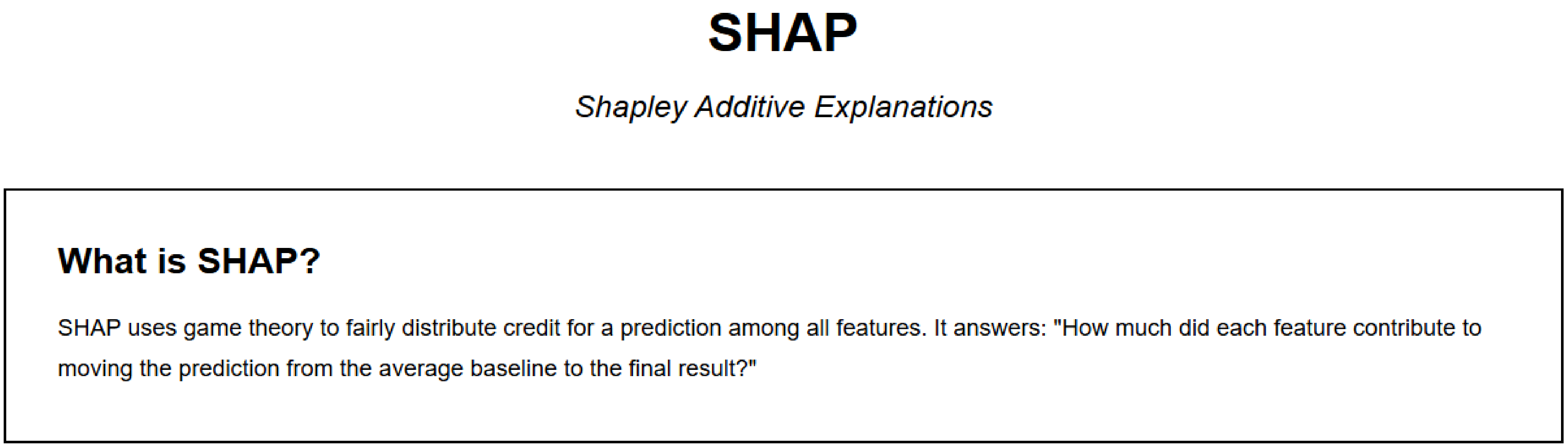
XAI: Human-AI Interaction

- Sharmil N



SHAP

Shapley Additive Explanations

What is SHAP?

SHAP uses game theory to fairly distribute credit for a prediction among all features. It answers: "How much did each feature contribute to moving the prediction from the average baseline to the final result?"

Classroom Analogy

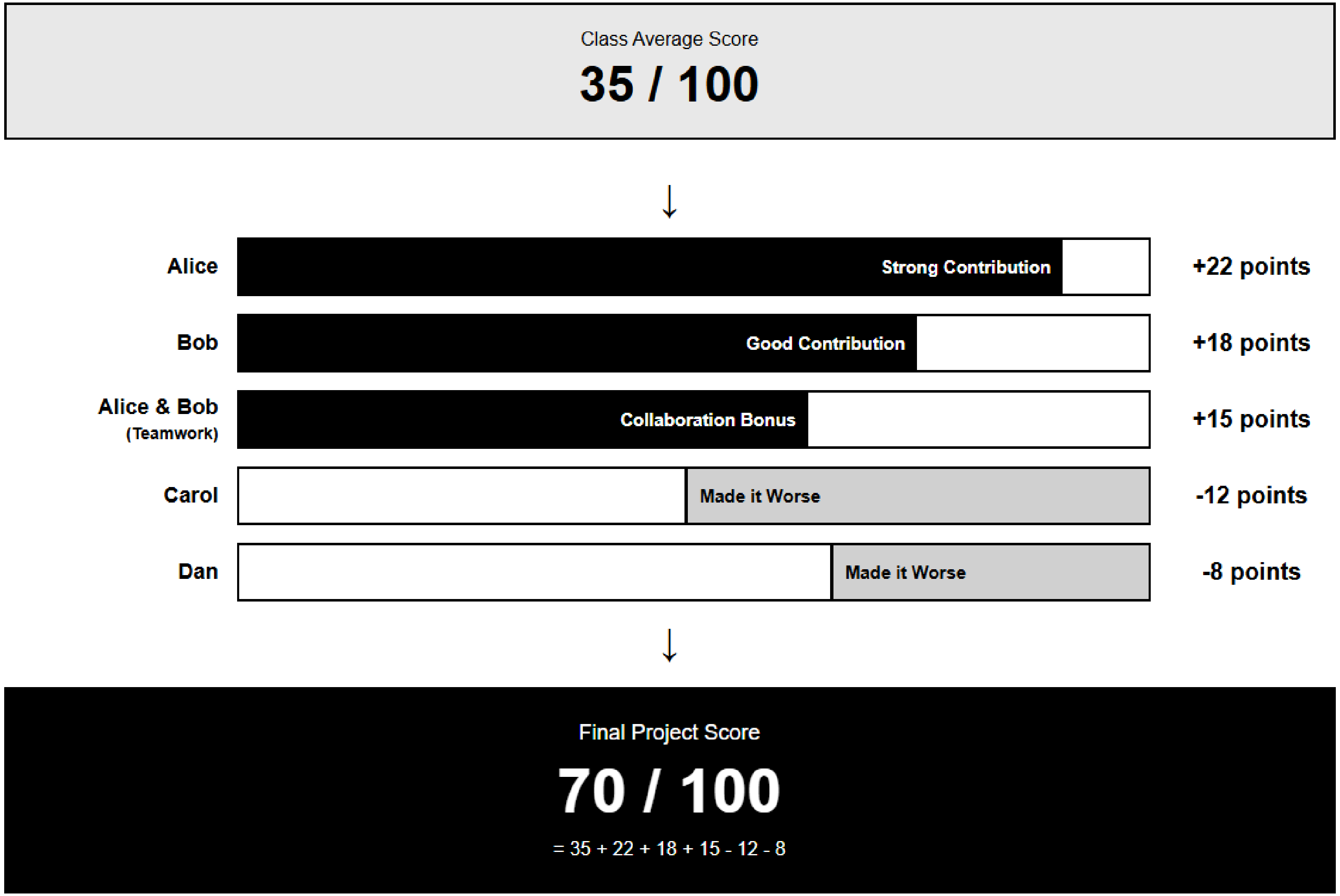
Imagine a group project where four students submit work that scores 70 out of 100.

The class average is 35 out of 100.

Question: How much did each student contribute to getting 35 points above average?

SHAP's Approach:

- Test what happens if only Alice contributes → Gets 45/100
- Test Alice + Bob together → Gets 60/100
- Test Alice + Bob + Carol → Gets 50/100
- Test every possible team combination
- Calculate each student's "marginal contribution" fairly across all scenarios



Class Average Score

35 / 100

↓

Alice — Strong Contribution — +22 points

Bob — Good Contribution — +18 points

Alice & Bob (Teamwork) — Collaboration Bonus — +15 points

Carol — Made it Worse — -12 points

Dan — Made it Worse — -8 points

↓

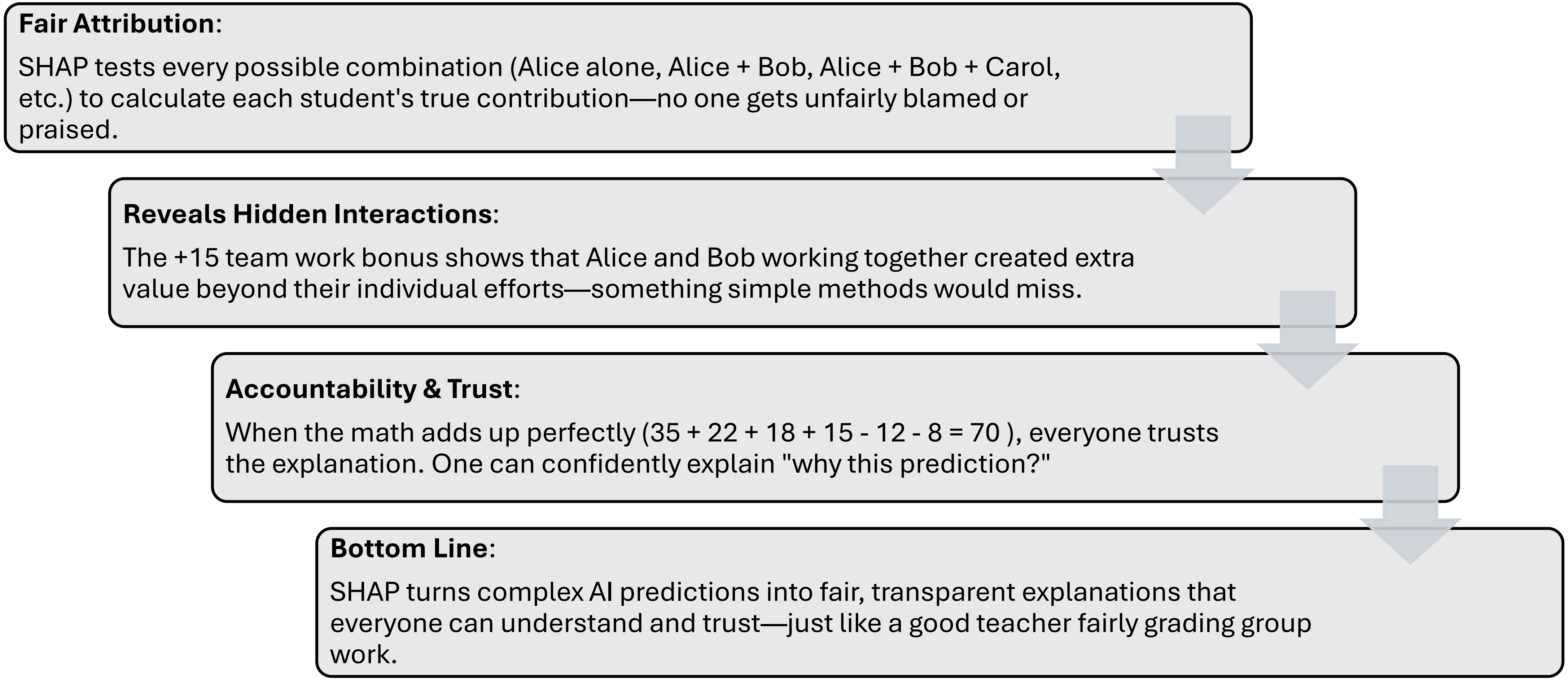
Final Project Score

70 / 100

= 35 + 22 + 18 + 15 - 12 - 8

Just like SHAP tests every combination of features (like testing every combination of students), it ensures fair credit distribution. The contributions always add up exactly to the difference between the prediction and the baseline.

35 points above average = Alice (+22) + Bob (+18) + Their teamwork (+15) - Carol's harm (-12) - Dan's harm (-8) ✓



Fair Attribution:

SHAP tests every possible combination (Alice alone, Alice + Bob, Alice + Bob + Carol, etc.) to calculate each student's true contribution—no one gets unfairly blamed or praised.

Reveals Hidden Interactions:

The +15 team work bonus shows that Alice and Bob working together created extra value beyond their individual efforts—something simple methods would miss.

Accountability & Trust:

When the math adds up perfectly (35 + 22 + 18 + 15 - 12 - 8 = 70), everyone trusts the explanation. One can confidently explain "why this prediction?"

Bottom Line:

SHAP turns complex AI predictions into fair, transparent explanations that everyone can understand and trust—just like a good teacher fairly grading group work.

Citation: All images were generated by Claude. The test case and idea was completely original; Claude was used to help guide the data representation. 10/19/25 9:15pm EST