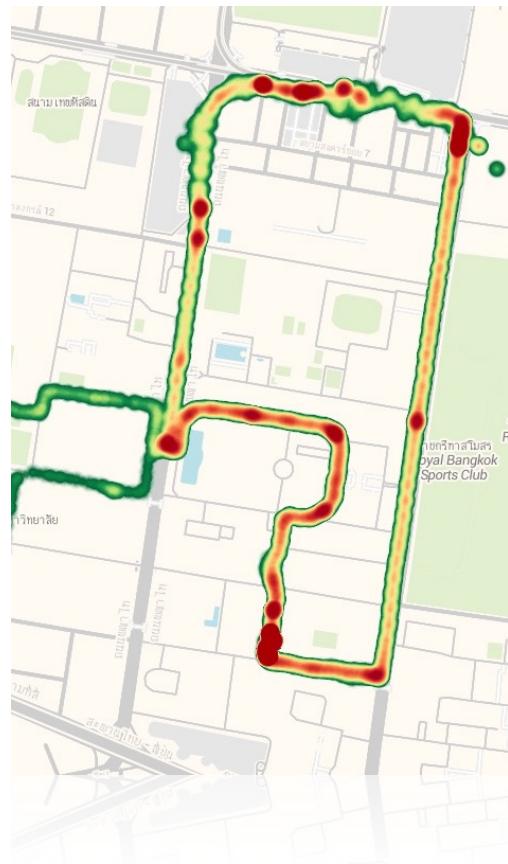
Example: Tracing Traffic





GPS Average Speed

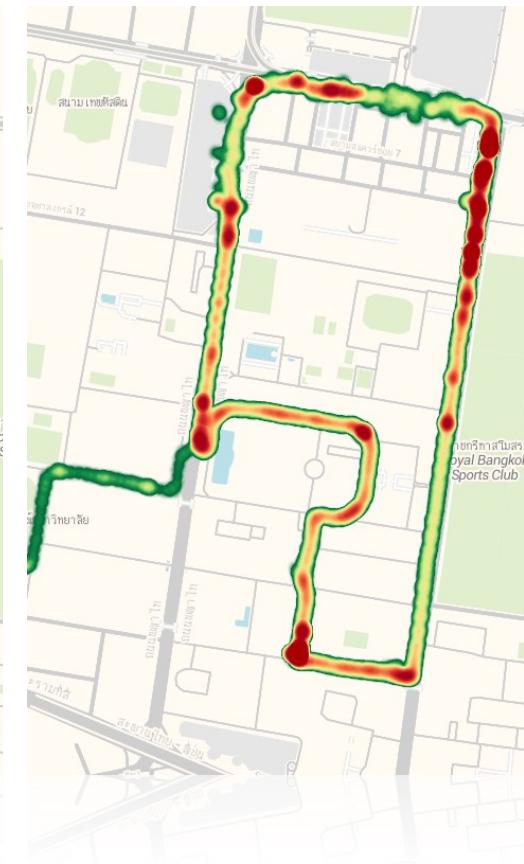
6:00-10:00



10:00-15:00



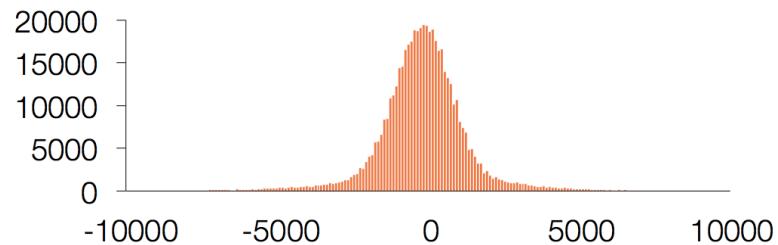
15:00-18:00



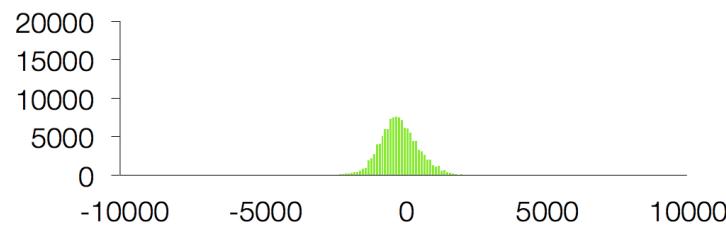
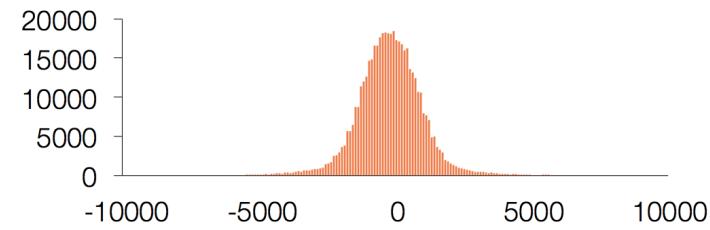
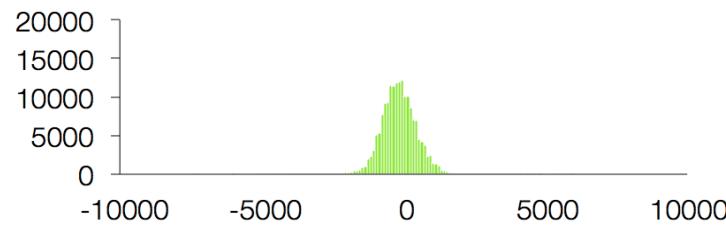
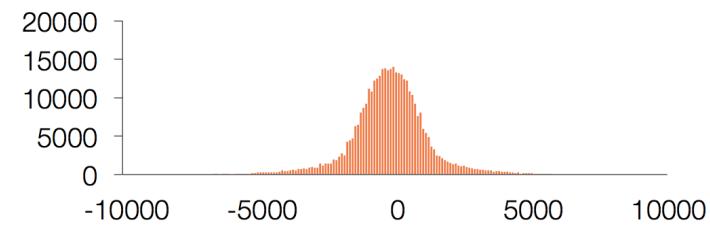
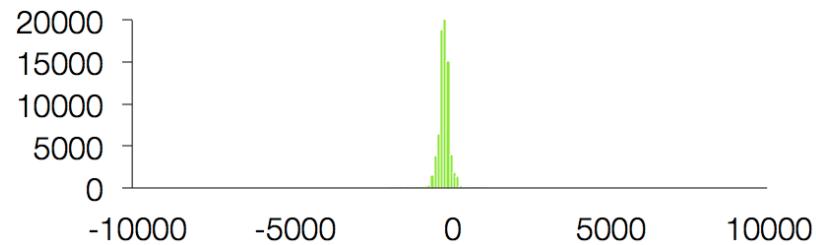


Bus Drivers' Behaviors

Bus A



Bus B





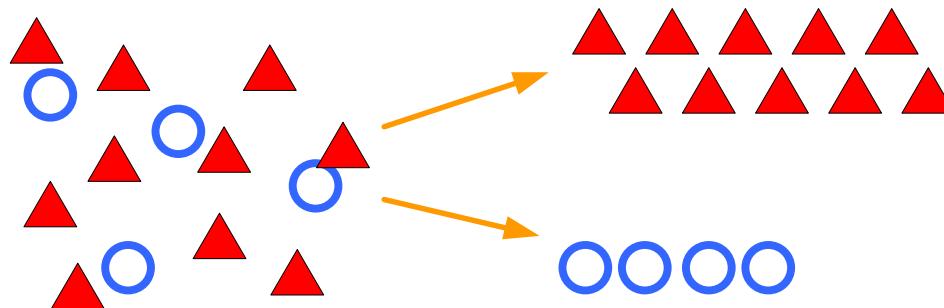
3) Data Products

- 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

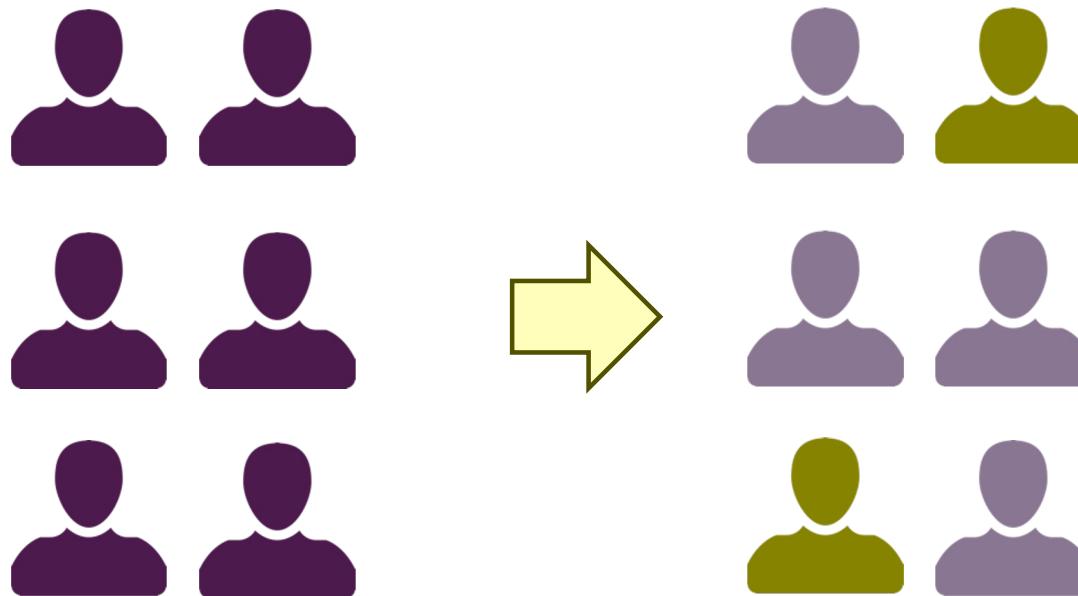
- Identify to which set of categories a new observation belongs

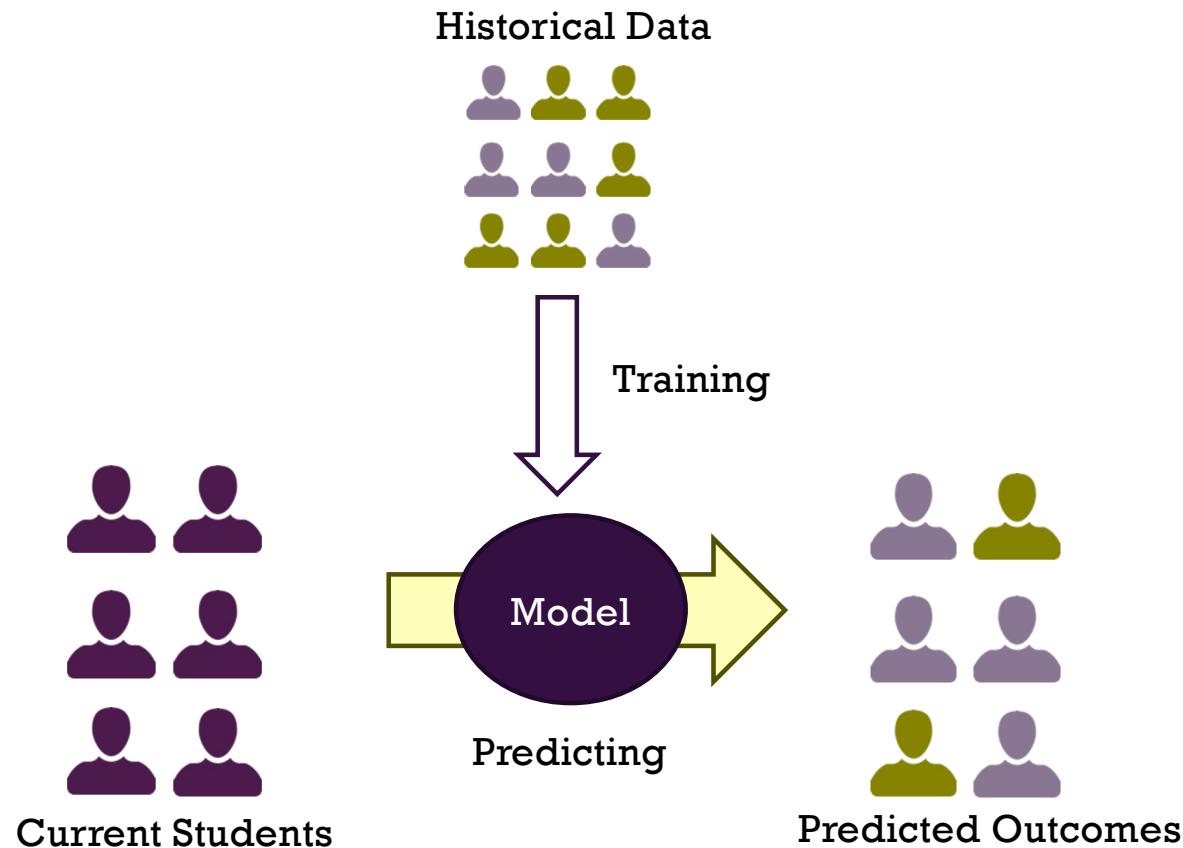


- Example: spam filtering, customer churn prediction, complaint classification

+

Example: Students Grade Prediction





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

Example: Amazon Recommendation

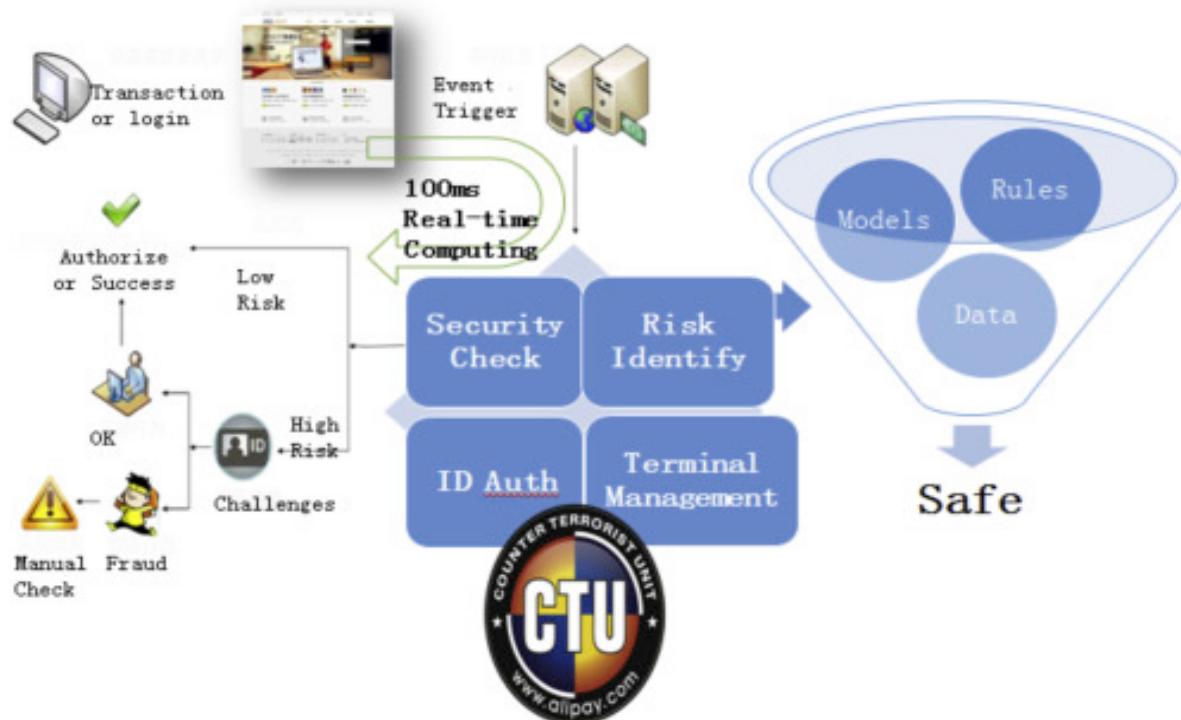
- Amazon sells 480M products (485k new products per day)
- Use recommendation systems to bring products to customers
- Analyze data from 300M customers
 - Purchase history
 - Reviews / Ratings
 - Search history
 - Views

The screenshot shows a portion of an Amazon website with a navigation bar at the top. Below the navigation, a banner for 'Natawut's Amazon' is displayed, along with a message encouraging users to sign in for order status and balances. The main content area features a grid of recommended products:

- Computer & Technology Books**: 92 items. Includes a large image of the book 'Hadoop Application Architectures' by Mark Grover, Ted Malaska, Jonathan Seidman, and Gwen Shapira.
- Science & Math Books**: 51 items. Includes a large image of the book 'Storytelling' by Robert J. Knell.
- Other recommendations**: Includes images of books like 'Own the Room' and 'Introductory Machine Learning'.



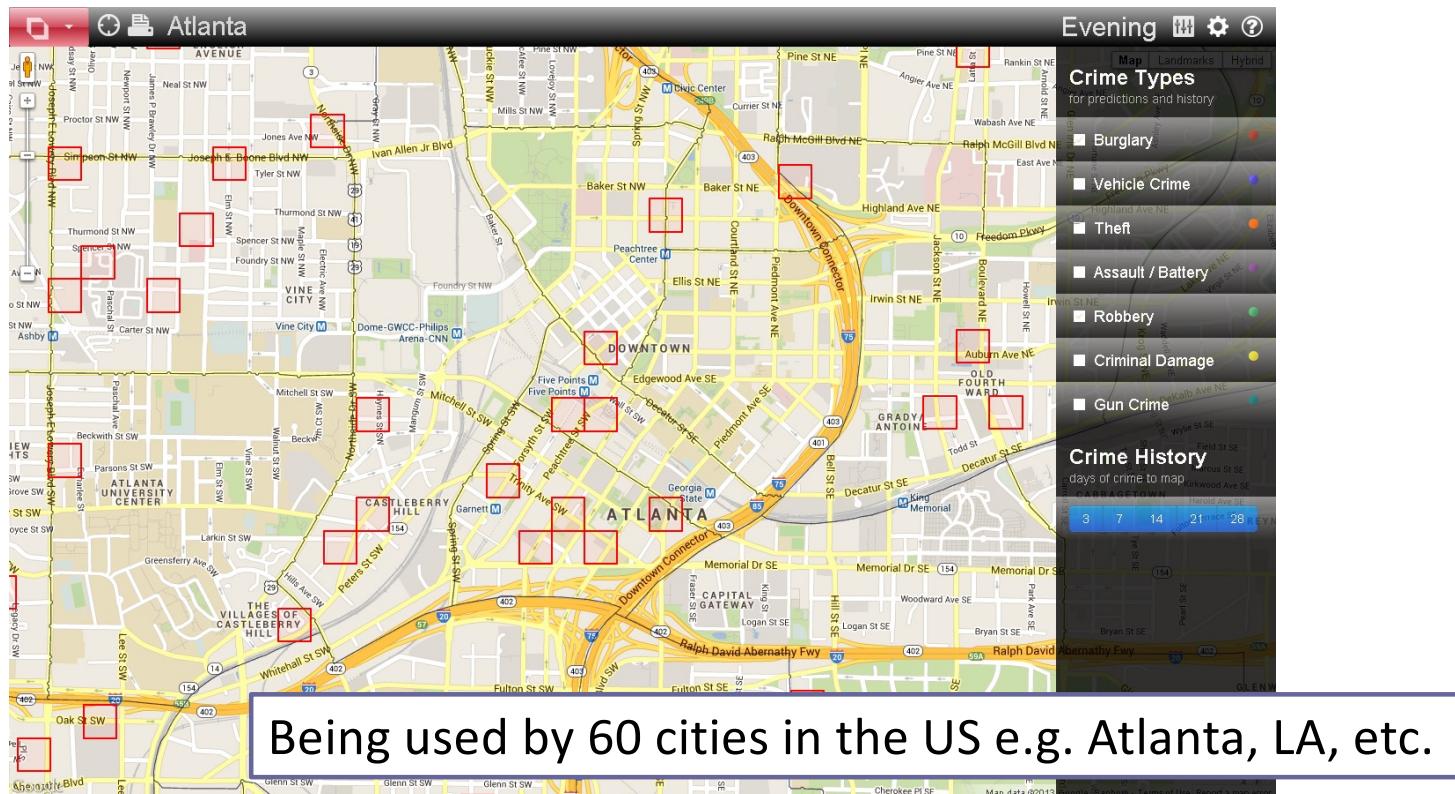
Case study: Alibaba Fraud Detection



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



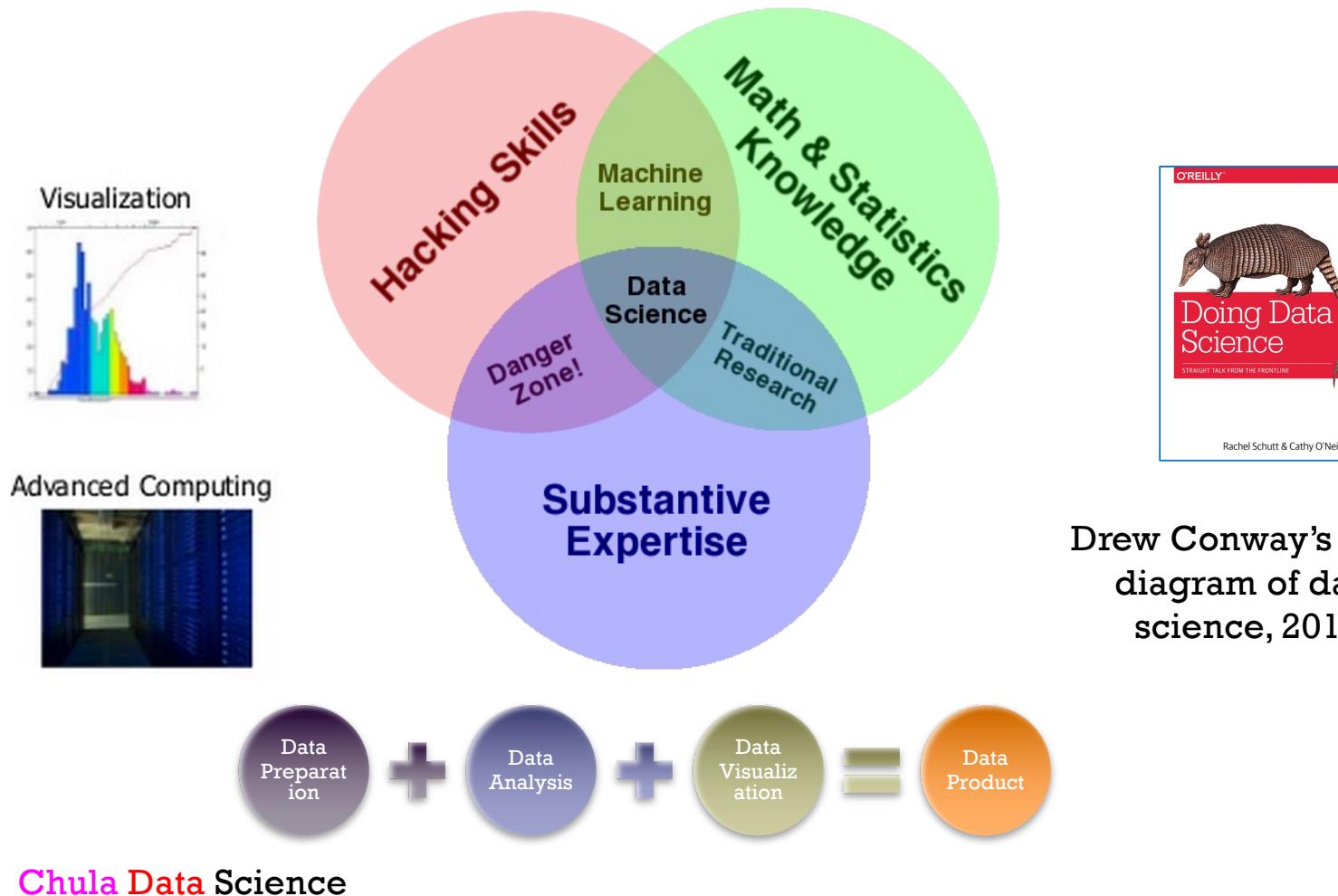
Case study: Predictive Policing

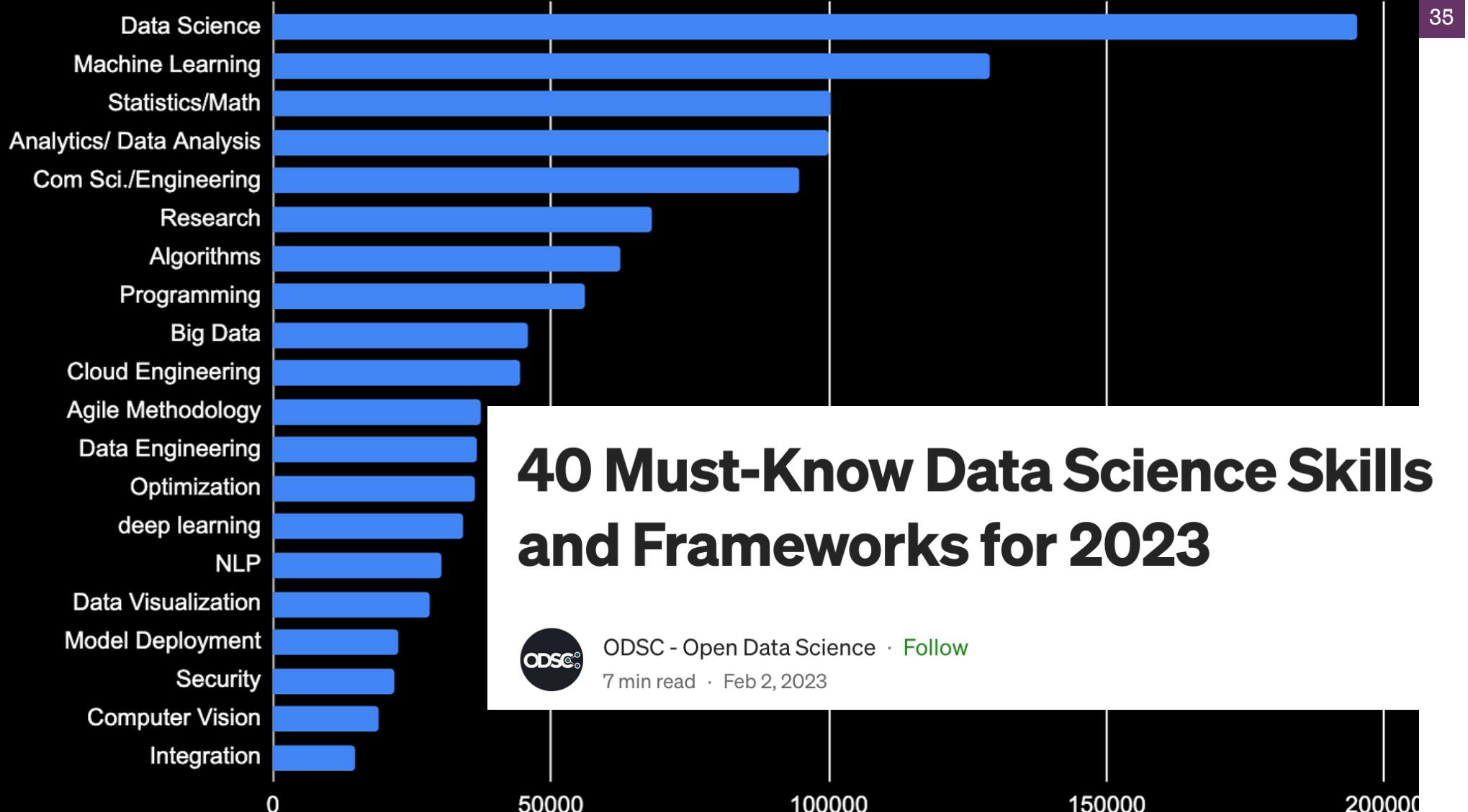


Source: <http://www.forbes.com/sites/ellenhuet/2015/02/11/predpol-predictive-policing>



Drew Conway's Data Science Venn diagram (Skills)





40 Must-Know Data Science Skills and Frameworks for 2023



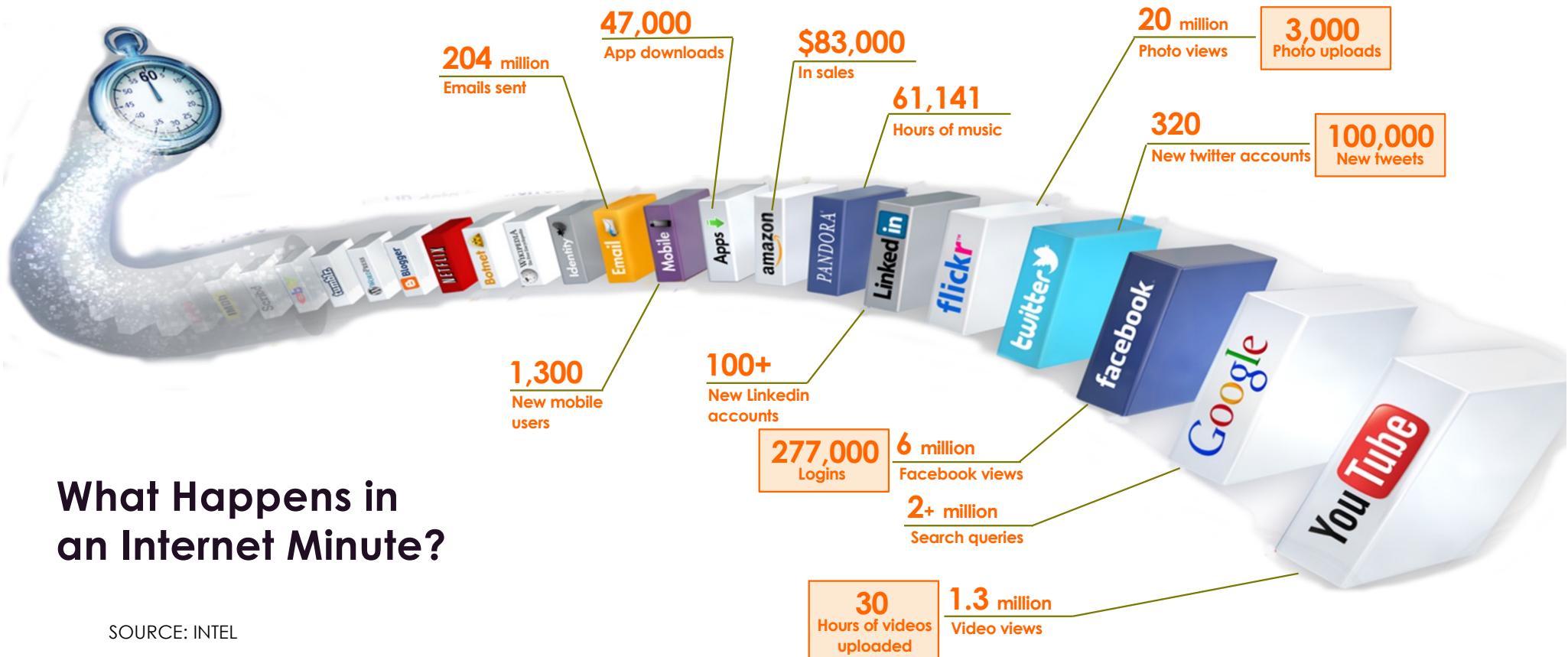
ODSC - Open Data Science · Follow

7 min read · Feb 2, 2023

+

Big Data

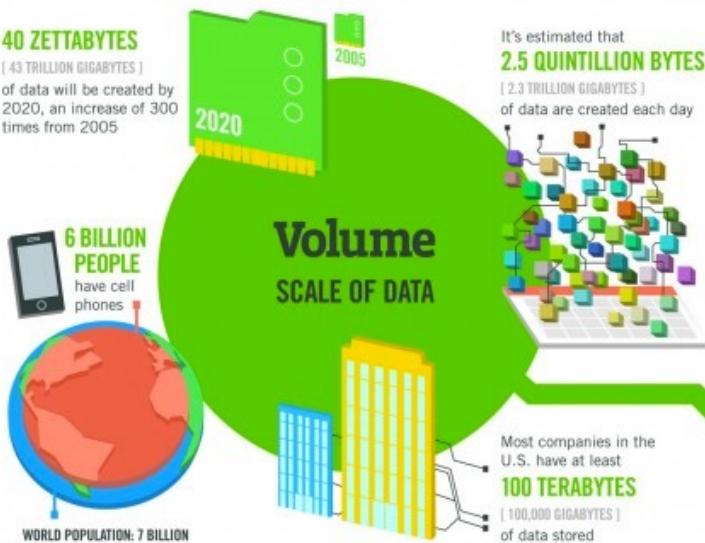
Big Data Explosion



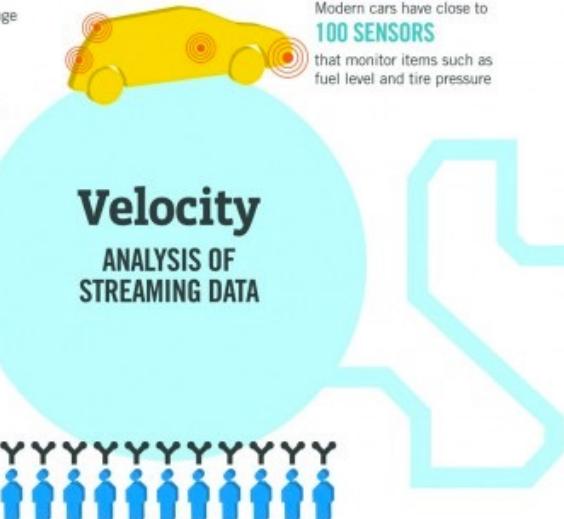
What Happens in
an Internet Minute?

SOURCE: INTEL

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by
2020, an increase of 300
times from 2005



The New York Stock Exchange captures
1 TB OF TRADE INFORMATION during each trading session



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

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MPTEC, QAS

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**.

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015, 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States.



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT

are shared on Facebook every month



Variety
DIFFERENT FORMS OF DATA

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

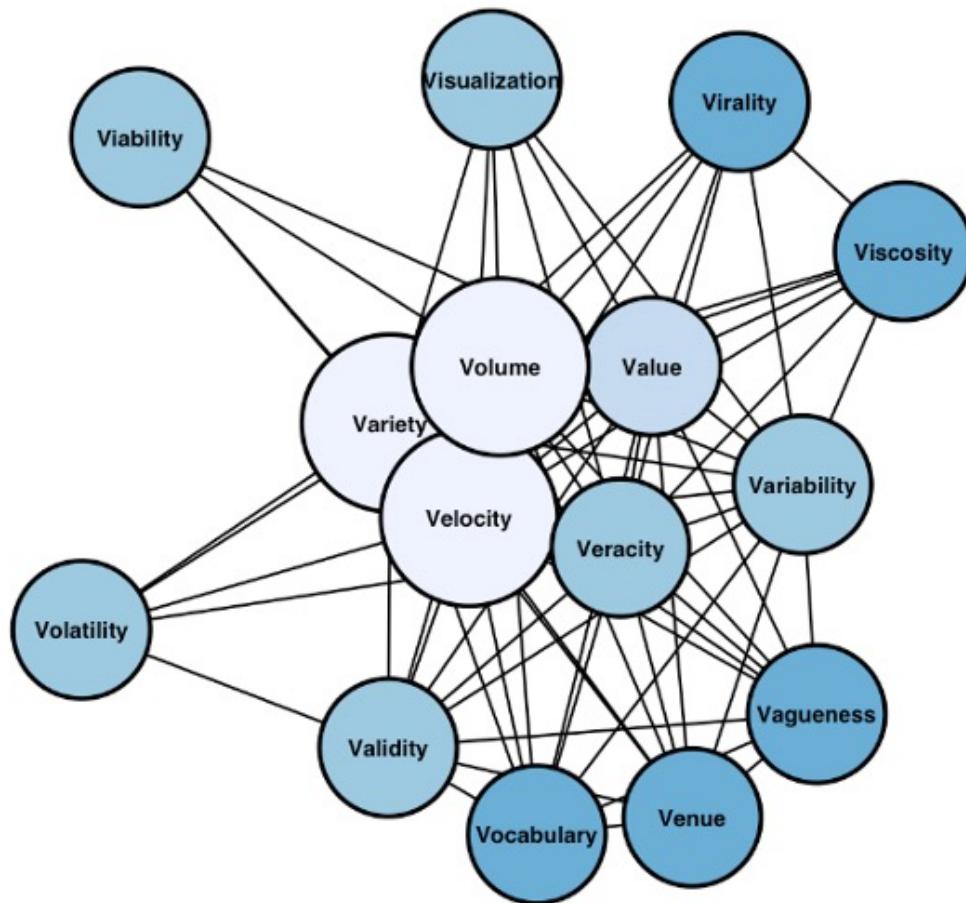


27% OF RESPONDENTS

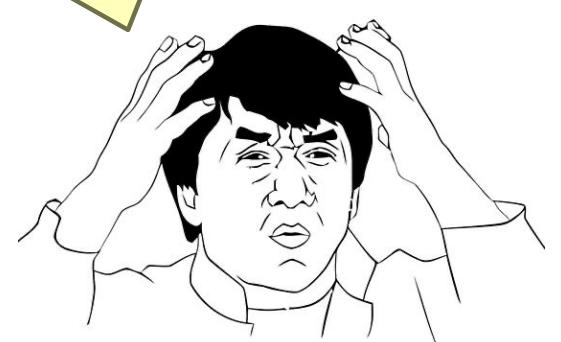
in one survey were unsure of how much of their data was inaccurate

Veracity
UNCERTAINTY OF DATA

Now 42 V of Big Data



42 V's?!?



Big Data Driver: Internal + External Data





KNOWING IS BEST

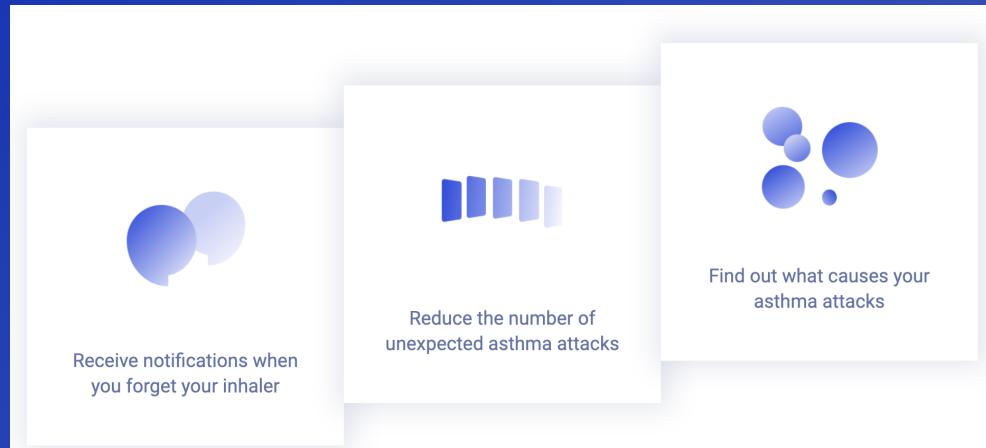
A Better Way to Know

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.

Every Beat. Every Breath.

Control your asthma with FindAir smart inhaler

Full Control. Less asthma attacks. Better life.



2nd prize
EIT HEALTH
INNOSTAR
AWARD



1st prize
UPC DIGITAL
IMAGINATION
CHALLANGE



Best Pitch
SMART
HEALTH
BUDAPEST



1st prize
INNOLABS
DIGITAL HEALTH
HACKATHON



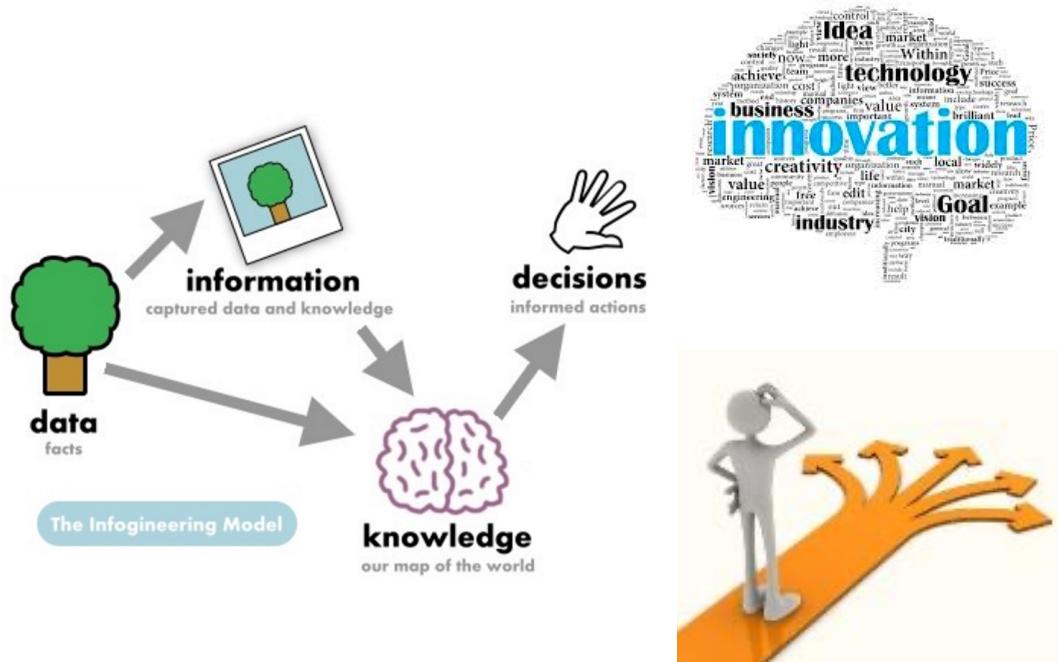
1st prize
INNOLABS
DIGITAL HEALTH
HACKATHON



1st prize
MEDTRENDS
TOP TRENDS

Big Data Analytics

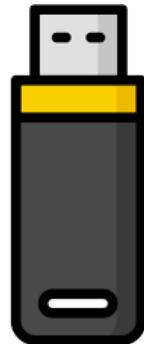
- It is a process of examining **Big Data** to uncover useful information and knowledge.
- More data means better decision!



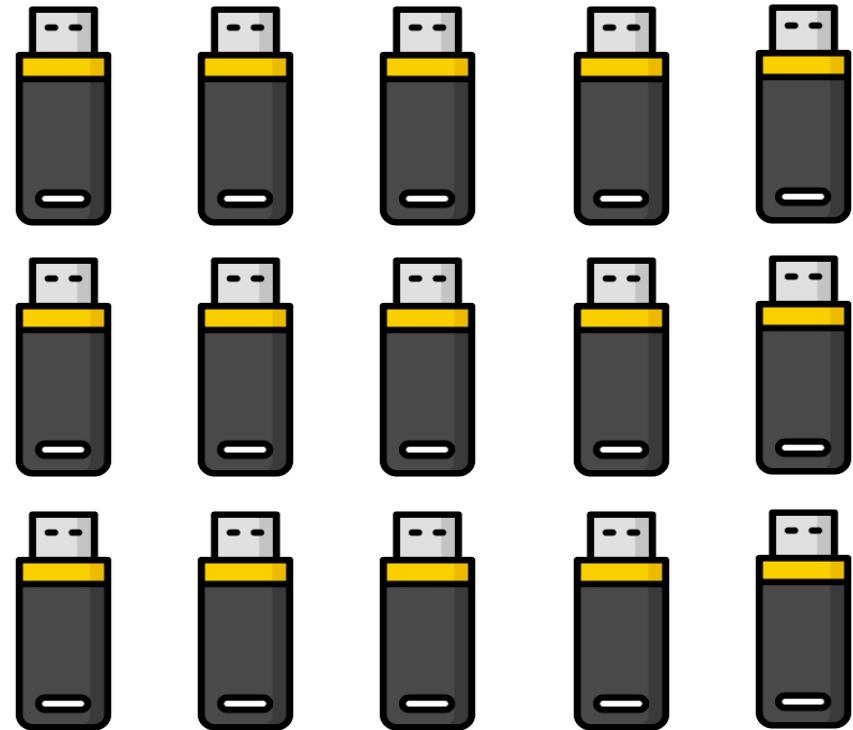
Big Data Challenges

Same tasks, but much more difficult!

2MB



200TB



Big Data Solution



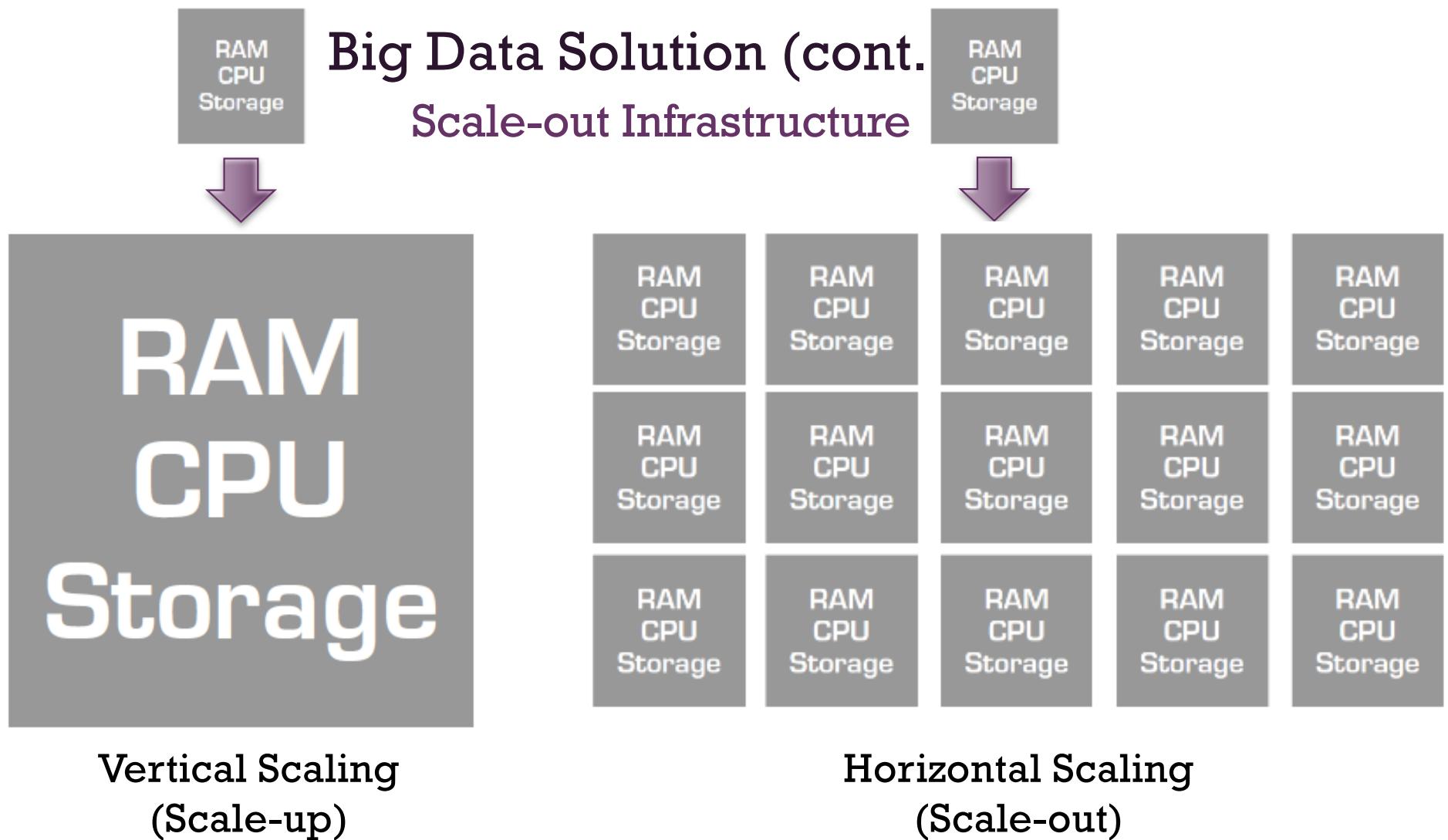
INFRASTRUCTURE



ALGORITHM

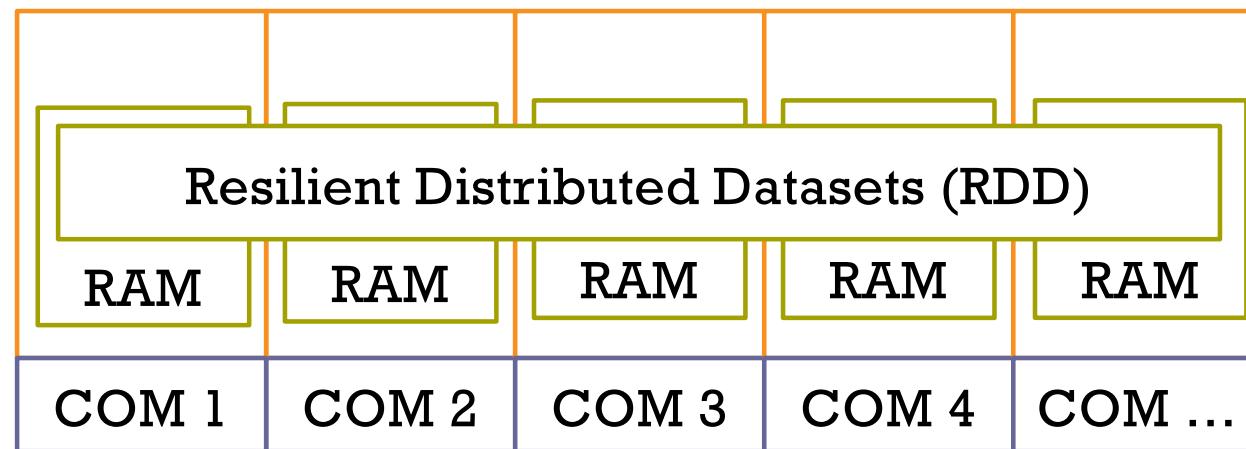
Big Data Solution (cont.)

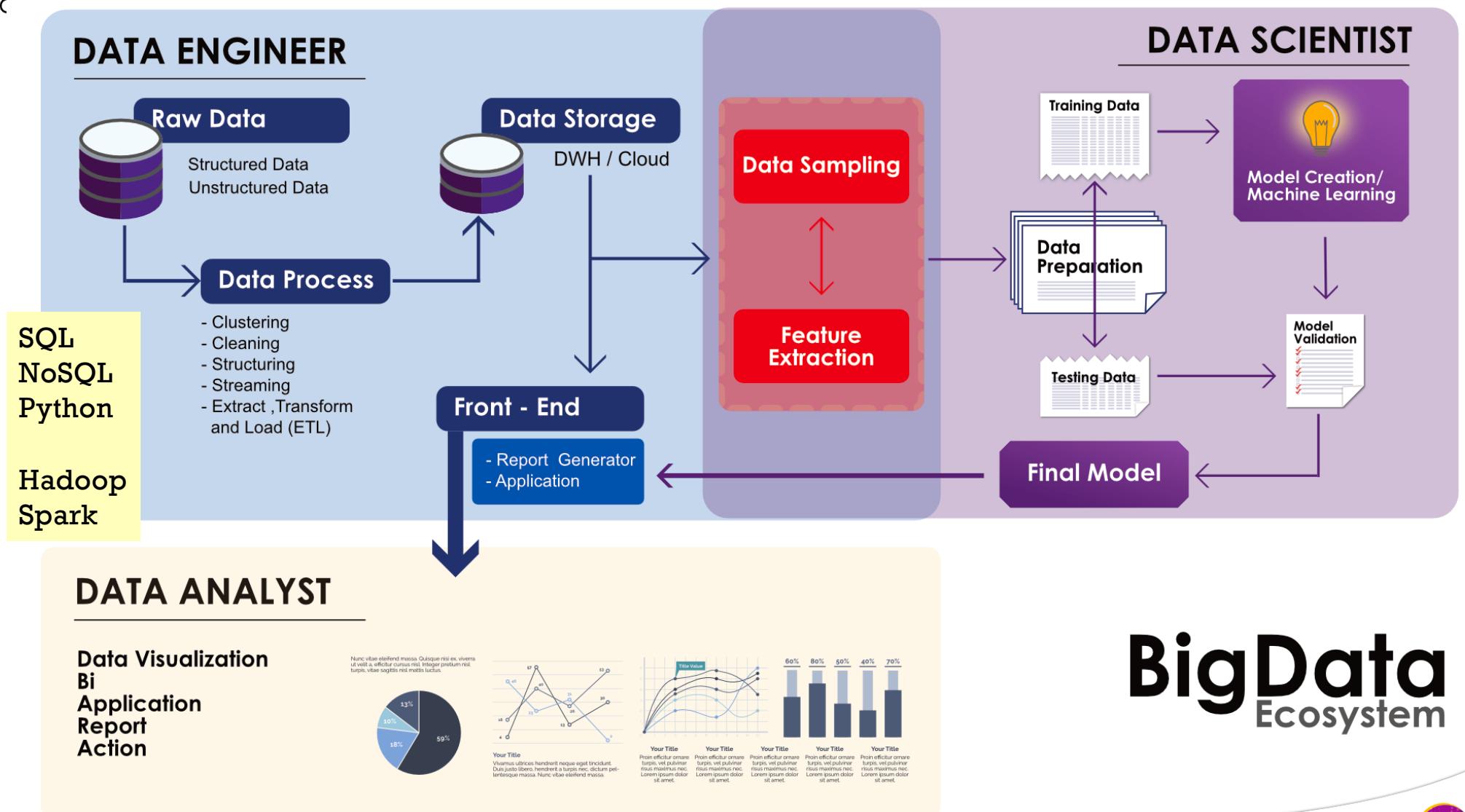
Scale-out Infrastructure



Big Data Solution (cont.)

In-memory & Distributed Computing





BigData Ecosystem

[LINK](#)



Business Intelligence By Coraline

<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 วันเดี๋ยวนี้ก็ล้มได้ครับ T_T)

Data Scientist + ML Engineer



Data Scientist

Datamites
Global Institute for Data Science

50

VS

Data Engineer



VS

ML Engineer



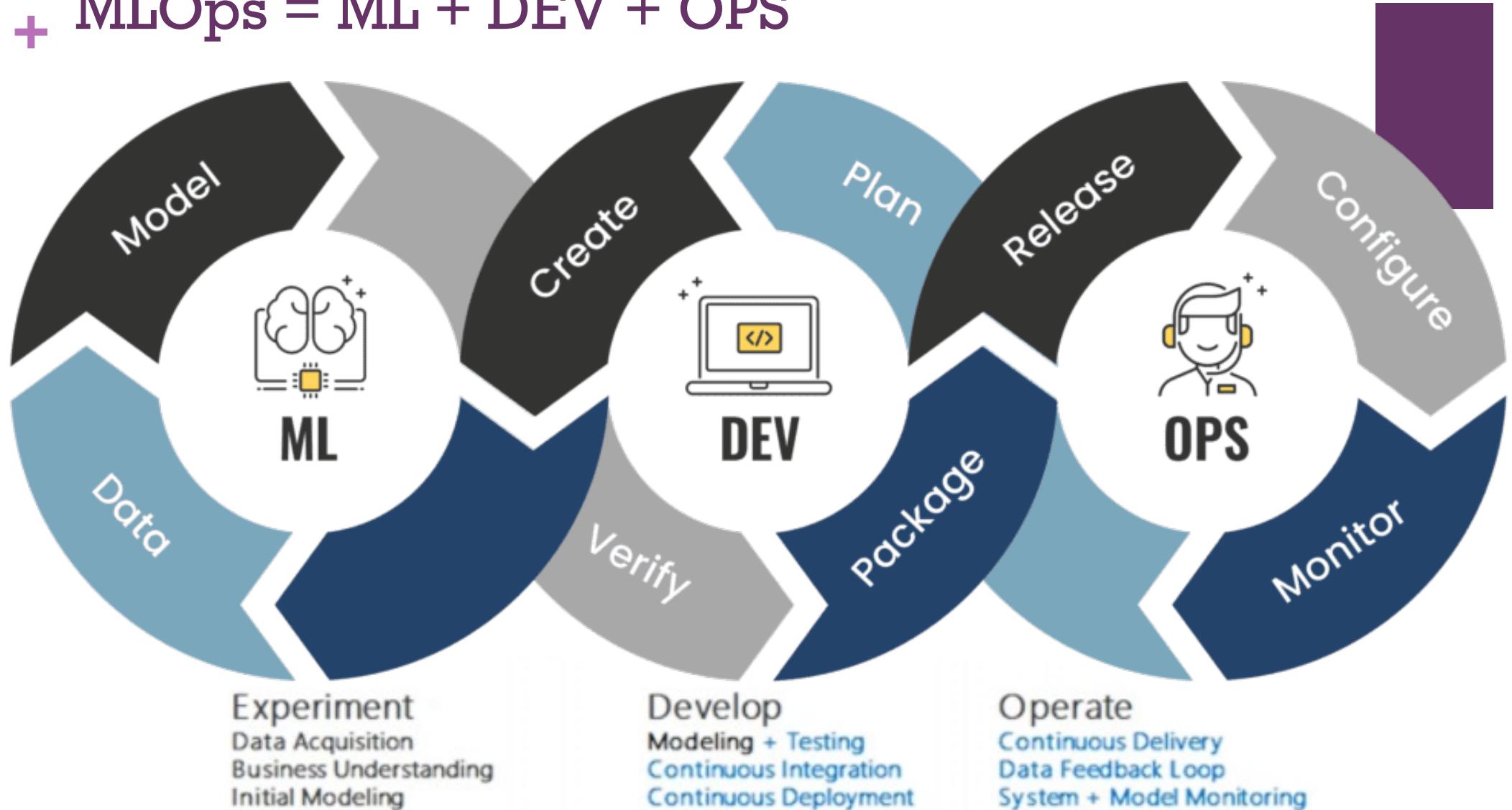
VS

MLOps Engineer



<https://vocal.media/education/data-scientist-vs-data-engineer-vs-ml-engineer-vs-ml-ops-engineer>
www.datamites.com

+ MLOps = ML + DEV + OPS





Data Science Process



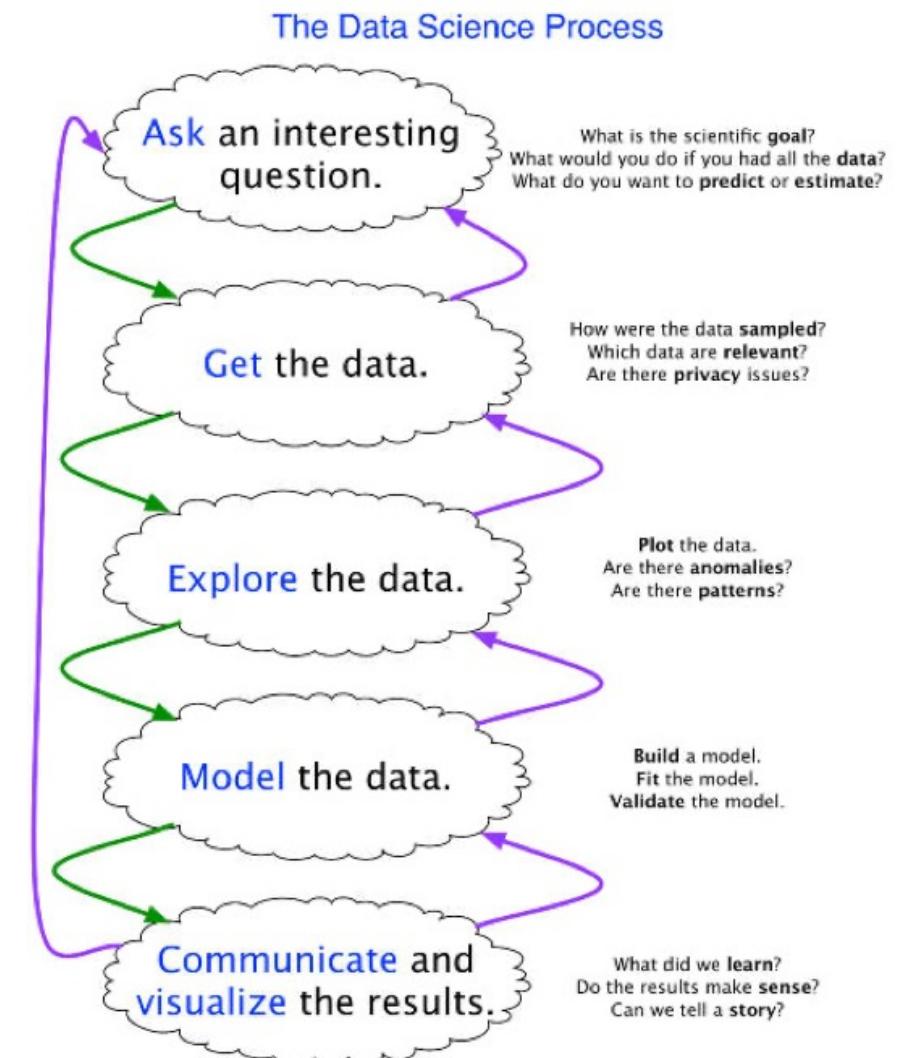
Data Science Process

Dr. Virote

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

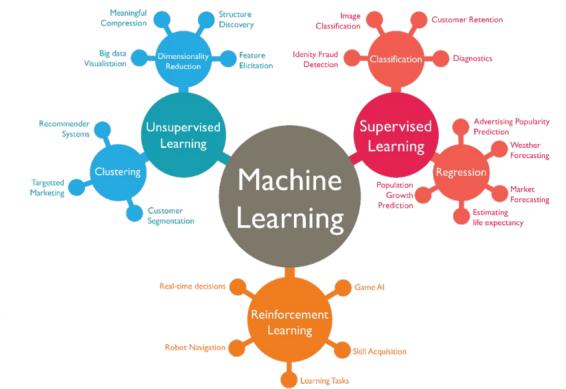
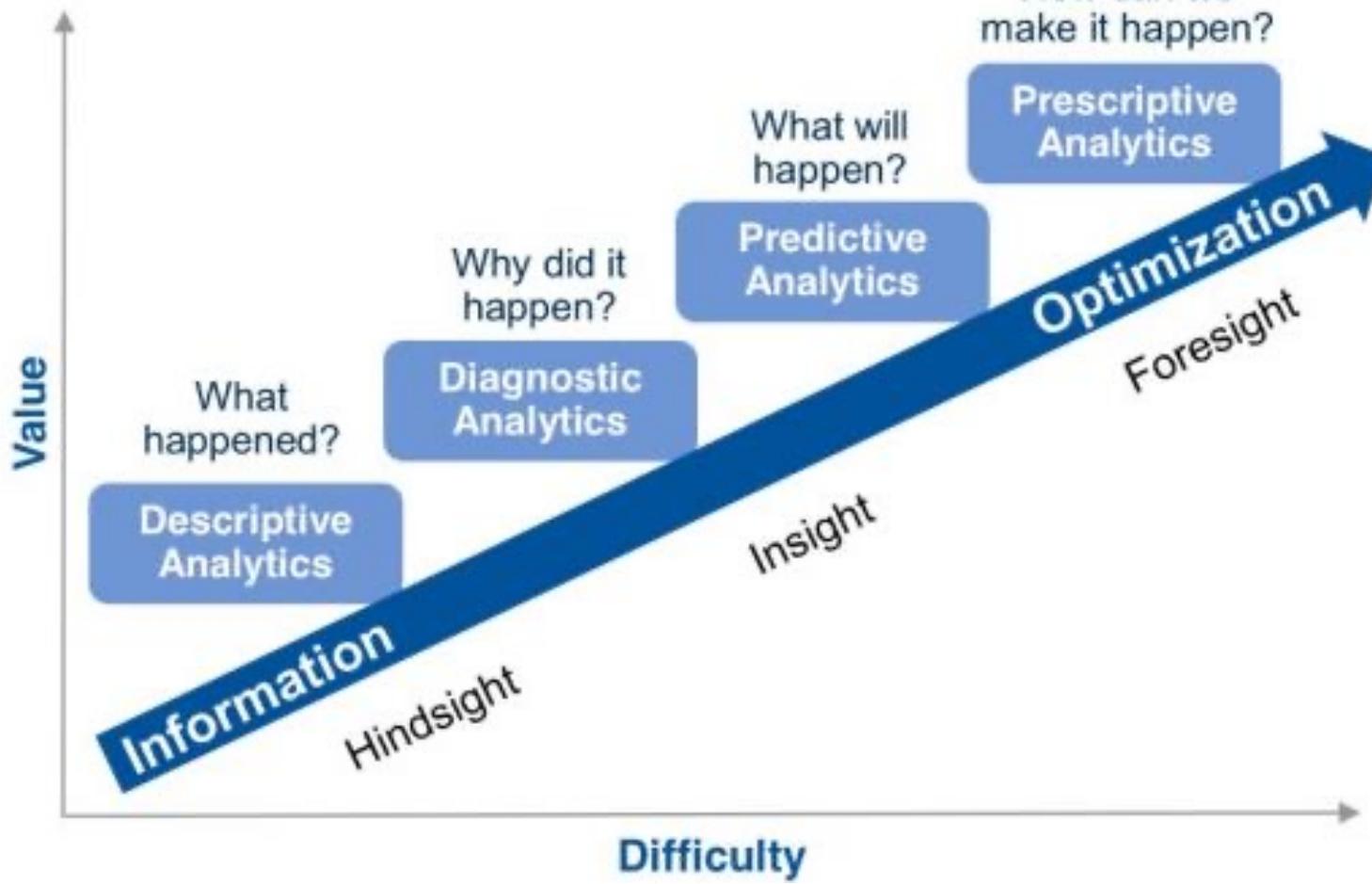
Aj.Natawut

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



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://cs109.org/>.

Data Analytics



BIG DATA





Types of Data Science Projects

Valuable insights

- Data visualization
- Analytical skills & storytelling
- Infographic

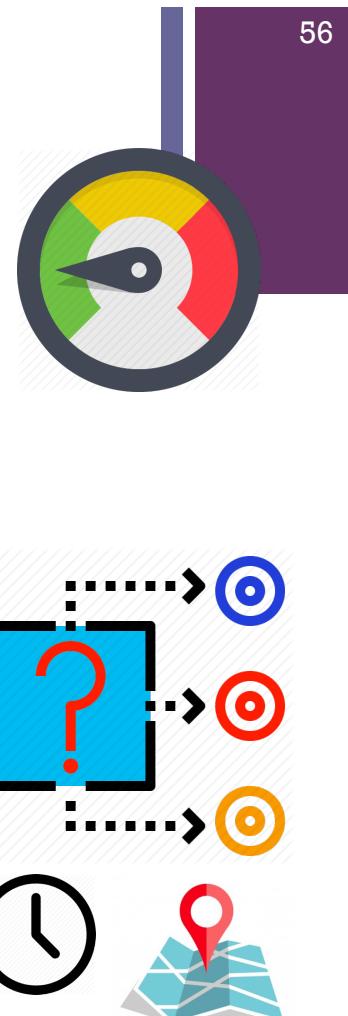
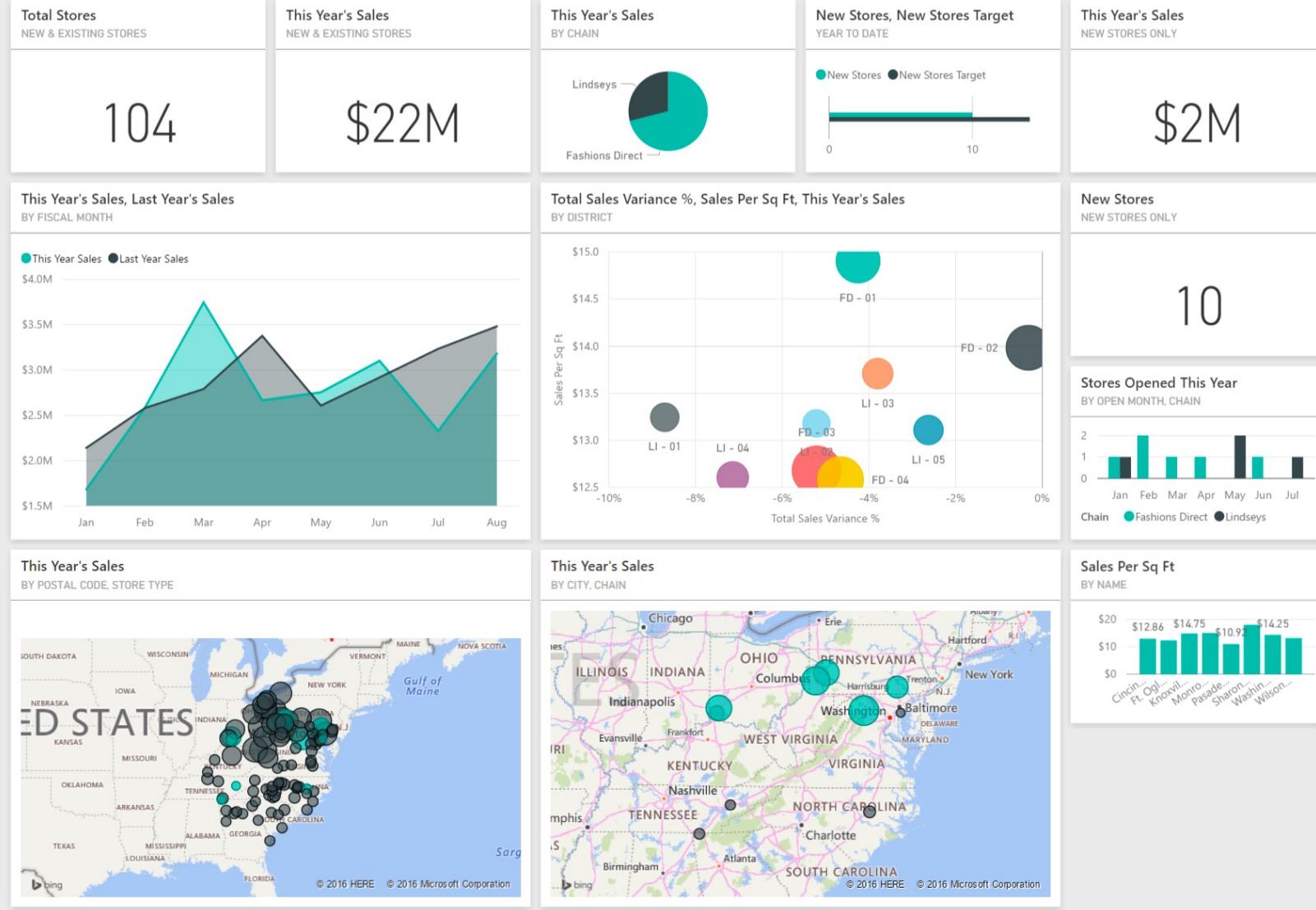


Advanced analytics

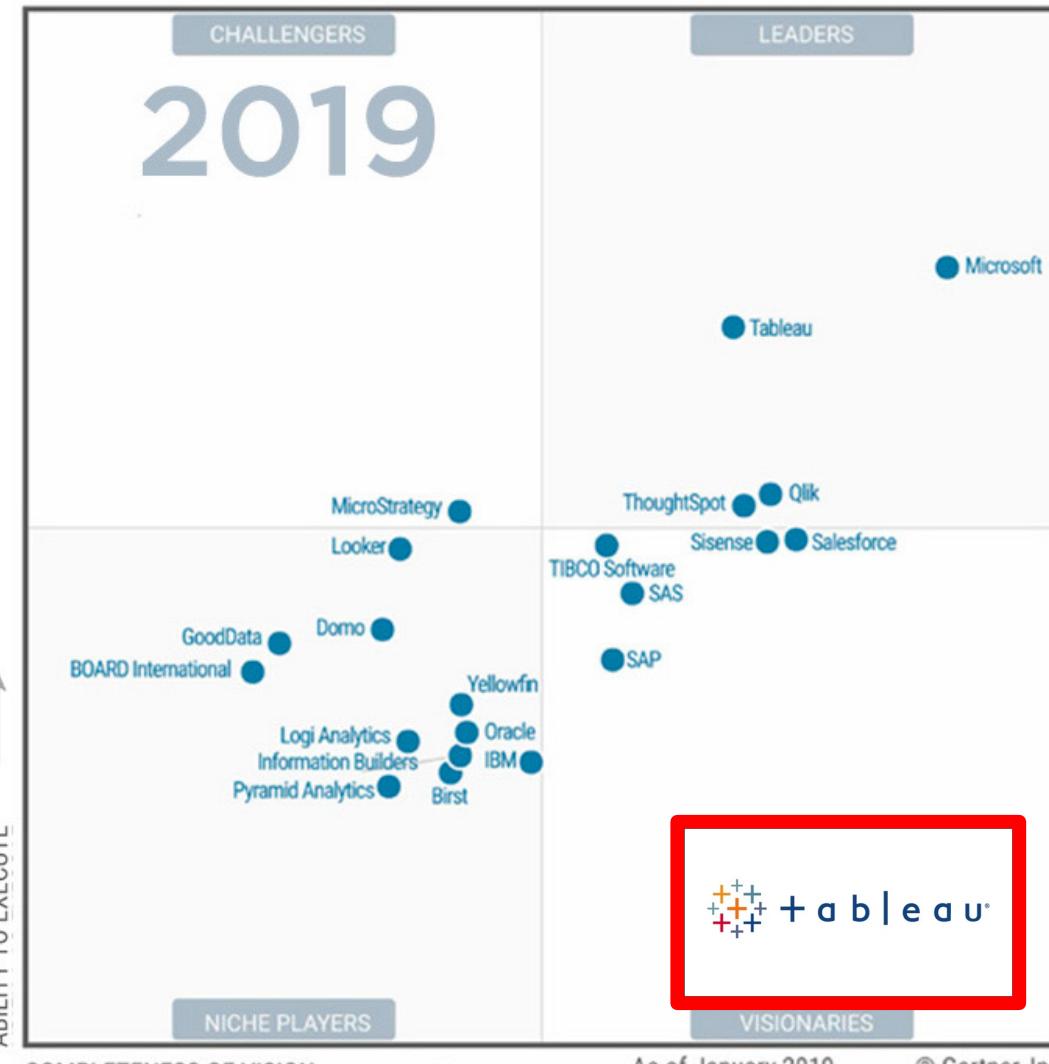
- AI/Machine Learning/Deep Learning
- Prediction, Forecasting, Clustering, etc.



Ask a question about your data



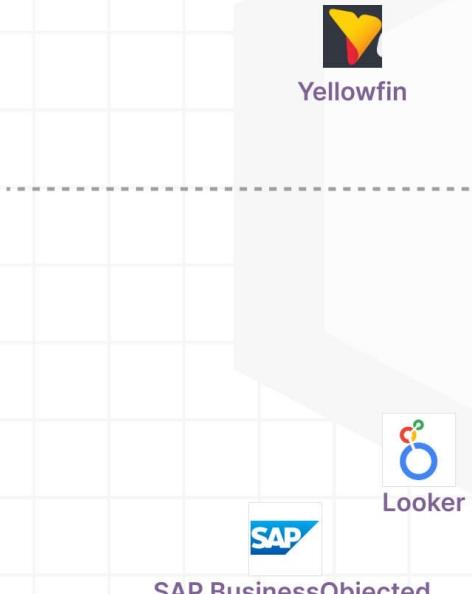
Magic Quadrant for Analytics and Business Intelligence Platforms.



Source: Gartner (Feb 2019 and 2020)



PRODUCT INNOVATOR



CHALLENGER

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



LEADER

58

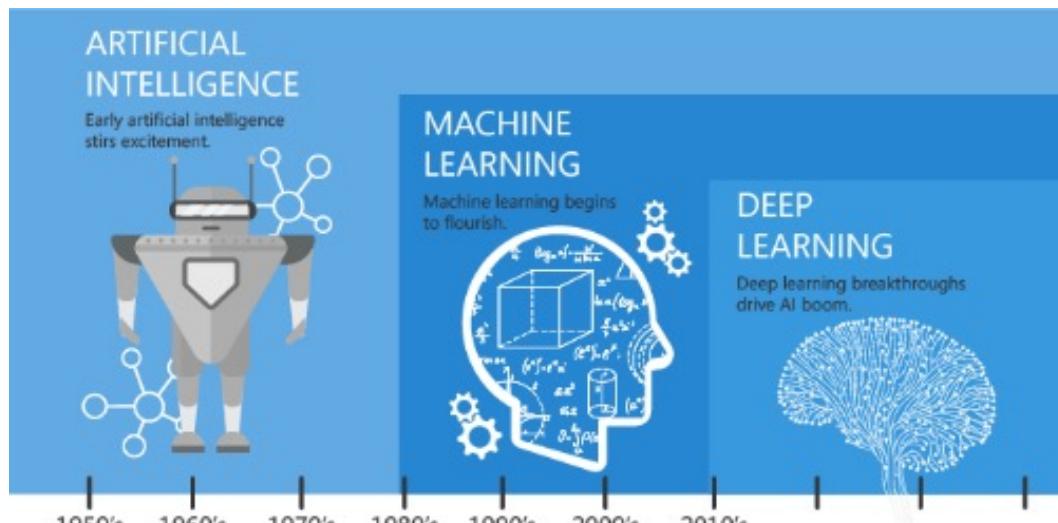
SERVICE STAR



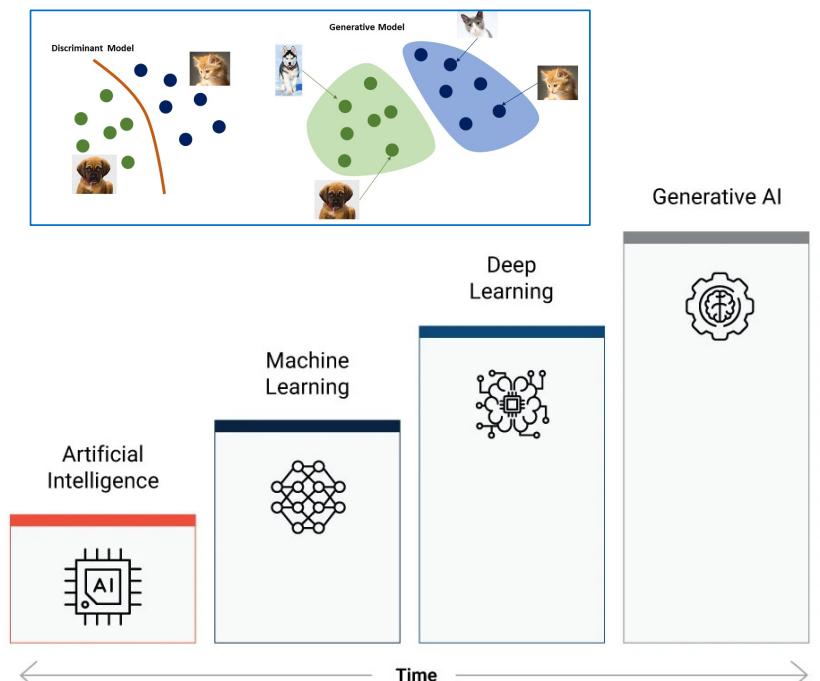
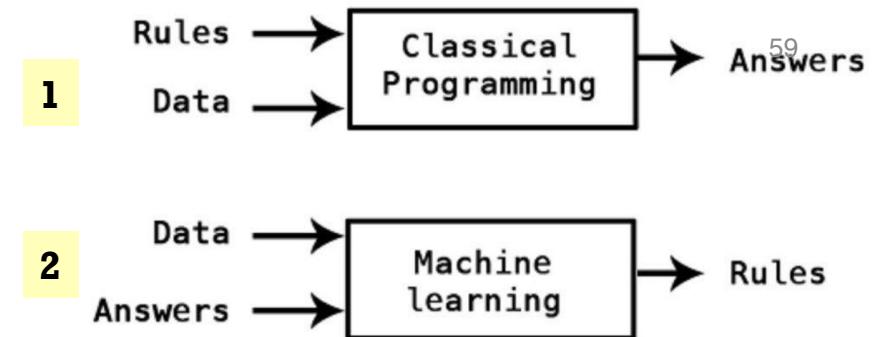


AI = Automation

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



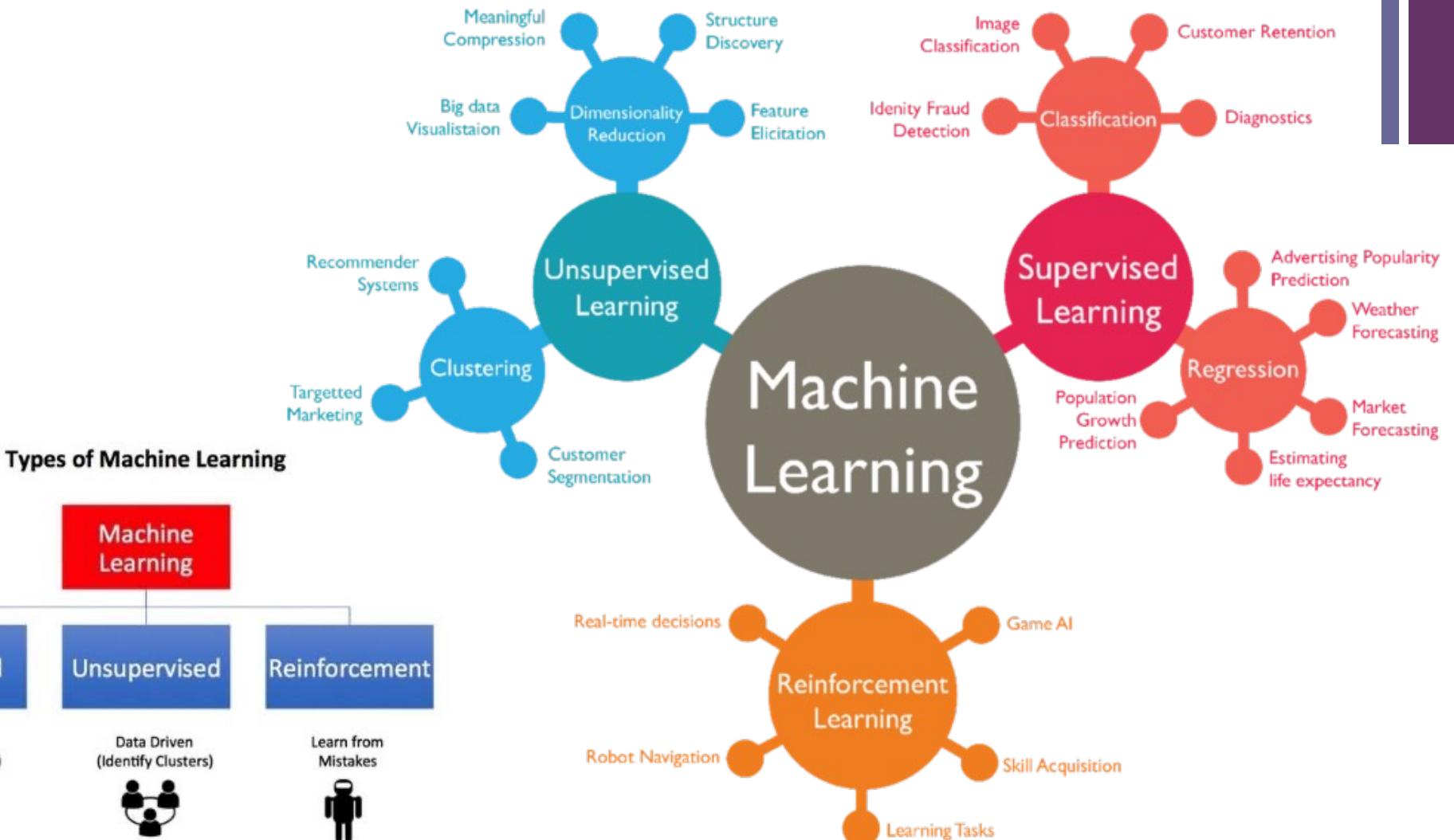
Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.



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

+ Machine Learning (ML)

60



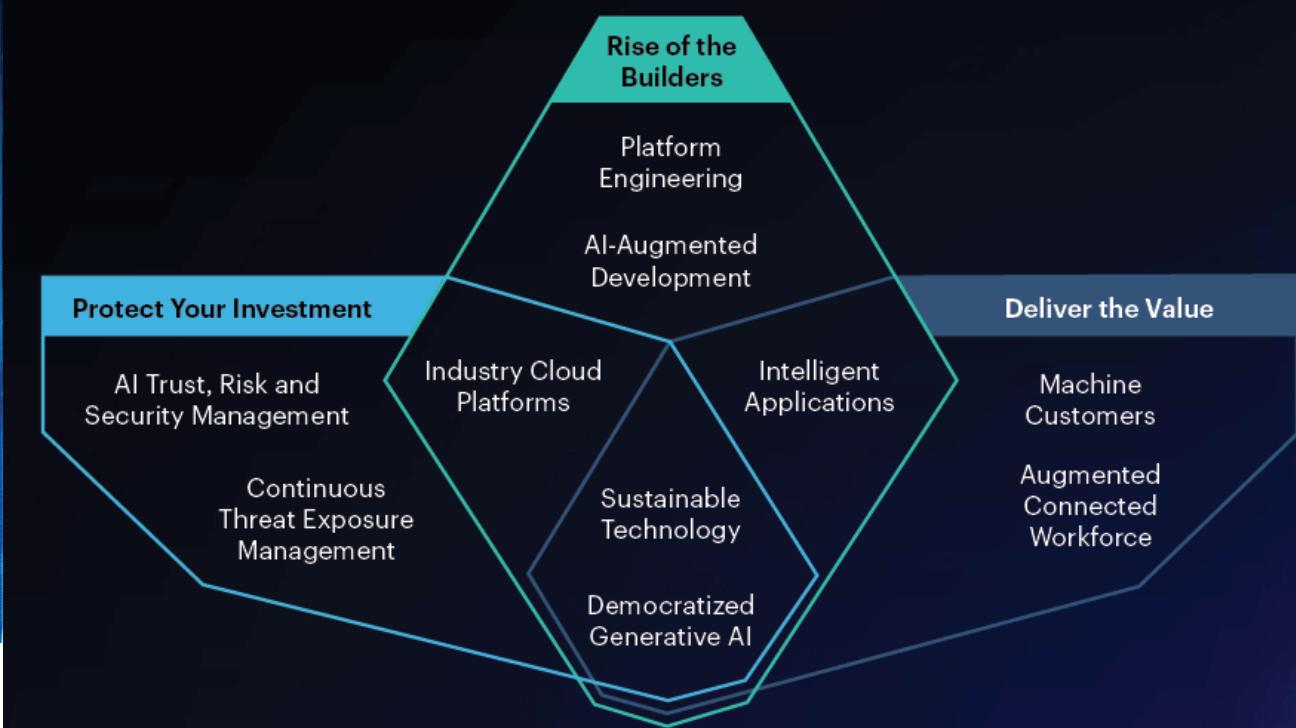
Top Strategic Technology Trends 2024

Top Strategic Technology Trends 2024

- | | |
|----|--|
| 1 | AI Trust, Risk and Security Management |
| 2 | Continuous Threat Exposure Management |
| 3 | Sustainable Technology |
| 4 | Platform Engineering |
| 5 | AI-Augmented Development |
| 6 | Industry Cloud Platforms |
| 7 | Intelligent Applications |
| 8 | Democratized Generative AI |
| 9 | Augmented Connected Workforce |
| 10 | Machine Customers |

Source: Gartner
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. CM_GTS_2080051

Gartner



Source: Gartner
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. CM_GTS_2080051

Gartner

<https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-trends-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 Quadrant



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



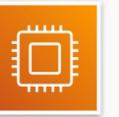
AR and VR



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
network



Robotics



Satellite



Security, Identity, and
Compliance



Storage

This document is protected under the copyright laws of Thailand and contains information that is proprietary and confidential to DailiTech, which shall not be disclosed outside the recipient's company or duplicated, used, or disclosed in whole or in part by the recipient for any purpose other than to the purpose of this document. Any other use or disclosures in whole or in part of this information without the express written permission of DailiTech are prohibited.

Data preparation for AI / ML and data science

Right



Vit Niennattrakul, Ph.D.



Amazon Forecast



Amazon Comprehend



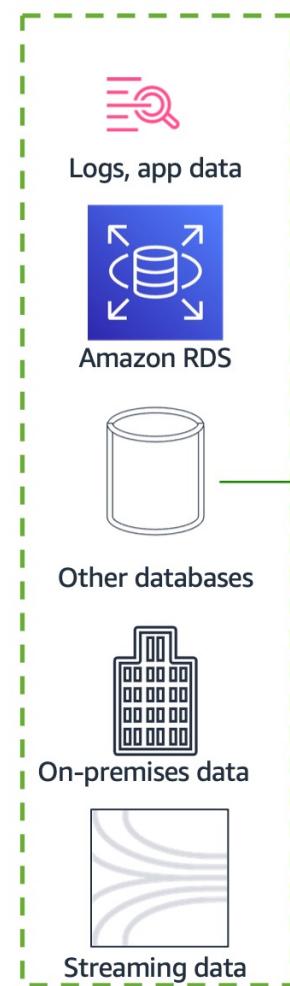
Amazon SageMaker



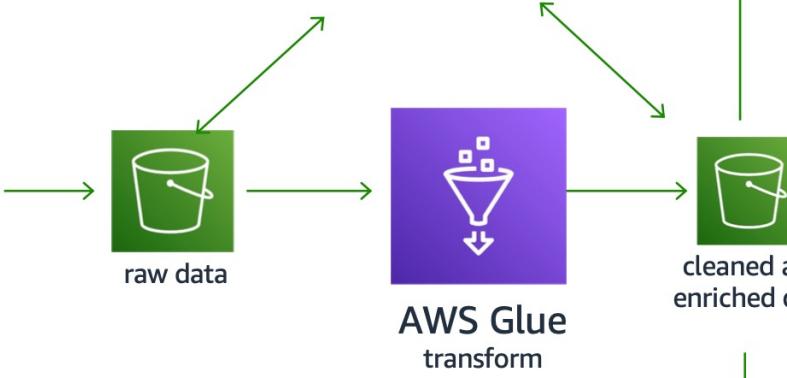
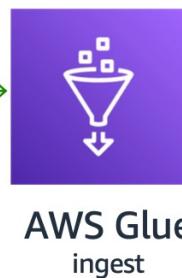
Amazon Rekognition



Amazon Lex



Notebooks:
data exploration,
experimentation



extracted features



training data

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

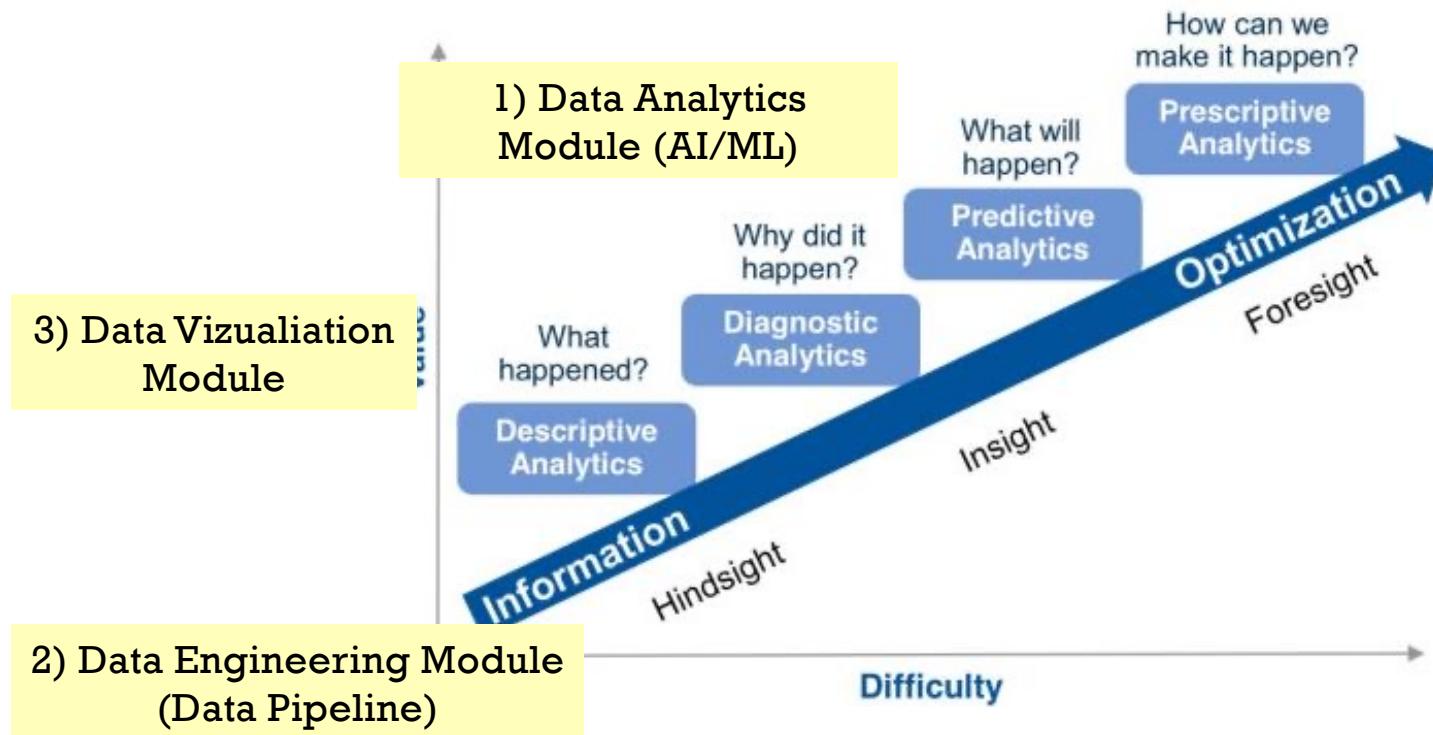
AWS Academy Lab Project - Cloud Data Pipeline Builder

- 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



+

Any questions? ☺