Understanding Feature Selection



Janani Ravi CO-FOUNDER, LOONYCORN www.loonycorn.com

Overview

Curse of dimensionality

Reducing complexity of data

Understanding feature selection

Filter methods

Embedded methods

Wrapper methods

Problems with Data

Insufficient data

Too much data

Non-representative data

Missing data

Duplicate data

Outliers

Too Much Data

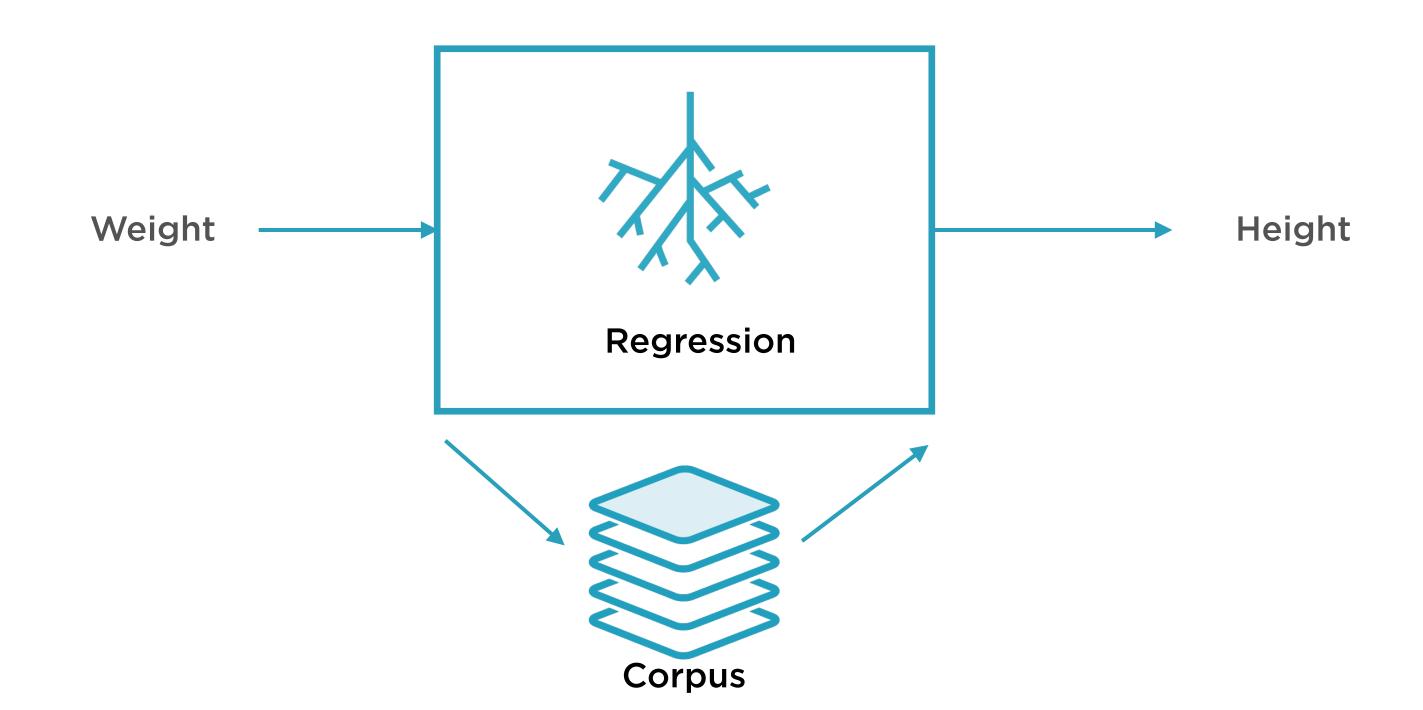


Data might be excessive in two ways

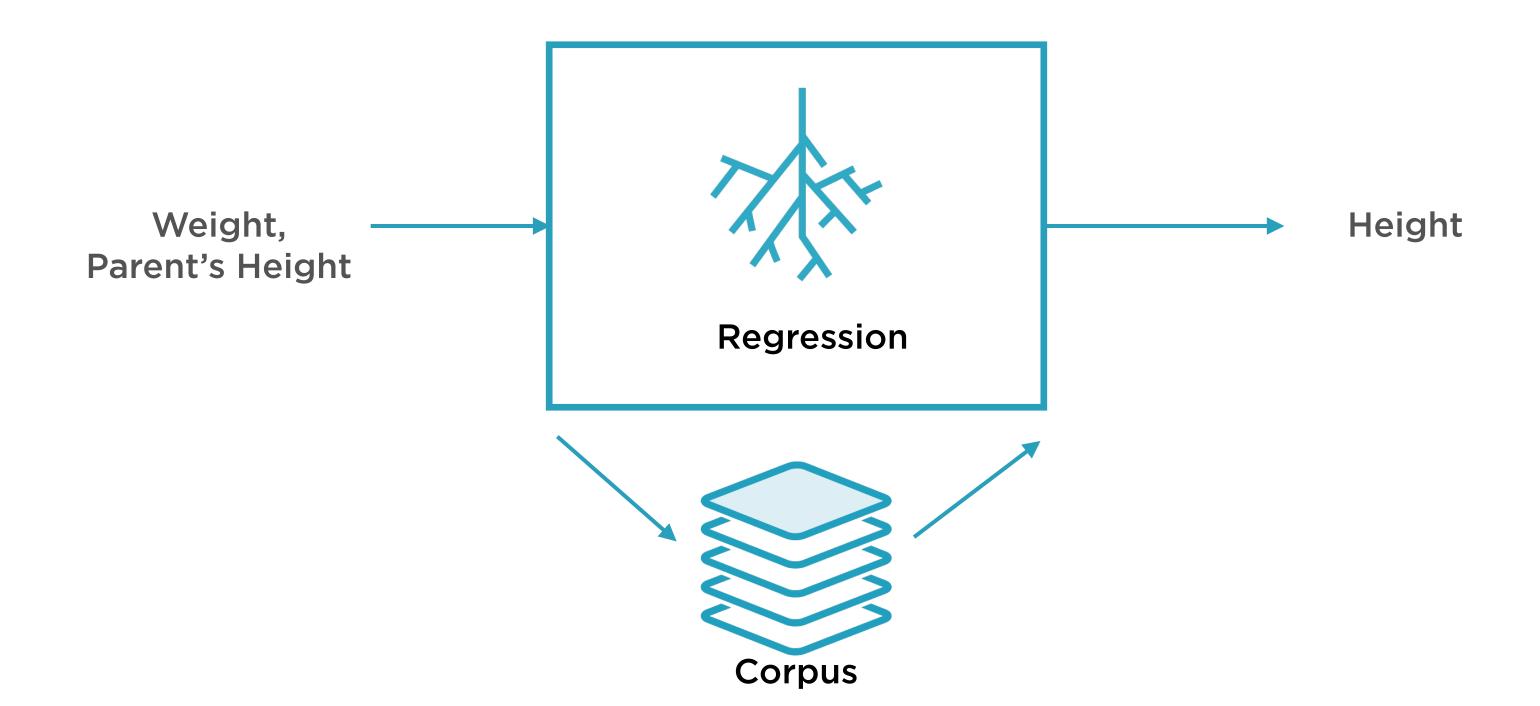
- Curse of dimensionality: Too many columns
- Outdated historical data: Too many rows

The Curse of Dimensionality

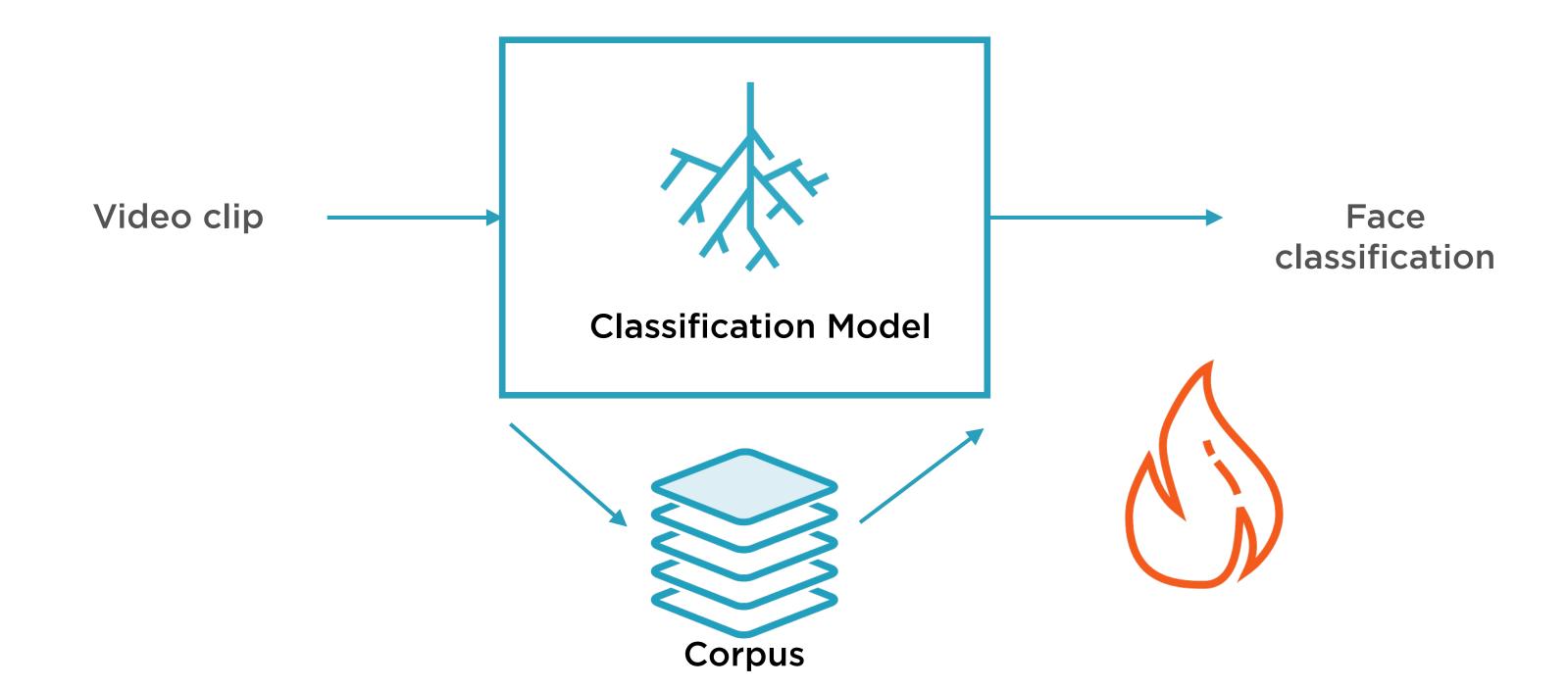
One X Variable



Two X Variables



Dimensionality Explosion



Curse of Dimensionality: As number of **x** variables grows, several problems arise

Curse of Dimensionality

Problems in Visualization

Problems in Training

Problems in **Prediction**

Curse of Dimensionality

Problems in Visualization

Problems in Training

Problems in **Prediction**

Problems in Visualization



Exploratory Data Analysis (EDA) is an essential precursor to model building

Essential for

- Identifying outliers
- Detecting anomalies
- Choosing functional form of relationships

Problems in Visualization



Two dimensional visualizations are powerful aids in EDA

Even three-dimensional data is hard to meaningfully visualize

Higher dimensional data is often imperfectly explored prior to ML

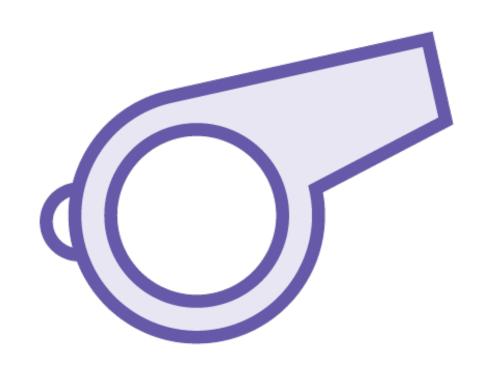
Curse of Dimensionality

Problems in Visualization

Problems in Training

Problems in **Prediction**

Problems in Training

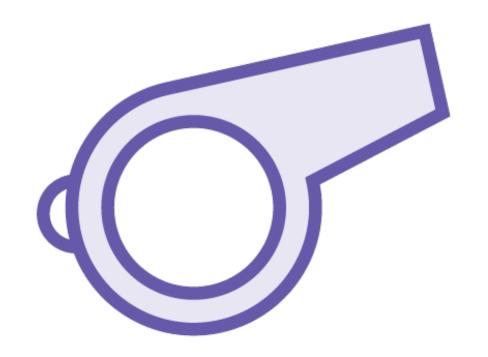


Training is the process of finding best model parameters

Complex models have thousands of parameter values

Training for too little time leads to bad models

Problems in Training



Number of parameters to be found grows rapidly with dimensionality

Extremely time-consuming

For on-cloud training, also extremely expensive

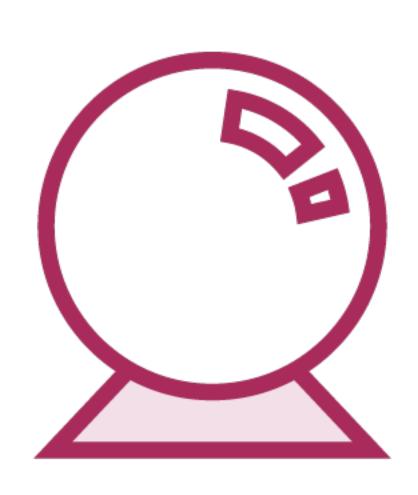
Curse of Dimensionality

Problems in Visualization

Problems in Training

Problems in Prediction

Problems in Prediction



Prediction involves finding training instances similar to test instance

As dimensionality grows, size of search space explodes

Higher the number of X variables, higher the risk of overfitting

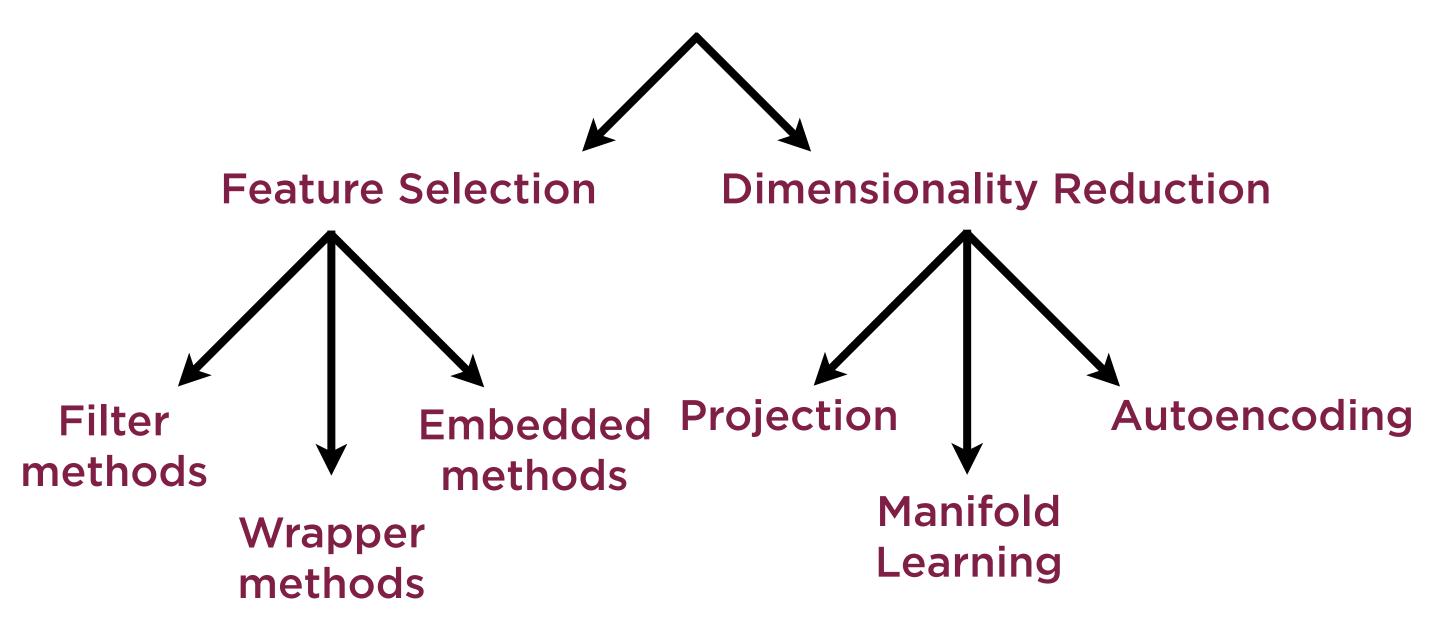
Curse of Dimensionality

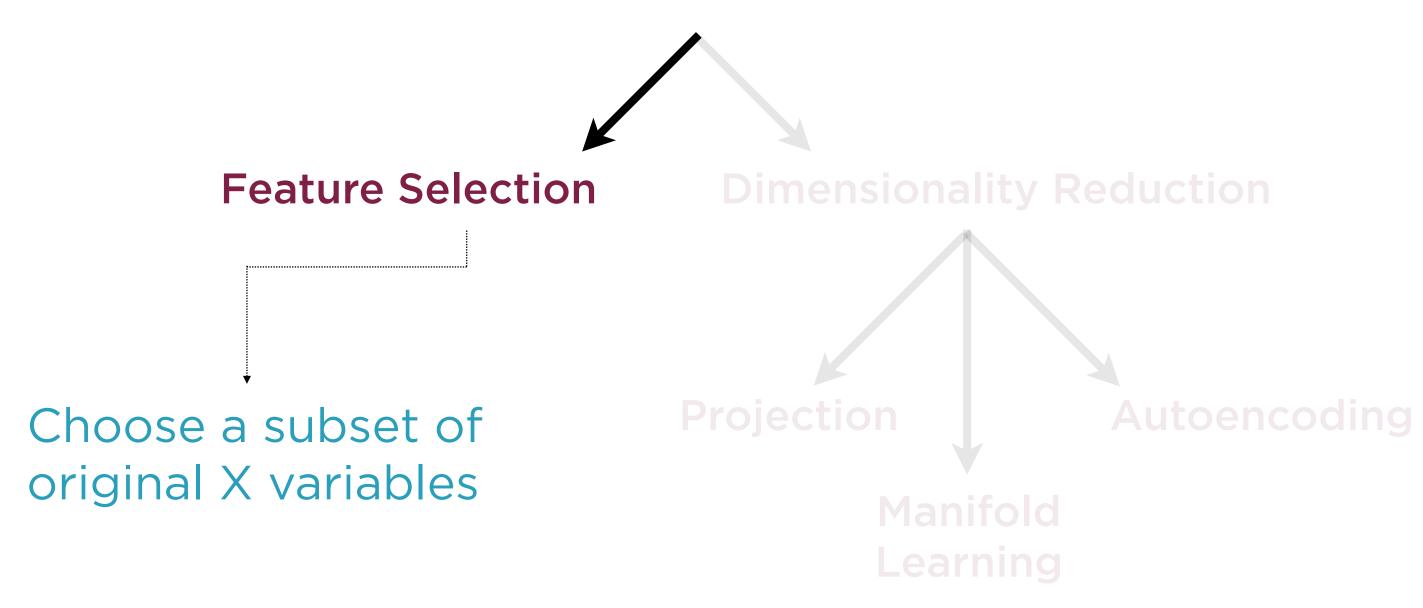


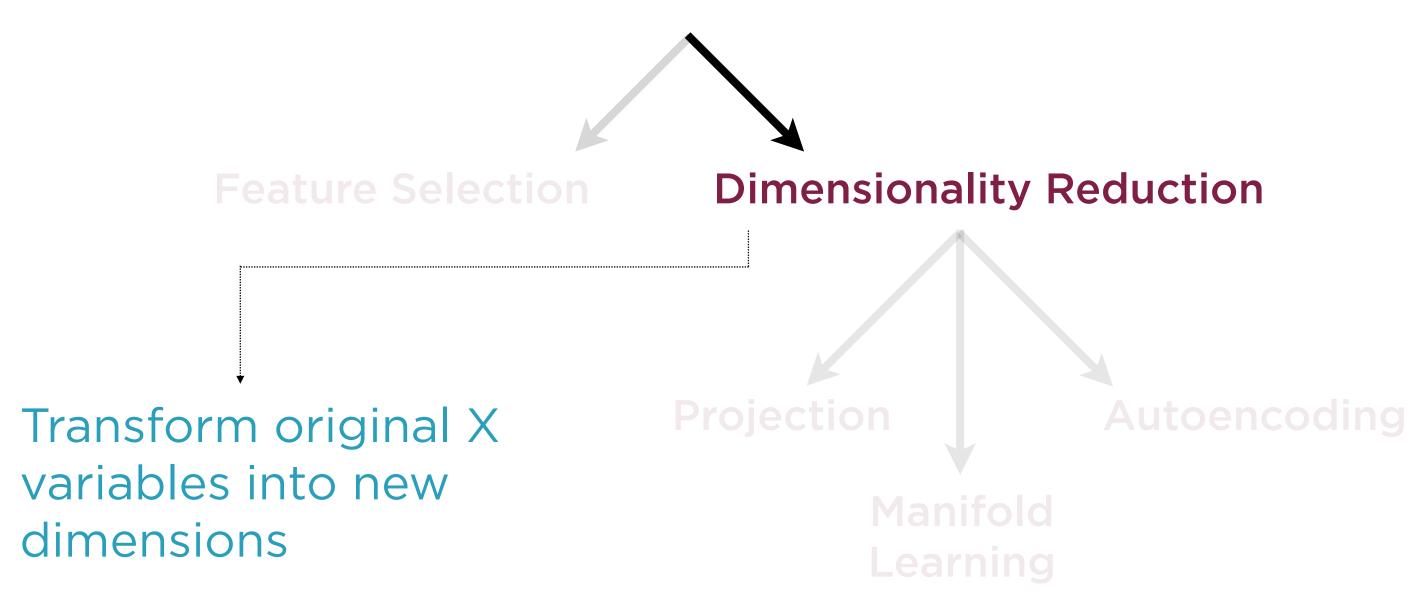
Easier problems to solve

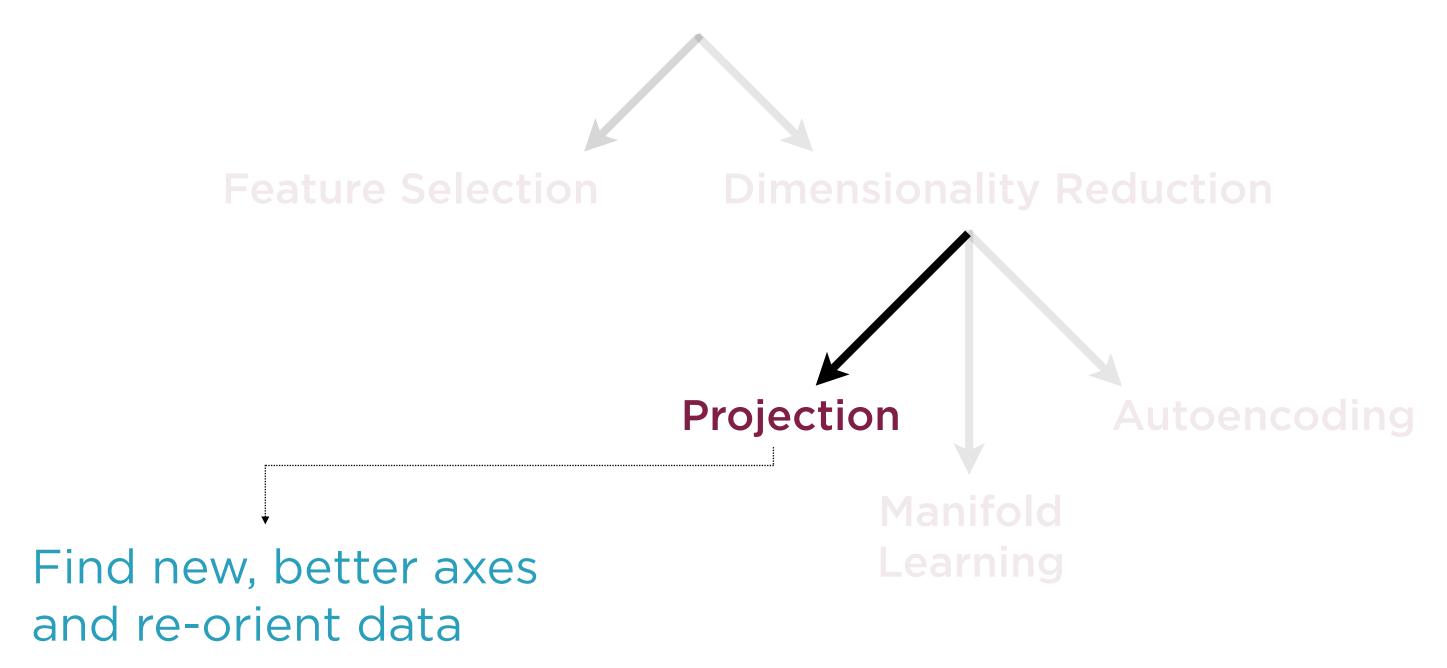
- Feature selection: Deciding which data is actually relevant
- Feature engineering: Aggregating very low-level data into useful features
- Dimensionality Reduction: Reduce complexity without losing information

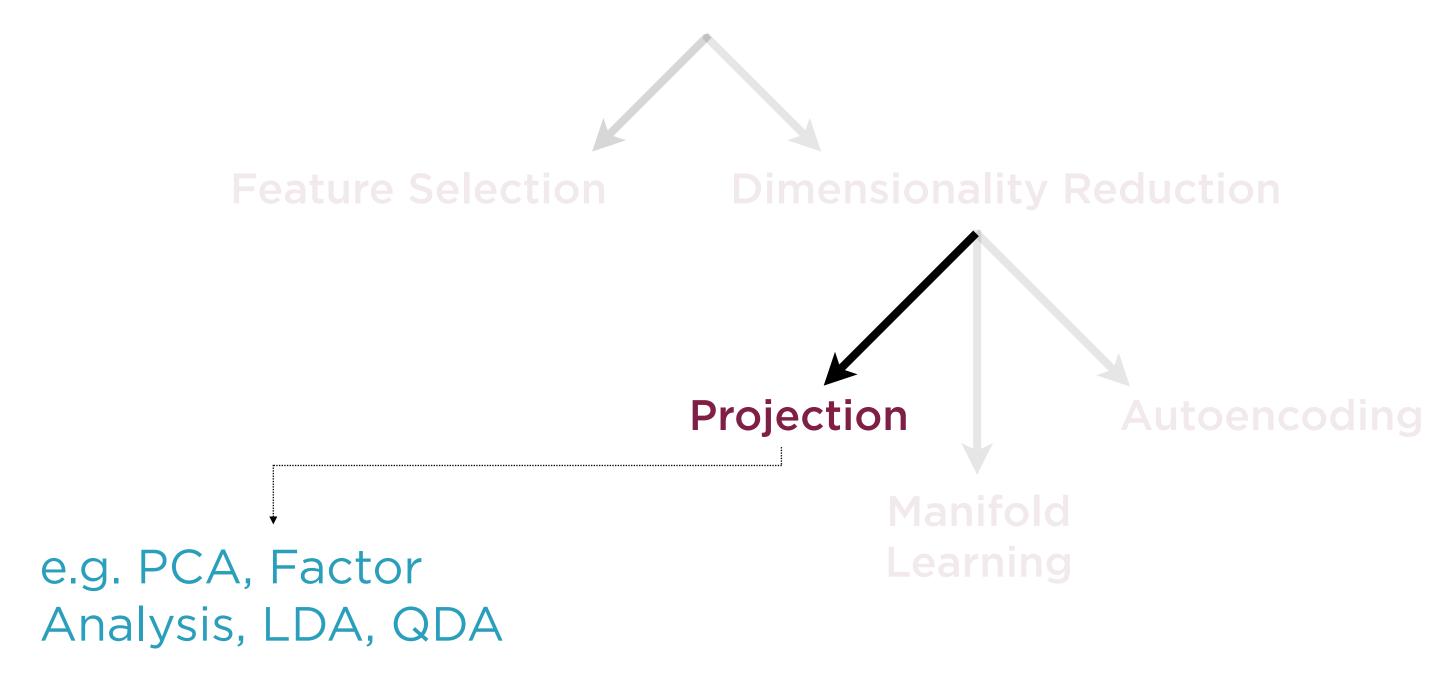
Solutions for Reducing Complexity

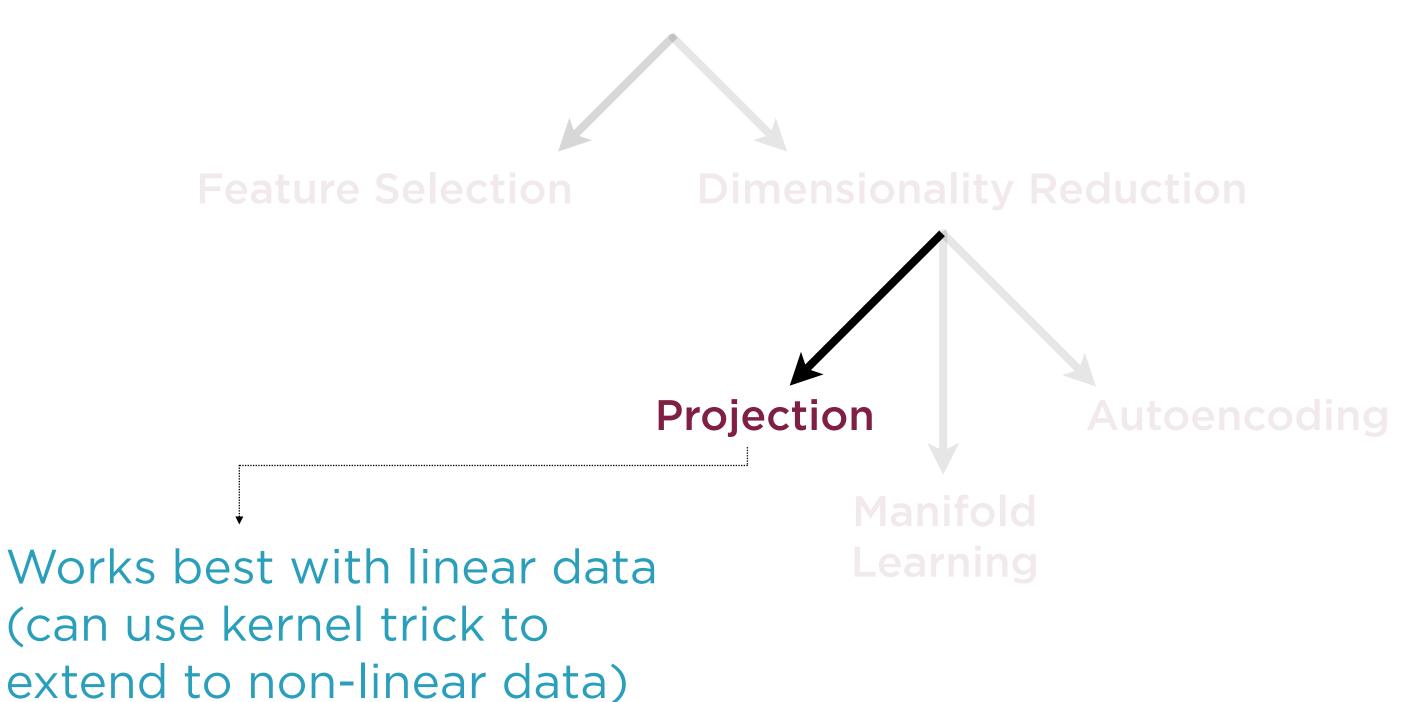


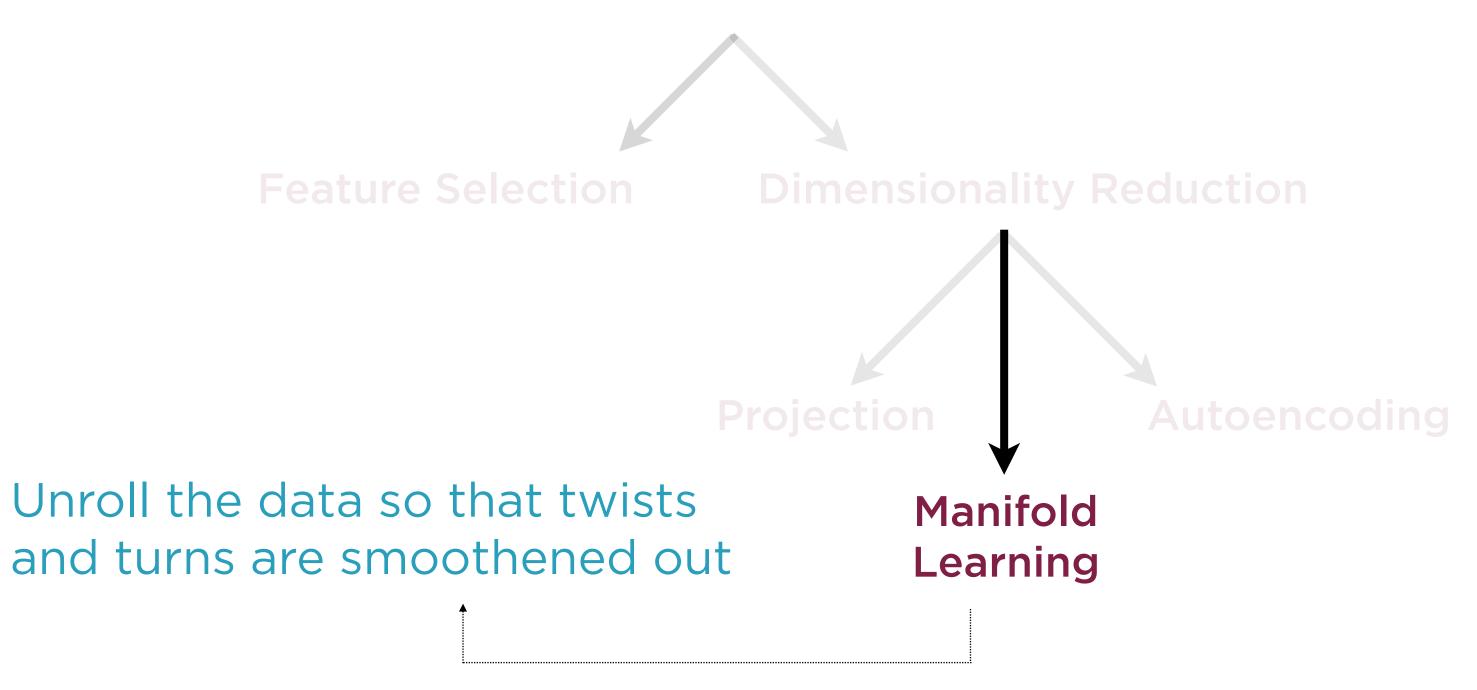


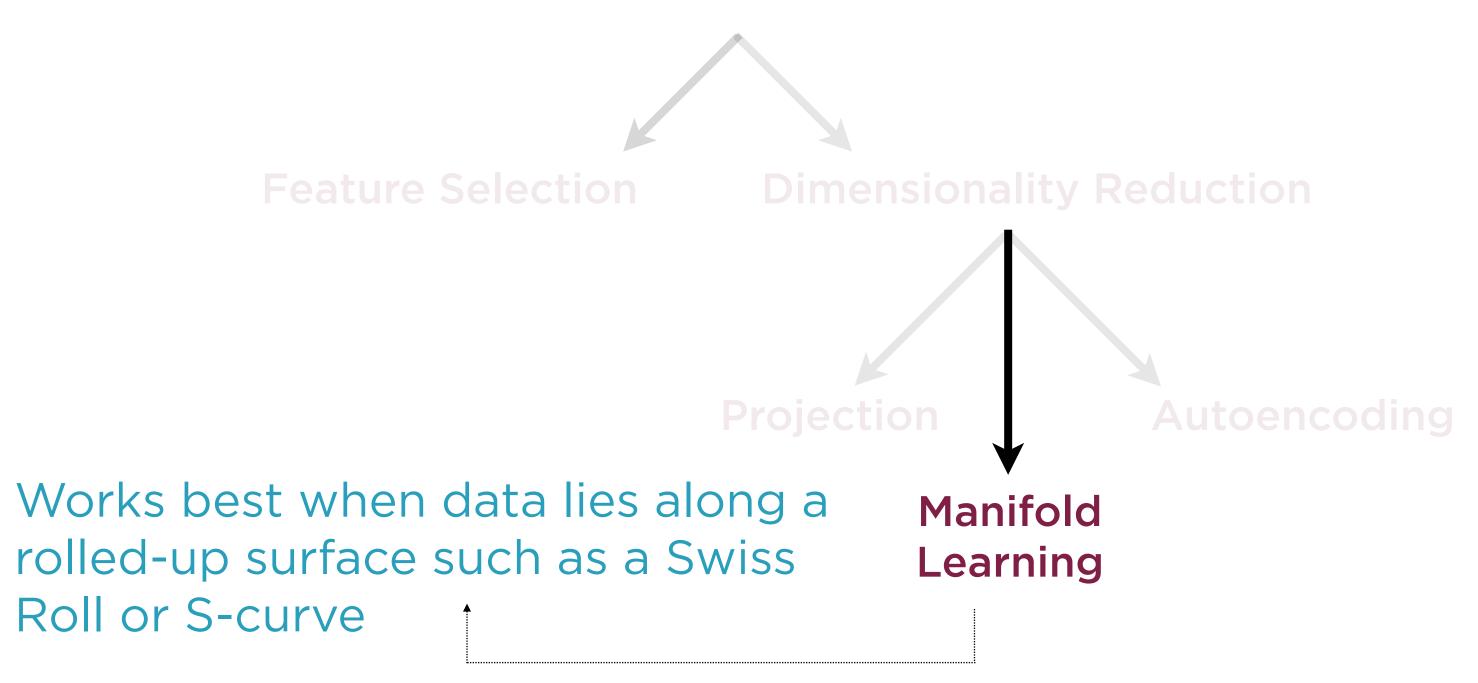


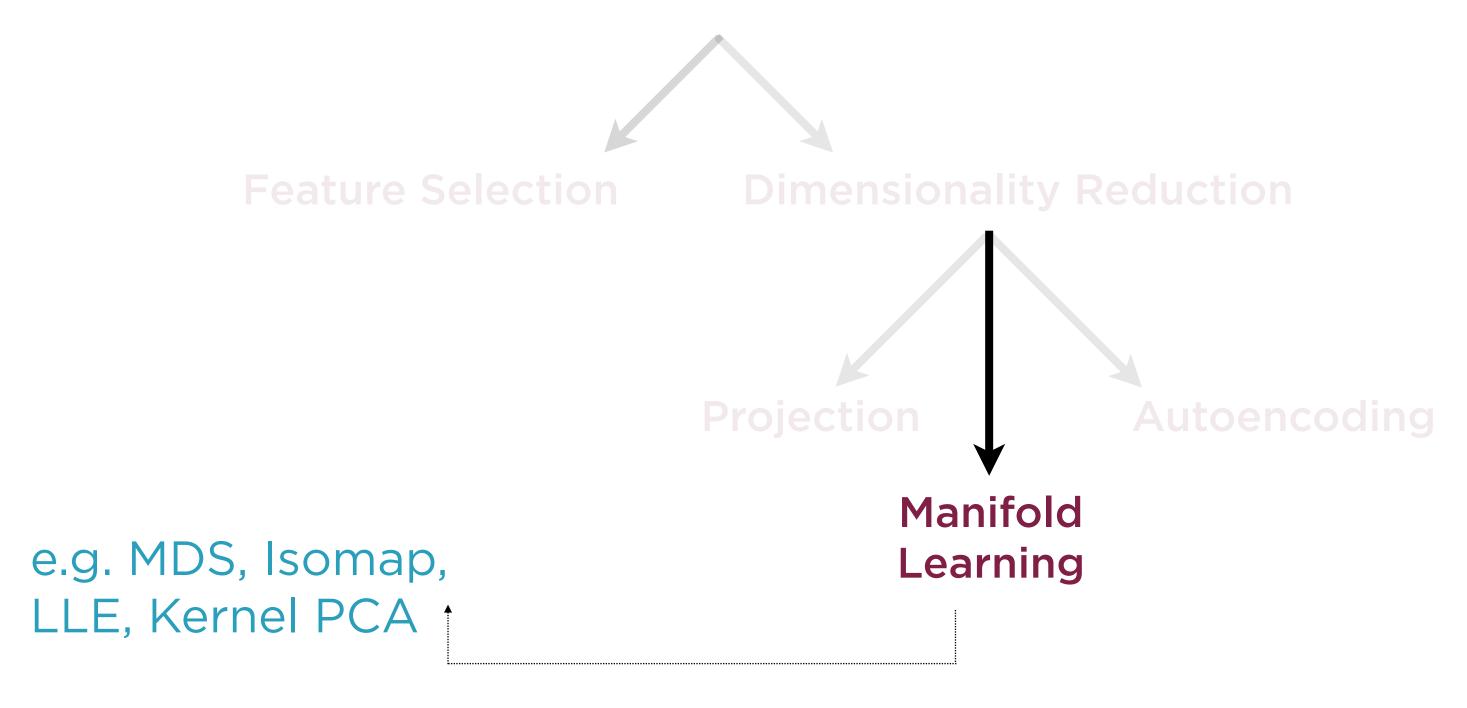


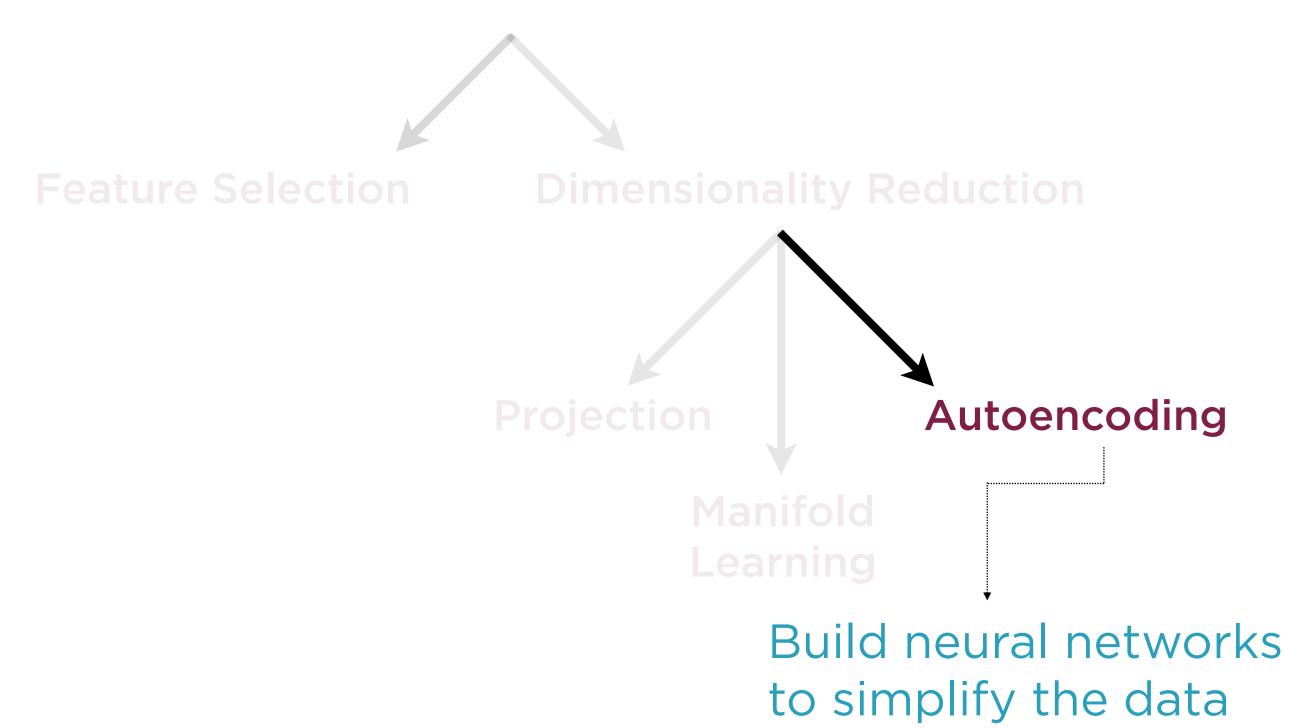


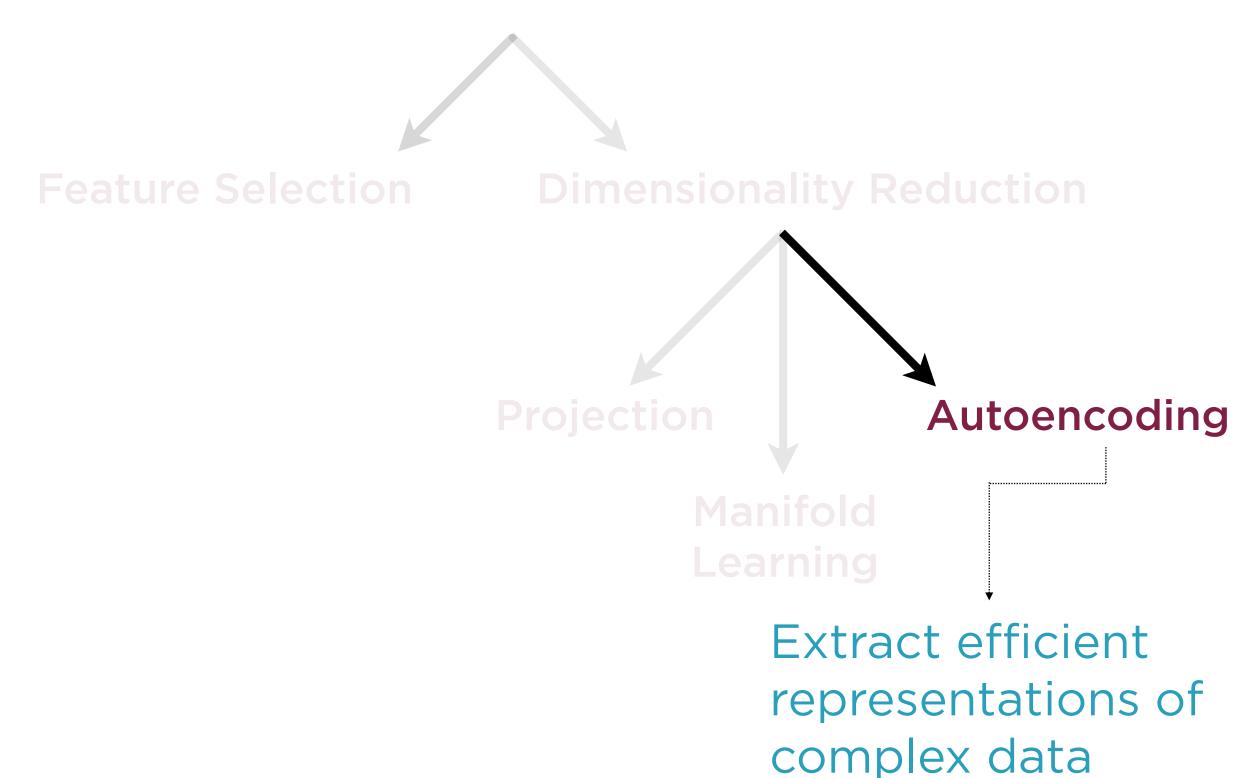












Drawbacks of Reducing Complexity

Loss of information

Performance degradation

Computational intensive

Complex pipelines

Transformed features hard to interpret

Choosing Feature Selection

Use Case

Possible Solution

Many X-variables

Most of which contain little information

Some of which are very meaningful

Meaningful variables are independent of each other

Feature selection

Choosing PCA and Factor Analysis

Use Case

Large number of X-variables

Most of which are meaningful

Highly correlated to each other

Linearly related to each other

For use in regression

Possible Solution

Principal Components Analysis (PCA) or Factor Analysis

Choosing PCA and Factor Analysis

Use Case

Large number of X-variables

Most of which are meaningful

Highly correlated to each other

Linearly related to each other

For use in classification

Possible Solution

Linear Discriminant Analysis (LDA) or Dictionary Learning

Choosing Manifold Learning

Use Case

Y not linearly related to X

Very high dimensionality of X (e.g. pixel counts in image data)

Many constraints on allowable values of X-variables (sparse features)

Three-dimensional plots of Y against pairs of X indicate manifold shape

Possible Solution

Manifold learning

Choosing Autoencoders

Use Case

Extremely complex feature vectors

Images, video, documents

Pre-processing before using in neural networks

Possible Solution

Autoencoders

Feature Selection

Choosing Feature Selection

Use Case

Possible Solution

Many X-variables

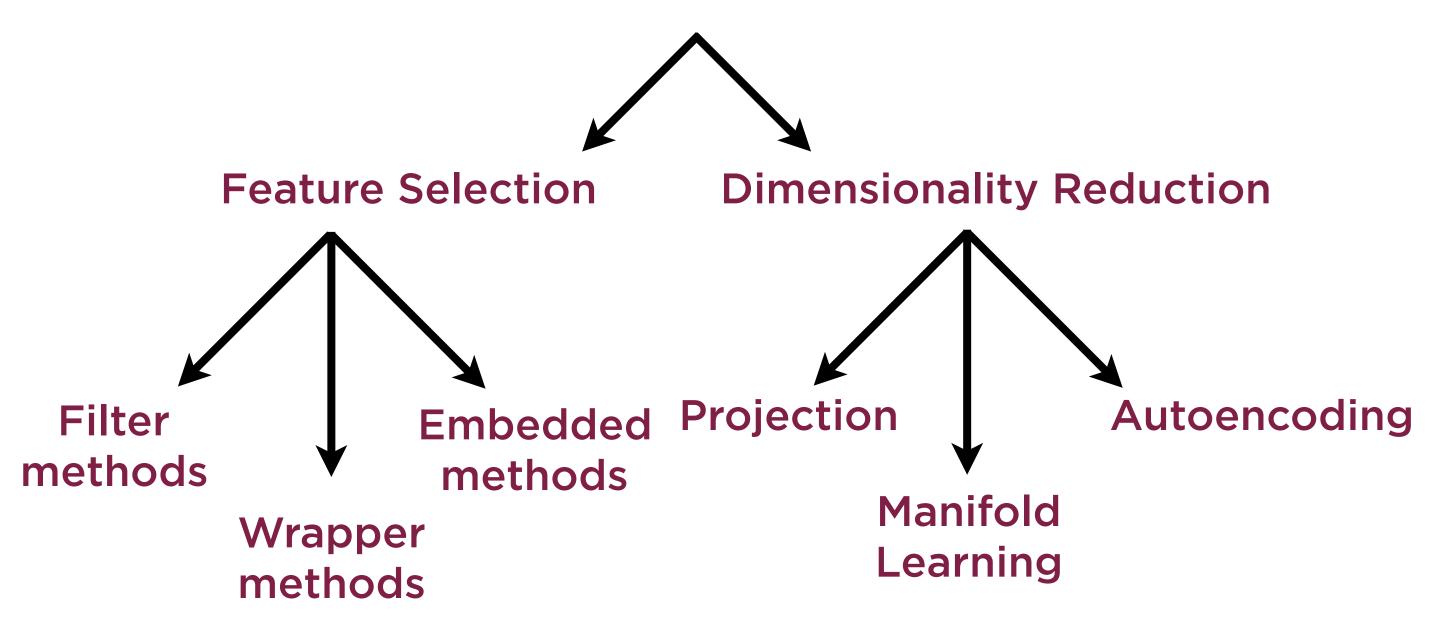
Most of which contain little information

Some of which are very meaningful

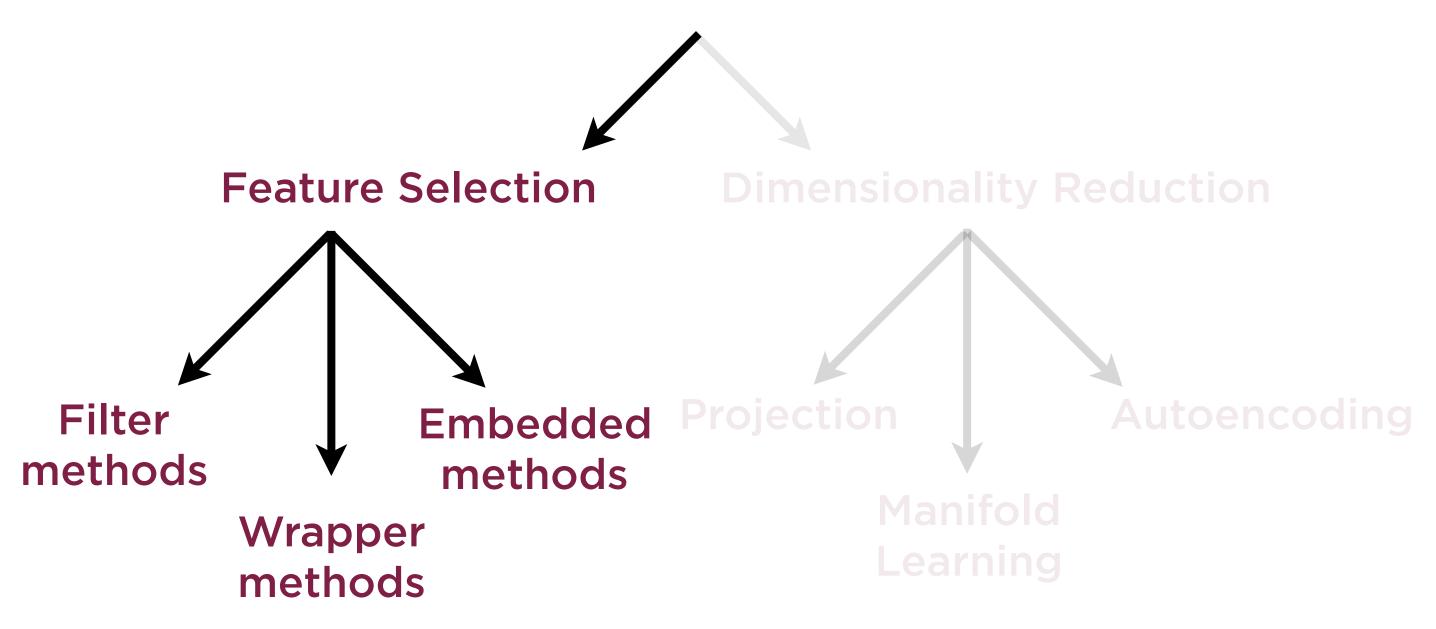
Meaningful variables are independent of each other

Feature selection

Reducing Complexity



Reducing Complexity



Filter Methods

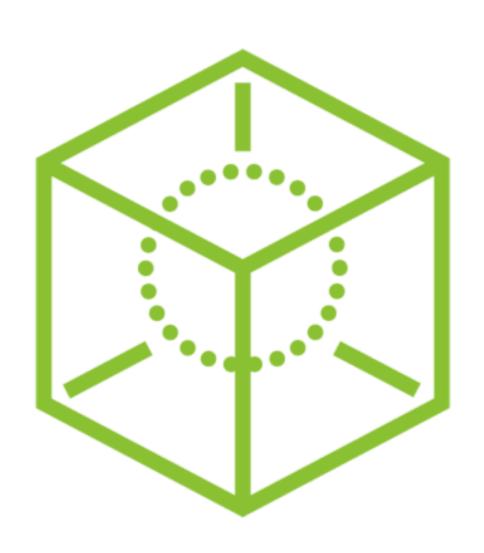


Features (columns) selected independently of choice of model

Rely on statistical properties of features

Either individually (univariate) or jointly (multi-variate)

Embedded Methods



Features (columns) selected during model training

Feature selection effectively embedded within modeling

Only specific types of models perform feature selection

Wrapper Methods



Somewhere between filter and embedded feature selection

Features are chosen by building different candidate models

Forward and backward stepwise regression are examples

Wrapper Methods



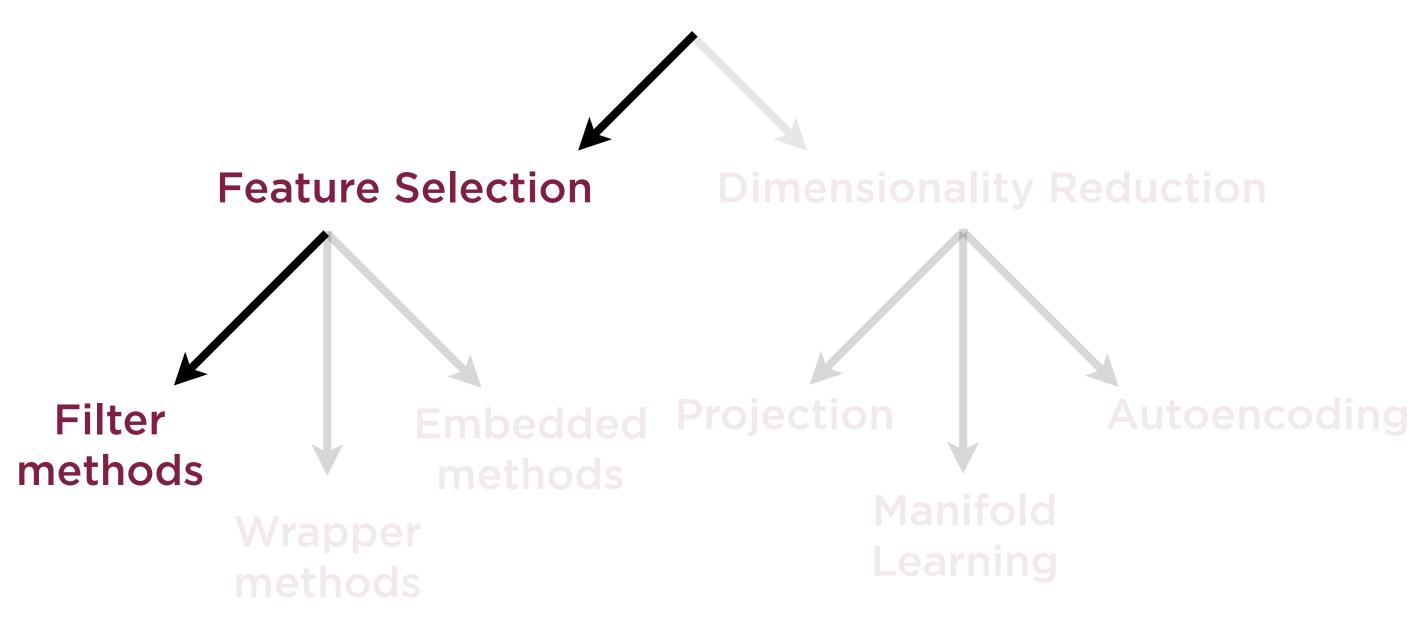
Each candidate model has different subset of features

However all candidate models are similar in structure

Features may be added or dropped to see whether the model improves

Filter Methods for Feature Selection

Reducing Complexity



Filter Methods



Features (columns) selected independently of choice of model

Rely on statistical properties of features

Either individually (univariate) or jointly (multi-variate)

Hypothesis

Proposed explanation for a phenomenon.

Lady Tasting Tea



Lady tasting tea: famous experiment
Was tea added before or after milk?
Muriel Bristol claimed she could tell

Lady Tasting Tea

Null Hypothesis
(H₀)

Alternate Hypothesis
(H₁)

The lady cannot tell if milk was poured first

The lady can tell if milk was poured first

Statistical Techniques

Variance Thresholding

Chi-square Test

ANOVA

Mutual Information

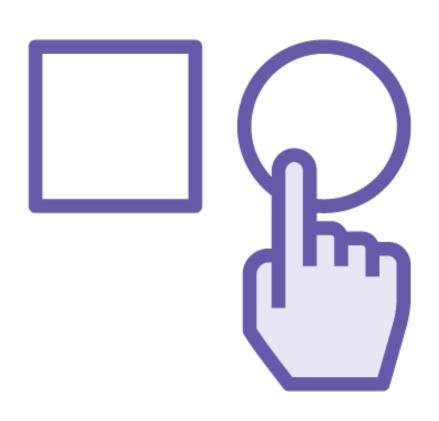
Variance Thresholding

If all points have the same value for an X-variable, that variable adds no information. Extend this idea and drop columns with variance below a minimum threshold.

Chi-square (x²) Feature Selection

For each X-variable, use the Chi-square test to evaluate whether that variable and Y are independent. If yes, drop that feature. Used for categorical X and Y.

Chi-square Feature Selection



Does observed data deviate from those expected in a particular analysis?

Tests the effect of one variable on the outcome, univariate analysis

Sum of the squared difference between observed and expected data in all categories

ANOVA

ANalysis **O**f **VA**riance

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

ANOVA Feature Selection

For each X-variable, use the ANOVA F-test to check whether mean of Y category varies for each distinct value of X. If not, drop that X-variable.

Diabetes Risk

Underweight patients Normal weight patients patients

Perform an ANOVA test to know whether the risk of diabetes is significantly different between these groups

ANOVA Hypotheses

Null Hypothesis

(H₀)

Alternate Hypothesis

(H₁)

H₀: All groups of patients are at an equal risk of diabetes

H₀: All groups of patients are NOT at an equal risk of diabetes

Mutual Information

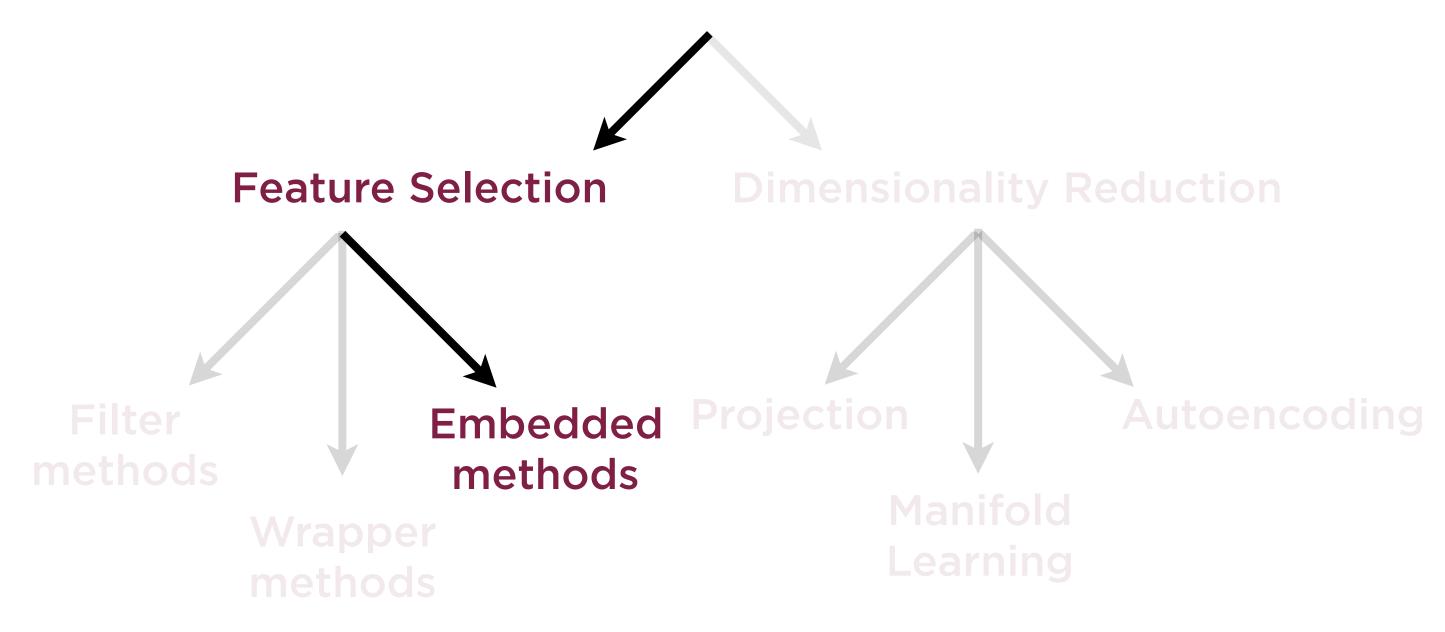
Measures the amount of information obtained on random variable by observing another.

Mutual Information

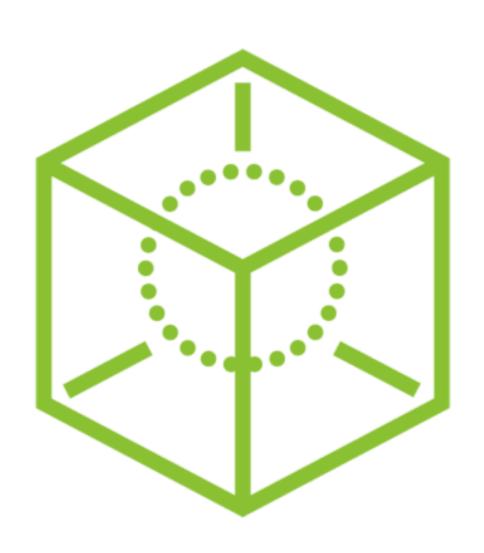
Conceptually similar to using ANOVA F-test for feature selection; superior as it also captures non-linear dependencies (unlike ANOVA-based feature selection).

Embedded Methods for Feature Selection

Reducing Complexity



Embedded Methods

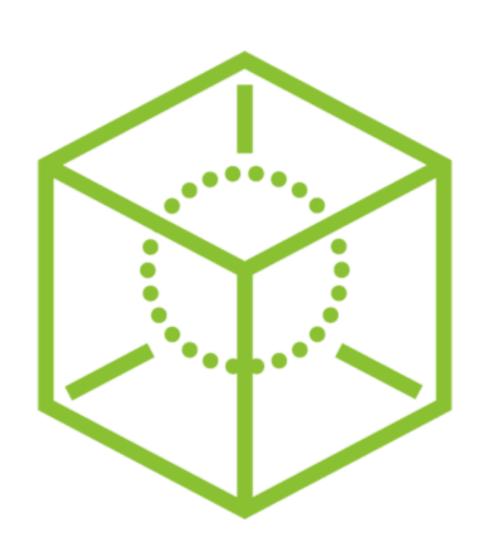


Features (columns) selected during model training

Feature selection effectively embedded within modeling

Only specific types of models perform feature selection

Embedded Feature Selection



Some machine learning algorithms automatically perform feature selection

- Decision trees
- Lasso regression

Jockey or Basketball Player?



Jockeys

Tend to be light to meet horse carrying limits



Basketball Players

Tend to be tall, strong and heavy

Jockey or Basketball Player?



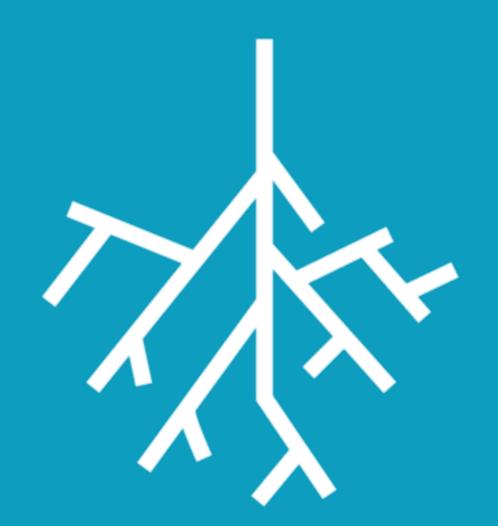
Intuitively know

Jockeys tend to be light

And not very tall

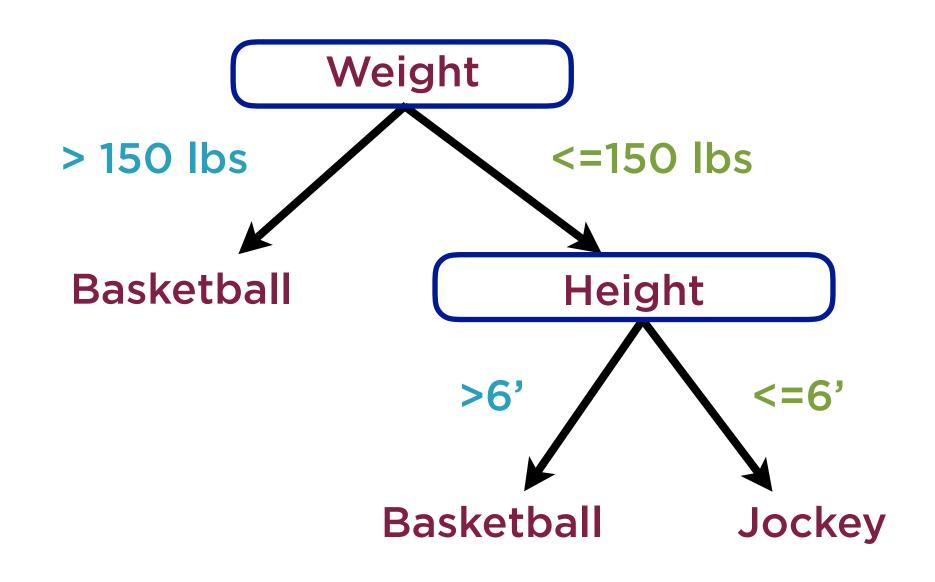
Basketball players tend to be tall

And also quite heavy



Decision trees set up a tree structure on training data which helps make decisions based on rules

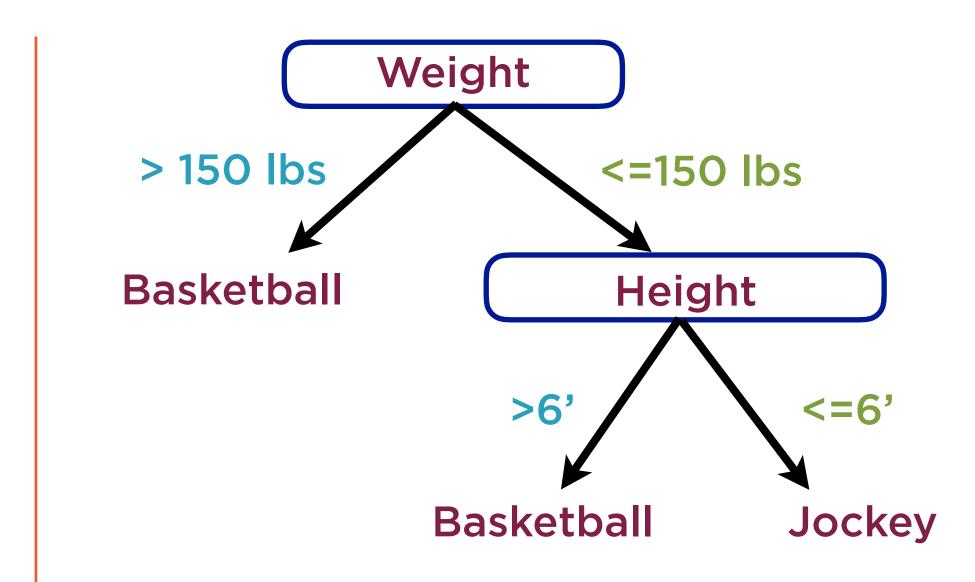
Fit Knowledge into Rules



Decision Tree

Fit knowledge into rules

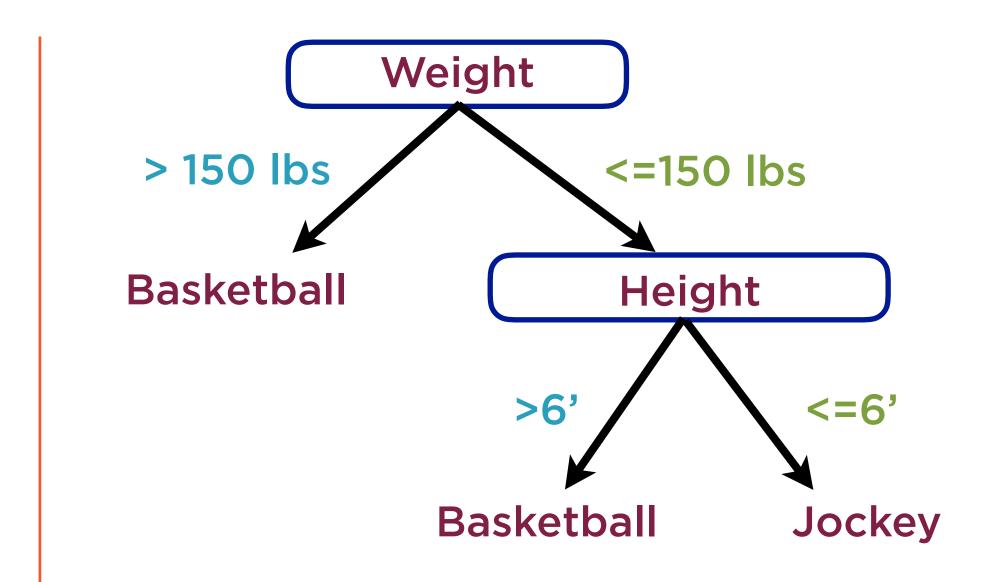
Each rule involves a threshold



Decision Tree

Order of decision variables matters

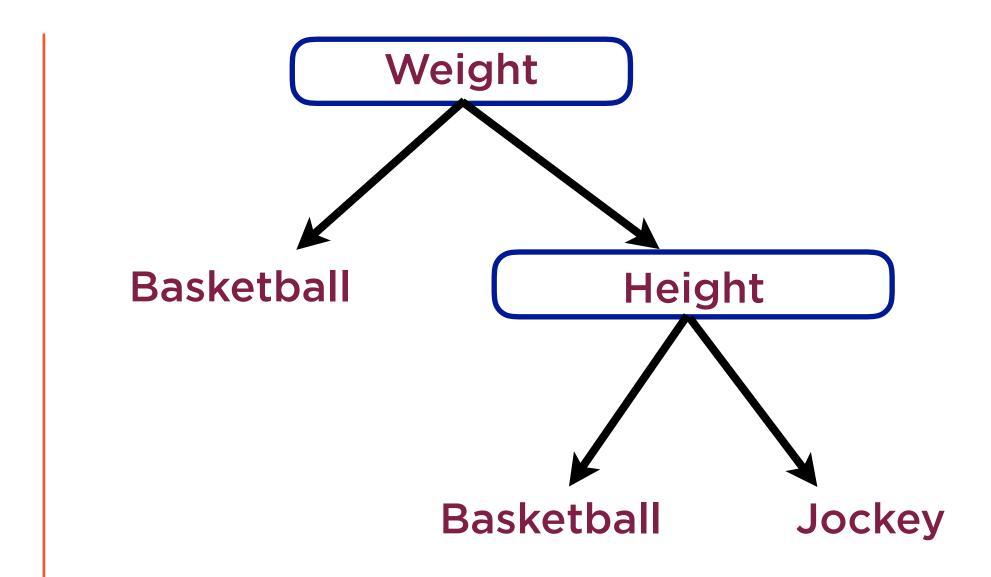
Rules and order found using ML

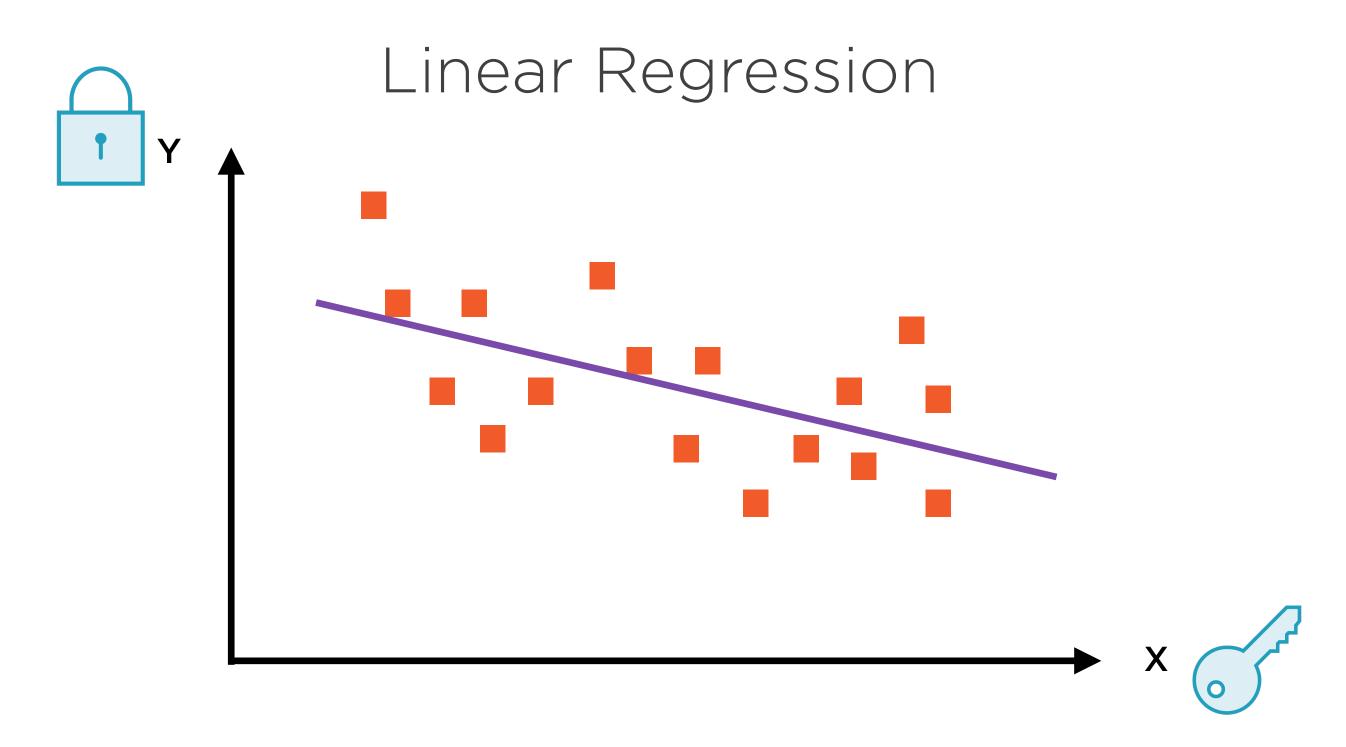


Decision Tree

Order of decision variables matters

Order determines feature importance





Find the "best fit" line that passes through this data

Linear Regression Line 1: $y = A_1 + B_1x$ Line 2: $y = A_2 + B_2x$

The "best fit" line is the one where the sum of the squares of the lengths of these dotted lines is minimum

Ordinary MSE Regression

Minimize

To find

A, B

The value of A and B define the "best fit" line

$$y = A + Bx$$

Minimize



To find

A, B

α is a hyperparameter

The value of A and B still define the "best fit" line

$$y = A + Bx$$

Minimize

 $+ \alpha (|A| + |B|)$

To find

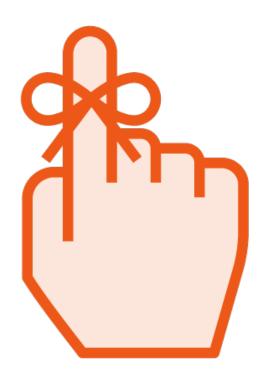
A, B

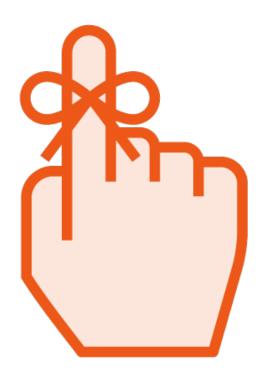
L-1 Norm of regression coefficients

α is a hyperparameter

The value of A and B still define the "best fit" line

$$y = A + Bx$$





 $\alpha = 0$ ~ Regular (MSE regression)

 $\alpha \rightarrow \infty$ ~ Force small coefficients to zero

Model selection by tuning α

Eliminates unimportant features



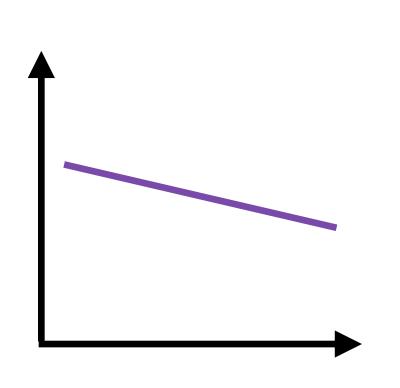
"Lasso" ~ <u>Least Absolute Shrinkage and</u> <u>Selection Operator</u>

Math is complex

No closed form, needs numeric solution

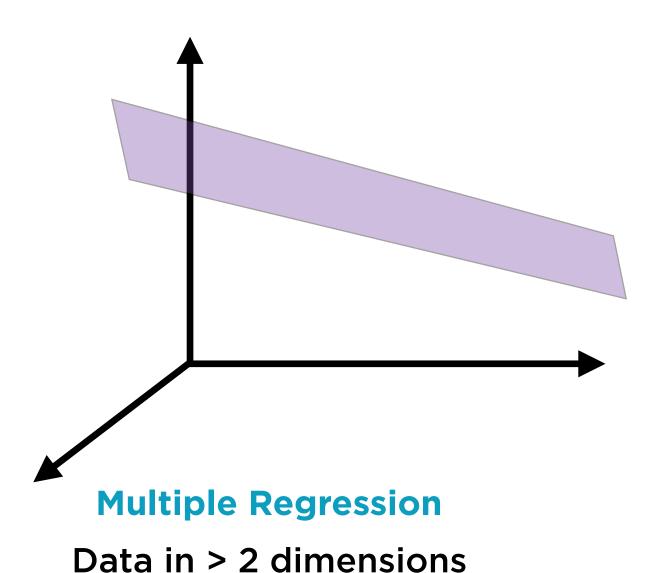
Multi-collinearity in Regression Models

Simple and Multiple Regression

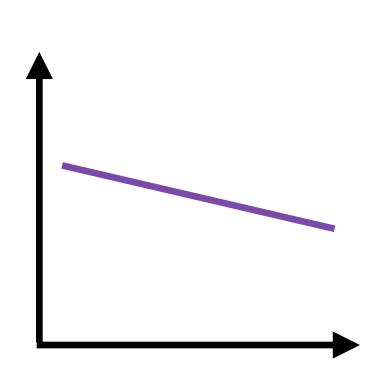


Simple Regression

Data in 2 dimensions

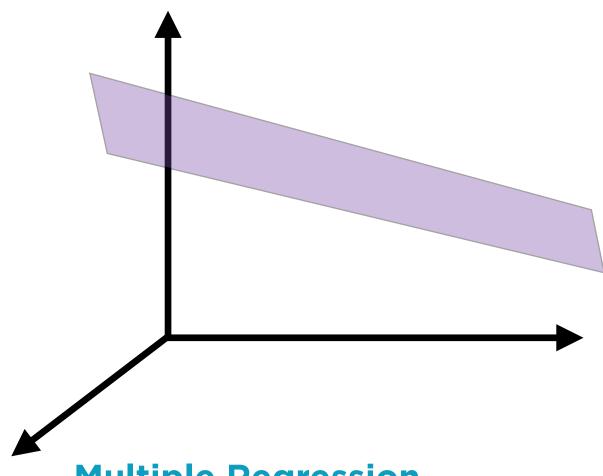


Simple and Multiple Regression



Simple Regression

Risks exist, but can usually be mitigated analysing R² and residuals



Multiple Regression

Risks are more complicated, require interpreting regression statistics

The big new risk with multiple regression is **multicollinearity**: X variables containing the same information

Multiple Regression

Regression Equation:

$$y = C_1 + C_2 X_1 + ... + C_k X_{k-1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{1k-1} \\ 1 & x_{21} & x_{2k-1} \\ 1 & x_{31} & \dots & x_{3k-1} \\ \dots & \dots & \dots \\ 1 & x_{n1} & x_{nk-1} \end{bmatrix} + \begin{bmatrix} C_1 \\ C_2 \\ \dots & C_k \end{bmatrix}$$

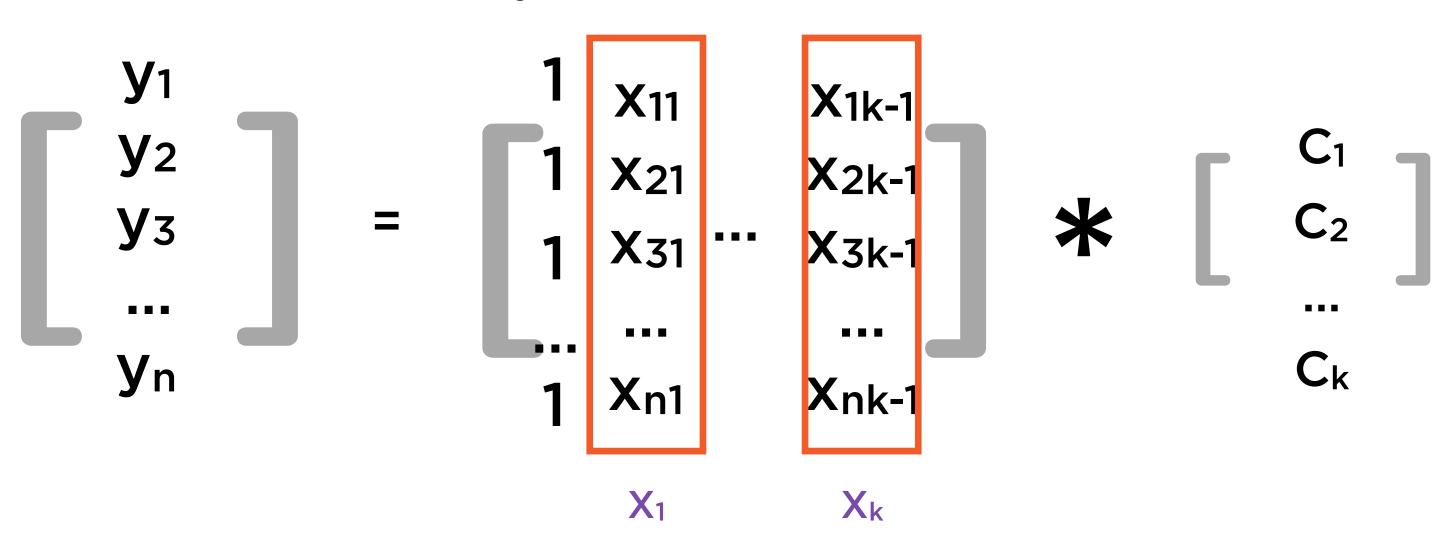
n Rows, 1 Column

n Rows, k Columns k Rows, 1 Column

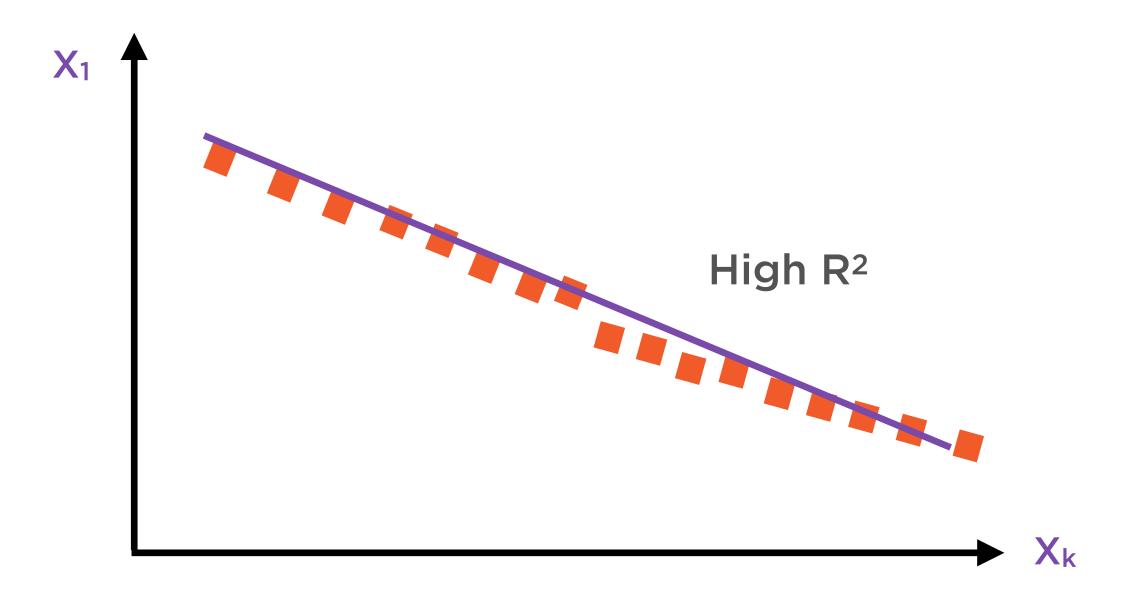
Multiple Regression

Regression Equation:

$$y = C_1 + C_2 X_1 + ... + C_k X_{k-1}$$

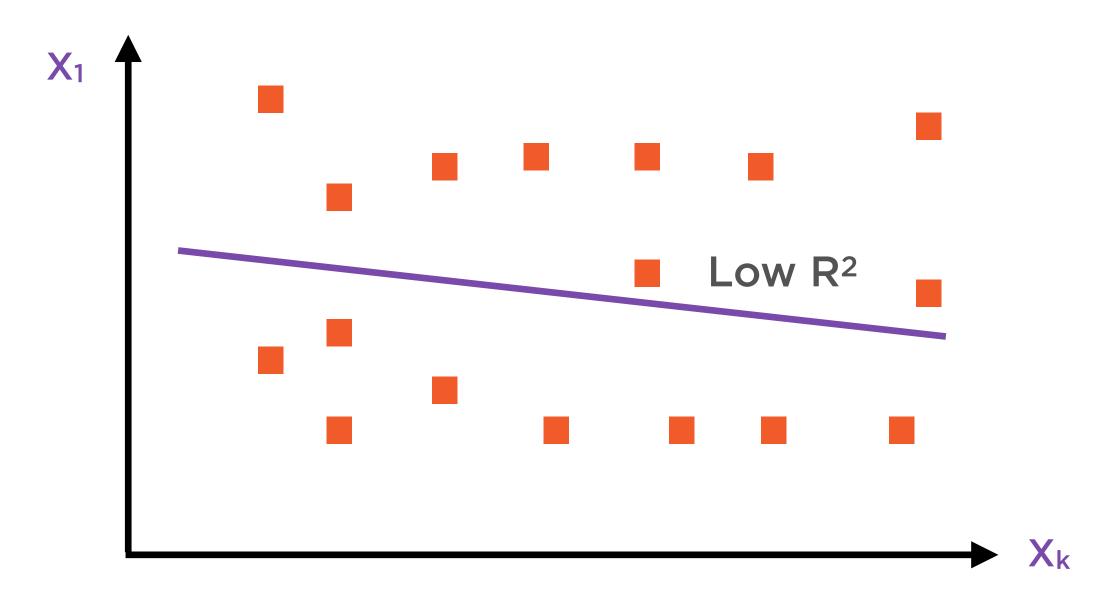


Bad News: Multicollinearity Detected



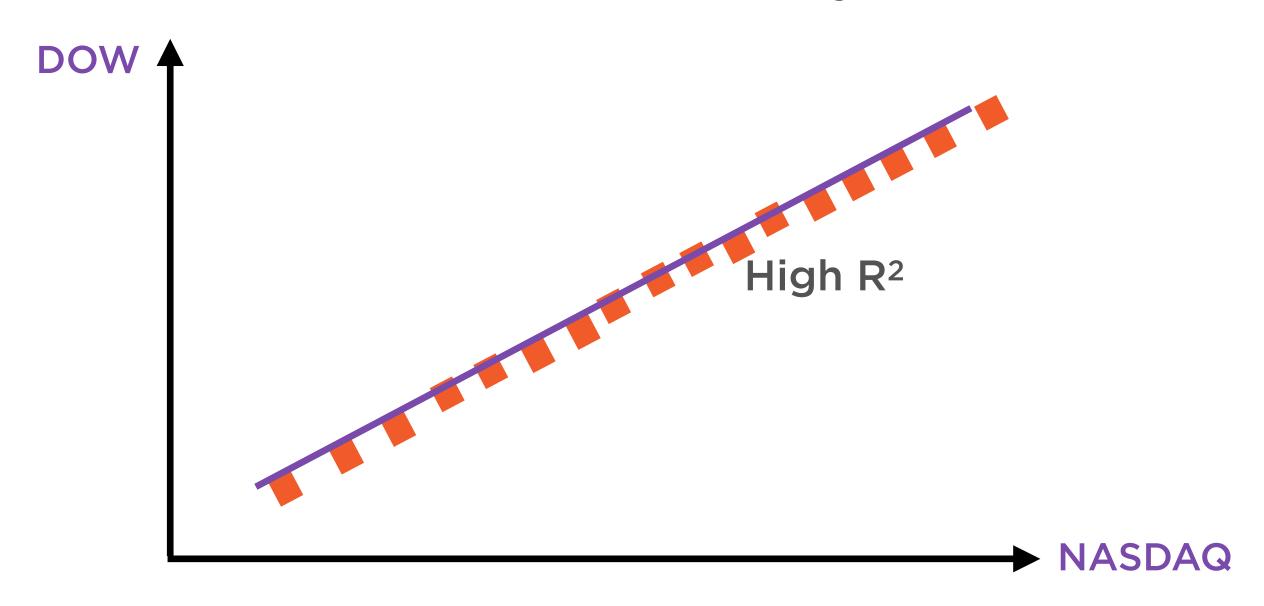
Highly correlated explanatory variables

Good News: No Multicollinearity Detected



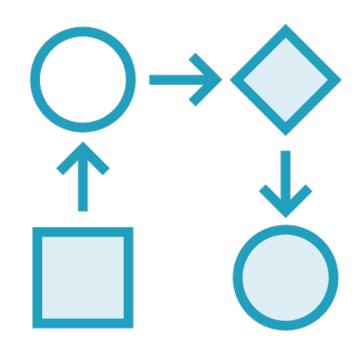
Uncorrelated explanatory variables

Bad News: Multicollinearity Detected



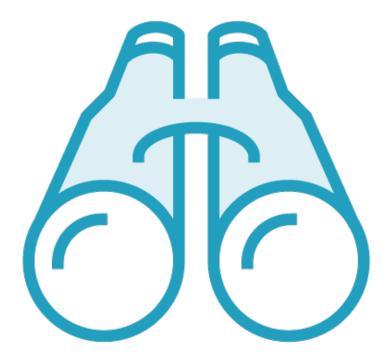
Highly correlated explanatory variables

Multicollinearity Kills Regression's Usefulness



Explaining Variance

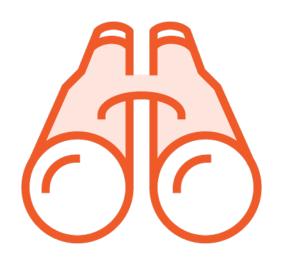
The R² as well as the regression coefficients are not very reliable



Making Predictions

The regression model will perform poorly with out-of-sample data

Multicollinearity: Prevention and Cure





Big-picture understanding of the data



Nuts and Bolts

Setting up data right



Heavy Lifting

Factor analysis, principal components analysis (PCA)

Summary

Curse of dimensionality

Reducing complexity of data

Understanding feature selection

Filter methods

Embedded methods

Wrapper methods