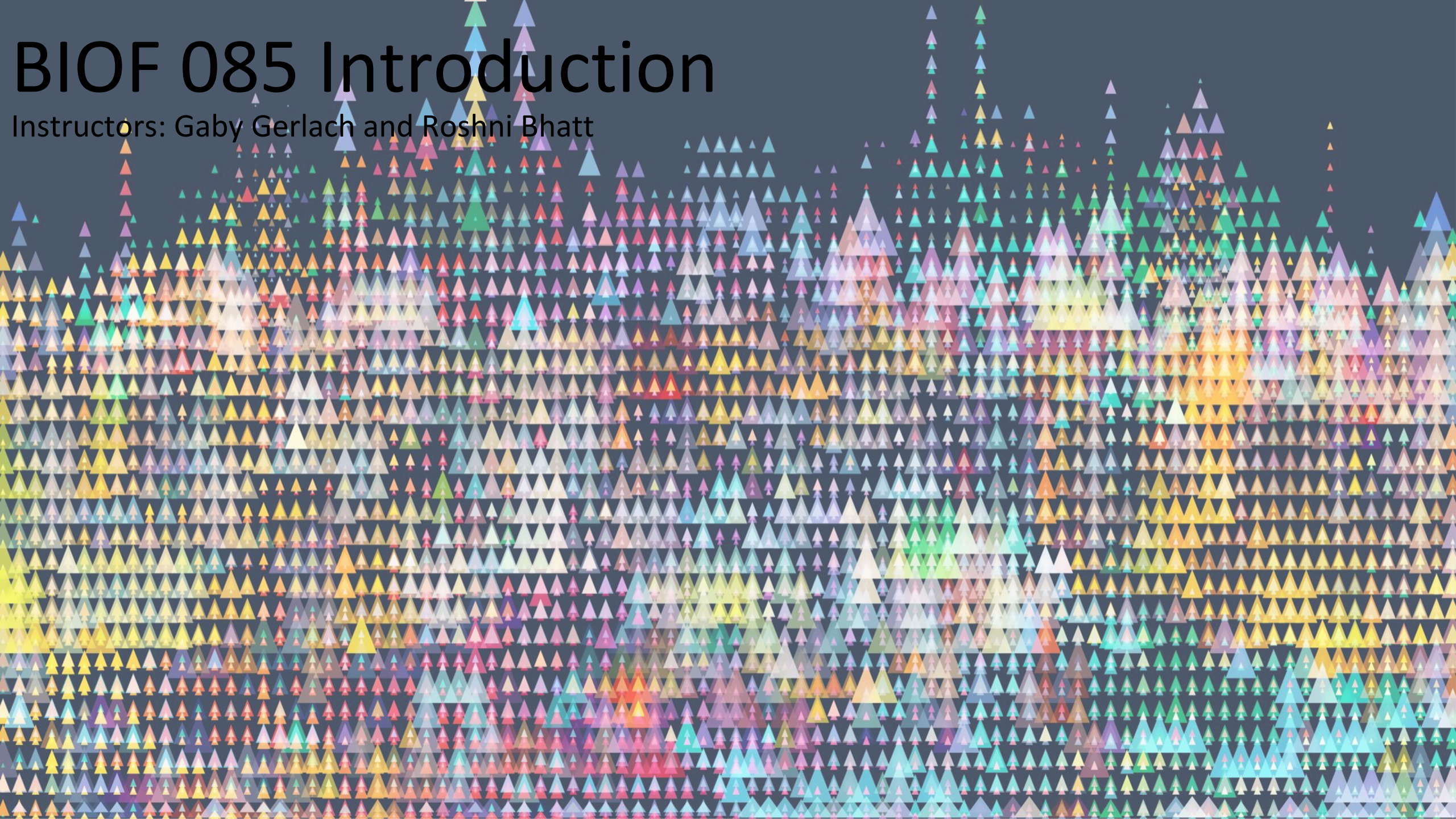
BIOF 085 Introduction

Instructors: Gaby Gerlach and Roshni Bhatt



Background

- PhD students in CMU-Pitt program in Computational Biology (CPCB)
- Use python to perform most of our analysis
- Interested mostly in Structural Biology
- Excited to be teaching this course!

Outline of Plan – Hybrid approach

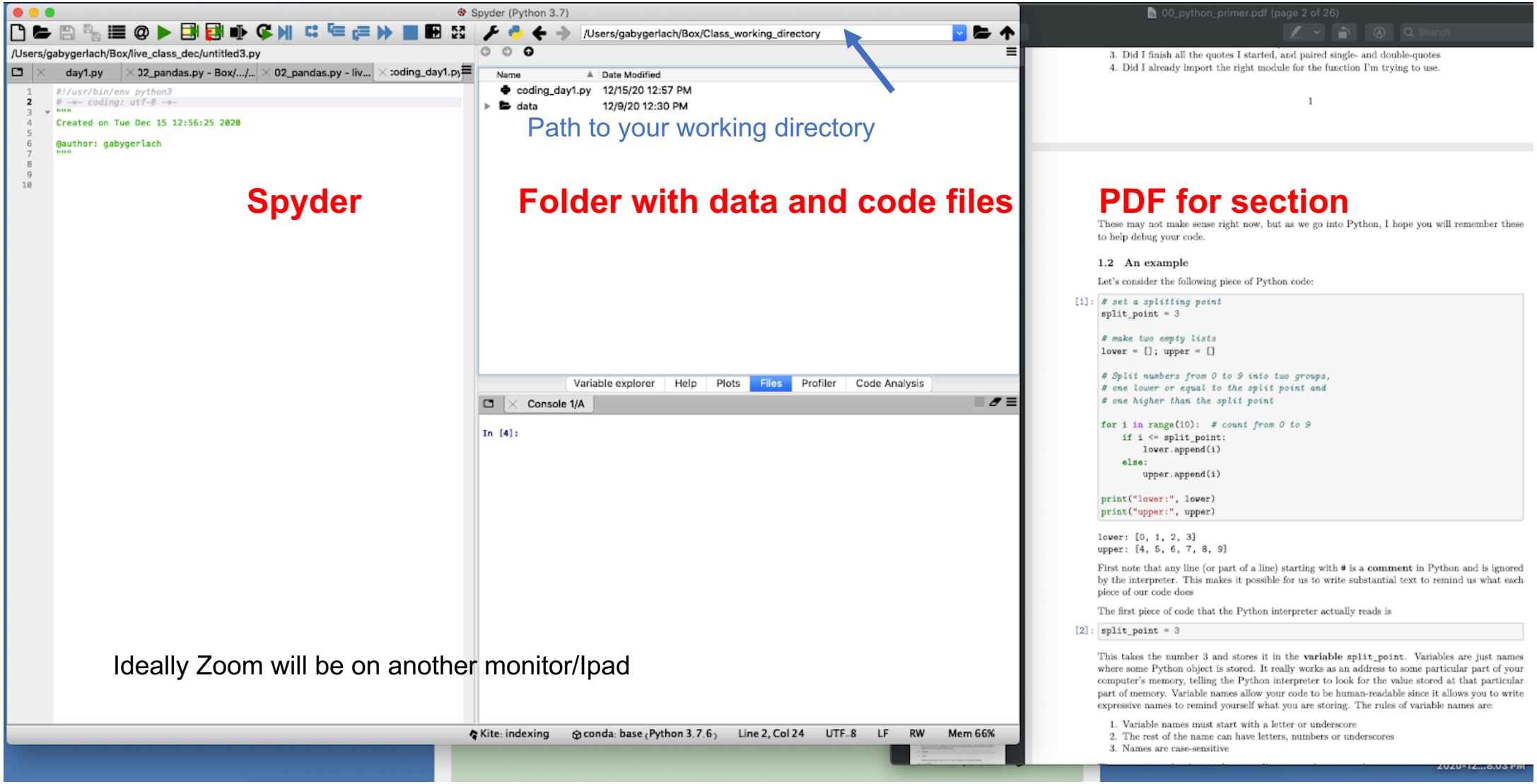
1. Face to face (about 50% of the time via Zoom)
    1. Lecturing and Q&A

2. Online Material (do this sequentially)
    1. Videos of screencasts
    2. Progress check assignments

Available material

- While we are live coding in either Spyder or Jupyter Notebooks
  - You are encouraged to follow along
- PDF versions of this are on Canvas and may be helpful to have open
- We will be pulling information from Jupyter Notebooks which have more explanation than we plan on saying

Effective work environment



Spyder

Folder with data and code files

PDF for section

Ideally Zoom will be on another monitor/Ipad

Some quick notes:

1. We are here to answer your questions, and provide the guidance the goal is for you to be coding in a guided manner not just listening to us talk
2. Feel free to interrupt us at any time
3. There is a lot of material in the notebooks we are not going to go through all of it – but it is all there for you too continue learning

Day 1

Time	Format	Topic	Instructor	Resource
9am – 11am	In-person via zoom	Why python? A python primer	Gaby	Intro slides, 00_python_primer
11 am - noon	In-person via zoom	Python tools for data science	Roshni	Python_objects intro 01_python_tools
1pm-2pm	In-person via zoom	Python tools for data science Data wrangling, cleaning, summarizing	Gaby	02_python_pandas
2pm -5pm	Asynchronous material	Data Munging	Gaby available via Slack	Day 1 independent work videos

Day 2

Time	Format	Topic	Instructor	Resource
9am-9:30	Q &A	Day 1 material	Gaby and Roshni	
9:30am –10:30am	In-person via zoom	Data Visualization/Stats	Roshni	Day 2 introduction
10:30am-12:30pm	Asynchronous material	Data Visualization	Roshni and Gaby available via Slack	Day 2 independent work videos
1:30pm - 2:30pm	In-person via zoom	Statistical Analysis pipeline	Roshni	04_stat_intro
2:30pm-4:00pm	Asynchronous material	Statistical Analysis using Python	Roshni and Gaby available via Slack	stats_python_practice
4:00pm-4:30pm	In-person via zoom	Q&A	Roshni	

Day 3

Time	Format	Topic	Instructor	Resource
9am –10am	In-person via zoom	Data analytics: Machine Learning	Gaby	ML slides, 06_python_learning
10am – noon	Asynchronous material	Data analytics: Machine Learning	Gaby and Roshni available via Slack	Day 3 independent work videos
1pm -2:00pm	In-person via zoom	String manipulation, RegEx, intro to bioinformatics	Gaby	07_python_appl
2:00-3:30pm	Asynchronous material	Use of skills learned in ML/bio application	Gaby and Roshni available via Slack	Day 3 Project - 08_genomics_project
4:00 – 4:30 pm	In-person via zoom	Explanation of tools, Q&A, miscellaneous	Gaby	07_python_appl

Adaptability

- There will likely be a range of abilities in this course
- It is likely there will be some people for who this is very quick and some for who it is a little slow
- We are doing our best to land in the middle, but this is just the reality

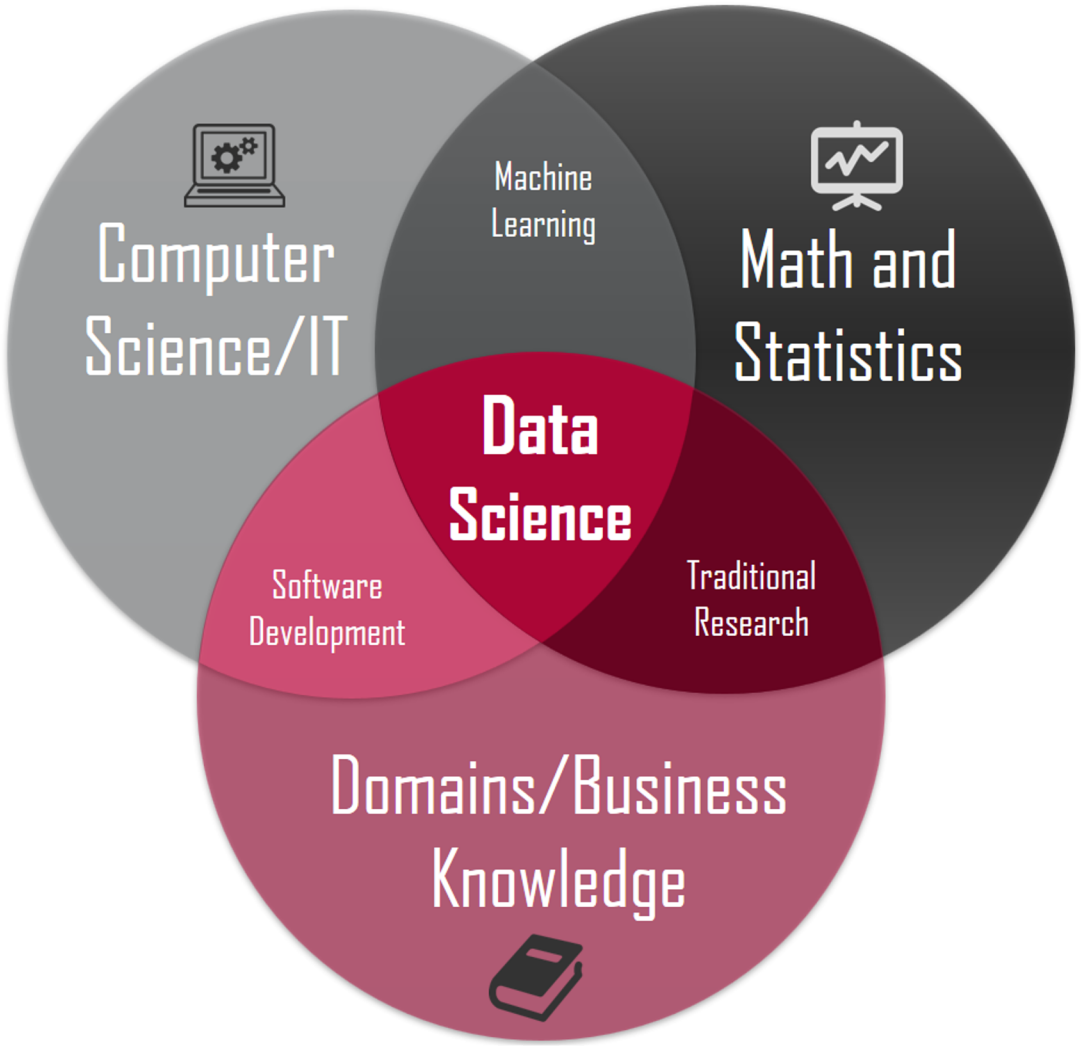
Slack

- This is the communication method will be available from ~8-6 each day
- If a question is asked outside of these hours there may be a delay in response
- Please use the #class-communications_mar2021 for questions during “lecture” so that everyone can benefit from your question
- Also feel free to interrupt us while we are talking
- During independent work you are encouraged to also use this chat, but are welcome to DM the instructors.

Now about you!

- Who are you?
- Why are you here?
- What is your background?
- Is there anything specific you came to learn about?
- What type of data are you interested in?

What is data science?

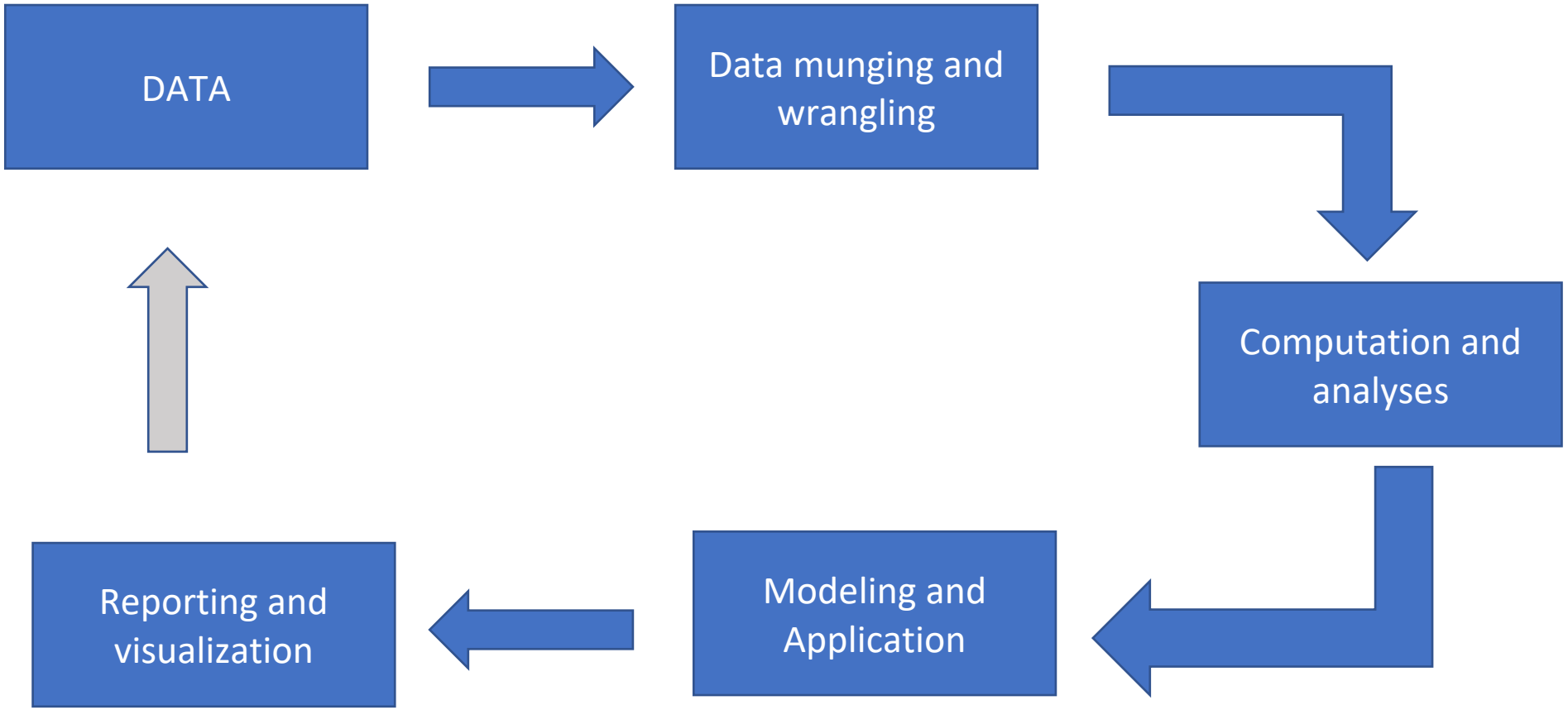


Data scientist?

- Statistician
- Computer scientist
- Database engineer
- Software engineer
- Mathematician

- I know good ones who are physics and geoscientists

What data science involves



What does it involve

1. Managing and cleaning data
2. Interest in exploring relationships between things, informed by domain knowledge
3. Statistical know-how
4. Computational skills
5. Tools

The goal here is the tools!

There are two major tools:

1. Python (https://www.python.org)
2. R (https://www.r-project.org)

Obviously here we are using Python.

Why Python?

- Popular general-purpose programming language
- Strong ecosystem through packages (230K+)
- Succinct, readable syntax
- Good balance between computational time and developer time
- Self-documenting
- Easier to integrate into production pipelines that already use python
- Increasingly strong Data science stack

Cons of Python

- Some places where the ecosystem is not rich enough
- More computer science-y, less statistical
- Version compatibility issues
- Poorer frameworks for display and dissemination of information
  - R tends to shine here

Python Data Science Stack

- To emulate Matlab
  - Numpy
  - Scipy
  - Matplotlib
- To emulate Maple
  - Sympy
- To add statistics/data science
  - Pandas
- Various visualization packages
  - Seaborn
  - plotly

Python Data Science Stack

- Philosophy has been to concentrate on a few large comprehensive packages
  - Statsmodels (statistics)
  - Scikit-learn (machine learning)
  - Pillow (image analysis)
  - Nltk (natural language processing)
  - Tensorflow & PyTorch (Deep learning)
  - PyMC3 (Bayesian learning)

Outline of Content

1. Python primer to get the basics of the language (day 1)
2. Pandas for data I/O, manipulation, cleaning and munging (day 1)
3. Using matplotlib and seaborn for data visualization (day 2)
4. Using pandas, scipy and statsmodels for statistics (day 2)
5. Using scikit-learn for basic machine learning (day 3)
6. Applications (day3)
    1. General examples
    2. High-level bioinformatics
    3. High-level string manipulation