STARS
SUMMER TRAINING ACADEMY FOR RESEARCH SUCCESS
UC San Diego
Graduate Division

Mini Course 1: Data Analysis in Python

Thinking about data, reading and manipulating data frames, statistics and visualization

Ian Jackson

Python Primer

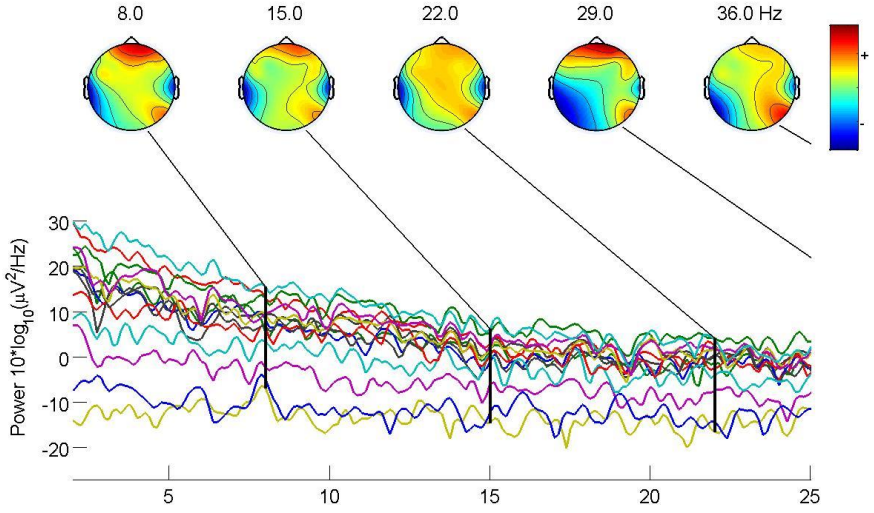
- These mini-courses will be taught in Python!
- Never worked with Python before?
  - Try out the free online tutorial from LearnPython!
  - UCSD Neuroscience's JC Gorman has great resources on her Github page
- Software requirements for this course:
  - Python3.8 or above
  - Jupyter Notebook
  - VSCode

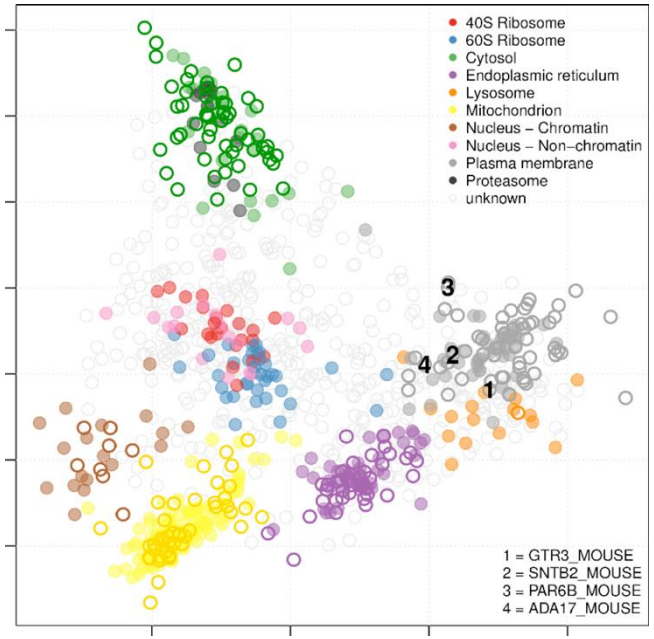
^ Link to NGP Python Bootcamp ^

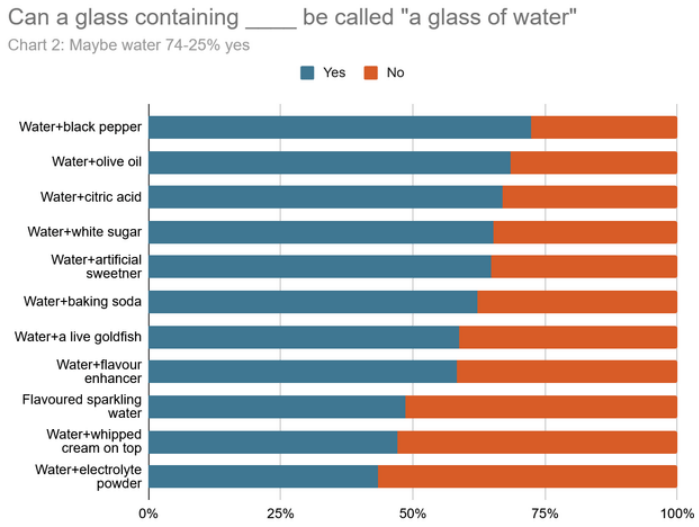
Now, a few questions for you!

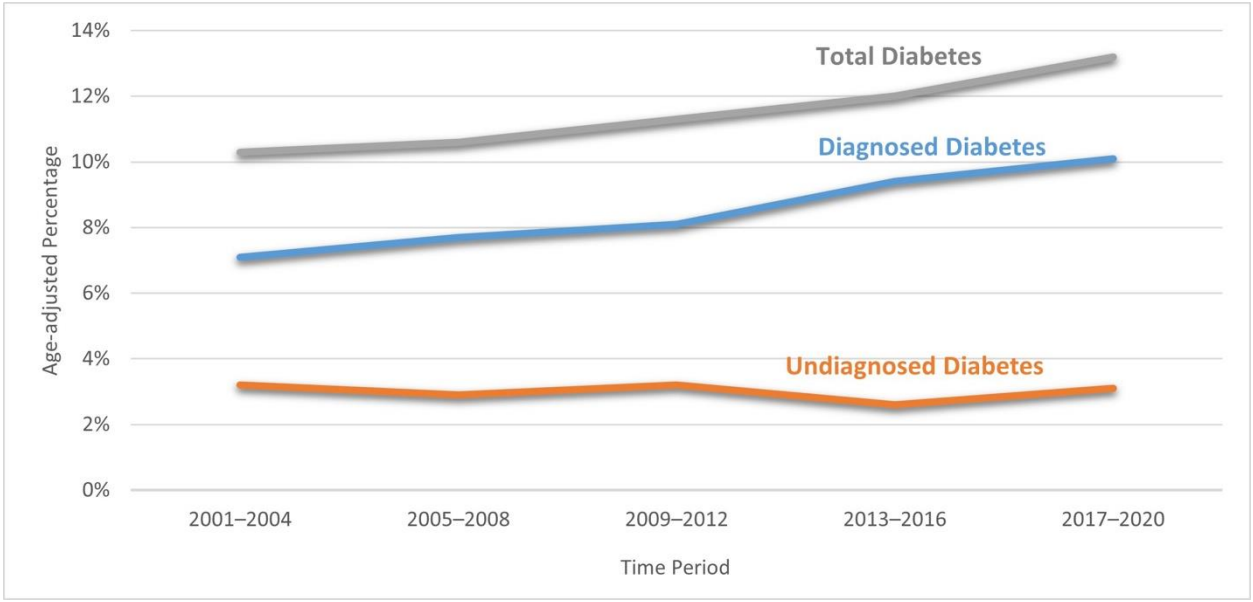
- What questions might you be working with this summer?
- What dataset might you be using?
- What is one thing you have learned so far from reading the literature?

So much data!


8.0
15.0
22.0
29.0
36.0 Hz
Power 10*log10(μV²/Hz)
30
20
10
0
-10
-20
5
10
15
20
25


40S Ribosome
60S Ribosome
Cytosol
Endoplasmic reticulum
Lysosome
Mitochondrion
Nucleus – Chromatin
Nucleus – Non–chromatin
Plasma membrane
Proteasome
unknown
1 = GTR3_MOUSE
2 = SNTB2_MOUSE
3 = PAR6B_MOUSE
4 = ADA17_MOUSE


Can a glass containing ____ be called "a glass of water"
Chart 2: Maybe water 74-25% yes
Yes
No
Water+black pepper
Water+olive oil
Water+citric acid
Water+white sugar
Water+artificial sweetner
Water+baking soda
Water+a live goldfish
Water+flavour enhancer
Flavoured sparkling water
Water+whipped cream on top
Water+electrolyte powder
0%
25%
50%
75%
100%


14%
12%
10%
8%
6%
4%
2%
0%
Age-adjusted Percentage
Total Diabetes
Diagnosed Diabetes
Undiagnosed Diabetes
2001–2004
2005–2008
2009–2012
2013–2016
2017–2020
Time Period

Generalized Steps of Data Analysis

- Establish a question (or just explore!)
- Find or create dataset
- Setup analysis environment
  - Programming language
  - File system setup
  - Identify packages
- Read and clean data
- Explore data to guide modeling
  - Run basic statistics
  - Visualization
- Build models of data to make
  - **Predictions:** estimates of new data points based on observations
  - **Inferences:** conclusions about the underlying distribution

What kind of model do I need to represent this data?

What is the overall ”shape” of each feature of the data?

What are my expectations of the relationships in the data?

It’s Jupyter Notebook Time!

- Go to: https://shorturl.at/A4MkG
- Clone the repository
  - Copy the URL (right)
  - Go to Terminal, navigate to a folder of your choice (e.g., Documents), and type in: **`git clone [enter URL]`**
- Open mc1.ipynb in VSCode and follow the instructions