Big Data for Disease Prevention and Precision Medicine

Marco Moscatelli[1], Matteo Gnocchi[1], Andrea Manconi[1], and Luciano Milanesi[1]

[1] Institute for Biomedical Technologies – National Research Council (CNR-ITB)

Corresponding Author:

Marco Moscatelli[1]

Via Fratelli Cervi, 93 - 20090 Segrate (MI) -Italy

Email address: marco.moscatelli@itb.cnr.i

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2186v1 | CC BY 4.0 Open Access | rec: 29 Jun 2016, publ: 29 Jun 2016

Motivation

Nowadays, advances in technology has arisen in a huge amount of data in both biomedical research and healthcare systems. This growing amount of data gives rise to the need for new research methods and analysis techniques. Analysis of these data offers new opportunities to define novel diagnostic processes. Therefore, a greater integration between healthcare and biomedical data is essential to devise novel predictive models in the field of biomedical diagnosis. In this context, the digitalization of clinical exams and medical records is becoming essential to collect heterogeneous information. Analysis of these data by means of big data technologies will allow a more in depth understanding of the mechanisms leading to diseases, and contextually it will facilitate the development of novel diagnostics and personalized therapeutics. The recent application of big data technologies in the medical fields will offer new opportunities to integrate enormous amount of medical and clinical information from population studies. Therefore, it is essential to devise new strategies aimed at storing and accessing the data in a standardized way. Moreover, it is important to provide suitable methods to manage these heterogeneous data.

Methods

In this work, we present a new information technology infrastructure devised to efficiently manage huge amounts of heterogeneous data for disease prevention and precision medicine. A test set based on data produced by a clinical and diagnostic laboratory has been built to set up the infrastructure.

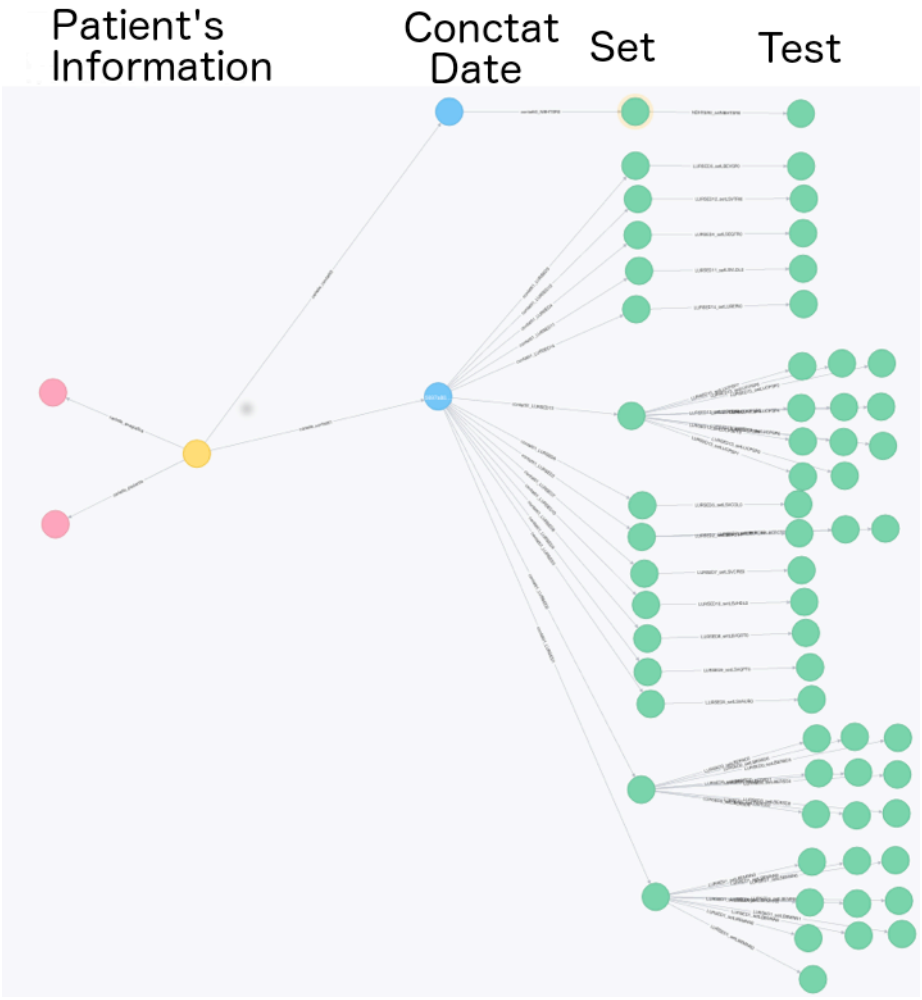
When working with clinical data is essential to ensure the confidentiality of sensitive patient data. Therefore, the set up phase has been carried out using "anonymous data". To this end, specific techniques have been adopted with the aim to ensure a high level of privacy in the correlation of the medical records with important secondary information (e.g., date of birth, place of residence).

It should be noted that the rigidity of relational databases does not lend to the nature of these data. In our opinion, better results can be obtained using non-relational (NoSQL) databases. Starting from these considerations, the infrastructure has been developed on a NoSQL database with the aim to combine scalability and flexibility performances. In particular, MongoDB [1] has been used as it fits better to manage different types of data on large scale. In doing so, the infrastructure is able to provide an optimized management of huge amounts of heterogeneous data, while ensuring high speed of analysis.

Results

The presented infrastructure exploits big data technologies in order to overcome the limitations of relational databases when working with large and heterogeneous data. The infrastructure implements a set of interface procedures aimed at preparing the metadata for importing data in a NOSQL DB. Moreover, data can also be represented as a graph using Neo4j [2]; The Neo4J DB allows you to emphasize and enhance the connections between the data and facilitate the retrieve and navigation of data (Fig 1).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2186v1 | CC BY 4.0 Open Access | rec: 29 Jun 2016, publ: 29 Jun 2016



Experimental tests on huge amount of data show that our infrastructure exhibits performances in terms of speed and scalability unachievable with relational databases. These performances are mainly related to ability of the infrastructure to index any type of field as well as to customize the queries. In particular, the high flexibility to customize the queries increases the search performance and specificity of the results.

As for future work, we planned to implement new functions and operators to perform specialized statistics analysis on big data.

References

[1] http://www.mongodb.org

[2] http://neo4j.com/

Acknowledgement

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2186v1 | CC BY 4.0 Open Access | rec: 29 Jun 2016, publ: 29 Jun 2016

This work has been supported by the “Fondazione Bracco”, the Italian Ministry of Education and Research Flagship (PB05) “InterOmics” and the European “MIMOMICS” (305280) projects.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2186v1 | CC BY 4.0 Open Access | rec: 29 Jun 2016, publ: 29 Jun 2016