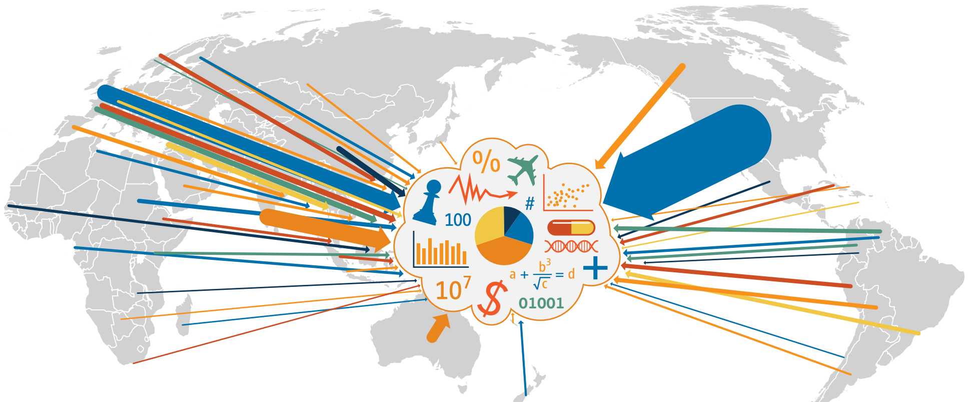

**Global contest.** Kaggle's competitions draw entries from many countries (arrow thickness reflects number of competitors from a country).

NEWS

# May the Best Analyst Win

**Exploiting crowdsourcing, a company called Kaggle runs public competitions to analyze the data of scientists, companies, and organizations**

LAST MAY, JURE ŽBONTAR, A 25-YEAR-OLD computer scientist at the University of Ljubljana in Slovenia, was among the 125 million people around the world paying close attention to the televised finale of the annual Eurovision Song Contest. Started in 1956 as a modest battle between bands or singers representing European nations, the contest has become an often-bizarre affair in which some acts seem deliberately bad—France's 2008 entry involved a chorus of women wearing fake beards and a lead singer altering his vocals by sucking helium—and the outcome, determined by a tally of points awarded by each country following telephone voting, has become increasingly politicized.

Žbontar and his friends gather annually and bet on which of the acts will win. But this year he had an edge because he had spent hours analyzing the competition's past voting patterns. That's because he was among the 22 entries in, and the eventual winner of, an online competition to predict the song contest's results.

The competition was run by Kaggle, a small Australian start-up company that seeks to exploit the concept of "crowdsourcing" in a novel way. Kaggle's core idea is to facilitate the analysis of data, whether it belongs to a scientist, a company, or an organization, by allowing outsiders to model it. To do that, the company organizes competitions in which anyone with a passion for data analysis can battle it out. The contests offered so far have ranged widely, encompassing everything from ranking international chess players to evaluating whether a person will respond to HIV treatments to forecasting if a researcher's grant application will be approved. Despite often modest prizes—Žbontar won just $1000—the competitions have so far attracted more than 3000 statisticians, computer scientists, econometrists, mathematicians, and physicists from approximately 200 universities in 100 countries, Kaggle founder Anthony Goldbloom boasts.

And the wisdom of the crowds can sometimes outsmart those offering up their data. In the HIV contest, entrants significantly improved on the efforts of the research team that posed the challenge. Citing Žbontar's success as another example, Goldbloom argues that Kaggle can help bring fresh ideas to data analysis. "This is the beauty of competitions. He won not because he is perhaps the best statistician out there but because his model was the best for that particular problem. … It was a true meritocracy," he says.

## Meeting the mismatch

Trained as an econometrician, Goldbloom set up his Melbourne-based company last year to meet a mismatch between people collecting data and those with the skills to analyze it. While writing about business for *The Economist*, Goldbloom noted that this disconnect afflicted many fields he was covering. He pondered how to attract data analysts, like himself, to solve the problems of others. His solution was to entice them with competitions and cash prizes.

This was not a completely novel idea. In 2006, Netflix, an American corporation that offers on-demand video rental, set up a competition with a prize of $1 million to design software that could better predict which movies customers might like than its own in-house recommendation software, Cinematch. Grappling with a huge data set—millions of movie ratings—thousands of teams made submissions until one claimed the prize in 2009 by showing that its software was 10% better than Cinematch. "The Netflix Prize and other academic data-mining competitions certainly played a part in inspiring Kaggle," Goldbloom says.

The prizes in the 13 Kaggle competitions so far range from $150 to $25,000 and are offered by the individuals or organizations setting up the contests. For example, chess statistician Jeff Sonas and the German company ChessBase, which hosts online games, sponsored a Kaggle challenge to improve on the player-ranking system developed many decades ago by Hungarian-born physicist and chess master Arpad Elo. Its top prize was a DVD signed by several world chess champions.

Still, Kaggle has shown that it doesn't take a million-dollar prize to pit data analyst against data analyst. Kaggle's contests have averaged 95 competitors so far, and the chess challenge drew 258 entries. "When I started running competitions, I found they were more popular and effective than I could have imagined," Goldbloom says. "And the trend in the number of teams entering seems to be increasing with each new competition."

Statistician Rob Hyndman of Monash University, Clayton, in Australia, recently used Kaggle to lure 57 teams, including some from Chile, Antigua and Barbuda, and Serbia, into improving the prediction of how much money tourists spend in different regions of the world. "The results were amazing. … They quickly beat our best methods," he says.

Hyndman suspects that part of Kaggle's success is offering feedback to competitors. Kaggle works by releasing online a small part of an overall data set. Competitors can analyze this smaller data set and develop appropriate algorithms or models to judge how the variables influence a final outcome. In the chess challenge, for example, a model could incorporate a player's age, whether they won their previous game, if they played

CREDIT: ADAPTED FROM KAGGLE

Published by AAAS

with white or black pieces, and other variables to predict whether a player will win their next game. The Kaggle competitors then use their models to predict outcomes from an additional set of inputs, and Kaggle evaluates those predictions against real outcomes and feeds back a publicly displayed score. In the chess challenge, the results of more than 65,000 matches between 8631 top players were offered as the training data set, and entrants had to predict the winners of nearly 8000 other already-played games.

During a competition, which usually lasts 2 months, people or teams can keep submitting new entries but no more than two a day. "Seeing your rivals, and that they are close, spurs you on," says Hyndman.

Kaggle encourages the sponsors of the competition to release the winning algorithm—although they are not always persuaded to do so—and asks the winning team to write a blog post about how they tackled the problem and why they think their particular approach worked well. Goldbloom hopes that this means other entrants get something out of the competition despite not winning. They not only hone analytical skills by taking part, he says, but also are able to learn from other approaches.

| KAGGLE COMPETITIONS | PRIZE | COMPETITORS |
|---|---|---|
| Predicting acceptance of grant applications for the University of Melbourne | $5000 | 90 |
| Predicting the "edges" of online social networks | $950 | 106 |
| Improving chess player rating system | $617 | 258 |
| Forecasting the movement of tourists around the globe | $500 | 57 |
| Predicting HIV progression in people taking different combinations of drugs | $500 | 109 |
| Predicting how far each country's football team will progress through the World Cup | $100 | 65 |
| Estimating travel time on one of Australia's main traffic arteries | $10,000 | 205 |
| Forecasting the final rankings of countries in the 2010 Eurovision Song Contest | $1000 | 22 |

**Business solution.** Anthony Goldbloom (*left*) founded Kaggle to run contests to solve data problems.

CREDIT: A. GOLDBLOOM; MELBOURNE (INSET)

### Predicting potential

Although only a handful of its competitions have finished, Kaggle has had promising results so far. Each contest has generated a better model for its data than what was used beforehand.

Bioinformaticist William Dampier of Drexel University in Philadelphia, Pennsylvania, organized the competition to predict, from their DNA, how a person with HIV might respond to a cocktail of antiretroviral drugs. This problem had been tackled extensively in academia, where the best models predicted the response of a patient to a set of three drugs with about 70% accuracy. By the end of the 3-month contest, the best entry was predicting a person's drug response with 78% accuracy. Dampier says even this improvement in accuracy could help doctors further improve their treatment strategies beyond the current "guess the drug and check back later" approach.

Dampier considers Kaggle's approach innovative, noting that it draws in data analyzers with various backgrounds and perspectives who are not shackled by a field's dogma. Such outsiders, he suspects, are more likely to see something different and useful in the data set. "The results talk, not your position or your prestige. It is simply how well you can predict the data set," says Dampier.

His point is well illustrated by Žbontar. Despite not tabbing Eurovision's actual winner, Germany, his overall prediction of the results beat a team from the SAS Institute—a data-mining company—and a team from the Massachusetts Institute of Technology. His submission incorporated both past national voting patterns—Eastern European countries tend to vote for each other, for example—and betting odds for the current contest.

Goldbloom also attributes Kaggle's success to crowdsourcing's capacity to harness the collective mind. "Econometrists, physicists, electrical engineers, actuaries, computer scientists, bioinformaticists—they all bring their own pet techniques to the problem," says Goldbloom. And because Kaggle encourages competitors to trade ideas and hints, they can learn from each other.

One sponsor of a Kaggle competition estimates that some entrants may have spent more than 100 hours refining their data analysis. This begs the question: What's the attraction, given the small prizes? Many data analysts, Goldbloom discovered, crave real-world data to develop and refine their techniques. Timothy Johnson, an 18-year-old math undergraduate at the California Institute of Technology in Pasadena, says working with the real data of the chess-ranking competition—he finished 29th—was more challenging, educational, and "fun" than analyzing the fabricated data sets classes offer.

For Chris Raimondi, a search-engine expert based in Baltimore, Maryland, and winner of the HIV-treatment competition, the Kaggle contest motivated him to hone his skills in a newly learned computer language called R, which he used to encode the winning data model. Raimondi also enjoys the competitive aspect of Kaggle challenges: "It was nice to be able to compare yourself with others; … it became kind of addictive. … I spent more time on this than I should."

What has proved tricky for Kaggle is persuading companies, agencies, and researchers to open up their data. Goldbloom tries to assuage companies' concerns about putting some of their data up on the Web by pointing out that they will get a competitive advantage if the Kaggle contestants produce a better solution to their data problems. So far, two private companies, one government agency, and three universities are among the groups to have used Kaggle.

As for researchers, Goldbloom says most reject his advances with an almost "visceral reaction." Overcoming such reluctance to expose data may be key to his company's survival. No one pays to enter a competition, so Kaggle depends on charging a fee to those running a contest—the sum changes from competition to competition. "We aren't profitable yet, but we have some huge projects coming up and we hope to be profitable by the end of the year," says Goldbloom.

Žbontar hopes Kaggle survives, as he's looking forward to bettering his prediction model for this year's Eurovision Song Contest and perhaps prying his friends out of more beer money. In a blog post analyzing his victory this past year, he issued this playful challenge: "I have many ideas for next year, which I will, for the moment at least, keep to myself."

**–JENNIFER CARPENTER**

Published by AAAS