# A Discipline-Enriched Dataset for Tracking the Computational Turn of European Universities

Federico Nanni
Data and Web Science Group
University of Mannheim
Germany
federico@informatik.uni-mannheim.de

Giulia Paci
Structural and Computational Biology Unit
EMBL Heidelberg
Germany
giulia.paci@embl.de

## ABSTRACT

In recent years, academic research appears to have been going through a methodological turning point. The discussion around the impact that computational methods will have on traditional fields of study has been the focus of many calls for papers and panels at established conferences. However, despite the high prevalence of this topic in the academic debate, it remains very challenging to assess whether academia as a whole has been actually adopting more digital resources and methods during the recent years. We are currently studying this topic by combining hermeneutic and text mining practices while analyzing one of the primary research output of European universities, namely doctoral theses. In this work, we present an enriched dataset we created for addressing this research questions and the first results of the analyses we have conducted so far.

**ACM Reference format:**
Federico Nanni and Giulia Paci. 2017. A Discipline-Enriched Dataset for Tracking the Computational Turn of European Universities. In *Proceedings of WOSP 2017, Toronto, ON, Canada, June 19, 2017,* 5 pages.
DOI: 10.1145/3127526.3127532

## 1 INTRODUCTION

During the last decades, academia seems to have experienced an unstoppable growth in the adoption of digital methods. Many argue that the impact of the use of computational resources, infrastructures and approaches has been challenging the traditional way we conduct research in sciences, social sciences and humanities [1, 8].

The "computational turn" that we can notice by looking at the programs of traditional conferences[1] or at the topics of research grants[2], is fostered by technological as well as cultural and socio-political reasons. As a matter of fact, on one hand the continuous growth in availability of digital datasets (from public genome data to open data provided by public administrations to collections of textual resources such as HathiTrust) together with the advancement in computational power and in machine learning approaches are playing a central role in fostering the adoption of digital technologies all around academia, from biology to sociology to history, and in supporting a re-thinking of our methodologies. On the other hand, the rhetoric that computer science companies such as Google and Facebook are fostering on the infinite potential of big data and artificial intelligence (which is constantly re-emphasized by the media) could have an impact on academia as well, by conditioning research focus, methods and collaborations.

**Overall Research Question.** While the rhetoric on the "computational turn" that academia is experiencing is easy to spot, what remains difficult to assess is whether academia as a whole has been actually adopting more digital resources and methods in research. The long-term goal of our study is to obtain a better understanding on whether, how and why different academic disciplines have been adopting computational resources and methods. We are currently working towards this goal by adopting a combination of text mining [13] and hermeneutic [12] approaches for the analysis of research practices at European Universities.

**Specific contribution.** In this work, we employ a text mining approach presented at the previous edition of WOSP [13] in order to examine a large collection of European doctoral dissertations (1980-2015) collected from the portal DART-Europe[3]. In particular, in this paper we present *a)* a large-scale discipline-enriched dataset we created and that we make completely available to the research community to allow further studies on the topic[4] and *b)* the first results of the analyses we have conducted so far on this annotated corpus.

In the next sections, we first give an overview of the related work. Next, we describe how we enriched the DART-Europe dataset and finally we present a few initial findings of our study.

## 2 RELATED WORK

Studying the recent past of higher institutions and understanding the role and influence of technological advancement over established research practices has already attracted the attention of different communities, which have addressed this topic with various methods and goals. In the next paragraphs we cover the most prominent areas and approaches.

**History of Higher Education.** The massive four-volume book series [17], directed by the European University Association, edited

[1] https://www.historians.org/annual-meeting/past-meetings/2015-resources-and-guides/digital-history-at-the-annual-meeting
[2] http://www.digitalmeetsculture.net/tag/horizon-2020/

[3] http://www.dart-europe.eu/basic-search.php
[4] https://federiconanni.com/computational-turn/

*WOSP 2017, Toronto, ON, Canada*

978-1-4503-5388-5/17/06…$15.00
DOI: 10.1145/3127526.3127532

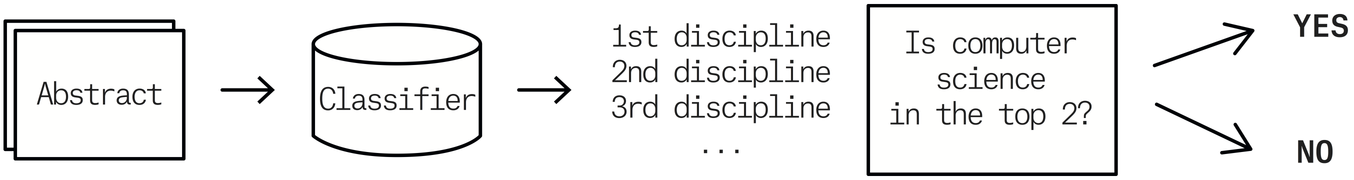

**Figure 1: Graphic representation of the adopted pipeline for interdisciplinarity detection.**

by Hilde de Ridder-Symoens and Walter Regg and published between 1992 and 2011, offers an unprecedented overview of how European universities have changed what they have taught and researched during the last centuries. To understand how new methods and practices spread across academia, in these studies, researchers adopt a large variety of sources and methods, from close reading of university-archive materials to the interpretation of the results of large-scale statistical analyses [7].

**History of Science and Technology.** The changes in teaching and research at academic institutions have also been studied by historians of science and technology, interested in understanding how scientific knowledge has moved back and forth between universities and the private sector and how political, economical and social actors have influenced scientific research in academia [16]. To address these research questions, researchers traditionally adopt a combination of methodologies, with particular attention to ethnography and anthropology practices in order to study, for example, the so-called "laboratory life" [10].

**Scientometrics.** A third perspective on universities and the changes in their research practices is offered by the scientometrics community [18]. The use of citation and co-citation measures has already given to researchers the possibility of comparing research trends in computer science across countries [5] or the advent of computational methods in biology [4].

**Text Mining and Scientometrics.** In addition to traditional bibliometric measures, more recently, a series of publications have focused on the use of word-based and topic-based approaches [9, 11], in order to expand the type of materials that could be analyzed and the perspectives of the topic. Text mining and machine learning techniques, such as Naive Bayes classifiers and LDA topic models, have been used to visualize and study interdisciplinary collaborations [2, 3, 14, 15] as well as to examine the change in research practices of specific domains, such as computational linguistic [6] and digital humanities [19].

## 3 ENRICHING DART-EUROPE

In this work we have adopted materials available on DART-Europe (Digital Access to Research Theses - Europe), a partnership of research libraries and library consortia who are working together in order to improve global access to European research theses. This portal offers over 700,000 theses from 28 European countries and 596 universities. In particular, in our study we have used a subcorpus of around 200,000 doctoral theses published between 1980 and 2015, which provide an abstract in English.

While this corpus presents an unprecedented amount of primary sources for researchers interested in the changes in research practices in academia, the available collection has a specific limitation. Only a very small part of the dataset has metadata regarding the discipline of the thesis. This limits both the navigation of the corpus and impedes any type of diachronic and discipline-based comparative study (such as studying the changes in biological research across the last thirty years).

**Table 1: Overall number of thesis and number of "computational" dissertations in our corpus, for the top 15 universities.**

| University | # of Theses | # of Comp. |
|---|---|---|
| Univ. of Groningen | 5254 | 344 |
| Univ. of Birmingham | 5217 | 411 |
| Univ. Aut. of Barcelona | 4871 | 448 |
| Wageningen Univ. | 4464 | 138 |
| Univ. of Uppsala | 4140 | 317 |
| Univ. of Rotterdam | 4072 | 325 |
| Univ. of Utrecht | 4031 | 318 |
| Univ. of Barcelona | 3846 | 205 |
| Univ. of Warwick | 3653 | 423 |
| Univ. of Padua | 3420 | 306 |
| Univ. of Athens | 3276 | 185 |
| Karolinska Institute | 3202 | 111 |
| Univ. of Manchester | 3177 | 485 |
| Univ. Of Thessaloniki | 2913 | 268 |
| Univ. of Helsinki | 2884 | 122 |

**Discipline Detection.** In order to address this issue and identify the most relevant disciplines of each thesis in our corpus, we have employed the Support Vector Machine presented by Nanni et al. [13]. As in previous work, each dissertation abstract has been represented as a TF-IDF vector.

We trained the supervised classifier on all theses abstracts from Italian universities (for a total of 11726): this subset of the collection provides information regarding the main discipline, as the abstracts are always paired with the subject area of the dissertation[5].

For assessing the quality of the classifier we evaluated it with 10-fold cross validation. We obtain a micro F1-Score of 0.72, which is consistent with the results obtained by Nanni et al. [13]. Moreover, as the addressed task can be also considered as a ranking problem, we report a Recall@1 of 0.72 and Recall@2 of 0.86. These results emphasize that the correct discipline appears to be in the top two results produced by our classifier in more than 85% of the cases.

**Macro Areas.** In order to support the identification of macro trends in the adoption of computational methods in academia, we first grouped disciplines considering the European Research Council

[5] More information regarding the disciplines is available in Nanni et al. [13].

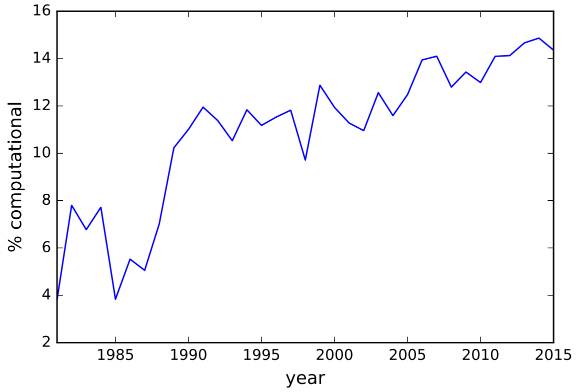

**Figure 2: Growth of computational theses, between years 1985 and 2015.**

(ERC) domains, namely: Physical Sciences and Engineering, Life Sciences, Social Sciences and Humanities[6].

**Computational Theses.** Finally, for determining whether a dissertation in our collection employs digital methodologies or not, we considered its main disciplines. We did so by employing the prediction confidence of the supervised classifier previously presented. If “Computer Science” (or “Computer Engineering”) appears to be one of the top two disciplines detected, we consider that thesis as having a “computational” aspect. From now on, we will call this thesis “computational”. Figure 1 illustrates the whole pipeline and in Table 1 we present some relevant statistics regarding the top 15 universities in our corpus.

# 4 QUANTIFYING THE COMPUTATIONAL TURN

Before going through our first results it is important to bear in mind that the dataset we employed offers only a partial view of the real output of European universities during the last thirty years. This is due to the fact that some universities do not appear in the DART dataset and for several (especially German) dissertations an abstract in English is not available. In addition to this, it is also important to consider that, even if the quality of our classifier is solid, it remains a machine learning approach with a margin of improvement. With this in mind, we present here the first results of our work.

## 4.1 Is Academia experiencing a computational turn?

The first analysis we conducted aimed at assessing whether academia has been generally going through a computational turn according to our enriched corpus. First of all, we detected that 13% of theses in our dataset have been labeled as “computational”. Moreover, as presented in Fig. 2, we noticed a constant growth of these theses over the decades, starting from less than 4% to around 15% of the total.

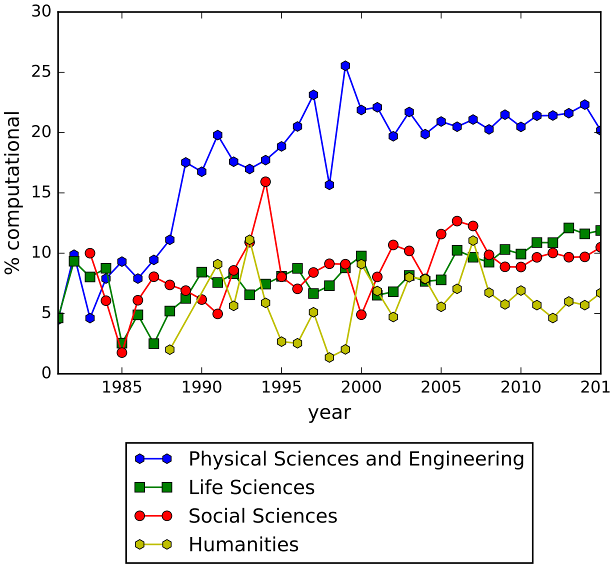

**Figure 3: Per-discipline growth, between years 1985 and 2015.**

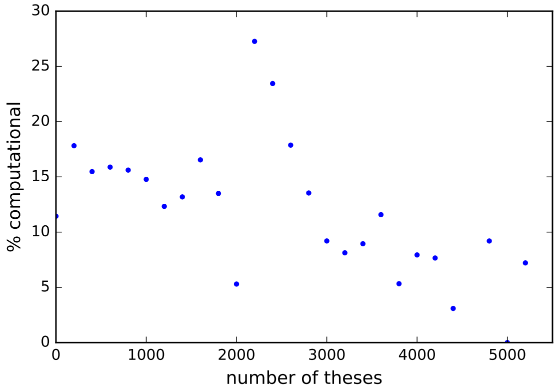

**Figure 4: Average percentage of computational theses (y axis) vs number of theses published in the institution (x axis), binned in intervals of size=200.**

## 4.2 How did different macro-areas react to it?

However, if we look at the mean percentage of computational theses per area, we can notice that this differs greatly, with Physical Sciences and Engineering having an average of 17% computational theses, Life Sciences 8%, Social Sciences 9% and Humanities 6%. More specifically, if we consider the single disciplines, the prevalence of computational theses in Physics, Biology, Linguistics and Economics over History, Classical Languages and Anthropology is evident. Furthermore, the time trend of computational methods

[6] In this work we distinguish between Social Sciences and Humanities, as opposed to the ERC domains, where they belong to the same one.

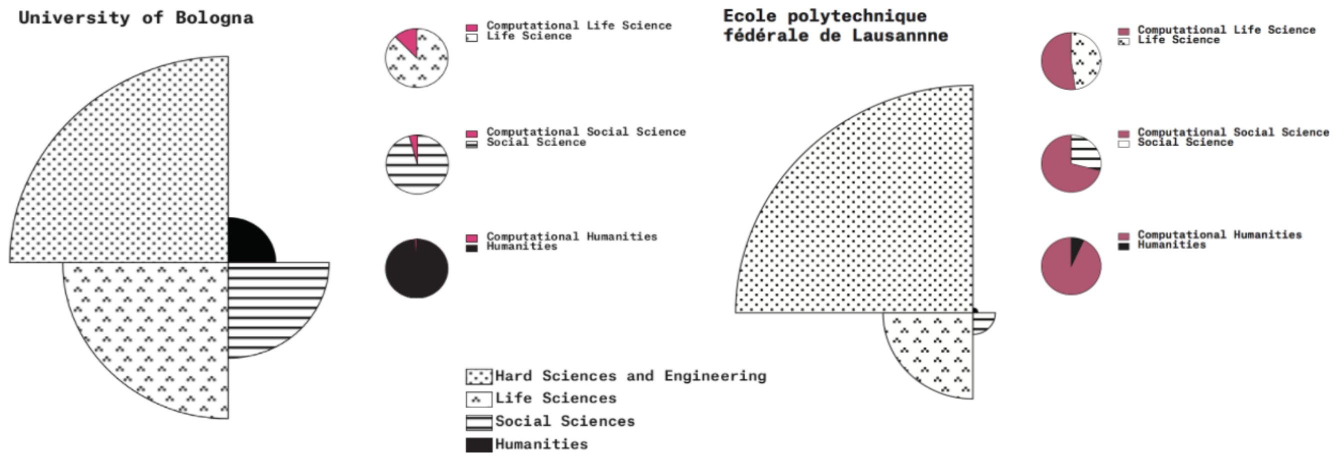

**Figure 5: Distribution of theses per ERC domain (top graph) and corresponding distribution of computational theses for relevant domains. Comparison between University of Bologna and Polytechnique of Lausanne.**

adoption also differs across macro-areas (see Figure 3): for instance, Physical Sciences and Engineering started the earliest, experienced a quick and steep growth between the 80s and the 90s, and have been relatively stable in the last ten years. The Humanities, on the other hand, started in later years and still present a very unstable profile, with less clear growth trends.

## 4.3 What kind of universities are fostering it?

The dataset at our disposal offers different points of view for answering this question. The results we report here are based on examining the relationship between the number of theses in Life Sciences, Social Sciences and Humanities published by each institution and the percentage of which are classified as computational. We focus on these three macro-areas as Physical Sciences and Engineering has already shown a clear pattern in previous analyses.

Two interesting aspects emerge from this preliminary study (see Fig. 4). First of all, we can notice an inverse relationship between the number of theses published in Life Sciences, Social Sciences and Humanities by a specific institution and the percentage of which are computational. This means that small and middle size institutions (which publish a smaller number of theses) seem to focus more on adopting computational methods in research. By examining closely the universities that publish more computational theses, we find that these are science and technology institutes such as the Telecom ParisTech, the Ecole Polytechnique Federale de Lausanne (see also Fig. 5), TU Delft, the Universitat Politecnica de Catalunya and the KTH Royal Institute of Technology in Stockholm. This seems to imply that, when research in Life Science, Humanities and Social Sciences is conducted at these tech-oriented institutions, it usually involves a computational aspect. These institutions would seem to be the driving force of the computational turn, as opposed to larger, more traditional, universities.

## 4.4 Towards a New Type of Comparative Study

In order to understand what kind of computational research is performed at these universities (is it really interdisciplinary research, or is it simply applying computational methods on a research task from a different field?) and which are the factors that foster these interdisciplinary projects, we envision further applications of the enriched dataset presented in this paper to support new types of comparative study between academic institutions (e.g., Fig. 5). This approach could also support the community in understanding the role that private and public research funds have played in orienting academic research in this direction and how traditional institutions have been dealing with the advent of computational methods and the growth of their application in academia.

# 5 CONCLUSIONS

During the last decades, academia seems to have experienced an unstoppable growth in adoption of digital technologies. Many argue that the impact of the use of computational resources, infrastructures and methods has been challenging the traditional way we conduct research in sciences, social sciences and humanities. In order to support studies on this “turn” in research practices, in this work we have presented how we enriched the DART-Europe dataset with disciplines and macro-area labels and how we used this new resource for an initial series of analyses on the topic. Our preliminary analyses support an increase in the adoption of computational methods in academia, albeit with large differences between research areas and types of institutions. Our enriched dataset allowed us to further investigate what institutions have fostered the computational turn the most, highlighting the important role of small and medium scale research centers in science and technology. The enriched database will allow further studies from scholars interested in better understanding the recent past of European academic institutions.

# REFERENCES

[1] David Berry. 2011. The computational turn: Thinking about the digital humanities. *Culture Machine* 12 (2011).

[2] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 443–452.

[3] Theodoros Giannakopoulos, Ioannis Foufoulas, Eleftherios Stamatogiannakis, Harry Dimitropoulos, Natalia Manola, and Yannis Ioannidis. 2014. Discovering and Visualizing Interdisciplinary Content Classes in Scientific Publications. *D-Lib Magazine* 20, 11 (2014), 4.

[4] Jiancheng Guan and Xia Gao. 2008. Comparison and evaluation of Chinese research performance in the field of bioinformatics. *Scientometrics* 75, 2 (2008), 357–379.

[5] Jiancheng Guan and Nan Ma. 2004. A comparative study of research performance in computer science. *Scientometrics* 61, 3 (2004), 339–359.

[6] David Hall, Daniel Jurafsky, and Christopher D Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of EMNLP*. Association for Computational Linguistics, 363–371.

[7] Susan Hockey. 2004. The history of humanities computing. *A companion to digital humanities* (2004), 3–19.

[8] Rob Kitchin. 2014. Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1, 1 (2014), 2053951714528481.

[9] Petr Knoth and Drahomira Herrmannova. 2014. Towards semantometrics: A new semantic similarity based measure for assessing a research publicationfis contribution. *D-Lib Magazine* 20, 11 (2014), 8.

[10] Bruno Latour and Steve Woolgar. 1979. *Laboratory life: The construction of scientific facts*. Princeton University Press.

[11] Kun Lu and Dietmar Wolfram. 2012. Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology* 63, 10 (2012), 1973–1986.

[12] Federico Nanni. 2017. Reconstructing a website's lost past - Methodological issues concerning the history of www.unibo.it. *Digital Humanities Quarterly* (2017).

[13] Federico Nanni, Laura Dietz, Stefano Faralli, Goran Glavaš, and Simone Paolo Ponzetto. 2016. Capturing interdisciplinarity in academic abstracts. *D-lib magazine* 22, 9/10 (2016).

[14] Leah G Nichols. 2014. A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics* 100, 3 (2014), 741–754.

[15] Francesco Osborne and Enrico Motta. 2013. Exploring research trends with rexplore. *D-Lib Magazine* 19, 9/10 (2013).

[16] Giuliano Pancaldi. 1993. *Le Università E Le Scienze: Prospettive Storiche E Attuali: Relazioni Presentate Al Convegno Internazionale, Bologna, 18 Settembre 1991*. Università di Bologna.

[17] Walter Rüegg. 2011. *A history of the university in Europe: Volume 4, Universities Since 1945*. Cambridge University Press.

[18] A Van Raan. 1997. Scientometrics: State-of-the-art. *Scientometrics* 38, 1 (1997), 205–218.

[19] Scott B. Weingart. 2015. Acceptances to DH2015 (pt. 2). (2015).