Assessing Amazon Turker and automated machine forecasts in the Hybrid Forecasting Competition

Predictive Heuristics
Andreas Beger and Michael D. Ward

5 January 2019
Asian POLMETH 2019, Kyoto, Japan

I'm going to show results of an exploratory analysis of the relative accuracy of volunteer, Amazon Mechanical Turker, and automated machine forecasts for a broad set of questions (IFPs)[1] in the HFC[2] during the first trial period (RCT-A[3]) that took place this year.

[1] Individual Forecasting Problem
[2] Hybrid Forecasting Competition
[3] Randomized controlled trial A

Disclaimer

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17071900005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

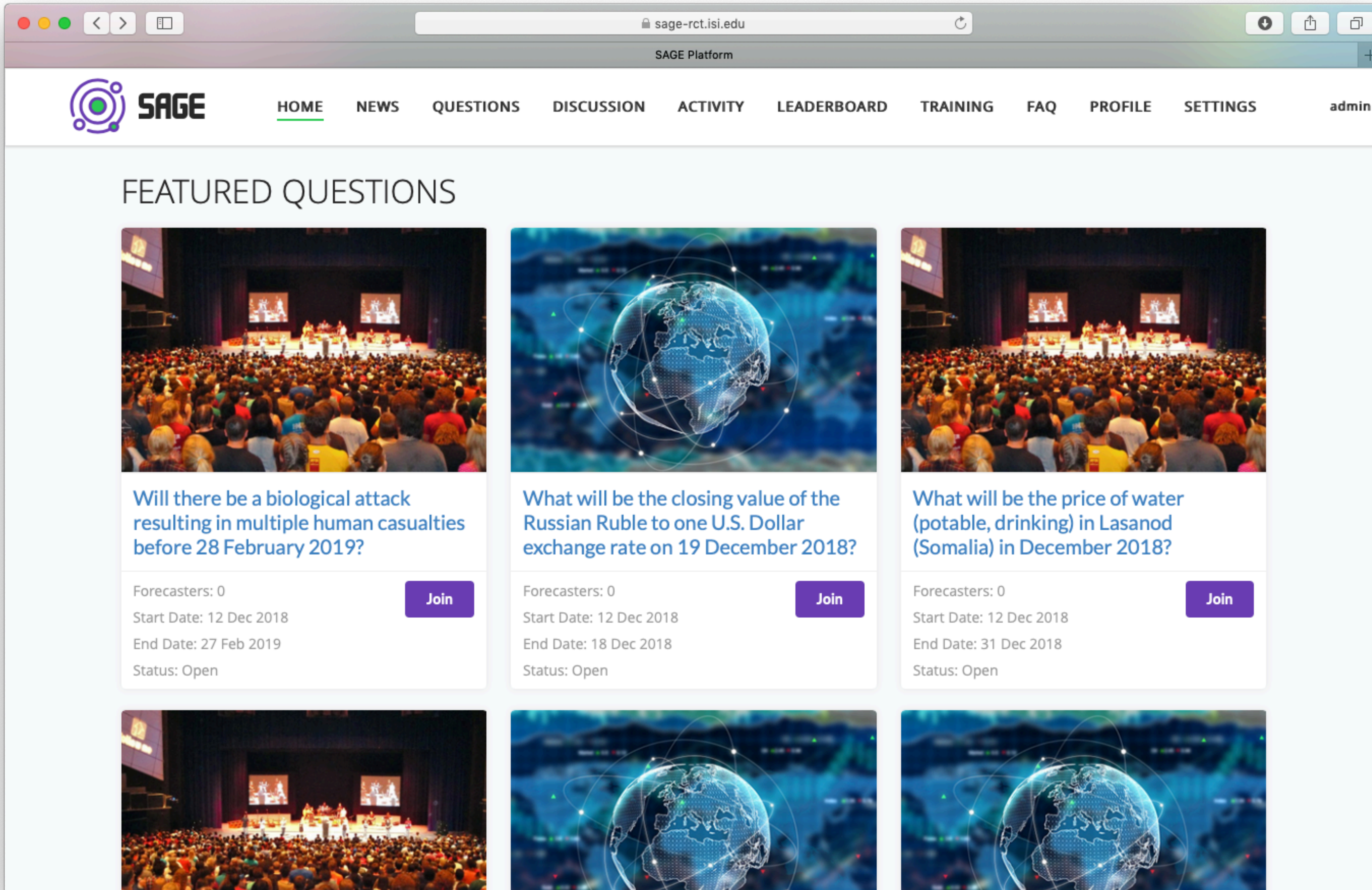
Hybrid Forecasting Competition

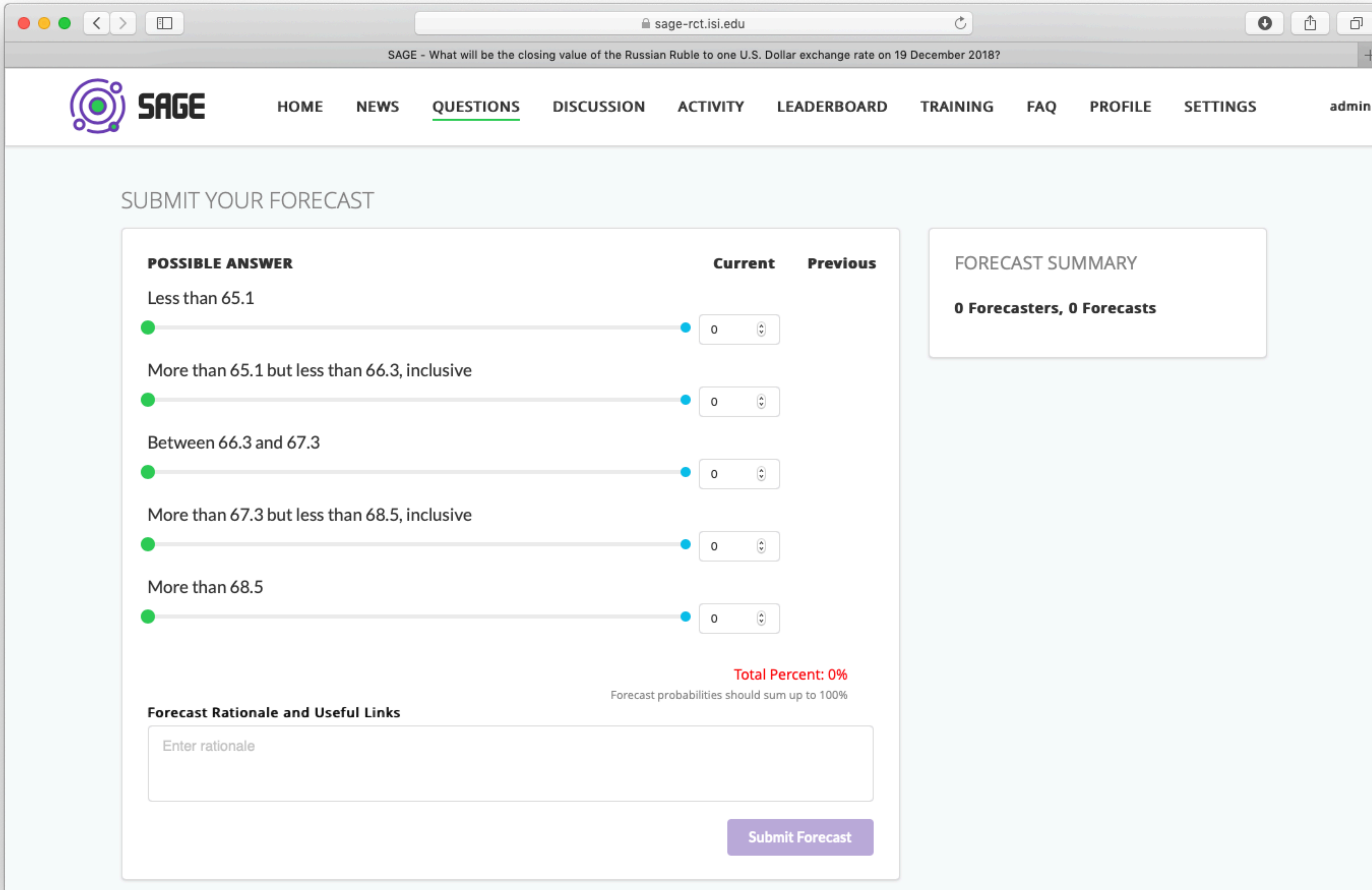
Project goal

> The HFC program is developing and testing hybrid geopolitical forecasting systems. These systems **integrate human and machine forecasting components** to create maximally accurate, flexible, and scalable forecasting capabilities.

From https://www.hybridforecasting.com

The idea is to overcome some of the respective weaknesses of human and machine-generated forecasts by combining them in some to be determined fashion.


sage-rct.isi.edu
SAGE Platform
SAGE
HOME
NEWS
QUESTIONS
DISCUSSION
ACTIVITY
LEADERBOARD
TRAINING
FAQ
PROFILE
SETTINGS
admin
FEATURED QUESTIONS
Will there be a biological attack resulting in multiple human casualties before 28 February 2019?
Forecasters: 0
Start Date: 12 Dec 2018
End Date: 27 Feb 2019
Status: Open
Join
What will be the closing value of the Russian Ruble to one U.S. Dollar exchange rate on 19 December 2018?
Forecasters: 0
Start Date: 12 Dec 2018
End Date: 18 Dec 2018
Status: Open
Join
What will be the price of water (potable, drinking) in Lasanod (Somalia) in December 2018?
Forecasters: 0
Start Date: 12 Dec 2018
End Date: 31 Dec 2018
Status: Open
Join


sage-rct.isi.edu
SAGE - What will be the closing value of the Russian Ruble to one U.S. Dollar exchange rate on 19 December 2018?
SAGE
HOME
NEWS
QUESTIONS
DISCUSSION
ACTIVITY
LEADERBOARD
TRAINING
FAQ
PROFILE
SETTINGS
admin
SUBMIT YOUR FORECAST
POSSIBLE ANSWER
Current
Previous
Less than 65.1
0
More than 65.1 but less than 66.3, inclusive
0
Between 66.3 and 67.3
0
More than 67.3 but less than 68.5, inclusive
0
More than 68.5
0
Total Percent: 0%
Forecast probabilities should sum up to 100%
Forecast Rationale and Useful Links
Enter rationale
Submit Forecast
FORECAST SUMMARY
0 Forecasters, 0 Forecasts

Scoring

Multinomial Brier score:

$$mBS = \sum_{i=1}^{R}(f_i - o_i)^2$$

- R is the number of answer options, f the vector of weights summing to 1, and o a 0/1 vector marking the correct option
- ranges from 0 (good) to 2 (bad)

Scoring

Multinomial Brier score:

$$mBS = \sum_{i=1}^{R}(f_i - o_i)^2$$

- R is the number of answer options, f the vector of weights summing to 1, and o a 0/1 vector marking the correct option
- ranges from 0 (good) to 2 (bad)

Ordered Brier score:

1. Split the ordinal categories (A-B-C-D) into cumulative binary pairs, aggregating the forecast probabilities for each grouping of categories (A-BCD; AB-CD; ABC-D).
2. Calculate the multinomial Brier score for each of the binary categories.
3. Average across the binary category scores to obtain the final Brier score.

- also ranges from 0 to 2
- "near misses" are penalized less than far misses

The hybrid part:

Some users could see time series charts and/or machine forecasts for some IFPs

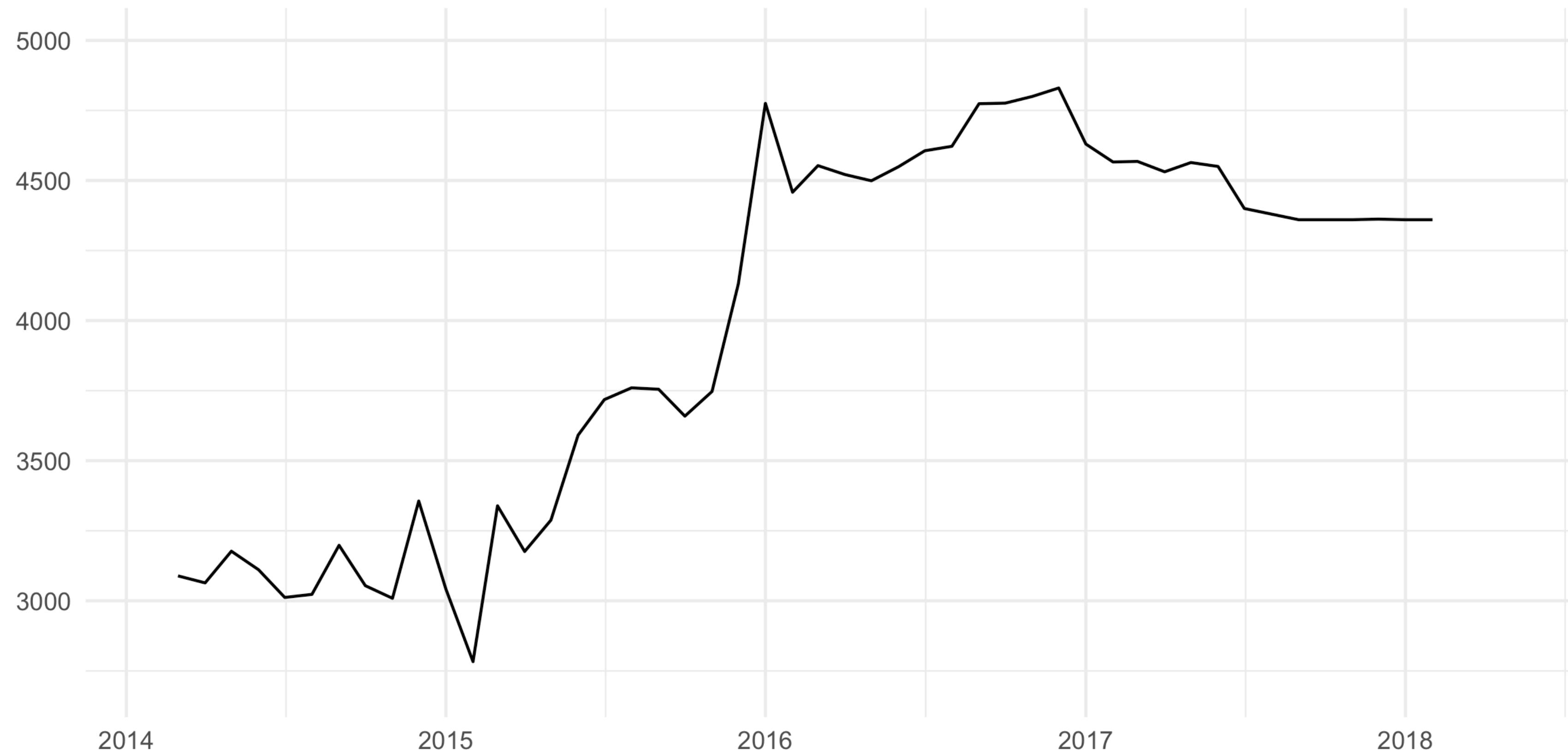
How this works

- An automated system is collecting and updating data from several data sources, and matching them up, if possible, to questions

If data for a particular question is available:

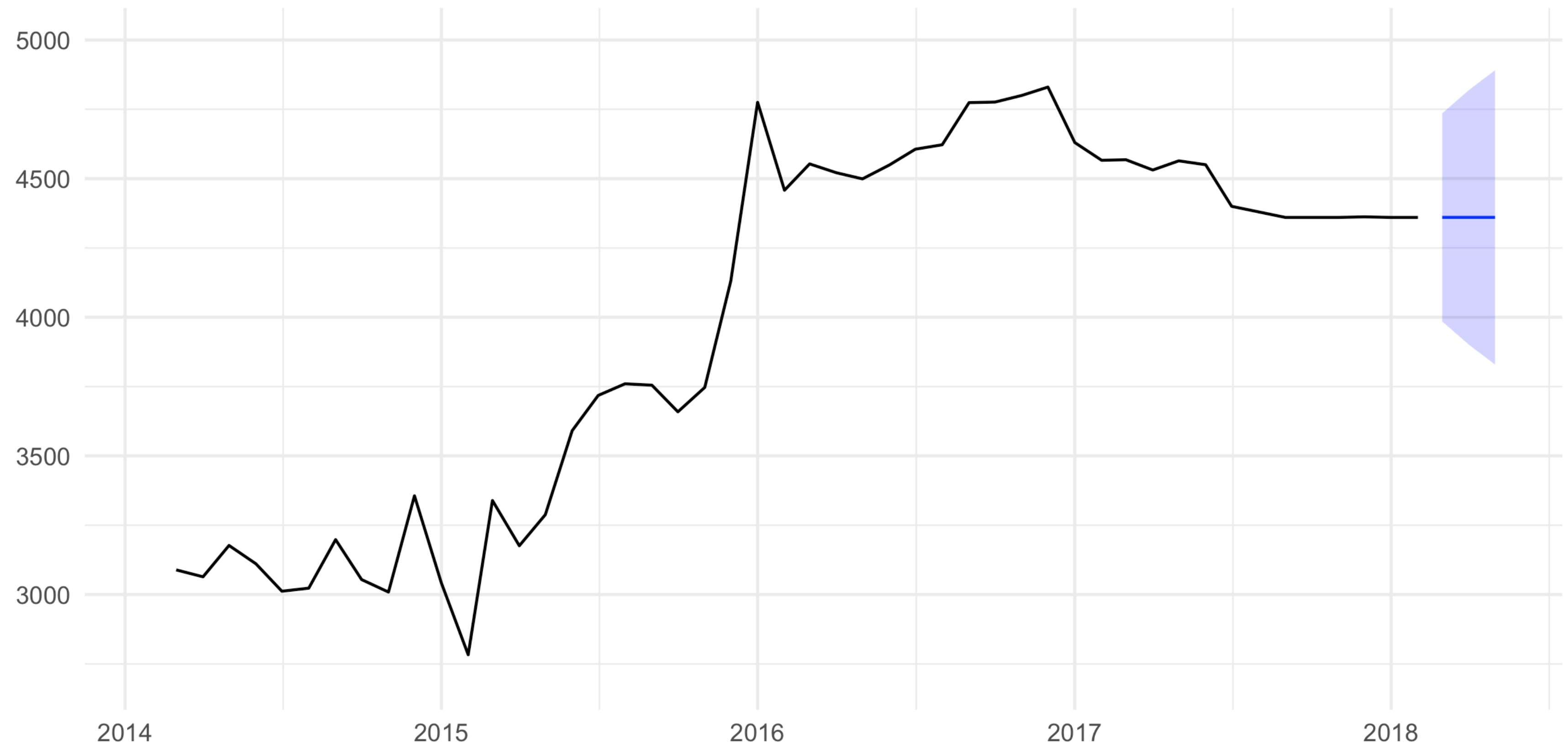
- Chart the data
- An automated system generates a machine forecast

How much crude oil will Iraq produce in May 2018?



- Less than 4,280
- Between 4,280 and 4,384, inclusive
- More than 4,384 but less than 4,473
- Between 4,473 and 4,576, inclusive
- More than 4,576

How much crude oil will Iraq produce in May 2018?



- Less than 4,280 *(38%)*
- Between 4,280 and 4,384, inclusive *(15%)*
- More than 4,384 but less than 4,473 *(13%)*
- Between 4,473 and 4,576, inclusive *(13%)*
- More than 4,576 *(21%)*

basil-ts

Automated time series forecaster microservice

Implemented in R + Python Flask + RESTful API

Sketch of the internals:

1. Parse incoming question and data
2. Produce a time series forecast using an automated ARIMA fitter[1]
3. Convert the time series forecast to answer option probabilities

Most of the complexity is related to automating the question/task parsing and handling edge cases.

[1] Hyndman & Khandakar 2008; R forecast package

Questions that were of primary concern to us

How well are the machine forecasts doing?

Are there particular question groups where performance is good or lacking? Basically, where do we need to focus improvements?

Let's start looking at data and results

Data

RCT-A, the first trial period, lasted from 7 March to 7 September 2018

- Use subset of forecasts from 2 May (turkers enter) to 1 August 2018 (change in tracking)

~49,000 forecasts

156 IFPs; 46 with machine forecasts

971 unique users

Forecasters:

- Volunteers who joined the platform
- Amazon Mechanical Turkers
- Machine (basil-ts)

Summary of some findings in the paper

Summary of some findings in the paper

Volunteers are better forecasts than Amazon Turkers

Machines are somewhere in the middle

Summary of some findings in the paper

Volunteers are better forecasts than Amazon Turkers

Machines are somewhere in the middle

Human forecasters who saw the machine forecasts did poorly; volunteer forecasts who saw only charts had overall the best performance

Summary of some findings in the paper

Volunteers are better forecasts than Amazon Turkers

Machines are somewhere in the middle

Human forecasters who saw the machine forecasts did poorly; volunteer forecasts who saw only charts had overall the best performance

It's not clear why though, e.g.:

Summary of some findings in the paper

Volunteers are better forecasts than Amazon Turkers

Machines are somewhere in the middle

Human forecasters who saw the machine forecasts did poorly; volunteer forecasts who saw only charts had overall the best performance

It's not clear why though, e.g.:

- they outperformed on questions that did not have data (charts and model)

Summary of some findings in the paper

Volunteers are better forecasts than Amazon Turkers

Machines are somewhere in the middle

Human forecasters who saw the machine forecasts did poorly; volunteer forecasts who saw only charts had overall the best performance

It's not clear why though, e.g.:

- they outperformed on questions that did not have data (charts and model)
- they outperformed volunteer forecasters who saw the machine forecasts even when the machine forecasts had good accuracy

Summary of some findings in the paper

Volunteers are better forecasts than Amazon Turkers

Machines are somewhere in the middle

Human forecasters who saw the machine forecasts did poorly; volunteer forecasts who saw only charts had overall the best performance

It's not clear why though, e.g.:

- they outperformed on questions that did not have data (charts and model)
- they outperformed volunteer forecasters who saw the machine forecasts even when the machine forecasts had good accuracy

Relative performance: where did the machine forecasts do better than human forecasters?

Performance by forecaster group

Forecaster	avg Brier	N
Machine	0.402	1975
Turker	0.433	39140
Volunteer	0.322	7816

Relative forecaster performance

Group questions by data source, N=14 groups

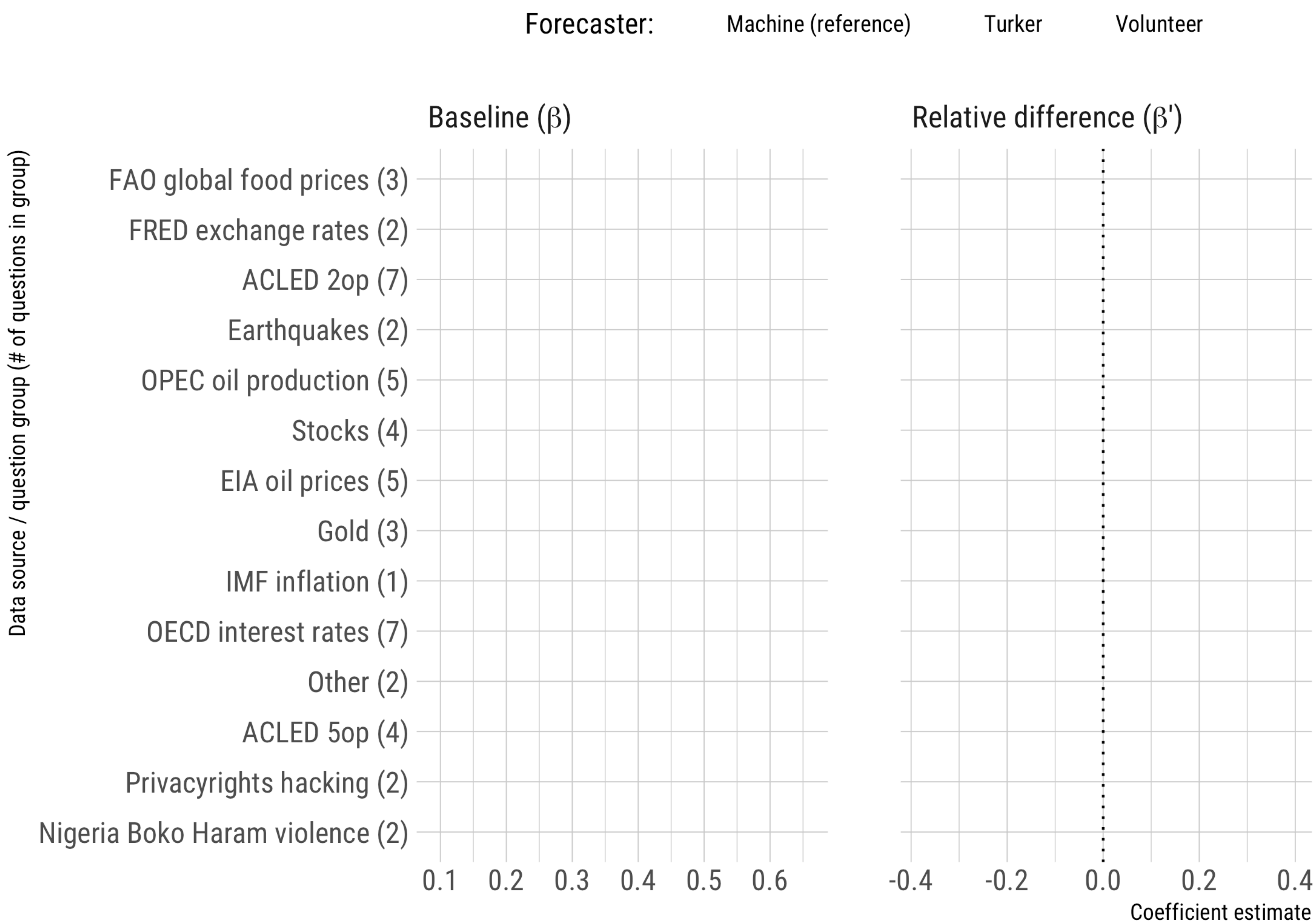
Linear model to compare average Brier scores of volunteer, turker, and machine forecasts (reference category)

$$\text{Brier}_{ij} = \beta_i \text{DataSource}_i + \beta'_{ij}(\text{DataSource}_i \times \text{Forecaster}_j)$$

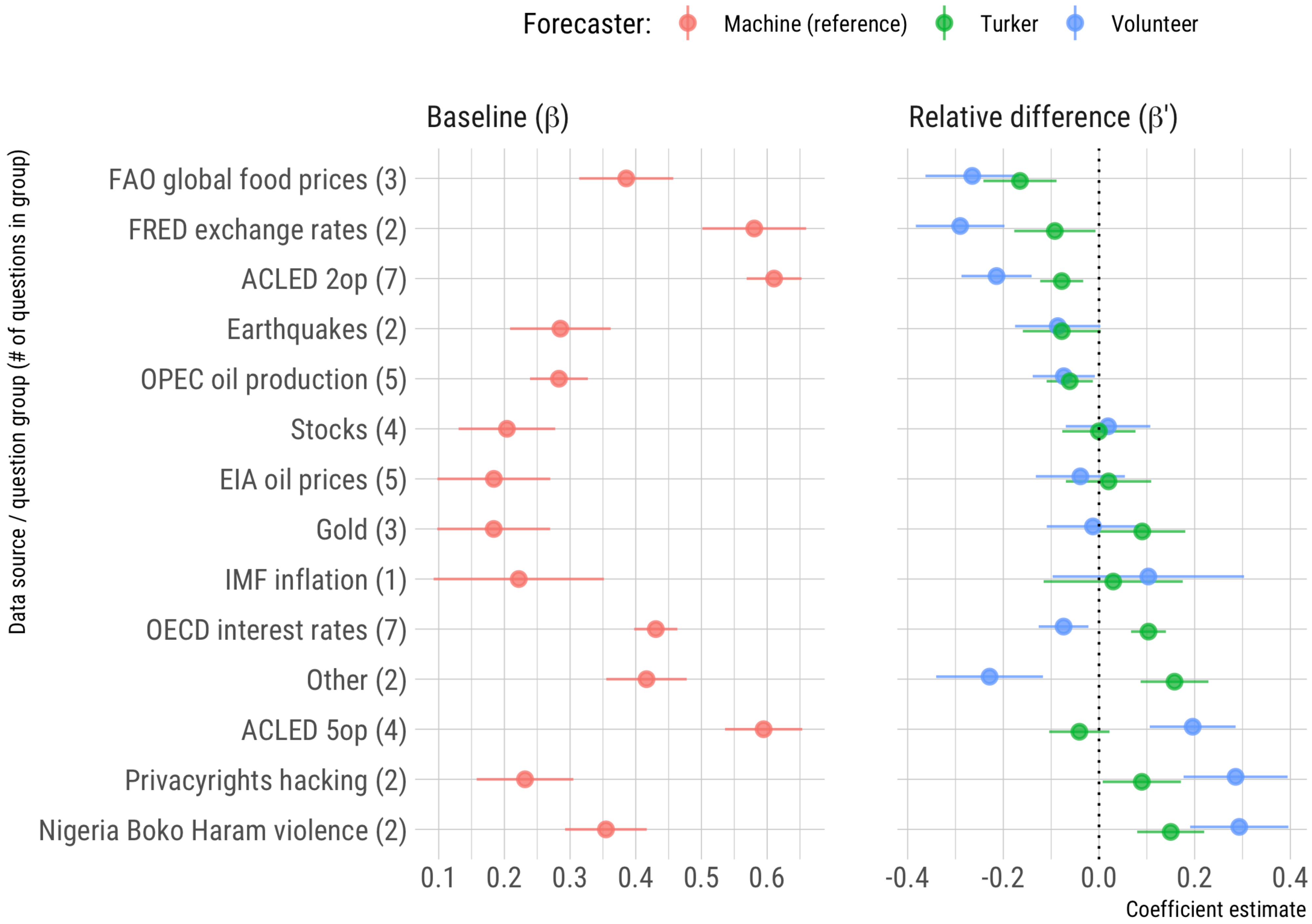
β_i = average Brier score for machine forecasts for question group i

β'_{ij} = average Brier score *relative to machine forecasts* for forecaster group j in question group i

- Negative values $\rightarrow$ human forecasters did better
- Positive values $\rightarrow$ machine forecasts did better



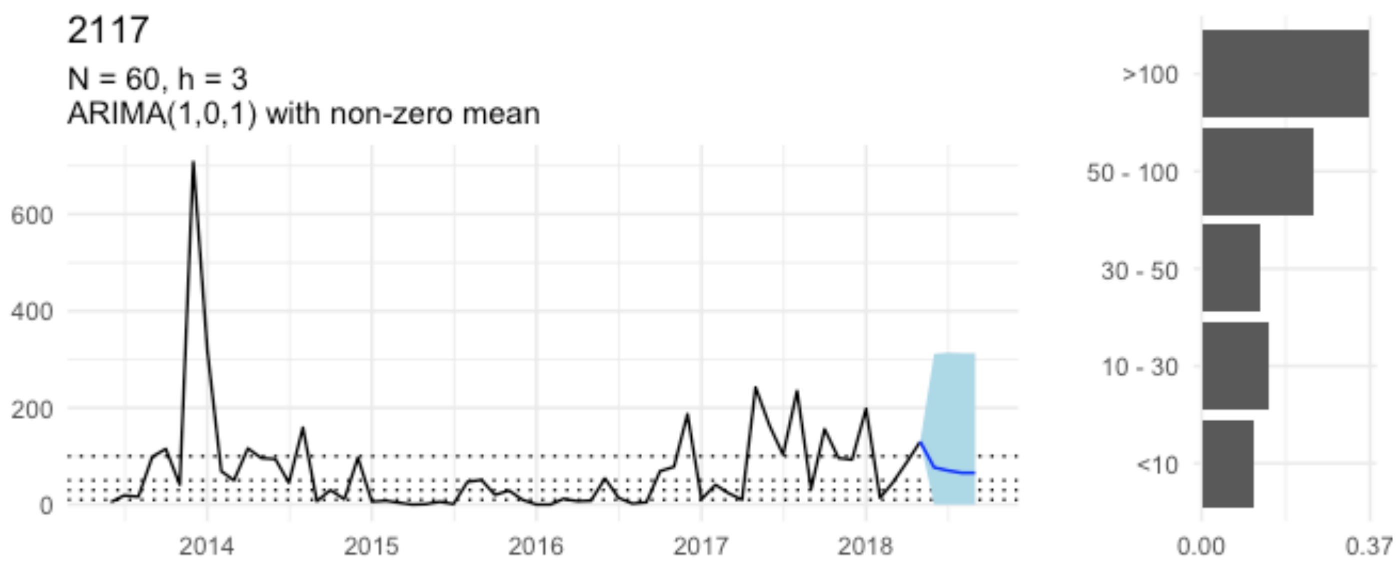
Forecaster:
Machine (reference)
Turker
Volunteer
Baseline (β)
Relative difference (β')
Data source / question group (# of questions in group)
FAO global food prices (3)
FRED exchange rates (2)
ACLED 2op (7)
Earthquakes (2)
OPEC oil production (5)
Stocks (4)
EIA oil prices (5)
Gold (3)
IMF inflation (1)
OECD interest rates (7)
Other (2)
ACLED 5op (4)
Privacyrights hacking (2)
Nigeria Boko Haram violence (2)
0.1
0.2
0.3
0.4
0.5
0.6
-0.4
-0.2
0.0
0.2
0.4
Coefficient estimate


Forecaster:
Machine (reference)
Turker
Volunteer
Baseline (β)
Relative difference (β')
Data source / question group (# of questions in group)
FAO global food prices (3)
FRED exchange rates (2)
ACLED 2op (7)
Earthquakes (2)
OPEC oil production (5)
Stocks (4)
EIA oil prices (5)
Gold (3)
IMF inflation (1)
OECD interest rates (7)
Other (2)
ACLED 5op (4)
Privacyrights hacking (2)
Nigeria Boko Haram violence (2)
0.1
0.2
0.3
0.4
0.5
0.6
-0.4
-0.2
0.0
0.2
0.4
Coefficient estimate

ACLED 5-option

Overall hard questions; machine outperformed human forecasts

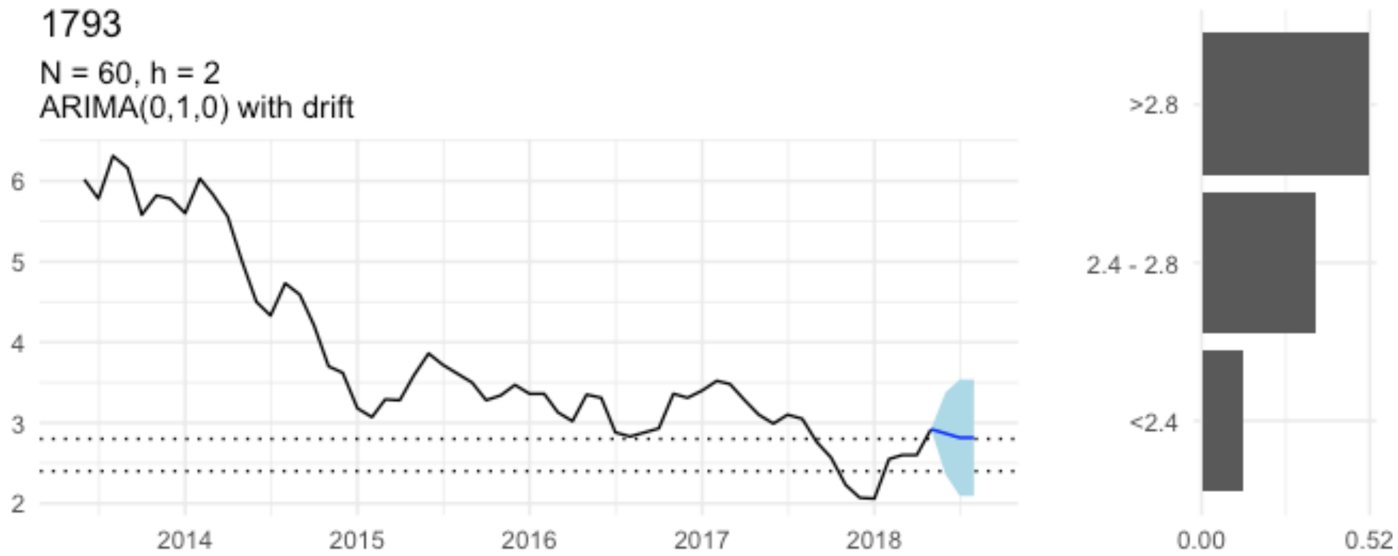
> How many battle deaths will ACLED record in Central African Republic in August 2018?



OECD interest rates

Overall hard questions; machine has middling performance

> What will be the long-term interest rate for Hungary (HUN) in July 2018?



ACLED binary question

Relatively easy questions, but machines underperformed a lot

> Will ACLED record any riot/protest events in Gambia in July 2018?

Conclusions

Machine forecasts did well on count questions that require data aggregation

- ACLED
- Privacyrights hacking
- Nigeria security tracker (Boko Haram)

Conclusions

Machine forecasts did well on count questions that require data aggregation

- ACLED
- Privacyrights hacking
- Nigeria security tracker (Boko Haram)

Some of the overall hardest questions, and where to some extent human forecasters did better, are economic/financial monthly series

- OECD interest rates
- FAO food price indices
- exchange rates
- oil production

Issues we did not address

Issues we did not address

What about selection issues; are volunteer forecasters able to self-select into questions they will do well on?

Issues we did not address

What about selection issues; are volunteer forecasters able to self-select into questions they will do well on?

Do the chart volunteers do better on questions requiring data aggregation; generally there are some inconsistencies on why/how the chart volunteers did so well.

Thank you!

Register to forecast at https://sage-platform.isi.edu/

: adbeger@gmail.com

: https://github.com/andybega/asia-polmeth-2019

: link to paper (on github under docs/pdf/)