WELCOME TO DATA SCIENCE

Alex Perrier

Data Scientist at Berklee Online, Contributor @ODSC

Twitter: @alexip

I blog at alexperrier.github.io

Datasets, Notebooks and Slides for this lesson are available on github.

WELCOME TO GA!

GENERAL ASSEMBLY

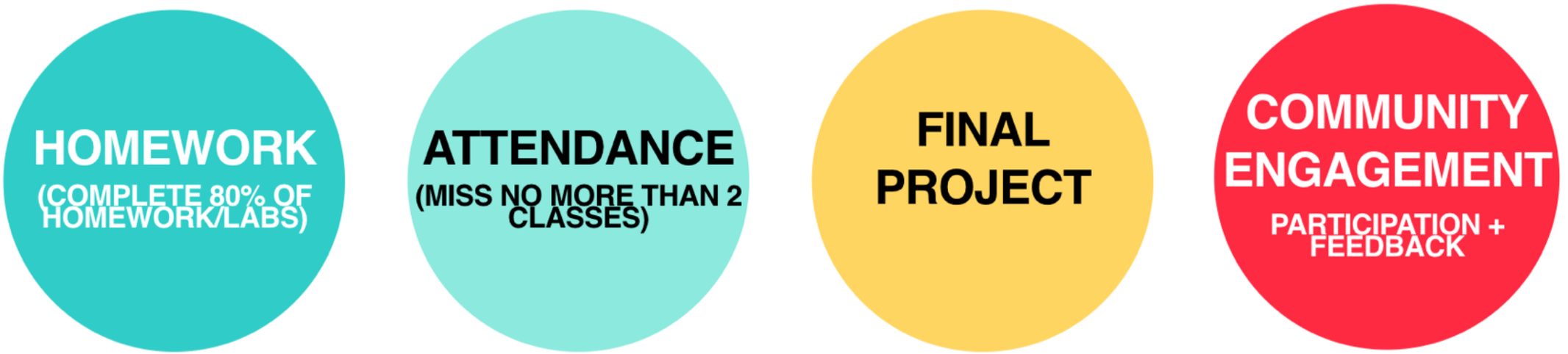
General Assembly is a global community of individuals empowered to pursue the work we love.

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



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

OFFICE HOURS

Questions comments are welcome anytime

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

INTRODUCTION

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

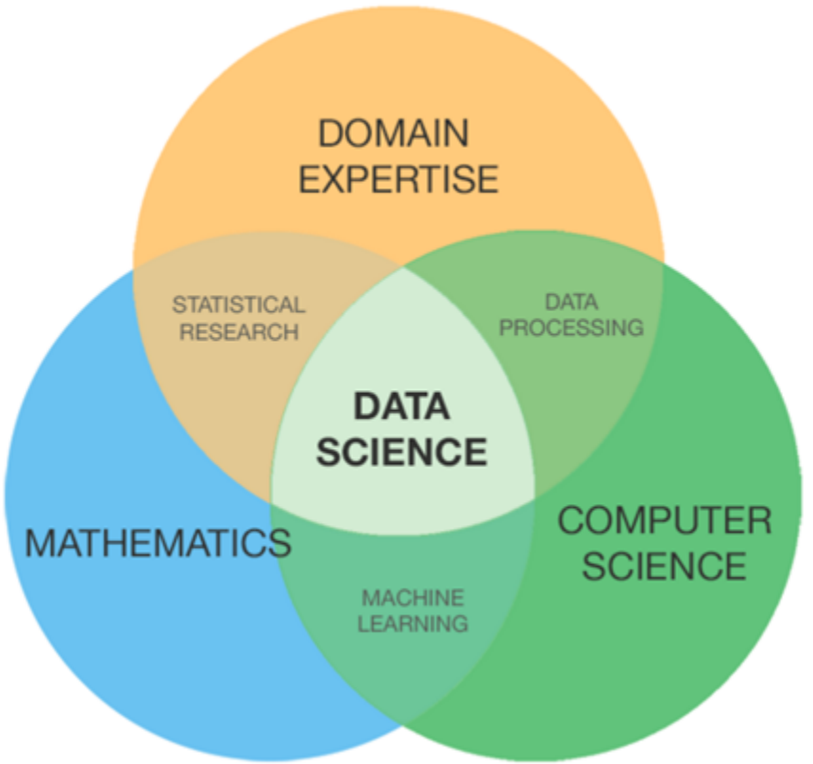
LEARNING OBJECTIVES FOR TODAY

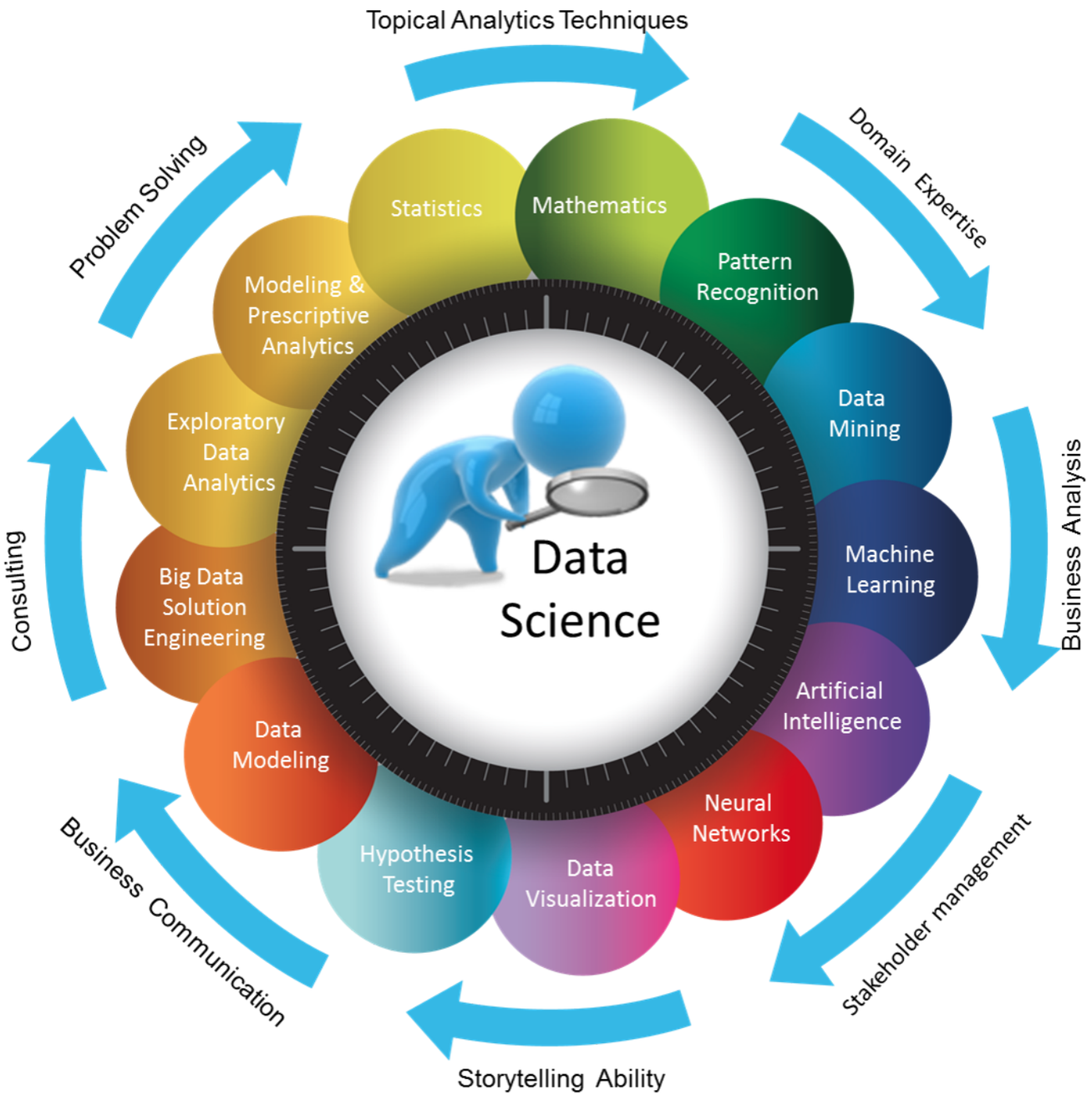
- Describe the roles and components of a successful learning environment
- Data science and the data science workflow
- Apply the data science workflow to meet your classmates
- Setup your development environment; review python and git basics

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,
  ...

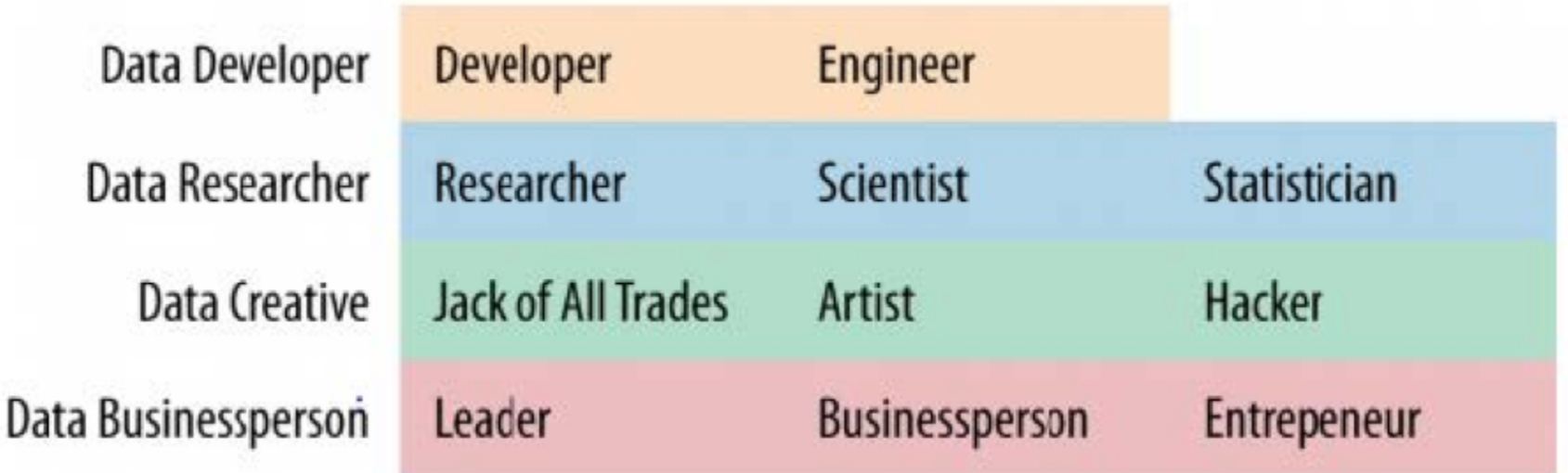
WHAT CAN YOU DO WITH DATA SCIENCE?

- **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



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

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: THE DATA SCIENCE WORKFLOW

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

THE DATA SCIENCE WORKFLOW

6 steps

1. **Identify** the Business Problem
2. **Acquire** Raw Data
3. **Parse and Mine** the Data: **data munging**
4. **Transform** the data: Feature engineering
5. **Select** and tune the Model: **Model Selection** and **Feature Selection**
6. **Present/ implement the results**: Visualization, deploy to production

STEP 1: IDENTIFY THE BUSINESS PROBLEM

- Identify Business or Product objectives,
- Identify and Hypothesize Goals
- Define Success Metrics,
- Find the right datasets

STEP 2: ACQUIRE RAW DATA

- Availability and Timeliness
- Security and Privacy
- Relevance, Bias, Sampling methods
- Sources: 3rd party platforms, in house, public data
- Heterogeneity: databases, files (csv, pdf), 3rd Party, API, ...
- Tools: 3rd party (Informatica, Jitterbit), scripts, spreadsheets, ...

STEP 3: PARSE AND MINE AKA DATA MUNGING

UNDERSTAND

- Documentation, Data dictionnaries

EXPLORE

- Perform exploratory surface analysis via filtering, sorting
- Exploratory Statistics and Visualizations
- Distribution? Trends? Outliers?

CLEAN

- Format and clean data in Python (dates, number signs, formatting)
- Invalid values
- Missing values, imbalanced sets, normalization

STEP 4: FEATURE ENGINEERING

CREATE NEW VARIABLES TO GAIN MORE INSIGHTS, MORE SIGNAL FROM THE DATA

For instance

- Date time Features: Number of days before event, week #, season, holiday, evening vs morning
- Combine, multiply, polynomial, log, inverse, ...
- Group by pattern
- Use domain knowledge
- One Hot Encoding
- Remove features to strengthen good features

ETL

Steps 3,4 are called ETL: Extract Transform Load

Evolved from batch processing in data warehouse environments

Creating the final dataset on which to apply models

- Combine
- Clean
- Complement
- Create

STEP 5: MODEL SELECTION

- What's a model?
- What's the simplest model?

STEP 5: MODEL SELECTION

What's a model?

One or a combination of algorithms

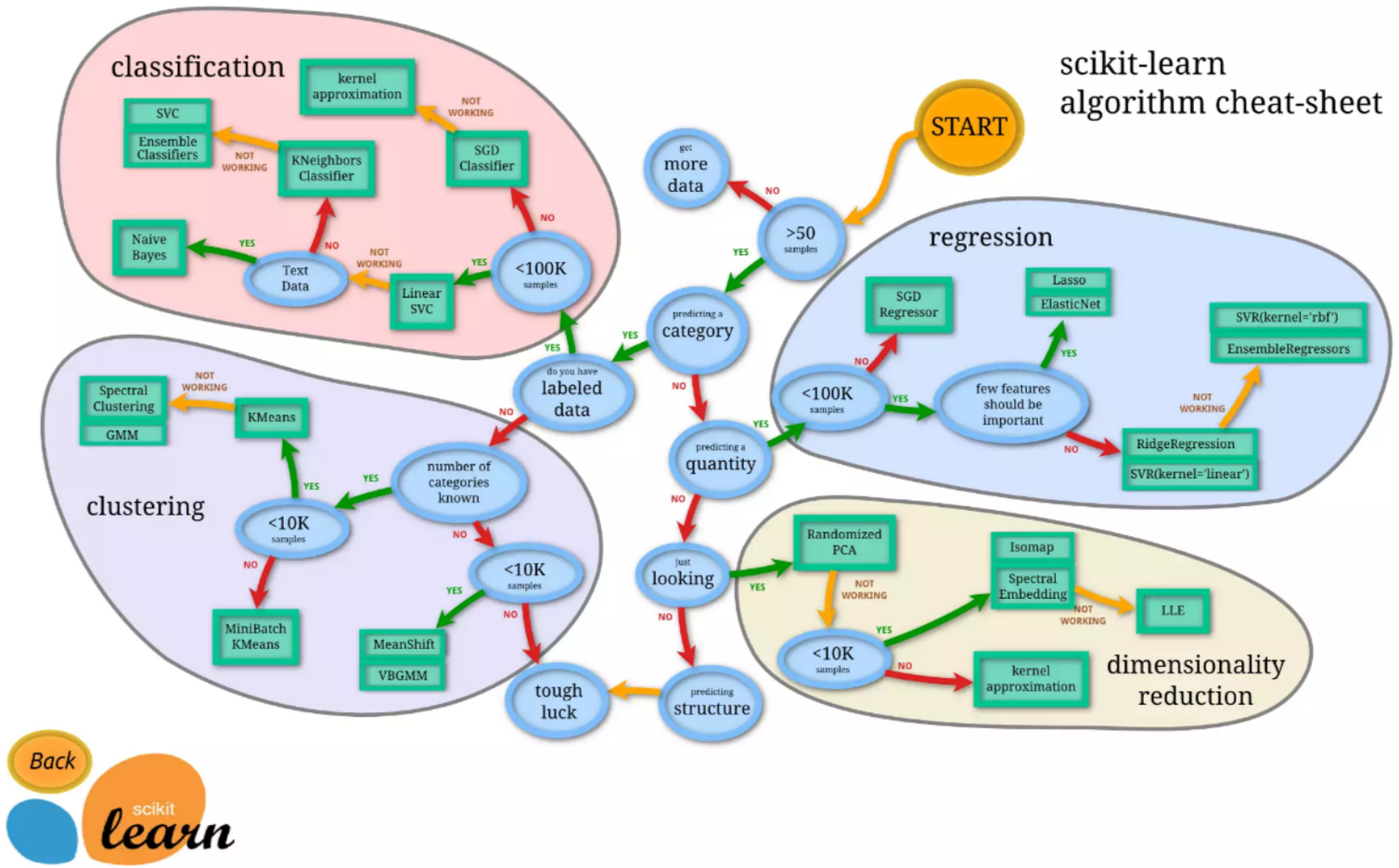
- Trained to the data
- With optimized parameters

Threshold, Linear Regression

STEP 5: MODEL SELECTION

- Select the appropriate type of models for the task: Regression, Classification, Clustering, Outlier Detection, ...
- Select the metric: precision, recall, accuracy, ..., RMSE, ranking
- Try different models, see how they perform,
- Fine tune their parameters

SCIKIT-LEARN ALGORITHM CHEAT-SHEET



STEP 6: THE RESULTS

Summarize findings with storytelling techniques

- Prediction scores
- Data visualization: plots, dashboards

Identify follow-up problems and questions

DATA VISUALIZATION

- Wind map
- What can UFO sightings tell us about extra terrestrials?
- Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance
- An analysis of the beatles
- Many other examples of great data visalizations on the Data is Beautiful Reddit

EXAMPLE: THE ONLINE RETAIL DATA SET FROM THE UCI MACHINE LEARNING REPOSITORY

The Online Retail Data Set

Problem Statement: *Using customer , determine how likely previous customers are to request a repeat delivery using*

- Order history
- Shopping carts composition
- Demographic data

Classic RFM model: Recency, Frequency, Monetary

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

ONLINE RETAIL: 1) 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.

ONLINE RETAIL: 2) IDENTIFY AND ACQUIRE THE DATA

- Ideal data vs. data that is available
- What data is available for this example? Limitations?
- What kind of questions might we want to ask about 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?
  - Secondary data = we didn't directly collect it ourselves
- Is this dataset aggregated / grouped? Can we use the aggregation or do we need to get it pre-aggregated?

ONLINE RETAIL: 3) PARSE, EXPLORE AND MINE THE DATA

1. Let's read the Data dictionary
2. First look
3. Load the data in a Notebook start exploring
4. Outliers? Valid Data?
5. Format and clean the data
6. Any missing values?
7. Normalize?

ONLINE RETAIL: 3) EXPLORE IN PYTHON

Open a new Notebook and follow along:

Jupyter Notebook: Exploration of the online retail dataset

ONLINE RETAIL: 4) FEATURE ENGINEERING

- Extract meaning and classes from product descriptions
- Define Categories
- Cancelled order
- Total amount per order
- Total amount per Customer, Country, Day ....
- Special Days: Holidays, week ends,
- One hot Encoding

=> Potential for hundreds, throusands of features

ONLINE RETAIL: 5) MODEL

- Find types of customers: Simple clustering
- Predict retention: Random Forests, Logistic Regression,
  ...
- Similarity between UK customers and Non UK

ONLINE RETAIL: 5) MODEL

The steps for model building are

- Select the appropriate model
- Build the model
- Evaluate and refine the model
- Predict outcomes and action items

=> back to step 2 (more data, other data), 3 (more cleanup), 4 (Add / Remove Features)

ONLINE RETAIL: 6) 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.

ONLINE RETAIL: 6) 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

THE DATA SCIENCE WORKFLOW

GUIDED PRACTICE

EXERCISE

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

PRE-WORK REVIEW

PRE-WORK REVIEW

- Data types
- Data structures and functions in Python
- Command line
- Git

ENVIRONMENT SETUP

DEV ENVIRONMENT SETUP

- Environment setup
- Create a Github account
- Install Python 3.5 and Anaconda
- Practice Python syntax, Terminal commands, and Pandas
- iPython Notebook test and Python review

CONDA ENVIRONMENTS

- List all the environments

```
conda info --envs
```

- Create a new environment

```
conda create --name gads python=3.5
```

- Activate environment

```
source activate gads
```

- Launch Notebook, should open in the browser at localhost:8888/

```
jupyter notebook
```

- Install missing modules. For instance:

```
conda install jupyter
```

TEST YOUR SETUP

Test your new setup using the lesson 1 starter code available in this jupyter notebook

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?

BEFORE NEXT CLASS

- Begin work on Project 1
- Read Yhat logistic regression
- Admission dataset

LINKS

- What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?
- Forbes: Data Science Falls Into Many Roles
- Read Gam Dias answer: What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data?
- The New Rules for Becoming a Data Scientist
- The Online Retail Data Set - UCI ML Repository

THANKS!

Alex Perrier

Twitter: @alexip