How a Citizen Scientist Will Win a Nobel Prize

Note: Read a fancier PDF version of this essay, complete with footnotes that academics like.

Summary

The ongoing democratization of data is helping science, but there is another, even more interesting revolution on the horizon. The dream of "Big Data" is that unprecedented insights are possible from very large data sets, often collected passively from ubiquitous devices or actively by "citizen scientists". Thanks to a deluge of data, hypothesis *testing* has become easier than ever, exposing another bottleneck: hypothesis *finding*.

"Open Data" initiatives help break barriers on how that data can be analyzed, but the real opportunity happens when the tools for *idea generation* and analysis are opened broadly. Properly enabled, the new world of data collection comes with (1) highly-motivated and enthusiastic subjects, who can (2) participate in the analysis and (3) generate and test far more hypotheses than has been feasible for trained experts. Future scientific researchers will think of themselves as designers and motivators for crowd-sourced projects, where research subjects are no longer passive data suppliers, but important and equal partners in the collection and creation of new knowledge.

The Data Deluge

Experimental data has always been the lifeblood of science, and new information technologies are exponentially increasing both the volume and kinds of data availability.

Consider the famous 1967 Steven Millgram experiment (cited best by Jeff Leek) to uncover the degrees of separation between two people in the US, giving rise to the popular internet meme "Six degrees of Kevin Bacon". An experiment that once required manual, expensive mailing of letters can now be reproduced in minutes on a much larger scale, with astounding results, an impressive feat that is repeated so regularly now as to be commonplace.

Much of this deluge comes from low-cost, ubiquitous smartphones that make new types of mobile data collection not just possible, but passively easy, often at no cost to the subjects or the experiment designers. Combined with the explosion of wearable computing devices like Fitbit or Apple Watch, the billions of smartphone-carrying people worldwide are a previously-unimaginable rich source of data about health (heart rate, exercise, sleep, glucose, etc.), environment (barometer, air quality, temperature), images, and much more to come. Furthermore, the ubiquity and ease of use makes many other scientific experiments easier to conduct, with apps that give users step-by-step guidance, notifications, and even built-in bar code reading capabilities that dramatically expand the number of compliant participants in any study involving human subjects.

Even the laborious process of finding properly-vetted research subjects has become easier, thanks to crowd-sourcing on powerful, widely-disseminated platforms like Apple's ResearchKit.

Meanwhile, crowd-funding platforms like Indiegogo, Kickstarter, and many others offer researchers a double blessing, with subjects who not only seek to join the study themselves, but willingly pay for project costs. One leading microbiome genetics company uBiome, for example, was able to bootstrap from nothing to an initial successful $350,000 fund-raising project into building and staffing its own multinational CLIA-certified lab.

It's not just *new* data that is now being usefully collected, but often *old* data is just as useful when combined into a mash-up, or reprocessed with new tools. Data collected for one purpose can often be useful for another, such as when the activity tracker company Jawbone detected an earthquake based on when people woke up.

The data explosion breeds opportunities for new technologies, especially in data science. For the analysis step, for example, powerful cloud computing hardware, deep learning algorithms based on repurposed GPU technology, and many other accumulated advances in the analysis of Big Data, have given scientists new tools to keep up with the flow of new data. Many of these techniques are already transforming the useful insights that can be obtained from vast reservoirs of existing biological data.

As the relative cost of experimental data declines, other phases of the scientific pipeline become more critical.

Generating and Pursuing Hypotheses

While full-time specialists have access to an unprecedented volume of data and techniques, they will always be limited in the number of new hypotheses they have time to pursue. Better time management and processes as well as increased budgets for staffing and education can help, but a finite team will always face constraints on the number and variety of questions it can think to ask. Because time is the most precious resource for any team, tradeoffs are inevitably made that limit the number of hypotheses.

How do we decide which questions to ask? The Millgram experiment, for example, is easy to perform by anyone empowered with the tools and expertise. But what other fascinating discoveries were never pursued simply because the questions were never asked by the people empowered to answer them?

Worldwide, there are about 7 million publishing scientists and staff (*2015 UNESCO data*), each devoted to answering only a few scientific questions. Even if the time and expense of data collection were eliminated, there are practical limits on the number of new insights of which each is capable. Furthermore, the inevitable specialization of each field puts another limit on the breadth of questions that can be asked, even by these professionals.

It's not that the new questions don't get asked. Every day, people worldwide confront puzzling new situations for which a scientific answer would be beneficial, but they often lack access to the necessary tools and data. In many cases, people discover ad hoc solutions that would could be proved scientifically if the proper resources were applied.

Health- and wellness related questions are among the most common examples. A person notices a symptom or a possible correlation between some action and a result: "I seem to sleep better after drinking a glass of warm milk", or "Diet soda seems to make me gain weight", or "The left lane always seems to be the slowest". Curious people make these sorts of observations multiple times each day, but virtually none of them receive consideration as serious scientific hypotheses, and except for the ones that attract the attention of a practicing scientist, rarely if ever result in analysis by a peer-reviewed publication.

Many of the people asking these questions are highly motivated to find answers. They, or a close friend or family member, may suffer from a serious ailment for which there is currently no treatment. An entrepreneur may see a business opportunity, perhaps in an area for which he or she already possesses most of the non-scientific skills or resources necessary. A specialist from one scientific discipline may wish to apply the results of a new technology or scientific insight to another field.

Millions of people already participate in many forms of personal data gathering through devices like Apple Watch or Fitbit, and many are also active in movements like Quantified Self or Biohacking, where they meet other like-minded individuals to discuss their data and discuss analysis tips. Some of the analysis can be quite sophisticated, such as decades-long longitudinal studies of sleep, food consumption, activity, financial transactions, and much more.

One early Quantified Self enthusiast, the late University of California Berkeley psychologist [Seth Roberts(http://sethroberts.net/), published many examples where his own self-experimentation gave him new actionable insights into important aspects of health, behavior, and more. Many of these insights are counter-intuitive (seeing faces in the morning improves mood, tasteless calories decrease appetite), but because the experiments are cheap and easy, many more of them can be performed.

Limitations of Citizen Science

Importantly, despite the passion of those involved, rigorous analysis will prove the vast majority of these speculations to be scientifically untenable. But like the deluge of data that is transforming the rest of science, if proper tools can be found to harness and improve these questions, it could bring a vast new source of interesting hypotheses and rigorous experimentation.

Data quality The limitations of self-reported data, and the hazards of relying on it, are well-documented. If the data is collected without regularized protocols, the results may be scientifically difficult to compare. A nutrition study that relies on food diaries, for example, may elicit different degrees of recall, different

definitions about portion size and many other factors, including simple forgetting or even cheating. Rigorous, professionally-designed studies take numerous precautions to make the data consistent.

Furthermore, by definition the kind of person who collects data about himself or herself is generally not representative of the population being studied, an obvious source of sample bias.

Crowd-sourced collection will be especially plagued by compliance issues. The passion that attracts a person to a study may fade, leaving the contributions incomplete.

But note that many important initial discoveries don't have to be scientifically or statistically rigorous to be worthwhile. The original research showing the value of fecal transplants to treat C Diff infections was based on an n=2 study. Reconstruction of the 1918 Spanish Flu virus was made possible thanks to a key finding by Johan Hultin, a 72-year-old retired pathologist. Any intelligent and curious person can make important contributions to science if given access to the appropriate tools.

Ethical Issues Informed consent is fuzzy when it's a self-experiment. Did the subject really understand the risks, or was he/she tricked into something by a nefarious third party? The highly competitive diet and supplements industry, for example, is often tempted to use hyperbole that may not be justified by the science. The First Amendment gives wide latitude to book publishers and authors, even to remedies that can be dangerous.

While there is no question that unsuspecting individuals can be duped into believing or doing odd things, especially if there is a financial incentive, scientists themselves are not immune. The peer review process, not to mention the many years of education and training provides formidable protection against scientists making some errors, but ultimately even professionals rely on the corrective properties of the scientific process itself.

Incidentally, it may be important to mention that the original term "snake oil" began as a not-so-subtle disparagement of Chinese immigrants whose traditional medicines competed against the established medical practitioners of the day, whose own remedies now look just as dubious. Rigorous scientists today should ask if their objections to citizen science stem from a similar set of prejudices against people and ideas that are vilified because they are different, not because they are scientifically invalid.

Professional vs. Amateur

The term "citizen scientist" is often used interchangeably with "amateur", implying untrained or unskilled. Despite the well-documented and shameful ignorance of scientific and statistical reasoning among much of the public, there remain substantial numbers of science-friendly people who *do* appreciate properly-grounded research and are eager to apply their skills to other domains.

Many "amateurs" are professionals in other disciplines who have simply not chosen to specialize professionally in a particular branch of science. There are many well-trained doctors, engineers, and others with formidable credentials whose advanced skills can bring world-class insights outside their specialization. The US Census estimates that more than 50 million Americans have degrees in STEM (science, technology, engineering, math). Engineers, market researchers, physicians, clinicians, and many other fields boast people highly trained in a scientific discipline, and whose skills can be applied outside their "day job", if sufficiently empowered.

Meanwhile, as professionals in any discipline understand, science at the cutting edge is messy, with many large gaps in knowledge that aren't fully appreciated by those without day-to-day contact with a laboratory. The so-called "crisis of reproducibility" that has affected the social sciences (and even some hard sciences) looms especially large in fields like health and nutrition, where mainstream consensus is regularly debunked. In many cutting-edge areas of science, such as the study of the microbiome, early results that seem to show one thing breed excitement that spills into the popular press, only to be overturned as additional data indicated another. (Examples include the idea of the Firmicutes/Bacteroidetes role in obesity, the proposal of the existence of enterotypes, the claim that bacterial cells outnumber human cells by ten to one, etc.)

The internet explosion has already blurred the distinction between professional and amateur in other fields. Restaurant critics, for example, traditionally were highly-skilled, often with culinary backgrounds and long experience rating and reviewing food. Yelp and its many competitors changed that, with results that cover far more locations, are updated more regularly, and can be much more customizable to each person than would be feasible by any other means. The situation is similar in countless other areas where the traditional role of experts has given way to crowd-sourced recommendation sites that offer guidance on

car or home purchases, movies, music, travel, and much more. While there remains a role for experts in each of these areas, consumers enjoy far more comprehensive coverage than was feasible before the wide availability of these services.

A similar transformation is happening in the practice of science.

The welcome trend toward open data and open publishing is already bringing more rigor to the scientific process. Much has been made of a graduate student who noticed an Excel error while trying to reproduce a highly-cited economics paper. Unlike the peer reviewers and others who hadn't caught the error, the student's motivation was simple self-education, like countless other non-professionals who just have questions of their own and want to pursue answers.

It can be humbling, but researchers who share data with their subjects can often learn new techniques and insights they wouldn't have learned in the cloister of the lab. At Stanford's Relman Lab, for example, the principals behind a multi-year study of the microbiome were motivated to introduce a new research perturbation partly because one of their (amateur) subjects had been self-experimenting with the effects of potato starch on the microbiota.

Empowering More People to Conduct Science

Nobel Prize-winning physicist and author Richard Feynman noted that "in science we are not interested in where an idea comes from". Especially in a mature field like physics, he adds, contrarian and even whacky perspectives are welcomed, because a long track record of steady zig-zag progress has given researchers the self-confidence to pit their existing tools against new data.

Crowd-sourced science projects work by finding subjects who are motivated to support the research by (1) the "perk" of seeing their own raw data at the cutting edge of a new field and (2) an altruistic desire to advance science for its own sake.

The missing pieces that can unleash new hypothesis generation include:

1. Access to the raw data, so that motivated people can go beyond the initially-anticipated purpose of the collectors.

2. Easy-to-use tools that let people explore and analyze their own data.

3. Forums to let people share and contribute to further analysis.

While the first of these, access to the data, is becoming more common as people and institutions demand it, the other requirements of hypothesis generation have not received enough attention.

All scientists should encourage non-specialists to share and analyze their own data, not simply submit it to professional experts.

Making raw data available is the first step; encouraging and enabling *analysis* and *sharing* is the game-changer.

The changing role of professional scientists: researchers as coaches

If the data and tools are available to anyone, and every motivated individual is empowered to ask and answer their own research questions, what is the role of a professional scientist?

There will always be a role for the traditional, specialized "bench" scientist who toils in a lab (or on a computer) working with complex data to find insights from complex phenomena. Tightly integrated teams working on a common problem will always be able to solve challenges that are beyond the abilities of a lone or disconnected individual.

But it's useful to return to the example of Yelp to see one way the profession can change. Deeply knowledgeable restaurant and other product critics are still reviewing, and there remains a place for centralized arbiters of standards, safety, and coordination, whether from government regulators (health inspectors and licensing) or market-driven private or public companies (e.g. Yelp, Consumers Union, Cochrane Review). As in the past, there remains a place for full-time experts, but many of these are associated with other activities such as education, publishing, or philanthropic organizations.

Even today, the most prominent researchers have busy schedules filled with teaching, consulting, and presenting and reviewing the work of others, either within their own labs or outside, as peer reviewers and conference organizers. Working with an empowered general public should become just another source of inspiration for their work.

Conclusion

The rapid explosion of new experimental data and data science tools to explore it has exposed a new opportunity to alleviate one of the biggest obstacles to scientific discovery: the generation and exploration of interesting new hypotheses. Whereas in the past, the means to uncover new knowledge was often limited to professionals and specialists within a given field, new platforms for crowdsourcing and citizen science are opening new opportunities for non-specialists as well. Professional scientists should embrace this trend by designing studies that collaborate with, rather than simply collect data from, the public. By encouraging open access not just of data, but where possible also the tools and expertise needed to analyze it, science will breed far more hypotheses and discoveries.