Computer Programming for DS and AI

Asst. Prof. Chantri Polprasert

Aug 2024

Dept. of ICT, AIT

Course LMS: Moodle

Computer Programming for Data Science and Artificial Intelligence (Aug 2024)

The course objective is to provide students with hands-on programming skills and best practices related to Data Science and Artificial Intelligence. It is a laboratory course in which students will develop programming skills in loading, cleansing, transforming, modeling, and visualizing data.

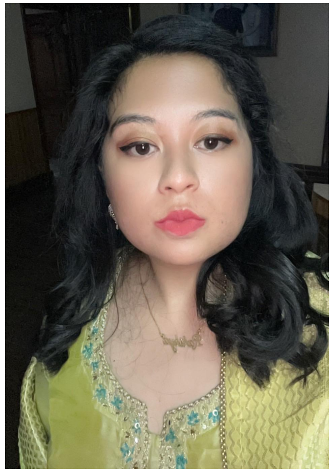
Teacher: Chantri Polprasert

https://teal2o.cs.ait.ac.th/teal_classroom/course/view.php?id=51

Enrollment key: cpdsai24

Instructor team

- Lecturer: Computer Programming in Data science, Algorithm, @Dept. of ICT, AIT
- Previously: Researcher @NECTEC, Instructor at Faculty of Science, Srinakharinwirot University
- Research Projects: Sign language, Stroke prediction, Human capability enhancement
- Email: chantri@ait.ac.th
- Office hour: Every Tuesday from 12:00-13:00 or upon request at CS210.



- TA: Rakshya Rama Moktan
- Office hour: Friday 13:00-14:00

Today’s Outline

- What is Data Science?
- Data Science Process
- Analytical Thinking, Asking Questions, Defining Problems
- Course Goal and Logistics

What is Data Science?

What is Data Science?

"Data Science is the exploration and quantitative analysis of all available structured and unstructured data to develop understanding, extract knowledge, and formulate actionable results."

- *Microsoft's* DAT203.1x Data Science Essentials

"Data science is the application of computational and statistical techniques to address or gain insight into some problem in the real world."

- *Zico Kolter, Carnegie Mellon University*

What is Data Science?

"Data science about drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference."

- ***Exploration*** *involves identifying patterns in information.*
- ***Prediction*** *involves using information we know to make informed guesses about values we wish we knew.*
- ***Inference*** *involves quantifying our degree of certainty: will the patterns that we found in our data also appear in new observations? How accurate are our predictions?*

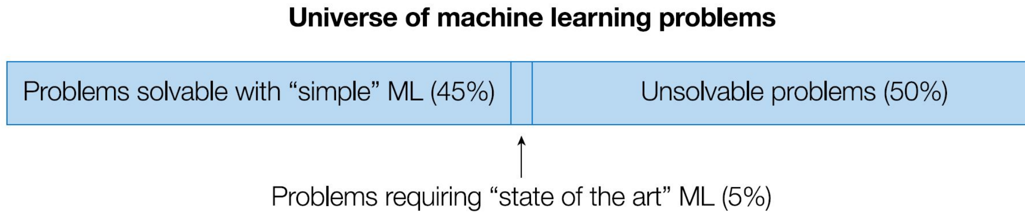
***Tools**:*

- ***Exploration**: visualizations and descriptive statistics*
- ***Prediction**: Machine learning and optimization*
- ***Inference**: statistical tests and models.*

Ani Adhikari and John DeNero and David Wagner, UC Berkeley

What Data Science is not

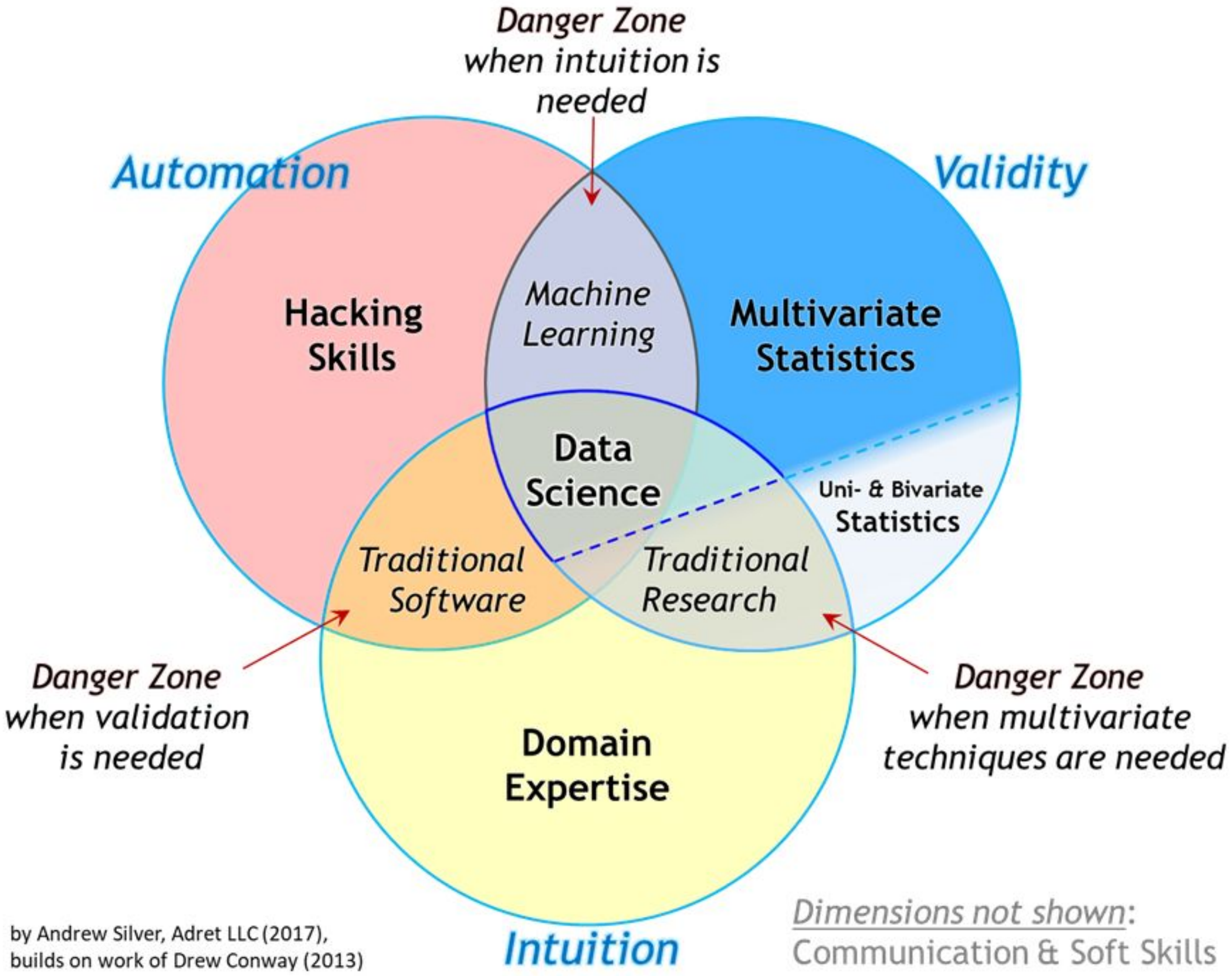
- Data science is not (just) machine learning. It involves
  - Defining the problem
  - Collecting data
  - Exploring, Interpreting and understanding results
  - Knowing what actions to take
  - Inference



- Data science is not (just) statistics
  - Historically, the academic field of statistics has tended more towards the theoretical aspects of data analysis than the practical aspects.
  - data science has evolved from computer science as much as it has from statistics.
- Data science is not (just) big data

How to become a Data Scientist?

- The Data Science Venn Diagram

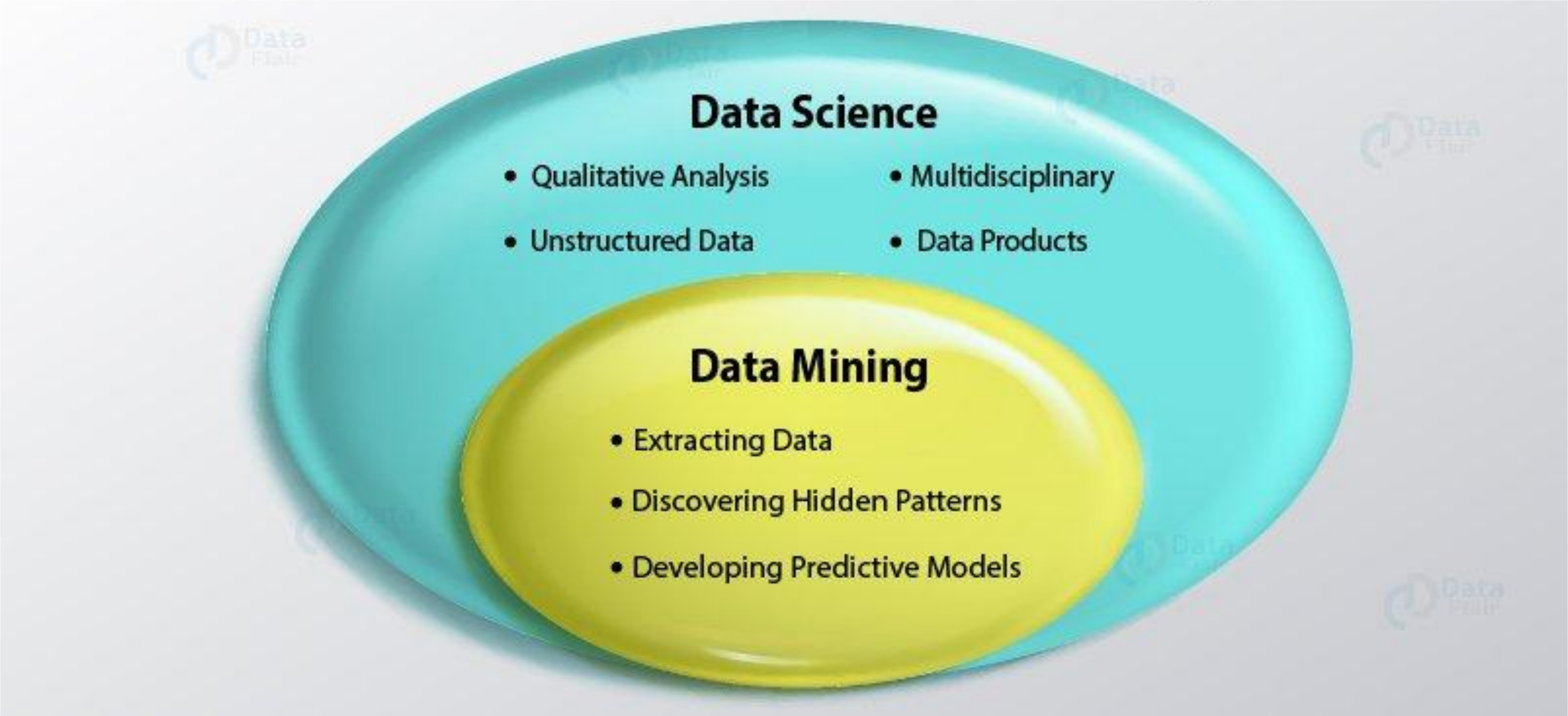


Domain Expertise:

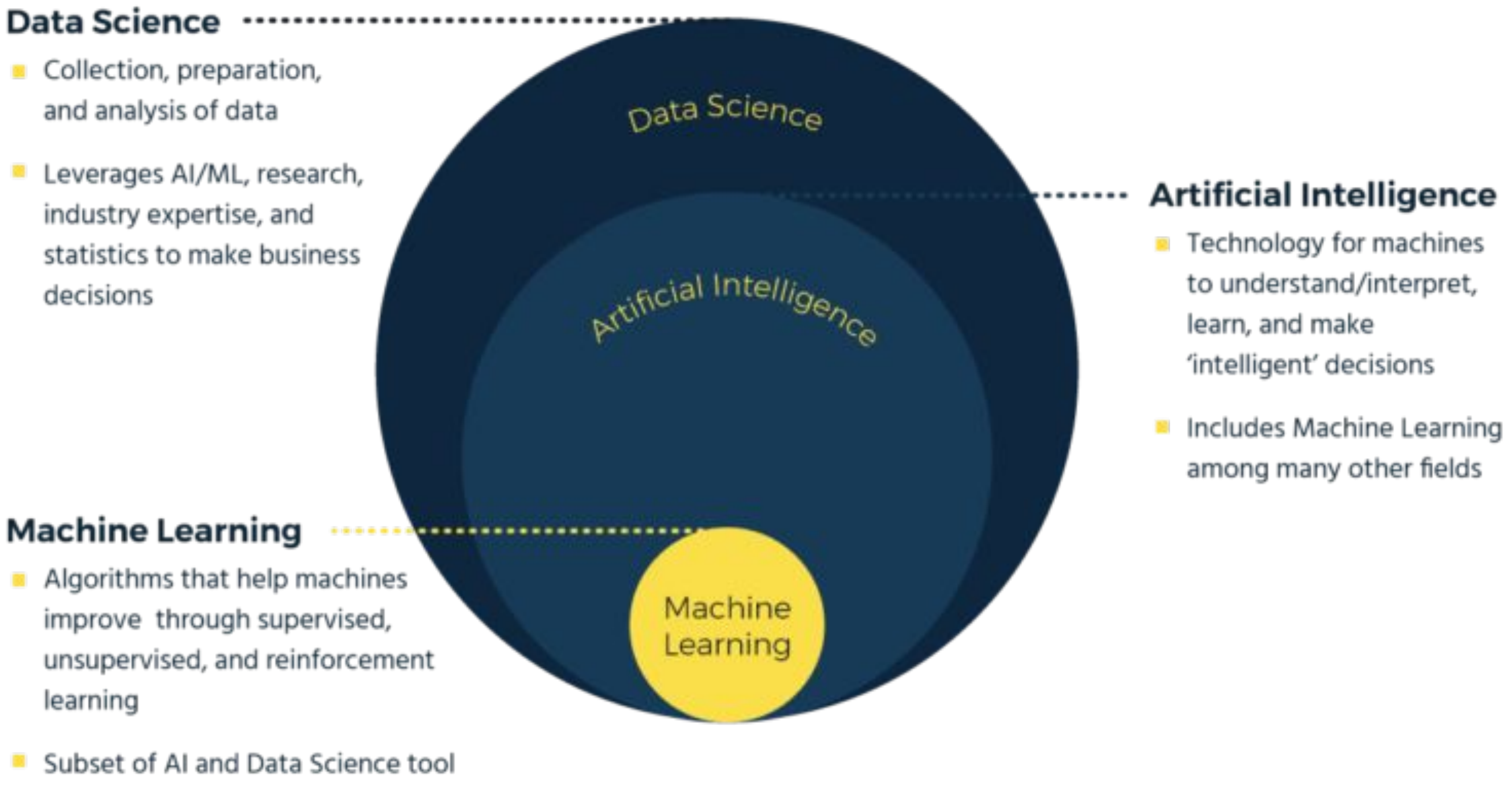
- Knows which questions to ask.
- Can interpret the data well.
- Understands the structure of the data.
- Work in teams.

Source : https://towardsdatascience.com/the-essential-data-science-venn-diagram-35800c3bef40

Data Science vs. Data Mining

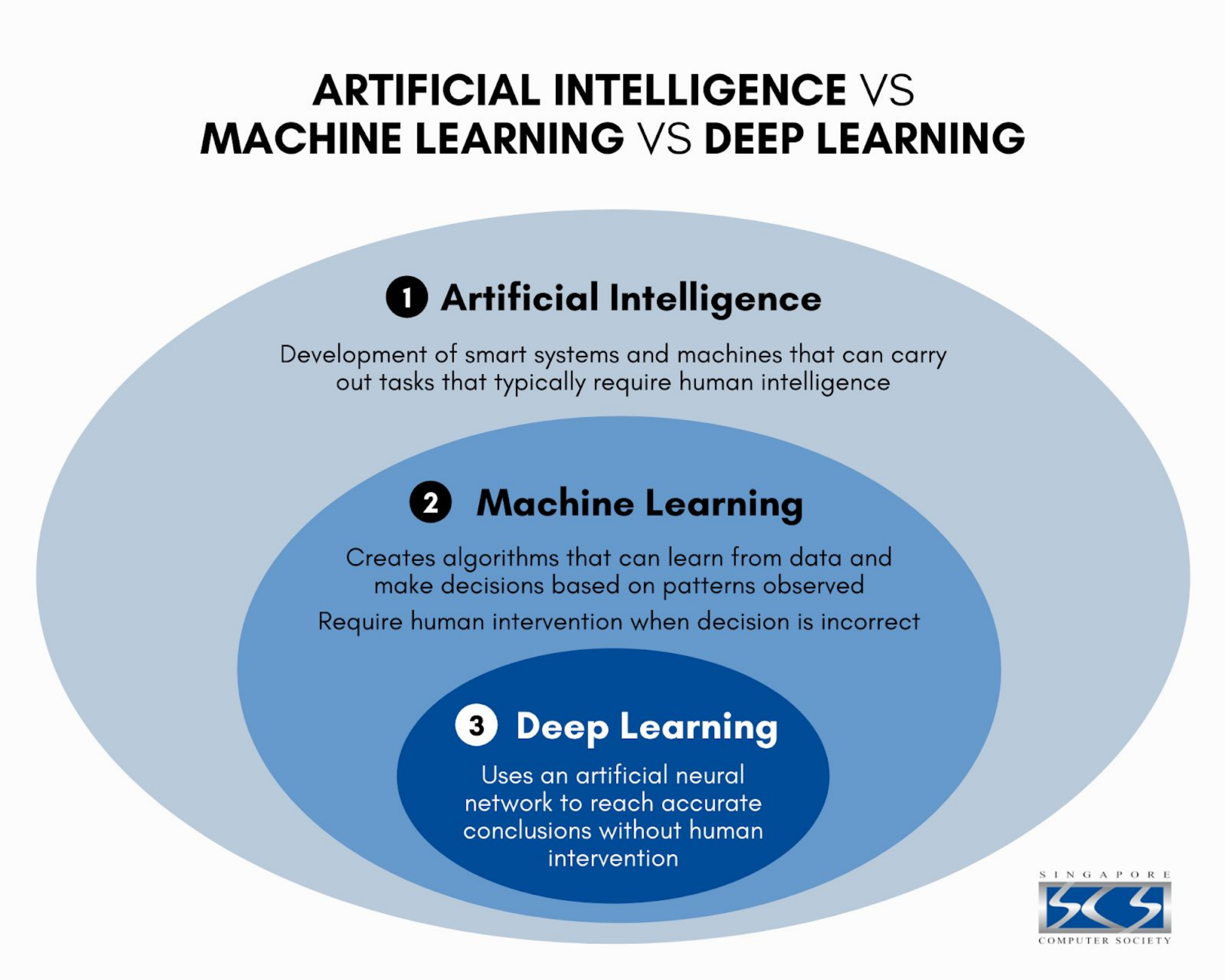


Data Science, Machine Learning and AI



https://braincube.com/resource/manufacturing-ai-vs-machine-learning-vs-data-science/

AI, Machine Learning and Deep Learning



https://www.scs.org.sg/articles/machine-learning-vs-deep-learning

Data-Driven Organization

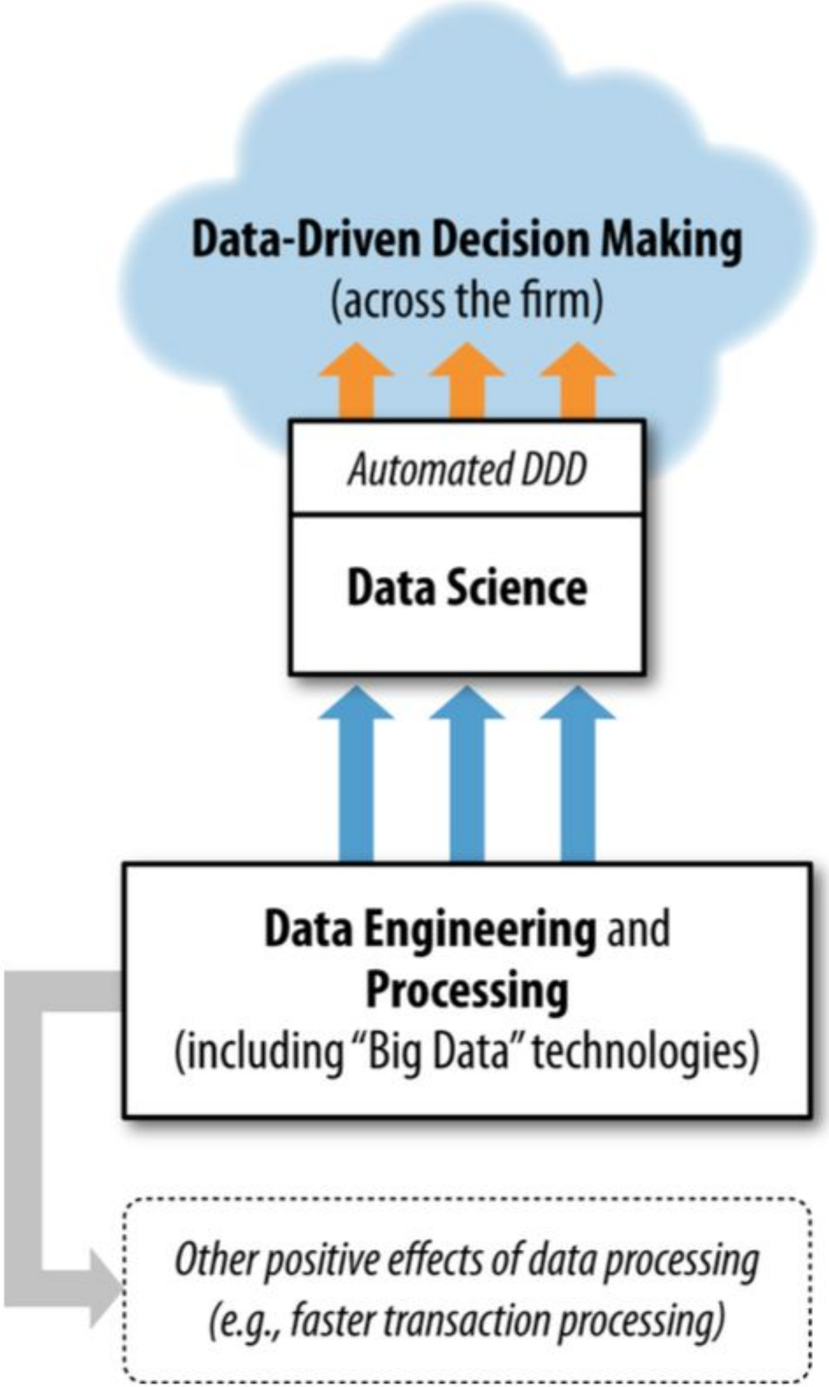
Data science in the Context of Various Data-Related Processes in the Organization

It distinguishes data science from other aspects of data processing that are gaining increasing attention in business.

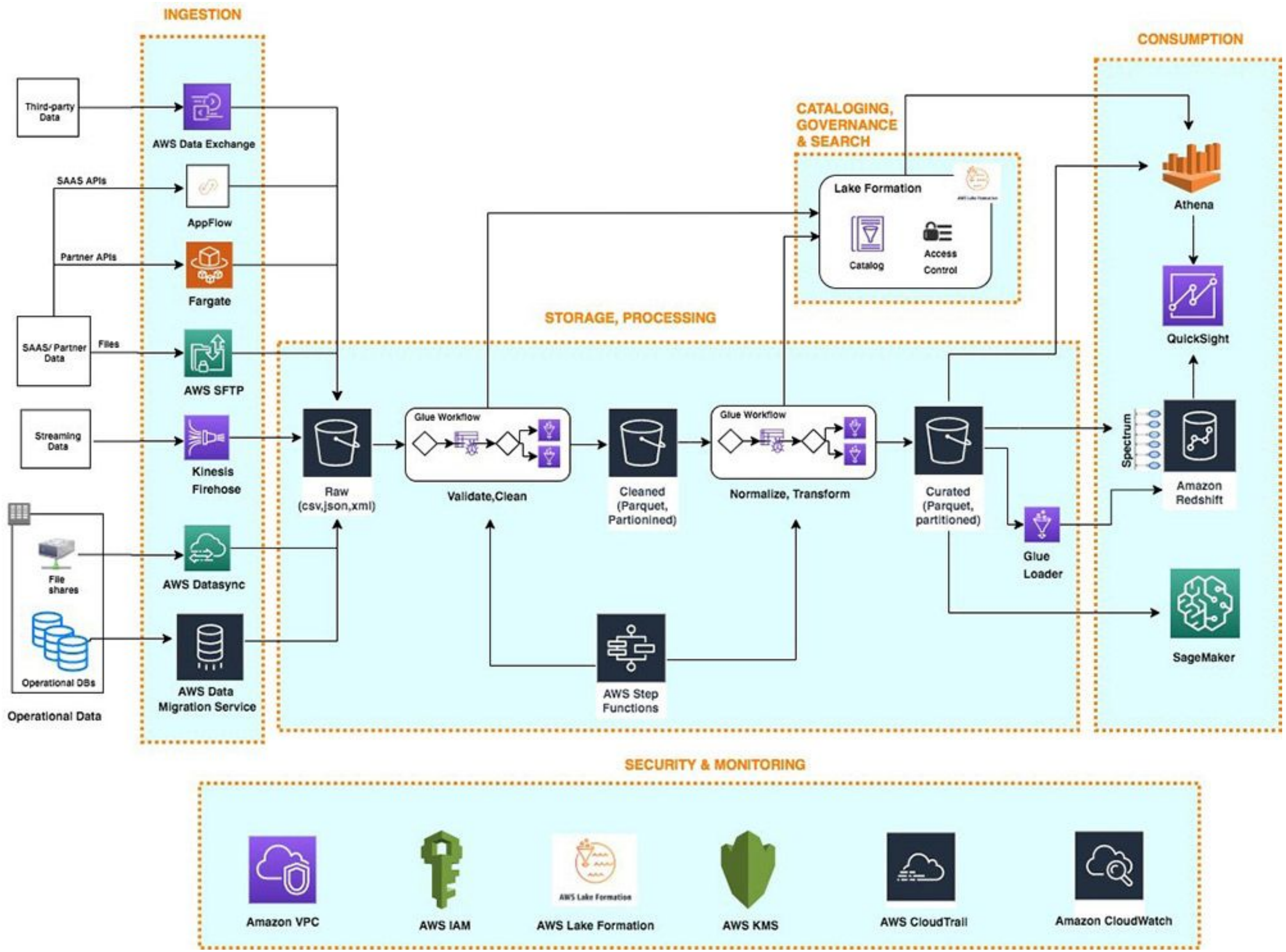
Data-driven organizations are 23 times more likely to acquire customers, six times as likely to retain customers, and 19 times as likely to be profitable as a result.

2020

McKinsey Global Institute

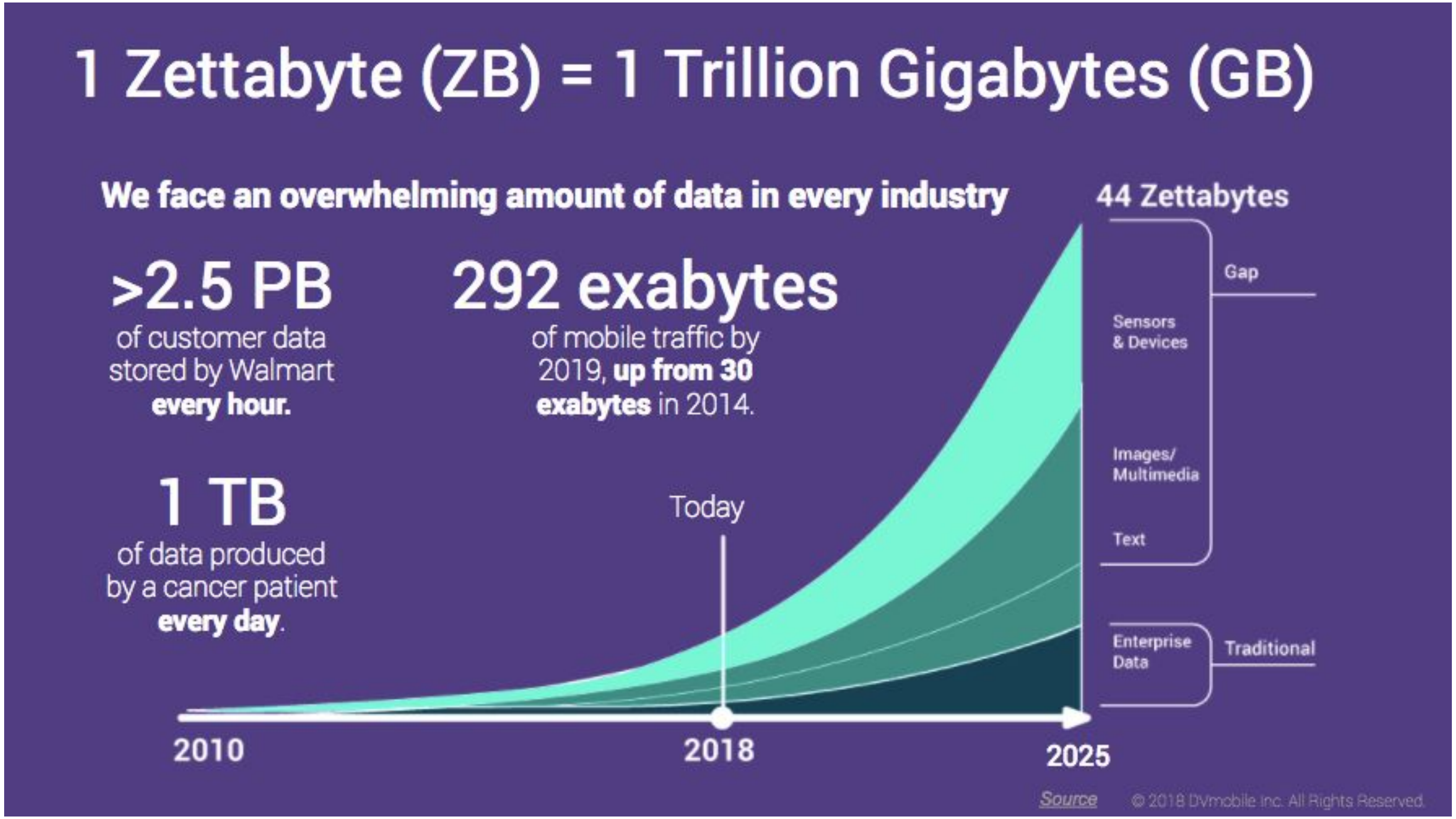


AWS serverless data analytics pipeline reference architecture

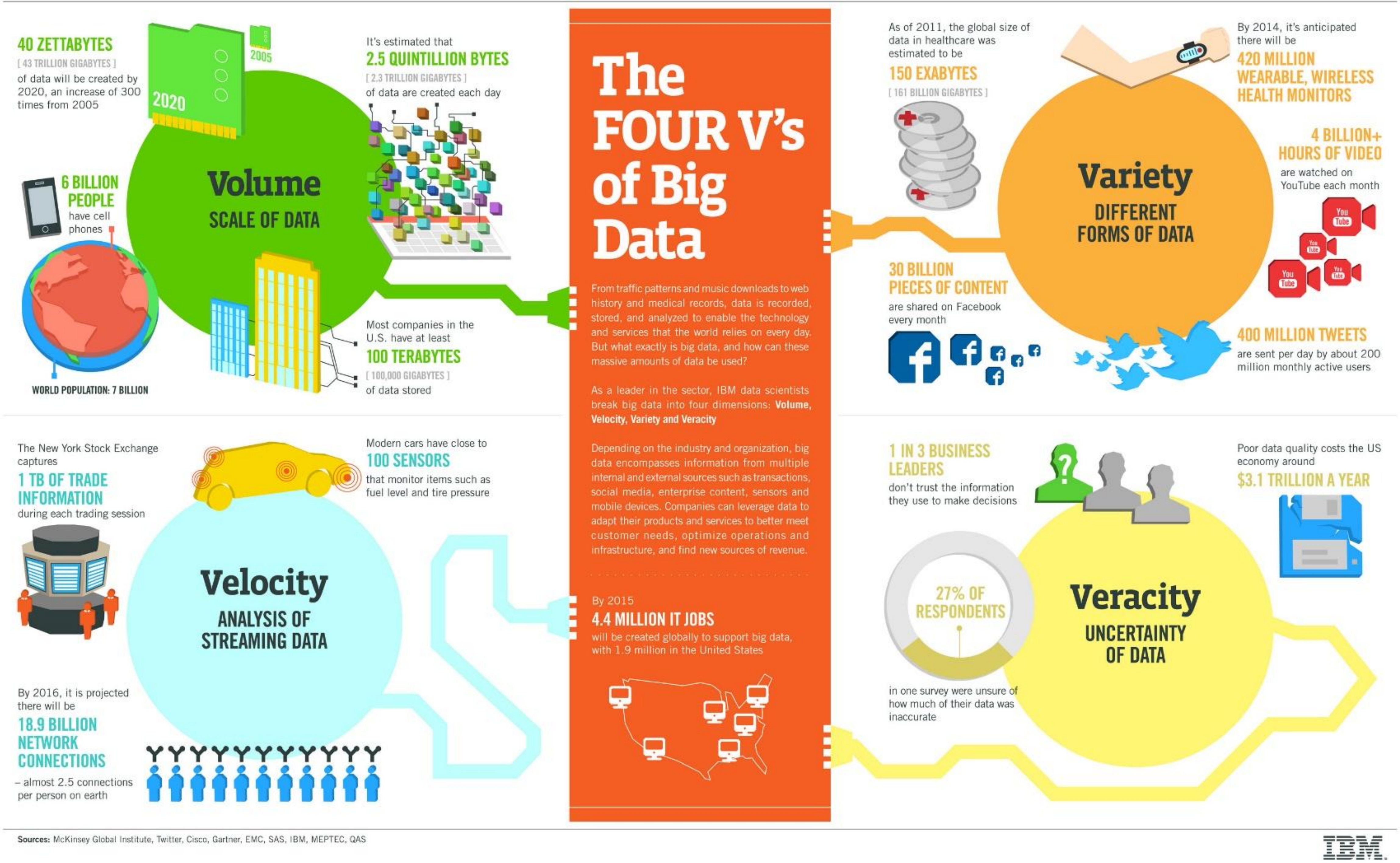


https://aws.amazon.com/blogs/big-data/aws-serverless-data-analytics-pipeline-reference-architecture/

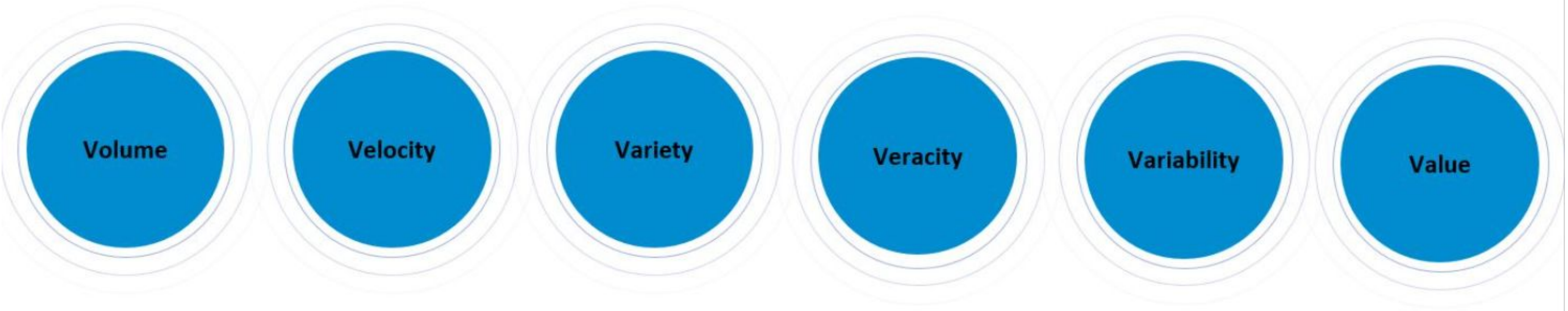
Why the Increased Interest in Data Science? :Exponential data growth!



- In the past, firms could employ teams of statisticians, modelers, and analysts to explore datasets manually, but the volume and variety of data have far outstripped the capacity of manual analysis.


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
2005
2020
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
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
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
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
27% OF RESPONDENTS
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

6Vs of Data



Volume
- How much data?
- Yottabytes
- Scalability of systems

Velocity
- How fast can I access?
- Hot Data path
- Cold Data path

Variety
- What type of data?
- Structured & Unstructured
- Heterogeneous formats

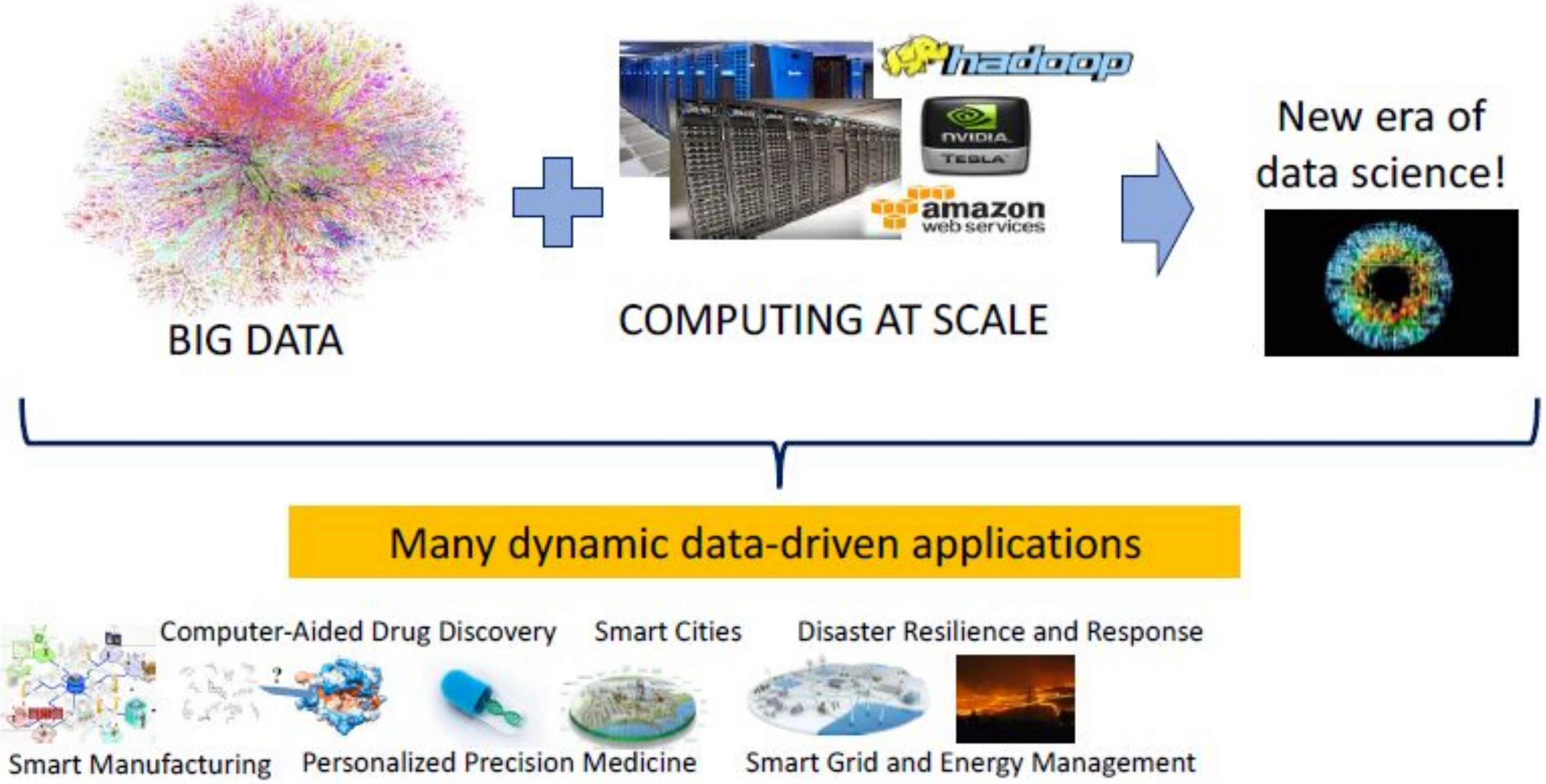
Veracity
- Is it reliable data?
- Accuracy of data
- Quality of data streams

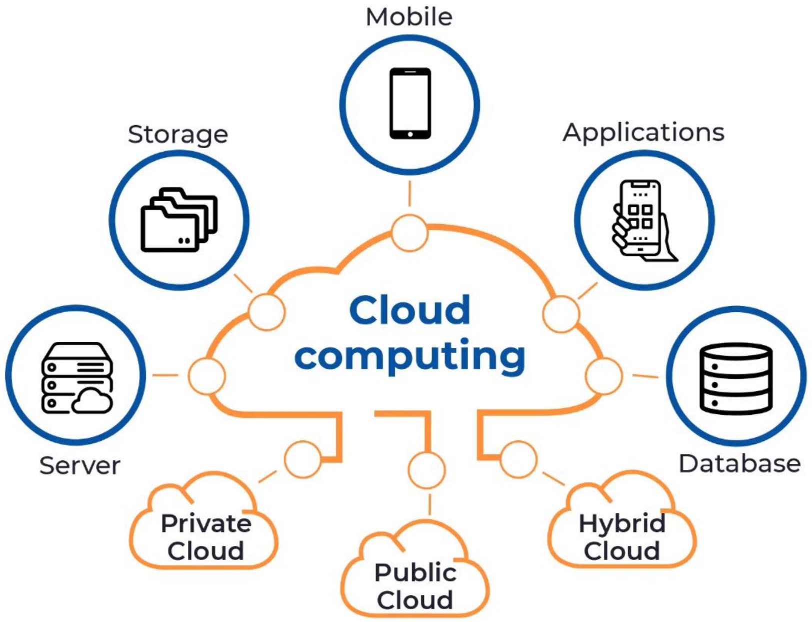
Variability
- How varied is the data?
- Different data sources
- Data Outliers

Value
- Usability of Data
- Utility of Data
- Usefulness of Data

Why the Increased Interest in Data Science? :Computing At Scale

- At the same time, computers have become far more powerful, networking has become ubiquitous, and algorithms have been developed that can connect datasets to enable broader and deeper analyses than previously possible.





Data Science Applications

- Medical and Health Care
  - Medical image analysis
  - Drug discovery
  - Stroke prediction
- Manufacturing
  - Supply chain optimization
  - Demand forecasting and inventory management
  - Robotics, automation and smart factory
  - Predictive maintenance, Industry 4.0
- Banking, Finance and Insurance
  - Credit risk modeling
  - Fraud claims
- E-commerce and Retail
  - Product recommendation
  - Online marketing
  - Product review analysis
  - Customer segmentation
- Transportation
  - Self-driving car
  - Logistics planning
  - Traffic delay planning
- Telecommunications
  - Churn prediction
  - Network traffic/quality management
- AI for Social Good (AI4SG)
  - AI for deaf people

Real Use Cases

Hurricane Frances

- From a *New York Times* story in 2004
- Walmart CIO, Linda M. Dillman, saw an opportunity to use ***predictive technology*** a week ahead of storm
- ***Discover patterns*** due to the past hurricane situations, they found that the stores would indeed need certain products, not just the usual flashlights.
- E.g. Strawberry Pop-Tarts,

  Beer!

Baby event

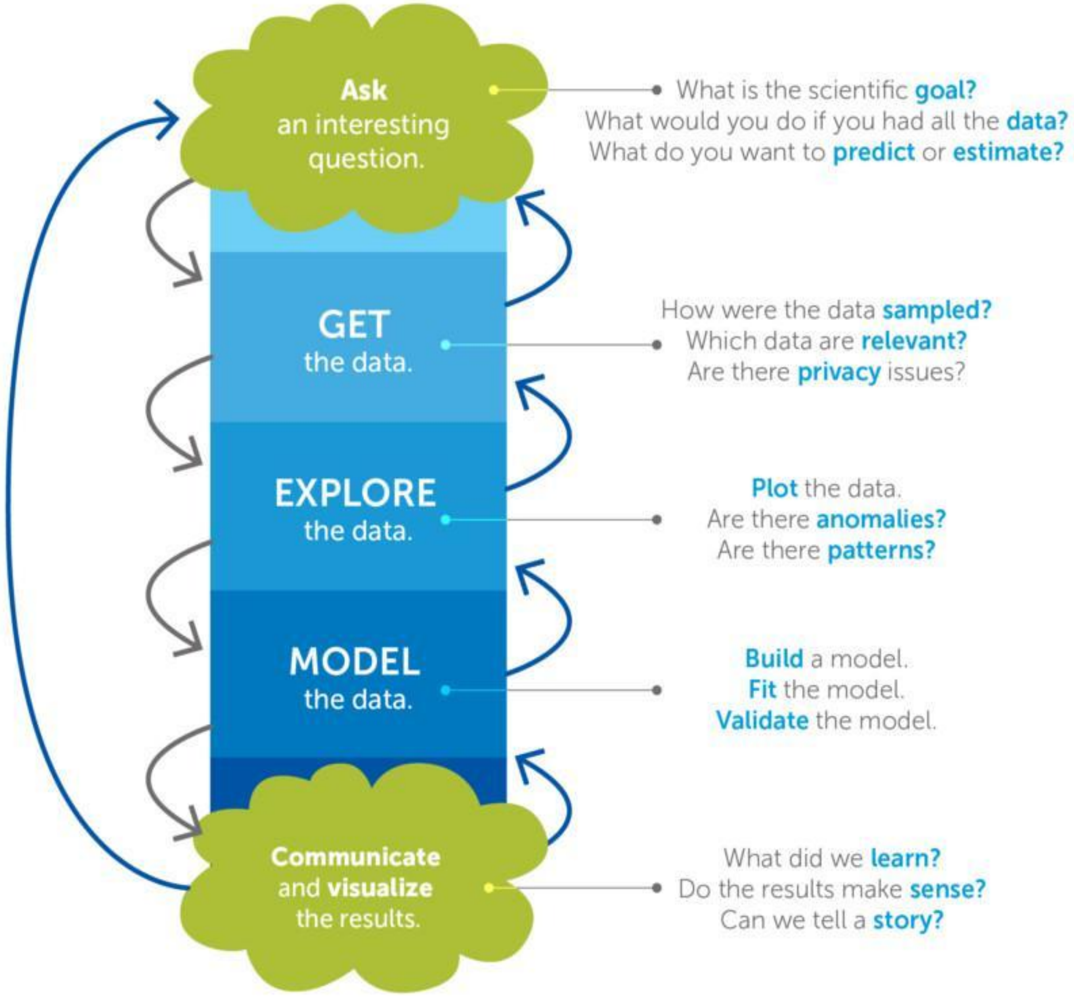
- In 2012, Target knew that the arrival of babies change buying habits.
- They were interested in whether they could ***predict*** that people *are* ***expecting*** a baby, **predict they are pregnants**, so they can make offers before competitors.
- E.g., pregnant mothers often change their diets, their wardrobes, their vitamin regimens, and so on.

Data Science Process

The Data Pipeline

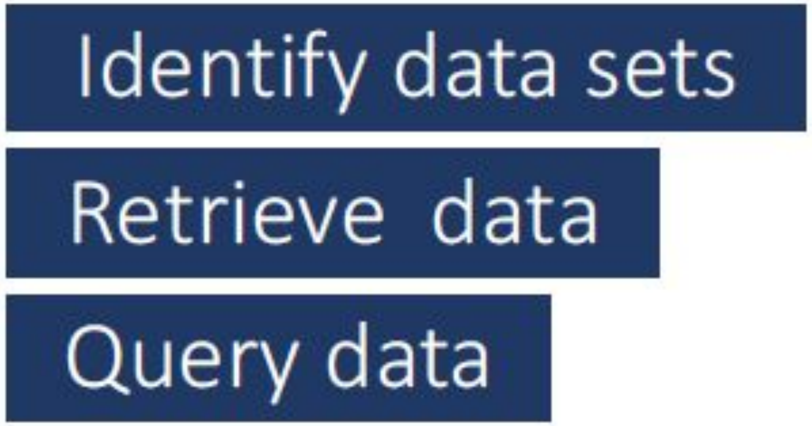
The

Data Science Process

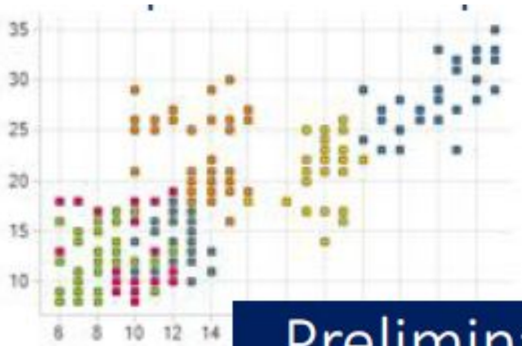


Derived from the work of Joe Blitzstein and Hanspeter Pfister, originally created for the Harvard data science course http://cs109.org/.

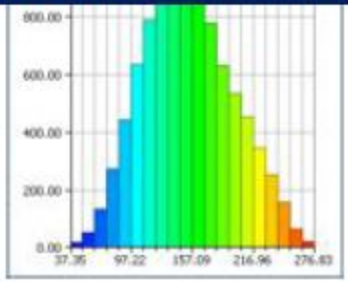
Step 1: Acquire Data



Step 2: Prepare Data



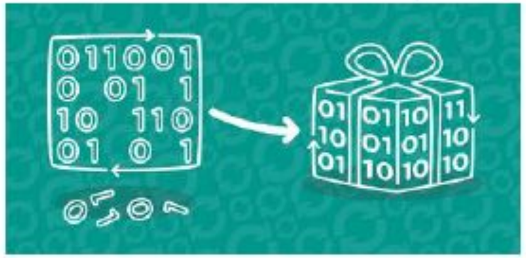
Understand nature of data



Step 2-A: Explore

Preliminary analysis

Step 2-B: Pre-process



Clean

Integrate

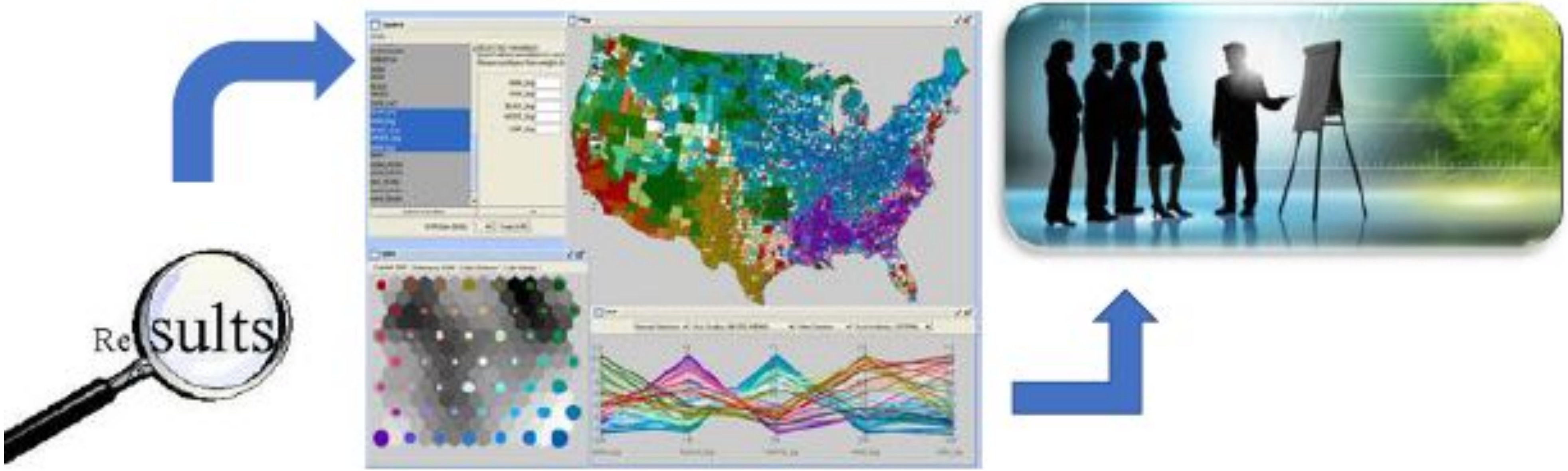
Package

Step 3: Analyze Data

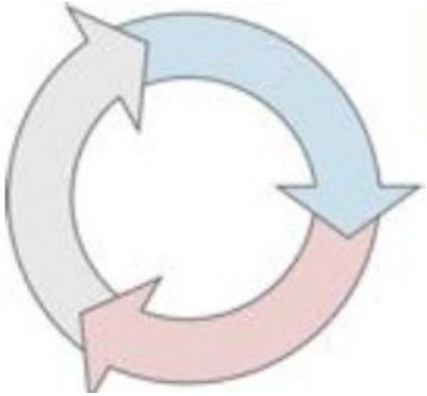
Select analytical techniques

Build models

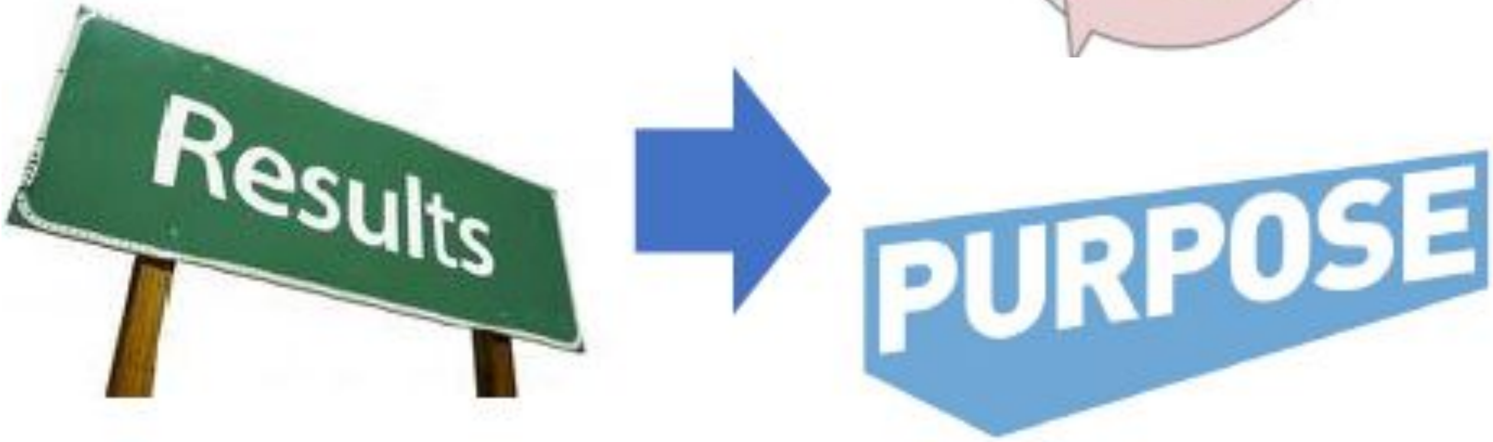
Step 4: Report/Communicate Results

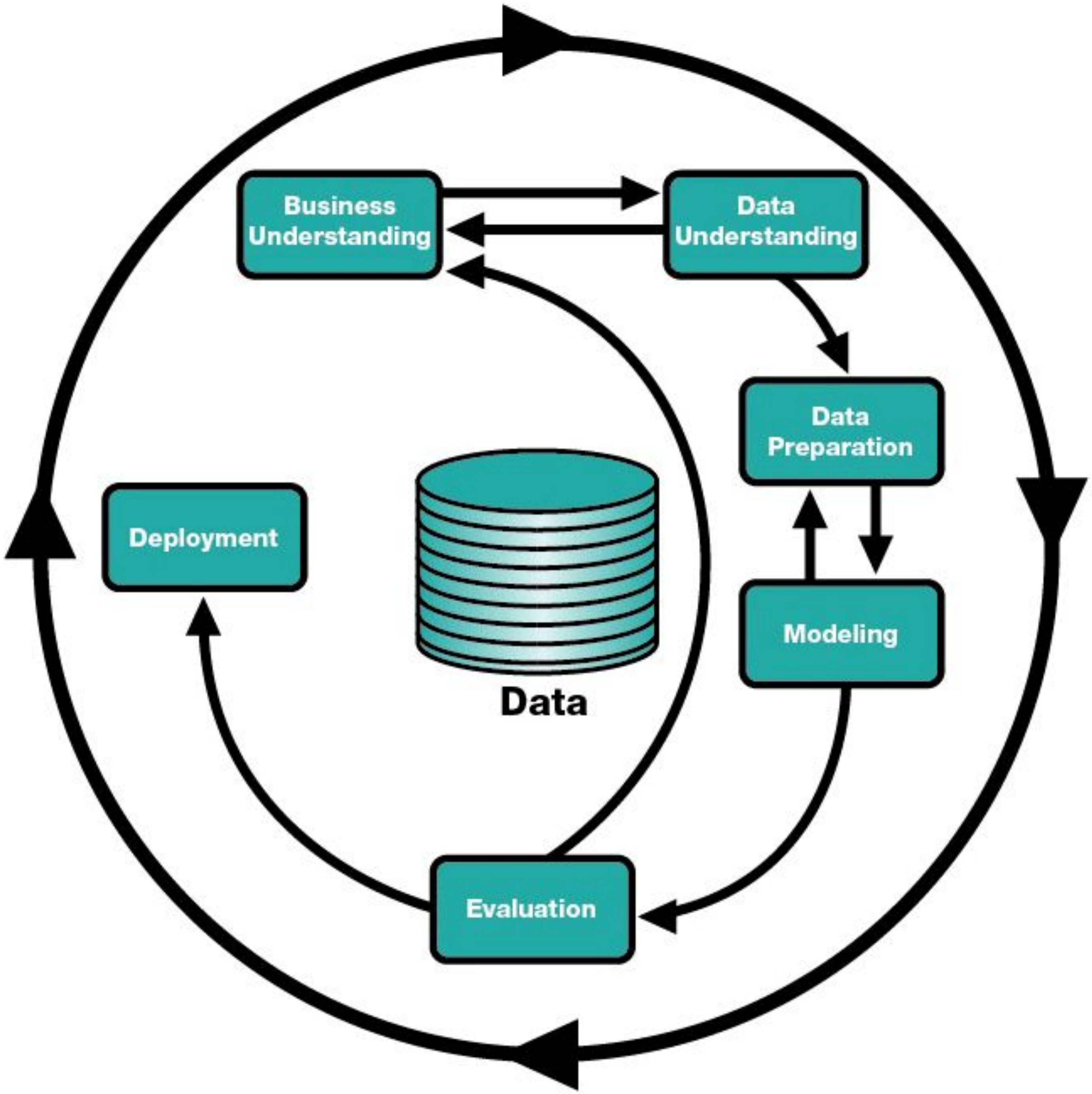


Step 5: Apply Results

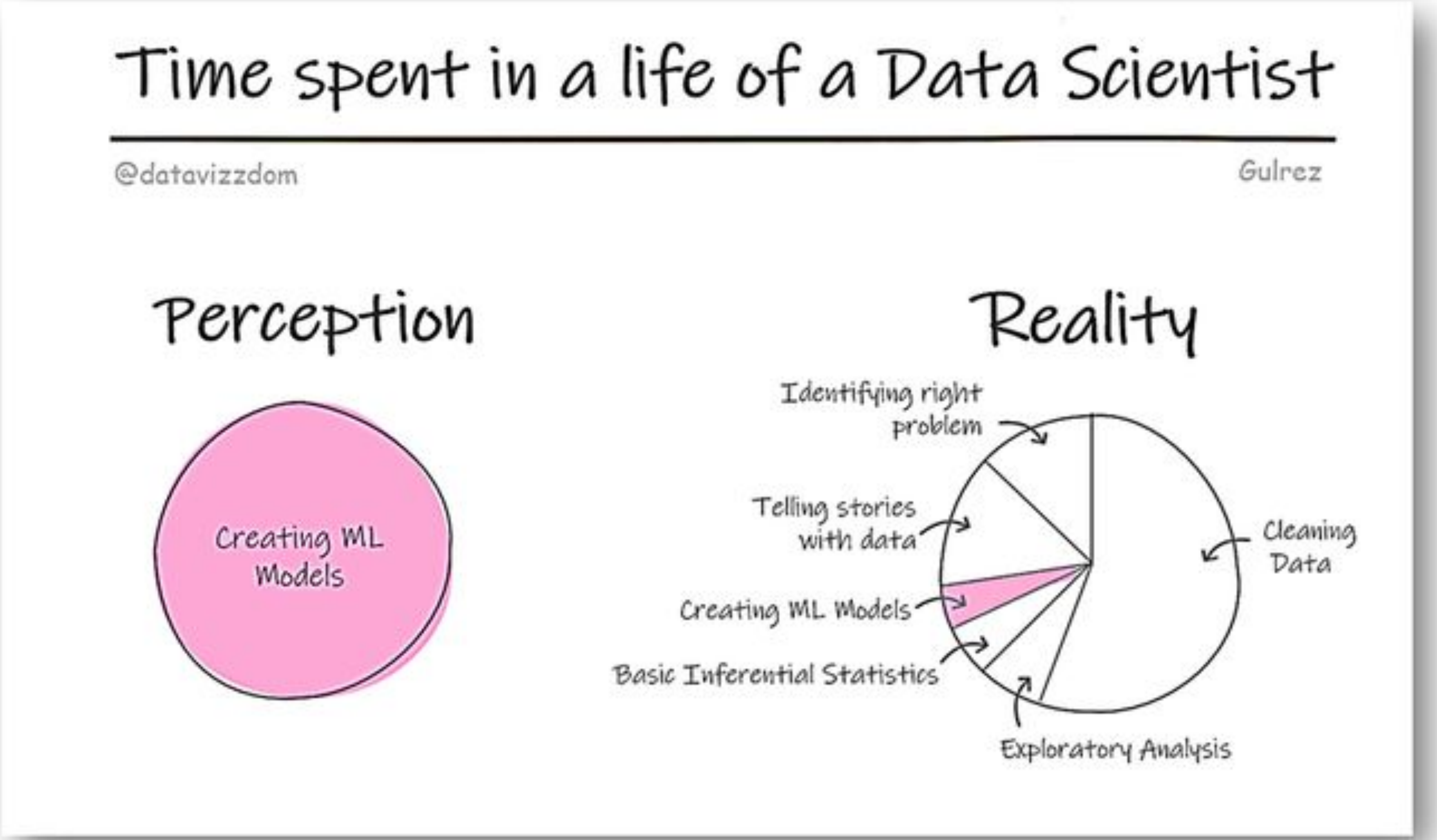


Iterative process





Cross Industry Standard Process for Data Mining (CRISP-DM)



- Twitter @DataVizzdom, Aug 25, 2020.

Analytical Thinking, Asking Questions, Defining Problems

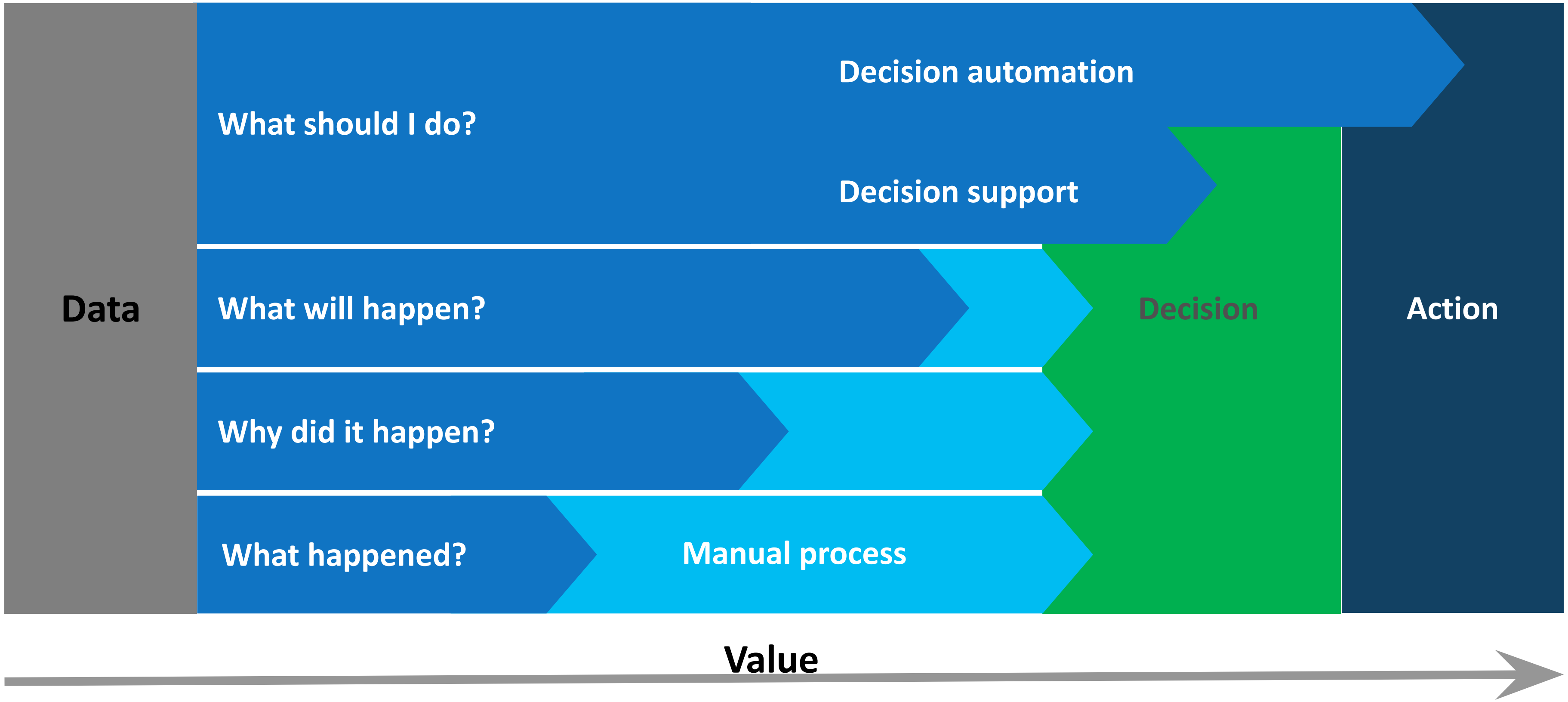
A problem well defined is a problem half-solved

— Charles Kettering

Type of Analytics

Source: IBM's Introduction to Data Analytics on Coursera

Data → Decisions → Actions



EXAMPLE: 4 Layers of Analytics Questions

Value

[Descriptive Analytics] What happened	• How was our sales last quarter?
[Diagnostic Analytics] Why did it happen	• Which factors influence our customers' purchase?
[Predictive Analytics] What will happen	• How can we identify if our customers are churning?
[Prescriptive Analytics] What should I do	• If we can predict our customers' churn, how can we design our customer's retention policy?

Activity: Lower churn

Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics

Question analysis

1. How many customers did we lose during the last 3 months?
2. How can we identify if our customers are going to churn during the next quarter?
3. If we can predict our customers' churn, how can we design our customer's retention policy?
4. How can we determine the factors that influence our customer churn?
5. Which area show highest customers' churn?
6. What type of customers that tend to have higher churn rate?
7. What's the cost of our customer retention plan?

Asking “Sharp” Questions

- What’s going to happen with my stock?
- How’s my car fleet is doing?

- Instead,

- What will my stock’s sales price be next week?
- Which car in my fleet is going to fail first?

- A sharp question can be answered with a name or a number.

Questions and Data Mining Tasks

- How much or how many? (regression/value estimation)
  - How much will a given customer use the service?
  - How much should we price the house for sales?
- Which category? (classification/probability estimation)
  - Among all the customers, which are likely to respond to a given offer?
  - Will some particular new customer be profitable?
- Which group? (clustering)
  - Do our customers form natural groups or segments?
  - What products should we offer or develop?
  - How should our customer care teams (or sales teams) be structured?

— Microsoft, The business understanding stage of the Team Data Science Process lifecycle

Questions and Data Mining Tasks

- Is this weird? (anomaly detection/profiling/behavior description)
  - What is the typical cell phone usage of this customer segment?
  - What kind of purchases a person typically makes on a credit card?
  - Is internet traffic atypical at a certain time?

- Which option should be taken? (recommendation/co-occurrence/market-basket analysis)
  - Which movies should Netflix recommend to subscribers?
  - Which products will likely be purchased together?

— Microsoft, The business understanding stage of the Team Data Science Process lifecycle

DATA SCIENCE PROJECT CANVAS

Designed by: Date:

Title:

1. Problem Statement

What problem are you trying to solve? What larger issues do the problem address?

2. Outcomes/Predictions

What prediction(s) are you trying to make? Identify applicable predictor (X) and/or target (y) variables.

3. Value Propositions

What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

4. Data Acquisition

Where are you sourcing your data from? Is there enough data? Can you work with it?

5. Modeling

What models are appropriate to use given your outcomes?

6. Model Evaluation

How can you evaluate your model performance?

7. Data Preparation

What do you need to do to your data in order to run your model and achieve your outcomes?

Modified from Bill Schmarzo's Machine Learning Canvas and Jasmine Vasandani's Data Science Workflow Canvas for DS511 Data Science @MSDS SWU

DATA SCIENCE PROJECT CANVAS (EXAMPLE)

Designed by: Chantri

Date: 4 Aug 2024

Title: Predict the level of PM2.5 24 hours in advance using data from weather monitoring stations in Bangkok

1. Problem Statement

What problem are you trying to solve? What larger issues do the problem address?

During the past 2-3 years, population in Bangkok suffered from high level of PM2.5 dust affecting their quality of life. By having an accurate prediction system, responsible authority can come up with proper measures to handle the situation. How do we accurately predict the level of PM2.5?

2. Outcomes/Predictions

What prediction(s) are you trying to make? Identify applicable predictor (X) and/or target (y) variables.

Predictor: past data from weather monitoring stations in Bangkok from Pollution Control Department, wind, humidity

Target: level of PM2.5 24 hours in advance (microgram/m^3)

3. Value Propositions

What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

By obtaining accurate prediction system, responsible authority from government sectors can issue measures that can effectively handle the situation.

4. Data Acquisition

Where are you sourcing your data from? Is there enough data? Can you work with it?

Request for data from Pollution Control Department

5. Modeling

What models are appropriate to use given your outcomes?

Supervised Machine Learning using Regression techniques such as Deep Learning-based Long Short-Term Memory (LSTM)

6. Model Evaluation

How can you evaluate your model performance?

Regression metric such as

-R^2 score

-Root mean-square error (RMSE)

-Mean square error (MSE)

-Mean absolute percentage error (MAPE)

7. Data Preparation

What do you need to do to your data in order to run your model and achieve your outcomes?

For time-series data, we may have to

- Handle missing data using moving average
- Use past data for 7 days to predict the level of PM2.5

Modified from Bill Schmarzo's Machine Learning Canvas and Jasmine Vasandani's Data Science Workflow Canvas for CP-DSAI @AIT

DATA SCIENCE PROJECT CANVAS (EXAMPLE)

Designed by: Chantri

Date: 4 Aug 2024

Title: WNBA K-Means Clustering to Find the Best Teams

1. Problem Statement

What problem are you trying to solve? What larger issues do the problem address?

Sports data analysis rarely includes women's sports.

How might I apply machine learning algorithms to women's sports data?

How might I design the best WNBA teams?

2. Outcomes/Predictions

What prediction(s) are you trying to make? Identify applicable predictor (X) and/or target (y) variables.

Outcomes: ranked teams comprised of all-time WNBA playes.

Predictor variables: player stats

Outcomes: ranked teams

3. Value Propositions

What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

To find the greatest female players across all teams and sports genres playing on the same team.

Make an American dream team (women side).

4. Data Acquisition

Where are you sourcing your data from? Is there enough data? Can you work with it?

Basketball-reference.com has comprehensive WNBA player stats, and it's relatively easy to scrape.

5. Modeling

What models are appropriate to use given your outcomes?

Unsupervised machine learning algorithm.

K-Means clustering, good that it clusters outliers.

6. Model Evaluation

How can you evaluate your model performance?

K-Means clustering evaluation metrics:

- silhouette score
- inertia

7. Data Preparation

What do you need to do to your data in order to run your model and achieve your outcomes?

- Get summary player stats.
- Divide players by position.
- Find best-performing characteristics per position to create ranked teams.

DATA SCIENCE PROJECT CANVAS (EXAMPLE)

Designed by: Chantri

Date: 4 Aug 2024

Title: Fake News Detector

1. Problem Statement

What problem are you trying to solve? What larger issues do the problem address?

WhatsApp deletes 2 million "fake news" accounts every month.

How do they do that?

How can I detect between fake news and real news?

2. Outcomes/Predictions

What prediction(s) are you trying to make? Identify applicable predictor (X) and/or target (y) variables.

Predictor variables: text, news headlines, news text

Target variables: fake news (1) or not fake news (0)

Want to predict if a news article is fake news or not fake news.

3. Value Propositions

What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?

I will have a better understanding of how WhatsApp might have created a model to detect fake news accounts by creating fake news detection.

Maybe I can use fake news detector for my own application.

4. Data Acquisition

Where are you sourcing your data from? Is there enough data? Can you work with it?

Reddit has subreddits called "The Onion" and "Not the Onion."

Scraping posts is relatively easy, and each subreddit has enough data for me to use.

5. Modeling

What models are appropriate to use given your outcomes?

My predictions will be discrete (1 for fake news, 0 for not fake news), and I have a labeled dataset.

So I'll test out some classification models, and will also rely on natural language processing vectorizers since I'm working with text.

6. Model Evaluation

How can you evaluate your model performance?

Depending on which models I use, I can interpret my coefficients and/or use a confusion matrix.

7. Data Preparation

What do you need to do to your data in order to run your model and achieve your outcomes?

Since I'm working with text, I need to use NLP methodologies to analyze my text.

- Count Vectorizer
- Term Frequency-Inverse Document Frequency (tf-idf)

Modified from Bill Schmarzo's Machine Learning Canvas and Jasmine Vasandani's Data Science Workflow Canvas for DS511 Data Science @MSDS SWU

Course Goal and Logistics

Course Goal

- By the end of the course you should be able to frame business/research questions , find useful datasets, perform basic and advanced data analysis using Python to help answer your questions, evaluate your results, and present your findings.

- Learning Objectives
  - Basic process of data science
  - An applied understanding of how to manipulate and analyse datasets
  - How to effectively visualize results
  - Basic statistical analysis and machine learning methods
  - Model Evaluation
  - Deployment and results monitoring

Why Python for Data Science?

- Easy-to-read and learn
- Vibrant community
- Growing and evolving set of libraries
  - Data management
  - Analytical processing
  - Visualization
- Applicable to each step in the data science process
- Notebooks

https://medium.com/@atillaguzel/popularity-of-data-science-python-and-pythons-major-libraries-f7146e202e5d

Course Logistics

Grade Distribution
- Quiz/Lab/Homework 25%
- Participation (lecture&lab) 5%
- Midterm Exam 25%
- Final Exam 25%
- Final Project 20%

Check attendance during the first 15 mins of the class.

Score	Grade
x >=80	A
80 > x >=75	B+
75 > x >=70	B
70 > x >=65	C+
65 > x >=60	C
60 > x >=55	D+
55 > x >=50	D
50 > x >= 0	E

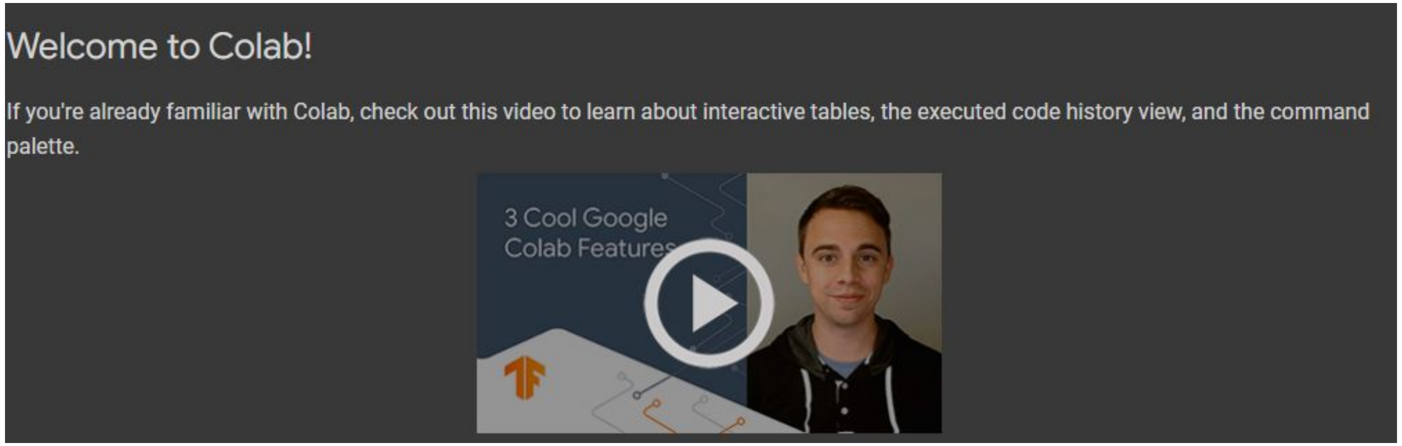
*May combine with group-based grading

Lecture: Tuesday (9:00-12:00), Lab: Tuesday (13:00-16:00)

	Date	Lecture	Lab/HW
1	6-Aug	Introduction to Data Science	Intro to Python
2	13-Aug	Data Acquisition	Numpy
3	20-Aug	Data Acquisition + Wrangling	Numpy + Pandas
4	27-Aug	Data Wrangling	Numpy + Pandas
5	3-Sep	Data Visualization	Matplotlib/Seaborn
6	10-Sep	Exploratory Data Analysis	
7	17-Sep	Introduction to Machine Learning	Sklearn
8	24-Sep	Midterm Exam	
9	1-Oct*	Model Selection	Proposal presentation
10	8-Oct	Supervised learning	Regression
11	15-Oct	Unsupervised learning	Classification
12	22-Oct	Data Preprocessing and Pipelining	
13	29-Oct	Evaluations	
14	5-Nov	Deep Learning*	Pytorch
15	12-Nov	Final Project	
16	19-Nov	Final Exam	

Google colab

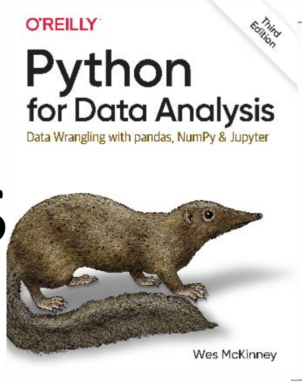
- https://colab.research.google.com/?utm_source=scs-index
- What is Colab?
  - Colab, or "Colaboratory", allows you to write and execute Python in your browser, with
  - Zero configuration required
  - Access to GPUs free of charge
  - Easy sharing

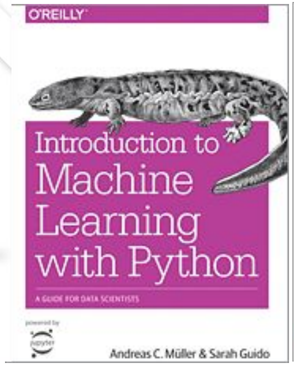


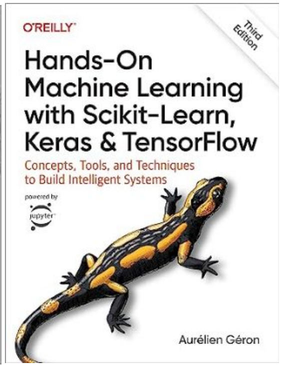
Anaconda

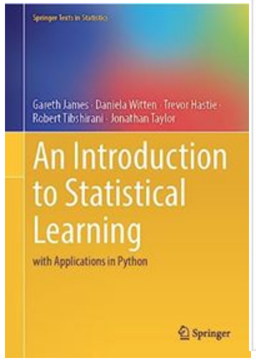
- A distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.)
- Aim to simplify package management and deployment.
- Distribution includes data-science packages suitable for Windows, Linux, and macOS.

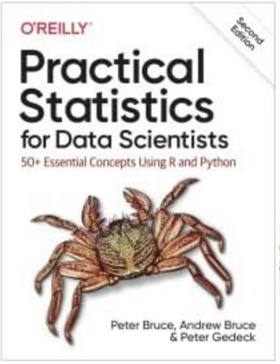
Book References

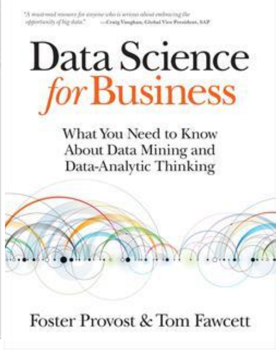








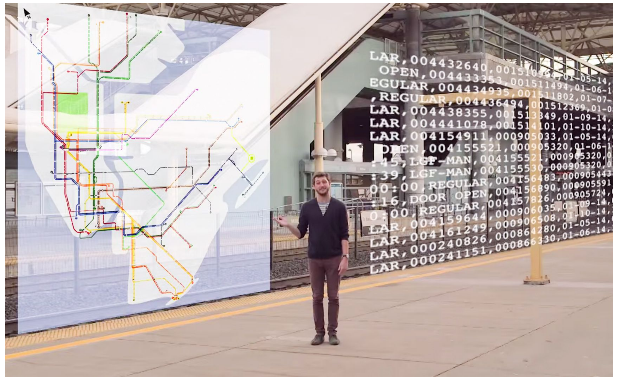




- Wes McKinney. ***Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython***. O'Reilly Media; 3rd edition (2022). Open Edition Online
- Jake VanderPlas. ***Python Data Science Handbook.*** ISBN: 978-1491912058 Free online
- Andreas C. Müller and Sarah Guido. ***Introduction to Machine Learning with Python: A Guide for Data Scientists***. O'Reilly Media; 1st edition 2017. ISBN: 978-1449369415
- Aurélien Géron. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. O'Reilly Media; 3rd edition 2022. ISBN: 978-1098125974
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. ***An Introduction to Statistical Learning with Applications in Python***. Springer, 2023 1st edition. https://www.statlearning.com/ (Free)
- Peter Bruce and Andrew Bruce. ***Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python.*** O'Reilly Media; 2nd edition 2020. ISBN: 978-1491952962
- Foster Provost and Tom Fawcett. ***Data Science for Business***. O'Reilly Media; 1st edition August 2013. ISBN: 978-1449361327
- Computational and Inferential Thinking: The Foundations of Data Science, UC Berkeley (Free online)

Course References

- **edX**: Python for Data Science, UCSanDiegoX (DSE200x). DAT203.1x Data Science Essentials, Professional Certificate in Python Data Science
- **Udacity**: Intro to Data Analysis (UD170), Intro to Machine Learning Udacity (UD120), Intro to Data Science (UD359).
- **Coursera**: Introduction to Data Science in Python; Applied Plotting, Charting & Data Representation in Python; Applied Machine Learning in Python, University of Michigan.
- **Data8**: Computational and Inferential Thinking: The Foundations of Data Science



Additional References

- Carnegie Mellon University: CMU Practical Data Science
  http://www.datasciencecourse.org/
- Machine Learning Canvas
  https://www.digitalistmag.com/cio-knowledge/2018/10/29/data-science-paint-by-numbers-with-hypothesis-development-canvas-06191989/
- Data Science Workflow Canvas
  https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0