Assessing scientific predictive power
Methodological crises
Diagnosis

- Biases
- Hype
- “Bad” statistics
- Bad metrics
- Misaligned incentives
- Group think
- ...

- Metascience research

“The scientific study of science itself’’
Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis

Christiaan H Vinkers assistant professor¹, Joeri K Tijdink psychiatrist⁴, Willem M Otte assistant professor³

BMJ 2015;351:h6467 doi: 10.1136/bmj.h6467 (Published 14 December 2015)
Scientific Predictions
Strategies for testing experts

• Record and quantify predictions
  – The Good Judgment Project
  – Metaculus
  – Prediction markets
Quantifying predictive power

• What does it mean to say “I’m 90% sure?”

• I need a proper scoring rule:
  – Brier scores
  – Surprisal
Quantifying predictive power

• Brier score:

\[ A = \begin{cases} 
1 & \text{if true} \\
0 & \text{if false} 
\end{cases} \]

\[ P_{\text{true}} = \text{Probability for "true"} \]

\[ P_{\text{false}} = \text{Probability for "false"} \]

\[ \text{Brier Score} = (P_{\text{true}} - A)^2 + (P_{\text{false}} - A)^2 \]

• Low scores good, high scores bad!
Brier

- Say I predict “yes” with 80% certainty and the result is “yes”:
  \[ \text{Brier} = .2^2 + .2^2 = 0.08 \]

- Say I predict “yes” with 80% certainty and the result is “no”:
  \[ \text{Brier} = .8^2 + .8^2 = 1.28 \]
Quantifying predictive power

• Logarithmic scores:

\[
\text{Surprisal} = \begin{cases} 
- \ln(P_{\text{true}}) & \text{if true} \\
- \ln(1 - P_{\text{true}}) & \text{if false}
\end{cases}
\]

• Low scores good, high scores bad!
Surprisal

• Say I predict “yes” with 80% certainty and the result is “yes”:

\[
\text{Surprisal} = - \ln P_{true} = - \ln 0.8 = 0.22
\]

• Say I predict “yes” with 80% certainty and the result is “no”:

\[
\text{Surprisal} = - \ln P_{false} = - \ln 0.2 = 1.61
\]
## All Time Stats

<table>
<thead>
<tr>
<th>Brier Score</th>
<th>No. of Questions Forecasted</th>
<th>No. of Forecasts</th>
<th>No. of Upvotes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0.312</strong></td>
<td>45</td>
<td>208</td>
<td>13</td>
</tr>
</tbody>
</table>

*(median: 0.301)*
### GJO

**My Brier Scores**

<table>
<thead>
<tr>
<th>Date</th>
<th>Question</th>
<th>Probability</th>
<th>Brier Score</th>
<th>Loss</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr 30, 2016 05:00PM</td>
<td>Will the government of Colombia and the FARC sign a final peace agreement before 1 May 2016?</td>
<td>92%</td>
<td>0.09</td>
<td>0.216</td>
<td>-0.116</td>
</tr>
<tr>
<td>Jul 26, 2016 07:00PM</td>
<td>Who will win the Democratic Party nomination for the US presidential election?</td>
<td>99%</td>
<td>0.039</td>
<td>0.051</td>
<td>-0.011</td>
</tr>
<tr>
<td>Nov 8, 2016 01:00PM</td>
<td>(Conditional) If Donald Trump wins the Republican nomination, which party will win the U.S. presidential election?</td>
<td>97%</td>
<td>1.507</td>
<td>1.172</td>
<td>0.326</td>
</tr>
<tr>
<td>Nov 8, 2016 01:00PM</td>
<td>Who will win the 2016 US presidential election?</td>
<td>99%</td>
<td>1.241</td>
<td>0.866</td>
<td>0.372</td>
</tr>
</tbody>
</table>

---

[https://www.gjopen.com](https://www.gjopen.com)
Metaculus

When will the Mars helicopter, Ingenuity, stop making successful flights for 6 months?

How high will the Haredi share of Israel’s population be at its peak?

1 dollar/watt solar energy by 2020?

Is the sunflower conjecture true?
Ex Quaerums

Meet our team!

Old Dominion
CS 420 class project!!

Craig Woodington  Nathan Livingston  Taylor Brett
Jena Singleton  Aaron Williams  Ashish Kondaka
Difficult Questions
A paradox about experts

• **Question:** How do I know that I can believe the experts?
  
  – **Answer:** Because they’ve shown themselves to be correct.

• **Question:** But how do I know what is correct?
  
  – **Answer:** Because I have consulted with experts.
When group-assessment doesn’t work

- Porter and Jick (# of citations: 1484)
  “Addiction Rare in Patients Treated with Narcotics,” New England Journal of Medicine, 1980, Letter to the Editor
Deep Questions

• Particle dark matter

VS

• MOND (Modified Newtonian Dynamics)

“...but only MOND has repeatedly predicted observational facts in advance of their discovery”
A prediction competition to incentivize good self-assessment?

- Did research actually confirm or refute a hypothesis?
- Did predictions come *before or after* the fact?
- Was the prediction non-trivial?
- Did the experts agree on the outcome?
- Were there incentives to check the outcome?
A reward system
An (silly) example question

Reference A allows for as many as 300 tornadoes in Kansas in 2024, while such a large number is excluded in the simulation used by reference B.
An (silly) example question

Reference A allows for as many as 300 tornadoes in Kansas in 2024, while such a large number is excluded in the simulation used by reference B.

Will the number of tornadoes observed in Kansas in 2024 equal or exceed 300?
## Predictions

<table>
<thead>
<tr>
<th>Scientist</th>
<th>Prediction</th>
<th>Claimed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>User #1 from Reference A</td>
<td>90% Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>User #2 from Reference B</td>
<td>90% No</td>
<td>No</td>
</tr>
</tbody>
</table>
# Predictions

<table>
<thead>
<tr>
<th>Scientist</th>
<th>Prediction</th>
<th>Claimed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>User #1 from Reference A</td>
<td>90% Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>User #2 from Reference B</td>
<td>90% Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

✗
## Predictions

<table>
<thead>
<tr>
<th>Scientist</th>
<th>Prediction</th>
<th>Claimed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>User #1 from Reference A</td>
<td>90% Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>User #2 from Reference B</td>
<td>90% No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

![Green check mark emoji](green_check_mark_emoji.png)

![Smiley emoji](smiley_emoji.png)
## Predictions

<table>
<thead>
<tr>
<th>Scientist</th>
<th>Prediction</th>
<th>Claimed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>User #1 from Reference A</td>
<td>90% Yes</td>
<td>No</td>
</tr>
<tr>
<td>User #2 from Reference B</td>
<td>90% No</td>
<td>No</td>
</tr>
</tbody>
</table>
## Predictions

<table>
<thead>
<tr>
<th>Scientist</th>
<th>Prediction</th>
<th>Claimed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>User #1 from Reference A</td>
<td>90% Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>User #2 from Reference B</td>
<td>90% Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>User #3</td>
<td>90% Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

✗
## Predictions

<table>
<thead>
<tr>
<th>Scientist</th>
<th>Prediction</th>
<th>Claimed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>User #1 from Reference A</td>
<td>90% Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>User #2 from Reference B</td>
<td>90% No</td>
<td>No</td>
</tr>
<tr>
<td>User #3</td>
<td>50% Yes</td>
<td>No answer</td>
</tr>
</tbody>
</table>
## Predictions

<table>
<thead>
<tr>
<th>Scientist</th>
<th>Prediction</th>
<th>Claimed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>User #1 from Reference A</td>
<td>90% Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>User #2 from Reference B</td>
<td>90% No</td>
<td>No</td>
</tr>
<tr>
<td>User #3</td>
<td>60% Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

(Sort of?)
<table>
<thead>
<tr>
<th>Scientist</th>
<th>Prediction</th>
<th>Claimed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>User #1 from Reference A</td>
<td>90% Yes</td>
<td>No</td>
</tr>
<tr>
<td>User #2 from Reference B</td>
<td>90% No</td>
<td>No</td>
</tr>
<tr>
<td>User #3</td>
<td>60% Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Good

• Before research:
  – Group starts with opposing views
  – Leads to a wide range of competing predictions

• After research:
  – Group comes together in consensus about outcomes
Bad

- Everyone agrees at the outset:
  - Then what was really learned?
  - No incentive for scrutiny/skepticism, checking results or questioning assumptions

- Widespread disagreement after research:
  - Then what was the actual outcome? What was learned?
  - No experts were convinced... nothing has changed
Incentivize “good” behavior within a reward system?

1) Continue to use surprisal for rewarding accuracy

2) But rescale total reward distribution on each question to reward

   – Range of predictions before outcome
   – Degree of consensus after outcome
The Math
3 Stages

• Question
  – Something to predict. “Will ‘A’ be true of false?”

• Prediction
  – Gather probabilistic forecasts (e.g., 85% “yes”)

• Resolution
  – Gather answers as to whether the outcome was true or false
Quantify each stage

- \( j \) labels questions
- \( i \) labels users (people making predictions)
- \( N_j = \text{Number of people making predictions on question } j \)
- \( p_{i,j} = \text{Probability for “yes” from user } i \text{ on question } j \)
Last stages

- Resolution:

\[ v_{i,j} = 1 \] if user \( i \) says the outcome is “yes”
\[ v_{i,j} = -1 \] if user \( i \) says the outcome is “no”
\[ v_{i,j} = 0 \] if user \( i \) supplies no answer

all on question \( j \)
Last stages

• **Mean resolution on question** \( j \):

\[
V_j = \frac{1}{N_j} \sum_{i}^{N_j} v_{i,j}
\]

\(|V_j| = "Consensus"

• **Question** \( j \) **outcome**:

\[
q_j = \begin{cases} 
+1 & \text{if } V_j > 0 \\
-1 & \text{if } V_j < 0 \\
0 & \text{if } V_j = 0 
\end{cases}
\]
Characterizing a question outcome

- \( s_{i,j} = \) Surprisal of user \( i \) on question \( j \)

\[
s_{i,j} = \begin{cases} 
- \ln p_{i,j} & \text{if } q_j = +1 \\
- \ln(1 - p_{i,j}) & \text{if } q_j = -1 \\
0 & \text{if } q_j = 0
\end{cases}
\]

- Mean and standard deviation on question \( j \)

\[
\langle s_j \rangle = \frac{1}{N_j} \sum_i^{N_j} s_{i,j} , \quad \Delta s_j = \sqrt{\langle s_j^2 \rangle - \langle s_j \rangle^2}
\]
What is a “good” prediction outcome

• Wide range of different predictions
  – Group reward should be proportional to $\Delta s_j$

• Group should agree on the results of the study
  – Group reward should be proportional to $|V_j|$
Characterizing a question outcome part II

• What is a “big” surprise on question $j$?

$$s_j^{\text{Big}} = \langle s_j \rangle + \Delta s_j$$

• My surprise is big if $s_{i,j} > s_j^{\text{Big}}$.

• $s_{i,j} - s_j^{\text{Big}}$ is a measure of my relative prediction accuracy.
Reward system

• If my (relative) predictions are always equally accurate, make my reward $\propto R_j$

• If my $R_j$ is always the same, make my reward proportional to relative prediction accuracy, $\propto s_j^{\text{Big}} - s_{i,j}$. 
**Reward system altogether**

- $r_{i,j}$ = award points given to user $i$ on question $j$

$$r_{i,j} = R_j \left( s_j^{\text{Big}} - s_{i,j} \right)$$

- All of user $i$’s reward points:

$$r_i = \sum_{j}^{n_i} r_{i,j}$$
**Simple examples**

<table>
<thead>
<tr>
<th>User 1</th>
<th>Predictions</th>
<th>Validate?</th>
<th>surprisal 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>question 1</td>
<td>0.9</td>
<td>1</td>
<td>0.10536052</td>
</tr>
<tr>
<td>question 2</td>
<td>0.07</td>
<td>-1</td>
<td>0.07257069</td>
</tr>
<tr>
<td>question 3</td>
<td>0.1</td>
<td>-1</td>
<td>0.10536052</td>
</tr>
<tr>
<td>question 4</td>
<td>0.96</td>
<td>1</td>
<td>0.04082199</td>
</tr>
<tr>
<td>question 5</td>
<td>0.99</td>
<td>1</td>
<td>0.01005034</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User 2</th>
<th>Predictions</th>
<th>Validate?</th>
<th>surprisal 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>question 1</td>
<td>0.7</td>
<td>1</td>
<td>0.35667494</td>
</tr>
<tr>
<td>question 2</td>
<td>0.4</td>
<td>-1</td>
<td>0.51082562</td>
</tr>
<tr>
<td>question 3</td>
<td>0.3</td>
<td>-1</td>
<td>0.35667494</td>
</tr>
<tr>
<td>question 4</td>
<td>0.88</td>
<td>1</td>
<td>0.12783337</td>
</tr>
<tr>
<td>question 5</td>
<td>0.99</td>
<td>1</td>
<td>0.01005034</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User 3</th>
<th>Predictions</th>
<th>Validate?</th>
<th>surprisal 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>question 1</td>
<td>0.17</td>
<td>1</td>
<td>1.77195684</td>
</tr>
<tr>
<td>question 2</td>
<td>0.8</td>
<td>-1</td>
<td>1.60943791</td>
</tr>
<tr>
<td>question 3</td>
<td>0.4</td>
<td>-1</td>
<td>0.51082562</td>
</tr>
<tr>
<td>question 4</td>
<td>0.8</td>
<td>1</td>
<td>0.22314355</td>
</tr>
<tr>
<td>question 5</td>
<td>0.95</td>
<td>1</td>
<td>0.05129329</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User 4</th>
<th>Predictions</th>
<th>Validate?</th>
<th>surprisal 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>question 1</td>
<td>0.2</td>
<td>1</td>
<td>1.60943791</td>
</tr>
<tr>
<td>question 2</td>
<td>0.95</td>
<td>1</td>
<td>2.99573227</td>
</tr>
<tr>
<td>question 3</td>
<td>0.7</td>
<td>1</td>
<td>1.2039728</td>
</tr>
<tr>
<td>question 4</td>
<td>0.65</td>
<td>1</td>
<td>0.43078292</td>
</tr>
<tr>
<td>question 5</td>
<td>0.85</td>
<td>1</td>
<td>0.16251893</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User 5</th>
<th>Predictions</th>
<th>Validate?</th>
<th>surprisal 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>question 1</td>
<td>0.19</td>
<td>1</td>
<td>1.66073121</td>
</tr>
<tr>
<td>question 2</td>
<td>0.95</td>
<td>1</td>
<td>2.99573227</td>
</tr>
<tr>
<td>question 3</td>
<td>0.6</td>
<td>-1</td>
<td>0.91629073</td>
</tr>
<tr>
<td>question 4</td>
<td>0.8</td>
<td>1</td>
<td>0.22314355</td>
</tr>
<tr>
<td>question 5</td>
<td>0.9</td>
<td>1</td>
<td>0.10536052</td>
</tr>
</tbody>
</table>
Simple examples

\[ R_j = |\Delta s_j V_j| \]
Simple examples
Key assumptions

• With on average “roughly reasonable behavior” on the part of users, outcome \( q_j \) is a good proxy for objective truth

• This is robust against biases, disagreements, etc

\[
q_j = \begin{cases} 
  +1 & \text{if } V_j > 0 \\
  -1 & \text{if } V_j < 0 \\
  0 & \text{if } V_j = 0 
\end{cases} \quad V_j = \frac{1}{N_j} \sum_{i} v_{i,j}
\]
Project

• Check these assumptions by simulating a group of predicting scientists!

Program for Undergraduate Research and Scholarship (PURS) grant from the Office of Research and Perry Honors College at Old Dominion University

3 semester stipend available!

Talk to me!
trogers@odu.edu
What to simulate

• Take $N$ questions with “God given” yes or no answers

• A group of scientists making predictions

• A group of scientists making resolutions, working toward consensus

• Biases, disagreements and other limitations

• Modification of beliefs as predictions are confirmed/refuted

• Question: Does $q_j$ remain a good approximation to objective truth?
What to simulate

Let X be a fundamental assumption/scientific belief

Belief in X means I am more likely to make accurate predictions

Disbelief means I am less likely to make accurate predictions

Does the reward system push the group toward or away from belief in X, and how quickly?
What happens to surprisal scores as collective belief turns to X?
Going forward

• What to model:
  – Forecasting by each user (many variables)
  – Probability to agree with consensus (more variables)
  – Probability to change minds
  – Various efforts to “cheat”

• Does the reward algorithm need to be tweaked?
<table>
<thead>
<tr>
<th>ID</th>
<th>Question</th>
<th>Resolved?</th>
<th>Name</th>
<th>Date</th>
<th>View More</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What is the chance Betelgeuse will supernova by 2030?</td>
<td>false</td>
<td>test testuser</td>
<td>2022-01-16T00:00:00</td>
<td>View More</td>
</tr>
<tr>
<td>2</td>
<td>Will there be more than 300 tornados in 2022?</td>
<td>false</td>
<td>marie curie</td>
<td>2022-02-20T00:58:42.29</td>
<td>View More</td>
</tr>
<tr>
<td>3</td>
<td>Will Norway win the most medals at the 2022 Winter Olympics?</td>
<td>false</td>
<td>bruce wayne</td>
<td>2022-02-20T00:58:42.34</td>
<td>View More</td>
</tr>
<tr>
<td>4</td>
<td>Will there be a human walking on the surface of Mars by 2025?</td>
<td>false</td>
<td>charles darwin</td>
<td>2022-02-20T00:58:42.36</td>
<td>View More</td>
</tr>
<tr>
<td>11</td>
<td>What are the chances this</td>
<td>false</td>
<td>charles darwin</td>
<td>2022-03-</td>
<td>View More</td>
</tr>
</tbody>
</table>
### Algorithm Testing

**Choose File:** qpr-test-file.csv

<table>
<thead>
<tr>
<th>Prediction ID</th>
<th>User ID</th>
<th>Prediction Text</th>
<th>Is Correct?</th>
<th>Prediction Confidence</th>
<th>Reward Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>Yes</td>
<td>true</td>
<td>95</td>
<td>0.02652377442235699</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>No</td>
<td>false</td>
<td>75</td>
<td>0.004185076615383915</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Yes</td>
<td>true</td>
<td>95</td>
<td>0.02652377442235699</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Yes</td>
<td>true</td>
<td>65</td>
<td>-0.009337927852887902</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>Yes</td>
<td>true</td>
<td>96</td>
<td>0.02652377442235699</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 5 entries