

Jon Neff: Machine Learning at Scale

textfitXL

How will your comment fit in the community?



Motivation



- Predict up/down votes of comments on Reddit
- Goal: scale up data science project from ~500k to 1.6B comments

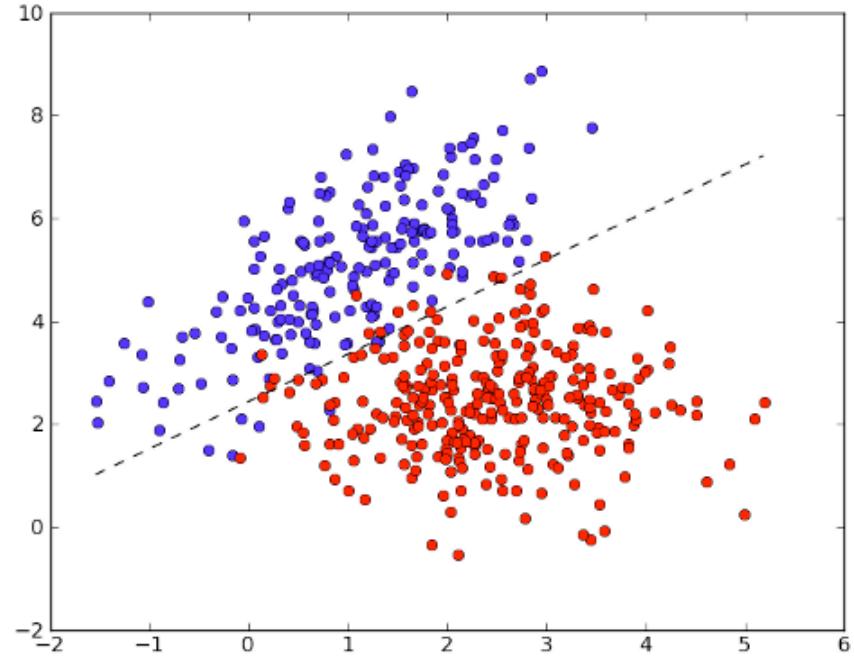
Demo

www.textfitxl.com

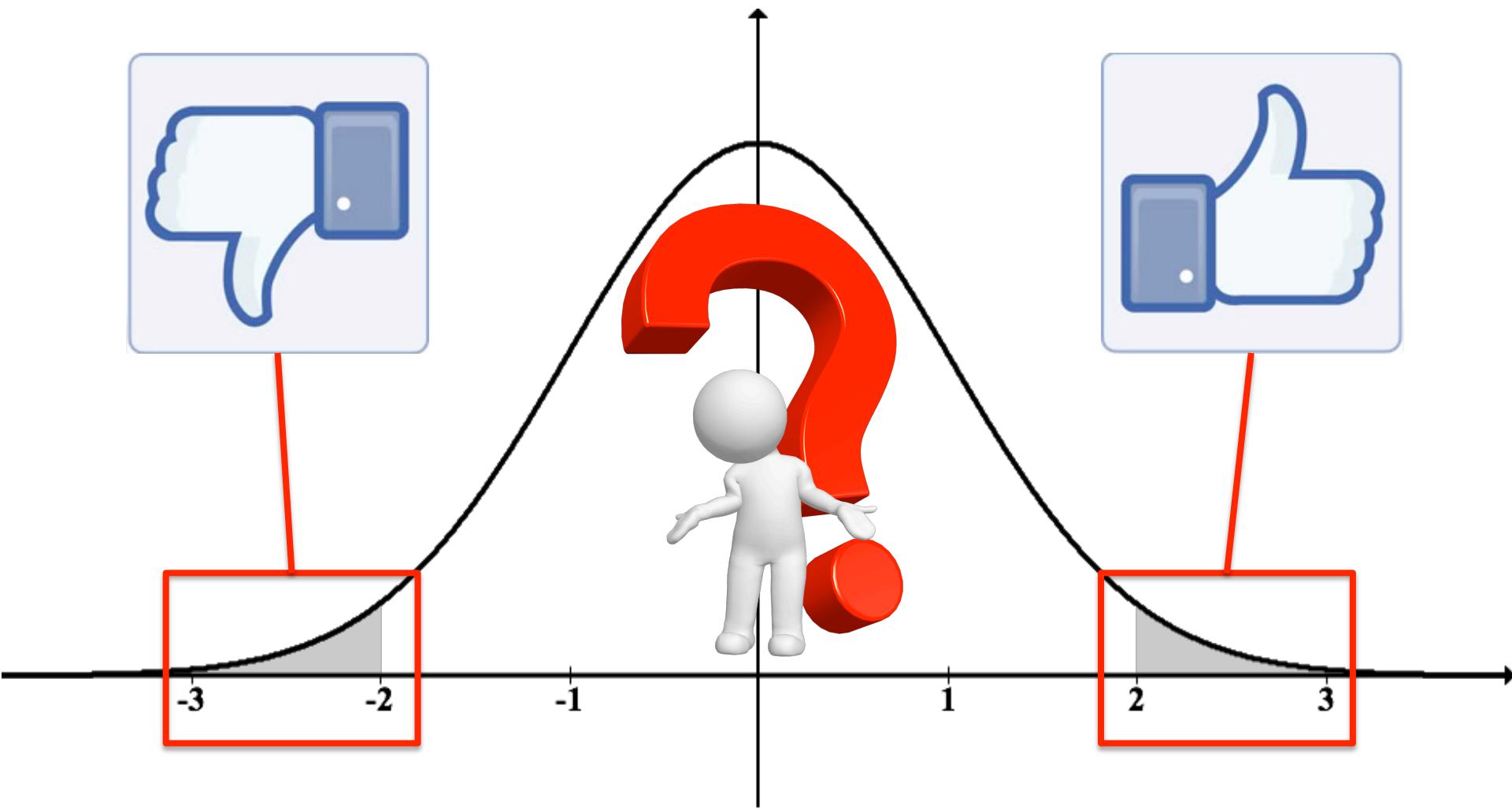
Thanks to Alyssa Fu for letting me re-use her code for the web app.

Algorithm, Features & Scaling

- Classification algorithm
 - Regularized logistic regression
- Features
 - Time since post
 - Comment length
 - Sentiment
 - Subreddit (categorical)
- Scalability
 - Spark: distributed, in-memory, DAG optimized
 - Sparse vector for OHE categorical: memory, time
 - Stochastic gradient descent: sample gradients

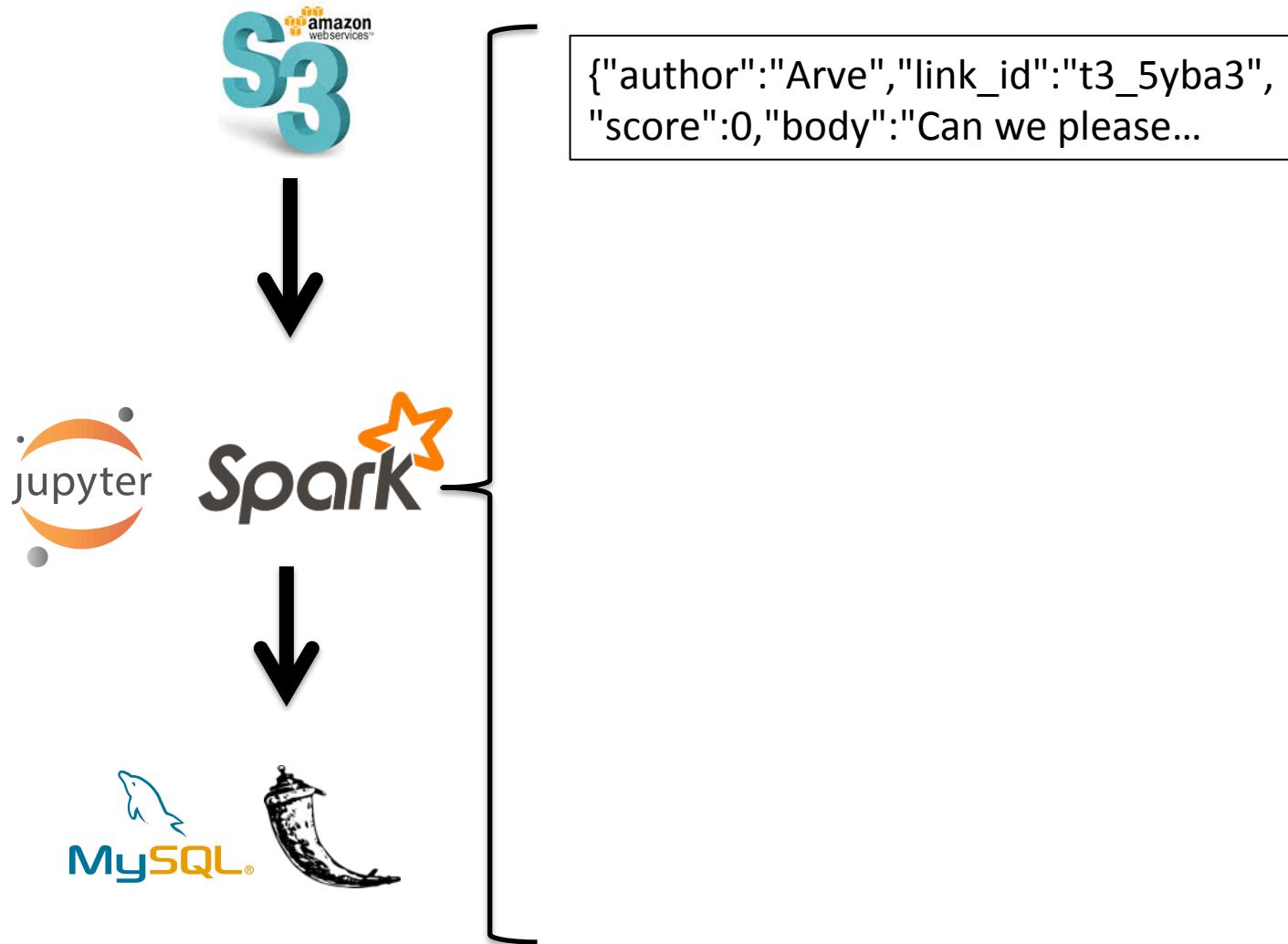


Filter to Top/Bottom 3%

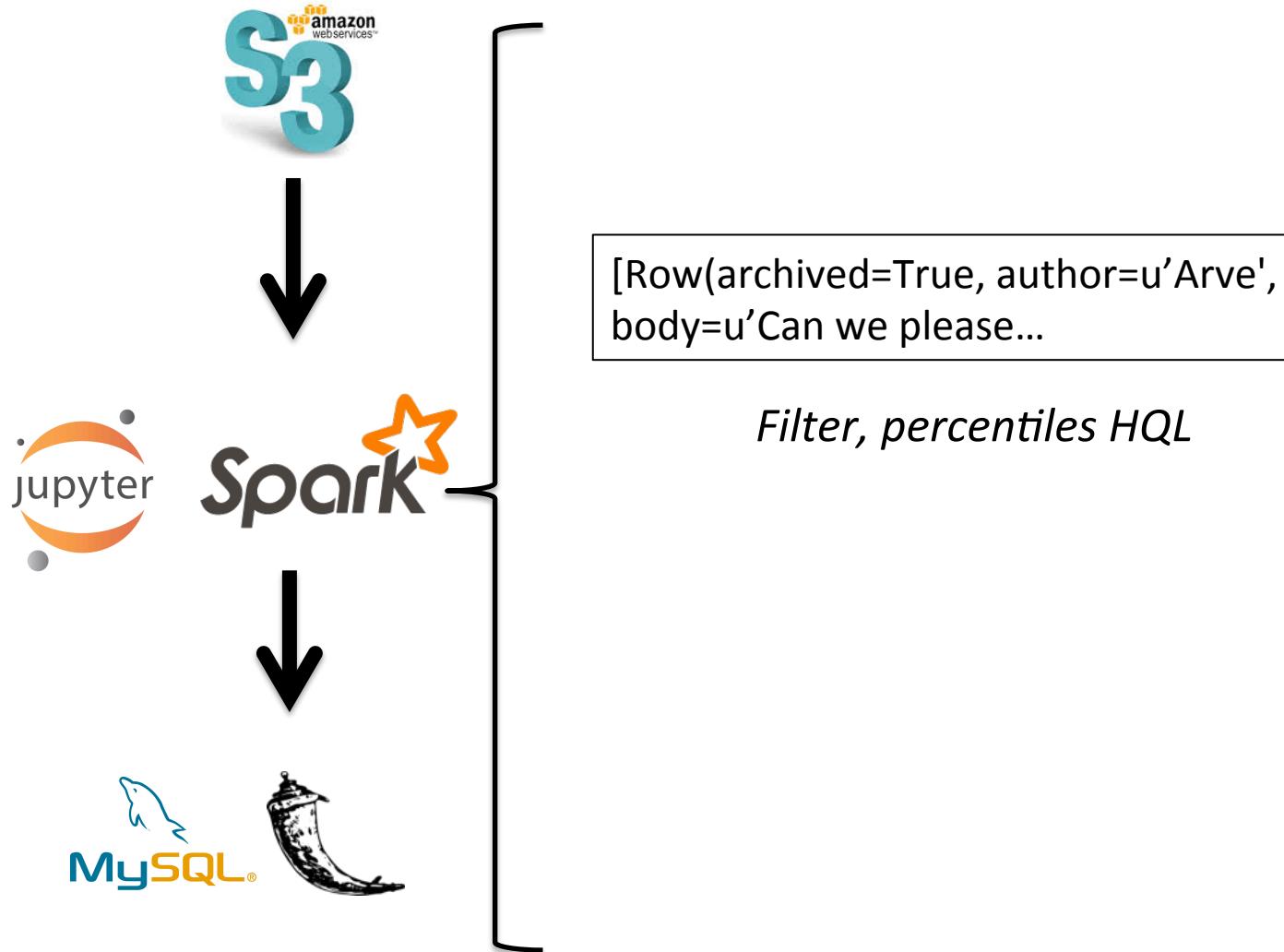


Data and Pipeline

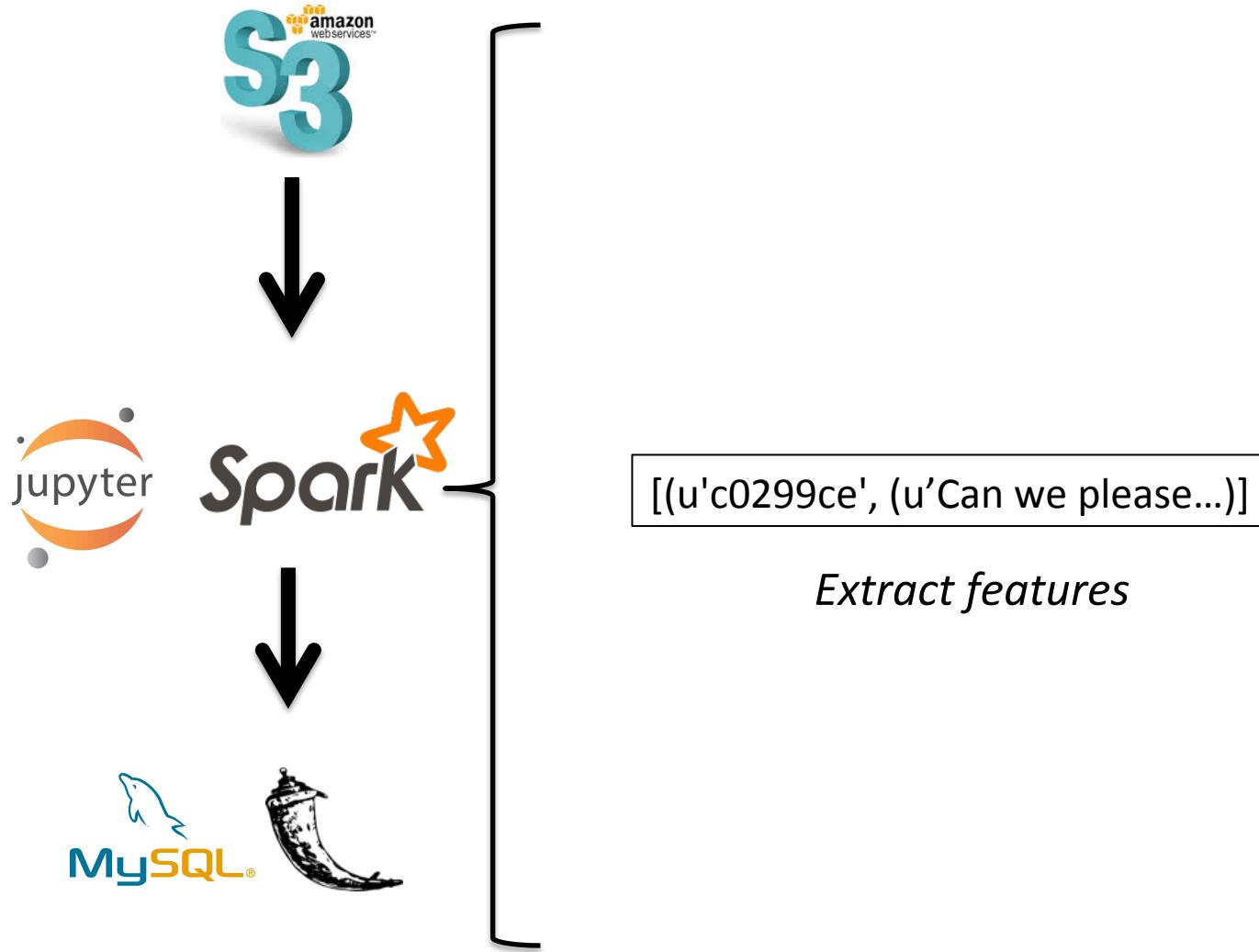
Source: Reddit comments in 0.9 TB JSON on S3.



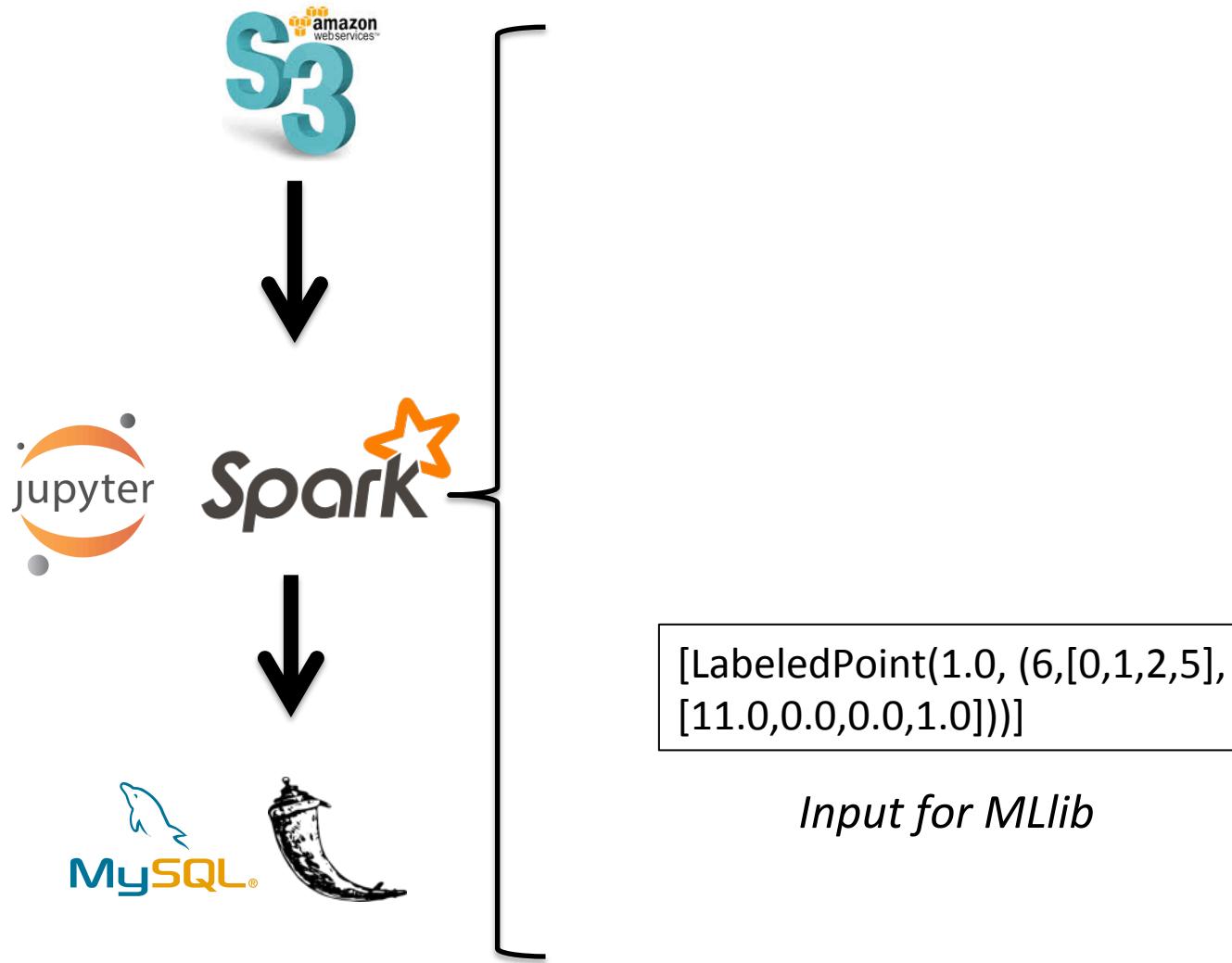
Data and Pipeline



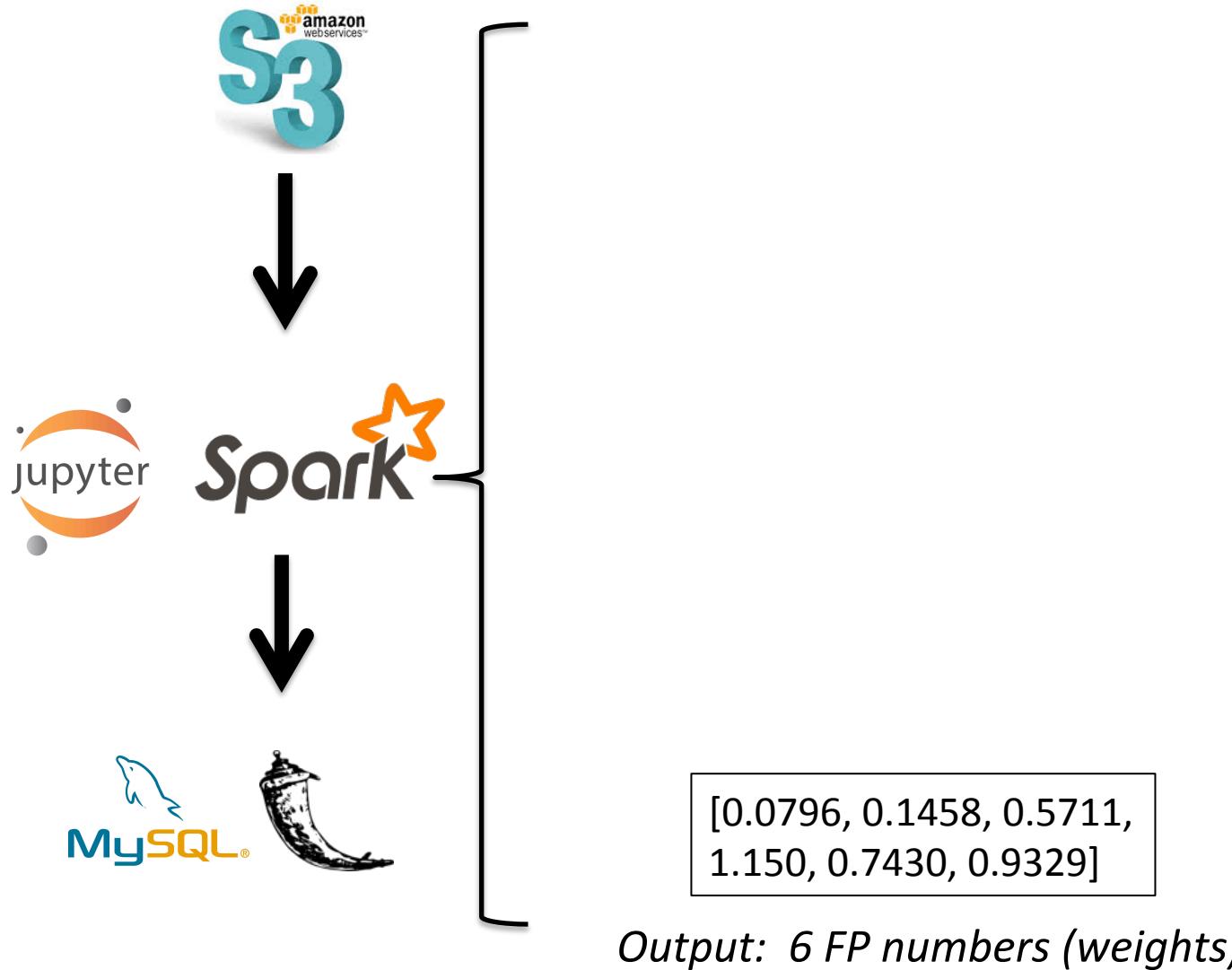
Data and Pipeline



Data and Pipeline



Data and Pipeline



Cluster and Cost

(1 year term)

Item	Number	Rate per node	Cost per month
t2.micro	1	\$6.57	\$6.57
m4.xlarge	9	\$126.29	\$1,136.61
1 TB EBS (mag)	9	\$50.00	\$450.00
Total			\$1,593.18

Accuracy and Run Time

- Accuracy: currently 52%
- Run time: 7.2 hours for entire 908 GB dataset.

Accuracy and Run Time

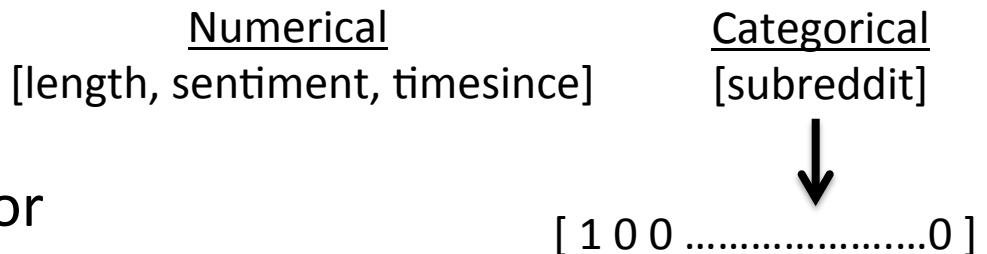
- Accuracy: currently 52%
- Run time: 7.2 hours for entire 908 GB dataset.



IS THAT THE BEST YOU CAN DO!!??

Challenges & Solutions

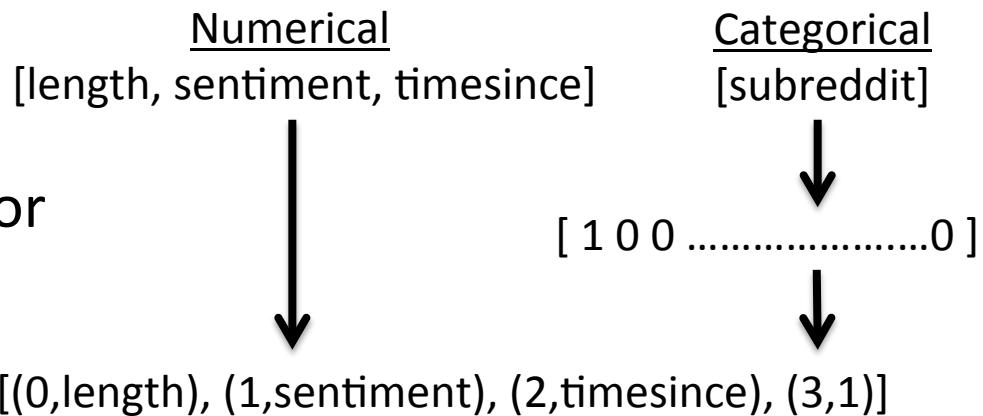
- Mixed categorical and numerical features in potentially large sparse vector



- *Sooooow* percentile, min time with map-reduce
- *Sooooow* brittle join

Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector



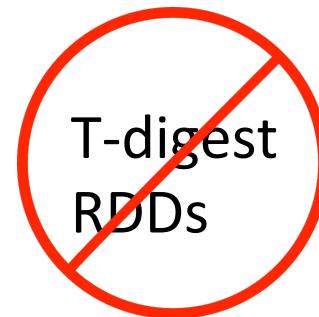
- Sooooow* percentile, min time with map-reduce
- Sooooow* brittle join

Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector
- *Slooooow* percentile, min time T-digest
RDDs
- *Slooooow* brittle join

Challenges & Solutions

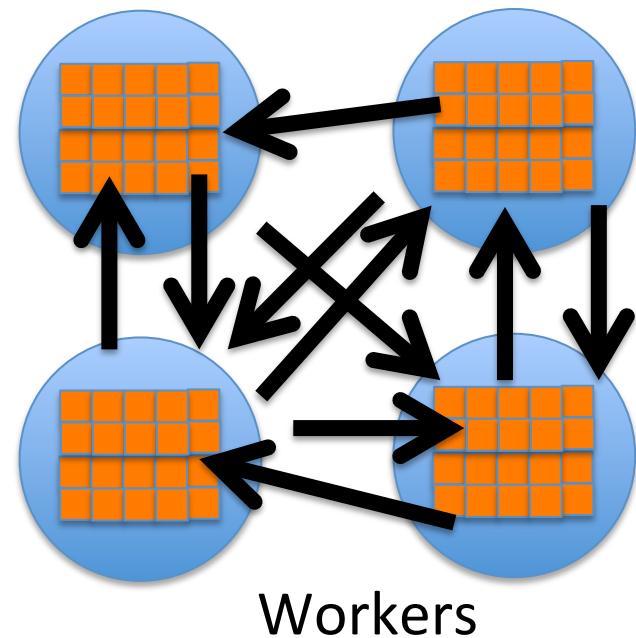
- Mixed categorical and numerical features in potentially large sparse vector
- *Slooooow* percentile, min time
- *Slooooow* brittle join



DataFrames
HiveQL

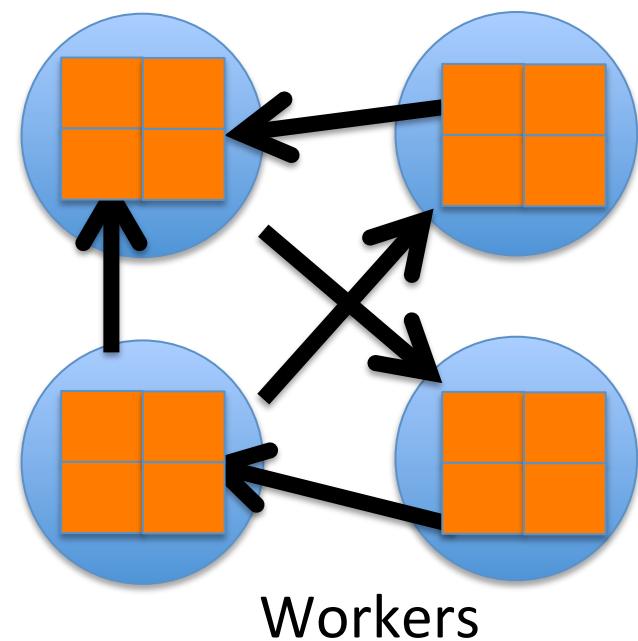
Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector
- *Slooooow* percentile, min time with map-reduce
- *Slooooow* brittle join



Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector
 - Reduce partitions after filter
 - Join with DataFrames
 - Size AWS instances for network performance
- *Slooooow* percentile, min time with map-reduce
- *Slooooow* brittle join



Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector
- *Slooooow* percentile, min time with map-reduce
- *Slooooow* brittle join

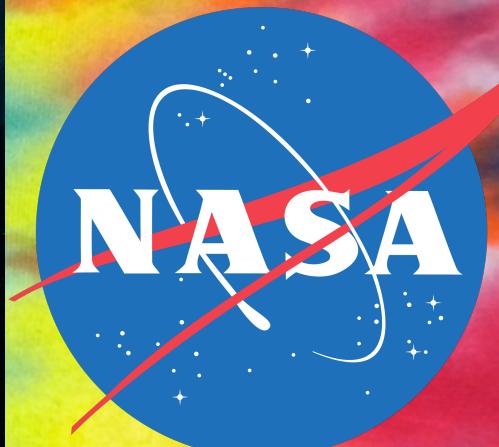
230% speed improvement!



About Me- Jon Neff



Ph.D.
MBA



Backup

Runtime

Activity	Runtime (minutes)
Read, filter subreddits	84
Create subreddit dict (percentiles)	138
Distinct, min (time), join, collect; create dict & broadcast	90
Extract features, train model	120
TOTAL RUNTIME	432