SAFEST: A Safeguarding Analytical Framework for dEcentralised Senstive daTa

Patricia Ryser-Welch, Leire Abarrategui, Soumya Banerjee

July 3, 2023

Abstract

An increasing demand and dependence of analyzing a data has been driven by *"Big Data"* and *"Internet of Things (IoT)"*. Scientific reproducibility, robustness and the cost of capturing new data has been improved through findable, accessible, interoperable, and reusable data sharing. Ethical and legal restrictions impose the use of privacy preservation and protection measures for any disclosive and sensitive information. We, therefore, present a possible model to support multi-disciplinary research team to protect against disclosure of individual-level data and large datasets used in other disciplines. We argue a technology reliance is not enough and a continuous collaboration that adapt to new cyber-security, and data inferential threat is needed. We consequently conclude some standards could lead to closer collaboration to support research and innovation in the long term.

1 Introduction

The fields of *"Big Data"* and *"Internet of Things (IoT)"* have positively responded to the increasing demand and dependence of analyzing a data for research purposes. Technological solutions have empowered scientific research with improved scientific reproducibility, robustness alongside a cost reduction of capturing new data [10,20]. To some extend, such endeavour has also addressed some need raised by the FAIR movement for data sharing; i.e., Findable, Accessible, Interoperable, Reusable [19]. Notwithstanding such positive outcomes, physical access to individual-level data may not always be possible; individuals may cede control over their privacy [63] and ethical, legal, and regulatory restrictions prevent sharing and accessing disclosive or sensitive information. Intellectual property or licensing issues associated with research can also impose some additional barriers [38].

Health-care, biomedical, and social sciences research depends on accessing ethically and legally large individual-level data. Transferring datasets becomes an impractical and challenging task, that a decentralisation approach relying on researchers' collaboration can overcome. The federation of existing systems

can cooperate with autonomous and heterogeneous computer systems [32]. The idea that data computations can be remotely-called behind firewalls becomes an attractive solution [38]. Data owner can open their data to a research community, without loosing control over the use of their data. However, suitable data preservation and protection measures should be put in place with sensitive and individuals data. A balance to (1) empower data governance, (2) prevent the inferential reconstruction of datasets, and (3) individual identification should be considered [49].

This paper, therefore, proposes a privacy-preserving and protecting decentralised architecture for the purpose of ethical and legal analysis. The latter enables data governers and custodians to share in a controlled manner highly-sensitive data with a reduced data governance burden. The architecture should also enable data analysts and researchers analysing data in an ethical and legal manner. Finally, research can be supported without lengthy transfer of datasets. The paper discusses various approaches and application of technologies applied to prevent disclosure of individual-level data. We finally suggest collaborative effort to adapt to societal needs as well as research needs.

2 Background

Privacy keeps information about individuals from being available to others [16]. This non-trivial task is becoming an even more challenging task; the by-product of *humans*, *Internet*, *WWW* and *devices* activities produce some intangible individuals' personal data at an alarming and increasing rate. Behaviour, preferences, and other sensitive details have become even more traceable, leading to predict future trends and product developments [47, 57]. Some platforms provide a dynamic approach for data capture from sizeable social media and geospatial data [5, 29, 41]. Vehicle sensors, and other spatio-temporal data has helped predict traffic flow in a city, support driverless vehicles, and quantify air pollution and traffic [15, 25, 37]. On the other hand, contact-tracing apps for fighting against the COVID-19 can collect and retain individuals' data [4]. The emergence of *"Internet of Medical Things" (IoMT)* is to integrate medical devices to bring a better connected health care provision. Bioinformatics and population-health medical research has also generated large quantities of data stored in various digital forms [12].

A growing concern is becoming more apparent this data is awaiting to be stored and analysed by the immense analytical power offered by Internet-Based services and become exploitable business assets to improve business activities, and products development, and medical research [2, 28, 34, 36, 42, 47, 57]. For example, the application of machine learning algorithms extends advanced statistical methodologies to analyse unstructured data (i.e., images, documents, social media entries). Amongst other, unknown patterns about individuals can potentially become identifiable without their knowledge and benefit others. Stolen individuals' details (i.e., a physical person or an organisation) can be used to commit a crime or damage reputation. Data analytics can also impact posi-

tively and negatively at a societal level; digital totalitarianism is arising in some countries [53]. Data mining techniques, and language models could also assist in predicting electoral exercise outcomes; the latter can empower campaigning, and political advertising [22, 23, 43, 44, 56]. Nevertheless, language models can also help us understand the propagation of fake news [3].

Despite these newly created phenomena, the COVID-19 pandemic - and the sudden adoption of some tracing technology to overcome the crisis - has globally engaged discussions about (1) data collection, (2) the potential future breaches of individuals' privacy, (3) confidentiality, and (3) possible identification of individuals without their knowledge. Before, individuals may have been oblivious of such issues [33, 68]. Consequently, multi-disciplinary research communities are increasingly discussing ethical considerations, privacy preservation, and protection of contact-tracing apps, social media usage, and other methods used to capture data [51, 60, 62].

A recent decentralisation movement of the world wide web is attempting to empower individuals with the data governance of their own data. An approach referred as *"Solid"* should offer individuals "pods" to keep and manage the data generated from social media, entertainment, shopping, and other web activities. Websites, web apps, and other web services should request data from these pods, rather than centrally storing the personal data. Individuals should allow organisations using the data [13, 45, 55]. Another approach has explored how the sharing of large files containing sensitive information with *InterPlanetery File System* (IPFS) and Blockchain technologies. Individuals appear to register files, grant and revoke access to them [6, 39, 59]. At the time of writing, these ideas remain at an early stage of development. It is likely to take some time and effort to bring them to a more mature option to the current implementation of the world wide web.

Previous decentralisation efforts benefited from the development of network interconnection technologies and service-oriented software architectures [7, 17, 21, 52]. These highly heterogeneous technologies led to the federation of systems capable of autonomous and distributed systems, capable of evolving and enabling data sharing [11, 32, 38]. Direct access remains unsuitable for highly-sensitive data without any data agreements. An alternative approach brings some hopes; it distributes the computations to the data using remote procedure calls to privacy-preserving computations. Referred as *the DataSHIELD approach*, it denies direct users access to individual levels of data, but offers the opportunity to analyse remotely harmonised individual level data without leaving their host site. The outcome of these calculations should lower the risks of inferential data reconstruction through some carefully thought disclosure controls [26, 46, 49].

3 Proposed framework

We propose a theoretical framework that decentralises secure multiparty computations in a context of federated analytical systems, to prevent any direct access

to the data. Trusted parties respond to some requests by completing some computations; it often employs some "need-to-know" standards that accomplish allowable tasks and only disclose computations' results [30, 67]. Referred as SAFEST, the proposed framework is inspired from network interconnection technologies, service-oriented software architectures, and the *DataSHIELD approach*. Each *server-side* component (shown in blue in figure 1) decentralises users' management, access right and privacy-preserving parametrisation, allowing restricted server-side computation call and execution. Some *client-side software* - shown in green in figure 1 - should enable the analysis and comparison of data, as if the data were outsourced to a central computer and jointly analysed. Communications between data analysts and some analytical servers can only occur using secured networking protocols and server-side connections.

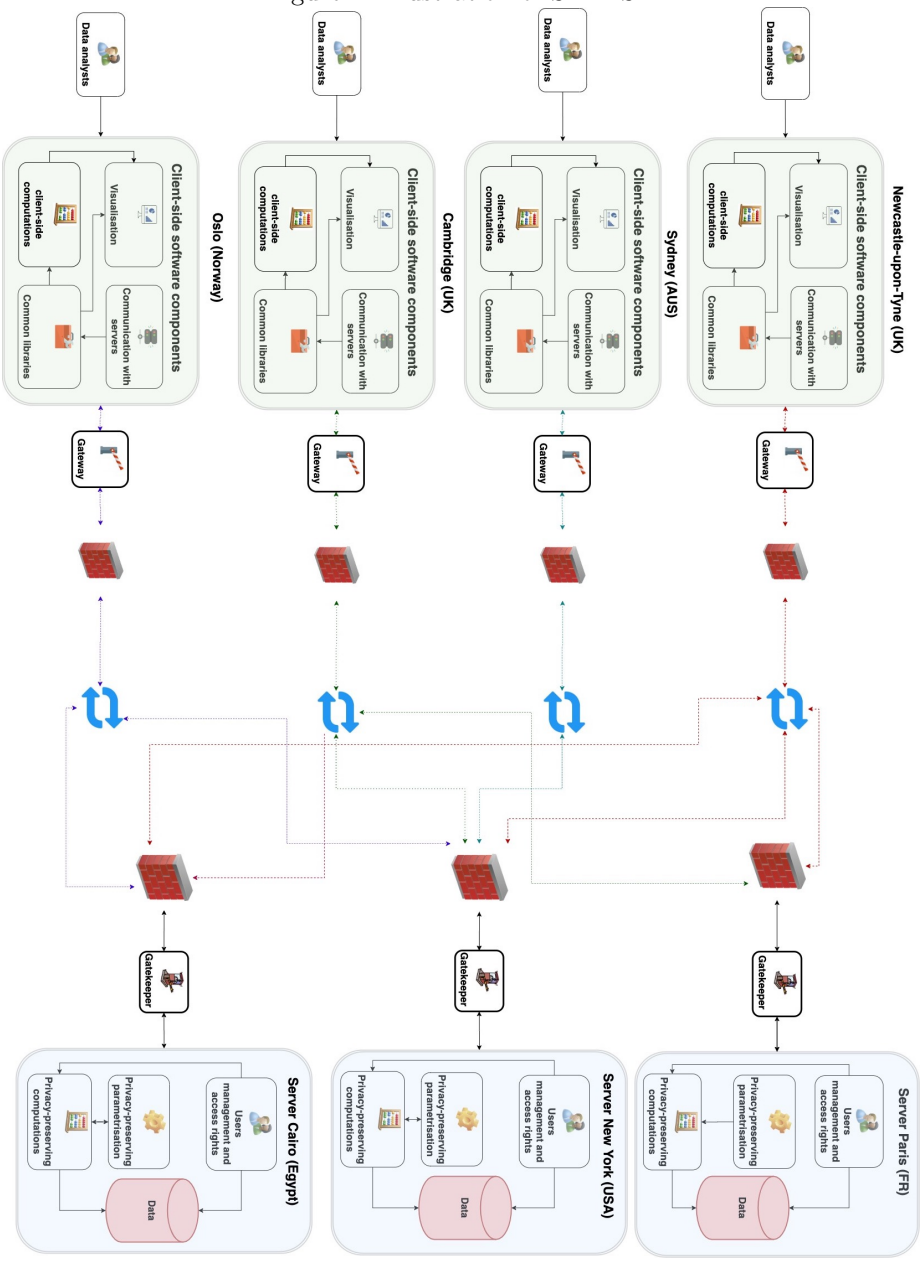
Some client-side components - visualisation, user-interaction, client-side computations, and networking tools - should bring some transparency, to obfuscate any remote-calculation calls to data analysts. Remote analysts should undoubtedly be informed of any server-side privacy-preserving and protective-features. A *Gateway* (client-side) should manage connections and process responses from any participating servers. A *Gatekeeper* (server-side) should authorises requests, validates users, and stops any disclosive responses. Invalid or unauthorised calls should be blocked.

Data governance should remain responsible to grant or revoke remote access to data analysts. Users' management and access rights should empower data governers authorising data analyst executing specific server-side computations on some specific datasets. Some privacy-preserving parameterisation control the server-side computations' behaviour and outcome. Consequently, only encryption or some server-side computations' results can be transmitted back to the client-side components. These privacy-preserving and protective features inform directly the *Gatekeeper* - only enabling authorised execution of computations on their own participating servers and preventing outcomes being returns under certain criteria.

4 Discussion

Individuals' identification may only occur with an in-depth domain knowledge and having access certain type of data. Time-series data related to pandemics and genomic data is unlikely to be considered as highly-sensitive, reducing some of the needs for user management, access rights, privacy-preserving parametrisation, and computations. The *UCSC Genome Browser* [1], and the *Covid-19 data hub* [31] are significant examples of such fully distributed systems, in which data needs to be outsourced to some storage and processing unit. Analysts can either (1) download and analyse data using a command-line interface provided by functional languages on their computer or (2) use a web interface providing some analytical and upload tools. BioConductor [27] enriches similar functionalities with analytical software libraries; any analytical computations relies on researchers computing facilities for their execution. No computation has yet

Figure 1: Illustration of SAFEST



to be brought to the data. Therefore, analysing large datasets can become an issue, if access to high-computation facilities become an issue.

The privacy-preservation and protection features - introduced in the previous section (see fig. 1) - may be needed for healthcare, biomedical and social sciences research. Phenotypic and other highly-sensitive data are likely to be analysed. Federation trends diverge in approaches, adopting emerging technologies and security standards. For example, some cloud-based solution relies on a specific platform to develop advanced analytical tools, such as machine learning algorithms. Such trend follows cloud technologies market leaders who have adapted software, data storage and analytical tools as a service to the privacy-protection needs of medical health care providers. Data analysts depend on a web interface to query directly the data and its meta-data. Cloud technologies also transfer heavily encrypted data across trusted consortia. Such approach has been adopted by MedCo [54]. There is little evidence that privacy-preserving parameterisations and computations have yet to be implemented. Data governance may not agree with the data being transferred outside their organisations and large transfer of data may be impractical.

Some specialised server-side warehouse software may offer some *"lighter-weight"* solutions, preventing any transfer of raw data, instead data is kept behind the firewall of an organisation. In such solutions, data can either be uploaded within the specialised software or linked to existing systems. Renku [69] offers liberal data sharing by data-repository use, with some privacy-protection features. The latter are dependent of privacy-protection offered by the chosen technology, and hence their suitability should be verified for the type of data shared prior use.

Personal-Health-Train [8], MOLGENIS [61] offers primarily the opportunity of discovering research datasets to the wider community. An organisation firewall appears to bring data-protection. However, very little evidences of requests being encrypted using of secure socket layer protocols or other mechanisms may be a concern. Especially, as batch data download appears to be feasible. *Cafe Variome* [40] functionalities include some specialised web server that obfuscates some data using privacy-preserving computations. Privacy levels can be set to manage access to some datasets, i.e., linked, public or private. The aim of *Cafe Variome* is data discovery rather than analysis, at the time of writing. Most of the aforementioned systems either offer a well-designed web interface or some scripts using programming languages designed for data analysis, i.e., R or Python. Some strict data agreements may need arranging before participating to any collaborations between research partners.

Remote procedure calls (RPC) can bring the computations to some data, where researchers remotely analyse some harmonised individual-level data. RPC can usefully implement a 'hub-and-spoke' design, bringing a step closer some implementation to fig. 1. Various data-protection and data-preservation techniques may need to be considered and implemented to prevent disclosure.

ViPaR [14] empowers data governers to authorise access to the data; no privacy-preservation settings are yet to be included. Encrypted anonymised data may be transferred but not saved permanently on a central server hard

disk. Researchers may not access directly the data, but the hosting organisation is likely to view data with appropriate tools. Unlike fig. 1 some federation of the data occurs at server side, rather than letting the researchers taking the decision to connect to specific data sources. Pooled data can therefore be analysed using server-level or virtually-pooled analysis, using an implemented analytical pipeline. An analytical web interface and some command line interface is available to conduct some remote analysis. To maintain a resilience to a low disclosure level encryption and anonymisation methods need continuously adapt to security threats and counter-measures. The use of ViPar within the same organisation may also improve the level of protection.

Another approach to 'hub-and-spoke' design brings more freedom to the data analysts. Some specialised client-side libraries, server-side libraries, and warehouse software brings some multi-platforms components to act as integration between existing systems and data analysts. Vantage6 [35] solution has yet to offers any gatekeepers, users' management, access right, and privacy-preserving parameterisation. It is yet unclear whether any forms of encryption, such as homomorphic data encryption at the server-level or secure-socket layer at transmission level, are required. Vantage6 has a "lighter-weight" use of technology to adapt to various needs dictated by the data and also supports multiple programming languages.

DataSHIELD [26, 46, 65] has developed from a prototype to a development platform for privacy-preserving federated analysis; it depends on the functional programming language R. Visualisation, client-side computations, and secured connections to the servers are implemented in R and Java. The data analysts have yet to use any web interfaces. Instead, data analysts can use R scripts, notebooks, vignettes, or create their own web interface to analyse and share their analysis results. Data governers can edit authorised requests, privacy-preserving parameterisation, user management, and access rights using specialised warehouse software. The latter can resource data kept outside the server, in the same organisation. Existing organisational security features, implementation of privacy-preserving computations, and also the settings of privacy-preserving parameters should lower disclosure. Data governance tools and their application should maintain resilience to data protection and data preservation.

Hub-and-spoke architecture has also been adopted in exploring the decentralisation of machine learning, such as deep learning, to multiple devices and cloud platforms. Machine learning algorithms are brought to some training and testing data sets distributed in nodes; some of them are distributed within consortia or across organisations [18, 24, 50]. Privacy-preservation using encryption techniques, such blockchain or homomorphic encryption are considered as state-of-the art, at the time of writing [48, 66]. However, the use a gatekeeper, users management and access rights as well as some privacy-preserving parameterisations has yet to be considered or acknowledged by the community.

Ongoing collaboration can act as gatekeepers (see fig. 1) to reduce disclosure. Verification of analytical results has yet to be automated, instead experts assessment is made; an analytical pipeline has integrated computing and human processes. Such endeavours appears to succesfully support COVID 19

and other causes. The number of publication using OpenSafely has rapidly increased exponentially as an impressive amount of raw NHS data can now be analysed [9, 58, 64]. Such process should lower suitably the level of disclosure to a certain optimum. Notwithstanding the positive contributions of this model, the human validation process may slow down the analysis pipeline. Adding disclosure controls alongside analytical tools may partially overcome such issues, while limiting disclosure. Automated and more traditional gatekeeping techniques may complement each other and preserve a low-level of disclosure.

5 Future research and conclusion

A real momentum exists in bringing computer scientists, medical practitioners as well as other disciplines together to federate systems and support researchers. The distribution of advanced statistical and machine learning methodologies has been distributed to address specific research questions. It would be beneficial attempting generalising the distribution techniques and algorithms for more than one purpose. With time, we anticipate some effort in this direction may naturally arise.

Collaboration is successfully integrating combinations of software to build distributed architectures. Little doubt exists that organisations IT services and researchers should work together to implement and maintain consortia close to some proposed framework. Additional research may need to explore the best approaches to bring together IT professionals, researchers, and their own perceptions of solving problems. Such collaboration may consider adaptability, flexibility and ability to integrate with existing systems. Some lightweight approaches or hub-and-spoke solutions appears to offer better opportunities for adapting to existing systems. It is possible lightweight approaches could become more resilient to a fast advancing technological ecosystems, we are experiencing.

Without the creation of standards and white papers distributions of computer systems over the Internet would have not been so successful. This model should be an inspiration for health-care, biomedical and other industries to drive innovations and resilience to future technological advancement.

We have presented a possible model to support multi-disciplinary research team to protect against disclosure of individual-level data. The model could also apply to large datasets used in other disciplines, such as astronomy. We conclude not all solutions should rely on technology on its own, but also continuous collaboration to adapt to new cyber-security, and data inferential threat. We have also concluded some standards could lead to closer collaboration to support research and innovation in the long term.

References

[1] Genome browser gateway. http://genome.ucsc.edu, 2020. [Online; accessed 18-december-2020].

[2] S. Abid, K. Keshavjee, A. Karim, and A. Guergachi. What we can learn from amazon for clinical decision support systems. *Studies in Health Technology and Informatics*, 234:1–5, 2017.

[3] O. D. Apuke and B. Omar. Fake news and covid-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, page 101475, 2020.

[4] M. A. Azad, J. Arshad, S. M. A. Akmal, F. Riaz, S. Abdullah, M. Imran, and F. Ahmad. A first look at privacy analysis of covid-19 contact tracing mobile applications. *IEEE Internet of Things Journal*, 2020.

[5] A. Bechini, D. Gazzè, A. Marchetti, and M. Tesconi. Towards a general architecture for social media data capture from a multi-domain perspective. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, pages 1093–1100. IEEE, 2016.

[6] J. Benet. Ipfs-content addressed, versioned, p2p file system. *arXiv preprint arXiv:1407.3561*, 2014.

[7] T. Berners-Lee, R. Cailliau, J.-F. Groff, and B. Pollermann. World-wide web: the information universe. *Internet Research*, 1992.

[8] O. Beyan, A. Choudhury, J. van Soest, O. Kohlbacher, L. Zimmermann, H. Stenzhorn, M. R. Karim, M. Dumontier, S. Decker, L. O. B. da Silva Santos, et al. Distributed analytics on sensitive medical data: The personal health train. *Data Intelligence*, 2(1-2):96–107, 2020.

[9] K. Bhaskaran, S. Bacon, S. J. Evans, C. J. Bates, C. T. Rentsch, B. MacKenna, L. Tomlinson, A. J. Walker, A. Schultze, C. E. Morton, et al. Factors associated with deaths due to covid-19 versus other causes: population-based cohort analysis of uk primary care data and linked national death registrations within the opensafely platform. *The Lancet Regional Health-Europe*, 6:100109, 2021.

[10] F. Breden, E. T. Luning Prak, B. Peters, F. Rubelt, C. A. Schramm, C. E. Busse, J. A. Vander Heiden, S. Christley, S. A. C. Bukhari, A. Thorogood, et al. Reproducibility and reuse of adaptive immune receptor repertoire data. *Frontiers in immunology*, 8:1418, 2017.

[11] S. Busse, R.-D. Kutsche, U. Leser, and H. Weber. Federated information systems: Concepts, terminology and architectures. *Forschungsberichte des Fachbereichs Informatik*, 99(9):1–38, 1999.

[12] A. J. Butte. Challenges in bioinformatics: infrastructure, models and analytics. *TRENDS in Biotechnology*, 19(5):159–160, 2001.

[13] R. Buyle, R. Taelman, K. Mostaert, G. Joris, E. Mannens, R. Verborgh, and T. Berners-Lee. Streamlining governmental processes by putting citizens in control of their personal data. In *International Conference on*

Electronic Governance and Open Society: Challenges in Eurasia, pages 346–359. Springer, 2019.

[14] K. W. Carter, R. W. Francis, K. Carter, R. Francis, M. Bresnahan, M. Gissler, T. Grønborg, R. Gross, N. Gunnes, G. Hammond, et al. Vipar: a software platform for the virtual pooling and analysis of research data. *International journal of epidemiology*, 45(2):408–416, 2016.

[15] C. Chen, K. Li, S. G. Teo, G. Chen, X. Zou, X. Yang, R. C. Vijay, J. Feng, and Z. Zeng. Exploiting spatio-temporal correlations with multiple 3d convolutional neural networks for citywide vehicle flow prediction. In *2018 IEEE international conference on data mining (ICDM)*, pages 893–898. IEEE, 2018.

[16] C. Clifton, M. Kantarcioglu, and J. Vaidya. Defining privacy for data mining. In *National science foundation workshop on next generation data mining*, volume 1, page 1. Citeseer, 2002.

[17] Ç. Cömert. Web services and national spatial data infrastructure (nsdi). In *Proceedings of Geo-Imagery Bridging Continents, XXth ISPRS Congress*, volume 4. Citeseer, 2004.

[18] Q. Dou, T. Y. So, M. Jiang, Q. Liu, V. Vardhanabhuti, G. Kaissis, Z. Li, W. Si, H. H. Lee, K. Yu, et al. Federated deep learning for detecting covid-19 lung abnormalities in ct: a privacy-preserving multinational validation study. *NPJ digital medicine*, 4(1):1–11, 2021.

[19] A. Dunning, M. De Smaele, and J. Böhmer. Are the fair data principles fair? *International Journal of digital curation*, 12(2):177–195, 1970.

[20] Editorial. Data sharing and the future of science. *Nature Communications*, 9(1):2817, 2018.

[21] C. Ferris and J. Farrell. What are web services? *Communications of the ACM*, 46(6):31, 2003.

[22] G. M. Fulgoni, A. Lipsman, and C. Davidsen. The power of political advertising: Lessons for practitioners: How data analytics, social media, and creative strategies shape us presidential election campaigns. *Journal of Advertising Research*, 56(3):239–244, 2016.

[23] S. Fuller. Brexit as the unlikely leading edge of the anti-expert revolution. *European Management Journal*, 35(5):575–580, 2017.

[24] M. N. Galtier and C. Marini. Substra: a framework for privacy-preserving, traceable and collaborative machine learning. *arXiv preprint arXiv:1910.11567*, 2019.

[25] C. K. Gately, L. R. Hutyra, S. Peterson, and I. S. Wing. Urban emissions hotspots: Quantifying vehicle congestion and air pollution using mobile phone gps data. *Environmental pollution*, 229:496–504, 2017.

[26] A. Gaye, Y. Marcon, J. Isaeva, P. LaFlamme, A. Turner, E. M. Jones, J. Minion, A. W. Boyd, C. J. Newby, M.-L. Nuotio, et al. Datashield: taking the analysis to the data, not the data to the analysis. *International journal of epidemiology*, 43(6):1929–1944, 2014.

[27] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):1–16, 2004.

[28] G. Gimpel. Bringing dark data into the light: illuminating existing iot data lost within your organization. *Business Horizons*, 2020.

[29] T. Gisselbrecht, L. Denoyer, P. Gallinari, and S. Lamprier. Whichstreams: A dynamic approach for focused data capture from large social media. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[30] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game, or a completeness theorem for protocols with honest majority. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, pages 307–328. 2019.

[31] E. Guidotti and D. Ardia. Covid-19 data hub. *Journal of Open Source Software*, 5(51):2376, 2020.

[32] D. Heimbigner and D. McLeod. A federated architecture for information management. *ACM Transactions on Information Systems (TOIS)*, 3(3):253–278, 1985.

[33] R. Herschel and V. M. Miori. Ethics & big data. *Technology in Society*, 49:31–36, 2017.

[34] M. Hobart. The 'dark data'conundrum. *Computer Fraud & Security*, 2020(7):13–16, 2020.

[35] T. Hulsen. Sharing is caring—data sharing initiatives in healthcare. *International Journal of Environmental Research and Public Health*, 17(9):3046, 2020.

[36] V. Khandelwal, A. Chaturvedi, and C. P. Gupta. Amazon ec2 spot price prediction using regression random forests. *IEEE Transactions on Cloud Computing*, 2017.

[37] Z. Koppanyi and C. Toth. Experiences with acquiring highly redundant spatial data to support driverless vehicle technologies. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4(2), 2018.

[38] S. Kowalczyk and K. Shankar. Data sharing in the sciences. *Annual review of information science and technology*, 45(1):247–294, 2011.

[39] R. Kumar and R. Tripathi. Large-scale data storage scheme in blockchain ledger using ipfs and nosql. In *Large-Scale Data Streaming, Processing, and Blockchain Security*, pages 91–116. IGI Global, 2021.

[40] O. Lancaster, T. Beck, D. Atlan, M. Swertz, D. Thangavelu, C. Veal, R. Dalgleish, and A. J. Brookes. Cafe variome: General-purpose software for making genotype–phenotype data discoverable in restricted or open access contexts. *Human mutation*, 36(10):957–964, 2015.

[41] A. Larkin and P. Hystad. Integrating geospatial data and social media in bidirectional long-short term memory models to capture human nature interactions. *The Computer Journal*, 2020.

[42] K. Law and F.-L. Chung. Knowledge-driven decision analytics for commercial banking. *Journal of Management Analytics*, 7(2):209–230, 2020.

[43] A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic. The party is over here: Structure and content in the 2010 election. *ICWSM*, 11(2011):17–21, 2011.

[44] T. Mahmood, T. Iqbal, F. Amin, W. Lohanna, and A. Mustafa. Mining twitter big data to predict 2013 pakistan election winner. In *INMIC*, pages 49–54. IEEE, 2013.

[45] E. Mansour, A. V. Sambra, S. Hawke, M. Zereba, S. Capadisli, A. Ghanem, A. Aboulnaga, and T. Berners-Lee. A demonstration of the solid platform for social web applications. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 223–226, 2016.

[46] Y. Marcon, T. Bishop, D. Avraam, X. Escriba-Montagut, P. Ryser-Welch, S. Wheater, P. Burton, and J. R. González. Orchestrating privacy-protected big data analyses of data from different resources with r and datashield. *PLOS Computational Biology*, 17(3):e1008880, 2021.

[47] L. M. Meier and V. R. Manzerolle. Rising tides? data capture, platform accumulation, and new monopolies in the digital music economy. *New Media & Society*, 21(3):543–561, 2019.

[48] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

[49] M. J. Murtagh, A. Turner, J. T. Minion, M. Fay, and P. R. Burton. International data sharing in practice: new technologies meet old governance. *Biopreservation and biobanking*, 14(3):231–240, 2016.

[50] R. Nasirigerdeh, R. Torkzadehmahani, J. Matschinske, T. Frisch, M. List, J. Späth, S. Weiß, U. Völker, N. K. Wenke, T. Kacprowski, et al. splink: A federated, privacy-preserving tool as a robust alternative to meta-analysis in genome-wide association studies. *BioRxiv*, 2020.

[51] M. J. Parker, C. Fraser, L. Abeler-Dörner, and D. Bonsall. Ethics of instantaneous contact tracing using mobile phone apps in the control of the covid-19 pandemic. *Journal of Medical Ethics*, 2020.

[52] D. G. Perry, S. H. Blumenthal, and R. M. Hinden. The arpanet and the darpa internet. *Library Hi Tech*, 1988.

[53] X. Qiang. The road to digital unfreedom: President xi's surveillance state. *Journal of Democracy*, 30(1):53–67, 2019.

[54] J. L. Raisaro, J. R. Troncoso-Pastoriza, M. Misbach, J. S. Sousa, S. Pradervand, E. Missiaglia, O. Michielin, B. Ford, and J.-P. Hubaux. M ed c o: Enabling secure and privacy-preserving exploration of distributed clinical and genomic data. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(4):1328–1341, 2018.

[55] M. Ramachandran, N. Chowdhury, A. Third, J. Domingue, K. Quick, and M. Bachler. Towards complete decentralised verification of data with confidentiality: Different ways to connect solid pods and blockchain. In *Companion Proceedings of the Web Conference 2020*, pages 645–649, 2020.

[56] R. Rathi. Effect of cambridge analytica's facebook ads on the 2016 us presidential election. *Towards Data Science*, 2019.

[57] M. Redi, L. M. Aiello, R. Schifanella, and D. Quercia. The spirit of the city: Using social media to capture neighborhood ambiance. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18, 2018.

[58] A. Schultze, A. J. Walker, B. MacKenna, C. E. Morton, K. Bhaskaran, J. P. Brown, C. T. Rentsch, E. Williamson, H. Drysdale, R. Croker, et al. Risk of covid-19-related death among patients with chronic obstructive pulmonary disease or asthma prescribed inhaled corticosteroids: an observational cohort study using the opensafely platform. *The Lancet Respiratory Medicine*, 8(11):1106–1120, 2020.

[59] M. Steichen, B. Fiz, R. Norvill, W. Shbair, and R. State. Blockchain-based, decentralized access control for ipfs. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 1499–1506. IEEE, 2018.

[60] R. Sun, W. Wang, M. Xue, G. Tyson, S. Camtepe, and D. Ranasinghe. Vetting security and privacy of global covid-19 contact tracing applications. *arXiv preprint arXiv:2006.10933*, 2020.

[61] K. J. van der Velde, F. Imhann, B. Charbon, C. Pang, D. van Enckevort, M. Slofstra, R. Barbieri, R. Alberts, D. Hendriksen, F. Kelpin, et al. Molgenis research: advanced bioinformatics data software for non-bioinformaticians. *Bioinformatics*, 35(6):1076–1078, 2019.

[62] S. Vaudenay. Centralized or decentralized. *The contact tracing dilemma*, 2020.

[63] A. E. Waldman. Cognitive biases, dark patterns, and the 'privacy paradox'. *Current opinion in psychology*, 31:105–109, 2020.

[64] E. Williamson, A. J. Walker, K. Bhaskaran, S. Bacon, C. Bates, C. E. Morton, H. J. Curtis, A. Mehrkar, D. Evans, P. Inglesby, et al. Opensafely: factors associated with covid-19-related hospital death in the linked electronic health records of 17 million adult nhs patients. *MedRxiv*, 2020.

[65] M. Wolfson, S. E. Wallace, N. Masca, G. Rowe, N. A. Sheehan, V. Ferretti, P. LaFlamme, M. D. Tobin, J. Macleod, J. Little, et al. Datashield: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, 39(5):1372–1382, 2010.

[66] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[67] A. C.-C. Yao. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, pages 162–167. IEEE, 1986.

[68] X. Zhang and H. Bi. Research on privacy preserving classification data mining based on random perturbation. In *2010 International Conference on Information, Networking and Automation (ICINA)*, volume 1, pages V1–173. IEEE, 2010.

[69] M. M. Ziehmer and R. Rees Mertins. Data management, open access and the eth research collection. In *Group Meeting Computer Engineering Group, ETH Zurich*. ETH Library, 2019.