Data Curation and Optimization Techniques: A Systematic Mapping Study

Jaqueline Alexsandra Azevedo Ferreira[1], Martin A. Musicante[1], and Umberto Souza da Costa[1]

[1]Universidade Federal do Rio Grande do Norte

October 28, 2022

Abstract

We develop a Systematic Mapping Study to observe trends and research opportunities around the concepts and techniques used in Data Curation for Big Data. Our work investigates scientific publications with the aim of identifying how data curation has been used recently, to organize and publish big data corpora. We are interested in browsing, identifying the mathematical and computational tools used in data curation. We focus on identifying how data curation has been modeled in different scenarios and which computational/mathematical techniques have contributed to improve data curation, with the aim of answering the following questions: *(i)* What mathematical fields have most contributed in the context of Data Curation? *(ii)* Which classes of optimization algorithms are used in the context of Data Curation? *(iii)* What application domains have benefited the most from Data Curation? While our main focus is on the definition of new methods and algorithms, we identified a large number of papers that concentrates just on applying known techniques to specific domains. Our study may be useful to identify challenges and opportunities for further theoretical studies, as well as to show the use of some formal techniques in real-life applications.

Expert Systems

Data Curation and Optimization Techniques: A Systematic Mapping Study

Authors:

Jaqueline Alexsandra Azevedo Ferreira
DIMAp - Computer Science Department
Federal University of Rio Grande do Norte - Brazil
CETENS - Center for Exact Sciences and Technology
Federal University of Recôncavo da Bahia - Brazil
email: jaquelineazevedo@ufrb.edu.br ORCID: 0000-0002-3184-6135

Martin A. Musicante
DIMAp - Computer Science Department
Federal University of Rio Grande do Norte - Brazil
email: mam@dimap.ufrn.br ORCID: 0000-0001-5589-3895

Umberto Souza da Costa
DIMAp - Computer Science Department
Federal University of Rio Grande do Norte - Brazil
email: umberto.costa@ufrn.br ORCID: 0000-0001-9291-2995

Conflict of Interest

The authors declare no conflicts of interest for this article.

Abstract

We develop a Systematic Mapping Study to observe trends and research opportunities around the concepts and techniques used in Data Curation for Big Data. Our work investigates scientific publications with the aim of identifying how data curation has been used recently, to organize and publish big data corpora. We are interested in browsing, identifying the mathematical and computational tools used in data curation. We focus on identifying how data curation has been modeled in different scenarios and which computational/mathematical techniques have contributed to improve data curation, with the aim of answering the following questions: *(i)* What mathematical fields have most contributed in the context of Data Curation? *(ii)* Which classes of optimization algorithms are used in the context of Data Curation? *(iii)* What application domains have benefited the most from Data Curation? While our main focus is on the definition of new methods and algorithms, we identified a large number of papers that concentrates just on applying known techniques to specific domains. Our study may be useful to identify challenges and opportunities for further theoretical studies, as well as to show the use of some formal techniques in real-life applications.

Graphical/Visual Abstract and Caption

Caption: We propose an overview on the applications of data curation, as well as the main models and techniques explored in the area, by summarizing the findings reported in the publications in this field over the last decade.

1. INTRODUCTION

Data Curation is a process of organizing and extracting information from a corpus of data. Data Curation produces *meta-data* that adds quantitative, semantic, and explicit information to a set of data. The goal of this addition is to ease the task of understanding and deducting new facts from raw data [65]. Data Curation has a multidisciplinary nature. It combines mathematical, computational, and statistical techniques, with semantic knowledge about the data in an application domain. In this context, Data Curation consists on retrieving, classifying, preserving, and reusing domain-specific data.

The main goal of this paper is to develop a systematic mapping study [51] on the use of tools for automatizing the process of Data Curation in any domain of application. This will help us to identify tendencies, opportunities, and challenges in Data Curation.

Systematic mappings are an useful tool to investigate how a given topic has evolved along time, unveiling quantitative data over the publications on a specific area. With the aiming of better presenting summarized information, systematic mappings use to employ graphical representations that points out over-explored aspects as well as research opportunities.

To the extent of our knowledge, there are only a few papers devoted to building systematic mappings, or similar studies for the activity of Data Curation. These papers include [53], which looks into standard policies and practices related to data curation programs in the UK. In [4], useful articles and books for understanding digital research data curation in academic and other research institutions are cataloged. It covers research topics such as data creation, acquisition, metadata, provenance, repositories, management, funding agency requirements, open access, peer review, publication, citation, sharing, reuse, and preservation. The works by Bailey *et al* [2,3] gather books and technical reports that are useful for curation and digital preservation. To the extent of our knowledge, our study is the first one devoted to survey the use of formal (mathematical, statistical, computational) tools in the data curation process.

We have structured the paper as follows: In Section 2, we present the methodology for systematic mappings reviews, including the definition of research questions, search strategy, and the procedure adopted. In Section 3, we answer our research questions and discuss the results. Finally, we discuss the conclusions of our findings and directions for future work.

2. Methodology

We adopted the procedure defined in [33] to perform our systematic mapping. Figure 1, taken from [51], shows the steps of the protocol (upper line) and the outcomes of each step (lower line). Each of these steps are detailed in the next sections.

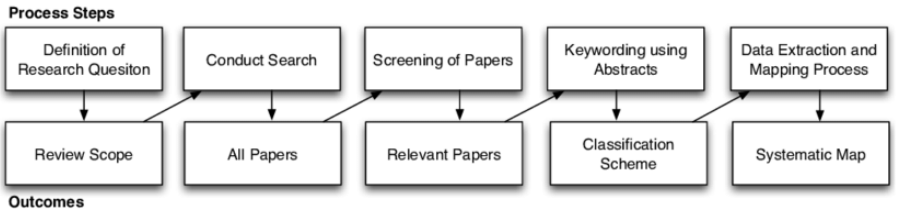

Figure 1: Systematic Mapping Process [51].

2.1 Definition of research questions

This activity defines the focus of the systematic mapping study, by specifying the scope of the study.

Research Question 01: "How Mathematics has contributed in the context of Data Curation?"

Data Curation is a multidisciplinary process. It may involve activities that require Mathematical and Statistical knowledge. Our first research question investigates to what extent these sciences are used in practice.

Table 1: Number of articles by digital library.

Database	Number of Papers
IEEE Xplore	07
Scopus	82
Mendeley	29
Web of Science	10
ACM	05
Science Direct	04
Total	137

Research Question 02: "Are there classes of optimization algorithms being used in the context of Data Curation? If yes, which ones?"

Since data curation consists on summarizing and extracting useful information from a corpus of data, the use of optimization algorithms seems to be a natural choice. Given a set of conditions, these algorithms look for optimal solutions. They can be used in the process of choosing the most relevant information, depending on domain-specific conditions.

Research Question 03: "Which are the most common application areas for Data Curation?"

Our third research question will identify the areas of application in which Data curation has been used.

2.2 Conduct Search: Search string definition and data extraction

The *search string* on a systematic mapping is the string used to query the bibliographic databases, looking for publications that will help in answering the research questions.

In our case, we identified the word "`optimization`", that is used in Computer Science, Mathematics, and Statistics to define processes that look for optimized solutions. In this way that word is enough to retrieve relevant papers in the three areas. We also included the term "`Data Curation`" to narrow the search to the area of interest for our study.

The search string was, thus, defined as: `"Data Curation" AND "Optimization"`

This search string was used to retrieve papers in six of the most used digital libraries, shown in Table 1. In that table we show the number of papers obtained by searching each database.

For each database, we retrieve the papers that contained our search string in their title and abstract. We performed this search for papers published between January 2011 and August 2022. A total number of 137 papers were found.

2.3 Publication screening

The next step in the construction of our systematic mapping consists on screening the publications, to select those that will be used to answer our research questions.

Inclusion and Exclusion Criteria were defined to help in building the corpus of papers in our study:

Inclusion Criteria: Initially, we included on our study all the 137 references resulting from the search. Other papers may have been included, provided that they are pertinent to our study. This option was not used in our mapping.

Exclusion Criteria: We excluded publications from our study, in acordance to the following criteria:

- Short papers (with less than four pages);
- Papers that are not available for download in full version;
- Papers that address manually performed Data Curation;
- Papers that are not available in English;
- Whole books or presentations;
- Duplicates (only one version of the paper is included).

After using the inclusion/exclusion criteria on the corpus of 137 publications retrieved by using the search string, we filtered out 70 of them. The remaining 67 documents were used as input to produce a classification scheme, as shown next.

2.4 Keywording using Titles and Abstracts

The Keywording activity consists on mapping the contents of each paper into a set of predefined categories. The definition of these categories depends on the goals of the systematic mapping study. In our case, we defined seven thematic facets containing categories.

For the definition of facets and categories we begin by using the My-SAE tool [37] to identify the most frequent words appearing in the title and abstract of the 61 papers included in the study. The result of this step is shown in Figure2. These words were used as an initial proposal to build the facets and categories in our study. From this list, we defined the following categories and facets:

1. *Computational Problem*.
   Characterizes the way in which the data search space is explored. The categories in this facet are: *Optimization*, *Classification*, *Modeling*, *Filtering*.
2. *Mathematical Domain*.
   This facet includes fields of mathematics that have made decisive contributions to solving the computational problems present in the articles. The categories in these facets are: *Statistics*, *Algebra*, *Analysis*, *Geometry and Topology*, *Dynamical Systems*, and *Differential Calculus*.
3. *Computational Approach*.
   Techniques and strategies used to solve the computational problem of the paper. The categories in this facet are: *Deep Mearning*, *Machine Learning*, *General AI*, *Discrete Optimization*, *Big Data Processing Techniques*, *Mathematical Optimization*, *Heuristics*, *Graph Algorithms*, *Game Theory*.
4. *Type of Algorithm*.
   Classifies the algorithms used to solve the computational problem. The categories in this facet are: *Genetic/Evolutionary Algorithms*, *Classification and Clustering*, *Neural Network*, *Integer Programming*, *Divide and Conquer*, *Not Specified*, *Greedy Algorithm*.
5. *Application Domain*.
   Includes the application areas to which Data Curation is performed. The categories in this facet are: Medicine, *Biomedicine*, *Neural Networks*,*Genomic*, *Data Curation*, *Big Data Analytics*, *Digital Humanities*, *Environmental Research*, *Biology*, *IoT*, *Pattern Recognition*, *Astronomy*, *Chemistry*, *Cloud Storage*.

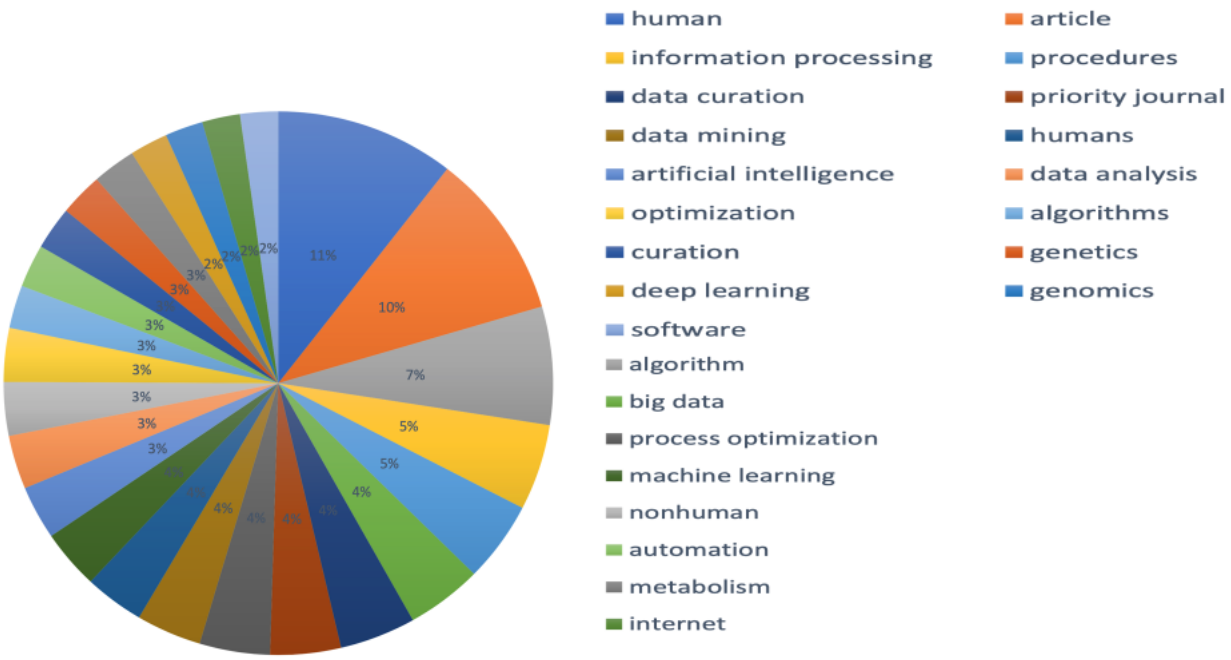

Figure 2: Keywords produced by My-SAE.

6. *Research Type*.
   The research approach of the paper, as defined by Wieringa et al [71]. The categories in this facet are: *Evaluation Research*, *Solution Proposal*, *Validation Research*, *Philosophical Paper*, *Opinion Paper*, *Experience Paper*.

7. *Origin of Authors*.
   We identified the countries of affiliation of each author. In the case of authors with more than one country of affiliation, we consider just the first country. The categories found are: *Brazil*, *USA*, *Greece*, *Japan*, *Netherlands*, *Portugal*, *Spain*, *Qatar*, *Denmark*, *Iceland*, *UK*, *Taiwan*, *Israel*, *Sweden*, *Switzerland*, *Belgium*, *Canada*, *Australia*, *New Zealand*, *Pakistan*, *Singapore*, *China*, *Luxembourg*, *Germany*, *France*, *Italy*, *Finland*, *Austria*, and *Korea*.

2.5 Data Extraction and Classification

Based on the research questions listed, and the categories and facets defined previously, we classified papers into categories. We followed the recommendation of Petersen [51] according to which, at least three evaluators must be part of the classification process. In this manner, all the authors participated from this step on.

The group analyzed the title and abstracts of all the 67 selected publications. The discussion about most papers also involved a brief reading of their body. We classify the papers into two groups (Table 2):

1. *Models/Techniques:* 44 papers having contributions related to the definition of optimization techniques in Data Curation; This group will be used as the main source of data to answer the research questions.
2. *Applications:* 23 papers applying optimization techniques to Data Curation, without proposing new methods.

The classification of each paper into their facets and categories was used next, to answer the research questions. This classification is available (in tabular form) at `shorturl.at/D1234`.

Table 2: Papers selected in this study.

Category	Reference
Models / Techniques	[6], [8], [10], [24], [15] [18], [17], [20], [22], [25], [27], [28], [43], [64], [30], [32], [34], [38], [40], [45], [46], [47], [49], [54], [55], [56], [57], [59], [60], [61], [66], [68], [69], [70], [73], [74], [75], [77], [26], [42], [44], [11], [13], [36].
Applications	[1], [5], [7], [9], [12], [14], [16], [19], [21], [23], [29], [31], [35], [39], [41], [48], [50], [52], [58], [63], [67], [72], [76].

3. Answering the Research Questions

In this section, we answer the research questions, carry out a bibliometric analysis and provide a synthesis of the results, trying to identify possible research niches. For answering our questions, only *Models/Techniques* articles were considered.

3.1 Answer to research question 01: "How Mathematics has contributed in the context of Data Curation?"

To answer this question, we focus on the corpus of 44 papers that define models and techniques for data curation. Figure 3 presents the distribution of publications that employ concepts and techniques from diverse fields of Mathematics. As we expected, *Statistics* appears as the most used mathematical field, due to its role in using methodologies capable of collecting, analyzing, interpreting, and predicting results over large data sets for better decision-making possibilities.

Let us now analyze the bubble chart in Figure 4. This kind of chart allows us to plot three dimensions on the same plane, being powerful enough to provide an overview of a field. In this figure, we show how the facets *Computation Problem* and *Applications* (horizontal axis, left and right sides respectively) relate to *Mathematical Fields* (vertical axis). The size of each bubble is proportional to the number of publications that simultaneously contribute to the intersecting categories. Notice that, as papers may be classified into more than one category, the sum of papers in the bubbles is greater than the number of *Models/Techniques* papers in our study.

From the right-hand side of the chart, we observe that *Biomedicine* and *Data Curation* are the most frequent application areas on the papers, each one representing, respectively, 8 (18.18%) and 9 (20.45%) of the selected publications (Figure 4). Concerning Biomedicine, 5 (62.5%) of these papers uses *Statistics*, 1 (12.5%) of them explores *Algebra*, or *Geometry and Topology*, and 2 (25%) of them employs mathematical *Analysis* techniques. Most of the papers that include *Data Curation* as application do not clearly identify the contribution of mathematical concepts. Just 4 papers (44.4%) specifically mention a field of Mathematics, while the other 5 did not make any reference to a mathematical formalism.

Concerning the facet *Applications*, the majority of papers either propose solutions based on Statistics or do not explicitly specify a mathematical field. We believe that this fact is due to the large adoption of statistical techniques to explore big data.

Concerning the left-hand side of Figure 4, it is possible to observe that 24 (54.54%) of the papers concentrates on *Classification* problems. The solutions proposed in these papers are frequently based on *Statistics* 15 (62.5%) for data clustering. To evaluate the performance of models used to solve this problem, metrics such as accuracy, precision, and recall need to be calculated.

We note that 20 (45.45%) of the selected publications concentrate on computational *Modeling* problems, with 8 (40%) having contributions from the *Statistics* category, 2 (10%) from the *Algebra* category, 2 (10%) from the

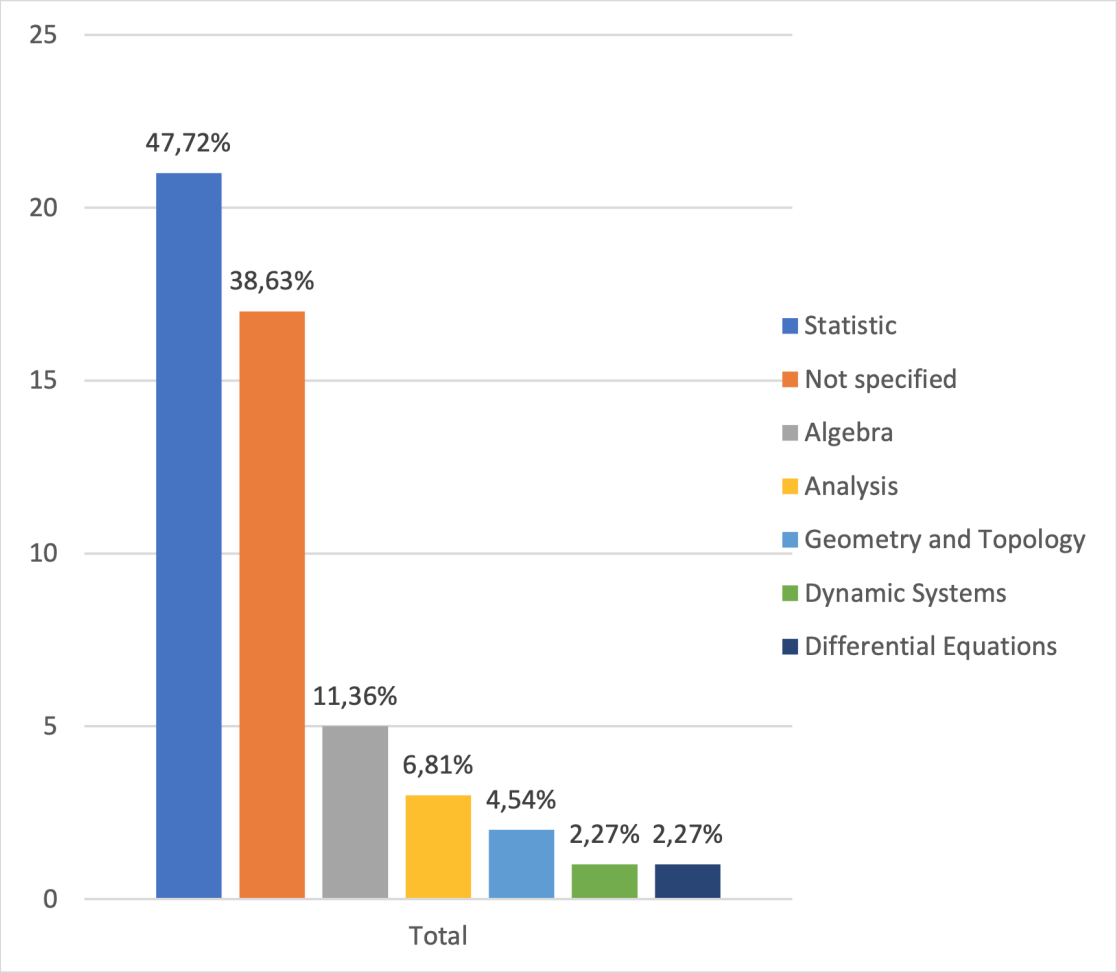

Figure 3: Distribution of *Models/Techniques* papers by field of Mathematics.

Analysis category, and 1 (5%) from the other areas.

Concerning the *Modeling* problem, we identified contributions from all the Mathematical fields that we consider. Notice that there are 8 papers involving modeling that do not explicitly state a mathematical area as main interest. Looking at the right-hand side of Figure 4, we notice that a significant number of papers explore statistical and numerical methods for proving models for biological sciences.

In our investigation, we found that the same paper may deal with more than one computational problem. At the same time, each paper normally have just one application area. This explains the lesser density of bubbles on the right-hand side of Figure 4.

Notice that the use of statistical methods is predominant from both the viewpoints of applications and formulation of mathematical problems in the context of data curation (Figure 4).

Summary: It is clear that Mathematics/Statistics has an important role in the selection and filtering processes in Data Curation.

Differential Calculus may be used to look for better parameters to algorithms so that intelligent models learn patterns through mathematical functions that seek to minimize or maximize. Despite the possible contributions of *Differential Calculus* in all computational problems, our work reveals a low level of exploration and immaturity

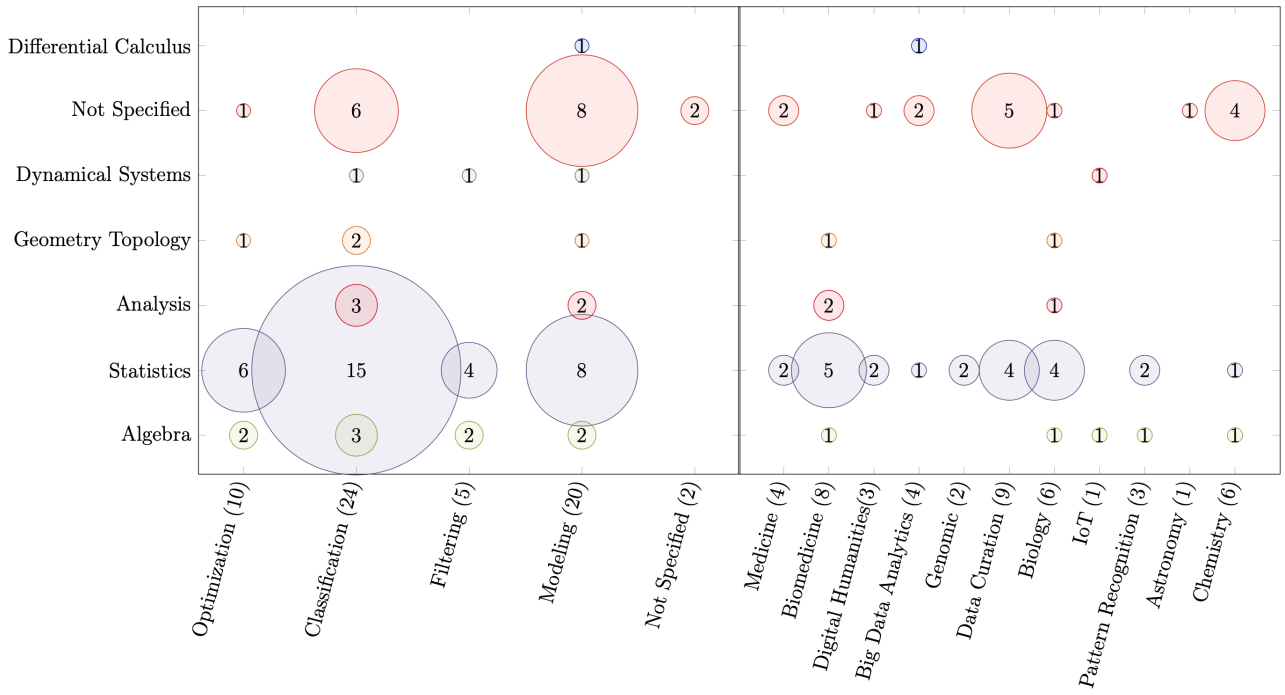

Figure 4: Mathematical Area vs. Computational Problems/Applications (*Models/Techniques* papers).

in the use of many fields of Mathematics in the construction of new results in the field of Data Curation.

The area of *Algebra* is also frequently used to solve computational problems in the context of Data Curation. Notice that Algebra is the only area of Mathematics explored for all the computational problems we consider. However, observe that only 9 of the 44 papers use Algebra elements as basis to provide computational solutions. This observation may suggests the potential to further exploration on the use of Algebra for Data Curation.

3.2 Answer to research question 02: "Are there classes of optimization algorithms being used in the context of Data Curation? If yes, which ones?"

Our second research question is intended to give a better understanding about the use of optimization algorithms in Data Curation. In Figure 5 we present the distribution of papers with concern to the category of algorithm explored. Regarding optimization techniques, about half of the papers either do not provide details about the algorithm used or provide *ad-hoc* solutions. The frequent use of tailored algorithms indicates the need to adapt or devise solutions to better fit the nature of the application data.

As expected, a significant portion of papers concentrates on optimizing clustering methods. In special, *Neural Networks* is one of the main approaches used to cluster data. It is relevant to inform that our study reported that 5.2% of the articles received contributions using more than one type of algorithm. In addition to the graph shown in Figure 5, we built the bubble chart shown in Figure 6.

Figure 6 confirms our intuition about the importance of clustering algorithms to data curation, specially by using neural networks. We observe that 15 (34.1%) of the *Models/Techniques* papers report the use of *Machine Learning* techniques, from which 8 of them use *Neural Networks*. Notice the small number of papers that report the use of *Genetic/Evolutionary* methods, *Integer Programming* and *Greedy Algorithms*.

The seldom use of *Integer Programming* methods follows our intuition. They deal with the optimization of

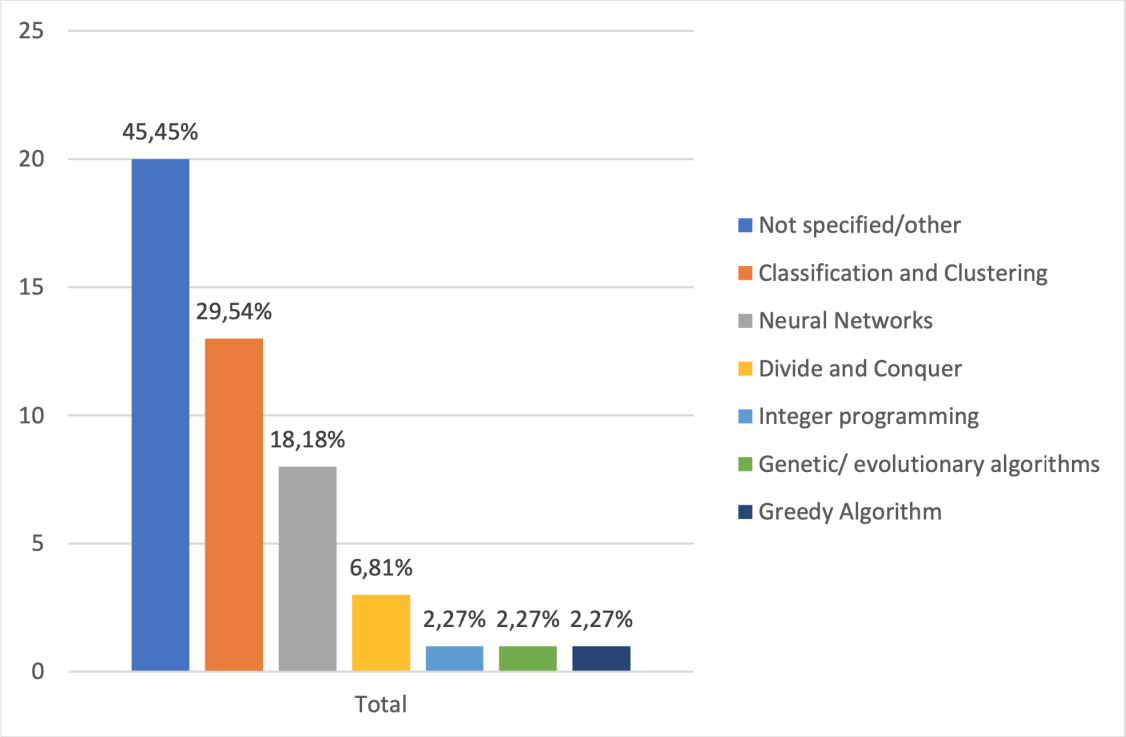

Figure 5: Distribution of papers by type of algorithms.

integer functions, which our study did not detect as frequent in the application domains.

In the case of *Genetic/Evolutionary* methods and *Greedy Algorithms*, both kinds of algorithms work to find solutions from a set of data items. They make locally optimal choices to find a globally optimal solution. They has a very low operating cost and may be applied for classification purposes in different application contexts. The small number of papers dealing with greedy and evolutionary algorithms may indicate opportunities for research in the area.

Summary: Our analysis shows that Artificial Inteligence and Machine Learning methods have contributed to Data Curation. Most algorithms identified in our study were tailored for specific application domains. We notice that there are some candidate algorithm classes (like evolutionary and greedy algorithms) that have not been extensively explored yet in the context of data curation.

3.3 Answer to research question 03: "Which are the most common application areas for Data Curation?"

Data Curation is a way of organize domain-specific data, allowing for a better analysis and visualization of them, as well as of derived knowledge.

The bubble chart presented in Figure 7 considers only the 44 papers classified as *Models/Techniques*. As expected, our study identified a large number of application domains for data curation. This figure presents a more regular distribution of papers than the previous bubble charts. This fact indicates the general interest of Data Curation on many application areas, tackling a variety of computational problems, and using a diversity of techniques.

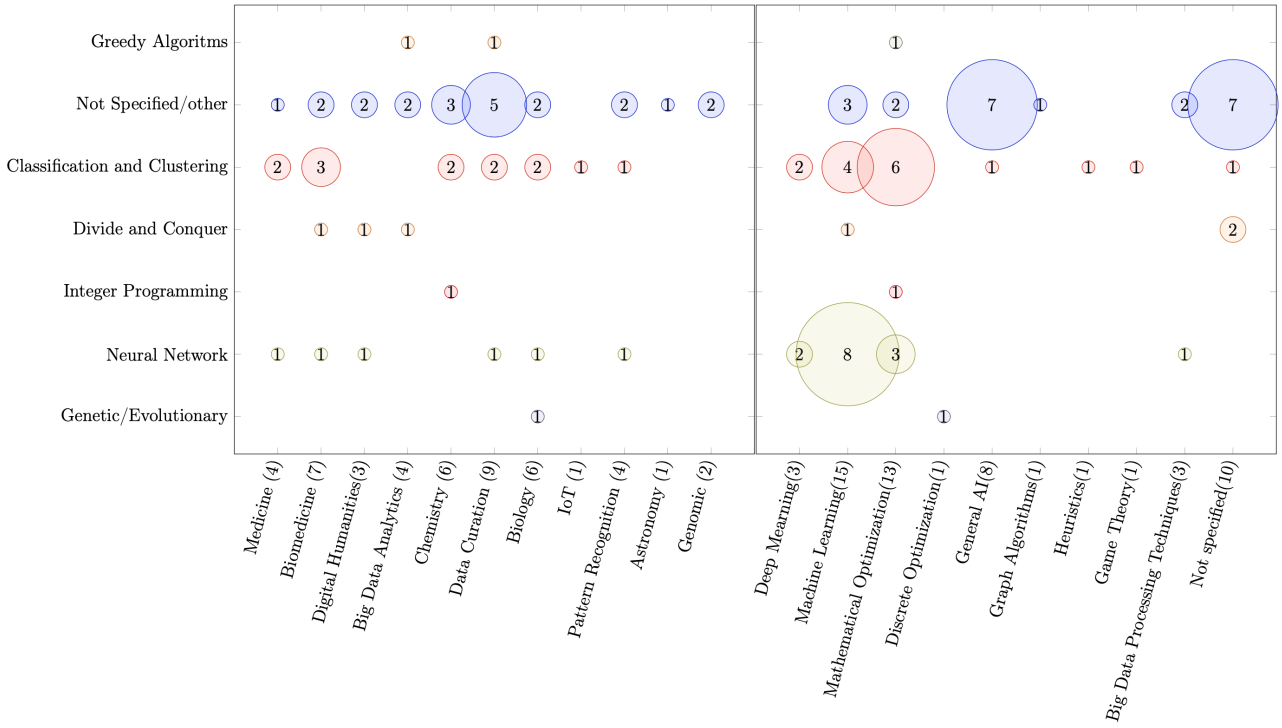

Figure 6: Type of Algorithm vs. Applications/Techniques (*Models/Techniques* papers).

Once again, we identified the prominence of Biological Sciences as popular domain of application of data curation. Most of clustering algorithms are focused on Medicine, Biomedicine and Biology.

We also identify a significant number of papers dedicated to improve data curation algorithms themselves. In this case, the research focuses on improving the efficiency and expressiveness of the algorithm, as well as the better use of meta-data in the curation process.

Notice that *Modeling* and *Classification* are the computational problems with the broader ranges of applications. The three *Models/Techniques* papers with applications of *Big Data Analytics* do not explore classification problems. This is due to the fact that those papers are not focused in the data analytics process, but on the proposal of Big Data platforms or on conceptual modeling.

We noticed the weak connection of the *Big Data Analytics* category with the mentioned techniques. At the same time, the relationship between the *Data Curation* category and the techniques used is much stronger.

As expected, almost all the applications depend on *Modeling* and *Classification*. Concerning the implementation of solutions, many of these applications uses AI, including *Deep Learning* and *Machine Learning* approaches, or *Mathematical Optimization* techniques.

The bar graph in Figure 8 include numbers for all the 65 papers in our study (44 *Models/Techniques* and 21 *Application* papers). The most prominent application area, with 17 papers (*i.e.*, about one fourth of our corpus) is Biomedicine. Notice that areas related to Biomedicine, such as Medicine and Biology are also frequent application fields.

The 10 papers having data curation as application domain correspond to works that propose original data curation methods or systems. The majority of those papers propose improvements and extensions to existing data curation approaches.

It is worth noticing that the fields of applications is quite broad. Moreover, applications like Big Data Analytics,

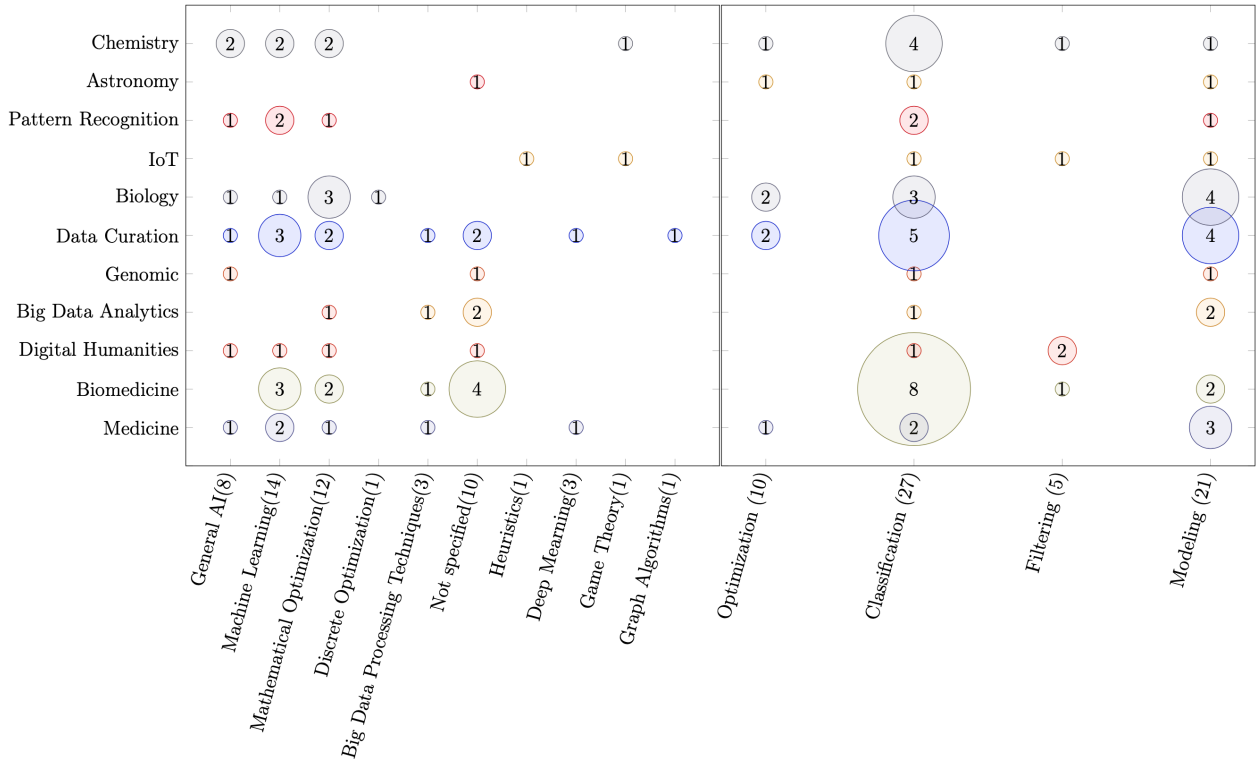

Figure 7: Applications vs. Techniques/Computational Problem (*Models/Techniques* papers).

Pattern Recognition and Cloud Storage may, themselves, be used in a vast range of applications.

Summary: By answering this research question, we confirmed our intuition about Data Curation being applicable to most domains dealing with large amounts of data. The most frequent application areas found by our study are those related to Biology and Medicine.

3.4 Bibliometric Analysis

In this section we discuss how the area of data curation has evolved along the period we consider in our analysis. Let us consider, again, the 65 publications marked *Models/Techniques* and *Application* papers of our study. The left-hand side of Figure 9 shows the number of papers by area and year of publication. We can see that, in general, recent years have a greater frequency. Moreover, fields of application related to the biological sciences have a larger participation in recent years. This is shown by the increasing size of the bubbles in the upper-left quarter.

The right-hand side of Figure 9 shows the affiliation of authors of the 65 papers in our study, by year of publication. We notice the larger size of bubbles on the column counting the number of USA authors, as well as the general increase of frequency for the last four years. This figure indicates, once again, that the field is gaining popularity in the research community.

Figure 10 shows the publications by year. Notice that both *Application* and *Models/Techniques* papers have a similar behavior. From 2018 on, we verify a significant increase in the number of *Models/Techniques* publications, so that 61.36% of the articles were published in the last five years and 45.45% in the last three years. The year 2020 had the highest publication rate, with ten *Models/Techniques* and five *Applications* papers, be-

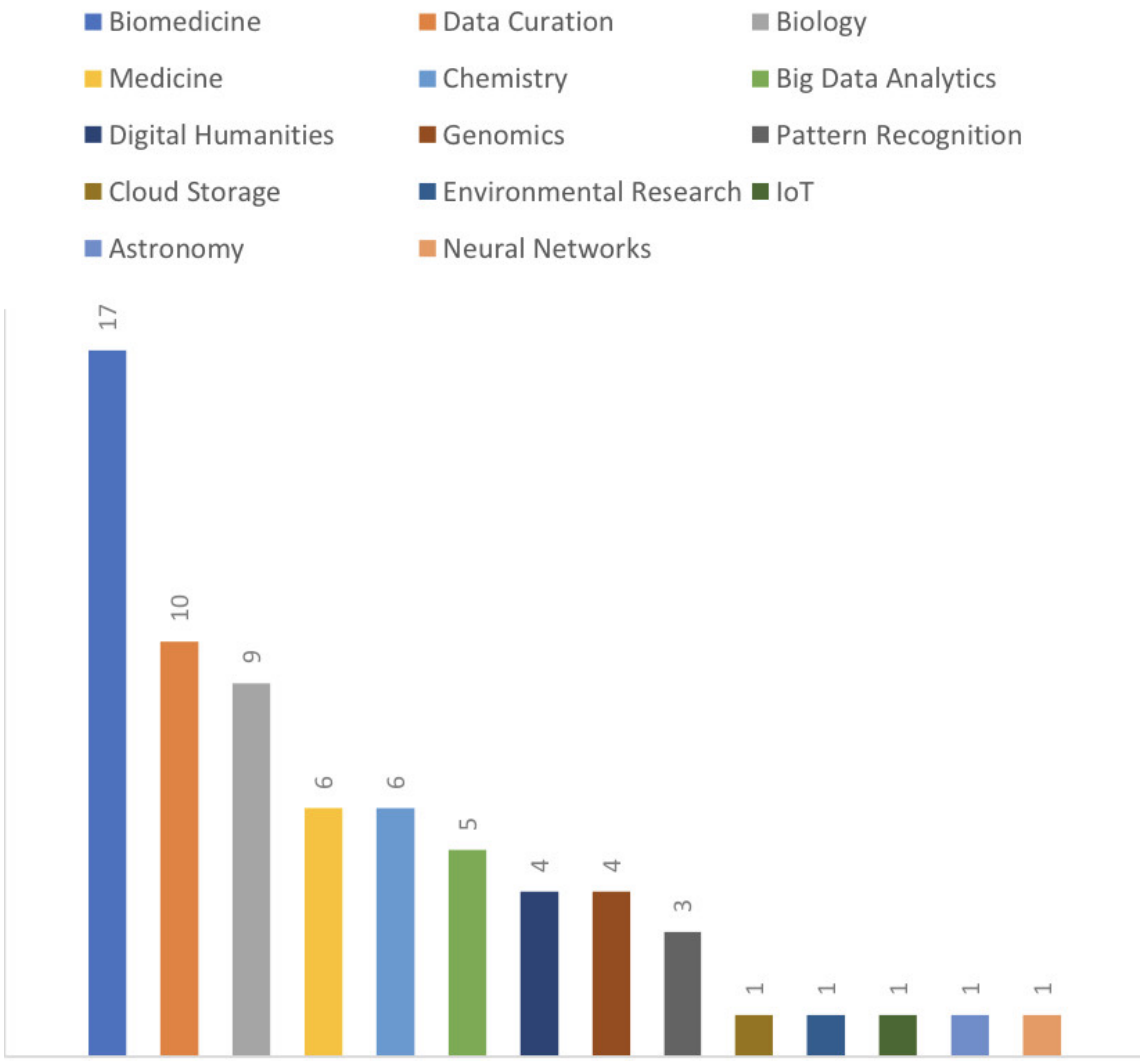

Figure 8: Most frequent applications of Data Curation.

ing 22.38% of the corpus. The corpus of publications was built in August of 2022, which explains the smaller number of papers in that year.

Conclusion

Our work reveals a growing interest of the academic and industrial communities in using data curation, as well as the use of mathematical and computational tools towards providing a useful experience to the user.

We notice a significant increase on the number of publications focused on the proposal of further concepts and techniques from 2017 on: we found a total of 11 papers in this category in the period from 2011 to 2016; this number raised to 33 in the period from 2017 to August 2022. The investigation we conducted indicates that data curation has been mainly applied to *Medicine*, *Biomedicine*, *Biology*, *Genomics* and *Chemistry*. The Data Curation process is capable of presenting data management strategies, for instance, during our analysis, we noticed that Data Curation was used to feed epidemiological studies and evaluate health profiling in hospitals. Comparing all the bubble charts, we noticed significant contributions from Statistics, and the use

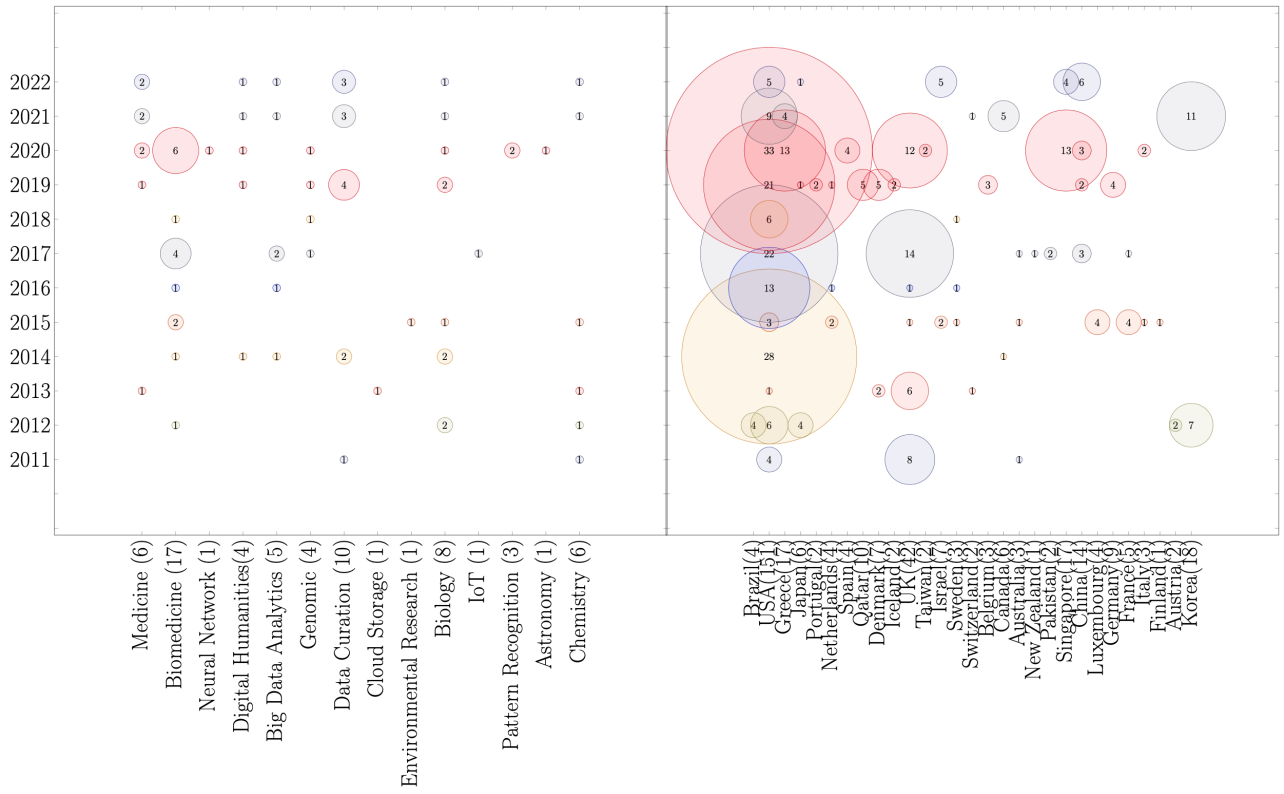

Figure 9: Year of publication vs. Applications/Origin of authors (*Models/Techniques* and *Applications* papers).

of classification and grouping algorithms in a well-distributed way in the health areas.

Many Data Curation solutions are based on computational *Modeling* and *Classification* techniques. These problems can be solved using, mainly, classification algorithms and Statistics. *Algebra* also can provide solutions to various computational problems related to Data Curation, while *Differential Calculus* is often effective in optimizing parameters to algorithms. We notice that the level of exploration of these areas may be improved in terms of maturity. Our analysis reveals the interest on optimizing big data through Machine Learning techniques with the support of Neural Network algorithms, classification, and clustering of data. Finally, despite the predominance of USA-based authors, our investigation corroborates the growing interest and contributions of several research groups around the world in using optimization strategies in the context of Data Curation.

Our work is intended to help the community to identify mathematical and computational tools that have been successfully applied to build Data Curation solutions. In this sense, there are algebraic structures such as Matroids [62] that could be used in together with greedy algorithms, to obtain optimal curated datasets. It may be a promising research area that we intend to investigate as future work.

References

[1] Athanasios Antonakoudis, Rodrigo Barbosa, Pavlos Kotidis, and Cleo Kontoravdi. The era of big data: Genome-scale modelling meets machine learning. *Computational and Structural Biotechnology Journal*, 18:3287–3300, 2020.

[2] Charles W Bailey Jr. Digital curation bibliography: Preservation and stewardship of scholarly works, 2012 supplement.

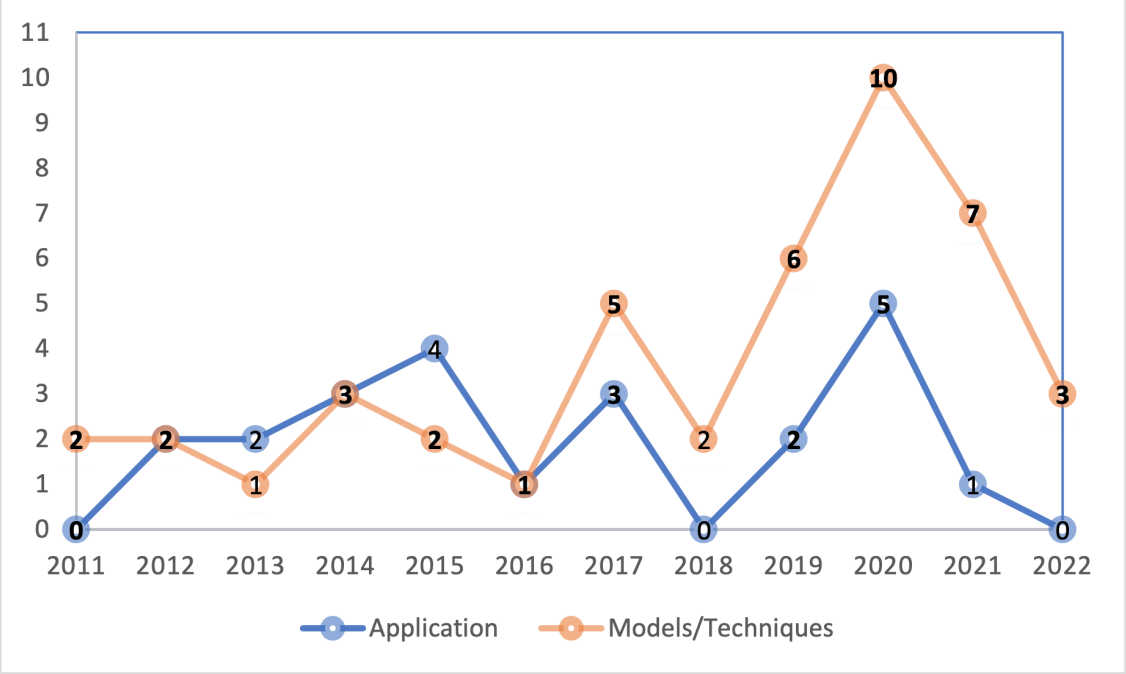

Figure 10: Distribution of *Models/Techniques* and *Application* articles over the years.

[3] Charles W. Bailey Jr. *Digital curation bibliography: Preservation and stewardship of scholarly works*. Digital Scholarship, 2012.

[4] Charles W. Bailey Jr. *Research Data Curation and Management Bibliography*. Digital Scholarship, 2021.

[5] Anita Bandrowski, Matthew Brush, Jeffery S Grethe, Melissa A Haendel, David N Kennedy, Sean Hill, Patrick R Hof, Maryann E Martone, Maaike Pols, Serena S Tan, et al. The resource identification initiative: A cultural shift in publishing. *Neuroinformatics*, 14(2):169–182, 2016.

[6] Richard Baran and Trent R Northen. Robust automated mass spectra interpretation and chemical formula calculation using mixed integer linear programming. *Analytical chemistry*, 85(20):9777–9784, 2013.

[7] Nam Bui, Solomon Henry, Douglas Wood, Heather A Wakelee, and Joel W Neal. Chart review versus an automated bioinformatic approach to assess real-world crizotinib effectiveness in anaplastic lymphoma kinase–positive non–small-cell lung cancer. *JCO clinical cancer informatics*, 1:1–6, 2017.

[8] Peter Buneman, James Cheney, Sam Lindley, and Heiko Mueller. The database wiki project: a general-purpose platform for data curation and collaboration. *AcM SIGMOD Record*, 40(3):15–20, 2011.

[9] Cyrielle Calmels, Andréa McCann, Laetitia Malphettes, and Mikael Rørdam Andersen. Application of a curated genome-scale metabolic model of cho dg44 to an industrial fed-batch process. *Metabolic engineering*, 51:9–19, 2019.

[10] Jianfang Cao, Aidi Zhao, and Zibang Zhang. Automatic image annotation method based on a convolutional neural network with threshold optimization. *Plos one*, 15(9):e0238956, 2020.

[11] Yushi Cao, David Berend, Palina Tolmach, Guy Amit, Moshe Levy, Yang Liu, Asaf Shabtai, and Yuval Elovici. Fair and accurate age prediction using distribution aware data curation and augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3551–3561, 2022.

[12] Nandish Chattopadhyay, Chua Sheng Yang Viroy, and Anupam Chattopadhyay. Re-markable: Stealing watermarked neural networks through synthesis. In *International Conference on Security, Privacy, and Applied Cryptography Engineering*, pages 46–65. Springer, 2020.

[13] Chen Chen, Zvi Yaari, Elana Apfelbaum, Piotr Grodzinski, Yosi Shamay, and Daniel A Heller. Merging data curation and machine learning to improve nanomedicines. *Advanced Drug Delivery Reviews*, page 114172, 2022.

[14] Ratul Chowdhury, Anupam Chowdhury, and Costas D Maranas. Using gene essentiality and synthetic lethality information to correct yeast and cho cell genome-scale models. *Metabolites*, 5(4):536–570, 2015.

[15] Noemi Deppenwiese, Petra Duhm-Harbeck, Josef Ingenerf, and Hannes Ulrich. Mdrcupid: A configurable metadata matching toolbox. In *MedInfo*, pages 88–92, 2019.

[16] Quinn Dombrowski and Ronelle Alexander. Digital humanities development without developers: Bulgarian dialectology as living tradition. In *DH-CASE II: Collaborative Annotations on Shared Environments: metadata, tools and techniques in the Digital Humanities*, pages 1–9. 2014.

[17] Ioannis Drivas, Dimitrios Kouis, Daphne Kyriaki-Manessi, and Georgios Giannakopoulos. Content management systems performance and compliance assessment based on a data-driven search engine optimization methodology. *Information*, 12(7):259, 2021.

[18] Ioannis C Drivas, Damianos P Sakas, Georgios A Giannakopoulos, and Daphne Kyriaki-Manessi. Big data analytics for search engine optimization. *Big Data and Cognitive Computing*, 4(2):5, 2020.

[19] Amanda K Dupuy, Marika S David, Lu Li, Thomas N Heider, Jason D Peterson, Elizabeth A Montano, Anna Dongari-Bagtzoglou, Patricia I Diaz, and Linda D Strausbaugh. Redefining the human oral mycobiome with improved practices in amplicon-based taxonomy: discovery of malassezia as a prominent commensal. *PLoS One*, 9(3):e90899, 2014.

[20] William MB Edmands, Lauren Petrick, Dinesh K Barupal, Augustin Scalbert, Mark J Wilson, Jeffrey K Wickliffe, and Stephen M Rappaport. compms2miner: An automatable metabolite identification, visualization, and data-sharing r package for high-resolution lc–ms data sets. *Analytical chemistry*, 89(7):3919–3928, 2017.

[21] Mohamed Y Eltabakh. Data organization and curation in big data. In *Handbook of Big Data Technologies*, pages 143–178. Springer, 2017.

[22] Maria Esteva, Weijia Xu, Nevan Simone, Amit Gupta, and Moriba Jah. Modeling data curation to scientific inquiry: A case study for multimodal data integration. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 235–242, 2020.

[23] Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309, 2014.

[24] X From. Mumal: Multivariate analysis in shotgun proteomics using machine learning techniques. 2012.

[25] Niall Gaffney, Christopher Jordan, Tommy Minyard, and Dan Stanzione. Building wrangler: A transformational data intensive resource for the open science community. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 20–22. IEEE, 2014.

[26] Angel Luis Garrido, Alvaro Peiro, Carlos Bobed, Eduardo Mena, and Cristian Morte. Icix: A semantic information extraction architecture. In *25th International Database Engineering & Applications Symposium*, pages 75–83, 2021.

[27] Yocheved Gilad, Katalin Nadassy, and Hanoch Senderowitz. A reliable computational workflow for the selection of optimal screening libraries. *Journal of cheminformatics*, 7(1):1–17, 2015.

[28] Michael Mu-Chien Hsu and Richard Jui-Chun Shyur. Auto curation on facenet embeddings with gamma and gaussian distribution to predict model performance in actual industrial deployment. In *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, pages 46–49. IEEE, 2020.

[29] Zhi-Liang Hu, James M Reecy, and Xiao-Lin Wu. Design database for quantitative trait loci (qtl) data warehouse, data mining, and meta-analysis. In *Quantitative Trait Loci (QTL)*, pages 121–144. Springer, 2012.

[30] Nesreen Hamdallah Jboor, Abdelhak Belhi, Abdulaziz Khalid Al-Ali, Abdelaziz Bouras, and Ali Jaoua. Towards an inpainting framework for visual cultural heritage. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pages 602–607. IEEE, 2019.

[31] Huan Jin and Hunter NB Moseley. Robust moiety model selection using mass spectrometry measured isotopologues. *Metabolites*, 10(3):118, 2020.

[32] Payam Karisani, Zhaohui S Qin, and Eugene Agichtein. Probabilistic and machine learning-based retrieval approaches for biomedical dataset retrieval. *Database*, 2018, 2018.

[33] Staffs Keele et al. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, Ver. 2.3 EBSE Technical Report. EBSE, 2007.

[34] Thordis Kristjansdottir, Elleke F Bosma, Filipe Branco dos Santos, Emre Özdemir, Markus J Herrgård, Lucas França, Bruno Ferreira, Alex T Nielsen, and Steinn Gudmundsson. A metabolic reconstruction of lactobacillus reuteri jcm 1112 and analysis of its potential as a cell factory. *Microbial cell factories*, 18(1):1–19, 2019.

[35] David Lagorce, Olivier Sperandio, Jonathan B Baell, Maria A Miteva, and Bruno O Villoutreix. Faf-drugs3: a web server for compound property calculation and chemical library design. *Nucleic acids research*, 43(W1):W200–W207, 2015.

[36] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021.

[37] Peterson Lobato. My-SAE. Accessed: April 1st, 2022.

[38] Yue Lu, Yuguan Li, and Mohamed Y Eltabakh. Decorating the cloud: enabling annotation management in mapreduce. *The VLDB Journal*, 25(3):399–424, 2016.

[39] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2017.

[40] Simeone Marino, Yi Zhao, Nina Zhou, Yiwang Zhou, Arthur W Toga, Lu Zhao, Yingsi Jian, Yichen Yang, Yehu Chen, Qiucheng Wu, et al. Compressive big data analytics: An ensemble meta-algorithm for high-dimensional multisource datasets. *Plos one*, 15(8):e0228520, 2020.

[41] Nolwenn Maudet. Dead angles of personalization: Integrating curation algorithms in the fabric of design. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 1439–1448, 2019.

[42] Marvin M Mayerhofer, Falk Eigemann, Carsten Lackner, Jutta Hoffmann, and Ferdi L Hellweger. Dynamic carbon flux network of a diverse marine microbial community. *ISME Communications*, 1(1):1–10, 2021.

[43] Mohammad Mazharul Islam, Vinai C Thomas, Matthew Van Beek, Jong-Sam Ahn, Abdulelah A Alqarzaee, Chunyi Zhou, Paul D Fey, Kenneth W Bayles, and Rajib Saha. An integrated computational and experimental study to investigate staphylococcus aureus metabolism. *NPJ systems biology and applications*, 6(1):1–13, 2020.

[44] Jintao Meng, Peng Chen, Mohamed Wahib, Mingjun Yang, Liangzhen Zheng, Yanjie Wei, Shengzhong Feng, and Wei Liu. Boosting the predictive performance with aqueous solubility dataset curation. *Scientific Data*, 9(1):1–13, 2022.

[45] John A Miller, Hao Peng, and Michael E Cotterell. Adding support for theory in open science big data. In *2017 IEEE World Congress on Services (SERVICES)*, pages 71–75. IEEE, 2017.

[46] Samuel Miravet-Verde, Raul Burgos, Javier Delgado, Maria Lluch-Senar, and Luis Serrano. Fastqins and anubis: two bioinformatic tools to explore facts and artifacts in transposon sequencing and essentiality studies. *Nucleic acids research*, 48(17):e102–e102, 2020.

[47] Syed Aun Muhammad, Waseem Raza, Thanh Nguyen, Baogang Bai, Xiaogang Wu, and Jake Chen. Cellular signaling pathways in insulin resistance-systems biology analyses of microarray dataset reveals new drug target gene signatures of type 2 diabetes mellitus. *Frontiers in physiology*, 8:13, 2017.

[48] Anna Papadopoulou, Douglas Chesters, Indiana Coronado, Gissela De la Cadena, Anabela Cardoso, Jazmina C Reyes, Jean-Michel Maes, Ricardo M Rueda, and Jesús Gómez-Zurita. Automated dna-based plant identification for large-scale biodiversity assessment. *Molecular Ecology Resources*, 15(1):136–152, 2015.

[49] Jimyung Park, Seng Chan You, Eugene Jeong, Chunhua Weng, Dongsu Park, Jin Roh, Dong Yun Lee, Jae Youn Cheong, Jin Wook Choi, Mira Kang, et al. A framework (socratex) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database: development and usability study. *JMIR medical informatics*, 9(3):e23983, 2021.

[50] Mi-Hyun Park, Soo Kyung Koo, Jin-Sung Lee, Han-Wook Yoo, Jong-Won Kim, Hae Il Cheong, and Hyun-Young Park. Kmd: Korean mutation database for genes related to diseases. *Human Mutation*, 33(4):E2332–E2340, 2012.

[51] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*, pages 1–10, 2008.

[52] Vasileios C Pezoulas, Konstantina D Kourou, Fanis Kalatzis, Themis P Exarchos, Evi Zampeli, Saviana Gandolfo, Andreas Goules, Chiara Baldini, Fotini Skopouli, Salvatore De Vita, et al. Overcoming the barriers that obscure the interlinking and analysis of clinical data through harmonization and incremental learning. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:83–90, 2020.

[53] Amy Pham. *Surveying the state of data curation: a review of policy and practice in UK HEIs*. M.sc thesis, Dept. of Computer and Information Sciences, University of Strathclyde, August 2018.

[54] Protiva Rahman, Arnab Nandi, Courtney Hebert, et al. Amplifying domain expertise in clinical data pipelines. *JMIR Medical Informatics*, 8(11):e19612, 2020.

[55] D Reker. Active learning for drug discovery and automated data curation. *Artificial Intelligence in Drug Discovery*, 75:301, 2020.

[56] Rodney T Richardson, Johan Bengtsson-Palme, Mary M Gardiner, and Reed M Johnson. A reference cytochrome c oxidase subunit i database curated for hierarchical classification of arthropod metabarcoding data. *PeerJ*, 6:e5126, 2018.

[57] Ana Rodriguez, Isaac Crespo, Ganna Androsova, and Antonio del Sol. Discrete logic modelling optimization to contextualize prior knowledge networks using prunet. *PloS one*, 10(6):e0127216, 2015.

[58] Christopher FL Saarnak, Jürg Utzinger, and Thomas K Kristensen. Collection, verification, sharing and dissemination of data: the contrast experience. *Acta tropica*, 128(2):407–411, 2013.

[59] Zeyuan Shang, Emanuel Zgraggen, Benedetto Buratti, Ferdinand Kossmann, Philipp Eichmann, Yeounoh Chung, Carsten Binnig, Eli Upfal, and Tim Kraska. Democratizing data science through interactive curation of ml pipelines. In *Proceedings of the 2019 international conference on management of data*, pages 1171–1188, 2019.

[60] Yao Shi, Paloma L Prieto, Tara Zepel, Shad Grunert, and Jason E Hein. Automated experimentation powers data science in chemistry. *Accounts of Chemical Research*, 54(3):546–555, 2021.

[61] Tianhong Song, Sven Köhler, Bertram Ludäscher, James Hanken, Maureen Kelly, David Lowery, James A Macklin, Paul J Morris, and Robert A Morris. Towards automated design, analysis and optimization of declarative curation workflows. 2014.

[62] Cormen Thomas H, Leiserson Charles E, Rivest Ronald L, Stein Clifford, et al. Introduction to algorithms, 2016.

[63] James E Tomkins, Raffaele Ferrari, Nikoleta Vavouraki, John Hardy, Ruth C Lovering, Patrick A Lewis, Liam J McGuffin, and Claudia Manzoni. Pinot: an intuitive resource for integrating protein-protein interactions. *Cell Communication and Signaling*, 18(1):1–11, 2020.

[64] Alvaro Sebastian Vaca Jacome, Ryan Peckner, Nicholas Shulman, Karsten Krug, Katherine C DeRuff, Adam Officer, Karen E Christianson, Brendan MacLean, Michael J MacCoss, Steven A Carr, et al. Avant-garde: an automated data-driven dia data curation tool. *Nature methods*, 17(12):1237–1244, 2020.

[65] Genoveva Vargas-Solar, Gavin Kemp, Irving Hernández-Gallegos, Javier Espinosa-Oviedo, Catarina Ferreira da Silva, and Parisa Ghodous. Exploring and curating data collections with curare. In *Proceeding of the 35ème Conférence sur la Gestion de Données–Principes, Technologies et Applications*, 2019.

[66] Randi Vita, Bjoern Peters, Zara Josephs, Paula de Matos, Marcus Ennis, Steve Turner, Christoph Steinbeck, Emily Seymour, Laura Zarebski, and Alessandro Sette. A model for collaborative curation, the iedb and chebi curation of non-peptidic epitopes. *Immunome Research*, 7(1):1, 2011.

[67] Simon Waddington, Jun Zhang, Gareth Knight, Jens Jensen, Roger Downing, and Cheney Ketley. Cloud repositories for research data–addressing the needs of researchers. *Journal of Cloud Computing: Advances, Systems and Applications*, 2(1):1–27, 2013.

[68] Ian Walsh, Matthew SF Choo, Sim Lyn Chiin, Amelia Mak, Shi Jie Tay, Pauline M Rudd, Yang Yuansheng, Andre Choo, Ho Ying Swan, and Terry Nguyen-Khuong. Clustering and curation of electropherograms: an efficient method for analyzing large cohorts of capillary electrophoresis glycomic profiles for bioprocessing operations. *Beilstein journal of organic chemistry*, 16(1):2087–2099, 2020.

[69] Xin Wang and Jinbo Bi. Bi-convex optimization to learn classifiers from multiple biomedical annotations. *IEEE/ACM transactions on computational biology and bioinformatics*, 14(3):564–575, 2016.

[70] Yijie Wang and Xiaoning Qian. Joint clustering of protein interaction networks through markov random walk. In *BMC Systems Biology*, volume 8, pages 1–13. BioMed Central, 2014.

[71] Roel Wieringa, Neil Maiden, Nancy Mead, and Colette Rolland. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. *Requirements engineering*, 11(1):102–107, 2006.

[72] Lei Xing, Daniel S Kapp, Maryellen L Giger, and James K Min. Outlook of the future landscape of artificial intelligence in medicine and new challenges. In *Artificial intelligence in medicine*, pages 503–526. Elsevier, 2021.

[73] Chi Yang, Deepak Puthal, Saraju P Mohanty, and Elias Kougianos. Big-sensing-data curation for the cloud is coming: A promise of scalable cloud-data-center mitigation for next-generation iot and wireless sensor networks. *IEEE Consumer Electronics Magazine*, 6(4):48–56, 2017.

[74] Wenjie Ye, Yue Dong, and Pieter Peers. Interactive curation of datasets for training and refining generative models. In *Computer Graphics Forum*, volume 38, pages 369–380. Wiley Online Library, 2019.

[75] Shuya Yoshida, Fumiyoshi Yamashita, Takayuki Itoh, and Mitsuru Hashida. Structure-activity relationship modeling for predicting interactions with pregnane x receptor by recursive partitioning. *Drug metabolism and pharmacokinetics*, 27(5):506–512, 2012.

[76] Zhiming Zhao, Paul Martin, Paola Grosso, Wouter Los, Cees De Laat, Keith Jeffrey, Alex Hardisty, Alex Vermeulen, Donatella Castelli, Yannick Legre, et al. Reference model guided system design and implementation for interoperable environmental research infrastructures. In *2015 IEEE 11th International Conference on e-Science*, pages 551–556. IEEE, 2015.

[77] Hao Zhu. Big data and artificial intelligence modeling for drug discovery. *Annual review of pharmacology and toxicology*, 60:573–589, 2020.