O'REILLY®

Real-World Active Learning

Applications and Strategies for Human-In-the-Loop Machine Learning

Ted Cuzzillo

SAN JOSE

LONDON

NEW YORK

SINGAPORE

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect—and merge.

- Learn business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # 15420

Real-World Active Learning

Applications and Strategies for Human-in-the-loop Machine Learning

Ted Cuzzillo

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo O'REILLY®

Real-World Active Learning
by Ted Cuzzillo

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Shannon Cutt
Production Editor: Melanie Yarbrough
Copyeditor: Amanda Kersey
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest
Cover Photo Credit: Jamie McCaffrey

February 2015: First Edition

Revision History for the First Edition
2015-01-21: First Release

978-1-491-91739-8

[LSI]

Table of Contents

Real-World Active Learning

Introduction

The online world has blossomed with machine-driven riches. We don't send letters; we email. We don't look up a restaurant in a guide book; we look it up on OpenTable. When a computer that makes any of this possible goes wrong, we even search for a solution online. We thrive on the multitude of "signals" available.

But where there's signal, there's "noise"—inaccurate, inappropriate, or simply unhelpful information that gets in the way. For example, in receiving email, we also fend off spam; while scouting for new employment, we receive automated job referrals with wildly inappropriate matches; and filters made to catch porn may confuse it with medical photos.

We can filter out all of this noise, but at some point it becomes more trouble than it's worth—*that* is when machines and their algorithms can make things much easier. To filter spam mail, for example, we can give our machine and algorithm a set of known-good and known-bad emails as examples so the algorithm can make educated guesses while filtering mail.

Even with solid examples, though, algorithms fail and block important emails, filter out useful content, and cause a variety of other problems. As we'll explore throughout this report, the point at which algorithms fail is precisely where there's an opportunity to insert human judgment to actively improve the algorithm's performance.

In a recent article on Wired ("The Huge, Unseen Operation Behind the Accuracy of Google Maps," 12/08/14), we caught a glimpse of the

massive active-learning operation behind the management of Google Maps. During a visit to Google, reporter Greg Miller got a behind-the-scenes look at Ground Truth, the team that refines Google Maps using machine-learning algorithms and manual labor. The algorithms collect data from satellite, aerial, and Google's Street View images, extracting data like street numbers, speed limits, and points of interest. Yet even at Google, algorithms get you to a certain point, and then humans need to step in to manually check and correct the data. Google also takes advantage of help from citizens—a different take on "crowdsourcing"—who give input using Google's Map Maker program and contribute data for off-road locations where Street View cars can't drive.

Active learning, a relatively new strategy, gives machines a guiding hand—nudging the accuracy of algorithms into a tolerable range, often toward perfection. In *crowdsourcing*, a closely related trend made possible by the Internet, humans make up a "crowd" of contributors (or "labelers," "workers," or "turkers," after the Amazon Mechanical Turk) who give feedback and label content; those labels are fed back into the algorithm; and in a short time, the algorithm improves to the point where its results are useable.

Active learning is a strategy that, while not hard to deploy, is hard to perfect. For practical applications and tips, we turned to several experts in the field and bring you the knowledge they've gained through various projects in active learning.

When Active Learning Works Best

The concept of active learning is simple—it involves a feedback loop between human and machine that eventually tunes the machine model. The model begins with a set of labeled data that it uses to judge incoming data. Human contributors then label a select sample of the machine's output, and their work is plowed back into the model. Humans continue to label data until the model achieves sufficient accuracy.

Active learning works best in cases where there's plenty of cheap, unlabeled data, such as tweets, news articles, and images. While there's an abundance of content to be classified, the cost of labeling is expensive, so deciding *what* and *how much* to label are key considerations. The trick is to *label only the data that will have the greatest*

impact on the model's training data and to feed the classifier an appropriate amount of accurately labeled data.

Real-World Example: The Spam Filter

Imagine a spam filter: its initial work at filtering email relies solely on machine learning. By itself, machine learning can achieve about 80–90% accuracy. Accuracy improves when the user corrects the machine's output by relabeling messages that are *not* spam, and vice versa. Those relabeled messages feed back into the classifier's training data for finer tuning of future email.

While one method may be to let the user label a random selection of the output (in this case, email), that takes a lot of time and lacks efficiency. A more effective system would use a classifier that estimates *its own certainty* of each verdict (e.g., spam or not spam), and presents to the user *only the most uncertain items*. When the user labels uncertain items, those labels are far more effective at training the classifier than randomly selected ones. Gradually the classifier learns and more accurately determines what is and is not spam, and with periodic testing continues to improve over time.

Real-World Example: Matching Business Listings at GoDaddy

A more complex example of active learning is found at GoDaddy, where the Locu team's "Get Found" service provides businesses with a central platform for managing their online presence and content (including address, business hours, menus, and services). Because online data can be riddled with inconsistencies (e.g., "Joe's Pizza" might be listed on "C Street" or "Cambridge St." or may even be listed as "Joe's Italian"), Get Found provides an easy way for businesses to implement a consistent presence across the web. While inconsistencies such as "Joe's Pizza" being listed as "Joe's Italian" could easily stump an algorithm, a human labeler knows at first glance that the two listings represent the same restaurant. Adam Marcus, the director of data on the Locu team, notes that a wide range of businesses, including restaurants, flower shops, yoga studios, and garages, rely on products such as Get Found for this type of business-listing service. To identify listings that are describing the same business, the Locu team allows algorithms to automatically match simple cases, like "Joe's Pizza" and "Joe's Pizzas," but reaches out to humans on

CrowdFlower for more challenging cases like "Joe's Pizza" and "Joe's Italian." This active learning loop has humans fill in the details and retrain algorithms to perform better in the future.

Real-World Example: Ranking Top Search Results at Yahoo!

Another real-world example of active learning involves the ranking of online search results. Several years ago at Yahoo!, Lukas Biewald, now CEO of the crowdsourcing service provider CrowdFlower, wanted to improve Yahoo!'s ranking of top search results. This project involved identifying the top 10 search results amongst millions. Biewald's team realized that the simplest strategy wasn't necessarily the best: rather than labeling a uniform sample from the millions of results (which would include pages that are *not* relevant), his team chose to use only the top results as training data. Even so, this had some bad outcomes: the top picks were a misleading sample because they were based on the algorithms' own work. For instance, based on the top results, the classifier might assume that a machine-generated page with "energy savings" repeated a thousand times is more relevant than another page with just a few mentions, which is not necessarily the case.

So how was the classifier to know which results belonged in the top 10 and which did not? The classifier had never seen many of the search results that were deep in the web and not included in the test data. So Biewald and his team addressed this by labeling and feeding back some of these uncertain cases to the model; after some repetition of this process, the model significantly improved its results.

Where Active Learning Works Best

Is crowdsourcing worth the trouble and expense? An experiment referenced by Biewald in his talk on active learning at the Strata Conference in February 2014 bears the dramatic result. The task was to label articles based on their content, identifying whether they covered baseball or hockey. Figure 1 shows the efficiency of two classifiers: one classifier (represented by the dotted line) worked with 40 randomly selected labels that were not generated via active learning; it achieves about 80% accuracy. The other classifier (represented by the solid line) worked with just 20 labels that were generated via active learning; it achieved the same accuracy, with only half the

labels. Biewald points out that the rise in efficiency (as shown in Figure 1) is still rising at the end, showing that there's a demand for even more labels.

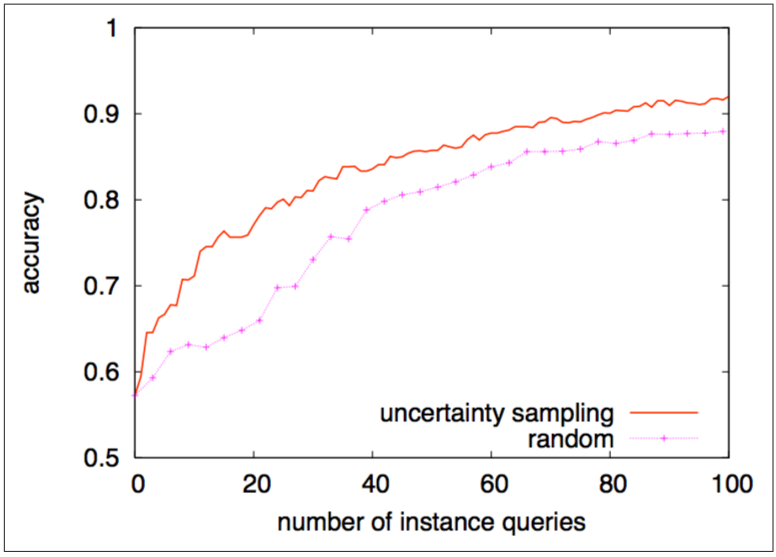

Figure 1. Comparing accuracy of selection methods: the dotted line represents randomly selected data, not generated via active learning; the solid line represents data generated via active learning. (Settles '10)

Basic Principles of Labeling Data

Imagine a new email classifier that has just made its first pass on a small batch of data and has found some email to be classified as spam and some as valid. Figure 2 shows red dots representing spam and green dots representing valid email. The diagonal line in-between represents the division between what is spam and what is not, and indicates the border between one verdict and another. In the figure, dots close to the center line indicate instances where the machine is *least certain* about its judgment.

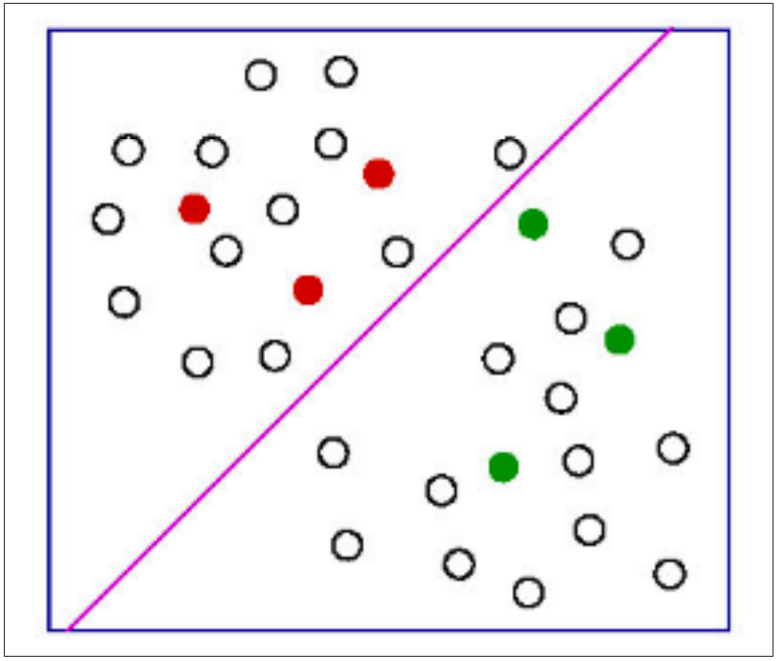

Figure 2. The colored dots near the center line represent judgments the machine is least certain about

At this point, the key consideration is *which dots* (in the case of spam, which email) should be labeled next in order to have maximum impact on the classifier. According to Lukas Biewald of CrowdFlower, there are several basic principles for labeling data:

- **Bias toward uncertainty.** Labels have the most effect on the classifier when they're applied to instances where the machine is the most uncertain. For example, a spam email classifier might confidently toss out an email with "Viagra" in the subject line, but it's less confident when a longtime correspondent uses the word.

  The machine's least certain judgments are likely to be based on content that the model knows little or nothing about. In instances where an email seems close to a known-good sample but also somewhat close to a known-bad sample, the machine is much less certain than in instances where an abundance of training data make the verdict clear. You'll make the biggest impact by labeling data that gives the classifier *more confidence,*

rather than labeling data that merely affirms what it already knows.

- **Bias toward ensemble disagreement.** A popular strategy in active learning is to use multiple methods of classification. Using multiple methods is an opportunity to improve the classifier because it allows it to learn from instances where the results of your different methods disagree. For example, a spam classifier may label an email with the words "Nigerian prince" as spam, but data from a second classifier might indicate that "Nigerian prince" is actually a long-term email correspondent; this helps the first classifier judge correctly that the message is valid email.
- **Bias toward labels that are most likely to influence the classifier.** Classifiers are generally uncertain about how to label data when random or unusual items appear. It helps to label such items because they're more likely to influence the classifier than if you were to label data that's similar to other, already labeled data.

  For instance, when Biewald's team at Yahoo! set out to improve the search engine's ranking of top 10 results, the algorithm showed odd results. It was so confused that it included web pages in the top 10 that were completely irrelevant and not even in the top 1,000. The team showed the classifier labeled data from the types of irrelevant pages that were confusing it, and this produced dramatically better results.
- **Bias toward denser regions of training data.** The selection of training data should be corrected in areas where the data volume is greatest. This is a challenge in part brought on by the other, previously mentioned principles, which usually result in a bias toward outliers. For example, labeling data where the algorithm is uncertain skews its training toward sparse data, and that's a problem because the most useful training occurs where data density is highest.

Beyond the Basics

For even greater accuracy, slightly more advanced strategies can be applied:

- **Active cleaning.** Look for the training data with the largest error. The one data point far off the norm (as in Figure 3) is by far the most influential data you can show the model. If the data is correctly labeled, it will teach the model about outliers. If it's incorrectly labeled (a common occurrence), it should be taken out.
- **Active cleaning using hand curation**. The more attention given to the data that goes into the model, the better the classifier will work. Look for "edge cases" to label by hand, and show the classifier as training data.

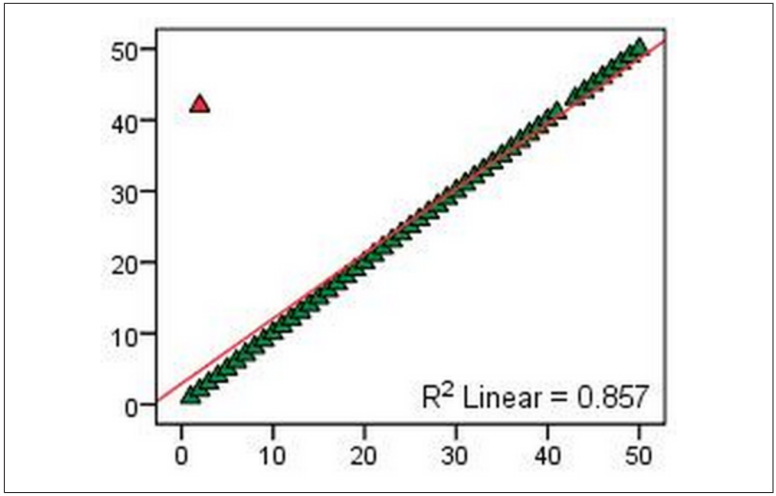

Figure 3. The red triangle represents data whose correct labeling would have good effect on accuracy of the classification model. (Lukas Biewald)

"Gold Standard" Data: A Best Practice Method for Assessing Labels

Patrick Philips, a crowdsourcing expert and data scientist at Euclid Analytics, describes a best practice method in active learning: formulating "gold standard" data. Before any crowd of contributors sees the data for a job, Philips spends one to two hours scoring a small subset of the data by hand. He adds that "gold standard" data can also be extracted from contributor-labeled data when there is strong agreement (on the labels) among the contributors. Creating and managing a set of "gold standard" data (e.g., four to five exam-

ples for each class of data) provides a standard for judging labels that come in from contributors; this can be done in several ways. First, it's an up-front filter: each worker's contributions are automatically compared with the "gold standard" data to measure understanding, ability, and trustworthiness for the job. Second, using the "gold standard" data allows for ongoing monitoring and provides a means to train and retrain workers and to offer corrections to improve performance. Third, the "gold standard" data allows you to score worker's accuracy and automatically exclude work that falls below a certain percentage of accuracy; in addition, this provides the opportunity to discover problems with the data itself.

Elements of "Gold Standard" Data

"Gold standard" data is the standard by which all other data in one application can be measured.

The Benefits

Setting up your own "gold standard" data gives you an overview of your data and helps you decide what labels you need. It also helps you avoid designing unhelpful/bad labels, which can result in severe mislabeling and problems later on. Your early work on a subset of "gold standard" data can save you time and money later.

Tips

- Start with just a small subset of your data, perhaps just four or five examples from each class.
- Use your "gold standard" data to measure the performance of each contributor so you know when to retrain workers. When a contributor's score falls below 70% accuracy, exclude his work and retrain.
- Continually review your "gold standard" data to ensure it's as accurate and useful as possible so that it maintains its purpose.

Managing the Crowd

The "crowd" solves a problem and has definite usefulness in active learning, but it also has a flip side: "Humans will sometimes give you wrong answers," explains Adam Marcus of GoDaddy. Mislabeled items sometimes result from boredom, perhaps even resentment.

Also, some questions might be unintentionally misleading, or the contributors might have raced through with little attention to the questions and answers. Whatever the cause, wrong answers can badly skew training data and take hours to correct.

One simple solution, explains Marcus, is to ask questions not once but several times. Redundant questioning establishes confidence in the labeling. An item that's labeled in one certain way by four different contributors is far more likely to be correctly classified than one that's labeled by just one contributor.

A more complex but beneficial solution is the creation of worker hierarchies. A hierarchy allows more than simple, redundant labeling—it sends items with low certainty up the ladder to more trusted workers. Hierarchies rely on long-term relationships with contributors. To enable hierarchies, organizations can recruit through companies such as oDesk, Elance, and other online marketplaces with a plentiful supply of customer-rated candidates. As workers become known and trusted, they're given more work and asked to review other workers. They might also receive recognition and bonuses, and they can even move up to more interesting tasks and even manage projects. "We have reviewers running jobs, doing way more interesting work than they were hired for," says Marcus. "These incentives give contributors a clear sense of upward mobility."

Contributors whose work falters, on the other hand, are given less work. The weaker their performance becomes, the more scrutiny they receive, and their work volume is incrementally reduced.

The hierarchy system also improves training. At GoDaddy's Locu team, a new worker recruited through oDesk, or other such agency, would have a week of training and practice; his work cleaning up the classifier's output would go first to a trusted worker, whose review would go back to the new worker. According to Marcus, within just a few weeks, the new recruit's work improves.

A Challenge: Overconfident Contributors

Redundant questioning and "gold standard" data are methods for helping to address a common problem identified by crowdsourcing expert Patrick Philips: overconfidence among contributors. Confidence bias, a phenomenon well known to psychologists, is the systematic overconfidence among individuals of their own ability to complete objective tasks accurately.

In one experiment by Philips, as described in his 2011 blog post "Confidence Bias: Evidence from Crowdsourcing," individuals were asked to answer a set of standardized verbal and math-related questions and to identify how confident they were in each answer. The difference between each individual's average confidence and actual performance was an estimate of confidence bias. Of the 829 people who answered 10 or more questions, *more than 75%* overestimated their abilities.

Philips found that confidence bias rises with a person's level of education and age, and also with the number of questions they answer accurately. In his experiment, US contributors were much more accurate *and* slightly more biased than the average. Individuals from India had average accuracy, but much higher confidence. In looking at gender, Philips found that women were more accurate and less biased than men.

More Tips for Managing Contributors

- **Pick your problems carefully.** Think about the problem you're trying to solve and structure it in a way that makes it easy to get meaningful feedback.
- **Pick a *solvable* problem (have your team try it first).** According to Philips, "If you can't do it on your team, it's probably not a solvable problem." If you find yourself with a seemingly unsolvable problem, consider using a parallel problem that can be solved more easily.
- **Make sure the task is clearly defined.** Whatever you want your labelers to do—test it with your team first. If your team has trouble, the labelers certainly will; this gives you the chance to make sure your task is clearly defined.
- **Use objective labels that reasonable people agree on.** In Philips' previous role at LinkedIn, the company set out to classify content in its newsfeed; he did this by having crowdsource workers label content using only a handful of descriptors, such as "exciting," "exhilarating," "insightful," and "interesting." The team sent out about 50,000 articles for labeling, and the task seemed easy enough until the labeled data came back. "It was a mess," says Philips. No one agreed internally on what each of these labels meant before sending the task out, so they couldn't agree whether the results were accurate when they came back

either. Results improved when the team switched to more objective labels, with a four-tier selection that accounted for overall quality based on coherence, spelling, and grammar; a second selection indicated the general content, such as nonfiction, fiction, and op-ed.

When to Skip the Crowd

You may discover that you need no crowd at all because the answer is right in front of you. In one project at LinkedIn, Philips' team wanted to refer people to job postings appropriate for their level of experience. The team hoped to have crowdsource workers classify members into one of three categories: individual contributor, manager, or executive manager. Though a seemingly straightforward task, it proved quite difficult. For starters, job titles vary wildly among companies; and even when they are the same title, the size of the company impacts the role itself; for example, the vice president at Google may not belong in the same seniority category as the vice president at a startup.

Other more indicative data, such as salary, wasn't available, so the team tried proxies. They looked at the number of years since graduation, which was useful, though not enough. Other proxies included endorsements within a network, the seniority of immediate connections, and maps that show whose profiles members have viewed.

Eventually, the team at LinkedIn found a solution based on data they already had: when LinkedIn members write recommendations, they explicitly indicate their relationship to that person—peer, manager, or direct report. With help from millions of LinkedIn recommendations, the team developed a system to rank employees by seniority within a company.

"Crowdsourcing is a great tool, but it's not without its challenges," says Philips. "Definitely look around first; you may already have the data that you need."

Expert-level Contributors

In cases where active learning requires expert-level knowledge or educated judgment, the recruitment and management of labelers becomes much more complex. This occurs when the task graduates

from simple, accurate assessments (such as spam or not, and human face or not), into tasks that only an expert crowd can perform.

In addition to finding the expert crowd, another challenge is that when you do find expert labelers, their labels can be wrong or random—and the non-expert would never know it. (Only specialists can distinguish, for example, an American tree sparrow from a white-crowned sparrow.)

Panos Ipeirotis, a leading researcher in crowdsourcing and associate professor at New York University, recalled one such instance when he asked contributors to give the name of Apollo astronaut Neil Armstrong's wife. The choices included "Apollo," "Gemini," "Laika," "None of the above," and "I do not know." Only one of these options is likely to have been a human name (Laika) and was actually the name of the dog sent into space by the Soviet Union, yet it was the answer chosen by some aspiring "experts." In these cases, "contributors are choosing an answer that is plausible," says Ipeirotis, "because they want to convey as much information as they can and don't want to admit that they don't know." (Ipeirotis found that in retrospect, replacing "I do not know" with "Skip" proved to be a much better choice.)

What complicates the matter is that a plausible-but-wrong answer can't be easily detected by a machine algorithm. If five people give the same plausible-sounding answer, for example, the algorithm becomes confident based on that inaccurate data, resulting in a bad classification that's reinforced by the workers' collective agreement.

In short, the best labelers are those who *admit when they don't know* the answer.

How to Find the Experts

For tasks that require expert knowledge, the usual crowdsource marketplaces offer little support; the challenge is that they usually cannot supply enough contributors with specialized knowledge. Ornithologists, historians, and fluent speakers of other languages, such as Swahili, Sicilian, or Southern Tujia, for example, all have to be recruited differently.

One promising method of expert recruitment, Quizz, is described in the research paper "Quizz: Targeted Crowdsourcing with a Billion (Potential) Users" by Panagiotis G. Ipeirotis and Evgeniy Gabrilo-

vich. The authors found that the best way to find subject-matter experts was to lure them into demonstrating their knowledge.

Ipeirotis and Gabrilovich began their experiment with eight quizzes that they placed as ads on popular websites. Each quiz challenged passersby with a question; for example, one question might be, "What is a symptom of morgellons?" (Those with medical knowledge know that morgellons involves delusions of having things crawling on the skin.) Each quiz question offered several plausible choices, as shown in Figure 4, and anyone who offered an answer learned instantly whether it was correct.

Figure 4. A sample Quizz question

In the background was an algorithm created by Ipeirotis and Gabrilovich that kept score and judged the expertise of each respondent. Participants who were judged to be sufficiently knowledgeable were invited to go further, and Quizz continued to measure their total contribution and the quality of their results.

In addition to scoring participants, the Quizz algorithm also used advertising targeting capabilities to score the websites where the ads

appeared. Sites that produced too few qualified candidates were dropped, as a way to continually optimize results. The algorithm also recorded the "origin" sites of those who gave good answers and began recruiting on those sites more heavily. For example, the recruiting algorithm quickly learned that consumer-oriented medical websites, such as Mayo Clinic and Healthline, produced many qualified labelers with medical knowledge, while ads on medical websites with a professional audience did not manage to attract contributors with sufficient willingness to participate.

Participants who clicked on an ad and answered the quiz questions constituted a "conversion" that was tallied by the algorithm. At the time Ipeirotis and Gabrilovich wrote their paper in 2014, the Quizz application began with a 10–15% conversion rate, which, over time, rose to a 50% conversion rate—by simply giving feedback to the advertising targeting algorithm.

Managing Expert-level Contributors

A key consideration in managing experts is how to get the most out of each contributor. According to Ipeirotis, the trick is to balance two types of questions: one type ("calibration") estimates the contributor's knowledge, and the other type ("collection") collects their knowledge. Balancing these two types of questions allows you to sustain the stream of collected knowledge as long as possible and explore the person's potential to give more.

The optimal balance of these two types of questions (calibration versus collection) depends in part on each contributor's recent behavior; it also depends on her *expected* behavior, which is based on that of other users. For example, the user who shows signs of dropping out is likely to be steered toward a proven "survival" mix: since contributors are motivated mainly by the contribution of good information, the "survival" mix lets them have more questions they are likely to answer correctly, followed by prompt acknowledgement of their work.

Payment is another factor to consider in managing expert-level contributors. In their research, Ipeirotis and Gabrilovich found that *paid* workers not only cost more, they often produced *poorer quality* data and were less knowledgeable than those who were unpaid. Ipeirotis and Gabrilovich describe an experiment in which one selection of contributors were paid piecemeal rates, with bonuses

based on scores; this group dropped out at a lower rate than a selection of unpaid workers. However, while the paid workers were staying on, *they were submitting lower-quality answers than those who were unpaid*. Interestingly, offering payment was not linked with high-quality answers; payment simply sustained workers, presumably in cases where unpaid workers, lacking the satisfaction of offering high-quality answers, would have given up.

Recruiting Expert-level Contributors

Recruiting contributors with expert knowledge is different from recruiting everyday crowdsource workers. For tips, we turned to researchers Panagiotis G. Ipeirotis and Evgeniy Gabrilovich.

The Essential Strategies

- The best approach to recruiting experts is to encourage them to lend their expertise.
- The best contributors are unpaid. Unlike everyday crowdsource workers, Ipeirotis and Gabrilovich found that expert contributors produce information in *more volume* and with *higher quality* when they weren't paid. The best motivator was the contribution of good information.
- By whatever means you are recruiting experts, keep track of your success: monitor the quality and quantity of recruitments, and modify your efforts accordingly.

A Real-World Example: Expert Stylists + Machine Learning

Expert contributors can do more than identify birds and medical symptoms. In one application, customers actually seem to trust the experts *more* than they trust themselves. Stitch Fix, an online personal styling and shopping service for women—relies on both expert contributors and machine learning to present customers with styles that are based on their own personal data.

The process at Stitch Fix begins with a basic model, an estimate of what customers will like based on their stated preferences for style and budget. Then, the model evolves based on information from actual purchases. Notably, a customer's model may be disrupted. For

example, the model may find that the customer who gave her size as 12 actually purchases items at a size 14 or that clothes she describes as "bohemian chic style" are actually what most people would call "preppy"; she may also buy clothes that reveal a higher budget than the one she gave.

Handling these types of disruptions and matching stated with actual preferences are the biggest challenges at Stitch Fix, says Chief Algorithms & Analytics Officer, Eric Colson. In addition, the lack of an industry standard for clothing sizes adds to the problem; for example, a size 10 at one store could be a size 6 at another. Customers also give bad data, such as their aspirational size (i.e., one that anticipates weight loss), rather than their true size. They may also misunderstand industry terms, confusing size with fit, for example.

Aside from customer-based data, a second set of data describes each item of clothing in fine-grain detail. Stitch Fix's expert merchandisers evaluate each new piece of clothing and encode its attributes, both subjective and objective, into structured data, such as color, fit, style, material, pattern, silhouette, brand, price, and trendiness. These attributes are then compared with a customer profile, and the machine produces recommendations based on the model.

But when the time comes to recommend merchandise to the customer, the machine can't possibly make the final call. This is where Stitch Fix stylists step in. Stitch Fix hands off final selection of recommendations to one of roughly 1,000 human stylists, each of whom serves a set of customers. Stylists assess unstructured data from images and videos of the merchandise and from all available customer comments (e.g., "I need clothes for a big meeting at work."). They may even reach outside of the machine's recommendations and use their own judgment to make final selections for which pieces will go to the customer. Before a shipment goes out, the stylist scrutinizes each piece to see how they look together and may even explain the selections to the customer.

According to Colson, occasional "smart risk" is also built in to the algorithm. Stitch Fix deliberately injects randomness to add value; to stay completely safe within a narrow range of customer preferences would truncate the possibilities. "On our 10th or 11th shipment," says Colson, "that's when you need to start mixing it up." A school teacher who dresses conservatively during the week, for example, probably has enough conservative clothing. "What's it going to take

to create a meaningful relationship?" asks Colson. "It might be to take her on the next part of her journey."

Stylists work anywhere that has Internet access, and though they are paid hourly, they often report intangible benefits, such as the satisfaction of happy customers.

Machines and Humans Work Best Together

Futurists once dreamed of machines that did everything, all guided by an unseen autopilot. Little did these visionaries know that the autopilot can do so much more with help from a crowd.

Active learning has put machines hand in hand with humans, and the success so far hints at huge potential. If this duo can choose clothing, thwart email spammers, and classify subtly different images, what else could it do?

About the Author

Ted Cuzzillo has covered technology as a journalist and industry analyst for 25 years. Topics have included telecommunications, computer networking, and environmental technology. Most recently, since 2007, he has been a regular contributor to business intelligence media that include TDWI "BI This Week," Information Management, Smart Data Collective, and his popular blog, Data-doodle. Current interests include the practice of storytelling and collaboration and technology that supports it.