Forensic Linguistics

Jason Baldridge
UT Austin
Language and Computers

Includes materials from Roger Shuy http://audiovideo1.law.umt.edu:8080/Roger-Shuy-Forensic-Linguistics

Author comparison: match texts on left to those by same author on the right

His manner was not effusive. It seldom was; but he was glad, I think, to see me. With hardly a word spoken, but with a kindly eye, he waved me to an armchair, threw across his case of cigars, and indicated a spirit case and a gasogene in the corner. Then he stood before the fire and looked me over in his singular introspective fashion.

He was not an ill-disposed young man, unless to be rather cold hearted and rather selfish is to be ill-disposed: but he was, in general, well respected; for he conducted himself with propriety in the discharge of his ordinary duties. Had he married a more amiable woman, he might have been made still more respectable than he was.

There are many theories about what happened, but two general narratives seem to be gaining prominence, which we will call the greed narrative and the stupidity narrative. The two overlap, but they lead to different ways of thinking about where we go from here.

For all the preposterous hat and the vacuous face, there was something noble in the simple faith of our visitor which compelled our respect. She laid her little bundle of papers upon the table and went her way, with a promise to come again whenever she might be summoned.

He was invited to Kellynch Hall; he was talked of and expected all the rest of the year; but he never came. The following spring he was seen again in town, found equally agreeable, again encouraged, invited, and expected, and again he did not come; and the next tidings were that he was married.

Our moral and economic system is based on individual responsibility. It's based on the idea that people have to live with the consequences of their decisions. This makes them more careful deciders. This means that society tends toward justice — people get what they deserve as much as possible.

Author comparison: match texts on left to those by same author on the right

His manner was not effusive. It seldom was; but he was glad, I think, to see me. With hardly a word spoken, but with a kindly eye, he waved me to an armchair, threw across his case of cigars, and indicated a spirit case and a gasogene in the corner. Then he stood before the fire and looked me over in his singular introspective fashion.

He was not an ill-disposed young man, unless to be rather cold hearted and rather selfish is to be ill-disposed: but he was, in general, well respected; for he conducted himself with propriety in the discharge of his ordinary duties. Had he married a more amiable woman, he might have been made still more respectable than he was.

There are many theories about what happened, but two general narratives seem to be gaining prominence, which we will call the greed narrative and the stupidity narrative. The two overlap, but they lead to different ways of thinking about where we go from here.

Brooks, New York Times, Apr 2, 2009

For all the preposterous hat and the vacuous face, there was something noble in the simple faith of our visitor which compelled our respect. She laid her little bundle of papers upon the table and went her way, with a promise to come again whenever she might be summoned.

He was invited to Kellynch Hall; he was talked of and expected all the rest of the year; but he never came. The following spring he was seen again in town, found equally agreeable, again encouraged, invited, and expected, and again he did not come; and the next tidings were that he was married.

Our moral and economic system is based on individual responsibility. It's based on the idea that people have to live with the consequences of their decisions. This makes them more careful deciders. This means that society tends toward justice — people get what they deserve as much as possible.

Brooks, New York Times, Feb 19, 2009

Author comparison: match texts on left to those by same author on the right

His manner was not effusive. It seldom was; but he was glad, I think, to see me. With hardly a word spoken, but with a kindly eye, he waved me to an armchair, threw across his case of cigars, and indicated a spirit case and a gasogene in the corner. Then he stood before the fire and looked me over in his singular introspective fashion.

Doyle, *Sherlock Holmes, A Scandal in Bohemia*, 1891

He was not an ill-disposed young man, unless to be rather cold hearted and rather selfish is to be ill-disposed: but he was, in general, well respected; for he conducted himself with propriety in the discharge of his ordinary duties. Had he married a more amiable woman, he might have been made still more respectable than he was.

There are many theories about what happened, but two general narratives seem to be gaining prominence, which we will call the greed narrative and the stupidity narrative. The two overlap, but they lead to different ways of thinking about where we go from here.

Brooks, New York Times, Apr 2, 2009

For all the preposterous hat and the vacuous face, there was something noble in the simple faith of our visitor which compelled our respect. She laid her little bundle of papers upon the table and went her way, with a promise to come again whenever she might be summoned.

Doyle, *Sherlock Holmes, A Case of Identity*, 1891

He was invited to Kellynch Hall; he was talked of and expected all the rest of the year; but he never came. The following spring he was seen again in town, found equally agreeable, again encouraged, invited, and expected, and again he did not come; and the next tidings were that he was married.

Our moral and economic system is based on individual responsibility. It's based on the idea that people have to live with the consequences of their decisions. This makes them more careful deciders. This means that society tends toward justice — people get what they deserve as much as possible.

Brooks, New York Times, Feb 19, 2009

Author comparison: match texts on left to those by same author on the right

His manner was not effusive. It seldom was; but he was glad, I think, to see me. With hardly a word spoken, but with a kindly eye, he waved me to an armchair, threw across his case of cigars, and indicated a spirit case and a gasogene in the corner. Then he stood before the fire and looked me over in his singular introspective fashion.

Doyle, *Sherlock Holmes, A Scandal in Bohemia*, 1891

He was not an ill-disposed young man, unless to be rather cold hearted and rather selfish is to be ill-disposed: but he was, in general, well respected; for he conducted himself with propriety in the discharge of his ordinary duties. Had he married a more amiable woman, he might have been made still more respectable than he was.

Austen, *Sense and Sensibility*, 1811

There are many theories about what happened, but two general narratives seem to be gaining prominence, which we will call the greed narrative and the stupidity narrative. The two overlap, but they lead to different ways of thinking about where we go from here.

Brooks, New York Times, Apr 2, 2009

For all the preposterous hat and the vacuous face, there was something noble in the simple faith of our visitor which compelled our respect. She laid her little bundle of papers upon the table and went her way, with a promise to come again whenever she might be summoned.

Doyle, *Sherlock Holmes, A Case of Identity*, 1891

He was invited to Kellynch Hall; he was talked of and expected all the rest of the year; but he never came. The following spring he was seen again in town, found equally agreeable, again encouraged, invited, and expected, and again he did not come; and the next tidings were that he was married.

Austen, *Persuasion*, 1818

Our moral and economic system is based on individual responsibility. It's based on the idea that people have to live with the consequences of their decisions. This makes them more careful deciders. This means that society tends toward justice — people get what they deserve as much as possible.

Brooks, New York Times, Feb 19, 2009

Quantitative features that help discriminate the authors

- Grammatical person: 1st (we/us/our, I/me/my)
- Grammatical tense: present, past
- Word frequencies: frequent use of “he”
- Punctuation: use of colons and semi-colons
- Average word and sentence length
- Syntax: prepositional adverbial phrases (“With hardly...”, “For all the...”)
- These must be counted in all texts. The texts of unknown authorship should then have values most similar to those of the texts of one of the known authors.

Qualitative features that help discriminate the authors

- Content: discussing ideas or fellows getting married
- Use of flowery language.
- Use of same/similar terms: “marriage”, “idea”/“theories”
- Dated words/forms of expression: “gasogene”, “stood before”, “laid ... upon”
- **Note**: if these can be reliably identified and counted, they could be used as the basis for quantitative features.

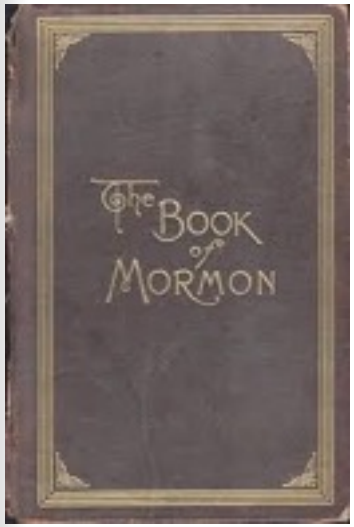
Forensic linguistics

- Forensic linguistics is a branch of applied linguistics that applies linguistic theory, research and principles to real life language in the legal context.
- Even more generally, it can be viewed as analyzing examples of language to discover properties that reveal more than just what is said.
  - authorship (same as other examples?, plagiarism)
  - psychological attributes of the author (deception, depression)
  - similarity to other examples (e.g., trademark disputes)

“Pre-forensic” forensic linguistics: authorship attribution

The Federalist Papers

Authors: Alexander Hamilton, James Madison, and John Jay
Period: 1787-1788



The Book of Mormon

Possible authors: John Smith, Solomon Spalding, Sidney Rigdon and others
Period: 1830

The Letters of St Paul

Authors: Paul the Apostle, Several unknown authors
Period: 0-200 A.D.

The writings of Thomas Wyatt

Possible authors: Thomas Wyatt, “Not” Thomas Wyatt
Period: early 1500s

Jan Svartvik, 1968

- First use of the term “forensic linguistics” in Svartvik’s analysis of the case of Timothy John Evans.
- in 1950, Evans was convicted and hanged for killing his wife and child
- in fact he had been duped by the killer, John Christie, into believing he was partly responsible for the deaths
- Svartvik analyzed four statements by Evans and showed that they had two very different **registers**, or varieties of language used for particular purposes or social settings. This indicated that **textual alternation** had occurred: the original statements by Evans had in fact been modified by the police.
- These variations called into question the authorship of the statements taken on two different occasions.

Malcom Coulthard, 1994

- Analyzed the statement of Derek Bentley, who was hanged in the 1950's for killing a police officer.
- Bentley was posthumously pardoned in 1998 based on multiple factors, including Coulthard's analysis of Bentley's confession.
- Coulthard argued, based on patterns of the language use, that the confession was largely edited by policemen.
  - Frequent use of "then", and in particular, use of "I then" rather than "Then I", which was inconsistent with Bentley's use of language in court testimony
- Other cases have brought up this sort of fabrication of confessions by police, e.g. of Patrick Molloy in the Bridgewater four case from 1978, appealed in 1997.

Molloy interview and statement (from Coulthard)

Extract from Disputed Interview with Molloy

P. How long were you in there Pat?

(17) **I had been drinking and cannot remember the exact time that I was there, but whilst I was upstairs I heard someone downstairs say 'be careful someone is coming'**.

P. Did you hide?

(18) Yes **I hid for a while** and then **I heard** the **bang** I have told you about.

P. Carry on Pat?

(19) I ran out.

P. What were the others doing?

(20) **The three of them were still in the room**.

P. What were they doing?

(21) **They all looked shocked and were shouting at each other**.

P. Who said what?

(22) **I heard Jimmy say 'it went off by accident'**.

P. Pat, I know this is upsetting but you appreciate that we must get to the bottom of this. Did you see the boy's body?

(23) Yes sir, he was **on the settee**.

P. Did you see any injury to him?

(24) Yes sir, **he had been shot in the head**.

P. What happened then?

(25) **I was appalled and felt sick**

Molloy interview and statement (from Coulthard)

Extract from Disputed Interview with Molloy

P. How long were you in there Pat?

(17) **I had been drinking and cannot remember the exact time that I was there, but whilst I was upstairs I heard someone downstairs say 'be careful someone is coming'**.

P. Did you hide?

(18) Yes **I hid for a while** and then **I heard** the **bang** I have told you about.

P. Carry on Pat?

(19) I ran out.

P. What were the others doing?

(20) **The three of them were still in the room**.

P. What were they doing?

(21) **They all looked shocked and were shouting at each other**.

P. Who said what?

(22) **I heard Jimmy say 'it went off by accident'**.

P. Pat, I know this is upsetting but you appreciate that we must get to the bottom of this. Did you see the boy's body?

(23) Yes sir, he was **on the settee**.

P. Did you see any injury to him?

(24) Yes sir, **he had been shot in the head**.

P. What happened then?

(25) **I was appalled and felt sick**

Extract from Molloy's statement

(14) Jimmy broke in through a window and loosed us in.

(15) They went downstairs and I went upstairs by myself.

(16) I searched the bedrooms I remember taking the drawers from some furniture and after searching them I stacked them one on top of the other.

(17) I had been drinking and cannot remember the exact time I was there but whilst I was upstairs I heard someone downstairs say be careful someone is coming.

(18) I hid for a while and after a while I heard a bang come from downstairs.

(19) I knew that it was a gun being fired.

(20) I went downstairs and the three of them were still in the room.

(21) They all looked shocked and were shouting at each other.

(22) I heard Jimmy say, "It went off by accident".

(23) I looked and on the settee I saw the body of the boy.

(24) He had been shot in the head.

(25) I was appalled and felt sick.

Anomalous aspects of the interview

- Forward-linking questions, not typically of usual interviews:
  - P: Who **said** what?
  - M: I heard Jimmy **say** “it went off by accident”.
  - Should have been “Who was **shouting**...”
- Grammatical misfits:
  - P: Did you see **the body of the boy**?
  - M: Yes sir, **he** was on the settee.
  - Should have been “**it**”.
- Process misfit
  - P: What **happened** then?
  - M: I **was appalled** and **felt sick**.
  - Describes two states rather than an action or event (e.g., I **vomited**)

The Unabomber

- The Unabomber carried out a mail bombing campaign between 1978 to 1995, killing 3 people and injuring 23.
- In 1995, he demanded publication of a 35,000 word manifesto in either the New York Times or the Washington Post (with a threat to kill more people if not complied with)
- FBI linguistic analysis: the Unabomber was young (under 25), a laborer, and possibly from the West Coast
- Linguist Roger Shuy's analysis was more accurate: about 50 years old, well-educated (possibly a doctorate, but not in the social sciences or humanities), from the Chicago area but had spent time in California.
- However, positive identification of the Unabomber's actual identity came from his brother, based on linguistic expression in the manifesto that recognized.

The Unabomber

- The Unabomber is Theodore Kaczynski: 53 years old (at the time), born in Chicago, PhD in mathematics from University of Michigan, and assistant professor at UC-Berkeley (for two years).
- His brother recognized Kaczynski's writing style and beliefs in the manifesto and notified the FBI.
  - Used the expression "cool-headed logicians"
  - Also the FBI's James Fitzgerald noted the unusual "you can't eat your cake and have it too"
  - See Ben Zimmer's Language Log post for more discussion:
    - http://itre.cis.upenn.edu/~myl/languagelog/archives/002762.html
- The evidence was sufficient for a judge to grant a warrant for Kaczynski's arrest.

Other roles for forensic linguistics

- Copyright infringement: McSleep (Quality Inn vs McDonald's, http://itre.cis.upenn.edu/~myl/languagelog/archives/003670.html)
- Intelligibility of company/government policy documents: Illinois Dept of Public Aid sent a letter with technical/ bureaucratic language informing recipients of lowering of benefits (Judith Levi, 1994)
- Analyzing intent: did a suspect actually agree to an illegal act, bribe, etc? Delorean trial (http://itre.cis.upenn.edu/~myl/languagelog/archives/003303.html)
- Transcription accuracy: figuring out what most likely was said in a noisy taped conversation. (Roger Shuy, http://audiovideo1.law.umt.edu:8080/Roger-Shuy-Forensic-Linguistics)

Government transcription	Defense transcription
I would take a bribe, wouldn't you.	I wouldn't take a bribe, would you.

Other roles for forensic linguistics

- Copyright infringement: McSleep (Quality Inn vs McDonald's, http://itre.cis.upenn.edu/~myl/languagelog/archives/003670.html)
- Intelligibility of company/government policy documents: Illinois Dept of Public Aid sent a letter with technical/ bureaucratic language informing recipients of lowering of benefits (Judith Levi, 1994)
- Analyzing intent: did a suspect actually agree to an illegal act, bribe, etc? Delorean trial (http://itre.cis.upenn.edu/~myl/languagelog/archives/003303.html)
- Transcription accuracy: figuring out what most likely was said in a noisy taped conversation. (Roger Shuy, http://audiovideo1.law.umt.edu:8080/Roger-Shuy-Forensic-Linguistics)

Government transcription	Defense transcription
I would take a bribe, wouldn't you.	I wouldn't take a bribe, would you.

Syllables ⟶

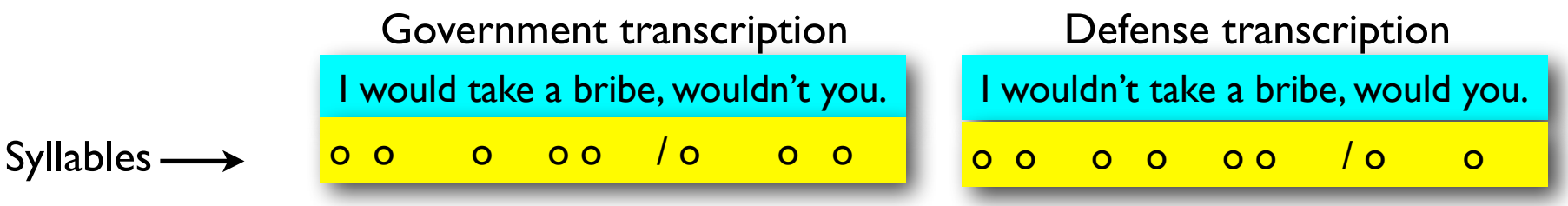
Other roles for forensic linguistics

- Copyright infringement: McSleep (Quality Inn vs McDonald's, http://itre.cis.upenn.edu/~myl/languagelog/archives/003670.html)
- Intelligibility of company/government policy documents: Illinois Dept of Public Aid sent a letter with technical/ bureaucratic language informing recipients of lowering of benefits (Judith Levi, 1994)
- Analyzing intent: did a suspect actually agree to an illegal act, bribe, etc? Delorean trial (http://itre.cis.upenn.edu/~myl/languagelog/archives/003303.html)
- Transcription accuracy: figuring out what most likely was said in a noisy taped conversation. (Roger Shuy, http://audiovideo1.law.umt.edu:8080/Roger-Shuy-Forensic-Linguistics)

	Government transcription	Defense transcription
	I would take a bribe, wouldn't you.	I wouldn't take a bribe, would you.
Syllables →	o o o o o / o o o	

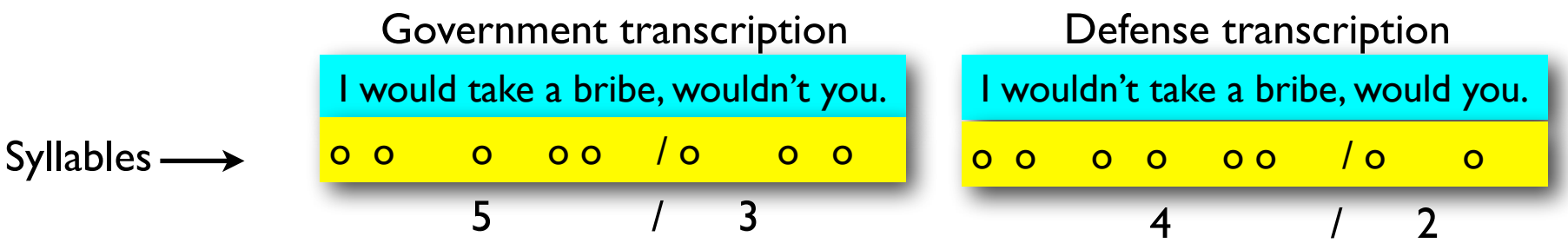
Other roles for forensic linguistics

- Copyright infringement: McSleep (Quality Inn vs McDonald's, http://itre.cis.upenn.edu/~myl/languagelog/archives/003670.html)
- Intelligibility of company/government policy documents: Illinois Dept of Public Aid sent a letter with technical/ bureaucratic language informing recipients of lowering of benefits (Judith Levi, 1994)
- Analyzing intent: did a suspect actually agree to an illegal act, bribe, etc? Delorean trial (http://itre.cis.upenn.edu/~myl/languagelog/archives/003303.html)
- Transcription accuracy: figuring out what most likely was said in a noisy taped conversation. (Roger Shuy, http://audiovideo1.law.umt.edu:8080/Roger-Shuy-Forensic-Linguistics)



Other roles for forensic linguistics

- Copyright infringement: McSleep (Quality Inn vs McDonald's, http://itre.cis.upenn.edu/~myl/languagelog/archives/003670.html)
- Intelligibility of company/government policy documents: Illinois Dept of Public Aid sent a letter with technical/ bureaucratic language informing recipients of lowering of benefits (Judith Levi, 1994)
- Analyzing intent: did a suspect actually agree to an illegal act, bribe, etc? Delorean trial (http://itre.cis.upenn.edu/~myl/languagelog/archives/003303.html)
- Transcription accuracy: figuring out what most likely was said in a noisy taped conversation. (Roger Shuy, http://audiovideo1.law.umt.edu:8080/Roger-Shuy-Forensic-Linguistics)



Daubert standards

- Forensic linguistics applies linguistic techniques to provide evidence in legal cases.
- The question of scientific validity of analysis techniques arises, and is governed by the *Daubert* standards.
  - **Knowledge and stature**: is the expert an acceptable person to apply the methodology?
  - **Testing**: is the methodology empirically testable? Can it be falsified or refuted?
  - **Peer review**: has the methodology passed the peer review process?
  - **Scientific method**: the methodology must have a known error rate
  - **Straightforwardness**: the methodology should be explainable in a clear manner that can be understood by judge and jury

(From Olsson, 2008, *Forensic Linguistics*)

Linguistic fingerprinting and identification

- There has been much interest in finding linguistic "fingerprints", but there are problems with the concept:
  - **language acquisition**: language is learned and continually changing
  - **linguistic homogeneity**: education, mass media
  - **register**: the same person speaks differently in different contexts, with different people
- No accepted definition of general linguistic fingerprint has so far been proposed, nor are we likely to see one for these reasons.
- Nonetheless, people do exhibit regularities in their speech and writing that could distinguish them from others.
  - This allows us to compare a limited set of authors/speakers in certain restricted conditions, just as we did with the first page of these slides.

Identifying style

- Every speaker uses language differently, leading to a unique style.
- Style is both:
  - a collection of markers which can be observed and measured
  - a set of unconscious habits which can be observed and measured
- Quantifying style:
  - word usage: presence/absence of words, relative word frequencies
  - type/token ratios
  - average word and sentence length
  - the number of unique words (hapax legomena)
- Clearly, computational linguistics can help out here!

Computers and forensic linguistics

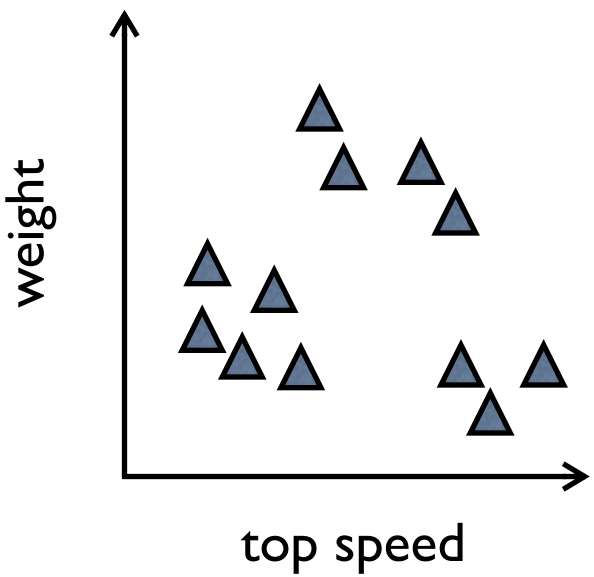
- Roles for computers:
  - efficiently count style markers in a corpus
  - web search to verify what are common versus less common forms of linguistic expression
  - machine learning algorithms that cluster and/or classify sets of authors based on a collection of texts and style markers that can be extracted from them
- Machine learning approaches have special appeal, in particular because they easily conform to the Daubert standards (if applied appropriately).
- However, they are best used as part of making an opinion informed by machine output and other information.
- Also, machine learning typically requires a substantial amount of text, and thus cannot be used for authorship determination of very short texts, like a single text message.

Authorship attribution with machine learning

- Machine learning provides a class of algorithms that perform **unsupervised clustering**.
  - They don't have labels for any of the data points (e.g., documents).
  - Based on properties measured from the data points, coherent clusters of documents with similar properties can be identified.
  - A cluster can correspond to many different things, including collections of documents by the same author.
- **Mixture models:** a popular class of probabilistic algorithms for clustering.
  - collections of probability distributions over the data
  - "soft" cluster membership: points are proportionally part of multiple clusters
  - a mixture of Gaussian distributions (a.k.a. the normal distribution) are one of the most commonly used type of mixture model
- The **K-means** algorithm is a related "hard" clustering algorithm that is simple and easy to understand.

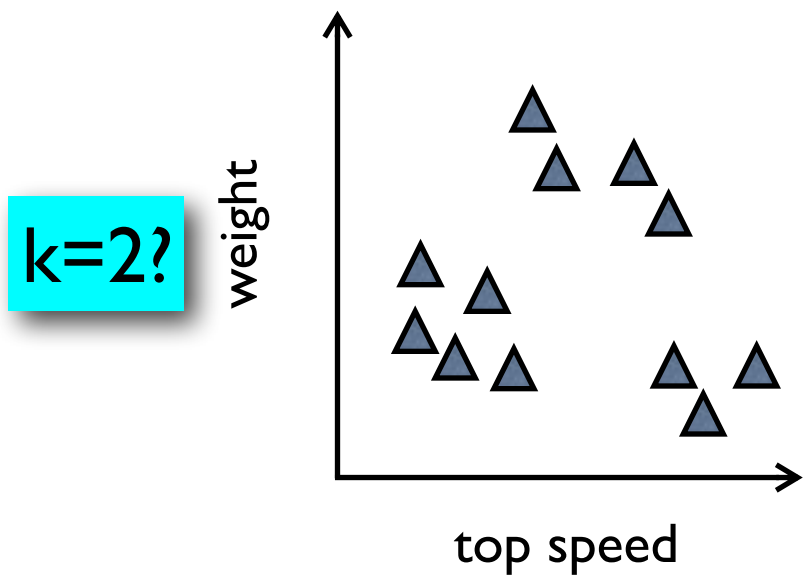
Hard clustering into *k* groups

- Assume you can measure various attributes for each data point, e.g.:
  - the weight and top speed of various vehicles
  - the average sentence length and average word length of various authors.
- Next, you want to identify *k* groups of similar items based on these attributes.
- How many groups? How to find them using an algorithm?



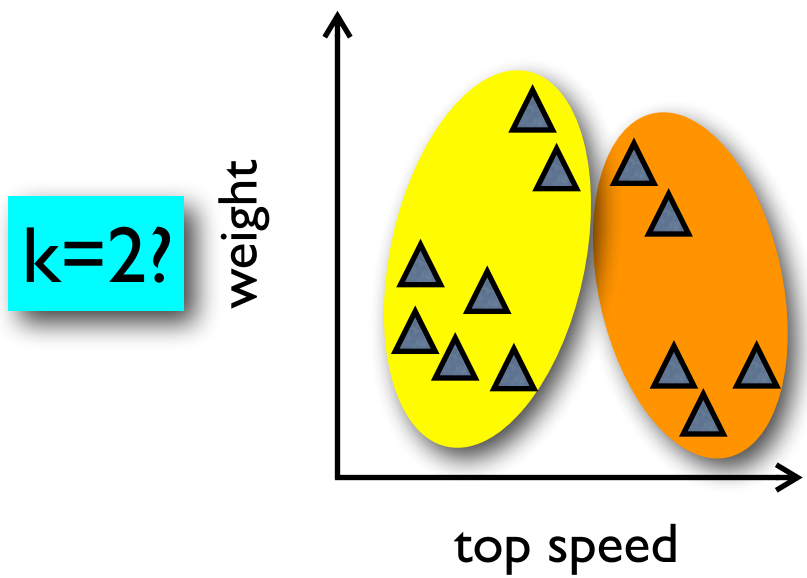
Hard clustering into *k* groups

- Assume you can measure various attributes for each data point, e.g.:
  - the weight and top speed of various vehicles
  - the average sentence length and average word length of various authors.
- Next, you want to identify *k* groups of similar items based on these attributes.
- How many groups? How to find them using an algorithm?



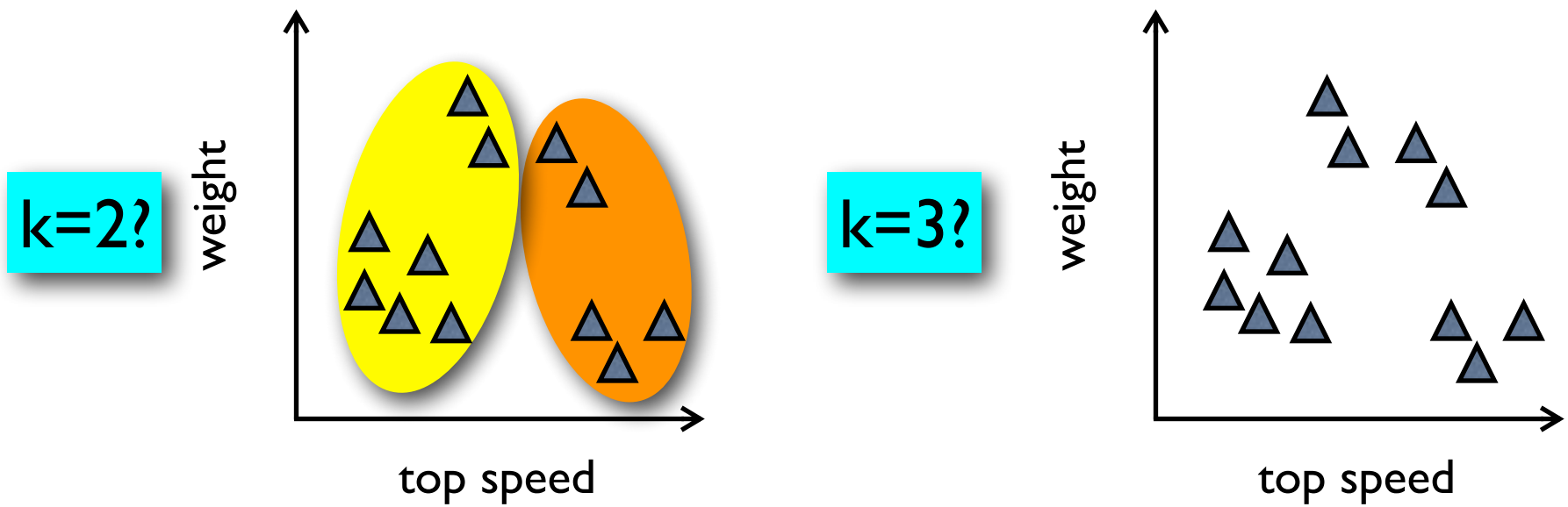
Hard clustering into *k* groups

- Assume you can measure various attributes for each data point, e.g.:
  - the weight and top speed of various vehicles
  - the average sentence length and average word length of various authors.
- Next, you want to identify *k* groups of similar items based on these attributes.
- How many groups? How to find them using an algorithm?



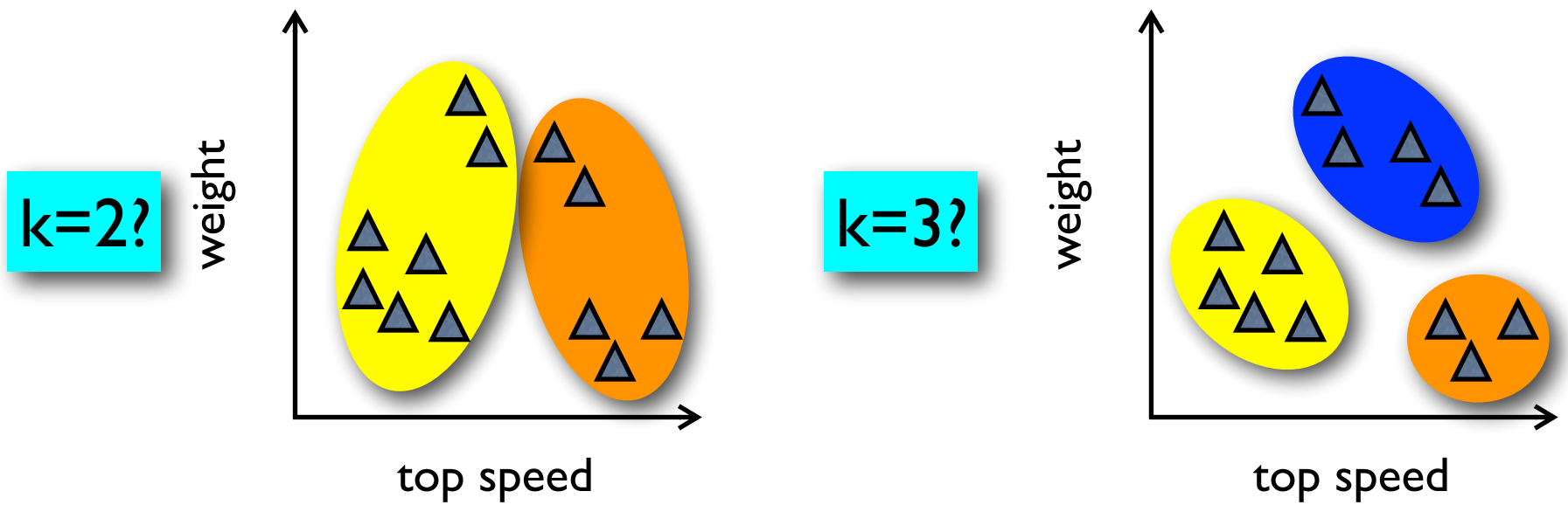
Hard clustering into *k* groups

- Assume you can measure various attributes for each data point, e.g.:
  - the weight and top speed of various vehicles
  - the average sentence length and average word length of various authors.
- Next, you want to identify *k* groups of similar items based on these attributes.
- How many groups? How to find them using an algorithm?



Hard clustering into *k* groups

- Assume you can measure various attributes for each data point, e.g.:
  - the weight and top speed of various vehicles
  - the average sentence length and average word length of various authors.
- Next, you want to identify *k* groups of similar items based on these attributes.
- How many groups? How to find them using an algorithm?



An authorship clustering problem

- Texts from three authors (five documents each)
  - Arthur Conan Doyle (obtained from Project Gutenberg)
  - Jane Austen (obtained from Project Gutenberg)
  - Paul Krugman (obtained from New York Times website)
- Measure the relative frequency of the words "I" and "the" in each document.

Austen

Document	I	the
Emma	1.8	3.2
Mansfield	1.5	3.9
Persuasion	1.3	4
Pride	1.7	3.6
Sense	1.6	3.4

Doyle

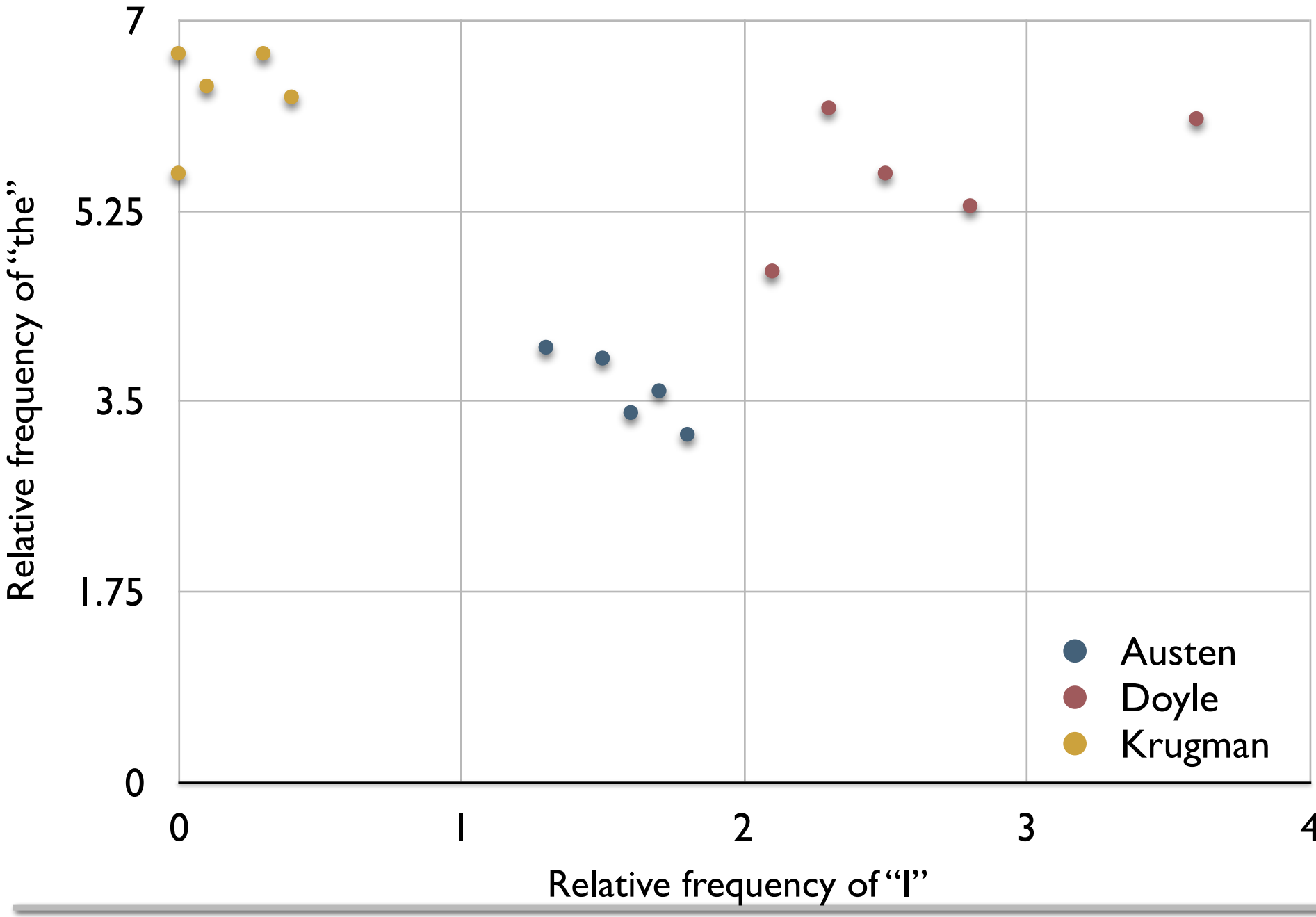
Document	I	the
City	2.1	4.7
Gerard	3.6	6.1
Holmes	2.8	5.3
Hound	2.5	5.6
Polestar	2.3	6.2

Krugman

Document	I	the
12-01-2008	0	6.7
12-07-2008	0	5.6
12-15-2008	0.3	6.7
12-19-2008	0.1	6.4
12-22-2008	0.4	6.3

- Now, "forgetting" who the authors are, let's see if they fall into distinct clusters based on these attributes.

Plot the attributes against each other



K-means: intuition

- We see the clusters quite clearly, but a computer doesn't and we need to specify an algorithm that allows it to identify them.
- The K-means algorithm is a simple algorithm for such tasks.
- The basic idea:
  - the values for the attributes in each dimension will be similar for each document of the same author
  - each author is represented as the averages for the attributes of all the documents he or she wrote
  - but: we don't know those averages since we "forgot" the authors!
  - so, we take a guess at the average for each author, and then see which documents each of our hypothesized authors were likely to have produced
  - these guesses will probably be wrong, but we can fix that by iteratively re-estimating them

K-means: algorithm

- We are given N documents: $D = d_1, d_2, ..., d_N$
- We need to output K centroids: $C = c_1, c_2, ..., c_K$
- These centroids partition the documents into clusters based on which centroid is closest to each document.

K-means(D,K)
C ← *SelectRandomCenters*(D,K)
while C does change
  for k ← 1 to K
    $g_k \leftarrow \{\}$
  for n ← 1 to N
    $j \leftarrow \text{argmin}_i$ *distance*(c_i, d_n)
    $g_j \leftarrow g_j \cup \{d_n\}$
  for k ← 1 to K

$$c_k \leftarrow \frac{1}{|g_k|} \sum_{d \text{ in } g_k} d$$

return C

[Based on Manning, Raghavan, and Schutze 2008]

K-means: algorithm

- We are given N documents: $D = d_1, d_2, ..., d_N$
- We need to output K centroids: $C = c_1, c_2, ..., c_K$
- These centroids partition the documents into clusters based on which centroid is closest to each document.

K-means(D,K)
C ← *SelectRandomCenters*(D,K)
while C does change
  for k ← 1 to K
    $g_k \leftarrow \{\}$
  for n ← 1 to N
    j ← argmin_i *distance*(c_i,d_n)
    $g_j \leftarrow g_j \cup \{d_n\}$
  for k ← 1 to K

$$c_k \leftarrow \frac{1}{|g_k|} \sum_{d \text{ in } g_k} d$$

return C

Pick K random points (could be some of the data points in D)

[Based on Manning, Raghavan, and Schutze 2008]

K-means: algorithm

- We are given N documents: $D = d_1, d_2, ..., d_N$
- We need to output K centroids: $C = c_1, c_2, ..., c_K$
- These centroids partition the documents into clusters based on which centroid is closest to each document.

K-means(D,K)
C ← *SelectRandomCenters*(D,K)
while C does change
  for k ← 1 to K
    $g_k \leftarrow \{\}$
  for n ← 1 to N
    $j \leftarrow \text{argmin}_i$ *distance*(c_i, d_n)
    $g_j \leftarrow g_j \cup \{d_n\}$
  for k ← 1 to K

$$c_k \leftarrow \frac{1}{|g_k|} \sum_{d \text{ in } g_k} d$$

return C

Pick K random points (could be some of the data points in D)

Stopping criteria (when does the algorithm stop?)

[Based on Manning, Raghavan, and Schutze 2008]

K-means: algorithm

- We are given N documents: $D = d_1, d_2, ..., d_N$
- We need to output K centroids: $C = c_1, c_2, ..., c_K$
- These centroids partition the documents into clusters based on which centroid is closest to each document.

K-means(D,K)
C ← *SelectRandomCenters*(D,K)
while C does change
  for k ← 1 to K
    $g_k \leftarrow \{\}$
  for n ← 1 to N
    $j \leftarrow \text{argmin}_i$ *distance*(c_i, d_n)
    $g_j \leftarrow g_j \cup \{d_n\}$
  for k ← 1 to K

$$c_k \leftarrow \frac{1}{|g_k|} \sum_{d \text{ in } g_k} d$$

return C

Pick K random points (could be some of the data points in D)

Stopping criteria (when does the algorithm stop?)

(Re)initialize the document clusters.

[Based on Manning, Raghavan, and Schutze 2008]

K-means: algorithm

- We are given N documents: $D = d_1, d_2, ..., d_N$
- We need to output K centroids: $C = c_1, c_2, ..., c_K$
- These centroids partition the documents into clusters based on which centroid is closest to each document.

K-means(D,K)
C ← *SelectRandomCenters*(D,K)
while C does change
  for k ← 1 to K
    $g_k \leftarrow \{\}$
  for n ← 1 to N
    $j \leftarrow \text{argmin}_i$ *distance*(c_i, d_n)
    $g_j \leftarrow g_j \cup \{d_n\}$
  for k ← 1 to K

$$c_k \leftarrow \frac{1}{|g_k|} \sum_{d \text{ in } g_k} d$$

return C

Pick K random points (could be some of the data points in D)

Stopping criteria (when does the algorithm stop?)

(Re)initialize the document clusters.

Find the closest centroid for each document; put the document in that group.

[Based on Manning, Raghavan, and Schutze 2008]

K-means: algorithm

- We are given N documents: $D = d_1, d_2, ..., d_N$
- We need to output K centroids: $C = c_1, c_2, ..., c_K$
- These centroids partition the documents into clusters based on which centroid is closest to each document.

K-means(D,K)
C ← *SelectRandomCenters*(D,K)
while C does change
  for k ← 1 to K
    $g_k \leftarrow \{\}$
  for n ← 1 to N
    j ← argmin_i *distance*(c_i,d_n)
    $g_j \leftarrow g_j \cup \{d_n\}$
  for k ← 1 to K

$$c_k \leftarrow \frac{1}{|g_k|} \sum_{d \text{ in } g_k} d$$

return C

[Based on Manning, Raghavan, and Schutze 2008]

Pick K random points (could be some of the data points in D)

Stopping criteria (when does the algorithm stop?)

(Re)initialize the document clusters.

Find the closest centroid for each document; put the document in that group.

Recompute centroids based on the new document clusters (the g_k's).

Calculating distance

- The documents are data points in some (possibly high-dimensional) space. We'll work with 2D here.
- Recall the Pythagorean theorem: $c^2 = a^2 + b^2$
- Here, the "a" is the distance on the x-axis and the "b" is the distance on the y-axis between points d_i and d_j.
  - $\text{distance}(d_i,d_j) = (x_i - x_j)^2 + (y_i - y_j)^2$
- Consider two data points $d_1 = (5,4)$ and $d_2 = (1,2)$.
  - $\text{distance}(d_1, d_2) = (x_1 - x_2)^2 + (y_1 - y_2)^2 = (5-1)^2 + (4-2)^2 = 4^2 + 2^2 = 20$
- Note: we could take the square root, but it doesn't matter since we are just comparing a bunch of squared distances.

Simplified problem: just two authors and four documents

- Let's apply the K-means algorithm to four documents
- Keep in mind that we are acting like we don't know who is the author of each document.

Austen

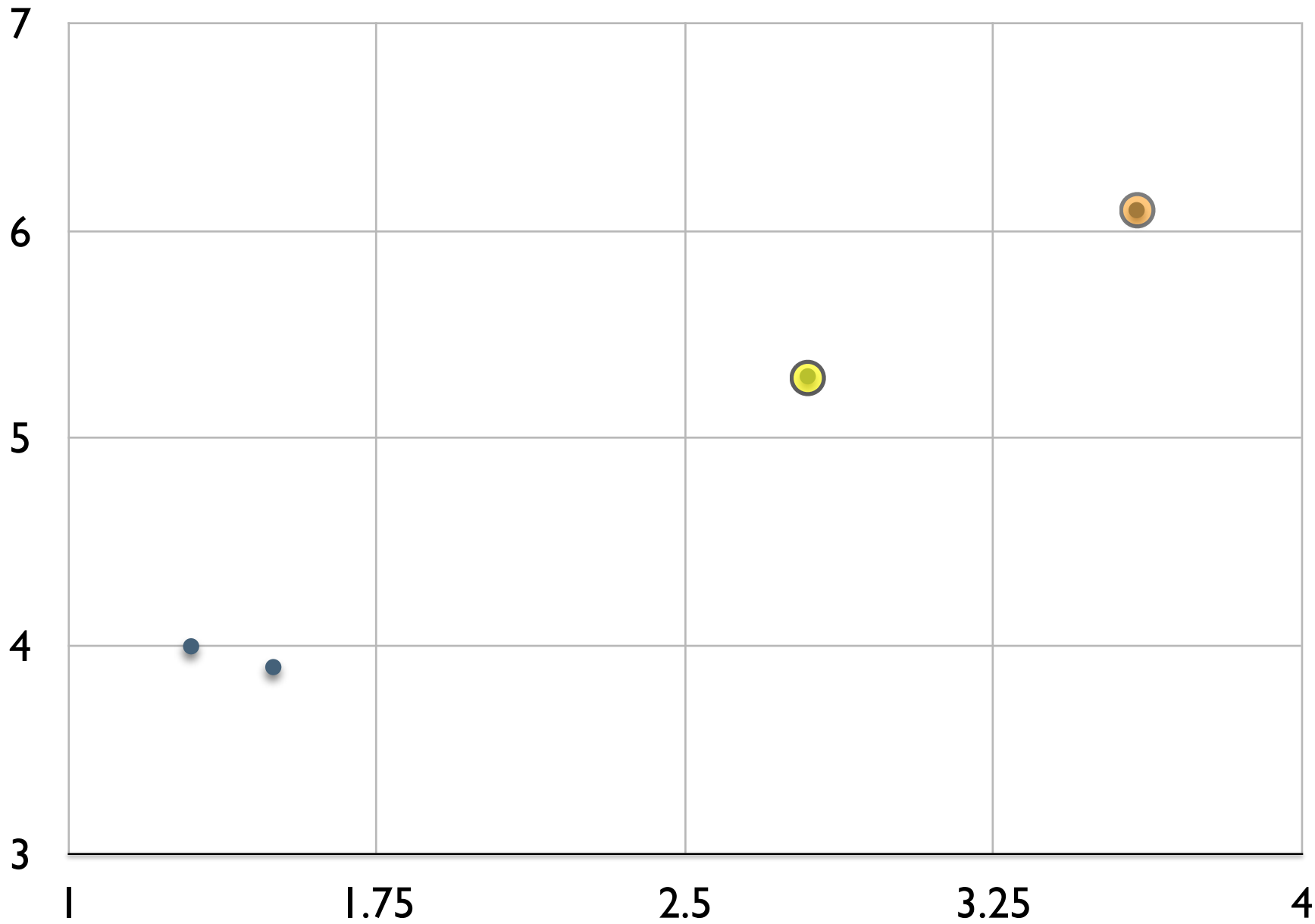
Document	I	the
Mansfield	1.5	3.9
Persuasion	1.3	4

Doyle

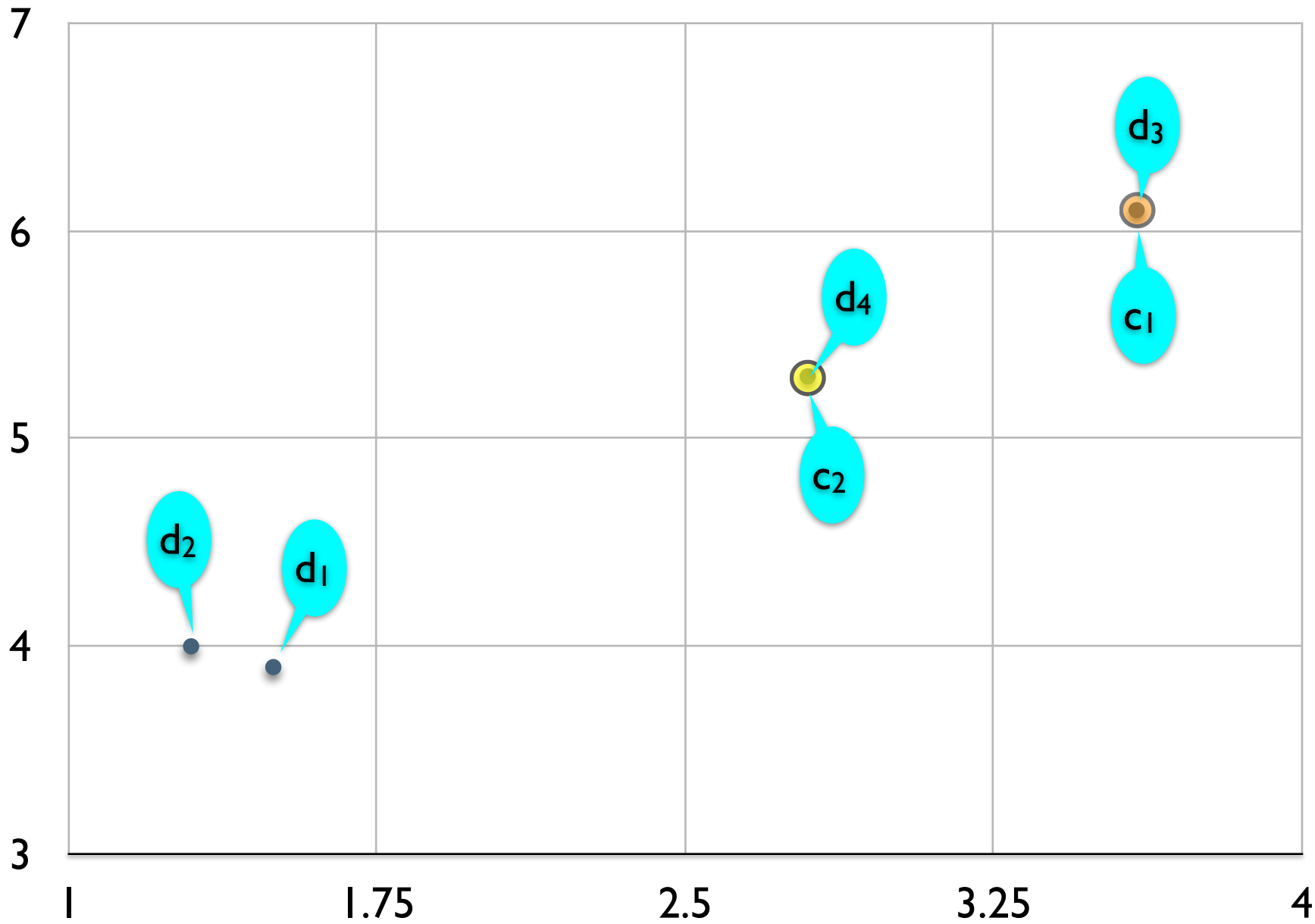
Document	I	the
Gerard	3.6	6.1
Holmes	2.8	5.3

- $D = \{d_1, d_2, d_3, d_4\} = \{ (1.5,3.9), (1.3,4.0), (3.6,6.1), (2.8,5.3) \}$
- Choose K = 2 (i.e., 2 authors)
- Choose $C = \{c_1, c_2\} = \{ (3.6,6.1), (2.8,5.3) \}$ as initial seed centroids.

Here's what it looks like



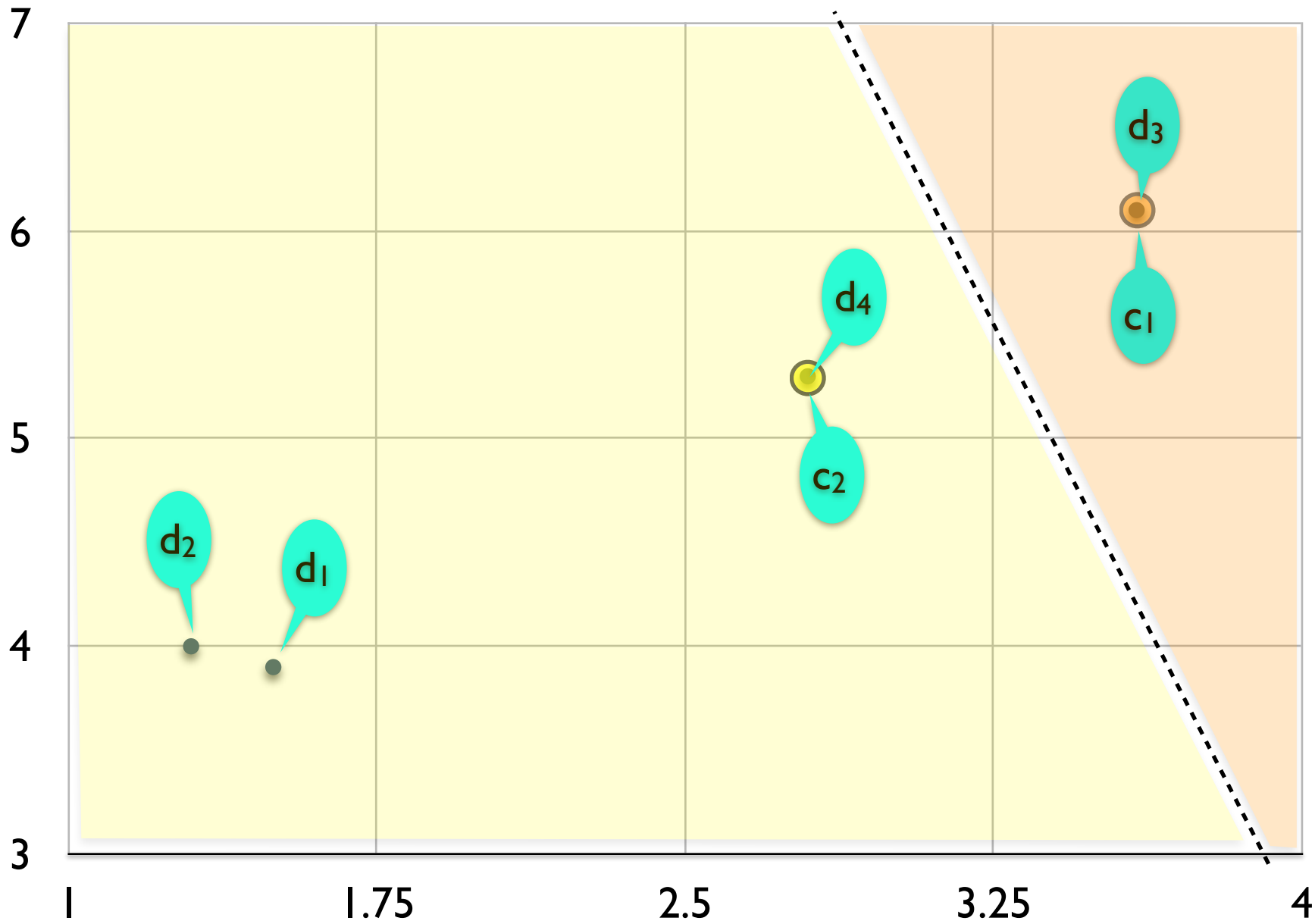
Here's what it looks like



Computing the groups

- Calculate the nearest centroid for each document and put it in the group for that centroid.
  - d1:
    - distance(c1,d1) = $(3.6-1.5)^2 + (6.1-3.9)^2 = 9.25$
    - distance(c2,d1) = $(2.8-1.5)^2 + (5.3-3.9)^2 = 3.65$
  - c2 is closer, so d1 is in g2
- Doing this for d2, d3, and d4, we find that:
  - g1 = {d3} and g2 = {d1,d2,d4}

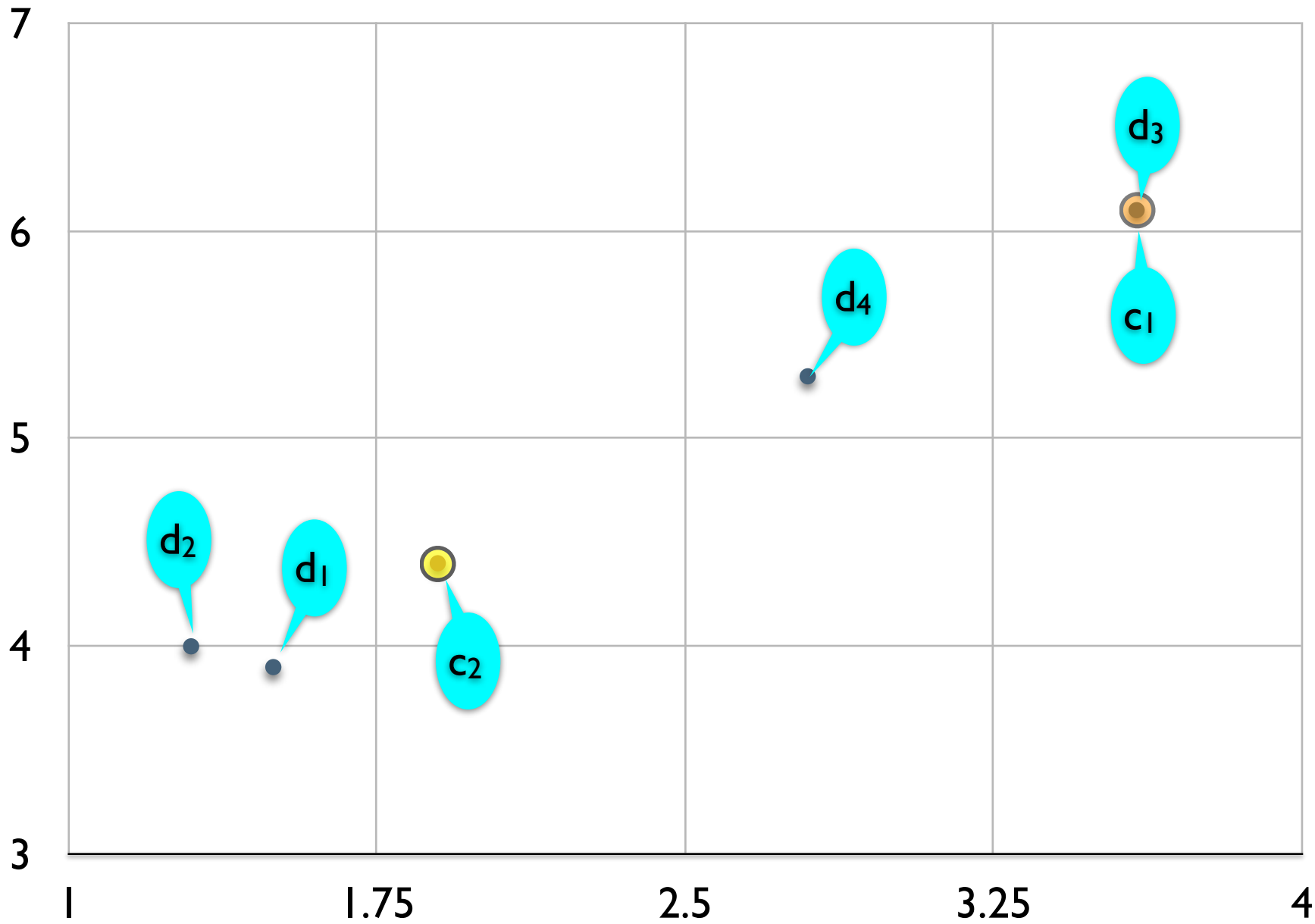
Here's what it looks like



New centroids

- Next, we need to compute the new centroids based on these groups.
- g2 has multiple elements:
  - sum of the x-values: 1.5+1.3+2.8 = 5.6
  - sum of the y-values: 3.9+4.0+5.3 = 13.2
  - size of g2 is 3, so c2 = (5.6/3, 13.2/3) = (1.9, 4.4)
- g1 stays the same:
  - size of g1 is 1, so we have c1 = (3.6/1, 6.1/1) = (3.6,6.1)

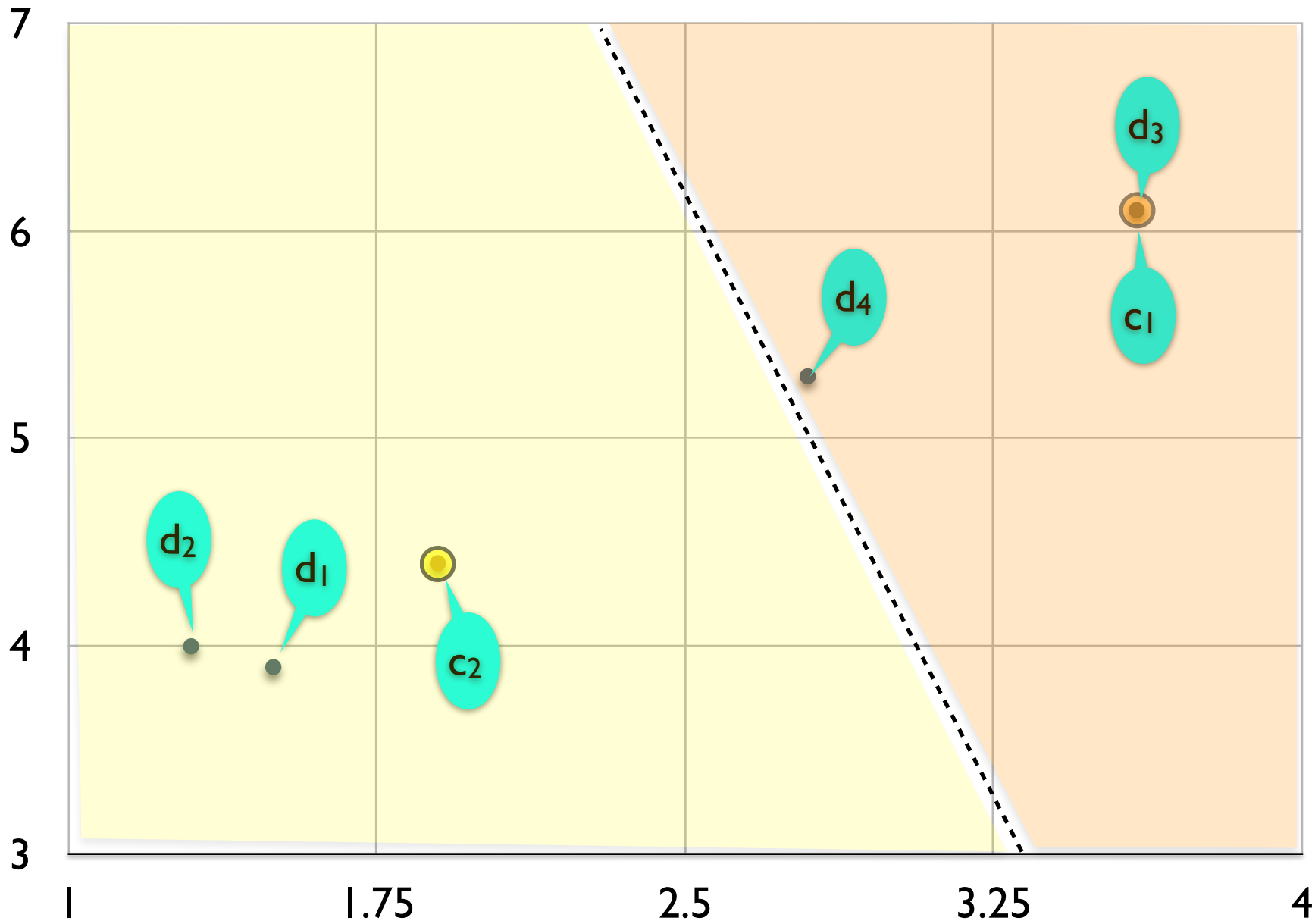
Here's what it looks like



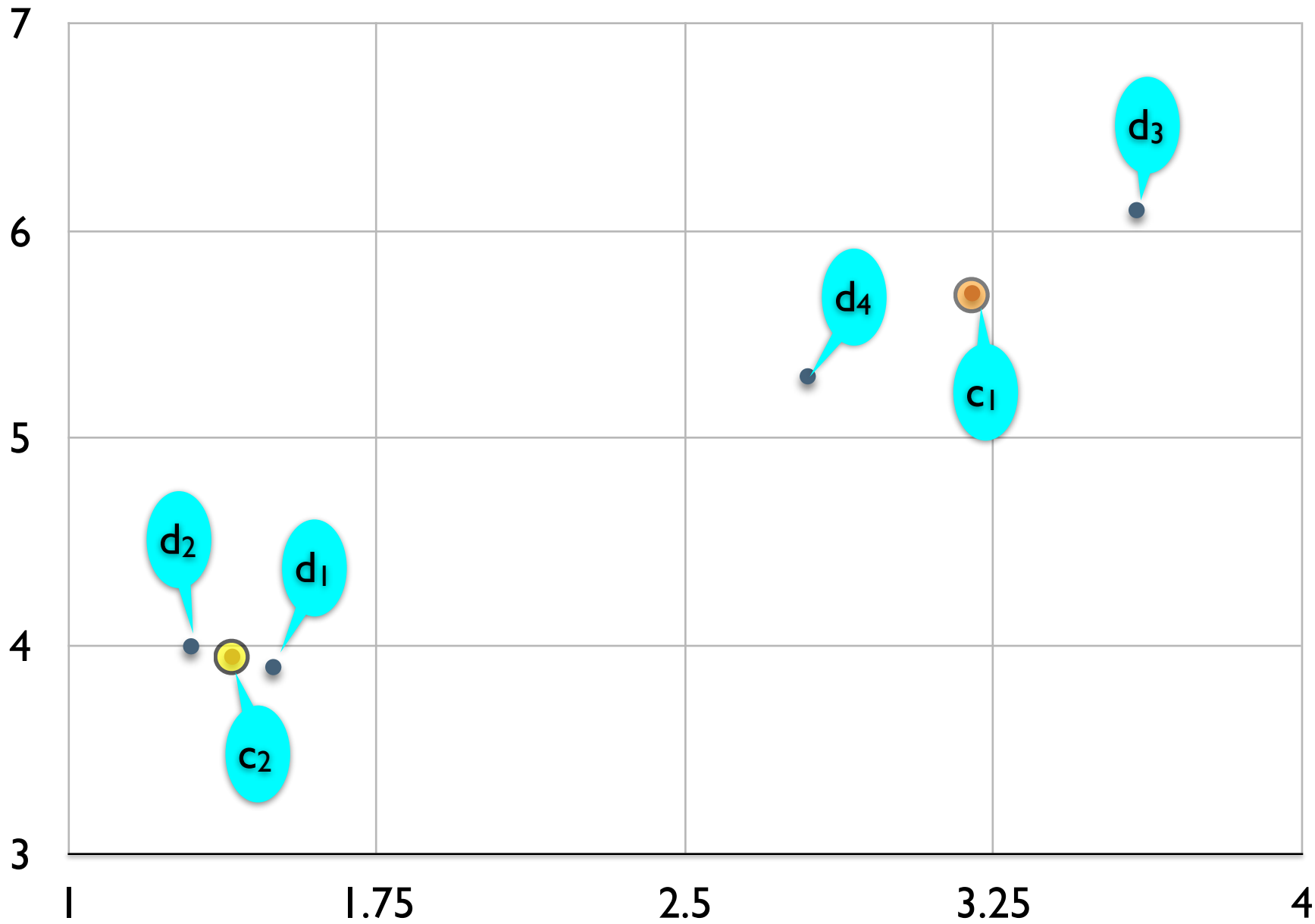
Re-assign groups based on new centroids

- We then keep iterating until the centroids stay the same
  - calculate nearest centroid for each document, put in in the group
  - recalculate centroids for new groups
- Notice that d4 is now closer to c1
  - distance(c1,d4) = $(3.6-2.8)^2 + (6.1-5.3)^2 = 1.28$
  - distance(c2,d4) = $(1.9-2.8)^2 + (4.4-5.3)^2 = 1.62$

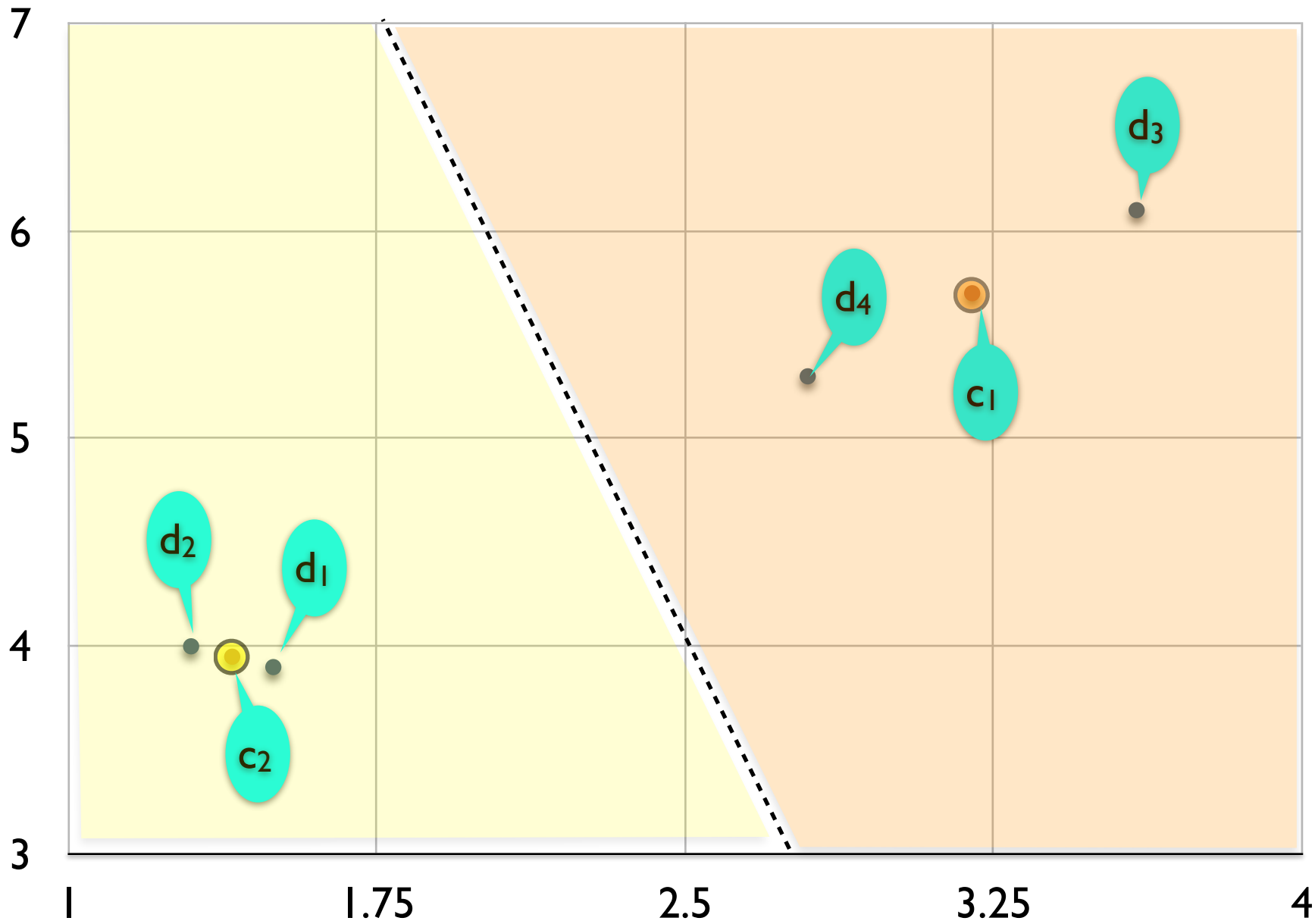
Here's what it looks like



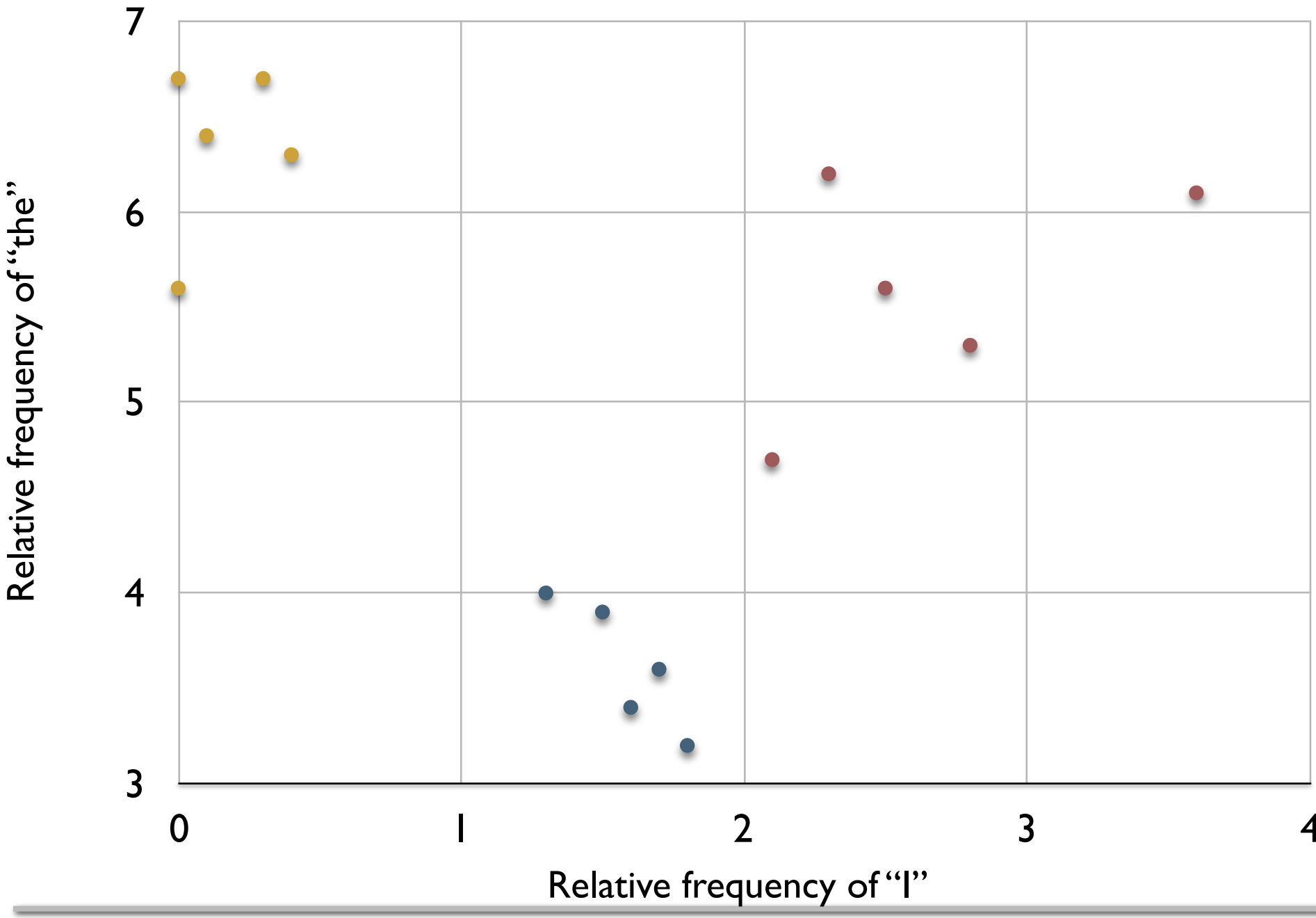
The next round would be...



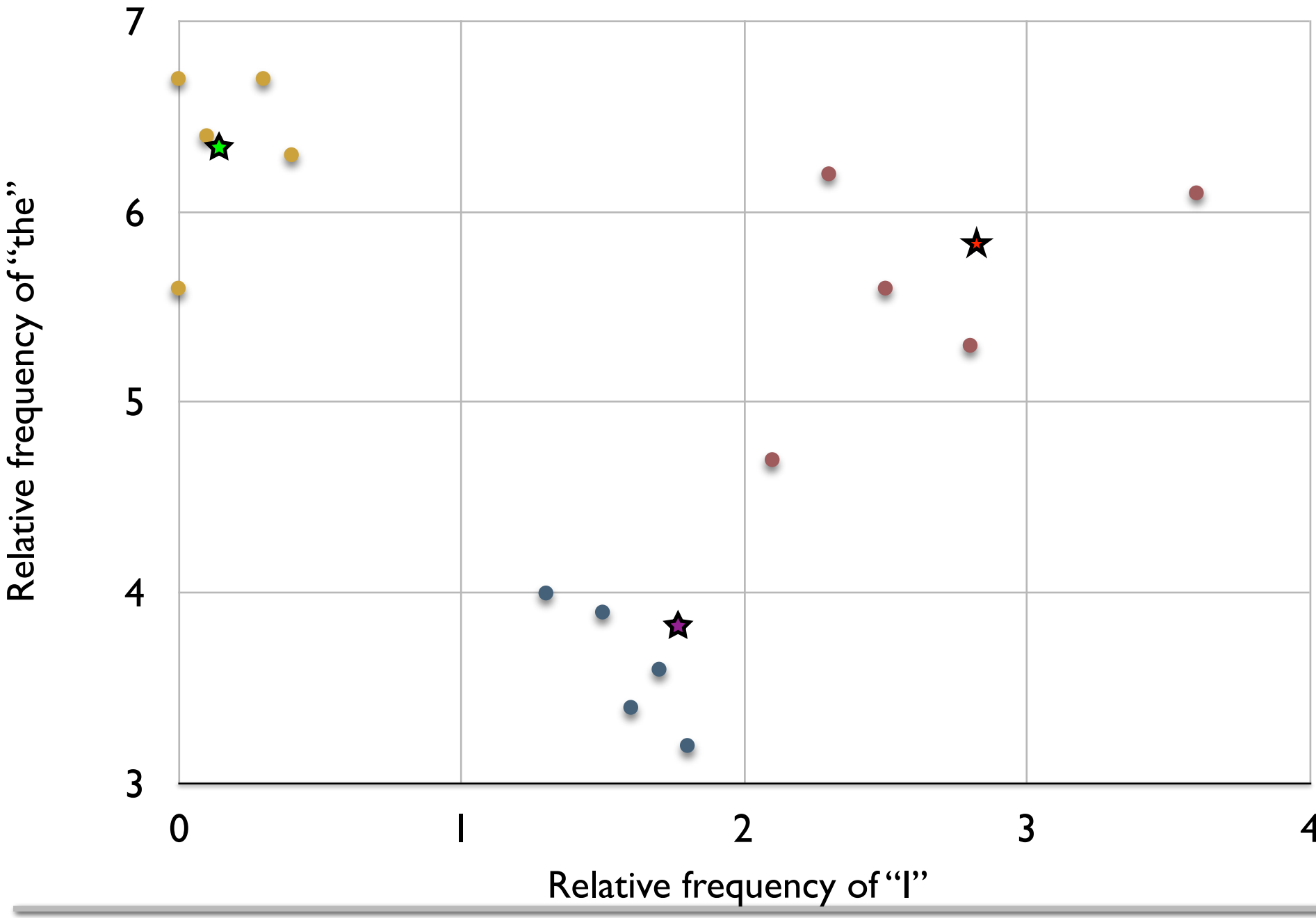
With the right groups



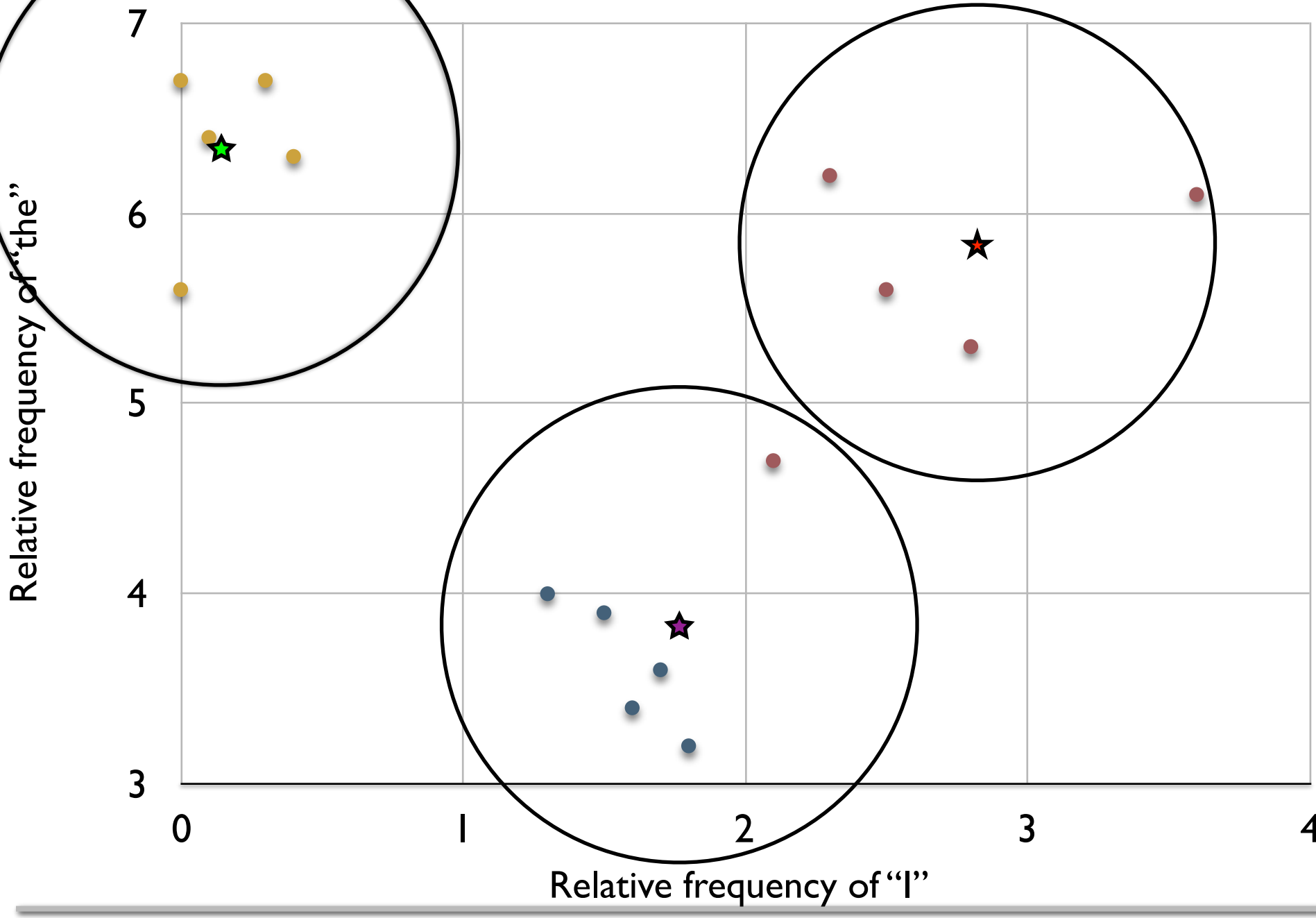
Running k-means on all the documents



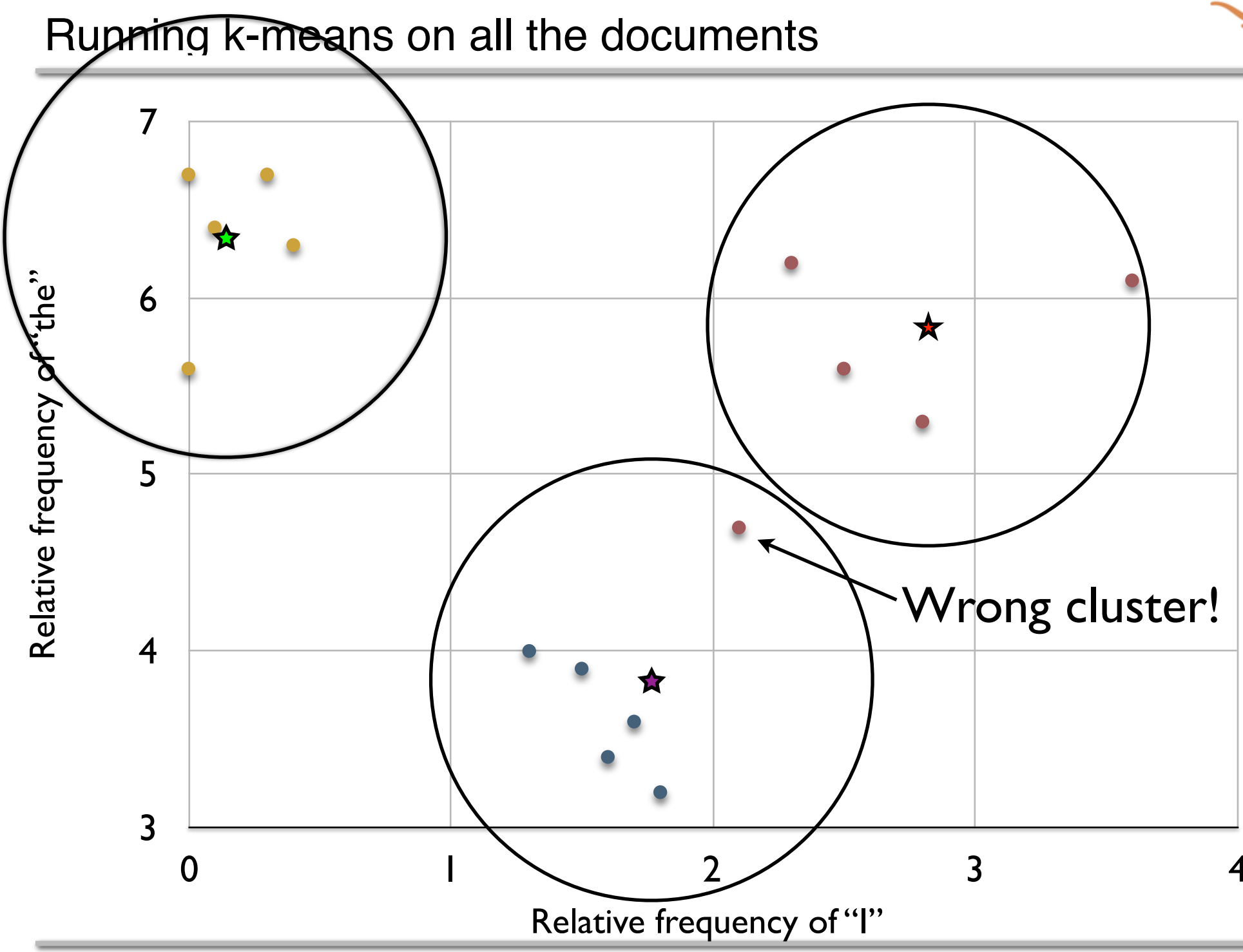
Running k-means on all the documents



Running k-means on all the documents



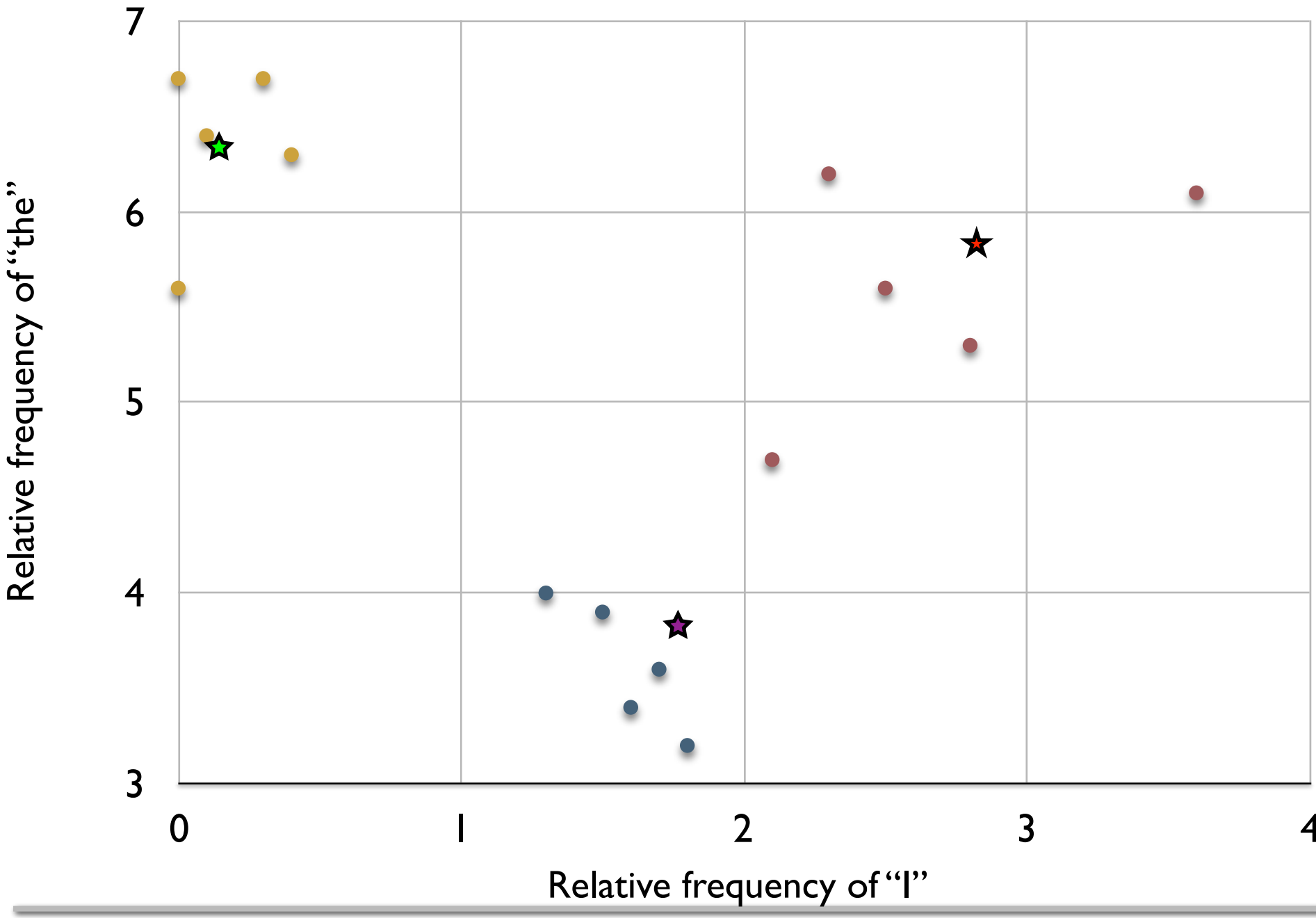
Running k-means on all the documents



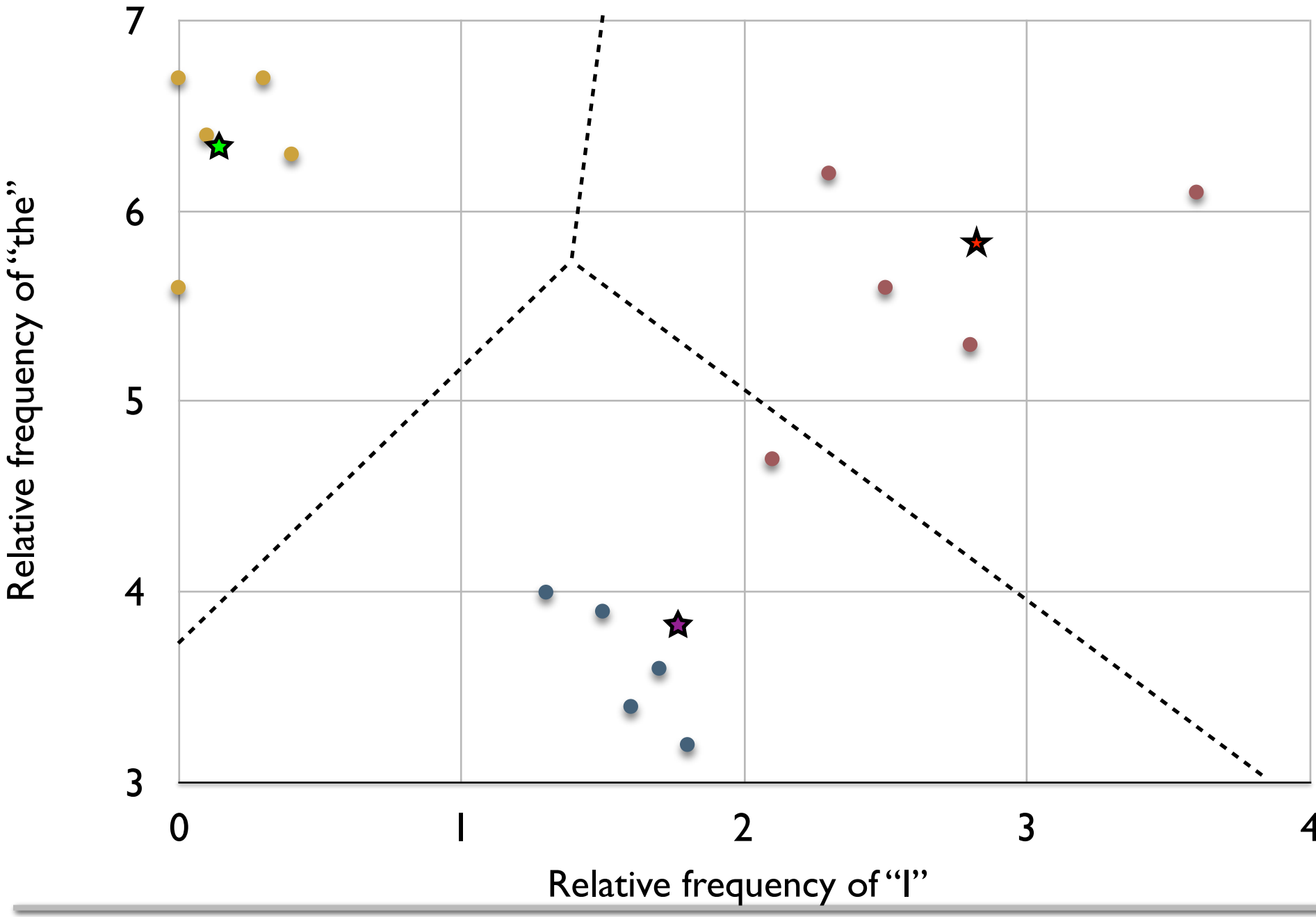
Back to authorship identification

- Features were extracted from the documents and used as values for plotting each document in a multi-dimensional space.
- Documents were then clustered according to K-means (other algorithms could be used).
- K-means gave us a set of centroids, so we can plot other documents into the same multi-dimensional space and compute which one is closest.

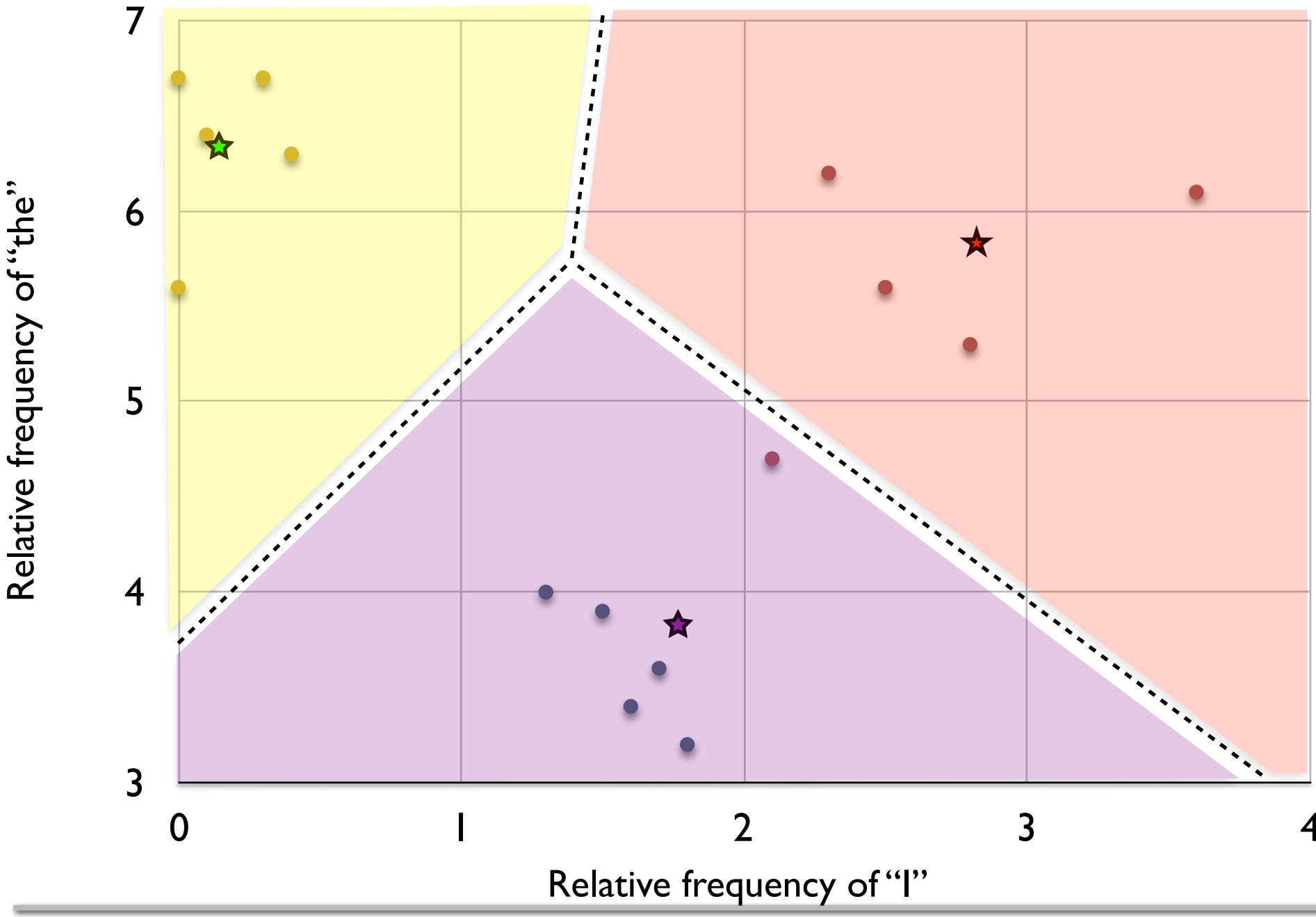
We are given new documents of known authorship



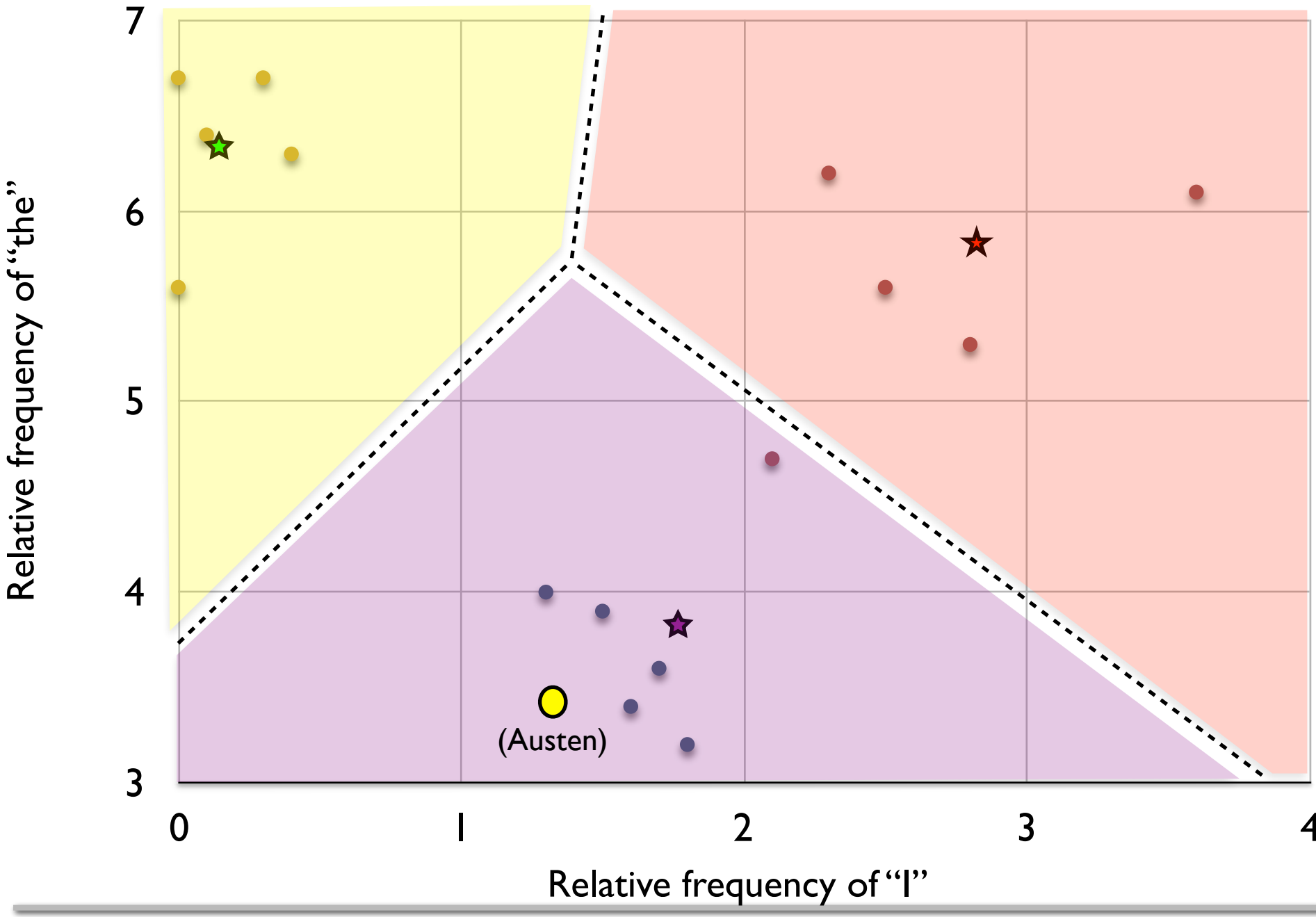
We are given new documents of known authorship



We are given new documents of known authorship



We are given new documents of known authorship

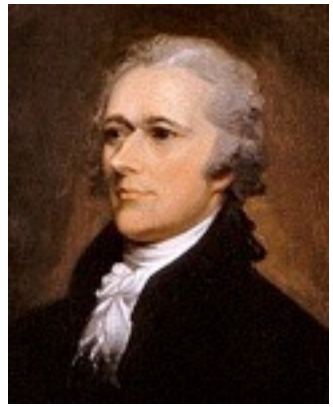


Attribution

- The known documents provide evidence that the clusters we found are sets of documents produced by the author(s) of the known documents.
- We can estimate the confidence in our authorship assignments based on how close the known documents are to each center.

A case study: The Federalist Papers

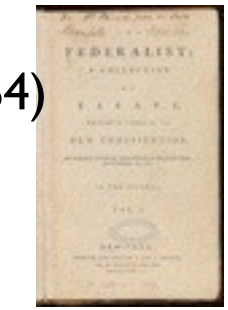
- 85 essays written in 1787 and 1788 arguing for the ratification of the new US constitution by the individual states.
- Three authors, Alexander Hamilton, John Jay and James Madison, all writing under the pseudonym "Publius".
- Later, Hamilton and Madison both claimed to have written a number of the same articles. Scholarship in the 20th century revealed most of them to be Madison's.



Alexander Hamilton
1st US Secretary of the Treasury.
51 articles (nos. 1, 6–9, 11–13, 15–17, 21–36, 59–61, and 65–85); co-authored 18, 19 & 20 w/ Madison.

James Madison
4th US President, "Father of the Constitution"
29 articles (nos. 10, 14, 37–58 and 62–63); co-authored 18, 19 & 20 w/ Hamilton.

John Jay
1st Chief Justice of the US.
5 articles (nos. 2-5 and 64)

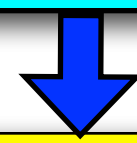


Authorship of the Federalist Papers

- Historian Douglas Adair in 1944 argued that Madison was the author of many of the disputed papers.
- This was confirmed by Mosteller and Wallace in 1964 using a Bayesian classification model.
- Adair's authorship determinations are still generally accepted, though twelve essays are still disputed over by some scholars.
- Experiment: cluster the documents based on all words that occur 5 or more times and k-means. (With some principal components analysis in between).

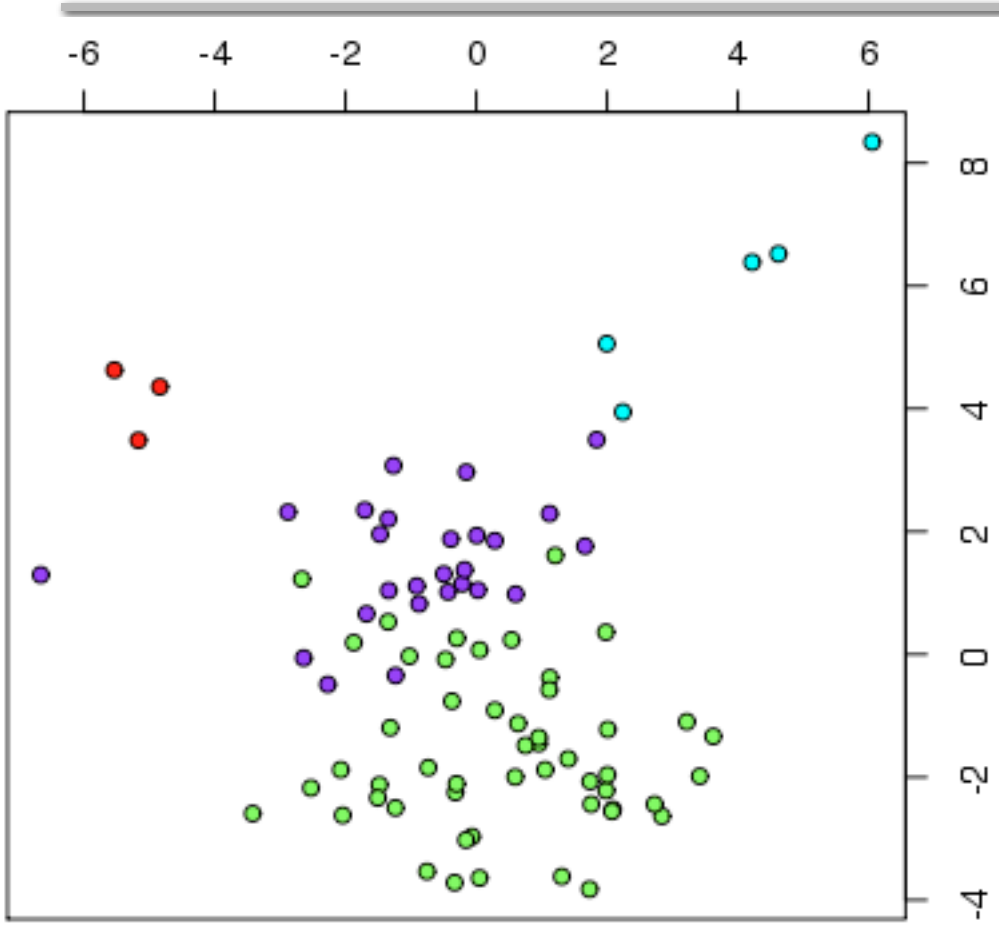
Extracting “features”: frequent words and their counts

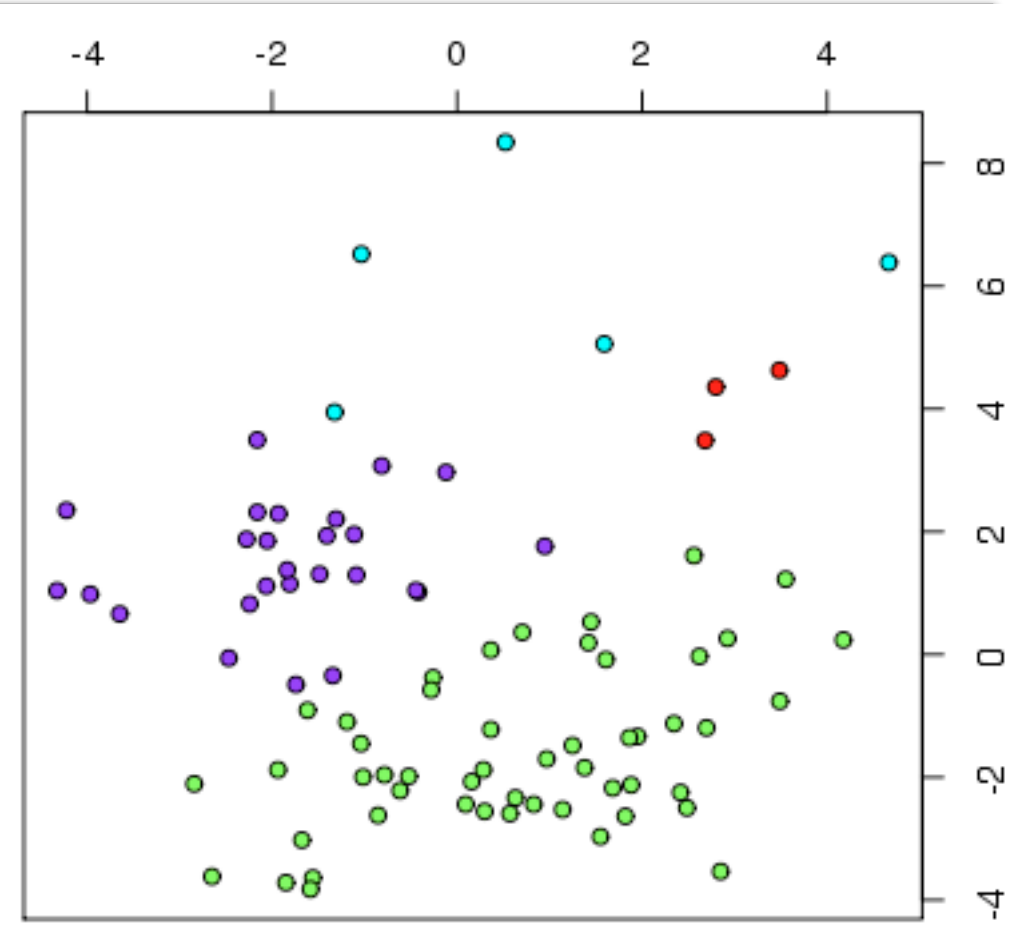
```
FEDERALIST No. 1

General Introduction

For the Independent Journal. Saturday, October 27, 1787


HAMILTON

To the People of the State of New York:

AFTER an unequivocal experience of the inefficacy of the subsisting
federal government, you are called upon to deliberate on a new
Constitution for the United States of America. The subject speaks its
own importance; comprehending in its consequences nothing less than the
existence of the UNION, the safety and welfare of the parts of which it
is composed, the fate of an empire in many respects the most interesting
in the world. It has been frequently remarked that it seems to have been
reserved to the people of this country, by their conduct and example,
to decide the important question, whether societies of men are really
capable or not of establishing good government from reflection and
choice, or whether they are forever destined to depend for their
political constitutions on accident and force. If there be any truth
in the remark, the crisis at which we are arrived may with propriety be
regarded as the era in which that decision is to be made; and a wrong
election of the part we shall act may, in this view, deserve to be
considered as the general misfortune of mankind.

.... <the rest of the essays> ....
```



```
ID Author NumWords . people for jury macedon one power with an as at to more states its be by this upon government them they has the not that than a ; but state courts , is it in
if may have executive would been no constitution and any on of or there all his are from their which will other
1 HAMILTON 1771 49 5 12 0 0 4 2 6 11 10 8 71 7 2 10 34 14 14 6 9 2 6 6 130 14 28 11 25 7 2 5 0 105 13 20 27 4 11 10 0 2 3 3 8 40 6 9 104 6 2 9 0 12 11 14 18 25 3
2 JAY 1841 40 21 13 0 0 10 1 13 1 15 10 52 5 2 5 15 10 14 1 9 4 22 6 105 10 44 5 29 11 8 0 0 122 16 38 34 3 4 17 0 5 8 1 0 83 1 8 81 10 0 4 0 6 4 21 11 2 4
3 JAY 1604 37 7 11 0 0 8 3 10 3 24 1 55 13 11 1 31 18 6 0 16 8 5 5 91 13 20 8 13 7 7 7 3 117 7 21 25 6 6 7 1 2 2 2 0 60 5 6 60 32 1 4 2 8 15 11 11 24 7
4 JAY 1806 32 7 12 0 0 13 2 12 3 20 2 50 13 1 9 26 14 1 0 16 12 17 1 84 14 17 9 16 10 10 5 0 125 10 28 24 12 10 9 0 17 2 1 0 90 5 11 70 24 3 4 2 11 8 19 10 15 11
5 JAY 1475 35 2 7 0 0 10 1 11 3 3 4 44 11 1 4 31 10 6 0 2 11 11 0 64 8 23 9 9 8 4 1 0 86 7 20 28 3 2 1 0 37 0 2 0 72 3 5 51 10 0 4 0 3 11 11 10 6 4
6 HAMILTON 2528 77 3 13 0 0 3 4 15 12 19 6 60 10 7 1 18 12 11 4 2 3 13 10 187 17 24 6 52 10 5 5 0 174 8 11 60 5 3 27 0 6 13 0 1 81 0 5 166 19 8 6 3 18 16 15 24 6 10
7 HAMILTON 2562 81 0 14 0 0 4 2 12 15 19 11 81 6 31 2 47 28 22 11 2 12 7 4 205 14 27 5 48 16 3 7 0 158 14 17 43 12 3 23 0 51 8 3 1 51 7 13 156 10 9 12 0 7 11 18 24 1 9
8 HAMILTON 2306 72 8 9 0 0 6 5 13 13 16 11 79 8 10 10 35 11 19 3 3 11 13 8 156 12 18 3 46 17 10 6 0 164 21 21 42 7 8 18 2 27 16 4 4 54 2 11 133 17 2 9 0 14 9 16 26 11 6
9 HAMILTON 2213 68 2 10 0 0 12 3 14 12 18 10 70 6 12 5 26 13 14 4 17 5 20 12 168 13 24 1 46 14 6 6 1 133 16 23 37 6 5 24 0 8 15 2 3 45 4 9 153 11 3 8 4 14 7 14 25 7 5
10 MADISON 3337 79 4 18 0 0 8 1 11 14 20 8 99 12 3 9 61 39 11 0 13 8 11 4 259 14 31 13 78 32 11 2 0 221 41 47 63 6 16 16 0 6 9 5 2 121 4 18 155 22 6 4 8 26 12 21 39 30 22
11 HAMILTON 2766 78 0 18 0 0 5 3 21 15 21 11 82 9 13 6 44 20 24 6 5 6 10 8 186 20 24 6 69 8 8 4 0 173 9 21 66 6 9 15 0 50 3 5 0 70 4 5 178 12 8 13 1 9 19 11 25 17 10
12 HAMILTON 2410 72 3 8 0 0 5 1 18 11 17 13 81 5 10 8 40 15 17 7 8 6 8 13 174 6 27 4 47 15 8 11 0 161 22 26 54 7 3 13 0 22 8 2 0 62 7 12 139 7 9 5 2 15 19 14 23 11 14
13 HAMILTON 1045 28 2 1 0 0 8 5 10 3 8 1 42 6 9 2 25 5 5 2 7 1 3 2 72 5 18 10 26 6 2 3 0 53 9 6 14 6 8 4 0 14 2 4 0 17 3 3 56 3 9 3 0 4 5 1 11 9 1
14 MADISON 2357 53 5 20 0 0 2 3 7 9 24 7 71 8 12 8 45 18 13 0 9 4 17 8 200 13 33 5 37 19 4 1 0 144 25 31 40 6 16 17 0 5 9 10 2 60 3 17 122 11 0 6 0 6 11 19 40 26 7
15 HAMILTON 3395 83 4 18 0 0 1 9 19 18 15 24 116 6 11 8 44 32 24 10 14 4 15 14 251 16 38 10 69 25 8 5 1 186 48 30 73 7 7 31 0 13    6 4 74 3 10 194 38 18 21 1 22 22 7 55 19 6
< ... 70 more lines for the other essays ...>
```

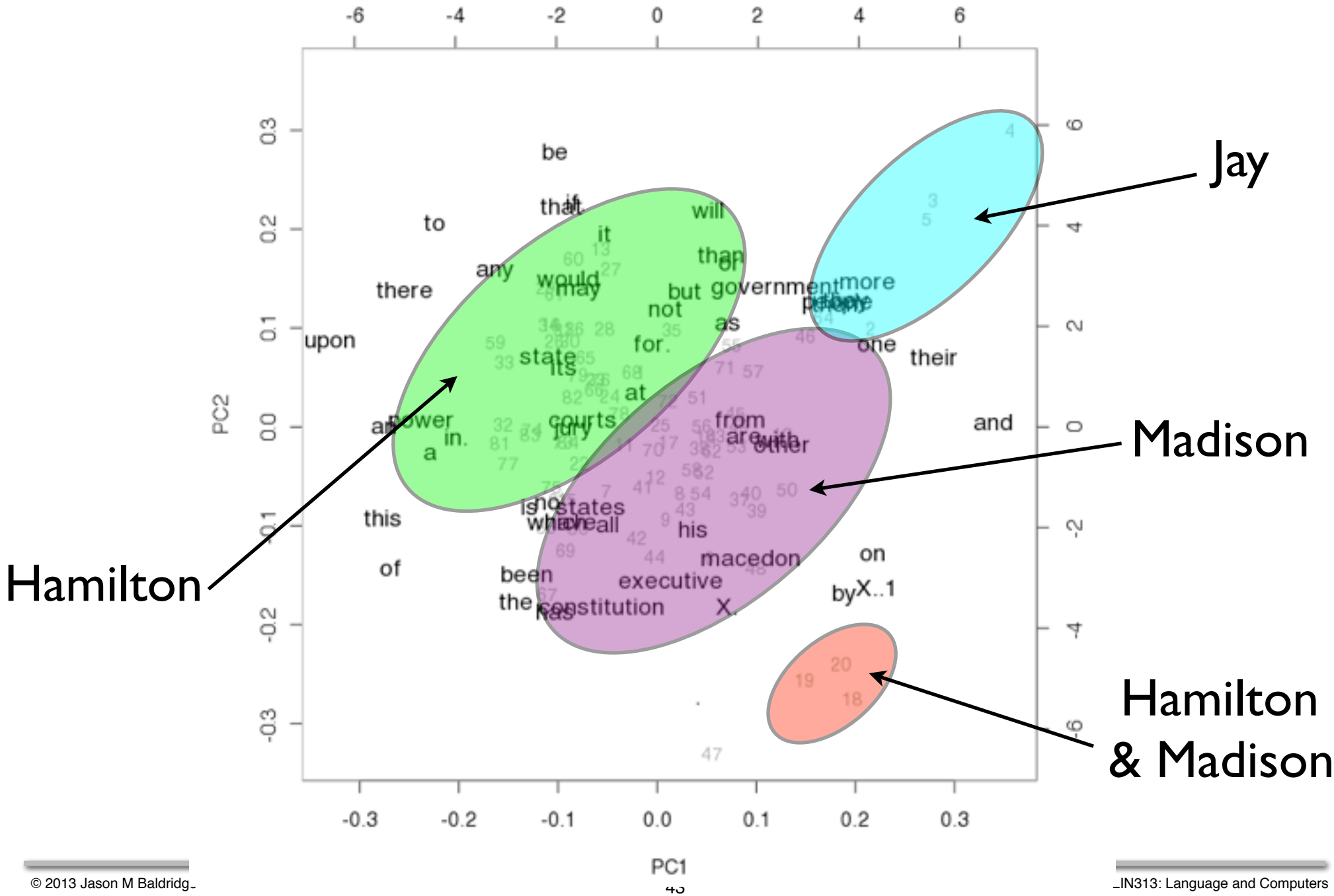
Dimensionality reduction: principal components analysis





Author	Color
Hamilton	green
Madison	purple
Jay	cyan
Hamilton & Madison	red

Biplot of features and documents (components 1 and 2)



K-means clustering

Three clusters (k=3): the model thinks the collaborations were most similar to Madison, plus three of Hamilton's.

	Author			
Cluster	COLLABMH	HAMILTON	JAY	MADISON
1	3	3	0	26
2	0	48	0	0
3	0	0	5	0

K-means clustering

Three clusters (k=3): the model thinks the collaborations were most similar to Madison, plus three of Hamilton's.

	Author			
Cluster	COLLABMH	HAMILTON	JAY	MADISON
1	3	3	0	26
2	0	48	0	0
3	0	0	5	0

Four clusters (k=4): Cleanly clustered. One article (#47) attributed to Madison is astray, but it sensibly goes with the collaborative articles.

	Author			
Cluster	COLLABMH	HAMILTON	JAY	MADISON
1	3	0	0	1
2	0	0	5	0
3	0	0	0	25
4	0	51	0	0

- With no changing of the features or parameters, amazingly close to the authorship attributions given by Adair.

K-means clustering of the Federalist Papers

- Try it yourself! (if you know Scala):
  - http://ata-s12.utcompling.com/assignments/clustering