Can Statistics Predict the Fields Medal Winners?

Adam Bannister and Olaf Teschke (FIZ Karlsruhe, Berlin)

With the upcoming ICM in Rio de Janeiro, the seasonal speculation about who will receive the 2018 Fields Medals at the opening ceremony is once more in full swing. With the big data industry measuring us in all possible ways, a natural question might be whether statistical approaches could possibly predict the committee's choice of the Fields Medallists. We undertake some experiments here to see which predictions are provided by standard approaches based on data from zbMATH and linked data sources.

Fields Medal is small data

One obvious obstacle, however, is that the set of Fields Medallists is small by its very nature and may easily defy statistics with all kinds of outliers. As a nice example, one could recommend reading Borjas-Doran's study on a statistical decline of Fields Medallists' productivity [BD14] and Kollár's amusing review [K15]. Without further discussing the fundamental problem of measuring a mathematician's productivity by publication and citation numbers – the fallacies of this approach have been frequently discussed in the newsletter, for example in [BT17] – we just note here the very last observation in [K15]:

> *"The limits of statistics are illustrated by the numbers contained in the penultimate line of [BD14, Table 1]. (It is not commented on in the paper.) While most of the Fields Medallists and contenders are happily alive, Figure 3 shows a disturbing pattern about those who have passed away […Namely, an average age of death of 74.0 for Fields Medallists compared to only 66.3 for contenders…] Thus, if you got a Fields Medal, you can expect to enjoy your extra US$120,000 per year for almost eight more years."*

Firstly, we may take this as an illustration of how seemingly exact science is often perturbed by possibly unreliable data. It was, for us, impossible to reproduce the average age of death of 74 from [BK14] ([K15] does not comment on this figure). Submitted in 2014 before the death of Grothendieck, the nine Fields Medallists deceased at that time reached an average age of 78.5. (A closer look at the appendix of [BK14] reveals that the 1936 medallists Ahlfors and Douglas seem to have been excluded from the study but that has almost no effect on this average). Secondly, as we are all sadly aware, this figure has been significantly affected since then by the passing of several medallists, reducing the average age at death by more than four years and wiping out a large part of this statistical effect.

Which data can reasonably be taken into account?

There is basically no formal limitation to being a Fields Medal candidate except for the famous age rule. However, even this simple condition requires some work – for most of the (approximately) one million authors in zbMATH, the age is simply unknown. Reasonable estimates are often possible based on publication history but in many cases may lead to gross errors: while quite a few mathematicians have already published relevant research in their teens, others may be over 30 at their first zbMATH entry. This happens frequently in border areas when most publications are outside the scope of the database or for mathematicians suffering from political suppression, e.g. from the Nazi regime, Stalinist terror or during the Cultural Revolution in China.

Needless to say, birth dates derived from publication history are therefore not suitable to check the rather clear-cut Fields Medal age limit. Fortunately, the zbMATH author database is linked to many other collections, some of which – like Wikipedia, MacTutor, GND and MathNet.Ru – provide birth date information. Additionally, the communities pursuing these services usually do a reliable job in relevance decisions. Overall, zbMATH contains links to data resources providing sufficient age information for almost 13,000 authors – only a fraction of the overall million but covering all Fields Medallists and likely candidates. Only 252 of them are at most 40 years old.

The harder task is to identify relevant features for the statistical model. Modern databases like zbMATH offer various facets that can be taken into account – not just quantitative publication and citation information but also granular subject information, or co-author and reference networks. Unfortunately, many of these quantities can be dual-edged: a high publication number may be obtained by people like Yau, Bourgain or Tao, or by more notorious representatives of the class of prolific writers.[1] High citation numbers may be related to a lasting impact of results but they are also much affected by subject and community custom. Even more importantly, they come with a massive time delay [BT17]. Subject information is certainly valuable (statistically, the Fields Medals are far from evenly distributed within MSC subjects) but will usually not reflect breakthroughs that may define new areas in the future. Publication sources are certainly meaningful – prize winners will almost inevitably have a distinctive record in the Annals, Inventiones, etc. – but are significantly limited by publication delay, with many relevant results appearing only after several years (the committee of course being aware of them). Close collaboration or citation distance to former prize winners may indicate that you are actively involved in pursuing cutting-edge research but could also be an indication of a supporting role rather than a unique individual effort qualifying for the medal.

Less ambiguous features would be existing prizes like the EMS Prizes (which are also connected to an age limit and have a distinct overlap with later Fields Medallists from Europe) but many prizes have a shorter history than the Fields Medal, as well as regional restrictions, thereby further complicating the involved statistics. Fortunately, a substantial list of prize winners is available for analysis via the zbMATH connection with Wikidata; others (like the EMS prizes) have been added manually. The same holds for the information on being an invited speaker to an ICM, which may reasonably be treated like winning a global prize.

Finally, we emphasise that no data generated by user searches were taken into account due to our strict data protection policy [HT14]. As outlined there, one could expect rather distinctive results, especially if IP information were analysed (which is ruled out). A rough approximation might be obtained by taking Google search data into account (although this would most likely not reflect the committee's procedure well). Currently, this would see Simon Brendle, Hugo Duminil-Copin, Alessio Figalli, Ciprian Manolescu, Fernando Codá Marques, Sophie Morel, Peter Scholze, Maryna Viazovska and Geordie Williamson as the most likely candidates (in alphabetic order, with Peter Scholze leading).[2] A closer look at the trends indicates that most of the queries are correlated to prize announcements, hence one might expect that this is covered by the above features.

Methodology

Educated humans will usually overcome most of these obstacles, e.g. a closer look will easily distinguish deep results from superficial mass publications with bulk references. Automatic recognition is, however, still limited in addressing such questions. Approaches like neural networks have made tremendous progress over the past years but still encounter problems, for example in distinguishing art from pornography (a somehow related question), despite the fact that technology in image processing has become more advanced and much more data are available. Some tools to recognise "maths pornography" might help editors, reviewers and readers but there has not been much activity toward this yet. Moreover, big data approaches would ideally require billions of samples as training data, far more than the currently available mathematical publications (although several groups of authors, in an often undervalued effort, are very active in enlarging the available datasets). The problem of scarce data applies not just to bibliometrics but even more to the other features mentioned, so there currently seems no hope of applying neural network technology to the Fields Medal prediction.

[1] Currently, Yau, the Fields Medallist with the most publications in the zbMATH database (authoring on average a paper every two weeks over the past few years), ranks only around 50th place in this list.

[2] This kind of crowd-sourced projection also agrees well with certain internet polls, e.g. https://poll.pollcode.com/44839318_result?v.

Instead, we just put the available data into a support vector machine model. We defined data slices for the information available at the time of the congress. First, we trained a model based on previous years' winners. We then used this model to analyse our candidates for this year. We repeated this procedure 20 times and averaged the results to remove any outliers created by an imbalance in the splitting of testing and training data. This averaging of the results is due purely to the small sample of data available to train a model; in some runs, we could be unlucky enough to get no positive examples in our training set.

As we have used prizes first awarded in the 1990s, we also had to limit our Fields years to 1994 or later. We took into consideration the EMS Prizes, the Bôcher Memorial Prize, the Coxeter-James Prize, the Fermat Prize, the SASTRA Ramanujan Prize, the Oswald Veblen Prize, the Clay Research Award, the Wolf Prize and the Salem Prize.

As a by-product, we obtained measures for the significance of the different features.

Results

Not unexpectedly, sole bibliometric features turn out to be almost non-predictive. In a model that takes just citation figures, journals or MSC subjects into account (the latter two features should be at least included to adjust citations numbers [BT17]), one can generate a high-dimensional (due to the variety of journals medallists have published in) linear model that is adjusted to the past but generates only individual winning probabilities of about 1% and less in the projection (with Jeremy Blanc, Anton Koroshkin, Luis Pedro Montejano and Evgeny Sevostyanov leading by slim digits a basically even field). Hence, citation-based hiring will most likely lead to missing a future Fields Medallist (actually, it will perform at most marginally better than randomly picking a mathematician younger than 40 years with a Wikipedia entry).

In contrast, other prizes and ICM invitations are the most predictive sole features, which produce distinctive projections and were more than 97% successful for past test sets. By adding further features like collaboration and citation distance to former Fields Medallists, prize winners and invited speakers, the success rate for test sets can be improved further (as is natural when dimensions are added) up to greater than 99.3% but with decreasing sharpness of prediction. The differences also indicate a possible bias toward more collaborative communities, after adding the distance features, and a bias against recent, yet unpublished achievements. Table 1 shows the figures for the leading contenders in the respective models.[3]

From this, one might reasonably predict that Peter Scholze is a strong favourite to win a Fields Medal but the others remain highly competitive, with different models producing very different outcomes. Geordie Williamson and Bo'az Klartag seem to have the most consistent statistical chances from the field.

Does the committee's composition matter?

Of course, the decision is solely made by the committee members, whom we can expect to weight mathematical achievement over superficial facets. Since assuming responsibility for the committee, the IMU has put much effort into creating a balanced composition of prize committee with respect to aspects like geography or research area, and the difficulty of obtaining significant projections may serve as a good illustration. Of course, the composition of the 2018 committee cannot be used for projections since it is revealed only at the ICM (except for IMU president Shigefumi Mori, who is an ex-officio

[3] Important caveat: Since we didn't use 4-years age slices of in the model to avoid more sparsity effects, the resulting probability reflects the chance of winning a Fields Medal in the future, not necessarily at the next Congress.

Table 1. Projected winners in different models

	Prize	Prize + Invitation	Prize + Invitation + Coauthor	Prize + Invitation + Coauthor + Citation	All features
Peter Scholze	64%	81%	91%	34%	86%
Geordie Williamson	56%	82%	25%	10%	2%
Bo'az Klartag	50%	38%	15%	14%	16%
Simon Brendle	49%	30%	2%	<1%	1%
Hugo Duminil-Copin	5%	6%	14%	11%	20%
Peter Pal Varju	5%	6%	11%	2%	16%
Sophie Morel	6%	6%	9%	<1%	6%
Alessio Figalli	5%	6%	9%	4%	2%
Ciprian Manolescu	5%	6%	9%	3%	<1%
Maryna Viazovska	1%	1%	9%	1%	<1%
Fernando Coda Marques	<1%	2%	1%	<1%	3%

member). However, one may ask whether the knowledge of the composition of past committees[4] would have significantly influenced the projections. The figures show only modest changes when adding the committee information, hence a significant "committee bias" cannot be confirmed via this statistical approach.

Conclusions

There are several facets of public information available that may serve as features for statistical predictions about Fields Medal winners but many come with certain disadvantages. Taking different reasonable models into account, the formal statistical approach may provide some educated guesses with reasonable probabilities but a rather high uncertainty remains, certainly sufficient to keep the tension about the disclosure of the winners at the ICM.

Perhaps the most important caveat is, however, that the statistical method will only succeed in carrying forward past trends to the future. As is well known, this is one of its major drawbacks, which may preserve or even worsen existing discriminations [O16]. Due to these effects, we didn't include available data features like gender or country of origin into the model because this would almost certainly generate further intrinsic bias. Since the composition of the Fields Medallists has grown significantly more diverse over the past few years (reflecting the development of the mathematical community), statistical predictions will most likely have a conservative bias compared to the actual decisions and the committee will likely succeed in proving statistical guesses at least partially wrong.

References

[BT17] A. Bannister, O. Teschke, An update on time lag in mathematical references, preprint relevance, and subject specifics. *Eur. Math. Soc. Newsl.* 106, 37–39 (2017; Zbl 06853068).

[BK14] G.J. Borjas, K.B. Doran, Prizes and Productivity: How Winning the Fields Medal Affects Scientific Output, *J. Human Res*. 50, No. 3, 728–758 (2015).

[K15] J. Kollár, Is there a curse of the Fields Medal? *Notices Am. Math. Soc*. 62, No. 1, 21–25 (2015; Zbl 1338.01086).

[O16] C. O'Neil, *Weapons of math destruction*. New York, NY: Crown Random House (2016; Zbl 06801031).

[HT14] J. Holzkämper, O. Teschke, Guarding your searches: data protection at zbMATH. *Eur. Math. Soc. Newsl*. 92, 54–55 (2014; Zbl 1302.68110).

See page 4 for a photo and biography of the authors.

[4] This information is available at the IMU site; the authors like to thank the MathOverlow community for clarifying a question related to the 1962 committee.