Rheinisch Westflische Technische Hochschule Aachen
Informatik 5 (Information Systems)

Master Thesis Proposal

Semantic Data Profiling in Data Lake for Cancer Genome

Jasim Waheed Ansari

Supervisor: Prof. Dr. Stefan Decker
Advisors: Dr. Oya Deniz Beyan, Dr. Michael Cochez, Naila Karim

Aachen, den July 4, 2017

Contents

Chapter 1

Introduction

Cancer is one of the leading causes of death around the globe. It is considered to be the most complex disease that human species has to deal with. It is characterized by undesirable growth in cells which are caused by changes in genomes and exposure to environment[5]. Each cancer type has unique architecture of genetic variation - somatic mutations, copy number alterations, gene expression profiles and different epigenetic aberration. Therefore, a strong need for better diagnosis, personalized medicines for patients have arose, which correlates with better understanding of changes at genetic level in tumors. All thanks to technological advancements in high-throughput techniques for genome-wide interrogations such as Microarray and Next-Generation Sequecing(NGS) platforms have provided wealth of multi-disciplinary, multi-institutional pilot projects in cancer studies[11] [7].

Various unique projects catering to cancer genome study such as 1000 Genome Project[2][3], Encyclopedia of DNA Elements Project (ENCODE)[4], Immunological Genome Project (ImmGen)[10], The Cancer Genome Atlas (TCGA)[6] which aims at investigating biological systems at several levels and creating humongous datasets with different data types. These projects motivates the bioinformaticians to develop novel data integration methodologies in order to explore hidden insights. With respect to current state-of-the-art for data integration platforms, data lake promises to addresses the onslaught of Big Data generated across the projects. However, one of the biggest challenges in data lake is to prevent turning into a data swamp or silos. Moreover, especially for cancer genomic data, where the users must have a first hand knowledge in biomedical area since its metadata alone cannot provide the full meaning of respective data. There are dearth of novel methodologies that can point out,for example - genes with name ”TP53” and ”BCC7” or ”Breast Cancer” and ”Breast Tumor” means the same. Additionally, in context of conceptual hierarchy, the patient’s clinical data set has a broader term as a Person or narrower terms such as Outgoing, Incoming or Doner. Thus, the use of knowledge organization system from both biomedical domain as well as generic is needed in order to understand the data much more unambiguously. This would assist the users to find, explore, understand, and trust the data better. Such domain knowledge and thesauri are captured as specification of conceptualization(or commonly known as ontologies) in BioPortal[9], Simple Knowledge Organization Systems(SKOS)[8].

The current data lake efforts that are on data quality have the following features - profiling, cleaning and data enrichment. Keeping in consideration of our work in relation to semantic recognition of descriptive data schema, we focus on profiling aspect. But, existing profiling capabilities revolves around statistical data profiling and do not offer semantic aspect. This motivates us to an idea of enabling semantic profiling in data lakes.

Semantic Profiling is a methodology that exploits semantic-based tools and ontologies in order to derive lucid understanding of the information being stored in current systems. This approach is much more rigorous than former traditional profiling techniques for the following reasons:
i) Unlike traditional profiling which captures as much anomalies as you can until you stop, whereas, in semantic profiling once you select the domain of study, you stick by the rules and guidelines.
ii) If a user perform semantic profiling on one system and would like to combine the results of first system with those from other system, it can easily be facilitated through its feature of reusability.
iii) Semantic profiling allows setting up of testing hypothesis to monitor the data lake in production scenario which can detect semantic drift.

Therefore, through this thesis work, we extend current profiling efforts in the data lake solutions in a much more rigorous way i.e. through semantical enrichment using ontologies[1.1]. This would open the door in answering questions about data sources like:

- Does the data adhere to particular standards such as gene and disease ontologies?
- Is there a way to capture controlled vocabularies, taxonomies and thesauri for clinical dataset of patients?

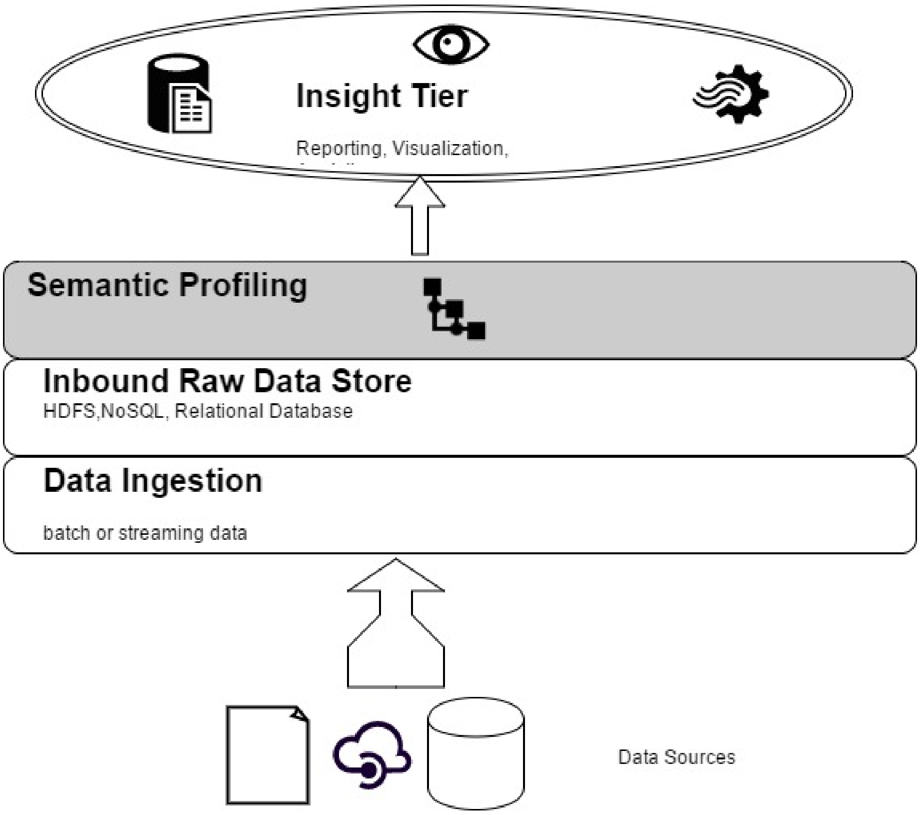

Figure 1.1: Need to extend the current semnantic tendencies of data lake

1.1 Thesis Goals & Outcome

Specific aims of my thesis are as follows:

- Extend current profiling efforts in our state-of-art data lake, which only offers statistical analysis, to semantic indicators that provides the following features:

  a) capture reports of all the bugs - different existing anomalies in the data source

  b) set of update actions applied to the data source - translation to the same concept or format

  c) semantic data structure with meta-information - in form of ontologies (for example, BioPortal, SKOS), data dictionary, regular expressions and flags to determine the type of profiling

- Develop a fully customizable dashboard with:

  a) provenance adhering to IT standards

  b) governance at entity level

Chapter 2

Background & Related work

2.1 What is Cancer Genome?

2.2 State-of-art cancer genome repositories

2.2.1 The Cancer Genome Atlas

2.3 Data Lake

2.3.1 Kylo: Data Lake solution

2.4 Knowledge Organization Systems

2.4.1 Simple Knowlege Organization Systems (SKOS)

2.4.2 Bioportal

Chapter 3

Solution

3.1 Development of semantic layer on top of Kylo, a data lake solution

3.1.1 Capturing mapping of ingested metadata to ontologies

3.1.2 Performing annotation, harmonization and canonicalization

3.2 Development of fully customizable dashboard with provenance and governance at entity level

Chapter 4

Evaluation

4.1 Distinguishing Breast Cancer Tumor Subclasses of Gene Expression Patterns with Clinical Implications

4.1.1 Background

4.1.2 Dataset in Detail

4.1.3 Identification of Tumor Subtypes by Hierarchical Clustering

4.1.4 Correlation to Clinical Results

4.2 Results & Discussion

Chapter 5

Time Plan

Bibliography

[1] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.

[2] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[3] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[4] Joseph R Ecker, Wendy A Bickmore, Inês Barroso, Jonathan K Pritchard, Yoav Gilad, and Eran Segal. Genomics: Encode explained. *Nature*, 489(7414):52–55, 2012.

[5] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, mar 2011.

[6] Ju-Seog Lee. Exploring cancer genomic data from the cancer genome atlas project. *BMB reports*, 49(11):607, 2016.

[7] Christoph Lengauer, Kenneth W Kinzler, and Bert Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643–649, 1998.

[8] Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. Skos core: simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*, pages 3–10, 2005.

[9] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl_2):W170–W173, 2009.

[10] Tal Shay and Joonsoo Kang. Immunological genome project and systems immunology. *Trends in immunology*, 34(12):602–609, 2013.

[11] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.