Eva Chen
Shrinidhi Lakshmikanth, David Budescu, Barbara Mellers and Phil Tetlock

GOOD JUDGMENT PROJECT AND THE CONTRIBUTION WEIGHTED MODEL

University of Pennsylvania

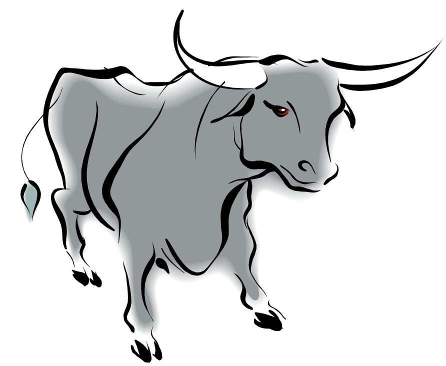
Harnessing the wisdom of the crowd to forecast world events

- IARPA created the ACE Program to dramatically enhance the accuracy, precision, and timeliness of intelligence forecasts
- Development of advanced techniques that elicit, weight, and combine judgments
- Five university-based teams enter the 2011-2015 tournament (GJP eliminated the other 4 teams after the second year)
- Each team submitted forecasts each day for each question, using methods of its choice
- IARPA has posed over 500 questions for the last 4 years:

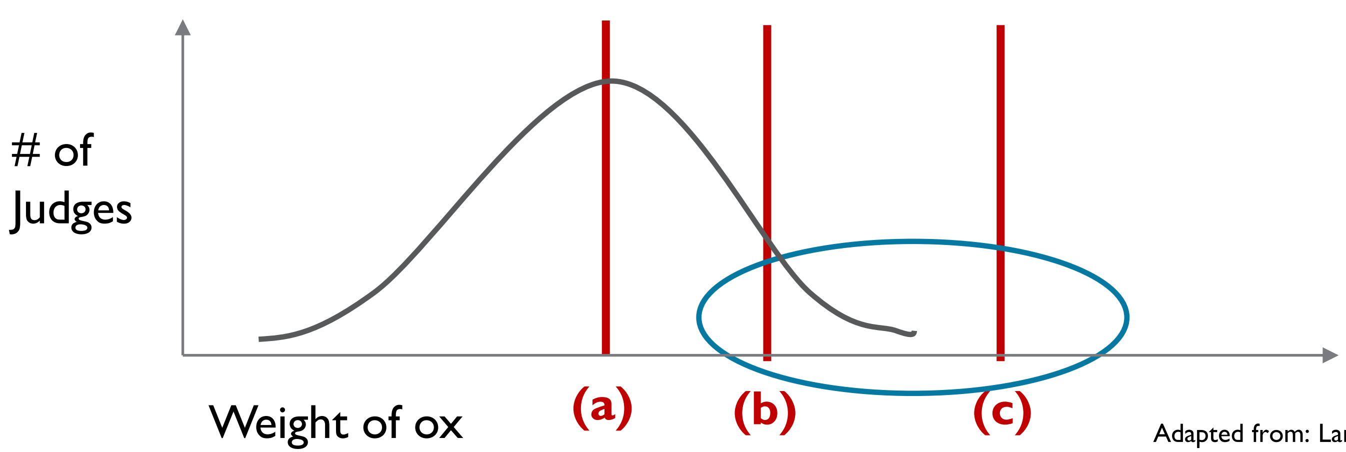
Wisdom of the crowd

Collective intelligence

- Average responses
- Diminish individual errors
- Knowledgeable and diverse
- Better than or equal to:
  - Average individual
  - Randomly selected individual



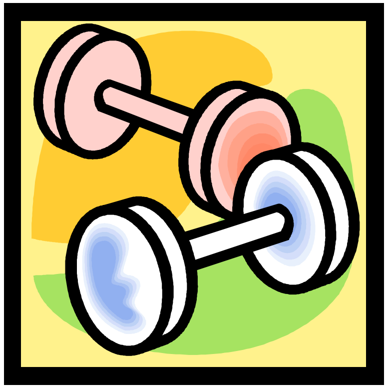
Sir Francis Galton's ox



Adapted from: Larrick, Mannes and Soll (2012)

Aggregation of judgment

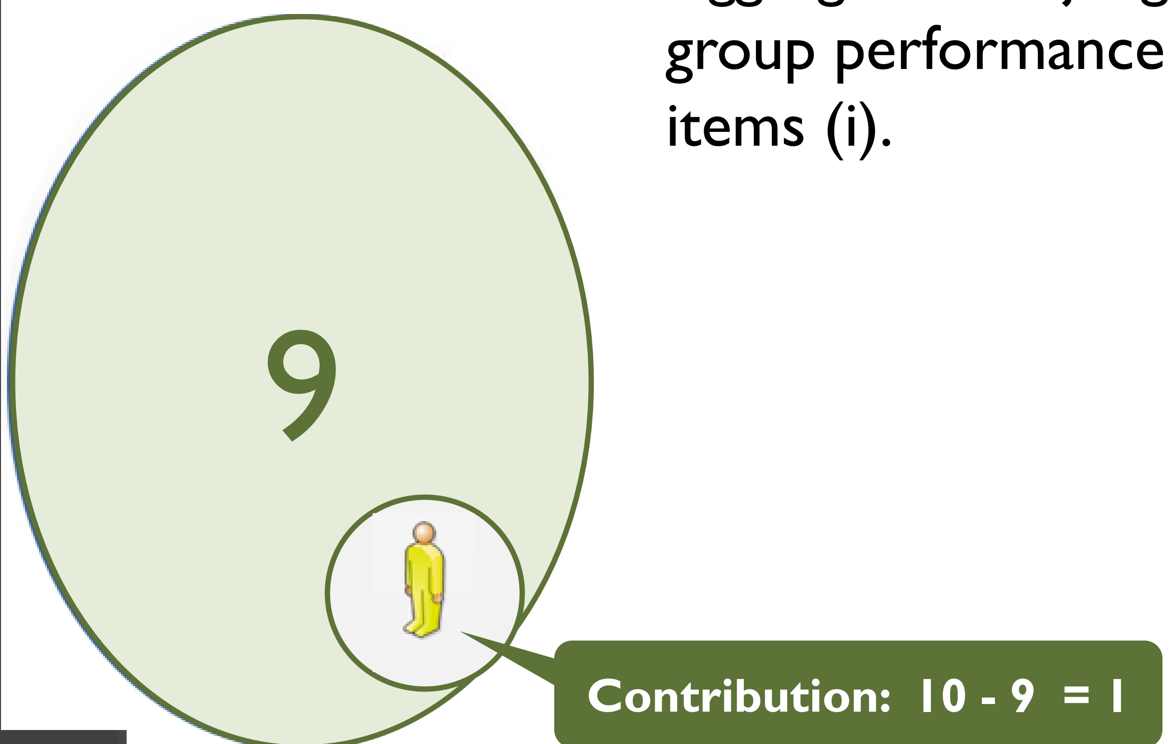
- Methods for aggregation
  - Behavioral (e.g., jury and committee)
  - Markets (e.g., prediction markets)
  - Mathematical
    - Bayesian models
    - Weighting models
- Bases for weights
  - Past performance
  - Test performance (Cooke)



Identifying experts

Contribution:

- Measure the expertise that the judge brings to the group.
- Aggregation of judge's impact on the group performance (Score) across all items (i).



The Aggregation Model

Group's aggregate forecast: $\mathbf{P_{Git} = A\ (P_{jit})}$ Forecast of judge (j) for event (i) at time (t)

Aggregation function

Re-calculate the group's forecast, excluding j: $\mathbf{P_{(G\,-\,j)it}}$

Merit score of the group : $\mathbf{S_{Git} = f\ (P_{Git})}$ **Merit function (e.g. Brier score)**

Judge's contribution to the group **item i at time t**: $\mathbf{C_{jit} = S_{Git} - S_{(G\,-\,j)it}}$

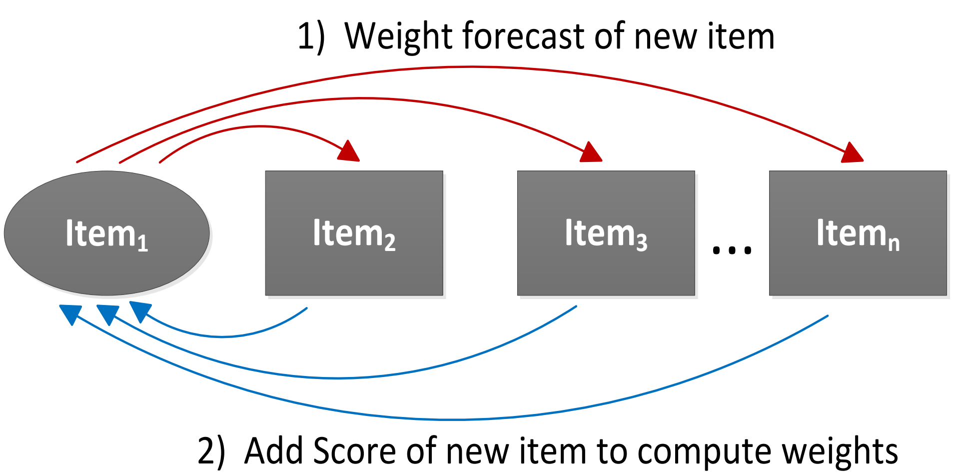
Judge's average contribution : $\mathbf{C_{jt} = \Sigma\ C_{ijt}\ /\ I_j}$ **All I_j items j answers at t**

$\mathbf{C_{jt,}}$

- **reflect the relative expertise of the various judges *in the context of the group***
- **can be positive, negative or 0**
- **can vary over time as more items are being forecasted**

Contribution Weighted Model

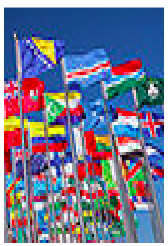
- Budescu and Chen (2015) proposed using a weighted aggregate of all positive contributors.
  - w_{jt}, are scaled such that all $w_{jt} \geq 0$, and $\Sigma\, w_{jt} = 1$.
  - $P_{Gi(t+1)} = A(w_{jt}, P_{ji(t+1)})$ for item i at time (t+1)
- CWM model:
  - weights are proportional to the contribution scores
  - only judges with positive contributions are used.
  - $w_{jt} = 0$ if $C_{jt} \leq 0$, and $w_{jt} = (C_{jt} / \Sigma C_{jt})$ if $C_{jt} > 0$.



Study 1: Geo-political forecasting tournament

Item/event

- Binary
- Ordered multinomial
- Unordered multinomial
- Conditional



#1417 Will Kim Jong Un meet a *head of state from one of the G7 countries, South Korea, China, or Russia **before 1 June 2015?

Opened on 08/27/14, Scheduled to close on 05/30/15 - 90 days

Your last forecast **None**

How likely is this event? 50 %

Probabilistic judgment

Experimental design

- Expertise (training & teaming)

Period1	No Training	Training
Individual	Ind-NT (157)	Ind-T (148)
Team	Team-NT (123)	Team-T (96)

- Facilitation (professional coaches)

Period 2	No Training	Training	Facilitation
Individual	Ind-NT (116)	Ind-T (105)	
Team		Team-NF (126)	Team-F (80)

Data collection

- Data from Jun'12-Jun'13 and from Jun'13-Jun'14
- Collect forecasts from voluntary judges.
- Items from international business, economy, military, policy, politics, etc.
- Judges answer items based on their interest (about 20% of items). We use those who answered ≥ 20 items
- Score (0-100), where 75 score = 0.5 probability

CWM compare to alternative models

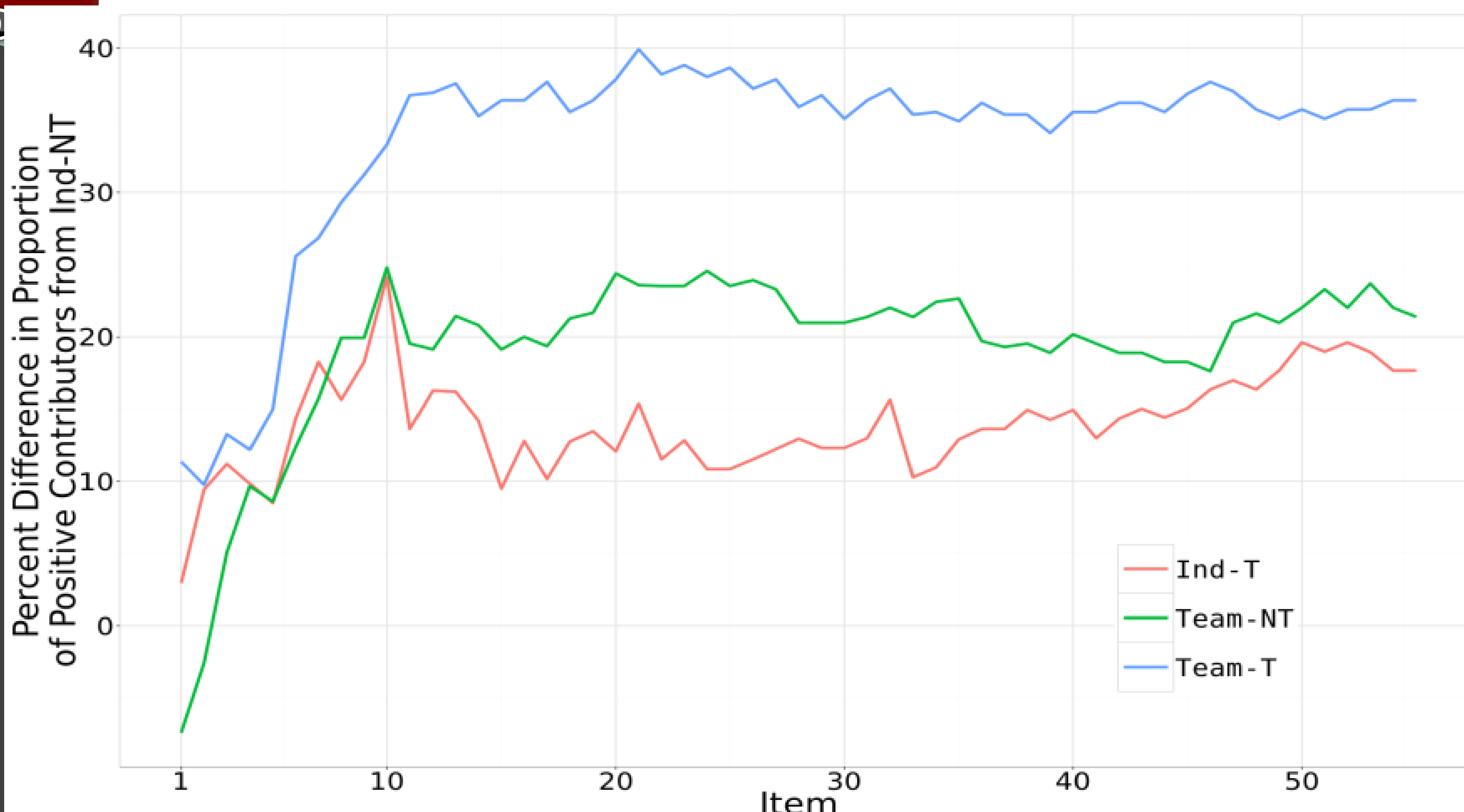
Models	Description
UWM	Unweighted mean of judges with 20 or more items
BWM	Weighted mean based on **past** Scores of judges who answered at least 20 items

- BWM was cross-validated.

CWM beats all models in Period 1

	Conditions			
	Independents		Teams	
	(Ind-NT)	(Ind-T)	(Team-NT)	(Team-T)
Mean Score of **CWM**	94.08	96.64	95.20	97.23
Mean Score of **UWM**	87.98	90.61	90.77	93.12
Mean Score of **BWM**	90.69	92.84	93.40	95.20
Proportion of relative improvement* (PRI) of CWM over UWM (in%)	**50.72**	**64.23**	**48.01**	**50.82`**
Proportion of items when CWM > UWM (in%)	96.43	98.21	91.07	96.42
PRI of CWM over BWM (in%)	**36.38**	**53.13**	**27.25**	**42.45**
Proportion of items when CWM > BWM (in%)	92.86	96.43	89.28	81.07

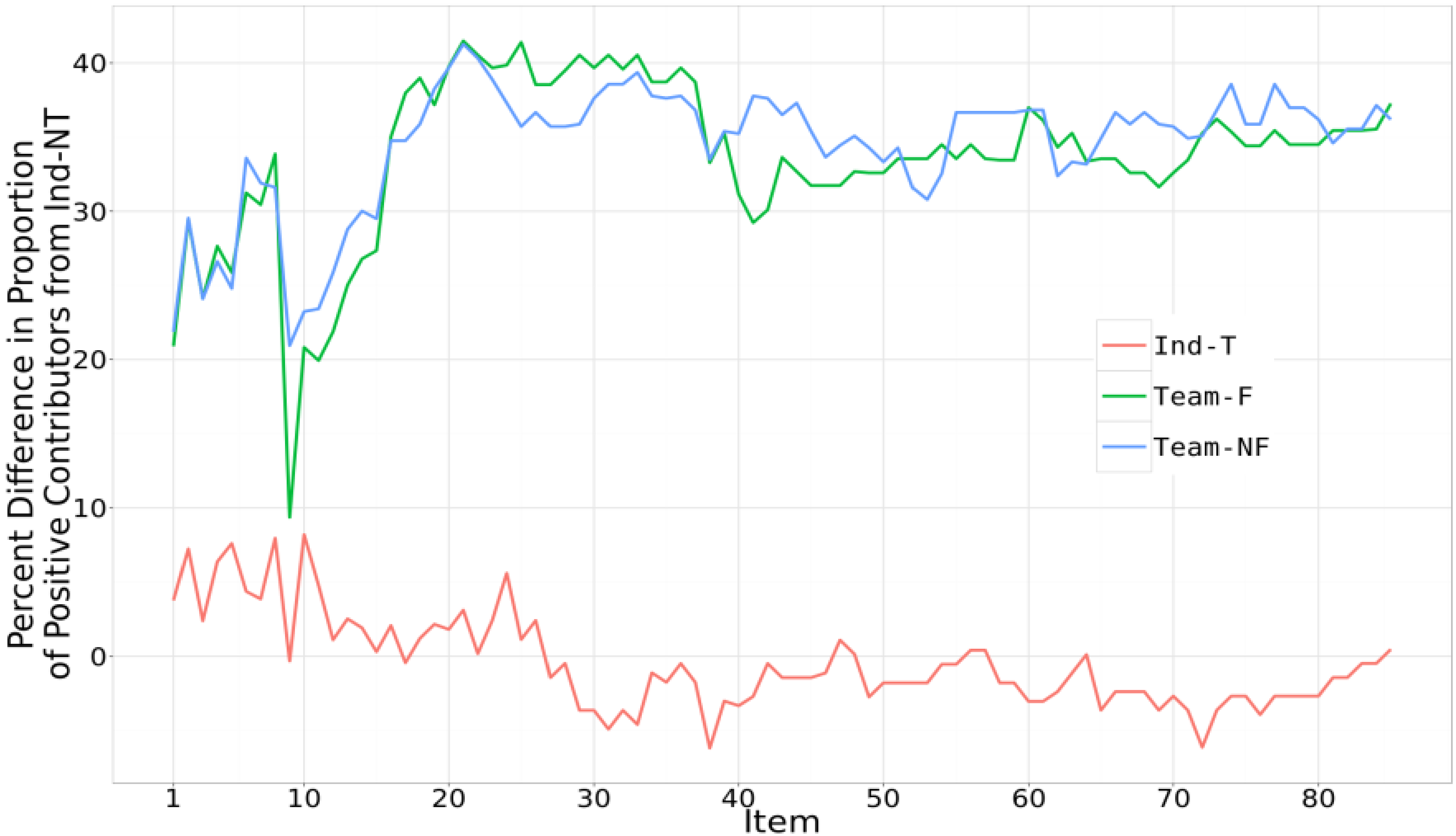
Effect of training and team



CWM beats all models in Period 2

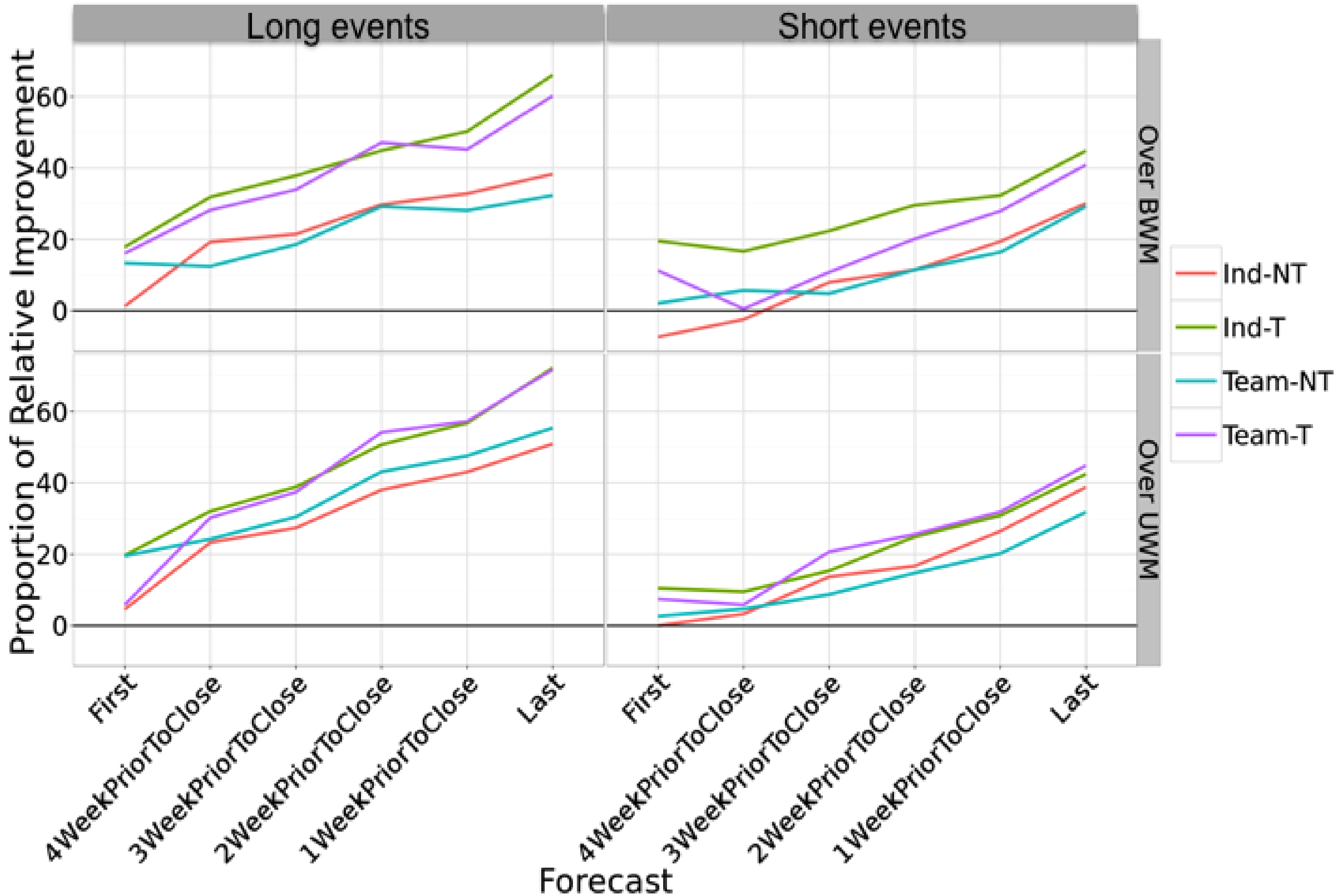
	Conditions			
	Independents		Teams	
	No training (Ind-NT)	Training (Ind-T)	No facilitation (Team-NF)	Facilitation (Team-F)
Mean Score of **CWM**	93.67	93.33	95.77	95.67
Mean Score of **UWM**	89.82	90.67	95.30	95.13
Mean Score of **BWM**	91.83	92.38	95.67	95.09
PRI of CWM over UWM (in%)	**37.84**	**28.51**	**10.00**	**11.02**
Proportion of events when CWM > UWM (in%)	90.70	84.88	75.58	87.21
PRI of CWM over BWM (in%)	**22.55**	**12.42**	**2.18**	**11.77**
Proportion of events when CWM > BWM (in%)	81.40	74.41	67.44	70.93

Effect of facilitation

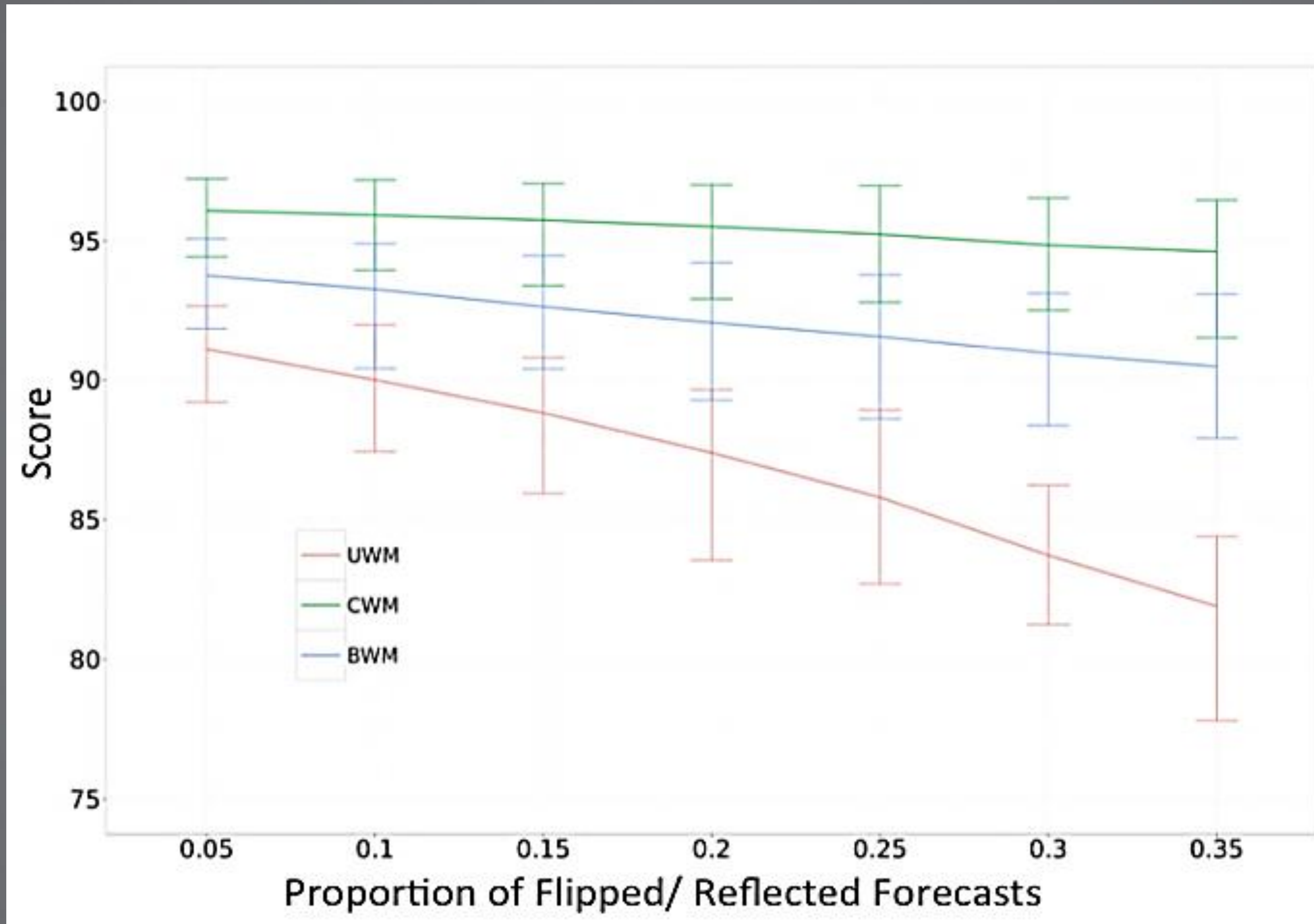


Discrimination

Effect of time



Robustness: Dishonest forecasters



* 50 run simulations using Teams form Period 1

Cost benefit analysis

Cost function= *Items (I)* * *Judges(J)* * *Cost (C)*

- Experts are costly (subset *J*, *w*, where $0 < w < 1$)
- Training questions require time (subset *I*, *p*, where $0 < p < 1$)

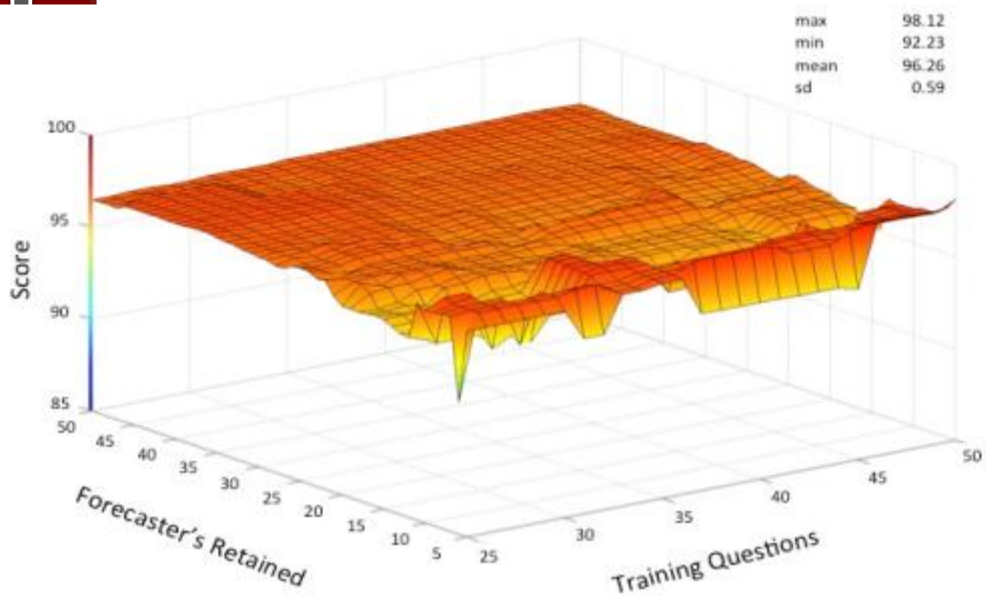
➔ **Maintain accuracy level**

Two scenarios (Ind-T from Period 2, to predict 36 items)**:**

1. Reduce cost by eliminating less contributing judges
2. Reduce cost by randomly eliminating judges

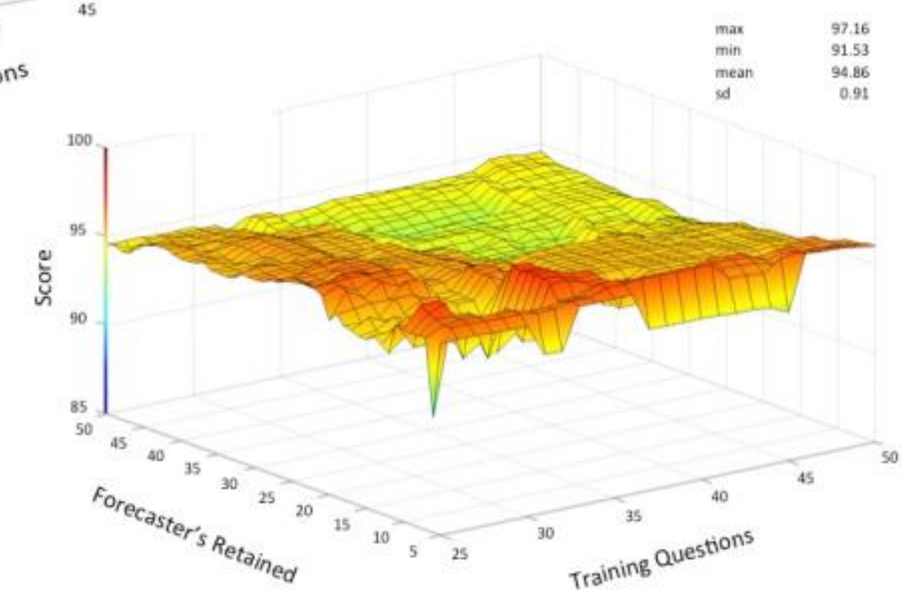
Reduce Cost function = $(p + (1-p)w)\, I\, J\, C$

Cost benefit analysis with top contributors

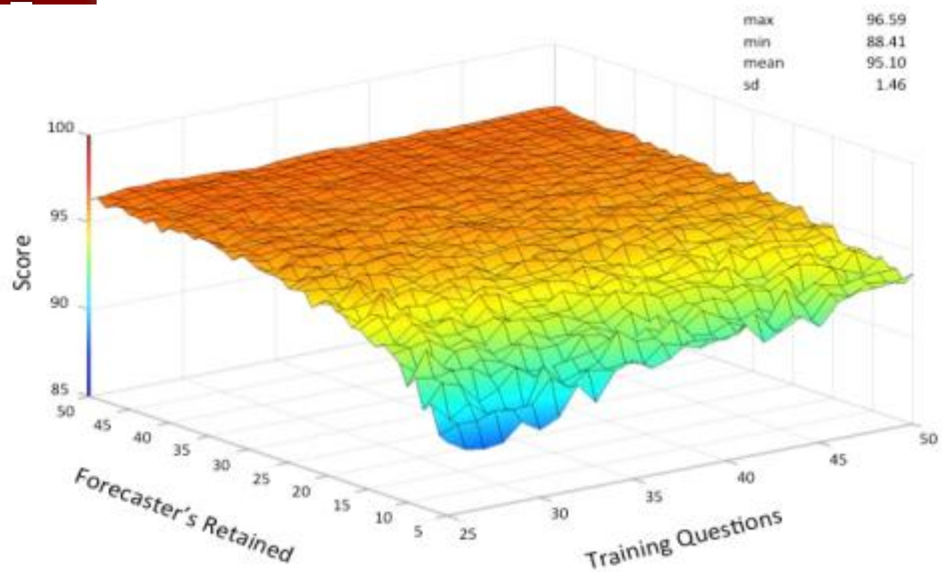


CWM:
Top 20 contributors
25 practice questions
57% saving => 95.29 Score

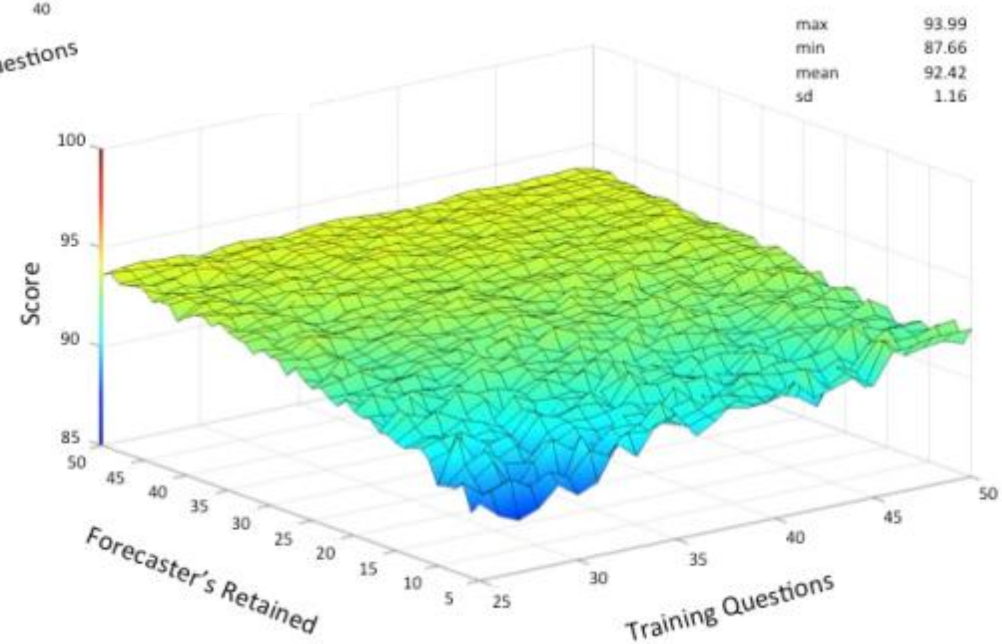
CWM vs. BWM:
PRI: 27.22% better than BWM
SD_{cwm}: 0.59
SD_{bwm}: 0.91



Cost benefit analysis with random forecasters



CWM:
15 contributors
40 practice questions
46% saving => 93.97 Score



* 50 run simulations

Summary of contribution

- Measure of contribution is simple, reliable and useful for assessing forecaster's performance.
- CWM is a better weighting tool in the aggregation process than those built solely on past, individual performance (BWM).

=> **weighting people who have knowledge against the crowd**

- CWM works best when there is expertise in the crowd: training or teaming
- CWM is robust (time, length of items and dishonest forecasters).
- CWM can reduce the cost of expert judgment.

GJ
P

www.goodjudgmentproject.com

Budescu, D.V. & Chen, E. (2015). Identifying expertise to extract the Wisdom of Crowds. *Management Science*, 61(2), 267-280.

This work was supported, in part, by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center.