Partnership for International Research and Education
A Global Living Laboratory for Cyberinfrastructure Application Enablement

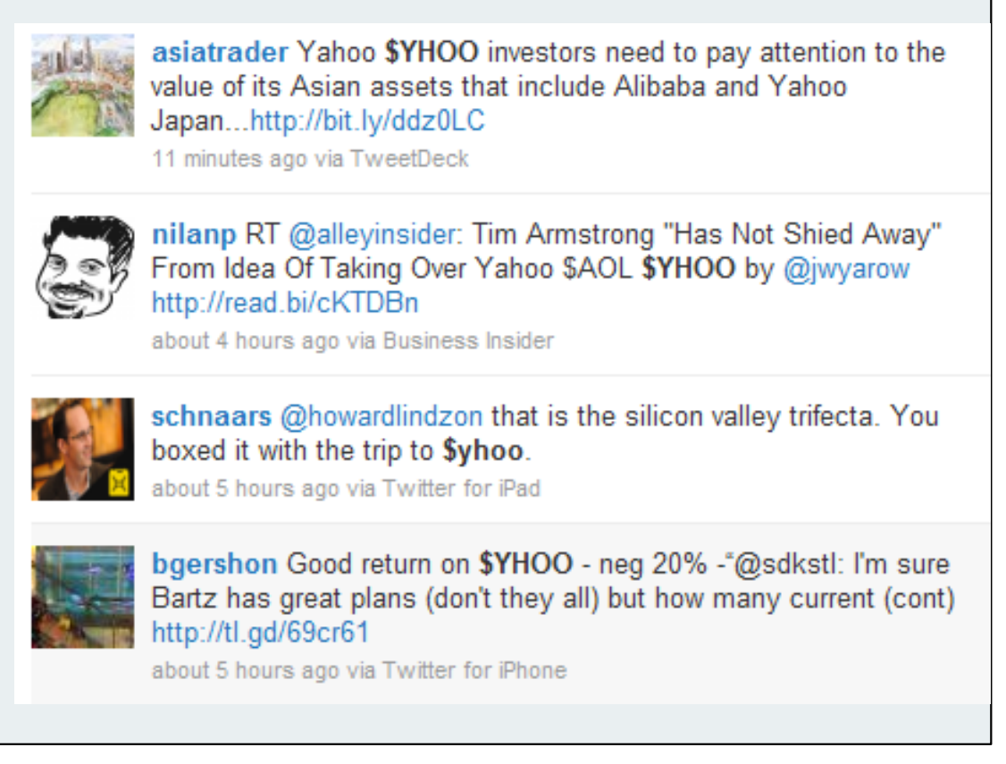
Correlating Real Time Series with Micro-Blogging data

Student: Eduardo Jose Ruiz, Florida International University
Research Advisor: Vagelis Hristidis, Florida International University
CI-PIRE Partner Advisors: Aris Gioannis, Carlos Castillo. Yahoo Research Barcelona

I. Research Overview and Outcome

Overview

- Investigate the possible correlation of time series and micro-blogging data.
- Previous studies show a correlation between news and query logs with real time series (stock, diseases).
- **Idea:** micro-blogging is rich on features (small text, sentiment, social network) . We can add relational features (related with the tweet graph)
- **Applications:** prediction, anomaly detection or entity extraction.
- Initial study shows that changes on a stock [price/ exchanged volume] are correlated with tweeter


asiatrader Yahoo $YHOO investors need to pay attention to the value of its Asian assets that include Alibaba and Yahoo Japan...http://bit.ly/ddz0LC
11 minutes ago via TweetDeck

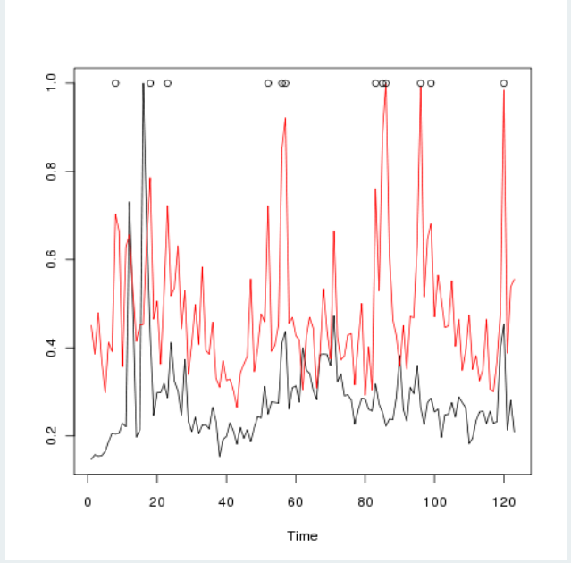
nilanp RT @alleyinsider: Tim Armstrong "Has Not Shied Away" From Idea Of Taking Over Yahoo $AOL $YHOO by @jwyarow http://read.bi/cKTDBn
about 4 hours ago via Business Insider

schnaars @howardlindzon that is the silicon valley trifecta. You boxed it with the trip to $yhoo.
about 5 hours ago via Twitter for iPad

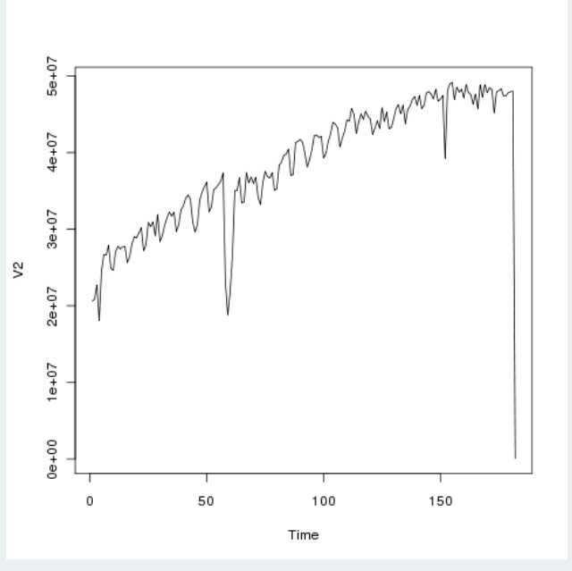
bgershon Good return on $YHOO - neg 20% -"@sdkstl: I'm sure Bartz has great plans (don't they all) but how many current (cont) http://tl.gd/69cr61
about 5 hours ago via Twitter for iPhone

Data Selection

- We select 150 random stocks from the S&P500 Index.
- We select changes on volume/price change such that $p(x > c) < 0.1$
- We process all the tweets and stocks published between 01/01/10 and 06/30/10.
- Filter the tweets using the ticker symbol ($YHOO) and company name (#Apple).
- Sample 30 tweets from the filtered tweets and check if they are related with the company and Biz domain.

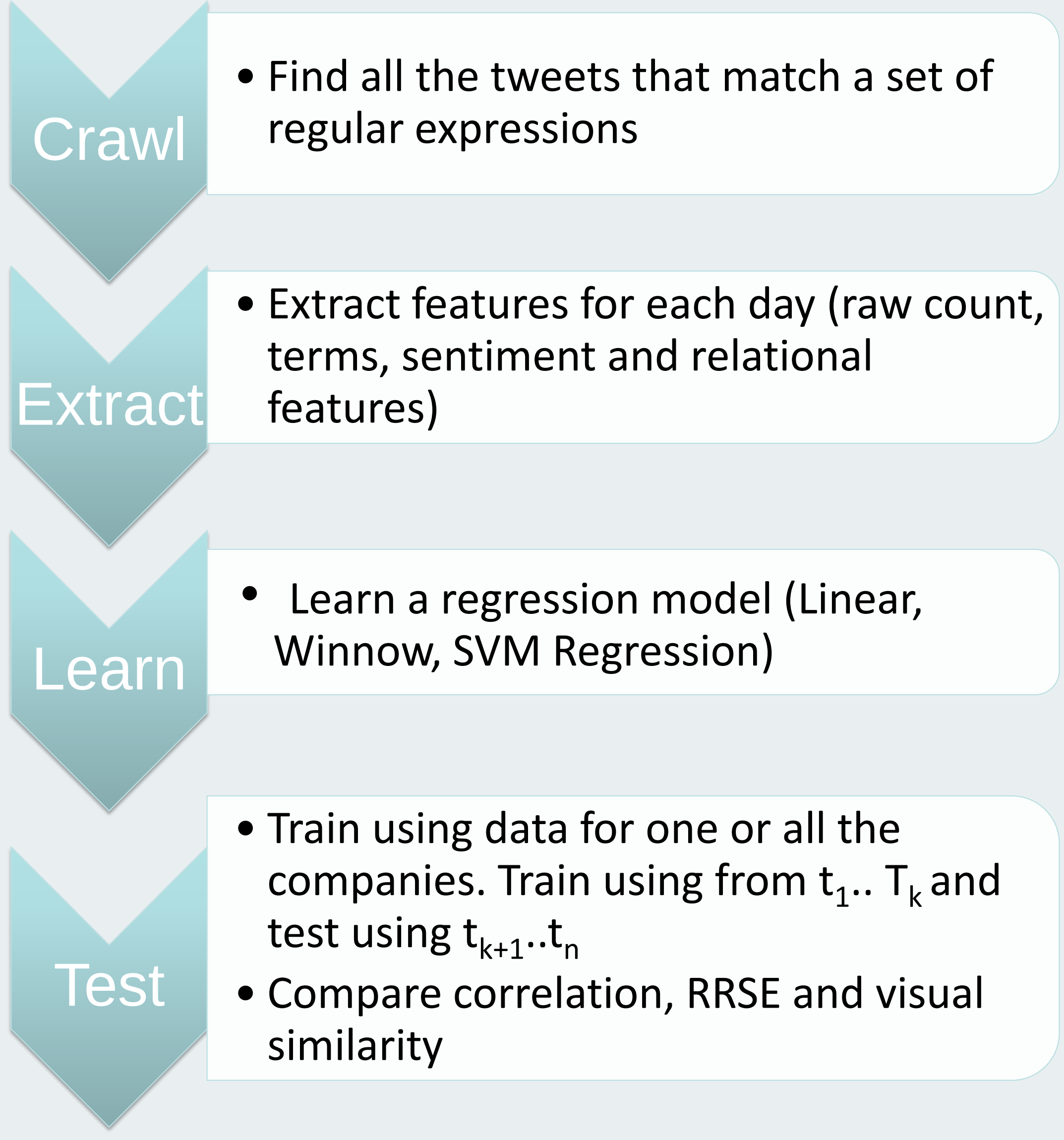


Volume change for ORCL + Tweets per day



Tweets per day between 01/10 – 07/10

Methodology



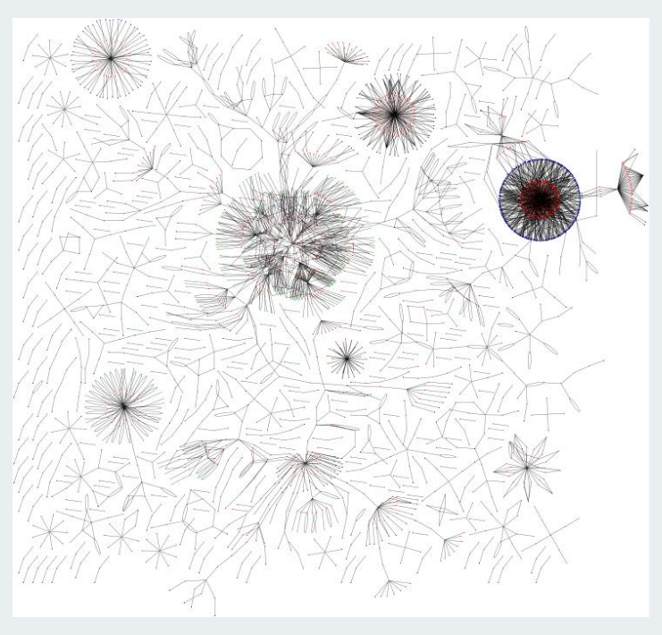
- Find all the tweets that match a set of regular expressions
- Extract features for each day (raw count, terms, sentiment and relational features)
- Learn a regression model (Linear, Winnow, SVM Regression)
- Train using data for one or all the companies. Train using from t_1.. T_k and test using t_{k+1}..t_n
- Compare correlation, RRSE and visual similarity

Implementation

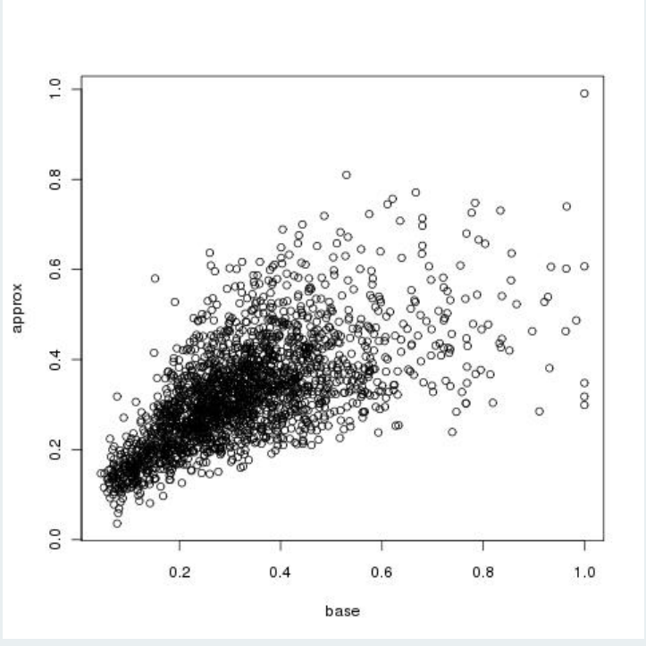
- The data extraction and processing was implemented using the map-reduce framework provided by Yahoo.
- Feature generation was implemented with Pig.
- Data Analysis and figures were implemented with R
- The regression models where implemented using WEKA

Results

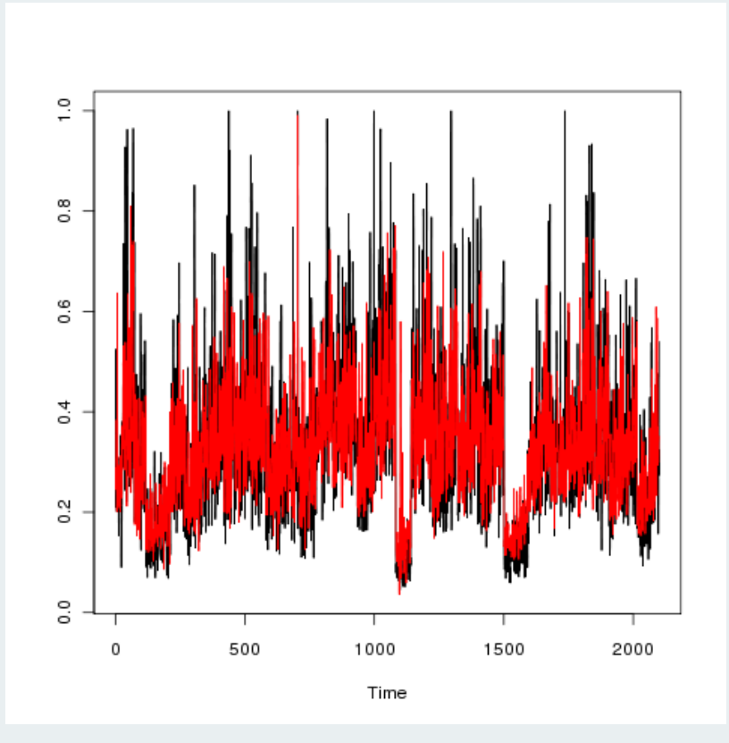
- We can **predict** the volume of stocks exchanged (price changes are more difficult)
- Simple tweet or user count is **not** enough (low correlation)
- **Best:** sentiment, top-50 tokens selected with Info. Gain, Raw Counts (user, re-tweets, tweets, hash-tags), Relational Features (ratio, #components, degree stats)
- Working on **improving** the graphical features? (Page Rank, skewed distribution, better graph)



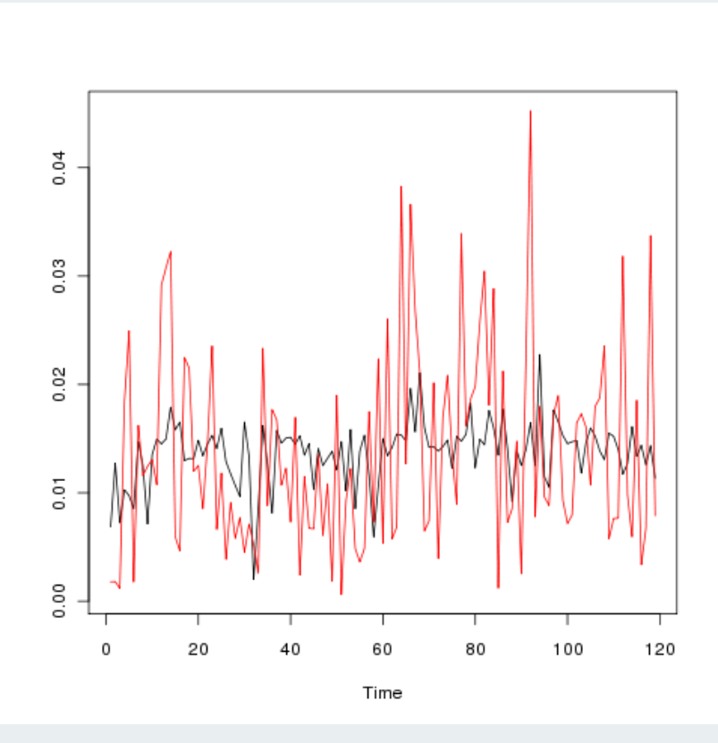
Graphical features for APPL (20100501)



Real vs. approximation



Approximation for the last month for all the tickers



Approximation for YHOO

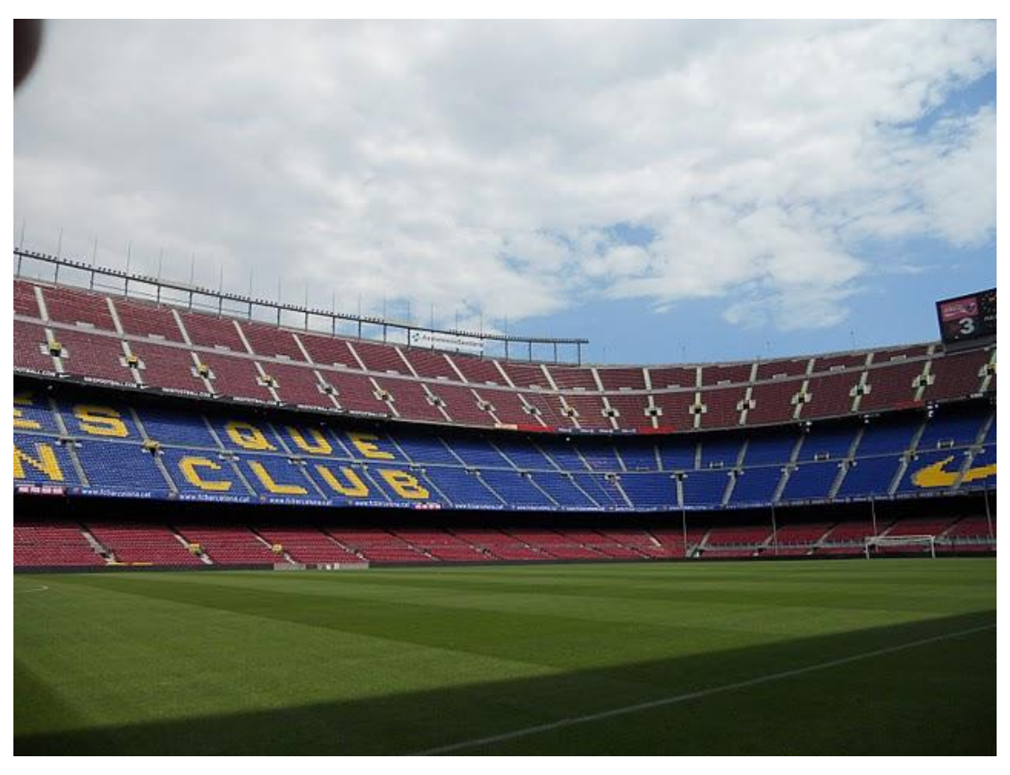
Features	Correlation
Single tweets	0.12
TSC	0.65
TSC+G	0.66

Conclusions and Future Work

- We present a framework to learn regression models that can explain real events using only the twitter data
- We show that some of the common features are not enough and
- We propose adding new **relational** features.
- **Future work:** integrate time series regression with twitter selection

II. International Experience

Shady, Xiao, Michael, Giussepe, Jeff, Me,Mike



Aris Gioannis

Carlos Chato Castillo

Sitges and Girona

- Some beautiful cities around Barcelona can be visited using the train system.
- A beach trip is obligatory in summer

Barcelona: my new favorite city

- A city to live in!! Culture, beach, parks, public transportation and people for everyone. Each walk was a discovery.
- Easier if you speak Spanish. Probably my background helped me to feel more comfortable
- I miss the food …..

Yahoo Research

- Excellent work environment!!!
- Collaborative research between multiple groups (NLP, search, distributed systems).
- Foosball Table!! Even Ricardo plays!!
- Yahoo has the biggest Hadoop cluster. They also have data that is difficult to access on academy.
- **Internships**: advanced students (PhD > 3 year, more than 3 months)
- Multi-lingual "long" lunch: at least 10 countries in my lunch group

Campeones, Campeones, OeeeOe

I had the chance to be in Spain for the 2010 World Cup. I watched the final in Recoletos and I was around for the big party

III. Acknowledgement

The material presented in this poster is based upon the work supported by the National Science Foundation under Grant No. OISE-0730065. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.