GENERAL ASSEMBLY

WELCOME TO DATA SCIENCE

Stefan Jansen
DAT-NYC

LEARNING OBJECTIVES

- Setup your development environment and review python basics
- 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

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

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

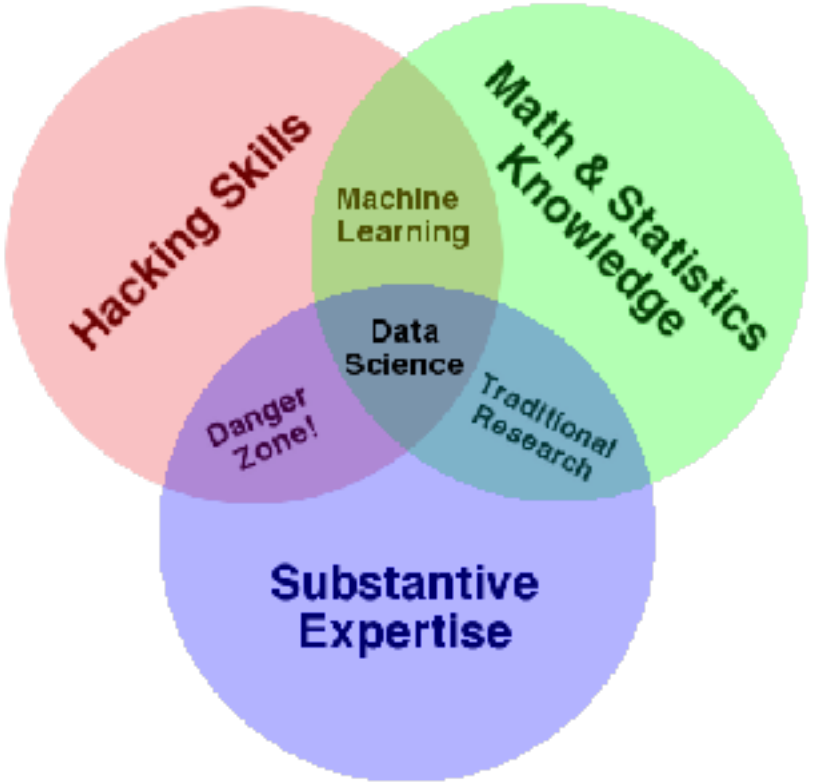
- Test your new setup using the lesson 1 starter code available at */lessons/lesson-1/code/starter-code/lesson1-starter-code.ipynb* in the Github repo

- Ask your classmates and instructor for help if you have problems!

WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

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



DATA SCIENCE BASED BUSINESS MODELS

WHO ARE DATA SCIENTISTS?

EXERCISE

DIRECTIONS (Teams of 3-4, 10 minutes)

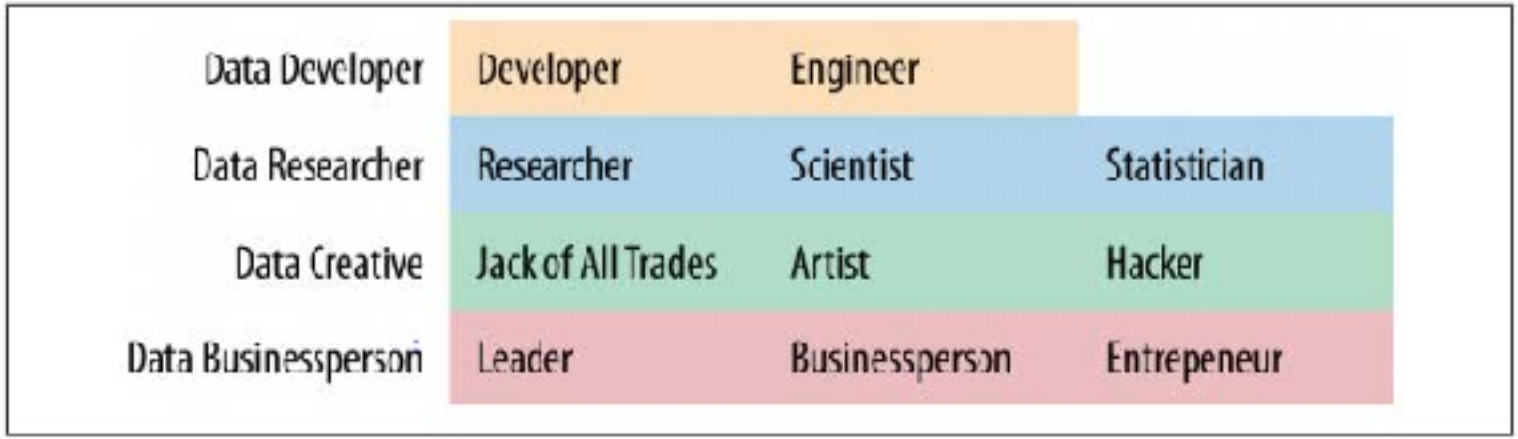
1. Who are Data Scientists?
2. How do Data Scientists add value?
3. What makes a good Data Scientist?
4. When finished, share your answers with the class

DELIVERABLE

Answers to the above questions

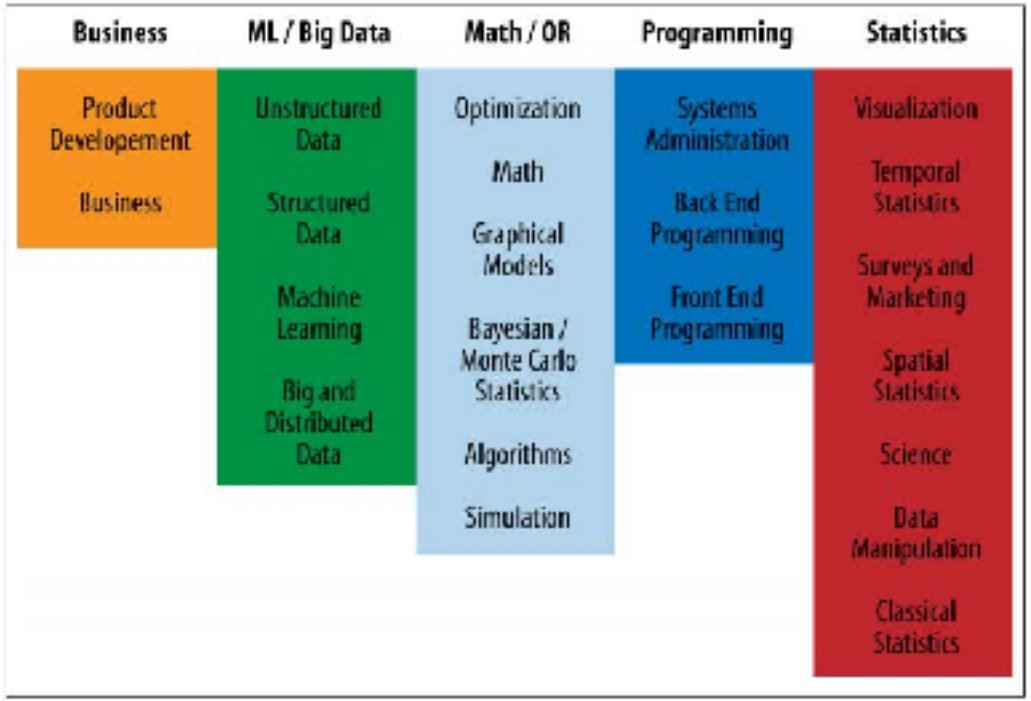
WHAT ARE THE ROLES IN DATA SCIENCE?

x`



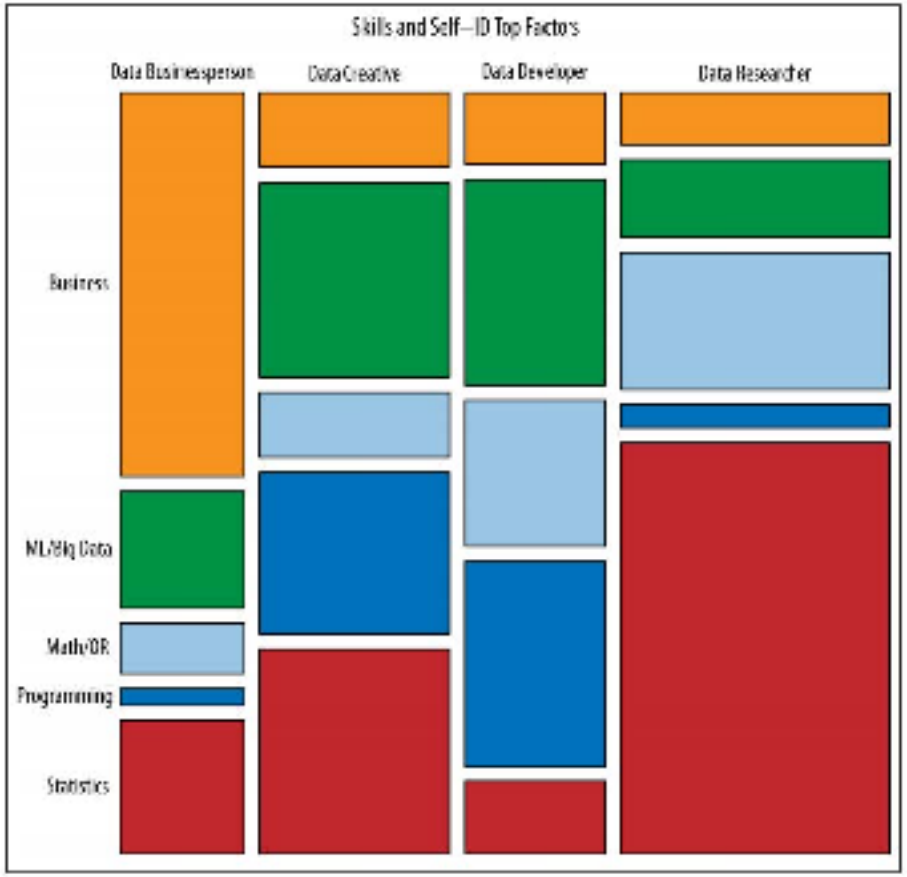
WHAT ARE THE ROLES IN DATA SCIENCE?

- 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?



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. 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 1. 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, kg m^{-2}					
All (n = 8169)	27.5 (27.1, 27.8)	27.9 (27.6, 28.2)	28.9 (28.7, 29.1)	28.8 (28.5, 29.2)	< 0.0001
Men (n = 4002)	27.6 (27.4, 28.3)	28.0 (27.6, 28.4)	28.5 (28.1, 29.0)	28.6 (28.2, 29.0)	0.05
Women (n = 4167)	27.2 (26.8, 27.5)	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 (195, 201)	198 (195, 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[c], mg dl^{-1}					
All (n = 3604)	113 (111, 116)	117 (113, 120)	113 (110, 116)	114 (113, 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, 125)	126 (121, 131)	118 (113, 123)	129 (122, 137)	0.5
Triglycerides, mg dl^{-1} (n = 3569)	103 (98, 109)	103 (99, 108)	110 (105, 115)	113 (108, 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, 103)	101 (99, 103)	0.1
Women (n = 1918)	97 (95, 98)	95 (94, 97)	97 (96, 98)	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 HDL-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 Satterthwaite-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.

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

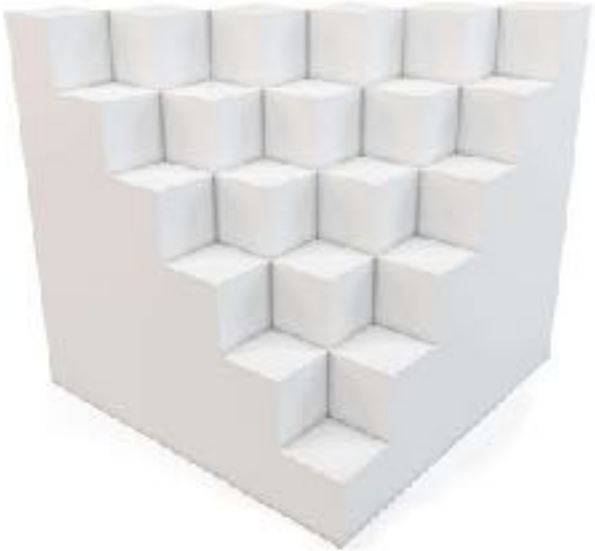
- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

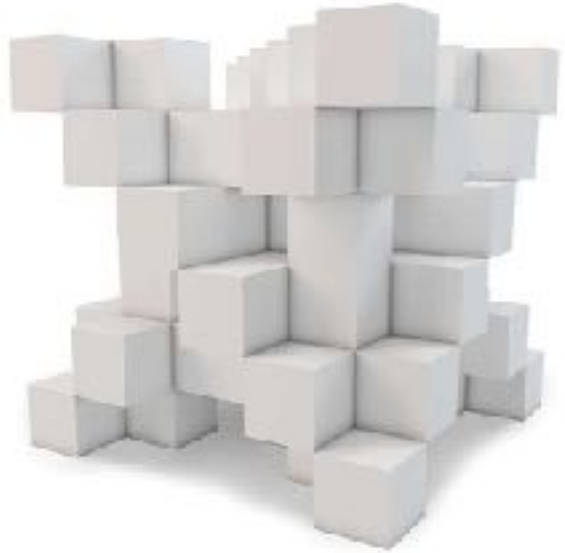
OVERVIEW OF THE DATA SCIENCE WORKFLOW

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)

DATA: STRUCTURED vs UNSTRUCTURED





dny3d ©
123RF.com

UNSTRUCTURED DATA

Bundit Chuangboonsri © 123RF.com

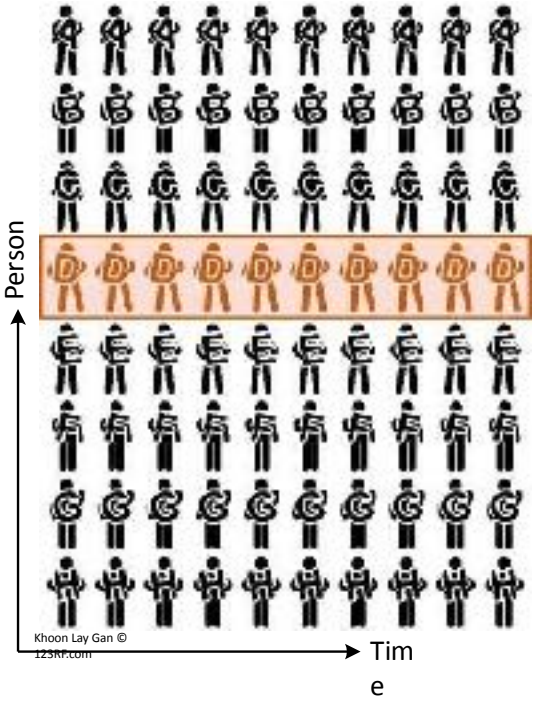
- Sessions 13 and 14 in Unit 3
  - Natural Language Processing

WE WILL MOSTLY LOOK AT STRUCTURED DATA

- Unit 2
  - Linear Regression (sessions 6 and 7)
  - Classification and Logistic Regression (session 8 and 9)
- Unit 3
  - Decision Trees and Random Forests (session 12)

milosb © 123RF.com

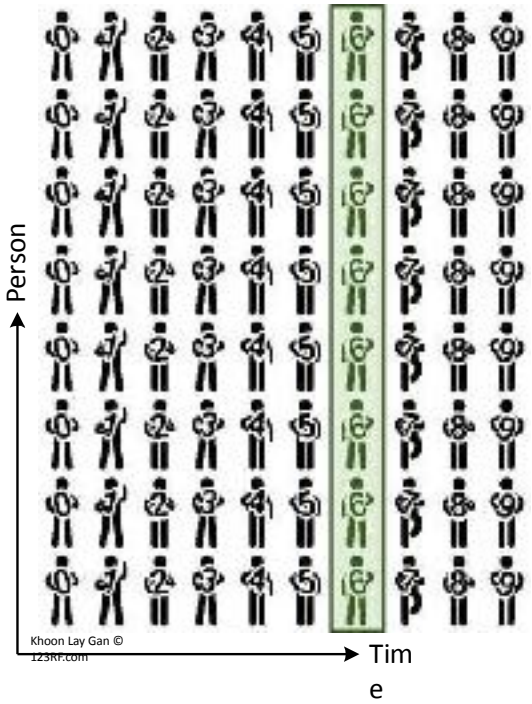
DATA CAN HAVE A TIME DIMENSION



- Sessions 15 and 16 in Unit 3
  - Time Series

OR CROSS-SECTIONAL

- And most of the course will focus on it

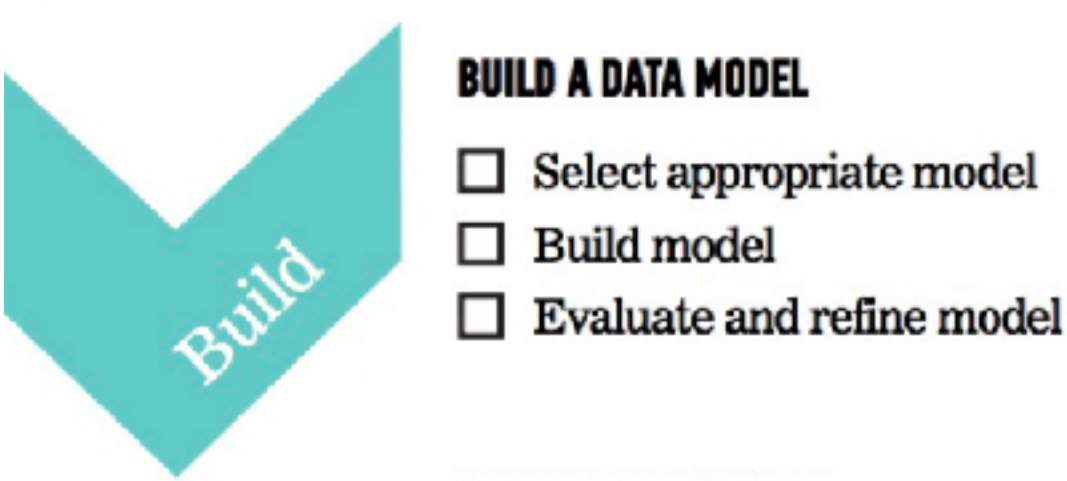


OVERVIEW OF THE DATA SCIENCE WORKFLOW

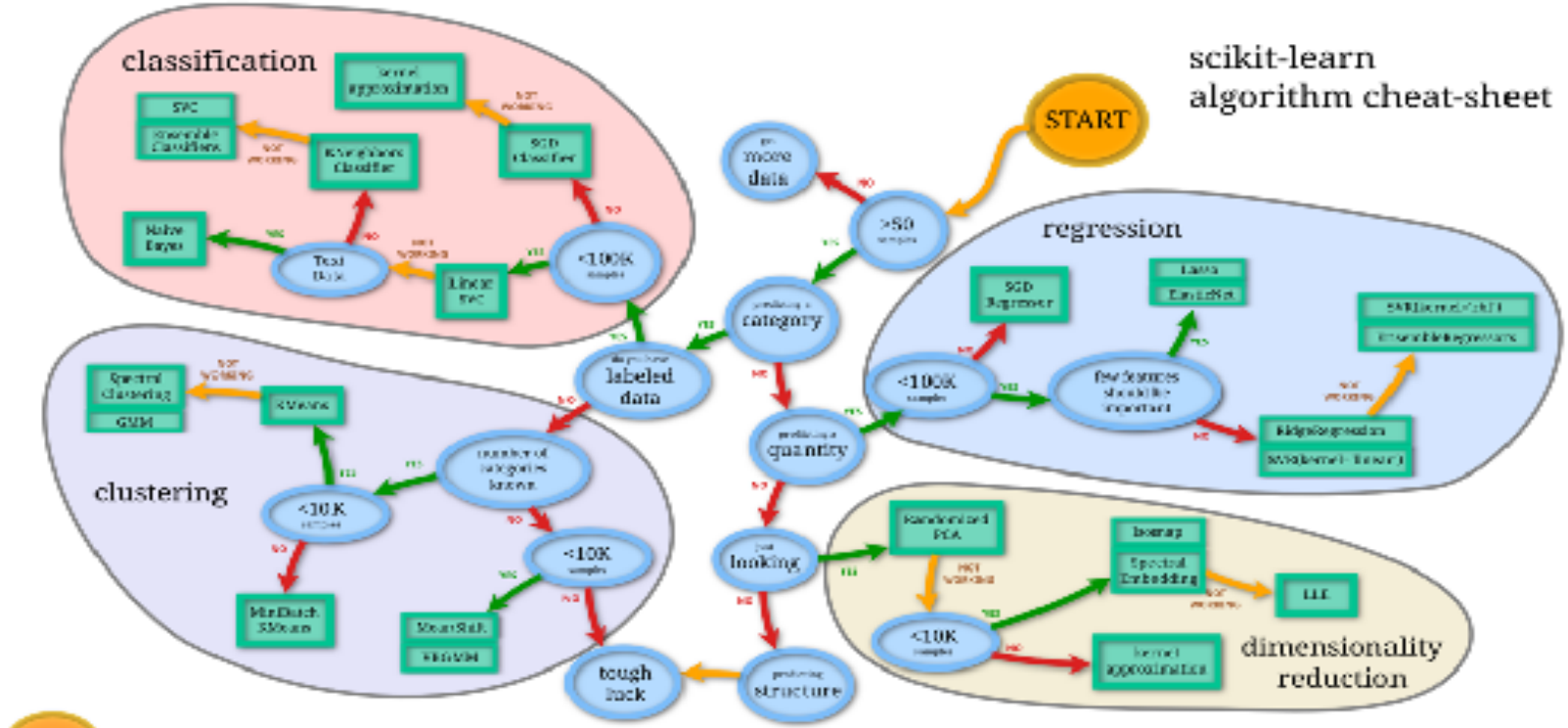
REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

OVERVIEW OF THE DATA SCIENCE WORKFLOW



ML ALGORITHMS ON OUR AGENDA



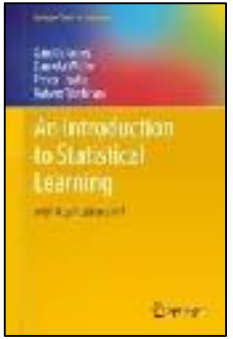
OVERVIEW OF THE DATA SCIENCE WORKFLOW

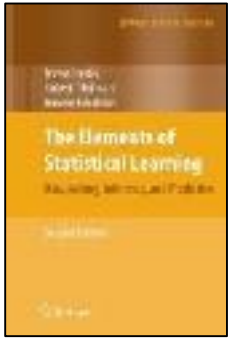
PRESENT THE RESULTS

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

GREAT FREE (OPTIONAL) RESOURCES

- An Introduction to Statistical Learning: with Applications in R (by James et al.).





- For a more advanced treatment of these topics, check out The Elements of Statistical Learning: Data Mining, Inference, and Prediction (by Hastie et al.)

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

CONCLUSION

REVIEW

CONCLUSION

`

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATES: Github Course page

- Project: Begin work on Project 1 & Start thinking about Final Part 1

Q & A

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET