Master's Thesis Defense

Analysis and Forecasting of Trending Topics in Online Media

Tim Althoff
German Research Center for Artificial Intelligence (DFKI)

Motivation

Large-scale Record of Human Behavior

- Primary method for communication by college students in the U.S.

- 340 million tweets each day by 500 million users

- One billion search requests and about twenty petabytes of user-generated data each day (2009)

Motivation

- People use web for news, information, and communication
- Online activity becomes indicative of the interests of the global population
- After Boston, e.g. Twitter has become a trusted channel for authorities

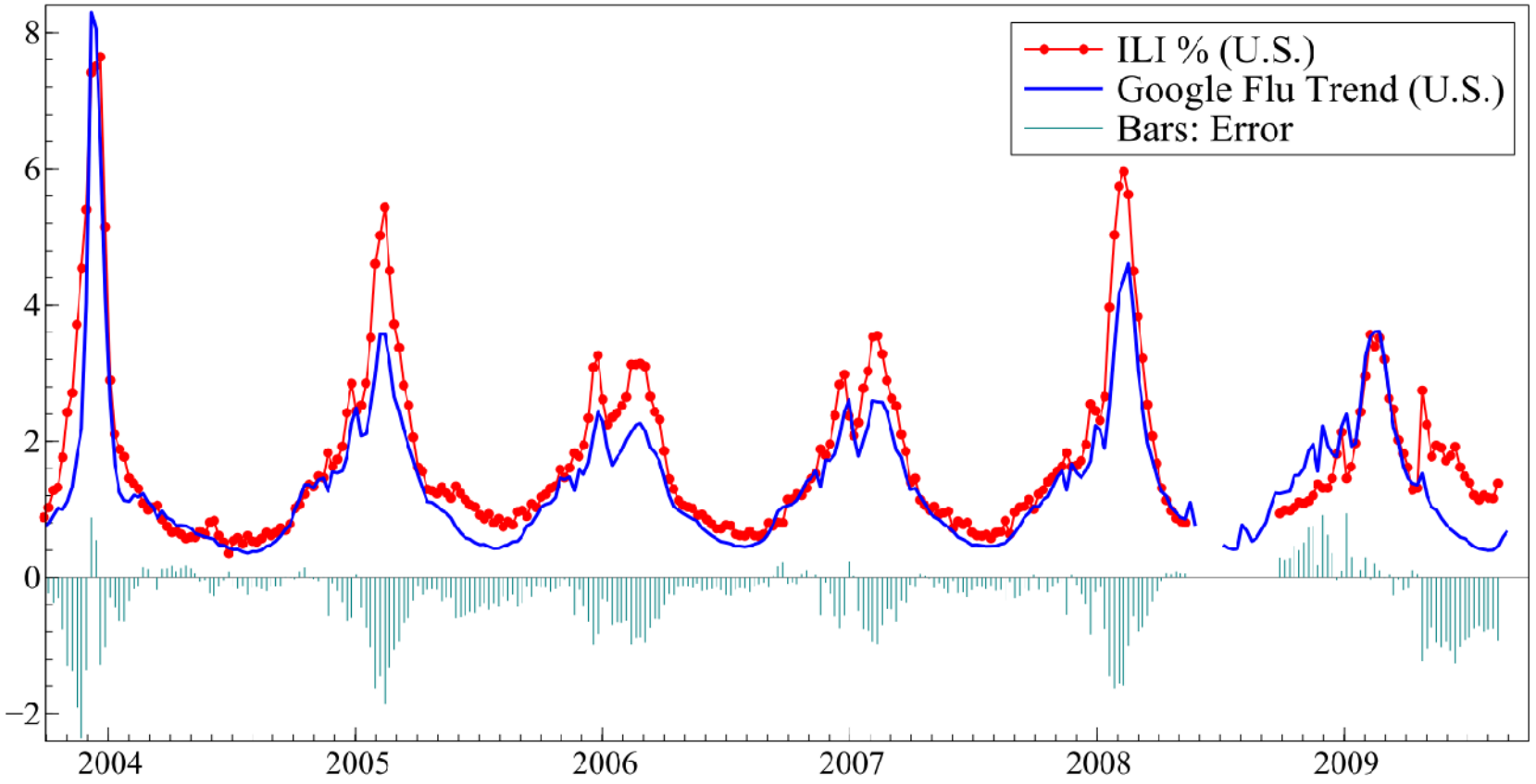
› **„Mirror of Society“**

[Steven Van Belleghem, 2012]

Use Case of Online Activity

Tracking Flu Infections

flu remedy



[Ginsberg et al. Detecting influenza epidemics using search engine query data. Nature. 2008.]

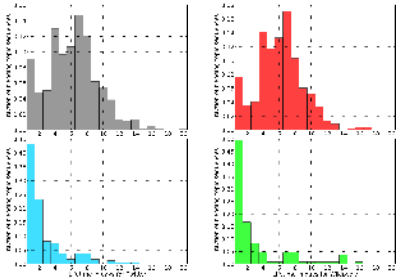
Shortcomings of Related Work

1. Focus on very specific use cases (e.g. flu)
2. Nowcasting instead of forecasting
3. Manual correlation analysis instead of fully automatic forecasting
4. No comparison/study of different online & social media channels wrt. emerging trends

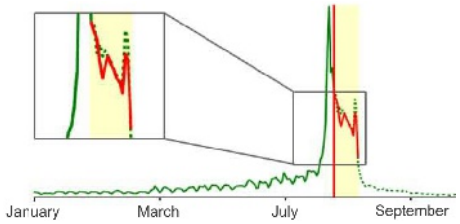
[Choi and Varian. Predicting initial claims for unemployment benefits. Google, Inc. 2009.]
[Choi and Varian. Predicting the Present with Google Trends. Economic Record, 2012.]
[Ginsberg et al. Detecting influenza epidemics using search engine query data. Nature. 2008.]
[Cooper et al. Cancer internet search activity on a major search engine. Journal of medical Internet research, 2005.]
[Goel et al. Predicting consumer behavior with web search. National Academy of Sciences, 2010.]
[Jin et al. The wisdom of social multimedia: using Flickr for prediction and forecast. ACM MM 2010.]
[Adar et al. Why we search: visualizing and predicting user behavior. WWW, 2007.]

Goals of Master's Thesis

1. Multi-channel analysis of trending topics



2. Fully automatic forecasting of trending topics

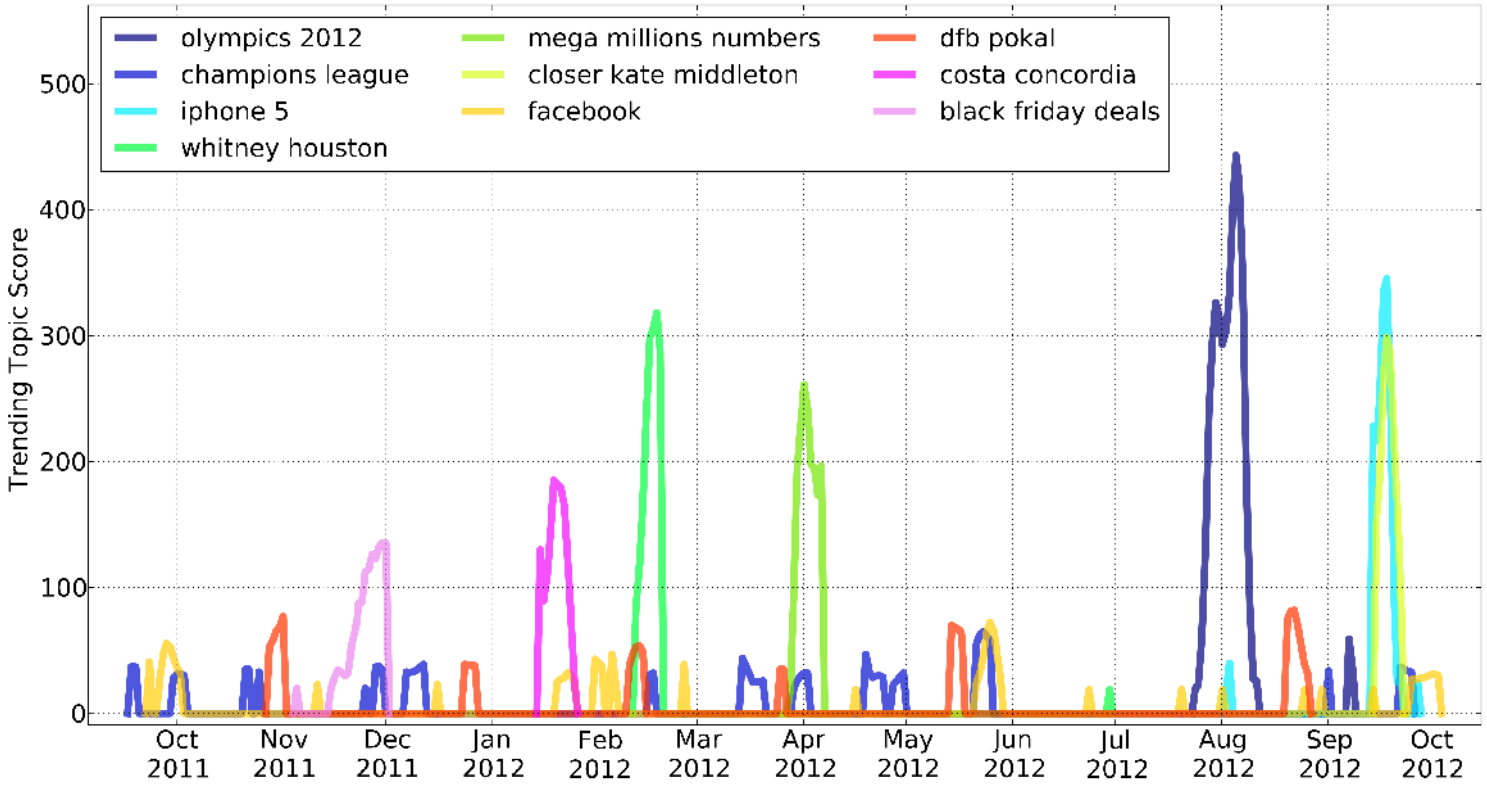


Trending Topics

Trending Topics

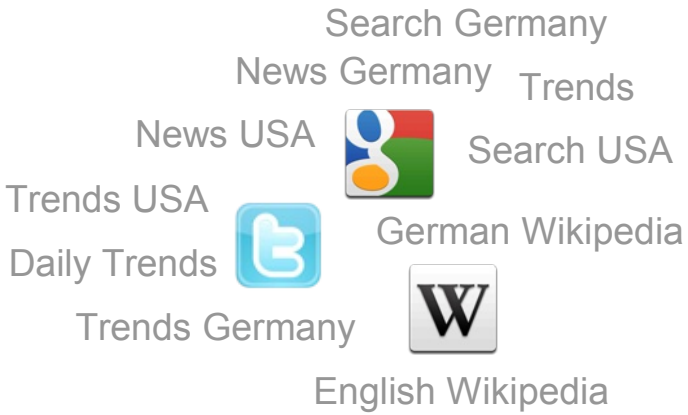
A **trending topic**

- is a subject matter of discussion agreed on by a group of people
- experiences a sudden spike in user interest or engagement



Trending Topics Discovery

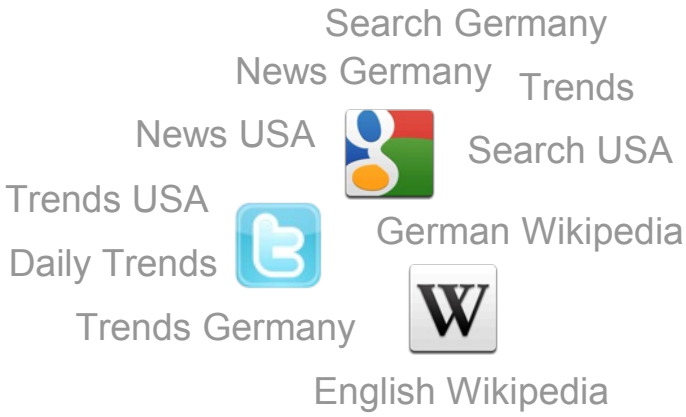
- Daily crawling of 10 sources
- Capturing
  - Communication needs
  - Search pattern
  - Information demand



[Damian Borth, Adrian Ulges, Thomas Breuel. Dynamic Vocabularies for Web-based Concept Detection by Trend Discovery. ACM Multimedia, 2012]

Trending Topics Discovery

- Daily crawling of 10 sources
- Capturing
  - Communication needs
  - Search pattern
  - Information demand

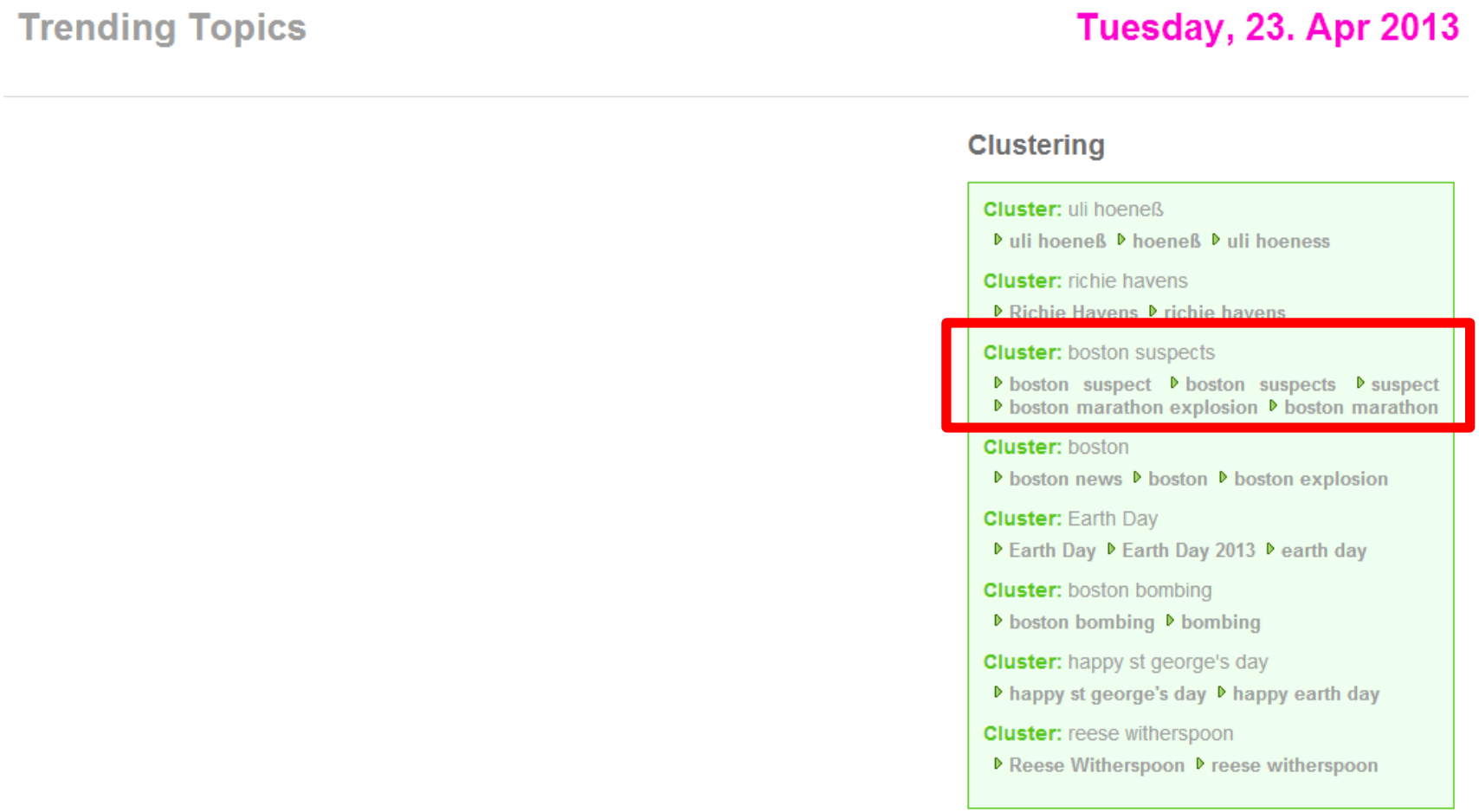


Dataset

- Observation period
  - Sept. 2011 – Sept. 2012
- 40,000 items analyzed
- ~3000 trending topics discovered

[Damian Borth, Adrian Ulges, Thomas Breuel. Dynamic Vocabularies for Web-based Concept Detection by Trend Discovery. ACM Multimedia, 2012]

Trending Topics Discovery

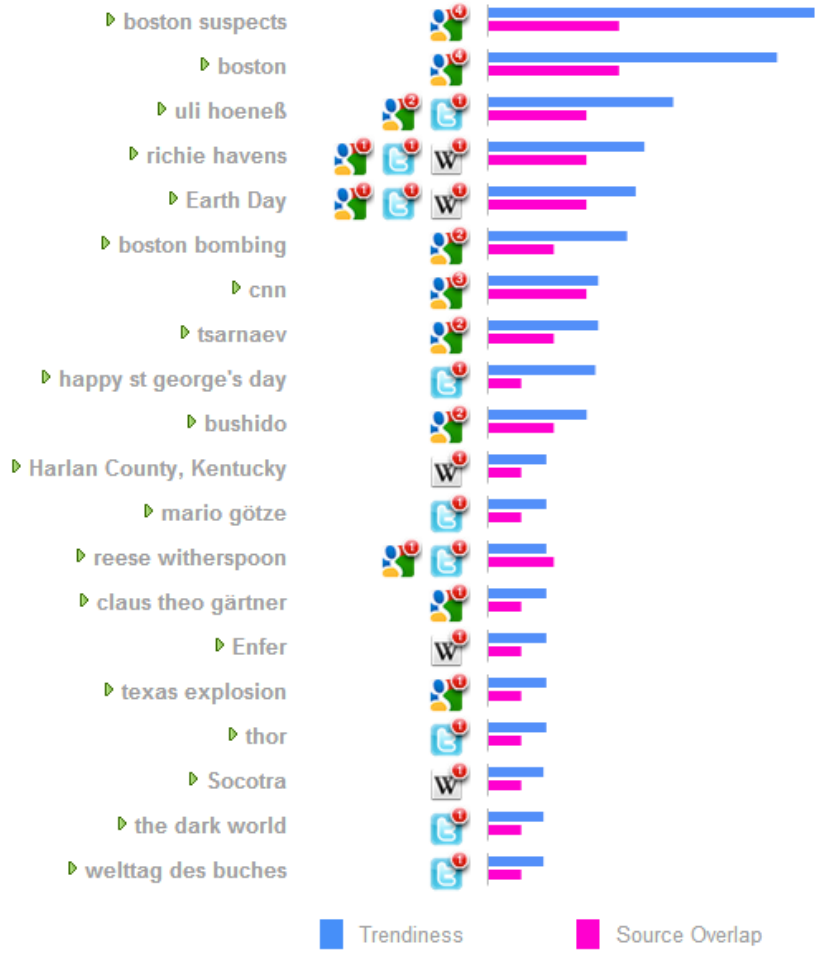


Trending Topics Discovery

Trending Topics

Tuesday, 23. Apr 2013

Top 20 Trends



Clustering

Cluster: uli hoeneß
- uli hoeneß
- hoeneß
- uli hoeness

Cluster: richie havens
- Richie Havens
- richie havens

Cluster: boston suspects
- boston suspect
- boston suspects
- suspect
- boston marathon explosion
- boston marathon

Cluster: boston
- boston news
- boston
- boston explosion

Cluster: Earth Day
- Earth Day
- Earth Day 2013
- earth day

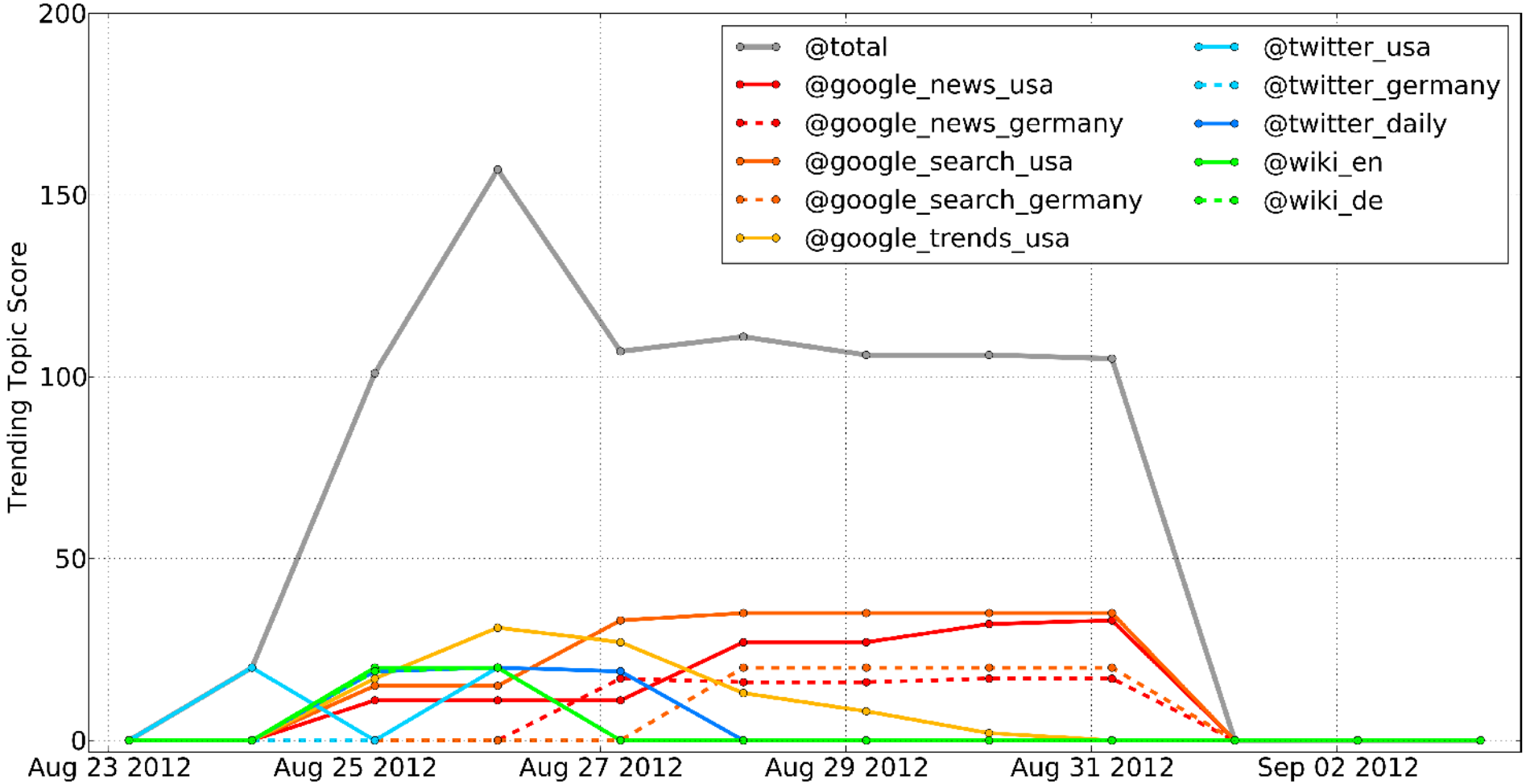
Cluster: boston bombing
- boston bombing
- bombing

Cluster: happy st george's day
- happy st george's day
- happy earth day

Cluster: reese witherspoon
- Reese Witherspoon
- reese witherspoon

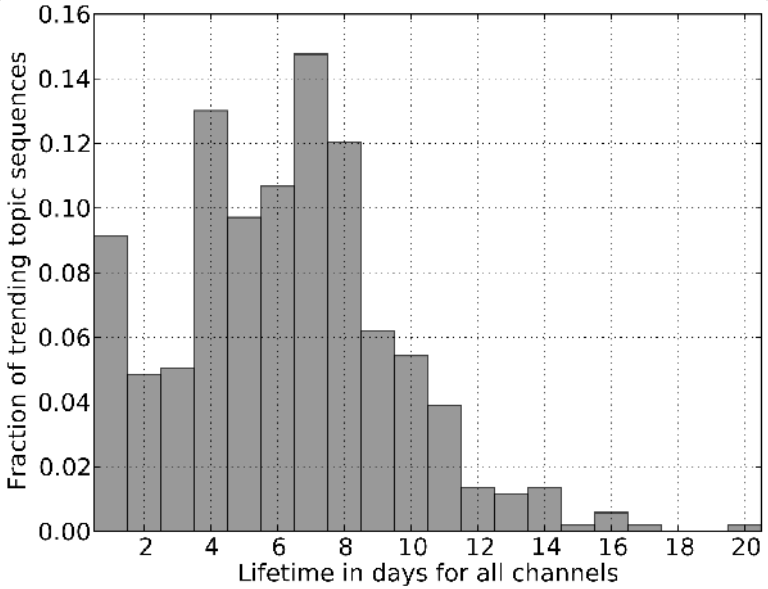
Multi-Channel Analysis

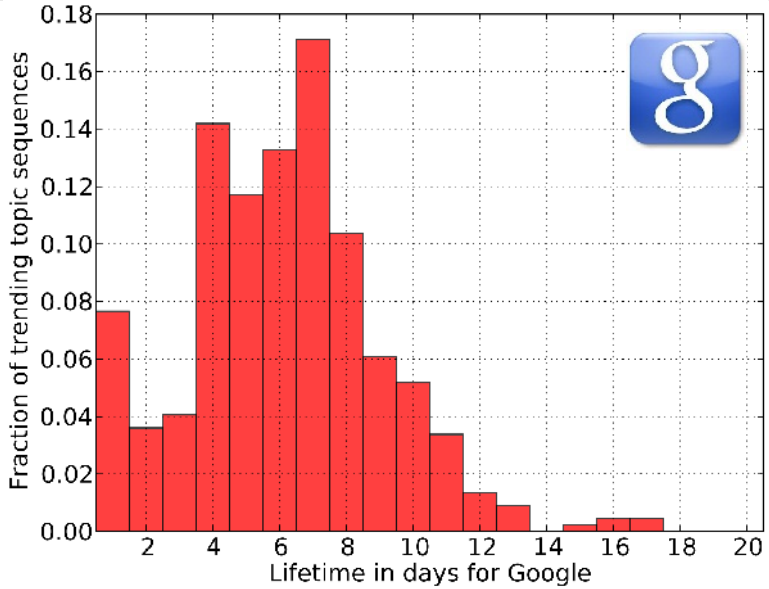
Example

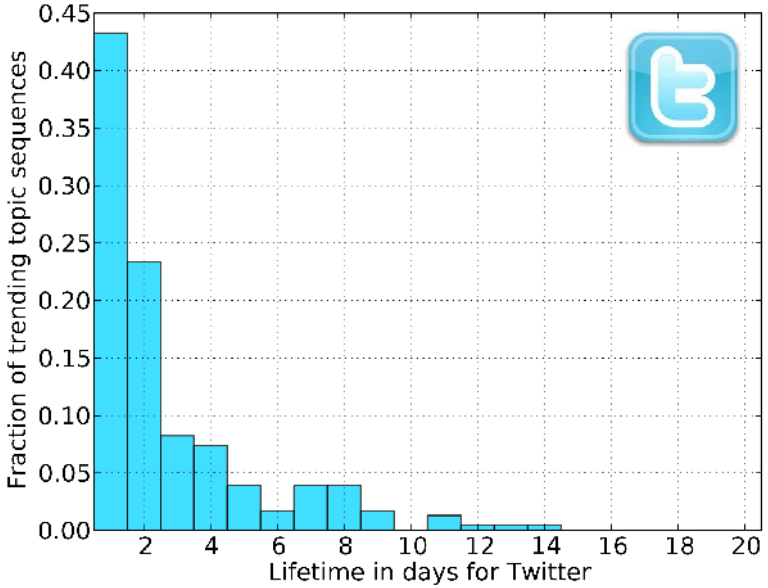


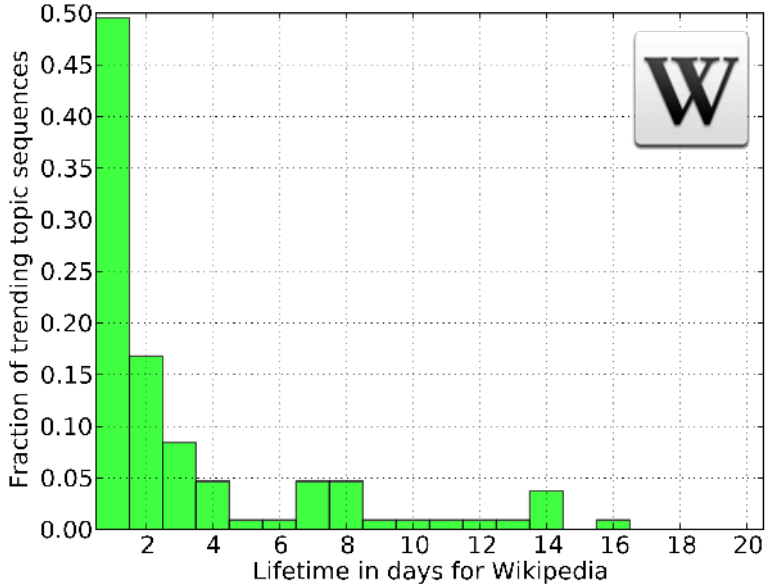
Neil Armstrong died on August 25, 2012.

Lifetime Analysis

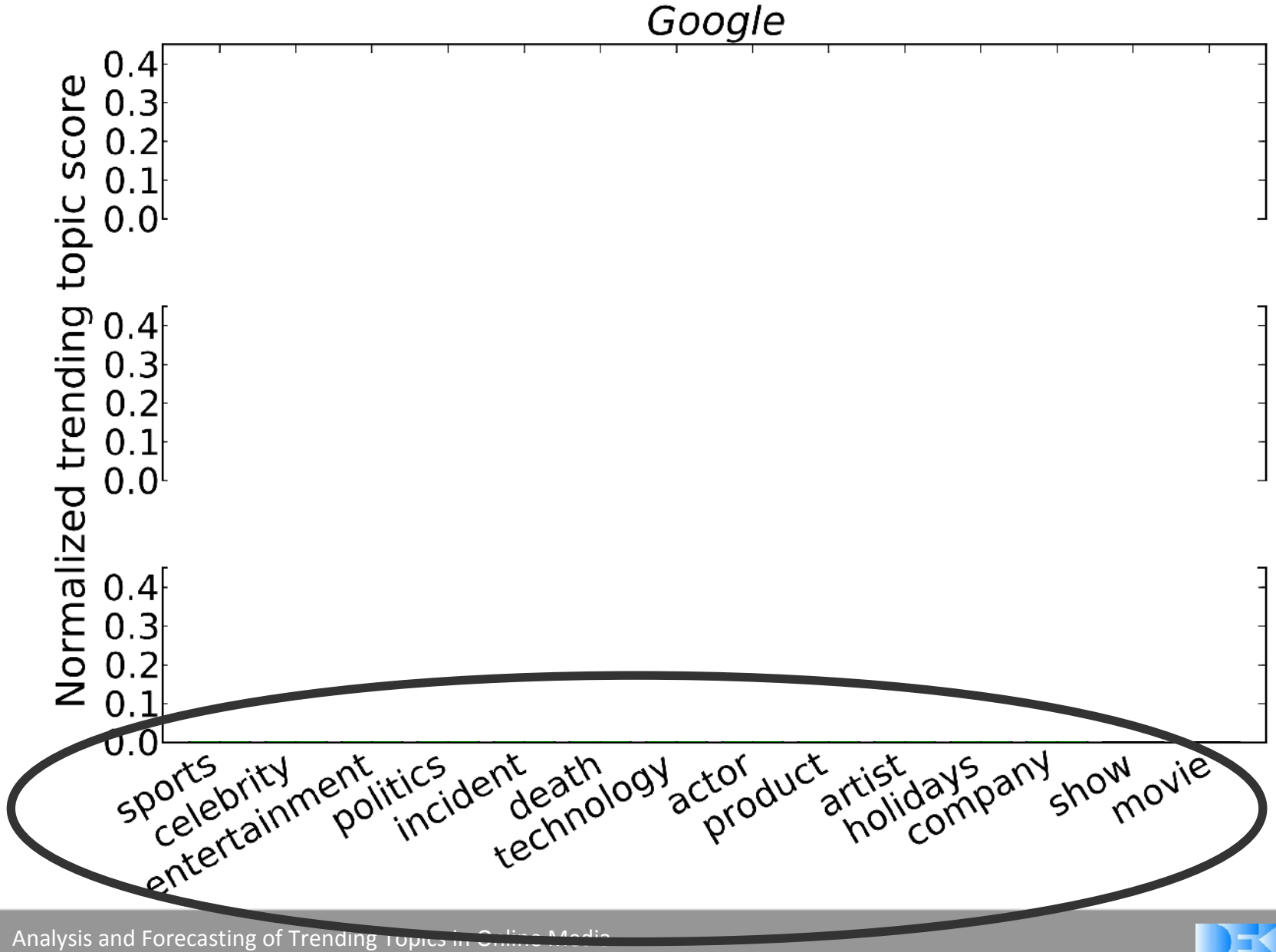




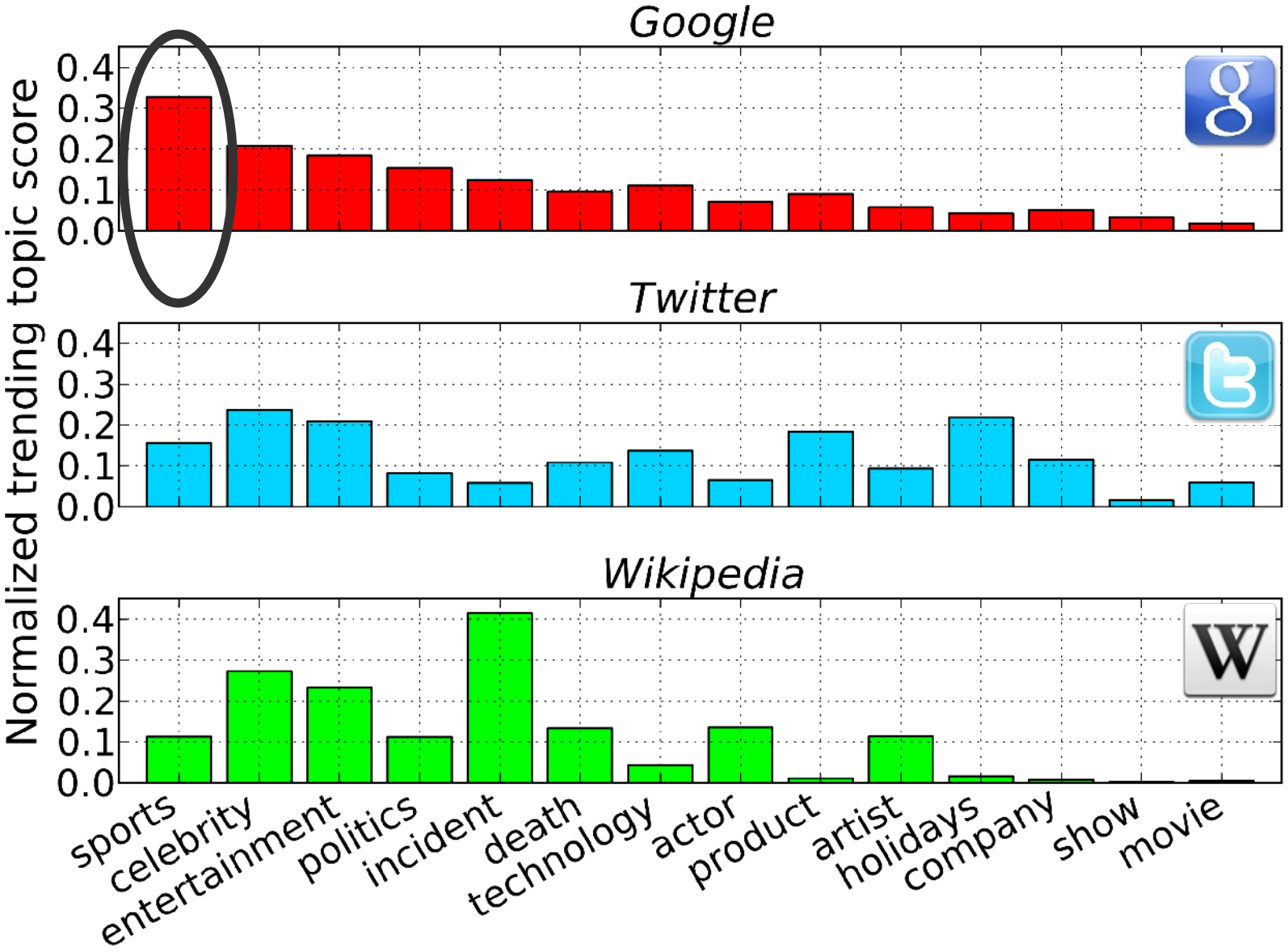




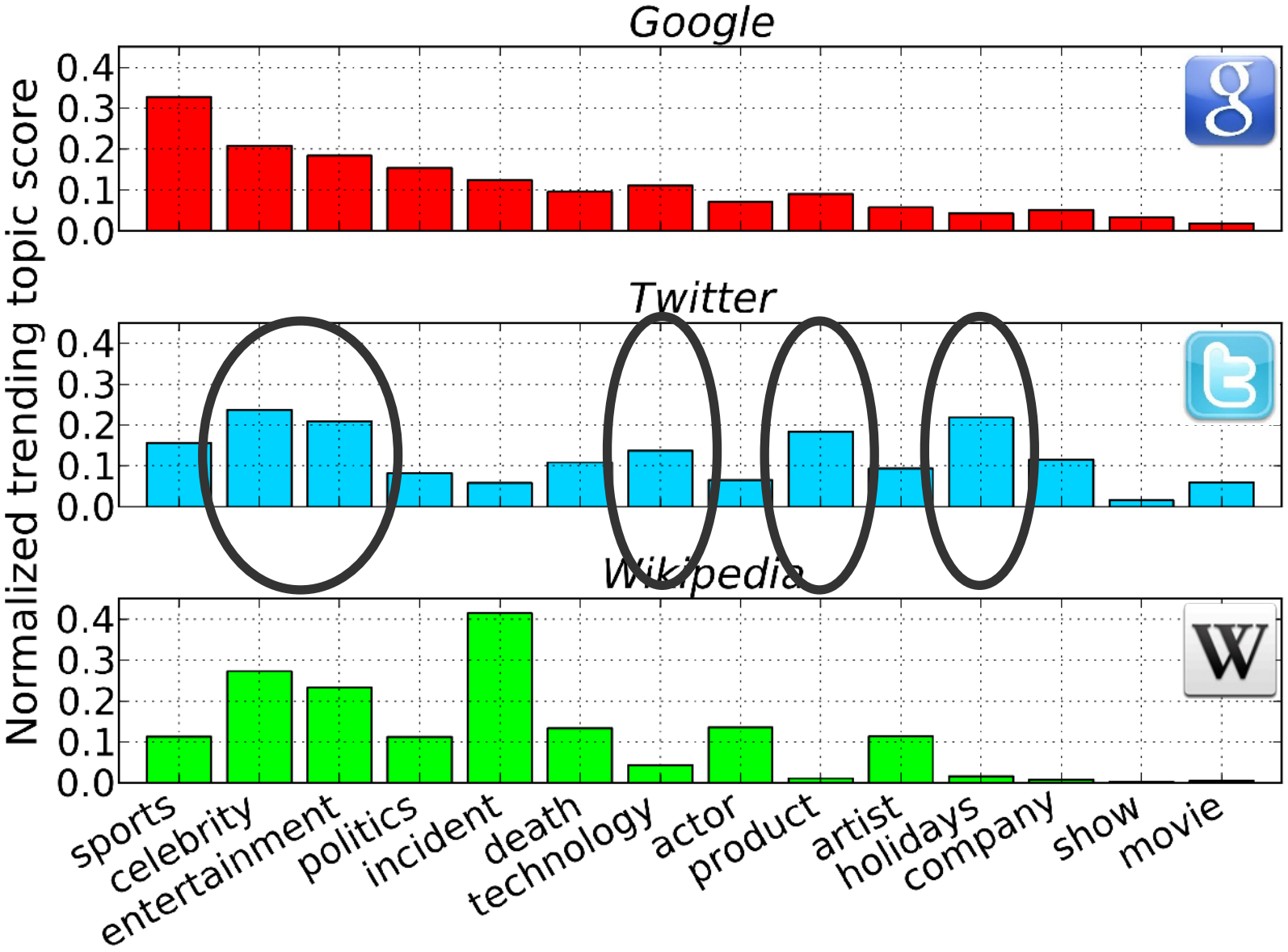
Topic Category Analysis



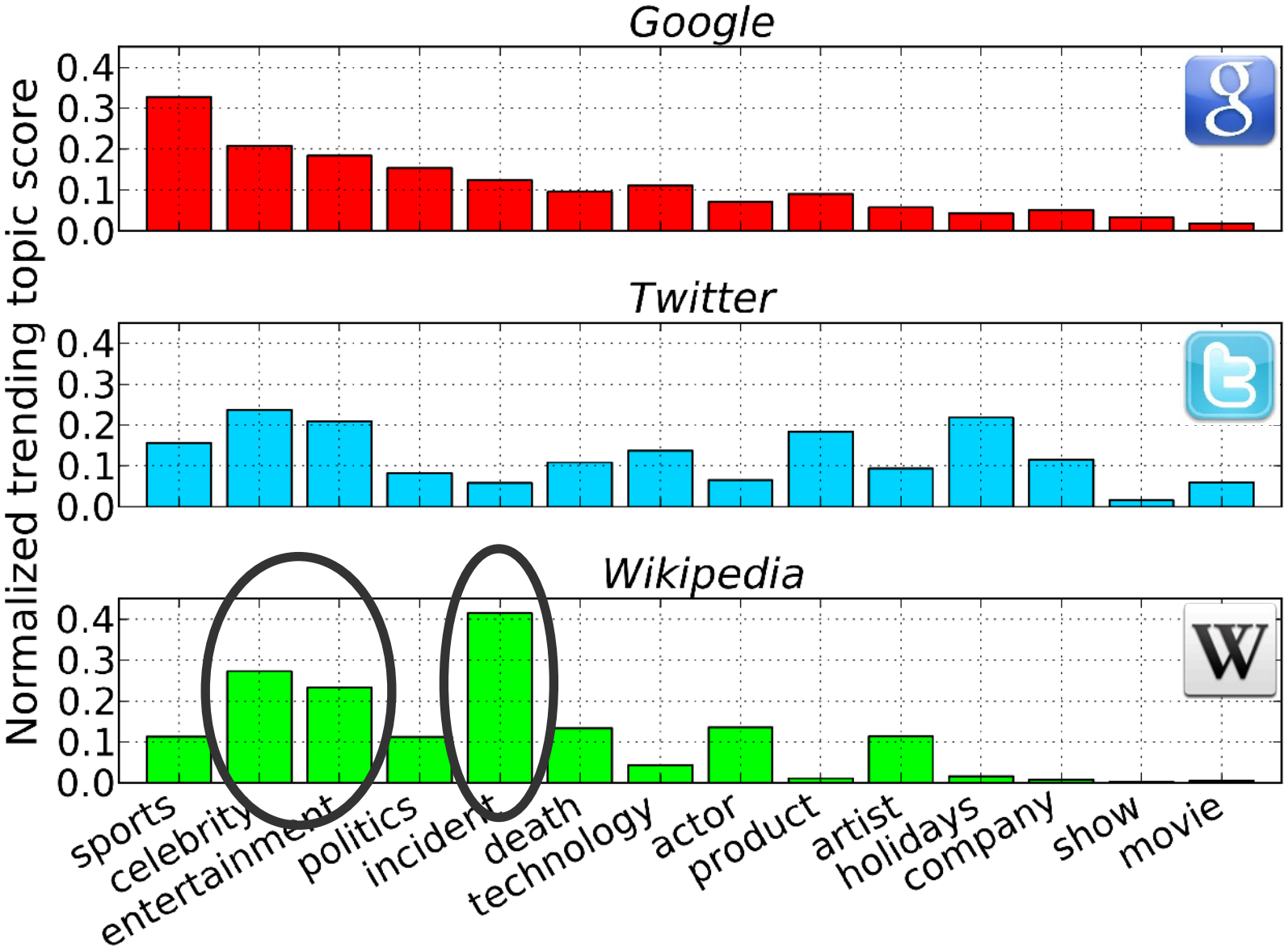
Topic Category Analysis



Topic Category Analysis

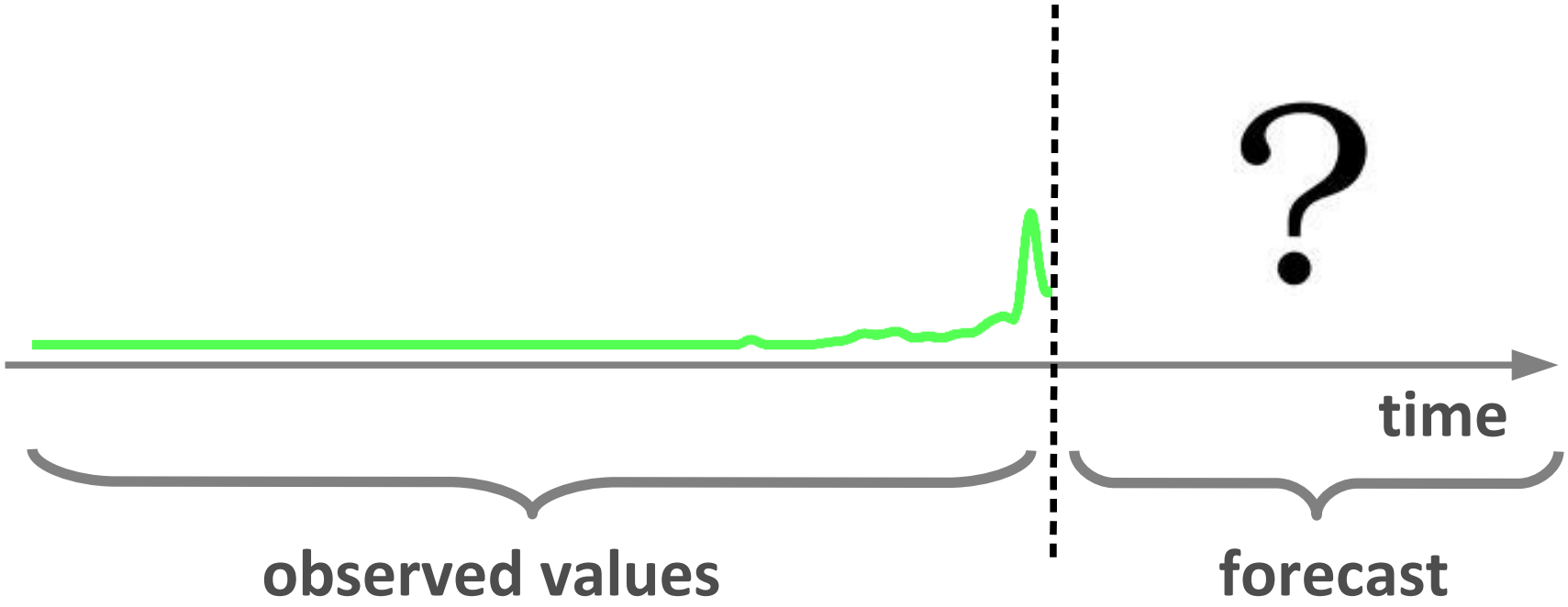


Topic Category Analysis



Forecasting

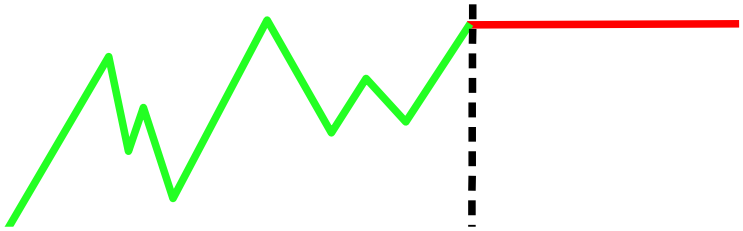
Forecasting



Basic Forecasting Techniques

Naive

$$X_t = X_{t-1}$$



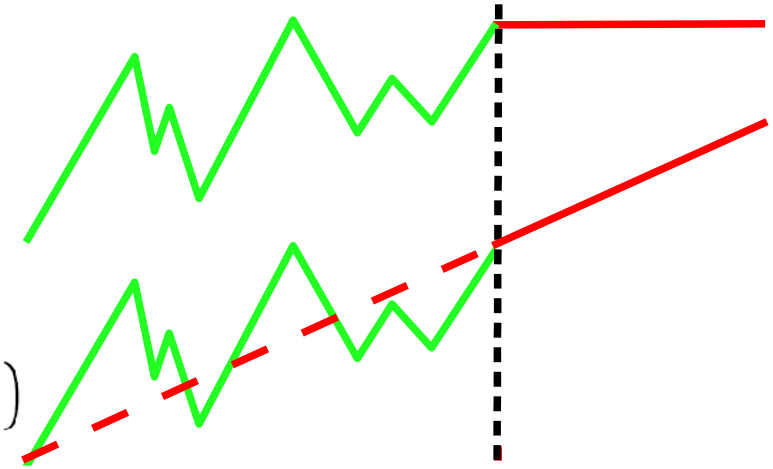
Basic Forecasting Techniques

Naive

$$X_t = X_{t-1}$$

Linear Trend

$$X_t = X_{t_0} + \frac{(X_{t_0} - X_{t_0-d})}{d} \cdot (t - t_0)$$



Basic Forecasting Techniques

Naive

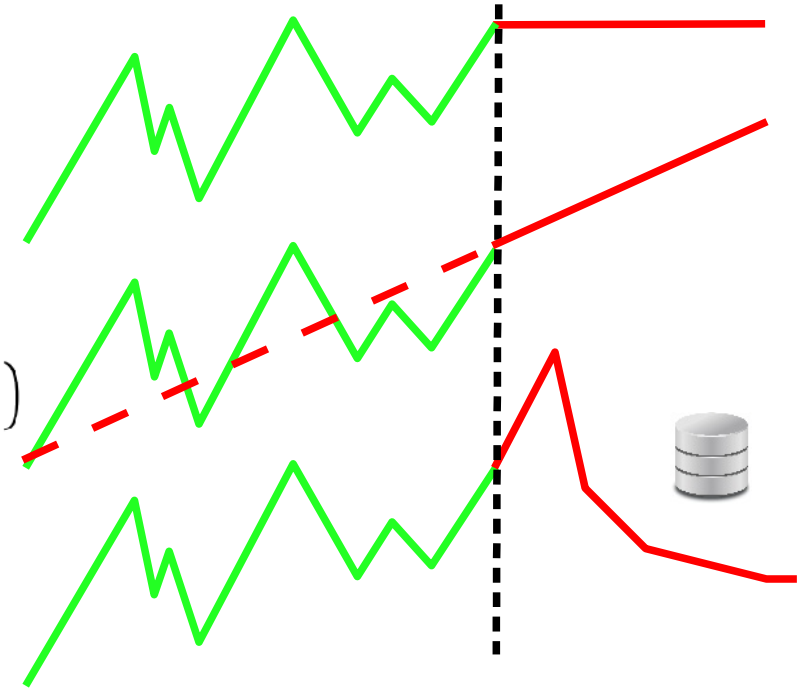
$$X_t = X_{t-1}$$

Linear Trend

$$X_t = X_{t_0} + \frac{(X_{t_0} - X_{t_0-d})}{d} \cdot (t - t_0)$$

Average/Median

$$X_t = \text{average}\left(X_t^1, \ldots, X_t^n\right)$$



Basic Forecasting Techniques

Naive

$$X_t = X_{t-1}$$

Linear Trend

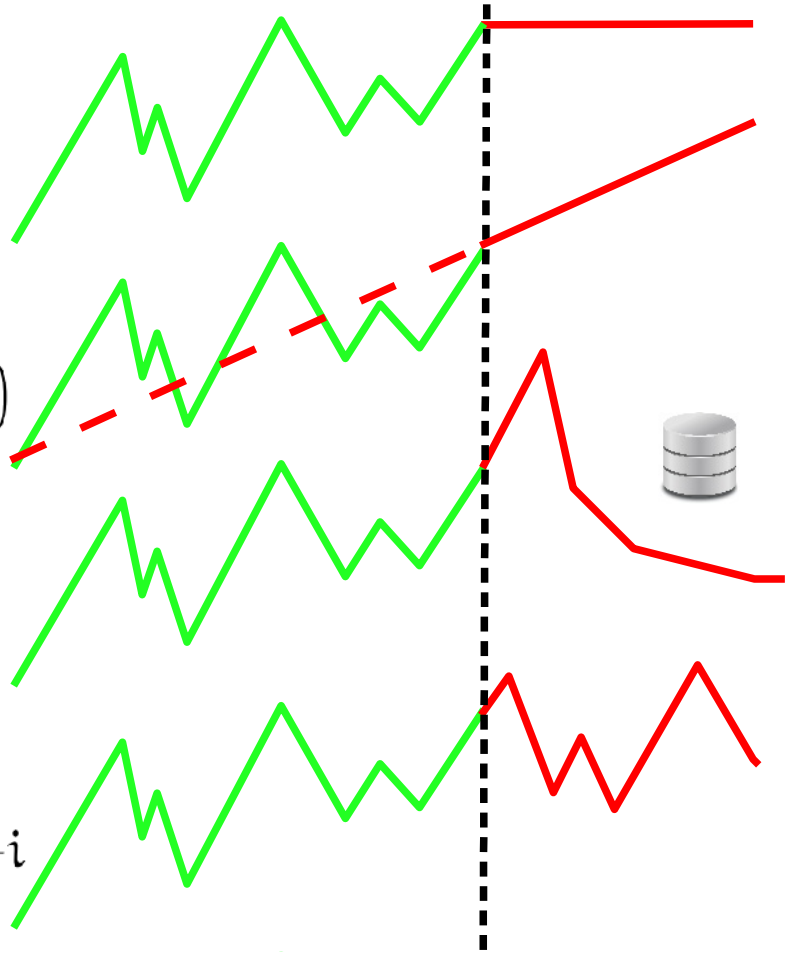
$$X_t = X_{t_0} + \frac{(X_{t_0} - X_{t_0-d})}{d} \cdot (t - t_0)$$

Average/Median

$$X_t = \text{average}\left(X_t^1, \ldots, X_t^n\right)$$

ARMA

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$



Basic Forecasting Techniques

Naive

$$X_t = X_{t-1}$$

Linear Trend

$$X_t = X_{t_0} + \frac{(X_{t_0} - X_{t_0-d})}{d} \cdot (t - t_0)$$

Average/Median

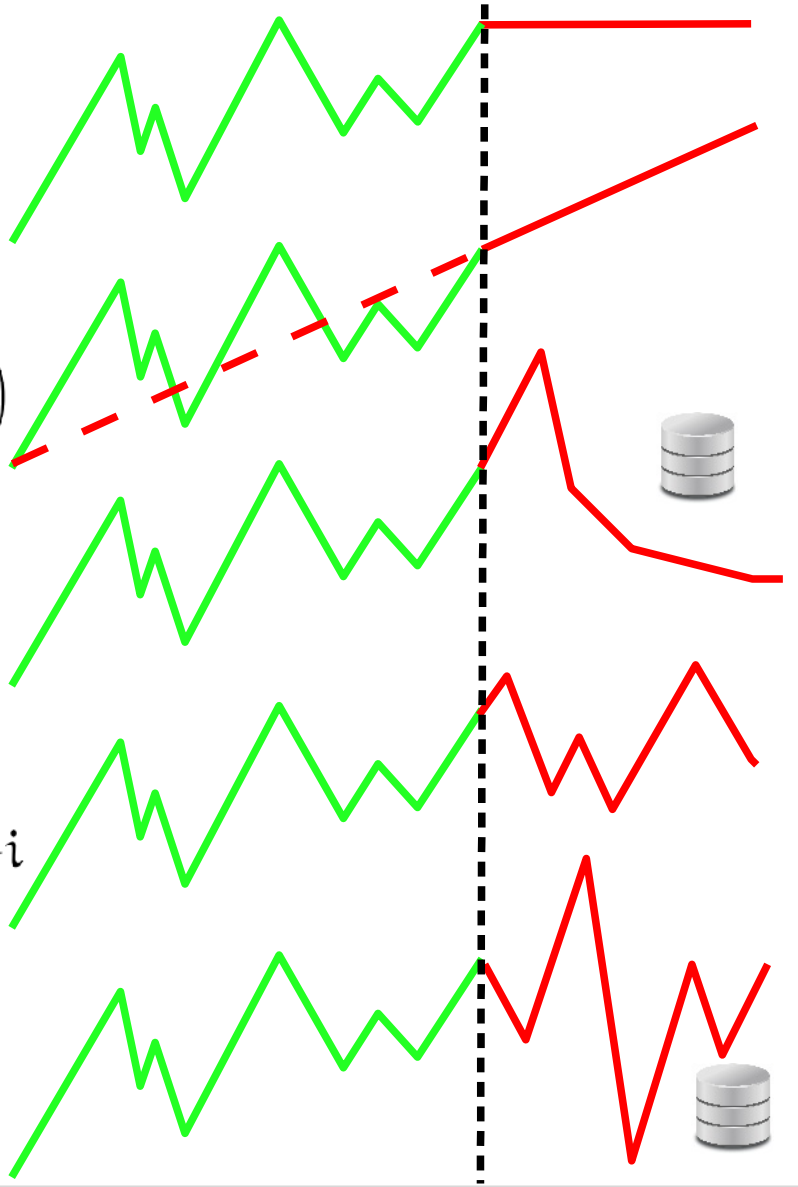
$$X_t = \text{average}\left(X_t^1, \ldots, X_t^n\right)$$

ARMA

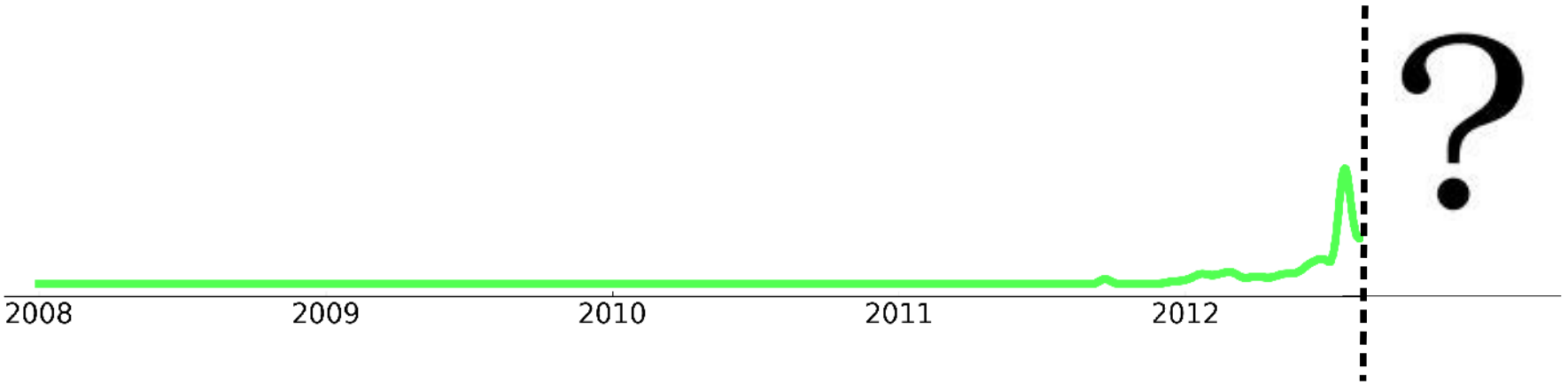
$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Nearest Neighbor

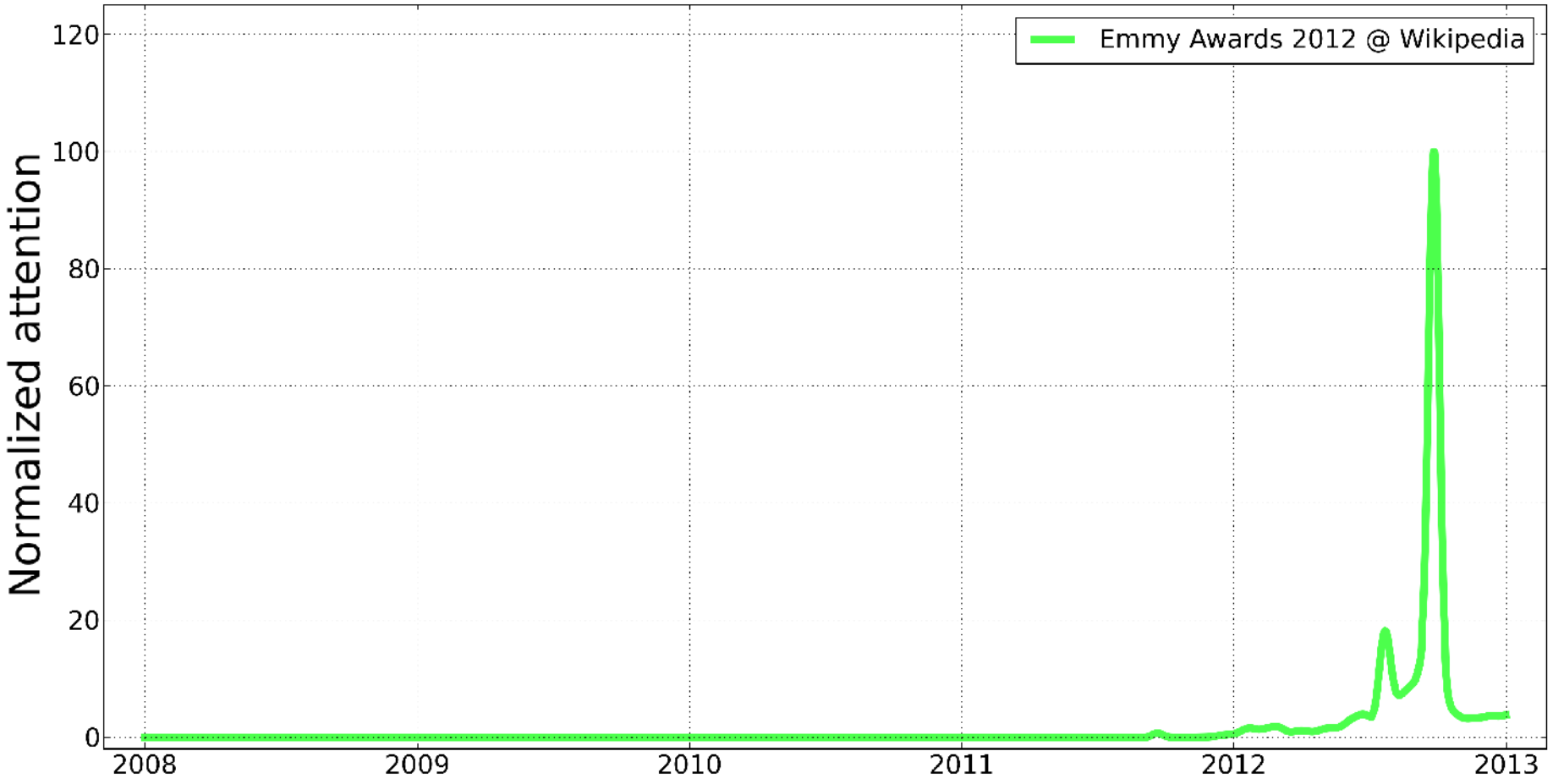
$$X_t = \text{operator}\left(X_{t_1}^1, \ldots, X_{t_k}^k\right)$$



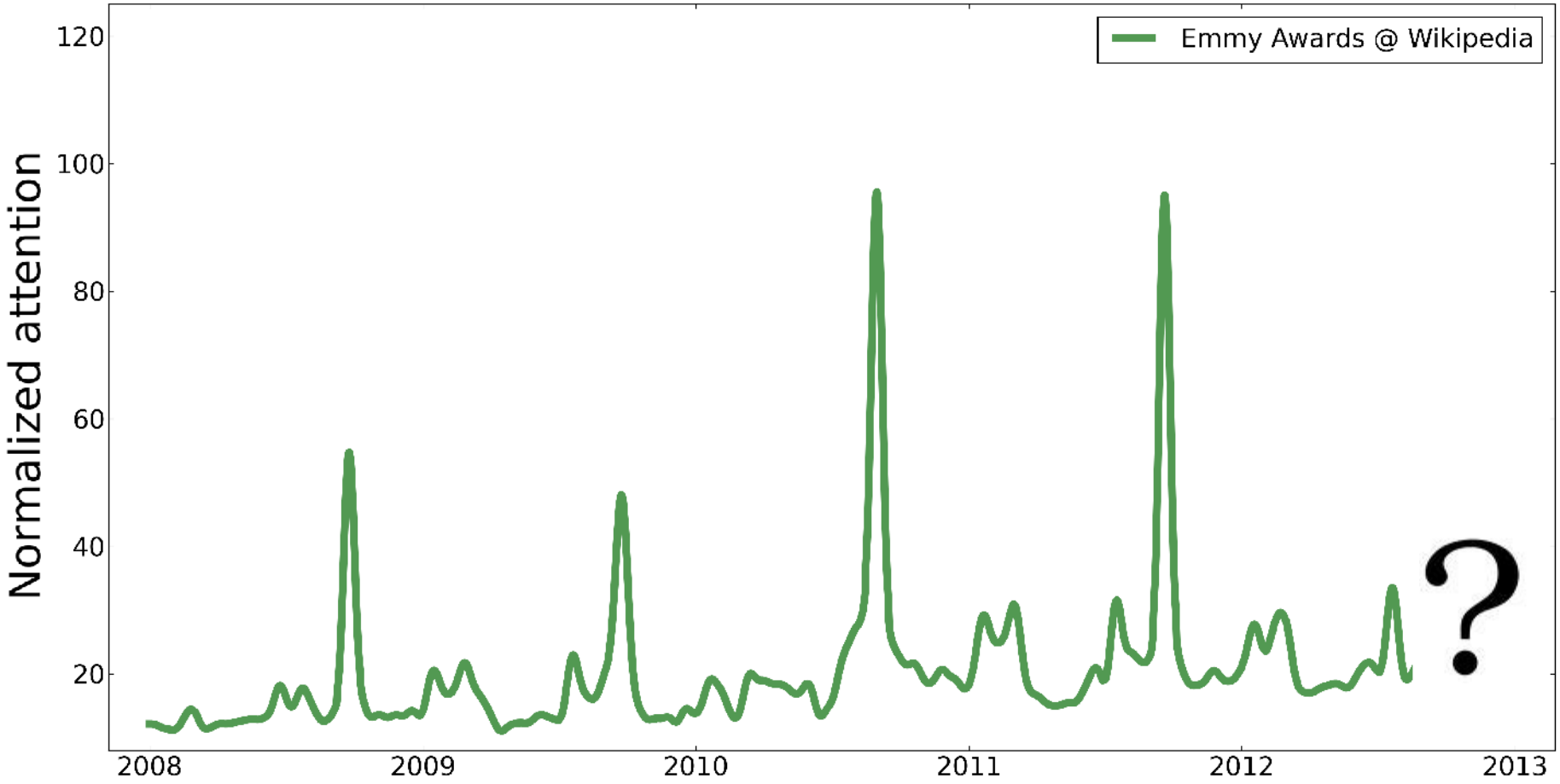
Example Signal



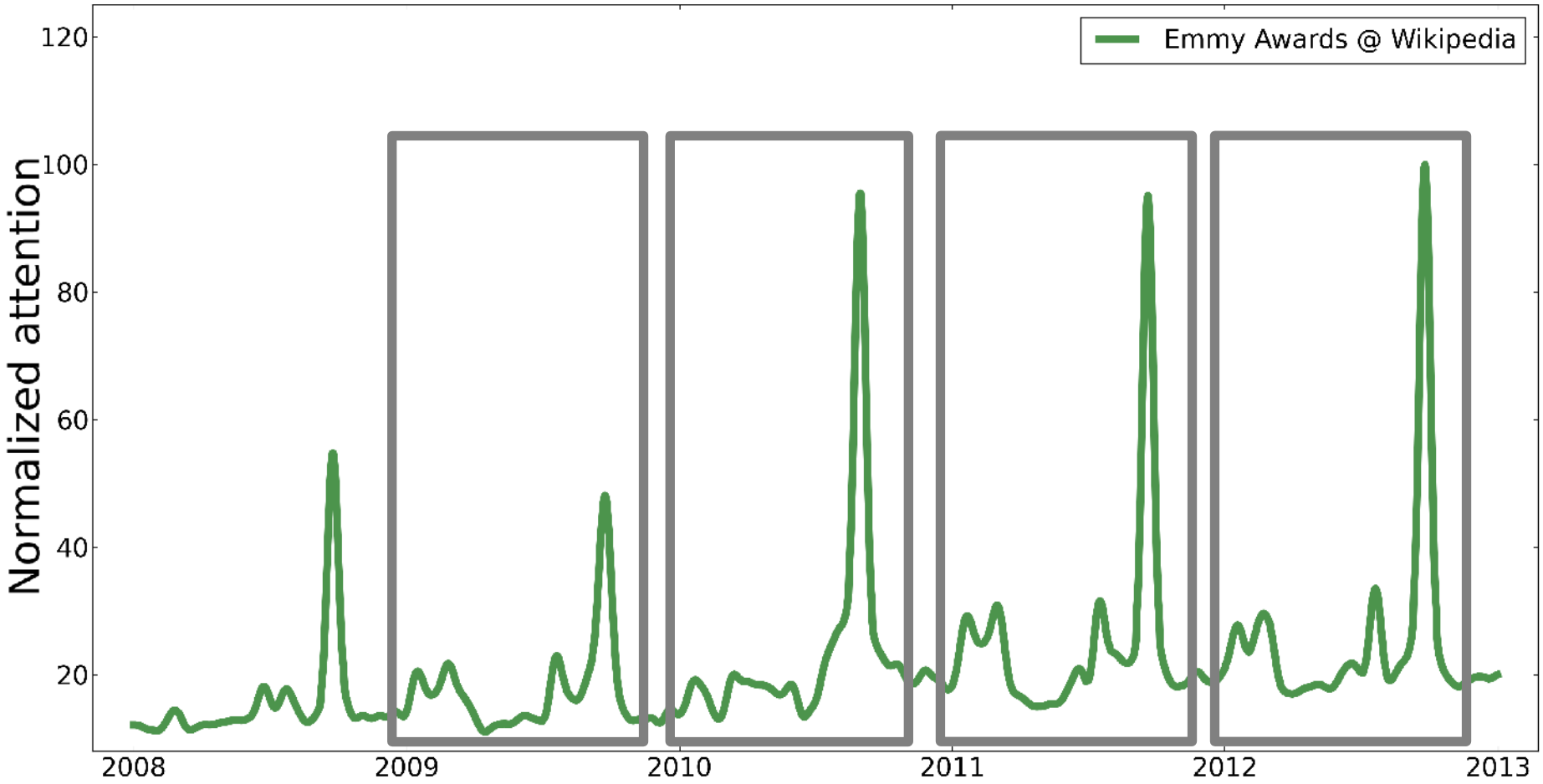
Forecasting Can Be Very Challenging



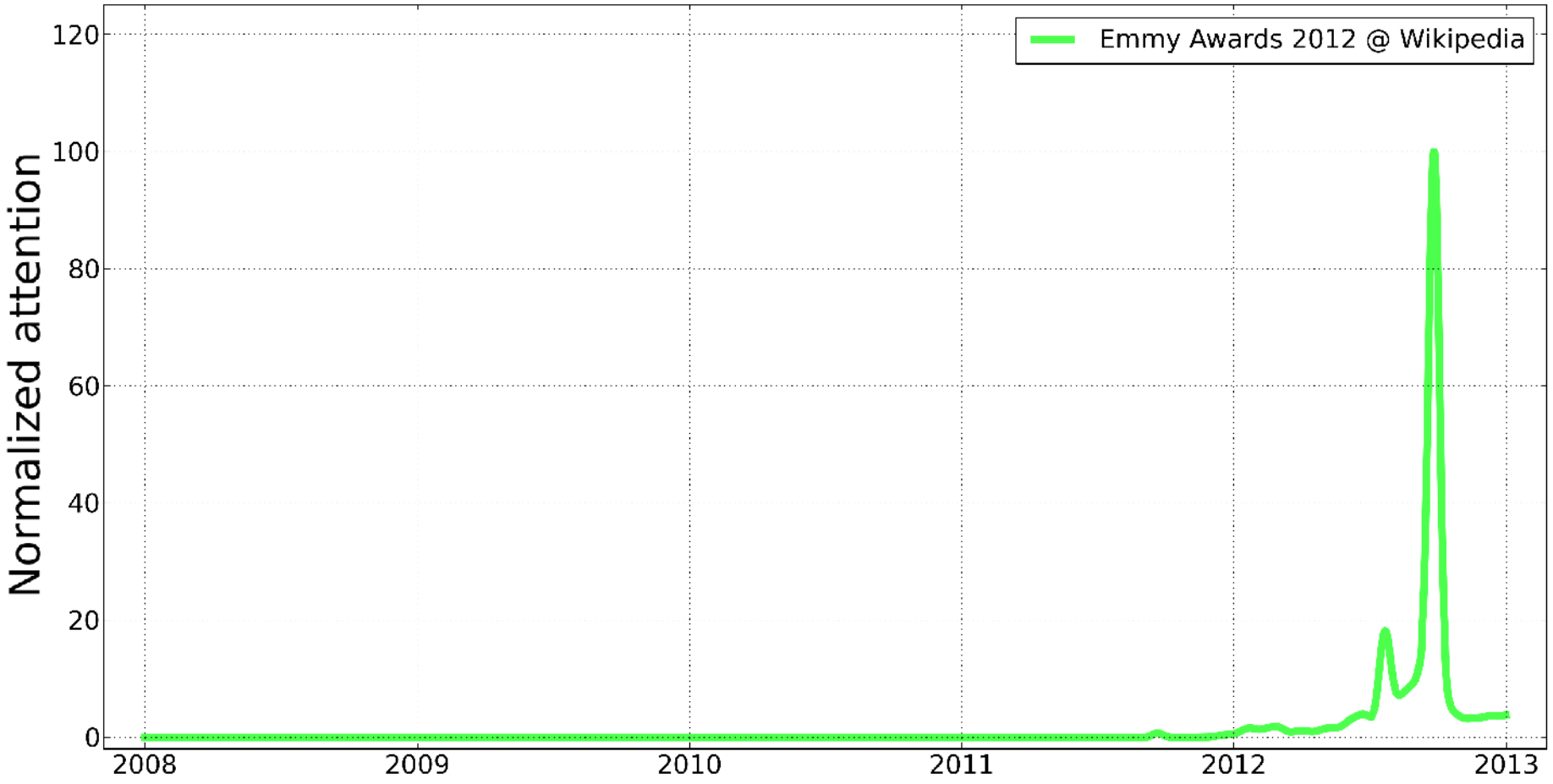
Recurring Signal



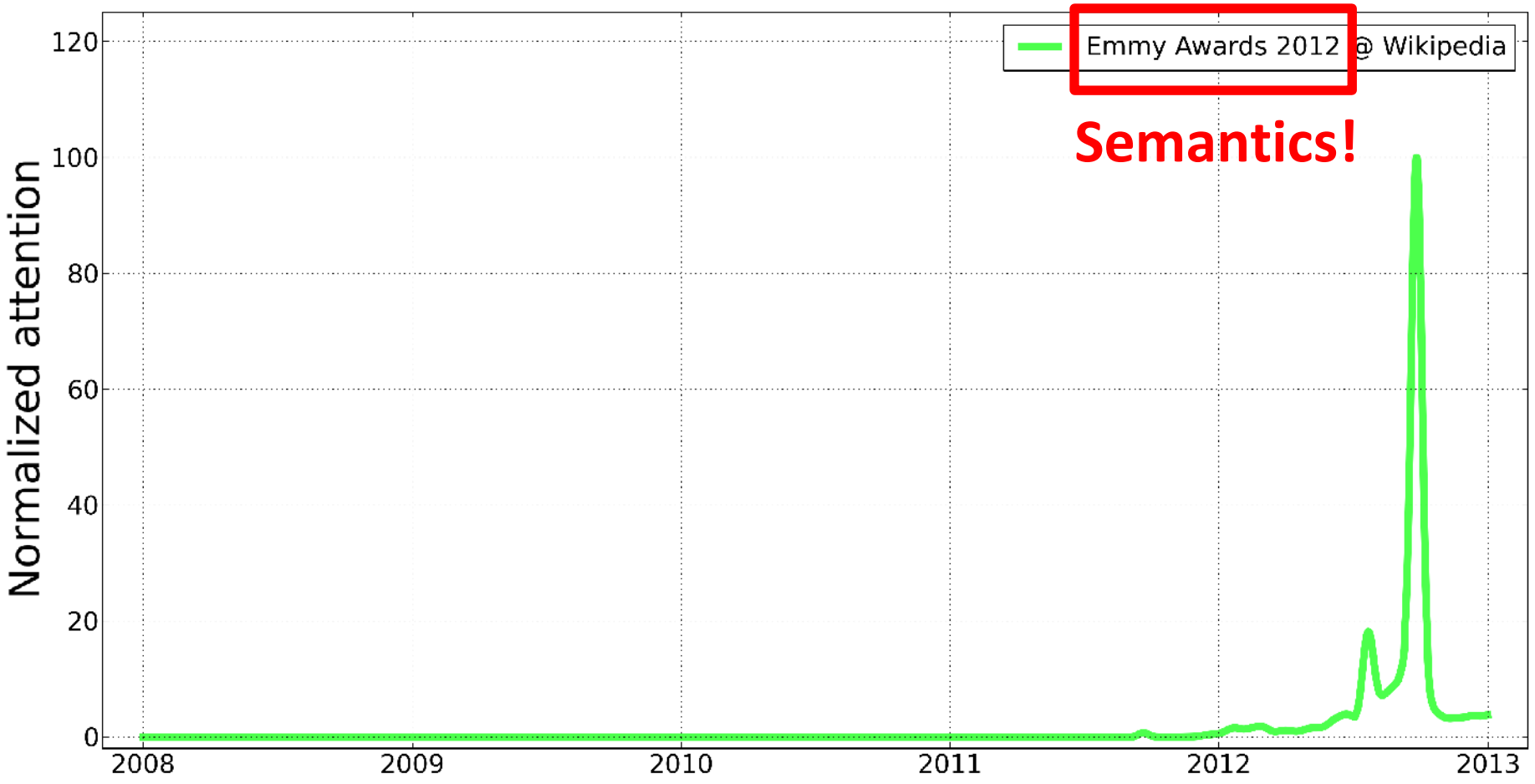
Recurring Signal



What To Do In This Case?

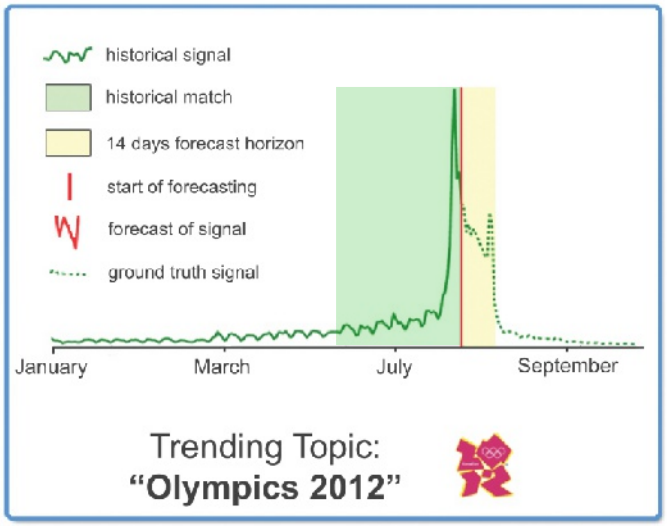


What To Do In This Case?

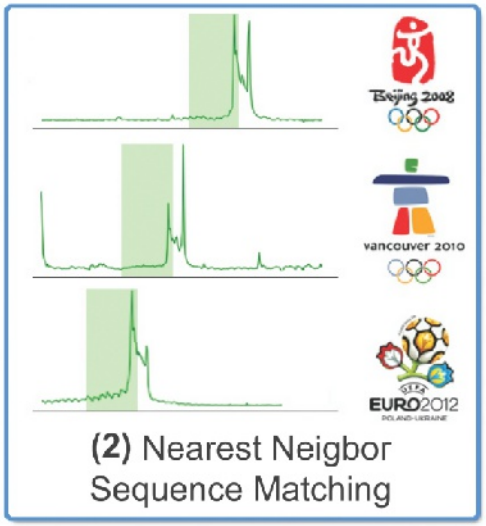


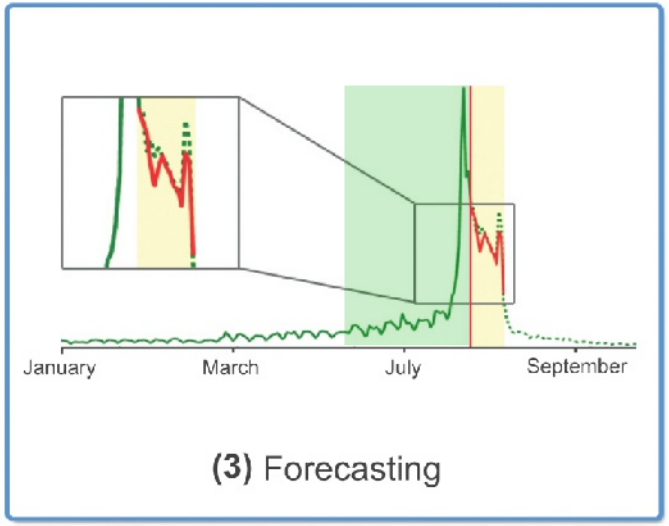
Exploiting Semantically Similar Topics

Approach

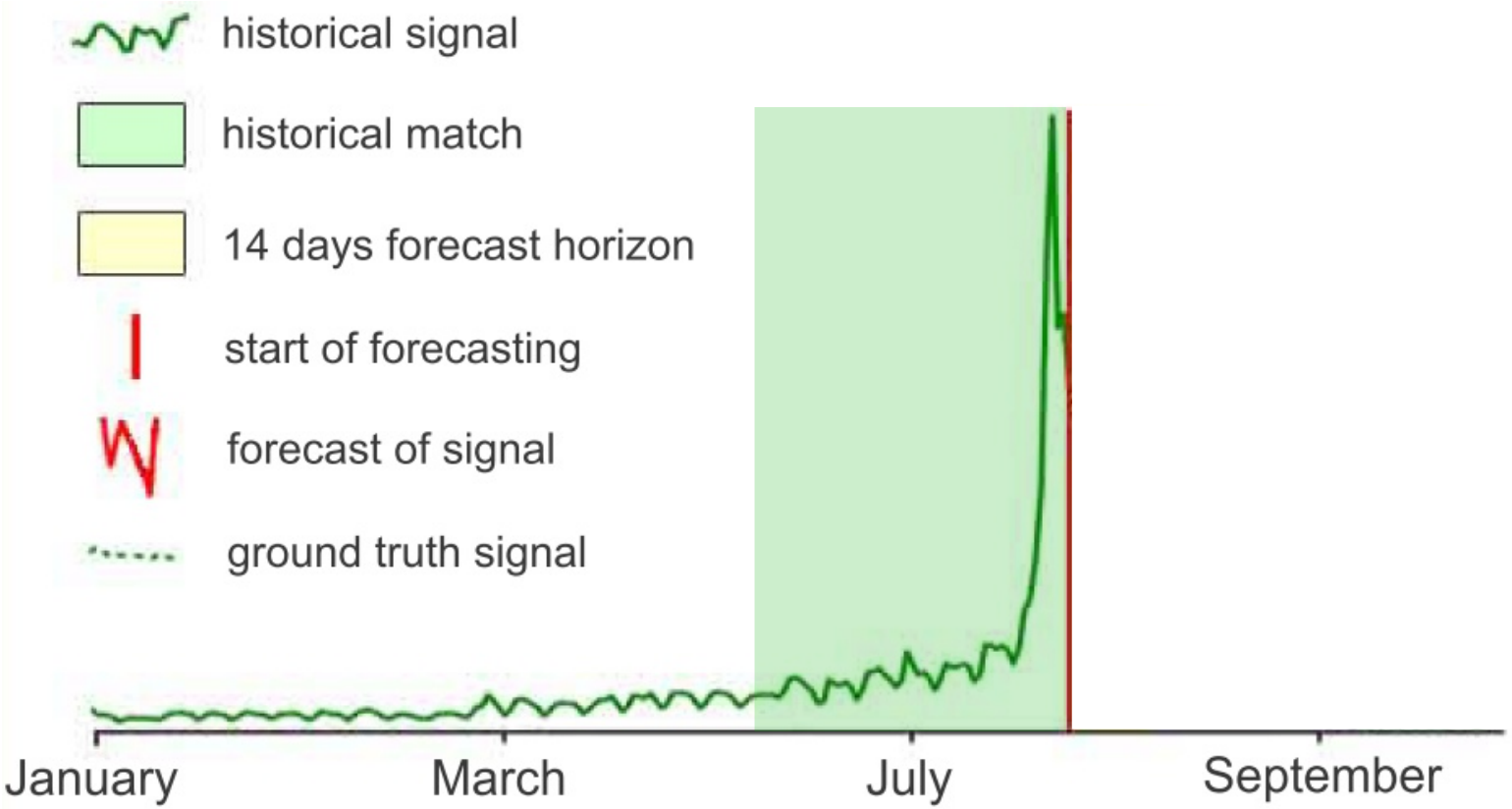

historical signal
historical match
14 days forecast horizon
start of forecasting
forecast of signal
ground truth signal
January
March
July
September
Trending Topic:
"Olympics 2012"


EURO2012
vancouver 2010
FIFA WORLD CUP
Brasil
Beijing 2008
torino 2006
(1) Discover Semantically
Similar Topics


Beijing 2008
vancouver 2010
EURO2012
(2) Nearest Neigbor
Sequence Matching

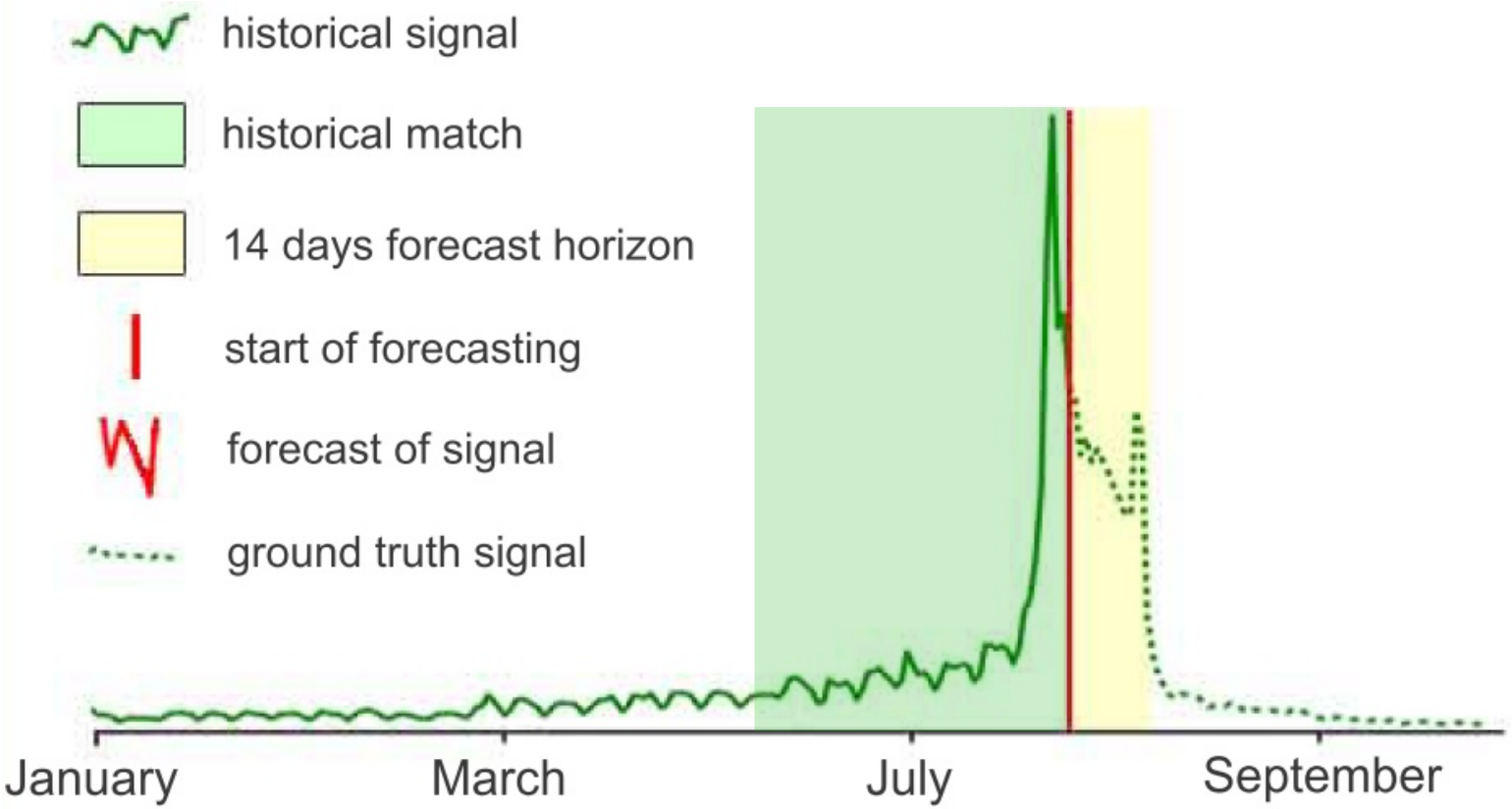

January
March
July
September
(3) Forecasting

Approach



Trending Topic:
“Olympics 2012”

Approach



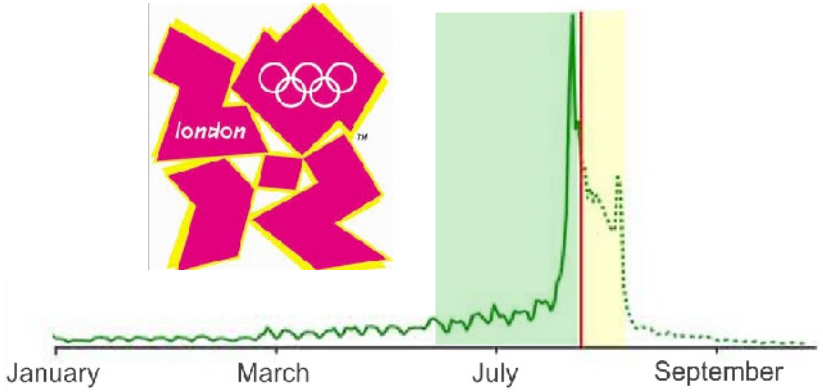
Trending Topic:
"Olympics 2012"

Discovering Semantically Similar Topics

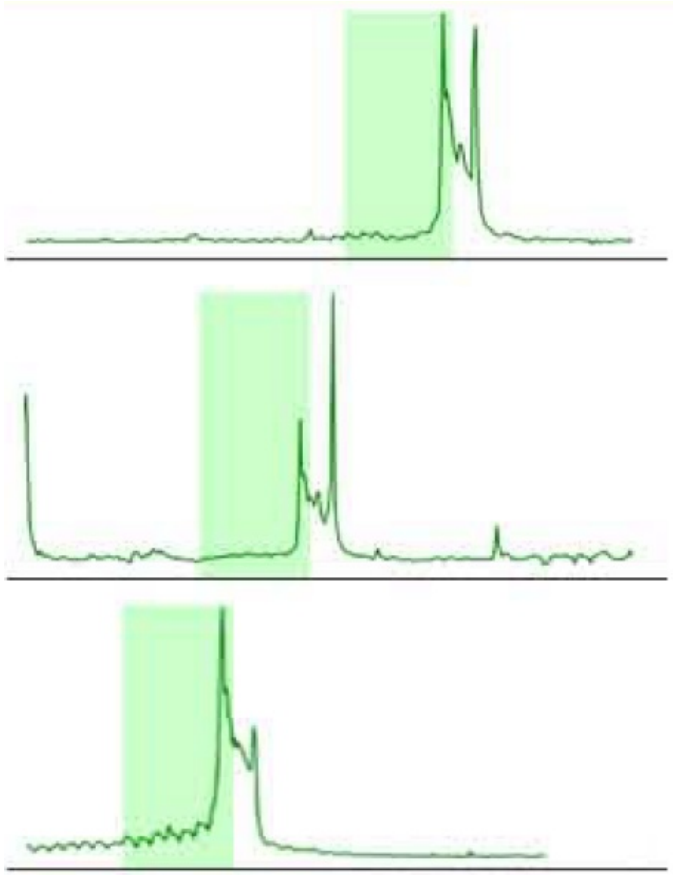
Trending Topic:
"Olympics 2012"

dcterms:subject	category:Olympic_Games_in_the_United_Kingdom category:Sports_festivals_in_London category:Scheduled_sports_events category:2012_Summer_Olympics category:2012_in_London
rdf:type	owl:Thing http://schema.org/SportsEvent dbpedia-owl:Event dbpedia-owl:SportsEvent http://schema.org/Event dbpedia-owl:Olympics yago:SportsFestivalsInTheUnitedKingdom yago:SportsFestivalsInEngland http://umbel.org/umbel/rc/OlympicGames http://umbel.org/umbel/rc/SportsEvent http://umbel.org/umbel/rc/Event yago:SportingEventsInEngland

Nearest Neighbor Sequence Matching

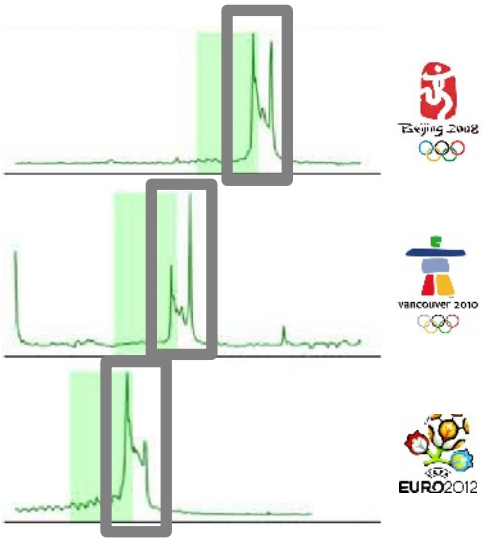


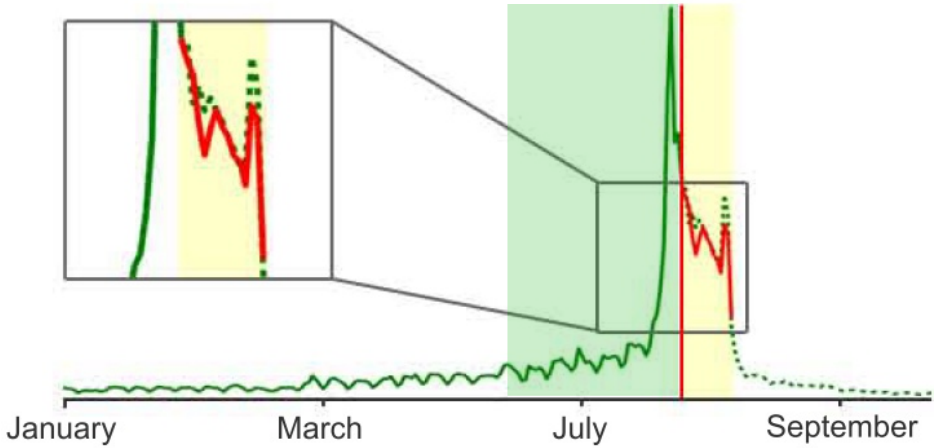
$$d(S^{t_0}_{topic}, S^{t_i}_{sim_i})$$



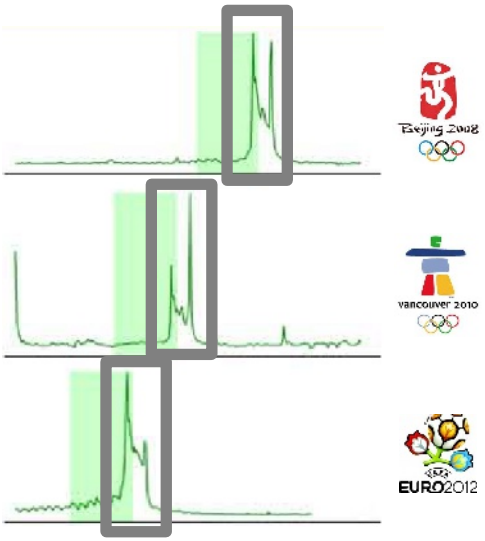
$$N^k(S^{t_0}_{topic}) = \{S^{t_1}_{sim_1}, \ldots, S^{t_k}_{sim_k}\}$$

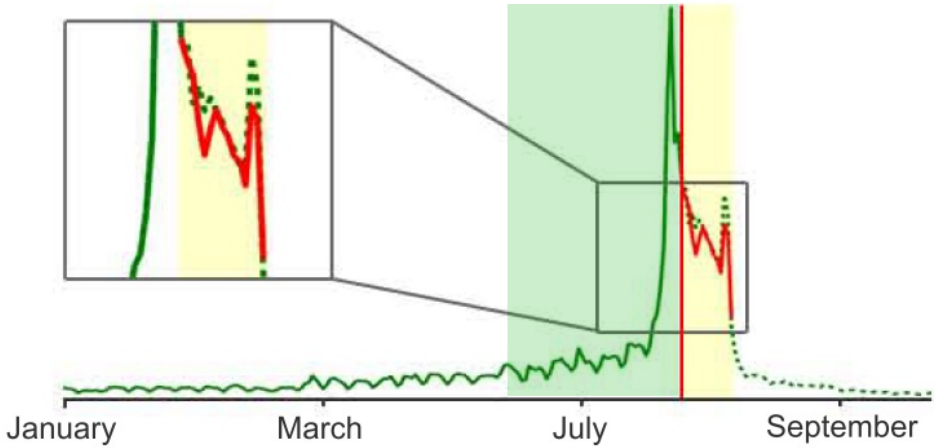
Forecasting





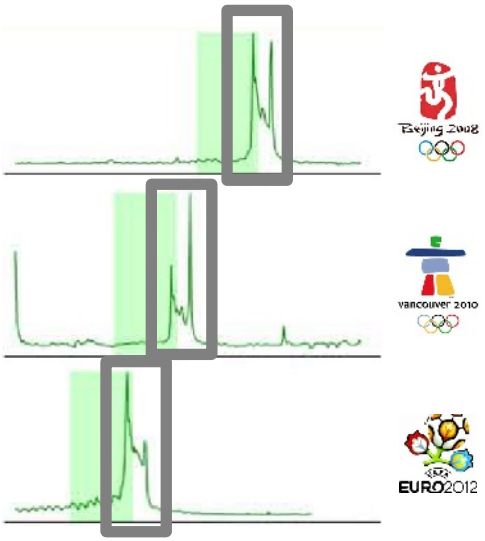
Forecasting

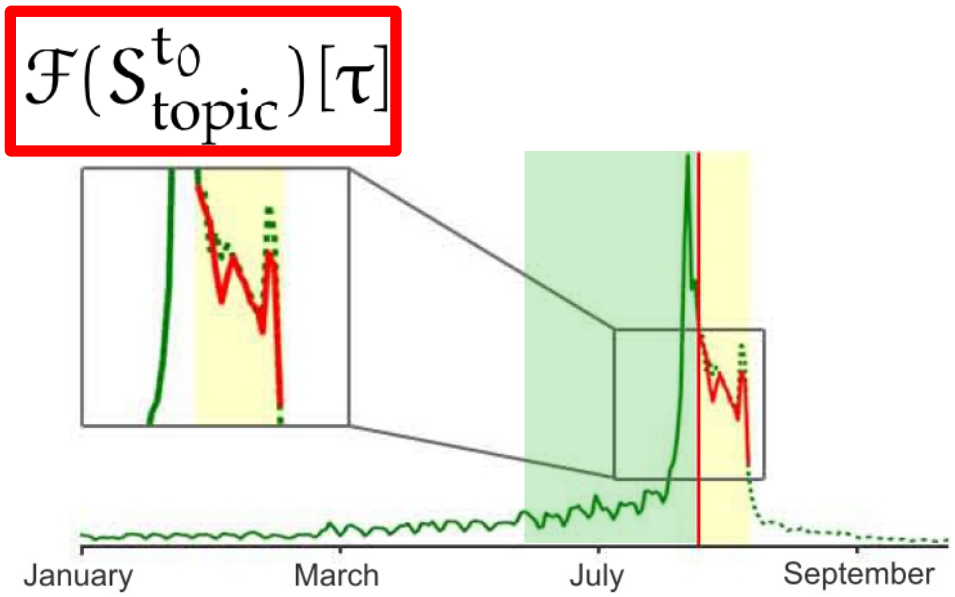




$$\mathcal{F}(S^{t_0}_{topic})[\tau] = \underset{S^{t'}_{topic'} \in N^k(S^t_{topic})}{\text{median}} \left(\alpha(S^{t_0}_{topic}, S^{t'}_{topic'}) \cdot S^{t'}_{topic'}[\tau]\right)$$

Forecasting

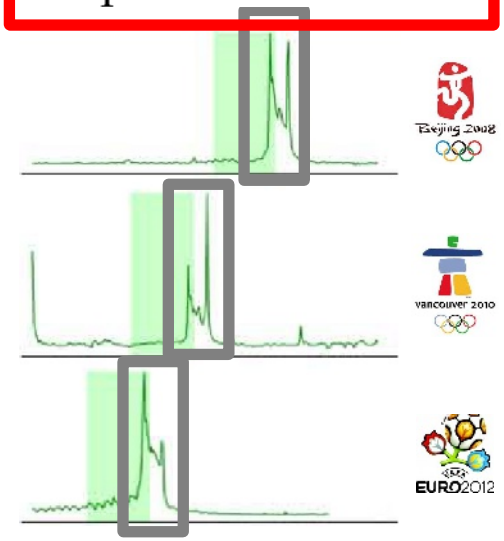




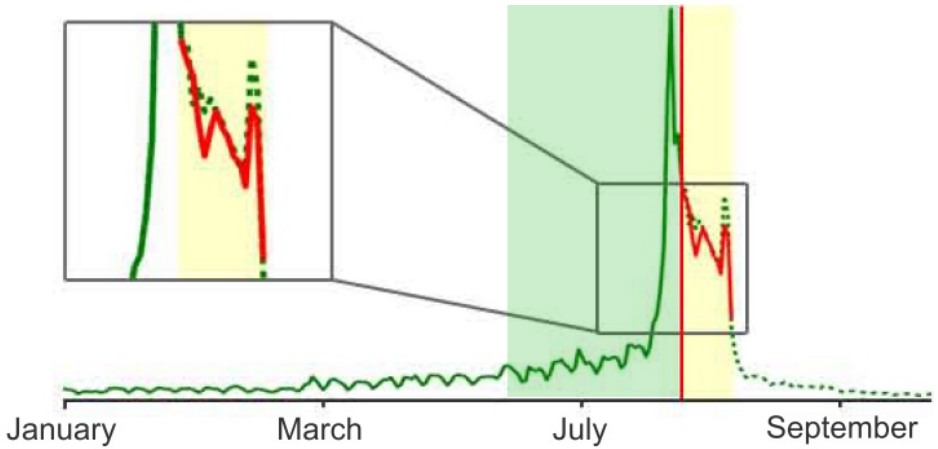
$$\mathcal{F}(S^{t_0}_{topic})[\tau] = \underset{S^{t'}_{topic'} \in N^k(S^t_{topic})}{\text{median}} \left(\alpha(S^{t_0}_{topic}, S^{t'}_{topic'}) \cdot S^{t'}_{topic'}[\tau]\right)$$

Forecasting

$$S^{t'}_{topic'} \in N^k(S^t_{topic})$$



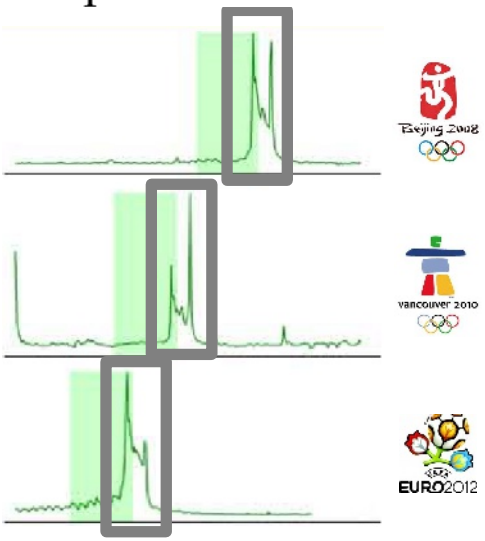
$$\mathcal{F}(S^{t_0}_{topic})[\tau]$$



$$\mathcal{F}(S^{t_0}_{topic})[\tau] = \underset{S^{t'}_{topic'} \in N^k(S^t_{topic})}{\text{median}} \left(\alpha(S^{t_0}_{topic}, S^{t'}_{topic'}) \cdot S^{t'}_{topic'}[\tau] \right)$$

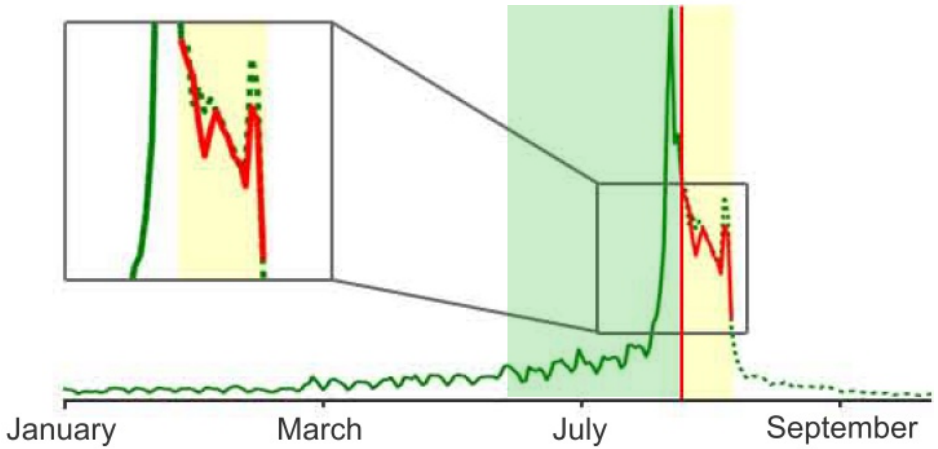
Forecasting

$S^{t'}_{topic'} \in N^k(S^t_{topic})$



median

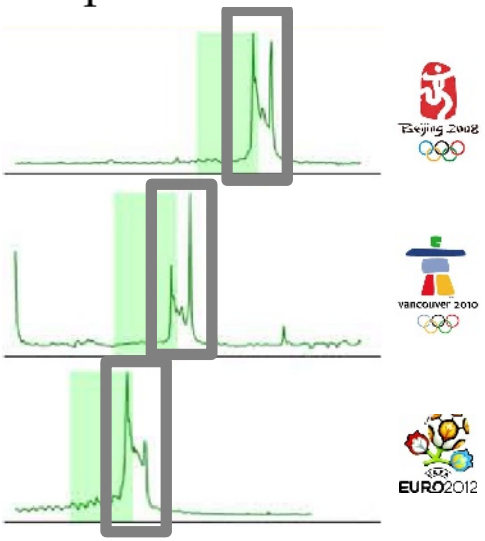
$\mathcal{F}(S^{t_0}_{topic})[\tau]$



$$\mathcal{F}(S^{t_0}_{topic})[\tau] = \underset{S^{t'}_{topic'} \in N^k(S^t_{topic})}{\text{median}} \left(\alpha(S^{t_0}_{topic}, S^{t'}_{topic'}) \cdot S^{t'}_{topic'}[\tau]\right)$$

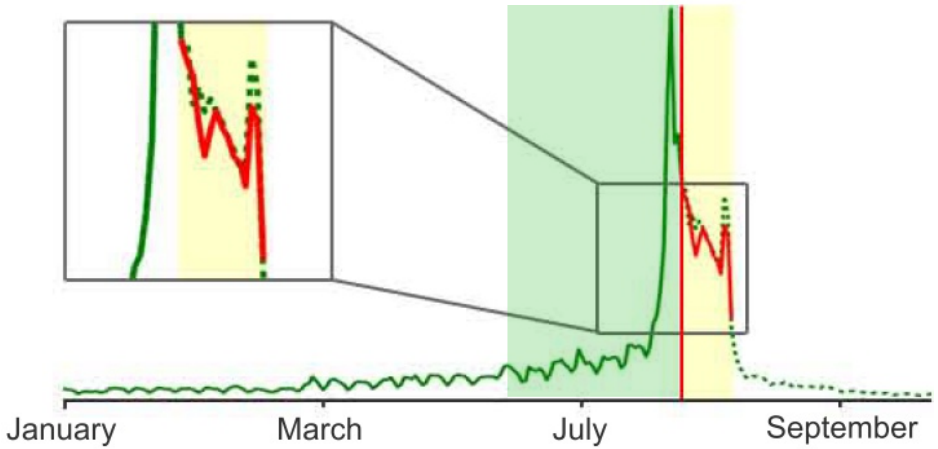
Forecasting

$$S^{t'}_{topic'} \in N^k(S^t_{topic})$$

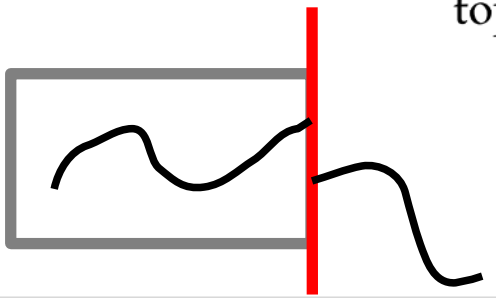


median

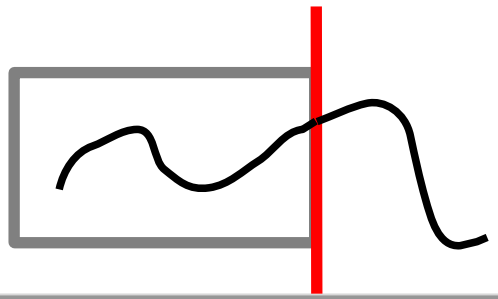
$$\mathcal{F}(S^{t_0}_{topic})[\tau]$$



$$\mathcal{F}(S^{t_0}_{topic})[\tau] = \underset{S^{t'}_{topic'} \in N^k(S^t_{topic})}{\text{median}} \left(\alpha(S^{t_0}_{topic}, S^{t'}_{topic'}) \cdot S^{t'}_{topic'}[\tau]\right)$$



$$\alpha(S^{t_0}_{topic}, S^{t'}_{topic'})$$



Evaluation

Dataset

- **Public** dataset of **Wikipedia** view statistics
- **Historical data** from 2008 until today
- **Large-scale:** 5 million articles attracting 870 million views each day
  - 2.8 TB compressed

→ Semantic similarity between 5 million articles

→ 9 billion sequences to choose from

- Use **200 top trending topics** for evaluation

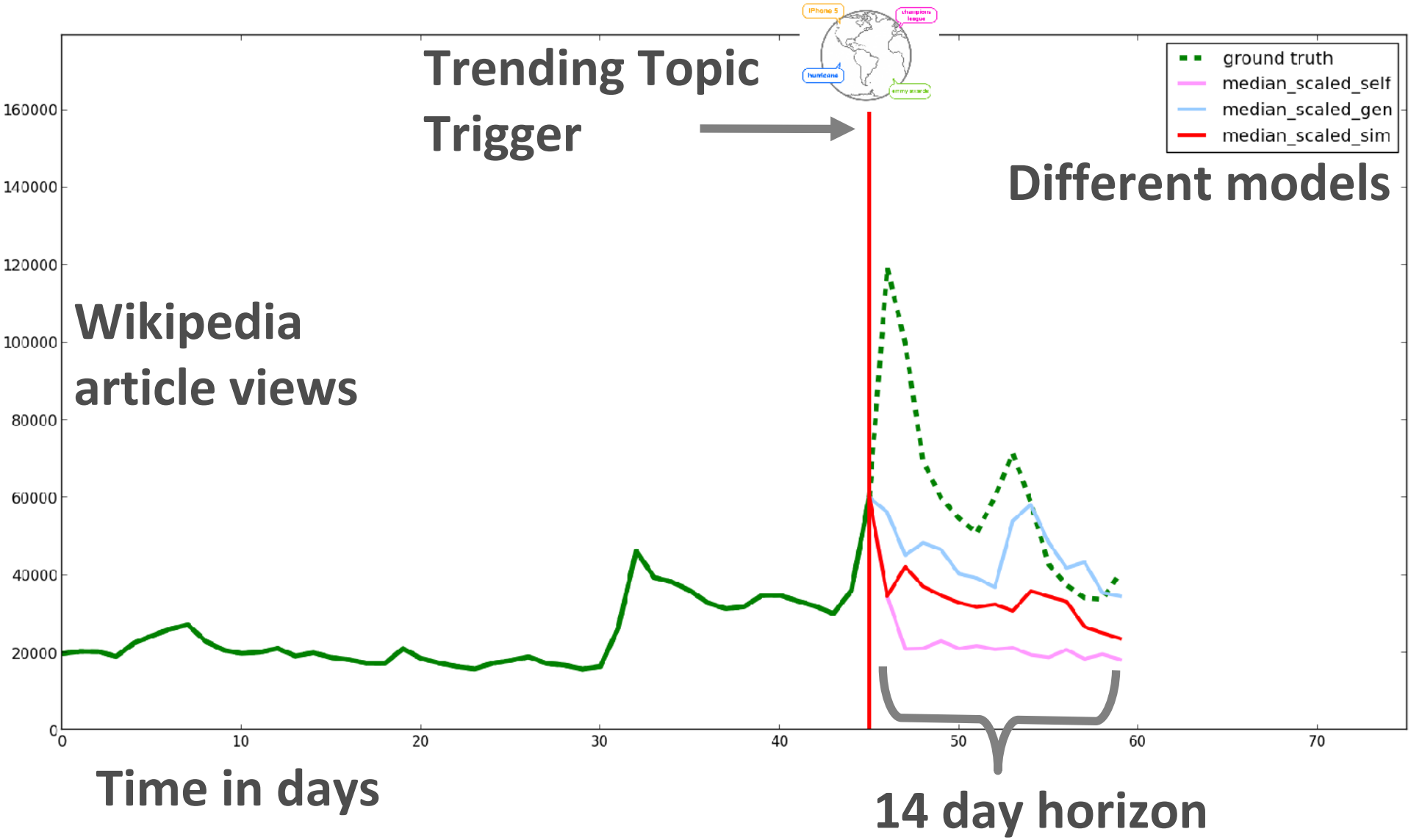
Thanks Jörn!

Discovering Semantically Similar Topics

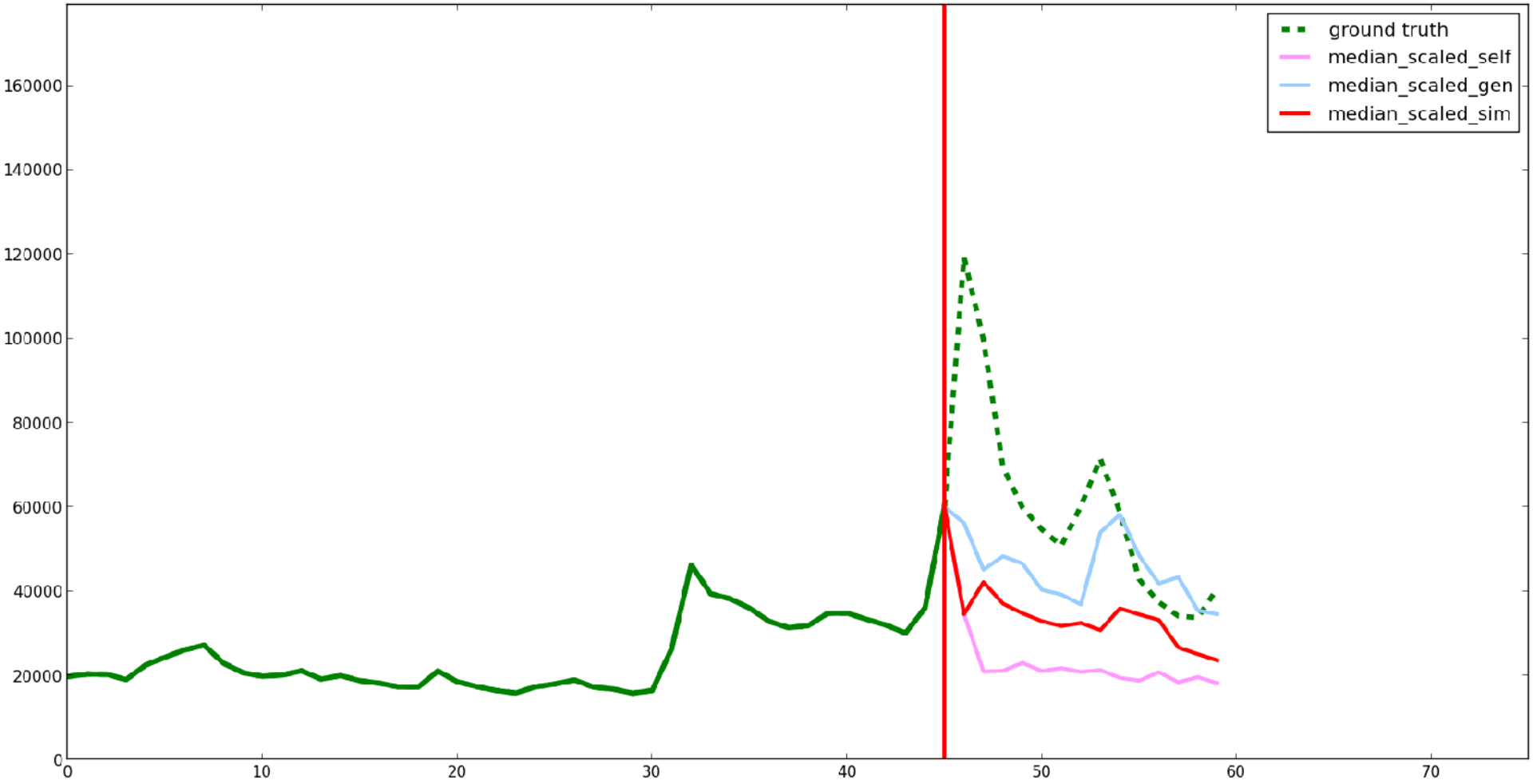
Trending Topic	Nearest Neighbor Topics
2012 Summer Olympics	**2008 Summer Olympics, UEFA Euro 2012, 2010 Winter Olympics** :: 2016 Summer Olympics, 2014 FIFA World Cup, 2006 Winter Olympics
Whitney Houston	**Ciara, Shakira, Celine Dion, Brittany Murphy, Ozzy Osbourne** :: Alicia Keys, Paul McCartney, Janet Jackson
Steve Jobs	**Mark Zuckerberg, Rupert Murdoch, Steve Jobs** :: Steve Wozniak, Bill Gates, Oprah Winfrey
Super Bowl XLVI	**Super Bowl, Super Bowl XLV, Super Bowl XLIV** :: Super Bowl XLIII, 2012 Pro Bowl, UFC 119
Justin Bieber	**Selena Gomez, Kanye West, Justin Bieber** :: Katy Perry, Avril Lavigne, Justin Timberlake
84th Academy Awards	**83rd Academy Awards, 82nd Academy Awards** :: List of Academy Awards ceremonies, 81st Academy Awards

Example: “The Hunger Games”

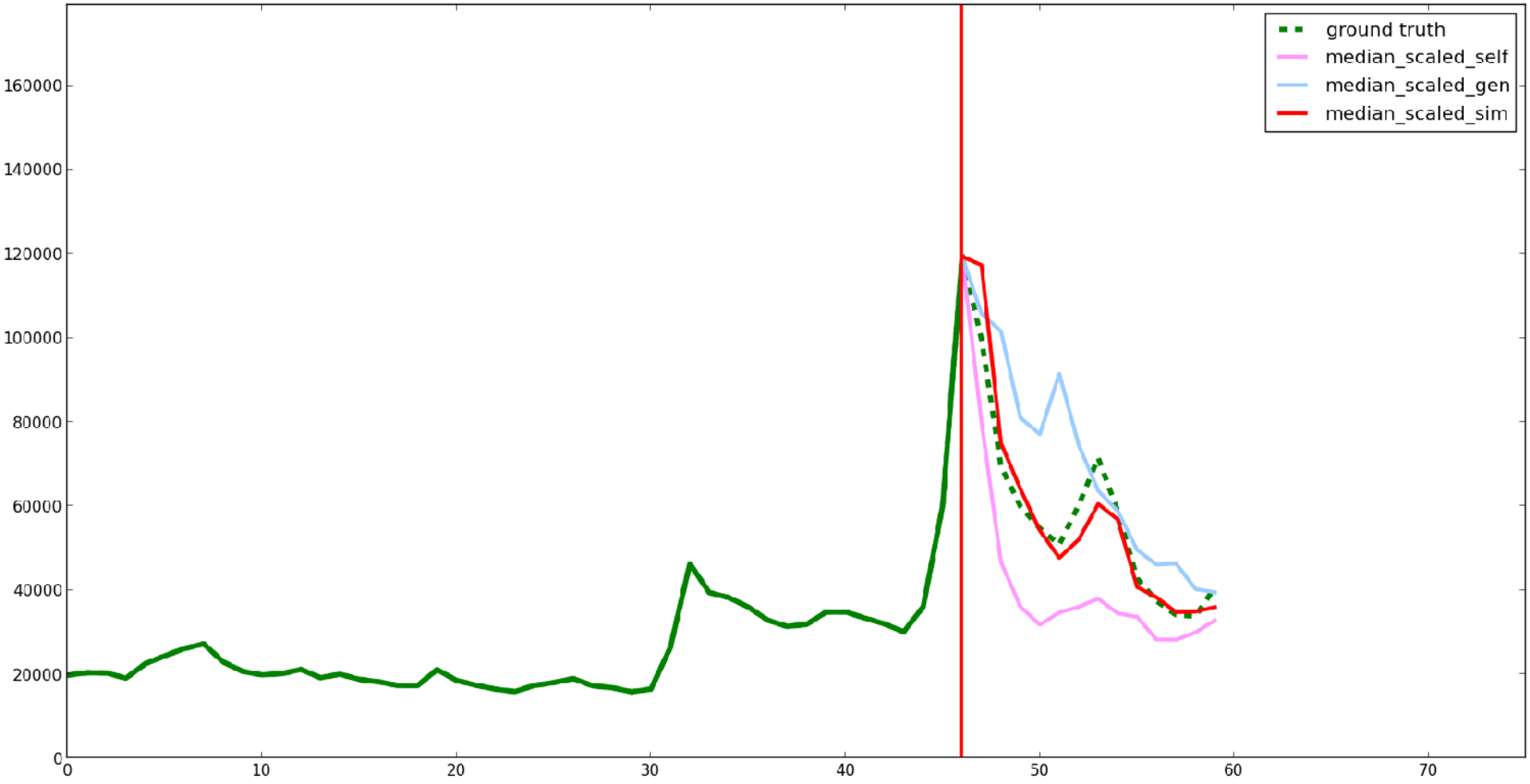
Example: “The Hunger Games”



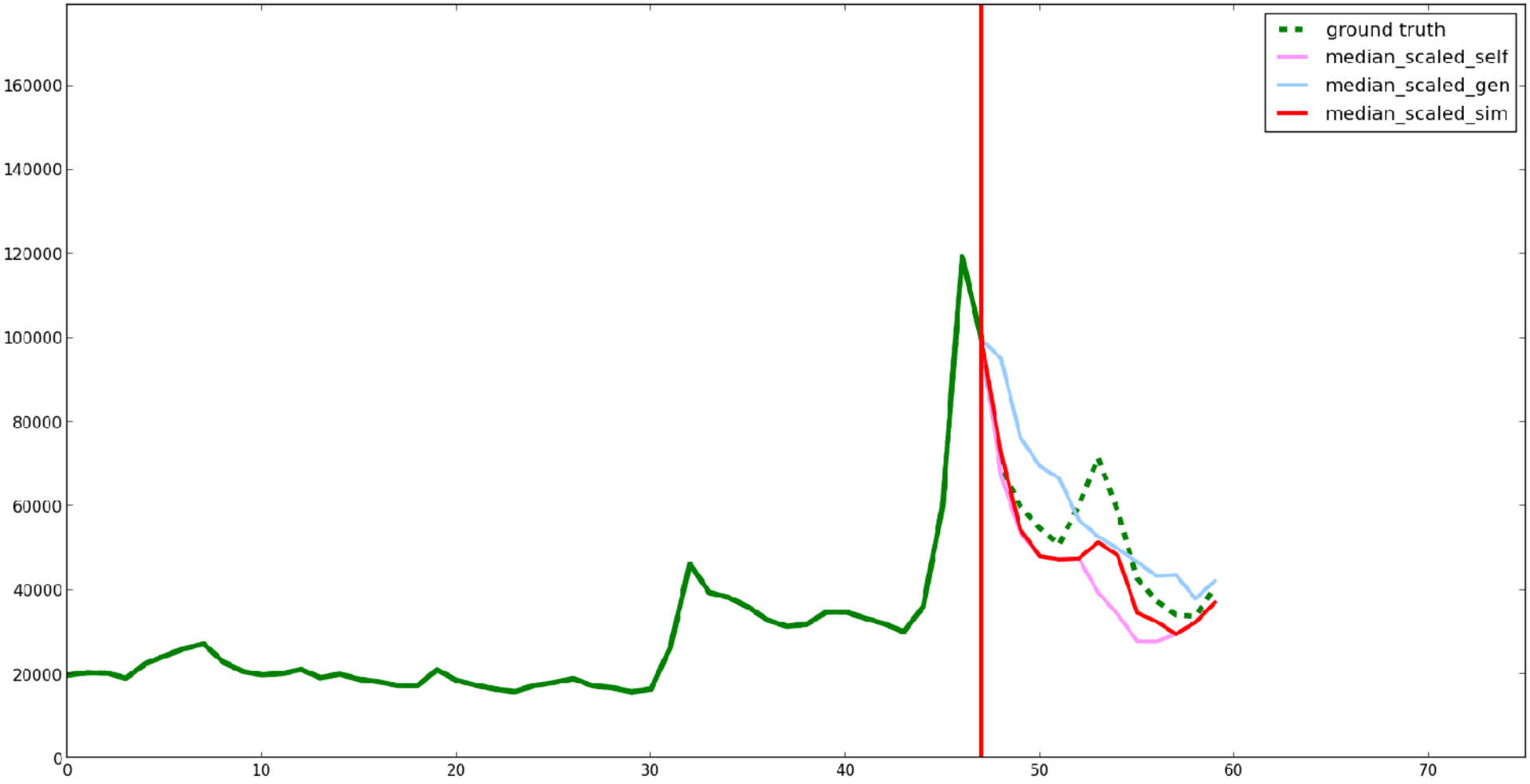
Example: “The Hunger Games”



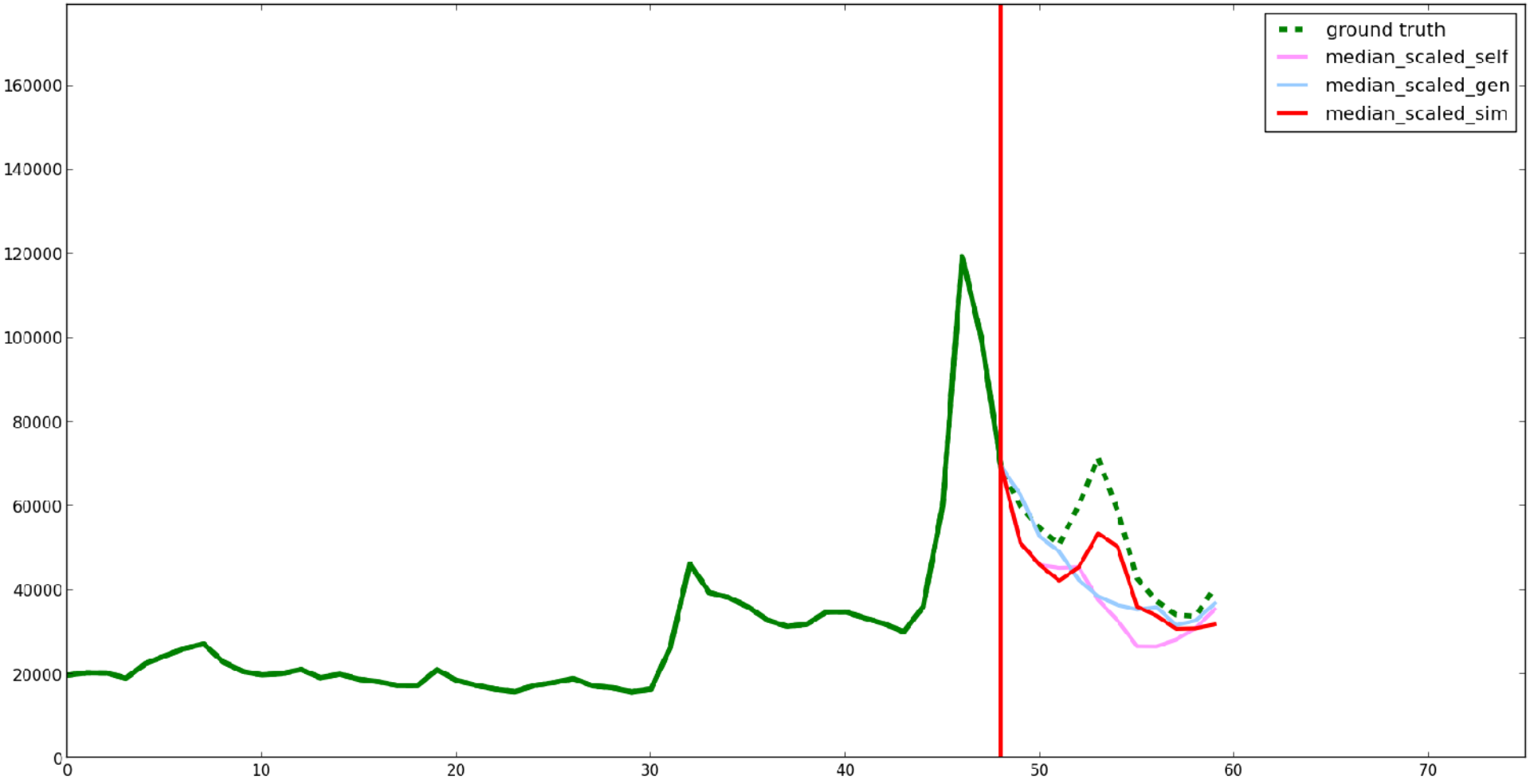
Example: “The Hunger Games”



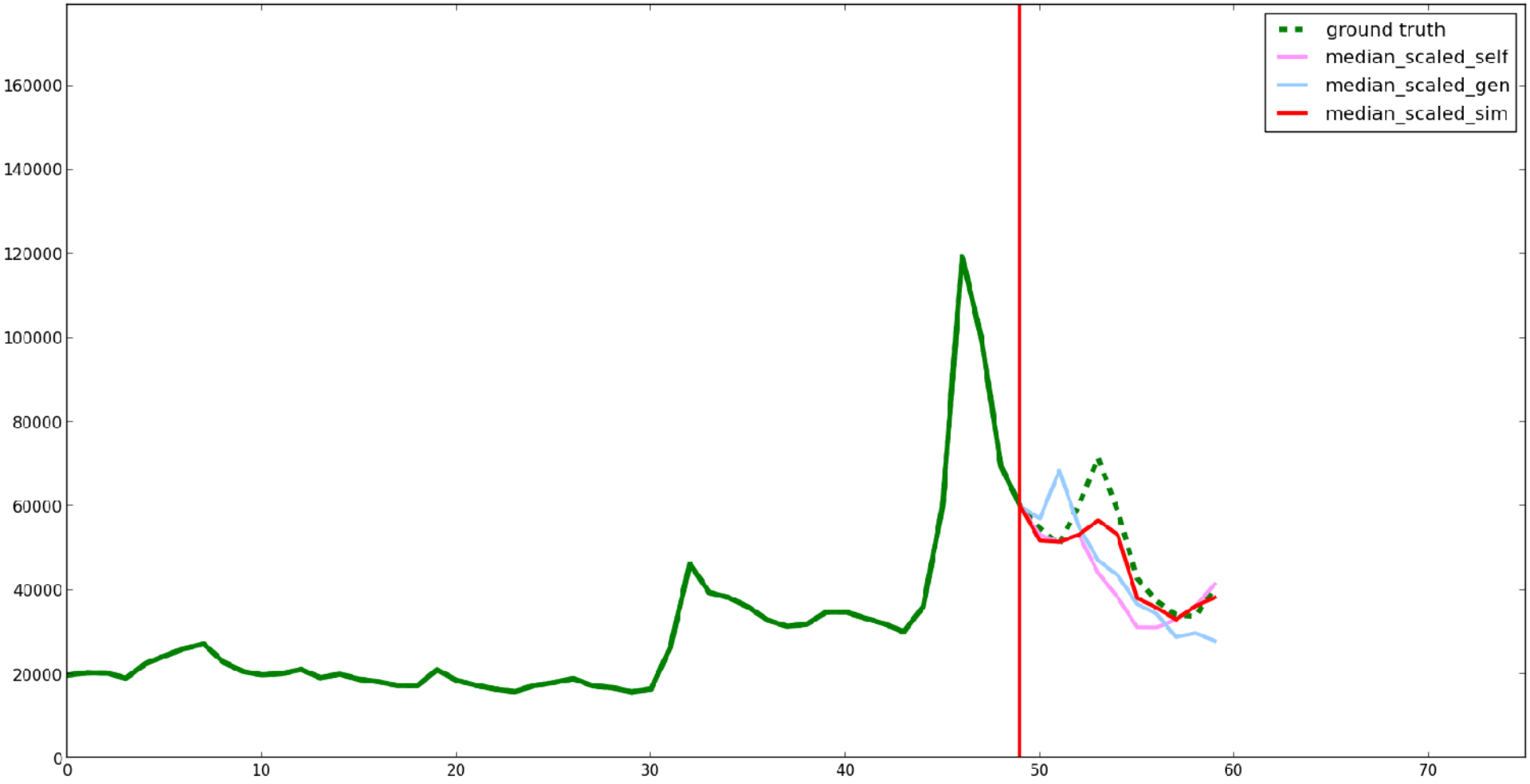
Example: “The Hunger Games”



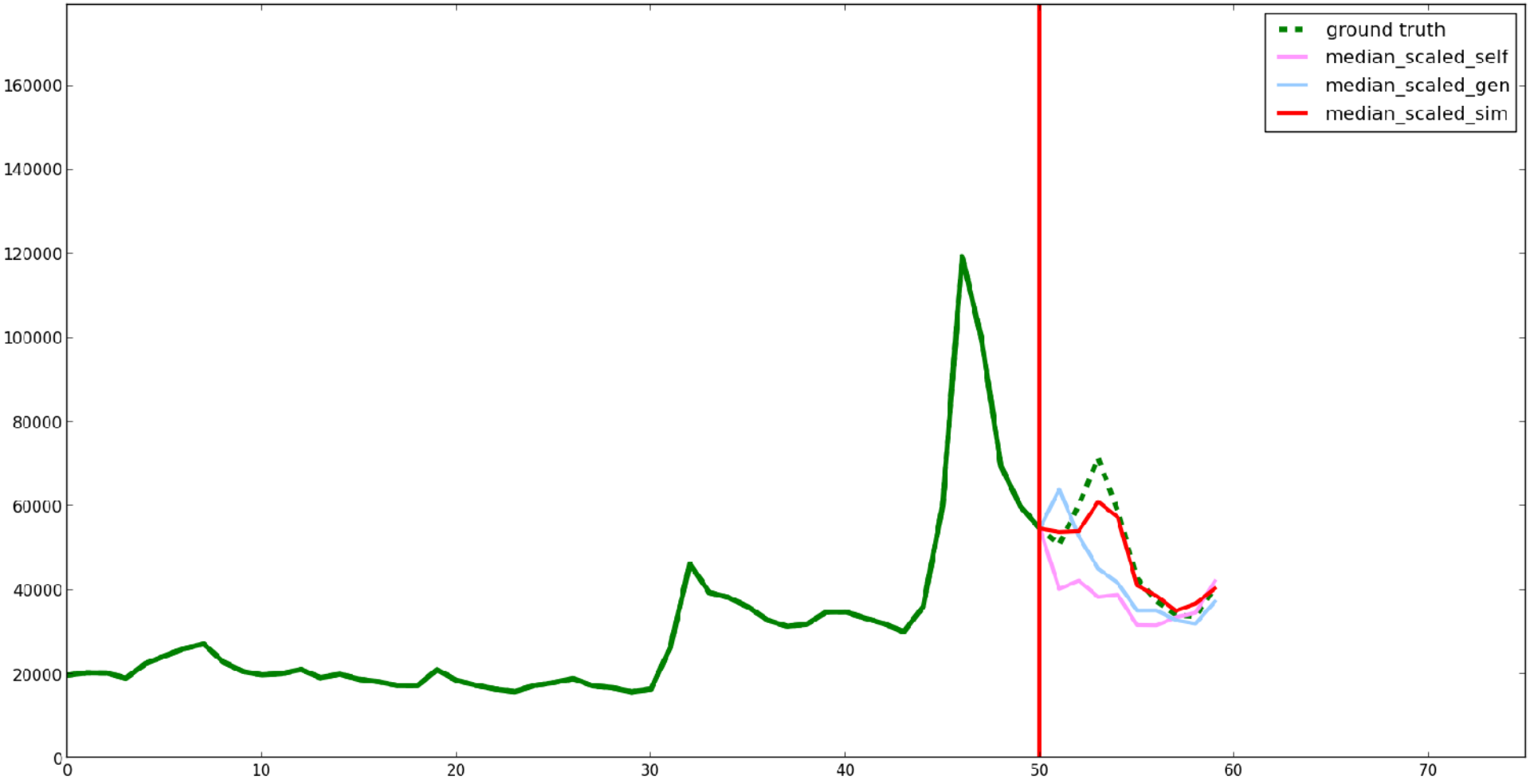
Example: “The Hunger Games”



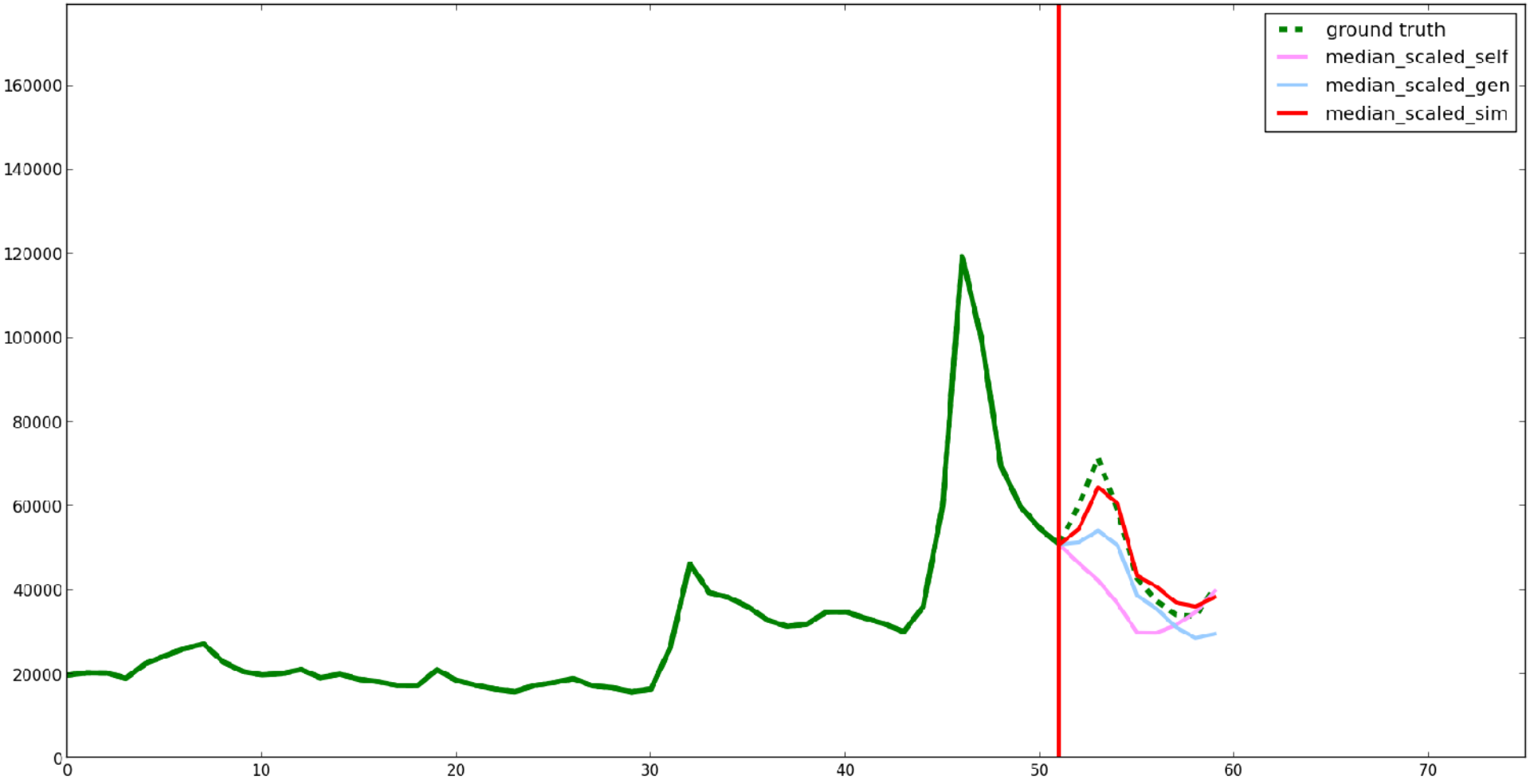
Example: “The Hunger Games”



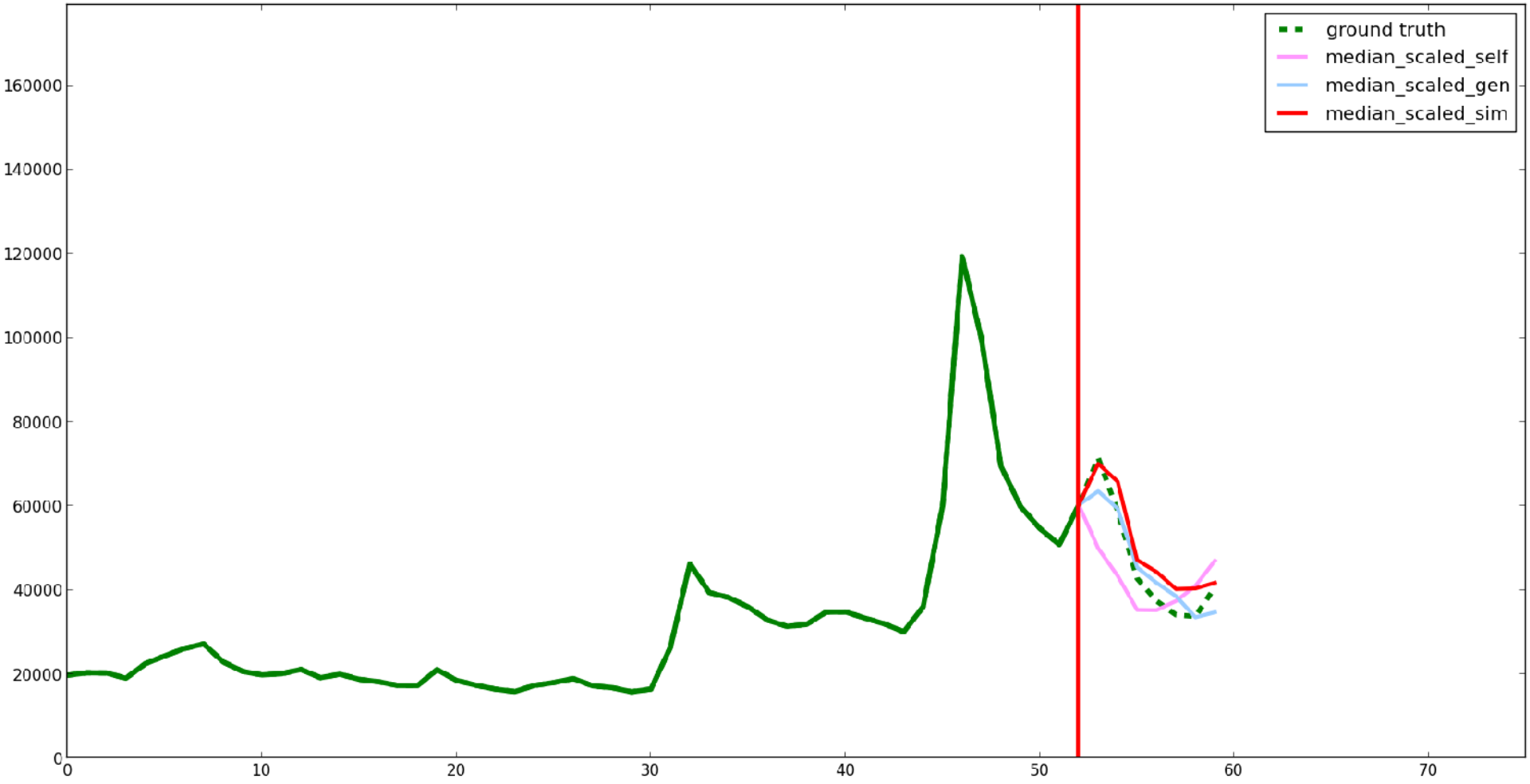
Example: “The Hunger Games”



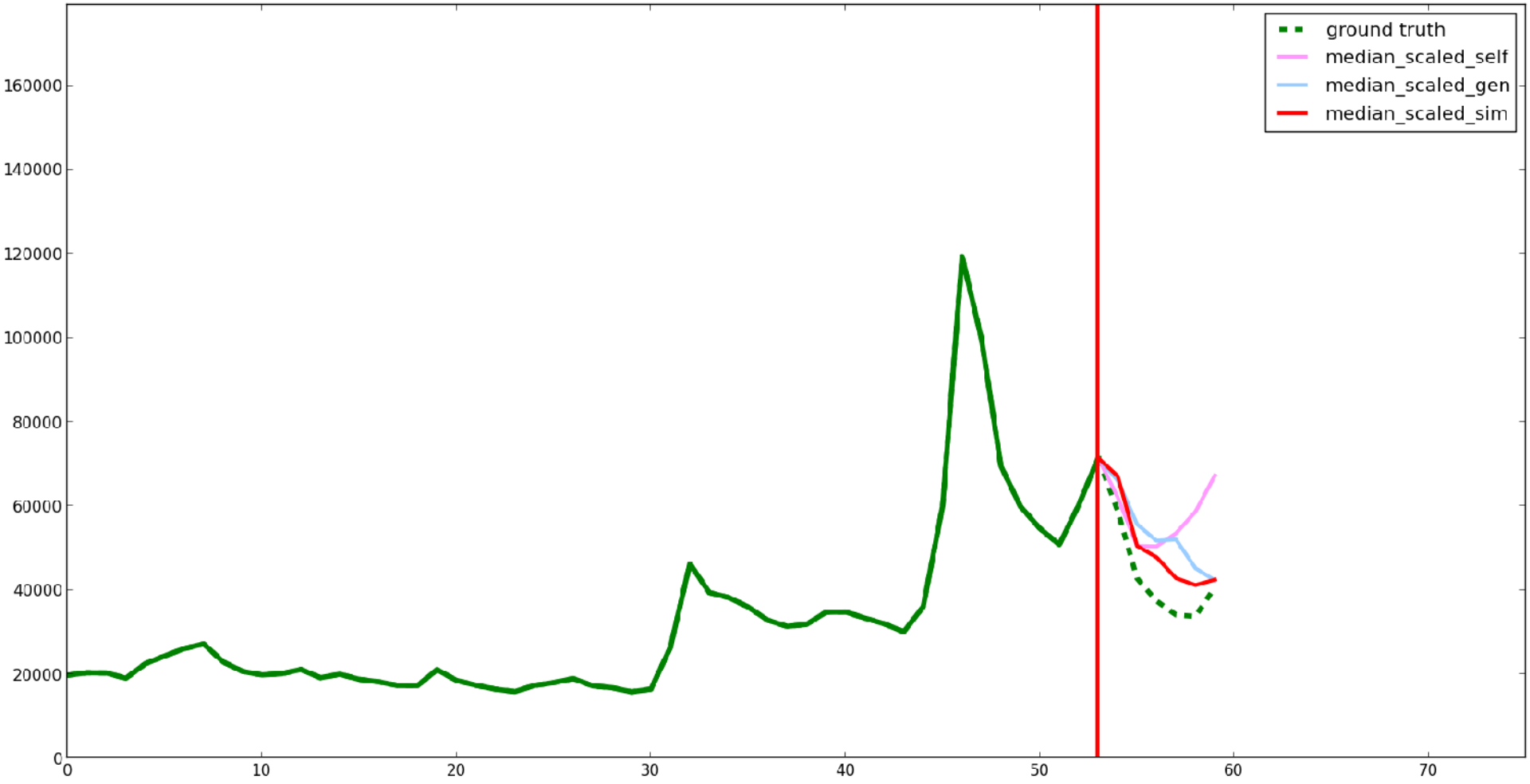
Example: “The Hunger Games”



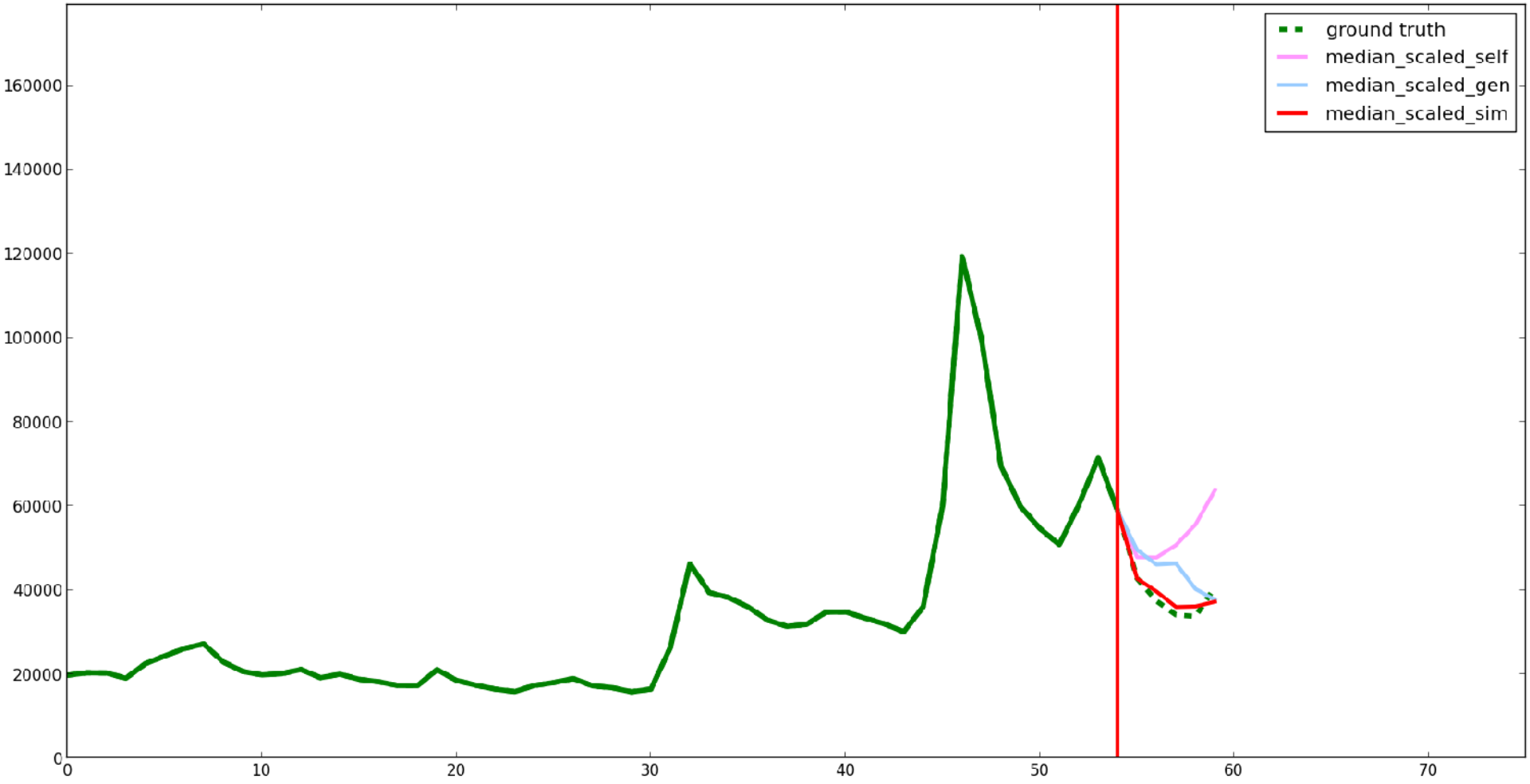
Example: “The Hunger Games”



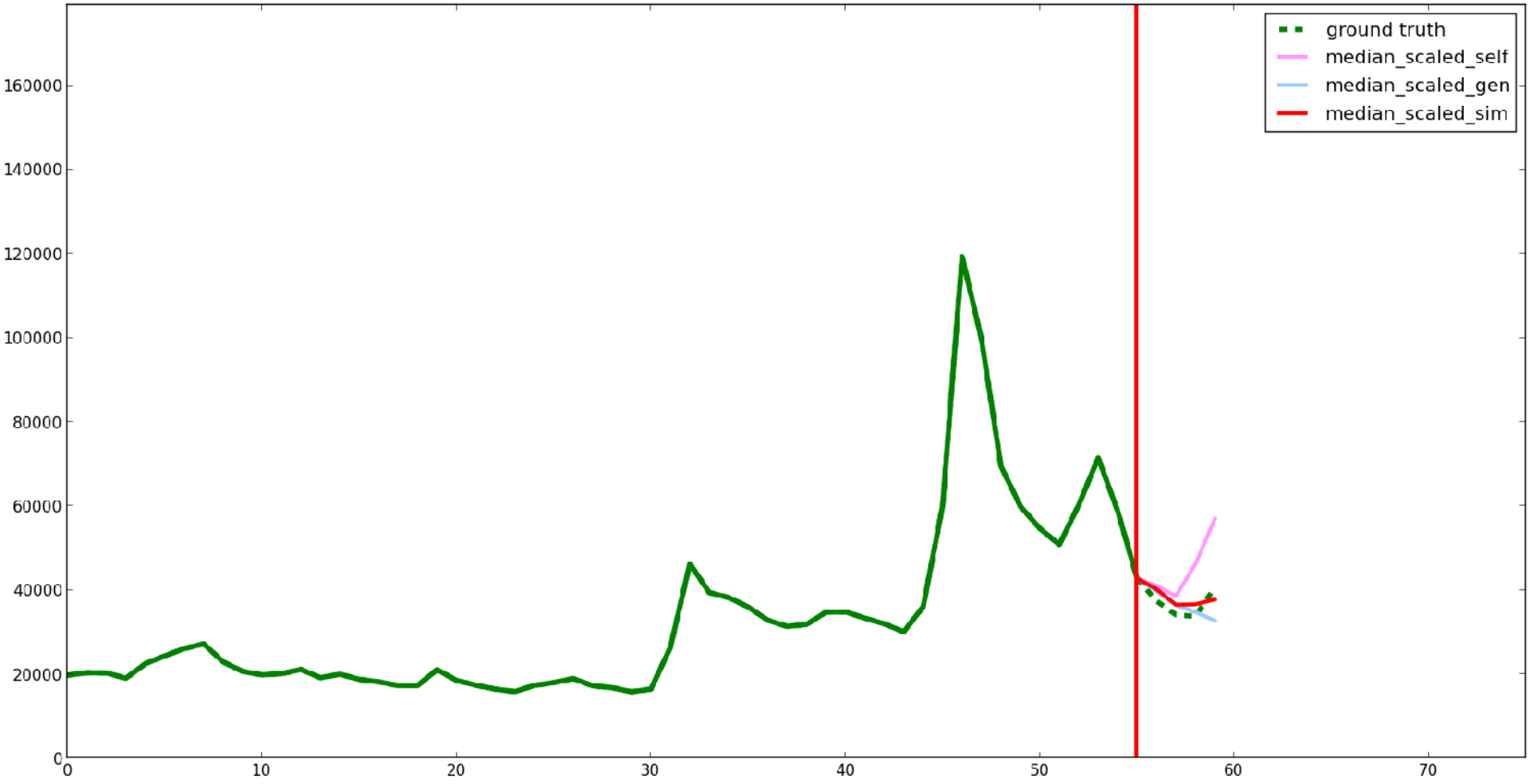
Example: “The Hunger Games”



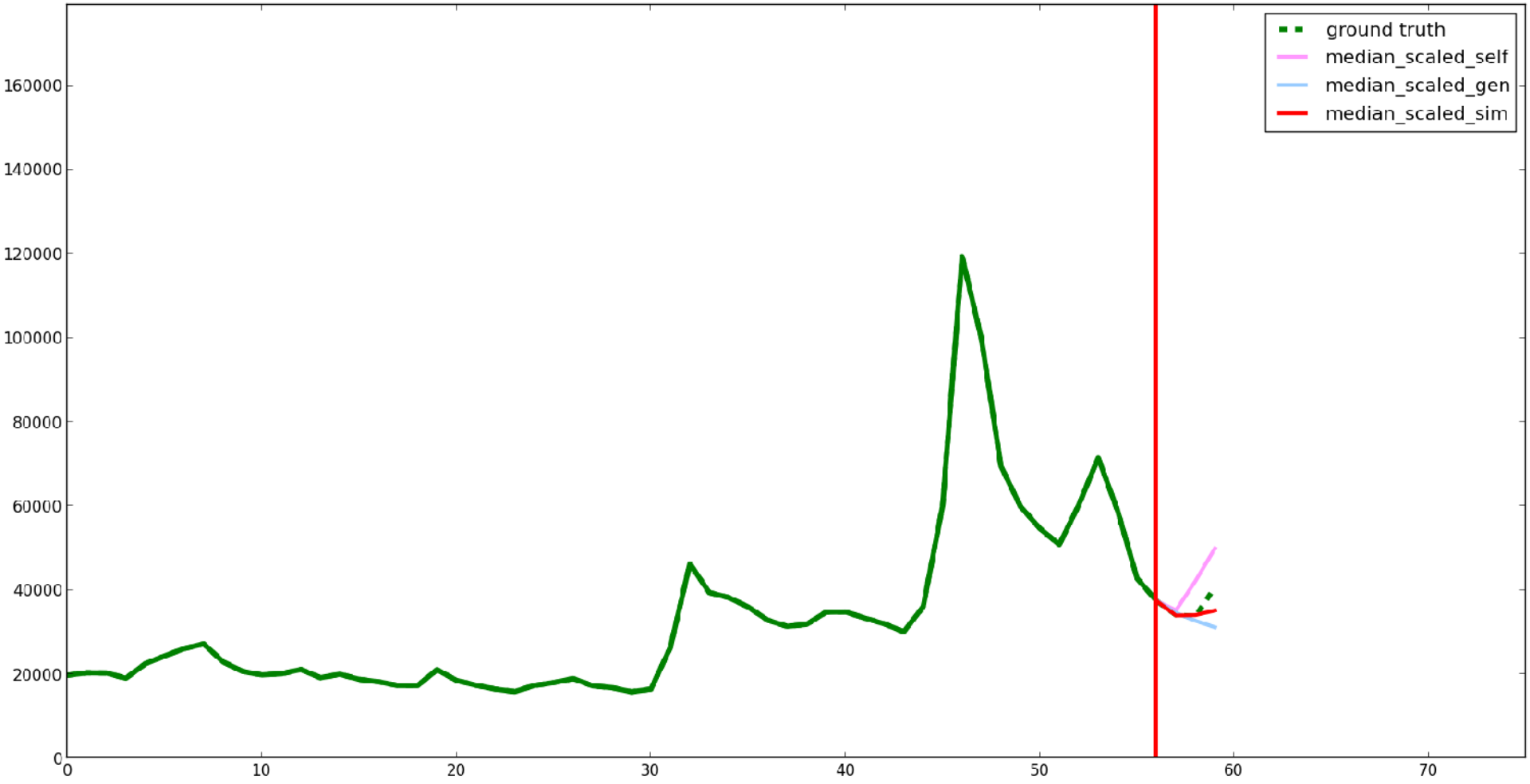
Example: “The Hunger Games”



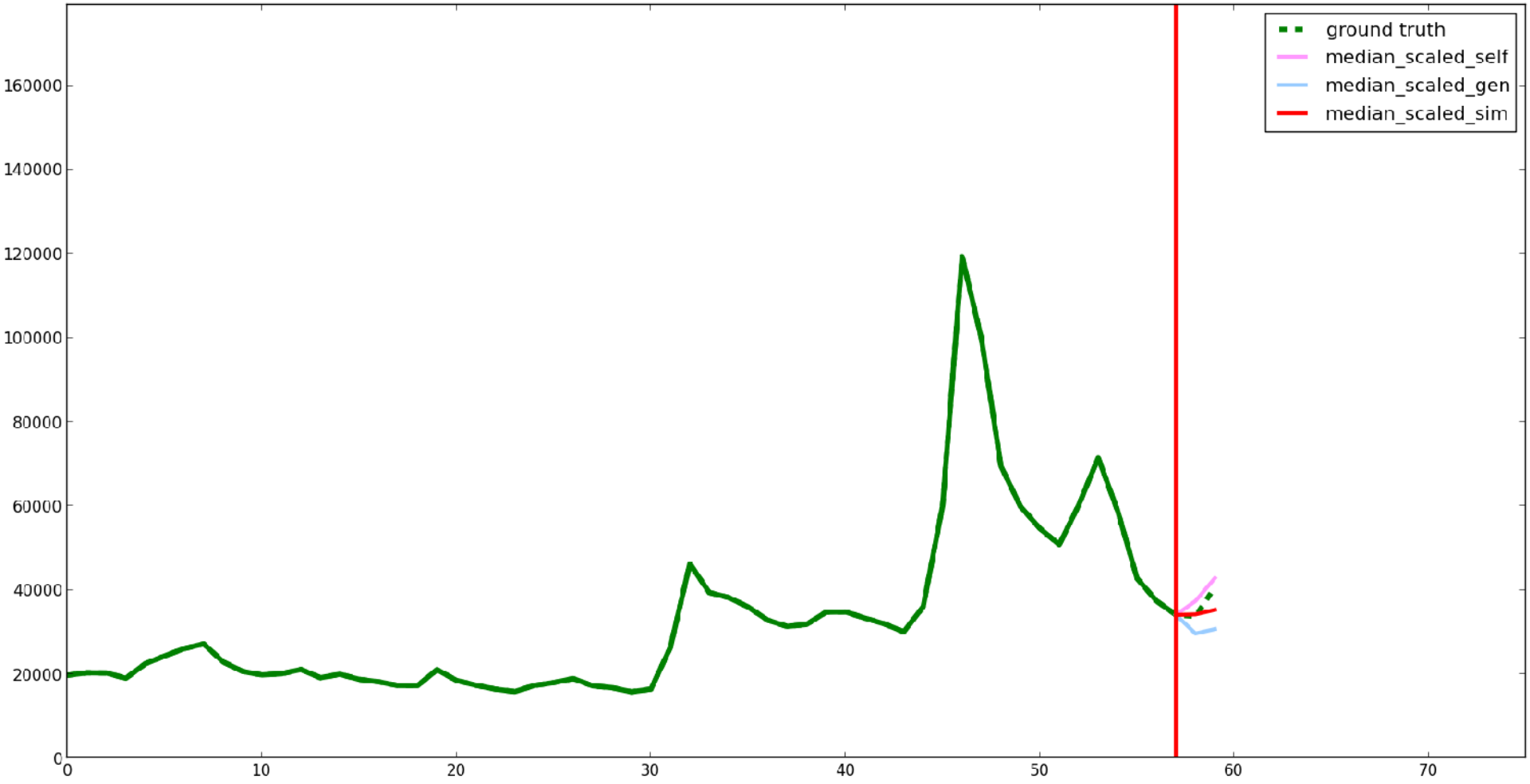
Example: “The Hunger Games”



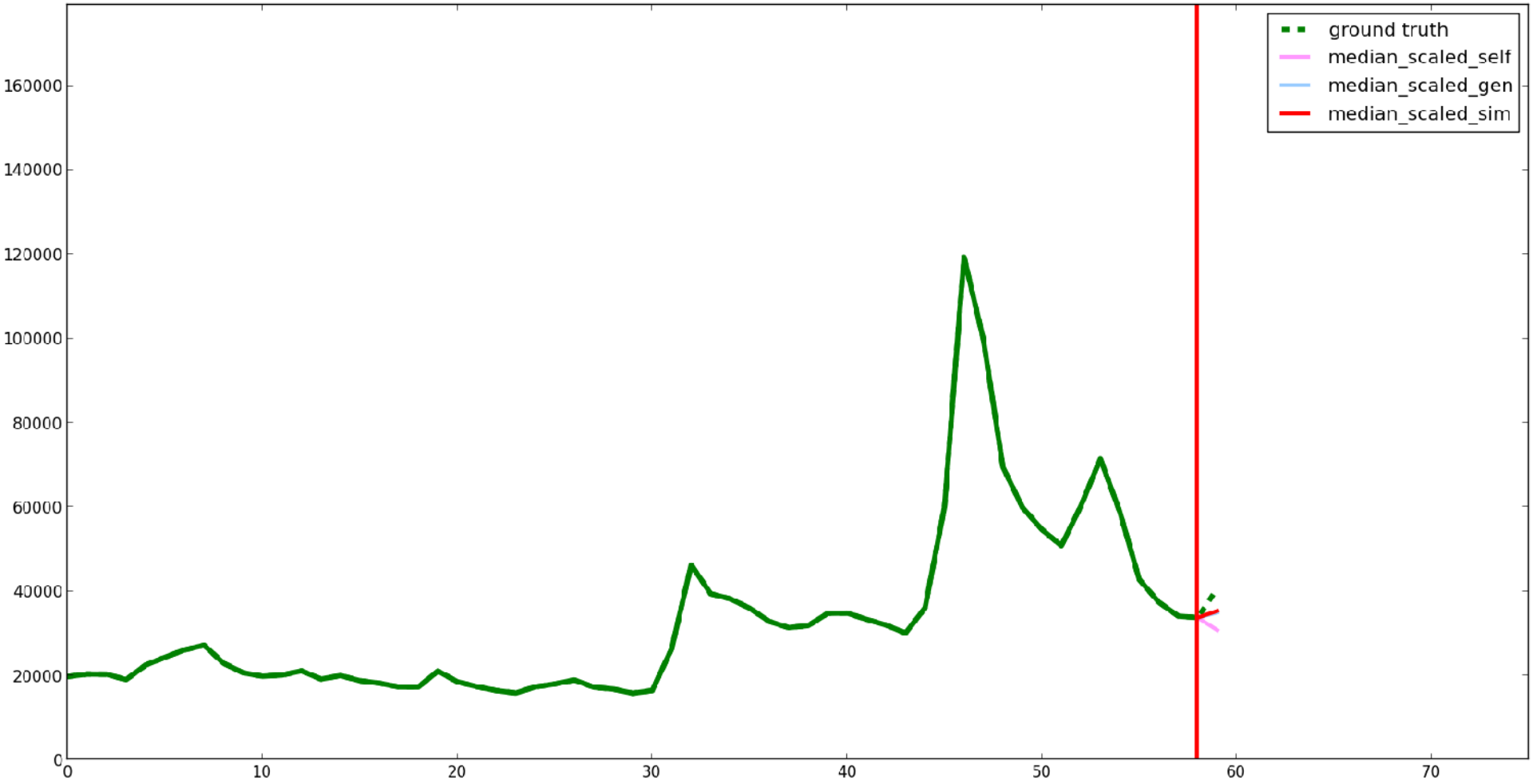
Example: “The Hunger Games”



Example: “The Hunger Games”



Example: “The Hunger Games”



Quantitative Results

Root Mean Square Error

	Method	RMSEs in 1000				
		$\tau = 0$	$\tau = 3$	$\tau = 5$	$\tau = 7$	$\tau = 9$
Baselines	naive					
	linear trend					
	average trend					
	median trend					
ARIMA	AR(1)					
	AR(2)					
	ARMA(1,1)					
	AutoARIMA					
Self	average					
	average_scaled					
	median					
	median_scaled					
Gen	average					
	average_scaled					
	median					
	median_scaled					
Sim	average					
	average_scaled					
	median					
	median_scaled					

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2}$$

Root Mean Square Error

	Method	RMSEs in 1000				
		$\tau = 0$	$\tau = 3$	$\tau = 5$	$\tau = 7$	$\tau = 9$
Baselines	naive					
	linear trend					
	average trend					
	median trend					
ARIMA	AR(1)					
	AR(2)					
	ARMA(1,1)					
	AutoARIMA					
Self	average					
	average_scaled					
	median					
	median_scaled					
Gen	average					
	average_scaled					
	median					
	median_scaled					
Sim	average					
	average_scaled					
	median					
	median_scaled					

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2}$$

Root Mean Square Error

	Method	RMSEs in 1000				
		$\tau = 0$	$\tau = 3$	$\tau = 5$	$\tau = 7$	$\tau = 9$
Baselines	naive	63.2	33.1	20.2	17.4	11.4
	linear trend	86.9	48.5	28.3	23.1	14.5
	average trend	49.3	25.9	22.0	19.9	18.3
	median trend	48.1	24.9	20.6	18.1	16.1
ARIMA	AR(1)	50.1	27.8	20.1	15.9	12.7
	AR(2)	75.1	31.7	22.6	16.0	13.4
	ARMA(1,1)	53.0	28.7	20.5	15.8	13.2
	AutoARIMA	58.9	30.7	26.9	19.5	16.7
Self	average	46.0	23.7	19.7	18.0	16.6
	average_scaled	44.6	21.9	17.6	15.5	13.8
	median	46.1	23.8	19.7	17.7	16.0
	median_scaled	44.9	22.3	18.1	15.5	14.4
Gen	average	45.7	22.9	19.2	16.1	14.1
	average_scaled	45.7	22.5	16.0	14.1	11.4
	median	41.4	21.2	17.6	15.4	12.9
	median_scaled	40.1	19.5	15.2	12.8	10.2
Sim	average	41.4	18.8	16.0	14.0	12.3
	average_scaled	39.6	17.1	13.7	11.6	10.0
	median	42.1	19.9	16.5	14.0	12.5
	median_scaled	41.0	17.9	14.2	11.5	9.8

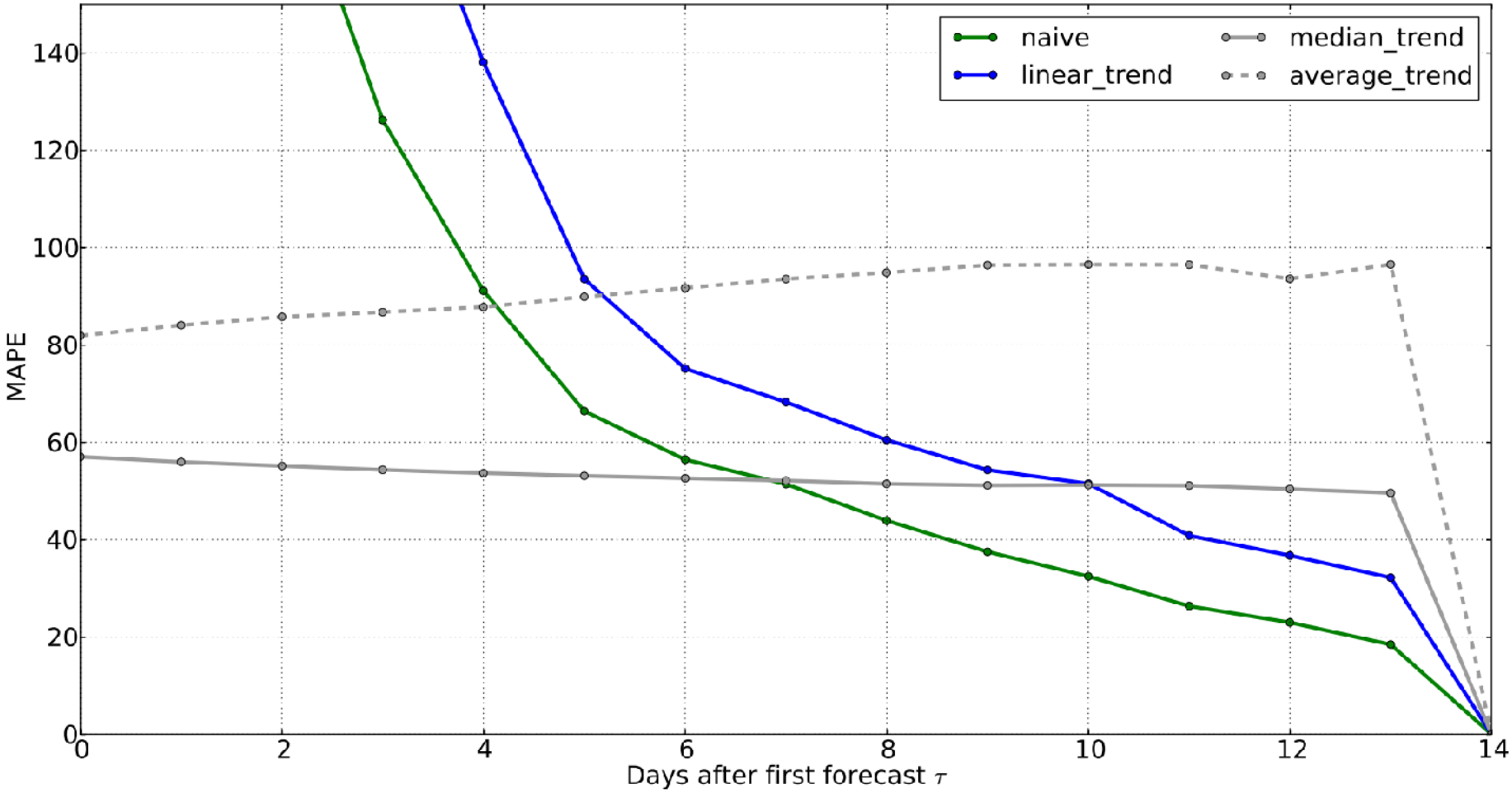
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2}$$

Mean Average Percentage Error

$$\mathrm{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

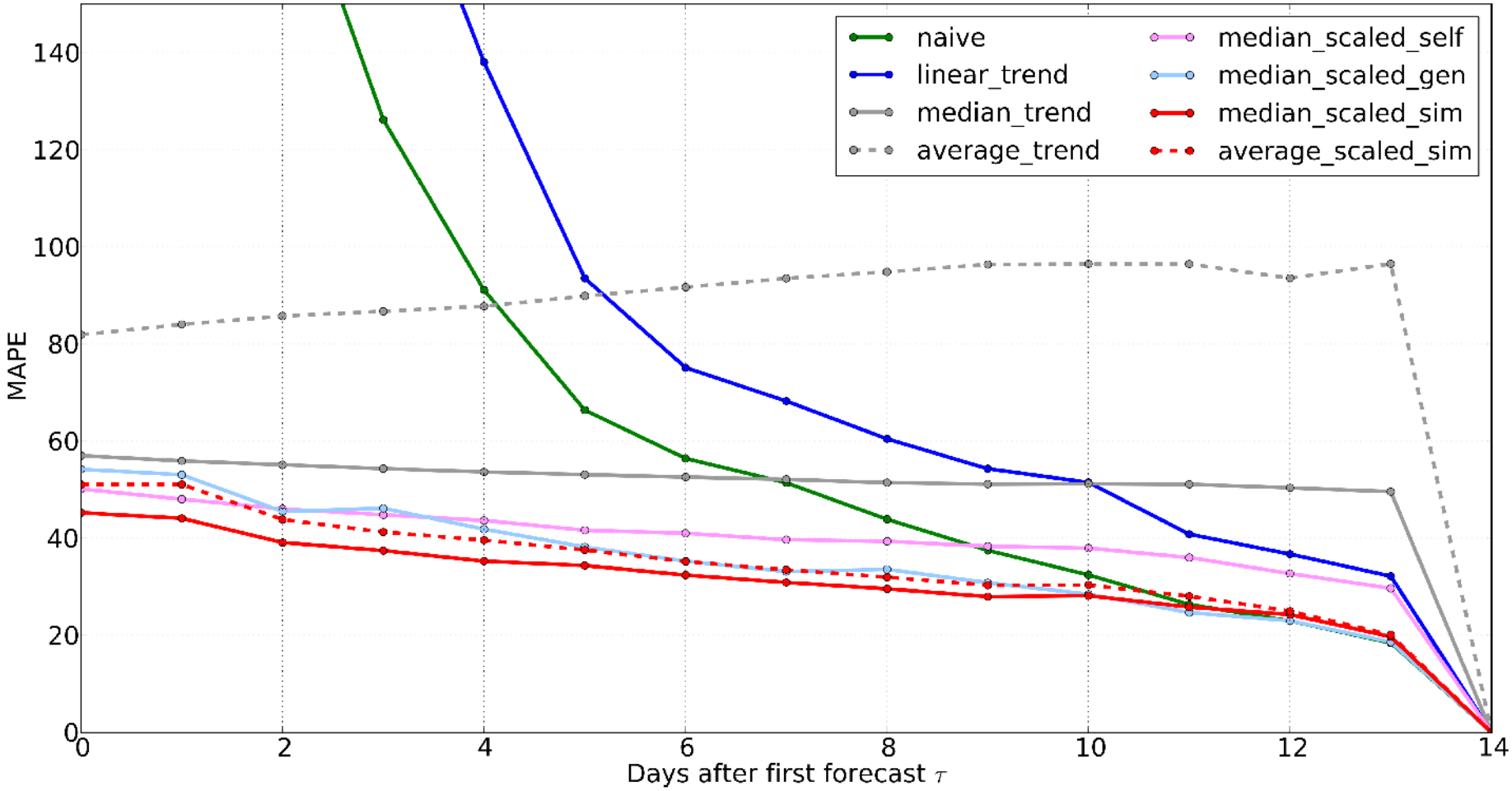
Mean Average Percentage Error

$$\mathrm{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$



Mean Average Percentage Error

$$\mathrm{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$



Conclusion

Main Contributions

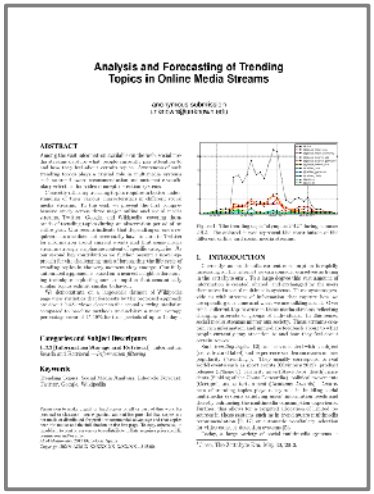
1. Multi-channel analysis of trending topics
2. Fully automatic forecasting technique exploiting semantic relationships between topics
3. Empirical evaluation on a large-scale Wikipedia dataset

Insights

1. Trends on Twitter and Wikipedia are significantly more ephemeral than on Google.
2. All observed media channels tend to specialize in specific topic categories.
3. Semantically similar topics exhibit similar behavior.
4. Exploiting this observation can significantly improve forecasting performance (by 20-90% compared to baselines).

Acknowledgments

PiCloud
CLOUD COMPUTING. SIMPLIFIED


Analysis and Forecasting of Trending Topics in Online Media Streams

ABSTRACT

Keywords

1. INTRODUCTION

Tim Althoff, Damian Borth, Jörn Hees, and Andreas Dengel.
Analysis and Forecasting of Trending Topics in Online Media Streams.
Submitted to ACM Multimedia, 2013.

Questions?

Thank you for your attention!

Backup Slides

Motivation

- **Big Data**: “High volume, high velocity, and/or high variety **information assets** that require new forms of processing to enable **enhanced decision making**, **insight discovery** and process optimization”

- Large-scale user behavior
  - „Mirror of Society“

[Mark A. Beyer and Douglas Laney. The importance of ’big data’: A definition. June 2012.]

Large-scale Record of Human Behavior

- 1.2 million video minutes per second (est. 2016)

- Primary method for communication by college students in the U.S.

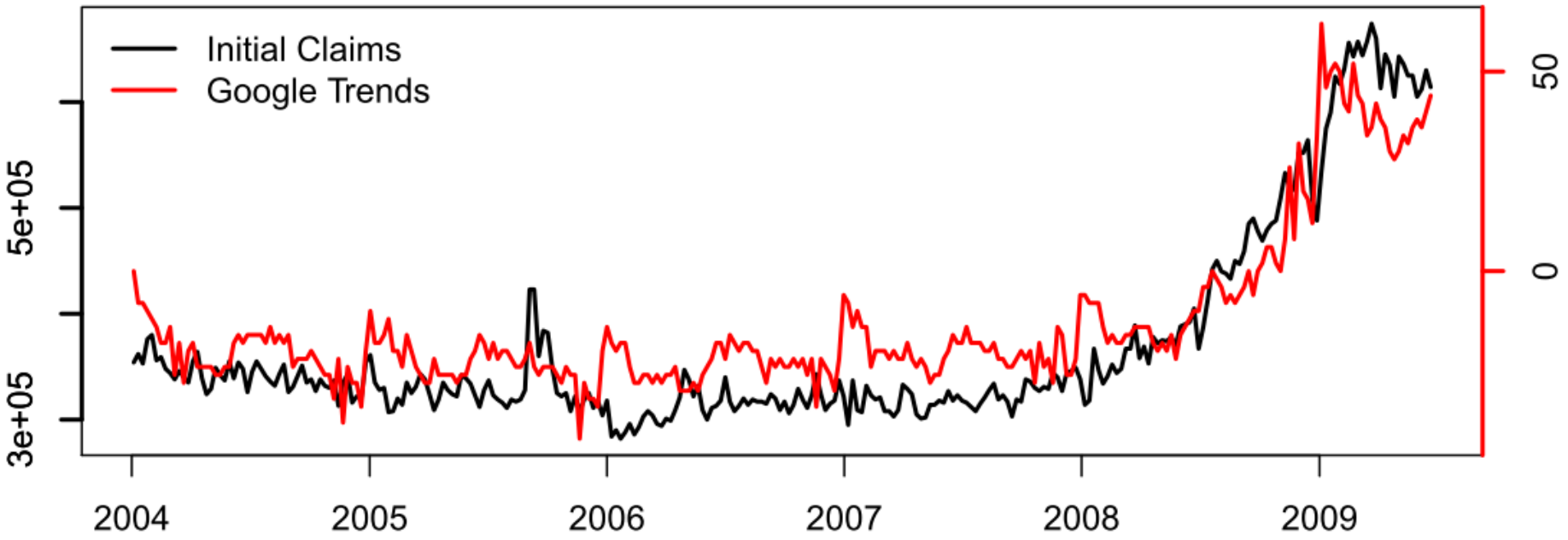
- 25 million articles and 1.8 billion page edits in 285 languages created by 100k contributors

- 340 million tweets each day by 500 million users

- One billion search requests and about twenty petabytes of user-generated data each day (2009)

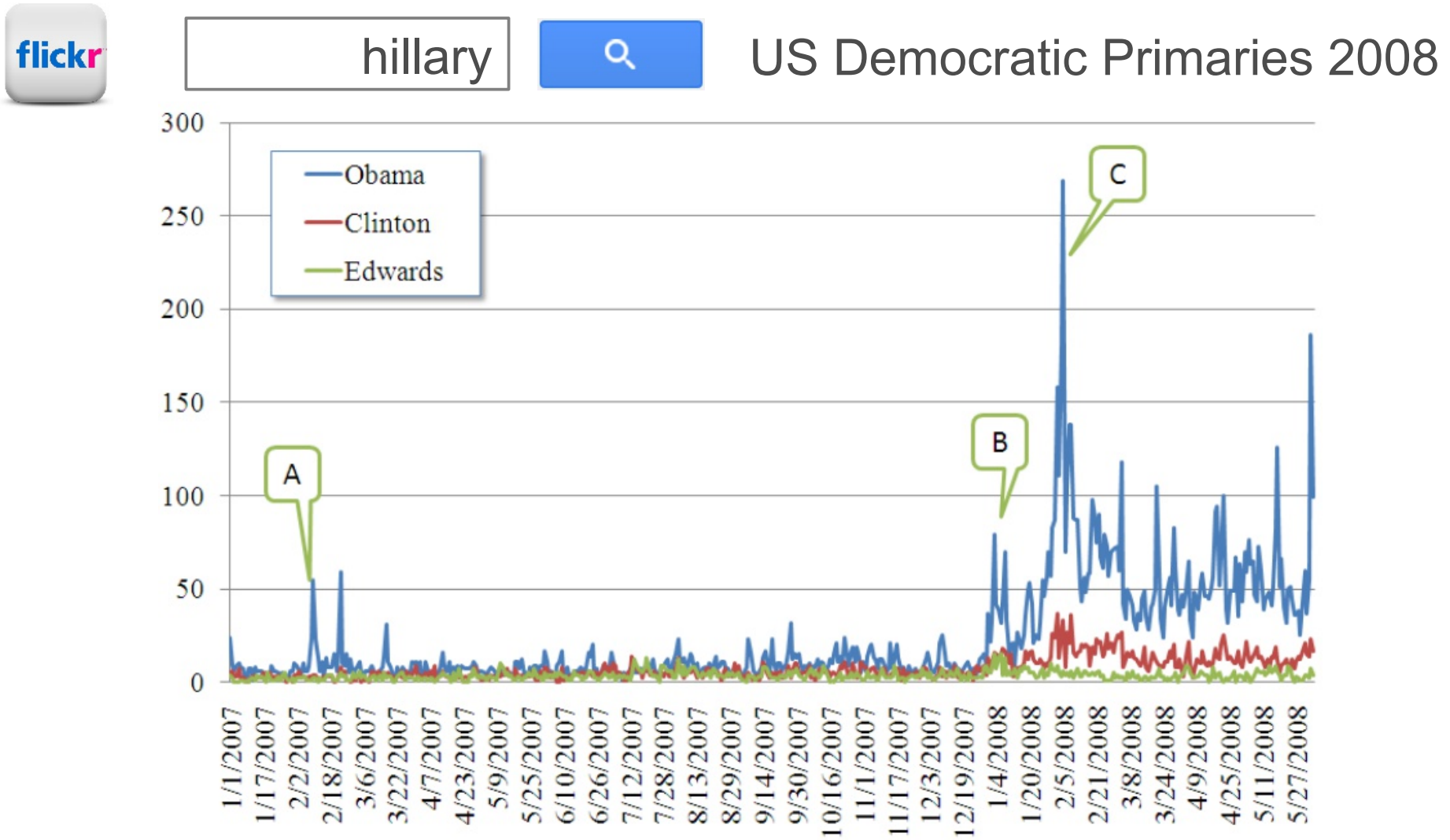
Unemployment Claims

jobs



[Choi and Varian. Predicting initial claims for unemployment benefits. Google, Inc. 2009.]

Political Elections



[Jin et al. The wisdom of social multimedia: using Flickr for prediction and forecast. ACM Multimedia 2010.]

What is the world talking about?

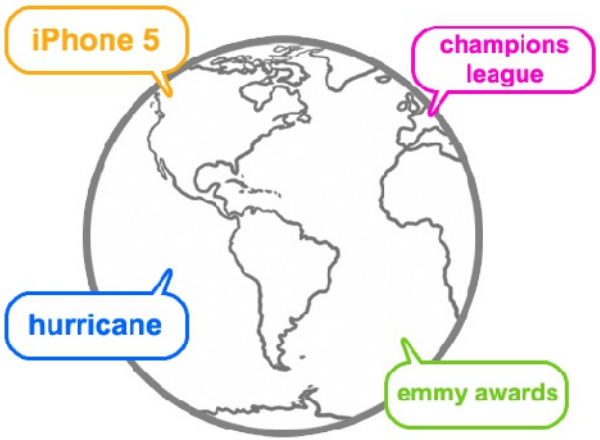
Social Media can offer insights in

- what the world is talking about
- how people feel about events, brands and products

resulting in an collective awareness for **trending topics**

A **trending topic**

- is a subject matter of discussion agreed on by a group of people
- experiences a sudden spike in user interest or engagement



[Damian Borth, Adrian Ulges, Thomas Breuel. Dynamic Vocabularies for Web-based Concept Detection by Trend Discovery. ACM Multimedia, 2012]

Trending Topics Discovery

Wednesday, 12. Oct 2011

Clustering

Cluster: apple
- apple
- ios 5
- lions 5-0

Cluster: phillies
- peyton hillis
- phillies

Cluster: the avengers
- the avengers trailer
- the avengers

Cluster: iphone 4s
- iphone 4s
- iphone 4

Cluster: Julija Tymoschenko
- Yulia Tymoshenko
- Julija Tymoschenko

Cluster: republican debate
- new hampshire debate
- gop debate
- republican debate
- presidential debate
- republican presidential candidates

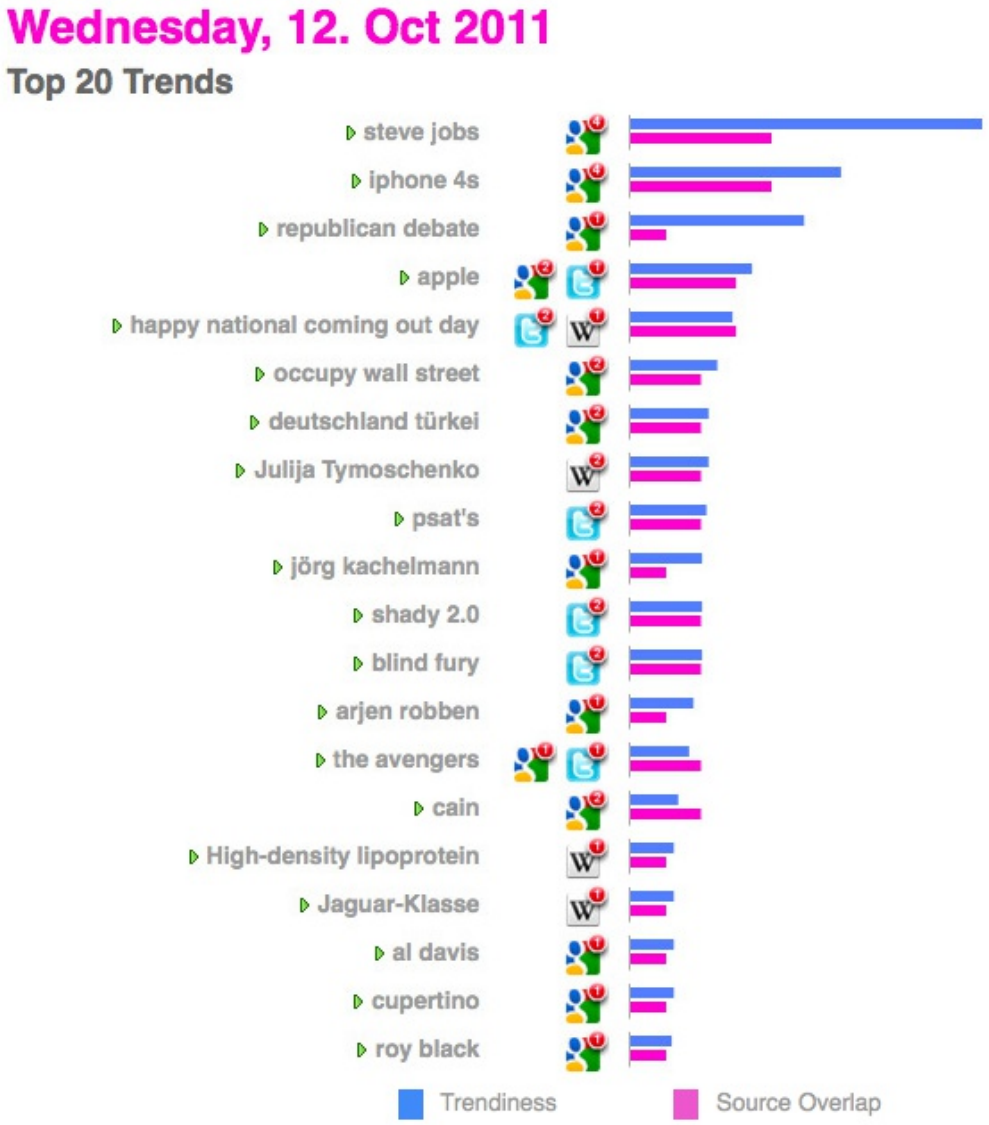
Cluster: occupy wall street
- occupy wall street
- occupy

Cluster: steve jobs
- steve jobs
- steve jobs dead
- steve jobs died
- jobs
- jobs steve

Cluster: happy national coming out day
- happy national coming out day
- national coming out day
- National Coming Out Day

Cluster: Phoenix Jones
- phoenix jones
- Phoenix Jones

Trending Topics Discovery



Clustering

Cluster: apple
- apple
- ios 5
- lions 5-0

Cluster: phillies
- peyton hillis
- phillies

Cluster: the avengers
- the avengers trailer
- the avengers

Cluster: iphone 4s
- iphone 4s
- iphone 4

Cluster: Julija Tymoschenko
- Yulia Tymoshenko
- Julija Tymoschenko

Cluster: republican debate
- new hampshire debate
- gop debate
- republican debate
- presidential debate
- republican presidential candidates

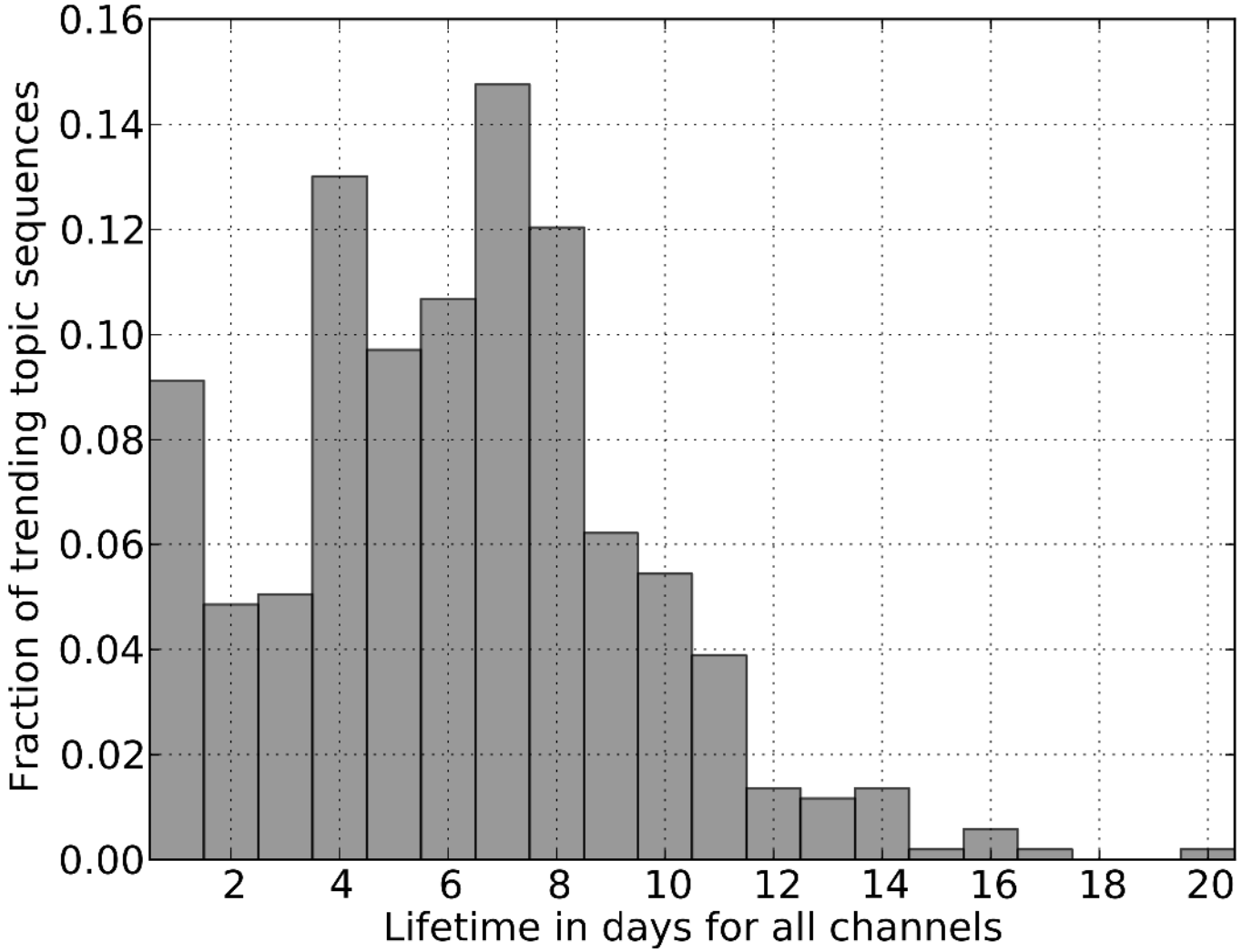
Cluster: occupy wall street
- occupy wall street
- occupy

Cluster: steve jobs
- steve jobs
- steve jobs dead
- steve jobs died
- jobs
- jobs steve

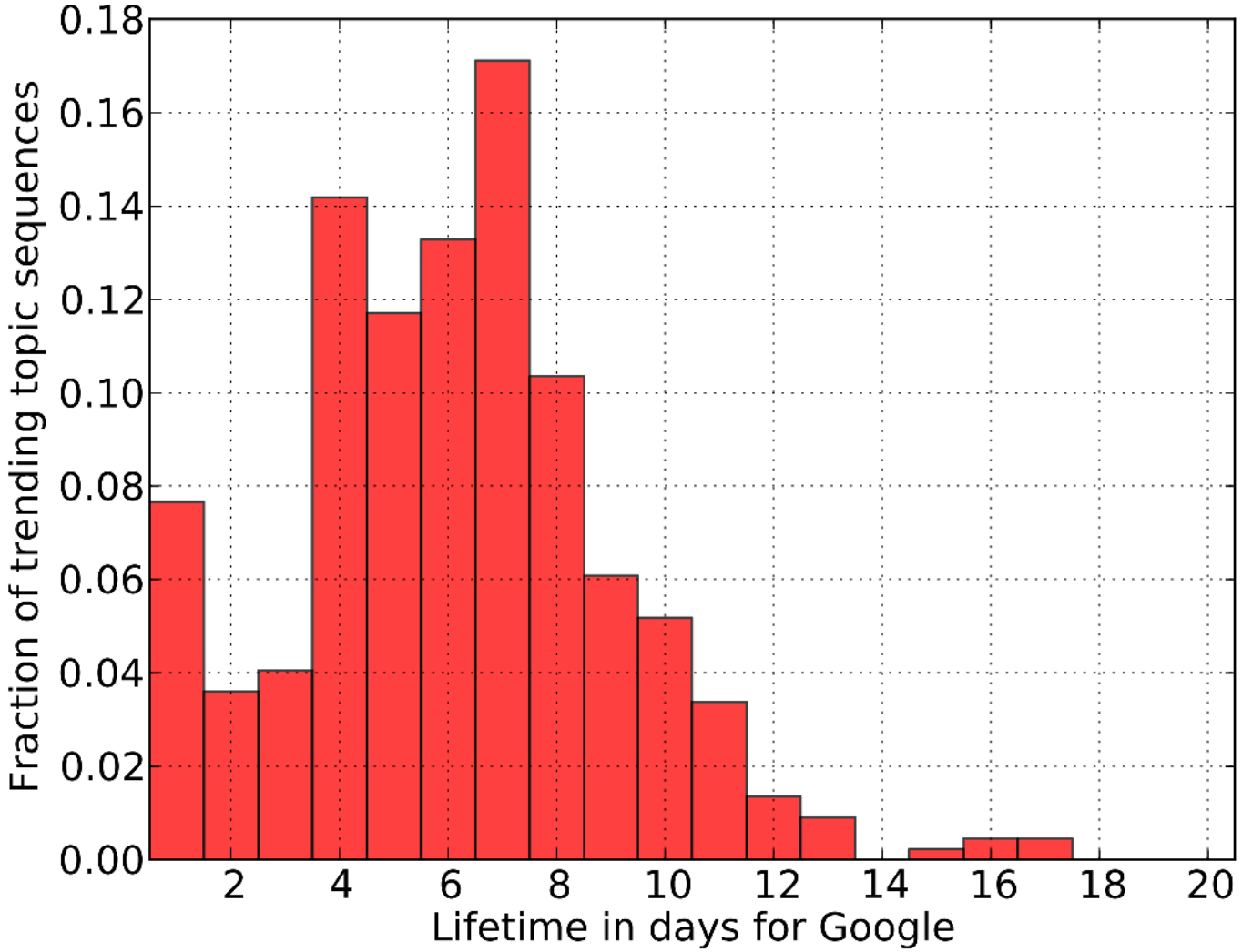
Cluster: happy national coming out day
- happy national coming out day
- national coming out day
- National Coming Out Day

Cluster: Phoenix Jones
- phoenix jones
- Phoenix Jones

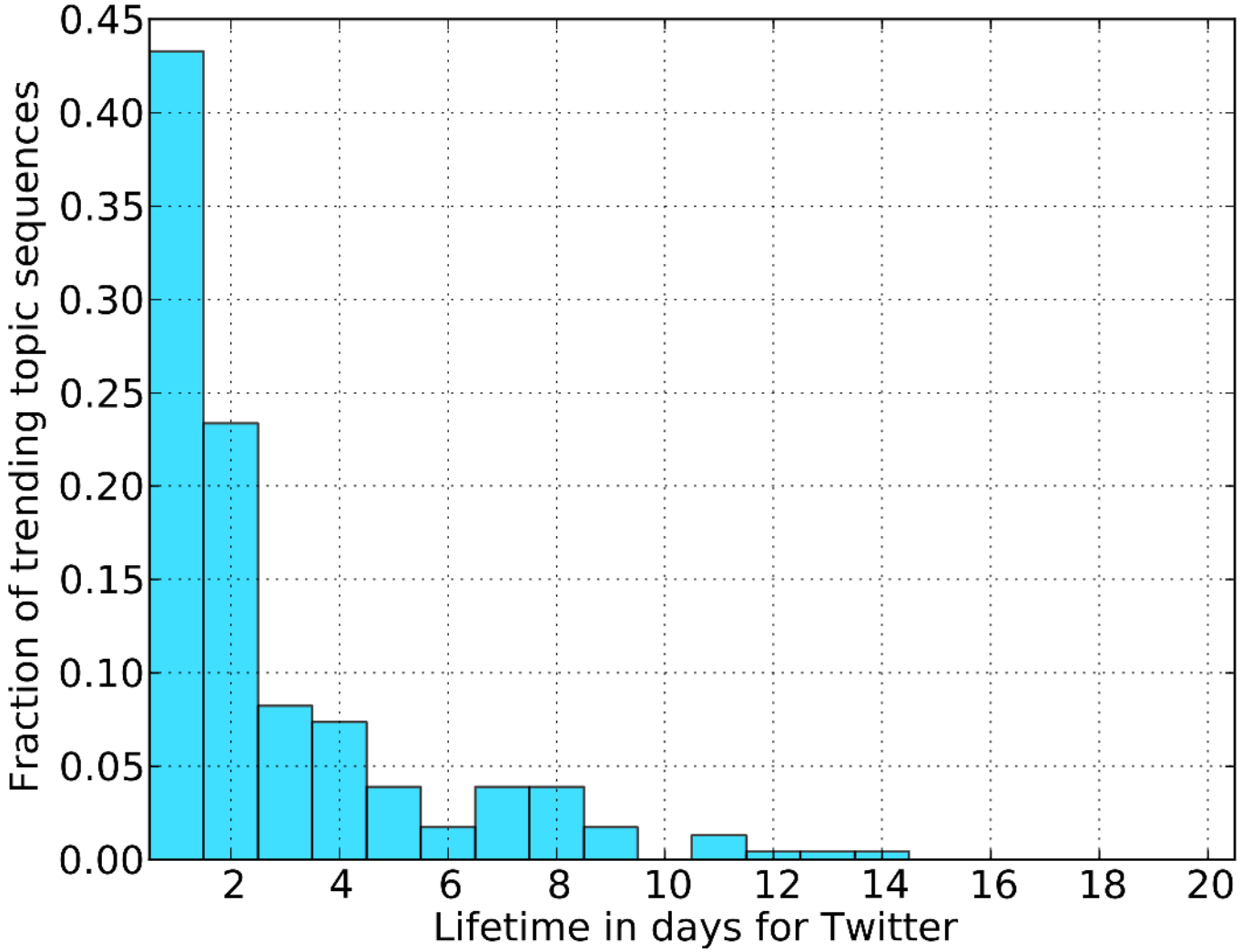
Lifetime Analysis



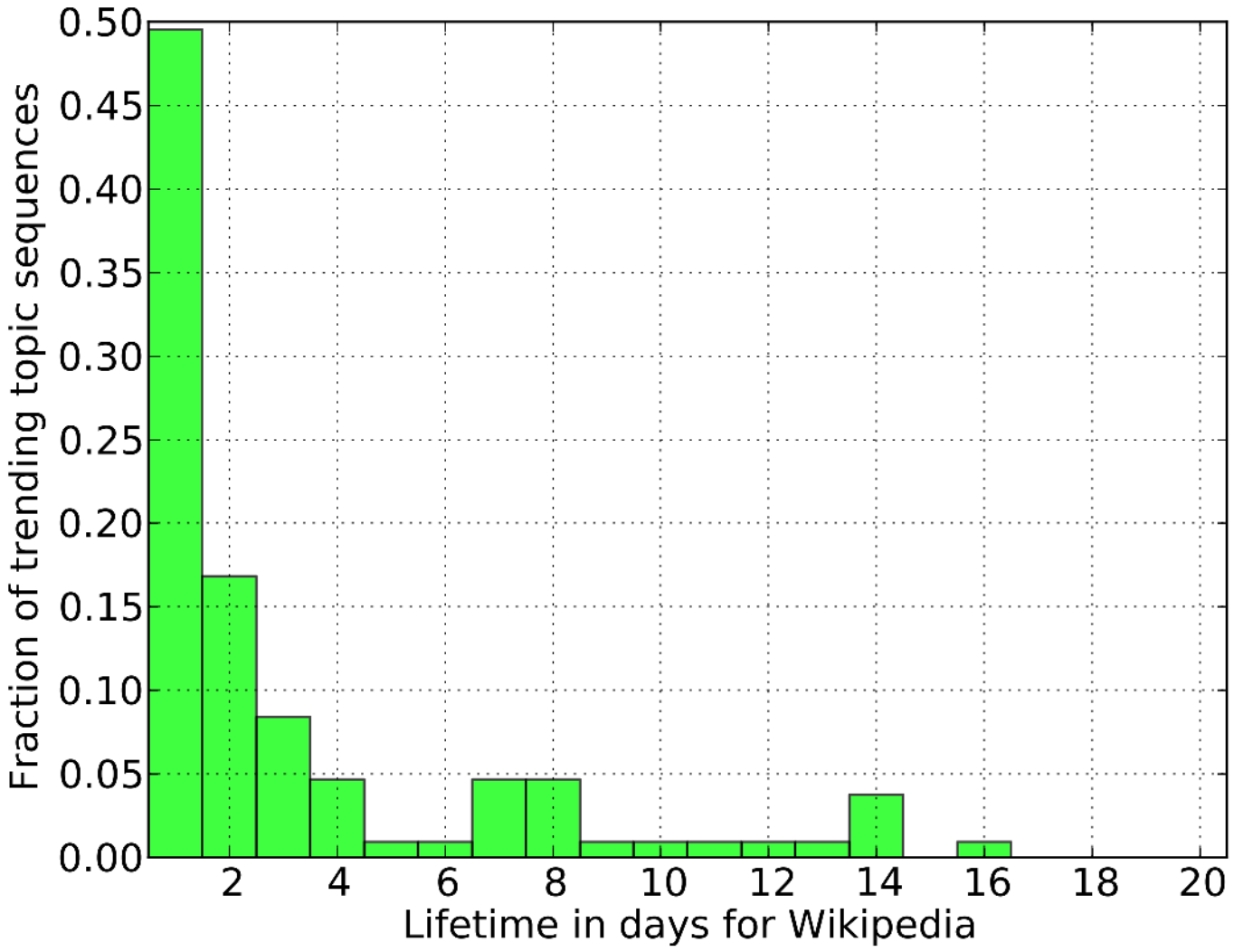
Lifetime Analysis



Lifetime Analysis



Lifetime Analysis



Basic Forecasting Techniques

Naive

$$X_t = X_{t-1}$$

Linear Trend

$$X_t = X_{t_0} + \frac{(X_{t_0} - X_{t_0-d})}{d} \cdot (t - t_0)$$

Average/Median

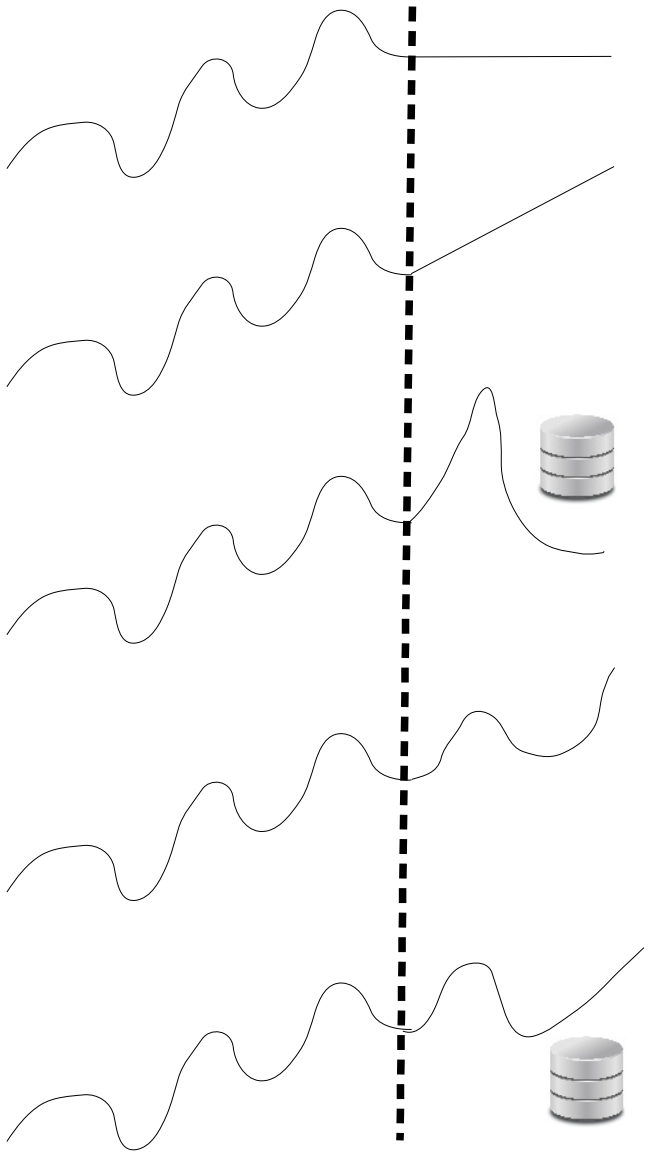
$$X_t = \text{average}\left(X_t^1, \ldots, X_t^n\right)$$

ARMA

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

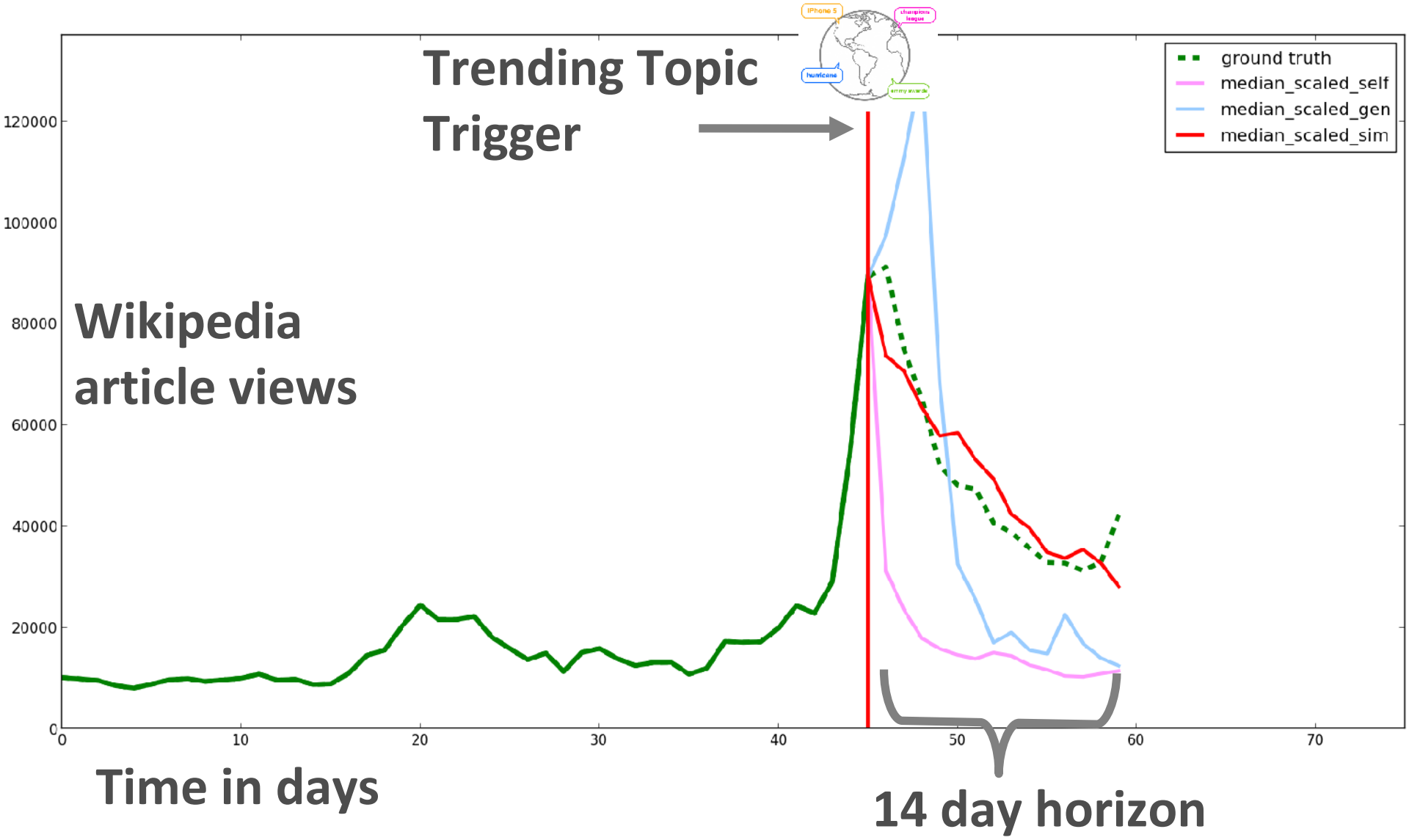
Nearest Neighbor

$$X_t = \text{operator}\left(X_{t_1}^1, \ldots, X_{t_k}^k\right)$$

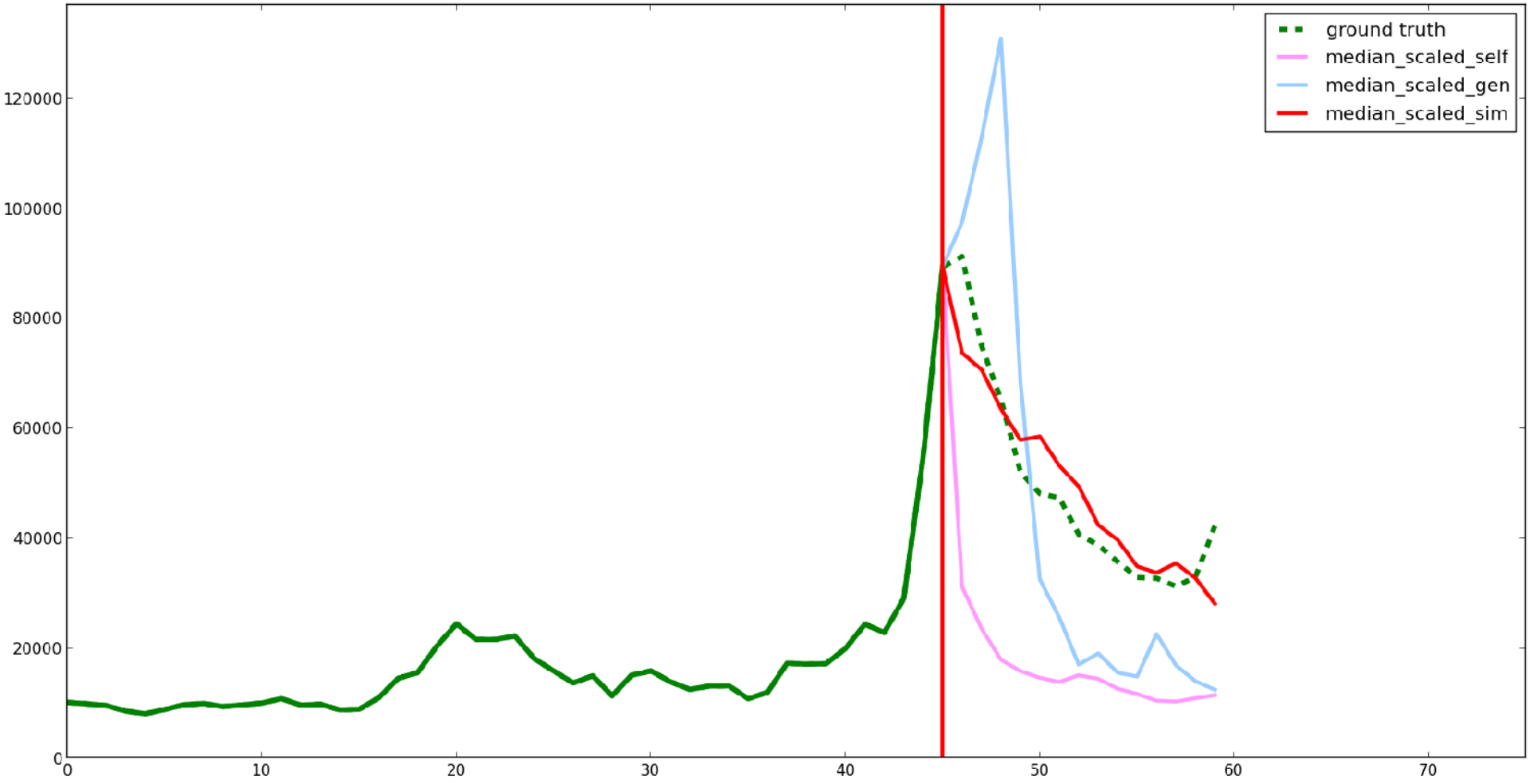


Example 2: “Battlefield 3”

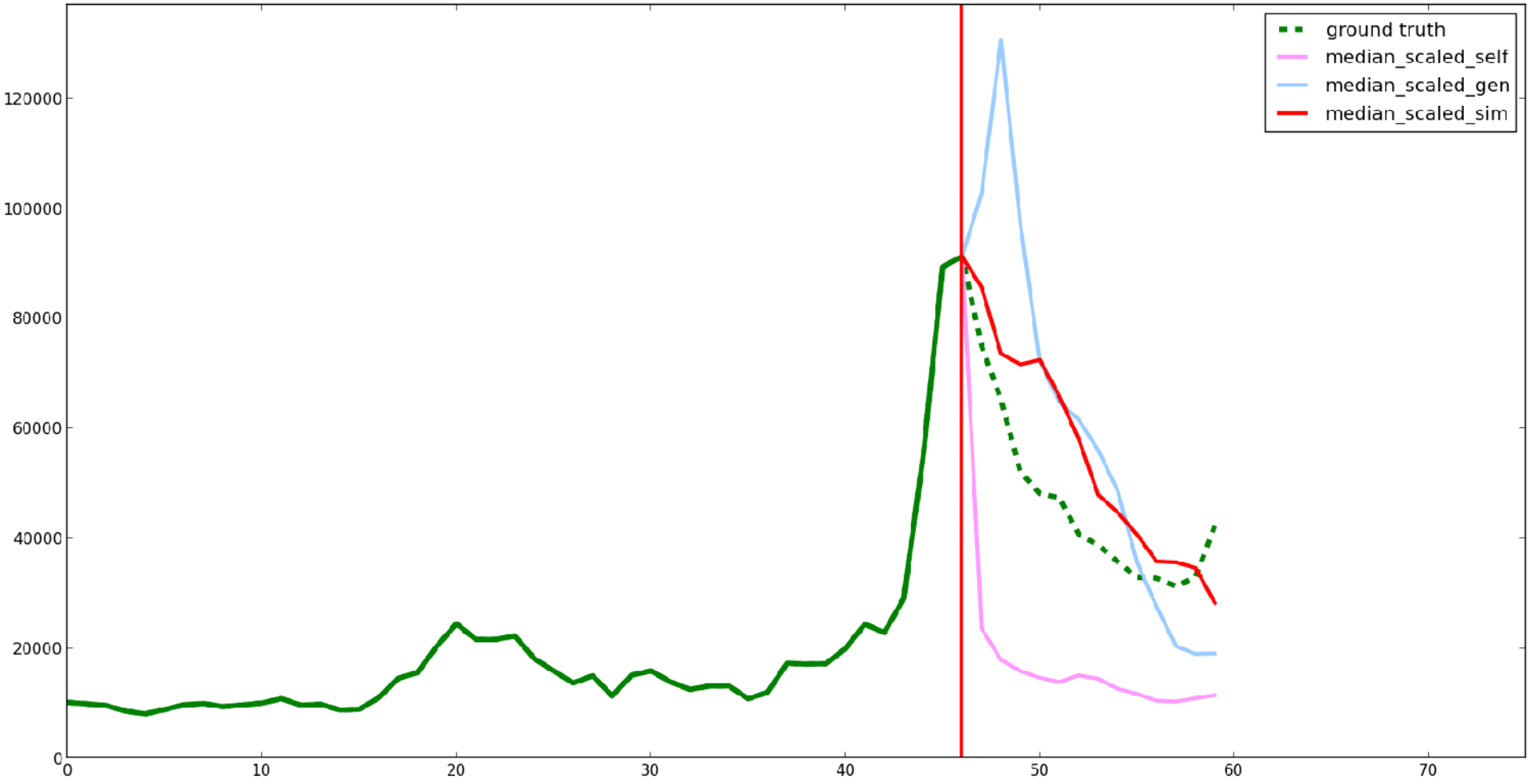
Example 2: “Battlefield 3”



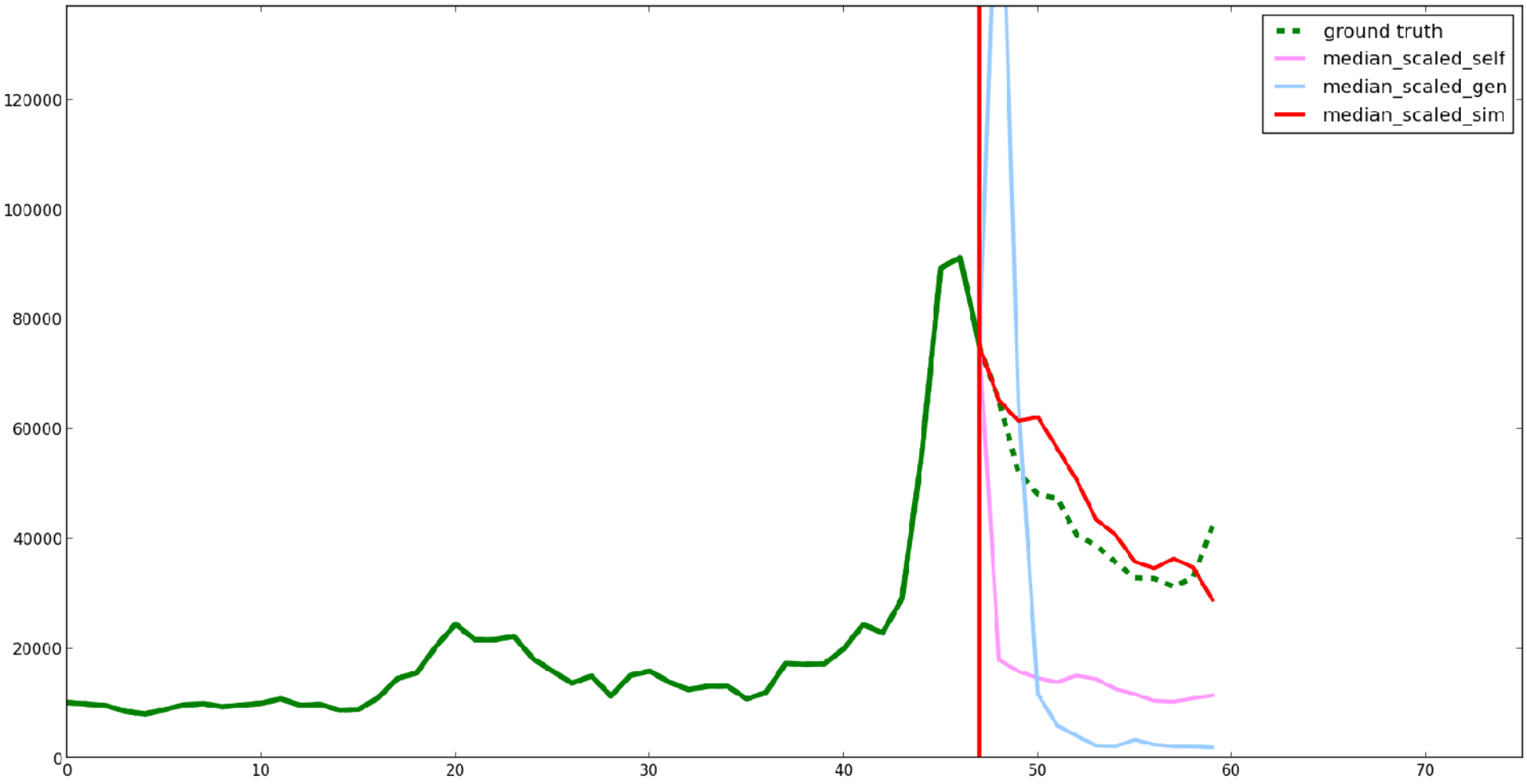
Example 2: “Battlefield 3”



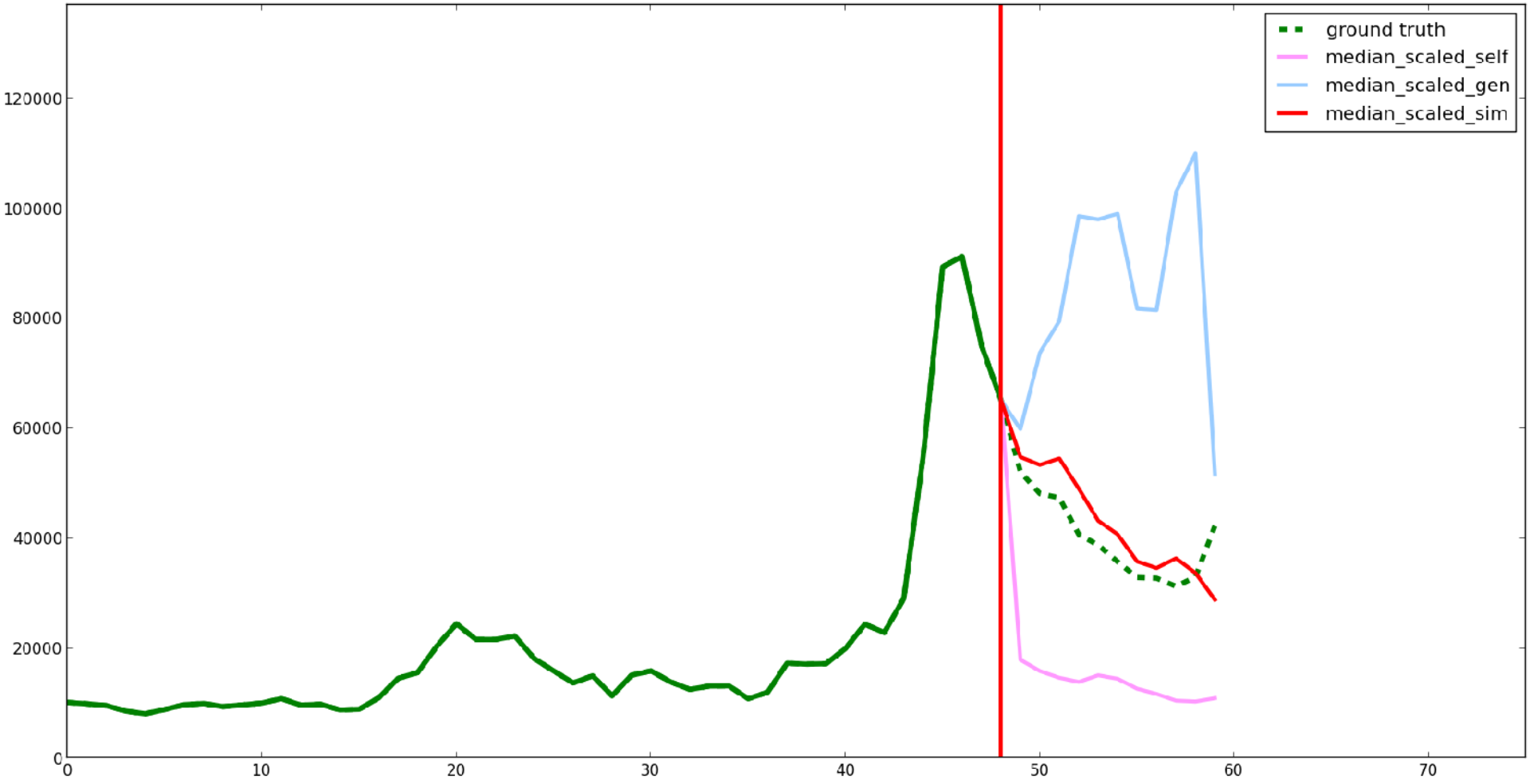
Example 2: “Battlefield 3”



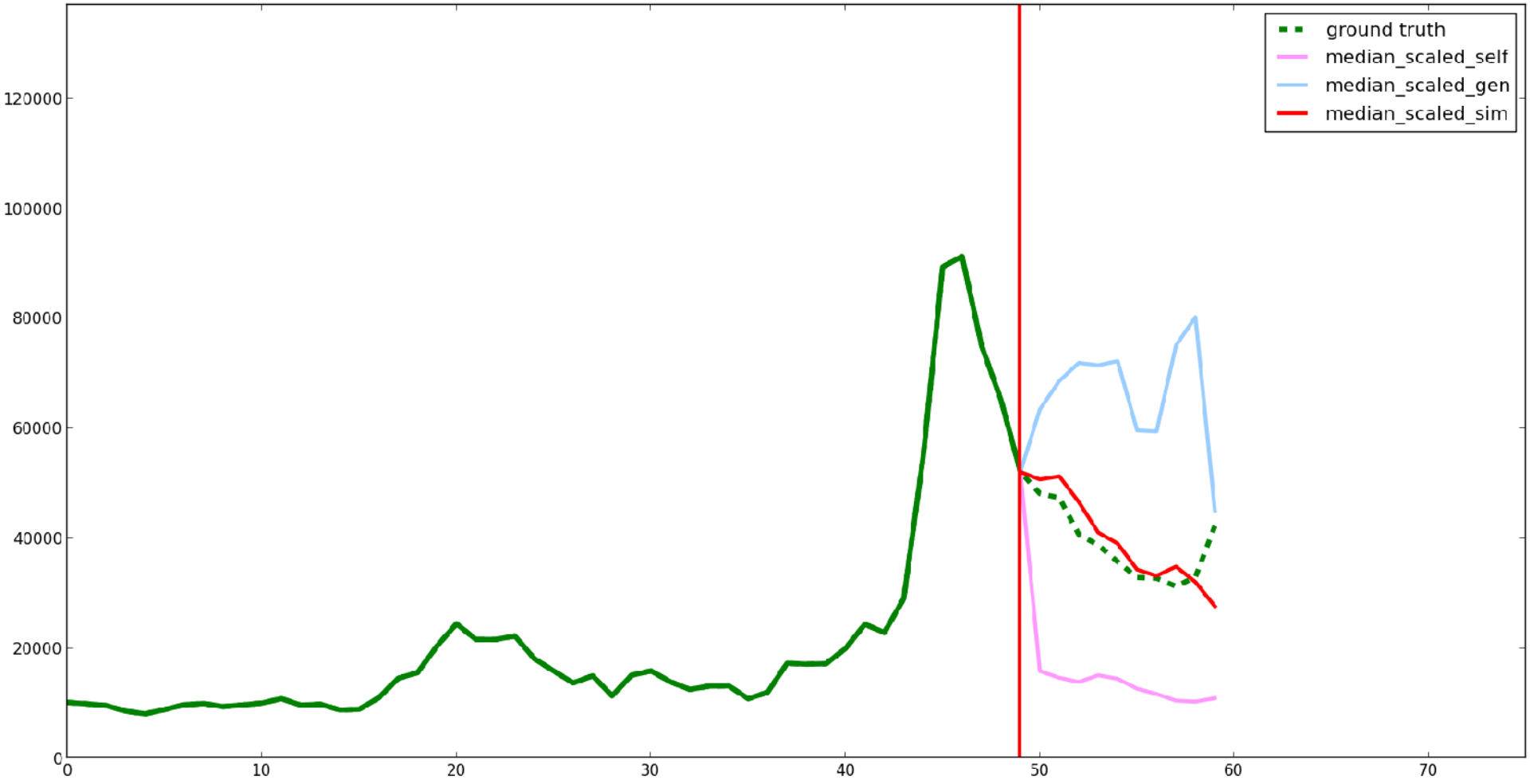
Example 2: “Battlefield 3”



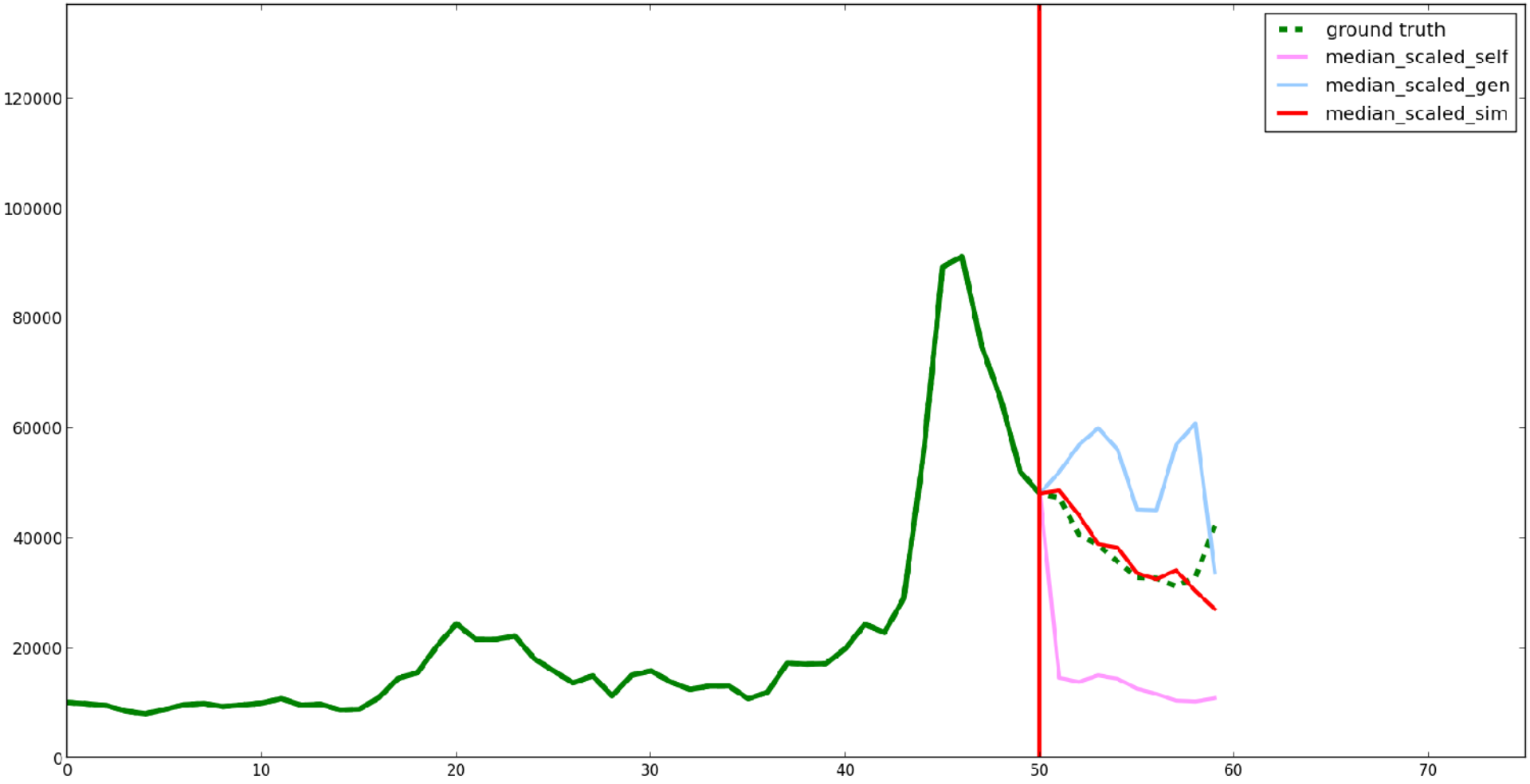
Example 2: “Battlefield 3”



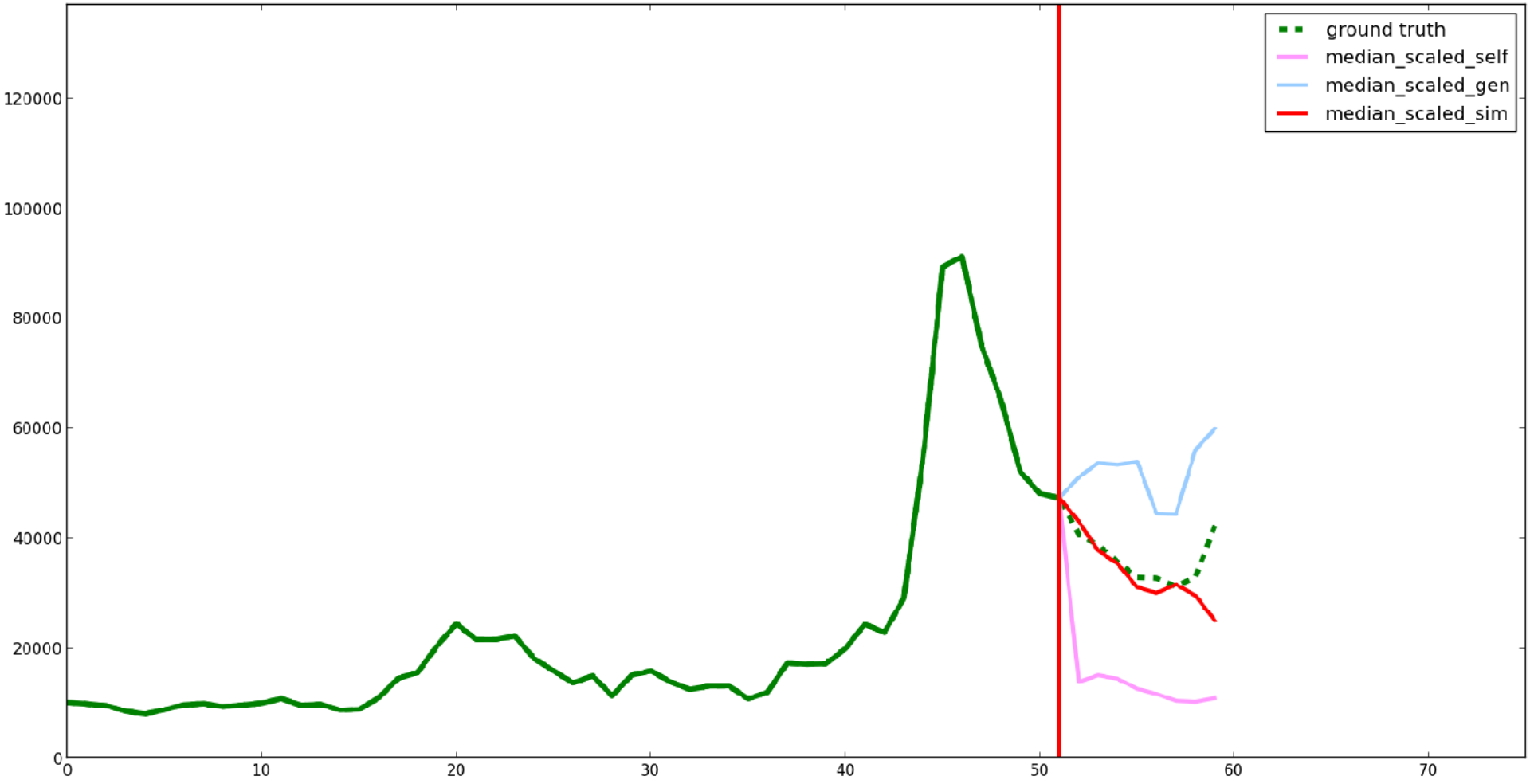
Example 2: “Battlefield 3”



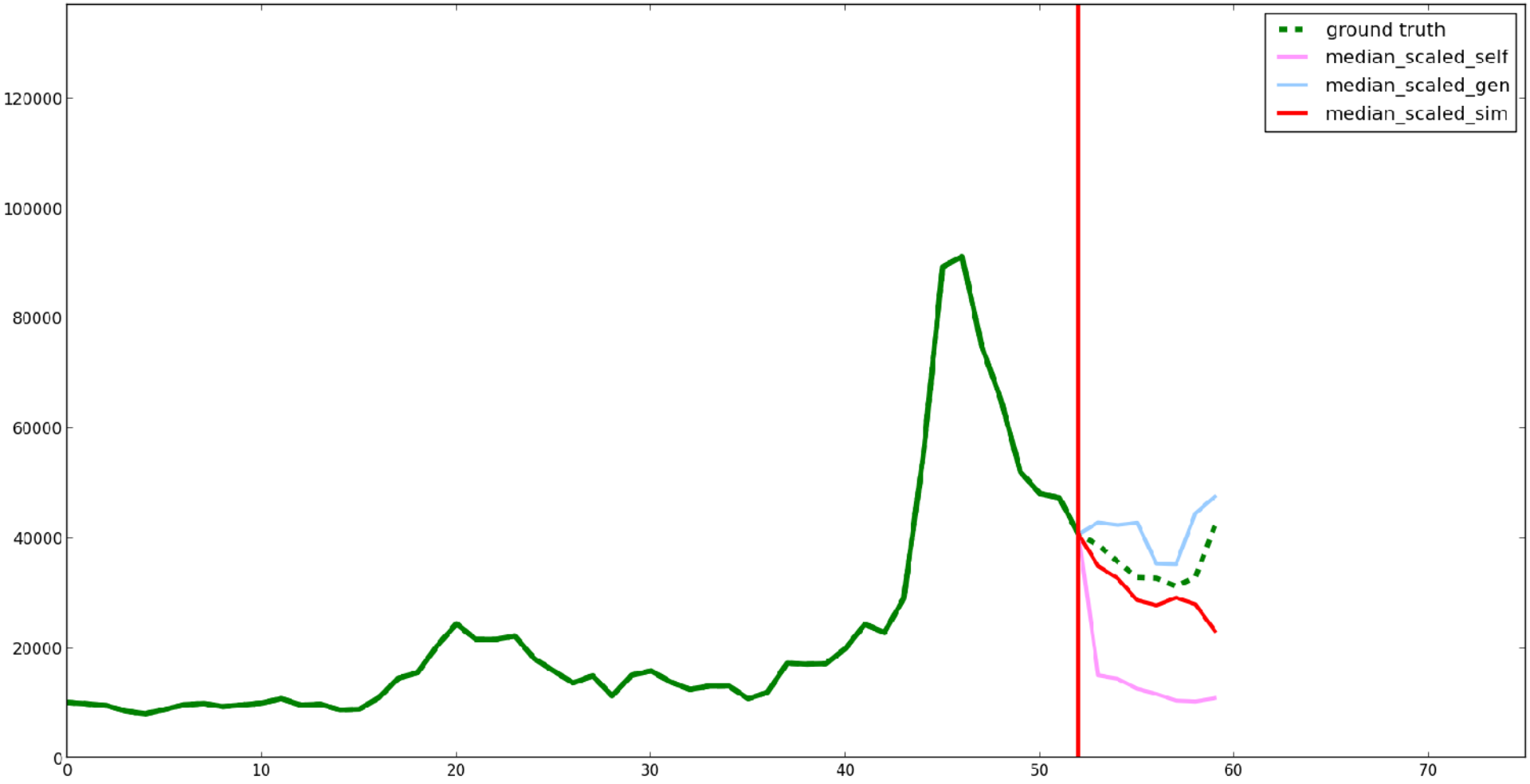
Example 2: “Battlefield 3”



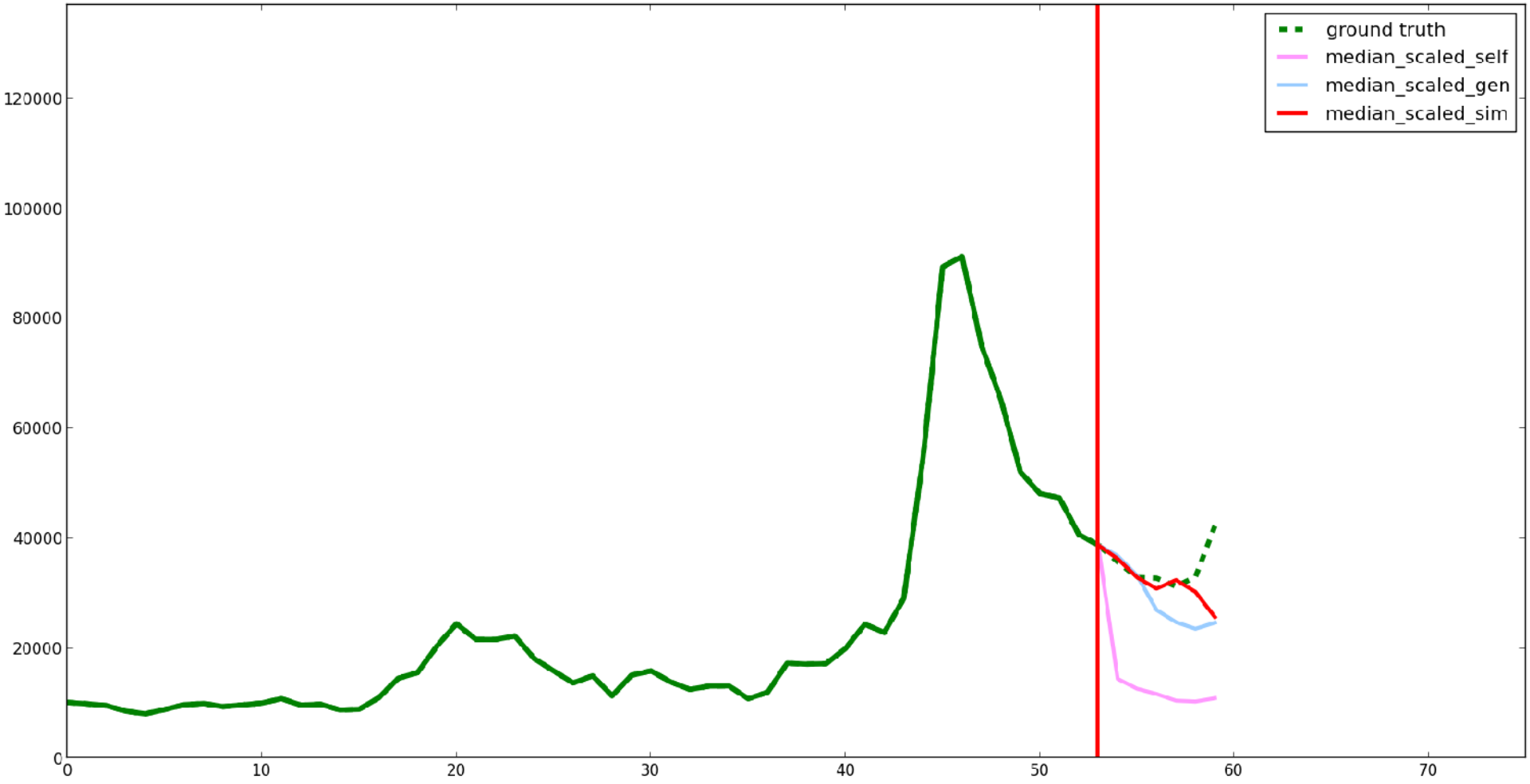
Example 2: “Battlefield 3”



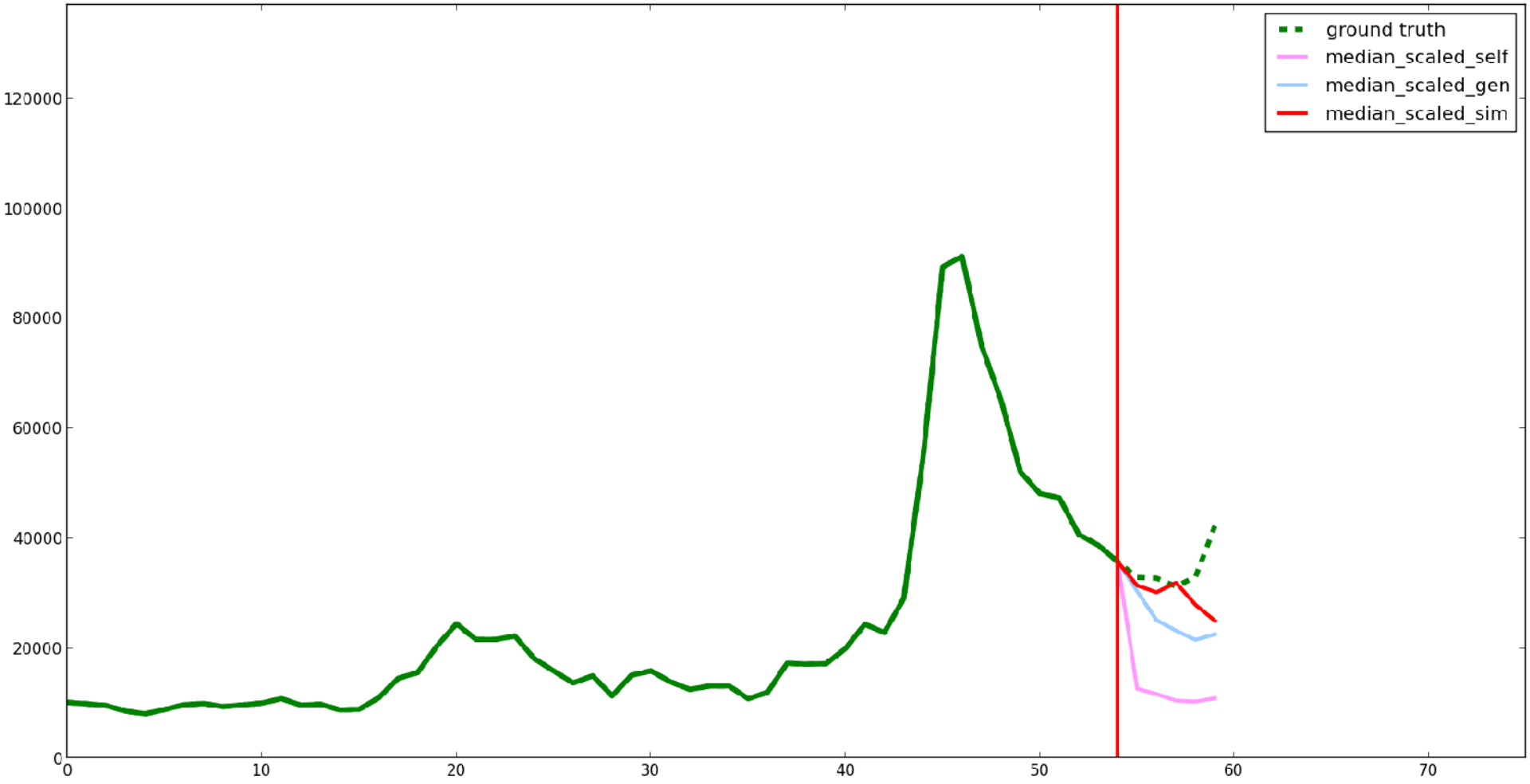
Example 2: “Battlefield 3”



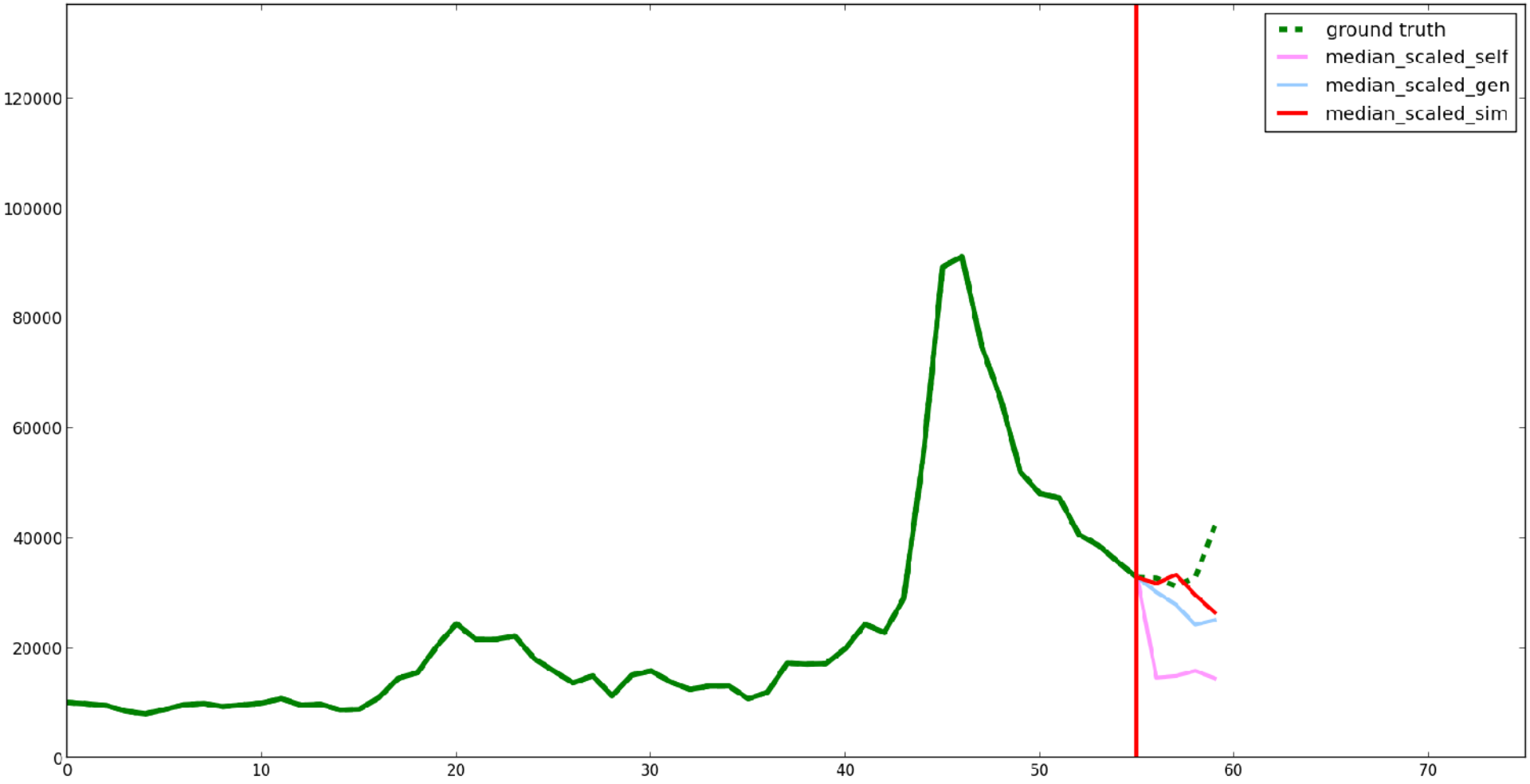
Example 2: “Battlefield 3”



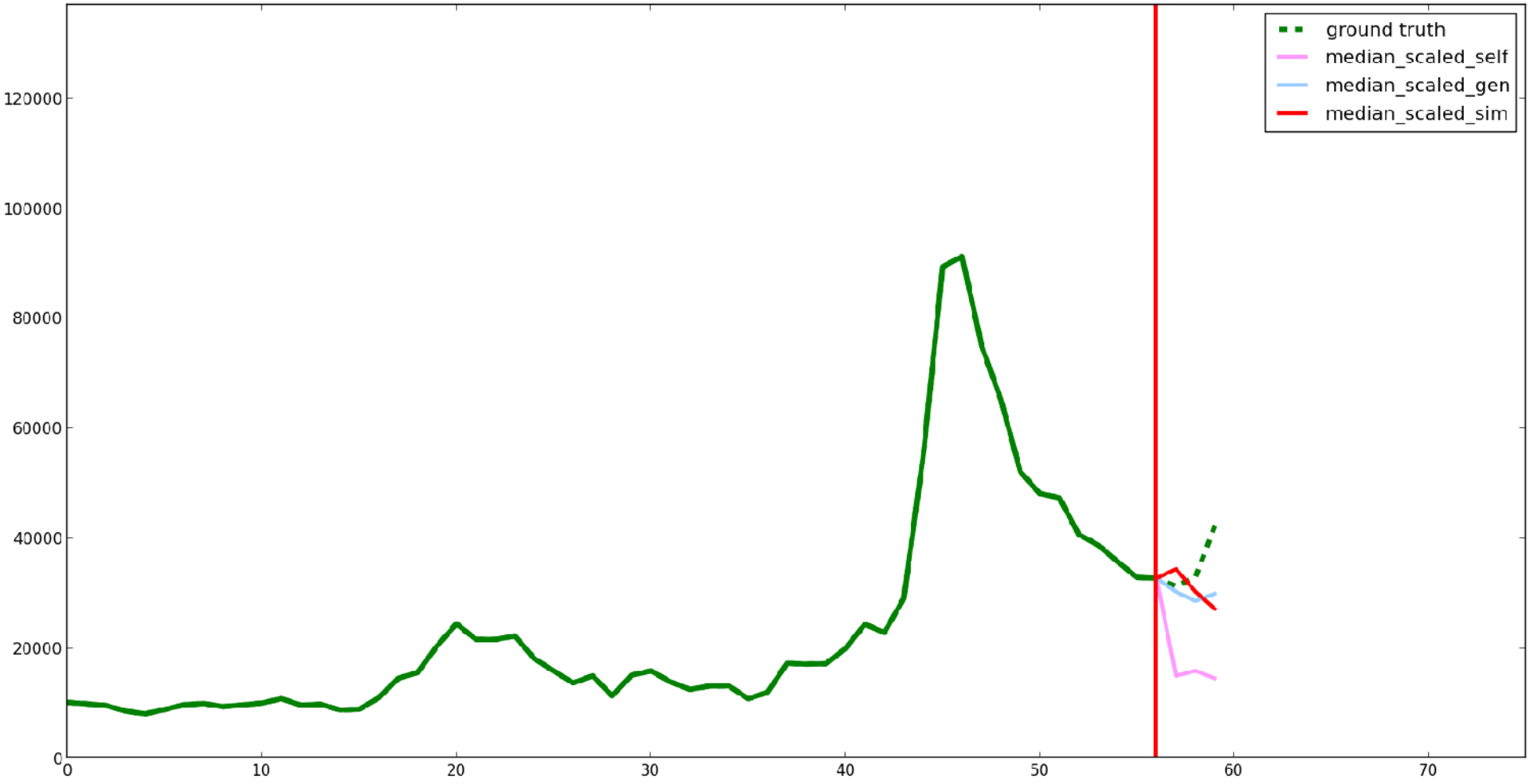
Example 2: “Battlefield 3”



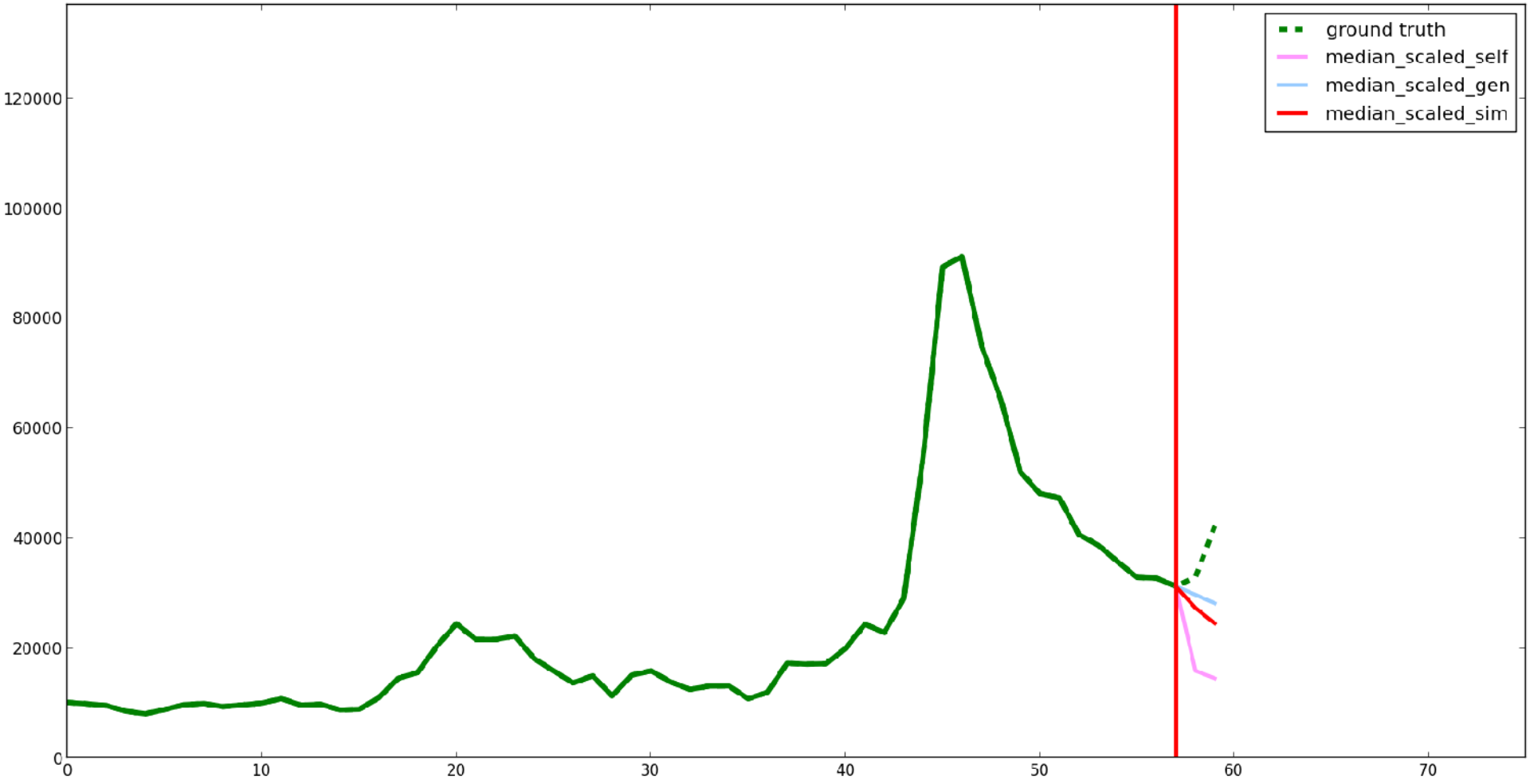
Example 2: “Battlefield 3”



Example 2: “Battlefield 3”



Example 2: “Battlefield 3”



Example 2: “Battlefield 3”