# RELATING NEWS ARTICLES SUMMARIES TO STOCK PRICES

MARCELO GROSSI FOR MCM - PRACTICUM

SCHOOL OF COMPUTING, DUBLIN CITY UNIVERSITY

EMAIL: MARCELO.GROSSI2@MAIL.DCU.IE

# PRESENTATION SUMMARY

- Part I
  - Problem and definitions,
  - State of the art,
  - A new approach,
  - Experiments,
  - Results and Conclusion
- Part II
  - Implementation details,
  - Technologies used

# PART I

# INTRODUCTION

- Stock price analyses
    - Can be highly lucrative!
    - Fundamental and Technical analyses
        - Econometrics
        - Chart and indicator analyses
- Increased computational power allows
    - Relating stock price to big external data sources
    - Such as News corpora

# RELATING NEWS TO STOCK PRICES

- How to relate textual data to numerical data?
  - Stock prices → numerical time series
  - News articles → unstructured textual data

# STATE OF THE ART

- Investor sentiment analysis
- Intuition: 'How will the investor react when confronted with news such as these?'
  - 'Economy **exceeds** expectations'
  - 'Market rallies after **unexpected** scenario'
  - 'Investment impacted after huge **loss**'

# INVESTOR SENTIMENT

- **Good**, **neutral** and **bad** news may influence investors into buying/selling stocks
- Techniques to classify investor sentiment
  - Lexicon-based approaches: dictionaries of financially good/bad terminology
  - ML/NLP: Supervised learning classification on annotated corpus
- Can investor sentiment be related to stock price movement?
  - *Gidfalvi, 2001* finds that news have predictive power in a 20 minutes window interval before and after publication
  - *Tetlock, 2007* shows that negative sentiment predicts downward pressure on prices
  - *Barber and Odean, 2008* correlate increase in trading volume with high news sentiment
  - Bloomberg sells document sentiment analysis to investors (*'Bloomberg Market Impact'*)

# PROBLEM SOLVED?

- Sentiment is not general (lacks stock context)
  - What is good news to one is not necessarily good to all
- Best results achieved through machine learning techniques
  - Needs big annotated corpus
  - If not applied to each individual stock also lacks context
- Relationship of sentiment and stock price movement is cumbersome
  - Assign probability of document moving prices
  - Create index and using threshold as a significant event – for event analysis

# OBJECTIVES

- Explore a different approach to relating news and stock price movement
- Does this new approach also impact stock prices?
- Can it be used as a viable trading strategy?

# DIFFERENT APPROACH TO RELATING NEWS CORPUS TO STOCK DATA

- More pragmatic and intuitive
- Use textual measure over time to directly transform textual corpus into time series!

- Easier to use – just another time series..
- Better for prediction

- What measure?
  - Relative Importance → daily TF-IDF
  - Gives more weight to ‘important terms’

# THE MEASURE: TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TFIDF)

- Measure associated with a term (or *n*-gram)
  - Term Frequency (TF) in its simplest forms is the number of times a term appears in a corpus
  - Inverse Document Frequency (IDF) is the logarithmic inverse of the number of documents (from the corpus) a term appears in
    - *TF = t (simple) or* $\frac{t}{T}$ *(scaled form); IDF* $= log\frac{D}{d}$
- Where t is the frequency a term appears in a corpus, T is total frequency of all terms in the corpus, d is the number of documents the term appears in and finally D is the number of documents in the corpus
  - *TFIDF = TF * IDF*

# TESTING FOR INFLUENCE

- How to test if term time series has influence over stock prices?
  - 'The effect does not precede its cause in time',
  - 'The cause has unique information about the series being caused that is not available otherwise'
    - Eichler, 2012

- Granger, 1969 and 1980 proposed a model for such a scenario
  - Based on fitting a VAR model with and without the 'causing' series
  - Granger-causality test or Granger-test

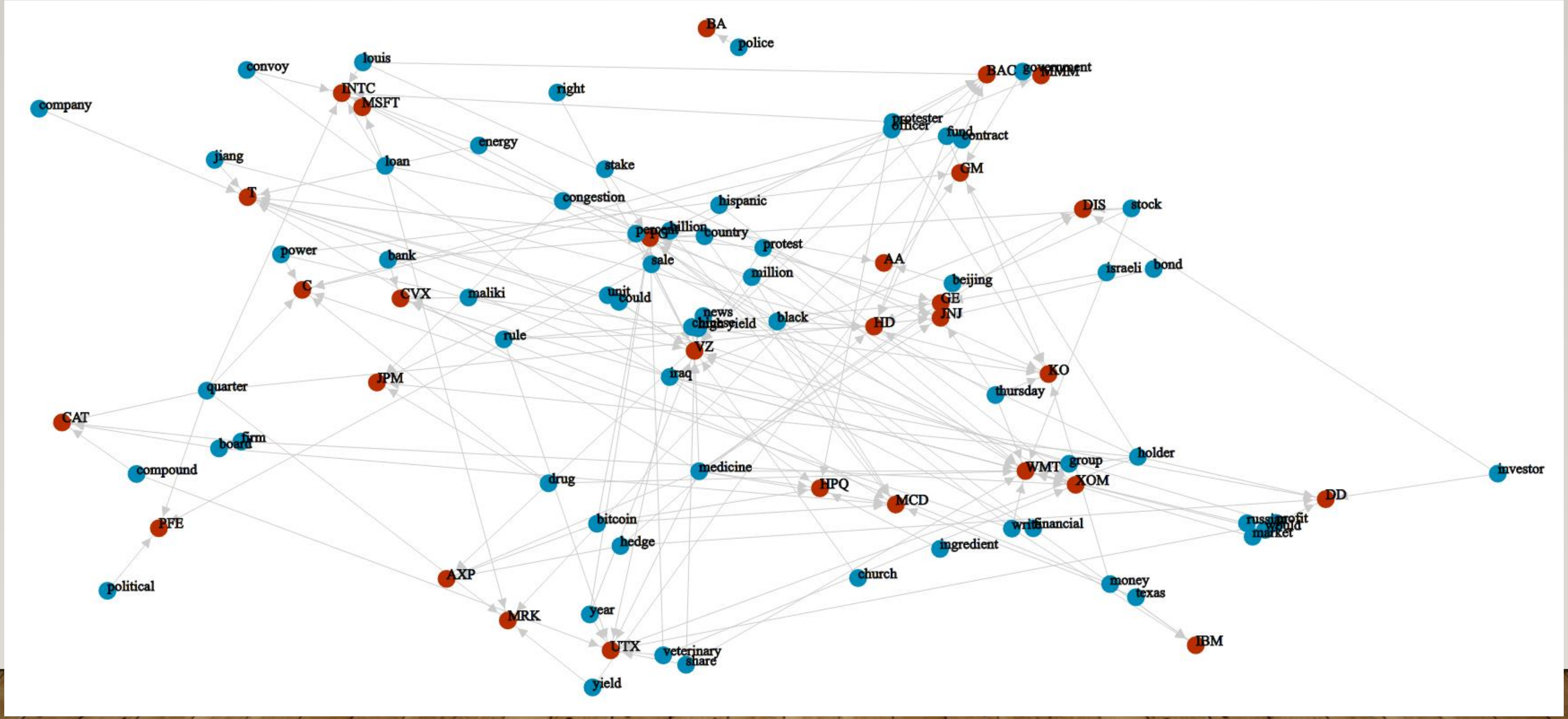
# GRANGER CAUSALITY TEST

- Fit target *y* to Vector Auto Regressive model without the causing time series
- $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \cdots + \beta_n y_{t-n} + \varepsilon_t$

- Add the causing time series *x* to the model and verify if it adds explaining power (*F-Test*)
- $y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_n y_{t-n} + \beta_0 + \alpha_1 x_{t-1} + \cdots + \alpha_n x_{t-n} + \varepsilon_t$

- Not uncommon for Granger-causality in both directions $y \leftrightarrows x$
  - Only accept as influencer if $x \rightarrow y$ and $y \nrightarrow x$

# GRANGER-CAUSING TERMS TO GRANGER-CAUSED STOCKS (FORCE DIRECTED GRAPH)



# VALIDATING AND FORECASTING

- Cross validate Granger-causality with another model
- Can the forecasts be used as a viable trading strategy?
- Widely used model for time series analysis
  - Variant of the ARMA model (Auto Regressive Moving Average)
  - The ARIMAX model (with Integration and eXogenous variable) allows for predictor variable to be included in the model
    - Differencing done prior to applying the model on both predictor and predicted variables if non-stationary (via KPSS test)
    - *ARMAX (p,q)* model
    - $y_t = \beta x_t + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 z_{t-1} + \cdots + \theta_n z_{t-q} + z_t$

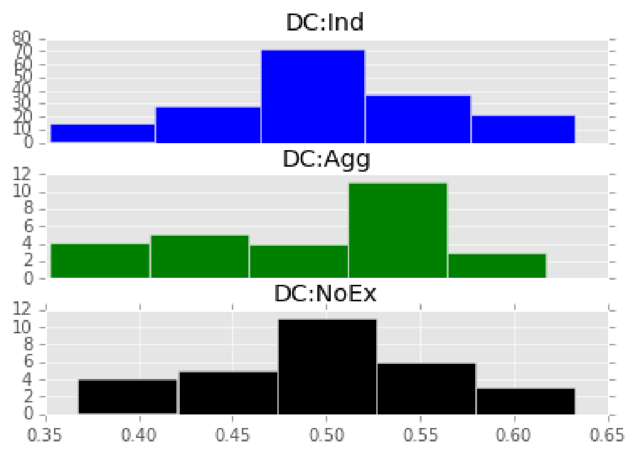
# RESULTS

- Do the new term time series corroborate the Granger-tests results?
  - Root Mean Squared Error
    - Comparison between models with individual predictor (IND), all identified predictors (AGG) and no predictors (NOEX)
    - Two tailed F-Test shows difference between IND, and NOEX, and also AGG and NOEX with 10% significance, confirming the Granger-causality tests

| | AGG | NOEX |
|---|---|---|
| **IND** | 0.564381 | 0.067272 |
| **AGG** | | 0.08543 |

# RESULTS (CONTD.)

- Can it be used as a viable trading strategy?
  - Directional Correctness Ratio
    - Simple measure that avoids pitfalls of convoluted trading strategies to validate theories
    - If not random (mean significantly different than 50%) means it is viable
    - Results show random behaviour and thus can not be used directly as a viable trading strategy




| | MEAN | VARIANCE |
|---|---|---|
| IND | 0.5004 | 0.0042 |
| AGG | 0.4961 | 0.0053 |
| NOEX | 0.4969 | 0.0043 |

# CONCLUSION

- Granger-causing terms not always intuitive
  - May be serving as a proxy for unknown variable
  - Can be improved by filtering allowed terms (use only nouns, named entities, etc.)

- ARMA models could fit the data better if more data was available
  - Re-test with 10+ years news and stock data and compare results
  - Could show improvement in directional correctness ratio (better fit!)

# CONCLUSION (CONTD.)

- Does this new approach also impact stock prices?
  - Stock price forecasts made using the related term time series show better results than not using it
  - Using all available term time series that are related to the same stock improve results further

- Can it be used as a viable trading strategy?
  - Directional correctness does not show deviance from random

# FUTURE WORK

- Repeat experiment using bigger data set
  - Spanning more years,
  - News corpora from different sources
  - Can compare results between different news providers
- Produce bag-of-words using filtered terms
  - Nouns, named entities, etc.
- Use the related term time series of different stocks and calculate distance
  - Can use cosine similarity and produce graph of related stocks (from their related terms)
  - May uncover non intuitive relationships that could prove useful as trading strategy

# PART II

# DATA SOURCES

- Daily historical stock data from Yahoo! Finance
  - Python Pandas Data Reader

- New York Times news portal (http://www.nytimes.com/)
  - Python Scrapy
  - Scraping over 148,000 news articles from January 1st 2013 to October 19th 2014
  - Filtered news articles by business related categories ending up with 49,227 articles
  - Transformed from HTML to JSON files (with title, text and date)

# TEXT PRE-PROCESSING

- Pre-process news article raw textual content into bag-of-words (*n*-grams)
- Python scripts
- Sequence of transformations
  - Expand contractions: ‘it’s’ → ‘it is’, ‘can’t’ → can not, ..
  - Stopwords removal: ‘be’, ‘but’, ‘by’, ‘each’, ‘for’, ‘and’, ‘them’, ‘we’, ..
  - Process sentences independently: So *n*-grams bigger than *unity* can only come from the same sentence.
  - Work tokenization: ‘What a great day!’ → [‘what’, ‘great’, ‘day’]

# TEXT PRE-PROCESSING (*CONTD.*)

- Sequence of transformations (contd.)
  - Lemmatization and Part-of-Speech Tagging:
    - 'saw' + 'verb' → 'see'
    - 'saw' + 'subject' → 'saw'
  - Word *n*-gram generation (*1* and *2*-grams)
- Bag-of-words saved to MongoDB for TF-IDF calculations

```
{
    "source": "source of news article",
    "date": "date in the format YYYY-MM-DD",
    "tags": ["list", "of", "tags"],
    "title": "title of news article",
    "url": "url of article",
    "bag_of_words": [["ngram1", freq1],..]
}
```

# TF-IDF CALCULATIONS

- Use MongoDB Map-Reduce-Finalize to calculate daily summaries
- Easy to get *n*-gram time series (saved as dictionary, so efficient to query)
- From this collection, it is easy to calculate TF-IDF for any period – as needed

```
{
    "date": "date in the format YYYY-MM-DD",
    "total_docs": number of unique news articles,
    "total_terms": number of unique terms,
    "term_counts": {
        "ngram1": [freq1, doc_freq1, tfidf1],
        "ngram2": [freq2, doc_freq2, tfidf2],
        …
    }
}
```

# DATA ANALYSIS

- Python Pandas
- Python Stats Models (http://statsmodels.sourceforge.net/)
  - Initiative to make Python a fully-featured statistical platform
- IPython Notebook for results and interactive data exploration

# THANK YOU!

Questions?