LECTURER: Nghia Duong-Trung

DATA SCIENCE

WHO I AM

- Name: Nghia Duong-Trung
- 09.2022 – present: The German Research Center for Artificial Intelligence (DFKI GmbH)
- 06.2022 – present: IU International University of Applied Sciences

- PostDoc in Machine Learning at Technische Universität Berlin
- PhD in Machine Learning at The Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim, Germany
- MSc in Software Engineering at Heilbronn University, Germany

- Profile: https://sites.google.com/ismll.de/duongtrungnghia/

- Course book: Data Science – DLMBDSA01, provided by IU, myStudies
- Reading list DLMBDSA01, provided by IU, myStudies
- Additional teaching materials:

https://github.com/duongtrung/IU-DataScienceCourse

SELF-LEARNING AND SELF-IMPROVING

- https://www.dataquest.io/blog/learn-data-science/
- https://blog.edx.org/7-learning-tips-for-data-science-self-study
- https://www.coursera.org/search?query=data%20science&
    - 2680 results for "data science“
    - https://www.coursera.org/specializations/introduction-data-science
    - https://www.coursera.org/specializations/data-science-python
    - https://www.coursera.org/specializations/data-science-fundamentals-python-sql

SELF-LEARNING AND SELF-IMPROVING

- Should read the course book before class
- *Optional*: reading list

TOPIC OUTLINE

Introduction to Data Science	1
Use Cases and Performance Evaluation	2
Data Preprocessing	3
Processing of Data	4
Selected Mathematical Techniques	5
Selected Artificial Intelligence Techniques	6

UNIT 1

INTRODUCTION TO DATA SCIENCE

STUDY GOALS

On completion of this unit, you will have learned…

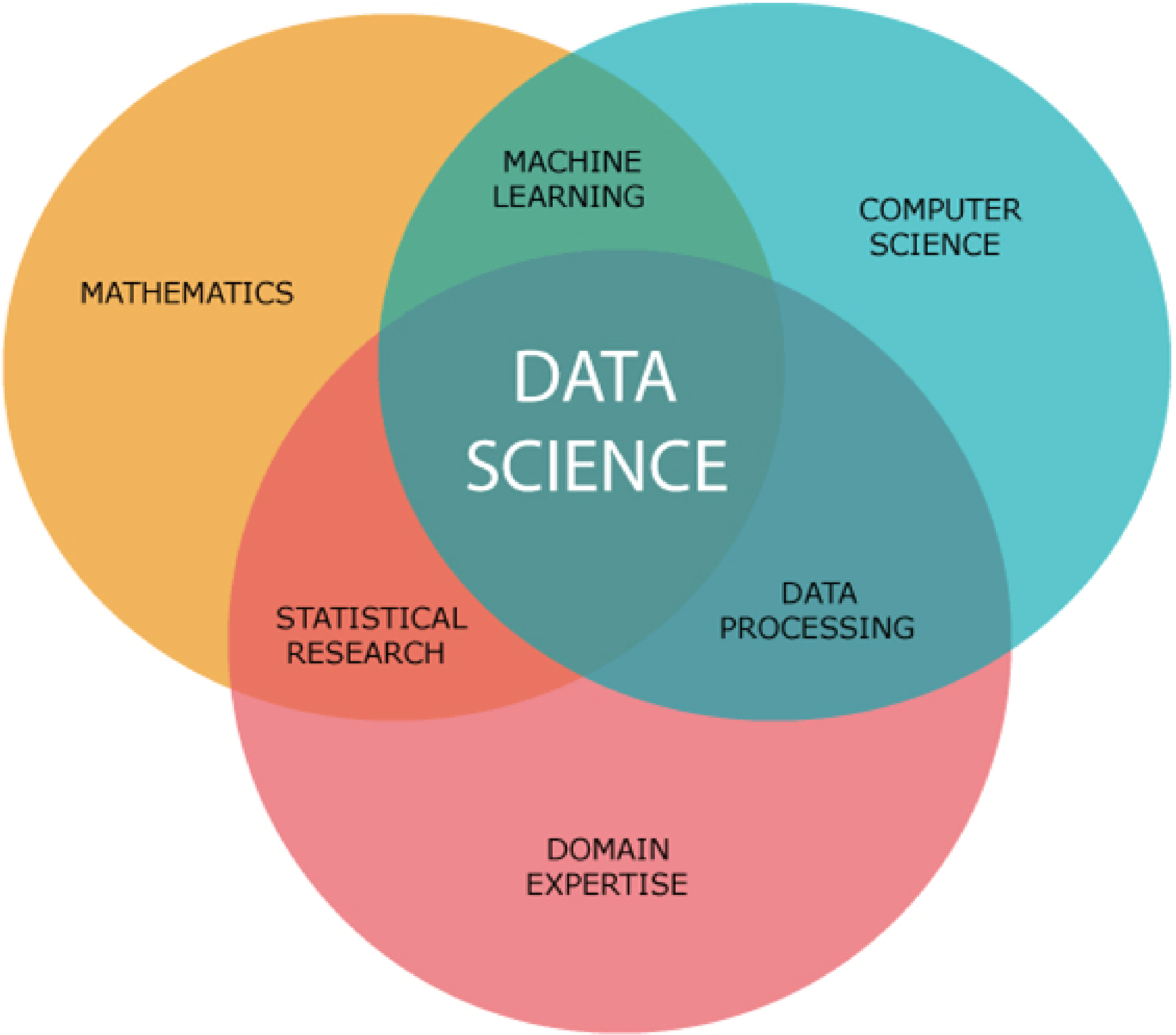
- the meaning of data science.
- common terms and definitions in data science.
- the different applications of data science.
- the typical sources of data.
- the types and shapes of data.
- probability distributions and Bayesian statistics.

EXPLAIN SIMPLY

1. Define the term data science in your own words.
2. Explain the difference between structured, unstructured and semi-structured data.
3. Identify two types of machine learning and give an application example for each type.

Data science

- analyze and explore the information contained in data
- incorporate domain knowledge
- create predictions to advise the decision-making process
- create value from data

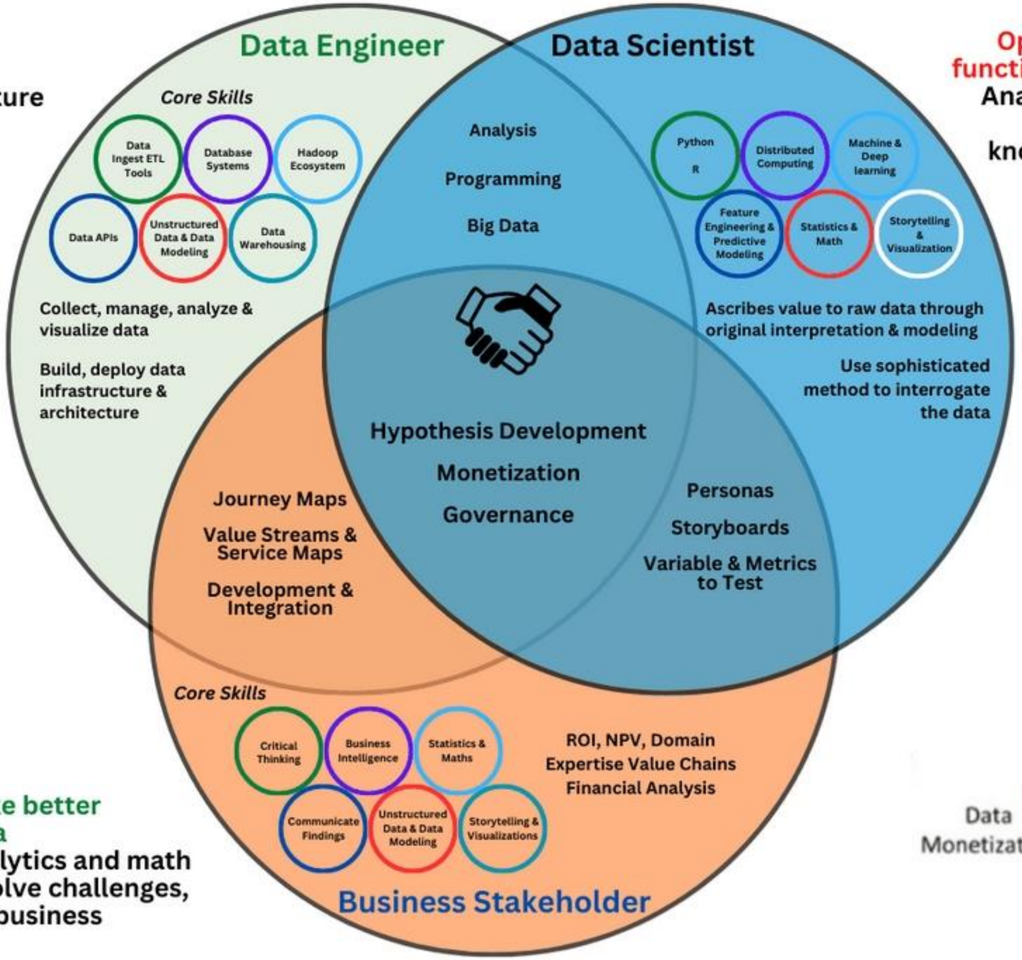


K21Academy
Learn Cloud from Experts

Navigating the Job Market for Data Science Professionals: Roles and Requirements.

Enable data access & utilization & enable value capture
Builds and support the infrastructure of data pipe and associated SW engineering infrastructure tasks.

Optimize & enable data for business & functional value capture & value creation
Analysis and interpretation of complex digital data to extract or discover knowledge and assist decision making.



Data Engineer

Core Skills

- Data Ingest ETL Tools
- Database Systems
- Hadoop Ecosystem
- Data APIs
- Unstructured Data & Data Modeling
- Data Warehousing

Collect, manage, analyze & visualize data

Build, deploy data infrastructure & architecture

Data Scientist

- Python
R
- Distributed Computing
- Machine & Deep learning
- Feature Engineering & Predictive Modeling
- Statistics & Math
- Storytelling & Visualization

Ascribes value to raw data through original interpretation & modeling

Use sophisticated method to interrogate the data

Analysis

Programming

Big Data

Hypothesis Development

Monetization

Governance

Journey Maps

Value Streams & Service Maps

Development & Integration

Personas

Storyboards

Variable & Metrics to Test

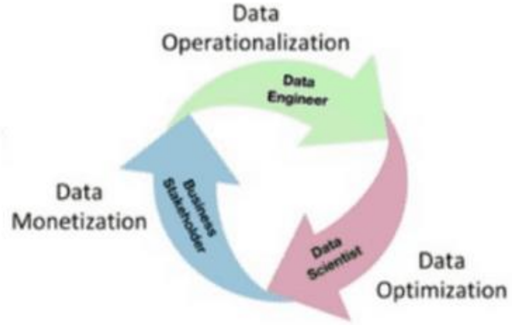
Core Skills

- Critical Thinking
- Business Intelligence
- Statistics & Maths
- Communicate Findings
- Unstructured Data & Data Modeling
- Storytelling & Visualizations

ROI, NPV, Domain Expertise Value Chains Financial Analysis

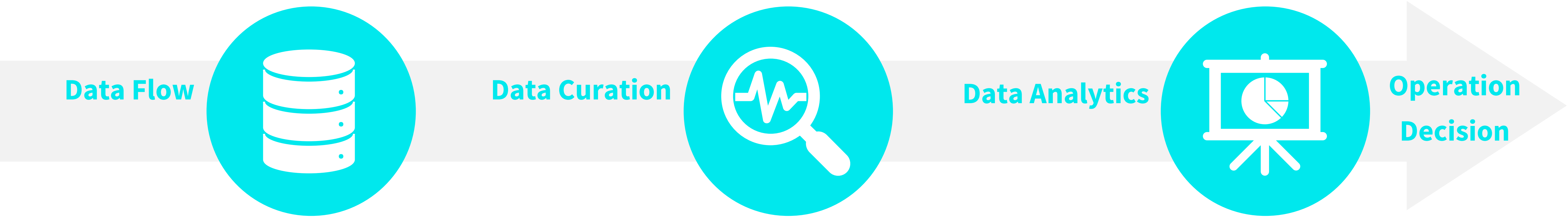
Business Stakeholder

Help the business make better decisions through data
Blend of business, analytics and math skills to explore and solve challenges, bridging the data and business communities.



https://k21academy.com/data-science/data-science-empowering-todays-world-businesses-and-careers/

DATA SCIENCE ACTIVITIES



Data Flow

- Data collection from different sources
- Data storage
- Data accessing

Example of customer churn: Combine data from historical marketing interactions and purchases with demographic data

Data Curation

- Data cleaning
- Data presentation
- Data evaluation

- Treat outliers and missing values
- Inspect visual patterns

Data Analytics

- Descriptive statistics & statistical analysis
- Modeling
- Visual techniques

- Build ML model to predict probability of customers leaving
- Create value from data insights
- Drive business decisions

Operation Decision

KEY TERMS

- Data, Database, Information, Data Science
- **Data mining**, Data Visualization and Statistics, **Knowledge Discovery** (KDD), **Pattern Recognition**
- Artificial Intelligence, **Machine Learning**
- Business Intelligence

- Two broad directions:
  - data engineer/scientist
  - or machine learning engineer/scientist

- https://blog.edx.org/data-science-analytics-career-guide

AI

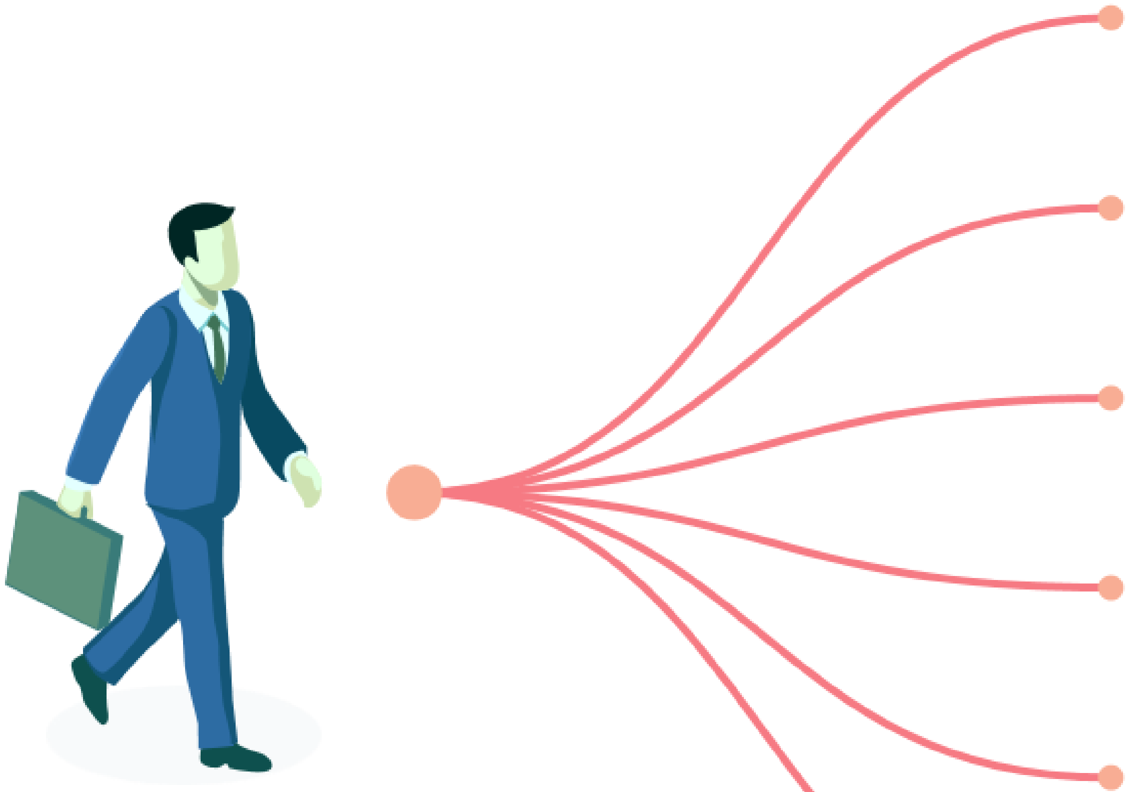
- sounds sexy
- gets us money from VCs
- what we all hope is the future

Machine Learning

- the only real “AI”
- traditionally an academic discipline
- not concerned with real-world software

Data Science

- applies machine learning to create actual products
- deals with real-world complexity



Male

70% of Data Scientists in our research were male

2 Languages

Data scientists speak at least 1 foreign language on average

2 years

This is a new profession. The median experience as data scientists of professionals in our research was 2 years

4.5 years

People who work as data scientists currently have a median work experience of 4.5 years (including previous positions)

R and/or Python

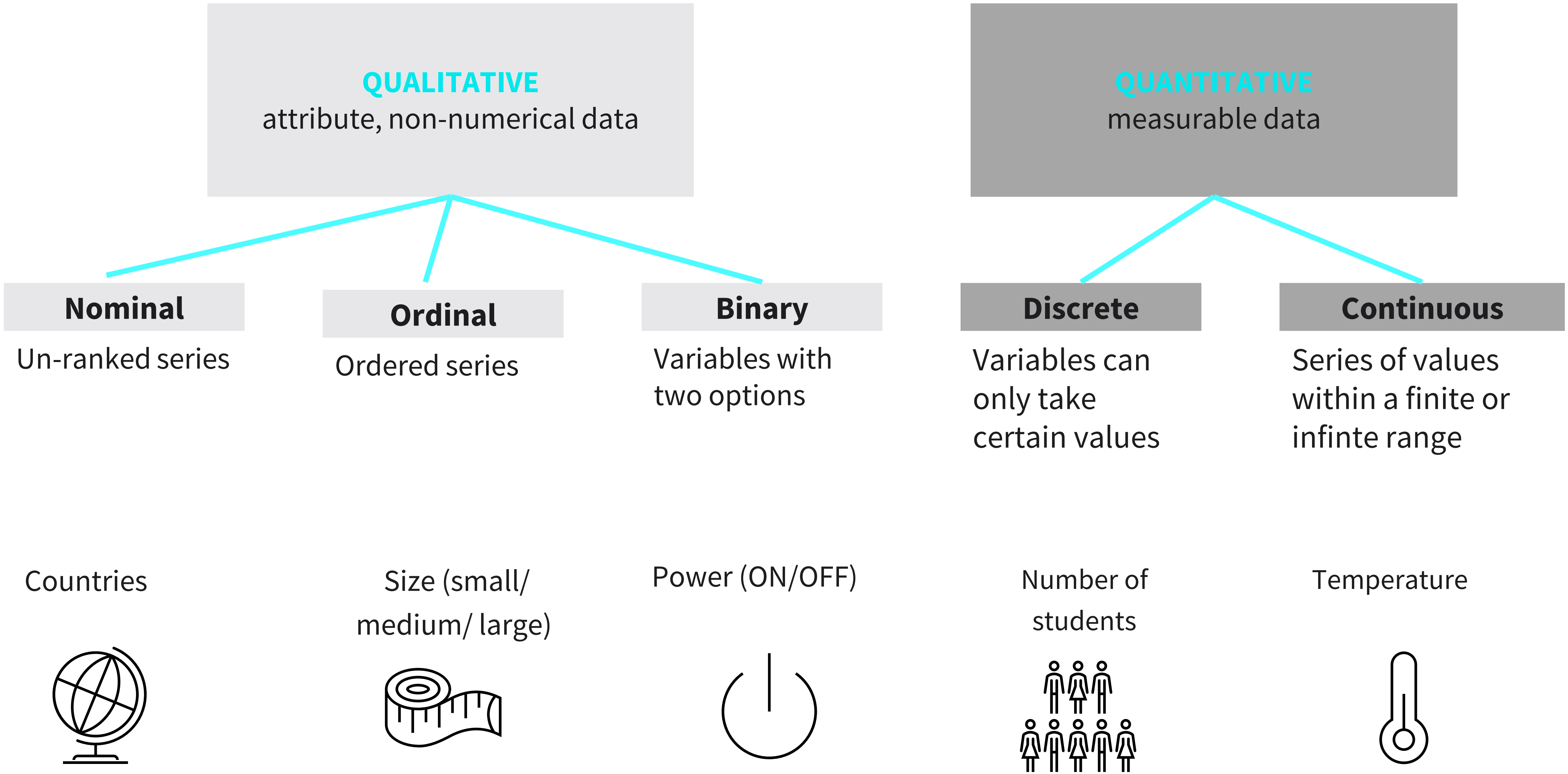
More than 50% of the data scientists in our research work in R and/or Python

Master or PhD

75% of data scientists have a PhD (27%) or a Master (48%) degree

365√DataScience

DATA TYPES



QUALITATIVE
attribute, non-numerical data

QUANTITATIVE
measurable data

Nominal

Un-ranked series

Countries

Ordinal

Ordered series

Size (small/ medium/ large)

Binary

Variables with two options

Power (ON/OFF)

Discrete

Variables can only take certain values

Number of students

Continuous

Series of values within a finite or infinte range

Temperature

DATA SHAPES

Structured Data

- Pre-defined data models
- Can be displayed in rows and columns
- Example: customer database (address, name, age etc.)

Name	Age	Address	Gender
John	30	City	m
Marie	4	Village	f

Semi-structured

- Contains some **tags**/attributes among unstructured data
- Example: Mails, Tweets

From: John Doe johndoe@mail.com
To: Marie Doe mariedoe@mail.com
Subject: Hello

Hi Marie,
How are you?

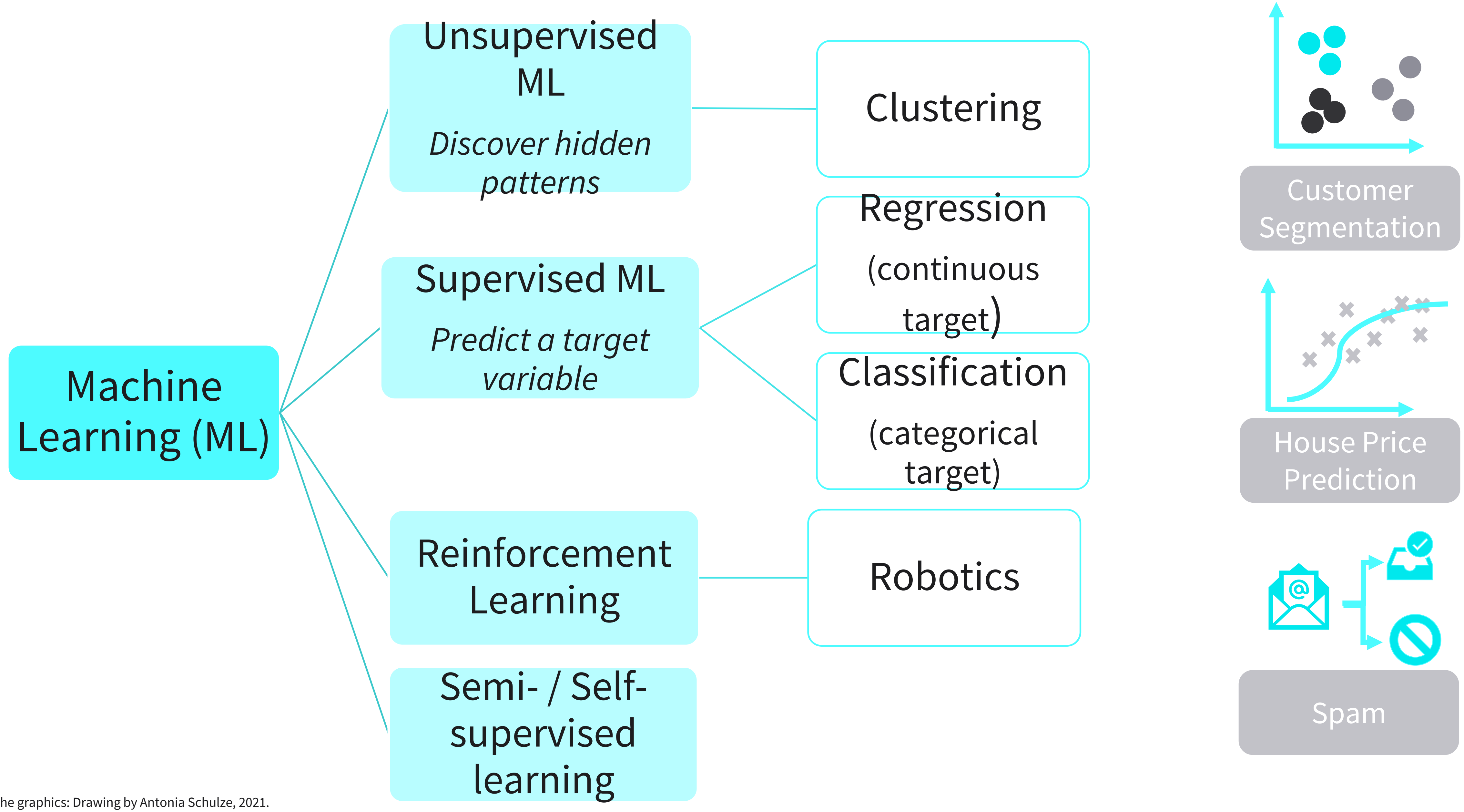
Unstructured Data

- Unknown form or structure
- Example: Online Reviews, Audio files, Videos, Images

★★★

The book is fabulous! I enjoyed it!

TYPES OF MACHINE LEARNING



Machine Learning (ML)

Unsupervised ML

Discover hidden patterns

Clustering

Customer Segmentation

Supervised ML

Predict a target variable

Regression (continuous target)

Classification (categorical target)

House Price Prediction

Reinforcement Learning

Robotics

Spam

Semi- / Self-supervised learning

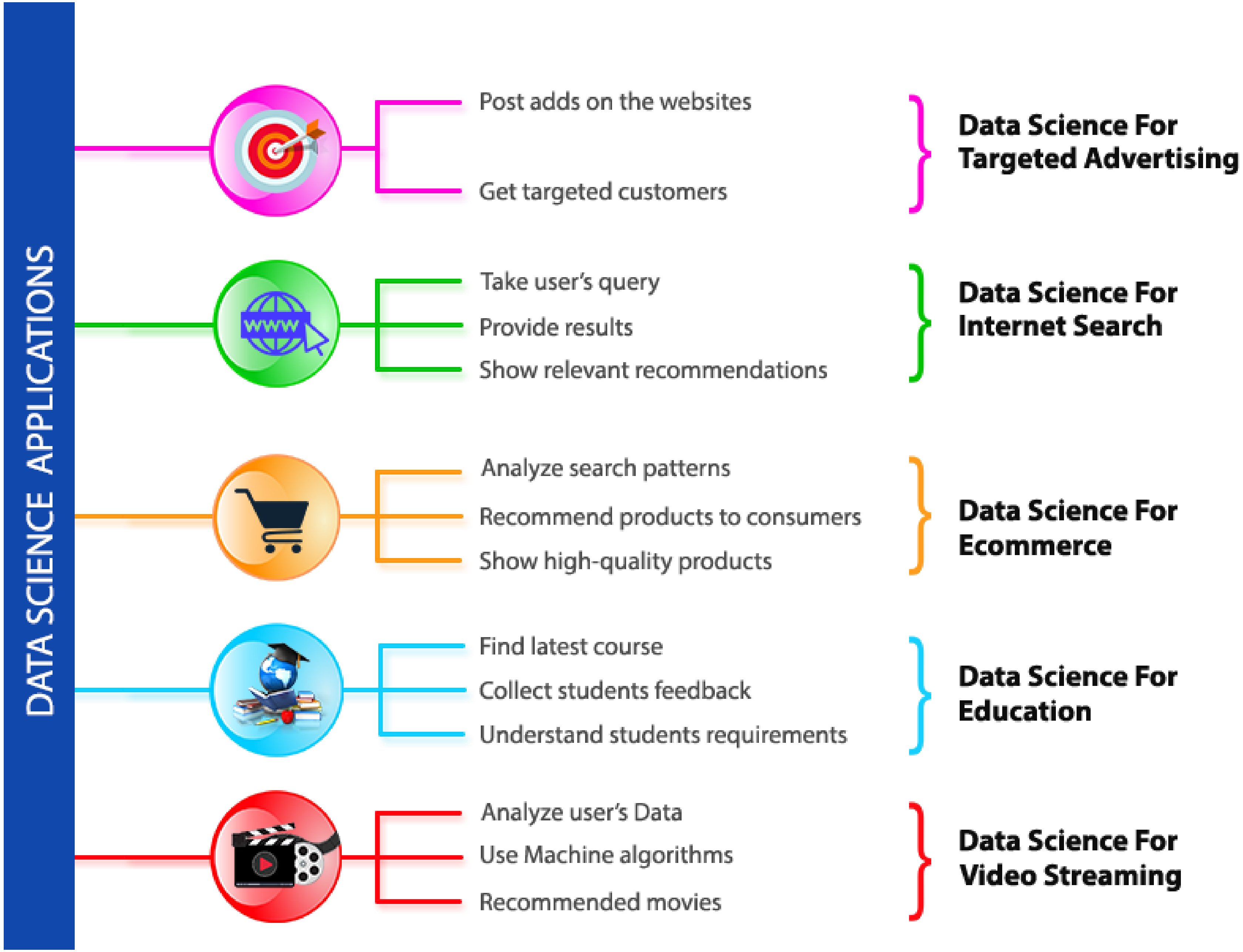
Source of the graphics: Drawing by Antonia Schulze, 2021.

TYPES OF MACHINE LEARNING

- https://elitedatascience.com/learn-machine-learning
- https://programmathically.com/how-to-learn-machine-learning-a-guide-for-self-starters/
- https://machinelearningmastery.com/start-here/

- https://www.coursera.org/search?query=machine%20learning&
  - 1292 results for "machine learning"
  - https://www.coursera.org/specializations/machine-learning-introduction
  - https://www.coursera.org/professional-certificates/ibm-machine-learning

DATA SCIENCE APPLICATIONS



DATA SCIENCE APPLICATIONS

Data Science For Targeted Advertising

- Post adds on the websites
- Get targeted customers

Data Science For Internet Search

- Take user’s query
- Provide results
- Show relevant recommendations

Data Science For Ecommerce

- Analyze search patterns
- Recommend products to consumers
- Show high-quality products

Data Science For Education

- Find latest course
- Collect students feedback
- Understand students requirements

Data Science For Video Streaming

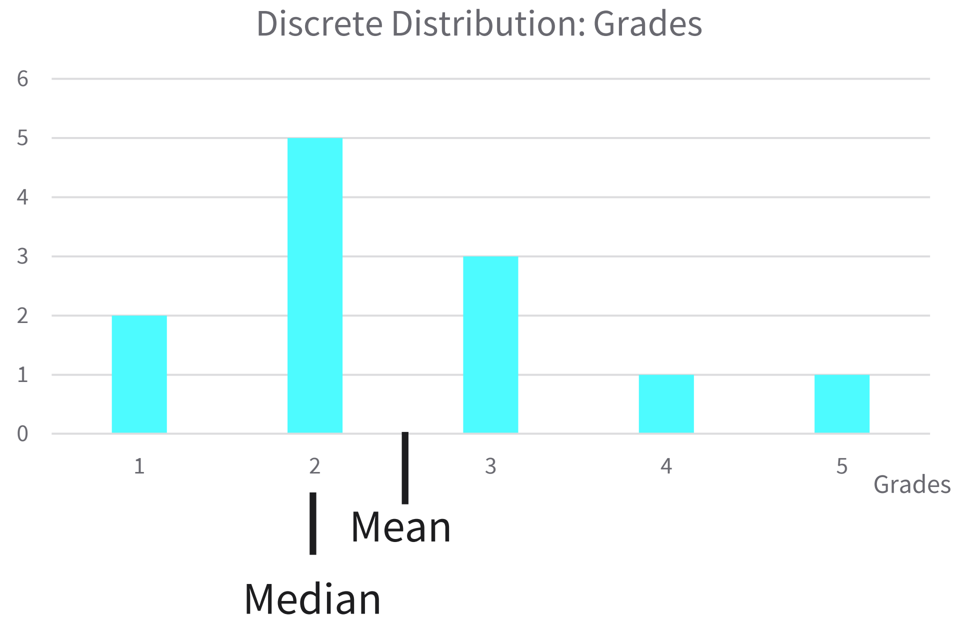
- Analyze user’s Data
- Use Machine algorithms
- Recommended movies

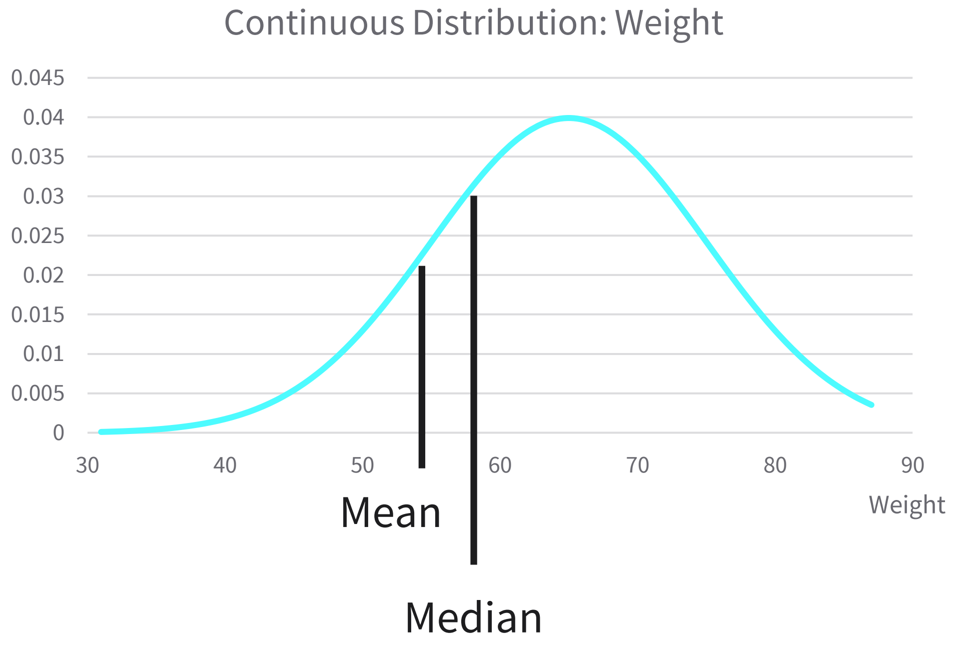
Value, Probability

Standard deviation = measure of spread

Mean = average

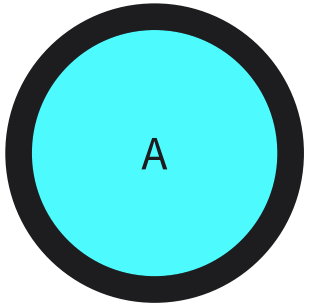
Median = 50% greater, 50% smaller values





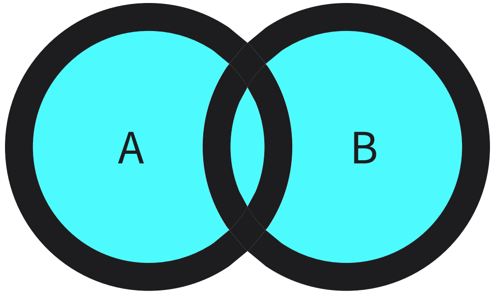
Source of the graphics: Drawing by Antonia Schulze, 2021.

DESCRIPTIVE STATISTICS – PROBABILITY THEORY



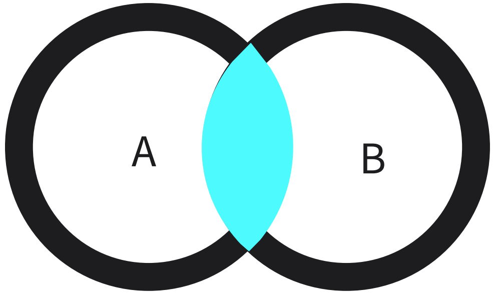
P (A)

Probability of an
event A happening



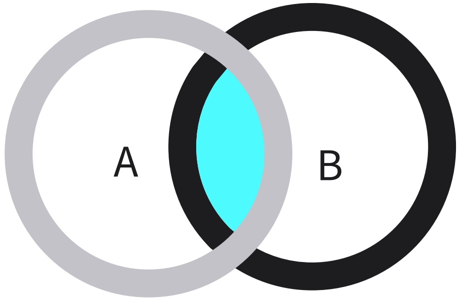
P (A ∪ B)

Probability of event
A or B happening



P (A ∩ B)

Probability of event
A and B happening



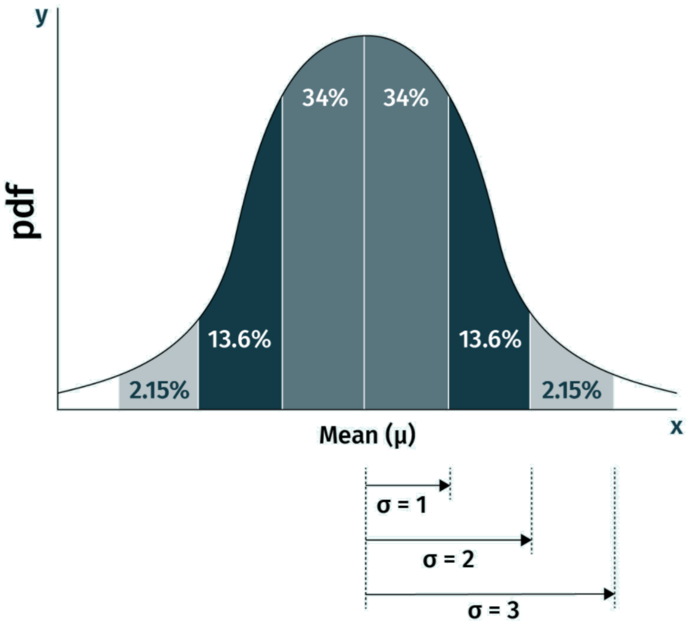
P (A | B)

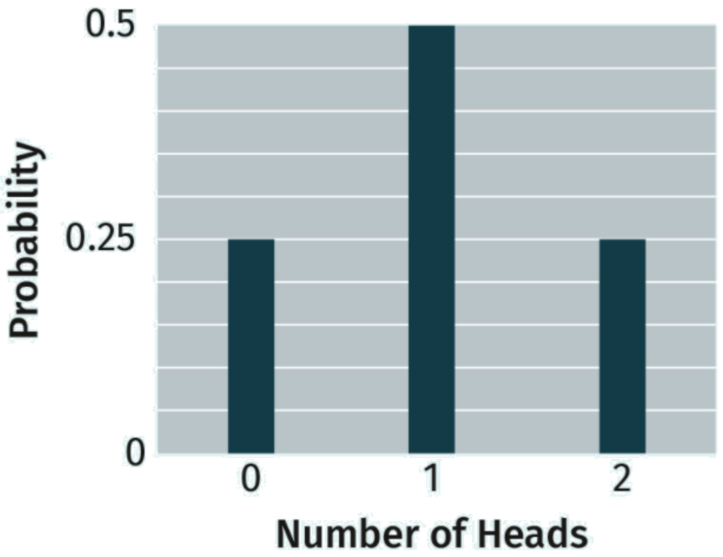
Probability of A, given that
event B already happened

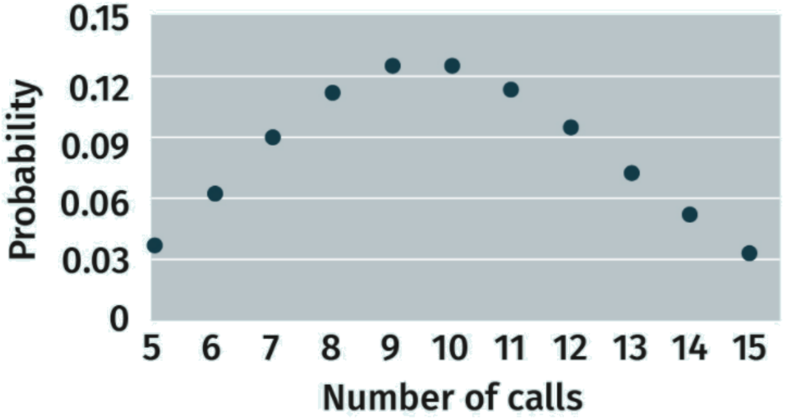
Conditional Probability

$$\frac{\mathbf{P\ (A \cap B)}}{\mathbf{P\ (B)}}$$

DESCRIPTIVE STATISTICS – PROBABILITY DISTRIBUTIONS







Normal Distribution

- Bell curve shape
- *Example: weight, height distribution*

Binomial Distribution

- Two possible outcomes
- *Example: P(# of heads) if toss coin twice*

Poisson Distribution

- Frequency of intervals between independent events
- *Example: P(# of calls per day) if average 5 calls per day*

Source of the graphics: IU International University, DLMBDSA01 Course Book.

BAYES THEOREM

Let us say P(Fire) means how often there is fire, and P(Smoke) means how often we see smoke, then:

- P(Fire|Smoke) means how often there is fire when we can see smoke
- P(Smoke|Fire) means how often we can see smoke when there is fire

Example:

- Dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

Probability of dangerous Fire when there is Smoke: P(Fire|Smoke)?

BAYES THEOREM

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(B|A)P(A) + P(B|not\ A)P(not\ A)}$$

http://allendowney.github.io/ThinkBayes2/chap02.html

REVIEW STUDY GOALSY GOALS

You have learned…

- the meaning of data science.
- common terms and definitions in data science.
- the different applications of data science.
- the typical sources of data.
- the types and shapes of data.
- probability distributions and Bayesian statistics.

SESSION 1

TRANSFER TASK

TRANSFER TASK

Prepare a case study to demonstrate the application of data science in an industry sector of your choice. Elaborate on potential data sources, the type and shape of data.

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

LEARNING CONTROL QUESTIONS

1. Which of the following is the blind machine learning task of inferring a binary function for unlabeled training data?

   a) Regression
   b) Unsupervised Learning
   c) Supervised learning
   d) Data processing

LEARNING CONTROL QUESTIONS

2. In which process are the data cleared from noise and the missing values are estimated/ignored?

   a) data preservation
   b) data security
   c) data publication
   d) data description

LEARNING CONTROL QUESTIONS

3. The probability $p(A|B)$ measures…

   a) the chance of event A given knowledge that event B has occurred.
   b) the chance of event B given knowledge that event A has occurred.
   c) the chance that events A and B occur at the same time.
   d) the chance of event A given knowledge that event B has not occurred.