Temporal Text Ranking and Automatic Dating of Texts

EACL 2014, Göteborg

Vlad Niculae (Max Planck Institute for Software Systems)
Marcos Zampieri (Saarland University)
Liviu P. Dinu (University of Bucharest)
Alina Maria Ciobanu (University of Bucharest)

1. Text Dating

Estimate the writing date of a text.

(Linguistic complement to *material dating*.)

1. Text Dating

Estimate the writing date of a text.

(Linguistic complement to *material dating*.)

- 1930? 1899? 1823? (Regression)
  (Preoţiuc-Pietro and Cohn, 2013)
- 18th / 19th century? (Classification)
  (de Jong et al, 2005)
  and our previous work

1. Text Dating

Estimate the writing date of a text.

(Linguistic complement to *material dating*.)

- Which is newer?

A Relation

Of ſome Trials of the ſame Operation, lately made in France.

1. *M. Denys*, Profeſſor of the *Mathematicks* and *Natural Philoſophy* at *Paris*, in a Letter of his to the *Publiſher* relateth, That they had lately tranſmitted the Blood of four *Weathers* into a *Horſe* of 26 years old, and that this Horſe had thence received much ſtrength, and more than an ordinary ſtomach.

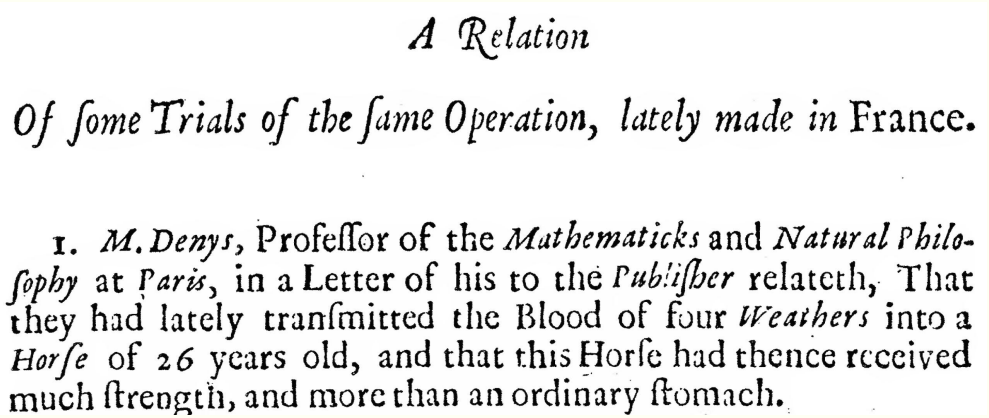
1. Text Dating

Estimate the writing date of a text.

(Linguistic complement to *material dating*.)

1899. W. Crane, A Floral Fantasy in an Old English Garden

- Which is newer?



A Relation

Of ſome Trials of the ſame Operation, lately made in France.

1. *M. Denys*, Profeſſor of the *Mathematicks* and *Natural Philoſophy* at *Paris*, in a Letter of his to the *Publiſher* relateth, That they had lately tranſmitted the Blood of four *Weathers* into a *Horſe* of 26 years old, and that this Horſe had thence received much ſtrength, and more than an ordinary ſtomach.

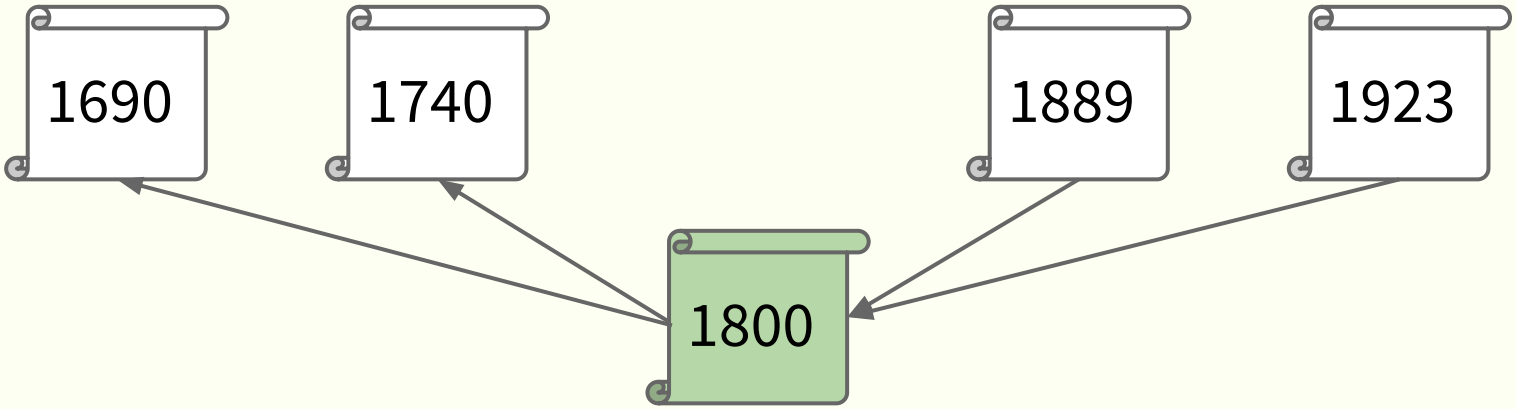
1667. An Account Of The Experiment Of Transfusion Practiced Upon A Man In London

2. This Work: Pairwise Ranking

Input: **pairs** of documents

Output: “$<$”, “$>$”

Not all input samples need to be comparable.

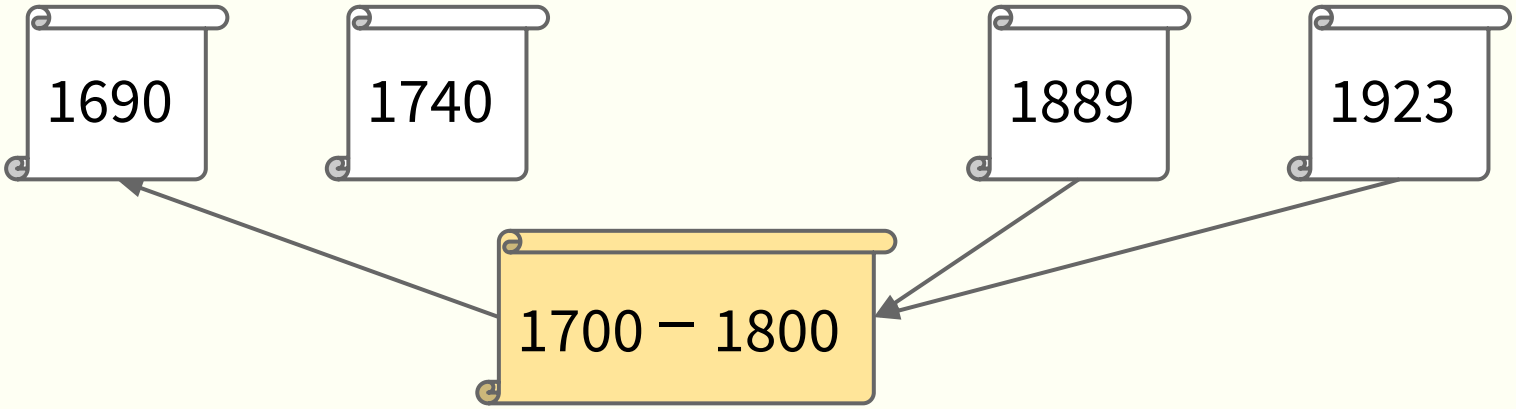


2. This Work: Pairwise Ranking

Input: **pairs** of documents

Output: “$<$”, “$>$”

Not all input samples need to be comparable.



3. Behind the Scenes

Binary classification of pairs.

$$g(d_1, d_2) > 0$$

But we want the dates, not a ranking!

3. Behind the Scenes

Binary classification of pairs.

$$g(d_1, d_2) > 0$$

But we want the dates, not a ranking!

$$w \cdot (d_1 - d_2) > 0$$

$$w \cdot d_1 > w \cdot d_2$$

3. Behind the Scenes

Binary classification of pairs.

$$g(d_1, d_2) > 0$$

But we want the dates, not a ranking!

$$w \cdot (d_1 - d_2) > 0$$

$$w \cdot d_1 > w \cdot d_2$$

Use a moment in time instead of a document:

$$w \cdot d_1 > \theta(1850)$$

Evaluation

4. Historical Corpora

Three languages:

- Colonia Corpus of Historical **Portuguese**
  (Zampieri and Becker, 2013)

- Corpus of Late Modern **English** Texts (CLMET)
  (de Smet, 2005)

- **Romanian** Historical Corpus
  (Ciobanu et al. 2013)

5. Simple Features

A. lexical (word counts)

B. naive morphological

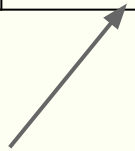
(character n-grams at the end of words)

+ feature transformation and selection

6. Results

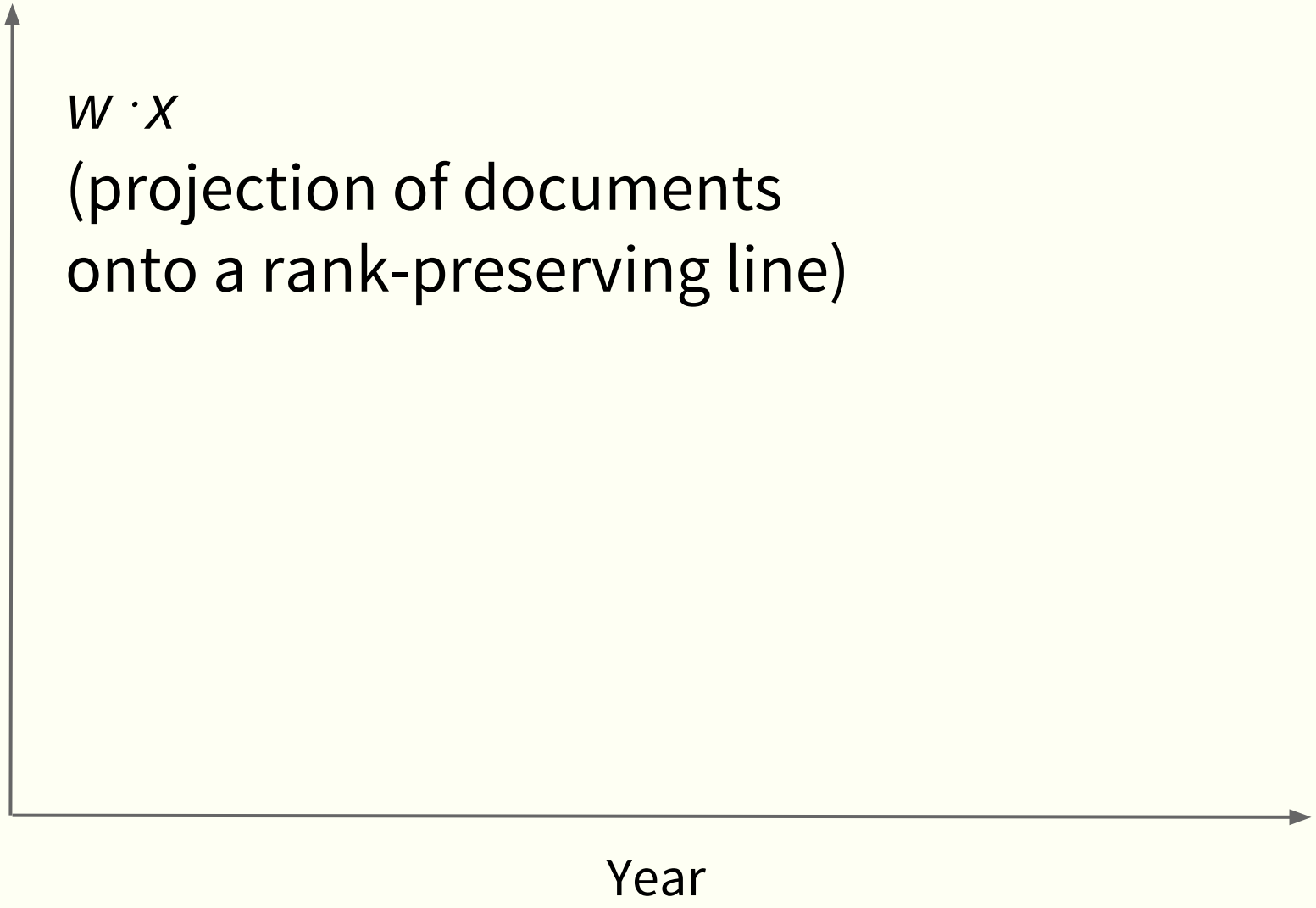
Comparable to the regression approach

	size	pairwise score	Ridge pairwise score
en	293	83.8%	83.7%
pt	87	82.9%	81.9%
ro	42	92.9%	92.4%

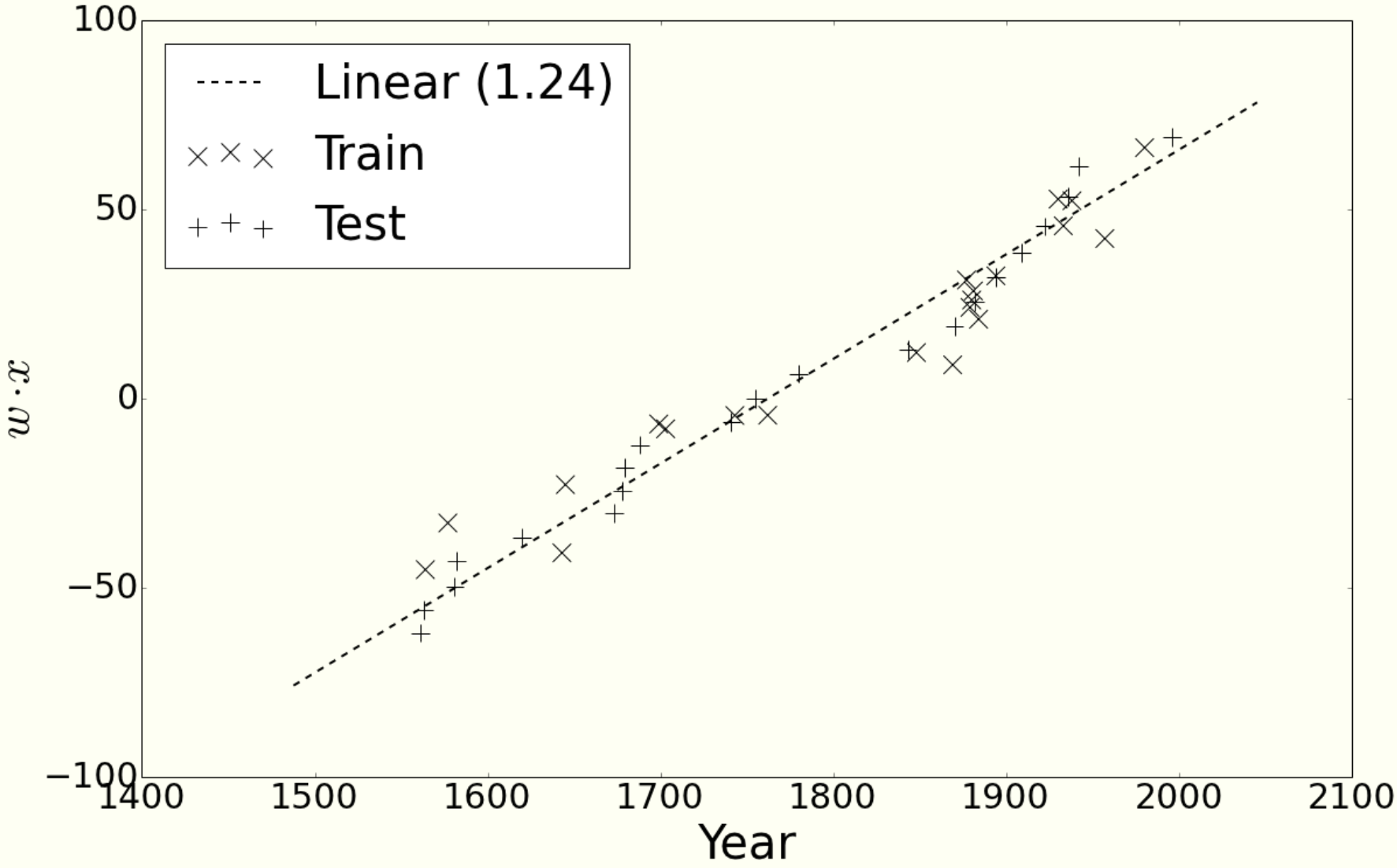


our system

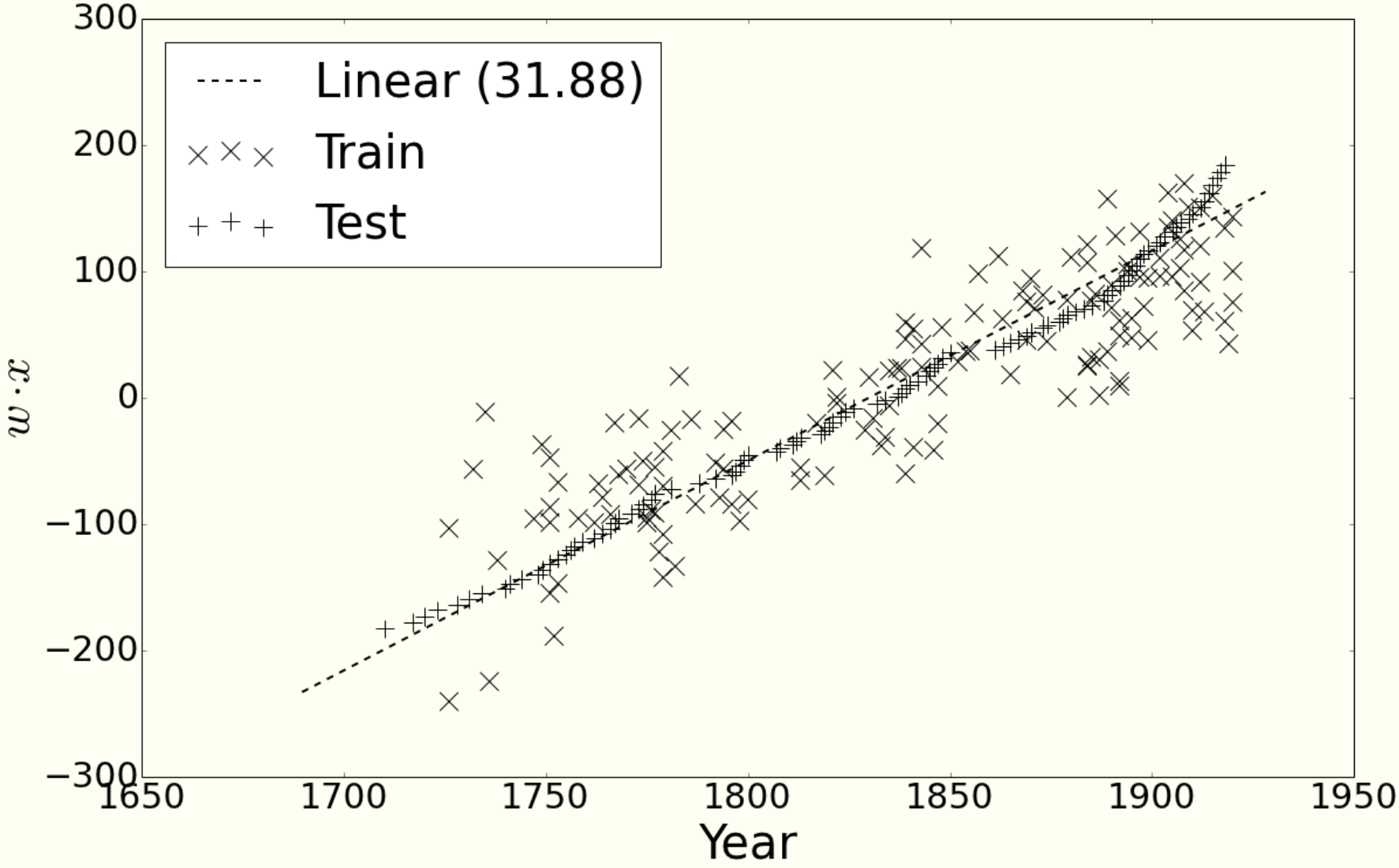
7. Function estimation (θ)



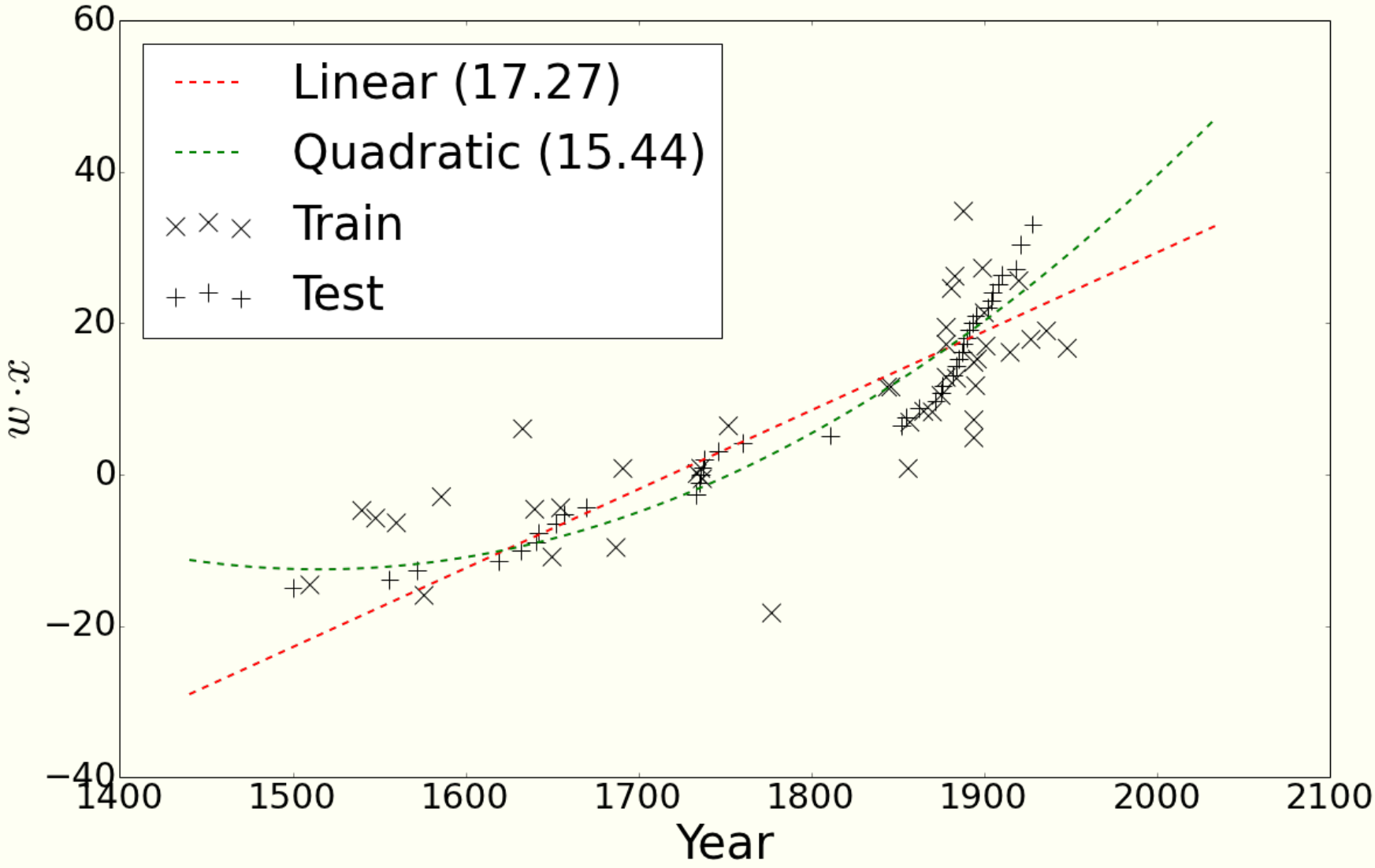
8. Function estimation (Romanian)



9. Function estimation (English)



10. Function estimation (Portuguese)



11. Dating uncertain texts

C. Cantacuzino (1650 – 1716), *Istoria Țării Rumânești*

Important historical work, contested writing time.

Published: 19th century.

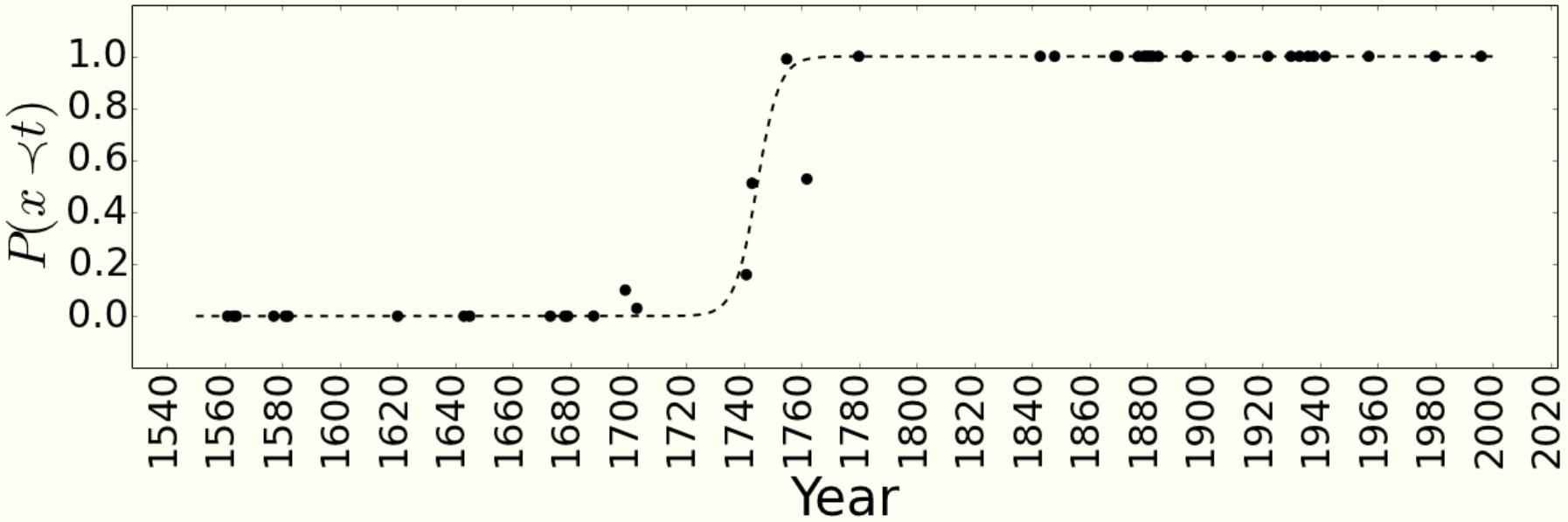
11. Dating uncertain texts

C. Cantacuzino (1650 – 1716), *Istoria Țării Rumânești*

Important historical work, contested writing time.

Published: 19th century.

We predict 1736.2 – 1753.2:



12. Conclusion & Future Work

- ranking approach to temporal modelling
- important gain on flexibility
- acceptable performance with simple features

12. Conclusion & Future Work

- ranking approach to temporal modelling
- important gain on flexibility
- acceptable performance with simple features

- application-specific feature engineering
- other historical corpora wanted!