Using Machine Learning to Predict Integrating Computation into Physics Courses

Nicholas T. Young[1], Marcos D. Caballero[1,2,3]

[1]Department of Physics and Astronomy, Michigan State University
[2]CREATE for STEM Institute, Michigan State University
[3]Department of Physics and Center for Computing in Science Education, University of Oslo

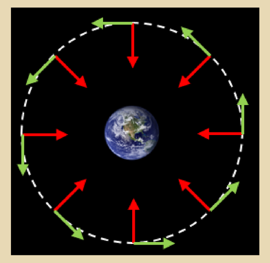
PERL @MSU

CREATE for STEM Institute

Department of Physics and Astronomy

Background

- Computation
  - Is central to the study of modern science & engineering
  - Can help students develop research skills, scientific ways of thinking, & deeper conceptual understanding [1]
- Calls to include computation in physics are both national[2] and locally emergent[3]
- **Goal: Determine factors that are predictive of whether faculty include computation in their physics courses**



Methodology

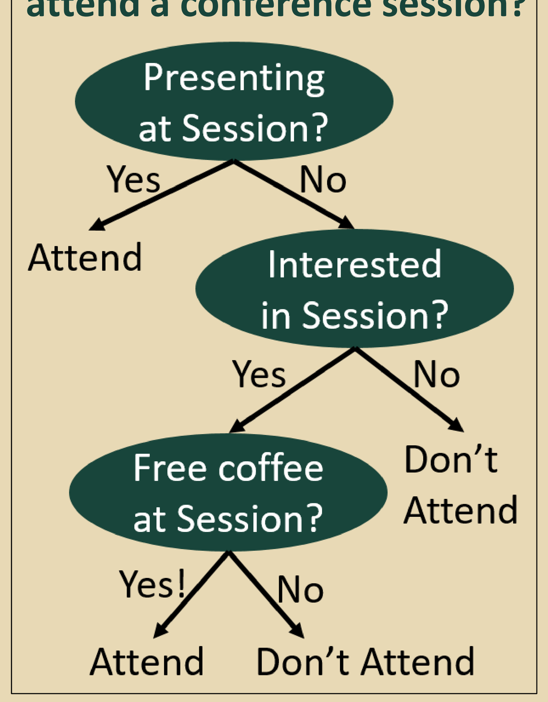
Survey
- AIP survey distributed to a random sample of all US physics faculty in fall 2016
  - Contains binary responses, Likert scale selections, and free response questions
- Areas of focus:
  - types of computational instruction
  - institutional resources and supports
  - faculty perceptions and motivations
  - perceived barriers
- Responses from 1246 faculty and 357 unique departments

Sample
- Use responses from 693 faculty on 44 items in this study
- Complexity of data and characterizing the analysis as a categorization problem suggests machine learning, e.g., random forests[4]

The Random Forest Algorithm [4,5]

Decision trees
- Segregate data based on binary features building rules to predict categories based on features[6]
- Overfit data: poorly predict because rules based on single decision tree instance

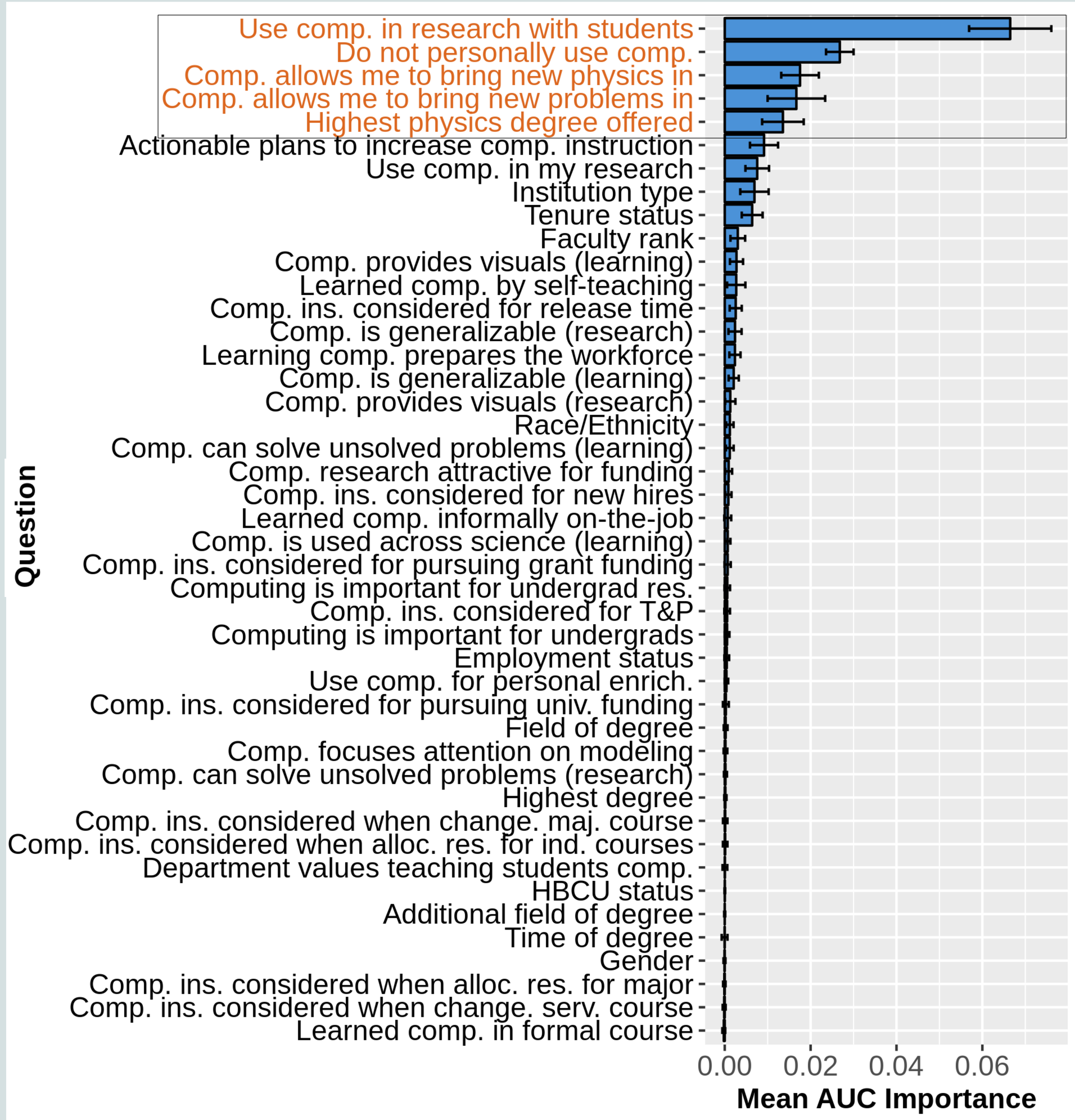
Decision tree: Does a person attend a conference session?



Random Forest
- Combine decision tress and get better results!
  - Good models add up; bad models cancel
- Important variables determined by changes to model when removing that variable[7,8]

Results

Which factors are important?



Is the model a good one?

Accuracy score
- Percent of correctly predicted categories

Confusion matrix
- Visual representation comparing model predictions to what the data says

ROC curve
- Illustrates diagnostic ability of the model in terms of false positive rate and true positive rate

Validation
- Ensure results do not change with different model parameters

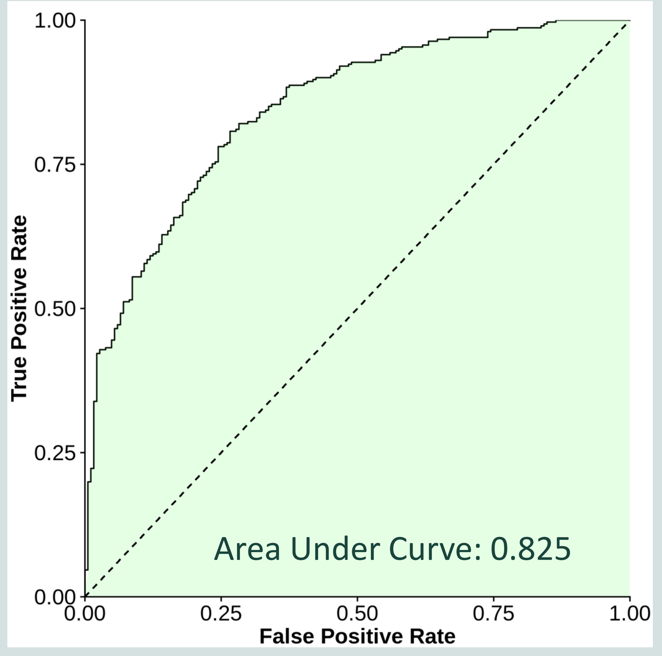
Models were developed by training from 70% of the data and using 30% for testing.

Do you teach computation?		Data Says	
		No	Yes
Model Predicts	No	43	12
	Yes	35	118

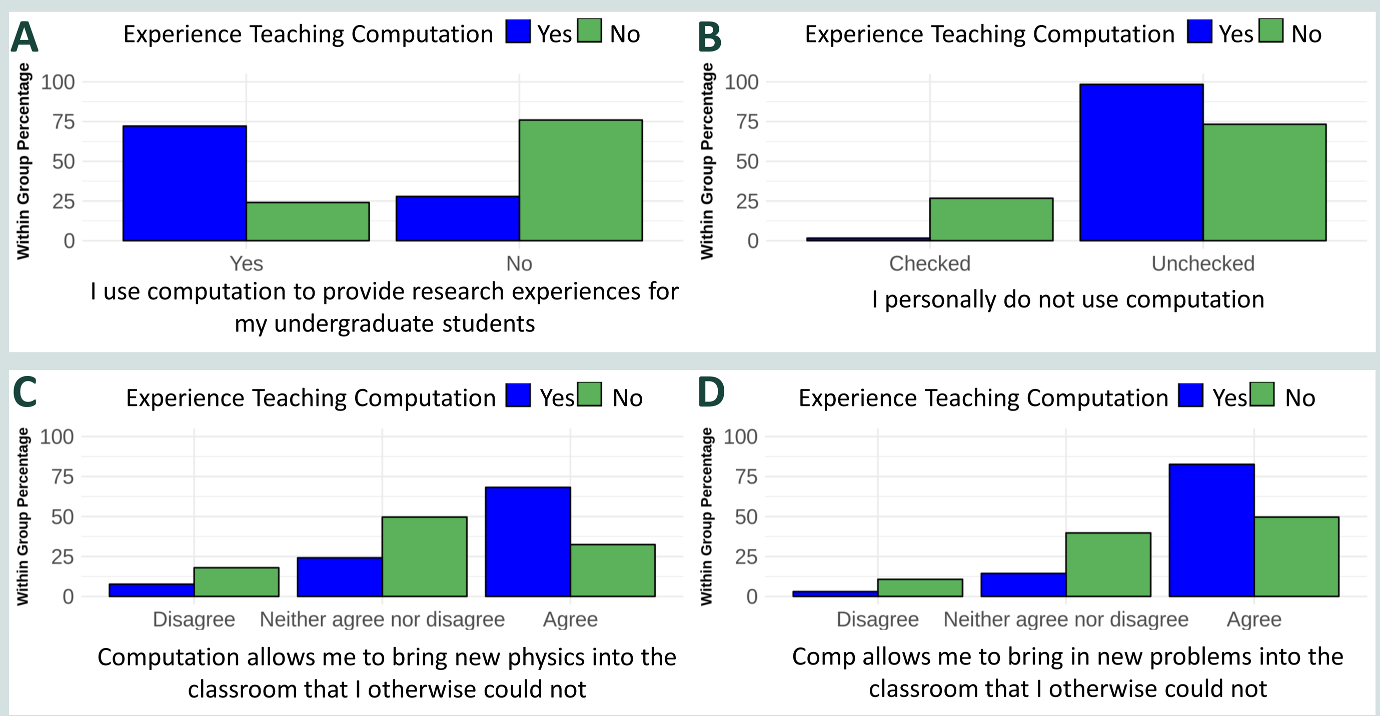
Accuracy: 0.774

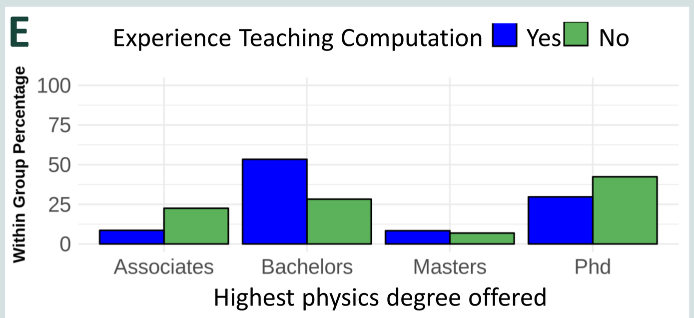
Average Accuracy: 0.774 ± 0.005
Average AUC: 0.838 ± 0.002

Accuracy and Area Under Curve values suggest a good model.



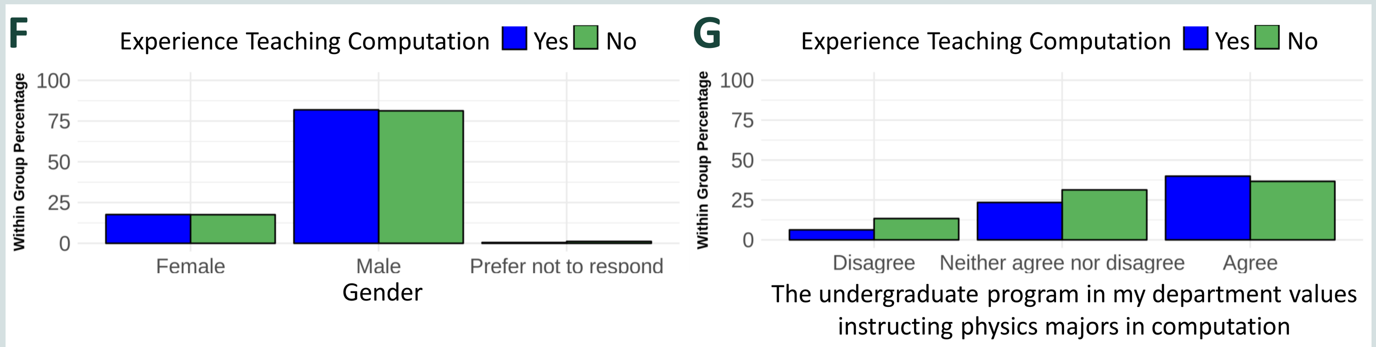
Are the important factors sensible?



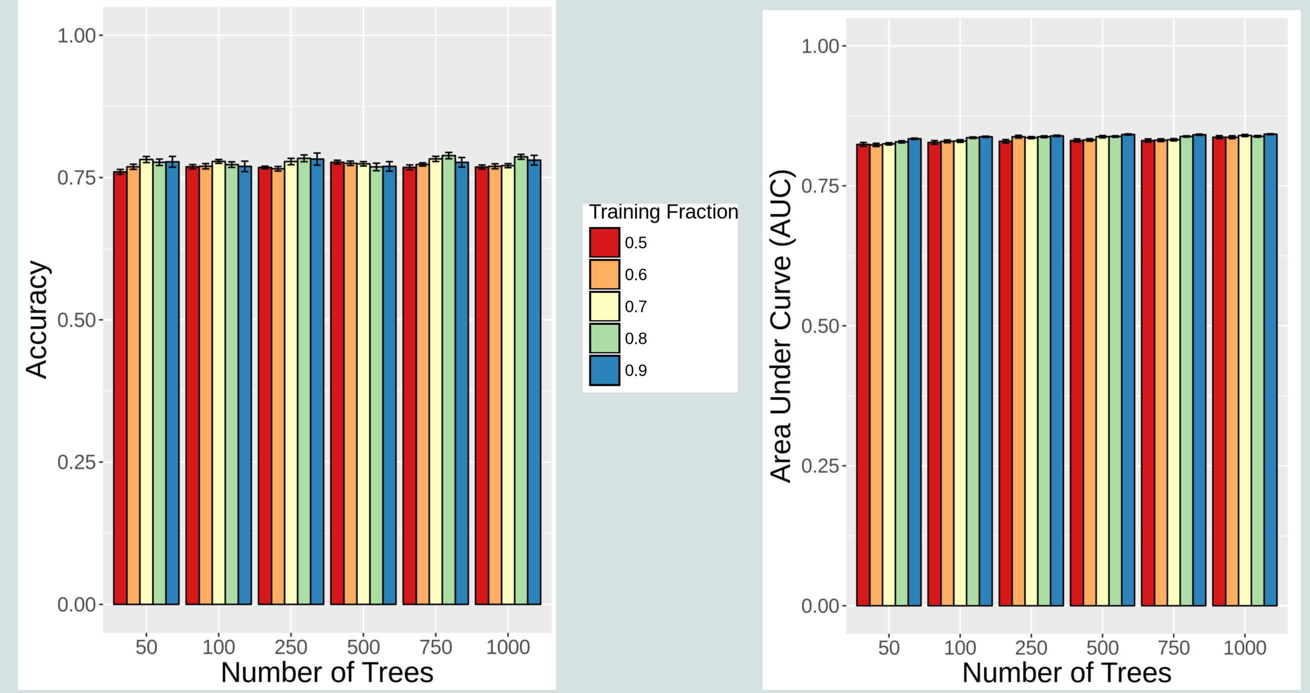


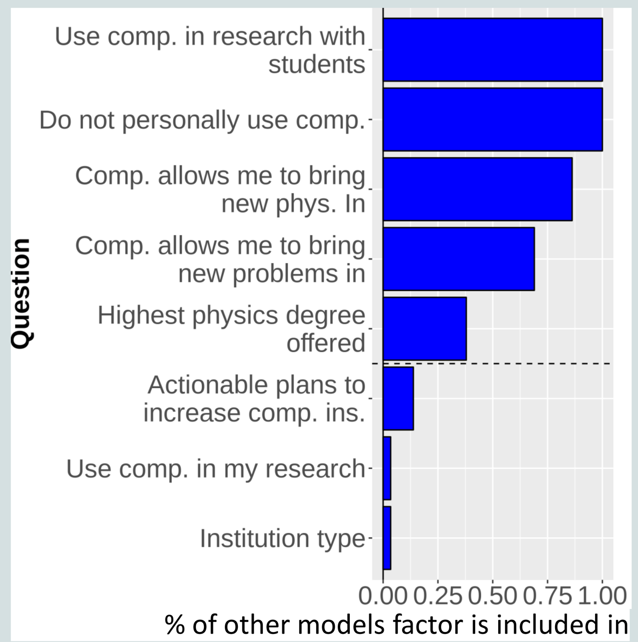
Plots A-E show differences in distributions of the important factors for faculty who do and do not have experience teaching computation.

Plots F and G show no such differences



Does the model change if the parameters change?





Accuracy and AUC are approximately the same regardless of the number of trees or training fraction

Some variation in selected factors, but same factors are selected again and again.

Discussion & Conclusions

Faculty that teach computation tend to:
- Use computation in their research with students or some other way outside of the classroom
- Believe computation brings new physics and problems into the curriculum
- Teach at institutions that offer up to a physics bachelor's degree

Factors that do not appear to be predictive:
- Demographic factors
- How computation is viewed by department

Conclusion: Faculty treat teaching computation as an individual choice

Comments on Random Forest model:
- Unbalanced classes may produce low accuracy value
- Selected variables do show differences in distributions between those who do and do not have experience teaching computation.
- Model appears stable against variations in the parameters such as size of the forest and fraction of data used for training

Impacts: Useful for groups like PICUP working to increase use of computational instruction.

Acknowledgements

This work was generously supported by the National Science Foundation (DUE- 1431776). Thanks to Laura Merner, Norman Chonacky, and Robert Hilborn for survey development and deployment. Thanks to John Aiken, Michelle Kuchera, and Tim McKay for their useful insights on the analysis.

References

[1] Chabay, Ruth, and Bruce Sherwood. "Computational physics in the introductory calculus-based course." American Journal of Physics 76.4 (2008): 307-313.
[2] AAPT Undergraduate Curriculum Task Force, AAPT Recommendations for Computational Physics in the Undergraduate Physics Curriculum, 2016
[3] Caballero, Marcos D., and Steven J. Pollock. "A model for incorporating computation without changing the course: An example from middle-division classical mechanics." American Journal of Physics 82.3 (2014): 231-237.
[4] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
[5] Strobl, Carolin, et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution." BMC Bioinformatics 8.1 (2007): 25.
[6] Rokach, Lior, and Oded Maimon. Data mining with decision trees: theory and applications. World scientific, 2014.
[7] Janitza et al. "An AUC-based permutation variable importance measure for random forests." BMC Bioinformatics 14:119 (2013)
[8] Diaz-Uriarte, Ramon and Sara Alvarez de Andres. "Gene selection and classification of microarray data using random forest." BMC informatics 7:3 (2006).