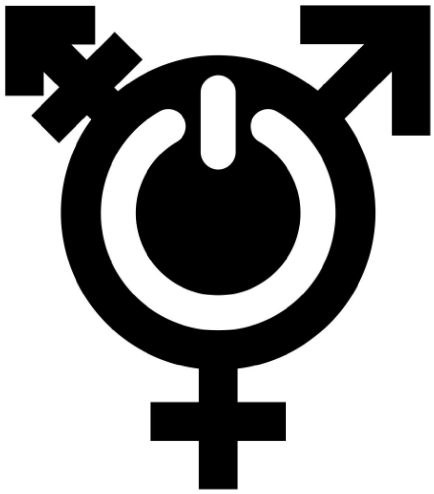
A Primer on Non-Binary Gender and Big Data

by kanarinka | Published June 3, 2016



Created by TNS
from Noun Project

This is a draft piece for a larger project called The Visual Catalogue of Uncertainty in Data that I'm just beginning with Mushon Zer-Aviv that seeks to catalog the ways in which the data never, under any circumstances, speak for themselves.

Gender data is often more complicated than it appears on the surface. Our received wisdom is that there are two categories: the world is made up of men and women. And yet, the historical record shows that there have always been more variations in gender identity than society has cared to outwardly acknowledge or collectively remember. These third, fourth and n-th genders go by different names in the different historical and cultural circumstances in which they originate, including female husbands, indigenous berdaches, Hijras, two-spirits, pansy performers, and sworn virgins.

In contemporary Western thought, there are several different but related concepts to untangle when it comes to sex and gender. *Biological sex* refers to a person's chromosomal and anatomical make-up whereas *gender identity* refers to the what gender a person feels themselves to be. *Gender expression* is the gender that a person presents to the outside world. Biological sex, gender identity and gender expression are aligned for most people, but not for people who identify as transgender, queer and gender-queer. And gender identity is different from *sexual orientation* which refers to being romantically attracted to people of a particular gender (or possibly more than one gender). Sexual orientations range from gay and straight to bisexual, pansexual and more. The site OKCupid defines 13 sexual orientations.

Moreover, rather than seeing either biological sex or gender as immutable and given, there are ways that these identities shift, change and may be directly manipulated over time. In the case of biological sex, this includes gender reassignment surgery for intersex infants born with "abnormal" genitals as well as medical and surgical transition for transgender people who are transitioning from one gender to another. Some individuals feel that their gender identity shifts from day to day or situation to situation – a concept known as *gender fluidity*. A newer term – *gender datafication* – can be used to refer to the external digital classification of gender and its representation in databases and code.

Non-Binary Gender and the State

Official, state-sanctioned acknowledgement and rights for sexual and gender minorities have been expanding in Western democracies over the past fifty years. There are nineteen countries that recognize same-sex marriage. Nations around the world have varied and uneven ability for individuals to officially amend their sex. Some, like Japan, mandate hormone therapy and surgery in order to be legally recognized as another gender. Only three countries as of 2015 allow individuals to self-determine their gender, including Ireland, Denmark, and Malta. Amnesty International considers LGBTQIA rights as human rights. In practice, sexual minorities and gender non-conforming people face harassment, discrimination and violence even in the most "enlightened" places despite the fact that they represent a significant sub-population. A study out of UCLA estimated that 0.3% of the US population is transgender and 3.5% is not straight in sexual orientation. This translates into 9 million individuals, about the population of New Jersey.

Non-Binary Gender and Social Media

Recognition of non-binary sex and gender differences has begun, slowly, to extend itself into the technological realm, primarily for social media companies responding to user pressure. In 2014, Facebook expanded their gender options from 2 to 58 for English speakers in the US and UK. The gender options they added were created in consultation with the LGBTQIA community and range from "gender non-conforming" to "two-spirit" to "trans female". The corporation later added the ability to identify as more than one gender and to input a custom gender. Other social networking and dating sites like Google+, OKCupid and Match.com have followed suit. While these changes may appear to be progressive, Facebook's databases still resolve custom and non-binary genders to Male and Female on the backend based on the binary gender that users select at sign-up where the custom option is not available. Here is how the Facebook Marketing API views gender: 1 = Male, 2 = Female. So while a user and her friends may see her presented as the gender she elects, she is a 1 = Male or 2 = Female to any advertisers looking to purchase her attention.

Automatically Detecting Gender

Computational, Big Data and artificial intelligence applications that deal with gender have invariably treated it as a binary. Competitions on Kaggle, a popular web platform for predictive modelling and analytics, have sought to predict gender from fingerprints and handwriting. Other work has sought to automatically identify the gender of bloggers, novelists, movie reviewers and Twitter users based on the style of their language. There are multiple libraries for predicting the gender of people based on their name, like OpenGenderTracker and the controversially named "Sex Machine" Ruby Gem (now called GenderDetector). Nathan Matias has a comprehensive account from 2014 of more of this research, including different uses, methodological choices and ethical guidelines.

While these applications seek to generalize about majority populations who largely do fall within the binary categories of male and female, they reinforce the idea that the world is *only* made up of these two groups which is categorically, empirically, and historically untrue. Moreover, these works tend to codify (literally, to write into code) essentialist, stereotypical characterizations of male and female communication patterns and present them as universal, context-free, scientific truths. For example: "women tend to express themselves with a more emotional language"; "men are more proactive, directing communication at solving problems, while women are more reactive". As we know from disciplines like Communication Studies, Geography and Science and Technology Studies (STS), representations do not always reflect reality but also have a role in producing it. This applies to code and statistical modeling just as it does to visualizations, images and videos. And when things are left un-represented they effectively do not exist.

Non-binary genders will always be outliers

Trans and gender non-conforming people will represent statistical outliers and minorities in any dataset that collects gender (around 0.3% if the UCLA study is correct) just as Native Americans will represent a small minority in any US-based data set that collects race. But this is not a good reason to simply ignore this group. As Brooke Foucault-Welles states, "When women and minorities are excluded as subjects of basic social science research, there is a tendency to identify majority experiences as 'normal,' and discuss minority experiences in terms of how they deviate from those norms". Minority experiences are relegated to the margins of analysis or, as mostly happens with trans people in relation to computation and gender, excluded altogether. Instead of ignoring these statistical outliers (which has dubious ethical and empirical implications and has even been called "demographic malpractice"), Foucault-Welles proposes that data scientists use minority experiences as reference categories in themselves. This means not just collecting more than two genders but also disaggregating any data processing, analysis and results based on these categories.

The risks and challenges of collecting non-binary gender data

At the same time, depending on what data is being collected and whether it is personally identifiable (or easily de-anonymized) it is important to recognize the potential risk of stating one's gender as something other than male or female. Because the sample sizes will be so small, these individuals may possibly be identified even within otherwise large data sets. Even when individuals do not volunteer this information to an application, it may be possible to algorithmically infer gender or sexual orientation from knowledge of their social networks. This can pose risks of repercussion, either in the form of personal shame for people who have hidden their gender identity or even discrimination, violence and imprisonment depending on the context and community where they live.

There are also challenges to collecting information about non-binary genders. As we can see from historical studies of non-binary gender, how many and which other genders exist depends heavily on culture and context. For example, the government of Nepal attempted to add to their census the category of "Third Gender" but gender minority communities, more likely to consider themselves *Kothi* or *Methi*, did not identify with this term. The Williams Institute at UCLA and the Human Rights Campaign provide short guides for collecting non-binary gender data. Just providing more choices in a drop-down menu is not always the best path. Depending on the circumstances, the most ethical thing to do might be to avoid collecting gender data, make gender optional or even stick with binary gender categories. Nathan Matias and Sarah Szalavitz chose to do this for their application FollowBias which detects gender from names in order to avoid outing someone's gender identity against their wishes. And if gender data is going to be used in processes with known structural inequalities such as hiring and promotion the most ethical action might be to entirely obscure a person's gender from either the human decision makers or the algorithms making discriminatory decisions in order to avoid bias.

In summary, non-binary gender and data represents complicated terrain for computational applications for numerous reasons. But we have an ethical and empirical imperative to tackle this complexity. The world is not and has never been comprised of only two genders. To assume gender is a simple binary is simply wrong.

What do you think? In the interest of strengthening this piece, I'd love to hear about others' work in collecting, securing and analyzing gender data beyond the binary. What are best practices? What are the urgent research questions and knowledge gaps? What potential insights might we derive from working with non-binary gender and data? What are the risks to gender minorities in relation to data? What kinds of variation do we see across culture, context and history? How might non-binary gender and data deal with intersectionality?

Post responses in the comments below or @kanarinka on Twitter.

‹ LIVEBLOGGING #ODR2016: AFTERNOON SESSI...

PRACTICING DATA SCIENCE RESPONSIBLY ›

Upcoming Events

THU 20 Open Virtual Civic Meeting (Spring 2020)
August 20 @ 1:00 pm - 2:30 pm

THU 27 Open Virtual Civic Meeting (Spring 2020)
August 27 @ 1:00 pm - 2:30 pm

View More...