



April 21, 2014

Mr. Donald S. Clark
Federal Trade Commission
Office of the Secretary
Room H-113 (Annex X)
600 Pennsylvania Avenue, NW
Washington, DC 20580

RE: Big Data: A Tool for Inclusion or Exclusion? Workshop, Project No. P145406

Dear Mr. Clark,

On behalf of the Center for Data Innovation (www.datainnovation.org), I am pleased to submit these comments in response to the Federal Trade Commission's (FTC) request for public comment on the recent FTC workshop "Big Data: A tool for inclusion or exclusion?"¹

The Center for Data Innovation is a non-profit, non-partisan, Washington-DC based think tank that formulates and promotes pragmatic public policies designed to enable data-driven innovation in the public and private sectors, create new economic opportunities, and improve quality of life. The Center is affiliated with the Information Technology and Innovation Foundation.

These comments supplement an earlier filing by the Center for Data Innovation dated August 15, 2014.² For your consideration, I am including a recently published report, as well as two articles, discussing some of the key issues raised in the workshop. Each of these is briefly summarized below.

¹ "Big Data: A tool for inclusion or exclusion," Federal Trade Commission, September 15, 2014, Washington, DC, <http://www.ftc.gov/news-events/events-calendar/2014/09/big-data-tool-inclusion-or-exclusion>.

² Daniel Castro, Comments to the FTC, Center for Data Innovation August 15, 2014, http://www.ftc.gov/system/files/documents/public_comments/2014/08/00026-92395.pdf.



Title: **The Rise of Data Poverty in America³**

Summary: Data-driven innovations offer enormous opportunities to advance important societal goals. However, to take advantage of these opportunities, individuals must have access to high-quality data about themselves and their communities. If certain groups routinely do not have data collected about them, their problems may be overlooked and their communities held back in spite of progress elsewhere. Given this risk, policymakers should begin a concerted effort to address the “data divide”—the social and economic inequalities that may result from a lack of collection or use of data about individuals or communities.

Title: **Big Data is a Powerful Weapon in the Fight for Equality⁴**

Summary: Many individuals have made the claim that big data might lead to more discrimination if algorithms unfairly disadvantage certain groups in their decision-making. While this type of discrimination is plausible, as discussed in this article, there are many compelling reasons why it may never come to pass. Moreover, the focus on how big data might be used to harm individuals has overshadowed the bigger opportunity to use data as a new tool in the fight for equality. Those working to fight discrimination should look to data as a way to further eliminate unjust biases in society and create a more fair and transparent society.

³ Daniel Castro, “The Rise of Data Poverty in America,” Center for Data Innovation, September 10, 2014, <http://www2.datainnovation.org/2014-data-poverty.pdf>

⁴ Daniel Castro, “Big data is a powerful weapon in the fight for equality,” *The Hill*, October 23, 2014, <http://thehill.com/blogs/pundits-blog/technology/221583-big-data-is-a-powerful-weapon-in-the-fight-for-equality>.



Title: **Wikipedia Edits Reveal America’s “Data Deserts”⁵**

Summary: User-generated data sets provide valuable information about many different communities. However, if some communities are not represented in the data, decisions may overlook members of these communities and their unique needs. These levels of inequality may even cluster geographically to give rise to “data deserts”—areas characterized by a lack of access to high-quality data that may be used to generate social and economic benefits. As an initial attempt to measure data deserts, the Center for Data Innovation analyzed which areas of the United States contributed the most to Wikipedia.

Thank you for considering the ideas outlined in these publications.

Sincerely,

Daniel Castro

Director, Center for Data Innovation
1101 K Street NW, Suite 610
Washington, DC 20005

dcastro@datainnovation.org

⁵ Travis Korte, “Wikipedia Edits Reveal America’s ‘Data Deserts’,” Center for Data Innovation, September 10, 2014, <http://www.datainnovation.org/2014/09/wikipedia-edits-reveal-americas-data-deserts/>.



The Rise of Data Poverty in America

By Daniel Castro | September 10, 2014

Data-driven innovations offer enormous opportunities to advance important societal goals. However, to take advantage of these opportunities, individuals must have access to high-quality data about themselves and their communities. If certain groups routinely do not have data collected about them, their problems may be overlooked and their communities held back in spite of progress elsewhere. Given this risk, policymakers should begin a concerted effort to address the “data divide”—the social and economic inequalities that may result from a lack of collection or use of data about individuals or communities.

Policymakers should begin a concerted effort to address the “data divide”—the social and economic inequalities that may result from a lack of collection or use of data about an individual or community.

INTRODUCTION

Every day, approximately 11,000 infants are born in the United States. Some of these babies are delivered in state-of-the-art hospitals where they not only receive excellent care, but their vital signs are also digitally monitored around the clock.¹ By the time they leave the hospital, some of them will have generated over 200 million data points. Others are born in far-less technically advanced settings and may leave the hospital with a keepsake footprint inked on a sheet of paper, but no digital footprint to speak of. As these children grow up, some will have access to the latest sensor-equipped mobile phones, wearable technology, intelligent vehicles, and other “smart” devices; some will attend schools that use cutting-edge data analytics to help them learn; and some will grow up in neighborhoods that routinely measure social, political, and environmental factors to foster more sustainable and resilient communities. But some will not.

For many years, policymakers have sought to address the “digital divide”—the social and economic disadvantages that may result from a lack of access to technology. Now policymakers should begin a concerted effort to address the “data divide”—the social and economic inequalities that may result from a lack of collection or use of data about an individual or community. Already gaps are appearing where certain groups of individuals do not have data collected about them or their communities because of where they live. If this trend towards a data divide continues we might even see the rise of “data deserts”—areas of the country characterized by a lack of access to high-quality data that may be used to generate social and economic benefits.

As the world races towards a more data-intensive future, individuals who come from data-rich environments may find that they have a comparative advantage over those who grow up in data poverty. These advantages may translate into better health care outcomes, increased access to financial services, enhanced educational opportunities, and even more civic participation. Conversely, if certain groups are routinely excluded from data sets, their problems may be overlooked and their communities held back in spite of progress elsewhere. To ensure that all individuals have access to the vast benefits offered by data-driven innovation and that no group is systemically disadvantaged, policymakers should pursue efforts to eradicate data poverty and close the data divide. Specifically, policymakers should do the following:

- Continue government data collection programs that focus on hard-to-reach populations
- Ensure that funding programs aimed at closing the digital divide consider the impact on data poverty
- Ensure that digital literacy programs help individuals understand data-producing technologies, such as social media and the Internet of Things
- Encourage civic leaders in low-income neighborhoods to understand the benefits of data and know how to integrate technology solutions into grant proposals

DATA POVERTY CAN AFFECT ECONOMIC AND SOCIAL WELFARE

Data has always had a tremendous impact on communities. The most visible example of this effect is the decennial census which is used to apportion congressional seats among the states. In addition, major government data collection initiatives such as the American Community Survey are used to determine how to distribute more than \$400 billion in federal funding annually, with a substantial portion of these funds going to

low-income households.² These surveys can have a negative effect on a community when they do not accurately count certain groups of individuals, such as homeless individuals, renters, or non-English speakers.

The impact that data has on individuals and communities will only grow, particularly with the emergence of the Internet of Things. The Internet of Things refers to the wide array of smart devices that collect data about the world around them and communicate with people and other computers over the Internet. Many of these devices, and the data generated by them, will be used to address important real-world problems, such as managing health care, protecting the environment, and improving public services.³ For example, sensor-equipped devices, like smart meters that track the real-time energy consumption of homes and smart thermostats that adjust heating and cooling in response to whether individuals are home, can help improve energy efficiency. The City of Chicago has even announced a project to mount sensors on lamp posts throughout the city to collect environmental data such as temperature, humidity, and air quality, among other information.⁴

Communities that are poor in data, as well as the individuals living in those communities, may fail to thrive. Rather than being the new oil, data may be the new oxygen.

Data is increasingly seen as a valuable resource in its own right, capable of creating enormous wealth and powering the economy, with commentators saying “data is the new oil.”⁵ However, the corollary to these dictums is that communities that are poor in data, as well as the individuals living in those communities, may fail to thrive. Rather than being the new oil, data may be the new oxygen. Indeed, the lack of access to certain key resources can significantly impact an individual’s quality of life. For example, individuals without official identification can face significant challenges in their daily lives, such as finding jobs, accessing financial services, and traveling on planes.⁶ When these access barriers disproportionately affect certain groups, it can leave them marginalized by society.

If disparities in data production and collection continue, they may lead not only to problematic policy decisions by the government, but also less-than-optimal decisions by individuals as well as by businesses. For example, commuters without access to real-time transit data might spend more time waiting for buses and trains than their peers with better access to that data, or for-profit transit alternatives might not have the data they need to invest in private transportation services. Sometimes this data may even mean the difference between life and death. Some carmakers are beginning to integrate collision notification features that automatically send data about the location and potential severity of a collision to emergency responders in the event of an accident, thereby allowing them to respond quicker, especially if a victim is unable to call for help.⁷

As shown in table 1, individuals produce a vast amount of data from many different sources from genetic data to electronic payments. As new types of data are generated about individuals, the disparity in data collected about different communities may be exacerbated.

Type of Data	Example
SENSORS	
Wearable tech	Nike FuelBand
Smart devices	Nest thermostat
GPS and accelerometers	Street Bump app
Smart infrastructure	Smart electricity meters
Genetic and genomic tests	23andMe genetic tests
Medical tests	Theranos blood tests
Satellites	Satellite images
COMPUTER	
Social media	Facebook posts
Transactional	Electronic payments
Web	Search patterns
Mobile	Bluetooth signals
Hardware	Network traffic logs
TRADITIONAL	
Surveys	American Community Survey
Records	Patient medical records
Tests	SAT scores

Table 1: Examples of different types of data

As described below, data is growing in importance to a wide variety of sectors, including education, health care, and financial services, and data poverty may have a serious impact on individuals obtaining many of the emerging benefits of using data in these sectors.

EDUCATION

Data is poised to have a disruptive impact on education. Data can help government leaders create more effective education policy, schools operate more efficiently, families find the best schools, teachers discover the most effective lessons, and students learn better.

There are many areas where data is being used to improve K-12 education. Some of these efforts to use data focus directly on helping students learn better or improve their behavior. For example, adaptive learning software can personalize the presentation of concepts based on each student's individual learning style and comprehension level, as well as provide students immediate feedback on their progress.⁸ The nonprofit testing firm

Educational Testing Services collects data from over 180 countries to analyze responses from test questions to determine if students have mastered the material, so that educators can customize assignments based on each student's progress.⁹ The Beaverton, Oregon School District uses student data about in-school and out-of-school suspensions, unexcused and excused absences, as well as demographic information to personalize disciplinary approaches to students to better help them succeed in the classroom.¹⁰ Eventually, schools hope to use social network analysis to predict and intervene in undesirable behavior like cheating or tardiness.¹¹

Other improvements come from using data to make better decisions about how to run a school. Some schools are implementing performance-based measures for their teachers, as well as using predictive analytics to screen potential new hires.¹² In Dallas, Texas, after the public school system began collecting better operational data, it realized it was spending more collecting payments from students for breakfasts and lunches than it cost to provide the meals. As a result, the Dallas Independent School District decided to offer all students free meals and eliminated its costly payment collection system, helping both its own financial standing and its students in the process.¹³ Some school districts make their operational data publicly available. The School District of Philadelphia, for example, has published a wide array of data sets about its public schools, including teacher pay, demographics, expenditures, and student performance.¹⁴ These open data efforts (i.e. efforts to make data freely available without restrictions) help ensure transparency and accountability in the educational system, allow parents to make informed decisions about which schools to send their children to, and promote equal access to a quality education.

Many universities are also beginning to use data to improve their students' experiences. For example, Arizona State University partnered with the learning analytics company Knewton to teach students remedial math skills so that they could succeed in college. Using a data-driven approach to personalized learning, the university was able to decrease the withdrawal rates by 56 percent and increase pass rates by 10 percent in the first year of the program.¹⁵ Similarly, after Georgia State University introduced adaptive learning software and a predictive analytics program to identify risky student behavior and setup appropriate interventions, it was able to reduce the rate of students failing or withdrawing from its introductory math classes from 40 percent to 20 percent.¹⁶ In addition, some schools are using predictive analytics to recommend courses for students based on their interests, learning habits, and past performance, much like Netflix recommends movies.¹⁷ And the University of Texas at Austin enters information about every incoming student into a database,

and compares it to past student performance, to identify which freshmen need extra support.¹⁸ Data is even being used to help match students to schools. One data analytics company is using its algorithms to help identify low-income students who are applying to colleges beneath their potential and suggesting better schools where they might want to apply.¹⁹ President Obama has even proposed linking student aid to schools' performance, so that students attending schools that have high graduation rates and produce graduates who get high-paying jobs, would be eligible for greater financial assistance.²⁰

While there are many opportunities to use data to improve education, students who come from data poor environments might lack access to many of these benefits, such as more efficient schools, personalized learning, and better guidance in making decisions about post-secondary education.

The coming years will likely see a rise in patient-generated health data, or information that is recorded by patients themselves or collected by a device outside of a clinical setting.

HEALTHCARE

Data is critical to the health and well-being of individuals, and it is being used to improve virtually every aspect of healthcare from developing new drugs to delivering care to patients. Increased use of data in health care offers a broad range of benefits, including more personalized and coordinated care, better quality, faster treatment development, and lower costs.²¹

One way that data is used to improve health care is to assess the efficacy of treatments. Regulators like the Food and Drug Administration (FDA) use data from clinical trials to decide whether drugs, vaccines, or other medical interventions are safe and effective. Historically racial and ethnic minorities, as well as women, have been underrepresented in clinical trials. For example, Hispanics represent approximately 16 percent of the U.S. population but only 1 percent of clinical trial participants.²² When certain groups are underrepresented in the data, the decisions made about the safety and efficacy of treatments for patients may be biased. For example, women are more likely than men to have an adverse reaction to a drug and may respond differently to medical devices. Some drugs have even been taken off the market because of effects found in women that were missed in clinical trials. Similarly, studies have found that various racial and ethnic groups respond differently to certain medications.²³

The coming years will likely see a rise in patient-generated health data, or information that is recorded by patients themselves or collected by a device outside of a clinical setting. The volume of patient-generated health data has increased rapidly in recent years due in part to the proliferation of wearable technology ("wearable tech") and other smart devices.²⁴ There are already a number of fitness and health devices, such as Nike Fuelband

and Fitbit, which track an individual's activity levels and sleep patterns, and the Withings scale which helps individuals track their weight, resting heart rate, and even the indoor air quality. The recently-introduced Apple Watch, a device with geo-tracking and health-tracking capabilities, will likely bring wearable tech even further into the mainstream. Data from these devices can help motivate individuals to lead healthy lifestyles or receive incentives, such as discounts or rebates on health-related goods and services. Individuals without access to this type of data about themselves may not benefit from incentives designed to promote healthy lifestyles.

Patient-generated data has considerable value, as it is typically much more granular than laboratory data, allows for real-time monitoring, and can capture small fluctuations that lab tests may miss. Health care providers may use the data to remotely monitor and manage their patients' care and intervene quickly if the need arises. For example, the FDA has recently approved mobile health startup AliveCor's algorithm for detecting whether an individual is having a heart attack by monitoring real-time electrocardiogram (ECG) data from a heart-rate monitor attached to a mobile device.²⁵ Or to take another example, Propeller Health has made a GPS-enabled device that tracks data about the usage of inhalers by asthma patients. Its system then integrates public information from the Centers for Disease Control and Prevention (CDC) about environmental asthma triggers so that health care providers can create personalized treatment plans.²⁶ Even simple interventions, such as smart pill bottles that record data about when patients take their medicine, can help ensure better health care outcomes by helping patients adhere to treatment plans.²⁷ Once again, individuals with access to data-driven health care technologies will likely have better outcomes than those without it.

A variety of health care initiatives have been launched that rely on public data sets. For example, the Baltimore-based startup Symcat uses machine learning algorithms to analyze medical records from the CDC to help patients identify potential diagnoses based on their symptoms.²⁸ Another company, mHealthCoach, has created an app that uses data from the Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project and warnings from the FDA's ClinicalTrials.gov website to send high-risk patients personalized messages and reminders about their medications.²⁹ Researchers are also experimenting with alternative sources of data for health care purposes. For example, local health officials have experimented with using online surveys of Facebook users to measure vaccine practices.³⁰ In another example, researchers evaluated personal ads on Craigslist to find the incidence of men who have sex with

men but do not openly identify as gay. The research suggested that public health initiatives to combat HIV/AIDS which rely on the size of the openly gay population may overlook certain communities where interventions may be necessary.³¹ These applications show the potential of using big data analytics in health care, but also show the limitations if certain groups are excluded from the data sets.

The importance of data in health care will only continue to grow as providers seek to use data to personalize prevention, diagnosis, and treatment options to patients' unique biological, environmental, and social conditions.

FINANCIAL SERVICES

Better data allows financial service providers to reduce costs and deliver better quality services to consumers. Many financial institutions are using big data to identify potential customers, help them resolve problems they might encounter, and ensure they are offering the financial services consumers need.³² Major banks, such as Bank of America and Wells Fargo, actively monitor social media to identify and resolve customer complaints.³³ Others use a variety of data, including customer profiles and transactions, to identify high risk transactions and cut down on fraud.³⁴ For example, credit card companies, like Visa and MasterCard, as well as payment processors like PayPal, use advanced algorithms and powerful computers to analyze hundreds of aspects of each transaction they process to reduce fraud and keep costs low for merchants and consumers.³⁵

Better data also helps creditors understand the risks associated with a particular activity and price their services appropriately. Deeper understanding can unlock new opportunities, such as loans and insurance that would otherwise be unavailable to underserved populations. For example, banks in Africa generally have not offered many commercial loans for agricultural purposes in the past because they do not have access to good data to build risk models. However, startups, such as Nairobi-based Gro Ventures, are collecting regular updates from small farmers throughout the region about crop yields and commodity prices, aggregating the data, and then using it to build risk models. Banks can then use these risk models to make loans to farmers or farm cooperatives.³⁶ In the United States, insurance companies like State Farm and Progressive have used data collection devices that plug into a vehicle's diagnostic port to capture information about individuals' driving behaviors and offer safe drivers lower rates.³⁷ The software-company Agero has even produced an app for a mobile phone that will collect data such as driving behavior, distance traveled, and hours driven during peak traffic times

which will allow consumers to purchase insurance based on how they drive rather than their demographic.³⁸

Data is making it possible for individuals and businesses to get access to credit who otherwise would be denied. For example, car dealers have begun to use Internet-connected devices to lower the risk of giving a car loan to consumers who lack credit or collateral. In the past, dealers would worry that customers might stop paying their loans and then abscond with the vehicle. However, by using a GPS-enabled mobile device to track the vehicle, the dealer can easily locate the car or even remotely disable it in the event of non-payment. By providing this data, an individual who might otherwise not get a car loan, such as someone who has been struggling to make ends meet but needs a car to get to a new job, is able to do so.³⁹ Similarly, startups like LendUp offer loans to individuals who might otherwise be denied credit or charged higher rates using traditional lending practices. By reviewing both traditional data from credit bureaus and social media data that captures previously hard-to-quantify aspects about individuals, such as the strength of their ties to the community, these startups are able to extend credit to more individuals.⁴⁰ Finally, many small businesses in the United States have had a difficult time obtaining credit from banks after the Great Recession. However, these small businesses, including many women-owned and minority-owned businesses, have been able to get credit from alternative lenders that look beyond collateral and credit scores, and instead use verified data about online sales, bank transactions, and online reviews to lend to companies with a positive cash flow.⁴¹

Data is shaping the future of many financial services and unlocking new opportunities for individuals and businesses. However, individuals and communities without rich data profiles may find that their needs are unmet and that they are unable to take advantage of new services.

RECOMMENDATIONS

Addressing the risk of data inequality should be a high priority for policymakers so that the benefits of data-driven innovation can be shared by all communities. Specifically, we recommend the following:

1. Continue government data collection programs that focus on hard-to-reach and underrepresented communities

Even traditional methods of data collection can underrepresent certain groups. With regards to traditional survey methods, there are many well-known causes of sampling biases. For example, with regards to telephone surveys, since pollsters are not permitted to use automated dialing technology to call cell phones, they often conduct polls only using

landlines. As a result, landline-only polls tend to undercount younger and non-White Americans who tend to live in cell phone-only households.⁴² And even attempts to count an entire population often miss certain groups. As noted earlier, the Census Bureau has had long-standing challenges collecting data about certain populations, as factors such as access to a telephone, poverty rates, high school graduation rates, language skills, mobility, and employment levels all impact how difficult or easy it is get a high response rate from a given neighborhood.

To ensure that all communities are well-represented in data sets, government-led data collection efforts should strive to be as inclusive as possible.

To ensure that all communities are well-represented in data sets, government-led data collection efforts should strive to be as inclusive as possible. For example, Sen. Brian Schatz (D-HI) recently introduced S. 989 the “Strengthening Health Disparities Data Collection Act” which would require the Department of Health and Human Services to collect data about the sexual orientation and gender identity in all of its federally-funded health-related programs.⁴³ As government agencies undertake new data collection tasks, part of their review process should consider whether their efforts are likely to underrepresent certain groups and disclose any known shortcomings when they publish their data sets.

2. Ensure that funding programs aimed at closing the digital divide consider the impact on data poverty

New methods of data collection may also result in biased results. For example, individuals without access to certain types of technology, such as smart phones and wearable tech, will not be able to contribute to aggregated data sets that are later used for research. There are a number of government programs in place to help close the digital divide, such as the Lifeline program to offer discounted phone services to low-income Americans.⁴⁴ Similarly, the Obama Administration has publicly backed the “Connect to Compete” program to provide computers and broadband Internet access to families that qualify for the National School Lunch Program.⁴⁵ Future funding programs aimed at closing the digital divide should also consider the impact that technology access has on data for certain communities. In particular, as the number of smart devices steadily grows, policymakers should ensure that low-income communities have affordable access to the Internet of Things. For example, public utility commissions should track smart meter deployment to ensure these devices are made available on an equal basis. In addition, policymakers should ensure that any federal or state programs aimed at spurring smart city development require citywide deployment, including in low-income neighborhoods.

3. Ensure that digital literacy programs help individuals to understand data-producing technologies, such as social media and the Internet of Things

Certain groups of individuals may choose not to use specific technologies that produce important data sets simply because they do not understand how to use the technology. For example, varying levels of digital literacy is one reason why some groups of individuals are underrepresented on social networks, use certain types of technologies at lower rates than others, or do not contribute user-generated content on platforms like Wikipedia.⁴⁶ Just as policymakers should update policies designed for the digital divide to address the data divide, policymakers should update policies designed to reduce digital literacy to help individuals understand data-producing technologies.

4. Encourage civic leaders in low-income neighborhoods understand the benefits of data and know how to integrate technology solutions into grant proposals

The Internet of Things offers many opportunities to use sensors to embed intelligence into everyday items and eventually create smart homes, smart cars, smart infrastructure, and smart cities. These devices will generate a substantial amount of data that will help communities learn how to be more productive, sustainable, and resilient. It is important that the Internet of Things does not only become part of certain types of cities or neighborhoods, but also is integrated into diverse communities. One way to ensure this happens is for federal grant-making agencies to work with civic leaders to understand how technology can be applied to various projects. For example, the U.S. Department of Transportation can provide training to ensure that both high-income and low-income communities have the capability to apply for funding to implement intelligent transportation systems. Other initiatives, including the U.S. Department of Energy's grants for energy efficiency and the U.S. Department of Education's technology grants, should take similar actions.

CONCLUSION

Some people might suggest that inequalities in data collection mean that data-driven solutions should not be used. This would be a mistake. Just as the solution to the digital divide was not to go limit the use of computers, the solution to the data divide is not to take steps back from using data. As described above, data-driven innovations have enormous opportunities to advance important societal goals such as improving the cost and quality of health care and education, as well as advancing access to financial services. The goal of policymakers should be to ensure that no groups are

systematically excluded from data collection activities so that all individuals have the opportunity to obtain the social and economic benefits of data.

REFERENCES

1. "IBM InfoSphere, Big Data Help Toronto Hospital Monitor Premature Infants," eWeek, September 26, 2013, <http://www.eweek.com/enterprise-apps/ibm-infosphere-big-data-help-toronto-hospital-monitor-premature-infants.html>
2. Lisa Blumberman and Philip Vidal, "Uses of Population and Income Statistics in Federal Funds Distribution – With a Focus on Census Bureau Data," 2009, Governments Division Report Series, Research Report #2009-1, <http://www.census.gov/prod/2009pubs/govsrr2009-1.pdf> and Rachel Blanchard Carpenter and Andrew Reamer, "Surveying for Dollars: The Role of the American Community Survey in the Geographic Distribution of Federal Funds," Brookings, July 26, 2010, <http://www.brookings.edu/research/reports/2010/07/26-acs-reamer>.
3. Daniel Castro and Jordan Misra, "The Internet of Things," November 2013, <http://www2.datainnovation.org/2013-internet-of-things.pdf>.
4. Whet Moser, "What Chicago's 'Array of Things' Will Actually Do," Chicago Magazine, June 27, 2014, <http://www.chicagomag.com/city-life/June-2014/What-Chicagos-Array-of-Things-Will-Actually-Do/>
5. "Venture Investing & Hiring in Silicon Valley," CNBC, February 22, 2012, <http://video.cnbcm.com/gallery/?video=3000074076>.
6. Peter Swire and Cassandra Butts, "The ID Divide," Center for American Progress, June 2, 2008, <http://www.americanprogress.org/issues/civil-liberties/report/2008/06/02/4520/the-id-divide/>.
7. Connie Pignataro, "Automatic crash notification: a promising resource for fire EMS," Fire Engineering, September 20, 2013, <http://www.fireengineering.com/articles/print/volume-166/issue-9/departments/fireems/automatic-crash-notification-a-promising-resource-for-fire-ems.html>.
8. Darrell West, "Big Data for Education: Data Mining, Data Analytics, and Web Dashboards," Brookings, September 2012, <http://www.brookings.edu/~media/research/files/papers/2012/9/04%20education%20technology%20west/04%20education%20technology%20west.pdf>.
9. Doug Guthrie, "The Coming Big Data Education Revolution," U.S. News & World Report, August 15, 2013, <http://www.usnews.com/opinion/articles/2013/08/15/why-big-data-not-moocs-will-revolutionize-education>, and "Who We Are," ETS, accessed September 3, 2014, <http://www.usnews.com/opinion/articles/2013/08/15/why-big-data-not-moocs-will-revolutionize-education>.

-
10. Darrell West, "Big Data for Education: Data Mining, Data Analytics, and Web Dashboards," Brookings, September 2012, <http://www.insidepolitics.org/brookingsreports/education%20big%20data.pdf>.
 11. Travis Korte, "Using Data for K-12 Education," Center for Data Innovation, July 28, 2013, <http://www.datainnovation.org/2013/07/using-data-for-k-12-education/>.
 12. "'Predictive' Tech Tools Aim to Help Districts Hire Better Teachers," Education Week, December 19, 2013, http://blogs.edweek.org/edweek/marketplacek12/2013/12/predictive_tech_tools_aim_to_hire_better_teachers.html.
 13. "All Dallas ISD students will not get free breakfast and lunch," The Dallas Morning News, October 1, 2013, <http://educationblog.dallasnews.com/2013/10/all-dallas-isd-students-will-now-get-free-breakfast-and-lunch.html/>.
 14. See for example, "Open Data Initiative," The School District of Philadelphia, n.d. <http://webgui.phila.k12.pa.us/offices/o/open-data-initiative> (accessed September 5, 2014).
 15. "Knewton Technology Helped More Arizona State University Students Succeed," Knewton, accessed September 3, 2014, <http://www.knewton.com/assets-v2/downloads/asu-case-study.pdf>.
 16. "Using Predictive Analytics, Adaptive Learning to Transform Higher Education," Government Technology, July 29, 2014, <http://www.govtech.com/education/Using-Predictive-Analytics-Adaptive-Learning-to-Transform-Higher-Education.html>.
 17. "The Netflix Effect: When Software Suggests Students' Courses," The Chronicle of Higher Education, April 10, 2011, <http://www.innovationfiles.org/cloud-computing-policy-challenges-for-a-globally-integrated/>.
 18. Dale Basye, "Big Data: What Can Be Learned from How We Learn?" Clarity Innovations, July 19, 2014, <https://www.clarity-innovations.com/blog/dbasye/big-data-what-can-be-learned-how-we-learn>, and Libby Nelson, "Big Data 101: colleges are hoping predictive analytics can fix their dismal graduation rates," Vox, July 14, 2014, <http://www.vox.com/2014/7/14/5890403/colleges-are-hoping-predictive-analytics-can-fix-their-graduation-rates>.
 19. "Civis Analytics's Dan Wagner on Data Solutions to Social Problems," BloombergBusinessweek, March 6, 2014, <http://www.businessweek.com/articles/2014-03-06/civis-analytics-dan-wagner-on-data-solutions-to-social-problems>.

-
20. "Obama Wants College Aid Tied to Rating System," Wall Street Journal, August 22, 2013, <http://online.wsj.com/news/articles/SB10001424127887323665504579028911262193186>.
 21. Daniel Castro and Travis Korte, "Comments to the Senate Finance Committee on Health Data Transparency," August 1, 2014, <http://www2.datainnovation.org/2014-senate-finance-health-data.pdf>.
 22. "Clinical Trials Shed Light on Minority Health," Food and Drug Administration, <http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm349063.htm>.
 23. "Successful Strategies for Engaging Women and Minorities in Clinical Trials," Society for Women's Health Research and U.S. Food and Drug Administration Office of Women's Health, September 2011, <http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/WomensHealthResearch/UCM334959.pdf>
 24. Michael Shapiro, et al., "Patient-Generated Health Data," RTI International, April 2012, http://www.healthit.gov/sites/default/files/rti_pghd_whitepaper_april_2012.pdf and "Patient-Generated Health Data," Health IT.gov, <http://www.healthit.gov/policy-researchers-implementers/patient-generated-health-data>.
 25. "Can AliveCor's Heart Monitor Predict Your Stroke Before it Strikes," Co.Labs, September 5, 2014, <http://www.fastcolabs.com/3035224/healthware/alivecors-predictive-heart-monitor-could-save-you-from-a-stroke>.
 26. Groves et al., "The 'big data' revolution in health care."
 27. Daniel Castro and Jordan Misra, "The Internet of Things," November 2013, <http://www2.datainnovation.org/2013-internet-of-things.pdf>.
 28. "A Q&A with Symcat," CNBC, n.d., <http://www.cnbc.com/id/100453499> and "About SymCAT," Symcat.com, n.d., <http://www.symcat.com/faq>.
 29. Peter Groves et al., "The 'big data' revolution in health care," Center for U.S. Health System Reform Business Technology Office, January 2013.
 30. "Facebook, Online Surveys Track HPV Vaccinations at Local Level," Health Data Management, September 2, 2014, <http://www.healthdatamanagement.com/news/Facebook-Online-Surveys-Track-HPV-Vaccinations-at-Local-Level-48728-1.html>.

-
31. Varoon Bashyakarla et al., “Harnessing Craigslist Personal Ads to Inform Federal HIV Prevention Funding,” Workshop on Data Science for Social Good, KDD 2014, August 224, 2014, <http://dssg.uchicago.edu/kddworkshop/>.
 32. Thomas Davenport and Jill Dyché, “Big Data in Big Companies,” International Institute for Analytics, May 2013, <http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf>.
 33. Amy Fontinelle, “Why Banks Are Scrambling To Hear Your Complaints,” Forbes, October 25, 2013, <http://www.forbes.com/sites/investopedia/2013/10/25/why-banks-are-scrambling-to-hear-your-complaints/> and Adam O’Daniel, “Can Bank of America, Wells Fargo learn to like social media?” Charlotte Business Journal, October 18, 2013, <http://www.bizjournals.com/charlotte/print-edition/2013/10/18/can-big-banks-learn-to-like-social.html>.
 34. Jason Bloomberg, “Three-Way Big Data Banking Battle Brewing,” Forbes, July 29, 2014, <http://www.forbes.com/sites/jasonbloomberg/2014/07/29/three-way-big-data-banking-battle-brewing/>.
 35. “Crunching the numbers,” The Economist, May 19, 2012, <http://www.economist.com/node/21554743> and “Visa Says Big Data Identifies Billions of Dollars in Fraud,” CIO Journal, March 11, 2013, <http://blogs.wsj.com/cio/2013/03/11/visa-says-big-data-identifies-billions-of-dollars-in-fraud/>.
 36. Ciara Byrne, “Data-Driven Lending Could Help African Farmers Feed The World,” Co.Labs, October 14, 2013, <http://www.fastcolabs.com/3019953/data-driven-lending-could-help-african-farmers-feed-the-world>.
 37. “Insurers will now be able to track driver behavior via smartphones,” ComputerWorld, September 3, 2014, <http://www.computerworld.com/article/2600344/telematics-insurers-will-now-be-able-to-track-driver-behavior-via-smartphones.html>.
 38. Ibid.
 39. “GPS devices help dealerships keep an eye on vehicles,” NBC News 10, February 4, 2013, <http://www.news10.com/story/20962823/gps-devices-help-dealerships-keep-an-eye-on-cars>.
 40. Evgeny Morozov, “Your Social Networking Credit Score,” Slate, January 30, 2013, http://www.slate.com/articles/technology/future_tense/2013/01/wonga_lenddo_lendup_big_data_and_social_networking_banking.html.

-
41. Paul Davidson, "Small businesses turn to alternative lenders," USA Today, November 13, 2012, <http://www.usatoday.com/story/money/business/2012/11/13/unconventional-business-loans/1650637/> and "Big Data provides Big Advantages in Small Business Lending," Business Insider, August 9, 2012, <http://www.businessinsider.com/big-data-provides-big-advantages-in-small-business-lending-2012-8>.
 42. "Researchers Warn of 'Bias' in Landline-Only Phone Polls," National Journal, June 18, 2013, <http://www.nationaljournal.com/blogs/hotlineoncall/2013/06/researchers-warn-of-bias-in-landline-only-phone-polls-18>.
 43. "S.989 - Strengthening Health Disparities Data Collection Act, 113th Congress," Congress.gov, n.d., <https://beta.congress.gov/bill/113th-congress/senate-bill/989>.
 44. "Lifeline Program for Low-Income Consumers," Federal Communications Commission, August 26, 2014, <http://www.fcc.gov/lifeline>.
 45. Note that the program is now called "EveryoneOn." Josh Gottheimer and Jordan Usdan, "Low-Cost Broadband and Computers for Students and Families," Federal Communications Commission, November 10, 2011, <http://www.fcc.gov/blog/low-cost-broadband-and-computers-students-and-families>.
 46. Travis Korte, "Wikipedia Edits Reveal America's 'Data Deserts'," Center for Data Innovation, September 9, 2014. "Lifeline Program for Low-Income Consumers," Federal Communications Commission, August 26, 2014, <http://www.fcc.gov/lifeline>.

ABOUT THE AUTHOR

Daniel Castro is the director of the Center for Data Innovation where he leads the Center's research efforts. Mr. Castro is also a senior analyst at the Information Technology and Innovation Foundation. Previously he worked as an IT analyst at the Government Accountability Office. He has a B.S. in Foreign Service from Georgetown University and an M.S. in Information Security Technology and Management from Carnegie Mellon University.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation at the Information Technology and Innovation Foundation conducts high-quality, independent research and educational activities on the impact of the increased use of data on the economy and society. In addition, the Center for Data Innovation formulates and promotes pragmatic public policies designed to enable data-driven innovation in the public and private sectors, create new economic opportunities/and improve quality of life. The Center for Data Innovation also sponsors the annual Data Innovation Day.

contact: info@datainnovation.org

datainnovation.org



October 23, 2014, 06:00 am

Big data is a powerful weapon in the fight for equality

By Daniel Castro, contributor

Many individuals have made the claim that big data might lead to more discrimination if algorithms unfairly disadvantage certain groups in their decision-making. For example, an online retailer might offer one price to Asian customers and another price to Latino customers. Most notably, this idea appeared in the White House big data **review** led by John Podesta, where the final report stated that "An important conclusion of this study is that big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups." This idea was also the subject of a recent Federal Trade Commission (FTC) **workshop** exploring what FTC Chairwoman Edith Ramirez **termed** "discrimination by algorithm." While this type of discrimination is plausible, there are many compelling reasons why it may never come to pass. Moreover, the focus on how big data might be used to harm individuals has overshadowed the bigger opportunity to use data as a new tool in the fight for equality.

One reason concerns about discrimination are likely overblown is because many laws, including the Americans with Disabilities Act (ADA), the Genetic Information Nondiscrimination Act (GINA), the Fair Credit Reporting Act (FCRA) and the Employee Retirement Income Security Act (ERISA), protect consumers from employers, creditors, landlords and others who may take adverse actions against them based on protected classes of information. Big data does not exempt businesses from following these laws. Earlier this year, for example, the FTC brought charges and entered into a **settlement** with Instant Checkmate for violating the FCRA.

Of course, just because a business might be able to create a racist or sexist algorithm, does not mean that it will do so. After all, most businesses are not actively seeking to discriminate against minorities, and in fact, **many companies** are actively championing a more inclusive worldview. Moreover, even where there are "bad apples," companies face significant market pressure to not engage in such behavior, a lesson that most executives have probably learned following the swift departure of sponsors after the disclosure of former Los Angeles Clippers owner Donald Sterling's racist remarks. Big data has not changed these factors.

But the bigger point is that the focus on preventing discrimination has diverted attention away from the bigger opportunity to use big data to create a more inclusive society. There are at least three ways this can happen. First, automated processes can remove human biases from decision-making. For example, while loan officers or apartment managers may discriminate, perhaps even unintentionally, on the basis of age or race, computers can be programmed to ignore these variables. Second, data creates feedback loops that encourage people to treat others as individuals. While **some taxi drivers** have notoriously refused to pick up passengers because of the color of their skin, apps like Uber allow drivers to decide whether to give someone a ride based on the passengers' ratings, which are mostly tied to whether the riders are **punctual and tidy**. Third, data is a useful way to identify latent racism, such as discriminatory hiring practices or racial profiling by police. For example, data collected about the disparate impact of New York City Police Department's controversial **stop-and-frisk policy** has helped change opinions on the approach.

In short, not only are fears that big data will lead to discrimination in the future likely overblown, but they have clouded the debate. Those working to fight discrimination should look to data as a way to further eliminate unjust biases in society and create a more fair and transparent society.

*Castro is director of the **Center for Data Innovation**.*

TAGS: Federal Trade Commission, FTC, Edith Ramirez, Big data

The Hill 1625 K Street, NW Suite 900 Washington DC 20006 | 202-628-8500 tel | 202-628-8503 fax
The contents of this site are ©2014 Capitol Hill Publishing Corp., a subsidiary of News Communications, Inc.



Search...

Go →

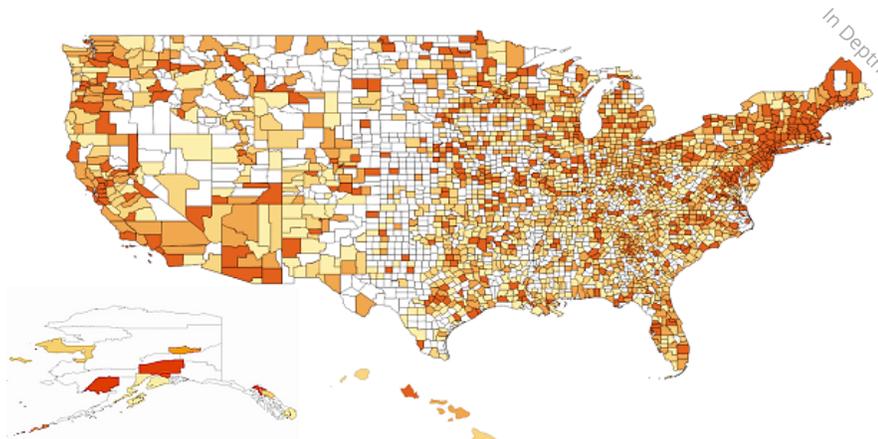
About Us

Publications

Events

Legislation Tracking

Resources



Published on September 10th, 2014 | by Travis Korte

0

Wikipedia Edits Reveal America's "Data Deserts"

For many, the concept of the "data revolution" conjures up images of online transactions, genomic medical data, and ubiquitous sensor networks. But user-generated data sets, such as social media updates, Flickr photos, and blog posts, are a major part of the data revolution as well. Internet users post **100 hours of video to YouTube every minute**, and tweet around 600 million times a day. These data sets reflect the lived experiences of millions of individuals, and collectively provide valuable information about many different communities.

User-generated data sets have been used for many purposes, from **using Twitter data to detect earthquakes** to **using Flickr and YouTube data to forecast political and economic attitudes**. For example, a number of recent initiatives have used user-generated data for public health purposes. The United Nations' Global Pulse initiative has **mined web search data** to detect non-communicable diseases such as cancer and diabetes, health officials in New York City have used Yelp reviews to **track and respond to public health outbreaks**, and Penn State University researchers have used Wikipedia search and click data to **predict outbreaks of illnesses**. The uses of user-generated data will only continue to grow in the future as they offer an enormous supply of real-time data.

But not all communities are equally represented in these data sets. Unequal access to broadband service, variations in access to technology, disparities in the level of digital literacy, and a host of other factors can influence who is included in the data and who is not. When communities are not represented (or underrepresented) in the data, decisions made after analyzing this data may overlook members of these communities and their unique needs.

As a result, the amount of user-generated information produced in a particular area of the country can serve as a bellwether for how much that community is able to realize the benefits of the data revolution. As my colleague Daniel Castro describes in a **recent report**, these levels of inequality may even cluster geographically to give rise to "data deserts"—areas characterized by a lack of access to high-quality data that may be used to generate social and economic benefits.

As an initial attempt to measure data deserts, the Center for Data Innovation analyzed which areas of the United States contributed the most to Wikipedia. Since Wikipedia is an entirely crowd-sourced project, the scope, depth, and accuracy of its articles depends on the engagement and interests of its users. Uneven contributions have had bizarre results: For example, **the Wikipedia article "List of**

[Back to Top ↑](#)
[Sign up for our weekly newsletter](#)

2,424
Followers53
Subscribers575
Fans

Latest

Popular Posts



Data-Driven Medicine in the Age of Genomics
December 11th | by Daniel Castro



Mapping the Internet's Sleep Cycle
October 22nd | by Travis Korte



Proposed EU Data Protection Regulations Could Impede Medical Research
October 21st | by Travis Korte



5 Q's for Education Data Expert Paige Kowalski
October 20th | by Travis Korte

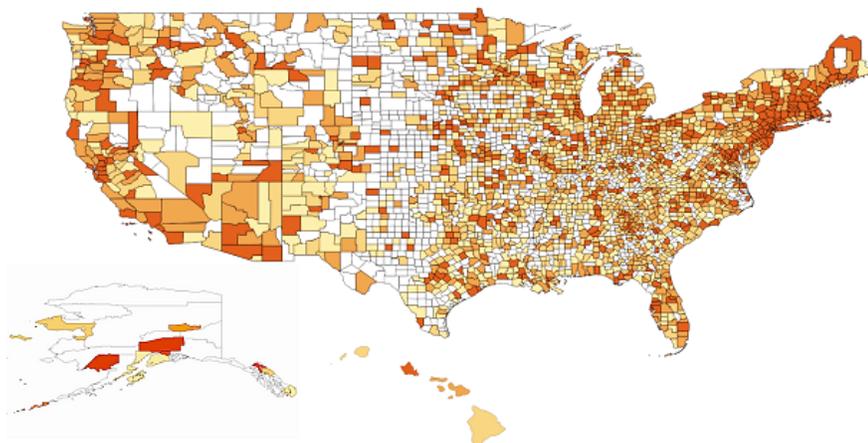


All Businesses Stand to Benefit from Big Data
October 17th | by Joshua New

Advanced Dungeons & Dragons 2nd edition monsters is longer than the article on **British literature**, and **the article on professional wrestler "The Undertaker"** has been revised **more times than has the article on global warming**. Similarly, a lack of data from particular communities might overlook important local insights and information known to members of that community.

To find Wikipedia's data deserts, we collected 134,958 edits to the English-language edition of the online encyclopedia Wikipedia made from July 19-23 and August 8-26 2014, and mapped them at the county level, adjusted for population. These edits represent public contributions from unregistered users in the United States. Since anyone can edit Wikipedia, gaps in this data set indicate areas where people have chosen not to participate in collaboratively editing the online encyclopedia.

Findings



White indicates that no edits were recorded for that county, and darker color indicates higher edit rates.

Although in some places our findings track with population and population density estimates, despite adjusting for population, other areas diverge from those metrics. For example, the non-coastal western states exhibit an edit rate that far exceeds what would be expected based on population metrics alone, as do northern counties in New York, Vermont, New Hampshire, and Maine. Some well-documented technology hubs, like the San Francisco Bay Area, the Pacific Northwest, and the Washington, D.C.-to-Boston corridor, share a high edit rate with regions with less high-tech reputations, such as Florida and southern Arizona. Particularly prolific individual editors can come from anywhere, which introduces some noise into areas with generally few edits, but the barren vertical strip extending from west Texas up through western Oklahoma, Kansas, Nebraska, and the Dakotas still appears stark. These areas have a relatively **low population density** and **high median age**, which may contribute to the low per capita edit rates.

Methodology

We created the data set of population-adjusted anonymous edits by county by downloading metadata for anonymous edits, geolocating the Internet protocol (IP) address associated with each edit, and identifying a county for each geolocation. We used the "recent changes" module of the **application programming interface** (API) for MediaWiki—the open source wiki software platform that contains the English Wikipedia as a subset—to download approximately 400,000 total edits. The API provides a range of information, including date, page title, and user data, for each edit. Edits submitted by anonymous users, i.e., those users who have not registered Wikipedia accounts, are also linked to an IP address. The IP address associated with each edit can be converted to a geolocation, i.e., latitude and longitude coordinates, using a range of online services. A very small fraction of the addresses we collected were formatted in IPv6, the newest version of the Internet protocol that **carries around four percent of Internet traffic** as of September, 2014. Available geolocation tools were not compatible with IPv6 addresses, so we filtered them out. We used freegeoip.net, a free API that takes IP addresses as input and outputs country, region, city, latitude and longitude, and other information. We filtered out IP addresses located outside the United States, as well as those addresses the geolocation tool was unable to resolve at the county level. IP geolocation is not exact, since some IP addresses reflect the locations of Internet service providers rather than users themselves. Still, geolocation services can typically place an IP address within a few miles of its origin, which is generally sufficient for

determining what county the address comes from.

After filtering out the unusable data, we then used the Federal Communications Commission's (FCC) **Census Block Conversions API** to match each latitude-longitude pair with a county, enabling mapping. We conducted all data processing in **R** and all mapping in **QGIS**.

Tags: [api](#), [counties](#), [data deserts](#), [wikipedia](#)

About the Author



Travis Korte is a research analyst at the Center for Data Innovation specializing in data science applications and open data. He has a background in journalism, computer science and statistics. Prior to joining the Center for Data Innovation, he launched the Science vertical of The Huffington Post and served as its Associate Editor, covering a wide range of science and technology topics. He has worked on data science projects with HuffPost and other organizations. Before this, he graduated with highest honors from the University of California, Berkeley, having studied critical theory and completed coursework in computer science and economics. His research interests are in computational social science and using data to engage with complex social systems. You can follow him on Twitter @traviskorte.

Related Posts



Create Maps of Wikipedia's Geotagged Articles →



White House Petitions API →



10 Bits: the Data News Hot List →



An API for Mobile Phone Accessibility →

0 Comments

Data Innovation Day

Login ▾

Sort by Best ▾

Share ↗ Favorite ★



Start the discussion...

Be the first to comment.

ALSO ON DATA INNOVATION DAY

WHAT'S THIS?

5 Q's for Applied Sports Scientist Gary McCoy

2 comments • 5 months ago



Travis Korte — It's a good question. A lot of pro teams in various sports have already started doing athlete ...

15 Women in Data to Follow on Twitter

1 comment • 8 months ago



Joe McCarthy — As someone who is interested in data science, and always happy to see women accorded more ...

100 Million Flickr Images for Download

2 comments • 4 months ago



preeti yadav — pls mail me the link on ypreeti4@gmail.com thnks

Will Open Government Be Accessible for People with Disabilities?

1 comment • 9 months ago



J. Albert Bowden II — implementing web standards is an afterthought for 90% of developers sadly. perhaps a ...

Subscribe

Add Disqus to your site

Privacy

© 2014

[About the Center for Data Innovation](#) | [Resources](#) | [Public Policy Issues](#) | [Events](#)

Back to Top ↑