GENERAL ASSEMBLY

DATA SCIENCE
COURSE OVERVIEW

AGENDA

- Course Philosophy
- What to Expect
- Course Tools
- Installfest!

DATA SCIENCE

COURSE PHILOSOPHY

TEACHING PHILOSOPHY

- Solve problems using **coding-oriented (**Python) techniques.
- Use **hands-on learning** alongside lecture.
- **Apply** concepts – leave each class with a new skill.
- Embrace what **diverse backgrounds** can bring.
- Success is **not a grade**.

COURSE CONTENT PHILOSOPHY

- Data science is not programming, mathematics, or statistics. It has its own **fundamental principles, workflow, and techniques**.
- Prefer **repetition and building upon fundamentals**.
- **Practice is necessary** to learn skills – pre-lesson materials, homework, and a course project are necessary for success.
- **Communication is key** – we want to hear your feedback!

HOW TO SUCCEED

- Be relentlessly curious – in exploring data AND materials!
- Be patient with yourself and keep trying. Effort > pre-knowledge.

- **Coding > watching videos and/or reading.**

- Study pre-lesson materials, do homework. Start the project early.
- Ask questions! Contact us on Slack at any time, even if we appear offline, and we will get back with you when we log in.
- Help classmates.

LOGISTICS

- Start and end class on time
- Missed classes
- Slack is preferred over email
- GitHub used for course content and homework
- Google+ community used for class recordings, discussions
- Office hours

THE DATA SCIENCE COMMUNITY

A community is people sharing with other people. Even if you are new, you have things to share!

- Write blog articles on your data science experiences.
- Put your data science projects on your website.
- Contribute to related open-source projects (e.g. on GitHub).
- Answer questions on Stack Overflow/Hacker News/etc.
- Give talks at local meetups.
- Get on Twitter and communicate with Python people.
- Go to local meetups.

http://datascience.la/

DATA SCIENCE

WHAT TO EXPECT

WHAT TO EXPECT

Core skills practiced daily:

- Python
- Python Data Science Libraries – matplotlib, scikit-learn, pandas
- Understanding, Selecting, and Validating Models

Survey of:

- Specific models – neural networks, clustering, dim. reduction, etc.
- Mathematical techniques/foundations
- Additional programming techniques and libraries

PRELIMINARY SYLLABUS

1. Intro
2. Command-Line Tools
3. Python
4. Git & Python Problem Solving
5. APIs & Web Scraping
6. Statistics & k-NN from Scratch
7. NumPy & Linear Regression from Scratch
8. Data Exploration with Pandas
9. scikit-learn & k-Means Clustering
10. Linear Regression II & Data Distributions
11. Logistic Regression & AUC
12. Neural Networks I
13. Complete Data Science Example
14. Image Data: Neural Networks II
15. Feature Selection & Dimensionality Reduction
16. Text Data: Natural Language Processing
17. Text Data: Naive Bayes
18. Recommendation Systems from Scratch
19. Deep Learning: Neural Networks III
20. Decision Trees & Random Forests
21. Big Data & Course Review
22. Project Presentations

NEW TO CODING?

- Expect to spend significant additional time learning Python.
- Take advantage of office hours and Slack.
- As you read, type in and execute the code. Do not copy and paste.
- Solve daily programming problems:
  - https://brilliant.org/, http://coderbyte.com/
  - https://www.hackerrank.com/, https://projecteuler.net/

NEW TO CODING?

- If you are stuck, start with code you know works. Challenge yourself to add small things to it.
- Students without coding experience often struggle translating ideas into code, which may mean less data analysis is ultimately done.
- At the end of the course, you will likely be a better programmer but likely will still feel you have much more to learn. However, this is a natural part of learning how to program – even experts frequently feel they could have done better.

DATA SCIENCE

COURSE TOOLS

WHAT IS ANACONDA?

Anaconda – package manager for scientific software. Includes:

- **Python 3.4.3** – latest version of the Python interpreter
- **IPython** – improved interactive Python shell
- **Spyder** – data science Python IDE
- **Jupyter** – "lab notebook" for coding (formerly IPython Notebook)

Recommended Supplemental Books (free online)

"Learn Python the Hard Way" http://learnpythonthehardway.org/book/ ← Note: Python 2

"Dive into Python 3" http://www.diveintopython3.net/

WHAT IS ANACONDA?

Anaconda – package manager for scientific software. We will use:

- **conda** – python package manager, for installing new packages
- **numpy** – ndarray, multi-dimensional array processing
- **pandas** – Series and DataFrame
- **matplotlib** –plotting, in the style of MATLAB
- **nltk** –Natural Language ToolKit
- **scikit-learn** – tools for modeling

Also:
scipy,
statsmodels,
theanos/keras,
(and more)

Recommended Supplemental Book

"Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython"
by Wes McKinney, the creator of Pandas.

GIT AND GITHUB

Git is a version control system widely used in industry.

- Retrieve latest versions of course materials.
- View exact changes made to course materials.

GitHub is a web-based Git repository hosting service.

Recommended Supplemental Book (free online)
"Pro Git"
http://git-scm.com/book/en/v2

QUESTIONS?

INSTALLFEST!

1. Install latest **Anaconda – PYTHON 3.4!** http://continuum.io/downloads
2. Install **Sublime Text 3** (or other text editor): http://www.sublimetext.com/3
3. Install **PyCharm Community Edition** (or other IDE): https://www.jetbrains.com/pycharm/download/
4. Install **Git**: http://git-scm.com/book/en/v2/Getting-Started-Installing-Git
5. Create a **GitHub** account: https://github.com/
6. Using Windows? Install latest **Gow** (GNU command-line tools):

   https://github.com/bmatzelle/gow/releases