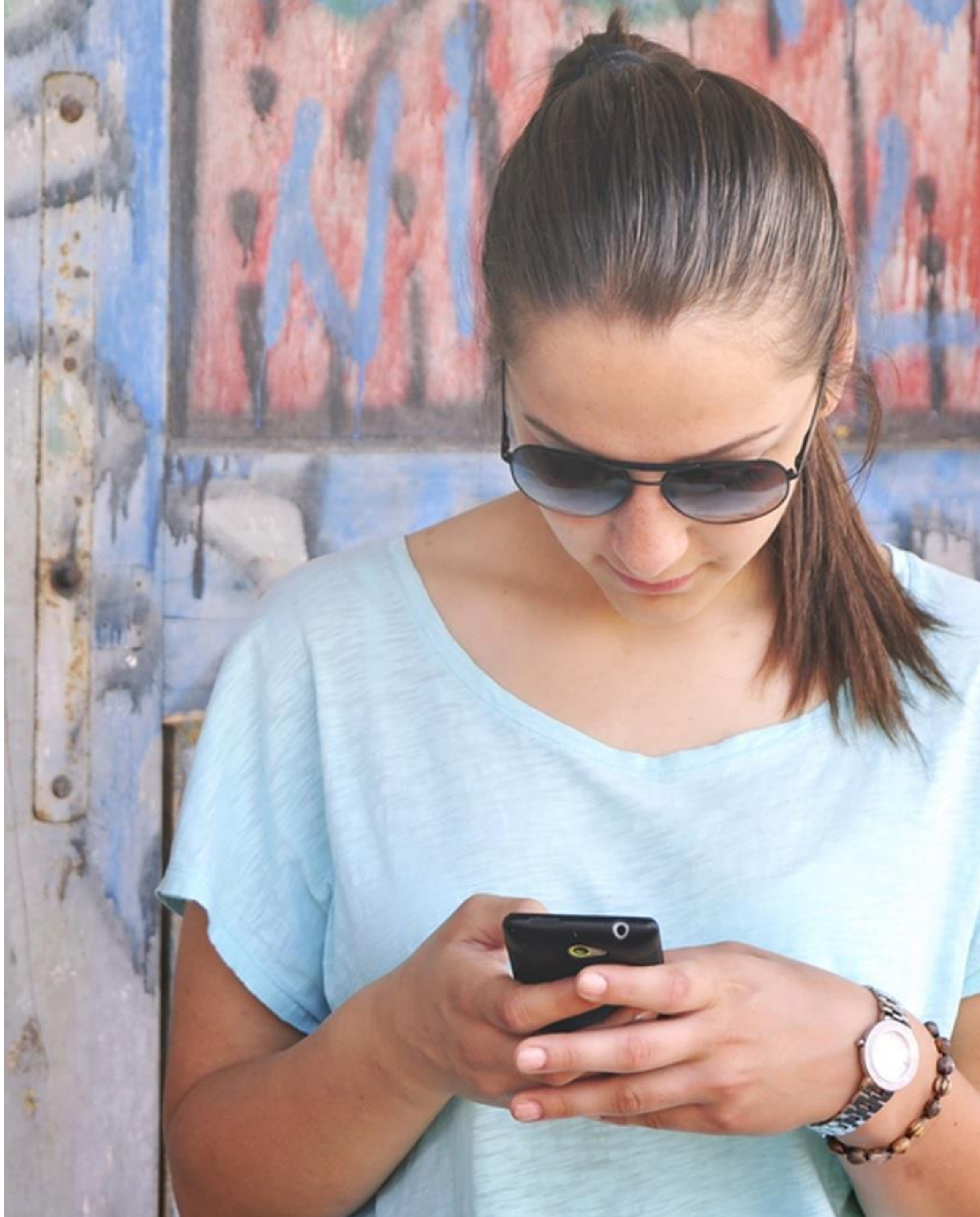
PREDICTING ONLINE BEHAVIOR

ASC CONFERENCE

LONDON 2019_05_16

respondi

CLOSE TO PEOPLE



THE PROJECT

MAKE TRACKING DATA SPEAK!

TRACKING

=

PANEL DATA
PERSON + ONLINE JOURNEY
(URLs VISITED)

PROS

GOOD FOR
CUSTOMER JOURNEYS or DECISION MAKING PROCESS ANALYSIS

CONS

NOT SO GOOD
TO RANK ARGUMENTS
TO UNDERSTAND WHAT MAKES THE DIFFERENCE
SMALL SAMPLE SIZE

WITH TRACKING DATA
CONTENT = BLINDSPOT

THE PROJECT

MAKE TRACKING DATA SPEAK!

WEB LISTENING

=

CRAWLING DATA
(TOPIC + CORPUS)

PROS

GOOD FOR
DESCRIBING THE ACTUAL CONTENT RELATED TO A TOPIC / WHAT IS SAID ABOUT (BY WHOM)?

CONS

NOT SO GOOD
TO ANALYSE WHO PAYS ATTENTION TO WHAT GRANULARITY

WITH WEB LISTENING
AUDIENCE= BLINDSPOT

THE PROJECT

MAKE TRACKING DATA SPEAK!

1ST STEP

DEMONSTRATE THE BENEFIT
OF MIXING
CRAWLING DATA AND PANEL DATA

TO PREDICT INTEREST IN NEWS

TRENDING NEWS

What‘s hot in the news?

Can it be predicted and how?

RESEARCH QUESTIONS

Is it possible to predict news trend at D+1?

What kind of data is useful beyond publication counts?

INTRODUCTION

NEWS OUTLETS

Top 60 news site in the UK
September 15 to December 31
134.917 visits (desktop)

PEOPLE

1311 people
52% Men, 48% Women
[18-24] 10% -- [25-34] 16% -- [35-44] 17% -- [45-54] 24% -- [55-64] 25% -- [65 +] 7%
[<1000 £] 36% -- [1000 £-2000 £] 38% -- [2000 £ +] 21% -- [No income] 6%

UNDER THE HOOD

ACCESS TO THE CONTENT VISITED BY PANELISTS

Topics

X

Websites

THE Sun

BBC NEWS

THE TIMES

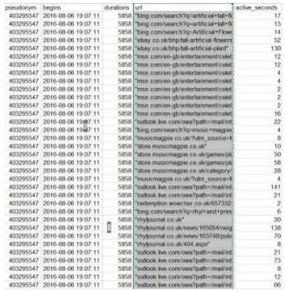
theguardian

Birmingham MAIL

X

Panelists

URLs visited



UNDER THE HOOD

DATA EXTRACTION

Machine learning – filtering of URLs

https://www.bbc.com/food

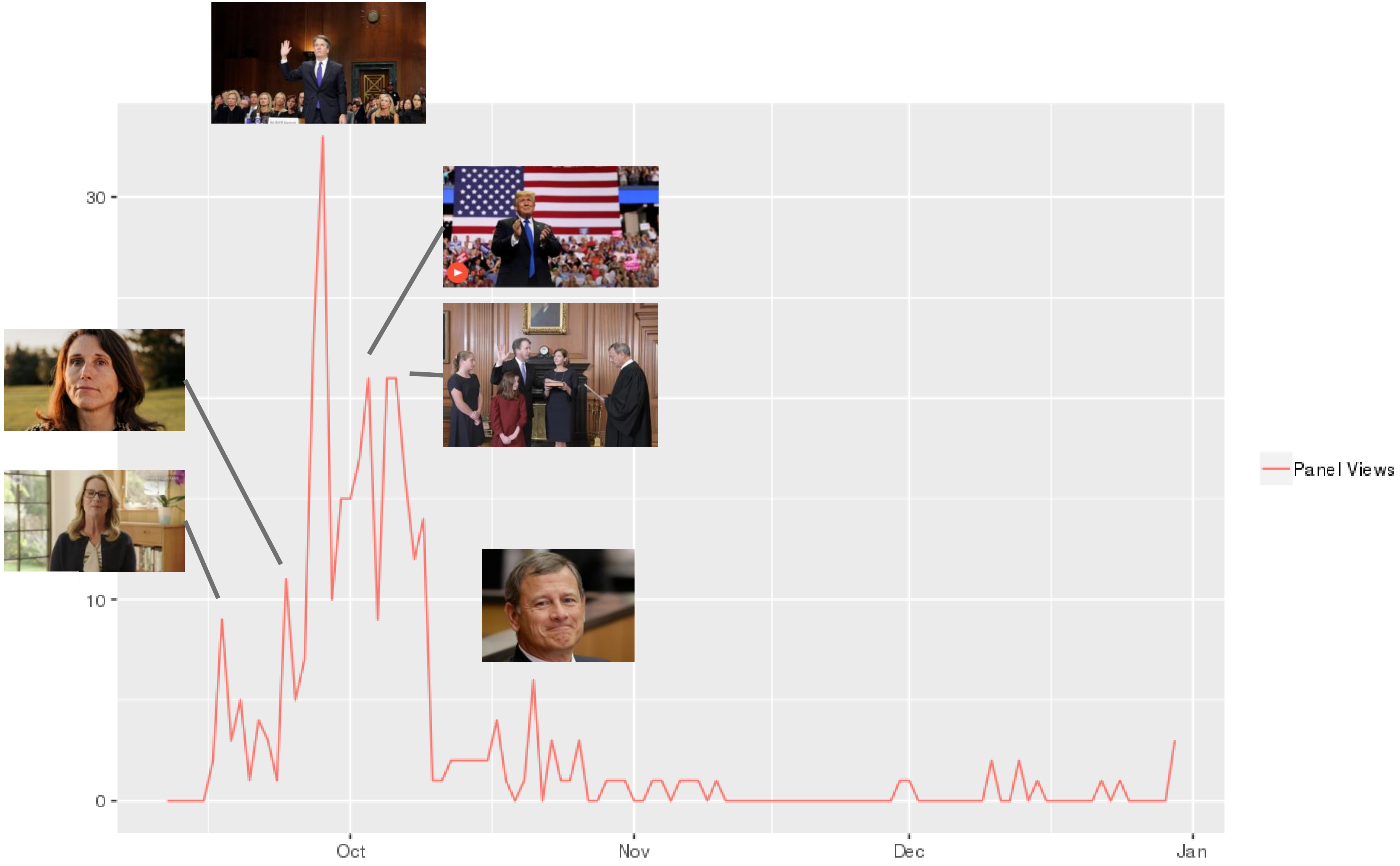
https://www.dailymail.co.uk/[...]/Pregnant-children-s-author-28-killed-roommate-money.html

Keywords extraction

IBM Watson

121.074 news viewed over the period by 1.311 panelists

TRENDS



Topic popularity across time
(articles views in the panel)

UNDER THE HOOD




TRAIN



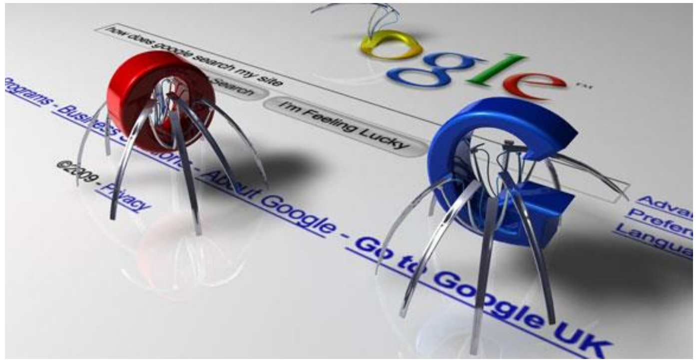





TEST


PREDICTORS

THREE FAMILIES OF PREDICTORS



respondi

CLOSE TO PEOPLE

CRAWLERS

Predictors based on publicly available information that anyone can get by scrapping webpages.

INSIDERS

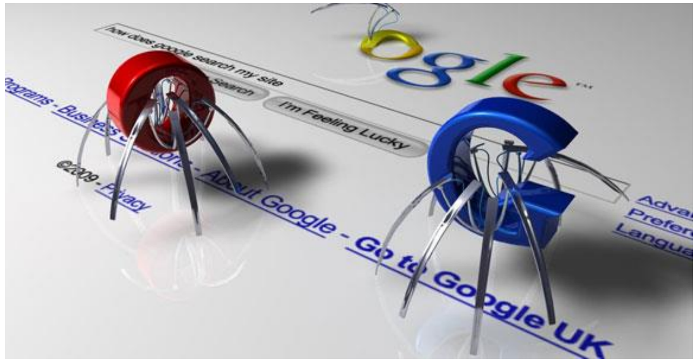
Predictors based on information available to the stakeholders, view counts on articles.

PANEL

Who reads what?

socio-demographics of viewers

MODEL 1



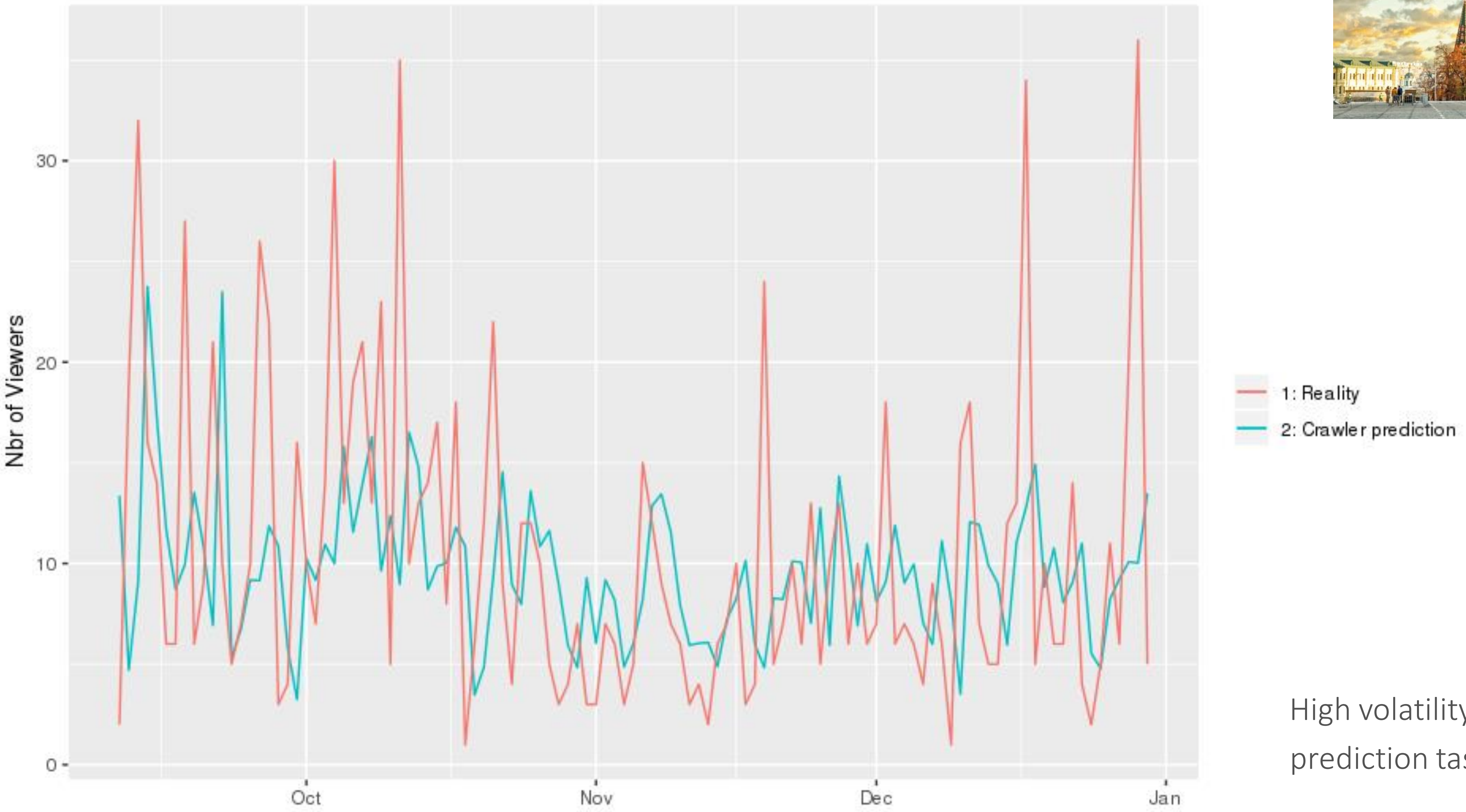
CRAWLERS

Predictors based on publicly available information

that anyone can get by scrapping webpages.

- SHORT TERM PUBLICATION COUNT
  how many articles published the day just before?
- MID TERM PUBLICATION COUNT
  how many articles published the three days before?
- RECENCY
  how long ago has the last article on the topic been published

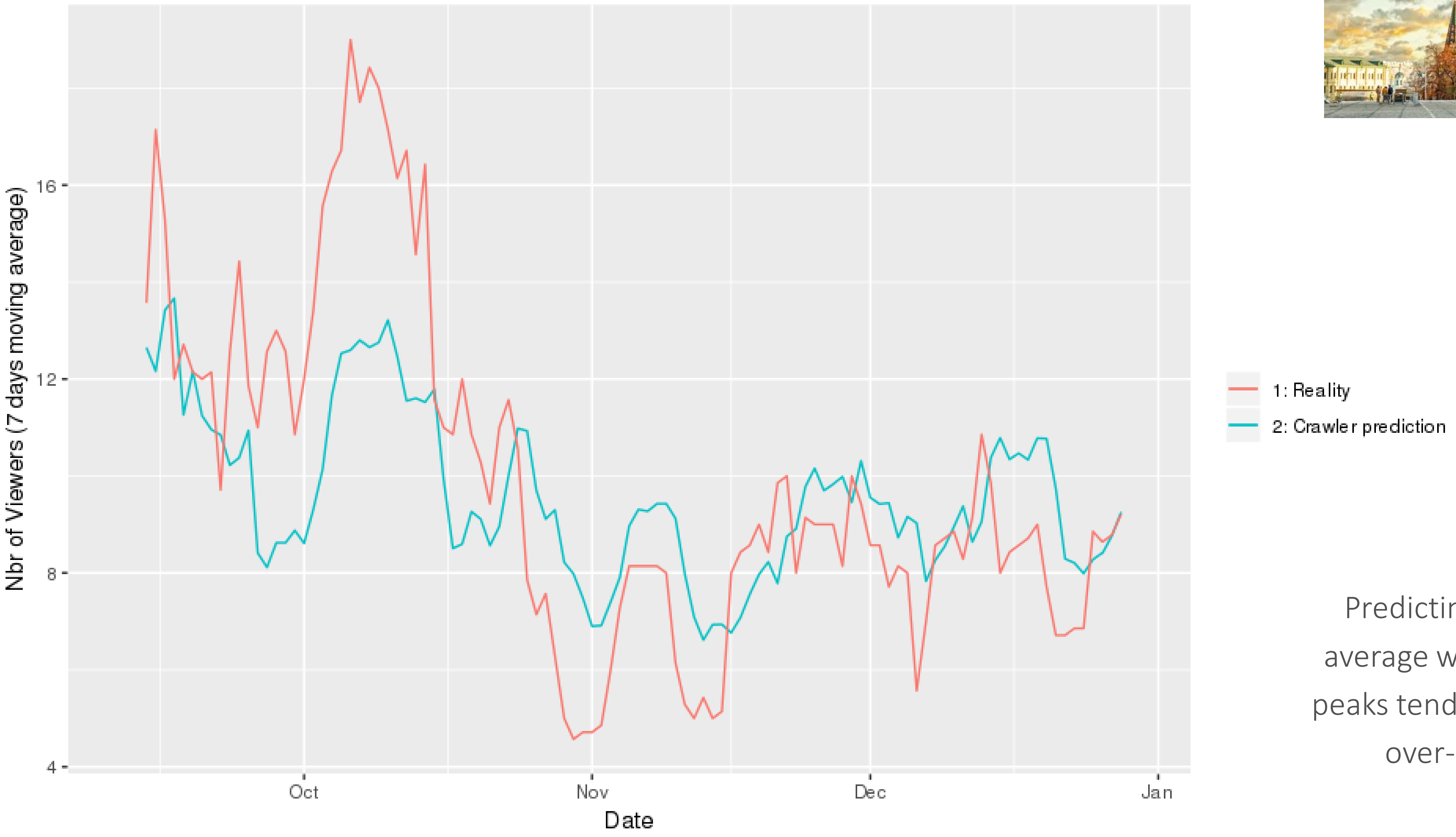
PREDICTION



News Trend
Prediction vs Reality

High volatility makes the prediction task daunting

PREDICTION



News Trend
Prediction vs Reality

Predicting the moving average works better but peaks tend to be under- or over-estimated.

MODEL 2

INSIDERS

Predictors based on information available to the stakeholders, view counts on articles.

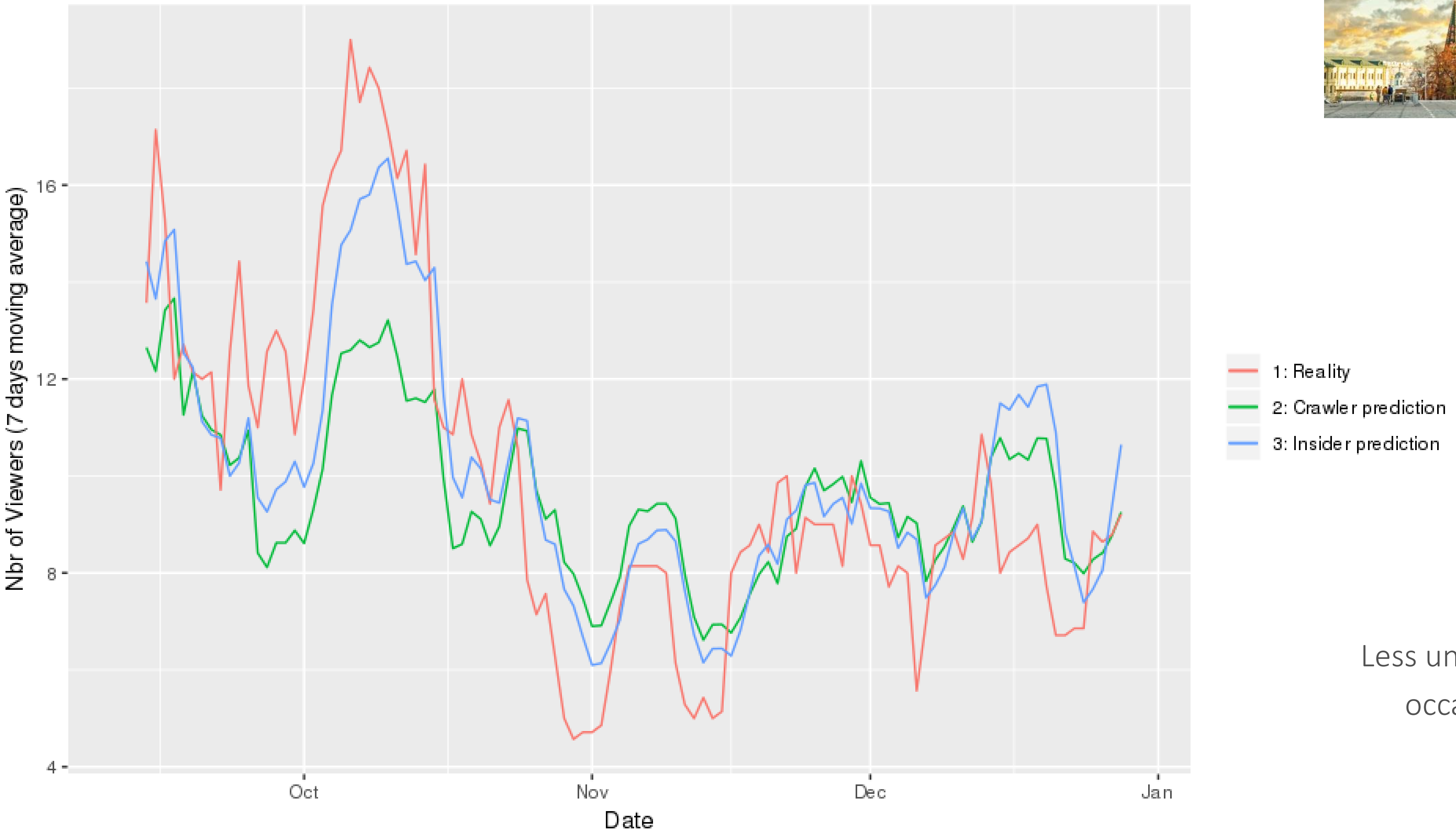
- ❑ MODEL 1 PREDICTORS

  (short and mid term publication count + recency)
- ❑ SHORT TERM VIEW COUNT

  how many people read about the topic the day just before?
- ❑ MID TERM VIEW COUNT

  how many people read about the topic the three days before?

PREDICTION



News Trend
Prediction vs Reality

Less undershooting but still occassionally misses the point respondi

CLOSE TO PEOPLE

PANEL

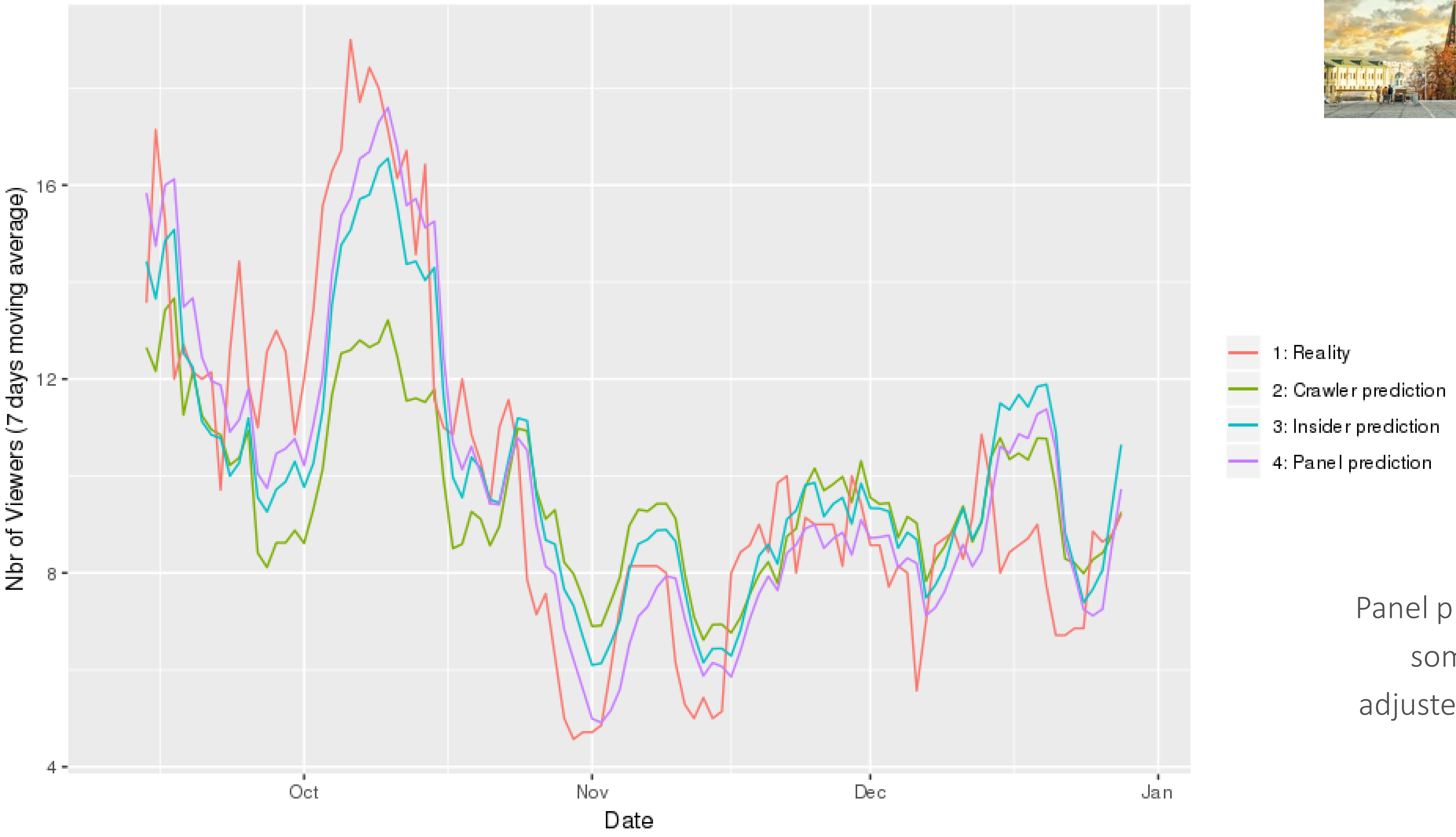
Who reads what?

socio-demographics of viewers

MODEL 3

- ❑ MODEL 2 PREDICTORS
  (model 1 predictors + short and mid term view count)
- ❑ INTEREST IN NEWS
  how many articles does this person read per day?
- ❑ INTEREST IN THE TOPIC
  did this person read about the topic over the last 5 days?
- ❑ INTEREST AMONG PEERS
  how many similar person read about the topic the day before?

PREDICTION

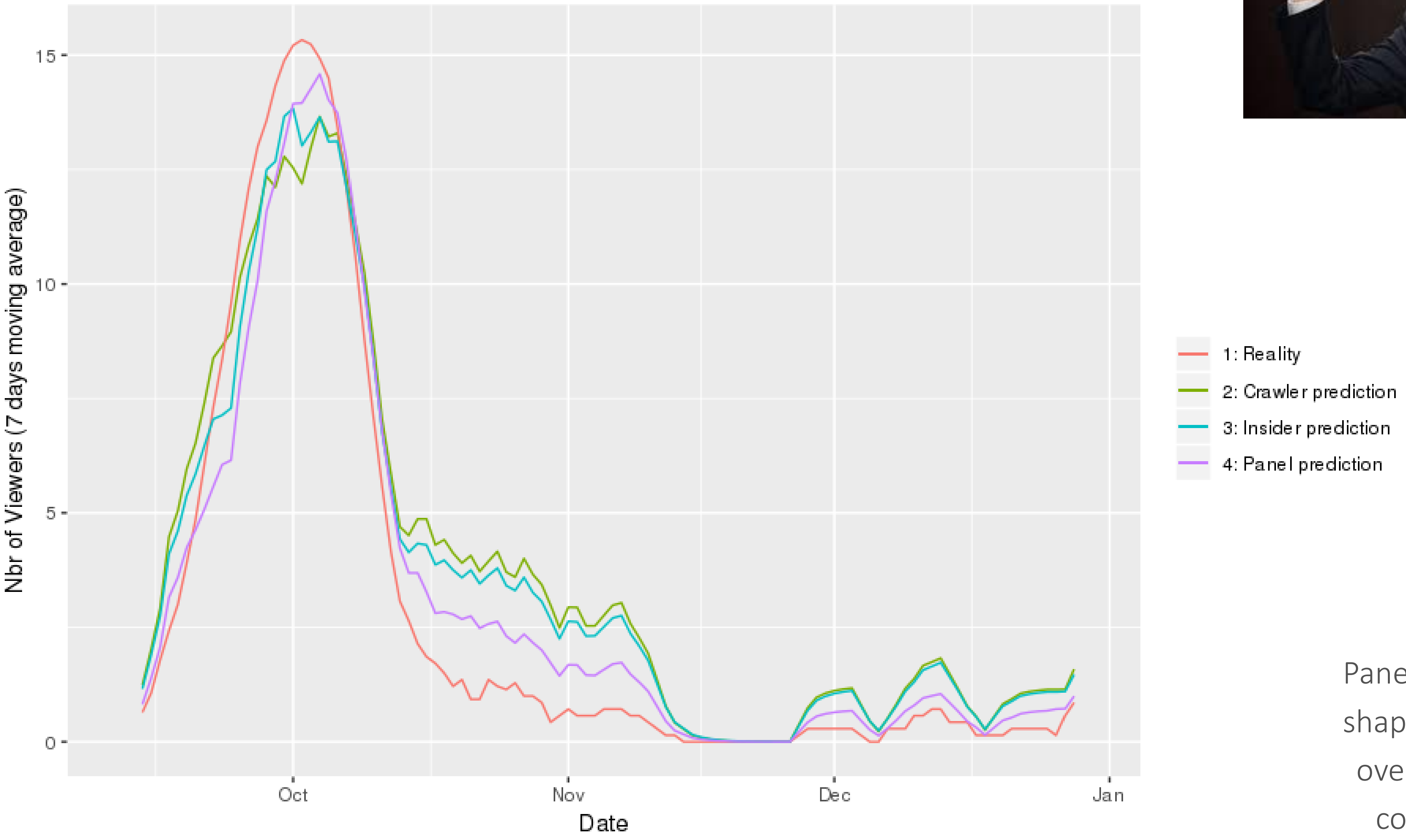


News Trend

Prediction vs Reality

Panel prediction allows for some fine-grained adjustements over insider prediction

PREDICTION

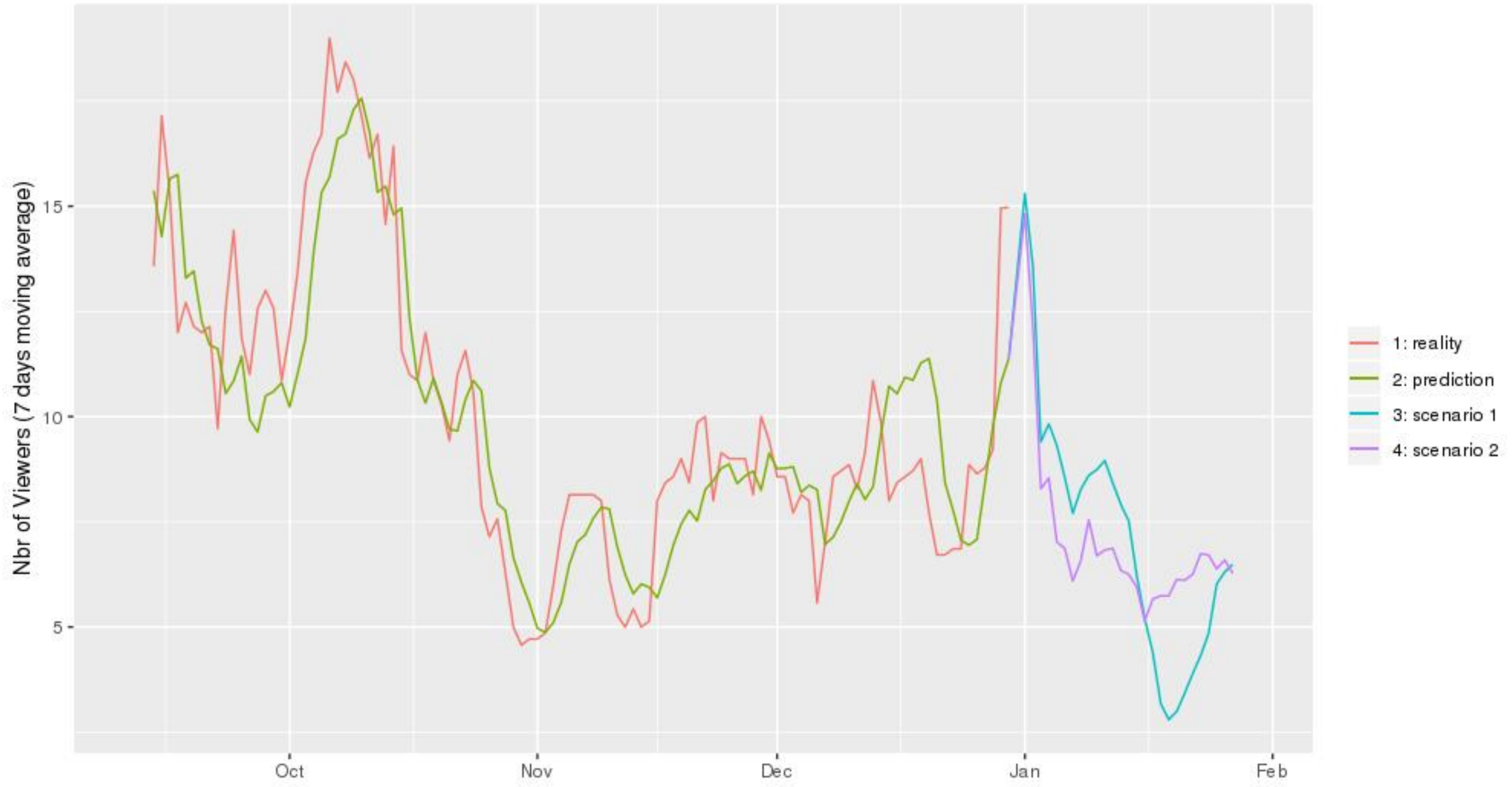


News Trend

Prediction vs Reality

Panel prediction gets the shape right and does not overshoot on the post-confirmation period

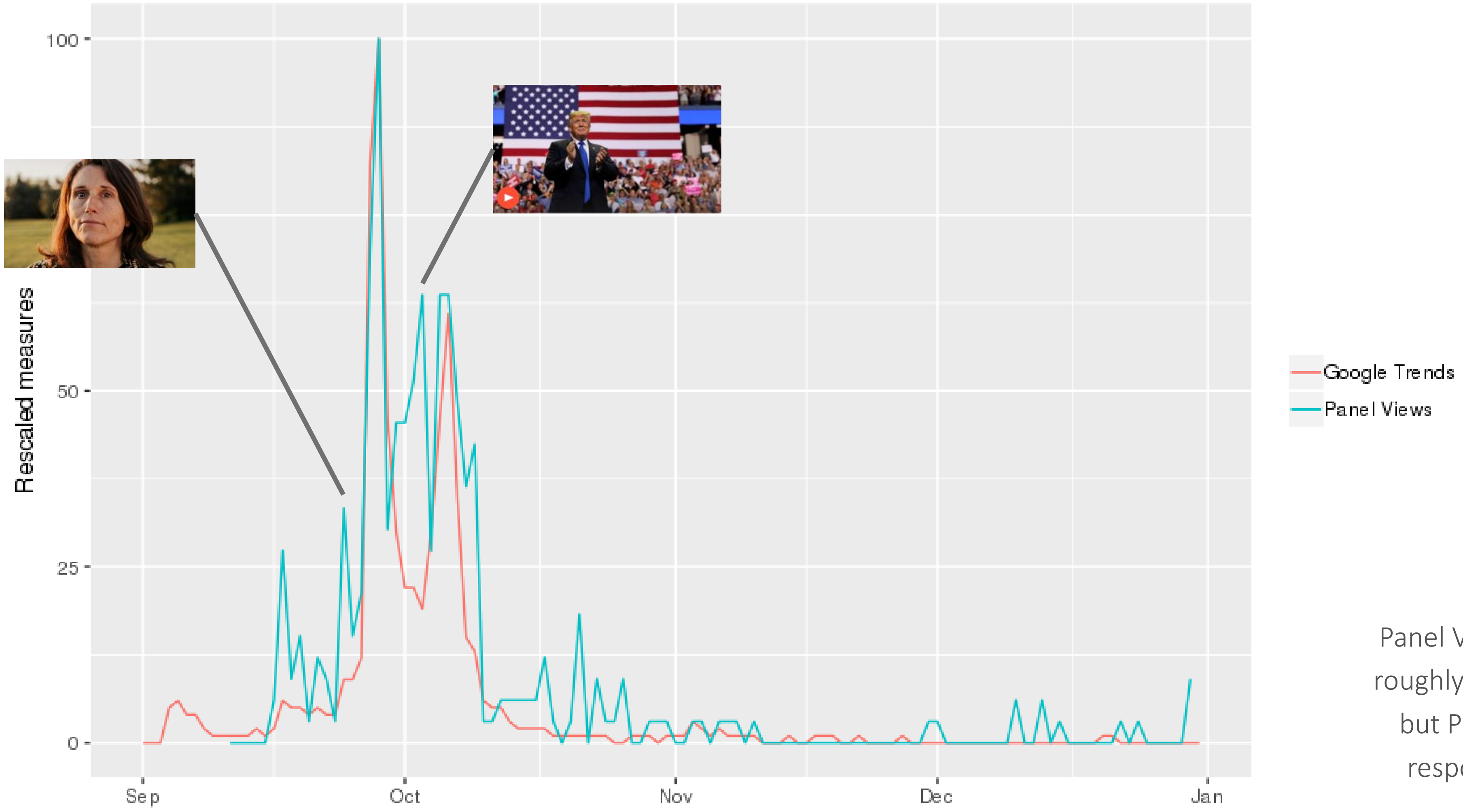
SIMULATION



One may use the models to simulate the effect of a media campaign. Inputs consist in a stochastic component (HMM fitted on past data) + a scenario (publication in the media on a given time frame). Imagine that you want to assess the effect of a media campaign promoting Russia:

- ❑ Scenario 1: media campaign from Jan 10 to Jan 20
- ❑ Scenario 2: no media campagin

TRENDS



Topic popularity across time
(Google Trends UK data)

Panel Views and Google Trends roughly follow the same pattern but Panel Views seem more responsive to actual news.

PERSPECTIVES

MUCH REMAINS TO BE DONE!

- ❑ Improve predictive accuracy
  - Fine tuning / Cross-topic modeling / Media sensitivity
- ❑ Improve outputs
  - Sentiments
- ❑ Improve simulations
  - Stochastic components / media types