Who Wrote This Document?

Authorship Attribution by Computer

Charles Nicholas

Department of Computer Science and Electrical Engineering

Revised November 13, 2014

UMBC

AN HONORS UNIVERSITY IN MARYLAND

Summary

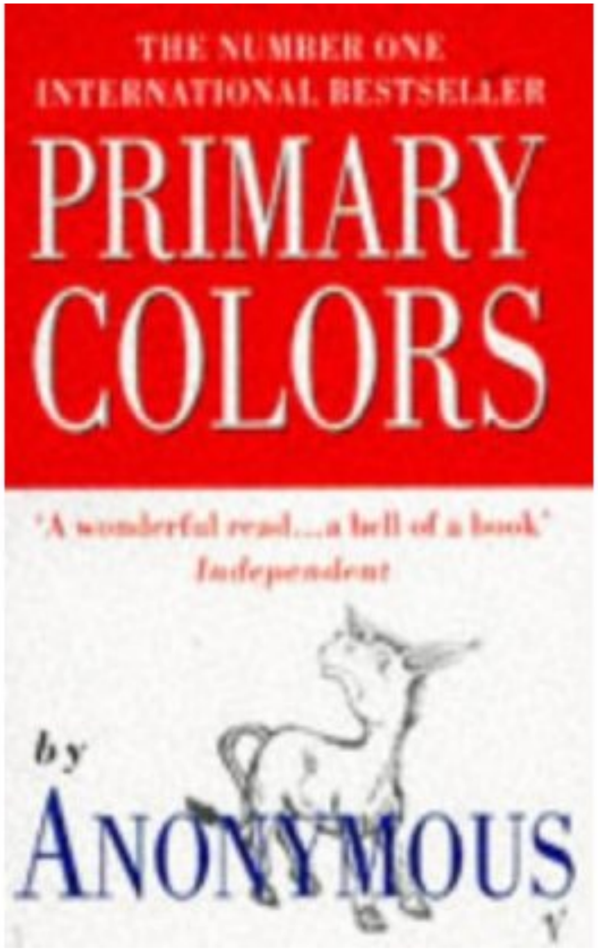
- Authorship questions are fascinating, but often complicated
- Linguistic or stylistic clues have been used for a long time
- Statistical and computer-based methods are now available
- Many questions remain!

Who cares?

- After all, documents usually list their authors
- But sometimes they don’t
- And sometimes they don’t tell the whole truth!

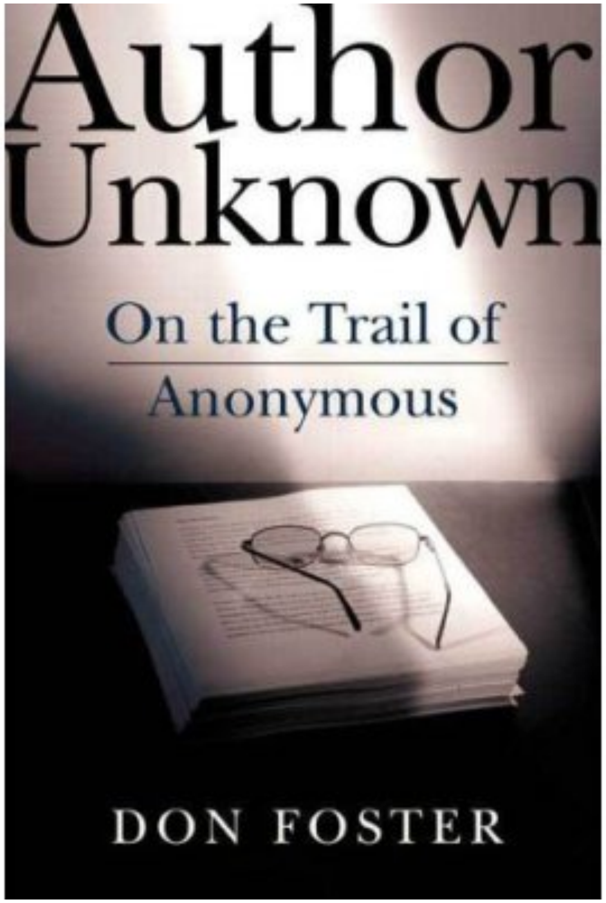
Example:

- The novel “Primary Colors” was in fact written by *Newsweek* columnist Joe Klein
- Professor Don Foster of Vassar College figured this out, and wrote his own book!



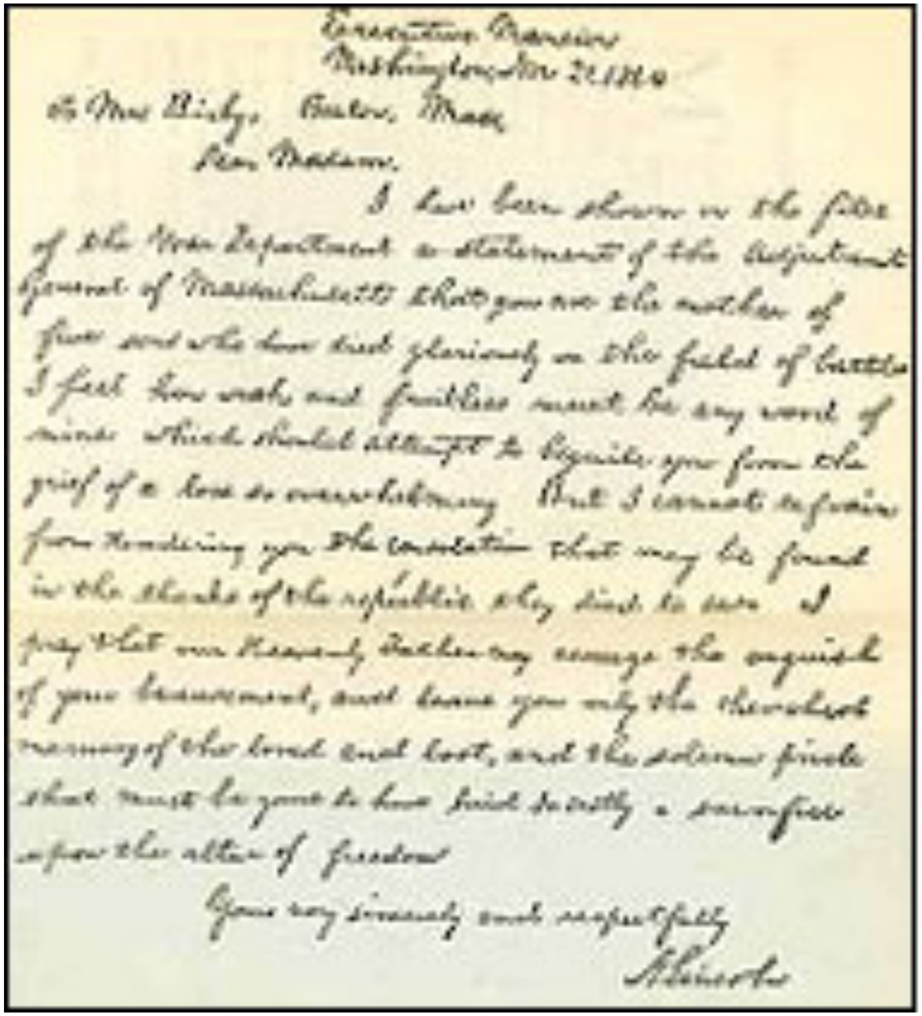
Foster Looks for Clues:

- Words and phrases repeatedly used
- Quirky expressions
- Patterns of punctuation
- Use of quotations
- Foster used on-line databases, (pre-WWW) but his methods were otherwise *not* automated



Lincoln's Letter to Mrs. Bixby

- Mrs. Bixby was thought to have lost five sons in the Civil War
- But maybe Lincoln didn't write this letter!



Executive Mansion
Washington, Nov. 21, 1864

To Mrs Bixby, Boston, Mass.

Dear Madam.

I have been shown in the files of the War Department a statement of the Adjutant General of Massachusetts that you are the mother of five sons who have died gloriously on the field of battle. I feel how weak and fruitless must be any word of mine which should attempt to beguile you from the grief of a loss so overwhelming. But I cannot refrain from tendering you the consolation that may be found in the thanks of the republic they died to save. I pray that our Heavenly Father may assuage the anguish of your bereavement, and leave you only the cherished memory of the loved and lost, and the solemn pride that must be yours to have laid so costly a sacrifice upon the altar of freedom

Yours very sincerely and respectfully
A. Lincoln

NORTHERN ILLINOIS UNIV.

Not So Recent Examples

- The works of Shakespeare
  - Some plays seem to have more than one author!
- From the Christian New Testament
  - Who wrote the Letter to the Hebrews? The letter itself doesn’t say!

How can we tell?

- Given a document, what forms of evidence can we use?
  - Knowledge of people, events or demonstrably earlier documents help us date documents
  - Linguistic evidence, such as vocabulary
  - Statistical evidence, such as consistency with other documents known to be by that author

Vocabulary

- In the Gospel of Mark, the Greek word *euthos* ("immediately") is used much more than in the rest of the NT
- More often than random chance would expect! χ^2=172, significant at p<0.001

	Mark	rest of NT
ευθεως	40	42
other words	11591	128640

One term or many?

- The frequency of a single term may be sufficient to suggest that document X was written by person Y, as in Mark's use of *euthos*
- But the use of many terms is likely to be more convincing

Function Words

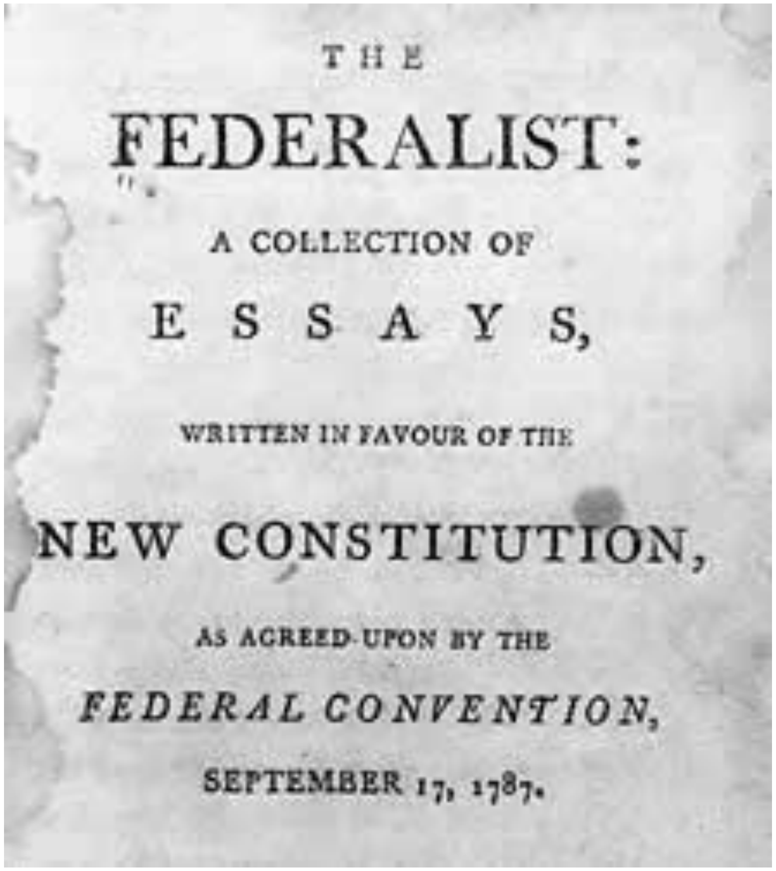
- Function words appear in most if not all documents written in a given language, regardless of topic
- Also known as “stop words” in Information Retrieval (IR)
- Since usage is independent of topic, patterns are likely to indicate authorship as opposed to other characteristics

Function Words Tell Us…

- **Inference and Disputed Authorship,** Mosteller and Wallace, 1964
- Using the Federalist papers as example, demonstrated how frequencies of function words can shed light on authorship questions.

Example: The Federalist Papers

- 85 essays written by James Madison, Alexander Hamilton, and John Jay under the pseudonym “Publius”
- Authorship of 11 has been disputed



THE

FEDERALIST:

A COLLECTION OF

ESSAYS,

WRITTEN IN FAVOUR OF THE

NEW CONSTITUTION,

AS AGREED UPON BY THE

FEDERAL CONVENTION,

SEPTEMBER 17, 1787.

Hamilton appears on the $10 bill

Hamilton appears on the $10 bill

Madison appears on the $5000 bill

Function Words in the Federalist Papers

- Hamilton uses the word “upon” much more often than Madison
- Hamilton uses “while” (in the sense of “at the same time as”) but Madison uses the (chiefly British) “whilst”
- The disputed papers never use “while”, and use “upon” and “whilst” in the same proportion as Madison

Matrix Methods Emerge

- Frequencies of these function words that distinguish one author from another can be analyzed using statistical tests, chi-square for example
- Methods such as singular value decompostion (SVD) and principal components analysis (PCA) can find combinations of terms with such distinguishing power
- Basic data structure is the Term-Document Matrix

Term-Document Matrix

- Create a matrix A, such that entry $a_{i,j}$ is the number of times term i occurs in document j
  - – Terms can be words or n-grams
  - – N-grams are best for noisy and/or multi-lingual
- The TDM is usually sparse; term weighting makes it more so
- Using only function words greatly reduces the rank of the TDM

Kjell and Frieder's Findings

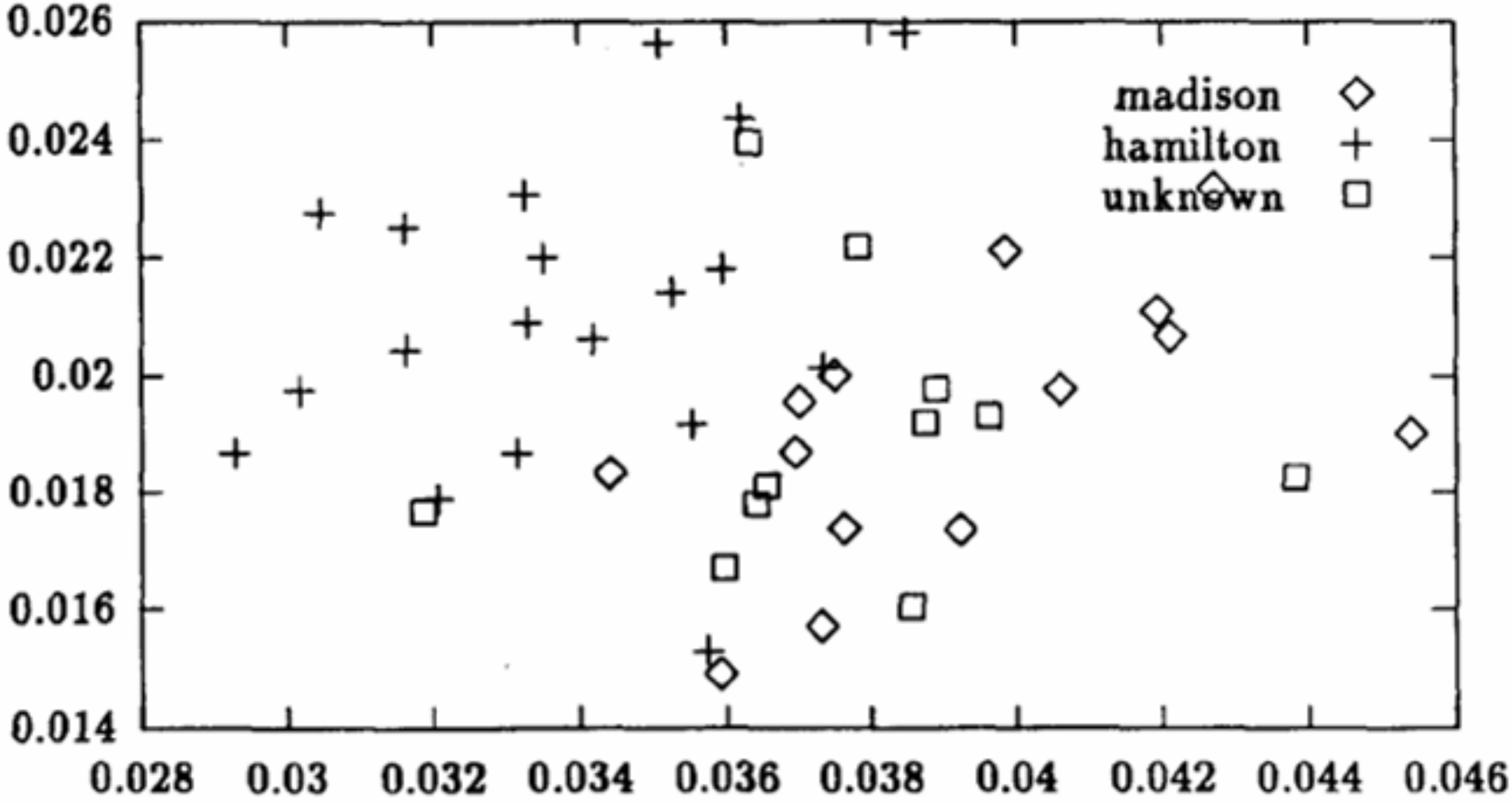


Figure 8: Plot of transformed feature vectors.

Observations on Kjell and Frieder

- The disputed documents are *mostly* in the Madison region, agreeing with other recent scholarship including Mosteller and Wallace
- Kjell and Frieder used a modest amount of data, i.e. the top ten most distinctive 2-grams
- Their analysis was computationally expensive at the time, but nowadays we have other options

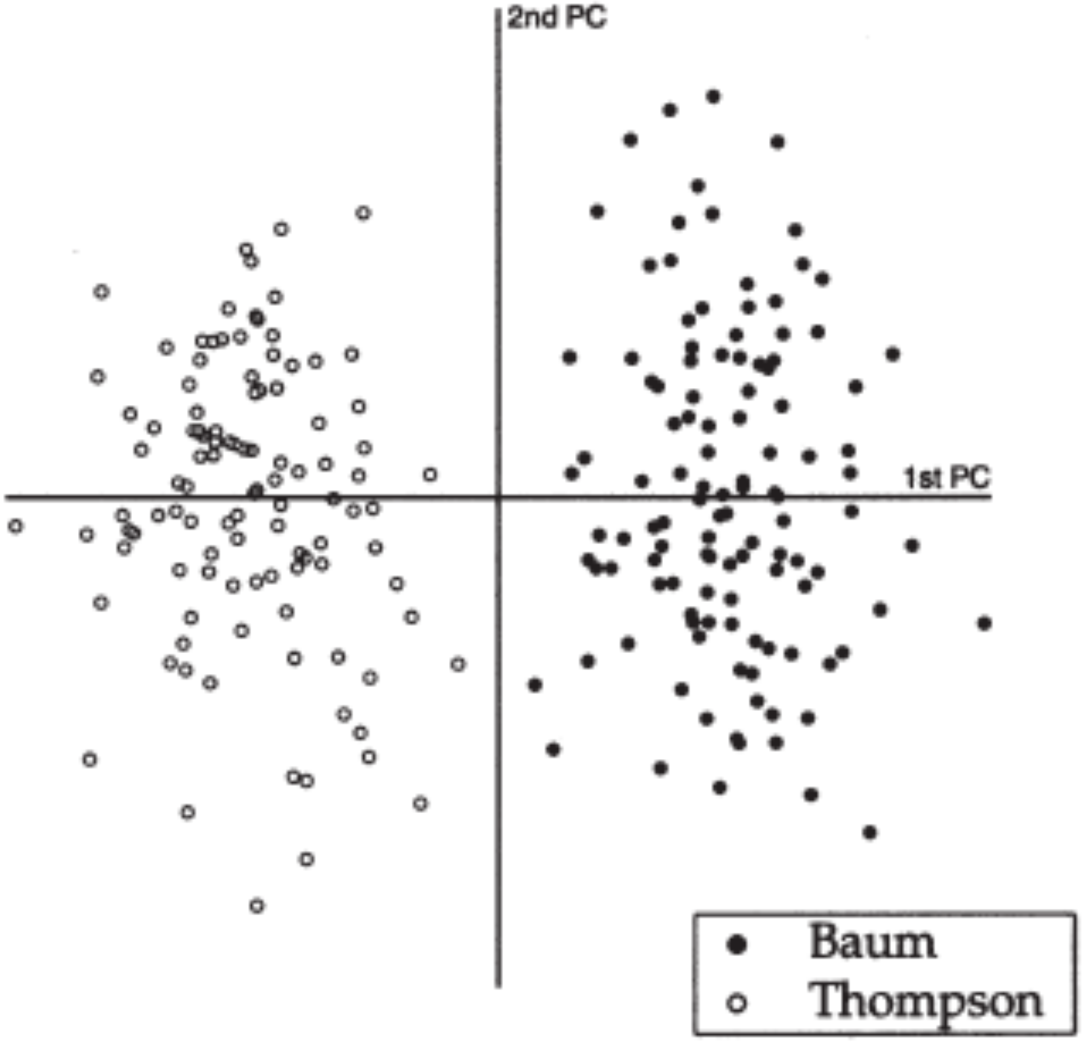
15th book of Oz

- L. Frank Baum created the Wizard of Oz books, and wrote the first 14
- Ruth Plumly Thompson wrote installments 16-31
- The authorship of the 15th book was unclear

Binongo's use of PCA

- José Binongo took the whole Oz corpus, and built a term-document matrix using 223 text segments (documents) and 50 function words as terms
- The resulting matrix was subjected to PCA
- Plotting the data on the space spanned by the first two principal components

Thompson wrote the 15th volume



Singular Value Decomposition

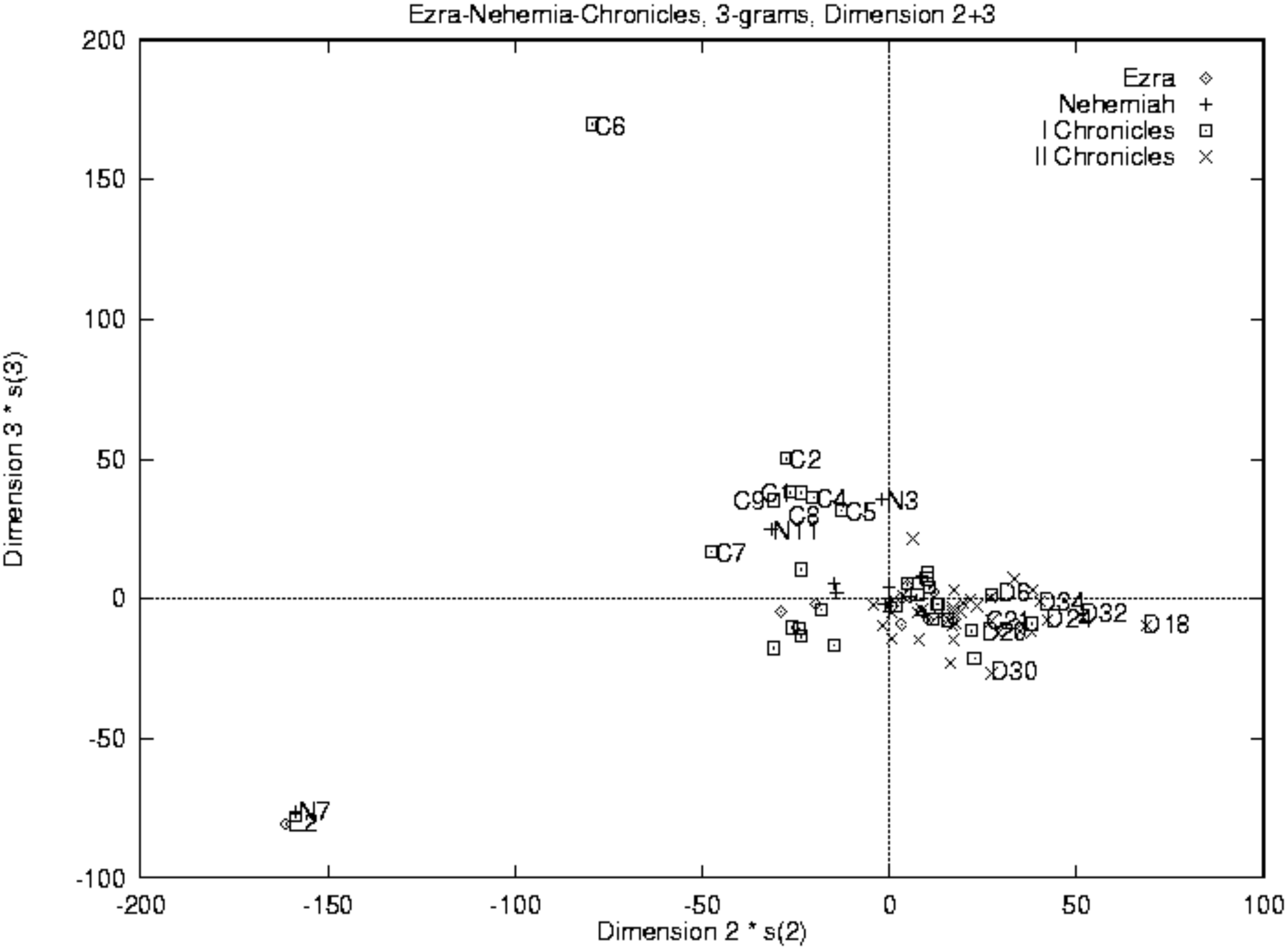
- The SVD is an alternative to Principal Components Analysis
  - Easier to calculate
  - Finds patterns of terms
- Basis for latent semantic analysis used in IR
- Patterns of terms become dimensions in a vector space

Can we spot other characteristics (besides authorship)?

- Soboroff and Nicholas looked at language, genre, and authorship as well as topic
- The SVD identifies patterns in the term document matrix, but the patterns still need interpretation
- Differences in language or dialect really stand out
- Examples from the Hebrew Bible

Ezra, Nehemiah, I and II Chronicles

- Attributed, by tradition, to Ezra
- We built a term-document matrix in which each chapter was a document, and Hebrew 3-grams were tabulated
- The SVD was calculated, and the first dimension (i.e. the X axis) was dominated by Hebrew function words
- So we projected the documents (chapters) onto the Y-Z plane



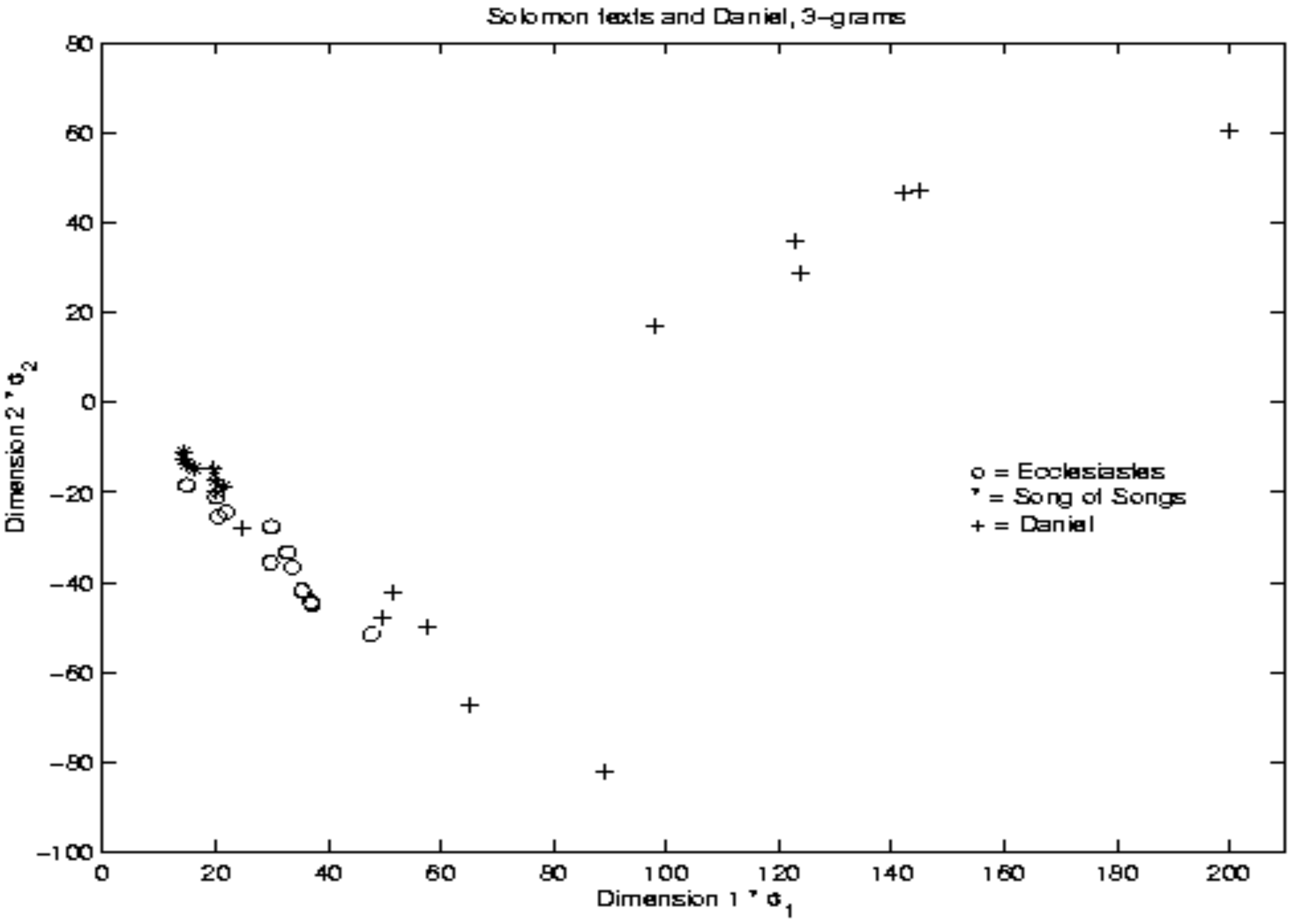
Ezra-Nehemia-Chronicles, 3-grams, Dimension 2+3
Ezra
Nehemiah
I Chronicles
II Chronicles
Dimension 3 * s(3)
Dimension 2 * s(2)
200
150
100
50
0
-50
-100
-200
-150
-100
-50
0
50
100
C6
C2
C9
C1
C4
N3
C8
C5
N11
C7
D6
D34
D32
D18
C21
D24
D20
D30
N7
2

What does this graph say?

- Some chapters, such as Nehemiah 7 and Ezra 2, are different from the rest
  - Most of the text is narrative
  - Ezra 2 is a census, as is Nehemiah 7
- This plot is consistent with the (traditional) hypothesis that these books were written by the same person

Ecclesiastes, Song of Songs, and Daniel

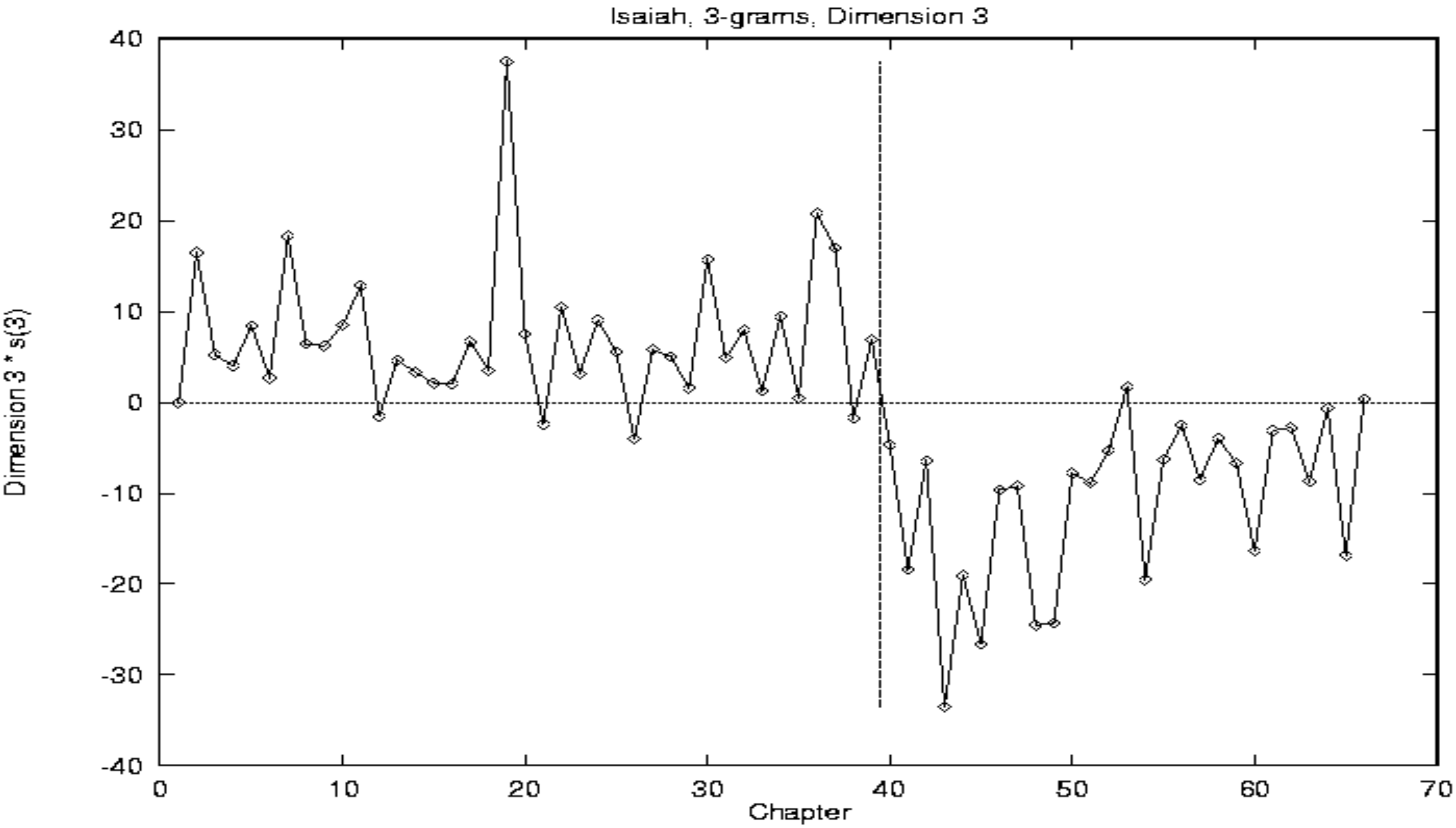
- Ecclesiastes and Song of Songs are traditionally attributed to Solomon, and are poetic in nature
- Daniel dates from much later, and is more narrative (and apocalyptic) in nature
- Modern visualization tools let us squeeze multiple dimensions into a single image


Solomon texts and Daniel, 3-grams
80
60
40
20
0
-20
-40
-60
-80
-100
0
20
40
60
80
100
120
140
160
180
200
Dimension 1 * σ1
Dimension 2 * σ2
o = Ecclesiastes
* = Song of Songs
+ = Daniel

What does this graph say?

- Song of Songs and Ecclesiastes are clustered together, consistent with their poetic nature (and/or Solomonic authorship!)
- Chapters 2-7 of Daniel are in Aramaic!
- Choosing which dimension(s) to look at can be important!

Was there one Isaiah or more?

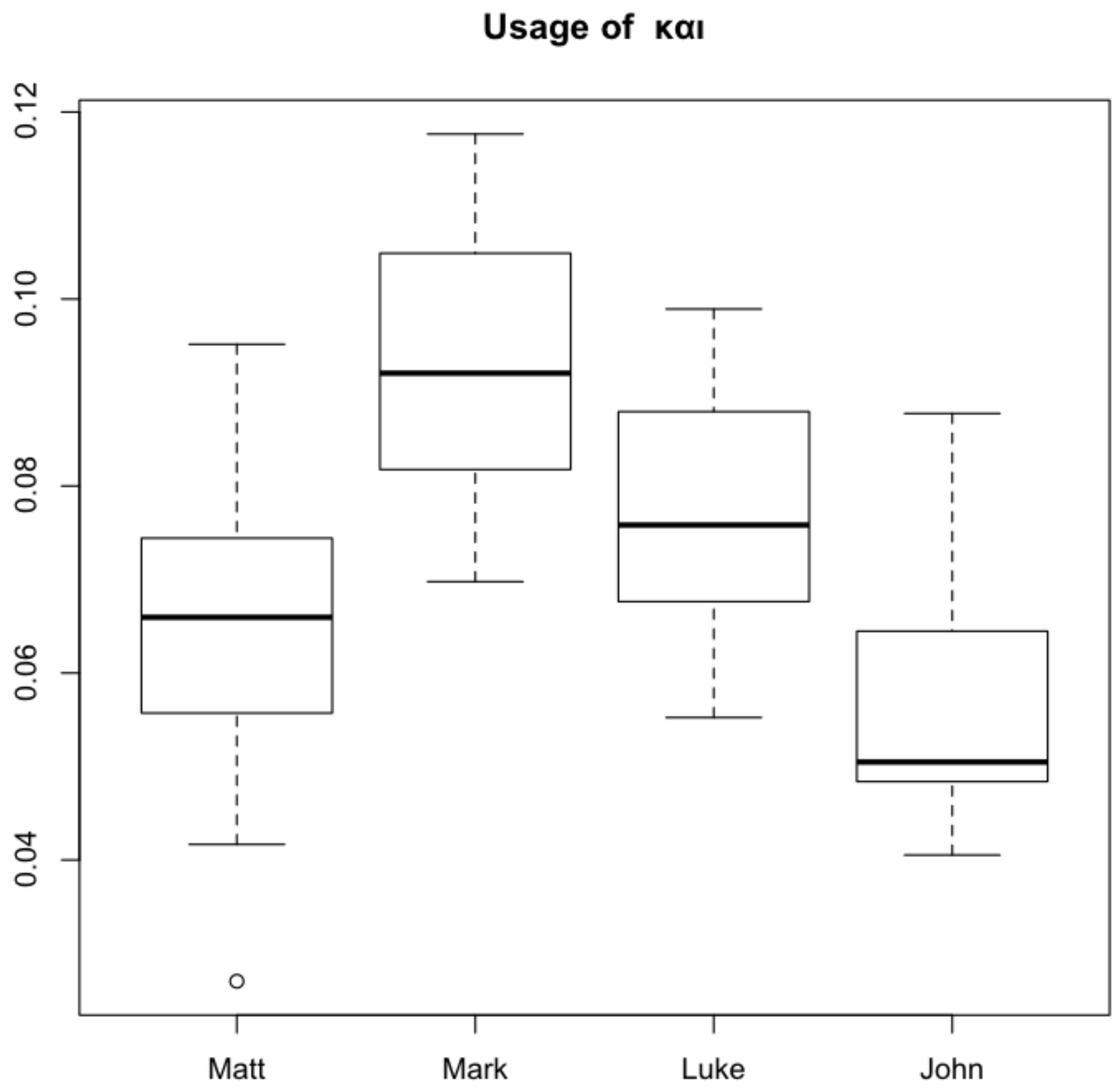


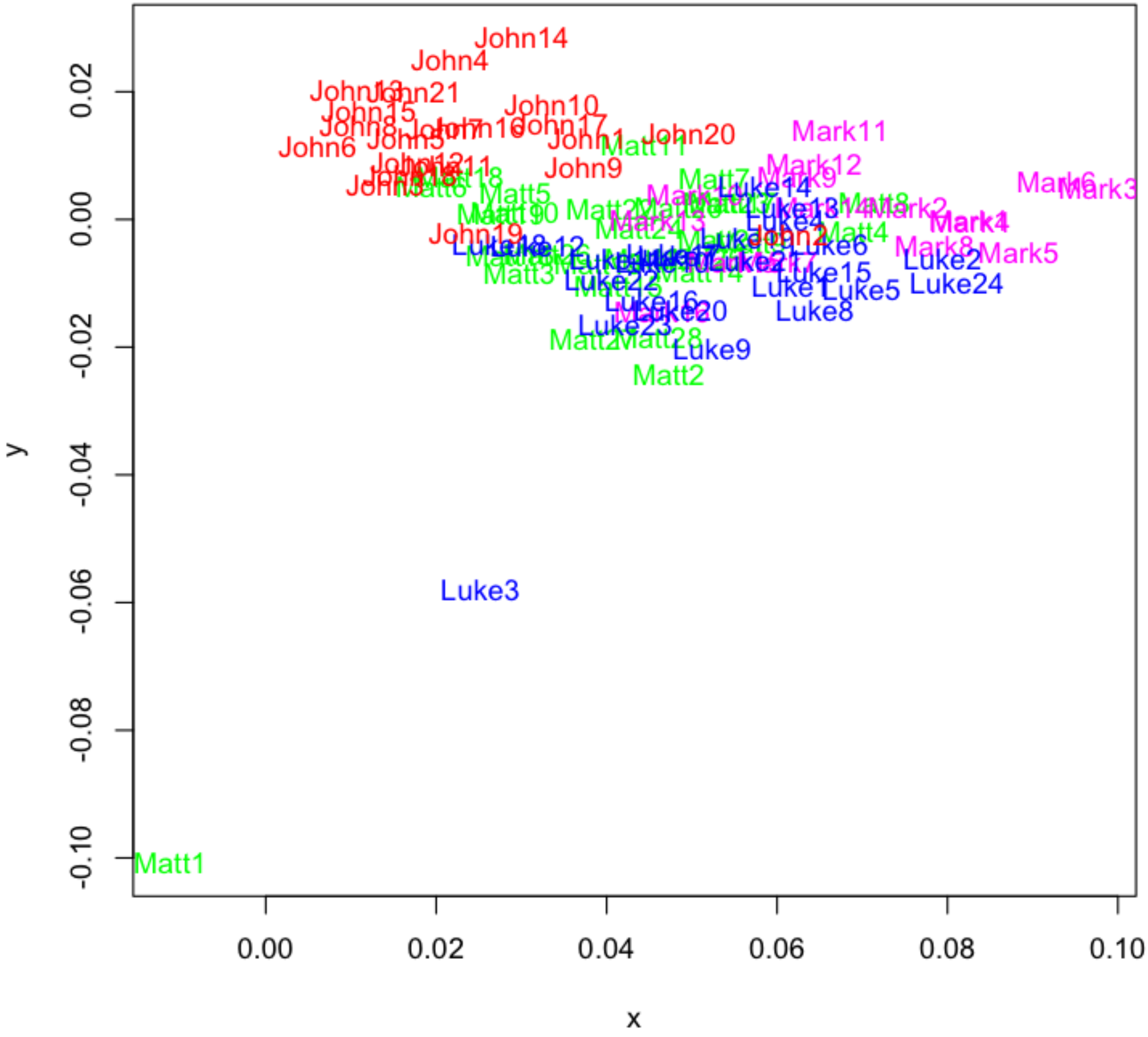
Dimensions of Isaiah

- In a monolingual corpus, the first dimension generated by the SVD will be dominated by function words
- The other dimensions can be inspected to see which terms are occurring together, or not, and in what proportion
- Some “new” pattern starts in Isaiah 40

Visualizing the New Testament

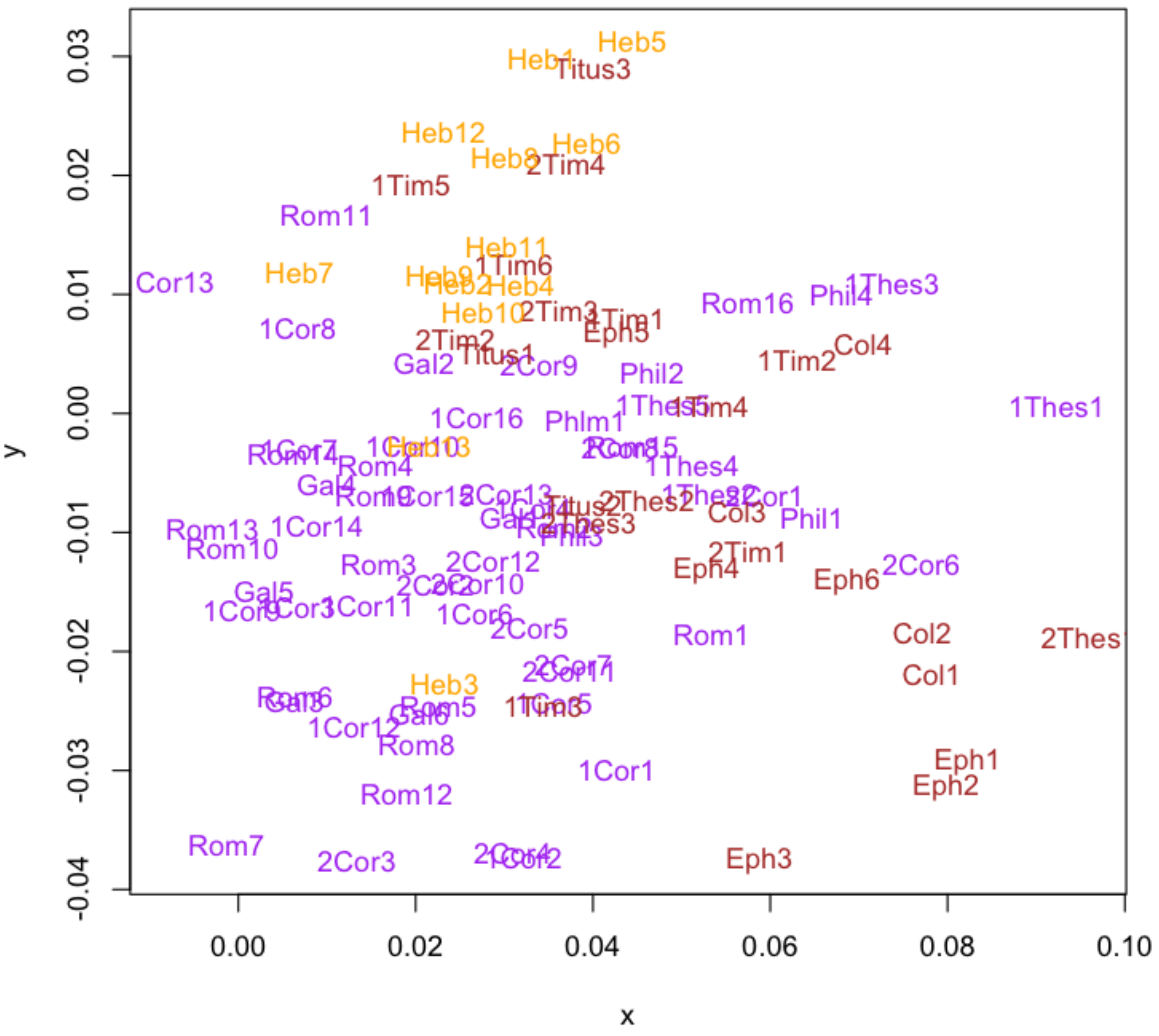
- The “synoptic problem” refers to the relationship between Matthew, Mark, and Luke
- We can build a TDM of the most common words used in 1st Century CE Christian writing
- Kai (‘and’) is by far the most common term in the corpus, but its frequency of use varies significantly (anova F=23.3, p=0)


Usage of καı
0.12
0.10
0.08
0.06
0.04
Matt
Mark
Luke
John


John14
John4
John13
John21
John15
John10
John8
John7
John16
John17
John6
John3
John1
John20
Matt11
Mark11
John12
John11
Matt18
John9
Mark12
Mark9
John18
Matt6
Matt5
Matt7
Luke14
Mark6
Mark3
Matt10
Matt9
Mark10
Luke13
Mark14
Matt8
Mark2
Luke4
Mark4
Mark13
Matt24
John2
Matt4
John19
Luke18
Luke12
Luke6
Mark8
Mark5
Luke17
Luke21
Luke2
Mark7
Matt3
Luke22
Matt14
Luke15
Luke24
Luke1
Luke5
Luke16
Luke20
Luke8
Luke23
Matt28
Luke9
Matt2
Luke3
Matt1
0.02
0.00
-0.02
-0.04
-0.06
-0.08
-0.10
y
0.00
0.02
0.04
0.06
0.08
0.10
x

Paul, and Paul

- Several NT books were undoubtedly written by Paul
  - Romans, 1&2 Cor, Gal, Phil, 1Thes, Phlm
- Some are attributed to Paul, but
  - Eph, Col, 2 Thes, 1 Tim, 2 Tim, Titus
- We don’t know who wrote Hebrews, but Paul is one of several candidates



Heb5
Heb1
Titus3
Heb12
Heb6
Heb8
2Tim4
1Tim5
Rom11
Heb11
1Tim6
Cor13
Heb7
Heb9
Heb2
Heb4
1Thes3
Phil4
Rom16
Heb10
2Tim3
1Tim1
Eph5
1Cor8
2Tim2
Col4
Titus1
1Tim2
Gal2
2Cor9
Phil2
1Thes5
1Tim4
1Thes1
1Cor16
Phlm1
1Cor7
1Cor10
Heb13
2Cor8
Rom15
Rom14
Rom4
1Thes4
Gal4
Rom9
1Cor15
2Cor13
Titus2
2Thes2
1Thes2
2Cor1
Col3
Phil1
Gal1
Col4
2Thes3
Rom2
Phil3
Rom13
1Cor14
Rom10
2Tim1
Rom3
2Cor12
Eph4
2Cor6
Eph6
2Cor2
2Cor10
Gal5
1Cor9
1Cor3
1Cor11
1Cor6
2Cor5
Rom1
Col2
2Thes
2Cor7
2Cor11
Col1
Heb3
Rom6
Gal3
Rom5
Gal6
1Tim3
1Cor5
1Cor12
Rom8
1Cor1
Eph1
Eph2
Rom12
Rom7
2Cor3
2Cor4
1Cor2
Eph3
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
-0.04
y
0.00
0.02
0.04
0.06
0.08
0.10
x

Limits of Existing Approaches

- Traditional methods of literary scholarship, based on history, language, or content, have limits
  - Patterns may defy easy description
  - Larger corpora are difficult
- Statistical evidence needs to be interpreted in light of human understanding of language and history

Research Questions

- Some questions which apply to authorship study:
  - How can we represent features of an author's rhetorical style, as opposed to just vocabulary?
    - e.g. Markan "sandwich"
  - How can we represent what an author knows?
    - e.g. Judges' reference to the (then future) monarchy "In those days Israel had no king, and everybody did as they pleased."

More Research Issues

- How to deal with authorship in large corpora
  - Can we build a search engine that finds documents with vocabulary or writing style similar to a given "query document"?
- How to represent more complicated features
  - Could a search engine find documents that mention first century CE people or events, but not second century?

Zooming Up to Today: Malware Analysis

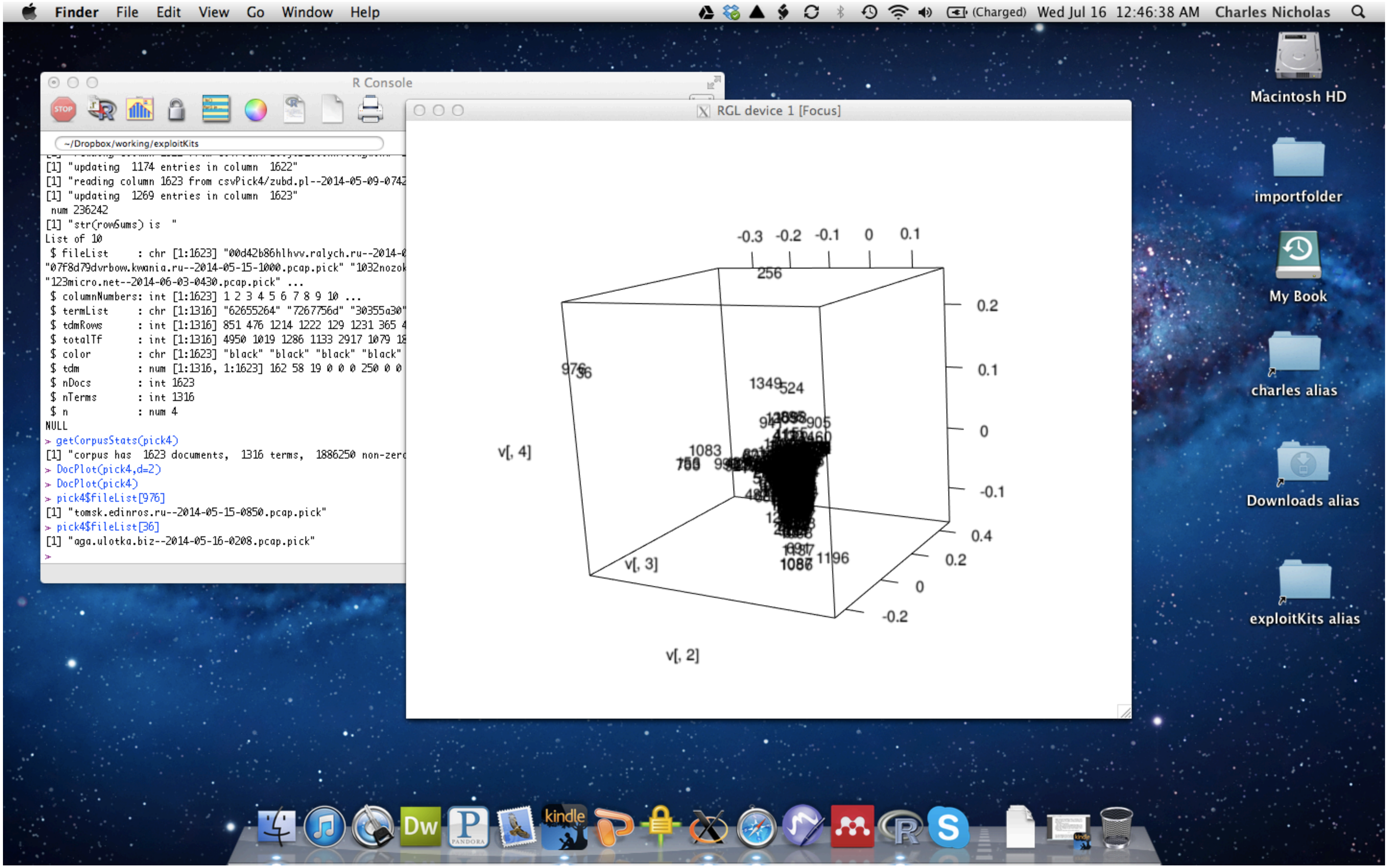
- Can we use techniques like these to figure out who wrote a malware specimen, such as CryptoLocker?
- People are looking at such questions, but so far no easy answers
- We can compare malware specimens, though, using compression. (How?)

Work in Progress

- Can we use compression-based similarity to compare malware specimens? Yes
- But isn’t compression kind of slow? Yes
- Can we cluster small malware collections anyway? Yes

Some Network Traffic

- Exploit Kits are a growth industry
- We have built a data set of TCP/IP sessions
- The raw data was processed through the tcpick utility, and the results were loaded into a TDM as described earlier...
- Ongoing effort sponsored by...



```
[1] "updating  1174 entries in column  1622"
[1] "reading column 1623 from csvPick4/zubd.pl--2014-05-09-0742
[1] "updating  1269 entries in column  1623"
 num 236242
[1] "str(rowSums) is  "
List of 10
 $ fileList     : chr [1:1623] "00d42b86hlhvv.ralych.ru--2014-0
"07f8d79dvrbow.kwania.ru--2014-05-15-1000.pcap.pick" "1032nozok
"123micro.net--2014-06-03-0430.pcap.pick" ...
 $ columnNumbers: int [1:1623] 1 2 3 4 5 6 7 8 9 10 ...
 $ termList     : chr [1:1316] "62655264" "7267756d" "30355a30"
 $ tdmRows      : int [1:1316] 851 476 1214 1222 129 1231 365 4
 $ totalTf      : int [1:1316] 4950 1019 1286 1133 2917 1079 18
 $ color        : chr [1:1623] "black" "black" "black" "black"
 $ tdm          : num [1:1316, 1:1623] 162 58 19 0 0 0 250 0 0
 $ nDocs        : int 1623
 $ nTerms       : int 1316
 $ n            : num 4
NULL
> getCorpusStats(pick4)
[1] "corpus has  1623 documents,  1316 terms,  1886250 non-zero
> DocPlot(pick4,d=2)
> DocPlot(pick4)
> pick4$fileList[976]
[1] "tomsk.edinros.ru--2014-05-15-0850.pcap.pick"
> pick4$fileList[36]
[1] "aga.ulotka.biz--2014-05-16-0208.pcap.pick"
>
```

"I think I've made one of the first steps toward unraveling the mysteries of the Old Testament. . . . I'm starting to read it!"

Selected References

- Applied Bayesian and Classical Inference: The Case of *The Federalist* Papers, Frederick Mosteller and David L. Wallace, Springer-Verlag 1984
- http://www.foundingfathers.info/federalistpapers/
- Who Wrote the Bible?, Richard Friedman, HarperSanFrancisco, 1997
- Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution, Jose Nilo G. Binongo, Chance 16(2) Spring 2003

More References

- Statistics for Corpus Linguistics, Michael Oakes, Edinburgh, esp. Chapter 5, Literary Detective Work
- Analyzing Worms and Network Traffic Using Compression, Stephanie Wehner, J. Comp. Security, 15(3), 2007, 303-320.

Still More References

- An article on the authenticity of Lincoln’s letter to Mrs. Bixby appeared in the January 2006 issue of American Heritage
- Charles M. Schulz, The Complete Peanuts, 1950-1952, Fantagraphics Books, 2004, p. 329

Additonal Slides

The Matrix Approach

- Select subset of document terms to be considered (all words, n-grams, function words, or whatever)
- Build a term-document matrix
- Transform as needed to make any patterns visible
- Figure out what the patterns mean!

Kjell and Frieder on the *FPs*

- Kjell and Frieder chose a set of 10 n-grams that most distinguished the sets of documents with known authorship in a training set
- Two clusters emerged in that term-document matrix, indicating Madisonian authorship of the eleven disputed Federalist Papers
- They used the KL-transform to reduce 10 dimensions to 2

Properties of the SVD

- SVD calculates matrices U, Σ, and V^T such that the term document matrix A = U Σ V^T
- The matrices U and V are *orthonormal*, i.e. the columns form a basis, and each column is length 1
- Complexity of full SVD is $O(n^3)$ for n non-zero entries in the matrix, so sparse is good

Interpreting U, Σ, and V^T

- The columns of U are sets (or patterns) of terms that occur (or not) together.
- The *singular values* are the main diagonal entries in Σ, and they give the relative importance of these patterns
- Entries in the rows of V^T are the coordinates of the documents in the space spanned by the columns of U

Dyadic Decomposition

- We can choose how much of the SVD to do
- For some k >= 1, we can calculate the rank k matrix $A_k \sim U_k \Sigma_k V_k^T$, where we compute only the first “k” of the singular values.
- The matrix A_k is the best (rank k) approximation to the original t-d matrix A.
- Choosing k=2 makes sense for a plot

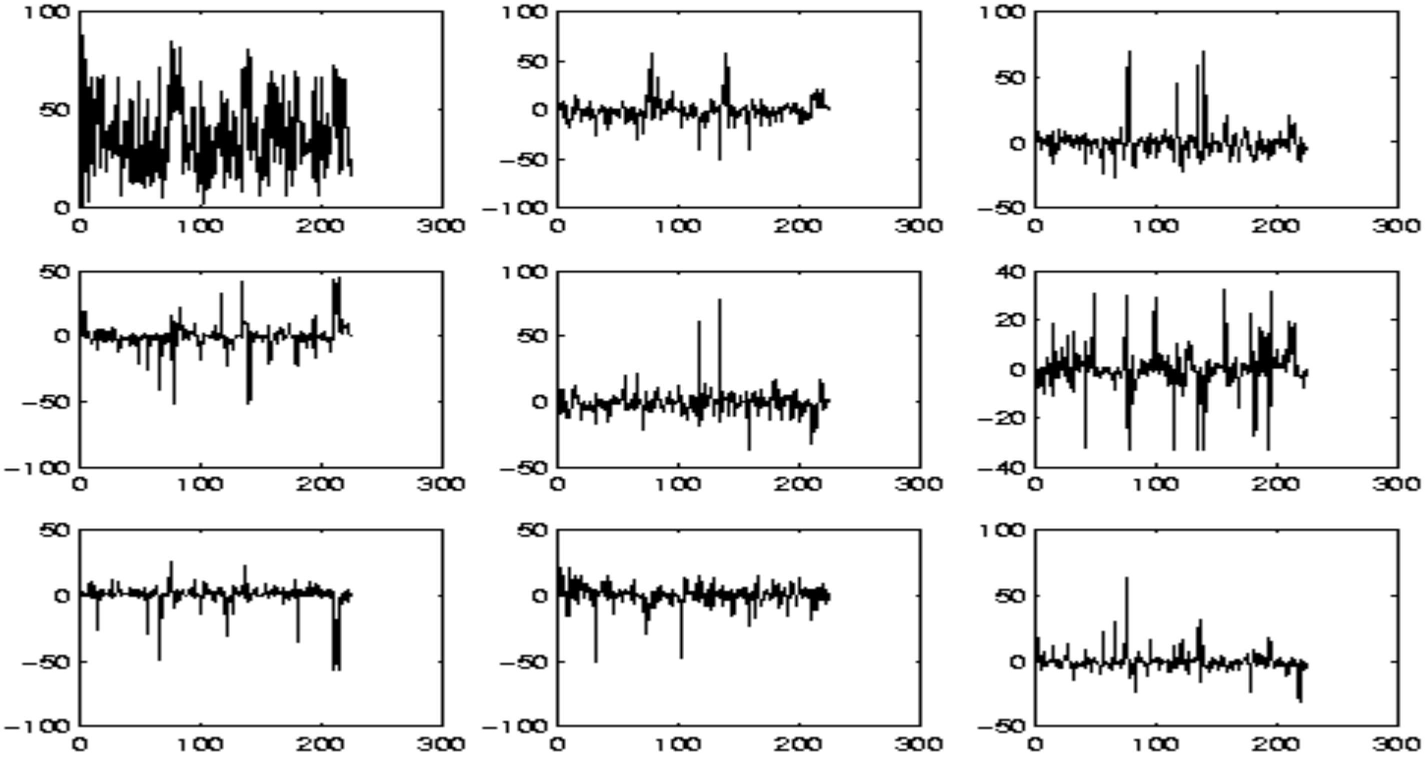
Interpreting U

- Each column U_1, U_2, …, U_k of U represents a pattern of terms that tend to occur together
- Terms common to all documents collect into U_1
- A frequency plot can show these patterns of terms occurrence
- In an AP News corpus, of almost 100,000 terms, a relatively small number really stand out, thereby helping to characterize these term patterns

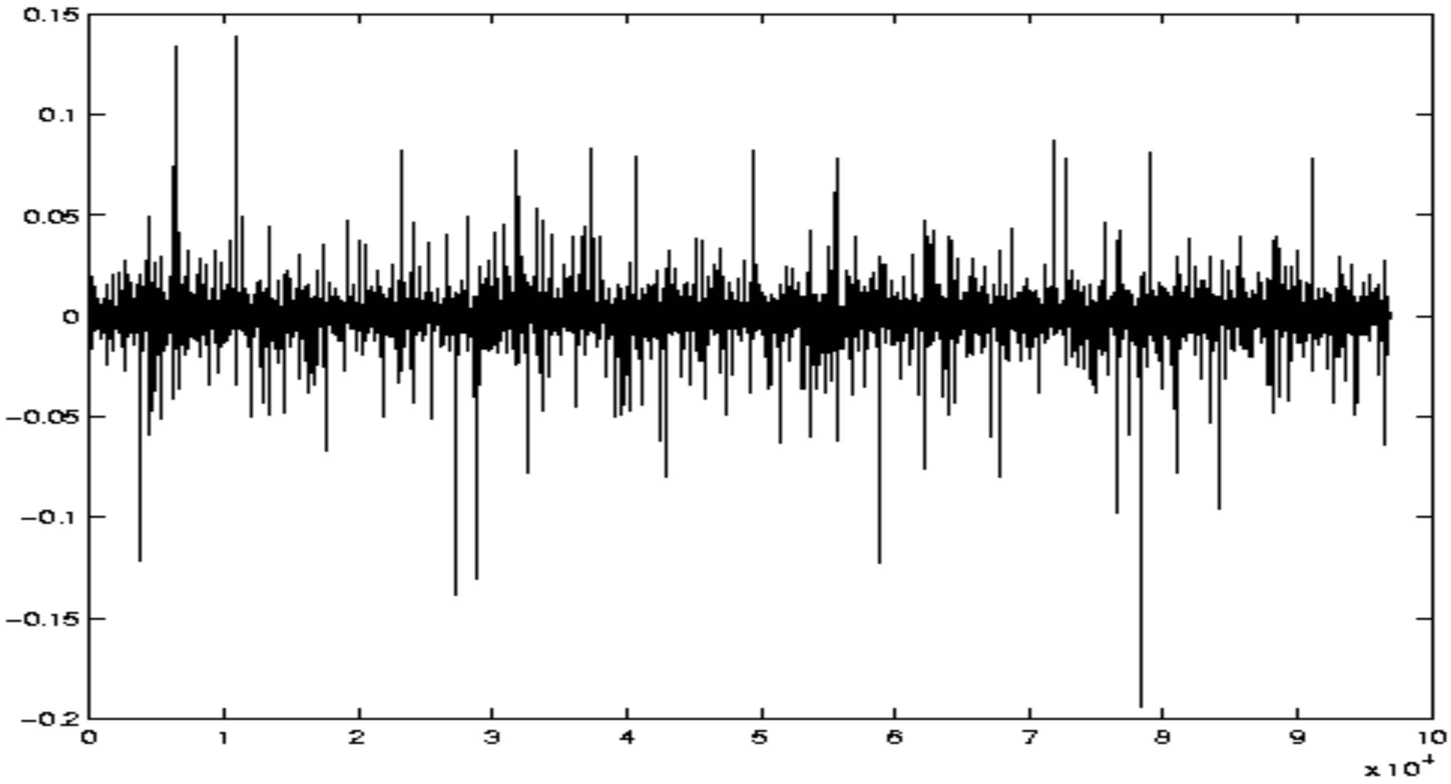
Interpreting V^T

- The columns of U form a basis, and the entries in row *i* of V^T are the coordinates of document *i* in the space spanned by the columns of U
- Documents that have large values in a certain dimension have many instances of the corresponding terms

Example: Coordinates of documents in various dimensions



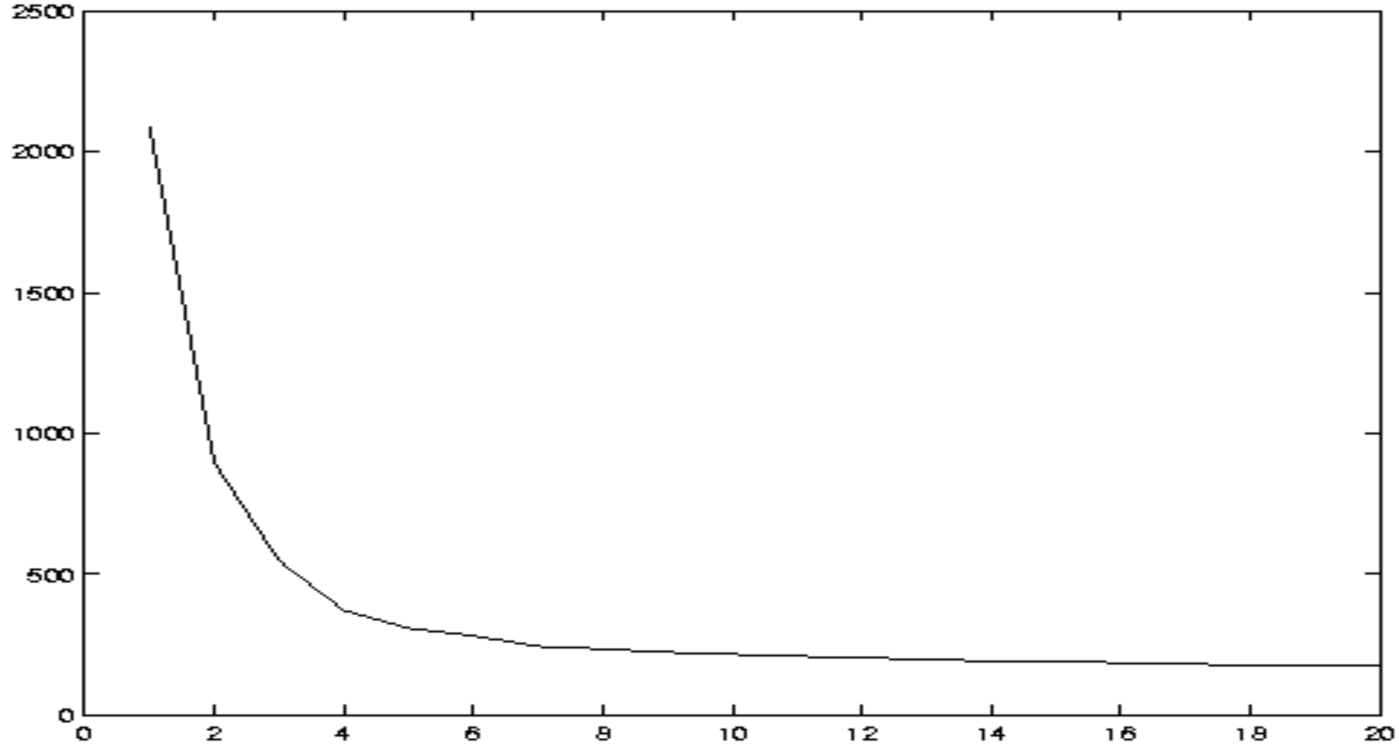
Example frequency distribution

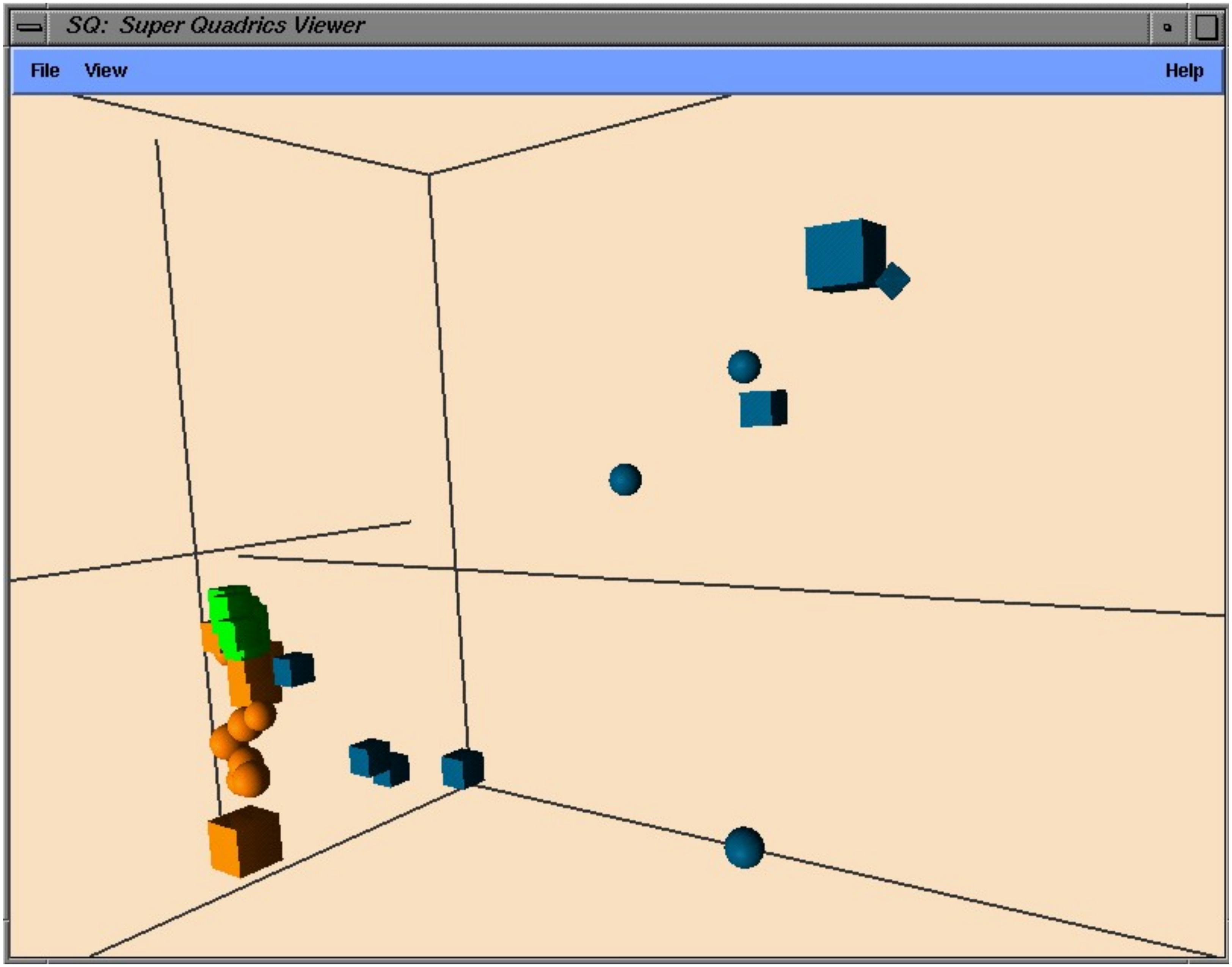


The Entries in Σ

- The singular values are the squares of the eigenvalues of the matrix AA^T
- A plot of the singular values is revealing
  - a steep left/downward slope indicates a homogeneous corpus
  - a "jagged" left side indicates a heterogeneous (multi-lingual?) corpus

Example plot of singular values




SQ: Super Quadrics Viewer
File
View
Help

Authorship as Text Classification

- TC relies on features, such as where and how often a term appears
- Probabilistic (e.g. Naïve Bayes) or Information Theoretic (e.g. Maximum Entropy) models are used
- Usually assumes a reliable training corpus