Three IARPA forecasting efforts: ICPM, HFC, and the Geopolitical Forecasting Challenge

Jonathan McHenry (Booz Allen Hamilton, on behalf of IARPA)

Acknowledgement: This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) via contracts 2009-0917826-016, 2015-14120200002-002, and other vehicles, and is subject to the Rights in Data – General Clause 52.227-14, Alt. IV (DEC2007).
Any views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA or the U.S. Government.

IARPA Mission and Method

IARPA's mission is to envision and lead high-risk, high-payoff research that delivers innovative technology for future overwhelming intelligence advantage

- **Bring the best minds to bear on our problems**
  - Full and open competition to the greatest possible extent, funding scientists and engineers in academia and industry, through contracts, grants, OTs, and prize challenges
  - World-class, rotational Program Managers
- **Define and execute research programs that:**
  - Have goals that are clear, measureable, ambitious and credible
  - Employ independent and rigorous Test & Evaluation
  - Involve IC partners from start to finish
  - Run from three to five years
  - Publish peer-reviewed results and data, to the greatest possible extent

The Hybrid Forecasting Competition

Human and machine forecasting systems each have relative strengths and weaknesses:

- Humans are **adaptive**, can reason about **new cases**, and apply their **real-world knowledge** to problems.
- However, they can be **slow**, **biased**, and are **subject to fatigue**.

- Machines are **fast**, **consistent,** and **tireless**.
- However, they tend to be **rigid**, and can be highly dependent on **training data**.

HFC is a 4-year competition to advance geopolitical forecasting by combining the strengths of humans and machines. HFC systems compete to produce accurate forecasts on large numbers of questions covering a wide range of topics. The breadth of topics and number questions will exceed the limits of either crowdsourced or machine forecasting systems, so that only hybrid systems can prevail.

IARPA provides each system with a stream of SotA crowdsourced forecasts, along with randomly assigned human participants.

https://www.hybridforecasting.com/, https://www.iarpa.gov/index.php/research-programs/hfc

GEOPOLITICAL FORECASTING CHALLENGE

- A competition to combine **public data** with an IARPA-provided ACE-like **stream of human forecasts** (state-of-the-art, from **HFC**) in order to **accurately forecast** a wide variety of **geopolitical events**, such as elections, conflicts, disease outbreaks, and macro-economic indicators.
- Runs for seven months, with $200k in prizes, using about **25 forecasting questions per month** (from **HFC**), like:
    - Will the WHO confirm >10 cases of Marburg in 2018?
    - Before March 2018, will South Korea file a WTO dispute related to solar panels against the United States?
    - Who will win the 2018 presidential election in Egypt?

CHALLENGE TIMELINE

DEC 2017
Registration opens

FEB 2018
Challenge launches

AUG 2018
Challenge closes

FALL 2018
Final winners announced

https://www.iarpa.gov/challenges/gfchallenge.html
https://www.iarpa.gov/index.php/working-with-iarpa/prize-challenges/1070-geopolitical-forecasting-challenge

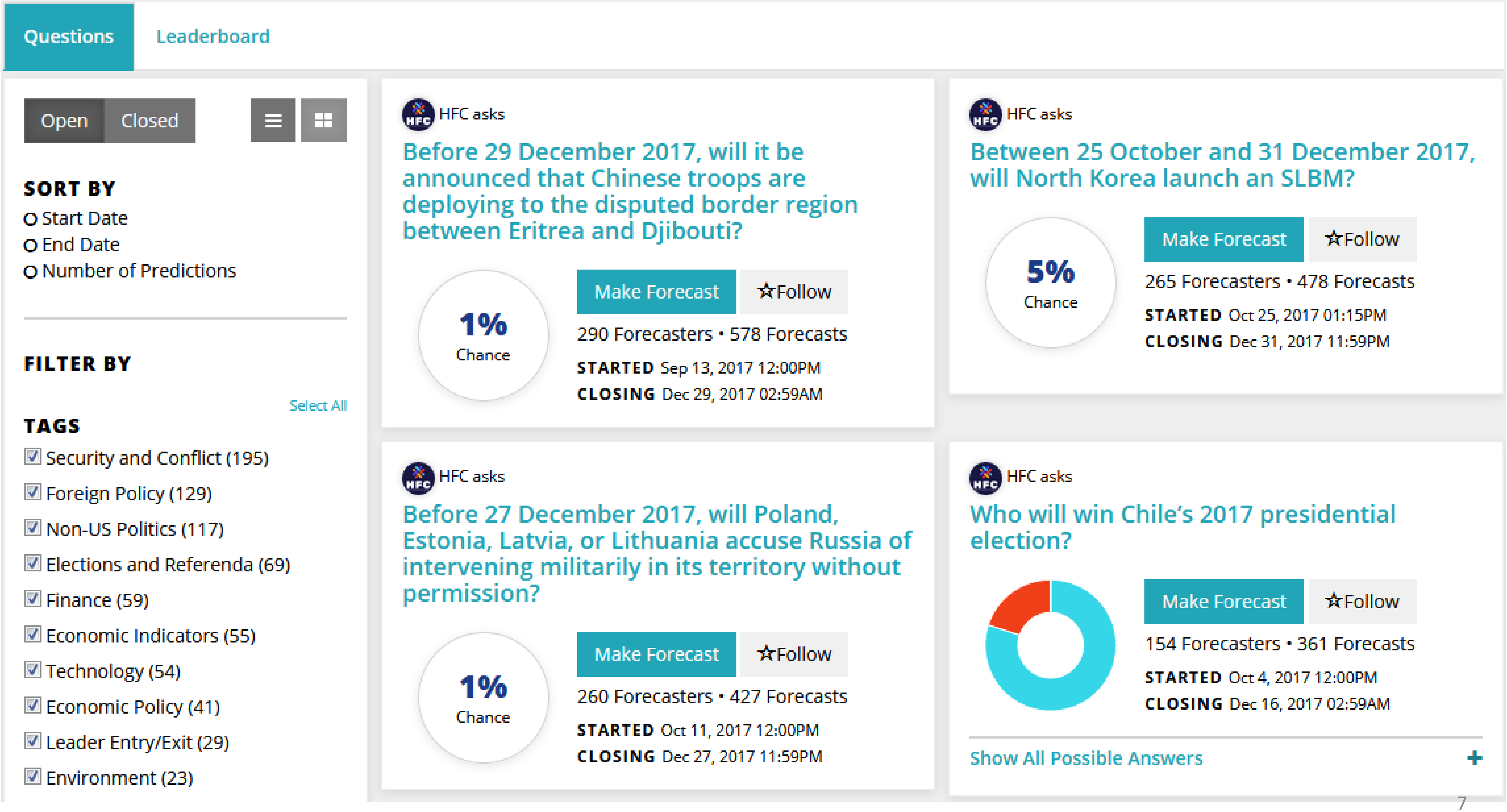
The Intelligence Community Prediction Market (ICPM)

- Since 2010, the US Intelligence Community has run ICPM on its classified network.
- ICPM users are Top Secret cleared gov't employees and contractors from across the IC.
- Participants use non-monetary points to buy and sell shares of answers to intelligence questions, such as potential event outcomes.
  → The resulting "price" serves as ICPM's consensus prediction for each question.
- Impetus behind ICPM: allow quick collaboration & settling on a numerical consensus.

- Participation is voluntary: no material (e.g., financial or administrative) benefit.
- ICPM has the largest dataset on the accuracy of analytic judgments in the history of the IC, including >190,000 predictions made by >4,300 users on a large array of geopolitical questions.

Backup

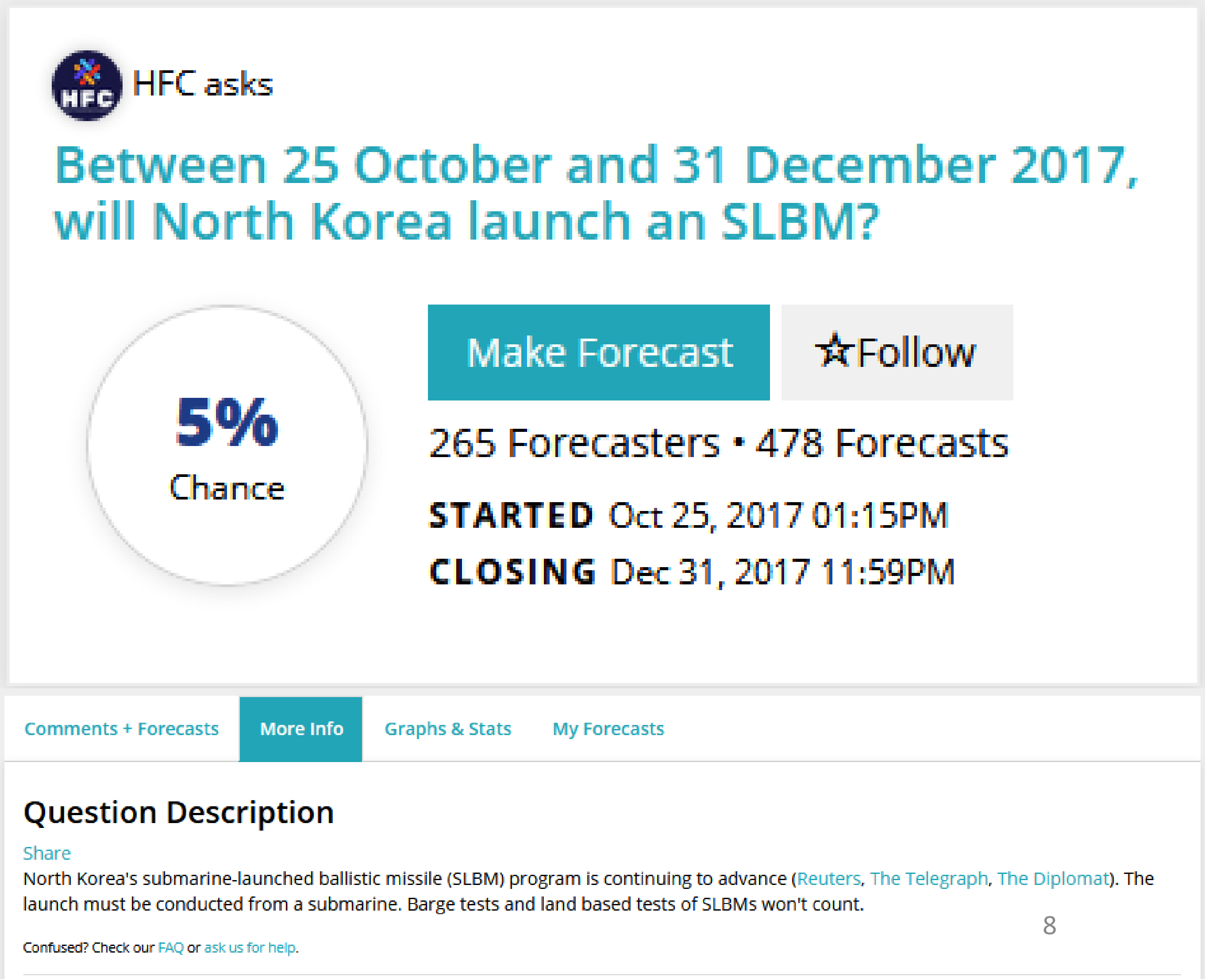
What is a prediction market? → Here is an example of a prediction market interface.



Source = https://www.gjopen.com/challenges/19-coming-soon-hfc-challenge?sort=prediction_sets_count -- retrieved 2017-12-01

Example prediction market forecasting question (FQ)

- Pay 0.05 to buy 1 “yes” share.
  - Receive 0.05 to sell 1 “yes” share.
- Pay 0.95 to buy 1 “no” share.
  - Receive 0.95 to sell 1 “no” share.
- Market “price” (probability) increases when shares of “yes” are purchased, decreases when shares of “yes” are sold, and does the opposite for “no” shares.
- When a FQ resolves, users receive one point for each share of the correct outcome owned.



Prediction markets generate probabilities over time

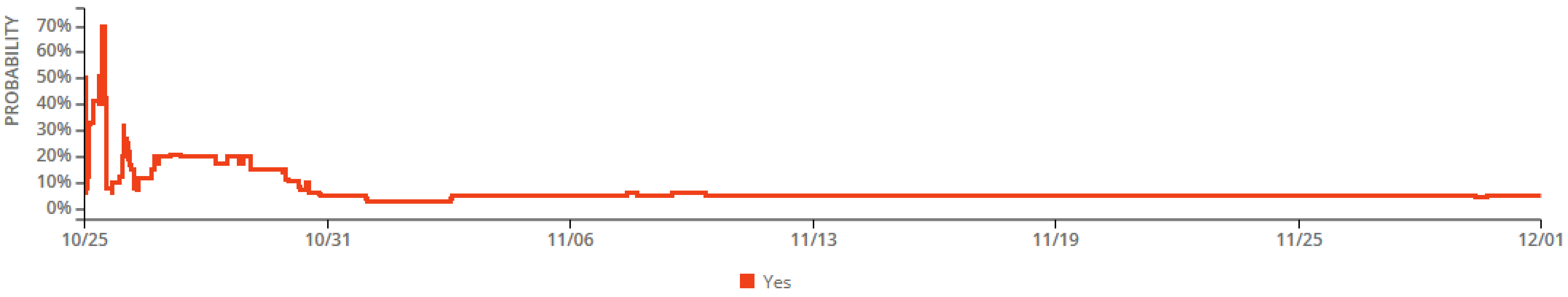
Comments + Forecasts | More Info | **Graphs & Stats** | My Forecasts

No. of Forecasters	Forecast Count	Forecasts in the last 24 hours	No. of Forecasts You've Made
265	478	8	0

Consensus Trend

Share

Click and drag on the small graph below to zoom



Source = https://www.gjopen.com/questions/666-between-25-october-and-31-december-2017-will-north-korea-launch-an-slbm -- retrieved 2017-12-01

User Comments

- Users give rationales for their forecasts, and give feedback to each other.

mwokasch made a forecast:

CHANCE 1%

ANSWER Yes

I think there is a misconception regarding the number of North Korean SLBM tests that have occurred in the past. This is partially because they have been testing SLBM missiles from land sites and have only recently transitioned to sea based launches. Instead of counting the number of SLBM launches, we should be analyzing the total number of tests of the Pukkuksong, which is North Korea's main SLBM.

Here is a list of the relevant launches:

Year: Launches (Sea based)
2014: 3 (1)
2015: 5 (3)
2016: 3 (2)
2017: 2 (1)

You'll notice a clear research testing distribution (starts with few launches, increases, and then declines when success are regular enough).

Moreover, following the May 2017 launch of the land-based Pukguksong-2, North Korea announced that:
1) this would be the upgraded version of the SBLM and
2) Mass production would soon begin soon

The only question that remains is: When the Pukguksong-2 comes off the assembly line, will North Korea find it necessary to conduct a Submarine test launch. As a reminder, the Pukguksong-2 has only ever been fired from land. However, I contest that it is not necessary to conduct a sea based test since its technical specs are similar and an upgraded version to the Pukguksong-1 that they have fired repeatedly. Moreover, I'm not certain they would want to test fire from a submarine in the near term since a failure would entirely change our perception of their SLBM arsenal.

As such, I have set my probability of a launch in 2017 incredibly low.

8 Upvote Reply Flag Link OCT 31, 2017 12:51PM

ebeeler made a comment:

Good analysis! But lets not forget the political dimension. It will be a very strong message towards the US to make successful SBLM launches. Kim has repeatedly shown that he's willing to do tests to show strength towards the US.

2 Upvote Flag Link NOV 1, 2017 03:02PM

Source = https://www.gjopen.com/questions/666-between-25-october-and-31-december-2017-will-north-korea-launch-an-slbm -- retrieved 2017-12-01

NATO Science and Technology Organization **System Analysis and Studies** (SAS) Panel, 2017-12-05

Comparative Evaluation of the Forecast Accuracy of Analysis Products and a Prediction Market

Jonathan McHenry (Booz Allen Hamilton, on behalf of IARPA)

presenting the work of Bradley J. Stastny and Paul E. Lehner,

with Steve Rieber (IARPA) joining for discussion

Acknowledgement: This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) via contract 2009-0917826-016, and is subject to the Rights in Data – General Clause 52.227-14, Alt. IV (DEC2007).

Extracting FQs from analytic products

As IC products were published, researchers reviewed them for FQs to use in this study.

A fictional example, representative of the selected statements:

> We assess with moderate confidence that StatLandia will be more at risk of widespread internal violence in 2018. We cannot rule out that Bayesian elements might seek to confront the Frequentist militia. Such efforts by Bayesians could prompt a violent response from Frequentists, leading to widespread fighting.

A derived FQ posted to ICPM would be:

> Will StatLandia will have widespread internal violence in the year 2018?

Research Questions

1. Can prediction markets yield more accurate forecasts than IC analysis products?
   [The elephant in the room] [spoiler: yes!]

I will probably not have time today to properly address other research questions.

2. How does reading an IC product influence personal probabilities?
   - Do analysts update, Bayes-style, in the direction of the product?
     [spoiler: not necessarily, they might update in the opposite direction]
3. After reading an IC product, are updated probabilities more accurate?
   [spoiler: the answer may be surprising…]

Data Collection

41 IC analytic products → 99 forecasting questions (FQs)

5 analysts → probabilities for each FQ imputed to the product

- Imputed probability implied by the contents of the entire product
- Imputed probability based on the product plus events that occurred after publication
    - Analysts were instructed not to consider their personal beliefs when imputing.

ICPM

- FQs posted to ICPM
- Tens to hundreds of users from across the IC forecast on each FQ over time

A sample of the Data

DocID	FQID	AnID	Init	Imp	Imp+E	Final	ICPM	Truth
D1	Q1	A1	0.40	0.20	0.20	0.20	0.48	0.00
		A2	0.40	0.25	0.95	0.90	0.40	0.00
		A3	0.40	0.75	0.75	0.40	0.43	0.00
	Q2	A1	0.20	0.60	0.50	0.50	0.40	1.00
		A2	0.50	0.95	0.75	0.80	0.40	1.00
		A3	0.75	0.75	0.75	0.80	0.29	1.00
D2	Q3	A2	0.30	0.90	0.90	0.95	0.89	1.00
		A3	1.00	1.00	1.00	1.00	0.90	1.00
	Q4	A2	0.20	0.90	0.90	0.95	0.80	0.00
		A3	0.70	0.30	0.30	0.30	0.78	0.00

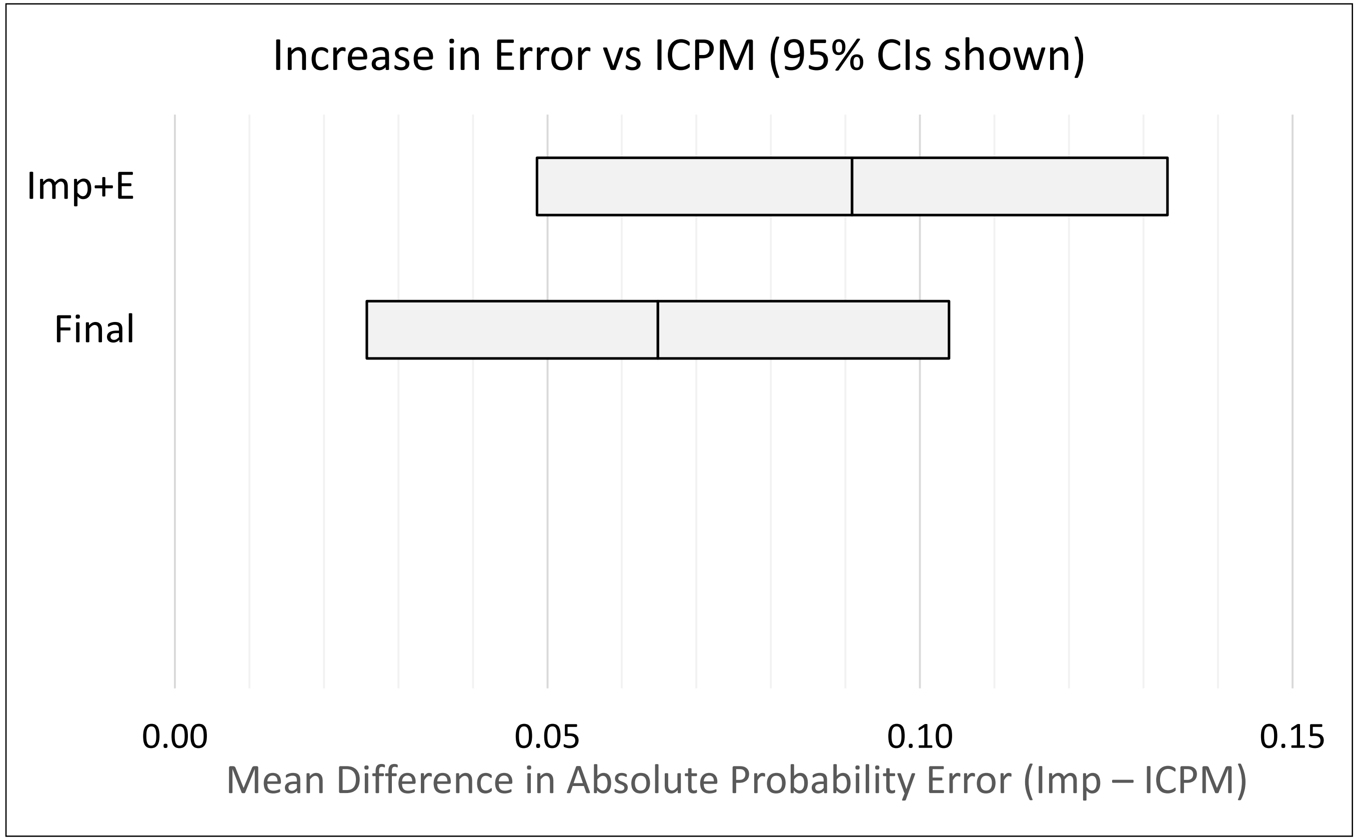
Color scale:

1.00	0.80	0.60	0.50	0.40	0.20	0.00

- Document ID (analytic product)
- Forecasting Question ID
- Analyst ID
- Initial personal probability
- Imputed product probability, based only on reading the product
- Imputed probability, based on product plus Events since publication
- Final updated personal probability
- ICPM probability (at the time of the analyst's imputation)
- Ground Truth (event outcome)

Results and Discussion

Accuracy comparison: ICPM vs Products



- ICPM was more accurate than probabilities imputed from IC products.
- ICPM was more accurate than probabilities provided by analysts.

	Mean Abs Error	Std.Dev.
Imp+E	0.39	0.23
Final	0.36	0.23
ICPM	0.30	0.21

Product Vagueness

Are forecasts from IC products clear or vague?

- If imputed probabilities **cluster tightly**, then the mean is a fair reflection of what is written in the product (irrespective of what the authors intended).
- If imputed probabilities **vary widely**, then that is evidence that the product did not make a meaningful forecast.

In 22 of the 83 questions answered by >1 analyst, the imputed probabilities differed by 0.5 or more.

➔ Clearly, the products left substantial room for substantially differing interpretations.

Qualitative language such as, "The probability is high that …", "It is likely that …", or "There is a fair chance that … ", are commonly used in IC forecasts, contrary to the preference of many consumers, who prefer numerical forecasts such as, "There is a 70% chance that…".

Did accuracy improve after reading products?

No statistical difference.

"Considerably more interesting than the [null] overall result, is the pattern of how analysts updated their personal probability judgments."

Table 3: Directional Accuracy of Updated Personal Probabilities partitioned by direction of update			
	Revised personal probability *more* accurate than initial	Revised personal probability *less* accurate than initial	Total
Personal probability revised in *same* direction as imputed probability	72	81	152
Personal probability revised in *opposite* direction of imputed probability	32	5	37
Total	104	87	

"Of particular note are the 37 forecasts where analysts updated their judgments by moving their personal probabilities in the opposite direction of the imputed probabilities."

Summary

Main results:

(1) ICPM forecasts were more accurate than analysis products.

(2) When analysts updated their probabilities opposite to what products implied, they were likely to update in the correct direction.

(3) 21% of product forecasts were so imprecise that analysts imputed probabilities that differed by more than 0.5.

Overall, these results suggest complementary benefits from traditional analysis and crowd wisdom approaches to forecasting.

Discussion

jonathan.mchenry@iarpa.gov 301-851-7730

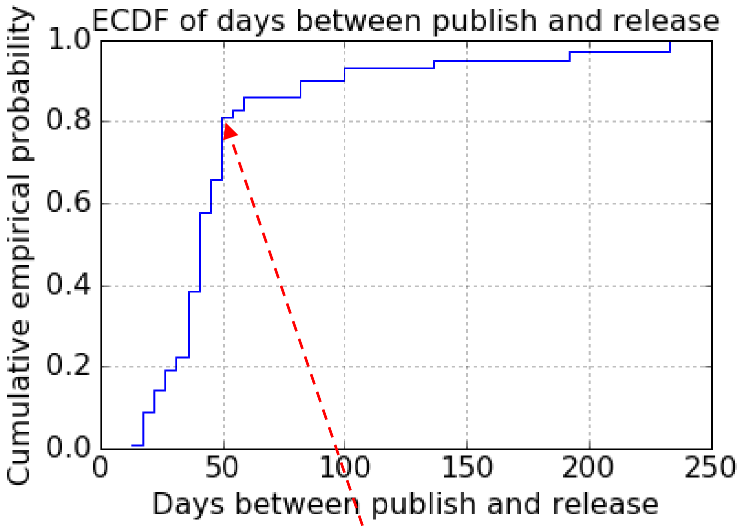
steve.rieber@iarpa.gov 301-851-7521

Stastny & Lehner (2017) has been accepted for publication by JDM, and is available on request.

Data will be available for download, after publication.

Backup

Lag distribution



- 80% of FQs were released < 50 days after product publication.
- Lags were always >0 due to the time required for the process of reviewing products and extracting forecasts.
- Longer lags are attributed to products selected earlier in the study.
- All products were considered to be the most current coverage of their subject matter.

Sample sizes

# FQs	# Analysts answering
17	4
26	3
40	2
16	1

Total FQs: 99

83 FQs ans. by >1 An.

Analyst ID	# FQs answered
Analyst 1	71
Analyst 2	69
Analyst 3	27
Analyst 4	48
Analyst 5	27

Total FQ answers: 242

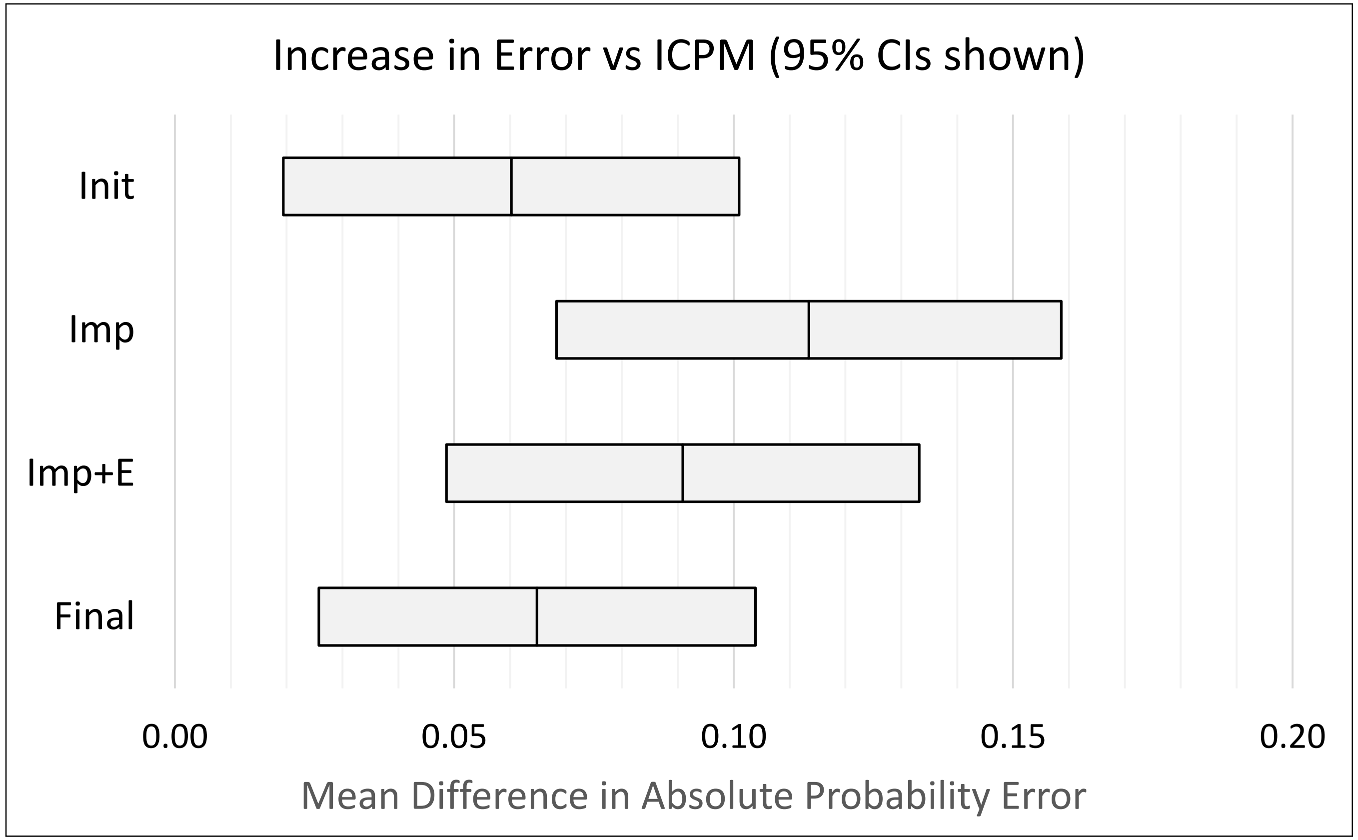
# Docs	# FQs in Doc
12	1
14	2
8	3
5	4
1	5
1	10

Total Docs: 41

FQ selection can be considered to be random.

- The five analysts answered the most recently released FQs, whenever they had time.
- They did not pick FQs based on their own subject matter expertise.

Accuracy comparison: ICPM vs Products



- ICPM was more accurate than the imputed probabilities.
- The average of Analysts' initial beliefs was more accurate than the average of their imputed forecasts.

	Mean Abs Error	Std.Dev.
Init	0.36	0.21
Imp	0.41	0.22
Imp+E	0.39	0.23
Final	0.36	0.23
ICPM	0.30	0.21

Does lag affect the comparative accuracy result?

- No.

Table 5: Comparison of ICPM and IC Product accuracy for different posting delays.

	Number of days until posted		
	10 to 35	36 to 50 days	More than 50 days
Number where ICPM more accurate	18	37	14
Number where IC product more accurate	4	16	10
Average difference in absolute error	0.170	0.117	0.017

- The ICPM advantage decreased with longer posting delays, so the accuracy advantage of ICPM can't be attributed to posting delay.

Calibration

Table 6: A Calibration Analysis of Imputed and ICPM Estimates.						
		Bin Midpoint				
		10%	30%	50%	70%	90%
Imputed Estimates	Number of questions contained in a bin.	19	29	22	24	5
	Percentage in bin that occurred.	21%	3%	9%	33%	60%
ICPM Estimates	Number of questions contained in a bin.	33	35	10	15	6
	Percentage in bin that occurred.	3%	11%	30%	27%	100%

- Both the product and ICPM forecasts exhibited poor calibration. Both exhibited overestimation of the likelihood of event occurrence.
    - for the most part, product writers, analysts, and ICPM participants overestimated the likelihood of event occurrences

Analyst beliefs affect imputed probabilities.

Table 1: Direction of initial personal probability relative to imputed		
Personal to Imputed	Same Direction	129
	Different Direction	82
	Sign Test	<.002
Personal to Imputed + Current	Same Direction	136
	Different Direction	81
	Sign Test	<.001

- Analyst interpretations are slightly biased toward individual beliefs, but they did a reasonable job of setting aside personal views.

Reading products changed analyst beliefs.

Table 2: Directional Changes in Updated Personal Probabilities		
		Shift in Personal probabilities
Change Relative to Imputed probabilities	In direction of Imputed?	152
	Away from Imputed?	37
	Sign Test	<.001
Change Relative to Imputed + Current	In direction of Imputed?	179
	Away from Imputed?	9
	Sign Test	<.001

- Analysts are taking what they learned in the products and using that information to update their personal beliefs.
- The influence that products have on analyst judgments is somewhat stronger than the influence their priors have on their interpretation of the products.

Imputed probabilities: personal bias

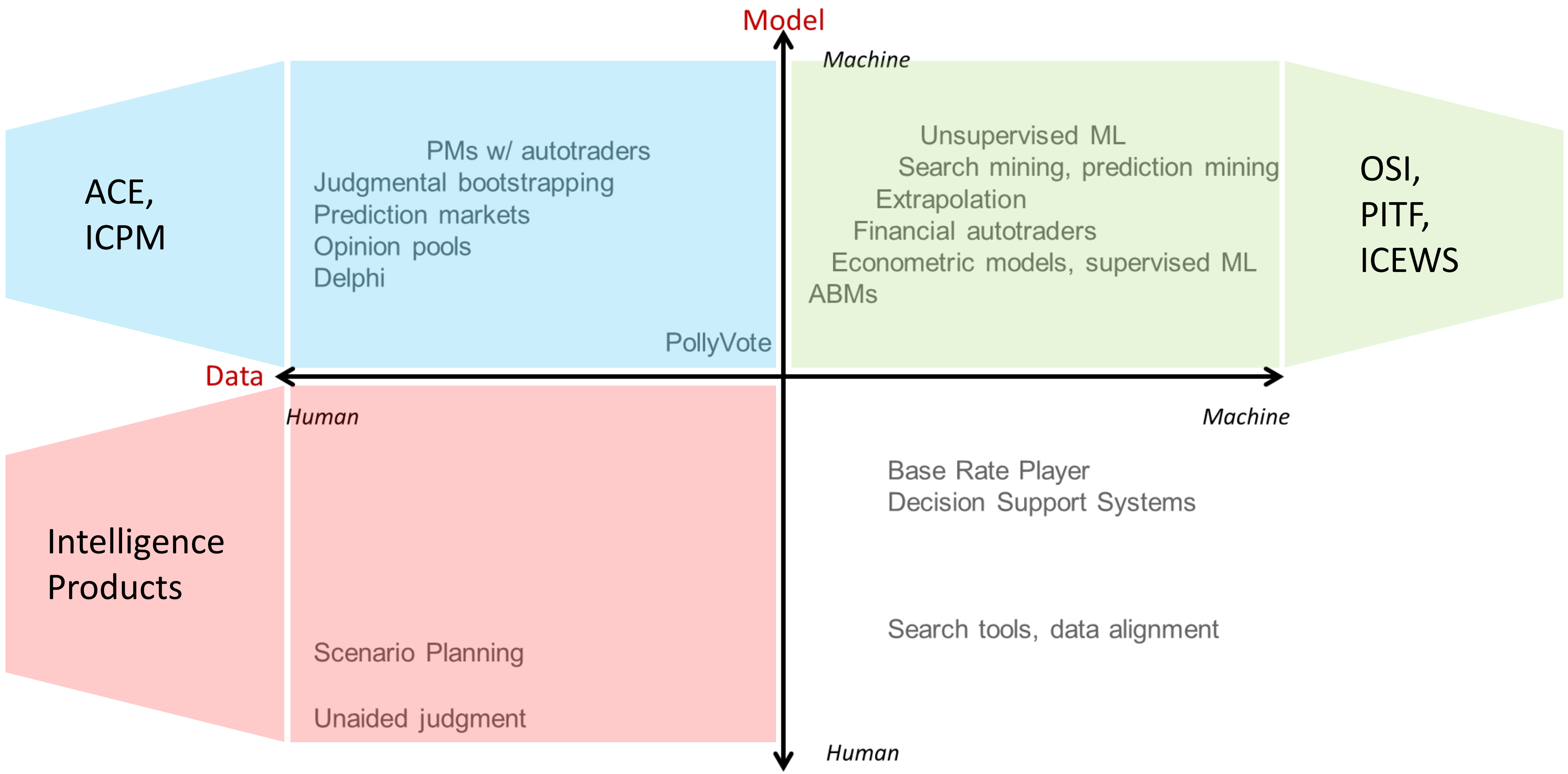
Analyst ID	# FQs answered
Analyst 1	71
Analyst 2	69
Analyst 3	27
Analyst 4	48
Analyst 5	27
Total FQ answers: 242	

- 83 FQs had 2 or more analysts answering.
- Did imputed probabilities agree?
  - 22/83 FQs had imputed probabilities differing by more than 0.5
    → these products left room for substantially differing interpretations!
- Were disagreements due to the bias of analyst priors?
  - 129/242 analyst personal probabilities were closer to that analyst's corresponding imputed probability than to the average imputed probability. 82/242 personals were closer to the average imputed than to their own imputed. 31/242 had personal or imputed equal to average imputed.
    → these results suggest analysts did a reasonable job of putting aside their personal views when making imputation judgments, but that they are not immune from this effect.
  - Similar results when imputing based on considering events since publication.

Other Complications

- 28 FQs were “fuzzy”. Fuzzy FQs did not have resolution language. All 28 fuzzy FQs were resolved (by ICPM admins).
- 103 of the extracted FQs had resolved
  - 4 FQs from one IC product were excluded "due to researcher error". One analyst's answer for one question was removed because "the analyst did not properly follow directions".
  - 96 binary FQs and 3 ternary FQs
- FQs and resolution language were reviewed by independent government assessors who had broad policy and analysis experience. Gov edits focused on the definitions of vague terms in the FQs and the res language.

The Forecasting Space



Humans

Strengths

- Adaptive
- Real-world knowledge

Weaknesses

- Cognitive Limitations
- Slow

Machines

Strengths

- Speed
- Consistency

Weaknesses

- Rigid
- Training Data Dependent

Hybrids

Humans vs. Machines

- Machines generally outperform humans when:
  - Well-structured training data are available
  - Large numbers of predictions are required
- Humans beat machines when:
  - Prediction tasks are noisy, complex, or diverse
  - Unclear reference classes or unique situations, when the “train of history hits a curve.”

Potential HFC Solutions

- Systems that integrate human and machine forecasts in novel ways.
- Approaches that enable humans to improve on machine forecasts (or vice versa).
- Machines that provide highly relevant content to human forecasters.
- Hybrid prediction markets.
- Machines that help humans work together in new ways.