Data Double

Announcer 1: You’re listening to –

Announcer 2: So What?

Announcer 1: The podcast that explores why library and information science research matters.

Announcer 2: We interview researchers about their work.

Announcer 1: And they connect the dots between what they do and its importance to your life.

Announcer 2: Okay, let’s get on it.

Michael Ridley: Digital exhaust, behavioural surplus, data shadows, data doubles. Recently, three professors at the Faculty of Information and Media Studies at Western University, Jacquelyn Burkell, Alissa Centivany and Alison Hearn, spoke at the London Public Library about the concept of your data double. The evening began with Professor Burkell introducing the topic of data doubles.

Prof. Burkell: Each one of us has, whether we realize it or not, a data double online, a reflection of who we are, what we do, what we like, what we say, what we think almost. That reflection includes information that we know that we’re providing, say when we register for an account on the New York Times.

It also includes records of activities that we carry out on an everyday basis, like searches that we do online, websites we visit, ads that we click through on. It includes information that’s inferred about us on the basis of what is known because we offered it or because it was recorded, thinks like inferences over things that we like on Facebook can be used to infer things like sexual orientation.

And this information is shared between and among data collectors. Although there are some regulations that cover the sharing of information, it turns out that in many instances organizations can share that information and can integrate that information. So it’s not all just in little separate packets; it becomes a very rich record of both our permanent identity and the activities that we’re doing on a day-to-day basis.

Michael: Here is Professor Alissa Centivany.

Prof. Centivany: Margaret Mead, the anthropologist, had a really nice quote which I think of when I think about data doubles, which is that she said we’re all a parliament of selves. And I think that holds true when we’re talking about data doubles, right?

In our offline worlds we have multiple identities that we might emphasize or foreground depending on particular contexts that we're in, or situations. The same is true online in many ways I think. I think some of the interesting things about data doubles is that we don't know a lot about what inferences are being drawn about us. There's a lot of mystery involved. The algorithms and machine processes that are compiling these packets are oftentimes quite opaque, and so we don't really know what our data doubles look like.

We also don't know much about what implications they have on our lives. I think that we can speculate that they have direct implications in terms of things like education and healthcare and finance, whether we're going to get credit or get a loan. But I think that this data is also used in sort of more subtle ways as well, that have serious implications on our online and offline lives in terms of how that information is used to architect online experiences and affect the choices that we make in subtle and important ways.

Michael: At this point, Professor Alison Hearn added some thoughts about agency and metrics.

Prof. Hearn: It's important to think about where is the space of agency in terms of creating the data double. People will say "Well, I don't mind…" If you make it known to them that people are collecting their data they'll say, "Well, I have nothing to hide so I don't mind." That's the standard line.

The problem with that is that it's not a question of what you're doing individually. It's a question of the aggregation of that data. So, we create the data by our online participation, and we may do it to express ourselves or create various versions of ourselves. But at the end of the day, we have very little control over how our data doubles are actually constituted and why and in what way.

And I'm interested really in questions of scoring and the idea that data doubles often take the form of a score, some kind of numeric assessment, which we don't see but is operating kind of on top of our everyday identities, working for example when we cross a border, to delineate what kind of risk we might pose, or when we get a speeding ticket. There are ways in which these scores pop up and used where we don't even understand, you know, we don't know that that's happening.

So, I think it's like who are the agencies behind the constitution of our data doubles? How are they operating? For what reason? To whose benefit? And I think it's true, there are pluses and minuses, but it's not necessarily… And there's not a lot of clarification, or clarity or transparency around how their constituted. So I think it's really a governance issue and really boils down to that, and it's a pretty serious one I think.

Michael: Professor Burkell again.

Prof. Burkell: As soon as we started putting personal information online there were principles developed to govern the collection and use of information. They don’t work all that well. They focus on consent, but they require that people pay attention to what they’re consenting to, and most of us don’t. But more importantly, they’ve gone by the wayside because the practices, the ways in which information is collected have changed so that in some cases no consent is required.

The EU is I think a long way ahead of us and the U.S. a long way behind us in terms of privacy regulation. The thing that’s on the radar now in the EU that’s most interesting and most problematic is the notion of a right to be forgotten. But even in the EU they don’t quite understand what it means to have a right to be forgotten. And even in the EU the right to be forgotten, as it plays out, is very limited in scope. And it actually doesn’t cover the kind of data that most concerns me, which is this behavioural tracking data of the record of all of my activities, where I’ve gone, what websites I’ve clicked on, what stores I visited. All of that isn't covered.

Michael: Professor Hearn again.

Prof. Hearn: So, as we’re trying to regulatory wise catch up with the technology, the technology has been so, completely reshaped our understandings of sociality, of work, so that younger generations don’t have a choice but to be online. So, even if they might want to opt into the right to be forgotten, they know that it’s militating against their chances to get a job, to be seen, to be, you know, to make money.

So what I was thinking about, about, you know, famous people not having a right to privacy, is that privacy now equals power. If you are in a position of power, you can choose to be invisible. You can choose to be forgotten. It doesn’t matter.

But, if you do not have those things, if you are not a major celebrity or powerful, rich billionaire, you must be visible. You must participate. And so, you know, regulation only goes so far if, in order to be on LinkedIn or be on Twitter, you have to agree to terms of service that basically make you completely transparent.

Michael: Professor Burkell again.

Prof. Burkell: We are talking about big data analytics, and recognize that there is a massive amount of information that’s being produced and collected about every single one of us, and that people are becoming very interested in aggregating this information and in looking for patterns over it. And what data analytics companies and individuals who are involved in data analytics have started to understand is that you can take that public-facing data, the information that you know you tell the world about you, and make inferences over it that will reveal things about you that you never said.

Kosinski, Stillwell and Graepel published this paper in 2013 where they took Facebook profile information and, on the basis of publicly-revealed Facebook profile information, were able to infer very personal and private characteristics, including sexual orientation, political affiliation and religiosity, things that those individuals never told. They never put them on their website. It wasn't just that they grabbed the data; it was that they inferred the data. So now things are knowable about you that you never revealed anywhere.

Michael: Professor Centivany spoke about data doubles in the context of education.

Prof. Centivany: Schools within Canada over the last couple of years have entered into an agreement with Google to offer these tools in classrooms. So the GSuite, if you're not familiar with that, is a set of products designed by Google. These include Gmail, Google Docs, Google Slides, Forms, Google Drive, Calendar, Sheets. The educational package also includes a name lookup function. So basically, these are tools, they're cloud-based tools, they're free, they're high-quality tools that can be used by students to create, to collaborate, to communicate and share content.

There's a lot of positives I think about this move. The G Suite is, well it's free. It's good quality. It does indeed facilitate students' work and productivity and their ability to collaborate with others. Google and the Ontario Ministry of Education have entered into an agreement that, not surprisingly, is confidential. So we don't actually know what the terms are of this agreement.

I first became aware of this actually last year or maybe the year before, when my third grader came home at the start of her school year with an authorization permission form for me to sign. And so, the way that this program works is all students in schools, generally starts around third grade and goes to 12 or 13, have an opt-in permission slip that gets sent home.

I have a copy of it here. It doesn't say much, okay. What it says is that this is a great tool for students. You know, it will help with their education. It will help with teachers being able to, you know, foster this collaborative, productive environment. It's good for digital literacy. It doesn't say what kind of data is going to be collected, at all. All it says is that no personal student information will be used for commercial purposes.

Okay, so what we know is that personally identifiable information is being collected, that Google is not using it for commercial purposes but they're probably using it for a lot of other things. We have no idea what those things are. It also says that Google will not collect or sell any information to third parties and there's no advertising.

So, that's basically all the information that we have. And then parents are given a choice. You can authorize your child's use of the G Suite for educational purposes, or you can say no and there are no alternatives provided, okay? So you either opt in or you're essentially left out. So this is the bargain that's being struck.

Michael: Professor Hearn again.

Prof. Hearn: What's being done with the data that you're generating? It's actually not about you specifically very much at all except as a way to target you to buy more stuff. Really, it's about the aggregation of the data, and the way and the effects of the selling of that aggregated data to certain different parties for different reasons that we should be – and the governance of that data that we should be concerned about.

It's not just what we purposefully put out there. It's how that data is then – there's metadata that's attached to that that is then, information that we're not purposely sharing, like our location or where we go or how long we spend there, who we're chatting with etcetera.

The metadata is then interpreted and analyzed by various algorithms and compared with other users' data for potentially useful correlations, saleable correlations. Like, some of my favourite, there are some great websites like useless correlations like 44% of people who check their horoscope regularly also floss regularly. Or, 50-year-old men who wear Converse shoes are more likely to enjoy roller coasters.

So these are the kinds of weird correlative insights that the parsing of big data can generate, and somebody's going to find them useful. And there's a whole set of industry arrangements, of businesses that have emerged, basically data brokers, in the wake of big data that we don't think about or know very much about.

But in this environment we are, in John Cheney-Lippold's terms, made of data. But it's only when that data is made useful that our algorithmic identities or data doubles take shape and work to govern us, I would argue in really unseen ways. They sort of float on top of our embodied selves and they accompany us wherever we're going, whether it's to the hospital or across a border, wherever it might be. So it's really the question of who's collecting and parsing and selling our data and for what ends that's central to, I think, where we should be looking.

Perhaps the most extreme example of this development in the re-socialization of credit scoring is of course the citizen score in China. And many of you probably have read about of this initiative on the part of the Chinese government, to implement one standard score for every Chinese citizen by the year 2020. And this score is meant to enhance trust, because it’s an emerging capitalist economy and there’s a lot of fraud, a lot of screwing with the system.

They’ve done a lot of prototypes of the project across the country. One of the private prototypes is the Sesame Credit Score, which was developed by Ant Financial which is a subsidiary of Alibaba, one of the biggest platforms on the globe, offers a score from 350 to 950 based on five general areas of your life. So this isn't just about your financial behaviour.

So it’s your credit and payment history; your fulfilment capacity, do you pay on time, stuff that we would expect; user verification details, so who you are, where you work, where you live; user behaviour and preferences, so that means your consumer habits, your tastes, your proclivities, which there, because in China mobile is the thing, they’re tracking you everywhere; and finally your connections, probably most scary of all, who do you know and what are their scores like. So these are your interpersonal relationships.

So they provide incentives to high-scoring users for access to loans and lower interest rates and cheaper products, and then severe disincentives to those with a lower score such as restricted access to retail services and travel. Sesame Credit incorporated the state’s blacklist of over six million people who had defaulted on court fines, so like you forgot to pay a parking ticket for example, within its scoring mechanism.

So, you know, people woke up one day and found that their score, their Sesame Credit Score was like in the dumpster. And there’s no mechanism for people to challenge their scores. And they’ve punished over 1.21 million defaulters with withholding access to travel. So they couldn’t get on a plane. They couldn’t get on a train. Even schooling, even access to schools. So, once it’s assigned you’re stuck with it.

So this kind of social credit scoring, which is happening to us, like let’s not delude ourselves that we’re in any better position actually, produces forms of what we could call reactivity on the part of users. So if you know you’re being assessed according to these various mechanisms, it produces conforming behaviour and generates its own kinds of contradictions, and attempts to gain the system as well.

So effects with the Sesame Credit Score that had been noted are people disconnecting from their low-scoring friends, changing or adapting their online social habits in order to be seen to be more trustworthy, stripping their posted content of anything that could be perceived as negative or anti-state. So, as Rachel Botsman writes, “Sesame Credit rewards users for sharing positive energy online. Nice messages about the government or how well the country’s economy is doing will make your score go up.”

Other unintended forms of reactivity to the Sesame score have arisen as well, including forms of data forgery, rich people hire hackers to hack into their score to make it seem like they go to better schools or they own more houses or whatever it is, to enhance their social profile, and predictably, new reputation consultancy businesses have emerged as well.

So, as I said, it’s easy for us to see that as an episode of the Black Mirror which I believe it was an episode of the Black Mirror. And there’s been a lot of really hand-wringing stuff written about the Chinese citizen score, but I want to warn us from being too judgy in relation to this, because the reality is, as I said, this kind of scoring practices are happening to us all the time in ways that we don’t know, arguably in more opaque ways than the Chinese credit score where at least they’re relatively transparent about the project and the prototypes that they’re using etcetera, etcetera.

Michael: Professor Burkell again.

Prof. Burkell: We’ve talked about a number of ways and places in which what we know is a very rich data double can be used, and a number of ways in which it can change us and the world around us. We’ve only touched the surface of it.

And if we put together all of what we knew, our creative imagination would probably only touch the surface of what corporations and governments are doing right now. And that I’m not being paranoid about. They’re just always one step ahead of us.

Michael: We hope that this episode of So What? on data doubles has helped you understand your data shadow and how it affects your life. My name is Michael Ridley.

Announcer 2: This has been another episode of So What?.

Announcer 1: The podcast about library and information science research and why it matters.

Announcer 2: So What? is created and produced by students at the faculty of information and media studies at Western University in London, Ontario.

Announcer 1: Find us online at sowhat.fims.uwo.ca