Beginner's guide to Machine Learning competitions

EuroPython 2015

Christine Doig

Slides bit.ly/ep2015-ml-tutorial

Notebooks bit.ly/ep2015-ml-tutorial-repo

Christine Doig

Data Scientist, Continuum Analytics

ch_doig

chdoig

chdoig.github.io

Concepts

- Data Science
  - Machine Learning
    - NLP
    - Sentiment analysis
  - Supervised learning
  - Classification

Setup

- Anaconda
- Kaggle
  - Competitions
  - Dataset

Process

- Feature preparation
  - Modeling
    - Optimization
      - Validation

45min

10 min

1h

5 min

1h

Concepts

- Data Science
- Machine Learning
- Supervised learning
- Classification
- NLP
- Sentiment analysis

Setup

- Anaconda
- Kaggle
- Competitions
- Dataset

Process

- Feature preparation
- Modeling
- Optimization
- Validation

Concepts

Data Science

Machine Learning

Supervised learning

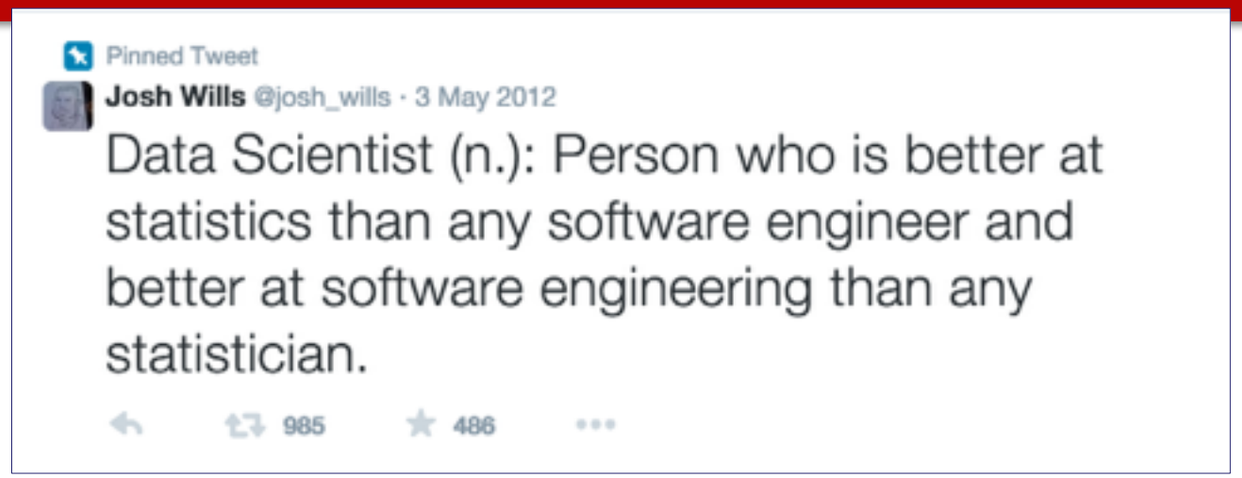
Classification

NLP

Sentiment analysis

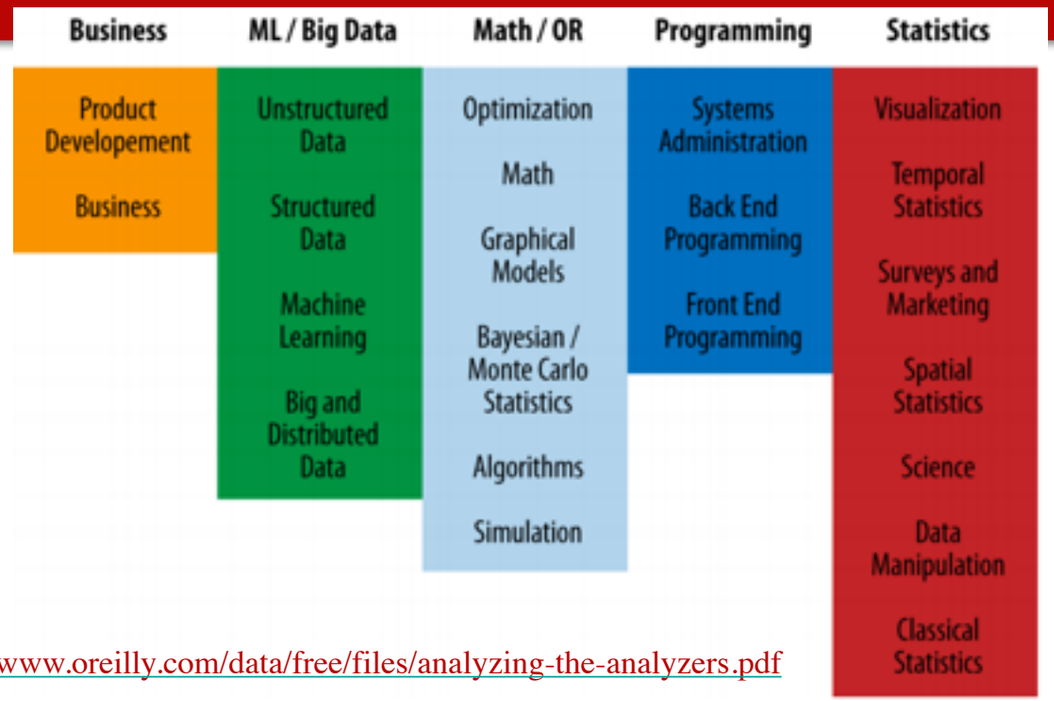
Data Science


I IS DATA SCIENTIST
Troll.me



From the lab to the factory - Data Day Texas

Slides: http://www.slideshare.net/joshwills/production-machine-learninginfrastructure
Video: https://www.youtube.com/watch?v=v-91JycaKjc



http://www.oreilly.com/data/free/files/analyzing-the-analyzers.pdf

data science

http://www.fico.com/landing/infographic/anatomy-of-a-data-scientist_en.html

http://www.experfy.com/blog/become-data-scientist/

Machine Learning/Stats

Scientific Computing

Analytics

data science

Distributed Systems

Web

Machine Learning/Stats

Data Scientists/
Modeler

Scientific Computing

Research/Computational
Scientist

data science

Analytics

Data/Business
Analyst

Distributed Systems

Data Engineers/
Architects

Web

Developer

data science

Machine Learning/Stats

Data Scientists/ Modeler

Model

Scientific Computing

Research/Computational Scientist

Algorithm

Analytics

Data/Business Analyst

Report

Distributed Systems

Data Engineers/ Architects

Pipeline/ Architecture

Application

Web

Developer

data science

Machine Learning/Stats

Data Scientists/ Modeler

Regression
Models
KNN
Supervised
Unsupervised
SVM
NN
Clustering
Crossvalidation
Classification
Deep Learning
Dimensionality reduction

Scientific Computing

Research/Computational Scientist

GPUs
arrays
HPC
algorithms
FTT
compute
DFT
optimization
graphs
algebra
performance
parallelism
parallelism

Analytics

Data/Business Analyst

min
max
summary statistics
Join
avg
Filter
Sort
databases
Groupby
Select
SQL
TopK
Reporting

Distributed Systems

Data Engineers/ Architects

hdfs
hadoop
spark
pipeline
consistency
batch
clusters
availability
stream processing
tolerance
jobs

Web

Developer

sematic web
apps
interactive data viz
databases
A/B testing
HTML/CSS/JS
NOSQL
frontend
servers
crawling
cloud
deployment
scraping
frameworks

data science

Machine Learning/Stats

Data Scientists/ Modeler

R

PyMC

scikit learn

StatsModels Statistics in Python

PyBrain

theano

Scientific Computing

Research/Computational Scientist

Numba

SciPy

NumPy

xray

Cython

HDF

PyTables

Analytics

Data/Business Analyst

PostgreSQL

pandas

xlwings

SQLAlchemy

Distributed Systems

Data Engineers/ Architects

Spark

Kafka

mrjob

Luigi

Storm

mrjob

hadoop

Web

Developer

D3

Scrapy

Bokeh

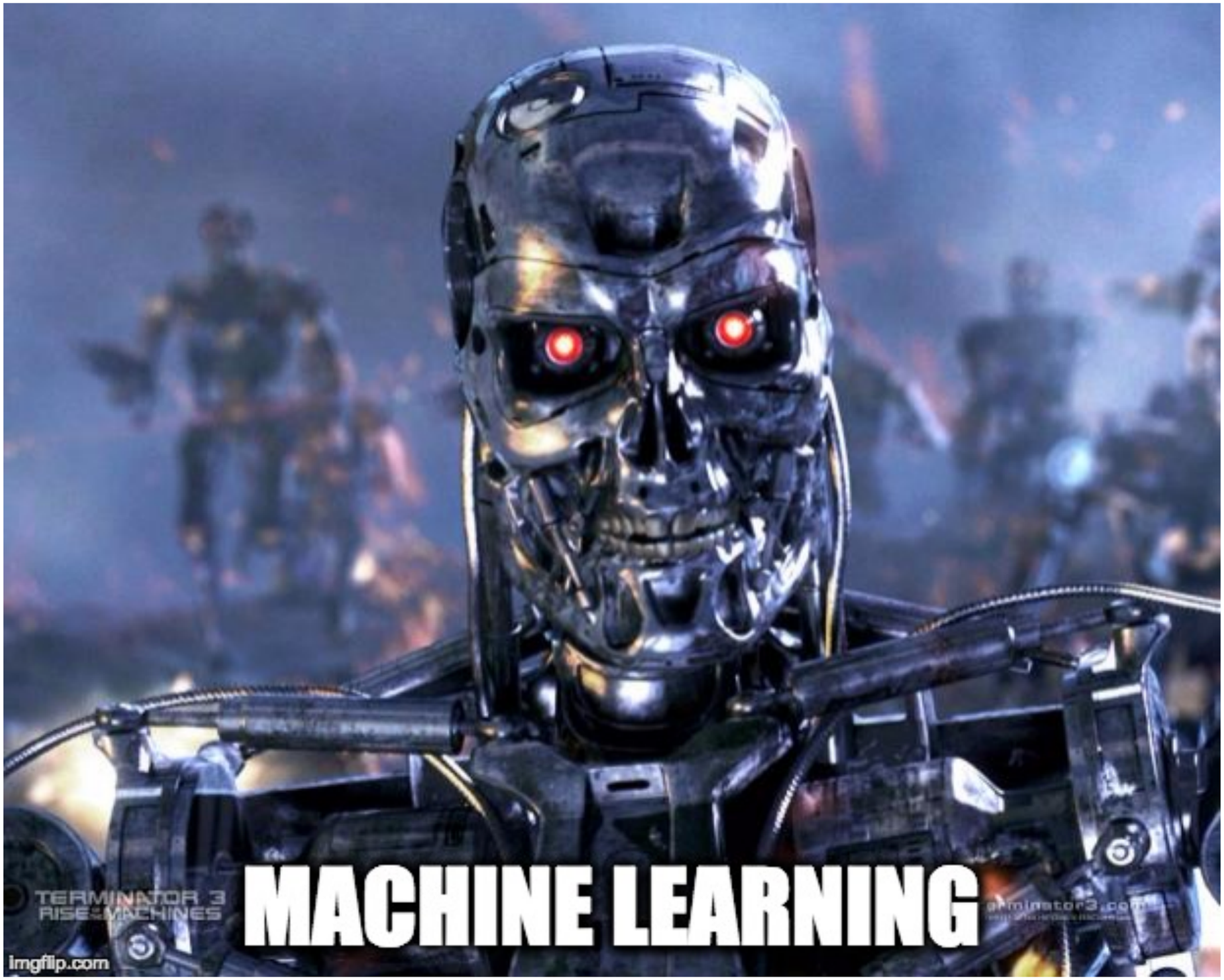
Flask web development, one drop at a time

jupyter

RDFLib

django

Machine Learning


TERMINATOR 3
RISE OF THE MACHINES
MACHINE LEARNING
imgflip.com

Machine Learning

no labels — *labels*

Unsupervised learning

Clustering

- K-means
- Hierarchical clustering
- *Topic modeling

Latent variables/structure

- Dimenstionality reduction
- *Topic modeling

Supervised learning

categorical — *quantitative*

Classification

- Logistic regression
- SVM
- Decision trees
- k-NN

Regression

- Linear regression

Machine Learning

no labels — *labels*

Unsupervised learning

Clustering

- K-means
- Hierarchical clustering
- *Topic modeling

Latent variables/structure

- Dimenstionality reduction
- *Topic modeling

Exploratory

Supervised learning

categorical — *quantitative*

Classification

- Logistic regression
- SVM
- Decision trees
- k-NN

Regression

- Linear regression

Predictive

Machine Learning

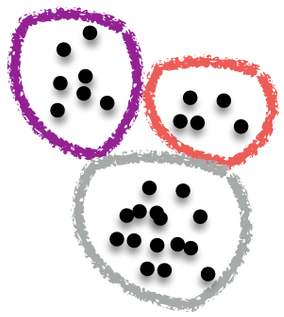
no labels

labels

Unsupervised learning

id	gender	age	job_id
1	F	67	1
2	M	32	2
3	M	45	1
4	F	18	2

Exploratory



group similar individuals together

Supervised learning

categorical

quantitative

Classification

Regression

id	gender	age	job_id	buy/click_ad	money_spent
1	F	67	1	Yes	$1,000
2	M	32	2	No	-
3	M	45	1	No	-
4	F	18	2	Yes	$300

Predictive

Classification

predict whether an individual is going to buy/click or not

Regression

predict how much is the individual going to spend

Natural Language Processing

Machine Learning

→ Natural language processing

field concerned with the interactions between computers and human (natural) languages

tasks

→ Sentiment analysis

Extract subjective information on polarity (positive or negative) of a document (text, tweet, voice message…)

e.g online reviews to determine how people feel about a particular object or topic.

Machine Learning

no labels

labels

Unsupervised learning

Clustering

- K-means
- Hierarchical clustering
- *Topic modeling

Latent variables/structure

- Dimenstionality reduction
- *Topic modeling

Supervised learning

categorical

quantitative

Classification

- Logistic regression
- SVM
- Decision trees
- k-NN

Regression

- Linear regression

Sentiment analysis

e.g.

Movie review → Positive 

Movie review → Negative 

I love you! → Positive 

Setup

Setup

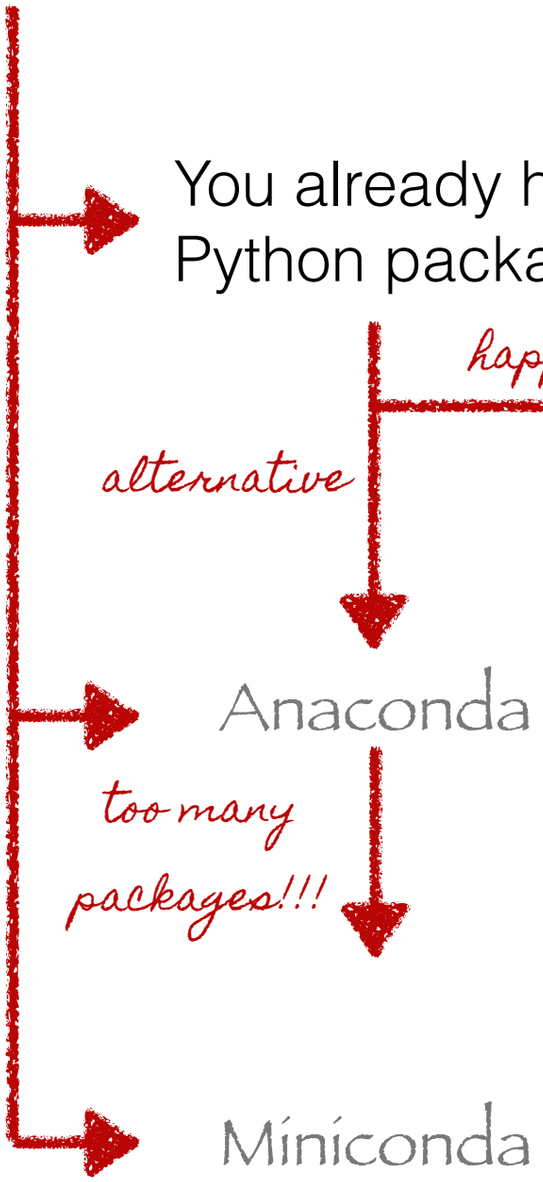
Anaconda

Kaggle

Competitions

Dataset

Setup options



http://continuum.io/downloads

Python + conda (package manager) + packages

Free Python distribution with a bunch of packages for data science

Miniconda + conda env

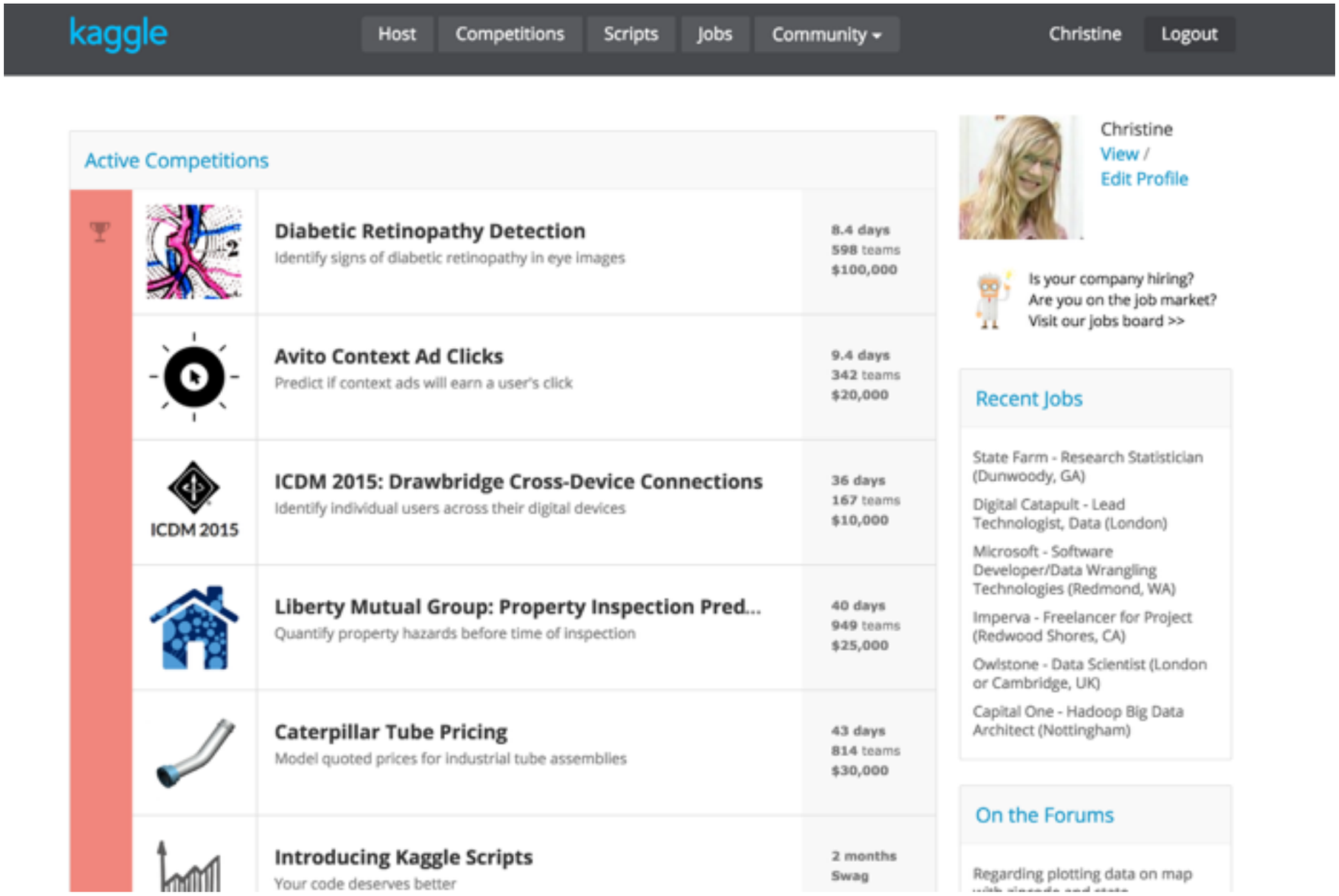
http://conda.pydata.org/miniconda.html

Python + conda (package manager)

```
git clone git@github.com:chdoig/ep2015-ml-tutorial.git
cd ep2015-ml-tutorial
conda env create
source activate ep-ml
```

Kaggle

https://www.kaggle.com/



hosts online machine learning competitions

Kaggle Competition

https://www.kaggle.com/c/word2vec-nlp-tutorial

kaggle | Host | Competitions | Scripts | Jobs | Community | Christine | Logout

Completed • Knowledge • 579 teams

Bag of Words Meets Bags of Popcorn

Tue 9 Dec 2014 – Tue 30 Jun 2015 (18 days ago)

Dashboard

Home
- Data
- Make a submission

Information
- Description
- Evaluation
- Rules
- Part 1: For Beginners - Ba...
- Part 2: Word Vectors
- Part 3: More Fun With Wo...
- Part 4: Comparing deep a...
- Setting Up Your System
- What is Deep Learning?

Forum

Leaderboard

My Team
- GitHub

My Submissions

Competition Details » Get the Data » Make a submission

Data Files

File Name	Available Formats
sampleSubmission	.csv (276.17 kb)
unlabeledTrainData.tsv	.zip (25.98 mb)
testData.tsv	.zip (12.64 mb)
labeledTrainData.tsv	.zip (12.96 mb)

Data Set

The labeled data set consists of 50,000 IMDB movie reviews, specially selected for sentiment analysis. The sentiment of reviews is binary, meaning the IMDB rating < 5 results in a sentiment score of 0, and rating >=7 have a sentiment score of 1. No individual movie has more than 30 reviews. The 25,000 review labeled training set does

Kaggle Competition

Bag of Words Meets Bags of Popcorn

https://www.kaggle.com/c/word2vec-nlp-tutorial

Data

50,000 IMDB movie reviews

labeledTrainData.tsv

25,000 rows containing an id, sentiment, and text for each review.

testData.tsv

25,000 rows containing an id and text for each review

Task

predict the sentiment for each review
in the test data set

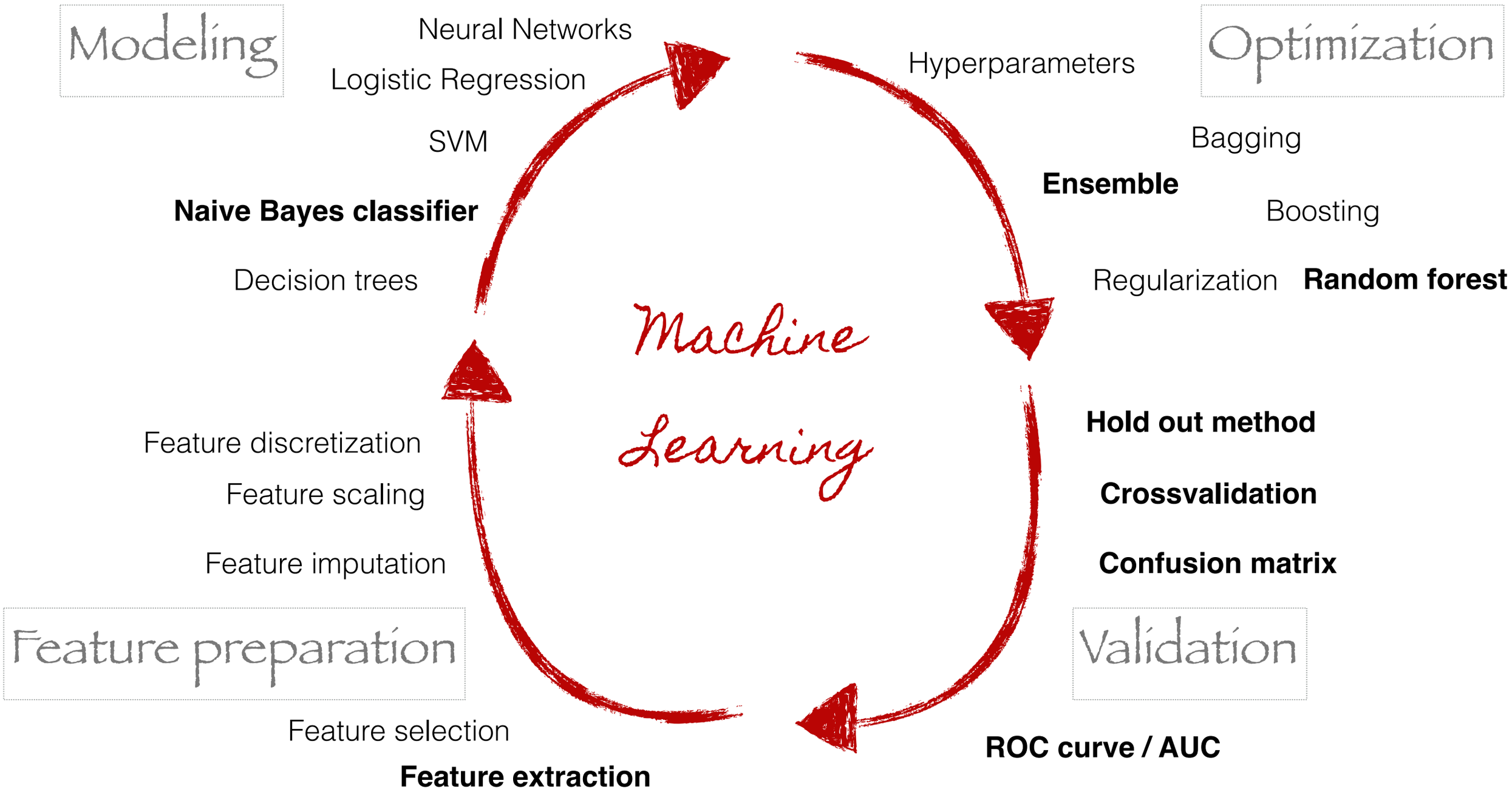
Process

Feature preparation

Modeling

Optimization

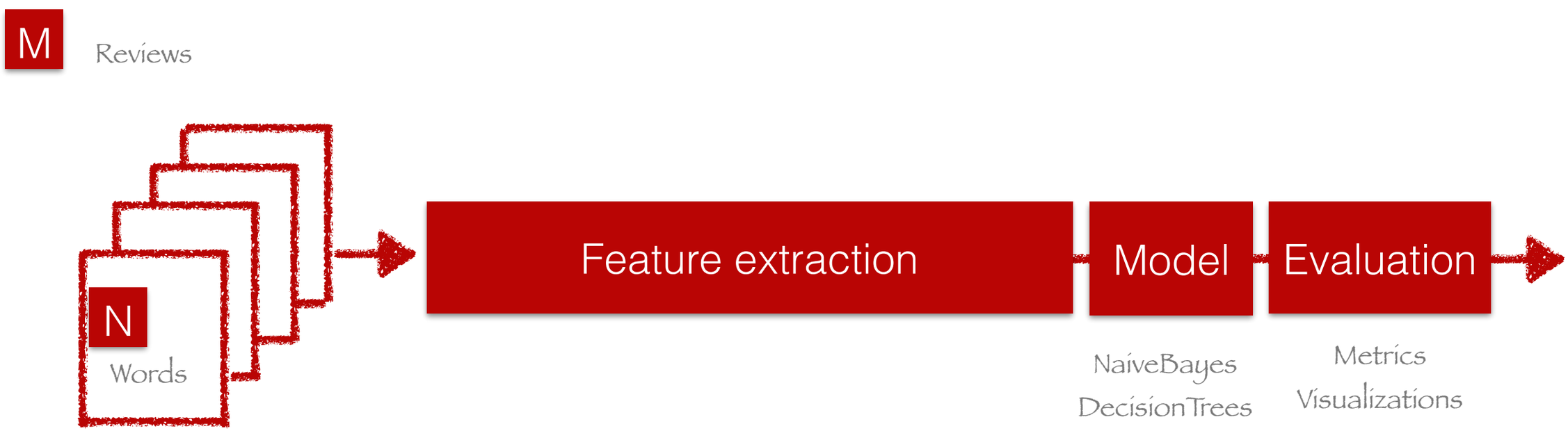
Validation


Modeling
Neural Networks
Logistic Regression
SVM
Naive Bayes classifier
Decision trees
Optimization
Hyperparameters
Bagging
Ensemble
Boosting
Regularization
Random forest
Machine
Learning
Hold out method
Crossvalidation
Confusion matrix
Validation
ROC curve / AUC
Feature preparation
Feature discretization
Feature scaling
Feature imputation
Feature selection
Feature extraction

Feature preparation

Feature extraction

the process of making features from available data
to be used by the classification algorithms



id	sentiment	review	count_words	terrible_word
1	0	the movie was terrible	4	1
2	1	I love it	3	0
3	1	Awesome! Love it!	3	0
4	0	I hated every minute	4	0

Feature extraction
Text

Tokenization

Simple

transition, metal, oxides, considered, generation, materials, field, electronics, advanced, catalysts, tantalum, v, oxide, reports, synthesis, material, nanometer, size, unusual, properties…

Collocations

transition_metal_oxides, considered, generation, materials, field, electronics, advanced, catalysts, tantalum, oxide, reports, synthesis, material, nanometer_size, unusual, properties, sol_gel_method, biomedical_applications…

Entities

transition, metal_oxides, tantalum, oxide, nanometer_size, unusual_properties, dna, easy_method, biomedical_applications

Combination

transition, metal_oxides, generation, tantalum, oxide, nanometer_size, unusual_properties, sol, dna, easy_method, biomedical_applications

Lemmatization

transition, metal, oxide, consider, generation, material, field, electronic, advance, catalyst, property…

Stopwords

language

generic

a
above
across
after
afterwards
again
against
all
…

domain

specific

material
temperature
advance
size
….

Vector Space

Dictionary

1 - transition
2- metal
3- oxides
4- considered
…

Corpus - Bag of words

[(0, 1), (1, 1), (2, 1)]
[(0, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1)]
[(2, 1), (5, 1), (7, 1), (8, 1)]
[(1, 1), (5, 2), (8, 1)]
[(3, 1), (6, 1), (7, 1)]
[(9, 1)]
[(9, 1), (10, 1)]
[(9, 1), (10, 1), (11, 1)]
[(4, 1), (10, 1), (11, 1)]

Feature_extraction.ipynb

Modeling

Naive Bayes Classifier

$$P(A|B) = P(B|A) * P(A) / P(B)$$

id	sentiment	review	count_words	terrible_word
1	0	the movie was terrible	4	1
2	1	I love it	3	0
3	1	Awesome! Love it	1	0
4	0	I hated every minute	4	0

What's the probability of the review being positive if the word love appears in the review?

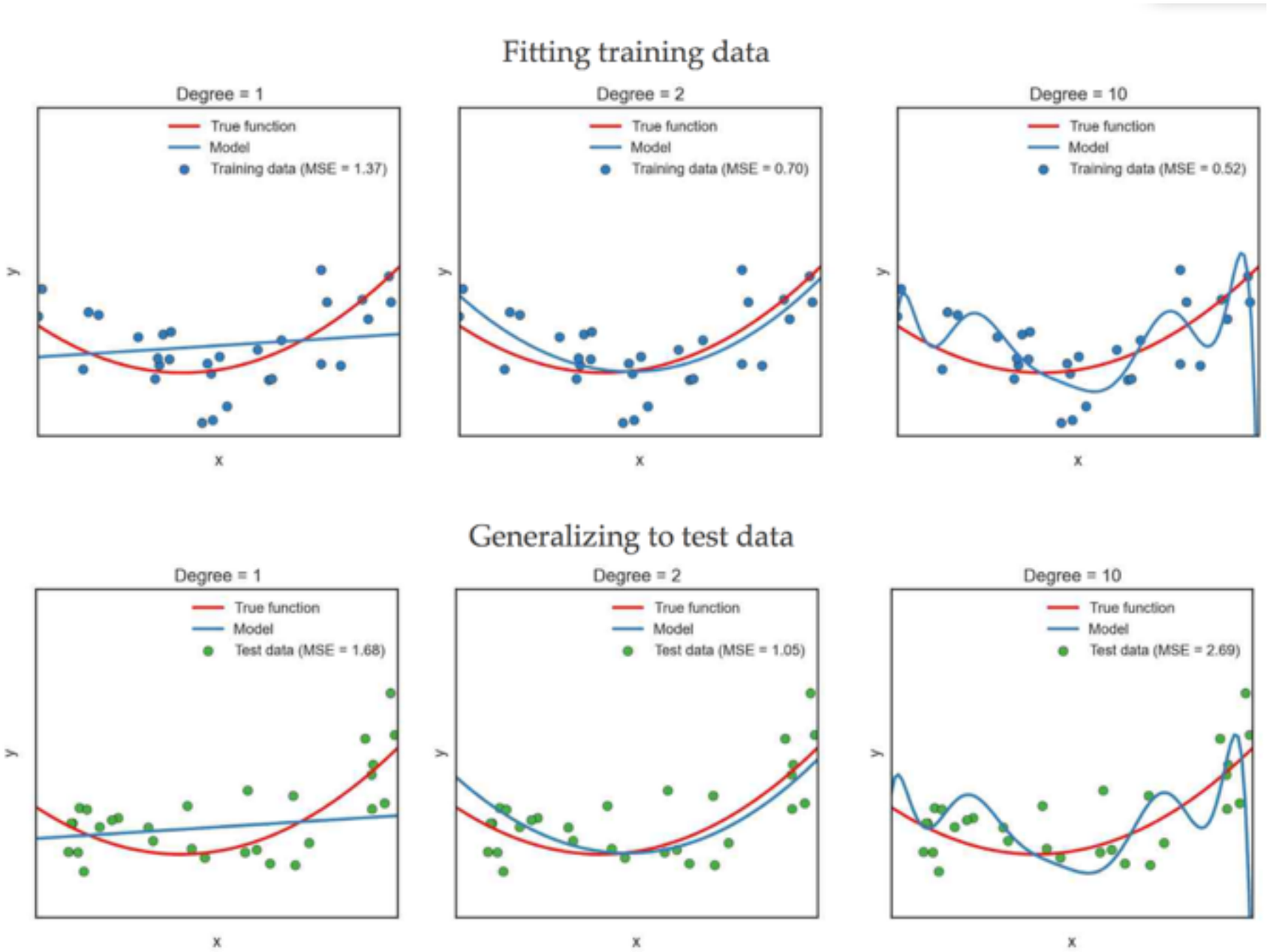
$$P(1 \mid love) = P(love \mid 1) * P(1) / P(love) =$$
$$(2/2 * 2/4)/(2/4) = 100\%$$

Modeling.ipynb

Validation

Overfitting

occurs whenever a model learns from patterns that are present in the training data but do not reflect the data-generating process. Seeing more than is actually there. A kind of data hallucination.



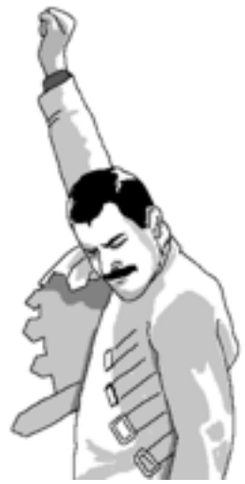
http://talyarkoni.org/downloads/ML_Meetup_Yarkoni_Overfitting.pdf

Validation

Training data

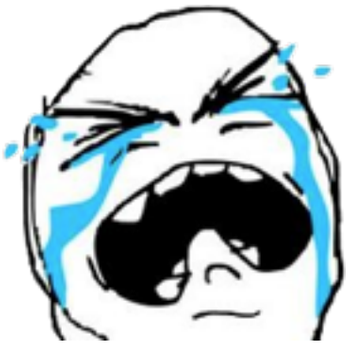
Model

Evaluate



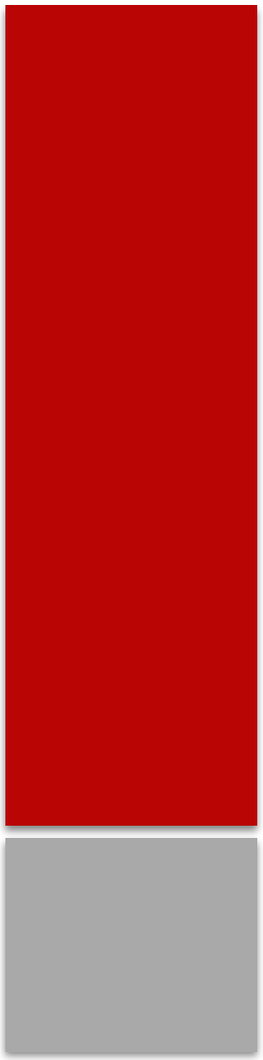
New data

Evaluate



Validation.ipynb

Hold out method



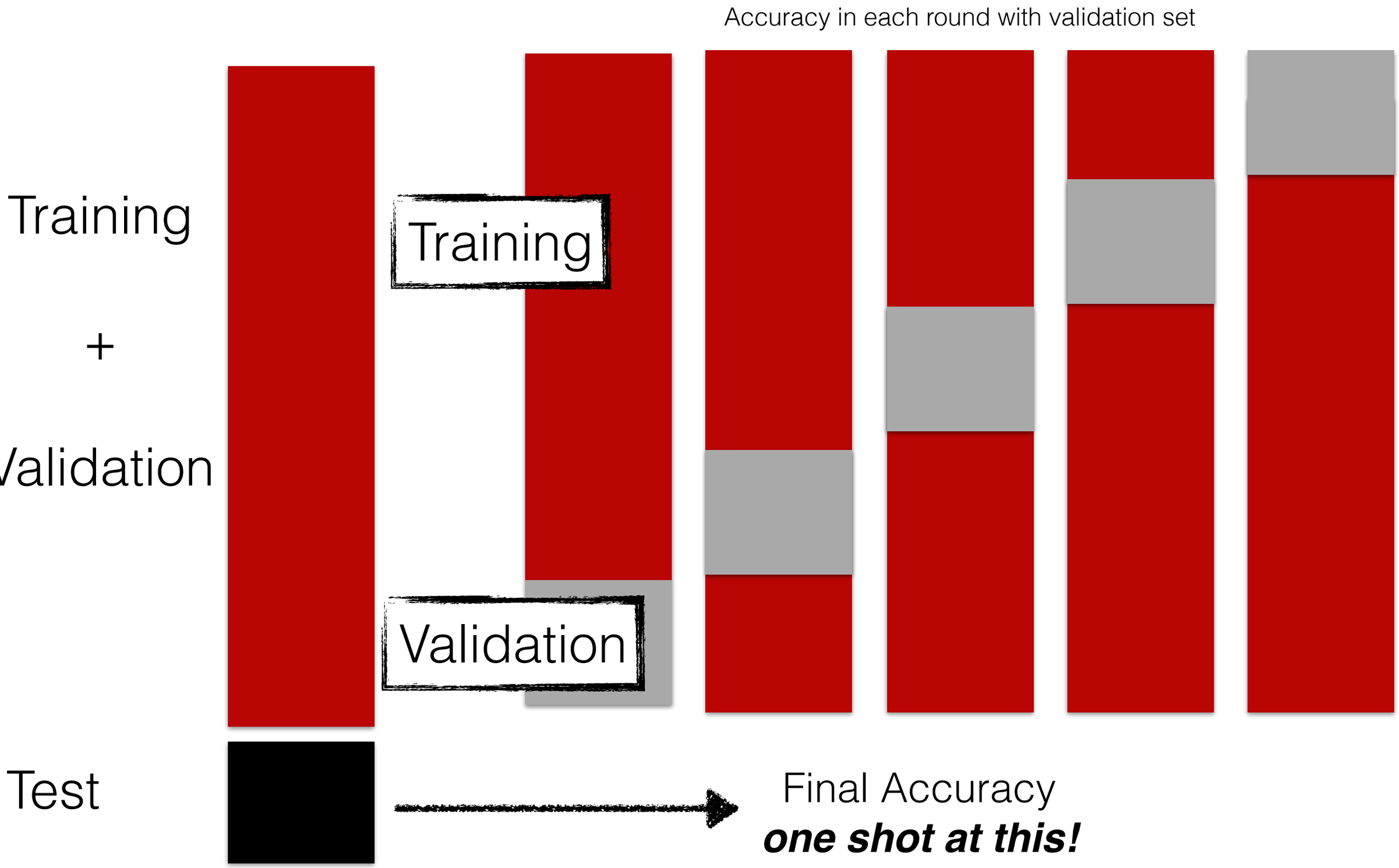
Training data

Test data

accuracy

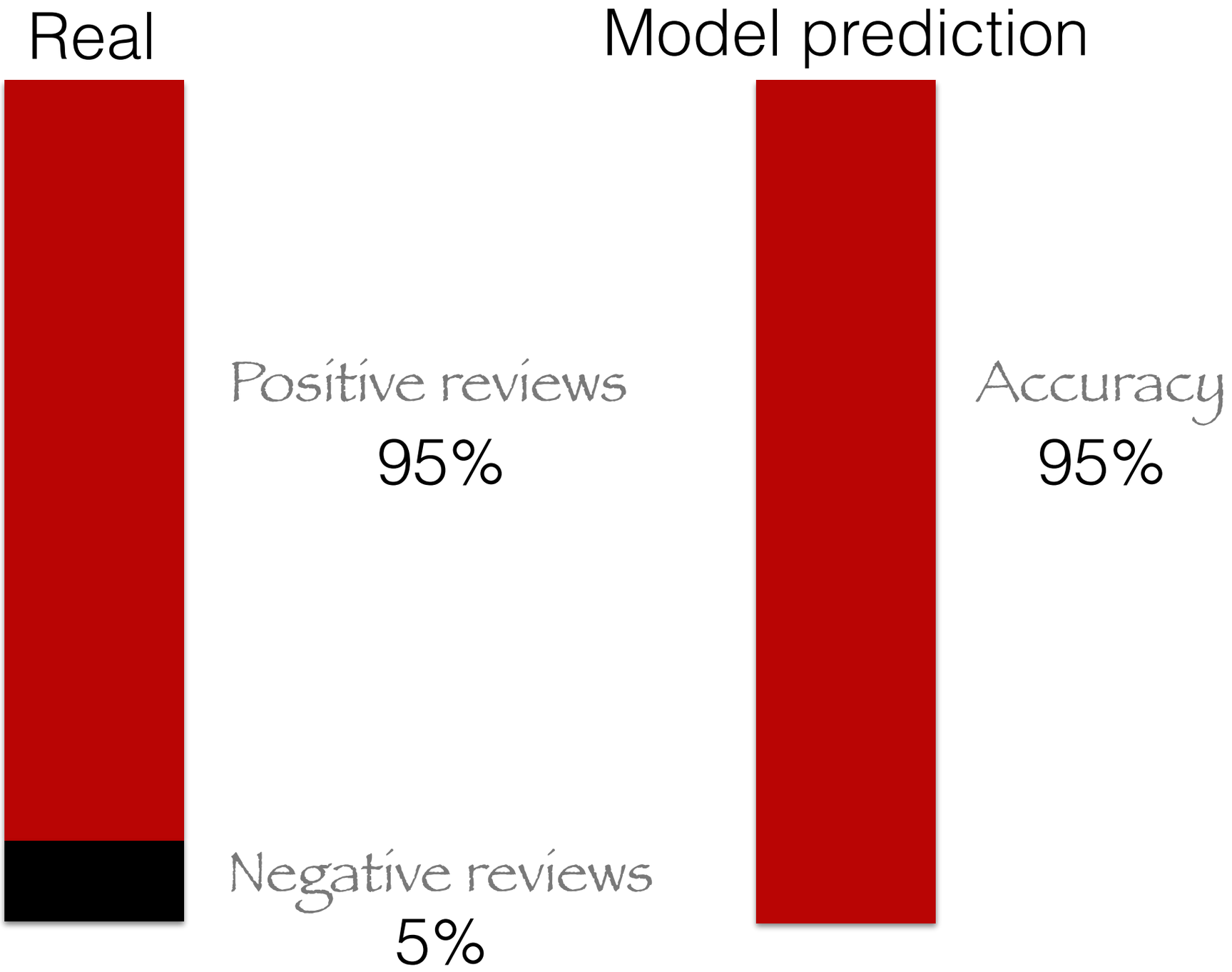
Crossvalidation

Accuracy = average(Round1, Round 2….)



Validation

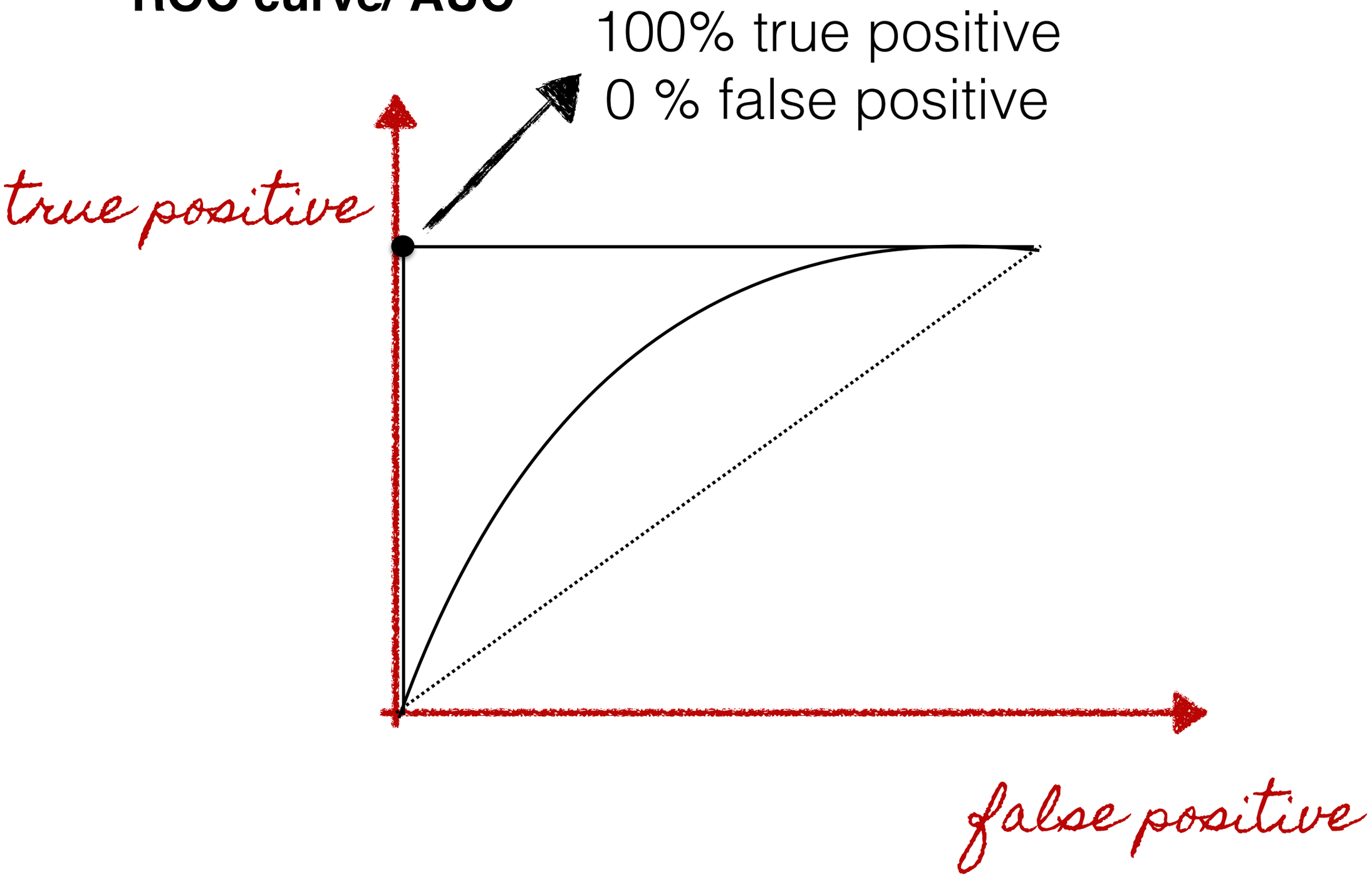
Confusion matrix



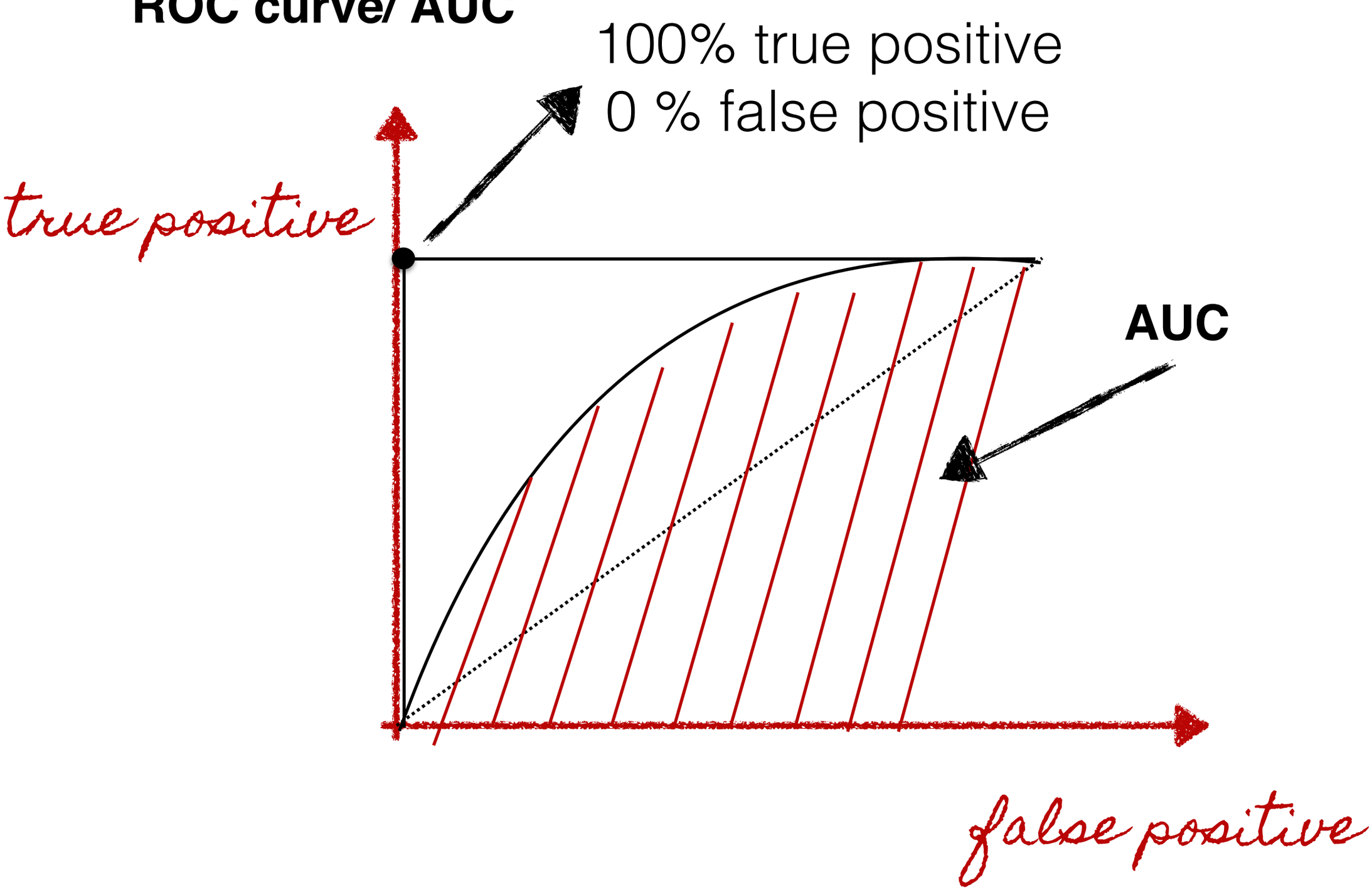
Confusion matrix

model/real	positive	negative
positive	95	5
negative	0	0

ROC curve/ AUC



ROC curve/ AUC



Kaggle leaderboard

kaggle | Host | Competitions | Scripts | Jobs | Community | Christine | Logout

Completed • Knowledge • 579 teams

Bag of Words Meets Bags of Popcorn

Tue 9 Dec 2014 – Tue 30 Jun 2015 (19 days ago)

Dashboard

Public Leaderboard - Bag of Words Meets Bags of Popcorn

This leaderboard is calculated on all of the test data.

See someone using multiple accounts?
Let us know.

#	Δ1w	Team Name * In the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Zhang Yang *	1.00000	2	Mon, 08 Jun 2015 06:01:54
2	—	Puma Gigliotti	1.00000	14	Sat, 20 Jun 2015 02:34:35 (-2.1d)
3	↑214	Abner	1.00000	36	Thu, 25 Jun 2015 07:19:13
4	↑8	Cristian	0.99999	7	Tue, 30 Jun 2015 23:39:07
5	↓2	HeChen	0.99996	1	Tue, 05 May 2015 13:59:05
6	↓2	Alejandro Peláez	0.99259	25	Sun, 10 May 2015 01:10:37 (-0.7h)
7	↓2	Sebastian Raschka_	0.99156	5	Fri, 19 Jun 2015 18:13:32
8	↓2	vgng	0.97663	22	Sun, 28 Jun 2015 23:24:02 (-7.8d)

Ensemble methods

Classifier 1 Classifier 2 Classifier 3

id	cls_1	cls_2	cls_3	ensemble
1	0	0	0	0
2	0	1	1	1
3	1	1	1	1
4	0	0	1	0

e.g. majority voting

w1 w2 w3

e.g. weighted voting

Ensemble.ipynb

Kaggle forums

kaggle | Host | Competitions | Scripts | Jobs | Community ▾ | Christine | Logout

Completed • Knowledge • 579 teams

Bag of Words Meets Bags of Popcorn

Tue 9 Dec 2014 – Tue 30 Jun 2015 (19 days ago)

Dashboard

Competition Forum

Post competition solutions

Start Watching

6

Hi all,

Since the competition is over, it would be great to have a topic to discuss the solutions.

My submission has the LB score of 0.97663. It is based on the ensemble of NBSVM, Paragraph Vector and Gated Recurrent Neural Network. The code is at: **https://github.com/vinhkhuc/kaggle-sentiment-popcorn.git**

I also would like to know about other approaches.

Concepts

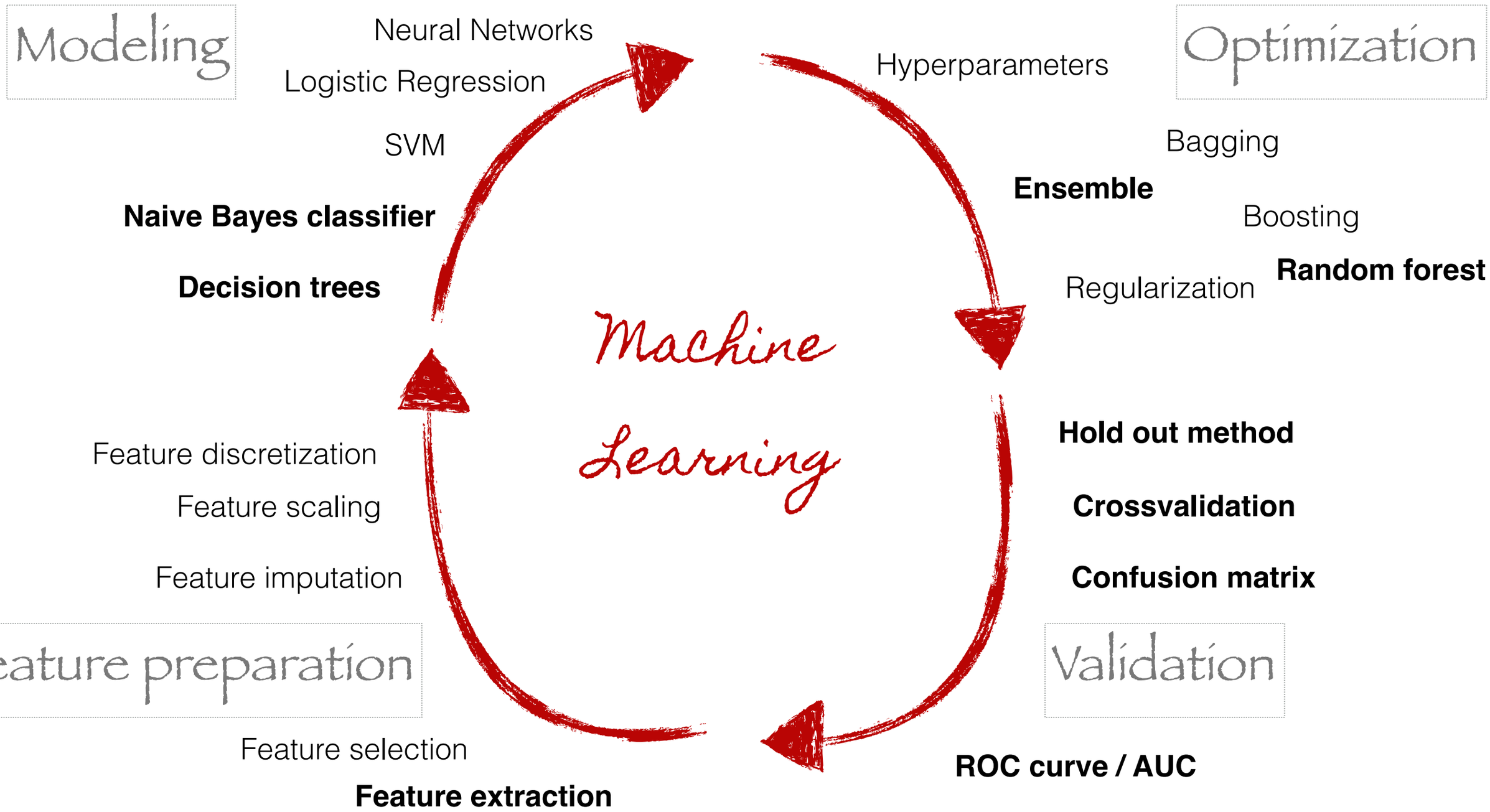
- Data Science
  - Machine Learning
    - NLP
    - Sentiment analysis
  - Supervised learning
  - Classification

Setup

- Anaconda
- Kaggle
  - Competitions
  - Dataset

Process

- Feature preparation
  - Modeling
    - Optimization
      - Validation


Modeling
Neural Networks
Logistic Regression
SVM
Naive Bayes classifier
Decision trees
Optimization
Hyperparameters
Bagging
Ensemble
Boosting
Regularization
Random forest
Machine
Learning
Hold out method
Crossvalidation
Confusion matrix
Validation
ROC curve / AUC
Feature preparation
Feature discretization
Feature scaling
Feature imputation
Feature selection
Feature extraction

Q&A