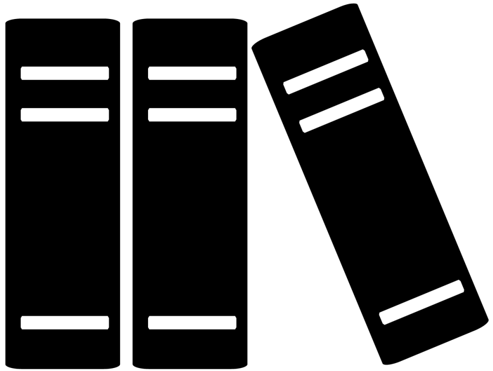
The Mysticism of Machine Learning

Data Boot Camp

Lesson 21.1

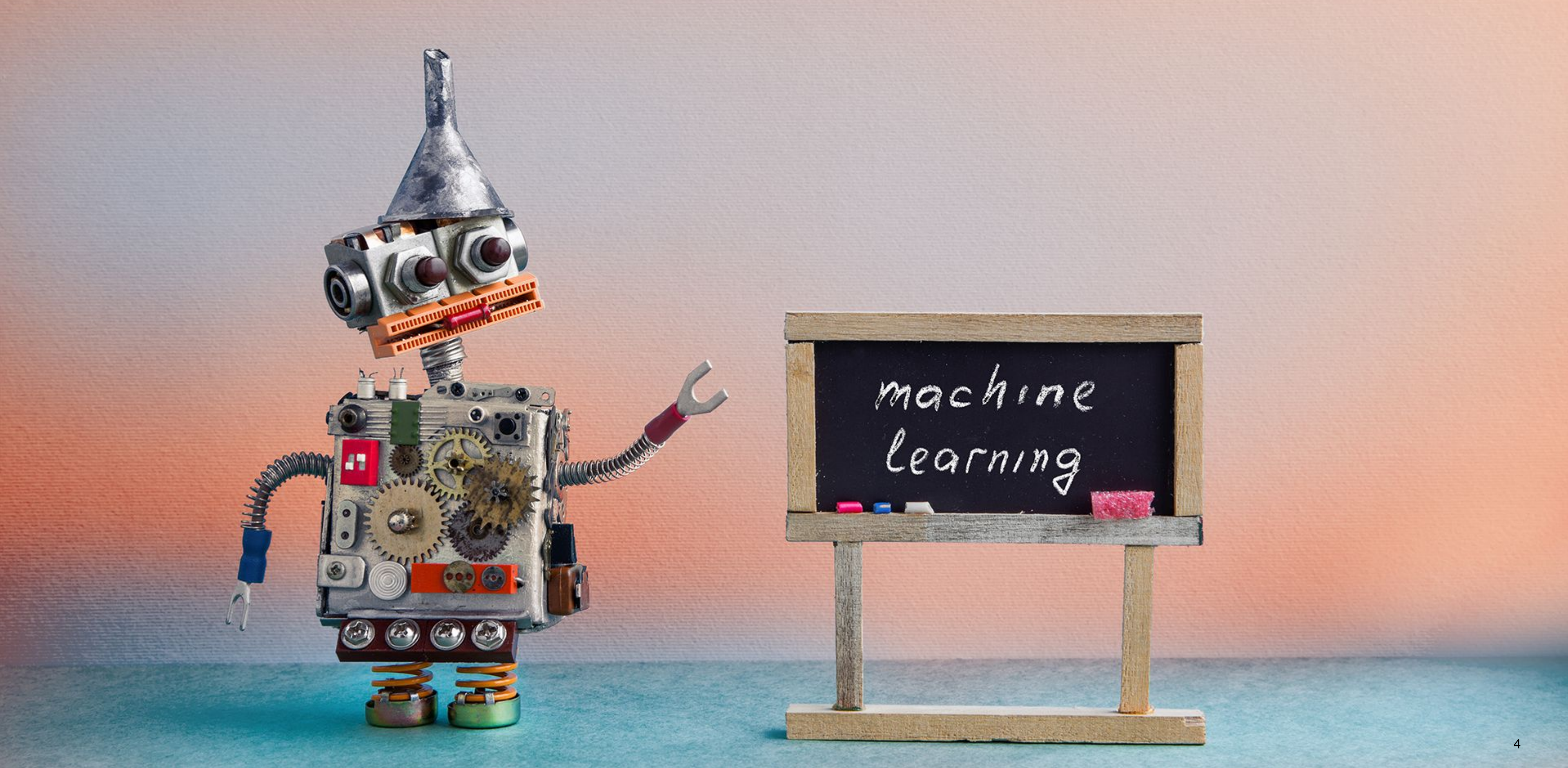
Machine Learning in a Nutshell

Libraries



Statistics

So It Begins...

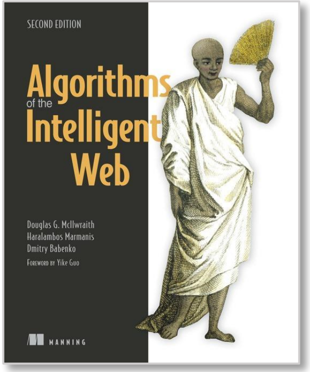


Basic Definitions

Intelligent Algorithms (Definition)

> Intelligent algorithms are ones that use data to modify its behavior. Intelligent algorithms differ in that they can change their behavior as they run, often resulting in a user experience that many would say is intelligent.

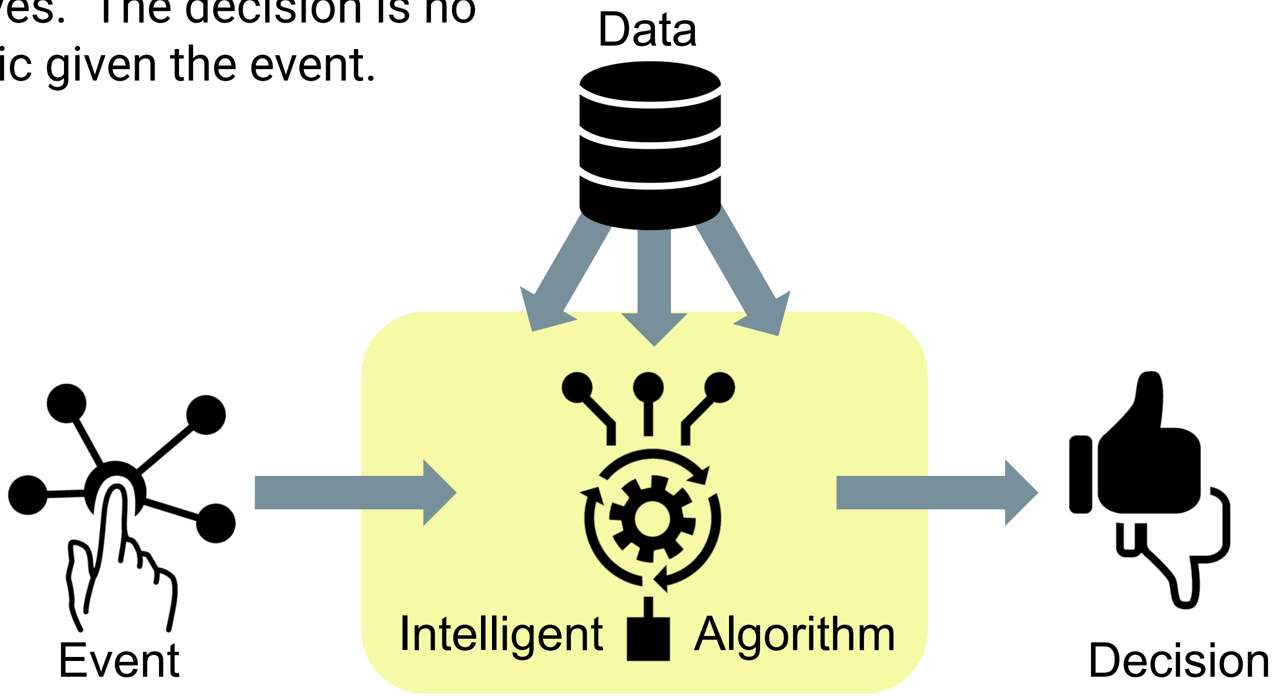
—*Algorithms of the Intelligent Web, Second Edition*



Algorithms of the Intelligent Web, Second Edition
by *Douglas G. McIlwraith Haralambos Marmanis Dmitry Babenko*
Publisher: Manning Publications
Release Date: August 2016

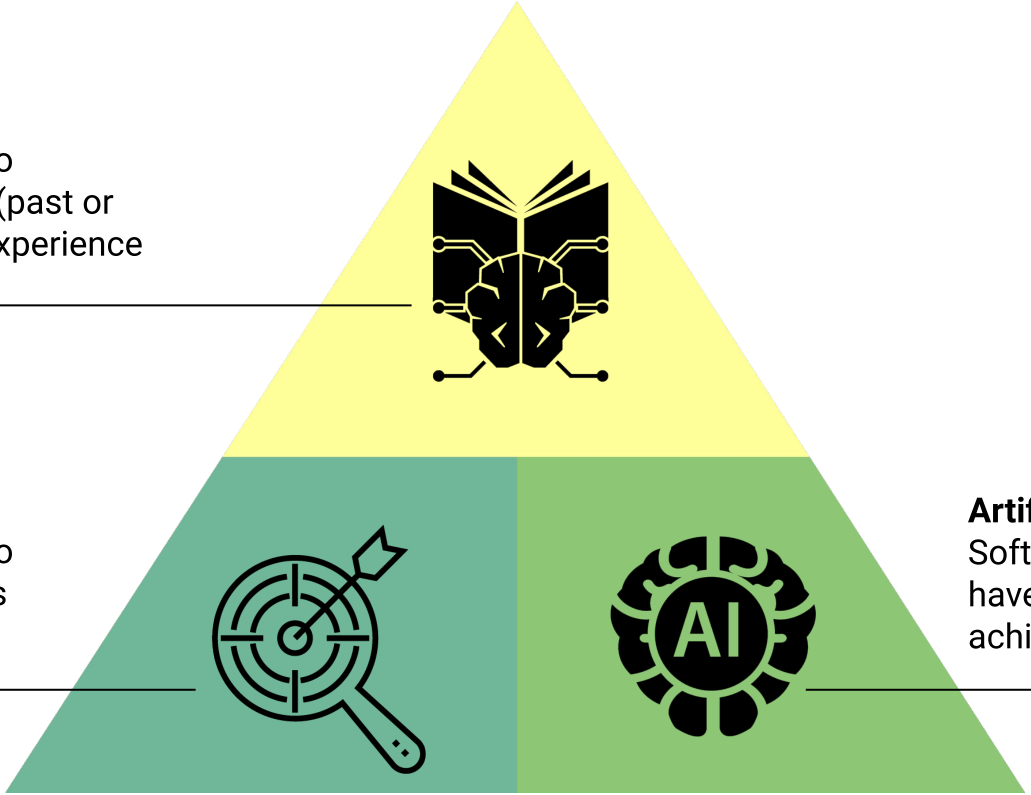
Intelligent Algorithms (Diagram)

Intelligent algorithms are ones that respond to data such that the algorithm gets better. It effectively “evolves.” The decision is no longer deterministic given the event.



Intelligent Algorithms (Triad)

Machine Learning
Capability of software to generalize phenomena (past or future) based on past experience



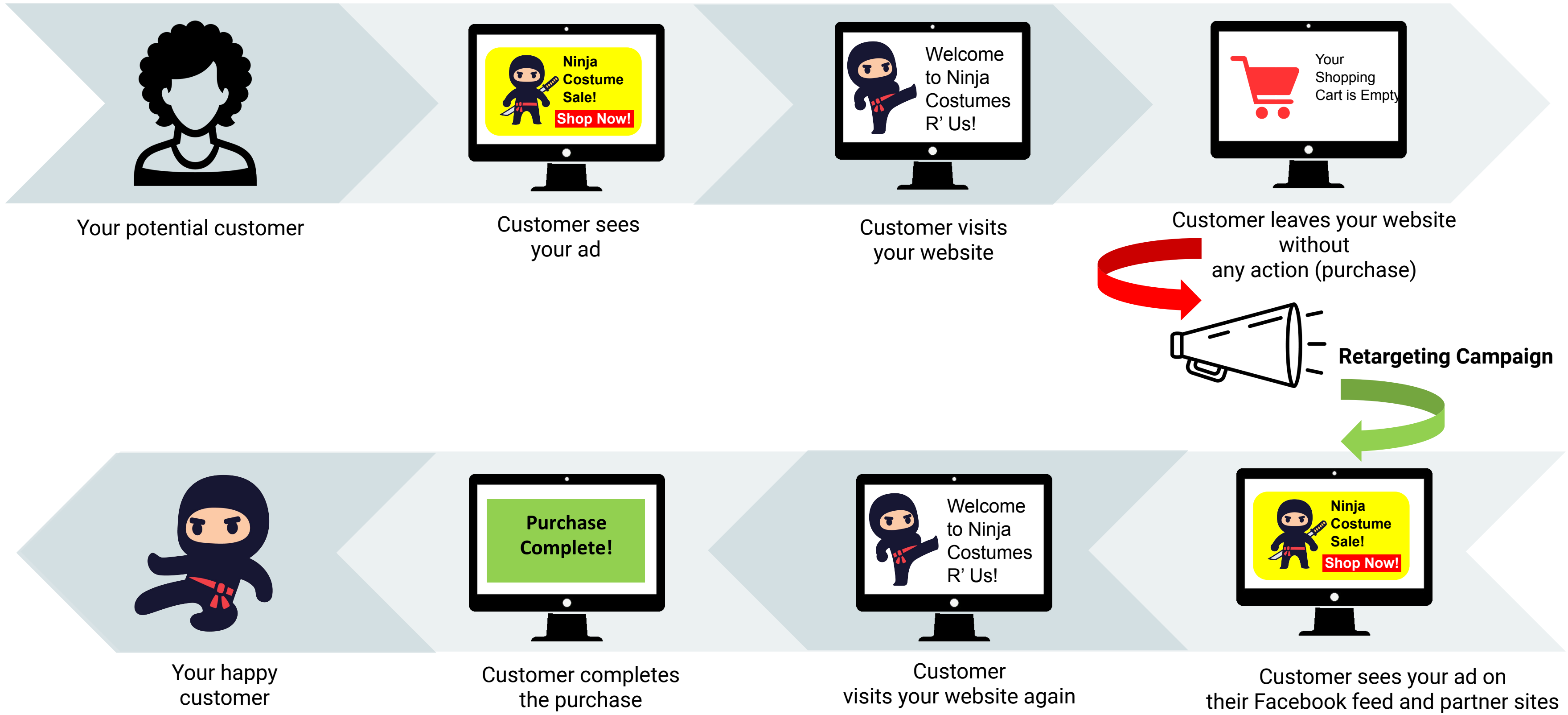
Predictive Analytics
Capability of software to predict future outcomes based on historic data

Artificial Intelligence
Software (and machines) that have a series of options to achieve a particular goal

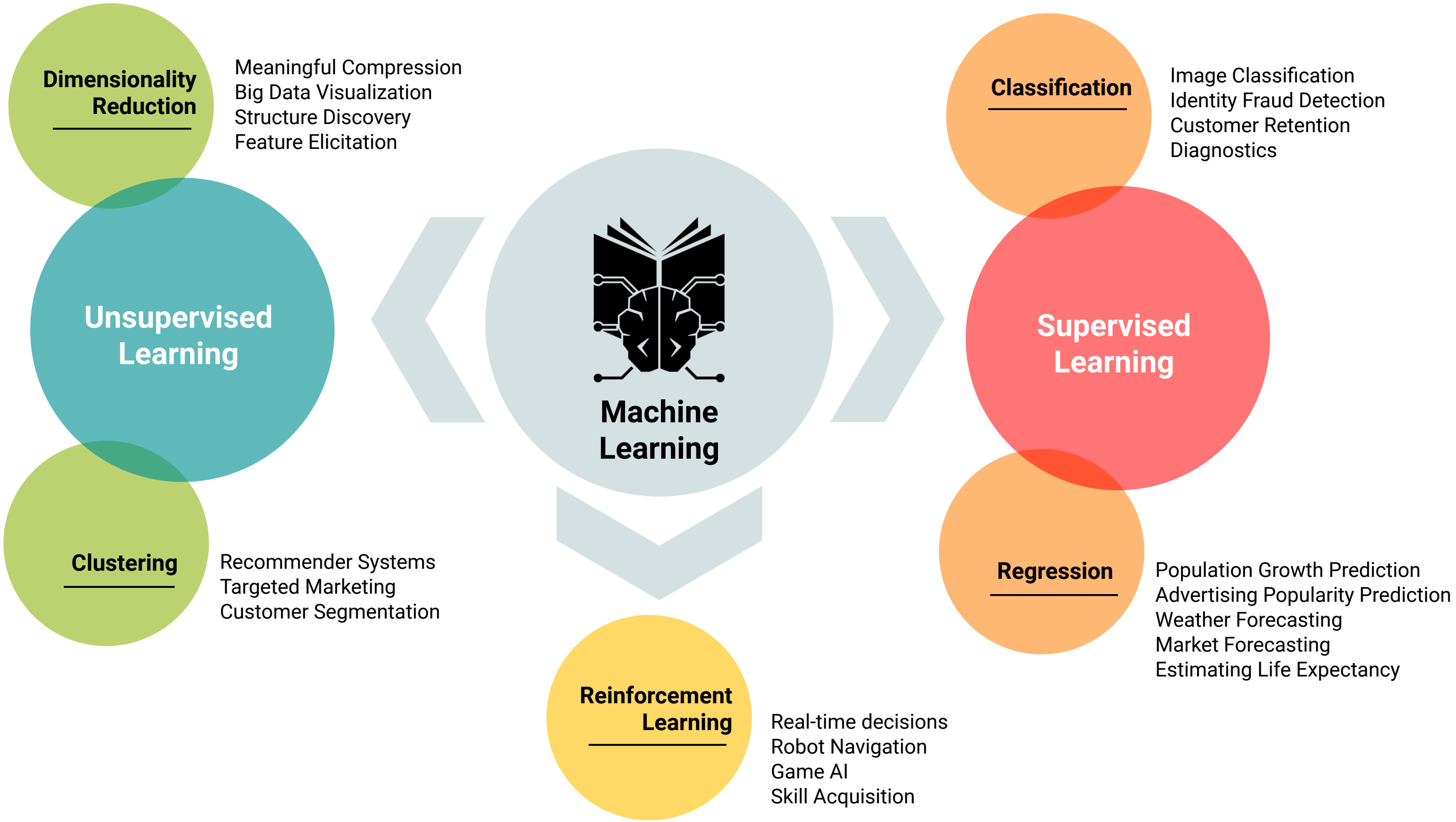
Artificial Intelligence (Example)

Predictive Analytics (Example)

How retargeting ads work:



Machine Learning (Categories)



Machine Learning (Categories)

Machine Learning

Unsupervised Learning

Dimensionality Reduction

Meaningful Compression
Big Data Visualization
Structure Discovery
Feature Elicitation

Clustering: Finding groups within a population

Recommender Systems
Targeted Marketing
Customer Segmentation

Supervised Learning

Classification: Applying labels to data

Image Classification
Identity Fraud Detection
Customer Retention
Diagnostics

Regression: Fitting data to predict where a new data point lies

Population Growth Prediction
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Estimating Life Expectancy

Reinforcement Learning

Real-time decisions
Robot Navigation
Game AI
Skill Acquisition

Machine Learning (Supervised)

Supervised Learning: Algorithms for which the potential outcomes are knowable in advance (i.e., category or numeric range) and can be used to correct the model's predictions.

01 Example

Using data such as credit score, credit history, income, etc., we are trying to predict whether an individual is a credit risk or not.

Known Category:
"Credit Risk" vs. "Not Credit Risk"

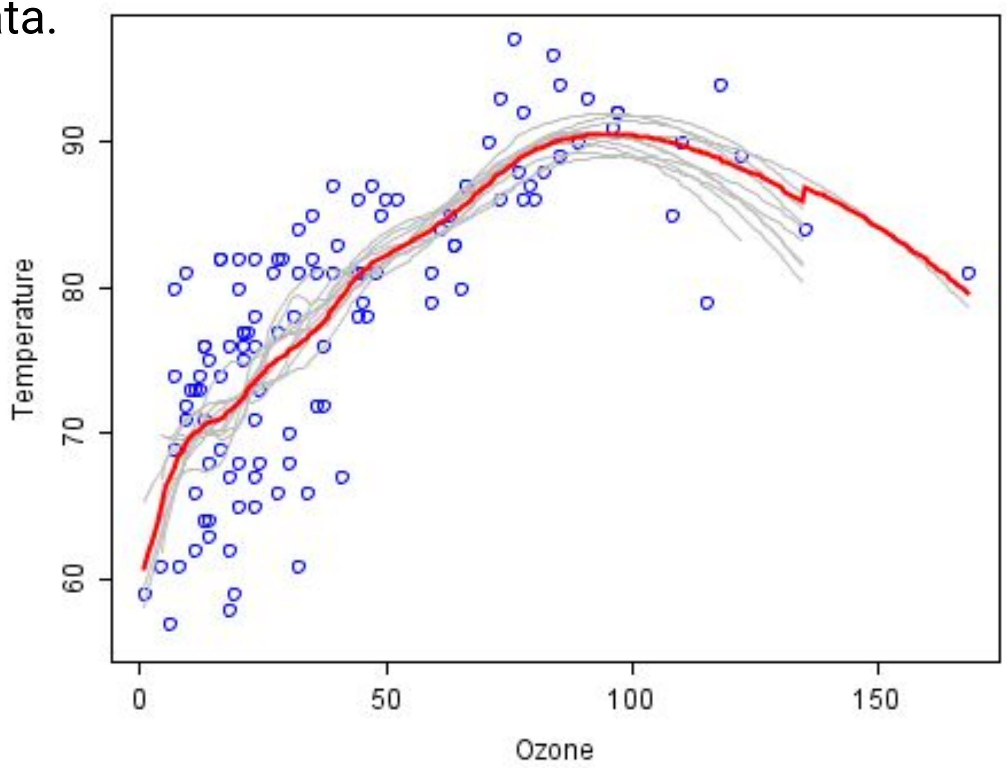
02 Example

Using features such as number of bedrooms, square feet, etc., we are trying to predict the market value of a house.

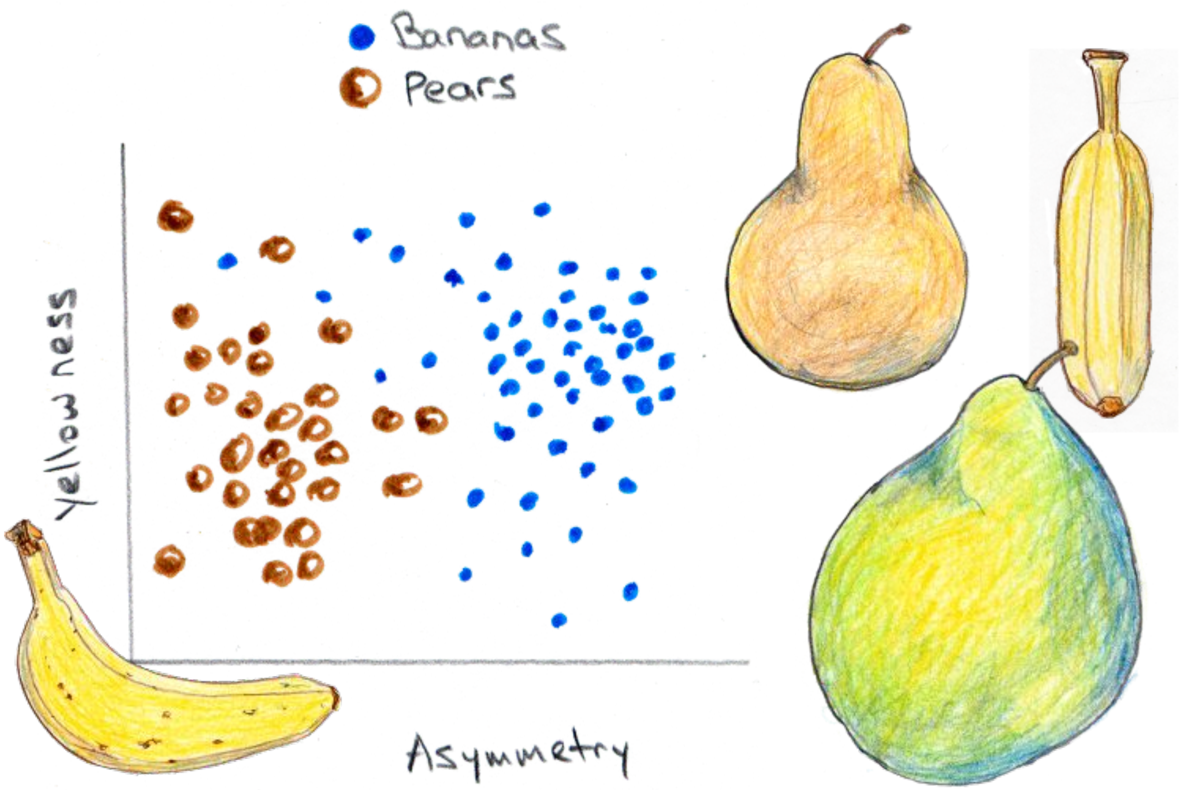
Numeric Range:
50,000—500,000

Machine Learning (Regression)

We'll be revisiting regression to predict the location of data points based on old data.



Machine Learning (Classification)



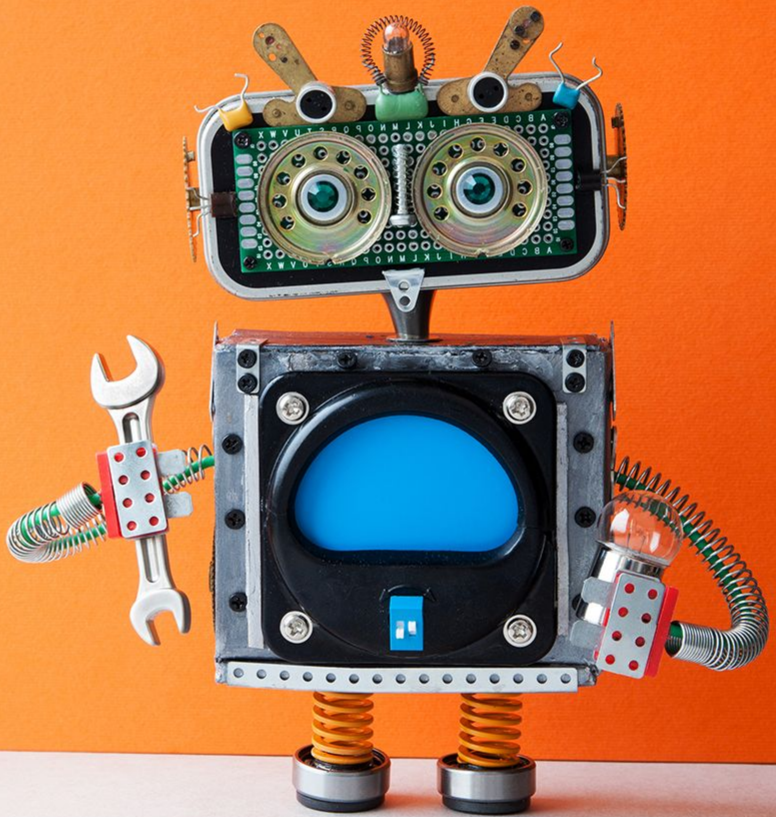
Machine Learning (Classification)

In classification problems, our focus is identifying which predefined label our data falls into, based on the **features** we have.

From: cheapsales@buystufffromme.com To: ang@cs.stanford.edu Subject: Buy now!	From: Alfred Ng To: ang@cs.stanford.edu Subject: Christmas dates?
Deal of the week! Buy now! Rolex w4tchs - $100 Med1cine (any kind) - $50 Also low cost M0rgages available.	Hey Andrew, Was talking to Mom about plans for Xmas. When do you get off work? Meet Dec 22? -Alf
Spam	Non-Spam

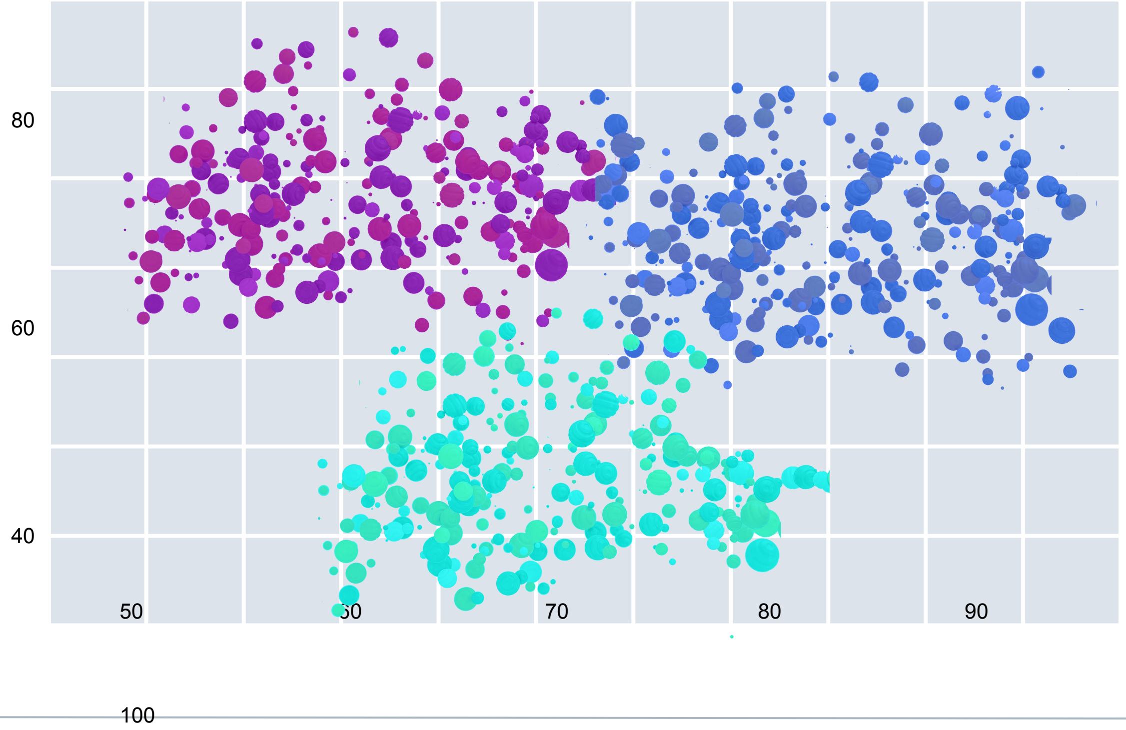
Machine Learning (Unsupervised)

Unsupervised Learning: Algorithms for which the potential outcomes are unlabeled. Inferences are made directly from the data without feedback from known outcomes or labels.



Machine Learning (Clustering)

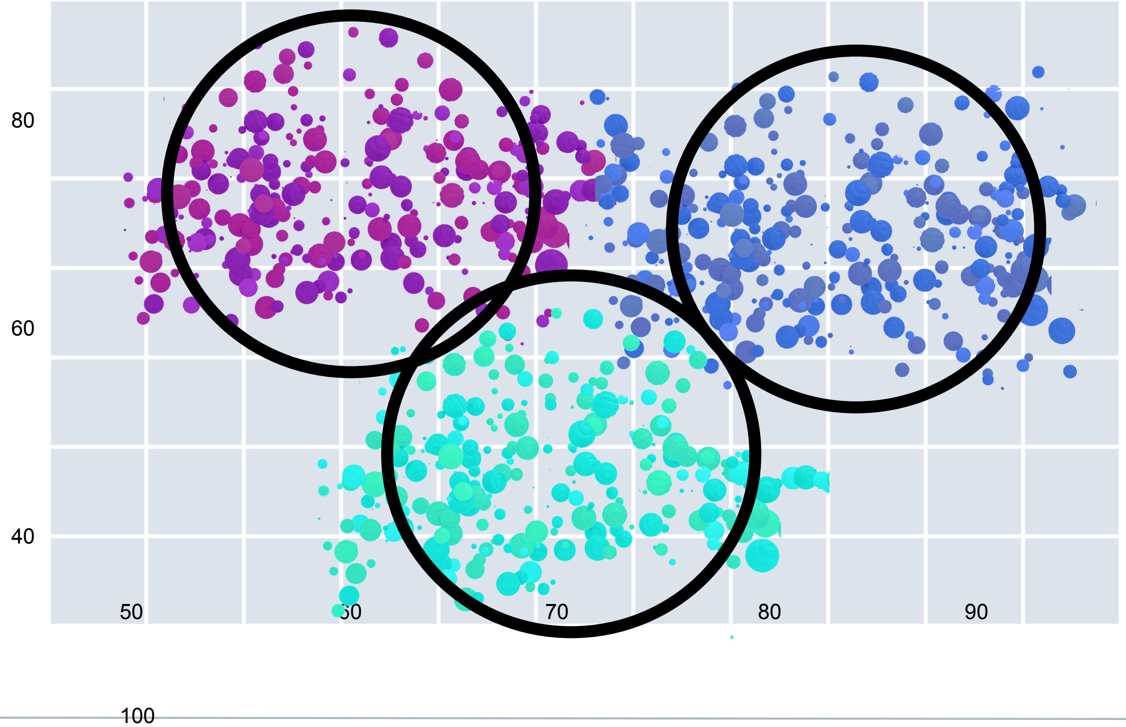
In this clustering problem, we expect our algorithm to find the groupings of data points based on location.



Machine Learning (Clustering)

In this clustering problem, we expect our algorithm to find the groupings of data points based on location.

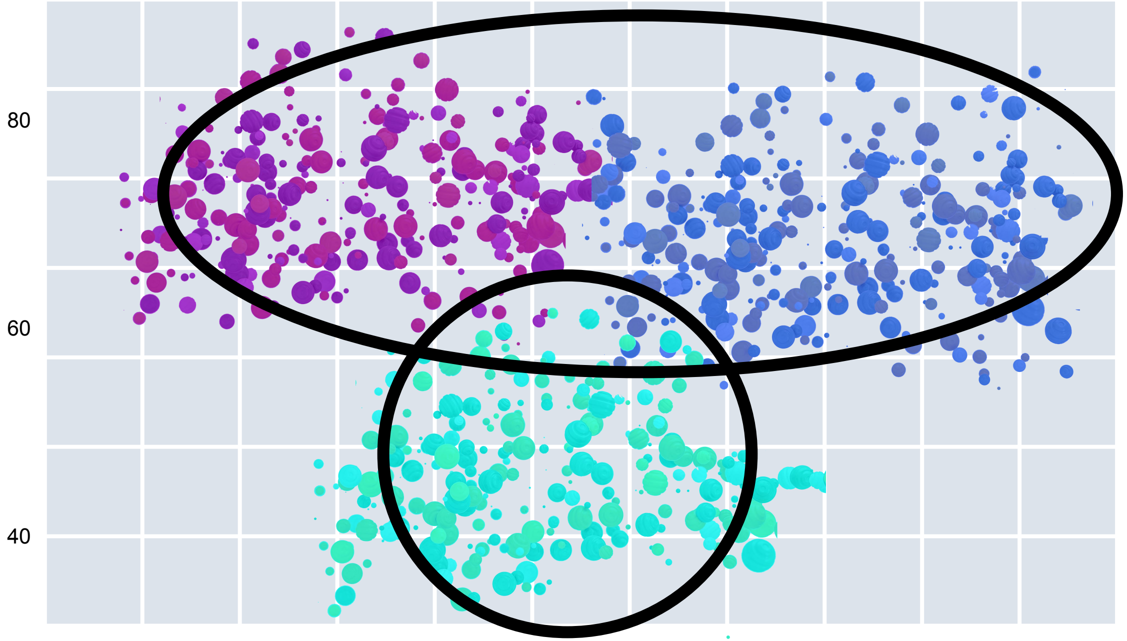
K=3



Machine Learning (Clustering)

But the problem is more complex:

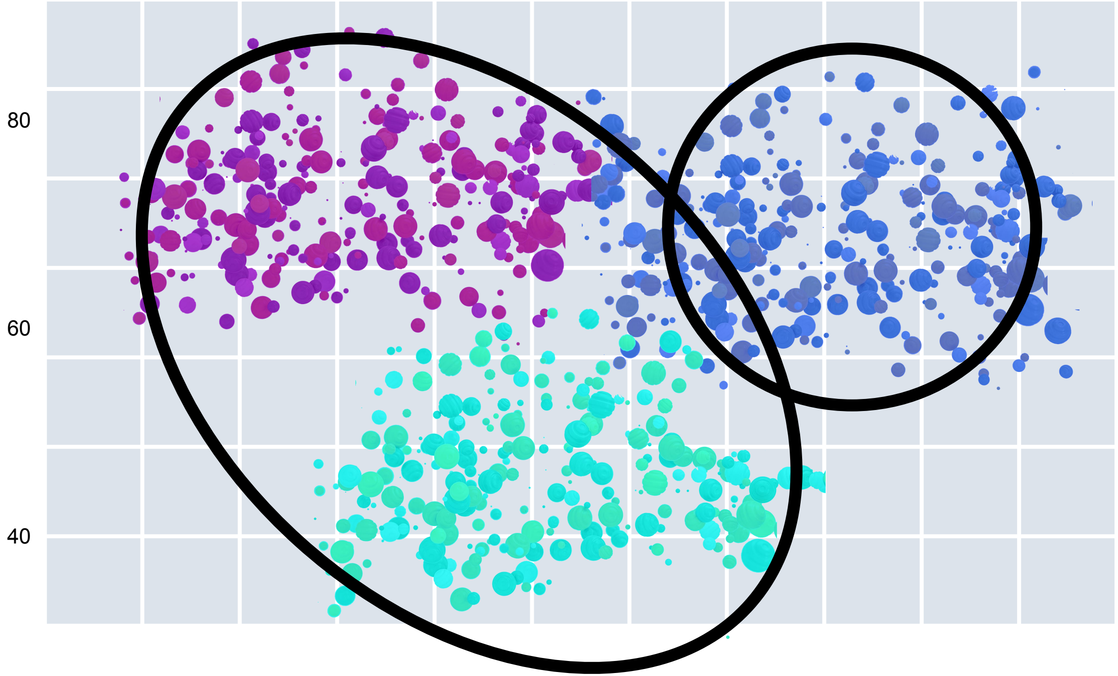
K=2



Machine Learning (Clustering)

Perhaps the clusters are not where we think they are.

K=2

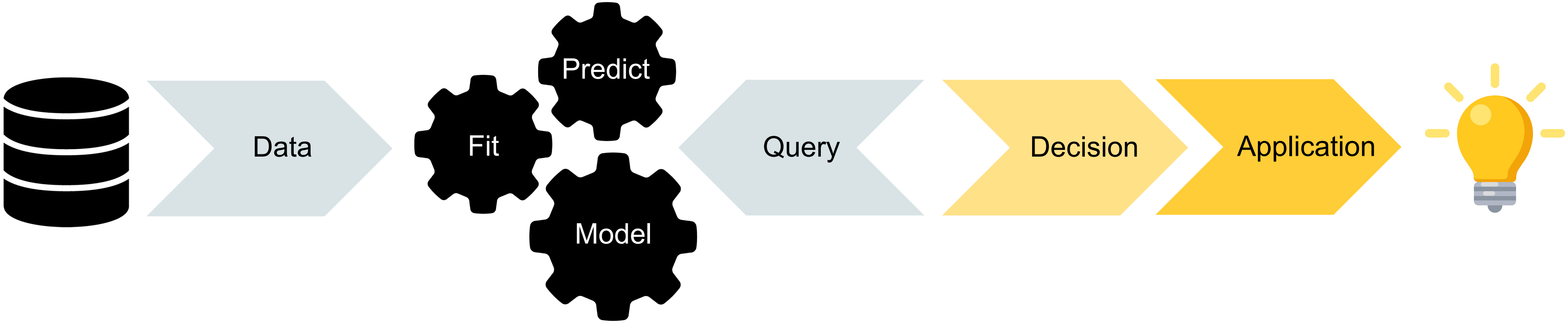


Training and Predicting

Regardless of the problem type, in Machine Learning we follow a familiar paradigm.

Model $\rightarrow$ Fit (Train) $\rightarrow$ Predict

Machine Learning



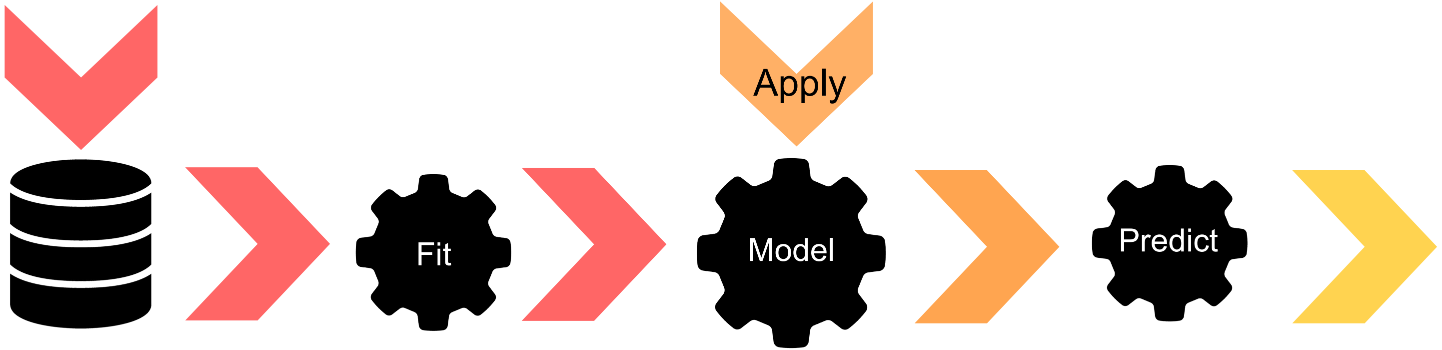
Training and Predicting

Regardless of the problem type, in Machine Learning we follow a familiar paradigm.

Model → Fit (Train) → Predict

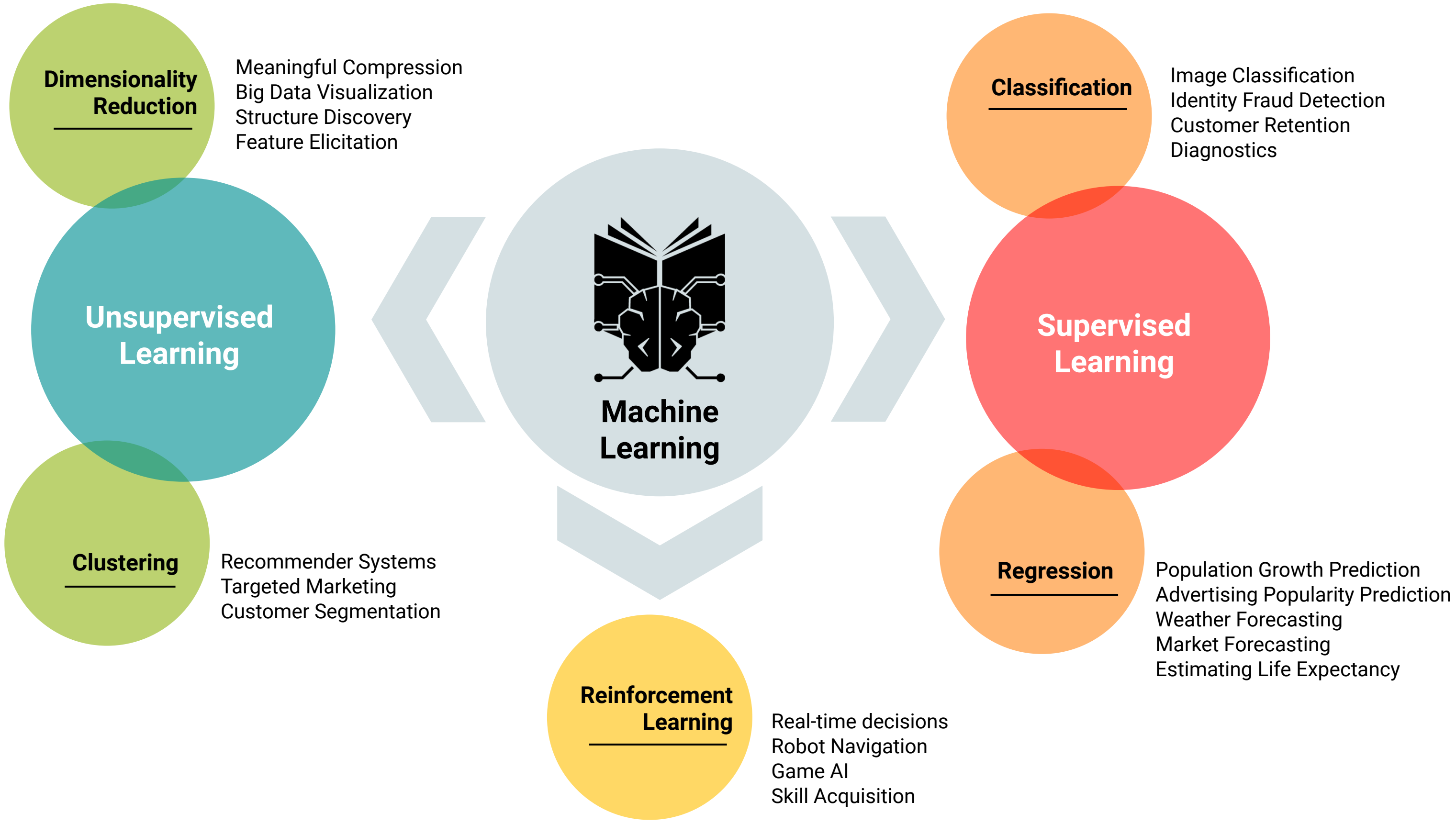
A	B	C	Class
11	16	22	1
10	8	4	2
…	…	…	…

A	B	C	Class
10	15	23	?



Class
1

Many Models: Which Do We Choose?



The Future of Machine Learning

Machine Learning will one day likely be relegated to a world of the off-the-shelf formulas readily available by the masses. We're already nearly there.

G2 | fx | =ROUND(C2,0)

	A	B	C	D	E	F	G
1	ROUND		No formatting		Formatted no decimal places		ROUND Formula
2			2.4		2		2.0
3			2.4		2		2.0
4			3.2		3		3.0
5			8.0		8		7.0

Questions?

Quantifying Machine Learning Models

Common Scoring Metrics

01

R^2 (R-Squared):

This is the baseline metric that many ML tools report on score. Higher R^2 values signify that the model is “highly predictive.” An R^2 value of >0.90 means that our model roughly accounts for 90% of the variability of the data.

02

MSE (Mean Squared Error):

This measures the average of the squares of the errors or deviations.

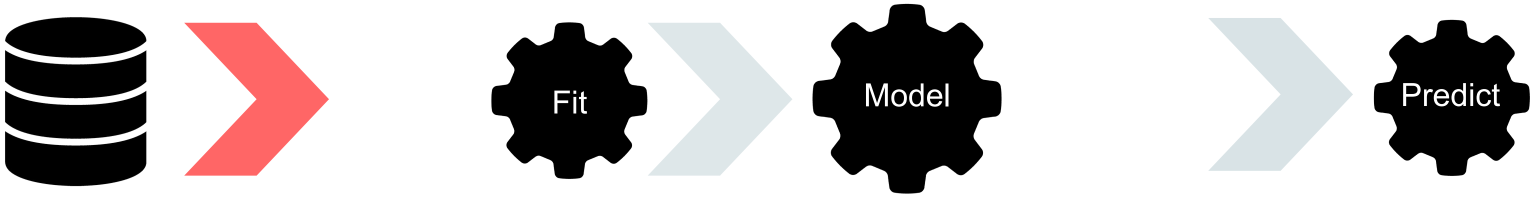
Basic Premise of Validation Using Training/Testing Data

We will cut a slice of this data (80%) to build our model, and then use this slice to predict the values for the remaining 20%.

Full Data Set (Historic)			
N=1000			
# bedrooms	# baths	Sq. feet (k)	Price (k)
2	1	1	200
3	2	1.5	250
...	...	...	...

Training Data Set			
N=800			
# bedrooms	# baths	Sq. feet (k)	Price (k)
4	3.5	3.2	450
2	2	1.5	220
...	...	...	...

Testing Data Set			
N=200			
# bedrooms	# baths	Sq. feet (k)	Price (k)
1	1	.5	60
5	3.5	4.2	780
...	...	...	...



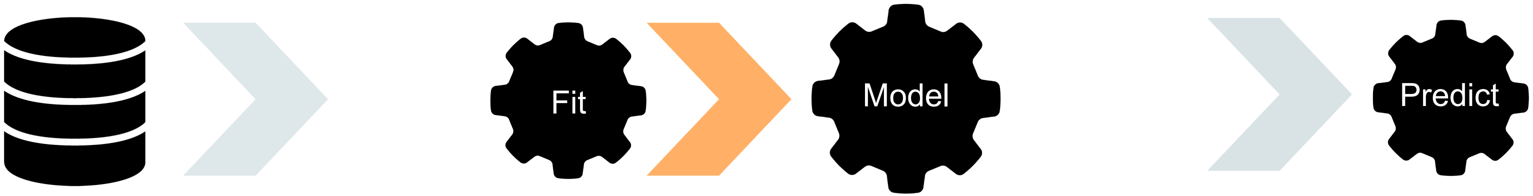
Basic Premise of Validation: Training

We use the training data to fit the model to the data. This is the training step where we build a model that can predict our output (home price) for a given set of features (# bedrooms, # baths, square feet). Once the model is trained, we can use the model to make predictions.

Full Data Set (Historic)			
N=1000			
# bedrooms	# baths	Sq. feet (k)	Price (k)
2	1	1	200
3	2	1.5	250
...	...	...	...

Training Data Set			
N=800			
# bedrooms	# baths	Sq. feet (k)	Price (k)
4	3.5	3.2	450
2	2	1.5	220
...	...	...	...

Testing Data Set			
N=200			
# bedrooms	# baths	Sq. feet (k)	Price (k)
1	1	.5	60
5	3.5	4.2	780
...	...	...	...



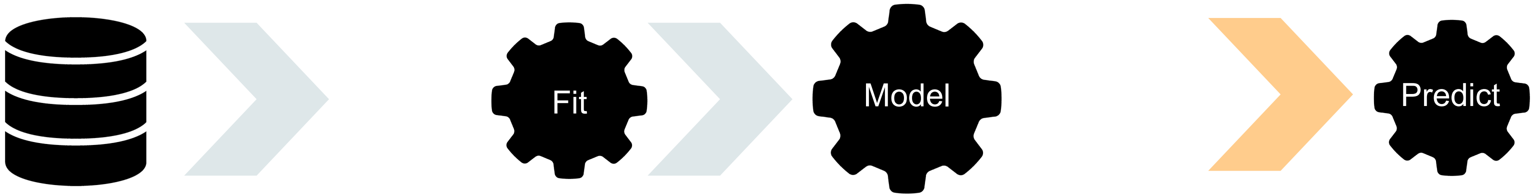
Basic Premise of Validation

We use the test data to make new home price predictions. We can then compare the home price of our prediction vs. the actual price. Based roughly on how often we are "correct," we get a score for the model as a whole. If the model scores well, we can trust it for future use. We train the model on the training data and score the model based on data that it has never seen before (test data).

Full Data Set (Historic)			
N=1000			
# bedrooms	# baths	Sq. feet (k)	Price (k)
2	1	1	200
3	2	1.5	250
...	...	...	...

Training Data Set			
N=800			
# bedrooms	# baths	Sq. feet (k)	Price (k)
4	3.5	3.2	450
2	2	1.5	220
...	...	...	...

Testing Data Set			
N=200			
# bedrooms	# baths	Sq. feet (k)	Price (k)
1	1	.5	60
5	3.5	4.2	780
...	...	...	...



<Time to Code>