O'REILLY®

Data and Social Good

Using Data Science to Improve Lives, Fight Injustice, and Support Democracy

Mike Barlow

HOLE #8
PAR 3
CURRENT SCORE
JOE
MATT
5
WINNING SCORE
SAN JOSE

25 FINSBURY CIRCUS
London, UK EC2M 7EE
2012 VACANCY RATES
2012 RENTS
LONDON

floor trader 788
TRADING LICENSE #34096785
€ 9322.50
acct JN44K03C
NEW YORK

SINGAPORE

Data and Social Good

Using Data Science to Improve Lives, Fight Injustice, and Support Democracy

Mike Barlow

Beijing • Boston • Farnham • Sebastopol • Tokyo
O'REILLY®

Data and Social Good
by Mike Barlow

Printed in the United States of America.

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

Editor: Tim McGovern
Production Editor: Kristen Brown
Proofreader: Kristen Brown
Interior Designer: David Futato
Cover Designer: Karen Montgomery

August 2015: First Edition

Revision History for the First Edition
2015-08-12: First Release

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-92291-0

[LSI]

Table of Contents

Data and Social Good

Several years ago, large management consulting firms began describing data as the "new oil"—a magically renewable and seemingly inexhaustible source of fuel for spectacular economic growth. The business media rapidly picked up on the idea, and reported breathlessly about the potential for data to generate untold riches for those wise enough to harness its awesome power.

At the same time, another story was unfolding. That story wasn't about a few smart guys getting rich. It was about people using data to improve lives and make the world a better place.

For many of us, it's an alluring narrative, perhaps because it supports our hope that deep down, data scientists and statisticians are nice people who value social good over crass materialism.

Megan Price, for example, is director of research at the Human Rights Data Analysis Group. She designs strategies and methods for using data to support human rights projects in strife-torn countries like Guatemala, Colombia, and Syria. "I've always been interested in both statistics and social justice," Price says. In college, she started off as a math major, switched to statistics, and later studied public health in grad school. "I was surrounded by people who were all really invested in using their math and science skills for social justice. It was a great environment for bringing those interests together."

In Guatemala, Price serves as lead statistician on a project in which she analyzes documents from the National Police Archive. She helped her colleagues prepare evidence for high-profile court cases involving Guatemalan officials implicated in kidnappings. By rigorously analyzing data from government records, Price and her colleagues revealed clear links between the officials and the crimes. In Syria, she was lead statistician and author on three reports commissioned by the United Nations Office of the High Commissioner for Human Rights (OHCHR).

“I’d like to think that many statisticians and data scientists would do that kind of work if they had the chance,” she says. “But it can be difficult to find the right opportunities. Doing pro bono work is a lovely idea, but there are limits to what you can accomplish by volunteering a few hours on nights and weekends. Many projects require full-time commitment.”

Price hopes to see an increase in “formal opportunities” for data scientists to work on non-commercial, socially relevant projects. “Right now, there are very few organizations hiring full-time data scientists for social justice. I’m hoping that will change over the next 10 to 15 years.”

Structuring Opportunities for Philanthropy

In many ways, DataKind is a harbinger of the future that Price envisions. DataKind is nonprofit that connects socially minded data scientists with organizations working to address critical humanitarian issues. “We’re dedicated to tackling the world’s greatest problems with data science,” says Jake Porway, DataKind’s founder and executive director. “We connect people whose day jobs are on Wall Street or in Silicon Valley with mission-driven organizations that can use data to make a positive impact on the world.”

DataKind’s programs range from short-term engagements done over a weekend to long-term, multi-month projects. All programs bring together data scientists and social-change organizations to collaborate on meaningful projects that move the needle on humanitarian challenges.

For example, when data scientists at Teradata were looking for new and improved ways to apply their skills to philanthropy, they teamed up with DataKind. The two organizations co-hosted a weekend “DataDive” that provided an opportunity for data scientists from DataKind and Teradata to work collaboratively with nonprofits and humanitarian organizations such as iCouldBe, HURIDOCS, GlobalGiving, and the Cultural Data Project on a wide range of data challenges, from improving an online mentoring program for at risk youth to tracking human rights cases in Europe.

“One thing we found is there is no lack of demand for these services. We have over 200 organizations that have submitted applications to receive some sort of data science services,” Porway says. “On the other side, we should mention, we have more than 5,000 people who have signed up to volunteer. There is demand on both sides.”

In many instances, the challenge is combining or integrating data from disparate sources. In London, DataKind UK, one of the organization’s six chapters worldwide, helped St Mungo’s Broadway, a charity that helps people deal with issues leading to homelessness, link its data with data from Citizens Advice, a national charity providing free information on civil matters to the public. Linking the data yielded a trove of new insights that made it easier for St Mungo’s Broadway to predict which clients were more likely to benefit from its support.

In India, DataKind works with Simpa Networks, a venture-backed technology company in India that sells solar-as-a-service to energy-poor households and small businesses. Simpa’s mission is making sustainable energy “radically affordable” to the 1.6 billion people at the “base of the pyramid” who currently lack access to affordable electricity.

In a six-month project financially underwritten by MasterCard, a team of DataKind volunteers is using Simpa Networks’ historical data on customer payment behavior to predict which new applicants are likely to be a good fit for its model. That will enable Simpa Networks to best serve its customers and better assess new customers to offer the most appropriate services and support.

“Our goal is offering energy services to everyone, which includes customers who otherwise would be ‘unbankable’ according to mainstream financial institutions,” says Paul Needham, Simpa Networks’ chairman and CEO.

Data analytics plays a major role in supporting Simpa’s ambitious mission. “Customer usage and payment behaviors are constantly tracked, and the data is fed into our proprietary credit scoring model. That helps us get smarter about selecting customers and allows us to take risks on rural farmers that some banks would be uncomfortable financing,” Needham says.

The energy situation is especially dire in India, where 75 million families have no access to electricity and enormous sums are spent on unclean fuels such as kerosene for lanterns. “The good news is that effective decentralized energy solutions already exist. Solar photovoltaic solutions such as solar home systems can be sized appropriately to meet the energy needs of rural households and small businesses,” Needham says. With data analytics, Simpa can make the case for loaning money that can be applied to clean-energy systems.

“Having learned from our past impact evaluation results, we have sufficient evidence to support the fact that Simpa’s clean energy service will significantly reduce the time needed to conduct farming work, household chores, cooking and cleaning,” Needham says. “We anticipate that overall health standards will improve in these households due to the improved quality of light and will encourage the move away from kerosene and other hazardous forms of energy usage. In our midline impact evaluation study, we have seen that 80 percent of customers surveyed suffered eye irritation due to smoke; after Simpa’s intervention, this figure dropped to 28 percent. Similarly, 10 percent of customers surveyed experienced fire accidents; after Simpa’s intervention, this figure dropped to zero. We also believe that shop owners in these energy-poor areas will be able to stay open longer hours, which is likely to increase their sales and overall productivity.”

Telling the Story with Analytics

DataKind also has collaborated with Crisis Text Line (CTL), a free service providing emotional support and information for anyone in a crisis. The process for accessing help is simple and efficient: people in need of help send texts to CTL and trained specialists respond to the texts with support, counseling, information, and referrals.

CTL is staffed by volunteers, and like all volunteer organizations, its resources are constrained. CTL’s mission is providing potentially life-saving support services for people in need—but it’s also critical for the organization to avoid overwhelming its volunteers.

“Repeat callers have posed a challenge for crisis centers since the 1970s,” explains Bob Filbin, CTL’s chief data scientist. “When you read through the academic literature, you see that repeat callers are a big difficulty for crisis centers.”

It’s not that CTL’s counselors don’t want to help everyone who texts them—it’s just that some people who contact CTL need a rapid intervention to avert a tragedy. The hard part is figuring out which people are experiencing acute, short-term crises requiring immediate attention and which people are suffering from less acute problems that can be dealt with over a slightly longer timeframe.

After analyzing data from thousands of texts and examining patterns of usage from academic literature, Filbin and his colleagues were able to make suggestions for managing the problem of repeat texters. “We realized that our counselors were spending 34 percent of their time with 3 percent of our texters. By rolling out new policies and new technical products, we were able to reduce the portion of time our counselors spent with repeat texters from 34 percent to 8 percent. It was a huge win for us because it allowed more people to use the service.”

In addition to freeing up more time for volunteers to interact with people experiencing acute problems, CTL was able to improve service for the repeat texters by guiding them toward helpful longterm resources.

Using data analysis to boost CTL’s ability to deliver potentially lifesaving services to people in need is especially gratifying, Filbin says. “It’s very exciting when we can use data to overturn existing assumptions or drive meaningful change through an organization. Bringing data to bear on the problem, measuring our progress and evaluating the effectiveness of our policies and products—it all makes an enormous difference.”

From Filbin’s perspective, it all comes down to good storytelling. “Data is only valuable when people act on it. Framing the data in terms of saving time was an emotional trigger than helped people understand its value,” he says. “By reducing the conversation minutes with repeat texters from 34 percent down to 8 percent, we suddenly saved a quarter of our volunteers’ time. That’s a powerful story.”

The idea of using data as a tool for storytelling is a recurring theme among data scientists working in philanthropic organizations. Most of the data scientists interviewed for this report mentioned storytelling as an important output of their work. Essentially, a good story makes it easier for managers and executives to make decisions and to take action on the insights generated by the data science team.

Data as a Pillar of Modern Democracy

Emma Mulqueeny, who writes a popular blog on data science, sees a larger trend evolving. Mulqueeny is the founder of Rewired State and Young Rewired State, a commissioner for the Speaker's Commission on Digital Democracy in the UK, a Google Fellow and a digital tech entrepreneur. Earlier in her career, while working for the UK government on digital communication strategies, she noticed a sea change in the way people responded to statements made by government officials.

"There was a huge scandal over expenses," she recalls, "and suddenly it seemed as though everybody lost their trust in everything the government was saying. Suddenly, everybody wanted facts. They didn't want your interpretation of facts, they just wanted facts."

Government officials were aghast. But as a result of the scandal, efforts were made to increase transparency. Data that previously had been off limits or difficult to obtain was made available to the public. Data.Gov.UK and Data.Gov, both launched in 2009, are prime examples of the "open data" trend in democratic societies. It's almost as if governments are saying, "You want data? We got your data right here!"

Mulqueeny sees those kinds of efforts as steps in the right direction, but she's adamant about the need for doing more. "The way people are operating online, the way they're learning, sharing and influencing is very much dependent on what's pushed into their space," she says. "We're all familiar with Google's machine learning algorithms. You search for 'blue trousers' and suddenly everywhere you go after that, you're seeing little adverts for blue trousers and other items to buy. Marketers know how to mark up data so it can be used for marketing."

Democratically elected governments, on the other hand, are still struggling with data. "Let's say you feel passionate about chickens. If the information is properly marked up, you are more likely to see when the government is discussing matters related to chickens," Mulqueeny says. "Now let's say the government decides to outlaw chickens in London. If the information is marked up, you'll probably see it. But if it's not properly marked up, you won't. Which means that you won't find out the government is considering banning chickens until you read about it in a newspaper or some other media outlet."

From Mulqeeny's perspective, real democracy requires more than just sharing data—it requires making sure that data is properly tagged, annotated, and presented to people when they are online. In effect, she is raising the bar for governments and saying they need to be as good as—or better than—online marketers when it comes to serving up information.

"People have expectations that their interests will be served in the space in which they choose to be online and that they will find out what's happening when they are online," she says. "That's the heart of everything at the moment."

No Strings Attached, but Plenty of Data

For as long as most of us can remember, charities have worked like this: People or organizations make donations to charities, and charities distribute the donations to people or organizations that need support. Recently, and for a variety of legitimate reasons, the validity of that model has been called into question. As a result, new models for charitable giving have emerged.

GiveDirectly is an organization that channels donations directly to the extreme poor in Kenya and Uganda. The money is distributed via mobile phones, which makes it relatively easy to keep very precise digital records of who's getting what from whom. GiveDirectly's model was inspired by programs initiated by the Mexican government in the 1990s. Those programs showed that direct cash transfers to poor people were often more helpful than benefits that were distributed indirectly.

The "secret formula" behind GiveDirectly's success is scientific discipline. Two of the group's co-founders, Michael Faye and Paul Niehaus, describe the differences between GiveDirectly and traditional charities:

> From the very beginning, we took a principled stand and decided to run randomized trials, which are the gold standard for discovering whether something works or doesn't. Some people can always find excuses for not running randomized trials. They will say they're too expensive or they take too much time or they might jeopardize the business model.

> Our response to those excuses is to ask, 'Would you buy drugs from a pharmaceutical company that doesn't run randomized trials of its drugs?' Of course you wouldn't. So why would you donate money to a charity that doesn't test its programs?

Although GiveDirectly distributes donations with no strings attached, its approach is the antithesis of just throwing money at problems. True to their roots as trained economists, Faye and Niehaus have devised an excruciatingly detailed system for making sure donations are used properly. After choosing a village or area to receive donations, GiveDirectly sends a team to the location. The team goes from house to house, creating a highly detailed, data-rich map of the location. Then a second team is dispatched to register local inhabitants and verify the data assembled by the first team.

No money is actually distributed until a third team has verified the information provided by the first two teams, and even then, only token payments are made to make absolutely sure the money winds up in the right hands. When all the tests are complete, additional payments are authorized, flowing directly to the local residents via mobile banking or other forms of digital cash transfer.

It's a rigorous approach, but it's an approach that can be scaled and audited easily. Transparency, redundancy, and continual analysis are crucial to the success of the overall process. "We think it's the future of charity in the developing world. In fact, we don't see ourselves as a charity—we see ourselves as service providers," Faye says.

GiveDirectly draws a distinction between data and evidence. "We emphasize that understanding impact requires not just knowing what happened, but what knowing *would* have happened if we hadn't intervened," Faye says. "We do that with randomized controlled trials."

Faye and Niehaus urge donors to ask basic questions of all charitable organizations:

- Where exactly does a donated dollar go? Who are the beneficiaries and how much money ultimately winds up in their hands?
- Beyond data alone, do the organizations have evidence showing the impact of their interventions?
- Are the organizations doing more good per dollar than the poor could do by themselves?

Collaboration Is Fundamental

When the New York City Department of Health and Mental Hygiene (DOHMH) realized that restaurant reviews posted on Yelp could be a source of valuable information in the ongoing battle to prevent foodborne illnesses, the department reached out to Yelp and to data scientists at Columbia University for help.

Over a nine-month period, roughly 294,000 Yelp reviews were screened by software that had been "trained" to look for potential cases of foodborne disease. According to an article posted on the Centers for Disease Control and Prevention (CDC) web site, "the software flagged 893 reviews for evaluation by an epidemiologist, resulting in the identification of 468 reviews that were consistent with recent or potentially recent foodborne illness."

The article notes that only 3 percent of flagged reviews described events that had been reported to the health department. While the absolute numbers involved were relatively small, the project represents a major victory for data science.

Expending all of that effort to identify a handful of potentially dangerous restaurants in New York City might not seem like a big deal, but imagine scaling the process and offering it to every health department in the world.

"Data is everywhere now, more so than ever before in history," says Luis Gravano, a professor of computer science at Columbia University who worked with the health department on the Yelp project. "Regular people now are leaving a rich trail of incredibly valuable information, through the content they post online and via their mobile devices." Increasingly, data that people generate over the course of their daily lives is picked up by sensors. That kind of passively generated data is "less explicit, but also potentially quite valuable," Gravano says.

The data generated by "regular people" represents a unique opportunity for data scientists. "Collectively, the data is a great resource for all of us who analyze data," he says. "But the challenge is finding the gold nuggets of information in these mountains of data."

Dr. Sharon Balter, an epidemiologist at the health department, says data science was the key to finding the important pieces of information hidden in the reviews. "The team from Columbia helped us focus on the small number of restaurant reviews that might indicate real problems. The challenge is sifting through thousands of reviews. We don't have the resources to investigate every one of them," Balter says. "The algorithms developed by the Columbia team helped us determine which leads to investigate, and that was incredibly helpful."

Here's how the process worked, according to the CDC article:

> Beginning in April 2012, Yelp provided DOHMH with a private data feed of New York City restaurant reviews. The feed provided data publicly available on the website but in an XML format, and text classification programs were trained to automatically analyze reviews. For this pilot project, a narrow set of criteria were chosen to identify those reviews with a high likelihood of describing foodborne illness. Reviews were assessed retrospectively, using the following criteria: 1) presence of the keywords "sick," "vomit," "diarrhea," or "food poisoning" in contexts denoting foodborne illness; 2) two or more persons reported ill; and 3) an incubation period ≥10 hours.
>
> Ten hours was chosen because most foodborne illnesses are not caused by toxins but rather by organisms with an incubation period of ≥10 hours (*1*). Data mining software was used to train the text classification programs (*2*). A foodborne disease epidemiologist manually examined output results to determine whether reviews selected by text classification met the criteria for inclusion, and programs with the highest accuracy rate were incorporated into the final software used for the pilot project to analyze reviews prospectively.
>
> The software program downloaded weekly data and provided the date of the restaurant review, a link to the review, the full review text, establishment name, establishment address, and scores for each of three outbreak criteria (i.e., keywords, number of persons ill, and incubation period), plus an average of the three criteria. Scores for individual criteria ranged from 0 to 1, with a score closer to 1 indicating the review likely met the score criteria.
>
> Reviews submitted to Yelp during July 1, 2012–March 31, 2013 were analyzed. All reviews with an average review score of ≥0.5 were evaluated by a foodborne disease epidemiologist. Because the average review score was calculated by averaging the individual criteria scores, reviews could receive an average score of ≥0.5 without meeting all individual criteria.
>
> Reviews with an average review score of ≥0.5 were evaluated for the following three criteria: 1) consistent with foodborne illness occurring after a meal, rather than an alternative explanation for the ill-

> ness keyword; 2) meal date within 4 weeks of review (or no meal date provided); 3) two or more persons ill or a single person with symptoms of scombroid poisoning or severe neurologic illness. Reviews that met all three of these criteria were then investigated further by DOHMH. In addition, reviews were investigated further if manual checking identified multiple reviews within 1 week that described recent foodborne illness at the same restaurant.

Gravano and Balter agree that the availability of "non-traditional" data was critical to the success of their endeavor. "We're no longer relying solely on traditional sources of data to generate useful insights," Gravano says. As a result, groups of people that were previously "uncounted" can now benefit from the work of data scientists. "We're setting up an infrastructure that will make those kinds of projects more routine. Our hope, moving forward, is that our work will become a continuous process and that we will continually refine our algorithms and machine learning tools," he says.

Recently, another group of researchers at Columbia used machine learning tools to better understand and predict preterm births, a healthcare issue affecting 12–13 percent of infants born in the U.S. That study relied on clinical trial dataset collected by the National Institute of Child Health and Human Development (NICHD) and the Maternal-Fetal Medicine Units Network (MFMU).

Conclusion

Most of the sources interviewed for this report highlighted the multidisciplinary and inherently collaborative nature of data science, and several expressed the belief that at some level, most data scientists see their roles as beneficial to society. That said, there still appears to be a clear need for organizations that provide structures and processes for enabling the collaboration and teamwork necessary for effective pro bono data science projects. In other words, doing data science for the good of humankind requires more than good intentions—it requires practical frameworks, networks of qualified people, and sources of funding.

Applying data science principles to solve social problems and improve the lives of ordinary people seems like a logical idea, but it is by no means a given. Using data science to elevate the human condition won't happen by accident; groups of people will have to envision it, develop the routine processes and underlying infrastructures required to make it practical, and then commit the time and energy necessary to make it all work.

Columbia University has taken a step in the right direction by launching the Data Science Institute, an interdisciplinary learning and research facility with dedicated faculty and six specialized centers: Cybersecurity, Financial and Business Analytics, Foundations of Data Science, Health Analytics, New Media, and Smart Cities.

"Whatever good you want to do in the world, the data is there to make it possible," says Kathleen McKeown, director of the Data Science Institute. "Whether it's finding new and unexpected treatments for disease or techniques for predicting the impact of natural disasters, data science has tremendous potential to benefit society."

How to Help

Crisis Text Line is looking for volunteers. If you are interested in becoming a crisis counselor, please visit http://www.crisistextline.org/join-our-efforts/volunteer/ for more information.

DataKind is also seeking volunteers. If you're a data scientist looking to use your skills to give back, you can apply to volunteer with DataKind at http://www.datakind.org/getinvolved/ or learn more at an upcoming event in your area: http://www.datakind.org/howitworks/dataevents/.

About the Author

Mike Barlow is an award-winning journalist, author and communications strategy consultant. Since launching his own firm, Cumulus Partners, he has represented major organizations in numerous industries.

Mike is coauthor of *The Executive's Guide to Enterprise Social Media Strategy* (Wiley, 2011) and *Partnering with the CIO: The Future of IT Sales Seen Through the Eyes of Key Decision Makers* (Wiley, 2007).

He is also the writer of many articles, reports, and white papers on marketing strategy, marketing automation, customer intelligence, business performance management, collaborative social networking, cloud computing, and big data analytics.

Over the course of a long career, Mike was a reporter and editor at several respected suburban daily newspapers, including *The Journal News* and the *Stamford Advocate*. His feature stories and columns appeared regularly in *The Los Angeles Times*, *Chicago Tribune*, *Miami Herald*, *Newsday*, and other major US dailies.

Mike is a graduate of Hamilton College. He is a licensed private pilot, an avid reader, and an enthusiastic ice hockey fan. Mike lives in Fairfield, Connecticut, with his wife and two children.