LECTURER: Nghia Duong-Trung

DATA SCIENCE

WHO I AM

- Name: Nghia Duong-Trung
- Current Employer: The German Research Center for Artificial Intelligence
  - Senior Researcher for Machine Learning Applications
  - Project: https://milki-psy.de/

- PostDoc in Machine Learning at Technische Universität Berlin
  - Project: https://kiwi-biolab.de/
- 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/
- Email: duong-trung.nghia.ext@iu.org

TUTORING SCHEDULE

- 6 weeks, Friday evenings, On Campus, Room BER 2.12 Spandau (Eingang 73 A)

	Date	Time	Title	Event type	Planning status	Attendance
	14.10.2022	17:30 - 20:00	Data Science - MSE_BER_DLMBDSA01_2022_WS_Q4_MAMAN-60+MAINTE-60+MAINTE-120+MADS-120	Tutorial (On Campus)	4. Published	Open
	21.10.2022	17:30 - 20:00	Data Science - MSE_BER_DLMBDSA01_2022_WS_Q4_MAMAN-60+MAINTE-60+MAINTE-120+MADS-120	Tutorial (On Campus)	4. Published	Open
	28.10.2022	17:30 - 20:00	Data Science - MSE_BER_DLMBDSA01_2022_WS_Q4_MAMAN-60+MAINTE-60+MAINTE-120+MADS-120	Tutorial (On Campus)	4. Published	Open
	11.11.2022	17:30 - 20:00	Data Science - MSE_BER_DLMBDSA01_2022_WS_Q4_MAMAN-60+MAINTE-60+MAINTE-120+MADS-120	Tutorial (On Campus)	4. Published	Open
	18.11.2022	17:30 - 20:00	Data Science - MSE_BER_DLMBDSA01_2022_WS_Q4_MAMAN-60+MAINTE-60+MAINTE-120+MADS-120	Tutorial (On Campus)	4. Published	Open
	25.11.2022	17:30 - 20:00	Data Science - MSE_BER_DLMBDSA01_2022_WS_Q4_MAMAN-60+MAINTE-60+MAINTE-120+MADS-120	Tutorial (On Campus)	4. Published	Open

PARTICIPANTS

Vorname	Nachname	E-Mail	MNR	Kohorte	Studiengang
Zoe	Detlefs	zoe.detlefs@iubh.de	9170306	Ber-MSE-MAINTE-120-2021-WS-Q4	MSE MAINTE-120
Marcos	Esteve Hernández	marcos.estevehernandez@iu-study.org	102202680	Ber-MSE-MADS-120-2022-WS-Q4-MM	MSE MADS-120
Niamatullah	Faizi	niamatullah.faizi1@iu-study.org	1222797	Ber-MSE-MAMAN-60-2021-WS-Q1	MSE MAMAN-60
Baris	Gümüs	baris.guemues@iu-study.org	102106169	Ber-MSE-MAINTE-120-2021-WS-Q4	MSE MAINTE-120
Ankita	Kamra	ankita.kamra@iu-study.org	42201667	Ber-MSE-MAINTE-60-2022-SS-Q2-EM	MSE MAINTE-60
Keerthana Reddy	Kolathur	keerthana-reddy.kolathur@iu-study.org	42201695	BH-MSE-MAMAN-60-2022-SS-Q2-EM	MSE MAMAN-60
Reyhane	Mohamadinejad	reyhane.mohamadinejad@iu-study.org	102210426	Ber-MSE-MADS-120-2022-WS-Q4-EM	MSE MADS-120
Suresh	Naburi	suresh.naburi@iu-study.org	42201812	Ber-MSE-MAMAN-60-2022-SS-Q2-EM	MSE MAMAN-60
Marydaphine	Ochoi	marydaphine.ochoi@iu-study.org	102210437	Ber-MSE-MADS-120-2022-WS-Q4-EM	MSE MADS-120
Agonis	Osmani	agonis.osmani@iu-study.org	102210265	Ber-MSE-MADS-120-2022-WS-Q4-MM	MSE MADS-120
Vinay	Oursang	vinay.oursang@iu-study.org	32111770	BH-MSE-MAMAN-60-2021-WS-Q1	MSE MAMAN-60

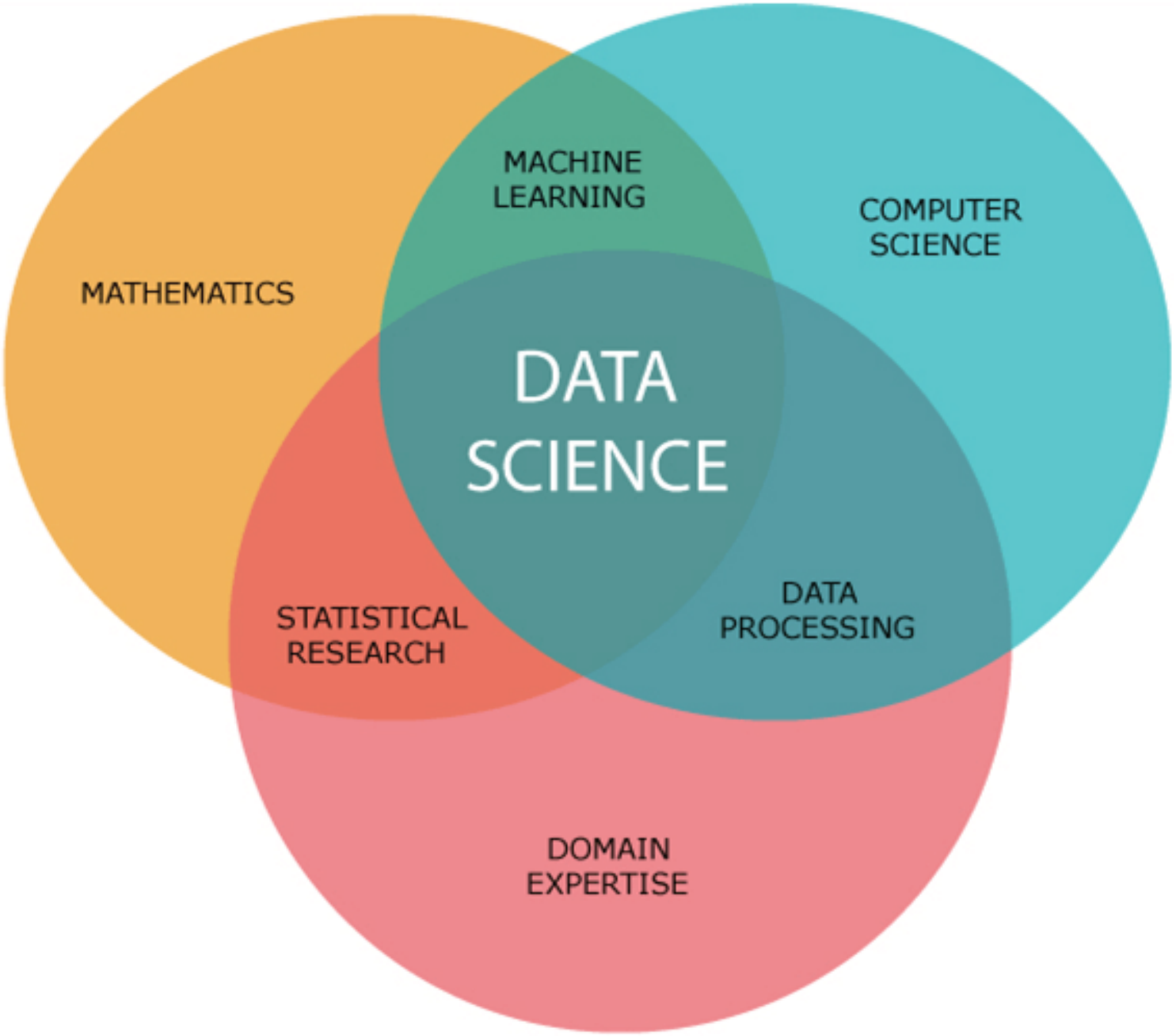
- Course book: Data Science – DLMBDSA01, provided by IU, myStudies
- Reading list DLMBDSA01, provided by IU, myStudies
- Additional teaching materials: https://github.com/duongtrung/IU-DataScienceCourse

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

UNIT 1

INTRODUCTION TO DATA SCIENCE

On completion of this unit, you will have learned…

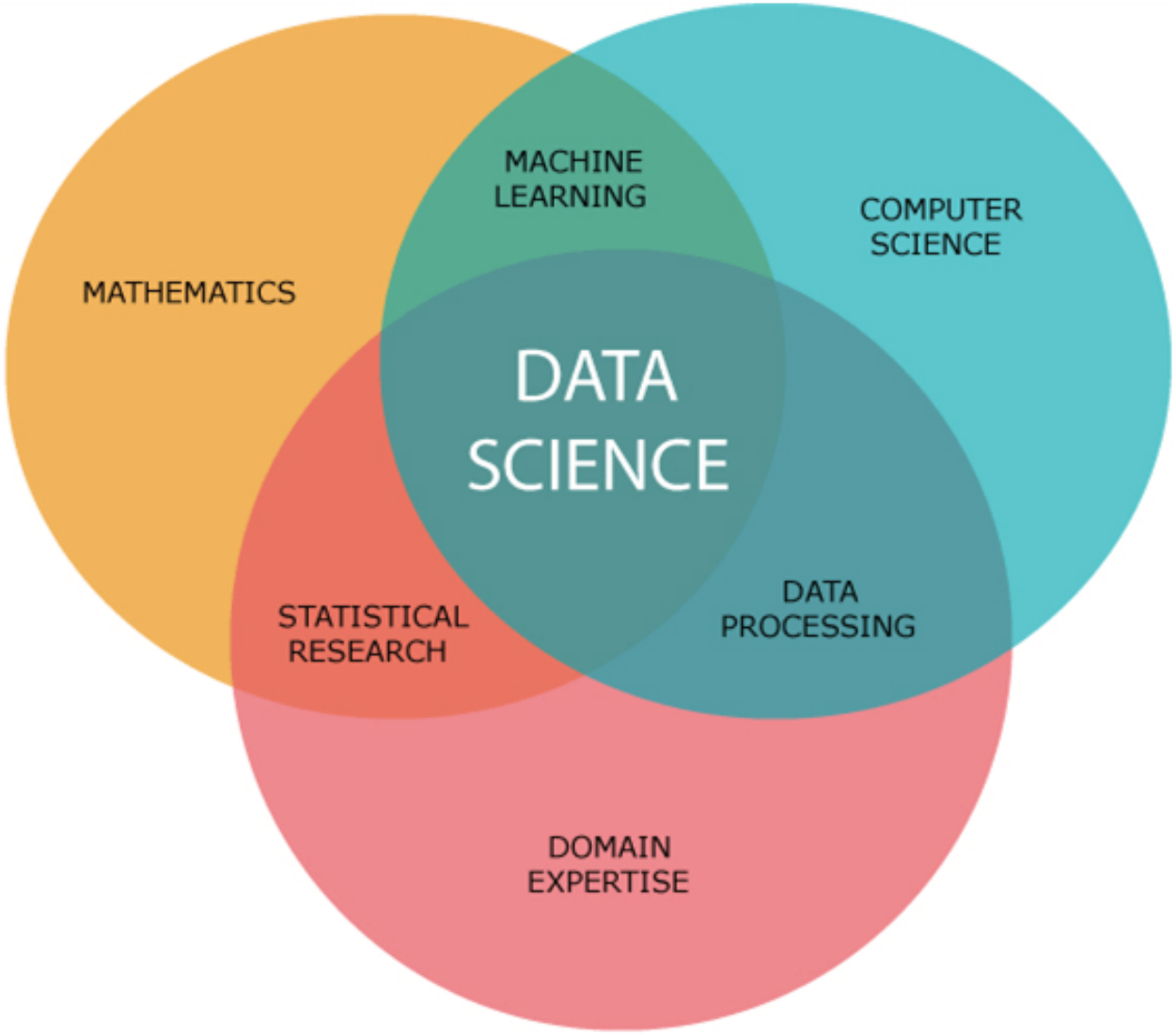
- the meaning of data science.
- common terms and definitions in data science.
- the different applications of data science.
- the typical sources of data.
- the types and shapes of data.
- probability distributions and Bayesian statistics.

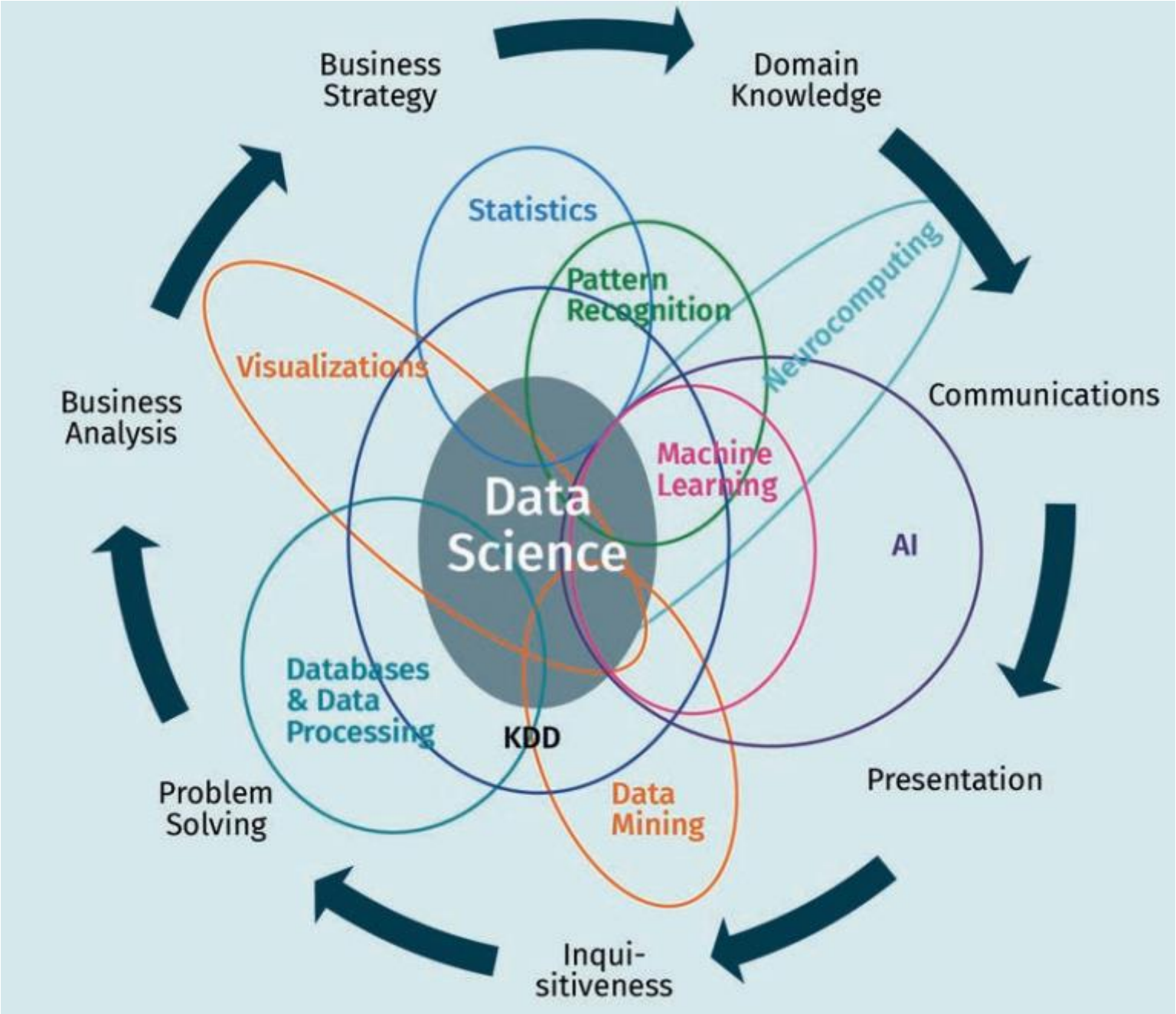
EXPLAIN SIMPLY

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




Business Strategy
Domain Knowledge
Communications
Presentation
Inqui-sitiveness
Problem Solving
Business Analysis
Statistics
Pattern Recognition
Neurocomputing
Visualizations
Machine Learning
Data Science
AI
Databases & Data Processing
KDD
Data Mining

DATA SCIENCE ACTIVITIES

- 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 cleaning
- Data presentation
- Data evaluation

- Treat outliers and missing values
- Inspect visual patterns

- Descriptive statistics & statistical analysis
- Modeling
- Visual techniques

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

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

Nominal
Un-ranked series

Ordinal
Ordered series

Binary
Variables with two options

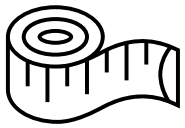
QUANTITATIVE
measurable data

Discrete
Variables can only take certain values

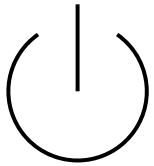
Continuous
Series of values within a finite or infinte range

Countries

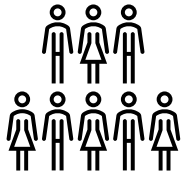
Size (small/ medium/ large)



Power (ON/OFF)



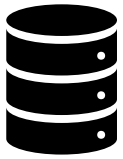
Number of students



Temperature

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

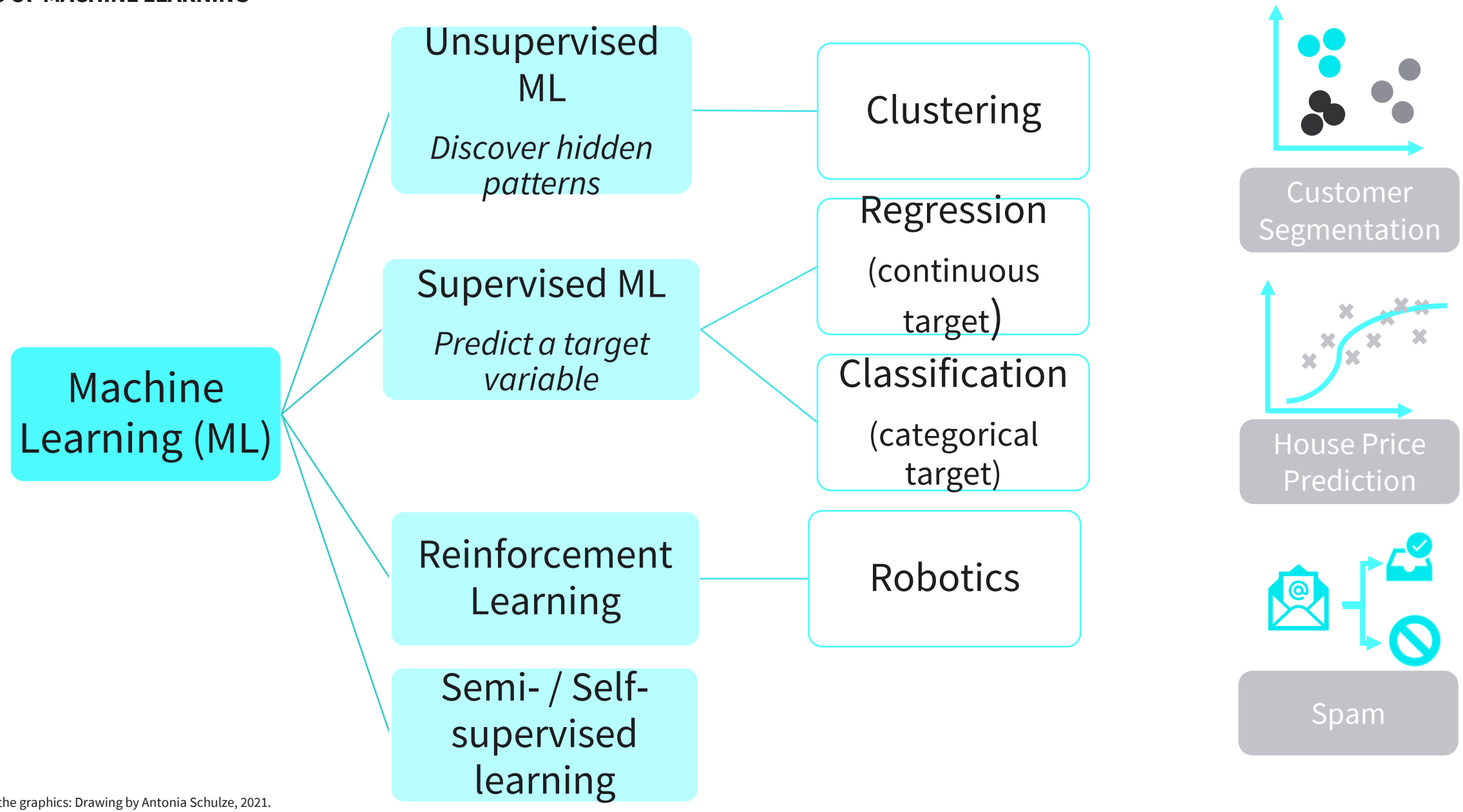
Hi Marie,
How are you?

Unstructured Data

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

The book is fabulous! I enjoyed it!

TYPES OF MACHINE LEARNING

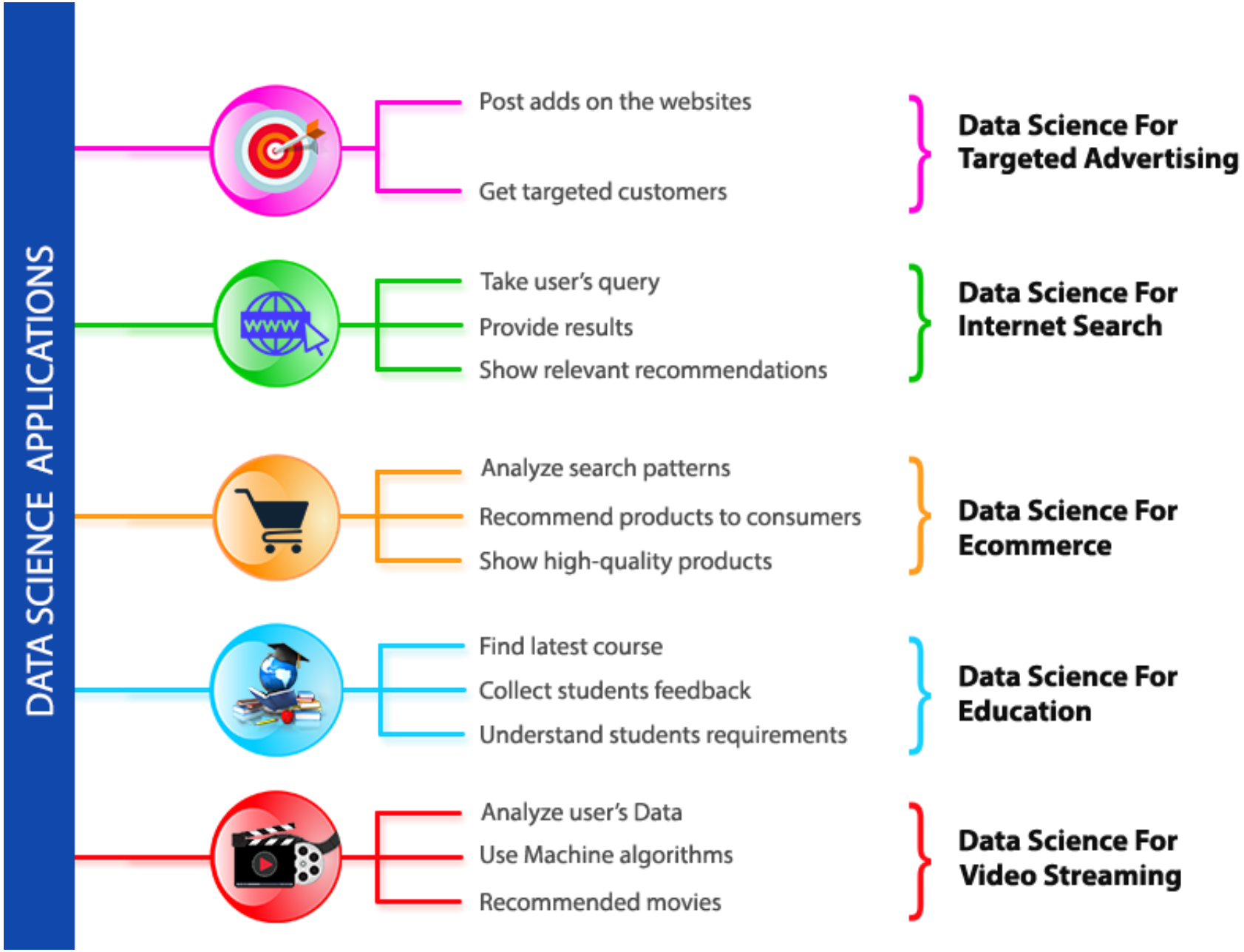


Source of the graphics: Drawing by Antonia Schulze, 2021.

- 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

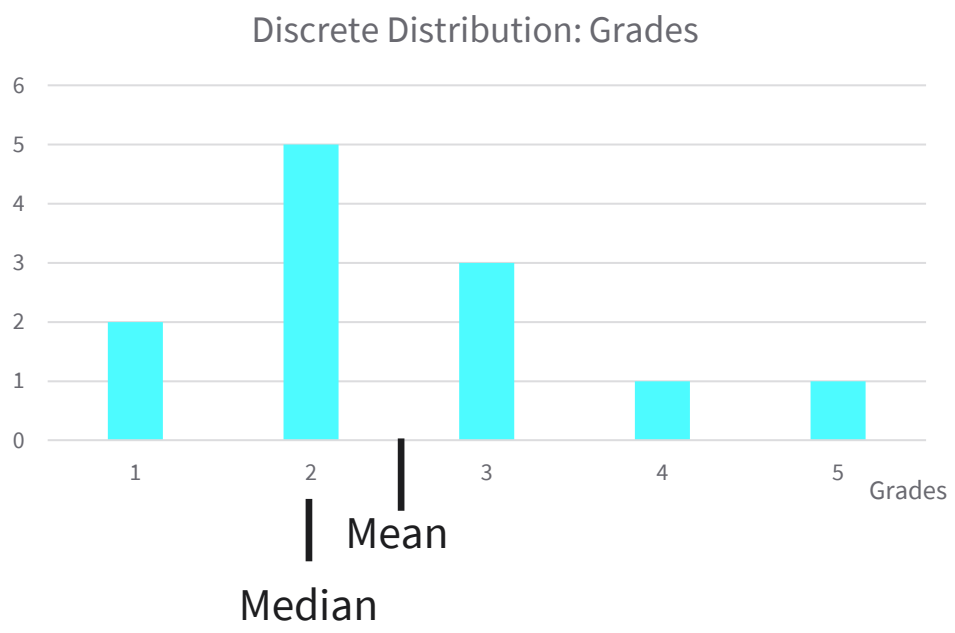


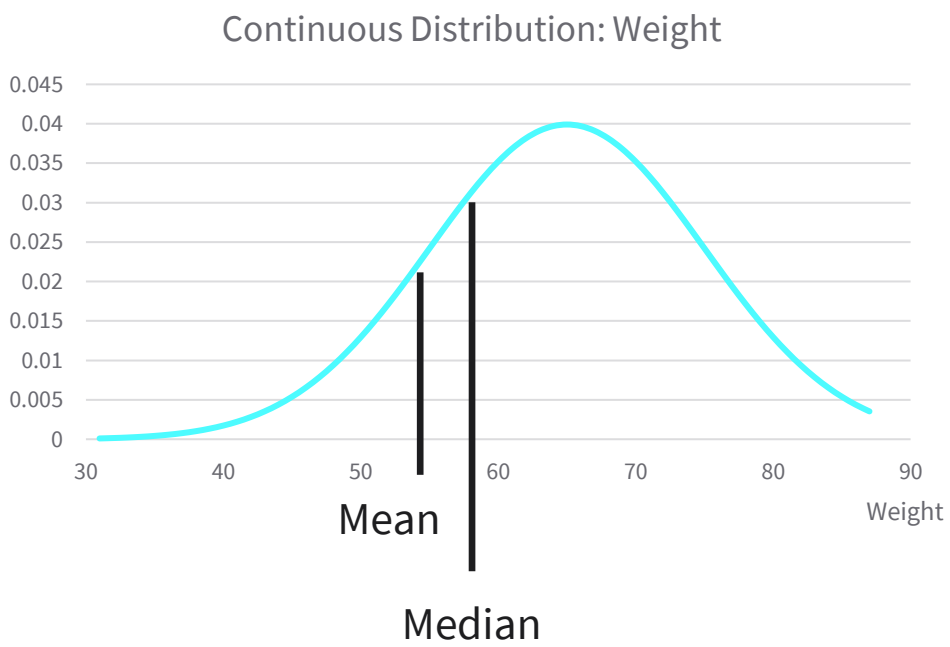
Value, Probability

Standard deviation = measure of spread

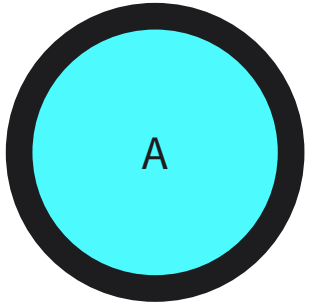
Mean = average

Median = 50% greater, 50% smaller values



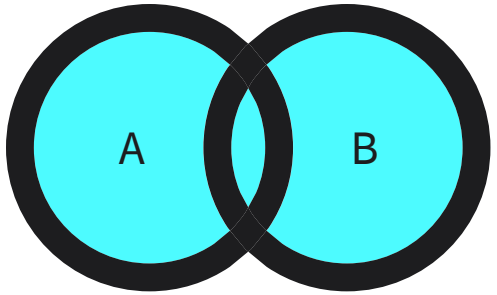


Source of the graphics: Drawing by Antonia Schulze, 2021.



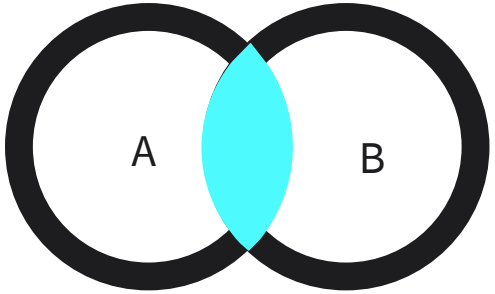
P (A)

Probability of an event A happening



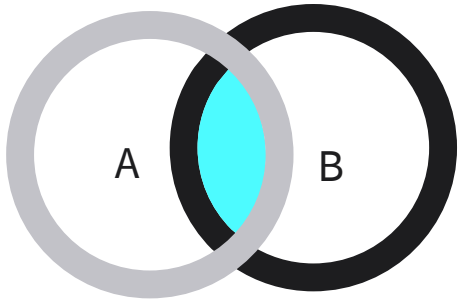
P (A ∪ B)

Probability of event **A or B** happening



P (A ∩ B)

Probability of event **A and B** happening

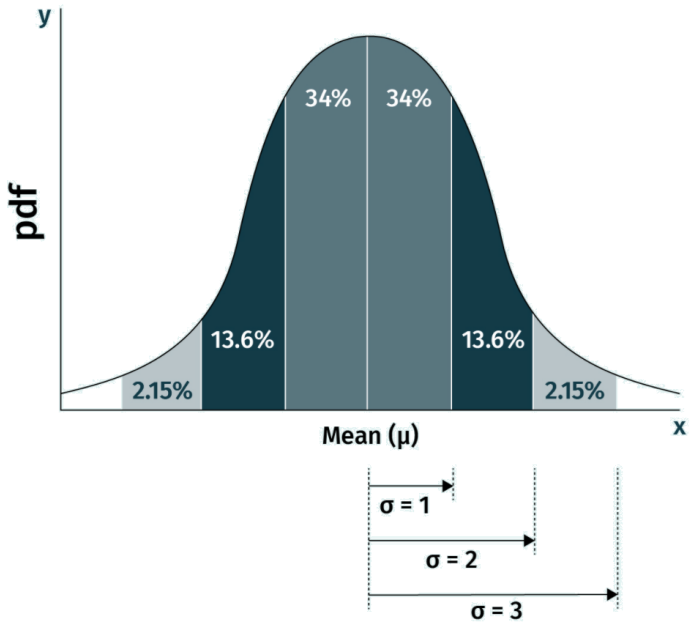


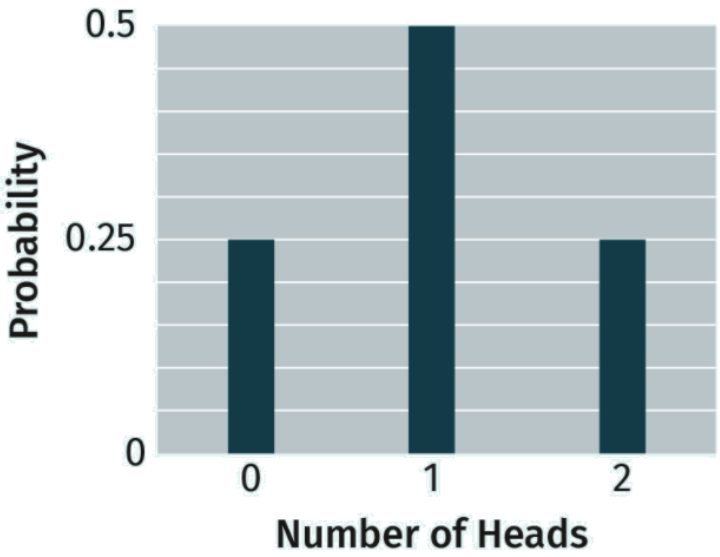
P (A | B)

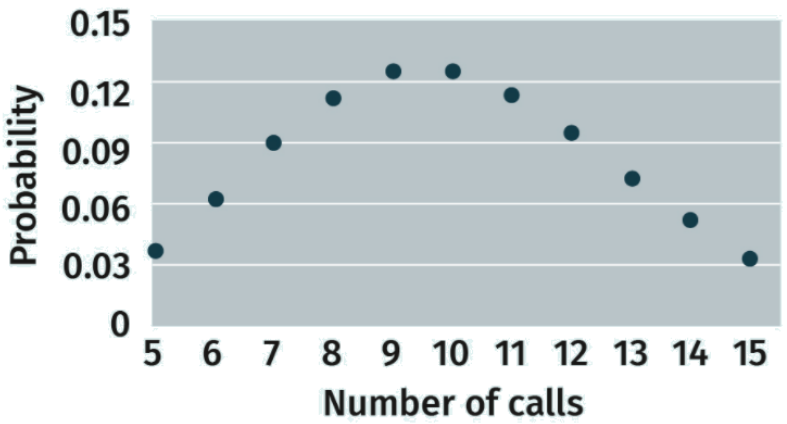
Probability of A, given that event B already happened

Conditional Probability

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







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.

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

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.

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

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

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.