CFR Miner

Richard Schwinn

8/1/2018 (updated: 2018-07-31)

Disclaimer

The views expressed in this presentation reflect the views of the author alone, and do not necessarily reflect those of the Office of Economic Research, the Office of Advocacy, nor the US Small Business Administration.

- This presentation was created for use with the *xaringan* package. Certain aspects are not capatible with the *pdf* format.

We are Mandated by Congress

- to research
  - the contributions,
  - status,
  - and needs of small businesses and
- to serve as
  - an independent voice for small business within the federal government, and
  - the watchdog for the Regulatory Flexibility Act (RFA)

What is the Code of Federal Regulations (CFR)?

- The CFR contains the final rules and regulations published in the Federal Register by the executive departments and agencies of the federal government.
  - e.g. The Federal Aviation Administration (FAA) regulates civil aviation to promote safety.
  - e.g. The Food and Drug Administration (FAA) is responsible for protecting the public health by ensuring the safety, efficacy, and security of human and veterinary drugs, biological products, and medical devices; and by ensuring the safety of our nation's food supply, cosmetics, and products that emit radiation.

CFR

- Navigation of the CFR is difficult
  - for the general public
  - AND FOR AGENCIES.
- In 1970, the CFR was about 50,000 pages and required a little more than a year to read.
- Today, it is 180,000+ pages and requires almost 4 years to read.

CFR

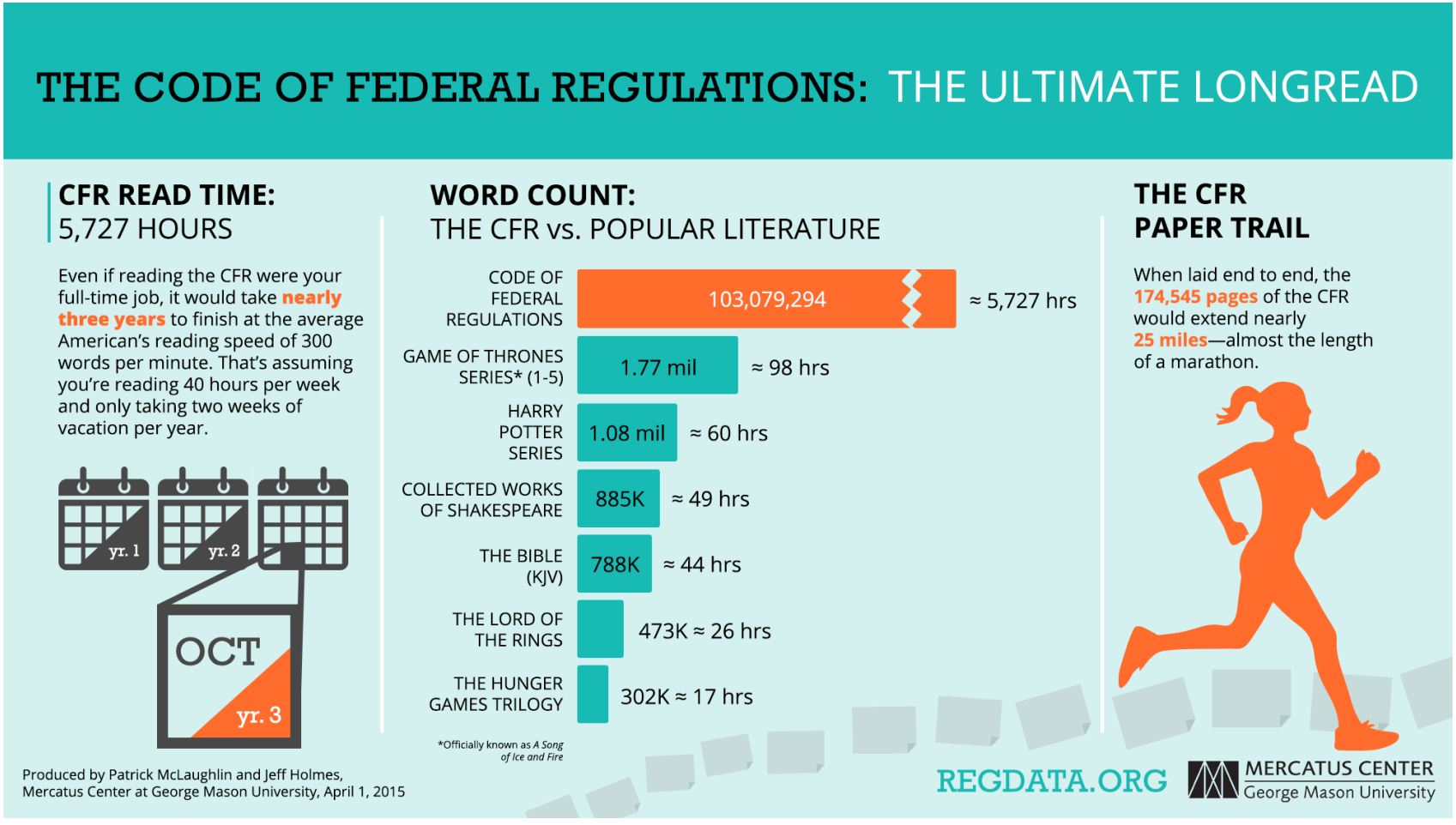
THE CODE OF FEDERAL REGULATIONS: THE ULTIMATE LONGREAD
CFR READ TIME:
5,727 HOURS
Even if reading the CFR were your full-time job, it would take nearly three years to finish at the average American's reading speed of 300 words per minute. That's assuming you're reading 40 hours per week and only taking two weeks of vacation per year.
yr. 1
yr. 2
OCT
yr. 3
WORD COUNT:
THE CFR vs. POPULAR LITERATURE
CODE OF FEDERAL REGULATIONS
103,079,294
≈ 5,727 hrs
GAME OF THRONES SERIES* (1-5)
1.77 mil
≈ 98 hrs
HARRY POTTER SERIES
1.08 mil
≈ 60 hrs
COLLECTED WORKS OF SHAKESPEARE
885K
≈ 49 hrs
THE BIBLE (KJV)
788K
≈ 44 hrs
THE LORD OF THE RINGS
473K ≈ 26 hrs
THE HUNGER GAMES TRILOGY
302K ≈ 17 hrs
*Officially known as A Song of Ice and Fire
THE CFR PAPER TRAIL
When laid end to end, the 174,545 pages of the CFR would extend nearly 25 miles—almost the length of a marathon.
Produced by Patrick McLaughlin and Jeff Holmes, Mercatus Center at George Mason University, April 1, 2015
REGDATA.ORG
MERCATUS CENTER
George Mason University

Figure 1:

Executive Orders & Statutes Requiring Review

1. Presidential Executive Order on Reducing Regulation and Controlling Regulatory Costs (EO 13771) requires that,
   - "[For] every new regulation issued, at least two prior regulations are to be identified for elimination…"
2. Section 6 of Executive Order 13563 requires agencies
   - conduct periodic, retrospective review and analysis of existing regulations that "may be outmoded, ineffective, insufficient, or excessively burdensome, and to modify, streamline, expand, or repeal them… so as to make the agency's regulatory program more effective and less burdensome in achieving regulatory objectives."
3. Numerous statutes require agencies to review their rules, such as Title 5, Section 610, which requires:
   - Agencies review all rules promulgated within the previous 10 years

CFR Organization (i.e. the real technical challenge)

Each of the 49 CFR Titles is (pseudo) organized by

- Volume
  - Subtitle
    - Chapter
- Subchapter
  - Part
    - Subchapter
- Part
  - Subpart
    - Subject
- Section

Although, all levels of nesting are optional.

Actual tree may be organized as

- Title
  - Part
    - Section

How to Explore and Analyze the CFR?

Use

- Natural language processing
- Dynamic visualizations
- Auto-CFR aggregator
- Integrate uploads and copy and pasted text, such as
  - small business proposals/plans
  - proposed regulations
  - popular media/transcripts

The CFR Miner

Link to CFR Miner

>The purpose of visualization is insight, not pictures. -Ben Schneidermen, 1999

Before We Get Started: A Primer on Text Analysis Visualizations

Text analysis involves preprocessing, classification, clustering, information extraction, and visualization.

1. After assembling a corpus, the first step to remove stop words (and, the, of) and other words that convey little topic distinctivity (such as *therefore*, *next*, *however*, etc.).
2. The next step is to stem the vocabulary, i.e. trim them to their roots so that words with the same root are combined (fight, fights, fought).
3. The third step is to create n-grams, which are sets of words that co-occur improbably often, and thus denote a single idea (White House, Supreme Court, etc.).

*See Olga Scrivner, David Banks, Julia Silge, and Simone Teufel for great introductory material.

Why We Need Visualizations *i.e. The problem with summary statistics*

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

Figure 3:

Why We Need Visualizations

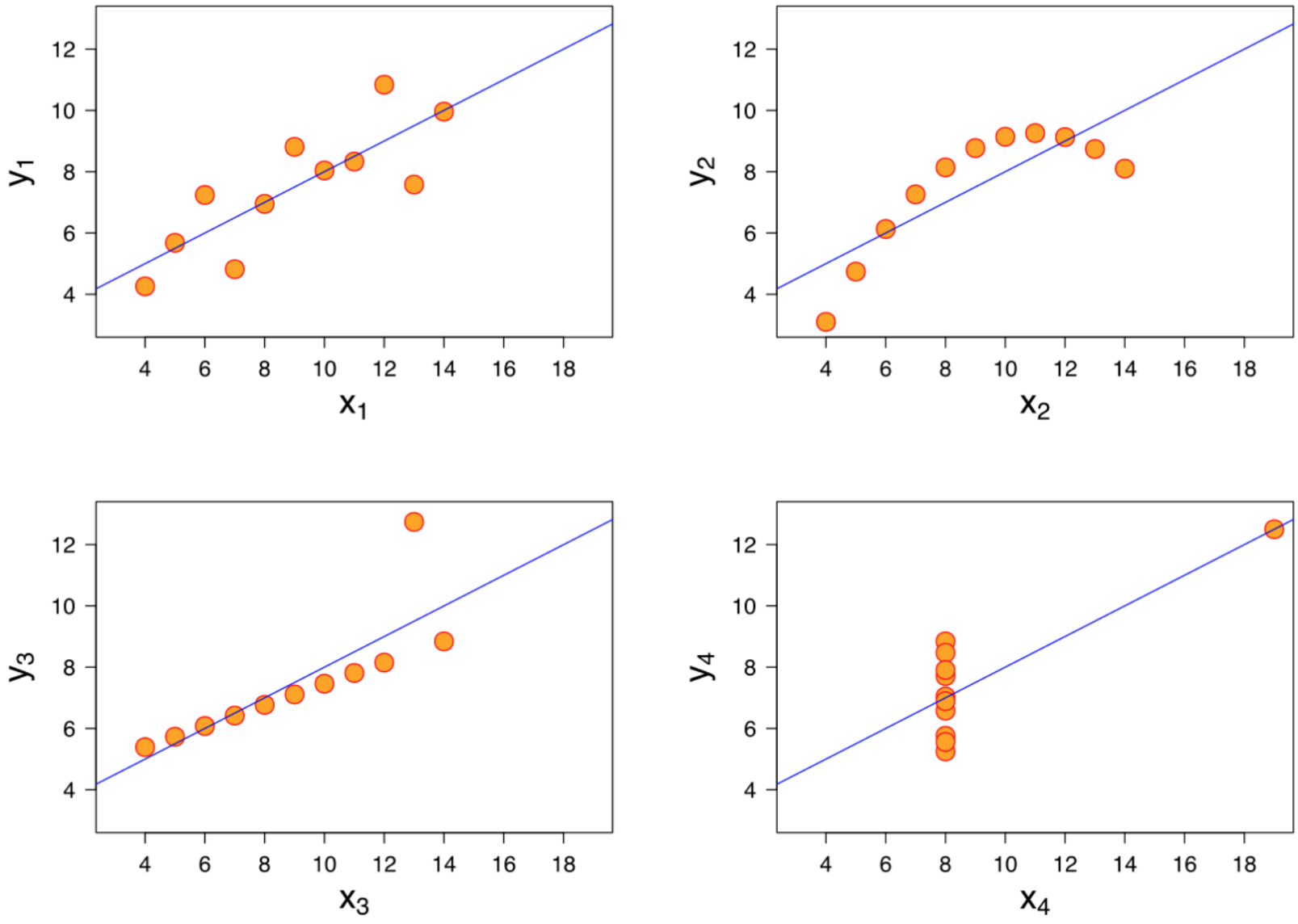

Figure 4:

Importance

Importance is measured by the θ or $tf * idf$. Suppose you are interested in CFR entries related to the word *prescription*:

- $tf_{i,j}$ measures the number of times term i appears in document j. e.g. $tf_{prescription,4} = 10$ means that the word *prescription* appears in document 4 on 10 occasions.
- df_i measures the percent of documents within which the word i appears across the entire corpus (i.e. all documents) of N documents. e.g. $df_{prescription} = 0.25$ means that the term *prescription* appears in 25% of the documents.
- $idf_i = \frac{1}{df}$ is the inverse document frequency. By inverting the percentage, weight is added to words that appear infrequently. e.g. $df_i = 0.25$ means that the tf is weighted by 4.On the other hand, if $df_i = 1$, and the term appears in all documents, the tf is weighted by 1.
- $\theta_{i,j} = \frac{tf_{i,j}}{df_i}$ provides measure of the importance of a word in

Zipf's Law

- Zipf's Law says that the i^{th} most frequent term in a given language, or set, has a frequency proportional to $\frac{1}{i}$.
- If the most frequent term *the* occurs 100 times, then the second most frequent term *of* should appear 50 times, and, the third most frequent, *and*, should appear 25 times, etc.
- The rule supposedly applies to
  - Notes in musical performances
  - Frequency of access to web pages
  - Income distributions among top earning 3% individuals
  - Korean family names
  - Size of earth quakes
  - Word senses per word
- While the law doesn't seem to hold very well for many English texts, it works quite well for regulatory texts.

Cosine Distance

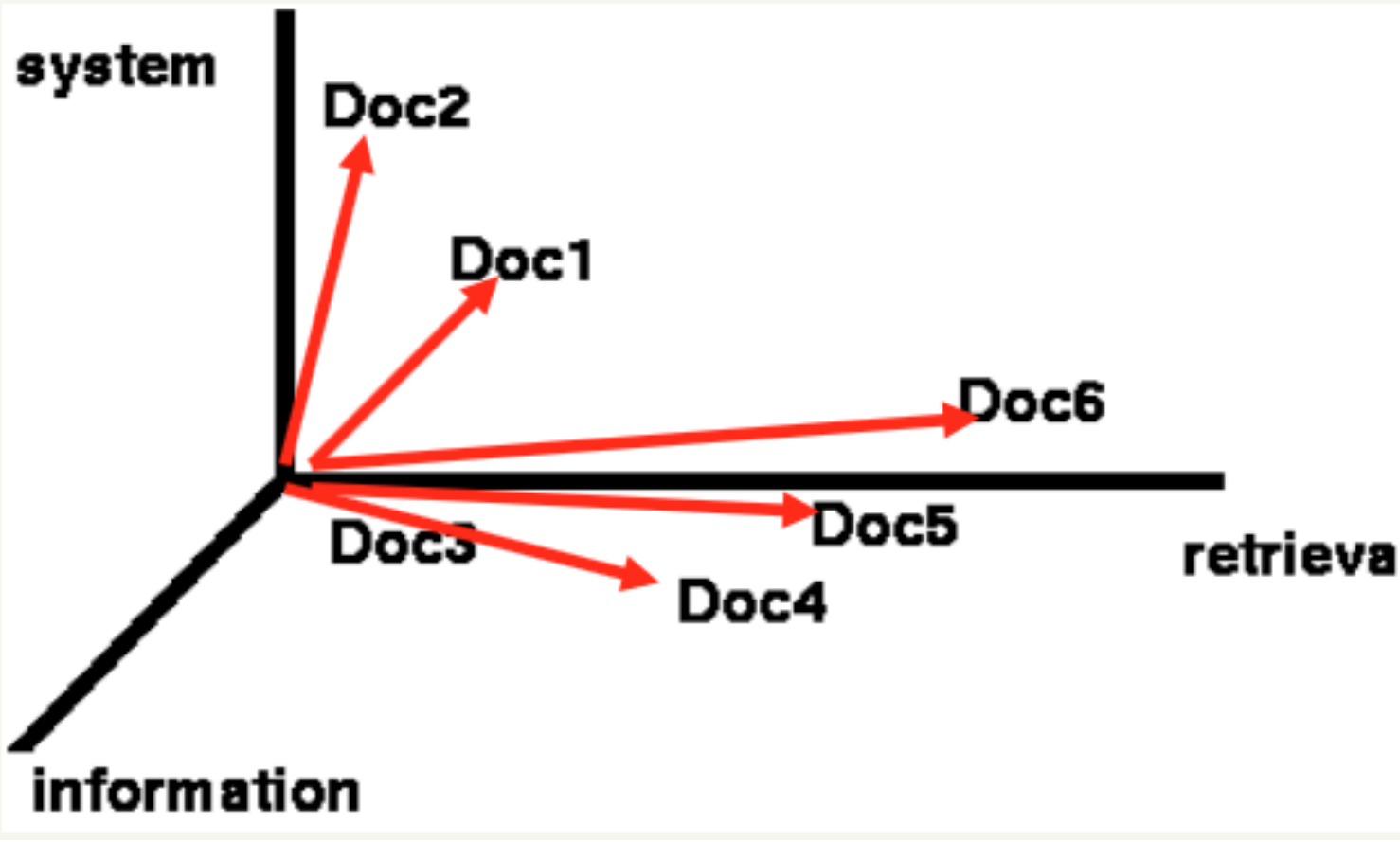

Figure 5:

Cosine Distance

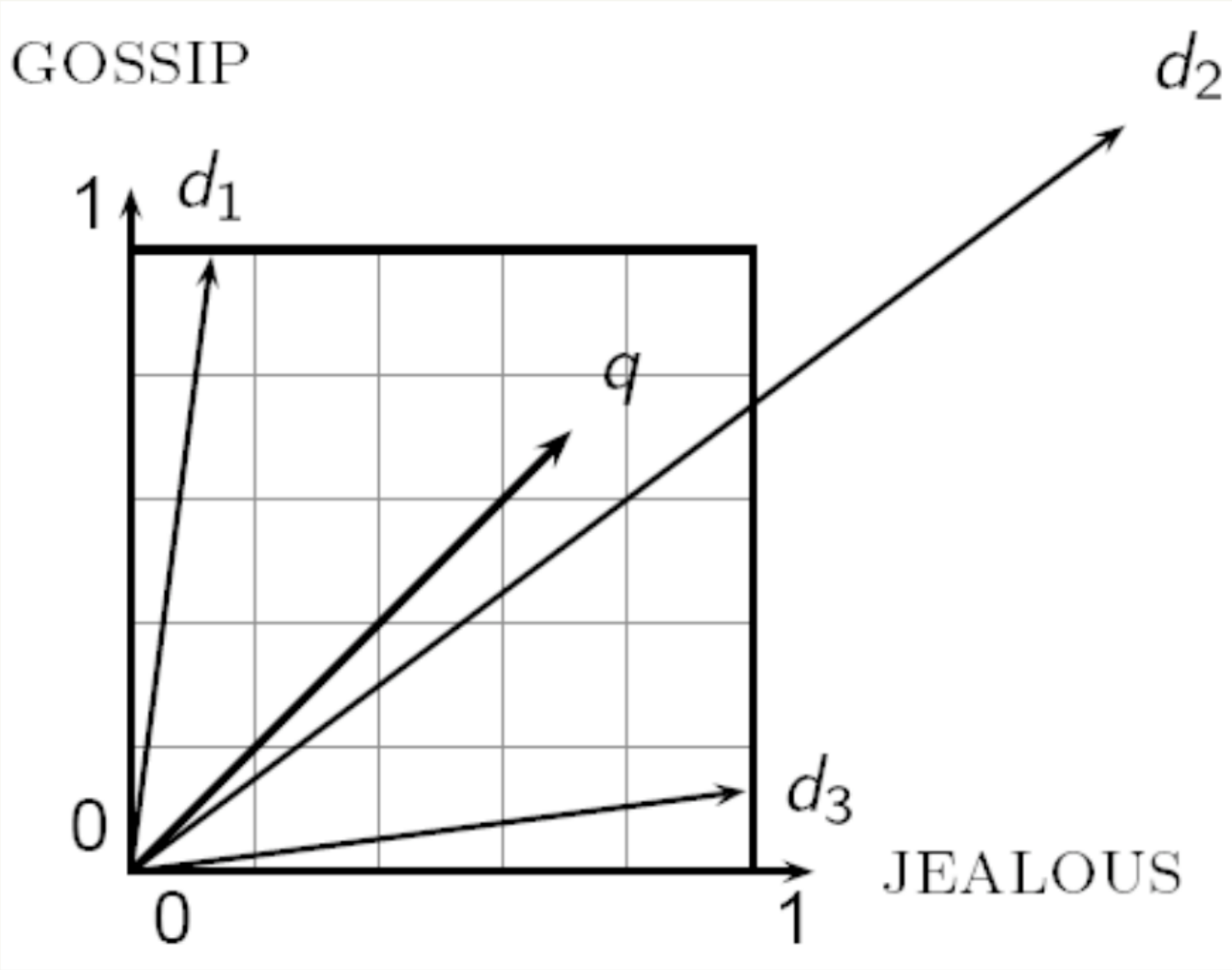

Figure 6:

Demo

Link to CFR Miner

Demo Style 1

- Explore
  1. Force vector maps and collapsible trees reveal structure and accentuate the interconnections between rules
- Preperation
  2. Solving the problem of granularity with automated aggregator
- Analysis
  3. Summarization algorithms provide succinct k-sentence summaries of any granularity of CFR content ranging from paragraphs to volumes https://www.positive.news
  4. Users can upload, copy & paste, or provide links to their own content, such as business plans or proposed rules, to identify related CFR entries
- Related Work
  5. Collaboration with the OIRA / NIST on IBRs
  6. Collaboration with NIST and ITC programs on TBT

Demo Style 2

- *Sally the Rule Writer* is tasked with reviewing Title 1, Part 457.[1] She would like to search for related rules promulgated by other agencies.[2]
- *Ralph the Researcher* is labor economist who would like to search the IRS code for information on the treatment of earnings of workers with and without dependents.
- *Contessa the Small Business Owner* would like to identify the CFR sections most related to the topics found in her business plan.

[1] Volume 1, Chapter 4...

[2] Names borrowed from W. Liberante.

Next Projects: Text Summarization of the FR Stream

Sumblr: Continuous Summarization of Evolving Tweet Streams

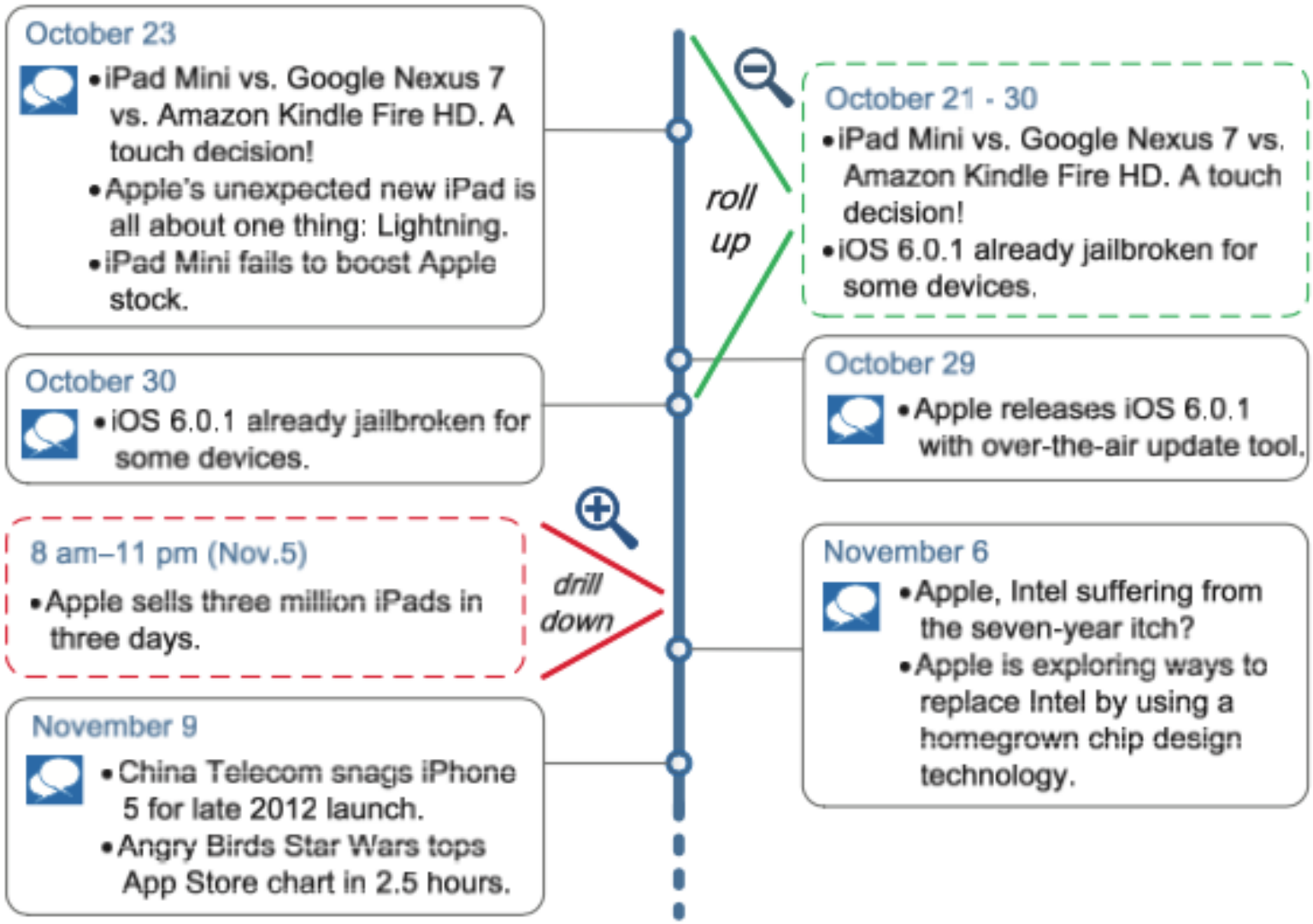

Figure 1: A timeline example for topic "Apple"

Other Updates

- Parallel processing is straight forward and basic tests show that it already promises a 7-fold speed improvement.
- GPU processing could realistically increase processing speeds 1000-fold+ although it is much more difficult to implement.
- Topic modelling using LSA/LDA is implemented at the command-line level but I'm still working on making the topic descriptions either graphically or semantically interesting.[3]
- Sentiment analysis plots are implemented but not obviously useful.

[3] I'd like to generate simple sentences that highlight the highly weighted words.