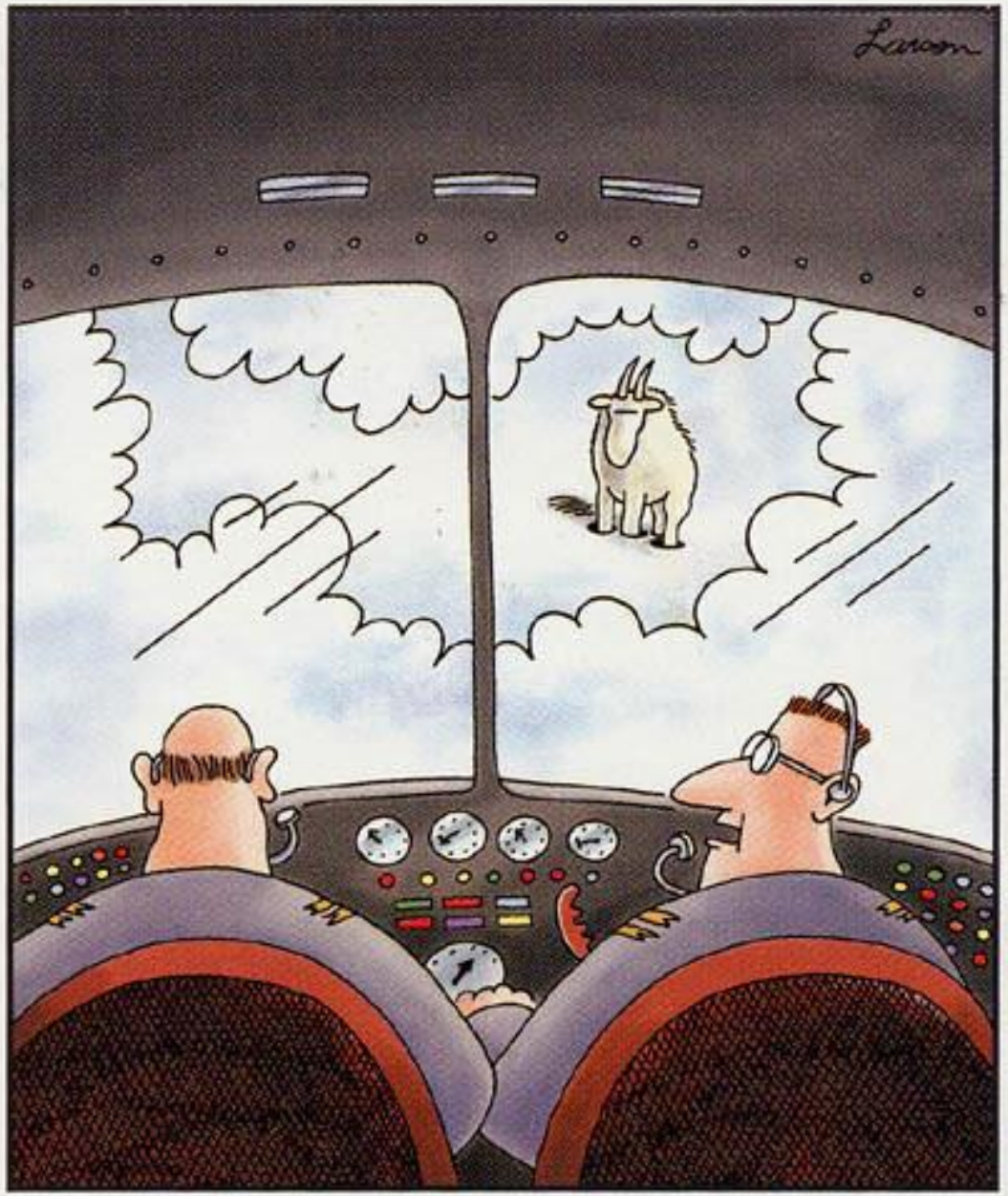


"Say … what's a mountain goat doing
way up here in a cloud bank?"

How to Win a Forecasting Tournament?

Philip E. Tetlock
Wharton School

CFA Asset Management Forum
Montreal, October 8, 2015

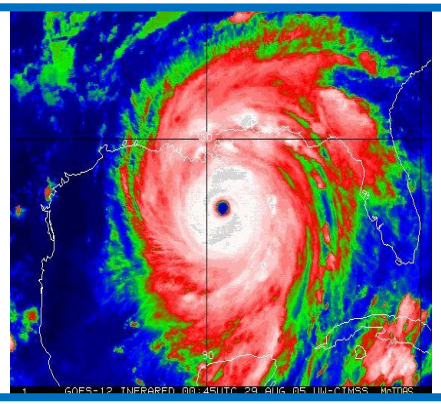
WHAT ARE FORECASTING TOURNAMENTS?

- *level-playing-field competitions to determine who knows what*
- *a disruptive technology that destabilizes stale status hierarchies*

How Did GJP Win the Tournament?

- By assigning the most accurate probability estimates to over 500 outcomes of “national security relevance”
- But how did GJP do that?

Winning requires picking battles wisely:



Where the pendulum swings

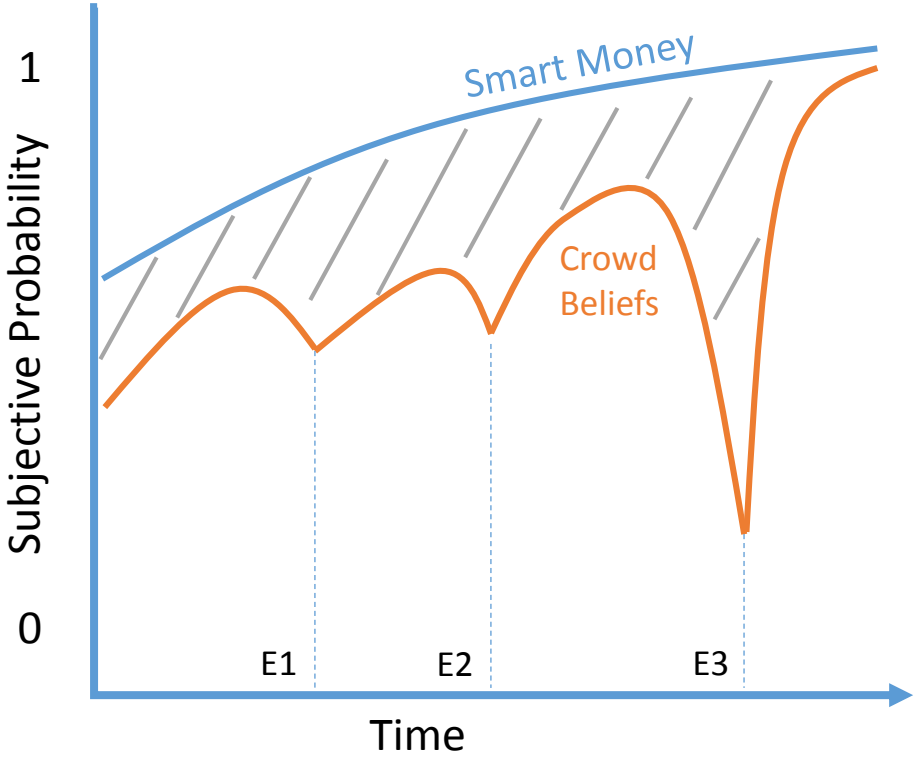
Where the hurricane meanders

Where the ball stops

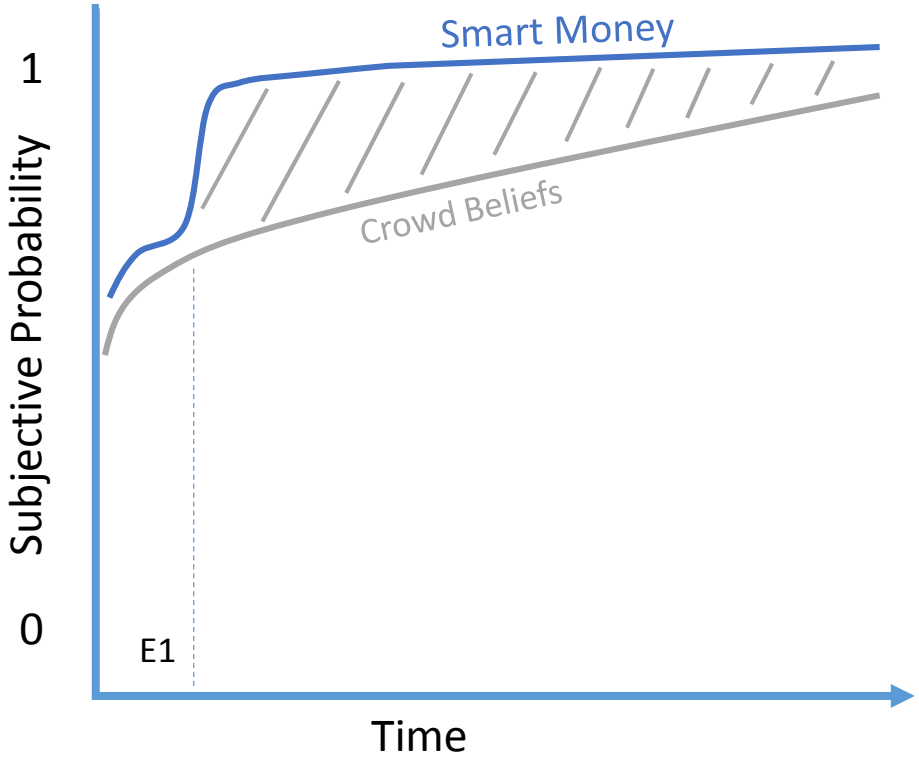
More Predictable ⟷ **Less Predictable**

Winning Requires Skill at:

Discounting Pseudo-Diagnostic News to Which Crowd Over-Reacts



Spotting Subtly-Diagnostic News to Which: Crowd Under-Reacts



And winning requires moving beyond blame-game ping-pong

WMD: Over-Connecting the Dots

False-Positives

Finding Osama Bin Laden

9/11: Under-Connecting the Dots

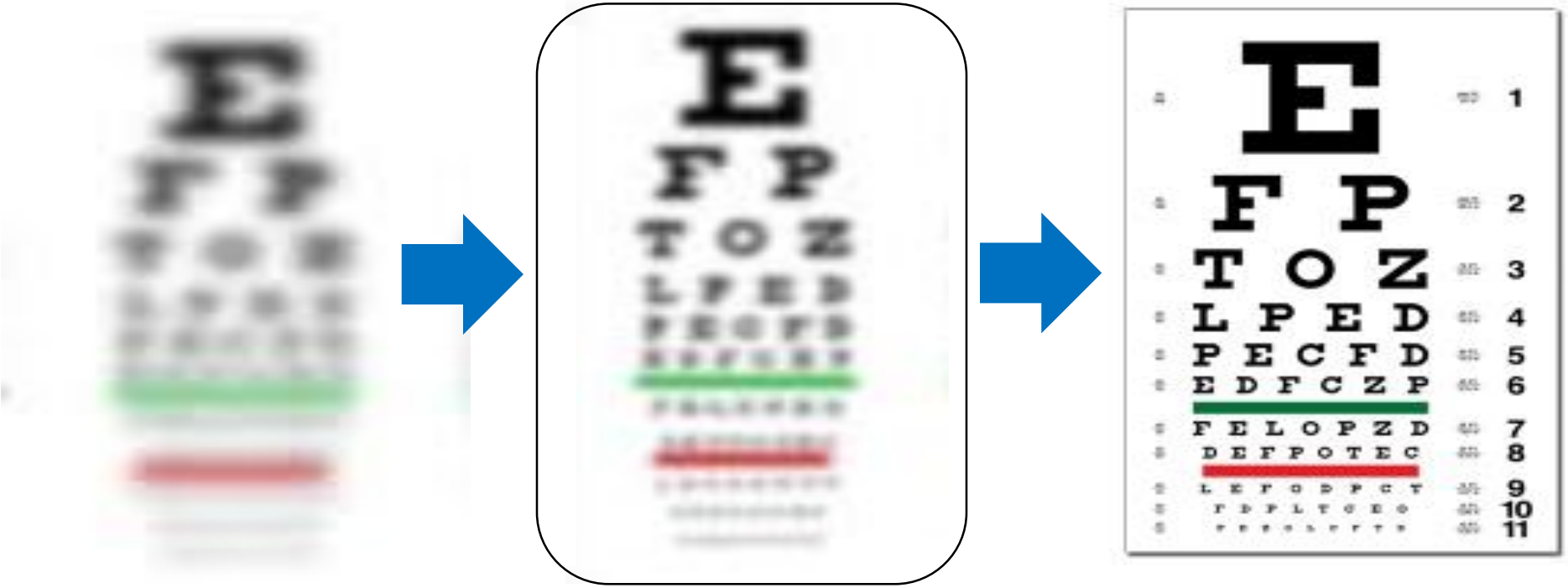
False-Negatives

But How Exactly Did GJP Pull it Off?

Get Right People on Bus	• Spotting/cultivating superforecasters (40% boost)
Teaming	• Anti-groupthink groups (10% boost).
Training	• Debiasing exercises (10% boost)
Elitist Algorithms	• Aggregation algorithms that up-weight shrewd forecasters AND extremize to compensate for conservatism of aggregates (25%--plus boost)

Obama's Osama Decision: Through a GJP Lens

- Hollywood vs. History (the myth and reality of Zero Dark Thirty)
- Two Thought-Experiment Variations on Reality
  - Clones vs. Silos
  - National Security vs. March Madness



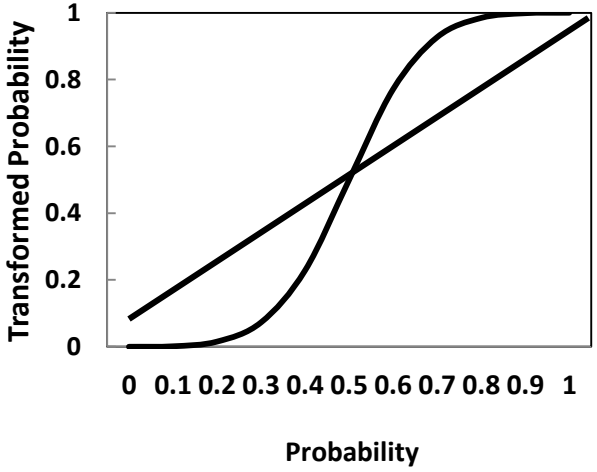
OPTOMETRY TRUMPS PROPHECY

- GJP's methods improve foresight using tested tools: personnel selection, training, teaming, incentives and algorithms
- Still a blurry world, just less so: GJP's best methods assign probabilities of 24-28% to things that don't happen/ 72-76% to things that do

END

Ungar's log-odds model beat all comers (including several prediction markets)

- Log-odds with shrinkage + noise
  - $\mu_j = a \log(p_j/(1-p_j)) + \varepsilon$
- Amount of transformation, *a*, depends on sophistication and diversity of forecaster pool



Measuring the accuracy of probability judgments

Day	Probability of Rain	Outcome of Rain	Brier Scores
1	90%	Yes = 100%	$(1-.9)^2 + (0-.1)^2 = 0.02$
2	50%	Yes = 100%	$(1-.5)^2 + (0-.5)^2 = 0.50$
3	50%	No = 0%	$(0-.5)^2 + (1-.5)^2 = 0.50$
4	80%	Yes = 100%	$(1-.8)^2 + (0-.2)^2 = 0.08$
Mean	68%	50%	0.28

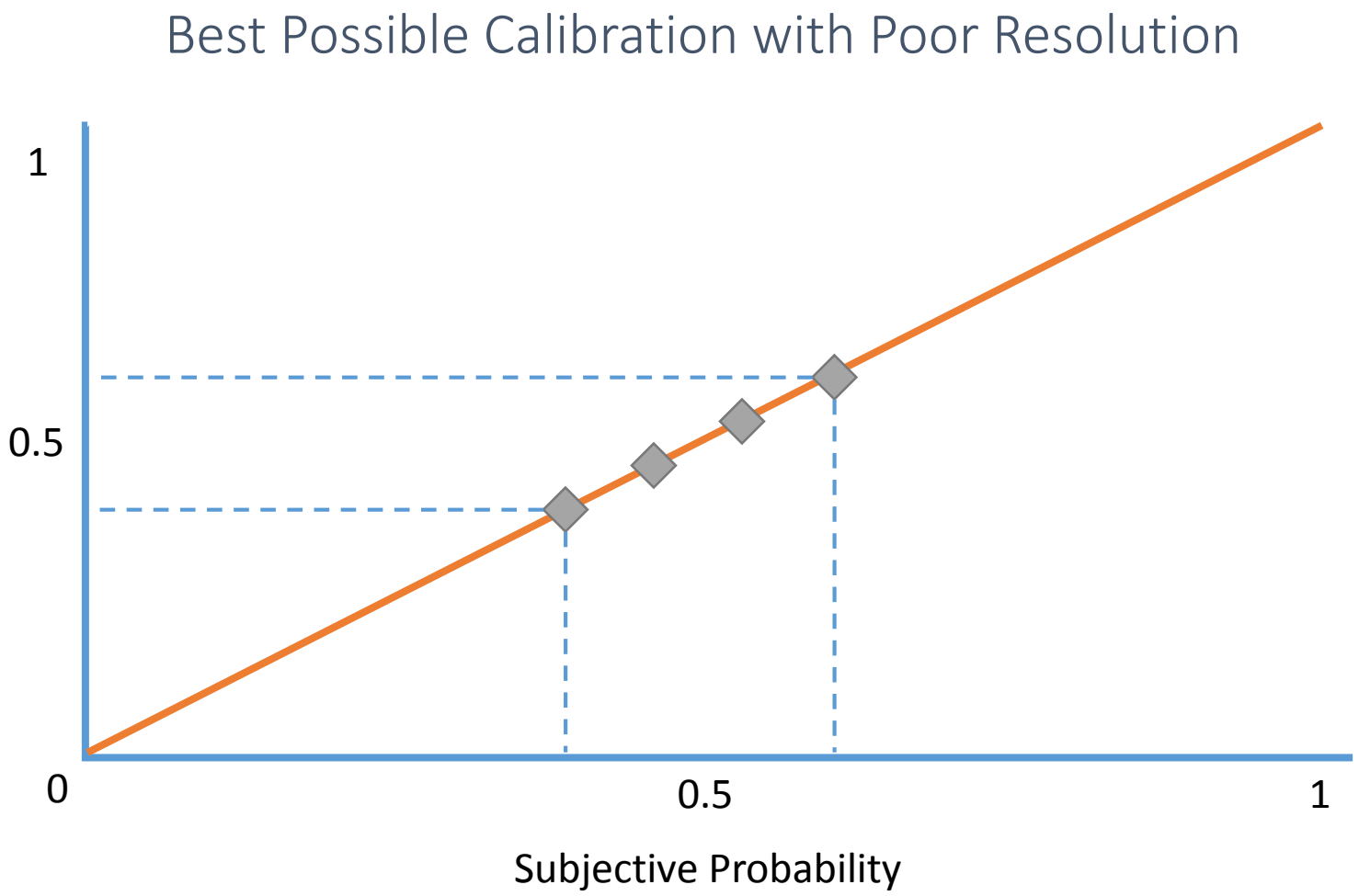
Measuring Accuracy: Brier Scoring

0	.5	2.0
Perfect theory of deterministic system	Just Guessing	Reverse Clairvoyance

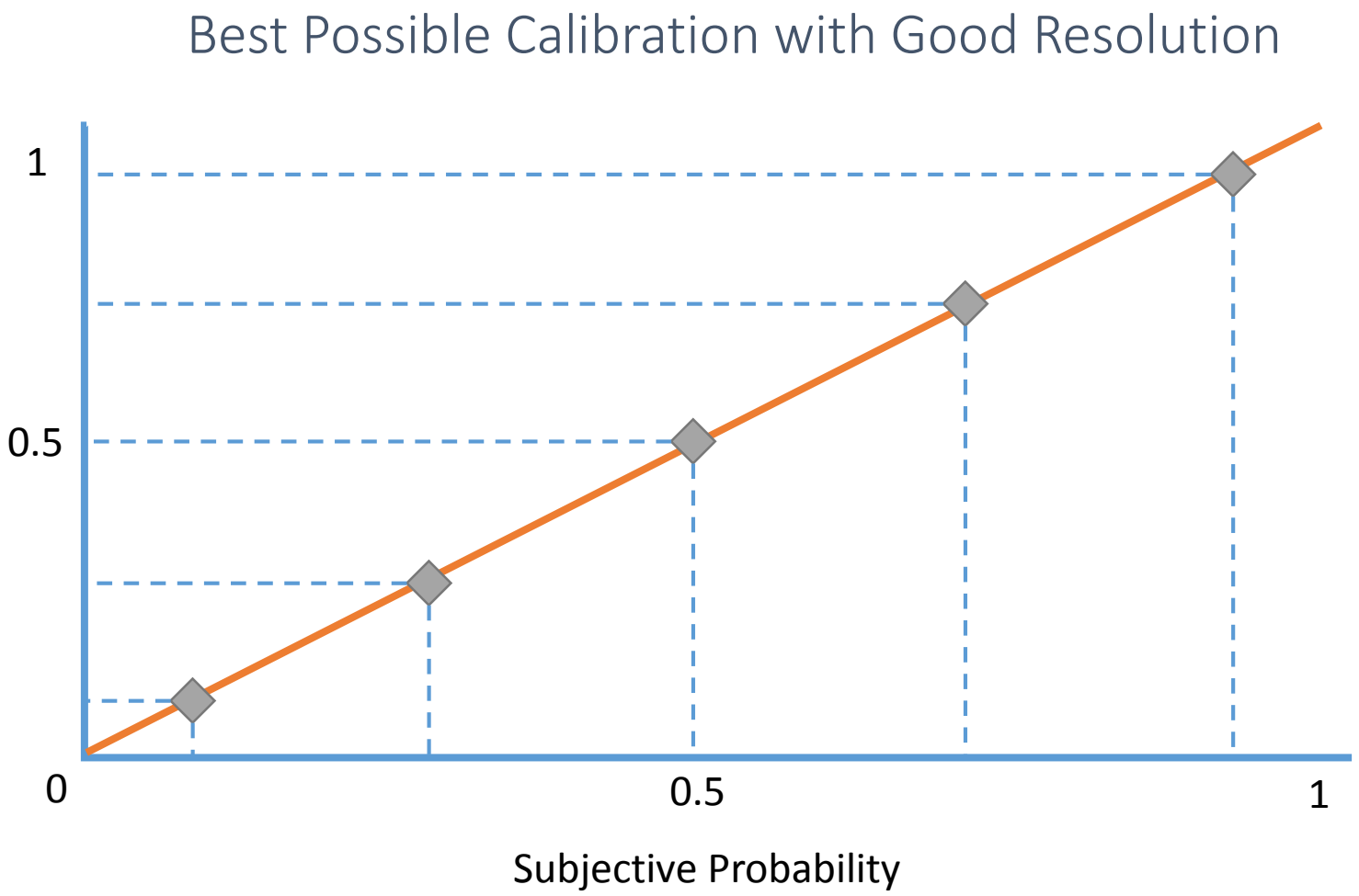
Breaking Brier Scores Down Into Two Key Metrics:

- Calibration
- Resolution

Examples of Calibration & Resolution

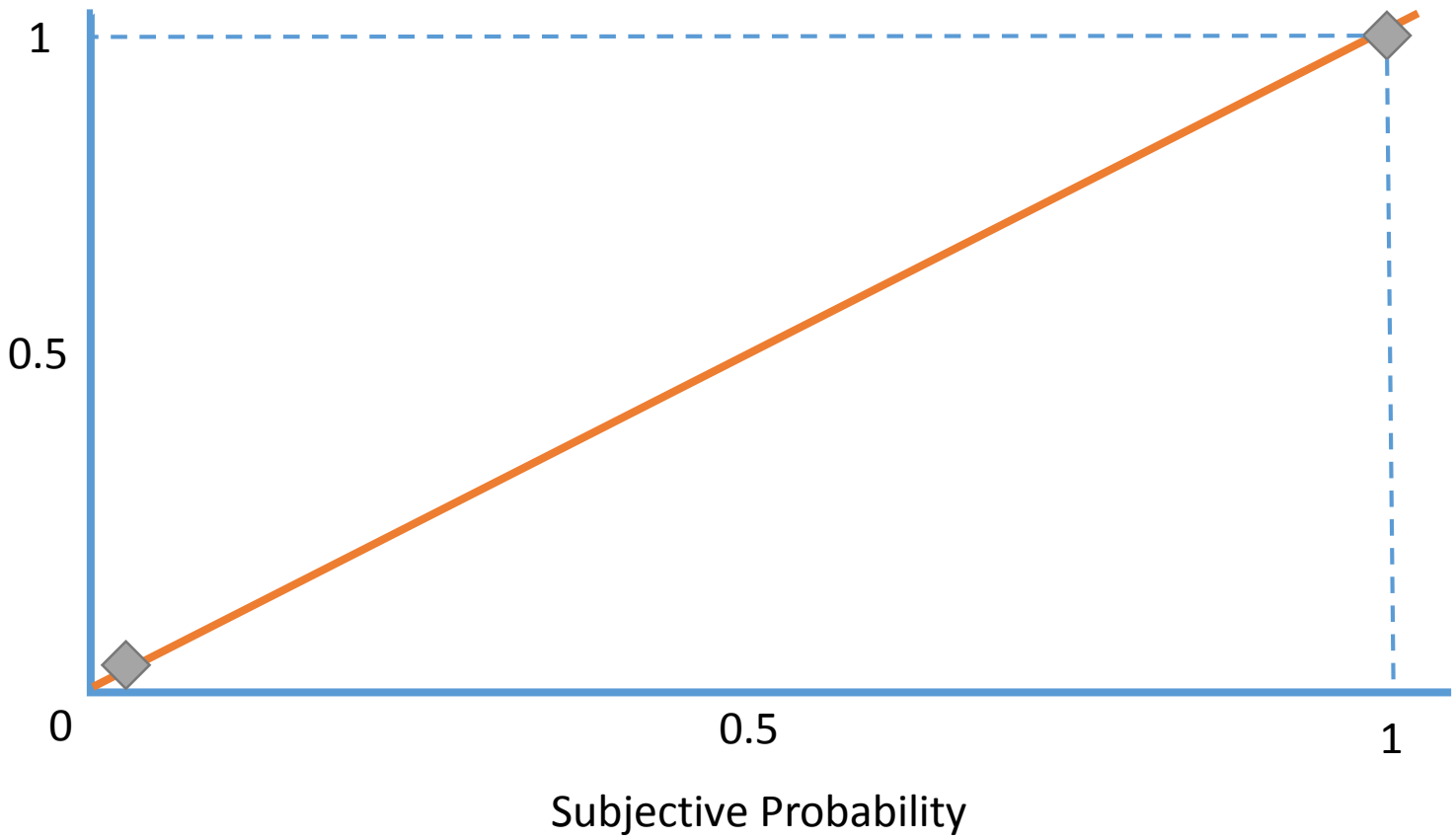


Examples of Calibration & Resolution



Examples of Calibration & Resolution

Best Possible Calibration with Best Possible Resolution



Benchmarking (what should count as a good brier score?)

- Minimalist:
  - Dart-throwing chimp
  - Simple extrapolation/time-series models

- Moderately aggressive:
  - Unweighted mean/median of wisdom of the crowd
  - Expert consensus panels (Central Banks, EIU, Bloomberg,…)

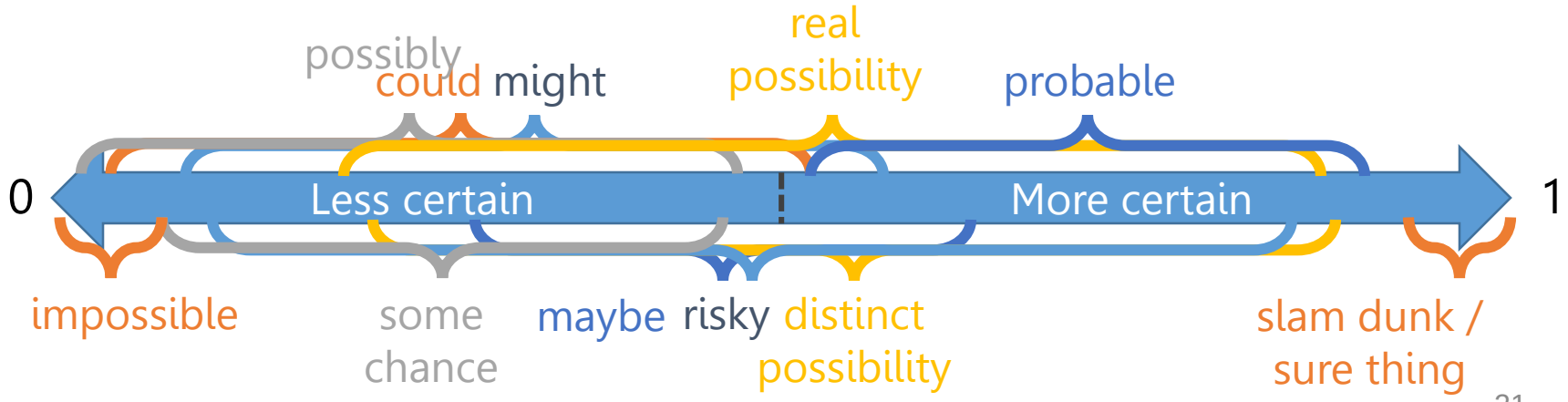
- Maximalist
  - Most advanced statistical/Big-Data models
  - Beating deep liquid markets

Other Take-Aways from the Tournaments

- We discovered:
  - Just how vague “vague verbiage” can be—and how it makes it impossible to keep score
  - The personality/behavioral profiles of superforecasters
  - The group-dynamics profiles of superteams
  - Designing debiasing training that boosts real-world accuracy

Vague verbiage can be very vague

- Watch what happens when we translate words into ***quant-equivalence ranges:***
  - 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*** (QER 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*** (QER, 0.95 to 1.0)



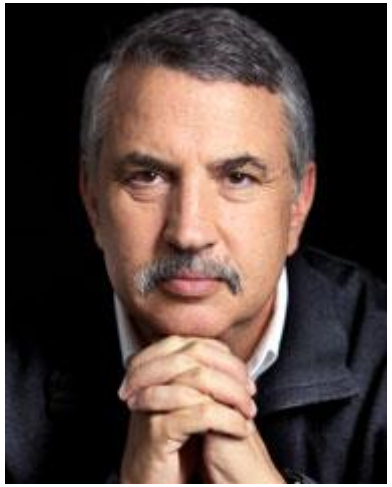
Krugman

Bremmer

Wolf

How Accurate Are Today's Thought Leaders?

Friedman



Kristol

Ferguson

Profiling Superforecasters

- Fluid intelligence helps but without…
- Active open-mindedness helps but without …
- And both combined count for little unless:
  - You believe probability estimation is a skill that can be cultivated—and is worth cultivating

Profiling Superteams

- Somehow manage to check groupthink via precision questioning and constructive confrontation without degrading into factionalism

Yet Goliath Decided to Lend David Slingshot Money

- In 2010, IARPA challenged five $5M-per-year research programs to out-predict a $5B-per-year bureaucracy in a 4-year tournament
- One of these programs, GJP, won the tournament—by big margins