Harvard Data Science Review • Issue 2.3, Summer 2020

Prediction, Machine Learning, and Individual Lives: An Interview With Matthew Salganik

Matthew J. Salganik[1], Lauren Maffeo[2], Cynthia Rudin[3]

[1]Department of Sociology, Princeton University, Princeton, New Jersey, United States of America,
[2]Steampunk Inc., District of Columbia, United States of America; OpenSource.com, Red Hat, Raleigh, North Carolina, United States of America,
[3]Department of Computer Science, Trinity College of Arts and Sciences, Duke University, Durham, North Carolina, United States of America; Department of Electrical and Computer Engineering, Pratt School of Engineering, Duke University, Durham, North Carolina, United States of America

Published on: Jul 30, 2020

DOI: 10.1162/99608f92.eecdfa4e

ABSTRACT

Machine learning techniques are increasingly used throughout society to predict individual's life outcomes. However, research published in the *Proceedings of the National Academy of Sciences* raises questions about the accuracy of these predictions. Led by researchers at Princeton University, this mass collaboration involved 160 teams of data and social scientists building statistical and machine learning models to predict six life outcomes for children, parents, and families. They found that none of the teams could make very accurate predictions, despite using advanced techniques and having access to a rich dataset. This interview of Matthew Salganik, the study's lead author and a professor of Sociology at Princeton University, was conducted by Lauren Maffeo, Associate Principal Analyst at Gartner, and Cynthia Rudin, a professor of Computer Science, Electrical and Computer Engineering, and Statistical Science at Duke University. It provides an overview of the study's goals, research methods, and results. The interview also includes key takeaways for policy leaders who wish to use machine learning to predict and improve life outcomes for people.

Keywords: machine learning, data science, social science, predictive modeling, policy, interviews, life course, mass collaboration

---

Lauren Maffeo (LM): Matthew, what is the Fragile Families Challenge and what were you trying to accomplish?

Matthew Salganik (MS): The Fragile Families Challenge is a scientific mass collaboration designed to answer one question: how predictable are life trajectories? That is, given some data about a person, how accurately can we predict what will happen to that person in the future?

Cynthia Rudin (CR): How did you attempt to quantify that?

MS: We measured the predictability of life outcomes by combining a high-quality dataset from the social sciences, a research design from machine learning, and 457

researchers from around the world (Salganik et al., 2020a).

More specifically, the Fragile Families Challenge builds on the Fragile Families and Child Wellbeing Study. This ongoing study is designed to understand families formed by unmarried parents and the lives of children born into these families. It collects rich longitudinal data about thousands of families who gave birth to children in large U.S. cities around the year 2000 (Reichman et al., 2001). The Fragile Families data—which have been used in more than 750 published journal articles—were collected in six waves: birth, child ages 1, 3, 5, 9, and 15. Each wave includes a number of different data collection modules. For example, the first wave (birth) includes survey interviews with the mother and father. Over time, the scope of data collection increased, and the fifth wave (age 9) includes survey interviews with the mother, father, child, and child's teacher.

Each data collection module is made up of about 10 sections. For example, the interview with the mother in wave one (birth) has sections about topics including child health and development, father-mother relationships, demographic characteristics, and education and employment. The interview with the child in wave five (age 9) has questions about topics including parental supervision and relationship, parental discipline, and school.

In addition to the surveys, at waves three, four, and five (ages 3, 5, and 9), interviewers traveled to the child's home to conduct in-home assessments that included psychometric testing, such as the Peabody Picture Vocabulary Test; biometric measurements, such as height, weight; and observations of the neighborhood and home. All together, the Fragile Families data includes thousands of variables about factors that researchers think are important predictors of child well-being.

To measure the predictability of life outcomes using these data, my co-organizers—Ian Lundeberg, Alex Kindel, Sara McLanahan—and I used a research design from machine learning called the common task method. The common task method is well-known to machine learning researchers—David Donoho (2017) even called it machine learning's 'secret sauce'—and it has been popularized by projects such as the Netflix Prize (Feuerverger et al., 2012).

The common task method requires many researchers to build predictive models with the exact same training data and then evaluate their predictions on the exact same holdout data using the exact same error metric. The standardization created by the common task method ensures that all the predictions can be compared fairly.

To set up a project using the common task method we picked six outcomes—such as the child's GPA and whether the family would be evicted from their home—and then recruited researchers from around the world to participate. During the Fragile Families Challenge, participants attempted to predict these six outcomes, which were measured in wave six (age 15), using data from waves one to five (birth to age 9).

CR: Why did you have a mass collaboration using the common task method instead of just doing it yourself?

MS: Although it is not typically used in the social sciences, the common task method is great for producing credible estimates of predictability. If predictability is higher than you expect, it can't be explained away by over-fitting, or by researchers selecting a metric that makes their performance look good. Alternatively, if predictability is lower than you expect, it can't be explained away by the failures of any particular researcher or technique.

For example, if I tried to predict these six outcomes and was unsuccessful, then people might say, "Well, maybe Matt is not very good at machine learning, or maybe he didn't try technique X or technique Y." But if hundreds of researchers try and *none* of them can predict very accurately, then maybe it is just not possible.

LM: What were the six life outcomes that researchers tried to predict?

MS: Participants were trying to predict the GPA of the child, the grit of the child, the material hardship of the household, whether the household was evicted, whether the parent was laid off of work, and whether the parent participated in a job training program. All these outcomes were measured when the child was 15 years old. We picked these six because they were interesting to domain experts, and we wanted some outcomes about the child, some about the parent, and some about the household. We did not pick these outcomes because we thought that they would be especially easy or hard to predict.

LM: Some of those, such as grit and material hardship, seem hard to quantify.

MS: That's true. Social scientists often have to measure hard-to-quantify concepts. In this case, all of our outcome data is self-reported in a survey. Grit, which is defined as a personality trait that combines passion and perseverance, is measured with a series of four survey questions. These questions are based on the grit scale originally proposed in Duckworth et al. (2007). Material hardship is measured with a series of 11 questions that focus on the experience of poverty, such as whether someone in the household could not afford to eat or visit the doctor. These questions were originally proposed in Mayer and Jencks (1989). The full details about how we measured each outcome are in Table S3 of our paper.

The definitions that we used, which came from scientific literature, might not perfectly map onto the outcomes of interest to policy makers. More generally, whenever you are trying to understand the performance of a predictive model, it is important to understand exactly how the outcomes were defined and measured.

LM: What do you think is the main finding of the Fragile Families Challenge?

MS: Despite having access to a rich dataset, and despite using modern machine learning methods that are optimized for prediction, none of the 160 teams were able to make very accurate predictions. The most accurate models had R^2 in the holdout data of about 0.2 for material hardship and GPA, and close to 0 for the other four outcomes.

To put these numbers in context, a model with R^2 of 1 would be perfectly accurate and a model with R^2 of 0 would be no more accurate than predicting the mean of the training data. So, one way to think about it is that the most accurate models were generally not much better than naive guessing.

I still vividly remember the first time we saw these results. I was really, really surprised. I had expected that the machine learning techniques, when combined with a very extensive dataset, would lead to more accurate predictions. I've had a lot of time to think about these results, and I'm still not sure how to make sense of them.

LM: Which specific machine learning techniques did the teams use to build predictive models?

MS: There were 160 teams that made valid submissions to the Challenge, and they used a huge variety of techniques for data preparation, feature selection, and statistical learning. Some examples of the machine learning techniques include: neural networks (Davidson, 2019), BART (Carnegie and Wu, 2019), LASSO (Stanescu et al. 2019), support vector regression (Roberts, 2019), and ensembles of various models (Rigobon et al. 2019). Beyond these specific examples, we have a sense of what all teams did because they open sourced their code at the end of the Challenge (Salganik et al. 2020b), and participants that were co-authors described their approach in our paper. So, if you are thinking, "Maybe Method X would lead to accurate predictions," I'm pretty sure that someone tried Method X.

Also, one other thing that was very surprising to me is that—to a first approximation—there was very little difference between the predictions made by the different approaches. Let's consider predicting GPA for a moment. There were some kids that were well predicted by basically all the teams, and there were other kids that were poorly predicted by basically all the teams. In other words, there seemed to be much more variability in kids than there was in approaches.

This consistency across approaches leads me to speculate that no approach is going to get substantially better predictive performance with these data and this task. Of course, I'd be happy to be proven wrong, and that's one great thing about the common task method. It creates a clear standard. If future researchers can beat that standard, that is a strong sign that they are doing something different and potentially interesting.

CR: Complicated ML models can be really helpful for computer vision, because of the structure of that kind of data. But the type of data in the Fragile Families Challenge doesn't seem likely to have that same structure. So, in that sense, the results might not surprise some ML researchers. But have these results been surprising to others?

MS: As I said, they were surprising to me. Before I did this project—and saw these results—I guess I must have thought that machine learning was somehow magic. I don't think that anymore.

The results are definitely surprising to some other people too. When I give talks about the Fragile Families Challenge, before I show the results I have people raise their hands to show what they expect. In my experience, machine learning/data science audiences often have higher expectations for predictive accuracy than social science audiences. Policy makers often have very high expectations, too.

CR: Were there any other results that surprised you?

MS: Yes, one more—but, Cynthia, you won't find this result surprising. We found that the most accurate models were only slightly better than a simple, four-variable regression model, where the four variables were chosen by a domain expert. When I look back at all the data that were collected about these kids and their social environments since birth, it is still hard for me to imagine that there is not a lot more information—in the predictive sense—than just the four variables in a linear model.

CR: I think the big question you bring up during your paper's discussion section is the study's policy implications. It has been shown now in several papers by various researchers (including both of us!) that simple, interpretable ML models are often as good as the best ML models (Rudin, 2019). Yet policy is being driven further and further towards black box ML methods. Why do you think that is?

MS: I don't know for sure, but my guess is that some people think that complex, black box methods will lead to better predictions. I certainly did. In some cases, like computer vision, black box methods do perform better. Conversely, in other cases, they don't. I think that policy makers don't have a good sense of when the complexity will be worth the cost, and I hope that as researchers, we can help develop some theory or empirical results that can provide guidance.

CR: You mentioned that policy makers should be concerned by the results, because essentially they imply that society is paying a cost for creating, testing, and trying to understand complicated models when they are not needed. Can you elaborate on how policy makers can be convinced to consider these costs?

MS: Just to make this a bit clearer, there was one team that submitted the winning models for three outcomes: GPA, grit, and layoff. They wrote about their approach in Rigobon et al. (2019), which is part of a special issue of *Socius* about the Fragile Families Challenge (Salganik et al., 2019). If you read that paper, it is basically impossible to understand how that model works; it would be very difficult to verify that the code is actually doing what the paper describes; and it would be very difficult to audit their approach to ensure that it didn't unintentionally discriminate against people in certain groups. To be clear, that's not a criticism of their paper. You could say something very similar about most of the papers in the special issue.

In fact, for many of the papers in the special issue, we struggled with even basic computational reproducibility. That is, we often could not reproduce the results in the paper, even when the authors gave us their code and data (Liu and Salganik, 2019). This is really troubling because basic computational reproducibility seems like a necessary—but certainly not sufficient—criteria for a model used for high-stakes decisions about people.

All of this is to say that the costs of creating, testing, understanding, and auditing complex models are real, and my sense is that policy makers will begin to notice these costs as they get more familiar with these models.

CR: Following up on the last question, there is so much lobbying by companies trying to sell black box models that it may be harder to reach policy makers with this important message. How can we (as scientists) reach them before the sales teams that these companies have?

MS: I think that we can—and should—communicate our findings to a wider audience, as you and others have done. Also, the fact that there are financial stakes can actually be helpful. I'm sure that there are some bad companies with incentives to obscure their model's performance with slick sales tactics, but there are probably also some responsible companies that have an incentive to make the process more clear and transparent.

The financial stakes also give governments leverage through procurement systems. For example, a government could say it will only buy systems that have been audited for fairness or that have been evaluated in a common task method setting. My hope is that over time governments will get better at assessing how useful these systems are before purchase, and they will get better at evaluating them during deployment.

LM: The dataset in the Fragile Families Challenge contained about 13,000 variables about 4000 families. How do you think the size of the dataset impacted the accuracy of the models?

MS: I don't know for sure how accuracy is related to size, but it is important to think about the size of the dataset both in terms of the number of variables and the number of families.

Starting with the number of variables, going from a simple, 4-variable model to a complex model that could draw on 13,000 variables led to a small increase in predictive performance. So, I doubt that there is some magical 13,001st variable that would make a huge difference. Also, I'm not sure that adding another 13,000 variables about each family will make a big difference in terms of predictive performance.

As far as adding more families, I'm not sure. Within our group, we've talked about whether 4000 is a lot or a little. I've said to Sara, who has been a PI [Principal Investigator] of the Fragile Families Challenge since it started in the late 1990s, "The Challenge *only* involved 4000 families." And she said, "Matt, that's *4000* families." You have to realize how hard it is to stay engaged with this many people over such a long time.

Since we can't increase the number of families in this dataset by a factor of 10 or 100, these kinds of questions are going to have to be addressed in future research with

other data. With this number of families, social science theory and domain expertise may be especially important, but if we had a million times more families, who knows?

CR: In the paper, you describe a tension between *understanding (*meaning sociological theory, perhaps) and *prediction*. Ideally, good sociological theory can substitute for low-quality data or a lack of data. Do you think the results of your study can inform future hypotheses about theory that would later be tested? Or do you think the data could support so many theories equally well that it would be difficult to find good ones to test?

MS: One of things that was most exciting to me about this project was trying to move beyond the tension between understanding and prediction, and come up with ways that prediction could help us get to understanding. That is, I personally didn't really see prediction as an end in itself, but I thought that it could be a good way to get understanding. Unfortunately, using prediction to get understanding has turned out to be harder than I expected, at least so far.

For example, although there are 13,000 variables in the Fragile Families dataset, only a relatively small number of them are frequently used in research. I was hoping that the Challenge might uncover some less-studied variables that turned out to be important for prediction. We could then do some subsequent research to figure out what is going on with these variables. Unfortunately, no surprising predictive variables have clearly emerged from this work yet.

However, I think the main results of the Challenge raise an important question for sociological theory: why are life outcomes unpredictable even when using what we would consider to be high-quality data and modern machine learning? One possibility is that there are other important social processes operating in the world that we are not currently theorizing or measuring. Finding those would be exciting, and we are already starting to use our results from the Challenge to help us look for them. For example, we have used the results of the Challenge to identify kids who are doing much better or worse than predicted by the best model. Then, we've conducted in-depth interviews with these kids and their parents to try to understand what is enabling some to beat the odds and what is leading some to struggle unexpectedly. By

understanding the lives of these hard-to-predict kids better, we hope to discover other important social processes that shape life outcomes.

Also, it could be that we need theories that lead us to expect unpredictability. In weather, for example, we don't expect perfectly accurate predictions far into the future because of the nature of weather. Likewise, we don't expect perfectly accurate predictions of stock prices, because that's not the way financial markets work. Maybe we should not expect perfectly accurate predictions of individual life trajectories because of the way that social systems work. To me, the Challenge raised this possibility as a big, open question. It shows how results about predictability can inspire us to develop new sociological theory.

CR: Switching from the theoretical to the very empirical, I know you said the dataset was really clean, but I admit that I have never seen a perfectly clean big dataset in my life. Is it really clean?

MS: Going into the Challenge, I thought that the dataset was clean. What I discovered, however, is that the whole idea of 'clean' is complicated.

In one sense, the dataset was clean. It was created by researchers to be used for research, so it's not administrative data or digital trace data that was created for some other purpose (Salganik, 2018). The families were selected based on a probability sampling design, the questions in the surveys were created to measure constructs that social scientists think are important, a lot of effort went into collecting the data, and there were hundreds of pages of documentation and codebooks. As I said earlier, the data had already been used in more than 750 published papers.

However, what we discovered during the Challenge was that many participants—particularly data scientists—thought the dataset was not clean. The dataset certainly was not one big rectangular matrix that was ready to go. Rather, many steps were needed to prepare the data for statistical learning. For example, there were different codes for different types of missing data. Participants needed to take this issue—and others—into account as they prepared the dataset for statistical learning.

Social scientists seemed to struggle less with this data preparation. In part, I think this is because social scientists are just more familiar with this type of dataset, but I also

think there is a more fundamental reason too. The data preparation needed for social science-style modeling is different than what is needed for data science-style modeling.

Many social scientists in the Challenge started with empty models and then added variables to them. They would prepare these variables—handling missing data, converting unordered categorical variables, such as race, into 0/1 variables, etc.—one at a time as they were added. On the other hand, many data scientists wanted to start with models that included all the variables and then use some kind of automated feature selection. This approach required them to prepare all the variables, and that turned out to be really hard. For example, some participants in the Challenge asked for a list of the variables that were unordered categorical variables, so that they could convert these into 0/1 variables. Not one of the hundreds of social scientists that had used the data before had ever asked for that. For a small number of variables, it is easy to manually decide which are unordered categorical variables, but it becomes basically impossible if you do it for 13,000 variables.

Fortunately, one of the participants in the Challenge—Greg Gundersen, who is a computer science graduate student at Princeton—came up with the idea of converting all the Fragile Families study codebooks into a machine-readable metadata API. With Greg's blessing, we took what he did during the Challenge and improved it so that future researchers can access the Fragile Families metadata in a machine-readable format. We even wrote a paper about our experience because we think it offers some important lessons to other social scientists that want to make their data more amenable to machine learning methods (Kindel et al., 2019).

CR: Do you have any last advice for policymakers considering using predictive models in settings like criminal justice and child protective services?

MS: For policymakers, the most important lesson from this project is that machine learning is not magic. If someone builds a complicated model using lots of data, that does not guarantee accurate predictions. In fact, it will take careful empirical work to figure out whether the predictive model will work well in your setting, and it will take careful thinking to figure out whether predictive models are the best way to advance your policy objectives.

LM: Which are the most important scientific questions that your research left unanswered?

MS: That's a hard question, but it is also a great question to end on because it highlights that mass collaborations using the common task method can be the first step in a larger research agenda.

We've already touched on a few of the unanswered questions that I find most interesting: is it possible to get much better predictive performance with these data for this task? How would the results be different if we used other datasets and other life outcomes? Why were the outcomes so hard to predict? In other words, are there important aspects of the life course that we just don't understand right now?

Beyond those specific questions, the Challenge also raises a more general question about how we work as scientists. I hope it gets people to explore other research questions that we can tackle collectively through a mass collaboration, but that none of us could address individually.

LM: Thank you so much for speaking with us. It's been fascinating to learn about this research.

Disclosure Statement

Matthew J. Salganik, Lauren Maffeo, and Cynthia Rudin have no financial or non-financial disclosures to share for this interview.

References

Carnegie, N. B., & Wu, J. (2019). Variable selection and parameter tuning for BART modeling in the fragile families challenge. *Socius, 5*, 1-10. https://doi.org/10.1177/2378023119825886

Davidson, T. (2019). Black-box models and sociological explanations: Predicting high school grade point average using neural networks. *Socius, 5,* 1–11. https://doi.org/10.1177/2378023118817702

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Feuerverger, A., He, Y., & Khatri, S. (2012). Statistical significance of the Netflix challenge. *Statistical Science*, *27*(2), 202–231. https://doi.org/10.1214/11-STS368

Kindel, A. T., Bansal, V., Catena, K. D., Hartshorne, T. H., Jaeger, K., Koffman, D., McLanahan, S., Phillips, M., Rouhani, S., Vinh, S., & Salganik, M. J. (2019). Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge. *Socius*, *5*, 1–24. https://doi.org/10.1177/2378023118817378

Liu, D. & Salganik, M.J. (2019). Successes and struggles with computational reproducibility: Lessons from the Fragile Families Challenge. *Socius*, *5*, 1–21. https://doi.org/10.1177/2378023119849803

Mayer, S., & Jencks, C. (1989). Poverty and the distribution of material hardship. *Journal of Human Resources, 24*(1), 88–114. https://doi.org/10.2307/145934

Reichman, N. E., Teitler, J. O., Garfinkel, I., & McLanahan, S. S. (1991). Fragile families: Sample and design. *Children and Youth Services Review, 23*(4–5), 303–326. https://doi.org/10.1016/S0190-7409(01)00141-4

Rigobon, D. E., Jahani, E., Suhara, Y., AlGhoneim, K., Alghunaim, A., Pentland, A., & Almaatouq, A. (2019). Winning models for grade point average, grit, and layoff in the Fragile Families Challenge. *Socius*, *5*, 1–10. https://doi.org/10.1177/2378023118820418

Roberts, C.V. (2019). Friend request pending: A comparative assessment of engineering- and social science–inspired approaches to analyzing complex birth cohort survey data. *Socius*, *5*, 1–8. https://doi.org/10.1177/2378023118820431

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Salganik, M.J. (2018). *Bit by bit: Social research in the digital age*. Princeton University Press: Princeton, NJ.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., . . . McLanahan, S. (2020a). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences, 117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Introduction to the special collection on the Fragile Families Challenge. *Socius*, *5*, 1–21, https://doi.org/10.1177/2378023119871580

Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2020b). Replication materials for 'Measuring the predictability of life outcomes using a scientific mass collaboration.' *Harvard Dataverse*. https://doi.org/10.7910/DVN/CXSECU.

Stanescu, D., Wang, E., & Yamauchi, S. (2019). Using LASSO to assist imputation and predict child well-being. *Socius, 5,* 1–21. https://doi.org/10.1177/2378023118814623

The preview image of this interview was created by Egan Jimenez, Princeton School of Public and International Affairs.