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

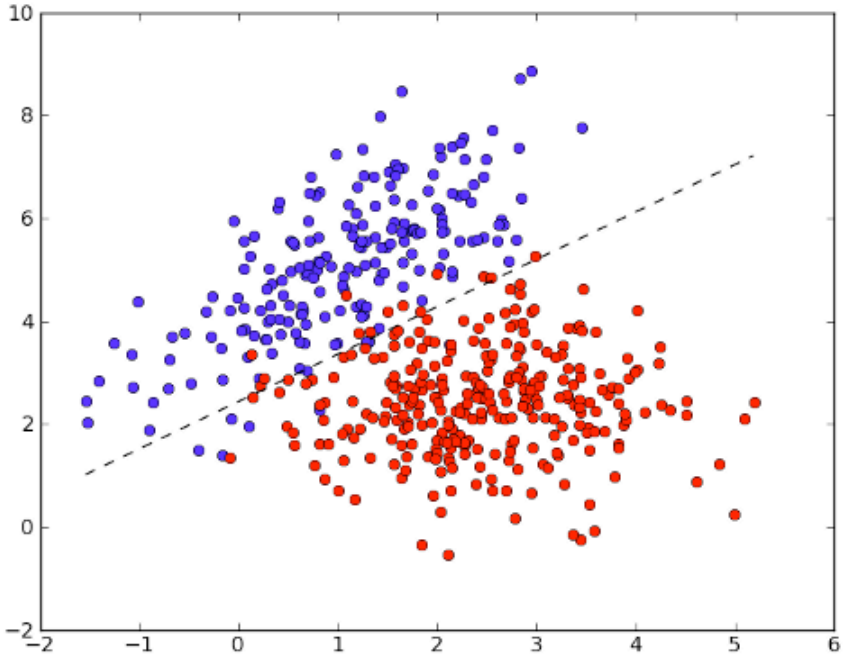
Demo

www.textfitxl.com

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

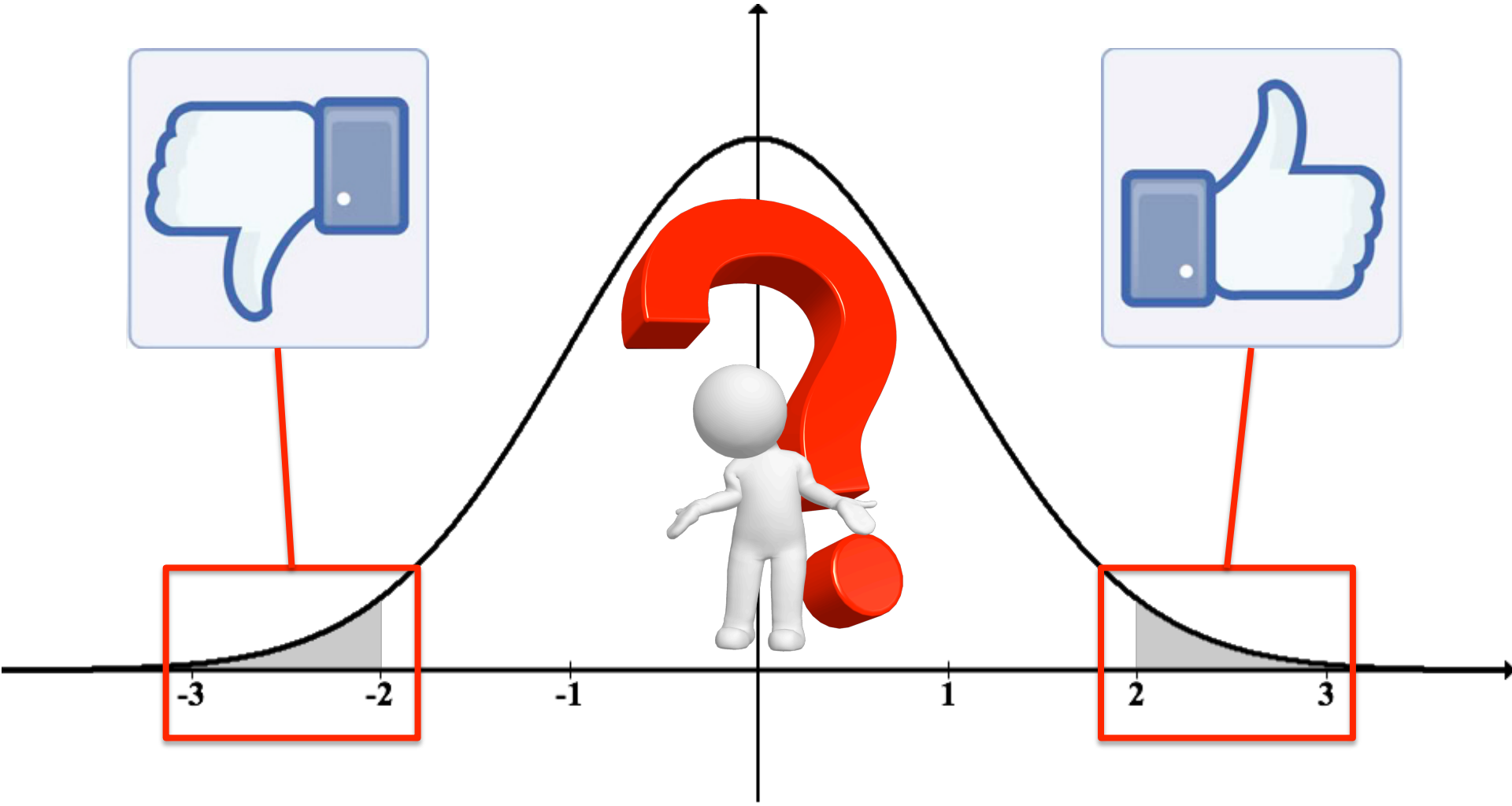
Algorithm, Features & Scaling

- Classification algorithm
  - Regularized logistic regression
- Features
  - Time since post
  - Comment length
  - Sentiment
  - Subreddit (categorical)



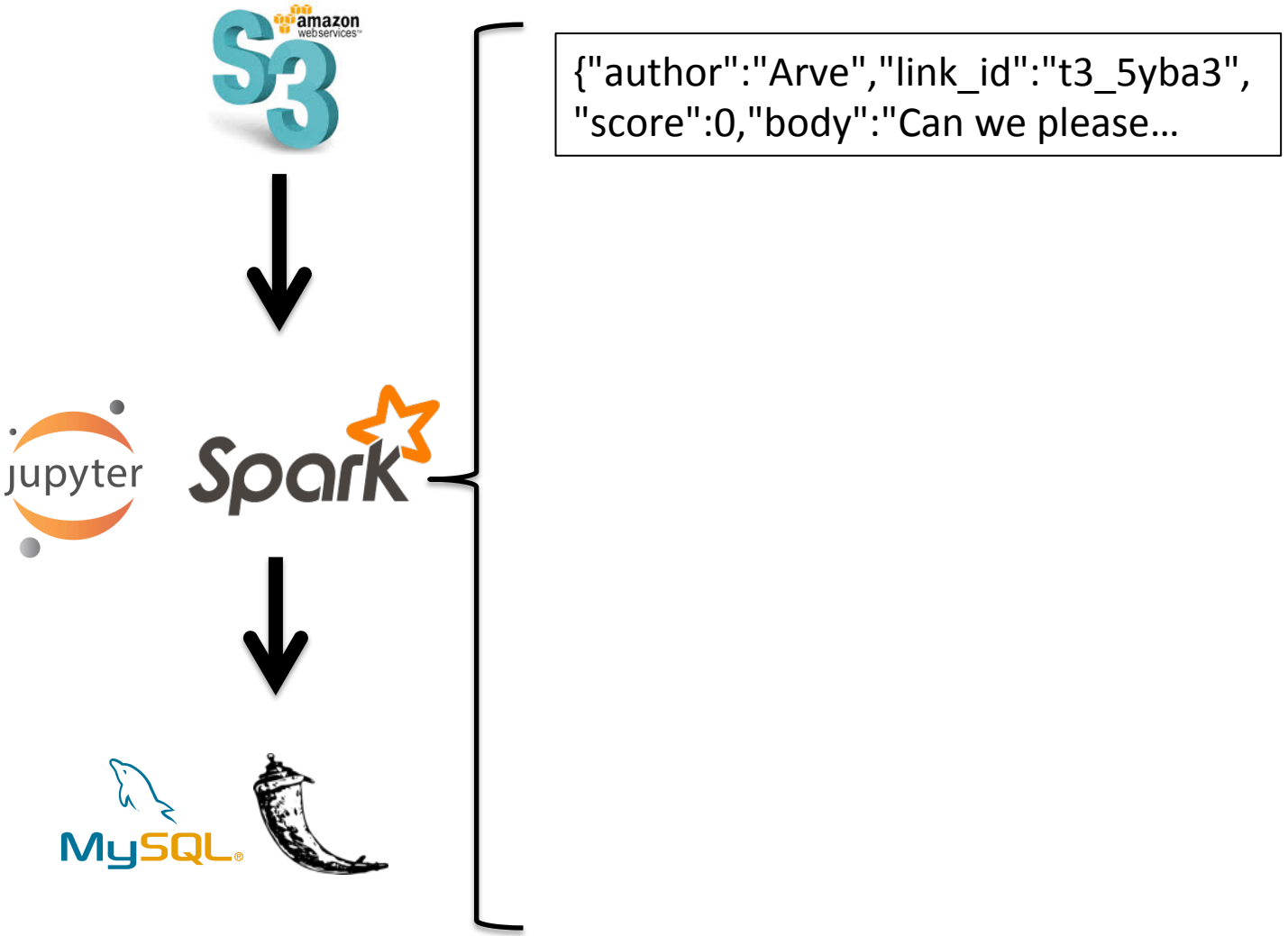
- Scalability
  - Spark: distributed, in-memory, DAG optimized
  - Sparse vector for OHE categorical: memory, time
  - Stochastic gradient descent: sample gradients

Filter to Top/Bottom 3%

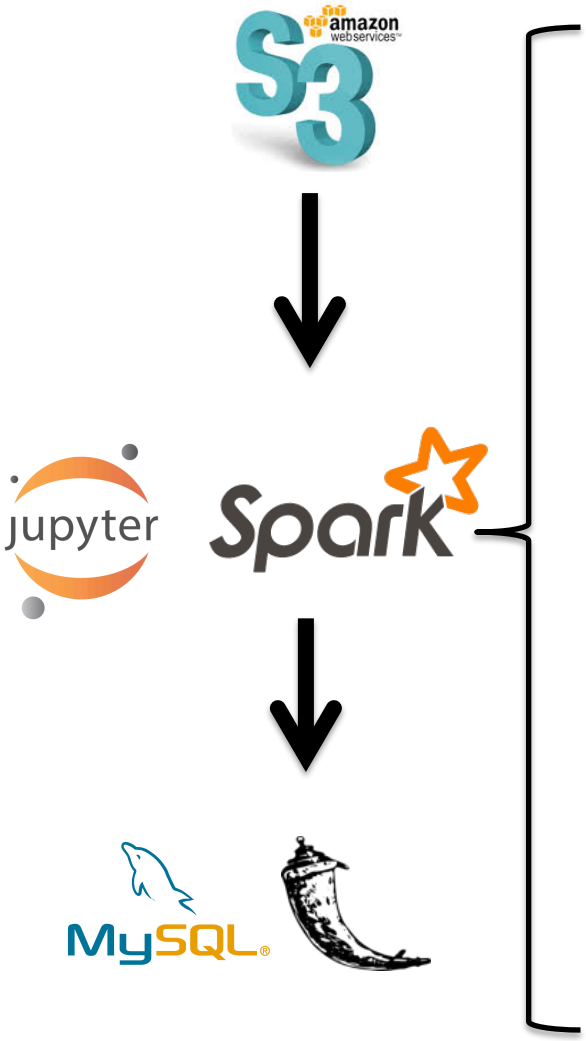


Data and Pipeline

Source: Reddit comments in 0.9 TB JSON on S3.



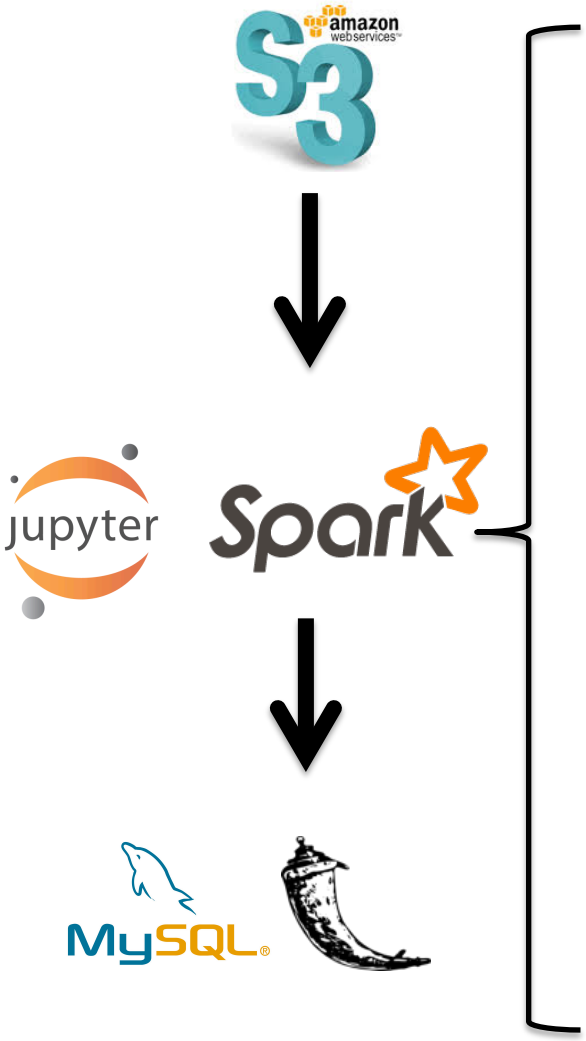
Data and Pipeline



[Row(archived=True, author=u'Arve',
body=u'Can we please…

Filter, percentiles HQL

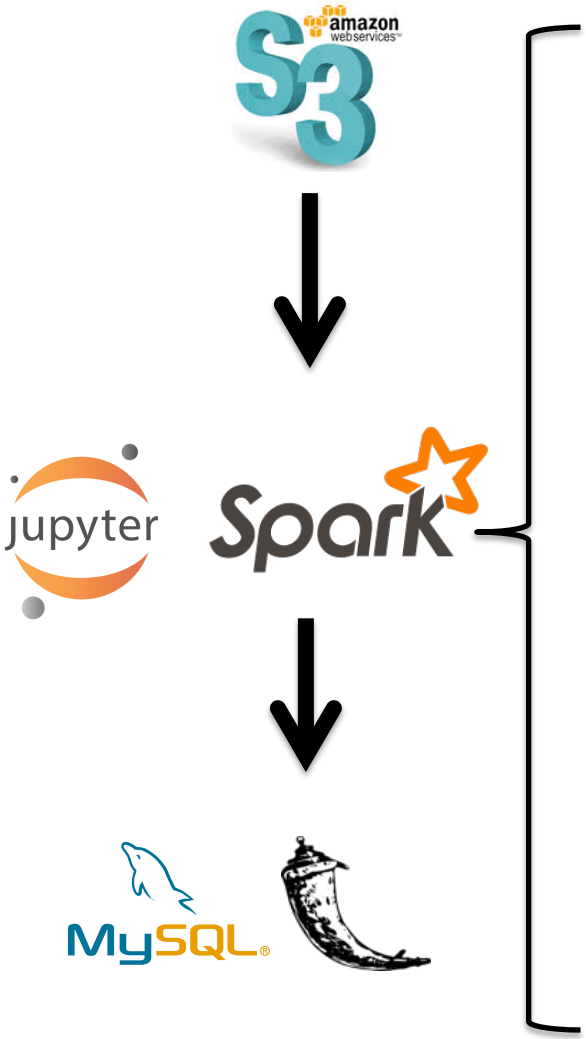
Data and Pipeline



[(u'c0299ce', (u'Can we please…)]

Extract features

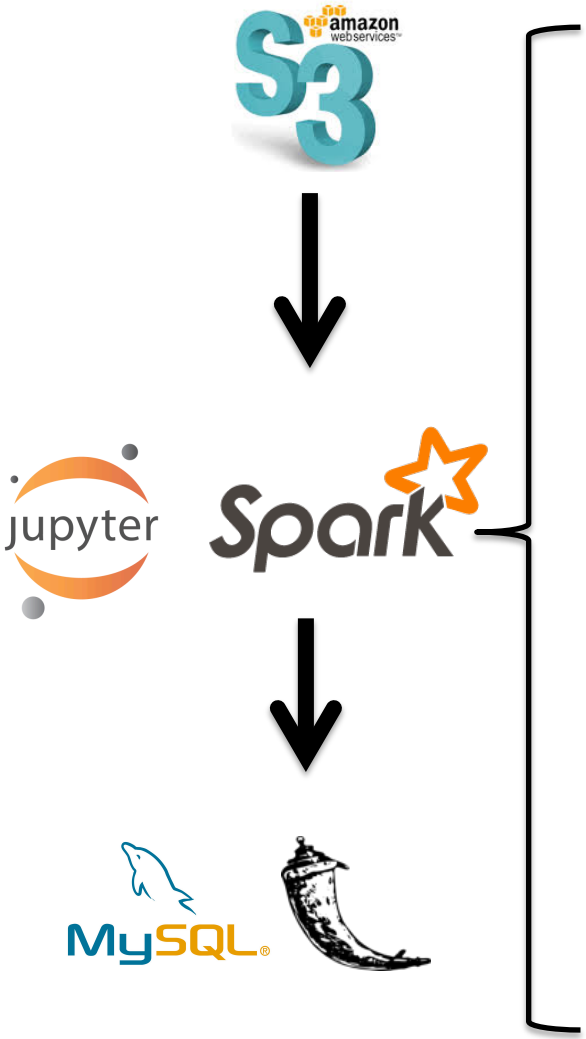
Data and Pipeline



[LabeledPoint(1.0, (6,[0,1,2,5], [11.0,0.0,0.0,1.0]))]

Input for MLlib

Data and Pipeline



[0.0796, 0.1458, 0.5711,
1.150, 0.7430, 0.9329]

Output: 6 FP numbers (weights)

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

- Run time: 7.2 hours for entire 908 GB dataset.

Accuracy and Run Time

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

Accuracy and Run Time

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

IS THAT THE BEST YOU CAN DO!!??

Accuracy and Run Time

- Run time: 7.2 hours for entire 908 GB dataset.
- Accuracy: currently 52%
  - vs. 65% accuracy for previous data science project

Accuracy and Run Time

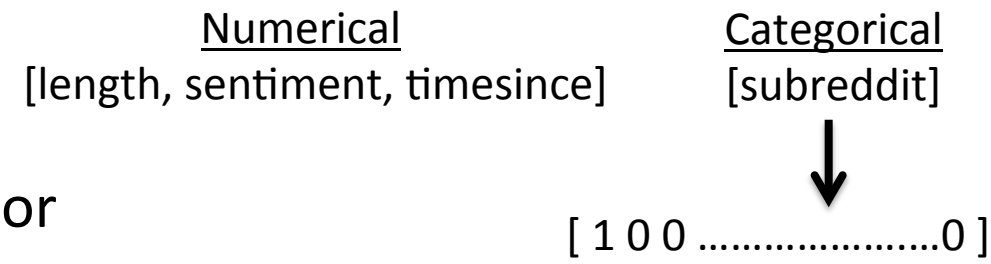
- Run time: 7.2 hours for entire 908 GB dataset.
- Accuracy: currently 52%
  - vs. 65% accuracy for previous data science project
  - Improve by excluding outliers (sentiment analysis) and deep nested comments

Accuracy and Run Time

- Run time: 7.2 hours for entire 908 GB dataset.
- Accuracy: currently 52%
  - vs. 65% accuracy for previous data science project
  - Improve by excluding outliers (sentiment analysis) and deep nested comments
- Pipeline makes further work possible
  - EDA
  - Cleaning
  - Different models
  - Hyperparameters

Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector

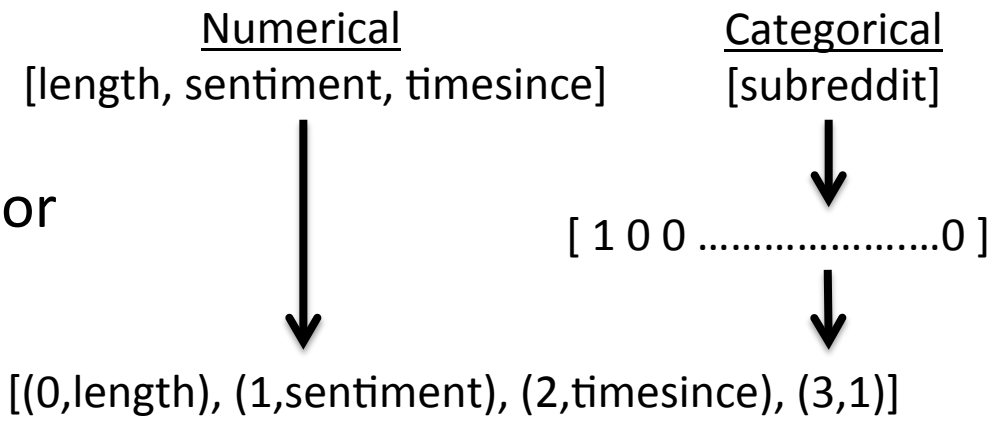


- *Slooooow* percentile, min time with map-reduce

- *Slooooow* brittle join

Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector



- *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

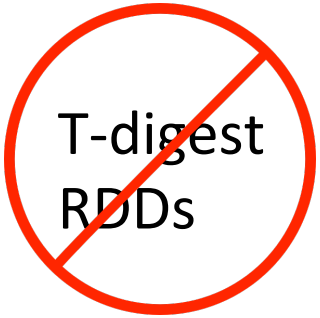
T-digest RDDs

- *Slooooow* brittle join

Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector

- *Slooooow* percentile, min time



DataFrames
HiveQL

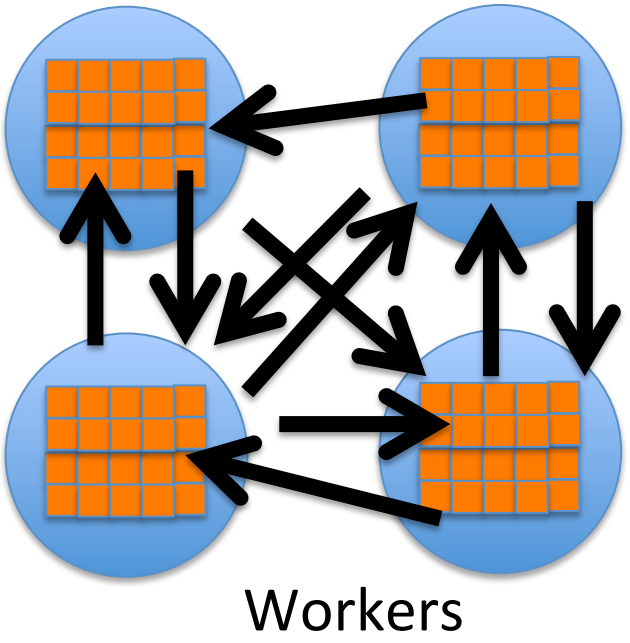
- *Slooooow* brittle join

Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector

- *Slooooow* percentile, min time with map-reduce

- *Slooooow* brittle join



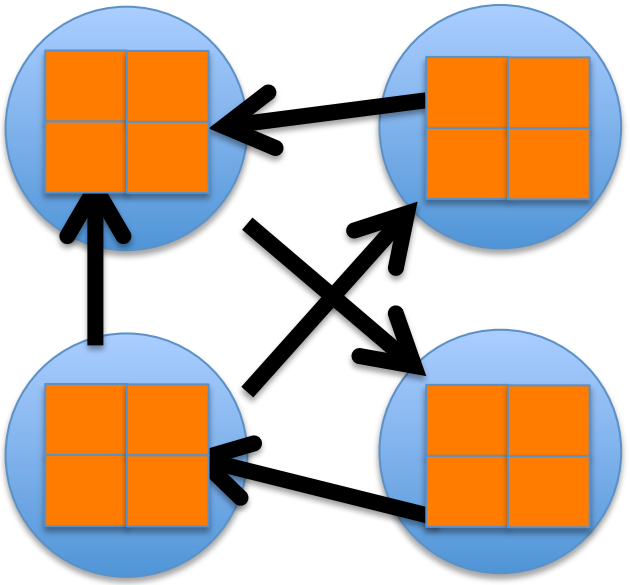
Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector

- *Slooooow* percentile, min time with map-reduce

- *Slooooow* brittle join

- Reduce partitions after filter
- Join with DataFrames
- Size AWS instances for network performance



Challenges & Solutions

- Mixed categorical and numerical features in potentially large sparse vector

- *Slooooow* percentile, min time with map-reduce

- *Slooooow* brittle join

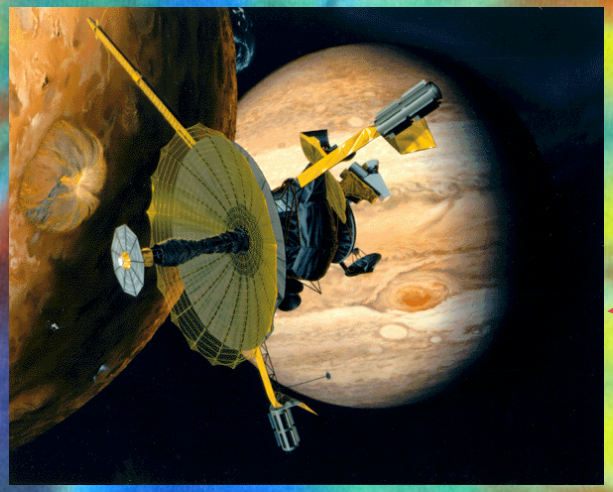
230% speed improvement!

About Me- Jon Neff

Ph.D.
MBA

TEXAS

PEPPERDINE UNIVERSITY



NASA

Backup

Runtime

Activity	Runtime (minutes)	Percent
Read, filter subreddits	84	19%
Create subreddit dict (percentiles)	138	32%
Distinct, min (time), join, collect; create dict & broadcast	90	21%
Extract features, train model	120	28%
TOTAL RUNTIME	432	100%