SOLVING ARTIFICIAL INTELLIGENCE'S PRIVACY PROBLEM

Yves-Alexandre de Montjoye[a], Ali Farzanehfar[b], Julien Hendrickx[b], Luc Rocher[ab]
[a] Imperial College London, Data Science Institute and Dept. of Computing
[b] Université catholique de Louvain, ICTEAM Institute

We are unlikely to see any 'general AI'—machines that could learn the way we do and successfully perform a large range of task—anytime soon

Yves-Alexandre de Montjoye is a lecturer (eq. Assistant Professor) at Imperial College London, where he heads the Computational Privacy Group. His research aims at understanding how the unicity of human behavior impacts the privacy of individuals – through re-identification or inference – in rich high-dimensional datasets such as mobile phone, credit cards, or browsing data. Yves-Alexandre was recently named an Innovator under 35 for Belgium (TR35). His research has been published in Science and Nature SRep. and covered by the BBC, CNN, New York Times, Wall Street Journal, Harvard Business Review, Le Monde, Die Spiegel, Die Zeit, El Pais as well as in his TEDx talks. His work on the shortcomings of anonymization has appeared in reports of the World Economic Forum, United Nations, OECD, FTC, and the European Commission. Yves-Alexandre worked for the Boston Consulting Group and acted as an expert for both the Bill and Melinda Gates Foundation and the United Nations. He is a member of the WEF network on AI, IoT and the Future of Trust; the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems; and the OECD Advisory Group on Health Data Governance.

KEYWORDS

- PRIVACY
- DATA ANONYMIZATION
- PSEUDONYMIZATION
- DE-IDENTIFICATION
- LARGE DATA-SETS
- IDENTITY
- K-ANONYMITY
- UNICITY

Artificial Intelligence (AI) has potential to fundamentally change the way we work, live, and interact. There is however no general AI out there and the accuracy of current machine learning models largely depend on the data on which they have been trained on. For the coming decades, the development of AI will depend on access to ever larger and richer medical and behavioral datasets. We now have strong evidence that the tool we have used historically to find a balance between using the data in aggregate and protecting people's privacy, de-identification, does not scale to big data datasets. The development and deployment of modern privacy-enhancing technologies (PET), allowing data controllers to make data available in a safe and transparent way, will be key to unlocking the great potential of AI.

INTRODUCTION

A world we could have only envisioned a few years ago is becoming a reality. Cars are learning how to drive themselves and are expected to heavily reduce traffic accidents and transform our cities[1]. Machine learning algorithms have started to reshape medical care and research. Physicians are already using them to identify high-impact molecules for drug development[2] and to accelerate skin cancer diagnosis, reaching an accuracy on-par with dermatologists in the lab[3]. A recent report by McKinsey found that 45 percent of all work activities could soon be automated using artificial intelligence (AI)[4]. AI is changing our economy and will have a radical impact on how we work, live, and interact.

Developing solutions allowing AI algorithms to learn from large-scale, often sensitive datasets, while preserving people's privacy is one of the main challenges we are facing today.

[1] https://www.wired.com/2016/10/heres-self-driving-cars-will-transform-city/

[2] https://www.technologyreview.com/s/604305/an-ai-driven-genomics-company-is-turning-to-drugs/

[3] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542* (7639); 115-118.

[4] McKinsey Global Institute (2016), The age of analytics: Competing in a data-driven world, *McKinsey.*

However, despite what the popular press would have us believe, AI bears very little resemblance to human intelligence (or Skynet for that matter). This is unlikely to change anytime soon. Instead, experts in its most popular branch, machine learning, have spent decades training a large ecosystem of advanced statistical models to **learn from data**. These are crafted for specific tasks such as inferring human emotions from text messages[5]; e.g. if a certain combination of words express a positive, negative or, neutral tone; or detecting and classifying cancerous lesions in pictures the way a dermatologist would. We are unlikely to see any 'general AI' — machines that could learn the way we do and successfully perform a large range of task — anytime soon[6]. Access to rich and large-scale datasets will thus be crucial to the development of AI in the coming decades.

This is particularly visible when considering the latest "advance" in AI: Deep Learning. Techniques very similar to Deep Learning (i.e. Deep Neural Networks), have been around for a long time. Neural Networks date back to the 1950s, and many of the key algorithmic breakthroughs occurred in the 1980s and 1990s. While the increase in computing power[7], in particular the advent of GPUs, has contributed to the recent success of deep learning, most of the increase in accuracy is arguably due to the availability of large-scale datasets[8]. As in Peter Norvig's seminal article in 2009[9], one can notice the unreasonable effectiveness of data: corpora of millions of speech records, hi-res images, and human metadata.

Other examples include the use of large-scale Facebook data to build "psychometric profiles" of 220M American citizens by Cambridge Analytica[10]. Their work in identifying an individual's gender, sexual orientation, political beliefs, and personality traits has been credited to have influenced the 2017 US presidential elections[11]. However, the research that underpins part of their work[12] as well as a lot of the analysis that has been made public[13] is fairly simple technically. Here again good accuracy e.g. on personality traits could be achieved with a lot of data and a simple linear regression.

While fueling fantastic progress in AI, this data and its collection and use by AI algorithms also raises privacy concerns that need to be addressed. The vast majority of this data, such as Facebook Likes, is personal. Produced by individuals going through their daily lives: making calls, visiting the doctor, using the GPS on their phone or car, etc. it contains detailed and often sensitive information about people's behavior, medical conditions, travel habits, and lifestyles and can be used to infer further information.

"DEVELOPING SOLUTIONS ALLOWING AI ALGORITHMS TO LEARN FROM LARGE-SCALE, OFTEN SENSITIVE DATASETS, WHILE PRESERVING PEOPLE'S PRIVACY IS ONE OF THE MAIN CHALLENGES WE ARE FACING TODAY."

AI has immense potential for good but the continuous access to always larger and richer datasets it requires will only be sustainable if this can be done while preserving people's privacy. Developing solutions allowing AI algorithms to learn from large-scale, often sensitive datasets, while preserving people's privacy is one of the main challenges we are facing today.

Historically, the balance between using the data and preserving people's privacy has relied, both practically and legally, on the concept of data anonymization. Data anonymization is achieved through a series of techniques used to disassociate an individual's record from their identity in a particular dataset. If the data cannot be associated with the individual to whom it relates, it cannot harm that person.

In practice, datasets are rendered anonymous through a combination of pseudonymization and anonymization (also called de-identification). The former, pseudonymization, is the process of replacing clear identifiers, such as names or account numbers, by pseudonyms. This is only the first line of defence as pseudonymization alone has been shown to not be sufficient. In the late 1990s, the Massachusetts Group Insurance Commission released "anonymized" data containing every hospital visit made by state employees. The then governor of Massachusetts, William Weld, assured that GIC had protected patient privacy by deleting identifiers. By using the public electoral rolls of the city of Cambridge, MIT student Latanya Sweeney was able to re-identify (linking data back to a person) the medical records of the governor using his date of birth, sex, and postcode and sent his medical records to his office[14].

The second line of defence, de-identification, was then developed to prevent re-identification, allowing once again for data to be used while preserving people's privacy. The first de-

[5] Liu, B., 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), pp.1-167.

[6] Etzioni, O. (2016), No, the Experts Don't Think Superintelligent AI is a Threat to Humanity, *MIT Technology Review.*

[7] Roger Parloff (2016), Why Deep Learning is Suddenly Changing Your Life, *Fortune*, http://fortune.com/ai-artificial-intelligence-deep-machine-learning.

[8] Sun, C., Shrivastava, A., Singh, S. and Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era. *arXiv preprint arXiv:1707.02968.*

[9] Halevy, A., Norvig, P. and Pereira, F., 2009. The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), pp.8-12.

[10] Green, J. and Issenberg, S. (2017), Trump's Data Team Saw a Different America—and They Were Right, *Bloomberg*, bloom.bg/2eEWfeO.

[11] Thompson-Fields, D. (2017), Did artificial intelligence influence Brexit and Trump win?, *Access AI*, http://access-ai.com/news/21/artificial-intelligence-influence-brexit-trump-win.

[12] Kosinski, M., Stillwell, D. and Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), pp.5802-5805.

[13] https://medium.com/@d1gi/cambridge-analytica-the-geotargeting-and-emotional-data-mining-scripts-bcc3c428d77f

[14] Sweeney, L., 2000. Simple demographics often identify people uniquely. *Health (San Francisco)*, *671*, pp.1-34.

identification criteria, k-anonymity[15], and an algorithm to achieve it, were proposed directly after Latanya Sweeney's attack. A dataset is said to be k-anonymous if no combination of user attributes (e.g. year of birth, sex, and postcode) are shared by fewer than *k* individuals. This makes it impossible to uniquely identify a specific person in the dataset as any information collected will always lead us to a group of at least *k* individuals. Datasets can be modified in various ways to make them k-anonymous: values in the dataset are coarsened (e.g. by recording the age range of a person rather than their exact age), certain attributes (columns) or users (rows) can be removed, etc. These principles of generalisation and deletion along with others underpin all algorithms designed to enforce k-anonymity. Extensions of k-anonymity, such as l-diversity[16] and t-closeness[17], have furthermore been proposed to protect against more complex inference attacks.

This combination of pseudonymization and de-identification worked quite well for about 15 to 20 years. However, modern datasets, and especially the datasets used by AI, are very different from those used in the mid 90s. Today's datasets, coming from phones, browsers, IoT, or smart-cities, are high-dimensional: they contain for each individual hundreds or thousands of pieces of information about him and the way he behaves. Mobile phone metadata contain all the places where an individual has used their phone, sometimes for years. Web browsing data contain every single pages you have visited while a human genome is composed of approx. 21,000 genes.

This fundamentally changes the ability of anonymization methods to effectively protect people's privacy while allowing the data to be used. Following several high-profile re-identification of behavioral datasets[18] [19], the concept of unicity was introduced in 2013 to evaluate the effectiveness of anonymization in modern datasets. Unicity, estimates the fraction of users that are uniquely identified by a number of randomly chosen pieces of information an adversary could have access to. A study based on mobile phone metadata, showed that just 4 points—approximate times and places—are sufficient to uniquely identify 95% of people in a dataset of 1.5 million individuals[20]. This means that knowing where and when an individual was a mere 4 times in the span of 15 months is, on average, sufficient to re-identify them in a simply anonymized mobile phone dataset, unraveling their entire location history.

Originally obtained in a European country, these results have now been replicated several times. A 2015 study looks at a dataset of 1M people in Latin America[21] while another replicates the results on a dataset of 0.5M individuals in a third country[22]. In 2015, the same methodology was applied to bank transaction data (credit and debit cards). This study, published in Science, concluded that 4 points — date and place of a purchase—were here again sufficient to uniquely identify 90% of people among one million credit card users[23].

While pseudonymization and simple anonymization utterly fail to protect people's privacy could generalisation, deletion, and other methods throw people off the scent again? Unfortunately, for both mobile phones and credit cards data, the answer is a resounding 'no'. The same is likely to be true for other large-scale behavioral datasets such as browsing, IoT data etc. The above studies demonstrate that adding noise or reducing the spatial or temporal resolution of data makes identification only marginally more difficult. Indeed, even in a very low-resolution mobile phone dataset[24], 10 points are enough to find a person more than 50% of the time[25]. Surprisingly perhaps, in the credit card study, knowing just 10 instances of when an individual has visited any one of 350 stores in a two-week period would result in a correct re-identification 80% of the time[26]. Deletion has mathematically the same marginal effect on the likelihood of re-identification.

These results has let researchers to conclude that "*we have currently no reason to believe that an efficient enough, yet general, anonymisation method will ever exist for high-dimensional data, as all the evidence so far points to the contrary. The current de-identification model, where the data are anonymised and released, is obsolete*"[27]. An opinion shared by President's [Obama] Council of Advisors on Science and Technology who concluded that anonymisation "*is not robust against near-term future re-identification methods. PCAST does not see it as being a useful basis for policy*"[28].

[15] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(05), 557-570.

[16] Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006, April). l-diversity: Privacy beyond k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on* (pp. 24-24). IEEE.

[17] Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on* (pp. 106-115). IEEE.

[18] Michael Arrington (August 6, 2006). "AOL proudly releases massive amounts of user search data". TechCrunch. Retrieved August 7, 2006

[19] Narayanan, A. and Shmatikov, V., 2006. How to break anonymity of the netflix prize dataset. arXiv preprint cs/0610105.

[20] de Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports, 3*, 1376.

[21] U.N. Global Pulse. Mapping the risk-utility landscape of mobile phone data for sustainable development & humanitarian action, 2015.

[22] Yi Song, Daniel Dahlmeier, and Stephane Bressan. Not so unique in the crowd:a simple and effective algorithm for anonymizing location data. ACM PIR, 2014.

[23] de Montjoye, Y. A., Radaelli, L., & Singh, V. K. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science, 347*(6221), 536-539.

[24] With the resolution reduced by a factor of 15 both temporally and spatially, approx. 15km² and 15 hours.

[25] de Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3, 1376.

[26] de Montjoye, Y. A., Radaelli, L., & Singh, V. K. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. Science, 347(6221), 536-539.

[27] de Montjoye, Y-A and Pentland, A, Response to Comment on "Unique in the shopping mall: On the re-identifiability of credit card metadata", 351, 6279, 1274--1274 (2016)

[28] https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

To make the matter worse, modern datasets are not only impossible to anonymize but also extremely rich. In the past, it was sufficient to look through the data to assess the potential damage of re-identification (e.g. whether these are medical records or fairly innocuous data). Sometimes sensitive information could even be removed to make the data "non"-sensitive (e.g. removing the fact that people might have watched specific movies). As we have seen in the Cambridge Analytica example, this doesn't work anymore with modern high-dimensional datasets. Their richness means that the sensitivity of the dataset might not be directly visible but instead come from what can be inferred from it. To assess the sensitivity of the data, one would needs to guess what an algorithm could possibly infer about an individual from his data, now or in the future. For instance, it has been shown that personality traits[29], demographics[30], socioeconomic status[31,32], or even loan repayment rates[33] can all be predicted from seemingly innocuous mobile phone data. This "risk of inference" in big data renders comprehensive risk assessments incredibly challenging — some would say impossible — to perform.

With the traditional de-identification model failing us how do we move forward training machine learning models on large-scale datasets in a way that truly preserves individuals' privacy?

Back in the 90s, when the first de-identification algorithms were developed, data transfer was exceedingly costly. Anonymizing the dataset once and for all and sending a copy of it to the analyst was the only feasible solution. 20 years later with internet, the cloud, and arrays of GPU powered machines, this is no longer the case. Data controllers can easily grant remote, tightly controlled and monitored access to datasets for training purposes instead of sharing the "anonymized" raw records — bringing algorithms to the sensitive data instead of the sending data to the algorithms.

For example, the OPen ALgorithms (OPAL) project[34], recently funded by the French Development Agency (AFD), is based on this framework. Led by the Computational Privacy Group at Imperial College London, in partnership[35] with Telefonica and Orange, OPAL aims to allow third parties to safely use the geolocation data through a questions-and-answers model. In short, the platform allows third-parties, such as researchers, to submit algorithms that will be trained on the data. The privacy of individuals is ensured through a series of control mechanisms put in place. For example, the platform validates the code before training the model; it ensures that only aggregated results sometimes with a little bit of noise are returned[36], ensuring that no single individual can be identified; and it records every interaction in a tamper-proof ledger ensuring auditability of the system. The combination of access-control mechanisms, code sandboxing, aggregation schemes, etc allows OPAL to guarantee that data is being used anonymously by machine learning algorithms and that even if the data itself is only pseudonymous.

Recognizing the issue, several other privacy-enhancing technologies (PET) are being developed to allow datasets to be used in a privacy-conscientious way through a mix of access-control, security based, and auditing mechanisms. Google's DeepMind is, for instance, developing an auditable system to train machine learning algorithms on individual-level health data records from the National Health Service[37] in the UK. Their 'Verifiable Data Audit' ensures that any interaction with the data is recorded and accessible to mitigate the risk of foul play. The French government also developed a similar solution, the Secure Data Access Centre (CASD)[38], to allow researchers to build statistical models using public surveys and national censuses through remote access and smartcard technologies.

AI and machine learning could revolutionize the way we work and live. Their potential is however crucially dependent on access to large and high-quality datasets for algorithms to be trained on. The way we have historically found a balance between using the data in aggregate and protecting people's privacy, de-identification, does not scale to the big data datasets used by modern algorithms. Moving forward, it is both crucial for our algorithms to be trained on the best available datasets out there and to do so in a way that truly protects the privacy of the individuals. The successful future of AI requires us to rethink our approach to data protection. Solutions like OPAL are at the forefront of this effort, forming the bedrock of safely using large-scale sensitive data for the public good.

> **"THE COMBINATION OF ACCESS-CONTROL MECHANISMS, CODE SANDBOXING, AGGREGATION SCHEMES, ETC. ALLOWS OPAL TO GUARANTEE THAT DATA IS BEING USED ANONYMOUSLY BY MACHINE LEARNING ALGORITHMS AND THAT EVEN IF THE DATA ITSELF IS ONLY PSEUDONYMOUS."**

[29] de Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. (2013, April). Predicting Personality Using Novel Mobile Phone-Based Metrics. In *SBP* (pp. 48-55).

[30] Felbo, B., Sundsøy, P., Pentland, A. S., Lehmann, S., & de Montjoye, Y. A. (2015). Using deep learning to predict demographics from mobile phone metadata. *arXiv preprint arXiv:1511.06660.*

[31] Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L., & de Montjoye, Y. A. (2017). Improving official statistics in emerging markets using machine learning and mobile phone data. *EPJ Data Science*, *6*(1), 3.

[32] de Montjoye, Y. A., Rocher, L., & Pentland, A. S. (2016). Bandicoot: a python toolbox for mobile phone metadata. *Journal of Machine Learning Research*, *17*(175), 1-5.

[33] Bjorkegren, D., & Grissen, D. (2015). Behavior revealed in mobile phone usage predicts loan repayment.

[34] Open Algorithms (2017), OPAL, www.opalproject.org/.

[35] *Other partners include: Data-Pop Alliance, MIT and the World Economic Forum*

[36] See e.g. differential privacy Dwork, C., 2008, April. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation* (pp. 1-19). Springer, Berlin, Heidelberg.

[37] Suleyman, M., Laurie, B, (2017), Trust, confidence and Verifiable Data Audit, *DeepMind Blog*, https://deepmind.com/blog/trust-confidence-verifiable-data-audit.

[38] Centre d'accès Sécurisé aux Données, CASD, https://casd.eu/en.