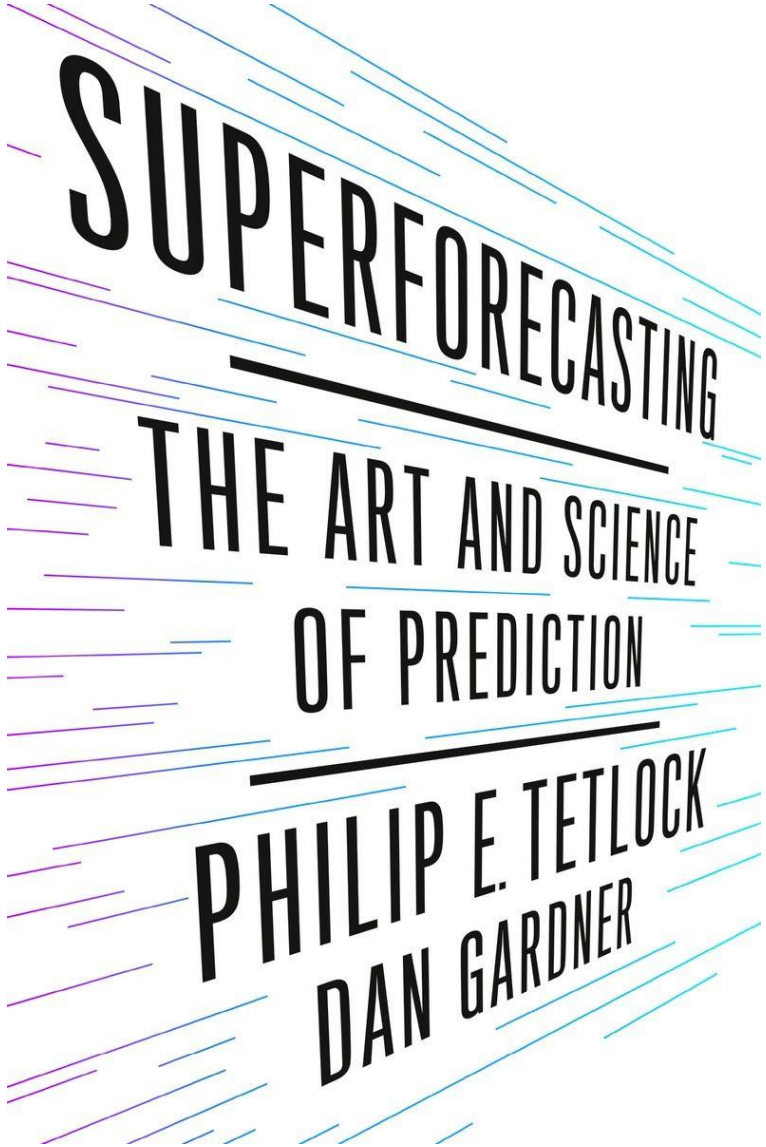
How Much Better Can We Get When We Get Serious About Keeping Score?

Philip E. Tetlock
Wharton & School of Arts & Sciences and
Good Judgment, Inc

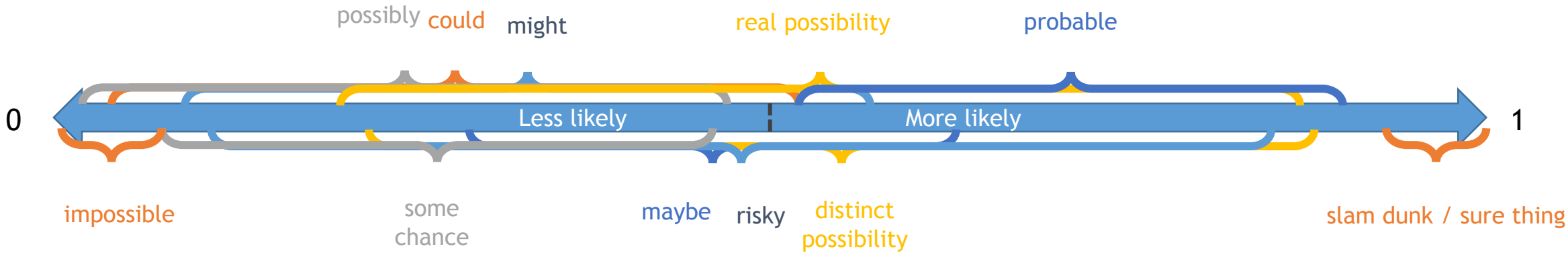
June 9, 2016



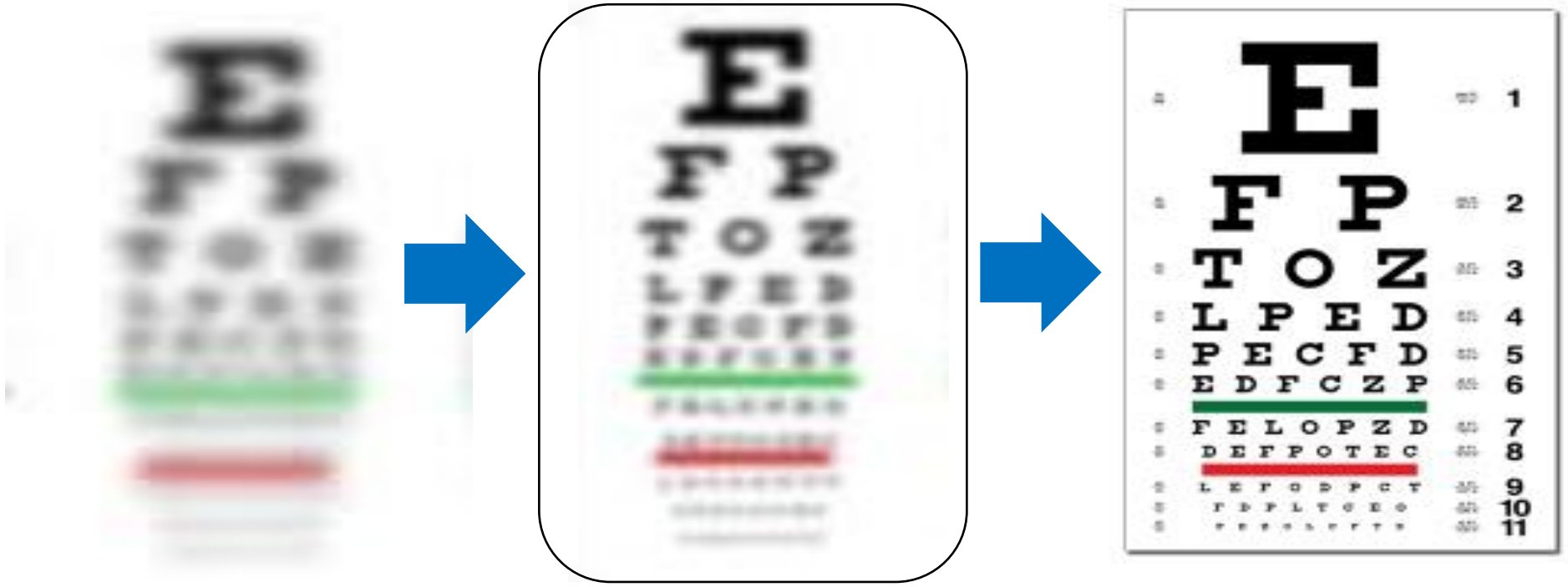
The problem of imprecision

- it ***might*** happen (0.09 to .64)
- it ***could*** happen (0.02 to .56)
- it's a ***possibility*** (0.001 to .45)
- it's a ***real possibility*** (0.22 to 0.89)
- it's ***probable*** (0.55 to 0.90)

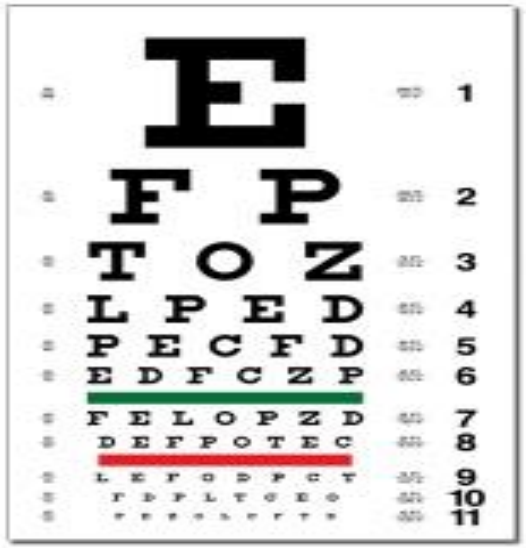
- ***maybe*** (0.31 to 0.69)
- ***distinct possibility*** (0.21 to 0.84)
- ***risky*** (0.11 to 0.83)
- ***some chance*** (0.05 to 0.42)
- ***slamdunk or sure thing*** (0.95 to 1.00)



Question 1: How much can we improve probabilistic foresight?



Superforecaster teams saw things 400 days out …



… as well as regular forecasters saw 150 days out

Question 2: How much should we value improved foresight?

Value in Jan 2003 of knowing probability of Iraqi WMD was not 1.0 (slamdunk) but between 0.6 and 0.8?

Sometimes subtler: In Dec 2014, that p(OPEC quota cut) was 0.1, not 0.5?

In July 2015 that p(Grexit) was 0.2, not 0.4?

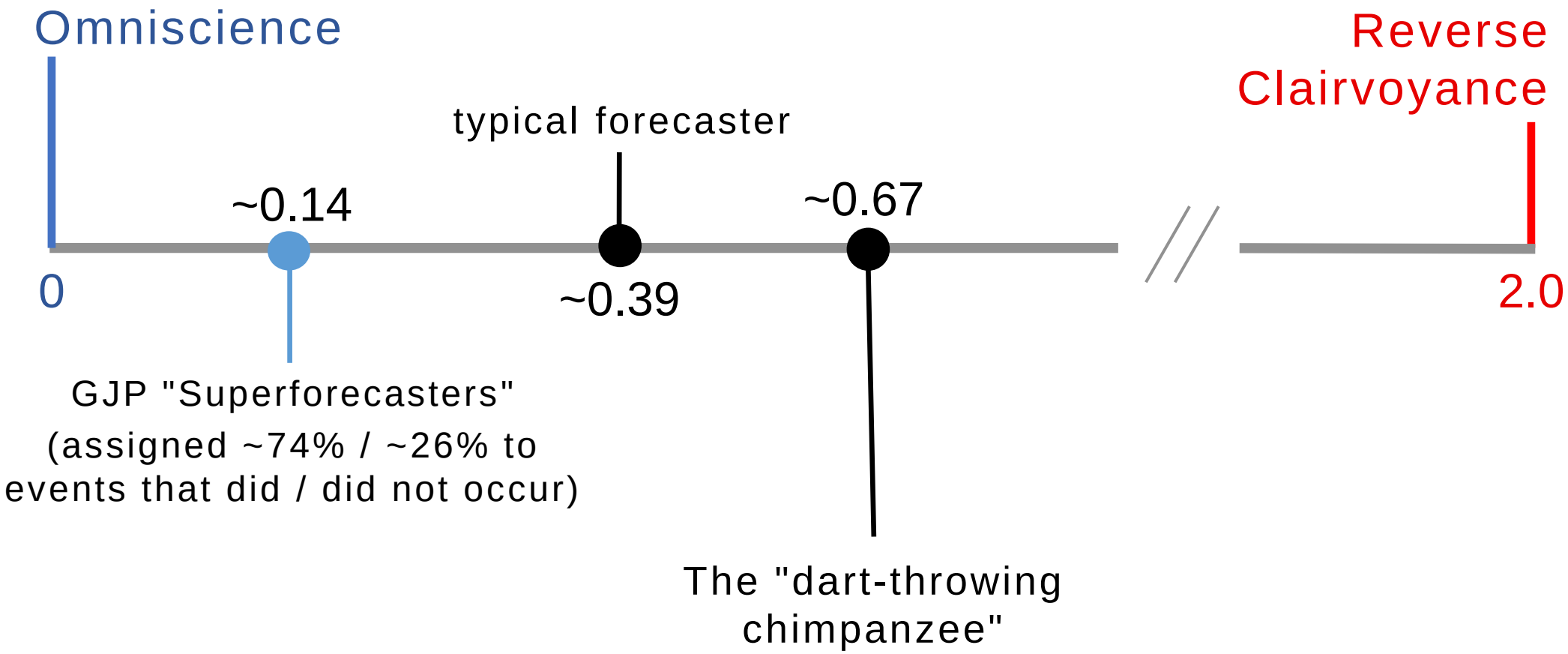
Key conclusion from IARPA Tournaments: Possible to get better

- Thousands of forecasters, hundreds of outcomes, >1mn forecasts
- GJP won by big margins against select scientific competitors
- But what exactly does “winning” mean?

Brier Scoring: Measures the accuracy of probability judgments

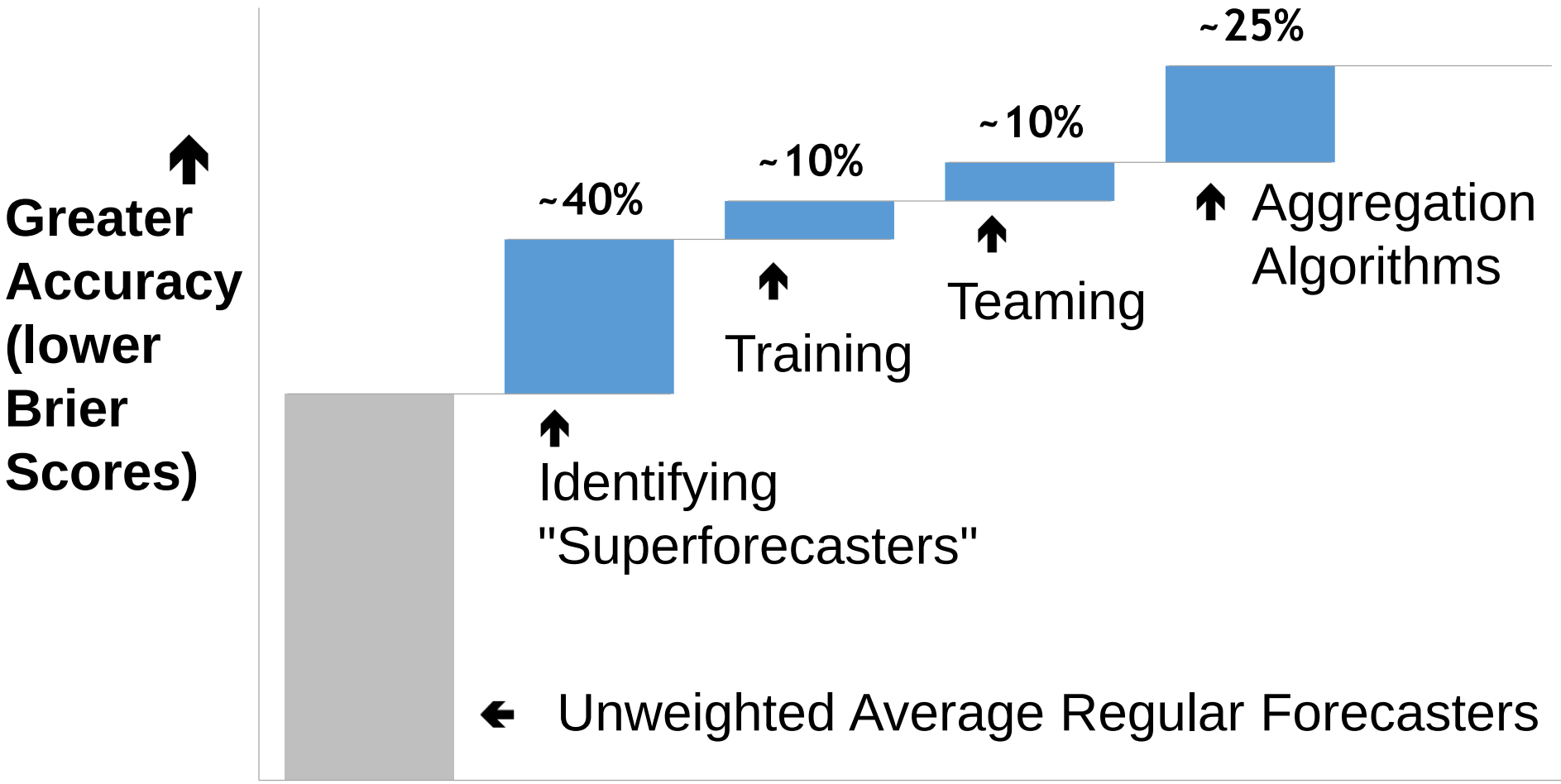
Day	Probability of Positive Earnings Surprise	Outcome	Brier Scores
1	90%	Yes, surprise	**$(1-.9)^2 + (0-.1)^2 = 0.02$**
2	90%	No surprise	**$(0-.9)^2 + (1-.1)^2 = 1.62$**
3	50%	Yes, surprise	**$(1-.5)^2 + (0-.5)^2 = 0.50$**
			Mean = 0.71

Where Do you Fall Along Foresight Continuum?



How GJP overcame human biases and won the tournaments

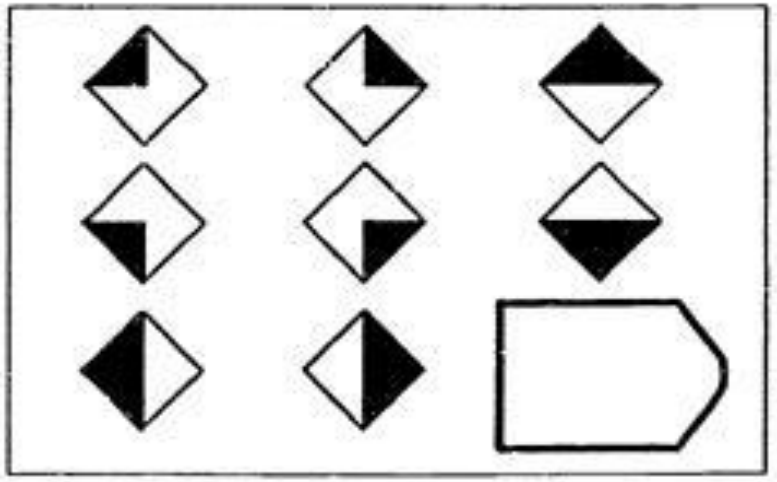
Four-Factor Winning Formula

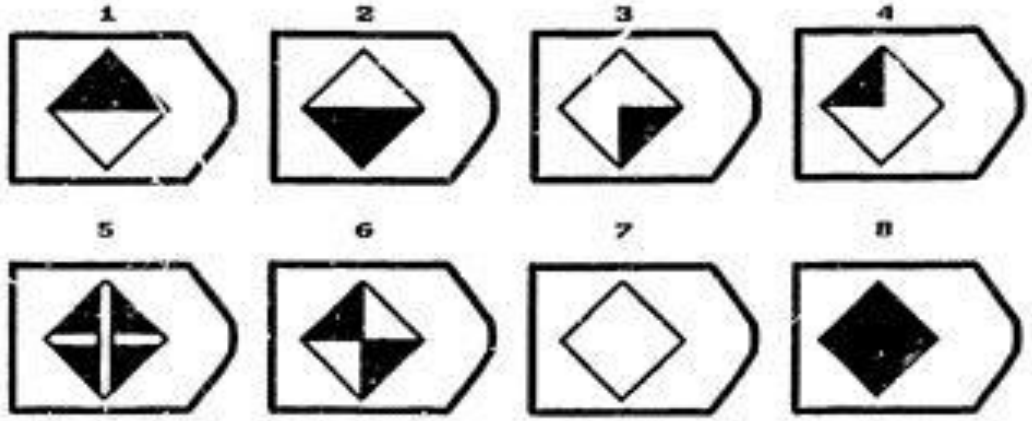


Factor 1: Identifying "Superforecasters"

- FLUID INTELLIGENCE
- NUMERATE—AND MORE SHADES OF MAYBE
- COGNITIVE "GROWTH MINDSET"

Fluid Intelligence (Raw Rapid Problem-Solving Ability)





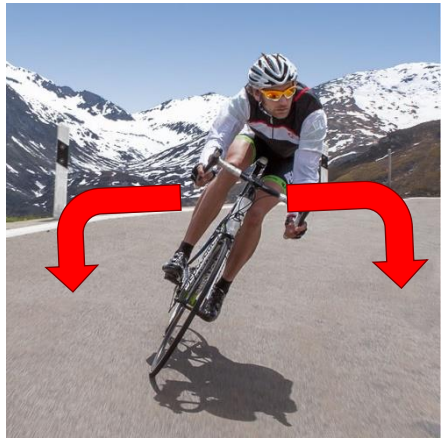
Factor 2: Training

Focus on error-balancing

Over-confidence

Over-adjusting to New Evidence

Over-Use Base Rates



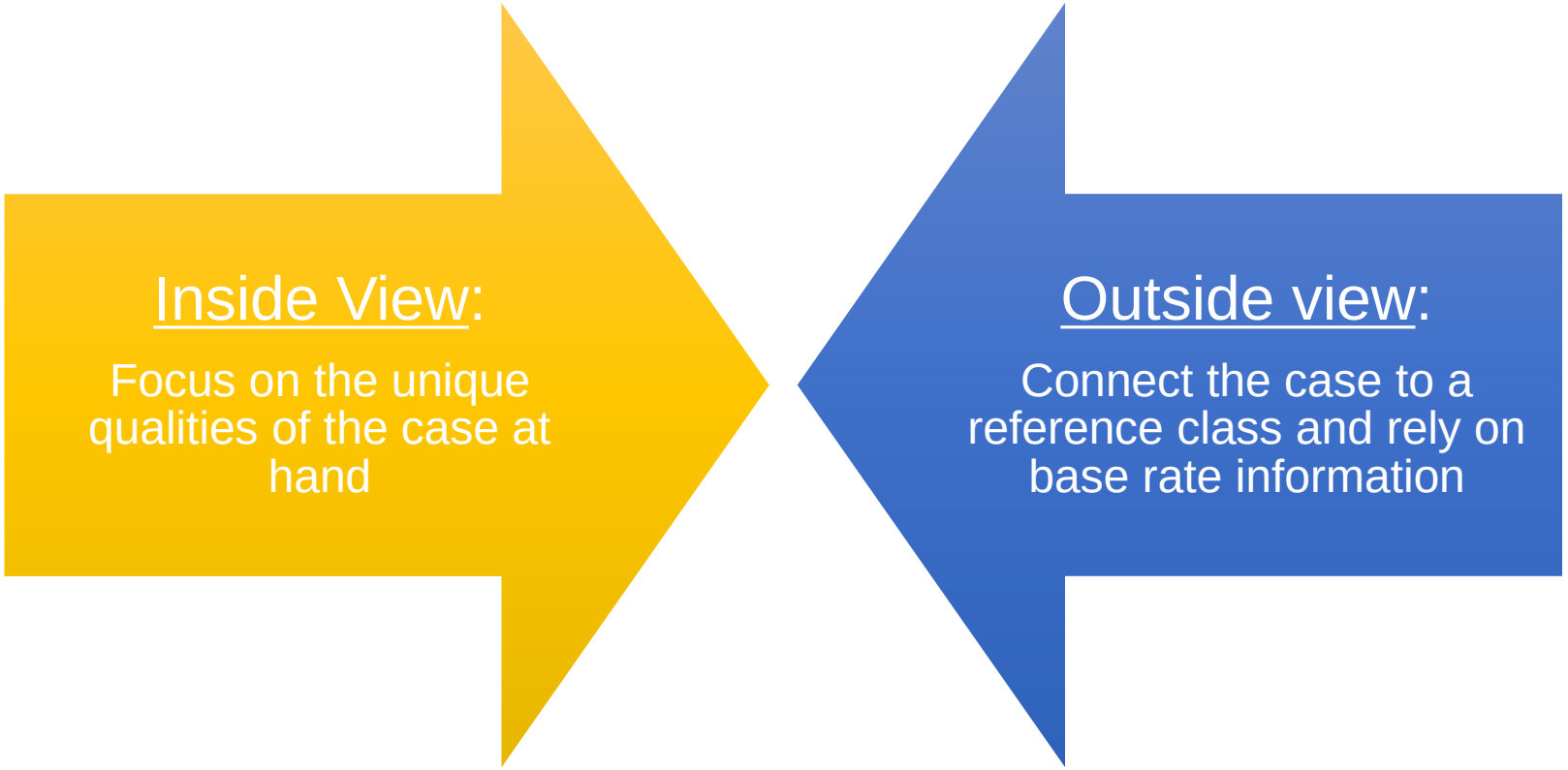
Under-confidence

Under-adjusting…

Under-Use Base Rates

Factor 2: Training on Balancing the Inside & Outside Views

Nobelist Daniel Kahneman warns: more common error is under-using base rates



Factor 2: Training on the Inside & Outside Views

Inside View

- Some people ask themselves: Does it look like a good fit (synergies, cultures)?
- Then they translate that feeling into a probability estimate

Outside View

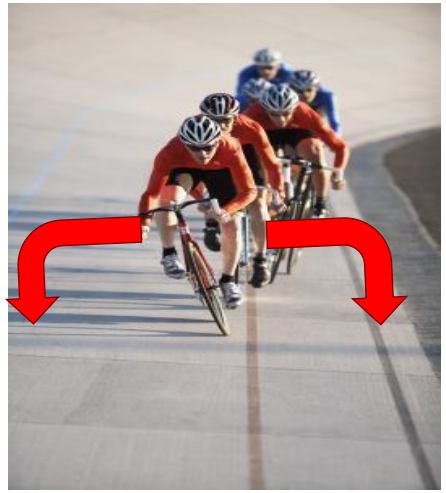
- Smart forecasters ask: How often do mergers fail to meet expectations?
- If 55% fail, they start from that estimate and adjust appropriately for the case at hand

Factor 3: Teaming (pursuing process gains)

Viewpoint Diversity

Dynamic Debate

Division of Labor



Avoid analysis-paralysis

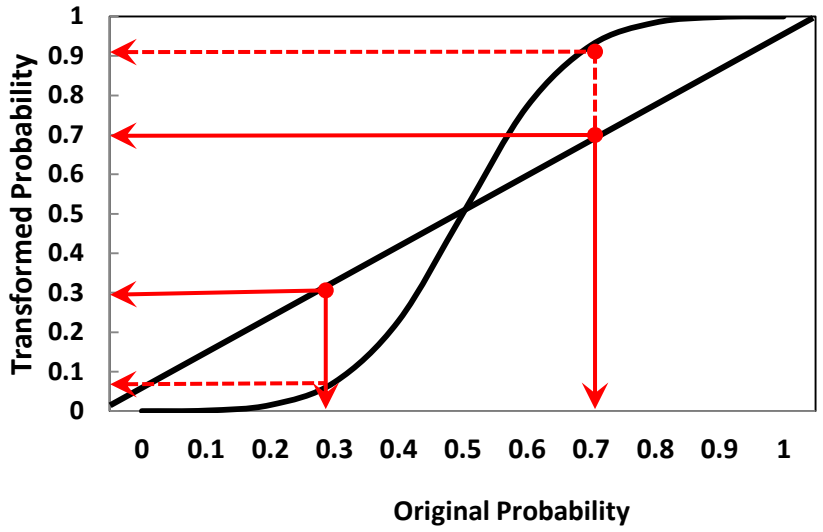
Disagree Without Being Disagreeable

Optimal Redundancy

Factor 4: Advanced wisdom-of-crowd algorithms

- Weighted average of best forecasters & log-odds transform
  - $m_j = a\ log(p_j/(1-p_j)) + e$

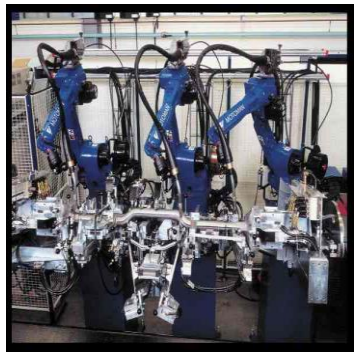
- Extremizing parameter, *a*, rises with diversity of forecaster pool



Next generation of tournaments focuses as much on quality of questions as on accuracy of answers

AlphaGo vs Lee Sedol in 2016

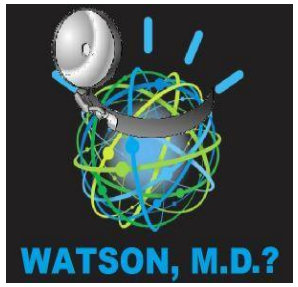
Robotics Industry Exceeds US$155 billion in 2020



Driverless Uber Cars in Las Vegas in 2017

4TH INDUSTRIAL REVOLUTION DRIVEN BY A.I. ?

Watson MD vs Best Medical Diagnostician in 2017



Half of Accounting Jobs are Automated

INVITATION: TEST YOURSELVES

Nominate pivotal questions with biggest ROI on improved foresight.

Forecast in your private in-house tournament—and compete against superforecasters by posing same questions to hatch@goodjudgment.com

Extra slides

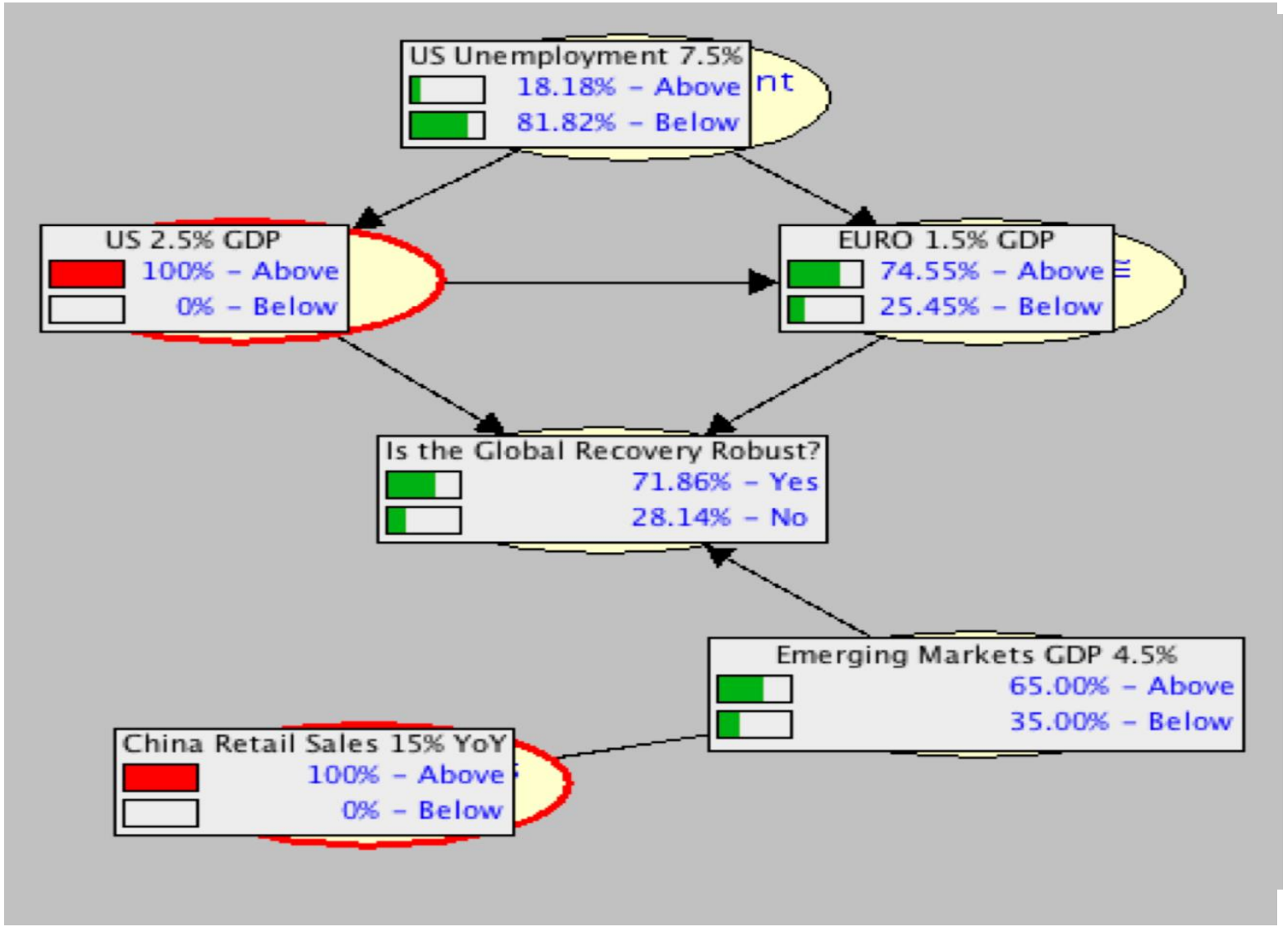
How Did GJP Pull it Off?

Get Right People on Bus	• Spotting/cultivating talent: superforecasters
Teaming	• Anti-groupthink groups: superteams
Training	• Debiasing exercises
Elitist Algorithms	• Aggregation algorithms that up-weight shrewd forecasters AND extremize to compensate for conservatism of aggregates

BAYESIAN SOFTWARE TOOLS FOR HELPING YOU THINK COHERENTLY ABOUT COMPLEXITY

- Evolving belief net…

1. Before We Observe Any Outcomes
2. After Observing
   - High US Unemployment
   - Low Eurozone Growth
3. After Observing
   - High US Growth
   - High Chinese Retail Growth



But You Don’t Create a Superforecasting Culture by Checking Off Checklists

- Requires changing deeply ingrained habits, changing how we think about thinking
- The parable of Zero Dark Thirty

HOLLYWOOD KNOWS US: DEEP DOWN DON'T WE FEEL…?

- Give me consensus: Disagreement is annoying
- Give me yes or no: Probabilities are for wimps
- Give me good outcomes—to hell with process: Silly to say you could be "right but unreasonable" or "wrong but reasonable"