How to Find New Industry Partners for Public Research: A Classification Approach

Karl Trela, Yuri Campbell, Friedrich Dornbusch, and Anna Pohle

***Abstract*—Finding new industry partners poses a challenge to many public research organizations. This article explores how statistical classification can support partner selection at the example of the Fraunhofer Society in Germany, Europe's largest public organization for applied research. We use internal cooperation data and feature sets based on unstructured data, i.e., text and industry codes, both of which describe business activities of firms. An important advantage of this data is that it is available for most companies in Germany, even small and medium enterprises, which allows for an almost complete screening of the market, in contrast to using other data sources, e.g., patents. In addition, we also include economic variables linked to firms, as turnover, number of employees/managers and firm age. We report the performance of various classification techniques such as logistic regression, support vector machines, and random forests in our dataset for diverse combinations of feature sets. Results show that simple methods with fewer parameters remain competitive in comparison to complex ones. Overall, the performance of most classifiers is high enough to support the decision process of finding new industry partners for public research.**

***Index Terms*—Collaborative research, knowledge and technology transfer, machine learning, natural language processing, partner selection, text mining, university–industry cooperation.**

I. Introduction

POLICY as well as society increasingly demand from academic institutions to transfer their knowledge to industry in order to leverage scientific discoveries and inventions into innovation [1]. Consequently, research organizations strive for ways to increase their knowledge and technology transfer activities in form of collaborations with industry partners [2], [3]. On the industry side, a rising complexity in production and its innovation leads to blurring cross-organizational boundaries. It forces companies to extend their networks way beyond their region for knowledge and technology sourcing [4], [5]. Hence, the search for new partners and the identification of the right partners is crucial for both industry as well as academia. While this is nowadays more true than ever, it is also long known to be a hard task to solve [6].

A number of studies have shown that partnerships between university and industry face a number of challenges. Besides building trust, overcoming cultural and organizational differences [2], [7], the search for suitable collaboration partners for firms and public research organizations alike is usually a hurdle, known as the "partner selection problem" [8]. As cooperation capability is limited in every organization, partner selection should be based on resource complementarity among partners [9], [10]. Hence, this comes down to a matching problem with regard to technological-scientific and economic fit. This, however, is also difficult to solve.

In both, academia and business, decision-makers are limited with regard to time, resources, and their capability to oversee all available cooperation opportunities. This favors *ad hoc* approaches for partner selection [8], which are usually based on past successful collaboration projects, personal contacts, or subjective opinion of experts [11], [12]. Particularly small and medium enterprises (SMEs) tend to employ social and often locally biased networks to find suitable partners [13]. This leaves innovation potential untapped, both from the perspective of a single organization as well as from society and policy.

Though human interaction is crucial in building trustful and stable relations within R&D alliances [14], we want to evaluate whether seeking for partners and making decisions can become faster, more efficient and better informed with the help of information technology. This is where the field of Tech Mining comes into play. Porter and Cunninghum define Tech Mining as "the application of text mining tools to science and technology information, informed by understanding of technological innovation processes" [15]. Therefore, Tech Mining has two main components, first, the use of text mining approaches, and therefore also of natural language processing (NLP), and second, the application of such tools and approaches to technology management. In turn, we argue that technology sourcing, and the knowledge sourcing related to it, are naturally tasks of technology management. Hence, the partner selection problem above is also a part of it, as long as the search for partners is based on technological fit.

Several studies before have tested the feasibility of defining and applying new systematic methods on R&D-partner identification in general, some of them are "intelligent" approaches based on machine learning techniques and NLP [16]–[18]. Our approach, however, differs in a crucial way. Existing approaches mainly employed patent or bibliometric

Manuscript received June 30, 2019; revised October 31, 2019 and February 22, 2020; accepted April 3, 2020. This work was supported by the German Federal Ministry of Education and Research BMBF under Grant 03IO1624. Review of this manuscript as arranged by Department Editor Y. Zhang. *(Corresponding author: Karl Trela.)*

The authors are with the Fraunhofer Center for International Management and Knowledge Economy, 04109 Leipzig, Germany (e-mail: karl.trela@imw.fraunhofer.de; yuri.campbell@imw.fraunhofer.de; friedrich.dornbusch@imw.fraunhofer.de; anna.pohle@imw.fraunhofer.de).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TEM.2020.2992060

data in order to create proxy-indicators for cooperation potential. The employed methods offer a way of summarizing and navigating possible matches of cooperation partners based on their technological portfolio, estimated through patent data. While this approach is sound for explorative analysis with focus on understanding technological proximity of potential partners, it has very restricted applicability in predictive analysis and no formal way of measuring performance, which are common problems to unsupervised learning approaches in general.

In contrast, our method relies on large-scale real historical cooperation data from a knowledge transfer context, in this case, provided by the Fraunhofer Society. Hereby, we have the opportunity to predict and test cooperation potential based on real cooperation behavior. While this enables us to cover a larger range of innovation activities in R&D in comparison with patent data based analysis, it also enables us to perform supervised learning, i.e., classification, and calculate objective performance measures on how the various partner selection algorithms generalize in reality.

As far as we know, this is the first article that addresses the sphere of cooperation between academia and industry from such a supervised learning perspective. For that, we extract information from data of qualitative nature, such as text and industry codes, using NLP techniques and integrate it with traditional econometric models. Such a combination of qualitative and quantitative data in partner selection has been called for since a long time [19], [20].

Our main goal is to test whether classification techniques along with NLP and economic data can help to find industry partners for public research. For this, we present a systematic approach, which includes prescreening a large number of companies for potential collaboration partners and ranking recommended companies according to a statistical measure. It makes use of supervised learning by stacking two classifiers, which are trained on past historical cooperation data as target variables together with text and economic indicators as dependent variables.

The rest of this article is organized as follows. The following section introduces the research questions by elaborating the search behavior in innovation activities and reviews previous quantitative approaches to address the partner selection problem. Section III presents the data, the methods, and the validation strategy. Section IV describes and discusses the results. Finally, we provide a discussion of limitations in Section V, and the practical use of the approach is described in see Section VI. Finally, Section VII concludes the article.

II. Partner Selection in Technology Transfer

The following subsection first aims to provide an in-depth understanding of the partner selection problem in technology transfer. It begins by outlining the structure of the industry-partner-selection process and the challenges it poses, thereby elaborating on the need for systematic approaches to support partner selection. The succeeding subsection presents an overview of systematic methodologies introduced so far. Then, it discusses the econometric determinants of university–industry cooperation and how these may contribute to a decision support system. Finally, we present our research questions as a synopsis of the advantages and drawbacks of the existing approaches.

A. Bounded Rationality and Heuristic Search Behavior in Partner Selection

The literature discusses partner selection in the context of collaborations, especially supply chains and research alliances [19], [20]. From a research organization point of view, *industry-partner selection* describes the process of searching and deciding for a partner to transform a research result into an innovation. This process can be divided into four stages [19]:

1) criteria formulation;
2) qualification;
3) final selection;
4) application feedback.

The *criteria formulation* stage involves the determination of criteria as well as preparation of subsequent steps. This stage tends to be characterized by uncertainty and vague information about partners and their cooperation potential. The *qualification* stage relates to reduction and narrowing down of potential partners to a smaller subset with higher cooperation potential. This involves the ranking of potential partners according to an either implicit or explicit matching profile. In a third step, the *final selection* takes place and the most relevant and suitable partners are chosen [19]. Finally, one evaluates the results of the selection process and makes adjustments in the *application feedback* stage.

This rather holistic description already shows that the main challenges in partner selection are, first, getting information and knowledge about potential partners and second, reducing the amount of available information to a level that supports efficient and well-grounded decision-making. The main reason here is that complexity in decision-making relates to the amount as well as quality—in a sense of validity and completeness—of information that needs to be gathered and processed. While scarce, vague, or biased information raises uncertainty [22], large amounts of information can also bring decision makers to the limits of their cognitive capacity.

Accordingly, bounded rational actors and individuals deal only with a biased set of information, because the search for complete information is costly and time consuming [13]. To deal with such limited or imperfect information, individuals often use heuristics. Those, besides being in some cases a good and efficient way to handle complexity can often lead to suboptimal or poor solutions [23], [24]. The basic argument here is that individuals do not perform exhaustive search processes across an entire search space, but prefer—if other more distant solutions are not known—the spatially and socially most proximate and cognitively satisfying solutions. Thus, applying heuristics in search processes directs the focus of decision-makers toward their existing social networks that are—favored by spatial proximity and face-to-face interactions—often biased. As a consequence, they tend to cooperate within existing networks and gathering information about new partners depends rather on word-of-mouth and relies mainly on subjective judgment, instead of quantitative factors [25].

Even though partner selection is based on existing social networks, trust and experiences are obviously the basis for network development and partner selection [3], [26], it can also be problematic. A long strand of literature has shown theoretically as well as empirically that it can cause path dependence on the one hand and local as well as social lock-in effects [27]–[29] on the other hand. Moreover, the source of new innovation-enabling knowledge, however, is most likely to be distributed worldwide and across industries. Organizations, as well as regions, that are able to identify new knowledge partners and tap new sources of knowledge efficiently, will gain a substantial competitive advantage. However, the information about adequate new partners and sources is still often collected and processed manually (from firm databases and web searches); therefore, it is difficult and time-consuming to obtain. Thus, we argue that quantitative and machine-learning-based approaches, using available data, can contribute to network development and partner selection by making the search process more efficient. A systematic process for searching technology partners is needed and would help organizations as well as regions to maintain competitiveness in an open innovation paradigm [12], [30].

B. Systematic Approaches With a Priori Assumptions

Recently, a considerable body of work dealing with the design of systematic, or semiautomated, approaches for supporting partner selection has been proposed.[1] With regard to the level of data-usage, these methods range from purely expert rating approaches, all the way to databased methods, which make heavy use of machine learning techniques. In contrast to the heuristic procedures presented in the preceding section, systematic approaches seek to operationalize criteria for partner selection in a formal way. Two methods can be distinguished within the systematic approaches: (expert) rating methods and bibliographic methods.

For partner selection with rating methods, defined criteria are applied to identify appropriate partners according to diverse perspectives (stage 1 of industry-partner selection). Those perspectives, and in turn related criteria, are often defined by experts and hard-coded in ranking strategies [16], [31]. They are fed into a weighting model or algorithm, which ranks options (stage 2).

Bibliographic methods make use of bibliographic data in order to support partner selection. These works apply patent and publication data statistics in order to find and assess how well two potential R&D partners fit together (stage 2). The statistics used for the fitness estimation varies from contribution to contribution, but falls usually into two main categories. The first is composed mainly of bibliometric analysis [11], [25], for example, bibliographic coupling, citation networks, coauthorship, keyword cooccurrence, etc. The second and most modern approach uses unstructured data processing methods such as text mining and semantic analysis [8], [12], [18] for this end. In the latter, keywords are preselected by experts and their frequency is encoded in vectors, one vector per document, e.g., a patent. The similarity between text documents is computed using cosine distance, a distance function between two keyword-frequency vectors, which is the elemental method in NLP and information retrieval. Based on the similarity between pairs of documents, partners are suggested. This suggestion depends partially on handcrafted criteria, such as keyword selection, and on how to use the similarity measure between documents. Overall, the goal of bibliographic methods is the same, to reduce information complexity in the bibliographic data in order to allow the exploration of partnership possibilities by the decision maker, usually also in a data visualization form, as a graph [17], [18], [25] or a lattice [8]. This is therefore an unsupervised learning approach.

The two approaches, rating methods and bibliographic methods, share that the partner preselection and ranking are essentially determined by *a priori* assumptions in the models. The various methods, however, differ in the expression of these assumptions. While rating methods rely solely on the operationalization of expert knowledge. Bibliographic methods convey assumptions on different levels. At times, it assumes that technological distance proxies technological fit for partnership. At other times, it relies on experts in order to compose keywords lists and their fitness.

In summary, a major issue with these methods are their explicit assumptions about the criteria that define technological fit i.e., that determine partner preselection. A complementary approach would apply observation data from cooperations to rely on statistical measures rather than assumptions in order to proxy technological fit in a supervised learning setting. Consequently, no efforts in criteria formation would be required any more. Such an approach, however, would not convey an explicit measurement of technological fit. On the other hand, a databased approach would allow to calculate performance measures to determine precisely the quality of the partner selection support solution. This would not only reduce complexity in the qualification stage (stage 2) but also in application feedback (stage 4). The former methods cannot be evaluated in a comparable way. Moreover, while these previous approaches can reduce complexity in the qualification stage (stage 2), they leave the hurdle in criteria formation untouched, as well as in application feedback.

Finally, the studies above focus on partner selection either in supply chain or in the larger context of open innovation. None of those deals directly with the issue of R&D partnership in form of contracted research between public research organizations and companies, in other words, in a knowledge transfer context.

Considering the drawbacks and research needs presented in this subsection, we want to examine Research Question 1:

RQ1: Can a supervised learning approach based on cooperation data support partner selection in public research?

[1]To perform the literature analysis, we have used Scopus (the largest source-neutral abstract and citation database for peer-reviewed literature) as well as the OECD library. In the first step, we mainly looked for the literature reviews in the topics "partner selection" and "university–industry collaboration" to gain a comprehensive overview. We used the search string: TITLE ((“partner selection”) AND (problem OR theory OR review OR innovation OR r&d OR decision)). Thereafter, we repeated the search with a focus on methods and success factors for partner selection and collaboration in the research about academic knowledge and technology transfer as well as the literature on innovation management. In addition, we followed the snowball principle to find more literature that is relevant and asked colleagues for paper recommendations.

C. Econometric Determinants of University–Industry Cooperation

The previous subsection has shown that systematic approaches have the potential to support the identification of industry partners and to moderate the cognitive bias inherent to the search process. In turn, several empirical studies deal with the identification of determinants of university–industry cooperation.[2] Instead of aiming at decision support in partner selection, they analyze real cooperation data and investigate the effect of a specific set of variables on the cooperation activity between universities and firms. For that end, they make use of standard statistic or econometric methods. By doing so, they are bypassing the shortcomings of present systematic approaches and can therefore be relevant for our endeavor.

Giunta *et al.* [32] investigated determinant indicators of university–industry cooperation in the biopharmaceuticals in Italy over a six years period between 2004 and 2010. In this article, a logistic regression (LR) model is constructed to estimate a binary dummy variable that signalizes if a given pair (university, industry) have coauthored an academic research article. The authors show an expected dependence of variables as for example prior partnership and geographical proximity on the probability of cooperation. Other works point out that the size, of institution and firms, and the firm's R&D and patents expenditure exert a significant impact on the cooperation activities [33], [34]. Further studies show that conducting R&D activities itself [35], [36] and R&D capacity [37] prove to be significant factors of cooperating companies across different countries. Furthermore, they show that firms's age and business operations sector impact significantly the likelihood of R&D cooperation.

A systematic review of the literature by Rybnicek and Koenigsgruber [38] showed that the list of determinants for successful R&D cooperations is by far larger than can be shown and discussed here. However, they show that partner selection in knowledge transfer context, and the overall performance of collaborations, depend on various factors. Besides quantifiable factors as shown above, a large part of those has qualitative dimensions. Thus, Wu and Barnes argue comprehensively in their literature review that future research should investigate how qualitative and quantitative objectives can be jointly considered in partner selection methods [20].

These approaches indicate that quantitative econometric data can be applied for understanding the econometric determinants of partner selection in a knowledge transfer context. However, no effort was made so far to combine these econometric measures with NLP approaches for supporting partner selection. In order to fill this gap, we ask Research Question 2:

RQ2: Is a combination of quantitative and qualitative information useful for supporting partner selection in public research?

[2]The literature search in Scopus for this chapter included variations of the following search string TITLE(((“university–industry” OR “industry–academy”) AND (collaboration OR link∗ OR cooperation)) OR ((“third mission” OR “triple helix” OR “technology transfer”) AND (university OR research)) AND (“success” OR determinants OR motiv∗ OR review)).

III. New Data-Driven Approach to Support Partner Selection

We propose an approach to support partner selection for public research organizations that is able to identify potential partners from a vast amount of companies (e.g., all SMEs in Germany). It is different from the previous ones in the sense that it does not require any expert knowledge, but only relies on the data of previous partners.

The advantage of such a data-driven approach is that it is easily transferable, scalable, and most importantly validatable. We will validate the performance of our approach by the example of the Fraunhofer Society, an organization for applied research with 73 institutes spread all over Germany.

Our approach consists of five parts, which we will describe in the following sections (see Fig. 1). We first provide some information on our data and its preprocessing in Section III-A. Then, we present the methodology of our approach (see Section III-B) and the way we measure its performance by means of validation (see Section III-C).

A. Data

1) Fraunhofer Industry-Partner Data: In order to assess the transfer potential of our approach, it is essential to understand the context of the Fraunhofer Society itself and the industry-partner data it provided us. The Fraunhofer Society is the largest nonprofit public research organization for applied sciences in the world [36]. It consists of 73 research institutions and has a goal to fill the gap between basic research and industrial applications; in other words, to facilitate and perform technology transfer from universities to industry. While only a part of its budget comes from the public funding, the majority comes from contracted research with industry partners [39]. This funding model clearly puts a lot of emphasis on acquiring industry partners for contracted research.

In the Fraunhofer Society, there is a clear focus on engineering and natural sciences among its research units. Additionally, fields related to health, social sciences, and economics are also found in the organization. Its research institutes are subsumed into eight Fraunhofer Clusters in order to enable R&D strategies coordination among institutes with related areas of technological expertise [39]. These clusters also give a glimpse of how diverse the portfolio of the organization is. These clusters are summarized in Table I.

Given the business areas in which the Fraunhofer Society is active, one expects that its industry partners are at least as diverse. The Fraunhofer Society documents all the past projects with its partners from the industry, because these are usually contracted research projects, which must be registered and reported for financial and controlling reasons. These internal project data are the cornerstone of our case study and the main enabling component of our approach.

The Fraunhofer Society provided us with a snapshot of its industry-partner data containing all R&D projects with external clients from 2012 to 2018, excluding classified projects. This amounts to roughly 100 000 projects carried out between its

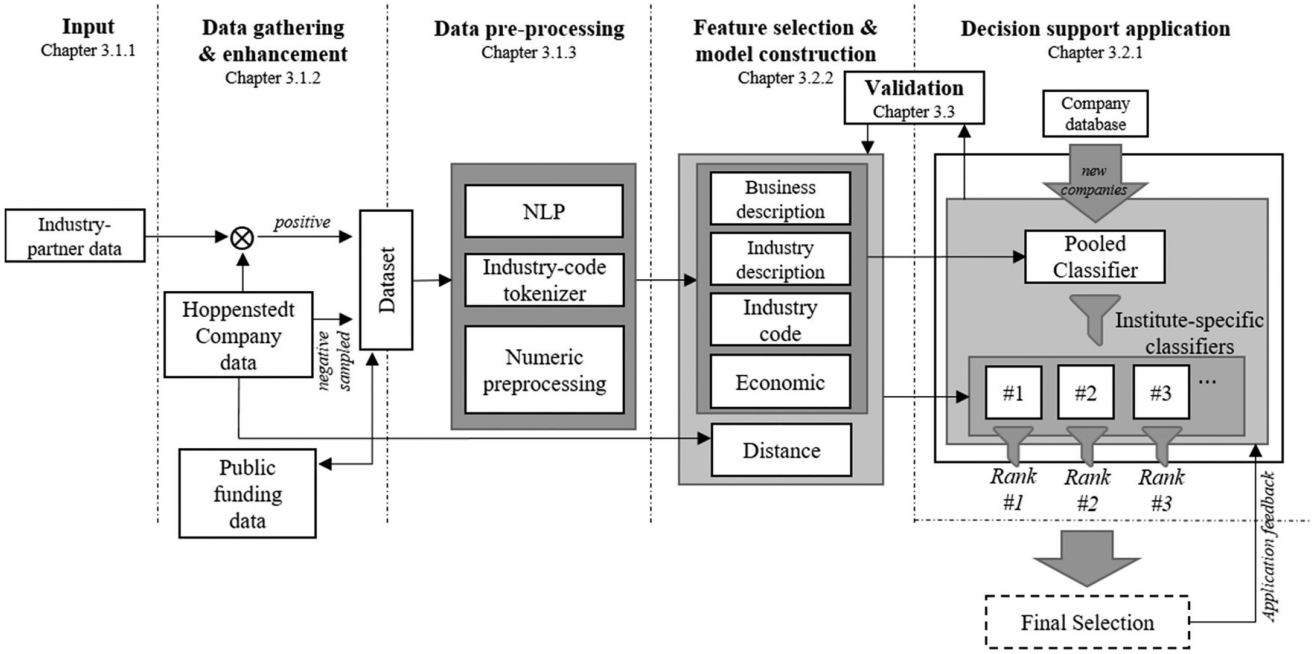

Fig. 1. New data-driven approach to support partner selection.

units and companies (see Table II). For each project, there is the name of the research institute of the Fraunhofer Society, the name of the industry partner, and an external unique identifier used for companies, the DUNS number.[3] This identifier allows combining the Fraunhofer dataset with proprietary external company databases in order to obtain more information on industry partners.

2) Company Data Gathering and Enhancement: With the unique identifier for every past partner at hand, we search those in a company database. For this end, we use a proprietary database, Bisnode's Hoppenstedt database. The Hoppenstedt database contains around 1.3 million companies in Germany, Austria, and Switzerland. In this source, additional information may be found for previous Fraunhofer partners. Besides very basic information such as name and address, one can find, for example, information as last registered *turnover*, number of *employees*, number of *managers,* and *firm age*. Additionally, there is a free text field with a short *business description*. Finally, almost every company in the Hoppenstedt database has one or more five-digit *industry-codes*[4] and a standardized *industry description*, both from a classification system for industry branches developed by the Federal Statistical Office of Germany, *Klassifikation der Wirtschaftszweige (WZ)*, which follows the same structure as the European NACE classification.

From all the past industry partners of Fraunhofer, we find an entry in Hoppenstedt for 9972 of them (see Table II). This constitutes the positive class, i.e., positive examples of past partners of Fraunhofer. In order to build a balanced supervised dataset, first, we sample randomly the same amount of companies from Hoppenstedt, excluding the 9972 found Fraunhofer partners. Second, we label this random sample as negative examples. This is a common technique from Recommender Systems literature, known as negative sampling, or sample-based learning [40], [41]. One performs this procedure in order to avoid overfitting and to guide convergence of the optimization algorithms in a positive-unlabeled supervised learning setting.

In our framework, it is natural that the interactions (or the partnerships in our case study) that did not take place are not always "unobserved signal," the *unlabeled* observations. At cases, it means that the researcher chose not to cooperate with the company in question. However, this negative decision is not registered anywhere. While this modeling technique in recommender systems is supported by the assumption that most of the available products are not of interest for one specific buyer, in our framework, the decision to use negative sampling is supported by the following.

According to the Leibniz Centre for European Economic Research (ZEW), 53 600 German companies engaged in R&D activities in 2017 [42]. This number amounts to 4.1% of the number of companies registered in the Hoppenstedt database. In other words, if all companies engaged in R&D were registered in Hoppenstedt, then this would still amount to only 4.1% of the total amount of registered companies. Now, even when assuming that all the R&D active companies were good matches for the Fraunhofer Society, which is extravagant, and that those companies were all listed in Hoppenstedt, the ratio of good potential matches wrongly labeled as false is bounded by 4.1%. This provides an upper bound on the falsely labeled companies.

[3]The Data Universal Numbering System, DUNS for short, is a proprietary system developed and regulated by Dun & Bradstreet with the goal to assign an unique numeric identifier (DUNS number) to every single business entity worldwide, which sometimes are whole companies, and others subunits of larger corporations.

[4]Companies have up to 19 WZ codes in our dataset.

TABLE I
FRAUNHOFER CLUSTERS AND THEIR AREA OF EXPERTISE [36], [39]

Fraunhofer Clusters	Institutes	Business areas
Information and communication technology	16	Mobility and transportation, e-government, public safety and security, manufacturing and logistics, media and creative sector, digital services, business and finance informatics, medical and healthcare systems, energy and sustainability
Life sciences	7	Medical technology, regenerative medicine, healthy foods, biotechnology, process, chemical, and herbicide safety
Light and surfaces	6	Surface and coating technologies, beam sources, micro and nano technology, materials processing, optical measuring techniques
Microelectronics	11	Smart and healthy living, energy efficient systems, mobility and urbanization, industrial automation
Production	12	Product development, manufacturing technologies, manufacturing systems, production processes, production organization and logistics
Defense and security	10	Crisis and disaster management, digital transformation, border security, combating terrorism and crime, resilience and protection of critical infrastructures, electronic warfare, protection and impact, networked operations and information gathering, provision of information and decision-making support
Materials	16	Health, energy and environment, mobility, construction and living, machinery and plant engineering, microsystems technology, safety
Innovation	5	Technology evaluation, structural change, transformation processes, future scenarios, innovation policy, decision support in policy making

TABLE II
INDUSTRY PARTNERS FOUND IN HOPPENSTEDT

	Companies	Associated Projects
Fraunhofer industry-partner data	17,274	105,435
Found in Hoppenstedt	9,972	66,364

Finally, this is the necessary condition on the dataset in order to use negative sampling [40], which justifies our design choice.

After sampling randomly the negative set, we have in our dataset 19 944 companies, or observations, half as positive examples of past industry partners and half of them negative. Moreover, every observation is associated with the above-mentioned variables: *turnover*, *employees*, *managers, firm age*, *business description*, *industry-code,* and *industry-branch-text.*

To complete the data enhancement phase, we integrate two more variables to our dataset for every observation. First, we search FÖKAT, a publicly available database with more than a 100 000 projects funded by the federal government of Germany for the 19 944 companies in the dataset.[5] This matching generates the binary variable *publicly funded*, which is positive if the search returned any hit and negative otherwise. Second, we keep the address of every company. This will be crucial to calculate the distance between every research unit of Fraunhofer and the respective company.

3) Data Preprocessing: We begin first by encountering the missing values in our data. In general, the quality of our data is satisfactory for most variables, merely *turnover* and *employees* have relatively high missing ratios, 56% and 33%, respectively

[5]For the identification in the FÖKAT database, the names of the companies were harmonized with regular expressions and then matched with harmonized names in Hoppenstedt. The information was extracted for both, Fraunhofer partners and randomly sampled companies from Hoppenstedt.

TABLE III
DESCRIPTIVE STATISTICS

Statistic	Unit / Format	Observations	Mean	St. Dev.	Min	Median	Max
Partner / Non-partner	Binary	19,944	0.50	0.50	0	0	1
Firm age	Years	19,919	32.91	45.34	0.00	18.00	1,818.00
Employees	Count	13,266	998.71	11,300.55	1.00	38.00	546,406.00
Turnover	M Euro	8,848	183.29	1,651.20	0.01	4.75	83,312.23
Managers	Count	18,981	4.84	7.02	0.00	2.00	163.00
Publicly funded	Binary	19,944	0.17	0.37	0	0	1
Industry codes	Character	18,876	-	-	-	-	-
Industry description	Character	18,876	-	-	-	-	-
Business description	Character	17,401	-	-	-	-	-

(see Table III), but still remain informative when imputed. We impute missing values of numeric variables (*turnover*, *employees*, *managers,* and *firm age*) by the median. Missing text variables (*business descriptions*, *industry-codes, or industry descriptions)* are marked as such for the latter text processing. Moreover, we log-transform numeric variables because their distribution is highly skewed to the right, as it can be seen from comparing the means and medians in Table III. Finally, the numeric variables are centralized and rescaled.[6]

Now, it remains to preprocess *business description*, *industry-codes,* and *industry descriptions*. *Business description* and *industry description* are free text variables, that is, unstructured data. In order to use such information in our model, we need to transform the unstructured data. We use the following common NLP-procedure. First, all words in the texts are scanned and stored in a list. This list is the raw corpus. We exclude words with less than ten occurrences and german stop-words. After that, the words in the corpus undergo stemming, which is the process of reducing inflectional forms of a word to a common base form. For this end, we use the Snowball algorithm[7] for the German language. Finally, we adopt one-hot-encoding as the preferred technique to encode the stems in text variable in our model, which converts the text variable for every company into many binary features indicating the occurrence of a specific word/stem.[8]

For *industry-code*, although it is technically a categorical variable of nominal order, if we apply a canonic one-hot-encoding, we would miss a critical point in an industry classification scheme like NACE or WZ. Their classes and subclasses are leafs in a tree structure, and therefore carry much more information than just the binary one of class affiliation, as shown in Fig. 2.

For example, WZ codes under the same division are more similar to each other than two WZ codes that are in different divisions. This logic holds on the whole classification tree, i.e., on all levels. Hence, it would be inattentive to miss this valuable information. In order to model in part this structural property of WZ codes, we split the code string and then encode it in a one-hot fashion. The developed tokenization scheme is shown in Fig. 3. If a company has multiple WZ codes, the same procedure is simply repeated for each code.

After encoding the variables *business description, industry-codes,* and *industry description* as explained above, the number of dimension of our model's feature set jumps from handful to a couple of thousand due to the one-hot-encoding of the text fields and WZ code(s) and the numerous binary variables that are generated. Moreover, for each of those last three variables, we create one binary feature to indicate a missing value. In this way, we are able to encode every observation, even those with a missing text-variable, without the need for imputation.

We remark here that the only similarity our approach shares with the initially presented bibliometric approaches in Section II-B is the encoding of texts into vectors. However, we use one-hot-encoding, which registers the presence of a word in a text. Differently from previous approaches, we encode every word and let the algorithm learn which words carry predictive information, instead of enforcing by hand so. The learned parameters assemble therefore an implicit measurement of technological fit of potential partners.

In order to finalize the construction of our feature set for every company, we compute the *geographic distance* from it to every Fraunhofer institute. This procedure generates 73 new features, representing the distance of the pair *(company, institute)*. For the calculation, we retrieve the geocoordinates from

[6]All the above numeric transformations were performed in the software framework R-Studio, with code developed for this article in conjunction with standard R libraries for data handling and descriptive statistics. For the following NLP tasks, the library tidytext [43] was used in combination with own code and standard R libraries.

[7][Online]. Available: https://snowballstem.org/algorithms/german/stemmer.html

[8]Alternatively, one can use word counts or normalized word-counts (e.g., TF-IDF). In our case, both methods however yield worse results with regard to the performance of the classifier.

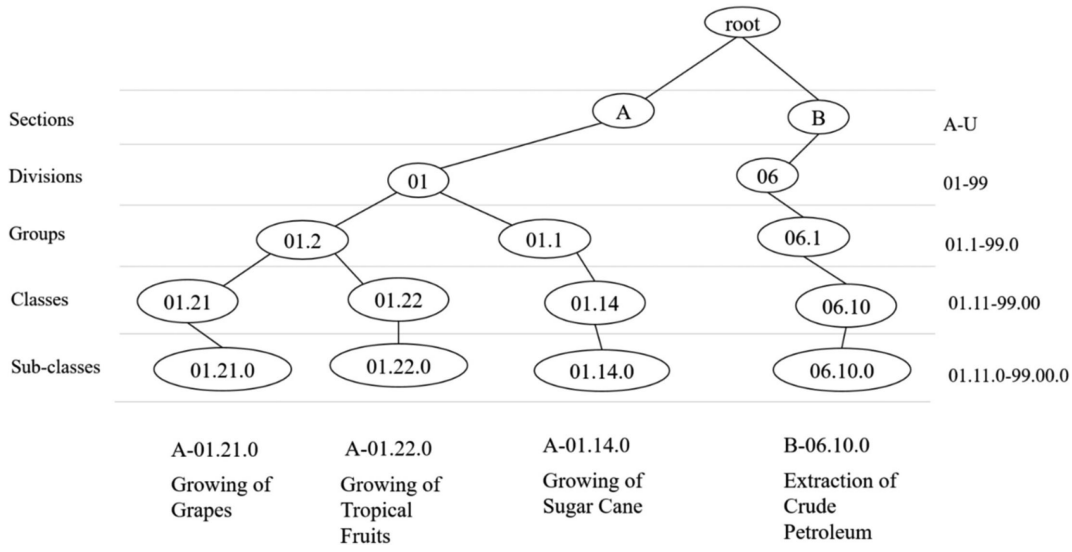

Fig. 2. WZ classification (example).

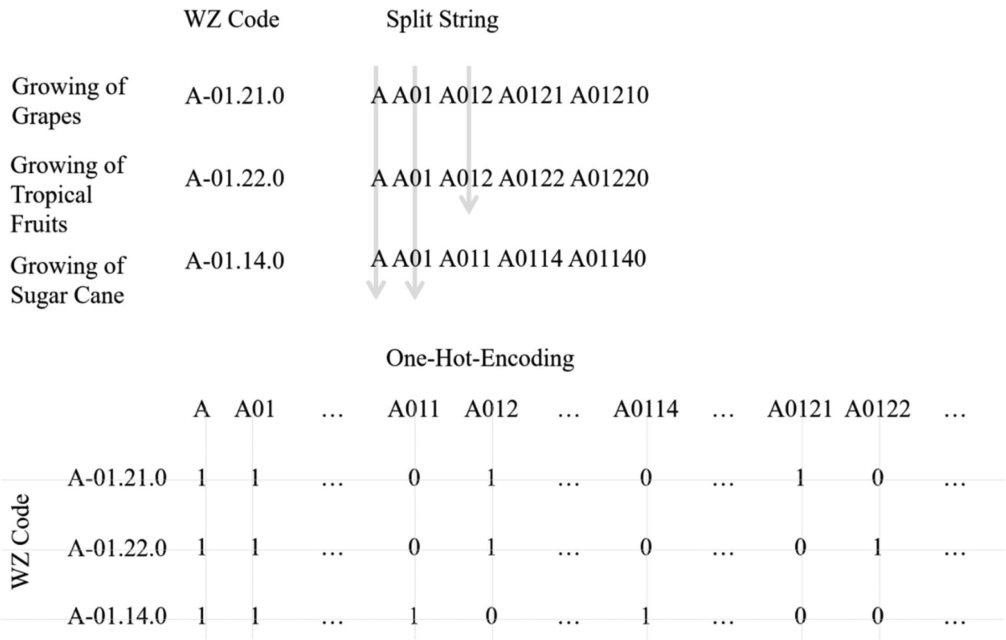

Fig. 3. WZ industry code tokenization scheme.

OpenStreetMap and, finally, we calculate the distance using the spherical law of cosines.

Finally, for every observation (firm) in our dataset, we have the standard features *turnover*, *employees*, *managers, firm age,* and *publicly funded*; the feature sets *business description*, *industry-code,* and *industry-branch-text,* assembled from unstructured data (text); and finally the feature set *distance.* In the following, we describe how we use this in order to support partner selection in a technology transfer context.

B. Methodology

1) Structure of the Decision Support Application: The purpose of our classification approach is to identify potential partners from a large set of companies. In our case, this means we need to assess each of the 1.3 million companies in our database. Since we are dealing with such a large amount of "new companies," our approach has two stages. First, we filter out irrelevant companies from our database with a rather generic *pooled classifier*. Then, we calculate a ranking of the remaining companies for each of the 70 Fraunhofer institutes with an *institute-specific classifier*. The provided ranking is supposed to provide the institutes with a prioritized list of potential matches. Experts' selection of suggested partners can then be used as feedback for the application, i.e., to retrain the classifiers.

As most other systematic partner-selection approaches, our approach supports the qualification stage of partner-selection

by providing a ranking of the most promising recommendations. Different from other approaches, however, it also provides an interpretable quality measure for the recommendation with probabilities, because it is a supervised prediction and not a mere unsupervised dimensionality reduction (see following sections). Moreover, our approach also supports the criteria formulation phase by automatically determining which features are relevant for the specific institute according to its partner-history. Thus, it is not necessary to model the preference of the decision maker explicitly, as needed in other approaches, because the classifier learns the preference from past data. Finally, while the selection of industry partners itself is left to the decision makers (e.g., managers at technology transfer offices), the selected firms can be used as feedback into the model to improve further recommendations.

2) Feature Selection and Model Construction: There are different requirements for the two classifier classes, which are stacked in our decision support application. For the first, it is important to cover as many of the existing industry partners as possible, while misclassifications to this regard are not as severe. For the second, it is highly undesirable to propose nonrelevant companies as high ranked recommendation.

Pooled classifier: First, we develop a pooled classifier as a prefilter, in order to discriminate good potential industry partners for the entire Fraunhofer Society from other companies. For this, we train one binary classifier to predict the target variable *partner/nonpartner* (see Table III). For model construction and validation, we use all the gathered data for all the 19 944 companies. In order to perform feature selection, we group the created features into four sets. The economic feature set contains the numerical variables (*turnover*, *employees*, *managers*, *firm age,* and *publicy funded*).[9] The *business description* set groups the cleaned, stemmed, and one-hot-encoded business description text. The *industry description* set assembles the binary variables generated after text preprocessing and one-hot-encoding procedure. Finally, the *industry codes* set collects the binary variables created by the industry code tokenizer. We do not consider the feature set *distance* for this classifier because the distance to the Fraunhofer Society as such cannot be calculated. The target variable of the pooled classifier denotes whether a company is a customer of any of the 73 Fraunhofer institutes, which are spread all over Germany.

The pooled classifier staging as a prefiltering step, i.e., preceding subsequent modules that are more complex (e.g., institute-specific recommendations), might be necessary for scalability reasons. For example, a recommendation web interface might not be able to handle 1.3 million company entries with a satisfactory latency or the model might be computationally too expensive to apply it to such a large dataset. Or if operated offline, one can leverage such classifier to sort out autonomously irrelevant companies from a large company databases like Hoppenstedt once it is trained, instead of doing this by hand.

[9]We choose these specific economic variables both, because they are commonly used in the partner selection literature (see Section II.C) and because they showed the highest performance for the purely economic pooled classification model (Model 1 in Table 5).

A company will be sorted out by the pooled classifier, when it is from an industry with no or few contacts to Fraunhofer. This is trivial for companies, which are assigned with industry codes that never appeared among Fraunhofer customers. It becomes more difficult if a company is assigned an industry code that has appeared among Fraunhofer customers. Should we consider this company as a potential customer? First, it depends on the code itself: How informative is this code in determining a Fraunhofer partner? Second, it depends on the other codes of the company: Are there more codes that indicate that this company is similar to previous Fraunhofer partners? Are there some that indicate the opposite? How informative are the other codes? Considering all these aspects can be quite complex, however our brain is the very good in making such fuzzy decisions and surely can do this as well. What a human cannot do easily however is to perform such a task for millions of companies and thousands of previous partners at a time. This is where the pooled classifier goes a long way.

Institute-specific classifier ensemble: Second, we build an ensemble of institute-specific classifiers in order to rank and recommend good potential industry partners to specific institutes or research units. This second level of our model provides each of the Fraunhofer research institutes with recommendations for *new* industry partners. For this, we train an ensemble of binary classifiers, one for each of the over 70 institutes, for discriminating between its own industry partners and industry partners of the other institutes, in a one-versus-rest manner. For each institute *i,* we construct a binary target variable ($partner/nonpartner_i$), which is positive if the company was an industry partner of the institute *i* and negative if the company was an industry partner of any other institute, but not institute *i*.

The institute-specific classifiers use only companies that were past industry partners of Fraunhofer, both for training and validation. This is because we want to capture characteristics of the companies that make them a good match for a specific institute compared to other institutes, which might be active in a completely different business area, but also have industry partners that are active in R&D. If we chose to discriminate again between industry partners of an institute and companies randomly sampled from Hoppenstedt**,** the classifier would put more weight on characteristics of innovative companies in general, like in the pooled regression, but not on the specific thematic characteristics of a certain business area. This labeling leads to imbalanced sample sets with regard to the target variable ($partner/nonpartner_i$), because most institutes worked only with a small fraction of the Fraunhofer industry partners. In extreme cases, too few positive observations do not allow for a proper estimation and validation. Therefore, we exclude institutes with less than 20 industry partners from the analysis, which leaves us with 70 out of 73 institutes, with one classifier each.

In this ensemble of institute-specific classifiers, we use the features sets *economic*, *business description, industry description,* and *industry code* as before, plus a new feature set, *distance*. Note that, every one of the institute-specific classifiers utilizes only one of the distance features. For example, the *kth* institute-specific classifier has as feature *distance (company,* $\mathrm{institute}_k$*)*.

3) Classifier Models and Libraries: Nonlinear classification models like neural networks have become very popular for text classification because of their good performance. Nevertheless, linear models, such as logistic regressions and support vector machines (SVM), remain strong baselines for text classification and often even reach up to nonlinear models' performance, while being much faster and able to process much larger datasets [44], [45]. Therefore, for both pooled and institute-specific classifiers, we use the library LIBLINEAR [46].[10]

LIBLINEAR contains various implementations of linear models for classification. Those models are based either on LR or SVM. SVMs may perform better, but LR has two main advantages. First, due to the simplicity of the model, it has straightforward coefficient analysis (see Section IV-A). Second, LR is a probabilistic classification method and, therefore, naturally outputs posterior probabilities that are well suited for ranking the observations on the test set, hence allowing the straightforward computation of rank-based performance measures. SVM, however, would require some additional calibration methods to emulate such probabilities.[11]

We use a L1 regularized variant of the LR classifier because it can handle high-dimensional data without overfitting. In addition, it can handle models with more features than observations. Moreover, to better assess the relevance of features, L1 regularization allows for exact zero coefficients, hence for feature selection during learning.

To test whether nonlinear models might still be worthwhile for our purpose, we also train a version of the Random Forrest [47] classifier. Since we are mainly interested in probabilities for ranking purposes, we train a so-called probability forest [48] with 500 randomized decision trees each. We use the *ranger* [49] library, which is specifically written for high-dimensional data.

C. Validation

To see whether the proposed classification approach has the predictive power to support industry partner selection for public research (RQ 1 in Section II-B.), we calculate its performance for the example of the Fraunhofer Society.

We perform ten-fold cross-validation[12] for all the classifiers and calculate various standard performance metrics, such as precision, recall, accuracy, f1-score, for each model specification by averaging across the folds.[13] We also compute the receiver-operating-curve, area under the curve (ROC AUC) metric, a standard performance metric for binary classification, which nowadays is provided with most validation implementations.[14] The receiver-operating-curve plots the relationship between false positive rate and recall for all possible classification thresholds.[15] ROC AUC is simply the area under this curve and is a measure of how well a classifier performs in different settings, for example, with a high threshold for a careful assignment of positive labels (e.g., clinical trials) or vice versa (e.g., cancer detection). At the same time, it is equal to the probability that a classifier ranks a random positive observation higher than a random negative observation.

While for the pooled classification all the above performance metrics are measured, for the institute-specific classification we just compute the ROC AUC score. This is because, while the latter should assess the performance of the classifier in ranking companies, the former should measure the performance of the model in deciding if a company is a prospective industry partner or not. Additionally, in the pooled classification task, a balanced dataset is used for training and testing, while this is not the case for the institute-specific classification. In fact, for each one of the 70 institute-specific classifiers in the ensemble, there is a distinct ratio of positive–negative observations. This all justifies the decision of using ROC AUC for comparison, as this measure is robust against unbalanced class distributions. ROC AUC, apart from being very useful in comparing classification models with distinct positive–negative class ratio, serves our purpose well for a second reason. ROC AUC is measuring in fact how well the models are ranking the industry partners, in comparison to the not-industry-partners. That is, given a group of companies to be tested by the model, ROC AUC scores higher if the model is able to rank higher the industry partners and lower the not-industry-partners. This is very important for the usability of the ranking. When, a researcher that is looking for a partner goes through a list of companies, clearly the true potential industry partners should be on the top of this list.

By comparing the performance metrics of different feature combinations, we can also test which of the proposed sources of quantitative (i.e., *economic*) and qualitative information (i.e., *industry codes*, *industry descriptions* or *business descriptions*) on the companies can be leveraged for classification (RQ 2 in Section II-C).

IV. Results and Discussion

In the following, we provide an analysis of the performance of the designed solution for supported partner selection. First, we assess how well the pooled classification approach works for identifying potential partners for research organizations in general, i.e., prefiltering our dataset. While on that, we conduct a coefficient analysis on our LR model for pooled classification and analyze how the used features influence on the probability of classifying a firm as a good potential partner. We compare the performance of three different classification techniques: regularized logistic regression, SVMs, and random forest. Second, we measure the performance of the ensemble of LR models for the

[10]LIBLINEAR is originally written in C/C++. We use LiblineaR, which is a wrapper around it for R. [Online]. Available: https://cran.r-project.org/web/packages/LiblineaR/index.html

[11]Usually Platt scaling is used for this purpose with SVMs (Platt, 1999). It transforms the binary SVM outputs into a probability distribution by fitting a logistic regression on the sample distances to the hyperplane estimated by the SVM. This makes it slower than a plain SVM estimation and adds another layer of uncertainty.

[12]Since we take a random sample for the negative class out of Hoppenstedt in the pooled classification, we needed to check our results for robustness beyond cross-validation. Therefore, we repeated cross-validation ten times for each model in the pooled regressions, with a different random sample each time. The results in fact remained very stable compared to those presented in this article.

[13]Apart from the model itself, the whole Machine Learning pipeline, from preprocessing to resampling, is set up using the MLR package [50] in R, which provides a convenient workflow for our purposes.

[14]In our case the MLR package for R.

[15]Precision, recall, accuracy, and f1-score are usually calculated with a probability threshold of 0.5.

TABLE IV
COEFFICIENTS OF LOGISTIC REGRESSIONS FOR POOLED CLASSIFICATION

Results of logistic regression models (pooled classification)

	Target variable: Partner/Non- partner						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Economic features							
Firm age	0.188				0.171	0.165	0.305
Employees	0.870				0.602	0.663	0.602
Turnover (M Euro)	0.183				0.259	0.200	0.195
Managers	1.360				1.026	1.021	1.036
Publicly funded	1.072				2.246	2.208	2.258
High-dimensional features							
Business description		X			X		
Industry description			X			X	
Industry codes				X			X
Obs.	19,944	19,944	19,944	19,944	19,944	19,944	19,944
Features	5	2,161	844	1,678	2,166	849	1,683

institute-specific classification. All the above analysis are also performed for various feature sets, which we compare among each other for the sake of completeness of the analysis.

A. Pooled Classification

Table IV shows the results of the pooled classification models with various feature combinations using a regularized logistic regression. Interestingly, the coefficients of all economic features show a positive value and none of the coefficients was regularized to zero. This shows that the chosen economic features are indeed relevant determinants of industry partners in the case of Fraunhofer, which supports the results of the previous literature presented in Section II-C.

However, from a machine learning perspective, we are rather interested in the performance of the respective models in cross-validation. Model (1) includes only economic features like *firm age*, size (*turnover*, *employees* and *managers*), and a binary feature, which indicates whether a firm received public funding or not, *publicly funded*. This simple model shows a reasonable performance. The precision performance metric indicates that 88% of the companies that were labeled as positive by the classifier in cross validation were indeed Fraunhofer industry partners. Because the dataset is balanced with regard to the target variable, this compares to a baseline of 50%, which would result from a naïve approach, for example, if one labeled all companies positive or negative. None of the high-dimensional models (2)–(4), which include several hundreds of features constructed from business or industry descriptions or industry codes, reaches such a high precision as in (1). However, the recall of the low-dimensional model shows that only 73% of the industry partners present in the dataset were identified as such by the classifier. In this regard, the high-dimensional models using industry descriptions and industry codes perform better, with a recall of 76% and 78%, respectively. In order to compare the different models, one would normally use the F1-score, being the harmonic mean of the precision and recall, because we are only interested in the performance with regard to the positive

class. Accuracy on the other hand assesses the performance for both classes, positive and negative. Both measures show that the economic model outperforms the high-dimensional models with a single source of information (2)–(4). However, a combination of economic indicators with high-dimensional features in models (5)–(7) dominates clearly all previous models with regard to all measures. Model (7) slightly beats all others with an accuracy and f1-score of 85%, precision of 86%, and recall of 84%.[16] This means that there is an increase of 11% in recall, the performance measure of main interest for pooled classification, between model (1) and model (7). However, the question is: Is this a considerable increase? To answer this question, consider how many partners one would miss if model (1) was used instead of model (7). Since there are 9972 confirmed industry partners in the dataset, we would miss almost 1100 companies.[17] We consider this as a substantial improvement.

As can be seen in Table V, more elaborate classification techniques, suitable for high-dimensional classification problems like SVMs with linear kernel or nonlinear random forests, do not perform noticeably better than the simple LR learner with regularization.[18] The best random forest model outperforms the best LR model (7) only by 1% of F1-Score, while the ROC AUC remains the same.

Hence, it seems that linear models indeed are sufficient and other nonlinear techniques are not very promising in our case. At the same time, LR models have the advantage of being interpretable to the decision makers and providing well-calibrated probabilities for ranking companies. Therefore, in the following, we only report the results of the regularized LR models.

B. Institute-Specific Classifications

Fig. 4 shows the performance distribution of the institute-specific classifiers for different model specifications using regularized LR learners.[19] For ease of representation, and since we are mainly interested in the quality of the probability rankings for this classification task, we only consider the ROC AUC performance metrics and no other metrics like precision or recall (see Section III-C). The jittered points show the averaged tenfold ROC AUC score of each of the 70 classifiers. The bell curves show the kernel-density estimates of the score distributions and the vertical lines mark the according median of the sample.

As expected, classifiers with low-dimensional economic features only (Model A) do not perform well in distinguishing industry partners between the different institutes. The best classifiers only reach up to a ROC AUC of around 70% and the median lies only slightly above 60%. Adding geographic distance between the companies and the respective institute (Model B) increases the performance of the classifiers considerably, with three classifiers reaching up to a ROC AUC of around 80%, but two institute-classifiers still remaining useless at below 50% ROC AUC. Classifiers using business descriptions from Hoppenstedt converted to high-dimensional word one-hot-encoding (Model C) perform better than classifiers using economic features only (Model A), but worse than geoeconomic features (Model B). However, most classifiers remain in the range between 60% and 70%, and many even below, which is not a satisfactory result.

Model (D) and (E) perform much better by using features constructed from text and code information on industry sector. Some classifiers range clearly above 80% ROC AUC, which compares to the performance of pooled classifier, despite the institute-specific classification being a much harder task. Most classifiers have a ROC AUC metric of above 70%, which is quite reasonable. There is no clear preference for one of the models, since the mean and median performances are almost the same. However, Model (D) seems to be more stable across institutes, as Model (E) has very low performance of below 50% for one institute.

As in the previous classification task, improved results are obtained with models that combine economic and high-dimensional features, as in Model (F), and even better results are achieved by adding geographic distance (Model G). The combinations of all three, however, yield even better results (Models H and I). The inclusion of this additional information helps us to increase the performance of most classifiers, but the best performers from Model (D) do not improve that much by adding economic or geographic information. A possible interpretation is that for some very diverse institutes, for whom it is difficult to make recommendations based on the business activity of their industry partners, location and economic information matter, while for the others information on business activity descriptions are sufficient. On average, classifiers including industry code features, economic features, and geographic distance perform the best.

The results show that those feature sets that describe the companies' industry, by either text or code, are the most important for the institute-classifiers in order to rank companies. However, adding economic or geographic information can improve the performance considerably.

V. Limitations

One limitation of our approach is that the information provided on Fraunhofer industry partners is positive-unlabeled. This means that we only know partners, but not *non*partners. This may hinder the training of the classifiers, because some of the companies marked as negative might still be good candidates for industry partners. The gravity of this effect differs between the *pooled* classification and *institute-specific* classifications.

[16]Note that we do not report all combinations of available features. A combination of economic features, business description, industry description, and industry codes indeed improves the f1-score, but only by 1% and it is computationally much more costly, because it has much more features (4688). None of the combinations however yields a better ROC AUC value than the best previous feature combination (model 7 with industry codes).

[17]For the other performance measure of interest in this article, ROC AUC, such an analogy is not possible and improvements can unfortunately only be assessed in a relative manner.

[18]Note that the ROC AUC measure is not available for SVM without further modifications and hence is not reported (see Section III.C).

[19]Note that these model specifications differ from those of the pooled classifier in Section IV.A, because some features, which are available for the institute-specific classifier (e.g., geographic *distance* between the company and the respective institute) are not available for the pooled classifier.

TABLE V
PERFORMANCE OF LOGISTIC REGRESSIONS, SVMS, AND RANDOM FORESTS

Performance (pooled classification)

	Target variable: Partner/Non-Partner						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Economic features	X				X	X	X
Business description		X			X		
Industry description			X			X	
Industry codes				X			X
Obs.	19,944	19,944	19,944	19,944	19,944	19,944	19,944
Features	5	2,161	844	1,678	2,166	849	1,683
	Logistic Regression						
Precision	88%	84%	82%	81%	88%	89%	86%
Recall	73%	73%	76%	78%	79%	80%	84%
Accuracy	82%	80%	80%	80%	84%	85%	85%
F1-Score	80%	78%	79%	79%	83%	84%	85%
ROC AUC	88%	86%	87%	88%	92%	93%	93%
	Support Vector Machine						
Precision	89%	83%	82%	80%	88%	89%	87%
Recall	72%	72%	76%	79%	78%	79%	83%
Accuracy	82%	79%	80%	79%	84%	85%	85%
F1-Score	80%	77%	79%	79%	83%	84%	85%
ROC AUC	-	-	-	-	-	-	-
	Random Forest						
Precision	86%	82%	81%	80%	88%	86%	86%
Recall	77%	78%	77%	80%	82%	86%	86%
Accuracy	82%	80%	80%	80%	85%	86%	86%
F1-Score	81%	80%	79%	80%	85%	86%	86%
ROC AUC	90%	87%	88%	88%	93%	93%	93%

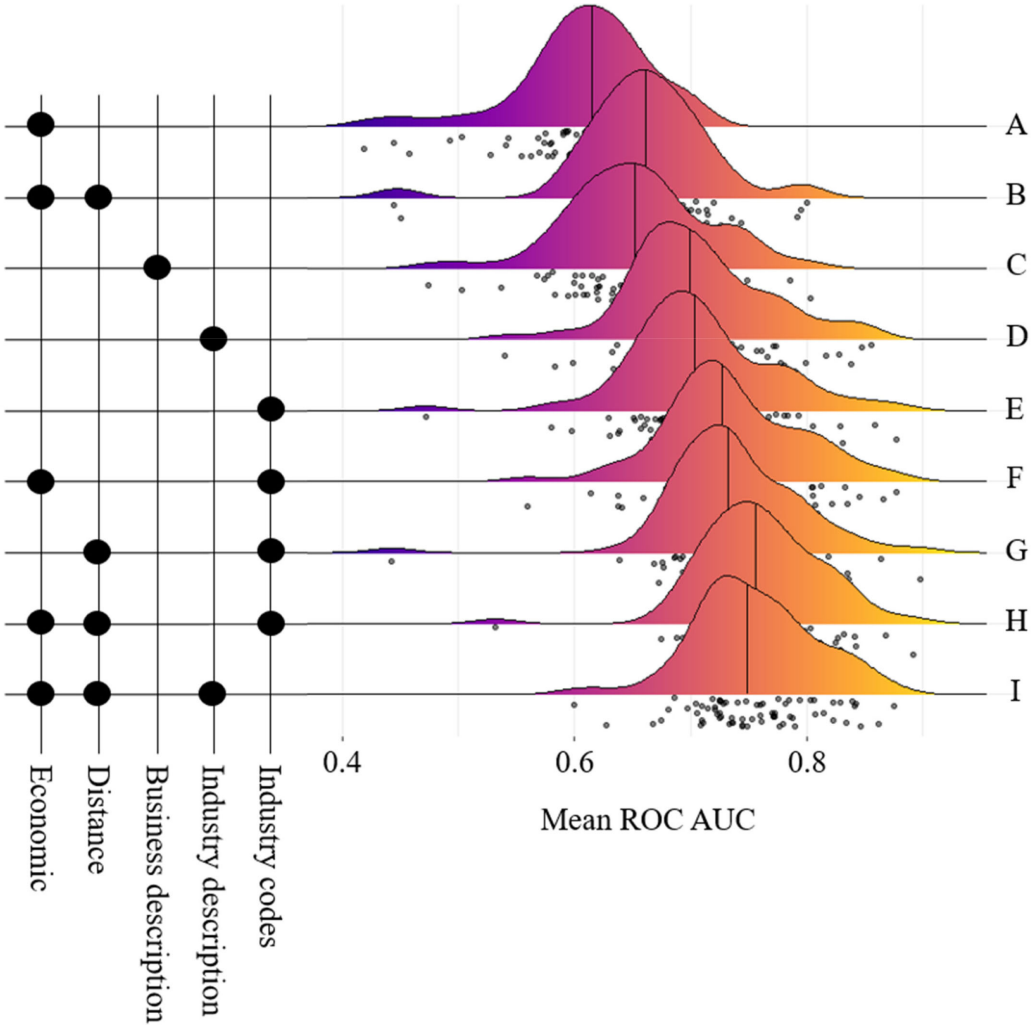

Fig. 4. Performance of the institute-specific classifiers (logistic regression).

For the *pooled* classification, we decided to randomly sample companies from Hoppenstedt as negative nonpartners. The argument why this shall not compromise the overall performance of our system is that, although this is indeed possible, it is also very unlikely. The database Hoppenstedt has around 1.3 million companies from which we sample randomly around 10 000, in order to balance our dataset. From all the companies listed in Hoppenstedt, the vast majority are not interesting for the Fraunhofer Society as a partner. First, because only a minority of all companies engages in innovation-oriented activities, as we address in Section III-A.2. Second, because an even smaller number among those could profit directly from the research portfolio the Fraunhofer Society has to offer. Hence, randomly sampling Hoppenstedt will return almost exclusively nonrelevant companies and we assume that it does not affect the training and validation of the pooled classifier. While one could argue that this is the reason this separation task is easy, it is certainly painstakingly hard to find manually a small group of interesting companies in a universe of 1.3 million different firms.

For the training of *institute-specific* classifiers, we choose to use Fraunhofer industry partners only. As mentioned in Section III-B.1, this helps the classifiers to learn to distinguish between specific thematic or technological characteristics of the companies rather than to distinguish between innovative and noninnovative or R&D active and nonactive companies in general. The downside of this approach is that it aggravates the problem that companies labeled as nonpartners might still be good candidates. Because the negative set encompasses Fraunhofer industry partners only. Therefore, the probability is high that the negative set includes companies, which are good candidates for institutes beyond those that have been industry partners according to industry-partner data. Note that, on the contrary to the *pooled classifier*, here one does not have a statistical measure to upper bound the ratio of falsely negative-labeled good potential partners.

On one hand, this "erroneous" labeling of nonpartners may affect the training of the classifiers considerably. For example, consider two institutes that do not share any industry partners (e.g., because of their geographic location), but their industry partners are very similar with regard to their business description, industry description or industry codes and represent potential partners for both institutes. Our two classifiers would assign

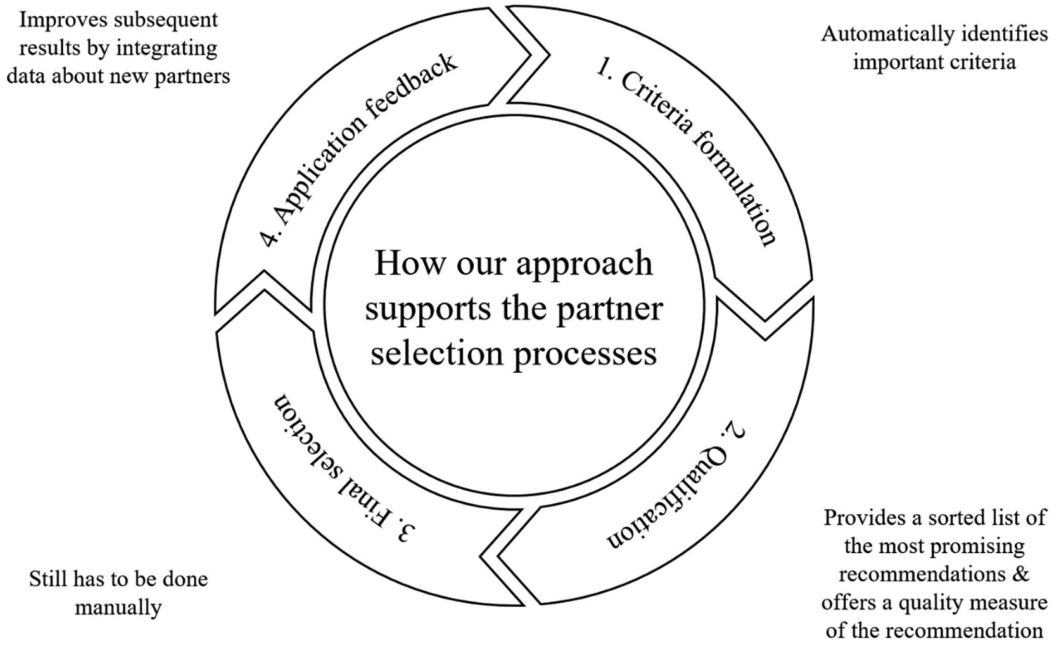

Fig. 5. Proposed support system for partner selection; adopted from [20].

low weights to features (i.e., words or codes), which the partners have in common because they appear for both industry partners and nonpartners. On the other hand, the classifiers will put more weight to features that are special for the respective institute and rank companies higher that match this particular institute only, which indeed is a desired effect for the decision support application.

An improvement of our results could potentially be reached by using observation weights in our classifications. Such weights could be, for example, constructed based on the number of projects or value of projects between a certain company and Fraunhofer. Moreover, it would be interesting to analyze if texts longer than the business activity, for example, from the homepages of the companies, would help the classifiers to make better predictions. In this case, however, a more thorough analysis of the various text-preprocessing options would be needed. Another interesting strand for further research could be how to make use of interinstitute correlations present in our data, for example, with recommender systems or multilabel problem transformation methods [51]. These methods usually use user–item interaction (in this case company-institute) data in order to predict preferences of other users by taking advantage of user–user correlation. Finally, it would be interesting to analyze empirically whether the use of our approach indeed improves the partner-selection process compared to the manual selection. This would be of great interest in order to settle, for example, the question if the pooled classifier is indeed helping researchers find new partners among the whole 1.3 million possibilities, or if they would prefer performing this step by hand.

Another practical limitation of our approach is that it makes only recommendations that are similar to previous partners, which is problematic for research institutes that want to explore a new branch of industry partners. On the other hand, our approach can be easily adapted to solve this problem. Instead of using previous customers to train a classifier one could also use a set of hand-selected examples of potential partners. The trained classifier would then be able to recommend further potential partners of this new type.

VI. Practical Use

The practical use of our article is to support research organizations at finding new industry partners by facilitating the partner selection process. Fig. 5 shows how managers in technology transfer offices and scientists can benefit from the proposed approach.

In Stage 1, our approach automatically identifies important criteria for the partner selection. Then, it provides in Stage 2 a ranked preselection of the best-fitting partners out of an overwhelming pool of organizations. Both phases base themselves exclusively on former cooperation behavior data and company data, reducing hence the entire initial hurdle in finding a partner. Proceeding, this recommendation is provided to the decision maker who performs the final selection. In the last phase of partner selection process, data from selected partners are added to the existing dataset and the models improve future recommendations.

Since our research is part of a research project for the Fraunhofer Society, we implemented the results in a web application called *SME-Match*[20] (see Fig. 6). Fraunhofer researchers can use this prototype tool to look for new industry partners. The focus of this application lies on the recommendation of SMEs, because there is an overwhelming number of potential SME partners, while most large corporations are already known.

[20]Originally *KMU-Match* in German.

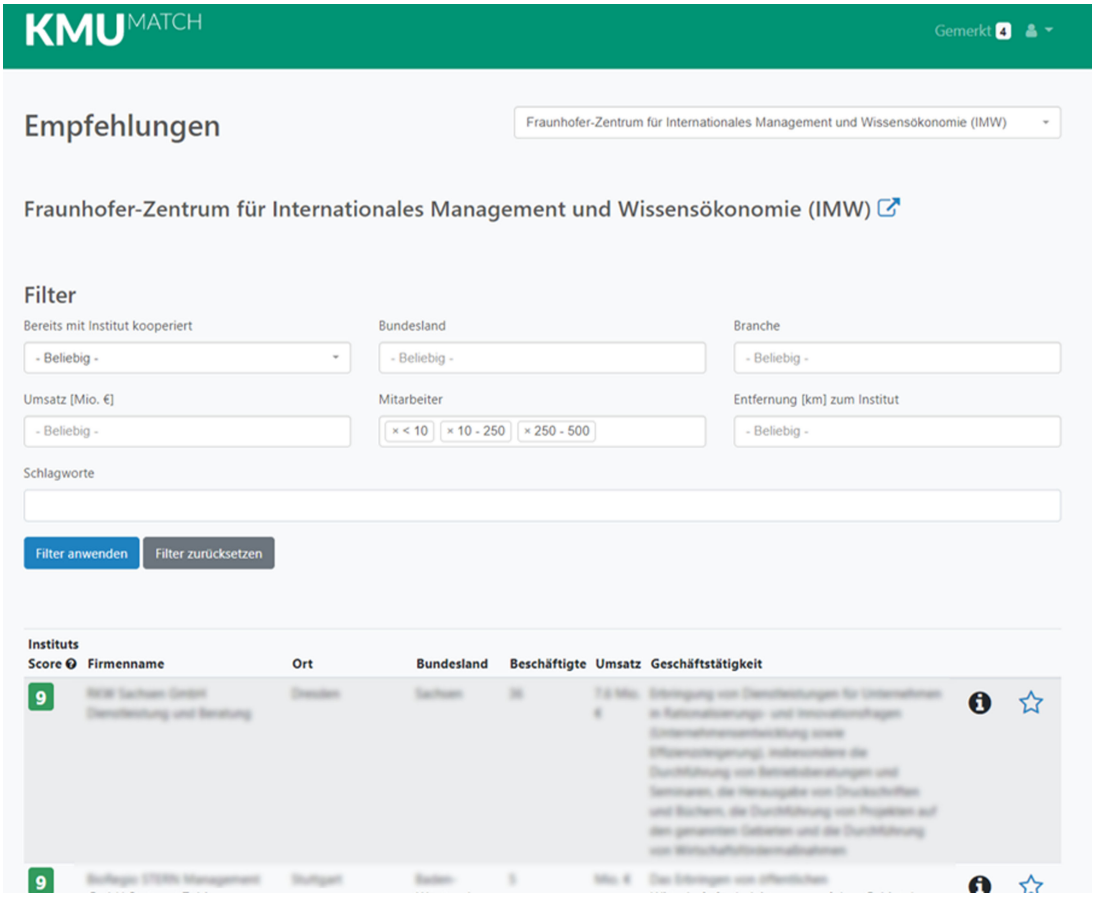

Fig. 6. Screenshot of prototype web-application SME-Match.

We prefilter companies coming from the Hoppenstedt database using a pooled classifier. Then, we calculate *Institute-Scores* for these prefiltered companies, which indicate how well a company fits a particular institute according to its industry-partner history.

Finally, we argue that this way of recommending potential good industry partners for public research organizations has a lot of transfer potential and is not restricted to our case study, the Fraunhofer Society. Considering the general nature of the partner selection problem, we point out that our model can be applied in various settings, both related to knowledge transfer and in a more general context. First, our approach can be easily transferred to any research organization that gathers historical cooperation data between firms and its internal research units. This is clearly the case for any university, with all its independent departments cooperating with companies. Although being independent, all those departments have to report their activities to the central administration, which therefore is in the position of using our method to find new good cooperation partners for its units. Any other public research organization with such a structure could also make use of our approach. Naturally, the performance could vary, depending on the quality of the available data. Second, any parent company with its independent subsidiaries could also take advantage of our approach to support partner selection. As before, the subsidiaries have also partners, for example, suppliers. This activity is sometimes reported to the parent company, which can benefit from its historical data with our approach. However, it is important to know what kind of partnership is registered in the historical data, because the mixture of historical cooperation data of different kinds, like suppliers and clients, can spoil the quality of the recommendations. In our case, the kind of partnership was clear; knowledge was transferred or acquired by the companies.

VII. Conclusion

Coming to the overall conclusion, we state that our classification approach indeed proved its feasibility and potential to support the industry-partner selection process at Fraunhofer, which answers our first research question. With precision and recall of around 85% for the pooled classification and ROC AUC scores between 70% and 90% for most institute-specific classifiers, the overall performance of the approach is certainly satisfactory for practical use. In doing so, it appears to be applicable to other public research organizations or even large corporations with a wide range of technologies and industry partners as well. Of course, given the necessary industry-partner data are available. Moreover, it allows for an almost complete screening of the

market, because it uses data that is available for almost all companies in Germany, which is easy to substitute with other texts or codes describing the business activity of companies.

In both steps of our classification approach, prefiltering with a pooled classifier and ranking relevant companies with institute-specific classifiers, the combination of economic features and industry-related features performs best. Economic information of a company on age, employees, and turnover supposedly helps to determine whether a company is able to cooperate with a research institute, with regard to financial capacity and absorptive capacity. Information on specific business activity of a company, extracted with NLP-techniques from texts or codes describing its industry affiliation, supposedly helps to determine whether this company fits the research area of a specific institute. With regard to our second research question, we therefore conclude that the combination of quantitative and qualitative information, economic data and business descriptions of companies in our case, is indeed useful for the partner selection support in public research.

References

[1] OECD, *Commercialising Public Research: New Trends and Strategies*. Paris, France: OECD, 2013.

[2] M. Perkmann and K. Walsh, "University-industry relationships and open innovation: Towards a research agenda," *Int. J. Manage. Rev.*, vol. 9, no. 4, pp. 259–280, 2007.

[3] M. Perkmann and K. Walsh, "The two faces of collaboration: impacts of university-industry relations on public research," *Ind. Corporate Change*, vol. 18, no. 6, pp. 1033–1065, 2009.

[4] H. W. Chesbrough, *Open Innovation: The New Imperative for Creating and Profiting From Technology*. Cambridge, MA, USA: Harvard Bus. Press, 2003.

[5] F. Battiston, J. Iacovacci, V. Nicosia, G. Bianconi, and V. Latora, "Emergence of multiplex communities in collaboration networks," *PloS One*, vol. 11, no. 1, 2016, Art. no. e0147451.

[6] D. Li, L. Eden, M. A. Hitt, and R. D. Ireland, "Friends, acquaintances, or strangers? Partner selection in R&D alliances," *AMJ*, vol. 51, no. 2, pp. 315–334, 2008, doi: 10.5465/amj.2008.31767271.

[7] J. Bruneel, P. d'Este, and A. Salter, "Investigating the factors that diminish the barriers to university-industry collaboration," *Res. Policy*, vol. 39, no. 7, pp. 858–868, 2010.

[8] B. Yoon and B. Song, "A systematic approach of partner selection for open innovation," *Ind. Manage. Data Syst.*, vol. 114, no. 7, pp. 1068–1093, 2014.

[9] A. Wirsich, A. Kock, C. Strumann, and C. Schultz, "Effects of university-industry collaboration on technological newness of firms," *J. Product Innov. Manage.*, vol. 33, no. 6, pp. 708–725, 2016.

[10] H. Hottenrott and C. Lopes-Bento, "R&D partnerships and innovation performance: Can there be too much of a good thing?" *J. Product Innov. Manag.*, vol. 33, no. 6, pp. 773–794, 2016.

[11] K. Lee, I. Park, and B. Yoon, "An approach for R&D partner selection in alliances between large companies, and small and medium enterprises (SMEs): Application of Bayesian network and patent analysis," *Sustainability (Switzerland)*, vol. 8, no. 2, pp. 1–18, 2016, doi: 10.3390/su8020117.

[12] J. Jeon, C. Lee, and Y. Park, "How to use patent information to search potential technology partners in open innovation," *J. Intellectual Property Rights*, vol. 16, pp. 385–393, 2011.

[13] T. Broekel and M. Binder, "The regional dimension of knowledge transfers—A behavioral approach," *Ind. Innov.*, vol. 14, no. 2, pp. 151–175, 2007.

[14] C. M. Beckman, P. R. Haunschild, and D. J. Phillips, "Friends or strangers? Firm-specific uncertainty, market uncertainty, and network partner selection," *Org. Sci.*, vol. 15, no. 3, pp. 259–275, 2004.

[15] A. L. Porter and S. W. Cunningham, *Tech Mining: Exploiting New Technologies for Competitive Advantage*. Hoboken, NJ, USA: Wiley, 2004.

[16] S. H. Chen, P. W. Wang, C. M. Chen, and H. T. Lee, "An analytic hierarchy process approach with linguistic variables for selection of an R&D strategic alliance partner," *Comput. Ind. Eng.*, vol. 58, no. 2, pp. 278–287, 2010.

[17] I. Park, Y. Jeong, B. Yoon, and L. Mortara, "Exploring potential R&D collaboration partners through patent analysis based on bibliographic coupling and latent semantic analysis," *Technol. Anal. Strategic Manage.*, vol. 27, no. 7, pp. 759–781, 2015.

[18] X. Wang *et al.*, "Identifying R&D partners through subject-action-object semantic analysis in a problem & solution pattern," *Technol. Anal. Strategic Manage.*, vol. 29, no. 10, pp. 1167–1180, 2017.

[19] C. Wu and D. Barnes, "Partner selection in agile supply chains: a fuzzy intelligent approach," *Prod. Planning Control*, vol. 25, no. 10, pp. 821–839, 2014, doi: 10.1080/09537287.2013.766037.

[20] C. Wu and D. Barnes, "A literature review of decision-making models and approaches for partner selection in agile supply chains," *J. Purchasing Supply Manage.*, vol. 17, no. 4, pp. 256–274, 2011, doi: 10.1016/j.pursup.2011.09.002.

[21] T. K. Das and I. Y. He, "Entrepreneurial firms in search of established partners: Review and recommendations," *Int. J. Entrepreneurial Behav. Res.*, vol. 12, no. 3, pp. 114–143, 2006, doi: 10.1108/13552550610667422.

[22] S. Baiman, P. E. Fischer, and M. V. Rajan, "Information, contracting, and quality costs," *Manage. Sci.*, vol. 46, no. 6, pp. 776–789, 2000, doi: 10.1287/mnsc.46.6.776.11939.

[23] S. Tello, S. Latham, and V. Kijewski, "Individual choice or institutional practice," *Manage. Decis.*, vol. 48, no. 8, pp. 1261–1281, 2010, doi: 10.1108/00251741011076780.

[24] G. Gigerenzer and W. Gaissmaier, "Heuristic decision making," *Annu. Rev. Psychol.*, vol. 62, pp. 451–482, 2011, doi: 10.1146/annurev-psych-120709-145346.

[25] Y. Geum, S. Lee, B. Yoon, and Y. Park, "Identifying and evaluating strategic partners for collaborative R&D: Index-based approach using patents and publications," *Technovation*, vol. 33, no. 6/7, pp. 211–224, 2013.

[26] J. Bruneel, P. d'Este, and A. Salter, "Investigating the factors that diminish the barriers to university–industry collaboration," *Res. Policy*, vol. 39, no. 7, pp. 858–868, 2010, doi: 10.1016/j.respol.2010.03.006.

[27] M. S. Granovetter, "The strength of weak ties," in *Social networks*. New York, NY, USA: Elsevier, 1977, pp. 347–367.

[28] M. Fritsch and M. Kauffeld-Monz, "The impact of network structure on knowledge transfer: An application of social network analysis in the context of regional innovation networks," *Ann. Regional Sci.*, vol. 44, no. 1, 2008, Art. no. 21, doi: 10.1007/s00168-008-0245-8.

[29] R. Martin and P. Sunley, "Path dependence and regional economic evolution," *J. Econ. Geography*, vol. 6, no. 4, pp. 395–437, 2006.

[30] H. Chang, J. Gausemeier, S. Ihmels, and C. Wenzelmann, "Innovative technology management system with bibliometrics in the context of technology intelligence," in *Lecture Notes in Electrical Engineering, 6, v.v. 6, Trends in Intelligent Systems and Computer Engineering*. O. Castillo, L. Xu, and S.-I. Ao, Eds., 1st ed., New York, NY, USA: Springer, 2008, pp. 349–361.

[31] M. Z. Solesvik and S. Encheva, "Partner selection for interfirm collaboration in ship design," *Ind. Manage. Data Syst.*, vol. 110, no. 5, pp. 701–717, 2010.

[32] A. Giunta, F. M. Pericoli, and E. Pierucci, "University-industry collaboration in the biopharmaceuticals: The Italian case," *J. Technol. Transfer*, vol. 41, no. 4, pp. 818–840, 2016.

[33] B.-Y. Eom and K. Lee, "Determinants of industry-academy linkages and, their impact on firm performance: The case of Korea as a latecomer in knowledge industrialization," *Res. Policy*, vol. 39, no. 5, pp. 625–639, 2010.

[34] U. Kaiser, "An empirical test of models explaining research expenditures and research cooperation: Evidence for the German service sector," *Int. J. Ind. Org.*, vol. 20, no. 6, pp. 747–774, 2002.

[35] K. Roigas, P. Mohnen, and U. Varblane, "Which firms use universities as cooperation partners? A comparative view in Europe," *Int. J. Technol. Manage.*, vol. 76, no. 1/2, pp. 32–57, 2018, doi: 10.1504/IJTM.2018.10009595.

[36] D. Comin, G. Licht, M. Pellens, and T. Schubert, "Do companies benefit from public research organizations? The impact of the Fraunhofer Society in Germany," ZEW—Leibniz Centre for European Economic Research, ZEW Discussion Papers, No 19-006, 2019. [Online]. Avialable: https://EconPapers.repec.org/RePEc:zbw:zewdip:19006

[37] P. Cardamone, V. Pupo, and F. Ricotta, "University technology transfer and manufacturing innovation: The case of Italy," *Rev. Policy Res.*, vol. 32, no. 3, pp. 297–322, 2015.

[38] R. Rybnicek and R. Königsgruber, "What makes industry–university collaboration succeed? A systematic review of the literature," *J. Bus. Econ.*, vol. 89, no. 2, pp. 221–250, 2019, doi: 10.1007/s11573-018-0916-6.

[39] Fraunhofer, "Annual report," Munich, Germany 2017. Accessed: Jun. 28, 2019. [Online]. Available: https://www.fraunhofer.de/en/media-center/publications/fraunhofer-annual-report.html

[40] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 452–461.

[41] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua, "Fast matrix factorization for online recommendation with implicit feedback," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2016, pp. 549–558.

[42] C. Rammer *et al.*, "Innovationen in der deutschen Wirtschaft: Indikatorenbericht zur Innovationserhebung 2018," ZEW Innovationserhebungen-Mannheimer Innovationspanel (MIP), 2019.

[43] J. Silge and D. Robinson, "Tidytext: Text mining and analysis using tidy data principles in R," *J. Open Source Softw.*, vol. 1, no. 3, 2016, Art. no. 37, doi: 10.21105/joss.00037.

[44] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics: Short Papers-Volume 2*, 2012, pp. 90–94.

[45] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, Vol. 2*, 2017, pp. 427–431.

[46] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Aug. 2008.

[47] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[48] J. D. Malley, J. Kruppa, A. Dasgupta, K. G. Malley, and A. Ziegler, "Probability machines: Consistent probability estimation using nonparametric learning machines," *Methods Inf. Med.*, vol. 51, no. 1, pp. 74–81, 2012, doi: 10.3414/ME00-01-0052.

[49] M. N. Wright and A. Ziegler, "Ranger: A fast implementation of random forests for high dimensional data in C++ and R," *J. Statist. Softw.*, vol. 77, no. 1, pp. 1–17, 2017, doi: 10.18637/jss.v077.i01.

[50] B. Bischl *et al.*, "MLR: Machine learning in R," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 5938–5942, 2016.

[51] P. Probst, Q. Au, G. Casalicchio, C. Stachl, and B. Bischl, "Multilabel classification with R package MLR," *R J.*, vol. 9, no. 1, pp. 352–369, 2017.