Data Ecology: Understanding and Designing Data Ecosystems

Raul Castro Fernandez
The University of Chicago
raulcf@uchicago.edu

Data profoundly shapes our economic, political, and social ecosystems, yet we have limited control over its influence. Unchecked dataflows among agents can distort or undermine these ecosystems. Analyzing dataflows helps us understand their use and misuse, revealing opportunities to harness their value. Beneficial dataflows, such as hospitals sharing patient data, enhance healthcare outcomes, while harmful dataflows, such as personal data sold to self-interested brokers, cause substantial damage. Equally critical are absent dataflows—such as banks or governments withholding data due to competition or mistrust—that lead to unrealized value. Despite their substantial impact, unified methods to manage dataflows effectively are lacking. Current legal (regulations), economic (incentives), and technical (privacy technologies) interventions are developed independently without a clear evaluation of their collective effectiveness.

Principles of Dataflows. To address this gap, I propose the concept of *Data Ecology*: a research agenda aiming to uncover principles that drive dataflows and to design comprehensive interventions to steer them positively. A central problem of Data Ecology is: Given a desirable outcome for a data ecosystem (such as in a company, city, or government), what interventions will lead agents' actions to that goal? The research agenda comprises three main areas: (i) formalizing the fundamental questions; (ii) designing new interventions (examples provided below); and (iii) evaluating interventions across diverse ecosystems. While some literature has explored these questions, data ecology provides a new lens that brings existing work into a common framework, helping us advance our understanding.

Preliminary and Ongoing Work. My group has explored numerous data ecosystems, including data sharing [1] and data markets [2, 3, 4, 5]. Data markets illustrate the practical value of the Data Ecology approach. Typically, data marketplaces suffer from Arrow's Information Paradox, a situation where sellers refuse to release data before payment (since data is non-rivalrous and easily duplicated), and buyers resist paying before validating the data's usefulness. This uncertainty leads to fewer transactions, limiting overall market efficiency.

Applying Data Ecology's dataflow lens, we identified this uncertainty as the critical bottleneck and developed a technical intervention called data escrow [6] to resolve it. Technically, a data escrow acts as a neutral intermediary: sellers securely register their datasets with the escrow platform, and buyers delegate specific computations to evaluate the dataset's value. For example, the escrow could run a buyer's machine learning model on the seller's dataset, returning only performance metrics (e.g., accuracy improvements) and results that pass a leakage filter. This approach enables buyers to verify data quality and relevance without compromising the seller's proprietary information, effectively eliminating data leakage risks.

Ongoing research extends this work to internal and external data-sharing markets, exploring foundational questions about data's intrinsic value [7], such as why specific dataflows emerge and what constitutes a "good" or beneficial data ecosystem.

Opportunities for the database community. The database research community is uniquely positioned to contribute to technical interventions in Data Ecology. Potential research avenues include: 1) Controlling Dataflows: Investigating privacy-preserving techniques beyond differential privacy and data escrows, such as secure multi-party computation and zero-knowledge proofs, to manage dataflows securely and flexibly. 2) Enhanced Provenance Techniques: Developing novel or adapted provenance systems that can reliably trace dataflows across complex ecosystems, providing controlled transparency where needed. 3) Internal Data Market Design: Designing effective internal marketplaces [2] to efficiently transfer tacit organizational knowledge about data use, quality, and context.

The Road Ahead. The ultimate goal of data ecology is to provide a general theory and mechanisms for understanding and controlling dataflows. Data shapes our world, but the final form need not be fixed. Data ecology's tools are more critical than ever as data's influence on our world broadens and intensifies.

REFERENCES

[1] Raul Castro Fernandez Data-sharing markets: model, protocol, and algorithms to incentivize the formation of data-sharing consortia. *SIGMOD*, 2023.

[2] Raul Castro Fernandez, Pranav Subramaniam, Michael Franklin Data market platforms: Trading data assets to solve data problems. *VLDB*, 2020.

[3] Raul Castro Fernandez Protecting Data Markets from Strategic Buyers. *SIGMOD*, 2022.

[4] Javen Kennedy, Pranav Subramaniam, Sainyam Galhotra, Raul Castro Fernandez. Revisiting Online Data Markets in 2022. A Seller and Buyer Perspective. *SIGMOD Record*, 2022

[5] Boxin Zhao, Boxiang Lyu, Raul Castro Fernandez, Mladen Kolar. Addressing Budget Allocation and Revenue Allocation in Data Market Environments Using an Adaptive Sampling Algorithm. *ICML*, 2023

[6] Siyuan Xia, Zhiru Zhu, Chris Zhu, Jinjin Zhao, Kyle Chard, Aaron Elmore, lan Foster, Michael Franklin, Sanjay Krishnan, Raul Castro Fernandez. Data Station: Delegated, Trustworthy, and Auditable Computation to Enable Data-Sharing Consortia with a Data Escrow. *VLDB*, 2022

[7] Raul Castro Fernandez What is the Value of Data?: A Theory and Systematization. *ACM/IMS Journal of Data Science*, 2025