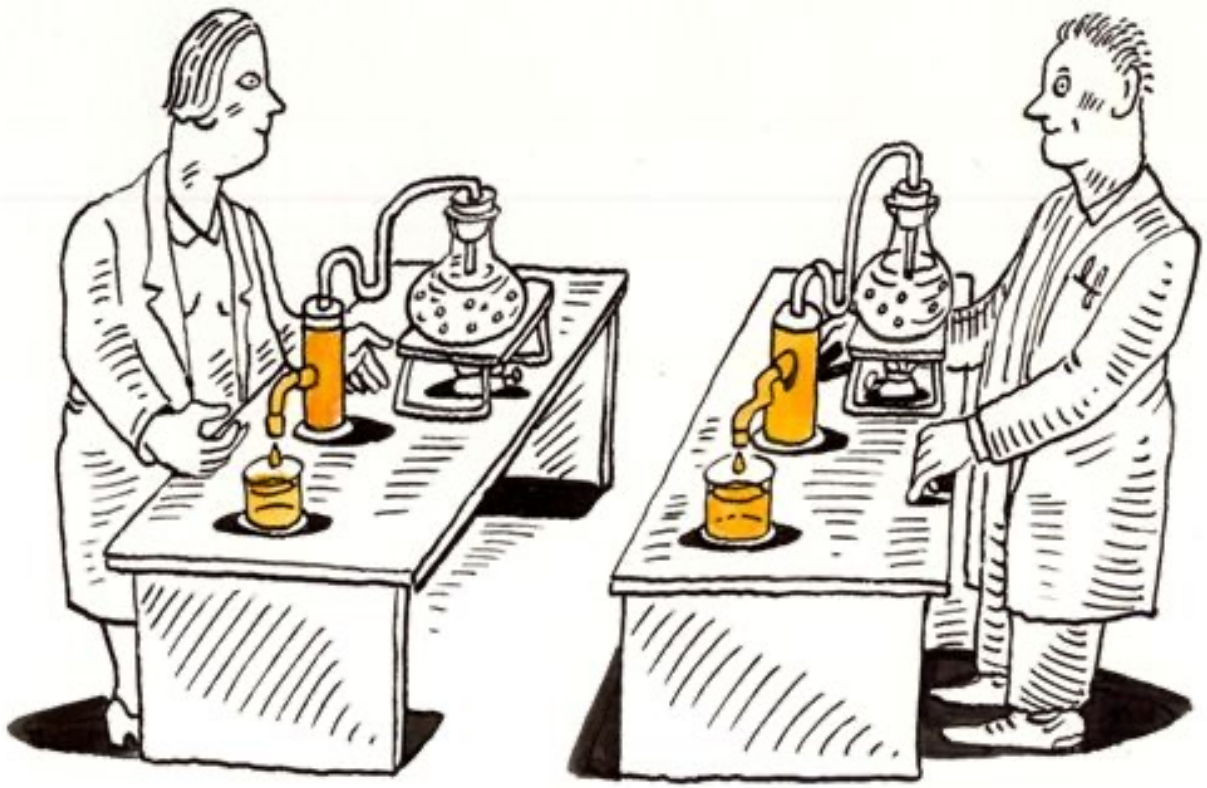
Introduction to Reproducible Science w/ Python

Nelson Roque, PhD
Assistant Professor, Penn State University

Scan for Slides & Code available on GitHub



Background

- Over the last decade I've engaged in professional development and research opportunities to become proficient across topics including data wrangling and modeling of text, image, video, and eye-tracking data, as well as more recently, sensor data.

- I look forward to training the next generation of scientists on code-based methods to enhance reproducibility and usability of their research results.

Workshop Materials & Required Software

- All slides, code, and recordings will be made available at the website listed below, as well as Github
  - Code & Slides
    - https://github.com/nelsonroque/contextlab_reproduciblescience_workshops
  - Textbook
    - https://nelsonroque.github.io/contextlab_introdatascience_webcourse/

- Cloud-based Python environment
- Provides free GPU and TPU resources
- Jupyter Notebook integration
- Real-time collaboration
- Accessible from any device
- Integrated with Google Drive
- Supports rich text and media
- Pre-installed popular libraries
- Ideal for data science and ML

Workshop Learning Objectives

1. Describe various principles, tools and techniques supportive of open and reproducible science.
2. Develop a code-only pipeline to allow reproducibility of data preparation and analyses.
    1. Learn about data wrangling principles (e.g., tidy data)
    2. Import, pre-process & visualize data using Python
3. Develop a long-term learning plan for practicing reproducible science tools and techniques.
    1. Share recommended readings, activities for continued learning

Agenda for Today

- What is Reproducible Science? Why?
- Tools Supporting Reproducible Science
- What is Data Science?
- Workflows Supporting Reproducible Data Science Science
- Orientation to Python, Jupyter Notebooks, Google Colab
  - **Skill 1: Python syntax primer**
  - **Skill 2: Data wrangling and visualization**
- Data Science: Latest trends
- Long-term Learning Recommendations

What is Reproducible Science?

Reproducible versus replicable

- **Reproducible:** when the exact results can be reproduced if given access to the original data, software, or code

- **Replicable:** research results can be reproduced by independent researchers using different methods.

What is reproducible science?

- Making entire scientific process transparent (as allowable; often required by law, grants).
- Sharing experiment, raw data, and analysis code
- Detailed methods (e.g., STROBE checklist)
- Sharing stimulus sets
- Being able to walk through a data analysis start (load raw data) to finish (manuscript analyses) in code
- Following FAIR Principles
- Pre-registering hypotheses and analysis plans (e.g., on OSF)

Why Reproducible Science?

Background

- A reproducibility crisis has emerged as a threat to the scientific enterprise.
  - Ioannidis, John P A. 2005. “Why Most Published Research Findings Are False.” PLoS Medicine 2 (8): e124. doi:10.1371/journal.pmed.0020124.
  - Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, *1*(1), 1-9.
  - Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” Science 349 (6251): aac4716–aac4716. doi:10.1126/science.aac4716.

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. “Negative” research is also very useful. “Negative” is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of “true relationships” to “no relationships” among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

Experimenter Degrees of Freedom

- Choices researchers have in experiment design, data analysis, and reporting
  - Researchers choose variables, sample sizes, and conditions.
  - Researchers select statistical tests and criteria for significance.

- Can lead to variability in research approaches and results

Benefits of reproducible research

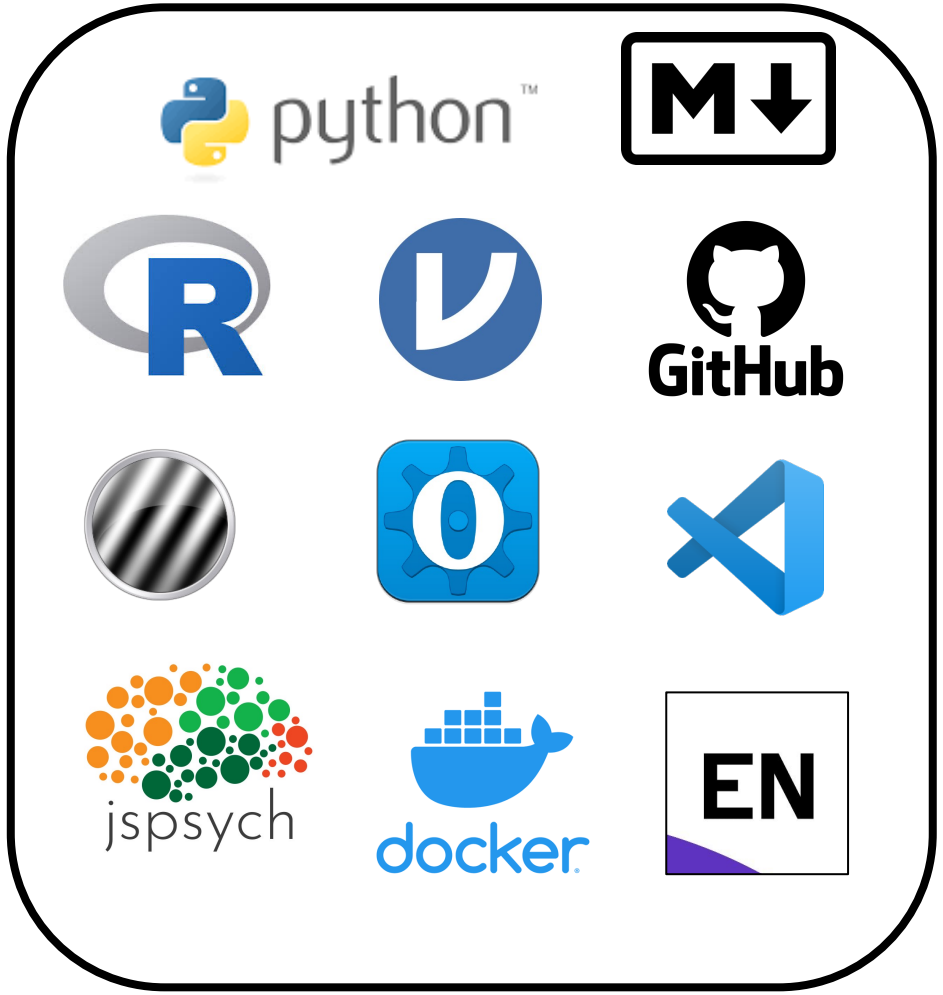
- increased likelihood that the research will be correct
- reproducibility makes it easier to check the research
- it is easier to reproduce the research independently
- easier to extend the research
- reusable code and instruction resulting in increased efficiencies

https://www.displayr.com/what-is-reproducible-research

Tools Supporting Reproducible Science

Tools Supporting Reproducible Science

- R
- Latex, Markdown
- Python, Anaconda
- Github, Github Pages: https://pages.github.com
- Docker, VMs
- Infrastructure as Code (IaC; e.g., AWS CDK)
- IDEs (e.g., Visual Studio Code)
- Open Science Framework
- Documentation software: Docusaurus, https://docusaurus.io
- Code-based experiment creation software: Opensesame, Psychopy, jsPsych



What is Data Science?

What is data science?

Data science is all about collecting, analyzing, and using data to solve mysteries, make predictions, and help the world work better.

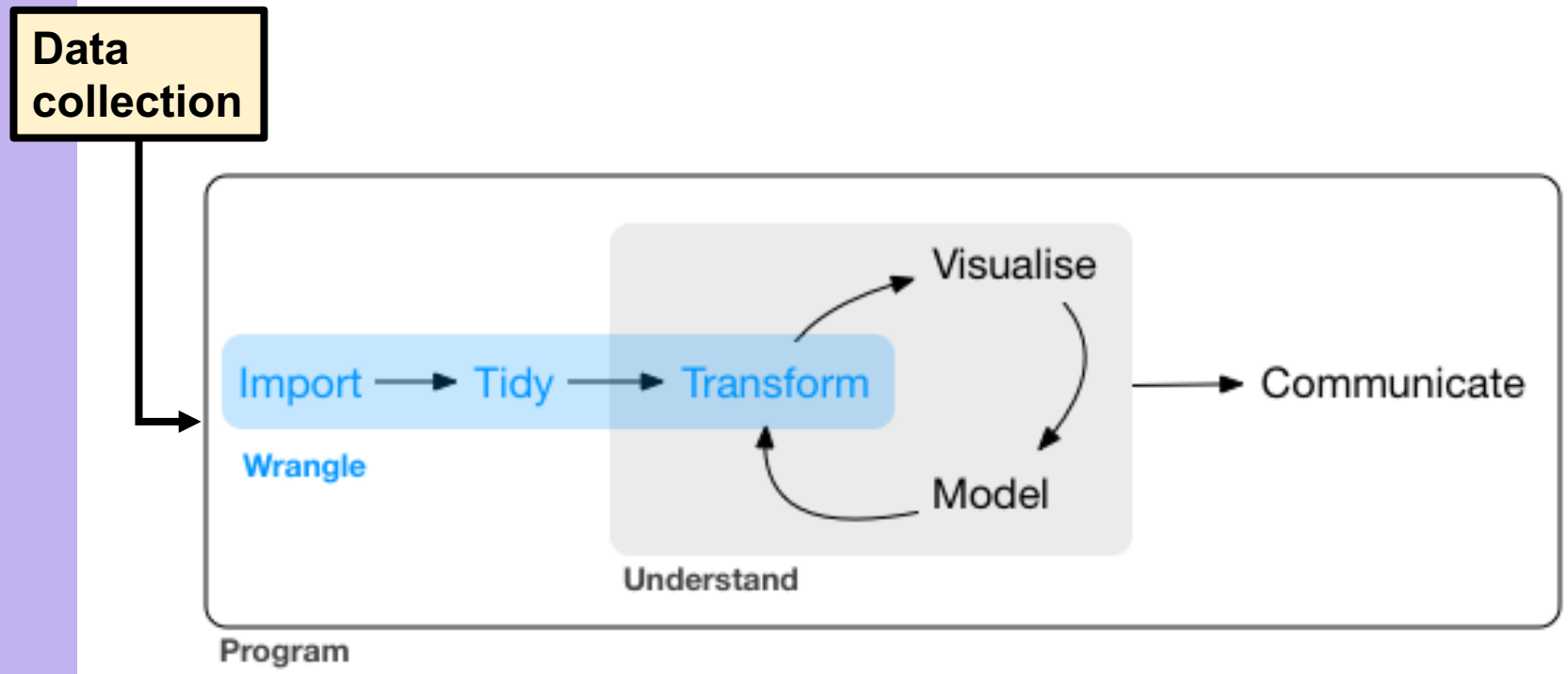
In a perfect data-generating system, data are already tidy.

Uncleaned Data Problems

- In practice, we must first determine data structure issues:
  - Column headers are values, not variable names.
  - Multiple variables are stored in one column.
  - Variables are stored in both rows and columns.
  - Multiple types of observational units are stored in the same table.
  - A single observational unit is stored in multiple tables.
- Data quality issues
  - Values out of range
  - Improperly/inconsistently coded response options
  - Inconsistent records counts

Workflows Supporting Reproducible Data Science

What is a typical data science workflow?



- The operations that carry your data from raw form, into something visualizable or analyzable
  - i.e., the data preparation phase of the research

Source: R for Data Science

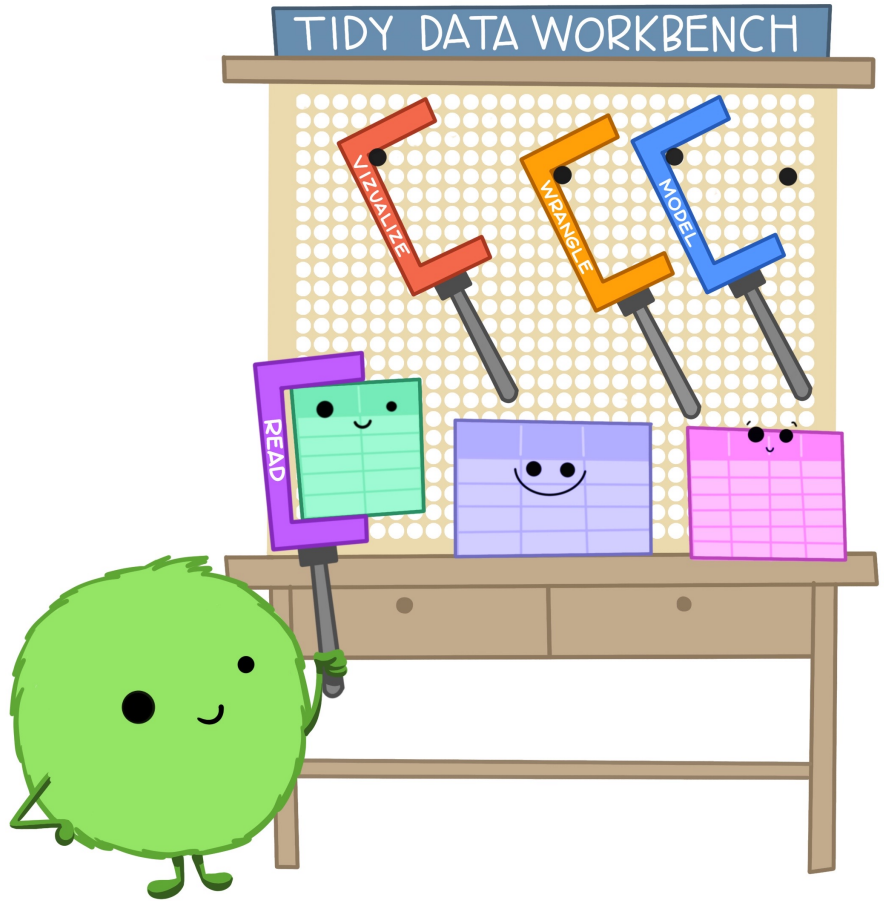
What is tidy data?

1. Every column is a variable.
2. Every row is an observation.
3. Every cell is a single value.

participant	condition	avg_response_time	perc_accuracy
9991	control	506	90
9992	active	516	96
9993	control	526	99

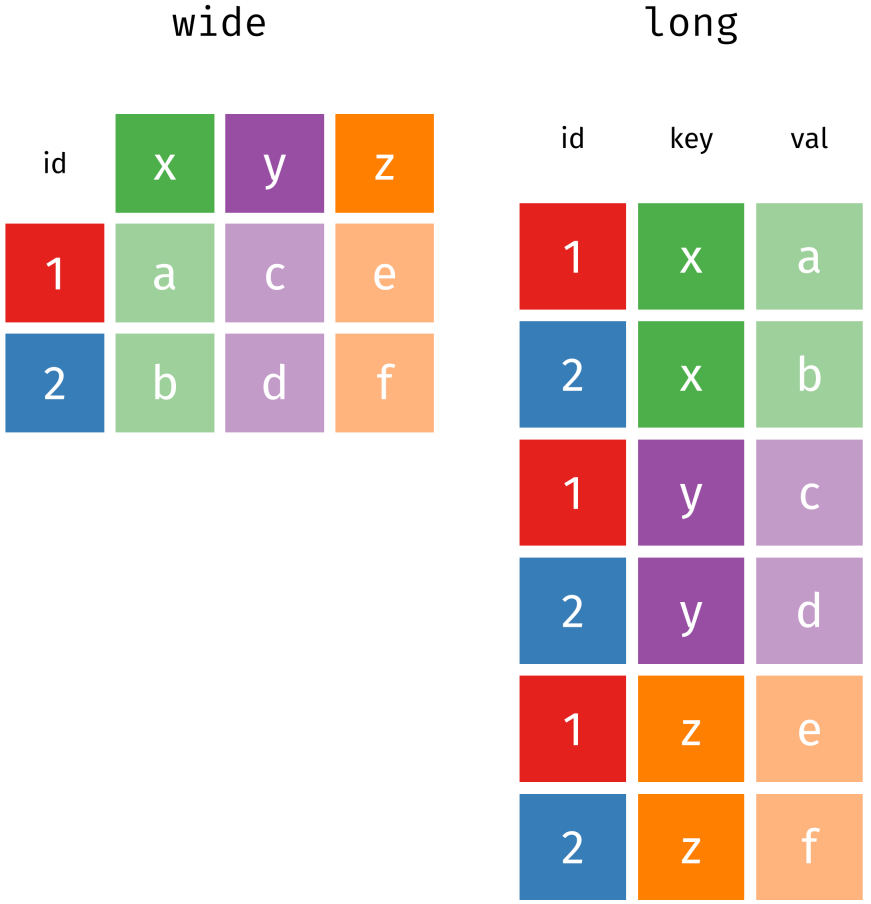
Source: R for Data Science

When working with tidy data,
we can use the same tools in
similar ways for different datasets...



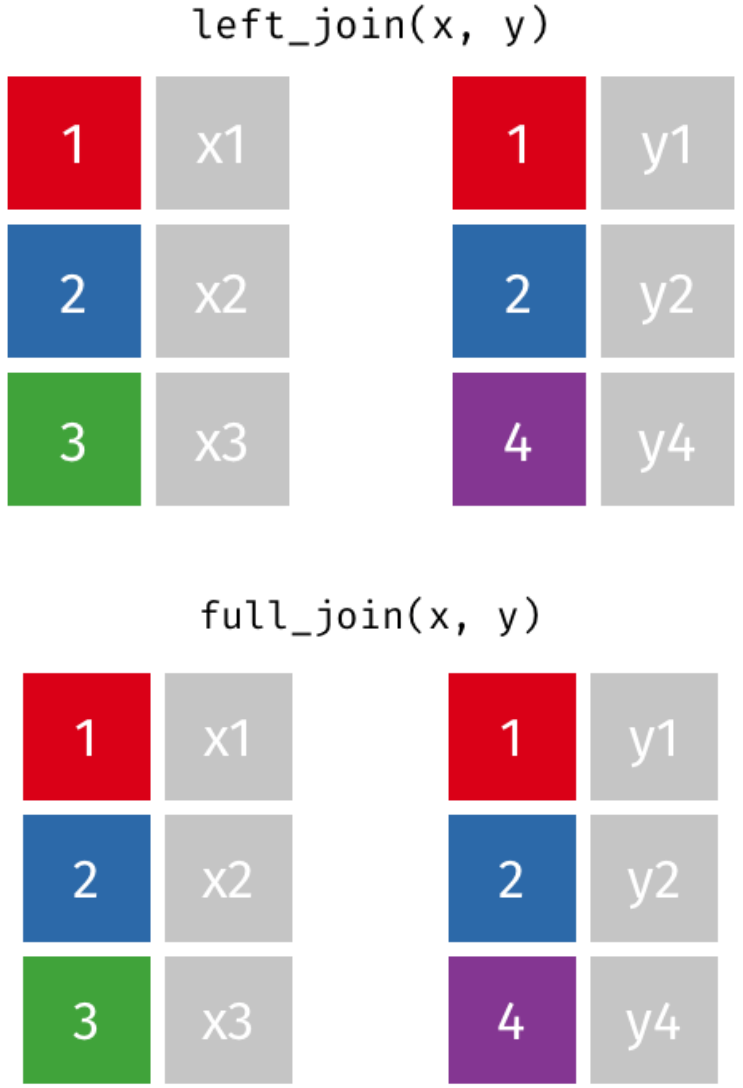
Data wrangling operations

- Data exploration
  - Plotting, descriptive stats
- Dealing with missing data
  - Impute missing data, insert missing codes (-999)
- Reshaping data
  - Add columns (e.g., create flag is missing more than 10 records)
  - Update column names (id = participant_id)
  - Converting between long and wide format



Data wrangling operations

- Filtering data
  - Remove specific observations, by column or row
- Merging/matching data from various sources
  - Join data by common id(s) in various ways (more on next slides)
  - https://dplyr.tidyverse.org/articles/two-table.html
- Other wrangling
  - Feature engineering
    - E.g., Add ‘features’ of date as new columns (e.g., what is day of week for 10/13/2020)



Orientation to Python, Jupyter Notebooks, Google Colab

Interactive Live Coding Segment Begins In 6 Slides!

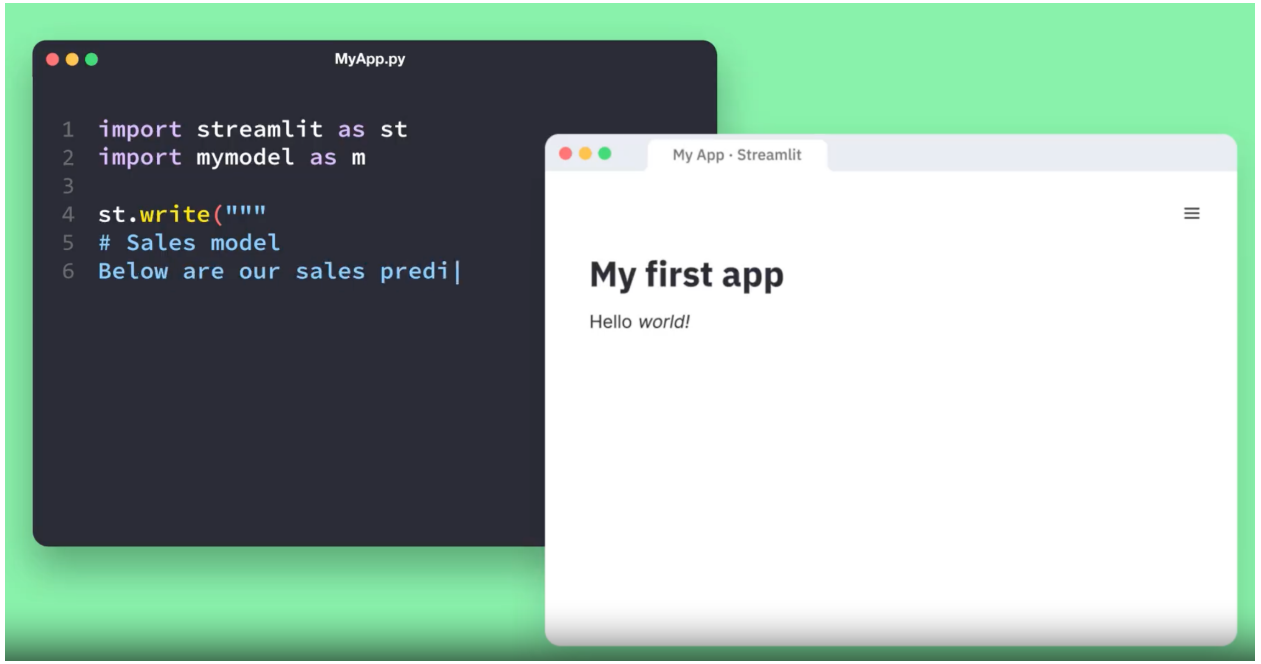
What is Python? Jupyter Notebooks?

- **What is Python?**
  - Python: High-level programming language
  - Versatile: Popular for data analysis, web development, machine learning and AI
  - Easy to learn

- **What is Jupyter Notebooks?**
  - Interactive documents: Combines code, text, and visualizations
  - Ideal for data exploration and analysis
  - Supports multiple programming languages
  - Popular among data scientists and researchers

What can you build with Python?

- Websites
- Mobile and desktop apps
- Machine learning models
- Bots: Chat Bot, Discord Bot
- Dashboards
- **Data science**



An example of Getting Started with Streamlit, a popular Python library for writing dashboards and other web apps: https://streamlit.io/

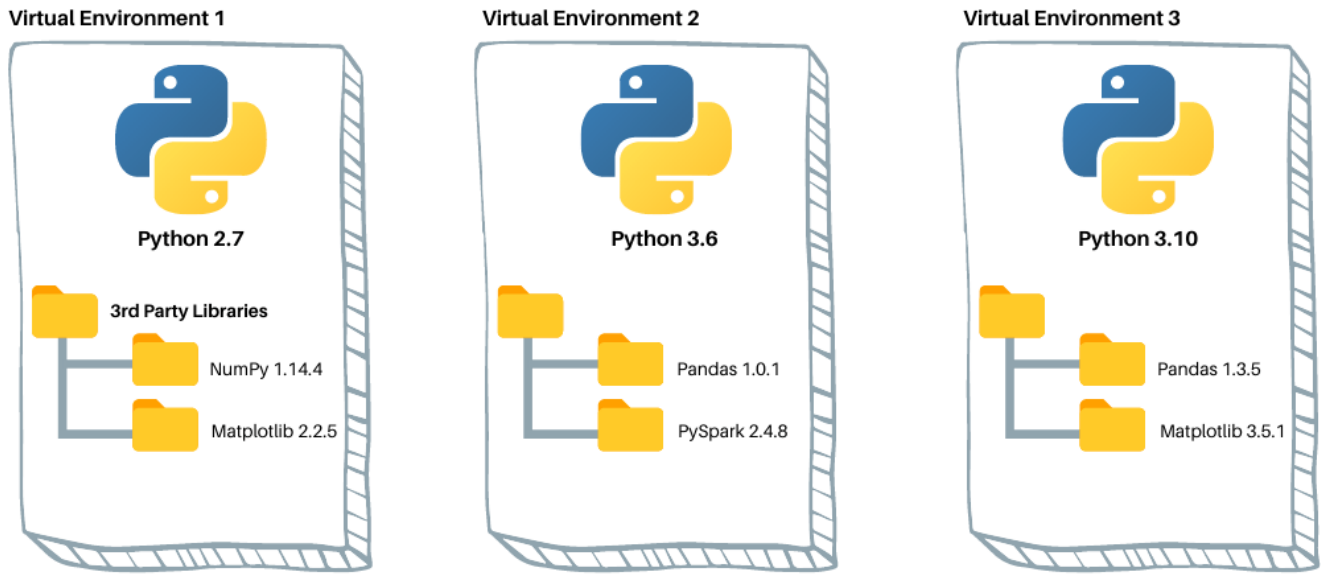
What are Python libraries?

- **Libraries of code to accomplish specific functions**
  - e.g., data viz, machine learning, data wrangling, compute effect sizes

- **Think of `pip` as an equivalent to iOS App Store, or Google Play Store**
  - CRAN is where you download published R packages (directly from R, RStudio)
  - pip is where you download published Python packages

What are Virtual Environments?

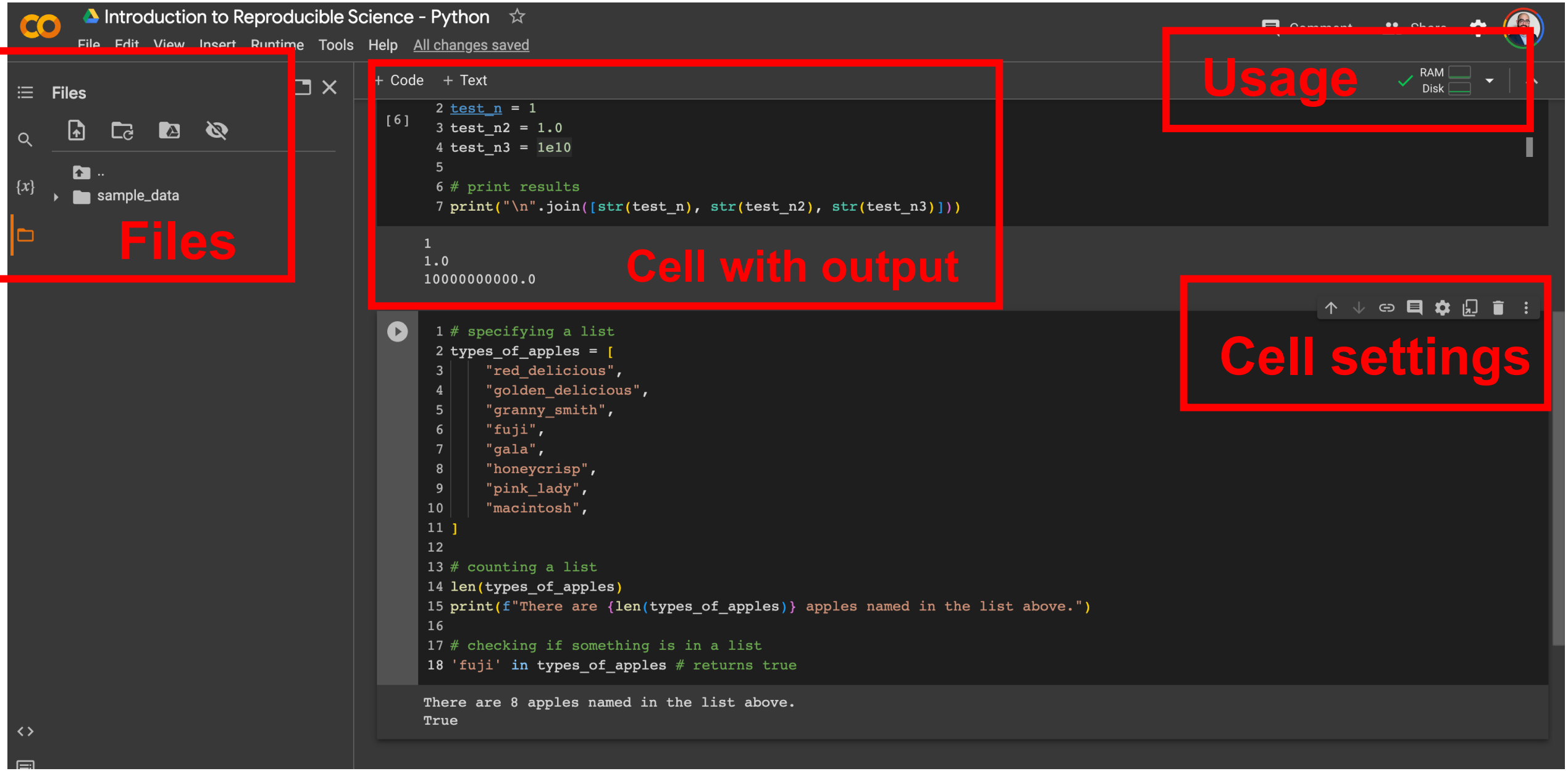
- **A tool to segment the libraries needed for a project, away from libraries needed for another project**
  - e.g., library A version 1.0 may be needed for Project A, but library A version 2.0 may be needed for Project B



Basic data types & structures

- **Integers (int):** Whole numbers (e.g., -3, 0, 42).
- **Floating-Point Numbers (float):** Numbers with decimal points (e.g., 3.14, -0.001).
- **Strings (str):** Text enclosed in quotes (e.g., "Hello, World!", 'Python').
- **Booleans (bool):** Represents True or False.
- **Lists (list):** Ordered collections of items (e.g., [1, 2, 3]).
- **Tuples (tuple):** Immutable ordered collections (e.g., (1, 2, 3)).
- **Dictionaries (dict):** Key-value pairs (e.g., {"name": "John", "age": 30}).
- **Sets (set):** Unique, unordered collections (e.g., {1, 2, 3}).
- **NoneType (None):** Represents the absence of a value.

Anatomy of Google Collab



Let’s jump into Python

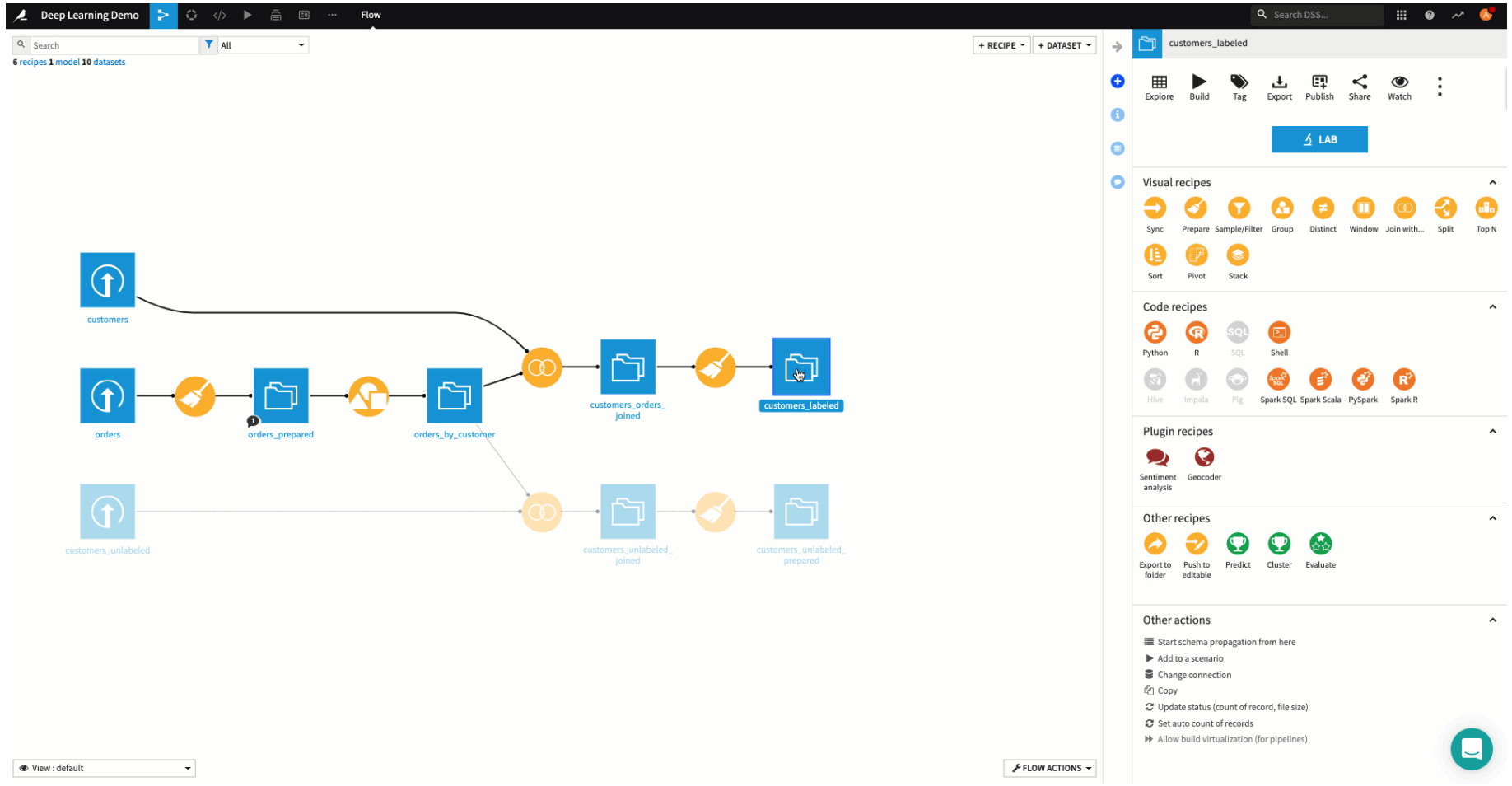
Scan the QR code to view the code we will work with

Data Science: Latest Trends

The Future of Data Workflows: No/Low Code

Seamless integration platforms, for example:

data iku



Long-term learning

Developing your Learning Plan

- Subscribe to Python, ML blogs (e.g., Towards Data Science)
- Create a personal library for ongoing reference
- Attend webinars, seminars, and conferences
- Pursue certifications (e.g., Coursera, Udacity) for validation
- Join local meetups and networking events
- Seek mentorship and mentor others
- Regularly read research papers and tech news (tldr).
- **PRACTICE WITH REAL DATA**

PRACTICE WITH REAL DATA

https://www.kaggle.com/datasets

https://www.data.gov/

You are not alone

ChatGPT

stackoverflow

Download Python Cheatsheets

- Tutorials
  - https://github.com/mGalarnyk/Python_Tutorials
  - https://www.geeksforgeeks.org/data-science-tutorial/
  - https://www.w3schools.com/datascience/ds_python.asp
- Books
  - https://jakevdp.github.io/PythonDataScienceHandbook/
  - https://www.knowledgehut.com/blog/data-science/python-data-science-books
  - https://wesmckinney.com/book/

THANK YOU

nur375@psu.edu

nelsonroque.com

https://github.com/nelsonroque