AI Bookie

At ai.sciencebets.org you can make your own predictions, challenge another prediction and turn it into a bet, or post a bet of your own. We are here to help. So, place your bets at ai.sciencebets.org!

■ *The AI Bookie column documents highlights from AI Bets, an online forum for the creation of adjudicatable predictions and bets about the future of artificial intelligence. Although it is easy to make a prediction about the future, this forum was created to help researchers craft predictions whose accuracy can be clearly and unambiguously judged when a prediction comes due. The bets will be documented online and regularly in this publication in The AI Bookie. We encourage bets that are rigorously and scientifically argued. We discourage bets that are too general to be evaluated or too specific to an institution or individual. The goal is not to continue to feed the media frenzy and pundit predictions about artificial intelligence, but rather to curate and promote bets whose outcomes will provide useful feedback to the scientific community. Place your bets! Please go to ai.sciencebets.org.*

Place Your Bets: Will Machine Learning Outgrow Human Labeling?

Mike Schaekermann, Christopher M. Homan, Lora Aroyo, Praveen Paritosh, Kurt Bollacker, Chris Welty

Some machine learning (ML) rhetoric seems to imply an assumption or expectation that, at some point, machines will outgrow the need for human labeled data. Today's reliance on such labeling is a sort of dirty little secret of artificial intelligence (AI), and some view it as a necessary means to a larger end. This bet is an attempt to formalize that attitude into a concrete question, whose answer can be measured over time.

The process of putting together this bet, which began a year ago at the Subjectivity, Ambiguity, and Disagreement workshop held in 2019 at The Web Conference, was quite long and involved. It goes, in part, to show why we have had such a hard time filling this column quarterly with bets, even though the support for AI bets has been tremendous. The bettors and bookies, however, found the entire process valuable — delightful even — and many potential side projects cropped up during the course of its specification. We came to believe this adversarial process, outlined in the first bookie column two years ago, is an important way to move science forward in an understandable way. Scientists are all, or should all be, skeptics, and there is nothing like having to justify yourself to a skeptic — not an arbitrary skeptic, but a knowledgeable one. We should be looking to move beyond our echo chambers.

The Bet

The reliance of ML on human labeled data will decrease in the next five years. To measure this, we have chosen a surrogate to quantify the reliance of ML on human labeled

ISSN 0738-4602

data: the proportion of research papers at major AI conferences, and reformulate the bet as: the percentage of ML papers at the Association for the Advancement of Artificial Intelligence conference, the Neural Information Processing Systems conference, and The Web Conference that use human labeled data for training or evaluating models will be lower in 2025 than in 2020.

Pro bet: Adjudication criteria accepted.
Con bet: Adjudication criteria accepted.

For: Mike Schaekermann and Christopher M. Homan

We Believe the Reliance on Human Labeled Data in AI Will Decrease, for the Following Reasons

Methods for collecting human data passively will improve. Over the next five years, we believe that technological advances in sensing and AI will make it possible to collect many of the kinds of labels that now require human judgement, and do so more cheaply. For example, two published market research reports[1,2] forecast a steady annual growth of at least 15 percent for the global market on wearable technologies over the next few years. We expect that this growth will be accompanied by an emergence of novel sensor modalities and the proliferation of longitudinal datasets precisely monitoring certain objective outcomes about the human condition (for example, cardiovascular, neural, or endocrine anomalies) for which human expert judgement is nowadays often used as an imperfect proxy.

Human interpretations are less reliable for scientifically valid tests. ML researchers, to more reliably obtain meaningful numbers from their evaluations, will avoid problems that require human labels, even if it means that the labels are in some cases less representative of human judgement. As one example, current AI systems for diagnostic support typically rely on doctors' interpretation of raw data, such as assigning diagnosis codes to medical images. However, inter-rater disagreement is a widely recognized issue across various medical subspecialties, requiring expensive procedures to ensure label quality. We project a shift toward an increased use of longitudinal datasets in which objective outcomes (such as future patient condition) are available, decreasing the need for even expert human judgement.

There are significant disciplinary boundaries between hard AI and human computation. Another pressure is that the so-called "gearheads" who understand the inner mechanics of complex mathematical-learning algorithms often lack the temperament for the messy and awkward work of collecting human annotations. So, while some may seek to reach across disciplinary boundaries for the sake of their work and take the time and trouble to collect human labels, many more will simply rely on data collection methods that are fully automated and thus much more convenient to them.

Demand for nonhuman intelligence will increase. Generally speaking, ML problems tend to fall into two categories: those that humans can already perform, and those that are beyond the capabilities of human intelligence. The latter category would seem to be much larger than the former (although its size is hard even to contemplate, given the same limits of human intelligence). Already, ML has helped us to discover adverse drug reactions by poring through and making sense of much more data than any human team could possibly manage. As another example, ML algorithms today can produce accurate weather forecasts orders-of-magnitude faster, at greater spatial resolution and with significantly less input data, than can human-driven simulations.[3] For most of these tasks human judgement will continue to be insufficient or irrelevant.

Against: Lora Aroyo and Praveen Paritosh

We Believe the Reliance on Human Labeled Data in AI Will Increase for the Following Reasons

Passive data are not enough. Many researchers believe that passively collected data will be widely and cheaply available in the future. However, it is already widely available in many fields, and we agree it will continue to grow, but it doesn't seem to accomplish what human labels do. For example, various types of recommender systems for movies, news, and shopping are still basic in terms of predicting human intent, desires, and needs. Understanding the utility of passive data for different user tasks and needs will always be guided by a human labeling process. For example, search generates a lot of passive data, in the form of clicks and queries, but the need for actively generated human data keeps increasing,[4,5] and there is no evidence that this is ever going to decrease as user needs keep getting more complex. The same holds for recommendation systems for shopping, movies, videos, news, social media, or decision-making systems for self-driving cars and digital assistants.

Demand for understanding subjective notions will increase. While some AI researchers might see the subjectivity and nuance of human cognition as reflected in the disagreement between raters as a problem, missing that nuance is an Achilles heel of today's AI systems.[6,7] This shows up as a lack of common sense or a lack of human understandability. We increasingly depend upon automatically detecting subjective things like toxicity, pornography, fake news, and multilingual or cultural perspectives. We believe that modeling that nuance will enable the next set of breakthroughs in AI systems that will increasingly rely on human labeled data.

Illustration courtesy James Gary.

Demand for multidisciplinary approaches will increase. The argument from the pro team, that "gearheads" who do math and "fuzzies" who understand humans won't talk to each other, is a sad but accurate state of the entire AI field "looking below the lamp." Various research initiatives around the world have been stimulating interdisciplinary collaborations in research projects — for example, the National Science Foundation, the Defense Advanced Research Projects Agency, the National Institute of Health, the Netherlands eScience Center in Amsterdam, and the Dutch Research Council — that we believe indicate a desire for understanding and modeling the phenomena, even if it involves interdisciplinary collaborations. AI has been a strongly multidisciplinary field, learning from and collaborating with linguists, neuroscientists, cognitive psychologists, and philosophers, and we believe that this will continue to increase.

Demand for fidelity to human behavior and explanatory requirements will increase. The pro team argues that the demand for nonhuman intelligence will increase, in part because human intelligence is just one instance of many possible intelligences. Over a large spectrum of diverse tasks, these nonhuman intelligences will have to be able to engage naturally with humans, and thus be able to model human understanding and communication satisfactorily. As AI systems grow beyond mere conveniences, they will have to explain their understanding to human stakeholders to be useful. However, the state of the art in machine explanation is a far cry from what human users and regulators will ultimately want. This, in turn, will rely on more human labeled data. Secondarily, we believe that human intelligence is a particularly important one and will be driving more growth and value to humanity than all the nonhuman intelligences put together. Automating aspects of human intelligence provides utility to humans by saving us time on tasks that we perform and thus know to be important.

Demand for dealing with bias will increase. Unlabeled data, such as web corpora, or other found datasets such as records of hiring decisions made in the past, have been shown to contain biases with social consequences that will become more apparent as predictive models trained on them are deployed. As we become more aware of different biases in our society and how they come bundled with any passively collected data, we will need more human labeled data to both discover and understand those biases automatically. This means there will be a continuous demand to deal with special cases that need human interpretation, and the demand for this will never cease (figure).

Methodology

When we first conceived this bet, *human labeling* meant to us the kind of conventional activity where trained researchers would annotate data with ground truth class labels for the purpose of supervised learning or qualitative coding, or delegate this task to crowd workers via microtasking websites such as Amazon Mechanical Turk. But in negotiating the terms of the bet, we realized that the simultaneous emergence over the last thirty years of the web and remote sensing means that high-quality human input can be obtained in any number of creative ways, often without the humans involved even realizing that they are generating data, such as games with a purpose such as ReCAPTCHA, as well as user interactions such as clicks, scrolls, and swipes. Moreover, straightforward ML problems such as classification, for which traditional labeling suffices, are increasingly less commonly studied, and more sophisticated problems, such as machine translation or autonomous driving, often require newer, less conventional data acquisition methods.

As one example, a common domain for ML is health, where the prediction task could be a disease given a patient's electronic health records. Medical diagnosis, like cancer or depression, usually requires a physician's judgment, and so this seems like a clear-cut case of human labeling. But what about hypertension? In most cases, a human takes blood pressure and enters the data, but once those numbers are known, a diagnosis of hypertension is relatively judgement-free. And blood pressure can now be measured automatically via drugstore kiosks or wearable sensors, with no human (except the patient) involved.

Edge cases such as these became more ominous as we began to think about how to adjudicate the bet. We came to realize — given that what constitutes *human labeling* is, ironically enough, itself a rather slippery concept — that the best way to determine the winners would be to survey research papers at the Association for the Advancement of Artificial Intelligence conference, the Neural Information Processing Systems conference, and The Web Conference, each year over the next five years (to see the trend rather than two data points), and simply count the number of papers reporting results using a human labeled dataset.

The process shall be to extract a random sample of 60 papers from each conference each year, then assign each paper to one pro and one con bettor. Each bettor will then answer one question per paper after skimming or reading it — does this paper use human labeled data for training or evaluating ML models? In the case of disagreement, the bettors will discuss, and the bookies will adjudicate, throwing out papers with no consensus

If the difference in the expected proportion of papers where the consensus answer is *yes* between responses in 2020 and 2025, satisfies a one-sided two-proportions z-test at the 5-percent level, the pro side wins. If not, the con side wins.

Notes

1. Wearable Technology Market Size Is Expected to Reach USD $57,653 Millions by the End of 2022: Valuates Reports, Bangalore, India. *PR Newswire,* Bangalore, India, February 10, 2020. (www.bloomberg.com/press-releases/2020-02-10/wearable-technology-market-size-is-expected-to-reach-usd-57-653-millions-by-the-end-of-2022-with-a-cagr-of-16-2-valuates).

2. Wearable Technology Market Growing at a CAGR of 15.5% and Expected to Reach $51.6 Billion by 2022: Exclusive Report by MarketsandMarkets. *PR Newswire,* Chicago, IL, July 1, 2019. (www.bloomberg.com/press-releases/2019-07-01/wearable-technology-market-growing-at-a-cagr-of-15-5-and-expected-to-reach-51-6-billion-by-2022-exclusive-report-by).

3. Using Machine Learning to "Nowcast" Precipitation in High Resolution, *Google AI Blog,* posted by Jason Hickey, January 13, 2020. (ai.googleblog.com/2020/01/using-machine-learning-to-nowcast.html).

4. Data Annotation: The Billion Dollar Business Behind AI Breakthroughs, Synced: *AI Technology and Industry Review*, San Jose, CA, August 28, 2019. (syncedreview.com/2019/08/28/data-annotation-the-billion-dollar-business-behind-ai-breakthroughs).

5. Data Annotation Tools Market Worth over $5bn by 2026, *Global Market Insights*, Selbyville, DE, January 28, 2020. (www.gminsights.com/pressrelease/data-annotation-tools-market).

6. Opinion: Artificial Intelligence Hits the Barrier of Meaning, Melanie Mitchell, *The New York Times,* November 5, 2008. (www.nytimes.com/2018/11/05/opinion/artificial-intelligence-machine-learning.html).

7. Why Understanding Ambiguity in Natural Language Processing is a Game Changer, *NewBaseQuid,* May 2, 2016. (netbasequid.com/blog/understanding-ambiguity-in-natural-language-processing).

Mike Schaekermann (mikeschaekermann@gmail.com) is a PhD candidate in computer science at the University of Waterloo's Human-Computer Interaction laboratory and a student researcher at Google Health.

Christopher M. Homan (cmh@cs.rit.edu) is an associate professor of computer science and director of the Lab for Population Intelligence at the Rochester Institute of Technology.

Lora Aroyo (lmaroyo@gmail.com) is a research scientist at Google Research in New York. Prior to joining Google, she worked at the Vrije Universiteit Amsterdam as a full professor in computer science.

Praveen Paritosh (pkp@google.com) is a research scientist at Google Research in San Francisco. Prior to joining Google, he worked as a knowledge scientist at Metaweb Technologies, Inc.

Kurt Bollacker (kurt@longnow.org) is the acting digital research director at the Long Now Foundation. Previously, he has been involved with multiple nonprofit organizations including the Internet Archive in the role of technical director.

Chris Welty (cawelty@gmail.com) is a research scientist at Google Research in New York and is an endowed professor of cognitive systems at the Vrije Universiteit Amsterdam.