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

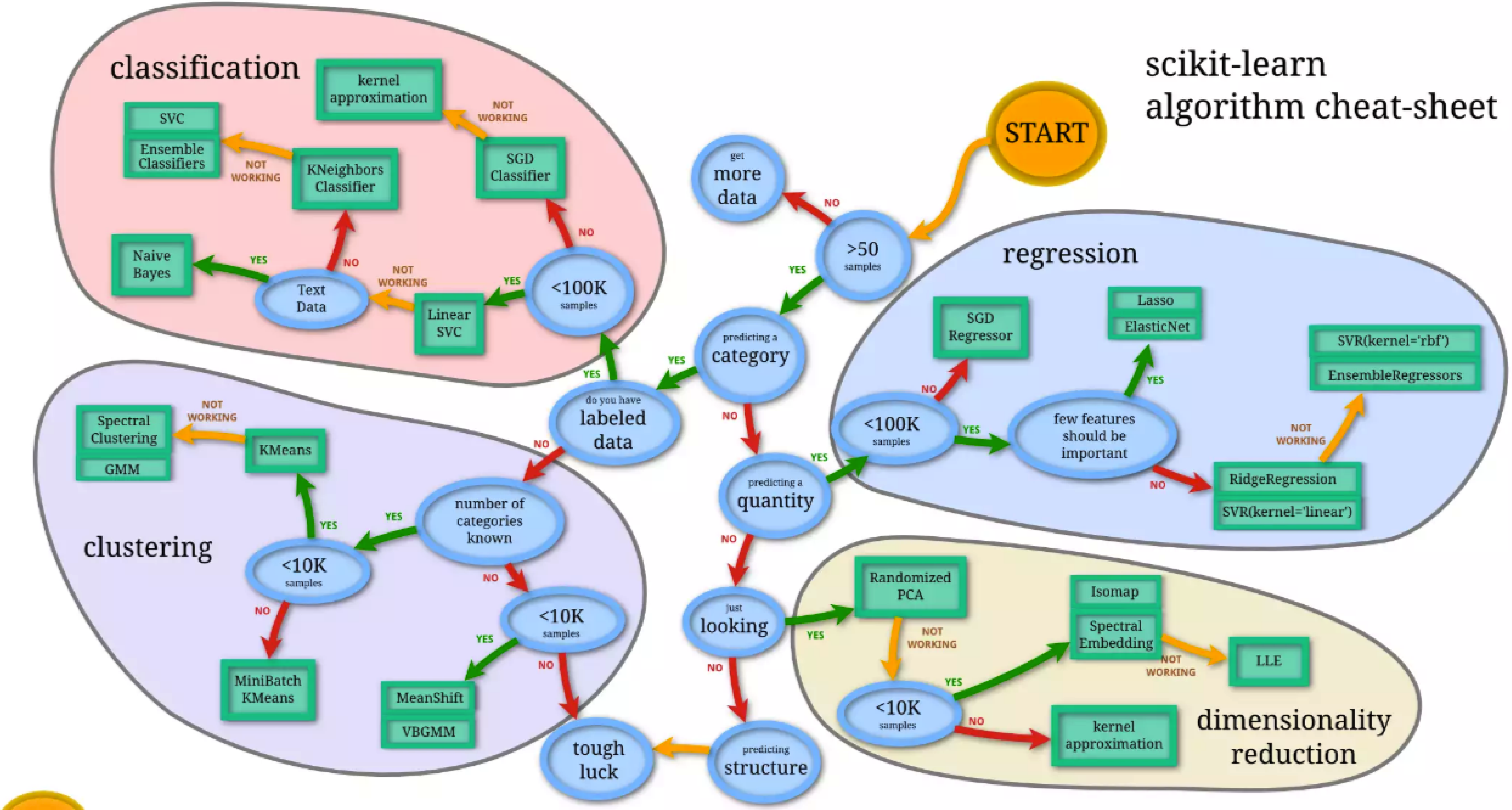
- Trained to the data
- With optimized parameters

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



START

>50 samples

NO: get more data

YES: predicting a category

YES: do you have labeled data

YES: classification

<100K samples

YES: Linear SVC

NOT WORKING: Text Data

YES: Naive Bayes

NO: KNeighbors Classifier

NOT WORKING: SVC, Ensemble Classifiers

NO: SGD Classifier

NOT WORKING: kernel approximation

NO: clustering

number of categories known

YES: <10K samples

YES: KMeans

NOT WORKING: Spectral Clustering, GMM

NO: MiniBatch KMeans

NO: <10K samples

YES: MeanShift, VBGMM

NO: tough luck

NO: predicting a quantity

YES: regression

<100K samples

NO: SGD Regressor

YES: few features should be important

YES: Lasso, ElasticNet

NO: RidgeRegression, SVR(kernel='linear')

NOT WORKING: SVR(kernel='rbf'), EnsembleRegressors

NO: just looking

YES: dimensionality reduction

Randomized PCA

NOT WORKING: <10K samples

YES: Isomap, Spectral Embedding

NOT WORKING: LLE

NO: kernel approximation

NO: predicting structure

tough luck

Back

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 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: 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

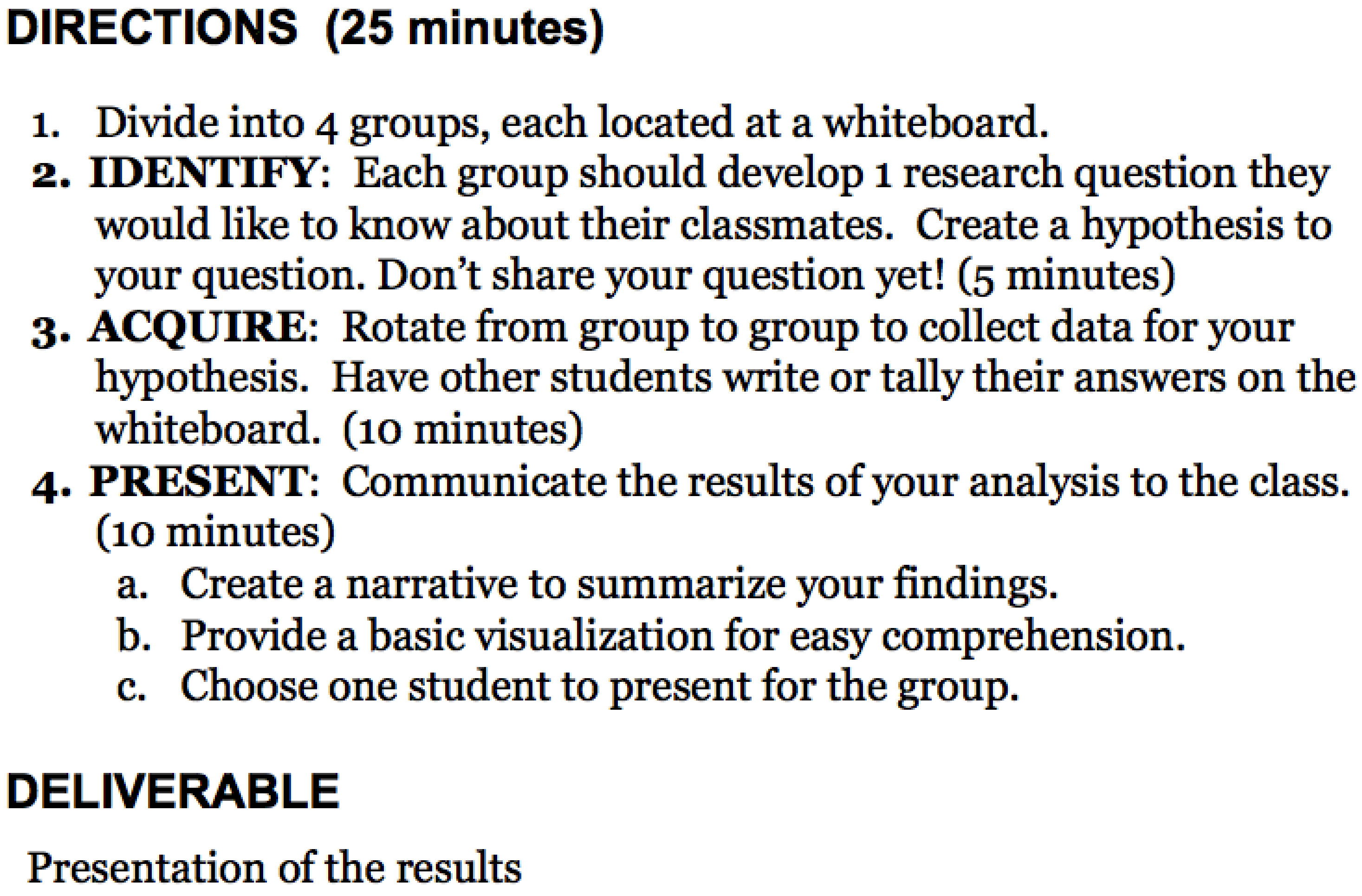
- Summarize findings with narrative and storytelling techniques
- Refine your visualizations for broader comprehension
- Present both limitations and assumptions
- Determine the integrity of your analyses
- Consider the degree of disclosure for various stakeholders
- Test and evaluate the effectiveness of your presentation beforehand

GUIDED PRACTICE

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

DEMO

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

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

https://github.com/generalassembly-studio/ds-curriculum/blob/master/lessons/lesson-01/code/starter-code/starter-code-1.ipynb

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

BEFORE NEXT CLASS

- Project: Begin work on Project 1 https://github.com/generalassembly-studio/ds-curriculum/blob/master/projects/unit-projects/project-1/starter-code/project1-starter.ipynb
- Yhat logistic regression http://blog.yhat.com/posts/logistic-regression-and-python.html
- Admission dataset https://github.com/generalassembly-studio/ds-curriculum/blob/master/projects/unit-projects/project-1/assets/admissions.csv

REFERENCES

- 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

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

THANKS!

Alex Perrier

twitter: @alexip

aperrier@berklee.edu

linkedin.com/in/alexisperrier