From text to vectors:
Automating the analysis of interview data

Bruce Sherin
Northwestern University
bsherin@northwestern.edu

TRUSE

June, 2010

The usual methods

- As researchers and practitioners, we'd like to know what's going on inside the mind of the science student.
- One way to do this: The medium of words.
- How do we get from the words to hypotheses about student thinking?
  - We use ourselves as scientific instruments.

The usual methods

- Interview students and videotape the interviews
- Code the videos:
  - Transcribe videos
  - Segment the transcripts
  - Induce a coding scheme
  - Apply the coding scheme to the segmented transcript

○ Tacitly assumed that humans must do the coding

  - To apply a coding scheme, need to understand natural language, pay attention to gestures, etc.
  - Inducing the coding scheme is even harder

Can any of this be automated?

- We want to automate both:
  - The induction of the coding scheme
  - Application of the coding scheme to code transcripts
- Why automation would be a good thing:
  - Coding is a lot of work
  - More importantly: Some support for human analysis
- Why it's a good time to investigate this:
  - Advances in computational linguistics
  - We have powerful computers

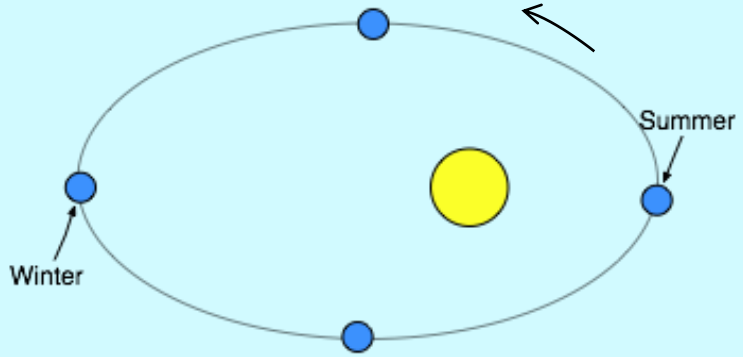
What's coming in this talk

- The data corpus: Interviews in which middle school students asked to explain the seasons.
- Computational linguistics: Vector space models
- Two categories of automated analysis:
  1. Given a coding scheme developed by human analysts, apply the scheme to a data corpus
  2. Both induce and apply a coding scheme
- This is super easy!

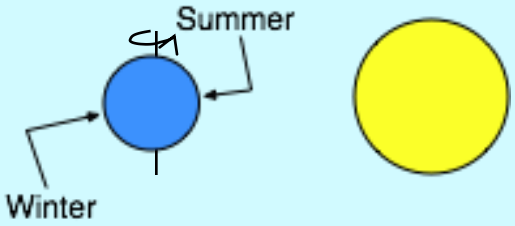
Data corpus: Explaining the seasons

- A collection of interviews in which middle school students were asked to explain the seasons.
- Our interview protocol, in brief:
  1. “Why is it warmer in the summer and colder in the winter?”
  2. Follow up questions for clarification
  3. Asked to make a drawing.
  4. Challenges for certain answers

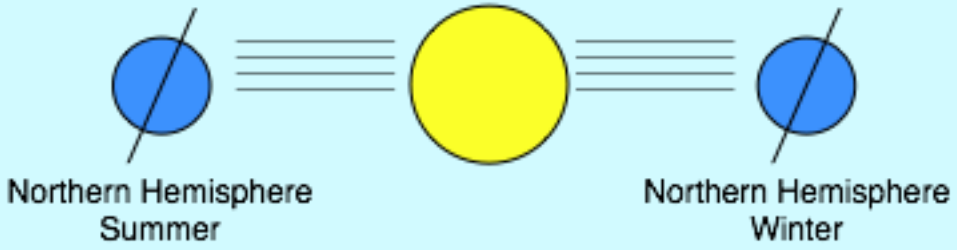
Three categories of explanations



Closer-farther explanations

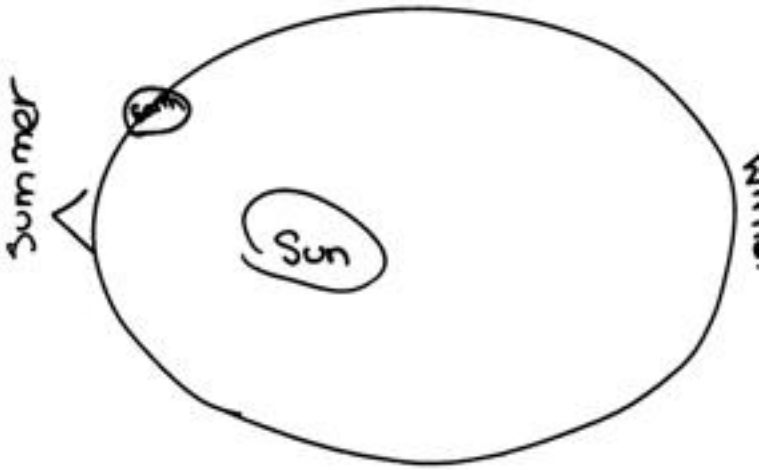


Side-based explanations



Tilt-based explanations

Jill gives a closer-farther explanation



Human coding

	CF	Side-based	Tilt-based	Shift	Total
First Coder	5	8	4	4	21
Second Coder	5	7	4	5	21

Kappa = .94 (almost perfect agreement)

Vector Space Models

Vector space models

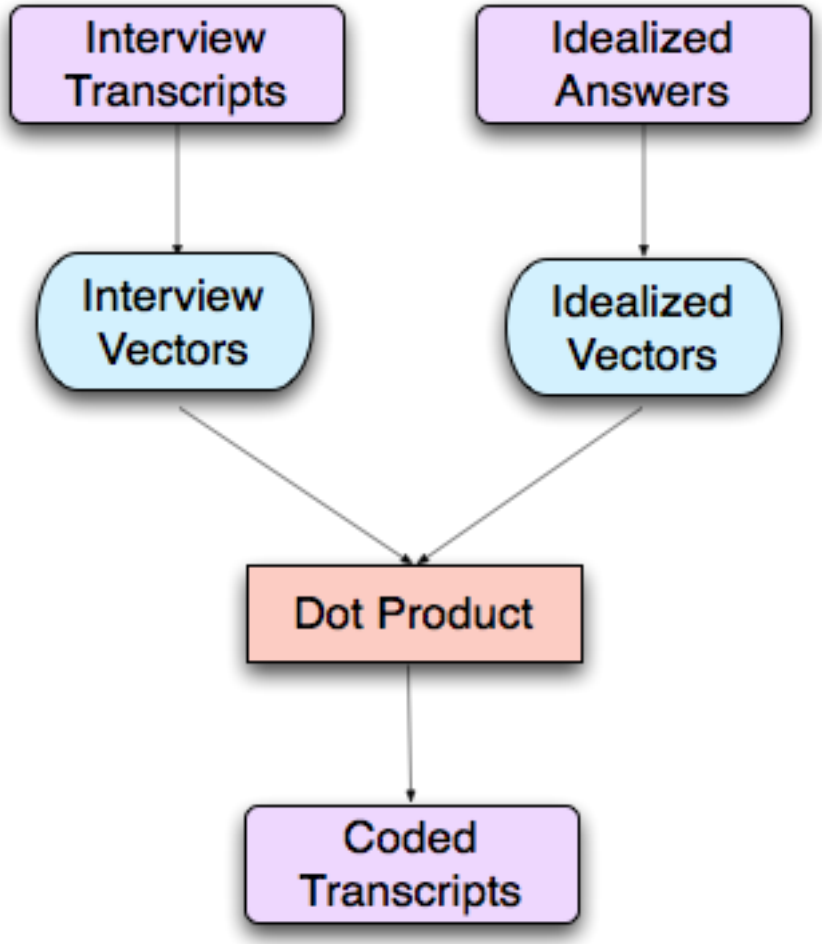
- A passage of text is mapped onto a vector, typically in a high dimensional space.
- The direction the vector points corresponds to the meaning of the passage
- To find the similarity in meaning between two passages, find the dot product of the corresponding vectors.

Map a passage to a vector

sun	3
earth	2
yeah	0
winter	1
summer	2
…	…

Its because the sun um we rotate around the sun like in an axis but its not a perfect circle and when and then like or not an axis like we orbit its like not a perfect circle its like egg shaped almost but not very noticeable and the sun the earth is on an axis on that orbit that when it goes around like there one part that closer and one part that farther so that kind of that explains why ...

Analyses that apply coding categories developed by humans

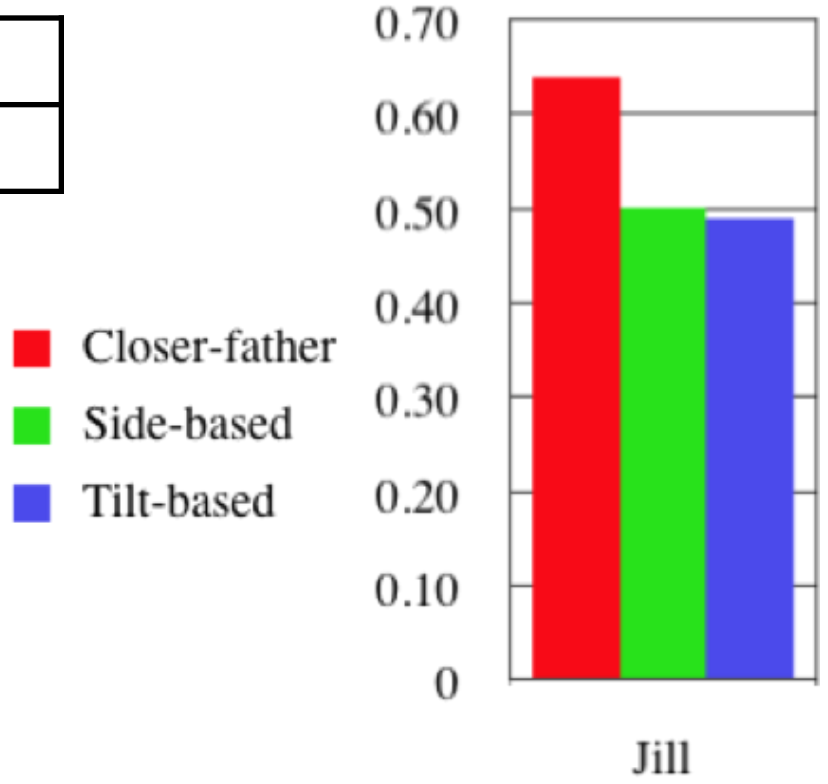


The earth orbits around the sun It takes one year for it to go around The earth orbits in an ellipse so that sometimes the earth is closer to the sun and sometimes it farther away from the sun When the earth is over here it closer to the sun it gets more heat so that makes it warmer and its summer When the earth is over here it farther from the sun it gets less heat from the sun and it colder So that when it winter

Coding the Jill transcript

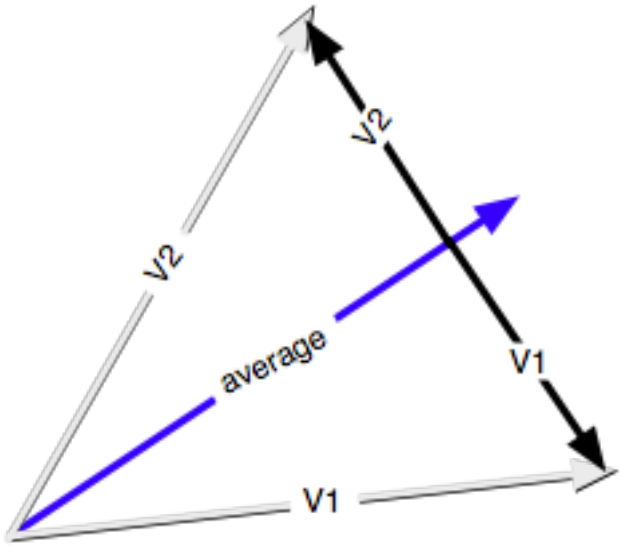
Dot products between Jill's transcript vector and the idealized response vectors.

Closer-father	Side-based	Tilt-based
0.64	0.50	0.49

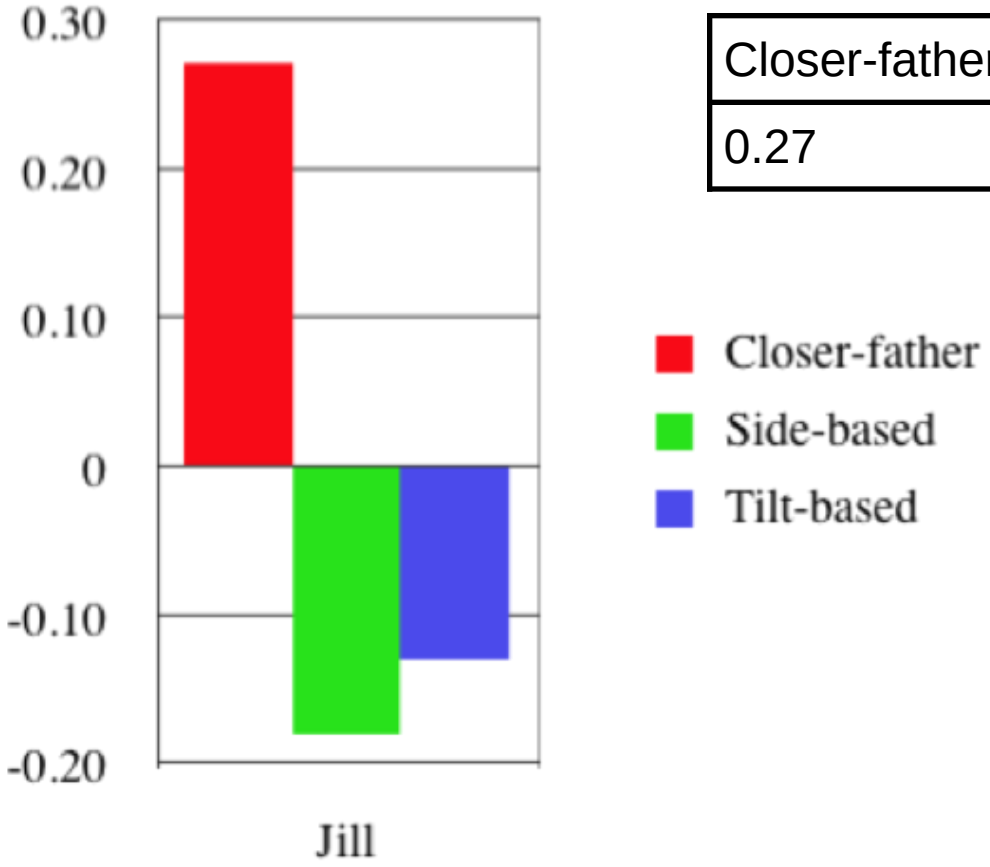


Deviationalization™

Deviationalize: Average the three vectors for the idealized answer documents and replace each one with their deviation from that average.

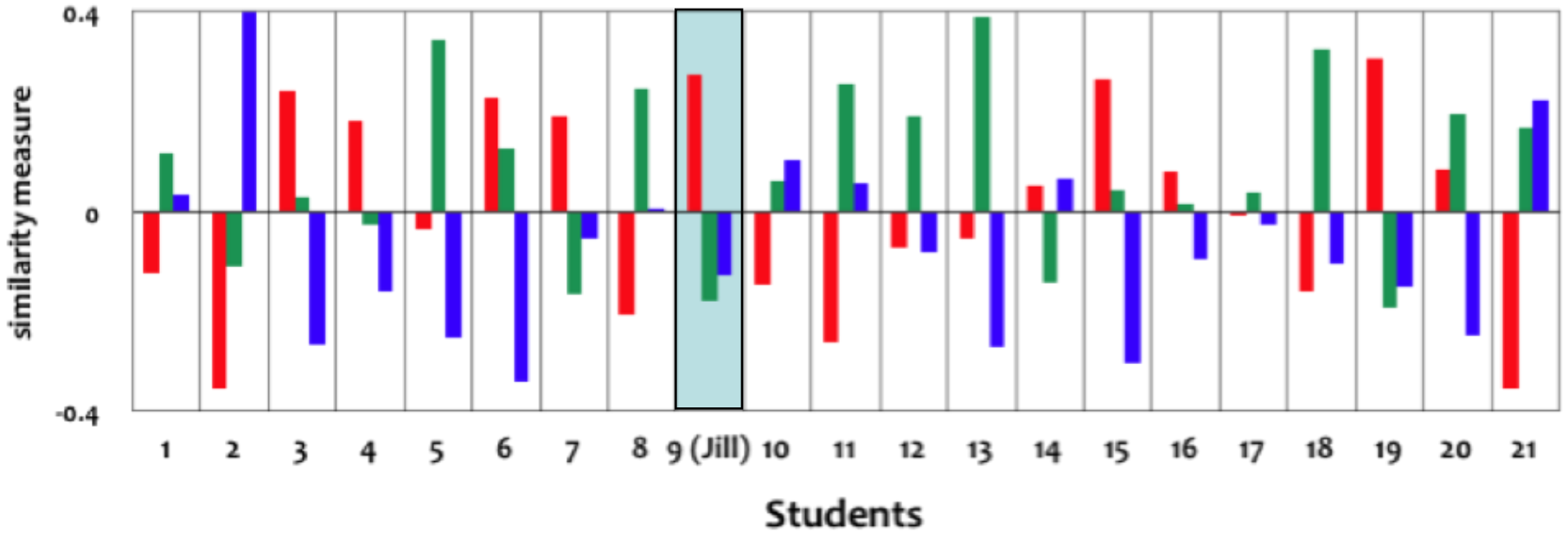


Jill Coded (after deviationalization)



Closer-father	Side-based	Tilt-based
0.27	-0.18	-0.13

Coding all 21 transcripts



	Agree	Disagree	Kappa
Student only transcript	14	2	0.81
Student+Interviewer	13	3	0.70

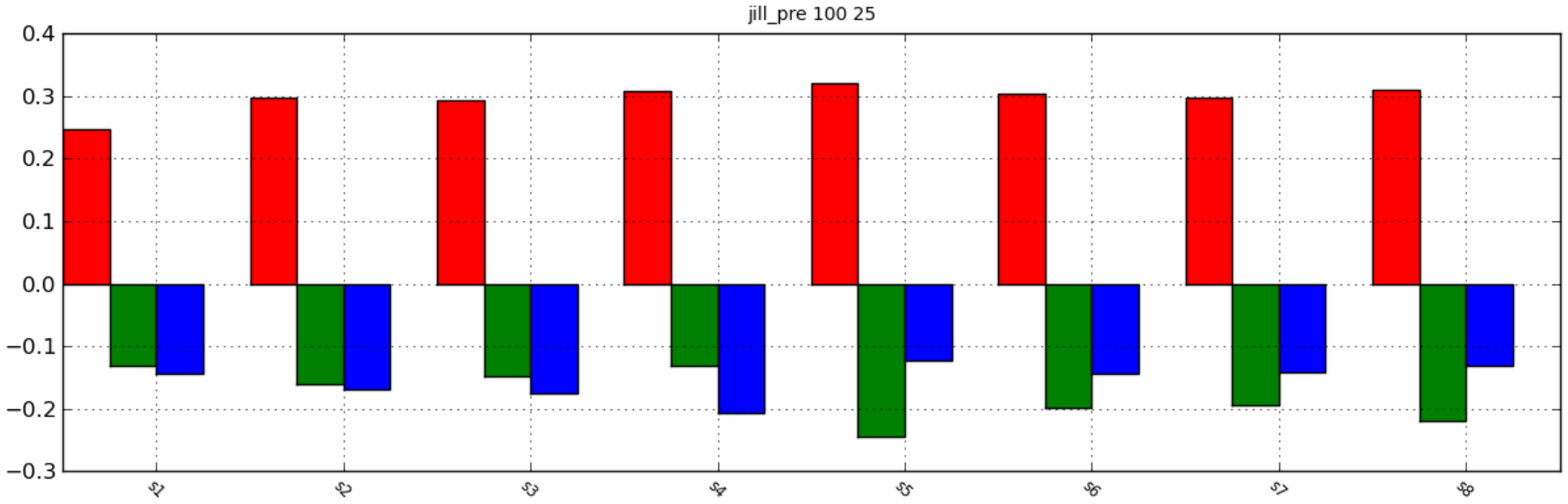
Leslie works it out

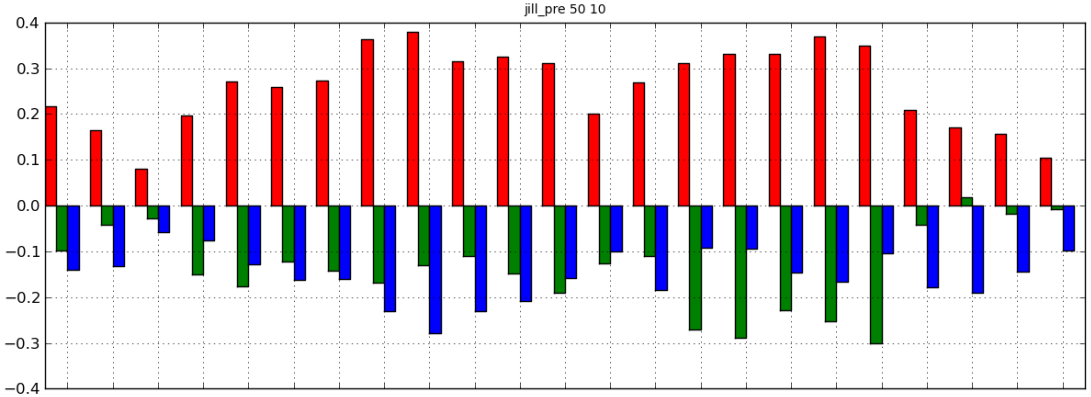
Coding segments of transcripts

- Assigning one code to an interview is a significant approximation
  - Students sometimes shift explanations
  - Students sometimes develop an explanation over a few minutes
- When researchers code, we frequently segment a transcript at a finer grain size.
- So: Slice the transcript documents into little segments and code those.
- Really my goal was to be able to code shifts. But I haven't succeeded in that yet.

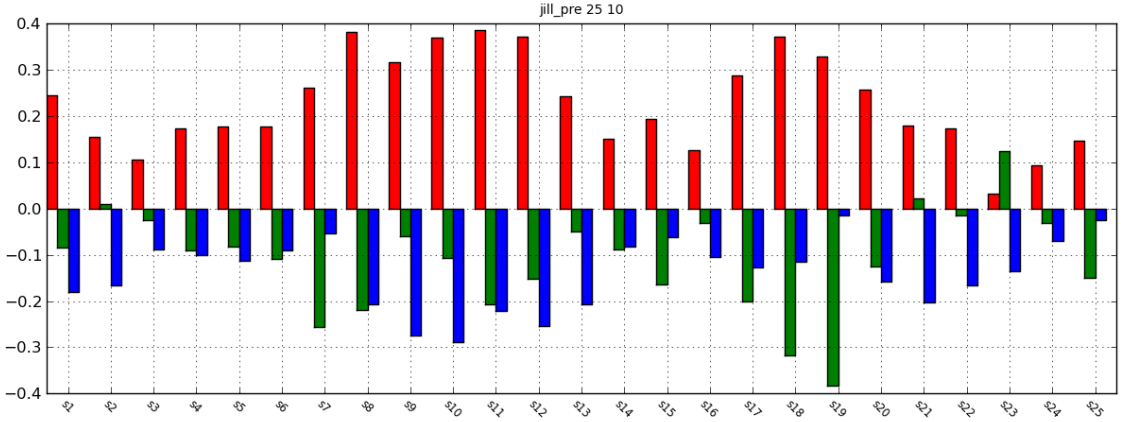
Jill segmented

- Slice Jill's transcript into:
  - 100-word segments
  - Stepping forward by 25 words

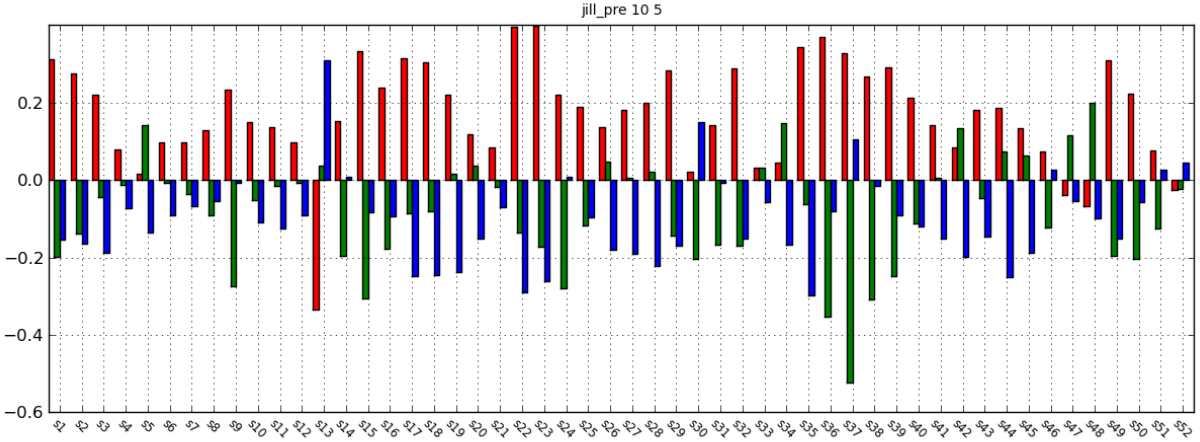




50 words
Step size: 10

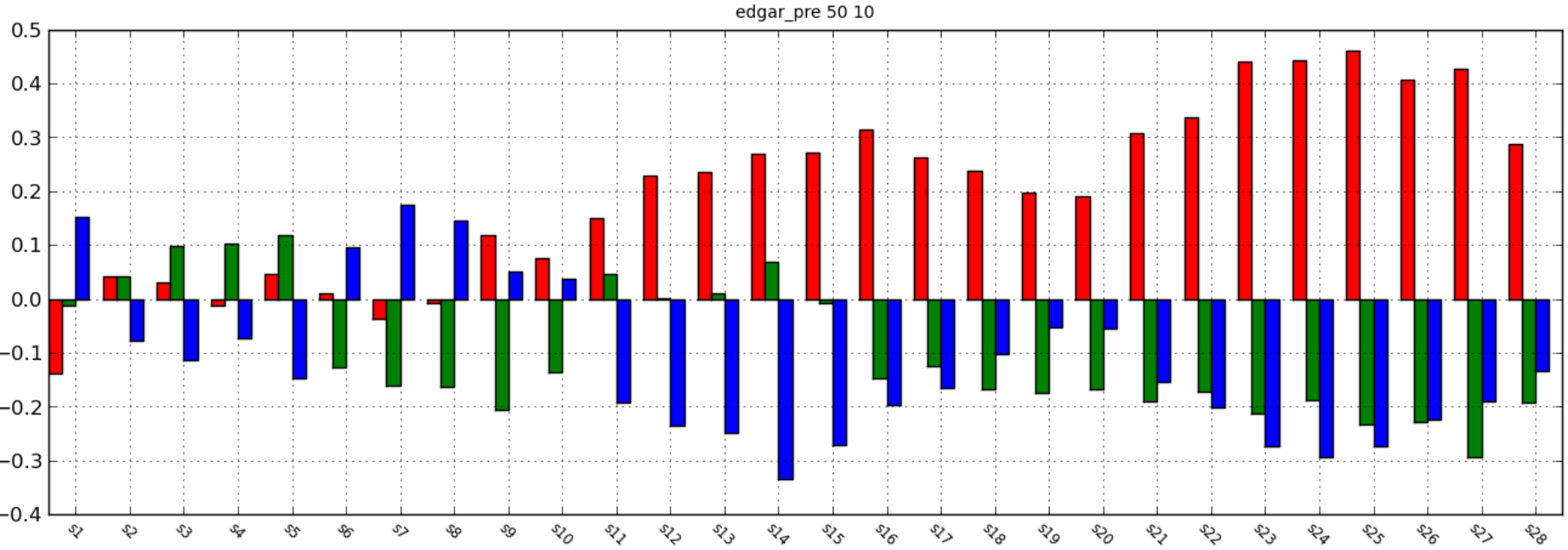


25 words
Step size: 10



10 words
Step size: 5
words

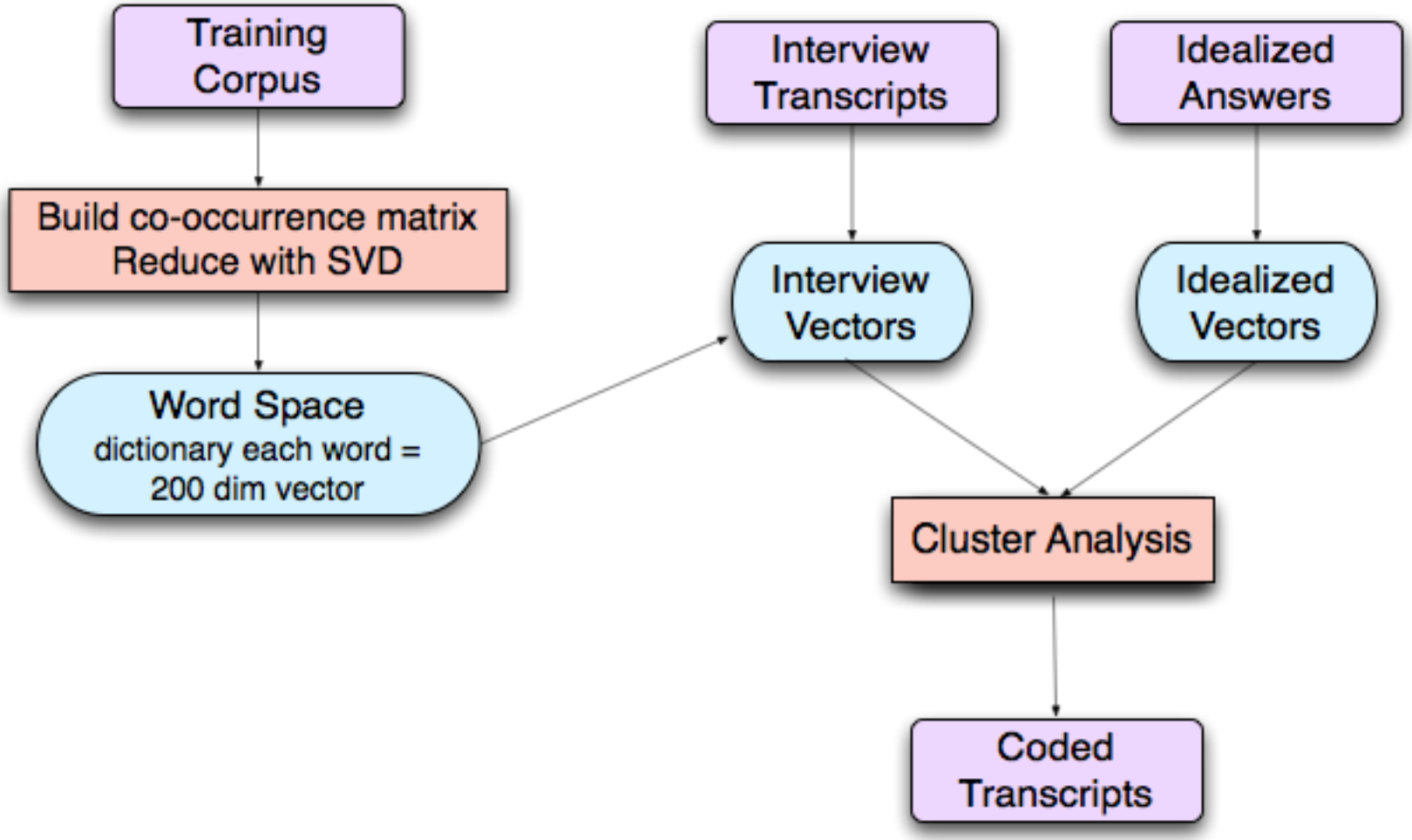
Edgar segmented

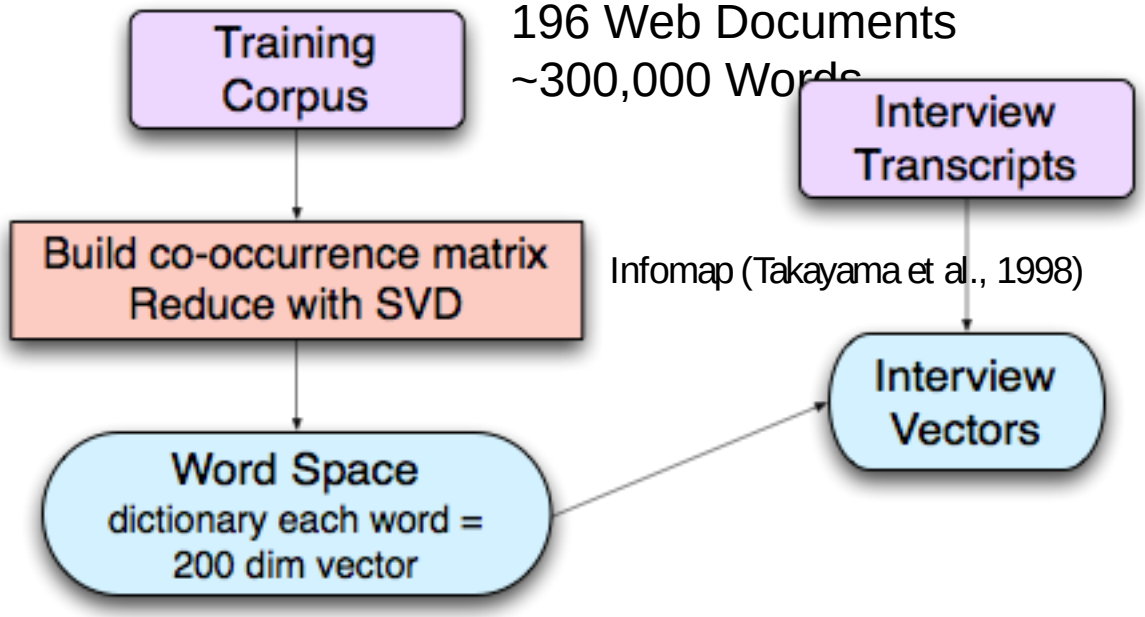


Inducing a coding scheme with cluster analysis

Clustering transcripts

- The preceding method relies on humans to develop the coding scheme, as embodied in the “idealized answer” documents.
- What we’d really like is for the computer to both invent the coding schema and code.
- This needs cluster analysis, plus some more elaborate methods to map passages to vectors.


Training Corpus
Build co-occurrence matrix
Reduce with SVD
Word Space
dictionary each word =
200 dim vector
Interview Transcripts
Idealized Answers
Interview Vectors
Idealized Vectors
Cluster Analysis
Coded Transcripts



Its because the sun um we rotate around the sun like in an axis but its not a perfect circle and when and then like or not an axis like we orbit its like not a perfect circle its like egg shaped almost but not very noticeable and the sun the earth is on an axis on that orbit that when it goes around like there one part that closer and one part that farther so that kind of that explains why ...

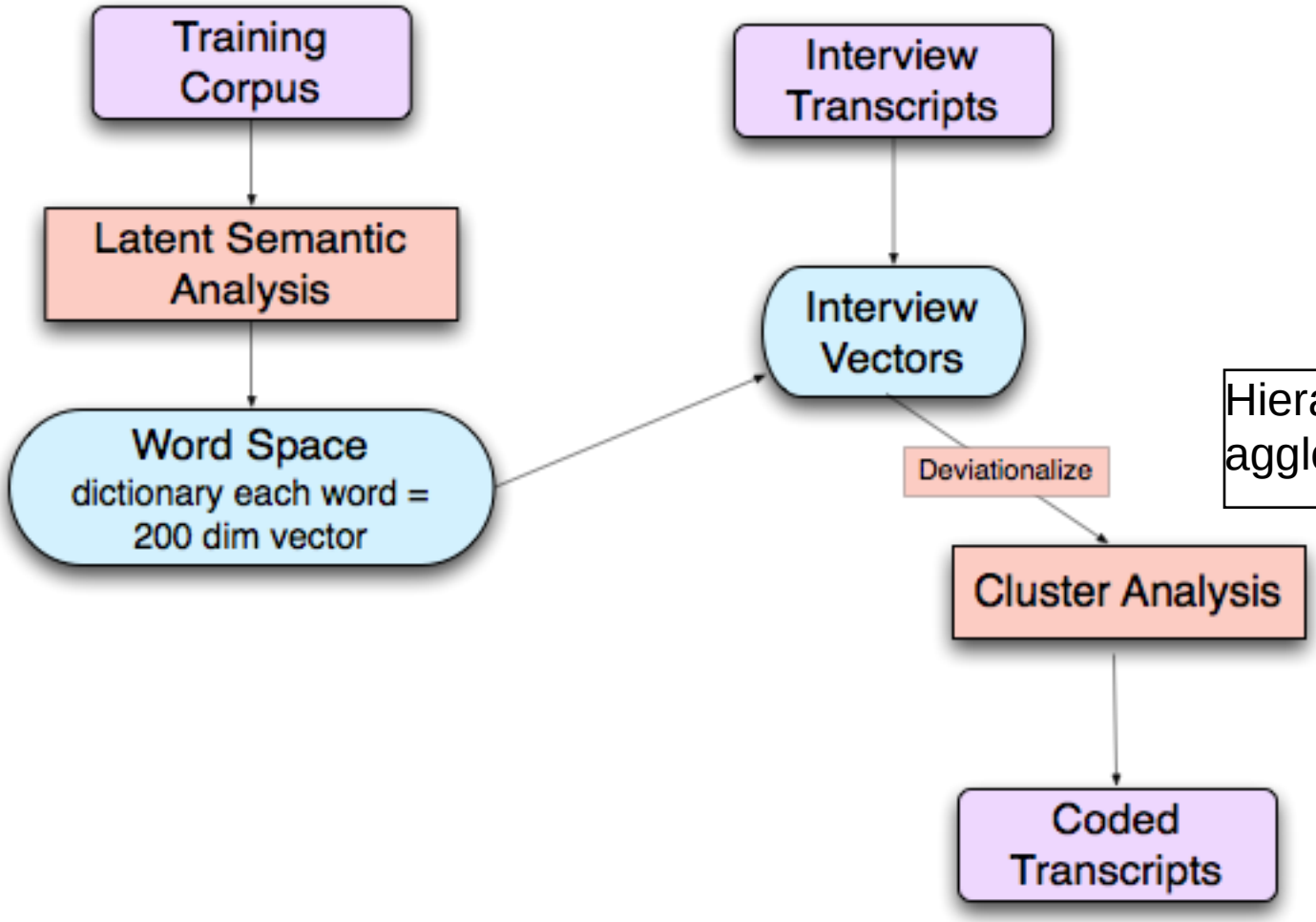
51th to 1050th most common words

200 dimensions (or whatever)

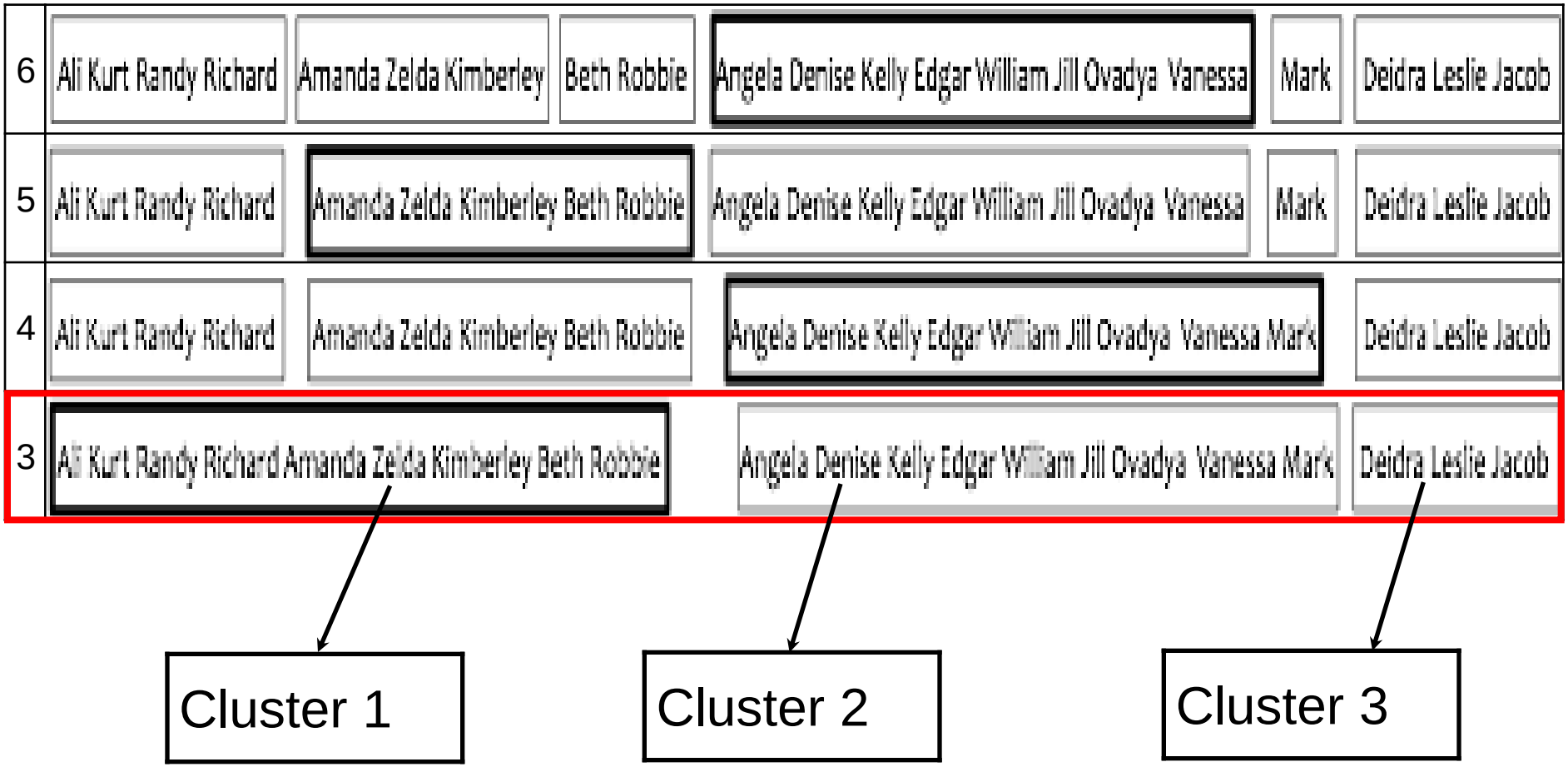
	rotate	kind	radiate	...
sun	25	12	2	...
	0.268472	0.304940	0.183209	...
earth	7	82	2	...
	-0.296295	0.209063	0.131982	...
solar	11	7	62	...
solar	0.231243	0.034267	-0.020067	...
...				
...				

200-dimension vector (list of 200 numbers)

"transcript vector"


Training Corpus
Latent Semantic Analysis
Word Space
dictionary each word = 200 dim vector
Interview Transcripts
Interview Vectors
Deviationalize
Cluster Analysis
Coded Transcripts
Hierarchical agglomerative clustering


21 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
20 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
19 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
18 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
17 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
16 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
15 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
14 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
13 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
12 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
11 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
10 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
9 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
8 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
7 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
6 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
5 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
4 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob
3 Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark Deidra Leslie Jacob



- We've effectively coded the transcripts by sorting into three categories. But what do the categories mean? Do they align with the human-derived scheme?

What do the clusters mean?

- Find the words in the training corpus with the largest dot product to the centroid of each cluster.

Cluster 1	
tilted	0.485
away	0.420
kind	0.329
towards	0.283
mentioned	0.250
angles	0.220
facing	0.220
hemisphere	0.219
axis	0.212
incident	0.212

Cluster 2	
closer	0.454
summer	0.333
winter	0.314
brings	0.214
farther	0.199
northwest	0.199
summers	0.194
pink	0.193
purple	0.192
reaches	0.191

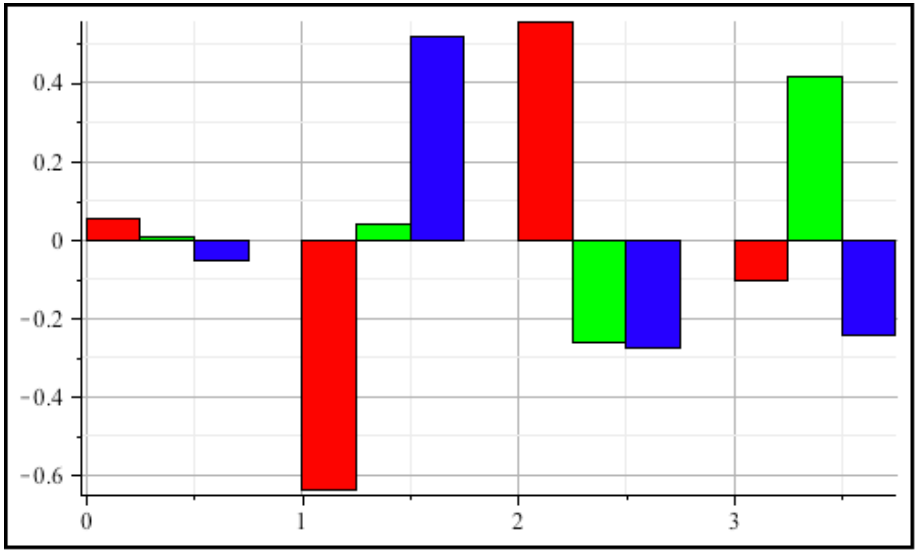
Cluster 3	
side	0.494
day	0.400
night	0.395
time	0.299
moon	0.289
rotates	0.271
poster	0.264
lunar	0.240
shadow	0.230
rotation	0.227

What do the cluster's mean?

Dot products between cluster centroids and idealized response vectors

	Agree	Disagree	Kappa
No Interviewer	10	6	0.55
With Interviewer	10	6	0.54

6	Ali Kurt Randy Richard	Amanda Zelda Kimberley	Beth Robbie	Angela Denise Kelly Edgar William Jill Ovadya Vanessa	Mark	Deidra Leslie Jacob
5	Ali Kurt Randy Richard	Amanda Zelda Kimberley Beth Robbie		Angela Denise Kelly Edgar William Jill Ovadya Vanessa	Mark	Deidra Leslie Jacob
4	Ali Kurt Randy Richard	Amanda Zelda Kimberley Beth Robbie		Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark		Deidra Leslie Jacob
3	Ali Kurt Randy Richard Amanda Zelda Kimberley Beth Robbie			Angela Denise Kelly Edgar William Jill Ovadya Vanessa Mark		Deidra Leslie Jacob



Cluster 1	
kind	0.486
planets	0.257
axis	0.249
motions	0.245
away	0.240
movement	0.232
ecliptic	0.223
learned	0.216
objects	0.214
planetary	0.198

Clustering transcripts

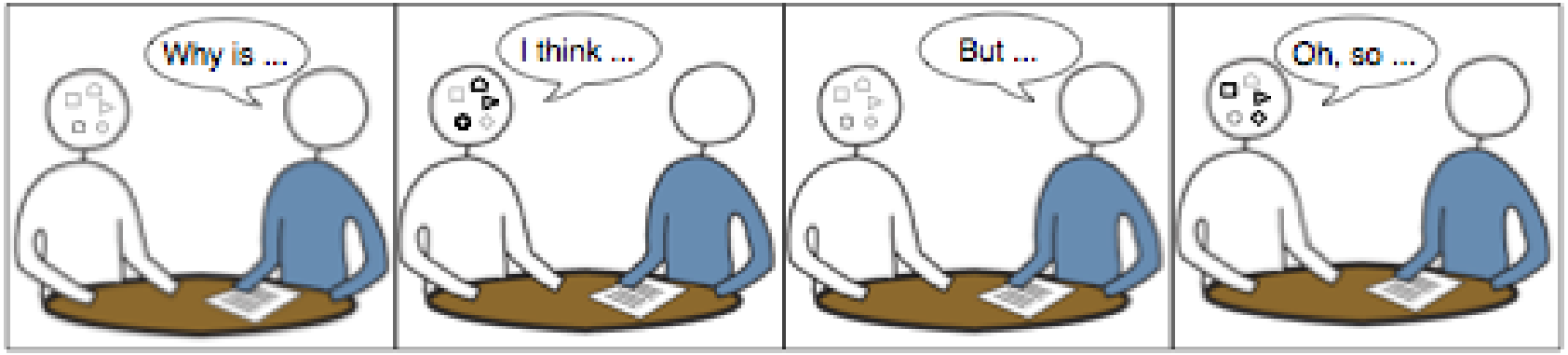
- Comparison to human coders:

	Agree	Disagree	Kappa
Four Clusters	11	5	0.54
Six Clusters	12	4	0.62

Bottom line:

- For coding individual transcripts, agreement with human coders is so-so.
- Ability to induce a coding scheme is tantalizing.

Another complication

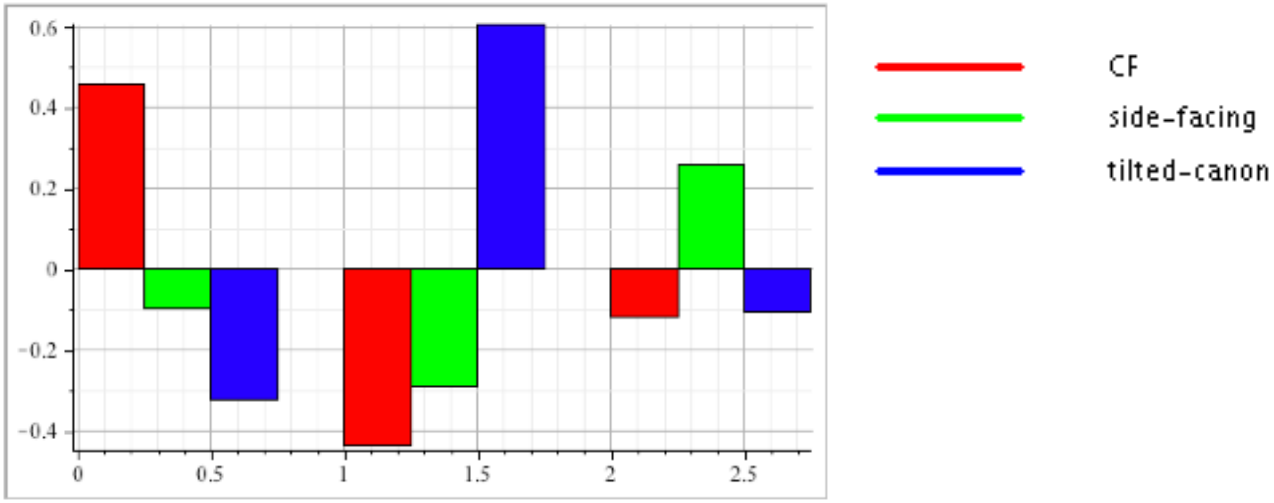


- What we really want to get at is underlying knowledge and processes of assembly.

Clustering segments of documents

- Try to find meanings at a smaller grain size in transcripts.
  1. Cut up all of the transcripts into small (25-word) segments.
     - End up with 606 segments.
  2. Compute the meaning of each in terms of their 200 dimension vector.
  3. Use cluster analysis to pull these into groups.

Three clusters of segments



sun	0.557854	2180
orbit	0.315278	548
earth	0.286494	1895
~	0.196327	25
solar	0.179732	1069
ellipse	0.176945	25
distances	0.168432	29
celestial	0.166542	126
evening	0.165143	11
morning	0.164609	16

tilted	0.475581	239
kind	0.460649	21
striking	0.349523	21
concentrated	0.304906	37
towards	0.280476	91
sunlight	0.279226	250
warm	0.278318	166
hotter	0.267920	32
hit	0.257450	42
angle	0.254329	219

winter	0.668527	622
summer	0.624533	623
hemisphere	0.312125	589
southern	0.291326	293
northwest	0.287605	10
northern	0.269531	551
day	0.265189	550
autumn	0.239734	53
night	0.229601	199
brings	0.208714	22

6 clusters of segments

Cluste	Cluste	Cluster 3		te	Cluster 5		
sun	tilted	winter	0.736		kind	0.893	.832
orbit	hitting	summer	0.702		planetary	0.234	.290
earth	area	hemisphere	0.351		axial	0.231	.265
ellipse	light	northern	0.349		tilt	0.229	.234
ecliptic	spread	southern	0.334		mentioned	0.221	.224
solar	sunlight	northwest	0.309		anticyclones	0.218	.214
~	hit	december	0.227		meteorological	0.209	.212
distances	rays	close	0.217		indication	0.203	.202
morning	concentrated	autumn	0.212		objects	0.199	.198
evening	striking	brings	0.209		manner	0.199	.188

Conclusions

Summary

- Typical methods
  - The field has tacitly assumed that you need humans to apply a coding scheme.
  - Inducing the coding scheme should be even harder
- The techniques described here:
  - Had no access to gestures, diagrams, facial expressions, etc.
  - They discarded still more information (e.g., word order)
- Nonetheless: Some relatively simple computational techniques can apparently do significant work

Summary

- Coded entire transcripts using idealized answers
  - Moderate agreement with human coders
- Coded segmented transcripts using idealized answers
  - Could not capture shifts reliably
  - Possible to work with segments as small as 10 words
- Induced a coding scheme by clustering transcripts
  - Coding scheme induced aligned with the human-derived scheme.
  - Coding of individual transcripts moderately good.
- Clustered segments of transcripts
  - It was possible to interpret clusters of 25-word segments.

Implications

- In the short term, the big win will not come from replacing humans in applying a coding scheme.
- Big win will come from *support* for human analysis
  - This is especially true for the automated analyses that induce the coding scheme
- Thus, paradoxically, the most immediate win may come from the computational analyses that seem more difficult, those that induce a coding scheme.

Outstanding issues and future work

- How should I deal with the huge parameter space?
- Some relatively easy extensions
  - Try some other areas of subject matter
  - Use this to answer a real research question
  - Automate the coding of shifts
- Techniques that go beyond Vector Space models
- The puzzle: Why does this work?

The End

Bruce Sherin

bsherin@northwestern.edu