Bias, Information, Noise: The Central Forces of Forecasting

Ville Satopää,
Assistant Professor of Technology and Operations Management (TOM)

Outline

1. Why Good Judgment Matters?
2. Bias, Information, and Noise (BIN)
3. BIN Analysis of Good Judgment Project Data
4. Conclusion

Judgmental Predictions

- Predictions made by people
- Why should we care about Good Judgment?
  - Previously unseen events (e.g., demand of new products)
  - Non-stationary environments
  - Most data analysis involves some level of judgment.

IARPA hosted four geopolitical forecasting tournaments from 2011 to 2015

5 University Research Groups

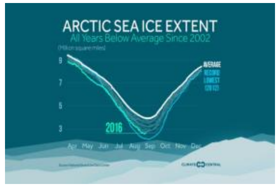
Over 500 Forecasting Questions

Over 20,000 Forecasters

Over 1,000, 000 Forecasts

Will China seize control of the 2nd Thomas Shoal before Jan 1, 2014?

Will the Colombian government and FARC begin talks before Jan 1, 2013?

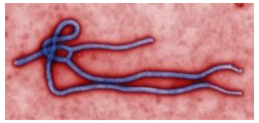


On Sept 15, 2014, will the Arctic sea ice extent be less than that it was on Sept 15, 2013?

Will any country announce its intention to withdraw from the Eurozone before Apr 1, 2013?

Will the number of Syrian refugees reported by the UNCHR exceed 250,000 before Apr 1, 2013?

Will the WHO report cases of Ebola in an EU state before June 1, 2015?



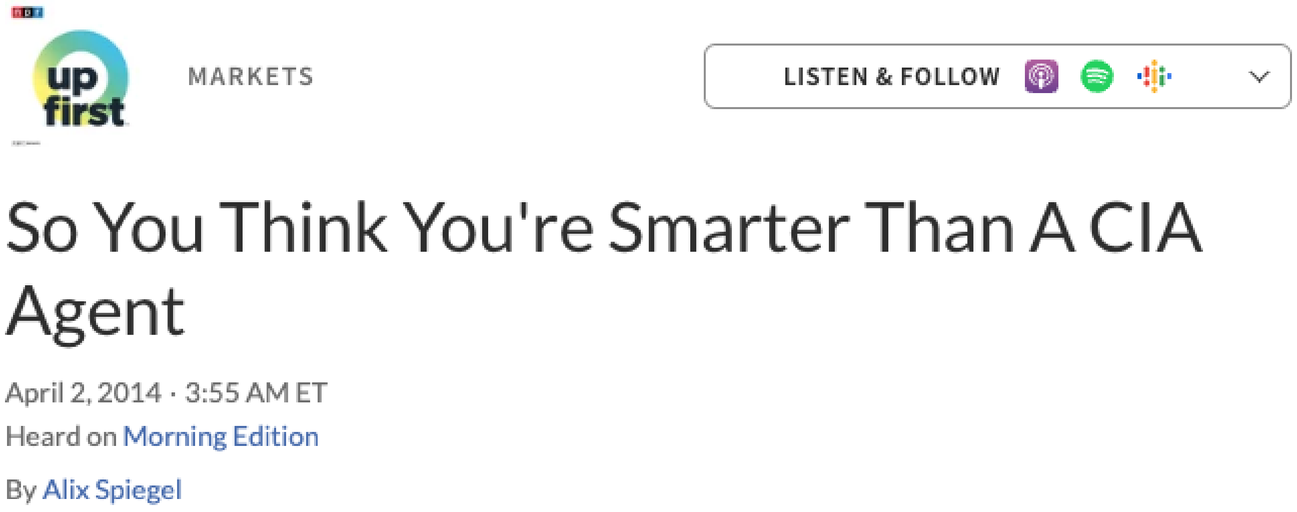
Accuracy Metric: Brier Score : 0 (Best) to 1 (Worst)

Imagine a “meteorologist who predicts the weather for 3 days.

Day	P(Rain)	Rain	Brier Score
1	0.9	1	(1-0.9)^2 = 0.01
2	0.4	0	(0-0.4)^2 = 0.32
3	0.9	0	(0-0.9)^2 = 0.81
Average	0.73	0.33	0.38

How accurate was the Good Judgment Project (GJP)?

1. GJP forecasters were 35-72% better than other teams
2. Superforecasters were 30% more accurate than Intelligence analysts in Prediction Markets who were forecasting the same questions using the same metric over the same period of time with access to classified information!



MARKETS

So You Think You're Smarter Than A CIA Agent

April 2, 2014 · 3:55 AM ET

Heard on Morning Edition

By Alix Spiegel

Outline

1. Why Good Judgment Matters?
2. Bias, Information, and Noise (BIN)
3. BIN Analysis of Good Judgment Project Data
4. Conclusion

Forecasting Accuracy

- Bias: systematic over/under estimation of probabilities
- Noise: uncorrelated variability with the outcome
- Information: correlated variability with the outcome

Example

- Repeatedly flip a **fair** coin twice
- Outcomes: TT, HT, TH, HH
- Each time predict the probability of seeing HH.
- Base rate = 0.25

Round	1	2	3	4	5
Flips	HT	TT	HH	HT	HH
Outcome	0	0	1	0	1

Example

Flip a fair coin twice: TT, HT, TH, HH

What is the probability of seeing HH?

Base rate = 0.25

1

No bias or noise:

No partial info.

- Predicts base rate, 0.25

Partial info.: the first flip.

- If T, then predicts 0; If H, then predicts 0.5.
- Mean is 0.25 = base rate
- Variance is 0.0625
- Covariance with outcome is 0.0625

Example

Flip a coin twice: TT, HT, TH, HH

What is the probability of seeing HH?

Base rate = 0.25

2

No noise:

Incorrectly believes that Prob of H is 0.6.

No partial info.

- Predicts 0.6 x 0.6 = 0.36 > base rate

Partial info. Sees the first flip.

- If T, then predicts 0; If H, then predicts 0.6.
- Mean is 0.3 > base rate

Example

Flip a coin twice: TT, HT, TH, HH

What is the probability of seeing HH?

Base rate = 0.25

3

No bias:

Sees unrelated coin flip.

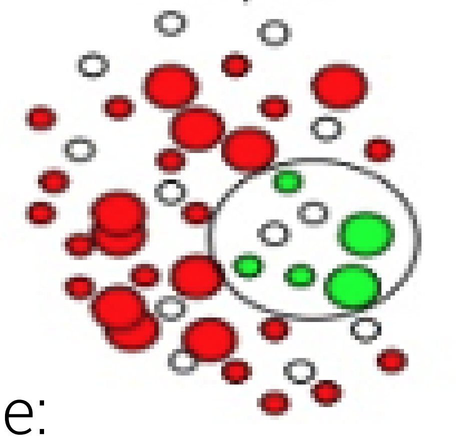
- If T, predicts 0; If H, predicts 0.5
- Mean is 0.25 = base rate
- Uncorrelated with the outcome

Sees one actual and one unrelated flip

- If HH, predicts 1; else, predicts 0
- Mean is 0.25 = base rate
- Variance is 0.1875
- Covariance with outcome is 0.0625

Model

Signal Universe



Following Satopää et al. (2016), we posit a signal universe:

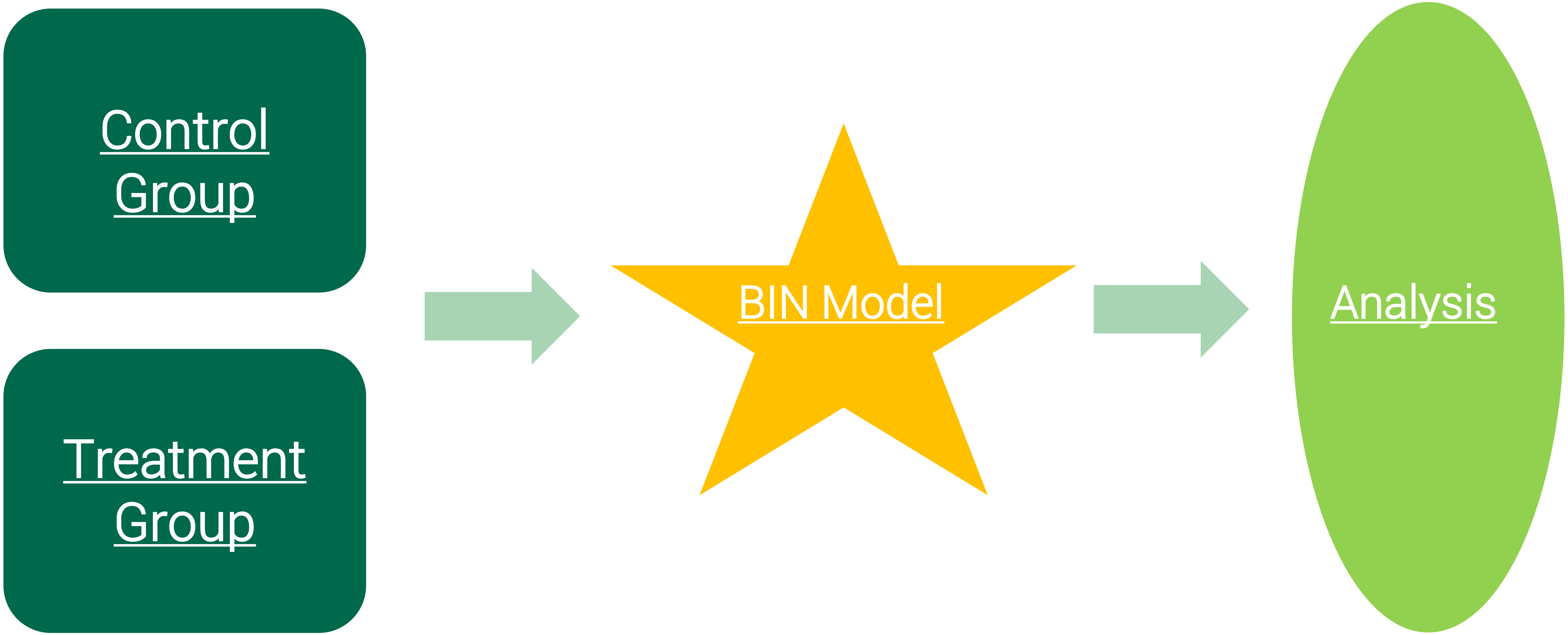
- Relevant and irrelevant signals
- Sum of relevant signals > 0 → The event happens; else it does not.

Forecasters sample signals from the universe:

- Irrelevant signals create noise
- Relevant signals increase partial information
- May center incorrectly, leading to bias.

Model

Workflow and Inputs



Model

Output

We estimate the model with Bayesian statistics.
Final output gives:

1. Posterior means of Bias, Information, and Noise;
2. 95% credible intervals for each component;
3. Probabilities that the treatment group outperforms control group;
   - E.g., ``Treatment group has less noise than the control group with probability 0.98.”
4. How much treatment improves accuracy through changes in Bias, Noise, and Information.

Outline

1. Why Good Judgment Matters?
2. Bias, Information, and Noise (BIN)
3. BIN Analysis of Good Judgment Project Data
4. Conclusion

Good Judgment Project Data

Description

- In 2011-2015 IARPA sponsored geopolitical forecasting tournament

> Would Serbia be granted EU candidacy by 31 December 2011?
> Forecasting began on September 1, 2011.
> Resolved as ``no”
> Question was open 4 months

- Good Judgment Project was the team from UPenn
- Full dataset contains hundreds of questions and thousands of forecasters

Good Judgment Project Data

Three Treatments

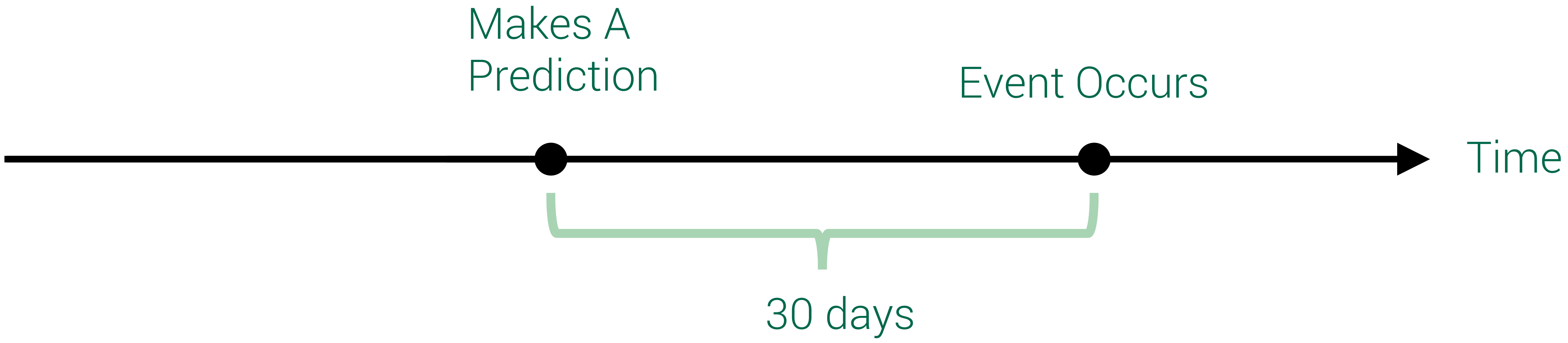
Probability Training: Forecasters completed a tutorial on probabilistic reasoning:

reference classes;
judgmental biases such as over-conf. or confirmation bias;
average multiple predictions from different sources.

Teaming: Forecasters worked in teams of 10-15.

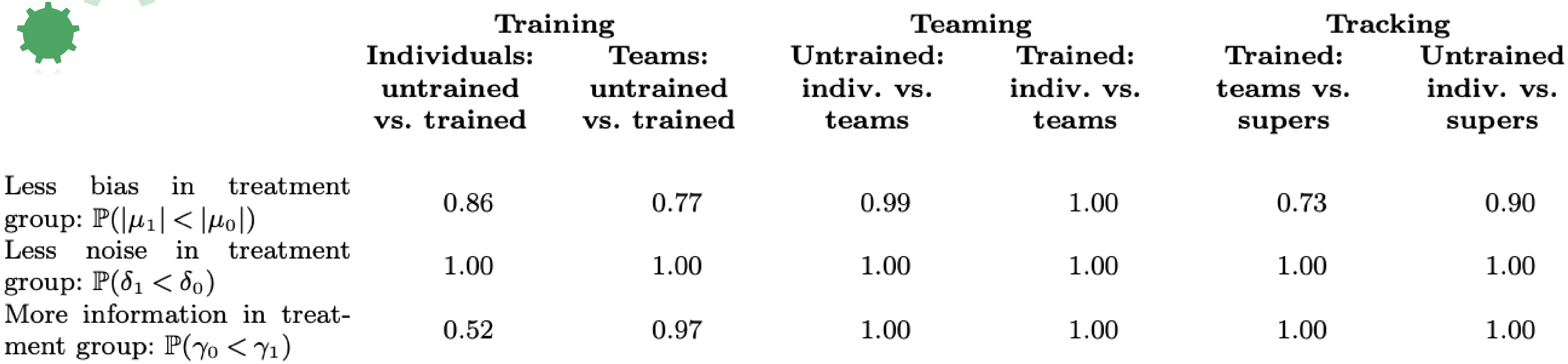
Tracking: At end of each year, the top 2% forecasters were designated as ``superforecasters” and allowed to work together next year.

Results: 30-Day Forecast Horizon



Good Judgment Project Data

Posterior Inference

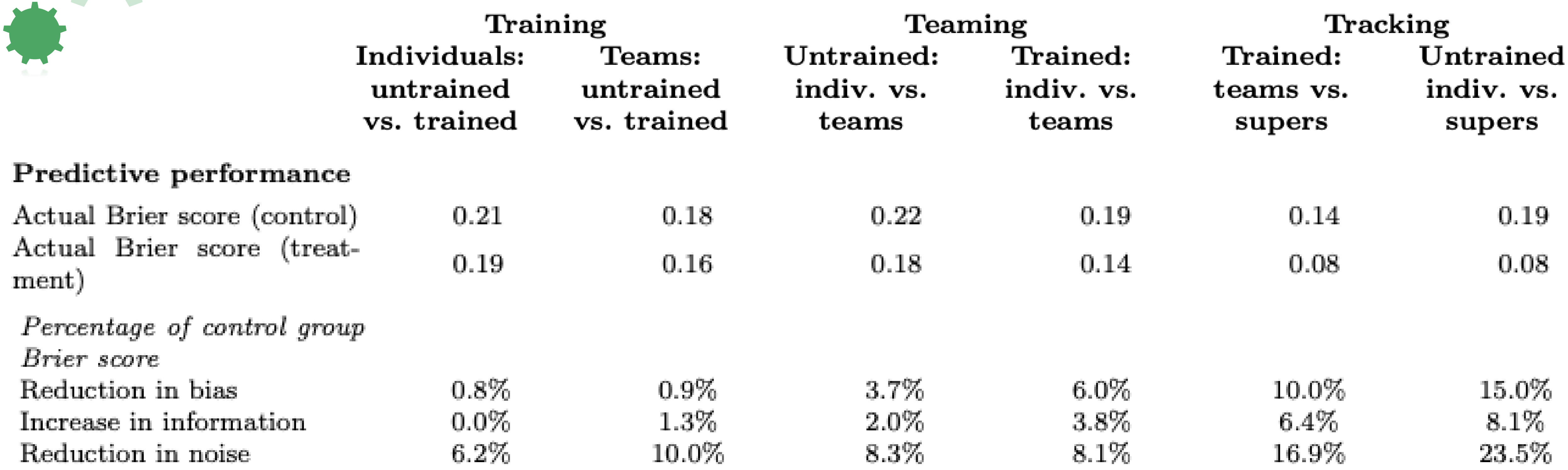


	Training		Teaming		Tracking	
	Individuals: untrained vs. trained	**Teams: untrained vs. trained**	**Untrained: indiv. vs. teams**	**Trained: indiv. vs. teams**	**Trained: teams vs. supers**	**Untrained indiv. vs. supers**
Less bias in treatment group: $\mathbb{P}(\lvert\mu_1\rvert < \lvert\mu_0\rvert)$	0.86	0.77	0.99	1.00	0.73	0.90
Less noise in treatment group: $\mathbb{P}(\delta_1 < \delta_0)$	1.00	1.00	1.00	1.00	1.00	1.00
More information in treatment group: $\mathbb{P}(\gamma_0 < \gamma_1)$	0.52	0.97	1.00	1.00	1.00	1.00

• Increased Info.	By all treatments, except prob. training on individuals
• Noise reduction	By all treatments
• Bias reduction	Only teaming

Good Judgment Project Data

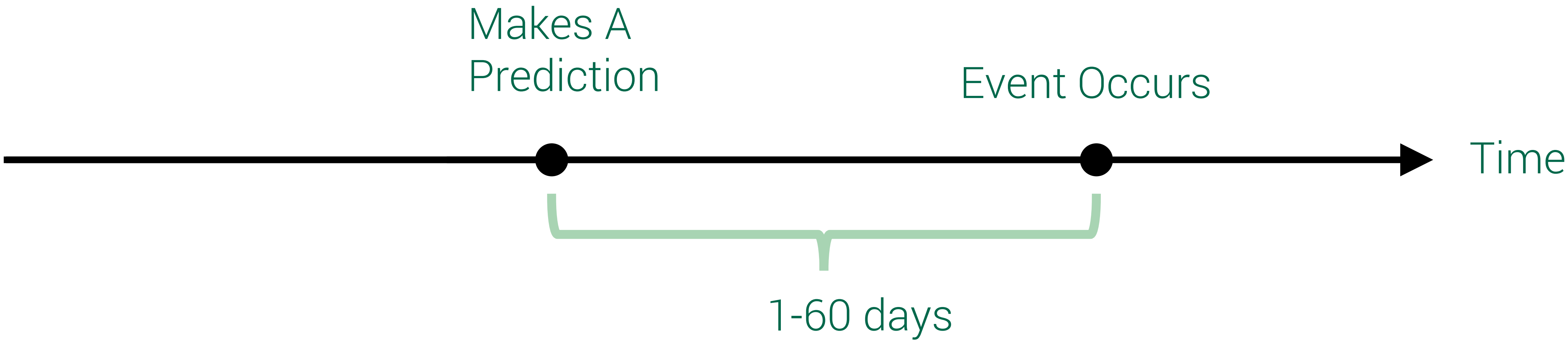
Predictive Performance



	Training		Teaming		Tracking	
	Individuals: untrained vs. trained	Teams: untrained vs. trained	Untrained: indiv. vs. teams	Trained: indiv. vs. teams	Trained: teams vs. supers	Untrained indiv. vs. supers
Predictive performance						
Actual Brier score (control)	0.21	0.18	0.22	0.19	0.14	0.19
Actual Brier score (treatment)	0.19	0.16	0.18	0.14	0.08	0.08
Percentage of control group Brier score						
Reduction in bias	0.8%	0.9%	3.7%	6.0%	10.0%	15.0%
Increase in information	0.0%	1.3%	2.0%	3.8%	6.4%	8.1%
Reduction in noise	6.2%	10.0%	8.3%	8.1%	16.9%	23.5%

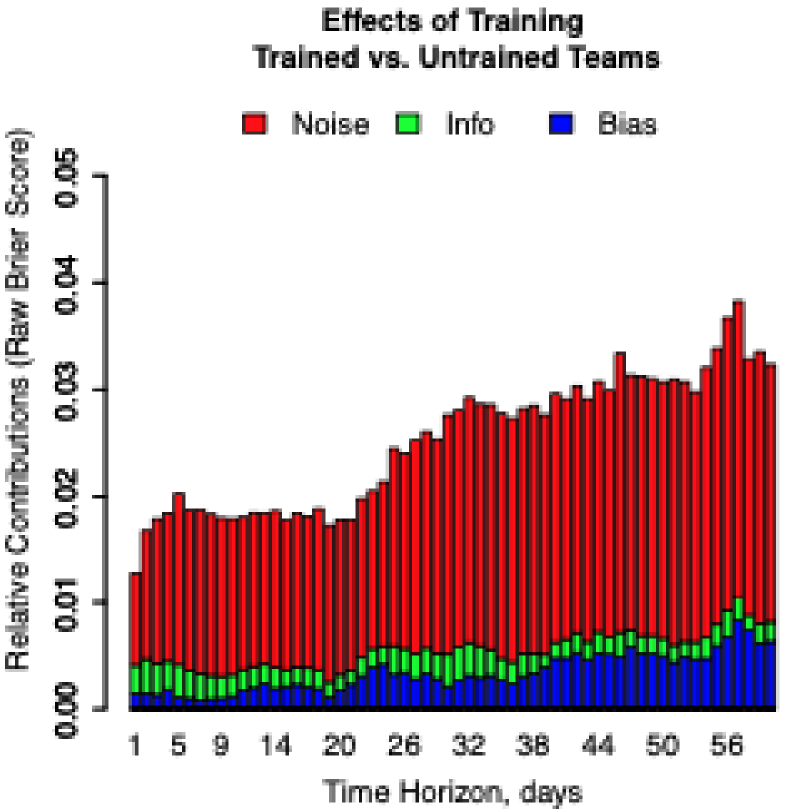
- Noise reduction is the most important!
- Training almost entirely noise reduction
- Teaming and tracking affect all three components.

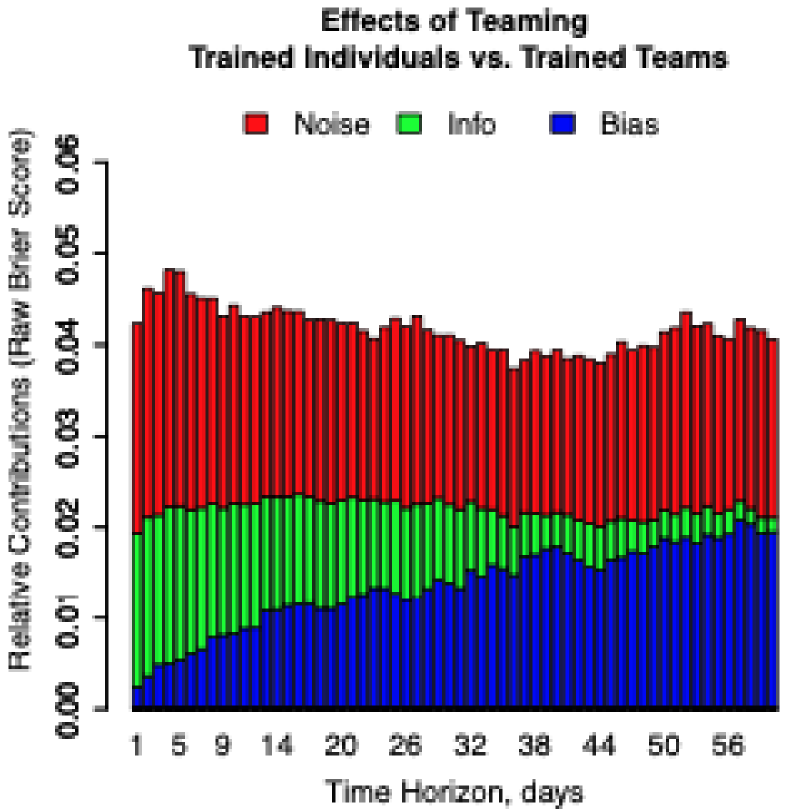
Results: Other Horizons

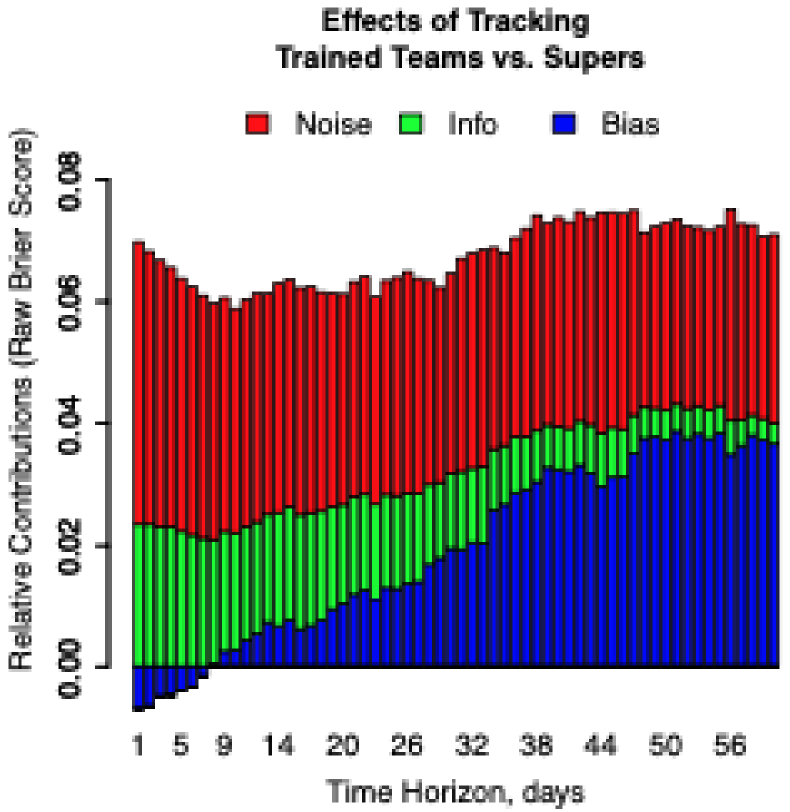


Good Judgment Project Data

Predictive Performance







- **Noise reduction consistently important.**
- Bias more important early on. Information important later on.
- Superforecasters have a ``bias blip” at the end.

Good Judgment Project Data

Key Findings

- Noise reduction emerged as the most consistent way to boost accuracy.
  - Not the original intent of the treatments.
  - In hindsight makes sense.
- Two observations about top performers:
  1. Discipline vs. creativity
  2. Tournaments may over-incentivize

Outline

1. Why Good Judgment Matters?
2. Bias, Information, and Noise (BIN)
3. BIN Analysis of Good Judgment Project Data
4. Conclusion

Conclusion

BIN Model offers a granular view into forecasting performance by

1. decomposing accuracy into bias, information, and noise, and
2. analyzing two groups (treatment and control) jointly and detecting significant differences.

The analysis of the GJP data revealed noise reduction as a key driver of accuracy.
How to reduce noise?

1. Discipline the internal judgment processes through noise audits (Kahneman et al. 2016) or other training exercises (Chang et al. 2016);
2. Aggregate judgments through prediction markets (Wolfers and Zitzewitz 2004, Atanasov et al. 2017) or statistical means (Larrick and Soll 2006, Budescu and Chen 2014, Satopää et al. 2014, Prelec et al. 2017);
3. Filtering out misleading or low-diagnosticity sources in the news environment and lightening the cognitive load on forecasters (Lazer et al. 2018);
4. Replacing human judges with machine-learning algorithms.

References

Papers:

Satopää, V. A., Salikhov, M., Tetlock, P., and Mellers, B.
"Bias, Information, Noise: The BIN Model of Forecasting”
Management Science, 2021

Satopää, V. A., Salikhov, M., Tetlock, P., and Mellers, B.
"Decomposing the Effects of Crowd-Wisdom Aggregators: The Bias-Information-Noise (BIN) Model.”
International Journal of Forecasting, 2022

Popular Press:

Want Better Forecasting? Silence the Noise. Podcast Interview with Knowledge@Wharton.
The Secret Ingredients of ‘Superforecasting’. INSEAD Knowledge.

Software:

R-package called BINtools available on CRAN

Ville Satopää
Assistant Professor of Technology and Operations Management
INSEAD
ville.satopaa@insead.edu