Personal and sensitive data

Pablo Diaz – UNIL

FORS

About personal data

Anyone who *processes* **personal data** must comply with *data protection* laws.

It is therefore essential to learn how to recognize personal data (as well as sensitive data, which is a subcategory of personal data)

What is data processing?

Processing refers to **any operation** with data, in particular the collection, storage, use, revision, disclosure, **archiving** or destruction of data.

What is personal data?

“Any information relating to an identified or identifiable natural person” (art. 5 let. a FADP)

Very broad notion: everything that can be related to a specific person is personal data !

The most common: names, addresses, pictures, recordings, etc.

More subtle: original / unique job, original idea, a rare decease, etc.

⚠ Combination of indirect identifiers may be personal data

Examples of personal data

Contact details

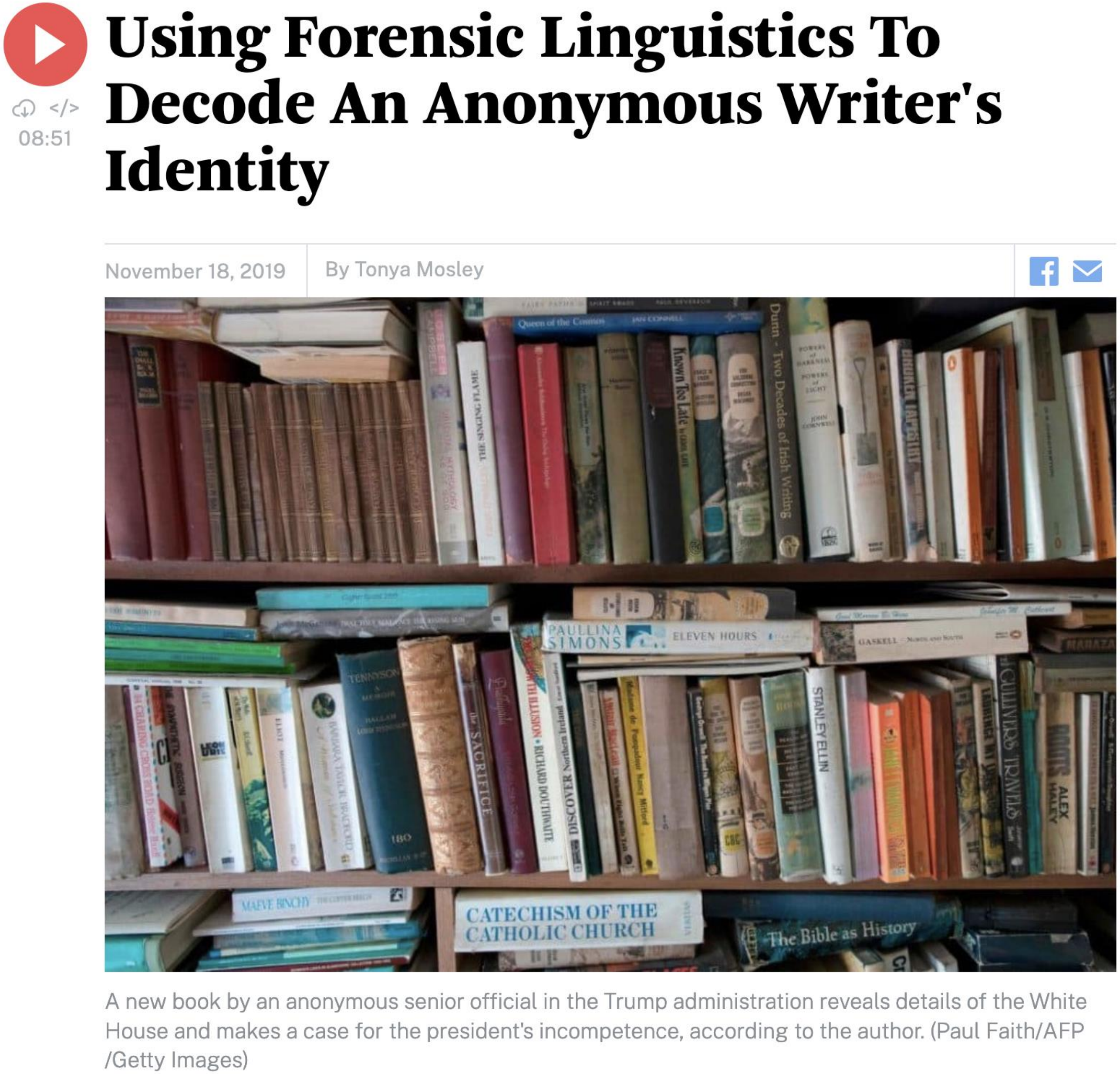
First name: Pablo
Last name: Diaz
Phone number:
123456
Email: pablo@diaz.ch

Combination of variables

- Male
- Chilean
- Ethics officer
- UNIL

Picture

FORS



Using Forensic Linguistics To Decode An Anonymous Writer's Identity

November 18, 2019

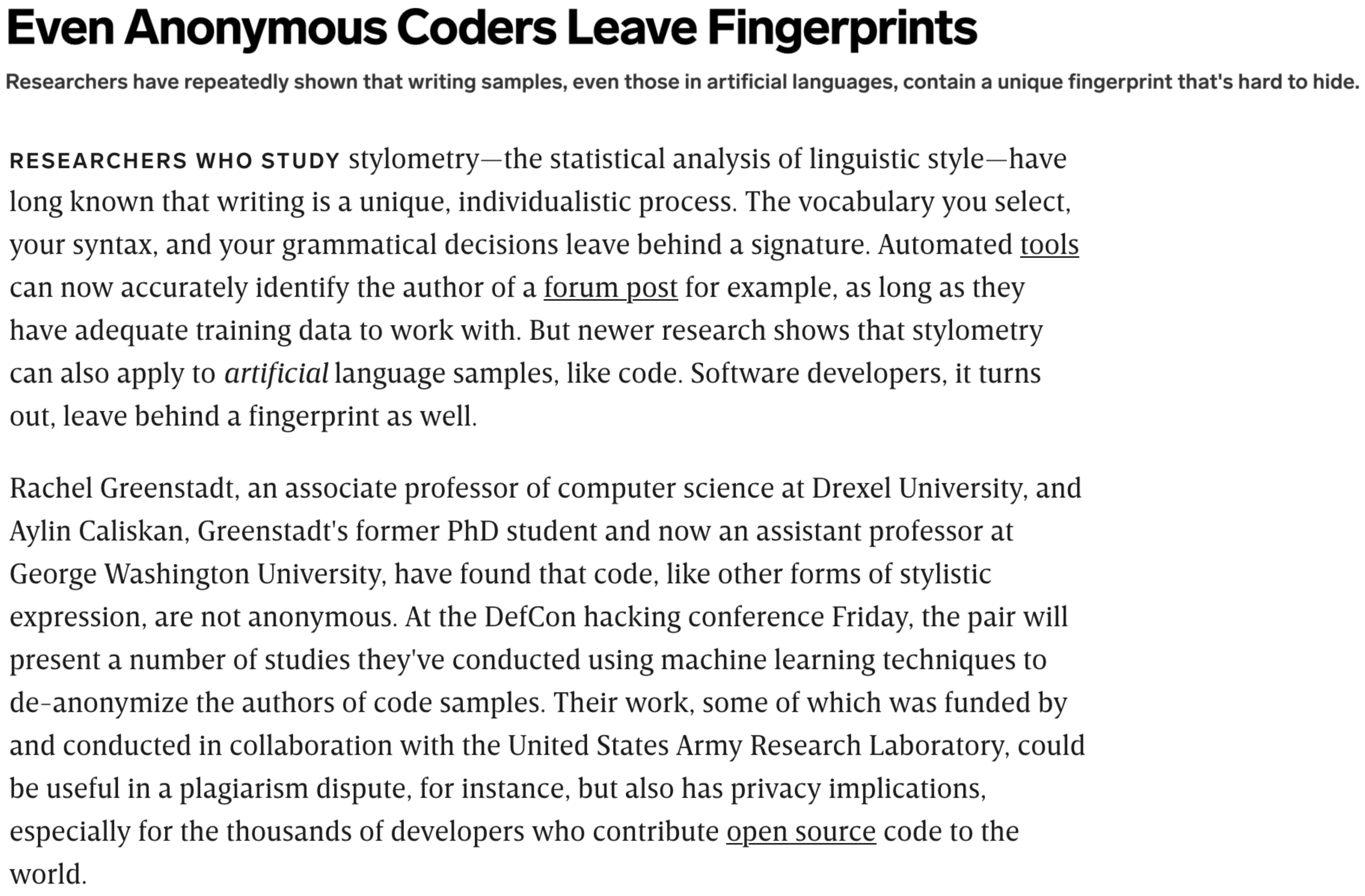
By Tonya Mosley

A new book by an anonymous senior official in the Trump administration reveals details of the White House and makes a case for the president's incompetence, according to the author. (Paul Faith/AFP /Getty Images)

On how forensic linguistics helped reveal the identity of the Unabomber

“[The Unabomber] was so careful not to leave fingerprints or any kind of DNA evidence, but he wrote this 35,000-word manifesto, and it became the strongest evidence against him. And it was the first time that a judge granted a search warrant based solely on linguistic evidence.

“He referred to women as broads. He referred to black people as Negroes. So it obviously put him in a sort of coming of age somewhere before the civil rights movement. He used particular phrases that linguists were able to say probably put him having grown up in Chicago. He also did use some linguistic smokescreens. Like this is a guy who went to Harvard and who had a Ph.D., and he used phrases like, I think there was once or twice where he was trying to portray himself as someone who did not have advanced degrees. And so linguists looking at that said, 'Does he not have an advanced degree or is it a smokescreen?' ”

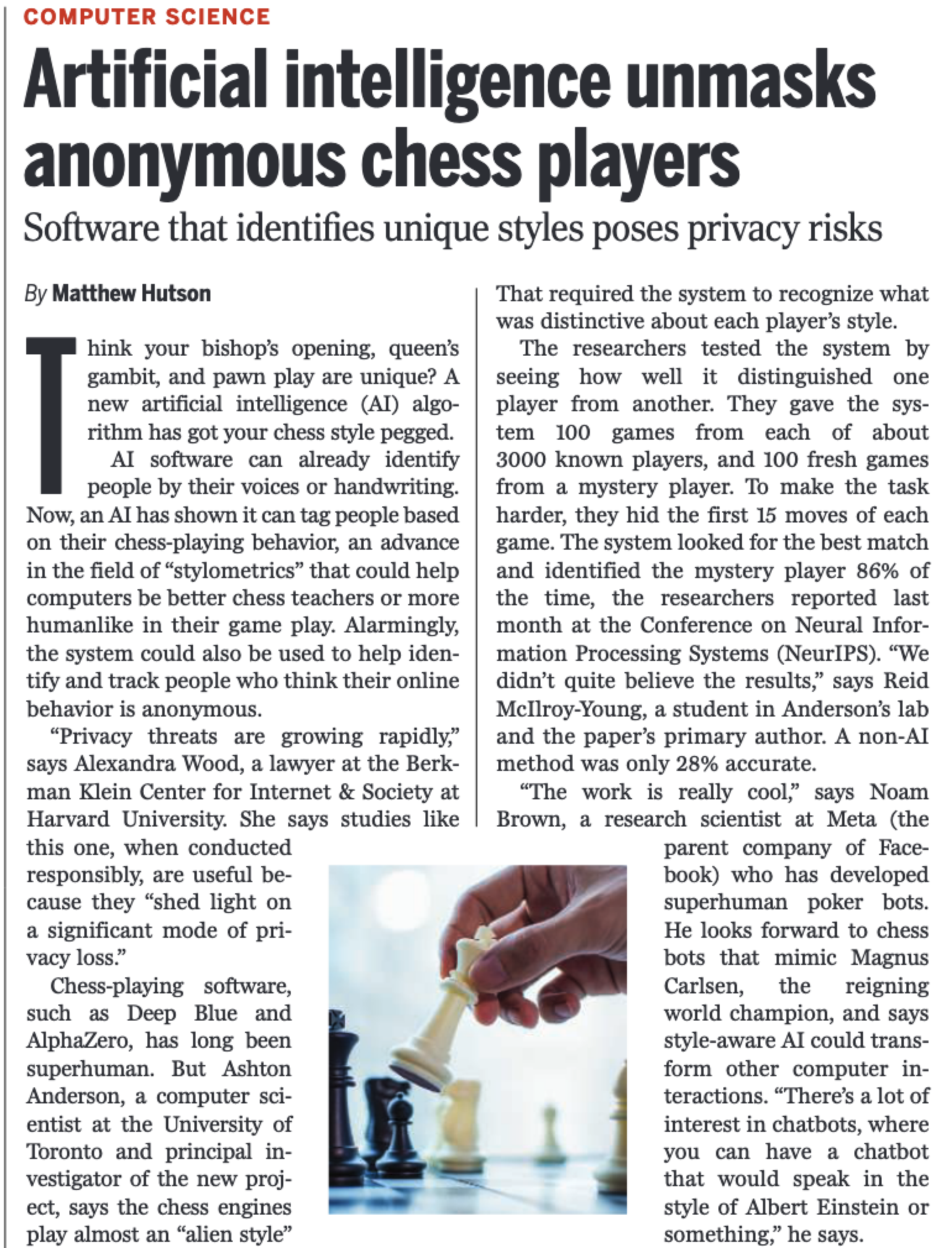


Even Anonymous Coders Leave Fingerprints

Researchers have repeatedly shown that writing samples, even those in artificial languages, contain a unique fingerprint that's hard to hide.

RESEARCHERS WHO STUDY stylometry—the statistical analysis of linguistic style—have long known that writing is a unique, individualistic process. The vocabulary you select, your syntax, and your grammatical decisions leave behind a signature. Automated tools can now accurately identify the author of a forum post for example, as long as they have adequate training data to work with. But newer research shows that stylometry can also apply to *artificial* language samples, like code. Software developers, it turns out, leave behind a fingerprint as well.

Rachel Greenstadt, an associate professor of computer science at Drexel University, and Aylin Caliskan, Greenstadt's former PhD student and now an assistant professor at George Washington University, have found that code, like other forms of stylistic expression, are not anonymous. At the DefCon hacking conference Friday, the pair will present a number of studies they've conducted using machine learning techniques to de-anonymize the authors of code samples. Their work, some of which was funded by and conducted in collaboration with the United States Army Research Laboratory, could be useful in a plagiarism dispute, for instance, but also has privacy implications, especially for the thousands of developers who contribute open source code to the world.



COMPUTER SCIENCE

Artificial intelligence unmasks anonymous chess players

Software that identifies unique styles poses privacy risks

By **Matthew Hutson**

Think your bishop's opening, queen's gambit, and pawn play are unique? A new artificial intelligence (AI) algorithm has got your chess style pegged.

AI software can already identify people by their voices or handwriting. Now, an AI has shown it can tag people based on their chess-playing behavior, an advance in the field of "stylometrics" that could help computers be better chess teachers or more humanlike in their game play. Alarmingly, the system could also be used to help identify and track people who think their online behavior is anonymous.

"Privacy threats are growing rapidly," says Alexandra Wood, a lawyer at the Berkman Klein Center for Internet & Society at Harvard University. She says studies like this one, when conducted responsibly, are useful because they "shed light on a significant mode of privacy loss."

Chess-playing software, such as Deep Blue and AlphaZero, has long been superhuman. But Ashton Anderson, a computer scientist at the University of Toronto and principal investigator of the new project, says the chess engines play almost an "alien style" That required the system to recognize what was distinctive about each player's style.

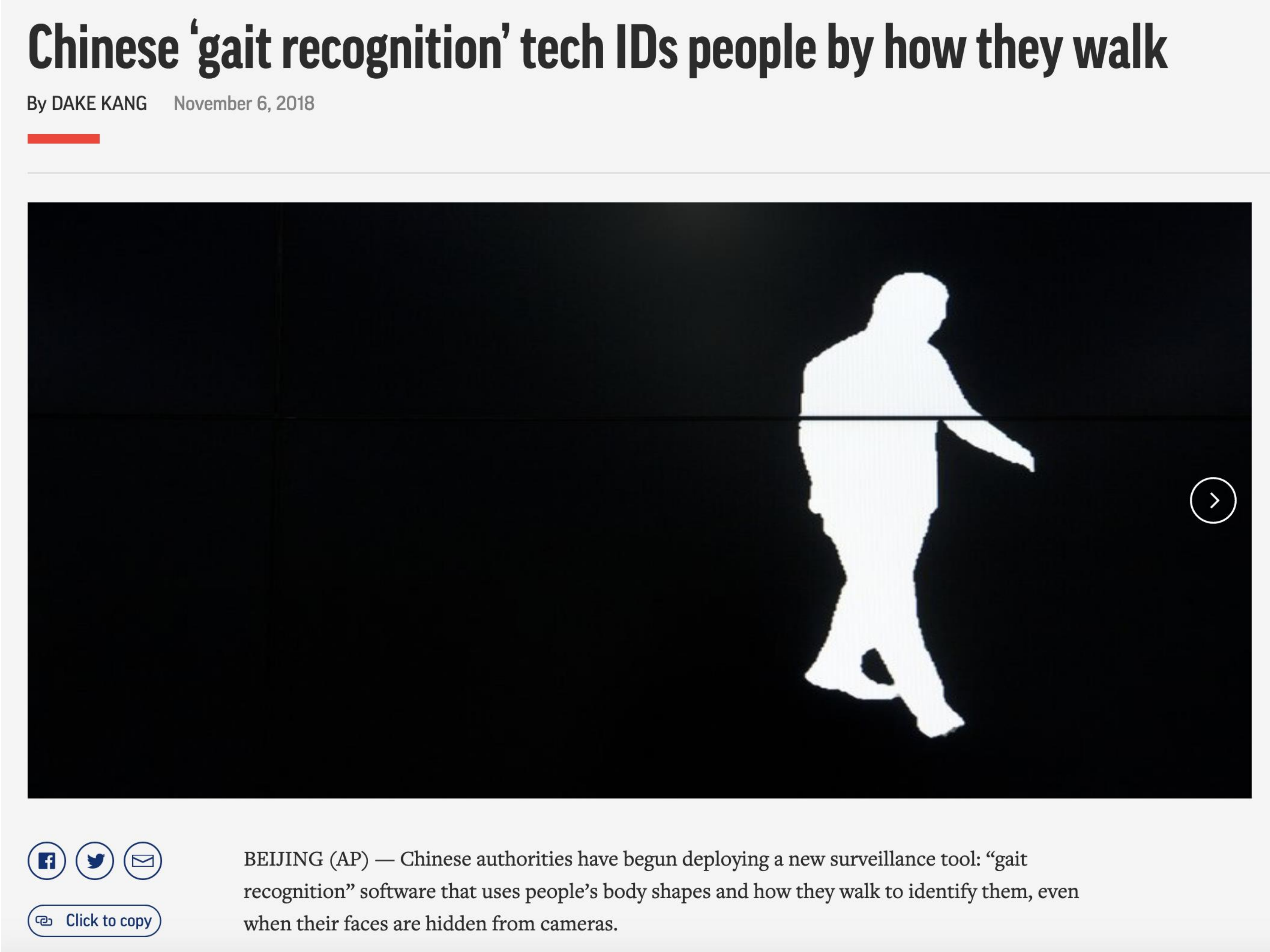
The researchers tested the system by seeing how well it distinguished one player from another. They gave the system 100 games from each of about 3000 known players, and 100 fresh games from a mystery player. To make the task harder, they hid the first 15 moves of each game. The system looked for the best match and identified the mystery player 86% of the time, the researchers reported last month at the Conference on Neural Information Processing Systems (NeurIPS). "We didn't quite believe the results," says Reid McIlroy-Young, a student in Anderson's lab and the paper's primary author. A non-AI method was only 28% accurate.

"The work is really cool," says Noam Brown, a research scientist at Meta (the parent company of Facebook) who has developed superhuman poker bots. He looks forward to chess bots that mimic Magnus Carlsen, the reigning world champion, and says style-aware AI could transform other computer interactions. "There's a lot of interest in chatbots, where you can have a chatbot that would speak in the style of Albert Einstein or something," he says.

The way you dance is unique, and computers can tell it's you

Nearly everyone responds to music with movement, whether through subtle toe-tapping or an all-out boogie. A recent discovery shows that our dance style is almost always the same, regardless of the type of music, and a computer can identify the dancer with astounding accuracy.

Studying how people move to music is a powerful tool for researchers looking to understand how and why music affects us the way it does. Over the last few years, researchers at the Centre for Interdisciplinary Music Research at the University of Jyväskylä in Finland have used motion capture technology—the same kind used in Hollywood—to learn that your dance moves say a lot about you, such as how extroverted or neurotic you are, what mood you happen to be in, and even how much you empathize with other people.



Chinese 'gait recognition' tech IDs people by how they walk

By DAKE KANG November 6, 2018

BEIJING (AP) — Chinese authorities have begun deploying a new surveillance tool: "gait recognition" software that uses people's body shapes and how they walk to identify them, even when their faces are hidden from cameras.

Sensitive personal data

1. Data relating to religious, philosophical, political or trade union-related views or activities,
2. data relating to health, the private sphere or affiliation to a race or ethnicity,
3. genetic data,
4. biometric data that uniquely identifies a natural person,
5. data relating to administrative and criminal proceedings or sanctions,
6. data relating to social assistance measures (art. 5 let. c FADP)

Depending on the **context**, almost all data can be considered sensitive (name, photo, job, etc.)

Examples of sensitive data

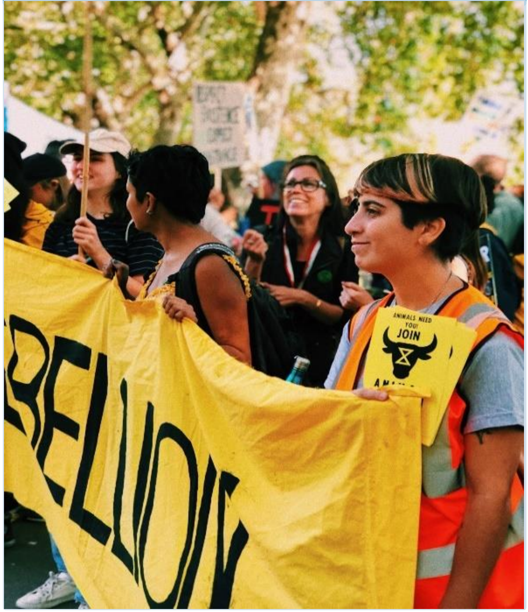
Contact details

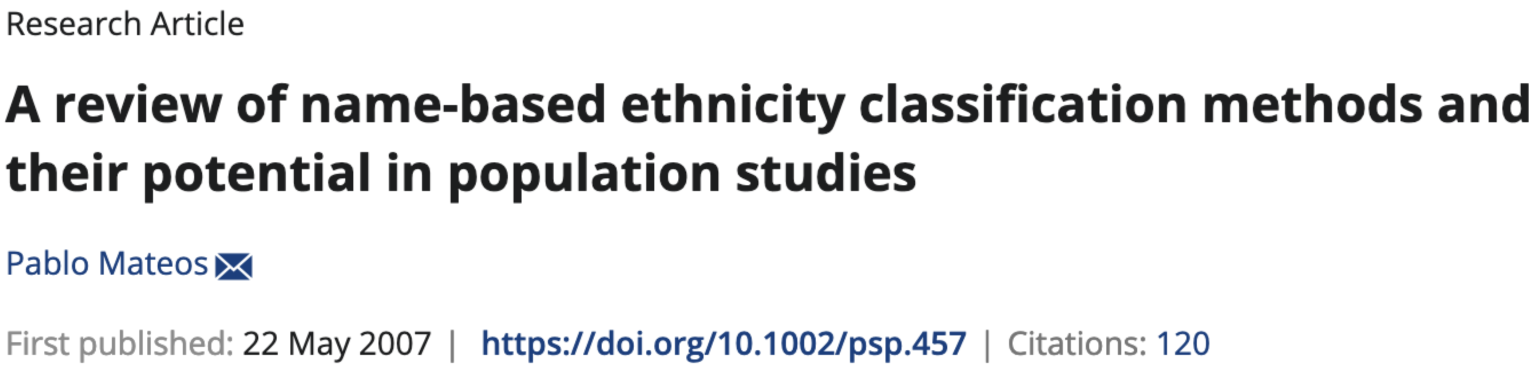
First name: Sebastian
Last name: Calfuqueo

Specific variables

- Gender : Trans
- Job : Trade-unionist
- Religion : Catholic

Picture





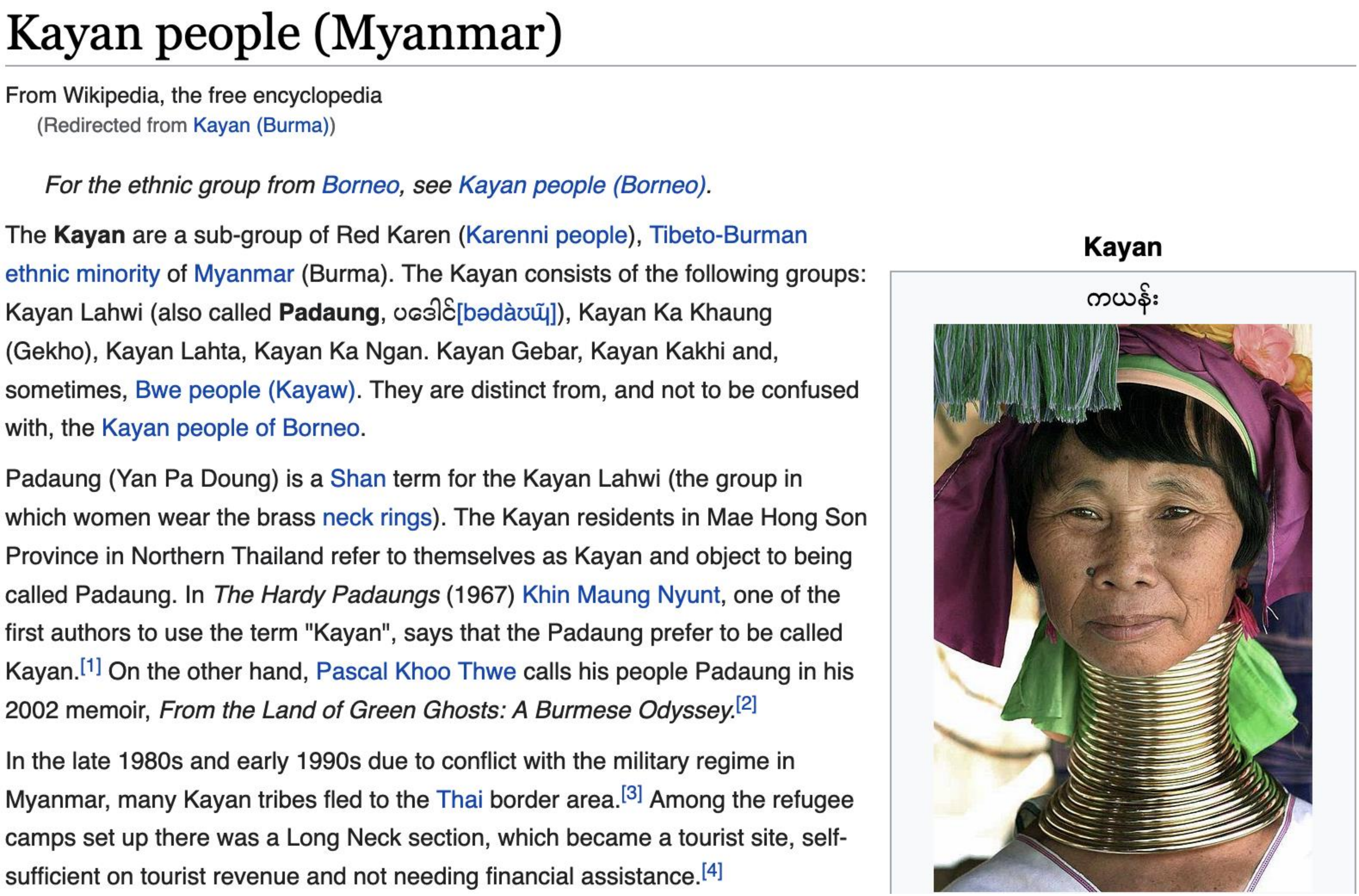
Research Article

A review of name-based ethnicity classification methods and their potential in population studies

Pablo Mateos ✉

First published: 22 May 2007 | https://doi.org/10.1002/psp.457 | Citations: 120

FORS



Kayan people (Myanmar)

From Wikipedia, the free encyclopedia
(Redirected from Kayan (Burma))

For the ethnic group from Borneo, see Kayan people (Borneo).

The **Kayan** are a sub-group of Red Karen (Karenni people), Tibeto-Burman ethnic minority of Myanmar (Burma). The Kayan consists of the following groups: Kayan Lahwi (also called **Padaung**, ပဒေါင်[bədàʊɰ̃]), Kayan Ka Khaung (Gekho), Kayan Lahta, Kayan Ka Ngan. Kayan Gebar, Kayan Kakhi and, sometimes, Bwe people (Kayaw). They are distinct from, and not to be confused with, the Kayan people of Borneo.

Padaung (Yan Pa Doung) is a Shan term for the Kayan Lahwi (the group in which women wear the brass neck rings). The Kayan residents in Mae Hong Son Province in Northern Thailand refer to themselves as Kayan and object to being called Padaung. In *The Hardy Padaungs* (1967) Khin Maung Nyunt, one of the first authors to use the term "Kayan", says that the Padaung prefer to be called Kayan.[1] On the other hand, Pascal Khoo Thwe calls his people Padaung in his 2002 memoir, *From the Land of Green Ghosts: A Burmese Odyssey.*[2]

In the late 1980s and early 1990s due to conflict with the military regime in Myanmar, many Kayan tribes fled to the Thai border area.[3] Among the refugee camps set up there was a Long Neck section, which became a tourist site, self-sufficient on tourist revenue and not needing financial assistance.[4]

Kayan

ကယန်း

⚠ The **voice** is biometric data
(= sensitive FADP)

About anonymity

In the social sciences, it is very difficult to have anonymous data.

It is therefore generally **safer to assume that we are dealing with personal data.**

Questions ?

Assessing the legality of sharing research materials including personal data via a third-party repository

Legal bases

Anyone who processes **personal data** must comply with **data protection** laws !

Data protection laws aim to protect people's **privacy**.

- Privacy is the “right to be left alone” (Warren & Brandeis)
- In the field of data protection, privacy is mainly apprehended through the notion of **informational self-determination.**

Legal bases

In Switzerland, there are data protection laws at two levels:

- Federal (e.g. FADP, HRA, ETH Act etc.)
- Cantonal (LPrD, etc.)

⚠ Universities, universities of applied sciences, etc. are subject to **cantonal law**.

Workshop approach

- In this presentation we will take the most general level possible in order to go beyond cantonal differences.

- The idea is to provide guidelines / best practices (inferred from an in-depth legal analysis) and not a “purely” legal analysis.

⚠ The following considerations do not apply to research falling within the scope of the HRA.

Assessing the legality of sharing

- Have the (personal) data been collected legally?
- Can the (personal) data be retained legally?
- Can the (personal) data be legally disclosed to third parties?

Have the data been collected legally?

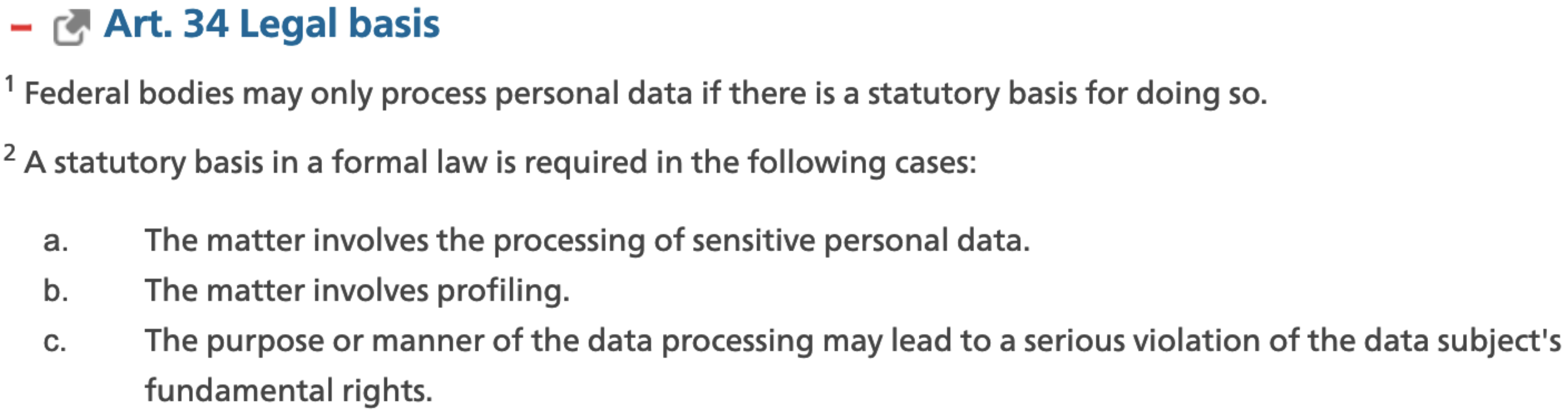
To consider that personal data have been collected lawfully, the following conditions must have been met (among others):

- the data must have been collected on a **legal basis**;
- the persons whose data have been collected must have been **informed** of the collection.

Legal basis for collection

- Public institutions (such as universities, etc.) always need a **legal basis** to collect personal data.
- In the case of sensitive data and profiling, a statutory basis in a **formal law** is required.

E.g. for federal bodies (FADP):



Art. 34 Legal basis

1 Federal bodies may only process personal data if there is a statutory basis for doing so.

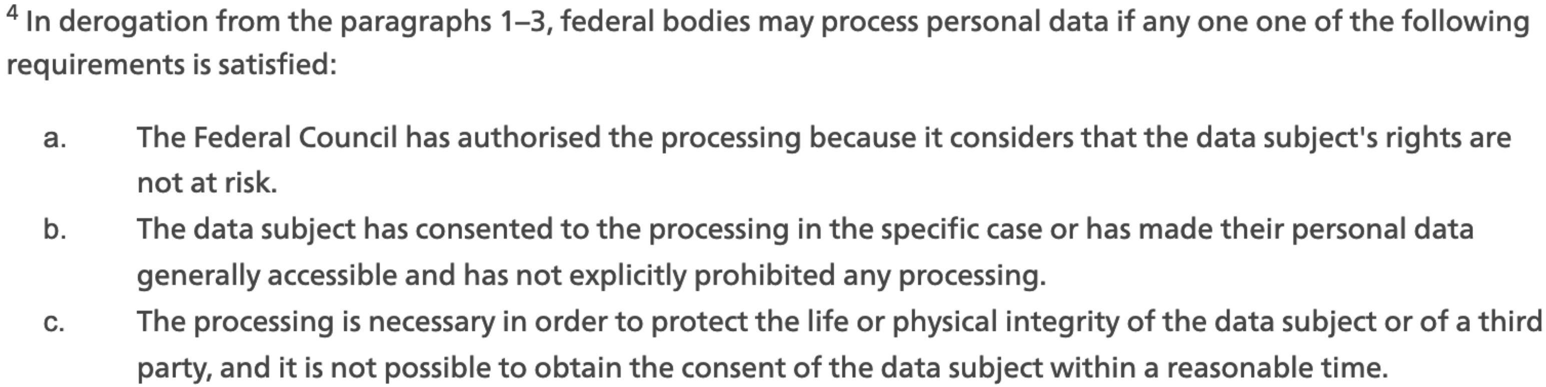
2 A statutory basis in a formal law is required in the following cases:

a. The matter involves the processing of sensitive personal data.
b. The matter involves profiling.
c. The purpose or manner of the data processing may lead to a serious violation of the data subject's fundamental rights.

Legal basis for collection

- At present, very few research institutions are able to rely on a formal law for the collection of sensitive data (with the notable exception of institutions in the ETH domain).
- The only solution open to researchers is to obtain the **explicit consent** of the people whose data they are collecting.

E.g. for federal bodies (FADP):



^{4} In derogation from the paragraphs 1–3, federal bodies may process personal data if any one one of the following requirements is satisfied:

a. The Federal Council has authorised the processing because it considers that the data subject's rights are not at risk.

b. The data subject has consented to the processing in the specific case or has made their personal data generally accessible and has not explicitly prohibited any processing.

c. The processing is necessary in order to protect the life or physical integrity of the data subject or of a third party, and it is not possible to obtain the consent of the data subject within a reasonable time.

Legal basis for collection

If the persons concerned have made their personal data accessible to everyone **without explicitly objecting to collection**, there is no need for a legal basis.

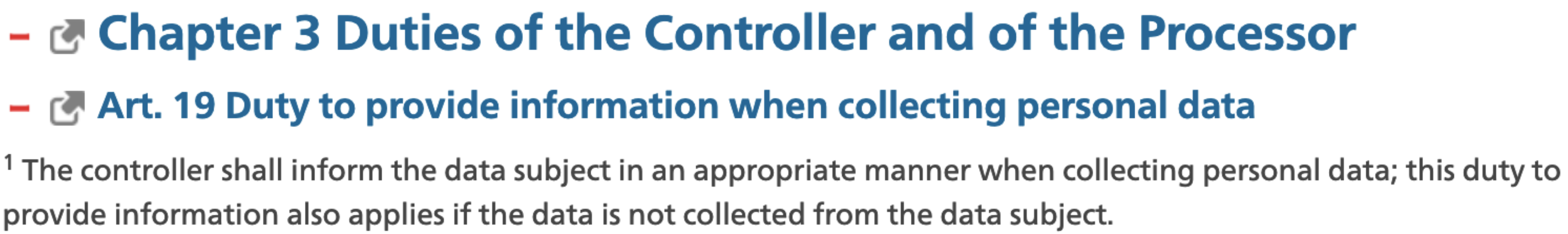
⚠ So-called "public" data legally remain personal data (and are therefore subject to data protection laws).

⚠ Just because data is easily accessible on the Internet does not mean that it can be collected without a legal basis or without consent. It is important to check (at least) the general conditions of use of the site concerned (which may prohibit certain uses).

Duty to provide information

Informing individuals of *any* collection of personal data about them is **mandatory** (even from third-parties),

E.g. for federal bodies (FADP):



– Chapter 3 Duties of the Controller and of the Processor

– Art. 19 Duty to provide information when collecting personal data

1 The controller shall inform the data subject in an appropriate manner when collecting personal data; this duty to provide information also applies if the data is not collected from the data subject.

What information should be provided?

Participants to a research project **must be notified** as a minimum of the following:

- the controller's (PI) **identity** and contact details
- the **purpose** of processing
- the **recipients or the categories of recipients** to which personal data are disclosed (+ *country* if disclosure abroad).

NB: This last obligation is central to open research data (see below).

Legality of data retention

- Personal data cannot be kept for no reason. They must be destroyed as soon as the purpose for which they were collected has been achieved.
- It is the **purpose** of the collection (as announced to the participants) that determines the time limit for the retention of personal data.

⚠ It is important to check that no specific time limit has been given to people (e.g. end of a research project) and that **no promise** of destruction has been made.

Legality of data sharing

There are two *main* ways to make the sharing of research materials including personal data possible:

- Sharing on the basis of *informed consent*
- Sharing on the basis of “*research privilege*”

Sharing on the basis of informed consent

Informed consent is the **best way** to allow the sharing of personal data.

This said, for consent to share personal data to be valid, a number of conditions must be met, including :

- Inform participants correctly about the **purpose** of data sharing (e.g. for research purposes).

⚠ *“General consent” does not exist under the general data protection regime. When providing information, a balance needs to be found between precision and generality. For example, someone taking part in a survey on eating habits may not want their data to be used in new research on religious beliefs. Access must always be controlled.*

Sharing on the basis of informed consent

Another condition for sharing personal data to be legal is:

- Participants must be informed about the categories of recipients (e.g. researchers, repositories, etc.).

⚠ It is important to inform participants about the type of people who will have access to the data. While it is not necessary to be specific, the **categories** of recipients must be **clearly** presented.

⚠ It should be noted that the people downloading the data must be **traceable**.

Sharing on the basis of informed consent

⚠ Repositories are **recipients** (subcontractors). To be allowed to share data through this type of infrastructure (on the basis of informed consent), it is therefore necessary to inform the participants and draw up a **subcontracting agreement**.

Sharing on the basis of research privilege

Without the participant's consent, it is still possible to share personal data under certain conditions (research privilege):

- To have the right to possess the data and not be prohibited from sharing them.
- To share the data for *research purposes* only.
- To require data recipients to publish the results of their analyses in a form that does not allow individuals to be identified.
- To require data recipients to destroy their data at the end of their analyses / processing..

Sharing on the basis of research privilege

That said, even when sharing under the “research privilege”, the **duty to provide information** remains.

From there, the following choices are open to researchers:

- Obtain informed consent to share
- Inform participants that their data will be shared via a repository (under conditions)
- Permanently delete all contact data

Some recommendations…

It is strongly recommended that researchers permanently destroy all direct identifiers (particularly contact data).

While anonymisation is not always possible, de-identifying datasets is a good layer of protection, especially when coupled with access control.

Some recommendations…

It is important to maintain a degree of control over data *access* and *re-use*. To achieve this, it is a wise decision to choose a repository that allows you to :

- control access to data;
- have user contracts that set out the conditions to be met for the re-use of data; and
- Monitor and keep track of data downloads, (re)use, publication, etc.

Thank you!

FORS