Week 1: Introduction

Cecilia Graiff

ALMAnaCH, Inria Paris; School of Public Affairs, SciencesPo

September 11, 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

Homework

- Assigned each week, uploaded on GitHub
- Purpose: practice Python concepts introduced in class
- **Not mandatory** (but recommended!)
- Corrected collectively at the beginning of following lecture

Course Evaluation

- Comprehensive **group project** due at the end of the course
  - Groups of **3 people**
- **Graded on pass/fail basis**

Course Evaluation

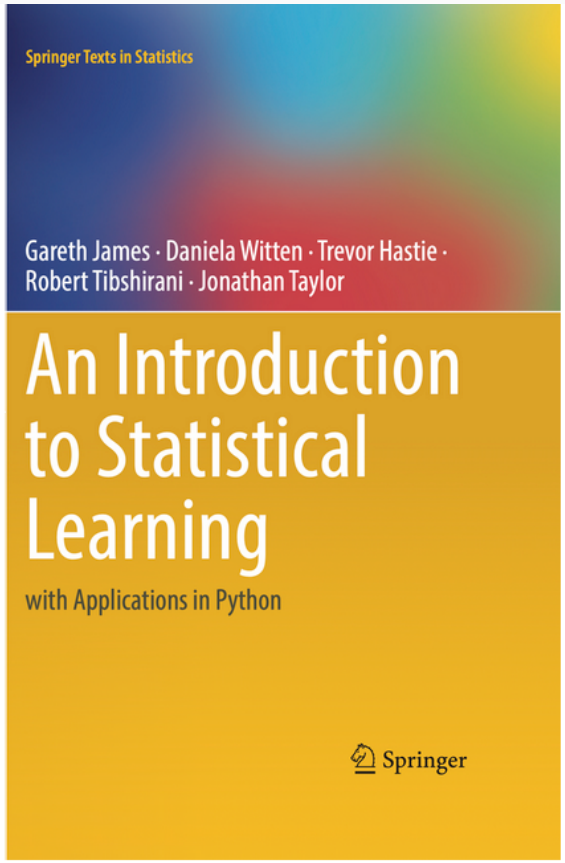
Submission Requirements

1. **Project Idea (Short Description)**
   - Submit by **October 30**
   - Write a short description of:
     - Team members
     - Chosen dataset
     - Research question
     - Planned pipeline
   - One person per group uploads the document on Moodle
2. **Final Project Report**
   - Submit by **December 20**
   - To submit (and to be evaluated):
     - **GitHub** repository with code
     - Project report in form of **scientific paper**

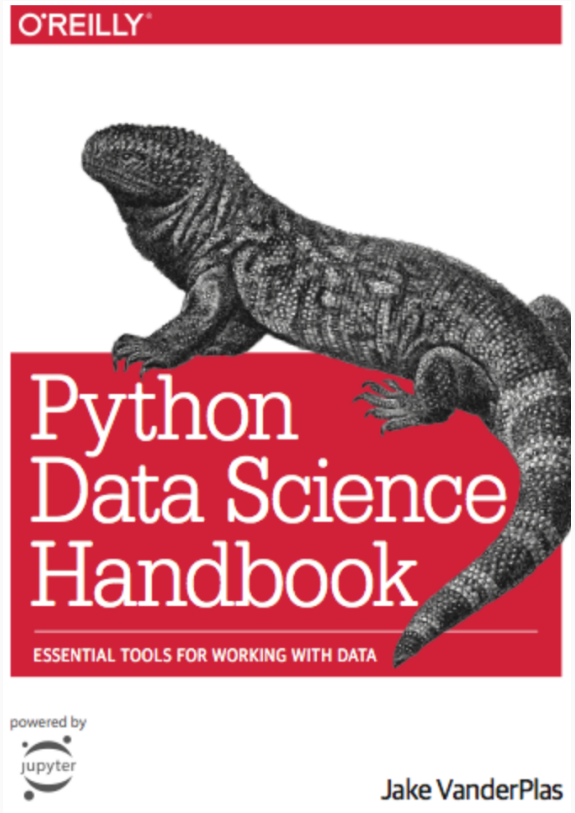
Course Materials

- PDF guide to Python (available on Moodle)
- Notebooks with code (available on GitHub)

Course Materials



An Introduction to Statistical Learning



Python Data Science Handbook

Possible datasets

- I will propose possible datasets during the course, depending on your needs and level
- Using a different dataset is possible, but you need to justify it to me.

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

- **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

Data Analysis: Beginner vs Advanced

- **From descriptive to predictive:**
  - Beginner: "What happened?"
  - Advanced: "What will happen?" and "What if we change something?"
  - E.g., **observing** vs **predicting** GDP growth, election outcomes, or patient survival.
- **Coding and math foundations:**
  - Compute basic data manipulation and summary statistics vs implement models from scratch
- **Handling more and richer data:**
  - Smaller and easier datasets vs complex data types: text (political speeches), spatial data (disease spread), networks.
- **Model evaluation and generalization:**
  - When you use **predictive models**, you need to evaluate their quality

Adapted from IBM Research.

Example: Regression

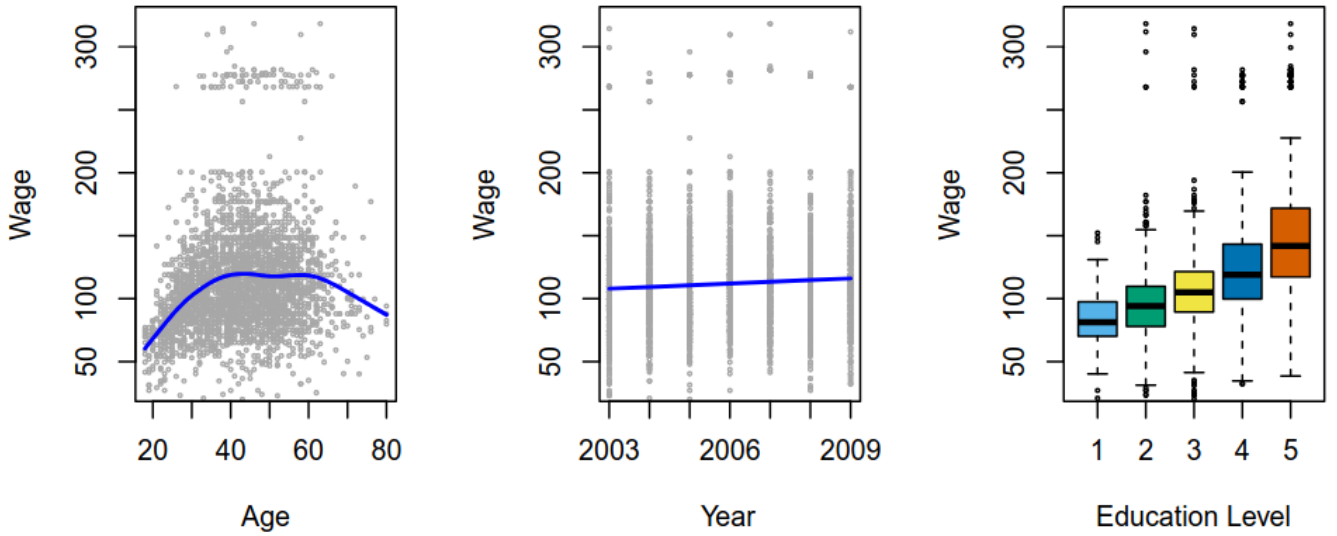


FIGURE 1.1. *Wage data, which contains income survey information for men from the central Atlantic region of the United States.* Left: *wage as a function of age. On average, wage increases with age until about* 60 *years of age, at which point it begins to decline.* Center: *wage as a function of year. 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

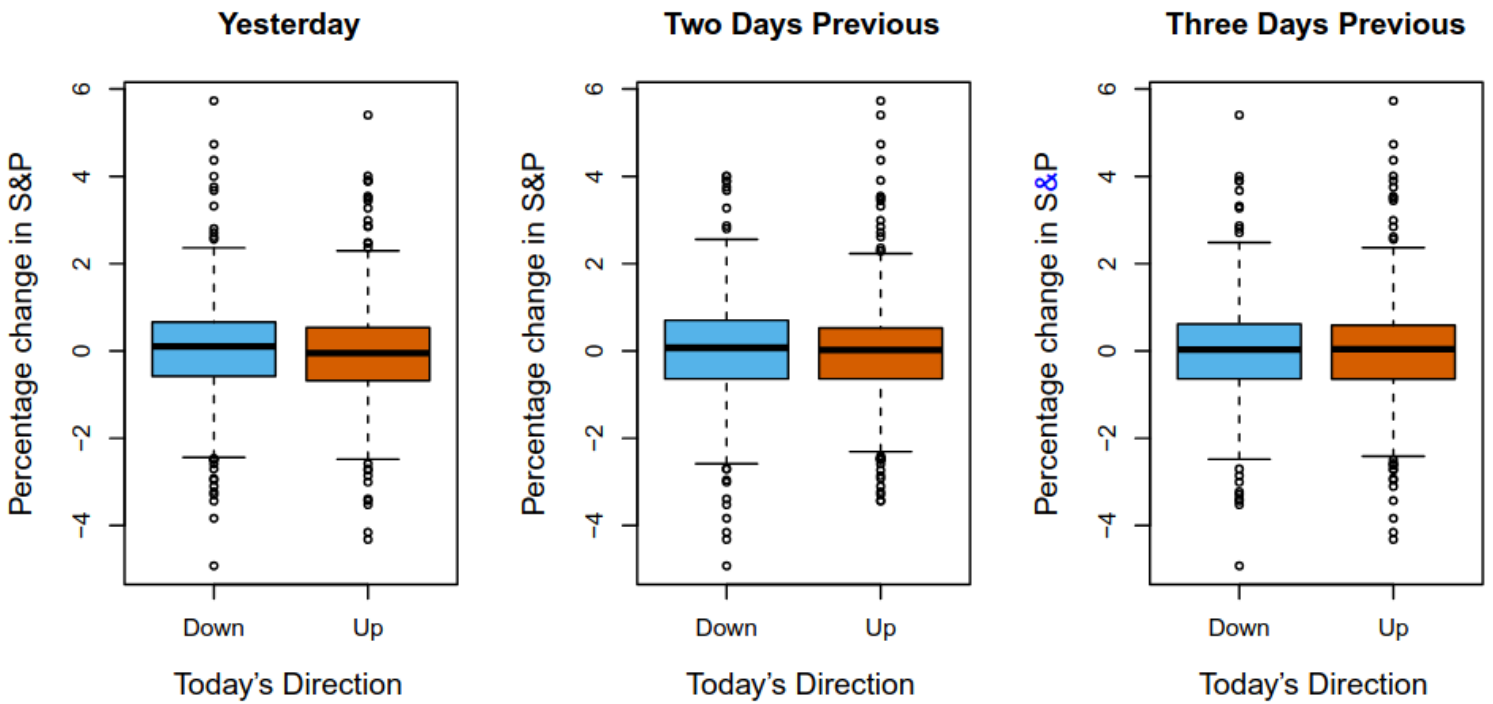


FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* `Smarket` *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

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

Example: Clustering

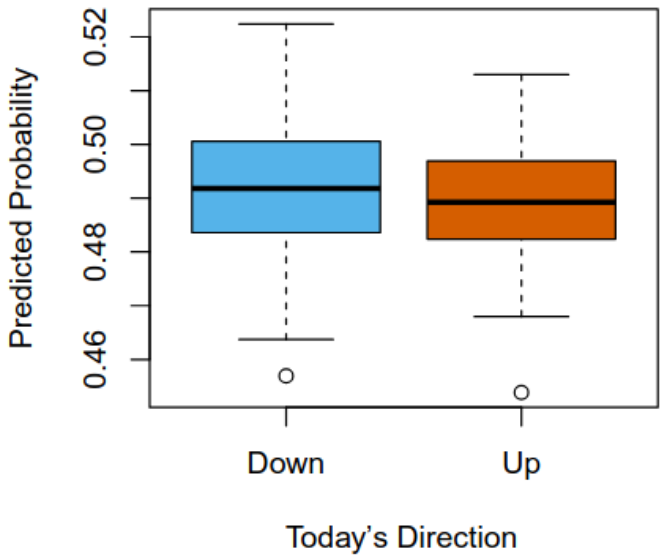


FIGURE 1.3. *We fit a quadratic discriminant analysis model to the subset of the* `Smarket` *data corresponding to the 2001–2004 time period, and predicted the probability of a stock market decrease using the 2005 data. On average, the predicted probability of decrease is higher for the days in which 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

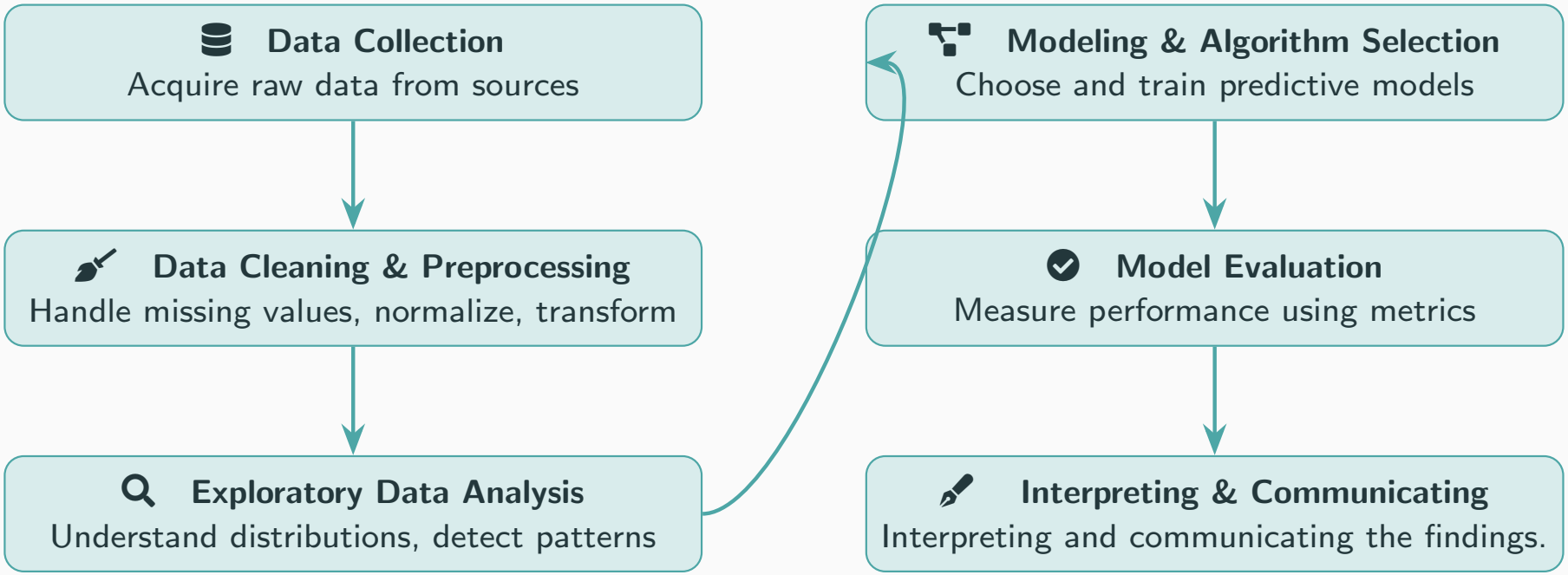
- **Quantitative data**: In the first example, we predict a **numerical value**.

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.

Data Analysis Pipeline



Section 3

Lab session: Python and Git Basics

Python: What and Why?

- Programming language of choice for **scientific computing** and **data science and machine learning**
- Readable syntax, suitable for beginners

Virtual Environments

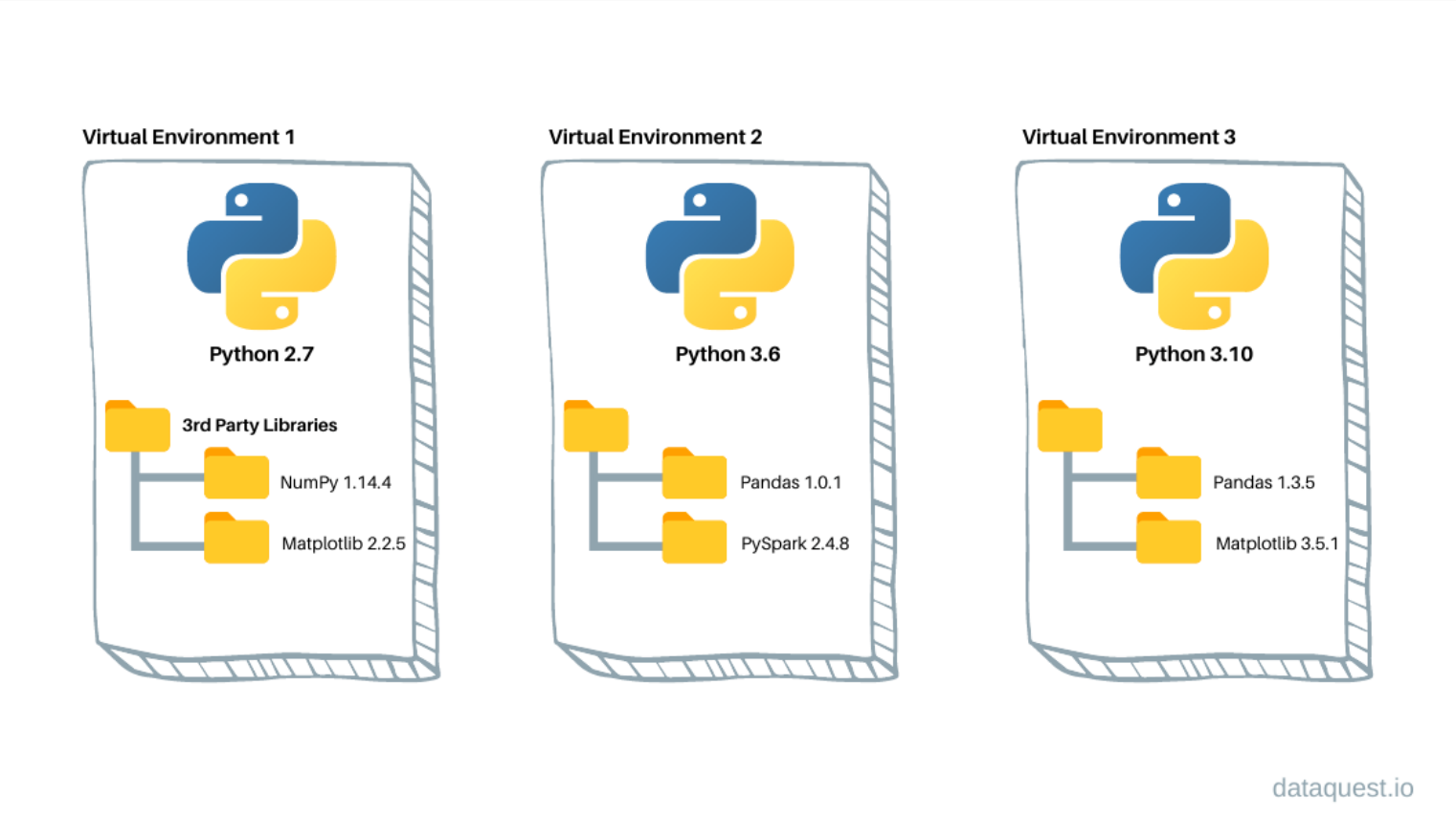


Figure 1: Virtual Environments in Python

Source: Dataquest

Git: Reproducibility and Version Control

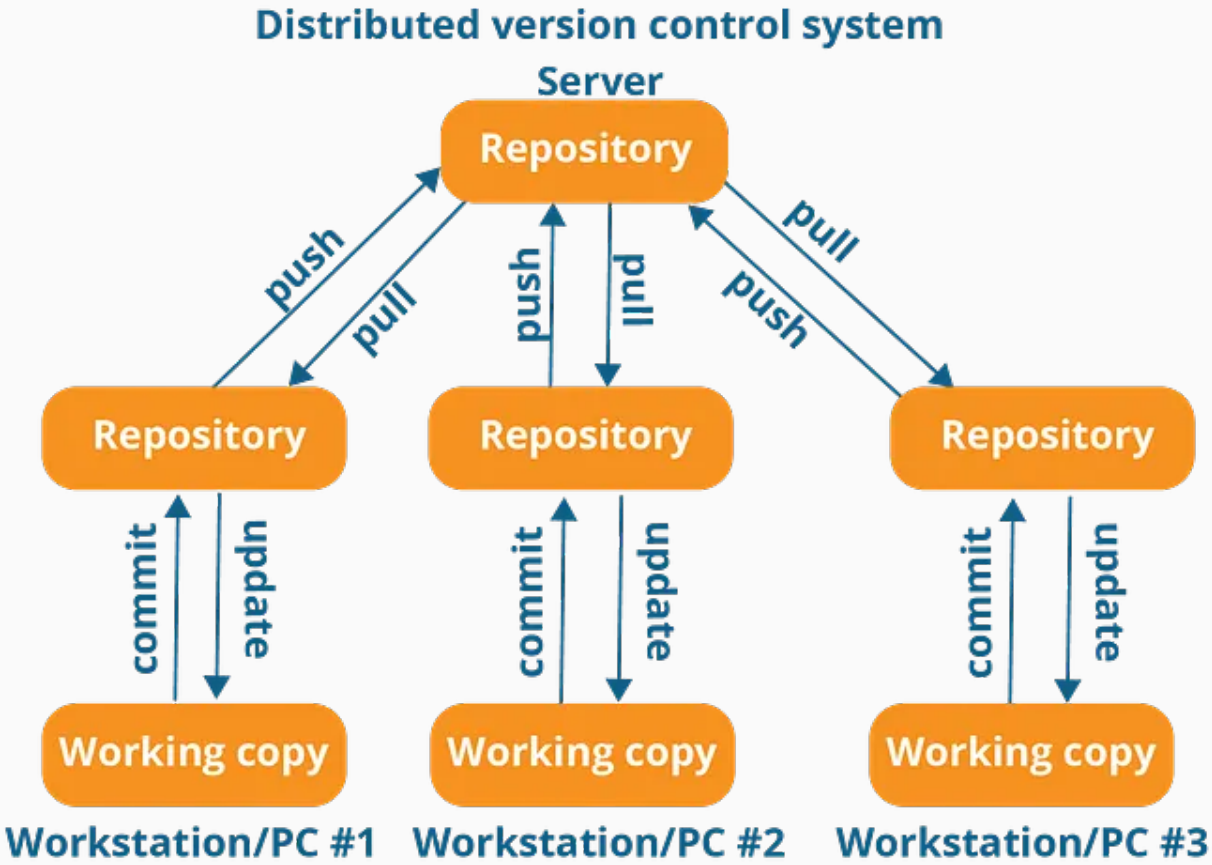


Figure 2: A simplified diagram of version control in Git.

Source: *Faun*