Your Electronic Self:
Surveillance and Social Sorting

Christian Eichenmüller
October 20, 2014

Christian Eichenmüller is Visiting Scholar and Research Assistant at the Baker Institute for Peace and Conflict Studies at Juniata College.

Surveillance has received a fair share of daily news coverage for more than a year now, stimulated most obviously by revelations regarding the US National Security Agency leaked by Edward Snowden in 2013. Today I want to focus on social sorting that does or could occur as a consequence of surveillance. Social sorting can be understood as a process of categorization that leads to differential treatment of people or differential access to services. I will address the following talking points: a comparison between the amount of information collected by the Stasi and the amount available to the National Security Agency (NSA); a brief discussion of so-called metadata; the question of a new information environment that might require new ways of sense-making, especially in light of too much information; and finally, a couple of questions we can explore together.

Let’s start with the Stasi, East Germany’s infamous secret police renowned for their meticulous, all-hearing, all-seeing surveillance regime. It is estimated that one in six people were Stasi informants in East Germany, observing their fellow citizens with prying eyes on behalf of the state. The state’s obsessive information collection led to file cabinets so huge that they would cover an area of 0.019 sq. km in the heart of Berlin. All reports and statements in these files are written by hand or by typewriter—millions of sheets of paper in hundreds of thousands of files.

After the revelations of former NSA contractor Edward Snowden, German president Joachim Gauck, who himself boasts a history of defending civil liberties in the former German Democratic Republic (GDR), and who was an ardent opponent of the Stasi at the time, concluded in an interview that the NSA was not to be compared with the Stasi. “We know, for example, that it is not like it was with the Stasi and the KGB – that there exist big filing cabinets in which all the content of our conversations are written down and nicely filed. This is not the case.”[1]

Of course, Joachim Gauck is correct. In the twenty-first century no one is scribbling down observation reports anymore, but Gauck obfuscates the problem by stating the obvious. Perhaps too deeply uncomfortable to admit his own implication in contemporary surveillance regimes, he hides behind a troublesome, but politically convenient, statement.

How then is surveillance conducted nowadays? If nobody is anymore writing down what you are doing, where you are, and with whom you are meeting, then how do authorities keep track of you? The answer is as uncomfortable as it is obvious: through automated systems and the devices you carry around and use on a daily basis.

Advances in information processing and storage capacities have made information collection ubiquitous. The amount of information that can be stored nowadays has far exceeded anything the Stasi could have imagined in their day. How many filing cabinets would the Stasi have had to fill up if they had collected the amount of information available to the NSA and printed it out? The answer: They would have needed file cabinets covering an area of seventeen million sq. km to hold the NSA's five zetabytes of information, printed out and filed away. That is more than the surface area of the entire United States.

A world in which everything you do is being tracked and stored is a world humans have no experience inhabiting. In this new world everything can be datafied—that is, it can be translated into quantifiable data. A rather extreme example, taking datafication on a day-to-day basis to unprecedented levels, is the Quantified Self Movement, "a group of fitness aficionados, medical maniacs, and tech junkies who measure every element of their bodies and lives in order to live better—or at least to learn new things they couldn't have known in an enumerated way before."[2] This kind of self-surveillance to "live better" and "learn more" can take various forms; whether pulse rate, blood-pressure, or brainwaves, nothing is beyond measurement. As a consequence, pseudo-scientific objective measures can now be integrated into people's lives, enabling a continuous do-it-yourself surveillance praxis. As bodies are turned into aggregate sets of numbers, they can be compared, assessed, and weighed against each other, bringing any deviation and or transgression from the norm into stark relief.

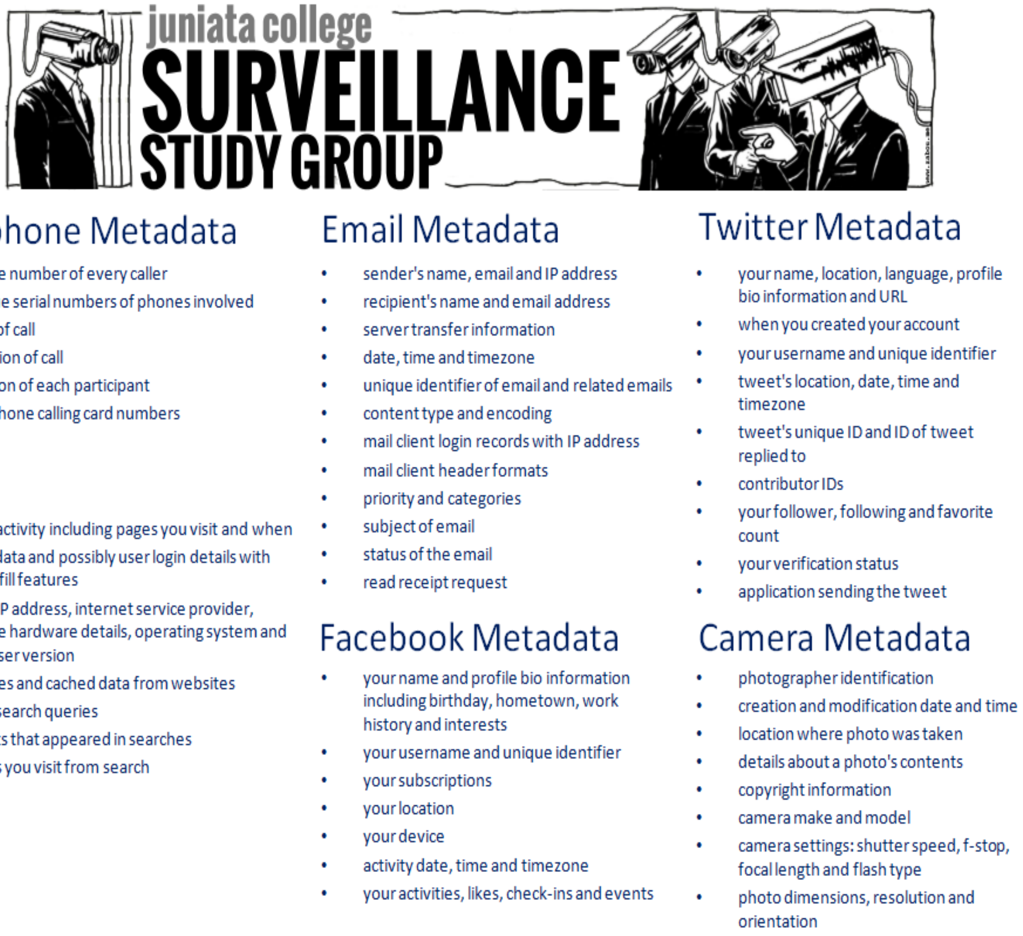
However, those who voluntarily participate in this movement are not the only ones who have provided troves of data on themselves to unaccountable entities. As long as you have a cell phone, you can count yourself as an accumulating data set, building what can be called your data double. If you surf the web; have a free email account; use apps on your phone, your iPad, or your personal computer; make purchases using credit cards; visit social media sites; or, increasingly, even attempt to avoid these things, you are still likely generating a trail of data as you bump up against digital society throughout your day.

I'd like to initially raise some key questions about this world of datafied reality conceptions. What happens with these incredibly large amounts of (personal) information? Who owns these large troves of information to which everybody contributes, but which are harvested by only very few entities that reap the benefits? What does it mean to have one's data collected, stored, and assessed by a few powerful state agencies and corporations?

Another question is whether your data double—that is, the sum of the information collected on you—is an accurate portrayal. Do data doubles match up with flesh and blood human beings? And can the gap between you and your data double be closed by ever more data? Or, in other words, by more intrusive surveillance? We might come back to these questions and the implications of various potential responses, but let's first look briefly at how your data double might be brought to life in the first place.

The Juniata College Surveillance Study Group, a group of students and sometimes faculty, have been examining for the past few months what exactly one refers to when speaking of metadata.[3] Metadata is all the information originating with the use of communication and information technologies, except for the content itself. Thus your telephone metadata would include the following: the phone number of every caller; the unique serial numbers of phones involved; the time of call; the duration of the call; the location of each participant; and the telephone calling card numbers.

There is no mention of the content of your phone call. Nobody would get to know the exact words you spoke (although as voice recognition software improves this could change quickly). Those state agencies and corporations collecting everybody's metadata will only see with which numbers you have been in touch, how often, how long, and from where.



juniata college
SURVEILLANCE
STUDY GROUP

Telephone Metadata

- phone number of every caller
- unique serial numbers of phones involved
- time of call
- duration of call
- location of each participant
- telephone calling card numbers

- your activity including pages you visit and when
- user data and possibly user login details with auto-fill features
- your IP address, internet service provider, device hardware details, operating system and browser version
- cookies and cached data from websites
- your search queries
- results that appeared in searches
- pages you visit from search

Email Metadata

- sender's name, email and IP address
- recipient's name and email address
- server transfer information
- date, time and timezone
- unique identifier of email and related emails
- content type and encoding
- mail client login records with IP address
- mail client header formats
- priority and categories
- subject of email
- status of the email
- read receipt request

Facebook Metadata

- your name and profile bio information including birthday, hometown, work history and interests
- your username and unique identifier
- your subscriptions
- your location
- your device
- activity date, time and timezone
- your activities, likes, check-ins and events

Twitter Metadata

- your name, location, language, profile bio information and URL
- when you created your account
- your username and unique identifier
- tweet's location, date, time and timezone
- tweet's unique ID and ID of tweet replied to
- contributor IDs
- your follower, following and favorite count
- your verification status
- application sending the tweet

Camera Metadata

- photographer identification
- creation and modification date and time
- location where photo was taken
- details about a photo's contents
- copyright information
- camera make and model
- camera settings: shutter speed, f-stop, focal length and flash type
- photo dimensions, resolution and orientation

Figure 1: Overview on metadata of various information & communication technologies.
Source: Juniata College Surveillance Study Group, https://surveillancestudygroup.wordpress.com

At first, metadata collection can sound innocuous, but add to this list e-mail metadata, or web browsing metadata, and you can know a person's social life to an unprecedented extent. The Stasi would have had a field day. In the old days it was extraordinarily difficult to determine who was talking to whom. Getting to know people's relationships and social networks was time-consuming hard work. In the old days people recorded observation reports on typewriters. Fast forward to today and we see that this old system has been replaced by automated systems that collect, store, and process everybody's information. Facebook and other social media companies are ideal surveillance platforms, providing those in charge of the infrastructure with maps of the social landscape.

What happens once the data are beyond your reach, on the web, or owned by Facebook, or Google, or LinkedIn?

Understanding digital data means getting used to the idea of function creep. Data used in one context can be used in a completely different context in the future. Function creep is not entirely new, as the following horrific example from Viktor Mayer-Schönberger's book *Delete* goes to show:

> In the 1930s, the Dutch government had put in place a comprehensive population registry containing name, birth date, address, religion, and other personal information for each citizen. The registry was hailed as facilitating government administration and improving welfare planning. Then the Nazis invaded the Netherlands and took possession of the registry, ruthlessly repurposing the personal information of millions of Dutch citizens to identify, persecute, and murder Jews and gypsies. Because of the information contained in the comprehensive registry, the Nazis were able to identify, deport, and murder a much higher percentage (73 percent) of the Dutch Jewish population than in Belgium (40 percent), France (25 percent), or any other European nation.[4]

Embedded in this example is a form of function creep—the information was repurposed and used in an originally unforeseen way. Compared to older forms, digital information is more transferrable, more mobile, and more modifiable. As a consequence the risk of function creep increases. Storage capacities have also been steadily rising. All of these factors make it much more likely that your data will be used for purposes you don't yet know or will never learn about. This trend is not just relevant in terms of consumer capitalism, but your information might become equally relevant for law enforcement institutions, political campaigns, or the military, to name a few.

And this is where processes of social sorting come to play a major role. Anybody who has personal information on large numbers of people will be incentivized by profit-making or other imperatives to make use of it. The nature of digital information will lead to you being described in a series of binary "forks in the road." If one value applies to you, you end up in this group; if a second value applies, you end up in another group. Numerical data sets leave no gray areas, and thus when they are employed for analysis they produce delineations, lists, bounded groups. They produce categories. Whoever, or whatever, uses your metadata will proceed to make decisions about which categories you do and do not belong to. Whether airport security, your health insurance provider, an amazon.com algorithm, or a watch list run by some intelligence agency, you have been and will be categorized based on your personal information. Or to put it differently, in the case of automated systems, your personal information categorizes you.

Perhaps you belong in the category that receives an insurance discount, or perhaps in the category that forces you to spend between eight to twelve hours in airport security before boarding an airplane. Needless to say, a world in which our digital footprints slot us into pre-conceived categories is not an egalitarian one.

How did we get here? How did this world come about, in which everything we do might be recorded, with potential repercussions five, ten, or fifteen years from now?

In his book *iSpy*, Mark Andrejevic argues that information collection and surveillance have largely been driven by the economic system.[5] Beginning with Taylorism in factories, information was collected to increase the efficiency of workers in the production process. These efficiency increases allowed for mass production on an unprecedented scale. However, producing for mass markets requires the respective consumers. Producers therefore had to get to know their customers – or better, they had to produce them. And so the consumer was born—the subject of consumption and object of study for market researchers. What will you buy? What will you spend money on? The corporation that knows the answer to these questions owns the recipe to get rich. We see how information collection has expanded from the sites of production to include the realm of consumption.

But we are far beyond the pesky customer survey. Staggering advances in information technologies have transformed the way that humans (and market researchers) interact with their surroundings, and how they observe and attempt to make sense of them. It is by the virtues of these information technologies that every transaction and interaction generates a recorded trail of data points, the endless haystack of information that continually tempts more elaborate analytics, smarter institutions, and an ever wider dragnet. Ultimately, we are dealing with a new information environment, and it is changing and challenging decision-making capabilities in novel ways. Unlike past eras where one might have faced an information scarcity when attempting to form judgments or make decisions, today there may actually be too much information. With a relevant database on every real and virtual corner, how are individuals and institutions to manage the process of sense-making?

For many, the answer is an increasingly popular buzzword: Big Data.

The promise of Big Data is to be able to see phenomena in a new way—in their totality. If I have data on every email sent, and every phone call made by everybody and at all times, wouldn't I have the most complete picture of social reality? And what would that mean? Politically? Socially? Technologically?

In his later book *Infoglut*, Mark Andrejevic writes about the challenges that come with new ways of sense-making and an overreliance on quantifiable reality conceptions.[6] Comprehension, in its traditional sense, gets more difficult. Thus we might be entering a period of post-comprehension, in which actionable intelligence trumps genuine understanding. If the data say that you vote Republican when you buy a particular razor brand, and if that assertion holds up repeatedly, it might be good enough. Causation in this example has been replaced by correlation. Will this become our general condition?

It is fascinating to see us all caught up in a large structural transformation, rapidly altering our lives and subjectivities. However, these developments should also stimulate us to think very hard and make conscious choices about what we give up, and what we take on.

In closing, I'd like to simply pose some provocative questions that point us in the direction of such critical inquiry. Our social realities are being fundamentally transformed by data collection and analysis, but we may be only vaguely aware of how these changes are manifesting themselves and what they might promise in the future.

How do you feel about traffic jam detection based on the handoff rate between cell towers and those cell phones being used in cars on the road? [7] This is a "yesterday" question. This has already been widely applied for a couple of years now. As cars and cell phones are all equipped with GPS technology this is, by now, standard practice.

How do you feel about auto insurance that is priced from a daily read-out of your automobile's black box? This question is a "today" question—chances are that your driving habits and propensity for taking risks are currently being assessed by your insurance provider. Let's put it another way: In what calendar year will car insurance be more expensive for drivers who insist on driving their car themselves rather than letting a robot do it?

How do you feel about urinanalysis every morning with a robot giving you recommendations for the day's diet? Would you want a computer to organize your daily life and decide what's best for you? How do you feel about access to medical and welfare benefits based on a readout of your consumer preferences, spending habits, and grocery shopping list?

Would you buy that Google Fridge that automatically sends out grocery orders of your favorite foods? The so-called *Internet of Things,* which, among other things, connects electric appliances with the Internet, allows respective companies to glean ever more information on our daily habits, movements, and preferences.

How do you feel about automatic assessment of students in educational institutions based on electronic test scores? The idea of standardized assessment is already with us. All that is needed is a last push for more efficiency and the process will be completely automated. Is this a student's dream or a student's nightmare?

What do you think about allocation of law enforcement resources through predictive analytics? Mapping crime statistics and demographics has been with us for a while already. However, given the amount of information now available, these logics can be taken to new levels. And perhaps there is a deeper question: What does that do to crime statistics? Are we producing an effect by measuring it on the basis of pre-existing biases?

How do you feel about the automation of politics itself? If you think this question to be completely absurd, we only have to think about the last Obama election campaign in which algorithms were used to determine how to allocate campaign resources most efficiently. As some Silicon Valley visionaries might argue, public administration would be much more efficient with algorithmic regulation—an idea that Evgeny Morozov has described as "the rise of data and the death of politics"[8].

Underlying all this is a more general question: How do we want to structure the complex interplay of technology and society? How do we create structures that allow for the benefits of technology without these technologies infringing upon our sovereignty and dignity? I would urge us all to take up these issues and to deliberate carefully about what world we want to live in, because a lot of these transformations are already with us, and they aren't likely to slow down on their own.

NOTES

1. The numbers comparing the surface area covered by file cabinets and Gauck's statement are from OpenDataCity at https://apps.opendatacity.de/stasi-vs-nsa/english.html.
2. Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (New York: Eamon Dolan/Mariner Books, 2014), p.94.
3. The basis for this overview is Guardian US Interactive Team, "A Guardian Guide to Your Metadata," theguardian.com, 12 June 2013, http://www.theguardian.com/technology/interactive/2013/jun/12/what-is-metadata-nsa-surveillance.
4. Viktor Mayer-Schönberger, *Delete: The Virtue of Forgetting in the Digital Age* (Princeton, NJ: Princeton University Press, 2009), p.141.
5. Mark Andrejevic, *iSpy: Surveillance and Power in the Interactive Era* (Lawrence: University Press of Kansas, 2007).
6. Mark Andrejevic, *Infoglut: How Too Much Information Is Changing the Way We Think and Know* (New York: Routledge, 2013).
7. This and the next two questions were originally posed by Dan Geer in "Personal Data and Government," 2013 Annual Conference & MIT-KIT Launch, MIT Consortium for Kerberos and Internet Trust, 7 October 2013, http://kit.mit.edu/sites/default/files/documents/Geer_MIT_KIT_2013_Conference.pdf.
8. Evgeny Morozov, "The rise of data and the death of politics," theguardian.com, 19 July 2014, http://www.theguardian.com/technology/2014/jul/20/rise-of-data-death-of-politics-evgeny-morozov-algorithmic-regulation.