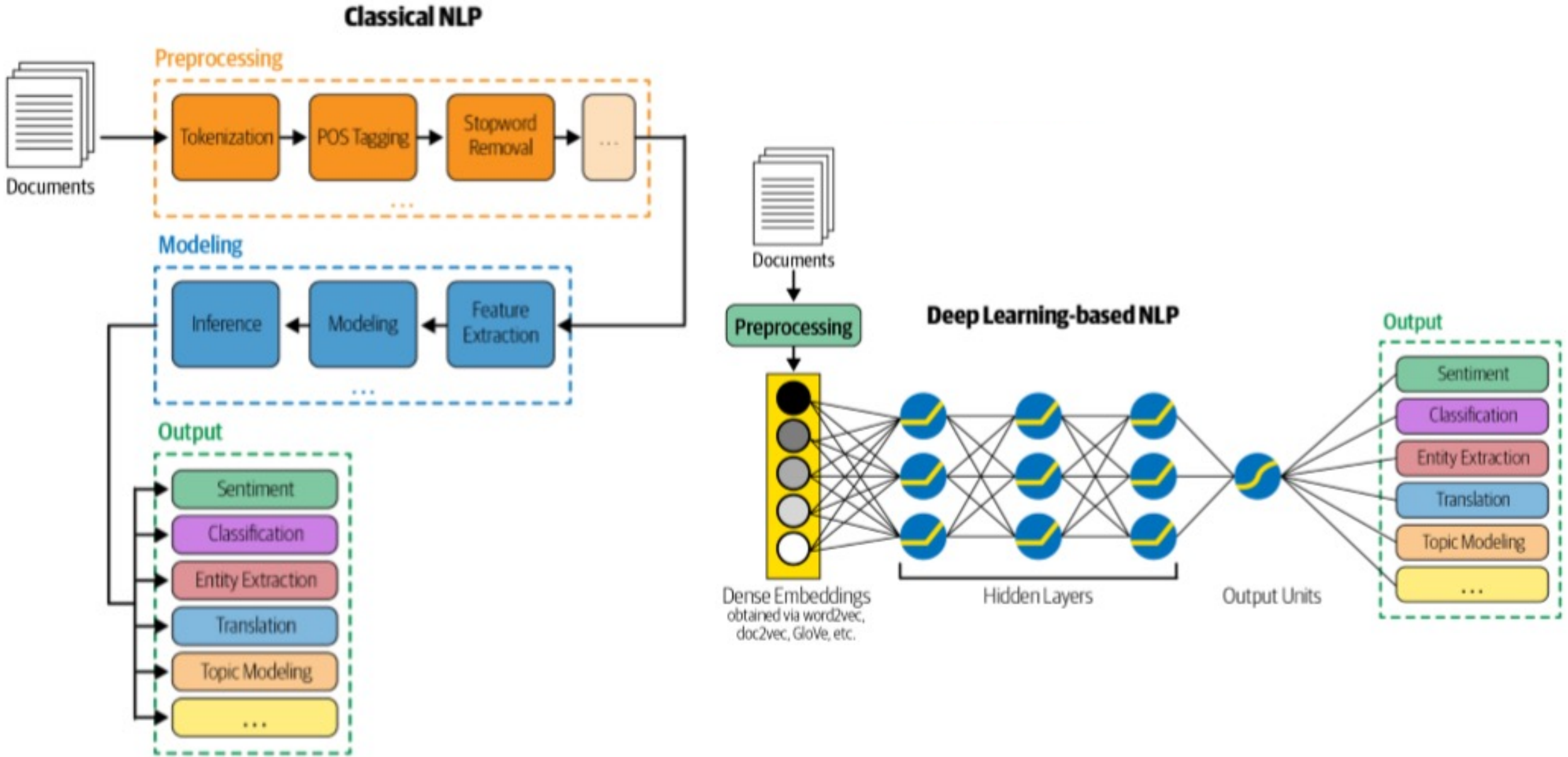
Topic 8
Classification Part 2


Classical NLP
Preprocessing
Documents
Tokenization
POS Tagging
Stopword Removal
...
Modeling
Inference
Modeling
Feature Extraction
...
Output
Sentiment
Classification
Entity Extraction
Translation
Topic Modeling
...
Documents
Preprocessing
Deep Learning-based NLP
Dense Embeddings
obtained via word2vec, doc2vec, GloVe, etc.
Hidden Layers
Output Units
Output
Sentiment
Classification
Entity Extraction
Translation
Topic Modeling
...

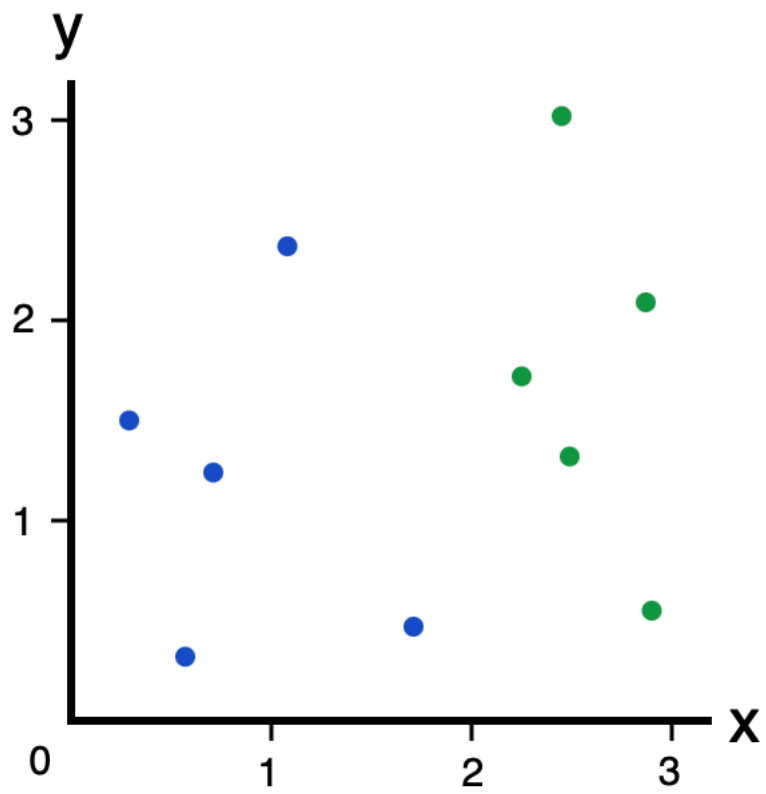
Many other options for classification

- SVM
- Logistic Regression
- KNN
- Decision tree
- Random Forest
- XGBoost
- etc.

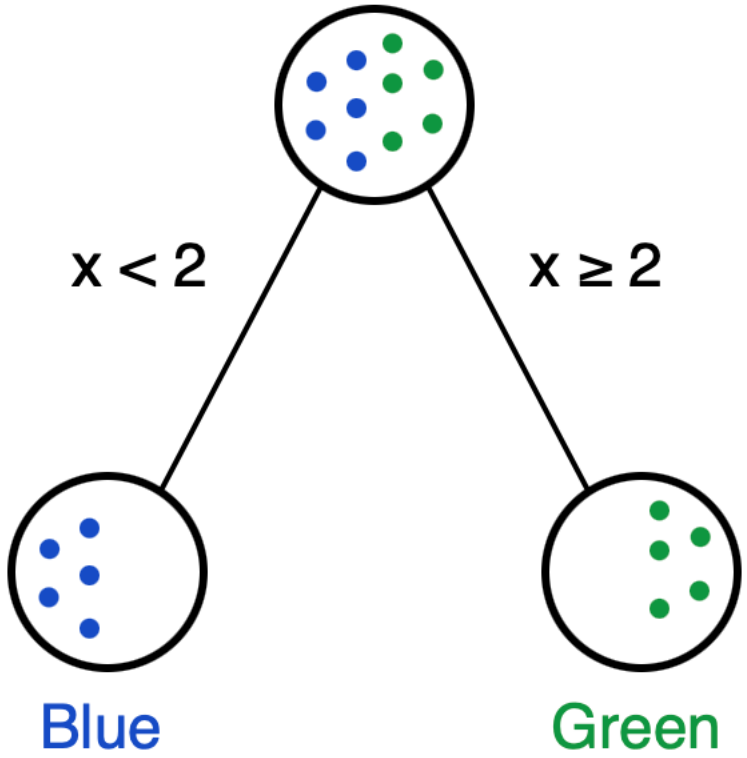
Decision Trees

- Decision trees are a popular machine learning algorithm for classification and regression tasks
- They are easy to interpret and visualize, and can handle both categorical and numerical features
- However, they can easily overfit to the training data, leading to poor performance on new data

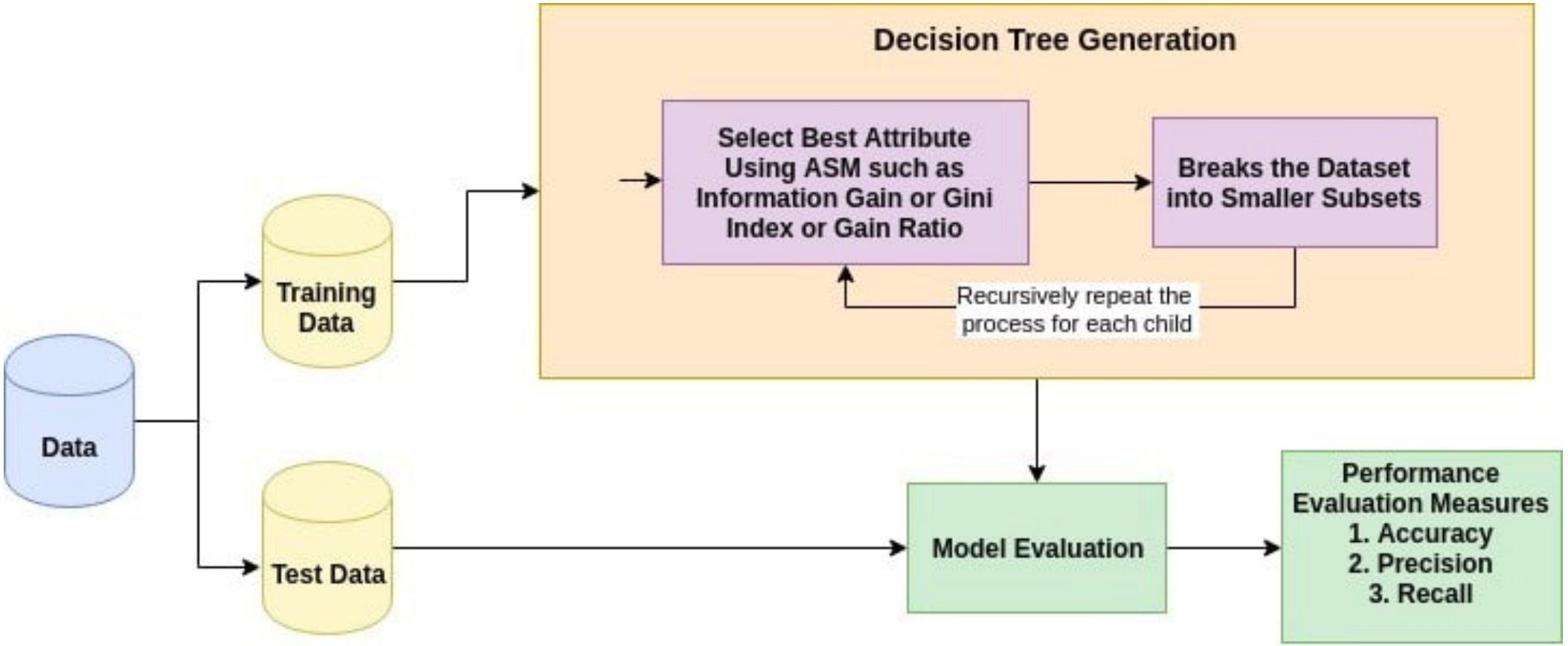
Decision Trees



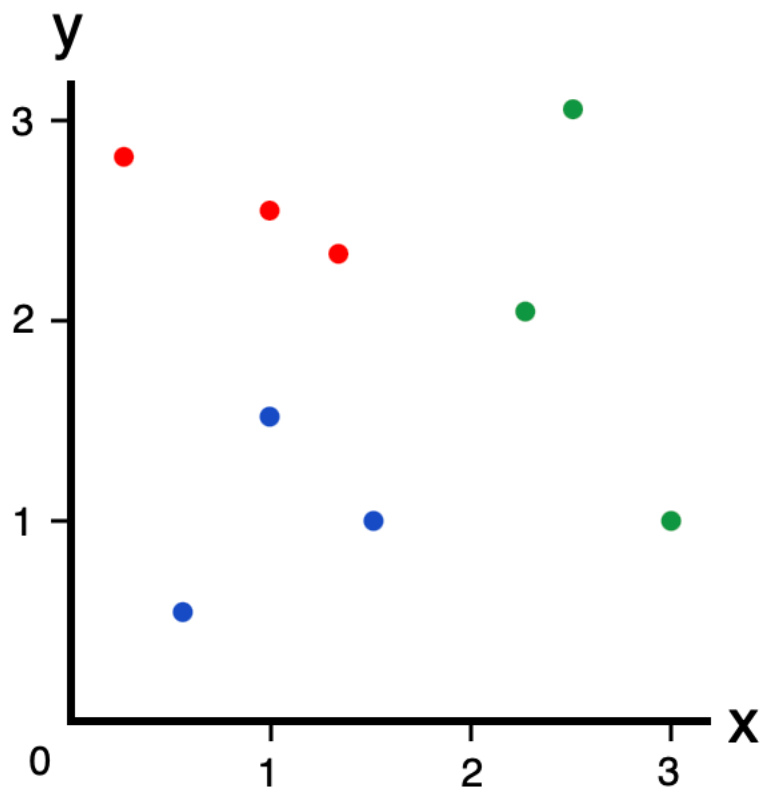
The Dataset

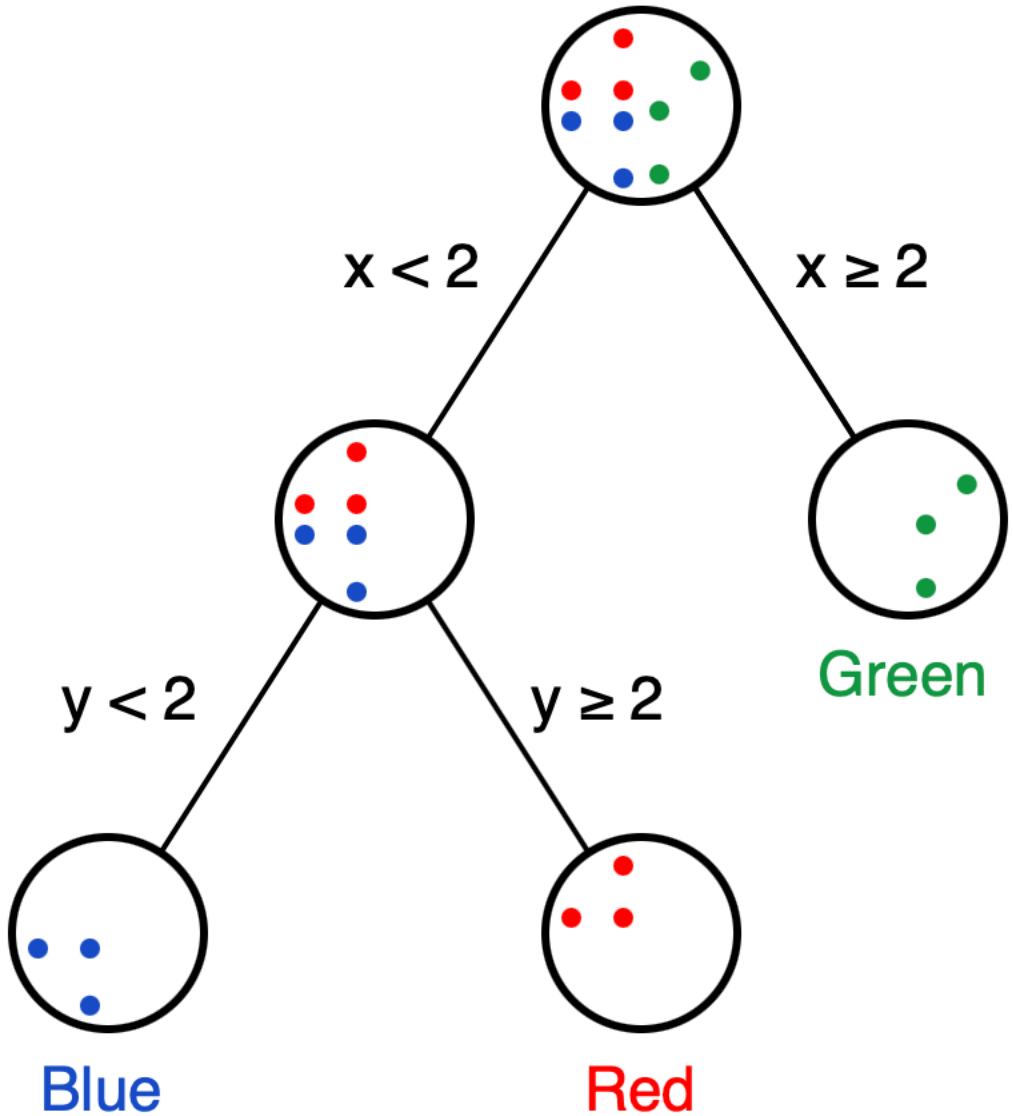


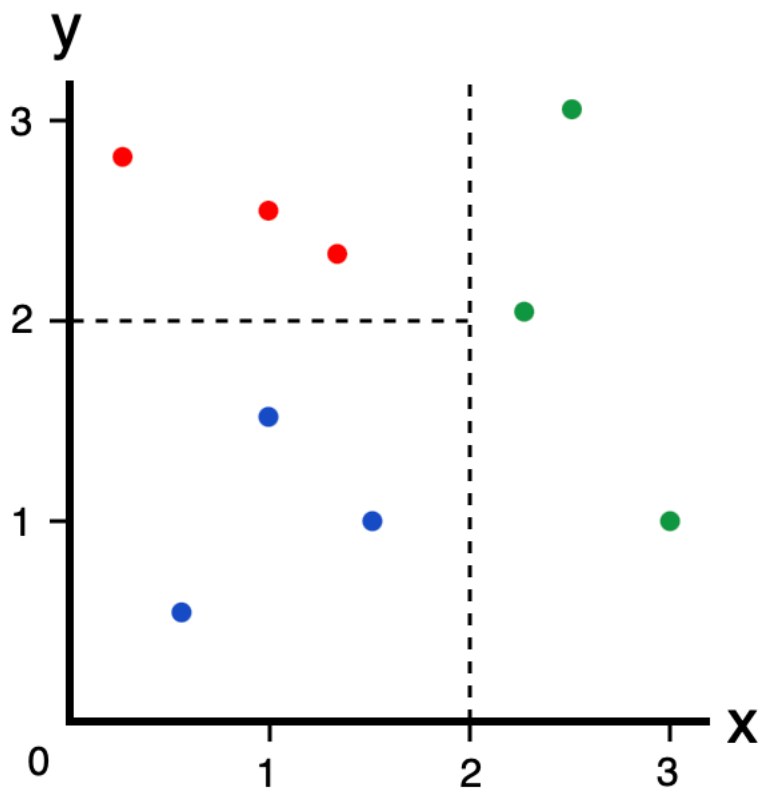
Decision Trees

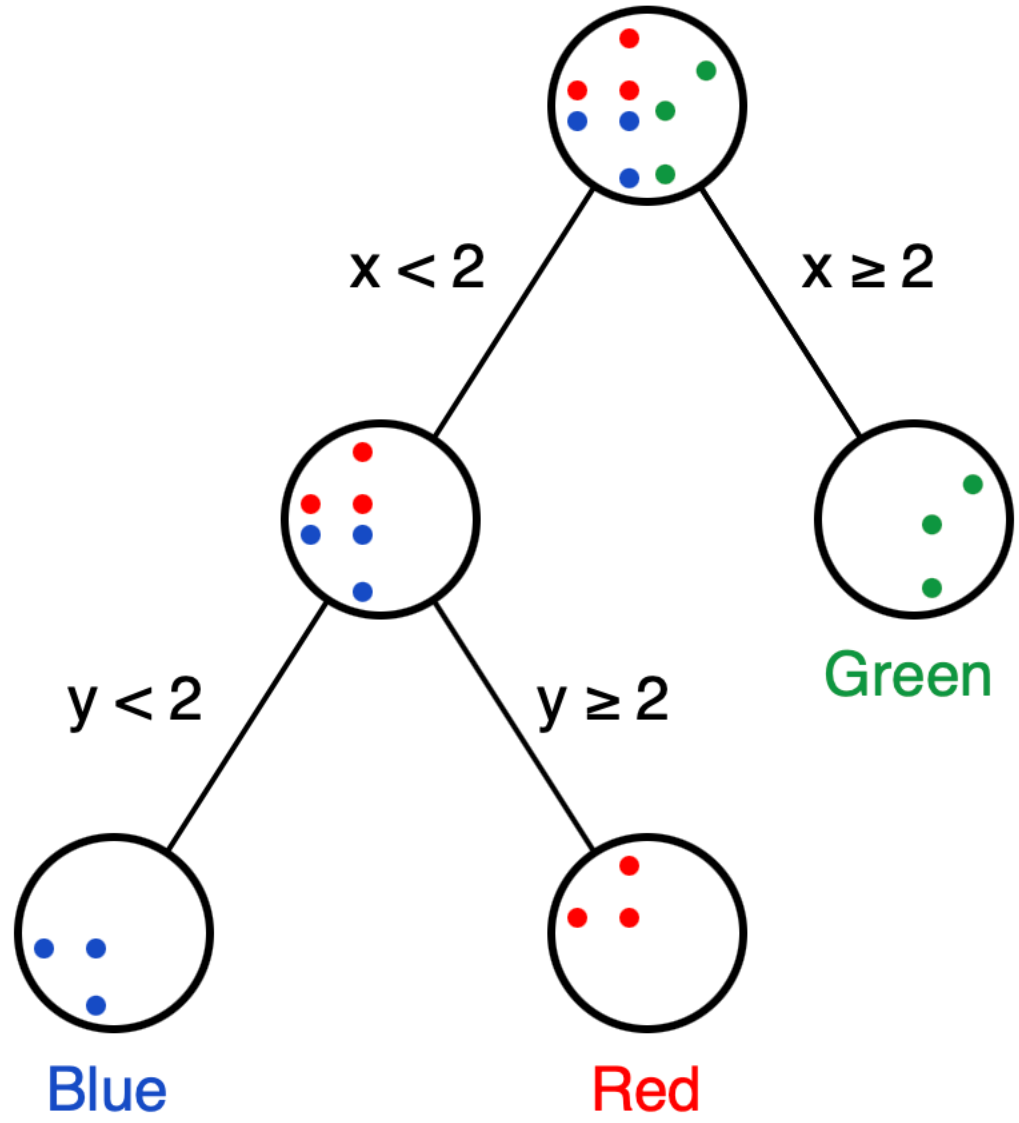


https://www.datacamp.com/tutorial/decision-tree-classification-python

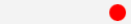

y
3
2
1
0
1
2
3
x


x < 2
x ≥ 2
y < 2
y ≥ 2
Green
Blue
Red


y
3
2
1
0
1
2
3
x


x < 2
x ≥ 2
y < 2
y ≥ 2
Green
Blue
Red

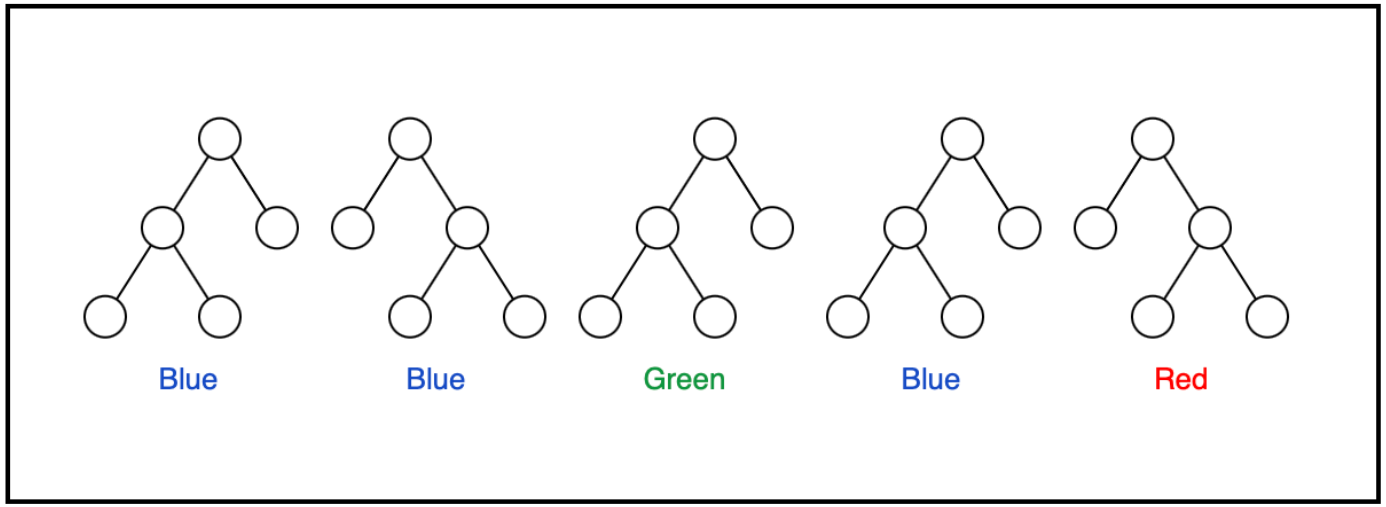
Decision making for splitting

Split	Left Branch	Right Branch	Gini Gain
$x = 0.4$			0.083
$x = 0.8$			0.048
$x = 1.1$			0.133
$x = 1.3$			0.233
$x = 2$			0.333
$x = 2.4$			0.191
$x = 2.8$			0.083
$y = 0.8$			0.083
$y = 1.2$			0.111
$y = 1.8$			0.233
$y = 2.1$			0.233
$y = 2.4$			0.111
$y = 2.7$			0.048
$y = 2.9$			0.083

Random Forest

- A bunch of Decision Trees
- Decisions are made by ensembling the decisions of these trees together

Bagging (Bootstrap Aggregating)



1. Sample, **with replacement**, *n* training examples from the dataset.

2. Train a decision tree on the *n* samples.

3. Repeat *t* times, for some *t*.

https://victorzhou.com/blog/intro-to-random-forests/

Next time

- Sentiment analysis
  - Rule-based and ML/DL solutions