Learning-based compressed data structures*

Paolo Ferragina
paolo.ferragina@unipi.it

Giorgio Vinciguerra
giorgio.vinciguerra@phd.unipi.it

Department of Computer Science, University of Pisa

The deluge of data in applications such as bioinformatics, information retrieval, and databases has made the use of compressed data structures, sometimes called succinct or compact data structures, indispensable [4]. The specialty of these structures is to represent data in compressed form while also supporting efficient access and queries on it. Traditionally, this is achieved by exploiting two particular sources of compressibility: statistical properties and repetitiveness of the data [5].

In this poster, we will discuss a different kind of data regularity based on geometric considerations: *approximate linearity*. Several recent results have shown that this geometric regularity is surprisingly frequent in real datasets, such as time series. We will expand the horizon of compressed data structures by presenting solutions that discover, or "learn", in a principled algorithmic way, these approximate linearities in the data to solve some fundamental and ubiquitous problems in computer science, such as predecessor search and rank/select primitives [1, 2, 3]. We will provide a walkthrough of these new theoretical achievements, also with a focus on open-source libraries and their experimental improvements in space of orders of magnitude compared to classical solutions (such as B-trees). We conclude by discussing the plethora of research opportunities that these new learning-based approaches to data structure design open up.

References

[1] Antonio Boffa, Paolo Ferragina, and Giorgio Vinciguerra. A "learned" approach to quicken and compress rank/select dictionaries. In *Proc. SIAM Symposium on Algorithm Engineering and Experiments (ALENEX)*, 2021.

[2] Paolo Ferragina, Fabrizio Lillo, and Giorgio Vinciguerra. Why are learned indexes so effective? In *Proc. 37th International Conference on Machine Learning (ICML)*, 2020.

[3] Paolo Ferragina and Giorgio Vinciguerra. The PGM-index: a fully-dynamic compressed learned index with provable worst-case bounds. *PVLDB*, 13(8):1162–1175, 2020. `https://pgm.di.unipi.it`.

[4] Gonzalo Navarro. *Compact data structures: a practical approach.* Cambridge University Press, 2016.

[5] Gonzalo Navarro. Indexing highly repetitive string collections, part I: Repetitiveness measures. *ACM Computing Surveys*, 54(2), 2020.

*Poster abstract submitted to the Stanford Compression Workshop 2021.