The ever growing need to efficiently store, retrieve and analyze massive datasets, originated by very different sources, is currently made more complex by the different requirements posed by users and applications. Such a new level of complexity cannot be handled properly by current data structures for Big Data problems.

To successfully meet these challenges, we propose a new generation of “Multicriteria Data Structures and Algorithms” that originate from some recent and preliminary results of the proponents. The “multicriteria” feature refers to the fact that we seamlessly integrate, via a “principled” optimization approach, modern compressed data structures with new, revolutionary, data structures “learned” from the input data by using proper machine-learning tools. The goal of the optimization is to select, among a family of properly designed data structures, the one that “best fits” the multiple constraints imposed by its context of use, thus eventually “dominating” the multitude of trade-offs currently offered by known solutions.

In this project, we will lay down the theoretical and algorithmic-engineering foundations of this novel research area, which has the potential of supporting innovative data-analysis tools and data-intensive applications.

Participants

Unit 1 - Università di Pisa

Paolo Ferragina (PI) paolo.ferragina@unipi.it
Davide Bacciu davide.bacciu@unipi.it
Antonio Carta antonio.carta@di.unipi.it
Andrea Valenti andrea.valenti@phd.unipi.it
Giorgio Vinciguerra (teleconferencing) giorgio.vinciguerra@phd.unipi.it
Francesco Tosoni francitosoni@gmail.com
Gemma Martini  
martini.gemma3@gmail.com  
Antonio Boffa  
a.boffa95@gmail.com  

Unit 2 - Università degli Studi di Palermo  
Raffaele Giancarlo (PI, teleconferencing)  
raffaele.giancarlo@unipa.it  
Domenico Amato  
domenico.amato01@unipa.it  
Mariella Bonomo  
mariella.bonomo@unipa.it  
Giosuè Lo Bosco  
giosue.lobosco@unipa.it  
Simona Ester Rombo  
simonaester.rombo@unipa.it  

Unit 3 - Università degli Studi del Piemonte Orientale "Amedeo Avogadro"-Vercelli  
Giovanni Manzini (PI)  
giovanni.manzini@uniupo.it  
Lavinia Egidi  
lavinia.egidi@uniupo.it  

Unit 4 - Università degli Studi di Milano  
Marco Frasca (PI)  
marco.frasca@unimi.it  

Program  
Thursday, February 6, 2020  

9:00  Welcome (10 min)  

9:10  Revisiting sorted array search procedures via machine learning [Task T1]  
Giosuè Lo Bosco (50 min)  

10:00 Multicriteria approaches for succinct complex networks representations in precision medicine [Tasks T4]  
Simona Ester Rombo (50 min)  

10:50  Coffee break (30 min)
11:20 Neural networks compression techniques for succinct classification models
[Tasks T2, T4]
Marco Frasca (50 min)

12:10 Prefix free parsing with applications [Tasks T4]
Giovanni Manzini (50 min)

13:00 Lunch at Pizzeria il Montino

15:00 On dynamisation of learned indexes & their theoretical ground [Tasks T1, T3]
Paolo Ferragina (50 min)

15:50 Wrap-up, discussion on next events

20:00 Dinner at Osteria di Culegna

Friday, February 7, 2020

9:00 Session on open problems (4 hours)

13:00 Closing

Main tasks of the project
[T1] Classic Data Structures vs Purely Learned Indexes.
[T2] Compressed ML models.

Some notes on the meeting
After the session on open problems, the group decided to collaborate as much as possible among all/most Units over the following issues:

- Multi-criteria PGM-index using binary-search with no branch, regression line, O’Rourke with PGM structure.
- PPM versus deep NN in predicting a sequence, studying the accuracy as the model size increases, and possibly adopting pruning techniques both on the NN and on the PPM trie.
- Study the efficacy of pruning and clustering techniques over NN weights.