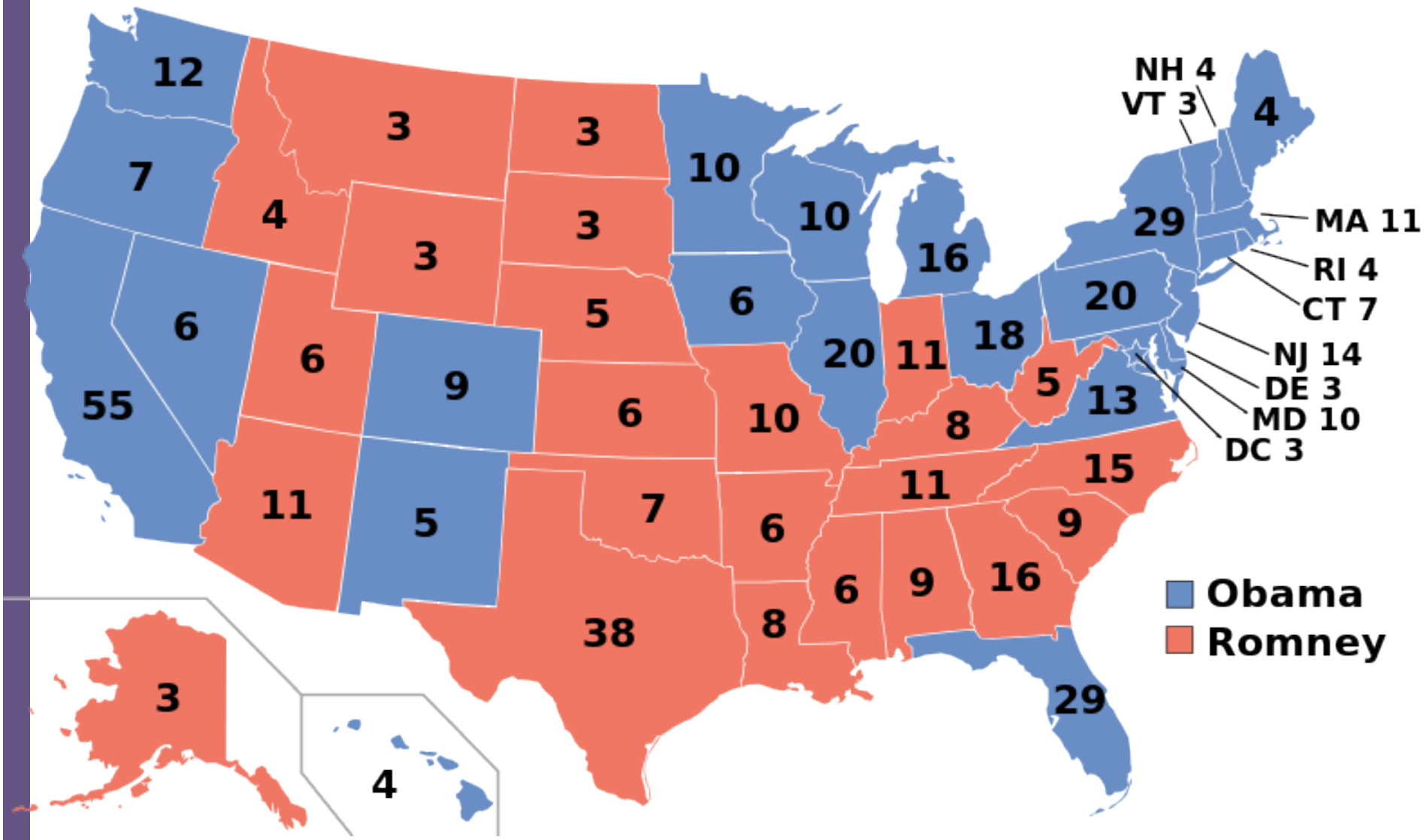
UNIVERSITY of WASHINGTON

Introduction to Data Science

Bill Howe, PhD
Director of Research,
Scalable Data Analytics
University of Washington
eScience Institute



http://commons.wikimedia.org/wiki/File:ElectoralCollege2012.svg
(public domain)

Nate Silver

source: randy stewart

"The intuition behind this ought to be very simple: Mr. Obama is maintaining leads in the polls in Ohio and other states that are sufficient for him to win 270 electoral votes."

Nate Silver, Oct. 26, 2012
fivethirtyeight.com

"…the argument we're making is exceedingly simple. Here it is: Obama's ahead in Ohio."

Nate Silver, Nov. 2, 2012
fivethirtyeight.com

"The bar set by the competition was invitingly low. Someone could look like a genius simply by doing some fairly basic research into what really has predictive power in a political campaign."

Nate Silver, Nov. 10, 2012
DailyBeast

Related: Obama campaign's data-driven ground game

"In the 21st century, the candidate with [the] best data, merged with the best messages dictated by that data, wins."

Andrew Rasiej, Personal Democracy Forum

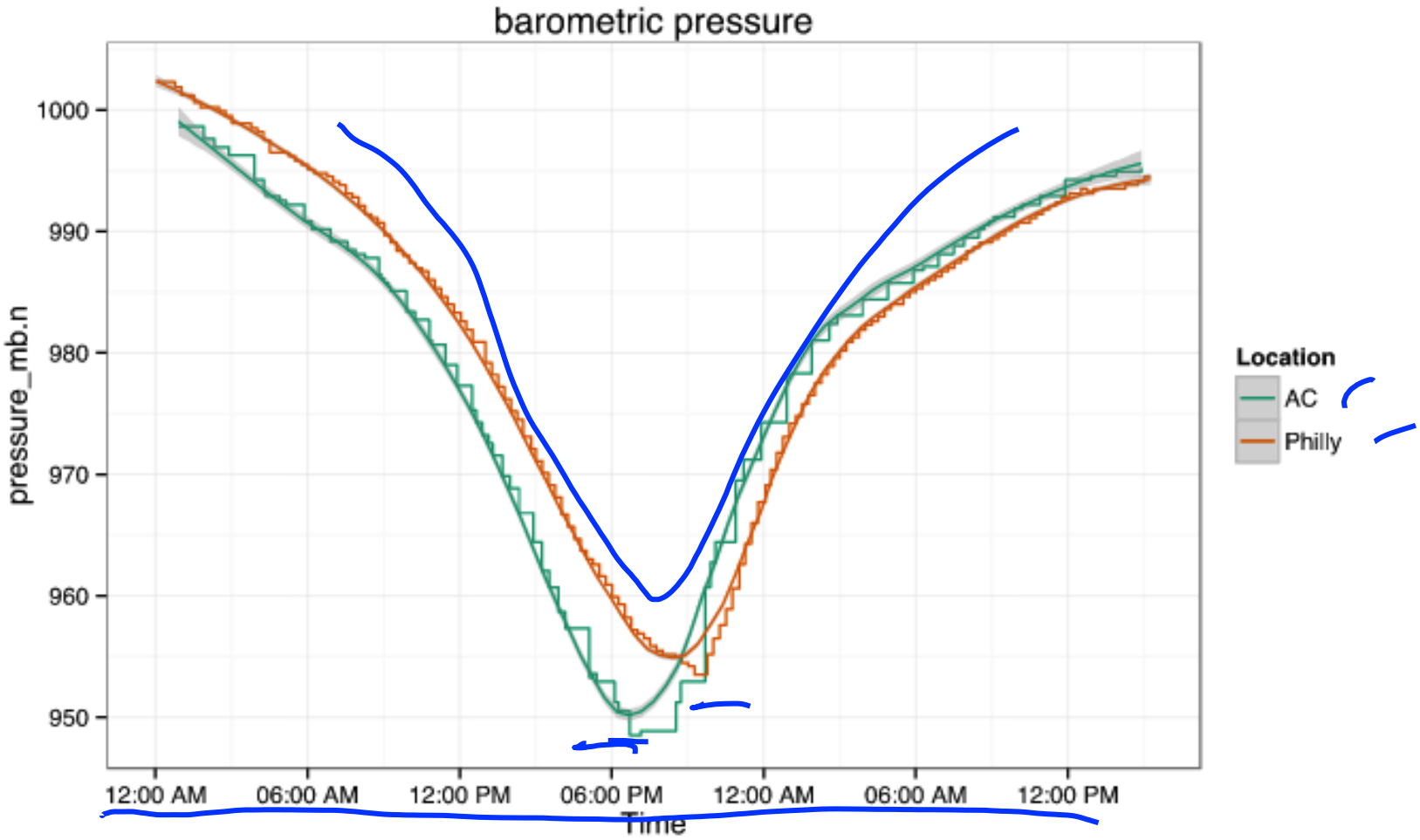
"…the biggest win came from good old SQL on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed."

Dan Woods
Jan 13 2013, CITO Research

"The decision was made to have Hadoop do the aggregate generations and anything not real-time, but then have Vertica to answer sort of 'speed-of-thought' queries about all the data."

Josh Hendler, CTO of H & K Strategies

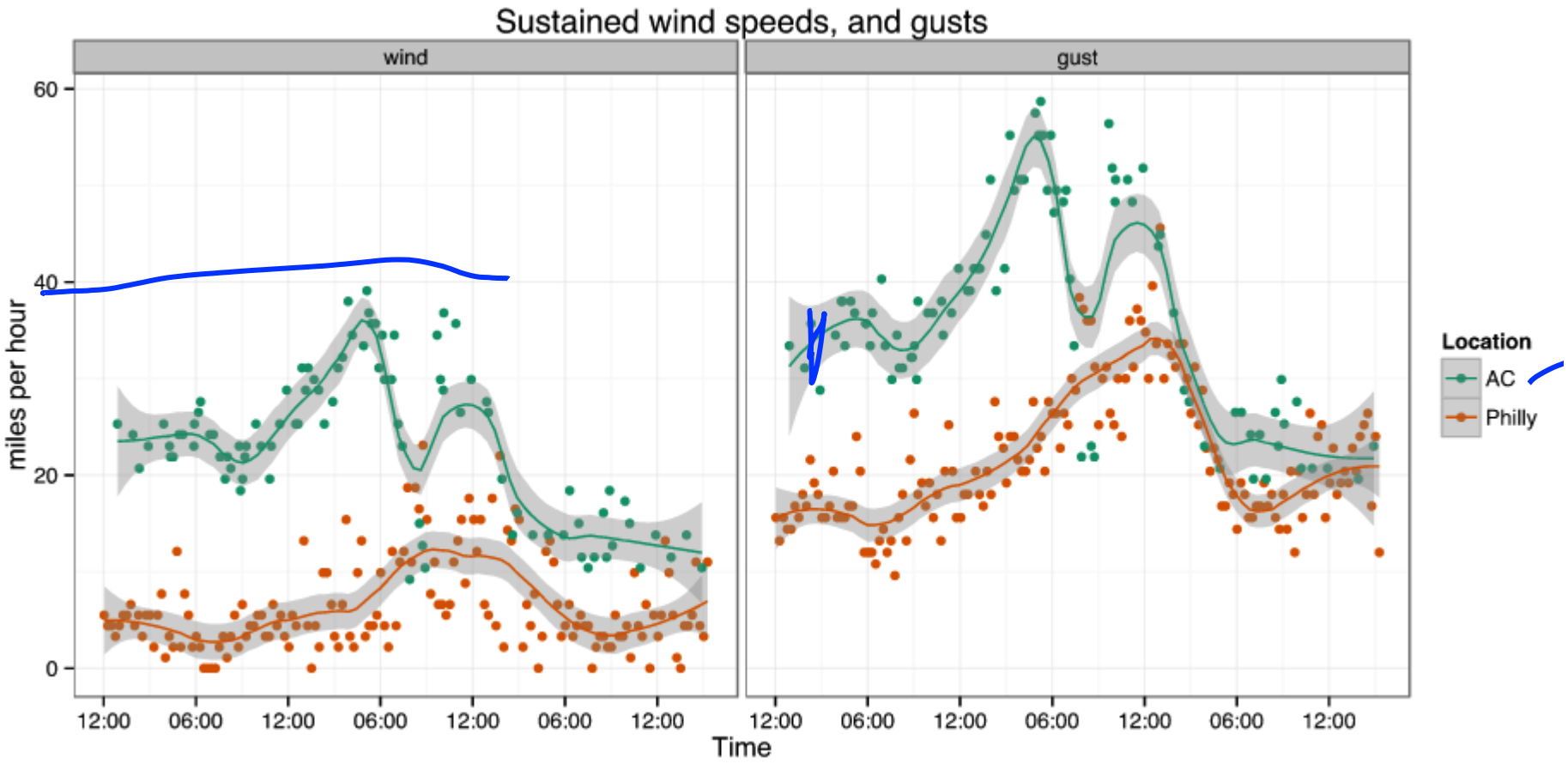
Hurricane Sandy



http://rpubs.com/JoFrhwld/sandy

Josef Fruehwald

Hurricane Sandy



http://rpubs.com/JoFrhwld/sandy

Josef Fruehwald

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030

1) Convert all the digitized books in the 20th century into n-grams
(Thanks, Google!)
(http://books.google.com/ngrams/)

A 1-gram: "yesterday"
A 5-gram: "analysis is often described as"

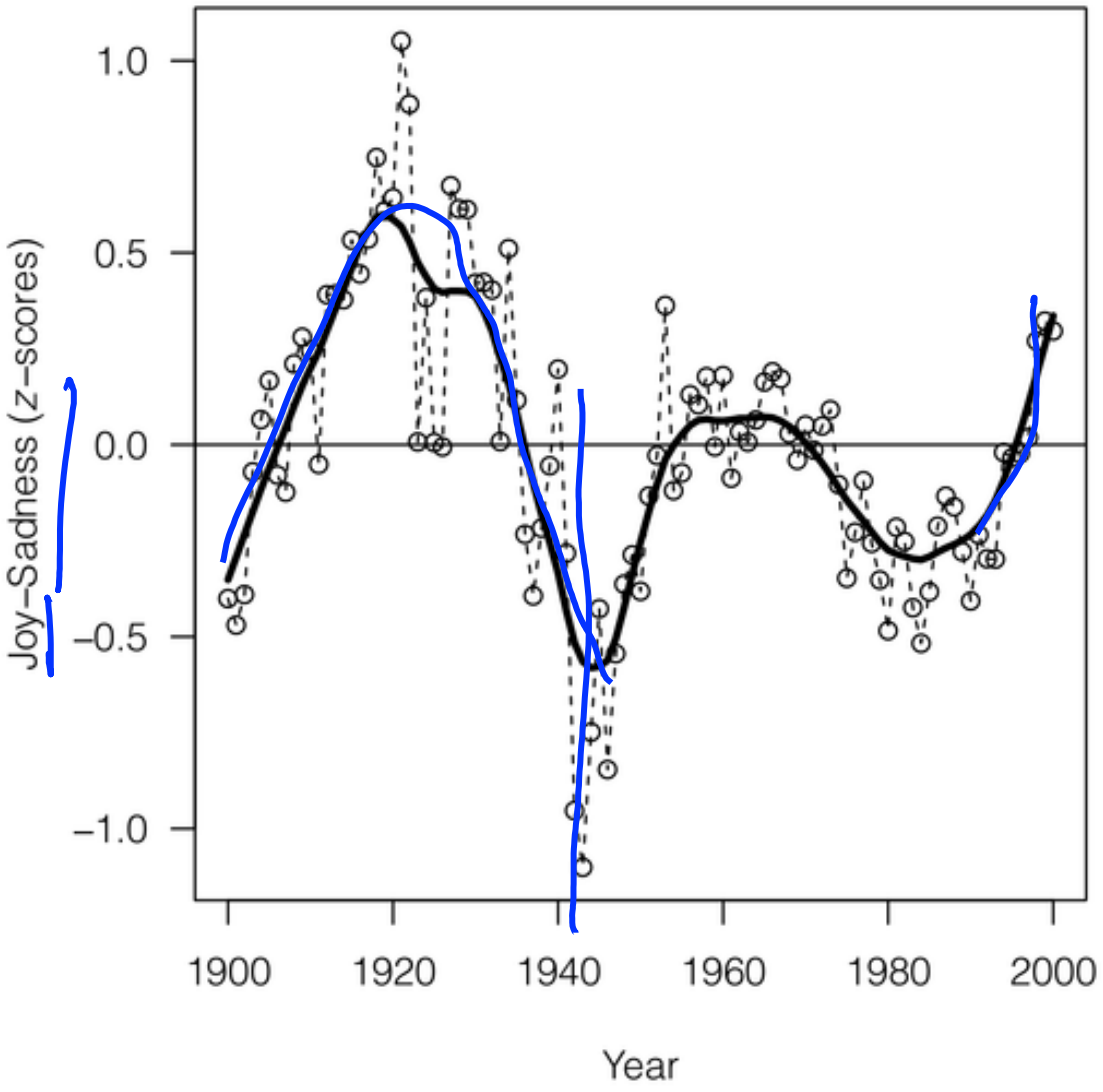
2) Label each 1-gram (word) with a mood score.
(Thanks, WordNet!)

3) Count the occurences of each mood word

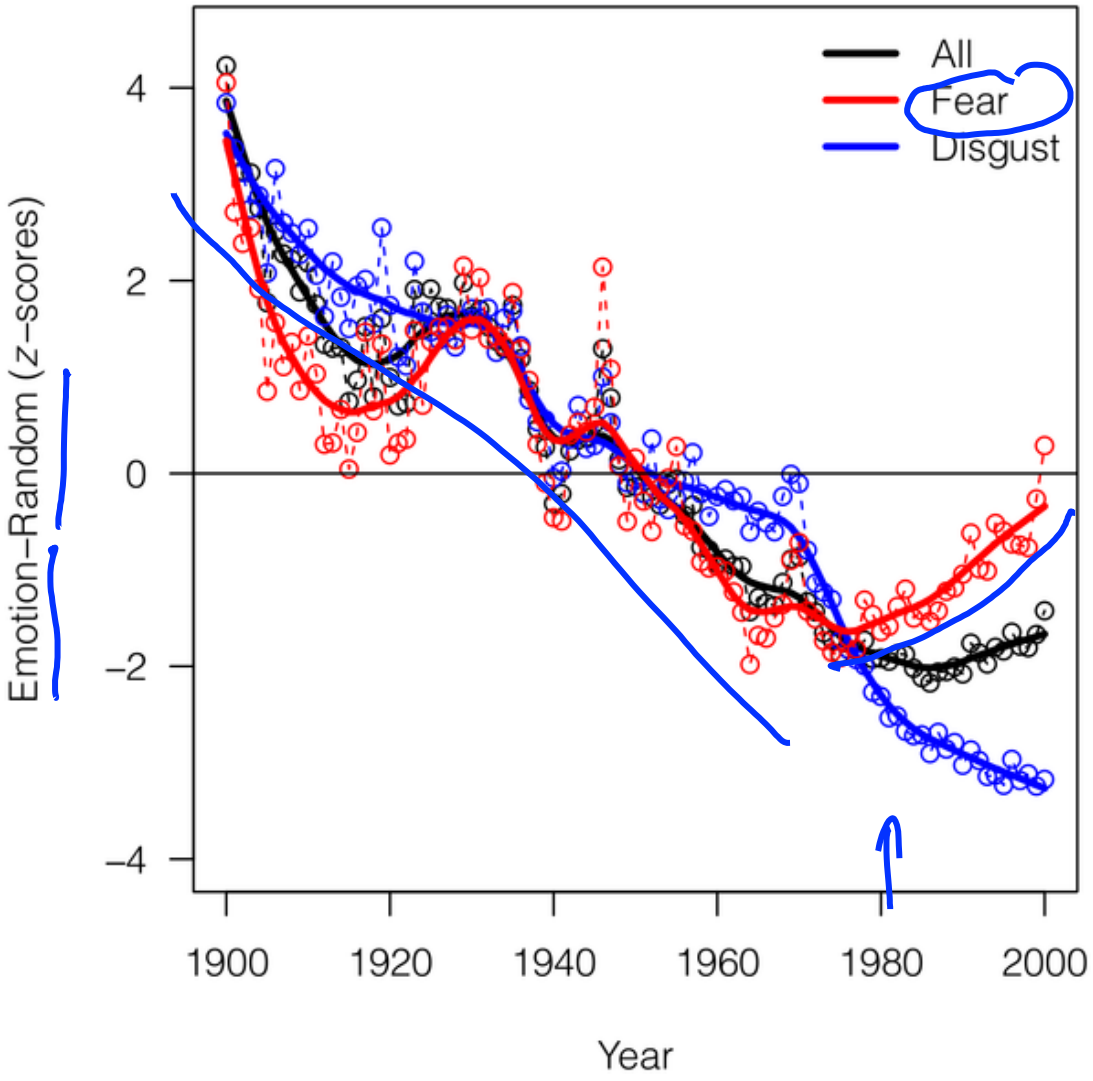
$$\mathcal{M}_Y = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{C_{\text{the}}},$$

$$\mathcal{M}z_Y = \frac{\mathcal{M}_Y - \mu_{\mathcal{M}}}{\sigma_{\mathcal{M}}},$$

Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi: 10.1371/journal.pone.0059030



Acerbi A, Lampos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi: 10.1371/journal.pone.0059030



…
2. Michel J-P, Shen YK, Aiden AP, Veres A, Gray MK, et al. (2011) ***Quantitative analysis of culture using millions of digitized books***. Science 331: 176–182. doi: 10.1126/science.1199644. Find this article online
3. Lieberman E, Michel J-P, Jackson J, Tang T, Nowak MA (2007) ***Quantifying the evolutionary dynamics of language***. Nature 449: 713–716. doi: 10.1038/nature06137. Find this article online
4. Pagel M, Atkinson QD, Meade A (2007) ***Frequency of word-use predicts rates of lexical evolution throughout Indo-European history***. Nature 449: 717–720. doi: 10.1038/nature06176. Find this article online
…
6. DeWall CN, Pond RS Jr, Campbell WK, Twenge JM (2011) ***Tuning in to Psychological Change: Linguistic Markers of Psychological Traits and Emotions Over Time in Popular U.S. Song Lyrics***. Psychology of Aesthetics, Creativity and the Arts 5: 200–207. doi: 10.1037/a0023195. Find this article online
…