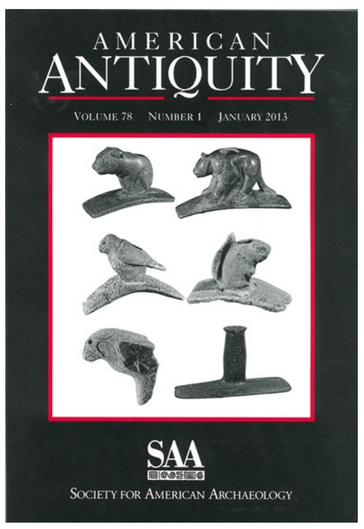
Text mining JSTOR

Quantitative approaches to histories of science

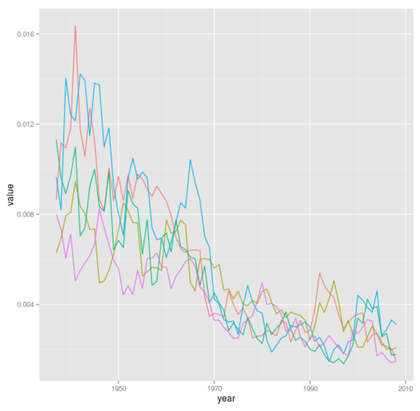
Ben Marwick & Ian Kretzler, UW Anthropology

Google's Ngram Viewer is a tool for visualizing the popularity of words over time in 5 million books digitized by Google. It has been described as the 'gateway drug' for data science in the humanities. While it is fun for casual searches, it has substantial limitations that prevent it from being a scholarly tool. The main problem is we simply don't know what is in the corpus. This makes validation through close reading of the texts impossible.

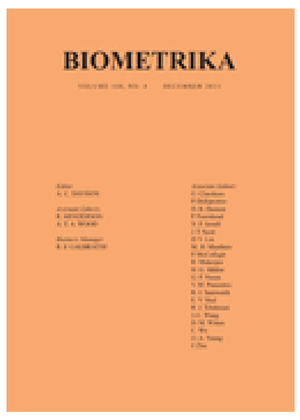
Inspired by Google's Ngram Viewer, we wrote JSTORr, an R package that visualizes ngrams for corpora held by JSTOR, a digital library of 2000 scholarly journals. JSTOR's Data for Research service allows users to download full text of journals. With JSTORr, the full text can be analysed for ngrams, word correlations, document clustering and topic modelling. This means that we can now carefully build sizable corpora of scholarly literature on specific topics and investigate them with text mining methods. Here we demonstrate some of the basic functions of the package to get insights into the histories of science.

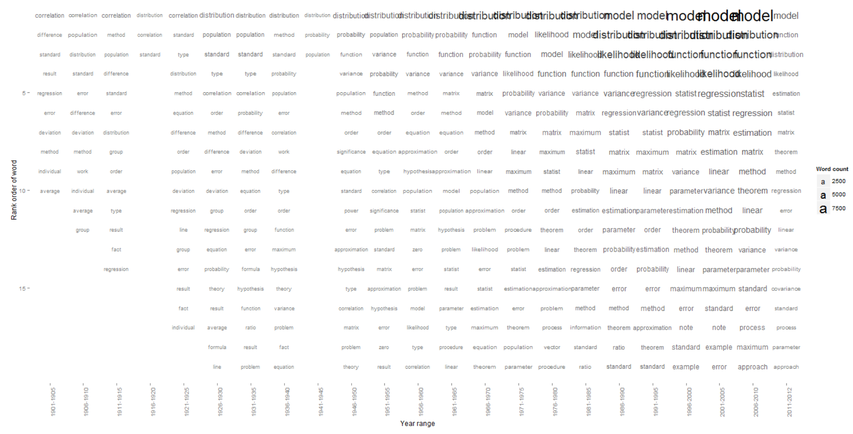


In archaeology we see a turn from typological analyses of pottery to work on population-level behaviors and settlement patterns. Dating becomes frequently discussed after radiocarbon methods become widely available.



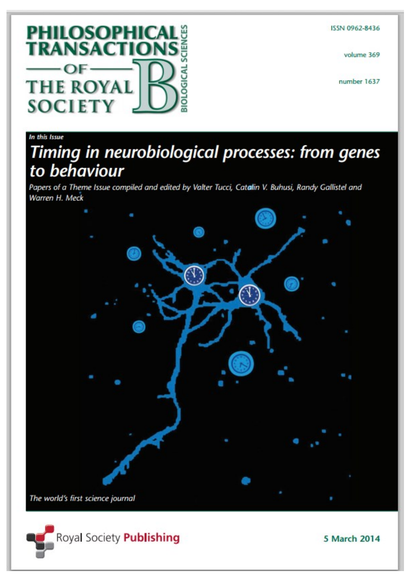
Topic models generated by latent Dirichlet allocation (above) show hot and cold topics over time in American archaeology. These validate and add context to the most frequent words over time in the top figure.

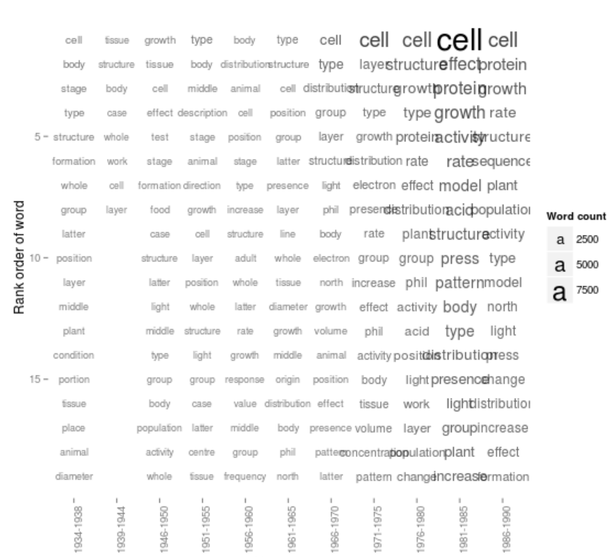




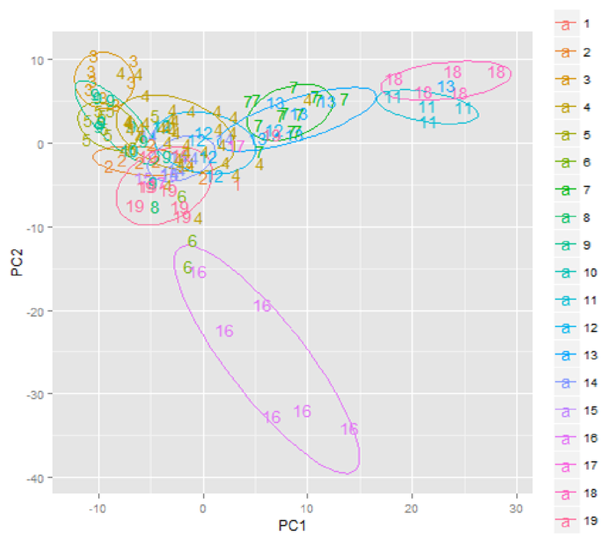
For example, we can quickly see intellectual shifts in a high-impact statistics journal (above) from correlation to model-based analyses. Similarly, a prominent biology journal (below) reveals a turn from macrostructures to a focus on cells and proteins. We might hypothesize that these intellectual trends are related to technological changes. This hypothesis could be tested by close reading of a small sample of articles.

Ngrams over time (left) show the complex relationship between gender and feminism in American archaeology. The two terms appear together then diverge in their frequencies.





K-means clustering of journal articles using word frequencies (right) reveals distinct approaches to gender in the American archaeological literature. In the lower cluster we have a cluster of articles on the archaeological record of the US Southwest. The upper right clusters are articles about gender theory and philosophy.



JSTORr is available from https://github.com/bmarwick/JSTORr Code for **reproducing the figures** on this poster is available at https://github.com/benmarwick/Data-Science-at-UW-Poster Data are available from http://dfr.jstor.org Thanks to Magdalena Balazinska (UW CS) for providing access to computing resources for development and testing. Thanks to Jiun-Yu Liu and Joss Whittaker for intensive testing and many useful suggestions