Week 2

Concepts in Machine Learning

fredhutch.io
Fred Hutchinson Cancer Research Center

Week 2 Learning Objectives

By the end of today’s class, you should…

CRISP-DM

- Review what each step generally entails

Machine learning paradigms

- Understand the essential difference between supervised and unsupervised learning

Loss functions

- Understand their basic purpose in fitting and evaluating models
- Recognize common loss functions

Bias-variance tradeoff

- Have working definitions for bias and variance
- Relate the concepts of “signal” and “noise” to high-bias and high-variance models

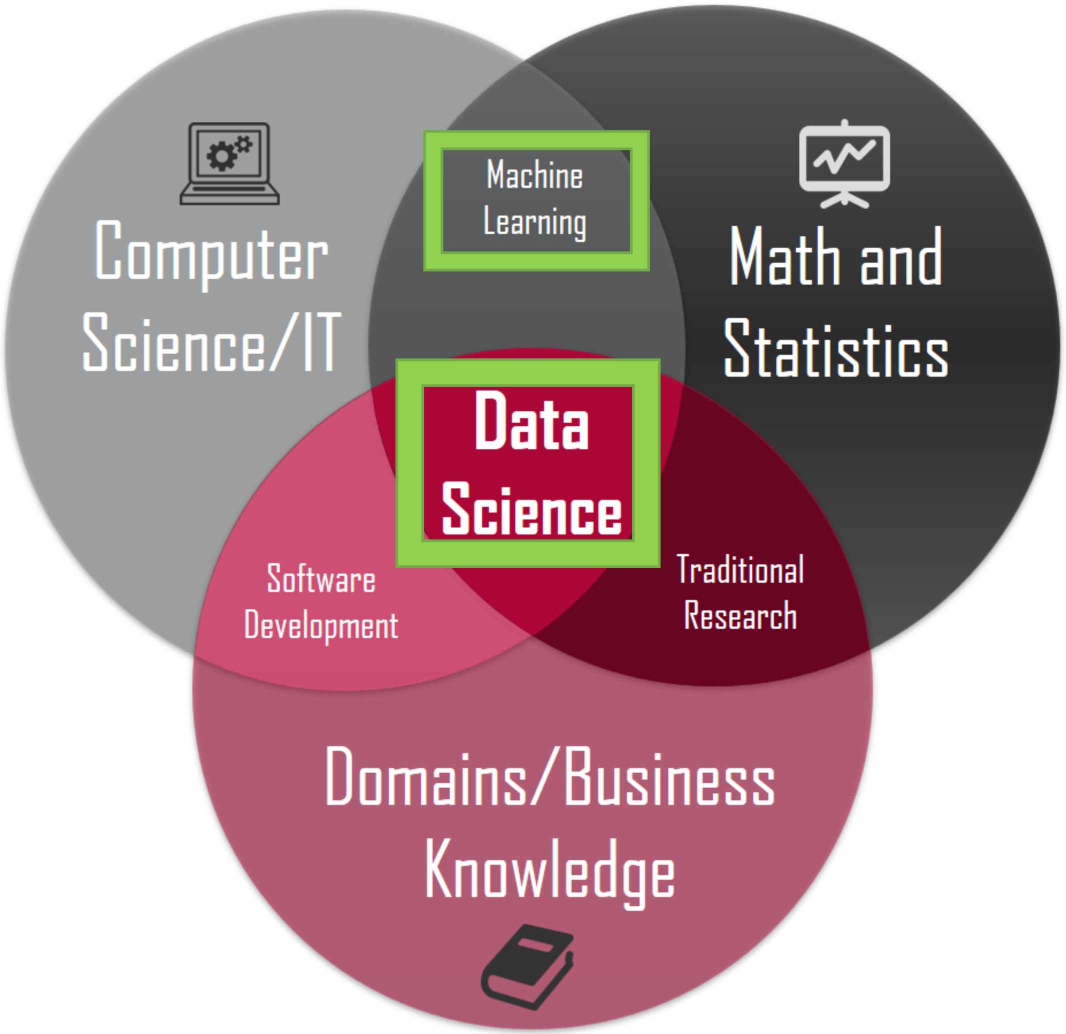
Regression

- Understand the distinction between regression and classification
- Recognize some common approaches to regression

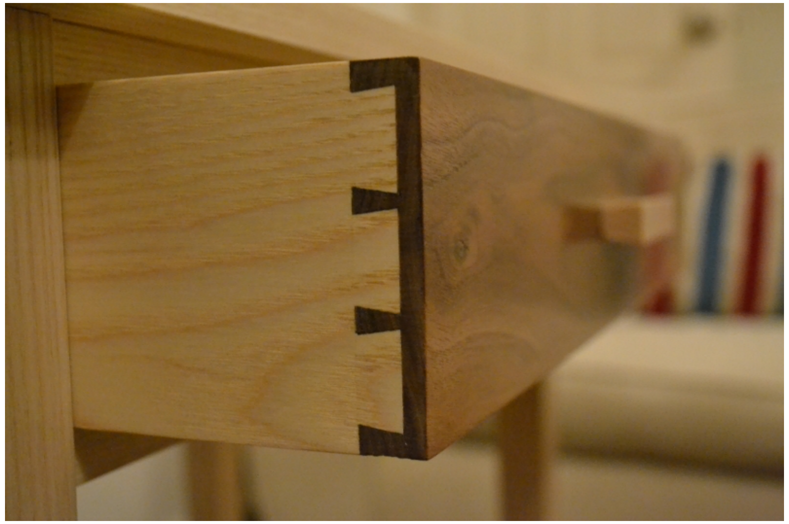
Definitions

- 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/)
- 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)
- Problem + Data + Algorithm(self-adjusting) + Compute ==> Insight

Data Science



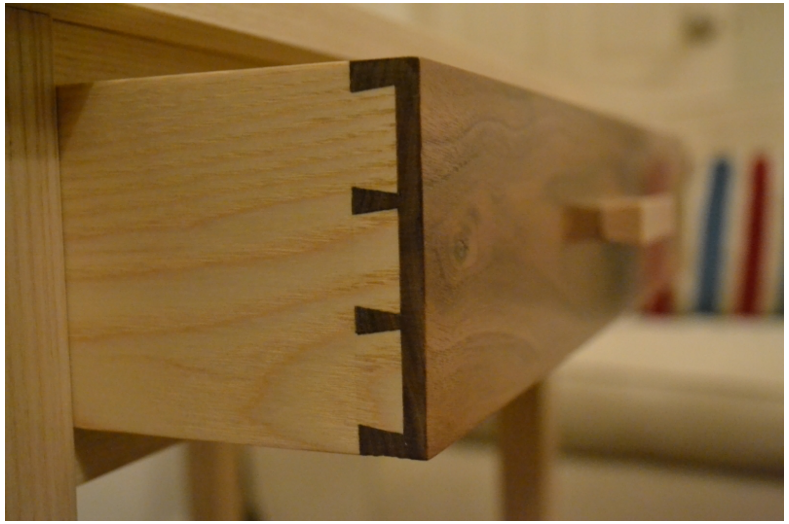
An Imperfect Analogy: Cabinet Making



(not technically a cabinet)

Capable Cabinet Maker…

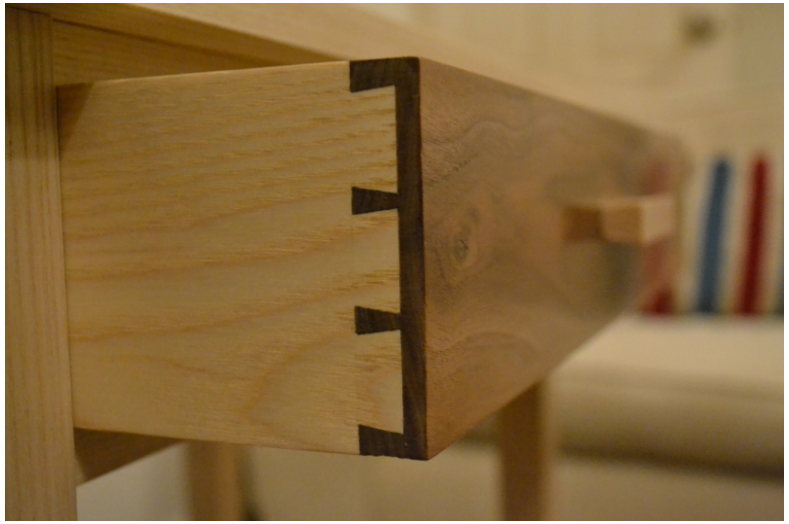
- Inspects and understands raw materials
- Uses the tools thoughtfully to shape and join materials
- Chooses approach and tools based on materials and goals
- Applies thoughtfulness born of experience



…but how does this relate to machine learning?

Capable ~~Cabinet Maker~~ ML Practitioner…

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



…and the tools?

SAT Style Analogy

Tools : Cabinet Making

as

Algorithms : Machine Learning

An Imperfect (Extended) Analogy

- Storage Need + Raw Materials + Tools + Work ==> Cabinet

- Problem + Data + Algorithm(self-adjusting) + Compute ==> Insight

Brief Aside: Experimental Design

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

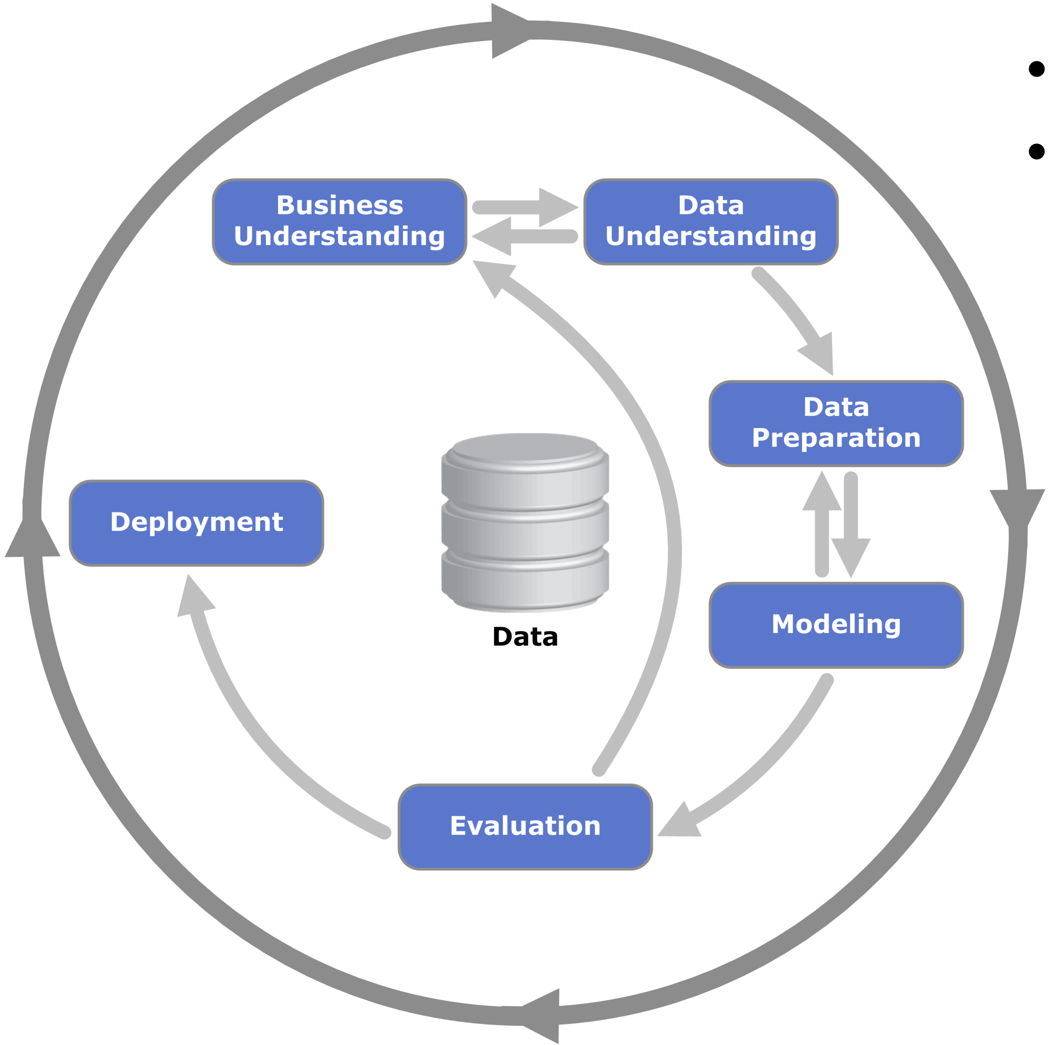
Brief Aside: Experimental Design

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

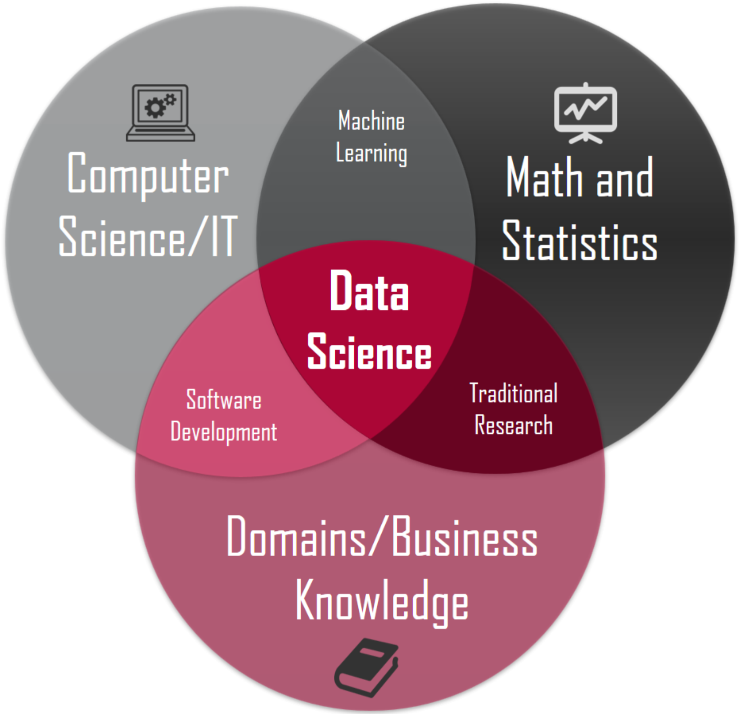
Because machine learning shares so many of these characteristics, it’s helpful to follow a process

Experts in machine learning often are guided by the steps laid out in…

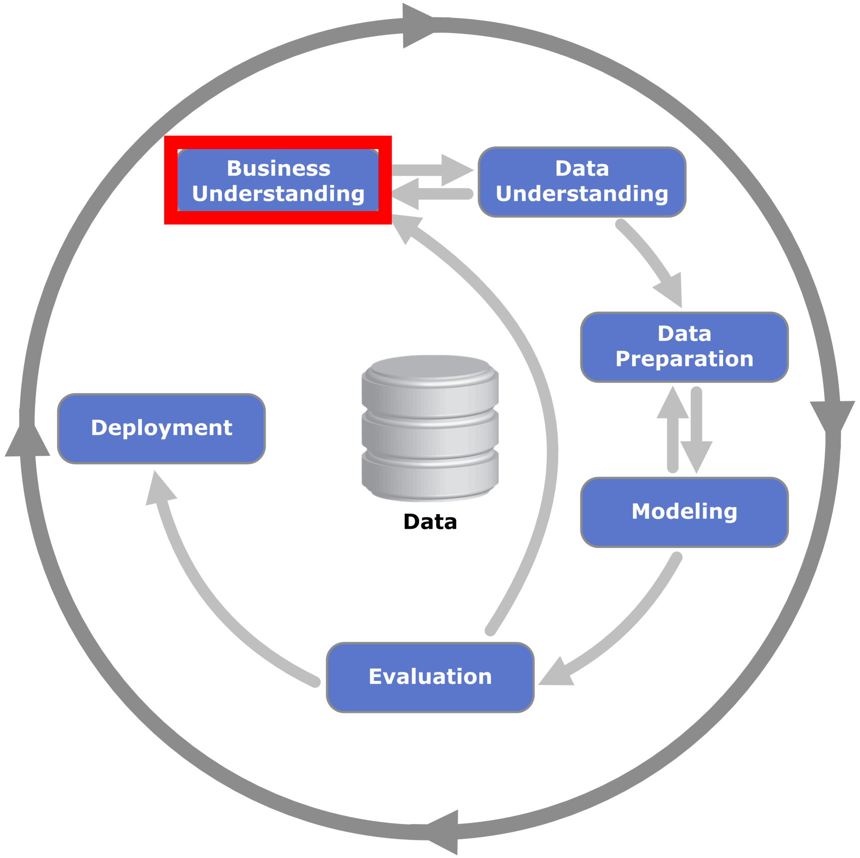
CRISP-DM
Cross-industry standard process for data mining



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

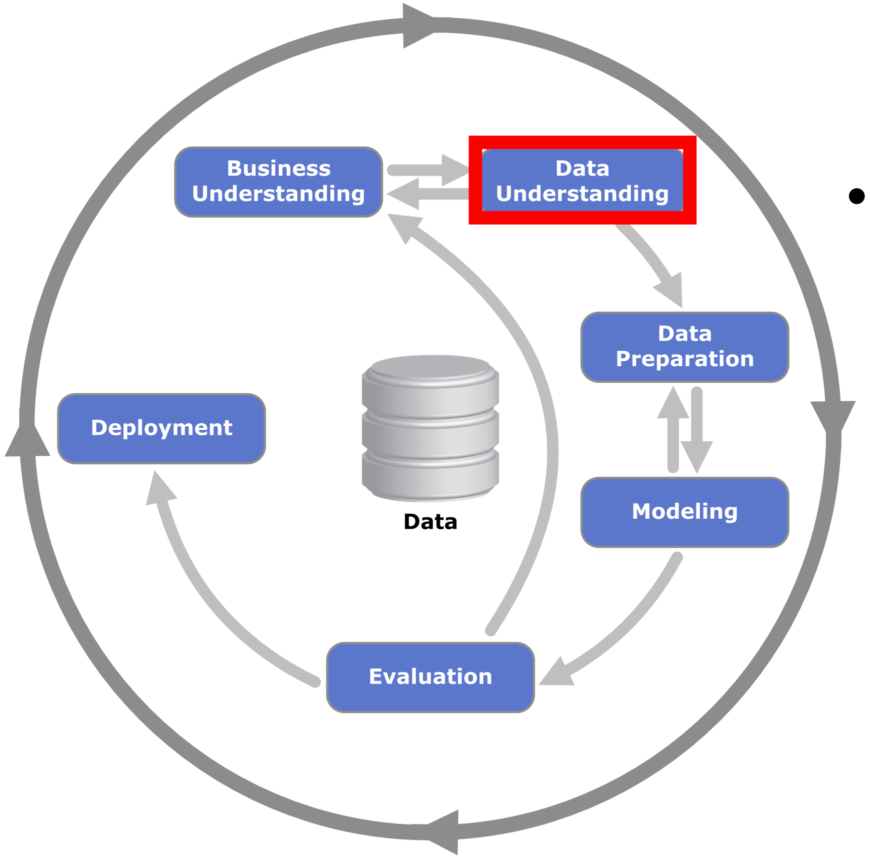


Business/Scientific Understanding



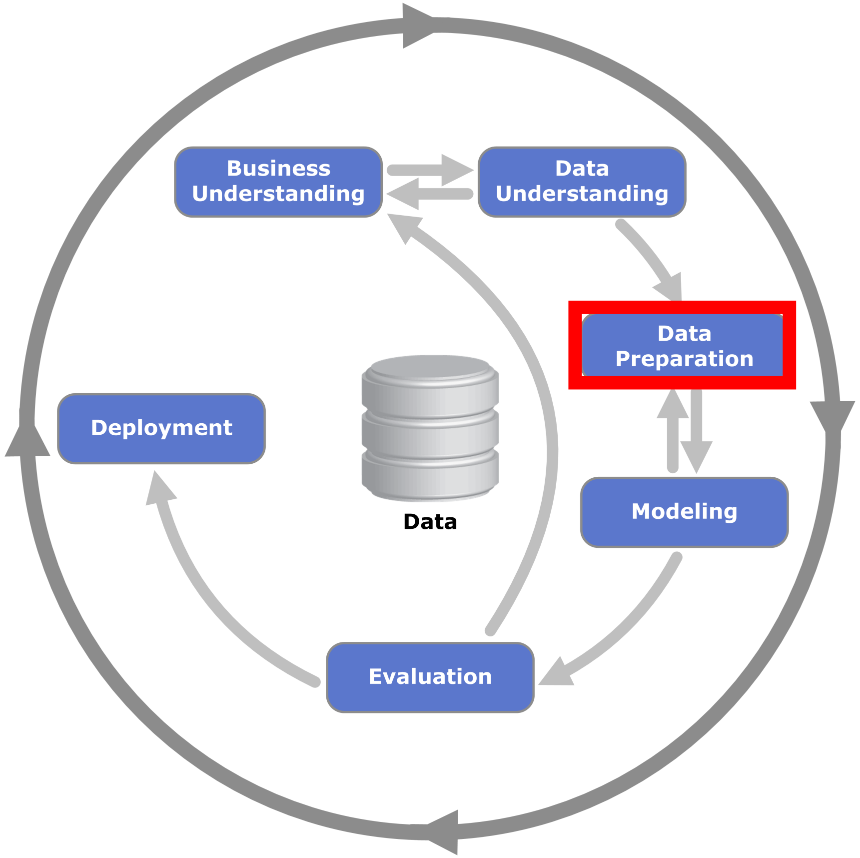
- Scientific domain knowledge
- What do we already know or believe?
- What are our research aims?
- How do we think we should proceed initially?
- Nearly every choice within any of the subsequent steps should refer back to this step!

Data Understanding



- Exploratory Data Analysis (EDA)
  - Data structure
  - Data quality
  - First insights
  - Interesting subsets
  - Form hypotheses for hidden information

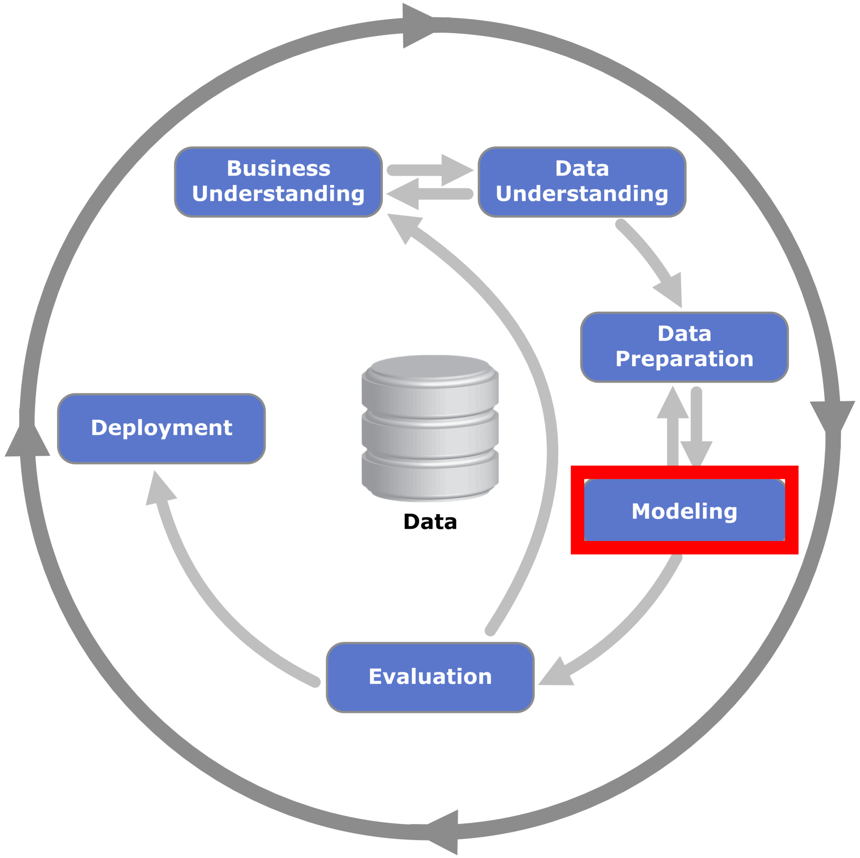
Data Preparation



- Data acquisition
- Data selection
- Data integration and formatting
- Data cleaning
- Data transformation and enrichment

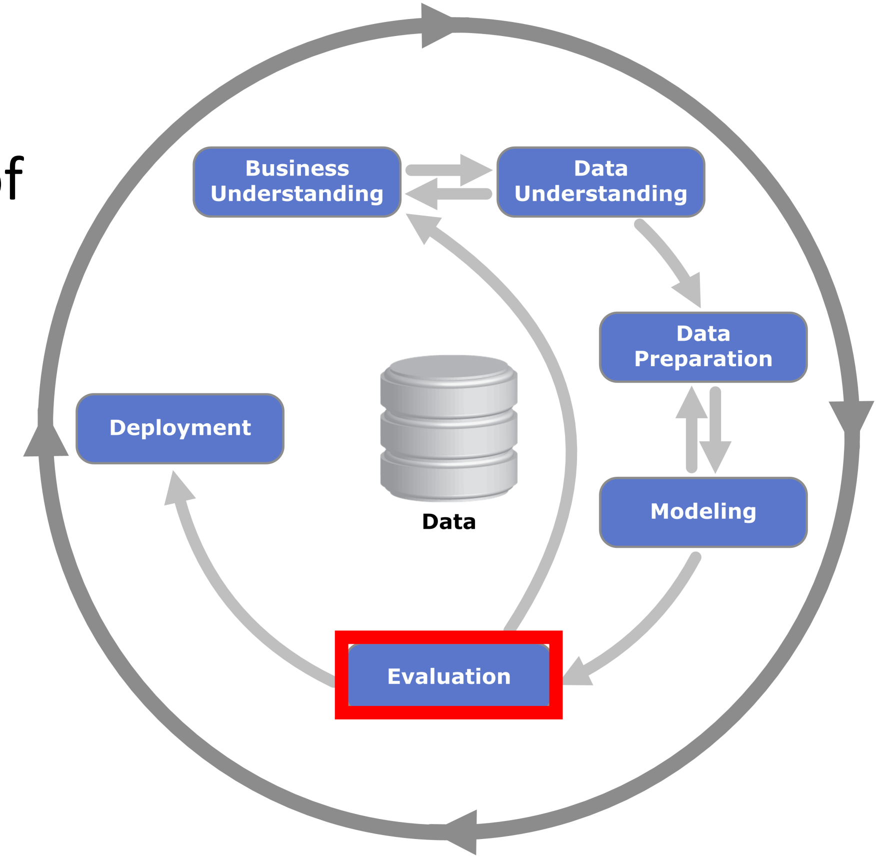
Modeling

- Selection of appropriate modeling technique
- Splitting of the dataset into training and testing subsets
- Examination of alternative algorithms and parameter settings
- Fine tuning of the model settings



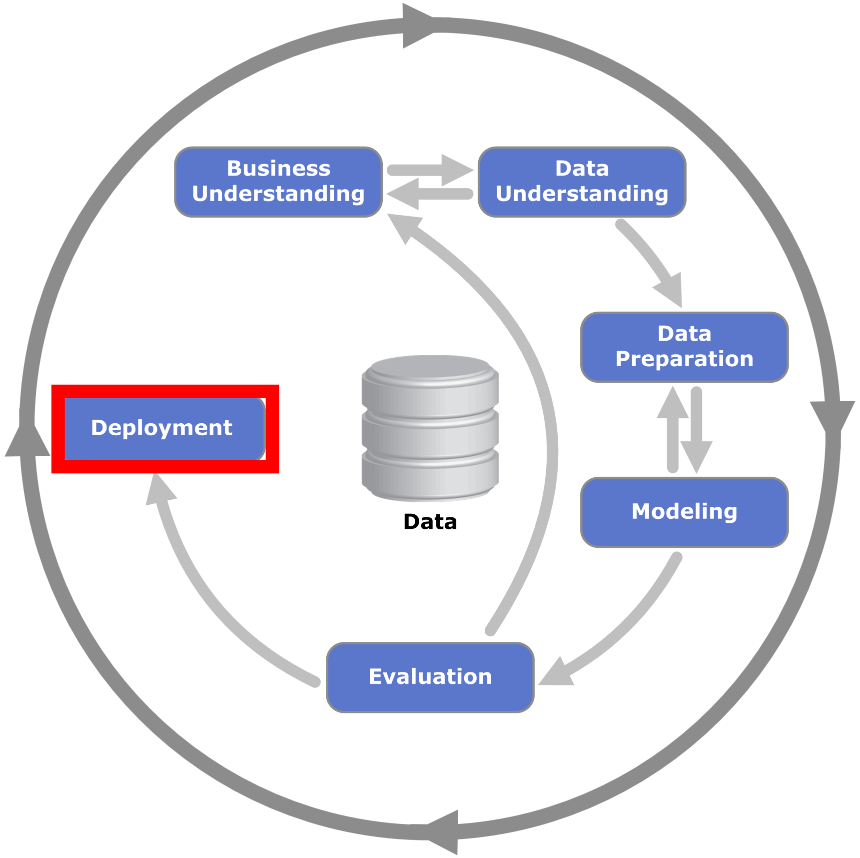
Evaluation

- Evaluation of the model in the context of the scientific success criteria
- Performance relative to TEST data with chosen loss function
- Balancing tradeoff between bias and variance



Deployment

- Will be specific to each problem space
- At FHCRC could relate to grants or publications



Machine Learning Paradigms

par·a·digm

/ˈperəˌdīm/

noun

1. a typical example or pattern of something; a model.
   "there is a new paradigm for public art in this country"

   Similar: model pattern example standard prototype archetype

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
- $\boxed{\text{Unsupervised Learning}}$
- 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 feature within the data that represents 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

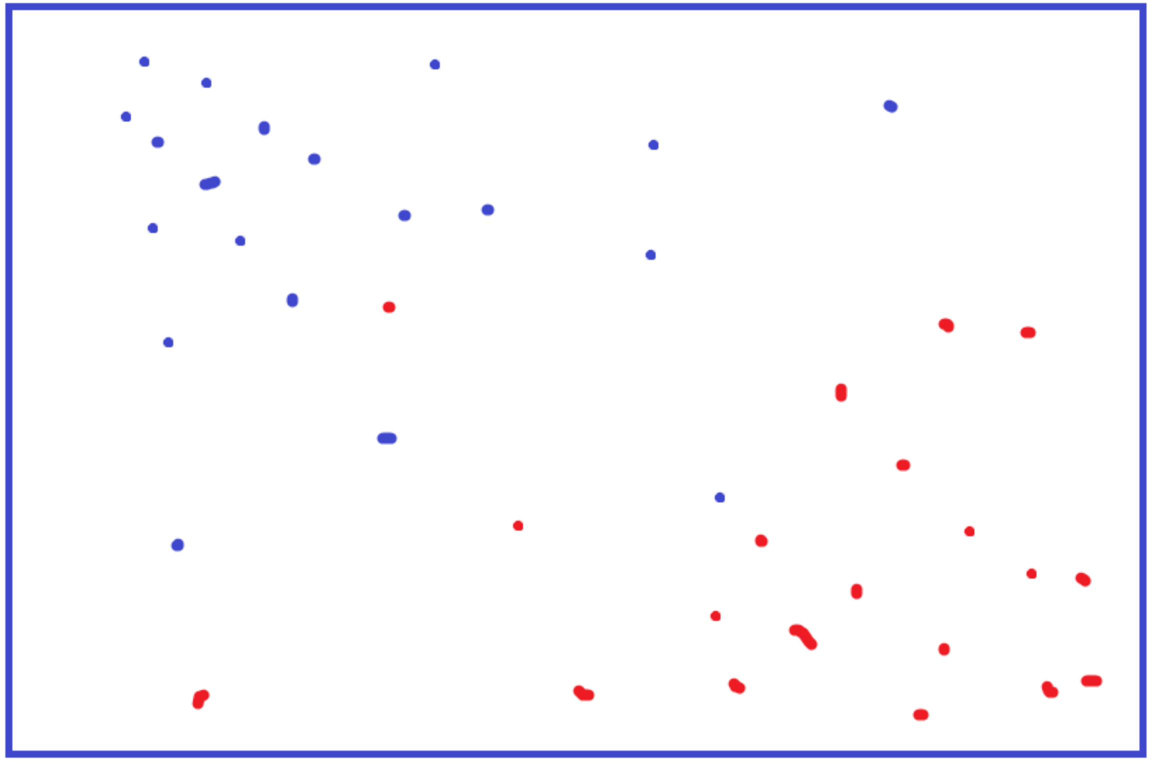
Question Time

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

Machine Learning Paradigms: Supervised Learning

- A way to group or predict new information, based on information we have seen before
- Some example questions:
  - “Given Age, and Height, what is someone’s Weight?”
  - “Given Petal Length and Width, what kind of flower is this?”
  - “Given a Patient’s clinical history*, what is the likelihood* they will have to enter the Emergency Department soon*?

What Problem Statement could we make?



Given a set of coordinates, will a point be Red or Blue?

Supervised Problems: Categorical Problems?

- Can we state our outcome as a choice of A **vs.** B?
  - No, we're not choosing a category

What Problem Statement could we make?

Given a set of descriptors regarding a specific fly, how long will that fly live?

Supervised Problems, cont’d: More things to consider

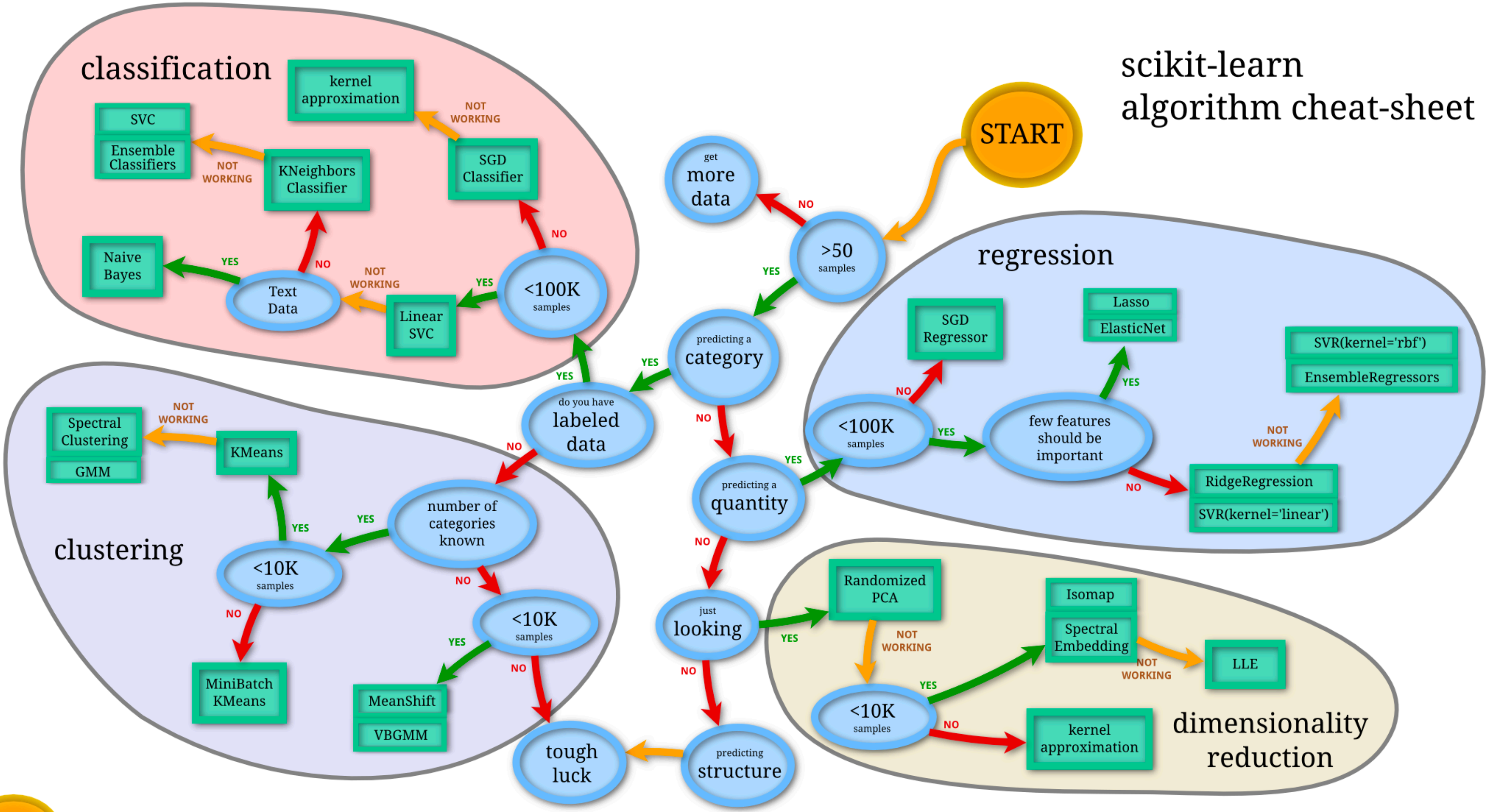
- What is the best performance we expect from this predictor?
  - “If a human being made these predictions given this information, how good would they do?”
- Is the data relatively sparse?
  - How much data is missing/has been imputed?
  - How many variables of input are there relative to the total number of examples?
- How ‘True’ is the target?
  - Does it represent an estimate?
- Are the targets we are trying to predict skewed?
  - Eg: 95% of all participants answered ‘No’, and 5% answered ‘Yes’

Supervised Learning: Regression Problems

- Can we state our target as a 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?
  - Do our data have outliers that are throwing off our model?

A Nifty Chart



Back

A Nifty Chart

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
    - Many models will struggle with features like these
    - What should we do?
- Create dummy variables/ one-hot encoding?

Color
Red
Red
Yellow
Green
Yellow

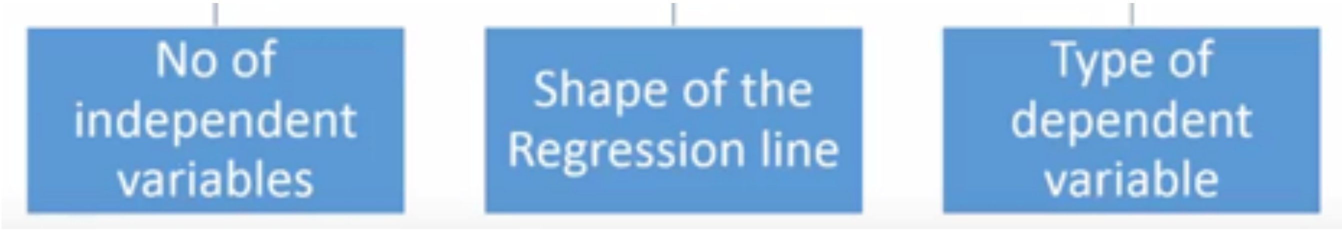
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

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:



Simpler than simplest first...

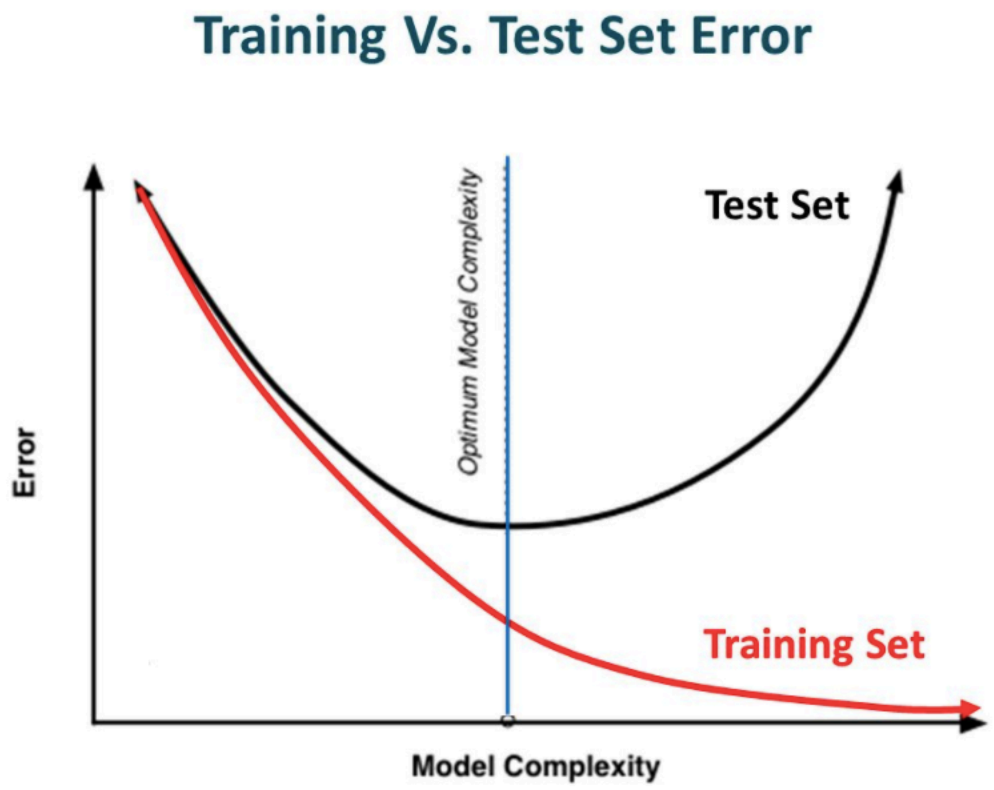
- Simpler than linear regression?
  - Predict the mean.
  - This is like a linear regression, but you force the slope to be 0
  - This will be our baseline a bit later...

Ordinary Least Squares (OLS)

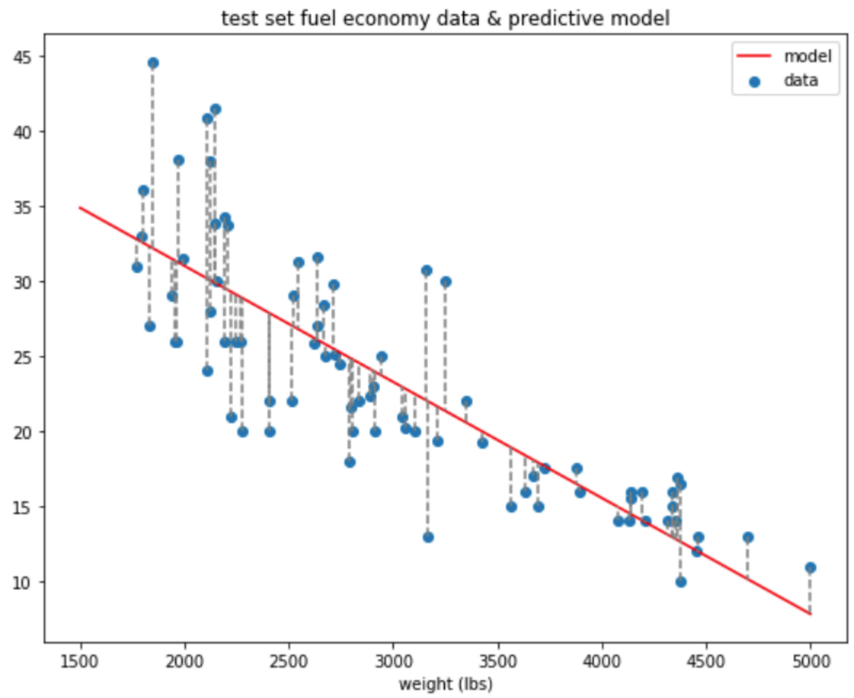
- This is the classic
  - Similar to y = mx + b, but with more terms

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

  - 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

Aspects of structure and wording drawn from https://statisticsbyjim.com

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…

Aspects of structure and wording drawn from https://towardsdatascience.com

Mean Squared Error (MSE)

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

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

Root Mean Squared Error (RMSE)

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

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

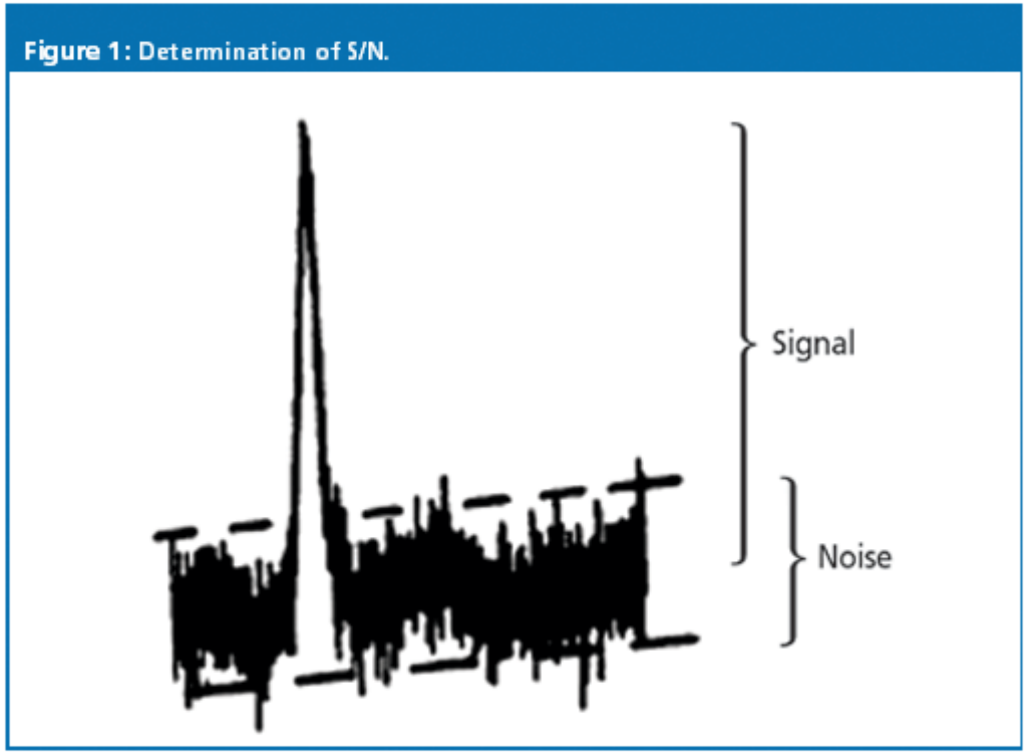
R Squared (R^2)

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

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

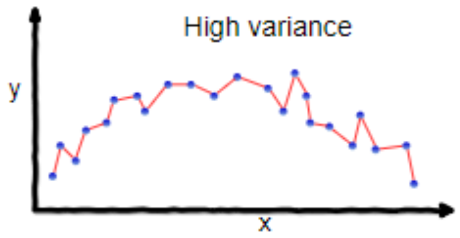
Bias-Variance Tradeoff

- 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

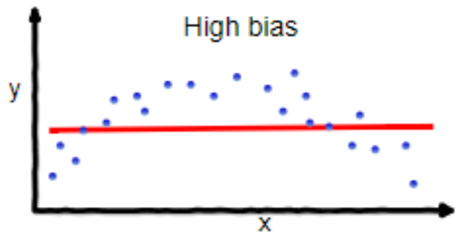


Bias-Variance Tradeoff

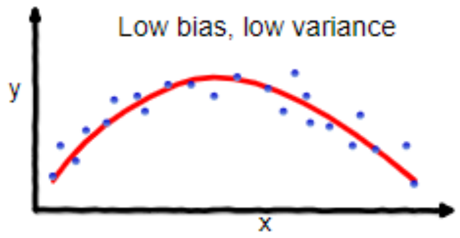
- High variance models
  - Can “overfit” to training data
    - Capture signal...
    - ...but also capture noise
  - Generalize poorly to unseen/test data



overfitting



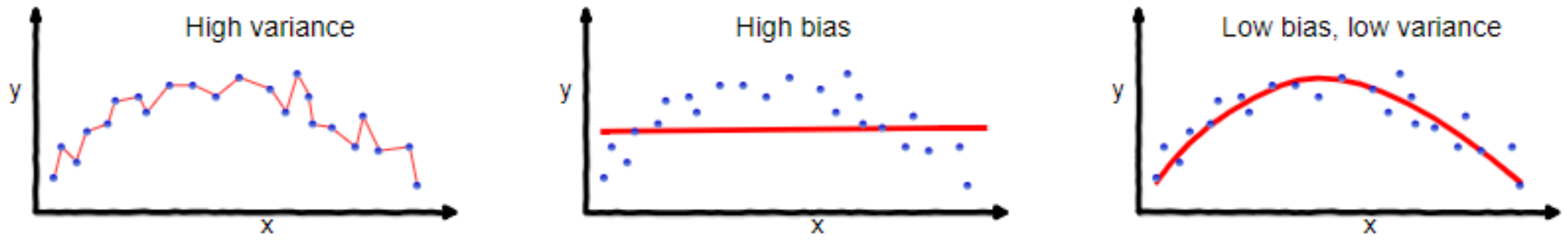
underfitting



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 a 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