Week 1: Introduction

Cecilia Graiff

ALMAnaCH, Inria Paris; School of Public Affairs, SciencesPo

September 4th, 2025

Section 1

Organisation

Course Objectives

- Learn basic Python syntax
- Learn basic concepts of data analysis and apply them to real-world data
- Gain understanding of the underlying statistical concepts
- Have an insight into Machine Learning

Course Content

- **Week 1**: Introduction to the course, Python installation, environment setting, basic Git tutorial, Jupyter Notebook tutorial, Python coding basics (data structures, loops, functions, basic operations).
- **Week 2**: Data collection and cleaning.
- **Week 3**: Data manipulation (`numpy`, `pandas`) and data visualization (`matplotlib`, `plotly`).
- **Week 4**: Exploratory Data Analysis (EDA)
- **Week 5**: Basic statistic and linear models.
- **Week 6**: Introduction to Machine Learning and to simple models with `sklearn`.

Course Evaluation

- Pass or fail basis
- Groups of 3 people
- Project ideas due by October 15th
- Project due by December 18th
- Elaborate a research question and the methodology to solve it based on real-world data

Course Materials

Possible datasets

-

- Using a different dataset is possible, but you need to justify it to me.

14:45 - 16:45

AI use regulation

Section 2

Introduction to Data Analysis

What is Data?

Data

Collection of **values** *that convey information*,
that help us **analyze**, **interpret**, and **make decisions**.

Types of data

Types of Data

- **Structured:** Organized, easy to search (e.g., tables, databases)
- **Unstructured:** Raw, complex (e.g., images, text, videos)
- **Semi-structured:** Mixed form (e.g., JSON, XML files)

See also: IBM - Structured vs Unstructured Data

Example: Regression

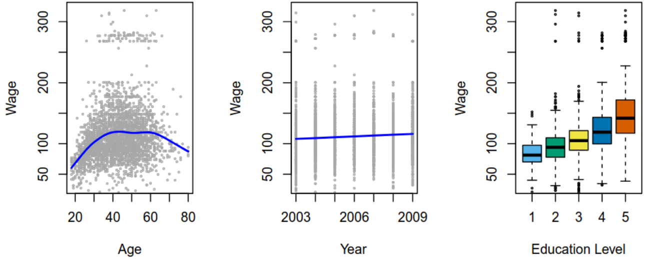
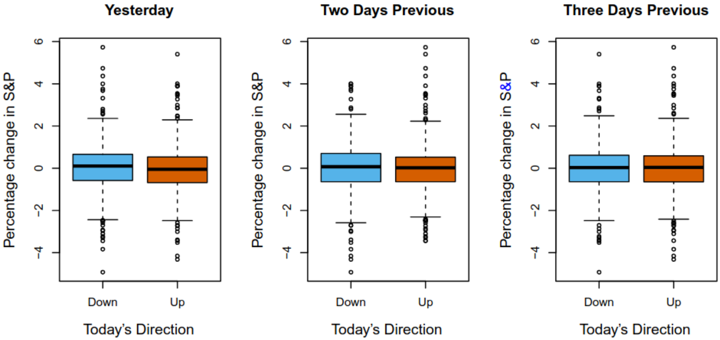


FIGURE 1.1. Wage *data, which contains income survey information for men from the central Atlantic region of the United States.* Left: wage *as a function of* age. *On average,* wage *increases with* age *until about* 60 *years of age, at which point it begins to decline.* Center: wage *as a function of* year. *There is a slow but steady increase of approximately* $10,000 *in the average* wage *between* 2003 *and* 2009. Right: *Boxplots displaying* wage *as a function of* education, *with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* wage *increases with the level of education.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Example: Classification



Yesterday

Two Days Previous

Three Days Previous

FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* `Smarket` *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Example: Clustering

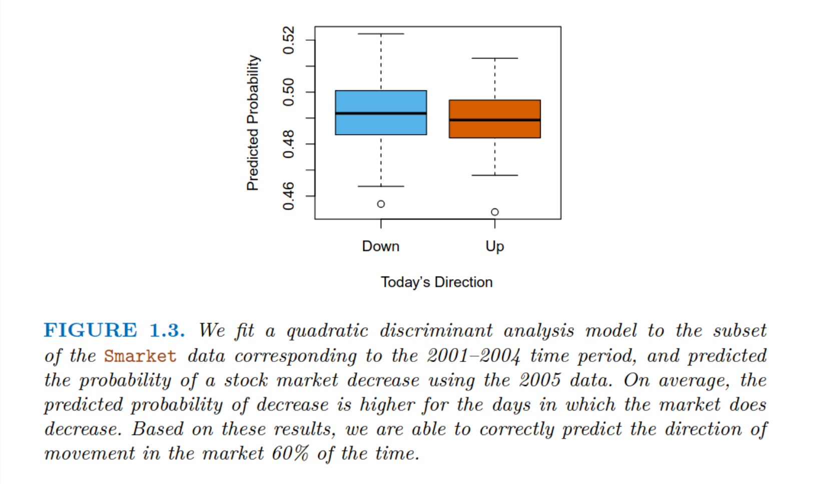


FIGURE 1.3. *We fit a quadratic discriminant analysis model to the subset of the* `Smarket` *data corresponding to the 2001–2004 time period, and predicted the probability of a stock market decrease using the 2005 data. On average, the predicted probability of decrease is higher for the days in which the market does decrease. Based on these results, we are able to correctly predict the direction of movement in the market 60% of the time.*

Source: *An Introduction to Statistical Learning with Applications in Python.*

Variable types

- **Quantitative data**: In the first example, we predict a **numerical value**.

Variable types

- **Quantitative data**: In the first example, we predict a **numerical value**.
- **Qualitative data**: In the second example, we predict a **label**. Qualitative variables are also referred to as **categorical**.

Disclaimer: For the purpose of this class, we will only deal with these two types of variables.

Learning Paradigms

Supervised Learning

Data: Labeled dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ are input features, and $y_i \in \mathcal{Y}$ are known target values or classes.

Goal: Learn a function $f : \mathbb{R}^d \to \mathcal{Y}$ to predict y from $\mathbf{x}$.

Examples:
$\mathcal{Y} = \mathbb{R}$ (regression) or
$\mathcal{Y} = \{1, \ldots, K\}$ (classification).

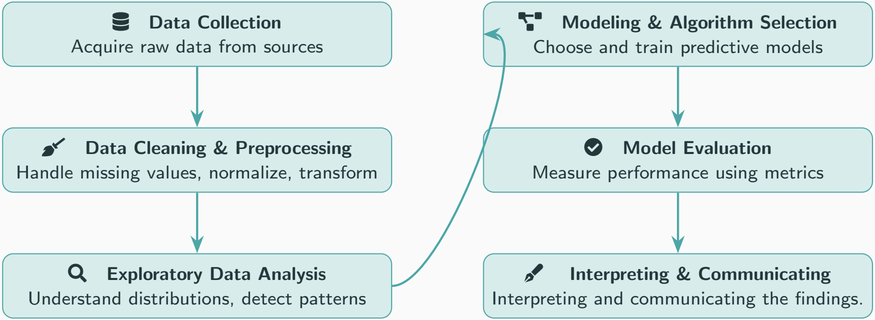
Unsupervised Learning

Data: Unlabeled dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$,
where $\mathbf{x}_i \in \mathbb{R}^d$ are input features only,
and no corresponding target values.

Goal: Discover underlying structure, distribution, or representation of $\mathbf{x}$.

Examples: Clustering, density estimation, dimensionality reduction.

Data Analysis Pipeline



Data Collection
Acquire raw data from sources

Data Cleaning & Preprocessing
Handle missing values, normalize, transform

Exploratory Data Analysis
Understand distributions, detect patterns

Modeling & Algorithm Selection
Choose and train predictive models

Model Evaluation
Measure performance using metrics

Interpreting & Communicating
Interpreting and communicating the findings.

Section 3

Lab session: Python and Git Basics

Git: Reproducibility and Version Control

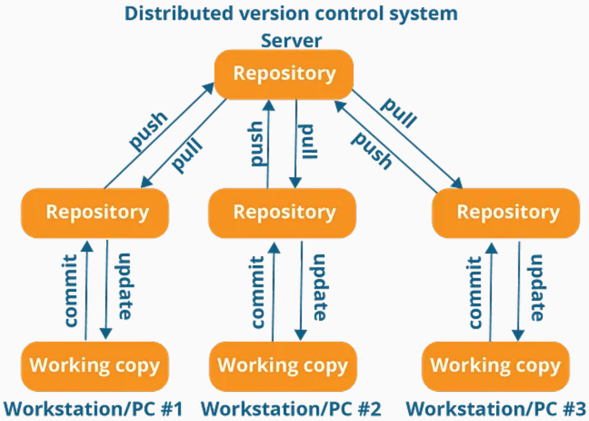


Figure 1: A simplified diagram of version control in Git.

Source: *Faun*

Git: Reproducibility and Version Control

- **Version Control**
  Keep a complete history of your project and track changes easily.
- **Collaboration**
  Multiple developers can work together without stepping on each other's toes.
- **Branching and Merging**
  Experiment safely with new features and merge only when ready.
- **Backup and Recovery**
  Revert to previous versions anytime to prevent data loss.

Git Setup & Installation

1. Download Git

Get the latest version at `https://git-scm.com/`

2. Cross-Platform Support

Available for Windows, macOS, and Linux.

3. Configure Your Identity

Set your username and email by running these commands in your terminal:

```
git config --global user.name "Your Name"
git config --global user.email "you@example.com"
```