Contextual Text Mining: Mining Causal Topics with Time Series Supervision

ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Contextual Text Mining:
Mining Causal Topics with Time Series Supervision

5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

Perceive

(Perspective)

Express

(English)

1. Natural language processing and text representation

2. Word association mining and analysis

4. Opinion mining and sentiment analysis

Text Mining for Understanding Time Series

What might have caused the stock market crash?

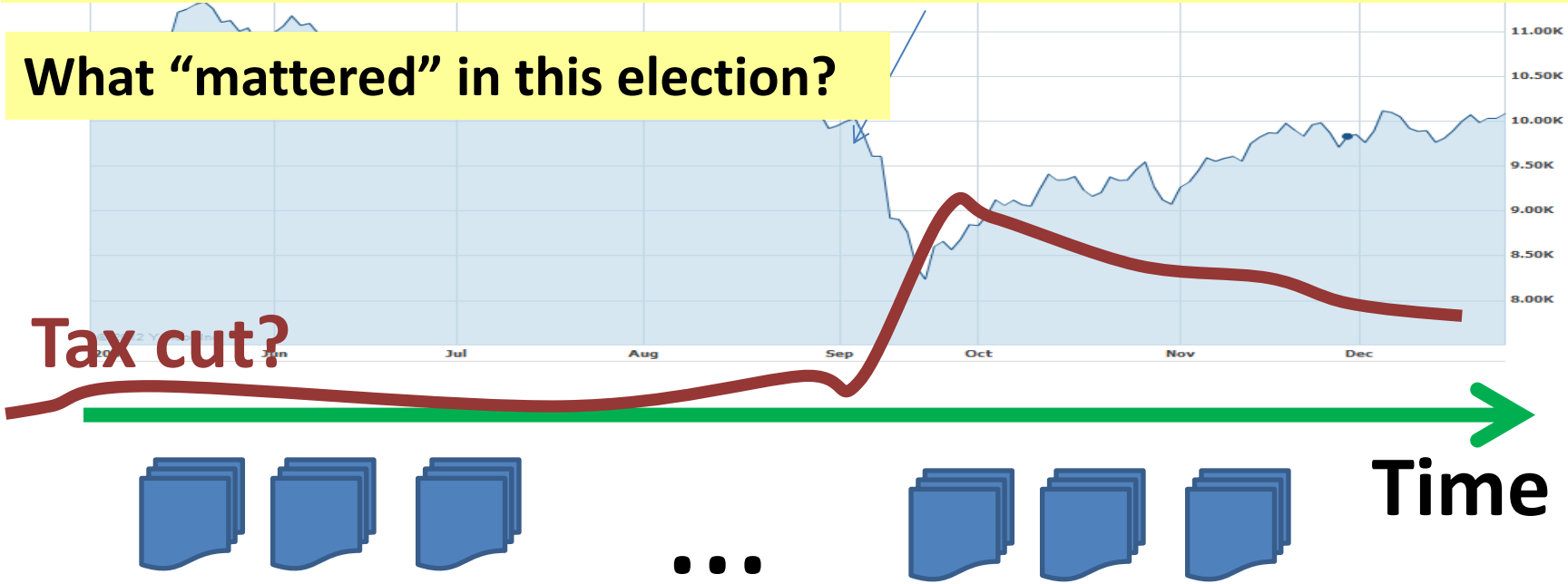
Any clues in the companion news stream?

Dow Jones Industrial Average [Source: Yahoo Finance]

Analysis of Presidential Prediction Markets

What might have caused the sudden drop of price for this candidate?

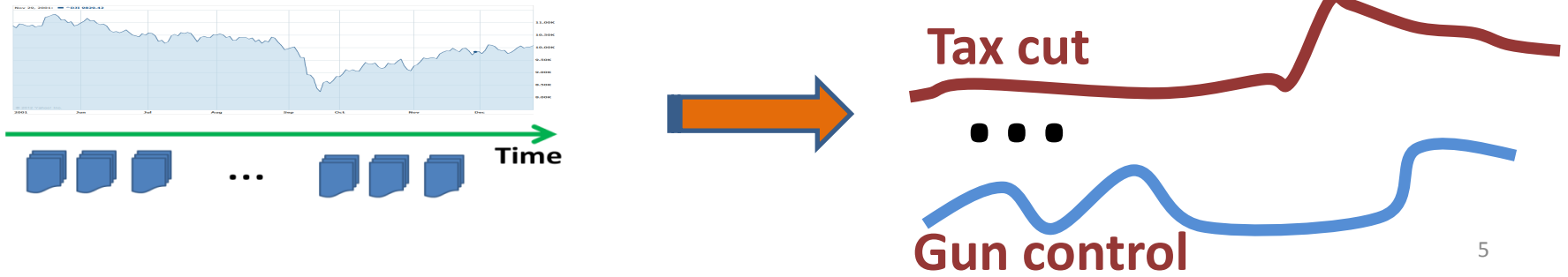
What "mattered" in this election?



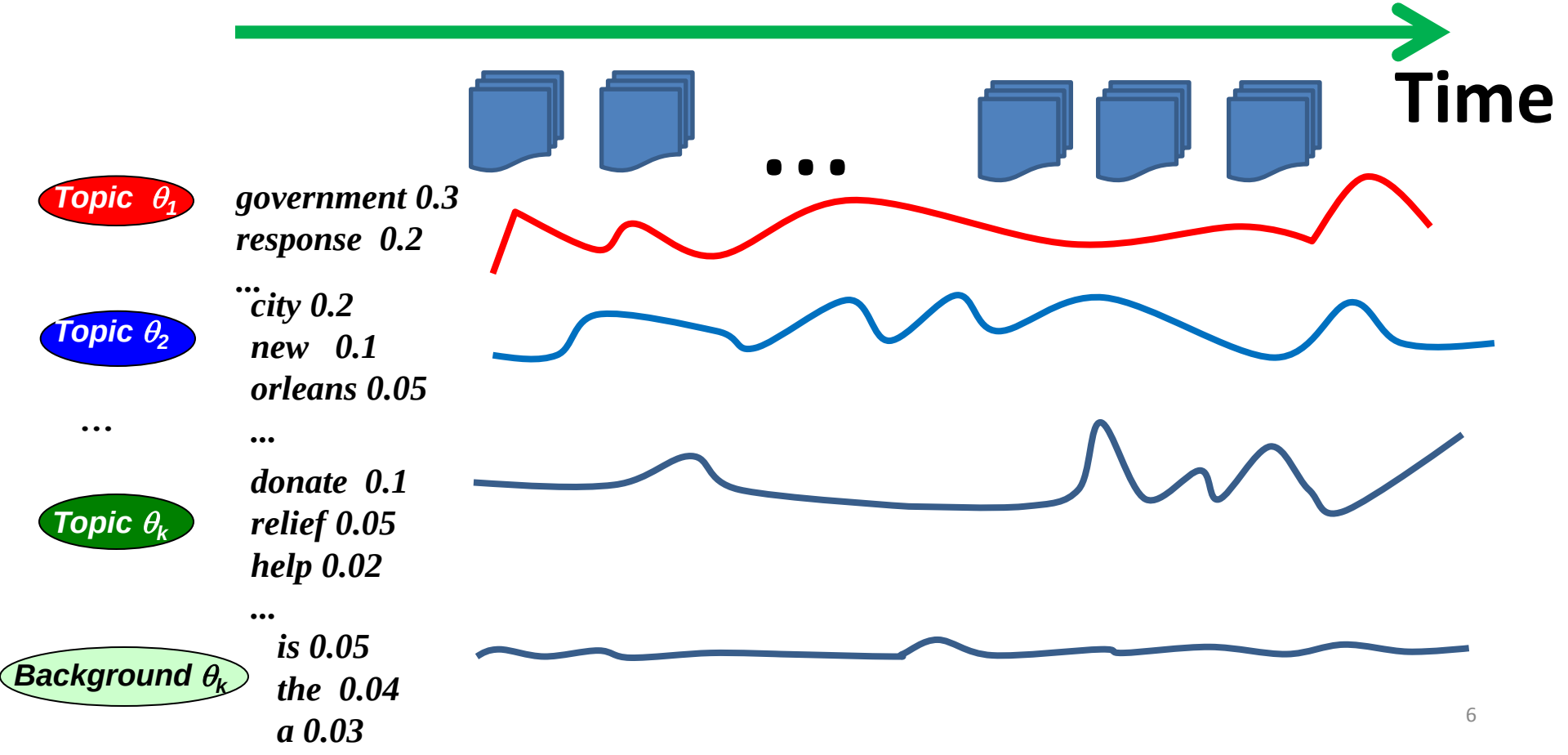
Any clues in the companion news stream?

Joint Analysis of Text and Time Series to Discover "Causal Topics"

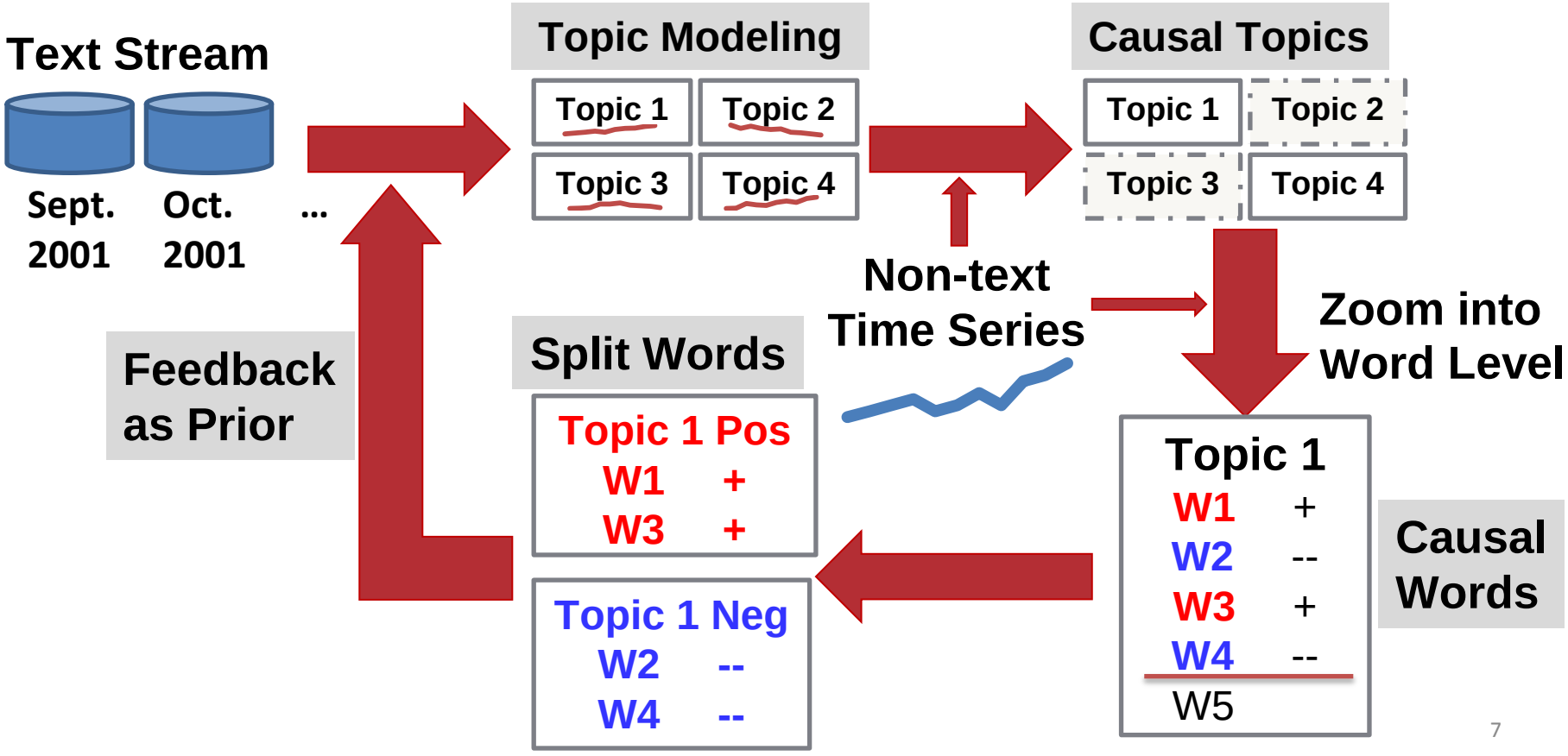
- Input:
  - Time series
  - Text data produced in a similar time period (text stream)
- Output
  - Topics whose coverage in the text stream has strong correlations with the time series ("causal" topics)



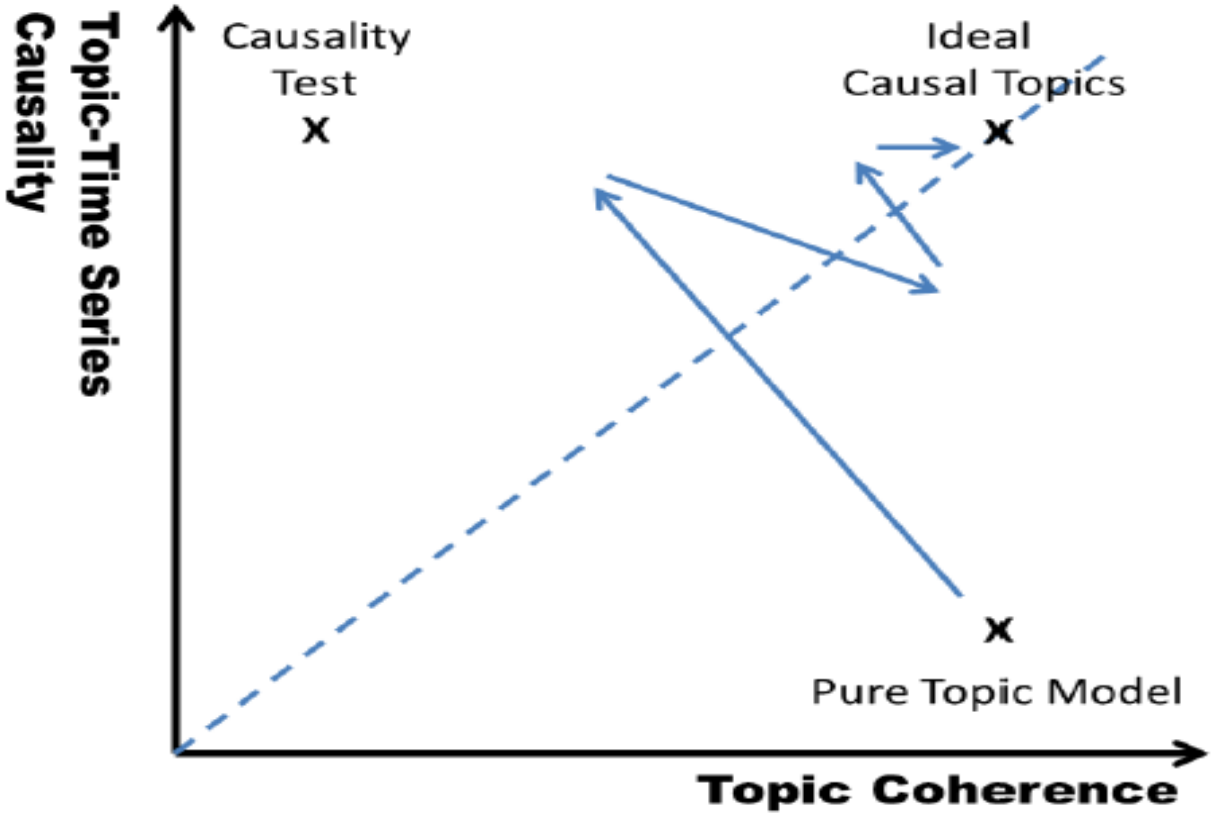
When a Topic Model Applied to Text Stream



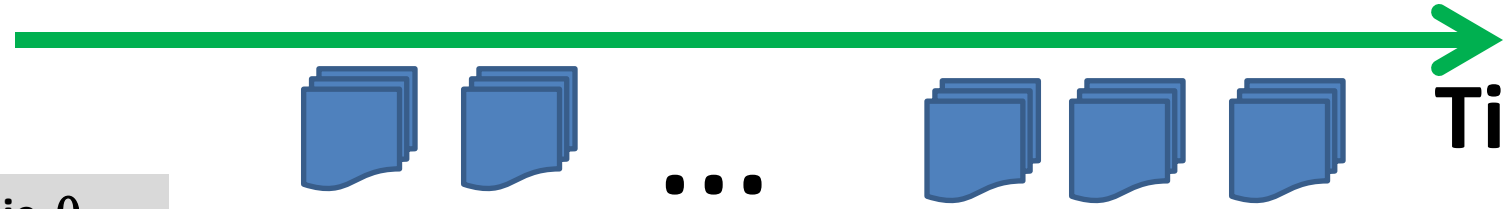
Iterative Causal Topic Modeling [Kim et al. 13]



Heuristic Optimization of Causality + Coherence



Measuring Causality (Correlation)



Topic θ_i

government 0.3
response 0.2
…

…

X_t

Causality(X_t,Y_t)=?

Does X_t cause Y_t?

Correlation(X_t,Y_t)=?

External Time Series (e.g. stock prices)

Y_t

Granger Causality Test is often useful [Seth 07]

Topics in NY Times Correlated with Stocks [Kim et al. 13]: June 2000 ~ Dec. 2011

AAMRQ (American Airlines)	AAPL (Apple)
russia russian putin	paid notice st
europe european	russia russian europe
germany	olympic games olympics
bush gore presidential	she her ms
police court judge	oil ford prices
airlines airport air	black fashion blacks
united trade terrorism	**computer technology software**
food foods cheese	**internet com web**
nets scott basketball	football giants jets
tennis williams open	japan japanese plane
awards gay boy	
moss minnesota chechnya	

Topics are biased toward each time series

Major Topics in 2000 Presidential Election [Kim et al. 13]

Top Three Words in Significant Topics from NY Times
tax cut 1
screen pataki guiliani
enthusiasm door symbolic
oil energy prices
news w top
pres al vice
love tucker presented
partial **abortion** privatization
court supreme **abortion**
gun control nra

Text: NY Times (May 2000 - Oct. 2000)

Time Series: Iowa Electronic Market
http://tippie.uiowa.edu/iem/

Issues known to be important in the 2000 presidential election

Suggested Reading

- **[Kim et al. 13]** Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas Rietz, and Daniel Diermeier. 2013. Mining causal topics in text data: Iterative topic modeling with time series feedback. In *Proceedings of the 22nd ACM international conference on information & knowledge management* (CIKM 2013). ACM, New York, NY, USA, 885-890. DOI=10.1145/2505515.2505612
- **[Seth 07]** Anil Seth, Granger Causality. 2007. *Scholarpedia*, 2(7): 1667, doi: 10.4249/scholarpedia.1667

Summary of Text-Based Prediction

- Text-based prediction is very useful for "big data" applications:
  - Inferring new knowledge about the world
  - Optimizing decision making
- Text data is often combined with non-text data for prediction
  - Joint analysis of text and non-text is necessary and useful
  - Non-text data provide context for mining text data (contextual text mining)
  - Text data help interpret patterns discovered from non-text data (pattern annotation)
- An active research topic with many open challenges