Lecture #1 | Introduction to big data and data science

SE271 Object-oriented Programming (2017)

Prof. Min-gyu Cho

Big Data?


ALSO
BIG
DATA
TOOLS
DISK
TARGET
APPLIED
SENSOR
SHARED
DEFINITION
CURRENT
MAY
MOVING
WITHIN
THOUGHT
EXAMPLES
EVERY
RECONSIDER
ZETTABYTES
PRACTITIONERS
CAPTURE
BUSINESS
INTERNET
SETS
GENOMICS
COMPLEXITY
DISTRIBUTED
DESCRIBING
RADIO-FREQUENCY
MANAGEMENT
TERABYTES
CASE
PETABYTES
INCLUDE
ABILITY
SYSTEMS
TOLERABLE
ELAPSED
FORMS
FC
STORAGE
SIZE
MPP
SAN
QUALITIES
PARALLEL
SINCE
MASSIVELY
GROW
SEARCH
DIFFICULTY
GARTNER
WORKING
OPPORTUNITIES
AMOUNT
CURRENTLY
SOLID
TYPES
RECORDS
NETWORKS
DATABASES
DESKTOP
HUNDREDS
CAPACITY
BURIED
WORLD'S
NOW
BIOLOGICAL
PROCESSING
UBIQUITOUS
ANALYTICS
TENS
COMPLEX
CITATION
INDEXING
LARGER
CONTINUES
SOCIAL
USE
SET
CONNECTOMICS
ORGANIZATIONS
RELATIONAL

4V of Big Data

- **Volume**: large amount of data
- **Variety**: various types of data
  - Numbers/text/pictures/videos
  - Web info, SNS, IoT
- **Velocity**: speed of data flow
- **Veracity**: uncertainty or inconsistency of data

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

Volume
SCALE OF DATA

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

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

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

Velocity
ANALYSIS OF STREAMING DATA

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

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

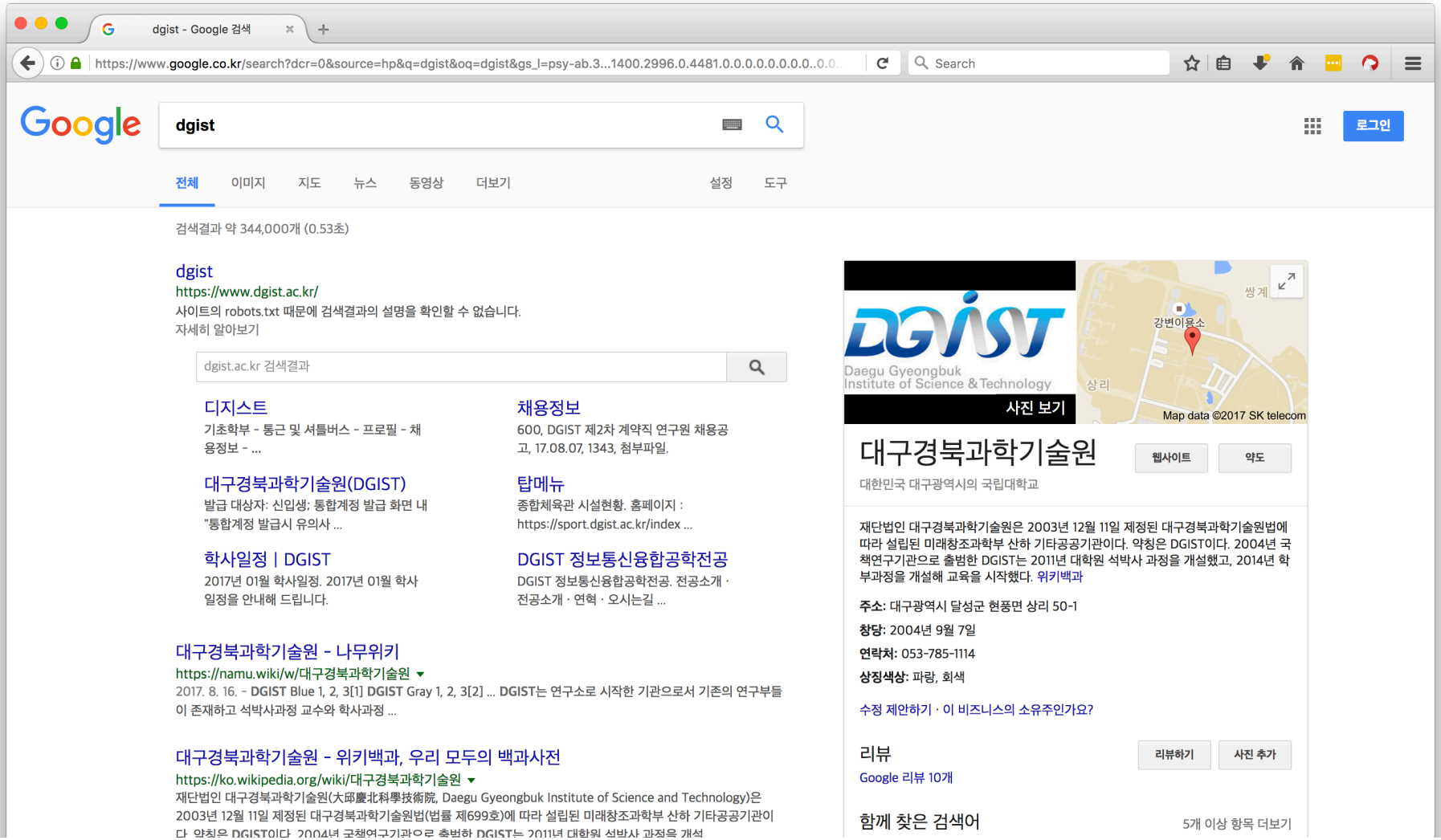
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

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

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

IBM

Example: Search engines



Example: Netflix recommender system

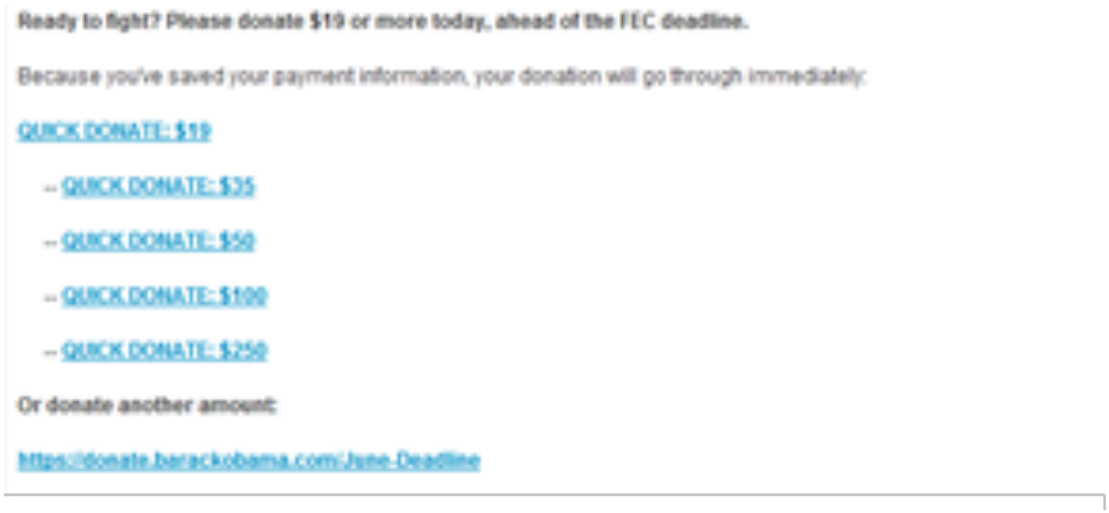
Everything is a Recommendation

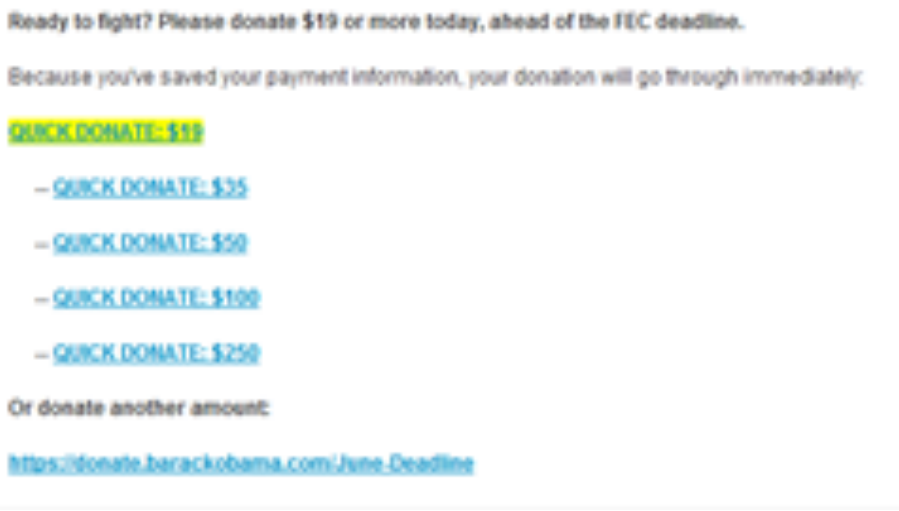
Over 75% of what people watch comes from our recommendations

Source: https://www.slideshare.net/justinbasilico/recommendation-at-netflix-scale

Example: 2012 US presidential elections

- Which e-mail is better?



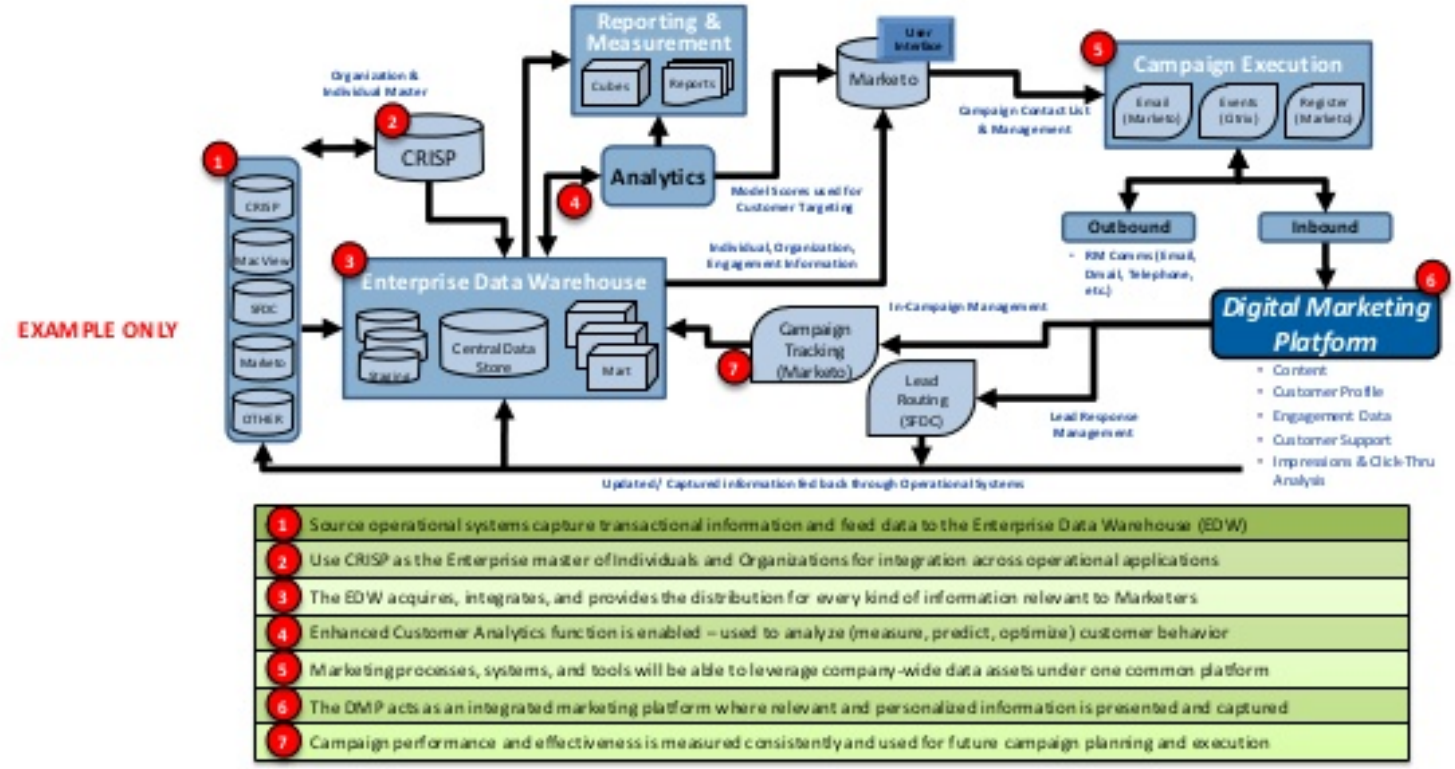


- A/B test and re-test is IMPORTANT!!!

Source: https://www.slideshare.net/jessday/obama-2012-lessons-from-a-datadriven-campaign

Example: Digital Marketing

Part of the Future Vision Would be to Map the Enablement of the *Marketing Process* through the Marketing Data Platform.



Source: https://www.slideshare.net/martinwalsh/example-digital-marketing-platform-strategy

Example: Deep Learning

- Used 22 mil. sentences as a base and generated simulated data sets to increase speech recognition accuracy of Google Home

Table 1: *Speech recognition Word Error Rates(WERs)*

	Trained with the room simulator	Baseline system
Original Test Set	11.97 %	12.02 %
Simulated Noise Set	19.55 %	47.88 %
Device 1	21.98 %	50.14 %
Device 2	22.23 %	48.65 %
Device 3	22.05 %	56.27 %
Device 3 (Noisy Condition)	34.83 %	76.01 %
Device 3 (Multi-talker Condition)	44.79 %	78.95 %

Source: https://research.google.com/pubs/pub46130.html

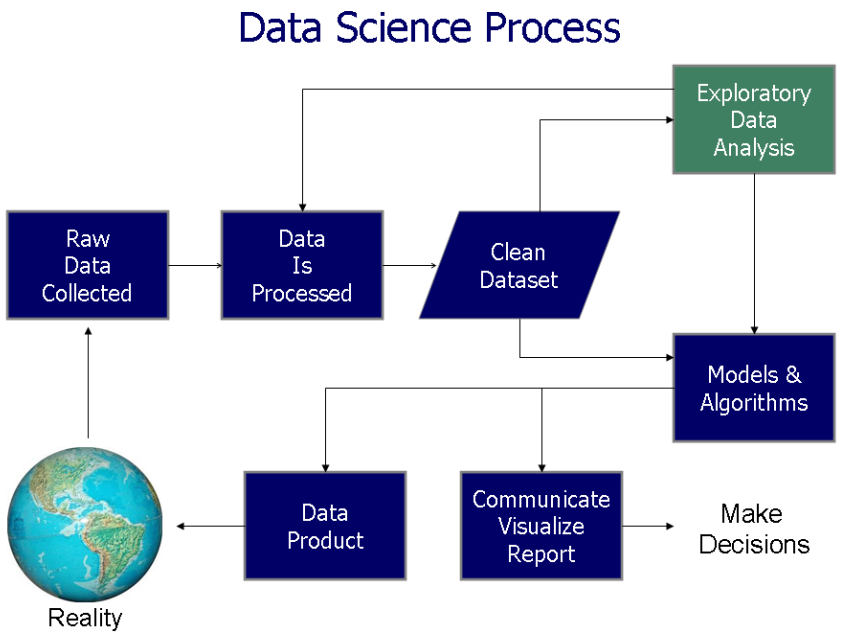
Why is big data analytics important?

- **Cost reduction**: Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
- **Faster, better decision making**: With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.
- **New products and services**: With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

Source: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html

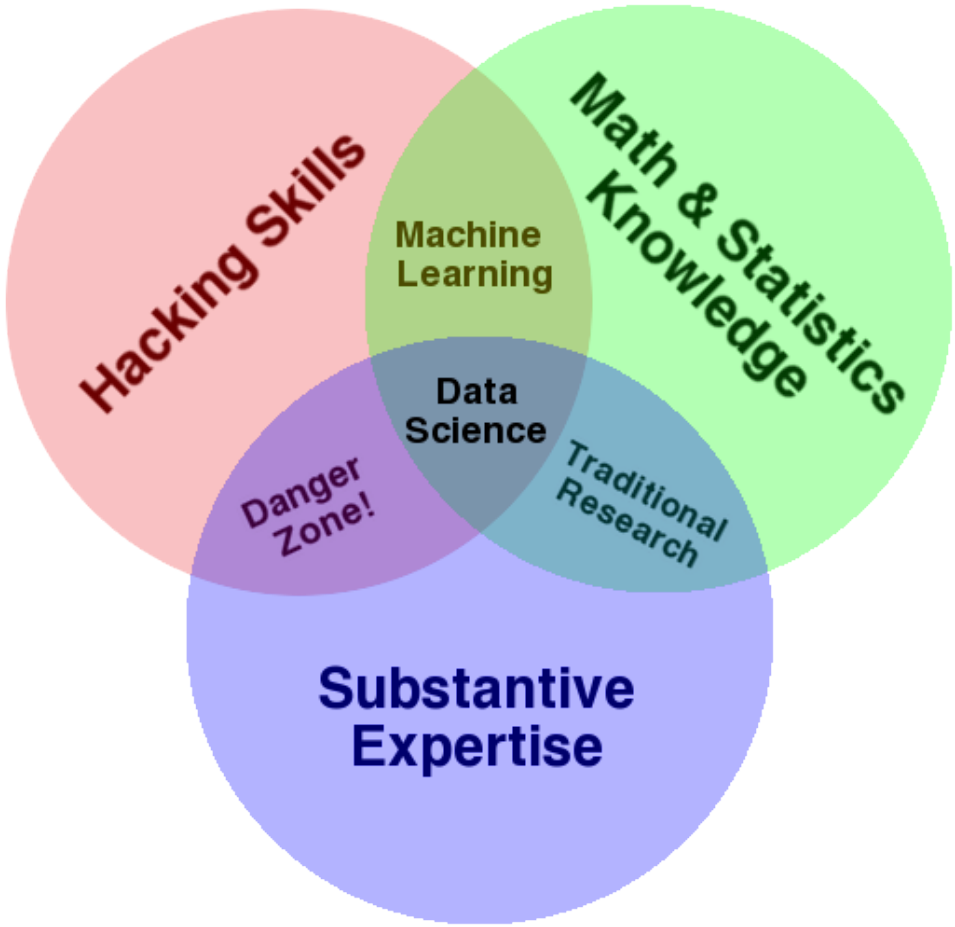
What is data science?

- Data science is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining for '**decision making**'

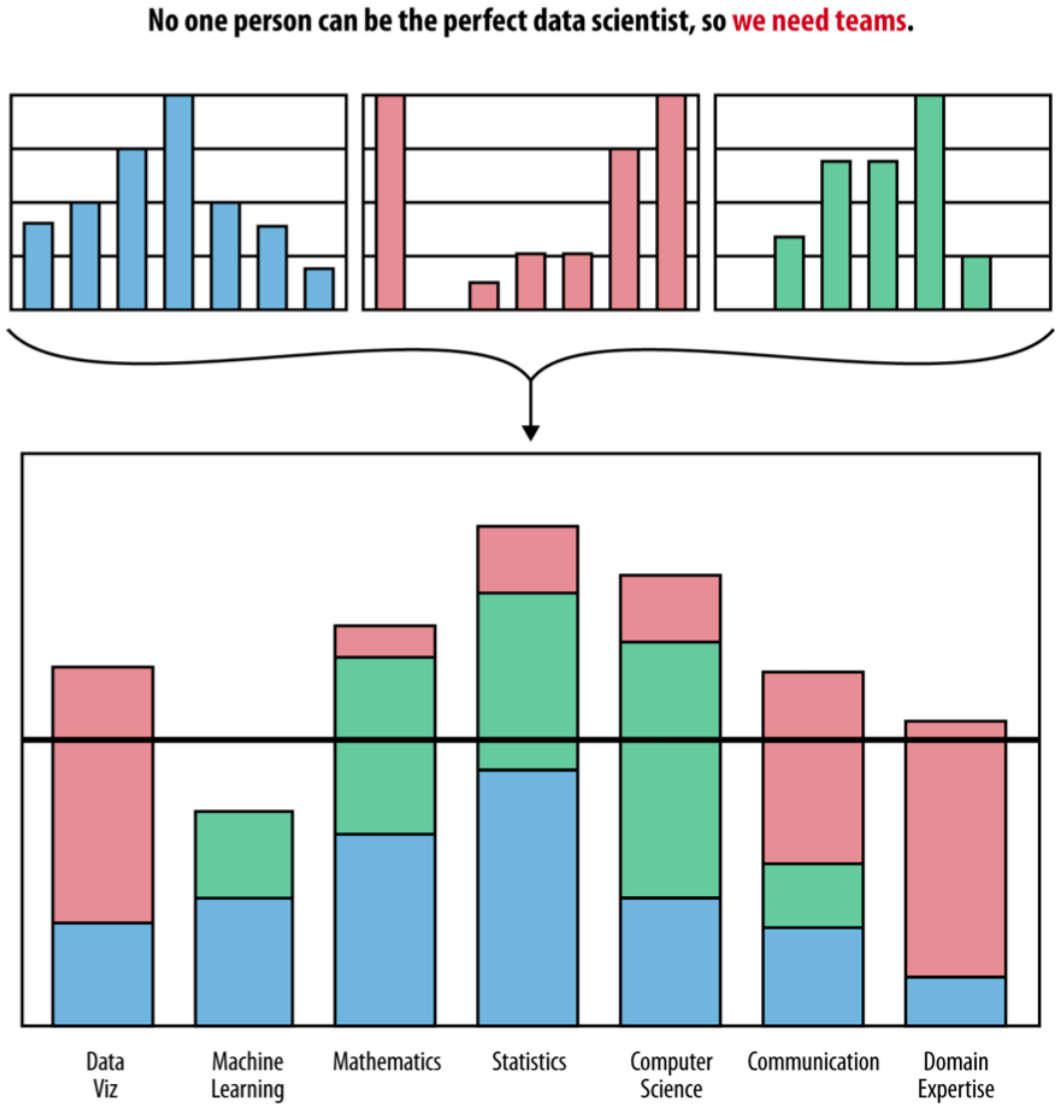


Source: https://en.wikipedia.org/wiki/Data_science

Another definition of data science



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



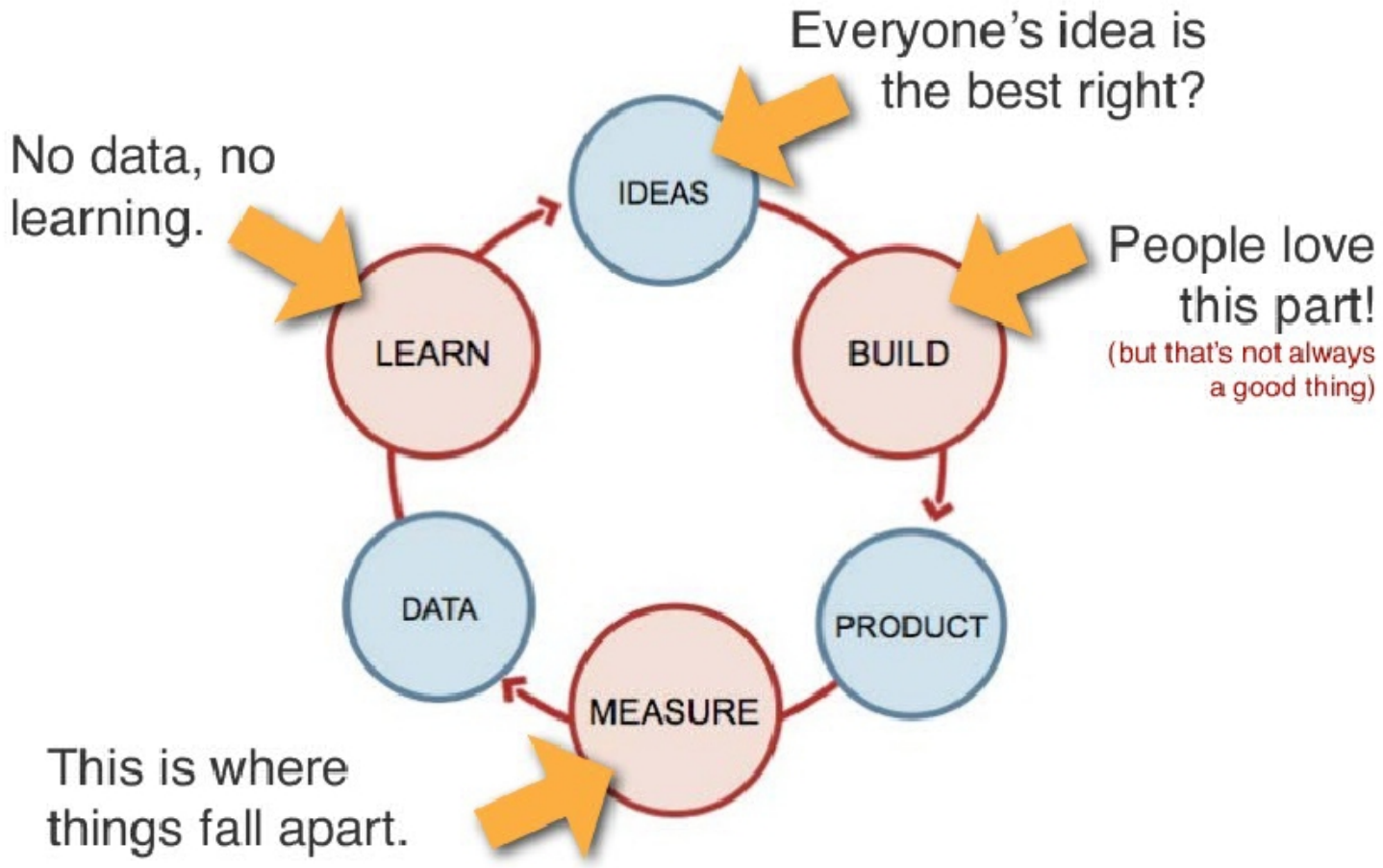
Source: “Doing Data Science”

Exploratory Data Analysis (EDA)

- Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set;
  - Uncover underlying structure
  - Extract important variables
  - Detect outliers and anomalies
  - Test underlying assumptions
  - Develop parsimonious models
  - Determine optimal factor settings
- The EDA approach is precisely that – an approach – not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out
- c.f., “Exploratory Data Analysis”, John W. Tukey (1977)

Source: http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

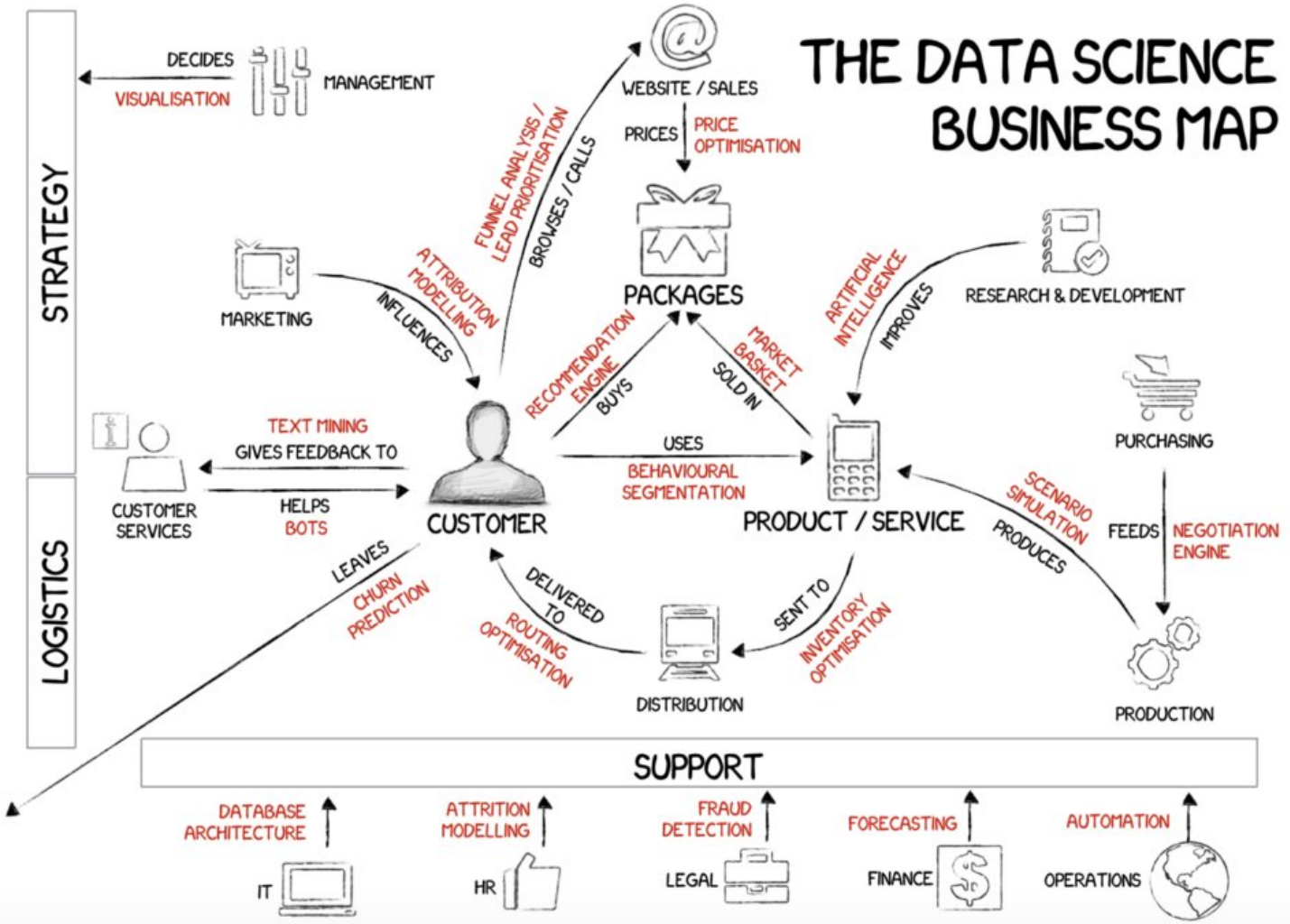
Lean Startup Cycle



Source: "Lean Analytics" by Alistair Croll

https://applied-data.science

APPLIED DATA SCIENCE

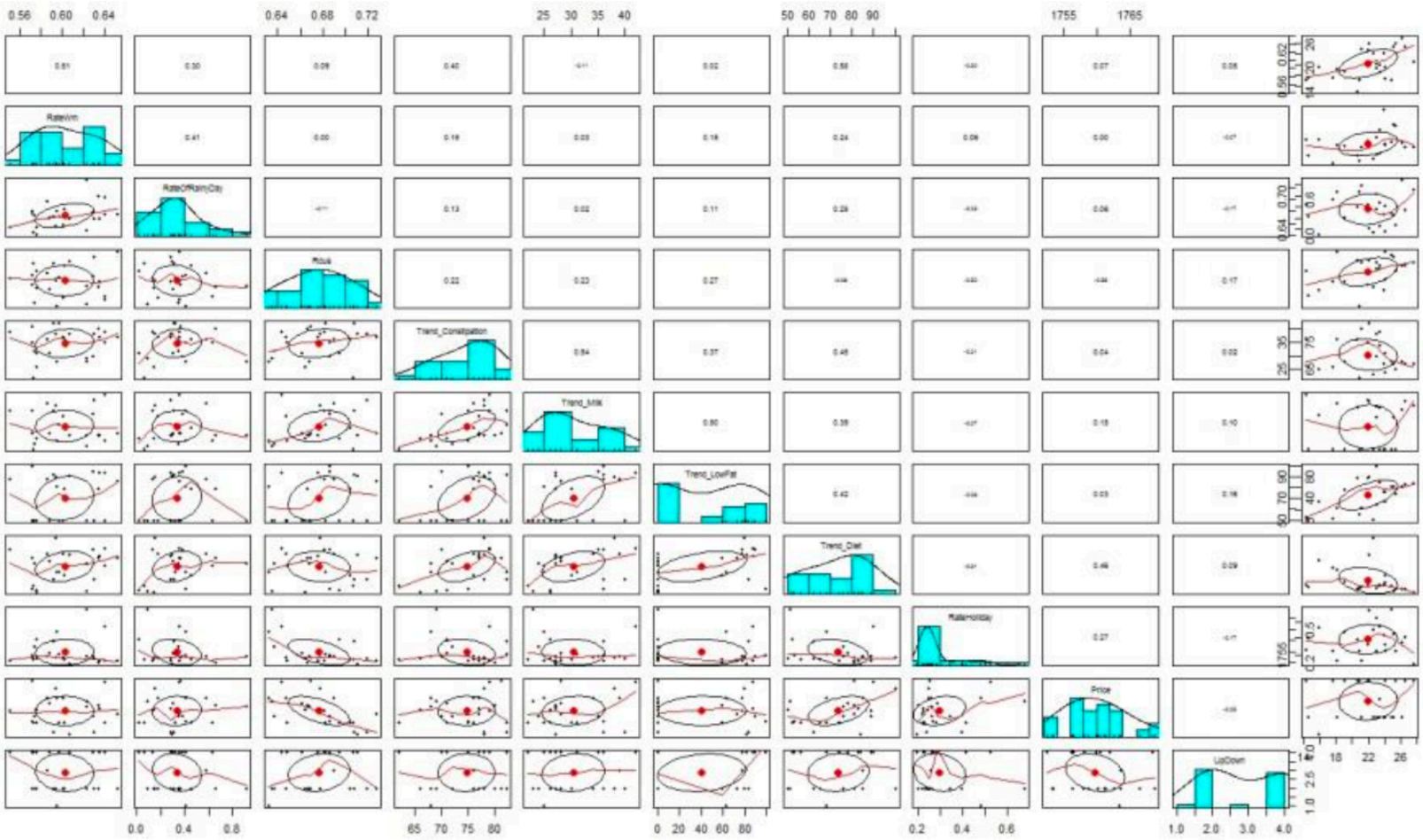


Data Visualization

- Data visualization is the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information"
- Two objectives of data visualization
  - Visualization for analysis: figure out hidden relationship between several features of given data, typically during EDA process
  - Visualization for communication: deliver insights to the targeted audience efficiently and effectively without any distortion of the facts, possibly using infographics or interactive visualization

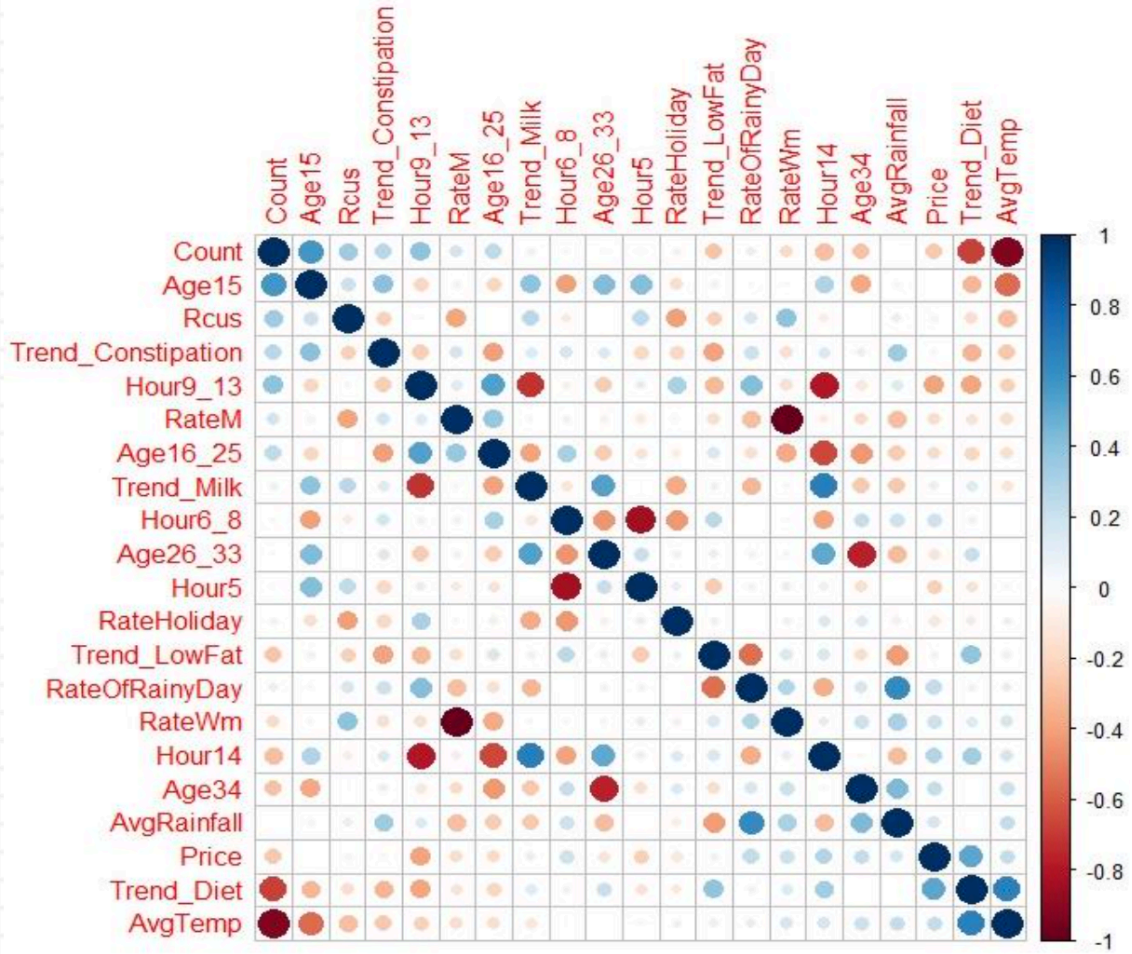
Source: “Milestones in the history of thematic cartography, statistical graphics, and data visualization”

Example of data visualization during EDA



Source: Sales prediction of dairy products by Hyunlim Yang

Example of data visualization during EDA (cont.)



Source: Sales prediction of dairy products by Hyunlim Yang

Examples of Data Visualization

- 축구 패스 분포도: http://news.joins.com/Digitalspecial/124
- 선거의 득표 수를 어떻게 지도에 표현할 것인가?
  - http://www.vw-lab.com/40
- 2017년 19대 대통령 선거개표결과, 중앙일보
  - http://news.joins.com/DigitalSpecial/179
- Curated list of data visualization
  - https://www.maptive.com/17-impressive-data-visualization-examples-need-see/
  - https://www.webdesignerdepot.com/2009/06/50-great-examples-of-data-visualization/

Data Science v.s. Data Engineering

- **Data Engineering**: manage automatic and scalable data flow (close to CS)
  - **Extract**: The process of reading data from a database
  - **Transform**: the process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data
  - **Load**: the process of writing the data into the target database
- **Data Science**: find insights from data (interdisciplinary domain)
  - **Discover**: Find, identify the sources of good data, and the metrics. Sometimes request the data to be created (work with data engineers and business analysts)
  - **Access**: Access the data. Sometimes via an API, a web crawler, an Internet download, a database access or sometimes in-memory within a database
  - **Distill**: Extract essence from data, the stuff that leads to decisions, increased ROI, and actions (such as determining optimum bid prices in an automated bidding system)

Assignment #0

- Due: 9/11 Monday
- Register to elice
  - Register dgist.elice.io
    - Use dgist email and Korean first/last name
  - Subscribe to “Introduction to Big Data Analysis and Visualization”
    - Subscription password: dgist_2017_ds
- Set up anaconda on your notebook

Reading List

- 데이터과학을 시작할 때 도움되는 것들
  - https://brunch.co.kr/@amangkim/37
  - https://brunch.co.kr/@amangkim/38
- 스타트업은 데이타를 어떻게 바라봐야 할까?
  - https://www.slideshare.net/yongho/ss-32267675
- Data Scientist versus Data Engineer
  - http://www.datasciencecentral.com/profiles/blogs/data-scientist-versus-data-engineer
- A very short history of data science
  - https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

ANY QUESTIONS?