Research Statement
Frederic Sala

Despite recent mass adoption and significant research efforts, core aspects of machine learning remain poorly understood. Practitioners fill in these gaps via trial-and-error or folk wisdom. However, for ML to continue its growth and be safely and widely deployed across domains—such as medicine and law—with large societal impact, a deeper theoretical understanding of core ideas in modern machine learning methods is required.

My research focuses on providing a fundamental understanding of data-driven systems. I derive theoretical results (tradeoffs and limits) and use these insights to extend and improve practical systems. I study core problems for such systems—including how to best represent data via embeddings, how to create data via weak supervision for machine learning, and how to efficiently store and reconstruct data. Some of the highlights of my research include:

- Embeddings are a core building block in ML pipelines. I discovered fundamental theoretical tradeoffs and proposed new algorithms for hyperbolic embeddings of data \[13\]: representations of given data that live in hyperbolic space. My work also extended non-Euclidean embeddings to a flexible class of spaces of mixed curvature \[4\]. These insights advanced our understanding of how non-Euclidean geometry can help better preserve the structure of data and led to state-of-the-art performance in graph neural network models \[3\].

- Obtaining labeled training data for machine learning models has become a major challenge. To tackle it, practitioners have increasingly turned to weaker forms of supervision. However, these weak sources are difficult to handle when they are correlated or have more complex underlying structure. I helped make theoretical advances that address these fundamental challenges in weak supervision, opening up new possibilities for these systems. My work includes efficient algorithms for synthesizing labels from weak supervision sources with theoretical guarantees \[7\], an approach to learn the structure of a model of such sources \[15\] that is up to 30× faster than previous techniques \[1\], and an efficient approach to handle hierarchical sources, extending weak supervision to handle large-scale video and time-series applications \[14\].

- I developed an algorithm for efficient data synchronization with theoretical guarantees that can outperform industry standard tools like rsync \[9\] in terms of bandwidth efficiency. Such tools are increasingly important, e.g., for federated learning, cloud storage, or in networks. My research also introduced tight bounds on the information required for reconstructing data from partial observations \[11\]. I developed hardware-inspired error-correcting codes for modern memories \[8\], robustified popular algorithms to hardware errors \[12\], and introduced a theoretical framework to evaluate broad ranges of error-correction techniques \[10\].

The common paradigm for my work is to study core challenges in emerging areas (how to best preserve the structure of data used by ML techniques, how to rapidly create labeled training data for ML pipelines, how to efficiently synchronize and store data), obtain a model for these problems, and to extract the key theoretical principles that govern the behavior of all solutions. This approach helps build our understanding and the resulting insights can be used to guide users, power new algorithms, and scale up systems and extend them to new settings.

This approach has a track record of success: my work has been published in the top venues in machine learning and information theory (NeurIPS, ICML, ICLR, AAAI, IEEE Transactions on Information Theory, ISIT), receiving a long talk at ICML \[13\] and a best-of award at SELSE \[10\]. It has also led to fruitful partnerships with industry; my and my collaborators’ contributions to the Snorkel system for weak supervision \[6\] are in use at organizations including Google, Intel, Ant Financial, and several others institutions.

Non-Euclidean Representations in Machine Learning

My research approach seeks out modern machine learning techniques that are widely used and impactful in practice, but are poorly understood from a theoretical point of view. I develop a better theoretical understanding and rely on it to enable extensions and use in new settings.
Hyperbolic Embeddings Modern ML methods require continuous representations while much of our data (words, graphs, relational data) is discrete. Embedding methods map discrete objects to continuous spaces and have thus become a core component in ML. For example, pretrained word embeddings such as word2vec and BERT have had an enorous impact on the space of ML techniques, but users still struggle to apply these strategies outside the core settings, or to understand when they are likely to be useful.

Embeddings seek to preserve the structure of data. Hierarchical data (e.g., Wiki categories, lexical databases like WordNet, biological relationships), represented by trees, cannot be embedded faithfully into Euclidean space, the workhorse space for ML. Instead, a non-Euclidean space called hyperbolic space fits these trees very well, even with only a few dimensions. This observation introduced a flurry of techniques to help embed tree-like data, but many practical questions were left unanswered: how can we choose hyperparameters like the number of dimensions and the numerical precision level for the embedding? Which trees benefit the most from hyperbolic embeddings? How well do graphs that aren’t strictly trees embed into hyperbolic space?

My work studies these questions and extracts fundamental results. For example, I derived a tradeoff between the embedding quality, the properties of the tree (depth, maximum degree), the dimension, and the level of numerical precision. These tradeoffs directly guide practitioners, who can, for example, plug in their graph properties, desired quality level, and system precision to obtain a suggested dimension. I established techniques to perform embeddings of any graph with distortion guarantees, showed how to recover the points in an embedding exactly from distances when an embedding exists, and implemented new algorithms to efficiently produce such embeddings.

Product Space Embeddings While hyperbolic spaces are a natural fit for hierarchical data, other forms of data have different types of structure and thus fit other non-Euclidean spaces; for example, a network of cities with flight times as edges fits spherical space. This led us to ask how to obtain a class of spaces that is flexible enough to fit many types of data while also being tractable to work with for ML applications. To address this challenge, I introduced a class of product manifolds made up of simple components—copies of hyperbolic, spherical, and Euclidean spaces—which offers a rich space to embed a variety of types of data. When used in concert with an algorithm we developed to take input data and estimate how many spaces and which kinds to use in the product, our approach offers better representations and downstream performance.

The insights into non-Euclidean representations developed in my work have had an impact on downstream machine learning models. For example, my collaborators introduced a hyperbolic version of the popular graph convolutional network model. They found that certain benchmark datasets, such as the PUBMED citation graph, have a hyperbolic structure that benefits from non-Euclidean models. This led to state-of-the-art performance for node classification (i.e., predicting the academic area of a paper in citation networks). More recently, we saw state-of-the-art performance for knowledge graph embeddings by exploiting symmetries in hyperbolic space.

Understanding and Extending Weak Supervision

While the quality of representations is critical to the ML pipeline, modern supervised ML also requires large quantities of labeled data. Obtaining unlabeled data is easy. However, hand-labeling large datasets is slow, expensive, and static, so that only the largest and wealthiest organizations can generally afford to curate these datasets.

To tackle the challenge of acquiring labeled data, practitioners use frameworks to assemble multiple sources of weak supervision. The Snorkel weak supervision framework that I have contributed to has had a significant impact, with over 30 organizations using it, including industry, government, and academic groups. For example, it has been used to fight human trafficking in the DARPA MEMEX program, as part of multiple collaborations with the Stanford Department of Radiology, and is used by the core ads team for a large search company.

In order to generate labeled data for training, Snorkel models and combines the weak supervision sources, leading to important theory and systems questions, including how to make label generation efficient and how to deal with complex structure in the sources and data. My research studied these questions, developed a better theoretical understanding, and leveraged these insights to propose algorithms that made weak supervision practical in new settings: multitask labels, data with complex dependencies, and high-dimensional scenarios like video.

Fast Multi-Task Label Generation A key challenge is to efficiently learn a model of the weak supervision sources, i.e., to learn the accuracies representing the expertise of different sources and the dependencies among them. Here, efficiency is both in terms of the number of samples and wall-clock time. Learning this model is particularly challenging when we wish to produce multi-task labels. My work leveraged recent advancements in random matrix

\footnote{For example, in graph embeddings, a pair of embedded nodes should be at the same distance in the continuous space as the shortest path between them in the graph.}
theory to produce a new algorithm for label generation that is much faster than previous approaches and has broader applicability. Its theoretical guarantees enable us to quantify the generalization performance of the multitask models we trained with our generated labels—and how different it is from a model trained with the true labels.

**Learning Dependency Structures**  Weak supervision approaches must model the dependencies (i.e., correlations) between the weak supervision sources, which may be complex. Ignoring these dependencies leads to significant performance degradation downstream. I applied a powerful statistical tool, robust PCA, [15] to learn the dependency structure directly from data with faster performance and better scaling than the existing approaches, enabling users to work with sources with far more complex structure.

**Extending to Complex Temporal Data**  While complex dependencies can arise among sources, the data itself may have a complicated structure. For a film, for example, we may wish to produce labels for frames, shots, scenes, or even the entire film. I extended the Snorkel framework to work with such multi-resolution data and introduced a new algorithm that uses a principled form of parameter sharing among the model of the sources [14]. These extensions allowed us to cheaply produce large datasets for activity and medical video models.

**Data Reconstruction and Storage**

Another focus of my research is on systems that either store data, and thus must be robust to errors, or reconstruct data from partial or noisy observations. The key challenge in such systems is efficiency: *how much redundant data must be stored or transmitted in order to tolerate a certain quantity of noise?* I study the limits of such systems and develop algorithms that approach these limits.

**Efficient Synchronization and Reconstruction**  Information is increasingly used and stored in a distributed fashion, e.g., data is stored locally and in the cloud, gradients are computed on edge devices and combined centrally in federated learning. This necessitates the repeated synchronization of, for example, files or model parameters. Sending an entire piece of data to a server with an out-of-date copy is expensive and wasteful. Observing that the differences between data copies can be modeled as insertion/deletion errors, I proposed a nearly-optimal synchronization algorithm that uses short codes to detect and transmit the changes between the parties performing synchronization. This approach can empirically outperform rsync, a popular tool for synchronization, by an order of magnitude in terms of bandwidth efficiency [9]. It can also be extended to apply to more complex networks.

Closely related is the problem of data reconstruction. In this setting, multiple copies of a single piece of data are observed, but all of these have been corrupted in a different way. A key question to ask is *for a given level of corruption, how many observations are necessary to ensure we can reconstruct the original data with no errors?* My work answers this question in the case of insertions, corruptions where arbitrary symbols are added to the data (and we do not know where). In particular, I refined prior work by determining the effect of the level of redundancy in the original data. My result fully characterizes the data reconstruction rates for insertion errors [11].

**Data Representations for Modern Hardware**  Representations play an important role in data storage. The traditional approach to representing data is via a vector of bits, but modern memories are not well-suited to this choice. For example, Flash cells store charge; if bits are determined by thresholding the charge level, over time, errors occur asymmetrically, since charge tends to leak out, only affecting the higher level. Instead, my work models these errors, the limits of techniques to handle them, and introduces new representations to deal with such problems. For example, when storing information in a permutation of the levels in multiple cells, charge leakage typically does not cause errors. My work generalizes this idea and builds additional error-correction on top [8] and studies how robust such representations are when even the hardware used to decode them is unreliable [12].

Finally, storage is also critical to ML systems. Most error correction strategies are orthogonal to the use of data, seeking only to ensure what is read is what was written. However, models are often robust to certain errors while highly sensitive to others—suggesting that for dedicated storage, error correction should be designed to minimize the impact on downstream outputs rather than inputs. My work introduced such strategies for linear models [5].

**Research Agenda**

A deeper theoretical understanding of core problems in machine learning will be one of the driving forces in making ML more useful, more reliable, and easier to apply. My goal is to lead a machine learning research group that studies fundamental problems in machine learning pipelines used in practical systems, reveals the theoretical limits and
tradeoffs of techniques, and leverages the resulting insights to improve and extend these systems. Some of the work I plan to pursue includes:

**Understanding and Improving Pretrained Embeddings**

Pretrained embeddings like BERT have revolutionized natural language processing by allowing users to fine-tune pretrained models while achieving excellent performance on a wide variety of problems. Such embeddings are trained on a number of related tasks and use large textual corpuses. While this approach has yielded significant gains in text-based tasks, we do not fully understand these results. My group will study these approaches, develop a theoretical understanding, then use it to build *multimodal pretrained embeddings* produced through large-scale multitask learning. Such embeddings have the potential to radically simplify training models for one or more modalities (text, image, video, graphs), opening up many applications to everyday ML users.

A major challenge is to produce sufficient labeled training data, since not all modalities offer large unsupervised corpora. We will use weak supervision approaches for generating large amounts of labeled training data. My work shows how to efficiently produce labeled datasets aimed at training a multitask model for a small set of tasks with a fixed logical relationship [7]. We will investigate how to extend this approach to large-scale multi-task models.

**Automating Non-Euclidean Representations**

A major difficulty to using non-Euclidean representations is that current models have to be modified to operate on points living in, e.g., hyperbolic space [2]. While it is possible to modify a model top-to-bottom to do so, it is tedious to repeatedly perform manual transformations. We will address this challenge by building a framework for *non-Euclidean learning that overloads operations* involved in popular models with their analogs in other geometric spaces. This framework enables the reuse of popular models with other geometries by writing only a few lines of code.

Another challenge is that users must currently either apply their intuition to predict which space is appropriate, or use heuristic approaches to sample data and predict a good space allocation, as we proposed in [4]. However, these heuristics are not guaranteed to work. My group will study procedures with theoretical guarantees to obtain the optimal geometry from data. Moreover, we will explore how to extend these ideas from only looking at the input data to the overall procedure of the algorithm to ensure we are obtaining appropriate representations for the task at hand. An efficient implementation of these steps will automatically learn and implement the right representation under-the-hood, improving performance for a variety of models with almost no work on the part of end users.

**Building Dynamic Knowledge Spaces**

Knowledge bases and knowledge graphs are used as the backend for popular ML systems, e.g., for question answering, giving them an enormous reach. Current knowledge bases are typically stored and accessed by a fixed schema and are updated by their owners at fixed intervals.

Instead, my research group will build a dynamic knowledge base that contains both structured information and a variety of pre-trained representations for its entities. This approach has several advantages. First, for users seeking to build models on top of the knowledge base or graph, it does not require a custom embedding step, since representations can be directly imported—and these representations are useful for a wide variety of tasks. Second, valuable signal is available in the downstream trained representations—but these are typically not used to update knowledge bases. That is, information flows only in one direction. Instead, we will learn from users by dynamically updating the representations. This symbiotic approach is the key to building more universal representations for knowledge.

Building such a knowledge space requires advances in several directions. Our knowledge space will be large, requiring distributed storage and operations for accessing and updating it. We will need improvements in distributed storage and optimization techniques. The locality of the information must inform the storage algorithm. Since the parameters storing the representations are frequently updated and may live in varying geometric spaces, we will also study how to efficiently perform non-Euclidean distributed optimization, a new area for optimization. In particular, we will develop theoretical results that reveal the fundamental limits of these approaches. If successful, this system has the potential to dramatically unify and improve the pretrained models available to ML users.

---

2 Non-Euclidean spaces are not vector spaces, so linear operations common to neural networks, like matrix-vector multiplies, must be replaced with appropriate analogs.
References


