


Intro to Text-mining: a content analysis method

Dr. J. Kasmire
Research Fellow at Cathie Marsh Institute and
UK Data Service

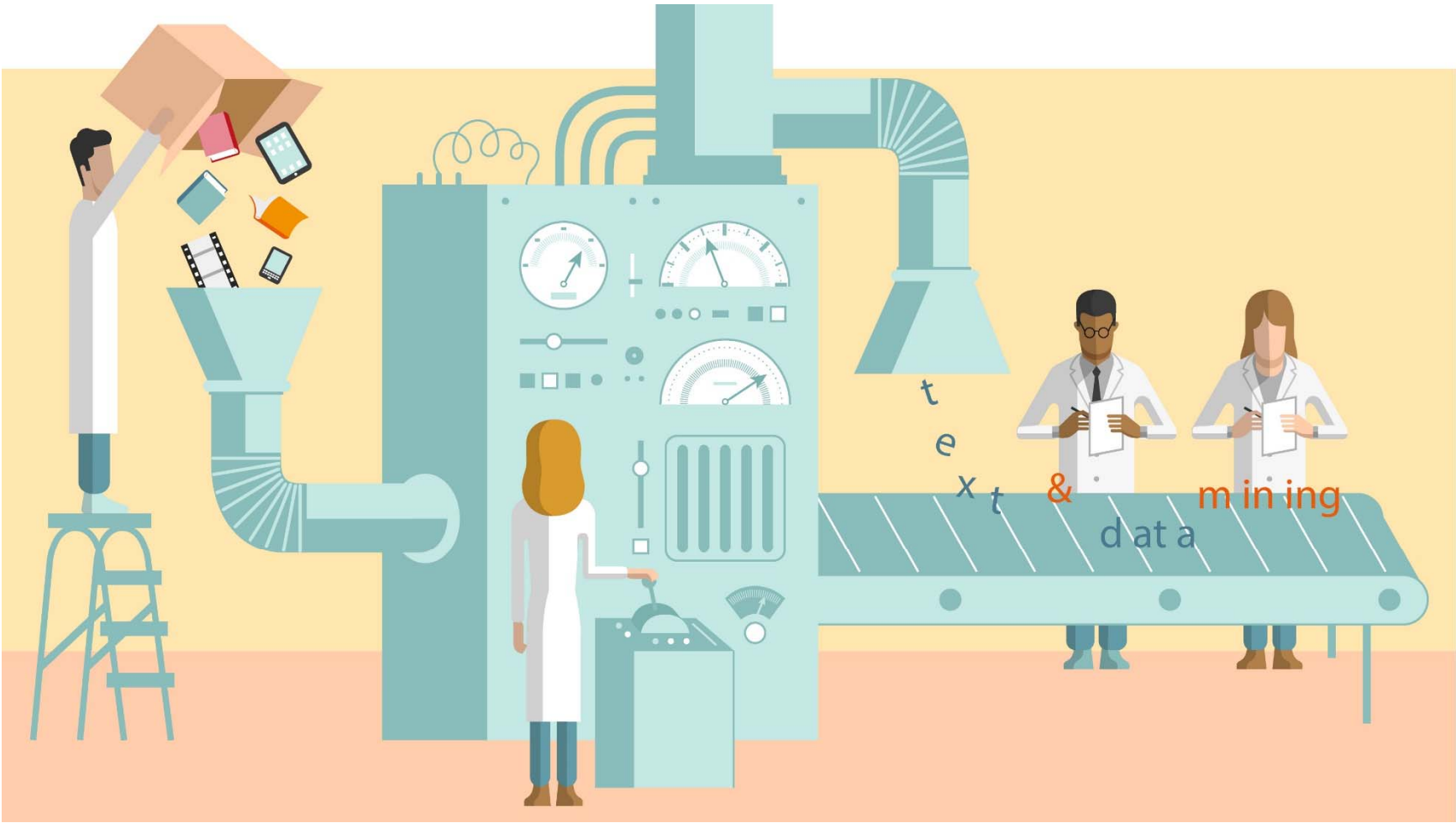


julia.kasmire@manchester.ac.uk

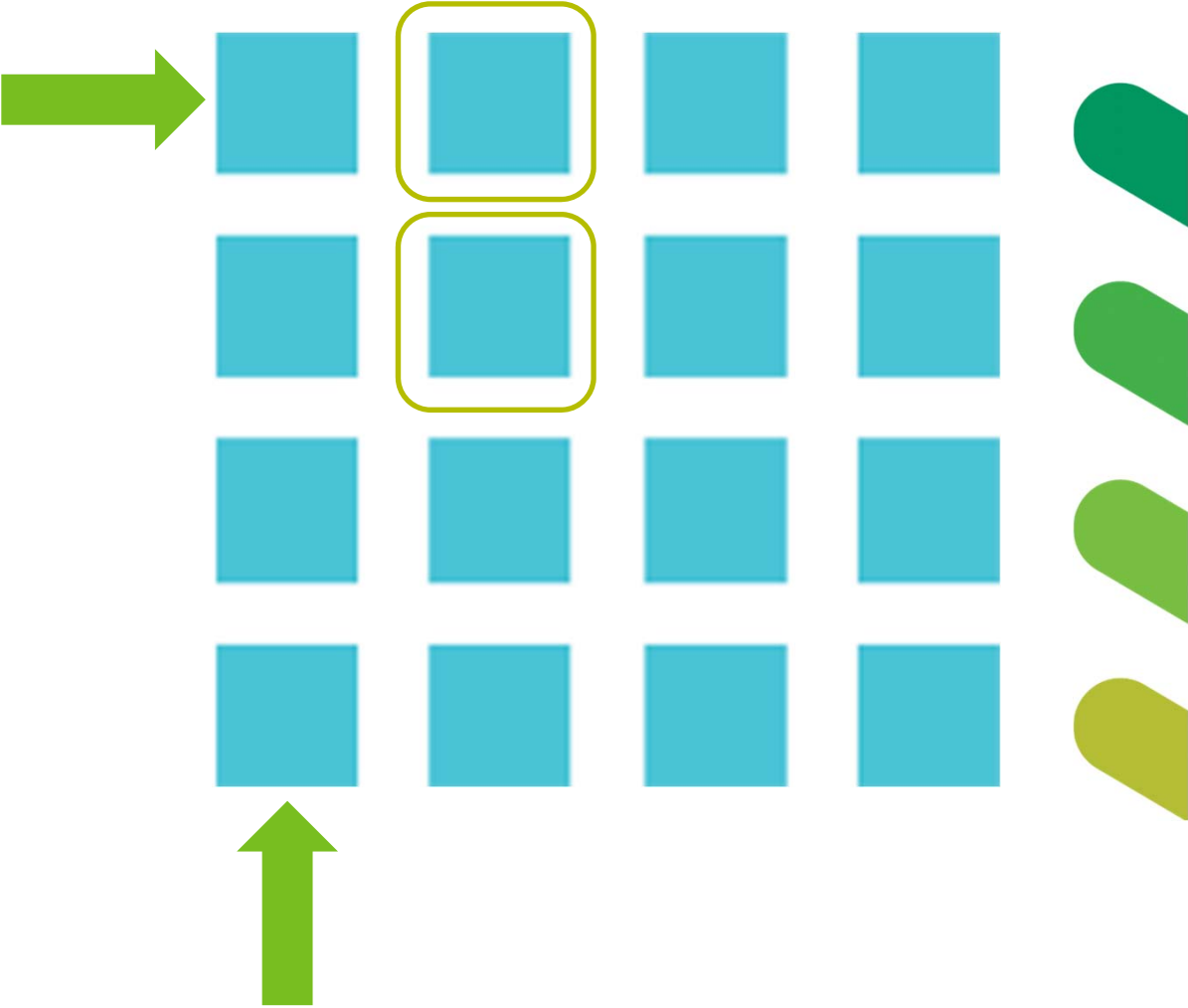
 @JKasmireComplex



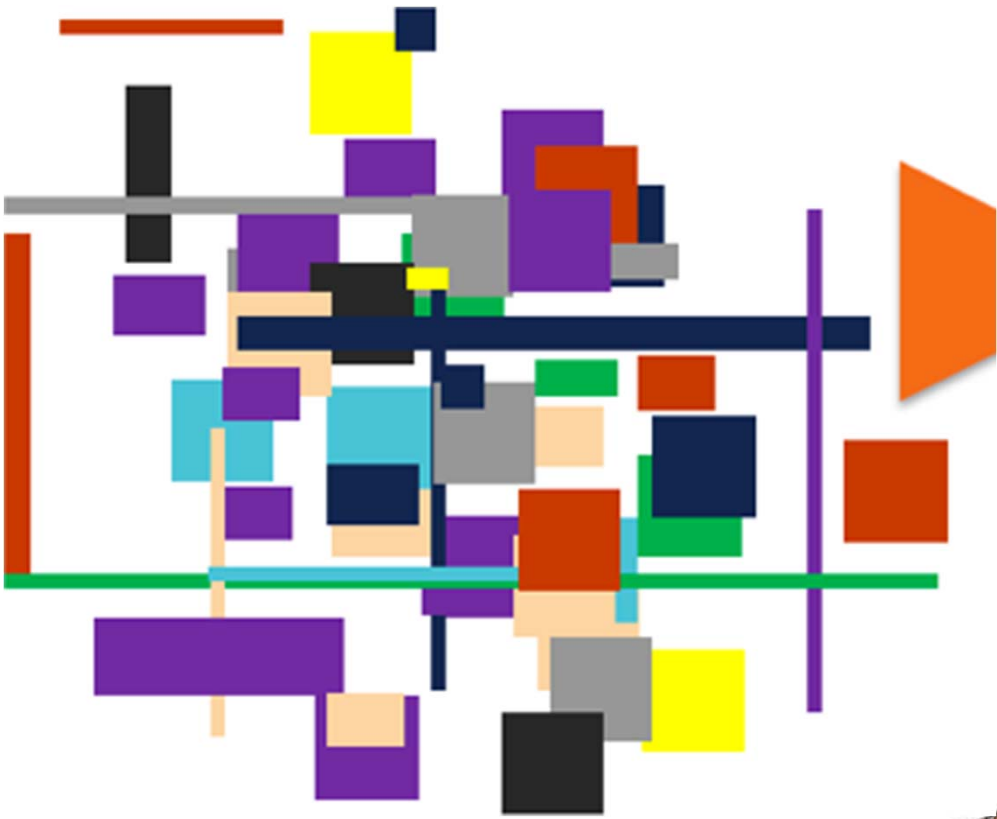
Text-mining is a kind of data-mining



What do I mean by structured data?



So what is unstructured or semi-structured data?



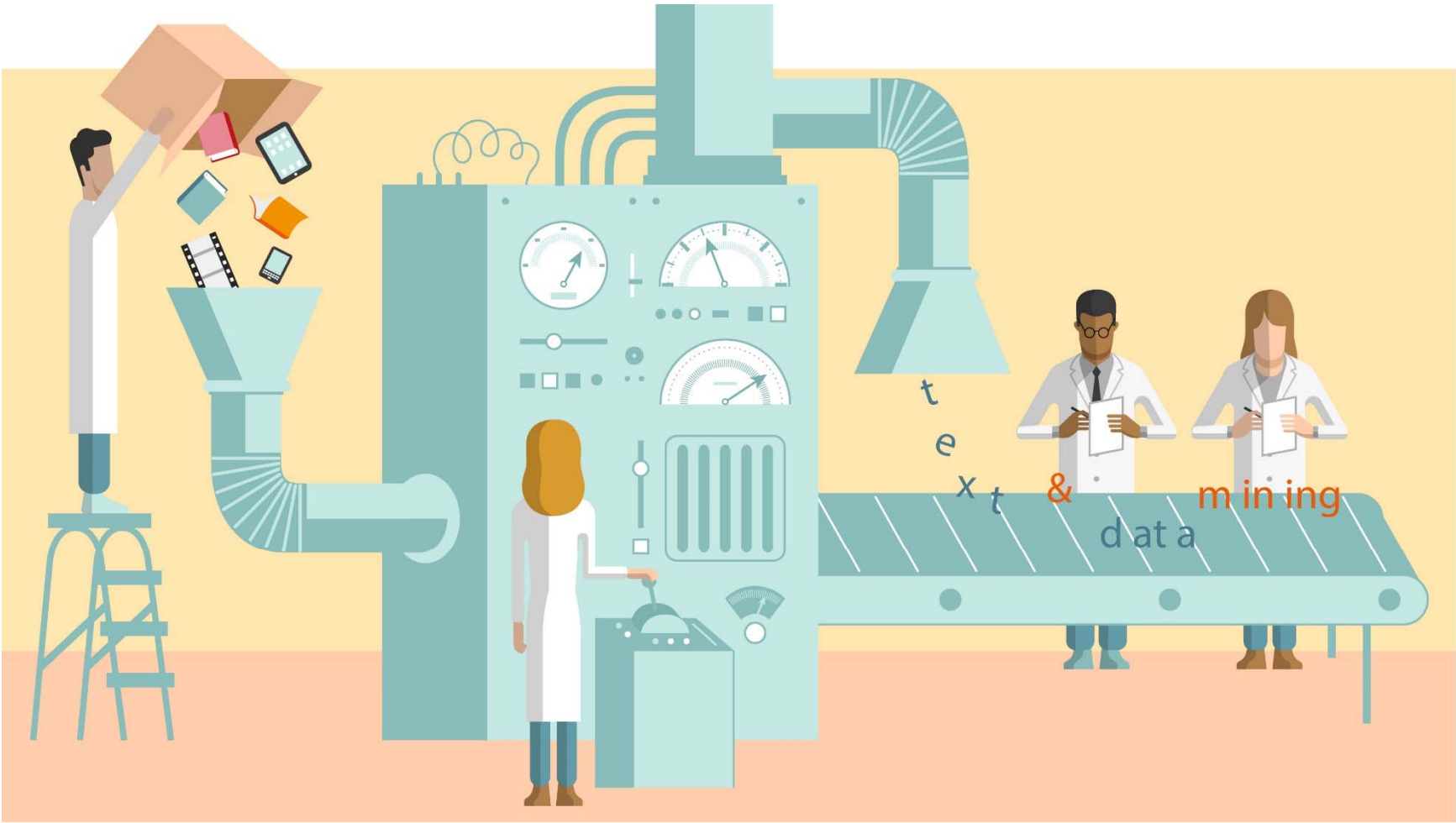
March 17 17/03
St. Patrick's Day 03-17
International Children's Day



Yeah... and?



Solution? Capture and reinforce the structure



Text-mining in four steps

1. Retrieval
2. Processing
3. Extraction
4. Analysis



Text-mining in four steps

1. Retrieval
2. Processing
3. Extraction
4. Analysis

SEARCH

Source = MANCHESTER EVENING NEWS

Date = 01/01/19700 to 31/12/2019

Keywords = “rail” AND “electrification” AND
“north” AND “England”



Text-mining in four steps

1. Retrieval
2. Processing
3. Extraction
4. Analysis



Text-mining in four steps

1. Retrieval
2. Processing
3. Extraction
4. Analysis

One big file file with all results - - - > one file, row or db entry per result

Basic NLP – correct spelling,
remove capitalisation
substitute acronyms or alternate references

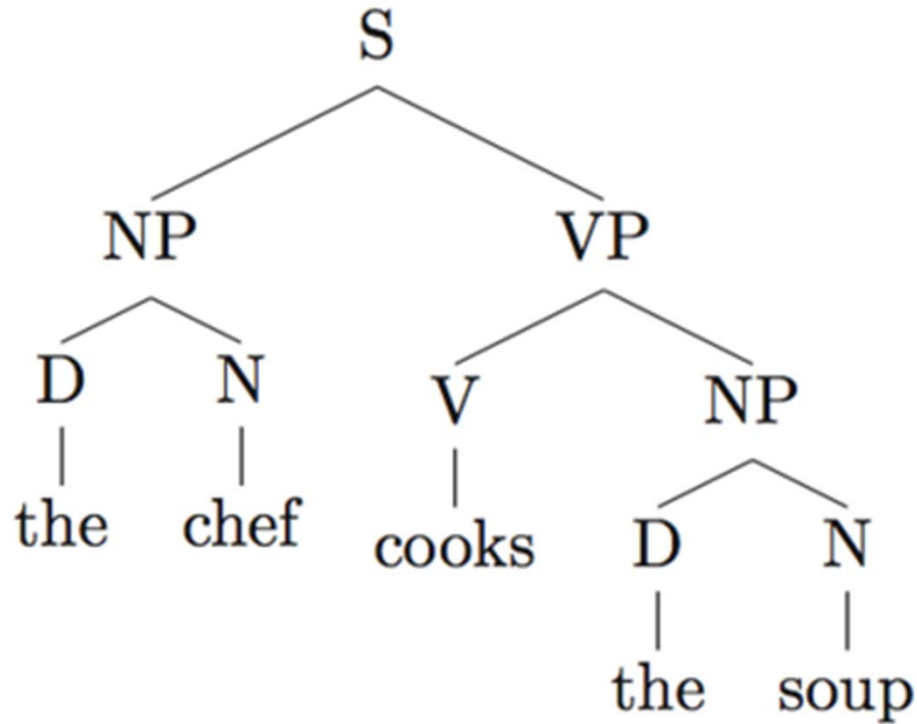
More NLP – classify words by grammatical category
disambiguate meaning by context
parse sentences and mark up structure

More NLP – classify words by grammatical category
disambiguate meaning by context
parse sentences and mark up structure

Text-mining in four steps

- 1. Retrieval
- 2. Processing
- 3. Extraction
- 4. Analysis

The chef cooks teh soup.



[S: [NP: [D: the] [N:chef]] [VP: [V: cook (singular, present) [NP: [D: the] [N:soup]]]]]

Text-mining in four steps

1. Retrieval
2. Processing
3. **Extraction**
4. Analysis



Text-mining in four steps

1. Retrieval
2. Processing (Relative) word counts
3. Extraction Equivalency suggestions
4. Analysis Relationship discovery
Timeline creation



Text-mining in four steps

1. Retrieval
2. Processing
3. Extraction
4. Analysis



Text-mining – One simple example

1. Retrieval
Download 10 days of tweets from 20 different users, also download trending hashtags for those 10 days
2. Processing
Remove everything that is not a hashtag
3. Extraction
Compare to hashtags from tweets to hashtags from trending list
4. Analysis
Calculate a “trendiness score” for each user from degree of match between their own hashtags and trending hashtags



Text-mining – A complex example (of mine)

1. Retrieval Download UK news articles with keywords like “Manchester” AND “commonwealth games”
2. Processing Articles -> sentences -> tokens -> custom processes that match proper nouns, dates, known structures and relationships, etc.
3. Extraction Compare extracted and processed tokens to identify events and the temporal relationships between them
4. Analysis Creates a timeline of events
Performance score against human analyst and state of the art AI



Text-mining - Applications

- Sentiment analyses
- Compare documents for author/style/etc.
- Change over time
- Automated systems
- Predictive modelling



Text-mining Pros and Cons

- Pros:
 - Large scale approach to difficult stuff
 - Can see detail of sub-groups
 - Novel application
- Cons:
 - Needs a large corpus
 - May need a lot of manually created training data
 - Lack of human interaction or supervision
 - Unclear what questions it can/cannot address
 - No creation of new techniques



Questions

Dr. J. Kasmire

julia.kasmire@manchester.ac.uk

 @JKasmireComplex

UKDS

 @UKDataService

 UKDataService

