Decoding Academic Departments: A Publication-Centric Methodology

Daniel Eskandar [*1] **Mika Thormann** [*2] **Peiru Fang** [*3] **Simon Hanrath** [*4]

Abstract

This paper introduces a methodology for analyzing and visualizing research dynamics within specific academic groups, like university departments, using publication data alone. The analysis, applied to the Machine Learning Department at the University of Tübingen, occurs in two stages: basic data analysis of publication metrics and advanced analysis using word embeddings to understand research areas and connections within the group. Using this method we highlight trends in Tübingen's Machine Learning Department, revealing dominant research fields and rising areas within the group, demonstrating the method's utility and potential for wider application.

1. Introduction

In the realm of academic research, particularly in computer science, platforms like CSRankings.org offer valuable high-level insights into the global research landscape, enabling users to assess universities' research output and standing. While these resources offer a broad overview, they lack detailed insights into the work and collaborations of departments or individual researchers. This gap in granular analysis is the primary focus of this work.

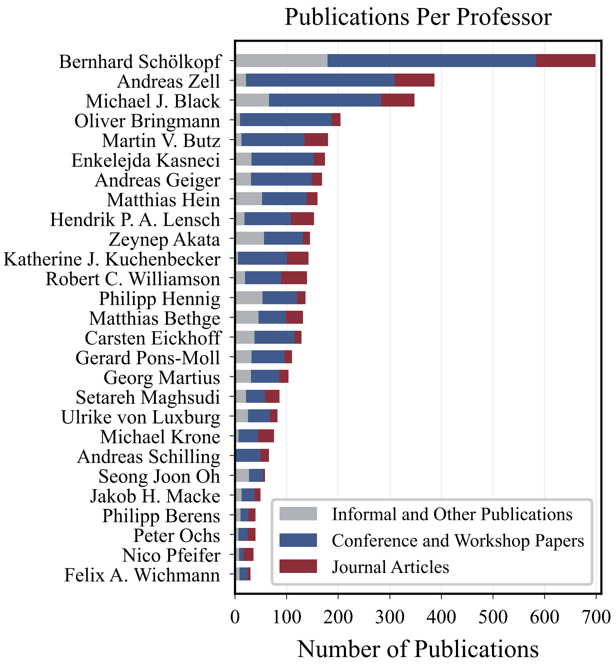

Figure 1. Overview of the selected scholars from the University of Tübingen's Machine Learning Department and the total amount of publication over their careers. The publication output among the scholars varies widely.

We present a methodology to analyze and visualize research dynamics within specific groups. Our approach relies entirely on the analysis of publication data, with an emphasis on extracting information from paper titles and co-authorship. The analysis is conducted in two stages. The first stage involves basic data analysis, focusing on straightforward metrics such as the number of publications per scholar and the department's paper output over time (Section 2.1). Here we also construct collaboration networks based on co-authorship data. The second stage involves generating word embeddings for scholars and their papers, aiming to derive detailed insights into their research areas and connections within the group (Section 2.2). This method promises a comprehensive understanding of the research themes, interconnections, and collaborative patterns among scholars, without relying on any other external data sources. The advantage of this methodology is its scalability and adaptability to different research groups and fields.

We apply this technique to the Machine Learning department at the University of Tübingen to analyze research themes and collaborations among its faculty members. By doing so, we illuminate the department's intellectual landscape and demonstrate the potential of our method.

[*]Equal contribution [1]6660442, daniel.eskandar@student.uni-tuebingen.de, MSc Machine Learning [2]5465950, mika.thormann@student.uni-tuebingen.de, MSc Machine Learning [3]5907175, peiru.fang@student.uni-tuebingen.de, MSc Informatik [4]5421174, simon.hanrath@student.uni-tuebingen.de, MSc Machine Learning.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023. Copyright 2023 by the author(s).

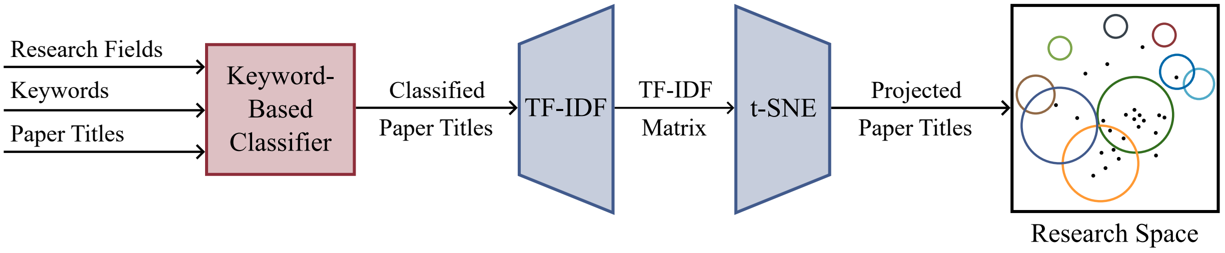

Figure 2. Method: Employing a keyword-based classifier, we classify paper titles, considering prior research fields and keywords. Subsequently, we construct a TF-IDF matrix, and apply t-SNE for 2D dimensionality reduction, forming a detailed research space.

2. Data and Methods

Using data from the Digital Bibliography & Library Project (DBLP) (Ley, 2023), we construct a dataset encompassing all 4,066 publications from 27 scholars at the University of Tübingen. This dataset includes publication titles, co-authors, venues, citation counts, and release years. Figure 1 illustrates the distribution of the publications among the selected scholars.

2.1. Initial Data Metrics & Co-authorship Analysis

The first step in our analysis is a direct examination of the dataset, focusing on readily extractable features. We focus on the number of publications over time for both the entire group of scholars and for each scholar, as this offers a high-level overview of the group's academic output. The next interesting feature to look at is the co-authors of each paper. With this information, we can construct a network mapping the collaborations between the selected scholars. This network visualization aims to show collaborative patterns and thematic connections among researchers.

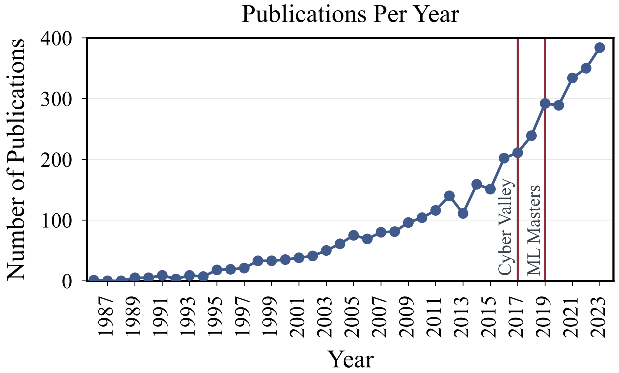

Figure 3. Trend of AI publications of the selected group of scholars, revealing a significant recent increase in academic output. The red lines mark the beginning of the Cyber Valley initiative and the Machine Learning Masters program respectively.

2.2. Paper Title Analysis

Collaboration patterns alone offer a limited view of scholarly contributions, particularly within the same institute. For a more comprehensive understanding, exploring researchers' chosen topics and their evolution over time is crucial. Traditional topic modeling techniques, such as Latent Dirichlet Allocation (Blei et al., 2003), struggle to capture meaningful research topics from short texts like paper titles.

To address this challenge, we manually compiled a set of research fields, drawing guidance from the fields associated with each scholar on CSRankings. This information serves as a prior for multi-label zero-shot classification of paper titles to research fields. A keyword-based approach is employed, associating a list of keywords with each research field. This allows for the determination of the probability of a paper belonging to each research field based on the presence of field-related keywords in the title.

With classifications established for paper titles, a comprehensive analysis of trends and prevalence in different research fields becomes feasible. Moving beyond this, a deeper understanding of the research landscape requires the exploration of scholars' work evolution alongside the dynamics of research fields.

To achieve this, a 2D research space is constructed by projecting papers, scholars, and research fields onto a plane. This involves the creation of a TF-IDF matrix (Spärck Jones, 1972) for the entire corpus incorporating both the paper titles and the concatenation of their respective classified research fields. Following this, dimensionality reduction is carried out using t-SNE (Van der Maaten & Hinton, 2008) on the TF-IDF matrix. The 2D vector representation of each paper constructs the research space. Colored points denote papers in diverse fields, with scholar and field points averaging associated and related papers, respectively. The field's radius signifies the mean distance to all points within it. Figure 2 encapsulates the method employed for transitioning from paper titles to the research space.

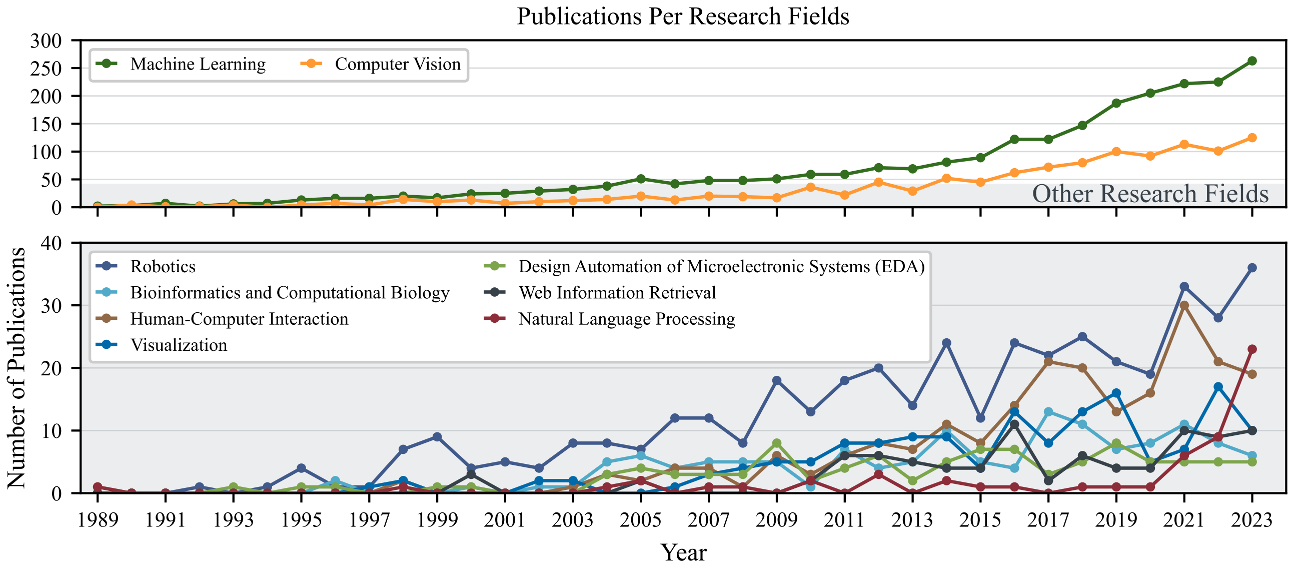

Figure 4. Break down of publications from the selected group over time into sub-categories. We see a strong focus in areas such as Machine Learning and Computer Vision, with steady, albeit lesser, activity in Robotics and Human-Computer Interaction. Additionally, the field of Natural Language Processing has shown increased activity over the past three years.

3. Results

We now apply our method to the selected scholars from the University of Tübingen's Machine Learning Department. Our analysis of publication trends, as detailed in Section 2.1, reveals an increase in academic output in recent years. This surge, particularly after the launch of the Cyber Valley initiative, is graphically represented in Figure 3. For an in-depth year-by-year breakdown of individual scholars' publications, readers are directed to our GitHub repository[1].

Figure 6 displays the co-authorship patterns within the observed group. While we see some strong connections like between M. Black, G. Pons-Moll and A. Geiger, the plot also reveals that most scholars have only co-authored less than 15 % of their publications with group members. Notable exceptions from this are F. Wichmann, J. Macke, and P. Berens. This is, however, explained by the fact that they generally have not published as many papers as the other scholars, as we can see in 1.

Using the title analysis methodology described in Section 2.2, we can categorize the total research output in Figure 3 into distinct fields. The selection of the sub-fields is based on a manual review of the scholars' profiles. With this approach, we can classify 91% of the data into one of the selected categories. Categorizing the individual publications into these fields leads to a more granular breakdown shown in Figure 4. Our findings indicate a focus on Machine Learning and Computer Vision, with steady, albeit lesser, activity in Robotics and Human-Computer Interaction. Notably, Natural Language Processing has increased in activity over the past three years.

The most revealing aspect of our analysis is the exploration of the research space, as discussed in Section 2.2. Figure 5 illustrates the research interests and connections among the selected scholars from the University of Tübingen. Additionally, a dynamic visualization of research interests in the group over time is available in our GitHub repository[1]. We see that the majority of scholars are concentrated in the two fields: Machine Learning and Computer Vision. However, we also observe some scholars focusing on other fields such as Robotics, Web Information Retrieval, or Bioinformatics. When looking at clusters such as in Computer Vision, we see that they correlate with increased collaboration among these researchers, consistent with the collaboration network in Figure 6. This is exemplified by the collaboration among M. Black, A. Geiger, and G. Pons-Moll. Generally, proximity in research space consistently translates into closely-knit collaborative partnerships.

4. Discussion & Conclusion

This paper introduced a method for examining academic output and collaboration trends within research groups. The analysis relies entirely on the publications of the selected group. This ensures access to up-to-date information. The approach also benefits from being broadly applicable, given that the majority of scholars publish papers following a uniform format. Our research offers insights into the publi-

[1] GitHub Repository: ScholarInsights. Available at: `https://github.com/SimonHanrath/ScholarInsights`

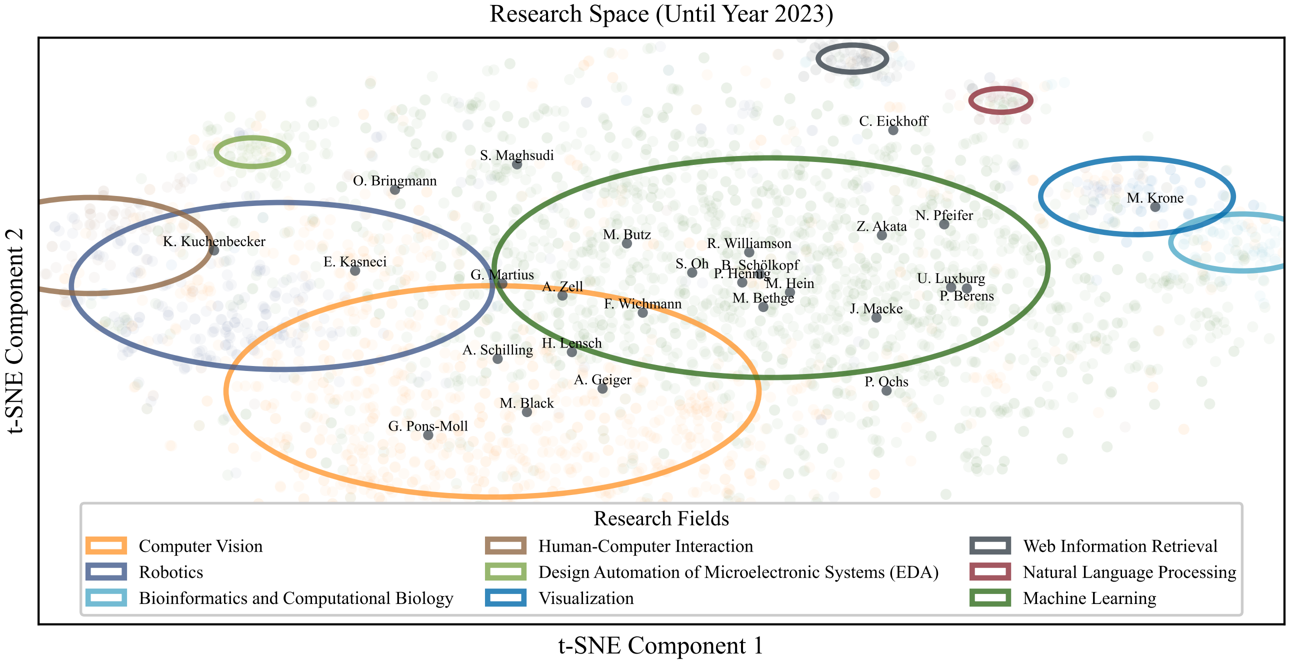

Figure 5. Figure 5. Visualization of the research space of Tübingen's Machine Learning Department until 2023. The closeness between scholar dots represents closeness in research themes. The transparent circles represent the published papers projected into the vector space. The large circles represent the research areas, showing that most scholars are working on Machine Learning or Computer Vision. However, exceptions such as M. Krone in visualization or O. Bringmann at the intersection of EDA, Robotics, and Machine Learning do exist.

cation trends, collaboration networks, and thematic focuses of scholars. The research space visualization, as illustrated in Figure 5, stands out as a particularly effective tool, encapsulating extensive data in a single, comprehensible plot. Although promising for broader application, this visualization method is contingent on the pre-definition of research fields. This limitation appears to be an inherent characteristic of the selected kind of data, presenting a challenge in unsupervised categorical label derivation, as mentioned in Section 2.2. For the direct application on the selected group of Tübingen's Machine Learning Department, we highlighted trends in overall research output as well as for individual research fields and scholars. We showed a concentration in Machine Learning and Computer Vision, while also identifying activity in smaller fields like Robotics, Bioinformatics, and Visualization, as well as emerging areas such as Natural Language Processing.

Contribution Statement

Peiru Fang collected and prepared the data. Mika Thormann and Daniel Eskandar mostly worked on the code for the analysis and the visualization. Simon Hanrath (with assistance from the team) worked on the text and the structure of the report.

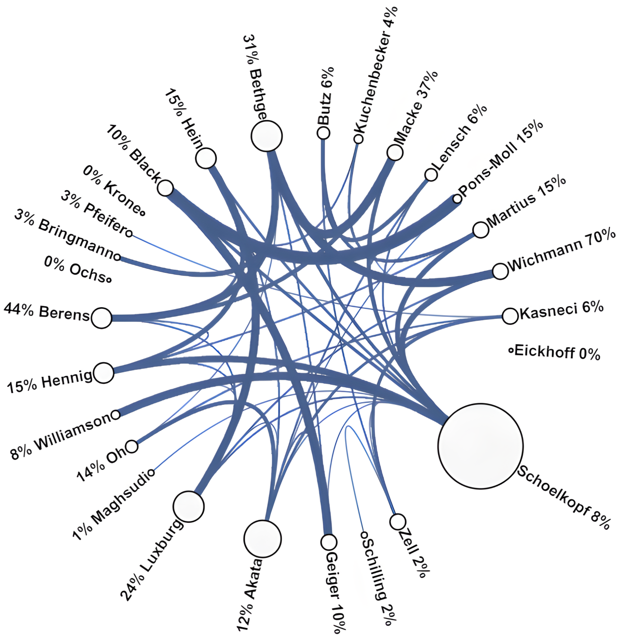

Figure 6. Collaborative publications among scholars. Circle sizes represent total joint publications, adjacent percentages indicate the proportion of each scholar's total publications from collaborations, and connection thickness shows the extent of collaboration between pairs on a scale from 0 to 16 publications.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3 (Jan):993–1022, 2003.

Ley, M. The dblp computer science bibliography. https://dblp.org, 2023. Accessed on: 03.12.23.

Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.