

Week 2

# Concepts in Machine Learning

[fredhutch.io](http://fredhutch.io)

Fred Hutchinson Cancer Research Center

# Week 2 Learning Objectives

- Quick CRISP-DM Review
- Supervised Learning vs Unsupervised Learning
  - Regression vs. Classification
- Loss functions
- Bias-variance tradeoff
- Common forms of regression
- Other forms of regression

# Definitions

- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.  
[\(https://expertsystem.com/machine-learning-definition/\)](https://expertsystem.com/machine-learning-definition/)
- Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. [\(https://en.wikipedia.org/wiki/Machine\\_learning\)](https://en.wikipedia.org/wiki/Machine_learning)
- Problem + Data + Algorithm(self-adjusting) + Compute ==> Insight

# Experimental Design

- Difficult to master or even do well
- Close interplay between
  - Goals
  - Methods
  - Data
  - Execution
- Requires thoughtful approach and broad understanding

# Capable Cabinet Maker

- Inspects and understands raw materials
- Uses the tools thoughtfully to shape and join materials
- Chooses approach and tools based on materials and goals
- Thoughtfulness born of experience



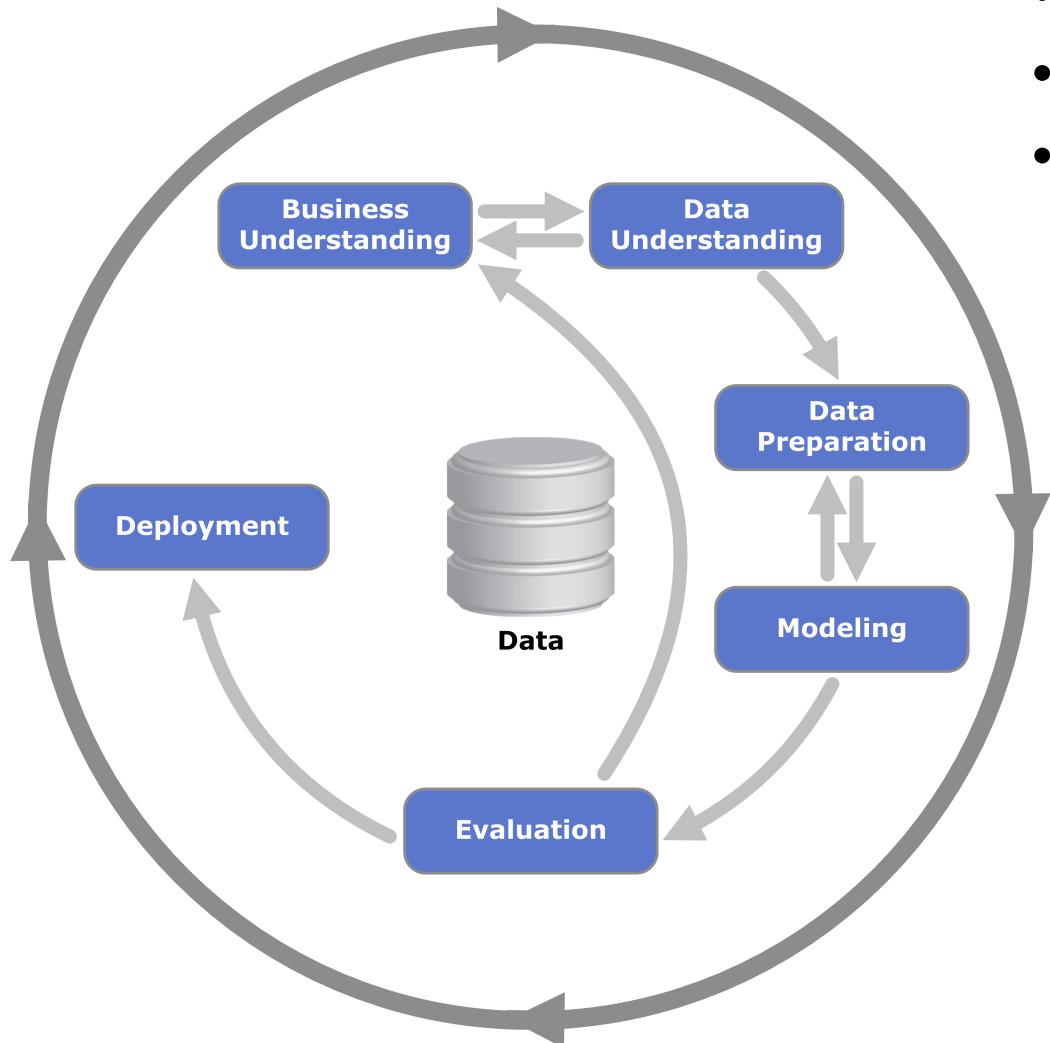
# Cabinet Making?

- Storage Need + Raw Materials + Tools + Work ==> Cabinet
- SAT analogy format

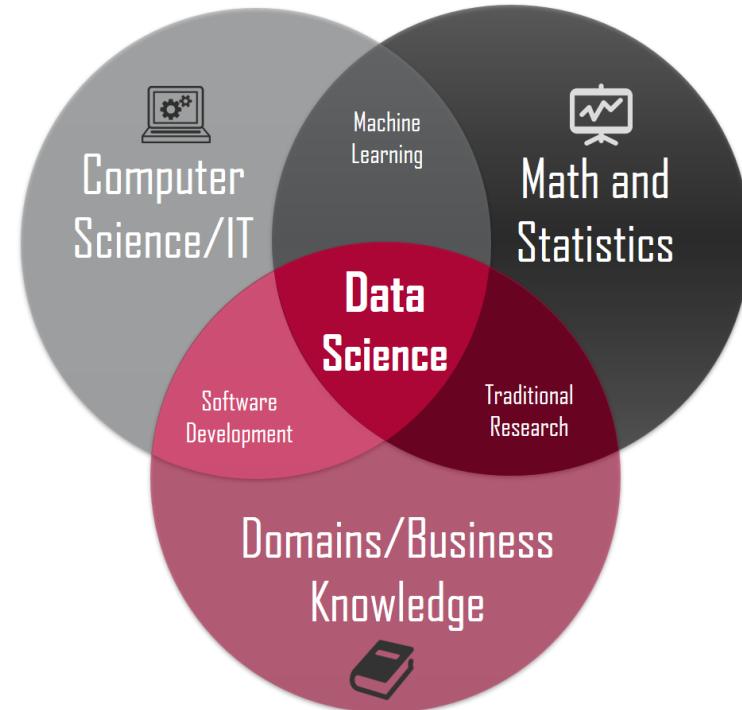
Tools : Cabinet Making as Algorithms : Machine Learning

# CRISP-DM

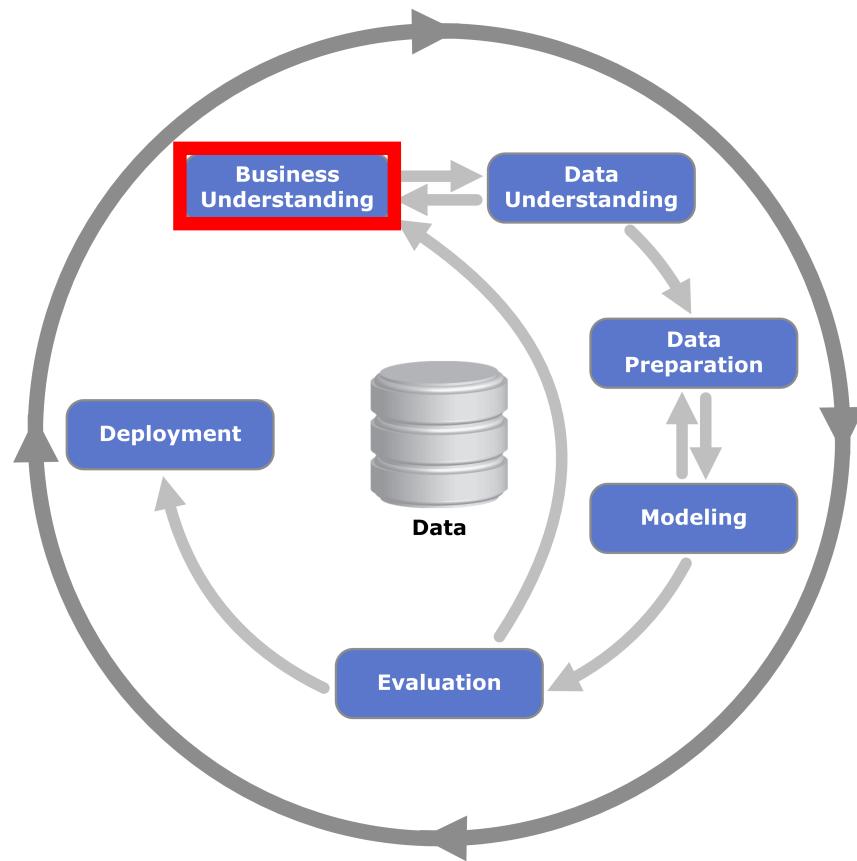
## Cross-industry standard process for data mining



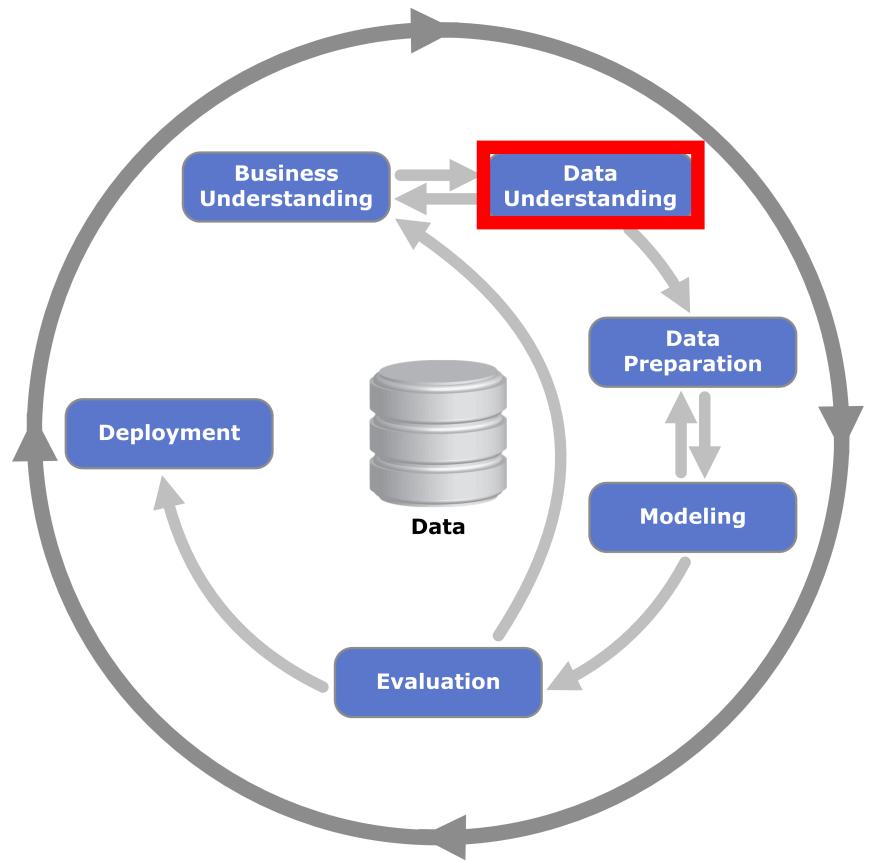
- Cyclical
- Iterative
- Connecting all 3 areas of the classic notion of “data science”



# Business/Scientific Understanding



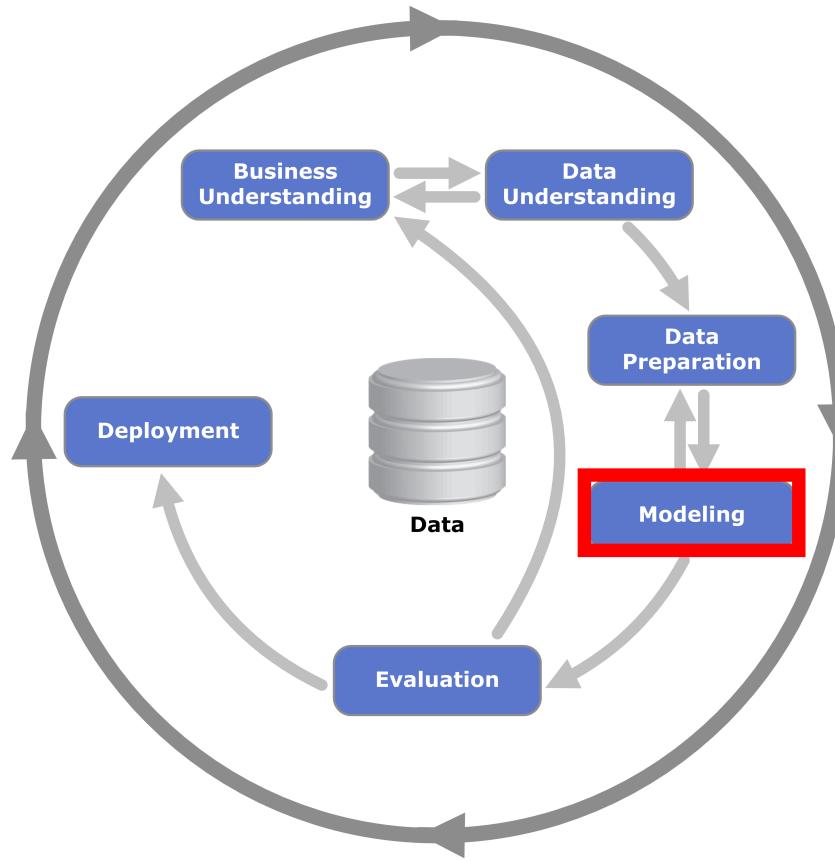
# Data Understanding



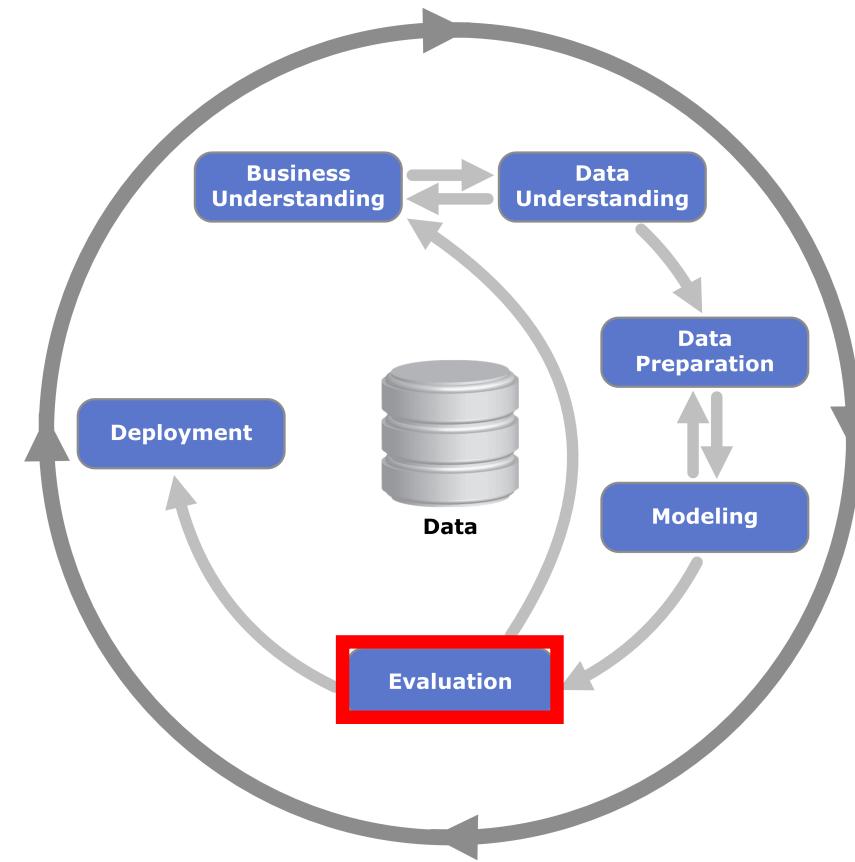
# Data Preparation



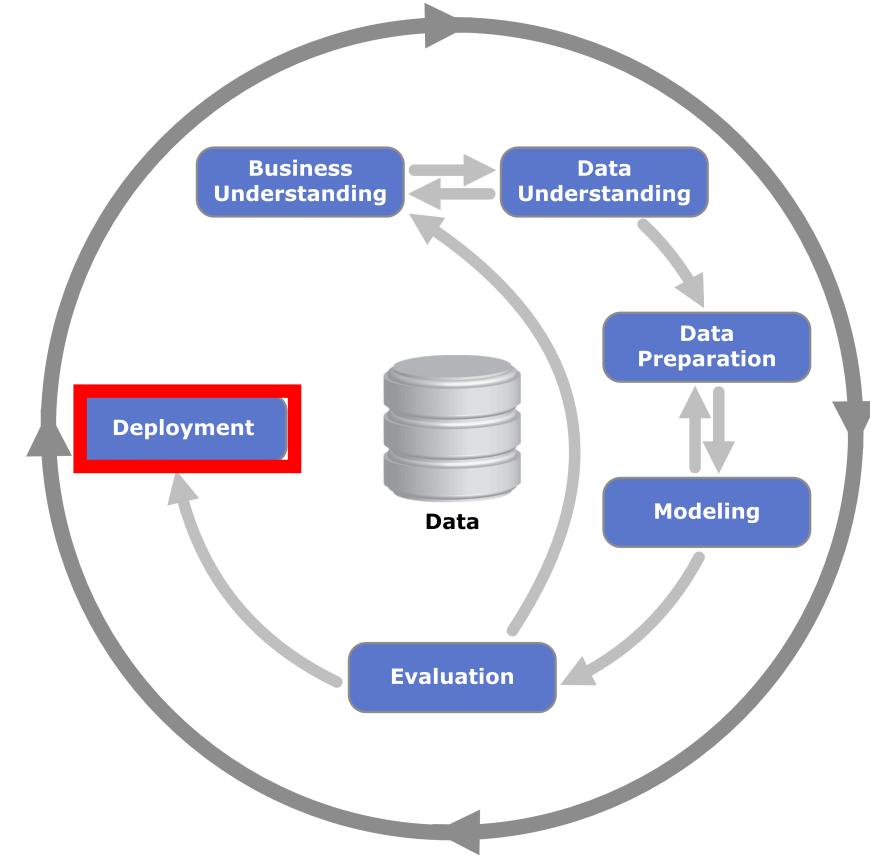
# Modeling



# Evaluation

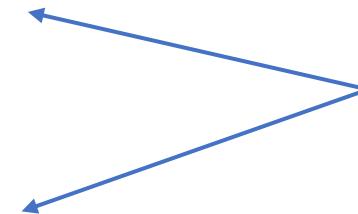


# Deployment



# 3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning



We'll focus on these

- Reinforcement Learning
- Semi-Supervised Learning

# 3 or 4 Machine Learning Paradigms

- Supervised Learning

- Unsupervised Learning

- Reinforcement Learning

- Semi-Supervised Learning

Data “Prediction”



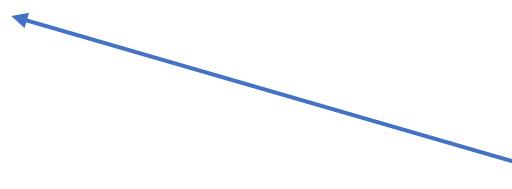
# 3 or 4 Machine Learning Paradigms

- Supervised Learning

- Unsupervised Learning

- Reinforcement Learning

- Semi-Supervised Learning



Data “Expression”

# Drosophila – The Fruit Fly



- Common model organism
- Data on mix of 500 wild-type and mutants
- Data on 10 classic genetic mutations including
  - Wing shape, size, color
  - Eye color
  - Fly size, color
- Data on 50 other metrics, including lifespan



# Key Considerations & The Task

- We are not generating these data
  - Observational data rather than Experimental Data
  - Model will not establish causal relationships (although it could suggest some to evaluate)
- We have one job: Try to predict the lifespan for new, unseen flies

# Is it a Supervised Problem?

- Do we have data?
- Do we have some features within the data that represent what we ultimately want to predict?
- If so, we can formulate it as a Supervised Problem
  1. “Train” a model by predicting the label and comparing to the correct answer. Update the model when we are wrong.
  2. “Test” the trained model by predicting the label of new data and evaluate
    - Our goal is a generalizable model- one that applies to new data well

# Question Time

- Why do we want to test on new data? What would happen if we didn't?

# Question Time

- Why do we want to test on new data? What would happen if we didn't?
  - We would already know the correct answer, since we have trained on it

# Supervised Problems, cont'd:

## More things to consider

- What is the best performance we expect from this predictor?
  - “If a human being made these predictions given this information, how good would they do?”
- Are the data relatively sparse?
  - How much data is missing/has been imputed?
  - How many variables of input are there relative to the total number of examples?

# Supervised Problems: Categorical Problems

- Can we state our outcome as a choice between of A **vs.** B? (could be any number of categories)

# Supervised Problems: Categorical Problems

- Can we state our outcome as a choice between of A **vs.** B? (could be any number of categories)
  - No, we're not choosing a category

# Supervised Learning: Regression Problems

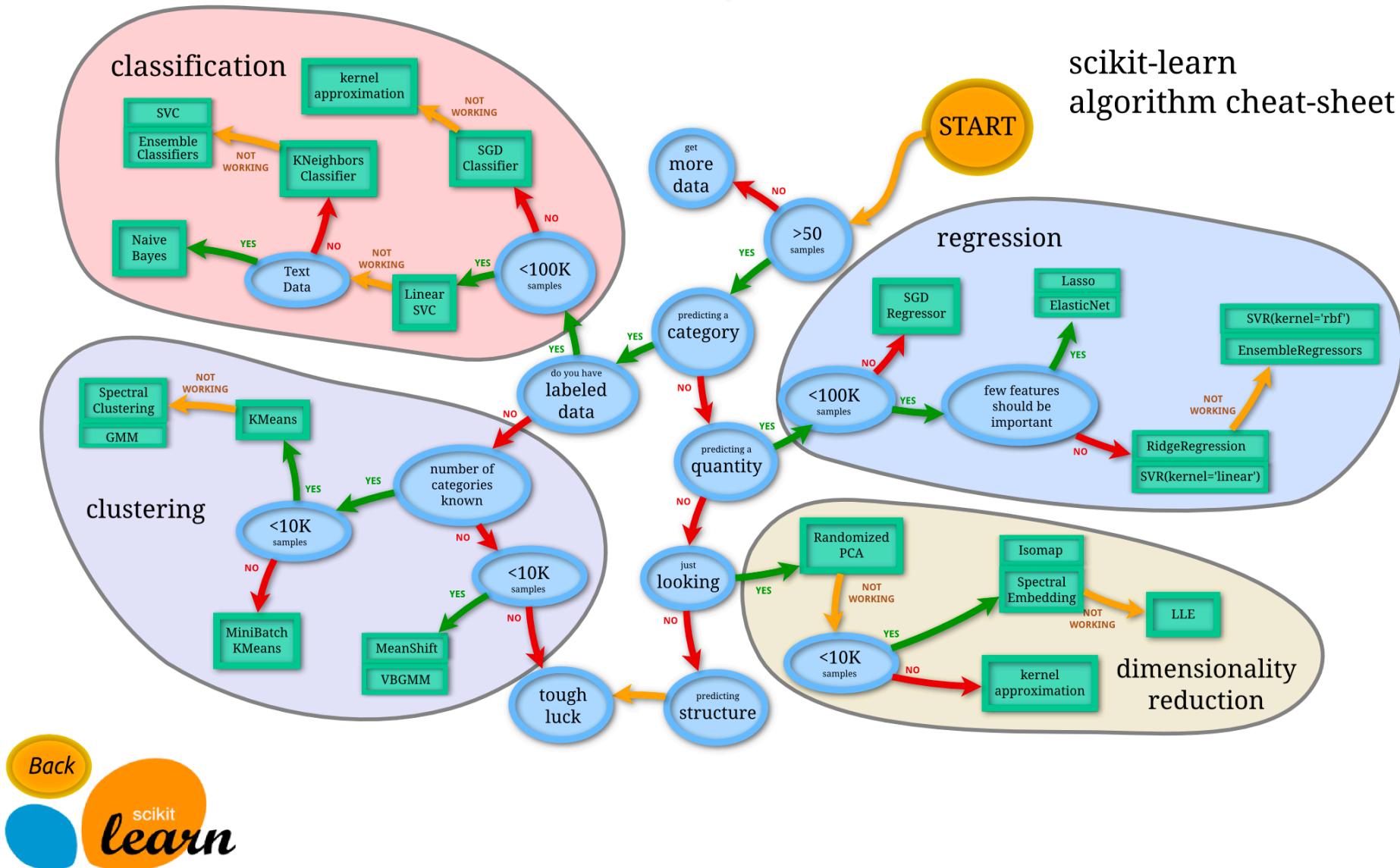
- Can we state our target as a real number?
  - [https://en.wikipedia.org/wiki/Real\\_number](https://en.wikipedia.org/wiki/Real_number)
- Since we don't have distinct categories, we don't have to worry how our data is binned (since there are no distinct bins!)
- What problems do we have to be aware of?
  - Is our data representative of the problem?
  - Are we fitting the wrong regression model to our data?
  - Does our data have outliers that are throwing off our model?

# Quiz Time

Which of these are Supervised Learning Problems?

1. If we know someone's demographics, can we guess their favorite movie genre?
2. What demographic sub groups exist of people that like Horror movies?
3. If we know someone's demographics, can we guess their weight?

# A Nifty Chart



# A Nifty Chart

[https://scikit-learn.org/stable/tutorial/machine learning map/](https://scikit-learn.org/stable/tutorial/machine_learning_map/)

# The chart is suggesting...

- We have >50 samples
- We are not predicting a category
- We are predicting a quantity
- We have <100K samples
- We're not yet sure how many features should be important...
  - So let's dive into some regression!

# Data are Messy

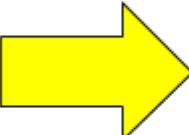
- Most effort will be spent on cleaning, imputing, and transforming the data to make new or better input.
- The second most effort will be spent on analyzing the results and figuring out if they are:
  - Meaningful
  - Good enough

# Issues with the data I

- Quite a few of the rows have one or more values missing
  - What should we do?
- Remove rows with missing values?
- Remove features with missing values?
- Impute mean values for numeric features?
- Impute majority category for categorical features?

# Issues with the data II

- Some of the features are categorical, with values like colors, numbers representing categories, names of genes
  - What should we do?
- Create dummy variables/ one-hot encoding?



Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow			

# You Will Not Have Enough Data

- The more, and more varied the information you have, the more useful your model and predictions will be
- Having too many variables (columns) and not enough observations (rows) leads to problems of **sparsity**
- Having too little information to train over leads to **ungeneralizable models** or **over-trained models** (these are essentially the same thing)

# When we say regression...

- Most people think of linear regression
  - We'll start here
- Logistic regression usually comes to mind next
  - We'll leave this for the classification case study next week
- Our approach will be dictated by lots of things, including:

No of  
independent  
variables

Shape of the  
Regression line

Type of  
dependent  
variable

# Simpler than simplest first...

- Simpler than linear regression?
  - Predict the mean.
  - This will be our baseline a bit later...

# Ordinary Least Squares (OLS)

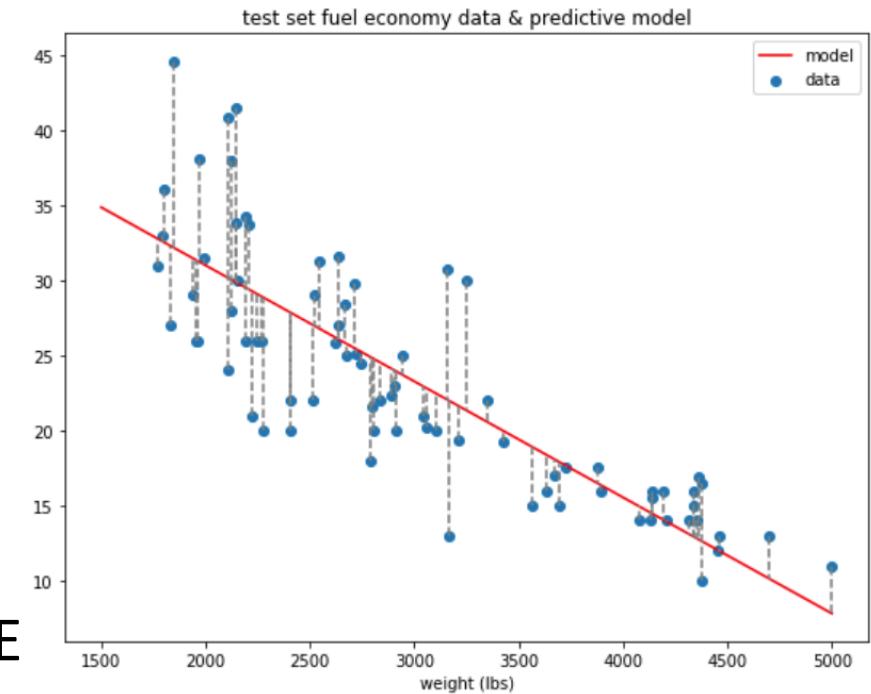
- This is the classic
  - Similar to  $y = mx + b$ , but with more terms

$$\hat{y}_i = \beta_0 \cdot 1 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p}$$

- Which terms should we include?
- Why not just include ALL the possible terms?
  - Hint: training data versus test data

# OLS cont.

- Ordinary Least Squares
  - Tries to tune coefficients by minimizing the SSE
    - sum of the squared errors
  - Lets us understand the mean change in a dependent variable given a one-unit change in each independent variable
  - You can also use polynomials to model curvature and include interaction effects
    - Despite the term “linear model” we can still model curvature



# More advanced versions...

- OLS is sensitive to outliers and has a few other issues
- Ridge Regression: similar, but has an additional term that helps prevent overfitting
- Lasso Regression: similar to Ridge, but also tries to increase accuracy by trying to reduce the number of features
- There are many variations on linear regression and handfuls of non-linear regressors that all have uses

# Loss Functions/ Evaluation Metrics

- Our model has to tune itself using some metric
  - Common choices
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - Mean Absolute Error (MAE)
    - R Squared ( $R^2$ )
    - Adjusted R Squared ( $R^2$ )
    - Mean Square Percentage Error (MSPE)
    - Mean Absolute Percentage Error (MAPE)
    - Root Mean Squared Logarithmic Error (RMSLE)
  - Classifiers will have different metrics
- Many of these are useful in different contexts. Choose carefully and see what makes the most sense...

# Mean Squared Error (MSE)

- Very common
- AKA Quadratic Loss, L2 Loss
- Emphasizes bad errors enough to lower the quality of the model overall

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

# Root Mean Squared Error (RMSE)

- Similar, a bit easier to interpret
- Interchangeable with MSE in many applications

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} = \sqrt{\text{MSE}}$$

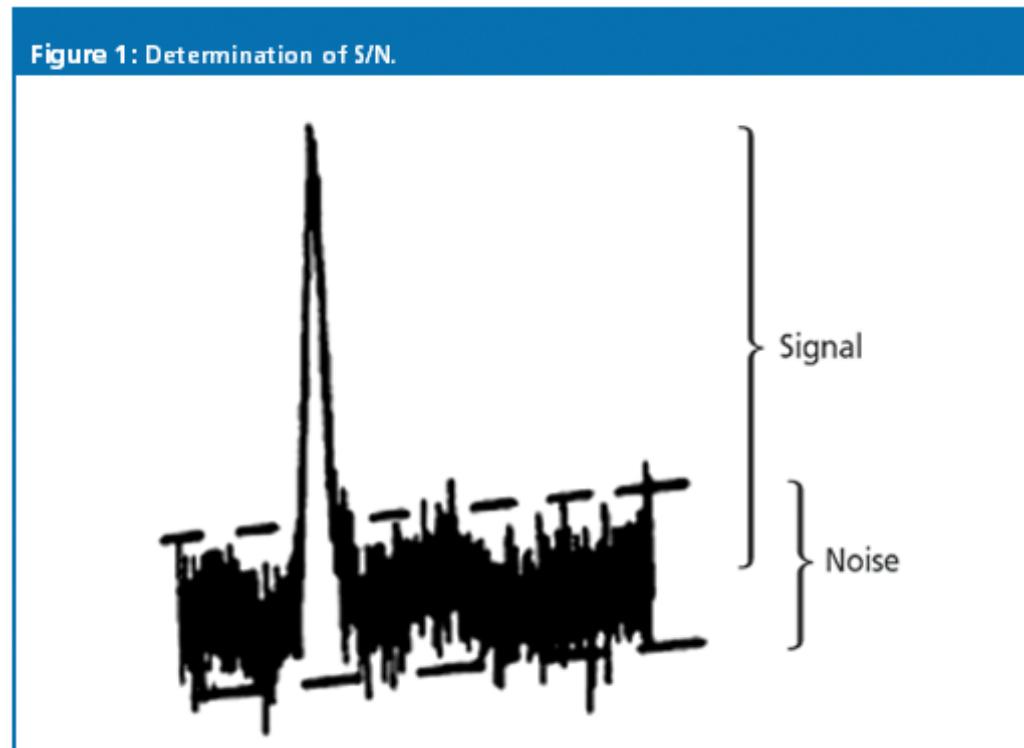
# R Squared ( $R^2$ )

- Like the two above, but normalized to have a best value of 1
- Uses mean as a baseline

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

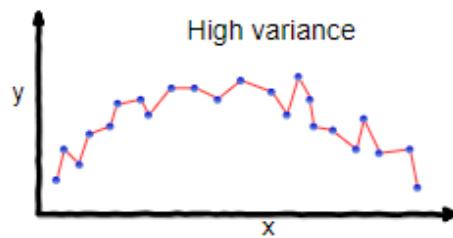
# Bias-Variance Tradeoff

- Fitted model should ideally
  - Capture all of the “signal” in the data
  - Ignore all of the “noise” in the data
- It’s generally hard to do both well

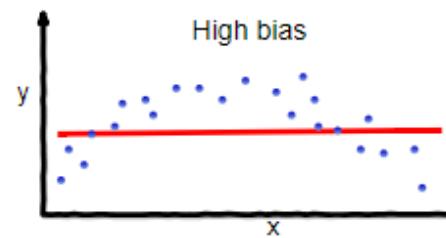


# Bias-Variance Tradeoff

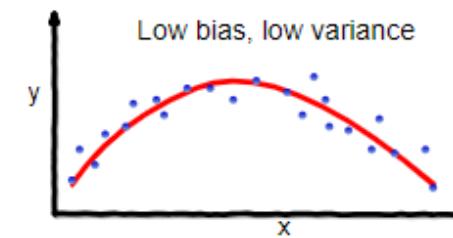
- High variance models
  - Can “overfit” to training data
    - Capture signal...
    - ...but also capture noise
  - Generalize poorly to unseen/test data



overfitting



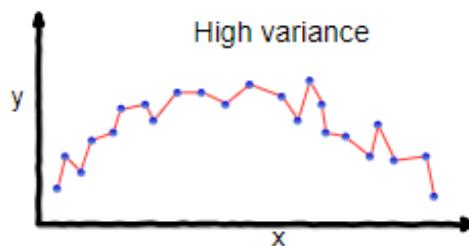
underfitting



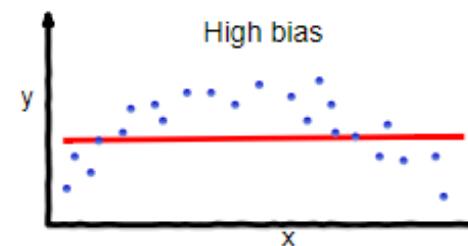
Good balance

# Bias-Variance Tradeoff

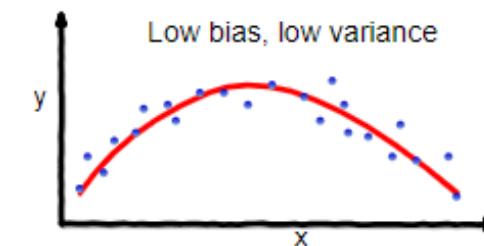
- High bias models
  - Can “underfit” to training data
    - Avoid capturing noise...
    - ...but also fail to capture all signal
  - Perform poorly on unseen/test data



overfitting



underfitting



Good balance

# There are many more regressions to discuss!

- Random Forests!
  - Many completely overfitted decision trees
  - The trees are decorrelated using
    - Bootstrap aggregation
    - Feature selection
  - They're wrong in different ways
  - Together, they can find lots of signal in the data!
- More on Random Forests next week in the Classifier Case Study

