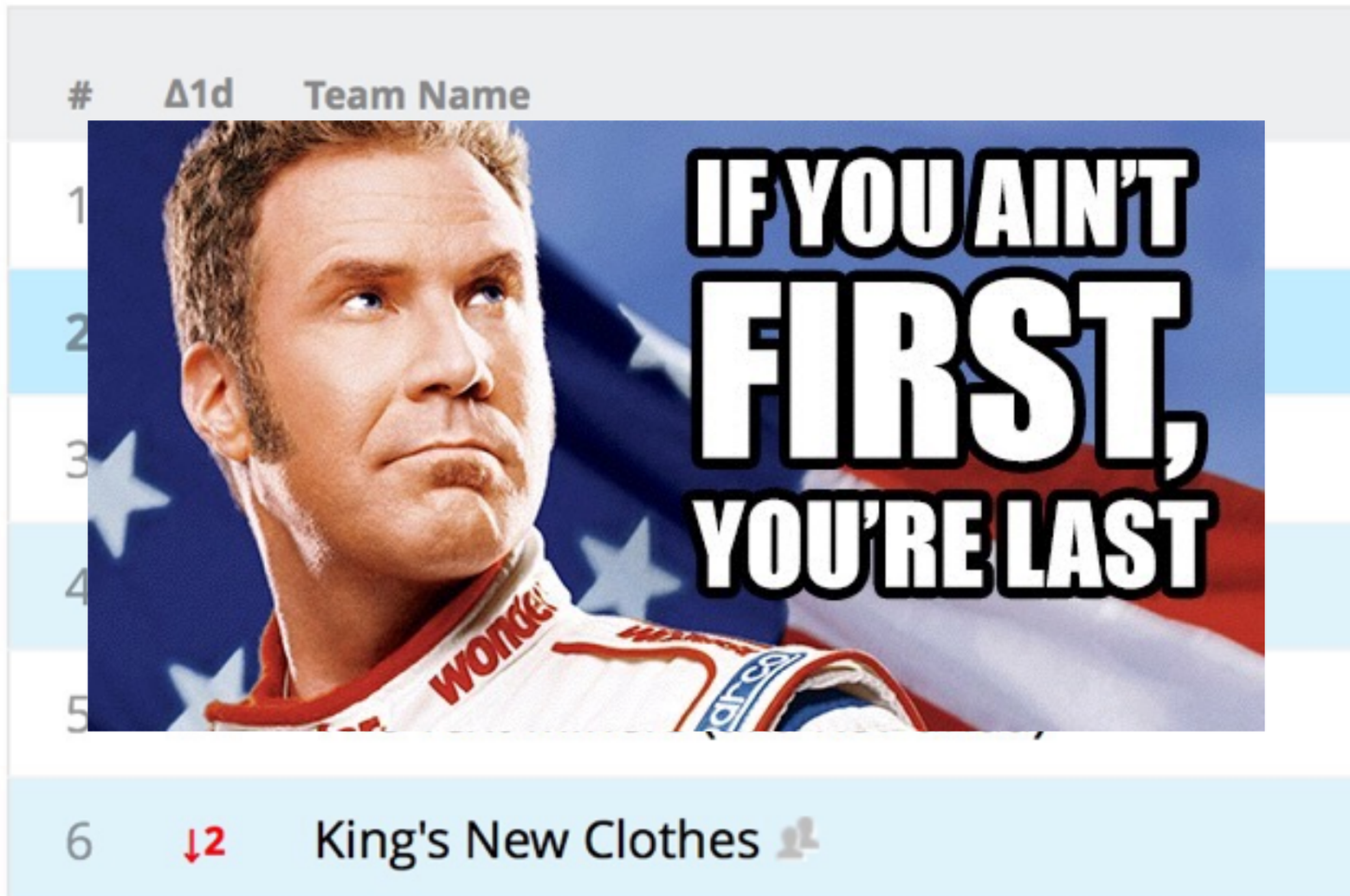
Kaggle text mining

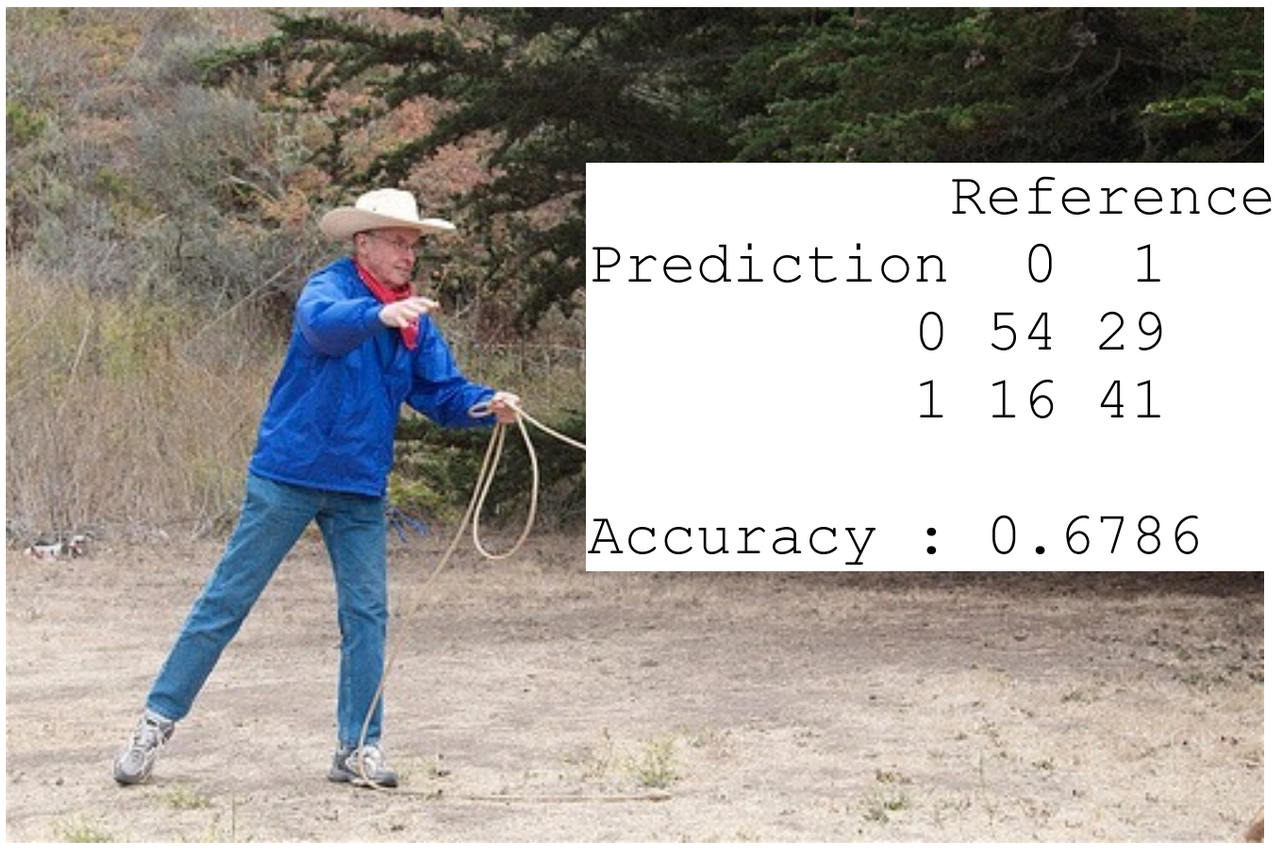
Winning solution: Anton Prokopyev

kaggle in Class

Motivation

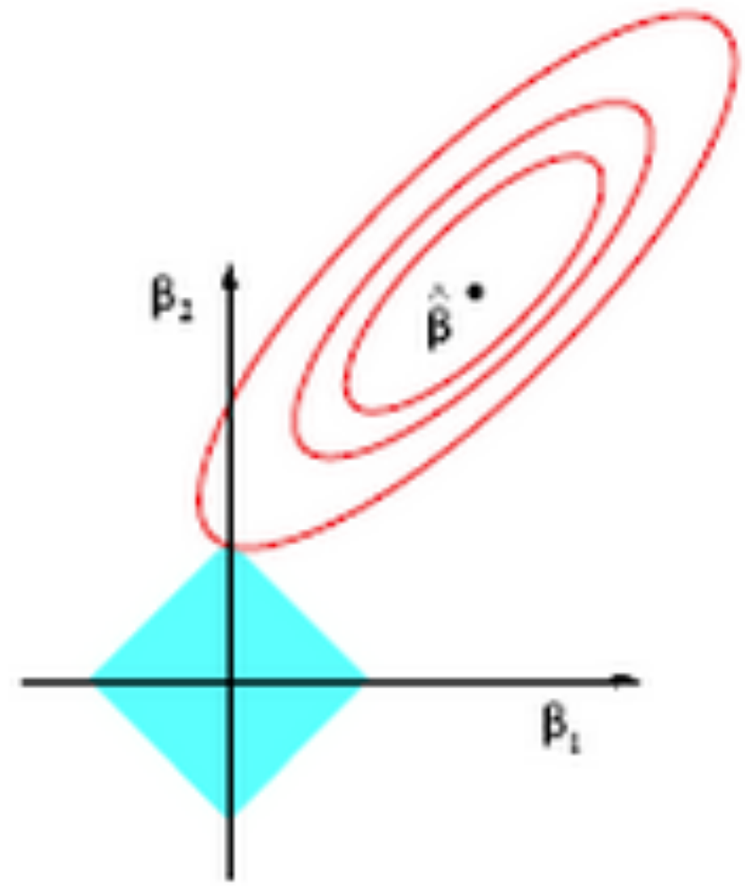

#
Δ1d
Team Name
IF YOU AIN'T
FIRST,
YOU'RE LAST
6
↓2
King's New Clothes

[1] LASSO Regression



```
          Reference
Prediction  0  1
         0 54 29
         1 16 41

Accuracy : 0.6786
```

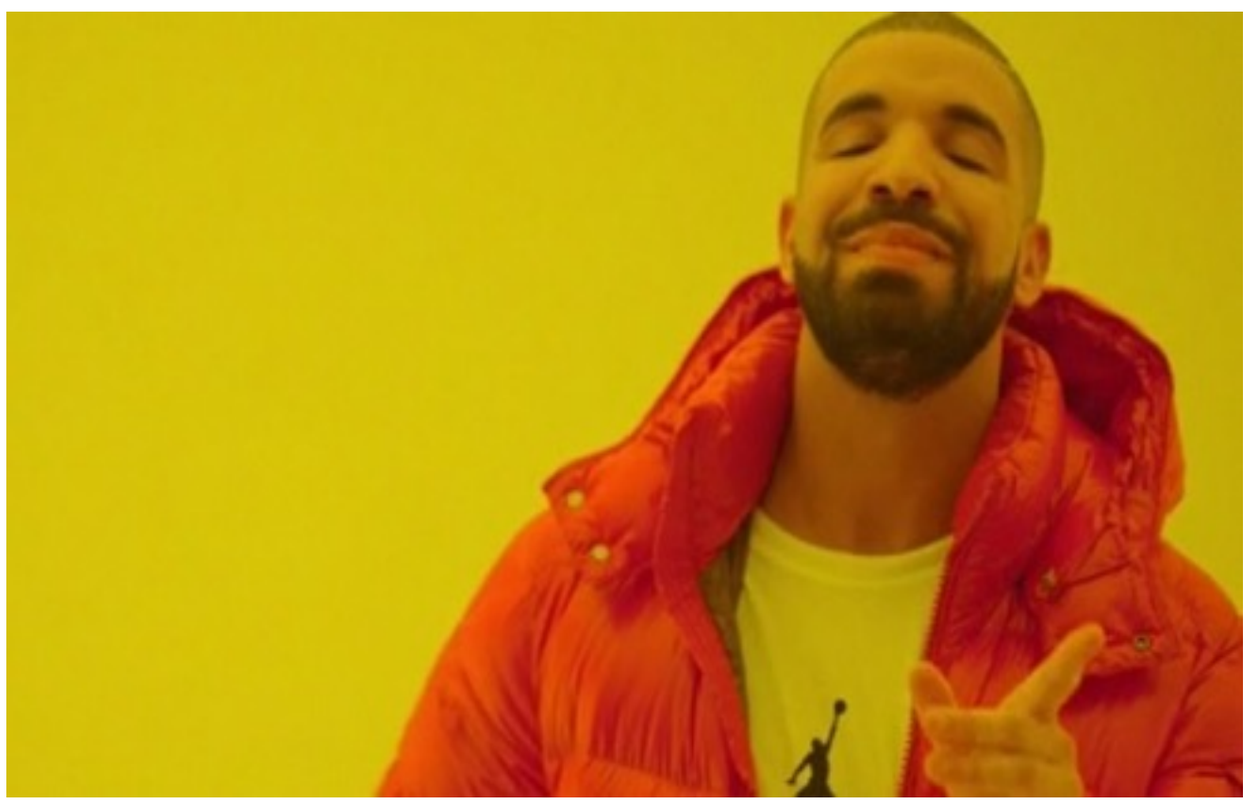


Least Absolute Shrinkage and Selection Operator

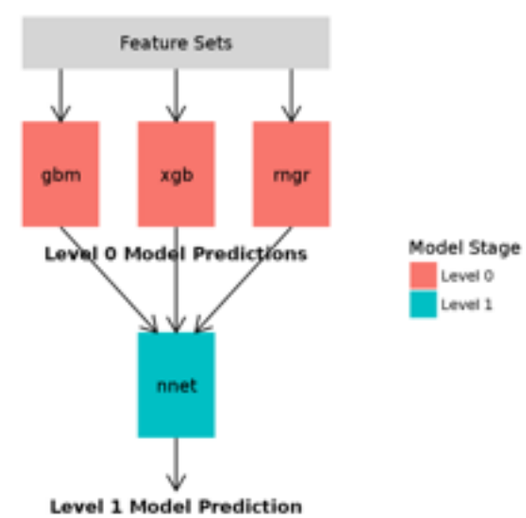
Competing on Kaggle with one model

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

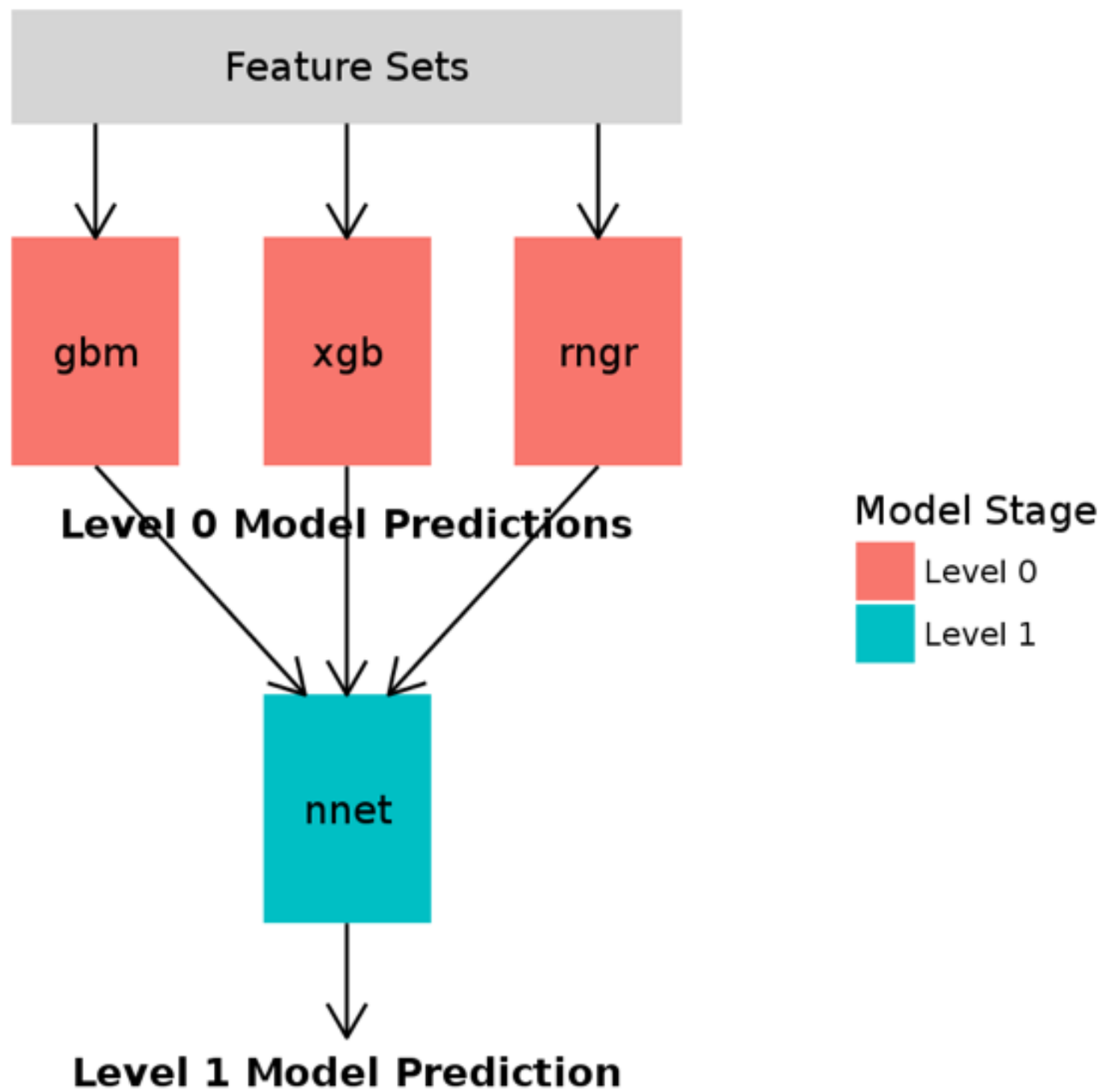
$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$



Competing on Kaggle with ensemble of models



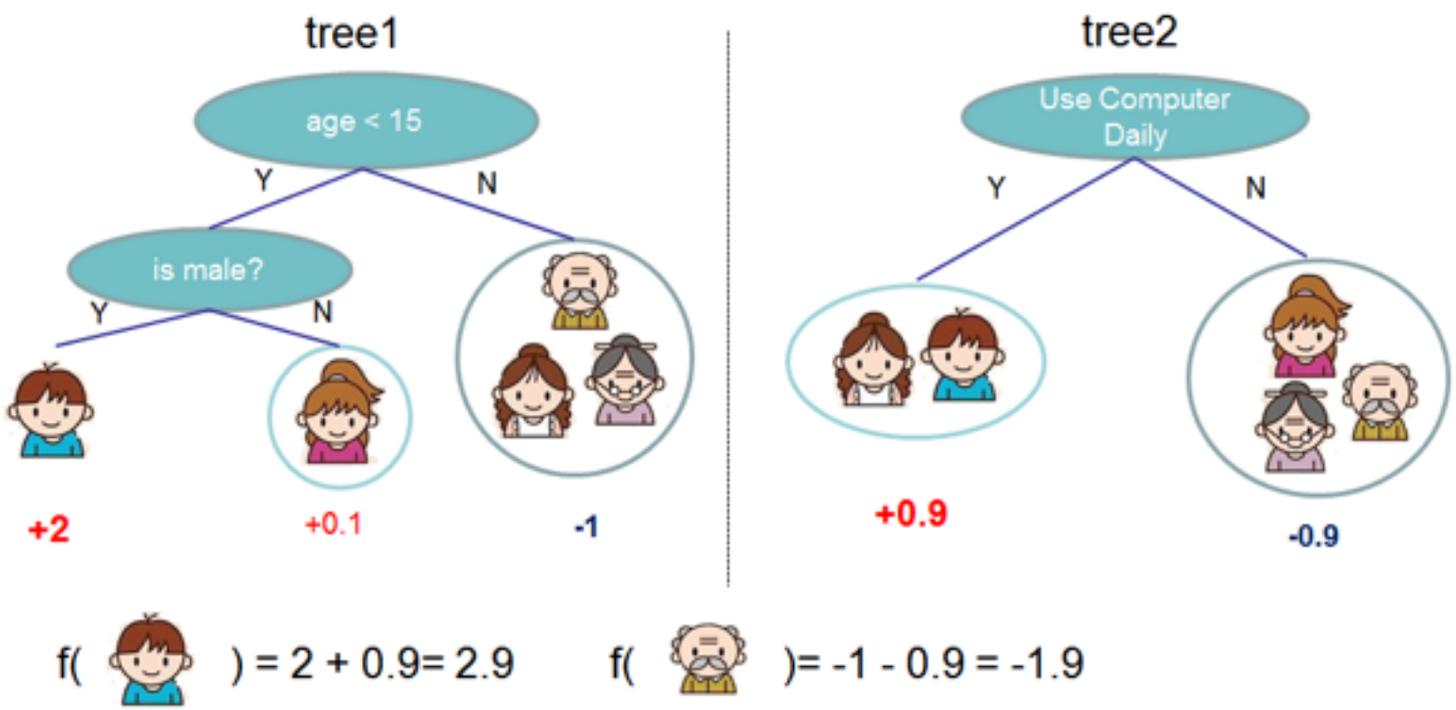
Ensemble modelling



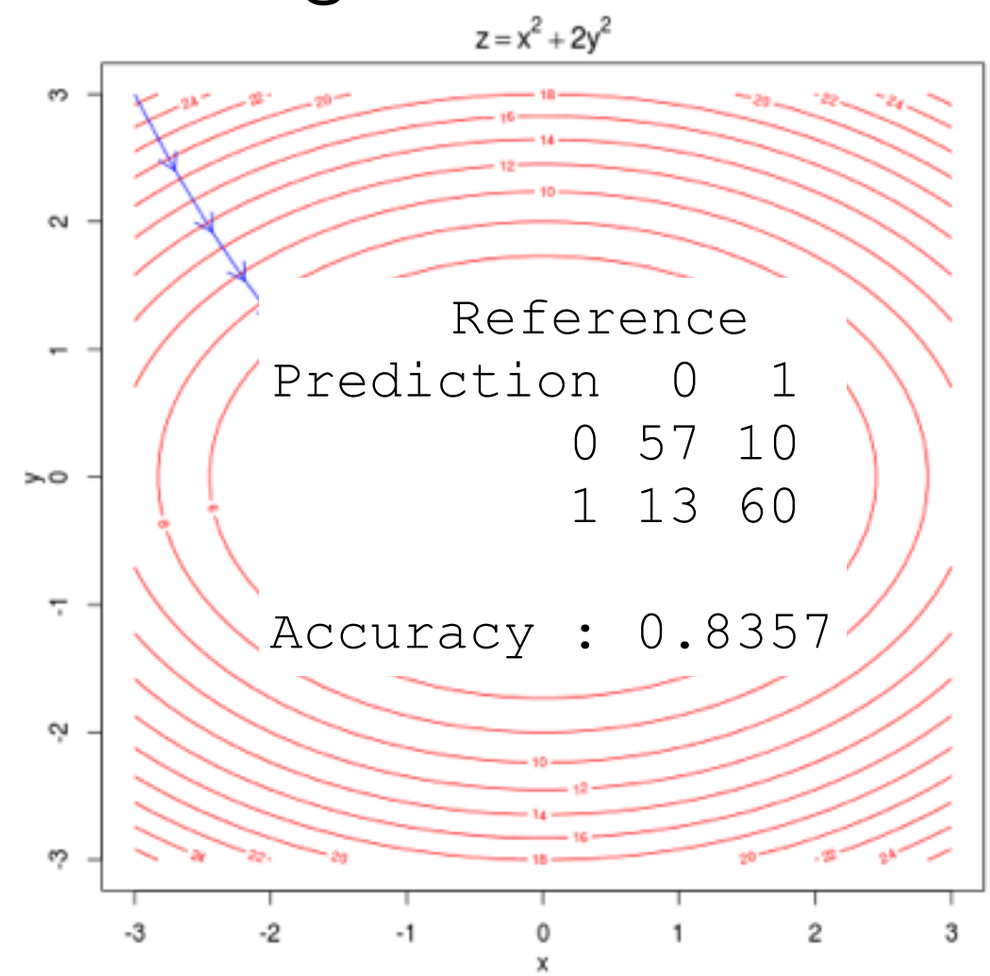
[2] XGBoost

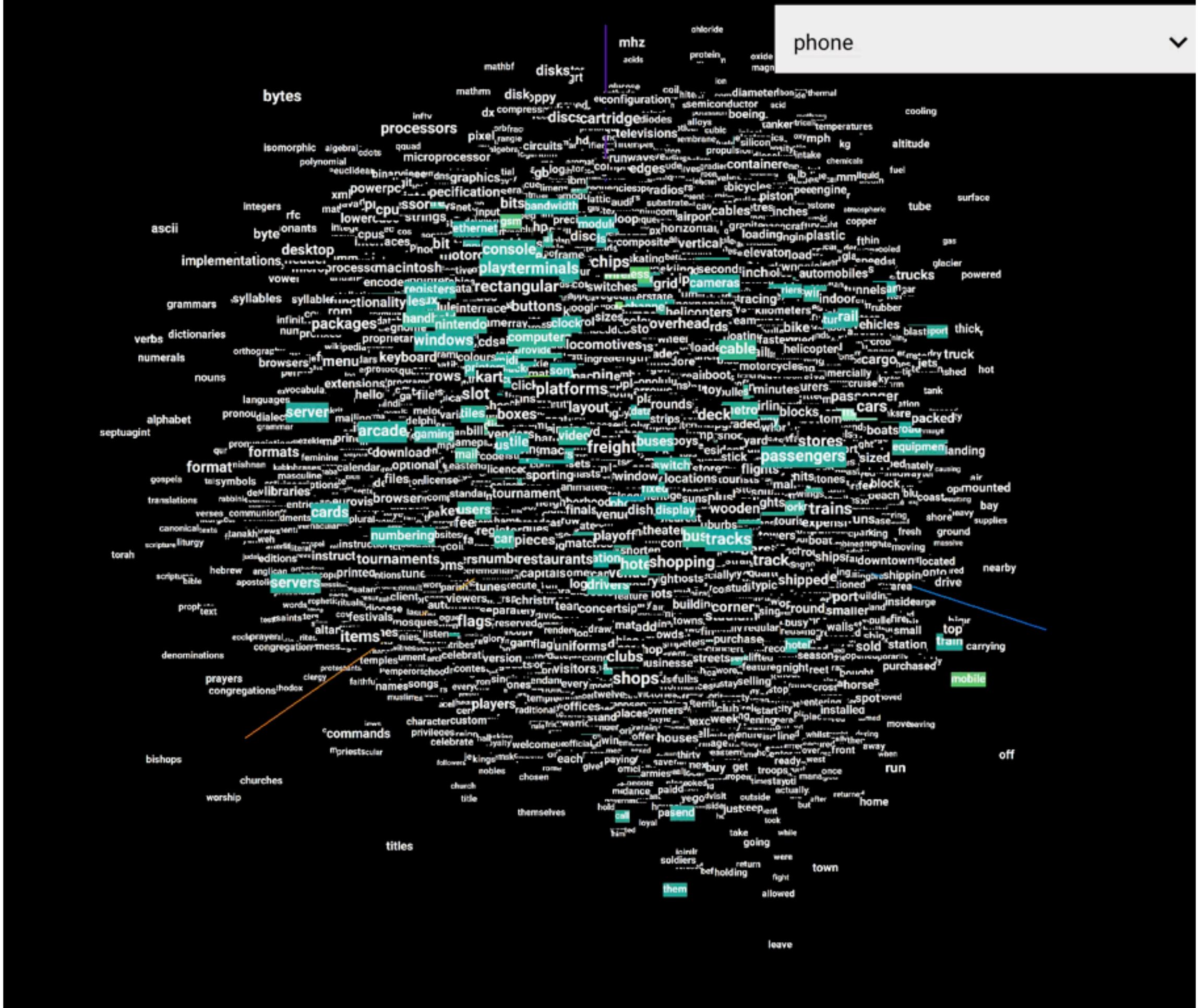
- Combines many weak trees

  Does the person like video games?

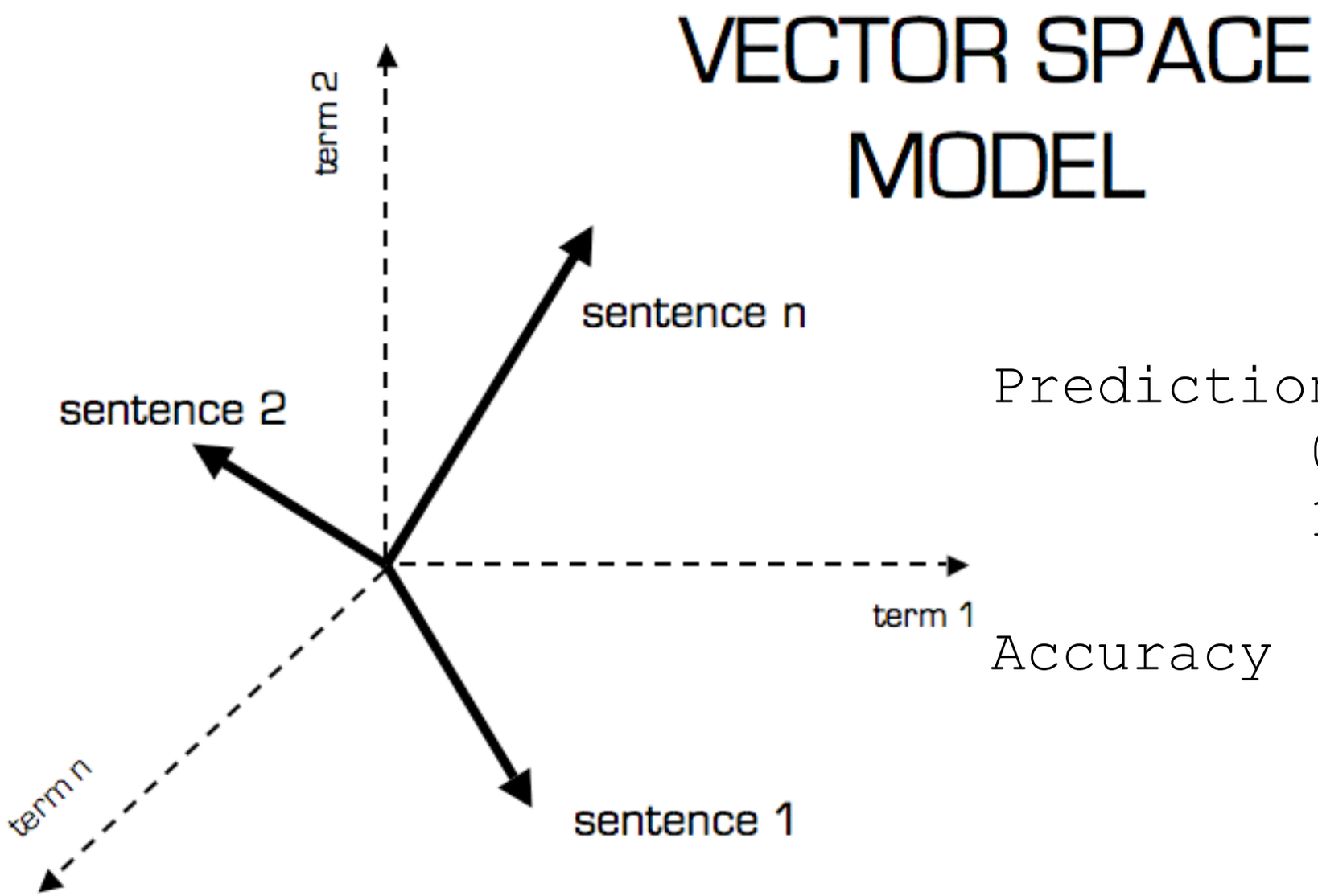


- Uses gradient descent




phone
mhz
bytes
processors
microprocessor
graphics
cartridge
televisions
containers
console
terminals
chips
rectangular
buttons
cameras
windows
nintendo
computer
locomotives
cable
keyboard
platforms
slot
server
arcade
gaming
freight
passengers
trains
tracks
formats
format
download
cards
numbering
tournaments
restaurants
shopping
servers
flags
items
clubs
shops
players
commands
bishops
churches
worship
titles
mobile
train
soldiers
them
town
leave
off
run
home

[3] text2vec




```
          Reference
Prediction  0  1
         0 60 20
         1 10 50

Accuracy : 0.7857
```

Combined performance

1st	Anton Prokopyev	0.80556

Select all of **Anton Prokopyev**'s submissions

Id	File	Description	Public	Private	Submission Date	Selected?
4660472	submit8_stack.csv	stack May 8	0.80556	0.81667	Tue, 09 May 2017 05:11:00	Yes
4657763	submit5.csv	text2vec only	0.80000	0.81667	Mon, 08 May 2017 18:35:19	Yes

Connect: linkedin.com/in/prokopyev

Thanks!

UC San Diego. POLI274: Text as Data, M. Roberts

High-dimensional word space: http://projector.tensorflow.org

XGBoost for Sentiment Analysis: https://github.com/wush978/FeatureHashing/blob/master/vignettes/SentimentAnalysis.Rmd

Text2Vec for Sentiment Analysis: https://www.r-bloggers.com/twitter-sentiment-analysis-with-machine-learning-in-r-using-doc2vec- approach