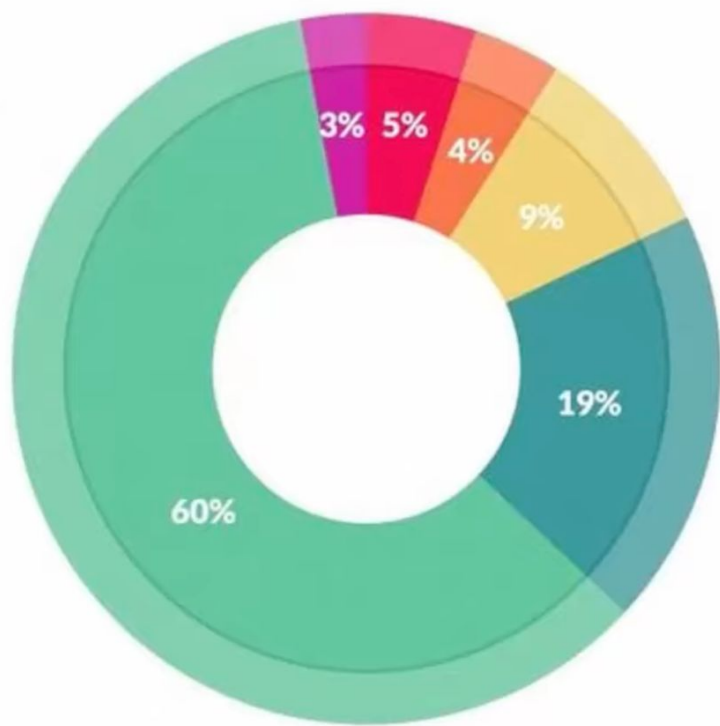


MIDS W207

Applied Machine Learning

Fall 2022




Week 3



## What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

# Prediction



**\$740,000** 4 bd | 3 ba | 2,028 sqft  
30198 La Primavera St, Temecula, CA 92592  
Est. payment: \$3,399/mo [Get pre-qualified](#)

[Message](#)

[Overview](#) [Facts and features](#) [Home value](#) [Price and tax history](#)

### Facts and features

**Type:** Single Family Residence  
**Parking:** 2 Attached Garage spaces  
**Year built:** 1974  
**HOA:** \$90 monthly  
**Heating:** Central  
**Lot:** 0.40 Acres  
**Cooling:** Central Air  
**Price/sqft:** \$365

### Interior details

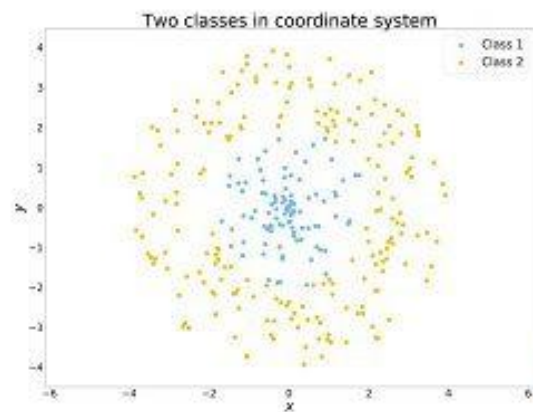
**Bedrooms and bathrooms**  
Bedrooms: 4  
Bathrooms: 3  
Full bathrooms: 2  
1/2 bathrooms: 1  
Main level bathrooms: 1

**Appliances**  
Appliances included: Built-in Range, Gas Cooktop, Disposal, Refrigerator  
Laundry features: Inside, Laundry Room

**Flooring**  
Flooring: Carpet, Laminate, Tile

**Interior Features**  
Interior features: Ceiling Fan(s), Granite Counters, All Bedrooms Up

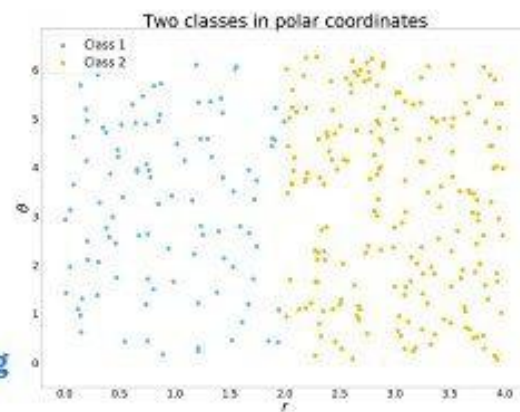
Features				Label
Size	Beds	Baths	Zip	Price
1100	1	1	64576	1.29
1900	3	1.5	78321	2.14
2800	3	3	98712	3.10
3400	4	3.5	25721	3.75



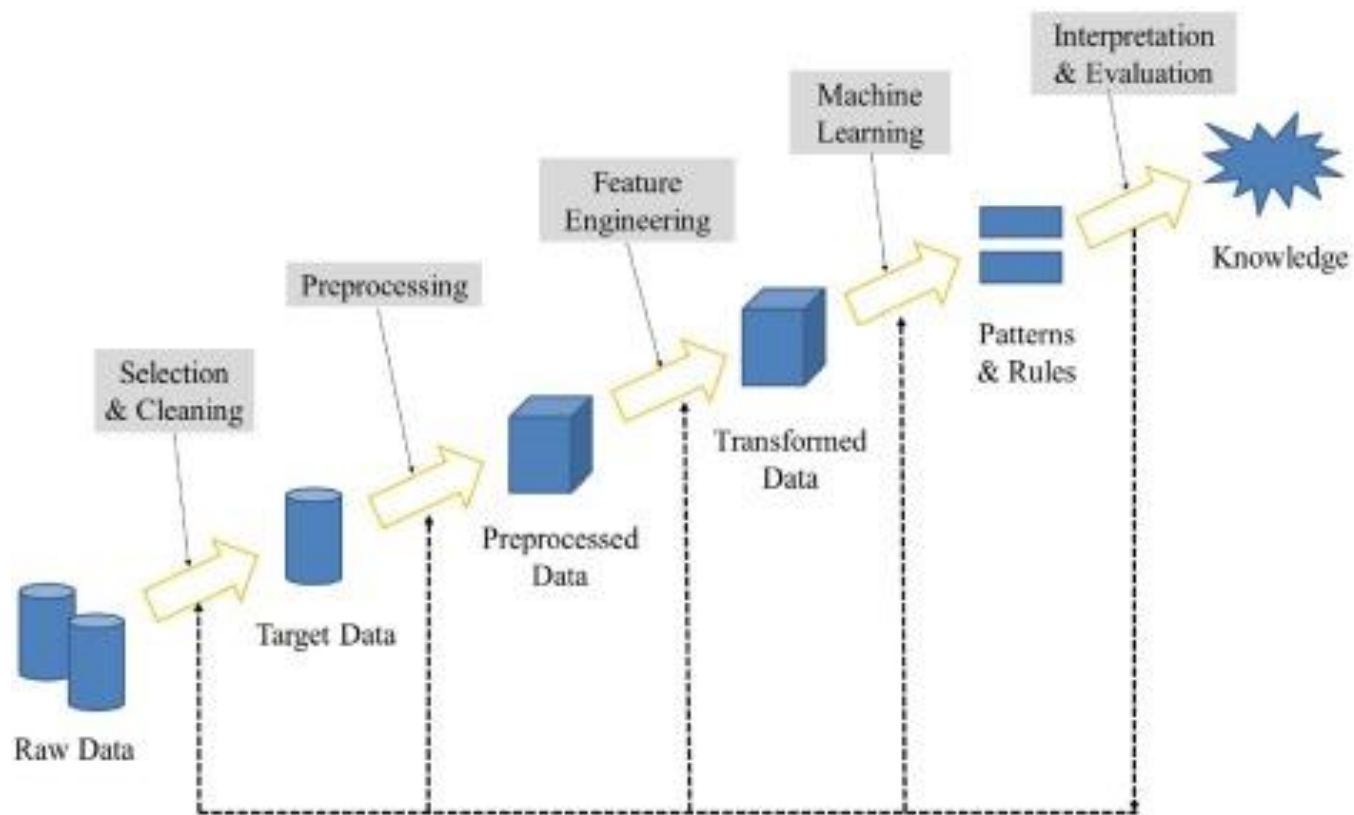
Tangled



Feature engineering



Transparent



# Missing Values

	col1	col2	col3	col4	col5
<b>0</b>	2	5.0	3.0	6	NaN
<b>1</b>	9	NaN	9.0	0	7.0
<b>2</b>	19	17.0	NaN	9	NaN

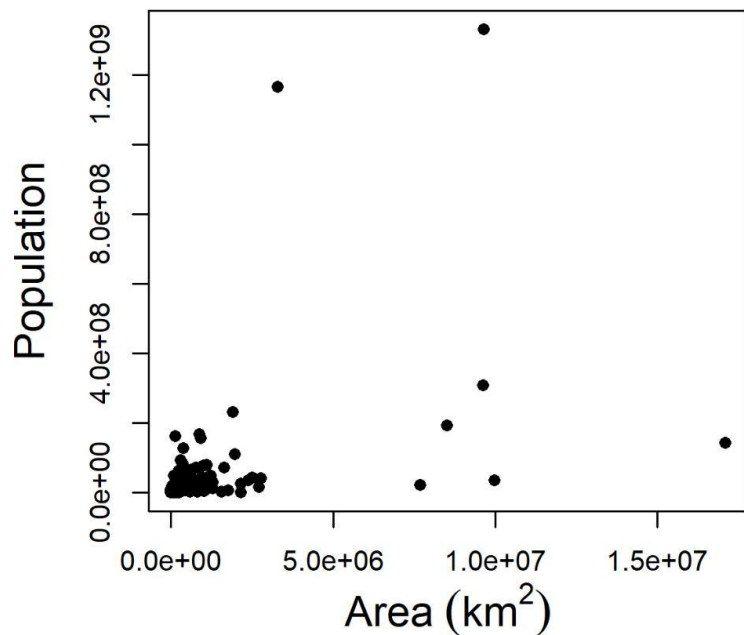
`mean()`



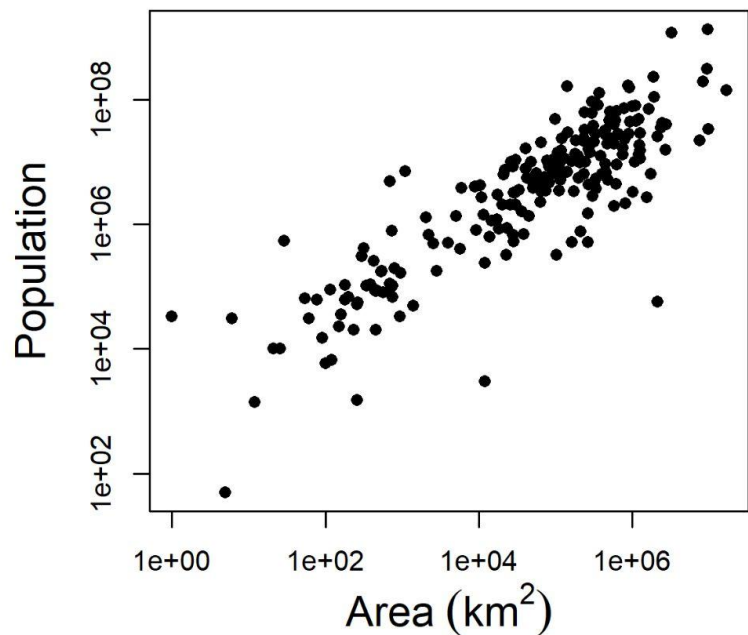
	col1	col2	col3	col4	col5
<b>0</b>	2.0	5.0	3.0	6.0	7.0
<b>1</b>	9.0	11.0	9.0	0.0	7.0
<b>2</b>	19.0	17.0	6.0	9.0	7.0

# Transforming Features

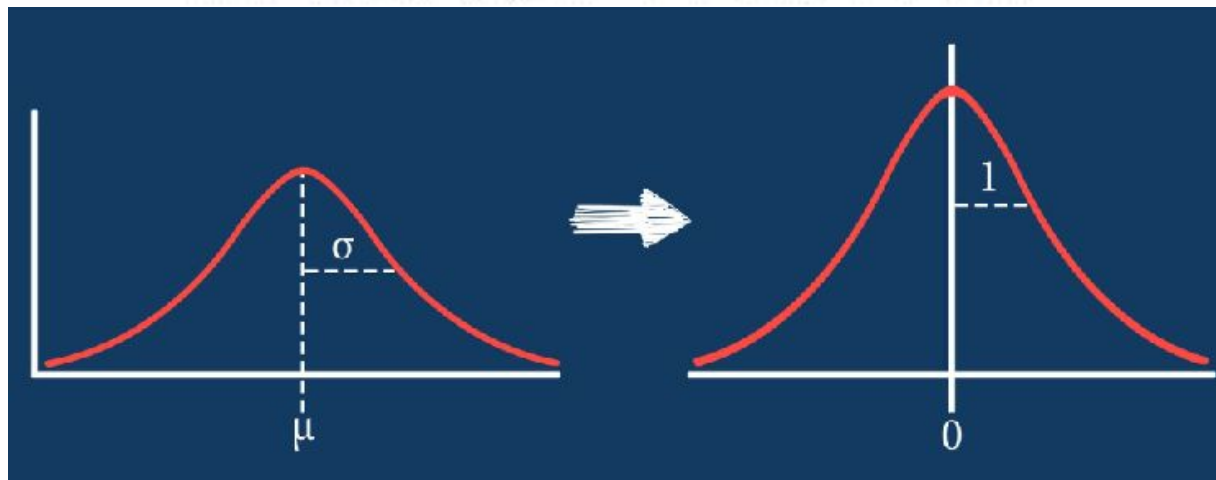
**Raw data**



**Log-transformed data**



# Scaling





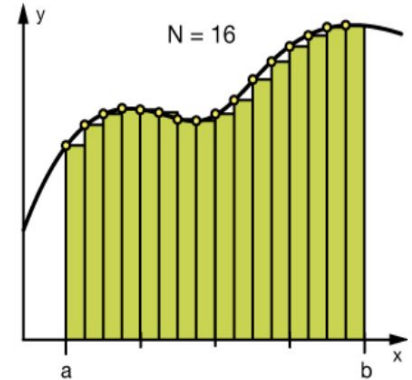
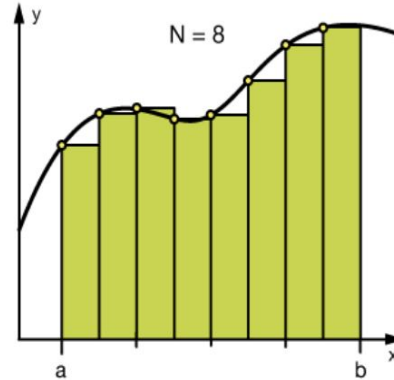
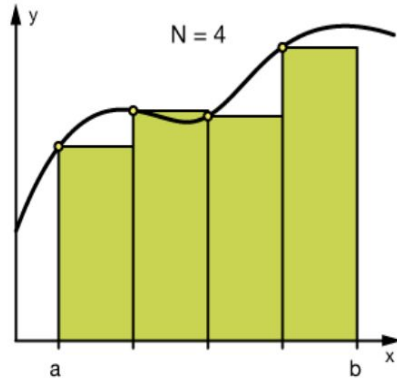
# Bucketing

## #Numerical Binning Example

Value	Bin
0-30	-> Low
31-70	-> Mid
71-100	-> High

## #Categorical Binning Example

Value	Bin
Spain	-> Europe
Italy	-> Europe
Chile	-> South America
Brazil	-> South America



# Encoding

## Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50



## One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

## Feature Selection

Subsetting the features

Ex: Using correlation with the dependent variable

## Feature Extraction

Creating new features when we could **NOT** have used raw features

Ex: from images to RGB values. Automatic methods such as PCA

## Feature Engineering

Creating new features when we could have used raw features

Ex: Creating a new dummy variable for working days

## Feature Learning

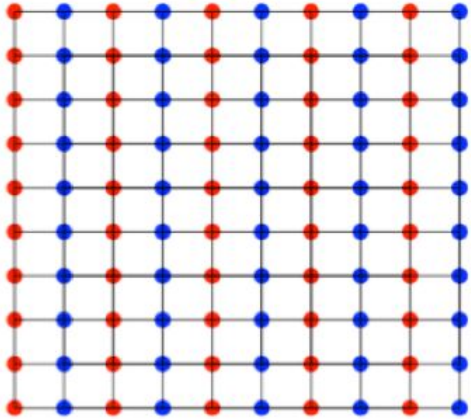
Constructing features automatically

Ex: Supervised neural networks, Independent component analysis

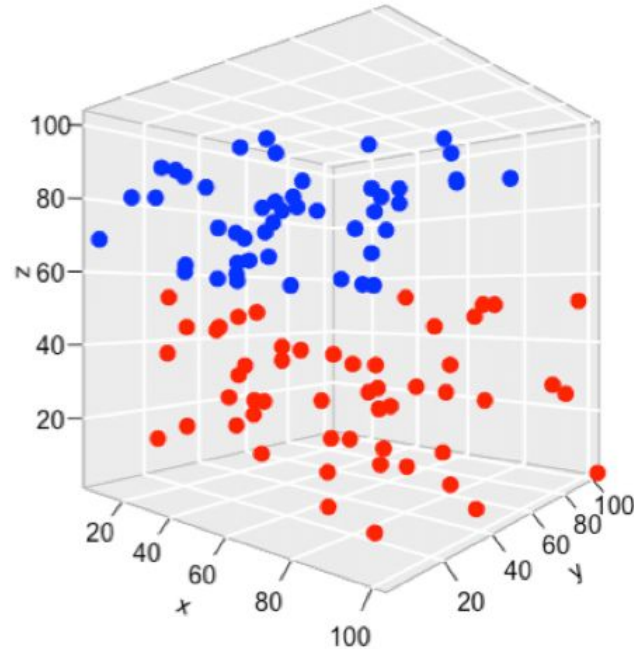
# Curse of Dimensionality



**(A) 1-D**

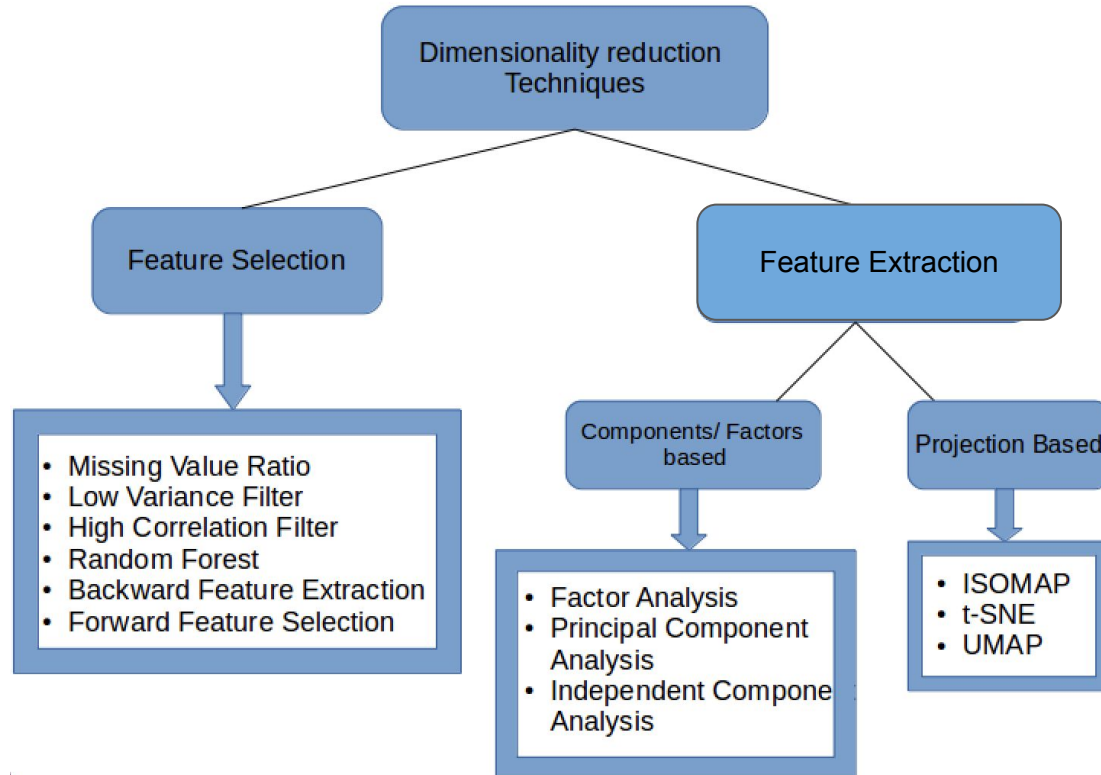


**(B) 2-D**



**(C) 3-D**

# Dimensionality Reduction



## Numerical

- Standardization

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Normalization

$$X_{\text{normalized}} = \frac{(X - X_{\text{minimum}})}{(X_{\text{maximum}} - X_{\text{minimum}})}$$

- Bucketing

Age<18	19<=Age<30	30<=Age<40	Age>=40
--------	------------	------------	---------

## Categorical

- One-hot encoding

Label Encoding						
Food Name	Categorical #	Calories				
Apple	1	95				
Chicken	2	231				
Broccoli	3	50				

→

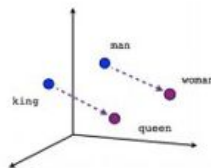
One Hot Encoding					
Apple	Chicken	Broccoli	Calories		
1	0	0	95		
0	1	0	231		
0	0	1	50		

- TF-IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

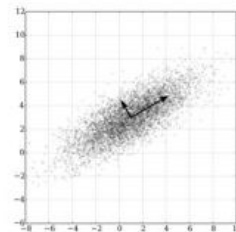
$tf_{i,j}$  = number of occurrences of  $i$  in  $j$   
 $df_i$  = number of documents containing  $i$   
 $N$  = total number of documents

- Word embeddings



## Dimensionality Reduction

- Principal component analysis (PCA)



- t-SNE

