Research Design and *pandas*

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

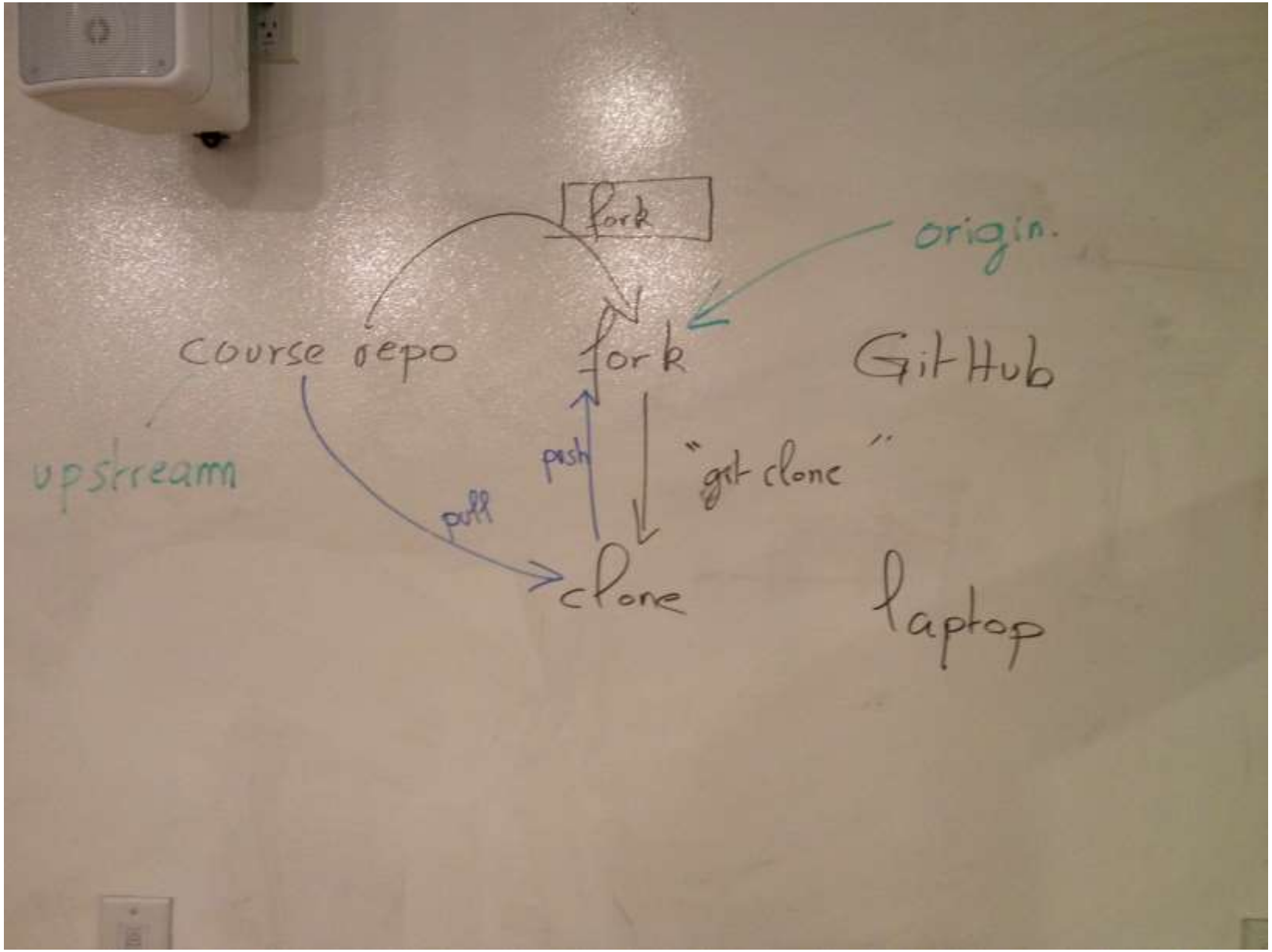
Learning Objectives

After this lesson, you should be able to:

- Define the data science workflow
- Define a problem and types of data
- Identify dataset types
- Apply the data science workflow in the *pandas* context
- Write an Jupyter notebook to import, format, and clean data using the *pandas* library

Announcements and Exit Tickets

Review


fork
origin.
course repo
fork
GitHub
upstream
push
"git clone"
pull
clone
laptop

Today

Today, we are covering Research Design and introducing the *pandas* library

Research Design and Data Analysis	Research Design	Data Visualization in *pandas*	Statistics	Exploratory Data Analysis in *pandas*
Foundations of Modeling	Linear Regression	Classification Models	Evaluating Model Fit	Presenting Insights from Data Models
Data Science in the Real World	Decision Trees and Random Forests	Time Series Data	Natural Language Processing	Databases

Here's what's happening today:

- Announcements and Exit Tickets
- Review
- Pre-Work
- The Data Science Workflow
- ❶ Identify the problem
  - The Why's and How's of a Good Question
  - The SMART Goals Framework for Data Science
- ❷ Acquire the Data
  - Data Types
  - Logistics of Acquiring Data
  - SF Housing Dataset from Zillow
  - Tidying Data
- ❸ Parse the Data
  - Documentation and Data Dictionaries
  - Codealong – Introduction to *pandas* and tidying up the SF housing dataset
- Lab – Research Design and *pandas*
- Review
- Exit Tickets

Pre-Work

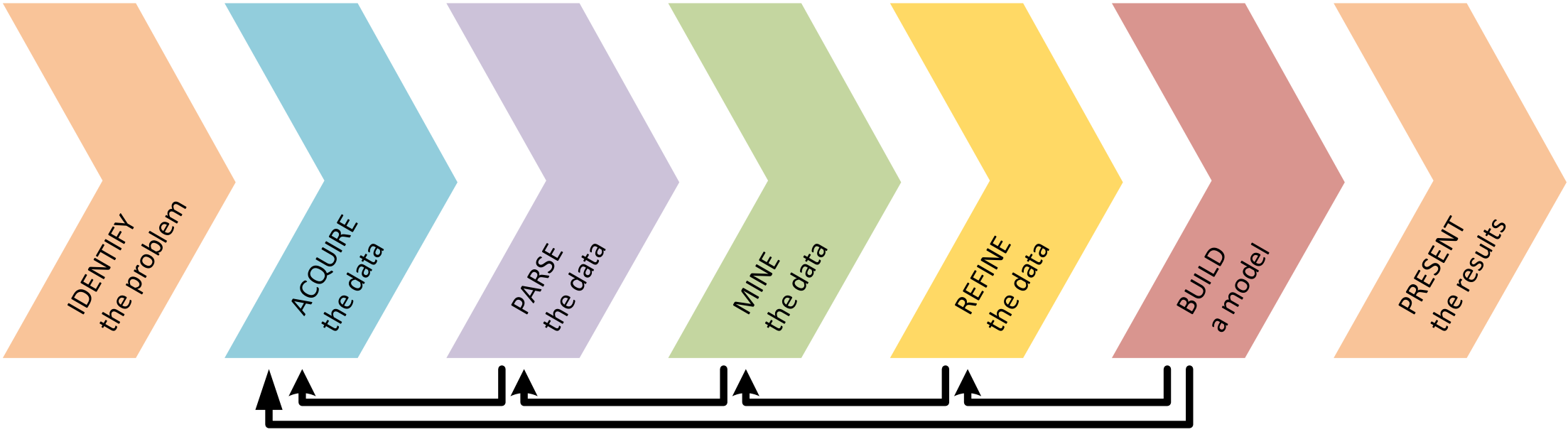
Pre-Work

- Complete your development environment setup and practice the different workflows used in the course
- Complete the onboarding pre-work
- Look into the first unit project
- Start ideating about your final project's topic

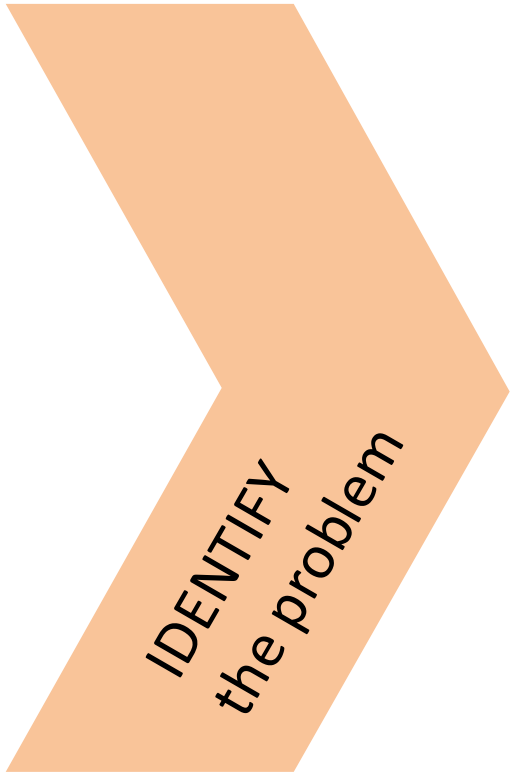
Q & A

The Data Science Workflow

The Data Science Workflow (a.k.a., the Data Science Pipeline)



❶ Identify the Problem



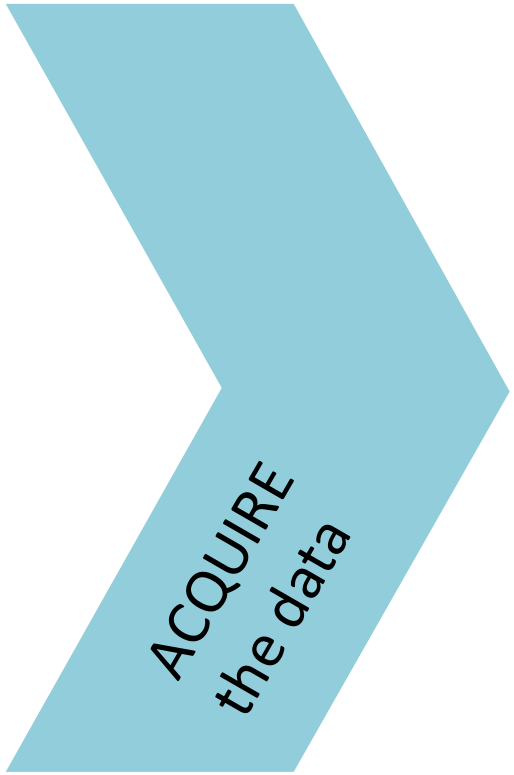
- Identify the Problem
  - Identify business/product objectives
  - Identify and hypothesize goals and criteria for success
  - Create a set of questions for identifying correct dataset

❶ Identify the Problem

The Why's and How's of a Good Question

Corina Rosu © 123RF.com

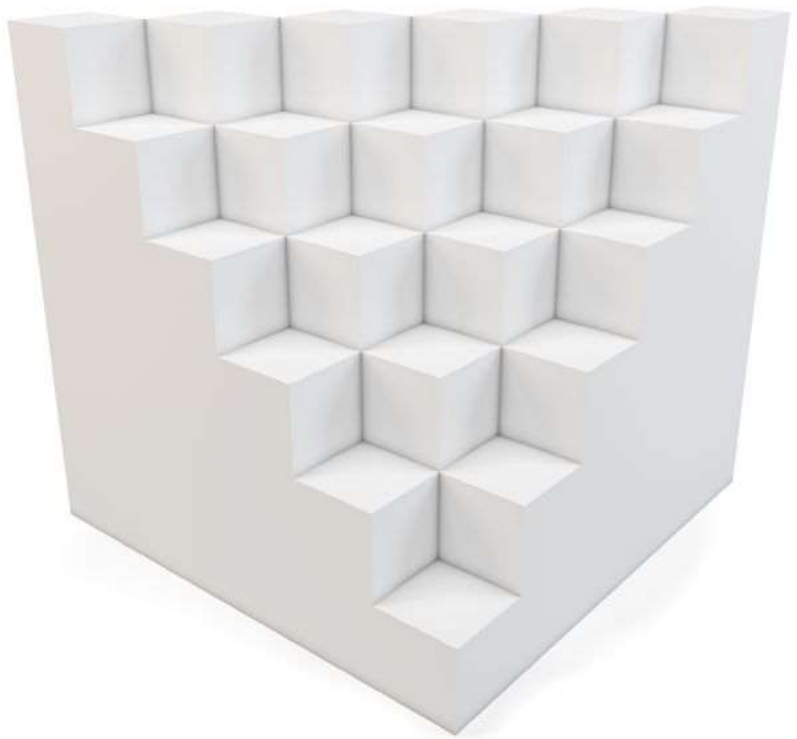
❷ Acquire the Data

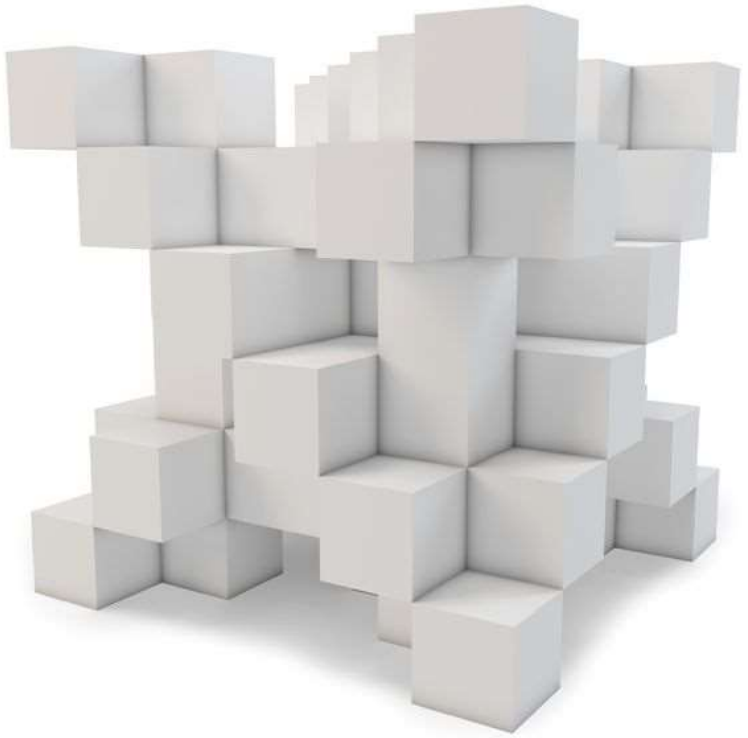


- Acquire the Data
  - Identify the “right” dataset(s)
  - Import data and set up local or remote data structure
  - Determine most appropriate tools to work with data

❷ Acquire the Data

Data can be either unstructured or structured data





dny3d © 123RF.com

❷ Acquire the Data

What's an example of unstructured data?

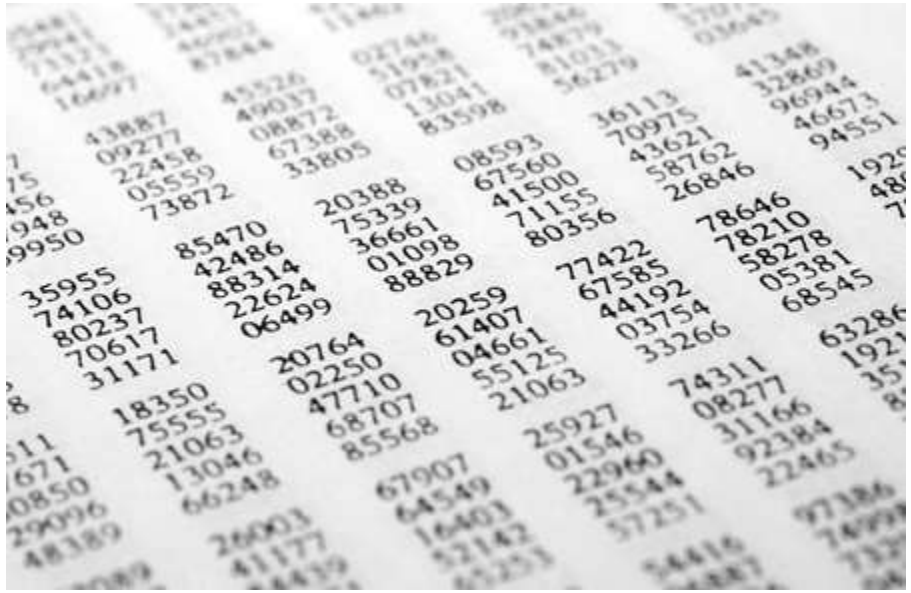
Bundit Chuangboonsri © 123RF.com

- Sessions 13 and 14
  - Natural Language Processing

❷ Acquire the Data

Most of the course will focus on structured data

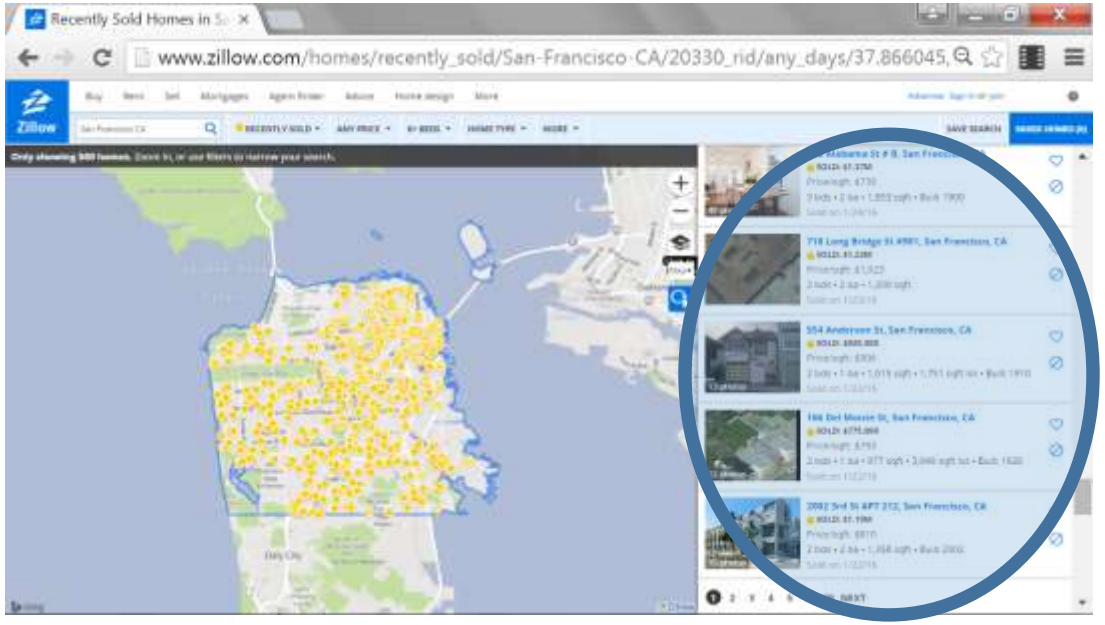
- Unit 2
  - Linear Regression (sessions 6 and 7)
  - Classification and Logistic Regression (session 8 and 9)
- Unit 3
  - Decision Trees and Random Forests (session 12)
  - Time Series (session 15 and 16)



milosb © 123RF.com

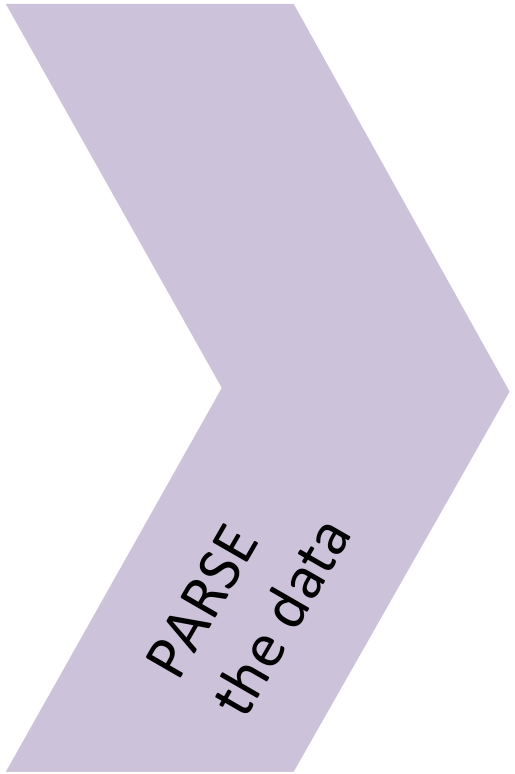
❷ Acquire the Data

Raw structured data is Messy™…




```
<div class="property-info"
id="yui_3_18_1_1_1456167242885_71870"><strong
id="yui_3_18_1_1_1456167242885_71869"><dt class="property-
address" id="yui_3_18_1_1_1456167242885_71868"><a
href="/homedetails/149-Shipley-St-San-Francisco-CA-
94107/15147894_zpid/" class="hdp-link routable" title="149
Shipley St, San Francisco, CA Real Estate"
id="yui_3_18_1_1_1456167242885_71873">149 Shipley St, San
Francisco, CA</a></dt></strong><dt class="listing-type zsg-
content_collapsed"
id="yui_3_18_1_1_1456167242885_71875"><span class="zsg-
icon-recently-sold type-icon"></span>Sold: $1.18M</dt><dt
class="zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71877">Price/sqft:
$1,116</dt><dt class="property-data"
id="yui_3_18_1_1_1456167242885_71880"><span class="beds-
baths-sqft">3 bds • 2 ba • 1,057 sqft</span><span
class="built-year" id="yui_3_18_1_1_1456167242885_71879"> •
Built 1992</span></dt><dt class="sold-date zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71975">Sold on
2/22/16</dt></div>
```

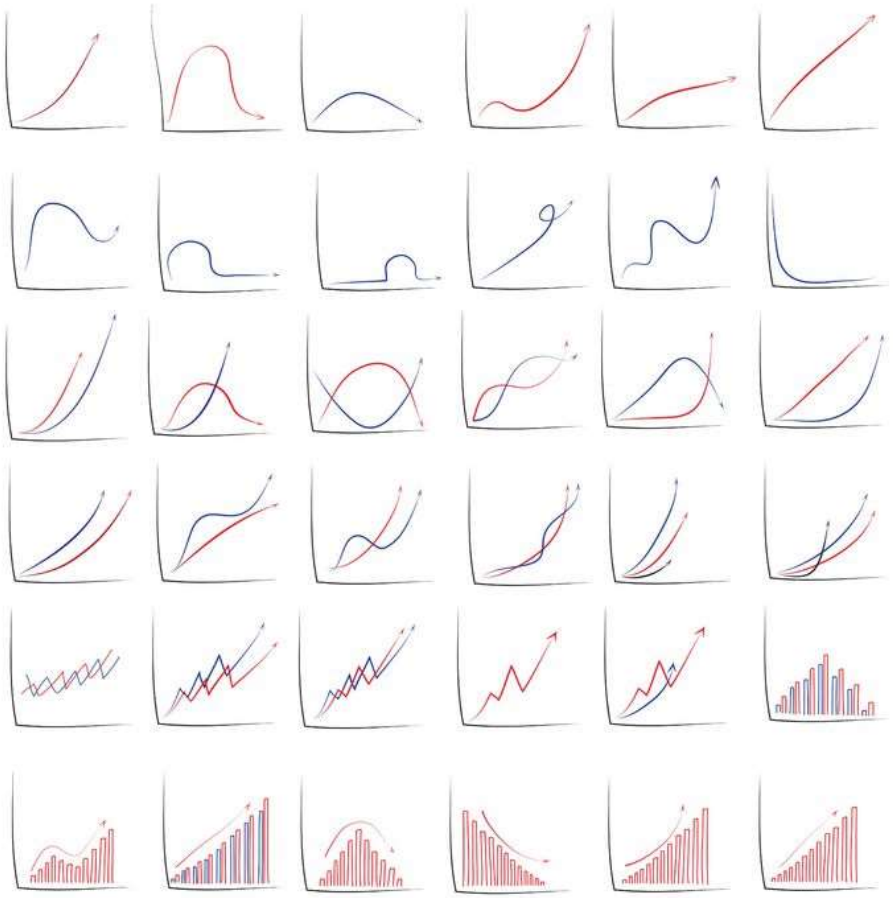
❸ Parse the Data



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

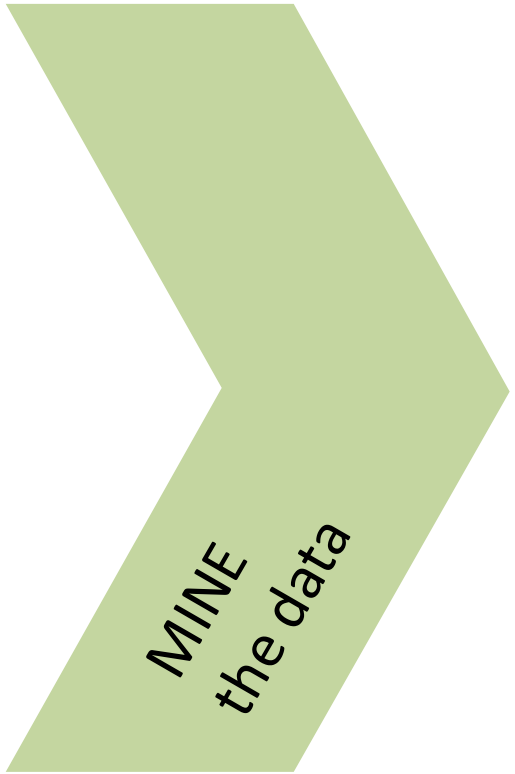
❸ Parse the Data

Exploratory Data Analysis



Napat Polchoke © 123RF.com

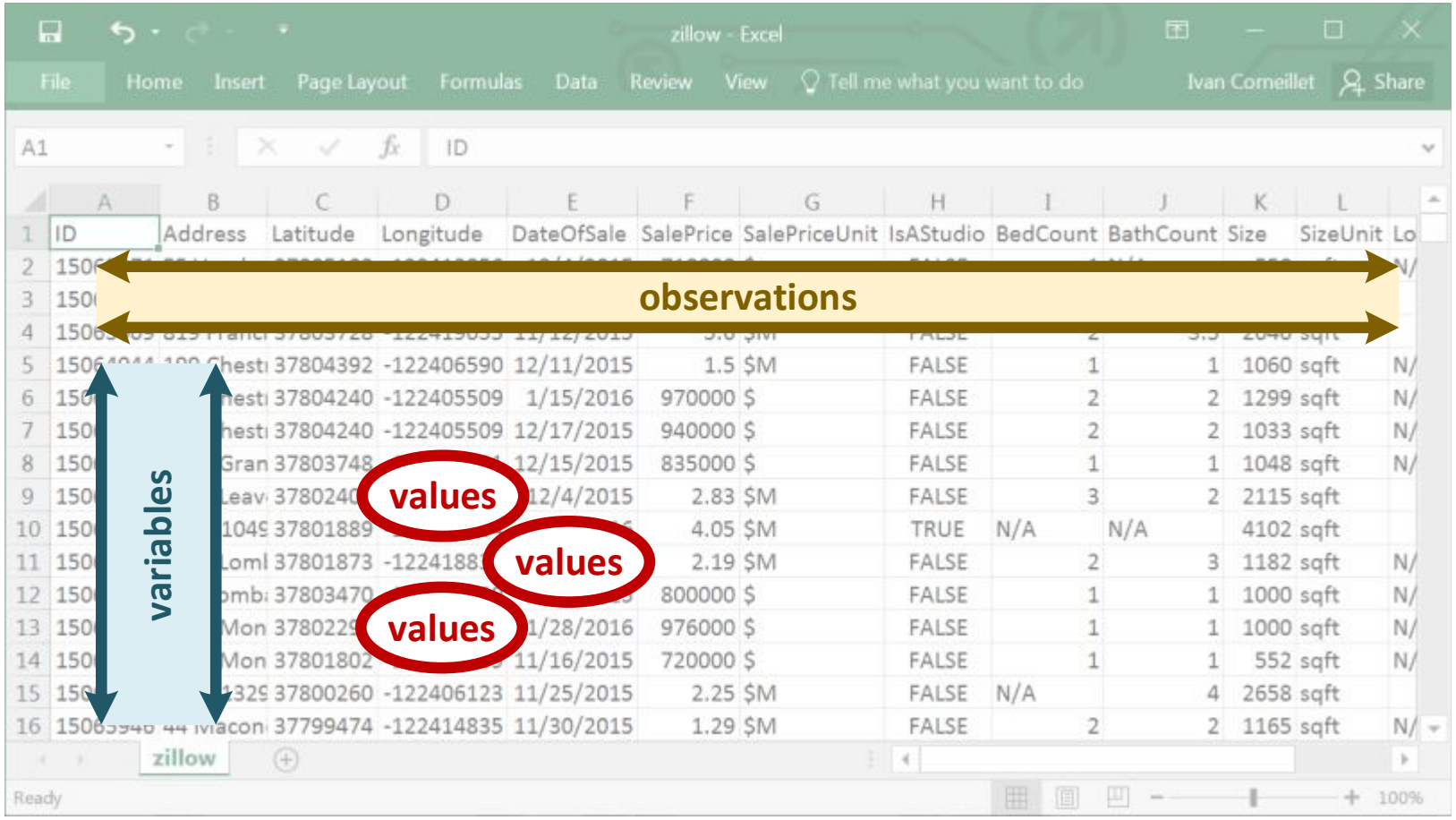
❹ Mine the Data



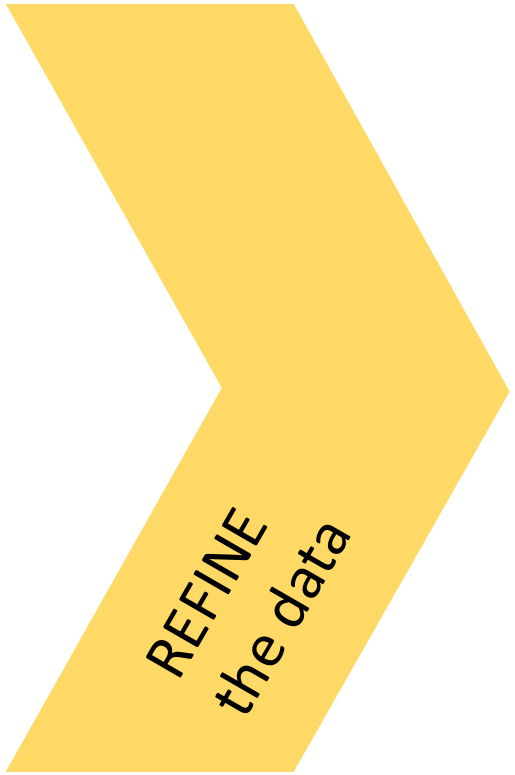
- Mine the Data
  - Determine sampling methodology and sample data
  - Format, clean, slice, and combine data in Python
  - Create necessary derived columns from the data (new data)

❹ Mine the Data

We will be tidying our data using the *pandas* library



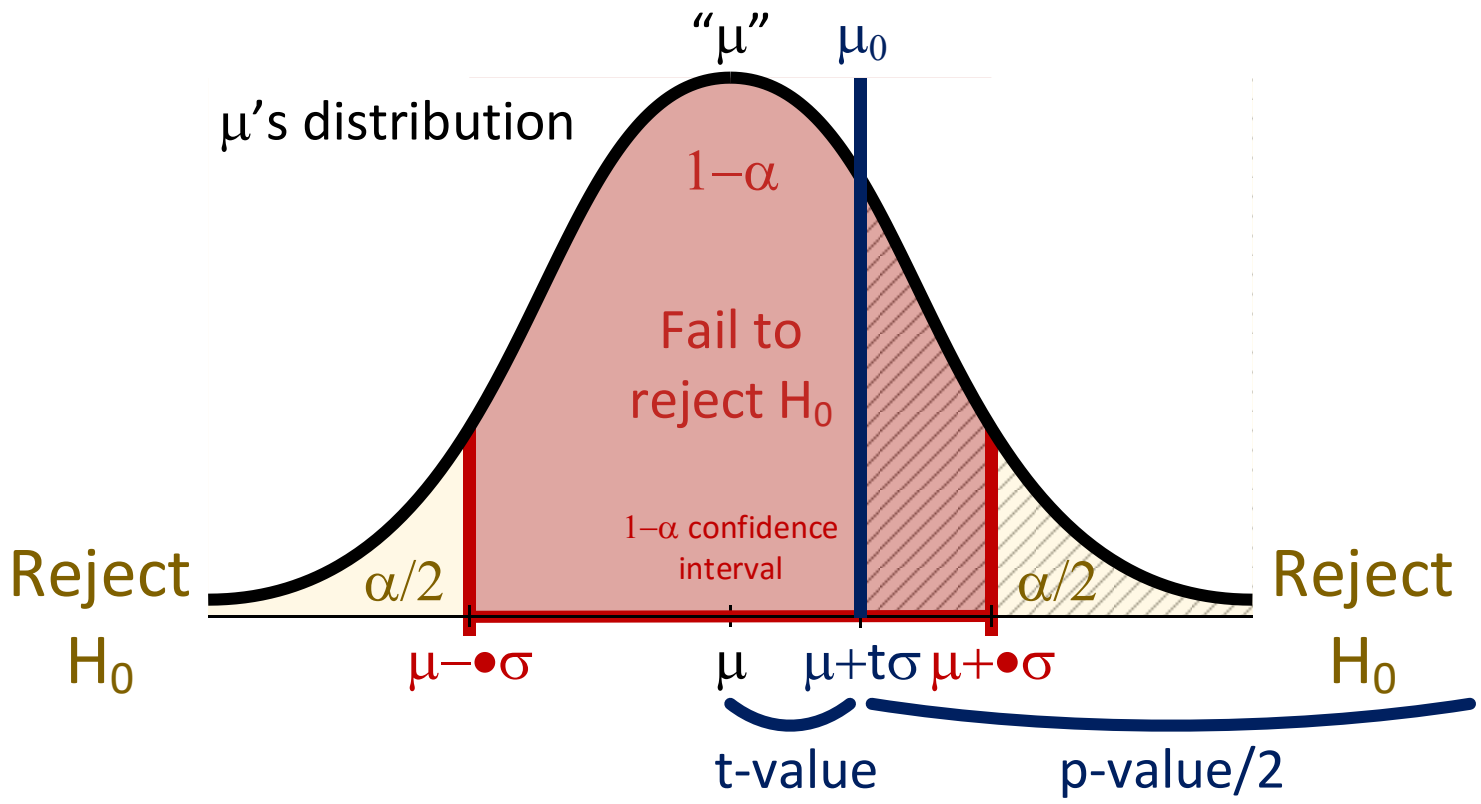
❺ Refine the Data



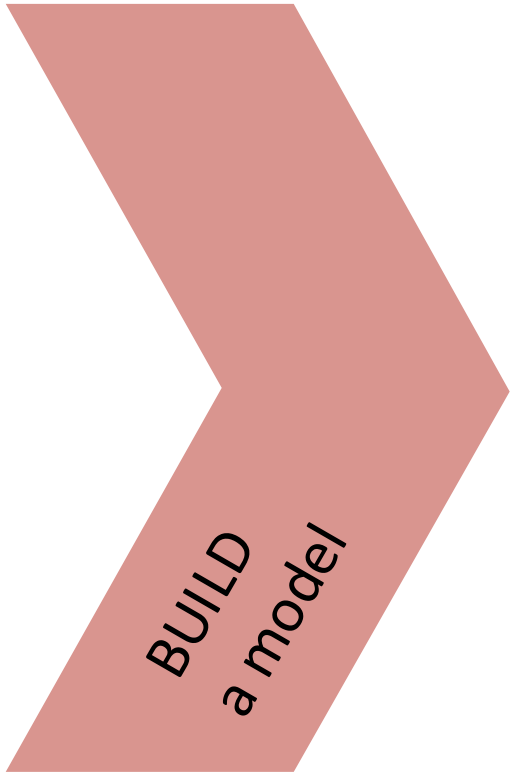
- Refine the Data
  - Identify trends and outliers
  - Apply descriptive and inferential statistics
  - Document and transform data

❺ Refine the Data

We will apply inferential statistics



❻ Build a Model



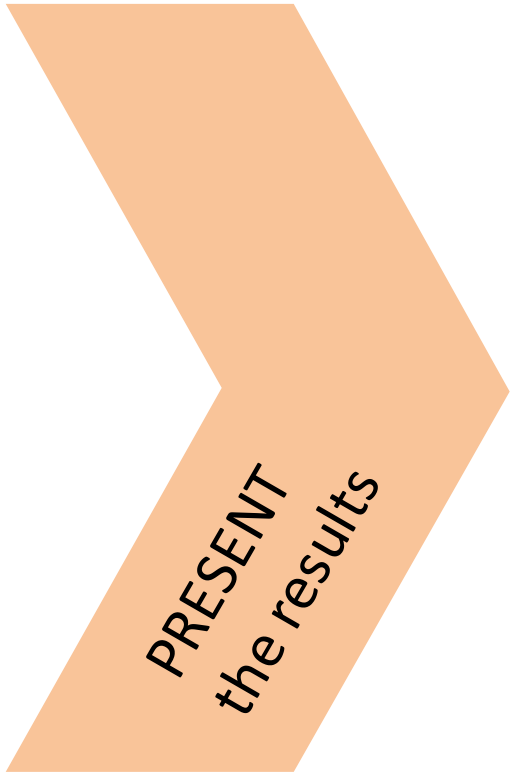
- Build a Model
  - Select appropriate model
  - Build model
  - Evaluate and refine model

❻ Build a Model

Types of machine learning algorithms we will study in this course (+ NLP)

	Continuous	Categorical
Supervised (a.k.a., predictive modeling)	Linear Regression k-Nearest Neighbors Decision Trees and Random Forests Time Series	Logistic Regression k-Nearest Neighbors Decision Trees and Random Forests
Unsupervised	*A machine learning model that doesn't use labeled data is called unsupervised. It extract structure from the data. Goal is "representation"*	

❼ Present the Results



- 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

❼ Present the Results

Know Your Audience

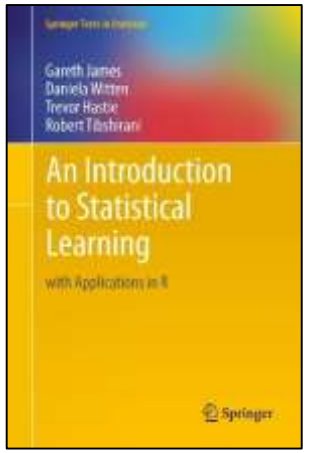
Corina Rosu © 123RF.com

A Note About Iteration

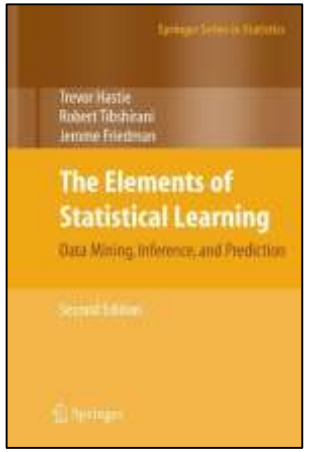
- Iteration is an important part of *every* step in the Data Science Workflow. At any given point in the process, you may find yourself repeating or going back and re-doing elements in order to better understand your data, clarify your model, and refine your presentation

- For example, after presenting your findings, you may want to:
  - Identify follow-up problems and questions for future analysis
  - Create a visually effective summary or report
  - Consider the needs of different stakeholders and how your report might be changed for them
  - Identify the limitations of your analysis
  - Identify relationships between visualizations

Some great resources to follow along the class (or afterwards) *(optional; not required for the course)*



- An Introduction to Statistical Learning: with Applications in R (by James et al.). The e-book is available free-of-charge here



- For a more advanced treatment of these topics, check out The Elements of Statistical Learning: Data Mining, Inference, and Prediction (by Hastie et al.). And yes, the e-book is also free… (here)

❶ IDENTIFY the Problem

❶ Identify the Problem

- Identify the Problem
  - Identify business/product objectives
  - Identify and hypothesize goals and criteria for success
  - Create a set of questions for identifying correct dataset

- The Why's and How's of a Good Question
- The SMART Goals Framework

❶ IDENTIFY the Problem

The Why's and How's of a Good Question

By asking a good question and setting a clear aim:

- You set yourself up for success
  - "A problem well stated is half solved" – Charles Kettering
- You help other data scientists learn from and reproduce your work
  - You establish the basis for making your analysis reproducible
- You also help them expand on your work in the future

❶ IDENTIFY the Problem

The SMART Goals Framework for Data Science

The SMART Framework for Data Science

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 you in 6 months!) can read your state and understand exactly how your analysis is performed
TIME-BOUND	You clearly state the time period and population for which this analysis will pertain

Trends often change over time and vary by the population of source of your data. It is important to clearly define who/what you included in your analysis as well as the time period for the analysis

❶ IDENTIFY the Problem

Activity | A SMART Goal for Your Final Project

Activity | A SMART Goal for Your Final Project

EXERCISE

DIRECTIONS (10 minutes)

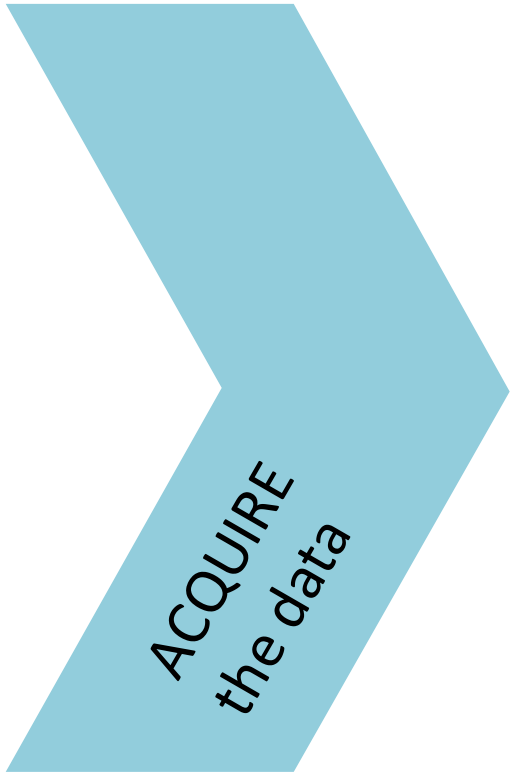
1. After the first class, you probably started brainstorming on an idea for your final project. If not, here's an opportunity!
2. If your idea is cool and interesting, that's great. But it is a SMART idea?
3. Assess your idea using the Data Science-tuned SMART Framework
    a. If you have just a couple of gaps, how can you close them?
    b. On the other end, if you have too many and closing these gaps would be difficult, you might want to consider something else
4. After 5 minutes, share your idea and gaps in pairs and offer advise to each other, again using the SMART Framework (2.5 minutes each)

DELIVERABLE

Answers to the above questions

❷ ACQUIRE the Data

❷ Acquire the Data



- Acquire the Data
  - Identify the “right” dataset(s)
  - Import data and set up local or remote data structure
  - Determine most appropriate tools to work with data

❷ Acquire the Data (cont.)

- 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 calibrated?
  - Is the dataset aggregated?
  - Do we need pre-aggregated data?

- Data Types
- Logistics of Acquiring Data
  - The SF housing Dataset
- Tidying Up Data

❷ ACQUIRE the Data

Data Types

Why Data Types Matter

- Different data types have different limitations and strengths, e.g., certain types of analyses aren't possible with certain data types
- There are 2 types of data which we may might use for analysis:
  - Cross-sectional data
    - Collect observations of many samples at the same point of time, or without regard to differences in time
  - Longitudinal data (i.e., time series)
    - Tracks the same sample (e.g., individual, household, or establishment) at different points in time

Data Types | Pros and Cons

	Pros	Cons
Cross-sectional data	❑ Often population-based ❑ Generalizable ❑ Less expensive compared to other types of data collection methods	❑ Separation of cause and effect may be difficult (or impossible) ❑ Variables/cases with long duration are over-represented
Longitudinal data (time series)	❑ Unambiguous temporal sequence; exposure precedes outcome ❑ Multiple outcomes can be measured	❑ Takes a long time to collect data ❑ Vulnerable to missing data ❑ More expense compared to other types of data collection methods

❷ ACQUIRE the Data

Activity | Knowledge Check

Activity | Knowledge Check

EXERCISE

DIRECTIONS (10 minutes)

1. What type of data is shown by Zillow? (http://www.zillow.com/san-francisco-ca/sold/)
2. Can you create a cross-sectional analysis from a longitudinal data collection? How? Is this applicable from the data above?
3. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

❷ ACQUIRE the Data

Logistics of Acquiring Data

Logistics of Acquiring Data

- Data can be acquired through a variety of sources
  - Web (e.g., Google Analytics, HTML)
  - Databases
    - SQL (Structured Query Language)
    - NoSQL (“Not only SQL”)
- File Formats
  - CSV (Comma-Separated Values)
  - TSV/TXT (Tab-Separated Values)
  - JSON (JavaScript Object Notation)
  - XML (eXtensible Markup Language)

❷ ACQUIRE the Data

Tidying Up Data

Raw data is Messy™…

EXAMPLE

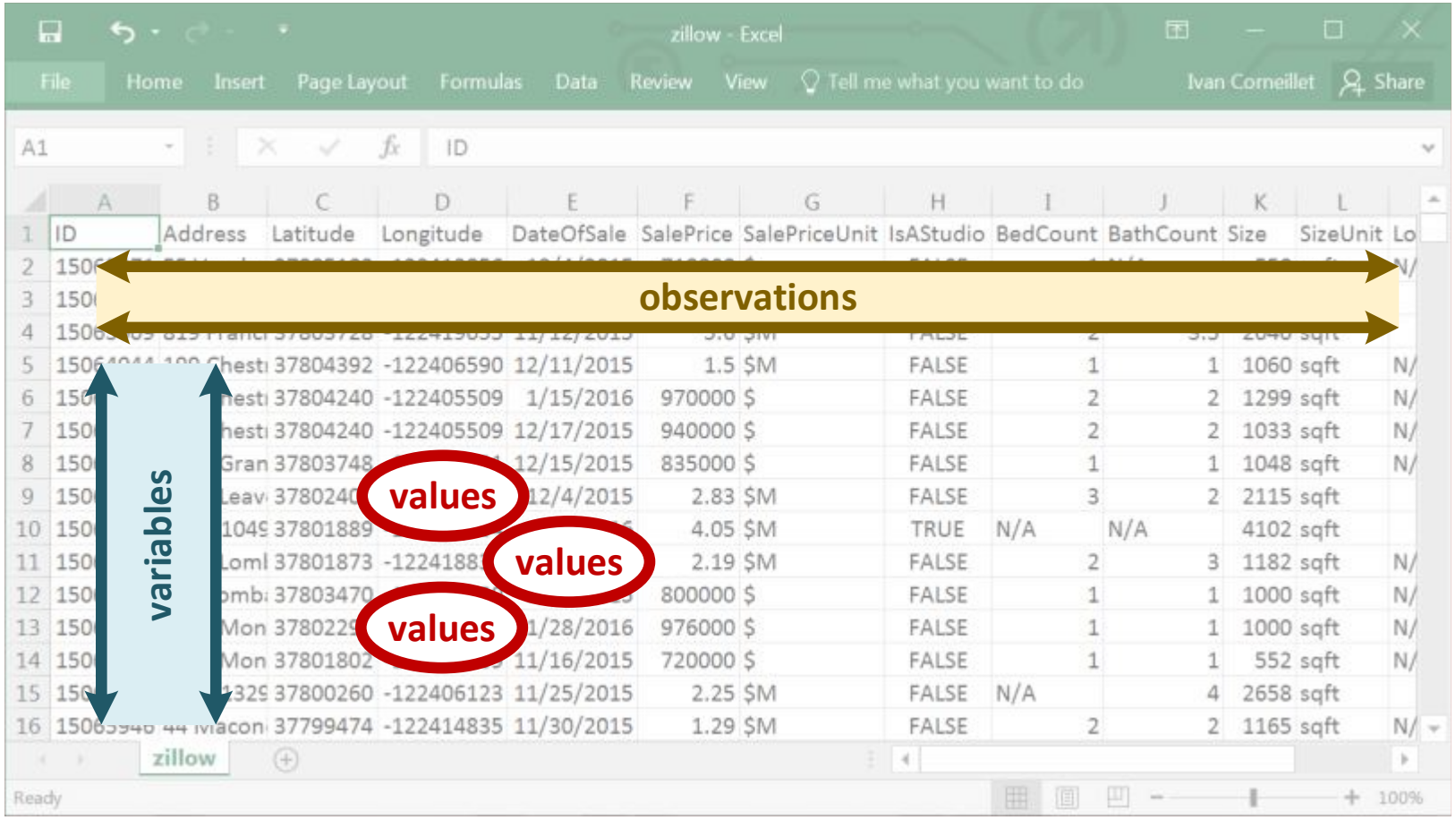
- Trouble tickets inspect and maintain manholes in New Year City
- "Service box," a common piece of infrastructure, had at least 38 variants, including `SB`, `S`, `S/B`, `S.B`, `S?B`, `S.B.`, `SBX`, `S/BX`, `SB/X`, `S/XB`, `/SBX`, `S.BX`, `S &BX`, `S?BX`, `S BX`, `S/B/X`, `S BOX`, `SVBX`, `SERV BX`, `SERV-BOX`, `SERV/BOX`, and `SERVICE BOX`

(Source: Big Data: A Revolution That Will Transform How We Live, Work, and Think)

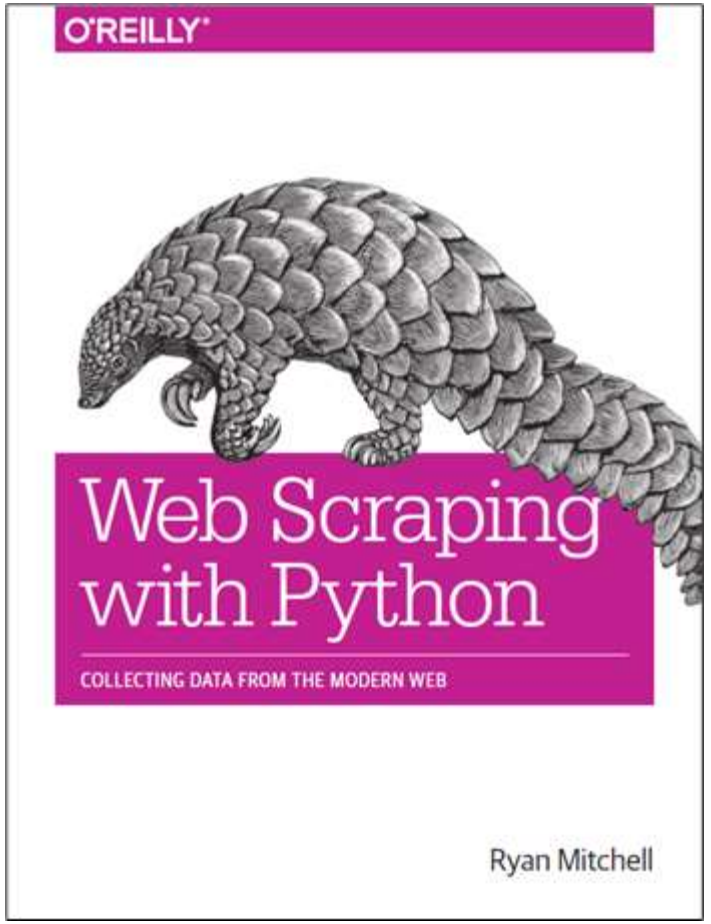
Tidying Up Data

- Tidying up data is the most fruitful skill you can learn as a data scientist
  - It will save you hours of time and make your data much easier to visualize, manipulate, and model
- Many data science tools follow a set of conventions that makes one layout of tabular data much easier to work with than others. Your data will be easier to work with if you follow three rules:
  - Each observation is placed in its own row
  - Each variable in the dataset is placed in its own column
  - Each value is placed in its own cell

Tidying Up Data (cont.)

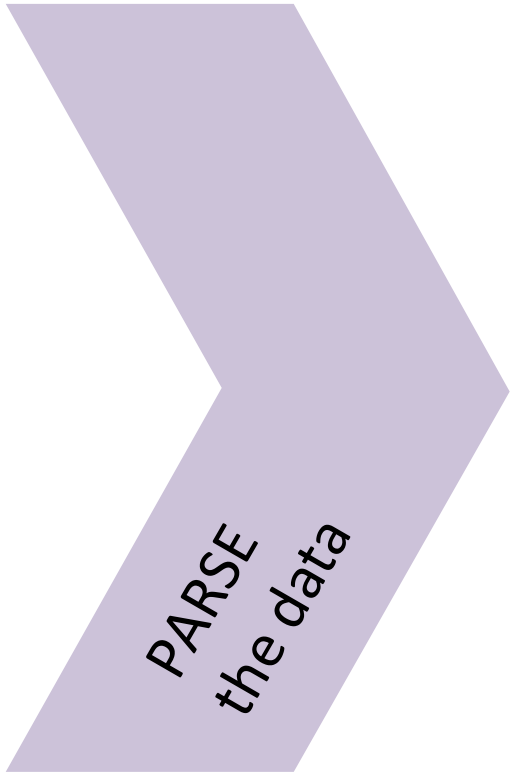


A good resource to get started with web scraping using Python *(optional; not required for the course)*



❸ PARSE the Data

❸ Parse the Data



- Parse the Data
  - Read any documentation provided with the data (session 2)
  - Perform exploratory data analysis (session 3)
  - Verify the quality of the data (sessions 2/3)

❷ Acquire the Data (cont.)

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

- Documentation and Data Dictionary
- Introduction to *pandas* + codealong
- Codealong: Tidying up (more) the SF housing dataset
- Lab

❸ PARSE the Data

Documentation and Data Dictionary

Documentation and Data Dictionary

- Data dictionaries
  - Help you judge the quality of the data
  - Also help understand how it's coded
    - Does "`gender = 1`" mean female or male?
    - Is the currency dollars or euros?
  - Help identify any requirements, assumptions, and constraints of the data
  - Make it easier to share data

Kaggle's Titanic Data Dictionary

EXAMPLE

```
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 Fiancés 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.
```

❸ PARSE the Data

Introduction to pandas

pandas is a Python library to manipulate and perform statistical and mathematical analysis on tabular and multidimensional datasets

- *pandas* provides the ability to index, retrieve, tidy, reshape, combine, slice, and perform various analyses on both single and multidimensional data
- It also includes loading and saving data from local and Internet-based resources
- We will use *pandas* to explore and manipulate the SH housing dataset

❸ PARSE the Data

Codealongs

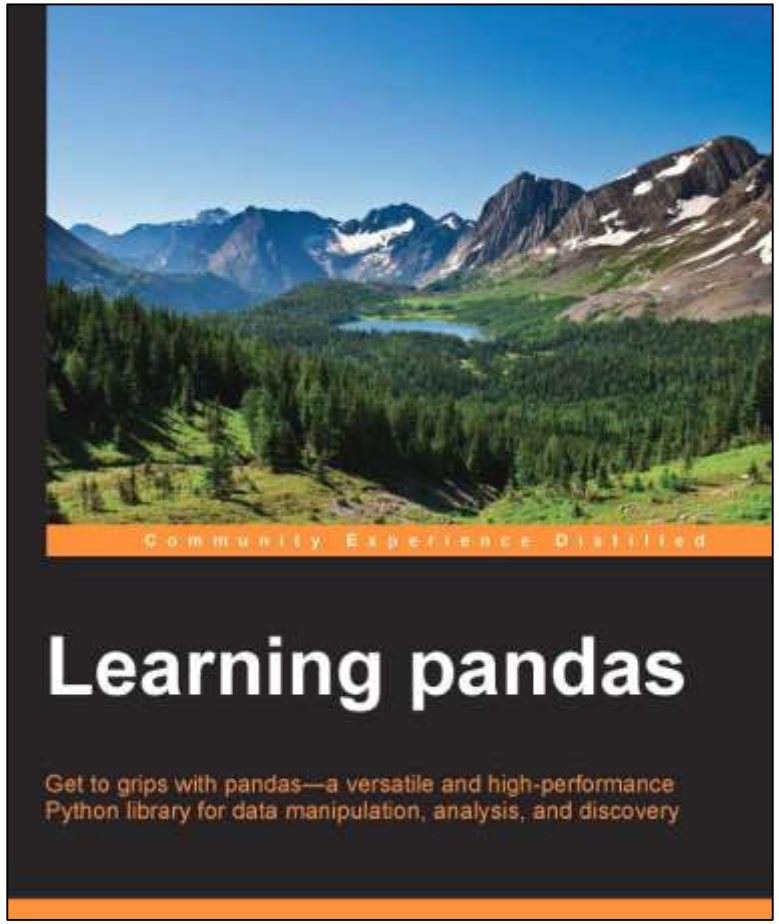
Part A – Introduction to pandas with the SF housing dataset

Part B – Use pandas to tidy up the SF housing dataset

The codealong was just the tip of the iceberg. There is much more. Check out the following:

- *pandas* documentation (which is very well written…)
  - http://pandas.pydata.org/pandas-docs/stable/

As well as a good book *(again optional; not required for the course)*



Unit Project 1

Lab

Introduction to pandas

Review

Review

You should now be able to:

- Setup and manage your personal GitHub repository for submitting assignments
- Define a problem and types of data
- Identify dataset types
- Apply the data science workflow in the *pandas* context
- Write an iPython notebook to import, format, and clean data using the *pandas* library

Q & A

Next Class

Descriptive Statistics for Exploratory Data Analysis

Learning Objectives

After the next lesson, you should be able to:

- Identify variable types
- Use the *pandas* and *NumPy* libraries to analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile range, variance, standard deviation, and correlation
- Create data visualizations – including boxplots, histograms, and scatter plots – to discern characteristics and trends in a dataset

Exit Ticket

Don’t forget to fill out your exit ticket here