GA GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE @ GA

Winston Featherly-Bean
Lead Instructor

Ellen Kim
Instructional Associate

John Sabini
Instructional Associate

LET’S START!

ABOUT ME

- Welcome to the part-time data science course at GA!
- We're Winston, Ellen and John.

ABOUT YOU

- How about you?
  - What's your name?
  - What are you doing now?
  - What do you want to be able to do after this course?

AGENDA

Timing	Topic
6.30 - 7:00	Opening and intros
7:00 - 7:10	Course outline
7:10 - 7:15	Tips for success
7:15 - 7:30	What is data science?
7:30 - 7:45	What is ML?
7:45 - 8:15	Example ML algorithm: K-nearest neighbors
8:15 - 8:25	Break
8:25 - 8:45	Tools and tech check
8:45 - 9:15	Python flash quiz!
9:15 - 9:30	Q&A / exit ticket

DATA SCIENCE @ GA

COURSE OUTLINE

COURSE LEARNING OBJECTIVES – YOU WILL BE ABLE TO…

- Define the language and approaches data scientists use to solve real world problems
- Perform exploratory data analysis with powerful programmatic tools, including Python
- Build and refine machine learning models to predict future outcomes
- Communicate data-driven insights to inform business decisions
- Start on your next data science project, alone or with a new community of collaborators!

COURSE SCHEDULE

Unit	Lessons	Class days
1 - Foundations	-What is data science? -Your development environment -Python foundations -Project workshop / FLEX	September 18, 20, 25, 27
2 - Working with Data	-Statistics review -Stats + plots in Python -Exploratory data analysis -Data visualization in Python -Project workshop / FLEX	October 2, 4, 11, 16, 18 (No class on October 9)

COURSE SCHEDULE

Unit	Lessons	Class days
3 - Data modeling	- Linear regression -Bias-variance + train-test split -Classification + KNN -Logistic regression -Project workshop / FLEX	October 23, 25, 30 November 1, 6
4 - Applications	-APIs and webscraping -NLP -Decision trees and random forests -Clustering -Project workshop / FLEX -Final project presentations	November 8, 13, 15, 20, 27, 29 (No class on November 22)

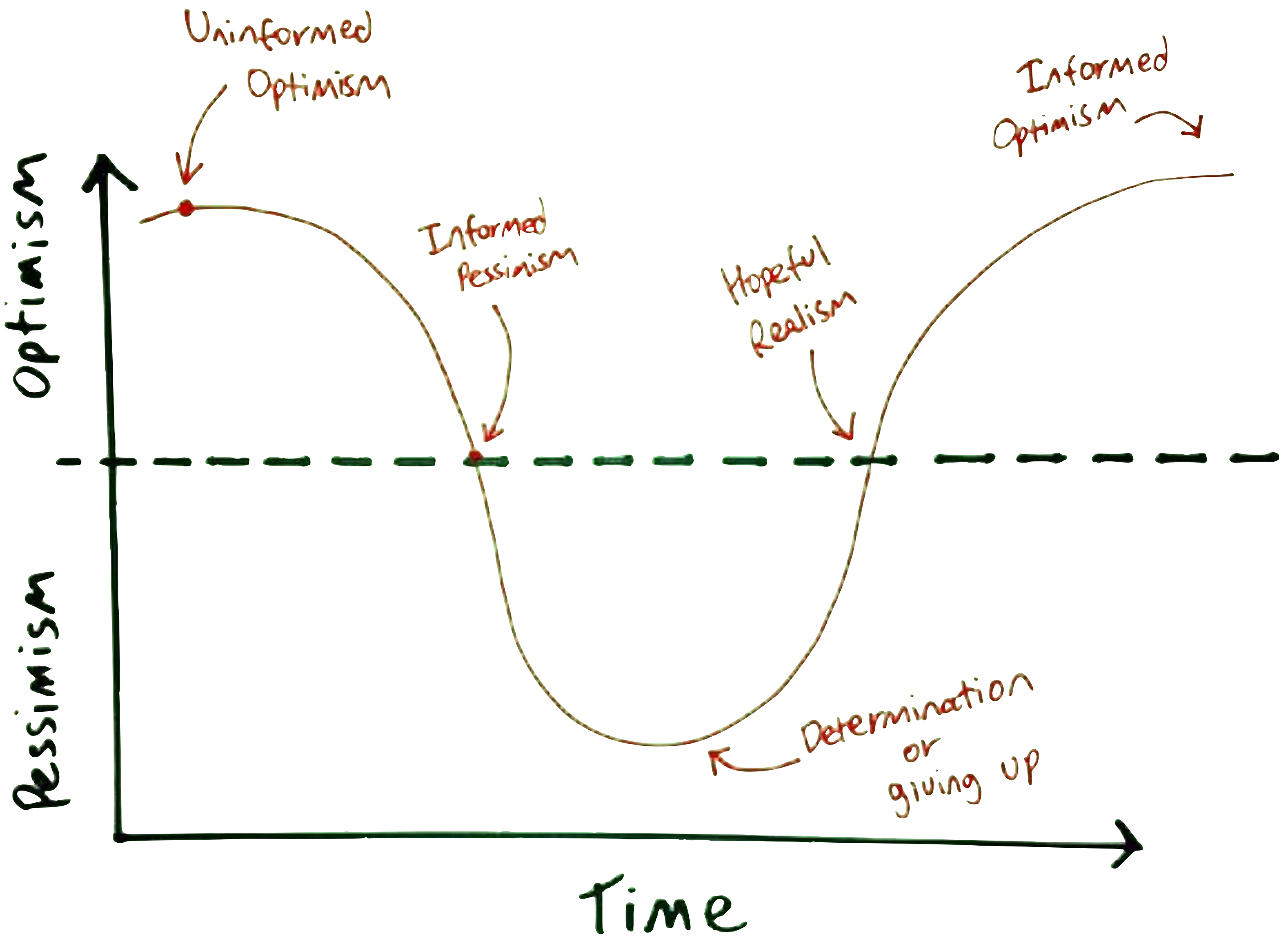
COURSE SCHEDULE

- There’s some flexibility built in — we can adjust the syllabus according to class needs and interest
- There are three unit projects, and one final project
- For the final project, you can use one of our datasets and problem statements or scope your own!

DATA SCIENCE

HOW TO GET THE MOST OUT OF THIS COURSE

TIPS FOR SUCCESS



TIPS FOR SUCCESS

YOU HAVE THE FREEDOM TO FAIL
EXPERIMENT
MAKE MISTAKES
TRY NEW THINGS

TIPS FOR SUCCESS

OWN YOUR LEARNING

PROJECT-BASED LEARNING WORKS

BE OPEN-MINDED

DON'T ISOLATE YOURSELF

COME PREPARED

RESPECT EACH OTHER

TIPS FOR SUCCESS

ORDER OF OPERATIONS

STACK OVERFLOW
SLACK YOUR PEERS
SLACK US
OFFICE HOURS
EMAIL US

ONWARDS!

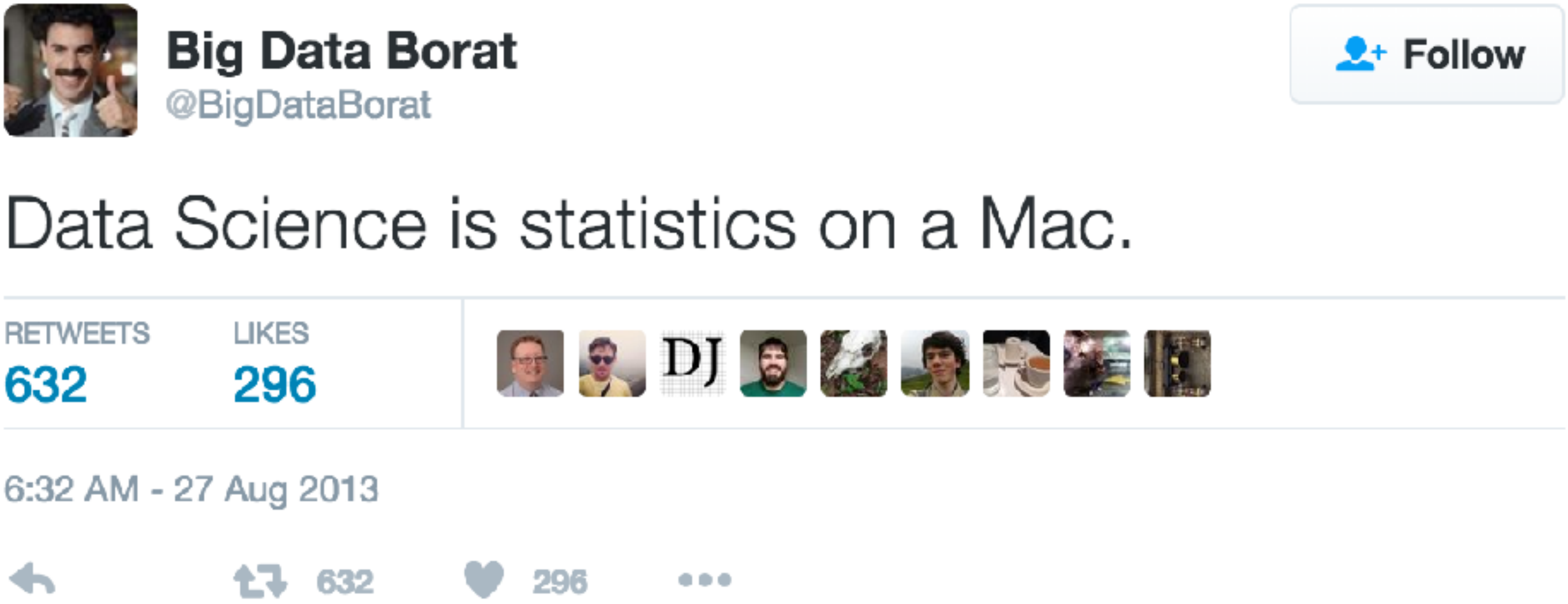
AGENDA

Timing	Topic
6.30 - 7:00	Opening and intros
7:00 - 7:10	Course outline
7:10 - 7:15	Tips for success
7:15 - 7:30	What is data science?
7:30 - 7:45	What is ML?
7:45 - 8:15	Example ML algorithm: K-nearest neighbors
8:15 - 8:25	Break
8:25 - 8:45	Tools and tech check
8:45 - 9:15	Python flash quiz!
9:15 - 9:30	Q&A / exit ticket

DATA SCIENCE @ GA

WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?



Big Data Borat
@BigDataBorat

Data Science is statistics on a Mac.

RETWEETS 632 LIKES 296

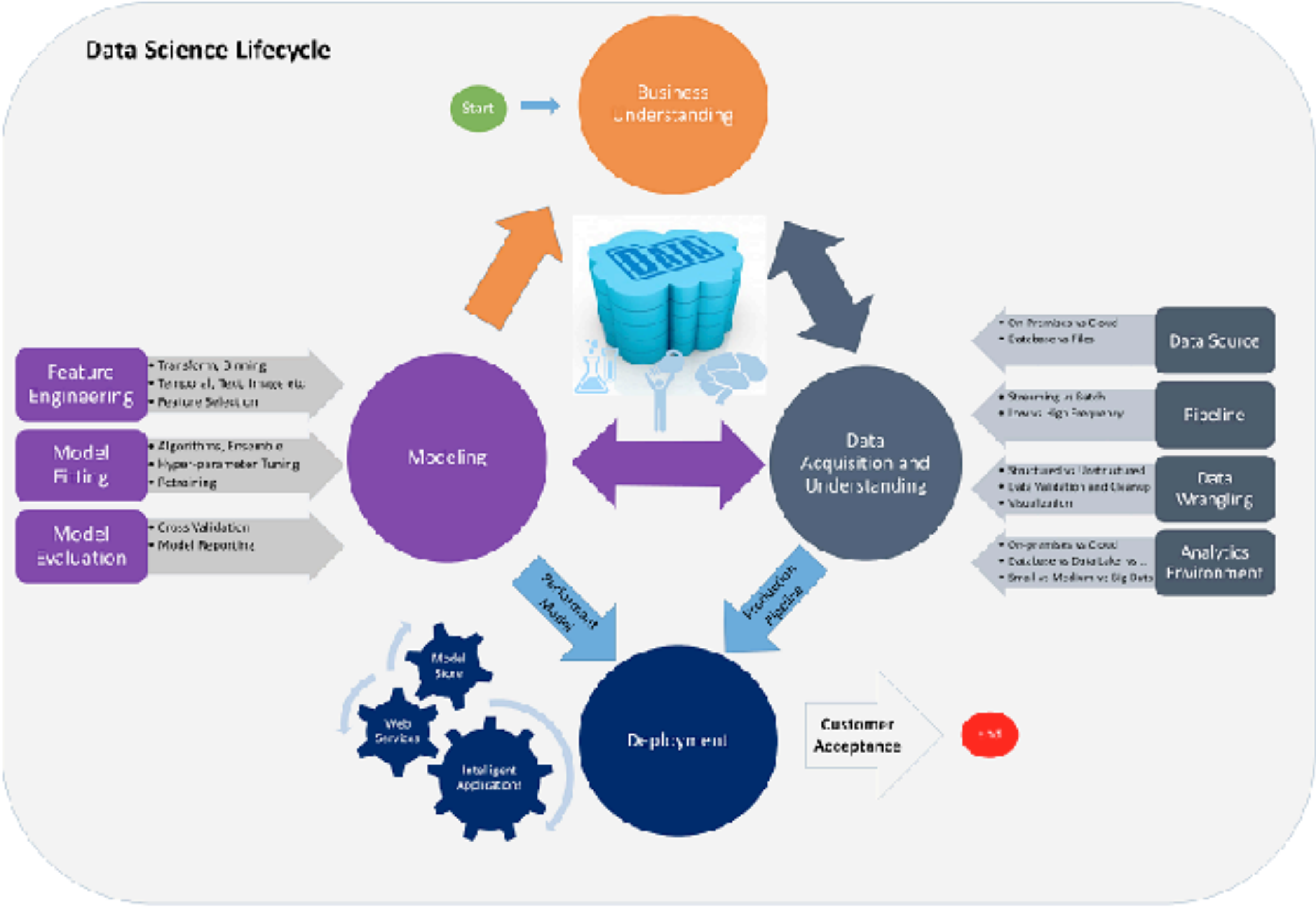
6:32 AM - 27 Aug 2013

WHAT IS DATA SCIENCE?

- "Data science"? A set of tools and techniques used to extract useful information from data
  - The application of scientific techniques to practical problems
  - An interdisciplinary, problem-solving-oriented subject
  - A rapidly changing space — and an overloaded term

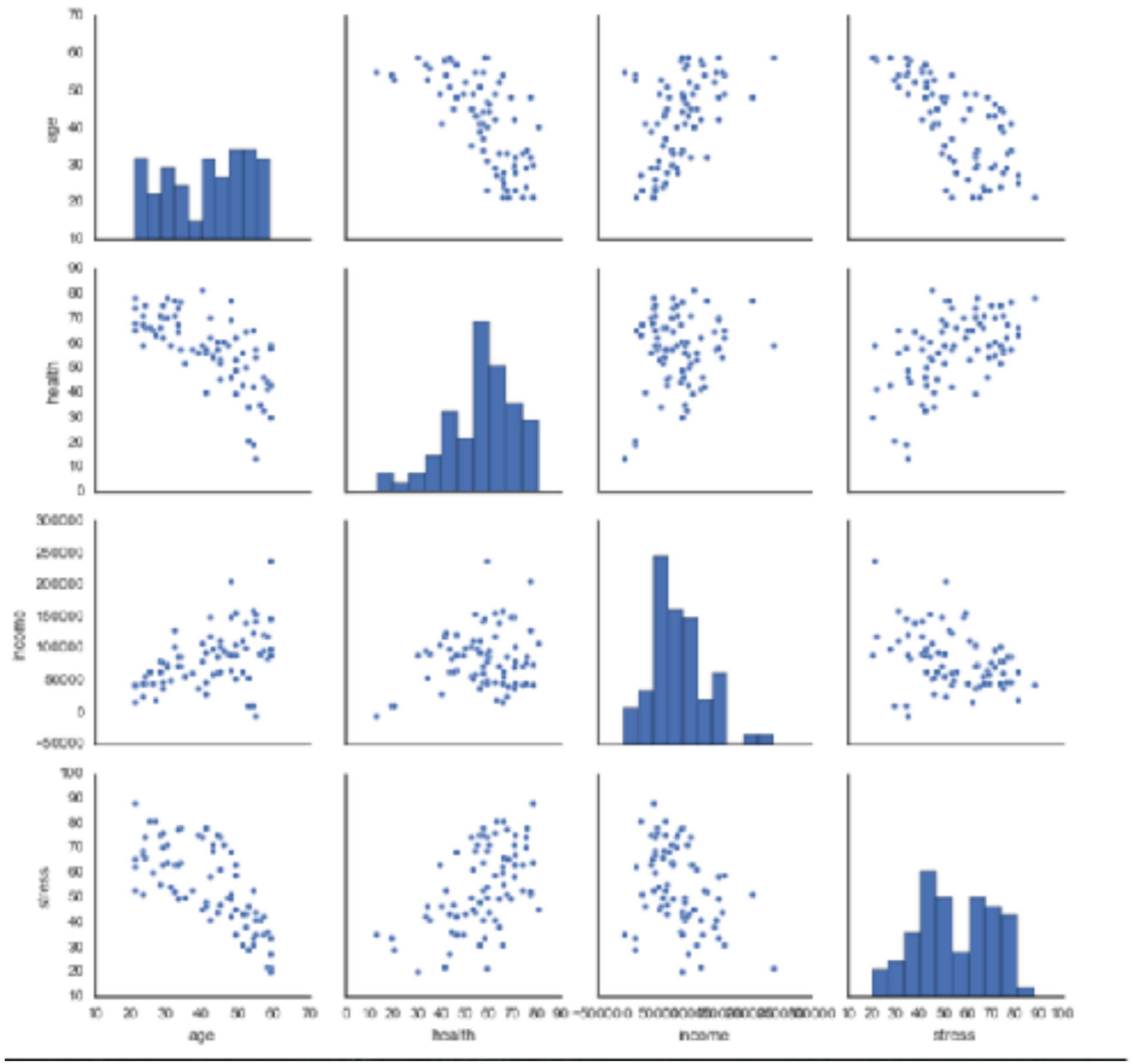
WHY DATA SCIENCE?

- Data too big for Excel
- Patterns too subtle for eyeballs
- Hypotheses to test
- Opportunity to automate



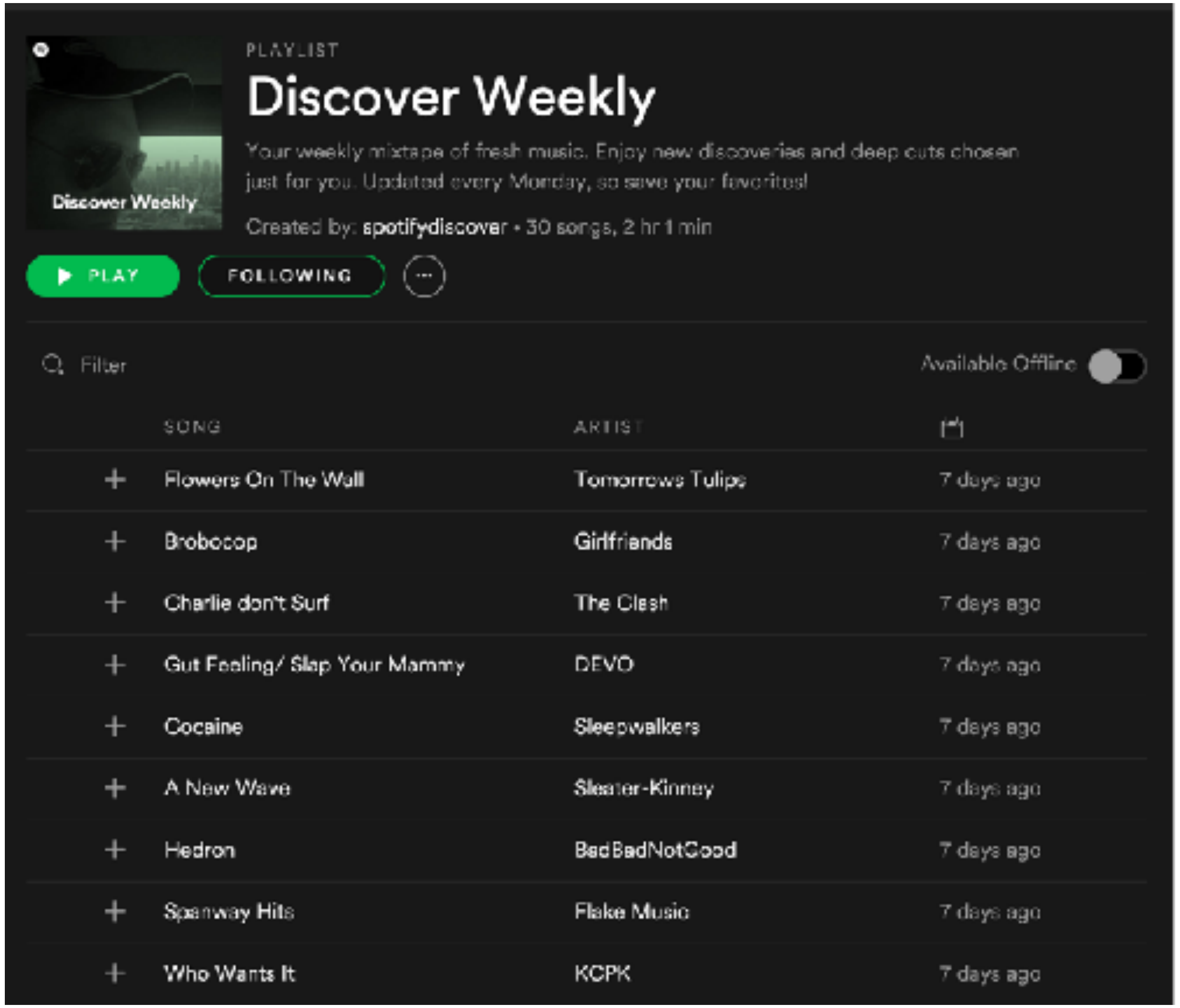
DATA ANALYSIS – DESCRIPTION, INFERENCE, PREDICTION (“TYPE A”)

- Scrape and collect business data, or generate data through experiments
- Sanitize and manipulate those data
- Visualize and describe the data
- Tell relevant business stories
- Infer relationships
- Build computerized models that predict and learn from data



DATA PRODUCTS – FUNCTIONAL OUTPUTS FROM STATISTICAL WORK (“TYPE B”)

- Build products from the output of data work and statistical analyses
- Might be…:
  - Structured datasets
  - Analytic dashboards
  - Predictions of machine learning models



Discover Weekly

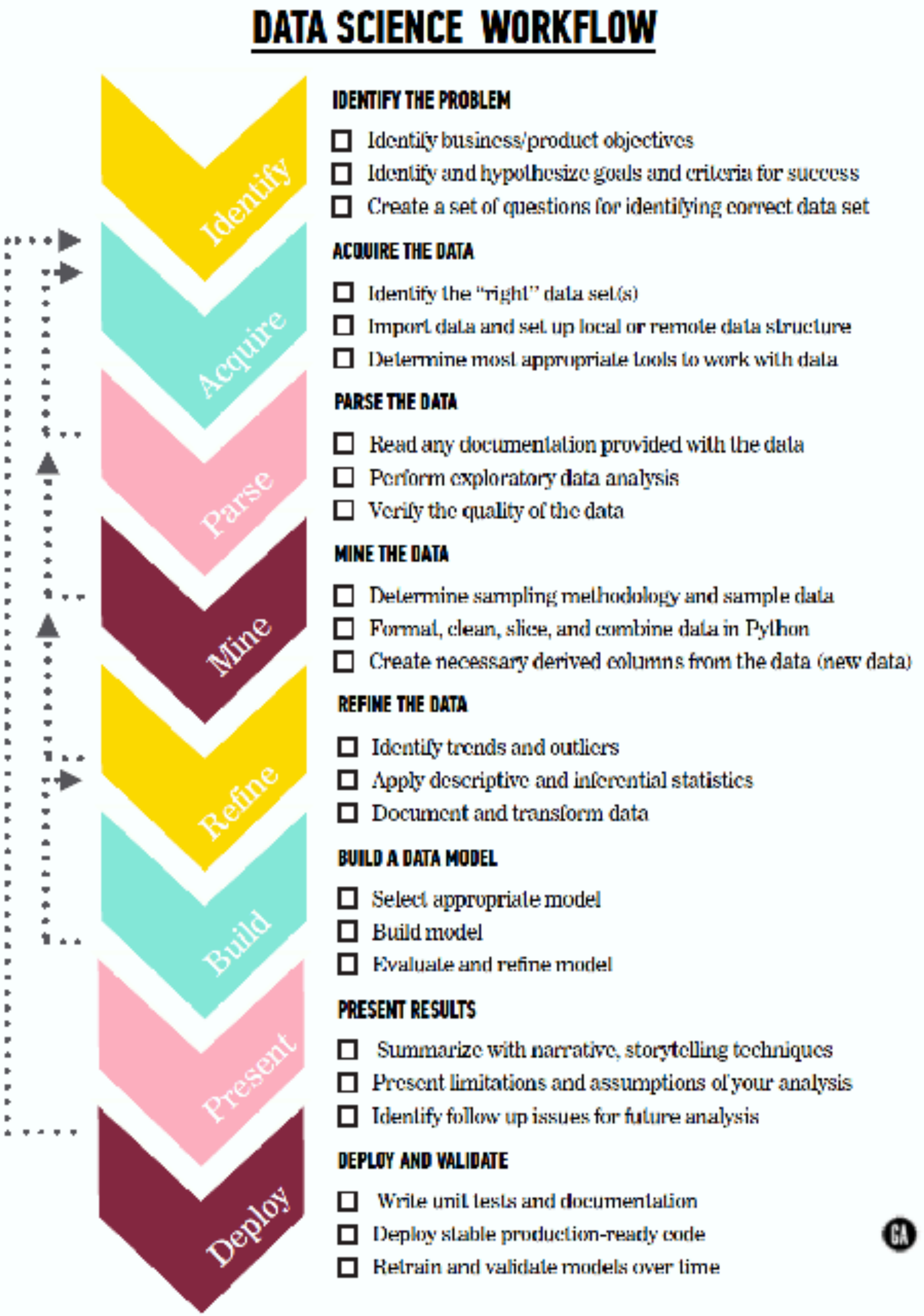
Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your favorites!

Created by: spotifydiscover • 30 songs, 2 hr 1 min

	SONG	ARTIST	📅
+	Flowers On The Wall	Tomorrows Tulips	7 days ago
+	Brobocop	Girlfriends	7 days ago
+	Charlie don't Surf	The Clash	7 days ago
+	Gut Feeling/ Slap Your Mammy	DEVO	7 days ago
+	Cocaine	Sleepwalkers	7 days ago
+	A New Wave	Sleater-Kinney	7 days ago
+	Hedron	BadBadNotGood	7 days ago
+	Spanway Hits	Flake Music	7 days ago
+	Who Wants It	KCPK	7 days ago

A DATA SCIENCE WORKFLOW

- Data science work should be reliable, reproducible and actionable
- We'll go deeper on workflow before kicking off your final projects



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

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

PARSE THE DATA

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

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)

REFINE THE DATA

□ Identify trends and outliers
□ Apply descriptive and inferential statistics
□ Document and transform data

BUILD A DATA MODEL

□ Select appropriate model
□ Build model
□ Evaluate and refine model

PRESENT RESULTS

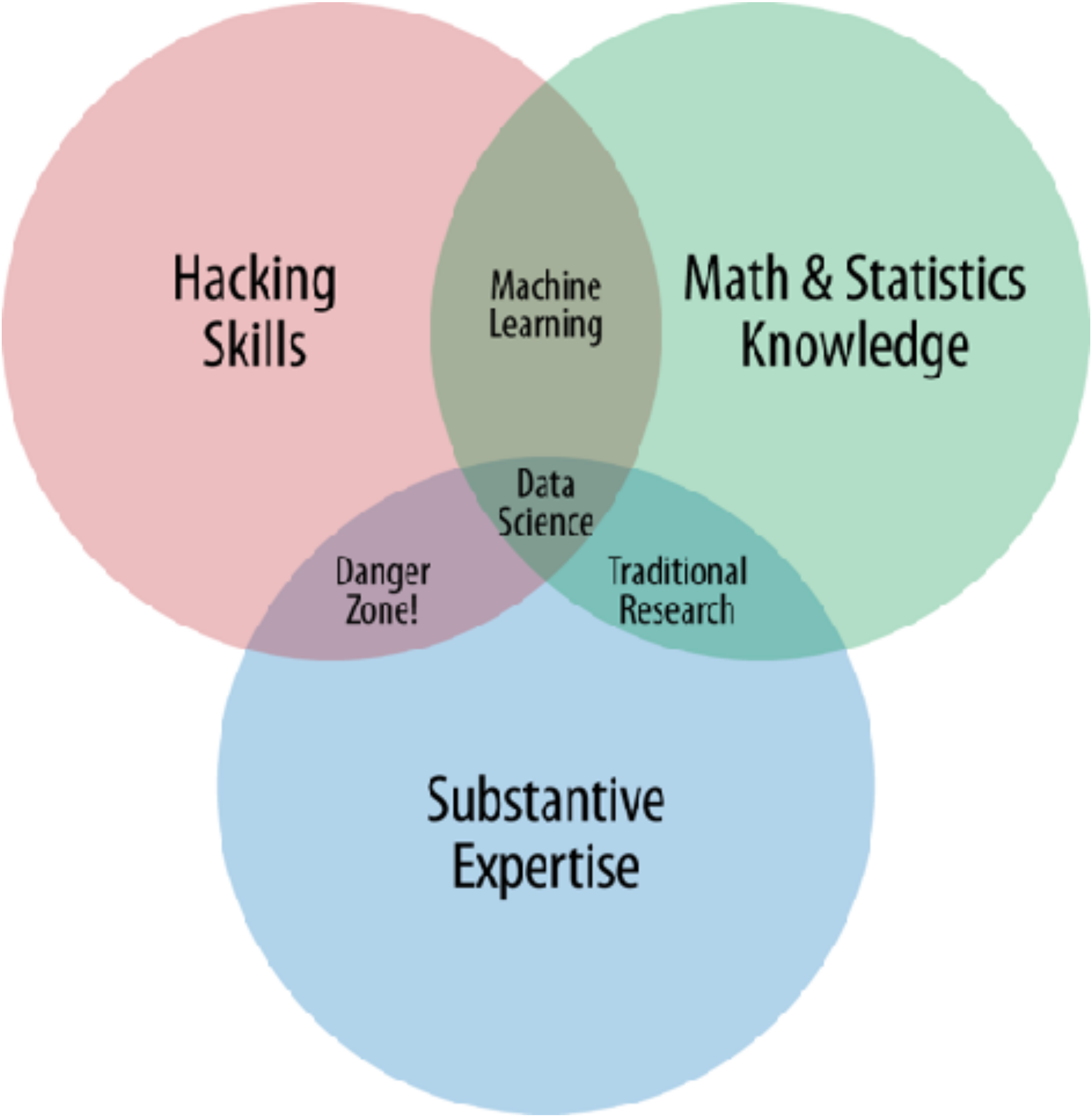
□ Summarize with narrative, storytelling techniques
□ Present limitations and assumptions of your analysis
□ Identify follow up issues for future analysis

DEPLOY AND VALIDATE

□ Write unit tests and documentation
□ Deploy stable production-ready code
□ Retrain and validate models over time

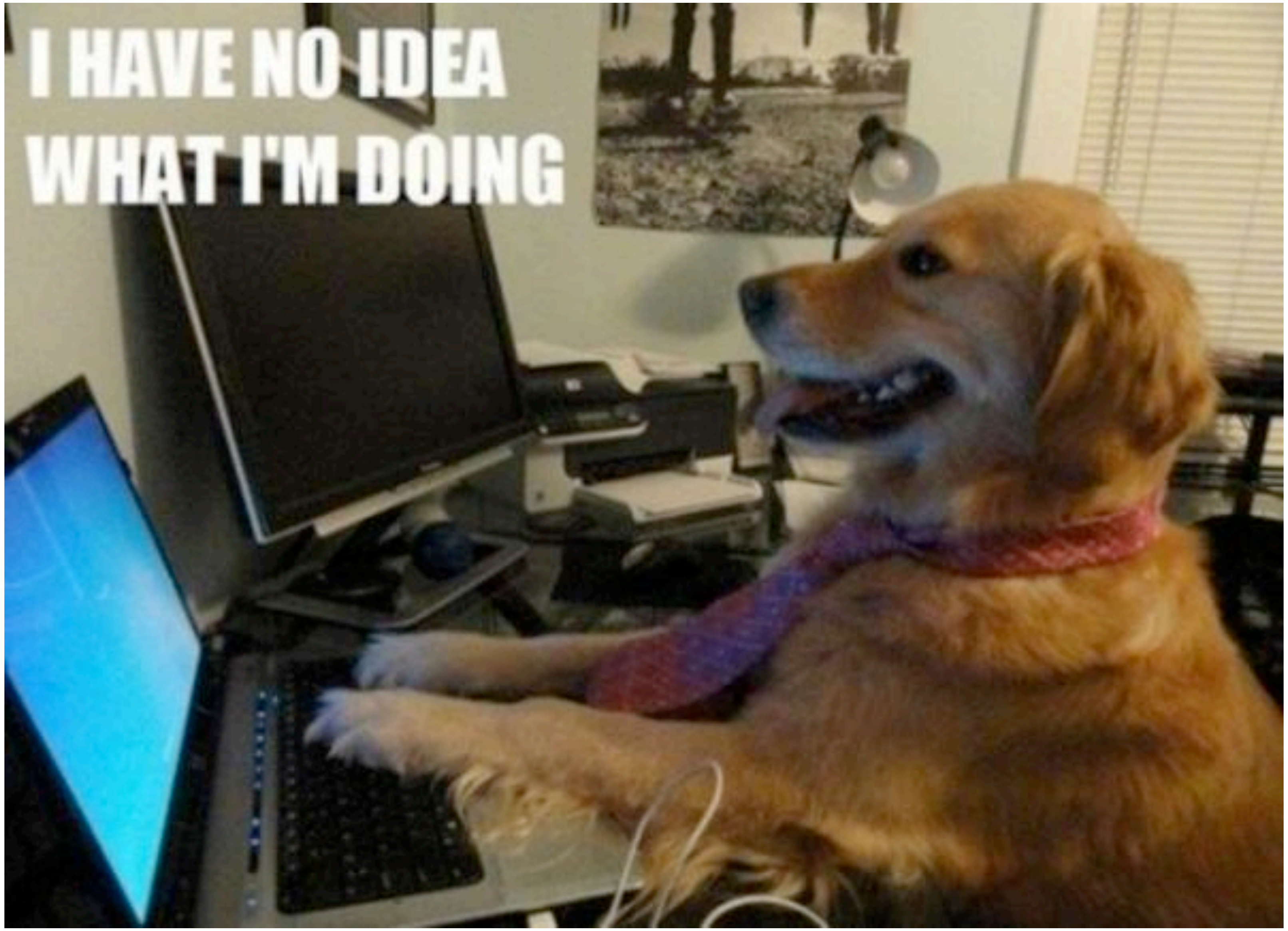
SKILLS OF A DATA SCIENTIST

- Programming skills
- Maths and Stats knowledge
- Business acumen (substantive expertise)
- Plus: Communication skills



SKILLS OF A DATA SCIENTIST

- No one knows everything
- The field changes rapidly
- Continuous learning required: it’s a feature, not a bug



MACHINE LEARNING

MACHINE LEARNING

- How to build systems that do not need to be explicitly programmed, and can improve with experience (more data)

- Two main branches: supervised and unsupervised machine learning

MACHINE LEARNING - TERMINOLOGY

- Supervised: there is a correct answer, we have examples of it, and we want to predict it

- Unsupervised: exploring possible structures in our data

MACHINE LEARNING – TERMINOLOGY

- Supervised machine learning can be used for two broad types of problems:
  - Regression - predicting a number
  - Classification - predicting a category
- The values we're predicting are our <u>targets</u>; we predict using <u>features</u>

MACHINE LEARNING - TERMINOLOGY

- Algorithms are a formal way of describing very precisely how to carry out certain computational tasks.
- Machine learning algorithms fit models to training data
- Try many iterations of those models by seeing how they perform on validation data
- We choose an error metric and use test data to assess our final model
- AI? ML + automation to perform specific tasks
- General AI… performs useful, novel tasks on command?

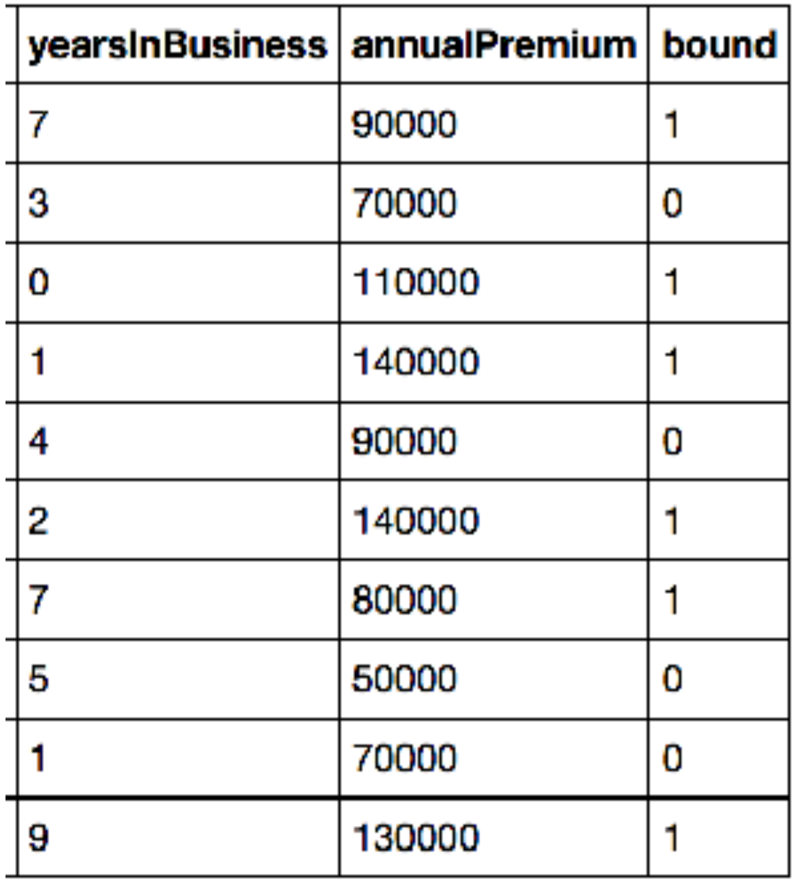
DATA SCIENCE @ GA

EXAMPLE ALGORITHM: K-NEAREST NEIGHBORS

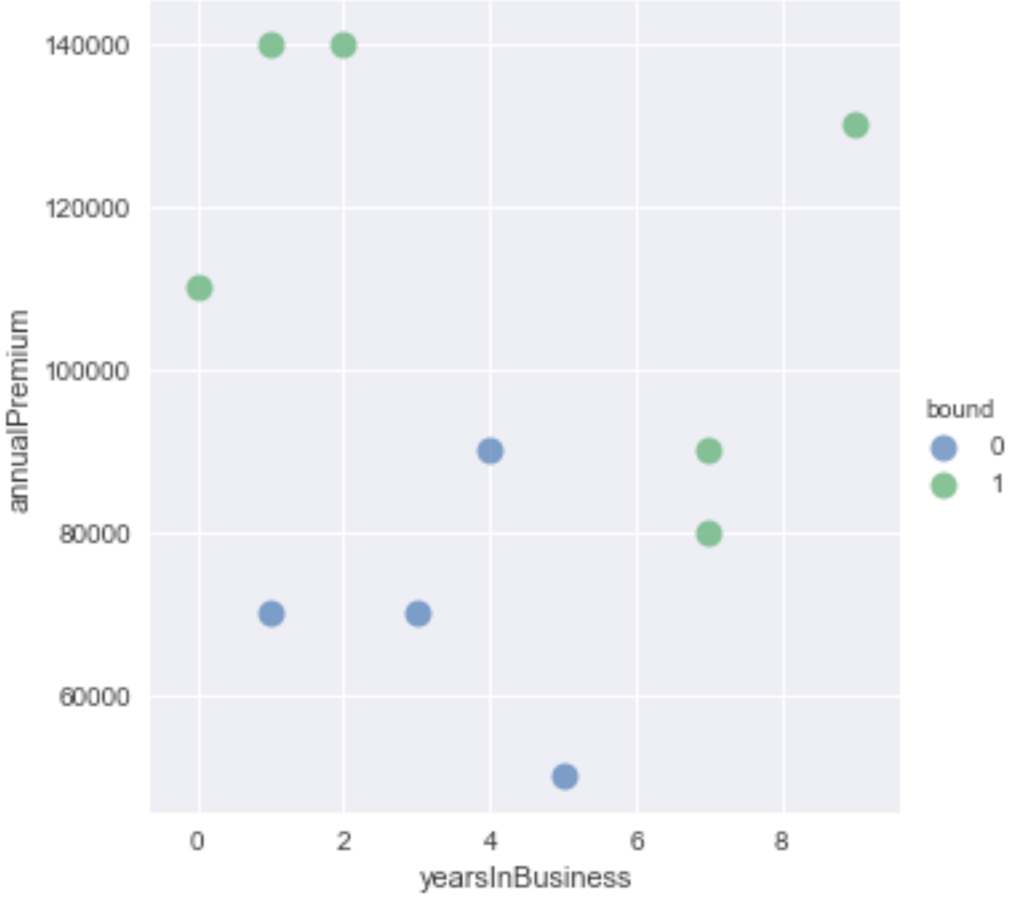
CLASSIFICATION WITH KNN

- K-Nearest Neighbors (kNN) is based on proximity to known data points with known classifications
- Cases are classified by a majority vote of its K-nearest neighbors, as measured by some distance function
    - E.g. if K = 1, then the case is simply assigned to the class of its nearest neighbor.

CLASSIFICATION WITH KNN



yearsInBusiness	annualPremium	bound
7	90000	1
3	70000	0
0	110000	1
1	140000	1
4	90000	0
2	140000	1
7	80000	1
5	50000	0
1	70000	0
9	130000	1



- Hypothetical results for quoting commercial insurance policies
- Is this a supervised or unsupervised problem?
- Do we have features and a target? If so, what kind of target?

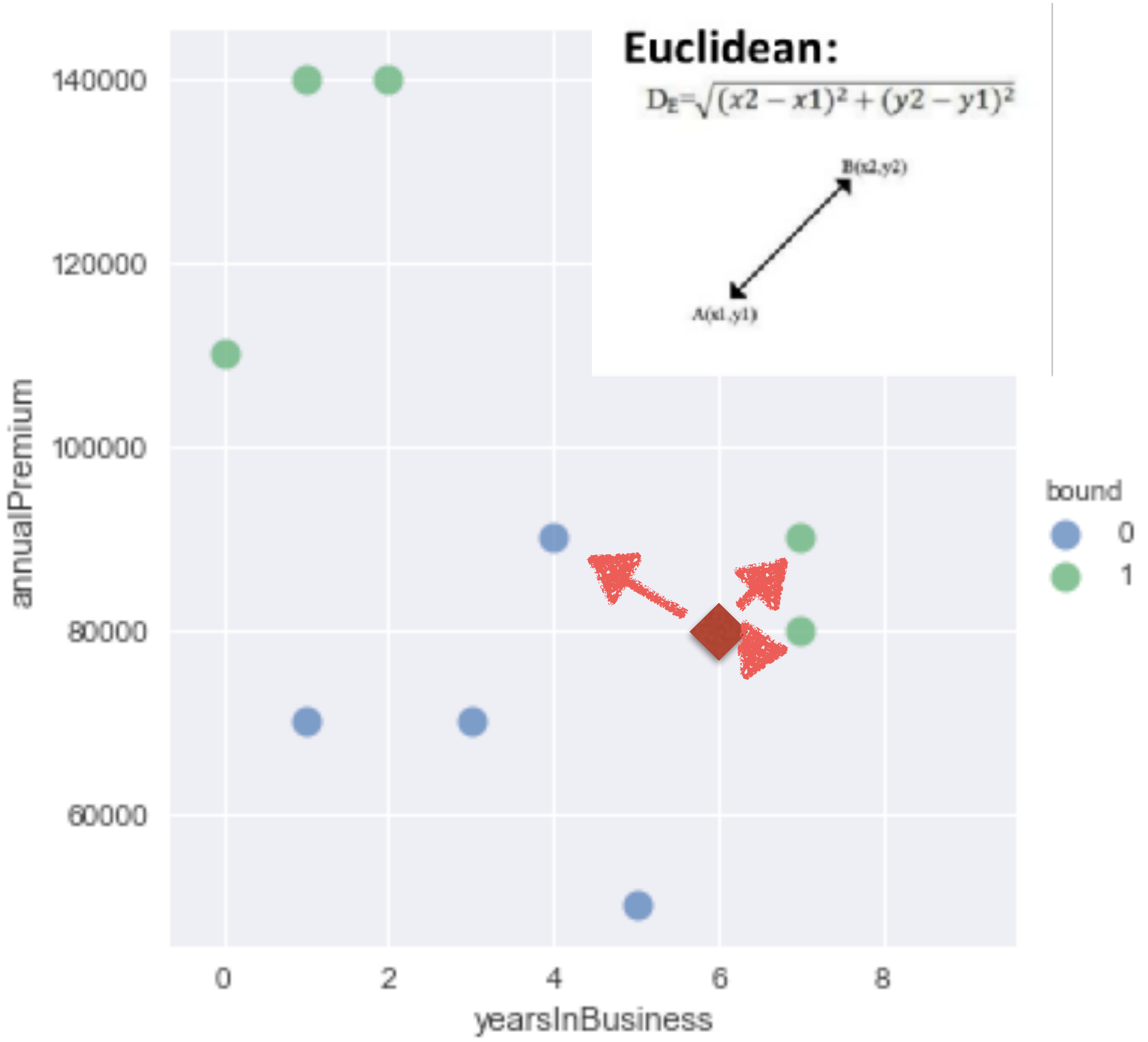
CLASSIFICATION WITH KNN

yearsInBusiness	annualPremium	bound	distance
7	90000	1	10000
3	70000	0	10000
0	110000	1	30000
1	140000	1	60000
4	90000	0	10000
2	140000	1	60000
7	80000	1	1
5	50000	0	30000
1	70000	0	10000
9	130000	1	50000

- NEW DATA:

6	80000	?

- Something is off…



Euclidean:

$$D_E = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

CLASSIFICATION WITH KNN

- Data science begins with interrogating the data
  - Where did it come from?
  - Is it correct?
  - Does it mean what we think it means, in its context?

CLASSIFICATION WITH KNN

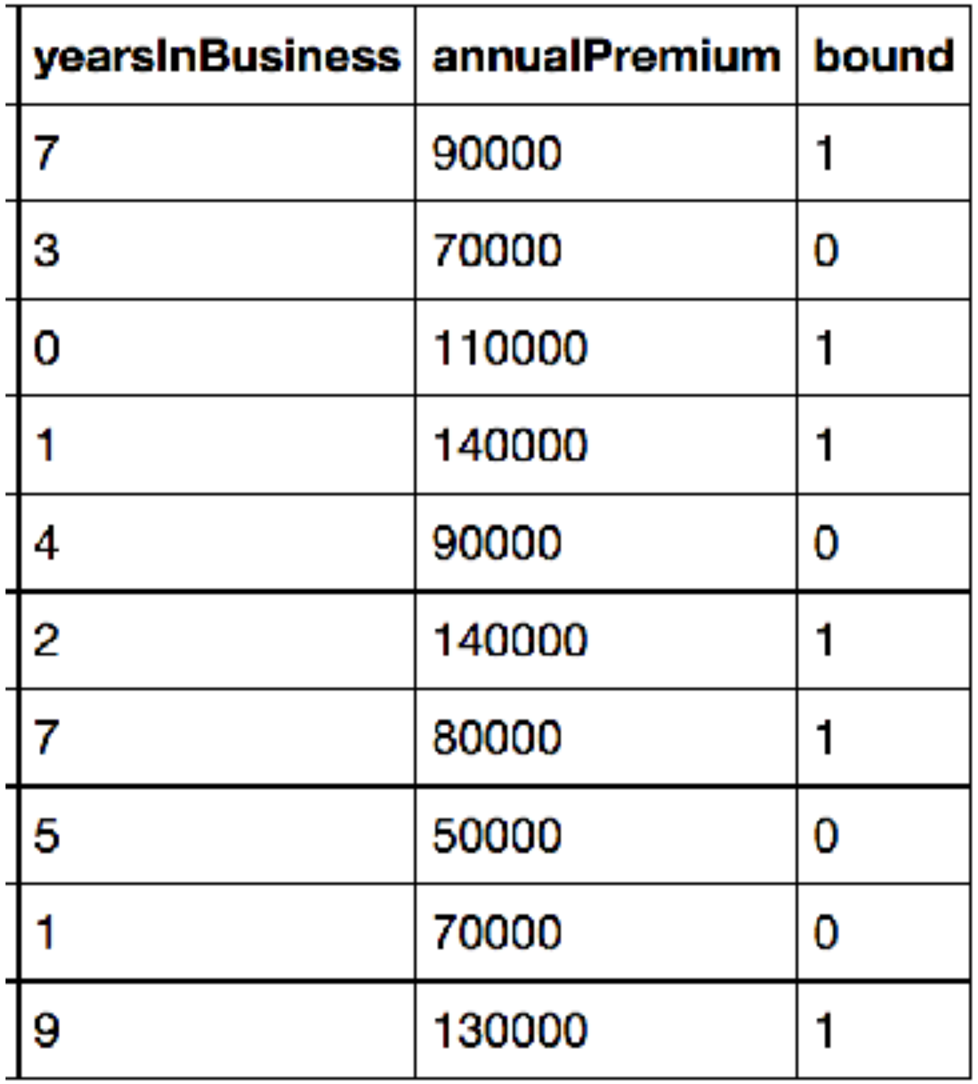
- Right now, annual premium is weighted vastly more than years in business
- Let’s rescale the data with “min-max scaling”, which adjusts feature values to be between 0 and 1
  - X` = (X - Xmin) / (Xmax - Xmin)
  - If our ‘years in business’ and ‘annual premium’ features had just three data points :

yearsInBusiness	scaled_years	annualPremium	scaled_premium
2	0	40,000	0
3	0.5	60,000	0.5
4	1	80,000	1

CLASSIFICATION WITH KNN

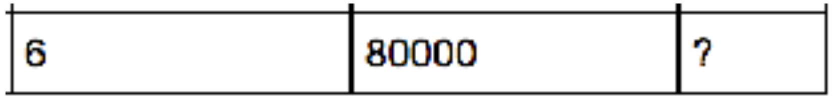
- Your turn! In pairs:
  - Scale the data
  - Plot the new data

X` = (X - Xmin) / (Xmax - Xmin)



yearsInBusiness	annualPremium	bound
7	90000	1
3	70000	0
0	110000	1
1	140000	1
4	90000	0
2	140000	1
7	80000	1
5	50000	0
1	70000	0
9	130000	1

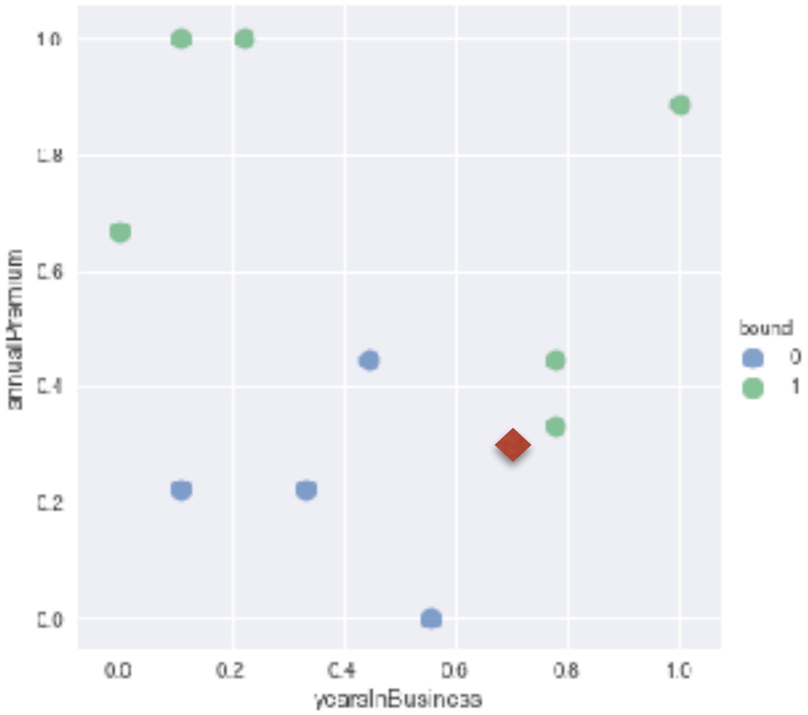
- NEW DATA:



6	80000	?

CLASSIFICATION WITH KNN

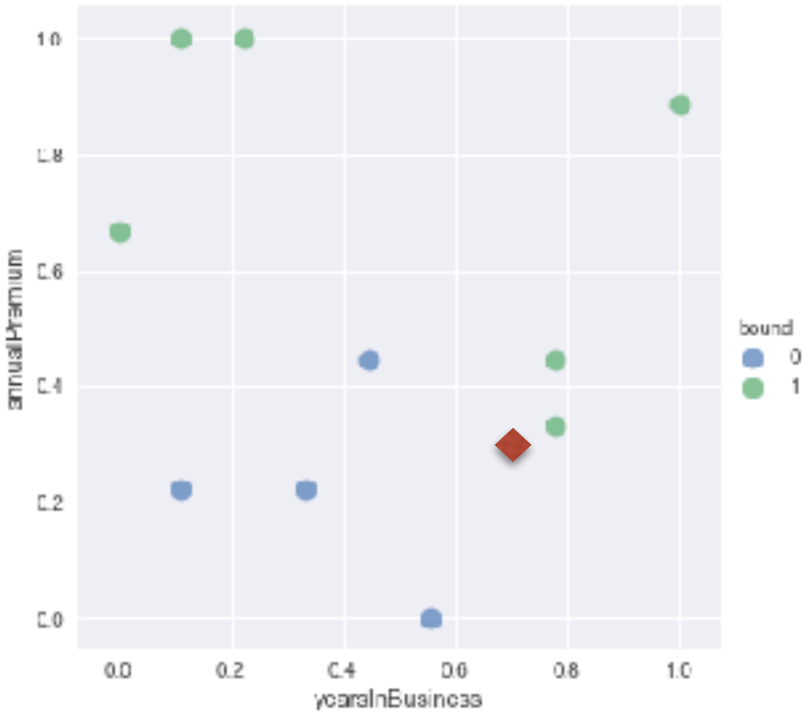
- If a potential insured has been in business 6 years, and we quote them $80,000, what does your algorithm predict when k = 3?
- What if k = 5?

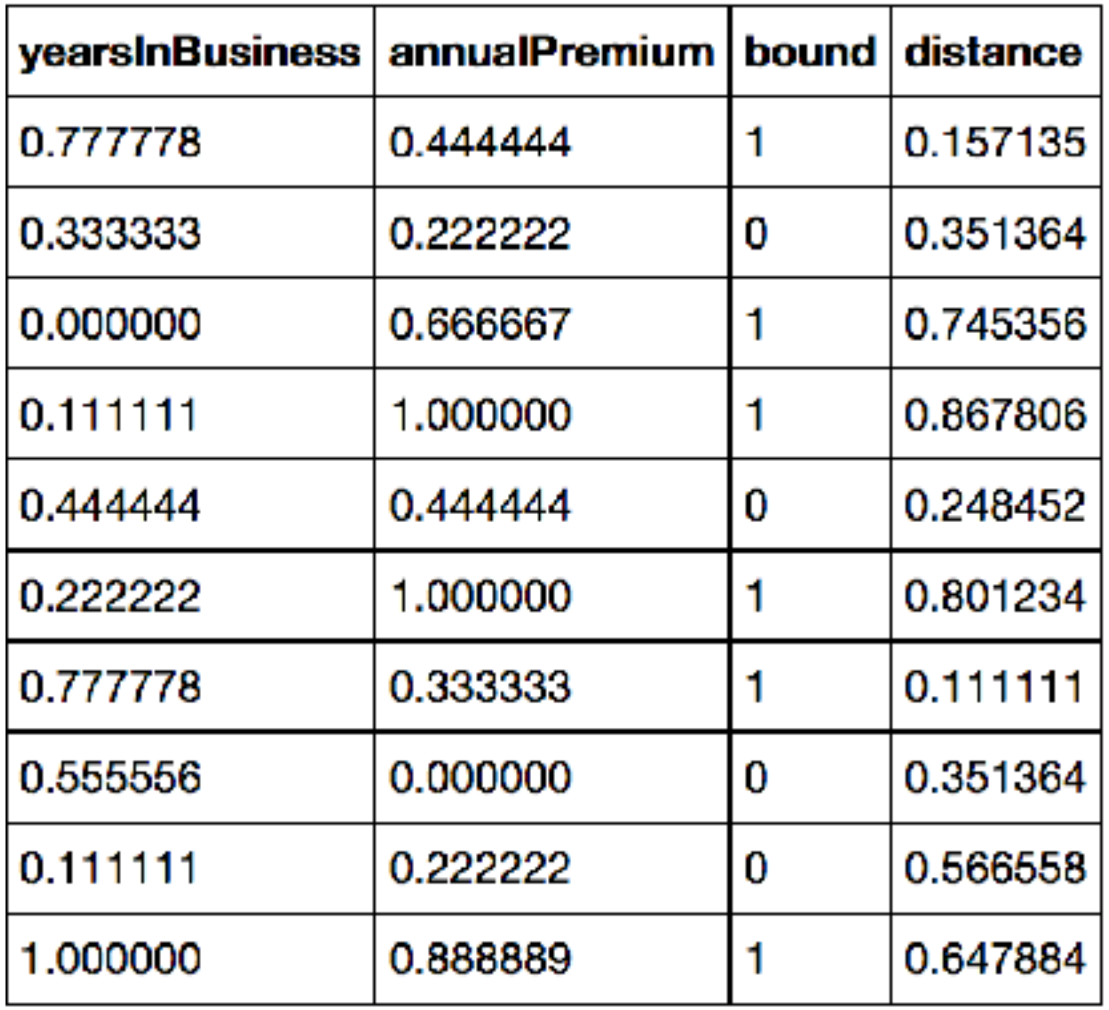


- (Slack me your prettier plots!)

CLASSIFICATION WITH KNN

- If a potential insured has been in business 6 years, and we quote them $80,000, what does your algorithm predict when k = 3?
- What if k = 5?





yearsInBusiness	annualPremium	bound	distance
0.777778	0.444444	1	0.157135
0.333333	0.222222	0	0.351364
0.000000	0.666667	1	0.745356
0.111111	1.000000	1	0.867806
0.444444	0.444444	0	0.248452
0.222222	1.000000	1	0.801234
0.777778	0.333333	1	0.111111
0.555556	0.000000	0	0.351364
0.111111	0.222222	0	0.566558
1.000000	0.888889	1	0.647884

BREAK

AGENDA

Timing	Topic
6.30 - 7:00	Opening and intros
7:00 - 7:10	Course outline
7:10 - 7:15	Tips for success
7:15 - 7:30	What is data science?
7:30 - 7:45	What is ML?
7:45 - 8:15	Example ML algorithm: K-nearest neighbors
8:15 - 8:25	Break
8:25 - 8:45	Tools and tech check
8:45 - 9:15	Python flash quiz!
9:15 - 9:30	Q&A / exit ticket

TOOLS + TECH CHECK

OUR TOOLSET

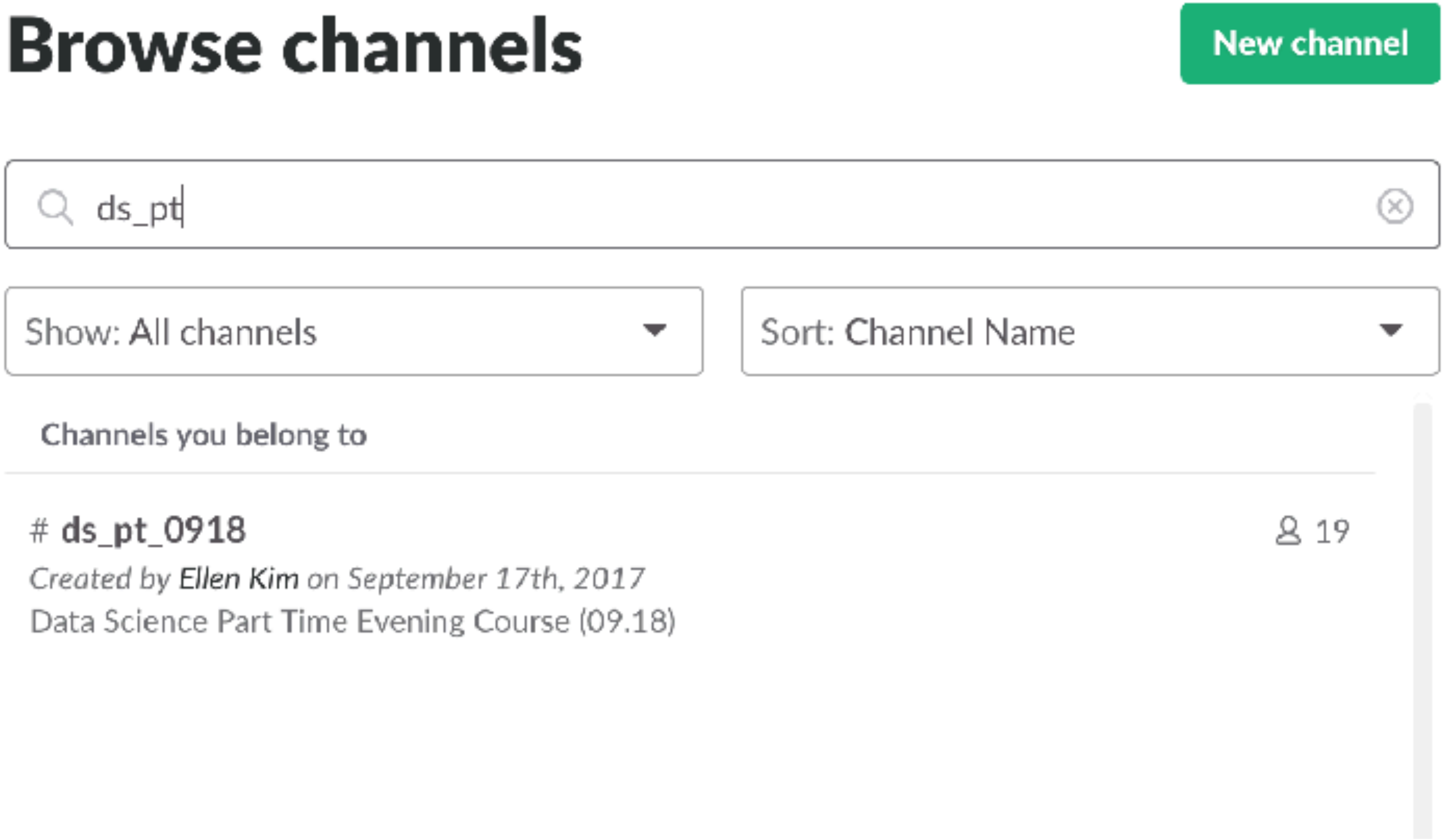
- Slack
- Git
- The command line
- Python 2.7*
- Jupyter notebooks

* Python 3.x is the right choice if you start a major project from scratch after this course.

OUR TOOLSET - SLACK

- Slack
  - You should have an invitation to GANYCEveningCourses.slack.com
  - It went to whichever email you registered with — possibly a work email
  - If you need help, flag down John or Ellen!

OUR TOOLSET - SLACK



ds_pt

Channels you belong to

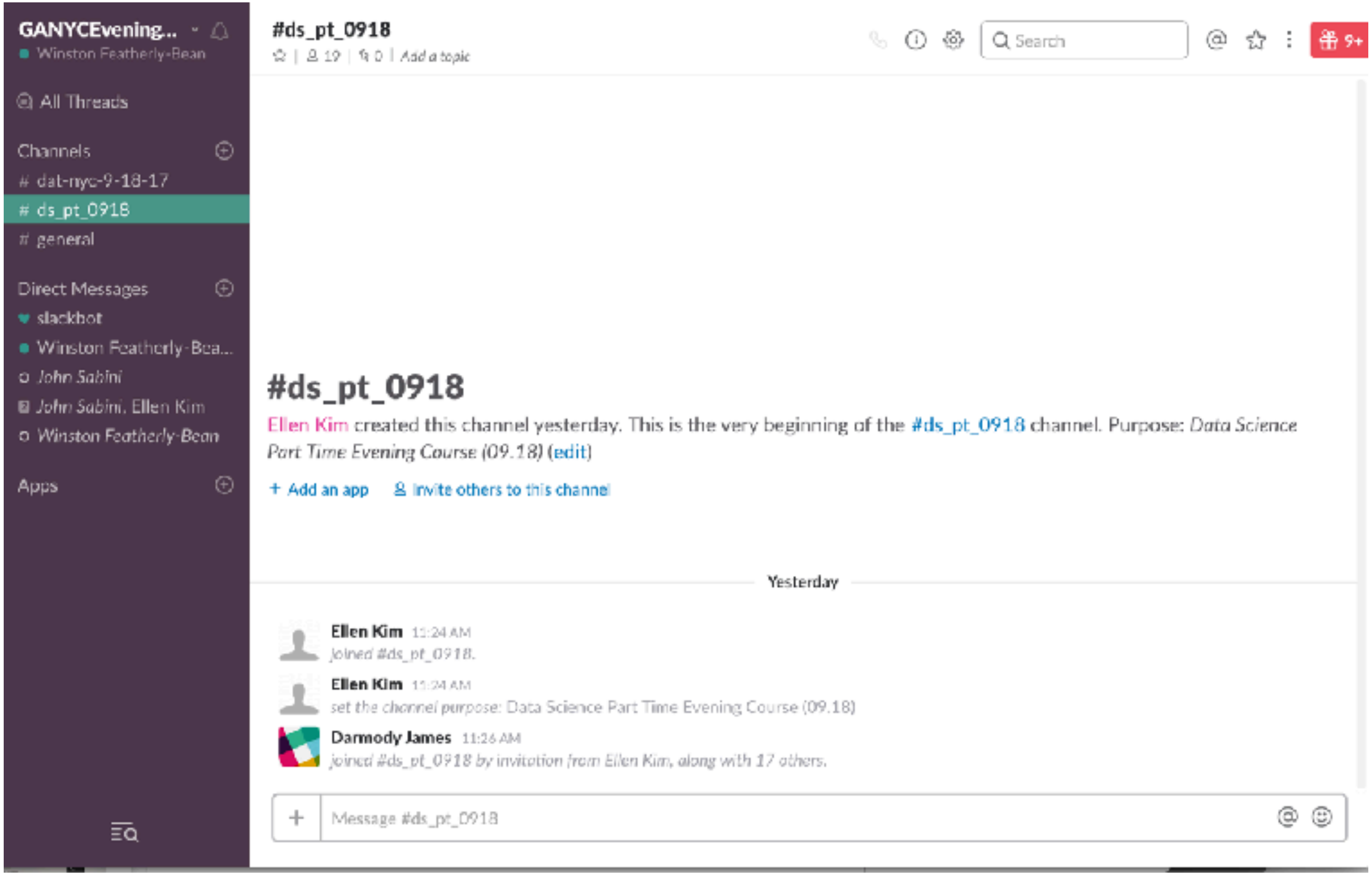
ds_pt_0918 19

Created by Ellen Kim *on September 17th, 2017*

Data Science Part Time Evening Course (09.18)

OUR TOOLSET – SLACK

- Say hi!



#ds_pt_0918

Ellen Kim created this channel yesterday. This is the very beginning of the #ds_pt_0918 channel. Purpose: *Data Science Part Time Evening Course (09.18)* (edit)

Ellen Kim 11:24 AM
joined #ds_pt_0918.

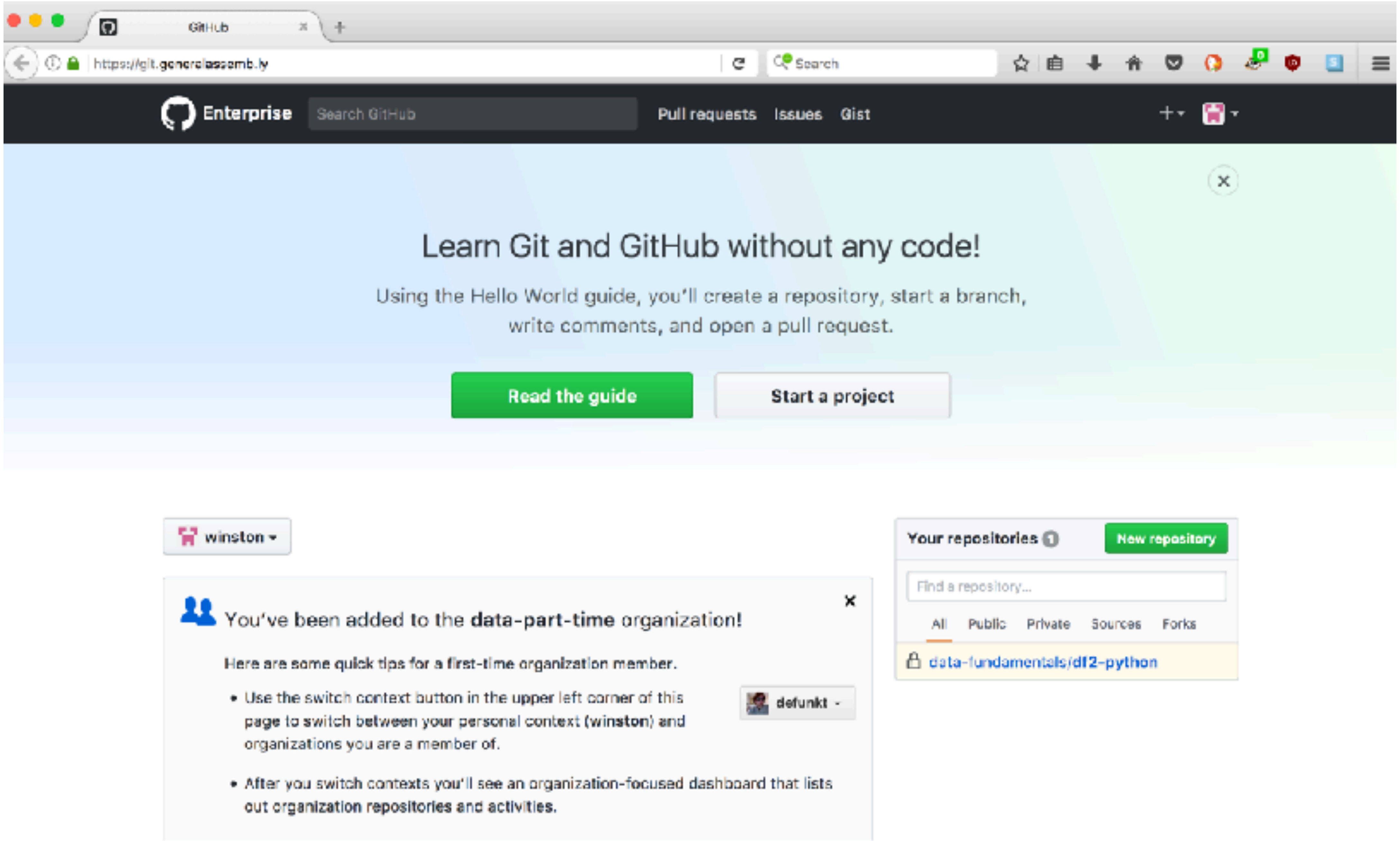
Ellen Kim 11:24 AM
set the channel purpose: Data Science Part Time Evening Course (09.18)

Darnody James 11:26 AM
joined #ds_pt_0918 by invitation from Ellen Kim, along with 17 others.

OUR TOOLSET - GIT

- Git
  - Create an account at http://git.generalassemb.ly/ and send Ellen your username.
  - You may have created an account at http://github.com/ too — that’s good, you’ll use it later.

OUR TOOLSET - GIT



Learn Git and GitHub without any code!

Using the Hello World guide, you'll create a repository, start a branch, write comments, and open a pull request.

You've been added to the **data-part-time** organization!

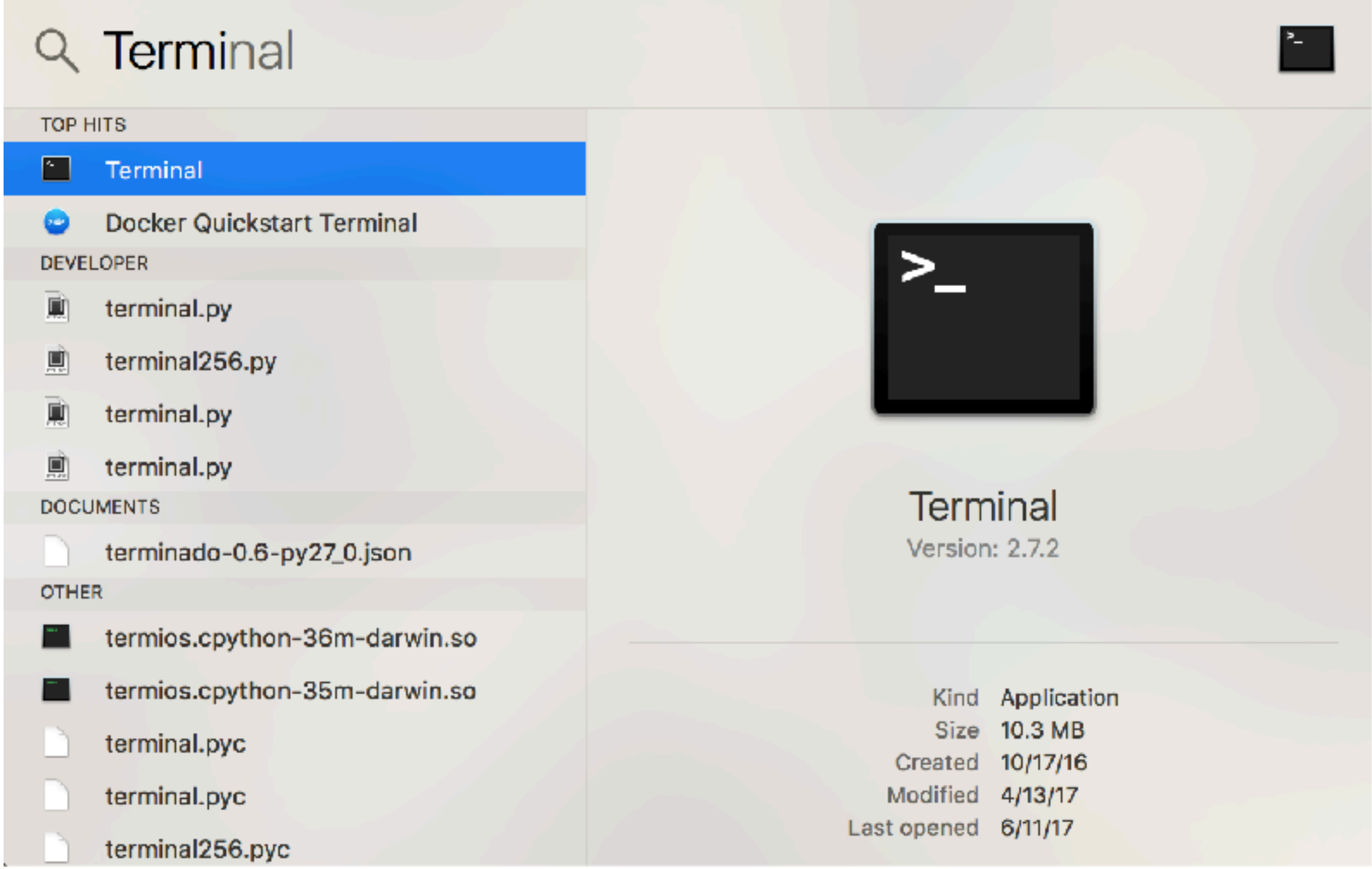
Here are some quick tips for a first-time organization member.

- Use the switch context button in the upper left corner of this page to switch between your personal context (winston) and organizations you are a member of.
- After you switch contexts you'll see an organization-focused dashboard that lists out organization repositories and activities.

OUR TOOLSET - THE COMMAND LINE

- The command line
  - OS X, *nix: search for Terminal
  - Windows: use Anaconda Prompt

OUR TOOLSET - THE COMMAND LINE



Terminal

TOP HITS

- Terminal
- Docker Quickstart Terminal

DEVELOPER

- terminal.py
- terminal256.py
- terminal.py
- terminal.py

DOCUMENTS

- terminado-0.6-py27_0.json

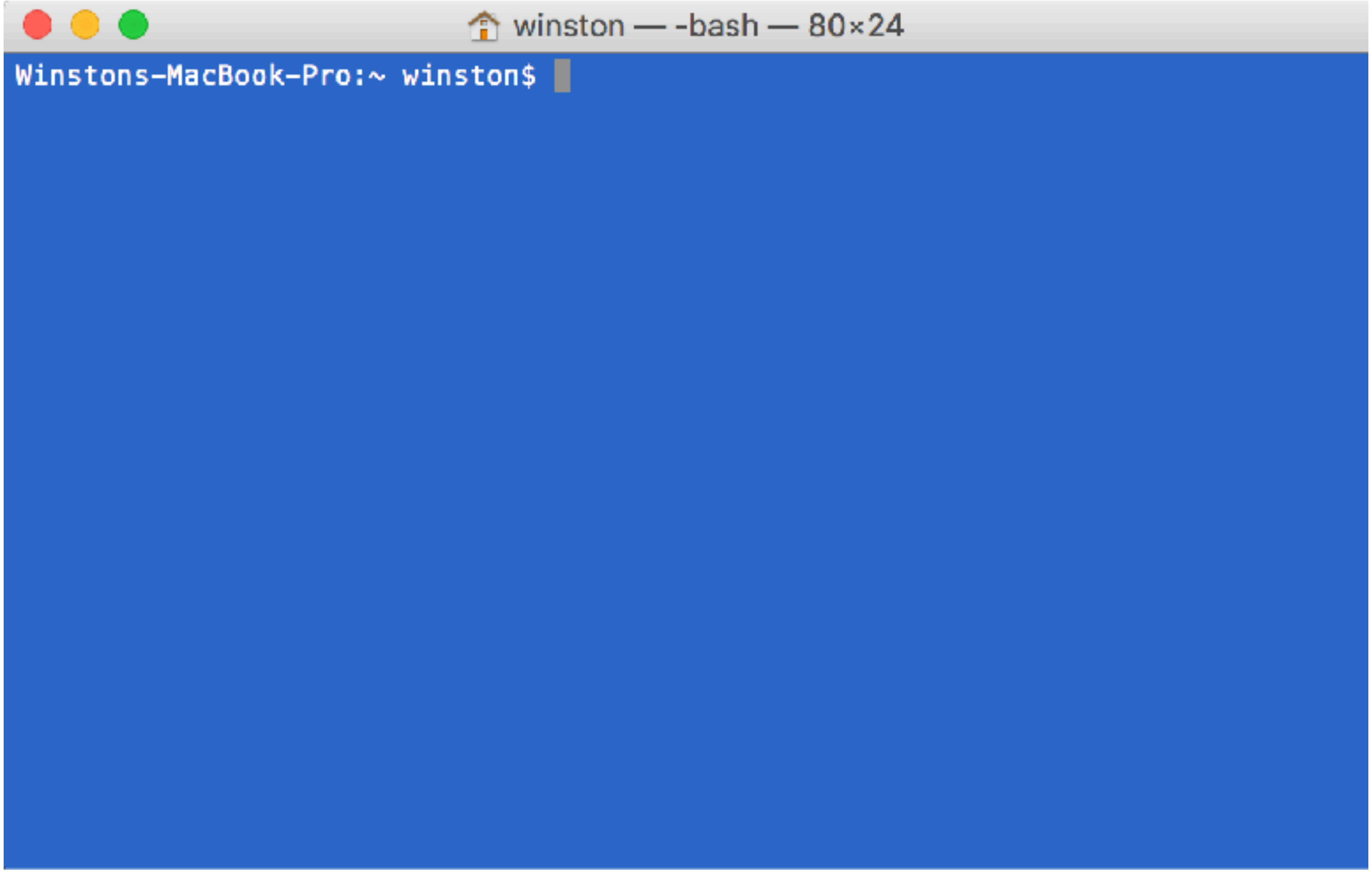
OTHER

- termios.cpython-36m-darwin.so
- termios.cpython-35m-darwin.so
- terminal.pyc
- terminal.pyc
- terminal256.pyc

Terminal
Version: 2.7.2

Kind	Application
Size	10.3 MB
Created	10/17/16
Modified	4/13/17
Last opened	6/11/17

OUR TOOLSET – THE COMMAND LINE

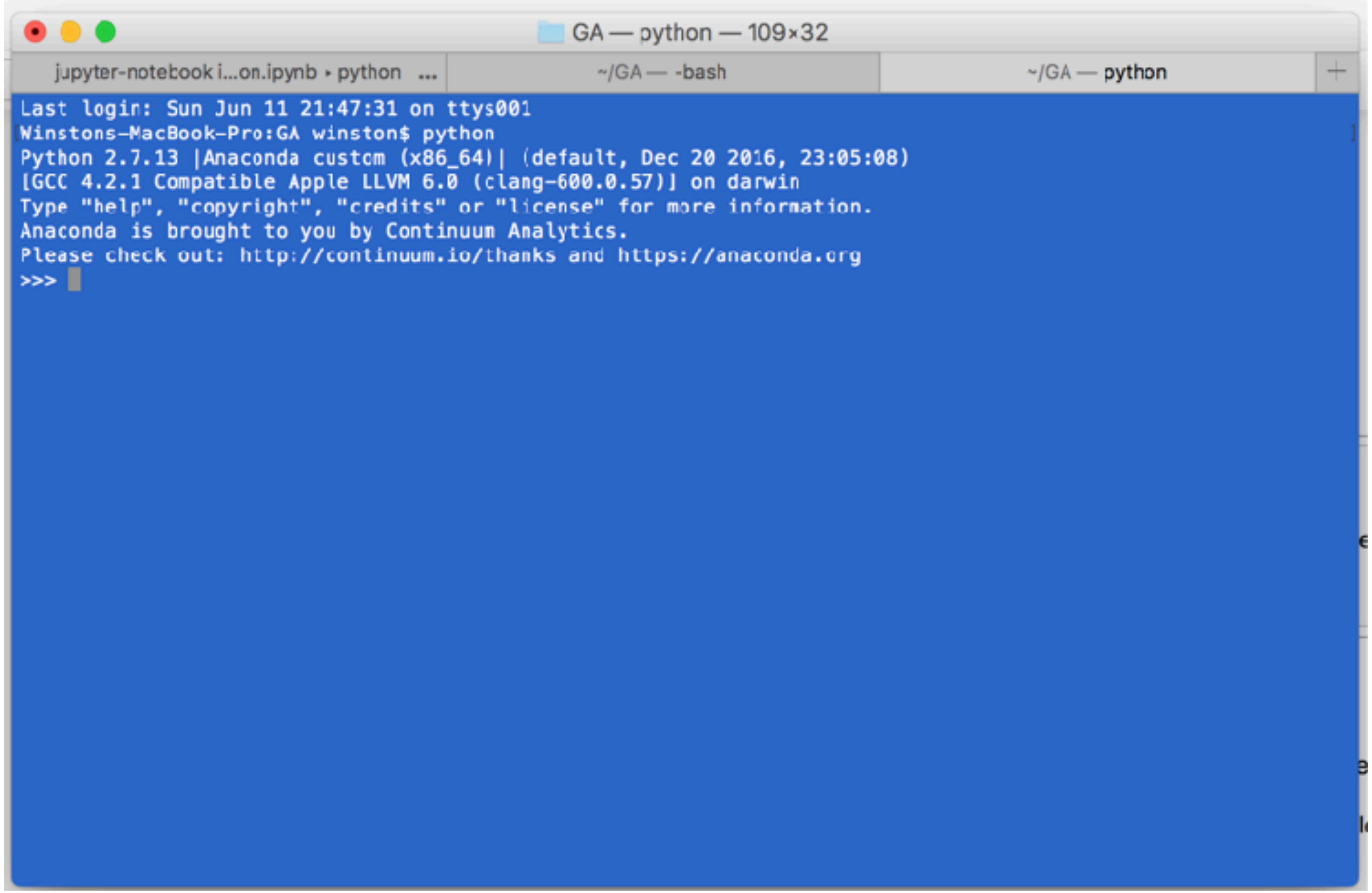


```
Winstons-MacBook-Pro:~ winston$ 
```

OUR TOOLSET - PYTHON

- Python 2.7
  - Just type ‘python’ at the command prompt
  - $ = command prompt
  - >>> = Python interpreter

OUR TOOLSET - PYTHON

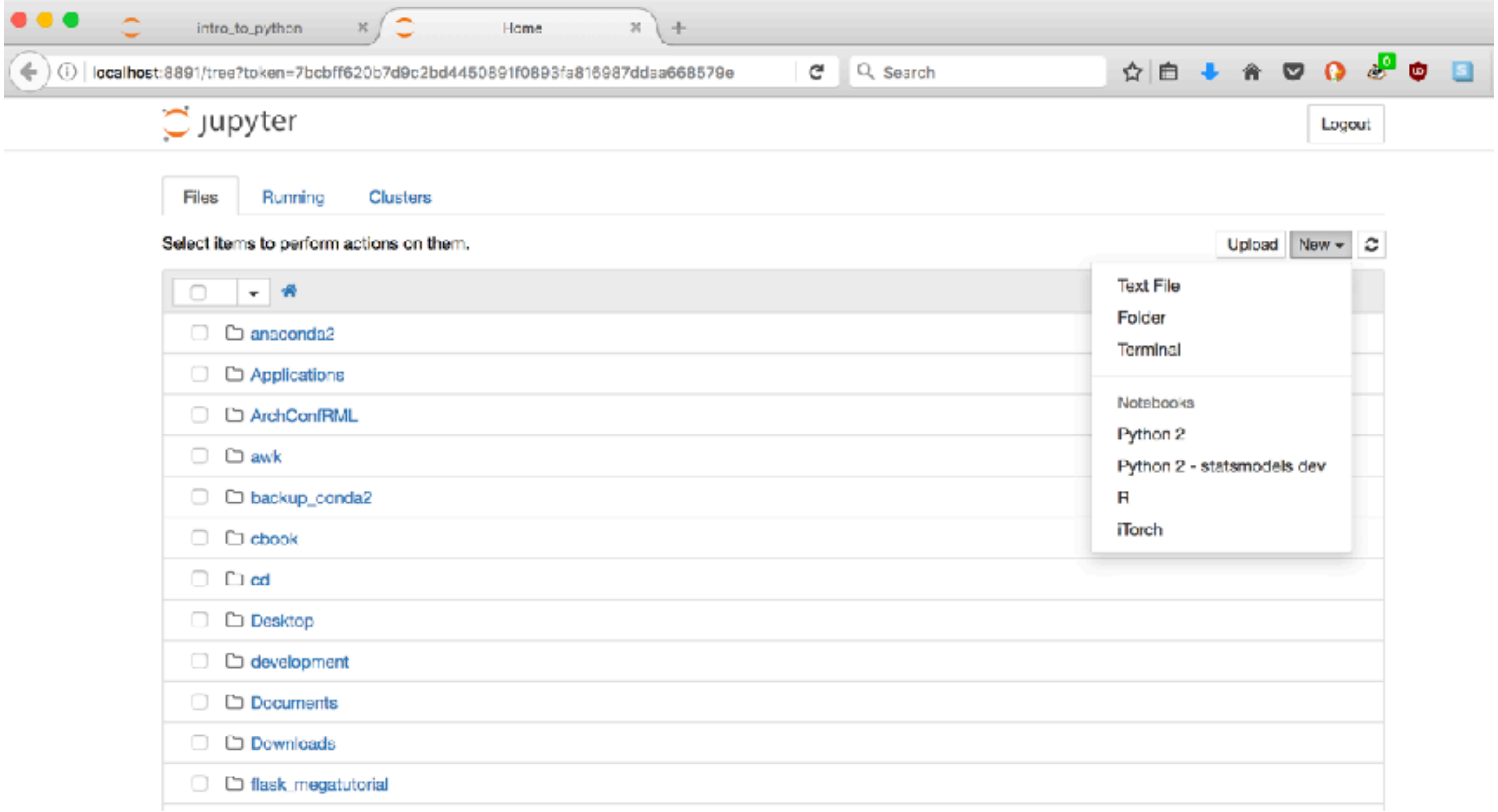


```
Last login: Sun Jun 11 21:47:31 on ttys001
Winstons-MacBook-Pro:GA winston$ python
Python 2.7.13 |Anaconda custom (x86_64)| (default, Dec 20 2016, 23:05:08)
[GCC 4.2.1 Compatible Apple LLVM 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://anaconda.org
>>> 
```

OUR TOOLSET - JUPYTER NOTEBOOKS

- Jupyter notebooks
  - Came with Anaconda
  - Great way of prototyping and sharing your code and analyses - after this class, most of our lessons will be in notebooks
  - With a browser open, type ‘jupyter notebook’ at the CLI, or open via the Anaconda GUI

OUR TOOLSET - JUPYTER NOTEBOOKS



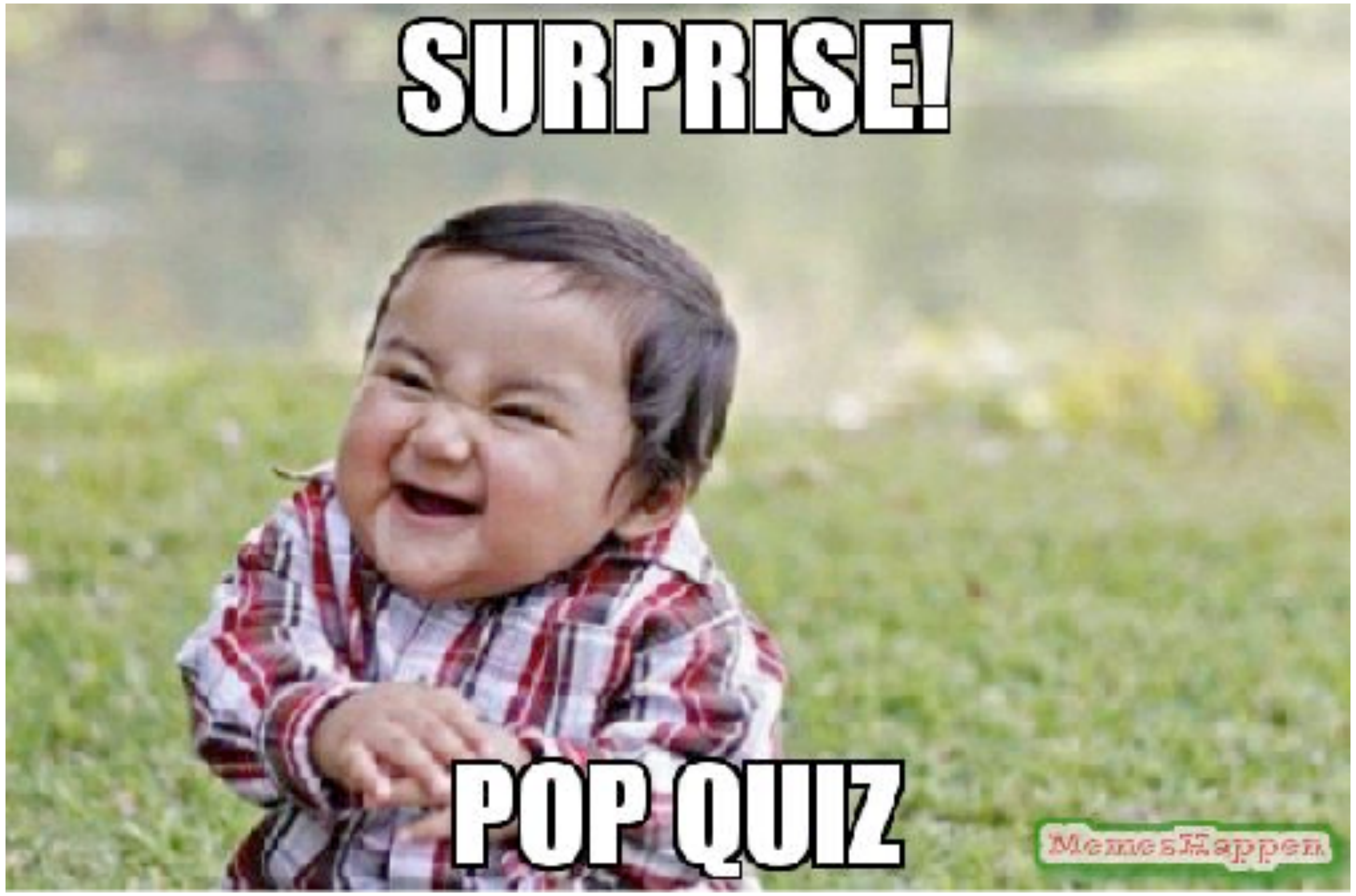
Select items to perform actions on them.

- anaconda2
- Applications
- ArchConfRML
- awk
- backup_conda2
- cbook
- cd
- Desktop
- development
- Documents
- Downloads
- flask_megatutorial

PYTHON FLASH QUIZ!

PYTHON QUIZ

- Create an account at www.HackerRank.com
- Go to https://www.hackerrank.com/dat-nyc click “Sign Up”, and start coding!



DATA SCIENCE @ GA

STUDENT PROFILE SURVEY

STUDENT PROFILE SURVEY

- Please take 10 minutes to fill out some information for the GA admins
    - (It includes important info like an emergency contact number)
    - Check Slack for the link!

DATA SCIENCE @ GA

Q&A

EXIT TICKET

EXIT TICKET

We’ll Slack you a link to the daily “exit ticket”.

This is your chance to give us rapid feedback on the lesson and the course. We’ll read these and try to address any questions or suggestions asap!

BYE!