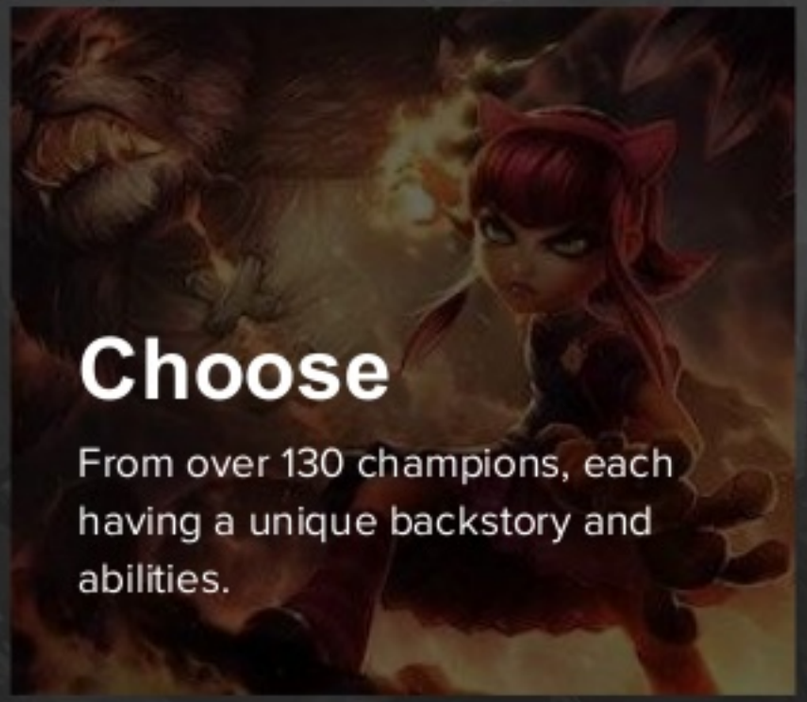
Combating Abusive Language in Chat with Apache Spark

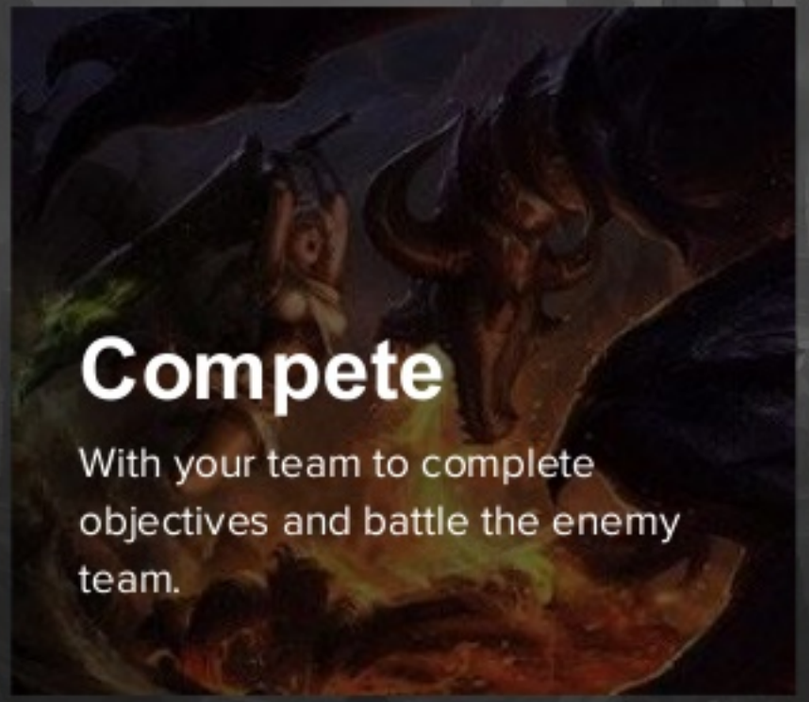
Wesley Kerr
Riot Games


LEAGUE OF LEGENDS



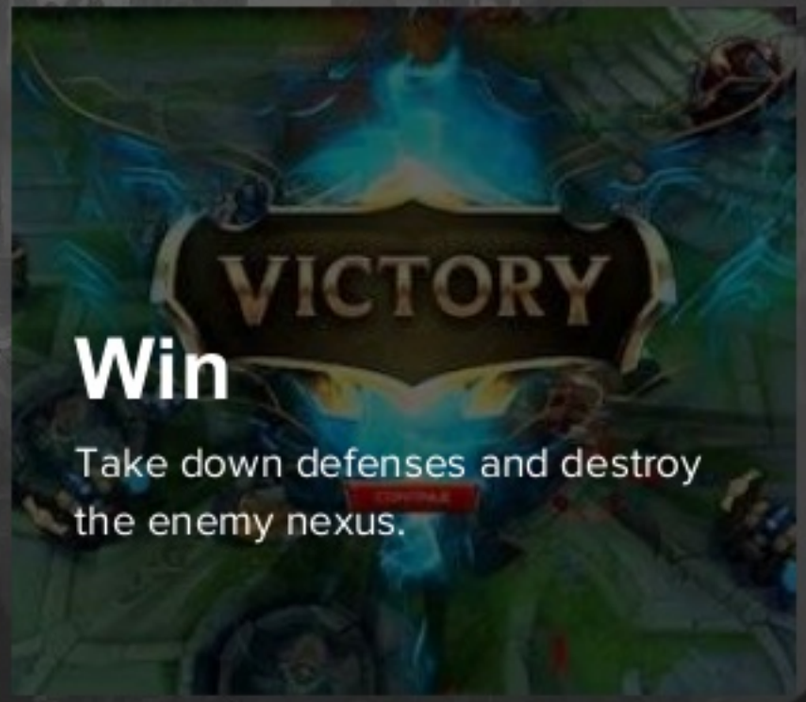
Choose

From over 130 champions, each having a unique backstory and abilities.



Compete

With your team to complete objectives and battle the enemy team.



Win

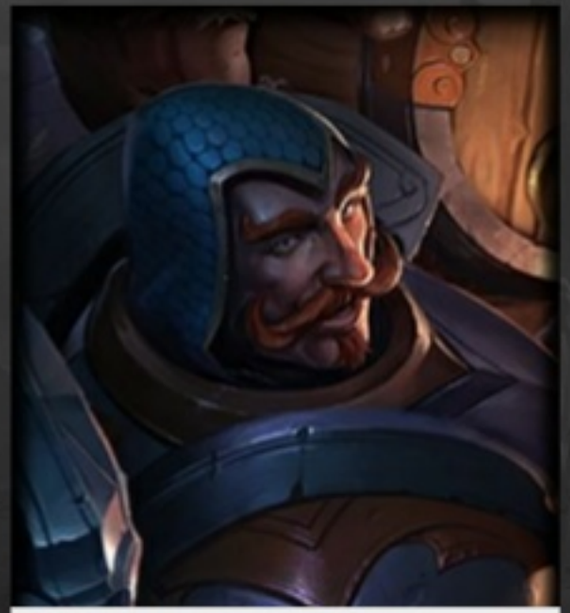
Take down defenses and destroy the enemy nexus.


Teamwork
94.5%
of players feel that
sportsmanship matters

In-Game Toxicity

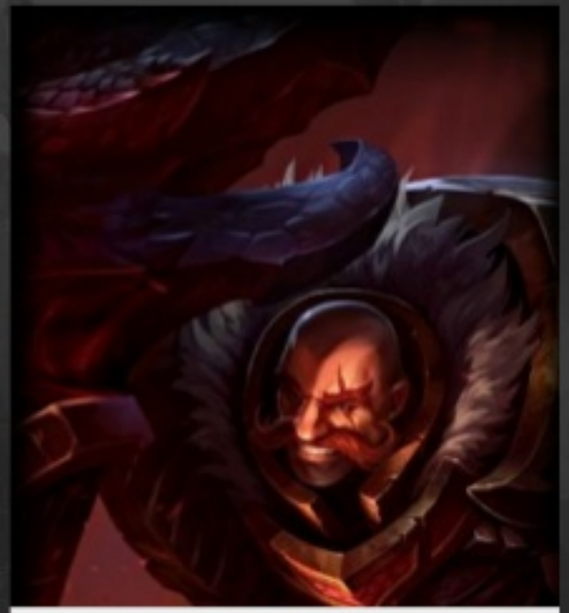
2%

of all games infected by serious toxicity



1%

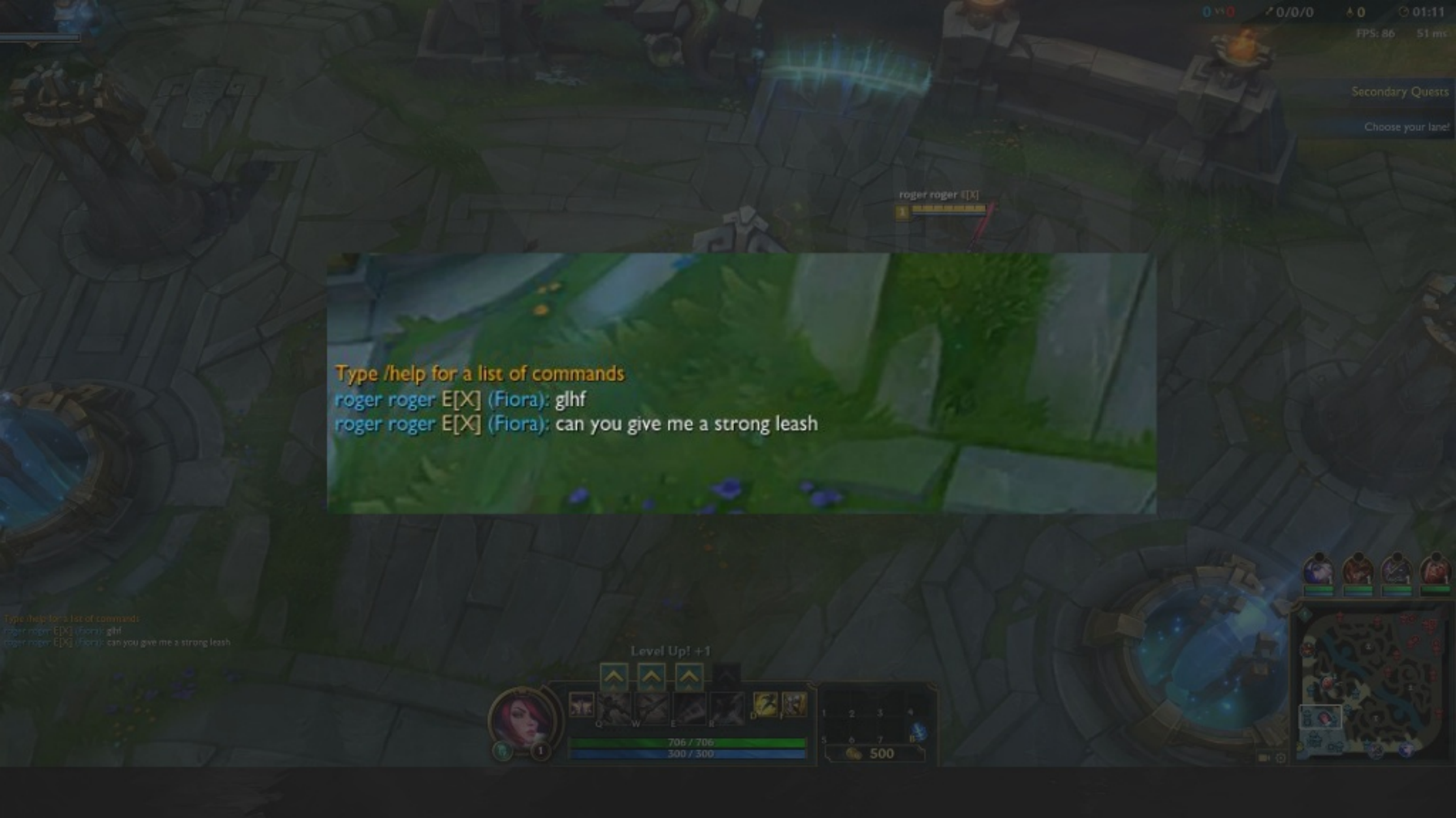
of all players are consistently unsportsmanlike



95%

of all serious toxicity comes from players who are otherwise sportsmanlike

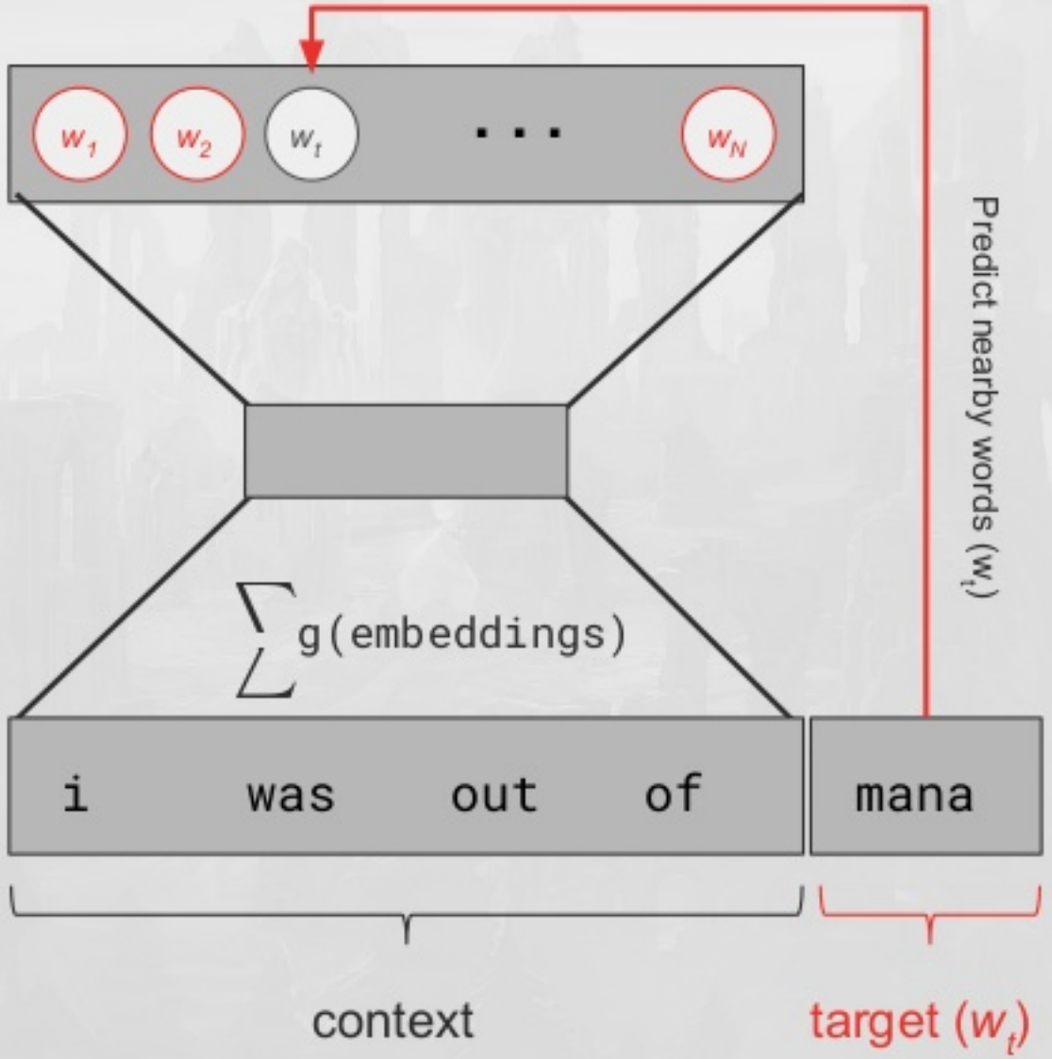

0 vs 0
0/0/0
0
01:11
FPS: 86
51 ms
Secondary Quests
Choose your lane!
roger roger E[X]
Type /help for a list of commands
roger roger E[X] (Fiora): ghf
roger roger E[X] (Fiora): can you give me a strong leash
Level Up! +1
706 / 706
300 / 300
500


0 vs 0
0/0/0
0
01:11
FPS: 86
51 ms
Secondary Quests
Choose your lane!
roger roger E[X]
Type /help for a list of commands
roger roger E[X] (Fiora): glhf
roger roger E[X] (Fiora): can you give me a strong leash
Level Up! +1
706 / 706
300 / 300
500

Exploration

Word2Vec

256 dimension embeddings
month of chat logs
each line of chat is a document
split on spaces and lower case



Exploration

gj	
nj	0.94
goodjob	0.83
gjj	0.82
gjh	0.81
\gj	0.79
gfj	0.78
gw	0.77
ty	0.77

qq	
t.t	0.74
q.q	0.74
t-t	0.73
q_q	0.72
:'(	0.72
t_t	0.72
;c	0.72
;(	0.71

noob	
nub	0.89
nooob	0.89
nob	0.83
n00b	0.80
nobb	0.79
noooob	0.79
noobb	0.78
nooooob	0.78

rekt	
rekted	0.94
wrecked	0.83
wrekt	0.82
owned	0.81
shrekt	0.79
clapped	0.78
bodied	0.77
roasted	0.77

Enhancements

- Desktop
- R/Python

Pros

Production

Extremely high precision

Cons

Limited data

Low recall

- AWS Clusters
- Apache Spark ML

- GPUs
- Tensorflow

Enhancements

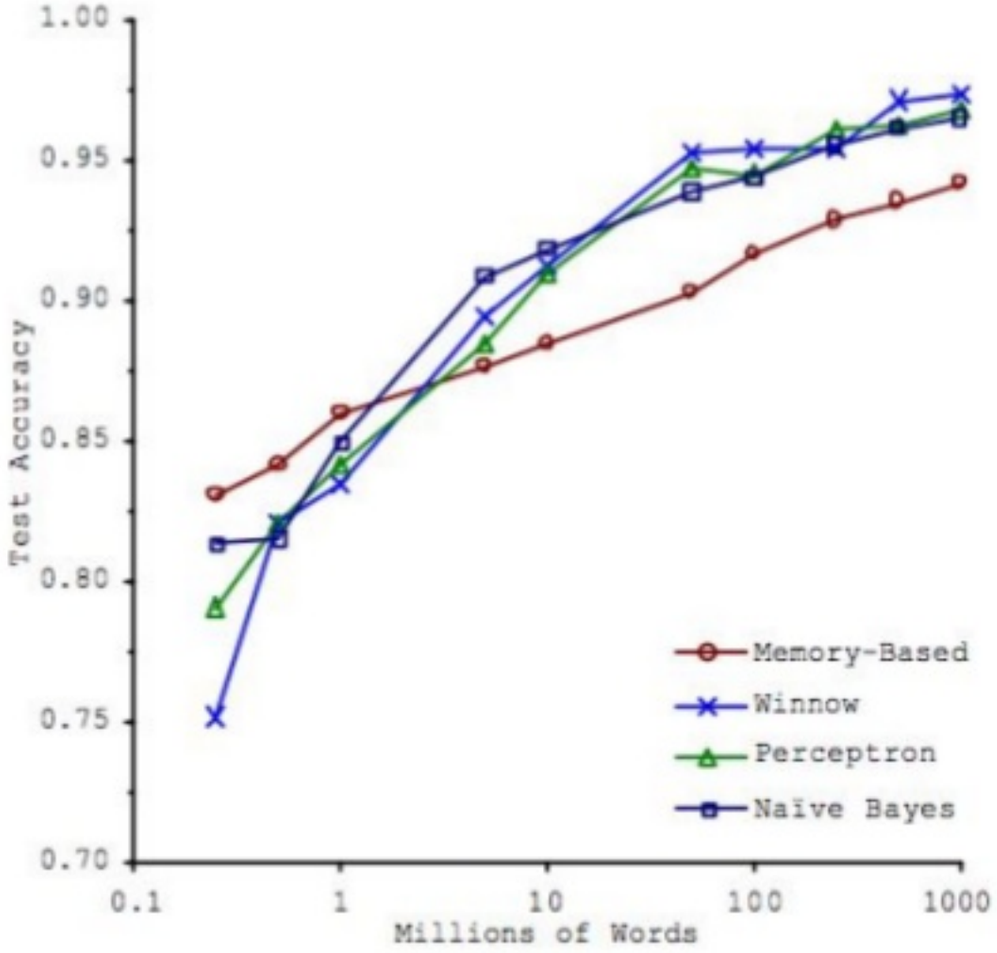
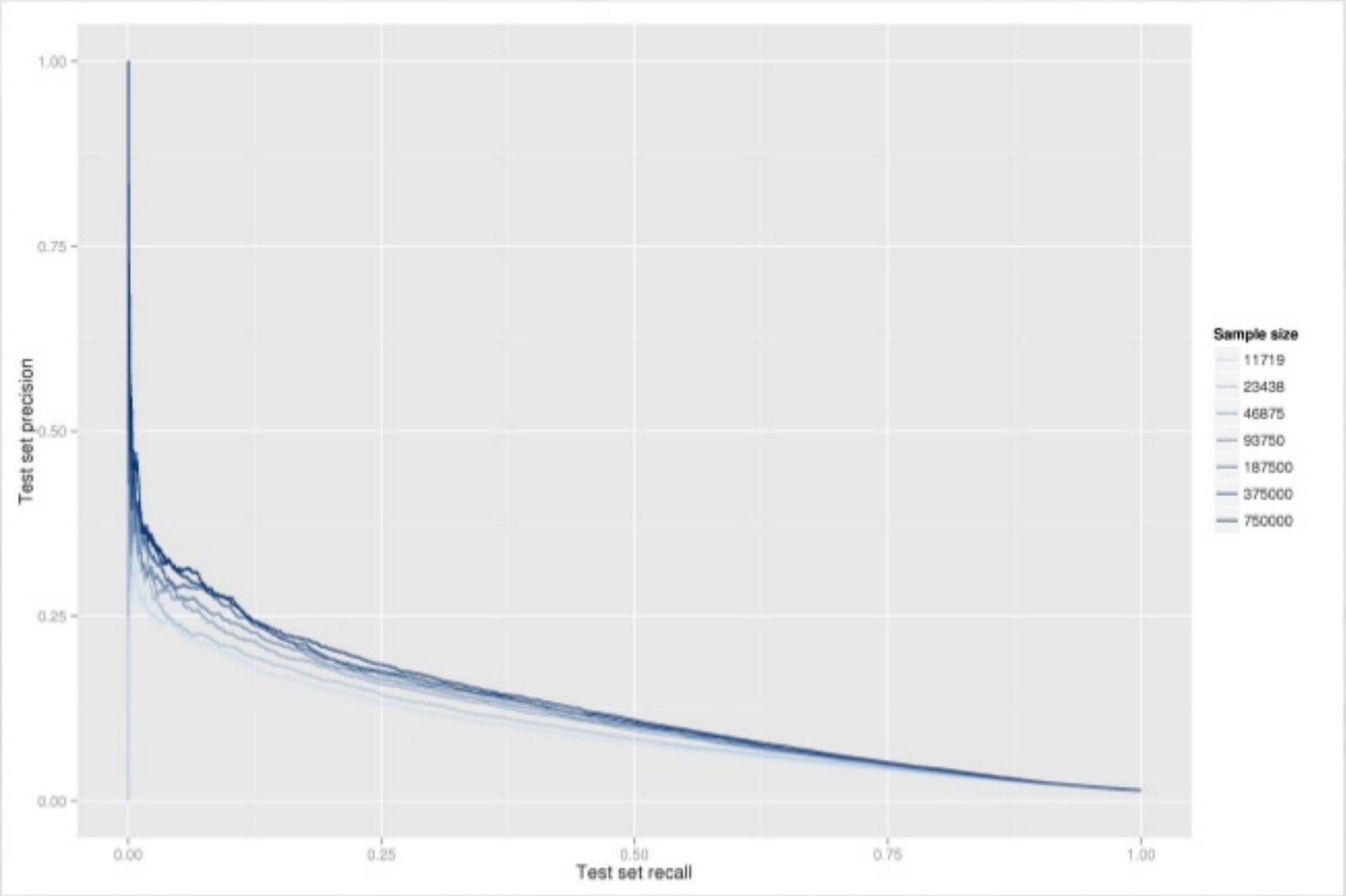


Figure 1. Learning Curves for Confusion Set Disambiguation

Banko and Brill [2001]. "Scaling to Very Very Large Corpora for Natural Language Disambiguation".

Enhancements



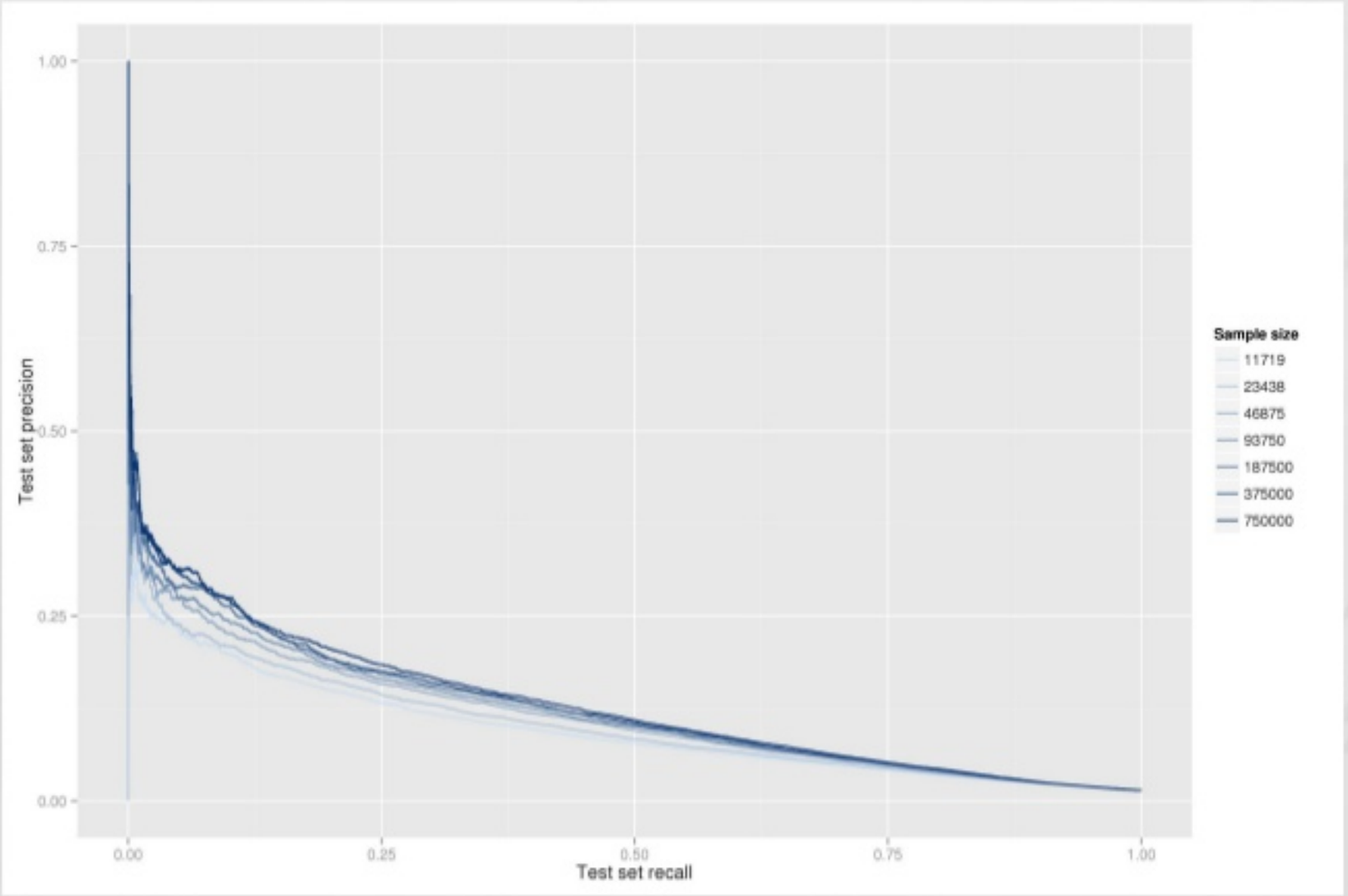
Enhancements

- AWS Clusters
- Apache Spark ML

Pros

Scale out model complexity

Scale out training data size



Enhancements

- AWS Clusters
- Apache Spark ML

Pros

Scale out model complexity

Scale out training data size

Spark Machine Learning Library

Transformers

n-grams
Tokenizer
Standard Scaler

Extractors

Word2Vec
TF-IDF
CountVectorizer

Algorithms

Logistic Regression
Random Forests
Gradient Boosted Trees

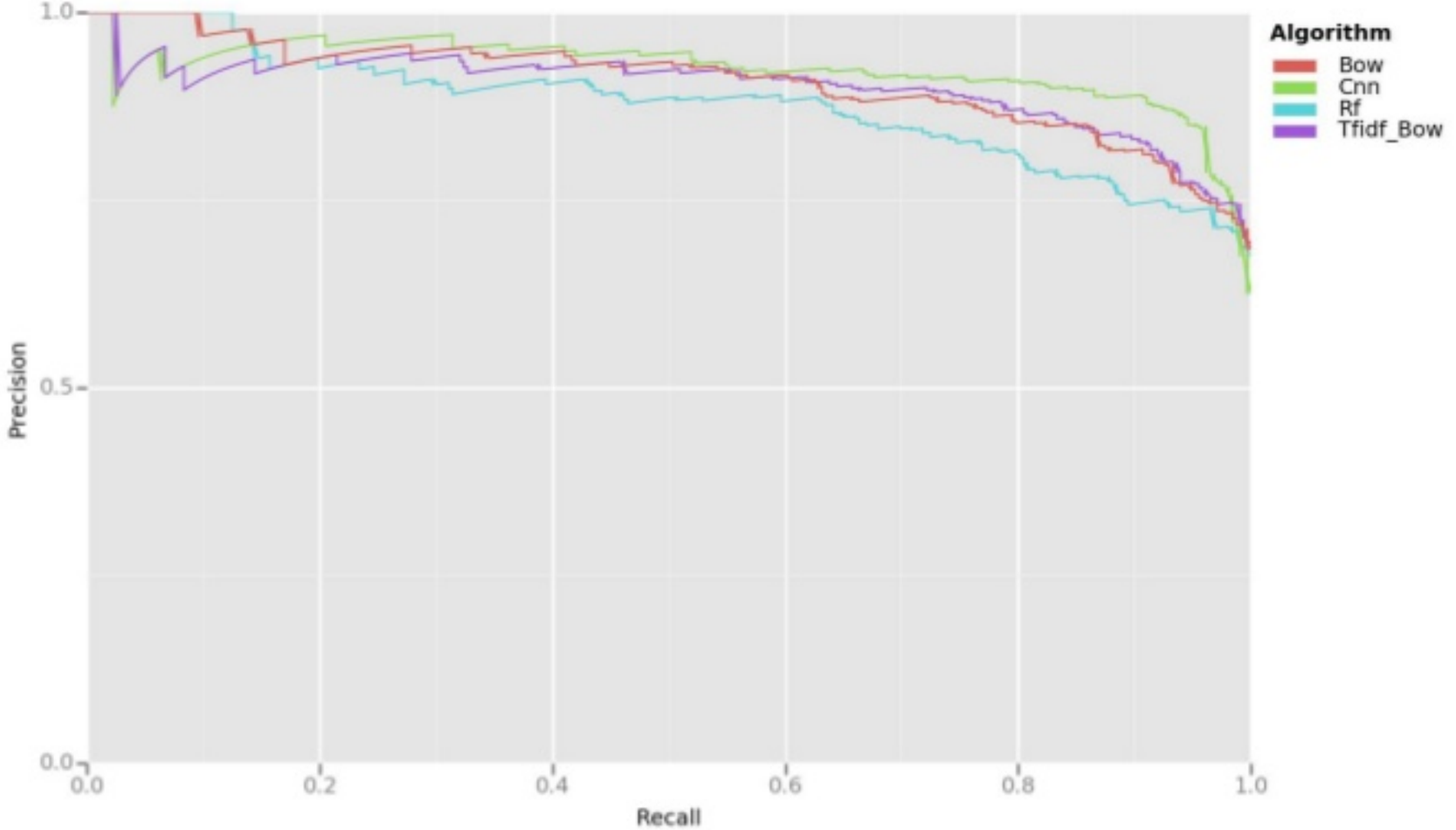
Enhancements

- AWS Clusters
- Apache Spark ML

Pros

Scale out model complexity

Scale out training data size



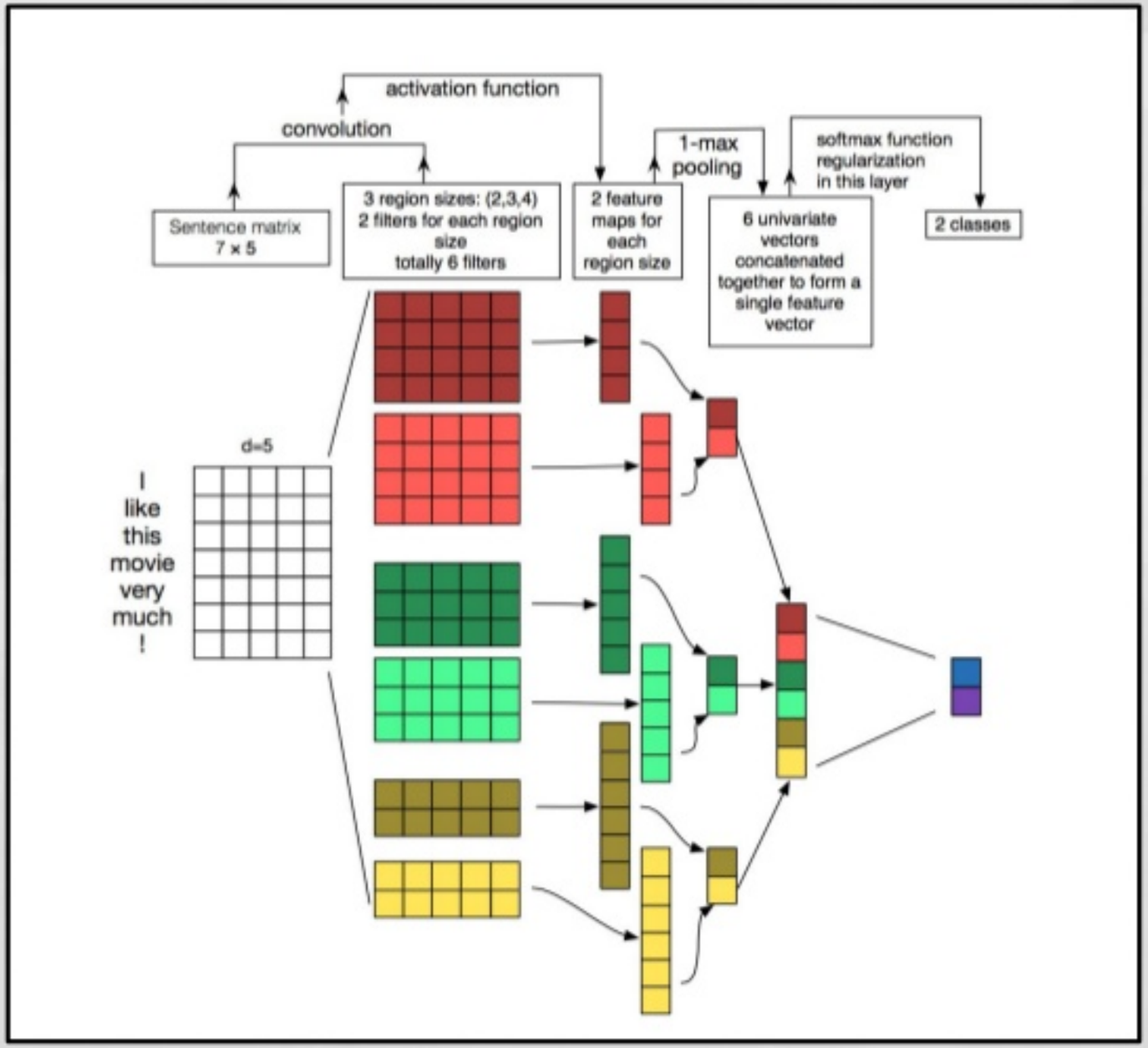
Future

- GPUs
- Tensorflow

Pros

Global Model

Easy tokenization



Source: Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification.

Future

- GPUs
- Tensorflow

Hyperparameters

Input

Vocab Size
Window Size
Stride

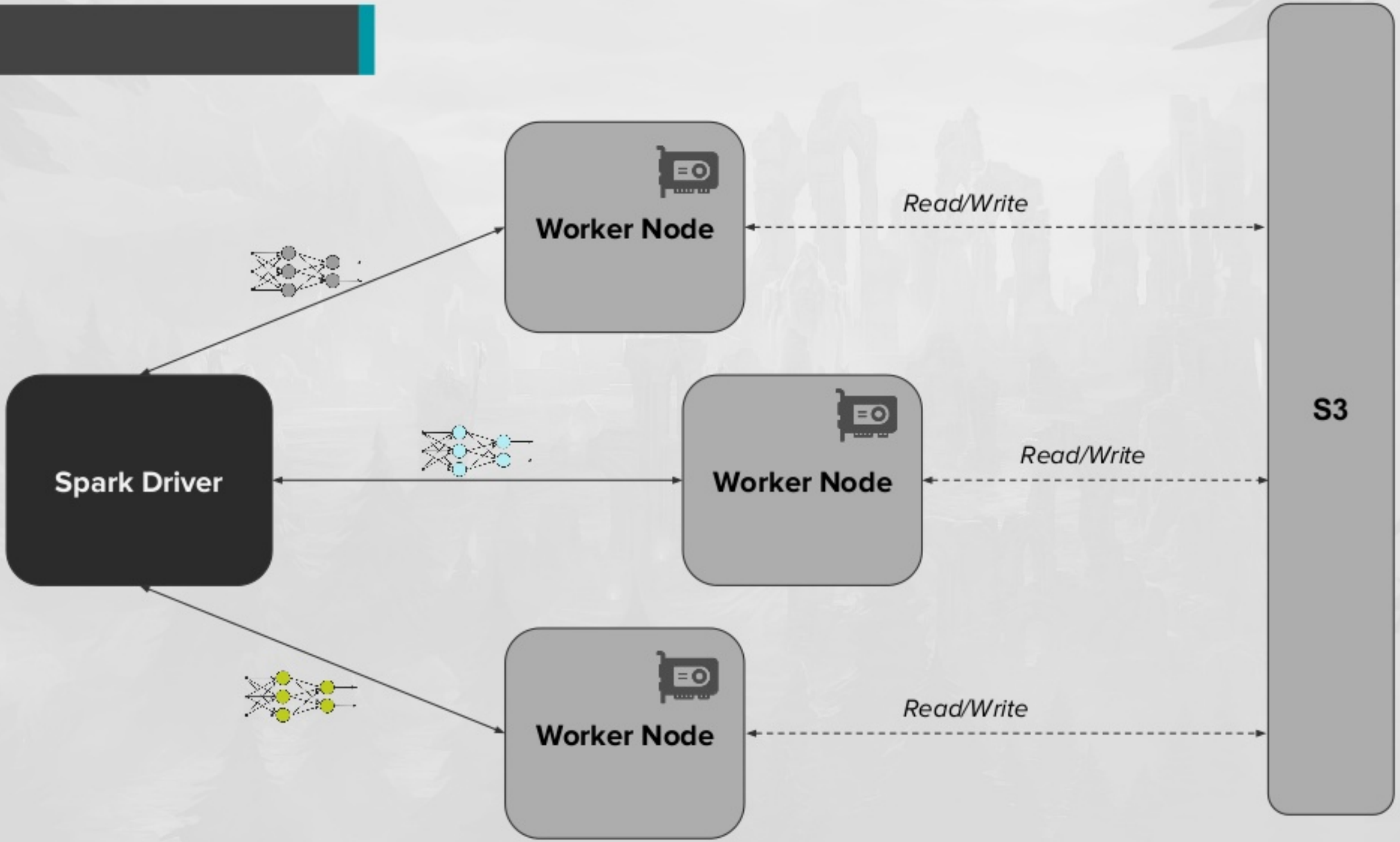
Convolutions

Activation Function
Window Size
# of Feature Maps
Depth

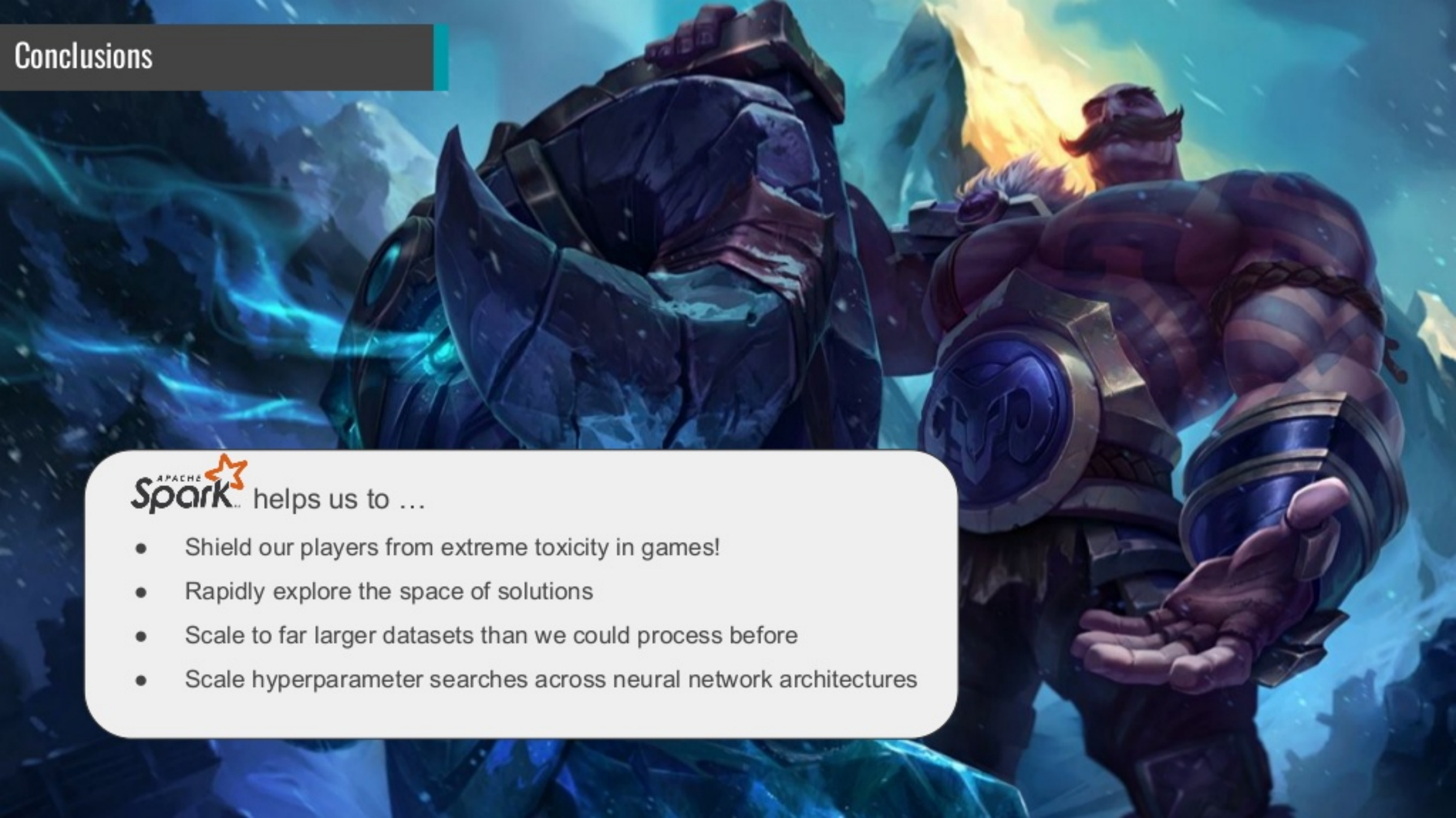
Connected

Activation Function
# of Hidden Nodes
Depth

Future


Worker Node
Read/Write
Spark Driver
Worker Node
Read/Write
S3
Worker Node
Read/Write

Conclusions



Apache Spark helps us to …

- Shield our players from extreme toxicity in games!
- Rapidly explore the space of solutions
- Scale to far larger datasets than we could process before
- Scale hyperparameter searches across neural network architectures

Thank you.

Wesley Kerr

WKERR @ RIOTGAMES.COM