GENERAL ASSEMBLY

EXPERIMENTAL DESIGN AND PANDAS

Stefan Jansen

DAT-NYC-43

LEARNING OBJECTIVES

- Define a problem and types of data
- Identify data set types
- Define the data science workflow
- Apply the data science workflow in the pandas context
- Create an iPython Notebook to import, format, and clean using the Pandas library

PRE-WORK

Command Line & GitHub Code-along

- Command Line (Bash / Terminal) review:
  - Open Terminal & try some new commands:
    - Get help: ‘man’
    - Use parameters: ‘ls’, ‘ls –l’, ‘ls –a’, ‘ls –al’; ‘ls *dir_name*’
    - Use ‘which’, ‘pwd’, ‘python’
    - Use ‘echo $PATH’

Command Line & GitHub REVIEW

- Pull from course repository:
  - cd *DAT-NYC-38*; 'git pull'
  - may need to commit your changes first;
  - see here how to overwrite local changes

- Set up homework repository, add collaborators and make changes

- Configure:
  - git config –-global user.name 'Your Name'
  - git config --global user.email your@email.com
  - Check: git config –list

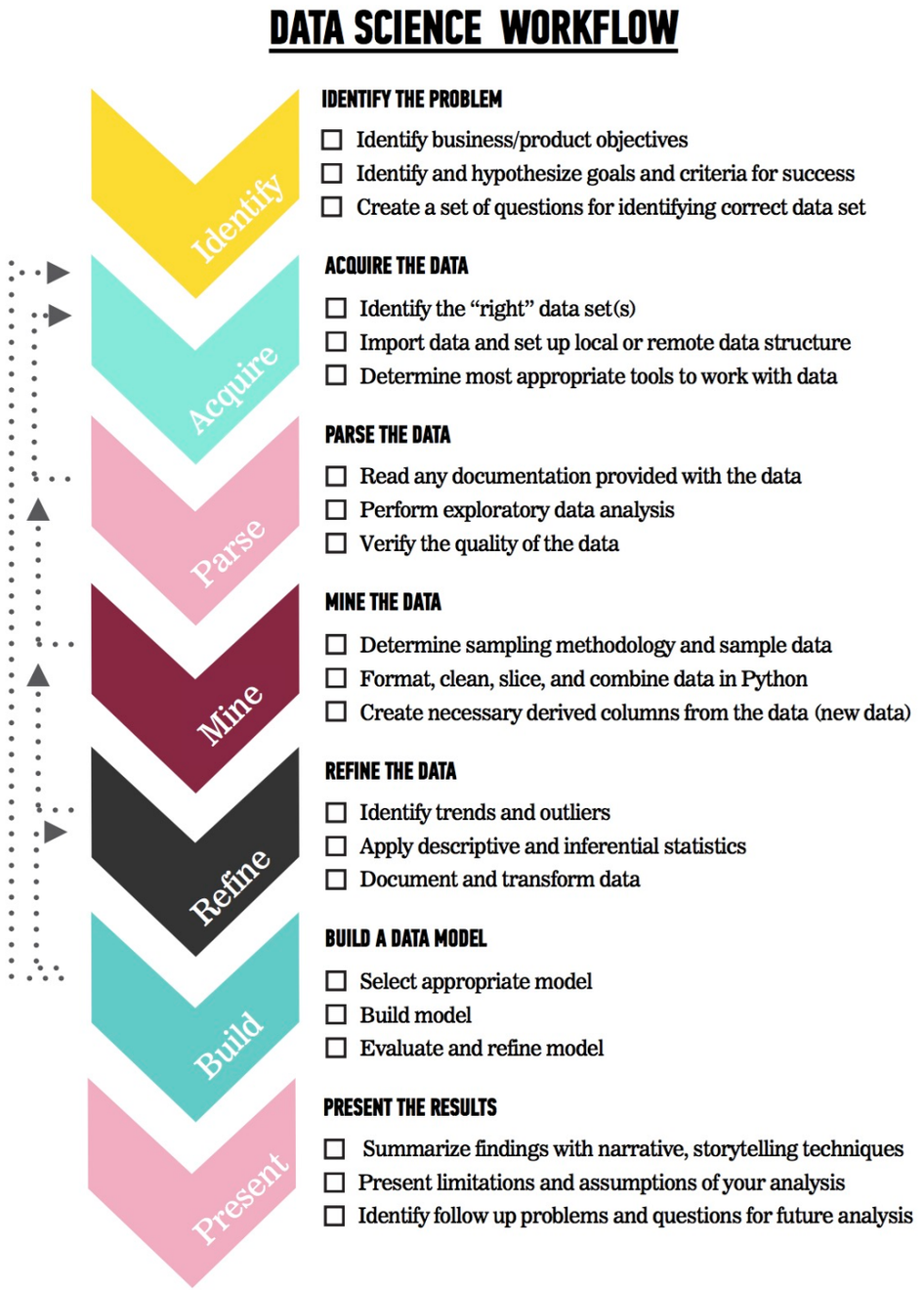
- Create ssh keys if you don't always want to type your username & pwd.

THE DATA SCIENCE WORKFLOW

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



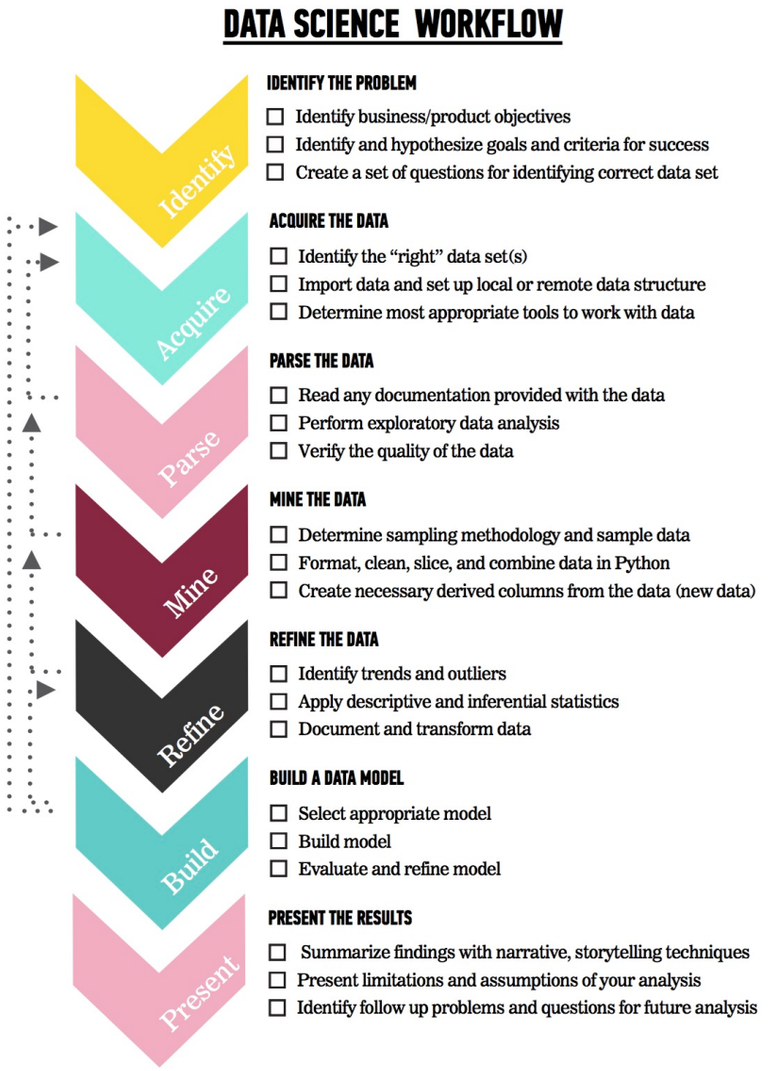
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- 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

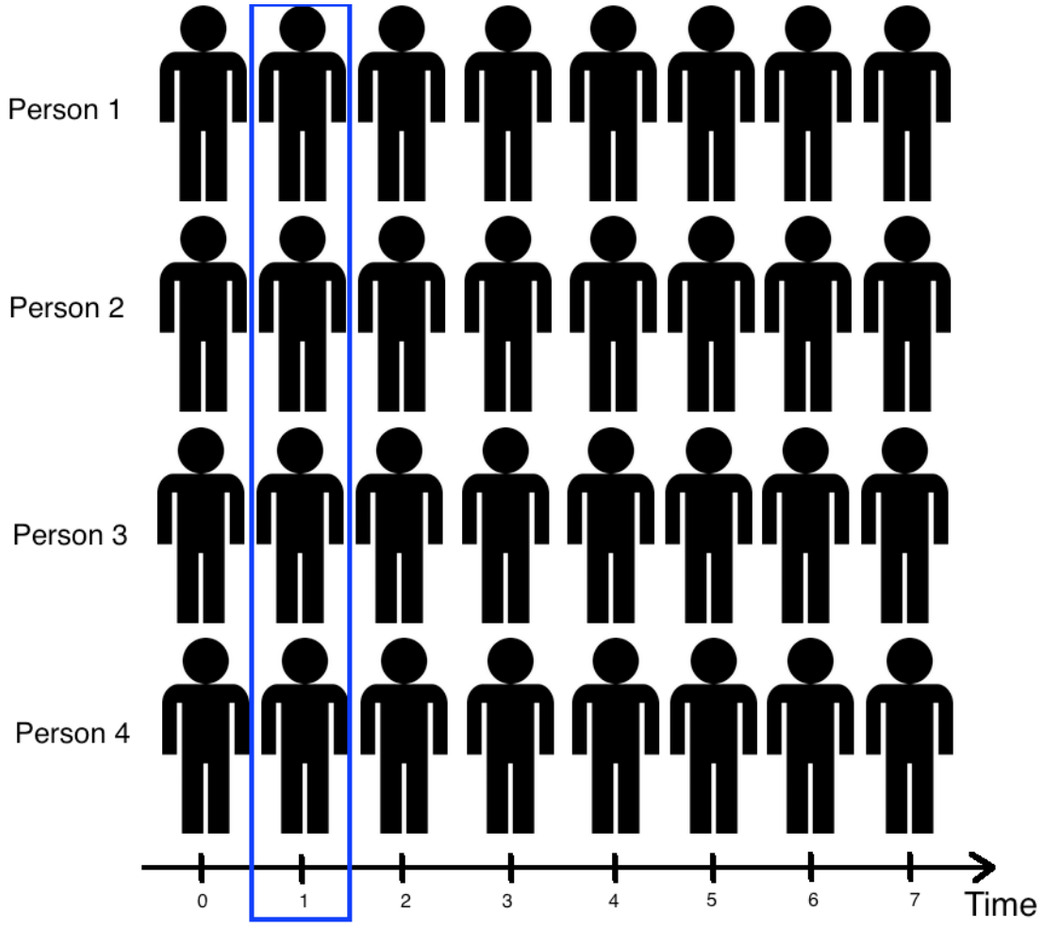
- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

WHY DATA TYPES MATTER

- Different data types have different limitations and strengths.
- Certain types of analyses aren’t possible with certain data types.

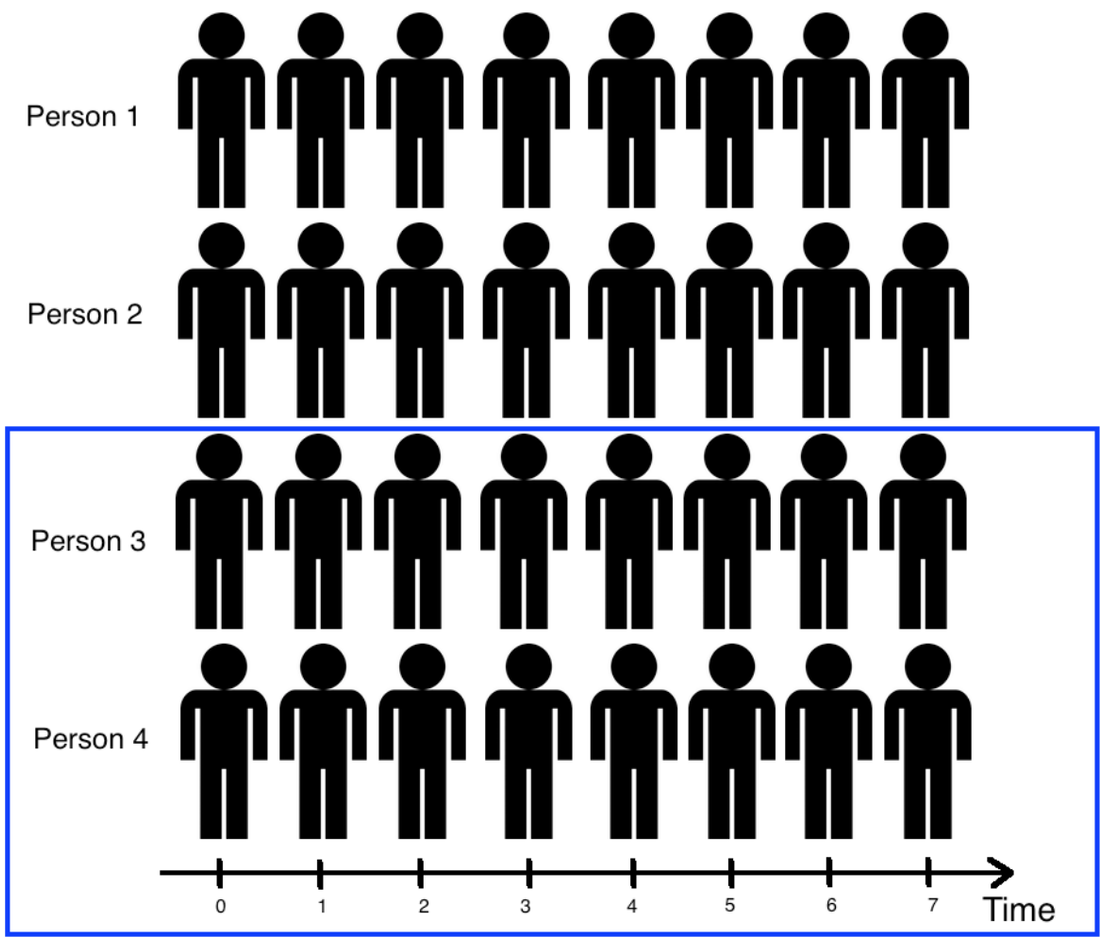
CROSS-SECTIONAL DATA

- All information is determined at the same time; all data from same time period.
- Issues: There is no distinction between exposure and outcome
- Strengths
  - Often population based
  - Generalizability
  - Reduce cost vs other data collection methods
- Weaknesses
  - Separation of cause & effect difficult (or impossible)
  - Variables/cases with long duration are over-represented (survivorship bias)

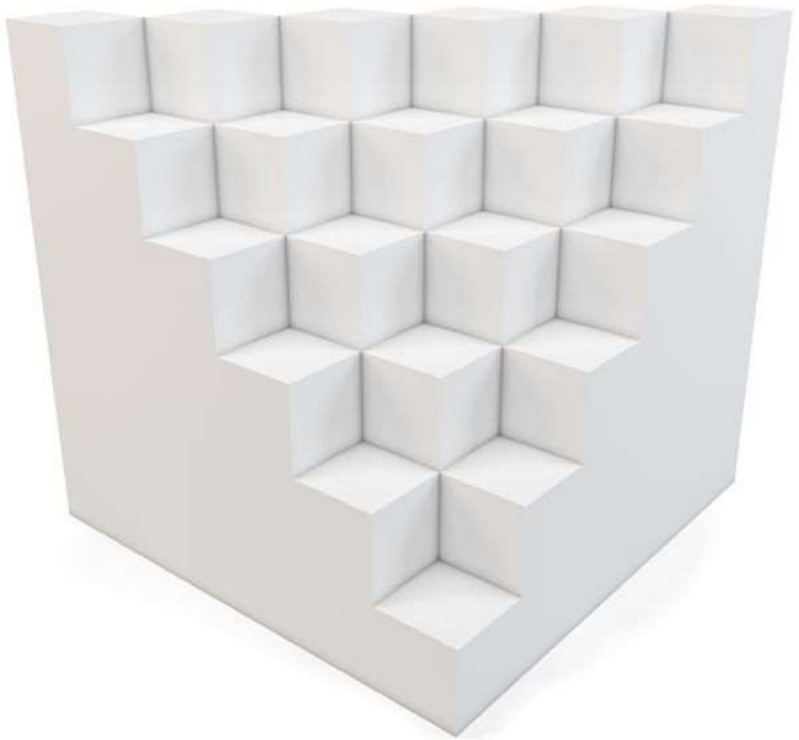


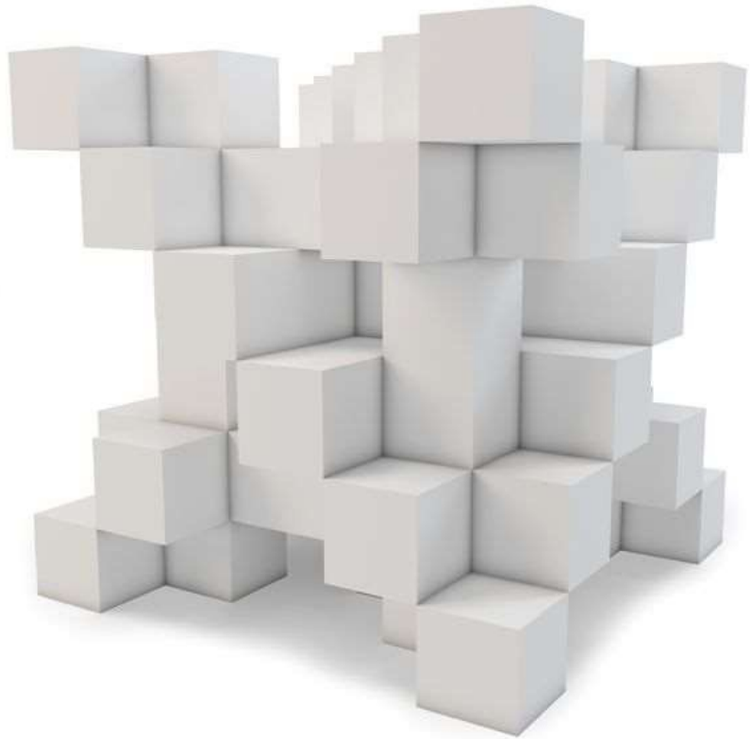
TIME SERIES/LONGITUDINAL DATA

- The information is collected over a period of time
- Strengths
  - Unambiguous temporal sequence - exposure precedes outcome
  - Multiple outcomes can be measured
- Weaknesses
  - Expense
  - Takes a long time to collect data
  - Vulnerable to missing data



DATA: STRUCTURED vs UNSTRUCTURED





dny3d ©
123RF.com

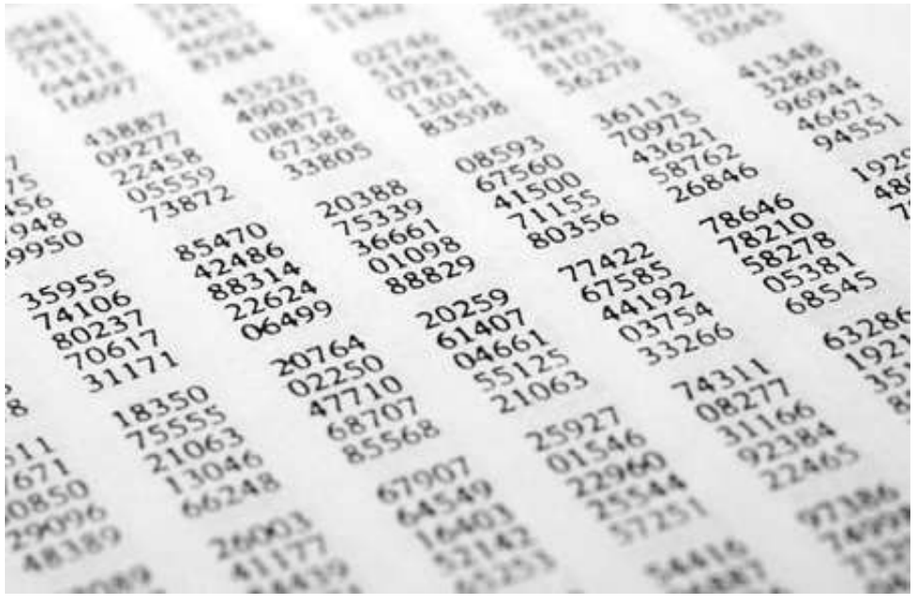
UNSTRUCTURED DATA

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

OVERVIEW OF THE DATA SCIENCE WORKFLOW

REFINE THE DATA

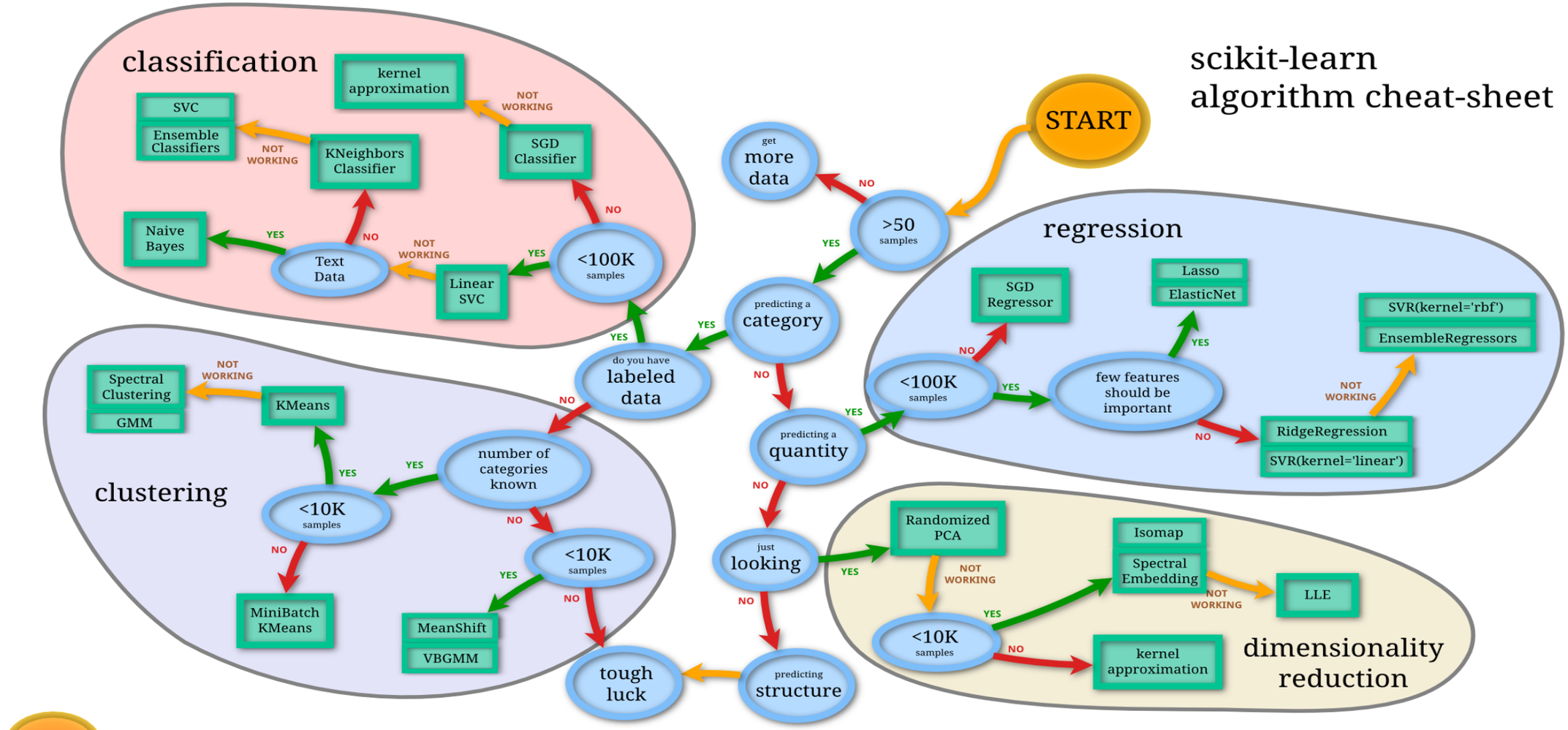
- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

OVERVIEW OF THE DATA SCIENCE WORKFLOW

BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

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

ACTIVITY: DATA SCIENCE WORKFLOW

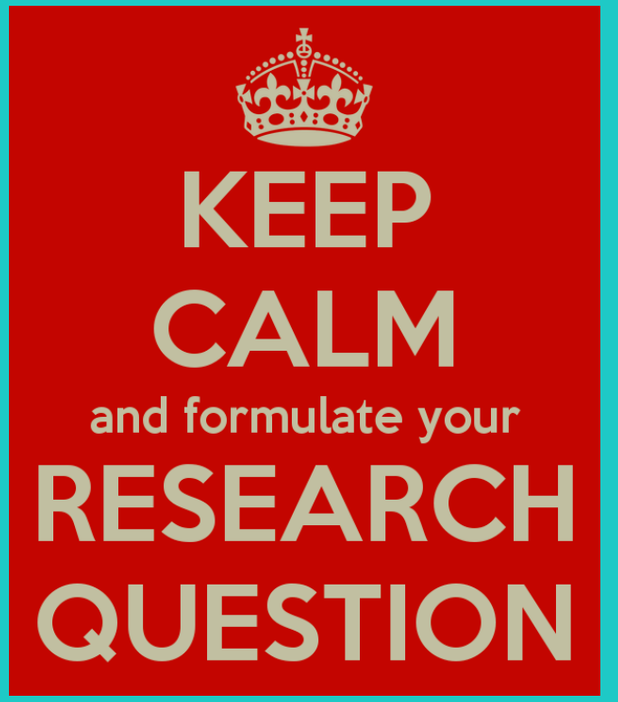
DIRECTIONS (20 minutes)

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. (5 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

ASKING A GOOD QUESTION



WHY DO WE NEED A GOOD QUESTION?

- “A problem well stated is half solved.” - Charles Kettering
  - 186 patents, Head of GM research 1920 - 47
- Sets yourself up for success as you begin analysis
- Establishes the basis for reproducibility
- Enables collaboration through clear goals
- Allows to better assess value of potential results

WHAT IS A GOOD QUESTION?

- Similar to SMART Goals:
  - S: specific
  - M: measurable
  - A: attainable
  - R: reproducible (and relevant)
  - T: time-bound

WHAT IS A GOOD QUESTION?

- **S** — **Specific:** State exactly what you want to accomplish (Who, What, Where, Why)
- **M** — **Measurable:** How will you demonstrate and evaluate the extent to which the goal has been met?
- **A** — **Achievable:** stretch and challenging goals within ability to achieve outcome. What is the action-oriented verb?
- **R** — **Relevant:** How does the goal tie into your key responsibilities? How is it aligned to objectives?
- **T** — **Time-bound:** Set 1 or more target dates, the “by when” to guide your goal to successful and timely completion (include deadlines, dates and frequency)

WHAT IS A GOOD QUESTION?

- Specific: The dataset and key variables are clearly defined.
- Measurable: The type of analysis and major assumptions are articulated.
- Attainable: The question you are asking is feasible for your dataset and is not likely to be biased.
- Reproducible: Another person (or future you) can read and understand exactly how your analysis is performed.
- Time-bound: You clearly state the time period and population for which this analysis will pertain.

DEMO

DIAGRAMMING AN AIM

EXAMPLE AIM

- Determine the association of foods in the home with child dietary intake.
- Using one 24-hour recall from the cross-sectional NHANES* 2007-2010, we will determine the factors associated with food available in the homes of American children and adolescents.
- We will test if reported availability of fruits, dark green vegetables, low fat milk or sugar sweetened beverages available in the home increases the likelihood that children and adolescents will meet their USDA recommended dietary intake for that food.

*NHANES: National Heath and Nutrition Examination Survey

HYPOTHESIS

- Children will be *more likely* to meet the USDA recommended intake level when food is always available in their home compared to *rarely or never*.

SPECIFIC

- How data was collected:
  - 24-hour recall, self-reported
- What data was collected:
  - Fruits, dark green vegetables, low fat milk or sugar sweetened beverages, always vs. rarely available
- How data will be analyzed:
  - Using USDA recommendations as a gold-standard to measure the association
- The specific hypothesis & direction of the expected associations:
  - Children will be more likely to meet their recommended intake level

MEASURABLE

- Determine the association of foods in the home with child dietary intake.
- We will test if the reported availability of certain foods increases the likelihood that children and adolescents will meet their USDA recommended dietary intake for food.

ATTAINABLE

- Cross-sectional data has inherent limitations; one of the most common is that causal inference is typically not possible.

- Note that we are determining association, not causation.

REPRODUCIBLE

- With all the specifics, it would be straightforward to pull the data from NHANES and reproduce the analysis.

TIME BOUND

- Using one 24-hour recall from NHANES 2007-2010, we will determine the factors associated with food available in the homes of American children and adolescents.

CONTEXT IS IMPORTANT

- The previous example laid out research goals.
- In a business setting, you will need to articulate business objectives.
- Example: Success for the Netflix recommendation engine may be if 70% of customers over the age of 18 select a movie from the recommended queue during Q3 of 2015.
- Regardless of setting, start your question with the SMART framework to help achieve your objectives.

ACTIVITY: KNOWLEDGE CHECK

IMPROVE THE FOLLOWING GOAL STATEMENT (5 minutes)

1. How is the following NOT using the SMART framework? Why? What is missing?

   I am looking to see if there is an association with number of passengers with carry on luggage and delayed take-off time.

   Think for yourself than results share with your table.

DELIVERABLE

Answers to the above questions

DATA SCIENCE WORKFLOW: ACQUIRE & PARSE

DATA SCIENCE WORKFLOW: ACQUIRE & PARSE

- For the remainder of class, we'll talk about steps 2 & 3 of the data science workflow: acquire and parse
- We'll be using iPython Notebook
- First a demo, then a codealong
- Finally, some hands on practice in a lab

ACQUIRE

- Do we have the “right” dataset for our problem?
- Questions to ask:
  - What type of data is it, cross-sectional or longitudinal?
  - How well was the data collected?
  - Is there much missing data?
  - Was the data collection instrument validated and reliable?
  - Is the dataset aggregated?
  - Do we need pre-aggregated data?

LOGISTICS OF ACQUIRING YOUR DATA

- Data can be acquired through a variety of sources
- Web (Google Analytics, HTML, XML)
- File (CSV, XML, TXT, JSON)
- Databases (SQL, NOSQL, etc)
- Today, we'll use a CSV (comma separated file)

PARSE: UNDERSTANDING YOUR DATA

- You need to understand what you're working with.
- To better understand your data
  - Create or review the data dictionary
  - Describe data structure and information being collected
  - Explore variables and data types

INTRO TO DATA DICTIONARIES AND DOCUMENTATION

- Data dictionaries help judge the quality of the data.
- They also help understand how it's coded.
    - Does gender = 1 mean female or male?
    - Is the currency dollars or euros?
- Data dictionaries help identify any requirements, assumptions, and constraints of the data.
- They make it easier to share data.

DATA DICTIONARY EXAMPLE: KAGGLE TITANIC DATA

```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)

SPECIAL NOTES:
Pclass is a proxy for socio-economic status (SES)
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch)
some relations were ignored.  The following are the definitions used
for sibsp and parch.

Sibling:  Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard
Titanic
Spouse:   Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances
Ignored)
Parent:   Mother or Father of Passenger Aboard Titanic
Child:    Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins,
nephews/nieces, aunts/uncles, and in-laws.  Some children travelled
only with a nanny, therefore parch=0 for them.  As well, some
travelled with very close friends or neighbors in a village, however,
the definitions do not support such relations.
```

GUIDED PRACTICE

WRITE A RESEARCH QUESTION WITH RAW DATA

WRITE A RESEARCH QUESTION WITH RAW DATA

DIRECTIONS (10 minutes)

1. Individually, look at the data from Kaggle's Titanic competition and write a high quality research question.
2. Make sure you answer the following questions:
   a. What type of data is this, cross-sectional or longitudinal?
   b. What will we be measuring?
   c. What is the SMART aim for this data?
3. When finished, split into pairs and share your answers with each other.

DELIVERABLE

Research Question

CODEALONG

NUMPY AND PANDAS INTRO

NUMPY AND PANDAS INTRO

- What are Numpy and Pandas? Python packages
- Pandas is built on Numpy.
- Numpy uses arrays (lists) to do basic math and slice and index data.
- Pandas uses a data structure called a DataFrame.
- Dataframes are similar to Excel tables; they contain rows and columns.

NUMPY AND PANDAS INTRO

	A	B	C	D
2014-01-01	0.731803	2.318341	-0.126191	-0.903675
2014-01-02	0.161877	-0.892566	0.967681	-1.514520
2014-01-03	0.776626	1.797420	0.916972	0.634322
2014-01-04	2.020242	-0.763612	1.239145	-0.919727
2014-01-05	0.772058	0.417369	-0.957359	-0.916665
2014-01-06	-1.670217	-3.249906	2.017370	1.674340

6 rows × 4 columns

NUMPY AND PANDAS INTRO

- With these packages, you can select pieces of data, do basic operations, calculate summary statistics.

- Follow along and code along as we learn about Numpy and Pandas.

DEMO

LAB WALKTHROUGH

LESSON 2 LAB WALKTHROUGH

- In this lab, you will look at meteorological data ozone and temperature.
- By the end of the lab, you will:
  - Load datasets
  - Check basic features of the data
  - Find and drop missing values
  - Find basic stats like mean and max

TOPIC REVIEW

REVIEW

- Let’s go through the lab. Any questions?
- Today, we’ve talked about
    - Defining a problem
    - Types of data
    - Acquiring and parsing data
    - Using Pandas

BEFORE NEXT CLASS

DUE BEFORE NEXT CLASS

- Project: Unit 1: Write a Research Question
- Any questions?

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET