Data Challenge

Machine Learning: Basic Principles

Andrea Nodari

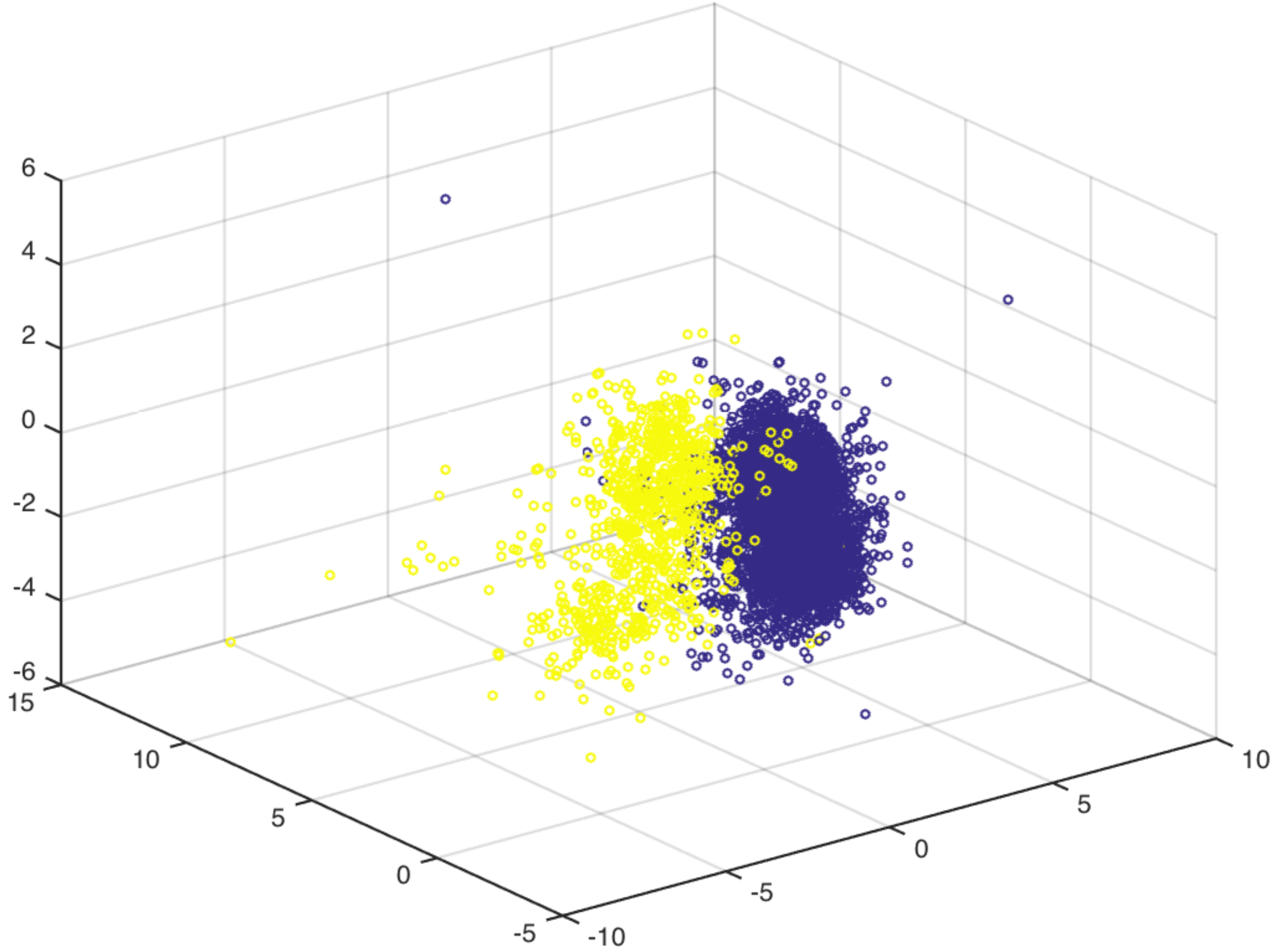
Outline

- Tried alternatives
- Visualizing the data with PCA
- Final Setup

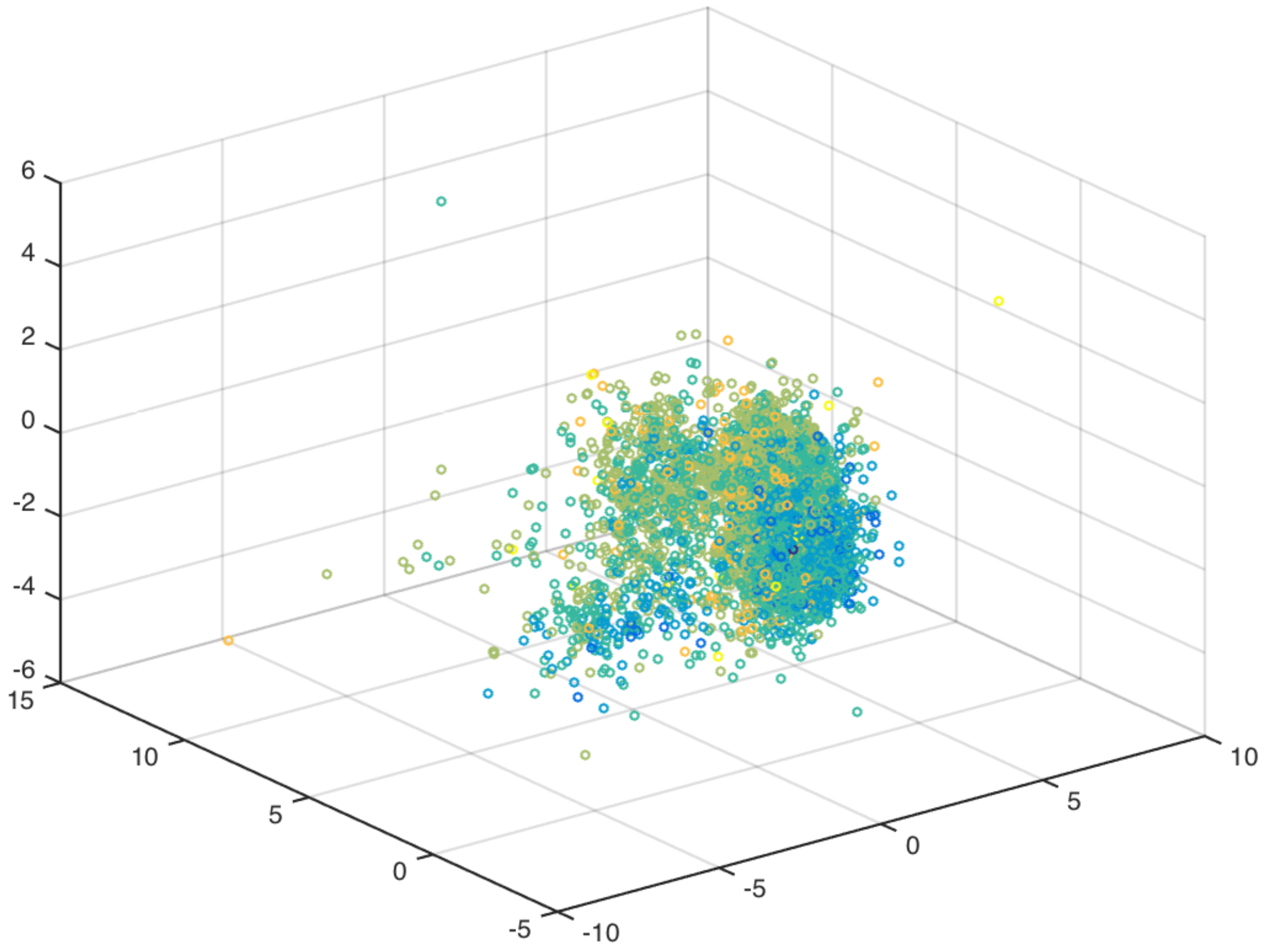
Tried Alternatives

- Logistic Regression
- Exhaustive Feature Selection (~1h)
- Principal Component Analysis
- Support vector machine

Visualizing data with PCA (types)



Visualizing data with PCA (qualities)



Final Setup for Types (1)

$$g(x) = \frac{1}{1 + e^{\theta^T x}}$$

$$Cost(\theta) = \Big(-\frac{1}{N} \sum_{t=1}^{N} r^t log(y^t) + (1 - r^t) log(1 - y^t) \Big) + \frac{\lambda}{2N} \sum_{j=1}^{N} \theta_j^2$$

fmiunc (unconstrained optimization by Octave)

Final Setup for Types (2)

- How to select lambda
    - Test the performance on the validation set
    - Pick the lambda which maximize the F-Score
- Results
    - About **99.1%** accuracy on test set

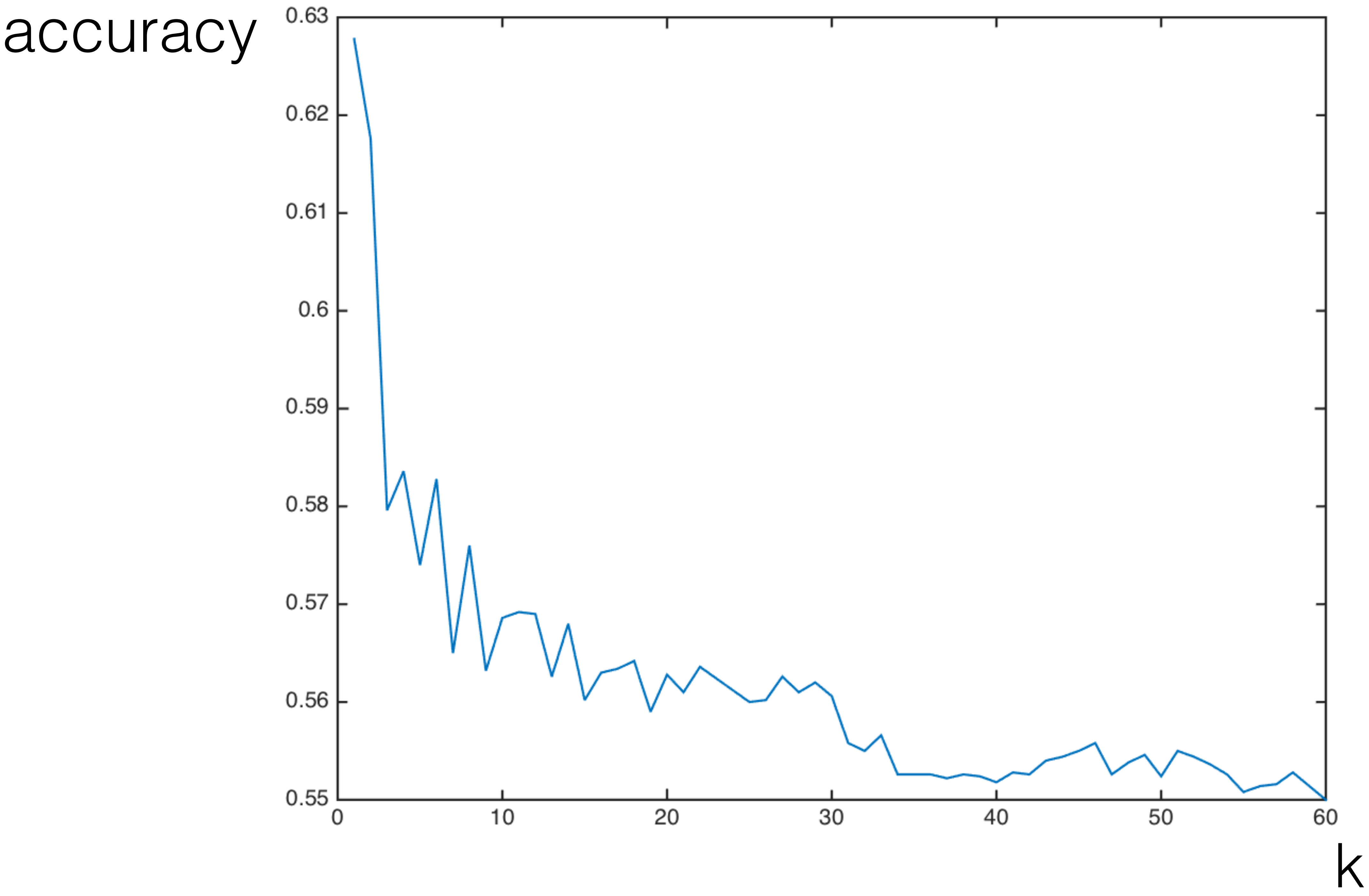
Final Setup for Qualities (1)

- k-Nearest-Neighbors using the *nearest point* as tie-breaker
- How to choose the best distance measure
  - Try different distances on validation set and pick the one with higher accuracy (the final one was Minkowski distance)

Final Setup for Qualities (2)

- How to choose the right *k* (number of neighbors)
  - Test the performance using cross-validation and pick the “k” that maximize the accuracy
  - ~**63%** on the test set

Picking the right *k*...



- My code is open-source, feel free to contribute
  - https://github.com/nodo/machine-learning-challenge

Thanks