It's Love Data Week! Don't stay alone with your Research Data.

2019 TOPICS

The main theme for the 2019 Love Data Week is data in everyday life, that is explored through two topics – open data and data justice. However, there are many aspects of this theme that may echo more strongly with your local community or organization. We encourage you to adapt and modify!

Stay informed. Click and read our take on the three topics proposed for this year's edition.

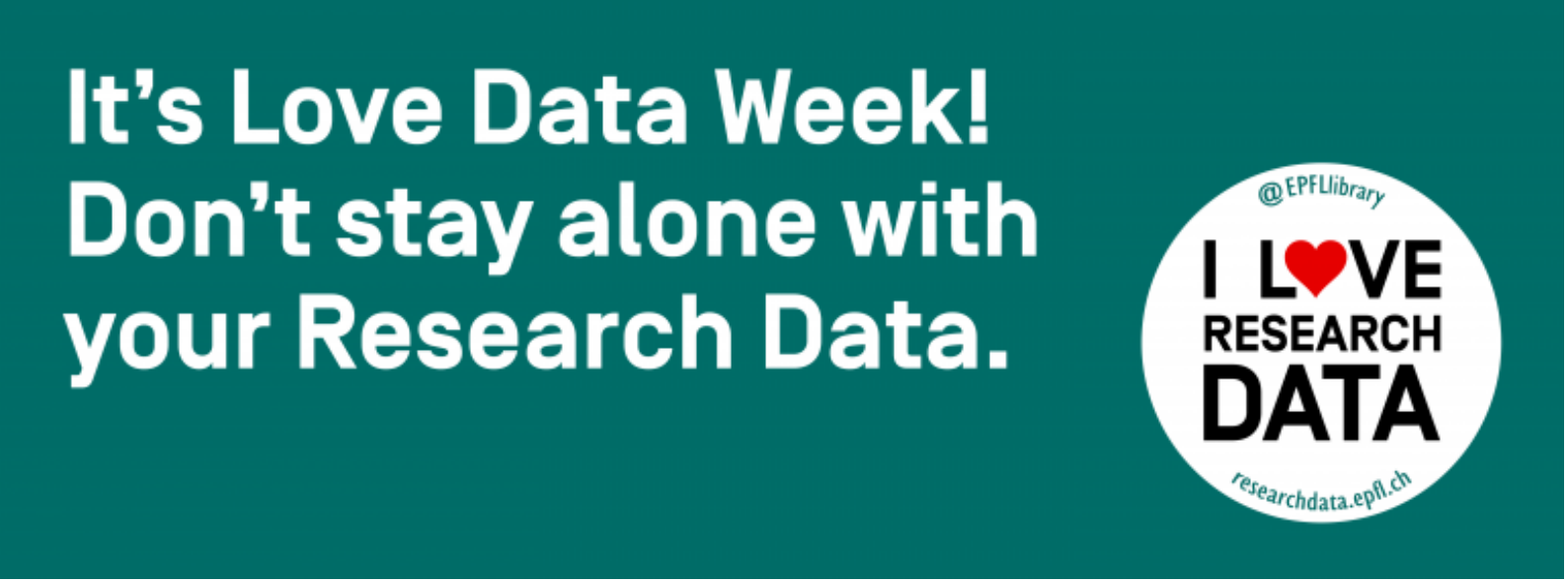
DATA IN EVERYDAY LIFE

The data dilemma

The impact of data is increasingly felt by anyone with an internet connection or who carries a mobile device. In Switzerland, the forecasted[1] rate of internet user penetration for 2019 is of 78%, in line with the European's rate[2] of 76%.
This digital penetration is a major driver for a flourishing digital industry, whose companies recently ascended in the top ranking of market value[3]. The digital industry is also set to revolutionize the manufacturing processings[4], via the so-called Industry 4.0 revolution.
Meanwhile, a visible and yet non-transparent part of the digital industry, able to quantify, aggregate, and sell[5] personal details of our everyday lives, makes its living from personal data[6]. Privacy concerns and many little and big scandals, brought the EU to adopt in 2018 the *General Data Protection Regulation*[7] (GDPR): all the main internet actors have tried to adapt to the new regulation, at least in the EU, with methods and results that surely need some improvements[8].

Top-down & Bottom-up

As technology enables us to easily create, analyze, and share data to improve our daily routine[9], data pervades our daily lives. However, we are still learning how to adopt (and adapt to) new policies and regulations designed to protect our privacy, as citizens as well as professionals. In Switzerland, a good

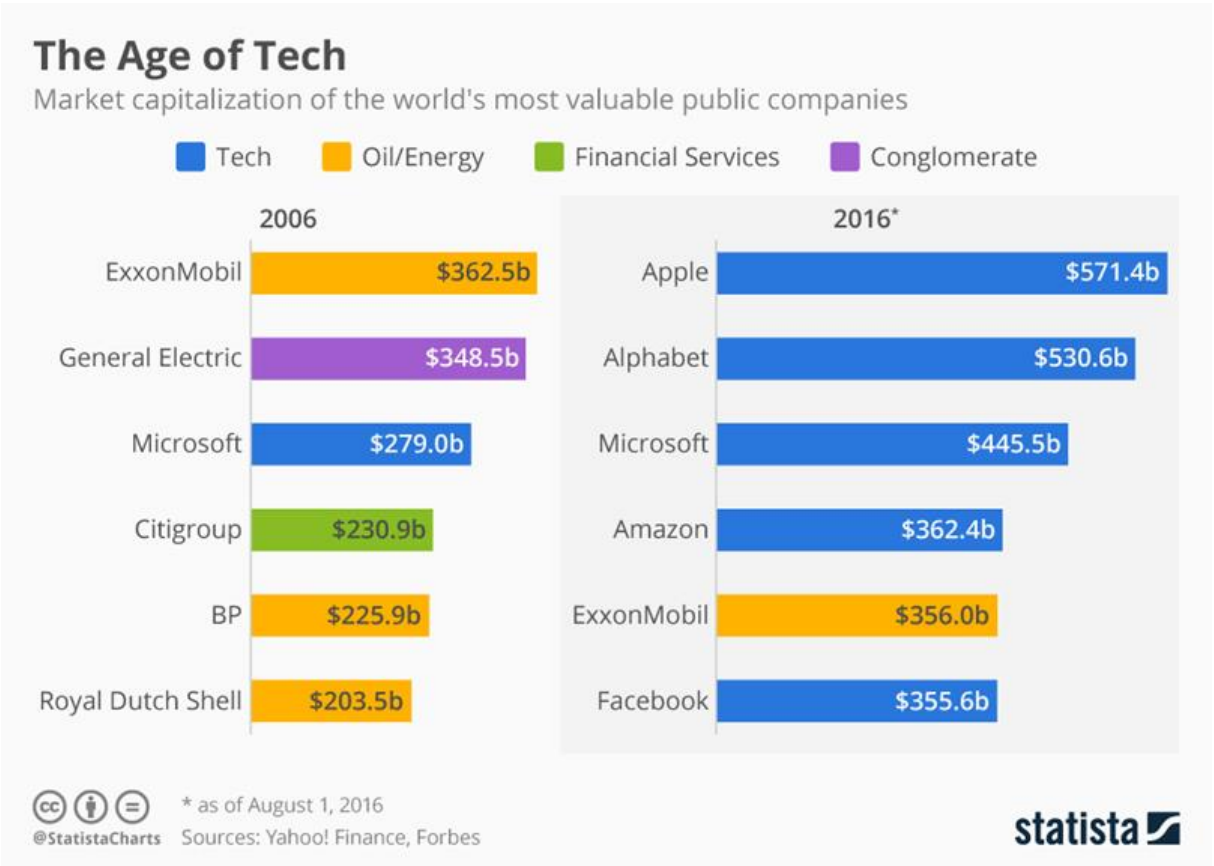


example of data privacy advocates is a nonprofit organization promoting digital rights, *personaldata.io*[10], which participates in the Facebook testimonies[11] following recent scandals. Of course, the security of data is another topic related to our daily use of phones, TVs, computers, connected refrigerators, routers, cloud storage accounts, social apps, etc. For instance, at the beginning of 2019 a huge data breach of 772 million emails addresses has already been exposed[12]. Clearly, the data protection in everyday life is not just about regulations, but a necessary cultural shift towards practices (password manager, 2-step authentication, biometrics authentication, etc.), along with large scale education on what one can or cannot share online.

Beyond the academia

It's impossible to completely separate new, arising business models in the digital industry[13], from topics like Open Data or Data Justice. Academic institutions are and should stay at the forefront of this change, by making research robust, accessible and reusable.
For instance, the EPFL proposed different events[14], initiatives[15] and trainings[16], with a focus on the media sector (especially fake news, data journalism, content personalization, AI, IoT, etc.) and Open Science. Also, the EPFL Research Data[17] team accompany researchers into managing their data life cycle[18], the Technology Transfer Office[19] advise thoroughly on the data licensing questions, and the Research Office assists with legal advice via the Research Ethics Assessment[20].
Funding initiatives, better policies, or even powerful technical tools are useless without ideas dissemination and public awareness. That's why universities like the EPFL participate to the *Love Data Week* initiative[21], with the ultimate goal of making the discussion percolate into many domains, as Open government data, Open software companies, publishing industry, etc.

It's Love Data Week! Don't stay alone with your Research Data.

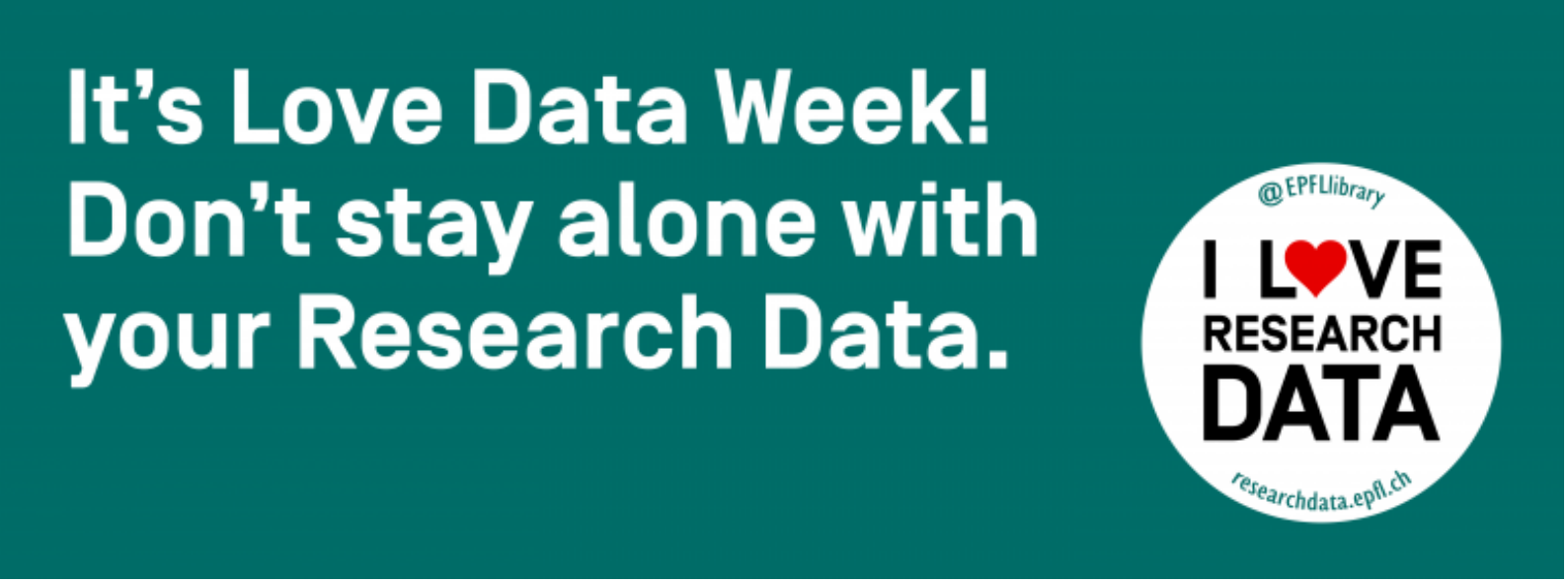


The Age of Tech[22], by Felix Richter (Aug 2, 2016). "Information is the oil of the 21st century, and analytics is the combustion engine." – Peter Sondergaard, Gartner Research

OPEN DATA

Open what?

While open-source software (OSS) seems bound to take over the cloud[1] and the global market[2], the topic of Open Data is slowly but certainly reaching the same level of awareness[3]. Taiwan, France,



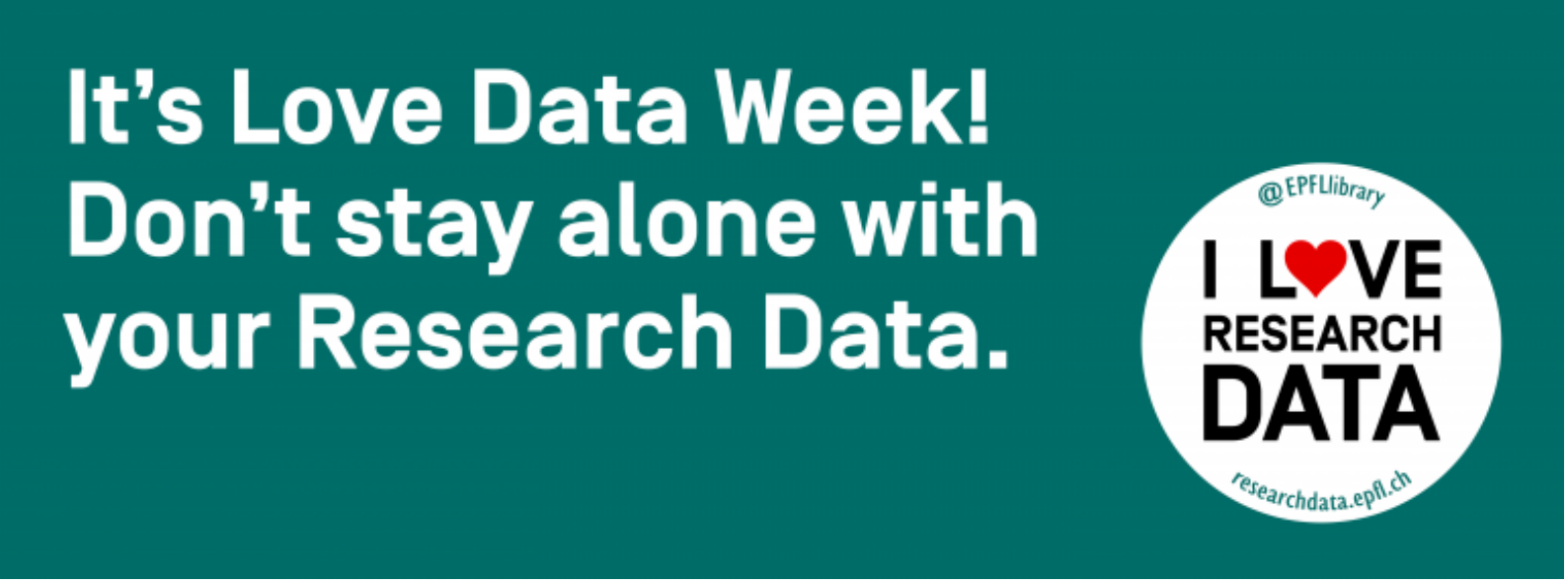
Italy, Spain and Colombia lead the way in interest for this topic (as ranked by Google search queries). As its adoption grows, a variety of academic and commercial use cases are monitored[4],[5]and mapped[6],[7].
But Open Data is still failing to become similarly discussed in the literature[8]. Maybe it depends on a fundamental confusion that still exists: its definition. The European Commission defines[9] Open data as ***"data that anyone can access, use and share"***, while the *Open Knowledge International*organization distinguishes the gratuitousness from the openness[10], as ***"Open data and content can be freely used, modified, and shared by anyone for any purpose"***.

Why researchers should care

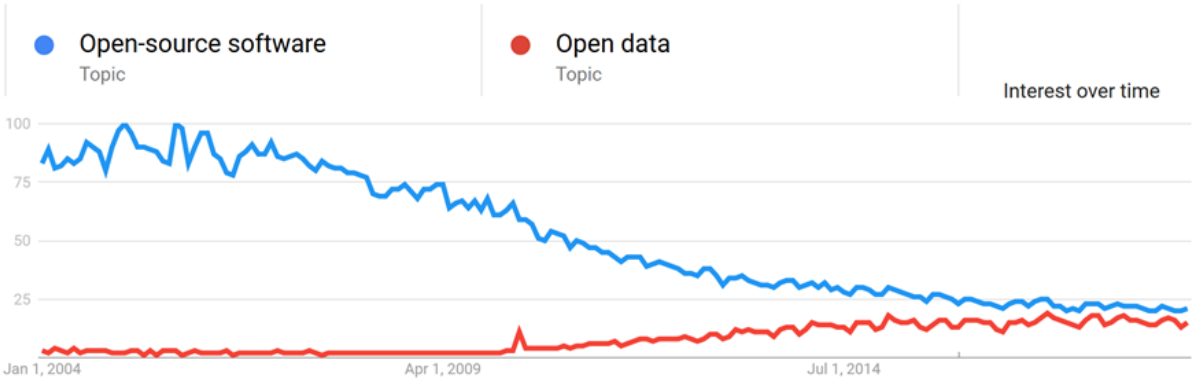
The European Commission clearly wants to leverage on Open Data envisioning a so-called Digital Single Market[11], and wants it to become common sense in the researchers' communities. But the first time many researchers confront with this topic is when writing the Data Management Plan[12](DMP) for funding requests. General concepts (eg. transparency, social or scientific values) are the usually listed as main drivers[13], but the long-term impact of Open Data is still unknown. Only recently it is making the jump from the policymakers[14] to the higher education institutions and, in most cases, the right policies are not yet in place, as the *Open Data Barometer* reports[15].
Real-world examples exist for Open Data use[16]: the *Human Genome Project* is probably the best-known, good example in which an openly accessible data repository is being used successfully[17]. Morever, the Open Data becomes fundamental for *Reproducibility Projects* as in Cancer Biology[18] or Psychology[19]. Open Data also promises to reduce research costs, with a 15% savings for projects using it to make research robust, accessible and reusable[20].
In addition, Open Data are easier to harvest and can make pulished results more easily discovered, with 20-50% more citation for articles linked to associated data[21]. This is technically possible by relying on data repositories discoverable via services like *re3data.org*[22], or data search engines like *Discover Mendeley Data*[23], *Datahub*[24], the European Union *Open Data Portal*[25], *Google Dataset Search*[26], etc.

As open as possible, as closed as necessary

Even if the Open Access publication of articles has sharply augmented in recent years[27], Open Data is sometimes subject to academic controversies[28] and over one-third of researchers (36%) rarely or never make their datasets openly available[29].
Nevertheless, closed (or private) data have many, well-known issues. For instance, the use of closed



data typically requires agreements on lengthy and complex terms and conditions, intricate access rights, restrictions about the storage, complex firewall technicalities, etc[30]. Finally, proprietary data services are typically also much more expensive.
The many problems with closed data are reasons for the governments and research funders to enforce more and more Open Data policies, as the European *Research Data Pilot*[31] or the Swiss *Open Research Data policy*[32], two main examples known by the EPFL scientific community.

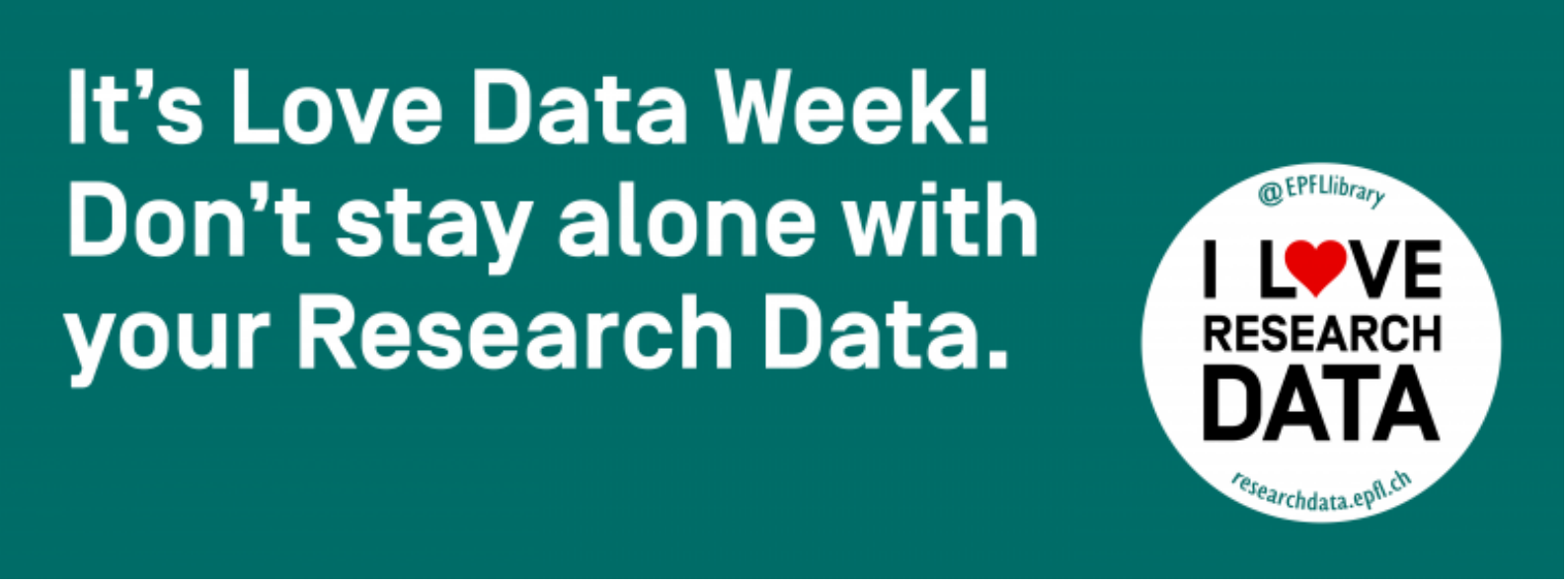


Interest over time[3]. Numbers represent search interest relative to the highest point on the chart for the given region and time. 100 is the peak popularity for the term, 0 means there was not enough interest.

DATA JUSTICE

The data dilemma

Take a minute and think about all the data about you, which you have online.
What kind of data is it? Do you feel it is an accurate representation of yourself, or does it represent only one part of you? Is this the part of you that you want to disclose online and share with the world? Or with the digital industry? Are you the owner of the data about yourself, and do you know how to claim its ownership?
In the context of the *Love Data Week*[1], we reflect on these questions by referring to the concept



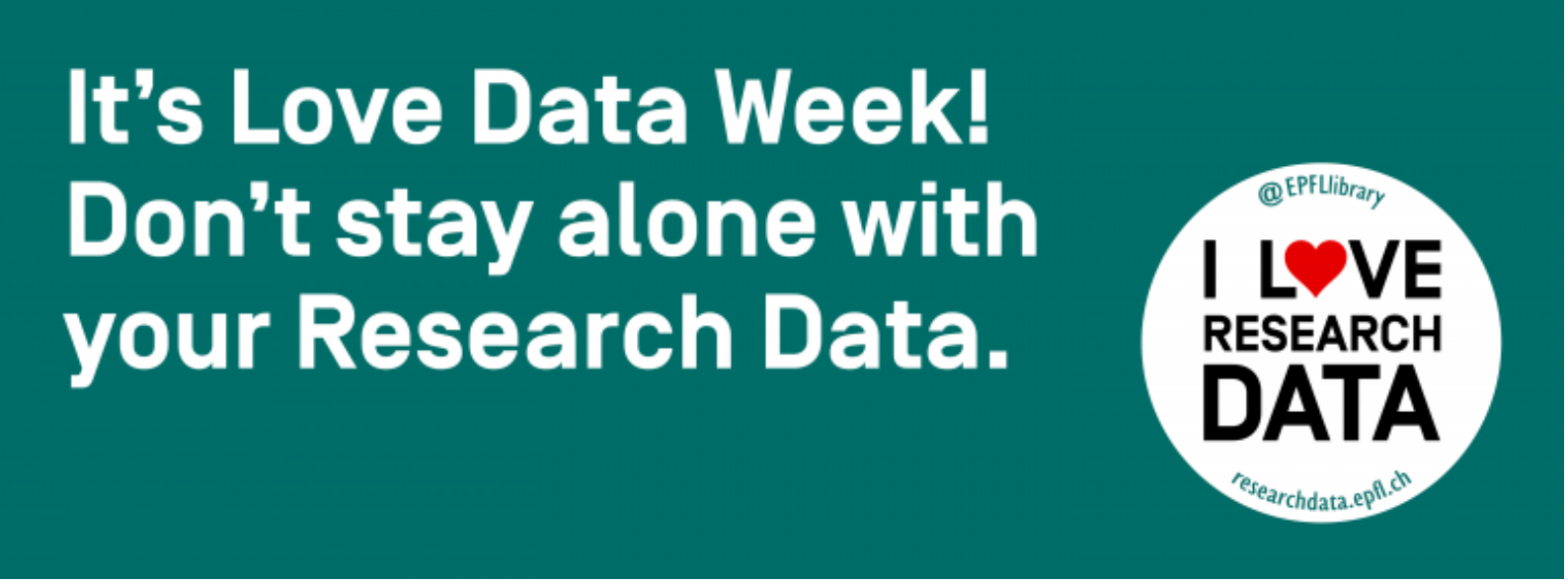
of *Data Justice*, ie the ***"datafication in relation to social justice"***[2], to think about how people are made visible or invisible, threatened or empowered, because of their digital lives[3].
At the EPFL Library, for instance, the exhibition *Data Detox* shed some light on this topic (content freely downloadable[4]). It clearly explains how and why it is essential to regain control over one's personal data (digital identity check, cleanup of accounts, tracking cookies, etc.). Other approaches other than detox exist, and the industry is waking up. For instance, initiatives range from the Facebook VPN for kids[5], or the deep-learning e-assistants like *Oyoty*[6] of the EPFL Innovation Park, or a paper on *Equality of Opportunity in Supervised Learning*[7] in collaboration with Google. Another approach relies on fooling online surveillance: to shift the balance of power between the trackers and the tracked, browser plugins like *TrackMeNot*[8] or *AdNauseam*[9] employ different obfuscation strategies.

The decentralization problem

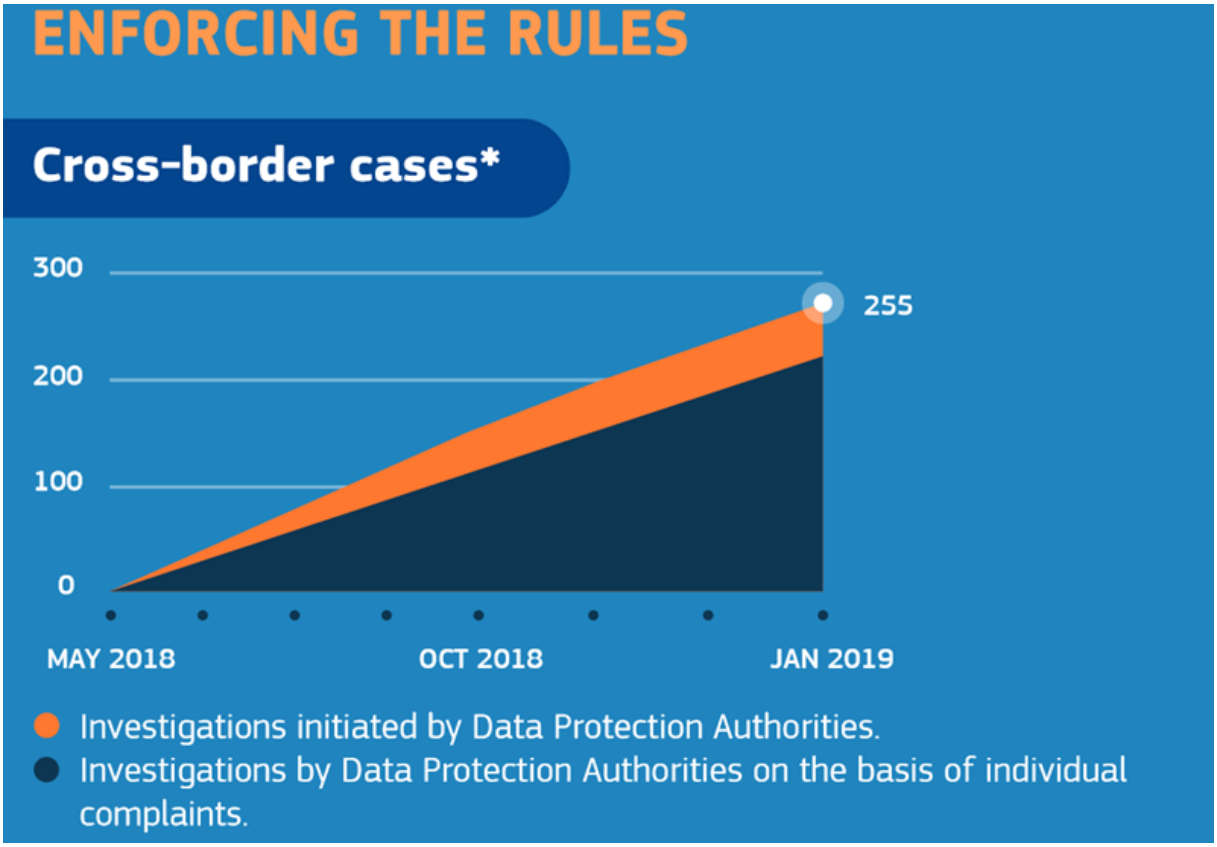
Many projects exist that harness big data or even *AI for good*[10],[11], like *crowdAI*[12], or help creating civic campaigns, like the *Fight for the Future*[13]. But why do they even exist?
Many risks concerning machine learning are recently being highlighted: while they drive the sales of these tech giants, the inherent biases of their algorithms (eg biased training datasets[14], extremization of suggestions, news rankings, echo-chamber, etc.) are not yet correctly addressed[15],[16].
Even neutral, open and community-driven initiatives like Wikipedia struggle with data justice[17],[18],[19],[20],[21], showing a polarization might seem contextual, but it is in fact a more general phenomenon. The well-studied rich-get-richer phenomenon persists and, paradoxically, decentralized structures amplify its scale[22].
The best example is the *World Wide Web* itself: deviating from its foundational scope of making academic information freely available to anyone[23], large digital enterprises monopolize the data circulating on it, especially our personal data (eg. Google, Facebook or Amazon in the West[24],[25], or Alibaba, Baidu or Tencent in the East[26]).

FAIR data, fair data

In the context of sharing and publishing data, it is import to keep in mind, who will eventually have access to data. But, how to use of the published data as fair and neutral as possible? Context (scientific research, private information, etc.) and documentation (metadata, licensing, etc.) are key, but a critical reflection to share is necessary, for each dataset. That is why knowledgeable policymakers, as well as researchers and scholars, are increasingly pushing for the application of the so-called *FAIR principles* of datasets, which is to manage data in such a way to make them be



Findable, Accessible, Interoperable, and Reusable[27].
Open Data alone will not solve the issues ultimately related to the way we, humans, want to use data, but it contributes to make the processes more transparent and just at the social scale.



GDPR in numbers[28]. Usually, national data protection authorities lead the investigations, while the other concerned authorities support them. If in disagreement, the European Data Protection Board will arbitrate. *Source: The European Data Protection Board.

It's Love Data Week! Don't stay alone with your Research Data.

2019 TOOLS

#11 RDM fast guides

On the occasion of Love Data Week 2019, EPFL Library launched new practical tools: the RDM fast guides.

10 FIRST RDM FAST GUIDES

NEW DMP FAST GUIDE

Any other questions?

Contact researchdata@epfl.ch

#lovedata19 @EPFLlibrary