

---

# WELCOME TO GA!

## ORIENTATION

---



---

# MEET YOUR INSTRUCTORS



---

## ABOUT ME

---

- PhD in EE (learning and control in systems with uncertainty), MS in Statistics
- Researcher at IBM Watson
- Chief Technology Officer at a Data Science Healthcare startup
- Love to travel, hike and take photos

Hi!

## STUDENT SERVICES

[StudentServicesLA@generalassemb.ly](mailto:StudentServicesLA@generalassemb.ly)

## STUDENT-FACING SUPPORT

**PAYMENTS  
+  
LOGISTICS**

**CAMPUS  
+  
TOOLS**

**COURSE  
COMPLETION  
STATUS**



# ABOUT GENERAL ASSEMBLY

---

**GENERAL ASSEMBLY'S MISSION IS  
TO BUILD OUR COMMUNITY BY  
TRANSFORMING MILLIONS OF  
THINKERS INTO **CREATORS.****

---

**GENERAL ASSEMBLY IS A GLOBAL  
COMMUNITY OF INDIVIDUALS  
EMPOWERED TO PURSUE THE  
WORK WE LOVE.**

---

# ALL YOU CAN LEARN BUFFET

---

## CLASSES & WORKSHOPS

Supplement and focus your learning during and after the course with a \$150 credit for 1 year from your start date

## STUDENT DISCOUNTS

Our students and alumni get free or discounted access to software and services from many vendors

## GA COMMUNITY

Expand your network at GA events, Meet-ups, and continue to collaborate and network with your classmates



# TOOLS



## FEEDBACK / SUPPORT

---

- Access to Instructional Team: office hours, in class support
- Exit Tickets
- Mid-Course Feedback
- End of Course Feedback



## Exit Tickets

### Welcome to Today's Exit Ticket

This template designed to help your instructor as well as the GA team to understand how you felt about the lesson and how effective the instruction was today. Your honest feedback will help you improve your learning experience for next time.

\* Required



Please, write your full name. \*

Course/Colori # \*

What's the lesson number? \*

If you're not sure, please check with your instructor or TA.

What was the topic of the lesson? \*

If you're not sure, please check with your instructor or TA.

My instructional team was effective in helping me achieve the learning objectives for this lesson. \*

1 2 3 4 5 6 7 8 9 10

Not Effective ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Highly Effective

The response was well-organized and relevant time was given for each activity. \*

1 2 3 4 5 6 7 8 9 10

Do Not Agree ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Strongly Agree

I feel prepared to continue practicing this skill outside of class. \*

1 2 3 4 5 6 7 8 9 10

Do Not Agree ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Strongly Agree



You've been added to the **generalassembly-studio** 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 (**malloryjacobs**) 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.



defunkt ▾



Welcome to GitHub! What's next? (15 days ago)

Create a repository

Tell us about yourself

Browse interesting repositories

Follow @github on Twitter

\_\_\_\_\_

# #ds-sm-15



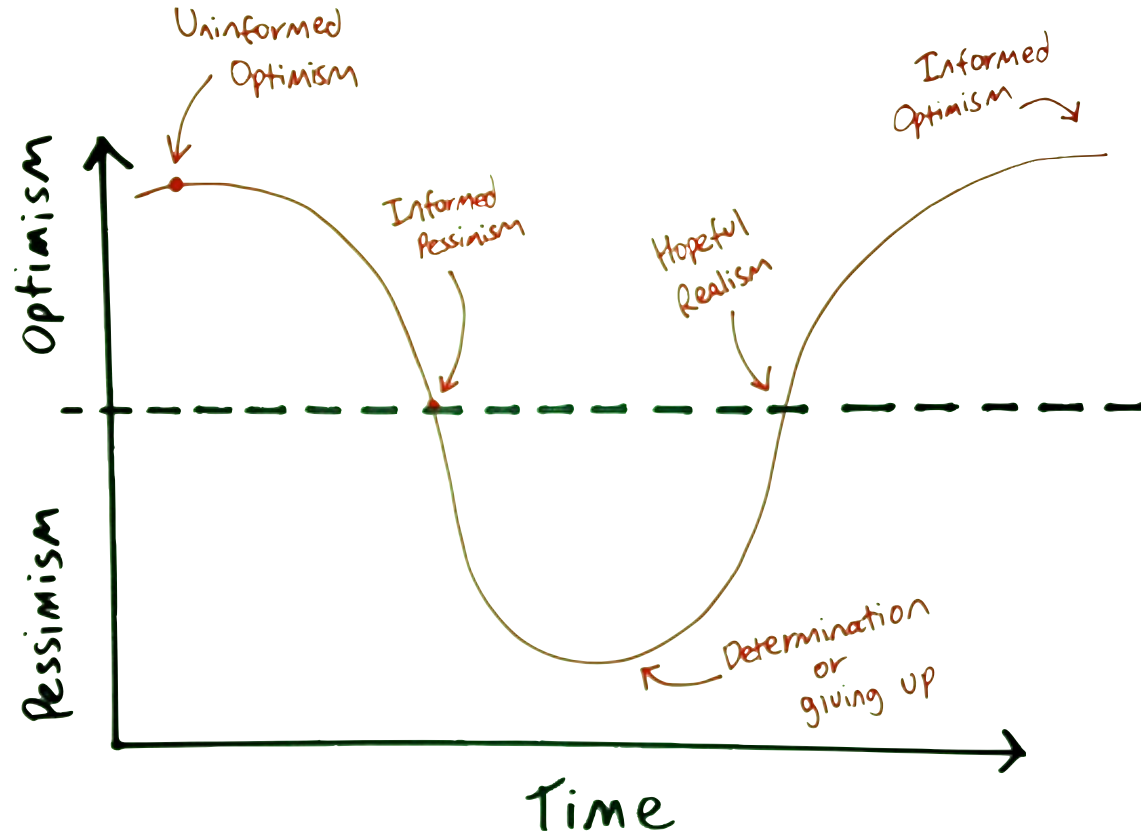
## Slack preferred over email for quick questions

---

# ROAD TO SUCCESS



# EMOTIONAL CYCLE OF CHANGE

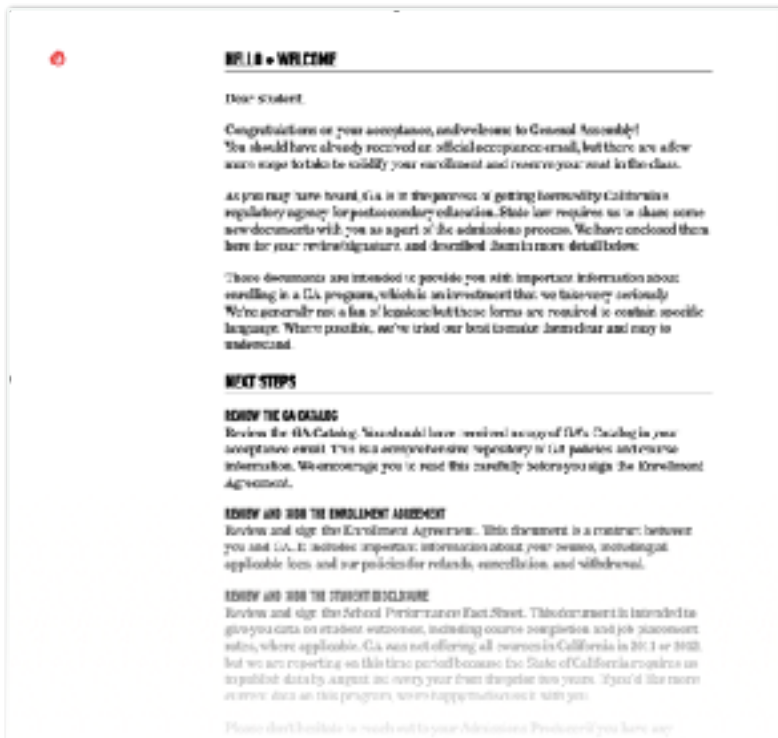


Don't Forget

# Have you signed your enrollment agreement?

## GA Catalog:

- ✓ Attendance Policy
- ✓ Cancellation/Withdrawal
- ✓ Tuition Payments
- ✓ Homework + Projects (80%)
- ✓ Letter of Completion







## GA Graduation Requirements

### **ASSIGNMENTS**

**(COMPLETE 80% OF  
WORK/LABS)**

### **ATTENDANCE**

**(MISS NO MORE THAN  
2 CLASSES)**

### **FINAL PROJECT**

### **COMMUNITY ENGAGEMENT**

**PARTICIPATION + FEEDBACK**



## Our Grads



### **BUILD YOUR NETWORK**

It's not just about altruism, your network is your most valuable asset. Alumni have started companies together and recruited other alumni to join their teams




### **ACCESS TO OUR ALUMNI COMMUNITY**

You're part of the alumni community forever



### **OFFICIAL LETTER OF COMPLETION**

Sent to your inbox via email after the course is completed. \*important if you're being reimbursed by your business



### **PERKS!** DISCOUNTS OFF CLASSES AND WORKSHOPS, TUITION CREDIT

We can't wait to have you back on campus

---

## HOW TO SUCCEED

---

- Ask
- Share
- Do
- Do
- Do

# ASK AWAY!



# WELCOME TO DATA SCIENCE

*Naumaan Nayyar*

*Chief Technology Officer, Vivace Systems*

## **INTRODUCTION**

---

# **INTRODUCTIONS**

---

## WELCOME TO DATA SCIENCE

---

# INTRODUCTIONS

- Who are you?
- Tell us about yourself
- Share your professional experiences
- What do you expect to get out of this course?

---

## WELCOME TO DATA SCIENCE

---

# LEARNING OBJECTIVES

- Describe the roles and components of a successful learning environment
- Define data science and the data science workflow
- Setup your development environment and review python basics



---

**DATA SCIENCE**

---

# PRE-WORK

---

## **PRE-WORK REVIEW**

---

- Define basic data types used in object-oriented programming
- Recall the Python syntax for lists, dictionaries, and functions
- Create files and navigate directories using the command line interface

## **INTRODUCTION**

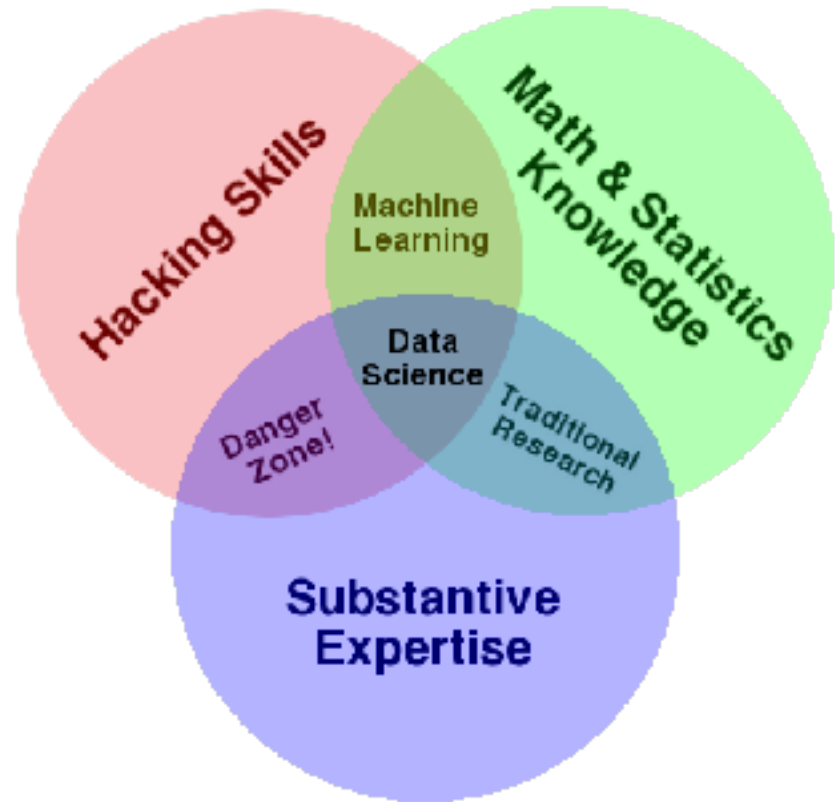
---

# **WHAT IS DATA SCIENCE?**

# WHAT IS DATA SCIENCE?

---

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



---

## WHAT IS DATA SCIENCE?

---

- ▶ Data Science
- ▶ Software Engineering
- ▶ Machine Learning (or Statistics... ?)

“...better statisticians than your average programmer and they’re better programmers than your average statistician”

## WHO USES DATA SCIENCE?

---

**NETFLIX**

**amazon.com**<sup>®</sup>

**Google**



 **FiveThirtyEight**



---

## WHO USES DATA SCIENCE?

---

► How does your company or industry use data science?

# WHAT ARE THE ROLES IN DATA SCIENCE?

► Data Science involves a variety of skill sets, not just one.

Business	ML / Big Data	Math / OR	Programming	Statistics
Product Development	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics



**QUIZ**

---

# **DATA SCIENCE BASELINE**

# ACTIVITY: DATA SCIENCE BASELINE QUIZ

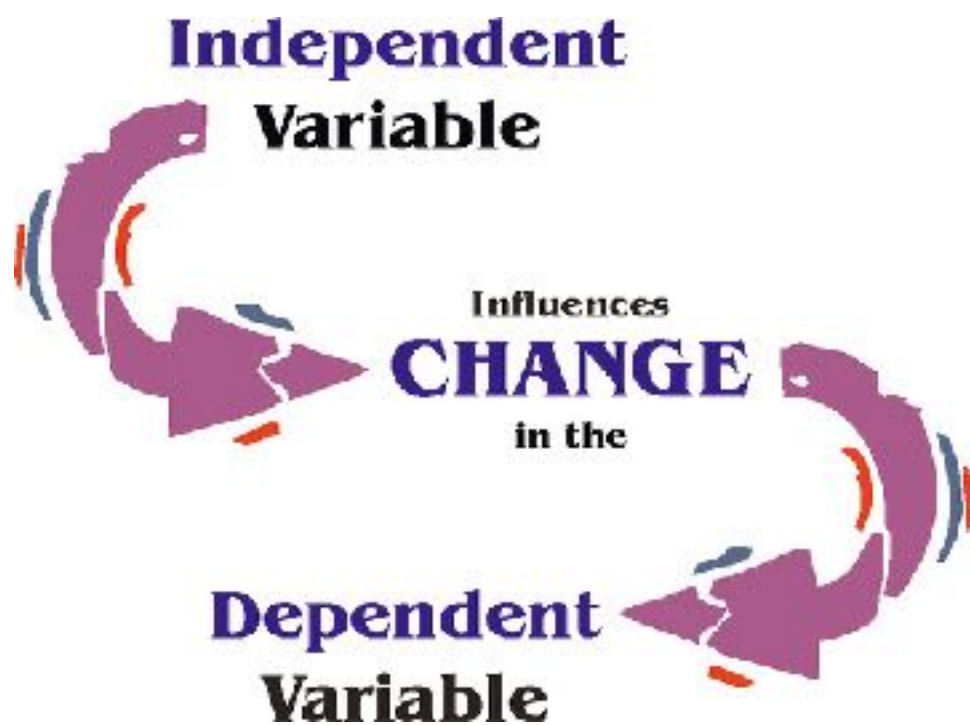
---



## DIRECTIONS (10 minutes)

1. Form groups of three.
2. Answer the following questions.
  - a. True or False: Gender (coded male=0, female=1) is a continuous variable.
  - b. Define the following:
    - i. Independent variable (Predictor)
    - ii. Dependent variable (Outcome)
    - iii. Covariate
  - c. Draw a normal distribution
  - d. True or False: Linear regression is an unsupervised learning algorithm.
  - e. What is a hypothesis test?

## ACTIVITY: DATA SCIENCE BASELINE QUIZ



I am what is  
The default, the status quo  
I am already accepted, can only be rejected  
The burden of proof is on the alternative  
**I am the null hypothesis**



## **INTRODUCTION**

---

# **THE DATA SCIENCE WORKFLOW**

---

## OVERVIEW OF THE DATA SCIENCE WORKFLOW

---

- A methodology for doing Data Science
- Similar to the scientific method
- Helps produce ? and ? results

---

## OVERVIEW 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

---



## **BUILD A DATA MODEL**

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

1. Problem definition 2. Data collection 3. Data cleaning 4. Exploratory data analysis 5. Feature engineering 6. Model selection 7. Model training 8. Model evaluation 9. Model deployment 10. Model monitoring

# 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

## FUTURAMA EXAMPLE

---

- ▶ **Problem Statement:** “Using Planet Express customer data from January 3001-3005, determine how likely previous customers are to request a repeat delivery using demographic information (profession, company size, location) and previous delivery data (days since last delivery, number of total deliveries).”



- ▶ We can use the Data Science workflow to work through this problem.

---

## **FUTURAMA EXAMPLE: IDENTIFY THE PROBLEM**

---

- ▶ Identify the business/product objectives.
- ▶ Identify and hypothesize goals and criteria for success.
- ▶ Create a set of questions to help you identify the correct data set.



---

## **FUTURAMA EXAMPLE: ACQUIRE THE DATA**

---

- ▶ Ideal data vs. data that is available
- ▶ Learn about limitations of the data.
- ▶ What data is available for this example?
- ▶ What kind of questions might we want to ask about the data?

---

## **FUTURAMA EXAMPLE: ACQUIRE THE DATA**

---

- ▶ Questions to ask about the data

- ▶ Is there enough data?

- ▶ Does it appropriately align with the question/problem statement?

- ▶ Can the dataset be trusted? How was it collected?

- ▶ Is this dataset aggregated? Can we use the aggregation or do we need to get it pre-aggregated?

---

## FUTURAMA EXAMPLE: PARSE THE DATA

---

- ▶ Secondary data = we didn't directly collect it ourselves
- ▶ Example data dictionary

Variable	Description	Type of Variable
Profession	Title of the account owner	Categorical
Company Size	1- small, 2- medium, 3- large	Categorical
Location	Planet of the company	Categorical
Days Since Last Delivery	Integer	Continuous
Number of Deliveries	Integer	Continuous

---

## FUTURAMA EXAMPLE: PARSE THE DATA

---

- ▶ Questions to ask while parsing

- ▶ Is there documentation for the data? Is there a data dictionary?
- ▶ What kind of filtering, sorting, or simple visualizations can help understand the data?
- ▶ What information is contained in the data?
- ▶ What data types are the variables?
- ▶ Are there outliers? Are there trends?

---

## **FUTURAMA EXAMPLE: MINE THE DATA**

---

- Think about sampling
- Get to know the data
- Explore outliers
- Address missing values
- Derive new variables (i.e. columns)

---

## **FUTURAMA EXAMPLE: MINE THE DATA**

---

- Common steps while mining the data
  - Sample the data with appropriate methodology
  - Explore outliers and null values
  - Format and clean the data
  - Determine how to address missing values
  - Format and combine data; aggregate and derive new columns

---

## FUTURAMA EXAMPLE: REFINE THE DATA

---

- ▶ Use statistics and visualization to identify trends
- ▶ Example of basic statistics

Variable	Mean (STD) or Frequency (%)
Number of Deliveries	50.0 (10)
Earth	50 (10%)
Amphibios 9	100 (20%)
Bogad	100 (20%)
Colgate 8	100 (20%)
Other	150 (30%)

---

## **FUTURAMA EXAMPLE: REFINE THE DATA**

---

- Descriptive stats help refine by
  - Identifying trends and outliers
  - Deciding how to deal with outliers
  - Applying descriptive and inferential statistics
  - Determining visualization techniques for different data types
  - Transforming data



---

## **FUTURAMA EXAMPLE: CREATE A DATA MODEL**

---

- Select a model based upon the outcome
- Example model statement: “We completed a logistic regression using Statsmodels v. XX. We calculated the probability of a customer placing another order with Planet Express.”
- Steps for model building

---

## **FUTURAMA EXAMPLE: CREATE A DATA MODEL**

---

- ▶ The steps for model building are
  - ▶ Select the appropriate model
  - ▶ Build the model
  - ▶ Evaluate and refine the model
  - ▶ Predict outcomes and action items

---

## **FUTURAMA EXAMPLE: PRESENT THE RESULTS**

---

- ▶ You have to effectively communicate your results for them to matter!
- ▶ Ranges from a simple email to a complex web graphic.
- ▶ Make sure to consider your audience.
- ▶ A presentation for fellow data scientists will be drastically different from a presentation for an executive.

---

## FUTURAMA EXAMPLE: PRESENT THE RESULTS

---

- ▶ Key factors of a good presentation include
  - ▶ Summarize findings with narrative and storytelling techniques
  - ▶ Refine your visualizations for broader comprehension
  - ▶ Present both limitations and assumptions
  - ▶ Determine the integrity of your analyses
  - ▶ Consider the degree of disclosure for various stakeholders
  - ▶ Test and evaluate the effectiveness of your presentation beforehand

---

## FUTURAMA EXAMPLE: PRESENT THE RESULTS

---

- ▶ Example presentations and infographics
  - ▶ [512 Paths to the White House](#)
  - ▶ [Who Old Are You?](#)
  - ▶ [2015 NFL Predictions](#)

**GUIDED PRACTICE**

---

# **DATA SCIENCE WORK FLOW**

---

## ACTIVITY: DATA SCIENCE WORKFLOW (TIME PERMITTING)

---



### EXERCISE

#### **DIRECTIONS (25 minutes)**

1. Divide into 4 groups, each located at a whiteboard.
2. **IDENTIFY:** Each group should develop 1 research question they would like to know about their classmates. Create a hypothesis to your question. Don't share your question yet! (5 minutes)
3. **ACQUIRE:** Rotate from group to group to collect data for your hypothesis. Have other students write or tally their answers on the whiteboard. (10 minutes)
4. **PRESENT:** Communicate the results of your analysis to the class. (10 minutes)
  - a. Create a narrative to summarize your findings.
  - b. Provide a basic visualization for easy comprehension.
  - c. Choose one student to present for the group.

#### **DELIVERABLE**

Presentation of the results

**DEMO**

---

# ENVIRONMENT SETUP



---

# DEV ENVIRONMENT SETUP

---

- Brief intro of tools
- Environment setup
  - Create a Github account
  - Install Python 2.7 and Anaconda
  - Practice Python syntax, Terminal commands, and Pandas
- iPython Notebook test and Python review

---

## DEV ENVIRONMENT SETUP

---

- ▶ Test your new setup using the lesson 1 starter code available at / *lessons/lesson-01/code/starter-code/* in the Github repo
- ▶ Ask your classmates and instructor for help if you have problems!

---

**CONCLUSION**

---

**REVIEW**

---

## CONCLUSION

---

- ▶ 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?

**DATA SCIENCE**

---

**BEFORE NEXT CLASS**

---

## BEFORE NEXT CLASS

---

- ▶ Complete Learn Python the Hard Way through Exercise 35, or as far as you have time
  - ▶ Remember the Python syntax for lists, dictionaries, and functions
- ▶ Create files and navigate through the CLI. GA Tutorial
- ▶ Go through Learn Pandas (up to Lesson 3)

---

## RESOURCES

---

- Python quick reference
- Probability and Statistics refresher

---

**WELCOME TO DATA SCIENCE**

---

**Q & A**



**WELCOME TO DATA SCIENCE**

---

# EXIT TICKET

**DON'T FORGET TO FILL OUT YOUR EXIT TICKET**