GENERAL ASSEMBLY

WELCOME TO

DAT11: DATA SCIENCE

Course Instructor: Paul Goodall

Astrophysicist, Data Scientist, Explorer

AGENDA

AGENDA

- AGENDA (infinite loop?)
- Introductions **– all!**
- About this course
  - Expectations – Course Structure
  - Expectations – Lesson Structure
  - Expectations - What the course IS
  - Expectations - What the course IS NOT
  - Expectations – Adult learning environment
- Icebreaker (15-mins) **– all!**
- ROE – Rules of Engagement **- all**
- Data Science lesson 1

ABOUT ME

- Welcome to Data Science 11 Sydney!
- Here's a bit about me:

Name | Background | Fun Fact

ABOUT ME

- Welcome to Data Science 11 Sydney!
- Here's a bit about James:

Name | Background | Fun Fact

ABOUT CLASS DAT11

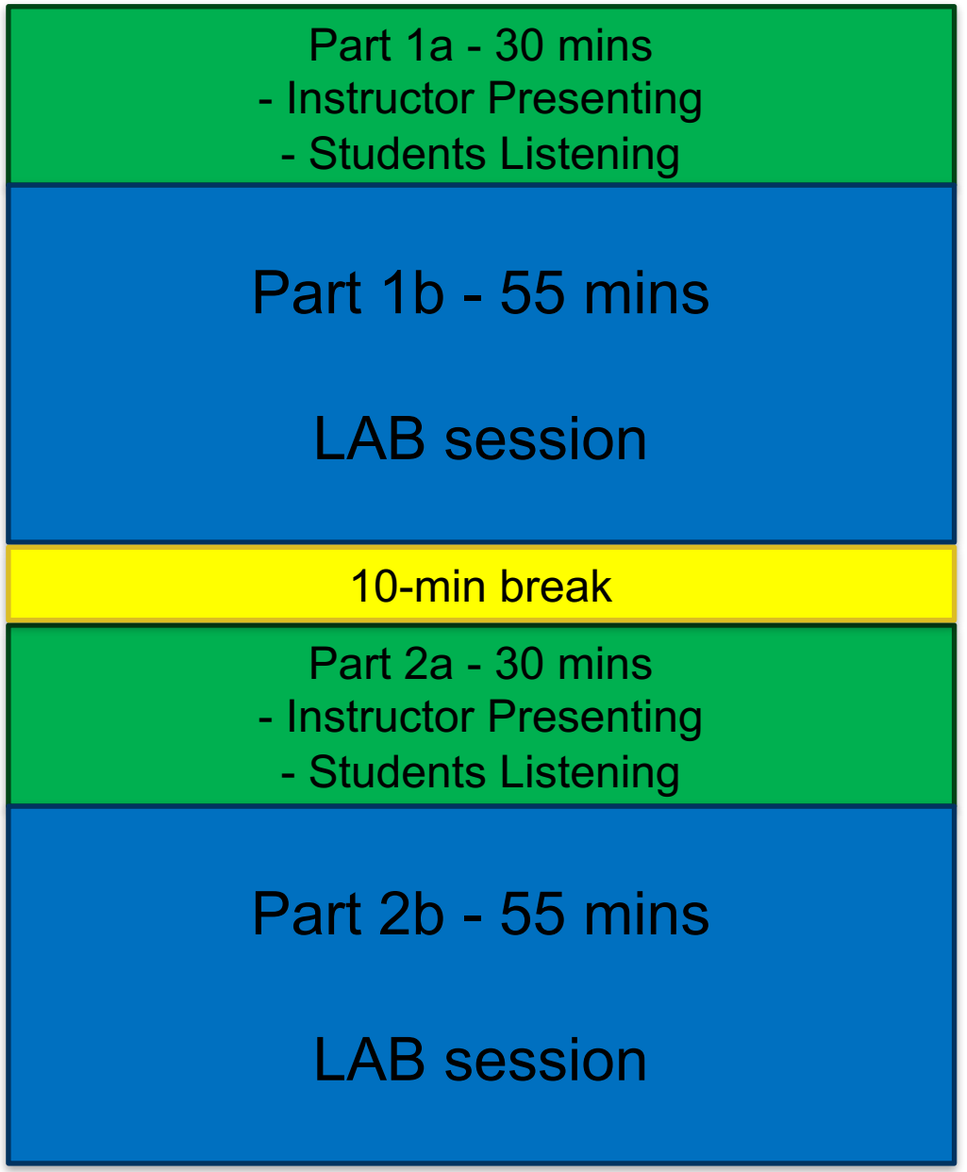
Round Robin for all:

Name | Background | Fun Fact

ABOUT THE COURSE

Structure of a Typical Lesson

- Statement of Learning Objectives
  - + overview
- Motivation
  - examples of practical problems
- Theory
- Lab
  - code walk-through
  - exercises
- Discussion
- Homework
  - projects (or project checkpoints), planning
  - reading



What the course IS

- **A whirlwind tour of the Data Science toolkit**
  a way for professionals who haven't done Data Science before, to get some exposure to what Data Science is all about

- **A Platform**
  an enabler for professionals to kick-start a successful career in Data Science

- For beginners / inexperienced technical professionals

What the course IS NOT

- Everything you need to instantly equip you as a world-class Data Scientist….

  This cannot be done in 60 hours!

  Takes many year experience of DOING Data Science

- NOT an advanced course – please be patient if others are less experienced and require more time

An efficient adult learning environment

- Learn by doing!
- Leverage experience of Course Instructor and TA
- Leverage experience of other students in class
- Core values:
  - Professionalism
  - Respect
  - Fun
  - Honesty
  - Motivation, Enthusiasm & Contribution
  - Anything else ??

Ice Breaker!

Rules of Engagement

GENERAL ASSEMBLY

DATA SCIENCE

Lesson 1, Week 1

Introduction to Data Science

What is data science and how do we apply it?

Learning Objectives

Already done in Intro!

- Review Pre-work
- ~~Describe the roles and components of a successful learning environment~~
- Define data science and the data science workflow
- Apply the data science workflow in a practical example
- Set up / test your development environment
- Get started with Python!

PRE-WORK

Pre-work Review

- Install **Anaconda** with Python 3.6 onto your computer.
  - Check that it runs by opening the Navigator application and launching a jupyter notebook from the navigator interface.

- Download and install a Github or Git client
  - Sign up for a public Github account.
  - Send your Github username to your course contacts.

- New platform – class server dat11:
  - http://www.paulgoodall.tech:8000/
  - http://128.199.71.237:8000/hub/login

New solution!

Everything already set up for you! ☺

- Introduce yourself on the course Slack channel
  - Put a picture of yourself in your profile.

Pre-work Review - cont'd

- Read any 2 chapters of the Data Science Handbook (www.thedatasciencehandbook.com/)
- Read the first 2 chapters of Introduction to Statistical Learning (http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf)
- Complete the Intro to Python for Data Science course on DataCamp (https://www.datacamp.com/getting-started?step=2&track=python)
  - We will expect to see the certificate.

DATA SCIENCE BASELINE

Initial Thoughts on a whiteboard:

What does “Data Science” mean to you?

Your First Assignment is: Yourself!

- What would you like to get out of this course?
- Are you using analytics in your work right now?
- What projects have you worked on?
- What projects are you planning?
- Are you using Python?
- Are you using any other data science tools?
- Is there a special topic or application area you would like us to cover in the lessons?

alteryx

Post your answers to our Slack channel: http://dat11syd.slack.com/ #general (meggan@ga.co)

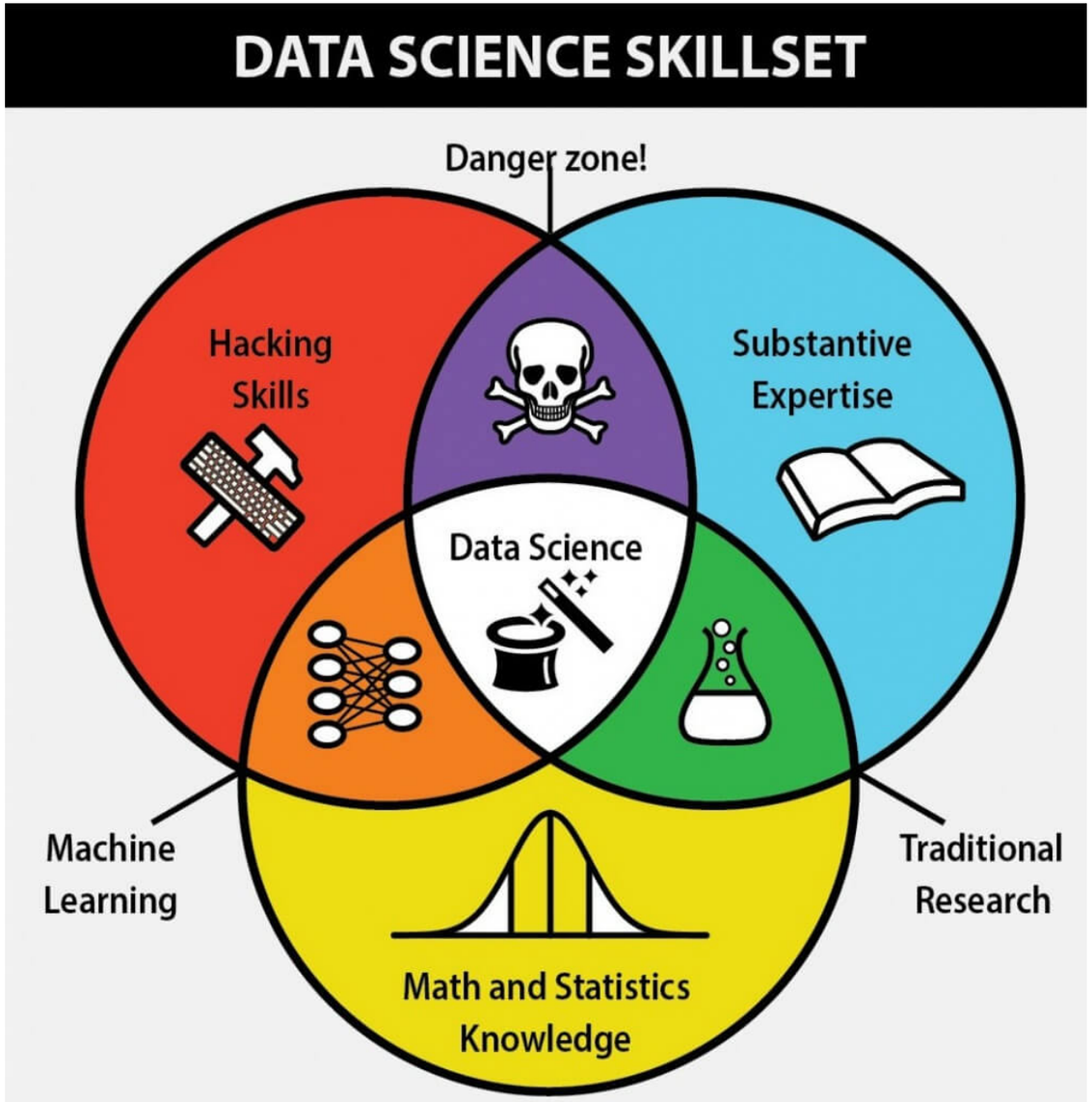
WHAT IS DATA SCIENCE?

What is Data Science?

Million-$ Question.

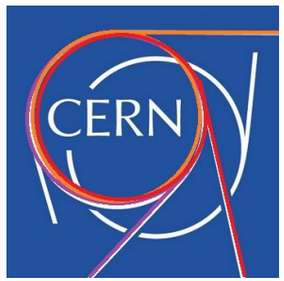
<u>GA's summary</u>:

- A set of tools and techniques for analysing data
- Interdisciplinary problem-solving
- Application of scientific techniques to practical problems



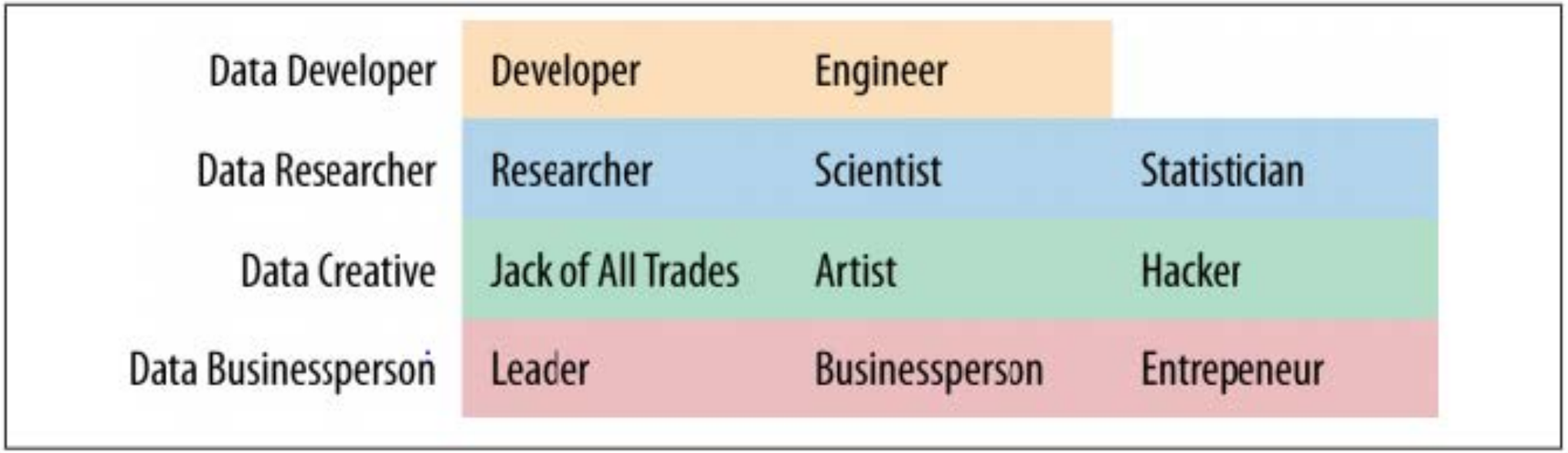
Source: Berkeleysciencereview

Who Uses Data Science?

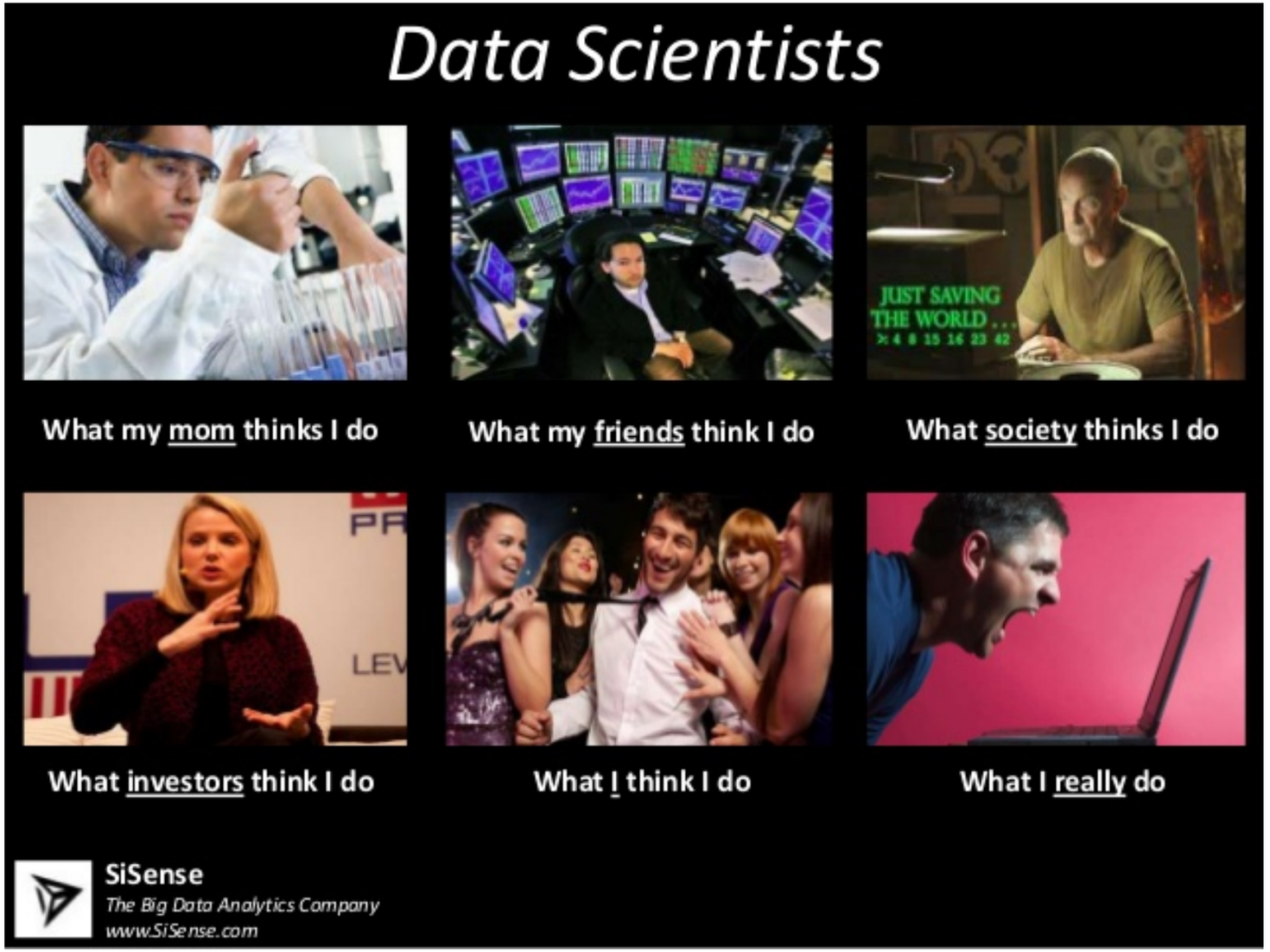


- ***Other examples?***

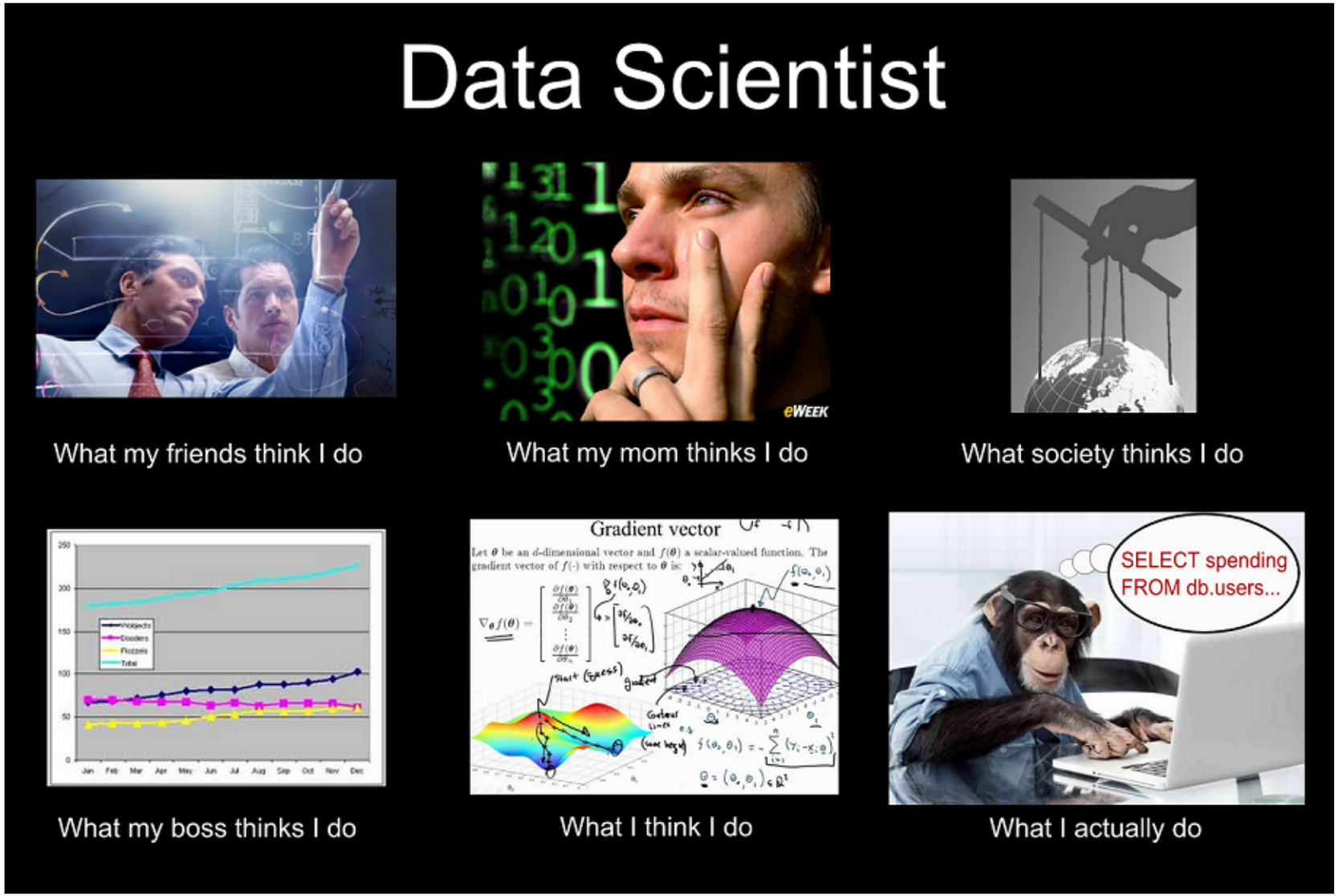
What Are the Roles in Data Science?



What do Data Scientists do?



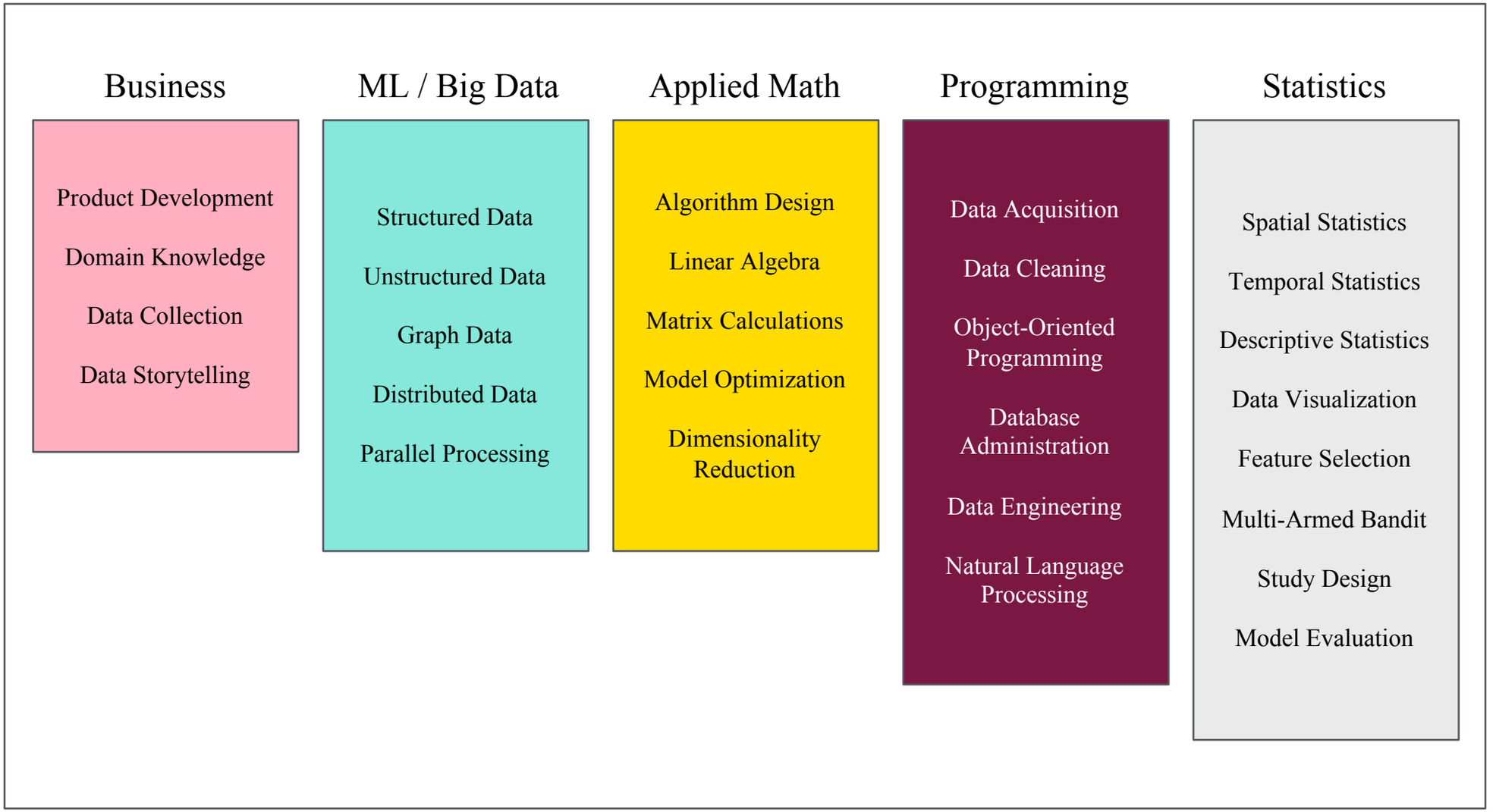
What do Data Scientists do?



Where Do Data Scientists Come From?
And what are their typical strengths?

	Hacking Skills	Math & Stats	Substantive Expertise	Methodology	Abstraction	Communication
Data Science program graduates						
Scientists (especially physics)						
Statisticians						
Developers						
Business Analysts						

Data Science Skill Sets



THE DATA SCIENCE WORKFLOW

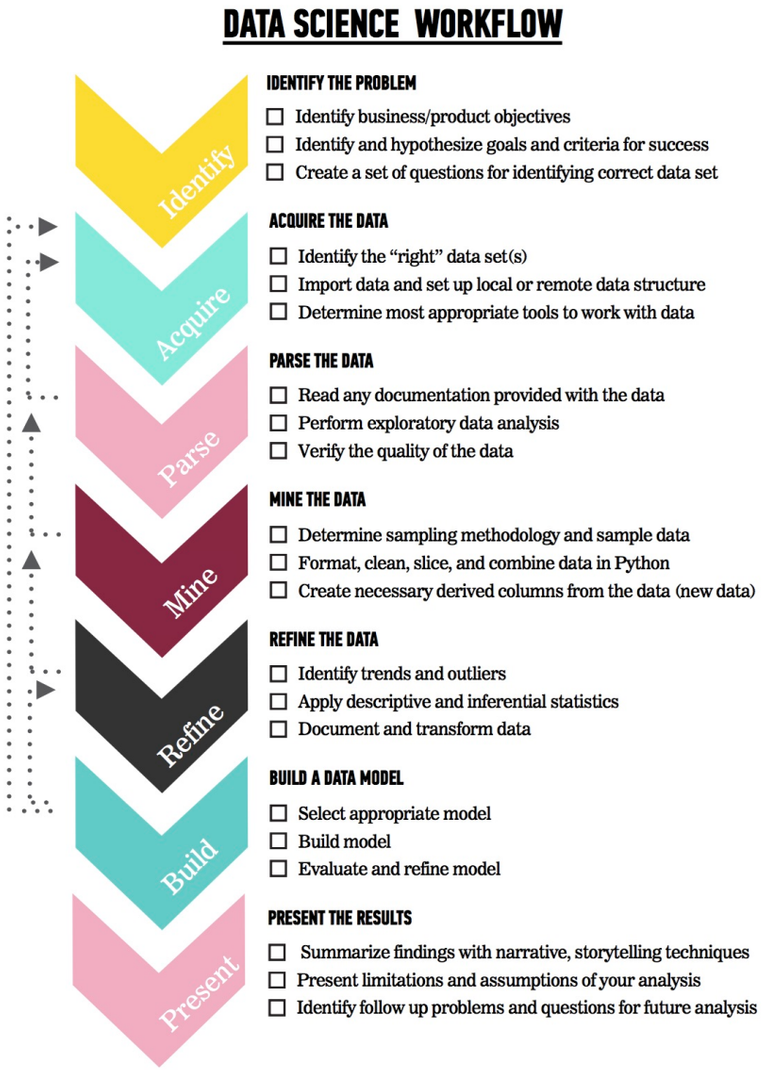
Features of the Data Science Workflow

- A methodology for doing Data Science
- Similar to the scientific method
- Helps produce *reliable* and *reproducible* results
  - *Reliable*: Accurate findings
  - *Reproducible*: Others can follow your steps and get the same results

Overview of the Data Science Workflow

The steps:

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results



Overview of the Data Science Workflow

IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

Overview of the Data Science Workflow

ACQUIRE THE DATA

- ☐ Identify the “right” data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

Overview of the Data Science Workflow

PARSE THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

Overview of the Data Science Workflow

MINE THE DATA

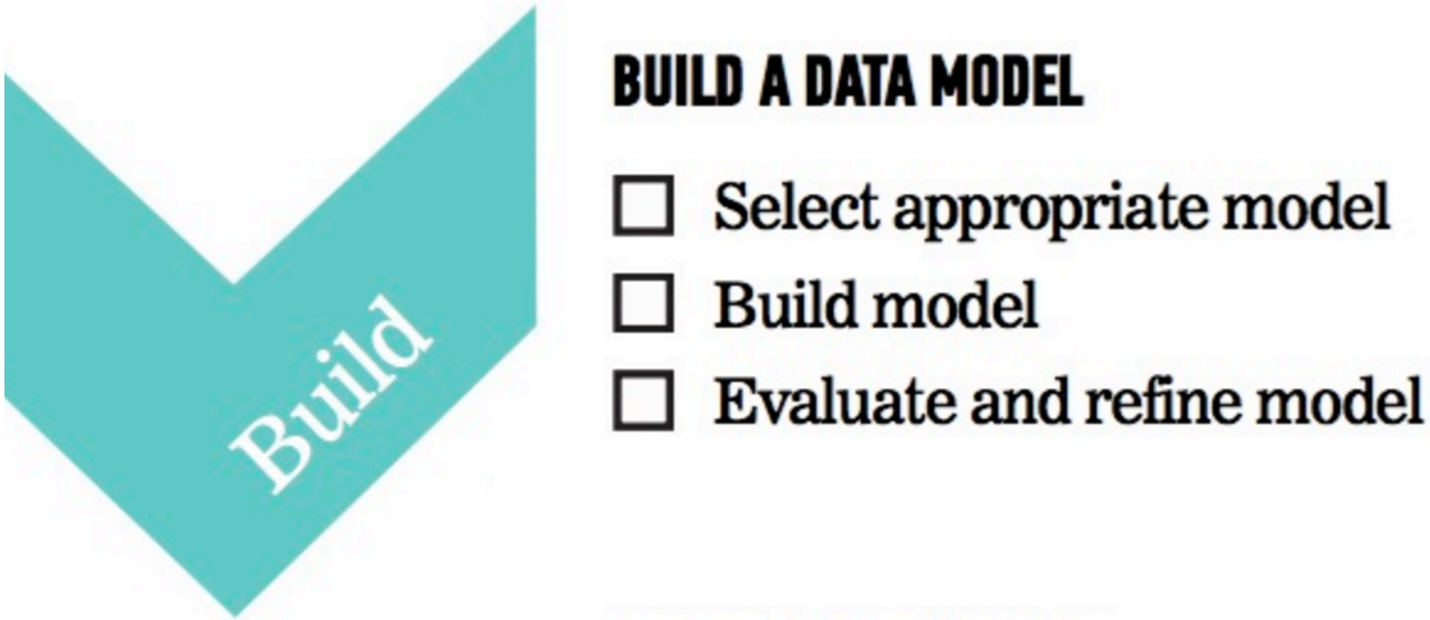
- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

Overview of the Data Science Workflow

REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

Overview of the Data Science Workflow



Overview of the Data Science Workflow

PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

GUIDED PRACTICE

DATA SCIENCE WORKFLOW

ACTIVITY: DATA SCIENCE WORKFLOW

DIRECTIONS (30 minutes)

There are 5 example presentations available for review at the link:
https://github.com/generalassembly-studio/dat11syd/tree/master/lessons/lesson-01/2_groupwork

1. Divide into 5 groups
2. Review one of the 5 final presentations selected from a past DAT course
3. Discuss with your team:

    a) Is it clear what the Data Science question is?
    b) Did the student communicate the problem space clearly?
    c) Did the student stick to the data science workflow as a framework?
    d) Did the student explain what their assumptions where and potential flaws?
    e) Did the student achieve?
    f) What would you do better if you did this project again?

DELIVERABLE

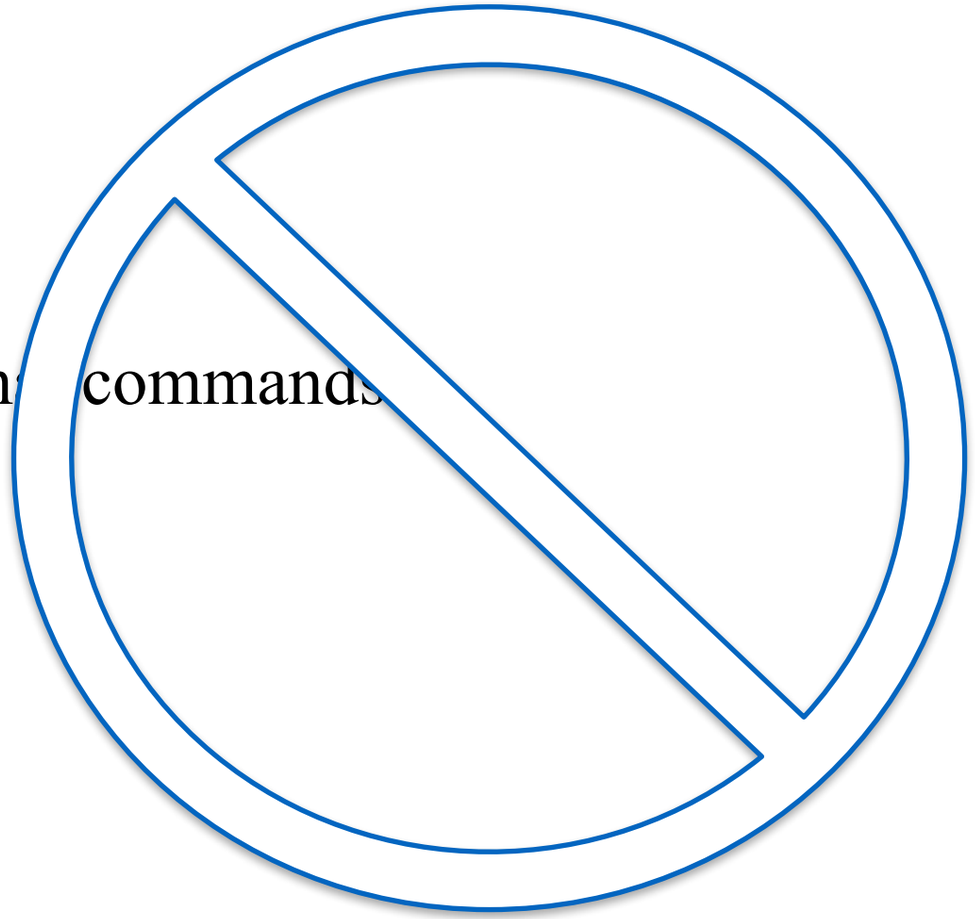
Presentation of the results

GUIDED PRACTICE

USING THE ENVIRONMENT

Dev Environment – old!

- Environment setup
  - Github account + Git client
  - Python 3.6 + Anaconda
  - Python syntax, Pandas, Termin commands
- iPython Notebook test
  - Python review
  - magics
- Github review / test



Dev Environment – New!

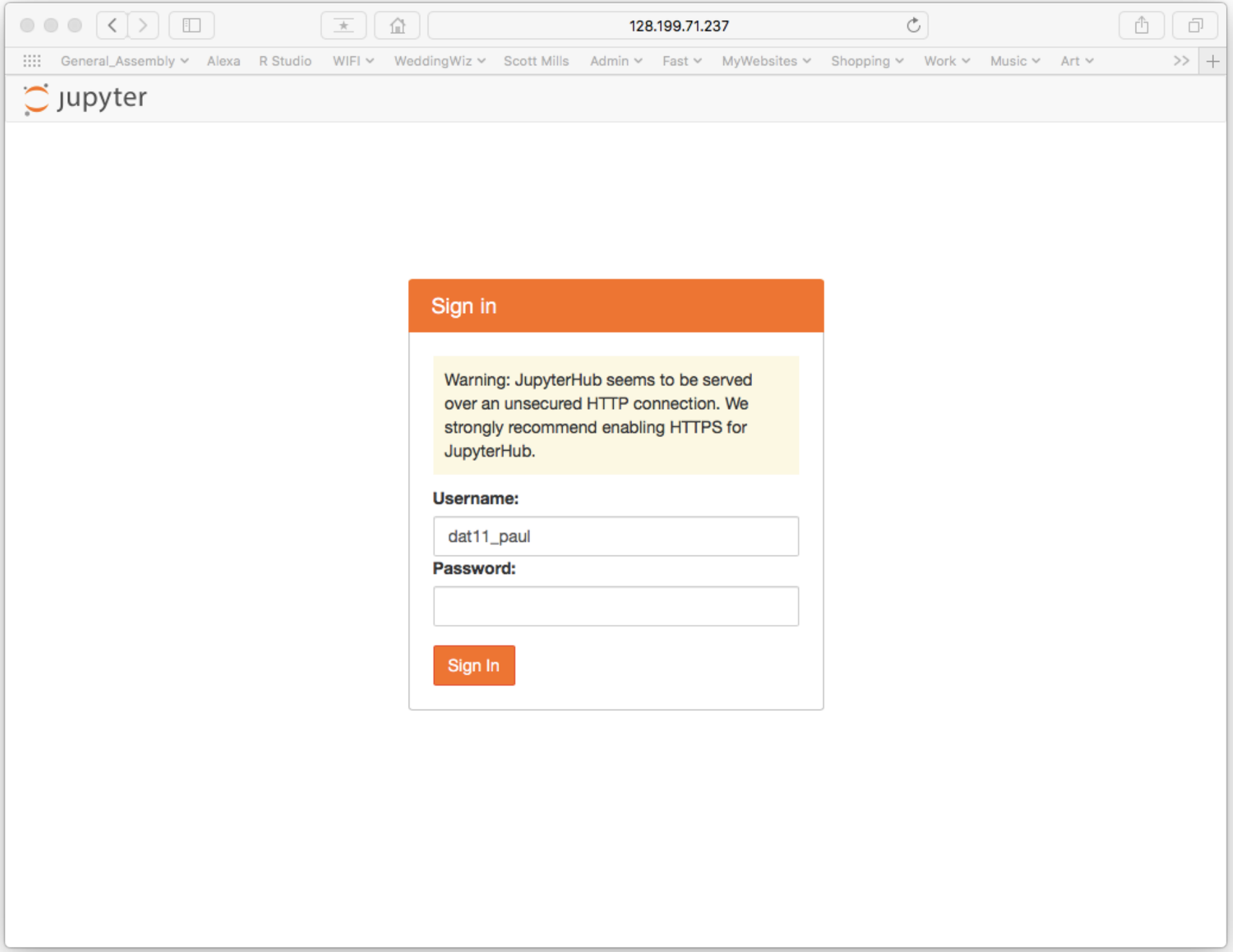
New platform – class server dat11:

http://www.paulgoodall.tech:8000/

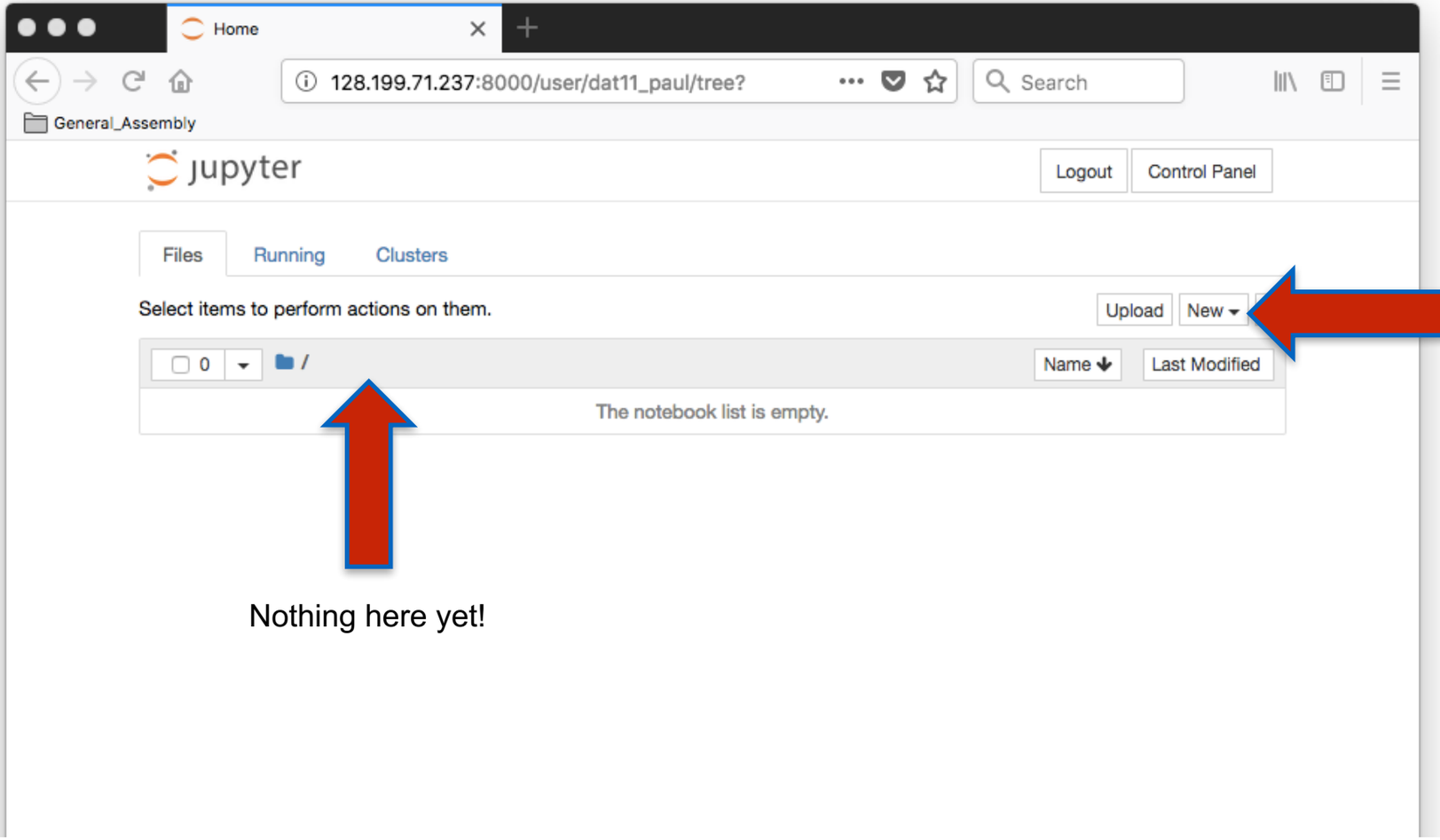
http://128.199.71.237:8000/hub/login

First name	Last name	Username	Password
Paul	Goodall	dat11_paul	
James	Lai	dat11_james	monkey_9
Akif	Alim	dat11_akif	monkey_9
Mai	Anh Ly	dat11_mai	monkey_9
Bernadette	Basha	dat11_bernadette	monkey_9
Swernim	Bhardwaj	dat11_swernim	monkey_9
Suren	Chandrasekera	dat11_suren	monkey_9
Renee	Checchin	dat11_renee	monkey_9
Sophia	Cherem Lopes	dat11_sophia	monkey_9
Sung	Choi	dat11_sung	monkey_9
Jesse	Imer	dat11_jesse	monkey_9
Bridget	Loudon	dat11_bridget	monkey_9
Alex	McCauley	dat11_alex	monkey_9
John	Niyonsaba	dat11_john	monkey_9
Andrew	Ren	dat11_andrew	monkey_9
Michael	Tostee	dat11_michael	monkey_9
Ricardo	Valencia	dat11_ricardo	monkey_9
Evan	Wong	dat11_evan	monkey_9
Jon	Workman	dat11_jon	monkey_9
Linna	TBC	dat11_linna	monkey_9

Logging in

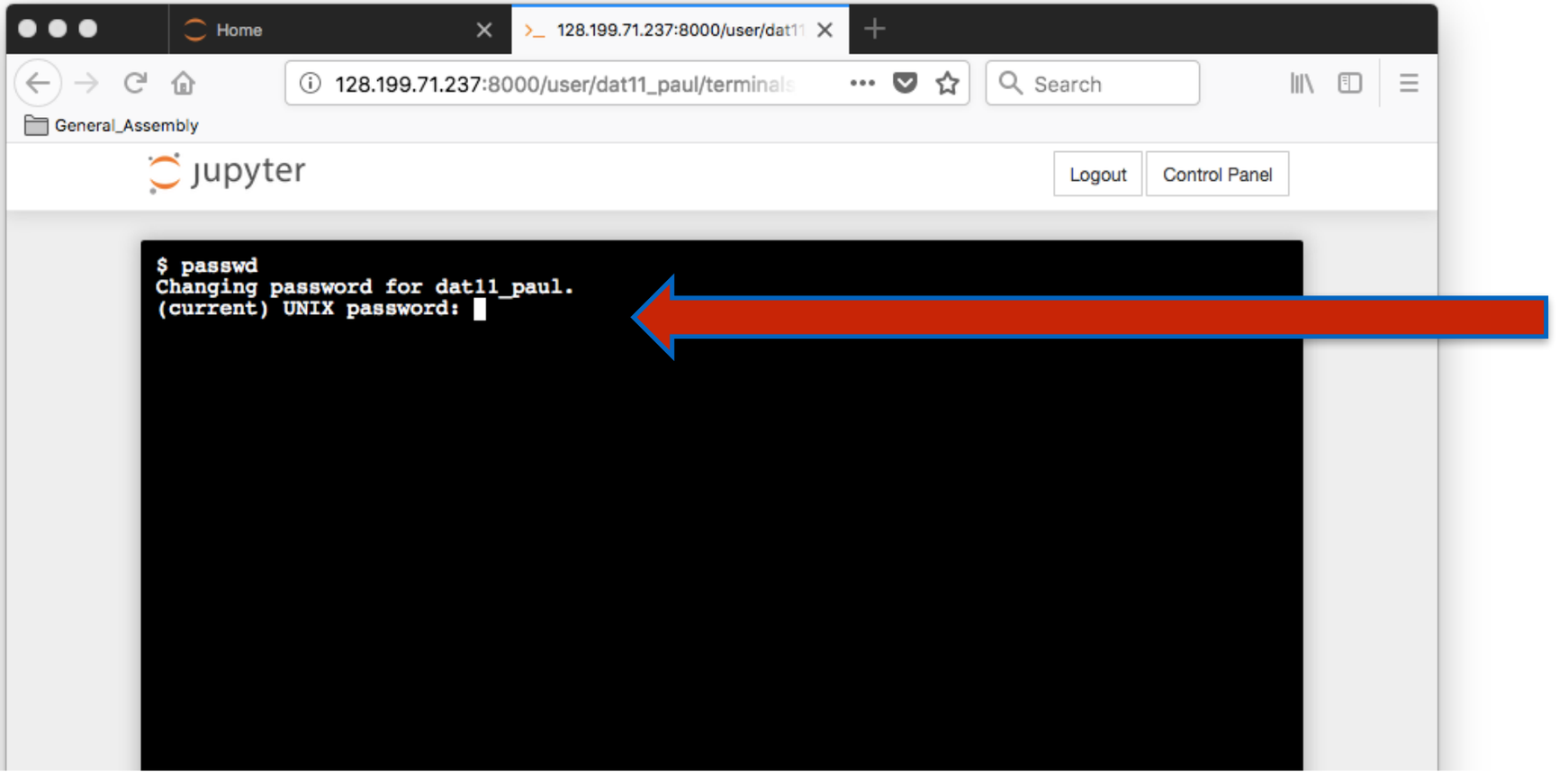

128.199.71.237
General_Assembly
Alexa
R Studio
WIFI
WeddingWiz
Scott Mills
Admin
Fast
MyWebsites
Shopping
Work
Music
Art
jupyter
Sign in
Warning: JupyterHub seems to be served over an unsecured HTTP connection. We strongly recommend enabling HTTPS for JupyterHub.
Username:
dat11_paul
Password:
Sign In

Welcome to Jupyter Hub



Select “Terminal”

Welcome to Jupyter Hub



Type the command:
passwd

and follow the prompts to change your login password.

WRITE IT DOWN.

Introduction to the Unix COMMAND LINE

- Begin this tutorial from slide 15
- Use the left and right arrow keys to navigate

https://software.rc.fas.harvard.edu/training/intro_unix/latest/#(15)

CONCLUSION

REVIEW

Conclusions

- You should now be able to answer the following questions:
  - What is Data Science?
  - What is the Data Science workflow?
  - How can you have a successful learning experience at GA?

Additional Resources

- Data Science Central
  - http://www.datasciencecentral.com/
- Python.org
  - https://www.python.org/about/
- Amazon
  - books on Python & Data Science
- LinkedIn
  - groups
- 538 (Nate Silver)
  - http://fivethirtyeight.com/

Optional Review

- **Probability & Statistics**
  - *Bonus: Additional Practice*
- **Python Syntax**
  - *Bonus: Additional Practice (Exercises 1-20)*
- **Git Tutorial**
  - *Bonus: Additional Practice*
- **Command Line Overview**
  - *Bonus: Additional Practice*

DATA SCIENCE

BEFORE NEXT CLASS

Before Next Class

- Finish course prework
- First Assignment
- Practice Python programming (Python for Data Science course on DataCamp)
- Join Kaggle
  - https://www.kaggle.com/

Q & A

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

https://docs.google.com/forms/d/e/1FAIpQLSdaBdMebhoXQpac2edDbIkHD-78HrTpBk3VKGH6lcDMVorSIQ/viewform