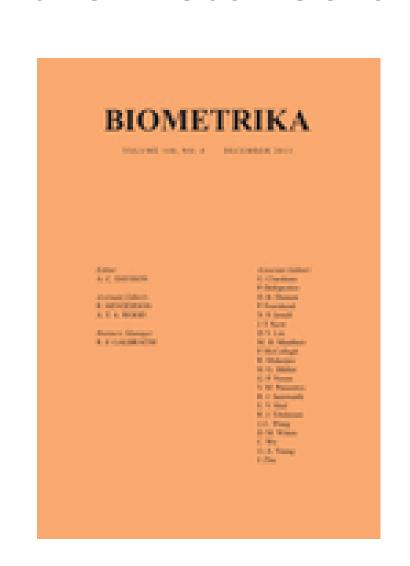
## Text mining JSTOR

## 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.

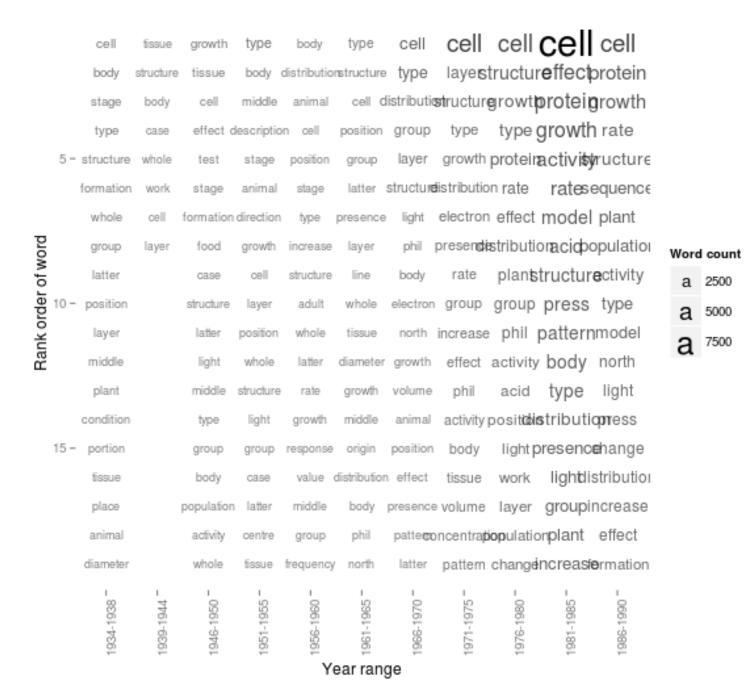
Inspired by Google's Ngram viewer, we wrote JSTORr, an R package that visualizes Ngrams for corpora held by JSTOR, a digital library of 2000 scholarly journals. JSTOR's Data for Research service allows users to download full text of journals. With JSTORr, the full text can be analysed for ngrams, word 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.

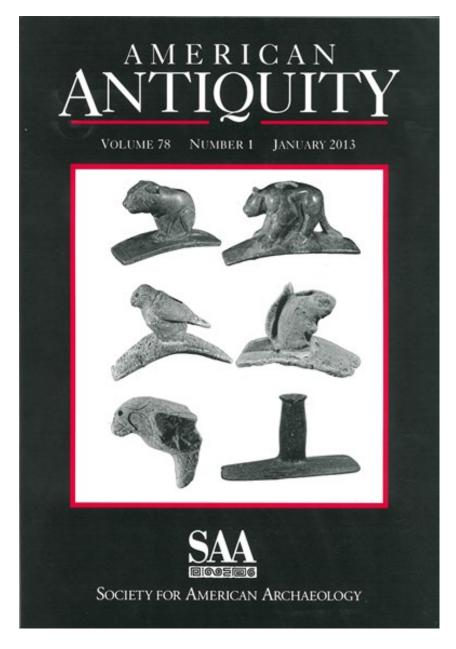


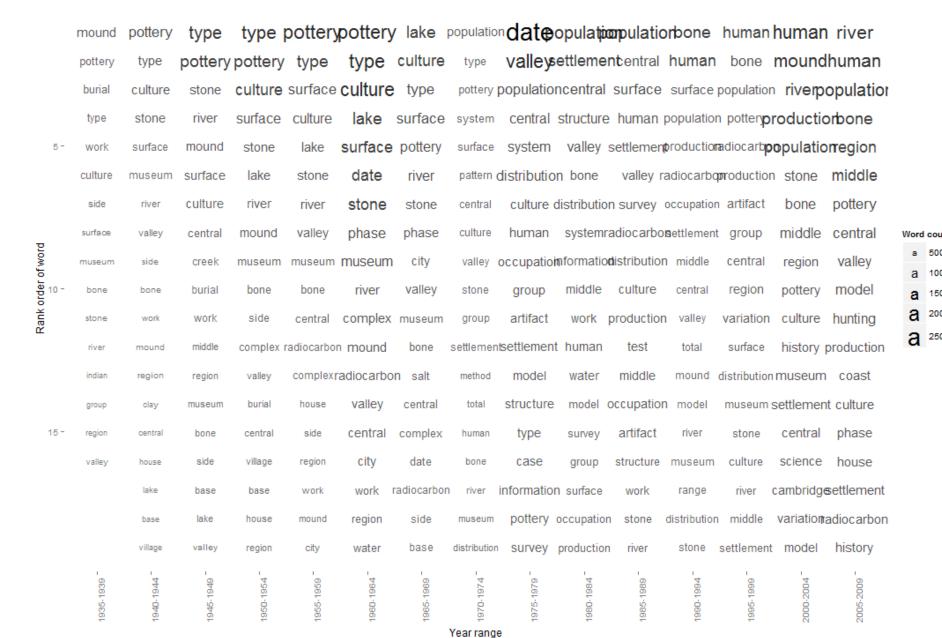
	correla	tion correlation	correlation	distribution	correlation	distribution	n distribution	distribution	distribution	distributio	histributio:	distributi <b>di</b>	stribu <b>d</b> a	stribu <b>t</b> i	<b>st</b> ribu <b>di</b>	<b>st</b> ribu <b>t</b> i	<b>st</b> ributi	omodel	model	mod∉	hode	hode	model	
Rank order of word	differer	ce population	n method	correlation	standard	population	population	method	probability	probability	population	probability	probability	function	model	likelihood	mod <b>ed</b>	istribu <b>t</b> li	isatribu <b>di</b> k	entrib <b>dt</b> i	atmibuti	<b>stn</b> ibutio	nction	
	standa	rd distribution	population	standard	type	standard	standard	standard	population	function	variance	function	function	probability	function	model	likelihook	ikelihodi	delihoo	<b>f</b> unction	iunction	iunction	listribution	
	resul	standard	difference		distribution	type	type	probability		variance	probability	variance	variance	variance	likelihood	function	function	function	function	ikeliho <b>di</b>	kelihodi	kelihood	likelihood	
	5 - regress	ion error	standard		method	correlation	correlation	population		population	function	method	matrix	matrix	probability	variance	variance	variance	regressio	n statist r	egressio	nstatist	estimation	
	error	difference	error		equation	order	probability	error	ror	method	method	order	method	model	variance	probability	matrix	regression	varianceregressi		n statist r	egression	statist	
	deviat	on deviation	distribution		difference	method	difference	correlation		order	order	equation	equation	method	matrix	matrix	maximum	statist	statist	orobabilit	/ matrix e	stimation	matrix	
	metho	d method	group		order	difference	deviation	work		significance	equation a	approximation	order	order	linear	maximum	statist	matrix	maximum	matrix	estimation	matrix	theorem	
	individ	ual work	order		population	error	method	difference		equation	type	hypothesisar	oproximatio	n linear	maximum	statist	linear	maximum	matrix	variance	linear	method	method	Word count a 2500
	0 - avera	ge individual	average		deviation	deviation	equation	type		standard	correlation	population	model	population	method	method	probability	linear	linear	paramete	variance	theorem	regression	<b>a</b> 5000
		average	type		regression	group	order	order		power	significance	statist	population a	pproximatio	n order	order		estimation				linear	error	<b>a</b> 7500
		group	result		line	regression	group	function		error	problem		hypothesis		procedure	theorem		parameter				orobability		
			fact		group	equation	error	maximum		approximation		zero	problem	likelihood	problem	linear		probability						
	<b>-</b>		regression		error	probability	formula	hypothesis		hypothesis	matrix	error	statist	error	statist	estimation			probability			parameter	probability covariance	
	5 -				result	theory	hypothesis	theory		-	approximation hypothesis	model	result	statist estimation	error	problem	method	error	error	error	standard		standard	
					individual	average	ratio	problem		correlation	error zero result	likelihood	type	maximum equation	theorem population	process	information standard	n theorem a	approximation theorem		note	process		
						formula	result	fact		problem														
						line	problem	equation		theory												approach	approach	
	1		ا ا		<u>و</u>	-	- 9	- 0	10		- Q	-	, Q		I ID		L Q		- 9	-0	ا ا		1 2	
	904-1905	1906-1910	1911-1915	916-1920	1921-1925	926-1930	1931-1935	1936-1940	1941-1945	946-1950	1951-1955	1956-1960	1961-1965	966-1970	971-1975	976-1980	1981-1985	986-1990	1991-1995	996-2000	2001-2005	2008-2010	2011-2012	
	-	<del>-</del>	-	-	-	<del>-</del>	-	-	-	<del>-</del>		Year ranne	<del>-</del>	-	-	<del>-</del>	-	<del>-</del>	-	-	Ø	Ø	Ø	

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.

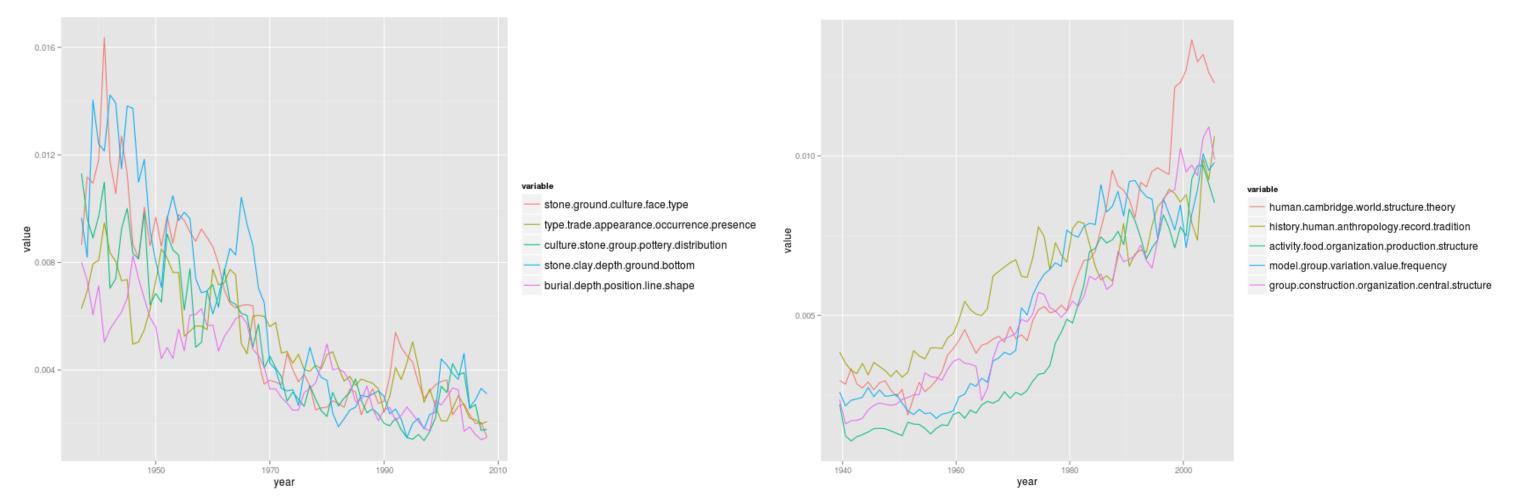




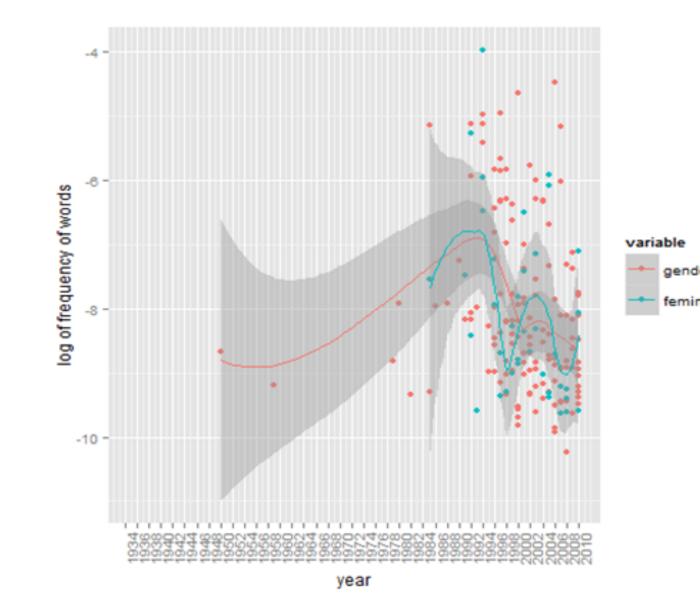




In archaeology we see a turn from typological analyses of pottery to work on population-level behaviors and settlement patterns. Dating becomes frequently discussed after radiocarbon methods become widely available.

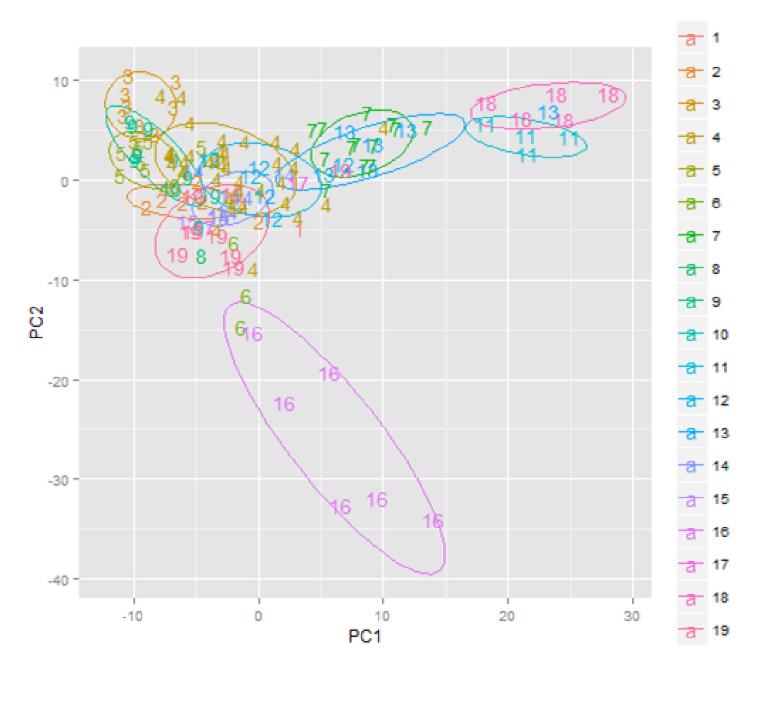


Topic models generated by latent Dirichlet allocation (above) show hot and cold topics over time in American archaeology. These validate and add context to the most frequent words over time in the top figure.



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 upper right we have a cluster of articles on the archaeological record of the US Southwest. The large lower cluster is articles about gender theory and philosophy.







JSTORr is available from <a href="https://github.com/UW-ARCHY-textual-macroanalysis-lab/JSTORr">https://github.com/benrarwick/Data-Science-at-UW-Poster</a> Data are available from <a href="http://dfr.jstor.org">http://dfr.jstor.org</a> 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