GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE

Rob Kelly

WELCOME TO DATA SCIENCE
Who am I?

- Data Scientist at Booz Allen Hamilton
- Work on environmental impact modeling for the FAA
- Mainly code in Golang, Javascript and Python

WELCOME TO DATA SCIENCE

Who are you?

- Who are you, what is your background, why are you doing this?
- Coding/Technical Background

WELCOME TO DATA SCIENCE
Agenda

- Overview of Course
  - (HOLIDAYS – 11/26, 12/24, 12/31)
- Learning Objectives
- Get setup on Slack
- Final Projects and Unit Projects

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
- Work on Github and Python fundamentals

QUIZ

STATISTICS WARMUP

ACTIVITY: DATA SCIENCE BASELINE QUIZ

DIRECTIONS (10 minutes)

1. True or False: Gender (coded male=0, female=1) is a continuous variable.
2. True or False: Linear regression is an unsupervised learning algorithm.
3. Elvis Presley had a twin brother who died at birth. What is the probability that Elvis was an identical twin? Historically, approximately 1/125 of all births were fraternal twins and 1/300 were identical twins.
4. Average soccer game has a goal every 4 minutes, odds of 6 goals in 8 minutes
5. Probability of getting heads 4 out of 6 on a coin biased to get heads 30% of the time

ACTIVITY: DATA SCIENCE BASELINE QUIZ

DIRECTIONS (10 minutes)

1. I personally think it's continuous, but it's a common debate
2. True, Regression implies Supervised Learning
3. 5/3000 are boy/boy identical twins, 6/3000 are boy-boy fraternal twins, so 5/11 is the odds that he was an identical twin.
4. You expect 2 goals in 8 minutes, so a Poisson distribution would be (math.e**-2 * 2**6) / math.factorial(6)
5. math.factorial(6) / (math.factorial(4) * 2) * .3**4 * .7**2

ACTIVITY: BAYESIAN BASELINE

MacKay's blood type problem

Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type O blood. The blood groups of the two traces are found to be of type O (a common type in the local population, having frequency 60%) and of type AB (a rare type, with frequency 1%). Do these data (the blood types found at the scene) give evidence in favor of the proposition that Oliver was one of the two people whose blood was found at the scene?

ACTIVITY: DATA SCIENCE BASELINE QUIZ

EXERCISE

S = the probability of the suspect and one random person
S' = two random people
D = the evidence

$$\frac{P(D \mid S)}{P(D \mid S')} = \frac{.01}{.01 * .6 * 2} = \frac{1}{1.2} = .83$$

WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

- "Essentially, **all models are wrong, but some are useful**." --- Box, George E. P.; Norman R. Draper (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley. ISBN 0471810339

- Can you name an example?

WHAT IS DATA SCIENCE?

- No Free Lunch Theorem
  - "Any two optimization algorithms are equivalent when their performance is averaged across all possible problems”

- In other words, averaged over all datasets of size n and uniformly averaged over all targets, the difference in error of any two algorithms is 0.

- All models are equally bad when considered against the set of all Pattern Recognition problems ☺

WHAT IS ARTIFICIAL INTELLIGENCE?

- “We deliberate not about ends, but about means. For a doctor does not deliberate whether he shall heal, nor an orator whether he shall persuade, nor a statesman whether he shall produce law and order, nor does any one else deliberate about his end. They assume the end and consider how and by what means it is attained, and if it seems easily and best produced thereby;”
  - - Aristotle, Nicomachean Ethics (Book III. 3, 1112b)

- What does this mean to you? How does it relate to Data Science?

WHAT IS ARTIFICIAL INTELLIGENCE?

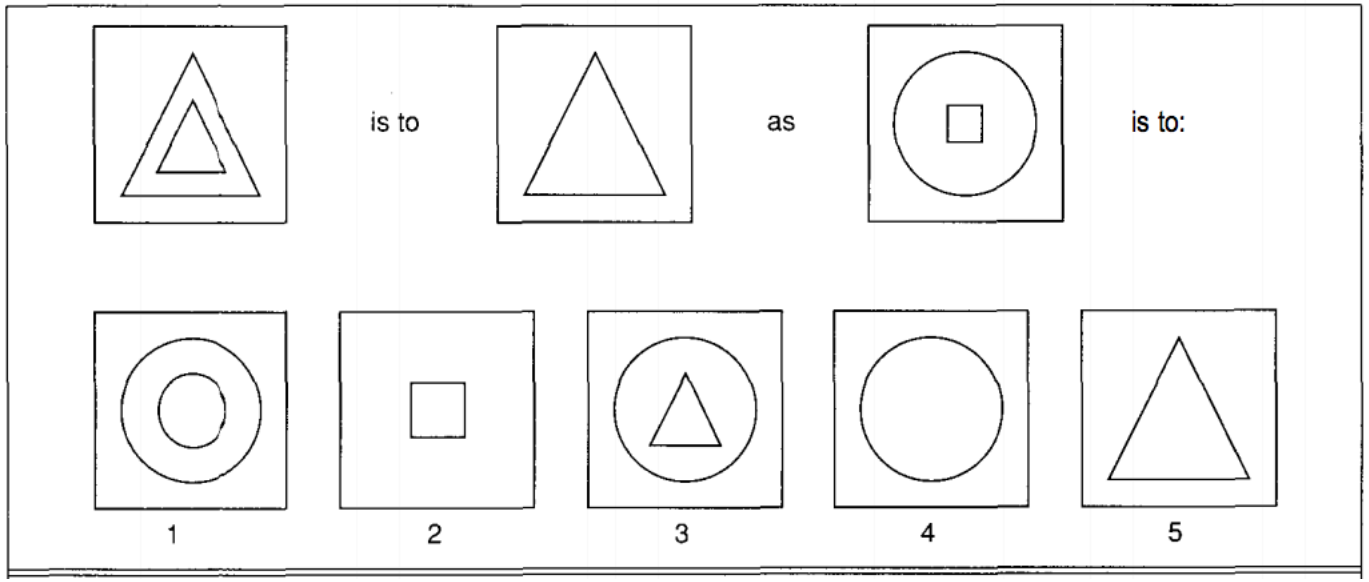
- To Achieve Artificial Intelligence, we need an “artifact” and intelligence
- A Turing Machine is a theoretical artifact designed to enable this intelligence
- You can't make a mechanical computing device that does anything that a Turing machine can't do
- This is known as Turing Equivalence

WHAT IS ARTIFICIAL INTELLIGENCE?

- In 1967, computers were solving problems like these:
  - Daniel Bobrow's STUDENT program

    If the number of customers Tom gets is twice the square of 20 percent of the number of advertisements he runs, and the number of advertisements he runs is 45, what is the number of customers Tom gets?

  - Tom Evans's ANALOGY program (1968)



WHAT IS ARTIFICIAL INTELLIGENCE?

- What they found was these sorts of solutions did not scale
- A program that was extremely effective at a very narrow type of problem would not generalize

- For example, Deep Blue won its first game against a world champion on February 10, 1996. This was 30 years later.

WHAT IS ARTIFICIAL INTELLIGENCE?

- Why did this happen?
- Combinatorial EXPLOSION!
- Quiz:
  - How many possible 10 X 10 pixel images are there?

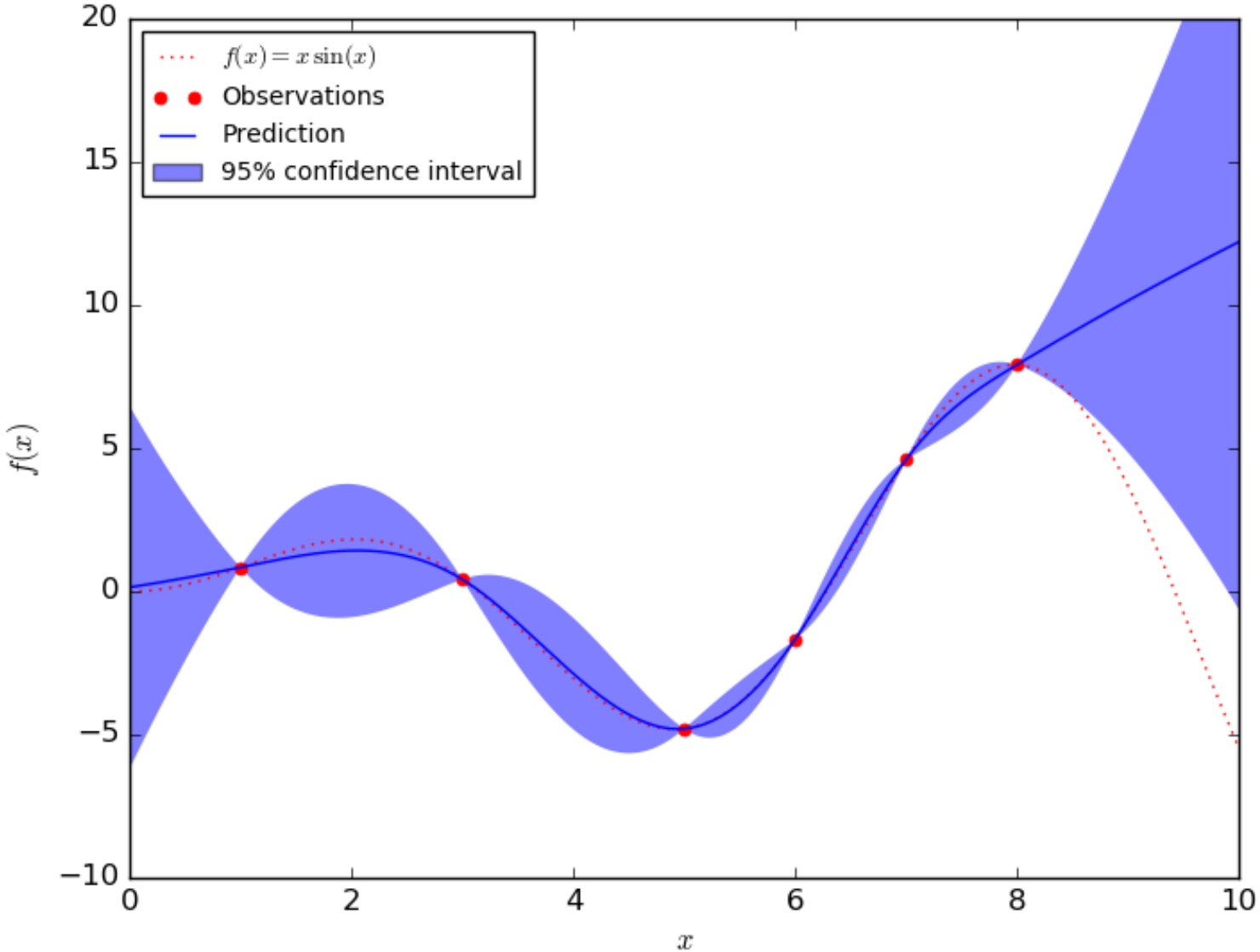
CONCLUSION

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

WHAT IS ARTIFICIAL INTELLIGENCE?

- Probability is a tool we use to cope with these crisis

- How can probability alleviate some of our concerns?


$f(x) = x\sin(x)$
Observations
Prediction
95% confidence interval
20
15
10
5
0
−5
−10
0
2
4
6
8
10
$f(x)$
x

GITHUB, ANACONDA, SLACK

EXERCISE

CODING EXERCISE

THE DATA SCIENCE WORKFLOW

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

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

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

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

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

CONCLUSION

REVIEW