# Data Science: An Introduction

Arun Kumar

IIT Ropar

July 23, 2025

# Outline of the Talk

1. Why Data Science?
2. Data Preparation
3. Algorithm Selection
4. Parameter Estimation
5. Evaluating Results

# Objectives of the Course

- Develop good understanding of the key concepts.
- Understand the mathematical formulations of the several key techniques in Data Science.
- Gain hands-on experience with Python to apply the concepts to data.

# Why Data Science ?

- Learning through personal experience and knowledge gained from different sources, which propagates from generation to generation, is at the heart of human learning and intelligence.
- In other words, we learn from experience/data.
- Different kind of data and different focuses on the data give rise to different scientific disciplines.

# Contd...

- Data can make a difference in an election.
- A business model success or failure depends on proper data analysis.
- Data can be very useful for social economic reforms.
- Data can provide better options to cure a disease.
- Data can be helpful in preventing crimes.
- In summary, data can help us in taking more informed decisions.

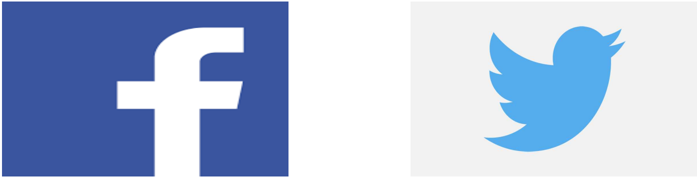
# Is data just a bunch of numbers ?

Gone are the days when data used to be just a bunch of numbers and categorical variables. People were happy to use MS-Excel to process the data.

### New Kind of Data

- Traditional: numerical, categorical, or binary
- Text: emails, tweets, news articles
- Records: user-level data, timestamped event data
- Geo-based location data: Housing data
- Network
- Sensor data: manufacturing companies to monitor data coming from machine sensors to know possible faults
- Images
- Audio & video data

# Unprecedented Data Growth in Recent Times

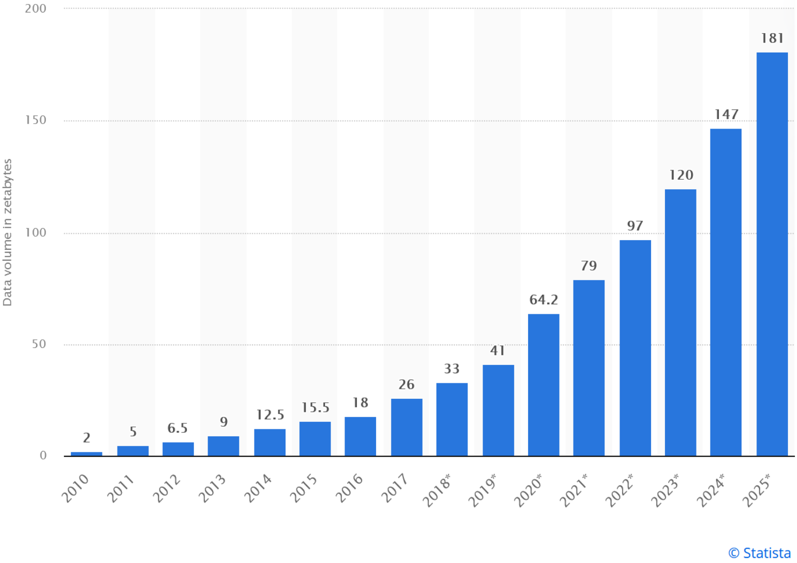


**(a)** Facebook

**(b)** Twitter

**Figure:** Social networking platforms are among the major source of data creation

- Snapchat users share half million photos per day.
- Several hundred professional join LinkedIn per day.
- Roughly 4 million videos are watched on YouTube per day.
- Around 50,000 photos are posted on Instagram per day.
- There are 2.5 quintillion bytes (2.5 Million TB) of data created each day
- Over the last 5 years alone 90% of the data in the world was generated.



1

[1] https://www.statista.com/statistics/871513/worldwide-data-created/

| Unit | Abbreviation | Approximate size |
|---|---|---|
| bit | b | binary digit, a single 0 or 1 |
| byte | B | 8 bits |
| kilobyte | KB | 1,024 bytes or 10^3 bytes |
| megabyte | MB | 1,024 KB or 10^6 bytes |
| gigabyte | GB | 1,024 MB or 10^9 bytes |
| terabyte | TB | 1,024 GB or 10^12 bytes |
| petabyte | PB | 1,024 TB or 10^15 bytes |
| exabyte | EB | 1,024 PB or 10^18 bytes |
| zettabyte | ZB | 1,024 EB or 10^21 bytes |
| yottabyte | YB | 1,024 ZB or 10^24 bytes |

# Why Data Science? - Job Prospects

- The Bureau of Labor Statistics US estimates that there will be a 36% job growth for data scientists rate over the next decade. The average of all job growth is just 5%.
- “Data Scientist: The Sexiest Job of the 21st Century"- ’Harvard Business Review’[2].
- Out of every 100 people with data science skills, only two are unemployed according to US News’s report.
- According to NITI Aayog report Data Science and AI has the potential to add USD 15.7 trillion to the global GDP by 2030 and roughly USD 1 trillion could be added to India GDP by 2035.

[2] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

# Different data!! Different scientific disciplines

- Statistics and Statistical Learning
- Pattern Recognition
- Signal and Image Processing and Analysis
- Data Mining
- Machine Vision
- Bioinformatics
- Industrial Automation
- Computer-Aided Medical Diagnosis
- Others

# Data Science Techniques

In spite of the different data and different disciplines, there is a common corpus of techniques that are used in all of them, and we will refer to such methods as

**“Data Science Techniques"**

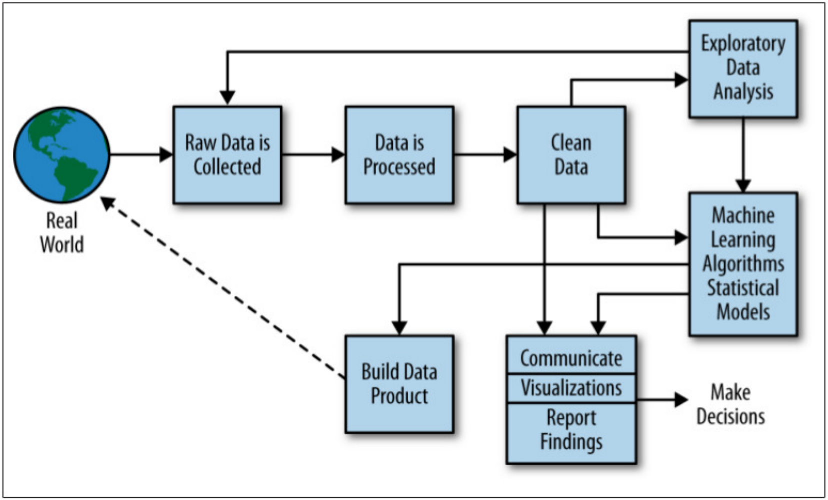
# Typical Profile

- Computer science
- Mathematics
- Statistics
- Machine learning
- Domain expertise
- Data visualization
- Communication and presentation skills

# Required Math and Stat

| Mathematics & Statistics | |
|---|---|
| **Mathematics** | **Statistics** |
| Linear Algebra | Descriptive Statistics |
| Calculus | Exploratory Data Analysis (EDA) |
| Optimisation | Parameter Estimation and Hypothesis Testing |
| Graph Theory | Probability Distributions Particularly Multivariate Gaussian and Multinomial |
| Probability/Combinatorics | Bayesian Analysis |

# Process



3

[3] Cathy O'Neil and Rachel Schutt (2014). Doing Data Science, Straight Talk From The Frontline. O'Reilly.

# Key Steps in Data Science

There are four main steps in a statistical learning:

1. Data processing and preparation for analysis.
2. Shortlisting of Algorithms/Models based on study's requirements.
3. Models' parameters estimation.
4. Evaluate models based on their accuracy.

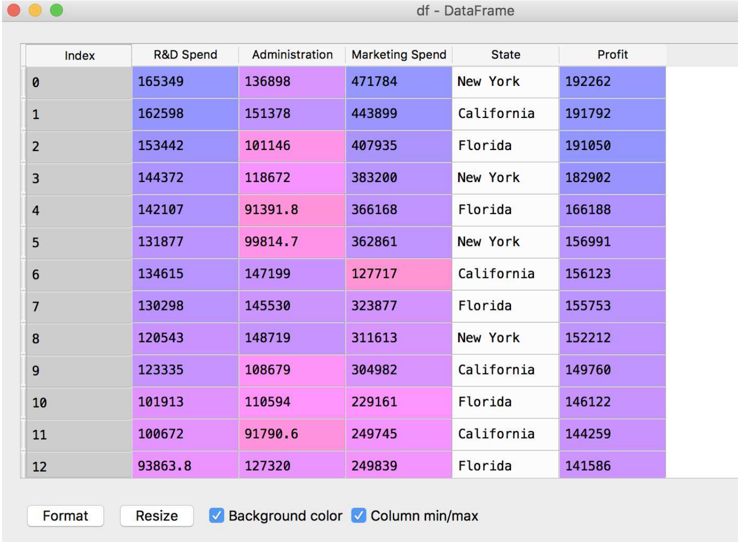
# Data Preparation

Statistical learning is all about the data. If the input data is of poor quality, the output will also be poor.

**Data format:** The tabular form is most commonly used to represent data for analysis. In tabular form each row represents a single observation and each column shows a variable describing the data point.

**Variables:** are also known as attributes, features or dimensions.

# Data Format



| Index | R&D Spend | Administration | Marketing Spend | State | Profit |
|---|---|---|---|---|---|
| 0 | 165349 | 136898 | 471784 | New York | 192262 |
| 1 | 162598 | 151378 | 443899 | California | 191792 |
| 2 | 153442 | 101146 | 407935 | Florida | 191050 |
| 3 | 144372 | 118672 | 383200 | New York | 182902 |
| 4 | 142107 | 91391.8 | 366168 | Florida | 166188 |
| 5 | 131877 | 99814.7 | 362861 | New York | 156991 |
| 6 | 134615 | 147199 | 127717 | California | 156123 |
| 7 | 130298 | 145530 | 323877 | Florida | 155753 |
| 8 | 120543 | 148719 | 311613 | New York | 152212 |
| 9 | 123335 | 108679 | 304982 | California | 149760 |
| 10 | 101913 | 110594 | 229161 | Florida | 146122 |
| 11 | 100672 | 91790.6 | 249745 | California | 144259 |
| 12 | 93863.8 | 127320 | 249839 | Florida | 141586 |

# Variable Types

There are four main types of variables, and it is important to distinguish between them.

**Binary:** This is the simplest type of variable, with only two possible options. For example the options could be Yes and No; Male and Female; Good and Bad etc.

**Categorical:** when there are more than two options, it is called a categorical variable.

**Integer:** These variables are used when the information can be represented as a whole number.

**Continuous:** This variable, represent numbers with decimal places.

# Variable Selection

- Variable selection is an important step in statistical leanring.
- While we might have a large number of variables in our original dataset, using too many variables in an algorithm might lead to slow computation, or wrong predictions due to excess noise.
- Variable selection can be done by using correlation and even by simple plots.
- Subject knowledge is very helpful in variable selection. For example, if somebody is interested in the prediction of currency price movement, it is important to understand the macro- and micro-economic factors, which drive the currency prices.

# Feature Engineering

Sometimes, however, the best variables need to be engineered. In feature engineering, we

- create new variables, which can provide more information;
- could also combine multiple variables in a technique called dimension reduction.
- Dimension reduction can be used to extract the most useful information and condense that into a new but smaller set of variables for analysis.

# Missing Data

It is not always possible to collect complete data. The data could be missed due to different reasons. Missing data can interfere with analysis, and can be handled in following ways:

**Approximated:** If the missing value is binary/categorical, it could be replaced by the mode or the more frequently occurring category. If data is integer/continuous the median can be used.

**Computed:** Missing values could also be computed using some models. For example in finance a missing value between two data points is calculated by using Brownian bridge.

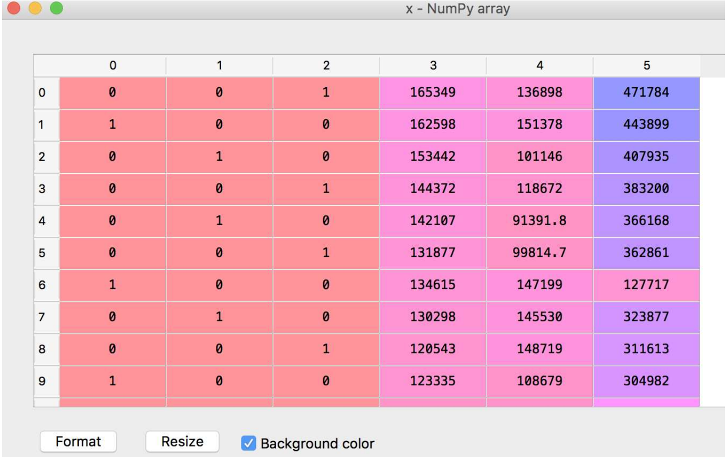
**Removed:** As a last option, rows with missing values could be removed from the dataset.

# Outlier Handling

Sometime, some values in data set may not be consistent with other data values. These data values are called outliers which could be too large or too small in comparison of other values. Outliers can greatly affect the analysis. Outliers can be handled by

- Cross checking the data with other sources.
- Understanding the event which could have lead to the outlier.
- For analysis part a model which is less sensitive to outliers may be chosen.

# Taking Care of Categorical Data



|   | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 165349 | 136898 | 471784 |
| 1 | 1 | 0 | 0 | 162598 | 151378 | 443899 |
| 2 | 0 | 1 | 0 | 153442 | 101146 | 407935 |
| 3 | 0 | 0 | 1 | 144372 | 118672 | 383200 |
| 4 | 0 | 1 | 0 | 142107 | 91391.8 | 366168 |
| 5 | 0 | 0 | 1 | 131877 | 99814.7 | 362861 |
| 6 | 1 | 0 | 0 | 134615 | 147199 | 127717 |
| 7 | 0 | 1 | 0 | 130298 | 145530 | 323877 |
| 8 | 0 | 0 | 1 | 120543 | 148719 | 311613 |
| 9 | 1 | 0 | 0 | 123335 | 108679 | 304982 |

☑ Background color

# Algorithm Selection

There could be various algorithms for modeling/prediction purpose. Some are listed here.

| **Category** | **Algorithm** |
|---|---|
| Supervised Learning | Regression Analysis<br>*k*- Nearest Neighbors<br>Naive Bayes<br>Decision Tree<br>Support Vector Machine<br>Random Forests<br>Neural Networks |
| Unsupervised Learning | *k*- Means Clustering<br>PCA and SVD<br>Social Network Analysis<br>Association Rules |

# Parameter Tuning

- Different algorithms have different parameters available for tuning.
- A single algorithm can generate varying results, depending on how its parameters are tuned/calibrated.
- A model’s accuracy suffers when the model is not suitable calibrated.

# Evaluating Results

After building a model, it must be evaluated. Evaluation metrics are used to compare how accurate the models are in their predictions. The following evaluations metrics are commonly used.

**Classification Metrics:**

- Percentage of correct predictions
- Confusion matrix

**Regression Metric:**

- Root mean squared error (RMSE)

# Validation and Cross-Validation

Metrics do not give a complete picture of a model's performance. Models giving very good result on current data might not do so for new data.

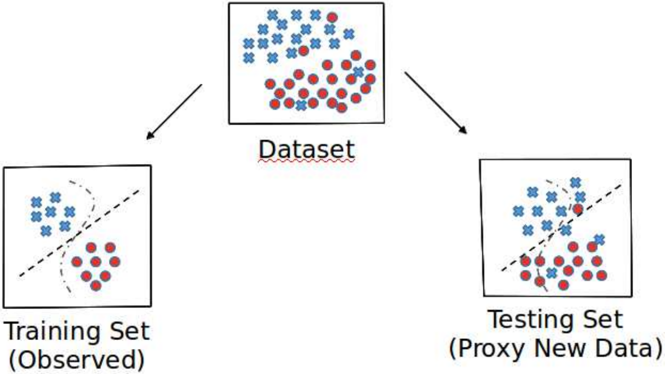
**Validation**

is an assessment of how accurate a model is in predicting new data.Generally, we split our current dataset into two parts: the first part would serve as a training dataset to generate and tune a prediction model, while the second part would be used as a proxy for new data and called testing dataset to assess the model's accuracy.

**Cross-Validation**

optimize the availability of data by dividing the dataset into several segments that are used to test the model repeatedly.

# Training and Testing Split



# Main Idea

### Example (Section 2.1 ISLR)

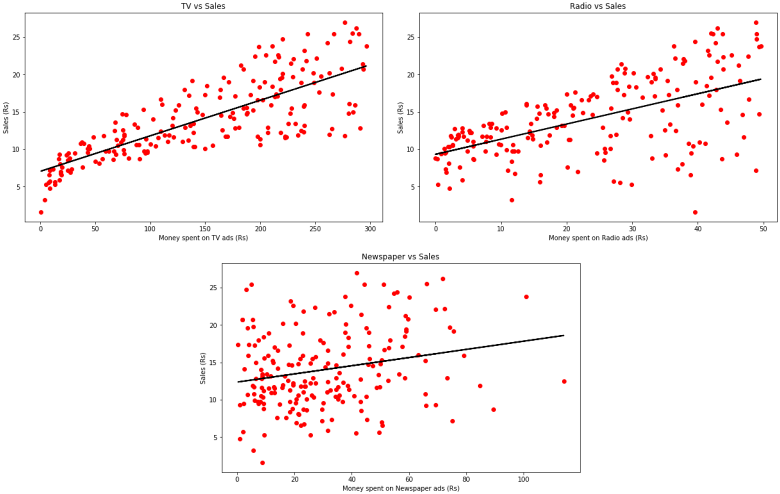
Suppose that a company hires a consultant to provide advice on how to improve sales of a particular product. The **Advertising data set** consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. In this case

- Sales is called output, response or dependent variable and is generally denoted by $Y$.

- TV, radio and newspaper budget are called input, predictors, features or independent variables and often denoted by variable $X$.

- We can denote TV, radio and newspaper budgets by $X_1, X_2$ and $X_3$ respectively.

# Advertisement versus sales



# The relationship between advertisement and sales

Essentially we want to find the function $f$ such that

$$Y = f(X_1, X_2, X_3) + \epsilon,$$

where $\epsilon$ is the error term independent of $X_1, X_2$ and $X_3$.

- $X_1$ is advertisement cost on TV
- $X_2$ is advertisement cost on Radio
- $X_3$ is advertisement cost on Newspaper
- $Y$ represents Sales

# The Main Task is to Estimate $f$

In general, suppose that we observe $p$ different predictors, $X_1, X_2, \ldots, X_p$ and a quantitative response $Y$. Suppose there is some relationship between $Y$ and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

$$Y = f(\mathbf{X}) + \epsilon = f(X_1, X_2, \cdots, X_p) + \epsilon.$$

In essence, data science includes a set of approaches for estimating $f$.

# How to estimate $f$?

- To estimate $f$, we need training data, which can be represented in matrix form as follows

$$A = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

- We have $p$ features and $n$ data points.
- Further, let $y_i$ represent the response variable for the $i$th observation.
- To estimate the unknown function $f$, we apply different learning method (Statistical & Machine) to the training data.

# Characterization of different models

Broadly the statistical learning methods can be characterized as

- Parametric: In this approach, we assume a **functional form** of the model and then estimate the parameters of the model. The model could be for example linear, non-linear, polynomial etc. The disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of $f$. If the chosen model is too far from the true $f$, then our estimate will be poor.

- Non-parametric: These methods **do not make** explicit assumptions about the **functional form** of $f$. Instead they seek an estimate of $f$ that gets as close to the data points as possible without being too rough. These methods generally needs large number of training data points and more computational resources.

# Characterization of different models contd...

| Category | Algorithm |
|---|---|
| Parametric | Simple Linear Regression |
| | Multiple Linear Regression |
| | Polynomial Regression |
| | LDA, QDA |
| | Naive Bayes |
| | Neural Networks |
| | LSTM |
| Non Parametric | *k*- Nearest Neighbors |
| | Decision Tree |
| | Support Vector Machine |
| | Random Forests |

# Prediction accuracy and interpretability

**Prediction Accuracy:** A model performance is assessed in terms of its accuracy to predict the occurrence of an event on unseen data. A more accurate model is considered as a more valuable model.

**Model Interpretability:** A Model interpretability assess the relationship between the inputs and the output variables. A more interpretable model can tell why the input is able to explain the output.

Generally a more complex (less interpretable) model has more accuracy.

# The trade-off between prediction accuracy and model interpretability

- Some models for estimating $f$ are less flexible, in the sense that they can produce just a relatively small range of shapes to estimate $f$.
- For example, the linear regression model is less flexible but it is more interpretable. We can easily talk about the change in dependent variable per unit changes in independent variables.
- The non-linear regression is however, more flexible and can generate different shapes but is less interpretable.
- When inference is the goal, there are clear advantages to using simple and relatively inflexible statistical learning methods.
- However, if prediction is the main goal and interpretability is not important, we can use more flexible models for the purpose.

# Supervised vs Unsupervised Learning

## Supervised Learning:

Both input and output data are provided. Supervised learning uses the patterns in the data to make predictions.

### Example

Many classical statistical learning methods such as linear regression, logistic regression and generalized linear models falls in this category.

## Unsupervised Learning:

Only input data is provided. Unsupervised learning tells the existing patterns in the data. These algorithms are unsupervised because we do not know what patterns to look out for and thus leave them to be uncovered by the algorithm.

### Example

Clustering analysis which ascertain on the basis of features $x_1, x_2, \cdots, x_n$, whether the observations fall into relatively distinct groups is an unsupervised learning technique.

# Supervised Learning: Regression vs Classification

### Regression

In regression problem, we try to predict a numerical value by using features variable.

### Example

Based on advertisement budget for TV, Radio and Newspaper we try to predict the sales. Based on education level and years of experience, we predict income.

### Classification

In classification problem, we categorize the data points in groups called the classes or categories.

### Example

Based on salary, age and gender we predict if a person is potential buyer of a product or not. Based on grades, GRE score and university ranking, we predict if a student will get admission or not.

# Basic Regression Template

- Importing the Libraries
- Importing Dataset
- Splitting into Testing and Training Set
- Feature Scaling
- Fitting the Regression Model to the Data Set
- Calculating RMSE
- Predicting
- Visualizing the Results

# Basic Classification Template

- Importing the Libraries
- Importing Dataset
- Splitting into Testing and Training Set
- Feature Scaling
- Fitting the Classification Model to the Data Set
- Predicting
- Making Confusion Matrix
- Visualizing the Results

# Measuring the quality of fit

To assess the quality of fit in regression setting, a commonly used measure is Mean Squared Error (MSE) defined by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

where $y_i$ is the actual value and $\hat{f}(x_i)$ is the predicted value.

- The MSE will be small if the predicted responses are very close to the true responses.
- The MSE calculated on **training data** is referred to as the **training MSE**.
- The MSE calculated on **test data** is referred to as the **test MSE**.
- The estimation part in the model minimize the training MSE and hence we prefer the models which has lower test MSE.

# Measuring the quality of fit in classification setting

In classification setting the most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the training error rate, the proportion of mistakes that are made if we apply error rate our estimate $\hat{f}$ to the training observations:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$$

- Here $y_i$ is the actual class and $\hat{y}_i$ is the predicted class.
- $I$ is the indicator function.
- Essentially, we calculate the average of wrong classifications.

# References

- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer New York.

- Cathy O'Neil and Rachel Schutt (2014). Doing Data Science, Straight Talk From The Frontline. O'Reilly.

- Mitchell, T. M. (2017). Machine Learning. McGraw Hill Education.

- https://www.superdatascience.com/