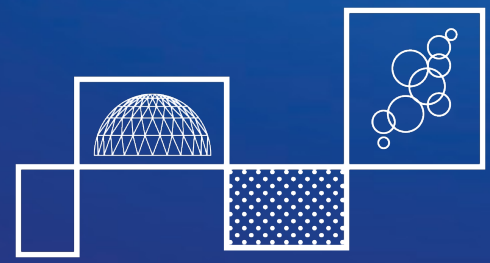
ML and Visualisation in COVID-19 research

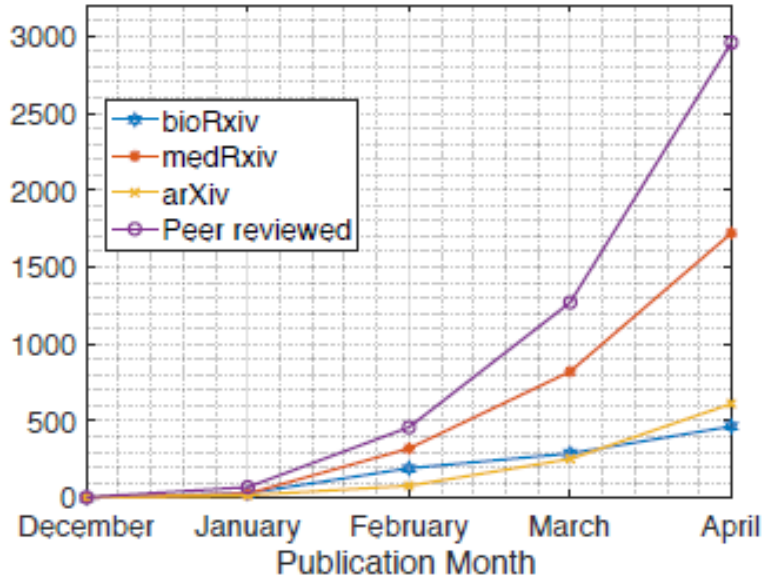
Ian Nabney

University of Bristol



Introduction

- Since December 2019, over 24,000 research papers from peer-reviewed journals as well as sources like medRxiv are available online (April 25th).
- "Data science, defined broadly, will play a central role in the global response to the COVID-19 pandemic."
  - Risk assessment
  - Screening and diagnosis
  - Simulation and modelling
  - Contact tracing
  - Understanding social interventions
  - Logistical planning
  - Automated patient care
  - Supporting vaccine and therapy development

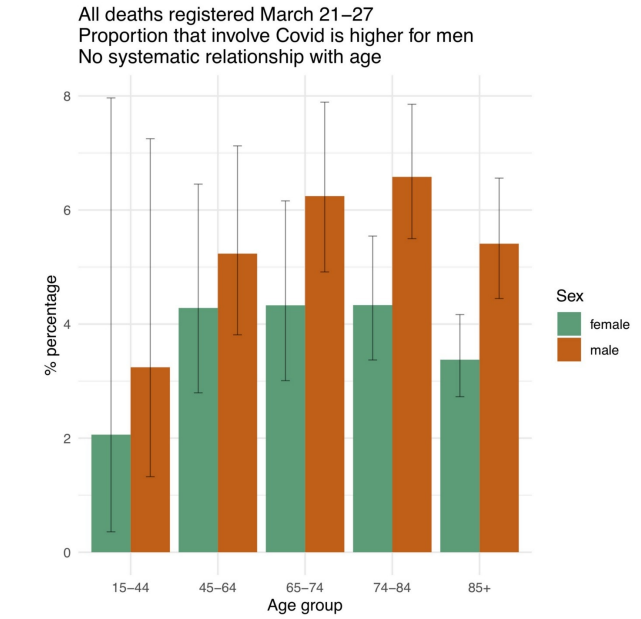


Challenges

- Data availability: for medical images and textual analysis, these datasets are small compared to the requirements of deep learning models. For example, in the case of biomedical data, sample sizes range from a few up to 60 patients.
- The scarcity of measured data is frequently due to the distributed nature of many data sources. For example, electronic healthcare records are often segregated on a national, regional, or even per-hospital level. A key challenge is therefore federating these sources, and overcoming practical differences across each source.
- A key challenge is balancing exigency vs. the need for well-evidenced and reproducible results that can inform policy. Important to capture (and communicate) uncertainty.

Challenges

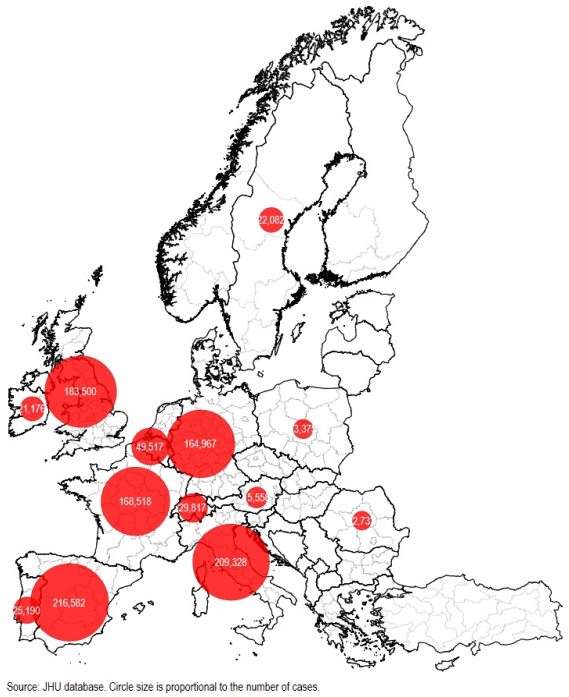
- Balancing exigency vs. the need for well-evidenced and reproducible results that can inform policy. Important to capture (and communicate) uncertainty.
- Security, privacy and ethics; e.g. debate in UK over tracking apps vs. South Korea
- Need for multidisciplinary collaboration



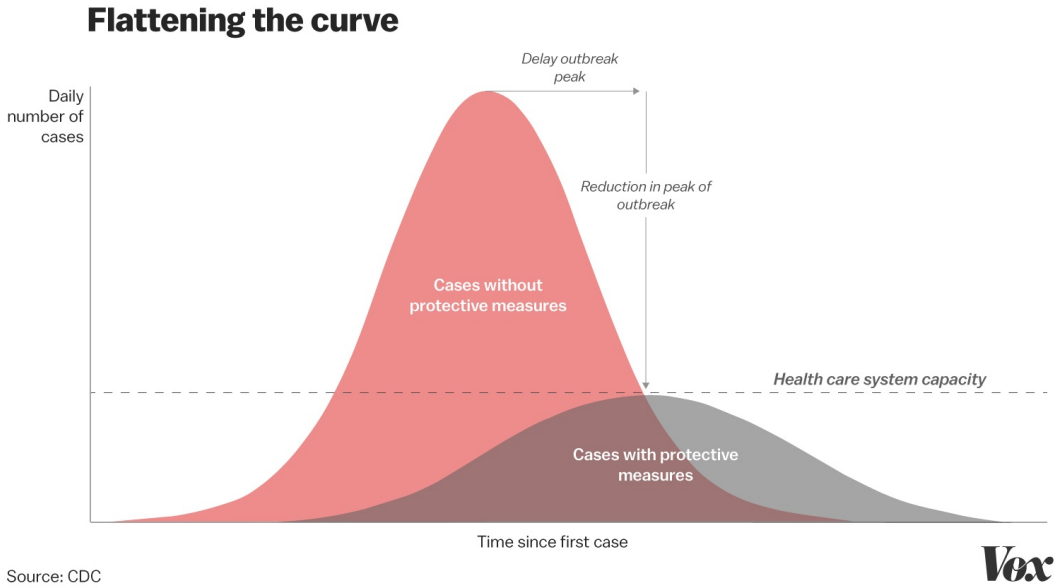
https://medium.com/wintoncentre/does-covid-raise-everyones-relative-risk-of-dying-by-a-similar-amount-more-evidence-e7d30abf6821

Information Visualisation

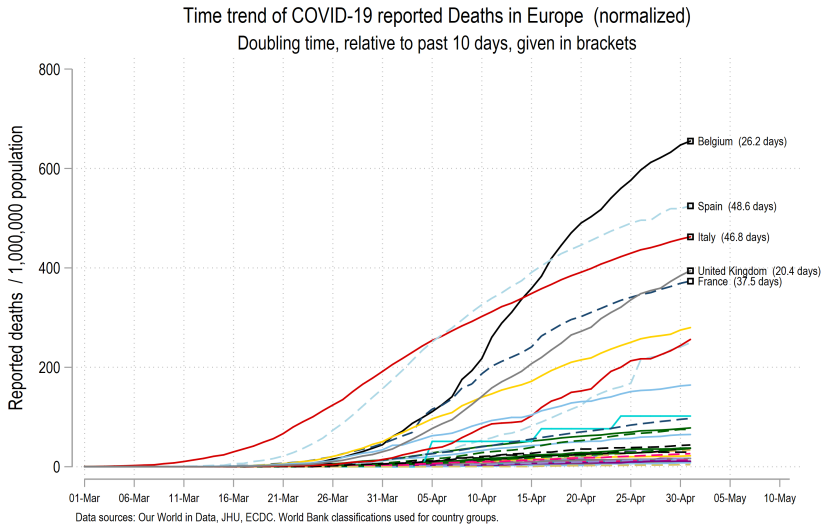
- "We are all epidemiologists now"



John Hopkins University (JHU) Coronavirus Resource Center (https://coronavirus.jhu.edu/map.html).

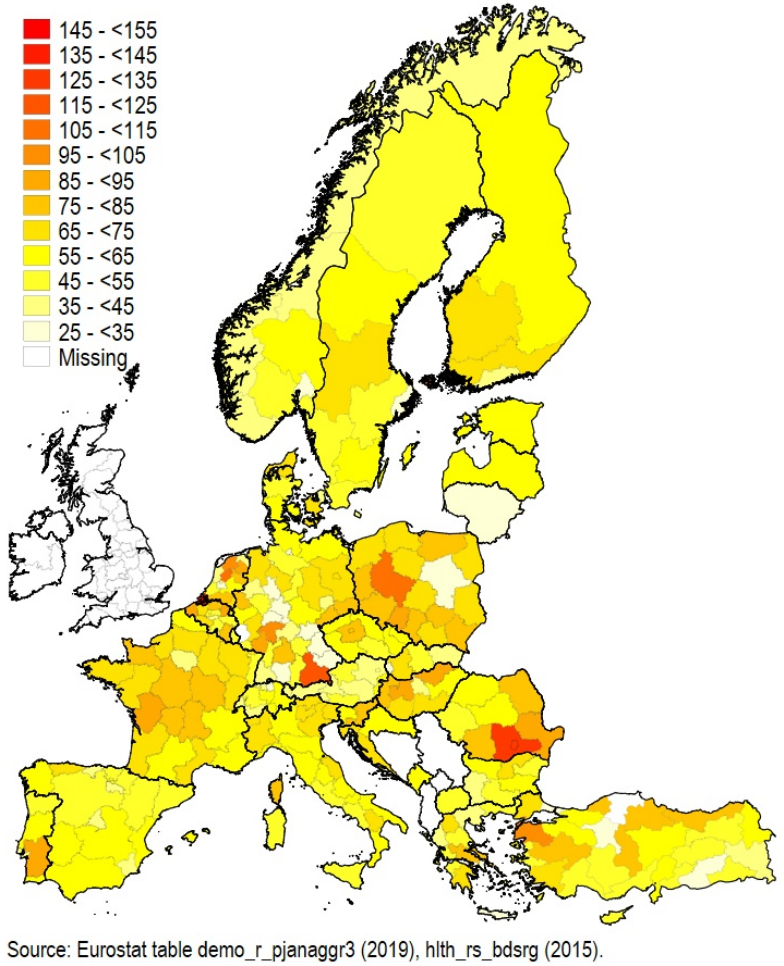


https://www.vox.com/2020/3/10/21171481/coronavirus-us-cases-quarantine-cancellation



Visualisation: topological data analysis

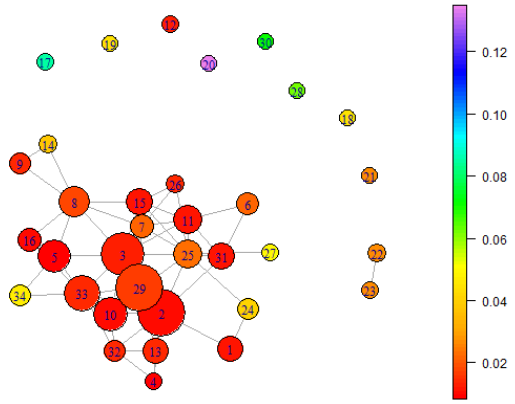
- Ball Mapper is a means of representing complex and high dimensional data sets as abstract graphs.
- BM graphs are constructed by covering the considered data set with a collection of overlapping balls of a radius ε. Central points of each ball are referred to as landmarks, the number of landmarks being a function of the density of the data and the choice of ε.
- Where there are points in the intersection of two balls those balls are connected by an edge on the graph. Numbers of points covered by a ball guide its size in the graph.
- Applied to UK regions: 6 variables (economic and population density). COVID-19 cases per day up to 17/04 – proportion used to colour nodes



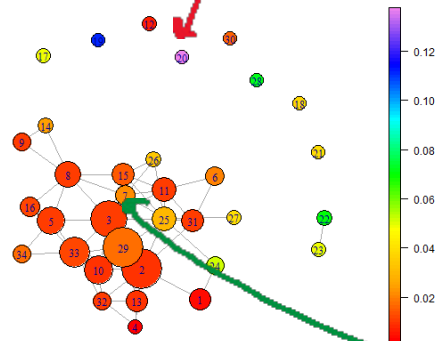
Share of 65+ to doctors

Plots

Ball 20 is Hammersmith, Kensington, Chelsea – outlier by very high earnings

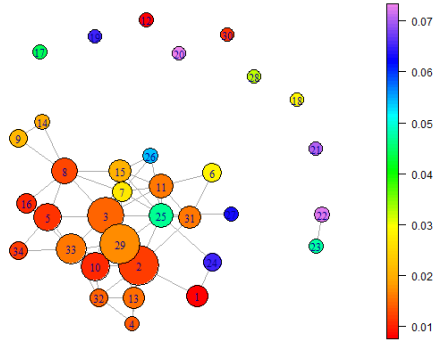


14th March

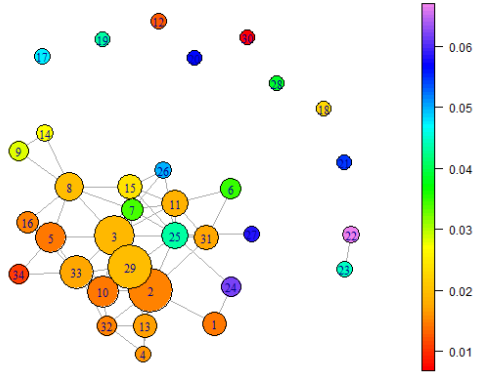


21st March

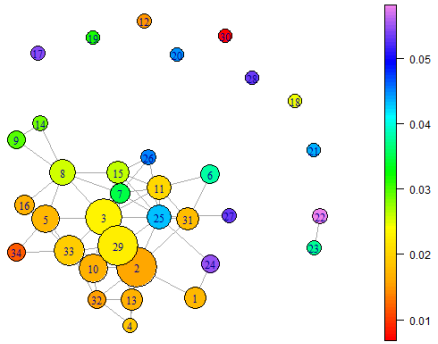
Ball 7 is Bristol, Birmingham, Nottingham, Manchester, Liverpool



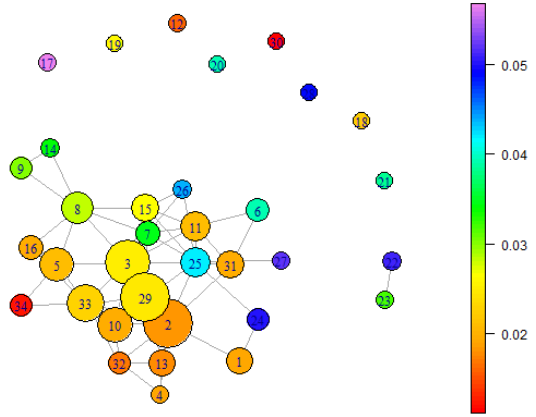
28th March



3rd April



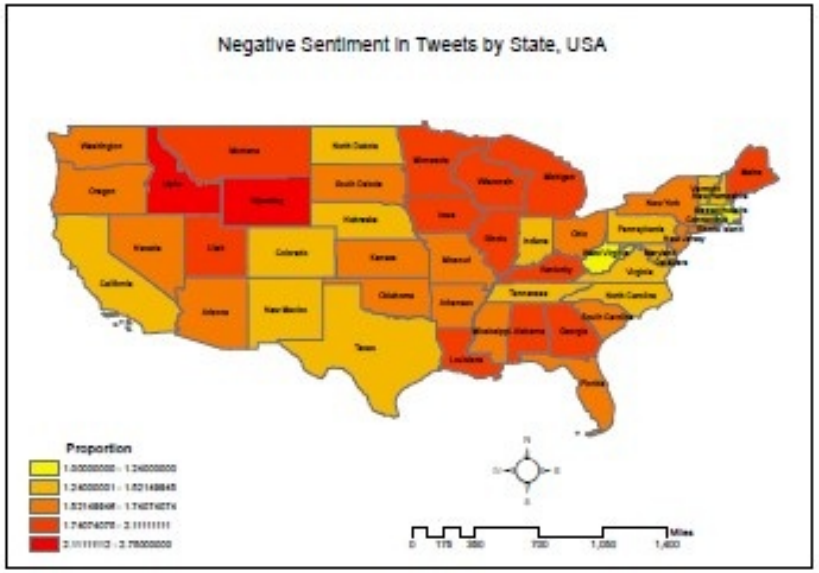
10th April



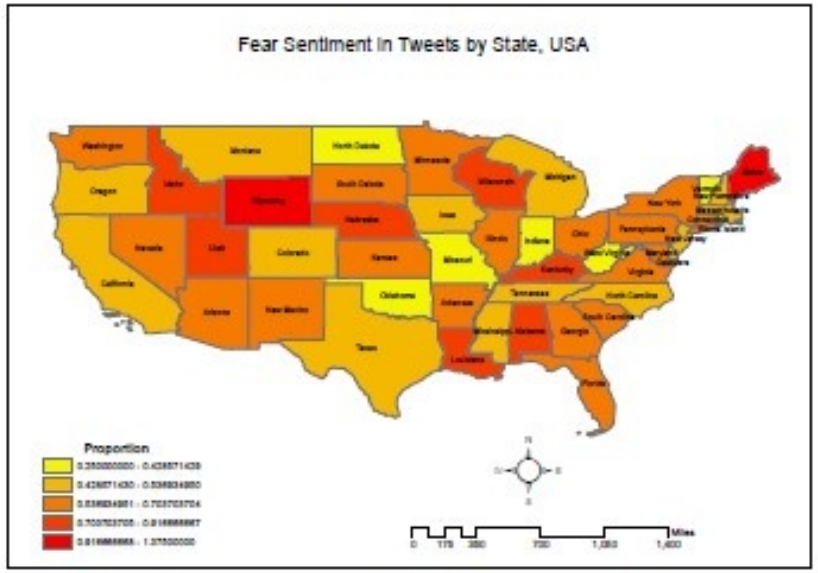
17th April

Geographic distribution of sentiment

- Sentiment analysis of Twitter to reveal ‘situational awareness’ (cf. wildfires)
- Negative sentiments not necessarily well correlated with hotspots.
- Followed up by word clouds (not very informative) and N-grams



(a) Negative sentiment.

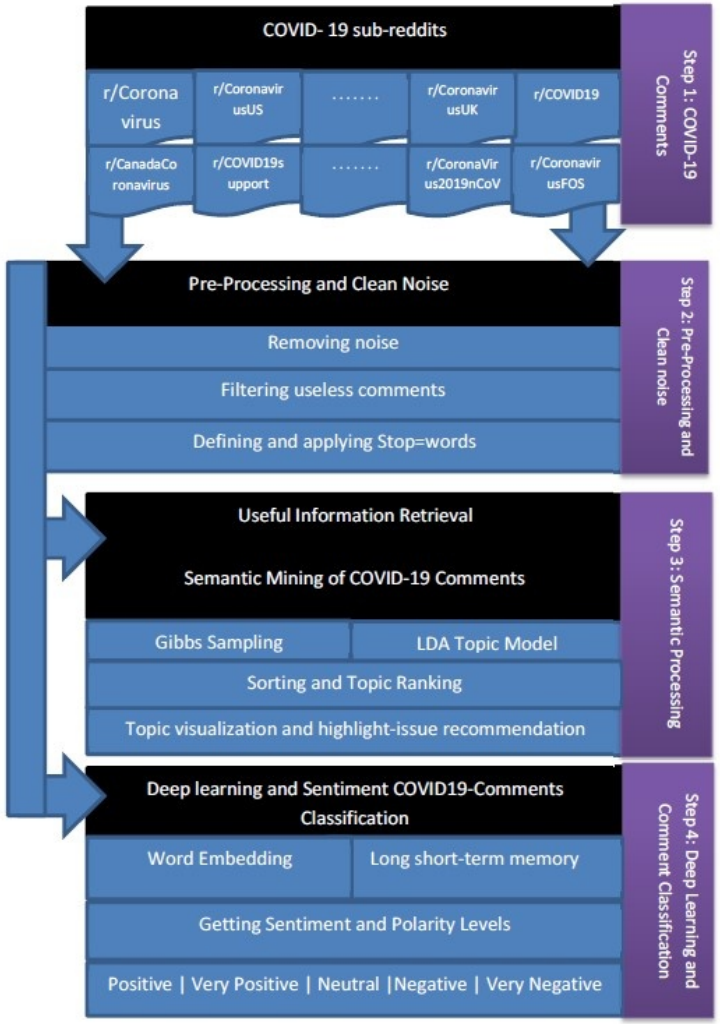


(b) Fear sentiment.

Deep Sentiment Classification and Topic Discovery

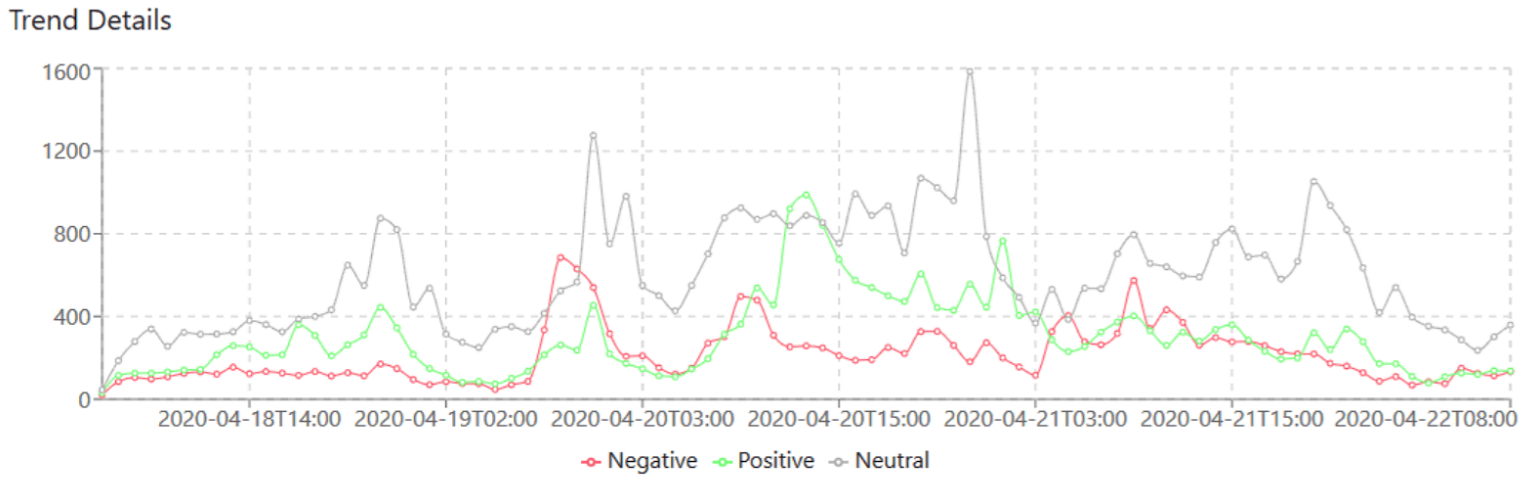
- 563,079 COVID-19–related comments from reddit. The dataset was collected between January 20, 2020 and March 19, 2020

Word cloud of most frequent topic



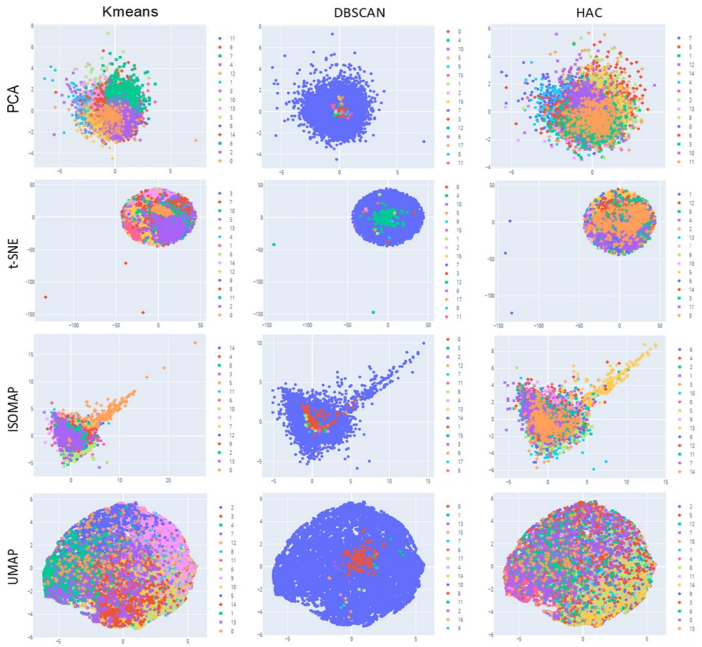
Response to lifting restrictions

- Sentiment analysis of Twitter to in terms of response to lifting lockdown restrictions
- From April 12th 2020 to April 21st 2020, a total of 208,220 tweets were scored and analyzed, this total number of tweets is growing daily as new tweets come in. The tweets are analyzed (sentiment scored) in real time and aggregated hourly.
- The scored tweets are not country specific but are captured globally, the reason being that less than 1-2% of Tweets are geo-tagged.
- Live dashboard https://dashboardliftrestriction.z6.web.core.windows.net/
- Of limited value since countries are at such different stages of the pandemic (and have imposed different restrictions).



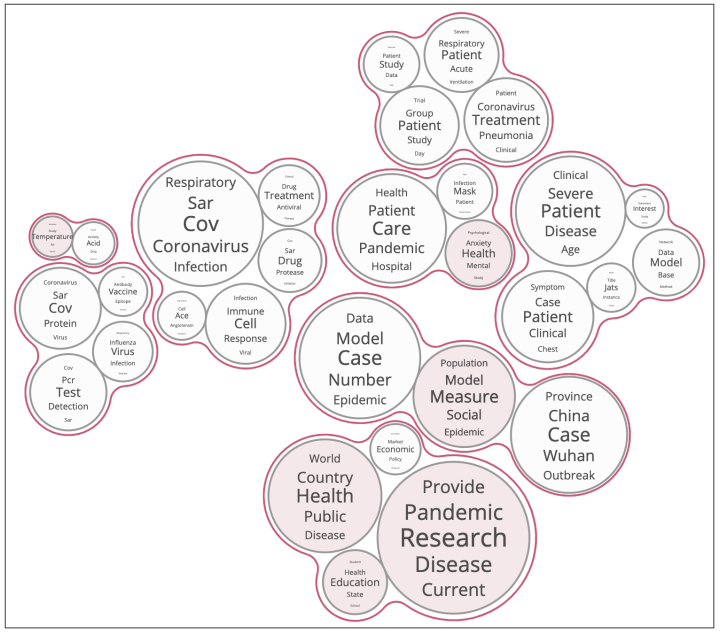
Meta-research on scientific papers

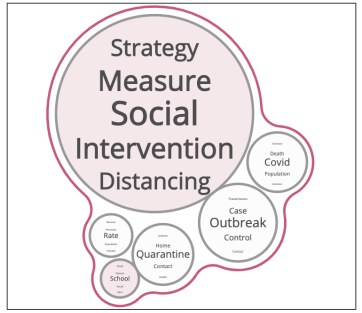
- One-class SVM for clustering and classification on COVID-19 open research dataset (CORD-19: 45,000 articles on corona virus). Nine defined target topics.

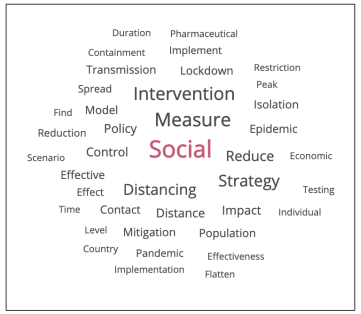


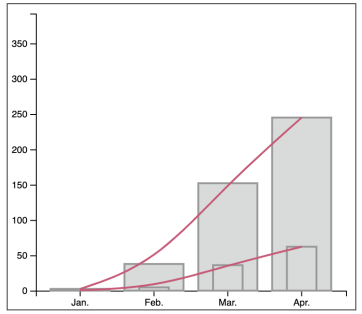
- Hierarchical topic visualization: Latent Dirichlet Allocation (LDA – see JP tutorial) to build a topic model from the research corpus titles and abstracts in Dimensions dataset

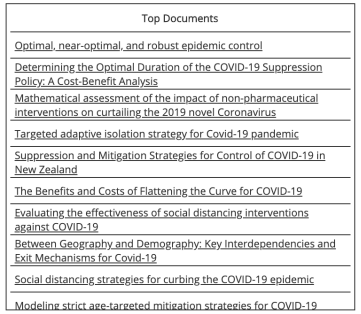
http://strategicfutures.org/TopicMaps/COVID-19/dimensions.html





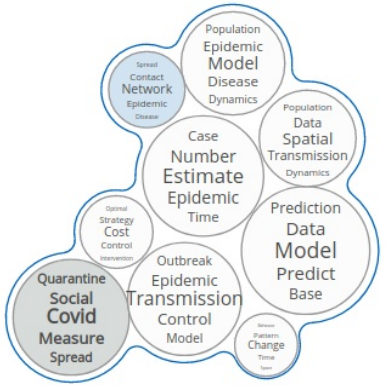


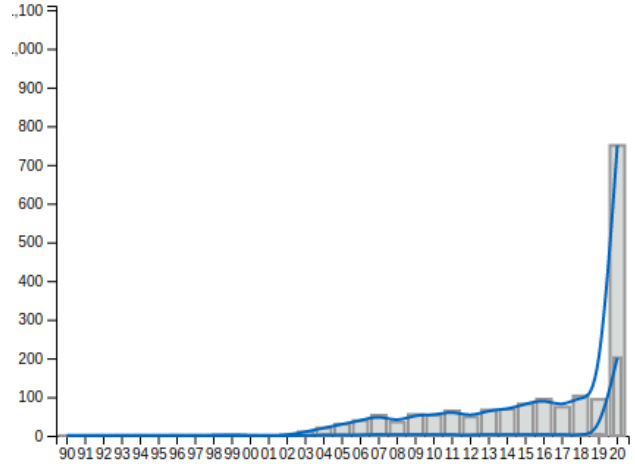




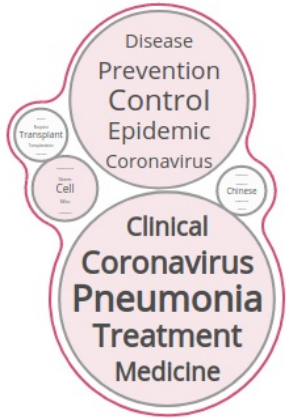
Visualising Research II

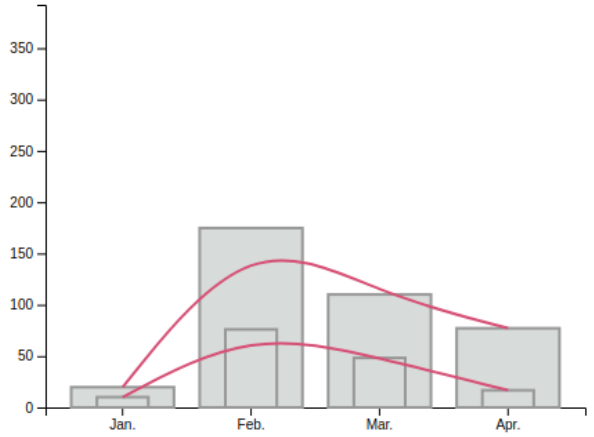
- Unprecedented social distancing



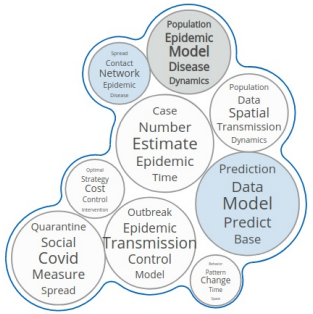


- Responding to a new disease





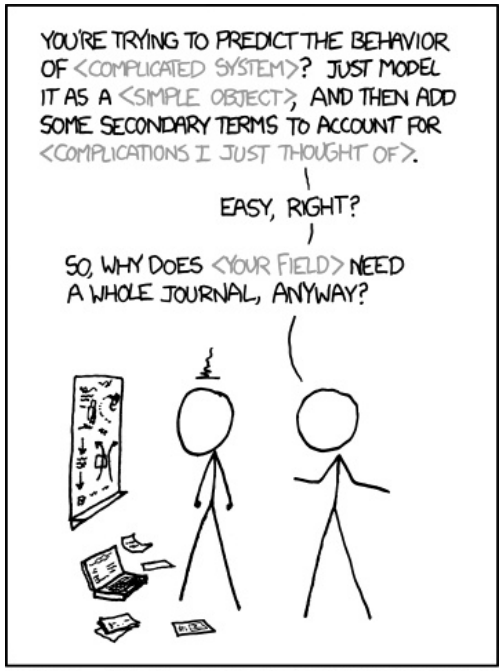
- Finding relevant work



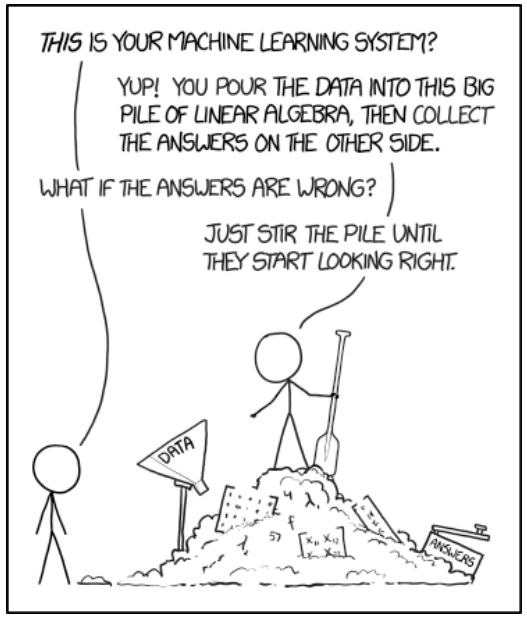
First thoughts

- A paper written one month ago ‘classical’, ‘old’, …
- Lack of definitive data nationally and internationally – cases, deaths, …
- Epidemiological models not always publicly available (UK model on github – but led to concerns about software engineering)
- Lots of analysis of social media data and scientific papers (now there’s a surprise)

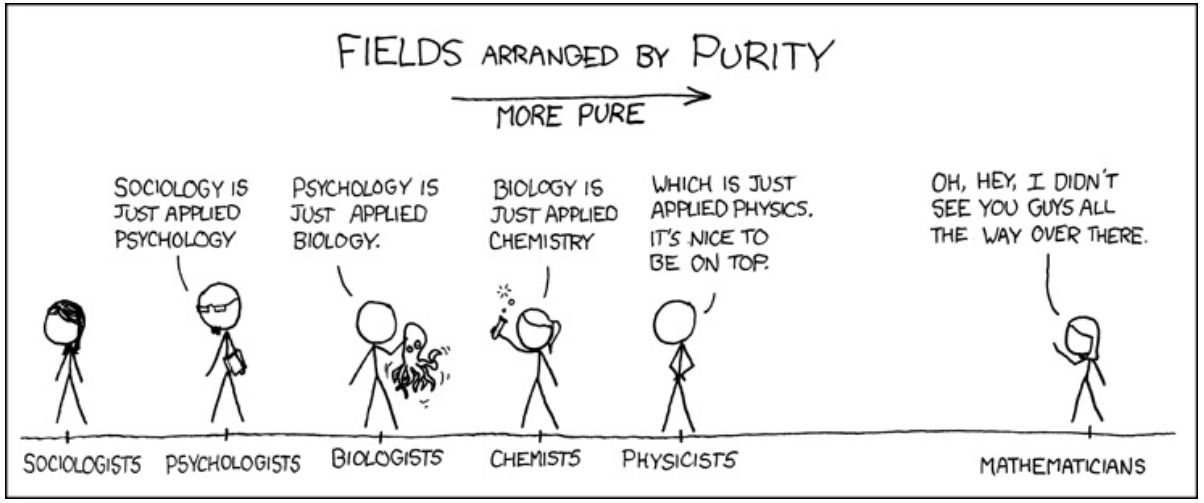
- Humility and responsibility: what we write/visualise needs to be clear and accurate (and we must state the uncertainties). There is a lot more hanging on the conclusions than an academic reputation.



https://xkcd.com/793/



https://xkcd.com/1838/



https://xkcd.com/435/

Bibliography

- https://medium.com/wintoncentre/does-covid-raise-everyones-relative-risk-of-dying-by-a-similar-amount-more-evidence-e7d30abf6821
- Leveraging Data Science To Combat COVID-19: A Comprehensive Review, Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, Adeel Razi, Maged N. Kamel Boulos, Adrian Weller, and Jon Crowcroft.
- Visualising the Evolution of English Covid-19 Cases with Topological Data Analysis Ball Mapper, Pawel Dlotko and Simon Rudkin.
- Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach, Hamed Jelodar, Yongli Wang, Rita Orji, Hucheng Huang.
- Target specific mining of COVID-19 scholarly articles using one-class approach, Sanjay Kumar Sonbhadra, Sonali Agarwal and P. Nagabhushan.
- Visualising COVID-19 Research, Pierre Le Bras, Azimeh Gharavi, David A. Robb, Ana F. Vidal, Stefano Padilla, and Mike J. Chantler.
- COVID-19: visualising regional socio-economic indicators for Europe, Asjda Naqvi.
- COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification, Jim Samuel , G. G. Md. Nawaz Ali , Md. Mokhlesur Rahman, Ek Esawi, and Yana Samuel.
- https://blog.iiasa.ac.at/category/data-and-methods/ Santosh Karanam

Thank you