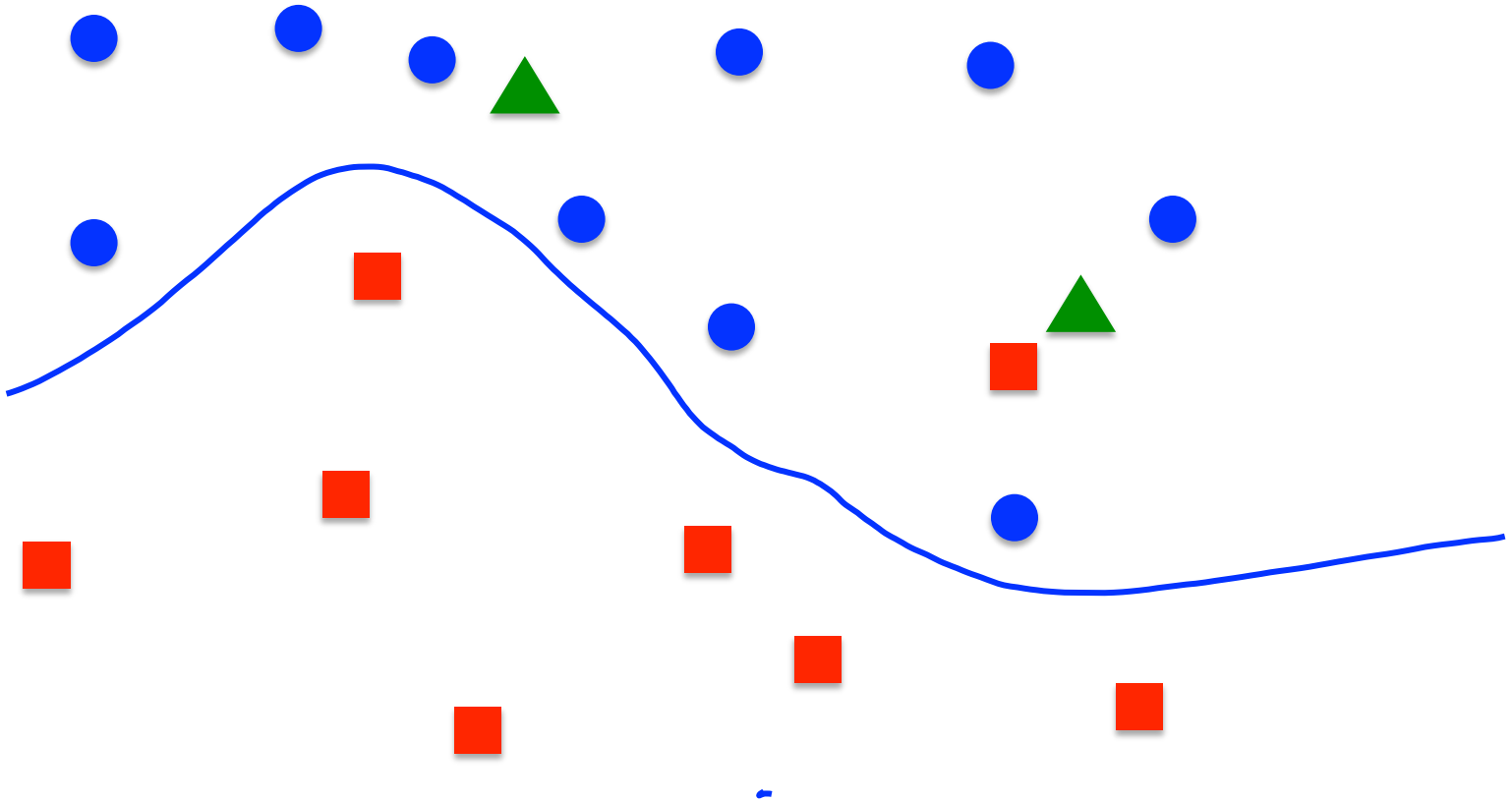
Where we are

- Supervised Learning
- Rules
- Trees
- Challenge: Overfitting
- Ensembles
  - Boosting Adaboost
  - Bagging
  - Bootstrap
- Random Forests

Nearest Neighbor



Nearest Neighbors Intuition

- The last document I saw that mentioned “Falcons” and “Saints” was about Sports, so I’ll classify this document as about Sports too

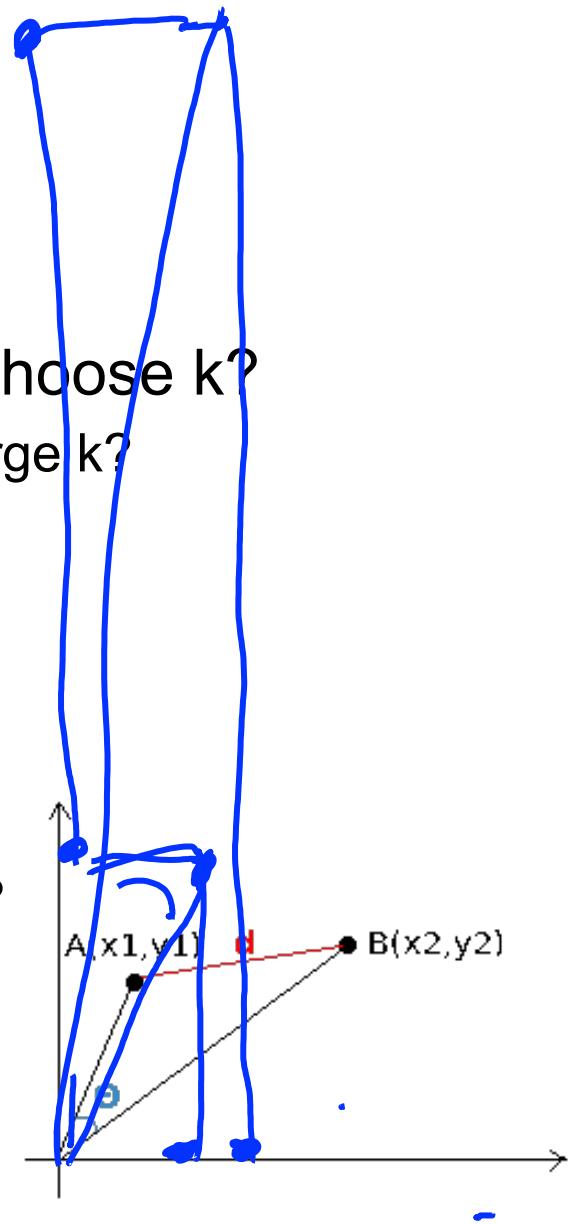
Nearest Neighbor choices

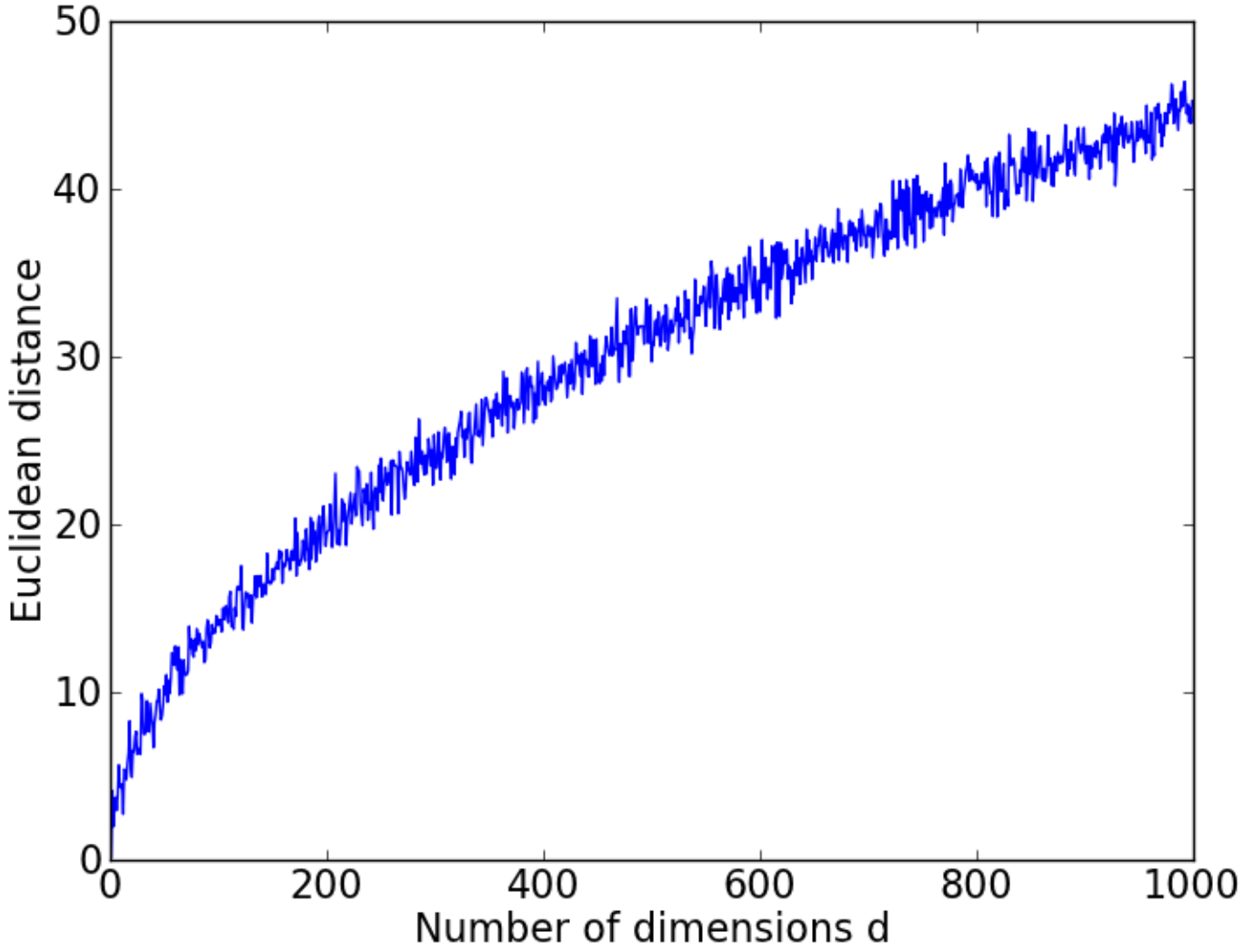
- k nearest neighbors – how do we choose k?
  - Benefits of a small k? Benefits of a large k?

    Large k = bias towards popular labels
    Large k = ignores outliers
    Small k = fast

- Similarity function
  - Euclidean distance? Cosine similarity?

Cosine = favors dominant components
Euclidean = difficult to interpret with sparse data, and high-dimensional data is always sparse




50
40
30
20
10
0
Euclidean distance
0
200
400
600
800
1000
Number of dimensions d