www.sbvimprover.com

Strategy for Scoring Prediction Performance

Raquel Norel, PhD

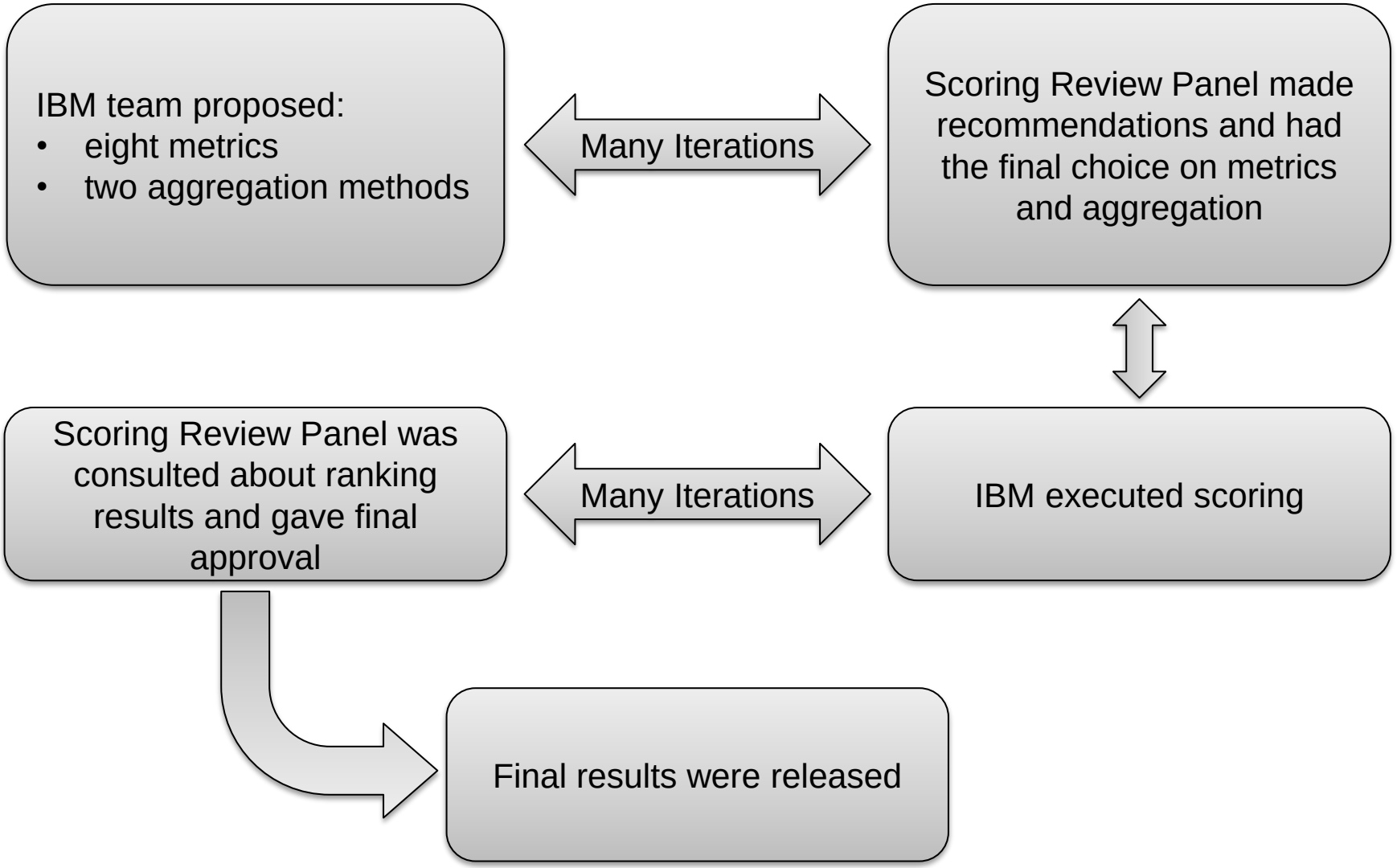
Functional Genomics & Systems Biology

IBM Computational Biology Center

Outline

- Scoring Process: The Scoring Review Panel
- Criteria for scoring and description of metrics

Scoring Process: at all times the identities of the teams were anonymized



Meet the Scorers

- Scoring Review Panel:
  - **Leonidas Alexopoulos**, National Technical University of Athens / ProtATonce
  - **Jim Costello**, Harvard Medical School
  - **Rudiyanto Gunawan**, Swiss Federal Institute of Technology (ETH) Zurich
  - **Torsten Schwede**, Biozentrum University of Basel & Swiss Institute for Bioinformatics
  - **Alfonso Valencia**, Spanish National Bioinformatics Institute (INB)

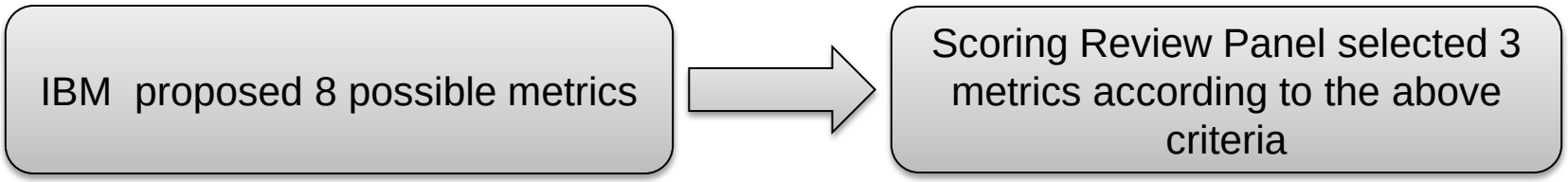
- IBM Scoring Team
  - Erhan Bilal
  - Raquel Norel
  - Gustavo Stolovitzky

Outline

- Scoring Process: The Scoring Review Panel
- **Criteria for scoring and description of metrics**

Rationale behind the chosen scoring methodology

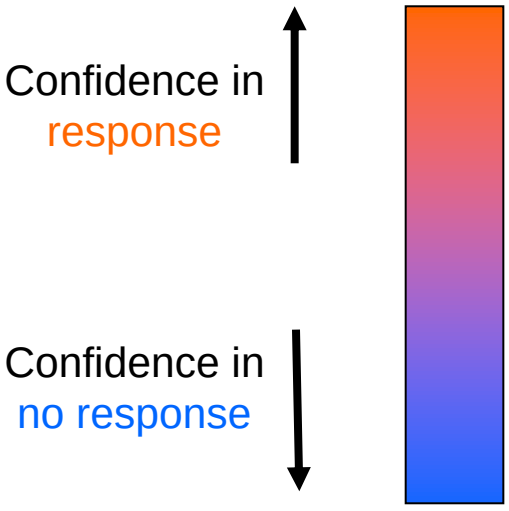
- **Basic premise**: no single metric can capture all the subtleties of a prediction
- We used non-redundant metrics that highlight different qualities of a prediction
  - ✓Threshold vs. non-threshold
  - ✓Order-based vs. confidence based
  - ✓Different ways of rewarding correct vs. incorrect predictions
- A metric should avoid rewarding pathological cases (e.g., predict all items to be one class)
- Unbalanced classes complicated choice of scoring metrics



Binary classification of ordered lists

Participants are required to give confidence values for their predictions of:

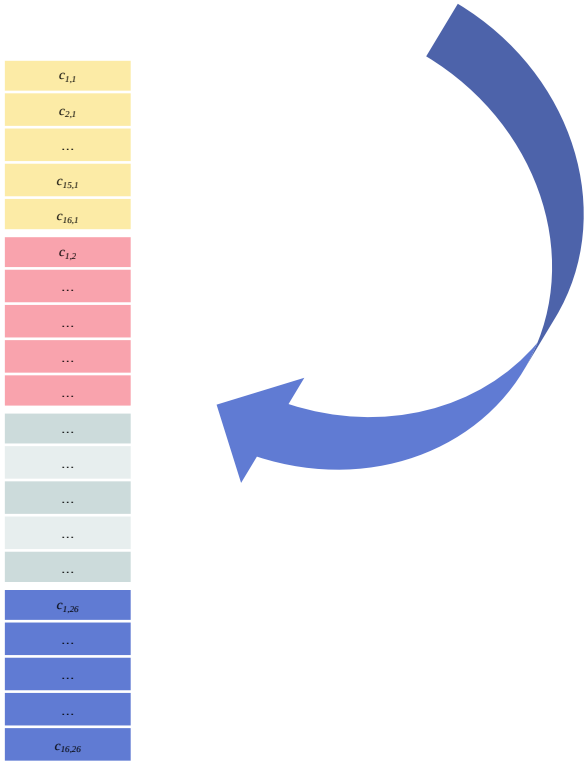
- Protein phosphorylation response to given a stimuli in SC1 and SC2
- Gene set activation for a given stimuli in SC3.



ITEM	Response confidence
item_13	1.00
item_4	0.99
item_52	0.95
…	
item_3	0.01

Vectorize the prediction matrix (few responses or positives calls)

Compound	Cmpd 01	Cmpd 02	…	…	…	…	…	…	Cmpd 26
AKT1	$c_{1,1}$	$c_{1,2}$	…	…	…	…	…	…	$c_{1,26}$
CREB1	$c_{2,1}$	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
TF65	$c_{15,1}$	…	…	…	…	…	…	…	…
WNK1	$c_{16,1}$	…	…	…	…	…	…	…	$c_{16,26}$



Confusion matrix: useful for binary classification

		actual value: *p*	actual value: *n*	total
prediction outcome	*p'*	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
total		P	N	

TP ≍ hit
TN ≍ correct rejection
FP ≍ false alarm
FN ≍ miss

Area Under Precision-Recall Curve (AUPR)

- The precision-recall curve explores, graphically, the tradeoff between these two complementary metrics as k is varied.
- **AUPR** curve presents a single number that summarizes the precision-recall tradeoff.

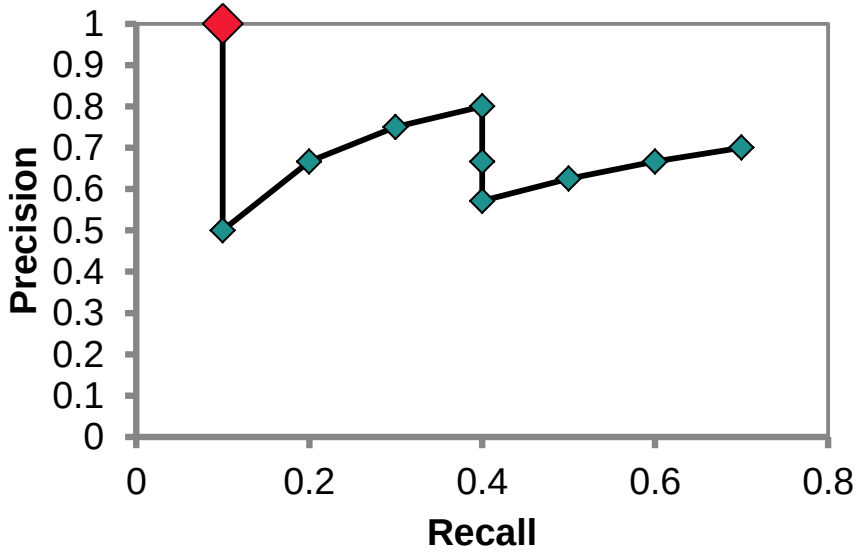
$$prec(k) = \frac{TP(k)}{TP(k) + FP(k)} = \frac{TP(k)}{k}$$

$$rec(k) = \frac{TP(k)}{P}$$

prec: Precision rec: Recall

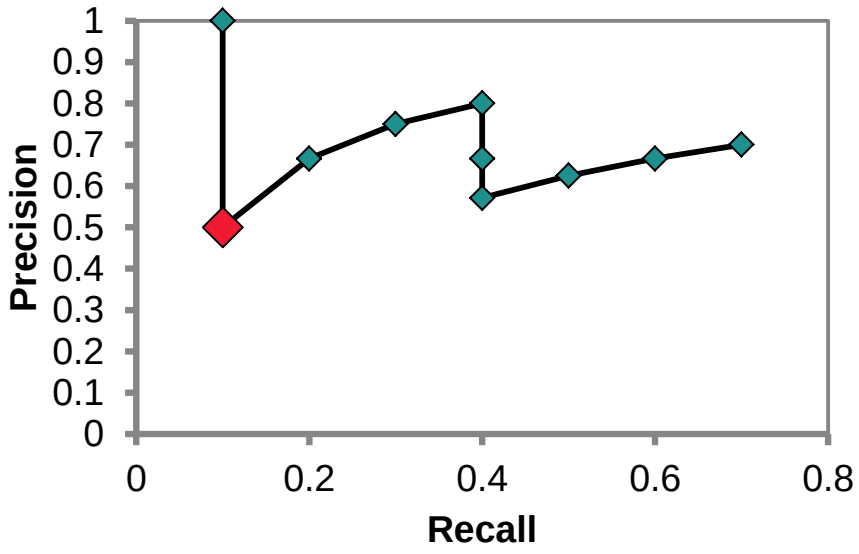
Scoring an ordered list as a binary classification

item ID	Participant predictions	Index k	Gold Standard	Precision TP(k)/k	Recall TP(k)/P
item_4	1.00	1	response	1.00	0.10
item_17	0.99	2	no response	0.50	0.10
item_2	0.95	3	response	0.67	0.20
item_45	0.94	4	response	0.75	0.30
item_13	0.82	5	response	0.80	0.40
…	…	…	…	…	…



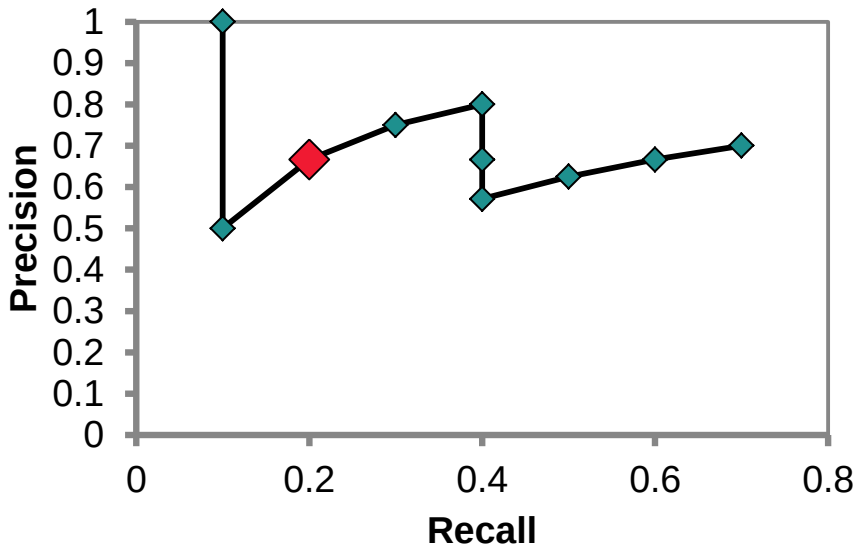
Scoring an ordered list as a binary classification

item ID	Participant predictions	Index k	Gold Standard	Precision TP(k)/k	Recall TP(k)/P
item_4	1.00	1	response	1.00	0.10
item_17	0.99	2	no response	0.50	0.10
item_2	0.95	3	response	0.67	0.20
item_45	0.94	4	response	0.75	0.30
item_13	0.82	5	response	0.80	0.40
…	…	…	…	…	…



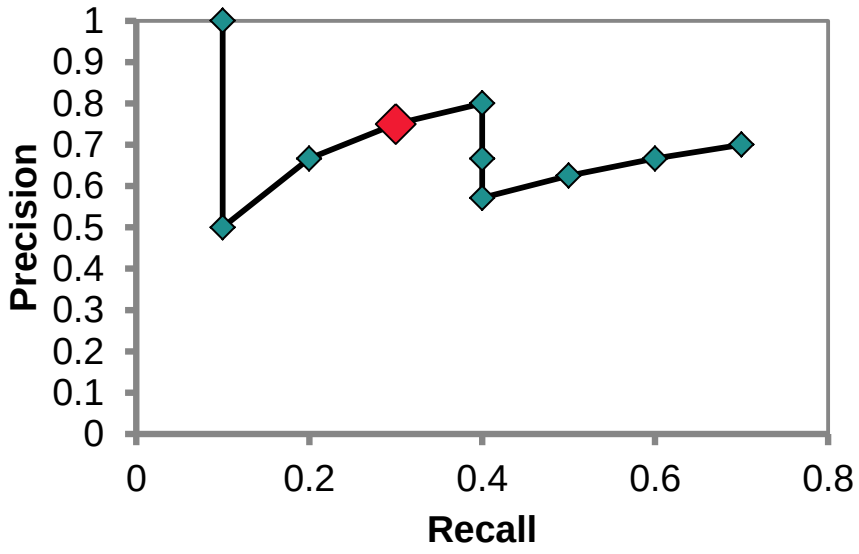
Scoring an ordered list as a binary classification

item ID	Participant predictions	Index k	Gold Standard	Precision TP(k)/k	Recall TP(k)/P
item_4	1.00	1	response	1.00	0.10
item_17	0.99	2	no response	0.50	0.10
item_2	0.95	3	response	0.67	0.20
item_45	0.94	4	response	0.75	0.30
item_13	0.82	5	response	0.80	0.40
…	…	…	…	…	…



Scoring an ordered list as a binary classification

item ID	Participant predictions	Index k	Gold Standard	Precision TP(k)/k	Recall TP(k)/P
item_4	1.00	1	response	1.00	0.10
item_17	0.99	2	no response	0.50	0.10
item_2	0.95	3	response	0.67	0.20
item_45	0.94	4	response	0.75	0.30
item_13	0.82	5	response	0.80	0.40
…	…	…	…	…	…



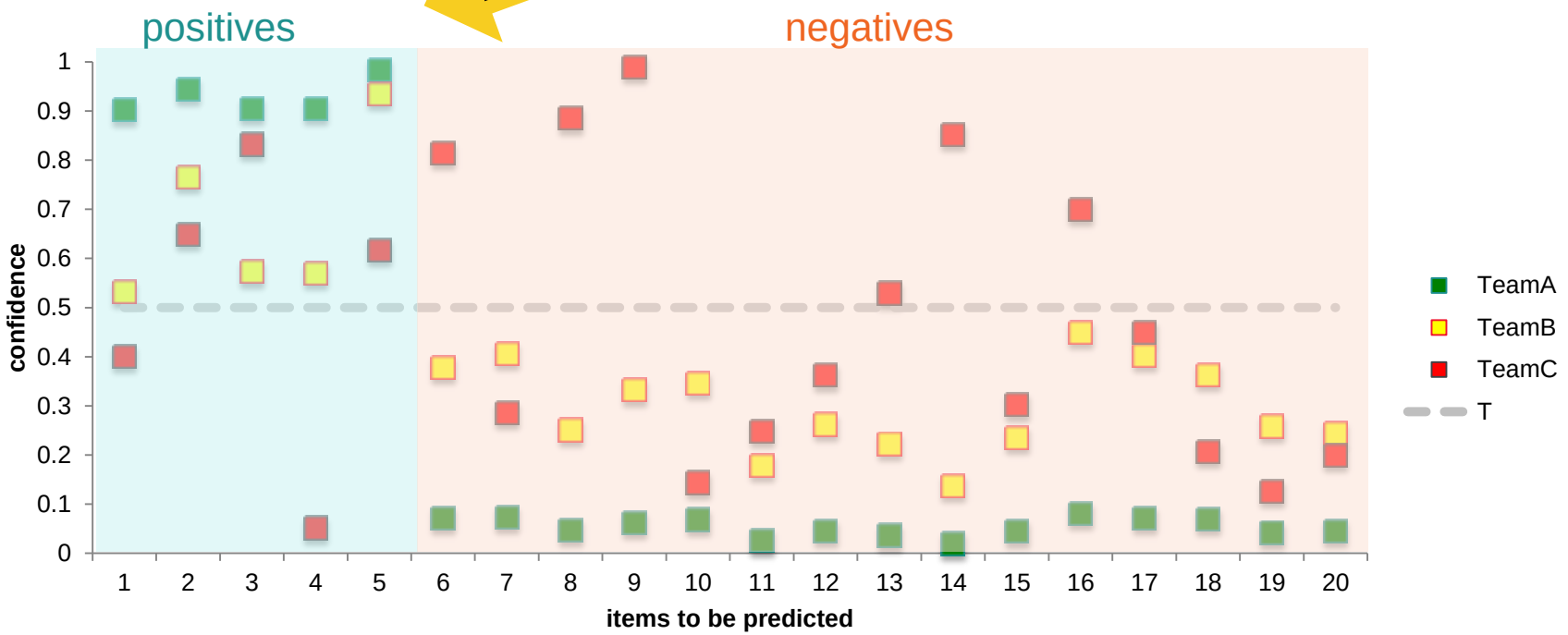
Balanced Accuracy (BAC)

$$BAC = \frac{1}{2}\left(\frac{1}{P}\sum_{\substack{Cij \geq 0.5 \\ |Zij| \geq 3}} 1 + \frac{1}{N}\sum_{\substack{Cij < 0.5 \\ |Zij| < 3}} 1\right) = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$$

BAC = 1.0
BAC = 1.0
BAC = 0.6

Results from toy example

Toy example

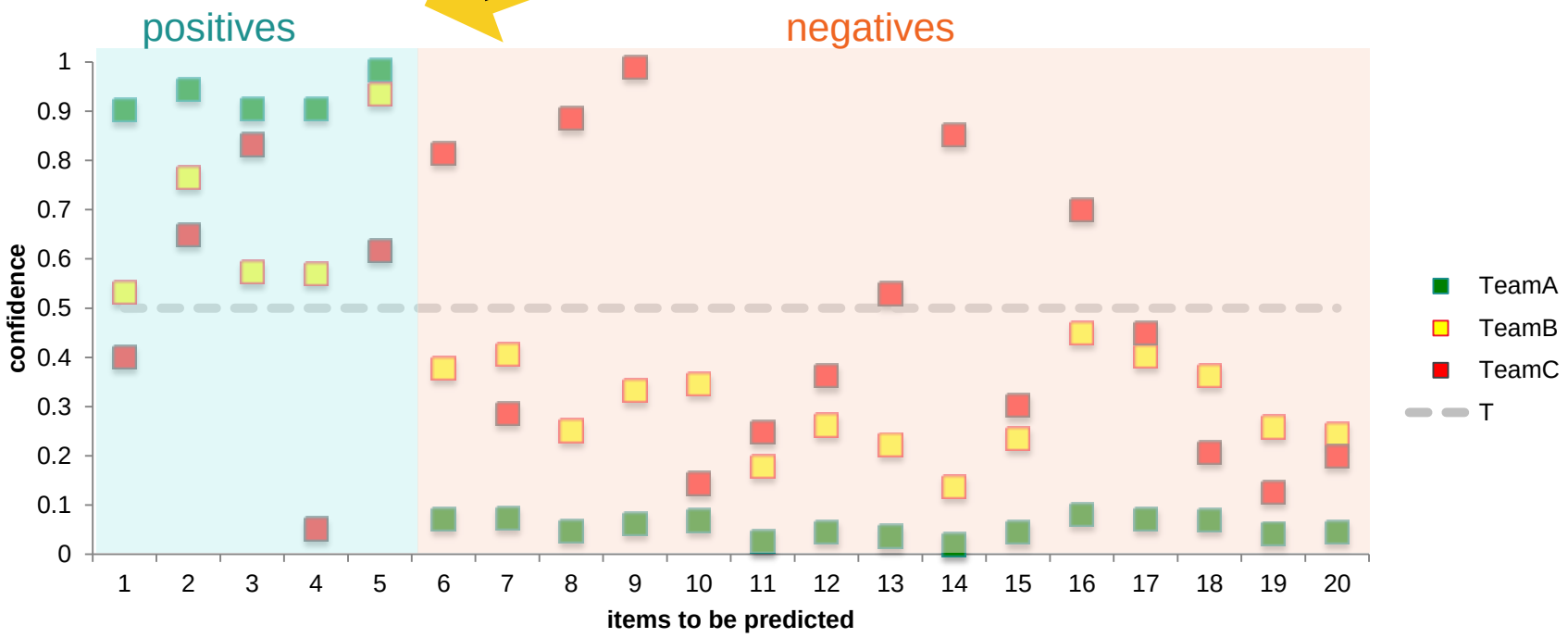


Balanced Accuracy (BAC) + Pearson Correlation (PC)

BAC = 1.0; PC = 0.9
BAC = 1.0; PC = 0.8
BAC = 0.6; PC = 0.1

Results from toy example

Toy example



Metrics' characteristics

	AUPR	BAC	Pearson
Threshold	☐	✔	☐
Confidence	☐	☐	✔
Rank	✔	☐	☐

- Random average: <AUPR> = P/(P+N)
- Random average: <BAC> = 0.5
- Random average: <Pearson> = 0
- Random average: $<Pearson_{normalized}> = 0.5$

To make comparison among metrics easier, we normalized the Pearson correlation:

$Pearson_{normalized} = 0.5*(Pearson + 1)$

With this modification to the Person correlation, all metrics range from [0,1], with 1 being a perfect score.

Methodology for aggregation of metrics

Criteria for aggregation of performance metrics:

- Final performance must reflect a consensus of best performers in majority of metrics
- All metrics carry the same weight

IBM team proposed two aggregation methods:
- Rank-based
- *p*-value based

Scoring Review Panel selected:
- Rank-based aggregation across metrics

Conclusions

- Combination of 3 scoring methods
- The scoring methods cover Confidence, Threshold and Rank
- Balanced Accuracy is especially important as it takes into account the imbalance of the sets
- Overall, methods avoid rewarding pathological submissions (“All on”, “All off”)

Thank you for your attention

The sbv IMPROVER project and www.sbvimprover.com are part of a collaboration designed to enable scientists to learn about and contribute to the development of a new crowd sourcing method for verification of scientific data and results. The project team includes scientists from Philip Morris International's (PMI) Research and Development department and IBM's Thomas J. Watson Research Center. The project is funded by PMI.