**Lucas Gomes Maddalena**

# On Semantically Aware Data Science:
# An Application of the Disease Ontology (DO)
# for Clustering COVID-19 Hospitalizations in Rio de Janeiro

PROJETO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
APRESENTADO AO DEPARTAMENTO DE ENGENHARIA INDUSTRIAL
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO
DO TÍTULO DE ENGENHEIRO DE PRODUÇÃO

Orientadora: Fernanda Baião

Departamento de Engenharia Industrial

Rio de Janeiro, 10 de Junho de 2022.

# ACKNOWLEDGEMENTS

I couldn’t start this acknowledge section without giving my advisor, Fernanda Baião, my most sincere appreciation for the opportunities she provided to me during the last two years and all the knowledge transferred to me.

Also, I would like to thank the staff, professors and colleagues from the Center for Healthcare Operations and Intelligence (NOIS), the EFFECT-Brazil project, the Tecgraf Institute, the Fiotec, and the Industrial Engineering Department of PUC-Rio for all the expertise and infrastructure provided during my time in the university as an undergraduate.

Thanks to my dad Armando, my mom Jucéa and my brother Pedro for making part of my family, without such, life would be much harder.

Thanks to my paternal grandma “vovó” Aida (*in memoriam)*, to my maternal grandparents “vovó” Ceinha and “vovô” Justino and to my godparents Marco and Maria Luiza for having you as part of family.

Last, but obviously not least, thanks Fernanda Maia for staying together with me during the entire time and for being a super friend, thanks to Claudia and Marcela for being my crazy friends who are together with me no matter what, it’s been almost ten years of friendship!

# ABSTRACT

Since the existence of human life, the act of storing acquired knowledge about facts and events has been establishing an important role regarding human development. However, for every individual, there is a unique perception of the universe they are living in. Therefore, artifacts made for knowledge storage and representation purposes, showed up in several different arrangements, influenced by many cultural, geographical and temporal factors, which is completely surmised.

However, on the 21st century, the exponential growth of technology, led the world facing a myriad of information coming from multitudinous sources. Then, finding ways of storing knowledge committed to certain rules became imperious.

Given this scenario, this work presents a brief explanation on Knowledge Organization Systems (KOSs) and how they showed up during the last centuries. An instance of a KOSs class are the ontologies, which have been playing an important role on, for example, making the semantics of the real world connected to data, data in which, without such ontological commitment, could be interpreted as representations of different entities than the one it is, leading to biased analysis and inaccurate prediction on data-driven projects.

This study will, based on works showing the benefits of bringing ontologies to the scenario of Data Science, make an application of the Human Disease Ontology, so enrichment on similarity measures, between group of diseases annotated in on Human Disease Ontology (DO) will be made. The step of collecting data will be done considering the SIVEP-Gripe Data Set.

Then, an analysis will be made on how better Machine Learning Algorithms can perform the analysis is made considering semantic rather than just numerical features.

**Keywords:** Data Science, Ontologies, Disease Ontology, Clustering, COVID-19, Semantic Similarities.

# RESUMO

Desde o começo da vida humana, o ato de registrar conhecimento adquirido sobre fatos e eventos, vem desempenhando um importantíssimo papel no que se refere ao desenvolvimento da humanidade. Portanto, cada pessoa tem sua visão própria no mundo em que estão inseridos. Portanto, foram várias as maneiras como artefatos com intuito de representar e guardar conhecimento, aparecem das mais diversas maneiras, por serem influenciados por fatores culturais, geográficos e temporais, o que é totalmente esperado.

Portanto, o crescimento exponencial da tecnologia, principalmente no século XXI, levou o mundo a frente de uma enorme quantidade de informação, vinda das mais diversas fontes. Portanto, procurar novas maneias de registrar conhecimento, baseado em certas regras, tornou-se crucial.

Dado esse cenário, este trabalho fornece uma breve explicação sobre Sistemas de Organização de Conhecimento (KOSs) e como eles vêm se apresentando nos últimos séculos. Uma instância desse tipo de sistema são as ontologias, que vem desempenhando um papel importantíssimo em, por exemplo, fazendo a semântica do mundo real vir a se conectar com os dados, dados esses, que sem esse comprometimento ontológico, podem ser interpretados de diferentes maneiras, como representações de entidades que não são as reais, tornando análise sobre esses dados enviesadas e predições ruins em projetos baseados em dados.

Esse estudo então, baseando-se em outras produções científicas que explicitam os benefícios de trazer as ontologias ao mundo do Data Science, fará uma aplicação da *Human Disease Ontology*, de tal maneira que será feito o enriquecimento semântico em medidas de similaridades entre grupos de doenças da ontologia em questão. A etapa de coleta de dados será feita usando a base de dados do SIVEP-Gripe.

Assim, uma análise de como algoritmos de aprendizado de máquina podem melhorar sua performance quando considerado a semântica dos dados ao invés de apenas suas variáveis categóricas e numéricas.

**Palavras-chave:** Ciência de Dados, Ontologias, Disease Ontology, Clustering, COVID-19, Similaridades semânticas.

# LIST OF TABLES

# LIST OF FIGURES

# SUMMARY

# 1 Introduction

In December 2019, the first case of coronavirus disease (COVID-19), caused by the SARS-CoV-2 virus, was reported. It did not take long for the disease to get enormous proportions and become a worldwide concern, and on March 11th, 2020, the World Health Organization (WHO) declared the disease outbreak a global pandemic (WHO, 2020).

By April 17, 2022, there have been already more than 6 million COVID-19 fatal cases and more than 504 million confirmed cases (Our World in Data, 2022) worldwide. Thus, a sudden appearance of an overwhelming amount of data for research and analysis occurred at an unprecedented rate, coming from a myriad of public and private institutions. However, the coexistence of semantically divergent and non-explicit definitions for data from distinct countries and periods that are being integrated and analyzed make the conclusions of such analysis and the extracted knowledge potentially questionable. The problem related to data coexistence (i.e., data coming from different sources) refers to situations in which data seems to represent the same concept at first sight but, when analyzing with a semantically aware approach, may represent different concepts of the real world.

In the pandemic scenario this is not different. Since the disease is affecting the four corners of the world, data comes from a thousand-and-one different data providers. Therefore, data integration in the COVID-19 domain can be compromised and semantic commitments shall be considered when treating pandemic data. As an illustration, in China, from Jan 15 until March 2, 2020, there have been seven different versions of the COVID-19 case definition issued by the government, and Tsang et al. (2020) estimate that the lack of a temporal consensus on the definitions led China official pandemic tracking to increase up to 7.1 times (IC 95%, 4.8 – 10.9) from one definition to another.

One of the main purposes of ontologies is to make the real-world data semantics explicit (Guizzardi, 2020); consequently, many benefits can be extracted by this kind of artifact, including its use as a communication artifact among different stakeholders, as a common data model to mediate data integration and access, or even as a formal specification to enable reasoning on data. In the COVID-19 domain, several works already proposed ontologies and applications (Babcock et al., 2021)(Wan et al., 2021)(Wu et al., 2021)(Sargsyanet al., 2020).

Maddalena and Baião (2021) proposed OntoCOVID, a domain ontology well-founded in the Unified Foundational Ontology (UFO) (Guizzardi et al., 2021). OntoCovid aims on making explicit semantical divergences in the COVID-19 case definition from two important health organizations, the World Health Organization (WHO) and the European Centre of Disease Prevention and Control (ECDC).

Recently, the multiple benefits of ontologies (including foundational ontologies, conceptual models, and other semantically aware artifacts) to enhance data analysis and knowledge extraction have been increasingly advocated. In this context, Amaral et al. (2021) present how ontologies, and specifically foundational ontologies, can have multiple benefits on every step of the internal cycle of the Data Science Life Cycle, while Maas and Storey (2021) show the benefits of pairing conceptual models with ontologies.

The present work focuses on data regarding the comorbidities (i.e., diseases) of patients who have been diagnosed with COVID-19 and were hospitalized in the state of Rio de Janeiro. The main objective is to analyze the impact of a semantically aware approach when finding similar subsets of hospitalizations in the dataset.

To this end, we apply a partition-based clustering technique and compared its results in two scenarios. The first scenario (semantic unaware) represented each hospitalization as a binary vector of comorbidities and applied the conventional cosine similarity metric. The second (semantic aware) scenario was proposed as follows.

During Data Pre-Processing step we matched each comorbidity found in the dataset with a corresponding concept in the Disease Ontology (DO) (Schriml et al., 2018). A total of 161 distinct diseases were linked to DO concepts, and we observed 465 different combinations of diseases, for all the patients in the dataset.

To compute similarities between individual comorbidities, we applied the measure proposed by Wang et al. (2004), which addressed semantics to find similarities between data, and specifically proposed a similarity metric in the bio-ontologies domain using DO terms. However, since each hospitalized patient was characterized by a (possibly empty) set of comorbidities in the dataset, the similarity between distinct hospitalizations required a groupwise similarity metric, i.e., measuring the similarity between two different groups of diseases, which represents the diseases a COVID-19 hospitalized patient has. For instance, while the pairwise metric performs a comparison between two terms such as “diabetes” and “asthma”, the groupwise similarity metric compares two sets of terms, such as “Diabetes, gilbert’s syndrome and flu” and “Psoriasis and AIDS”. Therefore, we applied the metric proposed by Mabotuwana et al. (2013) for calculating groupwise similarities between sets of DO terms. Mabotuwana et al. (2013) calculates groupwise similarities between terms on the SNOMED CT[1].

Hence, the semantic aware groupwise similarity between hospitalizations proposed in our work was computed by combining the groupwise metric of Mabotuwana et al. (2013) with the pairwise similarity between DO terms of Wang et al. (2004).

The impact of the proposed semantically aware approach when finding similar subsets of hospitalizations in the dataset is assessed in the Data Post-Processing step using metrics of cluster quality. An additional analysis was performed to show how well the resulting clusters from each scenario partitioned the subsets of diseases. Figure 1 displays each methodology used on the different steps of this work.

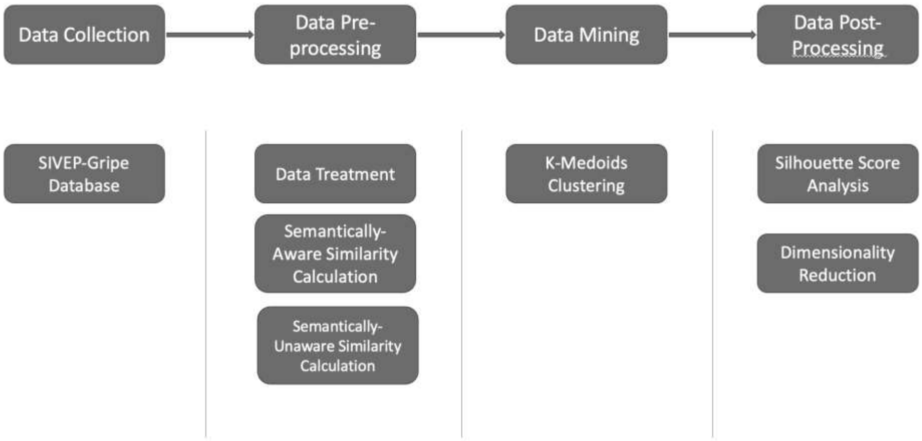


Figure 1: Resources used, and methodologies applied in each step of the study.
Source: Author

[1] SNOMED CT (The Systematized Nomenclature of Medicine Clinical Terms) is a comprehensive and multilingual clinical healthcare terminology, available at https://www.snomed.org/snomed-ct/why-snomed-ct

The remainder of this work will be organized in the following sections: Section 2 explains what Data Science and the Life Cycle of projects in this context. In section 3, semantics and ontologies are explained by firstly introducing artifacts which have been used for organizing knowledge. In section 4 more detailed explanation on how semantical awareness is imperious on the Data Science Life Cycle. Section 5 shows how each step of the Data Science Life Cycle was applied on data referring to COVID-19 hospitalizations in Rio de Janeiro and how an ontology may have been used. Findings of the research are show on Section 6. Conclusion and future works are detailed on Section 7. Finally, on Section 8, the bibliography which supported this research is listed.

# 2 Data Science and its Life Cycle

Drew Conway (2013) proposed that Data Science (DS) is the intersection of three different skills: hacking skills, mathematics and statistics knowledge and substantive expertise, this representation is represented by the Venn Diagram shown on Figure 2.

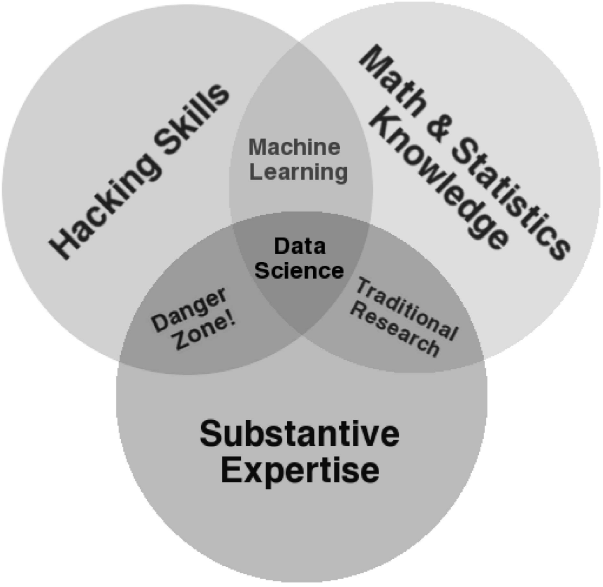


Figure 2: The Data Science Venn Diagram. Source: Conway (2013)

Math and statistics knowledge and hacking are undoubtedly critical skills for defining the success of a Data Science project. However, substantive expertise skill is often undervalued, leading to biased analysis and, thus, impacting negatively on decision-making. Therefore, this research proposes, inspired on both works of Amaral, Baião and Guizzardi (2021) and Mass et al. (2021), to make use of the benefits of conceptual modeling and ontologies on better problem understanding and other steps of a Data Science Project.

Also, Machine Learning play a fundamental role on Data Science projects, according to Taeho Jo (2021): “The machine learning (ML) is defined as the computation paradigm where the capacity for solving the given problem is built by previous examples”.

Machine Learning systems can be either supervised or unsupervised. In the first case, the algorithm learns from labeled data, also called training set, so the machine can predict the label of unlabeled data. On the other hand, in unsupervised learning data is unlabeled, and the algorithm aims to learn patterns based on data characteristics, usually the similarity between dataset instances. Example Machine Learning techniques that apply supervised learning are classification and regression, while unsupervised learning is used in clustering and association rules learning techniques (Han and Kamber, 2012).

Mass et al. (2021) cites different proposals for Data Science methodologies, including the ones by Kurfan and Musilek (2006), Shmueli and Koppius (2011), Chambers and Dinsmore (2014) and Goodfellow et al. (2014). Nevertheless, Shcherbarkov and Brebels (2014) proposes Lean Data Science Research Life Cycle, pairing the key principles of Lean Development, which came from Toyota Company to the Data Science Life Cycle and serves as the basis for our work.

The Lean Dara Science Research Lifecycle is divided in six steps, as in Figure 3: (1) Problem Understanding (Ask the right question), (2) Getting the Data, (3) Internal cycle of data science research, (4) Visualization of results, (5) Create actions based on results and (6) Feedback out of actions.

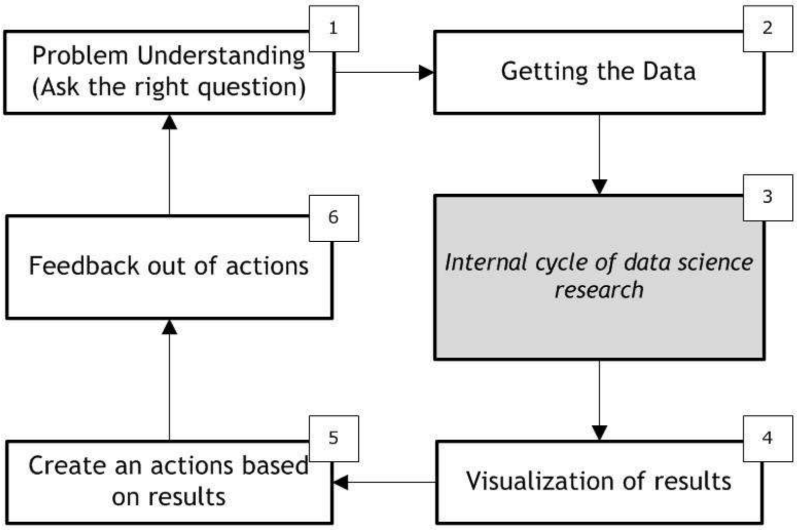

1
Problem Understanding (Ask the right question)
2
Getting the Data
3
*Internal cycle of data science research*
4
Visualization of results
5
Create an actions based on results
6
Feedback out of actions

Figure 3: Steps of lean data science research lifecycle. Source: Shcherbarkov and Brebels (2014)

1. Problem Understanding: Asking the right questions will better guide the project cycle, helps defining adequate type of data to be analyzed and computational and statistical methods further applied.

2. Getting the data: Data Science must live up to its name, therefore getting the right data is a crucial step on the project cycle. Also, the definition from Mealy G. (1967) that “data are fragments of a theory of the real world” suggests that data can be represented in several ways. Indeed, the real world is complex, and each individual has its own vision regarding reality. Hence, data must be carefully collected, considering which theory in the real world it represents. Also, data may come in many different formats, such as tables in CSV files, or as unstructured files such as photos and audio files and may be provided by different sources, such as internal data sources and external data sources Shcherbarkov et al. (2014).

3. Internal cycle of data science research: This step is mainly about data treatment, analysis, and application of Machine Learning (ML) models, which can be further divided in the three steps:

3.1 Data Pre-processing: During this step, data is analyzed so its properly manipulated and transformed; and therefore, better consumed on the Data Mining step. Such task is important because datasets in the real world is often corrupted i.e., has missing values and outliers[2] due to human errors, hardware failures, etc. There are several ways to tackle these issues, such as imputation of mean or median values, for example. Data elimination is also used for dealing with outliers and missing values. Also, data, depending on how it will be consumed during the project lifecycle, can carry unnecessary features, and leak important ones. Thus, feature selection and feature engineering consist of techniques aiming to solve these issues. In the case of unwanted features, basic elimination i.e., feature selection can be made. When facing lack of features, the second step on the lifecycle can be revisited so more data can be integrated, but also new features can be computed from the existing data, by calculating arithmetic and other statistical operations and therefore generate more useful data for the problem to solve, that would be the feature engineering case. Finally, data dimension (a.k.a. number of feature/variables) can be reduced by using techniques such as Principal Component Analysis (PCA) (Pearson, 1901), Uniform

[2] Outliers are mainly considered result of corrupted data. However, they can be useful information for fraud detection studies (Kou et al., 2004).

Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) and Multidimensional Scaling (Kruskal et al., 1964) which are also considered feature engineering/extraction techniques.

3.2 Data Mining: During this step, computational, mathematical, and statistical techniques are used, so knowledge can be obtained from the data. Supervised ML algorithms, such as classification, represented mainly by neural networks, decision trees and logistic regressions, plays an important role on discovering important data features to predict another feature. Often, when applying the mentioned algorithms, revisiting Data Pre-Processing step can occur, so data can be again manipulated and transformed, enhancing accuracy, sensibility, precision, and specificity of predictors. Also, unsupervised ML algorithms are widely applied on many DS projects, it is often represented by Clustering techniques, which is "the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such objects in a cluster are similar to one another, yet dissimilar to objects in other clusters" (Han et al., 2011). Since clustering relies on grouping data objects based on their pairwise (dis)similarities, such as Euclidian distance, Manhattan distance, revisiting Data Pre-Processing step may be needed to compute such kind of metric. Also, as data dimensionality increases, the identification of patterns become less trivial (the curse of dimensionality), therefore clustering aims to tackle this issue by resuming large amount of data into similar groups (Marmanis et al., 2009).

3.3 Data Post-Processing: Consists in interpreting patterns, information and other results obtained during Data Mining process, evaluate the quality of these results, considering the existence of biases, false correlations and maybe revisiting the previous step for retraining the selected model. In the case of clusters, the performance can be evaluated with the help of some metrics, such as Silhouette Coefficient (Rousseeuw, 1986).

4. Visualization of Results: This step can be done on the three following types: (1) Visualization of results as the initial point for decision support, (2) Visualizing the results of comparison with baselines or benchmark models and (3) Visualizing the information regarding the Data Mining process (Shcherbarkov and Brebels, 2014). Data

visualization. There are several methods of data visualization, each one for a certain purpose. For instance, pairwise dissimilarity can be represented by heatmaps, high-dimensionality data through scatter plots, with the help of dimensionality reduction techniques, and also the creativity on assigning visual artifacts so more information can be represented in a single and lean chart.

5. Creating Actions Based on Results: The results provided by the previous steps can be either considered a positive outcome or a negative outcome, on the first case, actions may be performed based on research results and, on the second case, Data Mining step may be revisited, restarting the cycle (Shcherbarkov and Brebels, 2014).

6. Getting Feedback from Action: Consists of the evaluation on how the obtained results were useful and trustful for the end-user and the decision maker. Using Key Performance Index (KPI) can be used to measure how the results impacted on the decisions made (Shcherbarkov and Brebels, 2014).

This work will focus on the first four steps of the described lifecycle, more specifically on the Internal cycle of data science research, where machine learning techniques will address domain semantics through the use of ontologies.

# 3 On Semantics and Ontologies

Knowledge Organization Systems (KOSs) are artifacts for organizing information and to explicit knowledge. There are several types of KOSs, Pieterse et al. (2014) classify these artifacts according to their complexity, as shown on Figure 4.

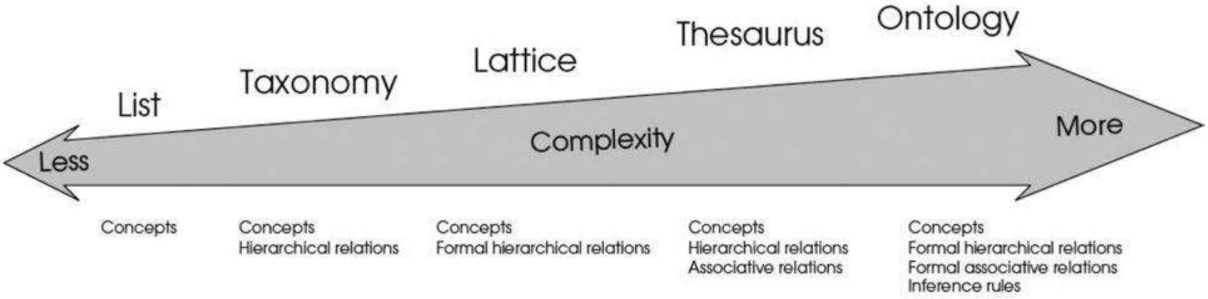


Figure 4: Classification of KOSs, Source: Pieterse et al. (2014)

According to the authors, a list is the least complex artifact to organize knowledge, it provides a linear structure of related things, together with some descriptions and/or properties owned by these things. Taxonomies are more complex structures than lists, due to the hierarchical relations representation (for example order is *subcategory-of* class), and biologists for years have been representing their knowledge about living organisms through taxonomic, such as the famous hierarchy proposed by Linné C. (1767) which described nature information with his book "*Systema naturæ per regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*", which is translated to "System of nature through the three kingdoms of nature, according to classes, orders, genera and species, with characters, differences, synonyms, places". In Pieterse et al. (2014) classification, "A lattice is a hierarchically organized collection that contains items and their attributes in which these items and their attributes are formally presented as a concept lattice". Thesaurus is a KOS that allows specification of the attributes of items as well as their equivalences, hierarchies, associations and/or contrast semantic relations between items. For example, Wordnet (Miller, 1995) (Fellbaum, 1998) is a thesaurus frequently used in several domains, and defines, for instance, that "**head** *is-part-meronym* of **human-body"** and that "**disease** is associated to **illness** through the *has-sister-term* relation". Finally, there are the so-called ontologies which, according to this classification, is the class of KOSs carrying more complex formalizations about knowledge on the domain of discourse, such as axioms, properties, associations of concepts. For example, the Human Disease Ontology (DO) specifies an axiom stating that *causes-or-contributes-to-condition* is **SubPropertyOf** *casually-related-to*. There are many ontologies being developed in several domains, and particularly in the biology domain (Bard et al., 2004).

Guarino (1997) suggests ontologies are classified in four different kinds, with respect to their level of generality, as shown on figure 5 below.

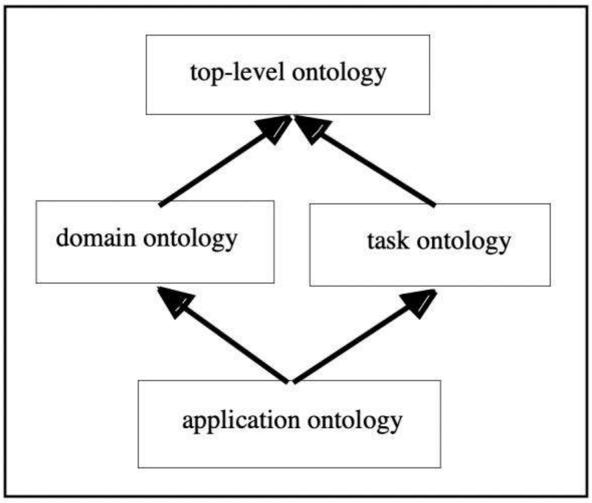


Figure 5: Abstraction levels of ontologies. Source: Guarino (1997)

According to Guarino (1997), top-level (or foundational) ontologies are the ones describing very general concepts, like space, time, matter, object, event, action etc., and the Unified Foundational Ontology (UFO) (Guizzardi, 2021) and the Basic Formal Ontology (BFO) (Arp et al., 2015) are examples of such kind of ontology. Domain and task ontologies describe, respectively, the vocabulary related to a generic domain (ex: Diseases) or a generic task (ex: Diagnosis), such descriptions are made by specializing concepts on a top-level-ontology. Lastly, application ontologies specialize the concepts of domain and task ontologies for describing conceptualizations in specific application contexts.

In this research, we make use of the Human Disease Ontology (DO), a domain ontology organized as a directed acyclic graph, representing the domain of ontologies and is mapped to uncountable others application ontologies.

DO makes the knowledge on the domain of human diseases explicit, by describing diseases through ontology properties, such as *is-a*, *has-material-basis-in* or *has-symptom*. For instance, DO states that:

**bone disease** *is-a* **connective tissue disease**
**congenital megabladder** *has-material-basis-in* **autosomal dominant inheritance**
**allergic conjunctivitis** *has-symptom* **allergic reaction**.

The Human Disease Ontology, in its last update on April 28th, 2022, comprises 17,840 classes and 45 properties (BioPortal, 2022) and is widely applied for several purposes in Academic and Industry contexts. In addition, it has been used by more than 50 other biomedical ontologies and there is a numerous list of software tools and other web resources that: (1) support the use of DO data, (2) have integrated or were built using DO data, or (iii) provide data linkages to the DO website (Disease Ontology, 2022).

# 4 On the Benefits of Semantics, Ontologies and Conceptual Modeling in the Data Science Lifecycle

Managing data cannot be accomplished solely by humans with their limited cognitive capabilities (Mass et al.,2021). Also, available data keeps growing and is becoming more important as a resource for decision-making. Thus, it is crucial to understand the domain which the data represents, to make a more precise usage from it.

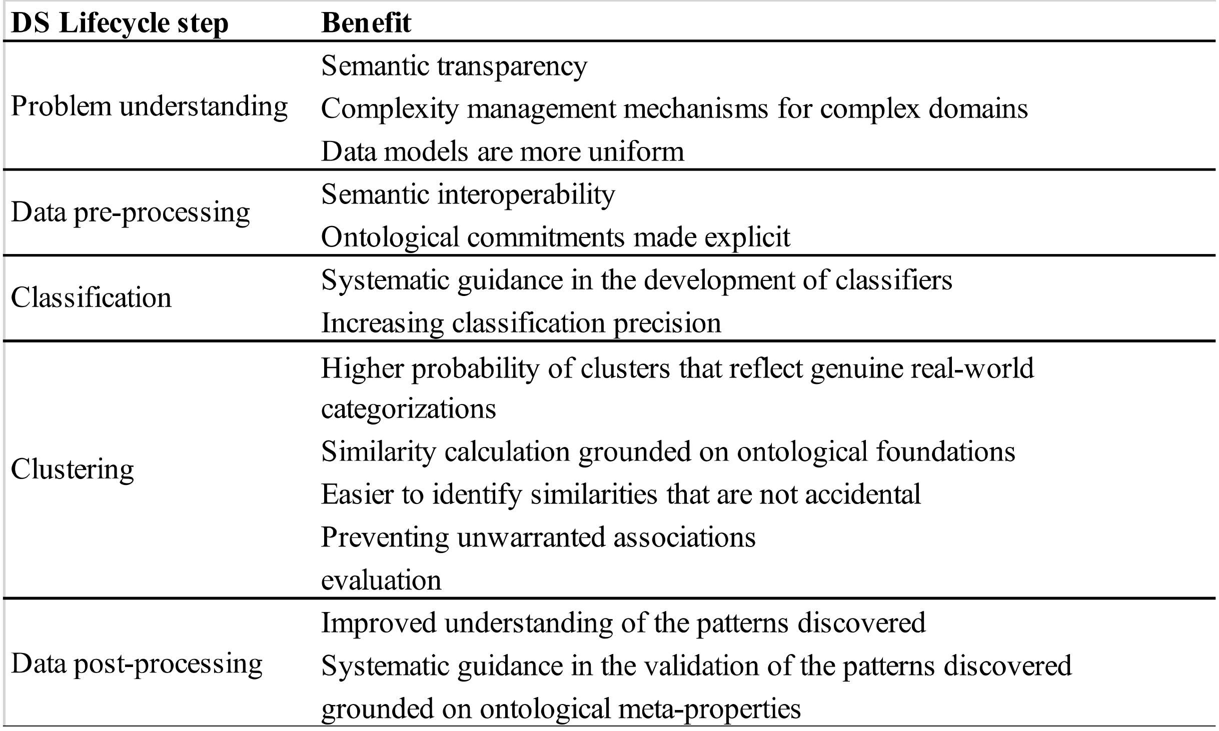
Mass et al. (2021) and Amaral et al. (2021) show that pairing conceptual modeling/ontologies artifacts with data science/machine learning techniques can not only enhance Data Science projects results but also support the development and evaluation of conceptual modelling approaches. However, this work will focus on the first mentioned kind of benefit, when semantical commitment helps on Data Science Projects.

In particular, Amaral et al. (2021) defend the benefits of using foundational and domain ontologies appears in each cycle of the Data Science Life Cycle, including Problem Understanding, Data pre- and post-processing, and Data Mining for different techniques (Classification and Clustering, for example). Such benefits are summarized on Table 1.

Problem Understanding, as aforementioned on section 2, relies on asking the right questions, so problems can be correctly solutioned. Ontologies can serve as a tool for this DS Lifecycle step. Indeed, ontologies can serve as tools by providing better understanding of the domain referring to the problem. Moreover, many methodologies of ontology engineering, such as SABiO: Systematic Approach for Building Ontologies (SABiO) (Falbo, 2014), takes as a step defining competency question i.e., natural language questions outlining and constraining the scope of knowledge represented in an ontology (Wiśniewski, 2019). Such commitment of ontologies with competency questions may lead to the right answers during the problem understanding step.

On the Data Pre-processing step, Amaral et al. (2021) defend ontologies could help on both on semantic interoperability and ontological commitment made explicit. These benefits refer to data integration which can be made not considering the ontological commitment of the sources providing the data and, therefore, joining data features which refers to different entities of the real world, leading to misinterpretations and false results on the DS project.

When clustering data, relying on foundational ontologies may lead to cluster results better reflecting real-world categorization. Moreover, calculating pairwise data similarity committed on ontological foundational can lead to similarities between data way more befitting to the domain where the treated data lays on.



| DS Lifecycle step | Benefit |
|---|---|
| Problem understanding | Semantic transparency<br>Complexity management mechanisms for complex domains<br>Data models are more uniform |
| Data pre-processing | Semantic interoperability<br>Ontological commitments made explicit |
| Classification | Systematic guidance in the development of classifiers<br>Increasing classification precision |
| Clustering | Higher probability of clusters that reflect genuine real-world categorizations<br>Similarity calculation grounded on ontological foundations<br>Easier to identify similarities that are not accidental<br>Preventing unwarranted associations<br>evaluation |
| Data post-processing | Improved understanding of the patterns discovered<br>Systematic guidance in the validation of the patterns discovered grounded on ontological meta-properties |

Table 1: Multiple Benefits of Foundational Ontologies and Domain Ontologies on Data Science. Source: Adapted from Amaral et al. (2021)

Traditional data mining methods and techniques treat data as merely “sums of attribute values”, and such approach can lead to biases and bad understanding of the patterns discovered (Amaral et al., 2021). Indeed, Clustering techniques mostly relies on calculating similarities – a data pre-processing step – which does not consider semantical attributes and are basically mathematical operations to calculate Euclidian distance and other kind of metrics. However, there have been for the past few years many proposals of considering ontologies on the calculation of object similarities, such as: Gilbert et. al (2013) and Lee et al. (2008). Also, on the biomedical field, especially for Gene Ontology (GO) (Ashburner et al., 2000; Gene Ontology Consortium, 2021), there are several similarity metrics considering many different ontologies, such as: Wang et. al (2004), Jiang and Conrath (1997), Resnik (1999) and Lin (1998). However, the metric proposed by Wang can also be extended for comparison between DO terms.

In this research scenario, ontologies will show up as a tool on the Data Pre-Processing step of de Data Science Life Cycle and, therefore, may enhance analysis results. The ontology terms (diseases) and taxonomic relations (*is_a*) will be considered when computing similarities between group of comorbidities, since each comorbidity is linked to a disease in the Disease Ontology. Similarities should be calculated following a groupwise approach, to enable a comparison between two groups of comorbidities. Pairwise similarities may be trivially computed by a simple application of a distance metric, either one of the four last mentioned metrics or any of the metrics available in HESML (Half-Edge Semantic Measures Library) (Lastra-Diaz et al., 2017).

Semantic aware groupwise metrics, however, are not that simple. According to Lastra-Diaz et al. (2017), “A groupwise semantic similarity measure is used to compute the degree of similarity between two sets of concepts defined into an ontology. This type of measure is commonly used to compare sets of GO terms in genomics, although they could also be used to compare sets of WordNet synsets evoked by two words”. Section 5 details the approach used to calculate DO terms groupwise similarities.

# 5 Semantically Aware Data Science Life Cycle: Applying the Disease Ontology on Clustering COVID-19 Hospitalizations Data

## 5.1 Associating comorbidities to diseases in the Disease Ontology

We analyzed the dataset from SIVEP-Gripe (*Sistema de Informação de Vigilância Epidemiológica da Gripe*), a nationwide surveillance database used to monitor severe acute respiratory infections in Brazil[3]. Each instance of such database represents a hospital admission due to COVID-19, characterized by several features related to the hospital where the patient was admitted, to the case evolution (Death or Recovery), to the patient vaccine administration, along with other features that were out of the scope of this work.

However, this dataset contains a lot of imprecise and missing data, mainly on data referring to the patient comorbidities, which this work aims to tackle. Hence, data selection followed a semantic aware methodology, described as follows.

Data Selection was made by filtering the first three thousand hospitalization of 2021 in the State of Rio de Janeiro. However, since this work will rely mostly on analyzing each patient set of comorbidities, the filtering also considered instances of data with noisy, inaccurate and missing information regarding this feature. Also, since this study focuses on the pairing of ontologies to the Data Science Lifecycle, rather than discovering new patterns, we did not prioritize analyzing larger datasets.

Patient comorbidities which appeared in the dataset were then mapped to the ontology. Each comorbidity on the dataset was associated with a DO disease. This step was performed manually, by searching for DO classes whose names were syntactically similar to the comorbidity name appearing in the dataset. Some of these associations can be seen on Table 2.

For example, if a hospitalization entry on SIVEP-Gripe dataset has, for instance, the word "DPOC" (short for *Doença Pulmonar Obstrutiva Crônica* in Portuguese) in *MORB_DESC*

[3] DATASUS Ministry of Health, SRAG 2020 - severe acute respiratory syndrome database - including data from COVID-19. Surveillance of severe acute respiratory syndrome (SARS). https://opendatasus.saude.gov.br/dataset/bd-srag-2020

column, we consider the patient has “Chronic Obstructive Pulmonary Disease”, which has the ID DOID:3083 in the DO.

| Name on SIVEP-Gripe Databae | DO Match |
|---|---|
| ALCOOLISMO | alcohol use disorder |
| ALZHEIMER | Alzheimer's disease |
| AMILOIDOSE | amyloidosis |
| ANEMIA | deficiency anemia |
| ANEMIA CRONICA | deficiency anemia |
| ANEURISMA DE AORTA ABDOMINAL | abdominal aortic aneurysm |
| ANOREXIA NERVOSA | anorexia nervosa |
| ANSIEDADE | generalized anxiety disorder |
| ARTRITE | arthritis |
| ARTRITE REUMATOIDE | rheumatoid arthritis |
| ARTROSE | osteoarthritis |
| ATAQUE ISQUEMICO TRANSITORIO | transient cerebral ischemia |
| AVC | cerebrovascular disease |

Table 2: Disease matching between SIVEP-Gripe names with DO terms. Source: Author

Hospitalization information regarding comorbidities of a patient is represent either by Boolean variables or by a free text field.

## 5.2 Calculating (dis)similarities between DO terms

There are several ways to calculate pairwise similarities between classes in an ontology. In this work, Wang et al. (2007) proposed to measure semantic similarity among DO terms. For computing such metric, Wang defines a term $A$ in DO as $DAG_A = (A, T_A, E_A)$, where $T_A$ is the set of all ancestors in DO graph and $E_A$ is the set of edges connecting DO terms to $A$. The S-Value of DO term $t$ related to term $A$ is defined as the contribution of $t$ to the semantics of $A$, such that, for any $t$ in $DAG_A$, its S-value related to term A is defined as:

$$S_A(t) = \begin{cases} 1, if\ t = A \\ \max\{w_e \times S_A(t') | t' \in children\ of(\mathrm{t})\}, otherwise \end{cases}$$

However, $w_e$ is a value representing the semantic contribution factor for edge $e \in E_A$ linking term t with its child $t'$, thus for every $e$, a corresponding weight $w_e$ may be predefined. Wang similarity measure for DO terms only considers *is-a* relationships, and the corresponding weight $w_e$ is preset to be 0.7.

Also, for a given term $A$, the total semantic contribution of $A$, $SV(A)$ in $DAG_A$ is computed as follows:

$$SV(A) = \sum_{t \in A} S_A(t)$$

Hence, for any pair $(A, B)$ of DO terms, $Sim_{Wang}(A, B)$ can be computed as follows:

$$Sim_{Wang}(A, B) = \frac{\sum_{t \in T_A \cup T_B}(S_A(t) + S_B(t))}{SV(A) + SV(B)}$$

For computing such metrics, the R software package DOSE (Yu et al., 2015) was used, which is part of the open-source software for bioinformatics Bioconductor. Figure 7 shows a heatmap representing pairwise similarities among some DO terms. For instance, let $A$ a vector of DO terms as follows:

$$A = (DOID{:}8398, DOID{:}409, DOID{:}2841, DOID{:}850, DOID{:}2914, DOID{:}7148, DOID{:}8857)$$

The six terms on vector $A$ represent, respectively, the diseases in the following set: (**osteoarthritis, liver disease**, **asthma**, **lung disease**, **immune system disease**, **rheumatoid arthritis, lupus erythematosus)**. We define a matrix $S$, such that the value on position $S_{A_i,A_j}$ represents the similarity $Sim_{Wang}(A_i, A_j)$, with the graphical representation on Figure 6.

Also, Figure 6 displays where in the ontology the terms on vector $A$ are placed, with respect to their relationships and hierarchies between other terms. Moreover, the relationship *has-subclass* is equivalent to *is-a* in the way that, if A *is-a* B, then B *has-subclass* A, such figure was made on Graphviz.[4]

[4] https://graphviz.org/

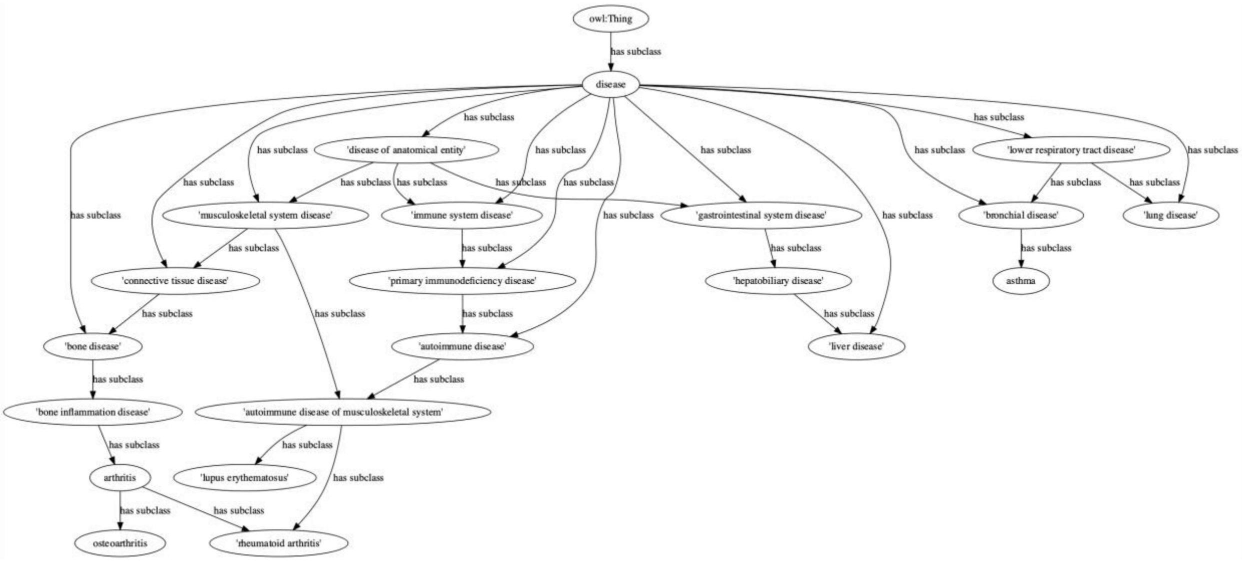

owl:Thing
has subclass
disease
'disease of anatomical entity'
'lower respiratory tract disease'
'musculoskeletal system disease'
'immune system disease'
'gastrointestinal system disease'
'bronchial disease'
'lung disease'
'connective tissue disease'
'primary immunodeficiency disease'
'hepatobiliary disease'
asthma
'bone disease'
'autoimmune disease'
'liver disease'
'bone inflammation disease'
'autoimmune disease of musculoskeletal system'
arthritis
'lupus erythematosus'
osteoarthritis
'rheumatoid arthritis'

Figure 6: Graph representing path-to-root concepts of six diseases in DO. Source: Author

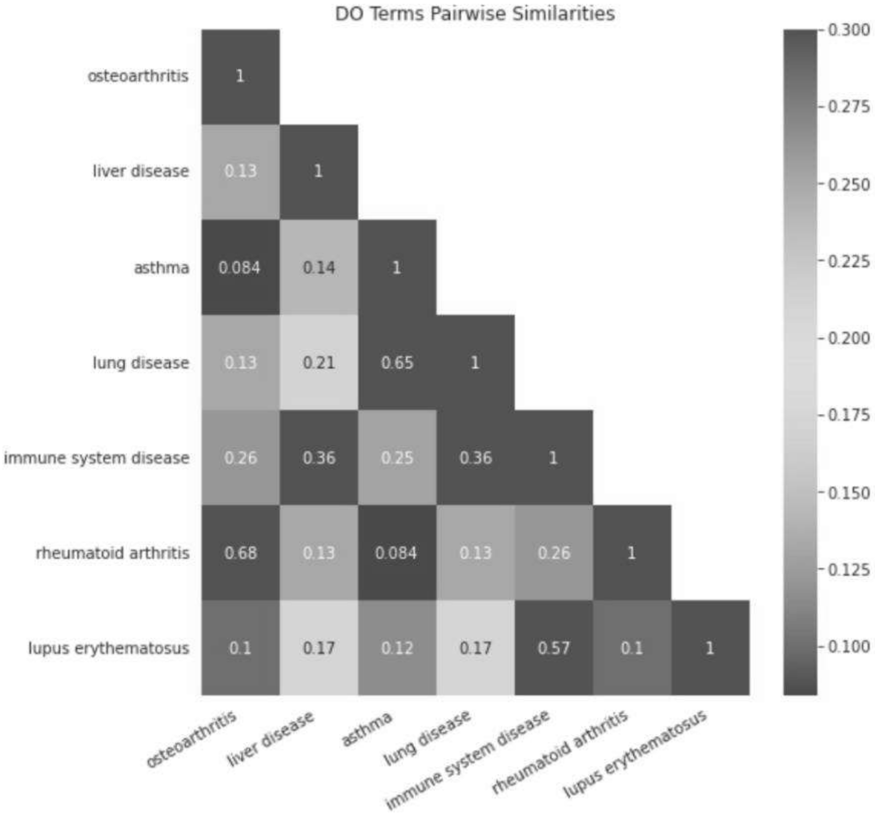


Figure 7: Pairwise similarities between DO terms. Source: Author

As can be seen on Figure 7, **rheumatoid arthritis** has a high similarity with **osteoarthritis** because both diseases have a relationship *is-a* (or, *is-subclass*) with **arthritis**. Also, since **rheumatoid arthritis** *is-a* **autoimmune disease of musculoskeletal system** together with **lupus erythematosus**, such DO terms have higher pairwise similarity when comparing **lupus erythematosus** with **osteoarthritis**.

Figure 8 illustrates a heatmap representing Wang pairwise dissimilarities between all 161 diseases.

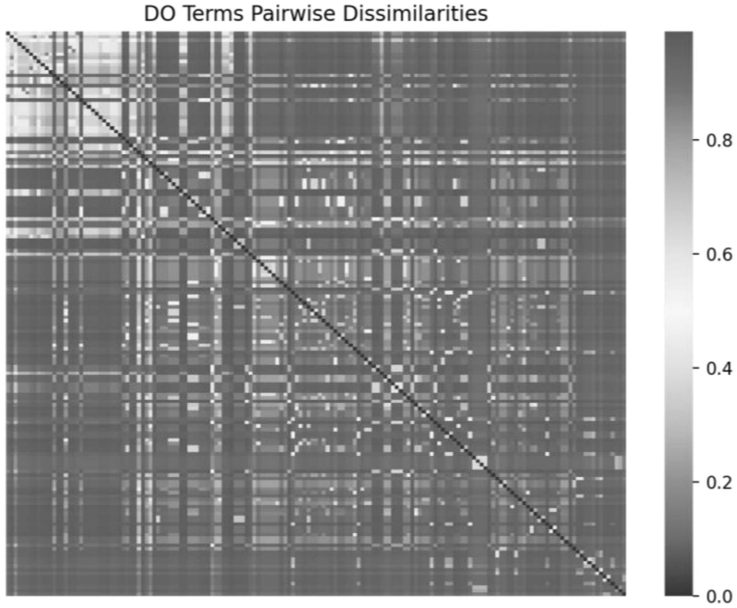


Figure 8: Heatmap Displaying Wang Pairwise Dissimilarities Between 161 DO Terms.
Source: Author

## 5.3 Calculating Groupwise (Dis)similarities

Each row in the hospitalizations dataset represents a hospital entry, which refers to a unique patient. As aforementioned, each entry contains data about the diseases a patient has. Hence, each instance on the dataset is characterized as a single group of DO terms. With the previous definitions, only pairwise similarity metrics between classes in the ontology can be computed. Then, for calculating similarities between set of diseases i.e., groupwise similarities, other approaches were required.

For instance, consider an ordered set $D$ containing $n$ terms from DO, and an example instantiation of $D$ in which $n = 4$:

$$D = \{\textbf{lupus erythemathosus}, \textbf{rheumatoid arthritis}, \textbf{liver disease}, \textbf{asthma}\}$$

Also, let $D' \subseteq D$ the subset representing the diseases a patient suffers, and an example instantiation of $D'$:

$$D' = \{\textbf{rheumatoid arthritis}, \textbf{asthma}\}.$$

Any subset of diseases in $D$ may be a represented as a document vector $v$, i.e., a $n$-dimensional binary vector, in which each coordinate represents if the concept of $D$ is in $D'$. Thus, in John's case, $v^T = (0 \quad 1 \quad 0 \quad 1)$. This representation is useful and broadly used in Natural Language Processing models and some machine learning techniques that rely on similarity measures between instances of data.

### 5.3.1 Cosine (Dis)similarity

Considering $x, y$ vectors in the n-dimensional space, cosine similarity between these vectors is represented as:

$$GSim_{cos}(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

The operation $x \cdot y$ represents the usual $\mathbb{R}^n$ inner product and $\|x\|$ represents the Euclidian magnitude of a vector $x \in \mathbb{R}^n$.

Also, this similarity metric follows the following property:

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n : 0 \leq GSim_{cos}(x, y) \leq 1$$

So, the cosine dissimilarity may be defined as:

$$GDSim_{cos}(x, y) = 1 - GSim_{cos}(x, y)$$

Even though this metric represents, at some way, groupwise disease similarities, ontologies are not considered as semantical enrichment artifacts. Therefore, according to

Amaral et al. (2021), data mining techniques relying in these metrics may lead to less genuine understanding of patterns discovered, due to the lack of semantics.

To tackle this, next section provides an ontologically well-founded (dis)similarity metric that may be considered as an extension of the original cosine similarity and is inspired on Mabotuwana et al. (2011) work, which applies the metric on the domain of radiology.

### 5.3.2 Semantically Aware Cosine (Dis)similarity

For introducing semantic similarity between document vectors, Mabotuwana et al. (2011) first define (in their words, in a loosely way) the similarity between two concepts $C1, C2$ in an ontology as:

$$Sim(C1, C2) = \frac{1}{d}$$

Where $d$ is the number of nodes in the shortest path between concept nodes (inclusive of) $C1$ and $C2$. However, the authors clarify that other similarity measures can be used, as long it preserves the basic property that increasing distance within the ontology is concomitant with a decrease in semantic similarity. Hence, the similarity measure defined by Wang et al. (2007) for DO terms will be used:

$$Sim(C1, C2) = Sim_{Wang}(C1, C2)$$

Henceforward, each term of the domain ontology brought up by the dataset, together with all the other concepts in their paths-to-root (a.k.a. seed concepts), will represent each coordinate of the document vectors which will be further analyzed. However, Wang pairwise similarity measure already represents the weight of seed concepts in its formula. Hence, in this work, only the Disease Ontology terms presented on the explored dataset will be considered, and such group of diseases will be represented as a set $C$, called context set.

Finally, with the definitions above, the DO terms groupwise similarities, $GSim_{Wang}(A, B)$, with respect to a context can now be computed. Hence, let $C = \{C_1, C_2, \dots, C_n\}$ be a set of diseases representing the context set and let two group of disease terms, namely, $A$ and B, which by definition, $A, B \subseteq C$. Then, we can the desired metric formula is represented below:

$$GSim_{Wang}(A,B) = \frac{\sum_{c \in C \cap (A \cup B)} \max_{a \in A} Sim_{Wang}(a,c) \cdot \max_{b \in B} Sim_{Wang}(b,c)}{\sqrt{\sum_{c \in C \cap A} \left(\max_{a \in A} Sim_{Wang}(a,c)\right)^2} \cdot \sqrt{\sum_{c \in C \cap b} \left(\max_{b \in A} Sim_{Wang}(b,c)\right)^2}}$$

Also, this similarity metric ranges from 0 to 1, therefore, dissimilarity is easily derived, such that:

$$GDSim_{Wang}(A,B) = 1 - GSim_{Wang}(A,B)$$

For instance, let's calculate the similarity between group of DO terms for context $C$, as in Table 3:

| | asthma | liver disease | lung disease | immune system disease | rheumatoid arthritis |
|---|---|---|---|---|---|
| $A = \{\text{asthma}, \text{liver disease}\}$ | 1 | 1 | 0.65 | 0.36 | 0.13 |
| $B = \{\text{rheumatoid arthritis}\}$ | 0.084 | 0.13 | 0.13 | 0.26 | 1 |

Table 3: Values for computing DO terms groupwise similarities

Now, similarity between groups of diseases $A$ and $B$ is calculated as follows:

$$GSim_{Wang}(A,B) = \frac{(1 \cdot 0.084) + (1 \cdot 0.13) + (0.65 \cdot 0.13) + (0.36 \cdot 0.26) + (0.13 \cdot 1)}{\sqrt{1^2 + 1^2 + 0.65^2 + 0.36^2 + 0.13^2} \cdot \sqrt{0.084^2 + 0.13^2 + 0.13^2 + 0.26^2 + 1^2}}$$

$$GSim_{Wang}(A,B) = 0.1824$$

Therefore, in this work, both groupwise dissimilarities were calculated to support on the Clustering during Data Mining Step. Figure 9 showing how smooth dissimilarity is when enriching data with semantics, while semantically unaware measures lead to false dissimilarities between data objects, which potentially may impact on further mining and analysis activities.

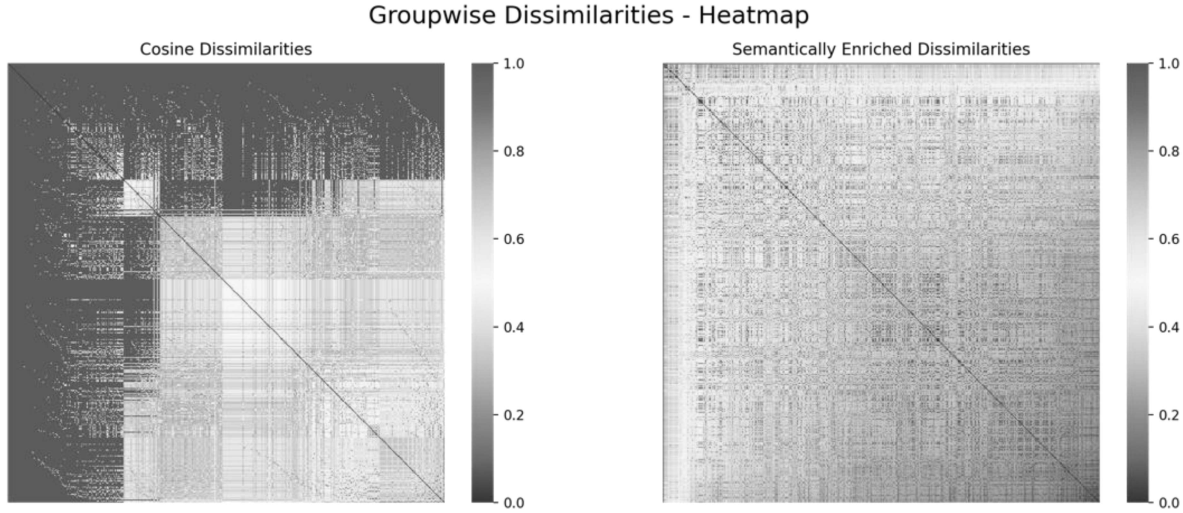


*Figure 9:* Heatmaps of groupwise dissimilarities using semantically unaware (left) and semantically aware (right) metrics. Source: Author

## 5.4 Clustering hospitalizations

Han et al. (2000) defines Clustering as the process of grouping a set of data objects into multiple groups or *clusters* so that the objects within a cluster have high similarity but are very dissimilar to objects in other clusters. Euclidian and Manhattan distance are often used as dissimilarity measure on clustering techniques. However, in this study, clustering analysis will rely on both cosine similarity and cosine similarity based on the prior mentioned Wang measure.

Clustering can be defined as an unsupervised machine learning algorithms and many approaches have been proposed for the last decades. Fahad et al. (2014) suggests that clustering methods are divided in five different types as shown in Figure 10.

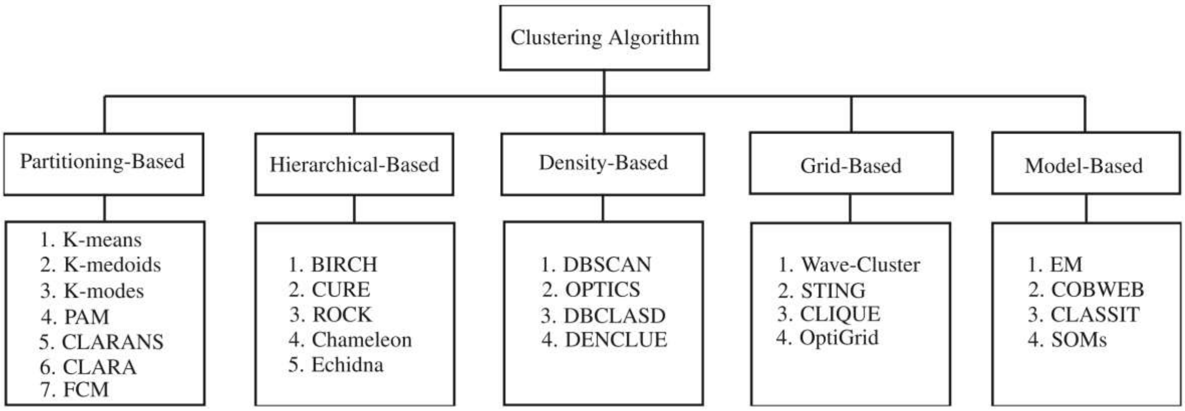


Figure 10: Types of clustering algorithms. Source: Fahad et al. (2014)

This work focuses on the use of the K-medoids clustering technique (Park et al.,2009), which is a Partitioning-Based clustering algorithm that is scalable and compatible to cluster objects upon precomputed dissimilarity metrics, which is the case of the data in this study.

Also, for choosing clustering algorithm parameters (such as the number of clusters) this work relies on the Silhouette Coefficient (Rousseeuw, 1986) as a metric which we want to maximize. Such metric, based on the intra-cluster and extra-cluster distances, provides information regarding the quality of the clusters. However, other techniques such as Calinski-Harabasz Index (Caliński, T., & Harabasz, J., 1974) and Davies-Bouldin Index (Davies, David L.; Bouldin, Donald W., 1979) can be used to evaluate performance of clusters.

K-medoids algorithm, according to Park et al. (2009), relies on three steps and in the case of this work follows the algorithm of Figure 11.

Step 0: Let $X = (X_1, X_2, \ldots, X_n)$ a vector of points in the dataset (a.k.a. sets of DO terms) and let $k (k < n)$ a pre-defined number of clusters.

Step 1: (Select initial medoids)

1-1. Calculate the distance between every pair of objects on the chosen similarity, so $d_{ij} = GDSim_{Wang}(X_i, X_j)$ when considering semantics and $d_{ij} = GDSim_{cos}(X_i, X_j)$ when using the usual cosine dissimilarity measure.

1-2. Calculate $v_j$ for object $j$ as follows:

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}}, j = 1, \dots, n$$

1-3. Sort $v_j$'s in ascending order. Select $k$ objects having the first $k$ smallest values as initial medoids.

1-4. Obtain the initial cluster result by assigning each object to the nearest medoid.

1-5. Calculate the sum of distances from all objects to their medoids.

Step 2: (Update medoids)

Find a new medoid of each cluster, which is the object minimizing the total distance to the objects in its cluster. Update the current medoid in each cluster by replacing with the new medoid.

Step 3: (Assign objects to medoids)

3-1. Calculate the sum of distances from all objects to their medoids.

3-2. Calculate the sum of distances from all objects to their medoids. If the sum is equal to the previous one, then stop the algorithm. Otherwise, go back to step 2.

Figure 11: K-Medoids Clustering Algorithm. Source: Park et al. (2009)

Notice that K-medoids predefines the number of clusters $k$ and depending on the desired result, different techniques may be applied. In this case, Silhouette Coefficient maximization is applied and further explained on next section.

For making use of such algorithm, Scikit-learn (Pedregosa et al,, 2011) implementation of K-Medoids on Python programming language (van Rossum et al., 2009) was used.

## 5.5 Clustering evaluation

To analyze clustering results, this study will rely on Silhouette Coefficient, proposed by (Rousseeuw, 1986). This metric, when taking its average, provides an evaluation of cluster validity and might be used to select an appropriate number of clusters, which is the exact approach of this study for making such selection.

When constructing silhouettes, two information are needed: the partition obtained by the clustering i.e., which cluster each data object is placed and the dissimilarities between these points. That way, let $I \in \{1,2,\dots,k\}$, where $k$ is the number of clusters and let $C_I$ be the set of the cluster $I$, let $i \in C_I$ a data points on cluster $I$ and consider

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

as the mean distance between all other possible data instances $i$ in the cluster. $|C_I|$ is defined as the cardinality of the cluster $C_I$. Also, $d(i,j)$ may be defined the distance, which is not necessarily the usual Euclidian distance. Then, consider

$$b(i) = \min_{J \neq I} \frac{1}{|C_j|} \sum_{j \in C_J} d(i,j)$$

as the minimal mean distance of $i$ to any other point $j$ on the data set. In hands of such definitions, the silhouette score for a given point $i$ in the dataset is defined by:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & if a(i) < b(i) \\ 0, & if a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & if a(i) = b(i) \end{cases}$$

Given the points $i$ of the dataset and the clusters where each datapoint is assigned, a point $i$ is well assigned to its cluster if $s(i)$ is a positive value, close to 1, as this score decreases and turns into a negative value, becoming closer to $-1$, means $i$ would better be on another cluster. Also, if $s(i)$ is on the neighborhood of zero, it the datapoint is considered indifferent regarding its assigned cluster and, would also be appropriate of staying on some neighbor cluster, implying that the cluster assigned to $i$ is overlapping another cluster.

Computing the mean of the present metric, for all points in a data set, on a Clustering result is often used as a framework on for choosing this kind of Machine Learning Algorithm parameter tuning. More specifically, the clustering algorithm is run for different number of clusters, and for each iteration, the associated average silhouette is stored. So, the number $k = k^*$ of clusters which maximizes the mentioned metric is selected as the optimal one (Rousseuw, 1997).

Scikit-learn (Pedregosa et. al, 2011) implementation of Silhouette Coefficient was again used for calculating such metric, it is important to point out the implementation makes possible using other distance/dissimilarities metrics, which was fundamental for this work.

The average silhouette coefficient was then calculated for each instance of K-medoids application, on both semantically aware and unaware dissimilarity data and for different numbers $k$ of clusters, ranging from 2 to 15. As seen on Figure 12, the optimal number of clusters $k = k^*$, which maximizes the average silhouette score was, on the semantic aware case was $k^*_{sem} = 5$ and on the semantic unaware case was $k^*_{no\ sem} = 5$, where each clustering obtained, respectively, scores of 0.277108 and 0.12143.

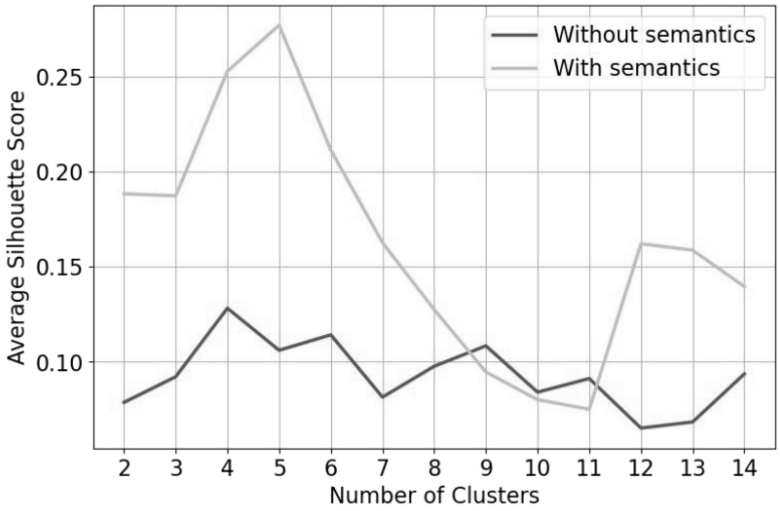


Figure 12: Comparing average silhouette score for different number of clusters. Source: Author

However, many times projects applying the DS Lifecycle, may lack on paying attention on silhouette scores of each single data point and relies on the analysis of just the average score computed above. Indeed, averages resumes several observations into to a single number. Hence, analysis on the distribution of this quality score should be made for each cluster and for each data point.

The obtained results regarding the quality of the clustering on both treated data are in fact encouraging, semantical enrichment on data benefits on the DS Life Cycle, at least on this work scenario. The bar plot displayed on figure 13 shows that not only the average silhouette is clearly higher, but the metric evaluated individually for each data point is clearly higher on the overall. Also, cluster 0 of the cosine dissimilarity clustering has mainly negative silhouette scores. Moreover, when clustering the semantically aware data, the average silhouette score was higher than 83% (386 out of 465) of the observations on the semantically unaware scenario.

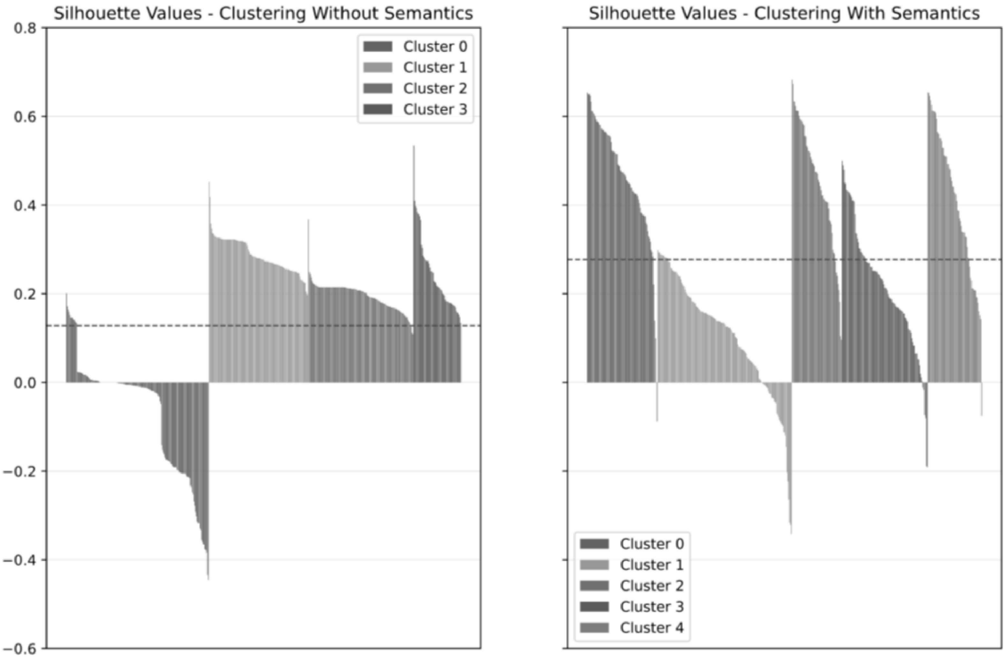


Figure 13: Bar plot with values of silhouette score for each data point. Source: Author

## 5.6 Applying Multidimensional Scaling (MDS)

Often, data comes as the results of distances or dissimilarities calculated between points. For instance, suppose a fictitious dataset containing information regarding all living persons of the world and their distance between each other. Such scenario would result, based on the United Nations (2019) estimation for worldwide population, on a dataset containing more than

$\frac{(7.7 \times 10^7)^2 - (7.7 \times 10^7)}{2} \sim 3.0 \times 10^{15}$ entries (distances matrix are symmetrical i.e., distance from A to B is the same as the distance from B to A, for every pair of points).

When considering this kind of data, distance information between points are provided. However, data regarding the position of points is suppressed. Ergo, position of arbitrary persons A and B could be, respectively, Duque de Caxias (municipality in Rio de Janeiro, Brazil) and Mendes (another municipality in Rio), at the same time that there are no constraints forbidding that A is in Manhattan Island and B is in in Stanford, which are also 50km apart. However, as data in the supposed dataset increases, the set of possible coordinates for person A gets more restricted i.e, there are less possibilities when estimating an individual position.

Such elucidation is made, because often on the Data Science Lifecycle, steps face both high dimensional data and dissimilarity matrices, which is the case of this work. In one side, this kind of data is useful, because Clustering techniques relies on (dis)similarities, but on the other dise, data dimensionality becomes quadratically higher as new instances become part of the dataset. Also, high-dimensional datasets are hard to manipulate since human sensorial capacities are only able to see the three spatial dimensions. However, more data features (i.e., dimensions) can be represented visual tools, such as shapes and colors (usually for categorical variables) or color ranges (usually for numerical features).

On the Data Science Lifecycle, and as mentioned on the second section, dimensionality reduction methods can be both applied on Data Pre- and Post-Processing (Xie et al., 2018). In the first case, such class of technique shows up as a feature extraction tool, which can help ML algorithms efficiency, by bringing data to lower dimensions, but aiming on the preservation of data global structure. Such technique can lead to other benefits during Data Post-Processing sted, the analysis of results obtained, when in lower dimensions, can lead to better understandings of the information obtained, by improving data visualization.

There are many methods for reducing features on a set of data, as already mentioned on Section 2, some examples are: (1) PCA (Pearson, 1901), which with the help of robust Linear Algebra techniques, compress the data but also minimizing information loss; (2) UMAP (McInnes et al., 2018) technique, in its framework relies on algebraic topological and Riemannian geometry – both advanced mathematical fields – to provide, according to the authors, dimensionality reduction with better performance i.e., preservation of data global structures and superior run timing; and also (3) MDS (Kruskal et al., 1964), which is used to translate “information about the pairwise ‘distance’ among a set of $n$ objects or individuals” (Mead et al., 1992) such technique could help on determining coordinates to population distance on earth population example introduced on the beginning of this section.

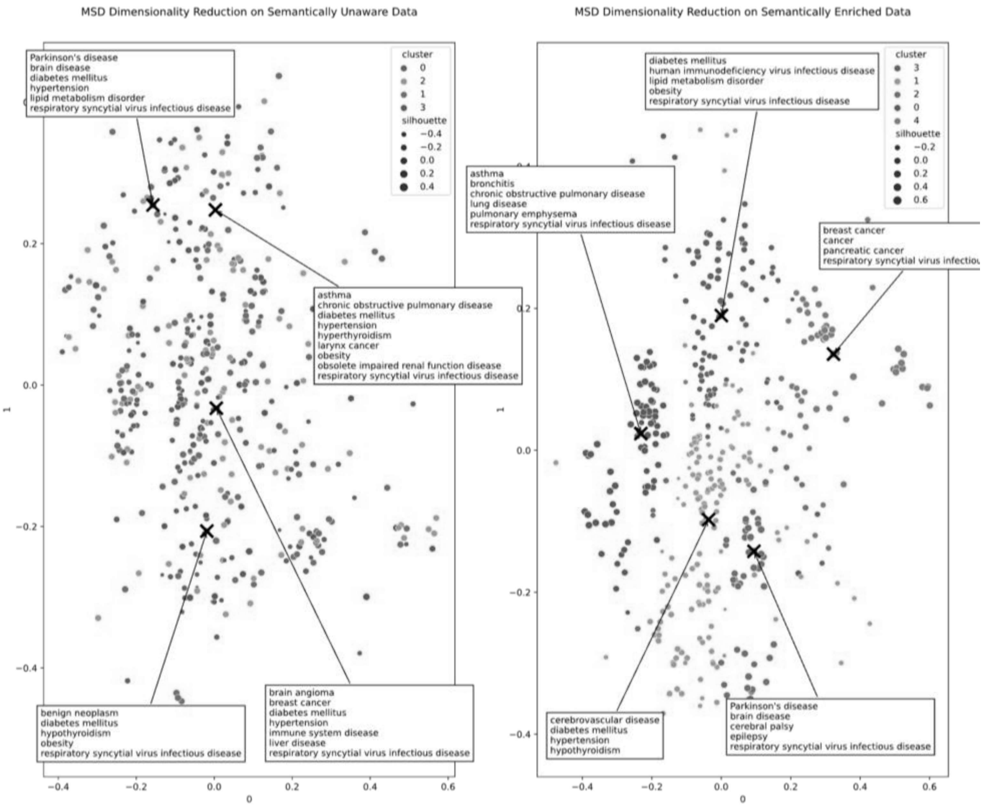


Figure 14: Approach for Graphical Representation of Difference between Clustering Techniques

In this work, MDS served well when transforming DO terms groupwise dissimilarity metrics into points in the cartesian plane, where each point represents a group of diseases on the ontology. The application of this dimensionality reduction technique helped during data post-Processing, when dissimilarity got by only two variables, instead of 461. Therefore, both charts displayed on Figure 14 were possible, it shows the transformed data in the cartesian plane. Moreover, information regarding both the clustering results and the obtained silhouettes scores were represented, respectively, by introducing different colors and radius sizes for each point. Also, for every cluster, the medoid point was represented with a black cross, where it emerged a box displaying all DO terms presented by the highest 4 silhouette scored group of diseases of each cluster.

The results shown in Figure 12 are crucial to make explicit how DS results are improved when adding semantics to data. While on the left chart clusters are overlapping (one more evidence to explain the low silhouette scores obtained), the one in the right, shows how the clusters were better separated, thus way closer to the main objective of this technique, which is to maximize intra-cluster similarities and maximize inter-cluster similarities. Lastly, our results evidenced the benefit of "Higher probability of clustering results that reflect real-world categorizations", exactly as mentioned by Amaral et al. (2021). When comparing both scenarios, the semantically unaware clusters grouped diseases which are, by common sense, dissimilar to each other; on the other hand, semantically aware clusters reflected real-world categorizations, i.e., diseases within the same cluster are clearly more similar to each other.

# 6 Conclusions

This work proposes a semantic awareness application of the Data Science Lifecycle on the COVID-19 domain and shows the benefits of considering ontologies and other semantic structures as tools for enhancing different steps of the DS Lifecycle.

Even though there are ontology terms groupwise metrics in the literature, they are not as present and accessible as the ones measuring pairwise similarities. So, in the context where groupwise distance between sets of objects are required, an adaption of the Mabotuwana et al. (2012) proposal for calculating groupwise similarities was made so Wang et al. (2007) was computed.

The level of details of all the metrics used in this work serves not only for better problem understanding, but also for making this research more committed on reproducibility, which is a fundamental tenet of science (Alston et al., 2021). Also, still on reproducibility aspect, the link to this study Source Code, can be accessed by clicking here.

The use as a comparison of semantically unaware metrics, the Cosine Similarity was used and the results on each step where both metrics were used, the benefits were clearly shown. Firstly, when calculating groupwise similarities, Cosine Similarity, as explained shown on Figure 8 led to false (dis)similarities between data objects and was pointed that could lead to bad results later, on the Data Mining step, which really occurred. Figures 10 and 11 shows how the overall silhouettes score (i.e.) on an overall are considerably higher when enriching data with semantics.

On Figure 12 aims giving the reader a visualization of the most important results in a nutshell. Which displays the overlapping clusters, that is a result of the semantically unaware similarity calculation. Also, such visual results agree with silhouette values found on the Data Pre-Processing step. On the other side of the graphic, which shows results of a semantically enriched Data Science project, intra-clusters distances are minimized, and inter-cluster similarities are maximized. Finally, a brief analysis on the quality of the grouping of diseases is made when presenting Figure 14.

Also, to make visualization of results clearer, this work relied on both Matplotlib (Hunter et al., 2007) and Seaborn (Waskom, 2021), which are environments for Data Visualization and helped me as results were better displayed.

This study was conducted in the context of the project "Effectiveness of COVID-19 Vaccination in Brazil Using Mobile Data" (EFFECT-BR), which is one out of ten, among 440 others worldwide, selected by the Grand Challenges ICODA COVID-19 Data Science, funded by the Bill & Melinda Gates foundation. Also, the Center for Healthcare Operations and Intelligence (NOIS[5]) which is part of PUC-Rio Industrial Engineering Department, together with Tecgraf institute, Fundação Oswaldo Cruz (FIOCRUZ) and Instituto D'Or (IDOR) gave the support needed so this study could be made.

## 6.1 Future Works

In this work, text treatment step on this work did not rely on modern Natural Language Processing (NLP) techniques. Leading, to manual tasks such as linking terms in the DO with data regarding comorbidities of the patients hospitalized. Therefore, as future work, such step can be automatized so more information can be considered.

Also, enriching the similarity pairwise metric by not only considering *is_a* relationships, but many others an ontology can provide. Also, such as the work of Glenda et al. (2021), the use of foundational ontologies and their associated metaproperties can also be applied for a project using the Data Science Lifecycle.

Since the computation of groupwise similarities relied on taking as an input regular pairwise similarity metrics, future works can also use other metrics such as Jiang and Conrath (1997), Resnik (1999) and Lin (1998) ones might be used, depending on the problem to solve and which one give better results. A combination of metrics, by assigning weights for each is also a possibility

[5] http://www.nois.ind.puc-rio.br

Gilbert et al. (2013) on their work presents the concept of semantical variables, which consider if a structure on a data set represents either an individual or a class, for assigning weights when calculating similarities. But also more important, semantic variables are paired with ordinary numerical and categorical ones, which wasn't attacked on this work and shall be done in future works inspired on this work.

Also, as a personal view, most of works making the connection with the world of ontologies with the world of data science is mainly on the bioinformatics, pairing ontologies to Data Science projects on other domains

This work based its Data Mining step on Unsupervised Machine Learning techniques. However, there are a plenty of opportunities to take the proposal of this work into Classification Algorithms.

# 7 References

1. ALSTON, Jesse M. ; RICK, Jessica A. A Beginner's Guide to Conducting Reproducible Research. **The Bulletin of the Ecological Society of America**, v. 102, n. 2, 2021.

2. ADIL FAHAD; NAJLAA ALSHATRI; TARI, Zahir; *et al.* A survey of clustering algorithms for big data: Taxonomy and empirical analysis. **IEEE Transactions on Emerging Topics in Computing**, v. 2, n. 3, p. 267–279, 2014.

3. AMARAL, Glenda; BAIÃO, Fernanda ; GUIZZARDI, Giancarlo. Foundational ontologies, ontology-driven conceptual modeling, and their multiple benefits to data mining. WIREs Data Mining and Knowledge Discovery, v. 11, n. 4, 2021.

4. ARP, Robert; SMITH, Barry D ; SPEAR, Andrew D. Building ontologies with basic formal ontology. Cambridge Mass.: The Mit Press, 2015.

5. ASHBURNER, Michael; BALL, Catherine A.; BLAKE, Judith A.; et al. Gene Ontology: tool for the unification of biology. Nature Genetics, v. 25, n. 1, p. 25–29, 2000. Disponível em: https://www.nature.com/articles/ng0500_25.

6. BABCOCK, Shane; BEVERLEY, John; COWELL, Lindsay G.; *et al.* The Infectious Disease Ontology in the age of COVID-19. **Journal of Biomedical Semantics**, v. 12, n. 1, 2021.
7. BARD, Jonathan B. L. ; RHEE, Seung Y. Ontologies in biology: design, applications and future challenges. **Nature Reviews Genetics**, v. 5, n. 3, p. 213–222, 2004.

8. CALINSKI, T. ; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics - Theory and Methods**, v. 3, n. 1, p. 1–27, 1974.

9. CHAMBERS, Michele. **Advanced Analytics Methodologies: Driving Business Value with Analytics**. [s.l.]: Pearson, 2014.

10. CONWAY, Drew. **Drew Conway**. Drew Conway. Disponível em: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram. Acesso em: 7 jun. 2022.

11. DAVIES, David L. ; BOULDIN, Donald W. A Cluster Separation Measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. PAMI-1, n. 2, p. 224–227, 1979. Disponível em: https://ieeexplore.ieee.org/document/4766909.

12. DISEASE ONTOLOGY. **Disease Ontology - Institute for Genome Sciences - Use Cases**. disease-ontology.org. Disponível em: https://disease-ontology.org/community/use-cases. Acesso em: 8 jun. 2022.

13. FALBO, Ricardo. SABiO: Systematic approach for building ontologies. **CEUR Workshop Proceedings**, v. 1301, 2014.

14. GENE ONTOLOGY CONSORTIUM. The Gene Ontology resource: enriching a GOld mine. **Nucleic Acids Research**, v. 49, n. D1, p. D325–D334, 2021. Disponível em: https://pubmed.ncbi.nlm.nih.gov/33290552/.

15. GIBERT, Karina; VALLS, Aïda ; BATET, Montserrat. Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. **Knowledge and Information Systems**, v. 40, n. 3, p. 559–593, 2013.

16. GOODFELLOW, Ian J; POUGET-ABADIE, Jean; MIRZA, Mehdi; *et al*. **Generative adversarial networks**. [s.l.: s.n.], 2014.

17. GUIZZARDI, Giancarlo. Ontology, Ontologies and the “I” of FAIR. **Data Intelligence**, v. 2, n. 1-2, p. 181–191, 2020.

18. GUIZZARDI, Giancarlo; BOTTI BENEVIDES, Alessander; FONSECA, Claudenir M.; *et al*. UFO: Unified Foundational Ontology. **Applied Ontology**, p. 1–44, 2021.

19. HAN, Jiawei; KAMBER, Micheline ; PEI, Jian. **Data mining : concepts and techniques**. Burlington, Ma: Elsevier, 2012.

20. HUNTER, John D., Matplotlib: A 2D Graphics Environment, **Computing in Science & Engineering**, v. 9, n. 3, p. 90–95, 2007.

21. JIANG, Jay J ; CONRATH, David W. Semantic similarity based on corpus statistics and lexical taxonomy. *In*: **The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)**. [s.l.: s.n.], 1997, v. 10, p. 19–33. Disponível em: https://aclanthology.org/O97-1002.

22. JO, Taeho. **Machine Learning Foundations**. Cham: Springer International Publishing, 2021.

23. KOU, Yufeng; LU, Chang-Tien; SIRWONGWATTANA, S.; *et al*. **Survey of fraud detection techniques**. IEEE Xplore. Disponível em: https://ieeexplore.ieee.org/abstract/document/1297040.

24. KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. **Psychometrika**, v. 29, n. 1, p. 1–27, 1964.

25. LASTRA-DÍAZ, Juan J.; GARCÍA-SERRANO, Ana; BATET, Montserrat; *et al*. HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. **Information Systems**, v. 66, p. 97–118, 2017.

26. LEE, WeiNchih; SHAH, Nigam; SUNDLASS, Karanjot; *et al*. Comparison of Ontology-based Semantic-Similarity Measures. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2008, p. 384–388, 2008. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655943/.

27. LIN, Dekang. An information-theoretic definition of similarity. *In*: **IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology (CIBCB)**. [s.l.]: Morgan Kaufmann Publishers Inc., 1998, p. 296–304.

28. LINNÉ, Carl von. Systema naturae, per regna tria naturae :secundum classes, ordines, genera, species cum characteribus, differentiis, synonymis, locis. 1767.

29. MAAS, Wolfgang ; STOREY, Veda C. Pairing conceptual modeling with machine learning. **Data & Knowledge Engineering**, v. 134, n. C, p. 101909, 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169023X21000367. Acesso em: 6 jul. 2021.

30. MADDALENA, L ; BAIÃO, Fernanda. OntoCovid: Applying SABiO to conceptual modeling well grounded in the COVID-19 domain. *In*: **CEUR Workshop Proceedings**. [s.l.: s.n.], 2021, v. 3050.

31. MARISCAL, Gonzalo; ÓSCAR MARBÁN ; COVADONGA FERNÁNDEZ. A survey of data mining and knowledge discovery process models and methodologies. **The Knowledge Engineering Review**, v. 25, n. 2, p. 137–166, 2010.

32. MARMANIS, Haralambos ; BABENKO, Dmitry. **Algorithms of the intelligent web**. [s.l.]: Manning Publications, 2009.

33. MCINNES, Leland; HEALY, John ; MELVILLE, James. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **arXiv:1802.03426 [cs, stat]**, 2020. Disponível em: https://arxiv.org/abs/1802.03426v3.

34. MEAD, A. Review of the Development of Multidimensional Scaling Methods. **The Statistician**, v. 41, n. 1, p. 27, 1992.

35. MEALY, G H. Another look at data. *In*: **Managing Requirements Knowledge, International Workshop on**. [s.l.: s.n.], 1967, v. 1, p. 525. Disponível em: http://doi.ieeecomputersociety.org/10.1109/AFIPS.1967.112.

36. MILLER, George A. WordNet: a lexical database for english. **Communications of the ACM**, v. 38, p. 39–41, 1995.

37. MILLER, George A. **WordNet: An electronic lexical database**. [s.l.]: MIT press, 1998.

38. NCBO BIOPORTAL. **Human Disease Ontology | NCBO BioPortal**. bioportal.bioontology.org. Disponível em: https://bioportal.bioontology.org/ontologies/DOID. Acesso em: 27 maio 2022.

39. OUR WORLD IN DATA. **Cumulative confirmed COVID-19 cases and deaths**. Our World in Data. Disponível em: <https://ourworldindata.org/grapher/cumulative-deaths-and-cases-covid-19>. Acesso em: 19 abr. 2022.

40. PARK, Hae-Sang ; JUN, Chi-Hyuck. A simple and fast algorithm for K-medoids clustering. **Expert Systems with Applications**, v. 36, n. 2, p. 3336–3341, 2009.

41. PEARSON, Karl. LIII. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 2, n. 11, p. 559–572, 1901.

42. PEDREGOSA, Fabian; GAËL VAROQUAUX; ALEXANDRE GRAMFORT; *et al.* Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Disponível em: <http://jmlr.org/papers/v12/pedregosa11a.html>.

43. PIETERSE, Vreda ; KOURIE, Derrick G. Lists, taxonomies, lattices, thesauri and ontologies: Paving a pathway through a terminological jungle. **KNOWLEDGE ORGANIZATION**, v. 41, n. 3, p. 217–229, 2014.

44. RESNIK, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. **Journal of Artificial**

**Intelligence Research**, v. 11, p. 95–130, 1999. Disponível em: https://doi.org/10.1613%2Fjair.514.

45. ROUSSEEUW, Peter J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987.

46. SARGSYAN, Astghik; KODAMULLIL, Alpha Tom; BAKSI, Shounak; *et al.* The COVID-19 Ontology. **Bioinformatics**, v. 36, n. 24, p. 5703–5705, 2020.

47. SCHRIML, Lynn M; MITRAKA, Elvira; MUNRO, James; *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. **Nucleic Acids Research**, v. 47, n. D1, p. D955–D962, 2019.

48. SHMUELI ; KOPPIUS. Predictive analytics in information systems research. **MIS Quarterly**, v. 35, n. 3, p. 553, 2011.

49. THUSITHA MABOTUWANA; LEE, Michael C ; COHEN-SOLAL, Eric V. An ontology-based similarity measure for biomedical data – Application to radiology reports. **Journal of Biomedical Informatics**, v. 46, n. 5, p. 857–868, 2013.

50. TSANG, Tim K; WU, Peng; LIN, Yun; *et al.* Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. **The Lancet Public Health**, v. 5, n. 5, 2020.

51. UNITED NATIONS. DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS. POPULATION DIVISION. **World population prospects : Highlights, 2019 revision**. New York: United Nations, 2019.

52. VAN ROSSUM, Guido ; DRAKE, Fred L. **Python 3 : reference manual**. United States: Sohobooks, 2009.

53. WAN, Ling; SONG, Justin; HE, Virginia; *et al*. Development of the International Classification of Diseases Ontology (ICDO) and its application for COVID19 diagnostic data analysis. **BMC Bioinformatics**, v. 22, 2021.

54. WANG, H.; AZUAJE, F.; BODENREIDER, O.; *et al*. Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. **IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing (IEEE Cat. No.04CH37612)**, 2004.

55. WANG, J Z; DU, Z; R. PAYATTAKOOL; *et al*. A new method to measure the semantic similarity of GO terms. **Bioinformatics**, v. 23, n. 10, p. 1274–1281, 2007.

56. WASKOM, Michael, seaborn: statistical data visualization, **Journal of Open Source Software**, v. 6, n. 60, p. 3021, 2021.

57. WIŚNIEWSKI, Dawid; POTONIEC, Jedrzej; ŁAWRYNOWICZ, Agnieszka; *et al*. Analysis of Ontology Competency Questions and their formalizations in SPARQL-OWL. **Journal of Web Semantics**, v. 59, p. 100534, 2019.

58. WORLD HEALTH ORGANIZATION. **WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020**. World Health Organization. Disponível em: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Acesso em: 22 abr. 2022.

59. WU, Huanhuan; ZHONG, Yichen; TIAN, Yingjie; *et al*. Automatic diagnosis of COVID-19 infection based on ontology reasoning. **BMC Medical Informatics and Decision Making**, v. 21, n. S9, 2021. Disponível em: https://doi.org/10.1186%2Fs12911-021-01629-0. Acesso em: 10 jun. 2022.

60. XIE, Haozhe; LI, Jie ; XUE, Hanqing. A survey of dimensionality reduction techniques based on random projection. **CoRR**, v. abs/1706.04371, 2017. Disponível em: http://arxiv.org/abs/1706.04371.

61. YU, Guangchuang; WANG, Li-Gen; YAN, Guang-Rong; *et al*. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. **Bioinformatics**, v. 31, n. 4, p. 608–609, 2015.