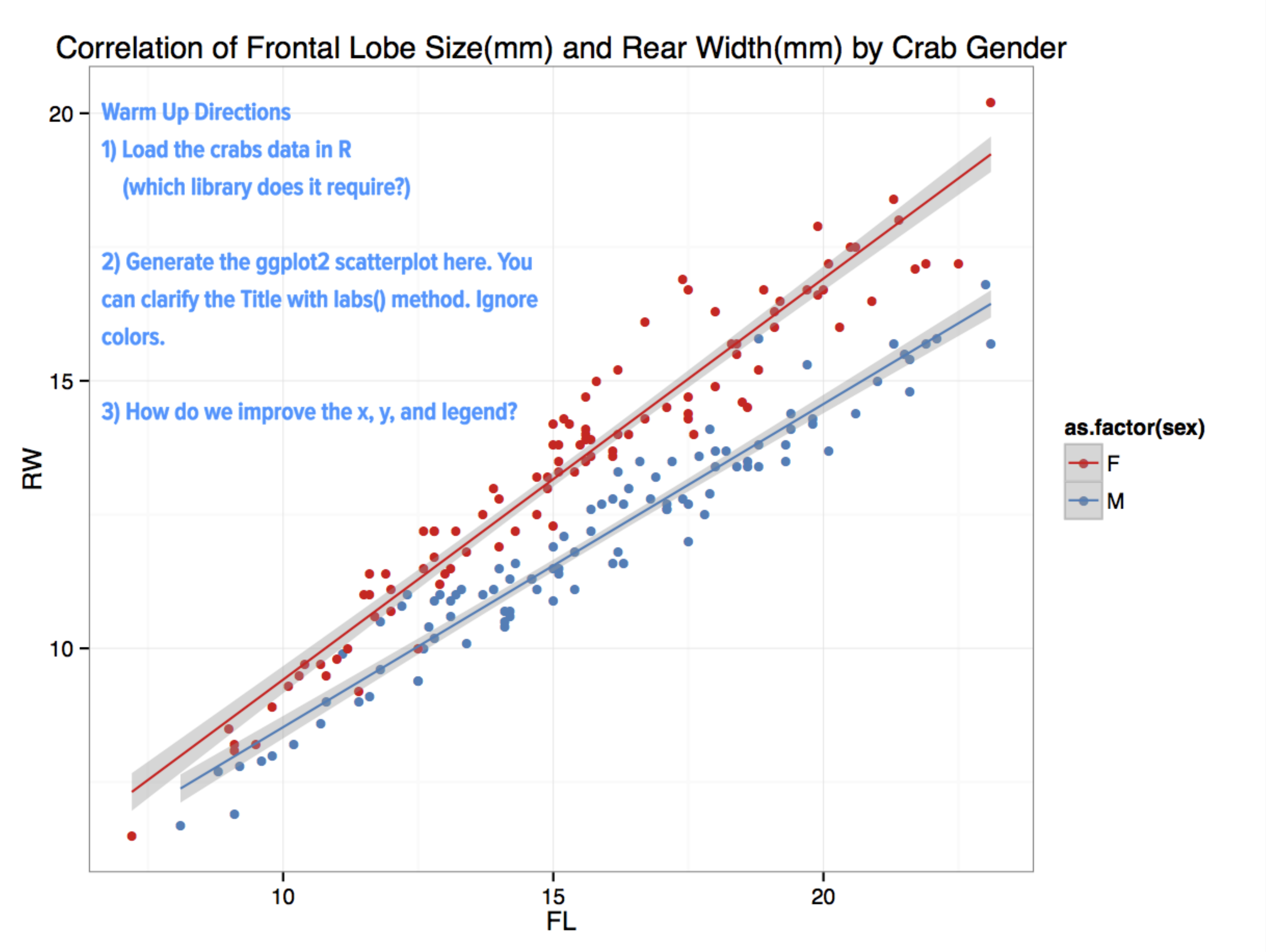
Intro to Data Science:

Machine Learning

Ed Podojil Data Engineer/Scientist, Animoto

warm up



Correlation of Frontal Lobe Size(mm) and Rear Width(mm) by Crab Gender

Warm Up Directions

1) Load the crabs data in R
   (which library does it require?)

2) Generate the ggplot2 scatterplot here. You can clarify the Title with labs() method. Ignore colors.

3) How do we improve the x, y, and legend?

Agenda

What is Machine Learning?

Machine Learning Problems

Multiple Regression & Feature Extraction

What is Machine Learning?

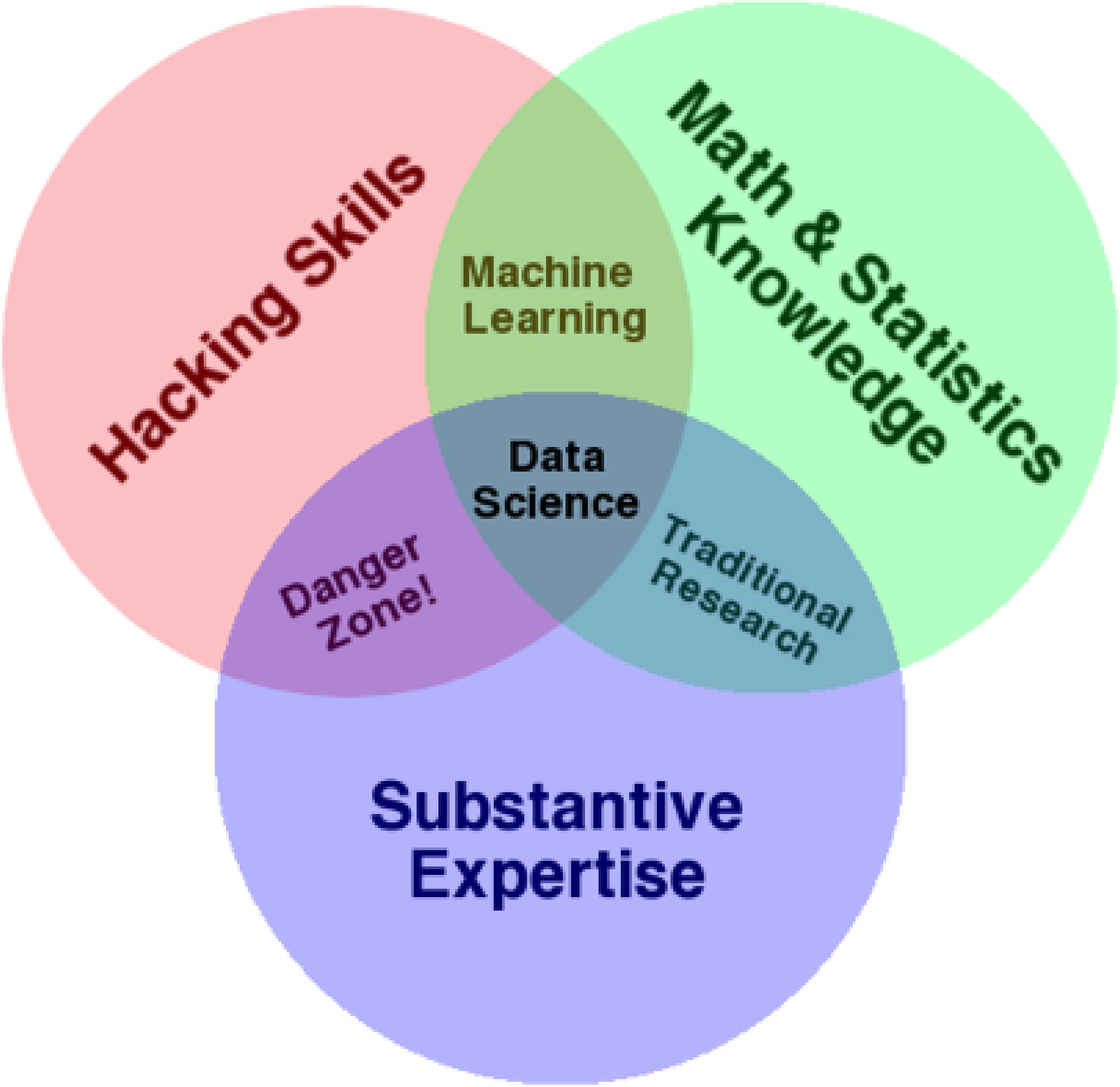
What is Machine Learning?

From WikiPedia:

> *“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data…The core of machine learning deals with representation and generalization…”*

Representation: extracting structure from data

Generalization: making predictions from data


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

III. Machine Learning Problems

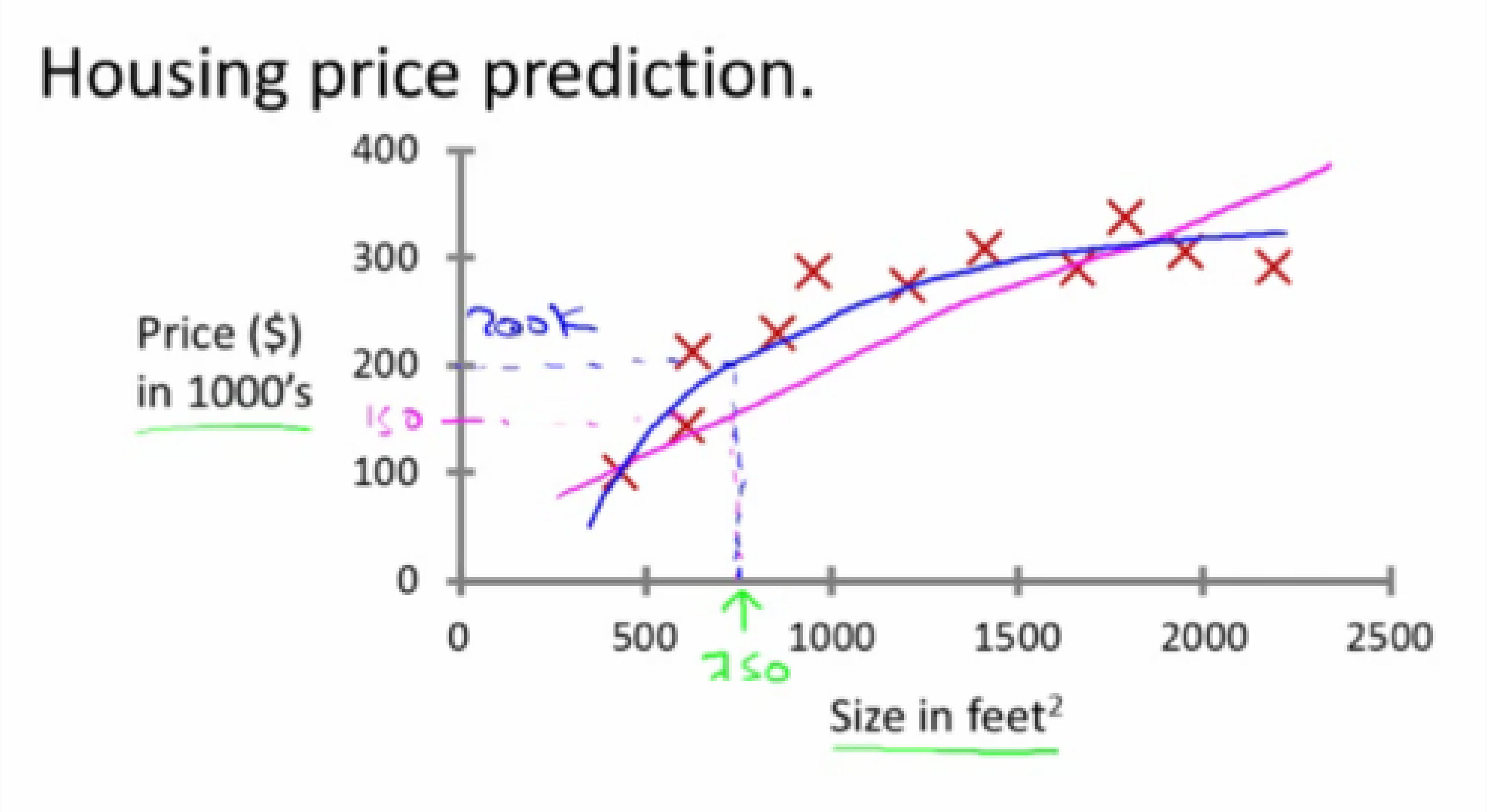
Types of Machine Learning Problems

Supervised Learning

Process used for making predictions

Sample data is already classified

Process uses pre-classified information to predict unknown space



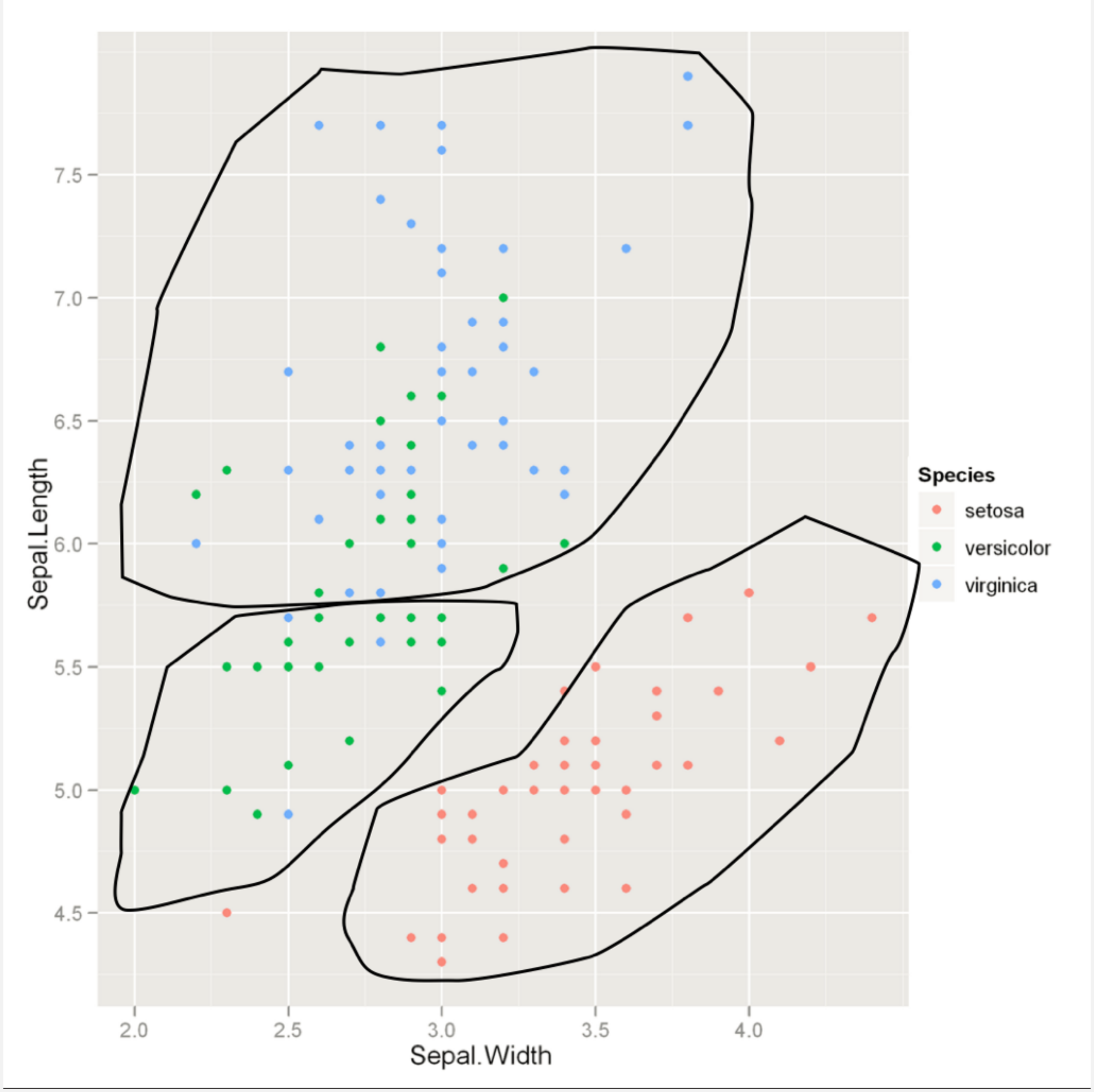
Credit: Andrew Ng, "Introduction to Machine Learning," Stanford

Unsupervised Learning

Process used for providing structure

No data was pre "structured", attempts to make sense out of independent variables

(you're making up, or the algorithm is making up, your dependent variable)



Credit: Thomson Nguyen, "Introduction to Machine Learning," Lookout

Features in Data

We call the space where data lives the "feature space"

Types of Features

Continuous/Quantative

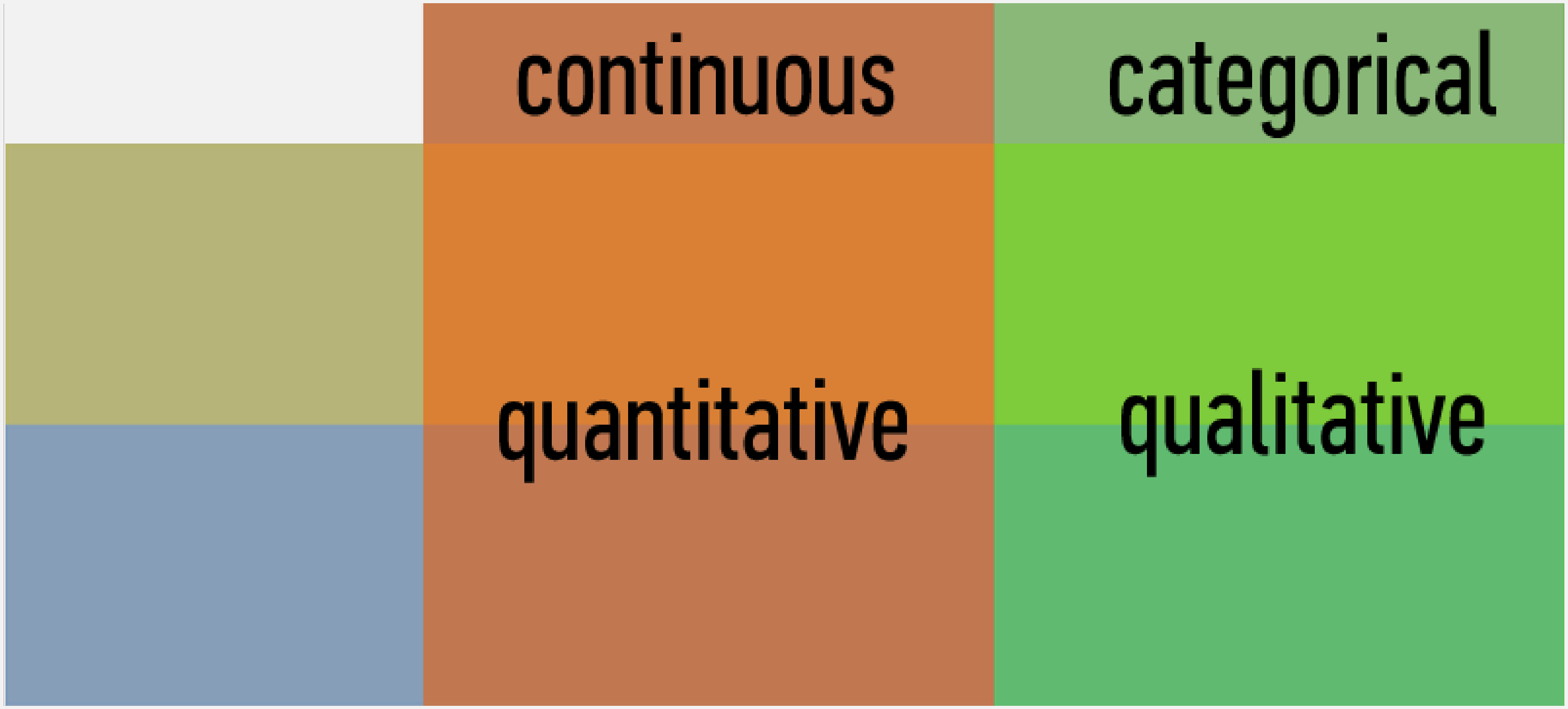
Categorical/Qualitative

Fitting it all Together

What's the goal?

Supervised	Predictions
Unsupervised	Structure

What data do we have?



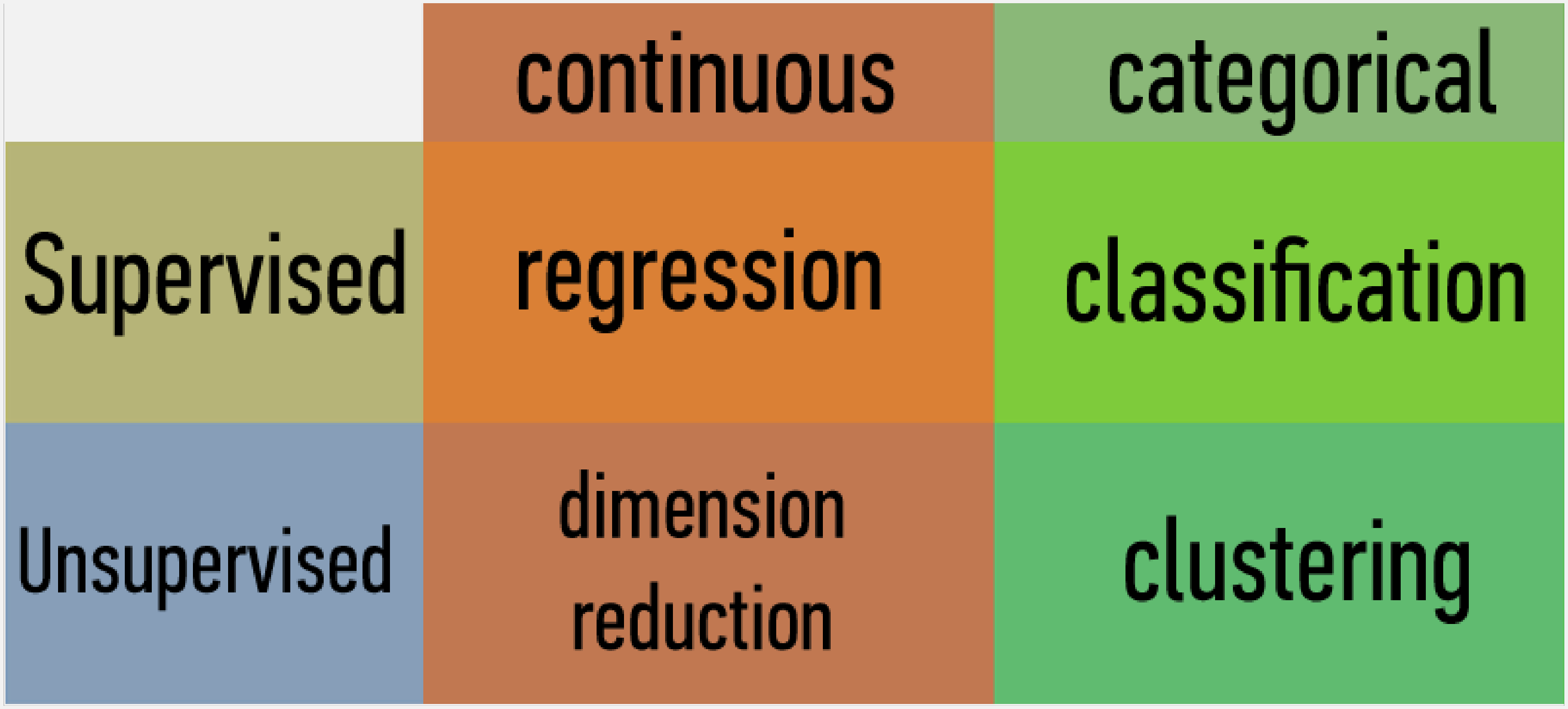
continuous

categorical

quantitative

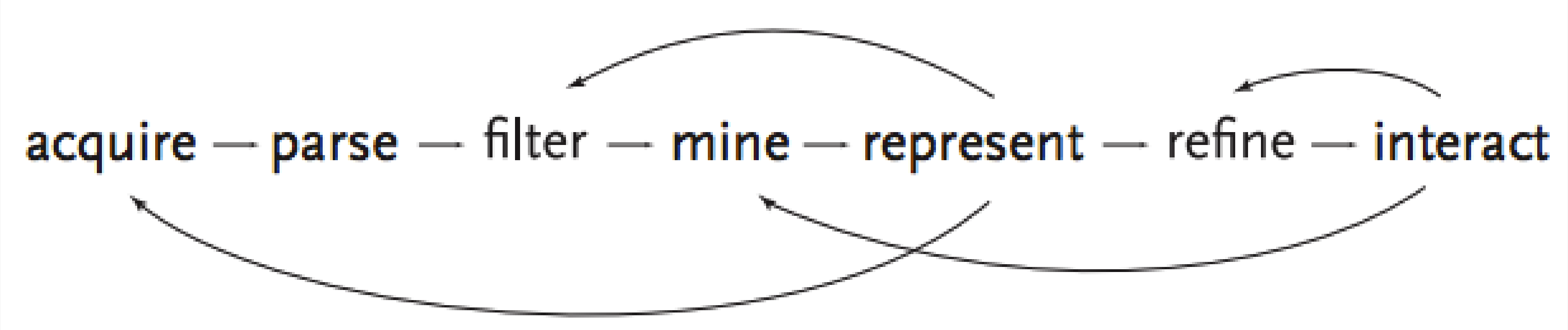
qualitative

How do we determine the right approach?



	continuous	categorical
Supervised	regression	classification
Unsupervised	dimension reduction	clustering

What do we do with the results?



acquire — parse — filter — mine — represent — refine — interact

Class break

In Class Reading/Discussion (25 minutes)

More about R functions

Directions

1. Open up your R Terminal
2. Use R Help (?) to find reading about the following functions
   - plot
   - lm
   - update
3. In a text/markdown file, write a short summary about the function
4. Go through an example from the help and explain what that example is doing

Reading Discussion

Exercise

Multiple Regression (Backward Elimination)

Objectives

Create a regression model using several independent variables

Extract meaningful features

Fit a linear model with multiple independent variables

To find out more about the data: `help(stackloss)`

Are all the independent variables necessary?

```
# Do these first
read.csv('http://heypodo.com/public/etc/enrollment.csv') # dependent var: ROLL
data(stackloss) # dependent var: stack.loss

# If you have more time
data(UScrime) # dependent var: R
data(swiss) # dependent var: Fertility
```

Final Discussion