DOI:10.1145/3416975

BY MARCELO ARENAS AND PABLO BARCELÓ

Chile's New Interdisciplinary Institute for Foundational Research on Data

THE MILLENNIUM INSTITUTE for Foundational Research on Data[a] (IMFD) started its operations in June 2018, funded by the Millennium Science Initiative of the Chilean National Agency of Research and Development.[b] IMFD is a joint initiative led by Universidad de Chile and Universidad Católica de Chile, with the participation of five other Chilean universities: Universidad de Concepción, Universidad de Talca, Universidad Técnica Federico Santa María, Universidad Diego Portales, and Universidad Adolfo Ibáñez. IMFD aims to be a reference center in Latin America related to state-of-the-art research on the foundational problems with data, as well as its applications to tackling diverse issues ranging from scientific challenges to complex social problems.

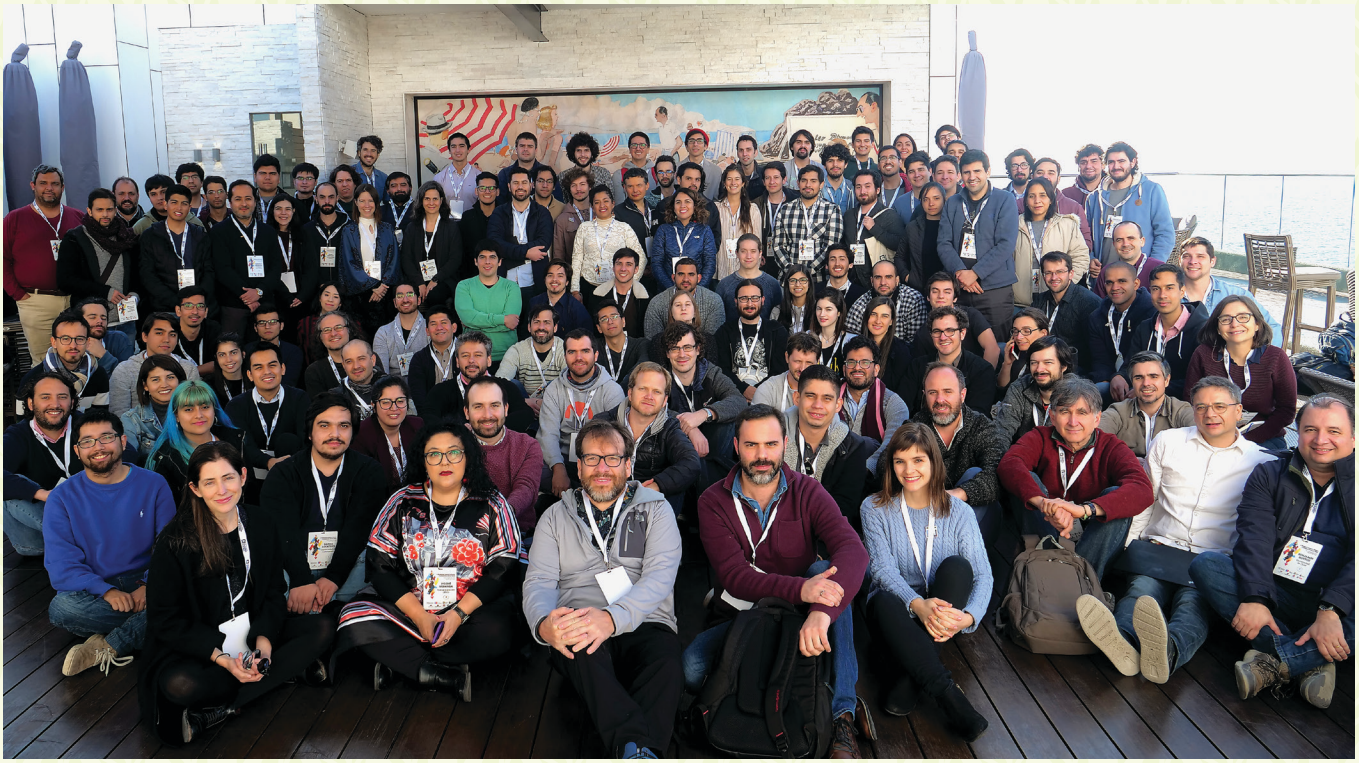
As tasks of this kind are interdisciplinary by nature, IMFD gathers a large number of researchers in several areas that include traditional computer science areas such as data management, Web science, algorithms and data structures, privacy and verification, information retrieval, data mining, machine learning, and knowledge representation, as well as some areas from other fields, including statistics, political science, and communication studies. IMFD currently hosts 36 researchers, seven postdoctoral fellows, and more than 100 students.

We at IMFD are convinced the development of a science of data requires producing a virtuous amalgamation of all the aforementioned areas, in order to cope with ever-increasing societal demands for taking full advantage of available information. A dramatic example of such needs has been given to humanity by the current COVID-19 pandemic, but there are clearly many others. At IMFD, we are carrying out several projects that aim to produce such interdisciplinary crossings. More importantly, we have worked to develop a common research agenda for the Institute, integrating the efforts of its members into interdisciplinary teams whose goal is solving some fundamental problems in data science. Specifically, the research agenda of IMFD has been organized into five long-term and transversal *emblematic* research projects, each of which requires input from several, if not all, the research areas mentioned previously. These emblematic projects are:

▸ *Data for the study of complex social problems.* This project seeks to encourage the combination of methodological strategies and techniques based on data analysis to develop novel diagnoses about relevant socio-political issues.

IMAGE BY ANDRIJ BORYS ASSOCIATES. USING SHUTTERSTOCK

a https://imfd.cl/en/
b http://www.iniciativamilenio.cl/en/home_en



Members of the Millennium Institute for Foundational Research on Data at a workshop held on Viña del Mar, Chile, in 2019.

▸ *Development of systems for graph and network analysis.* This project focuses on developing an efficient, correct implementation of standard languages to extract information from graph-structured data, as a way to retrieve valuable analytic information from large networks of interconnected data.

▸ *Query answering methods for emerging requirements.* This project seeks to develop theoretical foundations for information extraction tasks capable of integrating data management, data analysis, and machine learning techniques, while also making them scalable and verifiable.

▸ *Explainable artificial intelligence.* This project aims to develop new techniques that allow understanding of the inference processes of machine learning algorithms, where this understanding is embodied in the ability to provide explanations for learned patterns, as well as to create higher-level abstractions and procedures based on them.

▸ *Development of robust information structures.* This project pursues understanding of whether it is possible to create systems that allow people to have a more-accurate, less-biased view of reality, and whether the volume of data generated by users in digital platforms can be used to obtain a clearer picture of the social uses and communication phenomena of the Web.

The goal of this article is to show the achievements of IMFD along these lines of research, as well as to briefly reflect on the future of the Institute.

Some Achievements of the Institute

Data for the study of complex social problems. To meet the goal of this project, we have been developing a methodology for gathering and maintaining thick data,[c] which is data gathered by using qualitative methods, and for combining such data with large online sources to study the socio-political and economical dynamics within Chilean territory. Such thick data is usually collected from surveys, so we have focused on four territorial enclaves in contemporary Chile that concentrate a host of sociopolitical and socioeconomic challenges related to crucial dimensions of social life. Moreover, we have worked to produce prospective scenarios for public policy-making in each territorial enclave through computational social science techniques like agent-based modeling, drawing key parameters from the four territorial enclaves.

We currently are working on several lines of research in this project. In what follows, we explain one of them to exemplify the methodology mentioned earlier. In 2019, we designed and launched the "Monitor" project, which aims to produce a white paper each year on specific public policy challenges and their manifestations/implications in/for each type of territorial enclave. Our first coordinated fieldwork was conducted in the Quinteros bay area, one of the zones of environmental sacrifice in Chile. A sacrifice zone is a concept that emerged in the U.S. during Nixon's presidency as a result of the installation of coal and nuclear power plants in Utah, New Mexico, Arizona, and Colorado. Today, a sacrifice zone is a term used by different scholars to characterize specific places where the population (generally poor) coexists with industries whose activities are mainly based on coal, oil, gas, or nuclear energy. The aim of this fieldwork was twofold. On the one hand, we conducted in-depth interviews

c https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7

aimed at understanding how sociopolitical dynamics are carried out in these zones, identifying key challenges for public policies, and understanding why these challenges are not addressed. On the other hand, it served us as a way of producing more input to address some of the research questions we are working on, such as how environmental sacrifice areas are created. Faithful to our thick-data methodology, we combined this information with a large database of Chilean news about the area, in order to understand public opinion about zones of sacrifice. The results of this combination were promising; in particular, we have learned many lessons from them on how to combine thick data with online sources.

Development of systems for graph and network analysis. Graph databases are a fast-growing technology for modeling data and extracting information, in particular, because of the large number of applications where data can be naturally represented as a graph (as examples, consider social, crime detection, scientific, and bibliographic networks). This has created great interest from academia and industry in standardizing languages for extracting information from graph databases. The research background of our group, including several foundational papers in the area,[6] as well as our participation in standardization processes, gives us a comparative advantage to become key actors in the future of graph database theory and practice.

The main goal of this project is to develop languages to extract information from graph databases. In particular, we aim to develop a full-scale graph database system that encompasses state-of-the-art techniques for all aspects of data management, from data storage and indexing to concurrent access, querying, and graph analytics, and which can deal with large volumes of data. To this end, we have consolidated several lines of research, two of which will be discussed later.

Development of a standard graph query language. We actively led an effort between industry and academia to develop a standard graph database query language. As a result of this effort, we proposed the query language G-CORE,[1] which fulfills three fundamental principles for graphing such languages: to have a careful balance between expressiveness and evaluation complexity; to be composable (meaning that graphs are both the input and the output of queries), and to treat paths as first-class citizens. We are convinced that G-CORE will play a key role in shaping the future of graph query languages. In fact, the International Organization for Standardization (ISO) has launched an initiative to standardize graph query languages; two members of IMFD are participating, and G-CORE is considered one of the role models.

Creation of efficient algorithms for query answering. Joins are the costliest operation in query processing, and they have been the subject of intense research. Recent studies show that joins can be handled optimally by using appropriate data structures, which unfortunately are computationally hard to build.[2] One of the central lines of research of this project is the development of new techniques to handle joins within compact space and with provable performance guarantees. In particular, we have developed an algorithm to process join queries over a compressed representation of graphs and show that its running time is worst-case optimal,[3] and we have shown how worst-case optimal join algorithms can significantly speed up the performance of the graph query language SPARQL,[4] the standard query language for Semantic Web data. Moreover, we have explored new paradigms for answering queries over large volumes of data; in particular, we have studied the problems of enumerating, uniformly generating, and counting the answers to a query, proposing a simple yet general unifying framework to investigate these fundamental algorithmic problems, in particular for the case of graph databases.[5]

Query answering methods for emerging requirements. As the requirements for data management systems continue to evolve at an accelerating rate, the diversity of proposed solutions addressing these requirements likewise continues to grow.

IMFD aims to be a reference center in Latin America related to state-of-the-art research on the foundational problems with data, as well as its applications to tackling diverse issues ranging from scientific challenges to complex social problems.

We carried out the first empirical study of public opinion on false news in Chile. This work outlines who spreads, and how they spread, false news in the country.

While technology is rapidly advancing in sub-areas such as databases, machine learning, data analytics, information retrieval, privacy, and so on, the techniques developed are becoming increasingly specialized and divergent from each other. This project aims to help unify those currently disparate techniques by looking into several specific directions.

The ultimate goal of our project is to provide theoretical grounds for the next generation of database systems, trying to make them more flexible, scalable, secure, and robust. We have started, however, by pursuing some objectives that are more at-hand, and that can be seen as the first steps toward our more ambitious goals. These first steps have been achieved by bringing together researchers with different expertise in areas relevant to the project and making them work in specific projects that contribute toward a particular unexplored aspect of data management and its relationship with neighboring fields. Among others, we have started working on developing formal methods for verifying the semantics of query languages in systems like Coq;[20] building and studying languages for expressing analytical queries, like languages that combine features from relational and linear algebra;[7] characterizing the expressive and computational power of modern neural network architectures, including Transformers and Neural GPUs,[13] as well as Graph Neural Networks;[8] formalizing the notion of in-database classification of entities by developing and studying a framework that classifies entities based on features defined as relational database queries;[9] and, finally, building efficient algorithms for evaluating complex analytical queries over streams of data.[10,11,12]

An important stepping stone toward the full realization of the objectives of this project was the organization of an international workshop, called Emerging Challenges in Databases and AI Research (DBAI), in Santa Cruz, Chile, during November 2019.[d] The main objective of the workshop was to bring together a number of different communities to work on several of the problems mentioned here but from different angles, and to discuss ways in which all this work could be assembled toward the construction of more robust data management systems.

Explainable artificial intelligence. This project focuses its efforts on the development of new techniques that allow understanding of the inference processes behind some artificial intelligence (AI) algorithms, where this understanding is embodied in the ability to provide an explanation for such processes. Since the launching of our Institute, we have consolidated our progress into three main lines of research:

- *Visual Query Answering*, by developing an AI system capable of applying natural language to explain the reasoning behind the answer to a visual question.[14] This has been enhanced recently through the incorporation of a common-sense knowledge base, with promising results.[15]
- *Visualization*, by exploring the intersection between information visualization and explainable AI techniques; in particular, the effect of different types of explanations on the reliability of AI systems, both in recommender systems[16] and in document classification systems. Our work in visualization is receiving special attention for medical applications.
- *Social media*, by focusing on the extraction of information from social media, with the aim of providing explanations of when and why polarizing and controversial effects occur. In the work of this project, early detection of users' stances, harassment detection, early fake news detection, and the study of polarization dynamics in social media occupy central roles.[17,18]

Development of robust information structures. This project focuses on two prominent information disorders; that is, problems that work against the development of robust information structures. These disorders are the spread of misinformation (for example, rumors, conspiracies, hoaxes, and unverified news) and fabricated content (for example, "fake news," propaganda) on online social networks, and hate speech and incivility on digital media.

The spread of misinformation. Informed by social scientific theories,

d http://dbai2019.imfd.cl/

computational methods, surveys, and quantitative content analysis techniques, this research group continued examining the problem of the spread of incorrect information on social media platforms with several projects that bring together computer scientists, communication scholars, and political scientists. We carried out the first empirical study of public opinion on false news in Chile.[19] This work outlines who and how they spread false news in the country, among other issues related to disinformation. Also, via a project funded by The Social Science Research Council (SSRC) on fake news on Facebook during elections, we have started studying elections in three countries in Latin America: Chile, Colombia, and Mexico. This project aims to characterize the scope and diffusion of fake news in Spanish-speaking countries with respect to verified news, and to delineate the threat of digital disinformation in the region. Finally, we launched a project that takes a social scientific approach to the study of exposure, beliefs, and sharing of conspiracies and fake news in Chile during the social unrest that started in October 2019. Using longitudinal public opinion surveys, this project covered the social uprising and protests to study the sociodemographic, psychological, and media orientations that predict exposure to false information on social media.

Promoting healthy civic conversations on online social networks. Through an interdisciplinary perspective, we are studying how conversations take place on social media. Research into the area of incivility and social networks is organized around three projects. The first one analyzes incivility in commentaries posted by users on news media websites. The second project aims to train a classifier to tag comments and classify them as civil/uncivil, within a certain margin of error. The third project studies the relationship between the use of political memes and the presence of incivility in Chilean Twitter accounts.

The Future

We mentioned in the introduction the current COVID-19 pandemic as a dramatic example of the role that data is taking in contemporary society. Besides the classical tasks such as development of algorithms and tools to analyze data, deep debates have arisen on fundamental questions about data, such as its public availability; ownership; auditability of the processes used to collect, curate, integrate, analyze, and publish data; balance between transparency and privacy, and the roles that states, universities, research institutions, private organizations, and the general public should play in such issues.

We are convinced IMFD must play a key role in these discussions at the Latin American regional level. In particular, scientific discoveries and technological advances at IMFD must be placed at the service of the development of an integrated Chilean data infrastructure and governance. Hence, in the coming years, we will work on such development, paying special attention to the following five issues: to improve the ways in which data is collected by the Chilean government; to provide unified, integrated access to the data collected and the information developed from it, which should include data curation, but at the same time be auditable; to define different degrees of access to such information, taking into account the tension between transparency and privacy, as well as the different local uses of data; to improve the ways in which such information is analyzed, considering that such processes should be transparent and auditable; and to make this infrastructure one of the first places where the algorithms and techniques developed in IMFD are applied.

The virtual world, and its most basic support, data, came to the forefront with the COVID-19 crisis. This new world is the goal of IMFD research in the immediate future. We are dedicated to helping build the global data governance system (that is, the technical and legal infrastructure), and the social, political, and ethical practices to be observed in the virtual world. Our Institute will follow closely how this new reality is transforming the material practices of areas such as health and education, human life research, digital automation of work, and environmental research. We will focus on the development of areas that make the virtual world of data a contribution to improve people's lives.

Please join us on this ambitious project, IMFD is an open environment for collaboration!

Acknowledgments

We thank Claudio Gutiérrez and Sergio Toro for their many useful comments on this document. C

References

1. Angles, R., et al. G-CORE: A core for future graph query languages. In *Proceedings of SIGMOD Conf.*, 2018, 1421–1432.
2. Hung Q. Ngo, H.Q., Ré, C., and Rudra, A. Skew strikes back: new developments in the theory of join algorithms. *SIGMOD Rec. 42*, 4 (2013), 5–16.
3. Navarro, G., Reutter, J.L., and Rojas-Ledesma, J. Optimal joins using compact data structures. ICDT, 2020, 21:1–21:21.
4. Hogan, A., Riveros, C., Rojas, C., and Soto, A. A worst-case optimal join algorithm for SPARQL. *ISWC 1* (2019), 258–275.
5. Arenas, M., Croquevielle, L.A., Jayaram, R., and Riveros, C. Efficient logspace classes for enumeration, counting, and uniform generation. ACM PODS, 2019, 59–73.
6. Angles, R., et al. Foundations of Modern Query Languages for Graph Databases, *ACM Comput. Surv.* 50(5): 68:1-68:40 (2017).
7. Barceló, P., Higuera, N., Pérez, J., and Subercaseaux, B. On the expressiveness of LARA: A unified language for linear and relational algebra. ICDT, 2020, 6:1–6:20.
8. Barceló, P. et al. The logical expressiveness of graph neural networks. ICLR, 2020.
9. Barceló, P., Baumgartner, A., Dalmau, V., and Kimelfeld, B. Regularizing conjunctive features for classification. ACM PODS, 2019, 2–16.
10. Grez, A., Riveros, C., and Ugarte, M. A formal framework for complex event processing. ICDT, 2019, 5:1–5:18.
11. Grez, A., and Riveros, C. Towards streaming evaluation of queries with correlation in complex event processing. ICDT, 2020, 14:1–14:17.
12. Grez, A., Riveros, C., Ugarte, M., and Vansummeren, S. On the expressiveness of languages for complex event recognition. ICDT, 2020, 15:1–15:17.
13. Pérez, J., Marinkovic, J., and Barceló, P. On the Turing completeness of modern neural network architectures. ICLR, 2019.
14. Zhang, Y., Niebles, J.C., and Soto, A. Interpretable visual question answering by visual grounding from attention supervision mining. In *Proceedings of IEEE Winter Conf. Applications of Computer Vision* (2019).
15. Lobel, H., Vidal, R., and Soto, A. CompactNets: Compact hierarchical compositional networks for visual recognition. *Comput. Vis. Image Underst. 191*, 102841 (2020).
16. Dominguez, V., Messina, P., Donoso-Guzmán, I., and Parra, D. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of Intern. Conf. Intelligent User Interfaces* (2019).
17. Bugueño, M., and Mendoza, M. Applying self-attention for stance classification. CIARP, 2019, 51–61.
18. Bugueño, M., and Mendoza, M. Learning to detect online harassment on Twitter with the transformer. In *Proceedings of PKDD/ECML Workshops* (2019), 298–306.
19. Valenzuela, S., Halpern, D., Katz, J.E., and Miranda, J.P. The paradox of participation versus misinformation: social media, political engagement, and the spread of misinformation. *Digital Journalism 7*, 6 (2019), 802–823.
20. Diaz, T., Olmedo, F, and Taner, E. A mechanized formalization of GraphQL. CPP, 2020, 201–214.

Marcelo Arenas is a professor at the Universidad Católica and the Director of IMFD in Santiago, Chile.

Pablo Barceló is a professor at the Universidad Católica and the Deputy Director of IMFD in Santiago, Chile.