Applied Social Data Science - Coding Camp

Marina Schenkel
schenkem@tcd.ie

PhD Candidate

September 4 - 8, 2023

Introductions

- My background and projects

  Public Policy + Computational Methods

Examples: Thesis (super) preliminary results

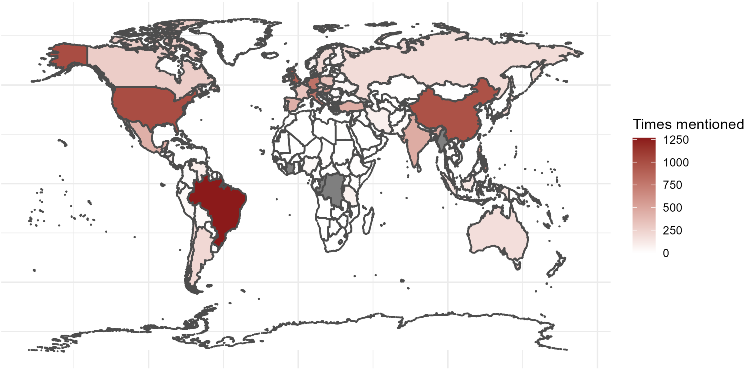


Figure: Number of times each county is mentioned in studies on populism and health/science included in the review.

Examples: Thesis (super) preliminary results

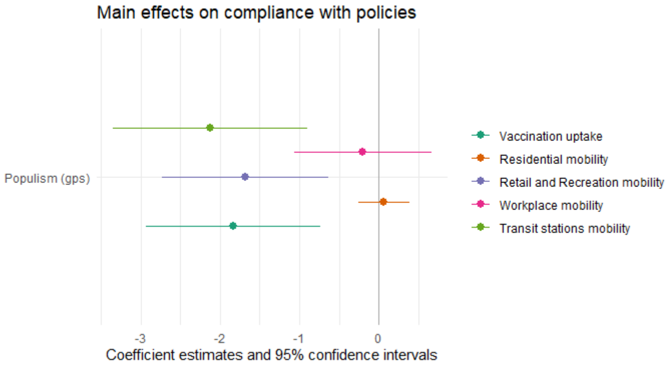


Figure: Estimating the main effect of populist rhetoric of the leader's party (GPS index) on compliance with COVID-19 policies. The figure shows the results from Random Effects models controlled for Presidentialism (binary), Party's ideology right-wing (binary), Liberal Democracy Index, Regional Index, Covid deaths and reproduction rates in the previous month, GDP per capita, Government effectiveness, Relative Political Reach, and Hospital beds/1000.

Examples: Master's projects with Twitter data

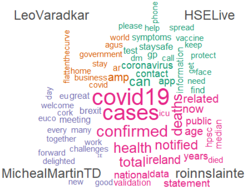


Figure: Irish government

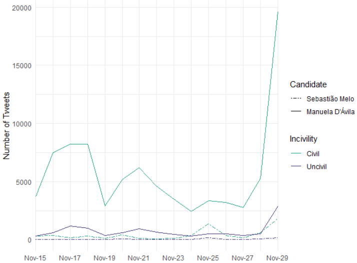


Figure: Brazilian candidates

Introductions

- Who are we

  Background and current course

  Coding familiarity?
- Expectations with the program/ Coding camp week

Schedule

- Monday: Introduction (Maxwell Theatre)
- Tuesday: R basics (Lecture Theatre LB01- O'Reilly Institute)
- Wednesday: Python basics (Maxwell Theatre)
- Thursday: Good practices and Latex (Maxwell Theatre)
- Friday: How to report and share results (Maxwell Theatre)

10am to 12pm. 10min break?

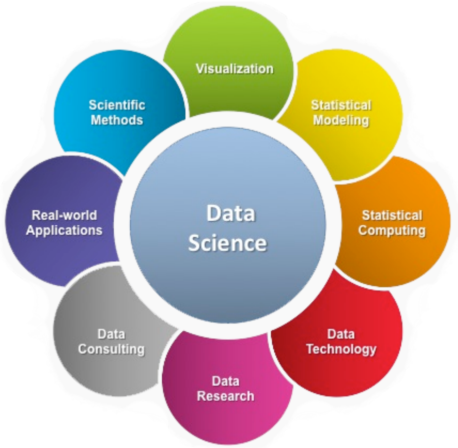
Today's class

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.*
- statistics $+$ computer science

What is data science?

- ‘The science of learning from data’ *Donoho, 2017.*
- statistics + computer science
- data mining, data analysis, knowledge discovery...

What is data science?

- ‘The science of learning from data’ *Donoho, 2017.*
- statistics + computer science
- data mining, data analysis, knowledge discovery...
- Involves principles, processes, and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data.

What is data science?

- ‘The science of learning from data’ *Donoho, 2017.*
- statistics + computer science
- data mining, data analysis, knowledge discovery...
- Involves principles, processes, and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data.
- Used for:

What is data science?

- ‘The science of learning from data’ *Donoho, 2017.*
- statistics + computer science
- data mining, data analysis, knowledge discovery...
- Involves principles, processes, and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data.
- Used for:

– Better decisions

What is data science?

- ‘The science of learning from data’ *Donoho, 2017.*
- statistics + computer science
- data mining, data analysis, knowledge discovery...
- Involves principles, processes, and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data.
- Used for:
  - Better decisions
  - Predictive analysis

What is data science?

- ‘The science of learning from data’ *Donoho, 2017.*
- statistics + computer science
- data mining, data analysis, knowledge discovery...
- Involves principles, processes, and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data.
- Used for:
- – Better decisions
- – Predictive analysis
- – Pattern discoveries, etc

A brief history...

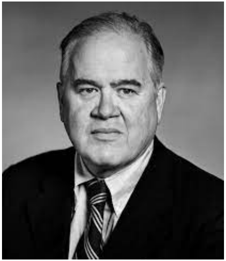


Figure: John Tukey, 1915-2000

'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.

A brief history...

'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

- 1960s - 1980s: advances in computer technology allow for new methods in processing and analysing data

Timeline

- 1960s - 1980s: advances in computer technology allow for new methods in processing and analysing data
- 1977: Exploratory Data Analysis, John Tukey

Timeline

- 1960s - 1980s: advances in computer technology allow for new methods in processing and analysing data
- 1977: Exploratory Data Analysis, John Tukey
- 1990s: ‘data mining’ and ‘knowledge discovery’ emerge as terms for finding patterns in increasingly large datasets

Timeline

- 1960s - 1980s: advances in computer technology allow for new methods in processing and analysing data
- 1977: Exploratory Data Analysis, John Tukey
- 1990s: ‘data mining’ and ‘knowledge discovery’ emerge as terms for finding patterns in increasingly large datasets
- 1996: ‘data science’ included for first time in International Federation of Classification Societies (IFCS) conference title

Timeline

- 1960s - 1980s: advances in computer technology allow for new methods in processing and analysing data
- 1977: Exploratory Data Analysis, John Tukey
- 1990s: ‘data mining’ and ‘knowledge discovery’ emerge as terms for finding patterns in increasingly large datasets
- 1996: ‘data science’ included for first time in International Federation of Classification Societies (IFCS) conference title
- 2000s: analytics becomes increasingly important to businesses, ‘big data’ becomes a thing

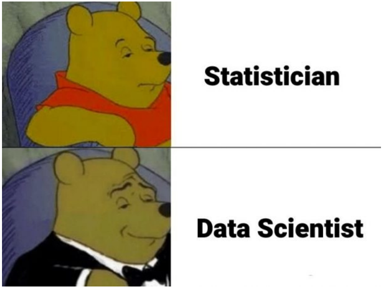
Timeline

- 1960s - 1980s: advances in computer technology allow for new methods in processing and analysing data
- 1977: Exploratory Data Analysis, John Tukey
- 1990s: ‘data mining’ and ‘knowledge discovery’ emerge as terms for finding patterns in increasingly large datasets
- 1996: ‘data science’ included for first time in International Federation of Classification Societies (IFCS) conference title
- 2000s: analytics becomes increasingly important to businesses, ‘big data’ becomes a thing

> *‘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

- **Statistics**: the mathematics associated with inference
- **Data science**: the practices associated with working with data
- Not everyone agrees with this distinction...
- Data science and statistics are essentially the same, but in practice they are coming to mean different things



Things to keep in mind

- There is a difference between **scientific** and **engineering** mindsets

Things to keep in mind

- There is a difference between **scientific** and **engineering** mindsets
  - the scientific mindset seeks to understand the underlying process (generative modeling)

Things to keep in mind

- There is a difference between **scientific** and **engineering** mindsets
  - the scientific mindset seeks to understand the underlying process (generative modeling)
  - the engineering mindset looks to find the best prediction (predictive modeling)

Things to keep in mind

- There is a difference between **scientific** and **engineering** mindsets
  - the scientific mindset seeks to understand the underlying process (generative modeling)
  - the 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**

Things to keep in mind

- **Data Science**

  Involves principles and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data - predict future outcomes, broader field

Things to keep in mind

- **Data Science**

  Involves principles and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data - predict future outcomes, broader field

  Programming, statistics, **machine learning** and algorithms towards combining, preparing and examining large datasets

Things to keep in mind

- **Data Science**

  Involves principles and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data - predict future outcomes, broader field

  Programming, statistics, **machine learning** and algorithms towards combining, preparing and examining large datasets

- **Data Analytics**

Things to keep in mind

- **Data Science**

  Involves principles and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data - predict future outcomes, broader field

  Programming, statistics, **machine learning** and algorithms towards combining, preparing and examining large datasets

- **Data Analytics**

  Analyses data to gain insights and inform decisions - past data for present decisions, specific questions.

Data science is a process

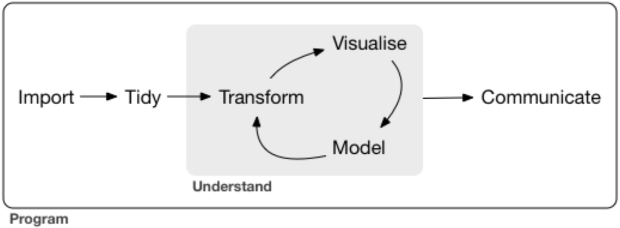


Figure: Data science tools and workflow, c/o Hadley Wickham (R for Data Science)

Data science is a process

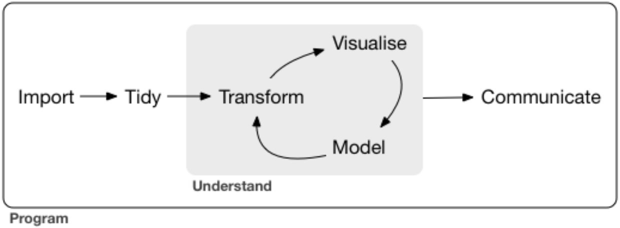


Figure: Data science tools and workflow, c/o Hadley Wickham (R for Data Science)

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

Data science is a process

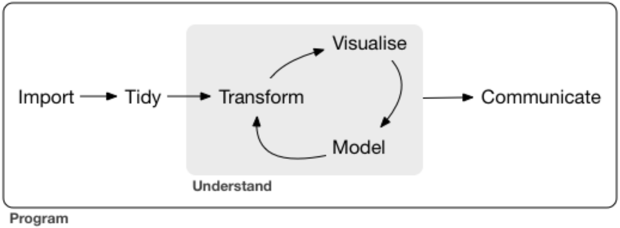


Figure: Data science tools and workflow, c/o Hadley Wickham (R for Data Science)

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

Six Divisions

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

1. Data Gathering, Preparation, and Exploration

Six Divisions

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

1. Data Gathering, Preparation, and Exploration
2. Data Representation and Transformation

Six Divisions

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!)

Six Divisions

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)

Six Divisions

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

Six Divisions

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

R and Python

YES, I WOULD LOVE TO SEE ANOTHER
R VS PYTHON FOR DATA SCIENCE COMPARISON
makeameme.org

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 `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 `rstudio.com/products/rstudio/download/`
2. Scroll down to "R Studio Desktop (Open Source License) Free" and click the "download" box underneath.
3. Follow the instructions for your system.

Open "R x64" app (console)

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

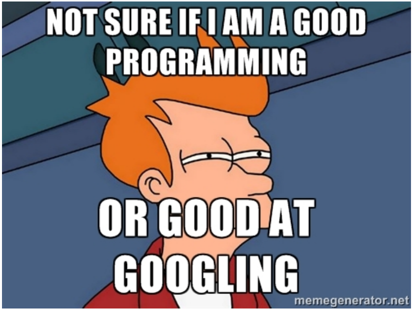
- There is a lot to learn
- There is a steep learning curve
- Slow and steady wins the race
- We are here to help

Expectations

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

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`