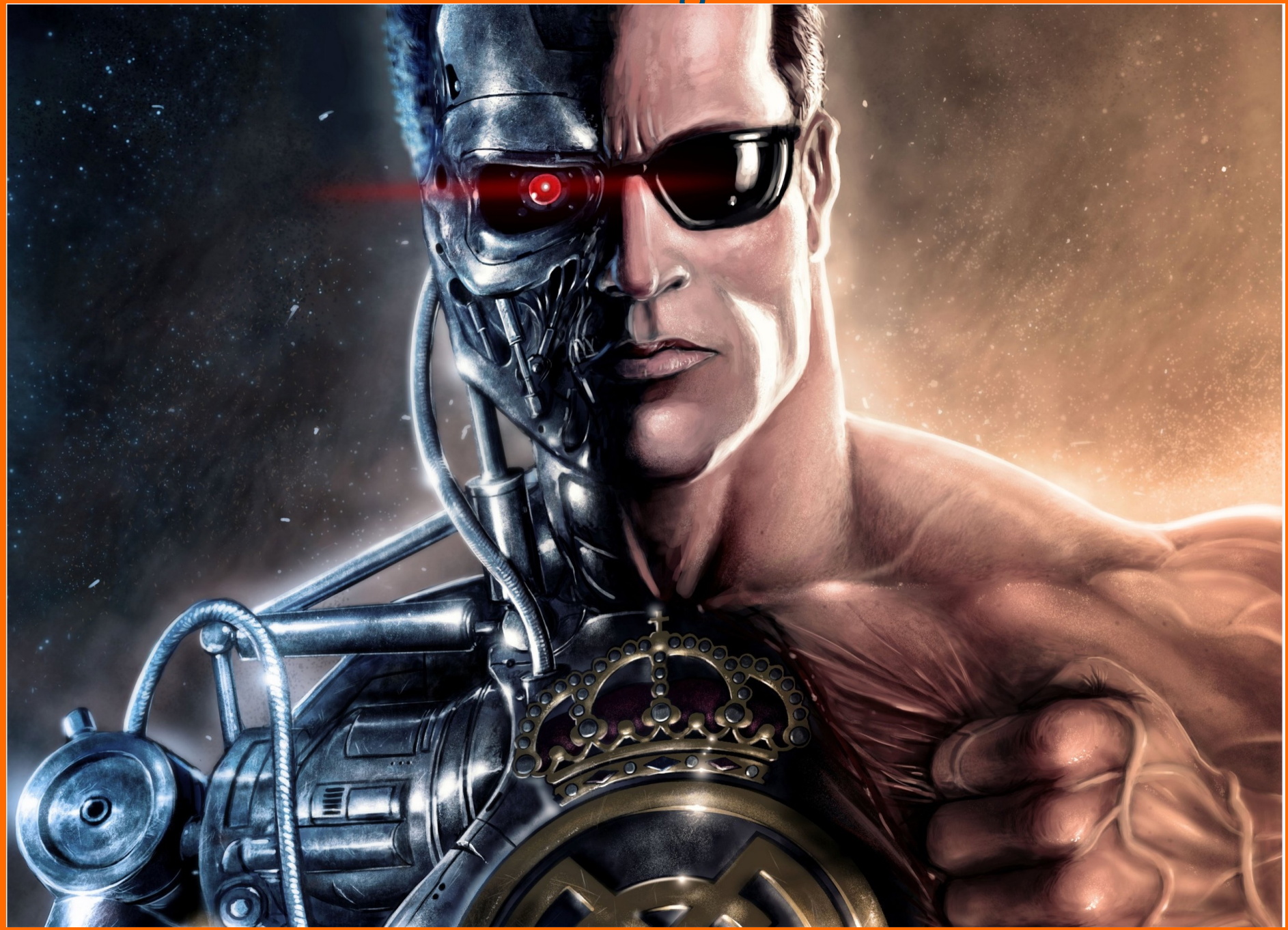
Machine Learning - Deep Learning

Farshid Alizadeh-Shabdiz, PhD, MBA

Spring 2022

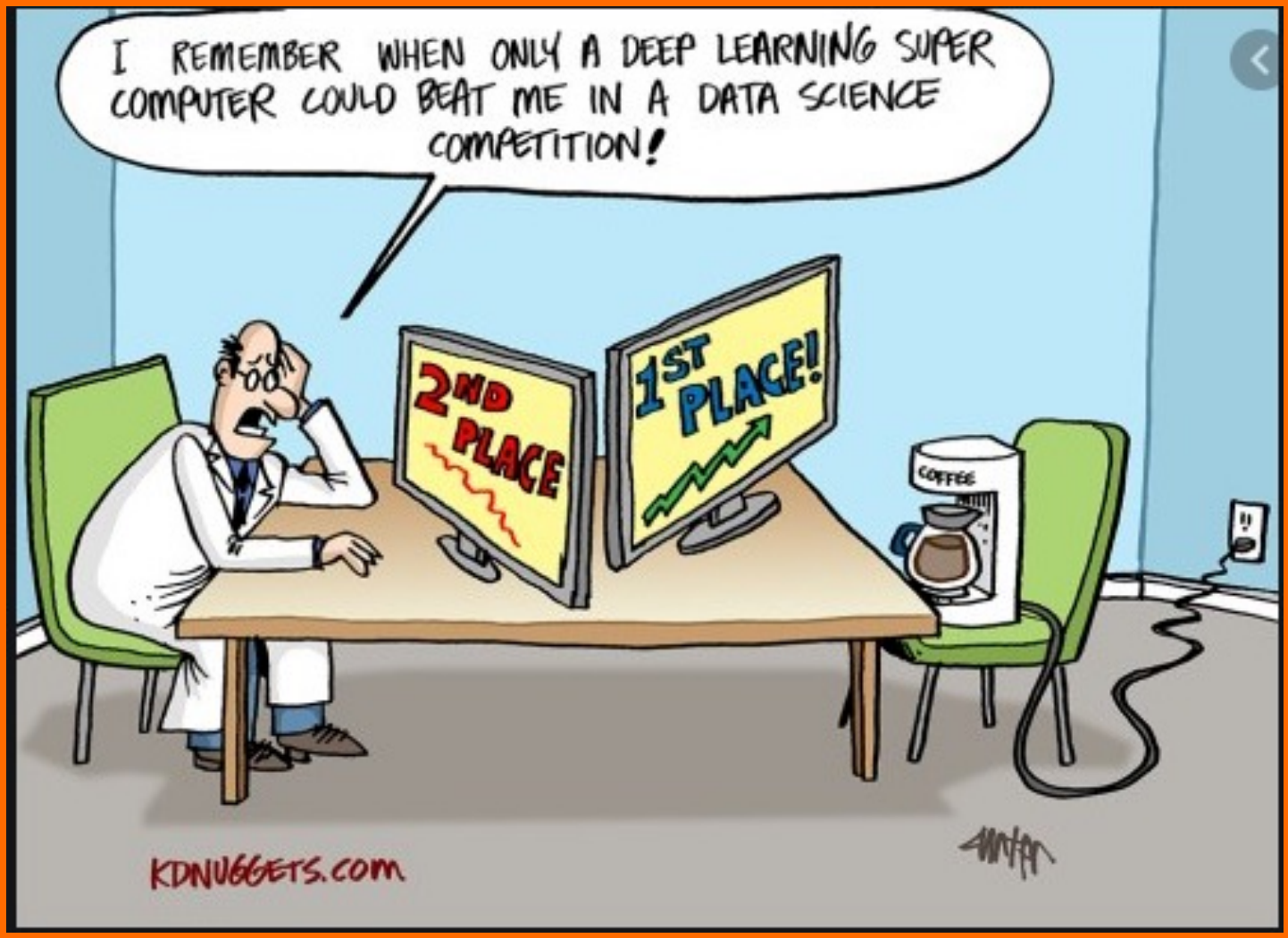
What is Machine Learning?

What is Machine Learning?



Deep Blue vs Gary Kasparov – May 11, 1997

In early 2011, an IBM computing system named Watson competed against the world's best Jeopardy champions!

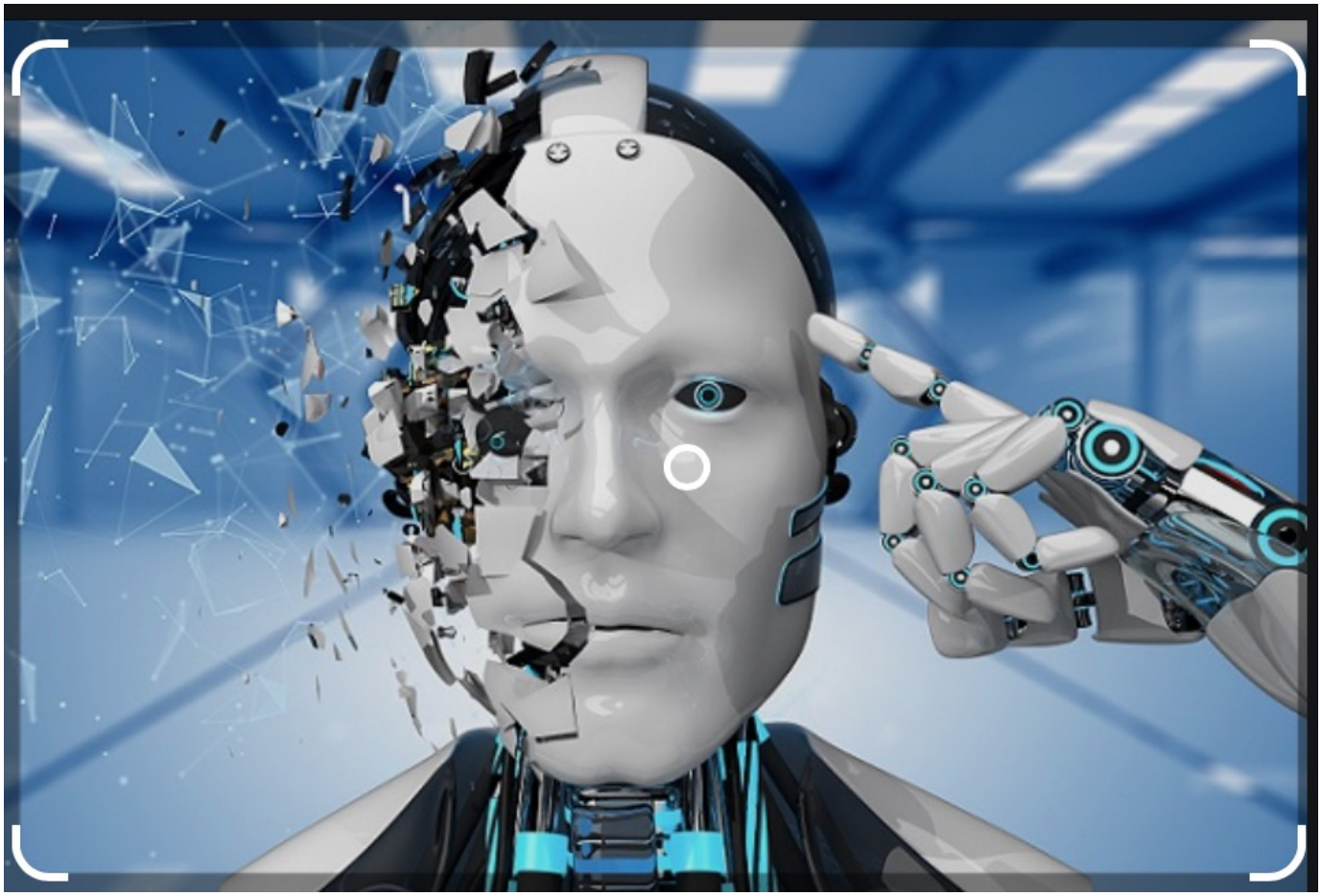

I REMEMBER WHEN ONLY A DEEP LEARNING SUPER COMPUTER COULD BEAT ME IN A DATA SCIENCE COMPETITION!
2ND PLACE
1ST PLACE!
COFFEE
KDNUGGETS.com

Statistical Machine Learning

vs

Deep Learning

Deep Learning



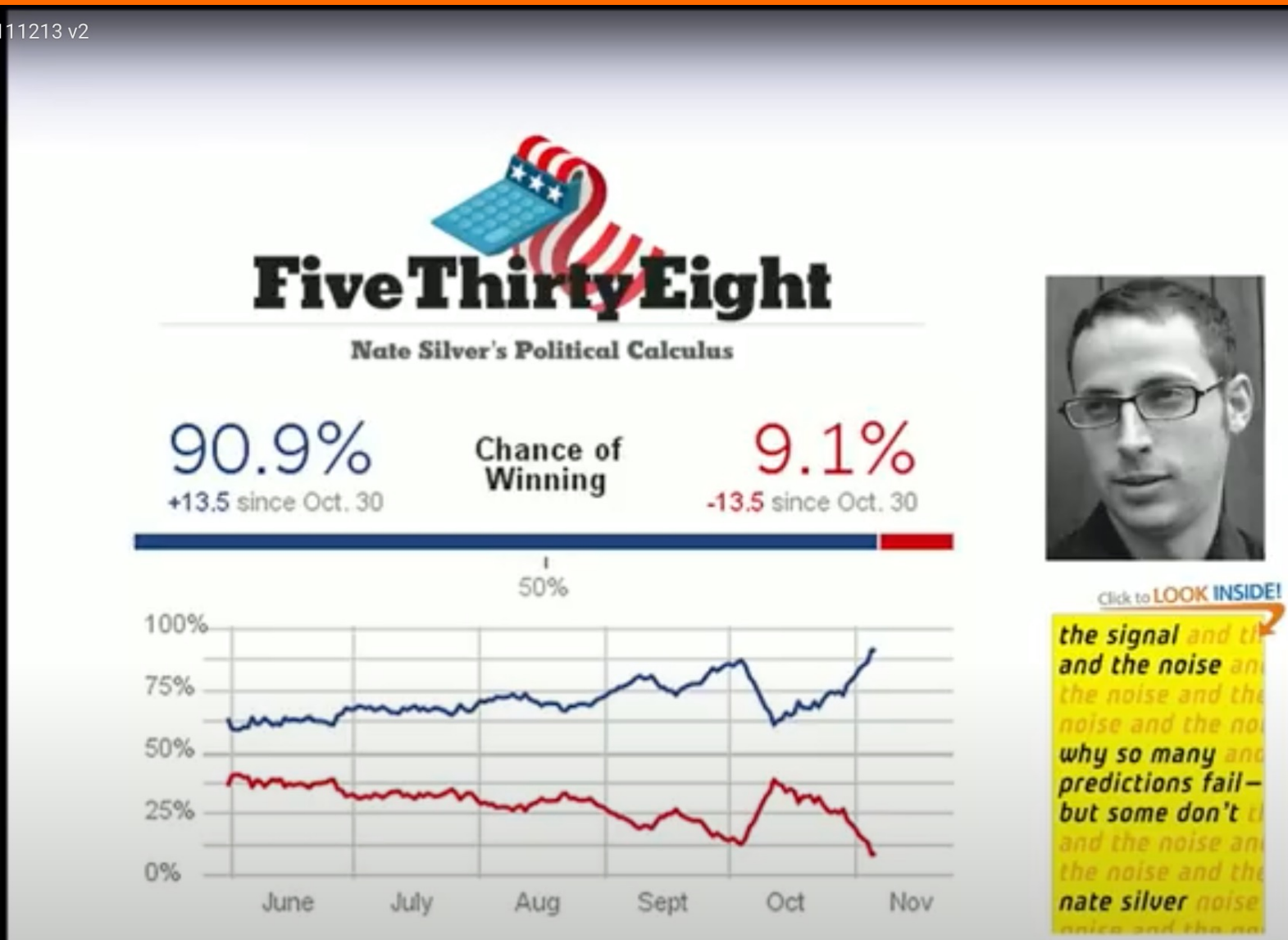
History of Neural Network

History of Neural Networks

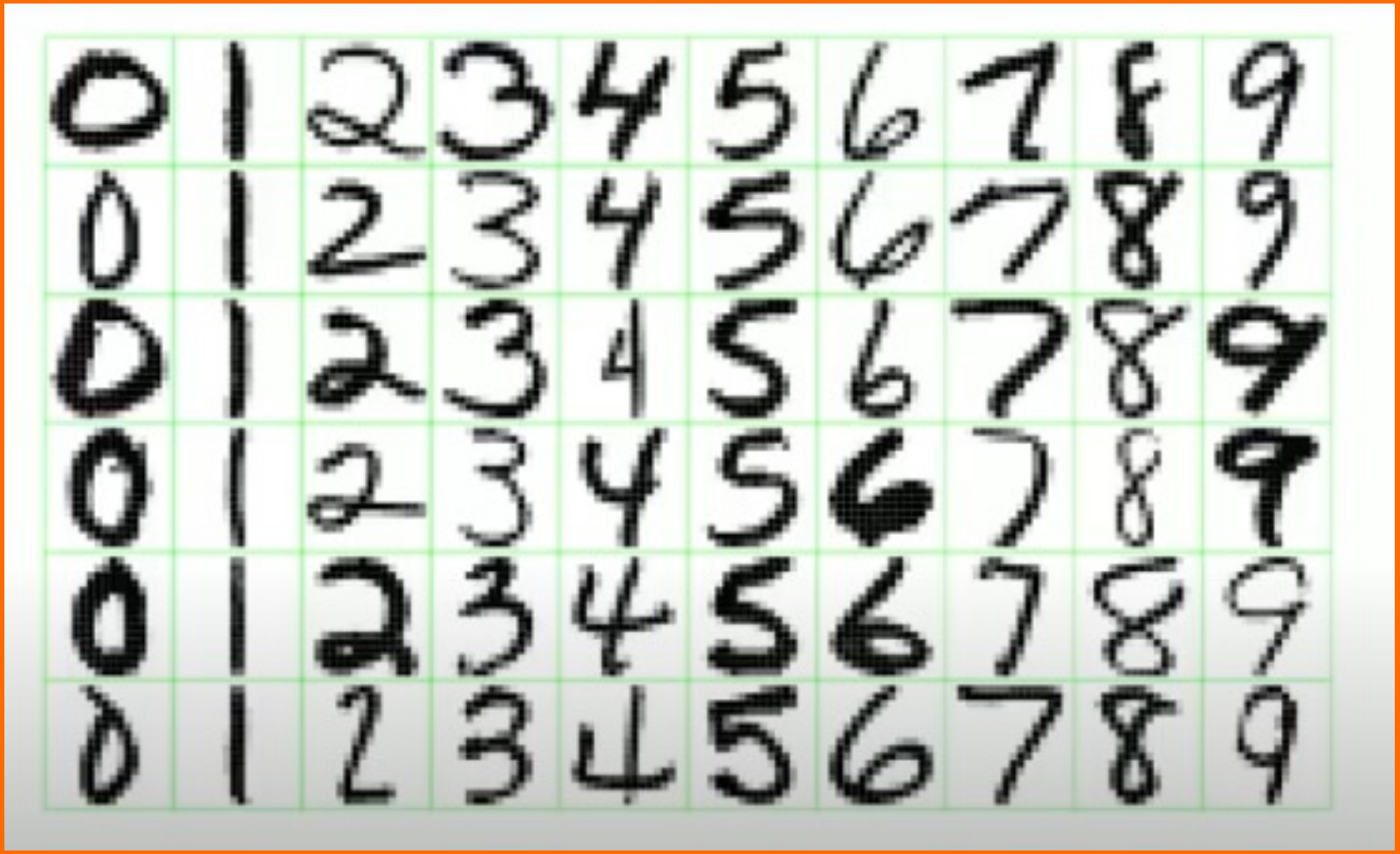
- 1943: McCulloch & Pitts created a computer model based on neural networks of brain
- 1949: Hebb's Rule – Hebb suggested how biological neurons work
- 1957: Perceptron – simple NN invented by Frank Rosenblatt
- Many promises!
- 1969: M Minsky & S. Papert showed limitation of perceptron
- 1970's: the first winter of AI
- 1986: D. Rumelhart, G. Hinton, and R. Williams that introduced backpropagation to train deep NN.
- 1985-1990: the 2nd winter of AI
- 1997: IBM Deep Blue beat Kasparov, world chess champion
- 1997: NN with memory - LSTM (long-short term memory) was developed
- 1999: GPU was developed
- 2001 : Winter of AI
- 2006: Prof Hinton trained a network to read handwritten numbers
- 2012: AlexNet CNN architecture won the ImageNet challenge with a large margin (17% error vs next best 26%)
  - 2009: Fei-Fei Li, AI prof at Stanford launched ImageNet (14 million labeled images).

2012 Senate and presidential selection prediction in New York Times blog

StatsLearning Lect1/2a 111213 v2

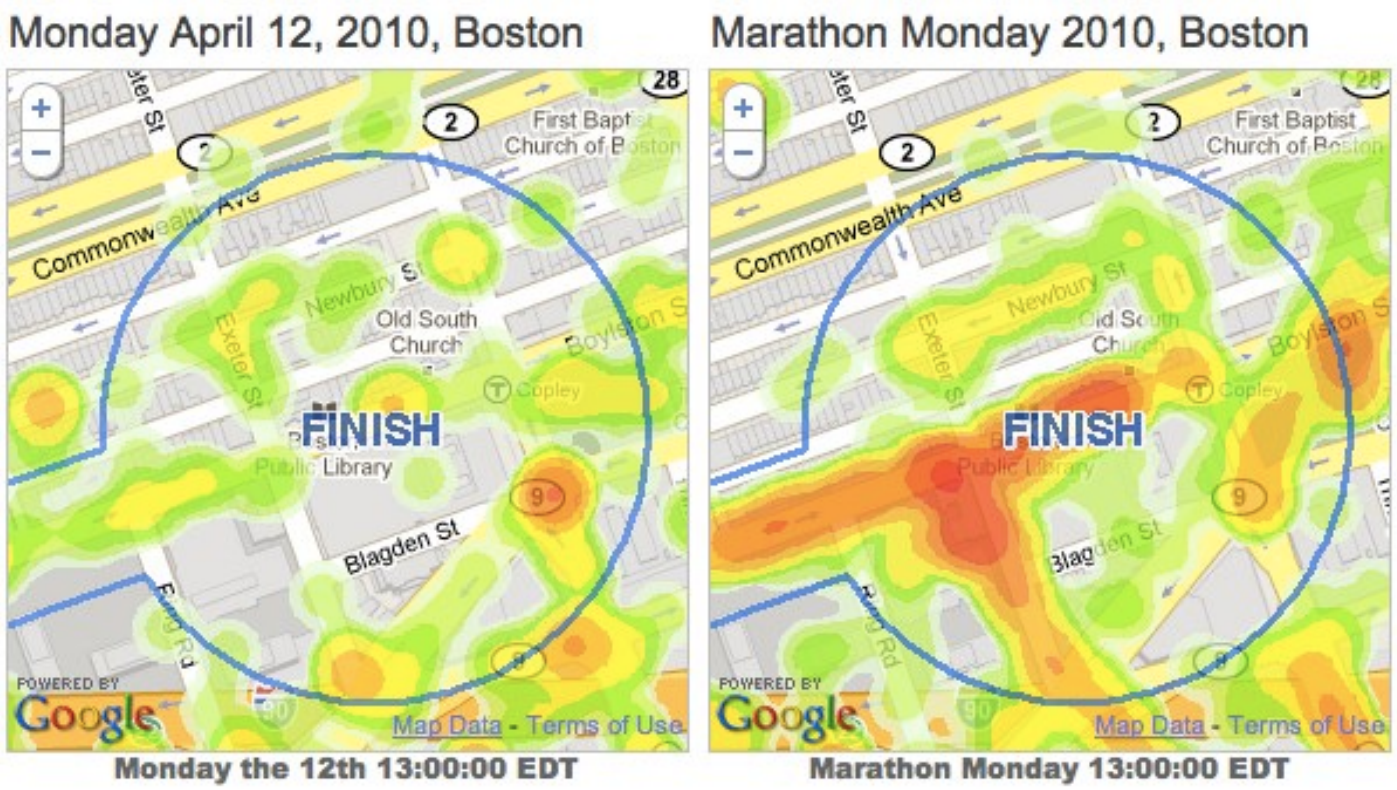


Identify handwritten numbers

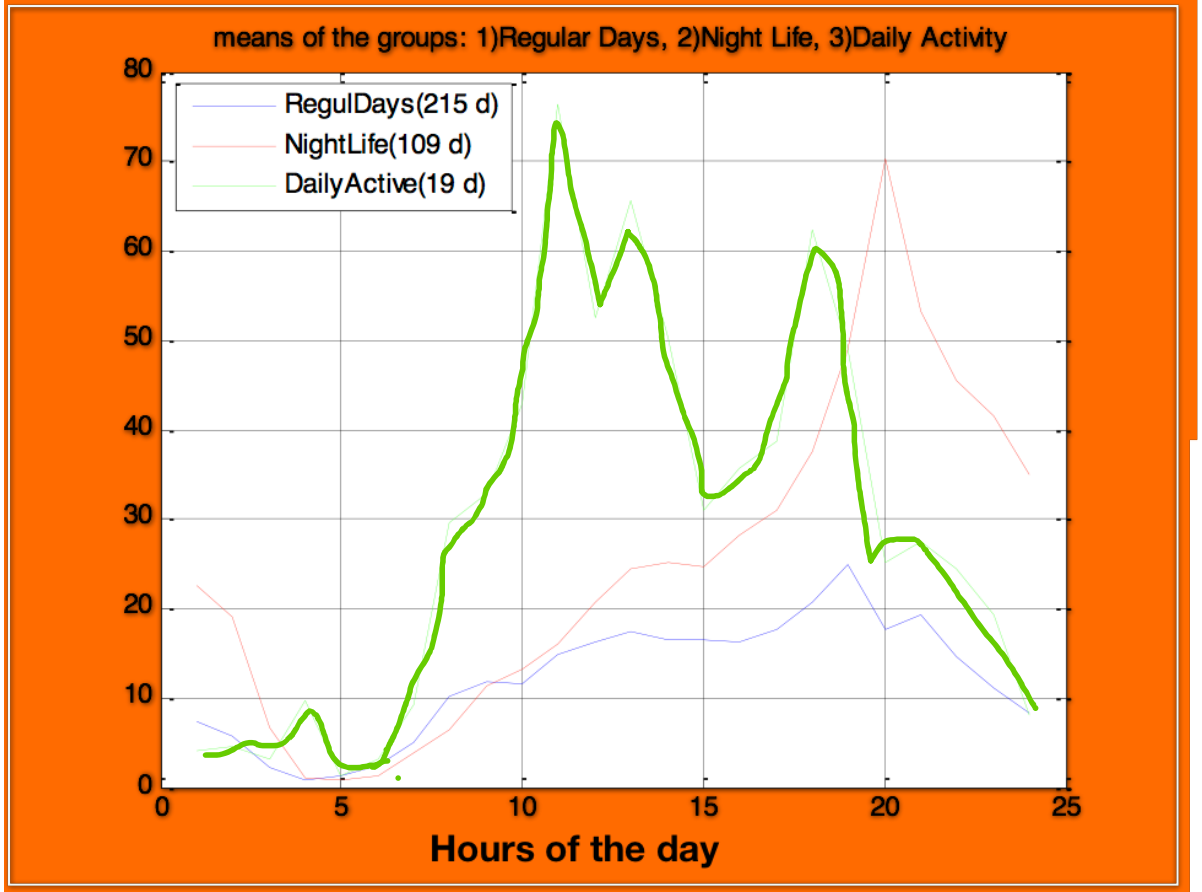


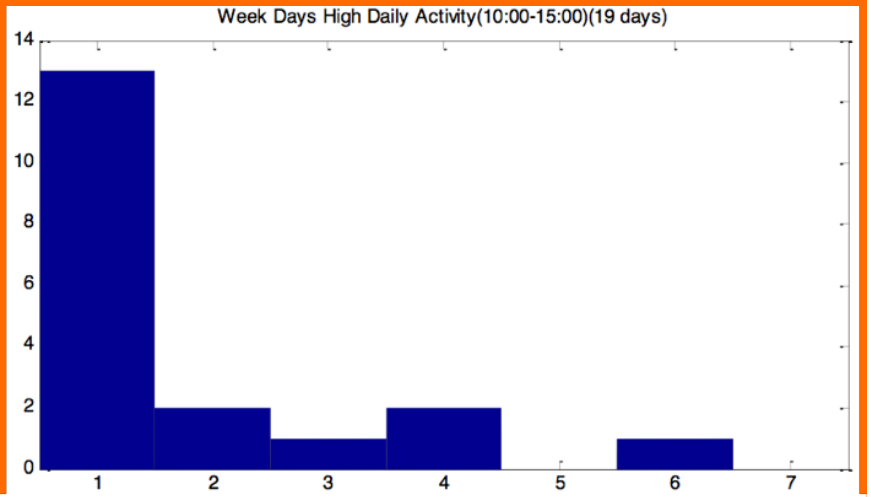
Example of Machine Learning

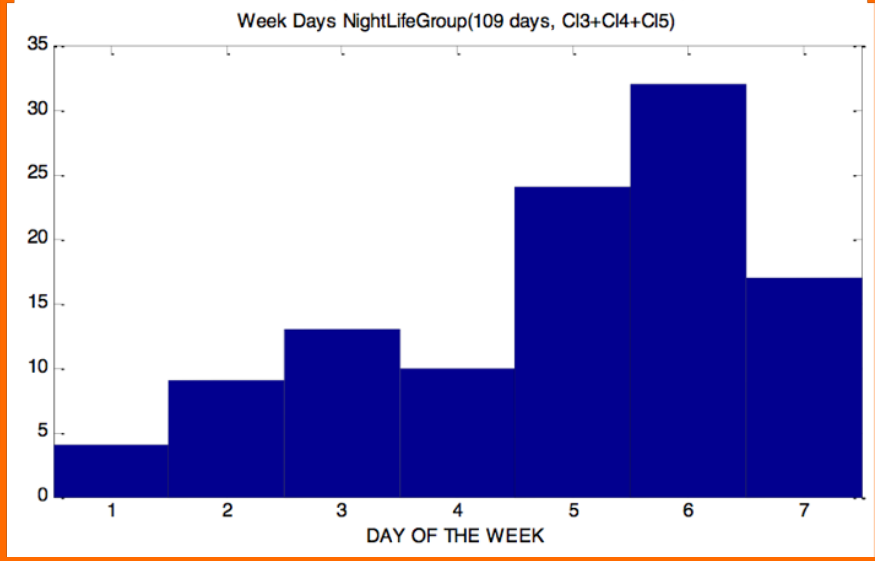
Real time population density based on location requests



Location Based Behavior Analysis – Using PCA

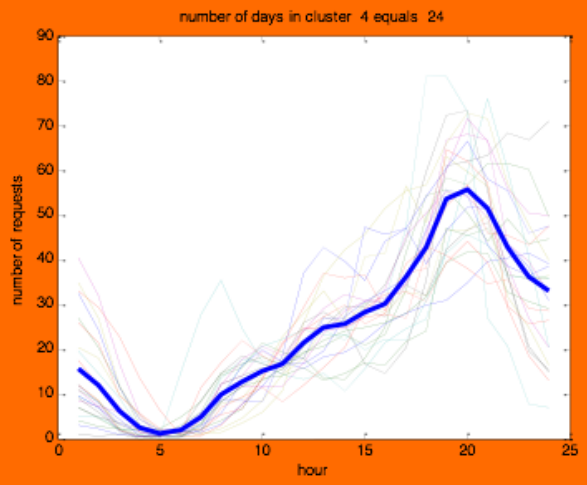




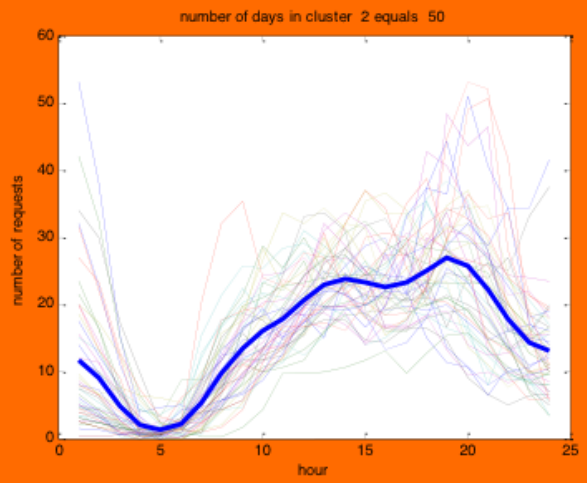


Location Based Behavior Analysis – Using K-Means

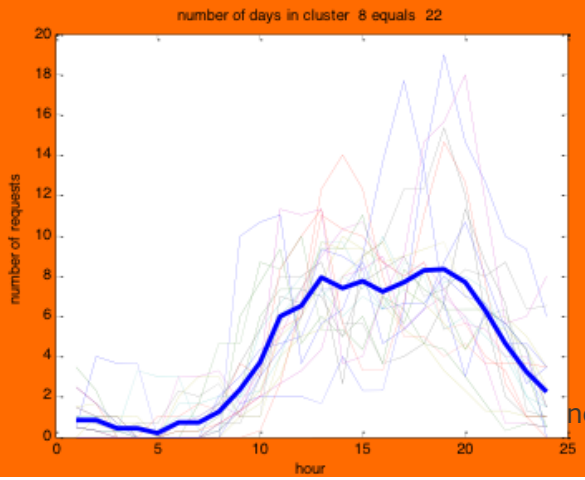
Friday & Sat



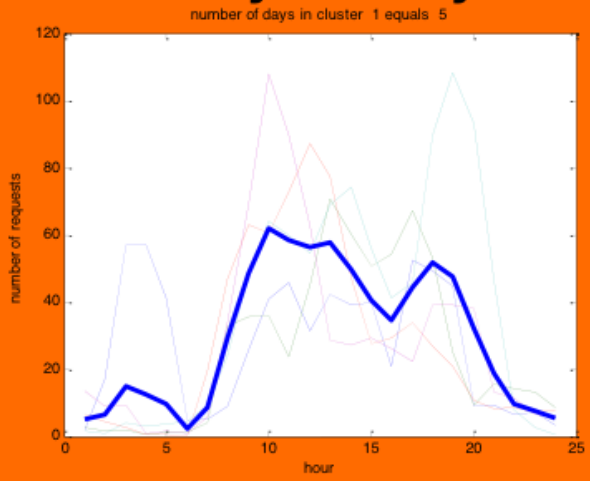
Normal weekday



Sunday - Suburb

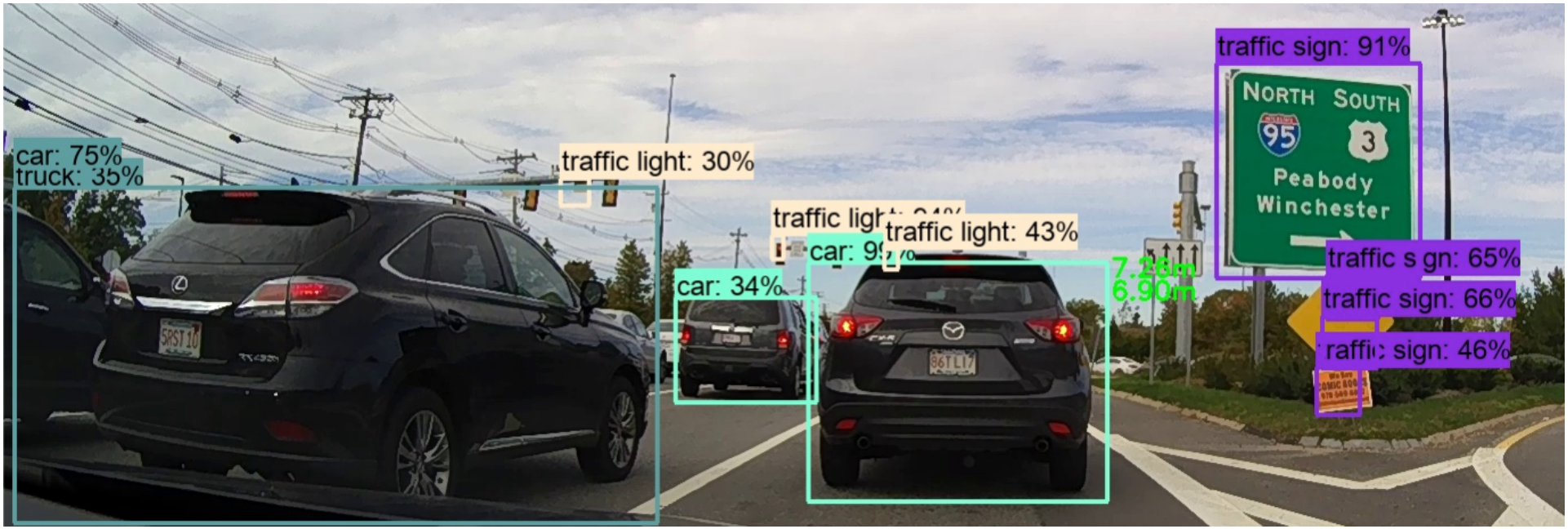


Crazy Monday

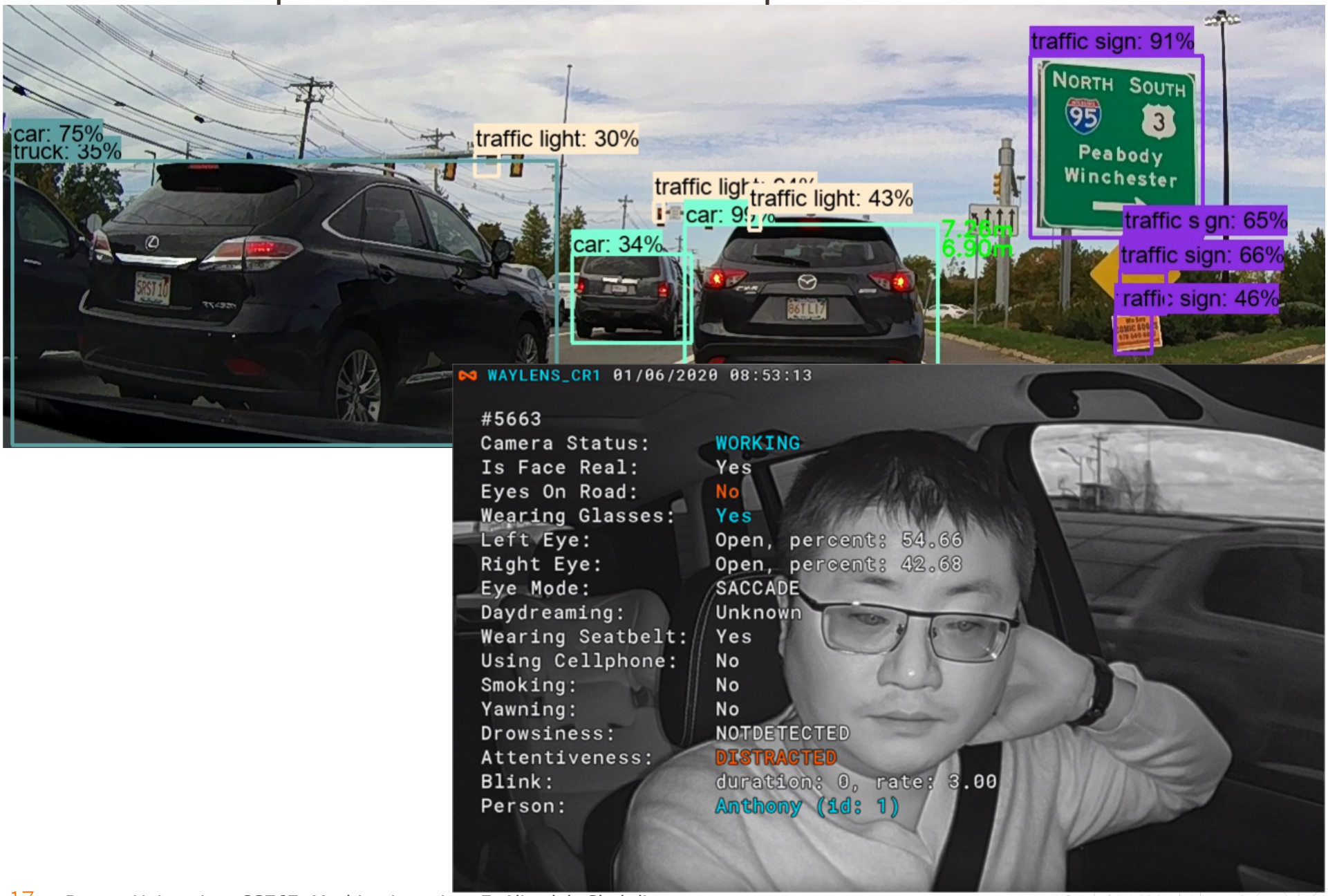


Computer Vision Object Detection

Edge-AI Cameras



Computer Vision Object Detection



Examples of Deep Learning

- Weather prediction
- USPS handwriting application
- Speech recognition
- Face recognition
- Object detection
- Medical imaging analysis
- Biogentics
- Extracting user' behaviors and user segmentation
    - Virtual behavior
    - Shopping behavior
    - Daily trip behavior

Machine Learning Tsunami

Social Networks and Ad Companies – User segmentation and users' personas creation

Autonomous vehicle industry

Recommendation engine

Biotech – drug development

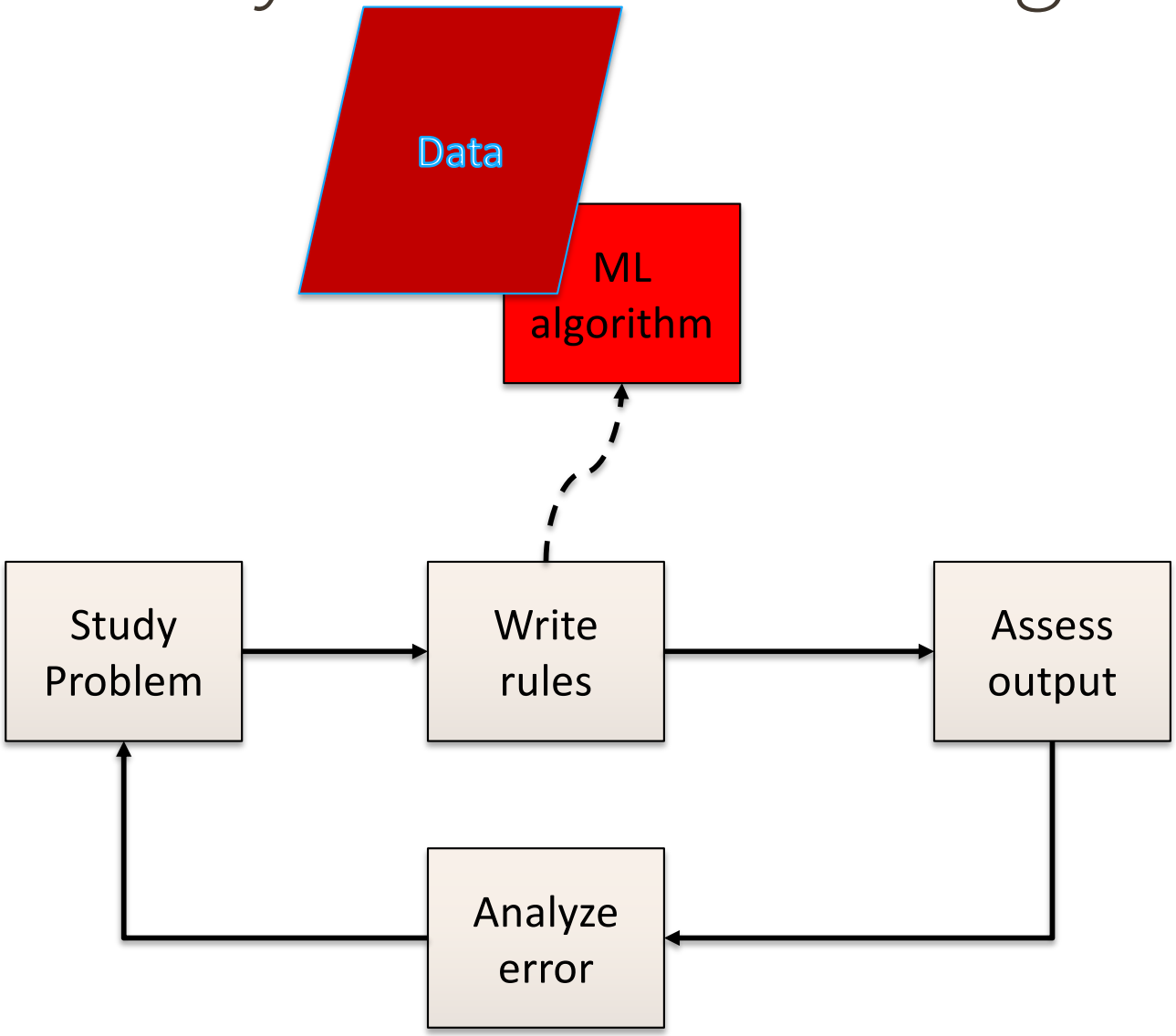
What is Machine Learning?

Learning from data

Why Machine Learning?

- Dynamically adjust an algorithm to get better or adopt.
- Data availability

Why Machine Learning?



About The Course

Philosophy of the Course

- Learn ideas behind the techniques
  - Helps to apply them correctly
- Being able to learn the new approaches, since this is a very dynamic field
- Being able to assess algorithms and help you to go for the simple ones. Ocham's razor!
- Going through simpler algorithms to understand the complicated ones
  - Learning concepts, tools and intuition

What does the Course Cover?

Applied and hands on approach to learning "Deep Learning"

- Concepts and intuition
- Tools
- Implementation with Keras and Tensorflow

Introduction

Course Material

- Course material
  - Notes and material are the main source
  - Notes always available on the same day of the class (WHY? ☺)
  - Text book is a good additional resource

    Hands-on Machine Learning with SciKit-Learn, Keras, and TensorFlow, By Aurelien Geron, Second edition.
  - So many material out there!

- Academic code of conduct

Admin

- Pre-requisite: knowing Python

- All the material will be posted on Blackboard
- There will be quizzes
- Assignments every week, which are due before the next class
  - No late assignment will be accepted
- Office hours: by appointment
- TA intro – TA will also announce their office hour

Grading

- Grading
  - Active class participation – 10%
  - Assignments and Quizzes – 30%
    - Each assignment and quiz has the same weight
  - Term project – 30%
    - To apply what we learn yourself - **PLUS** additionally you can apply a approaches
    - You have to choose a dataset
    - Projects are individual
  - Final exam – 30%
    - From the entire class

Deep Learning System

Flow Chart of Deep Learning System

- Data ingestion
- Data validation
- Data cleaning (Missing values, outlier detection, duplicates)
- Data sensing and visualization
- Re-sampling
- Data annotation and data labeling
- Data normalization
- Data preprocessing / Data Augmentation
- Deep learning model
- Model assessment
- Monitor the system

Machine Learning System – Data Cleaning

- Data cleaning
  - Remove missing fields
  - Remove the entire attribute of missing fields
  - Assign zero, or mean or median to missing fields
- Data preparation – how to deal with categorical attributes
  - Add dummy variable for categorical attributes (if there are N values, N-1 would be sufficient)
  - One-Hot encoding

Deep Learning System - Normalization

- Data normalization
  - Min-max: $\frac{x-min}{max-min}$
  - Standardization – Z-score: $\frac{x-Avg}{\sigma}$
  - Mean normalization: $\frac{x-avg}{max-min}$
  - Scaling to unit length: $\frac{x}{|x|}$
  - Max normalization: $\frac{x}{max}$

Deep Learning Problems

- ML types
  - Classification
  - Regression

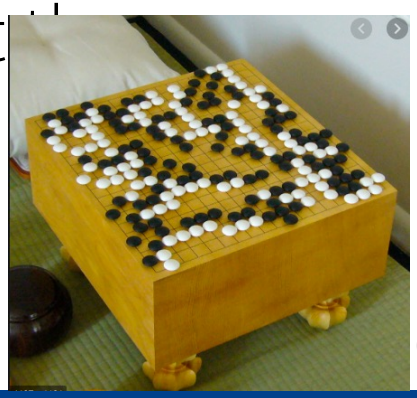
- Model
  - Instance based learning
  - Model based learning

Deep Learning Systems – Learning Approaches

- From training data's perspective
  - Supervised
    - Accurately predict general population
    - Understand impact of parameters on the outcome
    - Understand how parameters impact the outcome
    - Quality measurement of predictions and inferences
  - Unsupervised
    - Grouping with features
    - Objective is not Why?
    - Difficult to assess the results
    - Very powerful since no label data needed
    - Opens new dimensions that we don't even know!
  - Semi-supervised

Deep Learning Systems – Learning Approaches (cont)

- Reinforcement learning
  - Learning system with an "agent" who performs actions
  - Learning by observing the environment
  - Based on returned rewards or penalty of an action learns
- Example – Robotics
  - Game of Go when a program AlphaGo beat
    world champion in May 2017



Deep Learning Systems – Learning Sequences

- From learning sequence's perspective
  - Online Learning
  - Batch Learning

Deep Learning Model Design – An Iterative Process

Building a fresh model

1. Model selection
2. Hyperparameter selection
3. Model training
4. Model evaluation
5. Go to step one or two

Build upon an existing model

1. Base model selection
2. Hyperparameter selection
3. Model training
4. Model evaluation
5. Go to step one or two

Fitness Function - Regression

- Optimization measure – fitness function
  - Regression
    - Root Mean Square Error (RMSE)
    - Mean Absolute Error (MAE)
    - Generally, L_i norm
    - Huber Loss, which converges fast when it is close to the solution, and reduces impact of outliers when far from the solution

$$L_{\delta(x)} = \begin{cases} \frac{1}{2}x^2, & for \ \ |x| \leq \delta \\ \frac{1}{2}\delta^2 + \delta(|x| - \delta), & Otherwise \end{cases}$$

Reference: Wikipedia

Regression Assessment

- Fitness functions still can be used
- Or family of assessment like
  - R squared

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{Reg SS}}{\text{Total SS}}$$

  - Adjusted R squared

$$R^2{}_{adj} = 1 - \left(\frac{Res\ SS / (n-k-1)}{Total\ SS / (n-1)}\right)$$

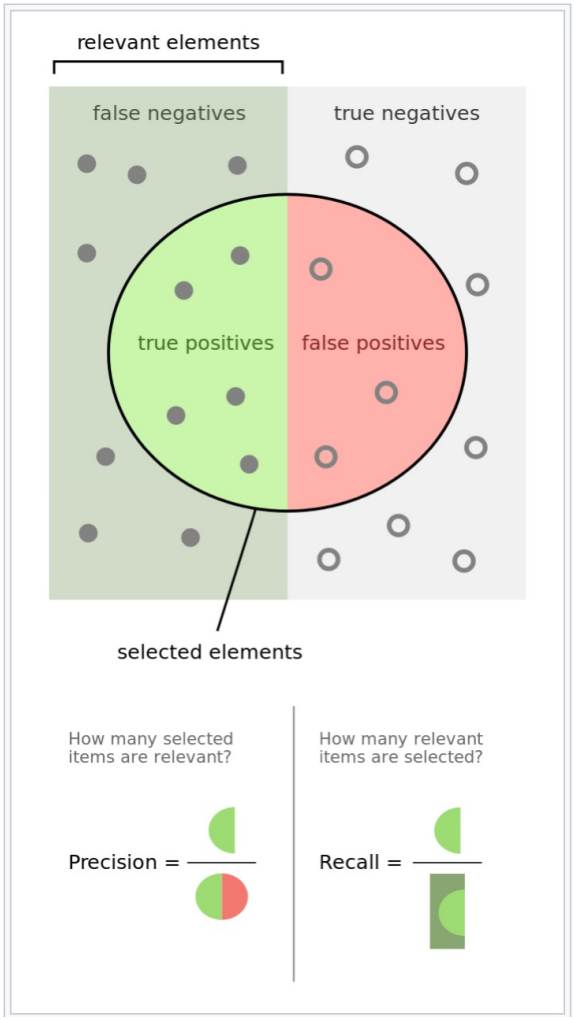
  - F-test

$$F = \frac{Reg\ SS / Reg\ df}{Res\ SS / Original\ model\ \ df}$$

Reference: Wikipedia

Fitness Function - Classification

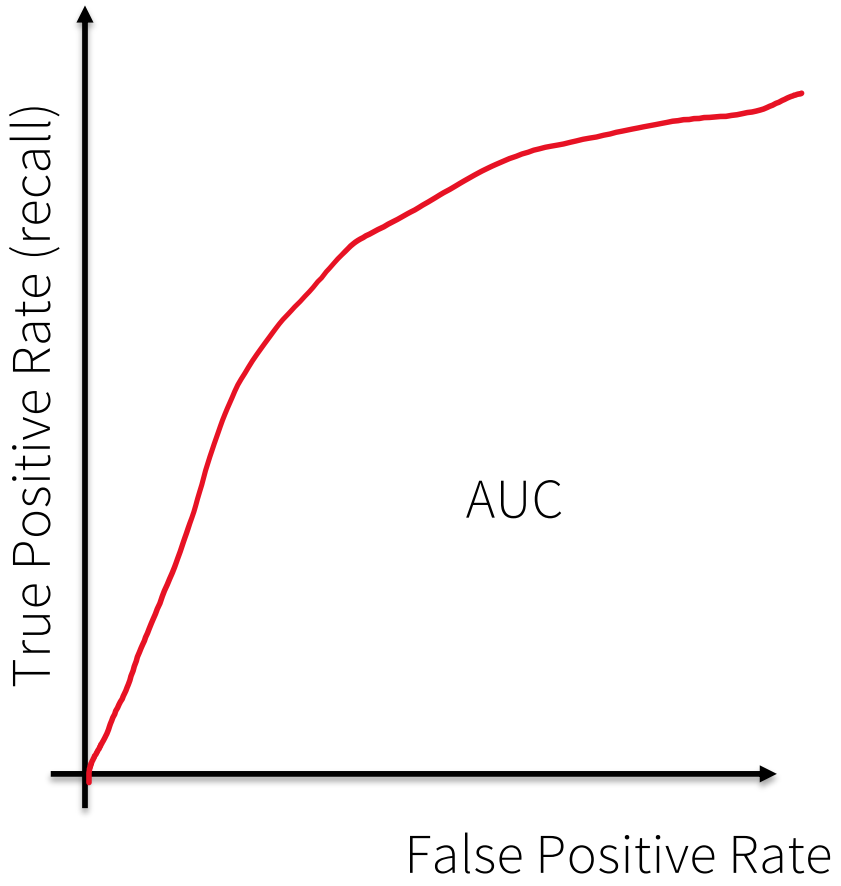
- Classification - Confusion matrix
  - Precision = TP / (TP+FP)
  - Recall = TP / (TP+FN)
  - Precision-recall tradeoff
  - F1 score
- Multi-classification / multilabel-classification
  - Confusion matrix
  - Average F1 score
  - Hamming loss: avg of total number of times label is wrong
  - Ranking loss: avg of proportion of label wrong



Reference: Wikipedia

Performance Measure of Classification

- ROC (Receiver Operating Characteristic):
  - True positive rate (recall) vs False positive rate
- AUC – Area under the curve



Machine Learning Algorithm

- Training, test and validation data set
- Performance characteristics
  - Accuracy
  - Sensitivity

Data or Machine Learning

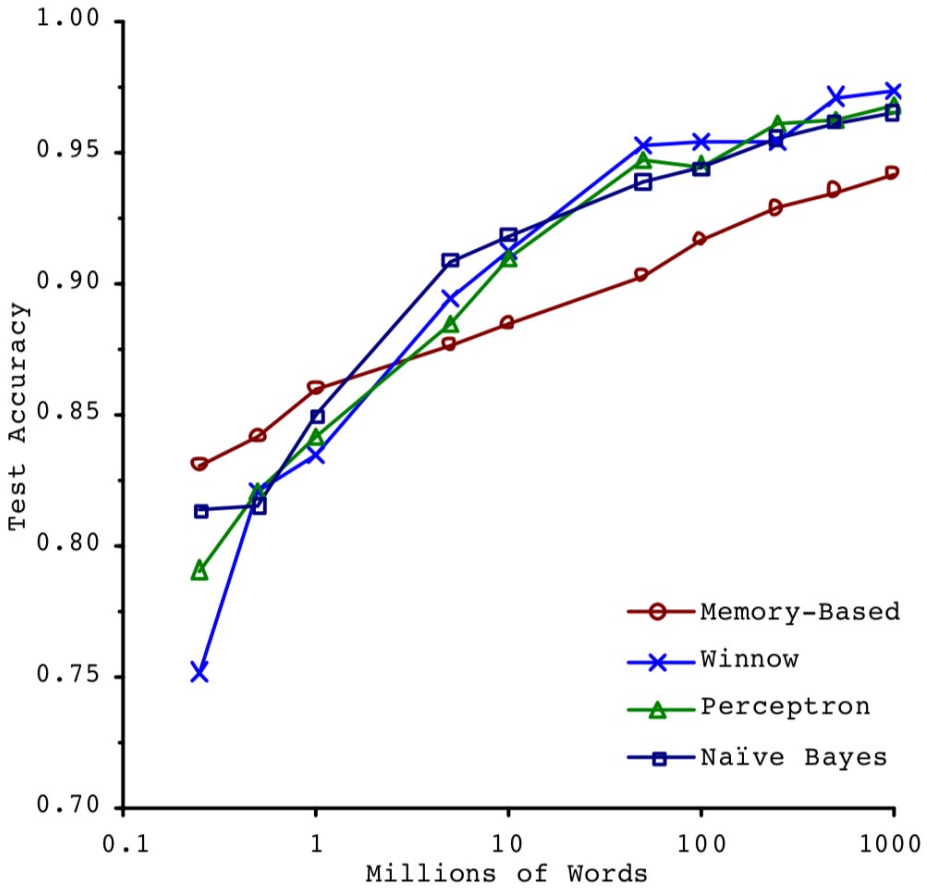


Figure 1. Learning Curves for Confusion Set Disambiguation

Michele Banko, Eric Brill, "Scaling a very very large Corpora for natural language disambiguation", Microsoft Research Lab, 2001

Data Traps

- Insufficient data
- Non-representative training data
- Poor quality data
- Irrelevant features
- Over-fitting
- Under-fitting

Example

Goal is:

predicting score based on # of hours studied

```
> student
       name hours score
1     Allen  8.00    80
2     Brown  2.00    67
3      Cole  9.00    98
4   Collins  9.00   100
5    Cooper  7.00    86
6       Cox  6.50    95
7      Hall  4.00    67
8      Hans  1.00    59
9    Howard  7.00    89
10  Jeffers  0.00    70
11  Johnson  1.00    50
12    Jones  1.50    66
13     King  4.50    81
14   Knight  4.00    69
15      Lee  3.50    71
16   Martin  3.00    66
17   Miller  3.00    75
18    Moore  3.00    60
19   Morris  4.00    67
20   Murphy  6.00    87
21     Reed  5.00    75
22    Smith  0.50    50
23  Stewart  5.00    68
24   Taylor  2.25    53
25   Thomas  2.00    54
26 Thompson  3.50    69
27   Walker  3.50    67
28     Ward  9.00   100
29 Williams  1.00    50
30   Wright  4.75    68
31    Young  4.00    79
```

Example Predicting Scores using Linear Regression

- Perfect training with zero error

  $R^2 = 1!$ PERFECT!

```
Residuals:
ALL 31 residuals are 0: no residual degrees of freedom!

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)        80         NA      NA       NA
nameBrown         -13         NA      NA       NA
nameCole           18         NA      NA       NA
nameCollins        20         NA      NA       NA
nameCooper          6         NA      NA       NA
nameCox            15         NA      NA       NA
nameHall          -13         NA      NA       NA
nameHans          -21         NA      NA       NA
nameHoward          9         NA      NA       NA
nameJeffers       -10         NA      NA       NA
nameJohnson       -30         NA      NA       NA
nameJones         -14         NA      NA       NA
nameKing            1         NA      NA       NA
nameKnight        -11         NA      NA       NA
nameLee            -9         NA      NA       NA
nameMartin        -14         NA      NA       NA
nameMiller         -5         NA      NA       NA
nameMoore         -20         NA      NA       NA
nameMorris        -13         NA      NA       NA
nameMurphy          7         NA      NA       NA
nameReed           -5         NA      NA       NA
nameSmith         -30         NA      NA       NA
nameStewart       -12         NA      NA       NA
nameTaylor        -27         NA      NA       NA
nameThomas        -26         NA      NA       NA
nameThompson      -11         NA      NA       NA
nameWalker        -13         NA      NA       NA
nameWard           20         NA      NA       NA
nameWilliams      -30         NA      NA       NA
nameWright        -12         NA      NA       NA
nameYoung          -1         NA      NA       NA
hours              NA         NA      NA       NA
id                 NA         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
```