

Text-Mining: Introduction and theory

Dr. J. Kasmire

Research Fellow at Cathie Marsh Institute and UK Data Service



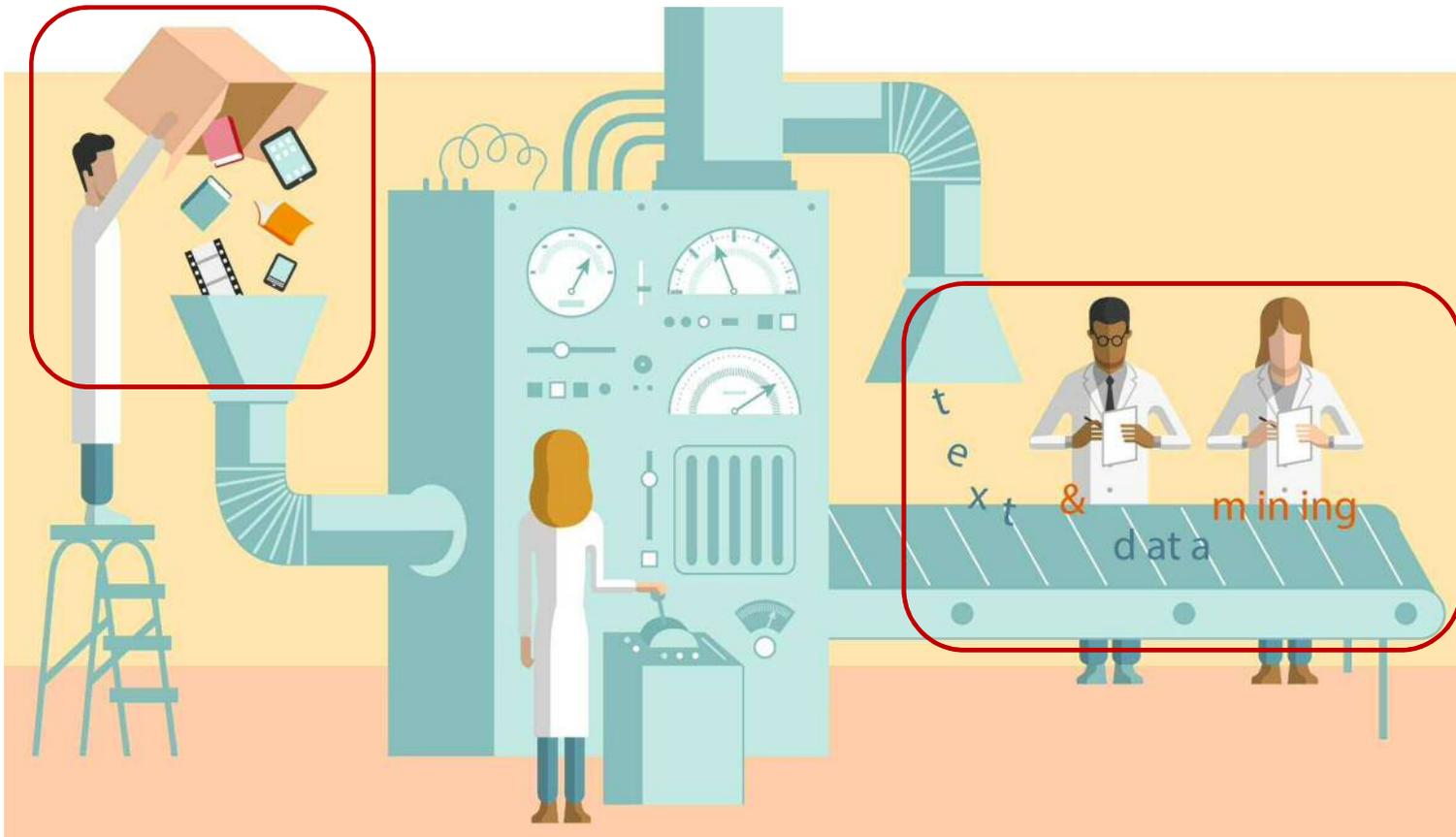
julia.kasmire@manchester.ac.uk



@JKasmireComplex



Text-mining is a form of data-mining



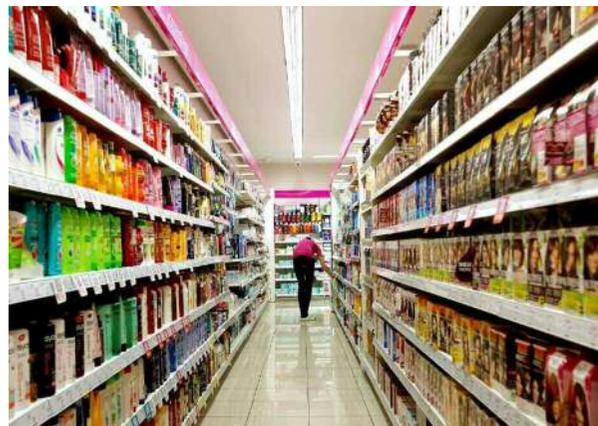
What do I mean by structured data?

Intro to
Text-Mining
02/06/20
75 attended

What do I mean by structured data?

Intro to Text-Mining	02/06/20	75
Text-Mining Basics	16/06/20	72
Text-Mining Advanced	29/06/20	85

Think about commonly found examples of structured data?



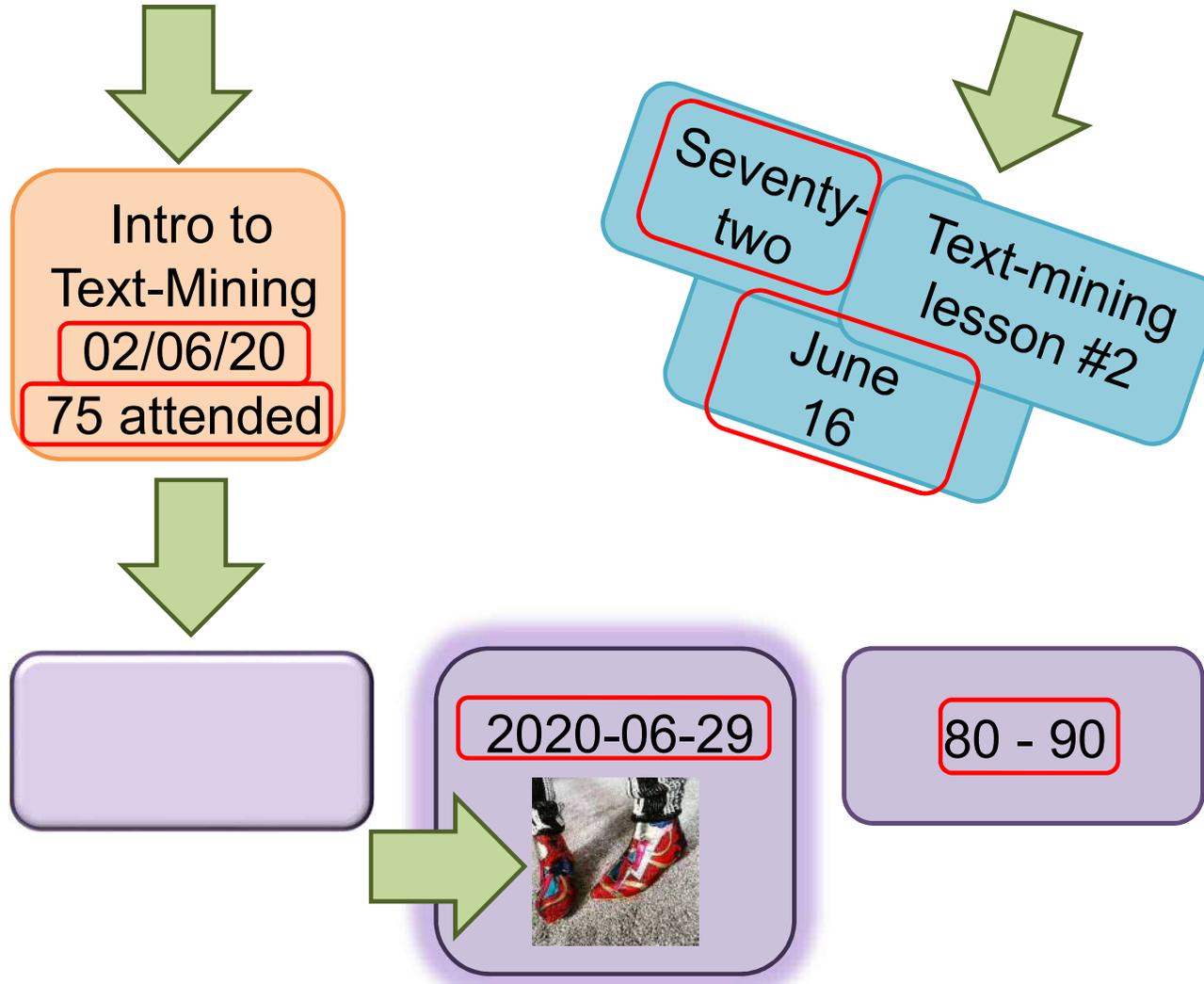
Structured data is...



Intro to Text-Mining	01/05/20	75
4 steps of Text-Mining	04/06/20	72
Text-Mining analysis	09/07/20	85

...familiar
...easy
...demonstrable

What about unstructured data?



What about semi-structured data?

Intro to
Text-Mining
02/06/20
75 attended

Seventy-
two
Text-mining
lesson #2
June 16

2020-06-29


80 - 90

Semi-(un)structured data is ...

Intro to Text-Mining
02/06/20
75 attended

Seventy-two
June 16
Text-mining
lesson #2

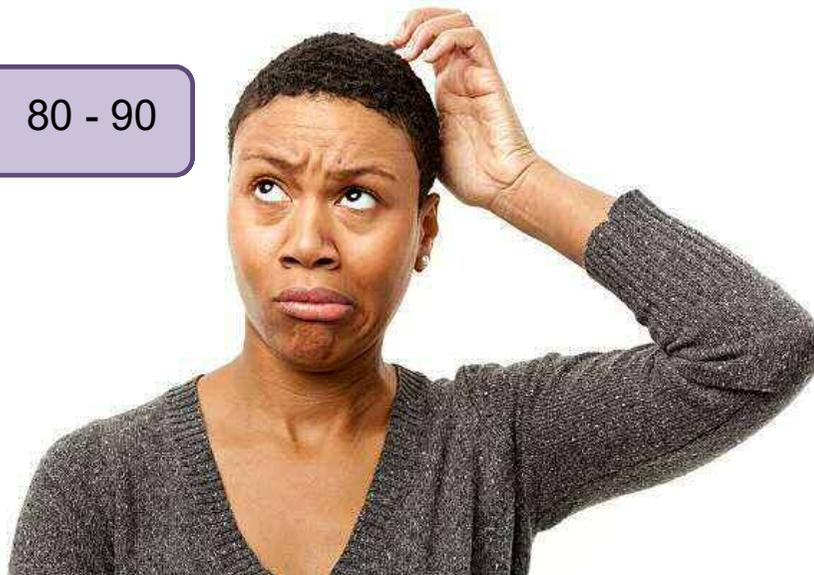
... less accessible
... difficult
... requires intuition
and “common sense”



2020-06-29

A small image showing a pair of red shoes with white patterns on a carpeted floor.

80 - 90



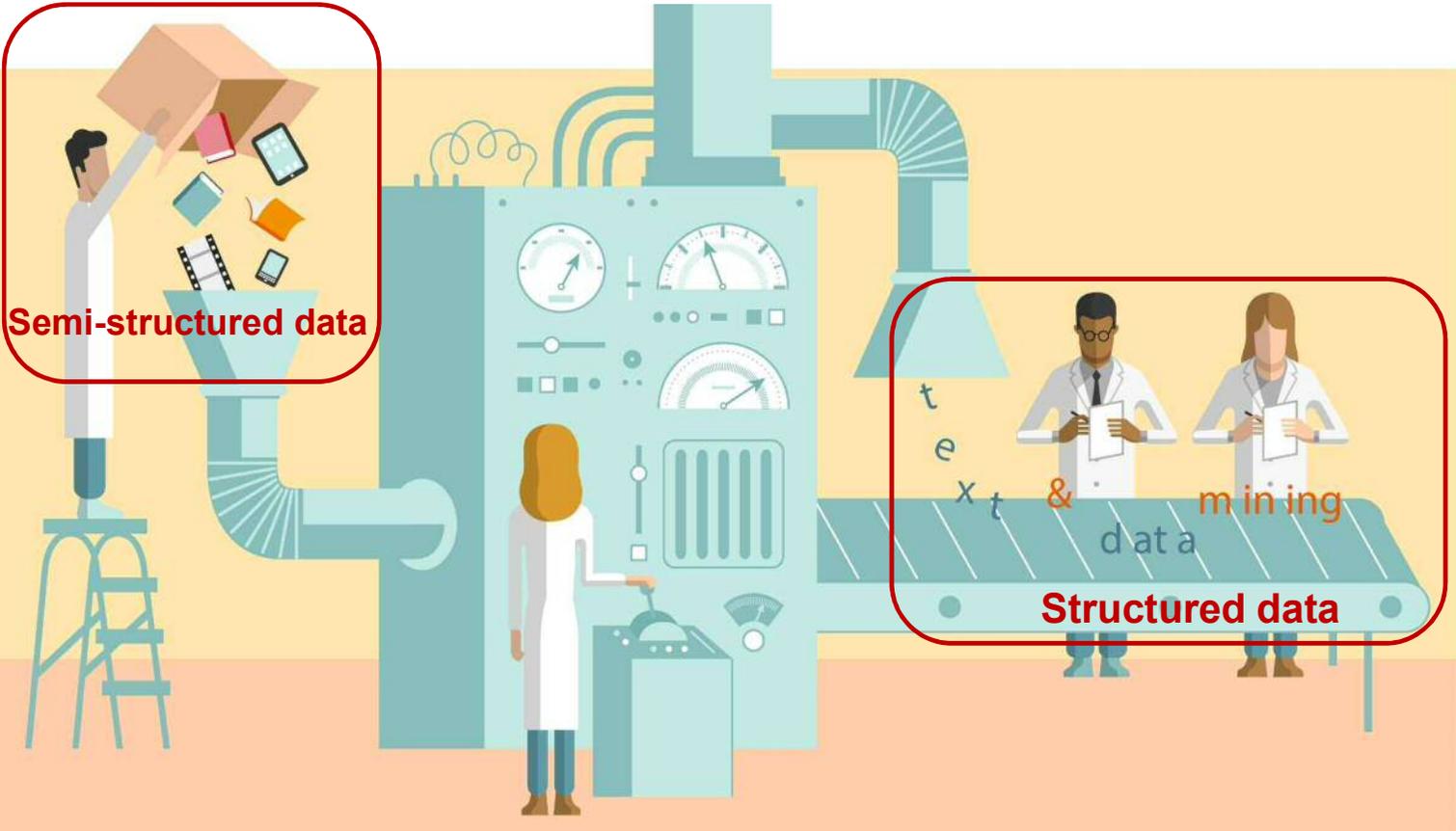
Yeah... and?



Why can't you just treat semi-structured and structured data the same?

- The tools won't work.
- Forcing the tools to “sort of” work can be a real pain and waste of time!
- Documenting the process is very difficult which makes for
 - Poor replicability
 - Hard to understand methods
 - Hard to visualise results
- First, you need to turn semi-structured data into structured data.
- There are tools to help you do that.

What are those tools?



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



Raw data - - - > 1 file/row/database entry per tweet/document/webpage

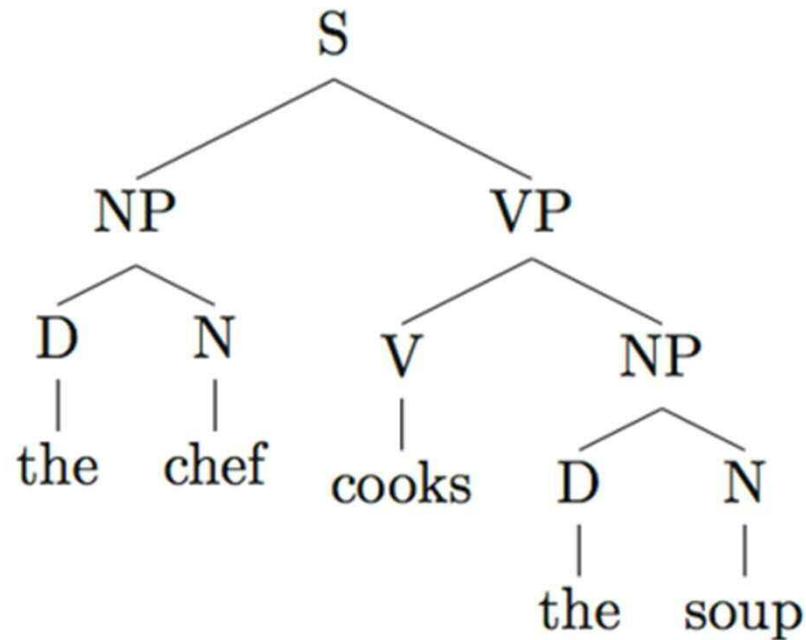
Basic NLP – correct spelling,
remove capitalisation,
substitute acronyms or alternate references

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

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



(Relative) word counts
Equivalency suggestions
Relationship discovery
Automatic categorisation
Prediction

Text-mining – One simple example

1. Retrieval
 - Download 10 days of tweets from 20 users.
 - Download trending hashtags for those same 10 days
1. Processing
 - Remove everything that isn't a hashtag (punctuation, trailing whitespace)
 - Store individual hashtags in data frame labelled by date and author
2. Extraction
 - Compare tweeted hashtags to trending list – by time, by volume, etc.
 - Calculate a “trendiness score” for authors based on degree of match and timing
4. Insight
 - Explain what a trendiness score measures –
Influencer status? Finger-on-the-pulse-ness?
Tendency to jump on bandwagons? Something else?

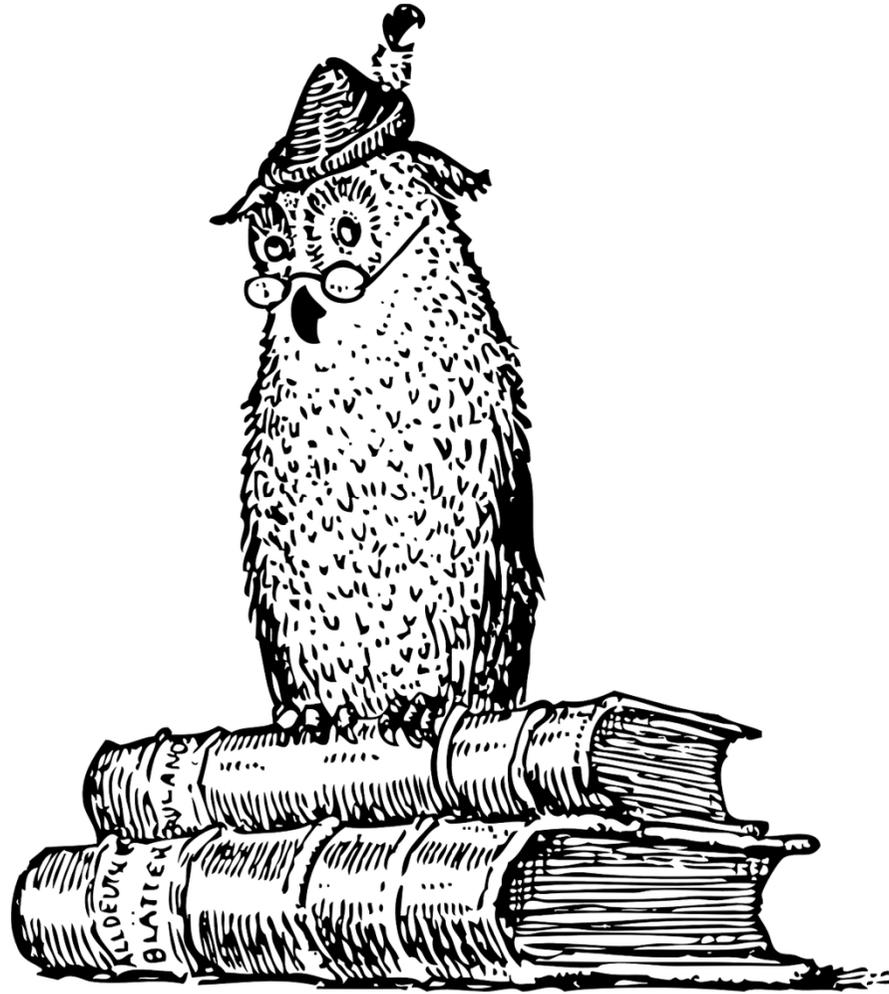
Text-mining – A complex example (of mine)

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
Create a timeline of events
Performance score against human analyst and state of the art AI
4. Insight Argue how automated event extraction and time line creation supports policies of event-based investment and regeneration

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
 - Lops off a bunch of structure or information that is hard to capture/amplify

Text-mining can't (yet) provide expert level insight

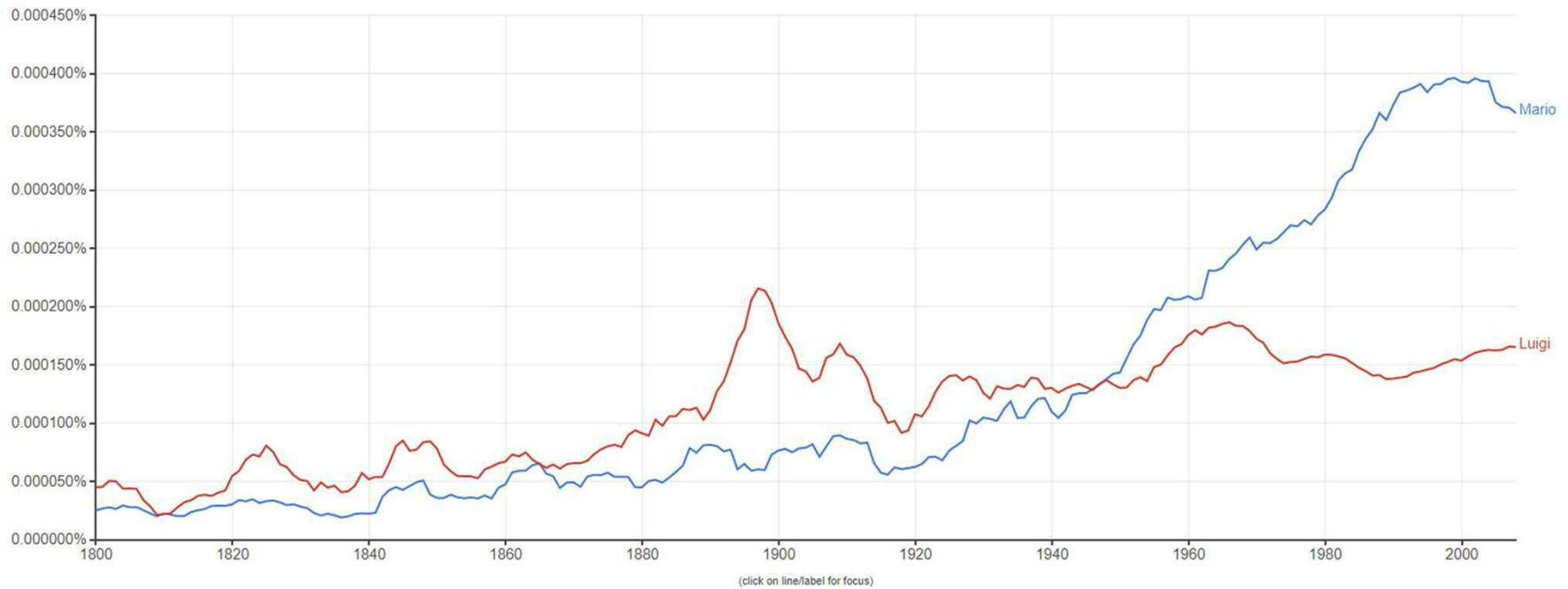


But it text-mining does...

Google Books Ngram Viewer

Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of [Search lots of books](#)



Citations and reading recommendations

Predicting the Present with Google Trends. Choi and Varian, 2012.
<https://doi.org/10.1111/j.1475-4932.2012.00809.x>

Programming with Python for Social Scientists. Brooker, 2020.
<https://study.sagepub.com/brooker>

Automate the Boring Stuff with Python: Practical Programming for Total Beginners, Sweigart, 2019. ISBN-13: 9781593279929



Questions

Dr. J. Kasmire

julia.kasmire@manchester.ac.uk

 @JKasmireComplex

UKDS

 @UKDataService

 UKDataService

