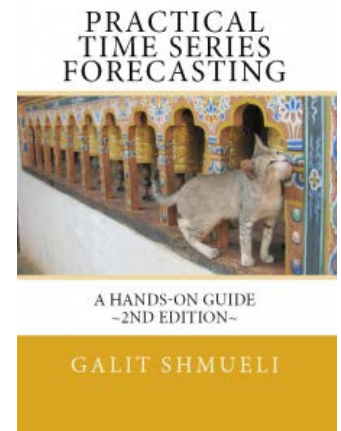


A pair of black binoculars with red-tinted lenses lies on a bed of smooth, grey stones. The background is a blurred, rocky landscape. The text "Including External Information" is written in white, sans-serif font across the middle of the binoculars, and "Forecasting Analytics" is written in a smaller, italicized white font below it.

Including External Information
Forecasting Analytics

Some methods, such as **linear regression** (as well as logistic regression and neural networks), allow the inclusion of external information in the form of predictors

$$(\text{Class attendance})_t = \beta_0 + \beta_1 (\text{Class attendance})_{t-1} + \beta_2 (\text{Interviews})_t + \varepsilon$$



Issue #1

Form of predictors: Lagged or forecasted

$$\text{Airfare}_t = \beta_0 + \beta_1 (\text{Petrol Price})_{t-1} + \varepsilon$$

$$\text{Airfare}_t = \beta_0 + \beta_1 (\text{Petrol Price})_t + \varepsilon$$



Must be forecasted

Issue #2

Choice of training/validation periods:
data availability and forecast horizon

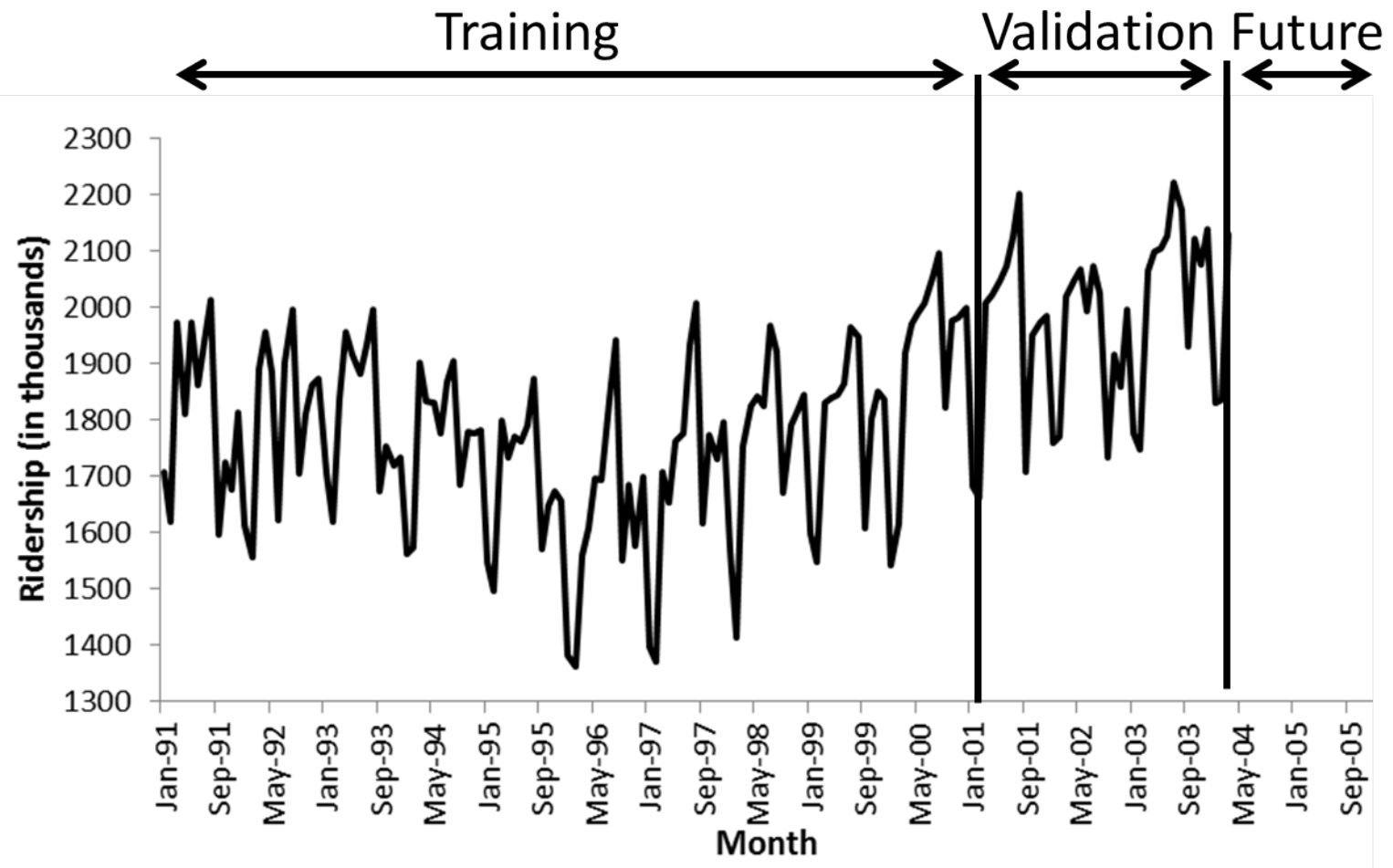
Must account for lag time in obtaining data

$$\text{Airfare}_t = \beta_0 + \beta_1 (\text{Petrol Price})_{t-1} + \varepsilon$$



Must be available at
time of prediction

Data lag time and forecast horizon



Predicting the Future With Social Media



Sitaram Asur



Social Computing Lab

The Social Computing Lab focuses on methods for harvesting the collective intelligence of groups of people in order to realize greater value from the interaction between users and information.



Bernardo A. Huberman

Published on arXiv Cornell University - March 2010

<http://arxiv.org/abs/1003.5699>

<http://www.slideshare.net/napo/predicting-the-future-with-social-media-3799074>

The question

How social media content can be used to predict real-world outcomes?

The case study:
predicting box-office revenues for movies using the chatter from Twitter

Why Twitter?

several tens of millions of users who actively participate in the creation and propagation of content

Why movies?

The topic of movies is of considerable interest among the social media user community

The real-world outcomes can be easily observed from box-office revenue for movies

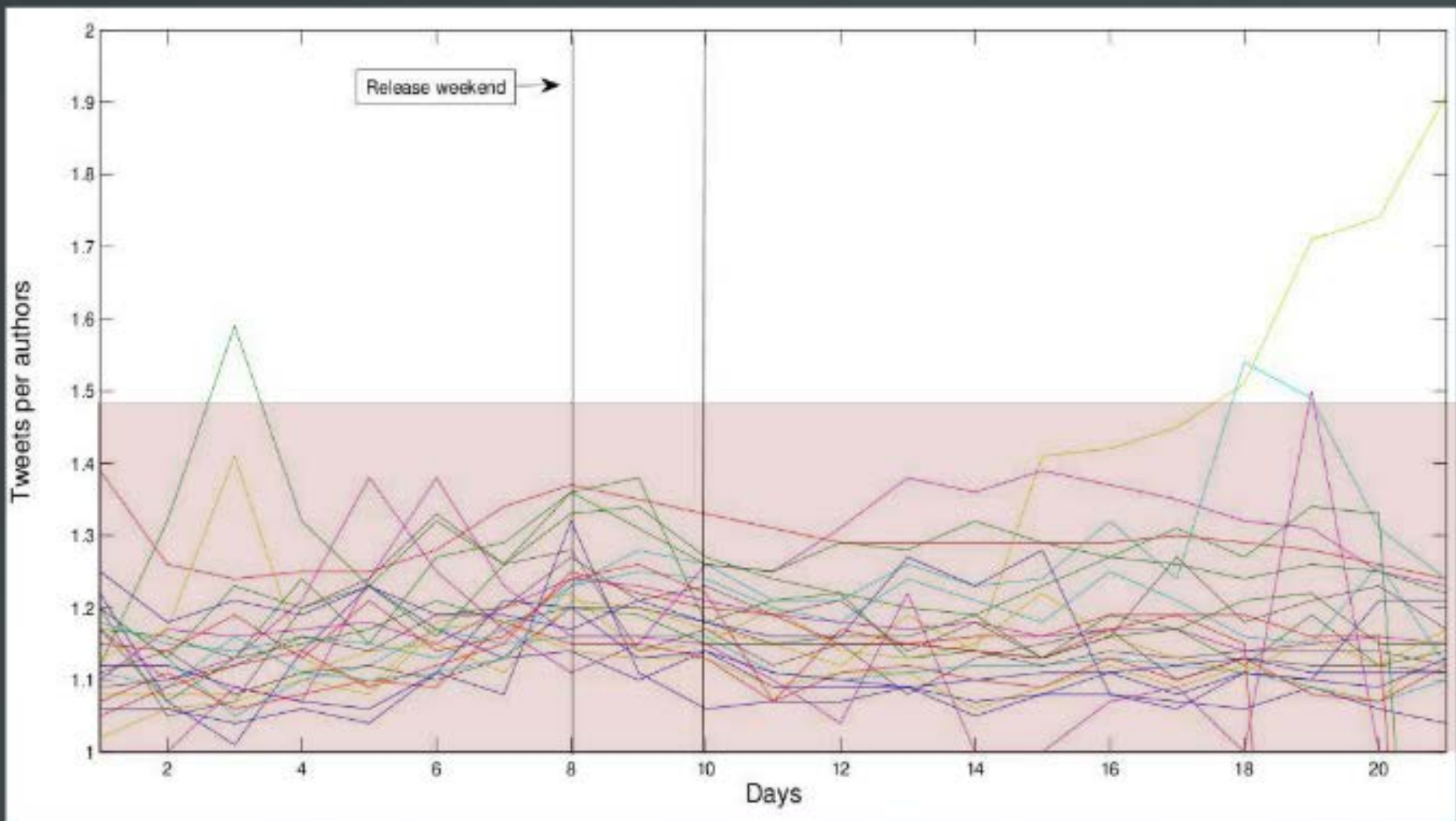


twitter



Dataset characteristics

Number of tweets per unique authors for different movies



y → tweets per authors
x → days
lines → movies

ratio remains fairly consistent between 1 and 1.5

Daily tweet rate (TR)



$$\begin{aligned} (\text{Box-office revenue})_t &= \beta_0 + \beta_1 \text{TR}_{t-1} + \beta_2 \text{TR}_{t-2} + \beta_3 \text{TR}_{t-3} + \beta_4 \text{TR}_{t-4} \\ &+ \beta_5 \text{TR}_{t-5} + \beta_6 \text{TR}_{t-6} + \beta_7 \text{TR}_{t-7} + \beta_8 (\# \text{ Theaters})_t + \epsilon \end{aligned}$$

“... in all cases, we are using only data available **prior** to the release to predict box-office for the opening weekend.”

“...since historical information on tweets are not available [retrospectively], we were able to use data on only the movies we have collected...”

In brief

Make sure the forecasting method(s) in your project use external information in a valid fashion

- Define required forecast horizon

- Determine data availability and timing

- Evaluate performance accordingly

(Pretend that you are deploying the model in practice)