Large-Scale, Computationally Intensive Forecasting in R

Farzan Rohani, Eric Tassone,
and Murray Stokely

Google

Overview

"Large-Scale" = Lots of time series

- Manual analysis impractical
- Diversity of time series
  - Growth
  - Seasonality
  - Shocks, regime change, emergence
- Automatic and robust methods needed
  - Ensemble method (a.k.a., "Many Models" approach)

"Computationally Intensive"...

- Two sources
  - Thousands of forecasts every day
  - Arising from statistical method
    - Fitting the "many models"
    - Quantifying statistical uncertainty
      - Simulation-based confidence intervals (requiring 1K to 10K forecasts per time series)

"Forecasting in R"

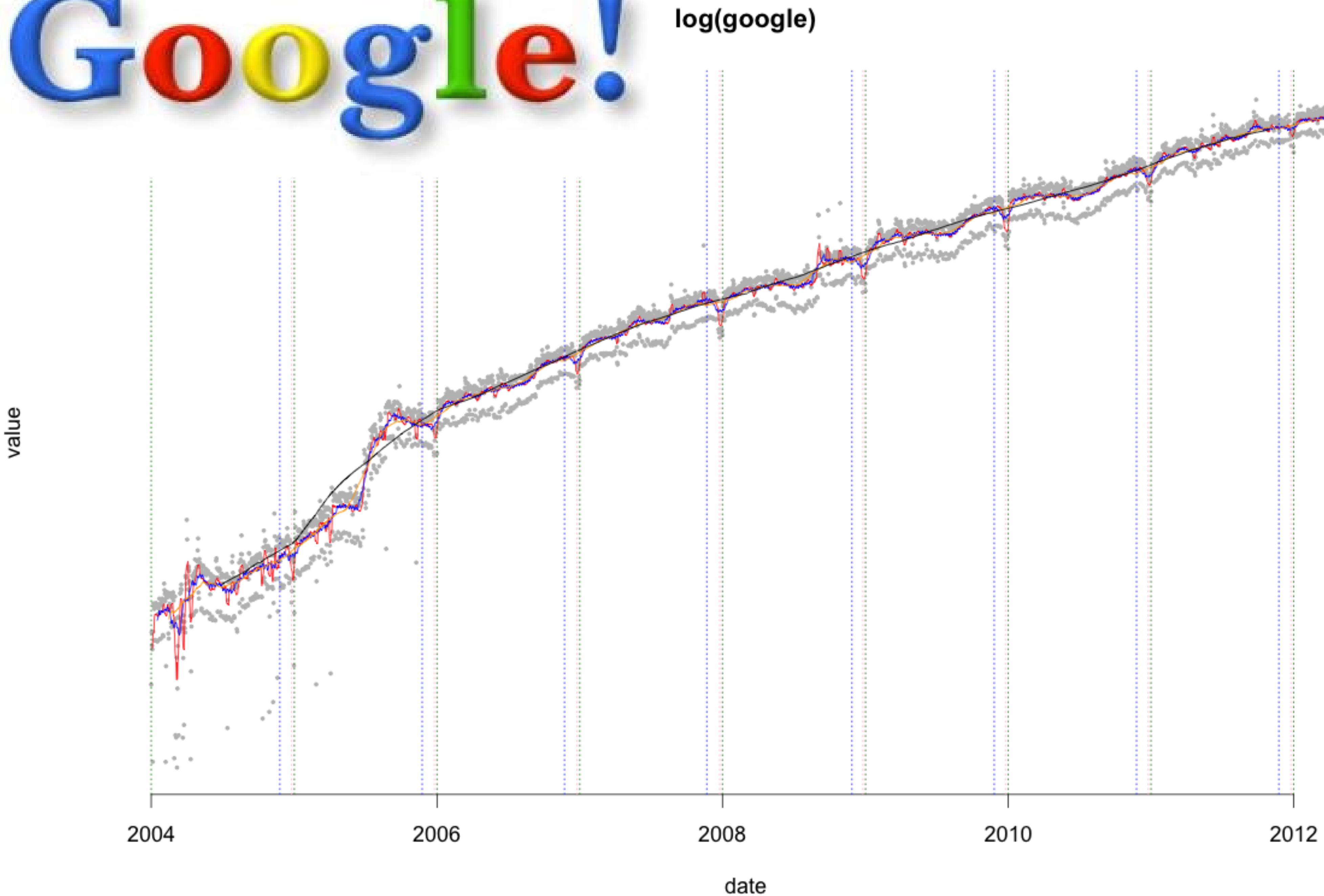
Just what it sounds like! :-)

Data

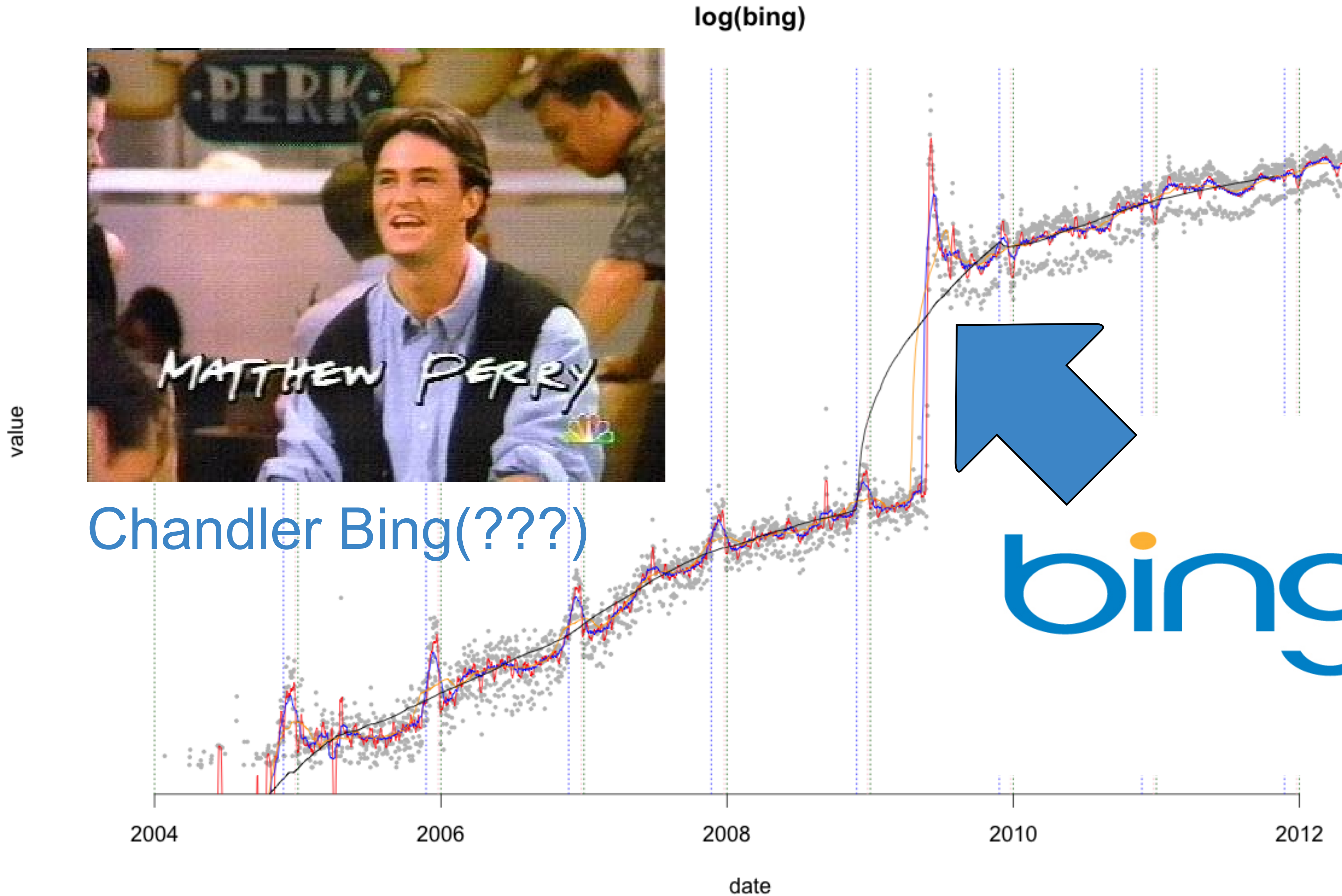
The Data

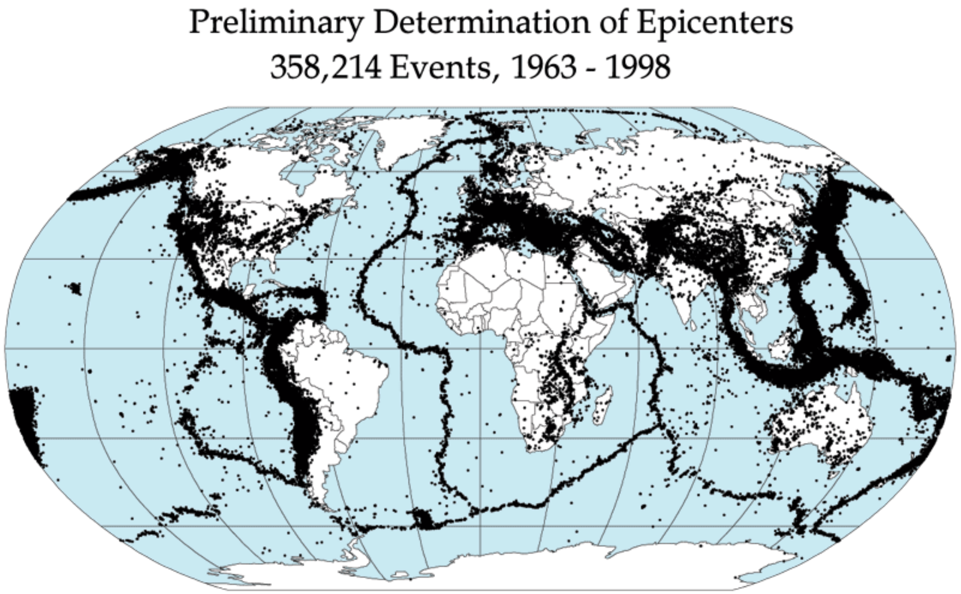
- 1000 time series from Google Trends:
  - Looked at past 5 years of data
  - Hence, persistently popular queries
  - Forecasts based on subset of 5 years
- Let's look at some of them...
  - Data since 2004, not just 5 years
  - Shown on logarithmic scale (y-axis)
  - Grey is data, 7-/31-/91-/365-day moving averages shown too
  - T-giving, New Years marked

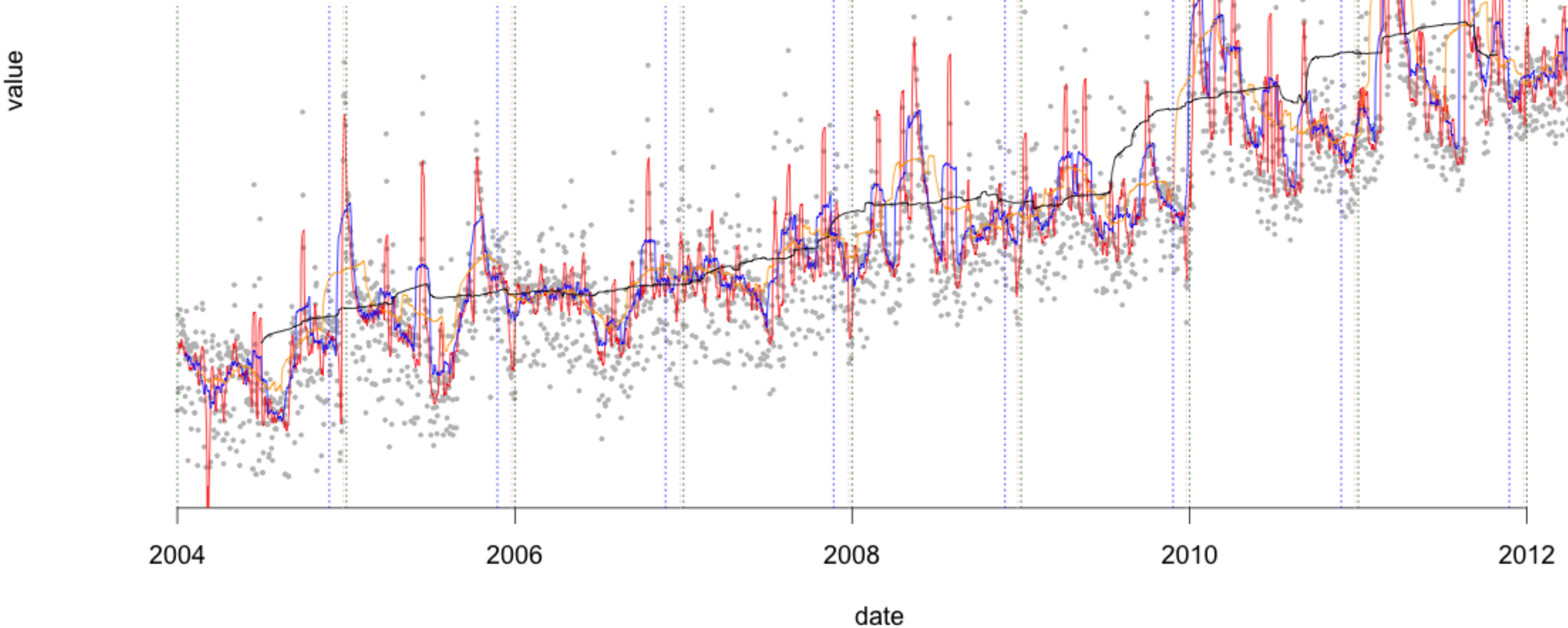
Google!

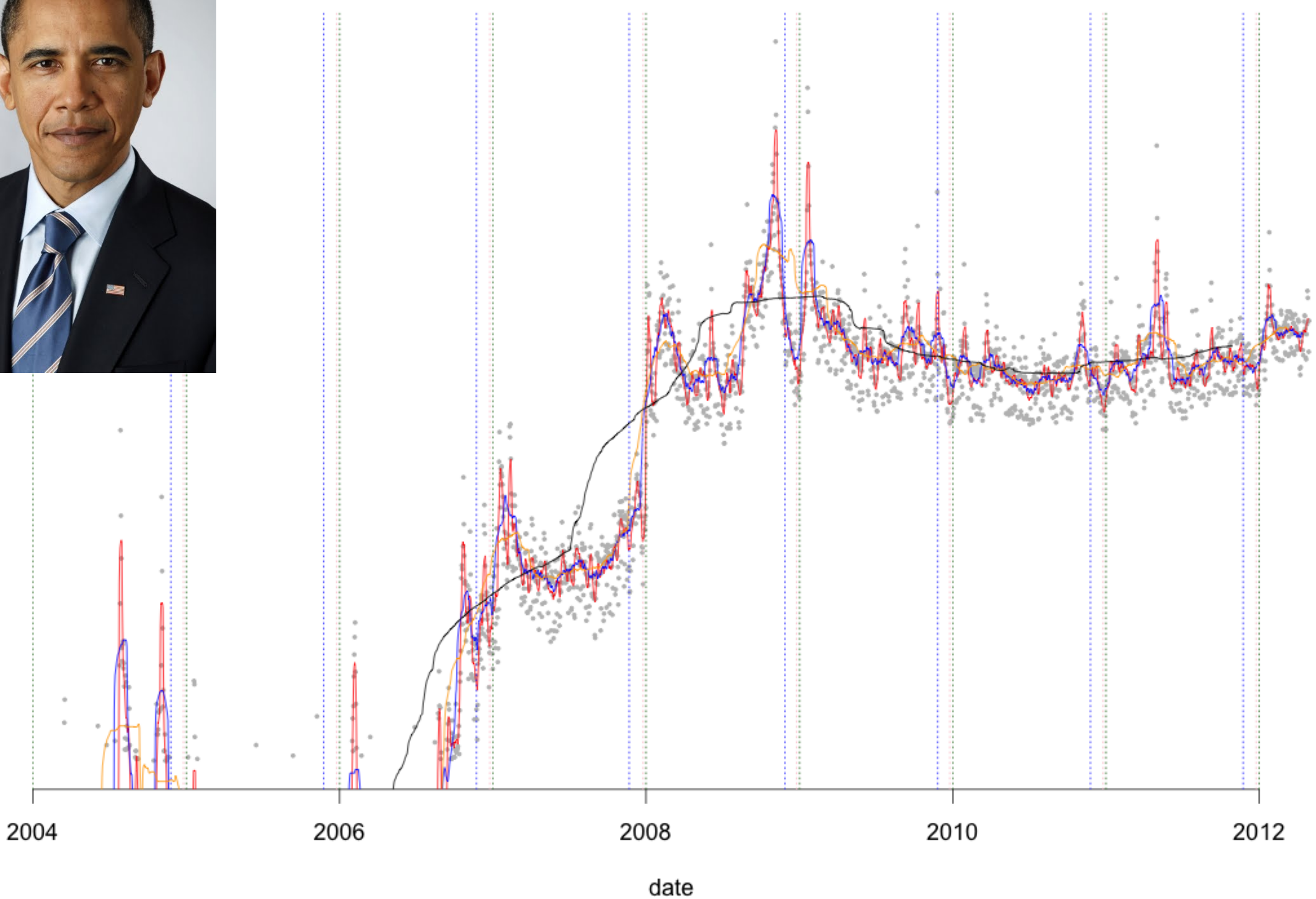


Google™

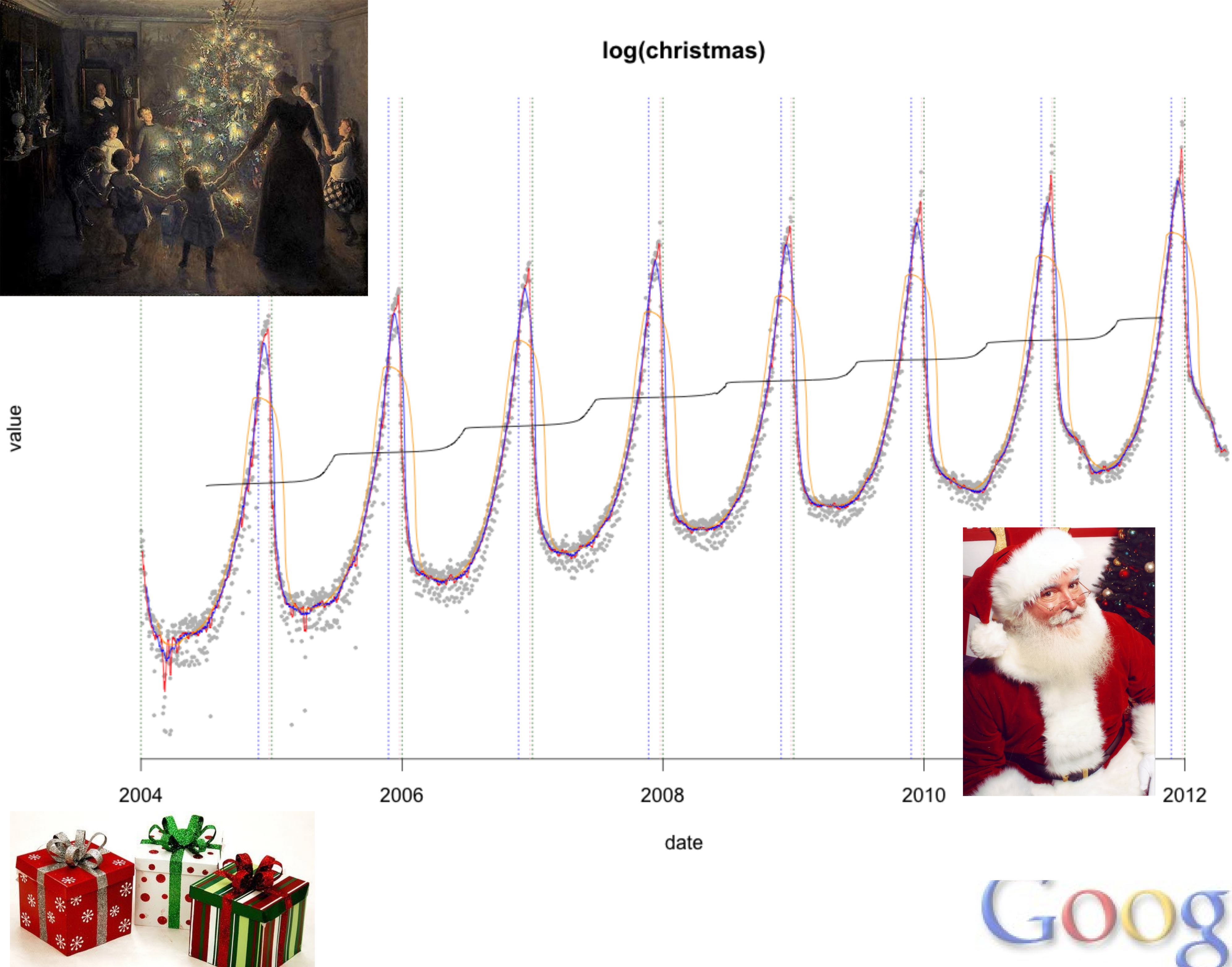

log(bing)
value
date
2004
2006
2008
2010
2012
PERK
MATTHEW PERRY
Chandler Bing(???)
bing
Google

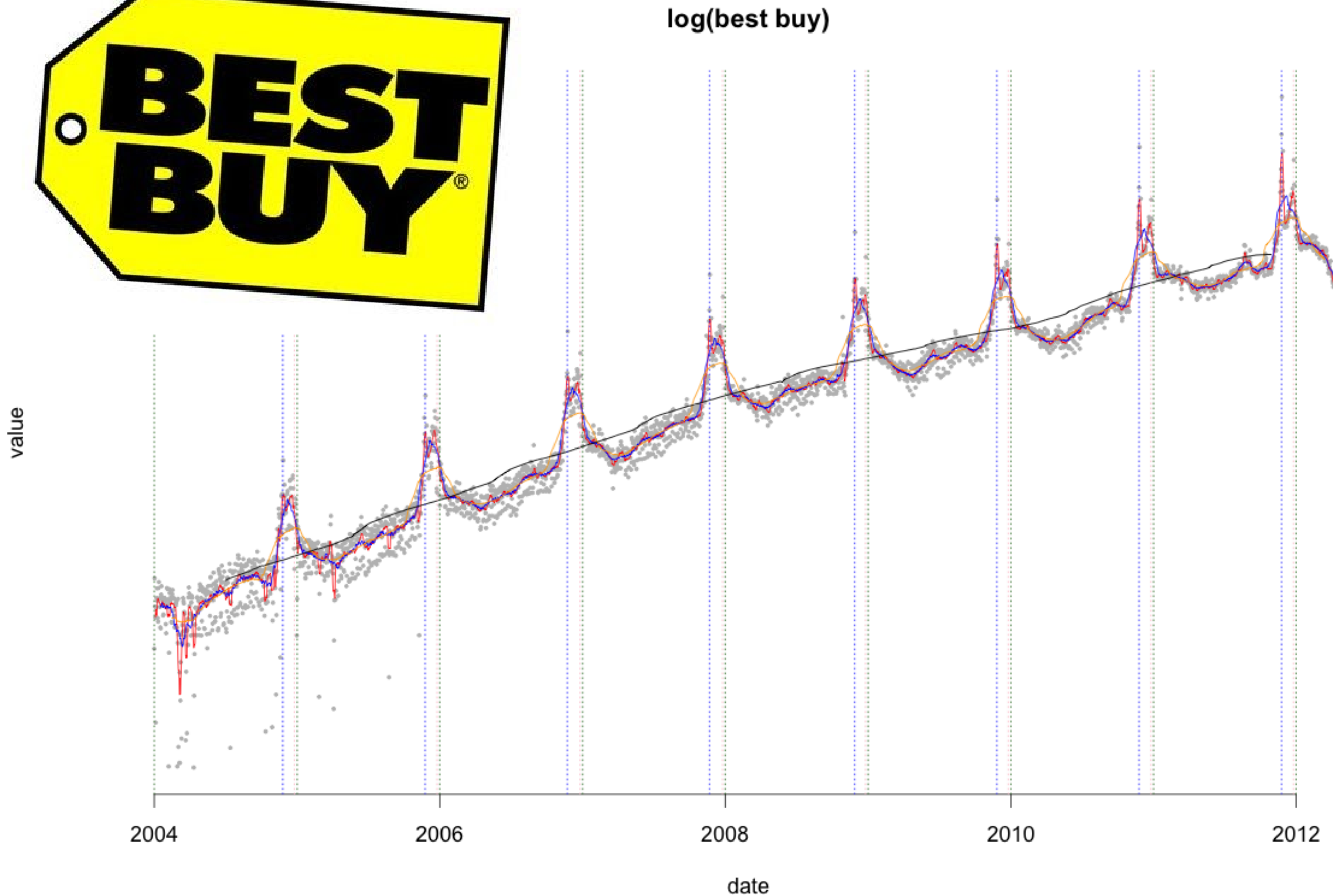

Preliminary Determination of Epicenters
358,214 Events, 1963 - 1998

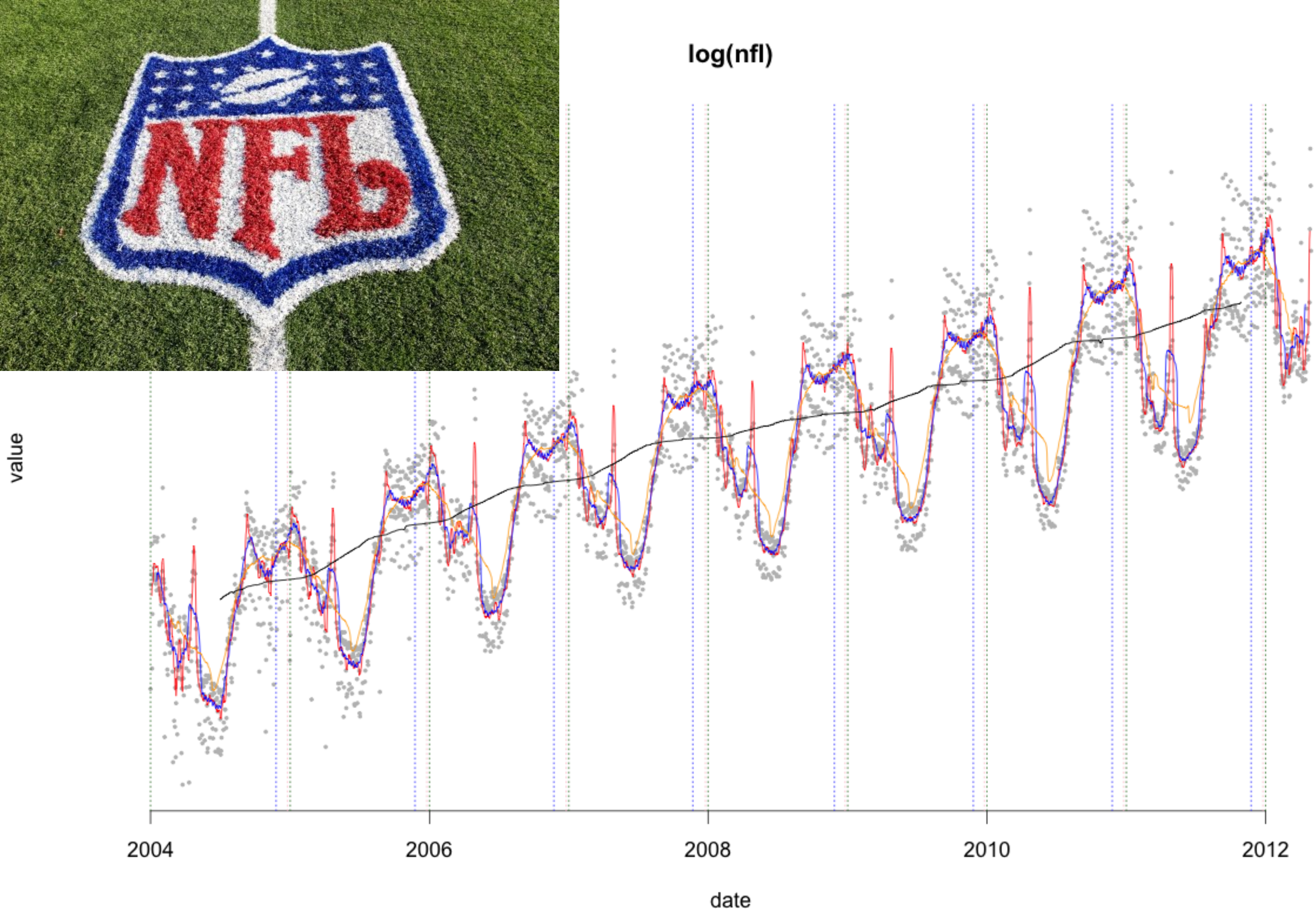

log(earthquake)
value
date
2004
2006
2008
2010
2012



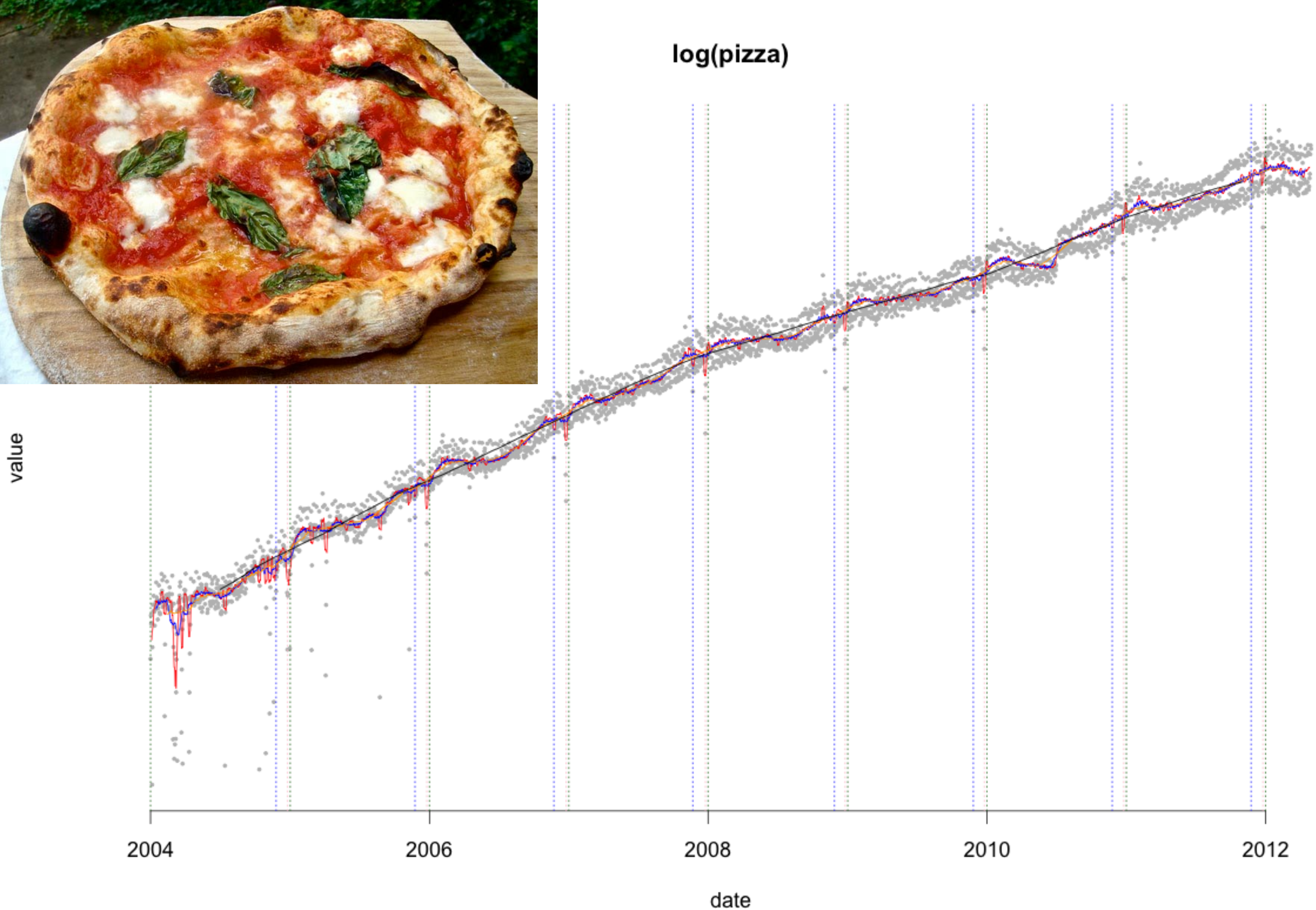
log(obama)
value
2004
2006
2008
2010
2012
date


log(christmas)
value
2004
2006
2008
2010
2012
date


log(best buy)
value
2004
2006
2008
2010
2012
date

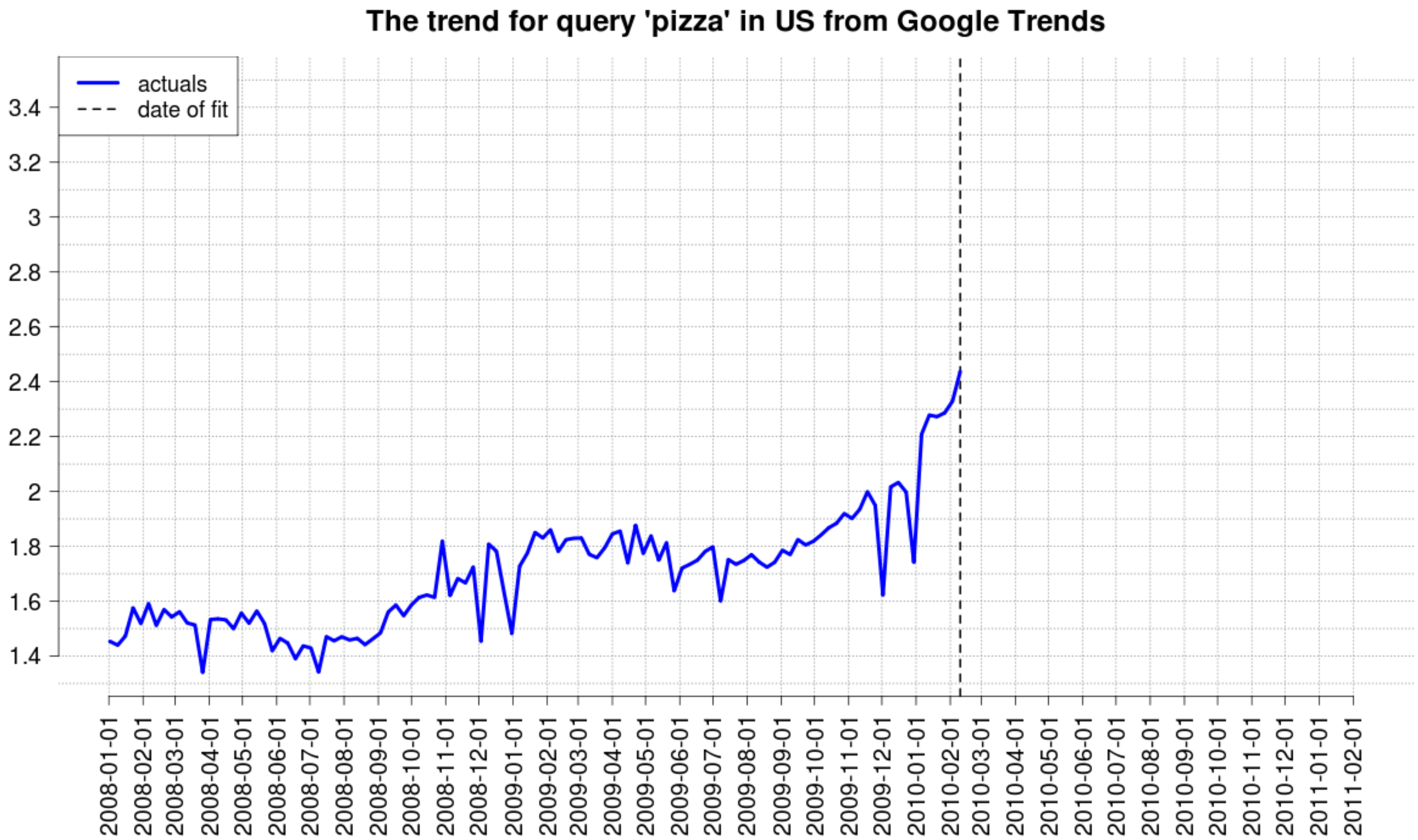

log(nfl)
value
2004
2006
2008
2010
2012
date

log(pizza)

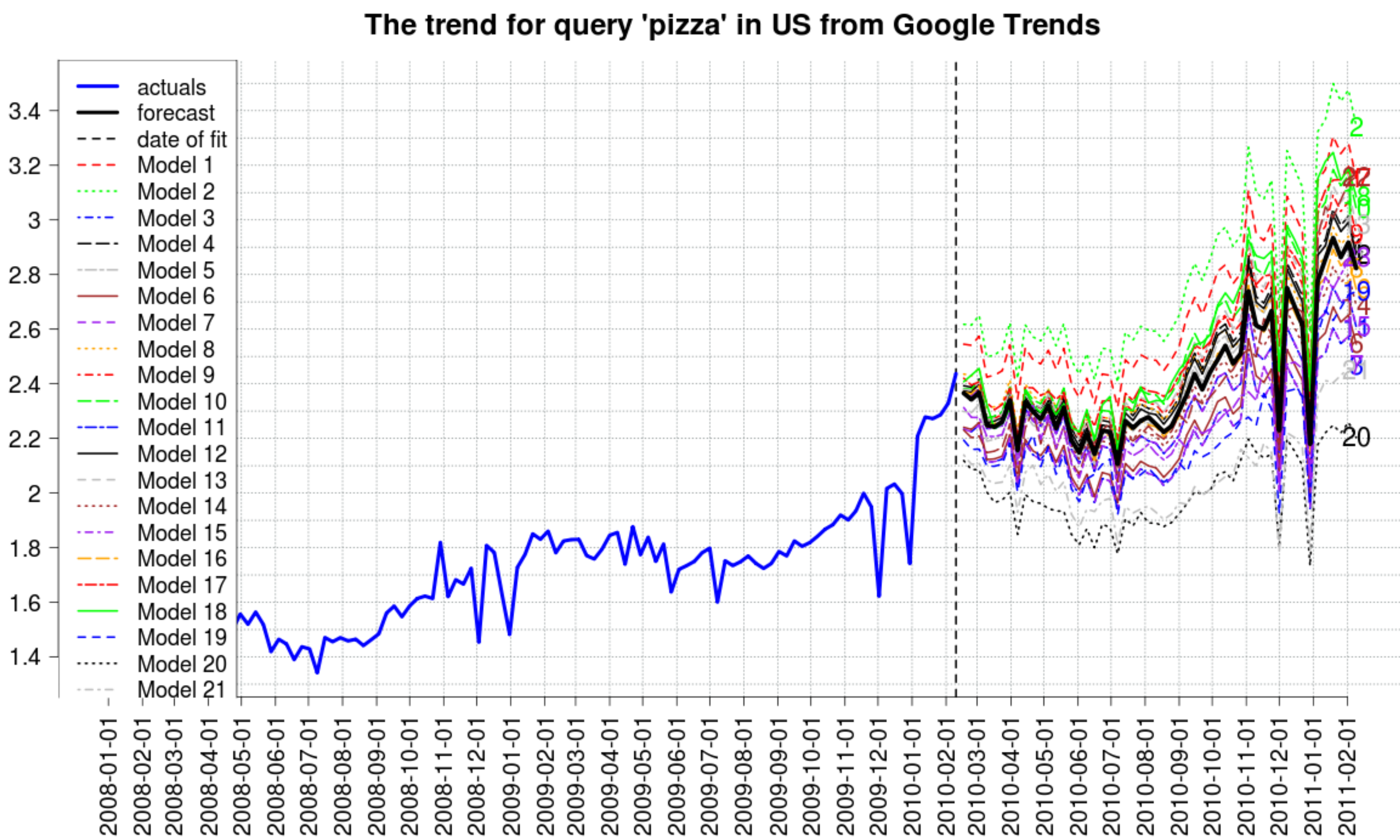


Methods

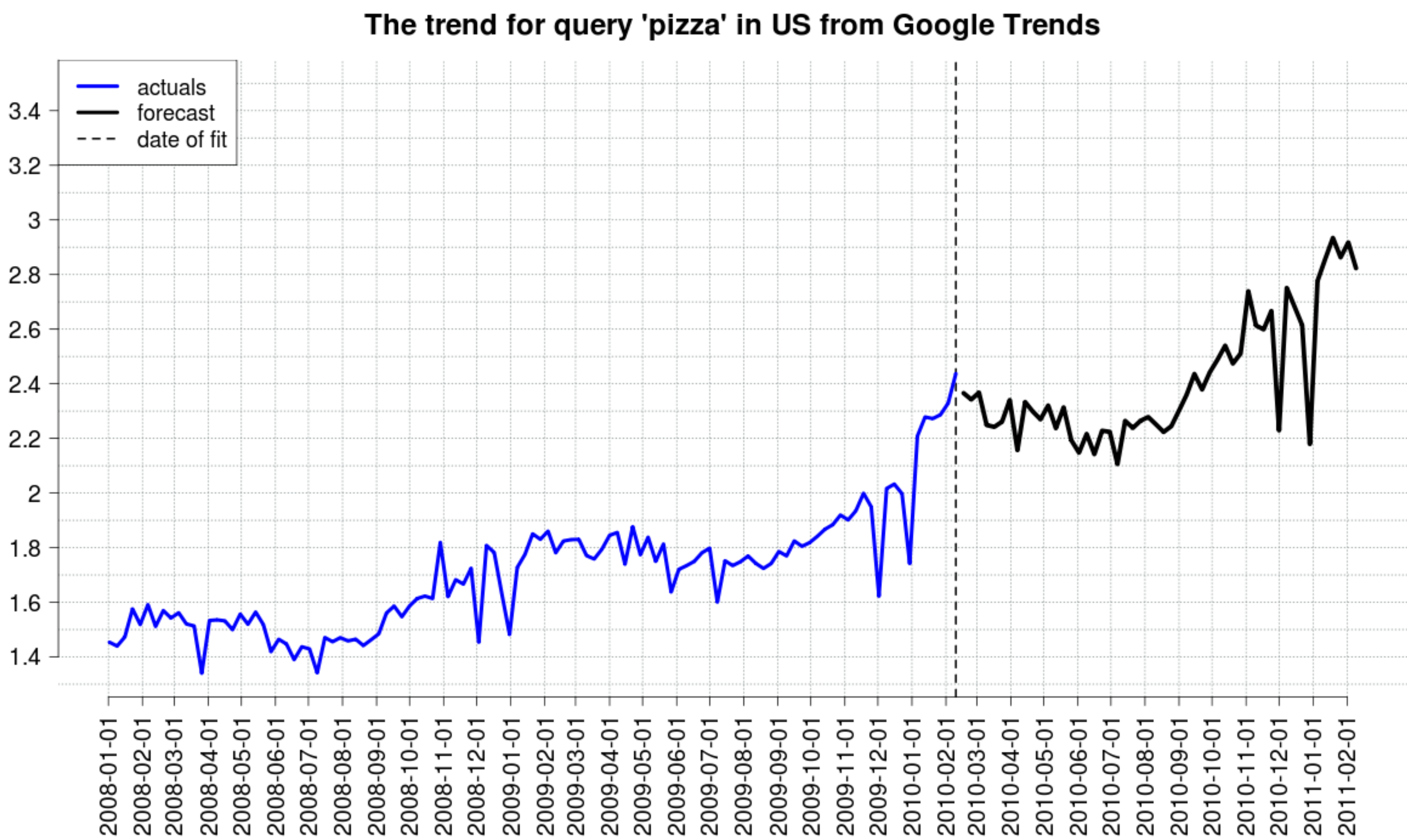
Many Models Approach



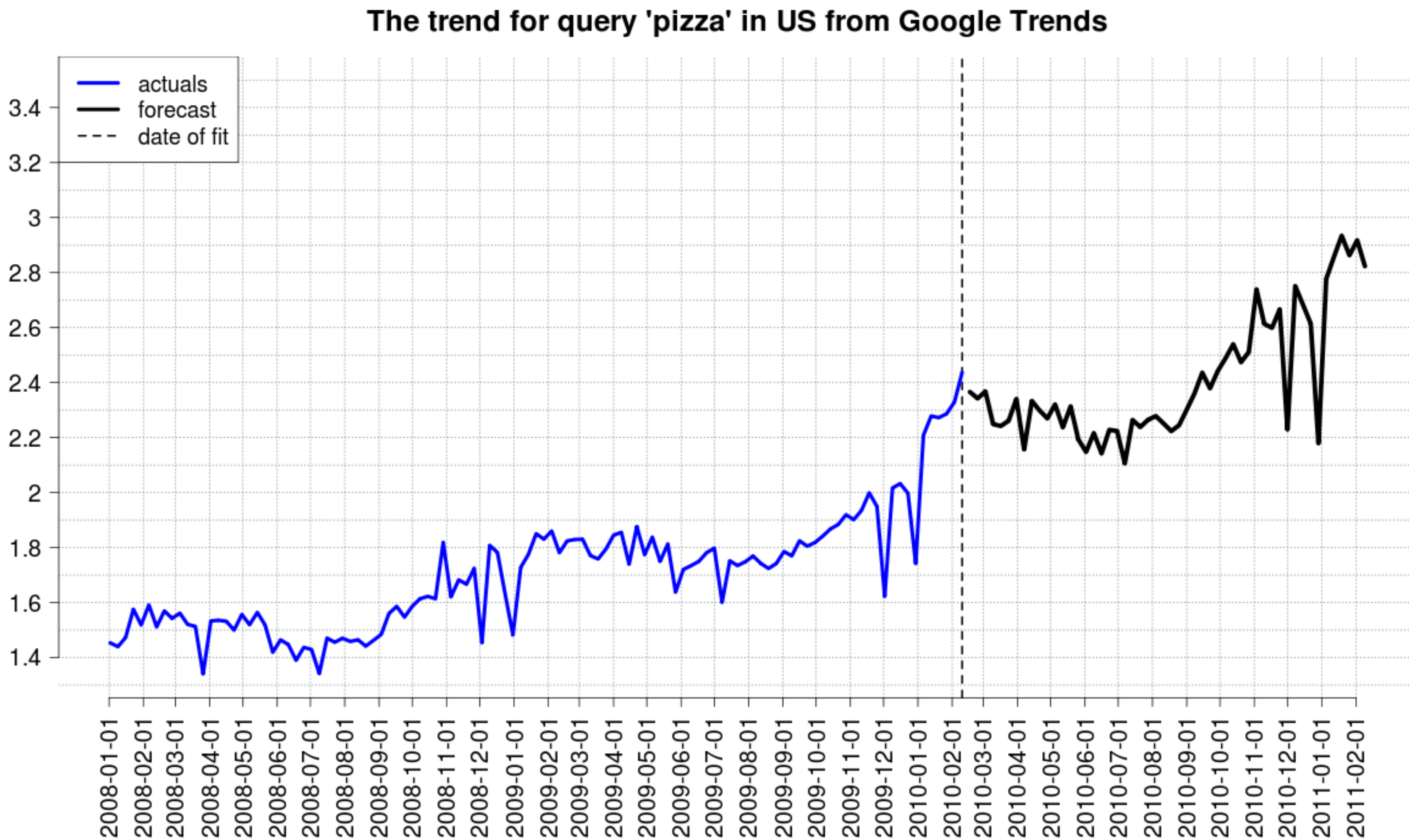
Many Models Approach



Many Models Approach



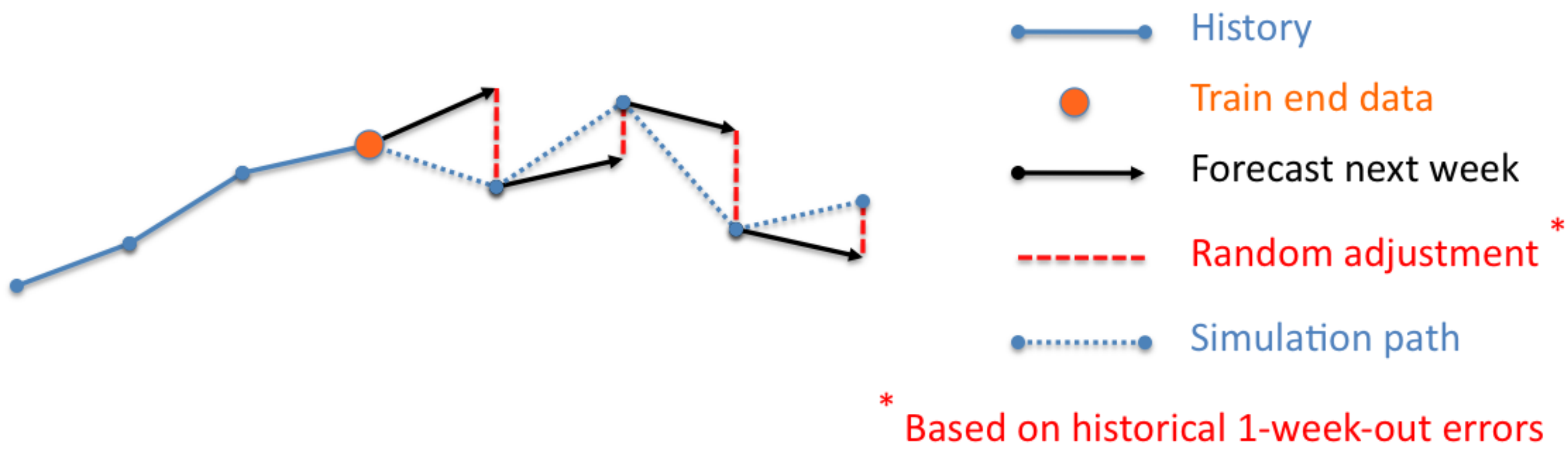
Many Models Approach

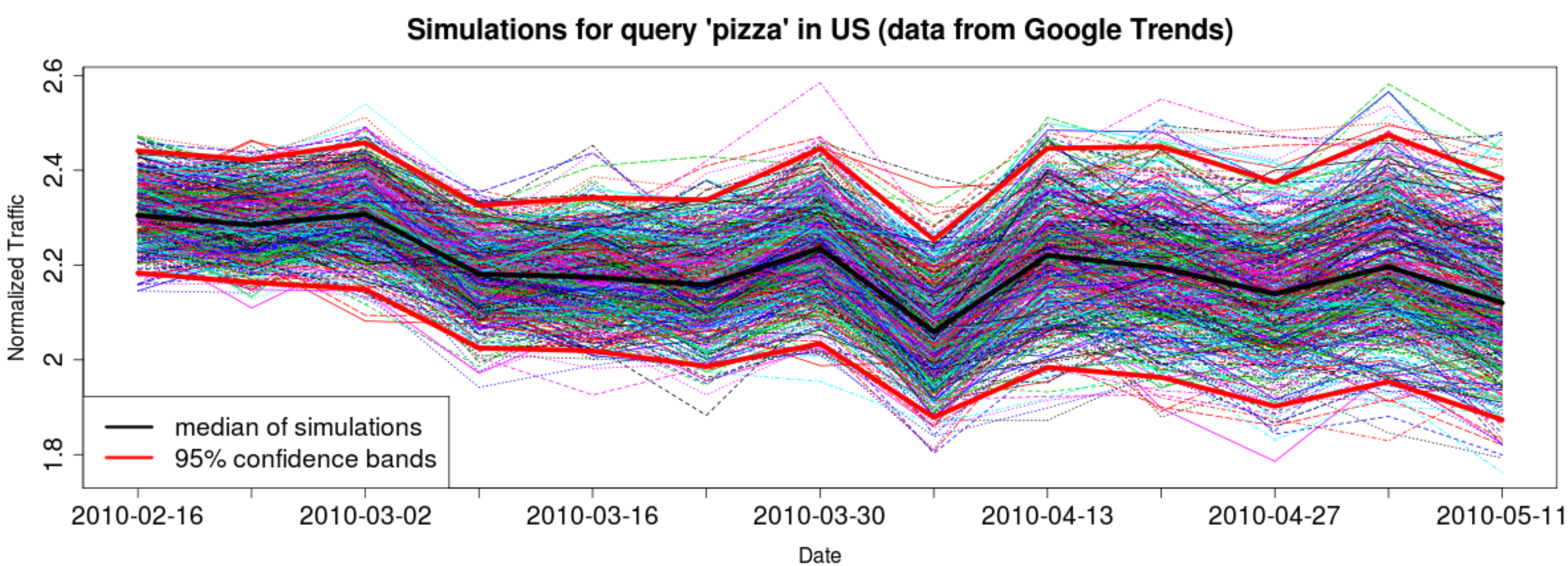


Armstrong, J. S. **Combining forecasts.** *Principles of forecasting: A handbook for researchers and practitioners , 417–439.*

Clemen, R. **Combining forecasts: A review and annotated bibliography**. *International Journal of forecasting 5, 559-583*

Confidence Intervals

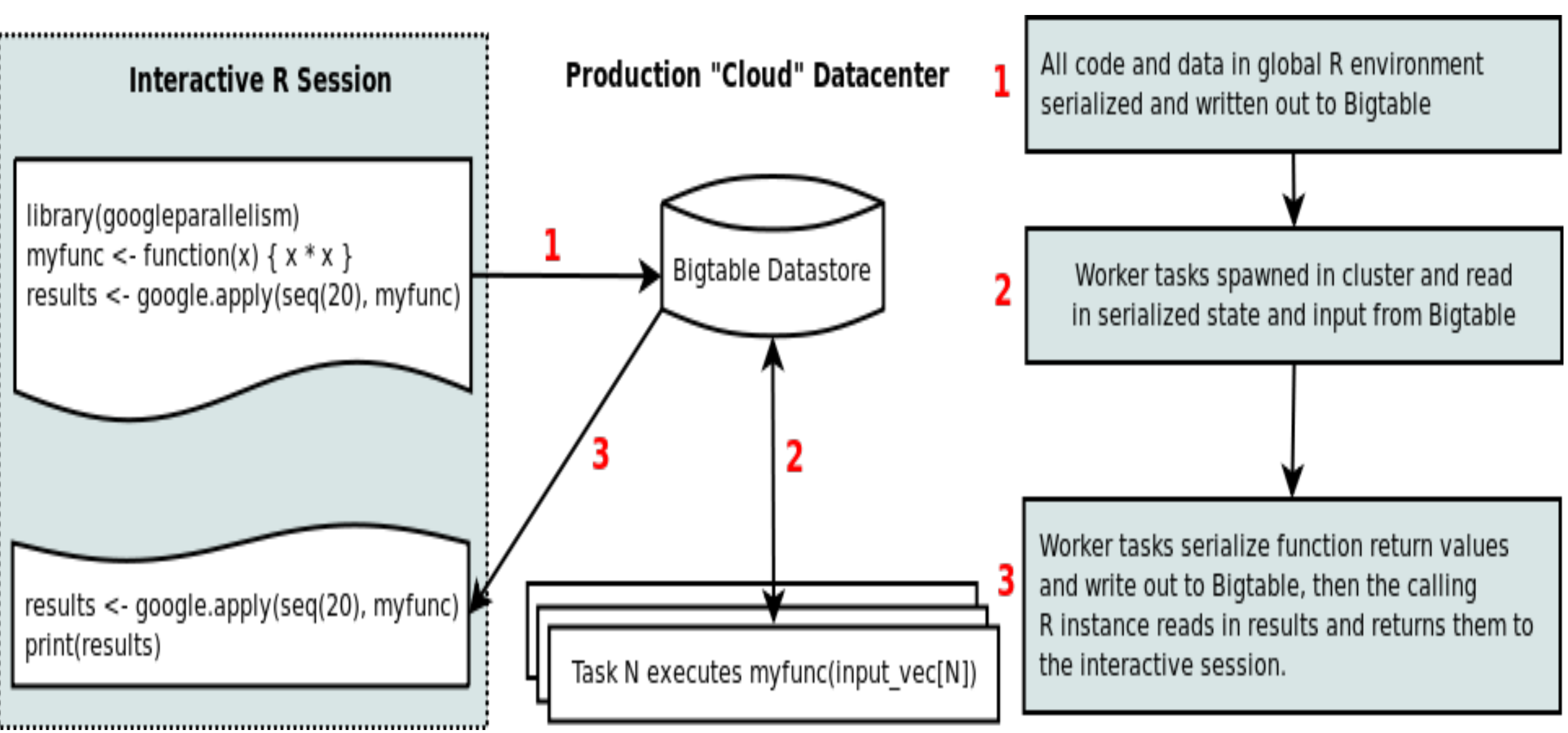




Parallelization in R: Why?

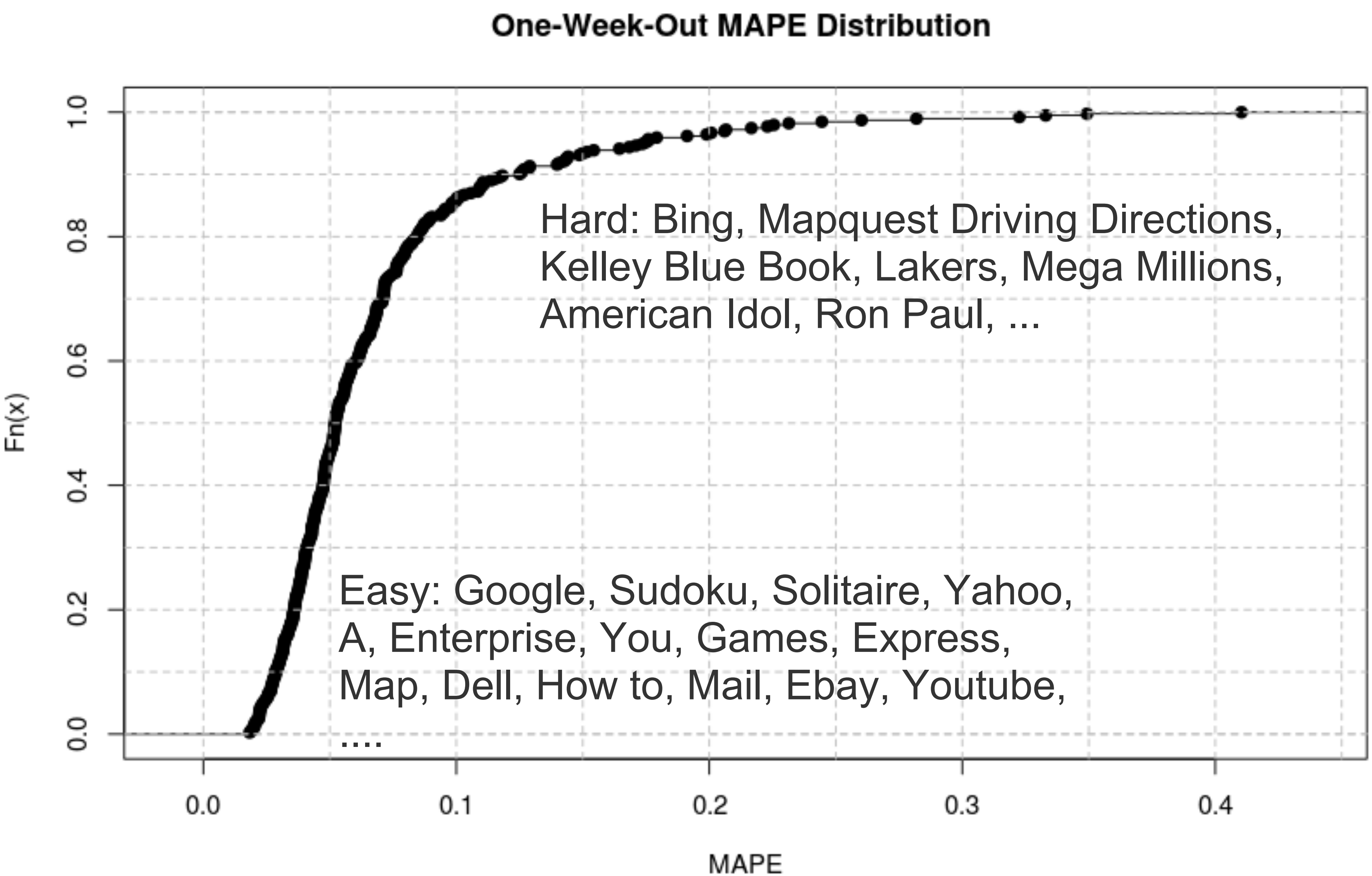
1,000 Trends t.s. X
1,000 realization/t.s. X
~10 sec/realization = 10^7 sec ~ 4 months

But 9 hours in parallel, cut by factor of ~300

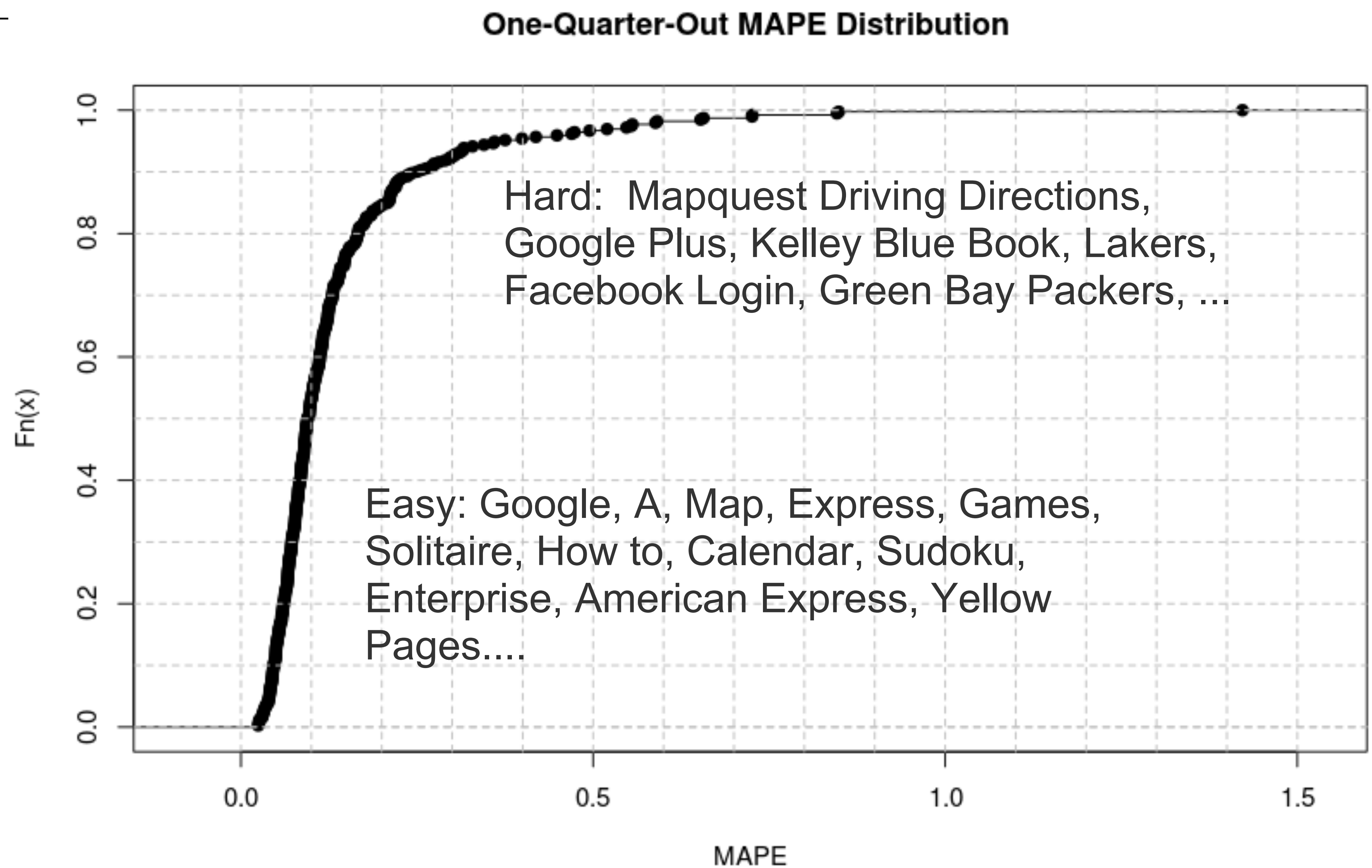


Overall Performance

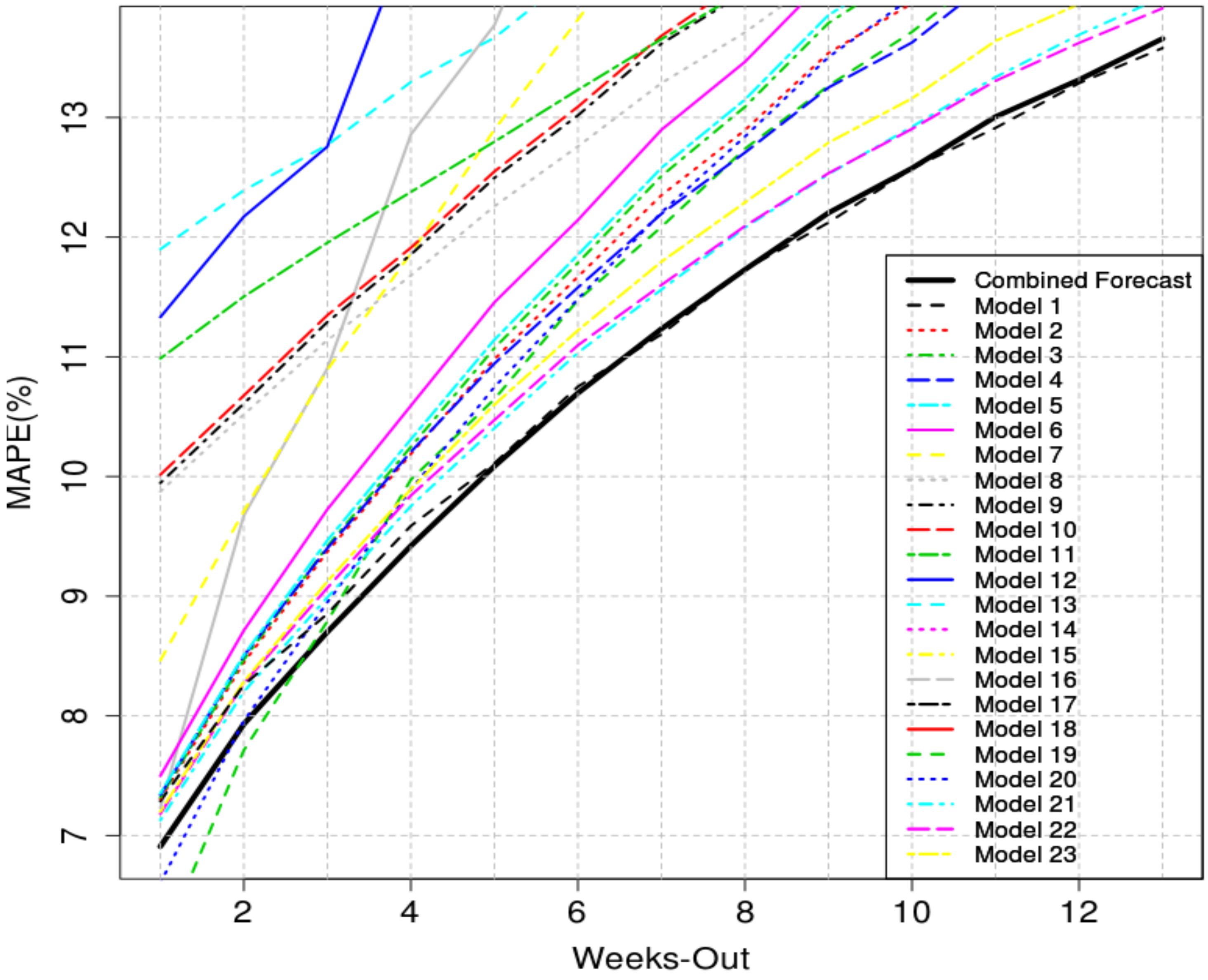
One-Week-Out MAPE



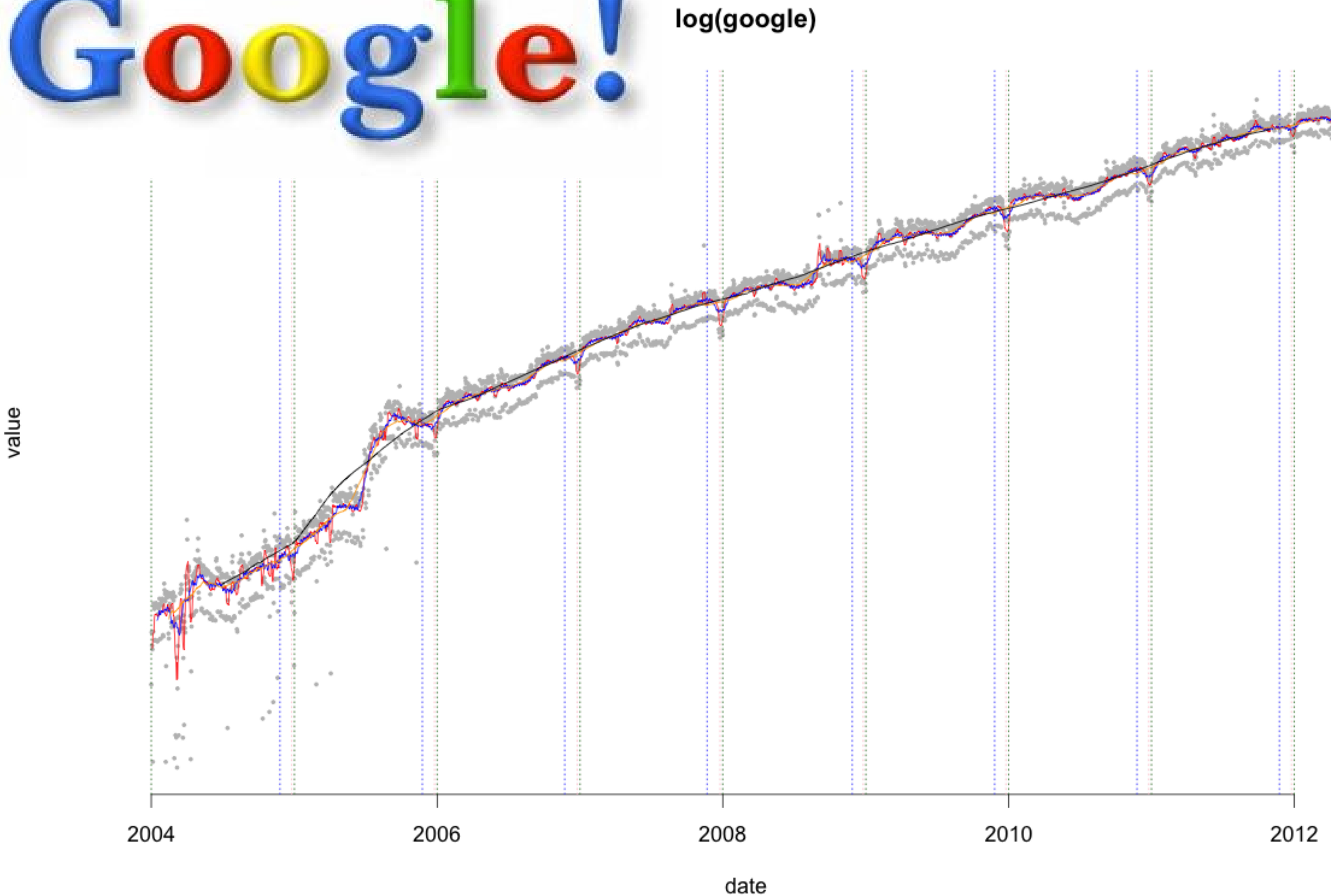
One-Quarter-Out MAPE



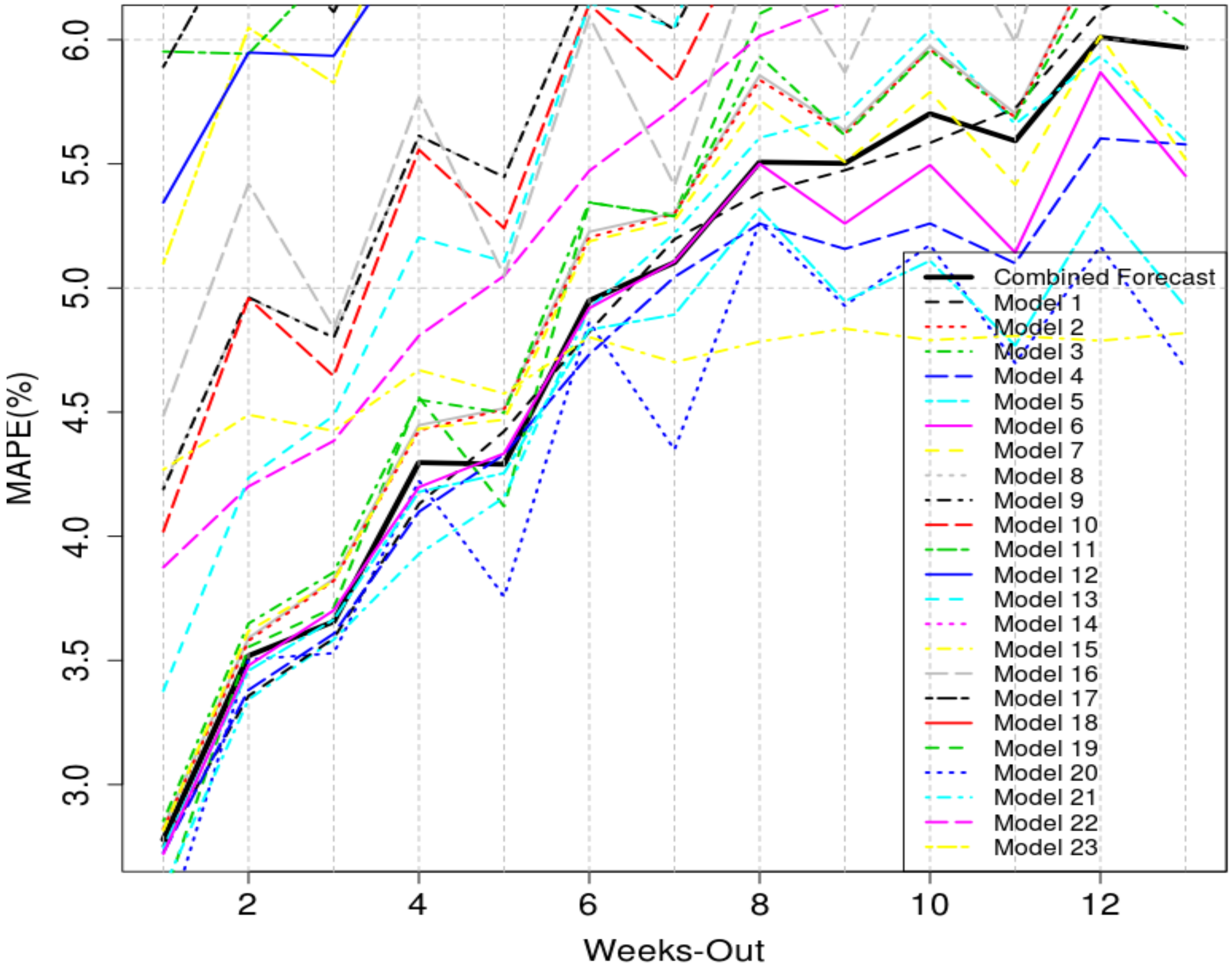
Performance of Individual Models (Overall, based on all queries)

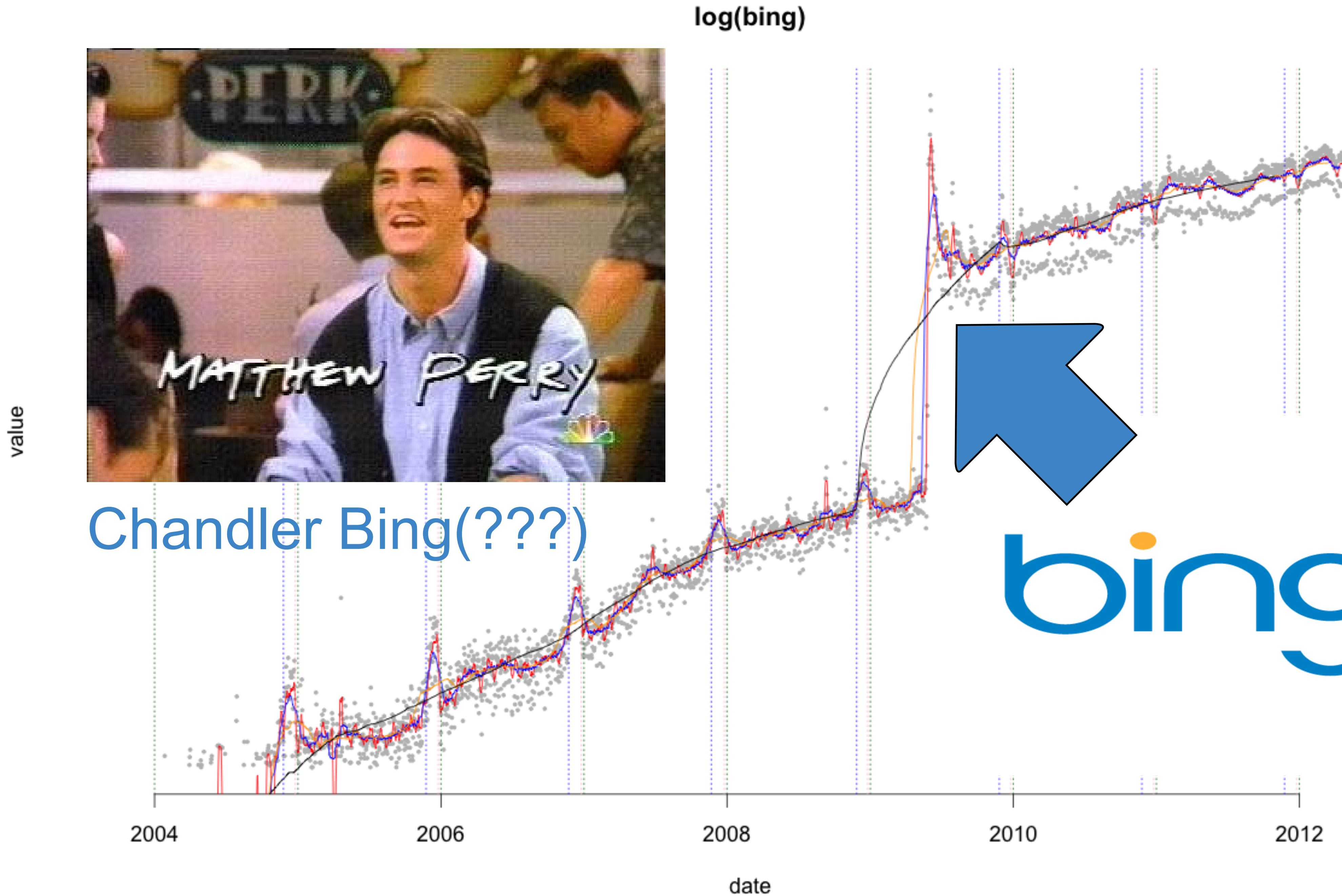


Performance on Individual Time Series


log(google)
value
date
2004
2006
2008
2010
2012

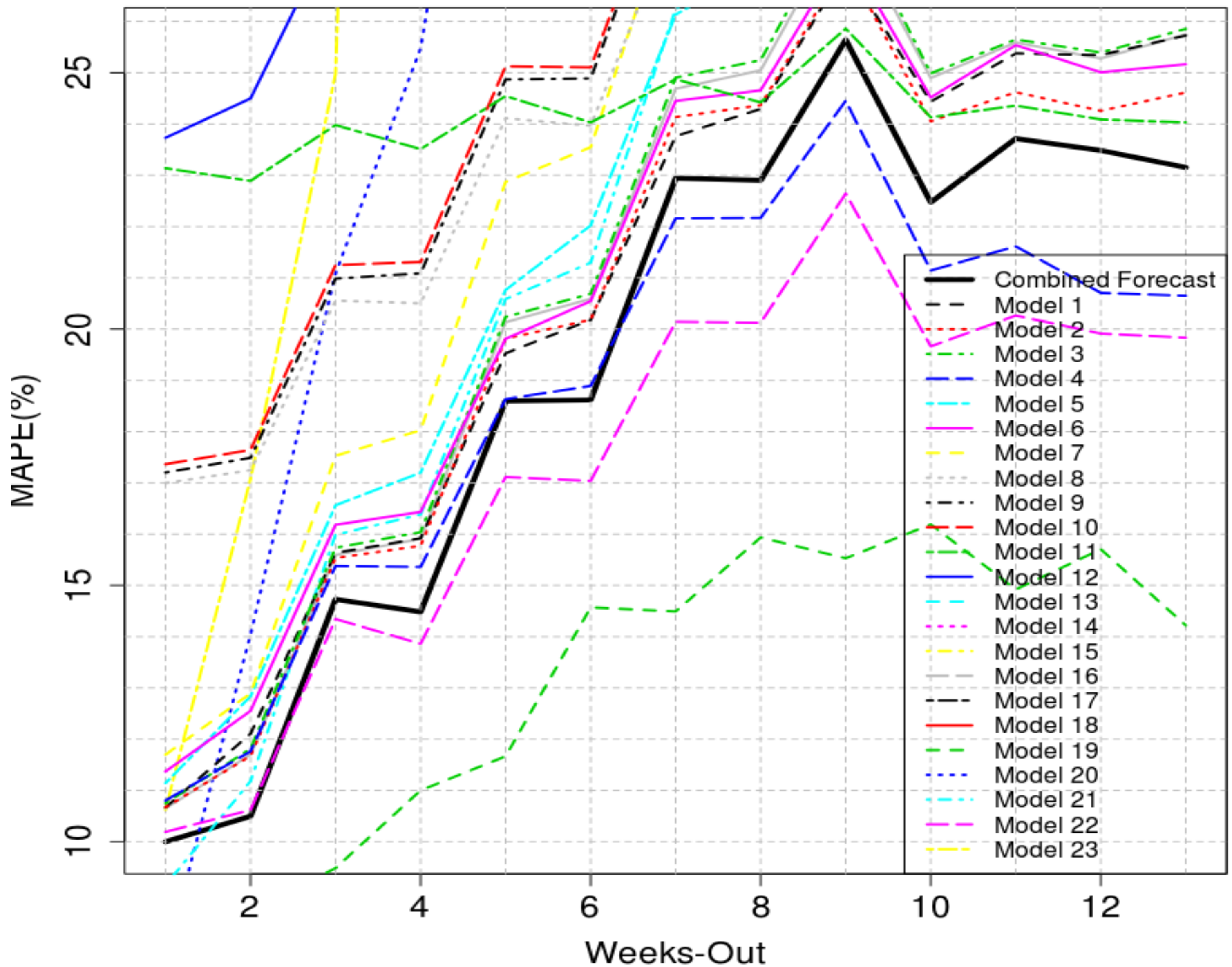
Performance on query "google"



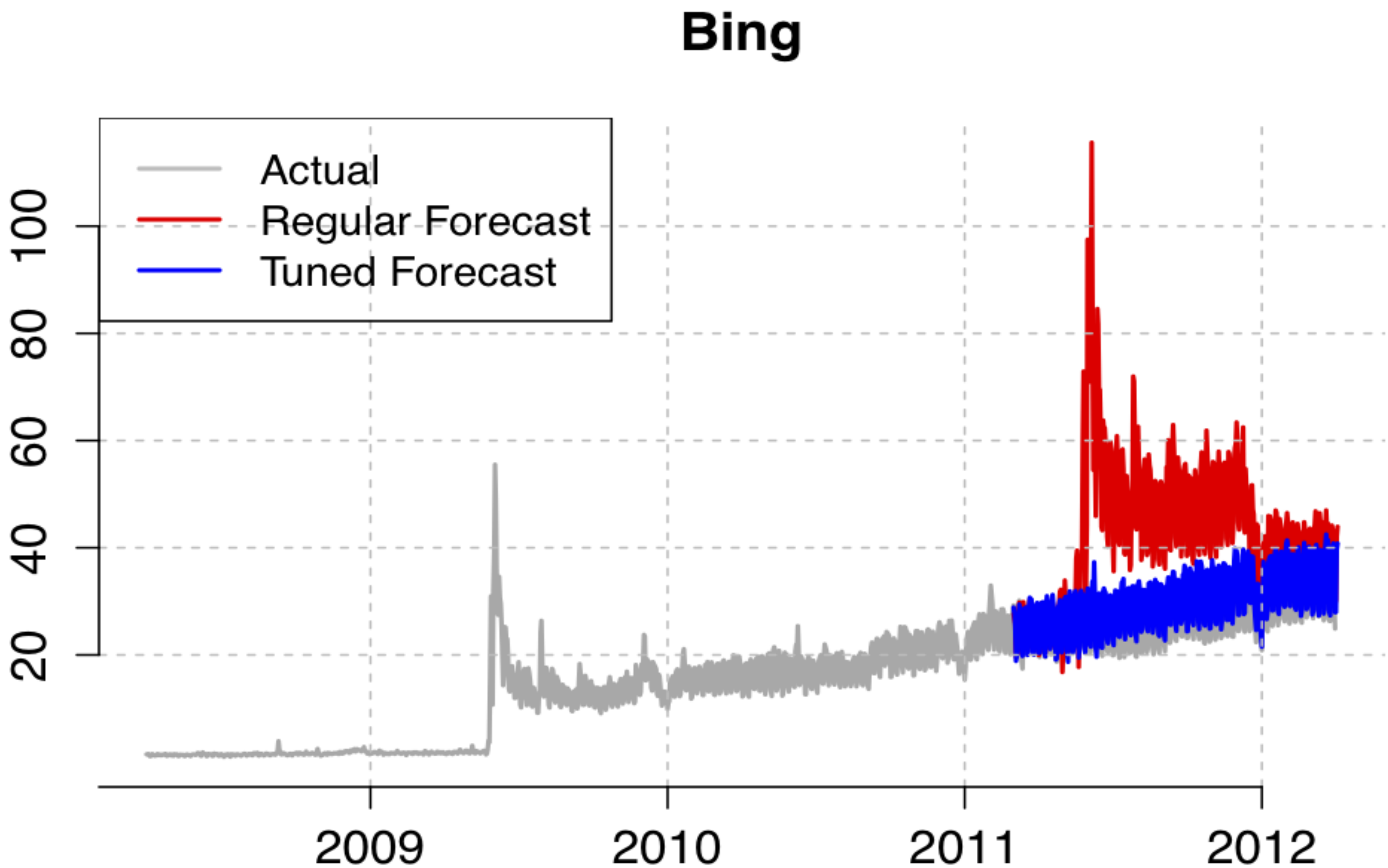


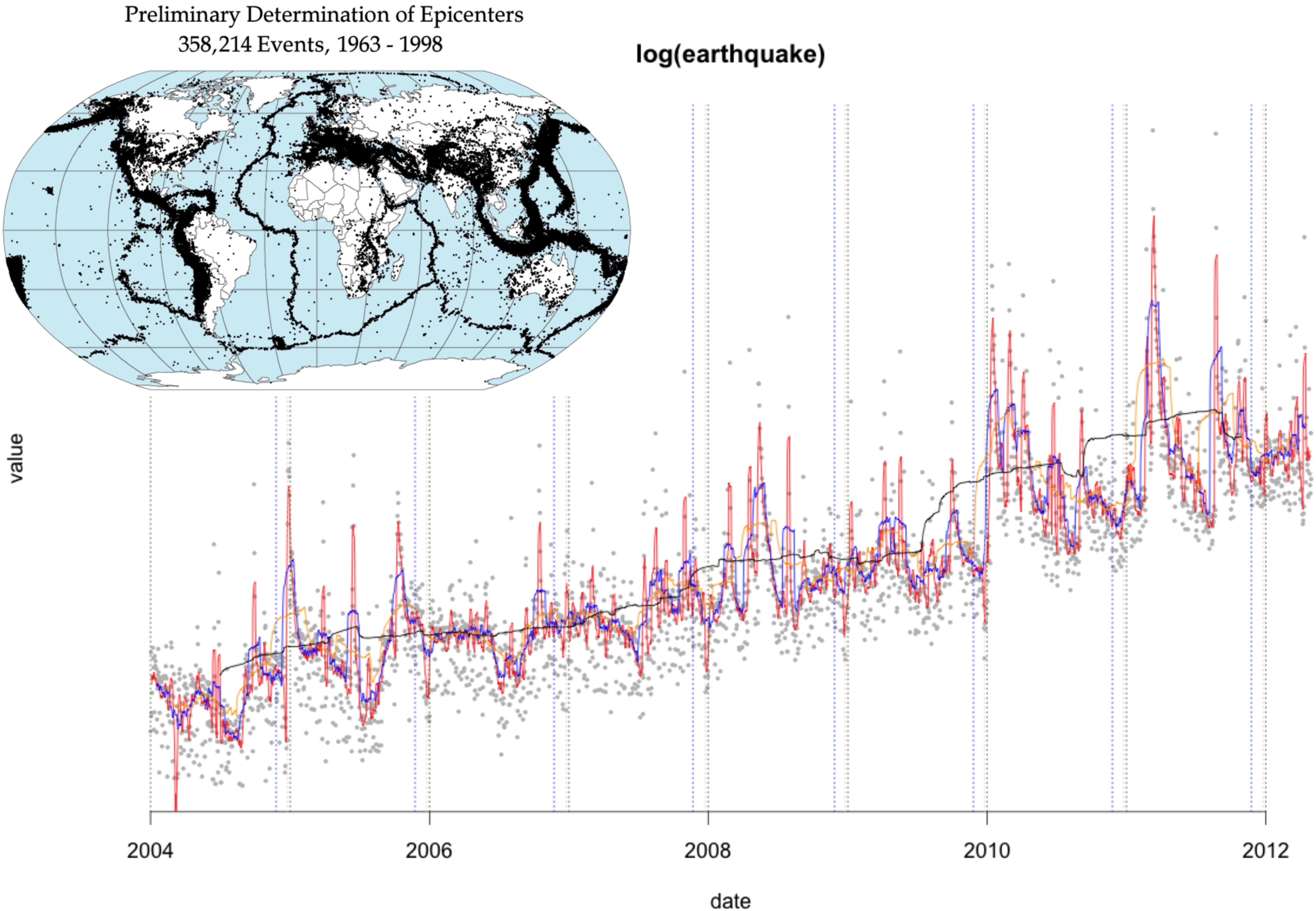
Chandler Bing(???)

Performance on query "bing"

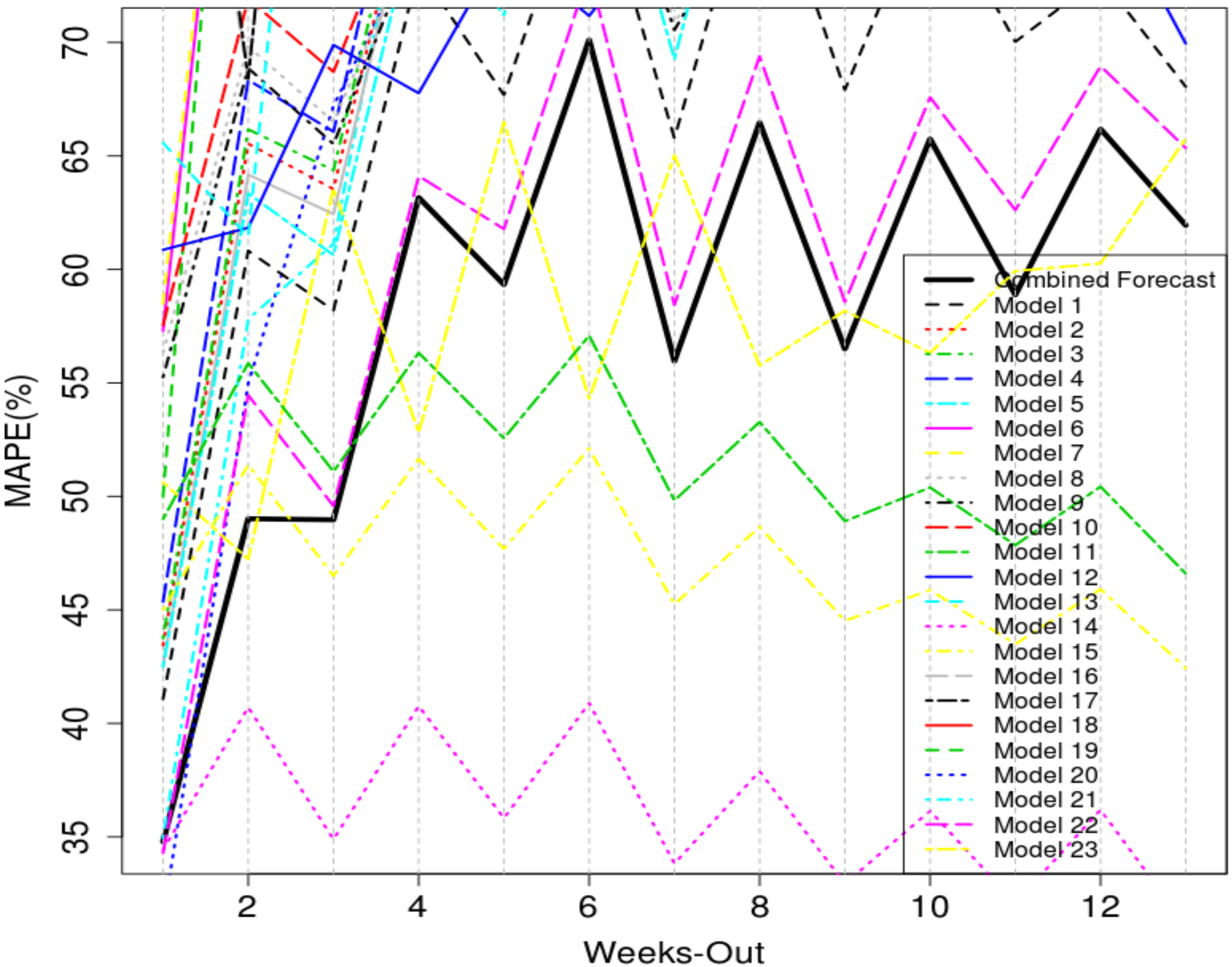


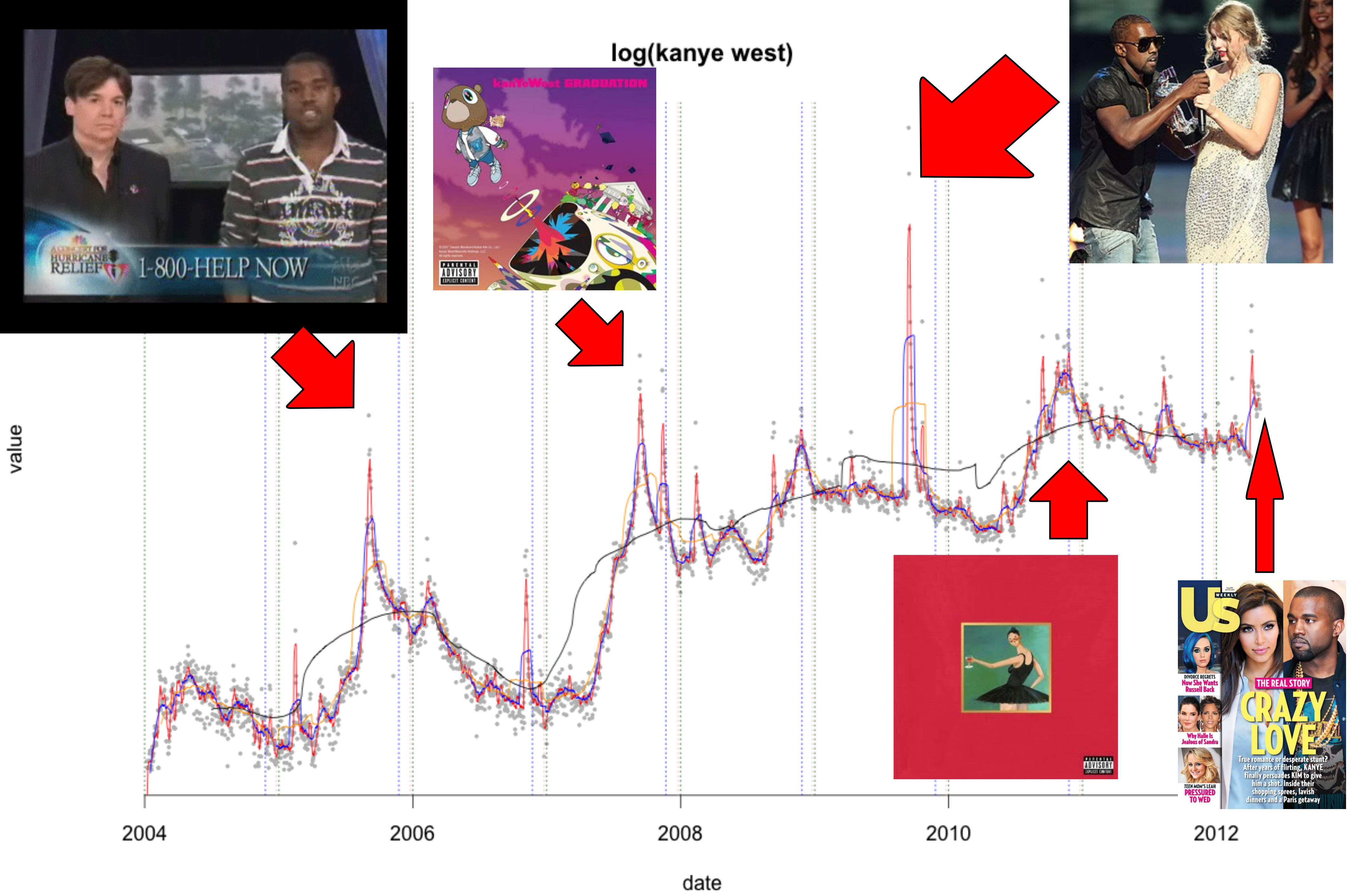
Tuned forecast (normalized values)




Preliminary Determination of Epicenters
358,214 Events, 1963 - 1998
log(earthquake)
value
2004
2006
2008
2010
2012
date

Performance on query "earthquake"





Celebrity earthquakes

Wrapping Up...

Conclusions, Statistical Methods

- Automatic, robust method needed for large number of diverse time series
- Many Models (MM) approach does well for large percentage of 1000 Trends queries
- Overall, MM beats individual models
- Some models do better on particular queries, but how to know *a priori*?

Conclusions, Computing

- Large number of time series & statistical method => large number of forecasts
- Number of forecasts => parallelization
- Parallelization: google.apply() internally, external options available
- Our version cut run time by factor of 300 (4 months -> 9 hrs)

See also:
JSM2011 Paper, http://research.google.com/pubs/pub37483.html) &
JSM2012 Presentation (K. Millar, "Scaling R to
Internet Scale Data", **upcoming** 8/1/12 @ 11:05am)

Future Work

- Classification?
- Correlations / multivariate methods?
- Aggregate / disaggregate forecasting?
- Dimension reduction? Then forecast?
- More models?

That's All, Folks!