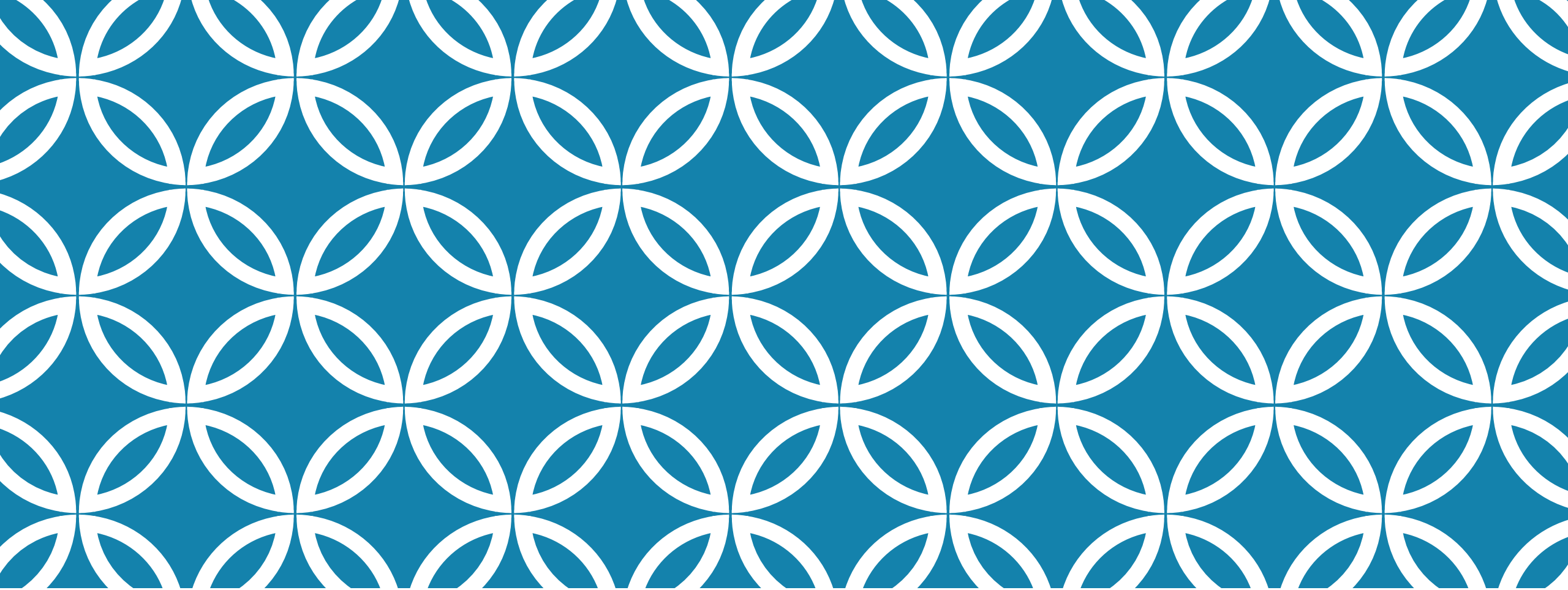


DATA SCIENCE OVERVIEW

OUTLINE

Overview

- Motivation
- Big Picture

Stages

- Selection of Topic / Problem
- Data Acquisition
- Exploratory Data Analysis
- Modeling and Analysis
- Experimental Design
- Conclusions

Documentation

- Example: RMD reports

MOTIVATION FOR DATA SCIENCE

Data is ubiquitous

- Recently we have the capability to acquire and analyze big data

Leveraging Data

- Analyzing data can yield evidence to support decisions
  - Data-Driven Decisions

DATA SCIENCE PIPELINE

I. Identify Problem to Solve
   I. Hypothesis
   II. Decision to support
II. Data Acquisition
   I. Cleaning
III. Exploratory Data Analysis
IV. Modeling and Analysis
V. Experimental Design
VI. Concluding Remarks

I. SELECTION PROBLEM

Identify a Problem to Solve

Categories

- Decision or Prediction based
  - EG. Prediction stock market based on twitter data
  - EG. Decide whether to buy / sell a stock based on twitter data
- Exploratory or Investigation
  - EG. Are there measureable patterns in twitter data related to stocks

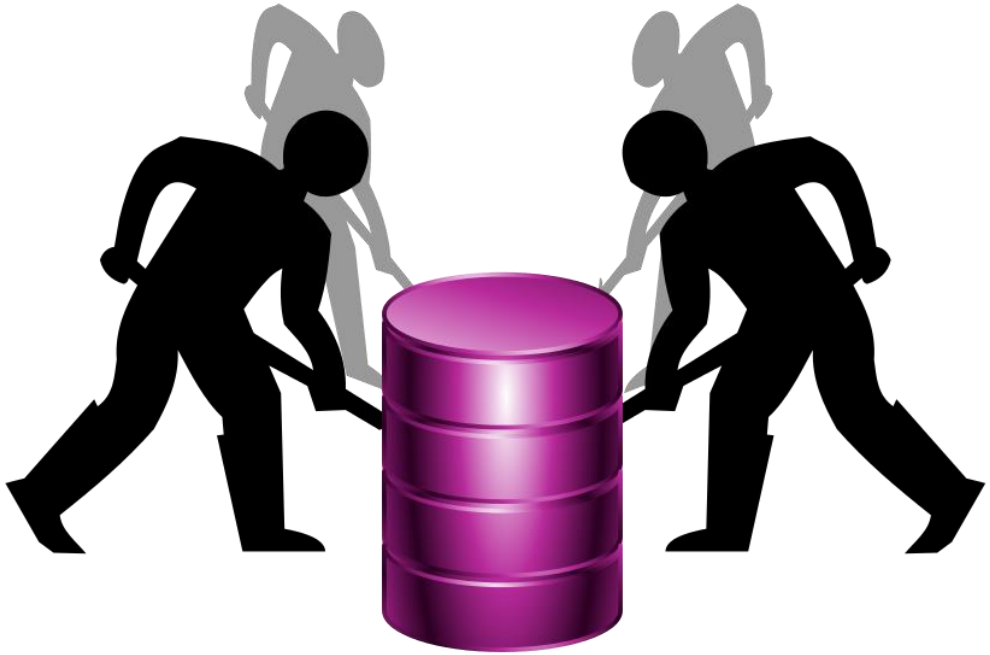
DATA ACQUISITION

Finding a data set

- Very important
- Must be able to support your investigation
  - Is there enough data?
  - Are the correct variables in the set?
  - Can you acquire the data?

Data is dirty … time to clean / tidy

- Missing Values
- Missing labels
- Text cleaning
  - Think twitter data ☹



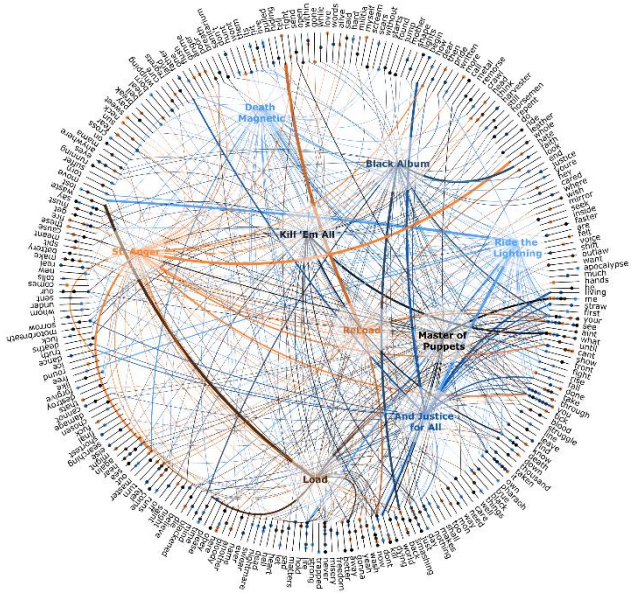
EXPLORATORY DATA ANALYSIS

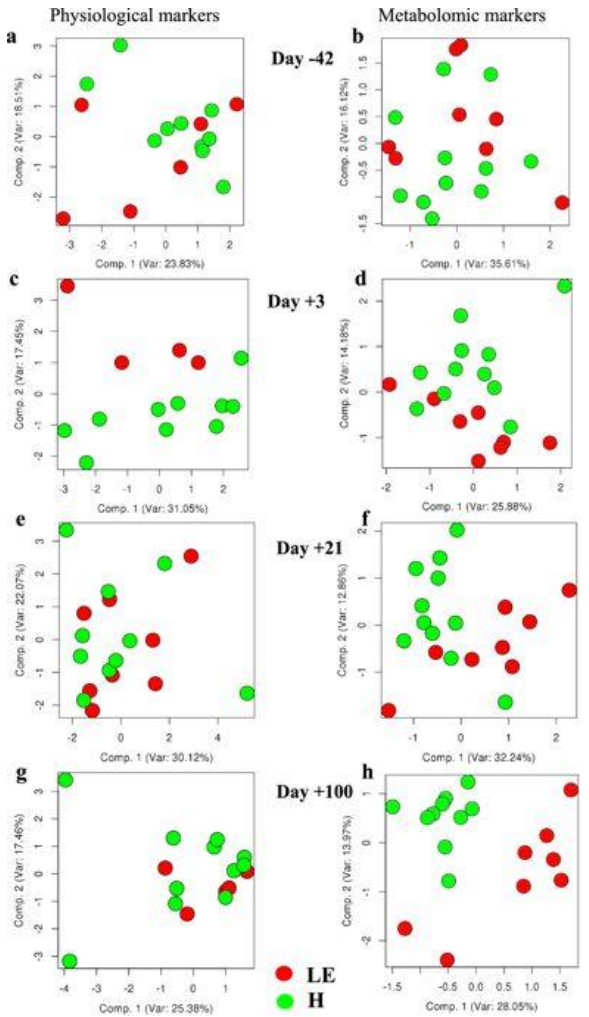
Finding patterns in your data

Focus on relevant variables

- Correlations
- Associations

Data Viz

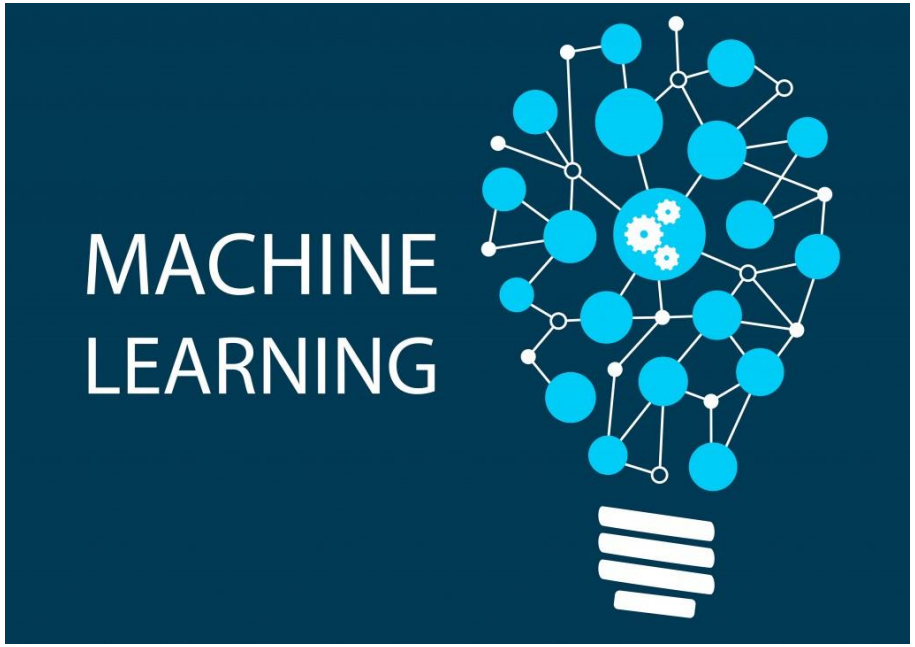




MODELING AND ANALYSIS

Based on EDA, determine a set of models that will help in your investigation

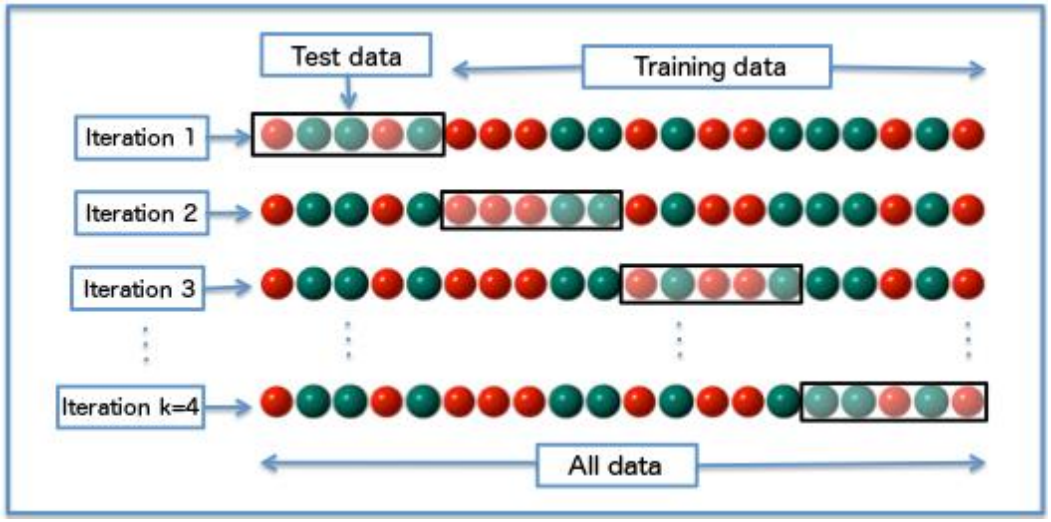
- Must determine which models / analytical techniques are appropriate
  - Based on data type: quantitative, qualitative, …
  - Based on EDA: which variables are relevant
  - Based on Category of Problem
    - Prediction
    - Investigation



EXPERIMENTAL DESIGN AND RESULTS

Experiments: Validation of Models / Analysis

- Design an appropriate validation scheme
  - Separate training and testing sets
  - Cross-validation Techniques
    - K-fold
    - Random Sample



Results

- Measures of error accuracy of each trial
  - RMSE, Confusion Matrix, ROC curve, …
- Measures of “repeatability” or “variability” of the trial results.
  - Assess variability across trial folds.

CONCLUSIONS

Relate Results back to Problem Statement

Do results support a decision / prediction?

Do results identify measurable patterns in the data?

DOCUMENTATION

Documenting a Data Science Project is VERY Important

- Document all stages
- Document “good” and “bad” results / findings
- Documentation should include good viz: figures, tables, plots, …

Reports

- R-markdown
  - Everything is contain in 1 file!!!
  - Otherwise multiple files: coding, report, figures, jpgs, …