Module #9b:
Visualization – Text as Data

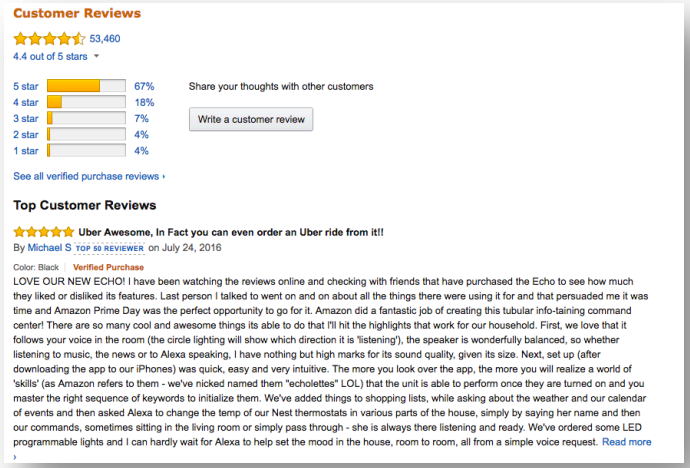
Visualization Techniques

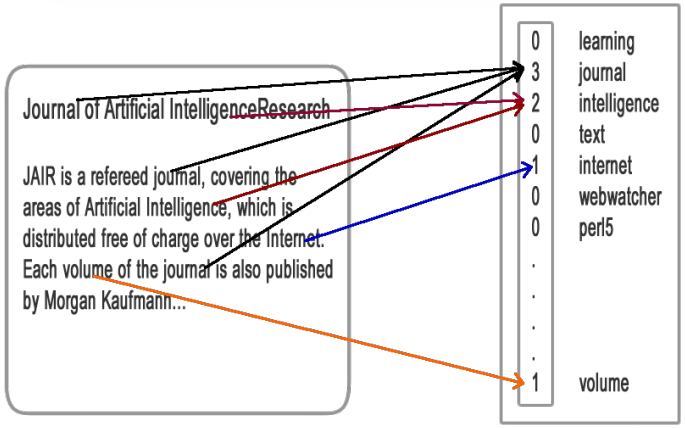
1. Text as Data
2. Visualizing Document Content
3. Evolving Documents
4. Visualizing Conversation
5. Document Collections

1. TEXT AS DATA

Bag of Words Model

- Ignore ordering relationships within the text
- A document ≈ vector of term weights
- Each dimension corresponds to a term (10,000+)
- Each value represents the relevance (e.x. simple term counts)
- Aggregate into a document-term matrix
- Document vector space model





Word Count

http://wordcount.org

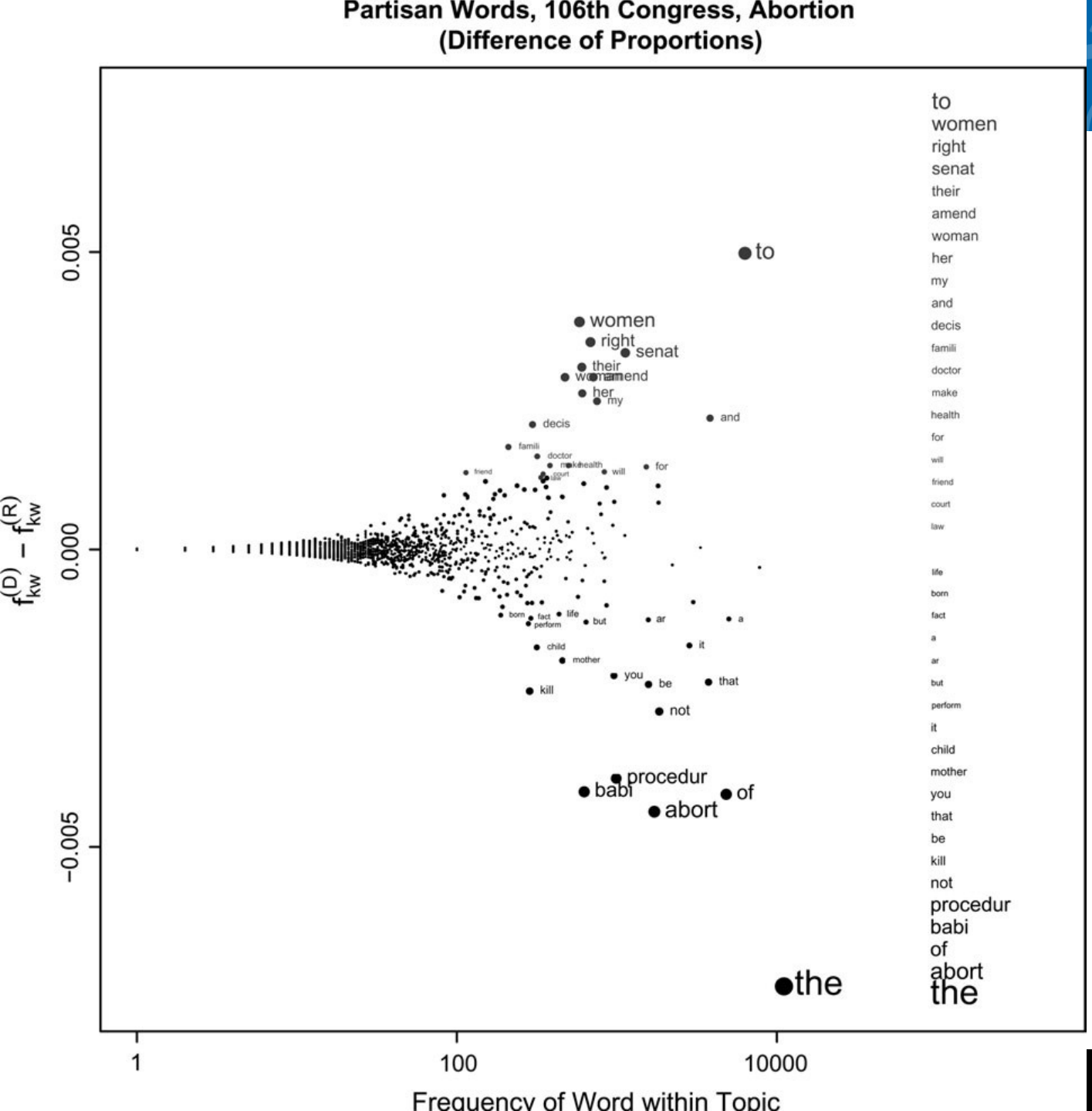
WordCounts (Harris '04)

Visualizations : Wordle of Sarah Palin RNC 9/3/2008 Speech

Creator: Anonymous
Tags:

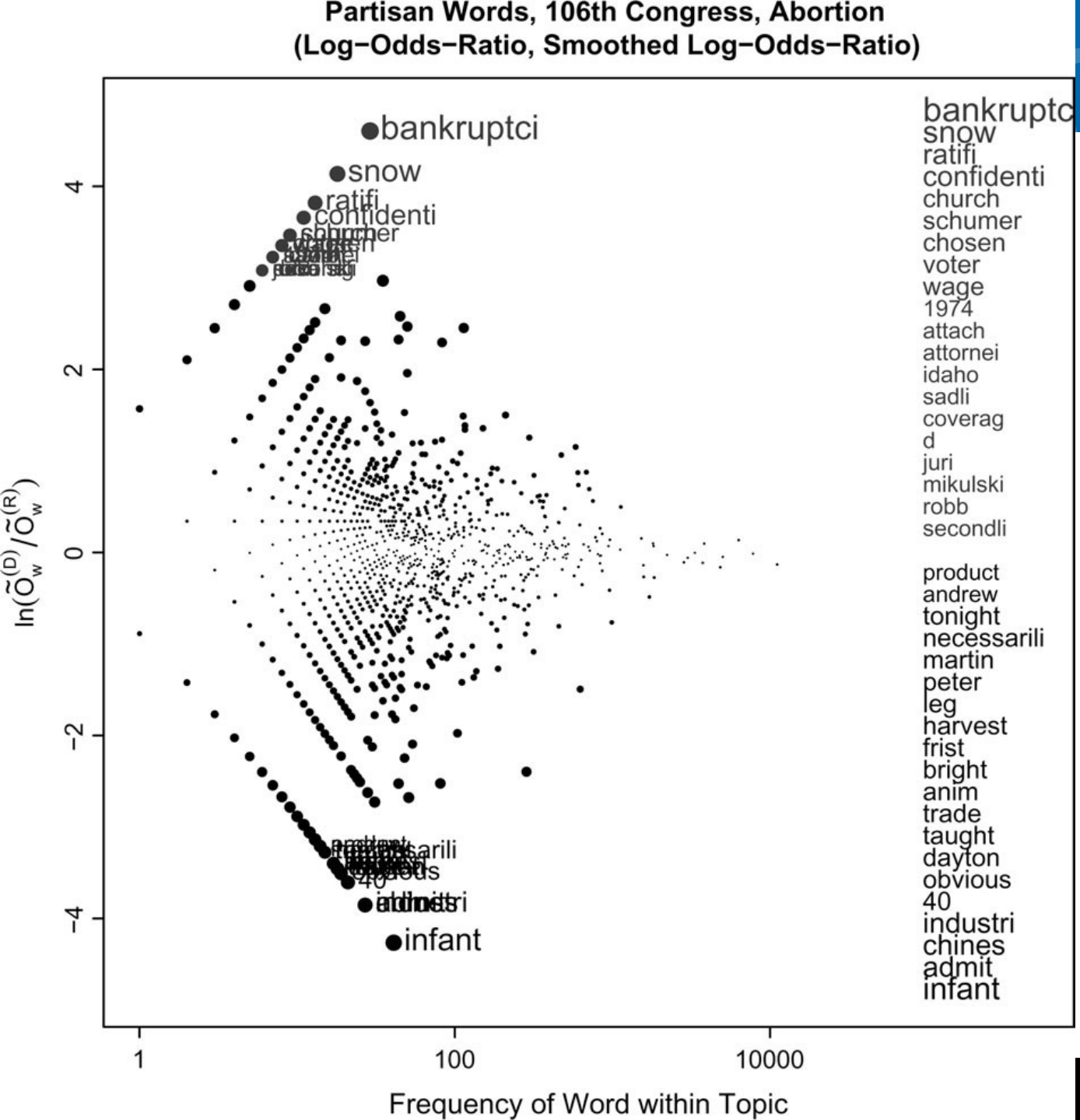
Edit Language Font Layout Color

Data file: Sarah Palin speaks at the Republican National Convention, 9/3/2008 Data source: SFGate / AP This data set has not yet been rated



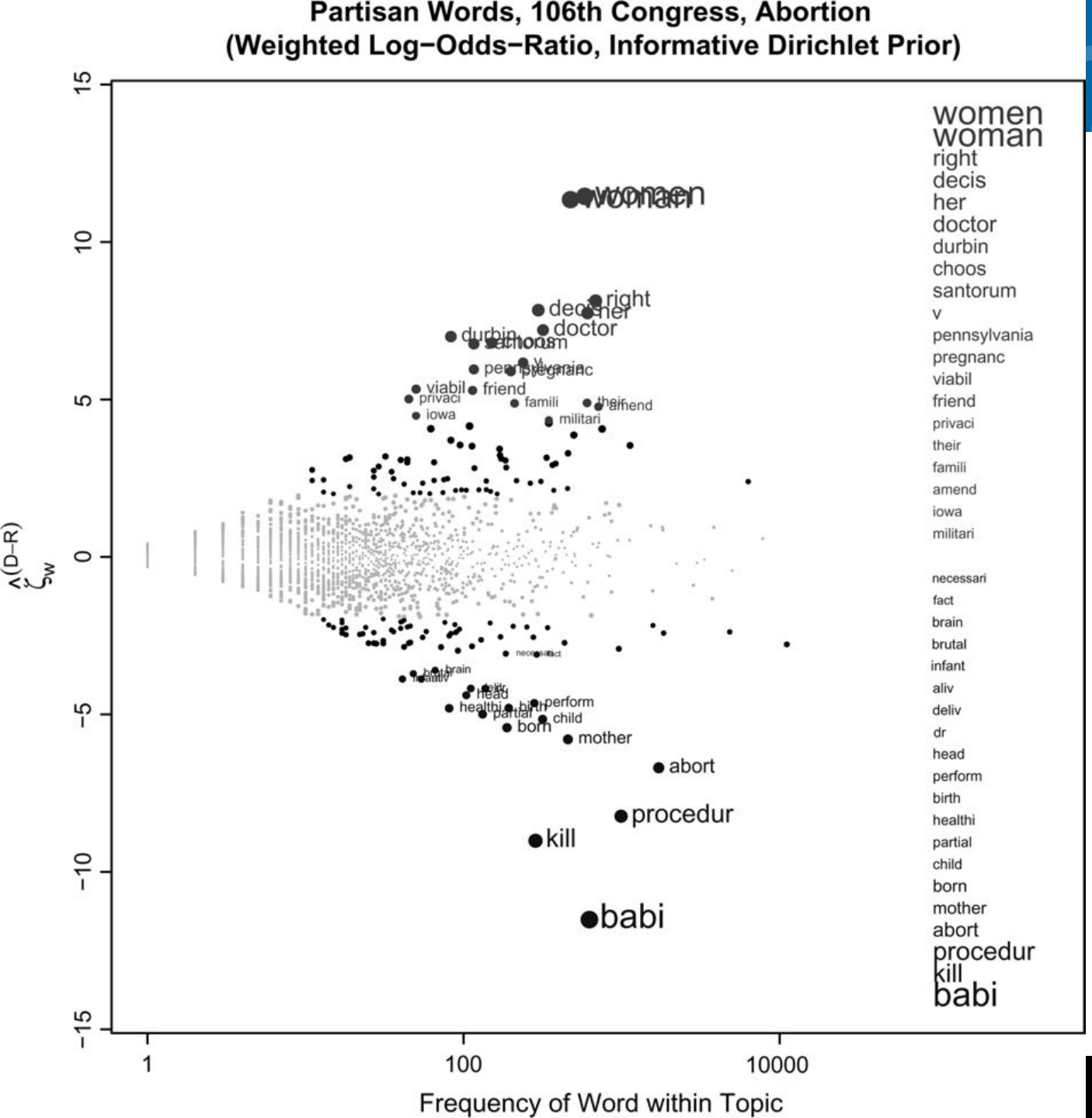
Senate speech from 1997 to 2004

Burt L. Monroe et al. ' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Polit Anal 2008; 16 (4): 372-403.



Senate speech from 1997 to 2004

Burt L. Monroe et al. ' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Polit Anal 2008; 16 (4): 372-403.



Senate speech from 1997 to 2004

Burt L. Monroe et al. ' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Polit Anal 2008; 16 (4): 372-403.

Keyword Weighting

Term Frequency

- tftd = count(t) in d
- Can take log frequency: log(1 + tftd)
- Can normalize to show proportion: tftd / Σt tftd

Term frequency–inverse document frequency (tf-idf)

- Reduce the weight of terms that occur very frequently in the document(s) and increase the weight of terms that occur rarely.
- Example: the vs. healthcare

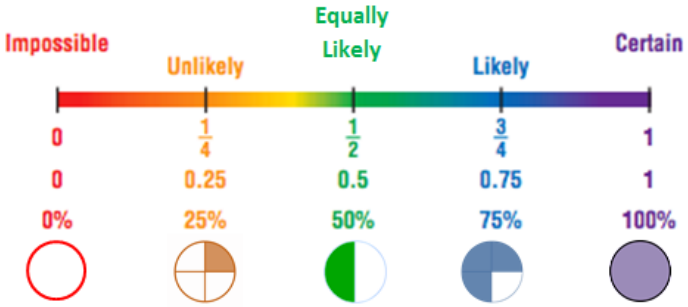
- TF.IDF: Term Freq by Inverse Document Freq

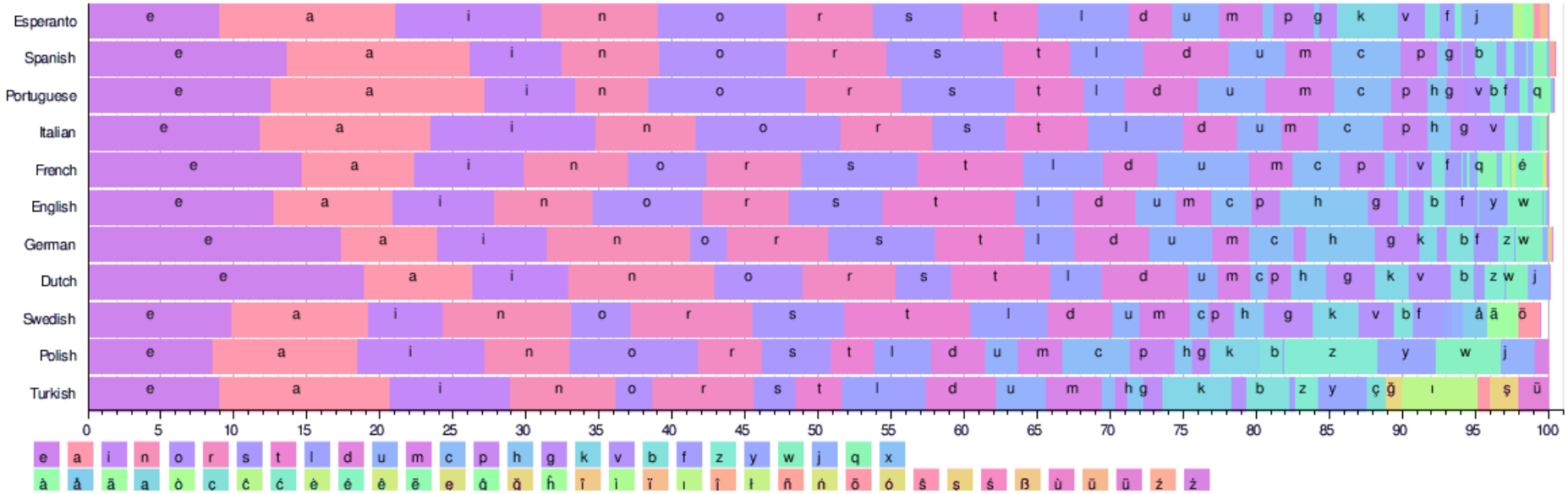
  tf.idftd = log(1 + tftd) × log(N/dft)

  dft = # docs containing t; N = # of docs

Keyword Weighting

- G2: Probability of different word frequency



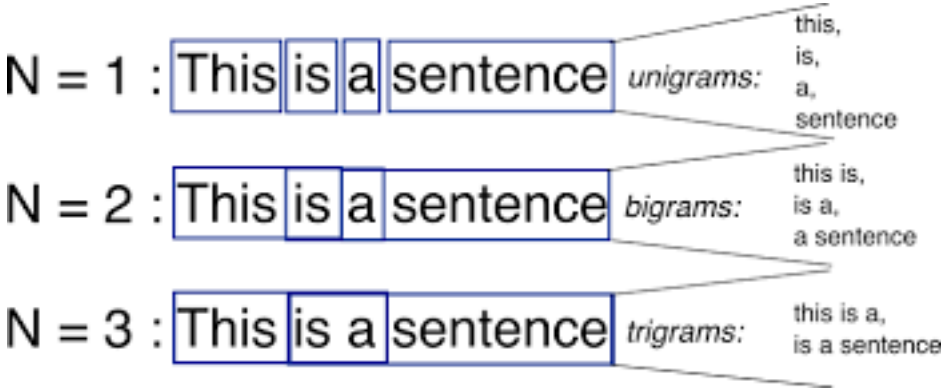


Limitations of Freq. Statistics

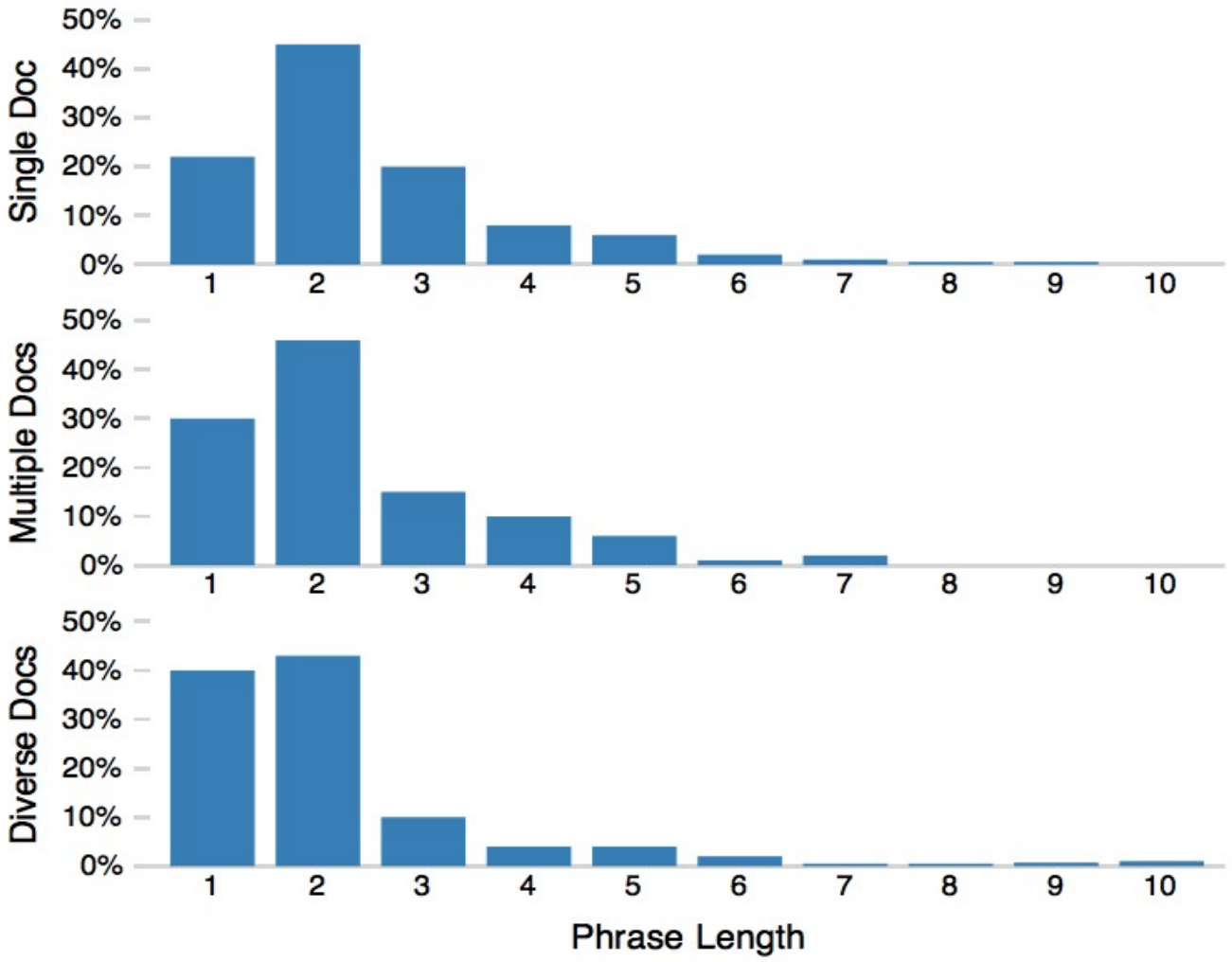
- Typically focus on unigrams (single terms)

- Often favors frequent (TF) or rare (IDF) terms
  - Not clear that these provide best description

- A “bag of words” ignores additional information
  - Grammar / part-of-speech
  - Position within document
  - Recognizable entities



N-grams

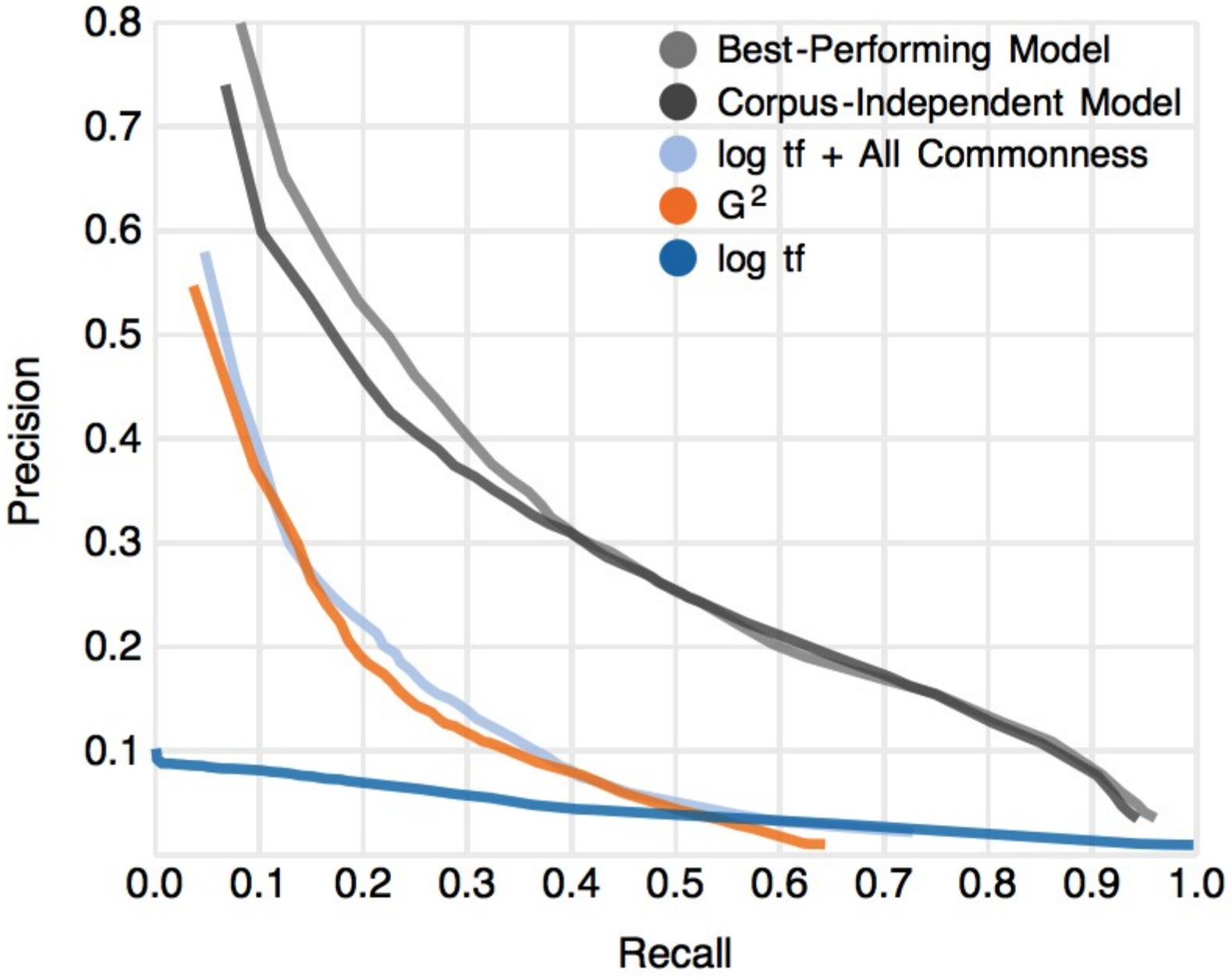

Single Doc
Multiple Docs
Diverse Docs
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
Phrase Length

Term Commonness

$$\log(tf_w) / \log(tf_{the})$$

- The normalized term frequency relative to the most frequent n-gram, e.g., the word “the”.
- Measured across a corpus or across the entire English language (using Google n-grams)

Scoring Terms with Freq, Grammar & Position



A fighter jet rain check

Story and video by Chamila Jayaweera

Have you ever thought about what it takes to make sure that sea-based fighter jets stay dry?

When it comes to the F/A-18 Super Hornet, Boeing engineers in St. Louis use a special process called the Water Check Test to rule out areas where moisture could seep into the aircraft and its electronics suite.

Program experts douse the jet with simulated rain at a 15-inch-per-hour rate for about 20 minutes inside an enormous hangar in St. Louis.

CHAMILA JAYAWEERA/BOEING

The Water Check team rolls in a large metal frame, which they affectionately call their "spray tree," over a Super Hornet inside a St. Louis hangar.

"Our ultimate customers are U.S. Navy fighter pilots, and we want to ensure their safety in flight and on the ground, and water-tight integrity of the aircraft also helps increase their effectiveness," said Boeing's Rich Baxter, F/A-18 Super Hornet final assembly manager.

To find out moreabout how the process works and watch the action unfold, click above to see the video story.

G2

Regression Model

fighter
F/A
Hornet
Super
Boeing
-18
rain
St.
jet
Louis
15-inch-per-hour
douse
hangar
water-tight
Check
Baxter
sea-based
aircraft
Rich
seep
click
Navy
sure
Water
moisture
watch
enormous
stay
want

Super Hornet
F/A -18
fighter jet
Boeing engineers
special process
rain check
electronics suite
Program experts
simulated rain
ultimate customers
enormous hangar
water-tight integrity
Rich Baxter
15-inch-per-hour rate
video story
aircraft
U.S. Navy fighter pilots
Super Hornet final assembly manager
U.S.
Navy fighter
fighter pilot
sea-based fighter

Tips for Effective Test-As-Data Visualizations

- Understand the limitations of your language model.
- Bag of words:
  - Easy to compute
  - Single words
  - Loss of word ordering

- Select appropriate model and visualization

- Generate longer, more meaningful phrases Adjective-noun word pairs for reviews

- Show key phrases within source text