CHAPTER 2:VOLUME

2.1 SOCIAL MEDIA DATA SIZES

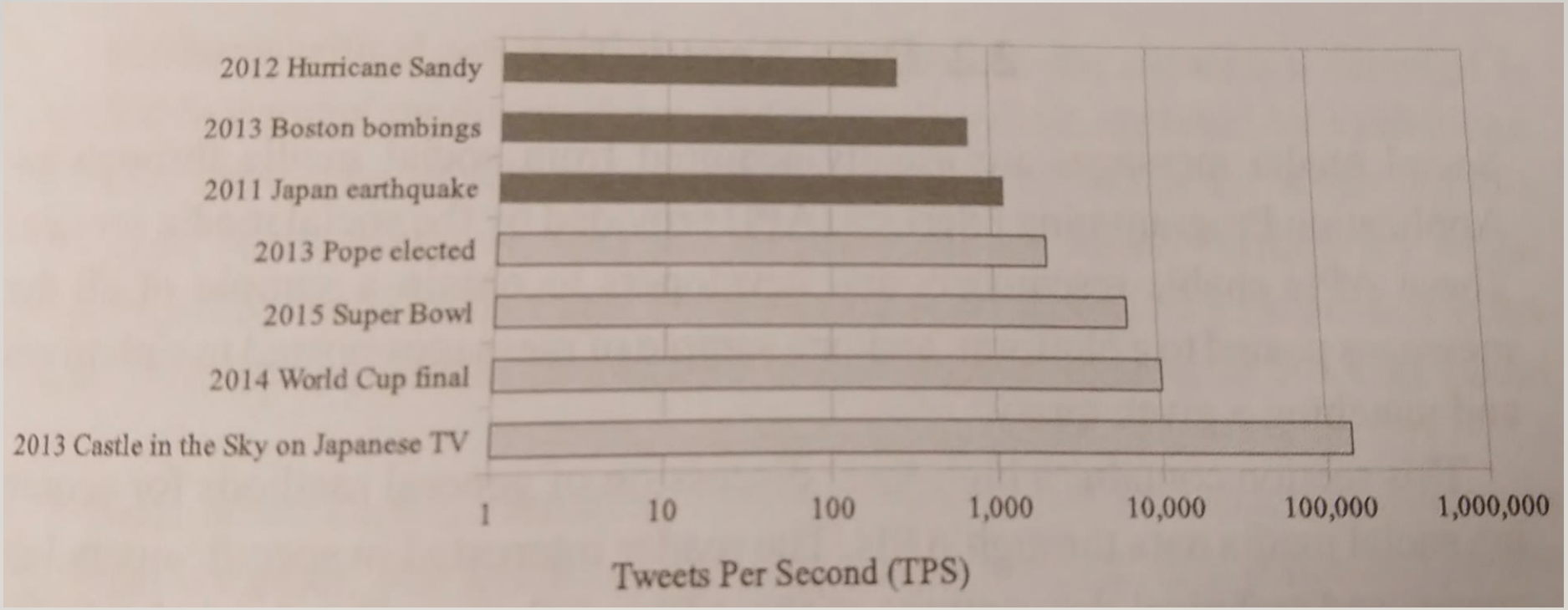
Intro

- Quickly outdated
- Immense (LiveData)
- 4 TB

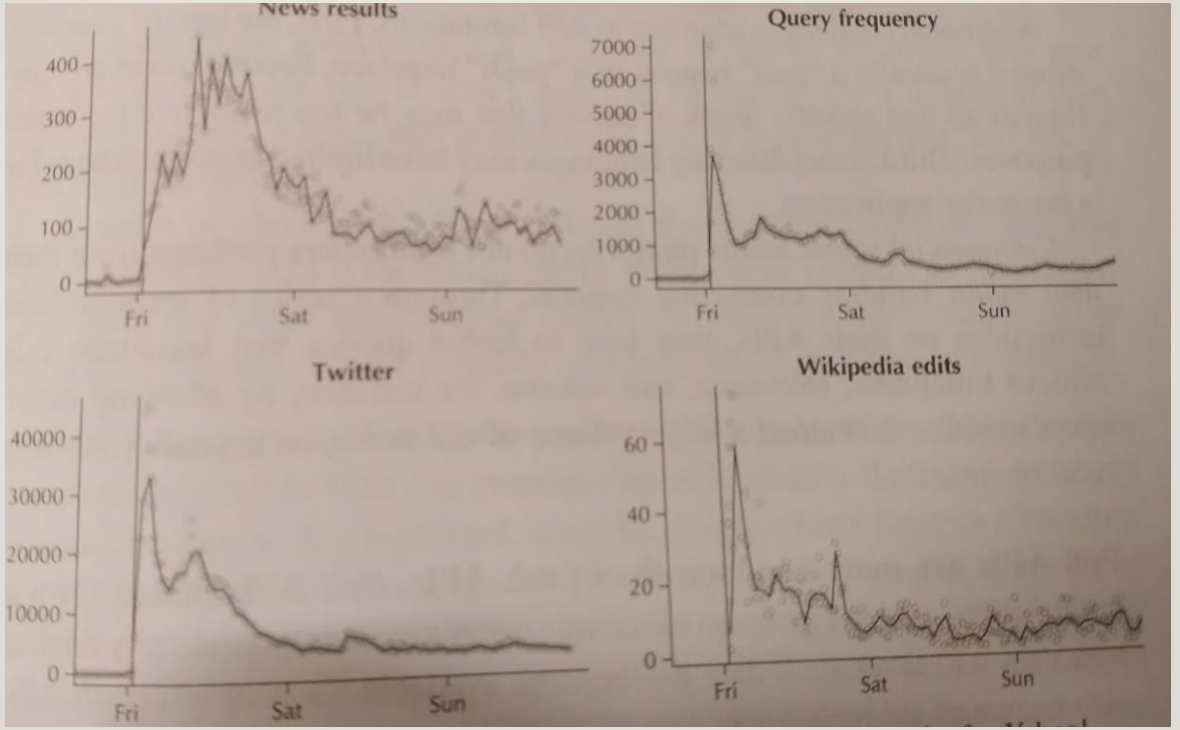
Social Media During Disasters

- Doubles During Disasters.
- Content relevant sites increase
- Forums to Social Media

TPS


2012 Hurricane Sandy
2013 Boston bombings
2011 Japan earthquake
2013 Pope elected
2015 Super Bowl
2014 World Cup final
2013 Castle in the Sky on Japanese TV
1
10
100
1,000
10,000
100,000
1,000,000
Tweets Per Second (TPS)

SENDAI EARTHQUAKE (2011)



2.2 DATA ACQUISITION

- Application Programming Interface (API)

Challenges

- Scale
- Resilience
- Commercial Restrictions

Pull and Push

- Pull is active, Push is passive query
- Pull more common than Push
- Same scalability
- Push is best for time sensitive info
- Twitter uses Push

2.2 DATA ACQUISITION

Rate limitations

- Short term granularities
- Long term granularities
- Max data

Query languages lack expressive power

- High level queries
- Boolean restrictions
- Sampling methods are not transparent.

2.2 DATA ACQUISITION

Query Construction

- Requires background knowledge
- Keyword-based predicats -> not trivial
- Hashtags useful, but conflicting

Adaptive Filtering

- Set of keywords that change over time
- Automatic or guided

Trade-off between precission and recall

- Depends on situation and task what to prioritize

2.2 DATA ACQUISITION - ENRICHING DATA CONTEXT

- Methods to collect messages tend to lose context of messages

Contextual Streams

- The set of messages
- Time and geographical

Time

- Time scope may limit context

2.3 POSTFILTERING AND DE-DUPLICATION

Postfiltering

- Relevancy filter
- Whitelisting

Spam and bot removal

- Remove spam and bots
- Not all bots are bad

De-duplication

- Repeated messages
- Redundancy is a form of relevance

2.4 DATA REPRESENTATION/ FEATURE EXTRACTION

- Data representation is key for mining and search tasks
- It maps input data to expected input of algorithm
- Messages represented as vectors
- Determining characteristiscs is known as feature engineering

TEXTUAL FEATURES

- **Word Features**
- Vector space model
- Single-word feature
- Multiword feature
- *N-grams*
- Further details later

- **Nonword features**
- Superficial text features
- Nonalphanumeric characters (Emoji)
- Systems are language specific

- **Metadata Features**
- Other relevant features
- Author, time, catagories
- User interactions

2.5 STORAGE AND INDEXING

- No need to store everything
- Unless retrospective analysis
- RDBMS and NoSQL

Indexing

- As datasets grow large, indexing is needed
- Ordered Indexes
- Spatial Indexe
- Textual/inverted indexes

2.6 RESEARCH PROBLEMS

- Expert input required
- Static queries
- Misses the big picture

Developing Architectures for real-time analysis.

- Batch analysis vs Real time

Reducing the overreliance on a single data source

- Twitter is great but...