UNIVERSITY OF AMSTERDAM

# Ad Hoc Monitoring of Vocabulary Shifts over Time

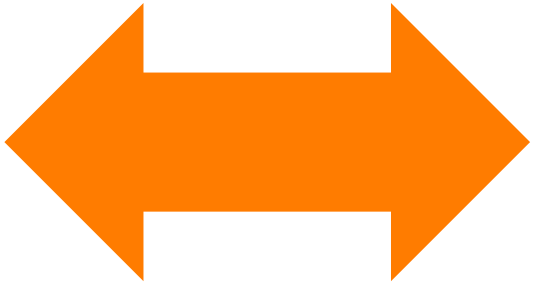
Tom Kenter, Melvin Wevers, Pim Huijnen, Maarten de Rijke

**CIKM 2015 - October 2015 - Melbourne**

# Ad hoc monitoring of vocabulary shifts over time

walkman

1986

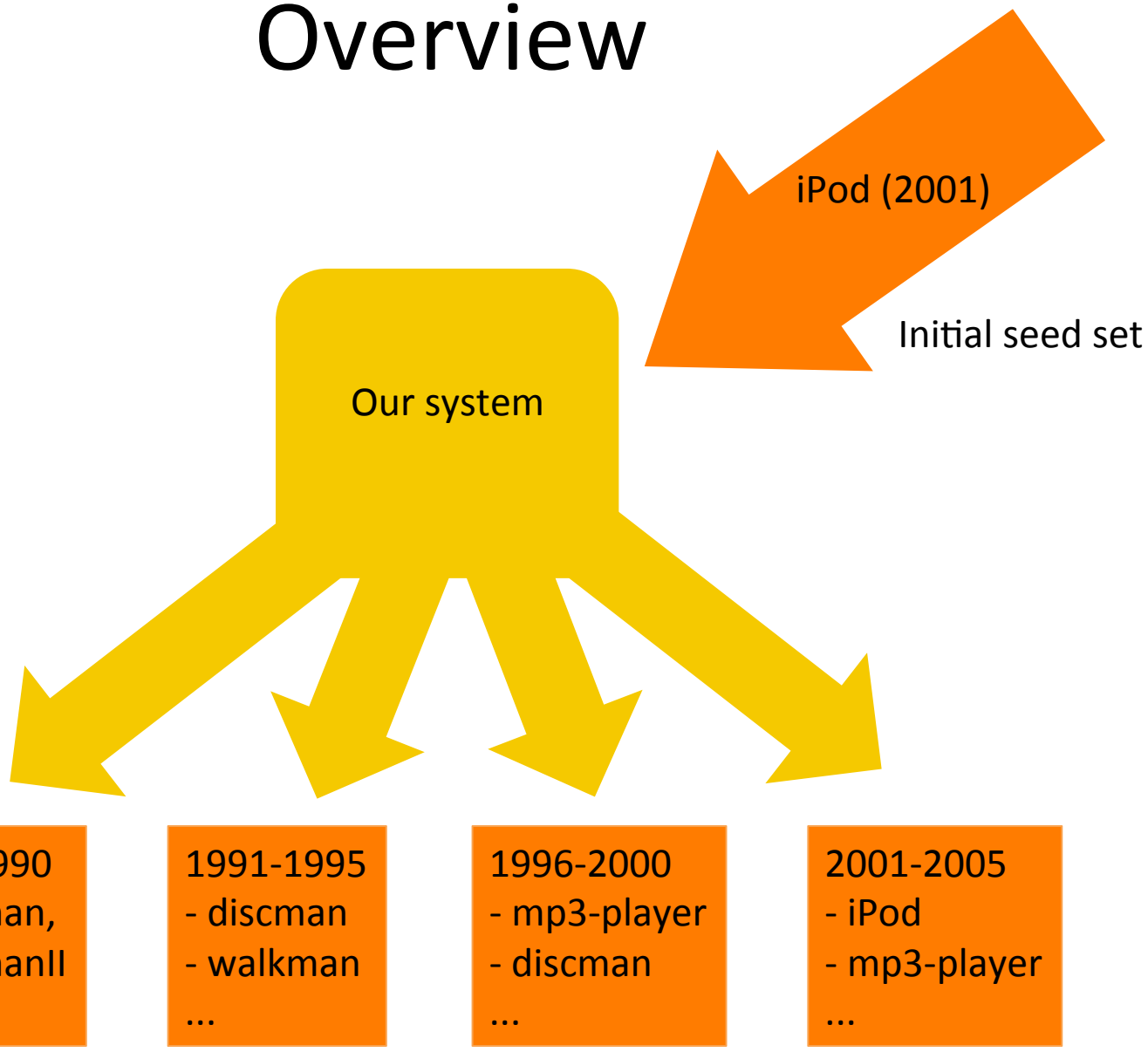


iPod

2000

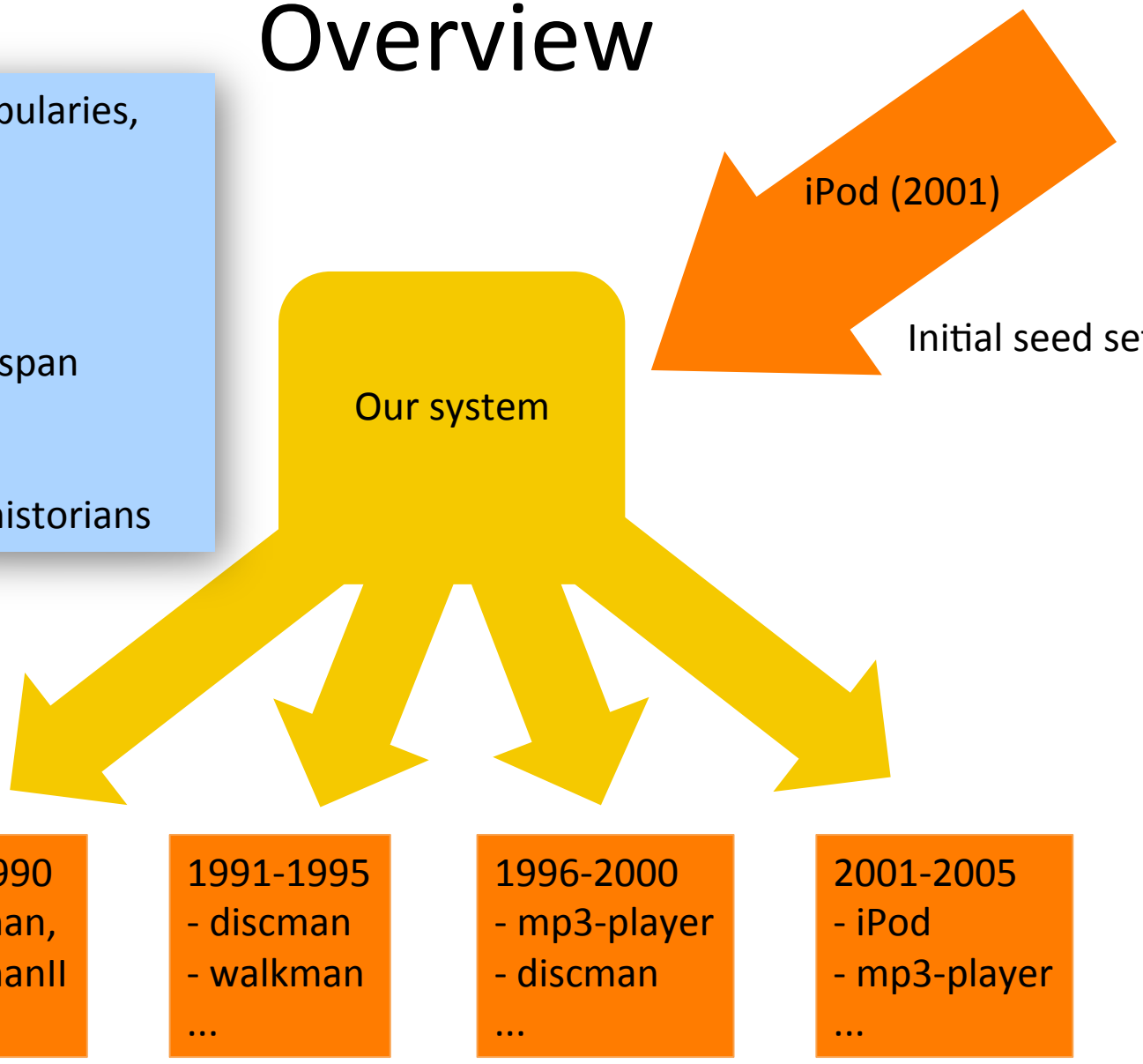
underlying concept *portable audio player*

# Overview


iPod (2001)
Initial seed set
Our system
10 word vocabularies
1986-1990
- walkman,
- walkmanII
...
1991-1995
- discman
- walkman
...
1996-2000
- mp3-player
- discman
...
2001-2005
- iPod
- mp3-player
...

# Overview

- Monitoring vocabularies, not literal words
- Ad hoc fashion
- Substantial time span (1950-1990)
- Joint work with historians

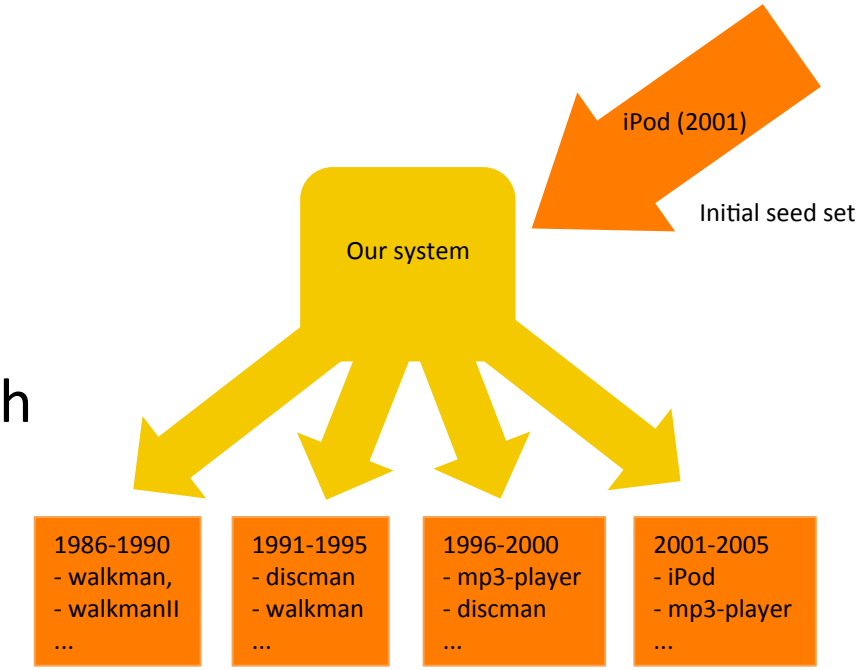


# Challenges

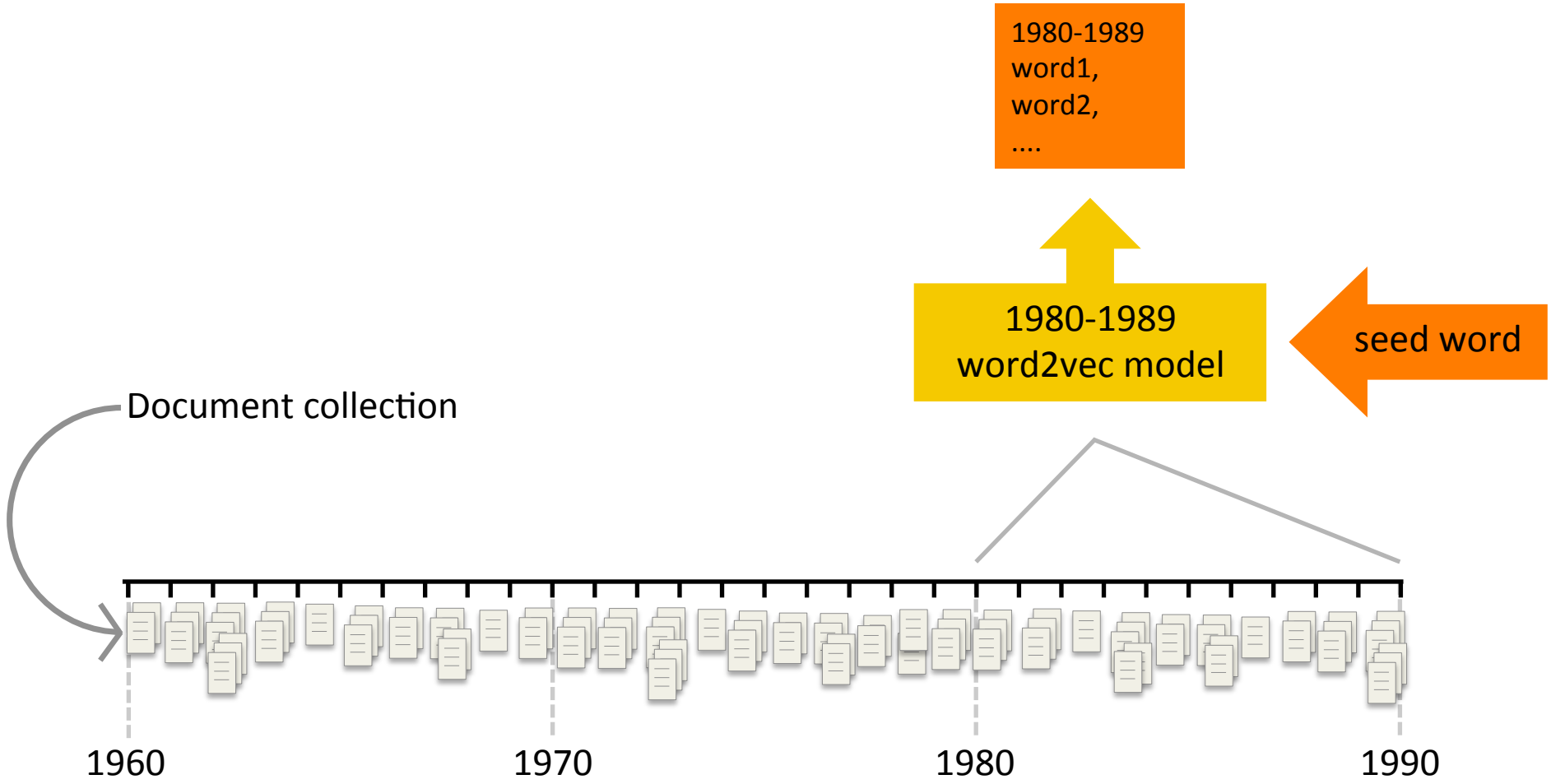
- The original seed word does not necessarily survive over time
- The underlying concept — *portable audio player* — might never be mentioned

Concept drift

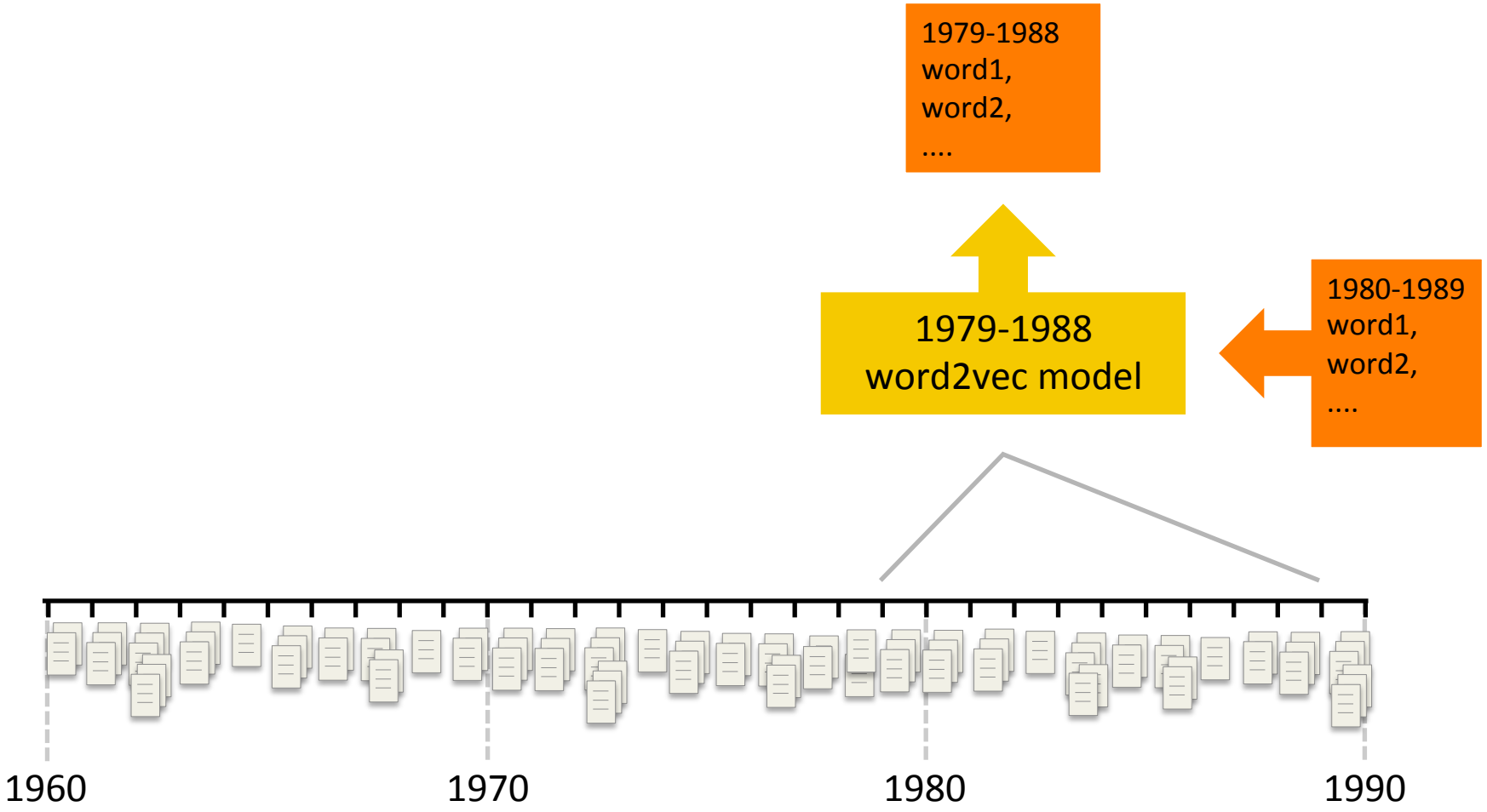
- You have to to adapt
- You should not adapt too much



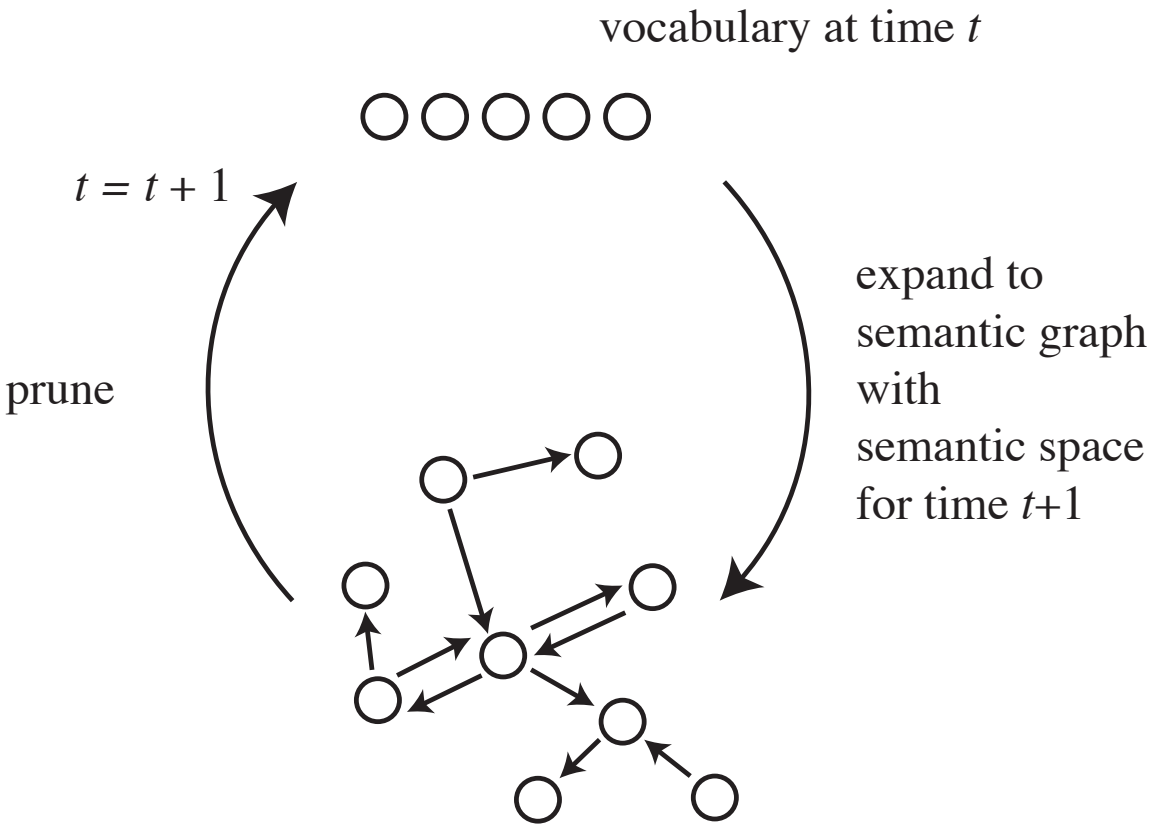
# Vocabularies from sliding time windows



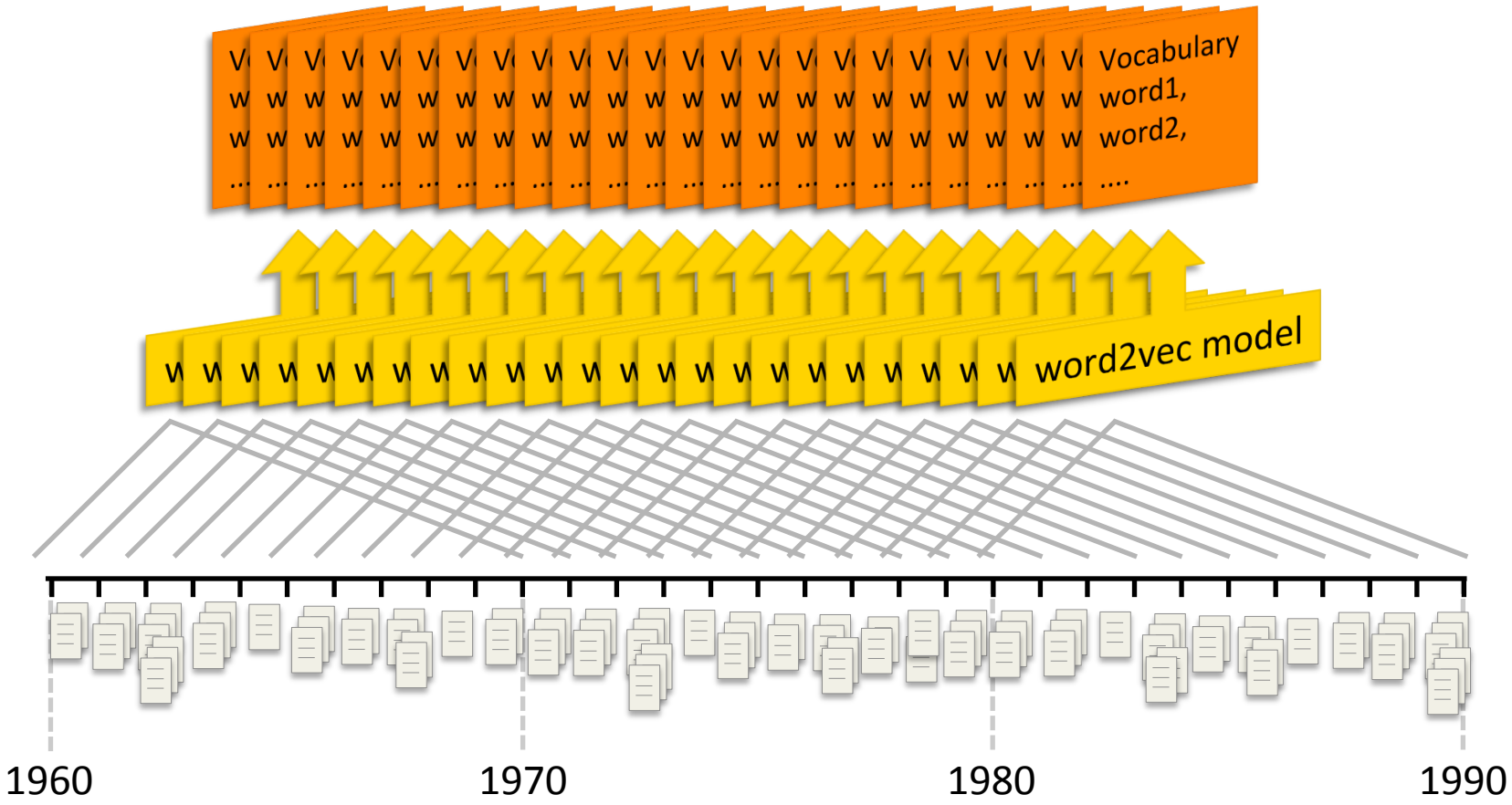
# Vocabularies from sliding time windows



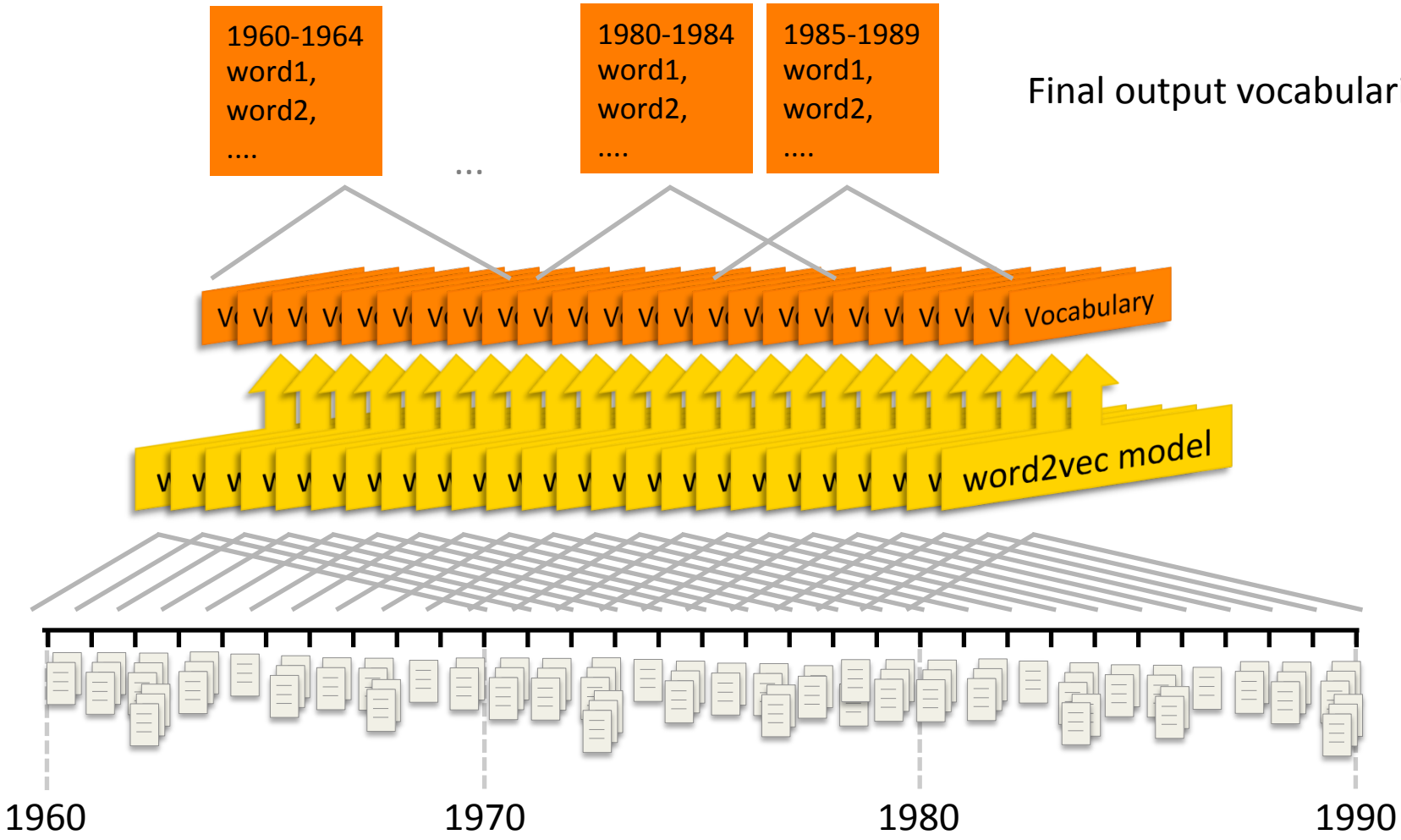
# Vocabularies from sliding time windows



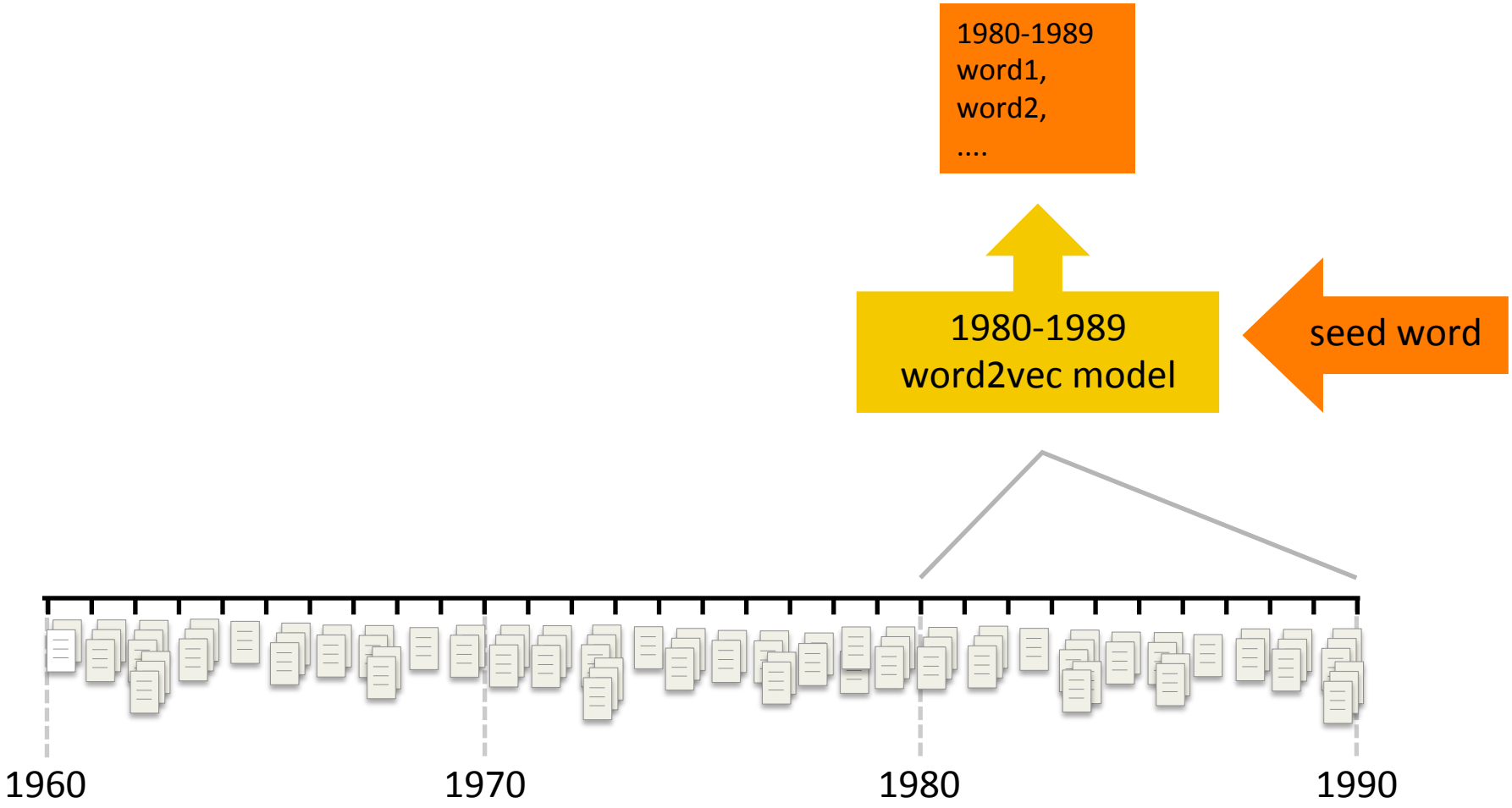
# Vocabularies from sliding time windows



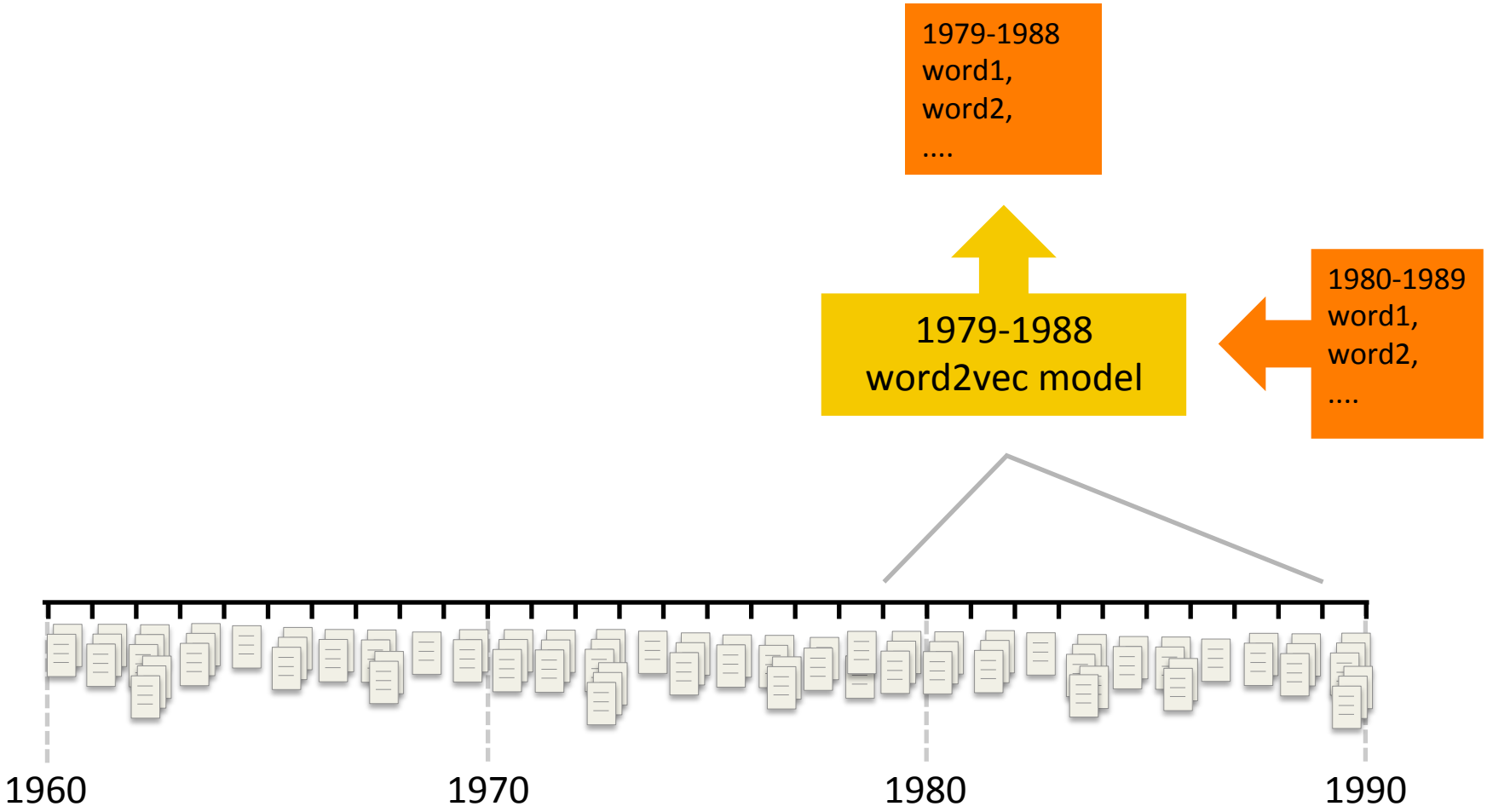
# Aggregate over vocabularies



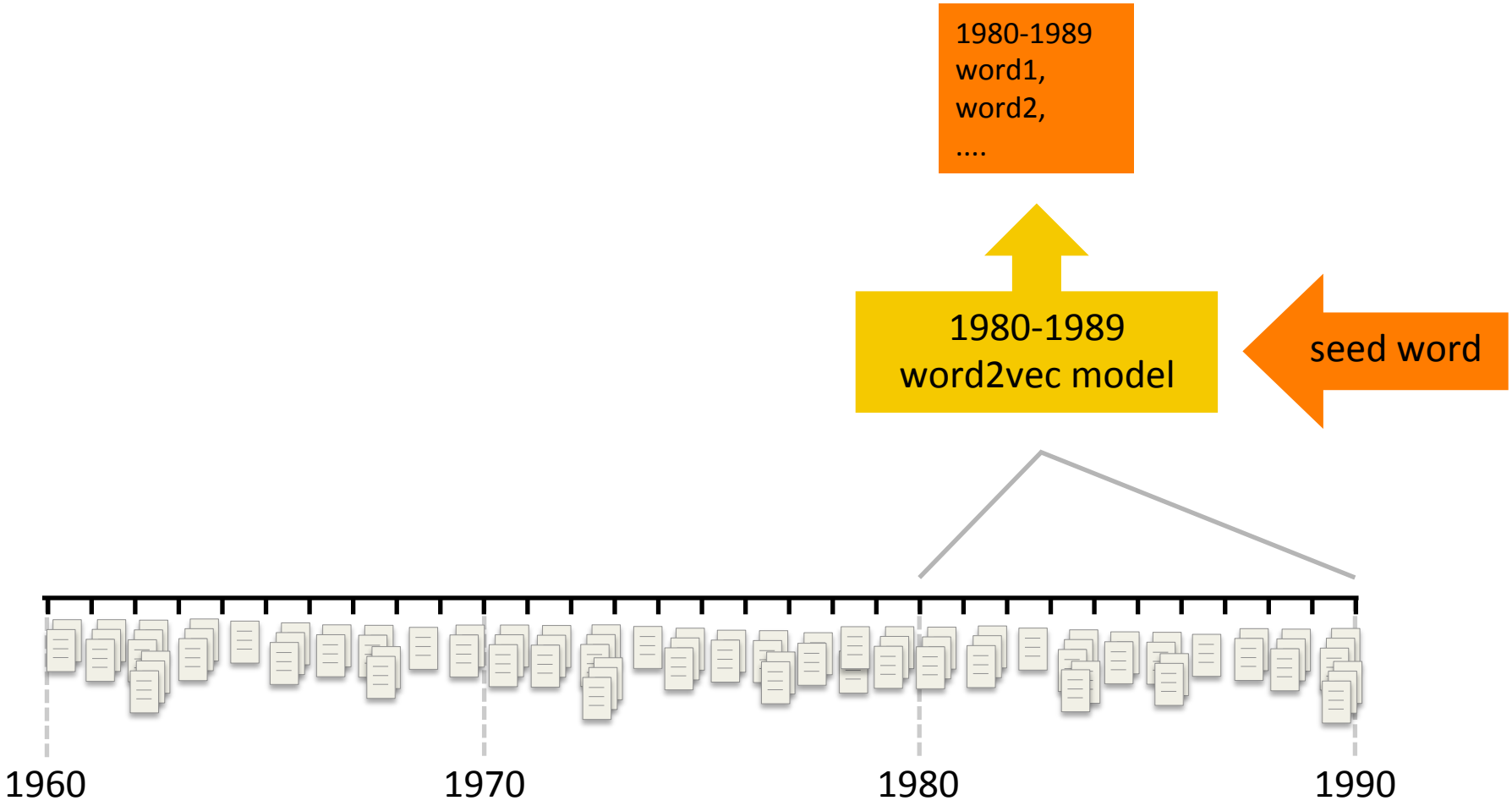
# Adaptive method



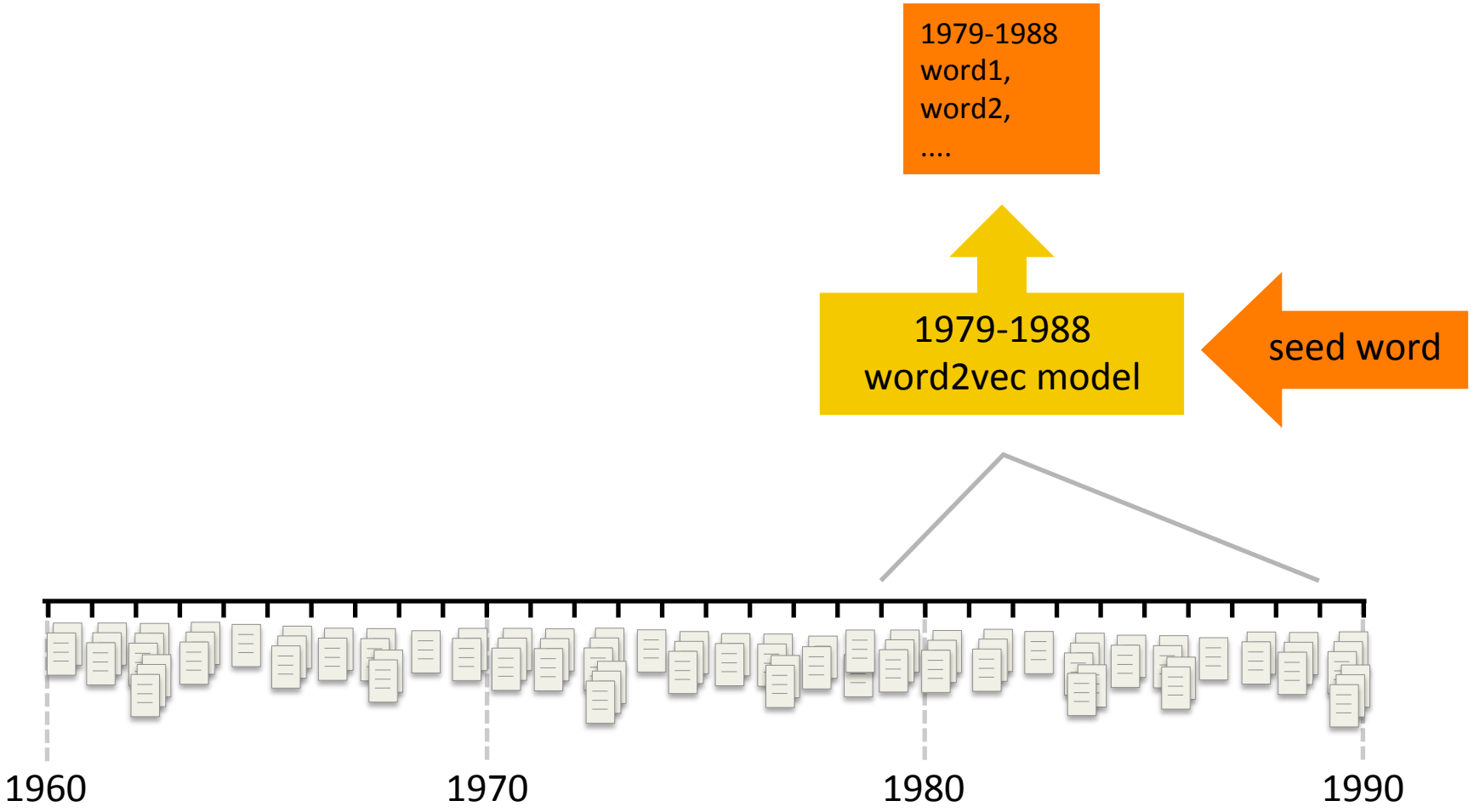
# Adaptive method



# Non-adaptive method



# Non-adaptive method



# Evaluation

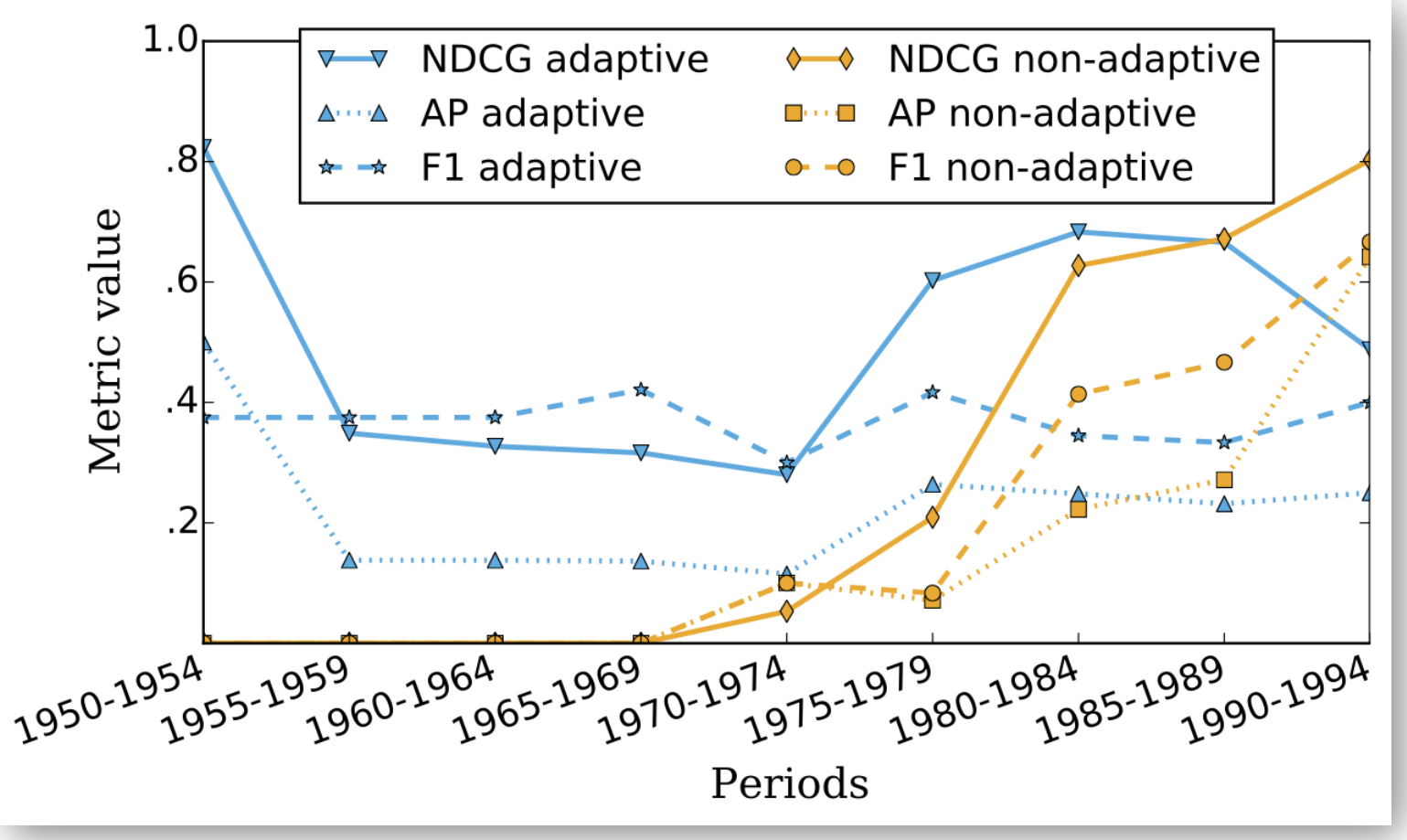
- Corpus
  4 decades digitized Dutch newspapers
- 21 seed sets
- Pooling approach
- Manually annotated data set
  Annotated by historians

- Dataset is publicly available
  http://ilps.science.uva.nl/resources/shifts/

# Evaluation

| Method | $F_1$ | p | r | NDCG | MAP |
|---|---|---|---|---|---|
| hybrid ($i = 1$) | $.384^{\ddagger}$ | $.537^{\ddagger}$ | $.406^{\ddagger}$ | $.646^{\ddagger}$ | $.343^{\ddagger}$ |
| hybrid ($i = 2$) | $.391^{\ddagger}$ | $.544^{\ddagger}$ | $\mathbf{.414}^{\ddagger}$ | $.650^{\ddagger}$ | $\mathbf{.346}^{\ddagger}$ |
| hybrid ($i = 3$) | $\mathbf{.392}^{\ddagger}$ | $\mathbf{.548}^{\ddagger}$ | $.411^{\ddagger}$ | $\mathbf{.653}^{\ddagger}$ | $.345^{\ddagger}$ |
| hybrid ($i = 4$) | $.389^{\ddagger}$ | $.545^{\ddagger}$ | $.405^{\ddagger}$ | $.651^{\ddagger}$ | $.343^{\ddagger}$ |
| hybrid ($i = 5$) | $.385^{\ddagger}$ | $.541^{\ddagger}$ | $.399^{\ddagger}$ | $.649^{\ddagger}$ | $.339^{\ddagger}$ |
| adaptive | .344 | $.551^{\ddagger}$ | .298 | $.514^{\dagger}$ | $.237^{\dagger}$ |
| non-adaptive | $.367^{\ddagger}$ | $.521^{\ddagger}$ | $.389^{\ddagger}$ | $.630^{\ddagger}$ | $.332^{\ddagger}$ |
| baseline | .303 | .450 | .296 | .554 | .266 |

# Example: *holocaust*



# Conclusion

- This task is new
- This task is hard

- A system has to adapt, but concept drift is hard to avoid

- There is room for improvement

# Thank you!

www.tomkenter.nl
tom.kenter@uva.nl
@TomKenter