Towards a Computational History of the ACL: 1980–2008

Ashton Anderson, Dan McFarland, Dan Jurafsky
Stanford University

Intro + Motivation

Simple data-driven methodology for computational history of science

What are the natural "periods" of a field's history?

How do people move from topic to topic?

Does a field's community develop over time?

Related work and our approach

Topic models have been used for computational history

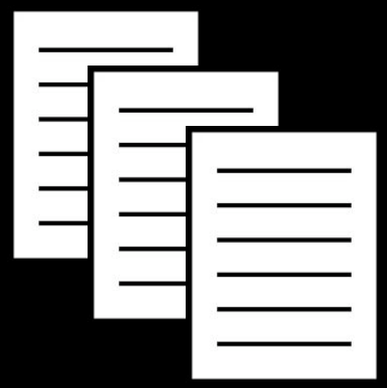
T.L. Griffiths and M. Steyvers. Finding scientific topics. PNAS 2004

David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. EMNLP 2008

C. Au Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. CIKM 2011.

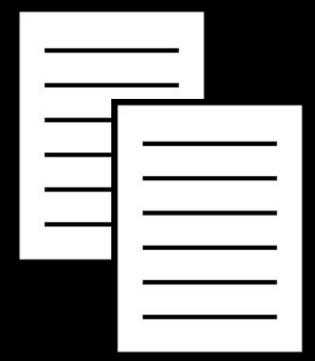
People are at the heart of our methodology

Topic X



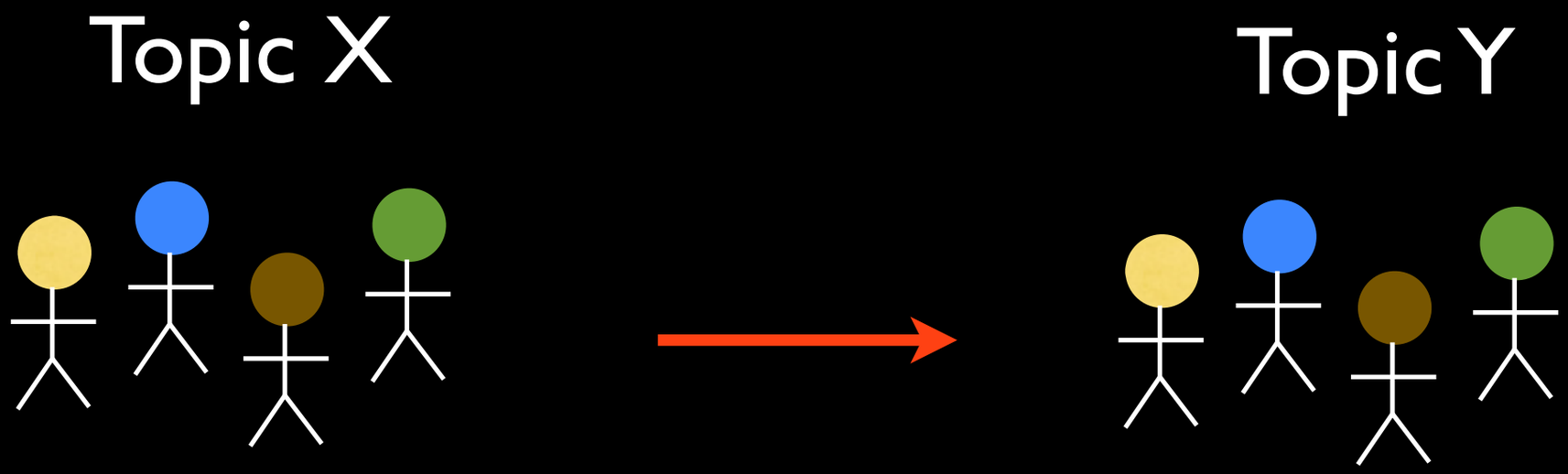
2002

Topic Y



2003

With topic models and counting alone, no hard evidence of a connection between rise and fall of topics X and Y



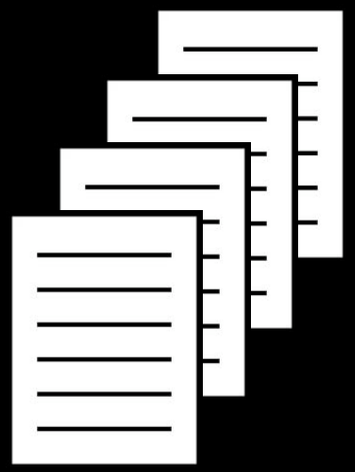
With topic models and counting alone, no hard evidence of a connection between rise and fall of topics X and Y

By tracking the movements of people over time, we can make stronger claims

Four components to our methodology:

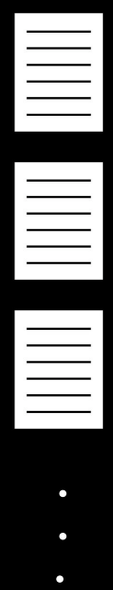
1. Identifying topics
2. Identifying epochs
3. Tracking participant flow
4. Examining author retention over time

1. Identifying topics
2. Identifying epochs
3. Tracking participant flow
4. Examining author retention over time



ACL anthology

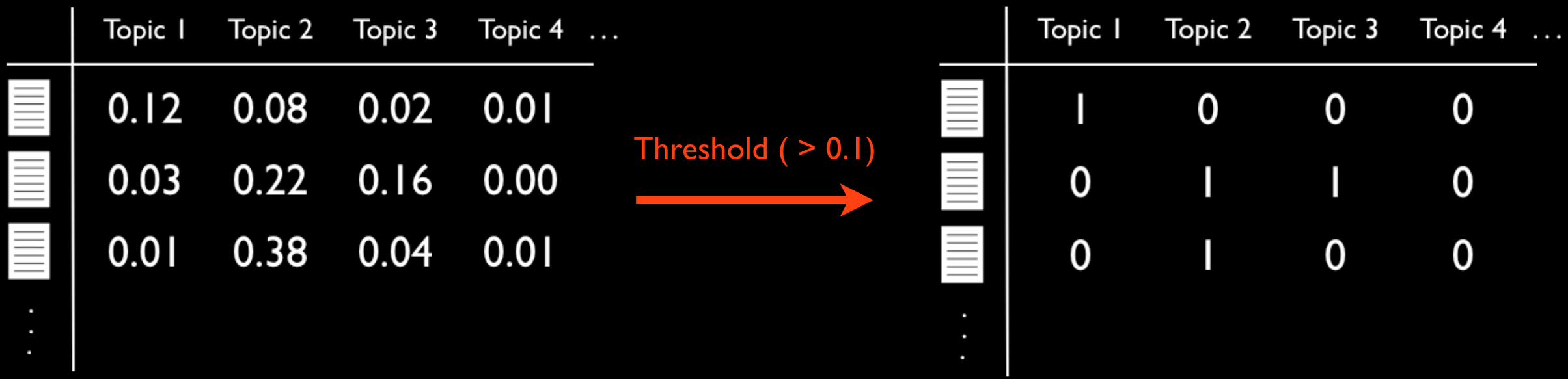
LDA



	Topic 1	Topic 2	Topic 3	Topic 4	. . .
	0.12	0.08	0.02	0.01	
	0.03	0.22	0.16	0.00	
	0.01	0.38	0.04	0.01	

LDA produces 100 topics
After expert hand-labeling and cutting non-substantive topics, we have 73 topics

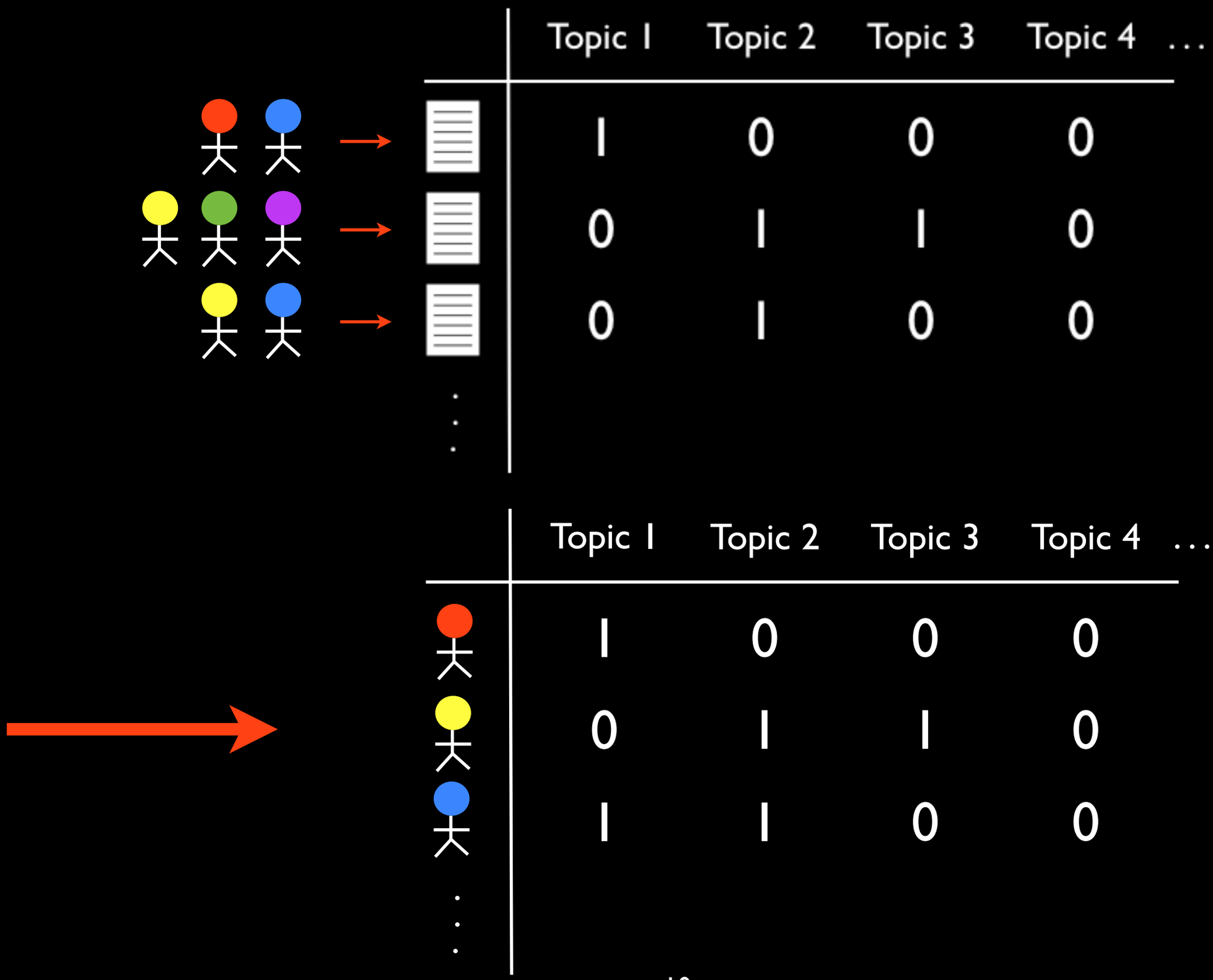
Thanks to Steven Bethard for the topic models



Convert soft to hard assignment

Now we have paper-to-topics assignment

This induces a naturally dynamic people-to-topics assignment:



Example Topics:

- Statistical Machine Translation (Phrase-Based): bleu, statistical, source, target, phrases, smt, reordering...
- Summarization: topic/s, summarization, summary/ies, document/s, news, articles, content, automatic, stories
- POS Tagging: tag/ging, POS, tags, tagger/s, part-of-speech, tagged, accuracy, Brill, corpora, tagset
- 70 more...

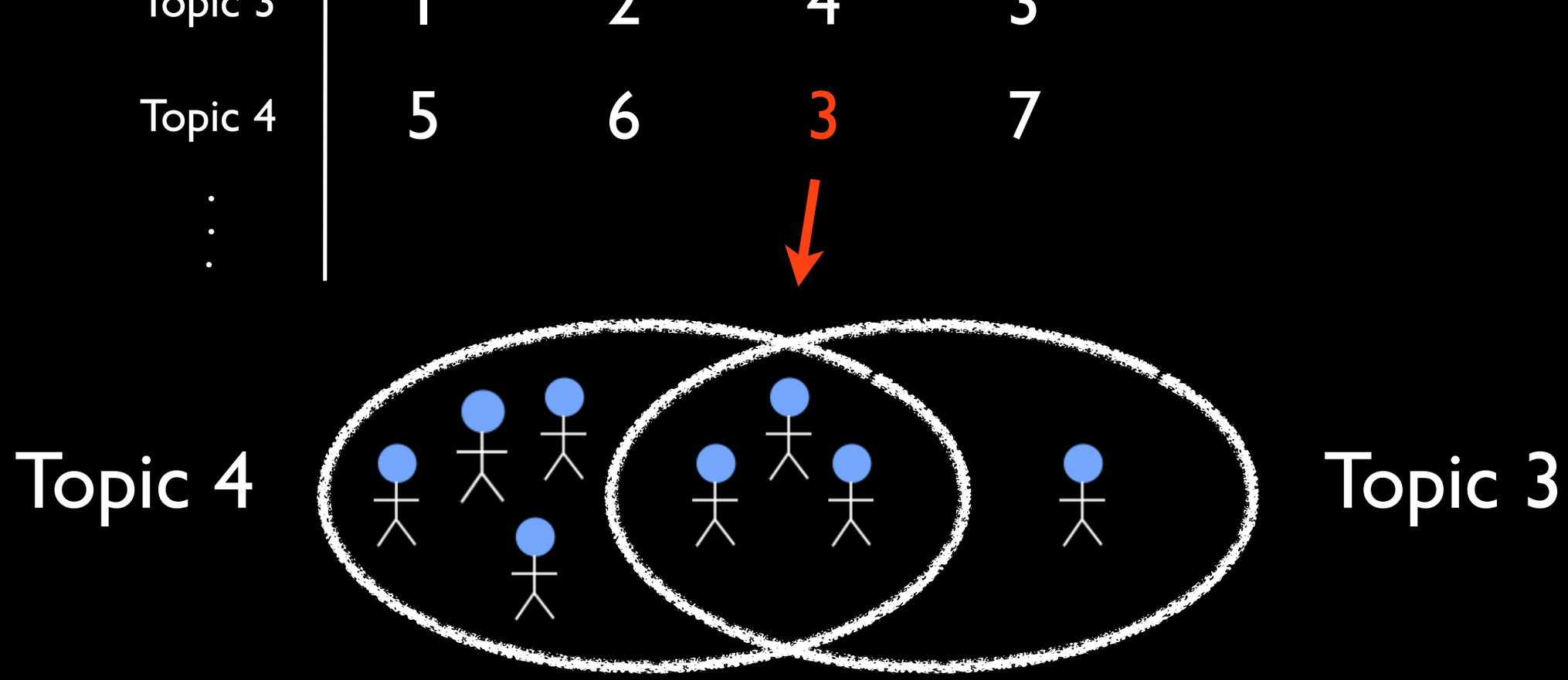
1. Identifying topics
2. Identifying epochs
3. Tracking participant flow
4. Examining author retention over time

Epoch: a sustained period of topical cohesion

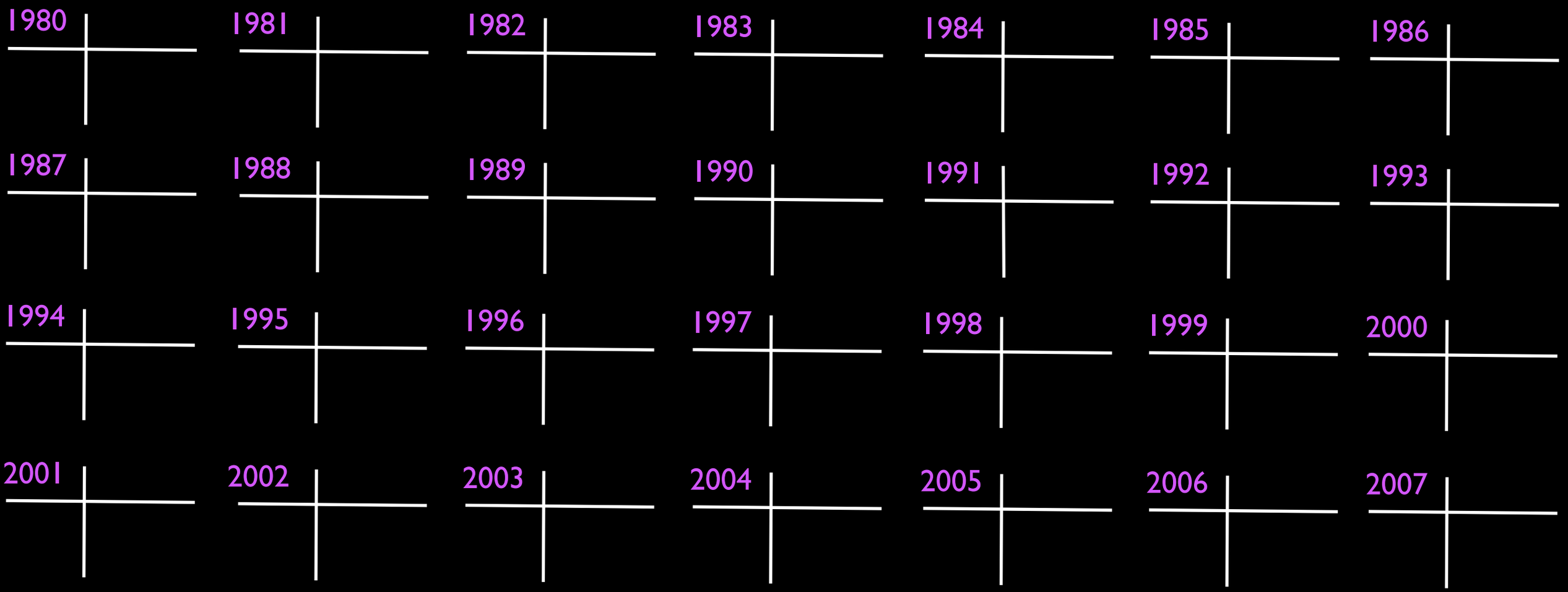
Our goal: partition the years spanned by the ACL's history into clear, distinct epochs

Our approach: first compute a topic co-authorship signature matrix to represent a particular year

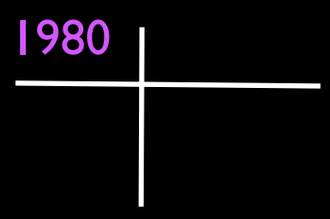
1980	Topic 1	Topic 2	Topic 3	Topic 4	…
Topic 1	7	2	1	5	
Topic 2	2	16	2	6	
Topic 3	1	2	4	3	
Topic 4	5	6	3	7	
⋮					

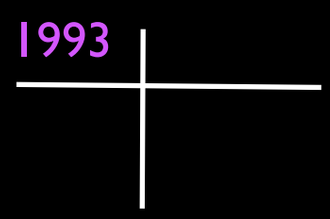


Do this for every year:

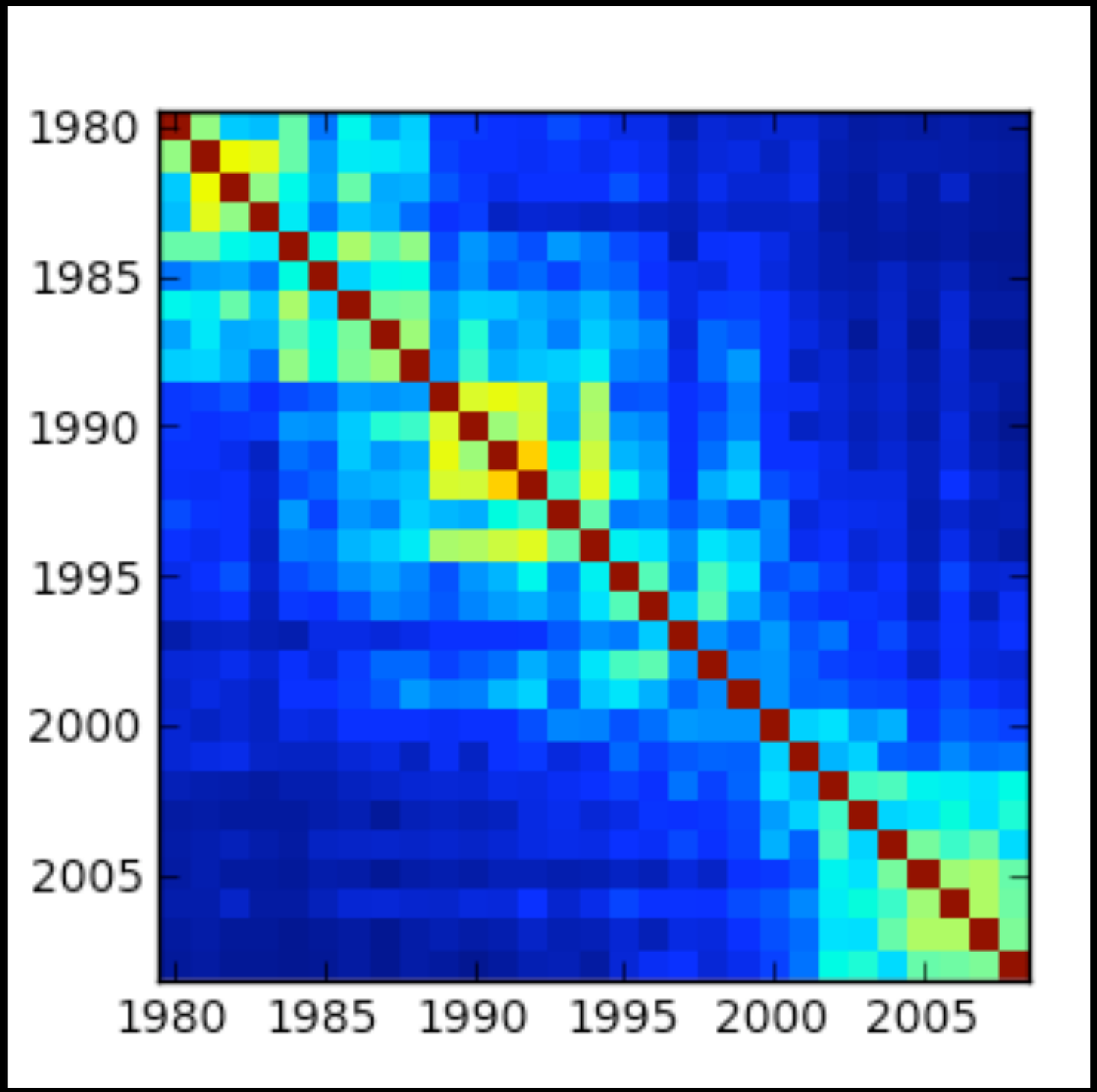


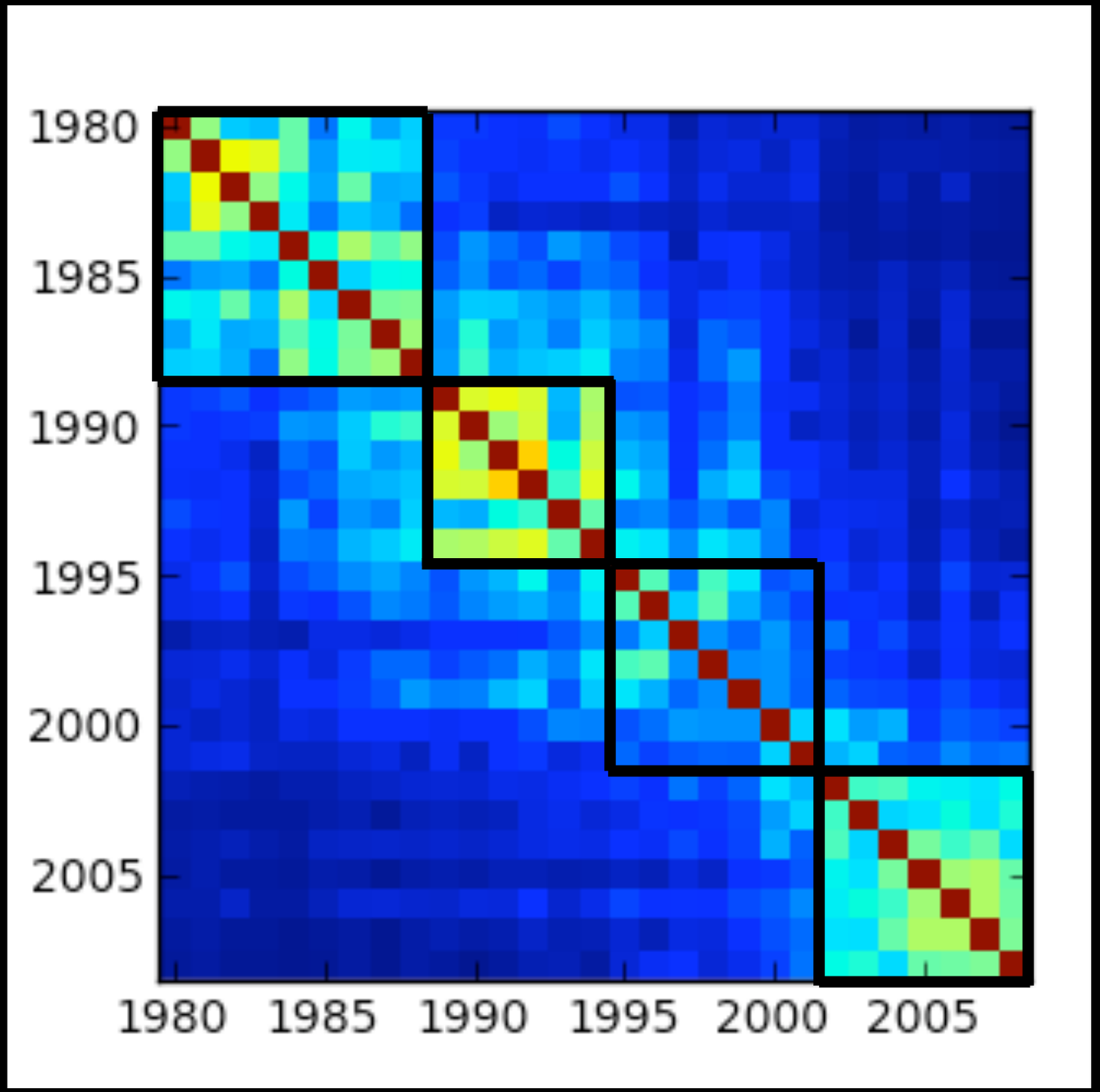
The similarity between years is then the correlation coefficient between their respective signature matrices:

Sim(1980,1993) = Corr. Coef.(

, 

)


1980
1985
1990
1995
2000
2005
1980
1985
1990
1995
2000
2005


1980
1985
1990
1995
2000
2005
1980
1985
1990
1995
2000
2005

Using this approach, we identified 4 natural epochs:

1. Early period	1980-1988
2. Bakeoff period (MUC, ATIS, DARPA)	1989-1994
3. Transitory period	1995-2001
4. Modern period	2002-2008

This method not constrained to return contiguous periods!

1. Identifying topics
2. Identifying epochs
3. Tracking participant flow
4. Examining author retention over time

How do scientific areas arise?
Which research areas developed out of others?

We answer these questions by tracing the paths of authors through topics over time, in aggregate.

First step: group topics into coherent clusters (for interpretability)

Define topic-topic similarity, then run clustering

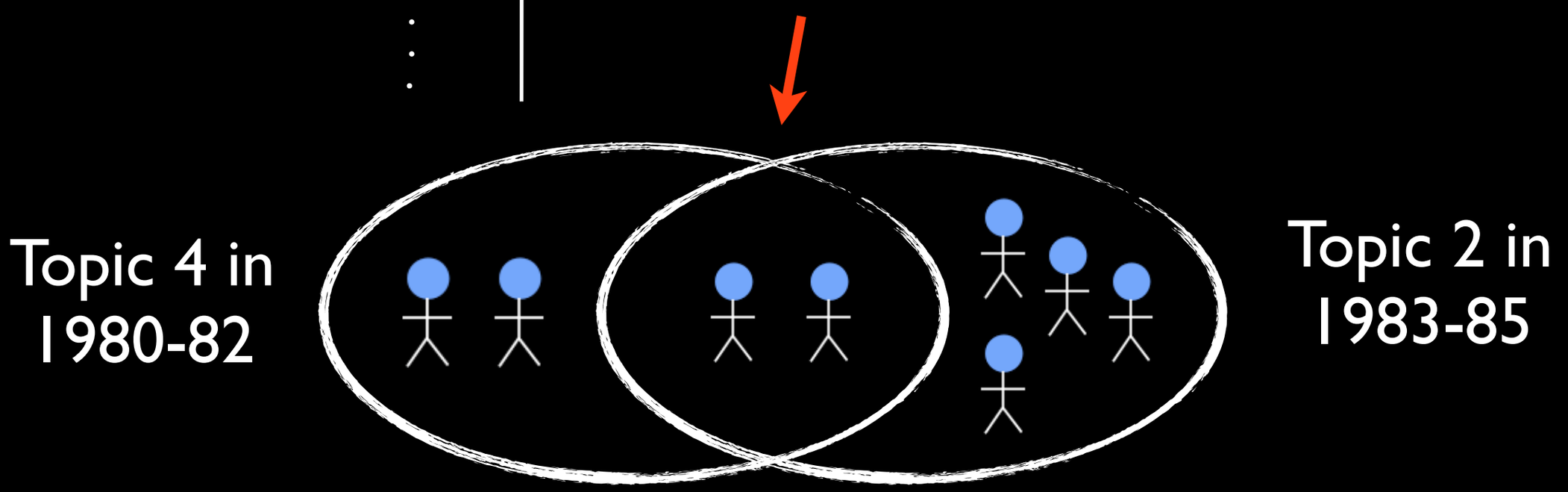
— Topics only need to be similar in how people move in and out of them

— Not necessarily similar in content

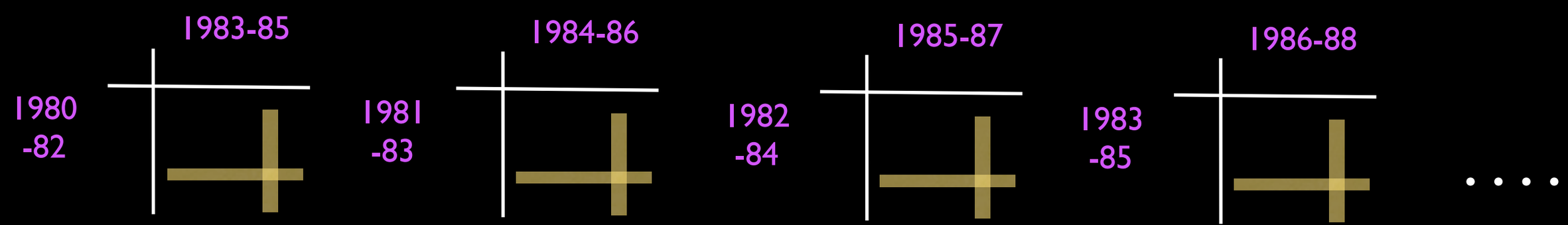
Our approach: Construct a flow profile for each topic, then topic-topic similarity is how correlated the respective topic profiles are

First compute how people moved in and out of all topics in adjacent time windows:

1980-82 \ 1983-85	Topic 1	Topic 2	Topic 3	Topic 4	…
Topic 1	15	5	1	3	
Topic 2	5	6	2	2	
Topic 3	1	2	2	3	
Topic 4	3	2	3	4	
⋮					



Then, a flow profile for topic *i* is the concatenation of the i^{th} row and i^{th} column of each matrix:



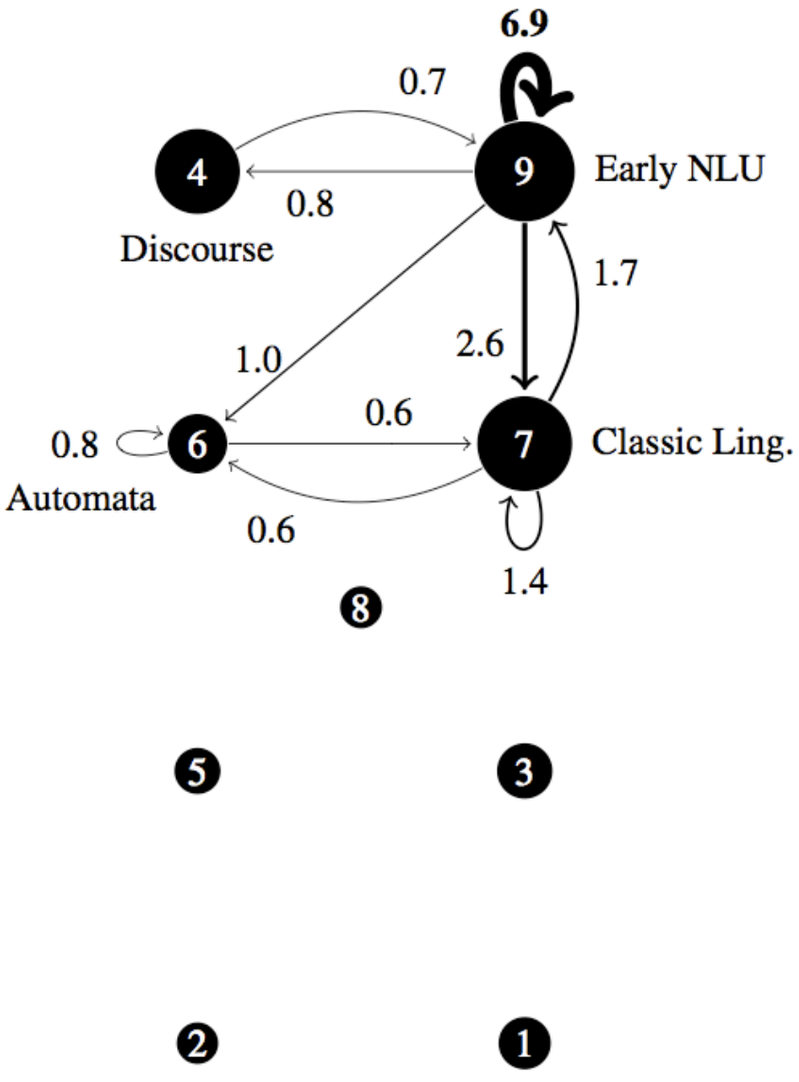
Flow profile for topic *i*

Using these flow profiles we can easily compute similarity between topics, and thus group topics into clusters

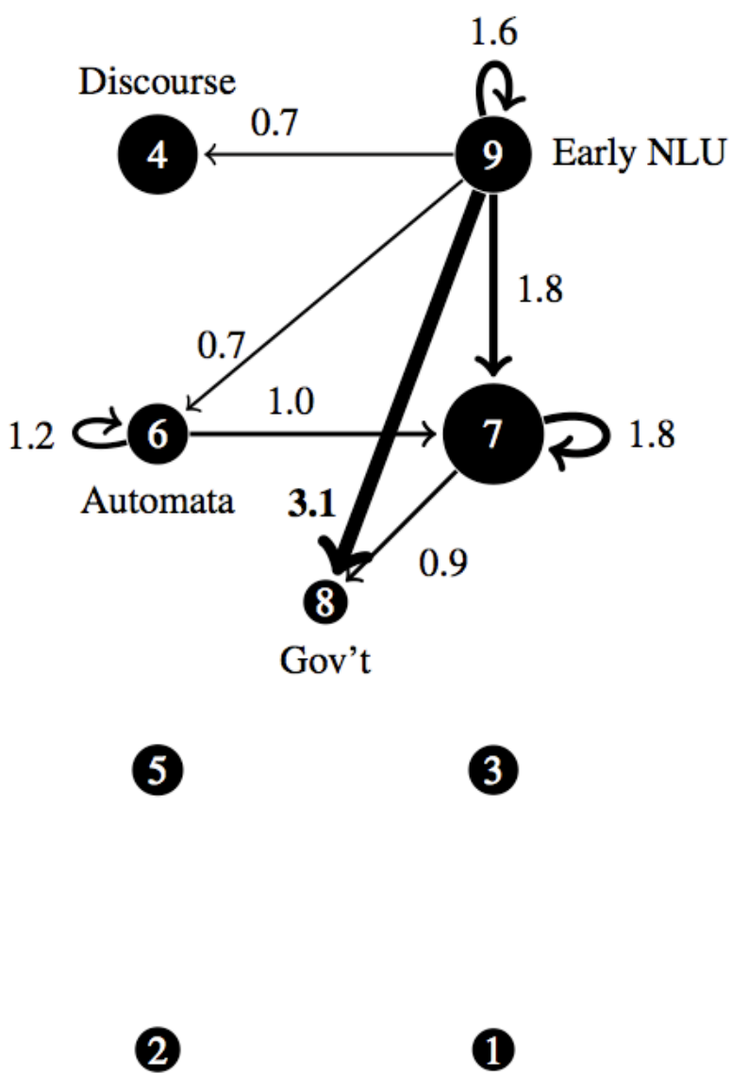
Our optimal cluster solution groups the 73 topics into 9 clusters:

1. Big Data NLP
2. Probabilistic Methods
3. Linguistic Supervised
4. Discourse
5. Early Probability
6. Automata
7. Classic Linguistics
8. Government Sponsored
9. Early NLU

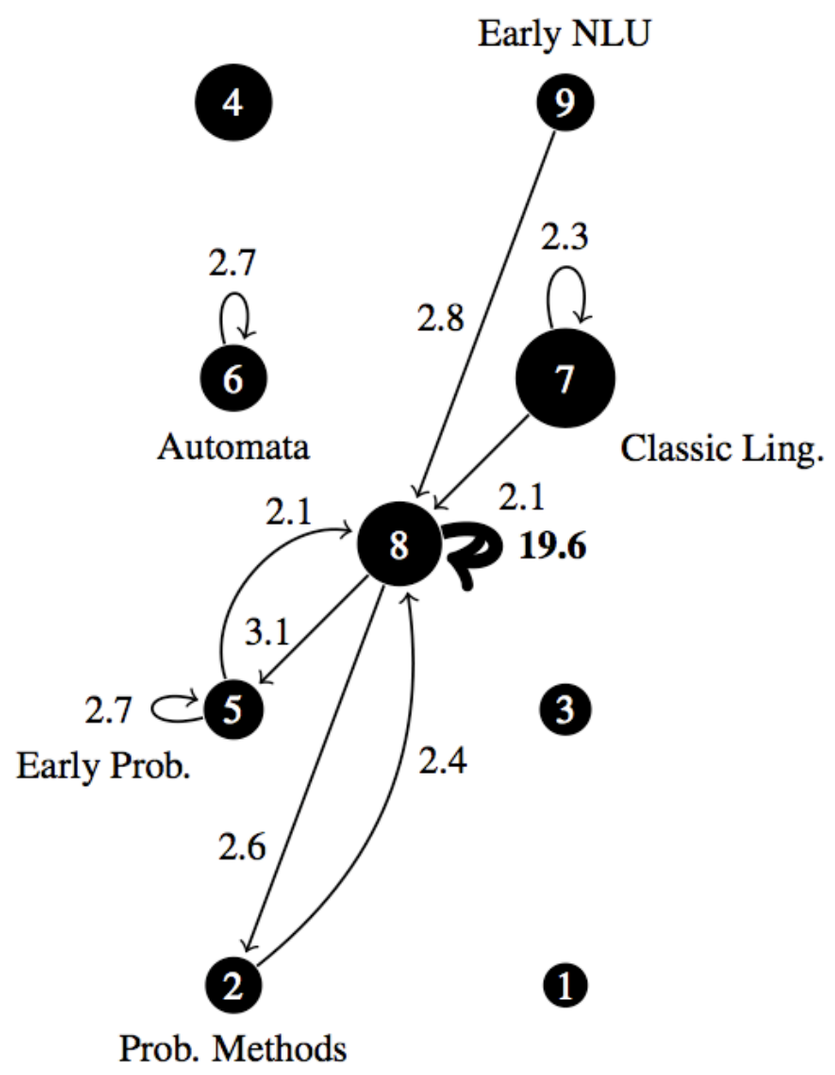
Finally, we define flow between clusters to be the average flow between topics in those clusters



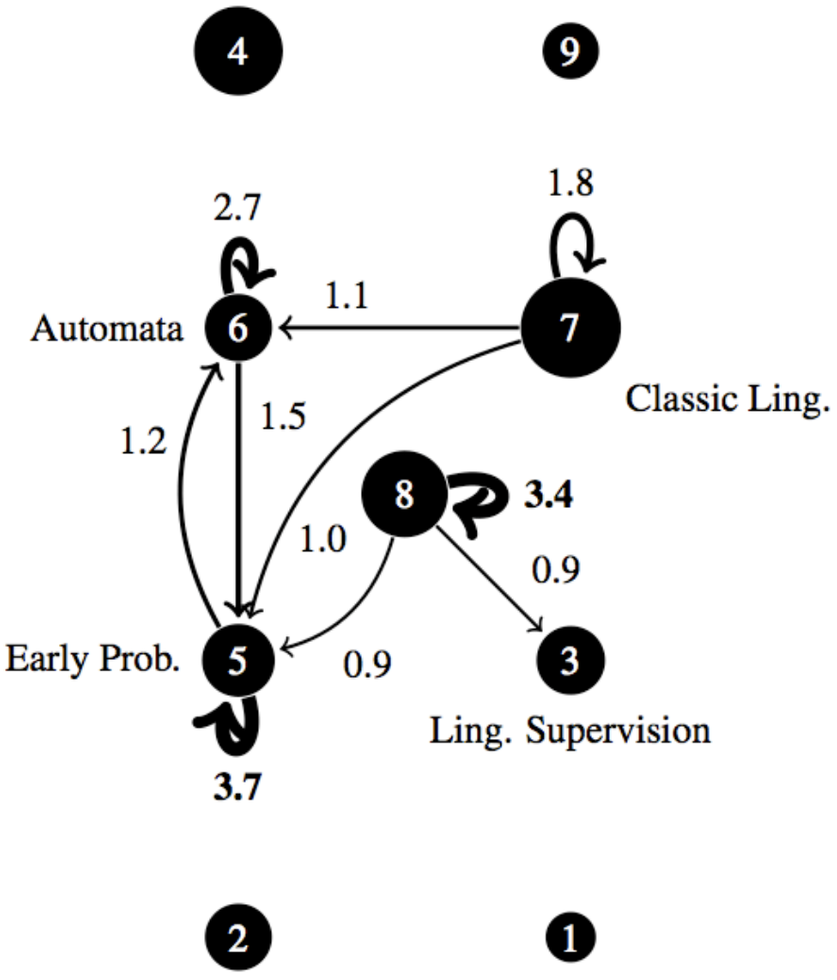
1980–83 — 1984–88



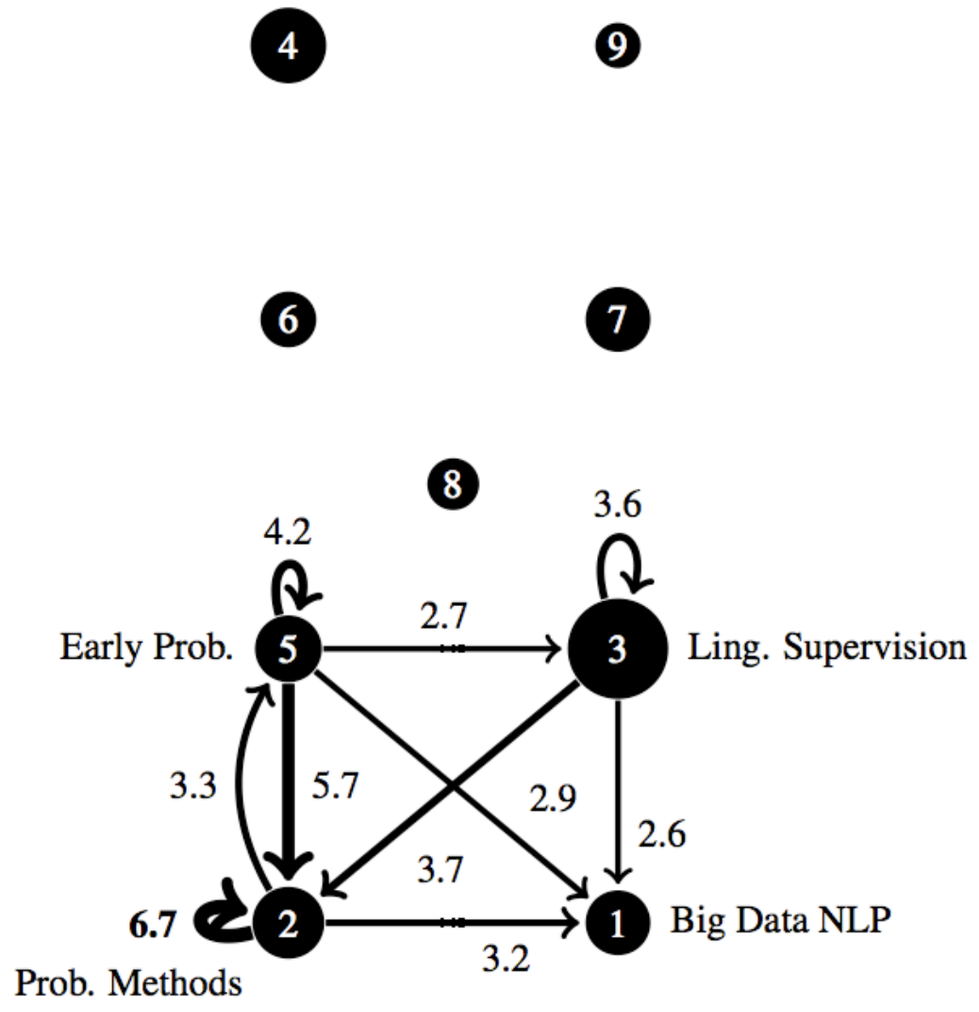
1986–88 — 1989–91



1989–91 — 1992–94



1992–94 — 1995–98



2002–04 — 2005–07

1. Identifying topics
2. Identifying epochs
3. Tracking participant flow
4. Examining author retention over time

Does a field's community develop over time?

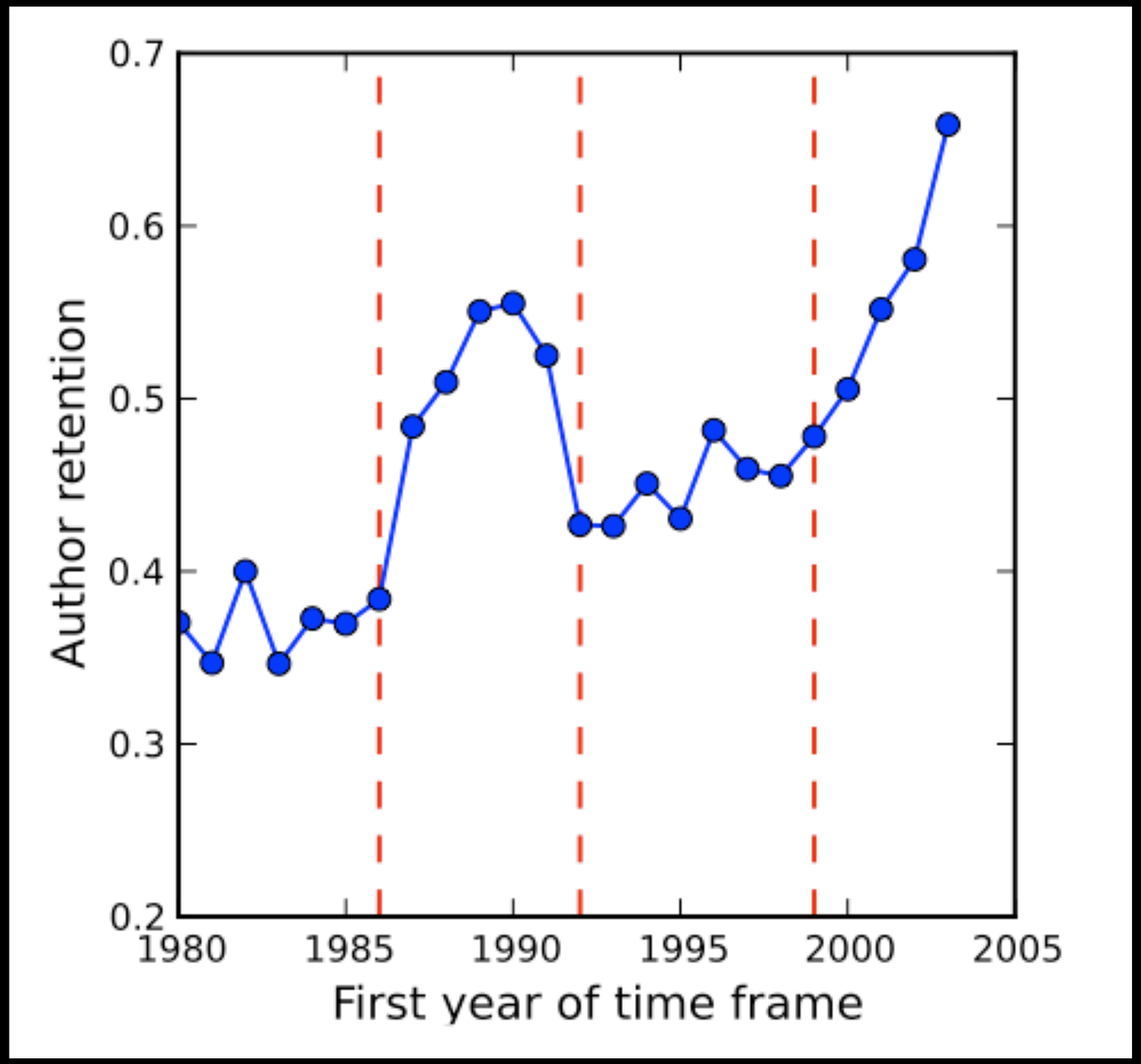
How has author retention varied over the course of the ACL's history?

Author retention: the Jaccard overlap between authors in neighboring time windows

Red dotted lines denote epoch boundaries

Field became integrated during bakeoffs period, then less so (but still higher than before)

In modern era field has become its most integrated ever



Conclusion

We developed a people-centric methodology for computational history and applied it to the ACL

— We identified 4 natural epochs in the ACL's history

— We traced the paths of authors through topics over time
  — Bakeoffs bridged early topics to modern ones

— We analyzed author retention over time
  — Bakeoffs helped integrate the field
  — In the modern era the field is the most integrated ever

Thanks!