The Landscape of Big Data and Gender

A Data2X Update

February 2021

data2x

The past, present, and future of big gender data

Data is everywhere. The shapes and textures of the things around you, the thoughts in your mind, these words you're reading: all are forms of data. We capture a small portion of this flood of data with our senses and our technologies; we analyze a subset of what is captured; and we use a fraction of that subset to guide our actions as individuals and societies.

The motivating belief of Data2X is that greater capture, analysis, and use of data catalyzes gender equality. Major gaps in knowledge persist about all aspects of women's and girls' lives: economic opportunity, education, environment, health, human security, public participation.[1] Filling these gaps makes inequality and discrimination visible, enabling public agencies, businesses, and civil society organizations to enact reforms that move societies closer to the ideals of equality and justice.

Over the last several years, Data2X has investigated the role of "big data" in filling gendered knowledge gaps. While there is no consensus definition of big data—the data science community continues to debate the topic[a]—at Data2X we think of big data simply in terms of its unique value: as information finely disaggregated over space and across time. When used in concert with traditional data, spatially and temporally rich information allows nuanced, effective response by public and private actors. Such capacity is of especially critical importance during fast-moving crises like pandemics, economic shocks, and natural disasters.

Data2X's big data work began in 2014 with *The Landscape of Big Data in Development*,[5] a report on how digital data—especially mobile phone records, Twitter posts, and satellite data—informs us about social and economic development in low-income countries. Data2X then funded a small set of proof-of-concept studies in 2016–2017 that used big data to explore the gendered aspects of nutrition, mental health, expenditures, and other themes, as summarized in our report *Big Data and the Well-being of Women and Girls: Applications on the Social Scientific Frontier*.[6] The success of this work led us, in partnership with other foundations, to launch a "Big Data and Gender Challenge" in 2018–2019. The ten winning research teams, selected from a pool of over one hundred applicants, used a gender lens to analyze mobility, financial services, education, physical security, the gig economy, and many other topics. In November 2019, we held a convening for the grantees to present their work; the proceedings are summarized in the Data2X report *Big Data, Big Impact? Towards Gender-Sensitive Data Systems*.[7]

a. See, for example, Laney's original "3 Vs" (volume, velocity, and variety) definition[2]; Letouzé's "3 Cs" (crumbs, capacities, and community) rejoinder[3]; and UN Women's categories of human-sourced, process-mediated, media-sourced, crowdsourced, and machine-generated data,[4] a classification adapted from Letouzé's work.

Just over a year after the convening, this brief reports on the ongoing work of five Data2X grantee partners: Girija Borker, formerly of Brown University and now at the World Bank; Jihad Zahir at Cadi Ayyad University in Marrakesh, Morocco; Stefaan Verhulst and colleagues at The Governance Lab (The GovLab) at New York University; the Flowminder Foundation and its partner organization WorldPop; and the research team of Ridhi Kashyap at Oxford University and Ingmar Weber at the Qatar Computing Research Institute (QCRI). We also review other innovative studies at the intersection of big data and gender carried out in the last several years, highlighting the trends most relevant to vulnerable women and girls in the world.[b]

Overall, this report draws out six observations about trends in big data and gender:

The current environment

1. COVID-19 and the global economic recession is stimulating groundbreaking gender research.

Where we're progressing, where we're lagging

2. Some gendered topics—especially mobility, health, and social norms—are increasingly well-studied through the combination of big data and traditional data.
3. However, worrying gaps remain, especially around the subjects of economic opportunity, human security, and public participation.
4. Capturing gender-representative samples using big data continues to be a challenge, but progress is being made.

Emerging opportunities

5. Large technology firms generate an immense volume of gender data critical for policymaking, and researchers are finding ways to reuse this data safely.
6. Data collaboratives that bring private sector data-holders, researchers, and public policymakers together in a formal, enduring relationship can help big data make a practical difference in the lives of women and girls.

b. The literature review was based on a Scopus database search on the titles of articles published between 2016 and 2020. We utilized two categories of search terms, one pertaining to gender and the second pertaining to big data. The former required one of the following search terms: gender, women, woman, girl, girls. The second utilized a broader set of terms associated with big data analysis (e.g., big data, machine learning). The search yielded several thousand results, of which we retained a subset for full-text review. We highlight a few notable articles among this subset in the pages that follow.

The Current Environment

COVID-19 and the global economic recession is stimulating groundbreaking gender research.

The COVID-19 pandemic and the consequent economic recession is illustrating the critical role of data in real-time policymaking—and also highlighting how little we know about the effects of the global crisis on gender inequality. The Sex, Gender and COVID-19 project, led by Global Health 50/50 at University College London, is tracking sex-disaggregated indicators directly relevant to COVID-19, including tests, cases, hospitalizations, and deaths.[9] As of November 2019, 32% of global COVID-19 cases and 25% of deaths were of unknown sex. In India and several other countries, the COVID-19 case fatality rate is for unknown reasons higher among women than men.[10]

The pandemic is affecting aspects of women's and girls' lives beyond health. Data2X, in partnership with Open Data Watch, recently published the brief *Tracking the Gender Impact of COVID-19: An Indicator Framework*, highlighting key data gaps relating to the consequences of COVID-19,[11] on economic well-being, access to safety nets, education, and other dimensions of human development. A few studies already suggest serious gender disparities emerging from the pandemic with respect to employment, childcare, and psychological distress.[12–15] Even the launch of effective vaccines brings uncertainties: clinical data on the relative efficacy of vaccines for men and women is scarce.[16]

The pandemic has increased the willingness of companies to share gender data; our grantee partner, The GovLab at New York University, has compiled a large repository of COVID-19-related datasets.[17] Telecommunications companies in particular are doing more in-house gender analysis, making anonymized call detail records available, and partnering with civil society on new research initiatives.[18] UN Women, for example, is working with mobile network providers to launch SMS-based rapid assessment surveys on the gendered impacts of the pandemic.[19] Facebook, with the World Bank and other partners, is leveraging its broad user base to conduct surveys on household gender dynamics in countries around the globe.[20] United Nations Global Pulse (UNGP) labs around the world are implementing innovative COVID-19 projects using digital data, including work on estimating transmission risk in Jakarta, Indonesia.[21] The pandemic is also stimulating the development of creative analytical methods that could be extended to gender questions. One study uses 164 million Google Street View images to identify predictors of COVID-19 in the built environment.[22] A similar approach could be used to examine the environmental predictors of gender inequality, for instance how urban planning decisions around lighting, transit, and zoning can influence women's physical security and access to markets, schools, and social services.

In the sections that follow, we discuss in more detail the ongoing pandemic-related projects of Data2X's grantee partners, especially the work of Flowminder to improve vaccination coverage in the Democratic Republic of the Congo using cell phone data—an initiative made possible by the urgency of COVID-19—and the work of Dr. Jihad Zahir and Dr. Girija Borker in analyzing the troubling global trends in domestic violence linked to COVID-19.[23] Dr. Zahir is also gathering online information on how the pandemic is posing economic and mental health difficulties for students in Morocco.

In addition, the technical approaches being developed for COVID-19-related studies are readily applicable to a wide range of gendered topics. Tracking the pandemic is fundamentally a question of improving the resolution of human well-being data in both time and space—exactly the type of challenge for which big datasets are useful.

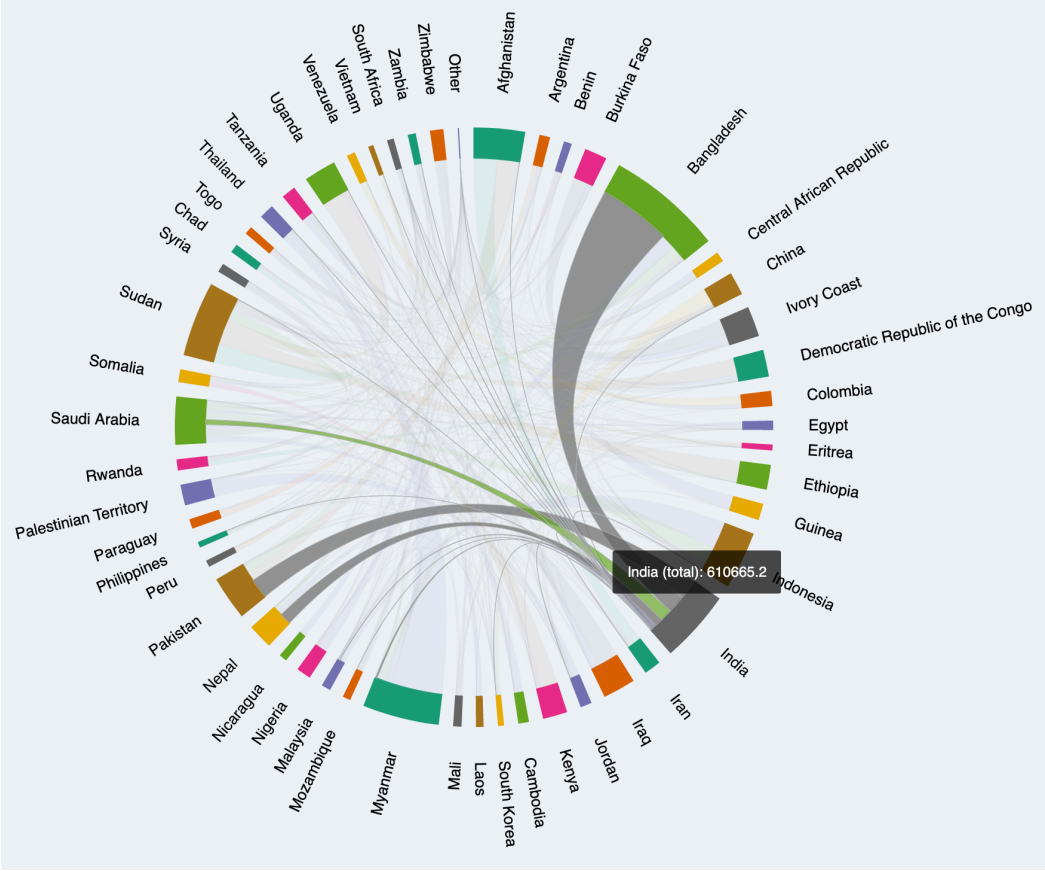
Where We're Progressing, Where We're Lagging

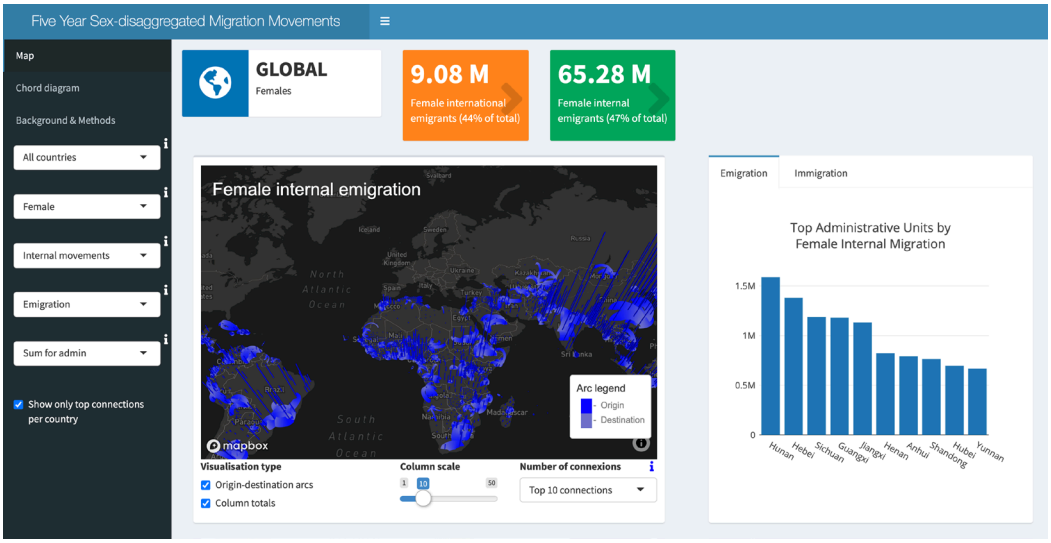
Some gender topics—especially mobility, health, and social norms—are increasingly well-studied through the combination of big and traditional data.

Data2X's grantee partners have in the past illustrated how call detail records[16,24] and credit card datasets[25] can be used to map mobility patterns. New work by WorldPop and the Flowminder Foundation compiles a variety of datasets, both big and traditional, to estimate internal and transnational migration patterns by sex (Figure 1).[26] Other researchers used cell phone data to look at population movements disaggregated by gender following natural disasters[27] as well as everyday patterns of mobility.[28] In addition, cell phone data can paint a portrait of how vulnerable groups like refugees are integrated with (or segregated from) the broader population, as well as internal and external communication links of displaced and migrant groups.[29] One recent study used Facebook data to show that women were disproportionately displaced in the U.S. states of Louisiana and Texas after Hurricane Laura in 2020.[8]

Big data health research is also expanding rapidly. Using data from social media platforms,[30] search engines,[31,32] online question and answer services,[33] and the Internet of Things (IoT),[34] advances in textual sentiment analysis have facilitated study of the gendered relationship between online expression and disorders like depression and anxiety.[35,36] Textual analysis is especially useful for mental health research, but social media datasets are are also being utilized for health studies more broadly: a recent review identified 105 social media health studies conducted between 2011 and 2017, with a nearly sixfold increase in the annual number of studies from the beginning to the end of that period.[37] Large-scale studies of electronic health records—digital versions of patient paper charts containing medical history, diagnoses, medications, and other information—also hold great potential for understanding gendered health trends. Recent studies using health records span a variety of topics, including prevention of mother-to-child transmission of HIV,[38] gendered patterns of ophthalmological disorders,[39] adherence to HIV treatment,[40,41] and evaluations of digital health programs themselves.[41] Other researchers are using unconventional forms of big data to study gendered topics. One study used a combination of georeferenced climate and crop calendar data, along with geotagged Demographic and Health Surveys data, to analyze how droughts affect the life courses of young women.[42]

Figure 1. Visualizations of (A) transnational migration, with a focus on out-migration from India, and (B) internal migration by sex. Source: WorldPop and Flowminder Foundation (https://gravier.shinyapps.io/fdfa_01_dev_v4/).





A great deal of work is also focused on changes in gendered social norms. The patterns of linguistic expression on Twitter effectively serve as a monitoring system for society-wide gender biases,[43] in addition to providing specific insights on issues like gendered religiosity.[44] A recent study found that web browsing behaviors and smartphone usage also accurately reflect the moral values of users.[45] Even interactions between Wikipedia editors offer a look at gendered norms: one study spanning fifteen years, nearly a million comments, and over one hundred thousand editors illustrated the gendered nature of communication styles on Wikipedia and its implications for the exercise of power.[46] Public comments on news articles also reveal how gender differences relate to engagement across various topics.[47] The UNGP labs and the Africa's Voices Project (AVP) are employing automated speech recognition techniques to analyze radio conversations on gender-based violence and other topics, allowing tracking of how social attitudes change over time.[30,48–52]

In all these cases, big data depends on the presence of complementary traditional datasets. Mobile operator data can illuminate mobility, but intra-household surveys are necessary to understand how the cultural and economic context affects phone and SIM card sharing between genders.[53] Inferring physical and mental health status from social media expression builds on background research about how language is shaped by gender, class, culture, and other factors. More generally, validation surveys reveal biases inside big datasets—for example, whether data is representative of lower socioeconomic classes, rural residents, or other groups. Big data is unique in many respects, especially in its temporal and spatial resolution, but it exists within a broader scientific environment, and its interpretability relies on the insights generated by traditional methods. We also note that the distinction between big and traditional datasets is heuristic, not objectively real—a useful tool to identify the unique potentials and risks of new data types. As methods of data integration evolve, however, the distinction between big and traditional datasets will grow increasingly blurry.

Worrying gender data gaps remain, especially around economic opportunity, human security, and public participation.

Despite the deluge of big data, many gendered topics remain relatively unexplored. Prominent among these are economic opportunity, human security, and public participation. With respect to economic opportunity, important knowledge gaps persist around childcare access, entrepreneurship, migrant worker conditions, pay inequity, and aspirations. Key human security-related gaps include information on conflict and gender-based violence, war-related mortality and illness, human trafficking, safety in public spaces, and violence against children. Information on local representation, professional leadership, voter registration and turnout, and violence against women in politics are key public participation-related gaps.

Progress is being made in some of these areas. With respect to economic opportunity, Data2X's grantee partner Dr. Girija Borker is using large-scale data scraping of business websites in India to compare how female entrepreneurs perform in comparison to male-run businesses.[54] Her study is analyzing performance over the life cycle of start-ups, constraints to success, and policy avenues to alleviate these constraints. Our grantee partners Ridhi Kashyap of Oxford and Ingmar Weber at QCRI are using LinkedIn data from the platform's marketing application programming interface (API) to assess gendered patterns of online professional networking across business sectors.[55]

Few other big data studies of business activity and gender exist, but one notable exception is a recent audit of the entire .uk domain—encompassing over 200 million web pages, 157 thousand organizations, and 2.3 million people—to investigate gender representation in the United Kingdom's economy.[56] The researchers find that men are overrepresented in roles, power, status, and titles. More broadly, the study illustrates the capacity of Internet data to map the landscape of gender bias in an economy as a whole. Another LinkedIn study using millions of user profiles found that women are underrepresented in leadership positions across a range of industries, especially health care, retail, and financial services.[57] A smaller sentiment analysis study of email communications finds gender differences in business advisory services in Sweden.[58] Our allies at UNGP have also done groundbreaking work on gendered bottlenecks to financial services in Cambodia,[59] the use of mobile money in Uganda,[60] and the financial impacts of natural disaster in the Mexican state of Baja California Sur.[61] Another UNGP study maps financial service access points across Indonesia; this analysis could be overlaid with mobility datasets to provide a sense of gendered access.[61,62]

Several of our Data2X partners are also working on research projects related to human security. Dr. Borker is continuing her previous work on women's mobility and safety by collecting app-based data on gender-based violence in Dar es Salaam, Tanzania, as well as analyzing the combination of cell phone call detail records and emergency helpline data to look at patterns of domestic violence in India.[54] Dr. Jihad Zahir, with the support of the United Nations Population Fund (UNFPA) and the National Union of Moroccan Women, recently launched a chatbot ("Najatbot") that links women who are victims of domestic violence to legal and psychological services (Figure 2).[63] This project, motivated by the increase in domestic violence in Morocco during the COVID-19 pandemic, provides information in the national Arabic

Figure 2. NajatBot, a Facebook community service that connects women who are victims of gender-based violence to legal and psychological services. Source: https://www.facebook.com/najatchatbot/

dialect, and is integrated into Facebook's Messenger app. The chatbot interacts with users through an easy-to-use menu and buttons interface, and will soon be updated with natural language processing (NLP) tools to allow a more flexible user experience. If Najatbot goes to sufficient scale, user interactions could be used to estimate the population-level prevalence of domestic violence. Other big data studies on gender-based violence also exist. For example, a partnership between the Data-Pop Alliance and GIZ Data Lab is developing a risk prediction model for domestic violence in Mexico City,[64] and an exploration of electronic health records from tens of millions of patients in the United States sheds light on the consequences of intimate partner violence.[65]

Some recent studies show the potential of big data to improve understanding of women's public participation. Several innovative analyses of hashtag activism on social media give insights on the underexplored question of how gendered political ideas spread across networks, how social media network structure reflects real-world structures of collective action, how these networks influence policy outcomes, and women's representation in social movements generally.[66–68] Another recent study on the donations of men and women to political candidates in the United States examines a rarely discussed aspect of gendered political voice.[69]

Despite these studies, large knowledge gaps on all of the above topics exist. For example, we know little about the prevalence and incidence of gender-based violence in hard-to-monitor (e.g., rural) populations. Mapping geographical patterns of female entrepreneurship faces a similar challenge, particularly with respect to women farmers and rural businesses. Very little information exists on the gendered aspects of civic engagement outside of elections.

Constructing gender-representative samples using big data is difficult, but progress is being made.

Three broad problems of representation complicate the use of big data to fill gender data gaps: 1) sex-disaggregated population estimates at high spatial and temporal resolution are rarely available; 2) gender disparities in access to digital technologies exist, leading to potential sampling bias when using big datasets; and 3) many big datasets are not gender-tagged.

Answering most gender questions requires accurate underlying population data. Unfortunately, censuses are too expensive to be done at the necessary frequency (or, in some countries, at all). To deal with this issue, our Data2X grantee partners WorldPop and Flowminder are aggregating a wide range of data sources, including censuses, vital registration datasets, health surveys, and satellite imagery, to construct reliable sex-disaggregated population estimates at very local levels (Figure 3).[70] These estimates then serve as the foundation for exploring other gendered topics at very high spatial resolution—for example, access to maternal health services[71] (Figure 4) and sex-disaggregated internal and transnational migration flows,[26] as noted earlier. Overall, the linkage of big and traditional datasets allows representative samples to be built, creating a foundation for novel and powerful gender-focused research designs.

Quantifying gendered access to technology is another important bottleneck; if women do not have equal access to technology, they will not be represented accurately in big data streams.[72] This gender invisibility can also interact with socio-economic class invisibility in many big data types (Table 1). Big data analysis may thus

Figure 3. Percentage of females aged 10-14 years old in 2020, subnational estimates. Source: WorldPop (https://www.portal.worldpop.org/demographics/)

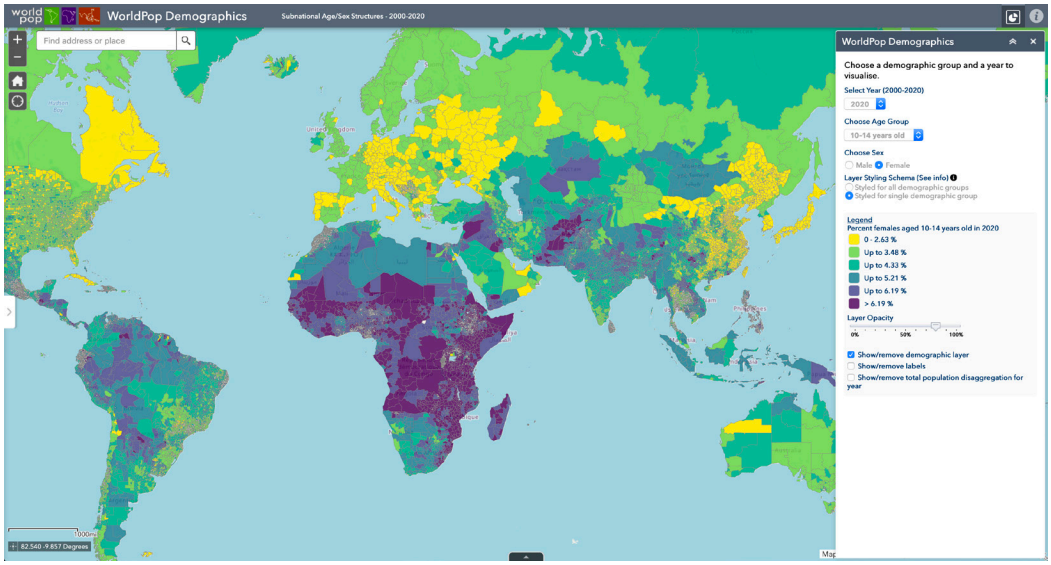


Figure 4. Percentage of women of child-bearing age within two hours of a health facility, subnational estimates from the Democratic Republic of the Congo and surrounding countries. Source: WorldPop (https://experience.arcgis.com/experience/8946bbc4090749c2aa1b6c1c80999bc6/page/page_14/?views=view_48)

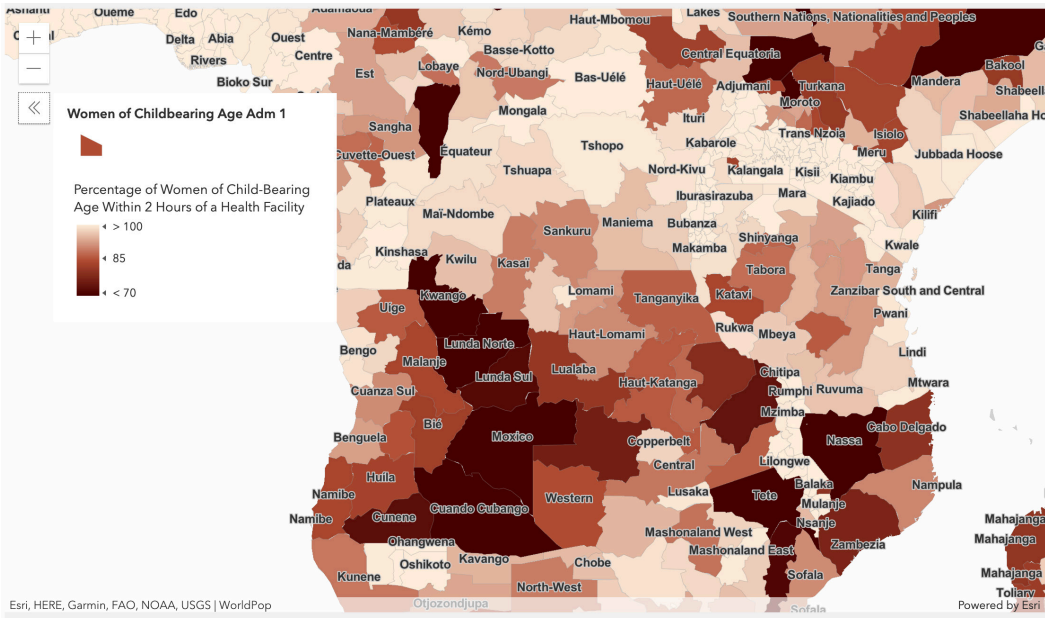
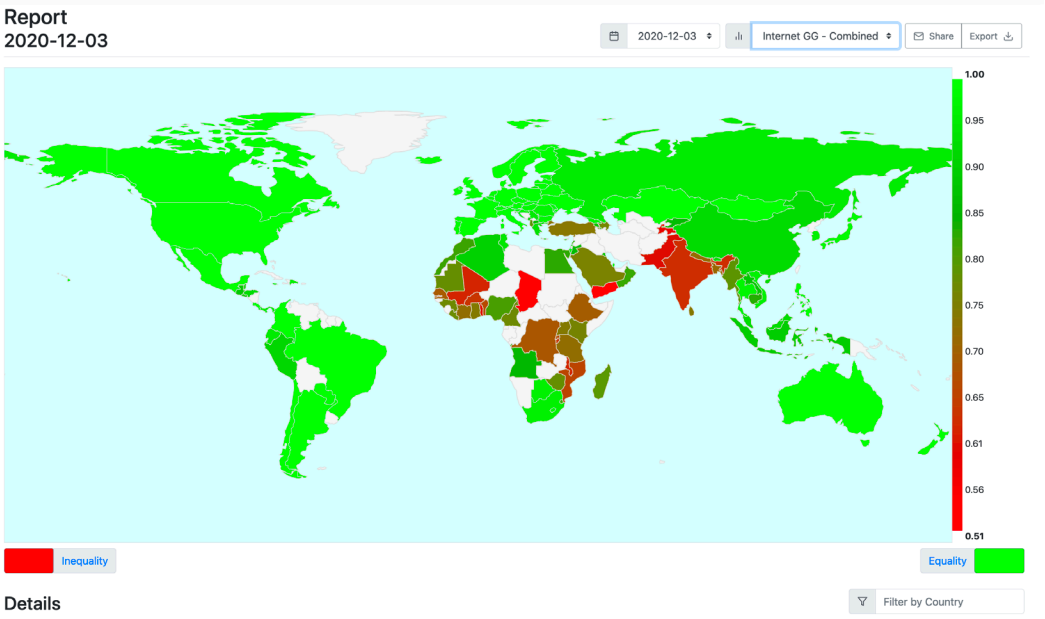


Table 1. Gender and socio-economic class label limitations in frequently used big data types.

		Gender tags generally present or easily inferred from content	
		Yes	**No**
Socio-economic class tags generally present or easily inferred from content	**Yes**	Call detail records from postpaid users (though gender accuracy is variable), electronic health records, email content, credit card data, financial transactions; some forms of aggregated social media data (e.g., advertising APIs)	Call detail records from prepaid users, road sensors, satellite imagery, electricity, public transit usage
	No	Individual-level Twitter, Weibo, Facebook, other social media user data (class can be inferred in some cases)	Search engines, web activity, audio recordings, crowdsourced app data (gender and class can be inferred in some cases for all of these types)

Figure 5. The ratio of female-to-male internet use estimated using the Facebook Gender Gap Index by country combined with other offline development indicators (e.g., the Human Development Index). Data shown for December 3, 2020; database updated daily. Values closer to one (shaded bright green) indicate female/male parity in digital access. Source: www.digitalgendergaps.org



be burdened by a "looking for our keys under the streetlight" problem: we tend to examine only those groups and phenomena that can be clearly seen by data, while ignoring others. In addition, technology use is heterogeneous within countries—across rural and urban populations, for example—and highly dynamic, changing from year to year. Standard surveys only sometimes capture these nuances.

Specific forms of big data itself, however, can shed light on these questions of digital access, and thereby help deal with the broader issue of making unbiased inferences from big data sources. Data2X grantee partners Ingmar Weber at QCRI and Ridhi Kashyap at Oxford University are working to indirectly estimate the gender digital divide using data from the online populations of LinkedIn, Facebook, Google, and Snapchat users (Figure 5).[73,74] Given the number of users on each of these platforms, estimates disaggregated by age, professional group (e.g., highly skilled professionals versus young adults entering the workforce), and other characteristics are also

possible. The high frequency of use over time allows a real-time portrait of digital gender inequalities—information that takes on critical importance during rapid onset crises like pandemics and recessions. In addition, the narrowing of the gender digital divide, especially with respect to mobile phone use in poor and remote areas of the world, serves not only as a proxy indicator for broad gender equality, but is also itself a catalyst for improved life outcomes for women and girls, including reduced mortality, greater reproductive and sexual health knowledge, and higher contraceptive use.[75]

Many big datasets are not gender-tagged; that is, the gender of the individual user or content-generator is not specified. Prepaid mobile phone datasets are perhaps the most important example, and even postpaid datasets often do not accurately record gender.[53] Given cultural differences in how different genders use mobile phones, gender inference models are still struggling to attain acceptable levels of accuracy. A few large-scale studies show some promise, however, including a recent effort by Mobilewalla researchers to infer gender among half a billion users.[76]

In addition, gender inference in text, voice, and image datasets are reaching high levels of accuracy.[77,78] Our grantee partner Jihad Zahir used YouTube comments to build an Arabic-language gender identification machine learning approach that attains an accuracy of 92% and a precision of 98%.[79,c] In another example, researchers triangulated gender using a combination of name lists, facial recognition based on profile pictures, and other techniques, achieving nearly 97% accuracy.[80]

Re-inserting sociological theory—that is, considering how and why gendered patterns manifest in different kinds of datasets—is critical to big data gender inference moving forward.[81] For example, understanding the cultural context of cell phone use within families could greatly improve gender prediction in mobile operator data. Overall, improved gender inference methods would open up new possibilities for existing datasets. For example, the Africa's Voices Project has created NLP tools to analyze hundreds of thousands of text conversations on multimedia platforms, including discussions among Kenyan and Tanzanian youth about contraception.[50] Performing textual analysis to infer gender in this dataset would lead to better understanding of the different ways in which men and women access and use contraception.

Serious problems of racial and gender bias, as well as privacy risks, still plague classification algorithms.[82–84] Thanks to the work of digital activists, these conversations are now more squarely in the public eye, which bodes well for both algorithmic accuracy and privacy protection. In addition, the detection of gender biases in big data training sets—for example, in newspaper articles and social media posts—can shed light on the nature and intensity of the underlying cultural biases that give rise to these flawed training sets,[43] ideally setting into motion a virtuous cycle of awareness and algorithmic reform.

c. *Accuracy* is a description of systematic errors, capturing the closeness of measurements to the correct gender class, for example. *Precision* is a description of random errors, capturing the closeness of repeated measurements to each other.

Emerging Opportunities

Large technology firms generate gender data critical for policymaking, and researchers are finding ways to use this data meaningfully and responsibly.

Large technology firms hold massive amounts of data that can illuminate gendered topics of global concern. For example, social media platforms are the world's largest databases of thoughts, emotions, and cultural norms. Because of the ease of accessing data, Twitter remains the most commonly analyzed social media data source, especially for textual and sentiment analysis studies. The scope of Twitter gender work is too large to summarize here, but recent studies have been conducted on topics as diverse as the interaction between public discourse and judicial decisions around same-sex marriage,[85,86] gendered strategies of evacuation after natural disasters,[27] and differences in religious attitudes between men and women.[44,87]

Facebook user data has traditionally been more difficult to access, but several studies show its potential for filling gender data gaps. A now-classic work showed that Facebook likes can be used to predict a wide variety of personal attributes, including religious and political views, ethnicity, and personality traits.[88,89] A more recent analysis of the liking practices of nearly 22 million Facebook users in 10 countries shows clear gender differences in civic and political expression,[88] while a study looking at German political party Facebook pages illustrates gendered differences in affiliation.[85] Another social media study of nearly six million users and 19 million posts on Facebook, Twitter, Weibo, and Baidu examines the relationship between gender and emotional responses to behavioral phenomena (in this case, procrastination).[90] Data2X's allies at UNGP, as well as other researchers, have done extensive analyses of Facebook and Twitter discussions about vaccination and other gender-relevant topics in Nigeria, India, and Indonesia.[32,91] Research teams from various universities in China have used data from Weibo, a microblogging site with over half a billion users, to explore gendered patterns of movement in urban areas,[92,93] including a detailed analysis of how public space interacts with gender to produce various mental and emotional states.[94]

The gig economy is also generating large amounts of gendered data. A recent study looked at work choices and earnings among a million Uber drivers in the United States, finding a 7% earnings gap between female and male drivers.[95] The study concludes that labor disparities can prevail even without overt discrimination, in the case of Uber due to gendered differences in work experience, desires to be closer to home and in safer locations, and driving speed. On the consumer side, in-progress work finds that providing Uber subsidies to women in Cairo greatly increases their travel distance and radius traveled, especially among those who have safety concerns about using public transport.[96] This suggests that economic and safety constraints hinder women's preferred mobility patterns. Ride-share companies are actively seeking other research collaborations in India, Brazil, and other middle- and low-income countries.[54]

Big tech companies are not the only large-scale private sector data holder, of course. As profiled throughout this report, valuable inferences about mobility, socio-economic well-being, and other topics can be made from mobile network provider data. The employment and salary records of large firms across sectors is another important and under-explored data source; such records hold key insights

about wage discrimination, "glass ceilings," and other gendered phenomena in the economy. Many other forms of socio-economic data are valuable windows into gender inequality.

Data privacy is the major obstacle to wider sharing of data by large companies. The privacy debate, formerly limited to academic researchers, civil society advocates, and industry personnel, is now in the public eye. Innovative solutions are emerging. For example, Data2X's partners at the Flowminder Foundation are creating secure protocols to access and analyze proprietary cell phone data. Flowminder's "FlowKit" suite of tools enables mobile network providers to share de-identified data on subscriber mobility patterns and network usage.[97,d] Similarly innovative protocols that "get code to the data" and receive results in return—instead of requesting the data itself—may help bring other key data-holders, including the large technology companies, to the table.

The broader point is that many users are willing to share personal data provided that the app brings value and user privacy is protected. The same applies to data-holder companies: if gender use cases powerfully demonstrate how big datasets are critical to policymaking, then data-holders will be more open to collaborations that allow access while ensuring security.

Data collaboratives show the path forward for how big data can make a practical difference in the lives of women and girls worldwide.

The mission of Data2X is broader than simply assuring that more gender data is generated and accessible. We are committed to ensuring that gender data has a tangible impact on the lives of women and girls; even the most innovative datasets and methodologies are worth little if scientific research does not move the world towards gender equality. In the coming years, Data2X will increasingly focus on ensuring that policymakers, civil society organizations, and other advocates for gender equality adopt best practices in the use of big data.

Many of our grantee partners are working towards the same goal. As mentioned earlier, Dr. Borker's work on female entrepreneurs aims at identifying gendered policies to alleviate constraints to small business performance. Dr. Jihad's chatbot is linking victims of domestic violence to necessary services; as smartphone and social media penetration extends to even the most remote areas, such tools may soon be used for estimating population-level estimates of key gender indicators. The Flowminder Foundation, via the GRID3 program, is currently partnering with a broad array of partners in the Democratic Republic of the Congo (DRC)—including DRC's Ministry of Health and other public agencies, Columbia University's Center for International Earth Science Information Network (CIESIN), WorldPop, the UNFPA, and the telecommunications provider Vodacom—to improve the effectiveness and equity of vaccination interventions in priority provinces of the country. The poor coverage of childhood vaccination in DRC is closely related to gendered mobility constraints; Flowminder is using a combination of geospatial, cell phone, and other data sources to develop algorithms that optimize the location of vaccination outreach sites, taking into account the spatial distribution and mobility of the population. Flowminder's

d. See https://flowkit.xyz/ for a detailed description of the FlowKit architecture.

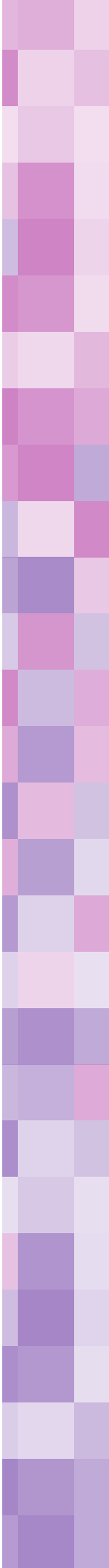
work has supported the DRC in vaccinating hundreds of thousands of additional children in the country over the past several years—a clear example of a big data policy success.

Another grantee partner, the GovLab at New York University, has created and begun to implement a coherent model of stakeholder cooperation that builds on successes like those described above. The model is called the data collaborative, a structure in which public agencies, private data-holders, civil society advocates, and academic researchers create a durable partnership to carry out research that has real-world impacts.[98] The data collaboratives are informed by GovLab's 100 Questions Initiative, an effort to collectively identify the most pressing global policy questions.[99] Defining and prioritizing questions in this way is essential to realizing big data's potential to have an on-the-ground impact, particularly when considering the immense volume of big data available.

Data2X partnered with GovLab to build a Gender domain within the 100 Questions Initiative, as well as to integrate gender into other existing domains (e.g., urban mobility and air quality). This multi-step process included consultations with 90 experts in both gender and data science to identify the most pressing gender-related global questions, and then releasing the questions for public voting. The Gender domain received significant attention—tallying more than one thousand votes from the public—and ten priority questions have been identified. These questions will help focus the efforts of researchers, advocates, and policymakers worldwide.

An open-access data portal that collects and standardizes the format of an array of big gender datasets would be of great value to researchers and create momentum for the creation of collaboratives. The volume of big data will continue to expand tremendously in the coming years. Facilitating access to this data will not only provide gender researchers with complementary information to explore their existing questions—including researchers who may not easily access a data collaborative—but also draw data scientists without a gender background into the orbit of gender work. The unique skills and fresh perspectives of such scientists would be powerful catalysts for the gender data revolution.

The vision of data as a force for policy impact, and social good generally, is at the heart of Data2X's mission. All forms of data, big and traditional alike, capture a slice of reality. They are thus value-neutral in their basic character; their impact on social and environmental well-being depends on the cultural frames used to interpret and deploy them. Data2X's frame is gender equality: specifically, the idea that women and girls have a right to access the collective resources of humanity, and to participate in the building of a better world. In the years to come, Data2X and its partners will work to advance this frame and ensure that gender data of all kinds produces real and lasting improvements in the lives of women, girls, and all people.



References

1. Grantham, K. (2020). Mapping Gender Data Gaps: An SDG Era Update.
2. Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META group research note 6, 1.
3. Letouzé, E. (2015). Big data and development: General overview primer. Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative.
4. Lopes, C. A. & Bailur, S. (2018). Gender equality and big data: Making gender data visible. UN Women.
5. Vaitla, B. (2014). The landscape of big data for development. Data2X.
6. Vaitla, B., Bosco, C., Alegana, V. et al. (2017). Big Data and the Well-Being of Women and Girls: Applications on the Social Scientific Frontier. Washington DC: Data2X.
7. Vaitla, B., Adler, N., Al Mouatamid, Y. et al. (2019). Big Data, Big Impact? Towards Gender-Sensitive Data Systems.
8. Schroeder, A., Kishore, N., Vembar, N., & Dresser, C. (2020). Displacement, gender disparities, and shelter utilization after Hurricane Laura. CrisisReady blog post.
9. Global Health 50/50. (2021). The sex, gender and COVID-19 project. Global Health 50/50, ICRW, African Population and Health Research Center.
10. Dehingia, N. & Raj, A. (2021). Sex differences in COVID-19 case fatality: do we know enough? *Lancet Glob Health* 9, 1: e14-e15. doi:10.1016/S2214-109X(20)30464-2.
11. Buvinic, M., Noe, L. & Swanson, E. (2020). Tracking the Gender Impact of COVID-19: An Indicator Framework. Data2X, Open Data Watch.
12. Zamarro, G., Perez-Arce, F. & Prados, M. J. (2020). Gender differences in the impact of COVID-19.
13. Alon, T., Doepke, M., Olmstead-Rumsey, J. & Tertilt, M. (2020). The Impact of COVID-19 on Gender Equality. NBER Working Papers 26947. doi:10.3386/w26947.
14. Xue, B. & McMunn, A. (2020). Gender differences in the impact of the Covid-19 lockdown on unpaid care work and psychological distress in the UK. doi:10.31235/osf.io/wzu4t.
15. Reichelt, M., Makovi, K. & Sargsyan, A. (2020). The impact of COVID-19 on gender inequality in the labor market and gender-role attitudes. European Societies. doi:10.1080/14616696.2020.1823010.
16. Zahir, J. (2020). Personal communication. Interview on November 6, 2020.
17. The Governance Lab, New York University. (2021). Data4COVID19: a living repository of data collaboratives seeking to address the spread of COVID-19.
18. Bengtsson, L. (2020). Personal communication. Interview on November 9, 2020.
19. Cheney, C. (2020). Will these 3 gender data trends outlast the pandemic? Devex. Published 23 November 2020.
20. Ladysmith. (2020). Survey on gender equality at home report: A gender data snapshot of life during COVID-19. Facebook Data for Good.
21. Pulse Lab Jakarta. (2020). Alternative use of traditional data against COVID-19. UN Global Pulse.
22. Nguyen, Q. C., Huang, Y., Kumar, A. et al. (2020). Using 164 million Google Street View images to derive built environment predictors of COVID-19 cases. Int. J. Environ. Res. Public Health 17, 17: 6359. Doi: 10.3390/ijerph17176359.
23. Taub, A. A. (2020). New COVID-19 crisis: Domestic abuse rises worldwide. The New York Times. Updated 14 April 2020.

24. Gauvin, L., Tizzoni, M., Piaggesi, S. et al. (2020). Gender gaps in urban mobility. Humanities and Social Sciences Communications 7:11.

25. Di Clemente, Luengo-Oroz, M., Travizano, M. et al. (2018). Sequences of purchases in credit card data reveal lifestyles in urban populations. Nature Communications 9: 3330. doi: 10.1038/s41467-018-05690-8.

26. WorldPop & Flowminder Foundation. (2020). Five-year sex-disaggregated migration movements.

27. Martín, Y., Cutter, S. L. & Li, Z. (2020). Bridging Twitter and survey data for evacuation assessment of Hurricane Matthew and Hurricane Irma. Natural Hazards Review 21, 2: 04020003. doi: 10.1061/(ASCE)NH.1527-6996.0000354.

28. Masso, A., Silm, S. & Ahas, R. (2019). Generational differences in spatial mobility: A study with mobile phone data. Popul. Space Place 25, 2: e2210.

29. Rhoads, D., Serrano, I., Borge-Holthoefer, J. & Solé-Ribalta, A. (2020). Measuring and mitigating behavioural segregation using Call Detail Records. EPJ Data Science 9: 5.

30. Saha, K., Torous, J., Caine, E. D. & De Choudhury, M. (2020). Psychosocial effects of the COVID-19 pandemic: Large-scale quasi-experimental study on social media. J. Med. Internet Res. 22, 11: e22600. doi: 10.2196/22600.

31. Jimenez, A., Santed-Germán, M.-A. & Ramos, V. (2020). Google searches and suicide rates in Apain, 2004-2013: Correlation study. JMIR Public Health Surveill 6, 2: e10919.

32. Tomeny, T. S., Vargo, C. J. & El-Toukhy, S. (2017). Geographic and demographic correlates of autism-related anti-vaccine beliefs on Twitter, 2009-15. Soc. Sci. Med. 191: 168–175. doi: 10.1016/j.socscimed.2017.08.041.

33. Priest, C., Knopf, A., Groves, D. et al. (2016). Finding the patient's voice using big data: Analysis of users' health-related concerns in the ChaCha question-and-answer service (2009–2012). Journal of Medical Internet Research 18, 3: e44. doi: 10.2196/jmir.5033.

34. Moreira, M. W. L., Joel J. P., Kumar, N. et al. (2019). Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. Information Fusion 47: 23–31.

35. Nambisan, P., Luo, Z., Kapoor, A. et al. (2015). Social media, big data, and public health informatics: Ruminating behavior of depression revealed through Twitter. 2015 48th Hawaii International Conference on System Sciences. doi:10.1109/hicss.2015.351.

36. De Choudhury, M., Counts, S., Horvitz, E. J. & Hoff, A. (2014). Characterizing and predicting postpartum depression from shared Facebook data. Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing - CSCW '14. doi:10.1145/2531602.2531675.

37. Nebeker, C., Dunseath, S. E. & Linares-Orozco, R. (2020). A retrospective analysis of NIH-funded digital health research using social media platforms. Digit Health 6: 2055207619901085.

38. Suryavanshi, N., Kadam, A., Kanade, S. et al. (2020). Acceptability and feasibility of a behavioral and mobile health intervention (COMBIND) shown to increase uptake of prevention of mother to child transmission (PMTCT) care in India. BMC Public Health 20, 752. doi: 10.1186/s12889-020-08706-5.

39. Donthineni, P. R., Kammari, P., Shanbhag, S. S. et al. (2019). Incidence, demographics, types and risk factors of dry eye disease in India: Electronic medical records driven big data analytics report I. Ocul. Surf. 17, 2: 250–256.

40. Vermund, S. H. (2019). Use of big data to identify risk of adverse HIV outcomes. The Lancet HIV 6, 8: e488–e489.

41. Mohan, D., Bashingwa, J. J. H., Dane, P. et al. (2019). Use of big data and machine learning methods in the monitoring and evaluation of digital health programs in India: An exploratory protocol. JMIR Res. Protoc. 8, 5: e11456.

42. Andriano, L. & Behrman, J. (2020). The effects of growing-season drought on young women's life course transitions in a sub-Saharan context. Population Studies 74, 3: 331–350.

43. Friedman, S., Schmer-Galunder, S., Chen, A. & Rye, J. (2019). Relating word embedding gender biases to gender gaps: A cross-cultural analysis. Proceedings of the First Workshop on Gender Bias in Natural Language Processing. doi:10.18653/v1/w19-3803.

44. Thomas, J., Al Shehhi, A. & Grey, I. (2019). The sacred and the profane: social media and temporal patterns of religiosity in the United Arab Emirates. Journal of Contemporary Religion 34, 3: 489–508.

45. Kalimeri, K., Beiró, M. G., Delfino, M., Raleigh, R. & Cattuto, C. (2019). Predicting demographics, moral foundations, and human values from digital behaviours. Computers in Human Behavior 92: 428-445.

46. Gallus, J. & Bhatia, S. (2020). Gender, power and emotions in the collaborative production of knowledge: A large-scale analysis of Wikipedia editor conversations. Organizational Behavior and Human Decision Processes 160: 115–130.

47. Lee, S. Y. & Ryu, M. H. (2019). Exploring characteristics of online news comments and commenters with machine learning approaches. Telematics and Informatics 43: 101249.

48. UN Global Pulse. (2019). Using radio broadcasts to augment early detection of health risks. (2019).

49. UN Global Pulse. (2019). Ending violence against women and girls in Uganda.

50. Africa's Voices Foundation. (2018). NLP of conversations in Swahili Slang (Sheng) (Well Told Story).

51. Africa's Voices Foundation. (2020). Using digital civic engagement to support land health interventions and the SHARED methodology in the Upper Tana river basin (ICRAF).

52. Africa's Voices Foundation. (2020). COVID19 – Kenya: Trusted two-way mass and individual health communications and rapid socio-epidemiological insights.

53. Flowminder Foundation. (2019). Towards high-resolution sex-disaggregated dynamic mapping.

54. Borker, G. (2020). Personal communication. Interview on November 6, 2020.

55. Verkroost, F. C. J., Kashyap, R., Garimella, K. et al. (2020). Tracking global gender gaps in information technology using online data. Digital Skills Insights 81–93.

56. Huluba, A.-M., Kingdon, J. & McLaren, I. (2018). The UK Online Gender Audit 2018: A comprehensive audit of gender within the UK's online environment. Heliyon 4, 12: e01001.

57. Murthy, S. (2015). Measuring gender diversity with data from LinkedIn. Linkedin Official Blog.

58. Widerstedt, B., Månsson, J. & Rosdahl, J. (2018). A warm welcome? Access to advisory services for men and women. Economic Analysis and Policy 58: 100–110.

59. UN Global Pulse. (2018). Big data for financial inclusion, examining the customer journey.

60. UN Global Pulse. (2017). Exploring the potential of mobile money transactions to inform policy.

61. UN Global Pulse. (2016). Using financial transaction data to measure economic resilience to natural disasters.

62. UN Global Pulse. (2019). Mapping financial services points across Indonesia.

63. Zahir, J. (2020). NajatBot.

64. Data-Pop Alliance. (2020). Using data to shed light on the shadow pandemic of domestic violence in Mexico. Data-Pop Alliance DataFeed.

65. Karakurt, G., Patel, V., Whiting, K. & Koyutürk, M. (2017). Mining electronic health records data: Domestic violence and adverse health effects. J. Fam. Violence 32, 1: 79–87.

66. Chenou, J.-M. & Cepeda-Másmela, C. (2019). #NiUnaMenos: Data activism from the global South. Television & New Media 20, 4: 396–411.

67. Larrondo, A., Morales-i-Gras, J. & Orbegozo-Terradillos, J. (2019). Feminist hashtag activism in Spain: measuring the degree of politicisation of online discourse on #YoSíTeCreo, #HermanaYoSíTeCreo, #Cuéntalo y #NoEstásSola. Communication & Society 32, 4: 207–221.

68. González, G. (2019). Escraches en redes feministas universitarias: una estrategia contra la violencia de género hacia las mujeres. Comunicación y Medios 28, 40. doi:10.5354/0719-1529.2019.53974.

69. Heerwig, J. A. & Gordon, K. M. (2018). Buying a voice: Gendered contribution careers among affluent political donors to federal elections, 1980-2008. Sociological Forum 33, 3: 805–825.

70. WorldPop. (2020). WorldPop demographics subnational age/sex structures - 2000-2020.

71. WorldPop. (2020). Access to health services. SDGs Today: The Global Hub for Real-Time SDG Data.

72. Vega Montiel, A. (2018). Gender equality and big data in the context of the sustainable development goals. Partecipazione e conflitto 11, 2: 544–556.

73. Kashyap, R., Fatehkia, M., Al Tamime, R. & Weber, I. (2020). Monitoring global digital gender inequality using the online populations of Facebook and Google. Demographic Research vol. 43 779–816.

74. Fatehkia, M., Kashyap, R. & Weber, I. (2020). Using Facebook ad data to track the global digital gender gap. Demographic Research 43, 27: 779-816. doi:10.31235/osf.io/rkvb3.

75. Rotondi, V., Kashyap, R., Pesando, L. M. et al. (2020). Leveraging mobile phones to attain sustainable development. Proc. Natl. Acad. Sci. USA. 117, 24: 13413–13420.

76. Sangaralingam, K., Verma, N., Ravi, A. et al. (2018). Predicting age & gender of mobile users at scale - a distributed machine learning approach. 2018 IEEE International Conference on Big Data. doi:10.1109/bigdata.2018.8621942.

77. Jia, S., Lansdall-Welfare, T. & Cristianini, N. (2016). Gender classification by deep learning on millions of weakly labelled images. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). doi:10.1109/icdmw.2016.0072.

78. Manik, L. P., Syafiandini, A. F., Mustika, H. F. et al. (2019). Gender inference based on Indonesian name and profile photo. 2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA). doi:10.1109/ic3ina48034.2019.8949589.

79. Zahir, J., Oukaja, Y. M. & Mousannif, H. (2019). Author gender identification from Arabic YouTube comments. 2019 15th International Conference on Signal-Image Technology Internet-Based Systems (SITIS) 672–676.

80. Yang, H. & Yuan, Y. (2016). A general gender inference method based on web. Proceedings of the 2016 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2016) doi:10.2991/aiie-16.2016.49.

81. Radford, J. (2017). Piloting a theory-based approach to inferring gender in big data. 2017 IEEE International Conference on Big Data. doi:10.1109/bigdata.2017.8258555.

82. Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research 81: 1-15.

83. Thelwall, M. (2018). Gender bias in machine learning for sentiment analysis. Online Information Review 42, 3: 343–354.

84. Zhao, J., Wang, T., Yatskar, M. et al. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: 2979-2989. doi:10.18653/v1/d17-1323.

85. Flesch, B., Vatrapu, R. & Mukkamala, R. R. (2017). A big social media data study of the 2017 german federal election based on social set analysis of political party Facebook pages with SoSeVi. 2017 IEEE International Conference on Big Data. doi:10.1109/bigdata.2017.8258236.

86. Clark, T. S., Staton, J. K., Wang, Y. & Agichtein, E. (2018). Using Twitter to study public discourse in the wake of judicial decisions: Public reactions to the Supreme Court's same-sex-marriage cases. Journal of Law and Courts 6, 1: 93–126.

87. Boy, J. D., Uitermark, J. & Wiersma, L. (2018). Trending #hijabfashion: Using big data to study religion at the online–urban interface. Nordic Journal of Religion and Society 31, 1: 22–40.

88. Brandtzaeg, P. B. (2017). Facebook is no "great equalizer": A big data approach to gender differences in civic engagement across countries. Social Science Computer Review 35, 1: 103–125.

89. Kosinski, M., Stillwell, D. & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. U. S. A. 110, 15: 5802–5805.

90. Chen, Z., Zhang, R., Xu, T. et al. (2020). Emotional attitudes towards procrastination in people: A large-scale sentiment-focused crawling analysis. Computers in Human Behavior 110: 106391.

91. UN Global Pulse. (2015). Understanding immunisation awareness and sentiment through analysis of social media and news content.

92. Rizwan, M. & Wan, W. (2018). Big data analysis to observe check-in behavior using location-based social media data. Information 9, 1: 257.

93. Ullah, H. et al. (2020). Spatiotemporal patterns of visitors in urban green parks by mining social media big data based upon WHO reports. IEEE Access 8: 39197–39211.

94. Ma, Y., Ling, C. & Wu, J. (2020). Exploring the spatial distribution characteristics of emotions of Weibo users in Wuhan waterfront based on gender differences using social media texts. ISPRS International Journal of Geo-Information 9, 8: 465.

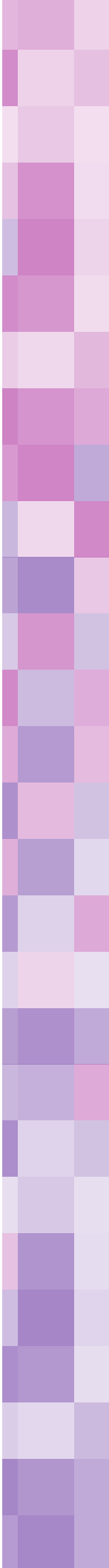
95. Cook, C., Diamond, R., Hall, J. et al. (2018). The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers. NBER Working Papers 24732. doi:10.3386/w24732.

96. Christensen, P. (2020). Mobility constraints: Evidence from an experiment with Uber riders. Presentation at Advances with Field Experiments, September 23-24, 2020.

97. The Flowminder Foundation. (2020). FlowKit. Flowminder/FlowKit.

98. GovLab, New York University. Data collaboratives.

99. GovLab, New York University. (2020). The 100 questions initiative.



data2x