WELCOME TO DATA SCIENCE

Alex Perrier

Data Scientist at Berklee Online, Contributor @ODSC

@alexip

aperrier@berklee.edu

linkedin.com/in/alexisperrier

WELCOME TO GA!

GENERAL ASSEMBLY

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.

FEEDBACK/SUPPORT

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

GA GRADUATION REQUIREMENTS

HOMEWORK
(COMPLETE 80% OF HOMEWORK/LABS)

ATTENDANCE
(MISS NO MORE THAN 2 CLASSES)

FINAL PROJECT

COMMUNITY ENGAGEMENT
PARTICIPATION + FEEDBACK

FOREVER AND EVER

BUILD YOUR NETWORK

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

FIND OPPORTU NITIES

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

13,000+ STRONG

You're part of the alumni community forever

PERKS!

15% OFF CLAASSES AND WORKSHOPS, $500 TUITION CREDIT

We can't wait to have you back on campus

OFFICE HOURS

Questions comments are welcome anytime

- Slack: DAT BOS 11
- aperrier@berklee.edu

- WHY DATA SCIENCE?
- ASPIRATIONS?
- DATA BACKGROUND?

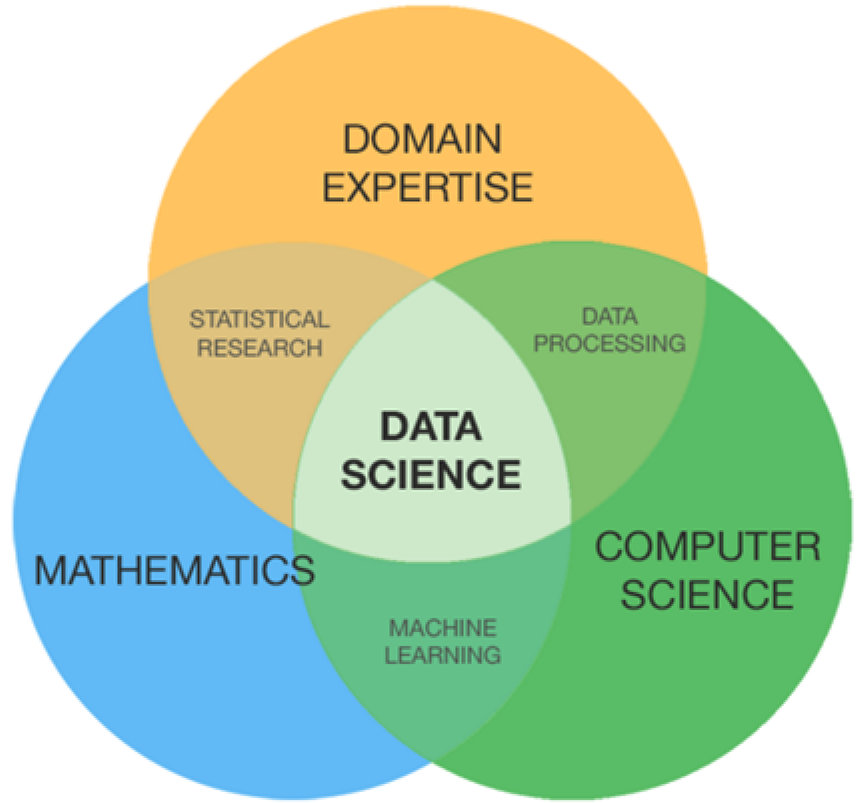
LEARNING OBJECTIVES FOR TODAY

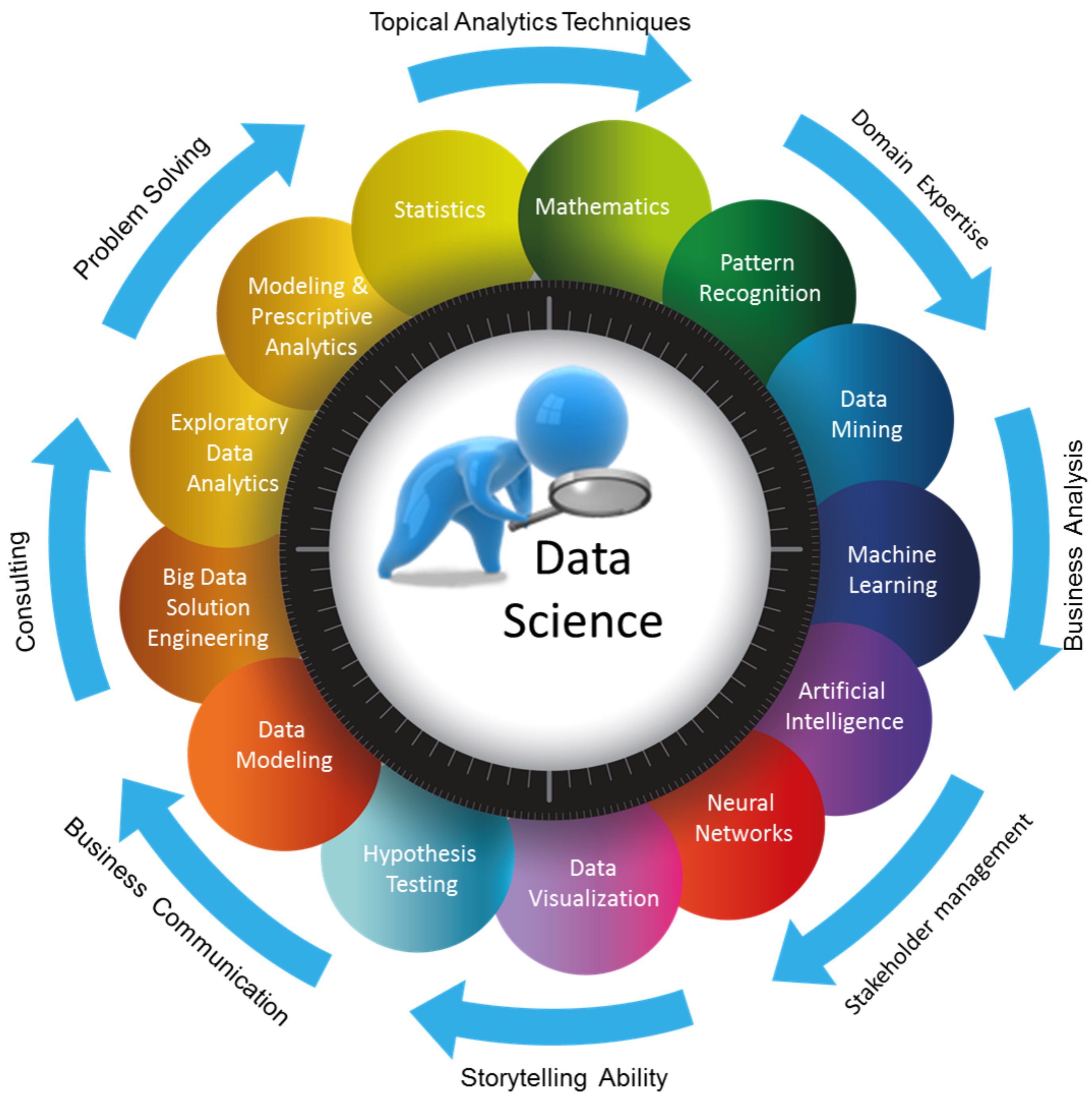
- Describe the roles and components of a successful learning environment
- Data science and the data science workflow
- Apply the data science workflow to meet your classmates
- Setup your development environment; review python and git basics

WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

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




Topical Analytics Techniques
Domain Expertise
Business Analysis
Stakeholder management
Storytelling Ability
Business Communication
Consulting
Problem Solving
Statistics
Mathematics
Pattern Recognition
Data Mining
Machine Learning
Artificial Intelligence
Neural Networks
Data Visualization
Hypothesis Testing
Data Modeling
Big Data Solution Engineering
Exploratory Data Analytics
Modeling & Prescriptive Analytics
Data Science

WHO USES DATA SCIENCE?

Companies:

- Facebook, Google,
- Amazon, Ebay,
- Spotify, AirBnB, Netflix,

Industries :

- Agriculture, Health, Transports, Astronomy, ...

WHAT CAN YOU DO WITH DATA SCIENCE?

- **Predictions**: market, demand, supply prices, population, weather, earthquakes, ...
- **Patterns**: customer behavior patterns
- **Detection**: Spam, Fraud, Failures, Cyber attacks
- **Extracting** meaning from large sets of data: handwritten health records, exoplanets
- **Streaming data**
- **NLP**: translation, speech to text, speech recognition, sentiment analysis, topic modeling, spell checking
- **Recommender systems**: Netflix, Spotify, Amazon

WHAT CAN YOU DO WITH DATA SCIENCE?

- **Ranking systems**: search results
- **Autonomous systems** (reinforcement learning / AI): playing games, self driving cars, drones
- **Time series**: algorithmic trading, signal processing, IoT
- **Image / Video**: automatic captionning, face and object recognition, ...

ROLES IN DATA SCIENCE

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

DIFFERENCE BETWEEN:

- **Data Analysis, Data Mining**: explore and find trends, anomalies and correlations.
  - DA: focuses on a subset
  - DM: looks at all the data (90's)
- **Statistics**: Finding the best model that fits the data
- **Machine learning**: The Math and the Algorithms.

  The model learns (auto-tunes) from the data
- **Predictive analytics**: Build models that can predict from past data
- **Data science**: All that and more

[Quora] What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data?

A TINY DROP OF HISTORY

Great article Forbes: A Very Short History Of Data Science

2001 Leo Breiman, Berkeley, publishes "Statistical Modeling: The Two Cultures":

"There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models.

This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."

GREAT CAREER CHOICE

- HBR: Data Scientist: The Sexiest Job of the 21st Century
- Burtch works: The Data Science Market: 2016 Compensation Insights
- Forbes: Machine Learning Is Redefining The Enterprise In 2016

ACTIVITY: DATA SCIENCE BASELINE

EXERCISE

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. According to the table on the next slide, BMI is the ______
      i. Outcome
      ii. Predictor
      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

EXERCISE

Table 3. Adjusted mean[a] (95% confidence interval) of BMI and serum concentration of metabolic biomarkers in American adults by categories of weekly frequency of fast-food or pizza meals, NHANES 2007–2010

BMI or serum biomarker	*Weekly frequency of fast-food or pizza meals*				P[b]
	0 Time	*1 Time*	*2–3 Times*	*⩾4 Times*	
BMI[c], kg m^{-2}					
All (N = 8169)	27.5 (27.1, 27.8)	27.9 (27.6, 28.2)	28.9 (28.4, 29.4)	28.8 (28.3, 29.2)	< 0.0001
Men (n = 4002)	27.9 (27.4, 28.3)	28.0 (27.6, 28.4)	28.5 (28.0, 29.0)	28.6 (28.2, 29.0)	0.05
Women (n = 4167)	27.2 (26.8, 27.6)	27.7 (27.3, 28.1)	29.3 (28.6, 29.9)	29.0 (28.1, 29.8)	< 0.0001
Total cholesterol, mg dl^{-1} (N = 8236)	199 (197, 202)	198 (196, 200)	199 (196, 201)	198 (196, 201)	0.5
HDL-cholesterol[c], mg dl^{-1}					
All (n = 8236)	54 (53, 55)	53 (52, 54)	52 (51, 53)	51 (50, 52)	< 0.0001
Men (n = 4042)	48 (47, 49)	48 (47, 49)	48 (46, 49)	46 (45, 47)	0.003
Women (n = 4194)	60 (59, 61)	58 (57, 60)	56 (55, 57)	56 (54, 58)	0.001
LDL-cholesterol[d], mg dl^{-1}					
All (n = 3604)	113 (111, 116)	117 (113, 120)	113 (110, 116)	114 (110, 118)	0.6
< 50 Years (n = 2151)	107 (105, 110)	112 (109, 116)	111 (107, 114)	108 (104, 112)	0.8
⩾50 Years (n = 1453)	123 (118, 129)	126 (121, 131)	118 (113, 123)	129 (122, 137)	0.5
Triglycerides, mg dl^{-1} (n = 3659)	103 (98, 109)	103 (99, 108)	110 (106, 115)	110 (104, 117)	0.2
Fasting glucose[c], mg dl^{-1}					
All (n = 3668)	99 (98, 100)	99 (98, 100)	99 (98, 100)	99 (98, 100)	0.5
Men (n = 1750)	102 (101, 104)	102 (101, 104)	101 (99, 102)	101 (99, 102)	0.1
Women (n = 1918)	97 (95, 98)	95 (94, 97)	97 (96, 99)	98 (96, 101)	0.2
Glycohemoglobin, % (N = 8234)	5.42 (5.39, 5.44)	5.39 (5.36, 5.42)	5.39 (5.36, 5.42)	5.40 (5.37, 5.44)	0.2

Abbreviations: BMI, body mass index; HDL, high-density lipoprotein; LDL, low-density lipoprotein; NHANES, National Health and Nutrition Examination Surveys. [a]Adjusted means were computed from multiple linear regression models with each biomarker as a continuous dependent variable. All biomarkers (except BMI, total- and HDL-cholesterol) were log-transformed for analysis; therefore, the back-transformed values for LDL-cholesterol, triglycerides, fasting glucose and glycohemoglobin are geometric means and their 95% confidence intervals. Independent variables included: frequency of fast-food meals (0, 1, 2–3 and ⩾4 times), age (20–39, 40–59 and ⩾60), sex, race/ethnicity (non-Hispanic white, non-Hispanic black, Mexican-American and other), poverty income ratio (⩽1.3, >1.3–3.5, ⩾3.5 and unknown), years of education (< 12, 12, some college and ⩾college), serum cotinine (continuous), hours of fasting before phlebotomy, (continuous), physical activity (none, tertiles of MET minutes/week), alcohol-drinking status (never drinker, former drinker, current drinker and unknown). N refers to observations used in the regression model for each biomarker. [b]P-value for the Sattherwaite-adjusted F test for frequency of fast-food meals as a continuous variable. [c]Significant interaction of fast-food meals with sex ($P_{interaction}$ < 0.05; thus, the results are stratified by sex [d]Significant interaction of frequency of fast-food meals with age ($P_{interaction}$ < 0.05); thus, the results are stratified by age categories.

PART II

Data Science WorkFlow