

Hands-On Federated Analytics

Daniel Kapitan

2025-07-06

Table of contents

Welcome	1
Image credit	1
1 Introduction	3
1.1 Key resources	4
2 The Paradox of Open	5
2.1 The engines of data science and the paradox of open	5
2.2 From closed digital platforms towards open, federated data plat- forms	6
2.3 Image credit	7
3 Summary	9
References	11

Welcome

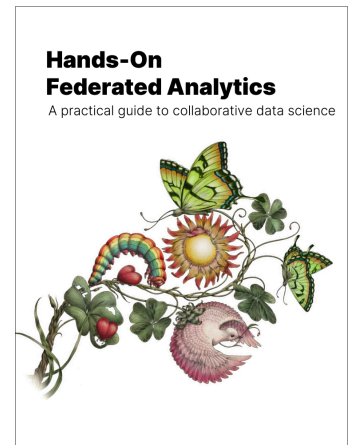
It is a truth universally acknowledged, that a data scientist in possession of a good skills, must be in want of more data.

Why I wrote this book ...

[link to some section](#)

Image credit

In the spirit of the [animal menagerie](#) of O'Reilly books, I have chosen the *Fortuna Fragilis* as the preliminary cover. This species can be found on [Terra Ultima](#) by [Raoul Deleo](#). I like to think that collaborative data science is fragile and fortuitous at the same time, and that it is worth kindling if we are to put data to use for the common good.



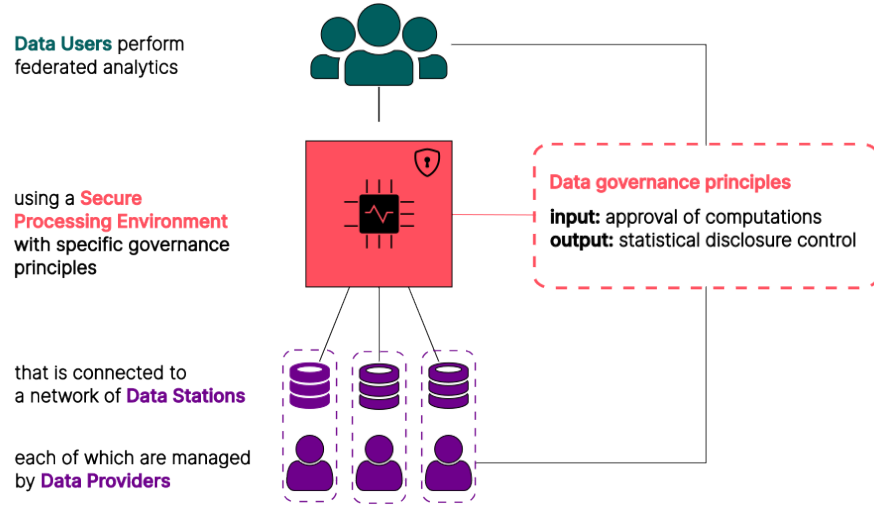
Chapter 1

Introduction

A survey on federated analytics (FA) defines it as “a paradigm for collaboratively extracting insights from distributed data that is owned by multiple parties (e.g., individual mobile devices or institutional organizations) under the coordination of a central entity (e.g., a service provider) without any of the raw data leaving their local parties or revealing information beyond the targeted insights. The core principles of this paradigm allow breaking the limitations for deriving analytics from limited centralized data, in terms of privacy concerns and operational costs.”¹

Discerning characteristics:

- Apply some form of statistical disclosure control (SDC) and/or privacy-enhancing technologies (PETs). Note we SDC can include differential privacy (DP), but we take a more generic approach that we want to protect the raw data as such, for example to provide guarantees to data holders with respect to commercial interest, trade secrets and the like. data privacy/security through PETs. Practically, all approaches to FA should have output control
- governance on the computation and not on data access as such



1.1 Key resources

Monographs, very detailed:

- Federated Learning Systems (2021) by Rehman and Gaber²
- Federated Learning (2022) by Ludwig and Baracaldo³

Review papers

- Elkordy, A.R. et al. (2023)¹
- Wang, Z. et al. (2025)⁴

Applications in healthcare

- Rieke, N. et al. (2020)⁵
- Joshi, M. et al. (2022)⁶, Fig 2 is nice diagram explaining horizontal, vertical partitioning and transfer learning

Chapter 2

The Paradox of Open

2.1 The engines of data science and the paradox of open

The [Computer History Museum](#) recounts that the history of the [Information Age](#) is driven by three ‘engines’, namely the silicon engine, the storage engine and the internet as the networking engine. Each of these engines have been instrumental in the development in of data science as we know today. Without the silicon engine, we would not be able to train foundational large language models with over 10^{23} FLOPs.[¹] Without the storage engine and the internet, we would not be able to store and collect petabyte-sized datasets such as [Common Crawl](#) as input for these models.

Besides the technological developments in and of themselves, the internet also gave birth to the open movement.[²] Numerous organisations and initiatives have been launched with a belief in openness and free knowledge. Their proponents placed their bets on the combined power of networked information services and new governance models for the production and sharing of content and data. Members of this broad movement believed it possible to leverage this combination of power and opportunity to build a more democratic society, unleashing the power of the internet to create universal access to knowledge and culture. Such openness meant not only freedom, but also presented a path to justice and equality.

We learned that open approaches flourish under two types of conditions:

1. **Projects where many people contribute to the creation of a common resource** – this is the story of Wikipedia, OpenStreetMap, Blender.org, and the countless free software projects that

provide much of the internet’s infrastructure.

2. **Circumstances where opening up is the result of external incentives or requirements, rather than voluntary actions** – this is the story of publicly-funded knowledge production like Open Access academic publications, cultural heritage collections in the Public Domain, Open Educational Resources (OER), and Open Government data.

However, the open revolution not happen. At least not on the scale that we and many other proponents of free culture expected. Although the copyright wars are almost over, conflicts about access to and control over informational resources have been superseded by conflicts about privacy, economic value extraction, the emergence of artificial intelligence, and the destabilising effects of dominant platforms on (democratic) societies. Instead of access to information, the control of personal data has emerged in the age of platforms as the critical contention.

Today, open is both a challenge to and an enabler of concentrations of power. Over the last decade, we have witnessed a wholesale transformation of the networked information ecosystem. The web moved away from the ideals and the open design of the early internet and turned into an environment that is dominated by a small number of platforms.

2.2 From closed digital platforms towards open, federated data platforms

7 8

Table 2.1: Types of data sharing and in relation to new standards and technology enablers to create openness. Taken from de Reuver et al. (2022).

	Type of data sharing	Technology enablers to create openness of health data platforms
1	Data at the most granular subject level, which is persisted and used to provide a longitudinal record.	Decentralized catalogs, semantic interoperability
2	Aggregated data, for example statistics for policy evaluation and benchmarking	Federated learning

	Type of data sharing	Technology enablers to create openness of health data platforms
3	Data analytics modules, that provide access to work and access the data.	Federated learning (FL) ⁵ and privacy- enhancing technologies (PETs) ^{9,10} : new paradigms that address the problem of data governance and privacy by training algorithms collaboratively without exchanging the data itself. Models can be trained on combined datasets and made available as open source artifacts for decision support. Data analysts can use FL and PETs to work with the data in a collaborative, decentralized fashion.
4	Trained models that have been derived from the data and can be used stand-alone for decision support.	ONNX, HuggingFace ...

2.3 Image credit

In the spirit of the [animal menagerie](#) of O'Reilly books, I have chosen the *Fortuna Fragilis* as the preliminary cover. This species can be found on [Terra Ultima](#) by [Raoul Deleo](#). I like to think that federated data science is fragile and fortuitous at the same time, and that it is worth kindling if we are to put data to use for the common good.

Chapter 3

Summary

In summary, this book has no content whatsoever.

References

1. Elkordy, A.R. et al. (2023). Federated Analytics: A Survey. *SIP* 12. [10.1561/116.00000063](#).
2. Rehman, M.H.U., and Gaber, M.M. eds. (2021). *Federated Learning Systems: Towards Next-Generation AI* (Springer International Publishing) [10.1007/978-3-030-70604-3](#).
3. Ludwig, H., and Baracaldo, N. eds. (2022). *Federated Learning: A Comprehensive Overview of Methods and Applications* (Springer International Publishing) [10.1007/978-3-030-96896-0](#).
4. Wang, Z. et al. (2025). A Survey on Federated Analytics: Taxonomy, Enabling Techniques, Applications and Open Issues. *IEEE Commun. Surv. Tutorials*, 1–1. [10.1109/COMST.2025.3558755](#).
5. Rieke, N. et al. (2020). The future of digital health with federated learning. *npj Digit. Med.* 3, 1–7. [10.1038/s41746-020-00323-1](#).
6. Joshi, M. et al. (2022). Federated Learning for Healthcare Domain - Pipeline, Applications and Challenges. *ACM Trans. Comput. Healthcare* 3, 40:1–40:36. [10.1145/3533708](#).
7. de Reuver, M. et al. (2018). The Digital Platform: A Research Agenda. *Journal of Information Technology* 33, 124–135. [10.1057/s41265-016-0033-3](#).
8. de Reuver, M. et al. (2022). The openness of data platforms: A research agenda. In *Proceedings of the 1st International Workshop on Data Economy DE '22*. (Association for Computing Machinery), pp. 34–41. [10.1145/3565011.3569056](#).

9. Scheibner, J. et al. (2021). Revolutionizing Medical Data Sharing Using Advanced Privacy-Enhancing Technologies: Technical, Legal, and Ethical Synthesis. *Journal of Medical Internet Research* *23*, e25120. [10.2196/25120](https://doi.org/10.2196/25120).
10. Jordan, S. et al. (2022). Selecting Privacy-Enhancing Technologies for Managing Health Data Use. *Front Public Health* *10*, 814163. [10.3389/fpubh.2022.814163](https://doi.org/10.3389/fpubh.2022.814163).