UK Data Service

Text-Mining: Introduction and theory

The webinar will begin at 1 pm

You now have a menu in the top right corner of your screen.

The red button with a white arrow allows you to expand and contract the webinar menu, in which you can write questions/comments.

Feel free to type questions as we go, we will answer as many as we can at the end

We can’t hear you.

Copyright © 2020 UK Data Service. Created by New Forms of Data Training, UK Data Service

cc BY-NC-SA

UK Data Service

Can you hear us?

Copyright © 2020 UK Data Service. Created by New Forms of Data Training, UK Data Service

cc BY-NC-SA

UK Data Service

Can you hear us?

If not:

- Check your speaker/headset is plugged in / volume is on.
- Click on audio to change to listening via phone
- We are recording this webinar and will post it on YouTube (https://www.youtube.com/user/UKDATASERVICE)

Copyright © 2020 UK Data Service. Created by New Forms of Data Training, UK Data Service

cc BY-NC-SA

UK Data Service

Text-Mining: Introduction and theory

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

julia.kasmire@manchester.ac.uk

@JKasmireComplex

Copyright © 2020 UK Data Service. Created by New Forms of Data Training, UK Data Service

BY-NC-SA

You might be interested in...

Recent -

- Being a Computational Social Scientist
- Web-scraping for Social Science Research (case study, from websites, and from API's)
- Code Demos
- https://www.ukdataservice.ac.uk/news-and-events/events/past-events.aspx
- https://www.youtube.com/user/UKDATASERVICE

Upcoming -

- Text-mining: Basic Processes 16 June 20
- Text-mining: Advance Options 29 June 20
- Health Studies User Conference 30 June 20
- Social Data and the Third Sector 2 to 16 July 20

You might be interested in...

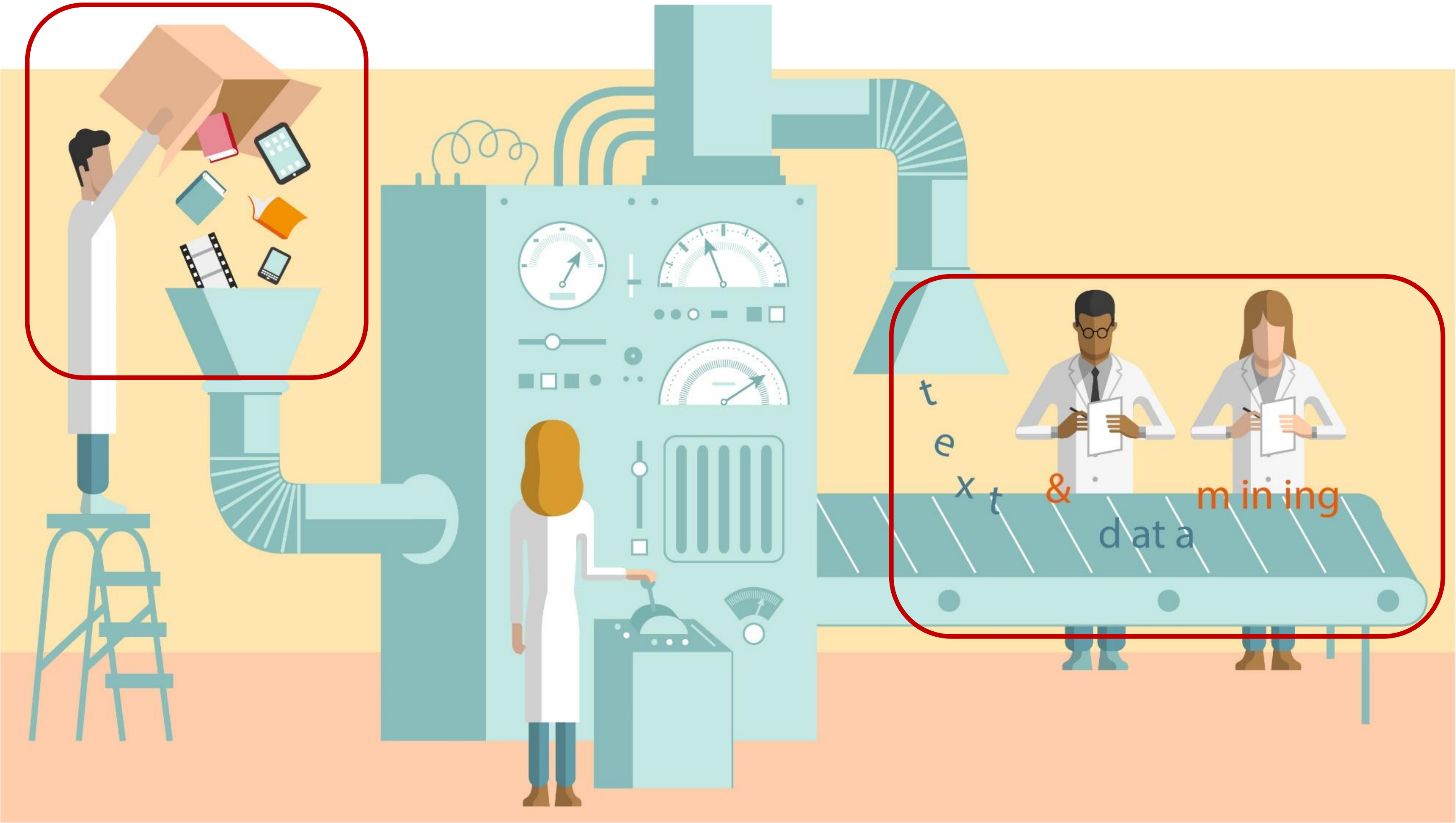
Recent -

- Being a Computational Social Scientist
- Web-scraping for Social Science Research (case study, from websites, and from API's)
- Code Demos
- https://www.ukdataservice.ac.uk/news-and-events/events/past-events.aspx
- https://www.youtube.com/user/UKDATASERVICE

Upcoming -

- Text-mining: Basic Processes 16 June 20
- Text-mining: Advance Options 29 June 20
- Health Studies User Conference 30 June 20
- Social Data and the Third Sector 2 to 16 July 20

Text-mining is a form of data-mining

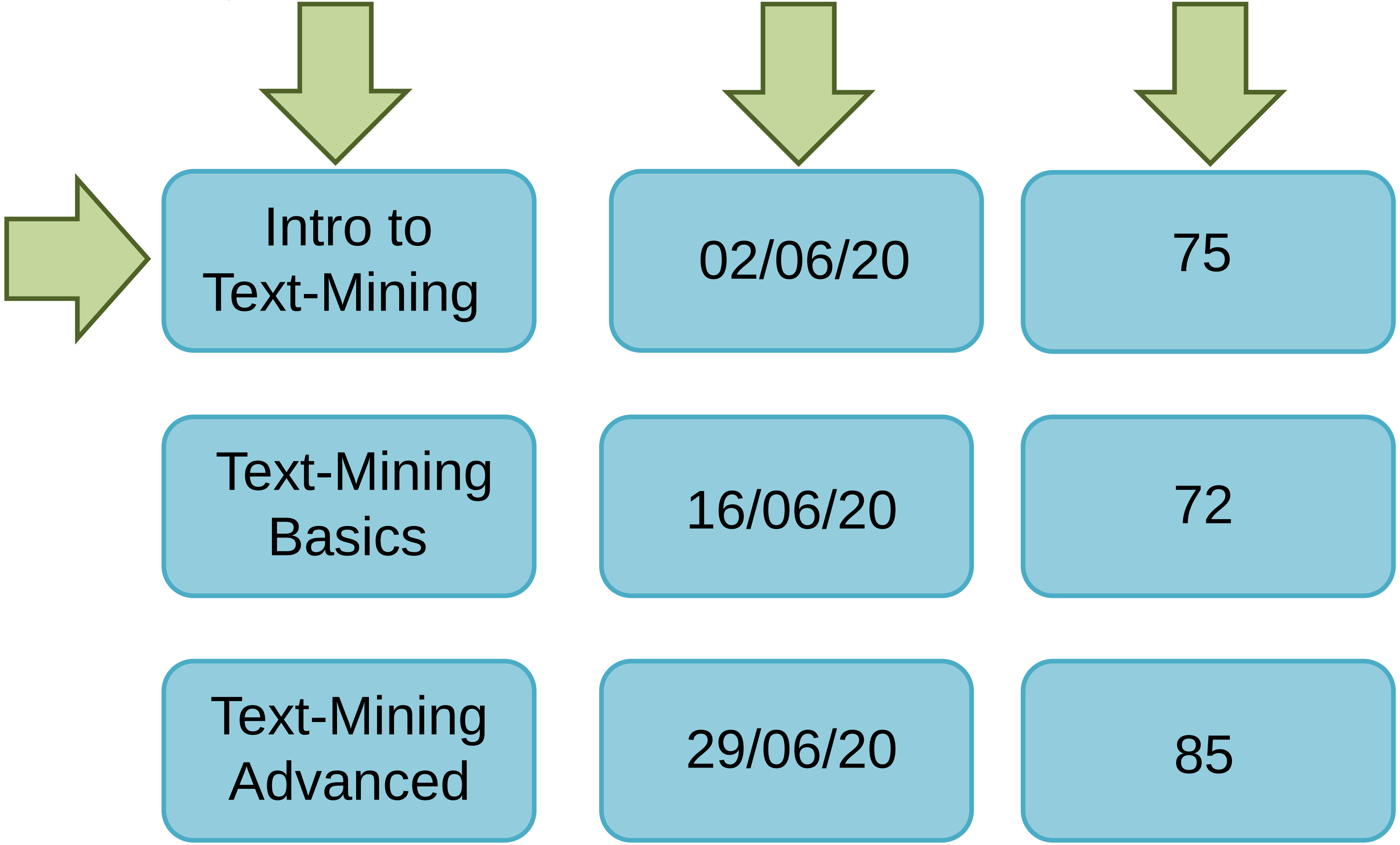


UK Data Service

What do I mean by structured data?

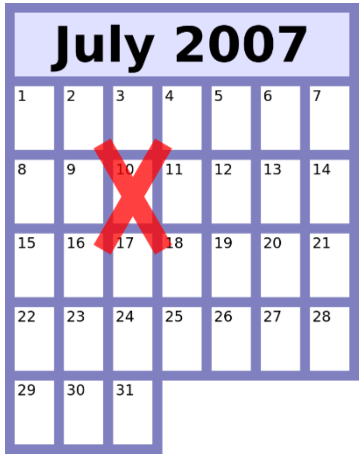
Intro to
Text-Mining
02/06/20
75 attended

What do I mean by structured data?



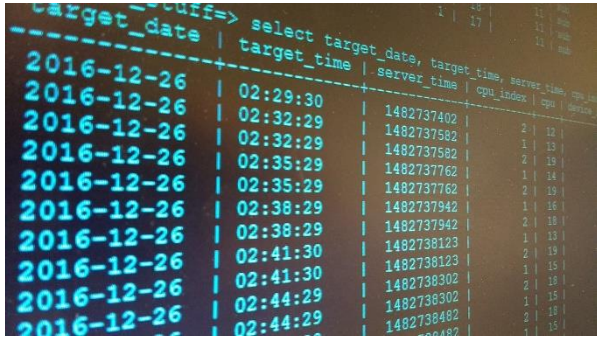
Intro to Text-Mining	02/06/20	75
Text-Mining Basics	16/06/20	72
Text-Mining Advanced	29/06/20	85

Think about commonly found examples of structured data?

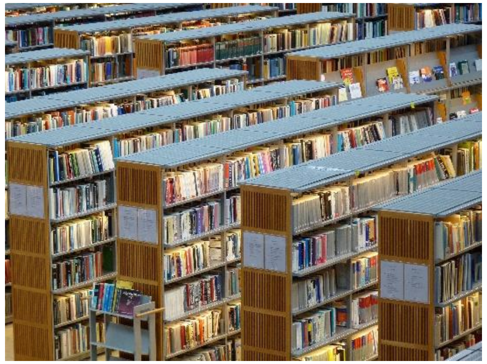


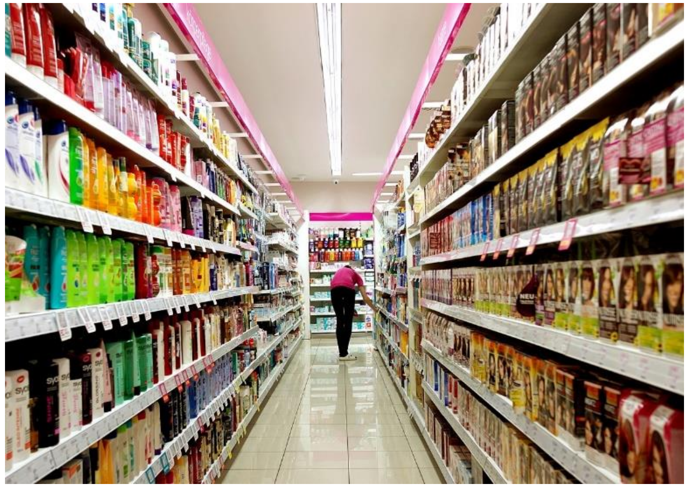
July 2007

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				



```
target_date  | target_time | server_time | cpu_index | cpu | device_
-------------+-------------+-------------+-----------+-----+--------
 2016-12-26  | 02:29:30    | 1482737402  |         2 |  12 |
 2016-12-26  | 02:32:29    | 1482737582  |         1 |  13 |
 2016-12-26  | 02:32:29    | 1482737582  |         2 |  19 |
 2016-12-26  | 02:35:29    | 1482737762  |         1 |  14 |
 2016-12-26  | 02:35:29    | 1482737762  |         2 |  19 |
 2016-12-26  | 02:38:29    | 1482737942  |         1 |  16 |
 2016-12-26  | 02:38:29    | 1482737942  |         2 |  18 |
 2016-12-26  | 02:41:30    | 1482738123  |         1 |  13 |
 2016-12-26  | 02:41:30    | 1482738123  |         2 |  19 |
 2016-12-26  | 02:44:29    | 1482738302  |         1 |  15 |
 2016-12-26  | 02:44:29    | 1482738302  |         2 |  18 |
 2016-12-26  |             | 1482738482  |         1 |  15 |
```





UK Data Service

Structured data is…

© Edward Massias

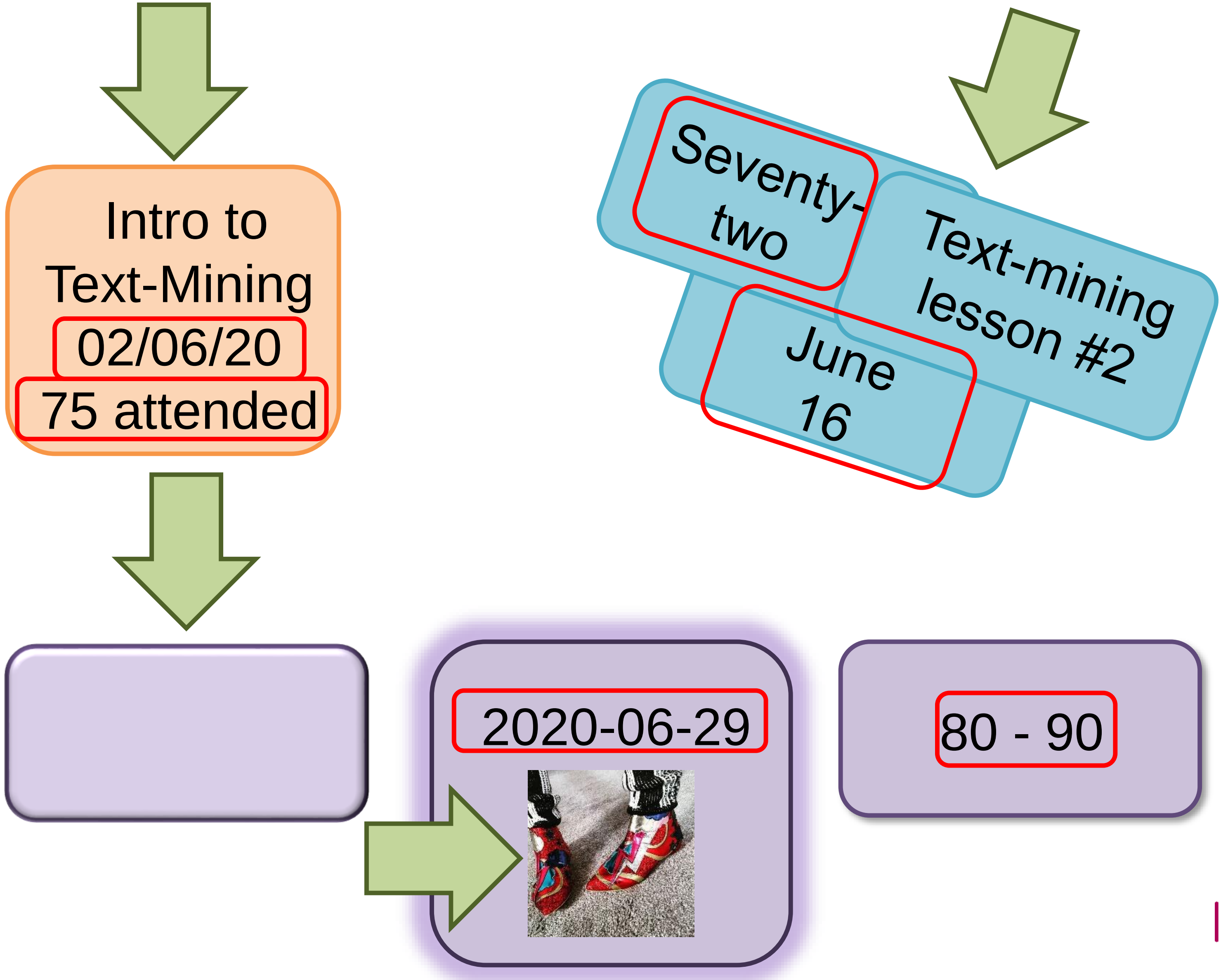
Intro to Text-Mining	01/05/20	75
4 steps of Text-Mining	04/06/20	72
Text-Mining analysis	09/07/20	85

…familiar

…easy

…demonstrable

What about unstructured data?



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

Seventy-two

Text-mining lesson #2

June 16

2020-06-29

80 - 90

What about semi-unstructured data?

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

Seventy-two

June 16

Text-mining lesson #2

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

80 - 90

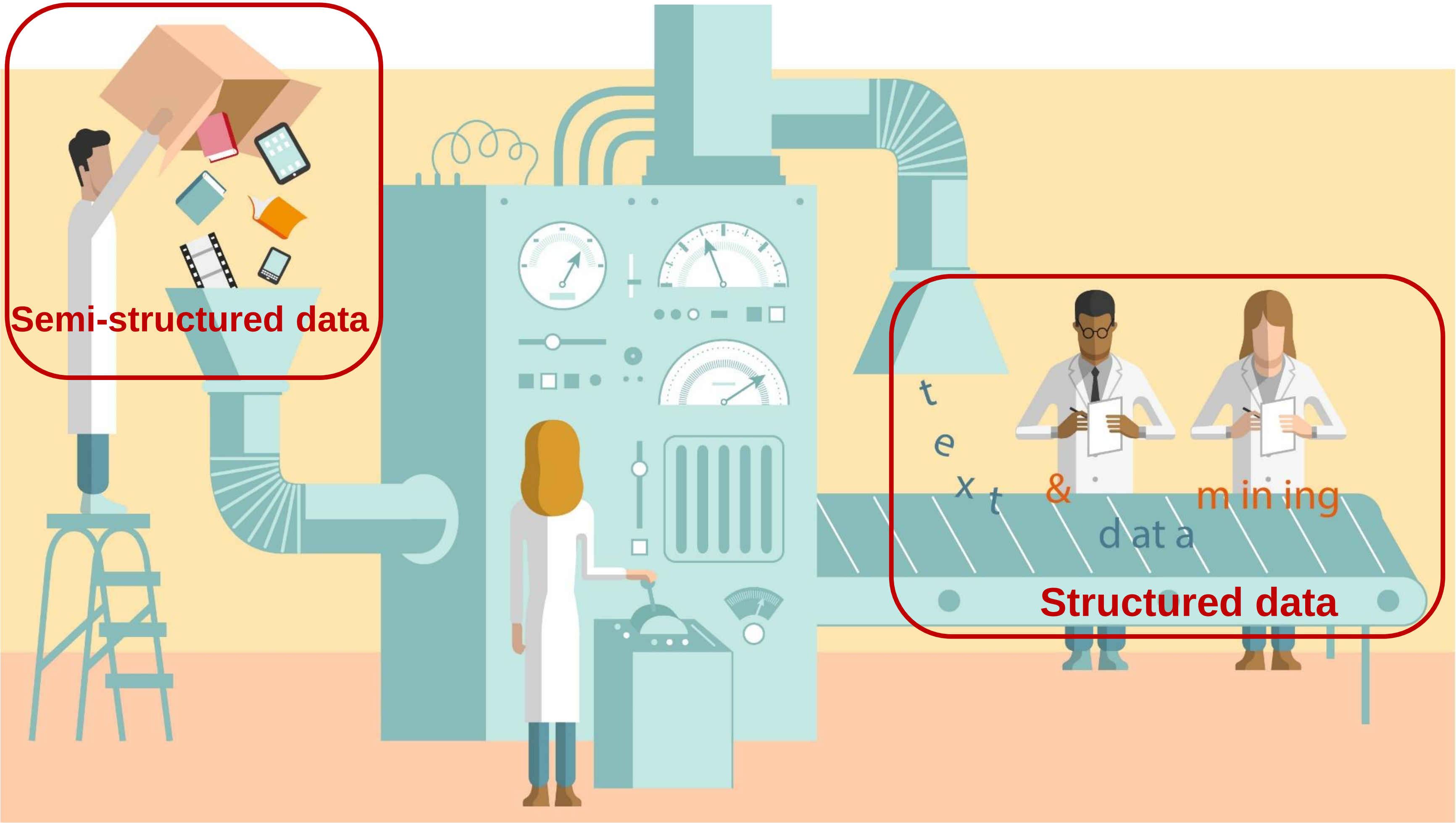
Yeah… and?

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?



UK Data Service

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

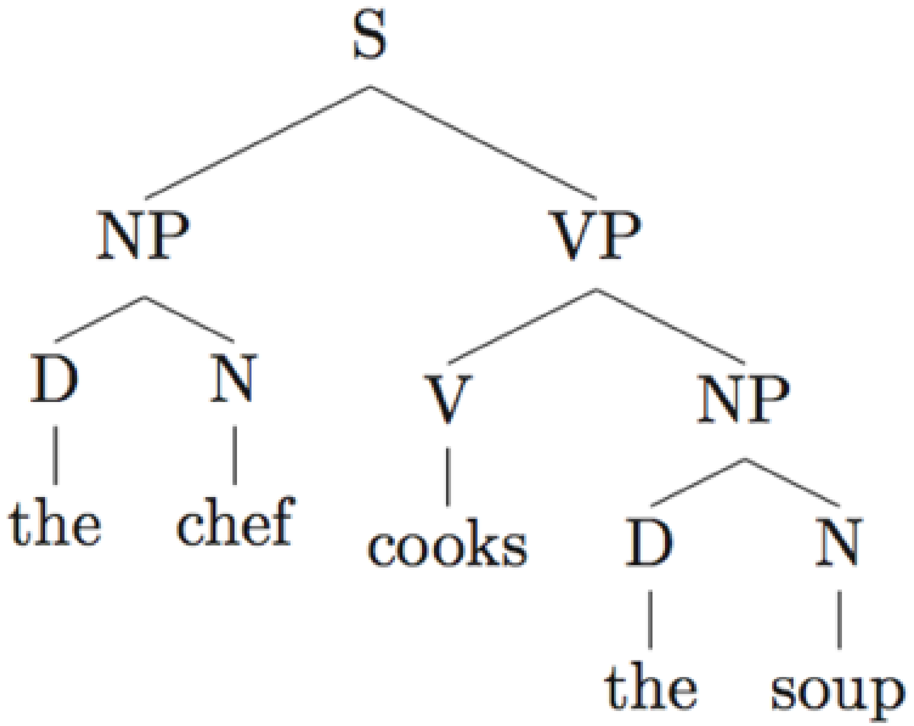
More NLP – classify words by grammatical category,
disambiguate meaning by context,
parse sentences and mark up structure

UK Data Service

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 in four steps

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

PANDERING
EXCELLENT POLITICAL SOCIALISTIC
INTELLIGENT PRESIDENT STRONG CITIZENSHIP
TRUE COUNTRY PROGRESSIVE POWERFUL IMPRESSIVE BS
FASCIST FORWARD
BOLD
CAMPAIGN PRESIDENTIAL AWESOME TRUTH PERFECT
INCLUSIVE SUCKS PASSIONATE
GOING BULL LIAR BRILLIANT PRAGMATIC CONFIDENT
DESERVE DECEPTIVE EMOTIONAL CHALLENGING GOOD PROUD
JOB GREAT INSPIRING HOPEFUL VOTE BEST
LIE IDIOT
PROPAGANDA NARCISSISTIC PARTY
AMBITIOUS AMERICAN BULLSHIT LIES BORING
LEADERSHIP CARING BULLSH
CHANGE
OUTSTANDING POSITIVE COMPASSIONATE THINKING RUBIO
CRAP CITIZEN
TIME BLAH
MILLER
BETTER
JOHNSON
AUTHENTIC
GARBAGE
WATCHING SCARY
FANTASTIC
TELL PEOPLE LEWIS SPEECH
DELUSIONAL RESPECT

UK Data Service

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 analysist 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

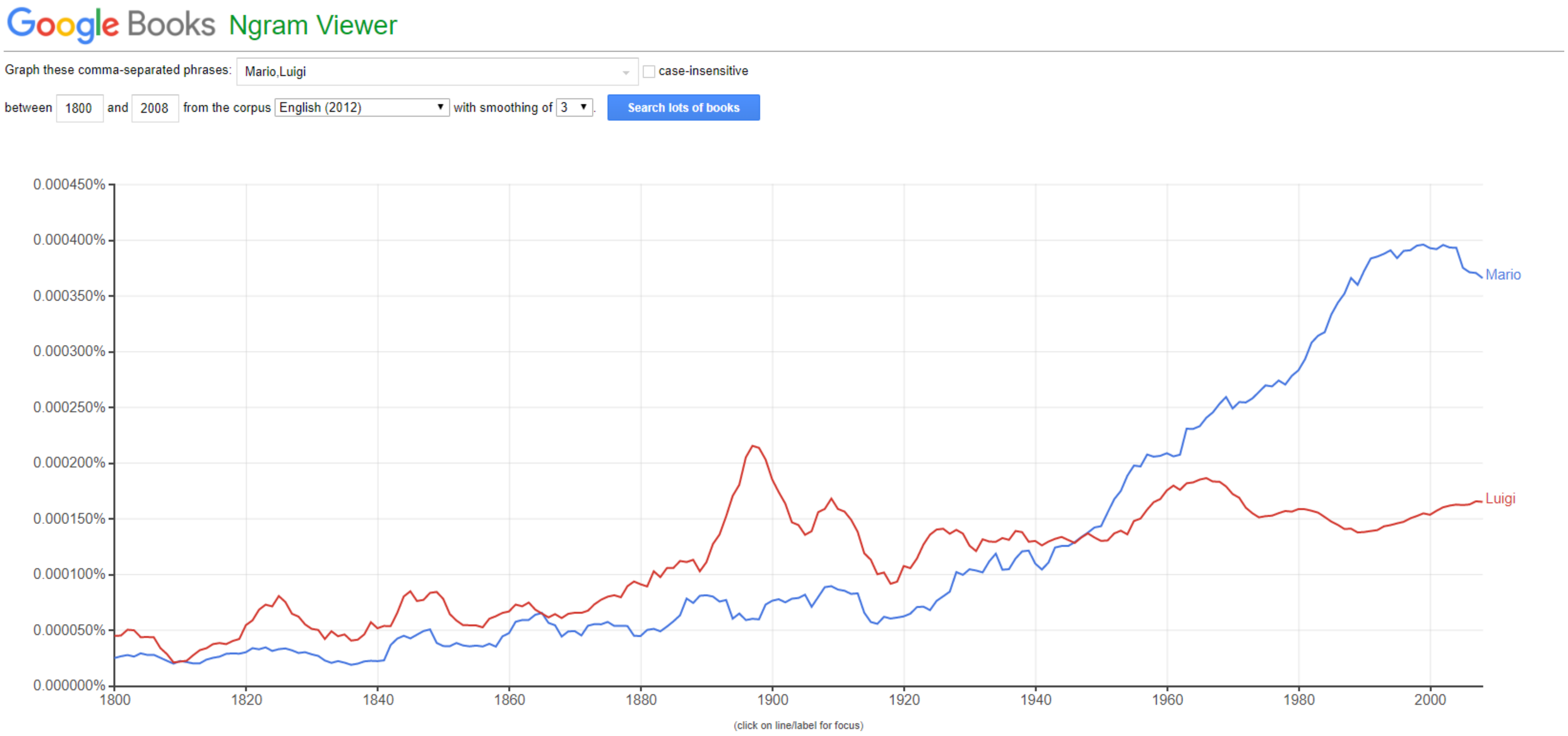
Text-mining Pros and Cons

- 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: Mario,Luigi ☐ case-insensitive

between 1800 and 2008 from the corpus English (2012) with smoothing of 3.

Citations and recommendations

Cited academic paper

- Predicting the Present with Google Trends. Choi and Varian, 2012.

https://doi.org/10.1111/j.1475-4932.2012.00809.x

Recommendations for Python

- 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
- SentDex, python programming tutorials on YouTube https://www.youtube.com/user/sentdex

Recommendations for R

- Quanteda, an R package for text analysis https://quanteda.io/
- Text Mining with R, a free online book https://www.tidytextmining.com/

Questions

Dr. J. Kasmire

julia.kasmire@manchester.ac.uk

@JKasmireComplex

UKDS

@UKDataService

UKDataService