WELCOME TO DATA SCIENCE

Abbas Chokor, Ph.D.

Staff Data Scientist, Seagate Technology

WELCOME TO DATA SCIENCE

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	What is Data Science	Lesson 1
	Research Design and Pandas	Lesson 2
	Statistics Fundamentals I	Lesson 3
	Statistics Fundamentals II	Lesson 4
	Flexible Class Session	Lesson 5
UNIT 2: FOUNDATIONS OF DATA MODELING	Introduction to Regression	Lesson 6
	Evaluating Model Fit	Lesson 7
	Introduction to Classification	Lesson 8
	Introduction to Logistic Regression	Lesson 9
	Communicating Logistic Regression Results	Lesson 10
	Flexible Class Session	Lesson 11
UNIT 3: DATA SCIENCE IN THE REAL WORLD	Decision Trees and Random Forests	Lesson 12
	Natural Language Processing	Lesson 13
	Dimensionality Reduction	Lesson 14
	Time Series Data I	Lesson 15
	Time Series Data II	Lesson 16
	Database Technologies	Lesson 17
	Where to Go Next	Lesson 18
	Flexible Class Session	Lesson 19
	Final Project Presentations	Lesson 20

WELCOME TO DATA SCIENCE

LEARNING OBJECTIVES

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

WELCOME TO DATA SCIENCE

AGENDA

- Welcome to GA: Meet & Greet
- Course pre-work
- What's data science?
- The data science workflow
- Environment setup
- Case study: NYC traffic analysis

DATA SCIENCE

WELCOME TO GA: Meet & Greet!

WELCOME TO GA!

- General Assembly is a global community of individuals empowered to pursue the work we love.

- General Assembly's mission is to build our community by transforming millions of thinkers into creators.

FOREVER AND EVER

It's not just about altruism, your network is your most valuable asset

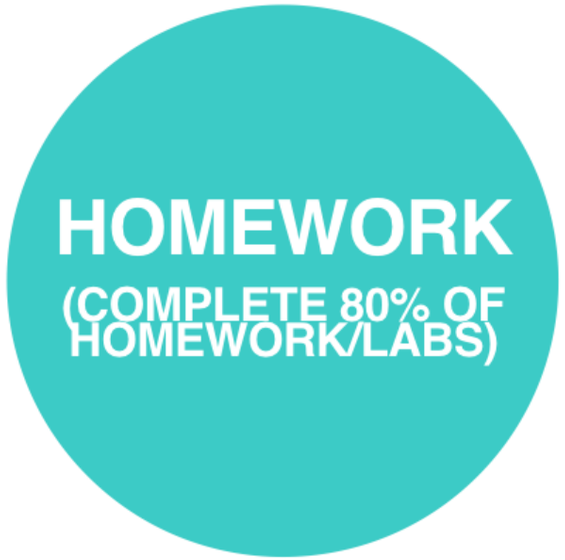
Alumni have started companies together and recruited other alumni to join their teams

You're part of the alumni community forever

We can't wait to have you back on campus

GA GRADUATION REQUIREMENTS

GA GRADUATION REQUIREMENTS



- Like any discipline, data science is a requires extensive practice in order to gain understanding of the concepts.
- It is imperative to submit complete homeworks in a timely manner to stay caught up in this course.
- Homework must be received on the day it is due in order to be eligible to receive full credit and consideration.
- Homework can be submitted to DAT-DEN-03 repository

GA GRADUATION REQUIREMENTS

- Attendance will be always tracked.
- You can miss up to 2 classes. But, what if it's an urgent circumstance?
- You have access to the building only 1 hour before the class and ***occasionally*** on weekends.
- The class is restricted to registered students.

GA GRADUATION REQUIREMENTS

- Your final project is a major step towards building a career in data science.
- You will work with your data! It's your project, it's your portfolio.
- I will have 1:1 meetings with each one of you to know more about your interests and come up with a plan for your project.

GA GRADUATION REQUIREMENTS

- Office hours: Tuesday and Thursday (5:00-6:30 pm)
- Join Slack - communication platform for students, alumni and instructors to connect.
- We value your feedback:
  - ❖ Exit tickets: classroom tool used to better understand your real-time learning
  - ❖ Mid-course & End of course feedback

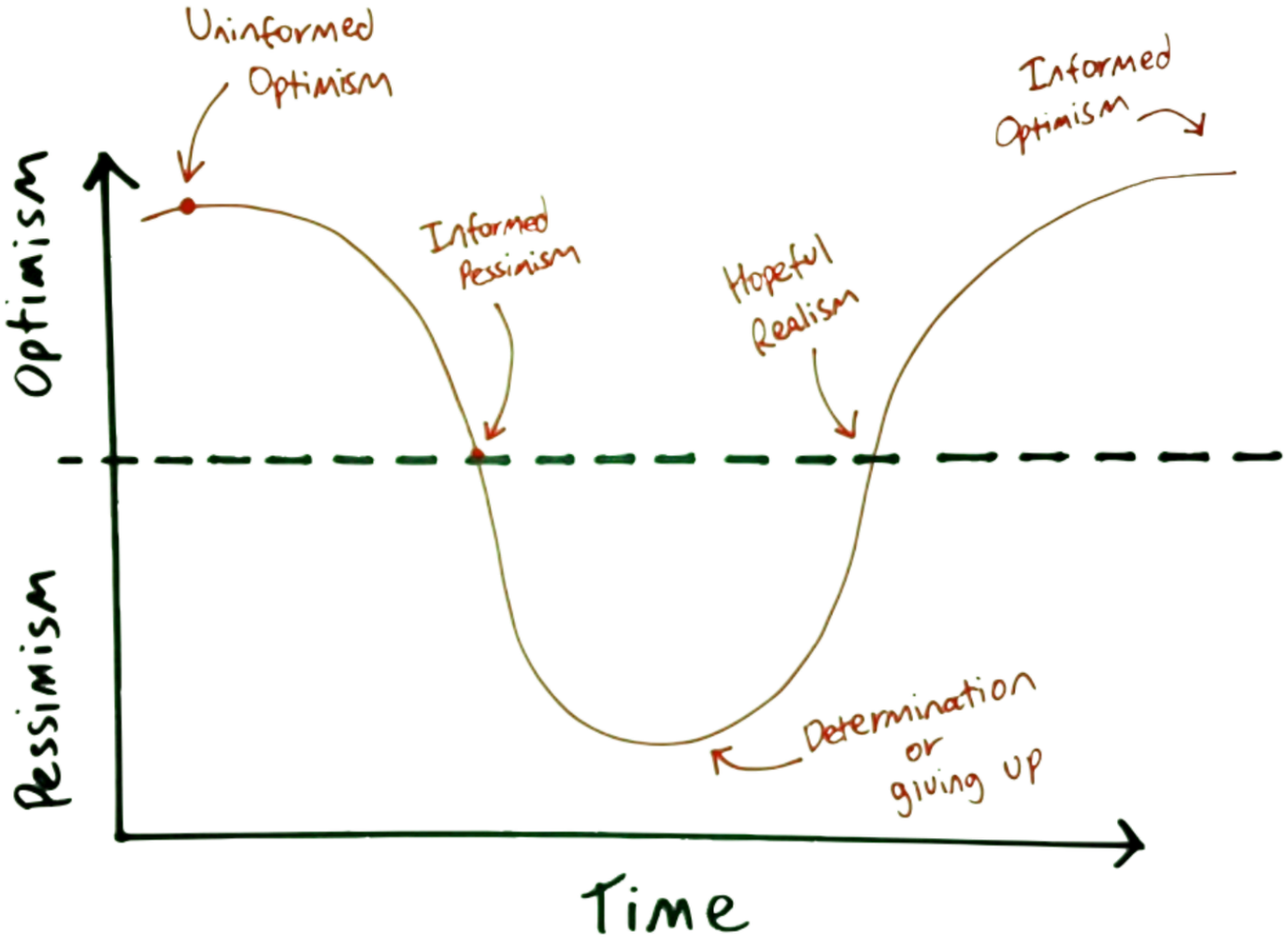
EMOTIONAL CYCLE OF CHANGE

HOMEWORK (COMPLETE 80% OF HOMEWORK/LABS)

ATTENDANCE (MISS NO MORE THAN 2 CLASSES)

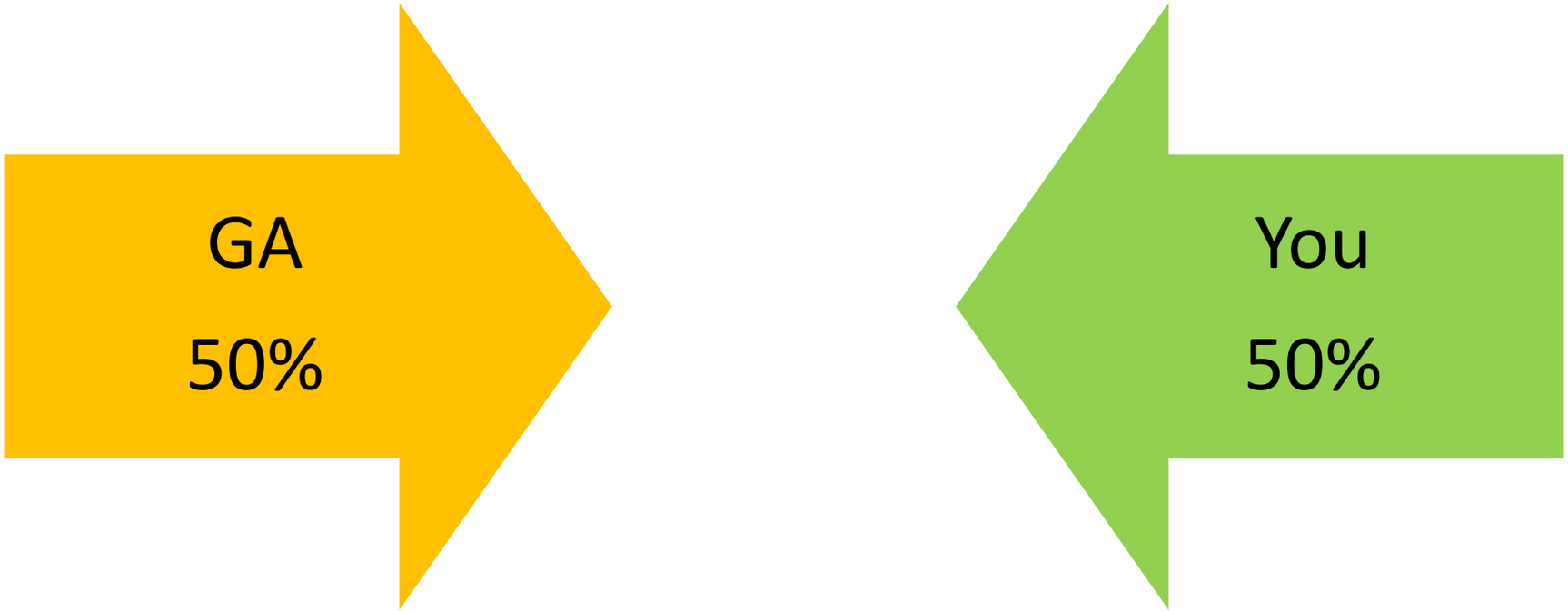
FINAL PROJECT

COMMUNITY ENGAGEMENT PARTICIPATION + FEEDBACK



STUDENT RESPONSIBILITIES

As a self directed program, we view students as a crucial part of the skill acquisition process.



INTRODUCE YOURSELF: WHO AM I?

Energy Consumption of Complex Facilities

Soiling Analysis in PV Systems

Management Team Member

Science Policy

Behavioral Change Effect on Patients Health

Safety Reports Analysis

IEQ and Productivity of Occupants

Social Media Mining for Adverse Drug Reactions

Teaching Engineering Concepts for non-Engineering Students

PHM for O&G Surface Machines

Incentives in U.S. Construction

PHM for Railcars

Civil Engineering Training

PHM & Energy Forecasting for PV Systems

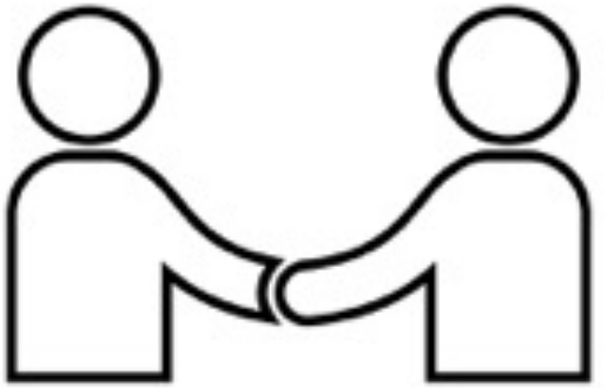
Risk Assessment for Contractors in Beirut Area

Edge Analytics & Video Surveillance

INTRODUCE YOURSELF: WHO ARE YOU?

Write on a piece of paper:

1. What are you up to these days?
2. What's your background/domain?
3. Why are you interested in this course?
4. What is your guilty pleasure?



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

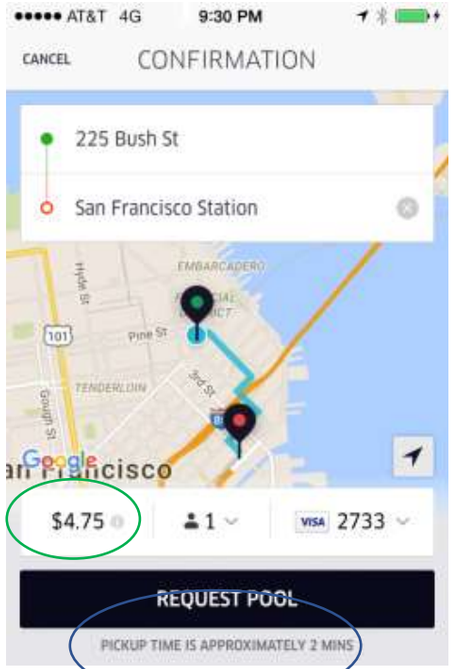
INTRODUCTION

WHAT IS DATA SCIENCE?

DATA SCIENCE: THE SEXIEST JOB OF 21st CENTURY

DATA SCIENCE IS ALL AROUND US

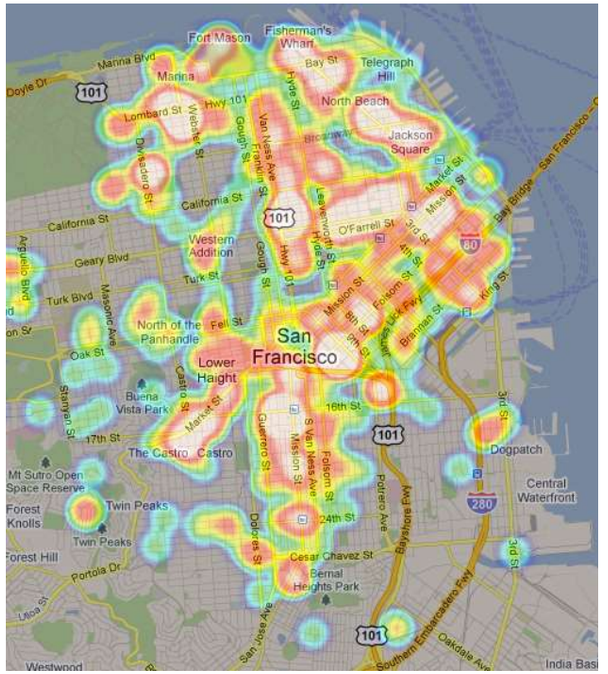
Rider



Fare Estimator and Dynamic Pricing (e.g., Surge Pricing) Algorithms

ETA (Estimated Time of Arrival) Algorithms

Driver



Heat maps for drivers on where to best position themselves in the city

DATA SCIENCE IS ALL AROUND US

Michael Jones	**THIS IS FOR YOUR ATTENTION.** - Sir/Madam, I am a Certified Public Accountant officer in my Bank so also account manager to a deceased customer who was a contractor with a pro	Nov 1
920ict	**Final Submission Due: Nov. 30, 2017** - Int'l Conference on Thermology (ICT 2018) January 13-15, 2018 Sanya, China Dear Colleagues, Int'l Conference on Thermology (ICT 2018) wil	Oct 31
Sara Maxwell	**Greetings.....** - Greetings Beloved in Christ, Greetings in the name of our Lord Jesus Christ,How are you doing hope you're having a wonderful weekend so far,I hope you don't really mi	Oct 31
Jana Wafa	**Urgent Help!** - Hello Dear, Sorry for any inconvenience, but I'm in a terrible situation. I came down here to kiev,Ukraine for a program, last night on our way back to the hotel room we wa	Oct 31
Ryan	**We sincerely invite you to be a Paper Reviewer-Welcome to submit your papers** - ESEM 2 0 1 8 ▶Invitation Letter Technical Program Committee Dear Professor, Wish you a nice (	Oct 31
Quentina	**The CMEE2017 first session has been a complete success|Second:Xiamen,12/24-25** - CMEE2017 INVITATION The 2nd International Conference on Computer, Mechetronics and I	Oct 30
abbas.chokor	**Re: unsubscribe NOW..** - To STOP receiving these emails from us Just hit relpy and let us know Thanks,	Oct 29
Rida Elias	**Account Shutdown**Last Warning**** - This is to notify all Students, Staffs in American University of Beirut that we are validating active accounts. Kindly confirm that your account (aic0	Oct 28
Hassan Baydoun	**Account Shutdown**Last Warning**** - This is to notify all Students, Staffs in American University of Beirut that we are validating active accounts. Kindly confirm that your account (aic0	Oct 28
Vivian	**Take the chance,December 24-25,discuss hot issues with colleague at 2017WCNE** - Hope to receive your "Wireless Communication|Network Engineering" paper The second intern	Oct 27
Kelsey	**Share your latest researches-Grasp the opportunity** - MENU Official Website Call For Papers Paper Submission About Shanghai 2017 Shanghai: the Location of International Semin	Oct 26
Ali Zein (Student)	**Account Shutdown**Warning**** - Dear Account Holder, This is to notify all Staffs that we are validating active accounts. Kindly confirm that your account is still in use by clicking the val	Oct 25
Hardy	**Participate in WCNE`17 with submission|collect papers with CPCI-S,EI,CNKI** - Homepage Theme Submission About Xiamen Colleagues are invited for joining WCNE'17 By attendi	Oct 23

DATA SCIENCE IS ALL AROUND US

DENVER SMART CITY

Clean & Healthy City | Connected Multimodal | Internet of Things | Entrepreneurial City

What is Denver Smart City?

A city founded by the gold rush pioneers, again pioneering. Smart City Denver is pioneering technology for a mobile, affordable and environmentally-sustainable city.

We are creating a holistic, scalable program to address the Denver's challenges: mobility, safety, and health.

Supported through strong industry and community partnerships, along with over $15 million of local and Federal funding, Denver Smart City is pioneering the path for a sustainable city model, protecting our beautiful Rocky Mountains and the states natural beauty we all love.

CLEAN & HEALTHY CITY

Learn about how we are reducing air pollution, limit greenhouse gas emissions and improve Denver's environmental conditions to protect public health and prioritize environmental sustainability.
Learn More...

CONNECTED MULTIMODAL CITY

A connected city moves constituents where they need to be fast, efficiently and affordably.

Learn More...

INTERNET OF THINGS PLATFORM

Smart devices supporting a Smart City. Denver's custom Internet of Things (IoT) platform supports safety, mobility, transportation and weather in one place.

Learn More...

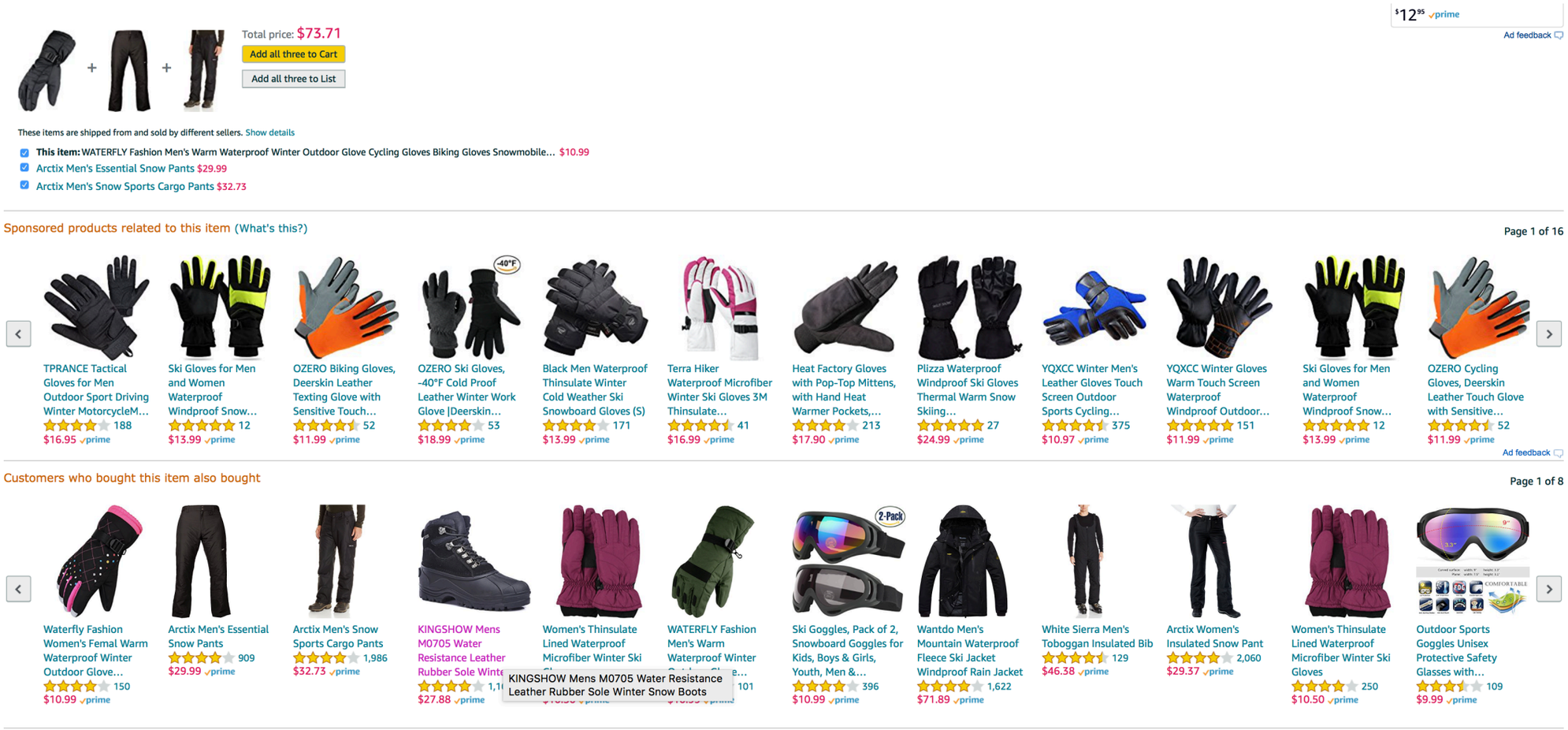
ENTREPRENEURIAL CITY

True to our pioneering roots, Denver is challenging the status quo, integrating our pioneer and entrepreneurial spirit bringing the best and brightest projects and collaboration to the Mile High City.

Learn More...

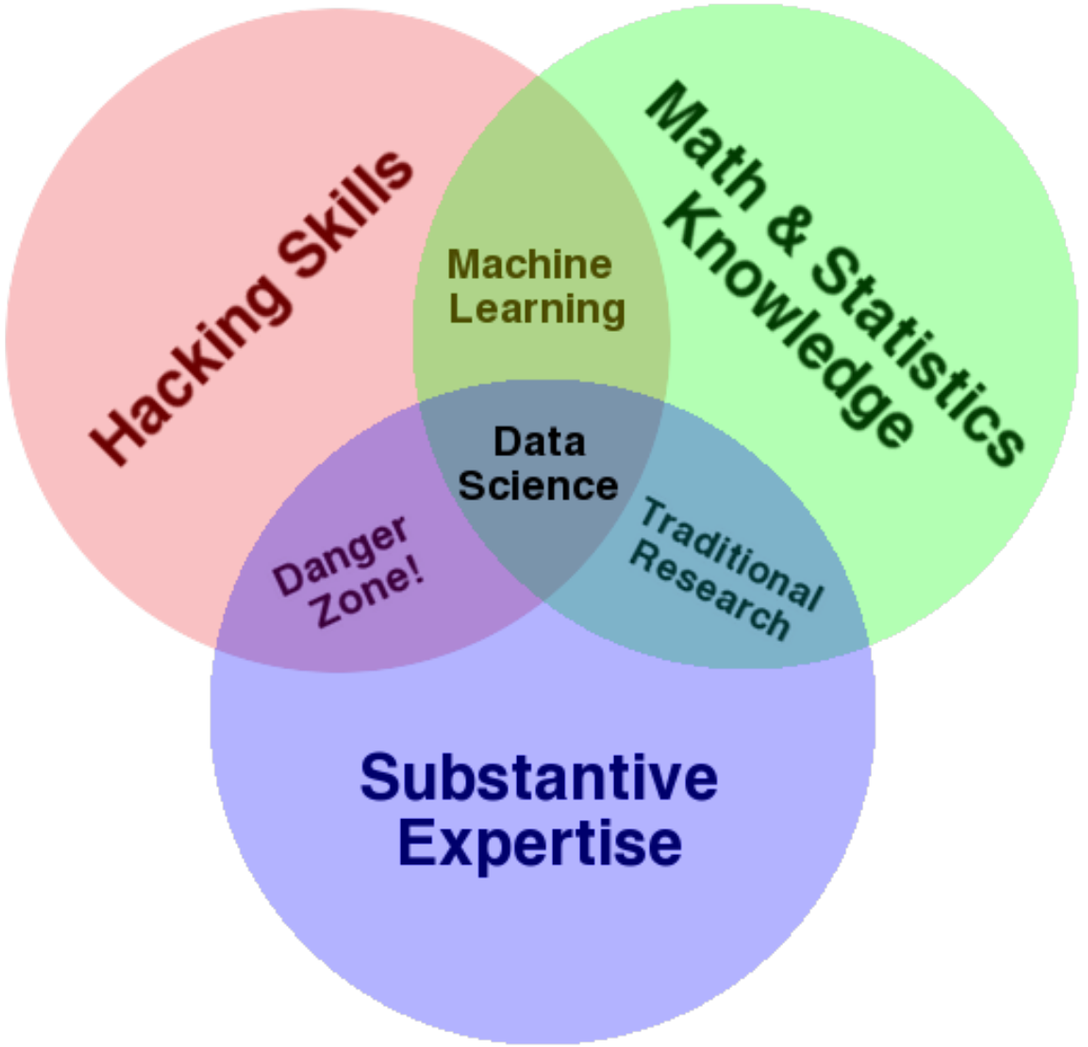
DATA SCIENCE IS ALL AROUND US

Amazon Recommendation System



WHAT IS DATA SCIENCE?

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

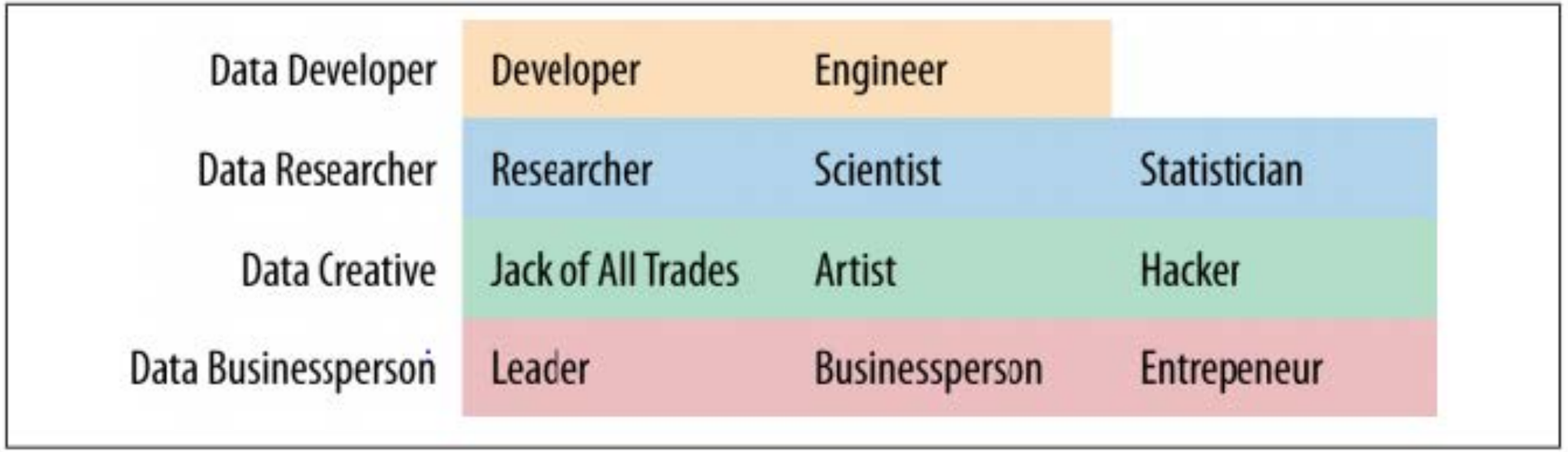


WHO USES DATA SCIENCE?

Can You Think of Others?

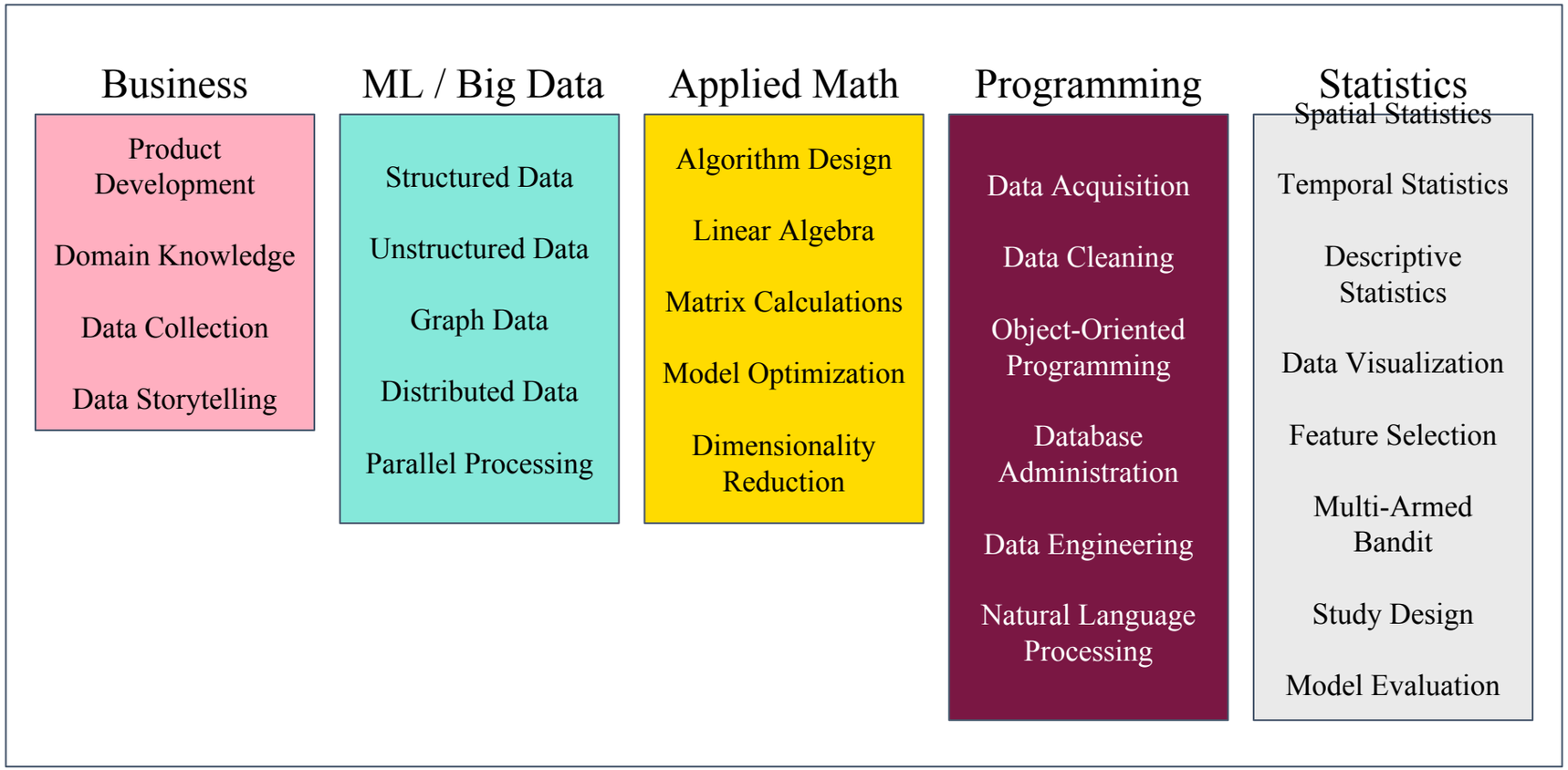
WHAT ARE THE ROLES IN DATA SCIENCE?

- Data Science involves a variety of roles, not just one.



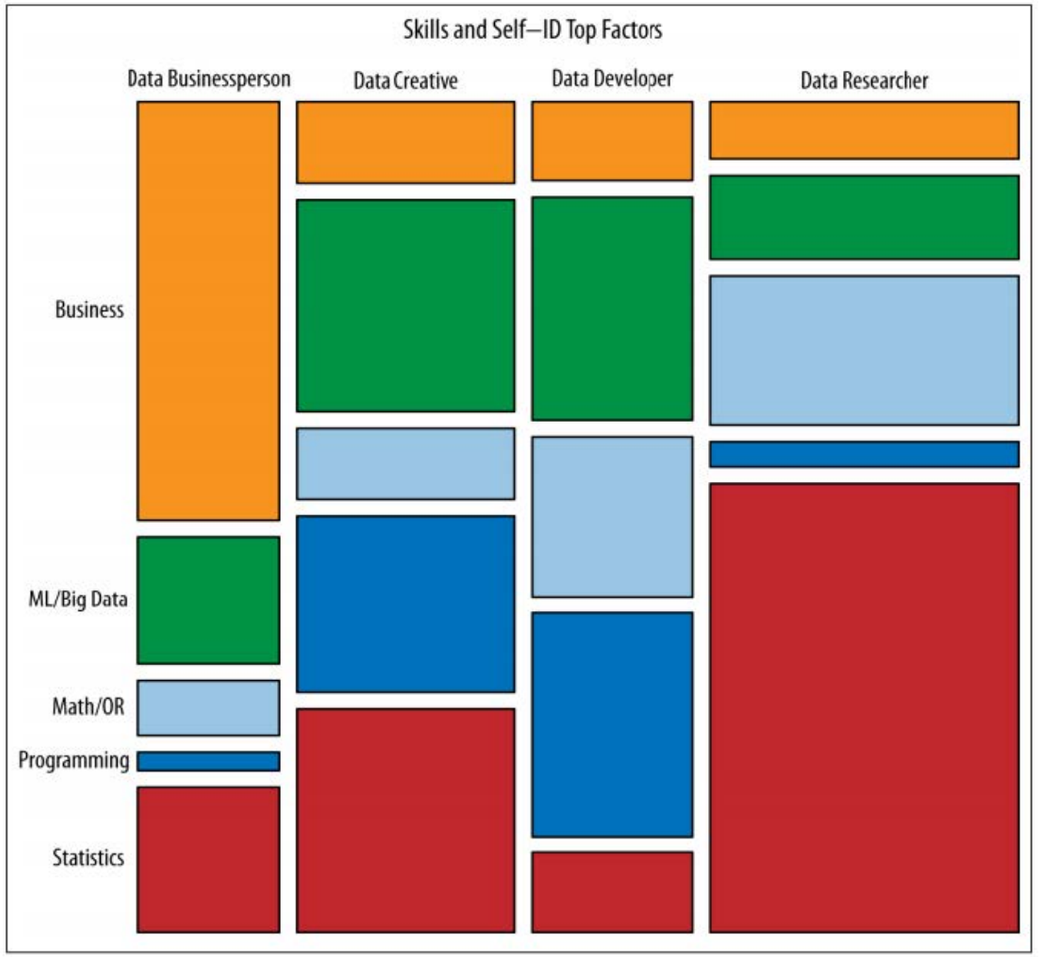
WHAT ARE THE ROLES IN DATA SCIENCE?

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

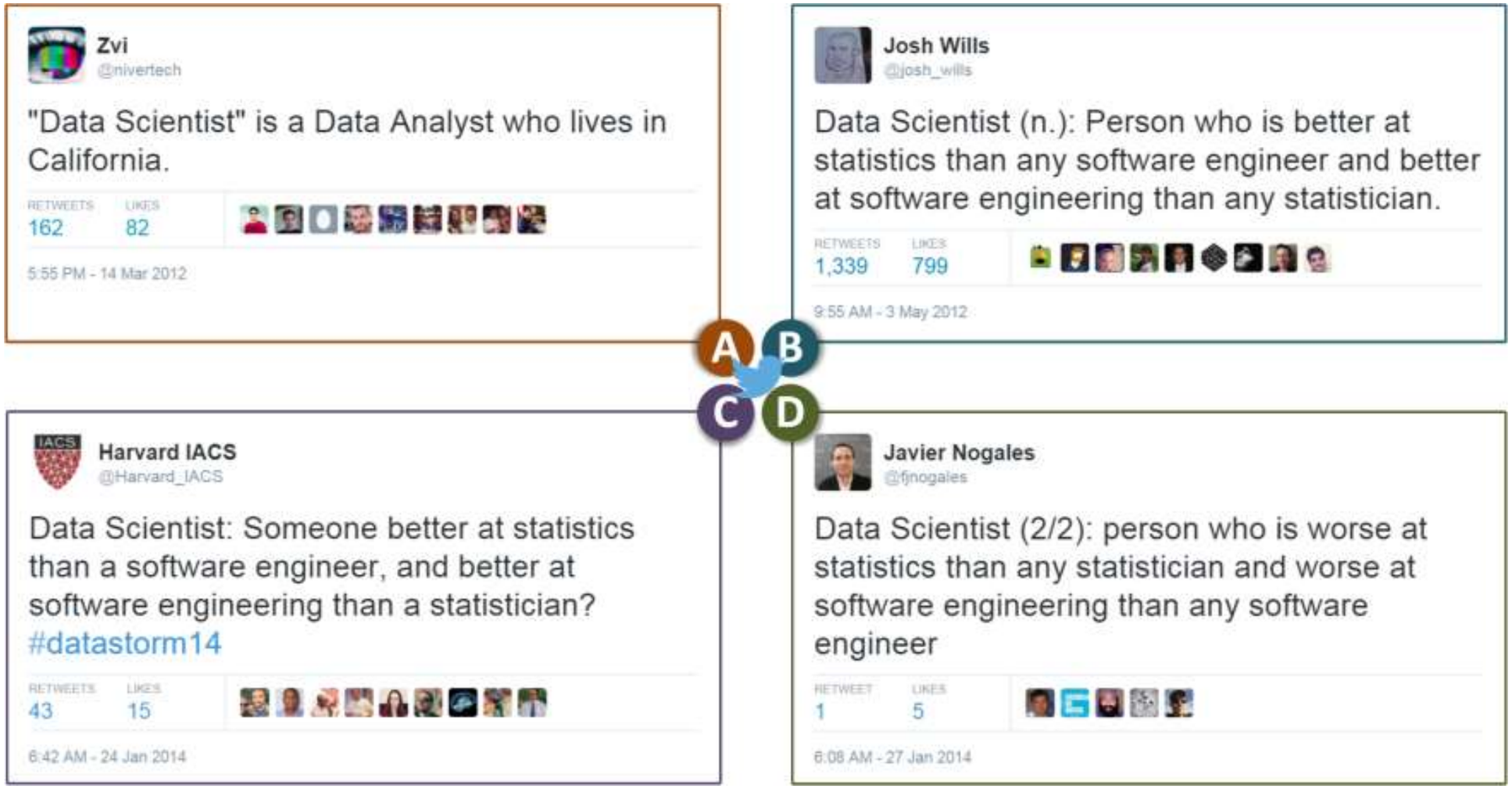


WHAT ARE THE ROLES IN DATA SCIENCE?

- These roles prioritize different skill sets.
- However, all roles involve some part of each skillset.
- Where are your strengths and weaknesses?



DATA SCIENTIST IN ONE SENTENCE



INTRODUCTION

THE DATA SCIENCE WORKFLOW

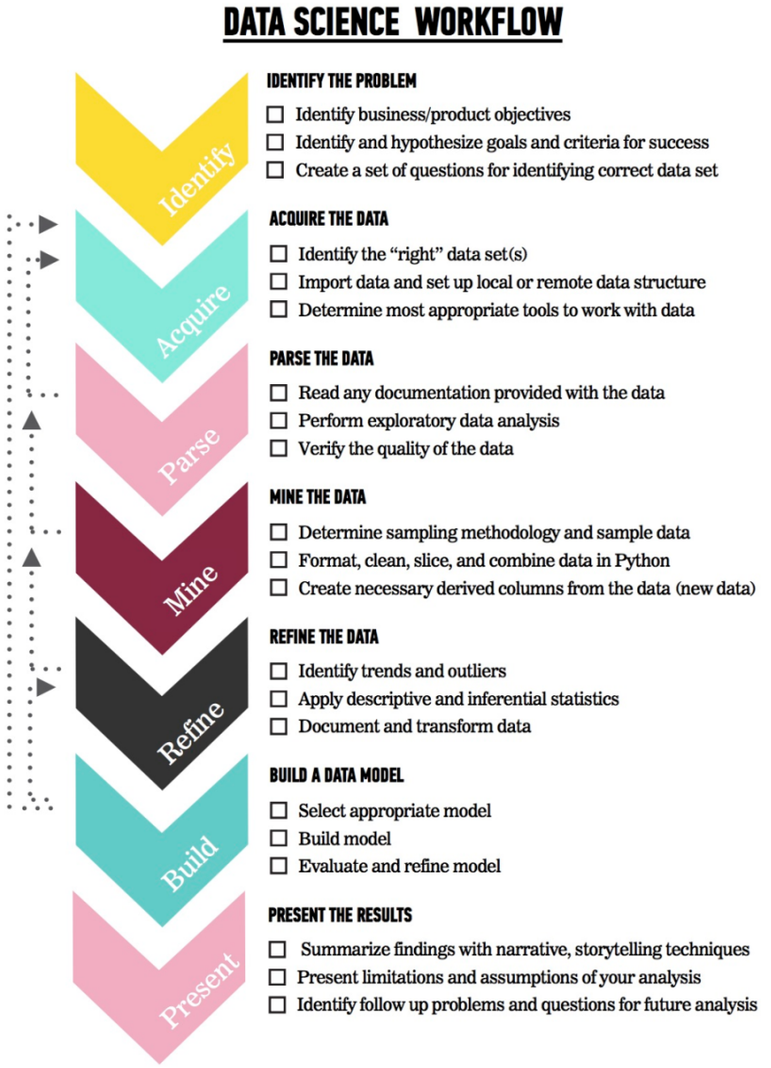
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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

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

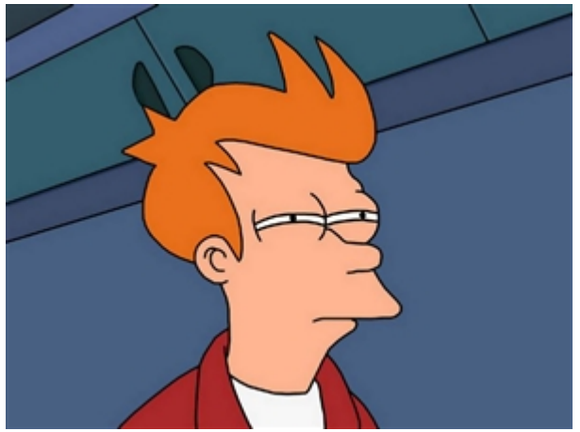
OVERVIEW OF THE DATA SCIENCE WORKFLOW

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

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
- https://github.com/ga-students/DAT-DEN-03

CASE STUDY

NYC TRAFFIC ANALYSIS

- Let's analyze the traffic in NYC:

https://github.com/ga-students/DAT-DEN-03/blob/master/lessons/lesson-01/Cabs.ipynb

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

DUE DATE

- Project: Begin work on Project 1

NEXT CLASS

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	~~What is Data Science~~	~~Lesson 1~~
	Research Design and Pandas	Lesson 2
	Statistics Fundamentals I	Lesson 3
	Statistics Fundamentals II	Lesson 4
	Flexible Class Session	Lesson 5
UNIT 2: FOUNDATIONS OF DATA MODELING	Introduction to Regression	Lesson 6
	Evaluating Model Fit	Lesson 7
	Introduction to Classification	Lesson 8
	Introduction to Logistic Regression	Lesson 9
	Communicating Logistic Regression Results	Lesson 10
	Flexible Class Session	Lesson 11
UNIT 3: DATA SCIENCE IN THE REAL WORLD	Decision Trees and Random Forests	Lesson 12
	Natural Language Processing	Lesson 13
	Dimensionality Reduction	Lesson 14
	Time Series Data I	Lesson 15
	Time Series Data II	Lesson 16
	Database Technologies	Lesson 17
	Where to Go Next	Lesson 18
	Flexible Class Session	Lesson 19
	Final Project Presentations	Lesson 20

WELCOME TO DATA SCIENCE

Q & A

WELCOME TO DATA SCIENCE

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

CLICK HERE