Ministero dell'Università e della Ricerca

Segretariato Generale
Direzione generale per il coordinamento e la valorizzazione della ricerca e dei suoi risultati

Ufficio III

Relazione Scientifica Intermedia - Prima Annualità
PRIN 2017 - protocollo: 2017WR7SHH

Principal Investigator

FERRAGINA Paolo
(cognome) **(nome)**

Università di PISA
(Università/Ente)

Risultati conseguiti

AMBITO DI VALUTAZIONE	RISPOSTA (spuntare)	DESCRIZIONE (max 3.000 caratteri spazi inclusi)
1) Personale appositamente da reclutare (titolare di contratti a tempo determinato, assegni di ricerca, borse di dottorato). **Sono stati stipulati contratti dal Gruppo di Ricerca?** ***Specificare, per ogni contratto, la data di attivazione, la tipologia di contratto e la durata. Segnalare altresì eventuali rescissioni o interruzioni, evidenziando le relative motivazioni.***	SI	Il Dott. Manuel Striani ha preso servizio il giorno 1/9/2020 con un assegno di ricerca finanziato dal progetto: durata 1 anno estendibile a 2 – Unità del Piemonte Orientale Contratto di collab. occasionale di 3 mesi del Dott. Giosuè Marino Cataldo presso l'unità UniMI dal 10/7/20, totalmente finanziata dal progetto Borsa di dottorato (primo anno su tre) della Dott.ssa Jessica Gliozzo presso l'unità UniMI, prenderà servizio il 2/11/20 nell'ambito del

		programma Collaborative Doctoral Partnership fra il dottorato in Informatica UniMI e il Joint Research Center di Ispra-Milano
2) Attrezzature, strumentazioni e software di nuovo acquisto. **Sono stati effettuati acquisti in tale ambito da parte del Gruppo di Ricerca?** ***Specificare la tipologia di bene acquistato e il relativo uso nell'ambito del progetto.***	SI	3 Workstation Dell 3630 Tower 1 Workstation Fujitsu 2 MacBook Pro 13' 2 stampanti HP LaserJet Asus Zenbook UX533FTC Attrezzature usate dalle unità di Palermo, del Piemonte Orientale e di Milano per lo sviluppo e la sperimentazione di prototipi software di "learned data structures", classificazione con modelli succinti e compressione. Le stampanti hanno un ruolo ausiliario.
3.1) Attività di divulgazione dei risultati. **E' stata sviluppata tale attività in ambito di convegni, seminari, ecc.?** ***Specificare per ogni partecipazione a convegno: il titolo del convegno, data, luogo, il titolo della ricerca presentata e il nome dello speaker. Se tale attività non è stata svolta, illustrarne la motivazione.***	SI	SEMINARI SU INVITO Italian Conference on Theoretical Computer Science (ICTCS), 14 settembre 2020, on-line. "The future of data structures: data-aware and self-designing". Speaker: P. Ferragina IEEE MELECON Conference, June 16-18, 2020, online. Round table on innovative startups and entrepreneurs: "Services of Big Data Analytics and Artificial Intelligence for Precision Medicine". Speaker: S. E. Rombo. Dipartimento di Scienze Della Vita, UniMore, 26 novembre 2019, Modena. "Life sciences and algorithmic design: speed and accuracy in small space". Speaker: R. Giancarlo. Dipartimento di Informatica, Università di Roma "La Sapienza", 29 ottobre 2019, Roma. "Algorithms, Theoretical Computer Science and Epigenomics: Mining Mathematical Laws for Predicting Chromatin Organization and Nucleosome Occupancy in Eukaryotic Genomes". Speaker: R. Giancarlo. Italian Bioinformatics Conference, 26-28 giugno 2019, Palermo. "Combinatorial Messages and Epigenomics: The case of Chromatin Organization in Eukaryotic Genomes". Speaker: R. Giancarlo. INNS Big Data and Deep Learning Conference, 18-19 aprile 2019, Sestri Levante. "Hybrid data structures and beyond". Speaker: P. Ferragina Spotify, 2 aprile 2019, London. "The evolution of searching data structures". Speaker: P. Ferragina.

		PRESENTAZIONI A CONVEGNI 46th Intl. Conference on Very Large Data Bases (VLDB), 31 agosto - 4 settembre 2020, online. "The PGM-index: a fully-dynamic compressed learned index with provable worst-case bounds". Speaker: G. Vinciguerra. 37th Intl. Conference on Machine Learning (ICML), 12-18 luglio 2020, online. "Why are learned indexes so effective?". Speaker: G. Vinciguerra. 1st Intl. Workshop on Artificial Intelligence for Health held in conjunction with the 18th Intl. Conference of the Italian Association for Artificial Intelligence (AIxIA 2019), 19-22 novembre 2019. "Deep Neural Networks' Architectural Issues and Data Representation Paradigms for Classification of DNA Sequences". Speaker: G. Lo Bosco 26th Intl. Symposium on String Processing and Information Retrieval (SPIRE), Segovia, 7-9 ottobre 2019. "Space-Efficient Merging of Succinct de Bruijn Graphs". Speaker: G. Manzini. Analysing Big Omics Data Workshop, 26 giugno 2019, Palermo. "A Spark Algorithmic Paradigm For Spaced Words Alignment-Free Classification, with Focus on Phylogeny". Speaker: R. Giancarlo Italian Bioinformatics Conference, 26-28 giugno 2019, Palermo. "Tripartite graph clustering for the prediction of lncRNA-disease associations". Speaker: S. E. Rombo Annual Symposium on Combinatorial Pattern Matching (CPM), 18-20 giugno 2019, Pisa. "A New Class of Searchable and Provably Highly Compressible String Transformations". Speaker: G. Manzini. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 14-18 settembre 2020, "Incremental Training of a Recurrent Neural Network Exploiting a Multi-Scale Dynamic Memory". Speaker: A. Carta.
3.2) Attività di divulgazione dei risultati. E' stata sviluppata tale attività nell'ambito della pubblicazione di lavori su riviste? *Specificare per ogni pubblicazione peer-reviewed su rivista: gli autori, il titolo del lavoro, il nome della rivista, il volume, l'anno della pubblicazione, il codice DOI e il tipo di open-access. Se tale attività non è stata svolta, illustrarne la motivazione.*	SI	R. Giancarlo, G. Manzini, A. Restivo, G. Rosone, M. Sciortino: The Alternating BWT: An algorithmic perspective. Theor. Comput. Sci., 812: 230–243 (2020). DOI 10.1016/j.tcs.2019.11.002 Green Open access: postprint su pagina personale

		L. Egidi, G. Manzini: Lightweight merging of compressed indices based on BWT variants. Theor. Comput. Sci., 812: 214–229 (2020). DOI: doi.org/10.1016/j.tcs.2019.11.001 Green Open access: postprint su pagina personale A. Kuhnle, T. Mun, C. Boucher, T. Gagie, B. Langmead, G. Manzini: Efficient Construction of a Complete Index for Pan-Genomics Read Alignment. Journal of Computational Biology, 27(4), 500–513 (2020). DOI: 10.1089/cmb.2019.0309 Open access P. Ferragina, G. Vinciguerra: The PGM-index: a fully-dynamic compressed learned index with provable worst-case bounds. PVLDB, 13(8), 1162–1175 (2020). DOI: 10.14778/3389133.3389135 Open access via pagina del progetto P. Ferragina, G. Vinciguerra: Chapter on "Learned Data Structures". In: Studies in Computational Intelligence, vol. 896, Springer, pages 5-41, 2020. DOI: 10.1007/978-3-030-43883-8_2 Green Open access: preprint su pagina del progetto

Relazione tecnica

Breve descrizione delle attività svolte da ciascuna unità di ricerca, nel periodo di riferimento.
Evidenziare, inoltre, con riferimento all'intero Gruppo di Ricerca:
a) se ci sono stati cambiamenti (aggiunte/eliminazioni o spostamenti temporali) rispetto al previsto, illustrando le principali motivazioni;
b) quale sia il reale progresso verso gli obiettivi previsti, indicando, altresì, gli eventuali risultati ottenuti;
c) come i risultati già ottenuti verranno sfruttati nell'ambito delle attività in corso di svolgimento, o se sia possibile prevederne uno sfruttamento diretto (brevetti, immissione di prodotti sul mercato, ecc);
d) se sono sopraggiunte particolari difficoltà che mettano a rischio il conseguimento degli obiettivi minimi previsti.

Una descrizione dettagliata delle attività scientifiche, divulgative e dei software sviluppati è presente nella pagina del progetto learned.di.unipi.it, sez. "Events".

Il progetto ha visto lo svolgimento di 2 eventi: un Kick-off in forma virtuale il 14.10.2019, e un meeting svoltosi in

presenza nei giorni 6-7.2.2020 presso UniPI. Il programma e le slide degli interventi sono disponibili sul sito su indicato.

L'attività dell'unità di Pisa ha riguardato il progetto e la realizzazione di strutture dati learned, compresse, dinamiche e multicriteria. Sono state sviluppate due librerie software open-source disponibili, con relativa documentazione e dataset, all'indirizzo pgm.di.unipi.it e pubblicate sulle piattaforme GitHub e The Python Package Index. Sperimentalmente, tali librerie hanno dimostrato miglioramenti di diversi ordini di grandezza su strutture dati allo stato dell'arte. Questi risultati sono stati pubblicati su una rivista, presentati ad alcune conferenze e hanno originato un capitolo di un libro Springer.

L'attività dell'unità di Palermo ha riguardato il progetto e la realizzazione di piattaforme software per Learned Binary Search; e la classificazione, compressione (multi-choice o basata su BWT) e analisi di BigData genomici via rappresentazioni succinte su Hadoop e Spark. Il software è disponibile sul sito del progetto e versioni preliminari di lavori su tali attività sono su arXiv. Questi risultati sono stati pubblicati su una rivista, presentati ad alcune conferenze, altri sono in corso di pubblicazione.

L'attività dell'unità del Piemonte Orientale è consistita nel progetto e realizzazione di metodi di compressione basati su grammatiche in vista di un loro utilizzo per comprimere le matrici di grandi dimensioni usate in Machine Learning. Sono in corso studi in collaborazione con le altre unità su varianti di compressori e indici (multi-criteria) basati su BWT e sulle capacità predittive di modelli learned rispetto a tradizionali modelli statistici usati in Data Compression. I risultati sono stati pubblicati su tre riviste e presentati a quattro conferenze.

L'attività dell'unità di Milano riguarda lo sviluppo di metodi di compressione per reti neurali profonde e di tecniche ad hoc per la loro rappresentazione in memoria. Parte di questo lavoro costituisce la base per lo sviluppo di classificatori multicriteria basati su reti neurali. I risultati dell'attività di UniMI sono stati sottomessi a conferenza internazionale.

RISPOSTE:

a) Nessun cambiamento rilevante rispetto al previsto; il lockdown ha però ridotto la collaborazione tra sedi e l'attività di diffusione dei risultati e il networking. Si segnala che l'estensione della posizione RTD-A presso UniMI del Dott. Frasca, inizialmente prevista alla sua scadenza e a valere sui fondi del progetto, non è più necessaria in quanto lo stesso ha vinto una posizione come RTD-B. La cifra così liberata sarà reinvestita per acquisizione di personale non strutturato.

b) Le attività del progetto hanno avuto un concreto sviluppo come dimostrato dai risultati ottenuti. Si segnalano inoltre una serie di nuove collaborazioni internazionali sui temi del progetto che ne dimostrano anche il suo impatto: con il gruppo di ricerca del Prof Idreos Stratos (Harvard University); con ricercatori della Dalhousie University, il Kyushu Institute of Technology, l'Università del Cile, l'Università della Florida e la Johns Hopkins University.

c) I risultati ottenuti costituiranno il punto di partenza di nuove ricerche, per avanzamenti scientifici e ulteriori sperimentazioni. Non è ancora possibile prevedere un loro sfruttamento diretto, ma i prototipi sviluppati fino a ora sono di sicuro interesse sia commerciale sia accademico, e possono diventare "prodotti" se saranno disponibili ulteriori finanziamenti e collaborazioni con aziende. Queste opportunità saranno investigate nei prossimi mesi.

d) Malgrado il lockdown si conta di riuscire a realizzare tutti gli obiettivi del progetto.