Meetup 1: Data Science Workflow and Toolkit

George I. Hagstrom

2024-08-28

What is Data Science?

CUNY School of Professional Studies

What is Data Science?

- Data science is a "discipline that allows you to transform raw data into understanding, insight, and knowledge"

CUNY School of Professional Studies

What is Data Science?

- Data science is a "discipline that allows you to transform raw data into understanding, insight, and knowledge"
- I hear often: "Data Science is just statistics with a clever brand name"

CUNY School of Professional Studies

What is Data Science?

- Data science is a "discipline that allows you to transform raw data into understanding, insight, and knowledge"
- I hear often: "Data Science is just statistics with a clever brand name"
- Is this a misconception?

CUNY School of Professional Studies

Data Science Workflow

Consider this visualization of the process for converting raw data into knowledge:

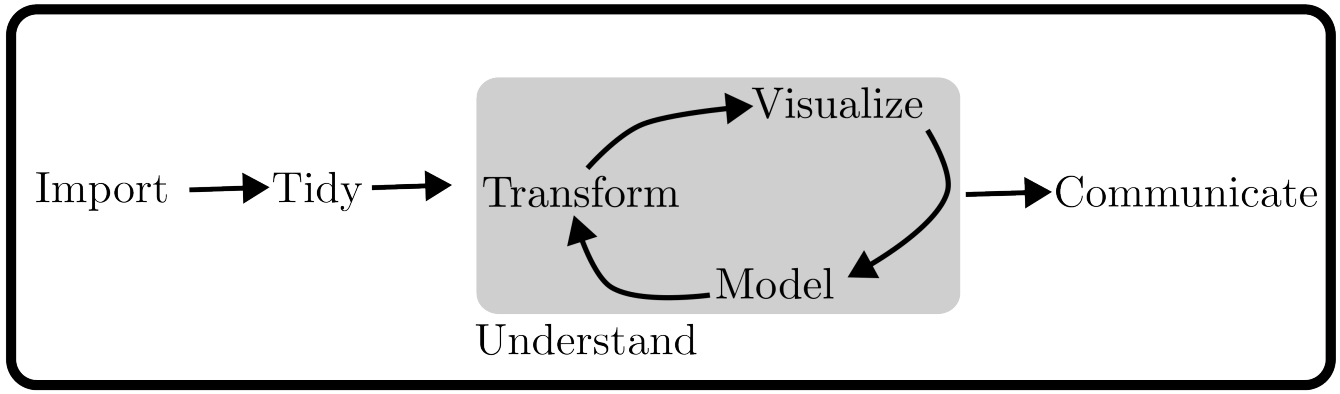


Figure 1: Figure from text

CUNY School of Professional Studies

Data Science Workflow

Consider this visualization of the process for converting raw data into knowledge:

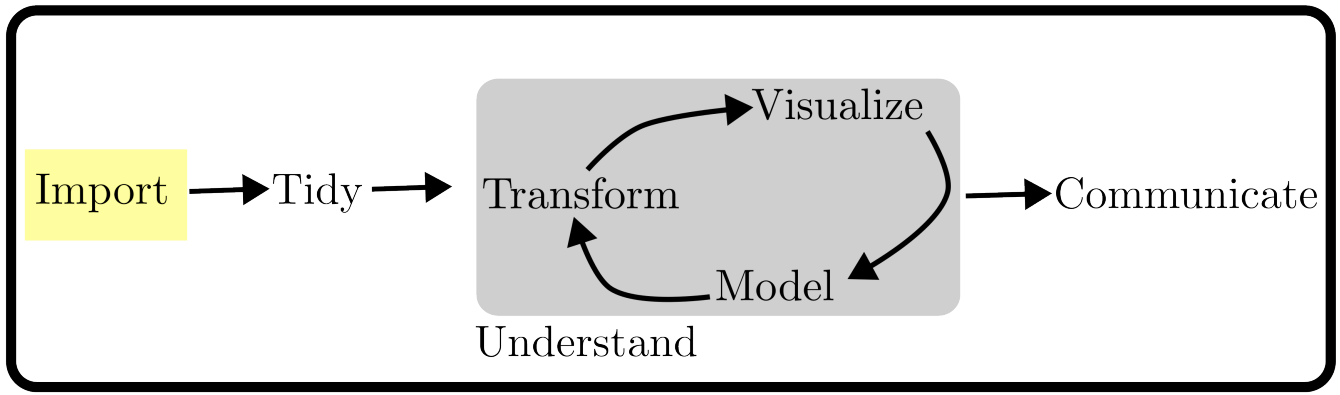


Figure 2: Figure from text

Load the data from files into software

Data Science Workflow

Consider this visualization of the process for converting raw data into knowledge:

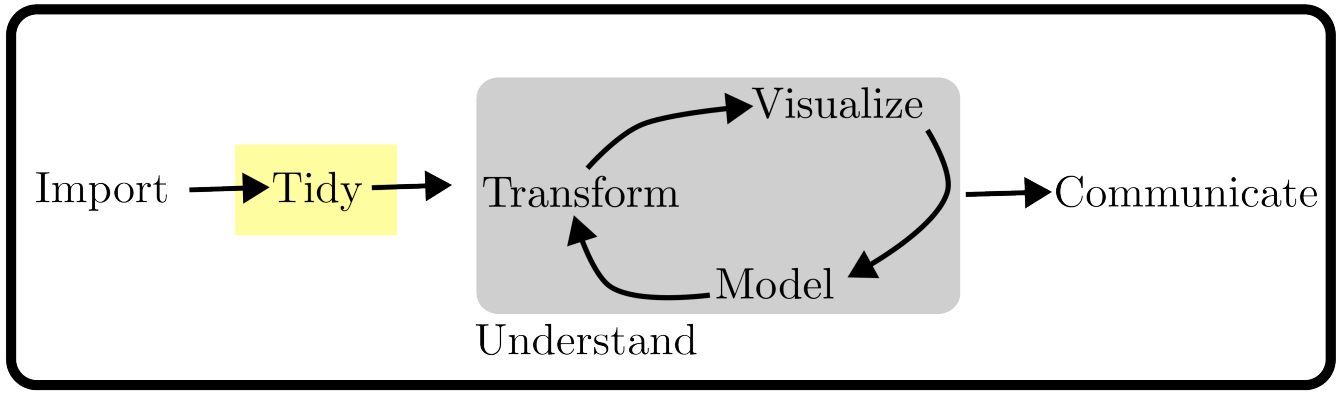


Figure 3: Figure from text

Tidy the data so it is stored in a consistent way

Data Science Workflow

Consider this visualization of the process for converting raw data into knowledge:

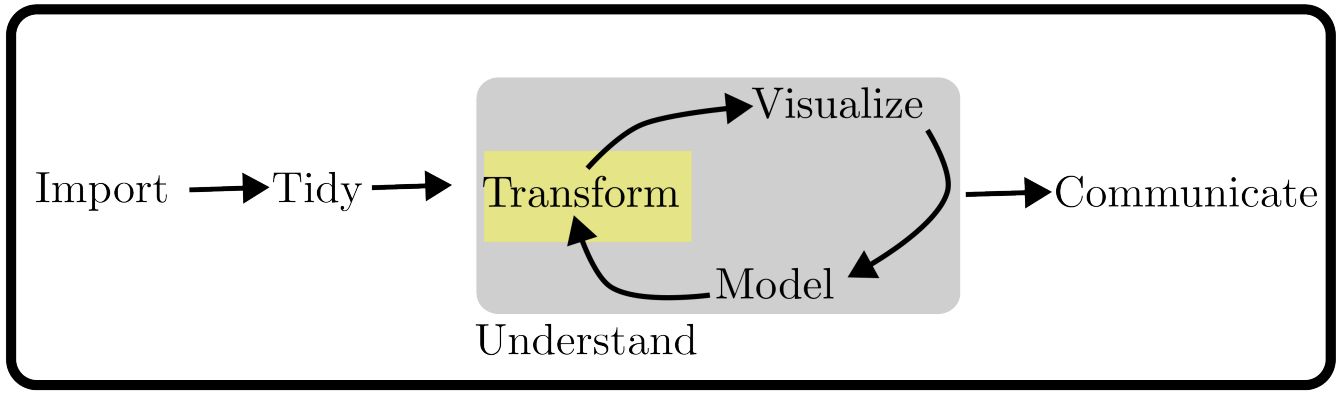


Figure 4: Figure from text

Transform the data to focus our analysis on observations of interest

Data Science Workflow

Consider this visualization of the process for converting raw data into knowledge:

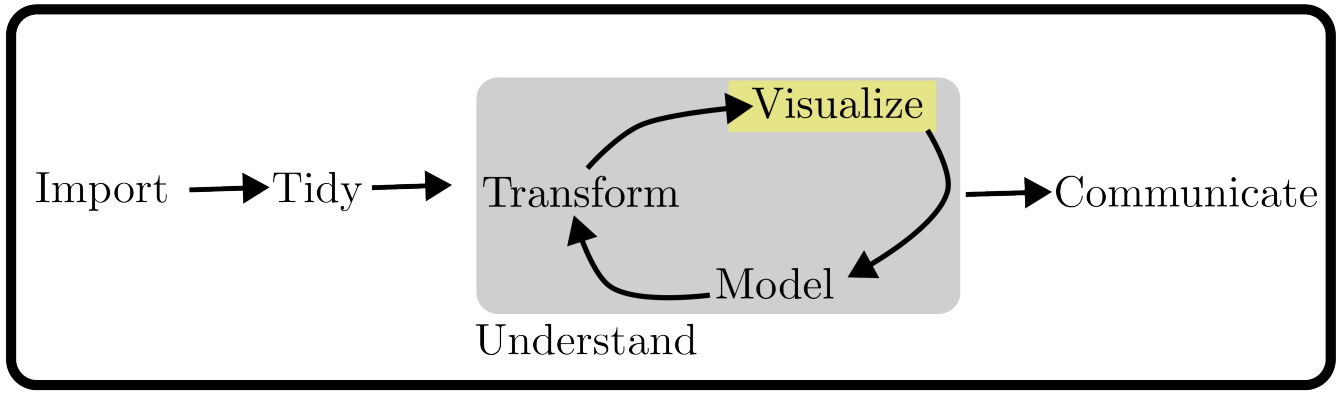


Figure 5: Figure from text

Visualize the data to find relationships, problems, and pose questions

Data Science Workflow

Consider this visualization of the process for converting raw data into knowledge:

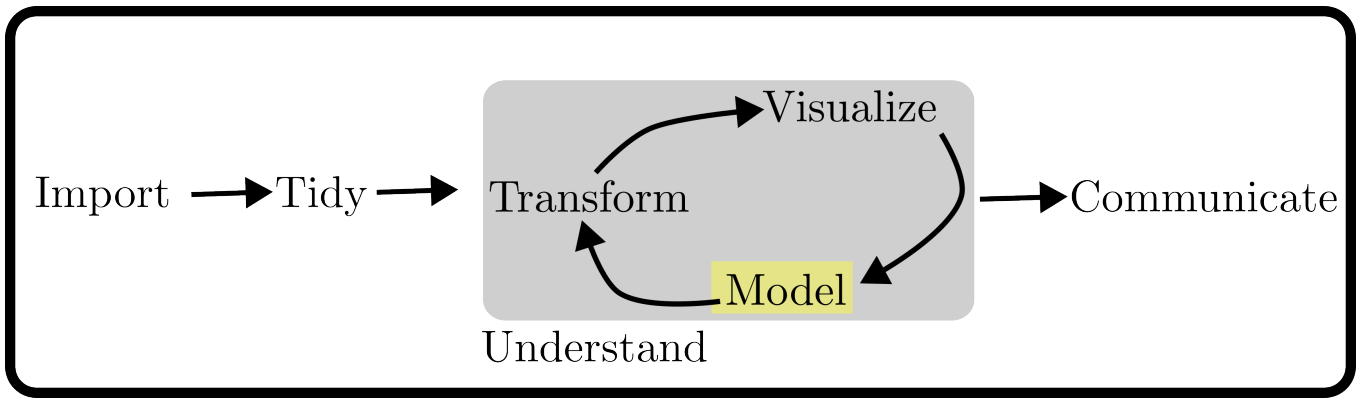


Figure 6: Figure from text

Model the data to answer questions precisely using statistics

Data Science Workflow

Consider this visualization of the process for converting raw data into knowledge:

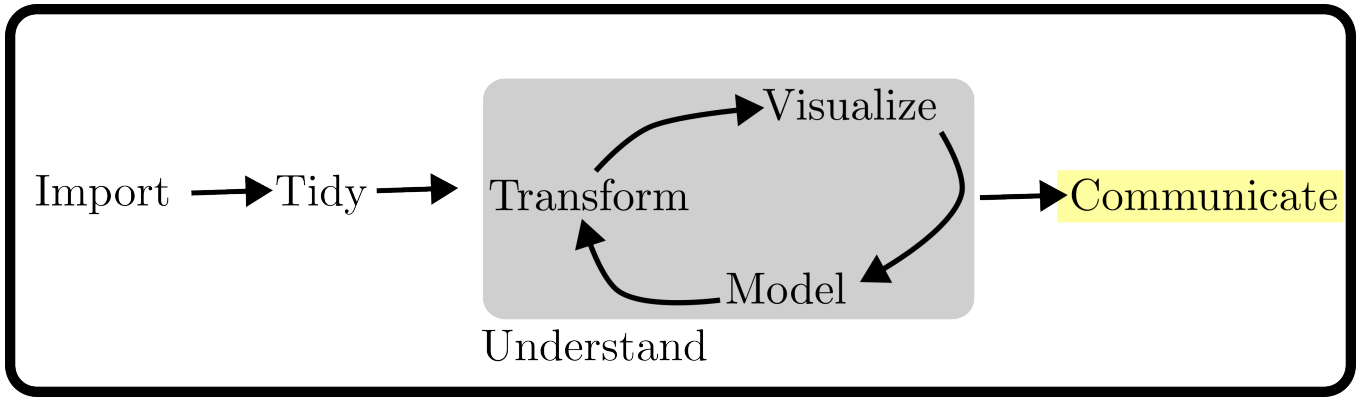


Figure 7: Figure from text

Communicate to share results with others

Data Science Workflow

Consider this visualization of the process for converting raw data into knowledge:

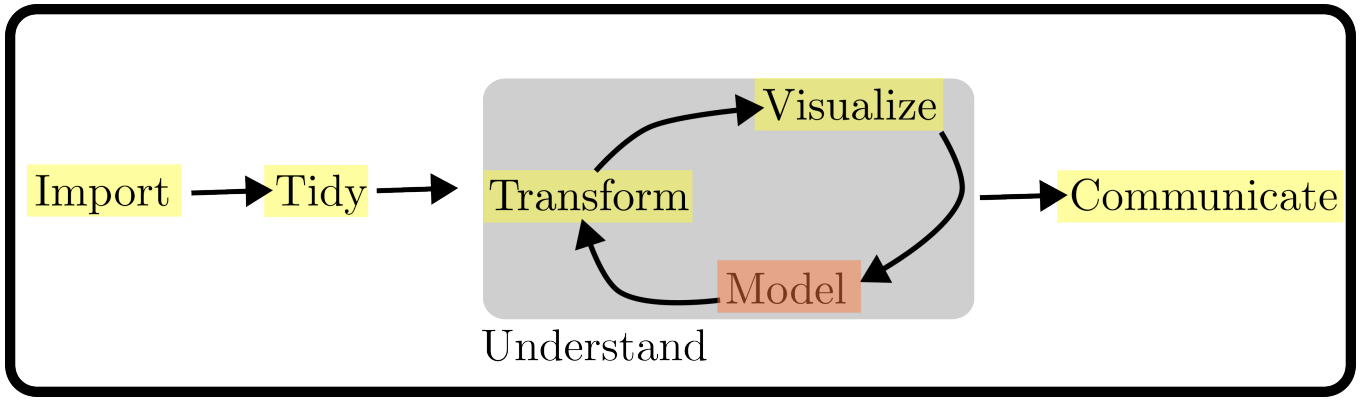


Figure 8: Figure from text

This class will focus on everything but modeling, i.e. the part of Data Science that isn't statistics

Modeling can be small part of Data Science projects

It is said that 80% of time in data science projects is spent on data mining, cleaning, tidying, exploratory data analsys, etc

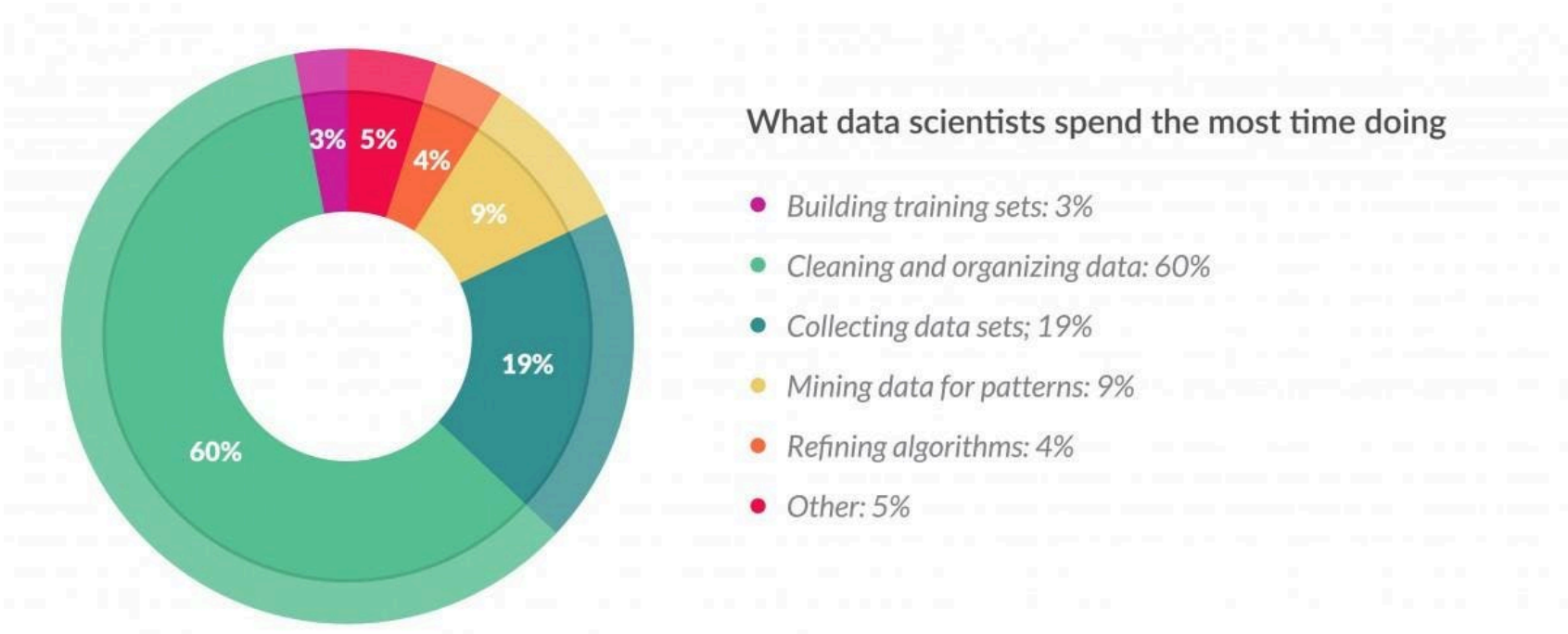


Figure 9: Figure from Forbes

Please forgive the Pie Chart

Intro/Case Study

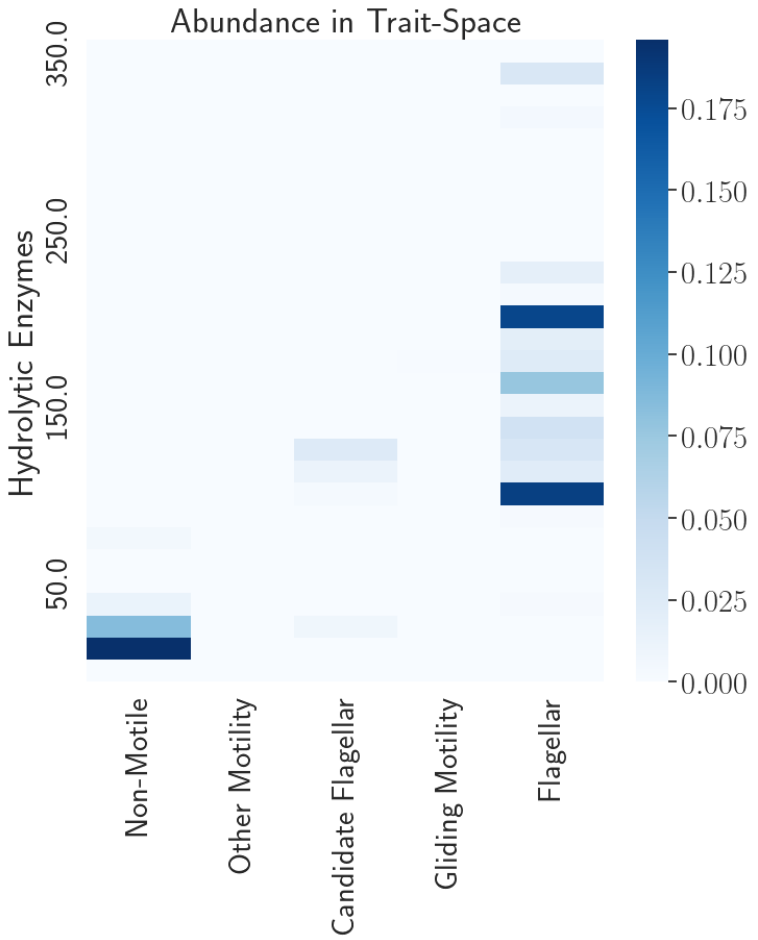


Figure 10: Trait Correlations in Marine Bacteria

- Data on how bacteria get their food in the ocean
- Getting data for this plot took months.....
- Many sources, data formats, quality issues, processing

Learning objectives

By the end of the course, you will have a foundation of skills in the Data Science Workflow

- Find data you need and do all steps to prep it for analysis
- Build expertise in R and the `tidyverse`
- Use and understand relational databases and SQL
- Collaborate with Git and GitHub
- Introduce you to distributed computing and other tools for large datasets
- Improve your programming ability

CUNY School of Professional Studies

Vignette: Electricity and CO2

Gas/Methane

Hydroelectric

Solar

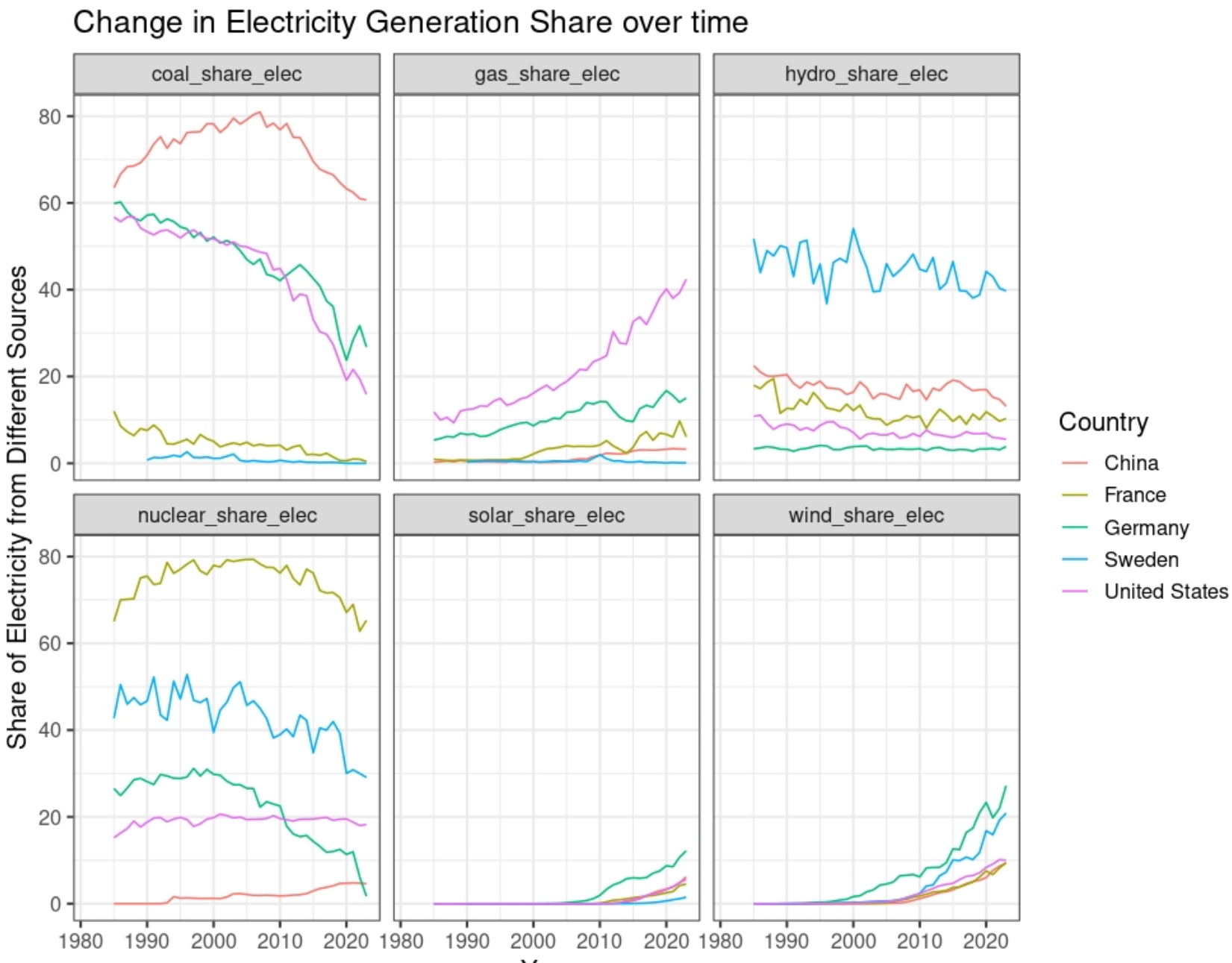
Coal

Nuclear/Fission

Wind

Figure 11: Sources of Power, refs last slide

Electricity Generation Over Time



Carbon Intensity of Electricity

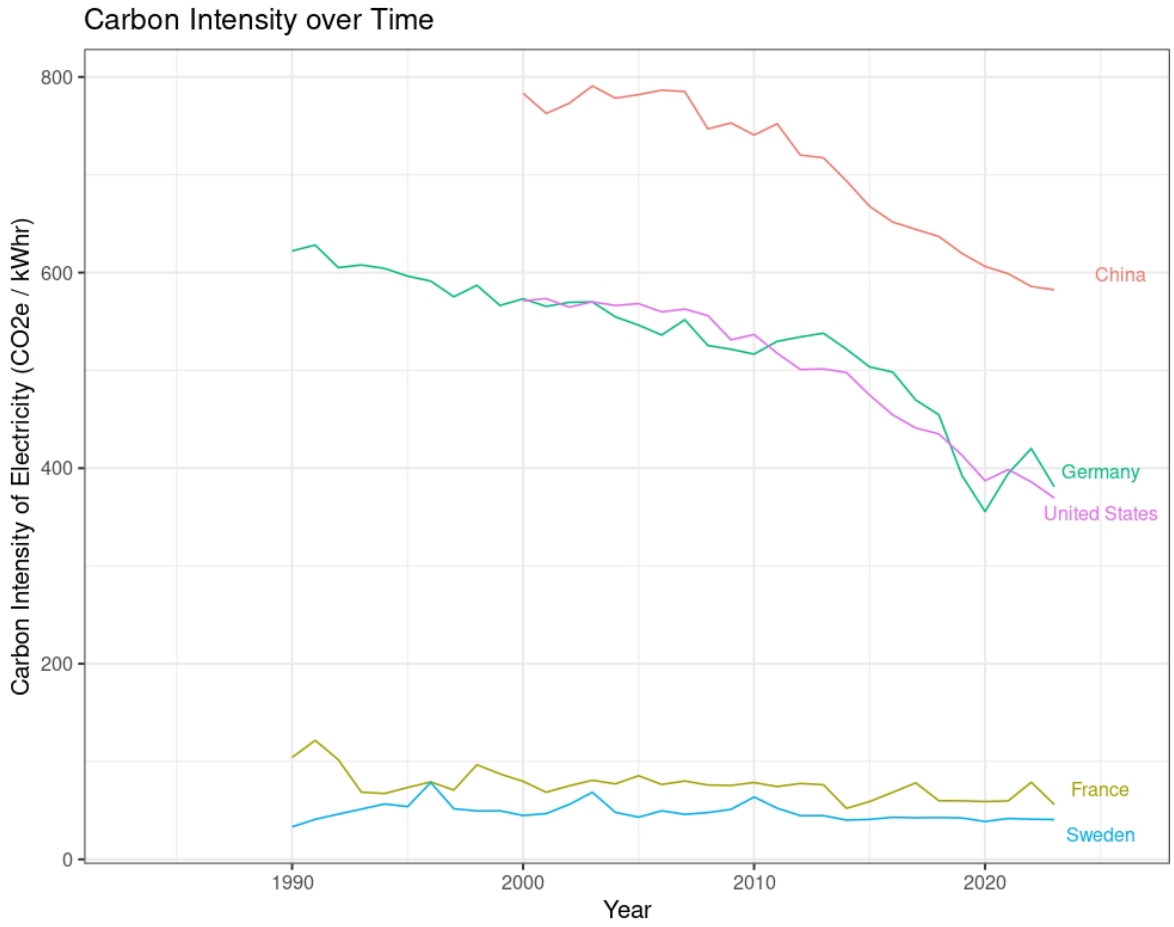
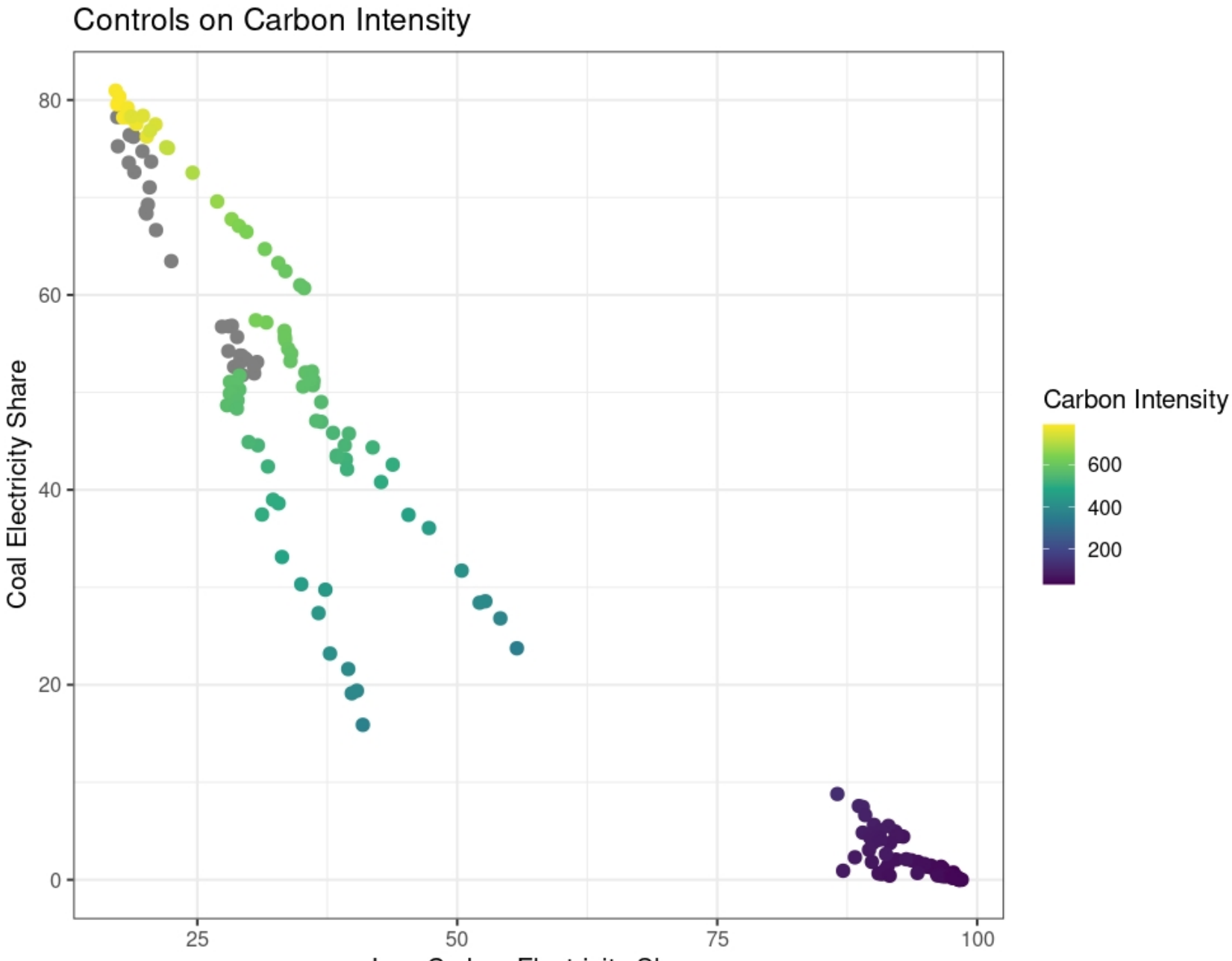


Figure 13: Source: Our World in Data

Controls on Carbon Intensity



Link to the Vignette

You can download the vignette from my github by clicking here

Remember to download the data if you want to render the file.

CUNY School of Professional Studies

Syllabus and Course Site

- Full Syllabus on the course website:
  - https://georgehagstrom.github.io/DATA607/
  - Course website contains links to weekly reading and homework assignments, meetup videos, course schedule, and other course materials
- Use the Brightspace page to submit assignments, either in `pdf` format or a link to an `html` on some site I can access (ie `github` or `rpubs`)

CUNY School of Professional Studies

Meetups

- 6:45-7:45 on Wednesday evening. Attending live preferred, watch video after if you can't
- Office Hours: On Zoom by appointment
- Communication and collaboration: https://fall2024data607.slack.com

CUNY School of Professional Studies

Assignments

- Labs (50%): Weekly Programming assignments
- TidyVerse Recipes (10%): Collaborative intro to Git
- Project (25%)
  - Assemble and explore a data set of your choosing
  - Explore your interests, build your portfolio!
- Data Science in Context Presentation (5%)
  - One 5 minute presentation, sign up for your presentation slot asap!
- Meetup Reflections and Introduction (10%)

CUNY School of Professional Studies

Schedule

Date	Start Time	Module	Video	Main Deliverables
Aug 28	06:45PM	Data Science Workflows and Toolkit		
Sep 4	06:45PM	Visualizing Data		Sep 8 Lab 1
Sep 11	06:45PM	Data Tidying and Wrangling		Sep 15 Lab 2
Sep 18	06:45PM	Exploratory Data Analysis		Sep 22 Lab 3
Sep 25	06:45PM	Data Transformations		Sep 29 Lab 4
Oct 2	06:45PM	Text and Strings		Oct 6 Lab 5
Oct 9	06:45PM	Databases and SQL		Oct 13 Lab 6
Oct 16	06:45PM	Advanced R Programming		Oct 20 Proj. Proposal
Oct 23	06:45PM	Webscraping and APIs		Oct 27 Lab 7
Oct 30	06:45PM	Git and Collaboration		Nov 3 TV Create
Nov 6	06:45PM	Tidy Text and NLP		Nov 10 Lab 8
Nov 13	06:45PM	Graphs and Graph Data		Nov 17 Lab 9
Nov 20	06:45PM	Big Data		Nov 24 TV Extend
Nov 27		No Meetup (Thansgiving)		
Dec 4	06:45PM	Cloud Computing		Dec 8 Lab 10
Dec 11	06:45PM			Dec 15 Final Project

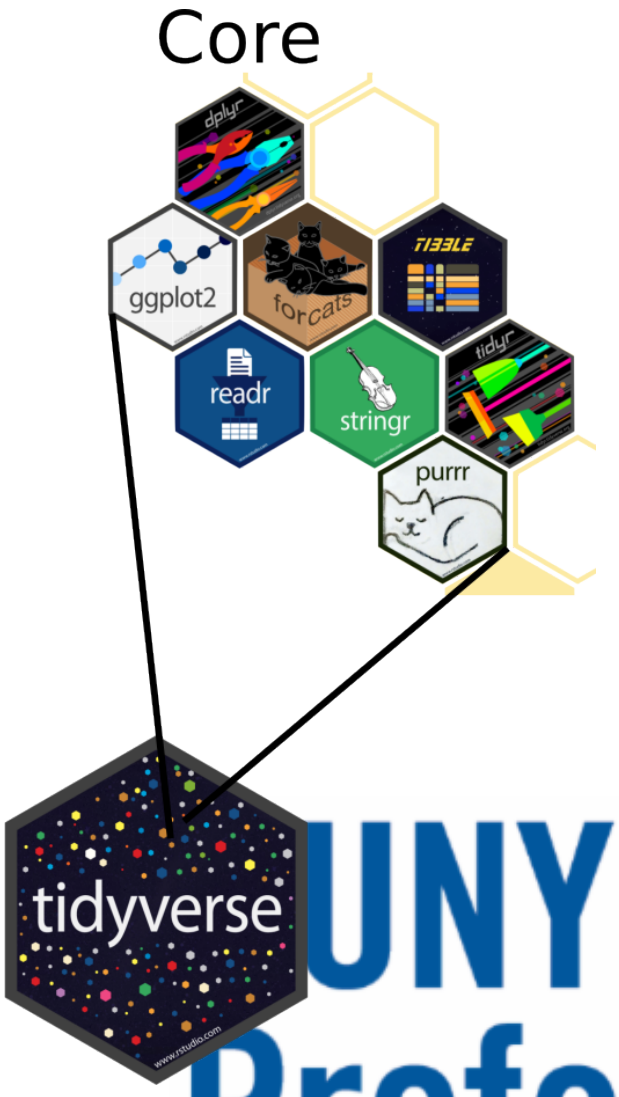
Textbooks

1. Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. (2023). *R for Data Science (2e)*. O'Reilly
2. Jennifer Bryan. *Happy Git and GitHub for the R User.*
3. Julia Silge and David Robinson (2017). *Text Mining with R.* O'Reilly

Recommended: Wickham, H. *Advanced R.* Baca Raton, FL: Taylor & Francis Group.

CUNY School of Professional Studies

Tidyverse: Opinionated Ecosystem



- Collection of compatible packages
- Shared philosophy, common grammar
- Strong Core, Many Extensions
- Advantages and Disadvantages

CUNY School of Professional Studies

What to do this week?

1. Readings:
   i) Intro and Chapter 28 of R4DS
   ii) Sections 1-15 of Happy Git
   iii) Quarto Tutorial
   iv) Appendix on R Help Files
2. Get software installed and configured:

i) R, RStudio, git, latex

3. Write a post introducing yourself on Brightspace Discussions
4. Sign up for your Data Science in Context Presentation

CUNY School of Professional Studies

Image References

1. Coal: By Morgre - Own work, CC BY-SA 3.0
2. Gas/Methane: By Georg Slickers - Self-photographed, CC BY-SA 3.0
3. Hydro: By Source file: Le Grand PortageDerivative work: Rehman - File:Three_Gorges_Dam,_Yangtze_River,_China.jpg, CC BY 2.0
4. Solar: By Parabel GmbH - Own work, CC BY-SA 3.0
5. Wind: By Erik Wilde from Berkeley, CA, USA - harvesting wind, CC BY-SA 2.0