NBER WORKING PAPER SERIES

CROWDSOURCING CITY GOVERNMENT: USING TOURNAMENTS TO IMPROVE INSPECTION ACCURACY

Edward L. Glaeser
Andrew Hillis
Scott Duke Kominers
Michael Luca

Working Paper 22124
http://www.nber.org/papers/w22124

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2016

The authors are deeply grateful for the collaboration of the City of Boston (especially Ben Batorsky, Matthew Mayrl, and Commissioner William Christopher), Yelp (Artem Avdacev, Luther Lowe, and Aaron Schur), and DrivenData (Peter Bull and Greg Lipstein), without whom the project described herein would not have been possible. Additionally, the authors gratefully acknowledge the helpful comments of Benjamin Edelman, Anthony Goldbloom, Mitchell Weiss, and especially Susan Athey, as well as the support of Yelp, the National Science Foundation (grants CCF-1216095, DGE-1144152, and SES-1459912), the Harvard Milton Fund, the Taubman Center for State and Local Government, the Rappaport Institute for Greater Boston, and the Wu Fund for Big Data Analysis. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w22124.ack

Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy
Edward L. Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca
NBER Working Paper No. 22124
March 2016
JEL No. C53,D04,D47,D8,L88,M50,R5

ABSTRACT

Can open tournaments improve the quality of city services? The proliferation of big data makes it possible to use predictive analytics to better target services like hygiene inspections, but city governments rarely have the in-house talent needed for developing prediction algorithms. Cities could hire consultants, but a cheaper alternative is to crowdsource competence by making data public and offering a reward for the best algorithm. This paper provides a simple model suggesting that open tournaments dominate consulting contracts when cities have a reasonable tolerance for risk and when there is enough labor with low opportunity costs of time. We also illustrate how tournaments can be successful, by reporting on a Boston-based restaurant hygiene prediction tournament that we helped coordinate. The Boston tournament yielded algorithms—at low cost—that proved reasonably accurate when tested "out-of-sample" on hygiene inspections occurring after the algorithms were submitted. We draw upon our experience in working with Boston to provide practical suggestions for governments and other organizations seeking to run prediction tournaments in the future.

Edward L. Glaeser
Department of Economics
315A Littauer Center
Harvard University
Cambridge, MA 02138
and NBER
eglaeser@harvard.edu

Andrew Hillis
Harvard University
Soldiers Field
Boston, MA 02163
ahillis@hbs.edu

Scott Duke Kominers
Society of Fellows
Harvard University
Soldiers Field
Boston, MA 02163
kominers@fas.harvard.edu

Michael Luca
Harvard University
Soldiers Field
Boston, MA 02163
mluca@hbs.edu

A data appendix is available at http://www.nber.org/data-appendix/w22124

1 Introduction

Big data and predictive analytics have the potential to make cities more efficient. Yet cities often lack the resources to use new data and methods effectively, and private researchers often lack the incentives to help solve the problems that cities face.

New platforms—such as Kaggle, TopCoder, and DrivenData—now enable governments and other organizations to outsource large-scale prediction problems via open tournaments. *But can open tournaments really help solve public problems?*

In this paper, we theoretically and empirically explore the potential of prediction tournaments to improve city operations and translate data science insights into practice. We provide a model that illustrates the tradeoff between tournaments and service delivery via consultancy. Then, we describe a tournament we co-ran to source algorithms that could use Yelp data to improve targeting of restaurant inspections in the City of Boston. Finally, we draw upon our experience in working with Boston to provide practical suggests for governments and other organizations seeking to run prediction tournaments in the future.

The open tournaments we describe here are quite different from the classic economic approach to tournaments introduced by Lazear and Rosen (1981). In the Lazear and Rosen (1981) model, tournaments take the form of labor contracts, under which workers within a firm compete to win prizes (like promotion) by performing parallel but non-redundant tasks. The tournaments that we discuss here, by contrast, are open to the public and necessarily involve redundant labor.[1]

In our setting, the city isolates a problem—in our application, predicting which restaurants are more likely to fail hygiene inspections. The city (potentially with private partner organizations—in our case, Yelp) then streams a body of relevant data and offers some sort of (possibly non-pecuniary) reward for algorithms that solve the city's problem. Citizens produce algorithms and submit them to a central server; the winning algorithms are then rewarded, publicized, and implemented.

Algorithm crowdsourcing tournaments are not (like in Lazear and Rosen (1981) tournaments) incentive contracts purged of common shocks; rather, they are a special form of labor outsourcing. The natural comparison for an open tournament then, is outsourcing to a single consulting company. (We do not examine the possibility of insourcing, i.e., providing the service in-house.[2]) Throughout most of recent history, the consultant model has been dominant. Theoretically, consultants can guarantee a threshold level of service

[1] The winning algorithms are implemented, and the losing algorithms are wasted labor. Thus, our tournaments roughly correspond the settings of most online tournament platforms, as well as tournaments used by companies like Netflix and researchers in fields like computational biology. Such tournaments are sometimes modeled as all-pay auctions, but the prior work has focused on questions of optimal mechanism/prize design, rather than on the types of problems best solved via contests (see, e.g., Che and Gale (2003); Siegel (2009)).

[2] We assume that the government does not want to hire the full-time labor needed to perform the algorithm design task. The public decision to outsource is motivated by the relative uniqueness of the task (designing prediction algorithms is not standard procedure in most City Halls), as well as significant rules governing public labor (such as local residency requirements).

quality, and avoid compensating redundant labor. While tournaments always involve redundant labor and the risk of a particularly poor outcome, they offer lower costs and a higher chance of upper tail events. In Section 2 of this paper, we develop a formal model comparing tournaments to consultancy. The model suggests that tournaments require (1) that the public sector is comfortable with project risk and (2) an abundant supply of low-cost labor willing to work in exchange for a chance of public recognition and a moderate prize.

In Section 3, we describe the design and initial results of an open tournament that we ran in collaboration with the City of Boston, Yelp, and DrivenData. The contest awarded financial prizes for the algorithms that most effectively used Yelp review text to predict Boston restaurant health and sanitation violations.[3] We believe that both of the conditions our model highlights were met in the case of the Boston competition: As prediction algorithms represent a completely new product, with limited public visibility, the downsides of failure were limited. Meanwhile, Boston has an abundance of smart data scientists eager for recognition and a chance to test, refine, and display their skills—and who occasionally have a low opportunity cost of time. Boston also has a high enough profile to encourage data scientists from elsewhere to compete.[4]

Over seven hundred people signed up for the tournament, and fifty-five ultimately contributed a total of 449 sets of predictions. We tested thirty-six "final" algorithm submissions out-of-sample, comparing their predictions to the true results of the 364 restaurant inspections conducted over a six-week period after the close of submissions; the winning algorithms were those that did the best in this "Evaluation Phase." The evidence suggests that using the winning algorithms to identify restaurants to inspect could increase inspection efficacy significantly: We estimate that the City of Boston could increase inspection productivity 30%-50% by allocating inspections as suggested by a top-performing algorithm from the tournament.

In Section 4, we discuss general lessons for cities seeking to implement prediction tournaments. In practice, most cities do not have data scientists or economists in house, and thus often lack the time and human capital to develop predictive algorithms internally. Tournaments provide a tool for outsourcing expertise—yet even designing and implementing tournaments requires some degree of technical competence that cities may lack. Currently, governments may be able to partner with academic or other non-profit partners that are willing to provide free tournament design advice. As the number of tournaments grows, however, the ability to lean on unpaid advisors may decline. Consequently, it is helpful to build a toolkit that enables cities to implement and evaluate tournaments on their own; Section 4 begins the process of developing such a toolkit.

[3] Earlier research has provided evidence suggesting that Yelp text could be used to make inspections more efficient (Kang et al., 2013); however, prior to our work, (to our knowledge) this insight had not been incorporated into city inspection processes.

[4] Moreover, the competition had the nonpecuniary benefits of being interesting to work on and valuable to the city.

2 A Brief Model of Tournaments

This section develops a framework for cities (and other organizations) that are deciding whether to use a tournament to develop a product—in our application context, a predictive algorithm. The tournaments we study have two essential features: (1) they are open to all, and (2) they task all participants with the same goal.[5]

As an alternative to running a tournament, the government can choose to contract with a consultancy that will receive fixed compensation as long it produces a product above some specified quality level. We constrain both the consultancy and the tournament to have simple contracting structures, in line with historical norms and the non-verifiable nature of innovation quality. A consultant's contract specifies a flat payment and a minimum quality threshold; the payment must be made if the produced quality exceeds the threshold. A tournament, by contrast, specifies a prize that must be awarded to the entrant that produces the highest-quality product.

The advantage of the tournament is that it will attract a wide range of workers, at potentially lower costs; the disadvantage is that the tournament will lead to work duplication. We show, intuitively, that tournaments make sense when the value of drawing on a wide pool of workers offsets the costs of duplication of effort.

The city chooses an option that maximizes the expected value of $V(q) - \text{Cost}$, where q is the produced quality level. All actors are assumed to be risk neutral. Consulting companies compete for government contracts, and earn no expected rents in equilibrium; hence, they deliver the lowest-cost means of achieving any fixed level of quality. In a tournament, workers will enter to the point at which their expected returns equal their opportunity costs of time.

For simplicity, we assume that each worker is of either *high* or *low* skill. High-skilled workers and low-skilled workers respectively have opportunity costs of time equal to $\overline{w}$ and $\underline{w}$. When performing the task, high-skilled workers offer a minimum quality level of $\overline{q}$ and low-skilled workers offer a minimum quality level of $\underline{q}$. With probability φ either type of worker can achieve a "breakthrough" that increases output quality to $q_{\max}$, which is greater than $\overline{q}$.[6]

In a consulting contract, there are three plausible values for the minimum quality level: $\underline{q}$, $\overline{q}$, and $q_{\max}$. As consulting companies offer to fulfill the contract at its expected cost, the contract that specifies $\underline{q}$ will cost $\underline{w}$; the contract that specifies $\overline{q}$ will cost $\overline{w}$; and the contract that specifies $q_{\max}$ will cost $\frac{\underline{w}}{\varphi}$. We assume that $\frac{\underline{w}}{\varphi} > \overline{w}$. If $V(\overline{q}) - V(\underline{q}) > \frac{\overline{w}-\underline{w}}{1-\varphi}$ and $\frac{\frac{\underline{w}}{\varphi}-\overline{w}}{1-\varphi} > V(q_{\max}) - V(\overline{q})$, then the city will prefer the high-quality consulting contract to either the low-quality contract or the maximum-quality contract.

[5] Thus, as discussed in the Introduction, our tournaments are quite different from those historically modeled by economists (Lazear and Rosen (1981)), in which workers compete to contribute to a firm's productivity by performing non-identical tasks, with promotions awarded based on relative achievement.

[6] Our risk assumption serves to make tournaments more attractive—the case for tournaments relies on the existence of workers with some upside potential who have low opportunity costs of time.

If the city runs a tournament, then it posts reward value R that is granted to the participant who delivers the highest-quality project.[7] A tournament with reward R attracts $N = N(R)$ participants. In theory, the participants may be all high-skilled workers (in which case $N = \frac{R}{\overline{w}}$ and $\underline{w} > \overline{w} - \overline{w}(1-\varphi)^{\frac{R}{\overline{w}}}$) or all low-skilled workers (in which case $N = \frac{R}{\underline{w}}$ and $\overline{w} - R(1-\varphi)^{\frac{R}{\underline{w}}} > \underline{w}$) or a mixture of both types (in which case $\frac{R(1-(1-\varphi)^N)}{N} + \frac{R(1-\varphi)^N}{N_H} = \overline{w}$ and $\frac{R(1-(1-\varphi)^N)}{N} = \underline{w}$).

We focus on tournaments that only attract low-skilled workers; such tournaments arise when wage inequality is large. In this case, the tournament sponsor chooses N (by choosing R) to maximize

$$(1-\varphi)^{\frac{R}{\underline{w}}} V(\underline{q}) + (1-(1-\varphi)^{\frac{R}{\underline{w}}}) V(q_{\max}) - R,$$

so that (ignoring integer constraints)

$$-\ln(1-\varphi)(1-\varphi)^{\frac{R}{\underline{w}}}(V(q_{\max}) - V(\underline{q})) = \underline{w}$$

determines the optimal R.

Now, we consider the tradeoff between a consulting contract and a tournament that both cost the same amount. If the consulting contract pays $\underline{w}$, then a tournament that pays the same amount yields the same result as the consultancy, as both draw only one low-skilled worker. If the consulting contract pays $\frac{\underline{w}}{\varphi}$, then the tournament is clearly dominated, as it does not guarantee maximal quality, while the consulting contract does.[8]

The most interesting comparison arises when the consultancy and the tournament both pay $\overline{w}$. One possibility (depending on parameter values) is that the tournament attracts one high-skilled worker, in which case the two contracts are again equivalent. Alternatively, when wage inequality is high, the tournament attracts $\frac{\overline{w}}{\underline{w}}$ low-skilled workers. In this case, the tournament dominates the consultancy if and only if $(1-\varphi)^{1-\frac{\overline{w}}{\underline{w}}} - 1 > \frac{V(\overline{q})-V(\underline{q})}{V(q_{\max})-V(\overline{q})}$. We then find:

Proposition 1. *There exists a value of φ, denoted φ^*, at which the returns to the tournament are the same as the returns to the consulting contract. For values of $\varphi > \varphi^*$, the tournament dominates the consultancy and for values of $\varphi < \varphi^*$, the consultancy dominates the tournament. The value of φ^* increases in $V(\overline{q})$ and decreases with $V(\underline{q})$, $V(q_{\max})$, and $\frac{\overline{w}}{\underline{w}}$.*

Proof. The value of $(1-\varphi)^{1-\frac{\overline{w}}{\underline{w}}} - 1$ is monotonically increasing in φ and goes from 0 to ∞ as φ goes from 0 to 1. Hence, there must exist a value of φ at which $(1-\varphi)^{1-\frac{\overline{w}}{\underline{w}}} - 1$ equals $\frac{V(\overline{q})-V(\underline{q})}{V(q_{\max})-V(\overline{q})}$, a constant. The value of $\frac{V(\overline{q})-V(\underline{q})}{V(q_{\max})-V(\overline{q})}$ is rising with $V(\overline{q})$ and falling with $V(\underline{q})$ and $V(q_{\max})$; hence, φ^* is rising with $V(\overline{q})$

[7] If multiple workers "win" the tournament by delivering the same highest level of quality, then they split the reward.

[8] In our setting, to provide a guarantee of the maximal outcome in a tournament, the reward would have to be infinite.

and falling with $V(\underline{q})$ and $V(q_{\max})$. For a given φ, the value of $(1-\varphi)^{1-\frac{\overline{w}}{\underline{w}}}-1$ is rising with $\frac{\overline{w}}{\underline{w}}$; hence, φ^* must be falling with $\frac{\overline{w}}{\underline{w}}$. □

Proposition 1 tells us that tournaments make sense when the probability of a breakthrough, φ, is relatively high. The range of values that make tournaments attractive increases with $V(\underline{q})$, $V(q_{\max})$, and $\frac{\overline{w}}{\underline{w}}$. That is, tournaments are more appealing when the baseline low-skilled outcome is not that bad, and when the best-possible outcome is particularly good. Wage inequality also makes tournaments more appealing, as tournaments attract workers with a particularly low opportunity cost of time. When cities want to ensure that they achieve at least the middle outcome, $\overline{q}$, tournaments are less attractive.

Our model suggests that the appeal of tournaments depends on wage inequality and the public sector tolerance for risk. Tournaments thus may be particularly attractive in the 21st century, because there are now many information technology workers, particularly in developing countries, with relatively low opportunity costs of time. The second factor that drives the appeal of tournaments is the public tolerance for risk. When running a tournament, the city must be willing to trade a reduced chance of getting a middling outcome for an increased probability of getting an outcome in the upper and lower tails.

3 A Restaurant Hygiene Prediction Tournament

Historically, cities have sometimes partnered with academics to solve technical problems, such as designing allocation mechanisms for assigning students to public schools (see, e.g., Abdulkadiroğlu, Pathak, and Roth (2005); Abdulkadiroğlu, Pathak, Roth, and Sönmez (2005)). Yet in practice, there are many important problems that cities face but that academics lack incentives to solve.[9]

Prediction tournaments provide a means for academics to assist city governments "lightly," providing design support but ultimately leaving the heavy lifting to the crowd; this type of partnership makes sense for many prediction problems, as the basic statistical approaches to prediction are by now well-understood. The missing ingredients for most cities' prediction problems are tuning and implementation, not blue-sky thinking.

Applying the framework described above, and building on preliminary evidence of Kang et al. (2013) suggesting that Yelp text could predict inspection failures, we partnered with the City of Boston, Yelp, and DrivenData to run an open tournament sourcing algorithms for predicting restaurant hygiene and sanitation violations from Yelp reviews.

Yelp was a natural partner, as it already provides public, crowdsourced restaurant review data that often include discussions of hygiene. Yelp has also shown remarkable willingness to partner with governments,

[9] Academics are most attracted to settings in which cities need to develop completely new solutions; by contrast, academics are less interested in settings where cities are looking to implement or scale existing insights.

reflecting both the company leadership's interest in problem-solving and its desire to cooperate with relevant local authorities.

The City of Boston was also a natural partner. The academic team is based in the Boston metropolitan area, and has a long history of cooperation with the Boston government. Moreover, the City of Boston has positioned itself on the cutting edge of intrapreneurship and technology in government, with initiatives such as the *New Urban Mechanics* (`http://newurbanmechanics.org/`). Moreover, we were confident that Boston would be able to attract tournament participants, both because of its abundance of skilled labor and because of its relatively high profile in the technology community.

DrivenData, finally, was a natural host for the tournament because it is Boston-based and focuses on data science contests with social missions. Working with DrivenData enabled us to attract data scientists specifically interested in public projects like the hygiene prediction tournament.

3.1 Tournament Structure

Participants in the tournament had twelve weeks to develop algorithms for predicting hygiene violations from Yelp data. While developing their algorithms, participants had access to a dataset recording 34,879 City of Boston hygiene inspections, dating back to April 2006, and a linked set of Yelp.com reviews, ratings, and business attributes for Boston restaurants recorded over the same time period.[10]

In Phase I of the tournament (the "Development Phase"), participants developed predictive algorithms based on historical data. During this phase, participants could share their predictive performance publicly on the DrivenData website, which ranked the highest performers to date. Over seven-hundred people registered for the tournament. Fifty-five competitors completed the Development phase, submitting a total of 449 sets of predictions. At the end of the Development Phase, participants submitted "final algorithms" for evaluation. Twenty-three competitors submitted a total of thirty-six separate final algorithms. In Phase II (the "Evaluation Phase"), the final algorithms were evaluated according to their effectiveness in predicting the outcomes of inspections conducted in a six-week test period that started after final algorithm submission. During the test period, the City of Boston conducted inspections in its usual manner, inspecting a total of 364 restaurants. The winning algorithm's designer received $3,000; the second- and third-place algorithms' designers each received $1,000; prize money was provided by Yelp.

[10]Each inspection record indicated the number of minor, major, and severe violations that an inspector found at the inspected restaurant. Yelp reviews include customers' ratings of restaurants, along with explanatory comments (for further discussion, see the work of Luca (2011, forthcoming)).

3.2 Results

Algorithm performance was measured by root mean squared logarithmic error (RMSLE).[11] Figure 1 shows the distribution of performance in the Development and Evaluation phases.

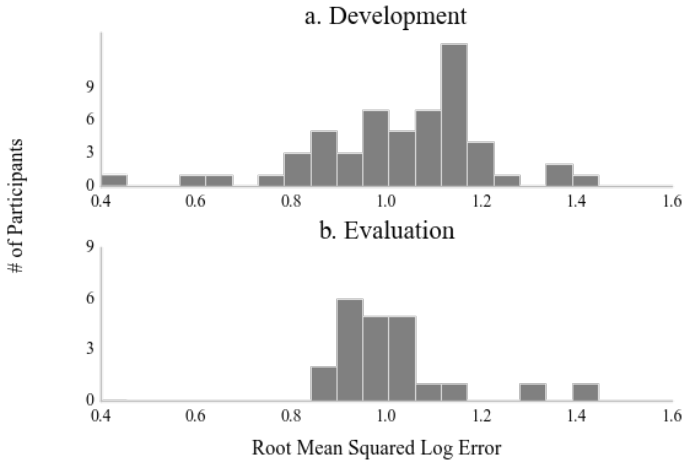

Figure 1: *Distribution of the top scores of each participant in each phase of the competition, as measured by root mean squared logarithmic error.*

Fifty-five participants submitted predictions in the Development Phase. Twenty-three participants competed in the Evaluation Phase. The testing data used during the Development Phase were based on the historical record of violations; participants had access to this data while training their algorithms. The testing data used during the Development Phase were the results of inspections conducted after submissions were complete; participants did not have access to this data.

The winner of the tournament was a data scientist based in the United Kingdom; her submission used the statistical program R to implement an average of predictions from a random forest model and gradient-boosted model with decision trees. The runner-up was a Ph.D. student in Marketing Analytics from the Netherlands; his submission used Python to implement a random forest model. Just as our theory model predicts, the tournament was effective in large part because it sourced contributions from a broad array of participants.

While the tournament itself scored submissions according to a standard prediction performance metric—RMSLE—the outcome for the City of Boston is best measured in terms of productivity. That is: *How much can using predictive algorithms improve inspector allocation?*

During the testing period, the City inspected 364 restaurants, uncovering 1,593 minor violations, 153 major

[11] To compute this scoring function, we first collapsed the prediction for each restaurant i into a unidimensional prediction $\hat{Y}_i$, by weighting the number of minor (1x), major (2x), and severe (5x) violations. We also computed the actual (weighted) numbers of violations Y_i that were found during the test period. The performance metric was

$$\text{RMSLE} = \sqrt{\frac{1}{\#\text{Restaurants}} \sum_i \Big(\log(\hat{Y}_i + 1) - \log(Y_i + 1)\Big)^2}.$$

violations, and 341 severe violations, for a total of 3,604 total weighted violations. We predict that if the City had used the algorithms to prioritize 364 restaurants to inspect from the universe of restaurants available in the data, it would have found 5,406 weighted violations (or 4,756 using the runner-up)—50% more violations than were found using the baseline inspector allocation system (or 32% more using the runner-up).[12] Thus, we estimate that the City of Boston would be 30%-50% more productive using a top-performing algorithm from the tournament. We are currently testing the winning algorithms' efficacy in practice, using a field experiment that integrates the winning algorithms into Boston's process for allocating inspectors.

4 Designing Prediction Problems and Tournaments

In this section, we draw upon our theory work and experience in Boston to offer general commentary on issues cities should consider when designing prediction tournaments.

4.1 Problem Selection and Setup

Prediction tournaments are most effective for solving well-defined prediction problems for which large data sets are available (either to the tournament organizer, or through external sources). For the case of hygiene prediction, for example, deciding which restaurants to inspect directly involves an element of prediction, yet inspections had not incorporated systematic predictive efforts in the past. Moreover, hygiene prediction provided us with an opportunity to incorporate new digital data sources (Yelp reviews), offering at least the possibility of significant improvements in predictive accuracy.

Cities can incorporate algorithms into many of their operational processes that involve prediction. Because algorithms are extremely literal and do not make implicit tradeoffs the way that a policymaker would, cities should be explicit about all of their tournament objectives, keeping in mind both intended and unintended consequences. A simple example from the hygiene prediction context is the need for specifying tradeoffs between minor, major, and severe violations. Through conversations with stakeholders, a complete set of design objectives can be identified and then formally integrated into the tournament scoring function (for further discussion, see Luca et al. (2016)).

[12] Alternatively, had the city used the winning algorithm to prioritize restaurants for inspection, it could have inspected only 219 restaurants (249 using the runner-up), reducing the number of inspections by 40% (32%) while identifying the same number of weighted violations and risks.

4.2 Choosing Data

Finding the relevant data for a prediction tournament requires a systematic approach to determining the value of different data sets, as well as collaboration with partner organizations. Beginning with internal data is sensible, but that data may lack critical information or be low-frequency. In the hygiene prediction competition, prior violations provided a good signal about future violations, but the frequency and scope of the Yelp data allowed for finer predictions.

4.3 Incentives and Information

Competitors are driven by a variety of motives, from prize money to job market signaling to just the opportunity to work on interesting and important problems; cities have considerable flexibility in leveraging all of these incentives. For example, a major incentive for participants is the ability to signal competence publicly, and city governments have the ability to generate significant publicity, partially because they have members (like the Mayor) who are covered regularly by the media. Tournament designers must decide how and at which stages to use publicity—upfront media attention can serve to inform prospective participants and generate competition, while *ex post* publicity serves to increase the signaling value of winning.

Entrant effort also responds to perceptions about the size and talent of the field. We witnessed a decline in the number of competitors by the end of the Development Phase. A likely explanation is that after observing relative performance during the Development Phase, many competitors opted not to submit final algorithms. Tournament designers must consider the timing and set of information delivered to participants about other competitors (for further discussion, see Boudreau et al. (forthcoming)).

4.4 Choosing a Platform

As our theory indicates, the talent pool largely determines the success of the tournament. Thus, as in many other domains, choosing the right platform is essential. At this point, the main prediction tournament platforms have developed their own distinct user followings. On larger platforms such as Kaggle, there are more than 450,000 registered data scientists who all receive an email when each competition begins. When setting up a tournament, the designer should look for a platform that has run similar tournaments in the past, and should look at prior tournaments' outcome statistics to get estimates of participation and expected performance.

4.5 Measuring Success

It is essential to choose metrics for evaluating both (1) tournament participants and (2) the overall value of the tournament itself. It is valuable to score entries according to information collected after the close of submissions—as we did in the Evaluation Phase—so as to have a true out-of-sample test.[13] And of course, tournaments are not free, and thus should only be used if they deliver value that exceeds costs. As we mentioned in Section 3.2, we have evaluated (and are continuing to evaluate) the value of the prediction tournament itself by asking how much the winning algorithms can improve the allocation of inspector time in Boston in practice.

5 Conclusion

Open tournaments are a new and exciting tool for leveraging latent, low-cost talent to solve cities' problems. However, tournaments are not a panacea, as they have downside risk and involve duplication of effort. Tournaments are thus most effective when (1) the organizers are comfortable with project risk and (2) an abundant supply of low-cost labor is available. In the case of Boston, the needed conditions were met, and the hygiene prediction tournament we co-ran successfully sourced algorithms that can improve inspector allocation.

Looking forward, the need is to develop a core set of tools for evaluating and implementing tournaments without heavy academic involvement. City governments should focus on tasks that are hard, but not too hard, and where there is data and interest from talented information technology workers. Even more importantly, cities need to develop good mechanism for judging success, so that we can learn where and when tournaments achieve the best results.

It is not clear whether tournaments will be complements or substitutes for in-house talent. In some cases, tournaments can eliminate the need for cities to hire knowledge workers. Yet tournaments themselves require a level of technological savvy that may increase the returns to having in-house talent.

Moreover, while we are optimistic about the potential of tournaments, we do not expect the open tournament model to eliminate traditional consulting contracts. There are many areas in which tournaments may fail to generate the basic quality level required by a city. Moreover, the supply of tech-savvy labor with relatively low opportunity costs of time may eventually dry up. But for the moment, tournaments seem like a valuable tool for connecting public consumers of technology products with producers who are willing to work for the prospect of a small reward and a bit of glory.

[13] Additionally, it is important to figure out a clear way of communicating the scoring system to tournament participants.

References

Abdulkadiroğlu, A., P. A. Pathak, and A. E. Roth (2005). The New York City high school match. *American Economic Review 95*(2), 364–367.

Abdulkadiroğlu, A., P. A. Pathak, A. E. Roth, and T. Sönmez (2005). The Boston public school match. *American Economic Review 95*(2), 368–371.

Boudreau, K. J., K. Lakhani, and M. E. Menietti (forthcoming). Performance responses to competition across skill-levels in rank order tournaments: Field evidence and implications for tournament design. *RAND Journal of Economics*.

Che, Y.-K. and I. Gale (2003). Optimal design of research contests. *American Economic Review 93*(3), 646–671.

Kang, J. S., P. Kuznetsova, M. Luca, and Y. Choi (2013). Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1443–1448.

Lazear, E. P. and S. Rosen (1981). Rank-order tournaments as optimum labor contracts. *The Journal of Political Economy 89*(5), 841–864.

Luca, M. (2011). Reviews, reputation, and revenue: The case of Yelp.com. Harvard Business School NOM Unit Working Paper No. 12-016.

Luca, M. (forthcoming). User-generated content and social media. In S. Anderson, J. Waldfogel, and D. Stromberg (Eds.), *Handbook of Media Economics*. Elsevier.

Luca, M., J. Kleinberg, and S. Mullainathan (2016). Algorithms need managers, too. *Harvard Business Review 94*, 96–101.

Siegel, R. (2009). All-pay contests. *Econometrica 77*(1), 71–92.