Feature Engineering

Chapter 6

Stephen Kimel

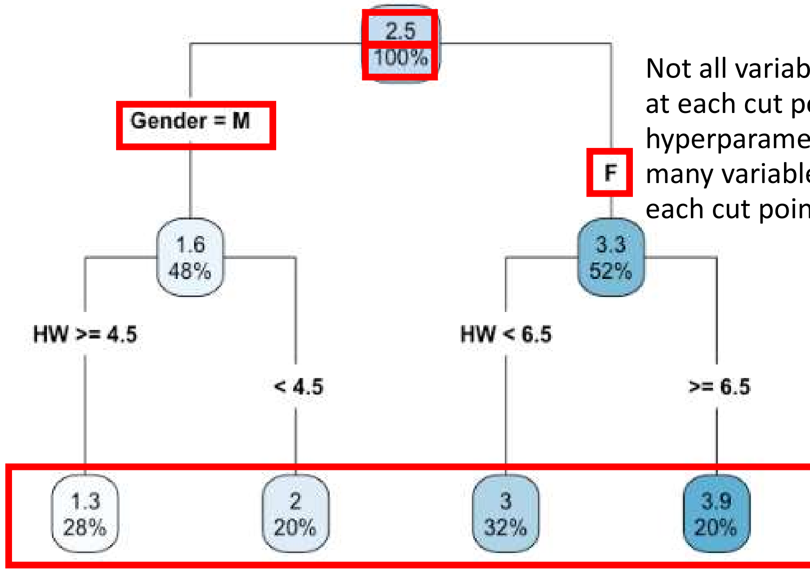
Continuous Variable Transformations

- “The goal of all of these approaches is to convert the existing continuous predictors into a form that can be utilized by **any** model and presents the most useful information to the model.”
- In my opinion (Stephen’s), the hardest model to mess up and the one where transformations are least likely to help is a Random Forest. Use RFs as a baseline.

Decision Tree to Random Forest

A Random Forest is made up of a lot of decision trees

Bootstrapped sample



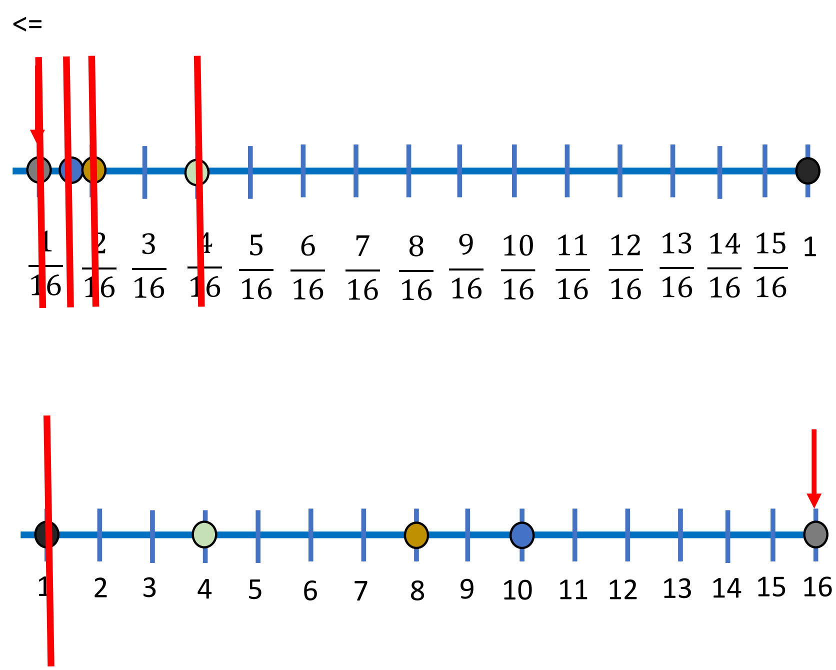
Not all variables are considered at each cut point. A hyperparameter of RFs is how many variables to consider at each cut point.

Continuous Variable Transformations

- "The goal of all of these approaches is to convert the existing continuous predictors into a form that can be utilized by **any** model and presents the most useful information to the model."
- In my opinion (Stephen's), the hardest model to mess up and the one where transformations are least likely to help is a Random Forest. Use RFs as a baseline.
  - RFs use cut points so scaling the data is not necessary and will not yield better results.
  - RFs are insensitive to outliers/extreme values.

1:1 Transformations & Random Forests

Example: Inverse Transformation



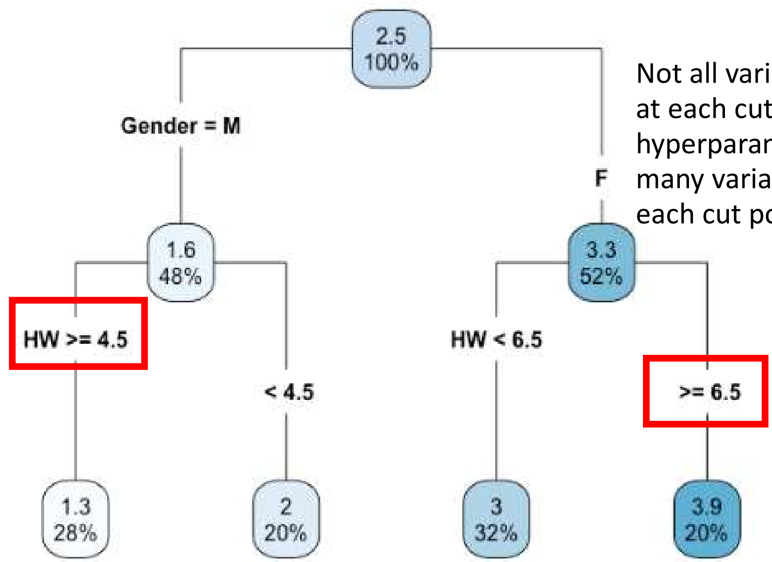
Continuous Variable Transformations

- "The goal of all of these approaches is to convert the existing continuous predictors into a form that can be utilized by **any** model and presents the most useful information to the model."
- In my opinion (Stephen's), the hardest model to mess up and the one where transformations are least likely to help is a Random Forest. Use RFs as a baseline.
  - RFs use cut points so scaling the data is not necessary and will not yield better results.
  - RFs are insensitive to outliers/extreme values.
  - RFs can pick up on interactions even if you don't specify these interactions before building the model.

Decision Tree to Random Forest

A Random Fest is made up of a lot of decision trees

Bootstrapped sample

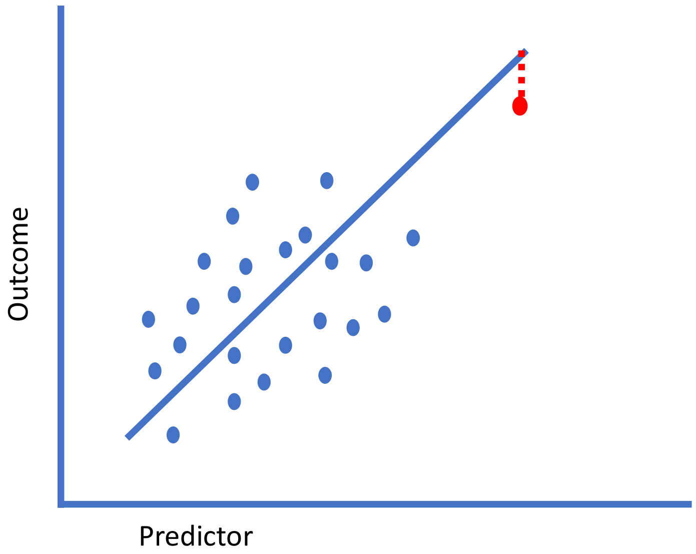


Not all variables are considered at each cut point. A hyperparameter of RFs is how many variables to consider at each cut point.

Continuous Variable Transformations

- "The goal of all of these approaches is to convert the existing continuous predictors into a form that can be utilized by **any** model and presents the most useful information to the model."
- In my opinion (Stephen's), the hardest model to mess up and the one where transformations are least likely to help is a Random Forest. Use RFs as a baseline.
  - RFs use cut points so scaling the data is not necessary and will not yield better results.
  - RFs are insensitive to outliers/extreme values.
  - RFs can pick up on interactions even if you don't specify these interactions before building the model.
  - RFs can handle lots of predictor variables (Chapter 10)
  - RFs don't have assumptions about the data (other than that it is representative) or residuals
  - RFs can handle missing data (Chapter 8)
  - *RFs can't extrapolate*
  - *RFs are not good for time series data*

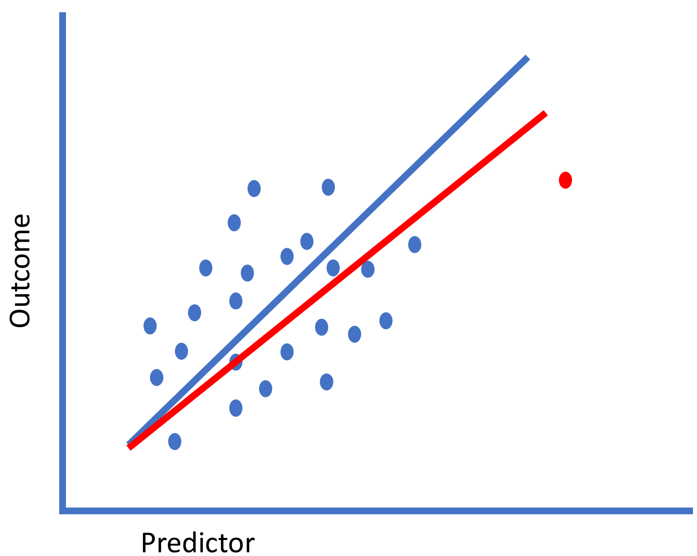
Extrapolation

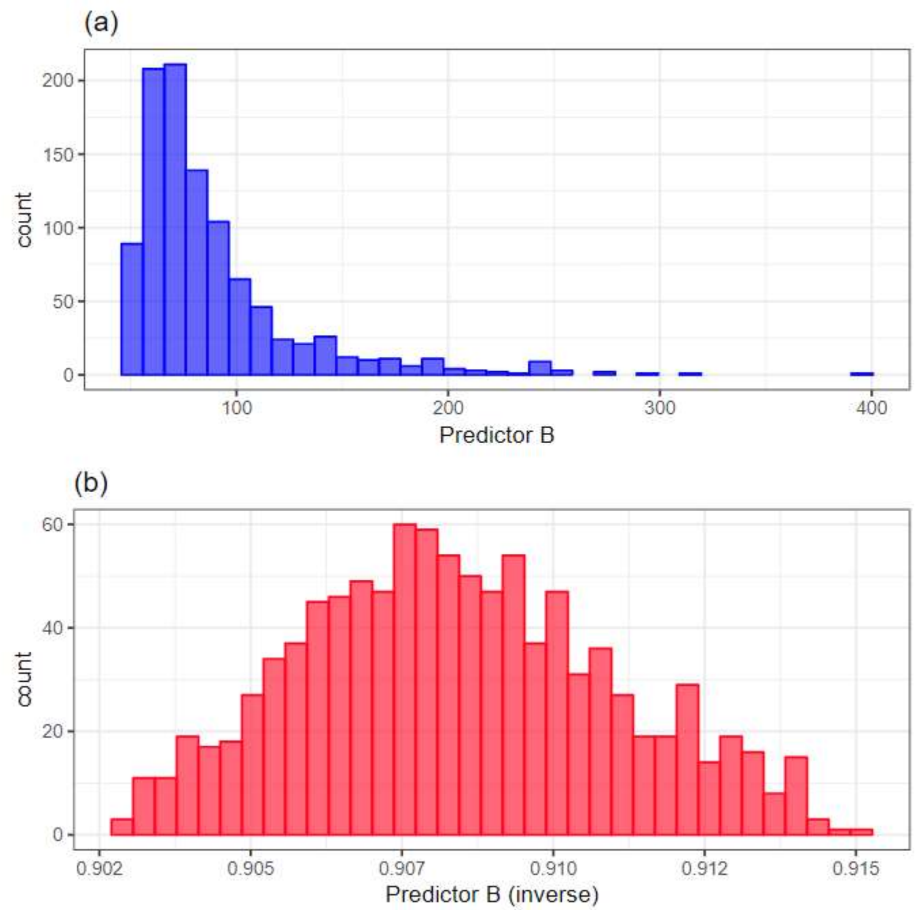


1:1

Box-Cox Transformation

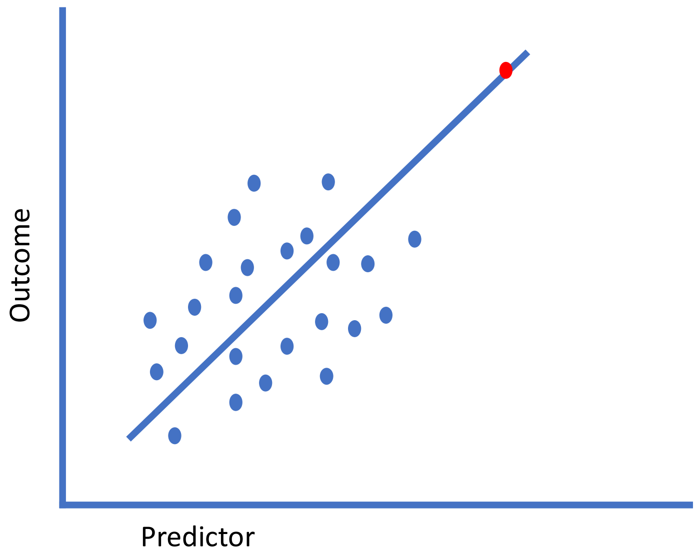
Goal: Make data normalish

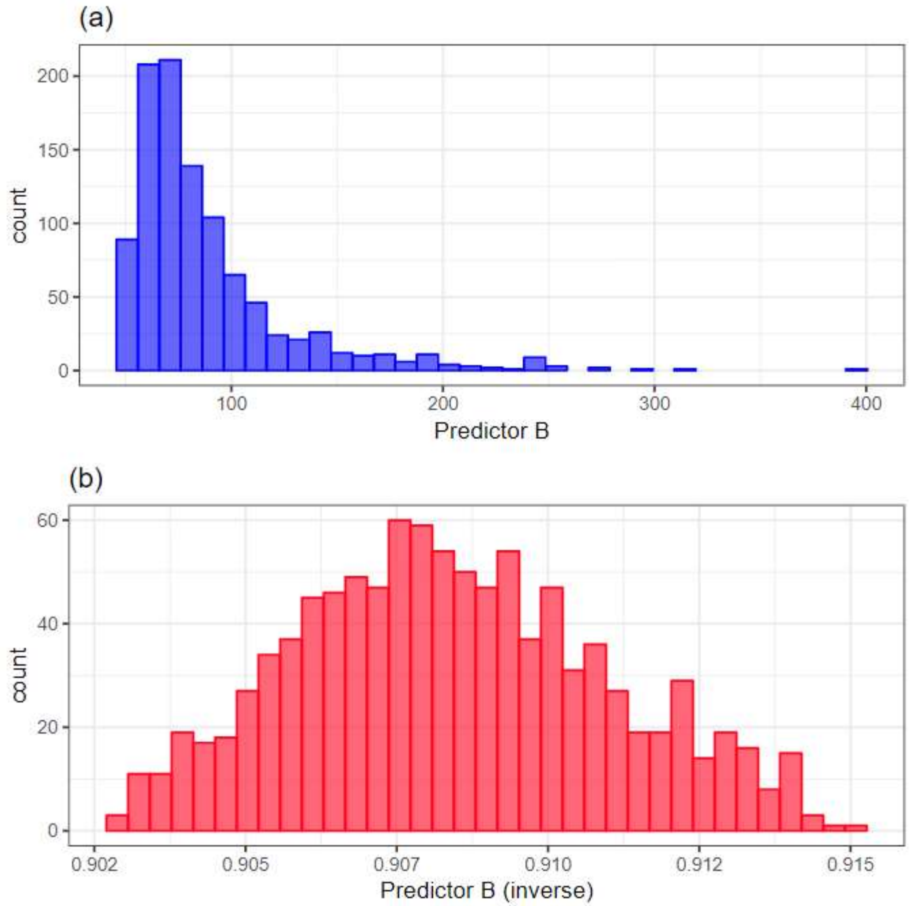




Box-Cox Transformation

Goal: Make data normalish





Centering a Variable

All variables that are centered will have a mean of 0

$$x_i - \bar{x}$$

Scaling a Variable

All variables that are scaled will have a mean of 0 and a standard deviation of 1

Needed when using K-nearest neighbors or lasso/ridge regression.

$$\frac{x_i - \bar{x}}{s_x}$$

“Again, it is emphasized that the statistics required for the transformation (e.g., the mean) are estimated from the training set and are applied to all data sets (e.g., the test set or new samples).”

Smoothing Time Series Data

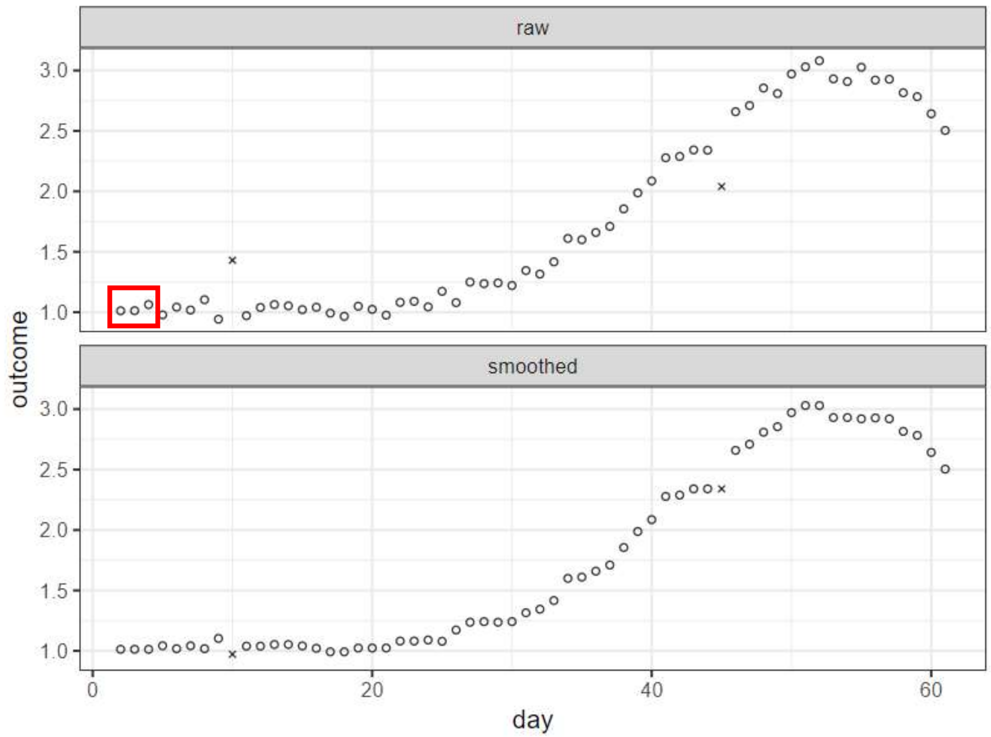
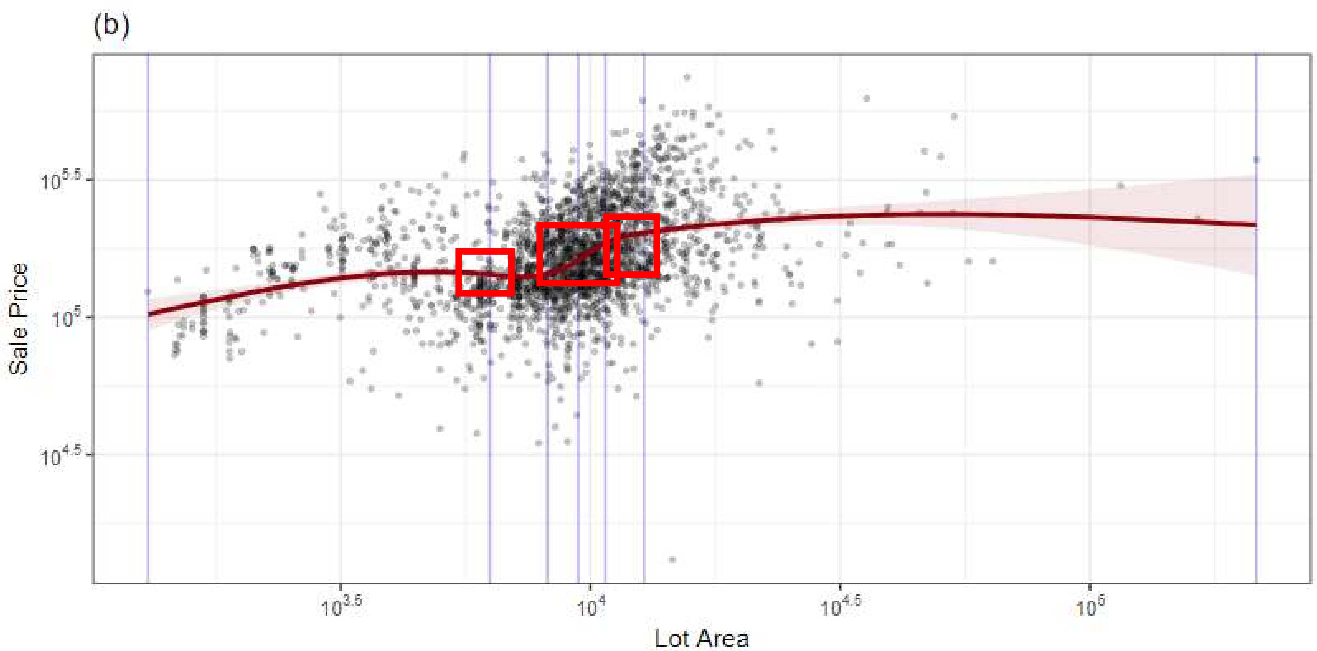


Figure 6.2: A sequence of outcome values over time. The raw data contain outliers on days 10 and 45. The smoothed values are the result of a 3-point running median.

1:Many

Natural Cubic Splines



Discretizing Continuous Variables

AKA Binning

- Pros
  - Simplify analysis
  - Avoid specifying relationship between predictor and outcome
- Cons
  - Unlikely that the actual trend is found with model
  - Removes some nuance to data
  - No objective cut-points
  - If there is no relationship between outcome and predictor, there is an increase in probability that and erroneous trend will be found

Many:Many

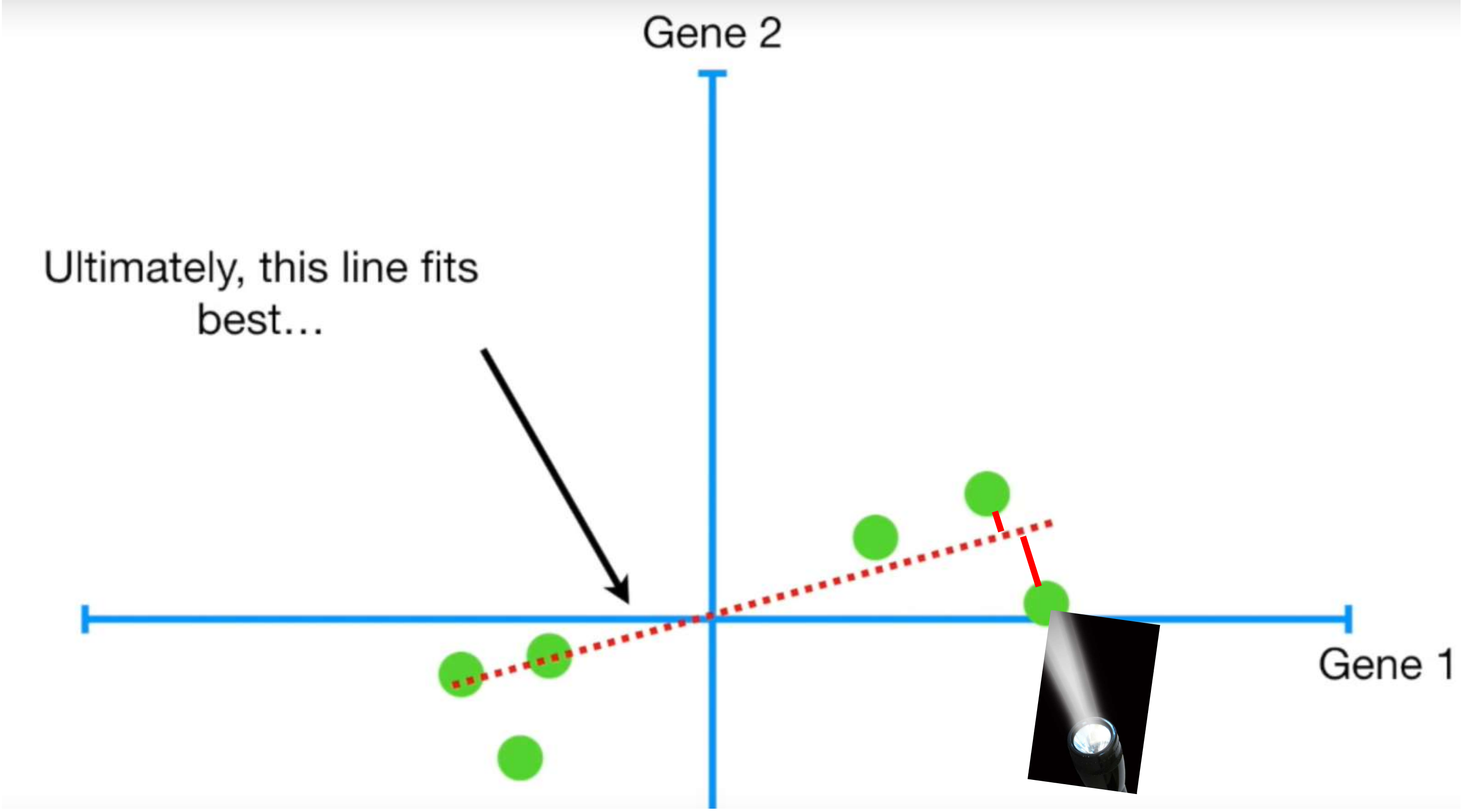
Principle Component Analysis (PCA)

- "The objective of PCA is to find linear combinations of the original predictors such that the combinations summarize the maximal amount of variation in the original predictor space."
- "An important side benefit of this technique is that the resulting PCA scores are uncorrelated."
- Is NOT variable selection

PCA

https://www.youtube.com/watch?v=FgakZw6K1QQ&vl=en

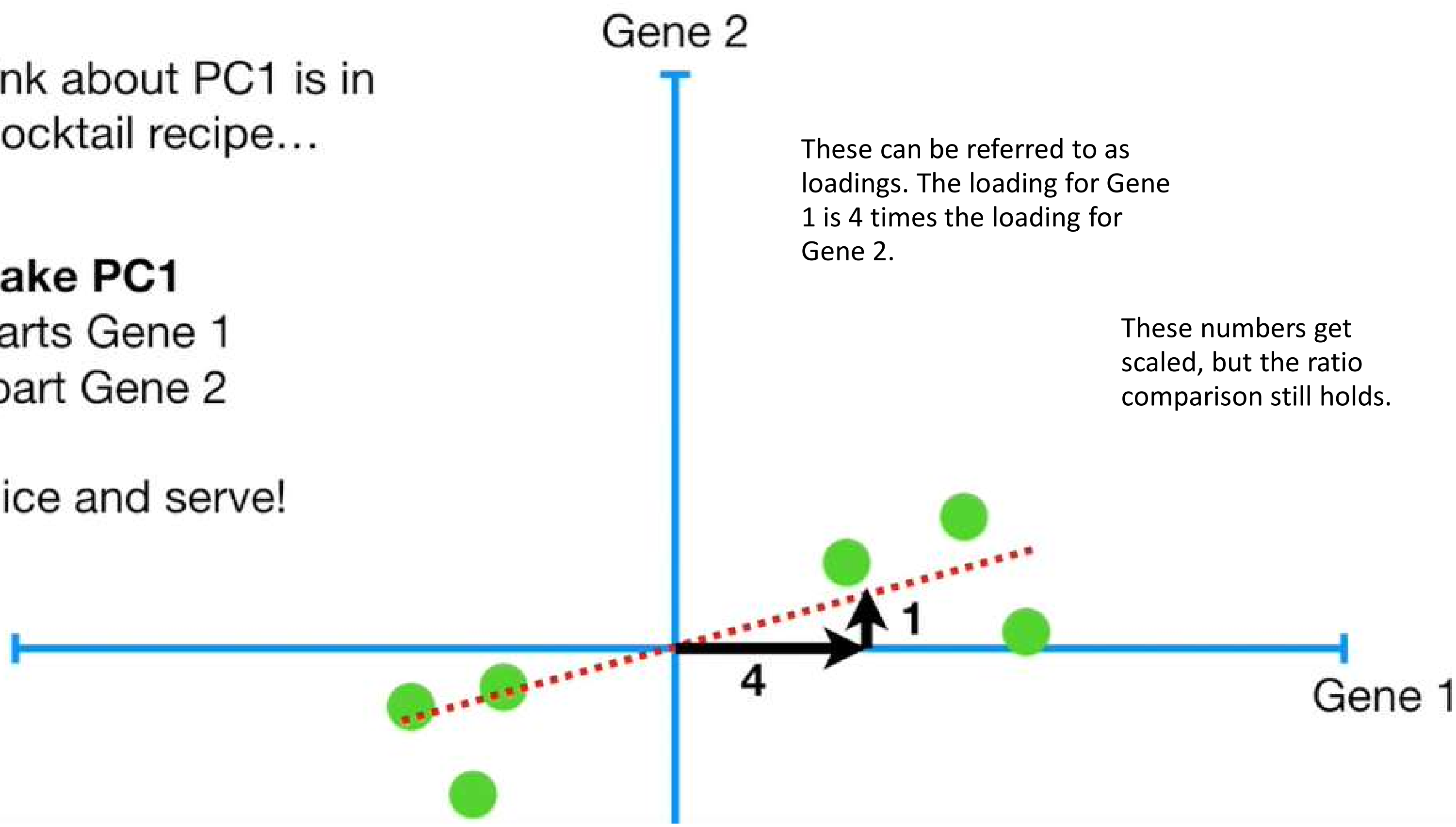
Goal: Retain the most information while reducing the number of dimensions



One way to think about PC1 is in terms of a cocktail recipe…

To make PC1
Mix **4** parts Gene 1
with **1** part Gene 2

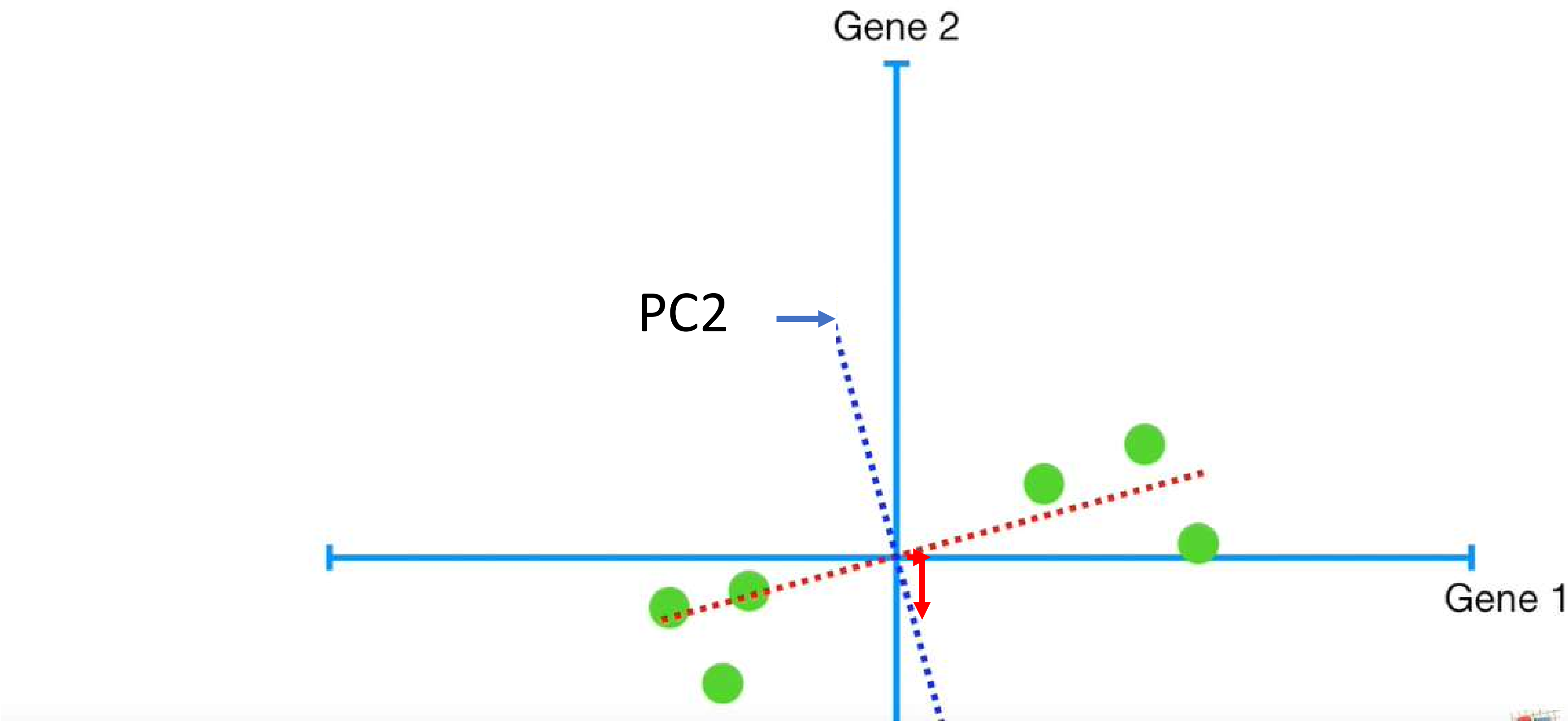
Pour over ice and serve!



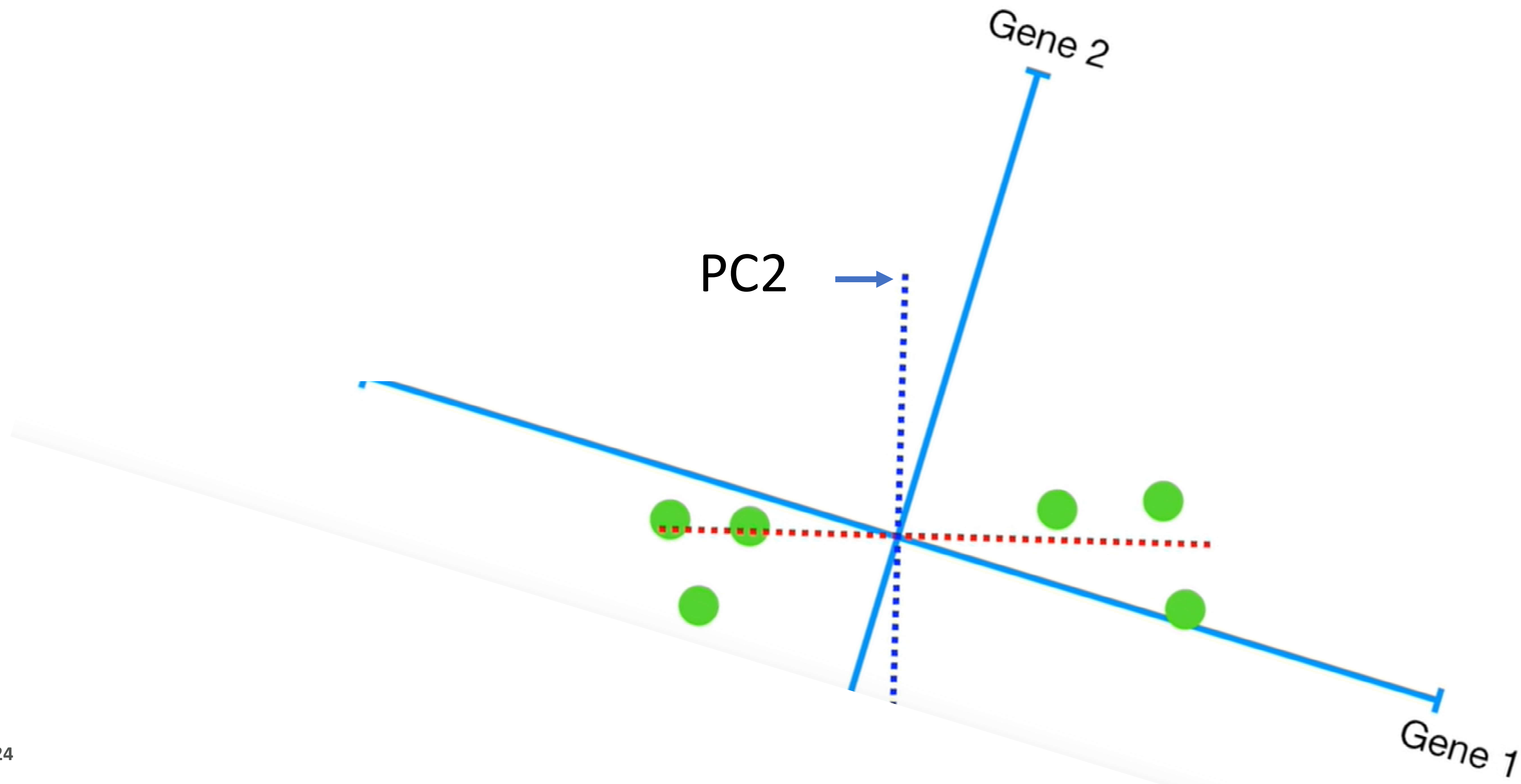
These can be referred to as loadings. The loading for Gene 1 is 4 times the loading for Gene 2.

These numbers get scaled, but the ratio comparison still holds.

PC2 is Perpendicular (Orthogonal) to PC1

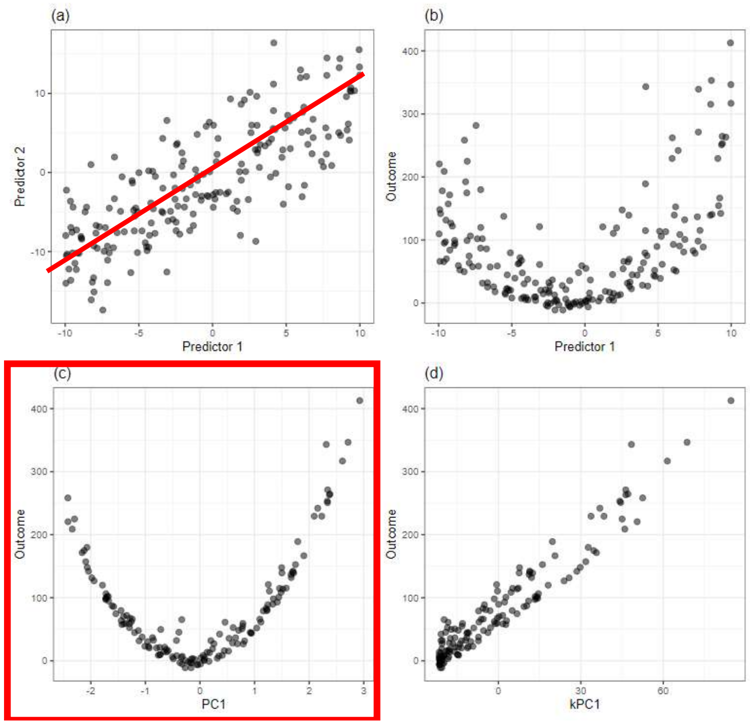


PC2 is Perpendicular (Orthogonal) to PC1



Kernel PCA

- Use Kernel PCA when the x variables are not linearly related to outcome variable.
- There are different kernels for di
  - Polynomial
  - Radial
  - Etc.



Linear Model

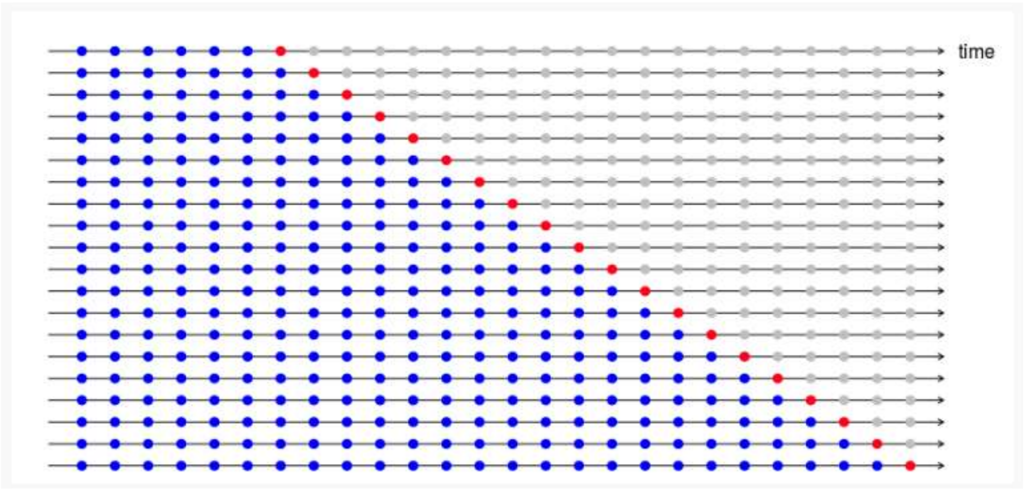
$$y = B_0 + B_1 x + \varepsilon$$

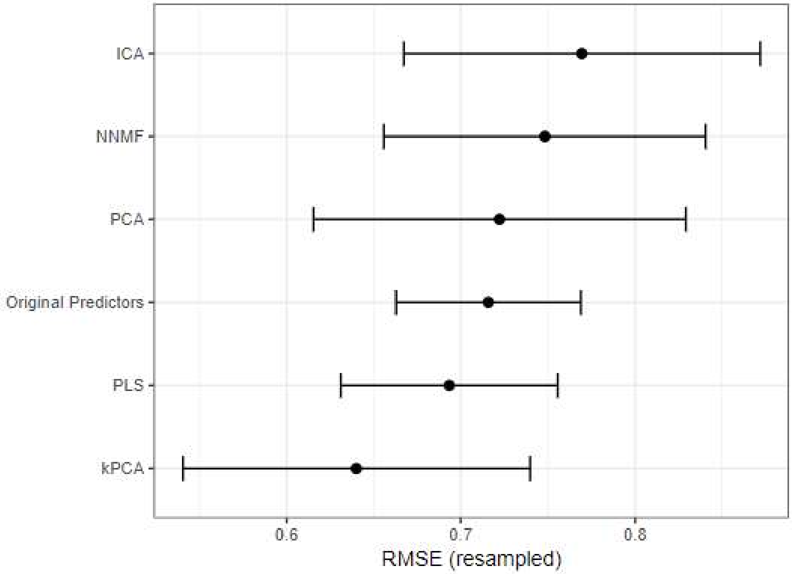
Other Unsupervised Methods

- Independent Component Analysis
  - When relationship is not linear
- Non-Negative Matrix Factorization
  - “Find the best set of coefficients that make the scores as “close” as possible to the original data with the constraint of non-negativity”
  - Features must be positive

Partial Least Squares

- Supervised version of PCA where components are optimally related to response variable
- Each component (dimension) is uncorrelated like PCA
- PLS can perhaps reduce dimensions better than PCA, but an assessment dataset must be used.

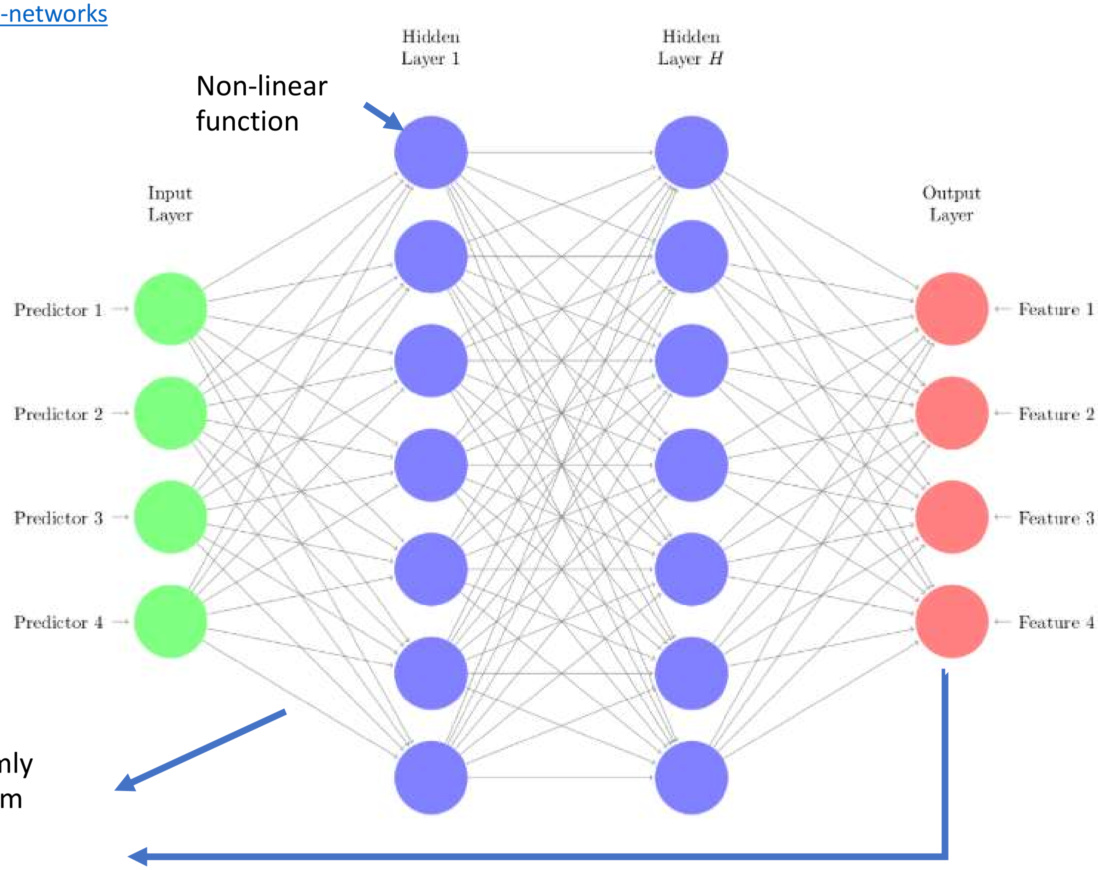




Autoencoders

https://www.3blue1brown.com/neural-networks

Non-linear function



Weights that are initially randomly assigned and then improved from information in the output layer

Other Techniques

- Spatial Sign
  - Make data into a circle
- Distance and Depth Features
  - Classification