# Canonical Trend Analysis for Social Networks

Felix Bießmann,
Jens-Michalis Papaioannou,
Mikio Braun, Matthias L. Jugel,
Klaus-Robert Müller, Andreas Harth

Berlin Institute of Technology
Department Machine Learning

Karlsruhe Institute of Technology

# Temporal Dynamics of Web Data

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

Understanding dynamic graphs [Leskovec et al, KDD, 2005]

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

- Understanding dynamic graphs [Leskovec et al, KDD, 2005]
- Causal Inference [Lozano and Sindhwani, NIPS 2010]

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

- Understanding dynamic graphs [Leskovec et al, KDD, 2005]
- Causal Inference [Lozano and Sindhwani, NIPS 2010]
- Diffusion of information [Gomez Rodriguez et al, ICML 2011/2012]

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

- Understanding dynamic graphs [Leskovec et al, KDD, 2005]
- Causal Inference [Lozano and Sindhwani, NIPS 2010]
- Diffusion of information [Gomez Rodriguez et al, ICML 2011/2012]

Canonical Trend Analysis

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

Understanding dynamic graphs [Leskovec et al, KDD, 2005]

Causal Inference [Lozano and Sindhwani, NIPS 2010]

Diffusion of information [Gomez Rodriguez et al, ICML 2011/2012]

Canonical Trend Analysis

- Exploits temporal structure to find trends

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

Understanding dynamic graphs [Leskovec et al, KDD, 2005]

Causal Inference [Lozano and Sindhwani, NIPS 2010]

Diffusion of information [Gomez Rodriguez et al, ICML 2011/2012]

Canonical Trend Analysis

- Exploits temporal structure to find trends
- Find web sources that precede/follow trends

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

Understanding dynamic graphs [Leskovec et al, KDD, 2005]

Causal Inference [Lozano and Sindhwani, NIPS 2010]

Diffusion of information [Gomez Rodriguez et al, ICML 2011/2012]

Canonical Trend Analysis

- Exploits temporal structure to find trends
- Find web sources that precede/follow trends

Examples:

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

Understanding dynamic graphs [Leskovec et al, KDD, 2005]

Causal Inference [Lozano and Sindhwani, NIPS 2010]

Diffusion of information [Gomez Rodriguez et al, ICML 2011/2012]

Canonical Trend Analysis

- Exploits temporal structure to find trends
- Find web sources that precede/follow trends

Examples:

- Spatiotemporal Dynamics of Retweets to News Articles

# Temporal Dynamics of Web Data

Web content is copied, repeated or rephrased (Trends/Memes)

This temporal structure contains important information

Growing interest in **temporal dynamics of graphs**

Understanding dynamic graphs [Leskovec et al, KDD, 2005]

Causal Inference [Lozano and Sindhwani, NIPS 2010]

Diffusion of information [Gomez Rodriguez et al, ICML 2011/2012]

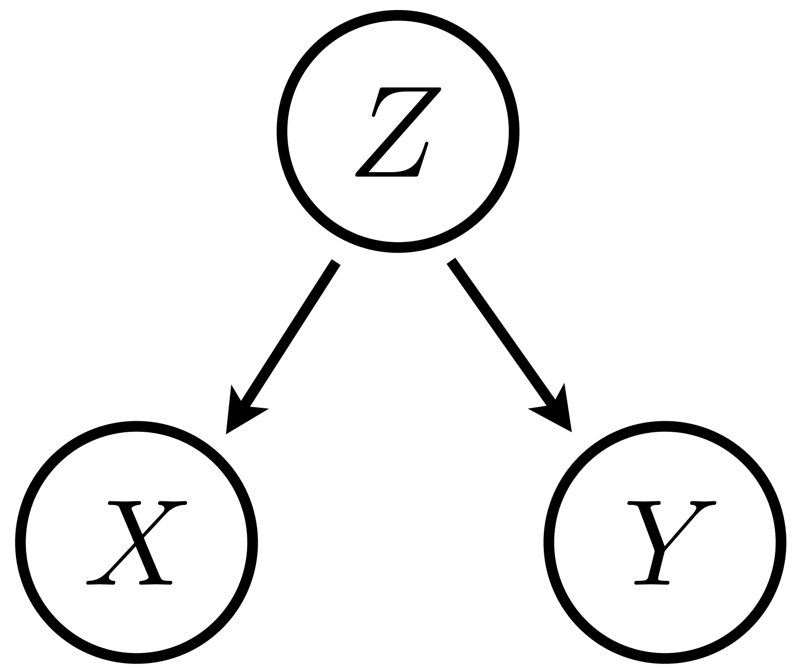
Canonical Trend Analysis

- Exploits temporal structure to find trends
- Find web sources that precede/follow trends

Examples:

- Spatiotemporal Dynamics of Retweets to News Articles
- Music trends on Last.fm

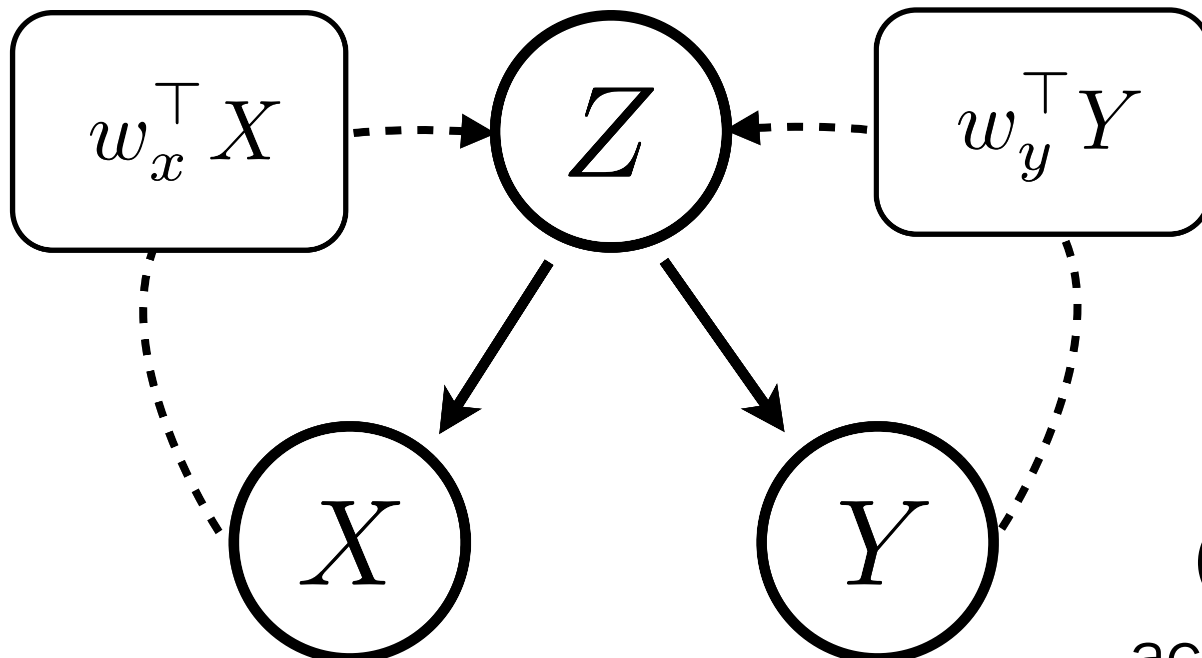
# Canonical Correlation Analysis




Latent Variable
(Trend)

Features
(e.g. Bag of Words, User actions, edge histograms, ...)

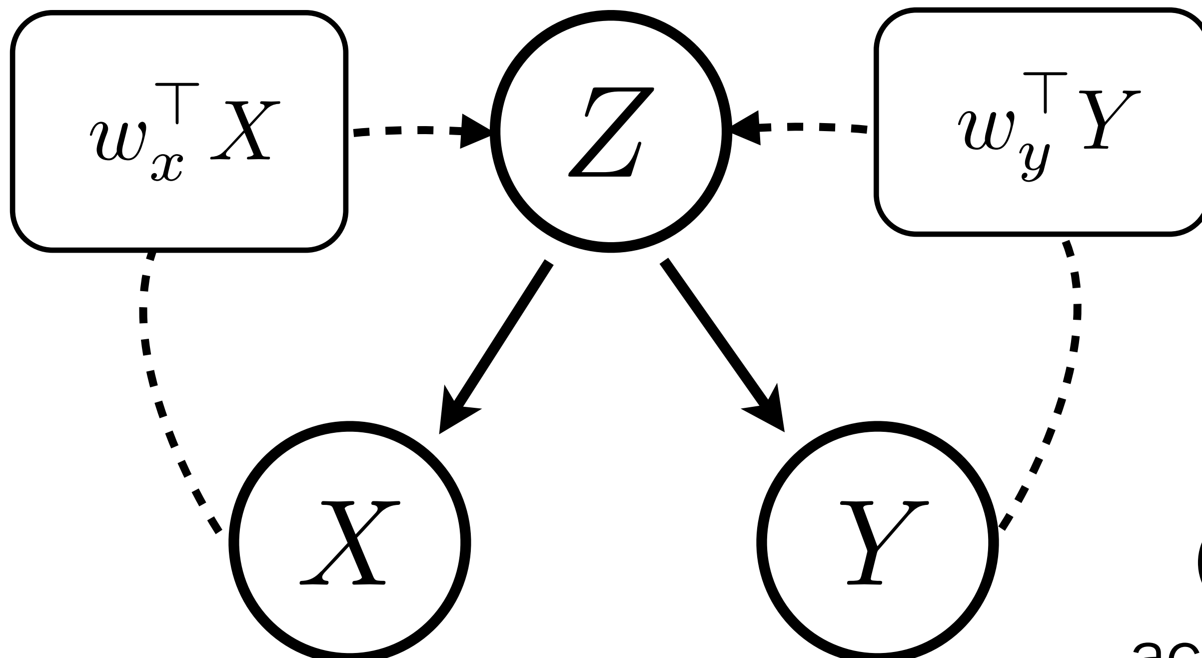
# Canonical Correlation Analysis




Latent Variable
(Trend)

Features
(e.g. Bag of Words, User actions, edge histograms, ...)

# Canonical Correlation Analysis



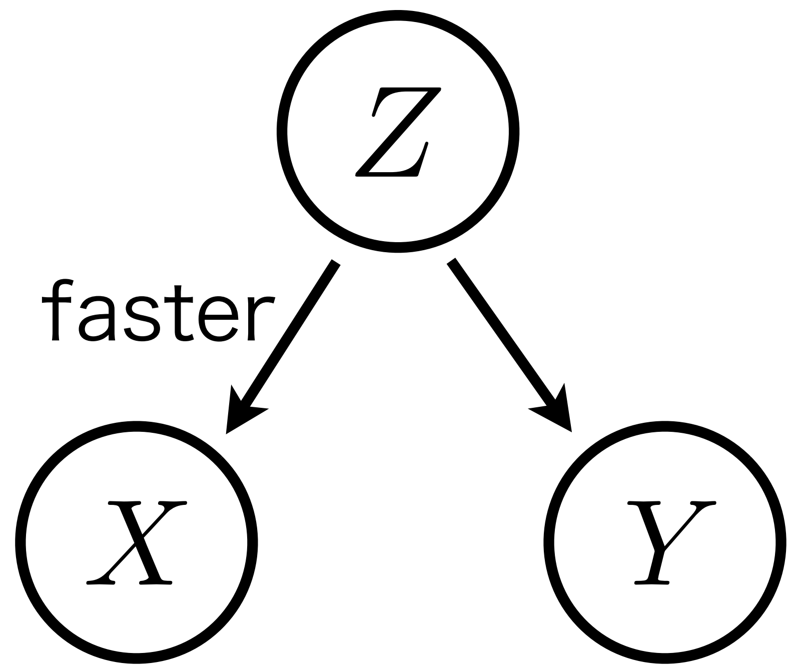

Latent Variable
(Trend)

Features
(e.g. Bag of Words, User actions, edge histograms, ...)

$$\operatorname*{argmax}_{w_x,\ w_y} \frac{w_x^\top X Y^\top w_y}{\sqrt{w_x^\top X X^\top w_x w_y^\top Y Y^\top w_y}}$$

[Jordan 1875], [Hotelling 1936], [Bach and Jordan 2006]

# Canonical Trend Model

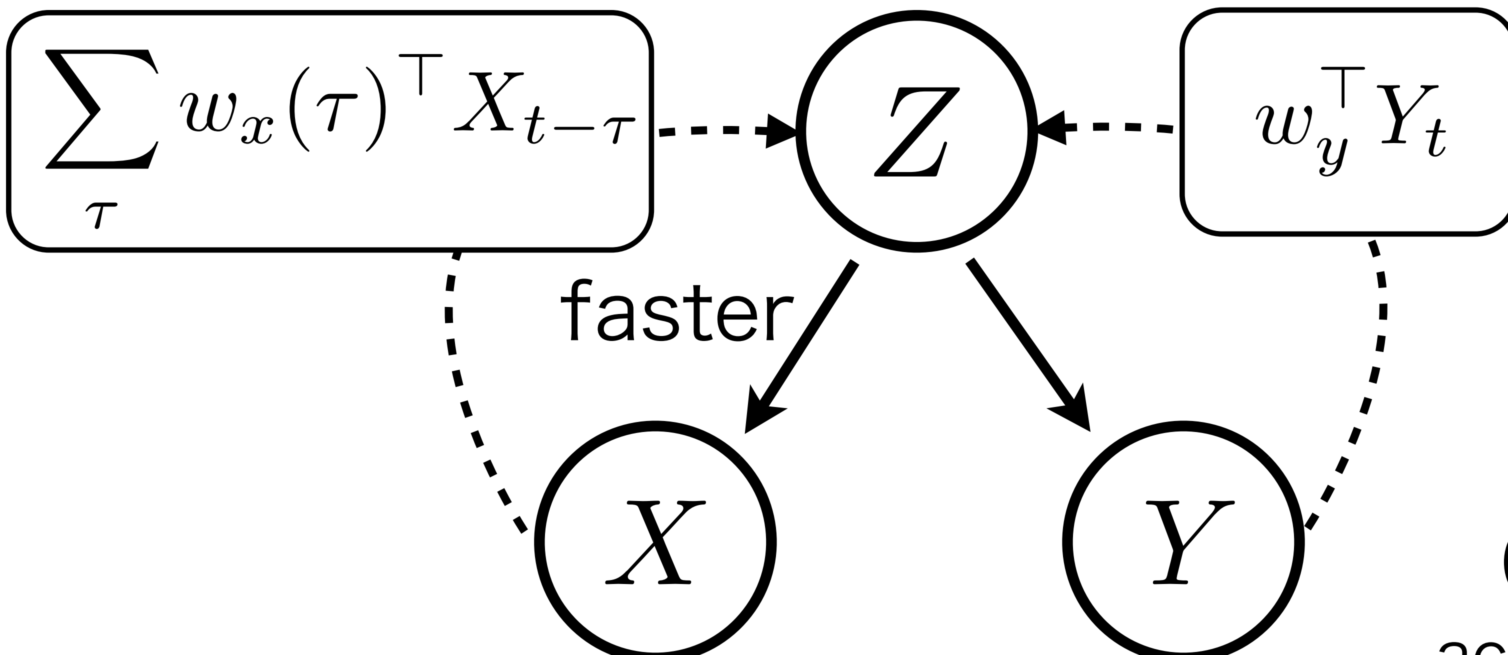



Latent Variable
(Trend)

Features

(e.g. Bag of Words, User actions, edge histograms, ...)

# Canonical Trend Model




Latent Variable
(Trend)

Features
(e.g. Bag of Words, User actions, edge histograms, ...)

# An Example on News Trends

# An Example on News Trends

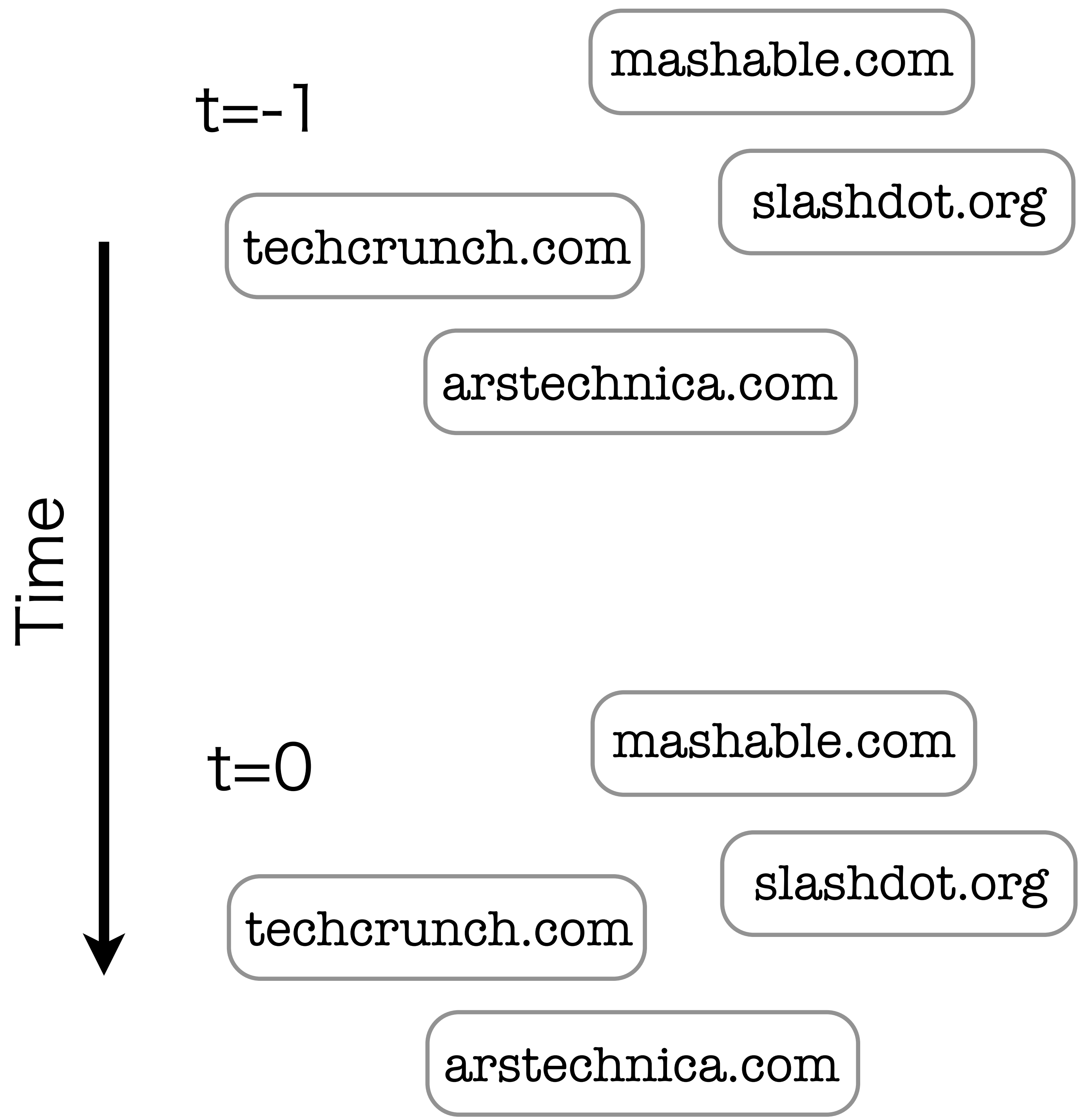
mashable.com

slashdot.org

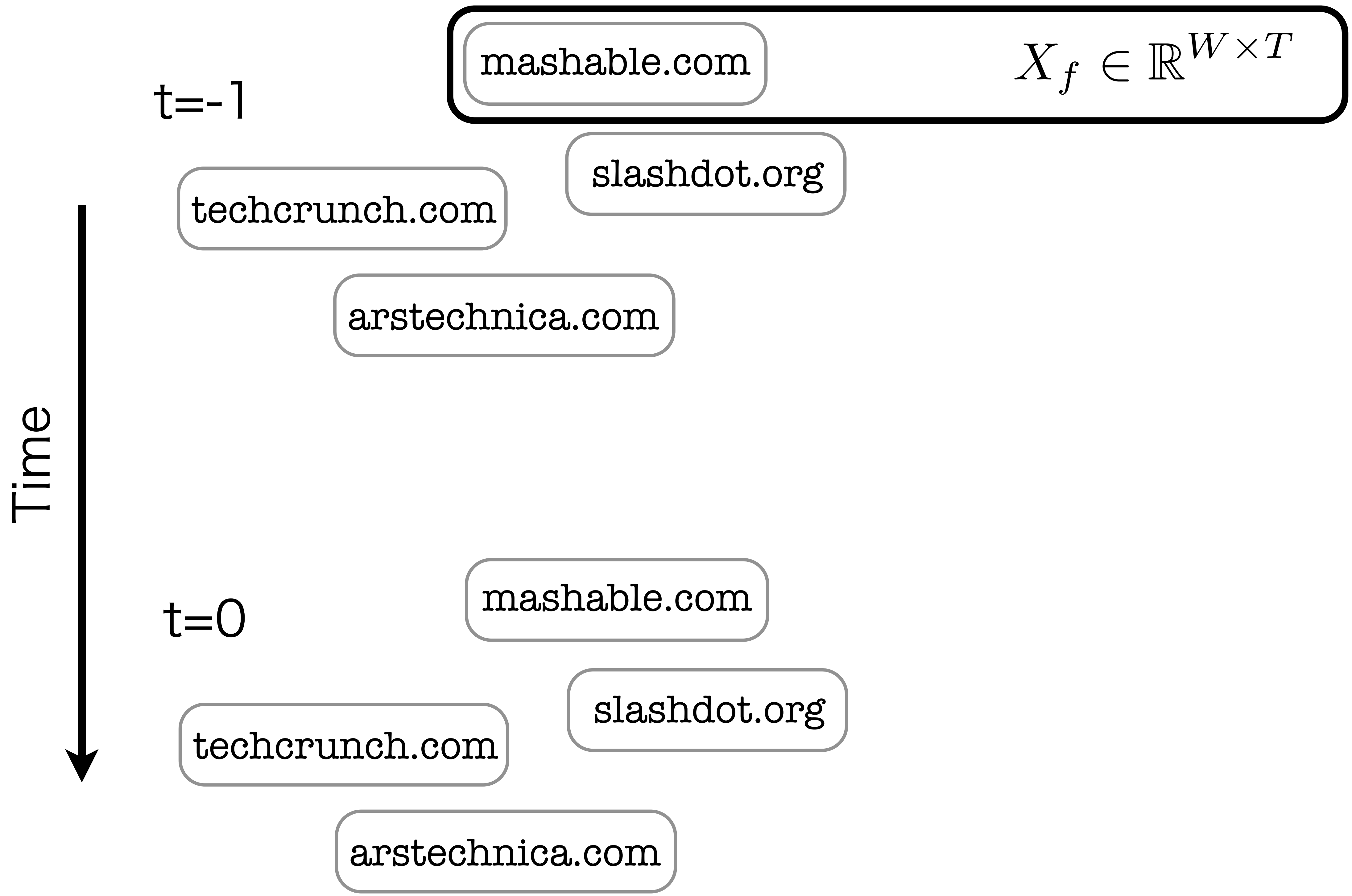
techcrunch.com

arstechnica.com

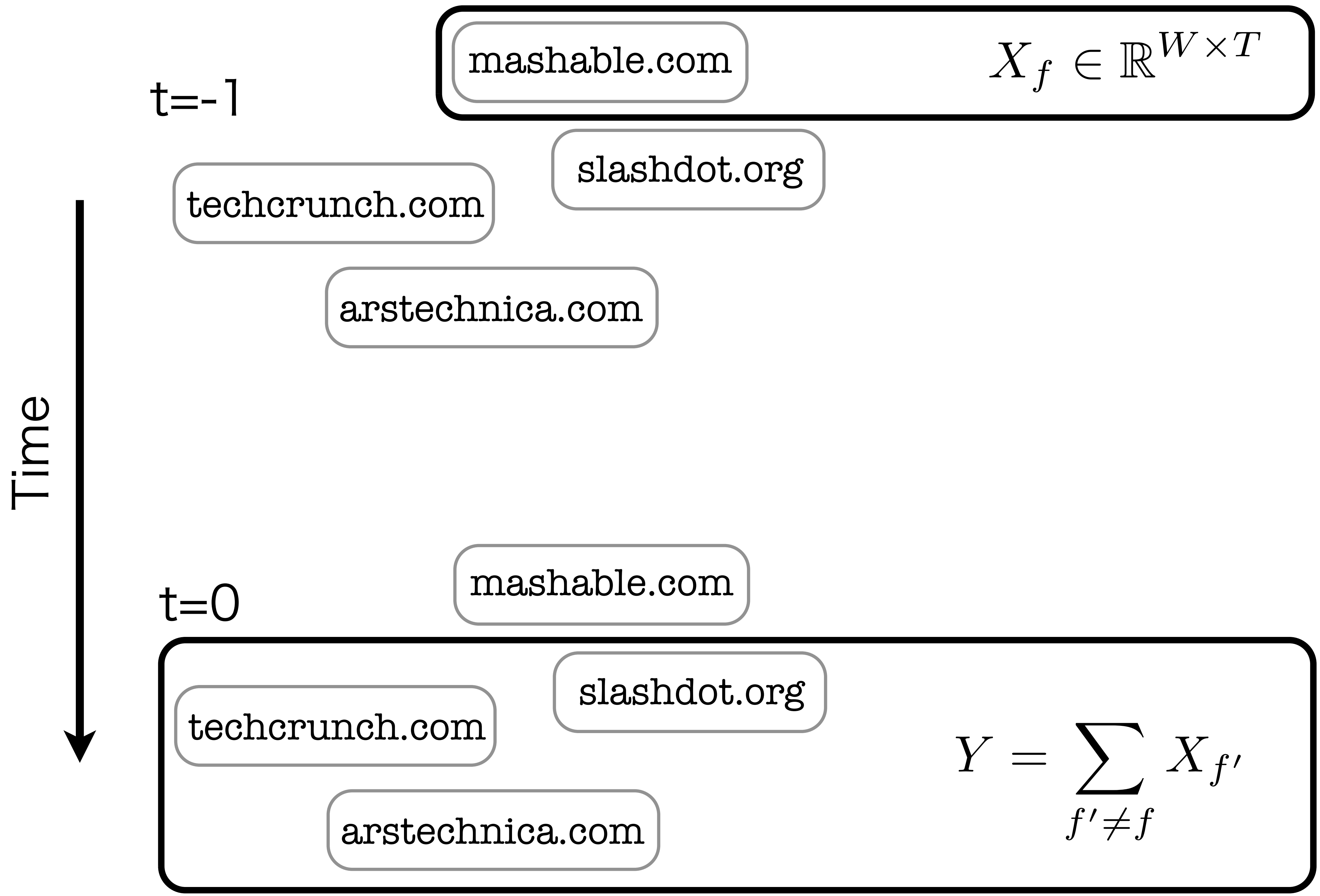
# An Example on News Trends



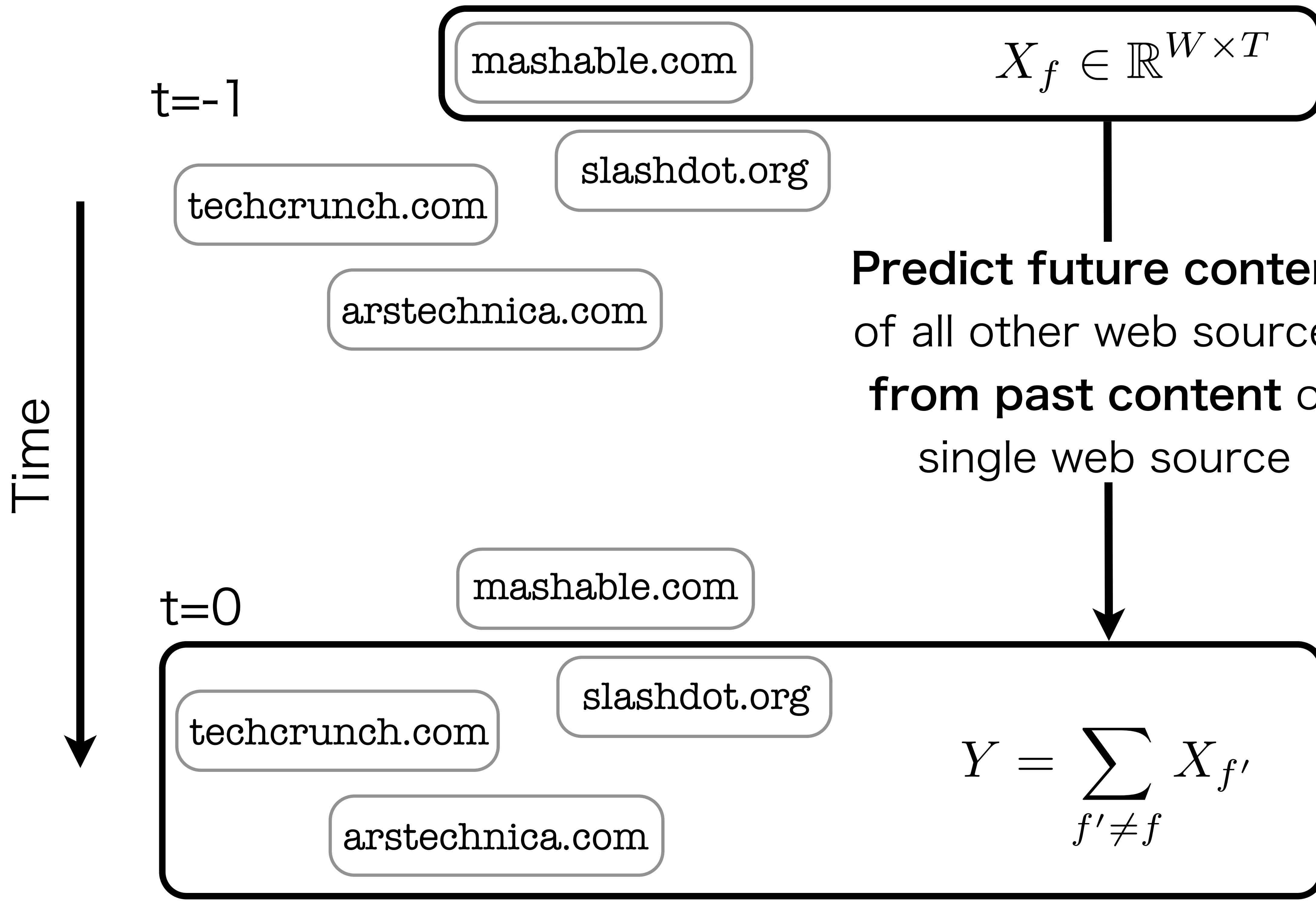
# An Example on News Trends



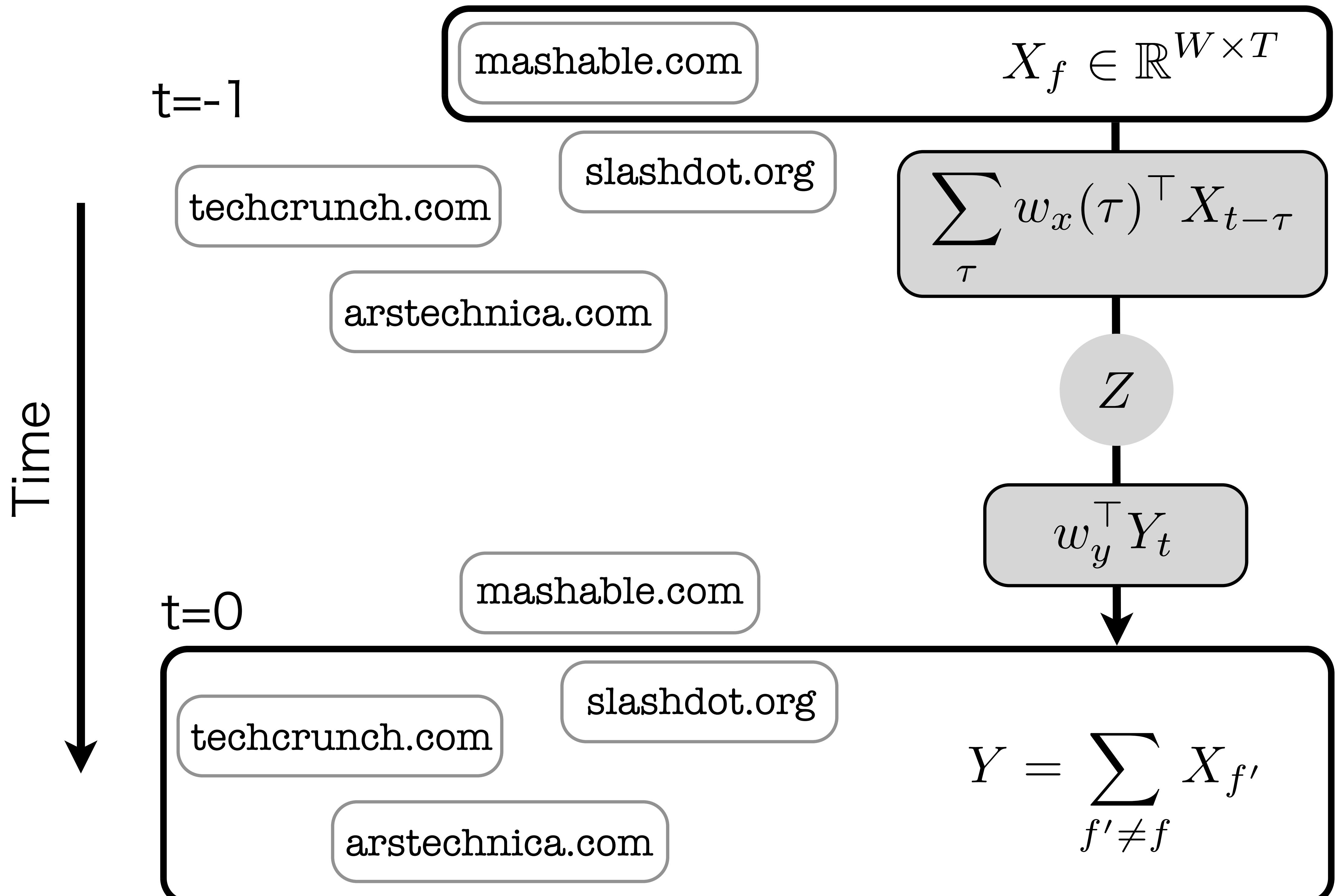
# An Example on News Trends



# An Example on News Trends



# An Example on News Trends



# Why Projecting to Canonical Subspace?

Easily interpretable: For Text data each canonical direction is a topic

[De Bie and Cristianini, 2004]

Information theoretic optimal compression

[Creutzig 2009]

Conversion of canonical correlations to granger causality index

[Otter 1991]

# Canonical Trend Analysis For Social Networks

Quantifying spatiotemporal **retweet** response to **news content**

# Canonical Trend Analysis For Social Networks

Quantifying spatiotemporal **retweet** response to **news content**

Finding users ahead and following music trends on **Last.fm**

# Canonical Trend Analysis For Social Networks


Some **news web site** publishes some content …

# Canonical Trend Analysis For Social Networks


Some **news web site** publishes some content ...

... which is **retweeted**

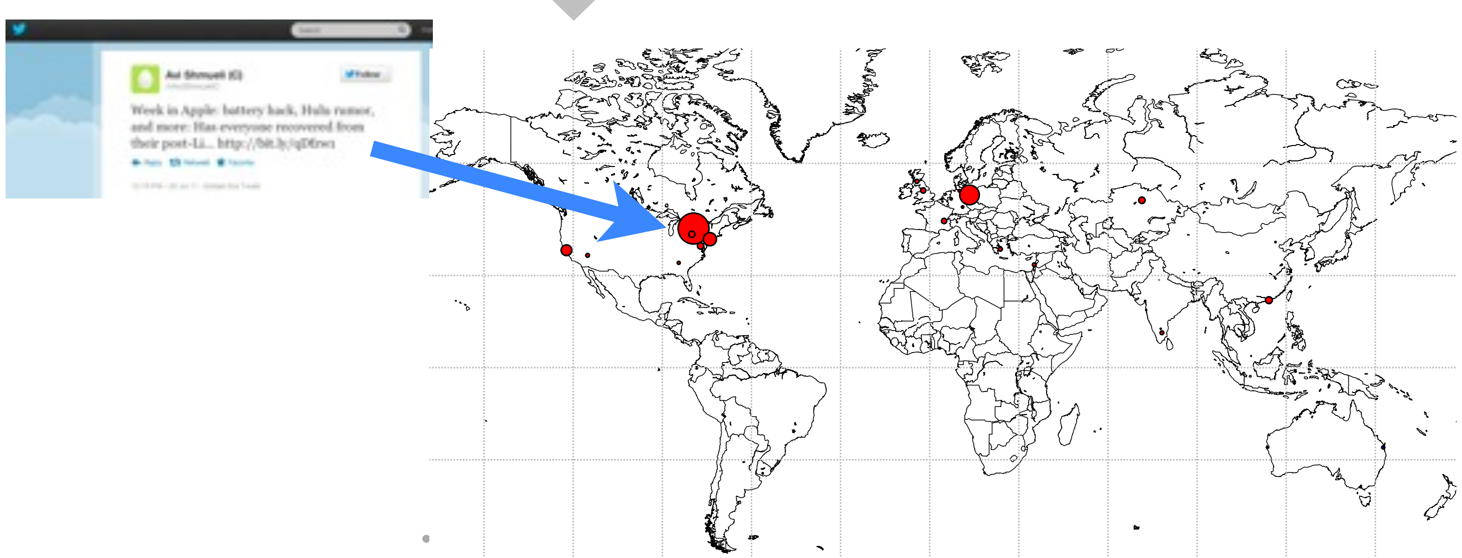

$t$ $t + \tau_1$ Time

# Canonical Trend Analysis For Social Networks


Some **news web site** publishes some content ...

... which is **retweeted**

... at different locations



$t$ $\quad t + \tau_1$ $\quad$ Time

# Data Extraction

# Data Extraction

For each news site $f \in \{1, 2, \ldots, F\}$ extract

For each news site $f \in \{1, 2, \ldots, F\}$ extract

**Bag-of-Words Features**

$$X_f = [x_f(t=1), \ldots, x_f(t=T)] \in \mathbb{R}^{W \times T}$$

# Data Extraction

For each news site $f \in \{1, 2, \ldots, F\}$ extract

**Bag-of-Words Features**

$$X_f = [x_f(t=1), \ldots, x_f(t=T)] \in \mathbb{R}^{W \times T}$$

**Retweet locations**

$$Y_f = [y_f(t=1), \ldots, y_f(t=T)] \in \mathbb{R}^{L \times T}$$

# Data Extraction: Retweet Locations

# Data Extraction: Retweet Locations

1. Extract URI of each news article in twitter stream

# Data Extraction: Retweet Locations

1. Extract URI of each news article in twitter stream
2. Retrieve Location from Twitter User Profile

# Data Extraction: Retweet Locations

1. Extract URI of each news article in twitter stream

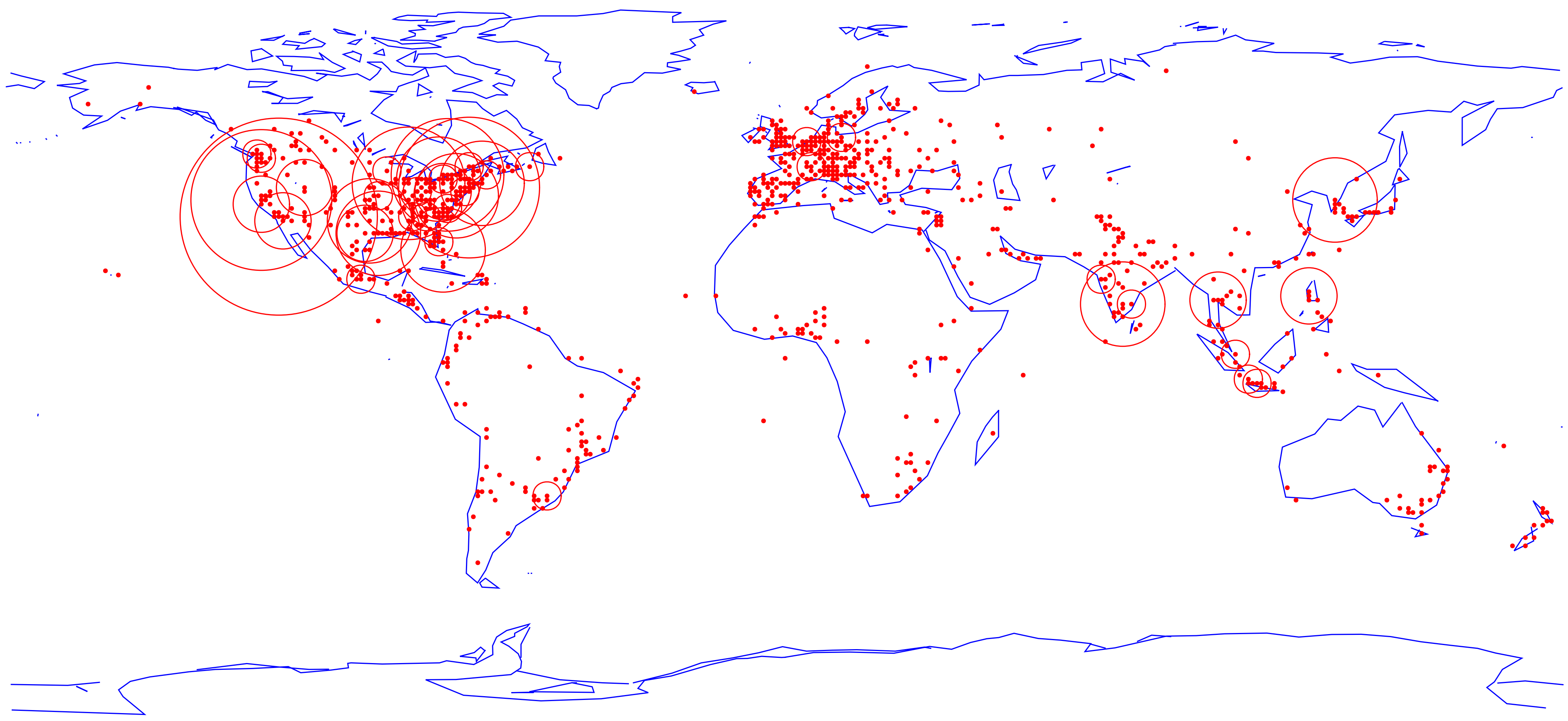
2. Retrieve Location from Twitter User Profile

3. Resolve Ambiguities / Remove non-sense Locations

# Data Extraction: Retweet Locations

1. Extract URI of each news article in twitter stream
2. Retrieve Location from Twitter User Profile
3. Resolve Ambiguities / Remove non-sense Locations
4. Downsample Geographic Locations

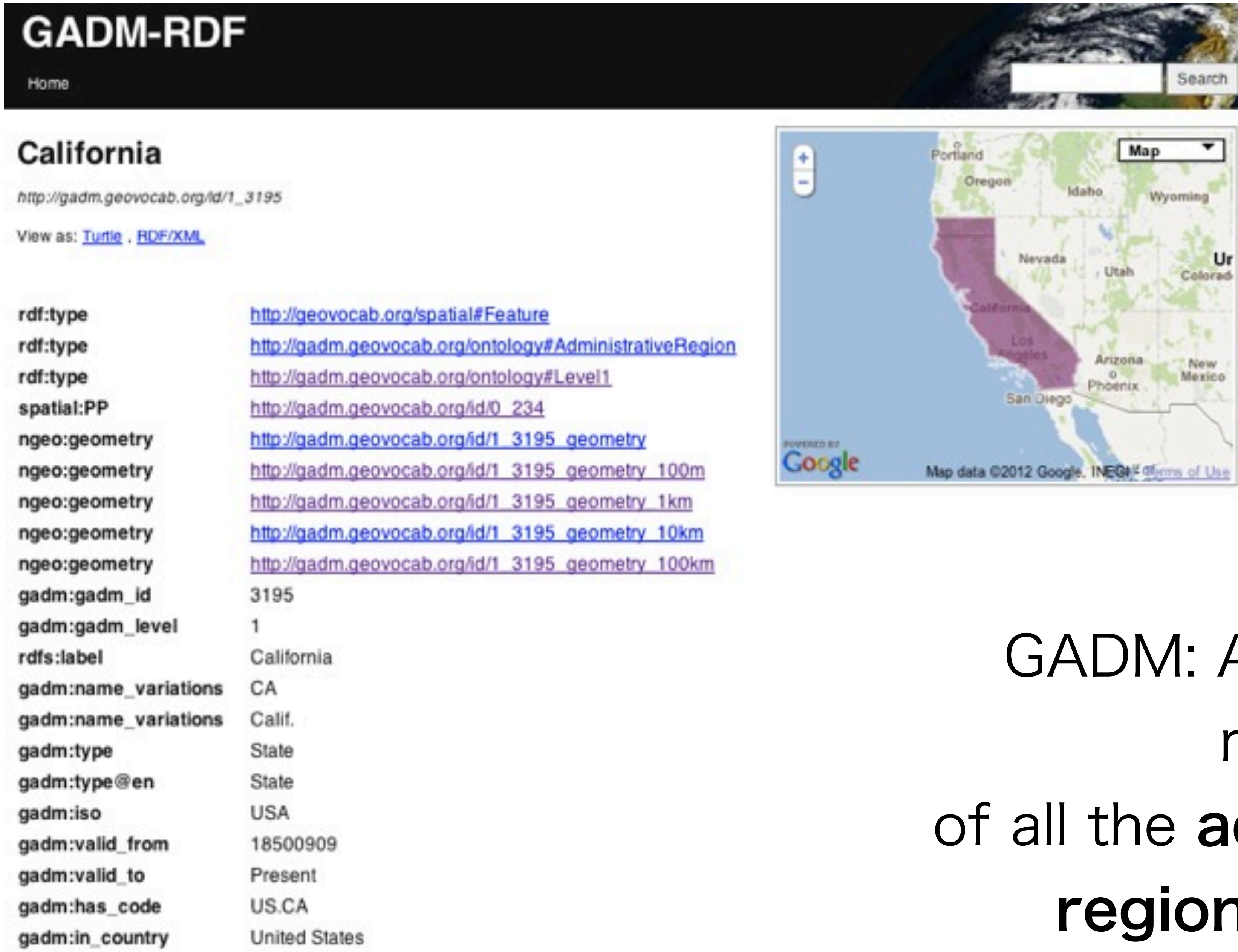
# Mean Locations of Retweeted News Articles





Canonical Trends

# Downsampling of Geographic Information



**GADM-RDF**

Home

Search

## California

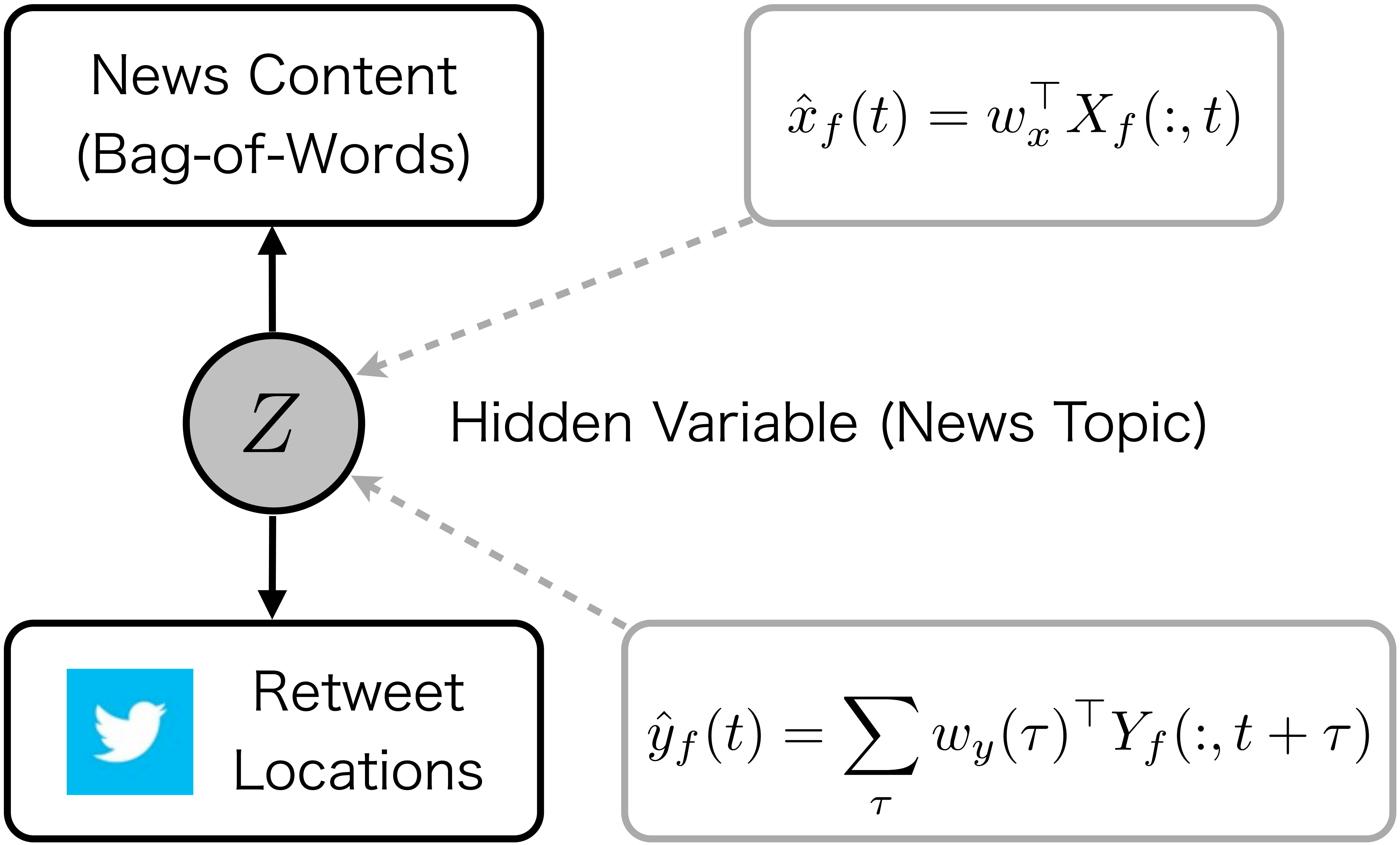
*http://gadm.geovocab.org/id/1_3195*

View as: Turtle , RDF/XML

| | |
|---|---|
| rdf:type | http://geovocab.org/spatial#Feature |
| rdf:type | http://gadm.geovocab.org/ontology#AdministrativeRegion |
| rdf:type | http://gadm.geovocab.org/ontology#Level1 |
| spatial:PP | http://gadm.geovocab.org/id/0_234 |
| ngeo:geometry | http://gadm.geovocab.org/id/1_3195_geometry |
| ngeo:geometry | http://gadm.geovocab.org/id/1_3195_geometry_100m |
| ngeo:geometry | http://gadm.geovocab.org/id/1_3195_geometry_1km |
| ngeo:geometry | http://gadm.geovocab.org/id/1_3195_geometry_10km |
| ngeo:geometry | http://gadm.geovocab.org/id/1_3195_geometry_100km |
| gadm:gadm_id | 3195 |
| gadm:gadm_level | 1 |
| rdfs:label | California |
| gadm:name_variations | CA |
| gadm:name_variations | Calif. |
| gadm:type | State |
| gadm:type@en | State |
| gadm:iso | USA |
| gadm:valid_from | 18500909 |
| gadm:valid_to | Present |
| gadm:has_code | US.CA |
| gadm:in_country | United States |



GADM: An RDF spatial representation of all the **administrative regions** in the world


News Content
(Bag-of-Words)
$\hat{x}_f(t) = w_x^\top X_f(:,t)$
$Z$
Hidden Variable (News Topic)
Retweet
Locations
$\hat{y}_f(t) = \sum_\tau w_y(\tau)^\top Y_f(:,t+\tau)$

# Canonical Trend Analysis

News Content (Bag-of-Words)

$$\hat{x}_f(t) = w_x^\top X_f(:,t)$$

Retweet Locations

$$\hat{y}_f(t) = \sum_\tau w_y(\tau)^\top Y_f(:,t+\tau)$$

Optimal $w_x \in \mathbb{R}^W$ and $w_y(\tau) \in \mathbb{R}^{WN_\tau}$

$$\underset{w_y(\tau),w_x}{\operatorname{argmax}} \operatorname{Corr}(\hat{x}_f(t), \hat{y}_f(t))$$

# Efficient Computation of Canonical Trends

$$\tilde{Y}_f = \begin{bmatrix} Y_{f,\tau=1} \\ \vdots \\ Y_{f,\tau=N_\tau} \end{bmatrix} \in \mathbb{R}^{LN_\tau \times T}$$

[Takens 1981]

**(linear) 'Kernel Trick'**

- Very efficient for high-dimensional feature spaces

# Efficient Computation of Canonical Trends

**Temporal Embedding**

[Takens 1981]

$$\tilde{Y}_f = \begin{bmatrix} Y_{f,\tau=1} \\ \vdots \\ Y_{f,\tau=N_\tau} \end{bmatrix} \in \mathbb{R}^{LN_\tau \times T}$$

- Standard CCA problem

[Jordan 1875], [Hotelling 1936], [Anderson 1999]

**(linear) 'Kernel Trick'**

$$w_y(\tau) = Y_{f,\tau}\alpha$$
$$w_x = X_f\beta$$

- Very efficient for high-dimensional feature spaces

[Fyfe 2000], [Fukumizu 2007]

# Efficient Computation of Canonical Trends

Objective function is maximized in the dual

$$\mathrm{Corr}(\hat{x}(t), \hat{y}(t)) = \frac{\sum_\tau (w_y(\tau)^\top Y_\tau)^\top X w_x}{\sqrt{\sum_\tau (w_y(\tau)^\top Y_\tau Y_\tau^\top w_y(\tau)) w_x^\top X X^\top w_x}}$$
$$= \frac{\alpha^\top K_{\tilde{Y}} K_X \beta}{\sqrt{\alpha^\top K_{\tilde{Y}}^2 \alpha \beta^\top K_X^2 \beta}}$$

where $$K_{\tilde{Y}} = \tilde{Y}^\top \tilde{Y}$$ $$K_X = X^\top X$$ are linear kernels

# Efficient Computation of Canonical Trends

$$\mathrm{Corr}(\hat{x}(t), \hat{y}(t)) = \frac{\sum_\tau (w_y(\tau)^\top Y_\tau)^\top X w_x}{\sqrt{\sum_\tau (w_y(\tau)^\top Y_\tau Y_\tau^\top w_y(\tau)) w_x^\top X X^\top w_x}}$$

$$= \frac{\alpha^\top K_{\tilde{Y}} K_X \beta}{\sqrt{\alpha^\top K_{\tilde{Y}}^2 \alpha \beta^\top K_X^2 \beta}}$$

Dual coefficients are solution to generalized eigenvalue equation

$$\begin{bmatrix} 0 & K_{\tilde{Y}} K_X \\ K_X K_{\tilde{Y}} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \lambda \begin{bmatrix} K_{\tilde{Y}}^2 + I\kappa_y & 0 \\ 0 & K_X^2 + I + \kappa_x \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

# Efficient Computation of Canonical Trends

$$\mathrm{Corr}(\hat{x}(t), \hat{y}(t)) = \frac{\sum_\tau (w_y(\tau)^\top Y_\tau)^\top X w_x}{\sqrt{\sum_\tau (w_y(\tau)^\top Y_\tau Y_\tau^\top w_y(\tau)) w_x^\top X X^\top w_x}}$$

$$= \frac{\alpha^\top K_{\tilde{Y}} K_X \beta}{\sqrt{\alpha^\top K_{\tilde{Y}}^2 \alpha \beta^\top K_X^2 \beta}}$$

Dual coefficients are solution to generalized eigenvalue equation

$$\begin{bmatrix} 0 & K_{\tilde{Y}} K_X \\ K_X K_{\tilde{Y}} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \lambda \begin{bmatrix} K_{\tilde{Y}}^2 + I\kappa_y & 0 \\ 0 & K_X^2 + I + \kappa_x \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

Bießmann et al, Machine Learning, 2010

# Comparisons: Mean, PCA and Canonical Trends

# Comparisons: Mean, PCA and Canonical Trends

**Canonical Trends**

**Mean**

**PCA**

# Comparisons: Mean, PCA and Canonical Trends

**Canonical Trends** $\underset{w_y(\tau), w_x}{\operatorname{argmax}} \operatorname{Corr}(\hat{x}_f(t), \hat{y}_f(t))$

**Mean**

**PCA**

# Comparisons: Mean, PCA and Canonical Trends

**Canonical Trends** $$\underset{w_y(\tau), w_x}{\operatorname{argmax}} \operatorname{Corr}(\hat{x}_f(t), \hat{y}_f(t))$$

**Mean** $$w_x^\top = \mathbf{1}_x/N, w_y(\tau) = \mathbf{1}_y/N$$

**PCA**

# Comparisons: Mean, PCA and Canonical Trends

**Canonical Trends**

$$\underset{w_y(\tau), w_x}{\operatorname{argmax}} \operatorname{Corr}(\hat{x}_f(t), \hat{y}_f(t))$$

**Mean**

$$w_x^\top = \mathbf{1}_x / N, w_y(\tau) = \mathbf{1}_y / N$$

**PCA**

$$\underset{w_y(\tau)}{\operatorname{argmax}} (w_y(\tau)^\top \tilde{Y}_f \tilde{Y}_f^\top w_y(\tau)),$$

$$\underset{w_x}{\operatorname{argmax}} (w_x^\top X X^\top w_x),$$

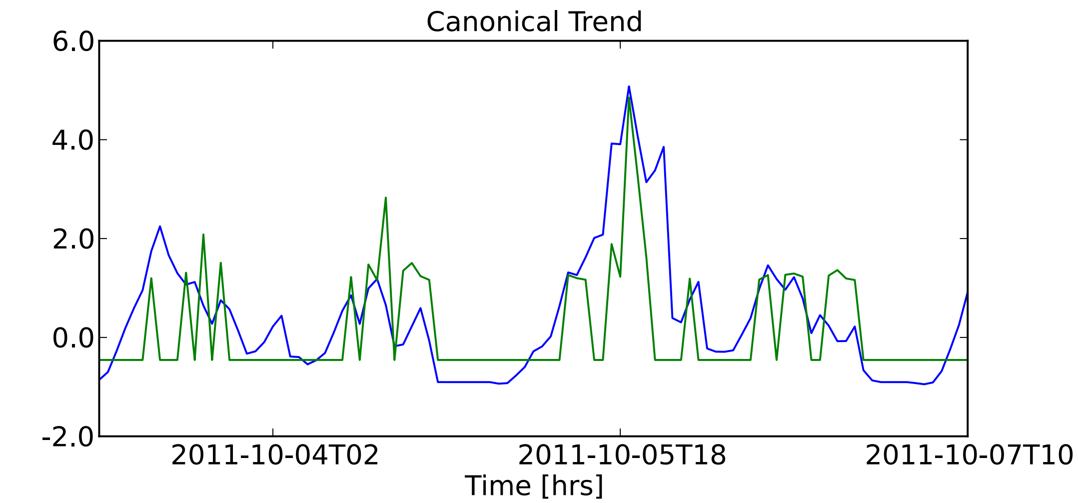
$$\text{s.t. } w_y(\tau)^\top w_y(\tau) = w_x^\top w_x = 1$$

# Comparisons: Mean, PCA and Canonical Trends

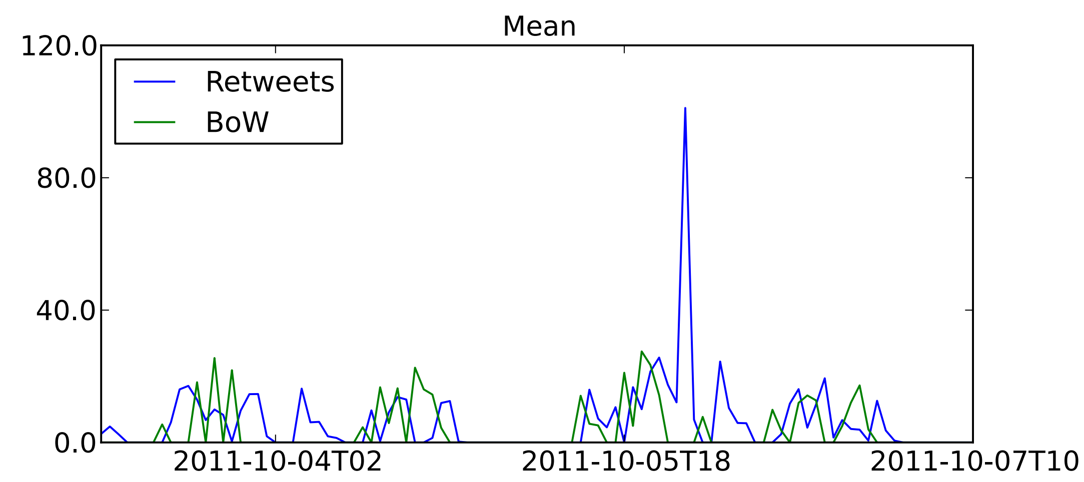
| | Hypothesis |
|---|---|
| **Canonical Trends** | News Content helps predicting retweet frequency |
| **Mean** | Mean Wordcount predicts mean tweet frequency best |
| **PCA** | Wordcount variance predicts tweet variance |

# Comparisons: Mean, PCA and Canonical Trends

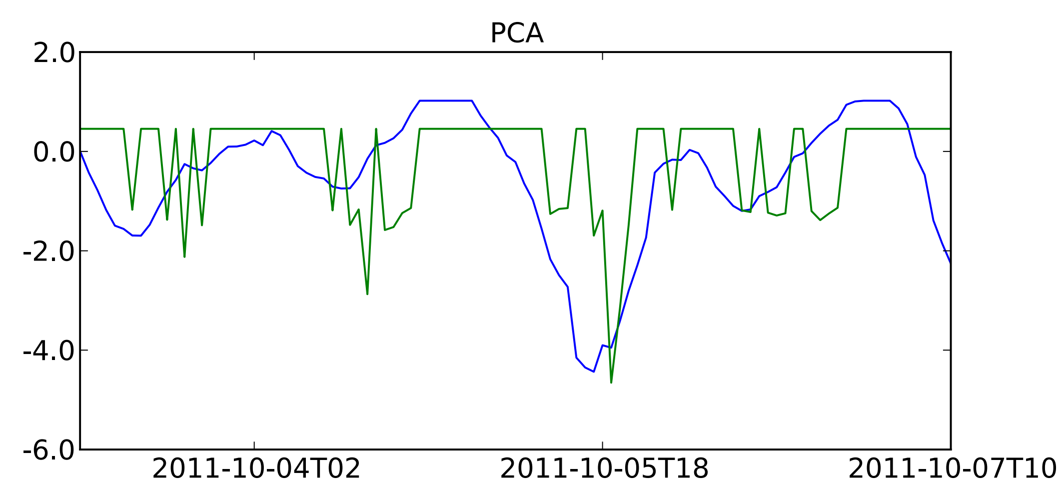
**Canonical Trends**



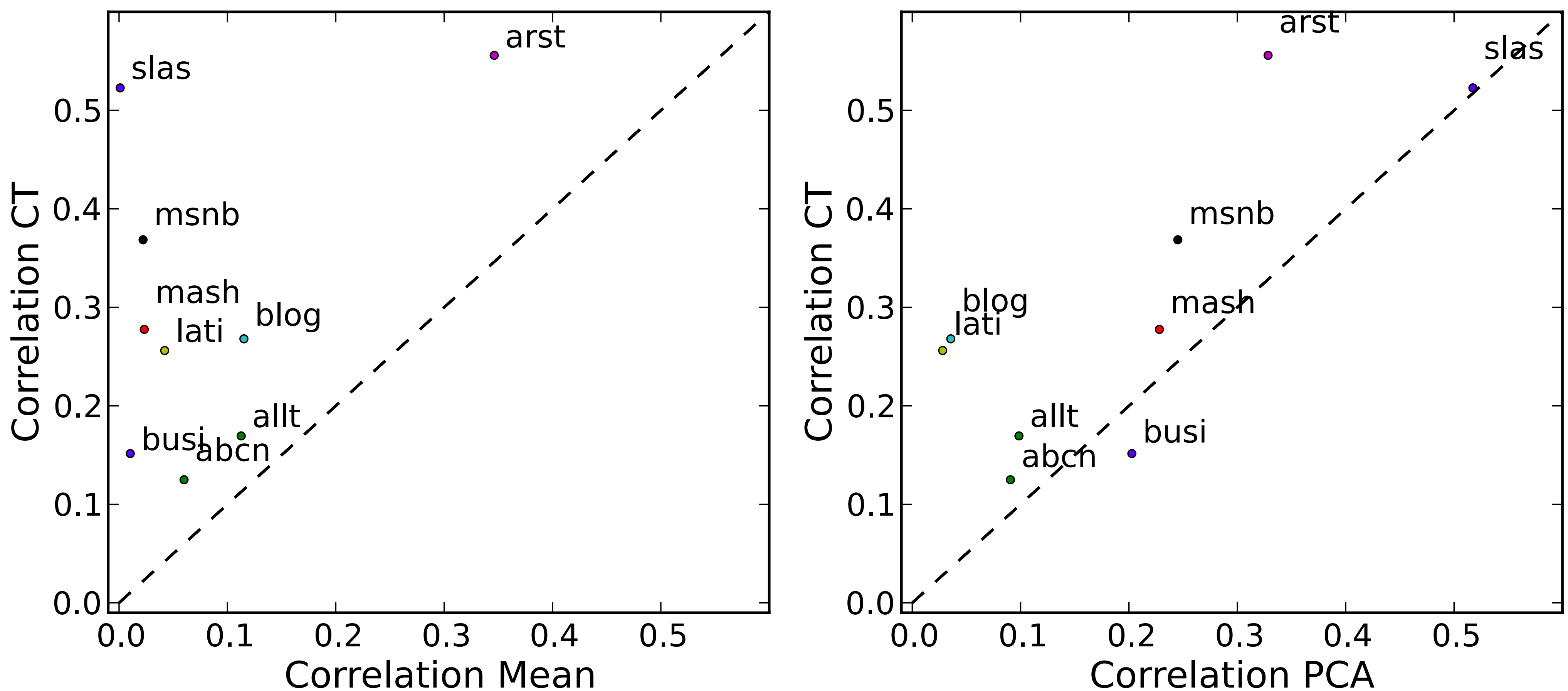

**Mean**



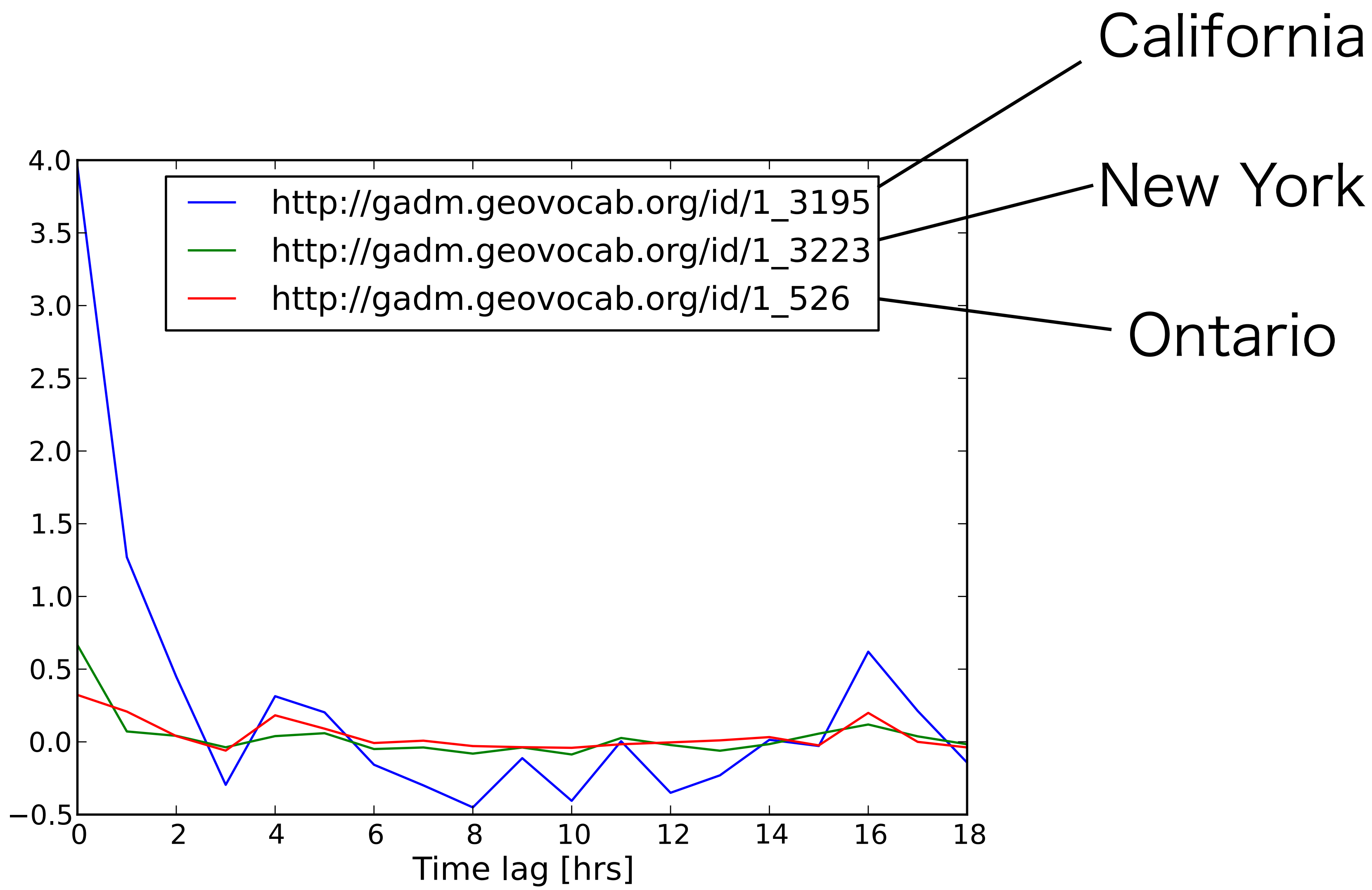

**PCA**



# Comparisons: Mean, PCA and Canonical Trends



# Canonical Convolution




Excerpts from LA Times

Spatiotemporal Response

# Spatiotemporal Analysis of Retweets of News

# Spatiotemporal Analysis of Retweets of News

We use canonical correlation analysis to compute

# Spatiotemporal Analysis of Retweets of News

We use canonical correlation analysis to compute

a Bag-of-Word subspace (topic) and

# Spatiotemporal Analysis of Retweets of News

We use canonical correlation analysis to compute

- a Bag-of-Word subspace (topic) and
- spatiotemporal twitter response patterns

# Spatiotemporal Analysis of Retweets of News

We use canonical correlation analysis to compute

a Bag-of-Word subspace (topic) and

spatiotemporal twitter response patterns

such that news content and retweets are maximally correlated

# Spatiotemporal Analysis of Retweets of News

We use canonical correlation analysis to compute

a Bag-of-Word subspace (topic) and

spatiotemporal twitter response patterns

such that news content and retweets are maximally correlated

Results can be interpreted w.r.t

# Spatiotemporal Analysis of Retweets of News

We use canonical correlation analysis to compute

a Bag-of-Word subspace (topic) and
spatiotemporal twitter response patterns

such that news content and retweets are maximally correlated

Results can be interpreted w.r.t

- How much impact has a news site on Twitter-Community

# Spatiotemporal Analysis of Retweets of News

We use canonical correlation analysis to compute

a Bag-of-Word subspace (topic) and
spatiotemporal twitter response patterns

such that news content and retweets are maximally correlated

Results can be interpreted w.r.t

- How much impact has a news site on Twitter-Community
- (Content that will lead to high retweet frequency)

# Spatiotemporal Analysis of Retweets of News

We use canonical correlation analysis to compute

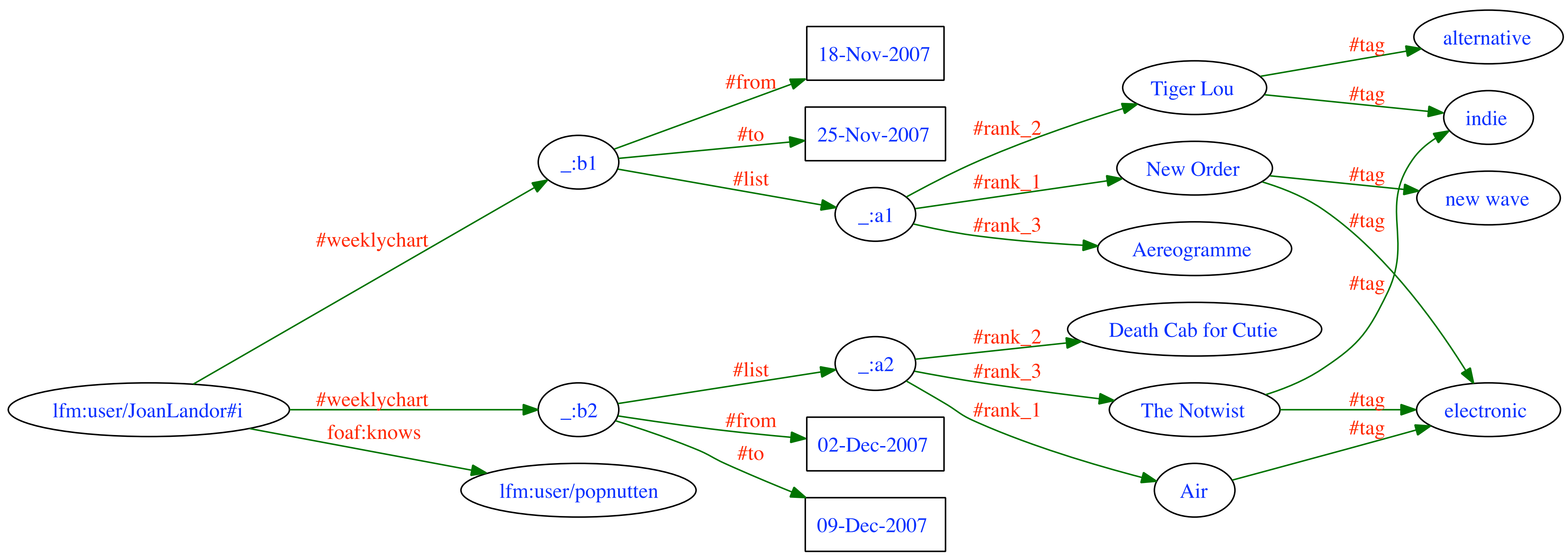
a Bag-of-Word subspace (topic) and
spatiotemporal twitter response patterns

such that news content and retweets are maximally correlated

Results can be interpreted w.r.t

- How much impact has a news site on Twitter-Community
- (Content that will lead to high retweet frequency)
- (Where and when maximal impact is reached)

# Users and Trends on Last.fm




A last.fm user subgraph

Extract Weekly Chartlist Last.fm-Music-tags

$$X_f = [x_f(t=1), \dots, x_f(t=T)] \in \mathbb{R}^{M \times T}$$

Single User Chart
Time Series

$$Y_f = \sum_{f' \neq f} X_{f'}$$

All Other Users

# Users and Trends on Last.fm

Extract Weekly Chartlist Last.fm-Music-tags

$$X_f = [x_f(t=1), \dots, x_f(t=T)] \in \mathbb{R}^{M \times T}$$

Single User Chart
Time Series

$$Y_f = \sum_{f' \neq f} X_{f'}$$

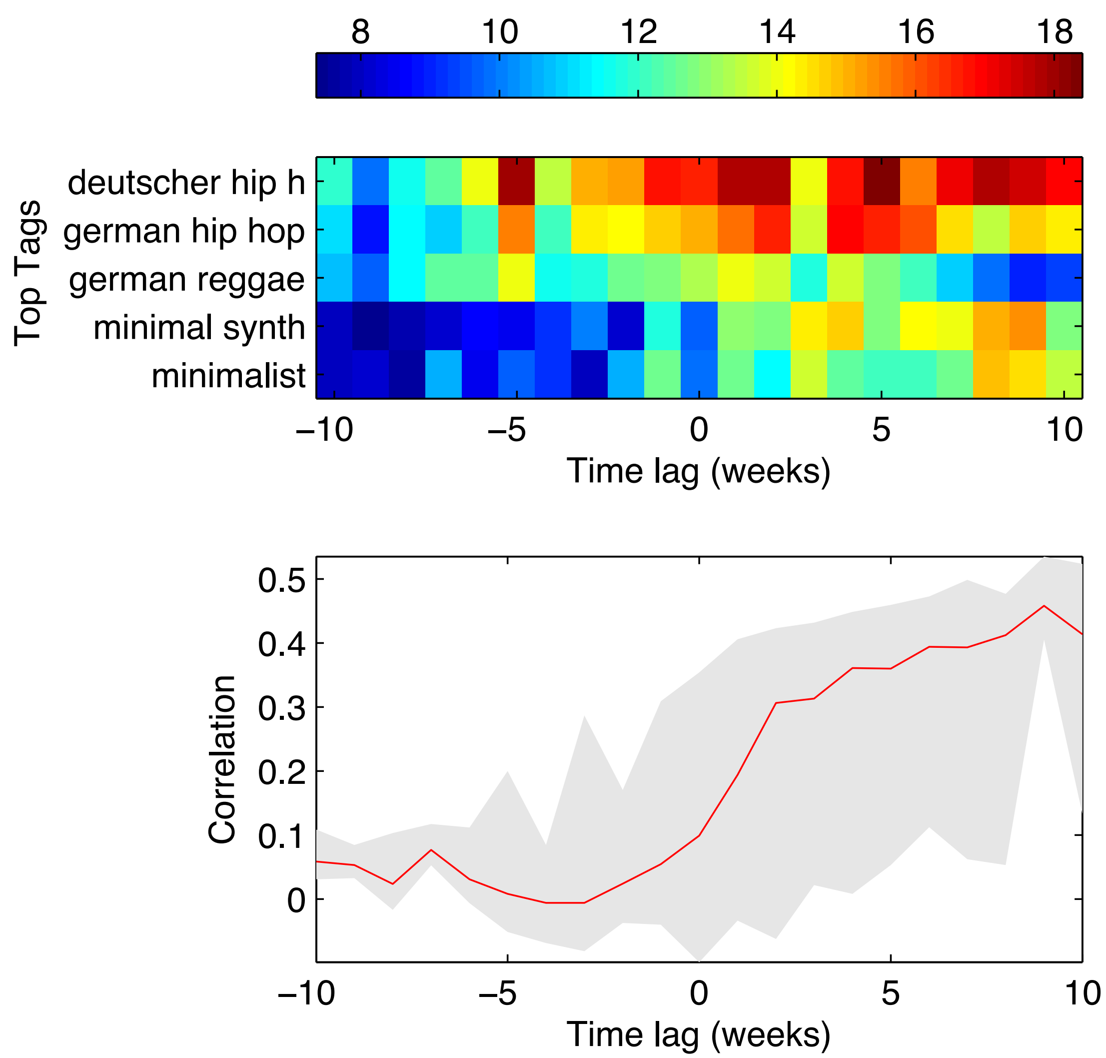
All Other Users

Canonical Correlogram

$$\rho(\tau) = \mathrm{Corr}\left(w_x(\tau)^\top X_\tau, w_y^\top Y\right)$$
$$= \frac{w_x(\tau)^\top X_\tau Y^\top w_y}{w_x(\tau)^\top X_\tau X_\tau^\top w_x(\tau) \cdot w_y^\top Y Y^\top w_y}$$
$$= \frac{\alpha^\top K_\tau K_Y \beta}{\alpha^\top K_\tau^2 \alpha \cdot \beta^\top K_Y^2 \beta}$$

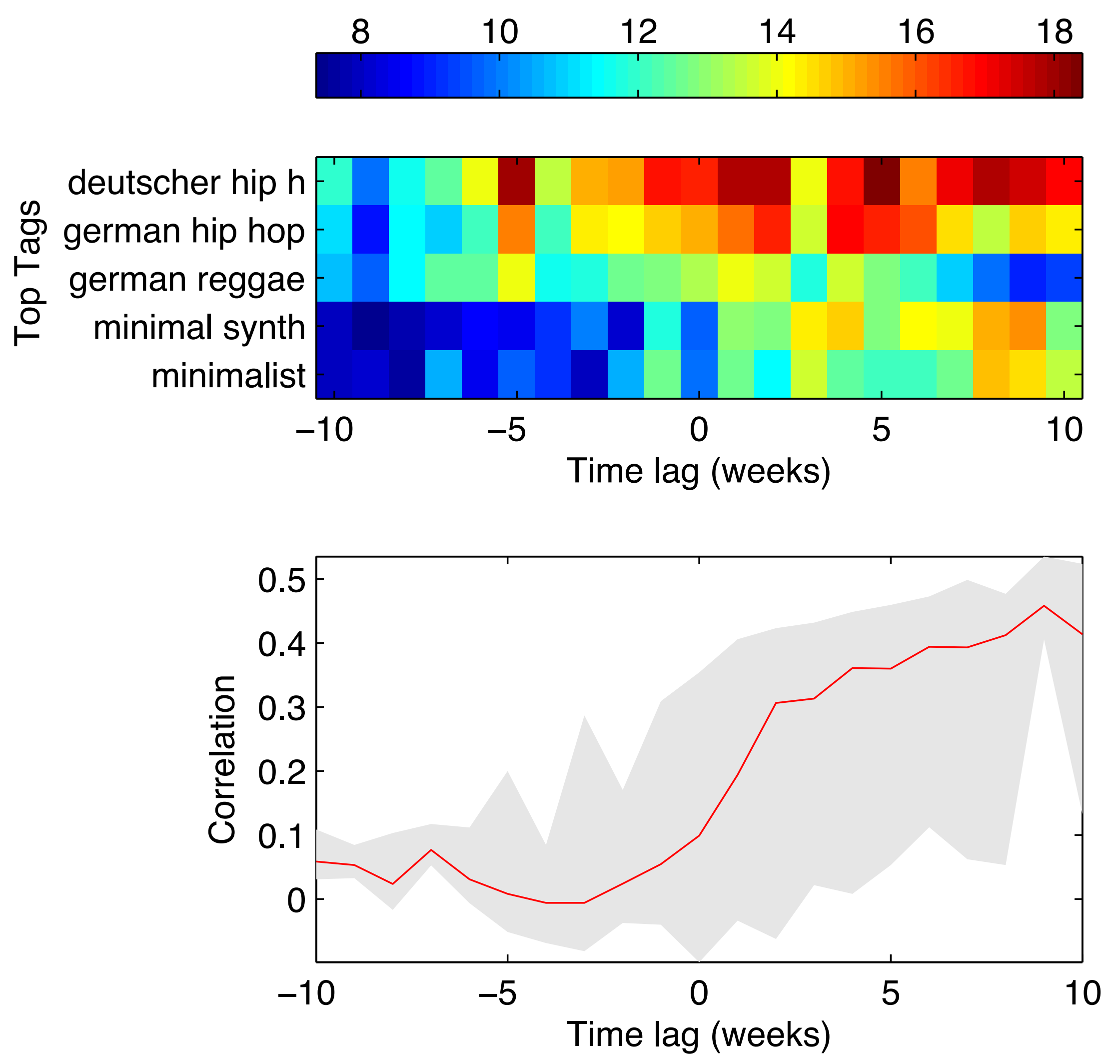
# Users and Trends on Last.fm

## Behind the Trend

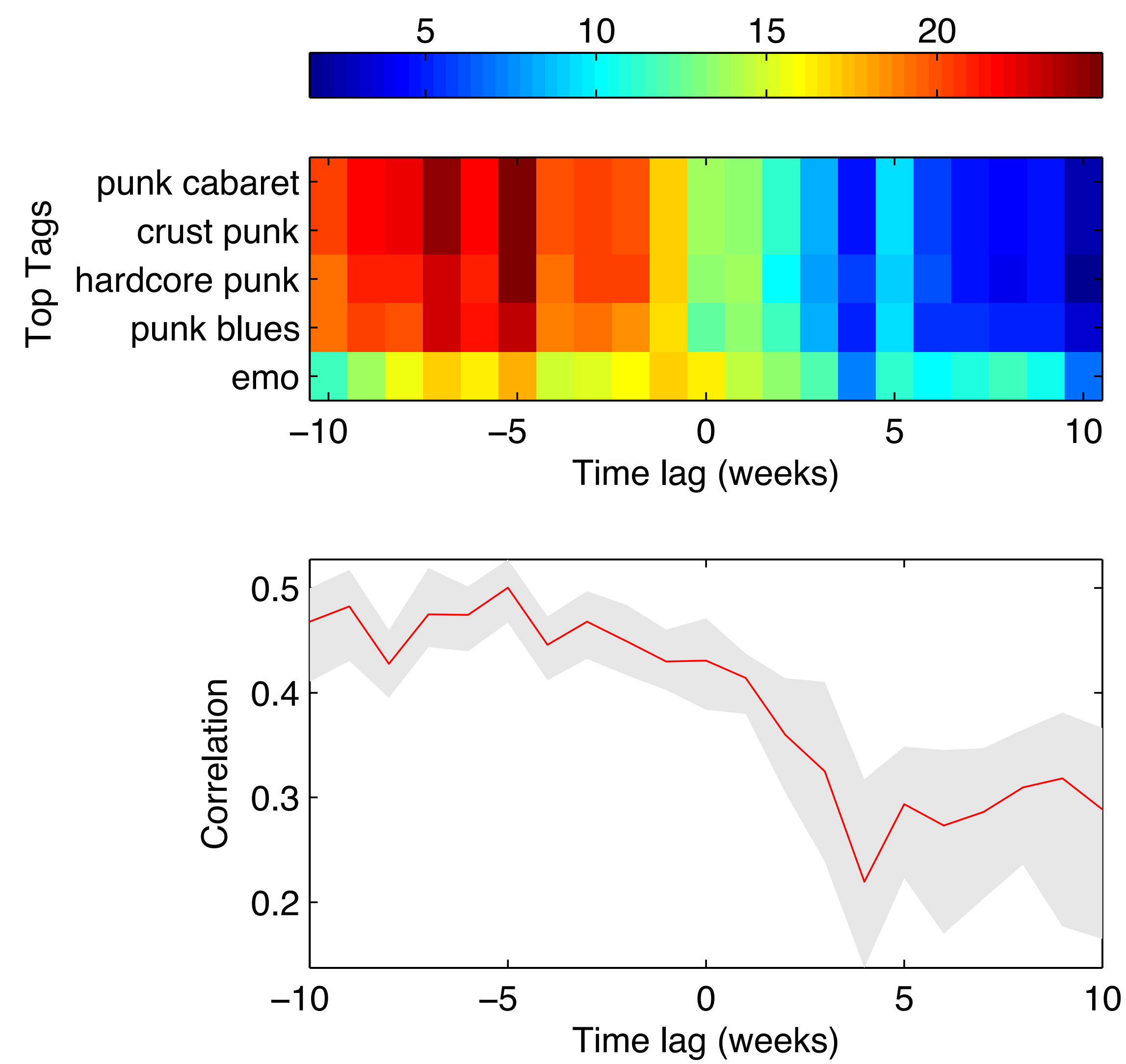


# Users and Trends on Last.fm

## Behind the Trend




## Ahead of Trend



# Summary

# Summary

Canonical Trend Analysis (CTA)

Canonical Trend Analysis (CTA)

Finds maximally correlated subspace of graph feature time series

# Summary

Canonical Trend Analysis (CTA)

- Finds maximally correlated subspace of graph feature time series
- Efficient computations via representer theorem

# Summary

Canonical Trend Analysis (CTA)

- Finds maximally correlated subspace of graph feature time series
- Efficient computations via representer theorem

CTA between news content and retweet location

# Summary

Canonical Trend Analysis (CTA)

- Finds maximally correlated subspace of graph feature time series
- Efficient computations via representer theorem

CTA between news content and retweet location

- Reveals ‘strongest’ topics and spatiotemporal tweet response

## Summary

Canonical Trend Analysis (CTA)

- Finds maximally correlated subspace of graph feature time series
- Efficient computations via representer theorem

CTA between news content and retweet location

- Reveals ‘strongest’ topics and spatiotemporal tweet response

CTA between users on Last.fm

# Summary

Canonical Trend Analysis (CTA)

- Finds maximally correlated subspace of graph feature time series
- Efficient computations via representer theorem

CTA between news content and retweet location

- Reveals ‘strongest’ topics and spatiotemporal tweet response

CTA between users on Last.fm

- Finds users ahead and behind musical trends

# Future Work

# Future Work

Sparse, non-negative canonical directions

# Future Work

Sparse, non-negative canonical directions

Other features than BoW

# Future Work

Sparse, non-negative canonical directions

Other features than BoW

Online optimization

# Future Work

Sparse, non-negative canonical directions

Other features than BoW

Online optimization

What about Nonstationarities?

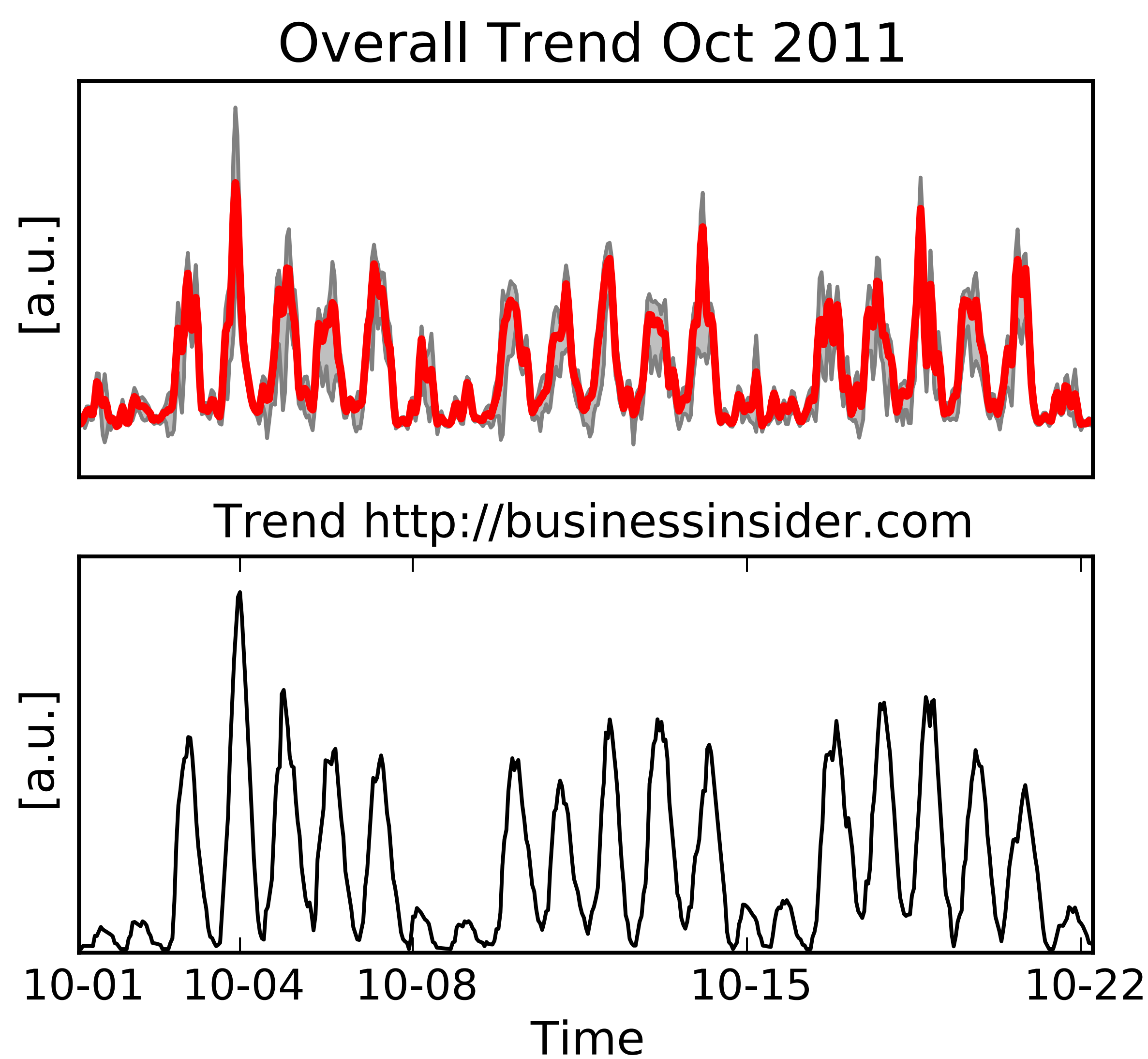
# Detecting ‘Trendsetting’ News Websites

Real Data Example:

BoW Features from 96 Technology News Feeds in October 2011

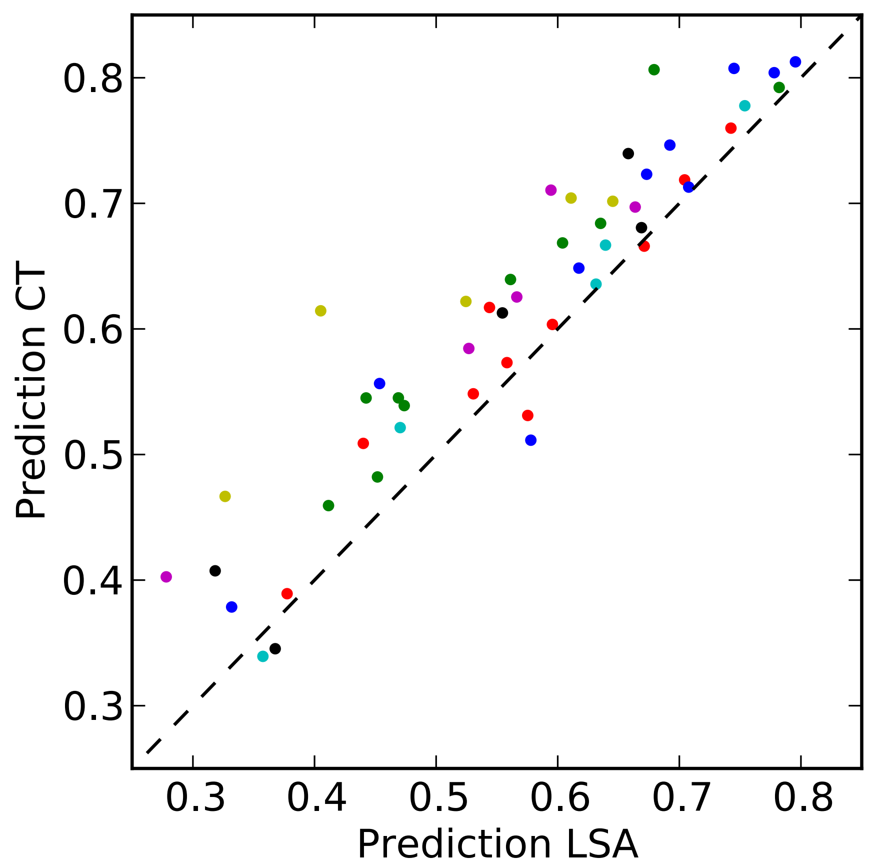
$$y_f(t)$$

$$\hat{y}_f(t)$$



# Comparison Canonical Trend Analysis and LSA

Canonical trend analysis **between** $X_f$ and $Y_f$
vs. LSA on $X_f$ and $Y_f$ **separately**




Canonical Topics predict overall topics
better than Latent Semantic Indexing