Interactive Data Exploration and Analytics

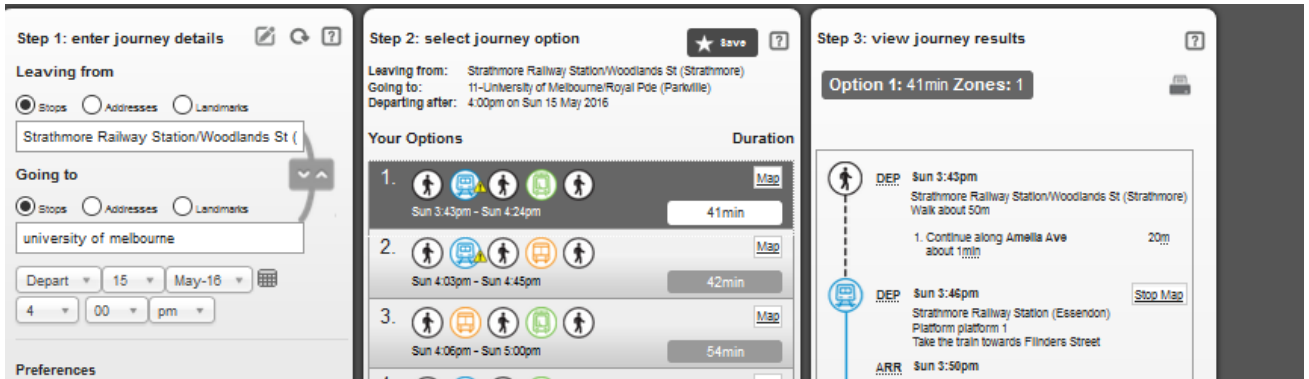
Anthony Liew 360459
Prakriti Giri 784333
Tolga Ozdogan 744166

Problem Background

- Massive amount of data i.e. Big Data and the need to gain knowledge from it.
  - Marketing purposes – targeted advertising based on trends and buying patterns
  - Scientific discoveries
  - Improved emergency services
- "Knowledge is power" – Francis Bacon
- What is the best way to make use of big data?
- Traditionally:
  - finite data within a database
  - Users have good knowledge of database schemas
  - Specific requirement for what the output needs to be
- This method WILL NOT work for big data.
- Need for an efficient method to explore big data in order to find interesting patterns or make new discoveries. Hence, "interactive data exploration and analytics"

Interactive Data Exploration and Analytics

- Enables the extraction of interesting patterns or knowledge from big data without knowing beforehand what is being sought for
- It involves continual analysis of a dataset, often one that typically exceeds the processing power of conventional databases, i.e. big data
- As defined by Idreos: *"to allow for instant access to the data, i.e. , without expensive initialisation steps , while at the same time, allowing the user to extract knowledge from the data by selectively and interactively investigating only parts of the data."*
- It is useful to researchers, scientists, policy administrators, charitable organisations, and just about any organisation interested in acquiring information from unprocessed data.
- Everyday examples:
  - Google Maps, PTV Journey Planner



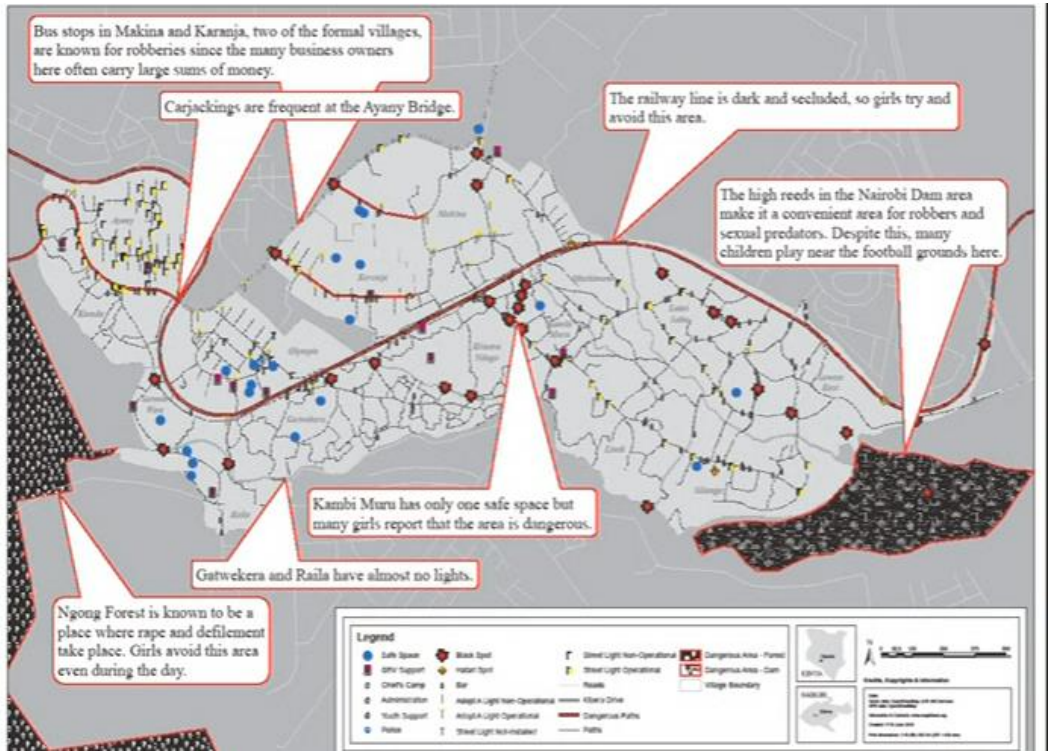
Facets of Data Exploration

- Visualisation Tools
  - Presenting data in a graphical form assists in gaining new insights and hypothesis about the data
- Exploration interfaces for user interaction
  - Facilitate user interaction with underlying data
  - Query Result Visualisation
  - Exploration Interfaces
    - Automate the process by discovering relevant data objects
    - Assist formation of queries

- Middleware
  - How can the performance of the data exploration task be improved?
  - Data Prefetching
    - Reduce overall exploration time
    - Challenge: Identifying promising data sets
  - Query Approximation
    - Approximate results improve response time
    - Query execution over subsets of data
- Database engine
  - Redesigning the database system architecture
  - Adaptive Indexing
  - Adaptive Loading

Advanced Projects/Applications

- Open Street maps(left) and iSpark(right)



- Applications allowing interactive data exploration can be useful for emergency services and humanitarian relief organisations as well

Building the application?

- The first and foremost stage before the whole data exploration and analysis process

- Data collection and manual data analysis
  - Selection of data sets critical for completeness, performance later
  - Involves web crawling or retrieving historical data from reputable sources
  - Human analysis of samples of underlying data, to determine features to start with
- Machine learning to organise and identify patterns
  - Algorithms vary widely according to applications
  - Include: Statistical Regression, K-Clustering, Neural Networks
- Caching the data

Vinem – A Data Mining Platform

- Visual Interactive Neighbourhood Mining on High Dimensional Data
- Purpose: Finds sub-space clusters over HD data i.e. finds sub-groups of objects that are related to each other
- Principle: Sometimes unsupervised methods of machine learning are more suitable for data exploration when little is known about the data in advance (particularly that which is usually raw or dirty)

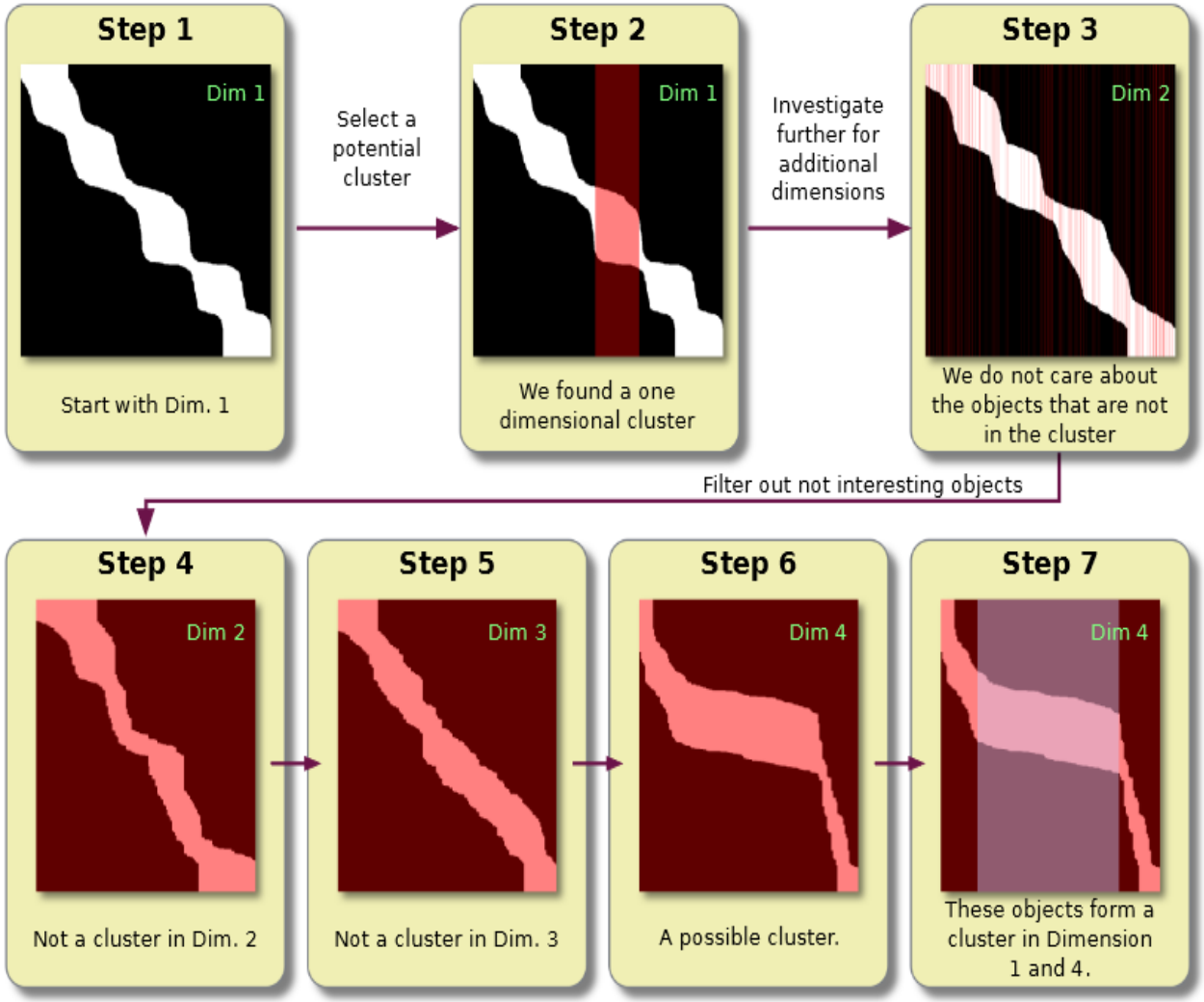
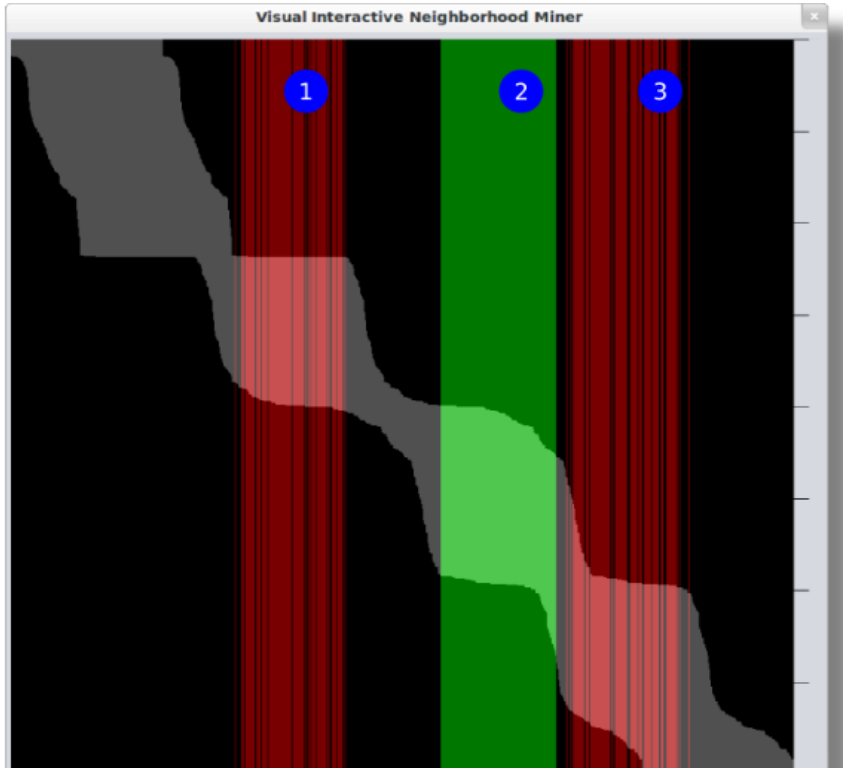
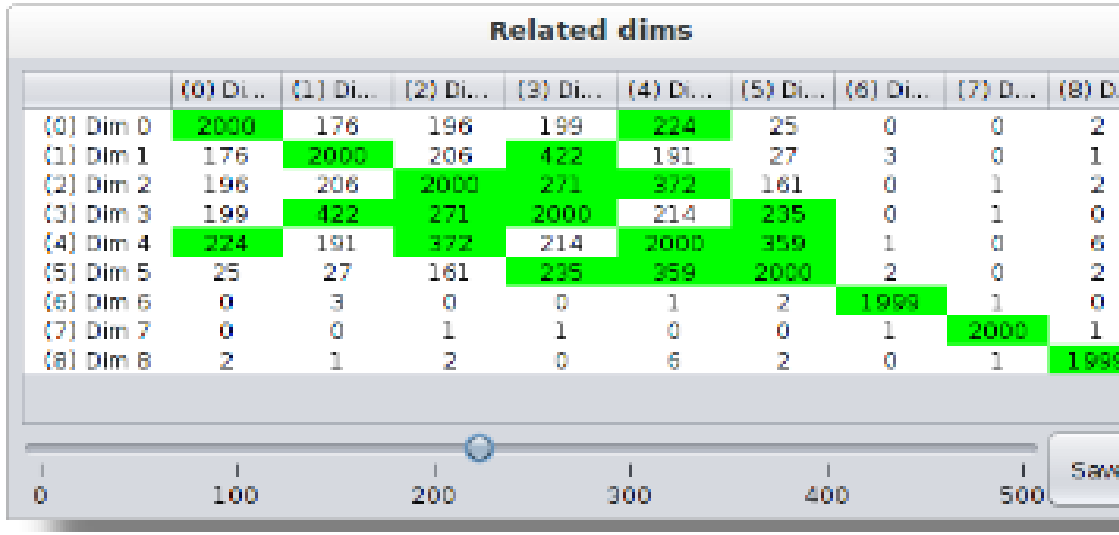


Fig: Data Mining Process in Vinem

Vinem – Visual Results

- In the visual interface users have control over:
  - the data sets to be included parameter selection
  - The type of neighbourhood (k-nearest of euclidean)
  - The mining method to focus on (fast or sampler.)
  - The required support weighting for individual attributes.
- There are different views available for displaying the results:
  - object neighbourhoods
  - member objects per cluster and the relevant attributes that were used for grouping them together.
  - E.g. right (Neighbourhood matrix, relevancy score of dimensions)





Text Data Exploration

Current Event Detection Practices

- *Detection of bioterrorism:* monitoring of health data such as sales of OTC medicines to detect events related to disease outbreaks like epidemics.
- *Detection of health related issues:* Monitoring of daily data feeds of 20,000 hospitals, pharmacies, medicine sales both temporally and geographically.
- **Shortcomings:** These applications lack the ability to handle text corpora which has the potential to include rich information.

Leadline Introduction

- The content of news*(CNN News)* and online social media *(Twitter microblogs)* embody important insights regarding the important events happening everyday.
- ***LeadLine***, added the capability of connecting topical themes with associating events on top of existing text data exploration tools.
- ***Leadline*** used 4W's approach which is What, Who, Where, When:
  1. Topical Modelling
  2. Early Event Detection
  3. Named Entity Recognition

- **Challenge:** An important obstacle is to keep the valuable information lost through the massive influx of text data and extract information in an event-driven approach.
- **Event Formulation:** In this context an event will be described with the tag *<Topic, Time, People, Location>*

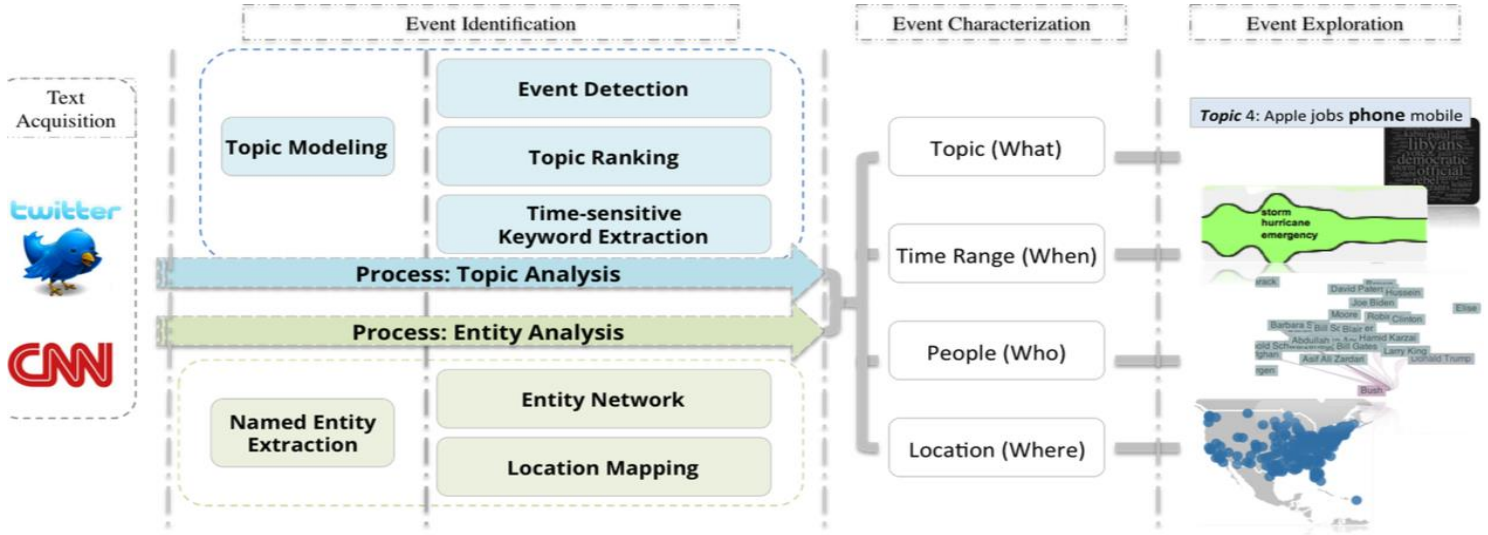
Leadline

Functionality of Leadline

- Leadline's main focus is to extract events from the text corpora.
- The text data is first organised into topical streams over time by *Latent Dirichlet Allocation (LDA)*.
- In order to detect the **"burstyness"** and the length of the events in temporal(time) scale an *Early Event Detection Algorithm* is used.
- *Named Entity Recognition* is used in order to discover the event's related People and Location attributes.

Leadline Visual Interface

- The visual interface provides the effective exploration of the events with related People and Location attributes of the events.
- It is used for building up the narratives of the event sequences in temporal(time) scale.
- It enables users to report their findings and provides capabilities for further analysis.



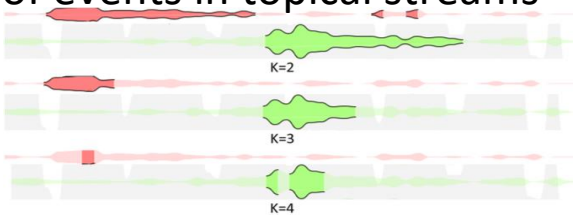
Leadline (contd.)

Data Acquisition

- There are no readily available datasets for CNN news and Twitter microblogs.
- So *web crawler bots* are used in order to acquire required data.
- The data acquisition process for *CNN News:*
  1. crawl the entire web domain
  2. download the pages,
  3. extract, parse and normalise the text articles, remove HTML formatting and noise from the text.
- For *Twitter microblogs*, Online Social Media(OSN) graph based web crawling technique is used. (Users as nodes, connections as edges)
- In total **5 billion** *Twitter microblogs* and **100K** *CNN news articles* are extracted over the course of 3 months.

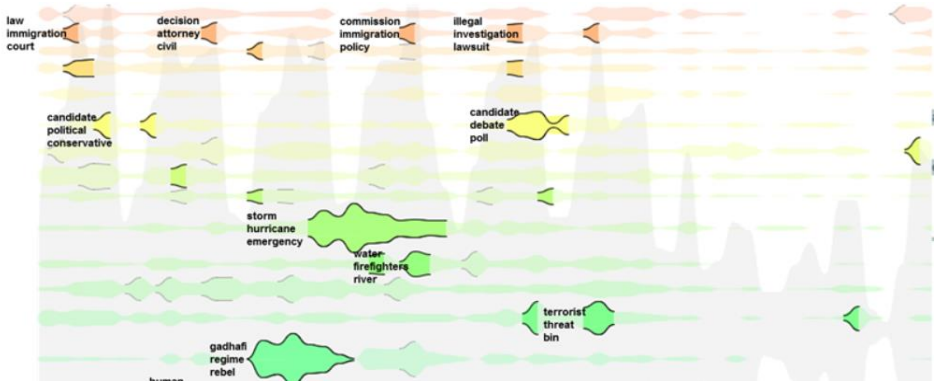
Leadline Introduction

- *Topical Level Event Detection Algorithm*: Find the "burstyness" of events in topical streams



- *Topic Ranking*: To visualise the related topics close to each other in topical stream visualisation

- *Time Sensitive Keyword Extraction*: To find what the event is actually about by using related keywords



Leadline (contd.)

Narratives

- *Leadline interactive system* uses attributes of events **(People, Place, Topic, Time)** to extend explorative capability of the user.
- *WordCloud* is used to summarise related events of Entities(People and Place).
- *Topic Streams* are used to visualise the Events on a river-like topical stream.
- *Geolocation Filtering* is used to extract meaningful information about the events.

Limitations and Improvements

- Limitations arise from the performance of the algorithms used such as *Topical Modelling* algorithm and *Entity Recognition* algorithm.
- *Entity Recognition Algorithm* gives results with uncertainty so the user should be noted.
- The system does not take into account inter-topic related events.
- If several bursts occur at the same time and same related entities, it can be assumed they are triggered by the same event.

Wrap Up

- Why Data Exploration is important?
  - People make decisions not computer, knowledge is power.
  - Big data surpasses conventional databases.
- Overview of Data Exploration facets such as User Interaction, Middleware, Database Engine
- How data exploration systems are built?
  - Data collection.
  - Selection of relevant attributes.
  - Machine learning algorithms.
- Vinem – a data mining platform
- Leadline - Event driven text data exploration tool

Questions

- Explain some ways that explorative data analytics is different from querying traditional databases.
- Discuss why data exploration and analytics is important.
- Give an outline of some steps involved in making an interactive data exploration application.

References

1. Idreos, Stratos, Olga Papaemmanouil, and Surajit Chaudhuri. "Overview of data exploration techniques." Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015. http://stratos.seas.harvard.edu/files/stratos/files/exploration-tutorial.pdf
2. Emin Aksehirli, Bart Goethals, and Emmanuel Müller, “Visual Interactive Neighborhood Mining on High Dimensional Data.” IDEA Workshop on KDD 2015 http://poloclub.gatech.edu/idea2015/papers/p10-aksehirli.pdf
3. Dou, Wenwen, et al. "Leadline: Interactive visual analysis of text data through event identification and exploration." Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on. IEEE, 2012. http://viscenter.uncc.edu/sites/viscenter.uncc.edu/files/CVC-UNCC-12-08.pdf
4. Subhajit Das, Andrea McCarter, Joe Minieri, Nandita Damaraju, Sriram Padmanabhan and Duen Horng Chau “ISPARK: Interactive Visual Analytics for Fire Incidents and Station Placement.” IDEA Workshop on KDD 2015 http://poloclub.gatech.edu/idea2015/papers/p29-das.pdf
5. Madden, Samuel. "Interactive data analytics: the new frontier." Proceedings of the Sixth ACM Symposium on Cloud Computing. ACM, 2015. (Abstract only) http://dl.acm.org/citation.cfm?id=2809956&CFID=614643800&CFTOKEN=25106038
6. Big data: The next frontier for innovation, competition, and productivity http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation
7. What is data analytics? http://searchdatamanagement.techtarget.com/definition/data-analytics
8. Introduction to Big Data Analytics: A Webinar https://www.youtube.com/watch?v=3SK9iJNYehg
9. Paul Conneally: How mobile phones power disaster relief https://www.ted.com/talks/paul_conneally_digital_humanitarianism?language=en