GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE

Ivan Hernandez, Ph.D

OUTLINE OF THE DAY

1. Welcome to GA / Course Information
2. Individual Introductions
3. Main Lesson (What is Data Science?)
4. Development Environment
5. Conclusions

DATA SCIENCE

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

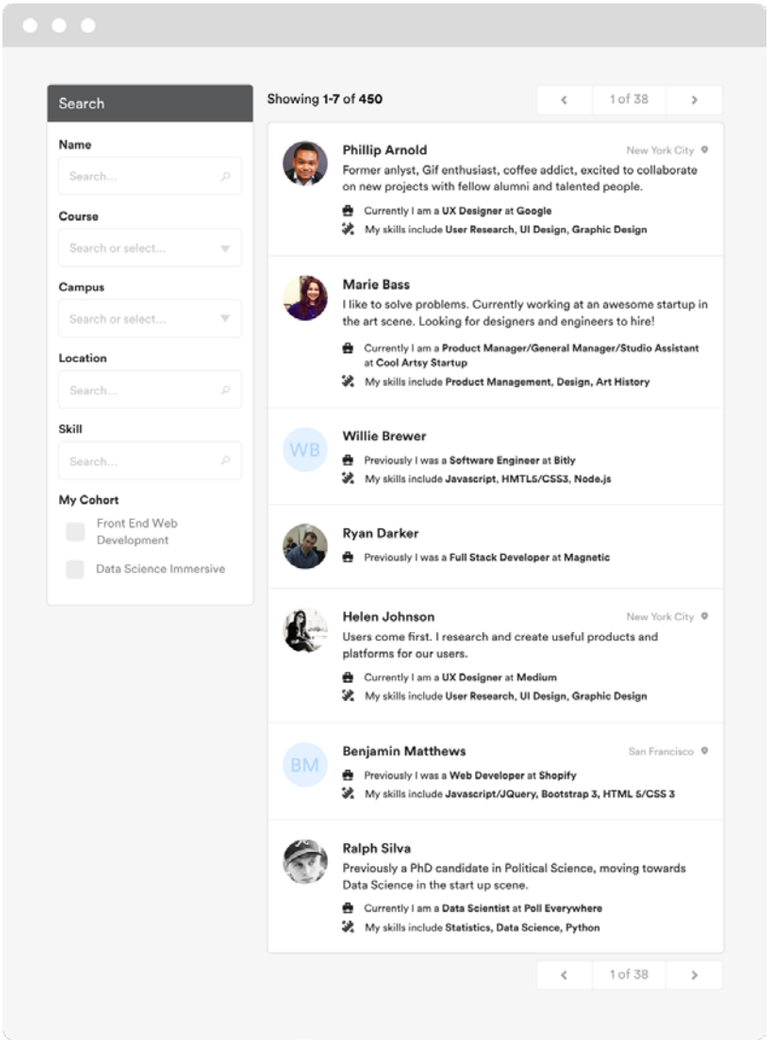
FEEDBACK/ SUPPORT

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

GA Directory

The GA Directory is a place for students, alumni and instructors to connect.

- Find your classmates
- Reach out to alumni
- Hire talent based on skill
- [Directory]



WELCOME

COURSE INFORMATION

ANNOUNCEMENTS AND COURSE MATERIAL

- All course related announcements will be made on Slack:
- https://chicagoeveningcourses.slack.com/messages/data-science/

- All course materials can be found on the course Github:
- https://github.com/ga-students/CHI-DS-3/

- All course descriptions and dates can be found in the course syllabus (on Github)

- Contact me either through slack or e-mail: ivan@ivanhernandez.com

GA GRADUATION REQUIREMENTS

* Arriving late by more than 10 minutes counts as ½ an absence

Projects

- Homework/Unit Projects
  - 4 Unit Projects in Data Science
  - Each builds on top of skills learned previous
  - Assigned approximately ~2 weeks during first half of course
  - Full timeline available in the syllabus (main Github folder)

- Final Project
  - Address a data-related problem in your professional field
  - Acquire a real-world data set, form a hypothesis about it, clean, parse, and apply modeling techniques and data analysis principles
  - 5 structured assignments
  - Presentation of results and written report

CLASSROOM RULES & EXPECTATIONS

- Open and focused discussion is encouraged
  - Be mindful of giving everyone an equal chance to talk
  - Raise your hand before you speak
  - Zero tolerance for discrimination or harassment

- Laptops are a required part of the class
  - Used during the lab sessions
  - Must be closed during the lecture portion.
  - Take notes using pen and paper

CLASSROOM ACCOMODATIONS

- WiFi is provided by General Assembly
  - Network Name: SPACE
  - Password (lowercase): w0rk5pac3
- Restrooms are located near the elevator
  - Feel free to use at any time
  - Try to minimize distraction when entering/leaving
- Power outlets
  - Located in the middle of the room
  - Located on the left and right sides of the room

LEARNING ACCOMODATIONS

- General Assembly will provide reasonable accommodations for religious obligations, disabilities, etc.
- If you have a special accommodation request, please do not hesitate to let me or Angie (angela@generalassemb.ly) know
- There is also an opportunity to make a request in the introductory survey (link at the end of slides)
- Requests can be made at any time during the course

CLASS INTRODUCTIONS

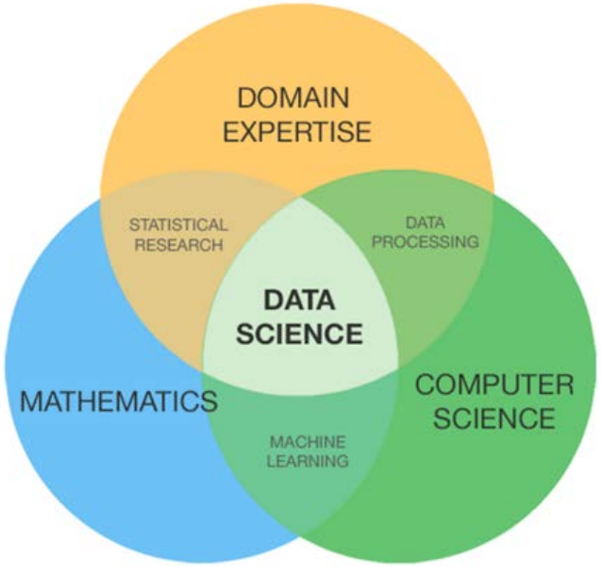
LEARNING OBJECTIVES

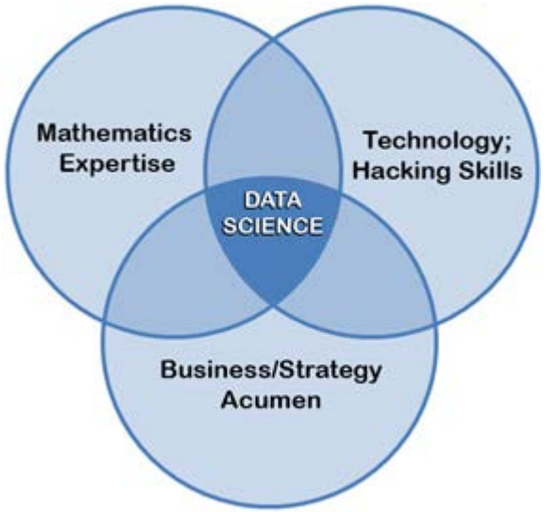
- Define data science and the data science workflow
- Apply the data science workflow
- Setup your development environment and review python basics

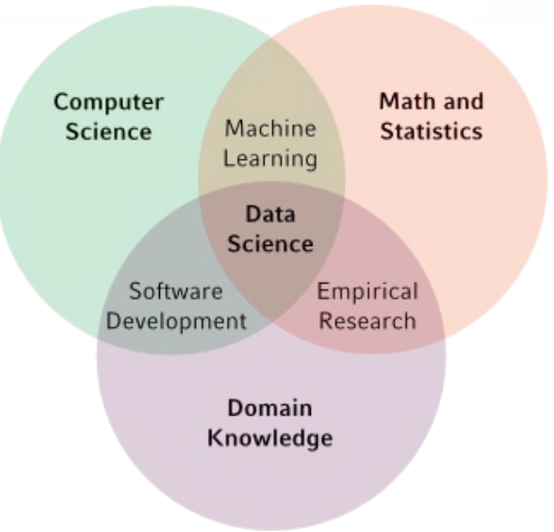
WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

- A set of tools and techniques for data
- Interdisciplinary problem-solving
- Multiple definitions
- Commonalities: Application of statistical and computational techniques to practical problems using the scientific method







WHAT IS DATA SCIENCE?: Illustrated Example

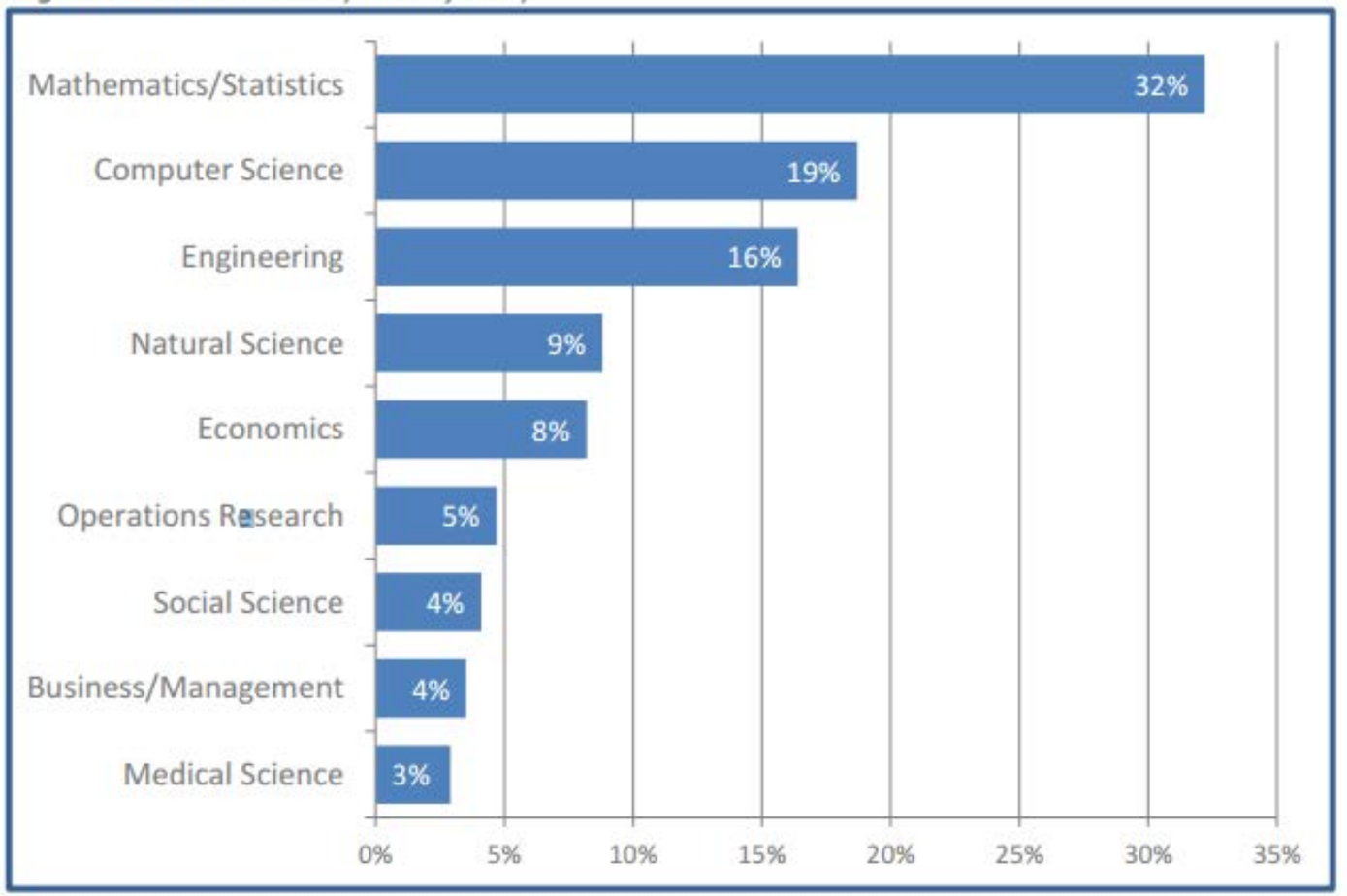
WORD CLOUD OF "DOING DATA SCIENCE: CHAPTER 1"

LATENT DIRICHLET ALLOCATION TOPIC MODEL OF "DOING DATA SCIENCE: CHAPTER 1"

Topic 1	Topic 2	Topic 3
Scientist	Hype	Statistic
Social	Mean	Scientist
Student	Scientist	People
Academia	Teach	Job
Define	Statistician	Skill
Question	Term	Industry
Field	Course	Google
People	Feel	Profile
Sense	Machine	Team
Solve	Leaning	Product

WHO ARE DATA SCIENTISTS?

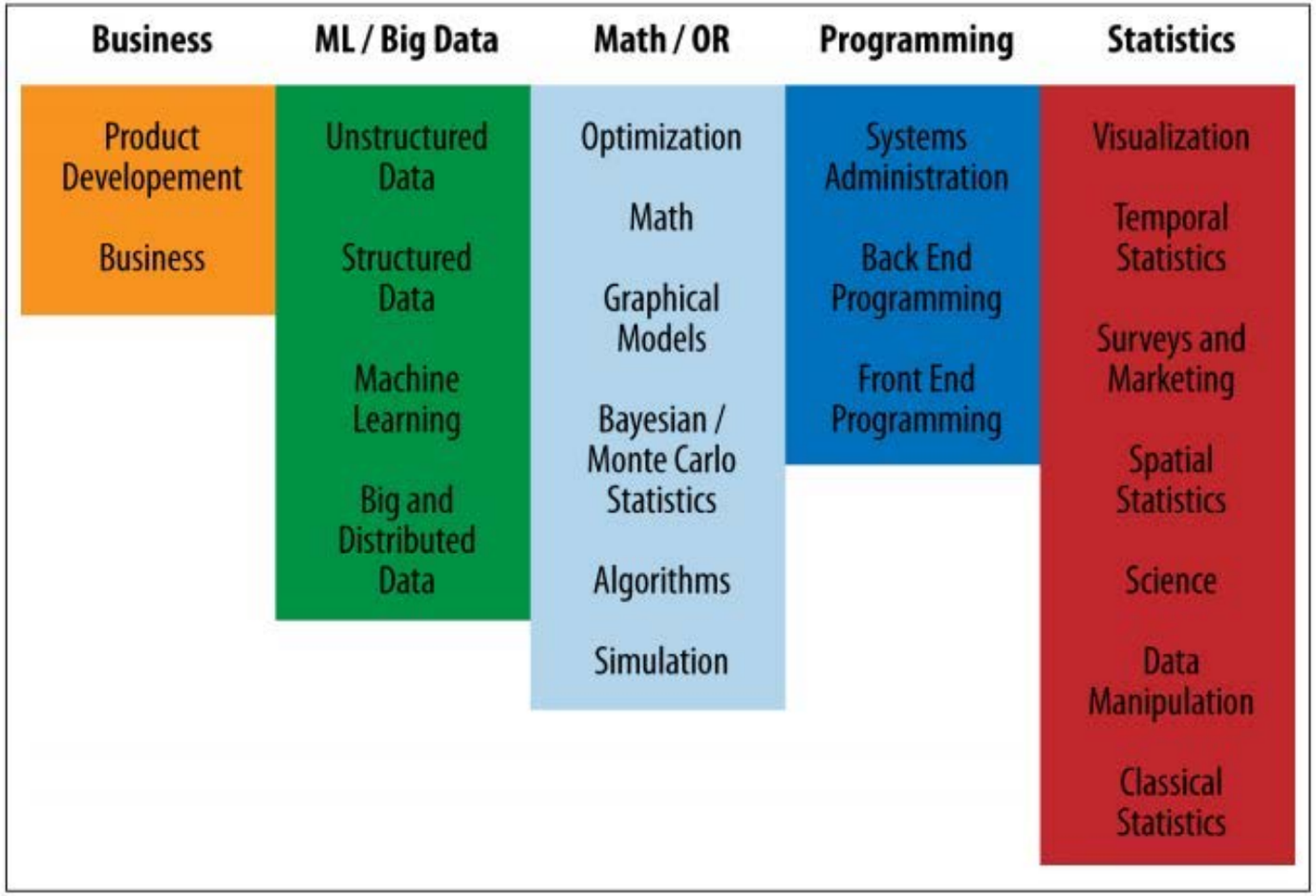
Figure 8. *Data Scientists by Area of Study*



Burtch Works Executive Recruiting study, 2014

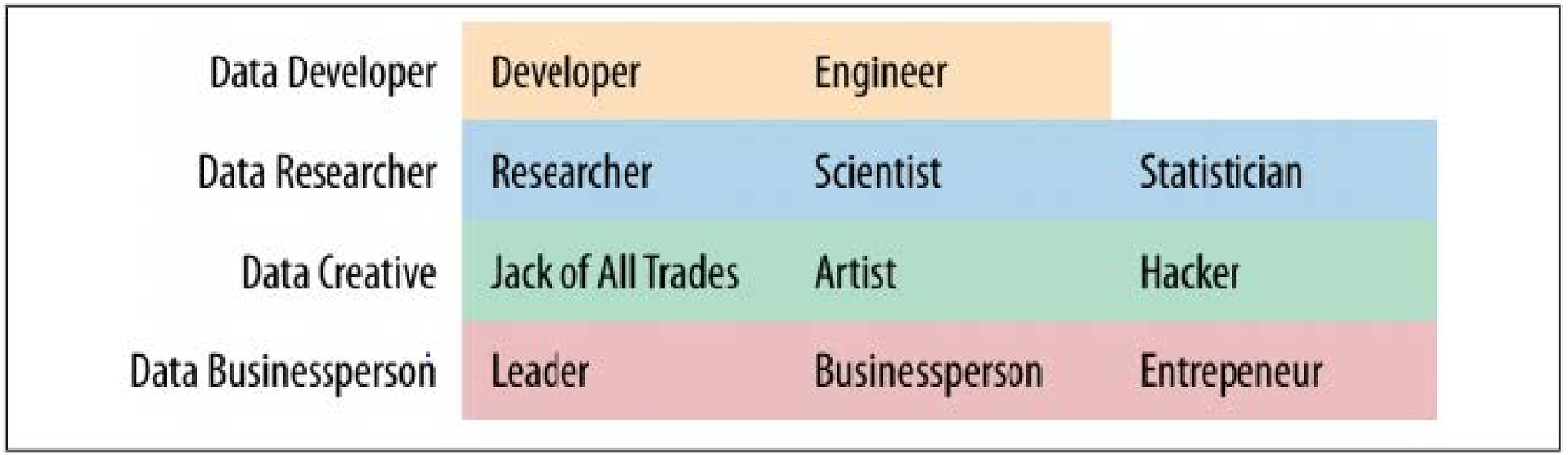
WHAT ARE THE ROLES IN DATA SCIENCE?

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



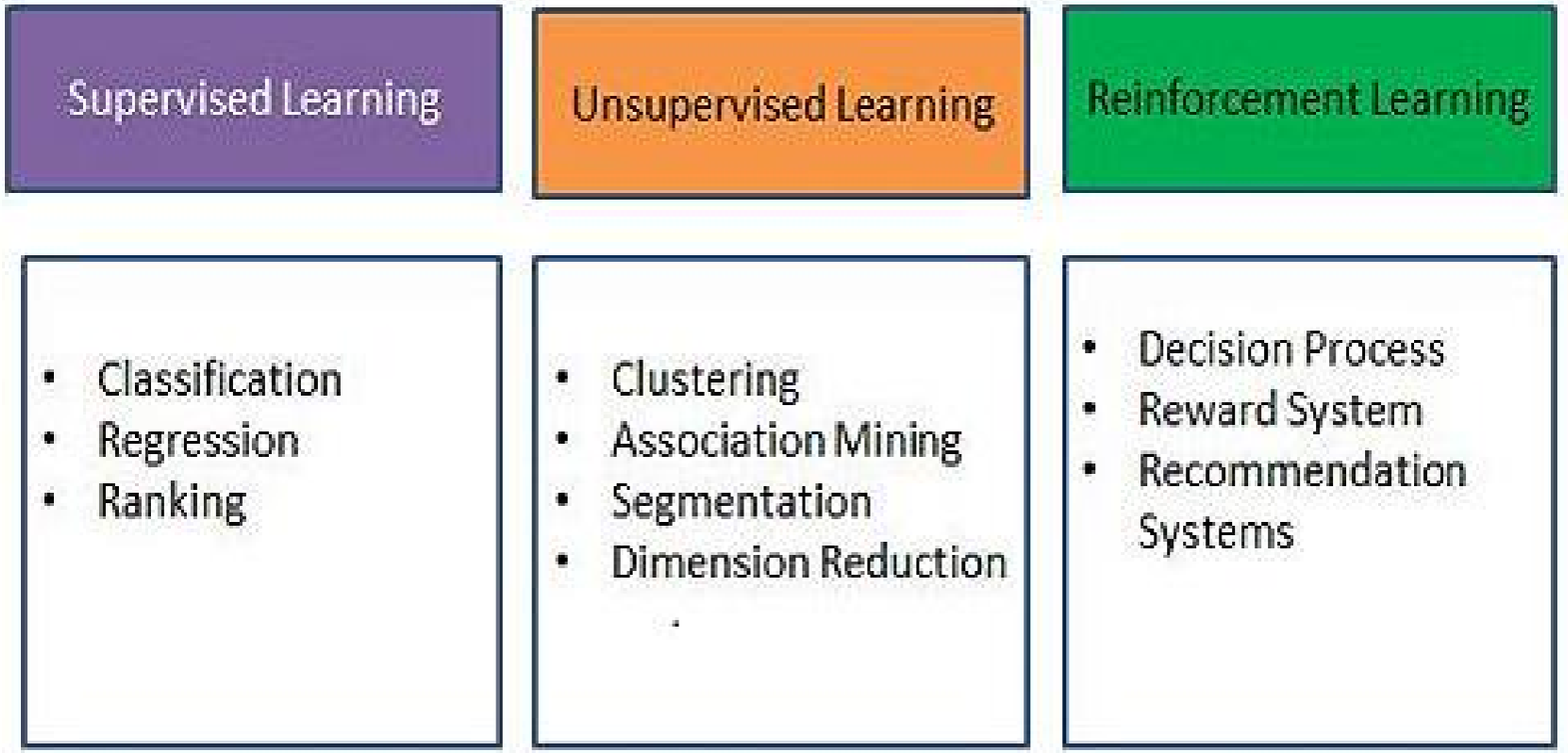
WHAT ARE THE ROLES IN DATA SCIENCE?

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



WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Data Scientists tend to use machine learning algorithms to address problems



WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Common questions answered by Data Scientists

1. Is this A or B? (Classification / Binary Prediction)
2. Is this A or B or C or D? (Recognition)
3. Is this Unusual? (Anomaly Detection)
4. How Much / How Many? (Regression / Quantative Prediction)
5. How is this Data Organized? (Grouping / Dimension Reduction)

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Is this A or B? (Classification): Predict events that have two possible outcomes
  - Will this customer default on their loan?
  - Is this an image of a cat or a dog?
  - Will this customer click on the advertisement?
  - Will this team win the basketball game?
  - Is this mole malignant or benign?

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Is this A or B or C or D? (Recognition): Predict which category a case belongs to
  - Which animal is in this image?
  - Which aircraft is causing this radar signature?
  - What is the topic of this news article?
  - What is the mood of this tweet?
  - Who is the speaker in this recording?

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- Is this Unusual? (Anomaly Detection): Determine if a phenomenon deviates from an expected range
  - Is this pressure reading unusual?
  - Is this internet message typical?
  - Is this combination of purchases very different from what this customer has made in the past?
  - Are these weather patterns normal for this century?

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- How Much / How Many? (Prediction): Predict a quantitative outcome
  - What will the temperature be next Tuesday?
  - What will my fourth quarter sales in Portugal be?
  - How many kilowatts will be demanded from my wind farm 30 minutes from now?
  - How many new followers will I get next week?

WHAT KINDS OF PROBLEMS DO DATA SCIENTISTS ADDRESS?

- How is this Data Organized? (Grouping): What are the categories or smaller dimensions within the data.
  - What are the different types of coffee drinkers?
  - Which viewers like the same kind of movies?
  - What kinds of car models does GM produce?
  - Are there common clusters of cable channels that customers tend to purchase together
  - What is a natural way to break these documents into five topics?

CLASS ACTIVITY

DATA SCIENCE QUESTIONS

ACTIVITY: DATA SCIENCE QUESTIONS

DIRECTIONS (10 minutes)

1. Break into pairs (person sitting next to you)

2. Pick a topic of interest to the both of you (e.g., music, finance, psychology, retail)

3. For each of the 5 kinds of data science questions, come up with a specific question you could ask for that topic.

    - Is this A or B? (Classification)
    - Is this A or B or C or D? (Recognition)
    - Is this Unusual? (Anomaly Detection)
    - How Much / How Many? (Prediction)
    - How is this Data Organized? (Grouping / Dimension Reduction)

WHO USES DATA SCIENCE?

Active Competitions

All Competitions

Active Competitions

Ultrasound Nerve Segmentation Identify nerve structures in ultrasound images of the neck	**58 days** **331** teams **376** scripts **$100,000**
Draper Satellite Image Chronology Can you put order to space and time?	**6.2 days** **403** teams **437** scripts **$75,000**
State Farm Distracted Driver Detection Can computer vision spot distracted drivers?	**41 days** **1064** teams **674** scripts **$65,000**
Grupo Bimbo Inventory Demand Maximize sales and minimize returns of bakery goods	**2 months** **626** teams **653** scripts **$25,000**
Avito Duplicate Ads Detection Can you detect duplicitous duplicate ads?	**20 days** **411** teams **252** scripts **$20,000**

Facebook V: Predicting Check Ins Machine Learning Software Engineer at **Facebook** Menlo Park, CA or Seattle, WA	**15 days** **943** teams **1934** scripts **Jobs**
Integer Sequence Learning 1, 2, 3, 4, 5, 7?!	**3 months** **70** teams **89** scripts **Knowledge**
Painter by Numbers Does every painter leave a fingerprint?	**4 months** **13** teams **52** scripts **Knowledge**
Shelter Animal Outcomes Help improve outcomes for shelter animals	**40 days** **1025** teams **671** scripts **Knowledge**

WHO USES DATA SCIENCE?

- Can you think of others?

THE DATA SCIENCE WORKFLOW

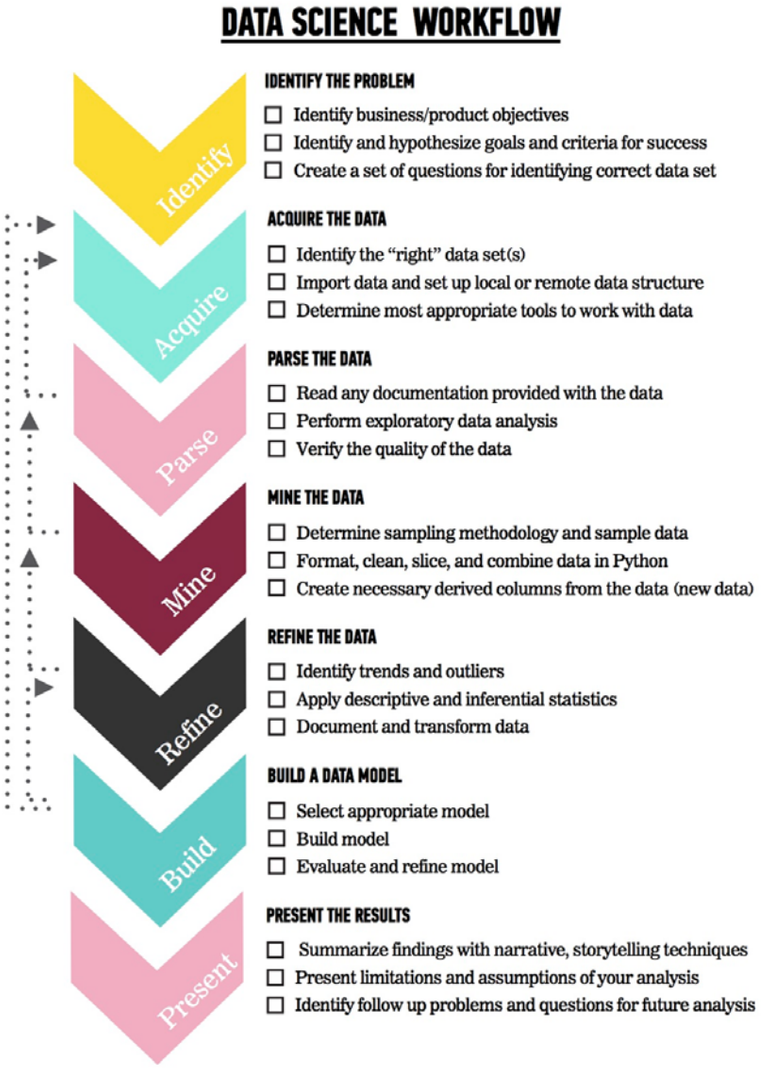
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- What process does a data scientist follow?
- Similar to the scientific method
- Helps produce *accurate* and *reproducible* results
  - *Accurate*: Describes a true consistent phenomenon or finding
  - *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

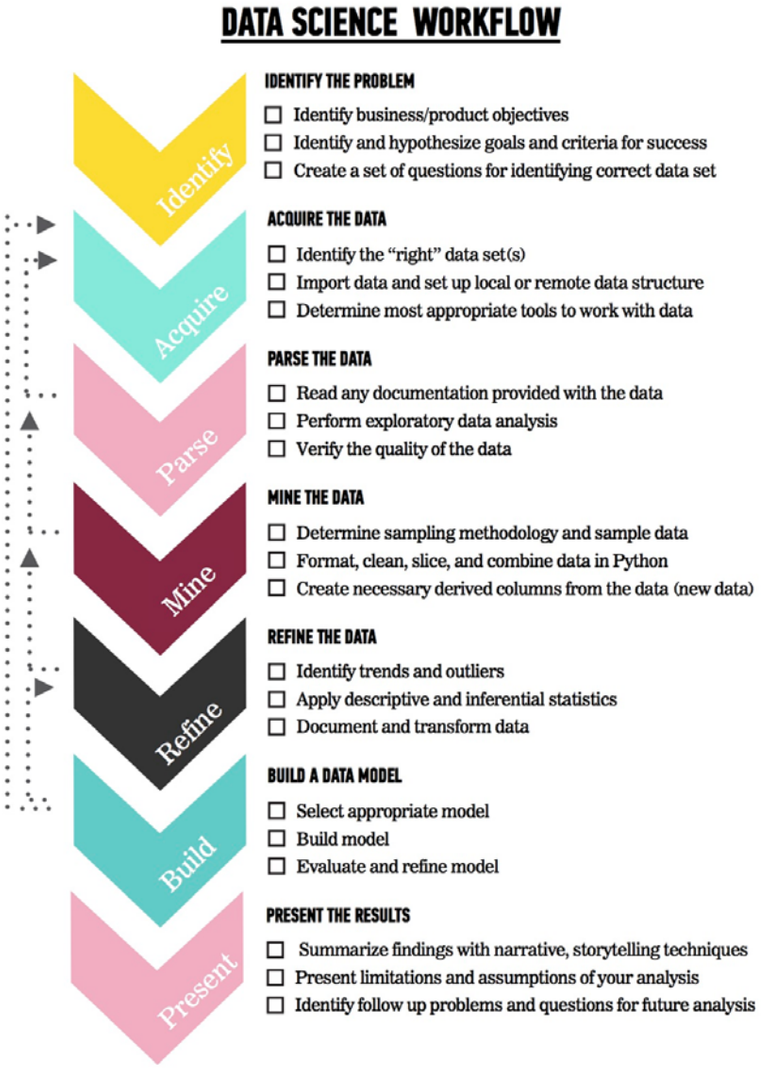
PRESENT THE RESULTS

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

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



THE DATA SCIENCE WORKFLOW: NETFLIX EXAMPLE

NETFLIX EXAMPLE

- Problem Statement: In 2006, Netflix Prize held a competition, open to anyone, to develop an algorithm that could predict user ratings for films, based on previous ratings without any other information about the users or films, i.e. without the users or the films being identified except by numbers assigned for the contest.

- Netflix offered a $1 million prize to any person/team that could improve the accuracy of its own recommendation engine by at least 10%

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

NETFLIX 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.

NETFLIX 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?

NETFLIX 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?

NETFLIX EXAMPLE: PARSE THE DATA

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

Variable	Description	Format
MovieID	A unique number indicating the movie	Categorical: Integer
CustomerID	A unique number indicating the customer who rated the movie	Categorical: Integer
Rating	Number of 'stars' assigned to a movie by a customer; integer from 1-5	Continuous: Integer
Title	English Language Title	Categorical: String
YearofRelease	Year a movie was released in the range [1890..2005].	Continuous: Integer

NETFLIX 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?

NETFLIX EXAMPLE: MINE THE DATA

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

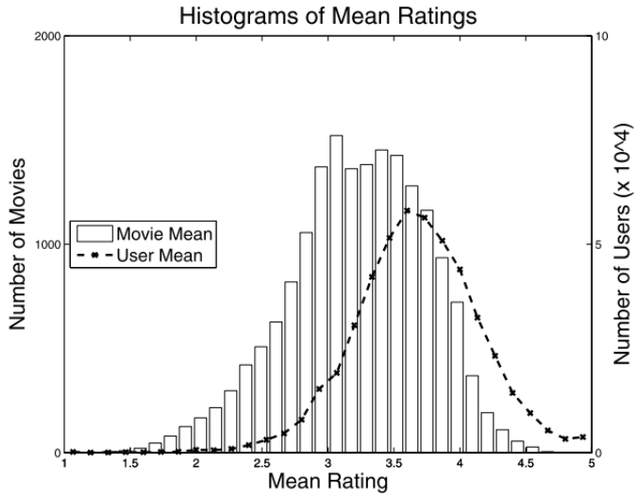
NETFLIX EXAMPLE: MINE THE DATA

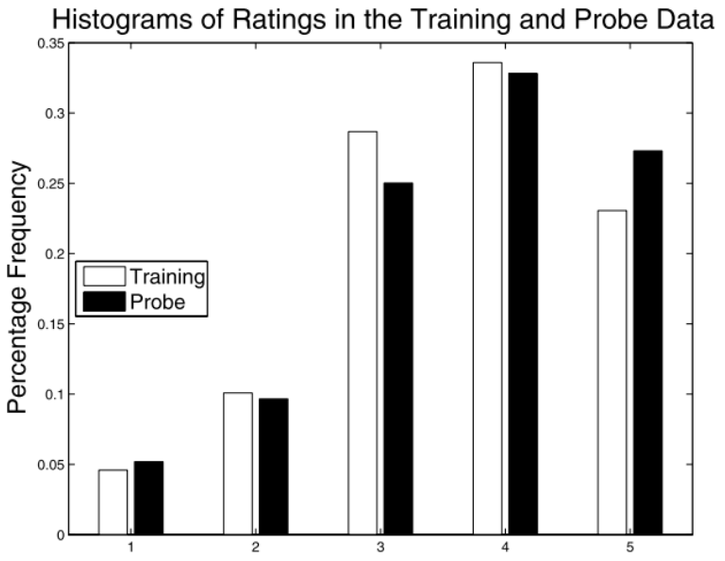
- Common steps while mining the data
  - Sample the data with appropriate methodology
  - Explore outliers and null values
  - Format and clean the data
  - Determine how to address missing values
  - Format and combine data; aggregate and derive new columns

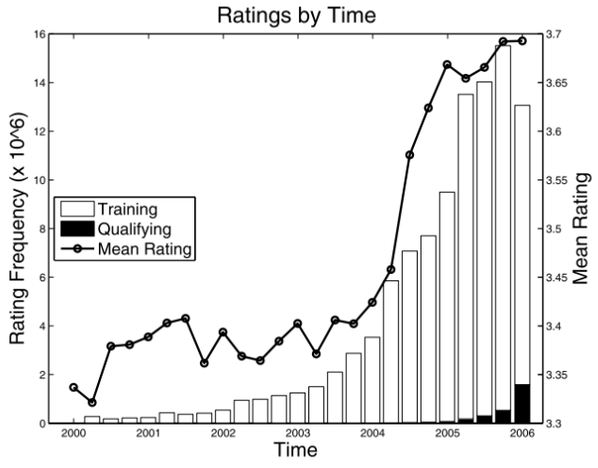
NETFLIX EXAMPLE: REFINE THE DATA

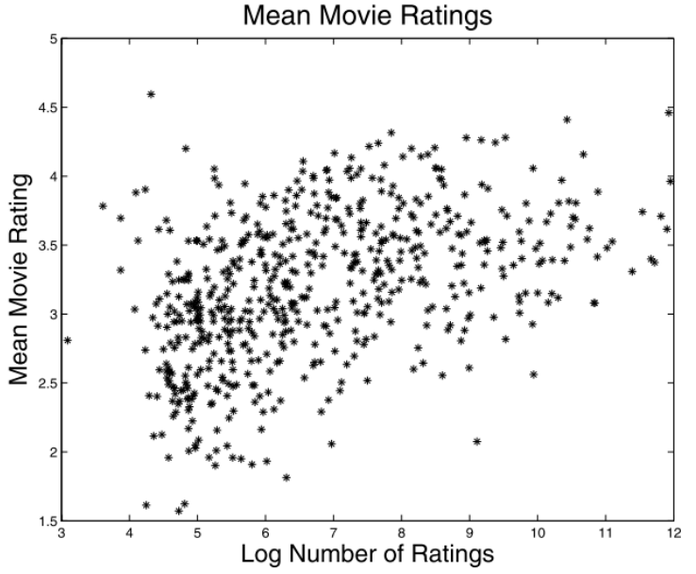
- Use descriptive statistics (mean, mode, standard deviation) to help:
  - Identifying trends and outliers
  - Deciding how to deal with outliers
  - Applying descriptive and inferential statistics
  - Determining visualization techniques for different data types
  - Transforming data

NETFLIX EXAMPLE: REFINE THE DATA









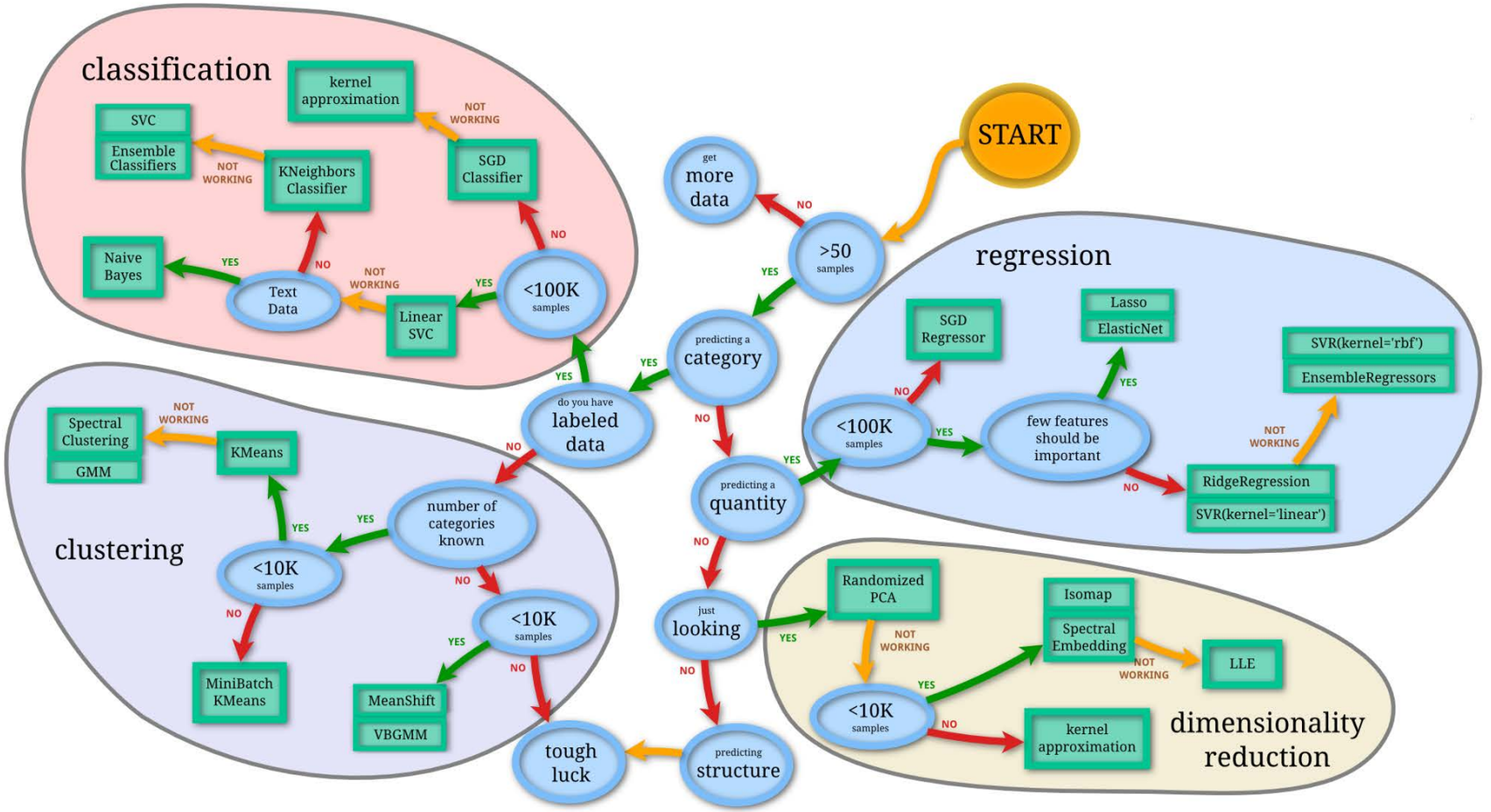
NETFLIX EXAMPLE: CREATE A DATA MODEL

- Select a model based upon the outcome
- Example models:
  - Linear regression where we predict a user's movie rating using release date, the movie's average rating, and user's average rating
  - Decision tree where we predict if a movie will receive 5 stars from a user based on the number of 5 stars a user has already given
  - K-Nearest Neighbor where we predict what rating a movie will receive based on the rating of similarly titled movies
- Steps for model building

NETFLIX EXAMPLE: CREATE A DATA MODEL

- The steps for model building are:
  - Select the appropriate model
    - Depends on many factors (type of research question, type of outcome, type of predictors, number of variables, number of cases)
  - Build the model
    - Select variables and parameters that go into the model
  - Evaluate and refine the model
    - See how model performs on a sample of data set aside, and make changes to improve performance
  - Predict outcomes and action items

NETFLIX EXAMPLE: SELECT THE APPROPRIATE MODEL



NETFLIX EXAMPLE: PRESENT THE RESULTS

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

NETFLIX 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

NETFLIX EXAMPLE: PRESENT THE RESULTS

- Example presentations and infographics
  - http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=8901D4FBEB41F4E5670517227ABC92DD?doi=10.1.1.142.9009&rep=rep1&type=pdf
  - http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf

DEMO

ENVIRONMENT SETUP

QUIZ

DATA SCIENCE BASELINE

ACTIVITY: DATA SCIENCE BASELINE QUIZ

DIRECTIONS (10 minutes)

1. Answer the following questions. On a sheet of paper
    a. True or False: Gender (coded male=0, female=1) is a continuous variable.
    b. Draw a Normal distribution
    c. Which measure of central tendency changes more in the presence of outliers: Mean or Median
    d. True or False: A small p-value is generally preferred when hypothesis testing
    e. True or False: Support Vector Machines are an unsupervised learning algorithm.

2. Break into groups of two (person sitting at your table and discuss your answers

DEV ENVIRONMENT SETUP

- Brief intro of tools
- Environment setup
  - Create a Github account (for homework)
  - 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-1/code/starter-code/lesson1-starter-code.ipynb* 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

- Create Github account for uploading projects: https://github.com/join?source=header-home
- Complete introductory survey: http://tiny.cc/chi-ds-survey
- Register on the GA directory:
- Read through final project instructions and start thinking about topic

Q&A

WELCOME TO DATA SCIENCE

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET LINK:

http://tiny.cc/chi-ds