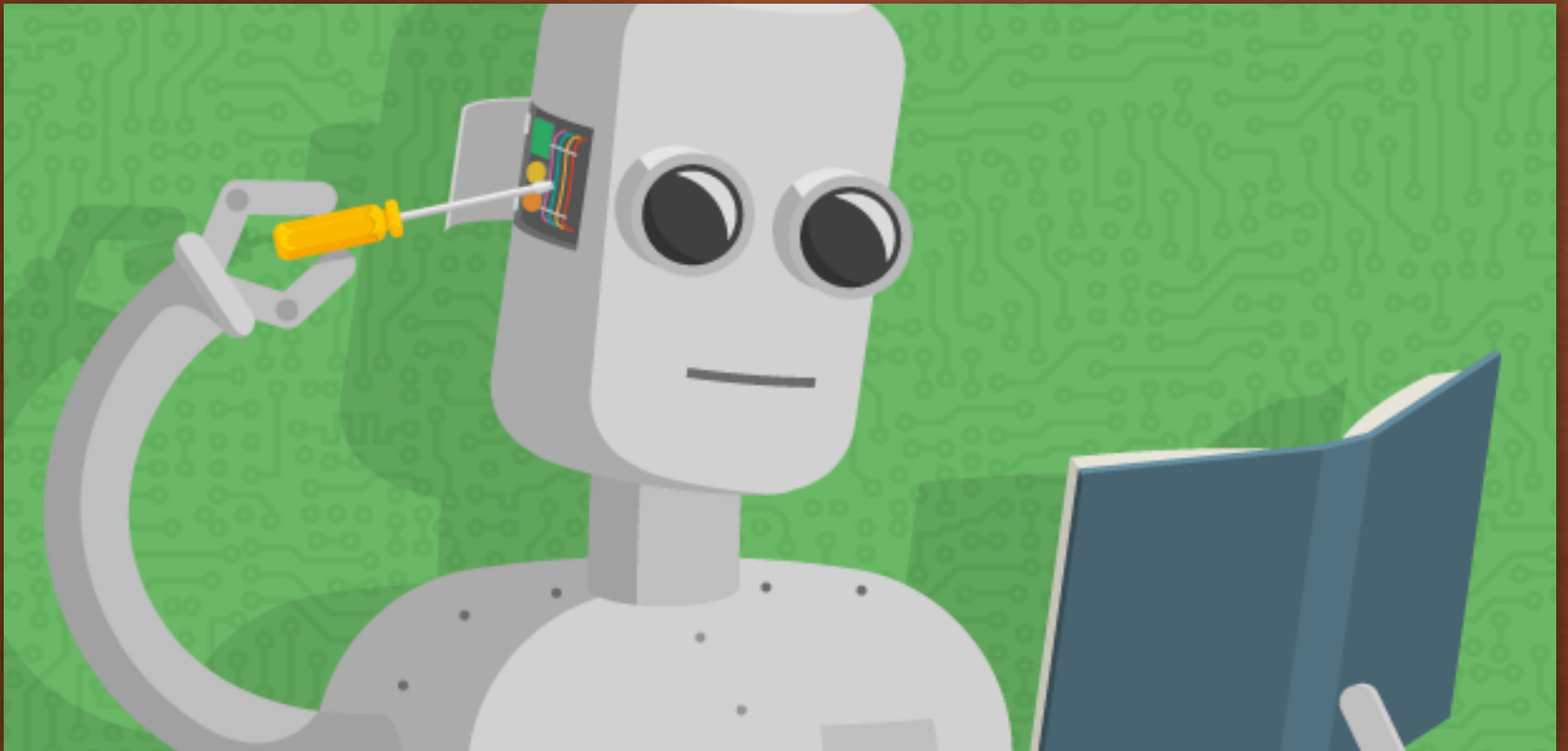


MACHINE LEARNING COURSE PROJECT

PREDICTION OF CREDIT DEFAULT



TEAM

MEMBERS AND MENTOR

Mentor: Suthirtha Das

Members:

Prasanth Manjunath 163050043

Sandeep Subramanian 163050036

Devendra Verma 163050061

Kurian Jacob 163050039

INTRODUCTION

PROBLEM STATEMENT

- Default: A major problem for credit card companies
- Need to reduce losses by reducing the number of customers who default on payments
- Try to classify customers who is likely to default

INTRODUCTION

SOLUTION

- Machine learning
- Use the available information about customers to predict whether this customer is likely to default

INTRODUCTION

DATA SET

- Collected data for 30000 people with
- 23 fields such as:
 - Credit limit
 - History of payments
 - Bill amount
 - Age
 - Sex
 - Education, etc

DATA SET PREPARATION

- Split into two sets in 2:1 for training and testing
- Applied PCA to find out which fields were actually relevant to us

DATA SET

ANALYSIS

- Applying PCA and running some temporary models, accuracy was found to increase with number of data fields
- Decided to take all 23 parameters into consideration

VARIOUS MODELS

OVERVIEW

- KNN Classifier
- Neural network classifier
- SVC
- Decision trees

KNN-CLASSIFIER

DESCRIPTION

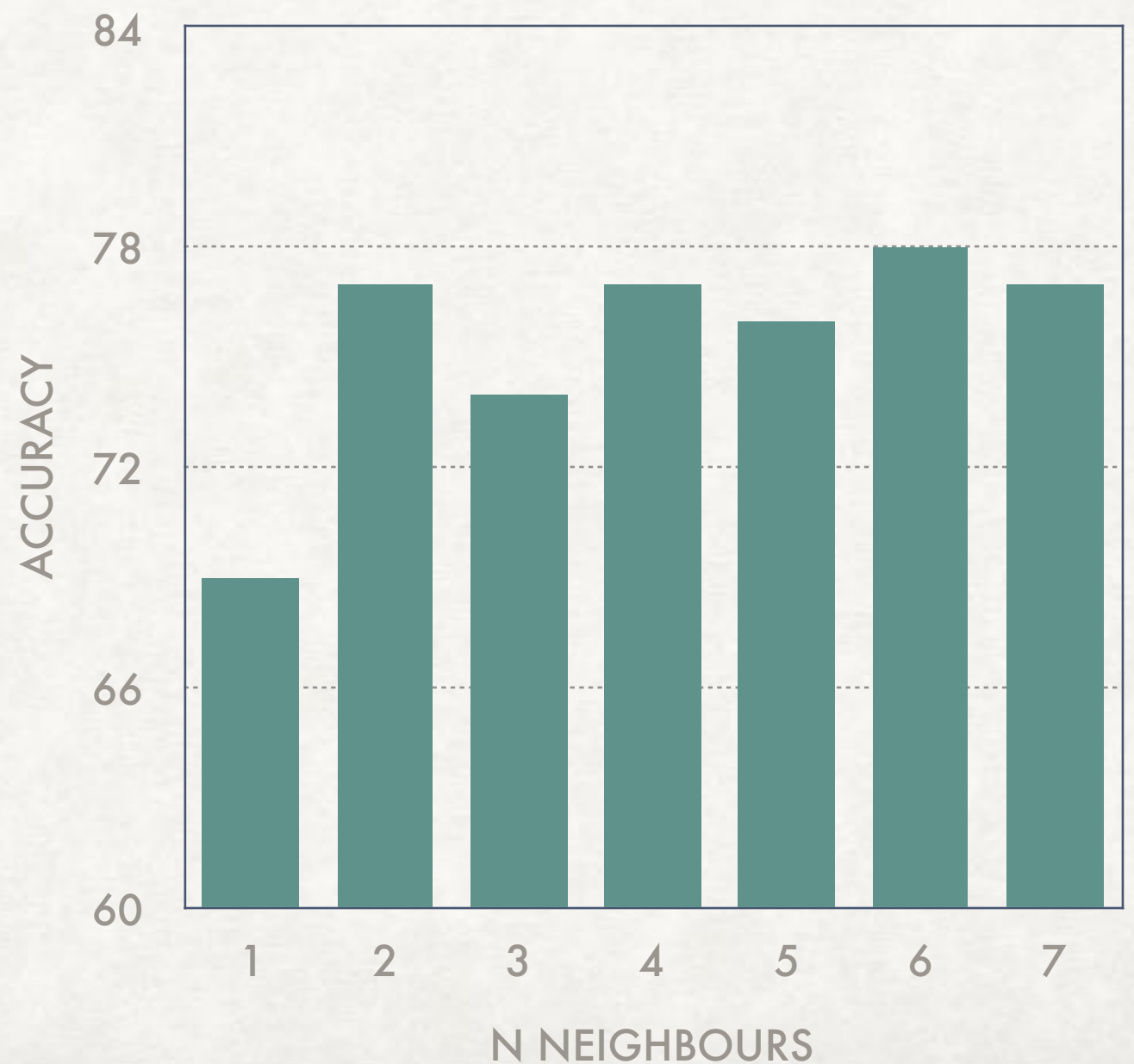
- k-nearest neighbour to predict the result
- K was set to a value of 4
- Minkowski distance metric

KNN CLASSIFIER

RESULT

Number of Neighbours	Accuracy
1	69
2	77
3	74
4	77
5	76
6	78
7	77

KNN ACCURACY AGAINST N-NEIGHBOURS



NEURAL NETWORK

DESCRIPTION

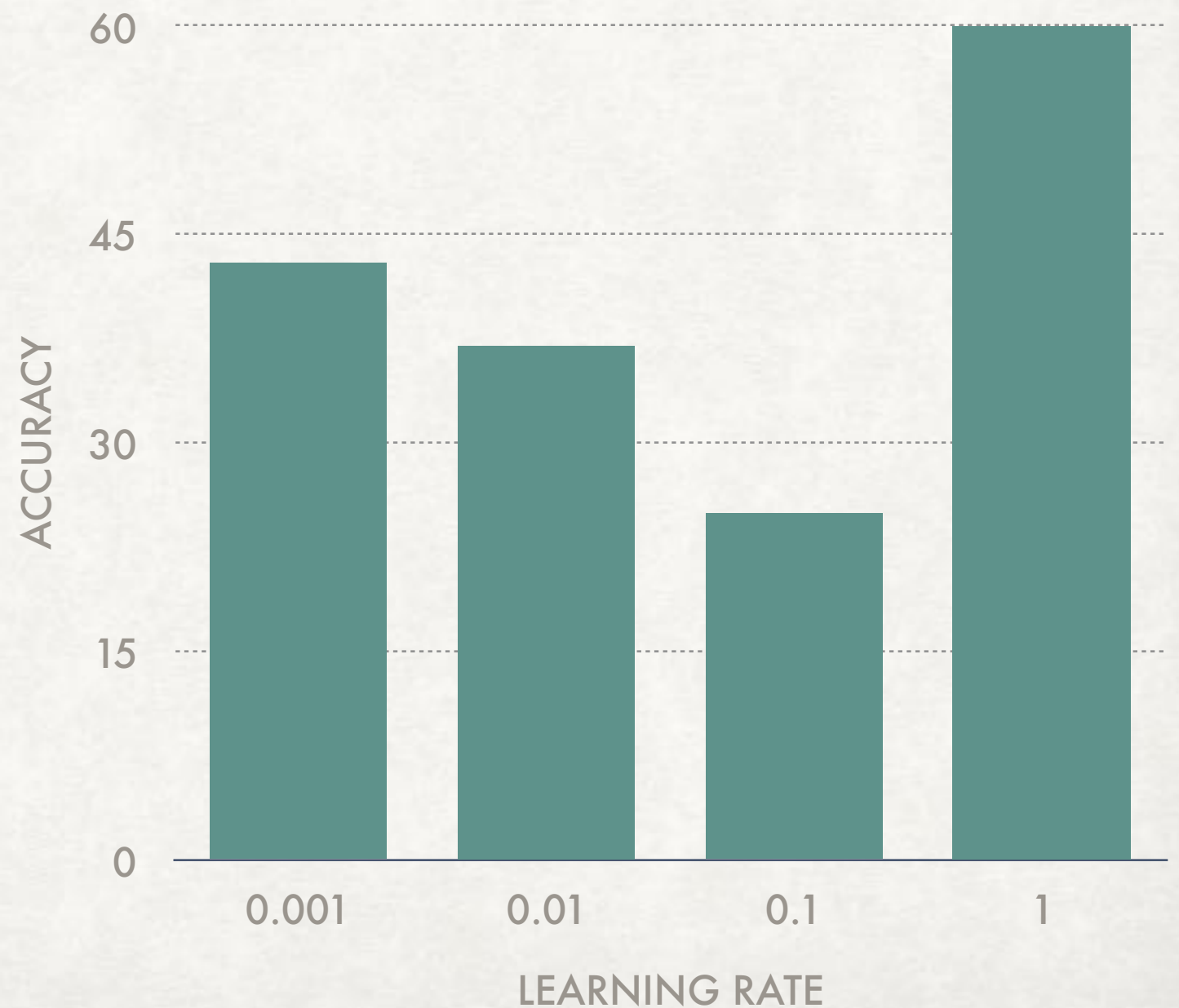
- Contains layers of units called neurons
- Multiple inputs, single output
- We used three hidden layers
- Input layer - 83 neurons
- Hidden layers have 160, 80, 23 neutrons respectively
- Output layer - 2 nodes
- Implemented from scratch

NEURAL NETWORKS

RESULT

Learning rate	Accuracy
0.001	43
0.01	37
0.1	25
1	60

ACCURACY FOR VARIOUS VALUES OF LEARNING RATE



SUPPORT VECTOR CLASSIFIER

DESCRIPTION

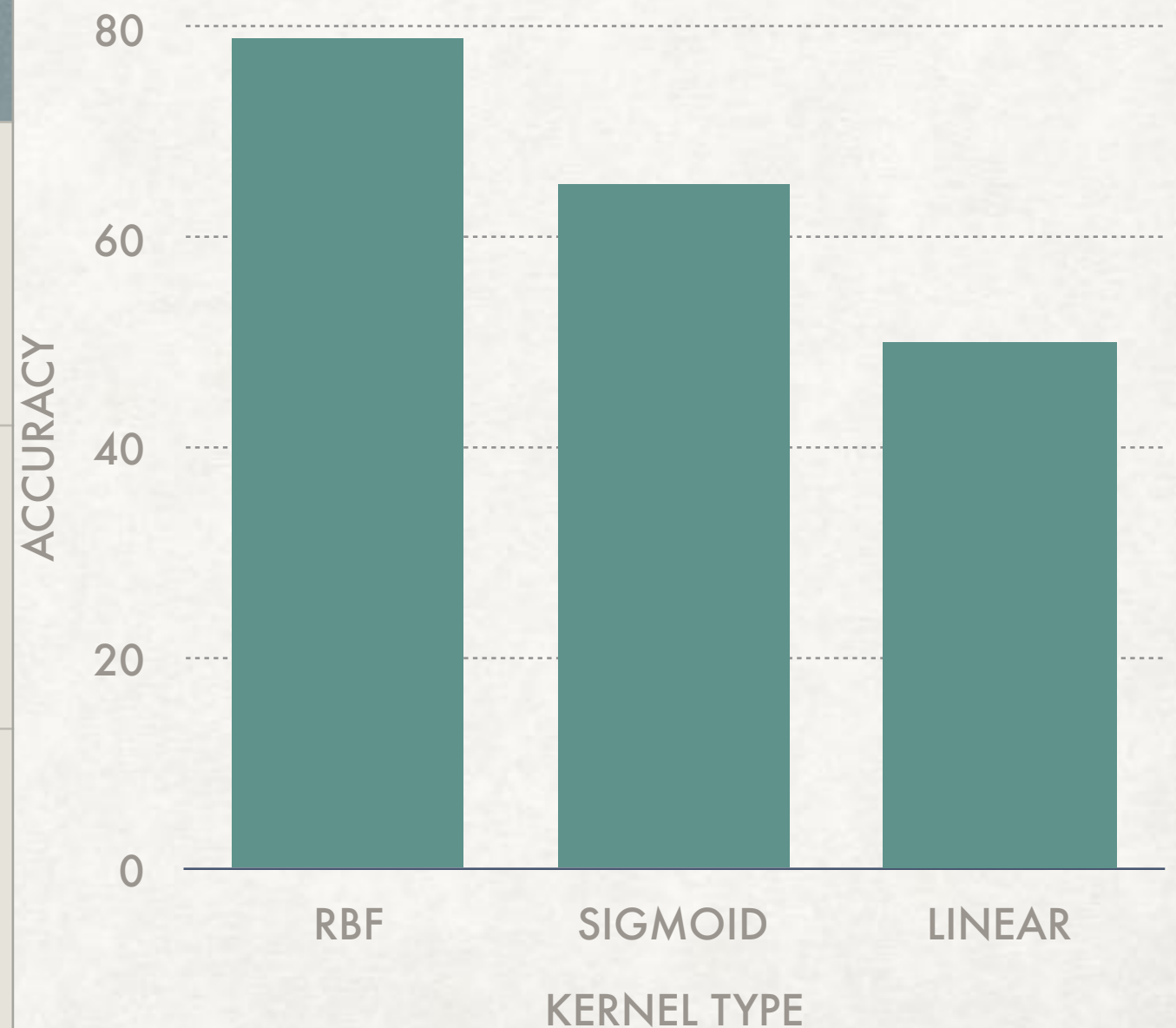
- Deterministic binary classifier
- Kernels implemented were RBF, Sigmoid and Linear

RESULTS

SVM CLASSIFIER

Kernel Type	Accuracy
RBF	79
Sigmoid	65
Linear	50

ACCURACY FOR SVM CLASSIFIER



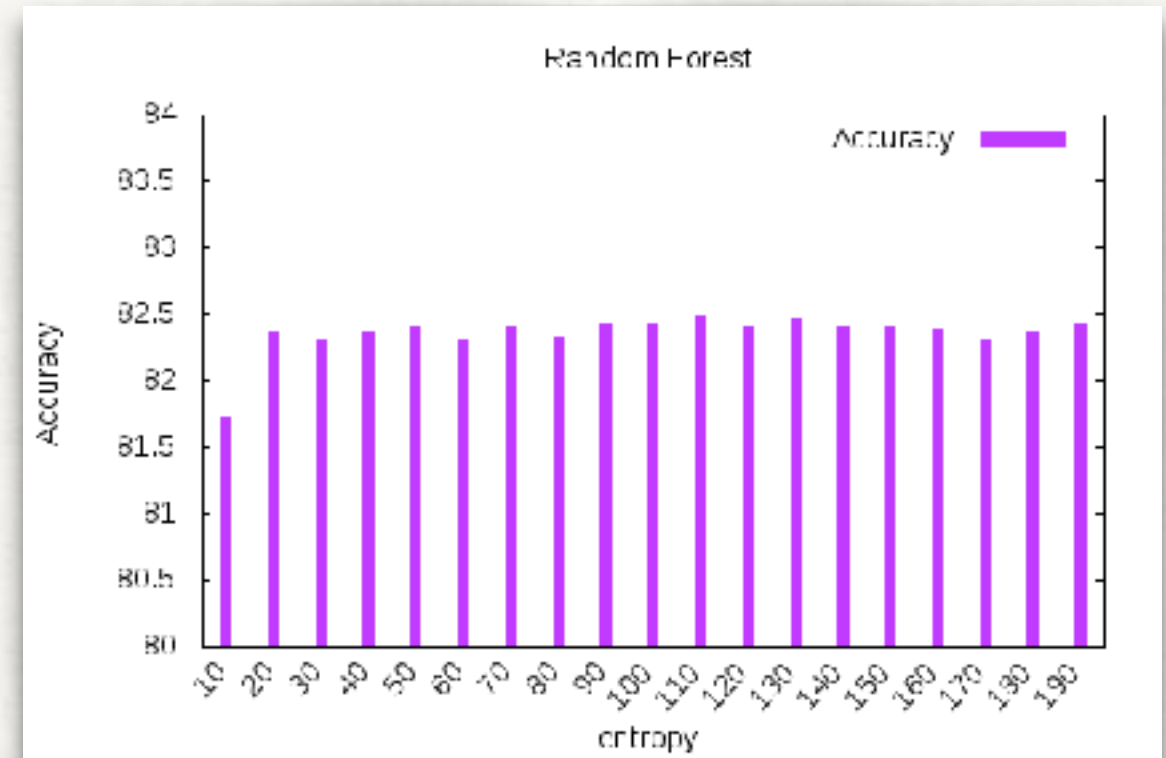
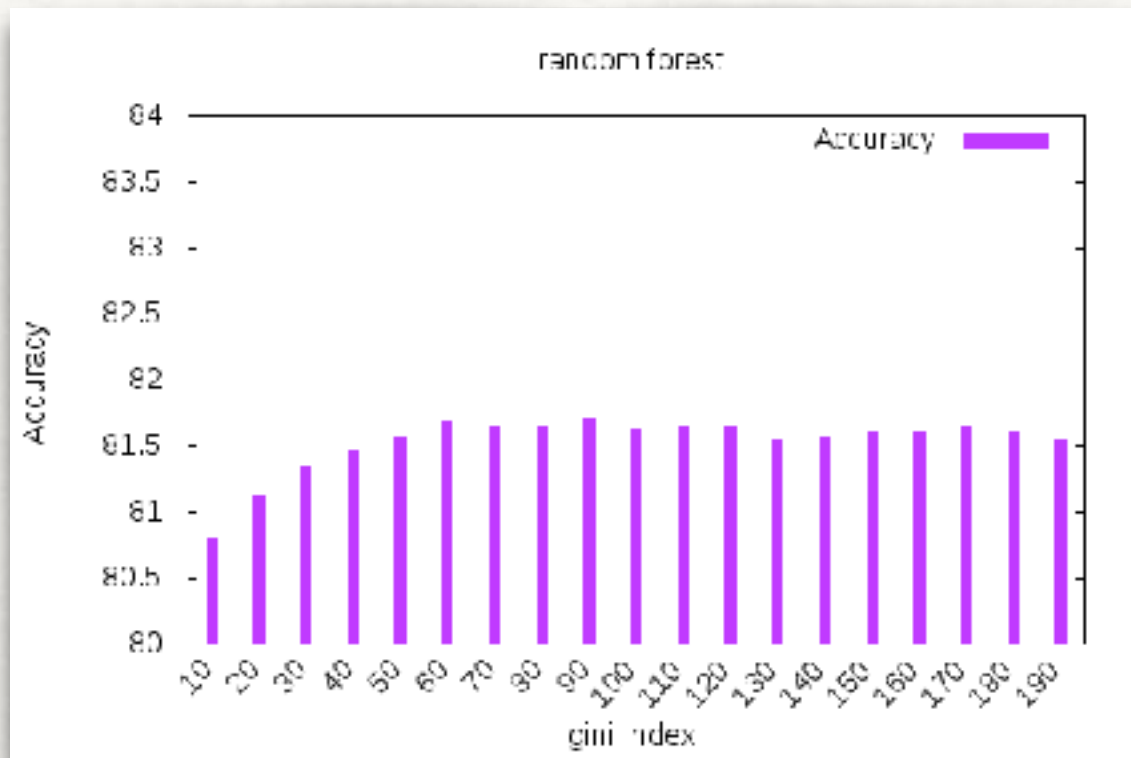
RANDOM FORESTS

DESCRIPTION

- Implemented random forests
- We prepared $\sqrt{20000} = 142$ trees
- Classes were weighed
- Entropy for impurity function

RANDOM FORESTS

RESULTS



ADABOOST

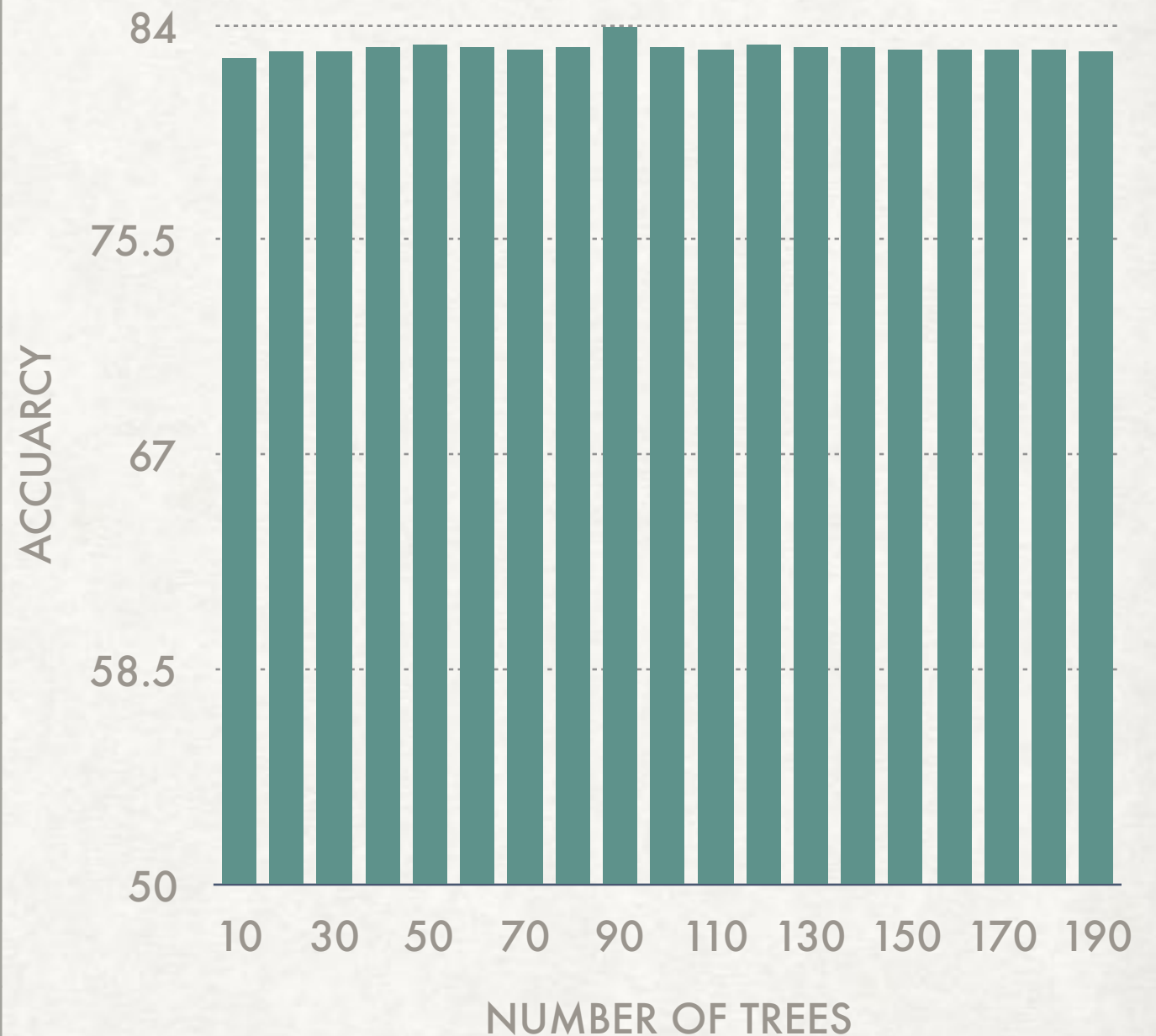
DESCRIPTION

- Similar to decision trees
- Can have varying weights

ADABOOST WITH DECISION TREES

RESULTS

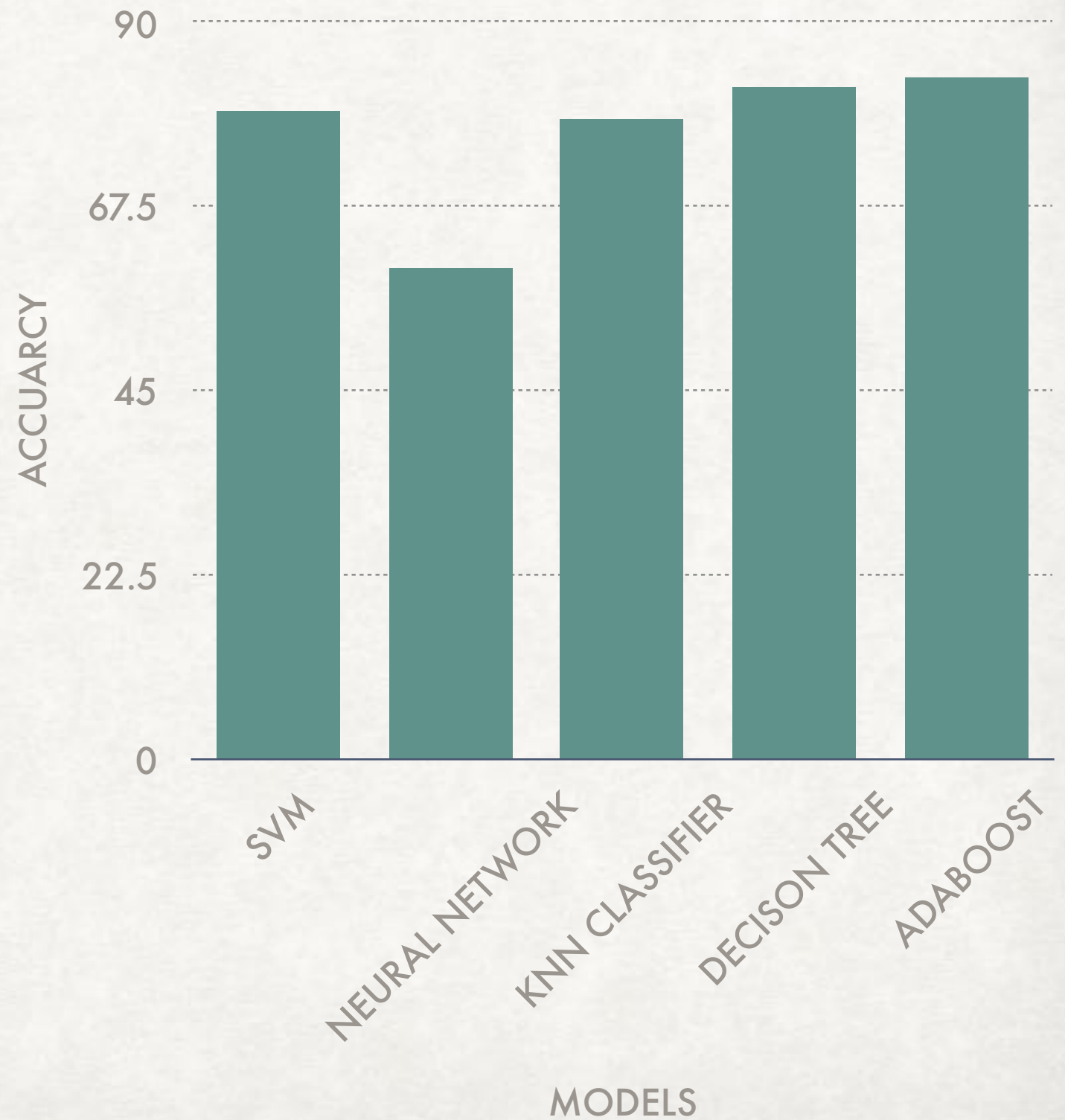
Number of trees	Accuracy
10	82.7
20	82.9
30	82.94
40	83.14
50	83.23
60	83.13
70	83.06
80	83.12
90	83.04
100	83.11
110	83.06
120	83.23
130	83.11
140	83.12
150	83.04
160	83.04
170	83.03
180	83.03
190	82.98



RESULTS

MODEL COMPARISON

Model Name	Accuracy
SVM	79
Neural Network	60
KNN Classifier	78
Decision tree	82
Adaboost	83.23



FURTHER WORK

- Explore accuracy using LSTM
- Since our dataset has only data from April to September, LSTM was not used, but if more data were available about customer, LSTM might be a better model.

CONCLUSION

After comparing four models, we have seen that decision tree with adaboost model have given the best accuracy.

REFERENCES

1. Yeh, I. C., & Lien, C. H. (2009).
The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.
<http://www.sciencedirect.com/science/article/pii/S095741740700671>
2. Bart Baesens , Rudy Setiono , Christophe Mues , Jan Vanthienen. *Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation, Management Science, v.49 n.3, p.312-329, March 2003.*
<https://www.jstor.org/stable/pdf/4101754.pdf>
3. Machine learning with scikit-learn.
<http://scikit-learn.org/stable/tutorial/basic/tutorial.html>
4. Neural network in 11 lines.
<http://iamtrask.github.io/2015/07/12/basic-python-network/>