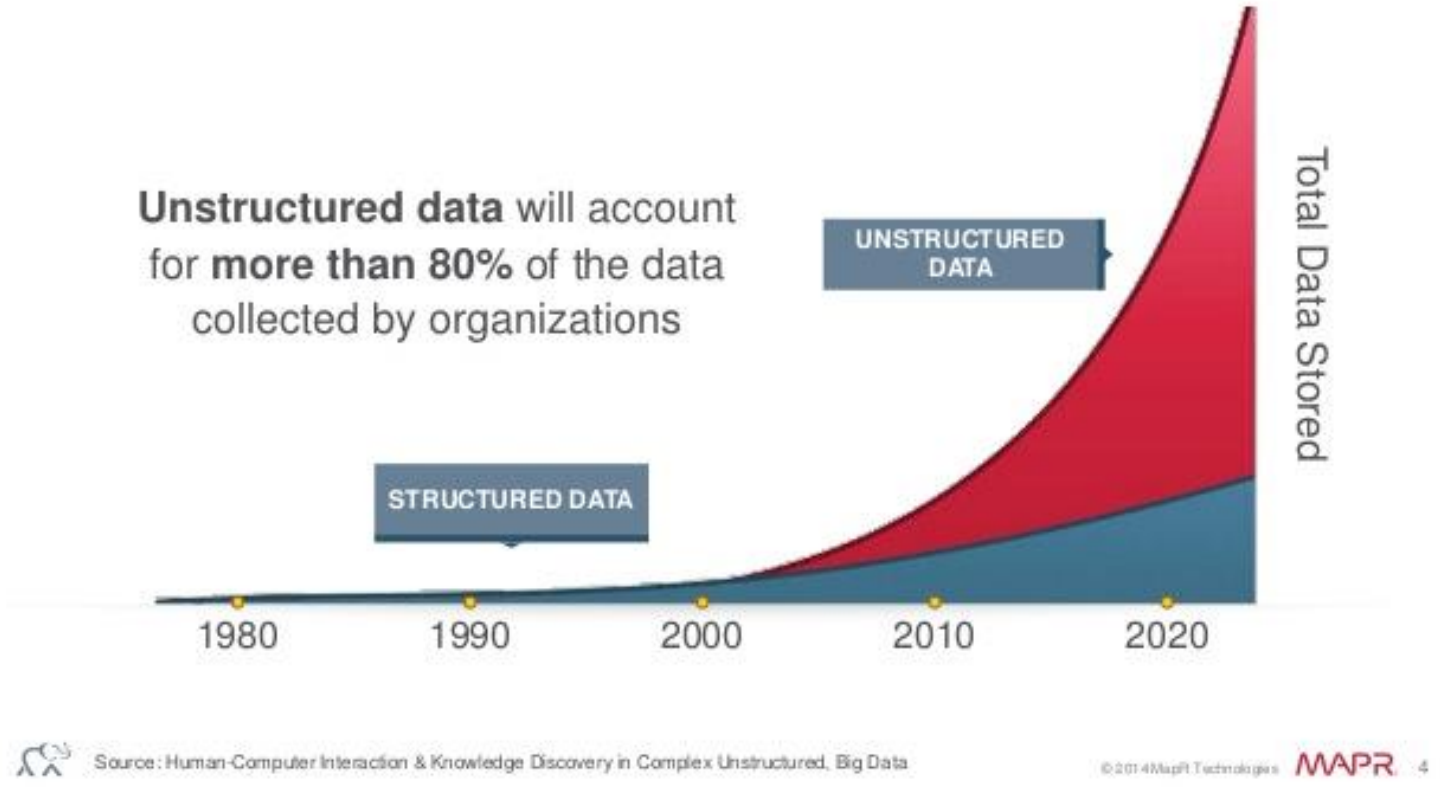
Unstructured Data Analysis

Introduction

By

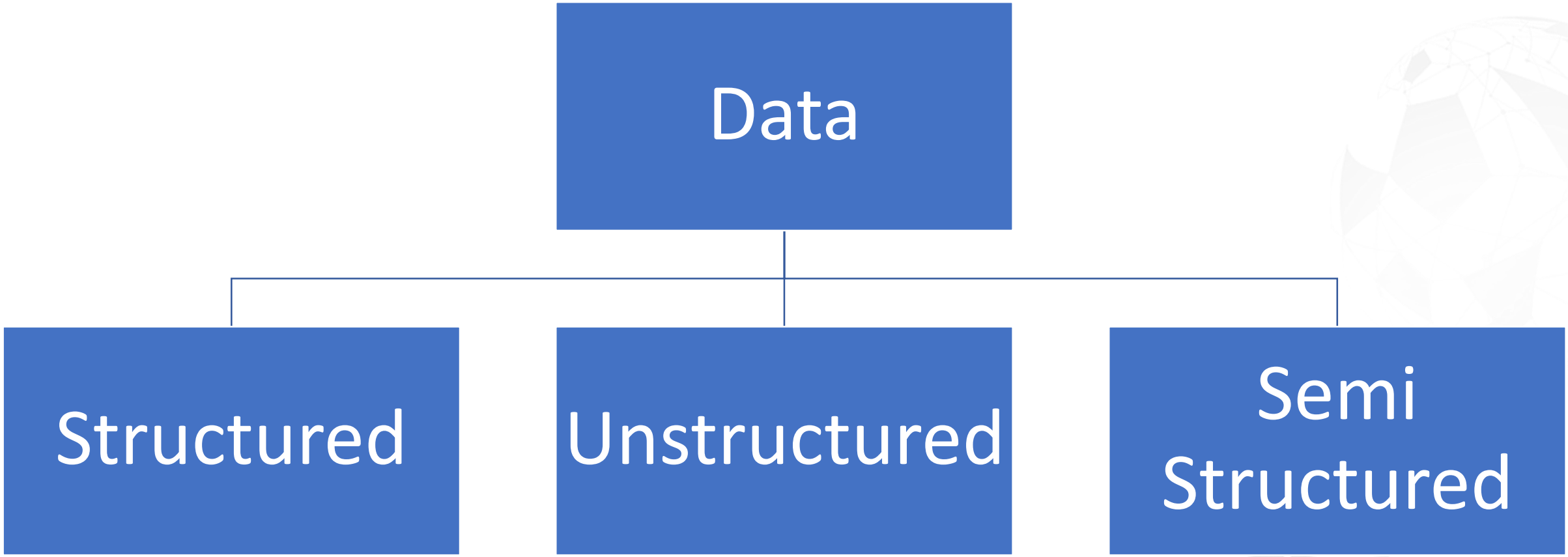
Kathirmani Sukumar

Growth of Unstructured Data



Companies started investing in technologies to work with unstructured data like text, images, videos, audios etc.

Types of Data



- Mostly tabular data
- JSON, web logs

- Text data
- Images,
- Video and audio files

- Tabular data along with large text column

Unstructured Data Analysis

Challenges

By

Kathirmani Sukumar

Sources of Text Data

- Social media platforms
  - Twitter, Facebook, Whatsapp, etc
- Documents
  - Memos, research papers, articles etc
- Email conversations
- Web pages
- SMS
- Transcripts

Humans vs Machines

Humans	Machines
Can understand numbers, text, images..	Can understand only numbers
Can easily spot spelling mistakes	Require algorithms to detect typos.
Can understand sarcasm	Cannot understand them
Can summarize large document	Complex algorithms are required to get basic summary of text
Polysemy: Can differentiate same word used in different context. Ex: bank (banking, river bank)	Only few algorithms can differentiate different meaning of same word
Once trained, manual effort to repeat a process. Ex: Classify 1000 tweets individually as positive, negative & neutral	Once trained, can easily repeat the process within short span (Can classify even 1L tweets within seconds)

Poor data quality

- Most text data are generated by humans – Prone to data entry
- Text data contains abbreviations (MoM, IMHO), short forms, poor grammar
- Mix of language – Regional language typed with alphabets
- Casual language – Fillers, Idioms etc
- Encoding issues when applications load text files

Feed me just numbers…

- Machines can understand only numbers
- Text data -> Structured data
- Possibility of missing the context during this process

High Dimension Data

- Conversion to structured data creates high dimension data
- Requires high end machines to process medium sized data
- Takes lot of time to train a model
- Performance of models comes down
- Redundant data – People write elaborate sentences

Unstructured Data Analysis

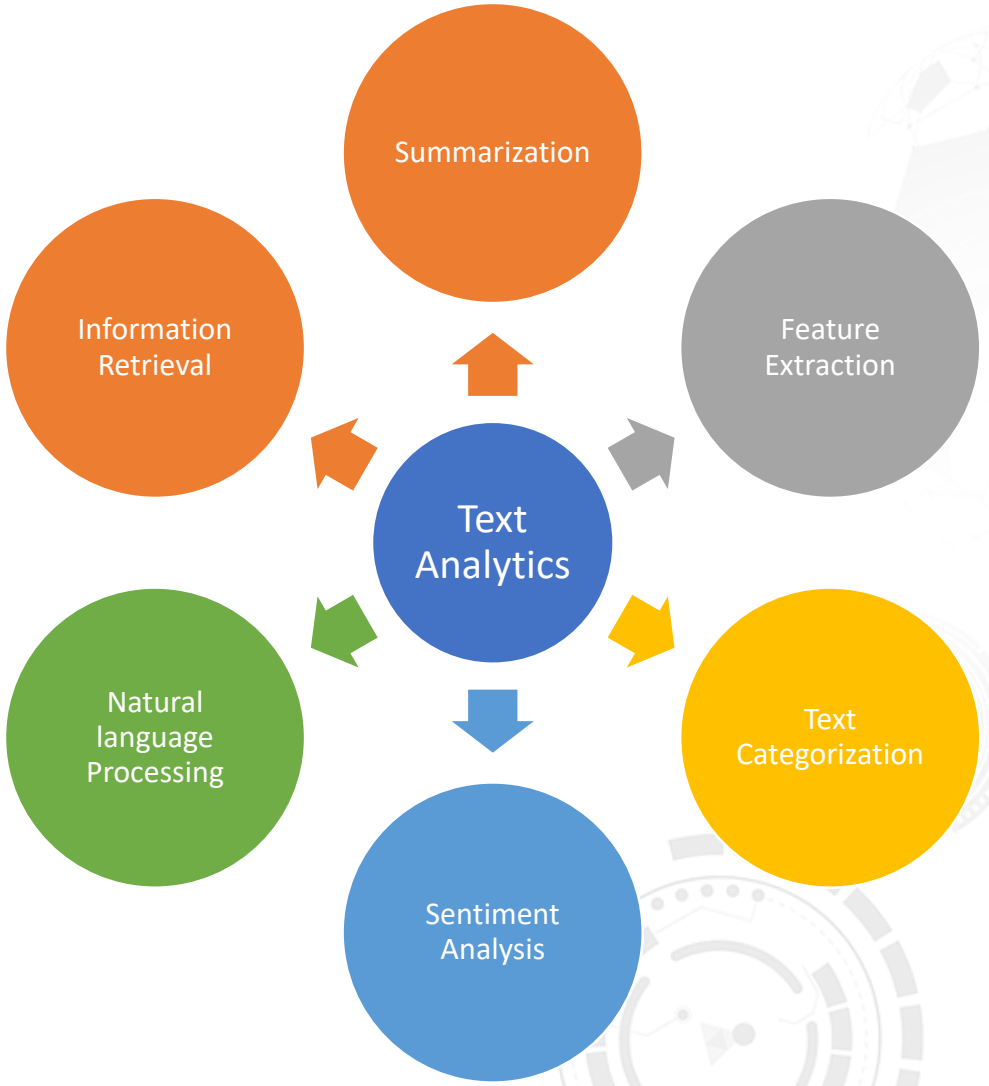
Applications

By

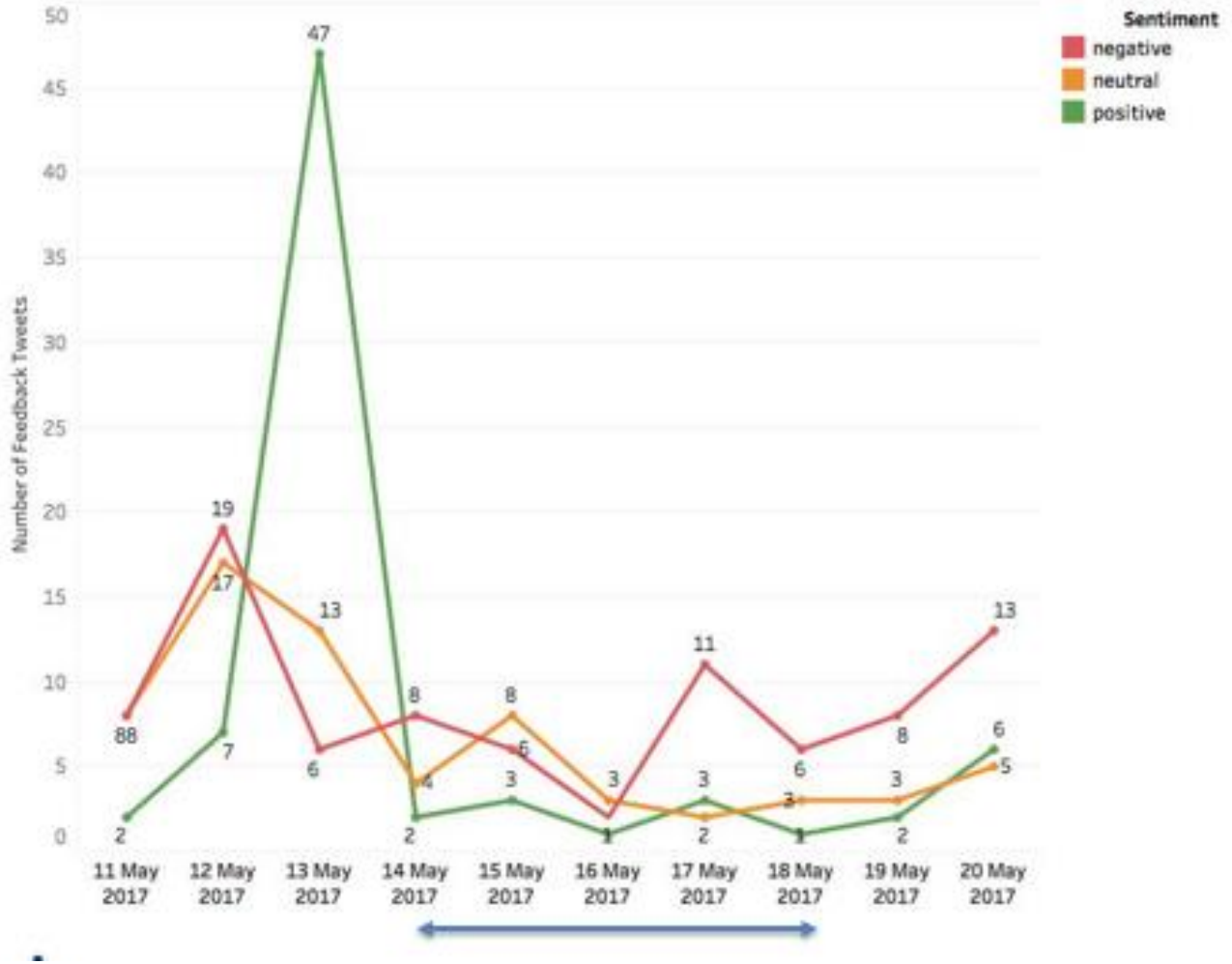
Kathirmani Sukumar

Text mining techniques

- Sentiment Analysis
  - Customer feedback & emotion
- Categorization
  - Redirecting complaints to concerned department
  - Organize documents
- Feature extraction
  - Extract phone, email ids, address etc
  - Identify themes, Identify relationships
- Text Summarization
  - One line news description
- Natural Language processing
  - Chat bots, personal assistant etc



High hopes before big sales day



Extract Topics

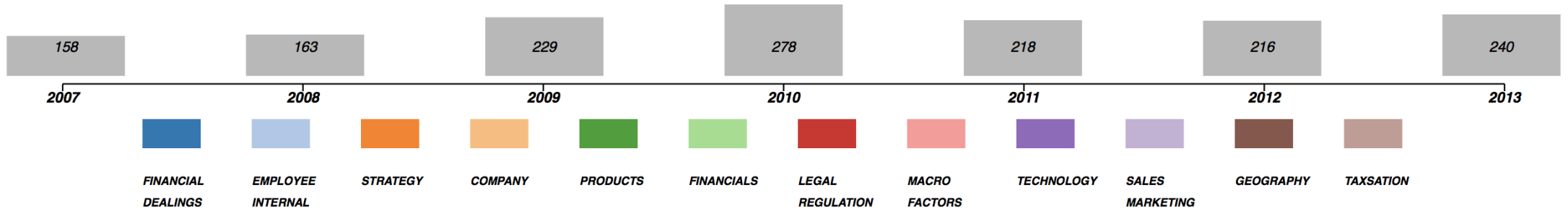
- The first group contains words related to tablet (ipad, android, app, device etc). Hence we can label it as ***Tablets***
- The second group contains words related to e-reader (nook, kindly, books, touch, page etc). Hence we can label it as ***E-Reader***
- The third group contains words related to battery problems. We can name it as "***Batteries***"
- The fourth group contain words related to wall mounting. We can name it as "**Wall mounting**"

Transcripts Analysis

Company: Goldman Sachs

Insight: Discussion over BASEL increased post 2008 Crisis *(Mouse-over)*

QUESTIONS ASKED



Source: Gramener.com

Document Similarity

- Fake news detection
- Identify plagiarism
- Recommend books
- Organize articles and books