



INMAS

Workshop Overview

- *Introductions*
- *Programming with Data*

James Balamuta

Inmas Fall 2021 Statistical Methods Workshop



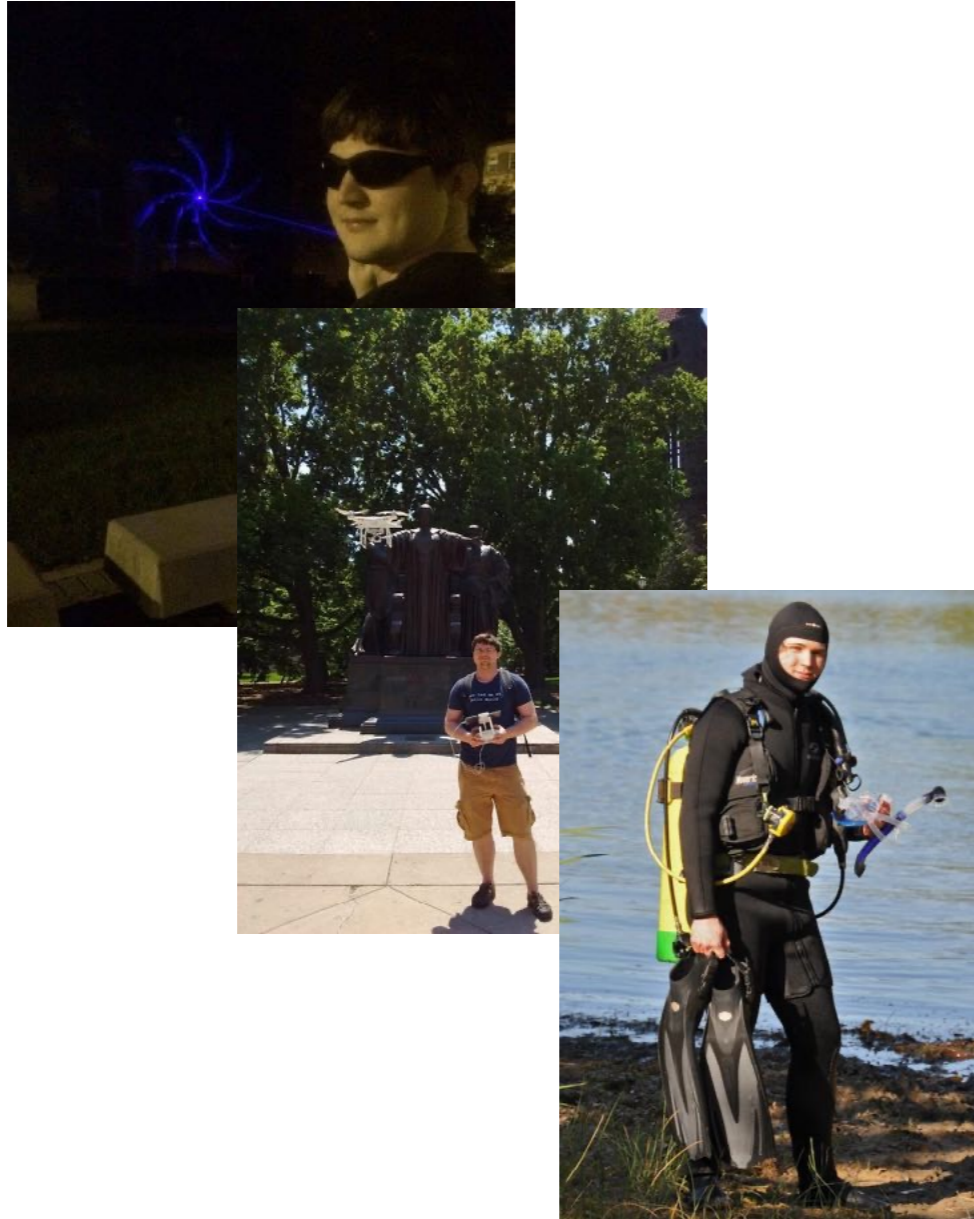
Lecture Objectives

- **Introduce** workshop structure
- **Compare** and **contrast** different forms of programming

Hello
my name is

James

Who am I?



James Balamuta
Visiting Assistant Professor
Department of Statistics
balamut2@illinois.edu

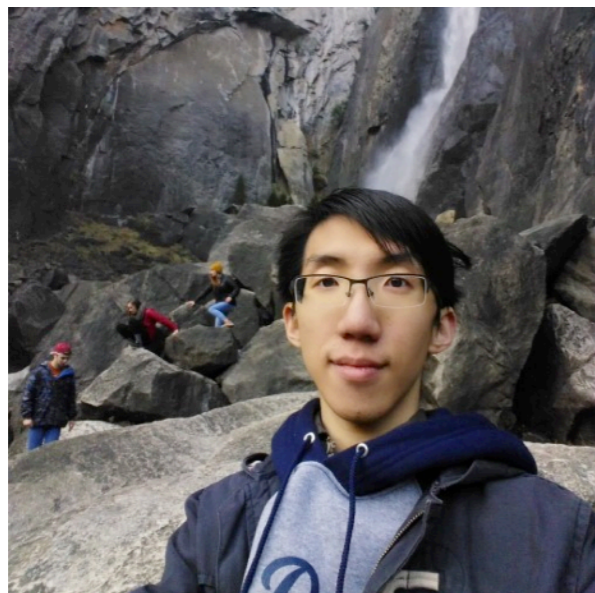
OCIO Cloud Innovation Scholar & Advisor
MLPACK Contributors Team
Google Summer of Code Mentor

Teaching Assistants (TAs)

UIUC



David Lundquist



Man Feng "Herman" Leung

JHU



Ningyuan (Teresa) Huang



Salma Tarmoun

How much data has been
generated since we started
talking?



DATA NEVER SLEEPS 8.0

How much data is generated *every minute*?

In 2020, the world changed fundamentally—and so did the data that makes the world go round. As COVID-19 swept the globe, nearly every aspect of life—from work to working out—moved online, and people depended more and more on apps and the Internet to socialize, educate and entertain ourselves. Before quarantine, just 15% of Americans worked from home. Now over half do. And that's not the only big shift. In our 8th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute—a trend that shows no sign of stopping.



The world's internet population is growing significantly year over year. As of April 2020, the internet reaches 59% of the world's population and now represents 4.57 billion people — a 6% increase from January 2019.



GLOBAL INTERNET POPULATION GROWTH 2014-2020 (IN BILLIONS)

As the world changes, businesses need to change with the times—and that requires data. Every click, swipe, share or like tells you something about your customers and what they want, and Domo is here to help your business make sense of all of it. Domo gives you the power to make data-driven decisions at any moment, on any device, so you can make smart choices in a rapidly changing world.

Learn more at domo.com

SOURCES: STATISTA, VITAL CAPITALIST, BUSINESS INSIDER, GAMESPOT, TECHCRUNCH, OMNICORE AGENCY, DOORDASH, BUSINESS OF APPS, NEW YORK TIMES, MUSIC BUSINESS WORLDWIDE, INC., THE VERGE, INC., HOOTSuite, JUSTIN STOUT, REDDIT, USER, AMAZON, VON



<https://www.domo.com/learn/data-never-sleeps-8>

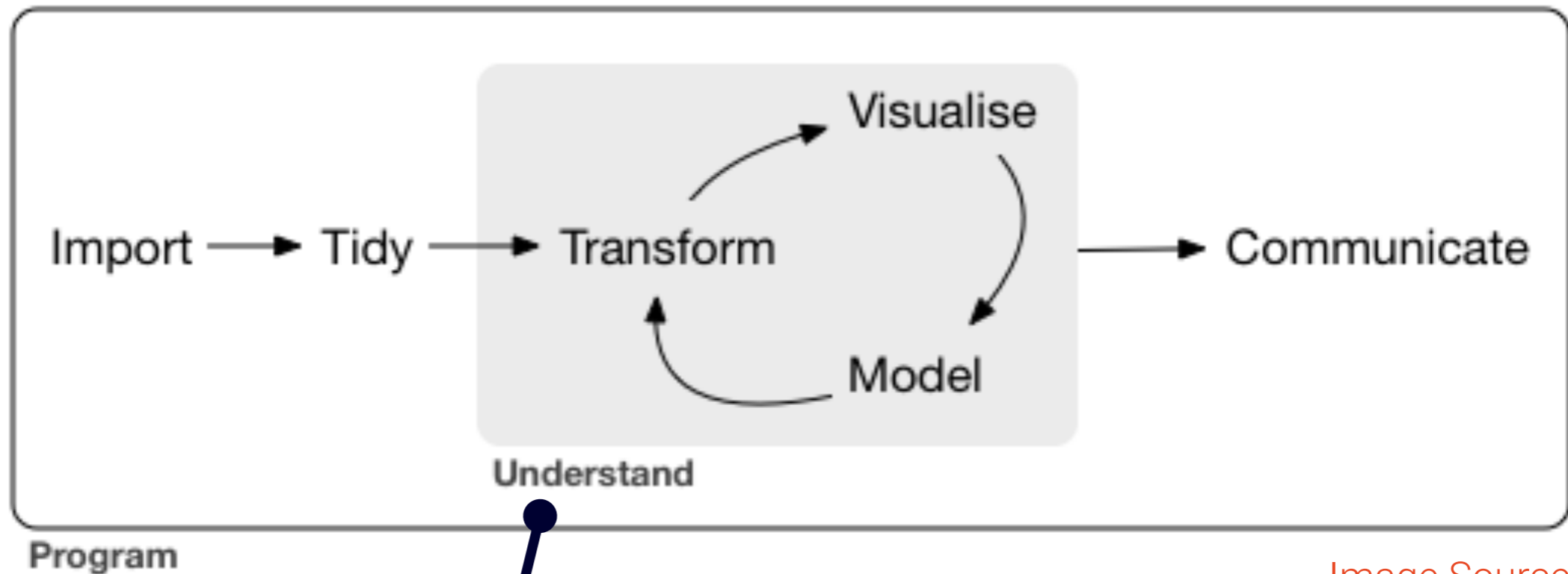
Meet & Greet

Name, Year, Major

Why Industry? Why Data Science?

Best part of summer break?

Agenda



[Image Source](#)

Focus is on **Understanding**

Topic Outline & Structure

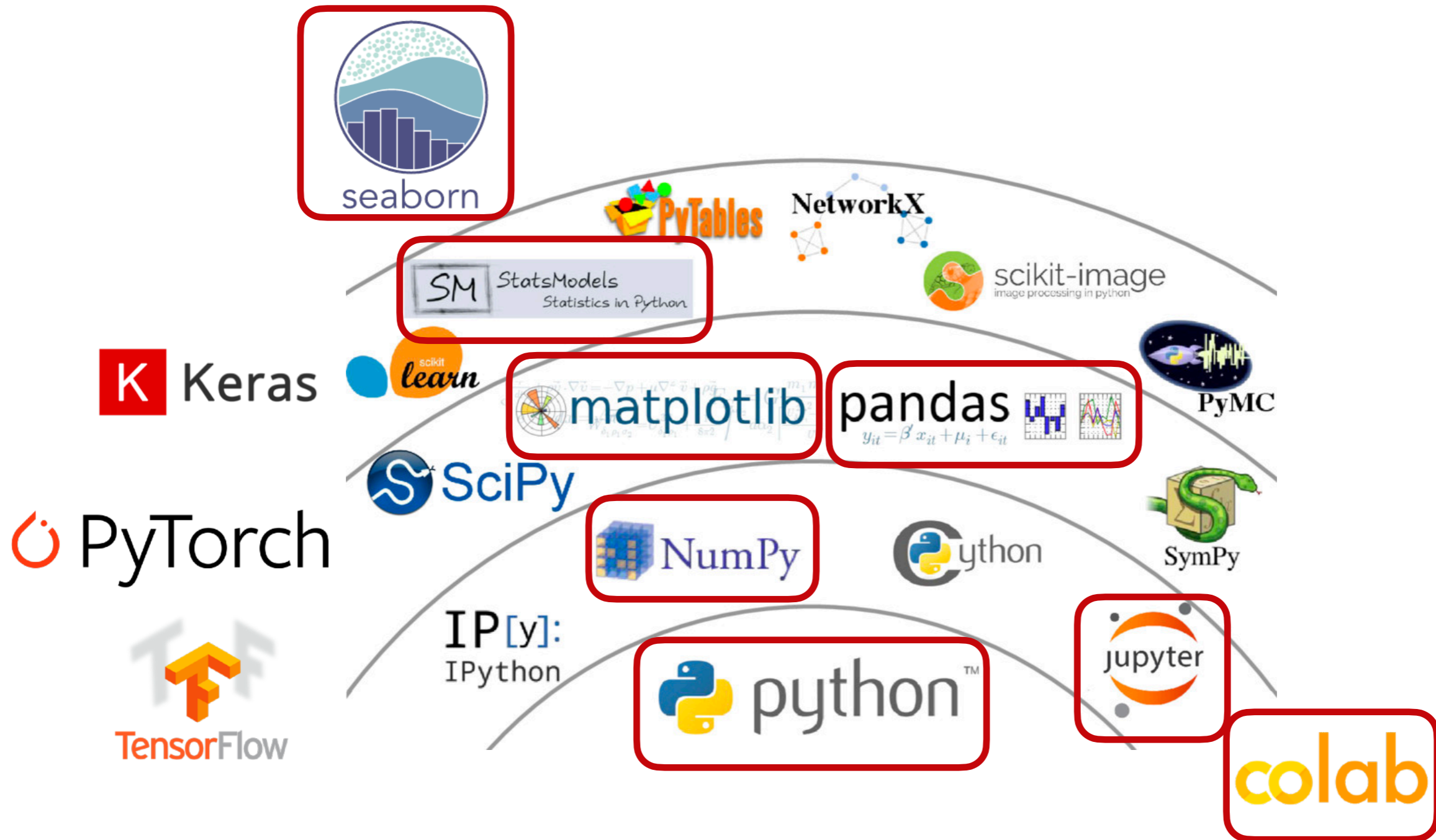
- **Topics**

- Data Wrangling
- Visualization
- Linear Regression
- Logistic Regression

- **Structure**

- ~10 - 20 Minutes Lecture
- 40 - 50 Minutes hands on notebook work within groups.

Workshop Software

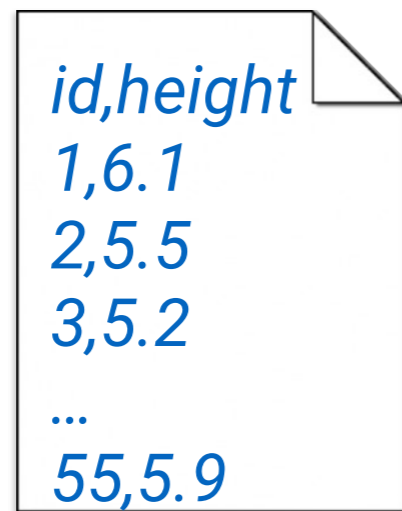


Source: [Jake VanderPlas Python's Data Science Stack \(JSM 2016\)](#)

Compute Options

tackling large-scale problems in a local context

Local
Import
and
Calculate



```
id,height  
1,6.1  
2,5.5  
3,5.2  
...  
55,5.9
```



Remote
Calculate
and
Retrieve



Server



Your Computer



Welcome To Colaboratory - Colab x +

colab.research.google.com/notebooks/welcome.ipynb#scrollTo=gJr_9dXGpJ05 Guest

Welcome To Colaboratory
File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

Connect Editing

Table of contents Code snippets Files X

- Introducing Colaboratory
- Getting Started**
- More Resources
- Machine Learning Examples: Seedbank
- + Section

Getting Started

The document you are reading is a [Jupyter notebook](#), hosted in Colaboratory. It is not a static page, but an interactive environment that lets you write and execute code in Python and other languages.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
seconds_in_a_day = 24 * 60 * 60
seconds_in_a_day
```

86400

To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command/Ctrl+Enter".

All cells modify the same global state, so variables that you define by executing a cell can be used in other cells:

```
[ ] seconds_in_a_week = 7 * seconds_in_a_day
seconds_in_a_week
```

604800

For more information about working with Colaboratory notebooks, see [Overview of Colaboratory](#).

Programming with Data

Definition:

Programming is the art of instructing a computer to do exactly what you say through an algorithm.

Definition:

Algorithms are a process or set of rules to be followed in calculations or other problem-solving operations.

Algorithm: Making Spaghetti

Ingredients

1 pound ground beef
1 large onion, chopped
2 garlic cloves, minced
1 (8-ounce) can tomato sauce
3 cups tomato juice
1 cup water
1 teaspoon salt
1 teaspoon sugar
2 to 3 teaspoons chili powder
1 teaspoon dried oregano
Dash of pepper
1 (7-ounce) package spaghetti, uncooked
Grated Parmesan cheese
Garnish: fresh Italian parsley sprigs

How to Make It

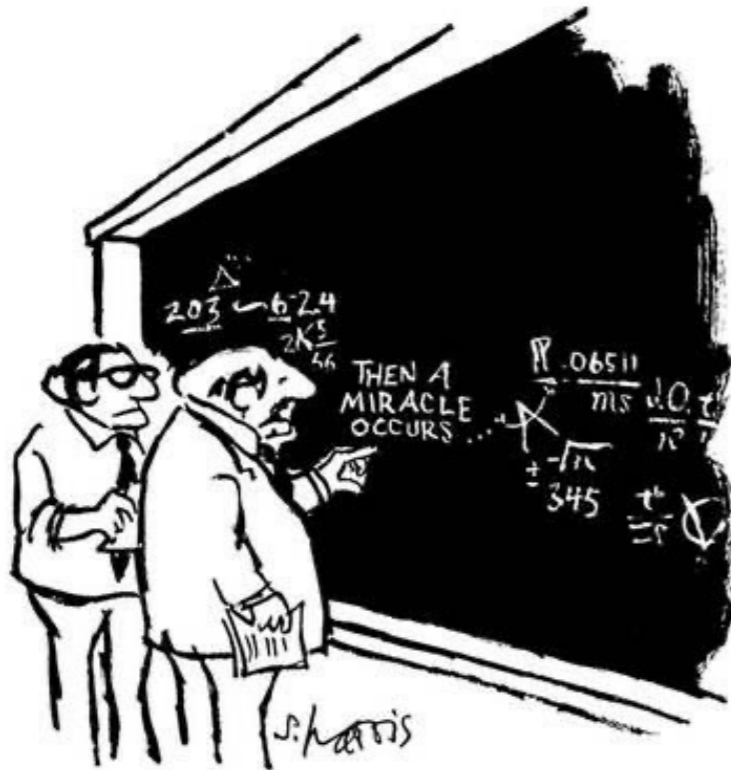
Step 1: Cook first 3 ingredients in a Dutch oven, stirring until beef crumbles and is no longer pink; drain well. Return beef mixture to pan. Stir in tomato sauce and next 8 ingredients; bring to a boil. Cover, reduce heat, and simmer, stirring often, 30 minutes.

Step 2: Add pasta; cover and simmer, stirring often, 20 minutes or until pasta is tender. Serve with cheese, and garnish, if desired.

[Recipe Source](#)

Bad Algorithms

... ambiguousness ...



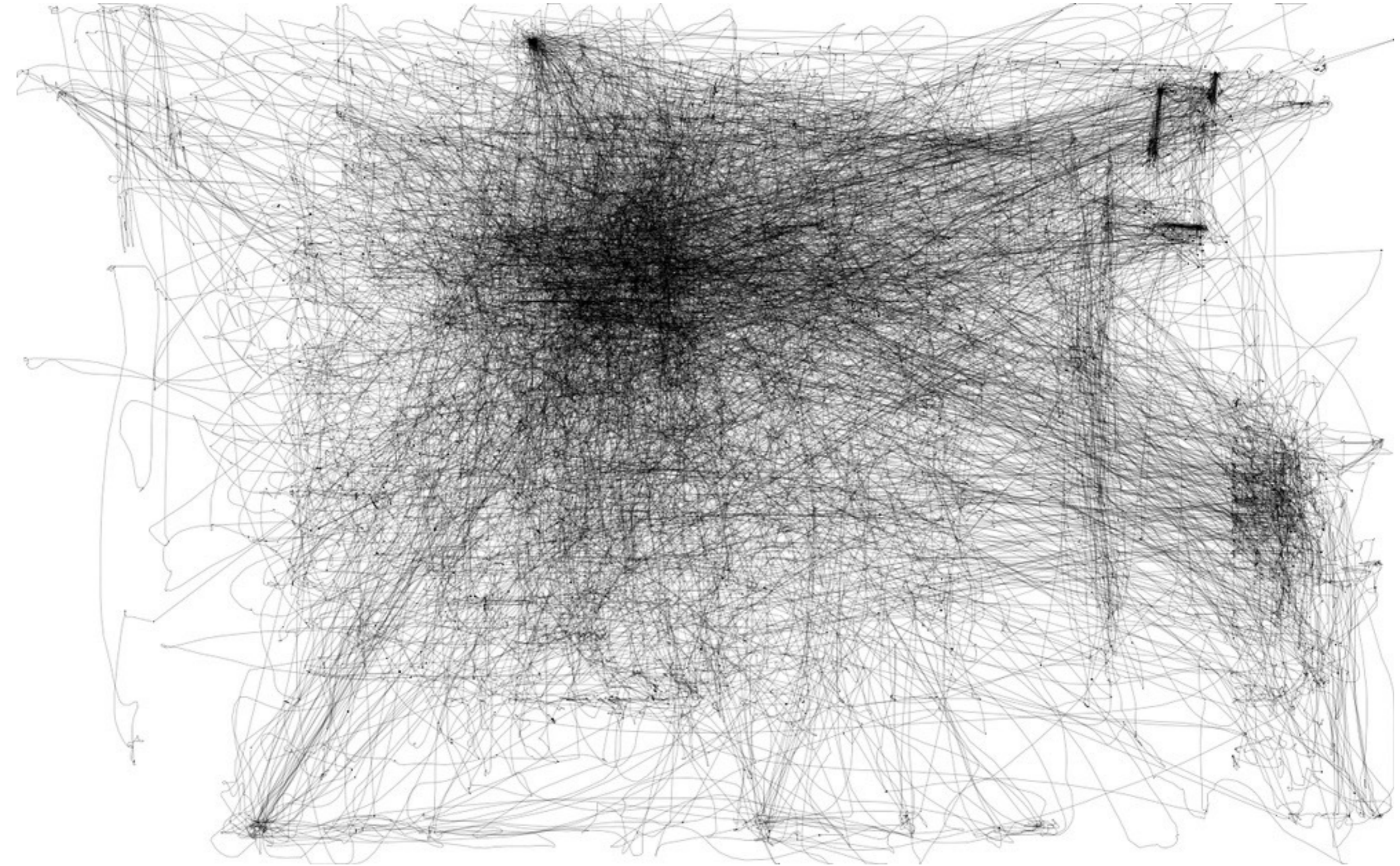
"I think you should be more explicit here in step two."

Comic Source

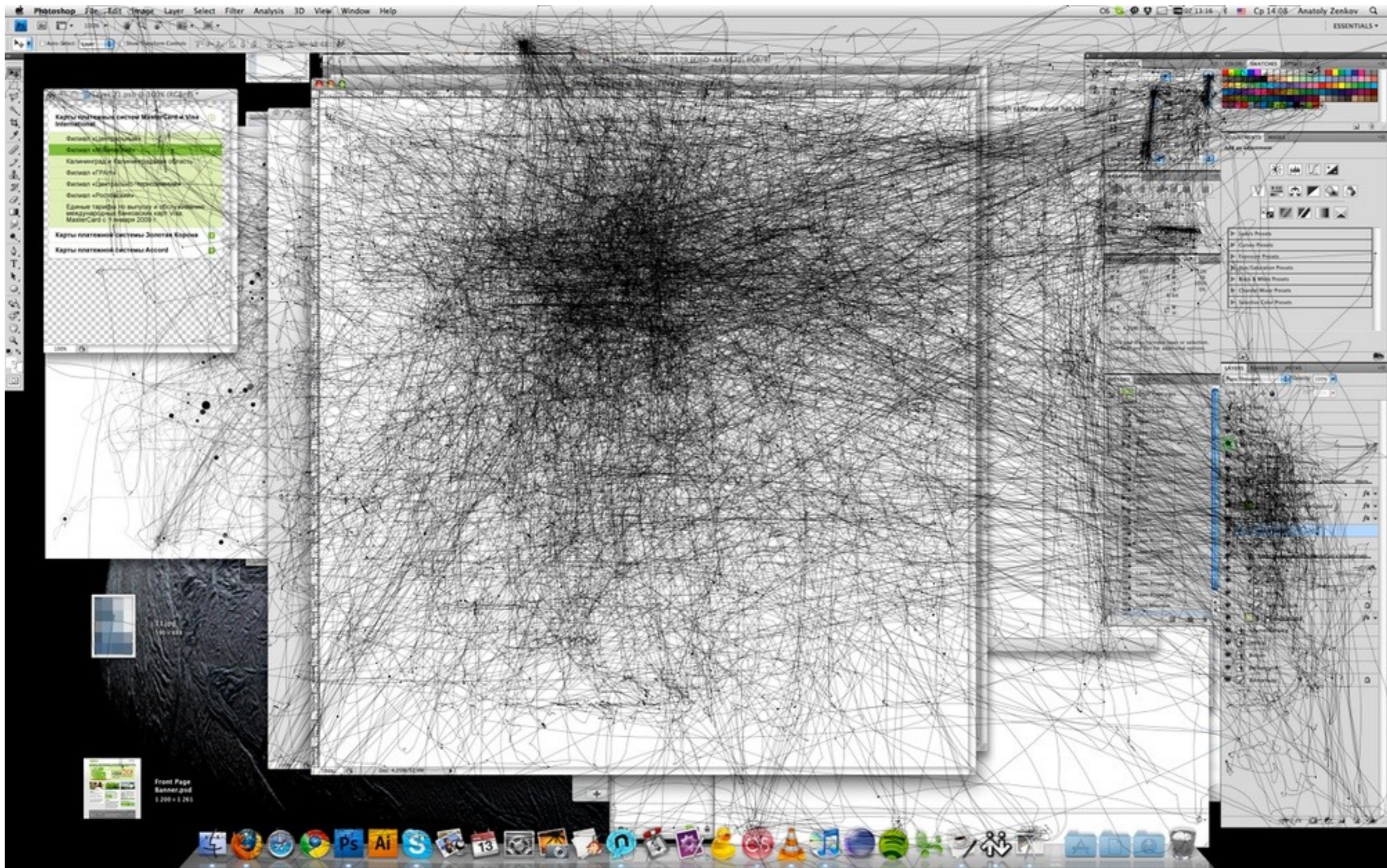


How Many Licks?

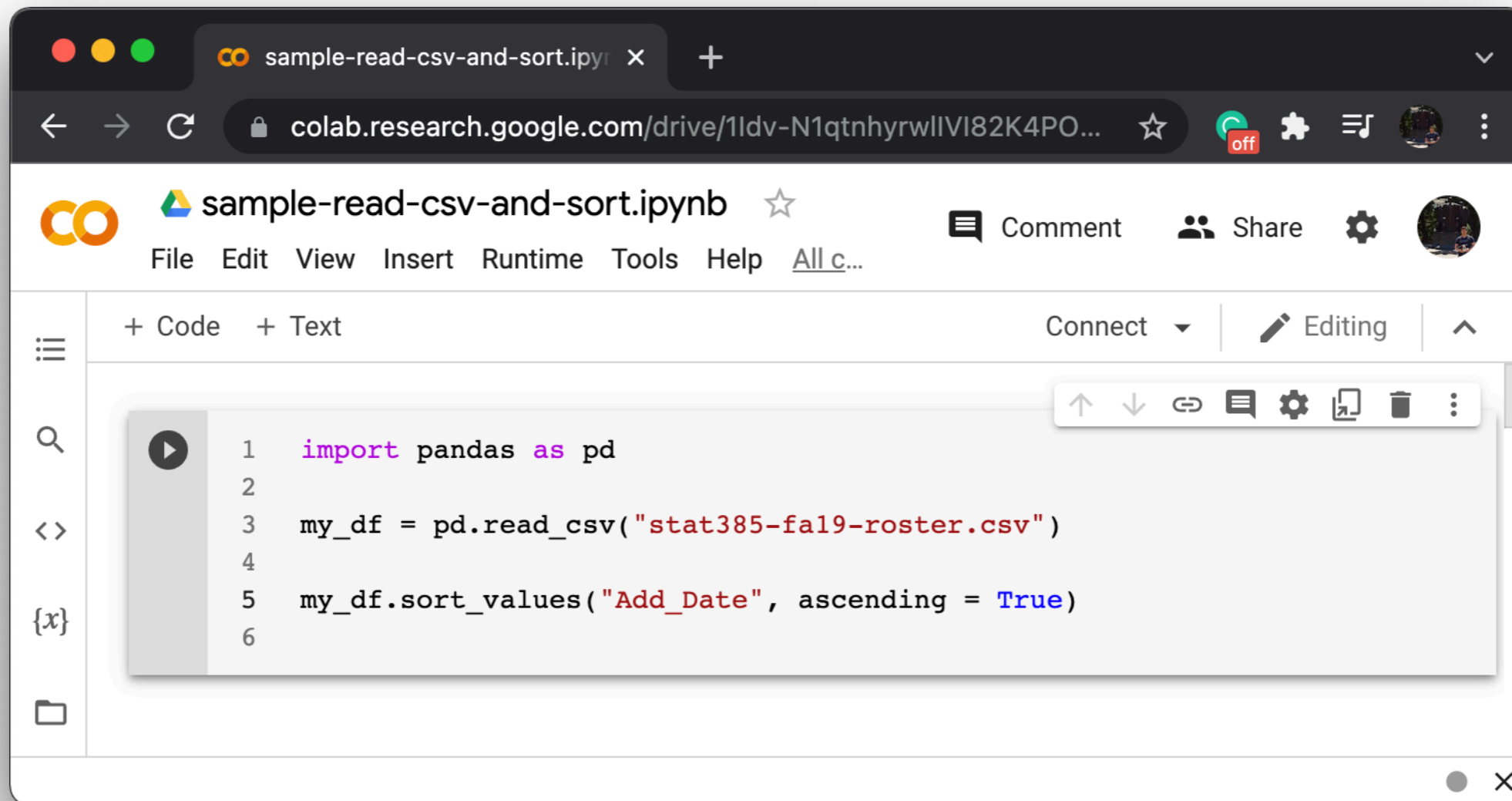
Is this programming?



Graphical User Interface (GUI)



Command Line Interface (CLI)

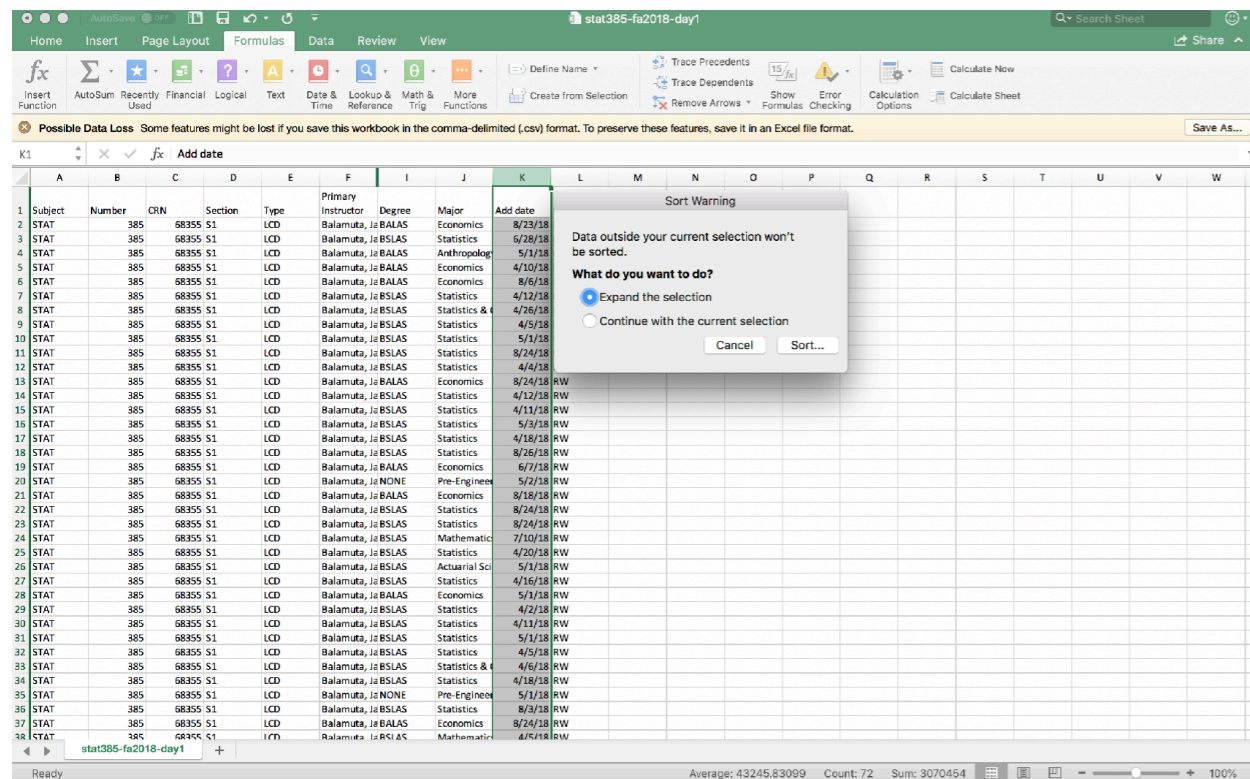


The image shows a screenshot of a Google Colab notebook interface. The browser address bar shows the URL `colab.research.google.com/drive/1ldv-N1qtnhyrwlIVi82K4PO...`. The notebook title is `sample-read-csv-and-sort.ipynb`. The code editor contains the following Python code:

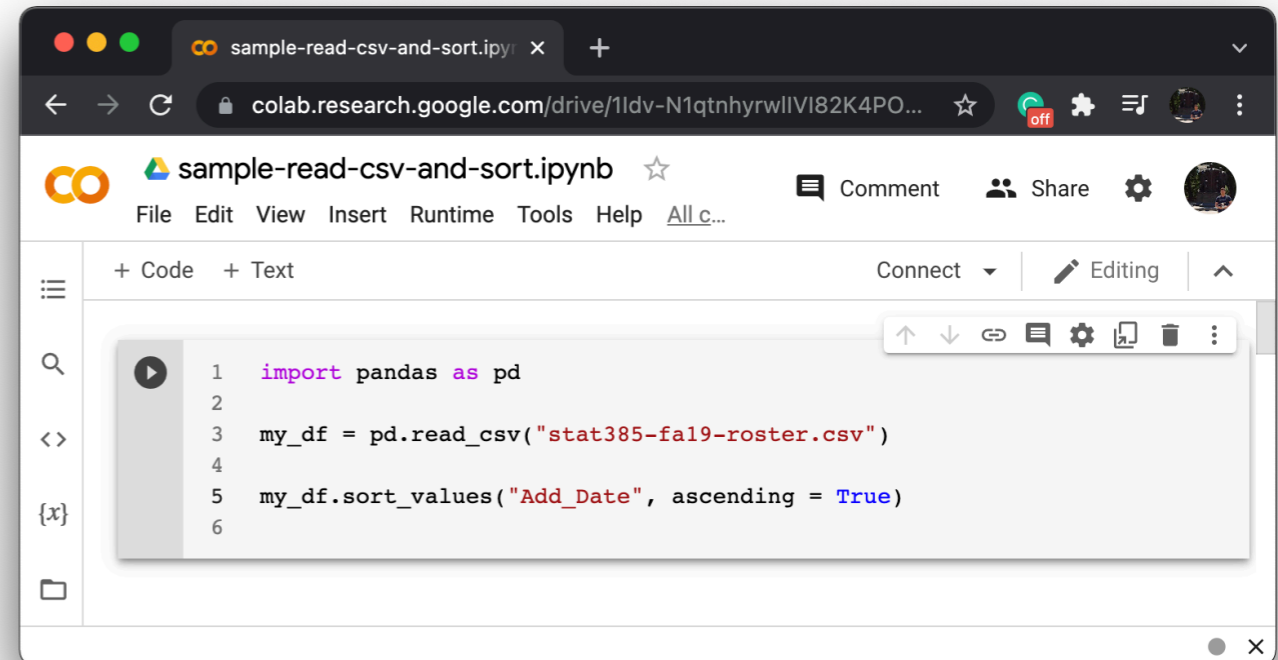
```
1 import pandas as pd
2
3 my_df = pd.read_csv("stat385-fa19-roster.csv")
4
5 my_df.sort_values("Add_Date", ascending = True)
6
```

The interface includes a menu bar with options like File, Edit, View, Insert, Runtime, Tools, and Help. There are also buttons for Comment, Share, and Editing. The code editor has a play button on the left and a toolbar on the right with icons for undo, redo, link, comment, settings, copy, and delete.

Comparison



GUI



CLI

This work is licensed under the
Creative Commons
Attribution-NonCommercial-
ShareAlike 4.0 International
License

