Introduction to Machine Learning

CptS 437

Spring 2022

Dr. Diane J. Cook

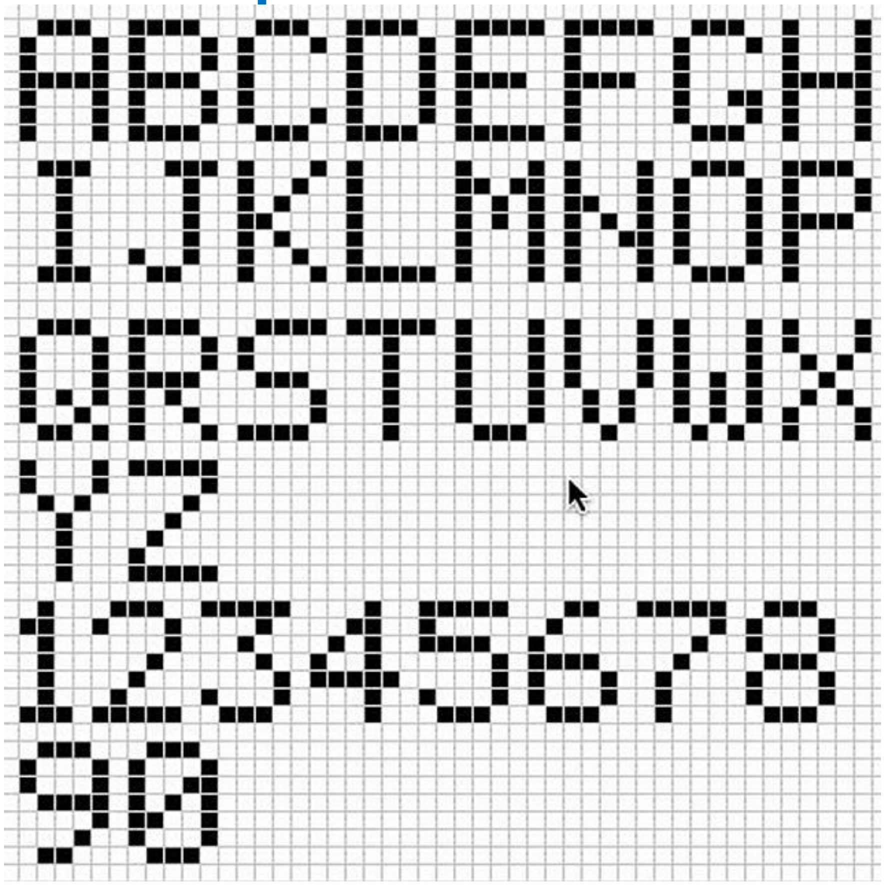
Why study machine learning?

- Data explosion
- Computer giants working in this field
- Harvard Business Review: Sexiest job of the 21st century

What is machine learning?

- A computer program that improves its performance at some task through experience

Example - Handwriting recognition



0.0	1.0	1.0	1.0	0.0
1.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	1.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0
1.0	1.0	1.0	1.0	1.0

Can this be automated?

https://webdemo.myscript.com/#/demo/write

0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0

0.0	0.0	1.0	0.0	0.0
0.0	1.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0

If (cell# mod 5) = 3 and cell value = 1.0

and

All other cell values = 0.0

?

0.0	0.0	1.0	0.0	0.0
0.0	1.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	1.0	1.0	1.0	0.0

If (cell# mod 5) = 3 and cell value = 1.0

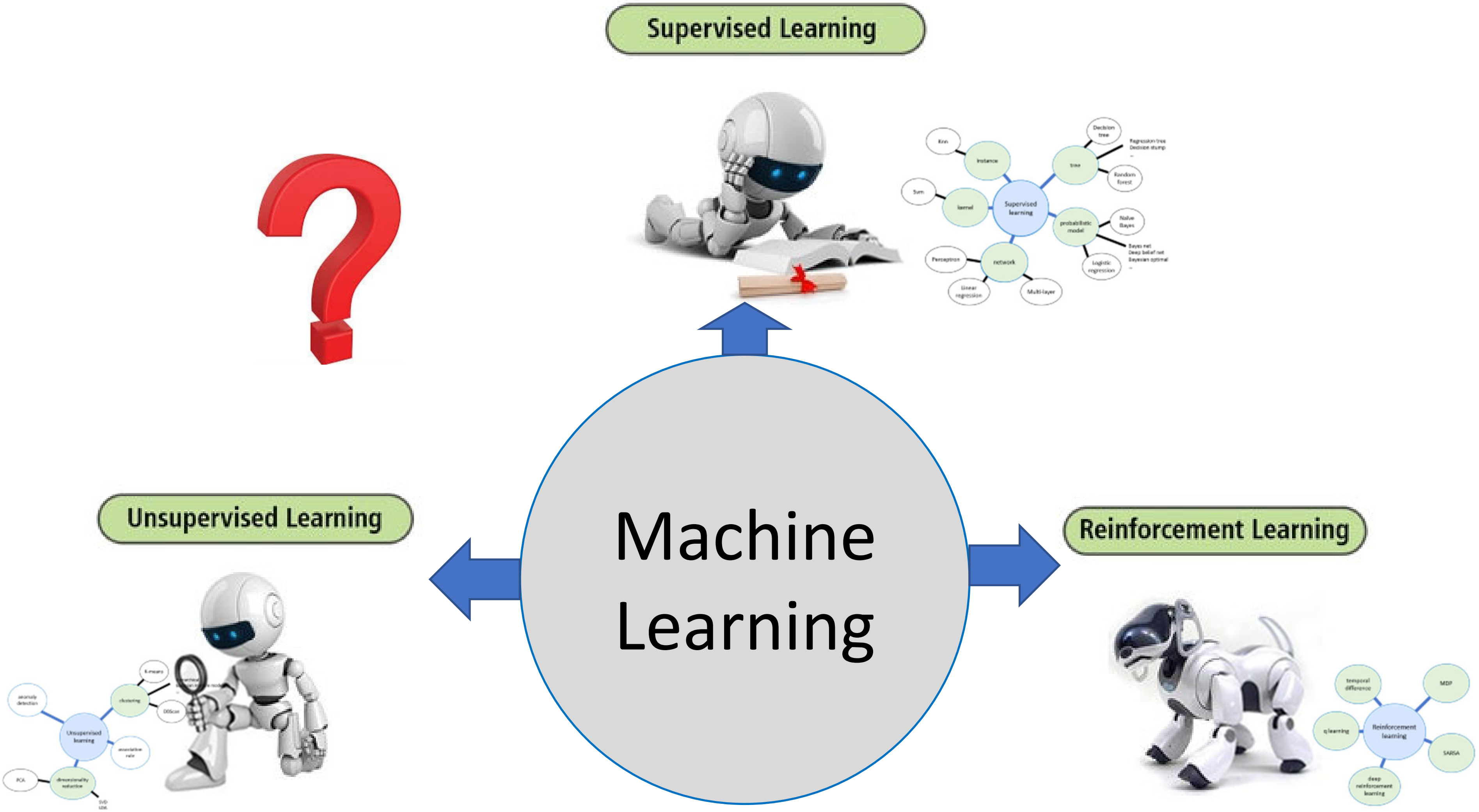
and

(If (cell# = 7) and cell value = 0.0 or 1.0

and

All other cell values = 0.0)

0.0	0.0	0.8	0.0	0.0
0.0	0.0	0.7	0.0	0.0
0.0	0.0	0.8	0.0	0.0
0.0	0.0	0.7	0.0	0.0
0.5	0.0	0.7	0.0	0.0
0.0	0.0	0.5	0.0	0.0
0.0	0.0	0.3	0.0	0.0

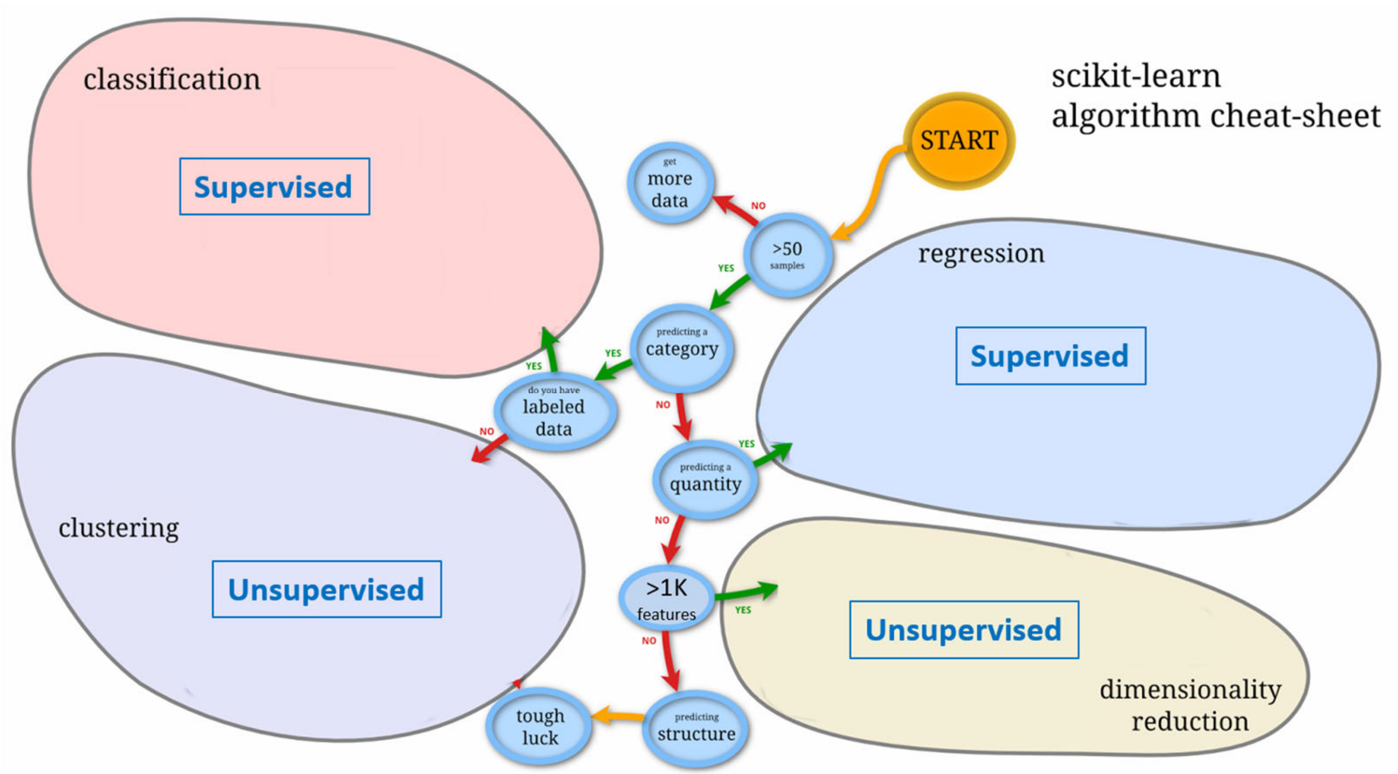


Machine Learning

Supervised Learning

Unsupervised Learning

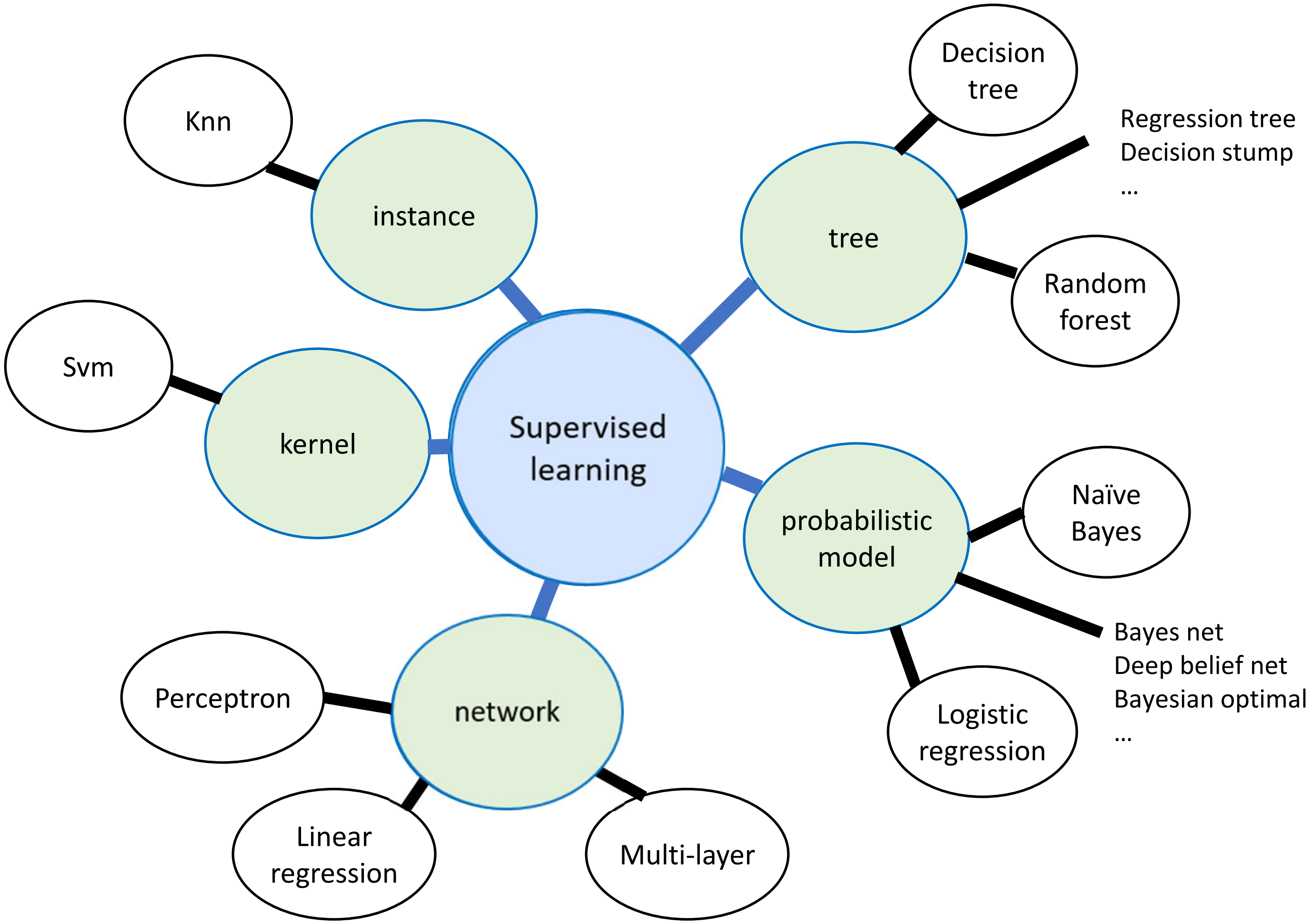
Reinforcement Learning



Supervised Learning

?

Unsupervised Learning



Supervised learning

instance

Knn

kernel

Svm

tree

Decision tree

Regression tree
Decision stump
…

Random forest

probabilistic model

Naïve Bayes

Bayes net
Deep belief net
Bayesian optimal
…

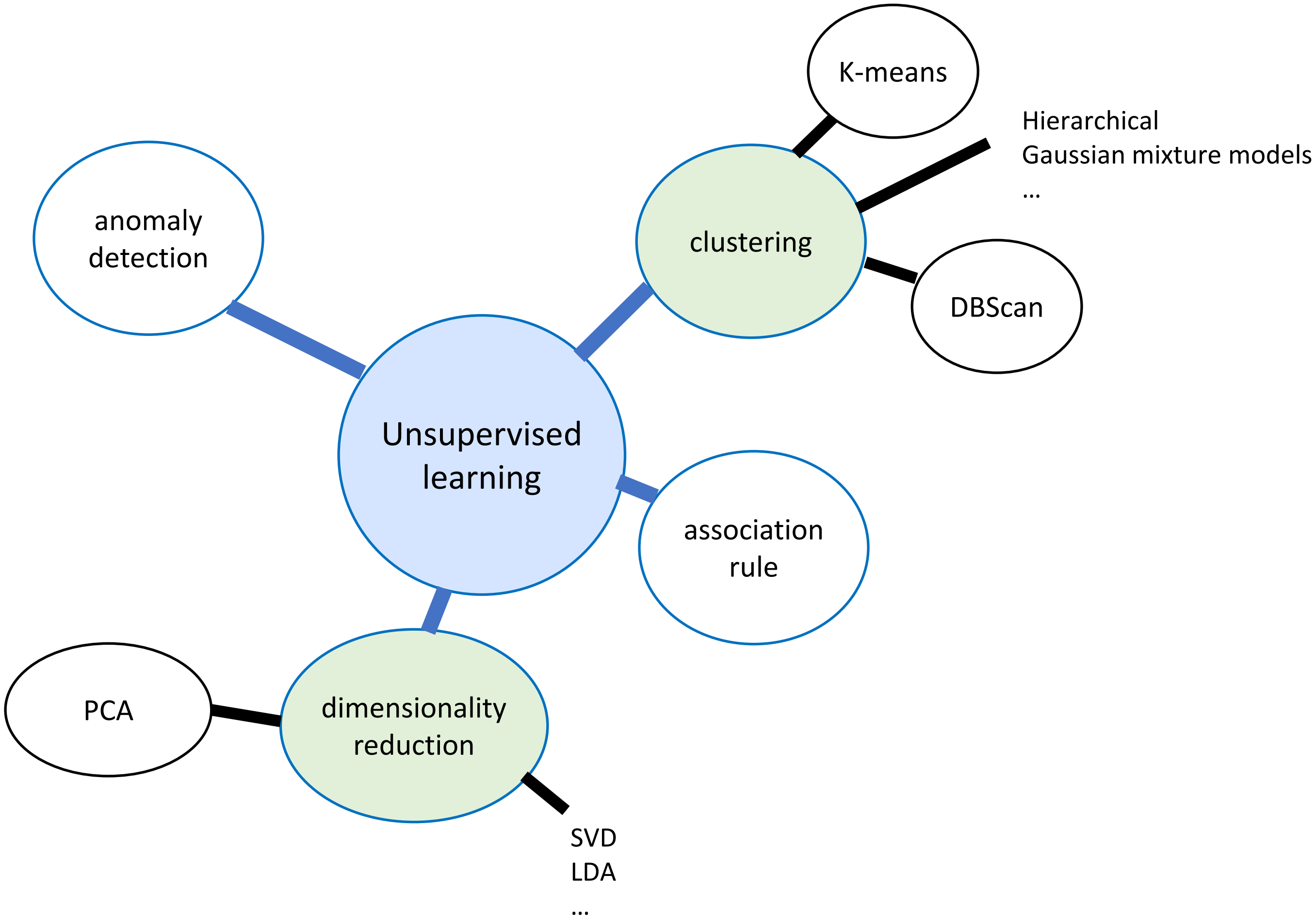
Logistic regression

network

Perceptron

Linear regression

Multi-layer



Unsupervised learning

anomaly detection

clustering

K-means

Hierarchical
Gaussian mixture models
…

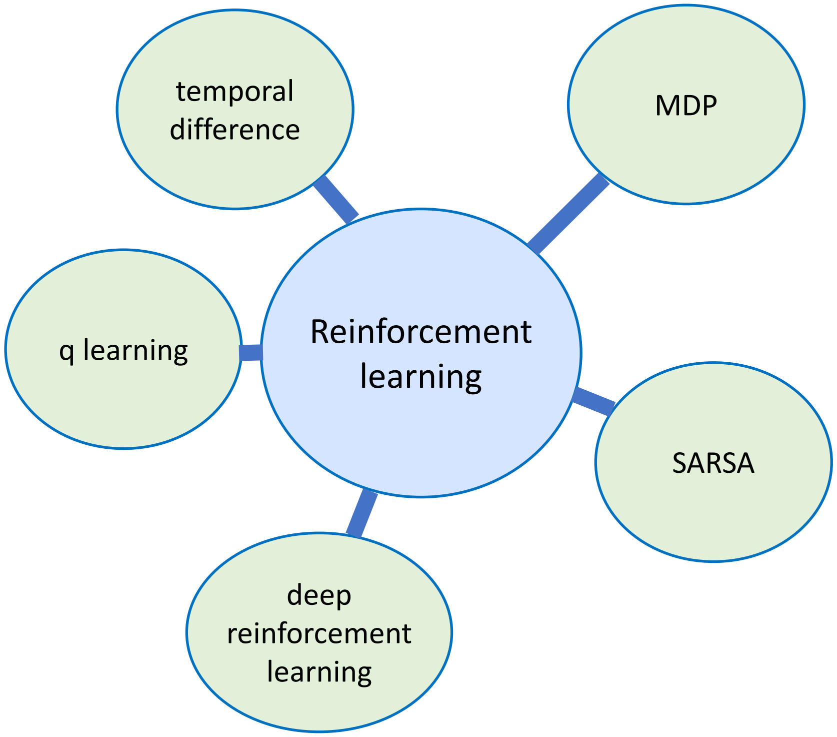
DBScan

association rule

dimensionality reduction

PCA

SVD
LDA
…



Reinforcement learning

temporal difference

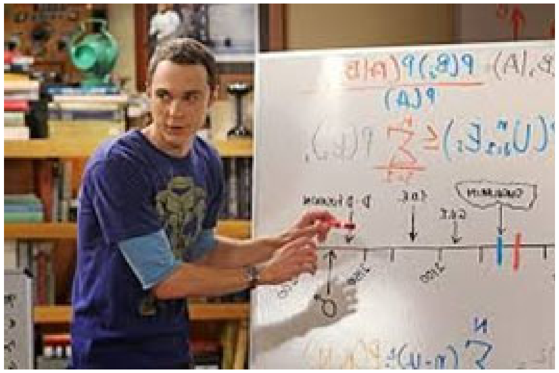
MDP

q learning

SARSA

deep reinforcement learning

How do we determine if learning has occurred?



Syllabus

Example (with decision tree)

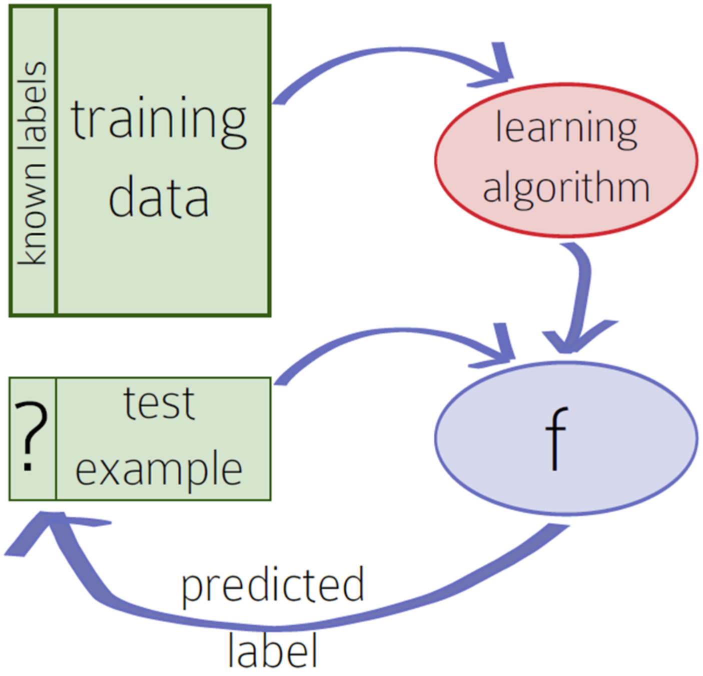
- Predict whether a home is in San Francisco or New York
- Given a set of past examples

Data representation

- Features
- elevation, price per sq ft, year built, #bathrooms, #bedrooms, square feet, price
- Feature values
- Class value (label)

elevation: 50’
price: $1777/sqft
year built: 1920
#bathrooms: 2
#bedrooms: 3
square feet: 800
price: $1,421,600
label: New York

Induction framework

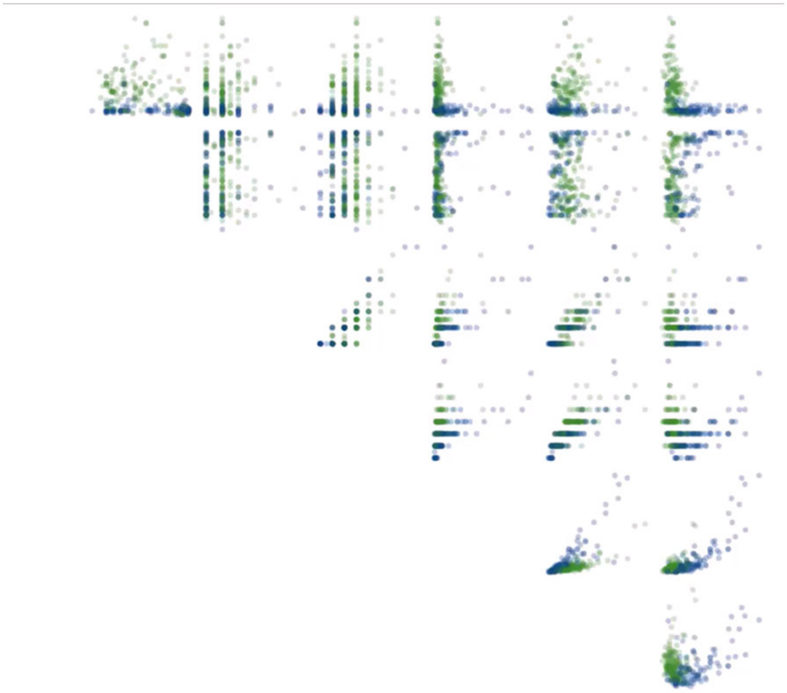


Types of inductive learning problems

- Regression
- Binary classification
- Multiclass classification
- Discovery
- Reinforcement learning

Supervised learning

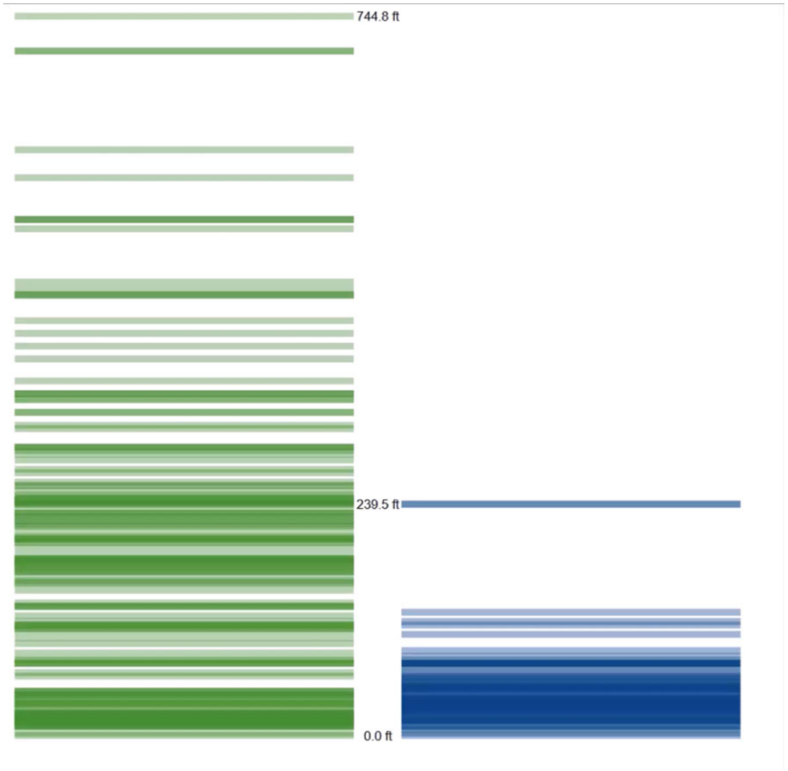
- Learning task
  - Learn to classify whether a home is in San Francisco or New York
  - Represent each home by elevation



Supervised learning

- Learning task
  - Learn to classify whether a home is in San Francisco or New York
  - Represent each home by elevation feature

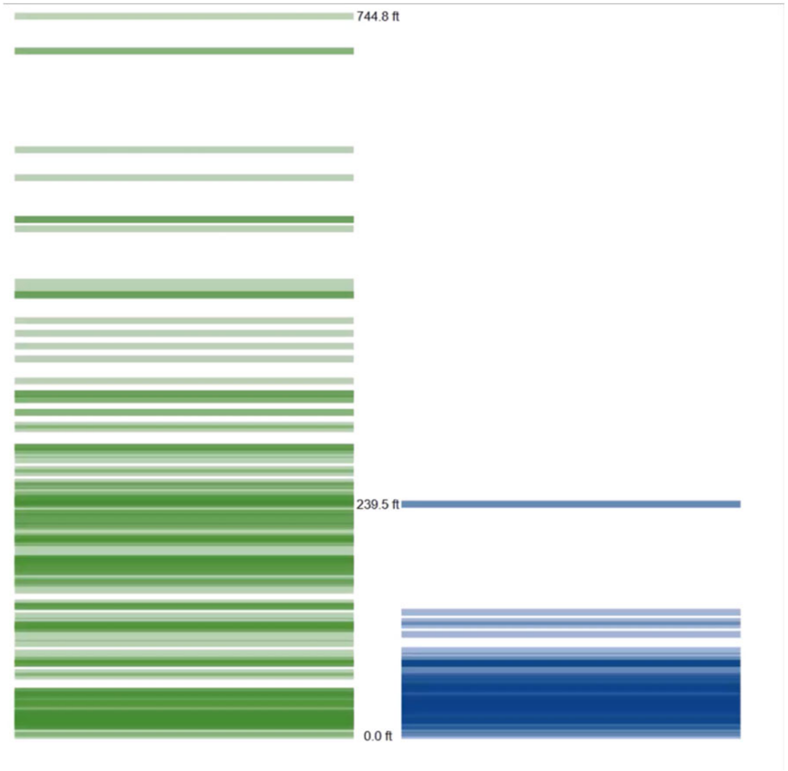
If elevation > 240’ then San Francisco



Supervised learning

- Add another dimension
  - Home is in San Francisco or New York
  - New feature: price per sq ft

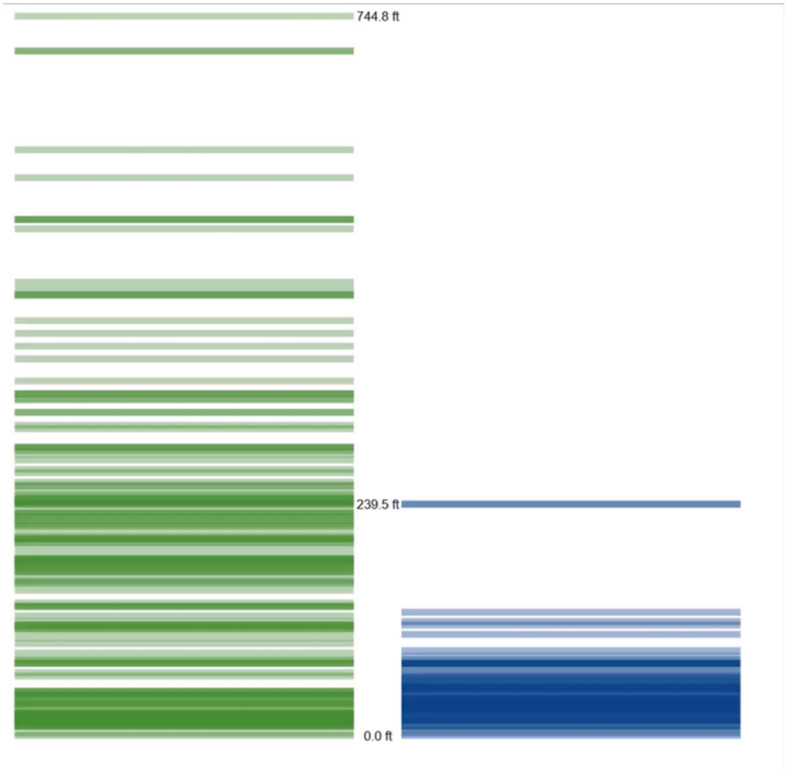
If elevation > 240’ then San Francisco



Supervised learning

- Add another dimension
  - Home is in San Francisco or New York
  - New feature: price per sq ft

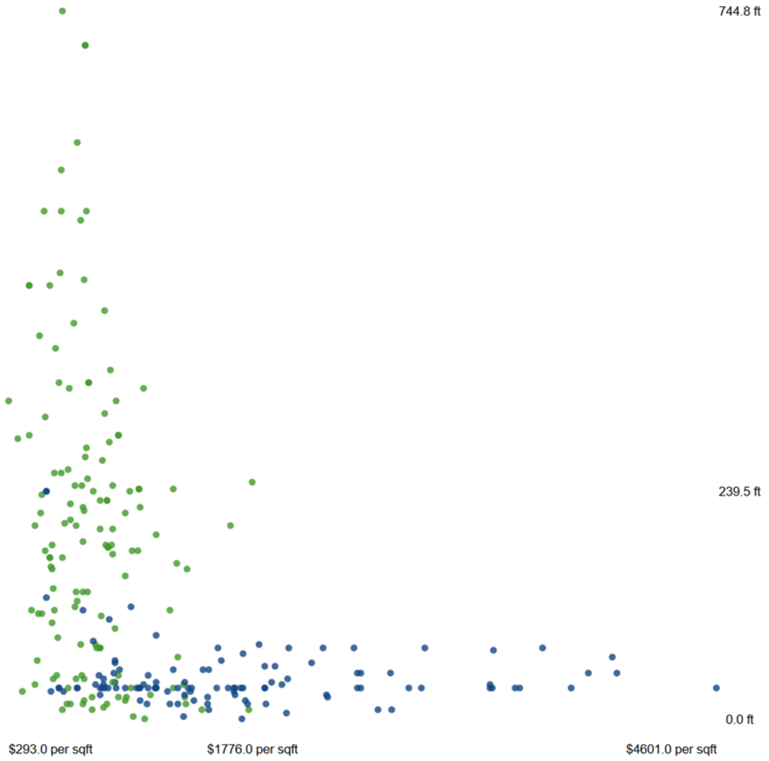
If elevation > 240'
    then San Francisco
else if price >$1777/sqft
    then New York



Drawing boundaries

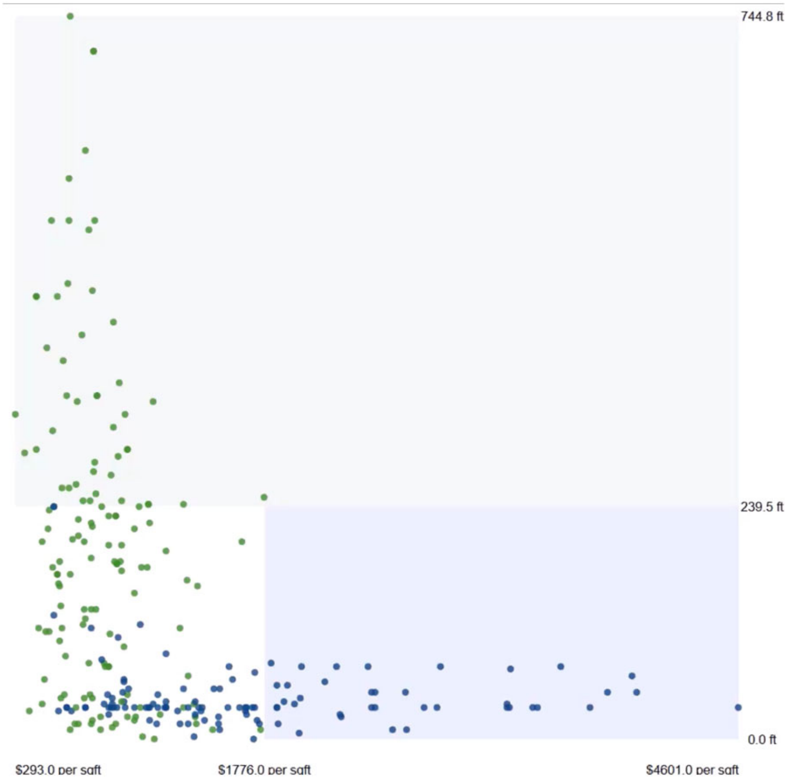
- Visualize rule as boundaries of regions in scatterplot

If elevation > 240’
    then San Francisco
else if price >$1777/sqft
    then New York

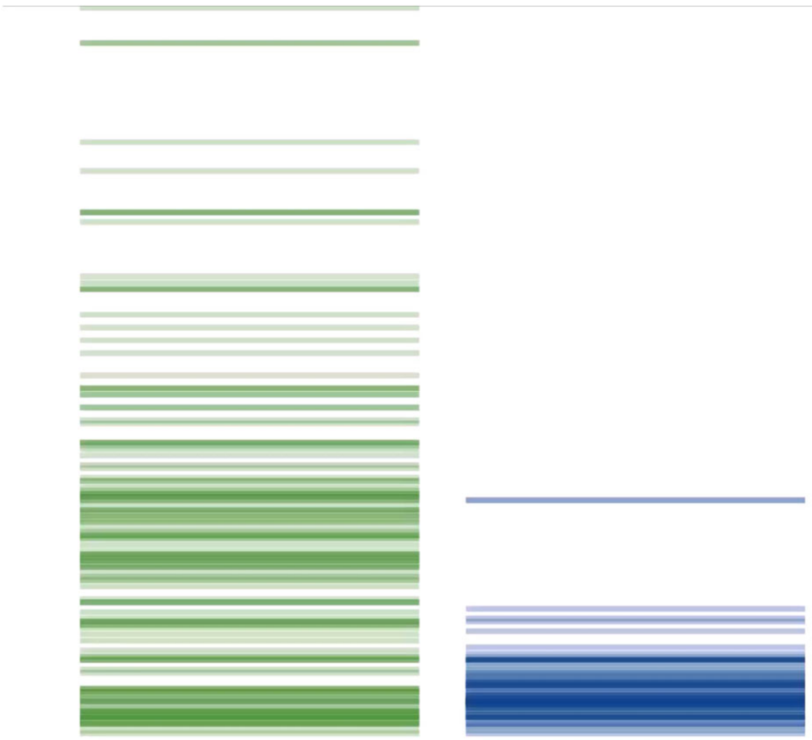


All features

- Seven dimensions

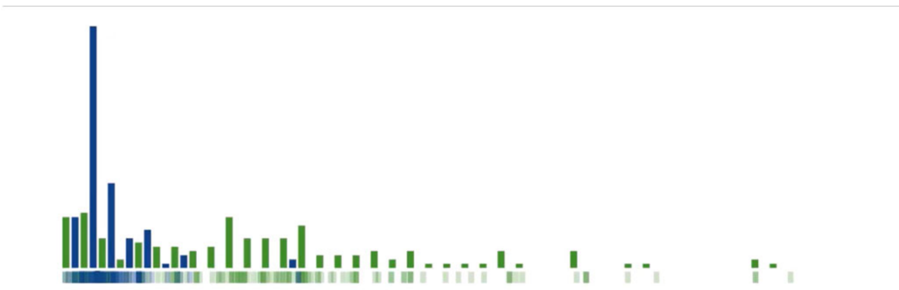


Classifier - Decision tree



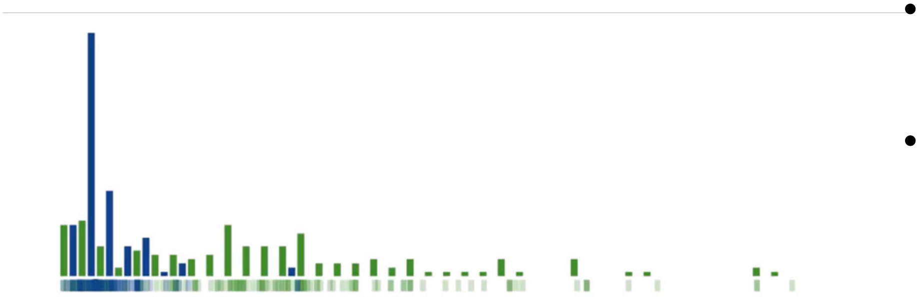
- View elevation histogram

Classifier - Decision tree



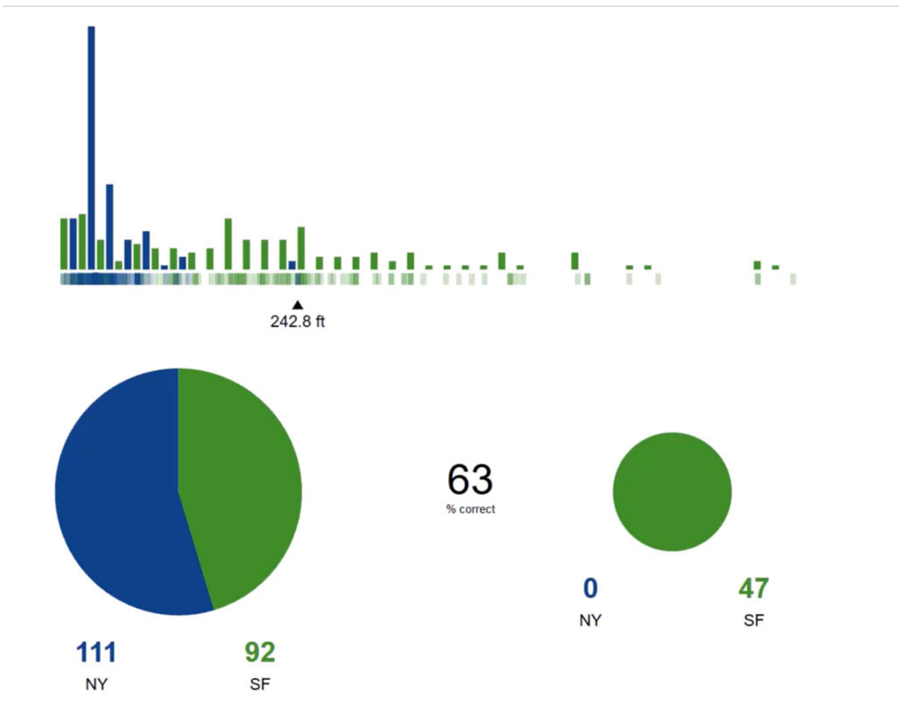
- Decision tree uses if-then statements to form boundaries
- If elevation > x then home in San Francisco
- These statements are represented by nodes in a decision tree
- Each node splits the data into branches based on feature values

Classifier - Decision tree



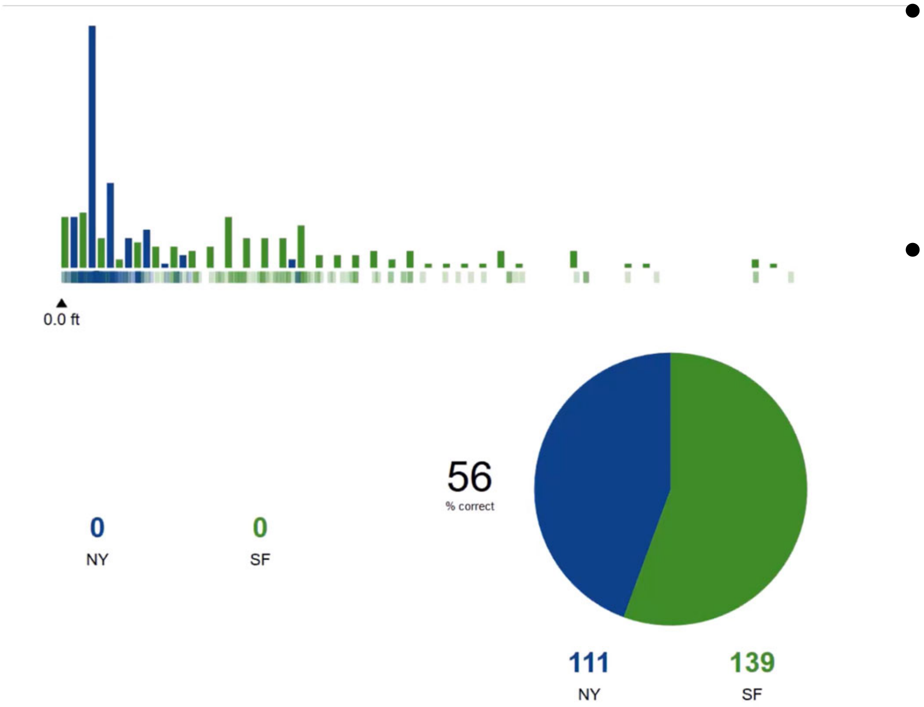
- If elevation > 240’ then home in San Francisco
- This split incorrectly classifies some San Francisco homes as New York homes
- Accuracy is 63% correct
- All the green incorrect labels are false negatives

Classifier - Decision tree



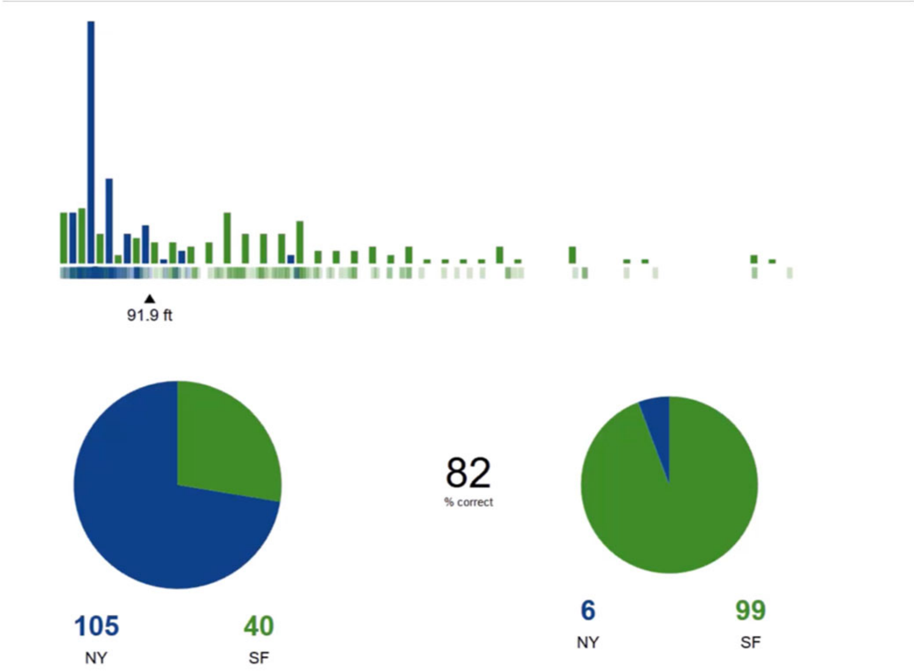
- If elevation > 0’ then home in San Francisco
- If we try to capture every San Francisco home, we will include New York homes
- These will be false positives

Classifier - Decision tree



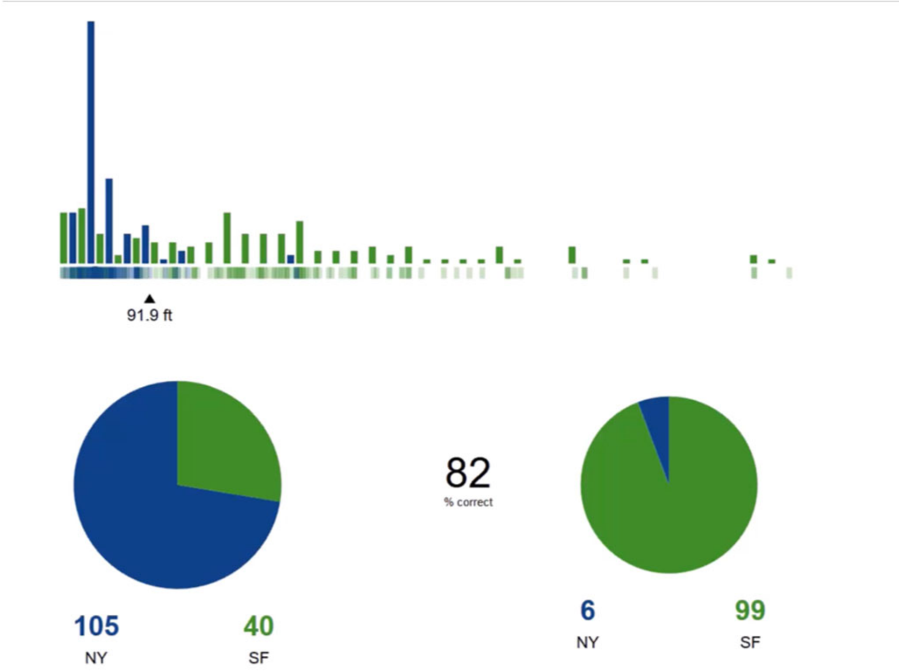
- The best split makes the groups as homogeneous as possible
- If elevation > 92’ then home in San Francisco

Classifier - Decision tree



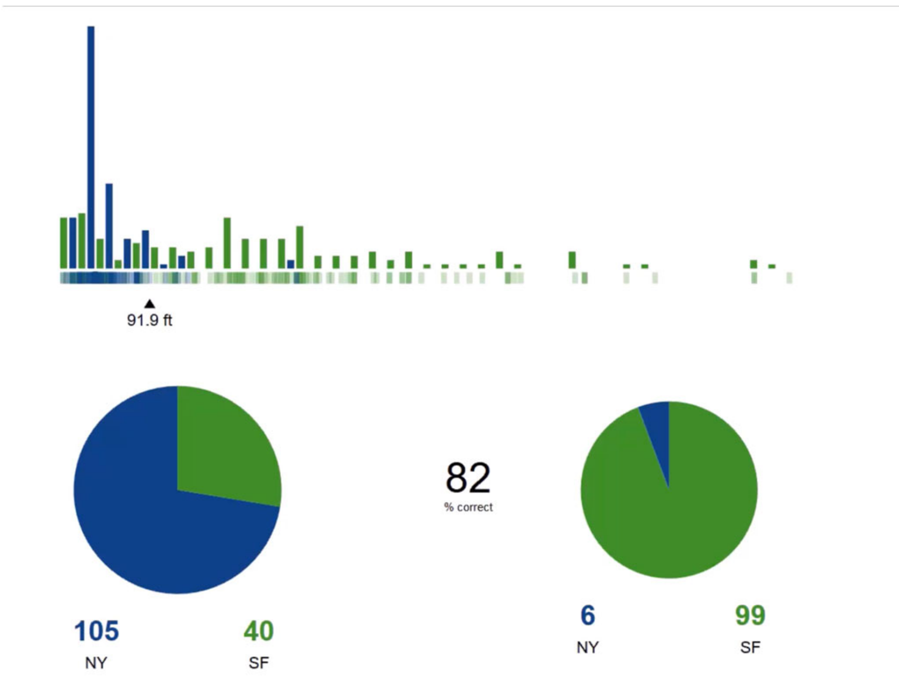
- Even the best split does not fully separate the classes

Classifier - Decision tree



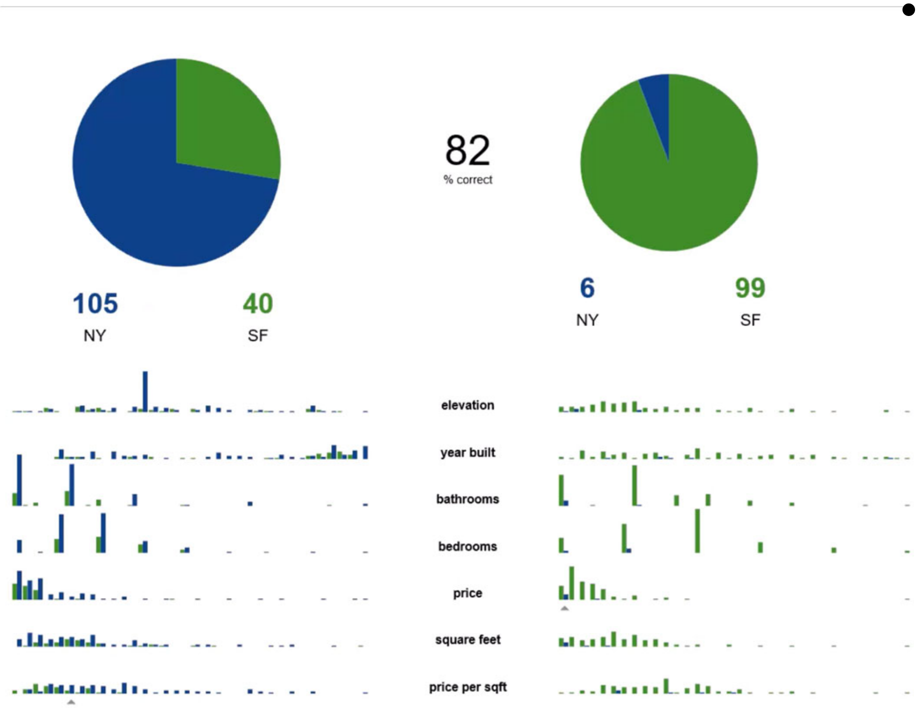
- Solution? Add another split point
- Repeat process on subsets of data
- Recursion

Classifier - Decision tree



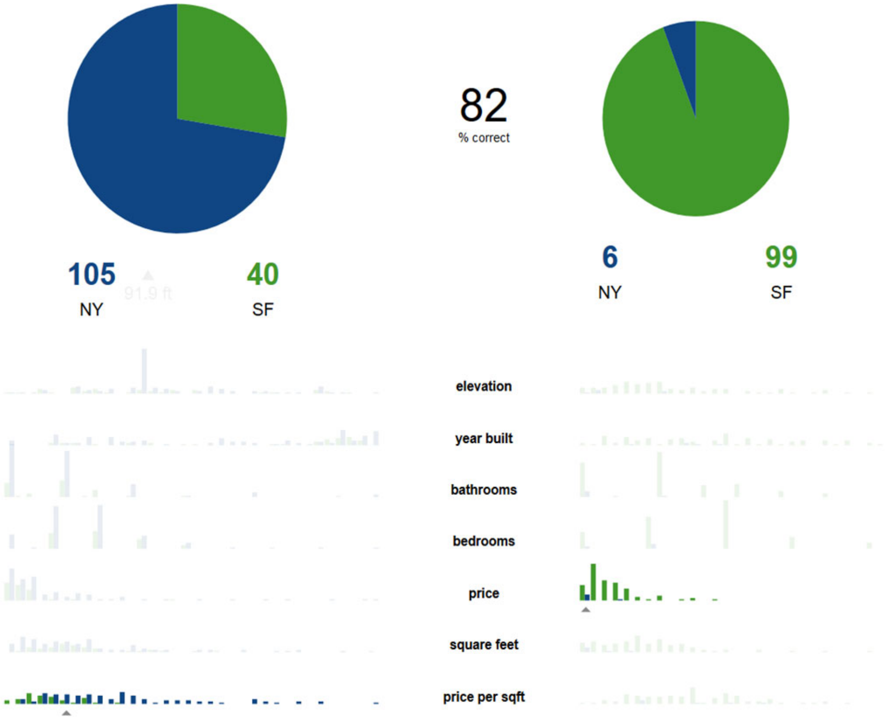
- Consider distribution for each subset

Classifier - Decision tree



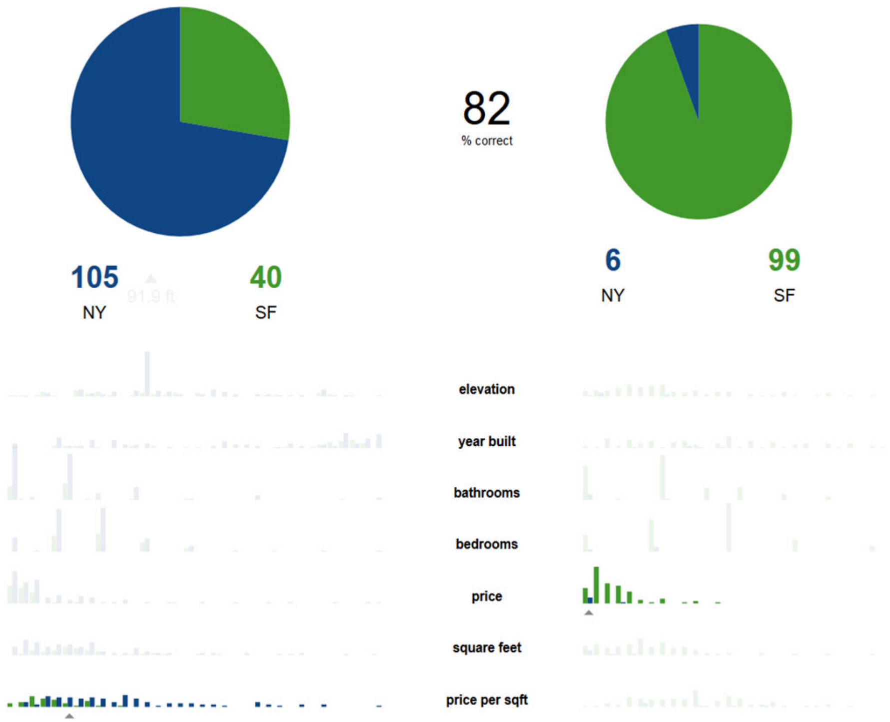
- Consider distribution for each subset

Classifier – Decision tree



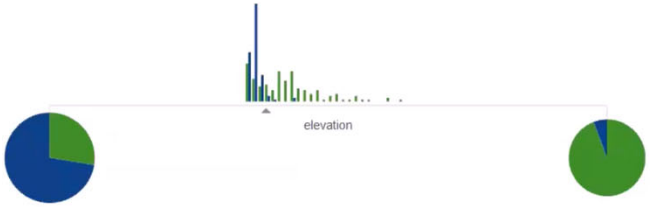
- Best split varies for each subset
- Lower elevation homes
  - Best split variable is price per square foot ($1061)
- Higher elevation homes
  - Best split variable is price of home ($514,500)

Classifier - Decision tree

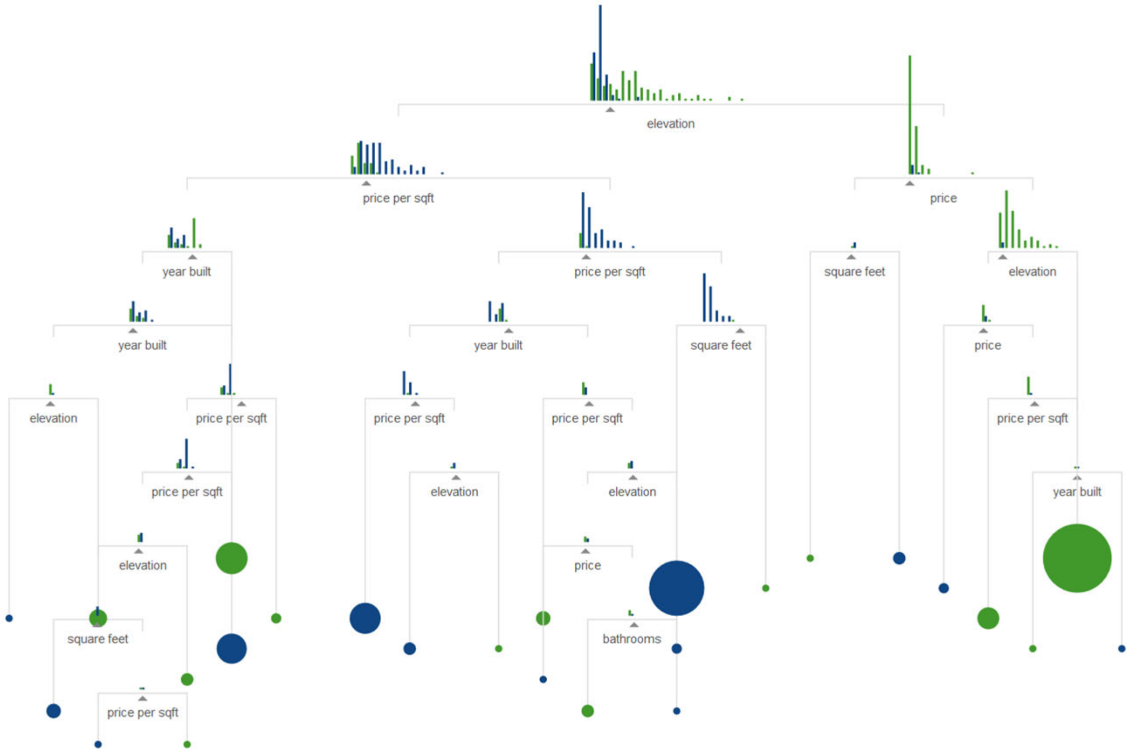


- Additional nodes add new rule details
- This can increase the tree’s accuracy

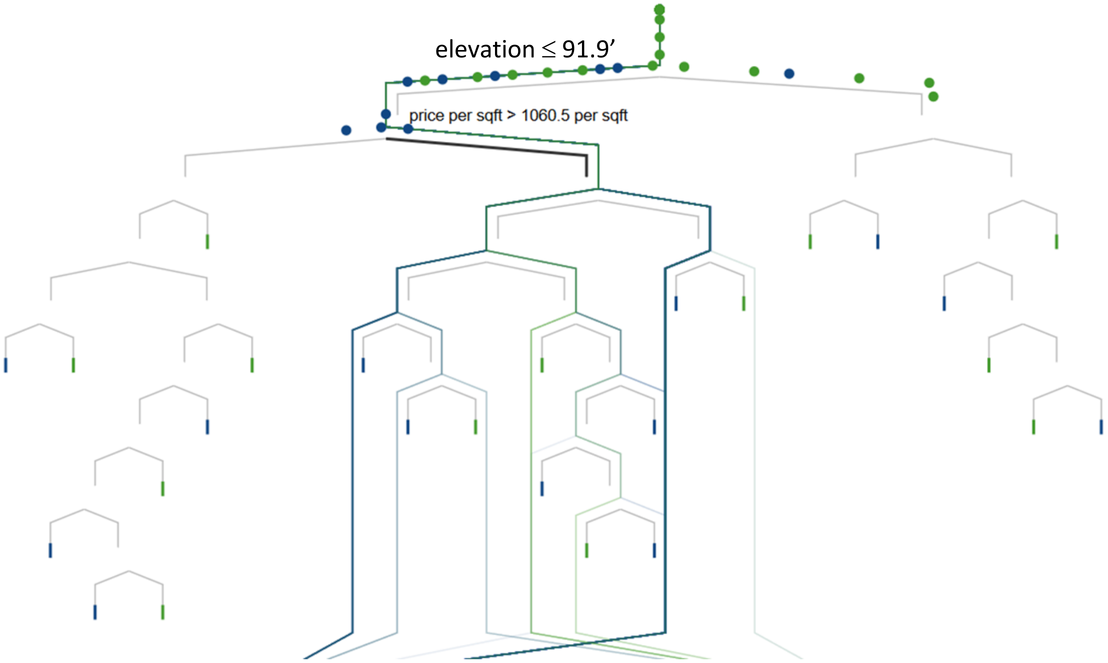
Classifier - Decision tree



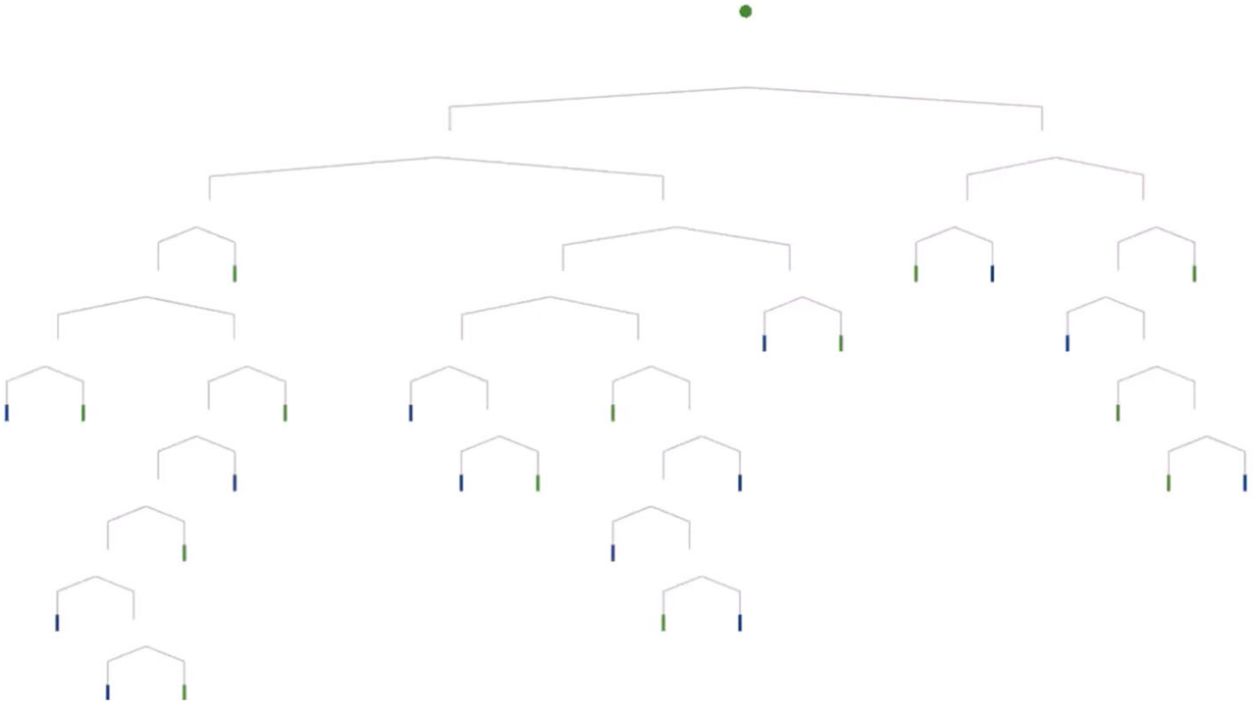
Predict class of new data point



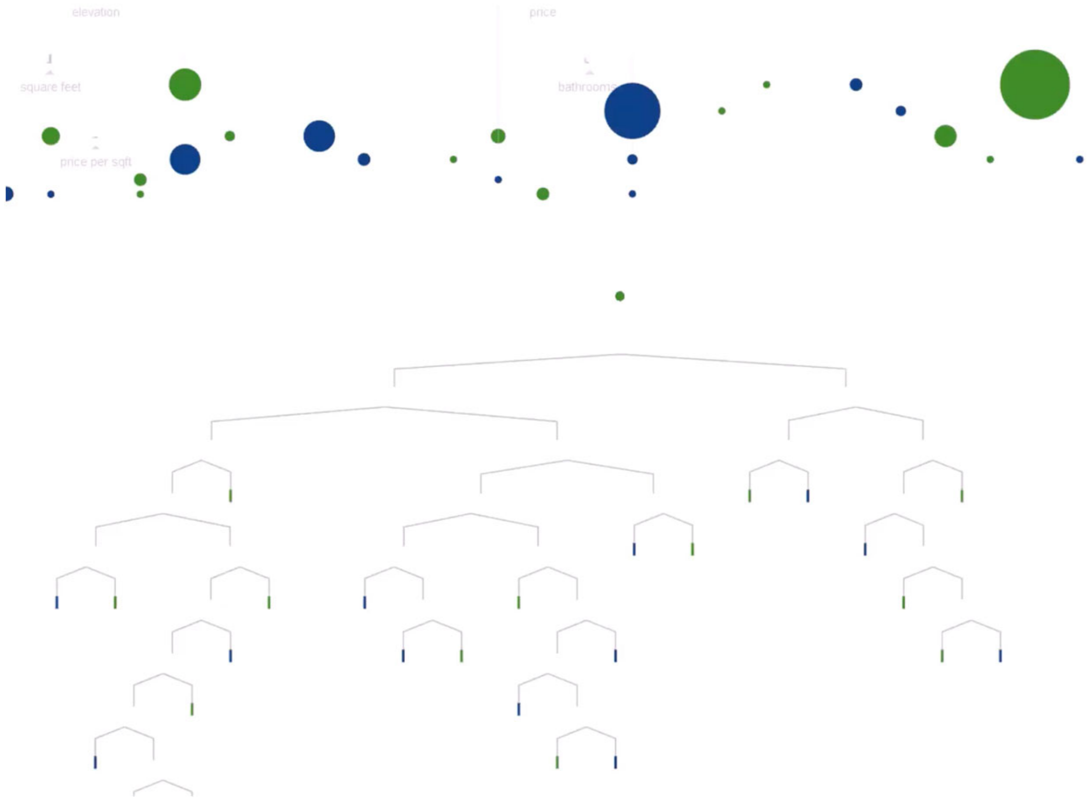
Predict using decision tree



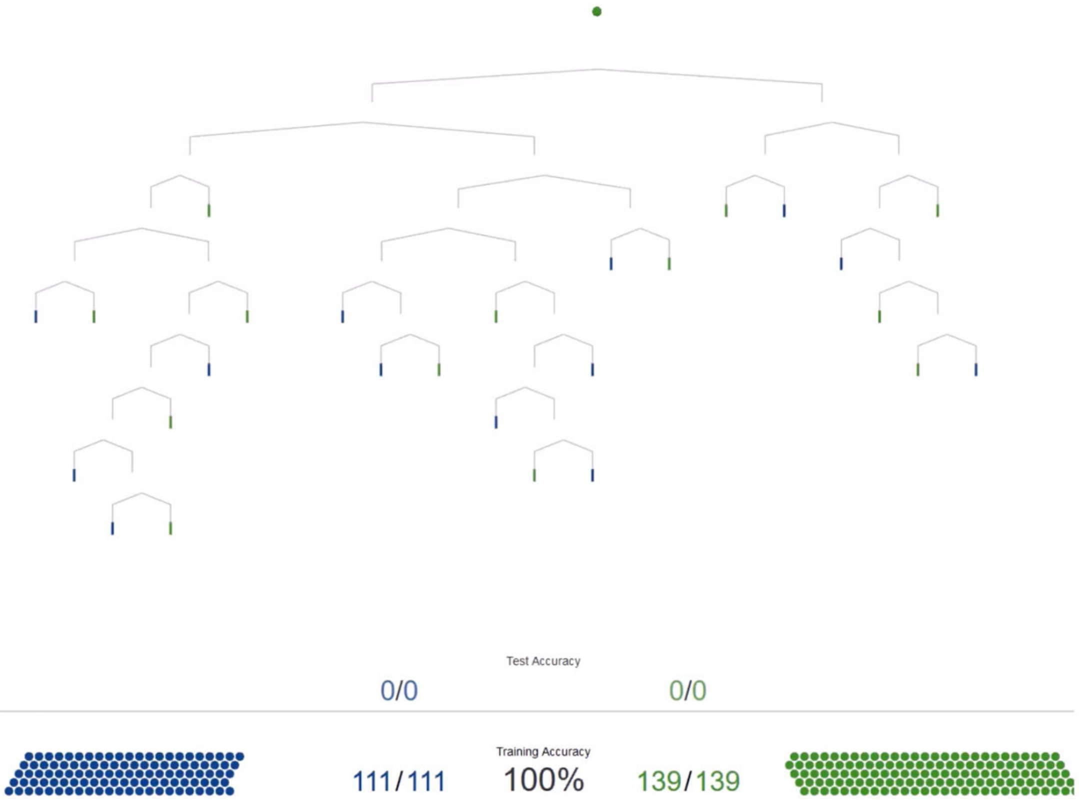
Predict using decision tree



Prediction accuracy



Bigger question - Prediction accuracy on new test data



Performance: loss function

- L(.,.)
- Regression
- Binary classification
- Multiclass classification
- Discovery
- Reinforcement learning

Review