

Applied Social Data Science

Coding Camp

Setup • Orientation

Elena Karagianni, PhD Candidate

karagiae@tcd.ie

 September 8–12, 2025

Department of Political Science, Trinity College Dublin

Introductions

Introductions

- Who *you* are
 - Background and current course
 - Coding familiarity
- Expectations with the program/coding camp week

Schedule

10am - 1pm

- **Monday: Introduction**
- Tuesday: R Basics + Good practices

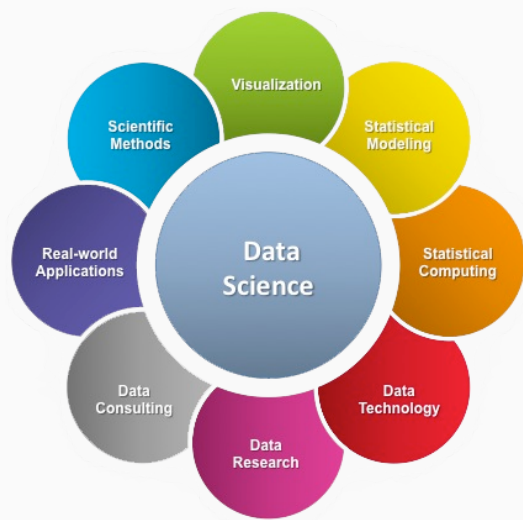
10am - 12pm

- Thursday: Python basics
- Friday: How to write up and report - \LaTeX

Today's class

- What is data science?
- Quantitative Programming Environments: R and Python
- Expectations

What is data science?



What is data science?

- “The science of learning from data” (Donoho, 2017).

What is data science?

- “The science of learning from data” (Donoho, 2017).
- A bridge between statistics and computer science

What is data science?

- “The science of learning from data” (Donoho, 2017).
- A bridge between statistics and computer science
- Aspects: data mining, data analysis, knowledge discovery




What is data science?

- “The science of learning from data” (Donoho, 2017).
- A bridge between statistics and computer science
- Aspects: data mining, data analysis, knowledge discovery
- **What it involves:**
 - Principles, processes, and methods
 - Automated or semi-automated analysis
 - Understanding real-world phenomena

What is data science?

- “The science of learning from data” (Donoho, 2017).
- A bridge between statistics and computer science
- Aspects: data mining, data analysis, knowledge discovery
- **What it involves:**
 - Principles, processes, and methods
 - Automated or semi-automated analysis
 - Understanding real-world phenomena

Used for:

-  Better decisions
-  Predictive analysis
-  Pattern discovery

A brief history



John Tukey, 1915–2000

*'All in all, I have come to feel that my central interest is in **data analysis**, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.'*

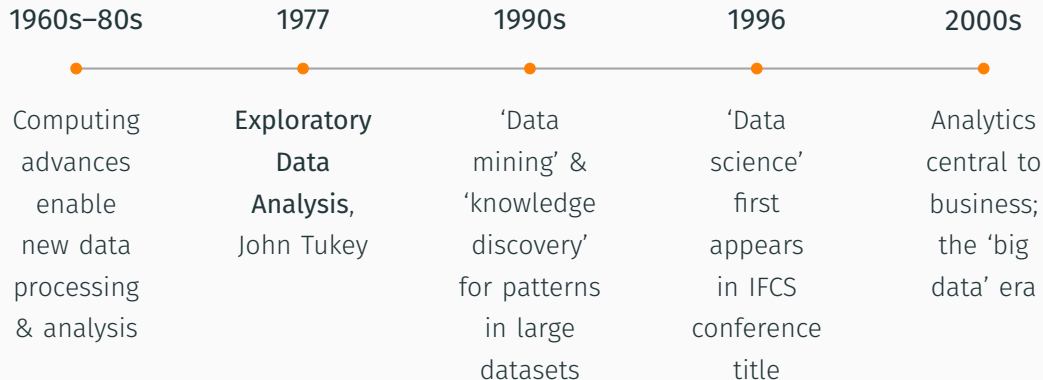
The Future of Data Analysis, 1962.

Four major influences act on data analysis **today**:

1. The formal theories of statistics
2. Accelerating developments in computers and display devices
3. The challenge, in many fields, of more and ever larger bodies of data
4. The emphasis on quantification in an ever wider variety of disciplines

Tukey, 1962!

Timeline



"I keep saying the sexy job in the next ten years will be statisticians [...] The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades."

Hal Varian, Google Chief Economist, Jan. 2009

Things to keep in mind

Key distinction

Statistics = mathematics of inference

Data science = practice of working with data

- Not everyone agrees with this distinction
- In practice, the terms are often used differently



Statistician



Data Scientist

Things to keep in mind

- There is a difference between **scientific** and **engineering** mindsets
 - **Scientific mindset** → seeks to understand the underlying process (generative modeling)
 - **Engineering mindset** → looks to find the best prediction (predictive modeling)
- In the social sciences, we often want to understand what's inside the 'black box', but not all data science methods are designed for this.

Things to keep in mind

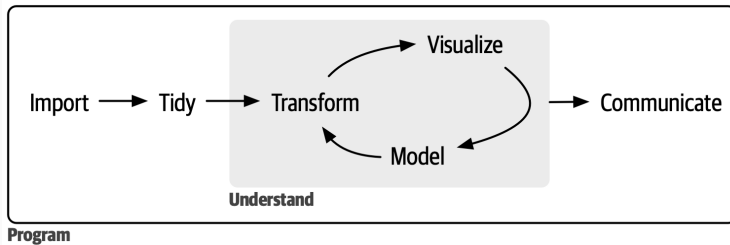
Data Science

- Broad field: principles & methods to identify and understand phenomena
- Uses programming, statistics, **machine learning**, and algorithms
- Works with large datasets to discover patterns and predict outcomes

Data Analytics

- Focused on insights for decision-making
- Analyzes past data to answer specific questions
- Supports present decisions with evidence

Data Science is a Process

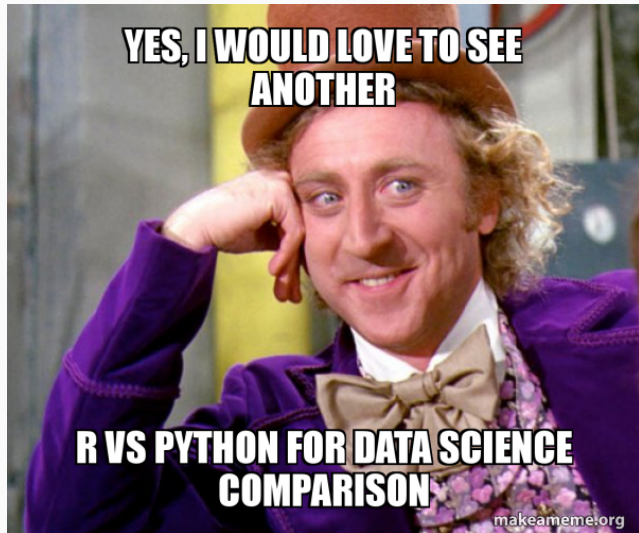


Introduction, R for Data Science, c/o Hadley Wickham

80 % of the data analysis process is spent on the process of cleaning and preparing the data!

The activities of 'Greater Data Science' are classified into six divisions (Donoho,2017):

1. Data Gathering, Preparation, and Exploration
2. Data Representation and Transformation
3. Computing with Data (several languages!)
4. Data Modeling (generative vs. predictive models)
5. Data Visualization and Presentation
6. Science about Data Science



Installing R

R is a programming language used for statistics. It is completely free and can be downloaded from **CRAN**, the comprehensive R archive network.

1. Go to <https://cran.r-project.org/>
2. In the box headed “Download and Install R”, click the link corresponding to your operating system.
3. Follow the instructions for your system.

Installing **R Studio** → integrated development environment

1. Go to <https://posit.co/download/rstudio-desktop/>
2. Scroll down to “2. Install RStudio” and click the button “Download RStudio Desktop for (your operating system)”.
3. Follow the instructions for your system.

Expectations

- There is a lot to learn

Expectations

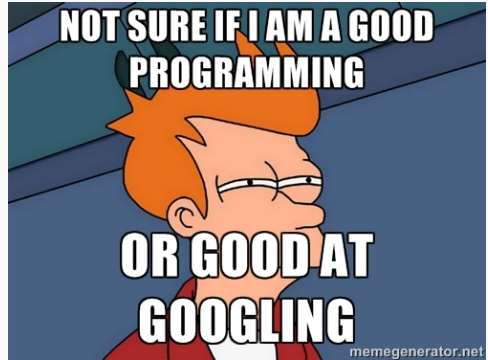
- There is a lot to learn
- There is a steep learning curve

Expectations

- There is a lot to learn
- There is a steep learning curve
- Slow and steady wins the race

Expectations

- There is a lot to learn
- There is a steep learning curve
- Slow and steady wins the race
- We are here to help



Useful Resources

- R for Data Science (2e): <https://r4ds.hadley.nz/>
- Python for Data Analysis (3e): <https://wesmckinney.com/book/>
- GitHub: <https://docs.github.com/en/get-started/quickstart/hello-world>
- Posit Primers: <https://posit.cloud/learn/primers>