DATA SCIENCE COURSE OVERVIEW

Adam Jones, PhD

Lead Instructor

AGENDA

‣ Course Philosophy

‣ What to Expect

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*.
  ‣ *Though, a combination can be most ideal.

HOW TO SUCCEED

‣ 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.

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.

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

‣ Feature Engineering

Survey of:

‣ Specific models – neural networks, clustering, dim. reduction, etc.

‣ Mathematical techniques/foundations

‣ Additional programming techniques and libraries
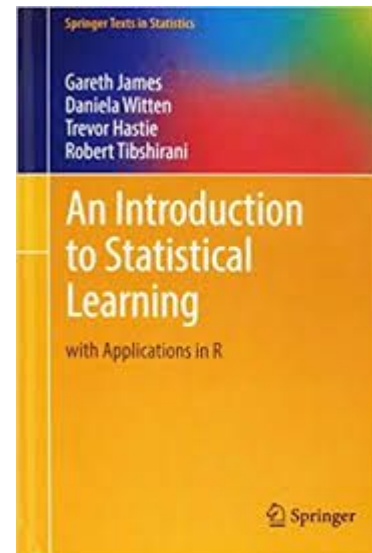
SYLLABUS

The **_class repository_** will be your 'one-stop-shop' for important class info:

https://git.generalassemb.ly/meccaLeccaHi/dat-sf-60

Let's check it out…

TEXTBOOKS

You'll also be assigned optional readings from this text:

An Introduction to Statistical Learning
Free copy here:
http://faculty.marshall.usc.edu/gareth-james/ISL/

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 problems:

> http://coderbyte.com/
> https://www.hackerrank.com/
> https://projecteuler.net/ (math-oriented coding)
> https://brilliant.org/ (for math refreshers)

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. This is a natural part of learning how to program – even experts frequently feel they could have done better.

OLD TO CODING?

‣ No problem!

‣ We want to maximize the value of this class for *every student*.

‣ Communicate with us regarding your areas of interest and we can keep you coding 'till the cows come home.

DATA SCIENCE
COURSE TOOLS

WHAT IS ANACONDA?

Anaconda – Python distribution for scientific software. Includes:
- **conda** – package manager, virtual environments
- **python** – latest version of the Python interpreter
- **ipython** – improved interactive Python shell
- **Spyder** – data science Python IDE
- **Jupyter** – "lab notebook" for coding (formerly IPython Notebook)

WHAT IS ANACONDA?

Anaconda – Python distribution for scientific software. Includes:

- ‣ **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
- ‣ **Many more…**

⦿ GIT AND GITHUB

‣ Git is a version control system widely used in open source and industry.

‣ Retrieve latest versions of course materials.

‣ View exact changes made to course materials.

‣ GitHub is a web-based Git repository hosting service.

‣ Our class repo is git.generalassemb.ly/meccaLeccaHi/dat-sf-60

COURSE TOOLS

⧉ SLACK

‣ We'll use Slack for our main class communications platform.

‣ Use it anytime you want to share information about the course, discuss lessons, and/or submit projects.

‣ Our Slack homepage is sf-dat-60.slack.com

DATA SCIENCE

QUESTIONS???