GENERAL ASSEMBLY

INTRODUCTION TO REGRESSION ANALYSIS

Abbas Chokor, Ph.D.

Staff Data Scientist, Seagate Technology

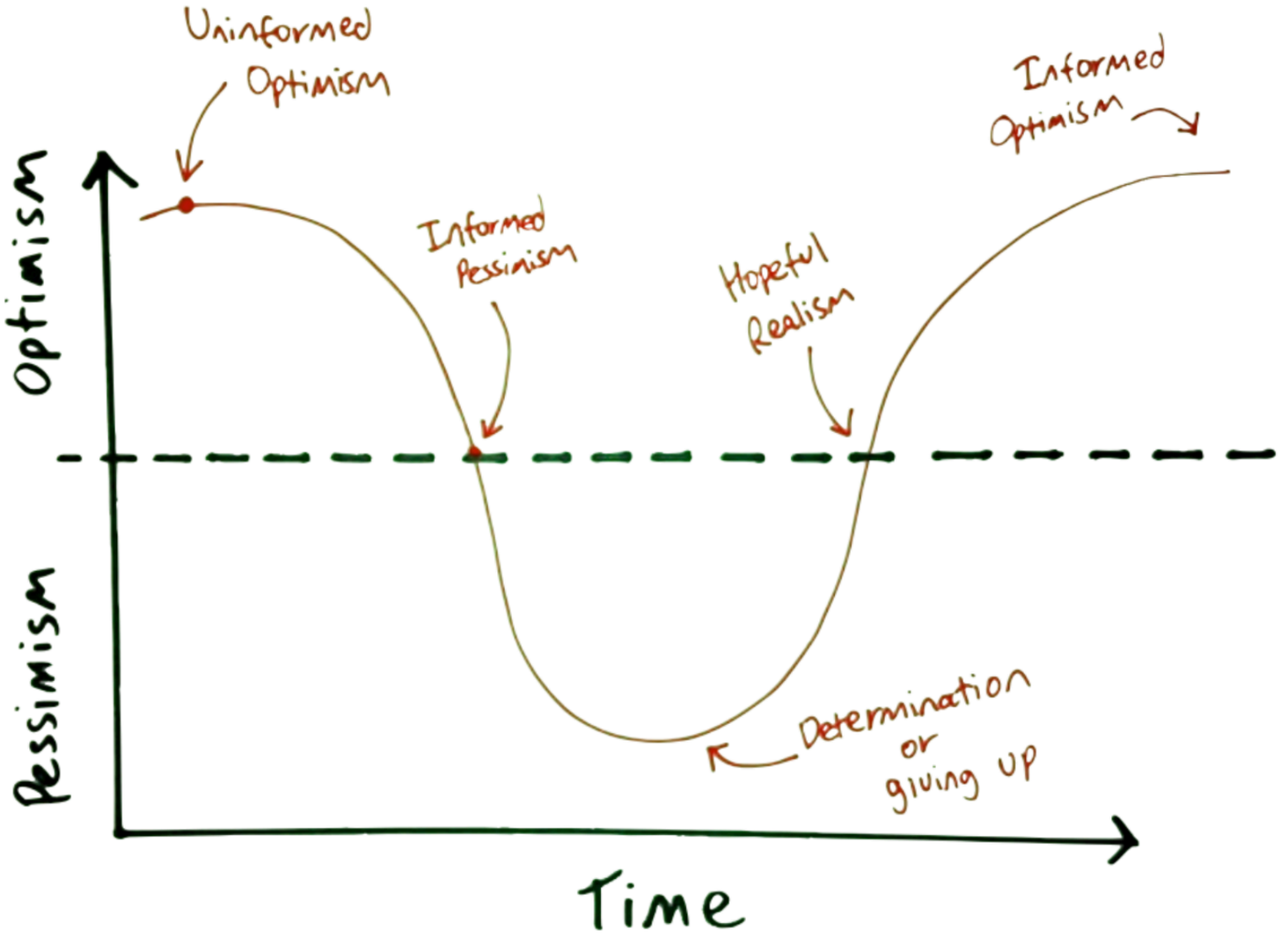
OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	~~What is Data Science~~	~~Lesson 1~~
	~~Research Design and Pandas~~	~~Lesson 2~~
	~~Statistics Fundamentals I~~	~~Lesson 3~~
	~~Statistics Fundamentals II~~	~~Lesson 4~~
	~~Flexible Class Session~~	~~Lesson 5~~
UNIT 2: FOUNDATIONS OF DATA MODELING	Introduction to Regression	Lesson 6
	Evaluating Model Fit	Lesson 7
	Introduction to Classification	Lesson 8
	Introduction to Logistic Regression	Lesson 9
	Communicating Logistic Regression Results	Lesson 10
	Flexible Class Session	Lesson 11
UNIT 3: DATA SCIENCE IN THE REAL WORLD	Decision Trees and Random Forests	Lesson 12
	Natural Language Processing	Lesson 13
	Dimensionality Reduction	Lesson 14
	Time Series Data I	Lesson 15
	Time Series Data II	Lesson 16
	Database Technologies	Lesson 17
	Where to Go Next	Lesson 18
	Flexible Class Session	Lesson 19
	Final Project Presentations	Lesson 20

Any questions about Unit 1 review?
Let's talk in the office hours

Today's Class

WHERE ARE YOU NOW?



Let's get to real work!

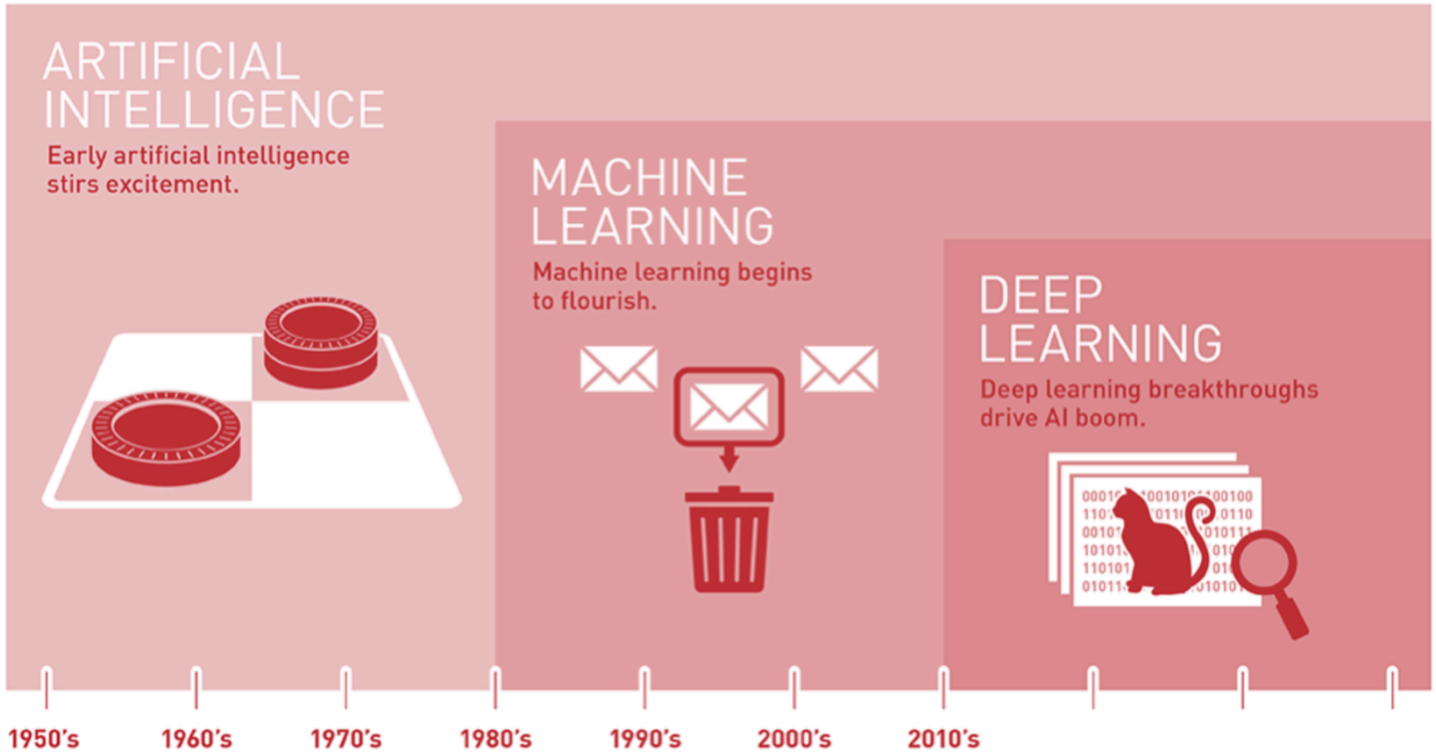
LEARNING OBJECTIVES

- Overview of machine learning modeling
- Define data modeling and simple linear regression
- Build a multivariate linear regression model using a dataset that meets the linearity assumption using the scikit-learn library
- Understand and identify multicollinearity in a multiple regression.

WHAT IS MACHINE LEARNING

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

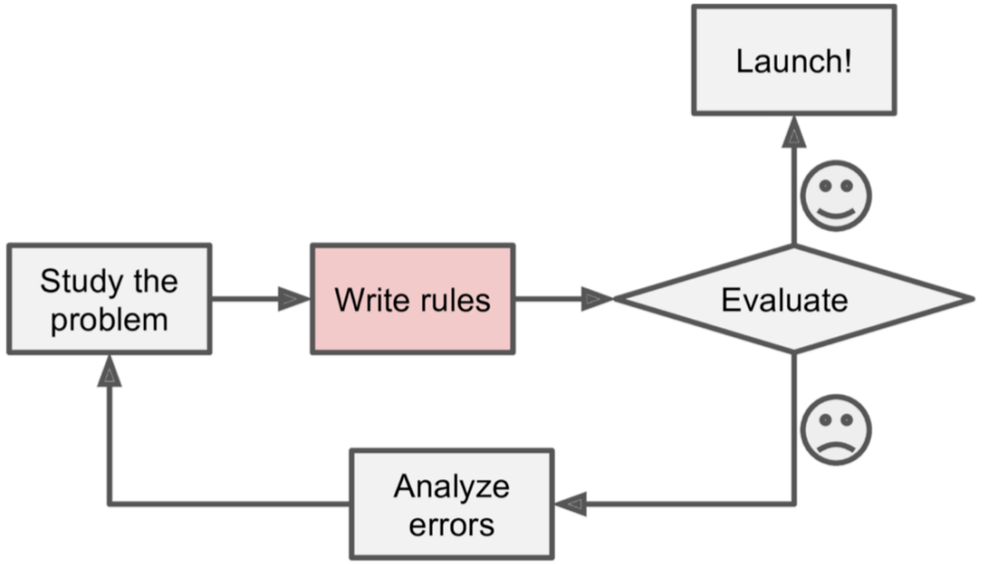
Arthur Samuel, *1959*

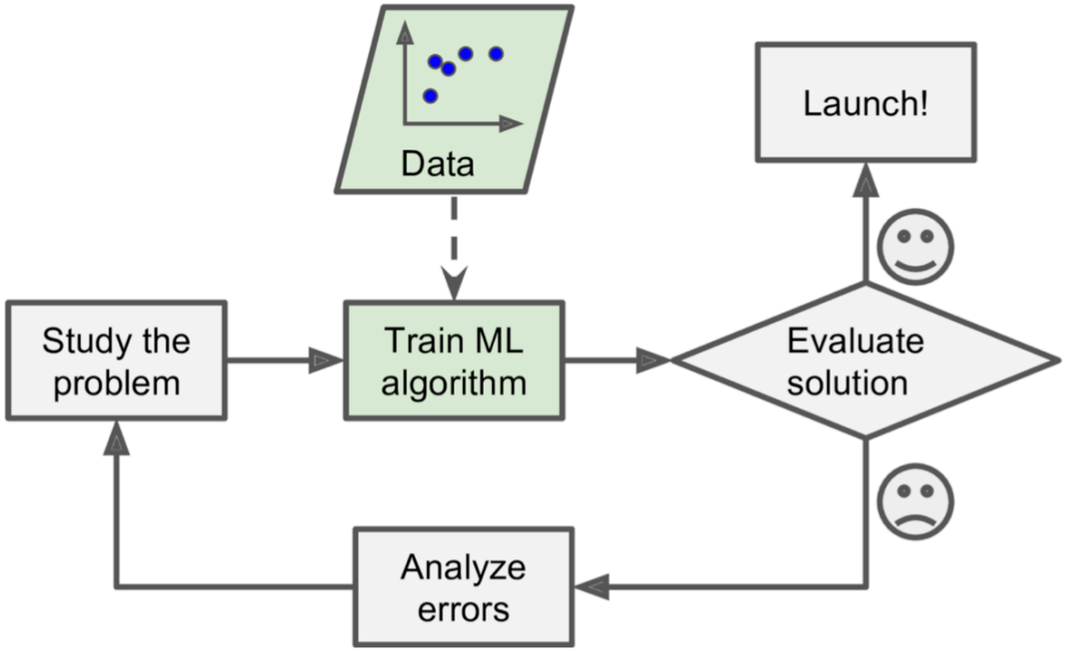


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

<u>Machine Learning</u>: is the science (and art) of programming computers so they can *learn from data*.

WHY TO USE MACHINE LEARNING?



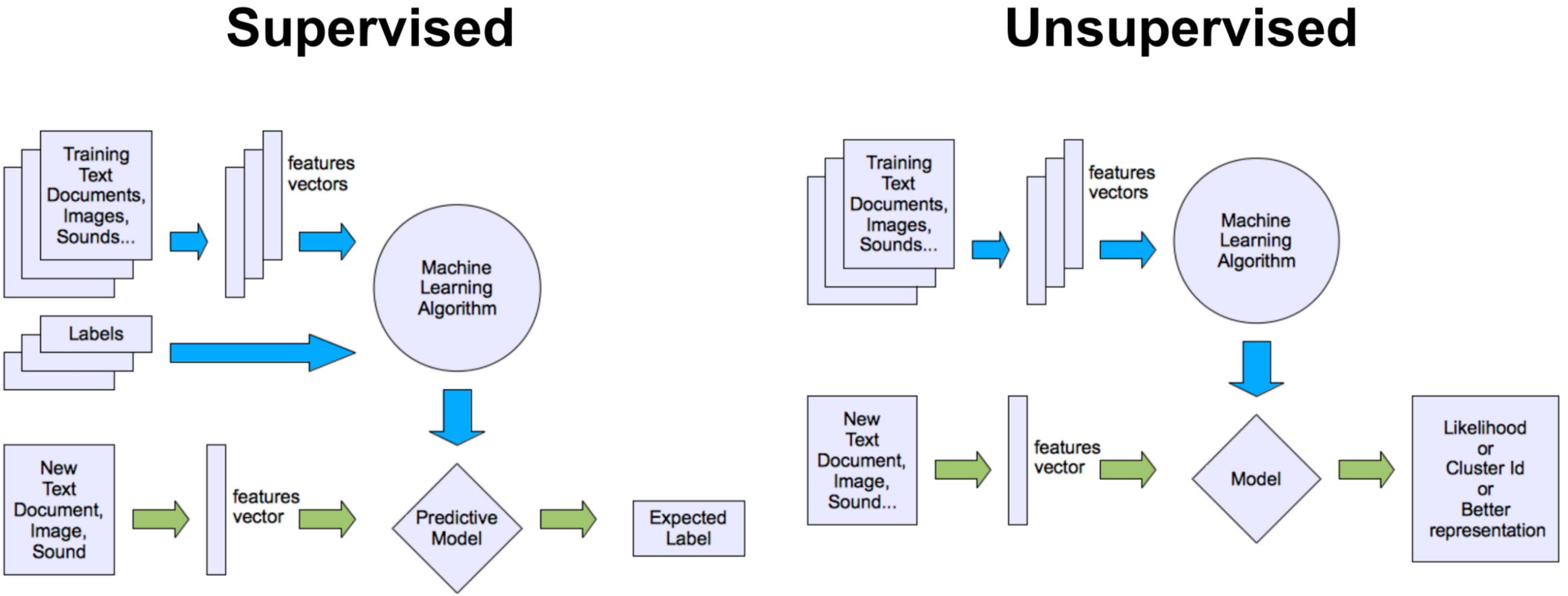


Traditional approach **Vs.** ***Machine Learning approach***

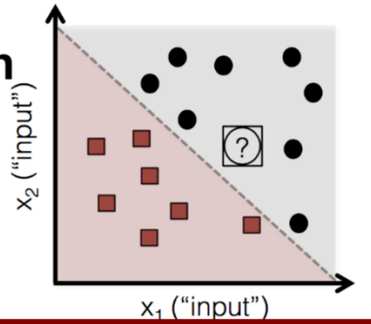
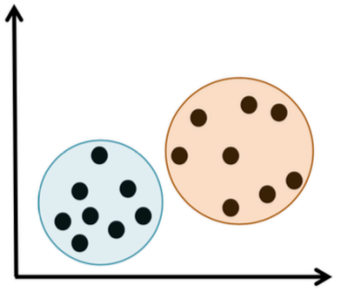
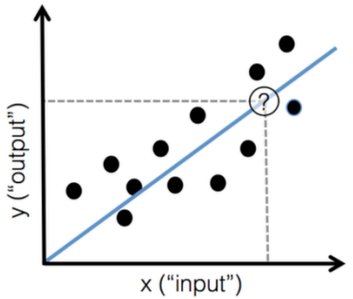
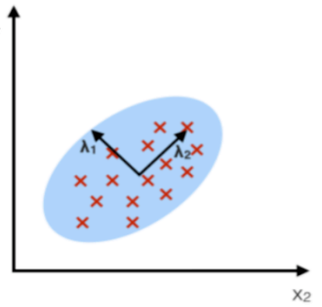
Did you know?

Machine learning algorithms are expected to replace 25% of the jobs across the world, in the next 10 years.

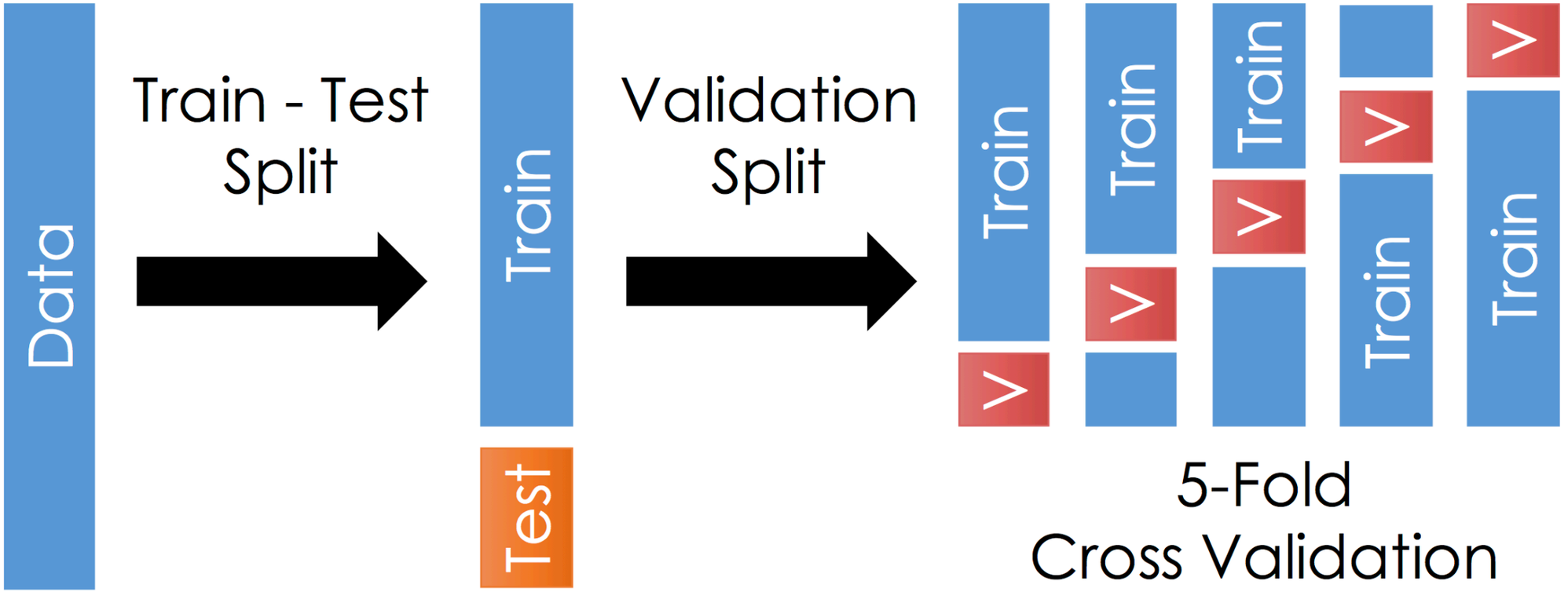
TYPES OF MACHINE LEARNING



TYPES OF MACHINE LEARNING

	Supervised Working with Labeled Data	**Unsupervised** Working with Unlabeled Data
Discrete Countable Data	**Classification** 	**Clustering** 
Continuous Infinite Data	**Regression** 	**Dimensionality Reduction** 

TRAINING – VALIDATING - TESTING



LEARNING OBJECTIVES

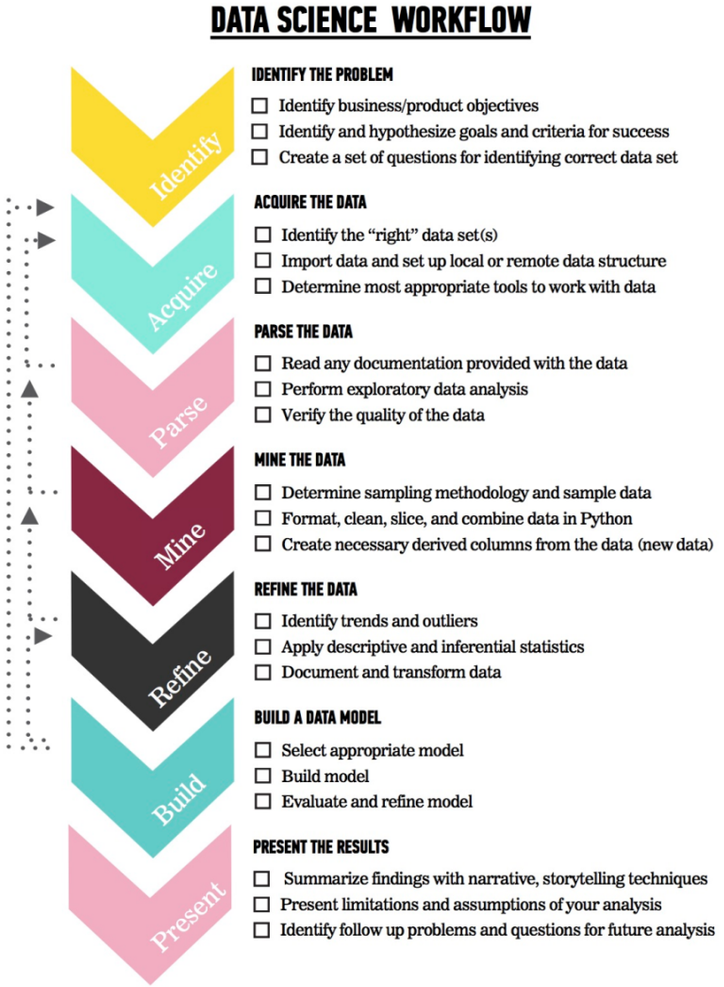
- ~~Overview of machine learning modeling~~
- Define data modeling and simple linear regression
- Build a linear regression model using a dataset that meets the linearity assumption using the scikit-learn library
- Understand and identify multicollinearity in a multiple regression.

OPENING

INTRODUCTION TO REGRESSION ANALYSIS

WHERE ARE WE IN THE DATA SCIENCE WORKFLOW?

- Data has been **acquired** and **parsed.**
- Today we'll **refine** the data and **build** models.
- We'll also use plots to **represent** the results.



INTRODUCTION

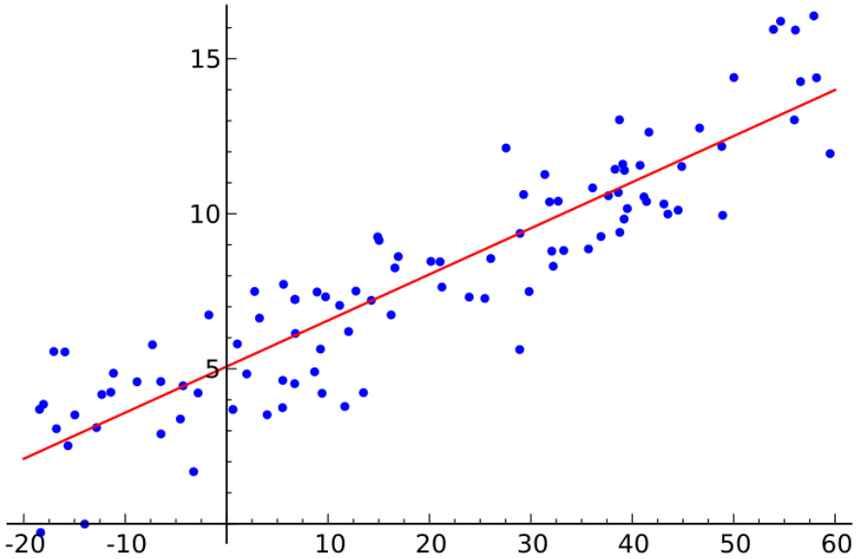
SIMPLE LINEAR REGRESSION

SIMPLE LINEAR REGRESSION

- Def: Explanation of a continuous variable given a series of independent variables

- The simplest version is just a line of best fit:
  $y = mx + b$

- Explain the relationship between **x** and **y** using the starting point **b** and the power in explanation **m**.



SIMPLE LINEAR REGRESSION

- However, linear regression uses linear algebra to explain the relationship between *multiple* x's and y.

- The more sophisticated version: y = beta * X + alpha (+ error)

- Explain the relationship between the matrix **X** and a dependent vector **y** using a y-intercept **alpha** and the relative coefficients **beta**.

SIMPLE LINEAR REGRESSION

- Linear regression works **best** when:
  - The data is normally distributed (but doesn't have to be)
  - X's significantly explain y (have low p-values)
  - X's are independent of each other (low multicollinearity)
  - Resulting values pass linear assumption (depends upon problem)
- If data is not normally distributed, we could introduce *bias*.

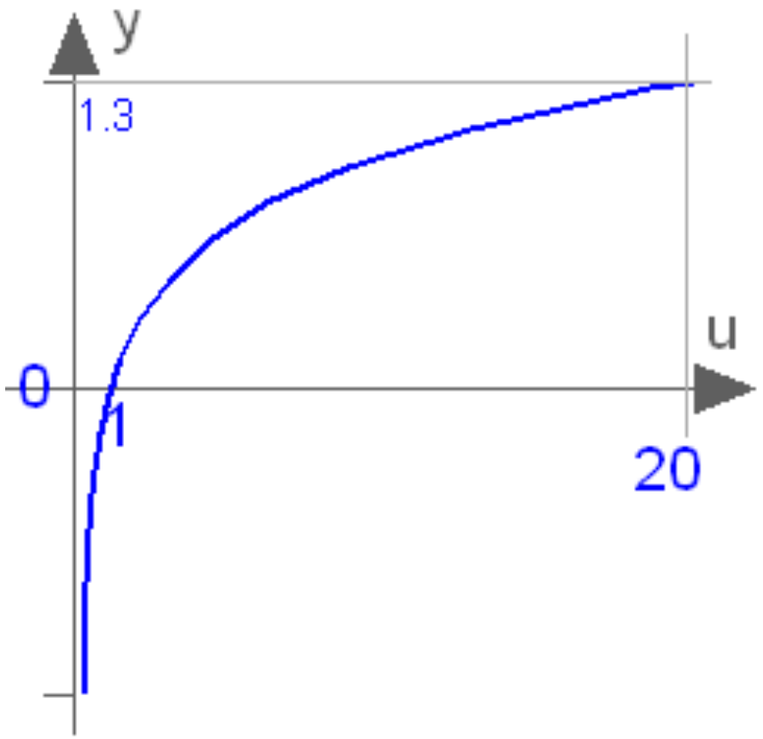
DEMO

REGRESSING AND NORMAL DISTRIBUTIONS

DEMO: REGRESSING AND NORMAL DISTRIBUTIONS

- Follow along with Part 1 of your starter code notebook while I walk through these examples.
- The first plot shows a relationship between two values, though not a linear solution.
- Note that lmplot() returns a straight line plot.
- However, we can transform the data, both log-log distributions to get a linear solution.

DEMO: REGRESSING AND NORMAL DISTRIBUTIONS



log function

GUIDED PRACTICE

USING SEABORN TO GENERATE SIMPLE LINEAR MODEL PLOTS

ACTIVITY: GENERATE SINGLE VARIABLE LINEAR MODEL PLOTS

DIRECTIONS (15 minutes)

1. Update and complete the code in Part 2 of the starter notebook to use **lmplot** and display correlations between body weight and two dependent variables: **sleep_rem** and **awake**.

DELIVERABLE

Two plots

INTRODUCTION

SIMPLE REGRESSION ANALYSIS IN SKLEARN

SIMPLE LINEAR REGRESSION ANALYSIS IN SKLEARN

- You can use the following principles:
  - All sklearn modeling classes are based on the base estimator. This means all models take a similar form.
  - All estimators take a matrix **X**, either sparse or dense.
  - Supervised estimators also take a vector **y** (the response).
  - Estimators can be customized through setting the appropriate parameters.

SIMPLE LINEAR REGRESSION ANALYSIS IN SKLEARN

- General format for sklearn model classes and methods

```
# generate an instance of an estimator class
estimator = base_models.AnySKLearnObject()
# fit your data
estimator.fit(X, y)
# score it with the default scoring method (recommended to use the metrics module in the future)
estimator.score(X, y)
# predict a new set of data
estimator.predict(new_X)
# transform a new X if changes were made to the original X while fitting
estimator.transform(new_X)
```

- With this information, we can build a simple process for linear regression.

DEMO

SIGNIFICANCE IS KEY

DEMO: SIGNIFICANCE IS KEY

- Follow along with Part 3 of your starter code notebook while I walk through these examples.
- What does the residual plot tell us?
- How can we use the linear assumption?

GUIDED PRACTICE

USING THE LINEAR REGRESSION OBJECT

ACTIVITY: USING THE LINEAR REGRESSION OBJECT

DIRECTIONS (15 minutes)

1. With a partner, generate in Part 4 of your starter code two more models using the log-transformed data to see how this transform changes the model's performance.
2. Use the code on the following slide to complete #1.

DELIVERABLE

Two new models

ACTIVITY: USING THE LINEAR REGRESSION OBJECT

DIRECTIONS (15 minutes)

```
X =
y =
loop = []
for boolean in loop:
    print 'y-intercept:', boolean
    lm =
linear_model.LinearRegression(fit_intercept=boolean)
    get_linear_model_metrics(X, y, lm)
    print
```

DELIVERABLE

Two new models

INDEPENDENT PRACTICE

BASE LINEAR REGRESSION CLASSES

ACTIVITY: BASE LINEAR REGRESSION CLASSES

EXERCISE

DIRECTIONS (20 minutes)

1. In Part 5 of your starter code, experiment with the model evaluation function we have (**get_linear_model_metrics**) with the following sklearn estimator classes.

    a. linear_model.Lasso()
    b. linear_model.Ridge()
    c. linear_model.ElasticNet()

Note: We'll cover these new regression techniques in a later class.

DELIVERABLE

New models and evaluation metrics

LEARNING OBJECTIVES

- ~~Overview of machine learning modeling~~
- ~~Define data modeling and simple linear regression~~
- Build a linear regression model using a dataset that meets the linearity assumption using the scikit-learn library
- Understand and identify multicollinearity in a multiple regression.

INTRODUCTION

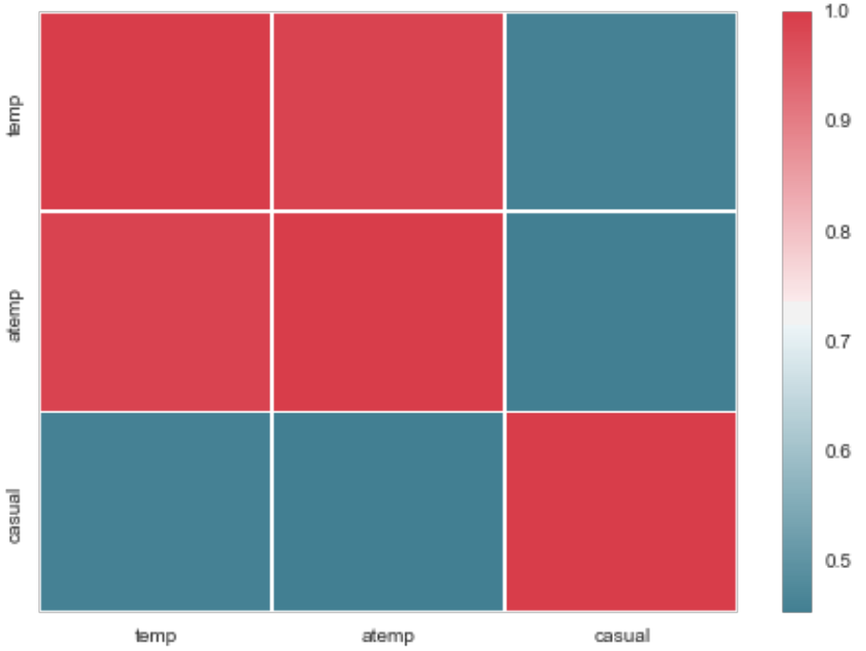
MULTIPLE REGRESSION ANALYSIS

MULTIPLE REGRESSION ANALYSIS

- Simple linear regression with one variable can explain some variance, but using multiple variables can be much more powerful.
- We want our multiple variables to be mostly independent to avoid multicollinearity.
- Multicollinearity, when two or more variables in a regression are highly correlated, can cause problems with the model.

BIKE DATA EXAMPLE

- We can look at a correlation matrix of our bike data.
- Even if adding correlated variables to the model improves overall variance, it can introduce problems when explaining the output of your model.
- What happens if we use a second variable that isn't highly correlated with temperature?



LEARNING OBJECTIVES

- ~~Overview of machine learning modeling~~
- ~~Define data modeling and simple linear regression~~
- ~~Build a linear regression model using a dataset that meets the linearity assumption using the scikit-learn library~~
- Understand and identify multicollinearity in a multiple regression.

GUIDED PRACTICE

MULTICOLLINEARITY WITH DUMMY VARIABLES

ACTIVITY: MULTICOLLINEARITY WITH DUMMY VARIABLES

DIRECTIONS (15 minutes)

1. Load the bike data in Part 7 of our starter code.
2. Run through the code on the following slide.
3. What happens to the coefficients when you include all weather situations instead of just including all except one?

DELIVERABLE

Two models' output

ACTIVITY: MULTICOLLINEARITY WITH DUMMY VARIABLES

DIRECTIONS (15 minutes)

```
lm = linear_model.LinearRegression()
weather = pd.get_dummies(bike_data.weathersit)
get_linear_model_metrics(weather[[1, 2, 3, 4]], y, lm)
print
# drop the least significant, weather situation  = 4
get_linear_model_metrics(weather[[1, 2, 3]], y, lm)
```

DELIVERABLE

Two models' output

GUIDED PRACTICE

COMBINING FEATURES INTO A BETTER MODEL

ACTIVITY: COMBINING FEATURES INTO A BETTER MODEL

DIRECTIONS (15 minutes)

1. With a partner, complete the code on the following slide in Part 8.
2. Visualize the correlations of all the numerical features built into the dataset.
3. Add the three significant weather situations into our current model.
4. Find two more features that are not correlated with the current features, but could be strong indicators for predicting guest riders.

DELIVERABLE

Visualization of correlations, new models

ACTIVITY: COMBINING FEATURES INTO A BETTER MODEL

DIRECTIONS (15 minutes)

```
lm = linear_model.LinearRegression()
bikemodel_data = bike_data.join() # add in the three weather situations

cmap = sns.diverging_palette(220, 10, as_cmap=True)
correlations = # what are we getting the correlations of?
print correlations
print sns.heatmap(correlations, cmap=cmap)

columns_to_keep = [] #[which_variables?]
final_feature_set = bikemodel_data[columns_to_keep]

get_linear_model_metrics(final_feature_set, y, lm)
```

DELIVERABLE

Visualization of correlations, new models

INDEPENDENT PRACTICE

BUILDING MODELS FOR OTHER Y VARIABLES

ACTIVITY: BUILDING MODELS FOR OTHER Y VARIABLES

DIRECTIONS (25 minutes)

1. Build a new model in Part 9 using a new y variable: registered riders.
2. Pay attention to the following:
   a. the distribution of riders (should we rescale the data?)
   b. checking correlations between the variables and y variable
   c. choosing features to avoid multicollinearity
   d. model complexity vs. explanation of variance
   e. the linear assumption

BONUS

1. Which variables make sense to dummy?
2. What features might explain ridership but aren't included? Can you build these features with the included data and pandas?

DELIVERABLE

A new model and evaluation metrics

LEARNING OBJECTIVES

- ~~Overview of machine learning modeling~~
- ~~Define data modeling and simple linear regression~~
- ~~Build a linear regression model using a dataset that meets the linearity assumption using the scikit-learn library~~
- ~~Understand and identify multicollinearity in a multiple regression.~~

CONCLUSION

TOPIC REVIEW

CONCLUSION

- You should now be able to answer the following questions:
  - What is simple linear regression?
  - What makes multi-variable regressions more useful?
  - What challenges do they introduce?
  - How do you dummy a category variable?

LESSON 7

UPCOMING WORK

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	~~What is Data Science~~	~~Lesson 1~~
	~~Research Design and Pandas~~	~~Lesson 2~~
	~~Statistics Fundamentals I~~	~~Lesson 3~~
	~~Statistics Fundamentals II~~	~~Lesson 4~~
	~~Flexible Class Session~~	~~Lesson 5~~
UNIT 2: FOUNDATIONS OF DATA MODELING	~~Introduction to Regression~~	~~Lesson 6~~
	Evaluating Model Fit	Lesson 7
	Introduction to Classification	Lesson 8
	Introduction to Logistic Regression	Lesson 9
	Communicating Logistic Regression Results	Lesson 10
	Flexible Class Session	Lesson 11
UNIT 3: DATA SCIENCE IN THE REAL WORLD	Decision Trees and Random Forests	Lesson 12
	Natural Language Processing	Lesson 13
	Dimensionality Reduction	Lesson 14
	Time Series Data I	Lesson 15
	Time Series Data II	Lesson 16
	Database Technologies	Lesson 17
	Where to Go Next	Lesson 18
	Flexible Class Session	Lesson 19
	Final Project Presentations	Lesson 20

UPCOMING WORK

Lesson 7

- Unit Project 2: due Thursday before class!
- Following Tuesday: Final Project 1

UPCOMING WORK: Unit Project 2

Project 2

In this project, you will implement the exploratory analysis plan developed in Project 1. This will lay the groundwork for our our first modeling exercise in Project 3.

Step 1: Load the python libraries you will need for this project

In [1]:
```
#imports
from __future__ import division
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
import pylab as pl
import numpy as np
%matplotlib inline
```

Step 2: Read in your data set

In [2]:
```
#Read in data from source
df_raw = pd.read_csv("../assets/admissions.csv")
print df_raw.head()
```

```
   admit  gre   gpa  prestige
0      0  380  3.61         3
1      1  660  3.67         3
2      1  800  4.00         1
3      1  640  3.19         4
4      0  520  2.93         4
```

Questions

Question 1. How many observations are in our dataset?

In [3]:
```
df_raw.count()
```

Out[3]:
```
admit       400
gre         398
gpa         398
prestige    399
dtype: int64
```

Answer:

Question 2. Create a summary table

In [ ]:
```
#function
```

In [ ]:
```

```

Question 3. Why would GRE have a larger STD than GPA?

Answer:

Question 4. Drop data points with missing data

In [ ]:
```

```

Question 5. Confirm that you dropped the correct data. How can you tell?

Answer:

Question 6. Create box plots for GRE and GPA

In [ ]:
```
#boxplot 1
```

In [ ]:
```
#boxplot 2
```

Question 7. What do this plots show?

Answer:

Question 8. Describe each distribution

In [ ]:
```
# plot the distribution of each variable
```

Question 9. If our model had an assumption of a normal distribution would we meet that requirement?

Answer:

Question 10. Does this distribution need correction? If so, why? How?

Answer:

Question 11. Which of our variables are potentially colinear?

In [ ]:
```
# create a correlation matrix for the data
```

Question 12. What did you find?

Answer:

Question 13. Write an analysis plan for exploring the association between grad school admissions rates and prestige of undergraduate schools.

Answer:

Question 14. What is your hypothesis?

Answer:

INTRODUCTION TO REGRESSION ANALYSIS

Q & A

INTRODUCTION TO REGRESSION ANALYSIS

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET!