



Article

Inside Production Data Science: Exploring the Main Tasks of Data Scientists in Production Environments

Arno Schmetz ^{1,*}  and Achim Kampker ^{1,2} ¹ Fraunhofer Research Institution for Battery Cell Production FFB, Bergiusstraße 8, 48165 Münster, Germany² Production Engineering of E-Mobility Components, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany

* Correspondence: arno.schmetz@ffb.fraunhofer.de

Abstract: Modern production relies on data-based analytics for the prediction and optimization of production processes. Specialized data scientists perform tasks at companies and research institutions, dealing with real data from actual production environments. The roles of data preprocessing and data quality are crucial in data science, and an active research field deals with methodologies and technologies for this. While anecdotes and generalized surveys indicate preprocessing is the major operational task for data scientists, a detailed view of the subtasks and the domain of production data is missing. In this paper, we present a multi-stage survey on data science tasks in practice in the field of production. Using expert knowledge and insights, we found data preprocessing to be the major part of the tasks of data scientists. In detail, we found that tackling missing values, finding data point meanings, and synchronization of multiple time-series were often the most time-consuming preprocessing tasks.

Keywords: data science; production data; smart manufacturing; Industry 4.0



Citation: Schmetz, A.; Kampker, A. Inside Production Data Science: Exploring the Main Tasks of Data Scientists in Production Environments. *AI* **2024**, *5*, 873–886. <https://doi.org/10.3390/ai5020043>

Academic Editor: Gianni D'Angelo

Received: 30 April 2024

Revised: 22 May 2024

Accepted: 5 June 2024

Published: 12 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Modern production is undergoing the transition to Industry 4.0 and smart manufacturing. In this kind of production, advanced digital technologies are integrated into manufacturing processes, aiming for insights, optimizations, and sustainability throughout the production line [1,2]. New products and components like industrial Internet of things platforms (IIoT) [3], artificial intelligence frameworks (AI) [4], fog computing [5], and services like predictive maintenance [6] have arisen worldwide. For example, in the battery cell production domain, digital transformation is key to the competitive production of European manufacturers. Regarding the expected growth in the market and announced factories, several digitalization use cases will allow billions of dollars in savings in the coming years [7].

The central and crucial requirement for smart manufacturing integration are production data. All improvements in production systems are based on data and therefore directly depend on the availability and quality of production data. Therefore, data pipelines are not just upgrades of the operations of production but strategically significant aspects of the success of the whole enterprise.

For the tasks of data exploration, analytics, and model development, as well as data-based optimization of production, *data scientists* as a profession provide their specialized expertise. While data science as a profession itself is independent of the manufacturing domain, this domain itself has specific challenges and contexts. In consequence, data scientists specializing in the field of production are required, and they are a scarce resource at the moment [8,9]. While the term and a basic understanding of data scientist could first be found back in the 1970s, their relevance and recognition has increased in recent years with the higher availability and complexity of data [10].

As modern production heavily relies on data and as data scientists work with data from production, several tasks arise for data scientists to tackle during their work. Depending on the company or institution, these may start from the acquisition of data or importing data using an ETL (extract transform load) process [11] or other starting points, while other data scientists may conclude different tasks besides the core analytics part. In particular, tasks impacting or impacted by data quality are crucial aspects of that core analytics work [12]. Further, the expectations of the actual work and increasingly relevant aspects like ethical, legal, or societal aspects have further changed the understanding of a data scientist's work [13].

While these changes and the differences in real tasks hold for data scientists in all fields, no current distinct observation on the actual main tasks and their load is known to the authors for the domain of production data science. In this paper, we present the results of expert interviews and surveys regarding the tasks of current production data scientists and the potential to leverage the efficiency of data scientists in the production domain.

2. Background and Related Work

The term “data scientist” itself can be traced back to the 1970s [10], including the first description of the work performed by such a person. In the time since then, with more data available, better computational resources and software, and especially the rise of artificial intelligence, the working definition of data scientists is changing and is not consistent. When looking for current positions for a data scientist online, we found more than 600 open positions in Germany. The technical tasks described in those positions varied, but some terms and skills were present in the vast majority of the positions. These common skills and tasks included machine learning or artificial intelligence, alongside statistical analytics, software development and engineering, and deriving optimization possibilities. Further tasks were less common throughout the positions and included data acquisition, including sensor selection, database management, data and IT security, visualization tasks, teaching and training, model deployment and monitoring, creating new software platforms, automation and CI/CD, data preparation, large language model (LLM) query engineering, and others. From this, the variety in the definition of the job data scientist can be seen, while some aspects show a common core [14].

In [15], the authors performed a survey of the literature, deriving the most relevant skills required by data scientists. Their key findings described the need for a variety of different domain skills for data scientists. Technical skills like programming, statistics, and math were joined by business understanding and communication skills. For the data focus, the skill to deal with structured and unstructured as well as low data quality was required. The authors proposed the distinction of data scientists dealing with modeling, decision support, and deriving new product features from the role of *data analyst*, which focuses on data queries, pure programming, and data quality assessment. Nevertheless, they concluded that this distinction was not made by most of their observed sources. In the end, they presented a table of 44 relevant skills for data scientists, highlighting business, communication, and statistics as the most critical skills.

In [16], the authors performed a classification of data scientists' work and performed expert interviews with 21 experts from different fields and backgrounds. Further, they described common data practices named by the experts and investigated the requirements for time, domain knowledge, representation, and pattern detection skills. They showed that the integration and cleaning of data, as well as the feature engineering, are time-consuming but also require expertise in pattern detection and specific domain knowledge. They concluded that the impact and requirements of human expertise in the data process make real data science highly dependent on the people performing the tasks and they are not transparent enough currently.

The authors in [17] performed a survey of tasks for a data science pipeline. They used 71 proposals from research, 105 Kaggle implementations, and 21 mature data science projects available on GitHub. The detailed ideal pipeline they found is shown in Figure 1.

In the beginning, a general study or case design is created, which results in the data acquisition step, which leads to the data preparation, with the data storage concluding the pre-processing layer. In the model building layer, feature engineering starts and one can also go back to the preprocessing (dashed arrows). After feature engineering, the actual modeling, training, and evaluation are performed. This feeds back to the feature engineering and forward to the prediction step. The post-processing layer starts with the interpretation of the results and model, which also includes visualization for communication. The final step describes the deployment, which optionally feeds back to the model evaluation. Using this pipeline, the authors investigated the frequency of the steps in their sources. They found the modeling and the preparation steps to be the most significant steps, while the interpretation and communication steps were the least common steps. Layer-wise, almost all sources included preprocessing and modeling layers (96% each), while post-processing layers were included infrequently (52%). The authors concluded that, depending on the project size, some stages can be omitted, but some significant steps are crucial, and a practical investigation should be performed in the future.

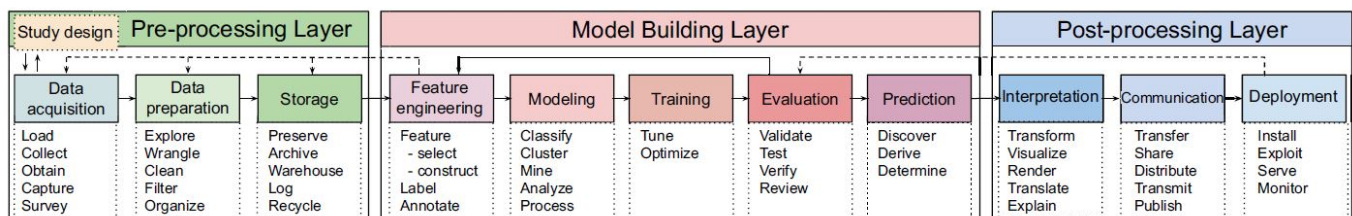


Figure 1. The data science pipeline from preprocessing steps (green block) via the modeling layers (red block) to the post-processing steps (blue block). Taken from [17].

CrowdFlower performed an industry survey on the actual tasks and the current problems data scientists see in their work [18]. They showed that the major part of the work deals with the cleaning and organizing of data and the collection of datasets, as shown in Figure 2. Compared with pipelines like those presented in Figure 1, these tasks may aggregate multiple stages found there. Further, the survey noted that 83% of the respondents said there was a lack of data scientists at their site to perform all the work required. Several years later, the parent company published a new survey [19] that focused on AI instead of data scientists as a whole. While the survey was less detailed about workloads than the CrowdFlower report, it showed that the main steps of data sourcing, data preparation, modeling, and human evaluation were all significant in terms of time spent on them, including a high variety in the answer ranges. The authors concluded that the time spent in managing and preparing data is high but slightly trending down.

Anaconda Inc. regularly performs a user survey and compiles this into a *state of data science* report. In 2023, the survey showed the order of the most time-consuming steps for data scientists. Data preparation and cleaning, along with visualization, were denoted the most time-consuming steps, while deployment and reporting were at the other end [20]. In the study for 2022, less qualitative and more quantitative results were shown. In that survey, the most significant steps were data cleansing (~26% of time) and visualization (~21% of time), while the deployment and model selection steps were the least time-consuming steps in comparison [21].

The different scientific publications showed different viewpoints, since [15] comes from a mobile computing and business background, [16] from a human–computer interaction background, and [17] from a software engineering background. At the same time, practical studies [18–22] tried to present a user-centered overview on the whole domain of data scientists. The actual numbers and ordering of the most relevant and time-consuming steps were widely inconsistent, making the surveys and research hard to compare. While all sources are united in their understanding of the crucial role of domain-specificity, there is a lack of publications dealing with the specific views of manufacturing data scientists.

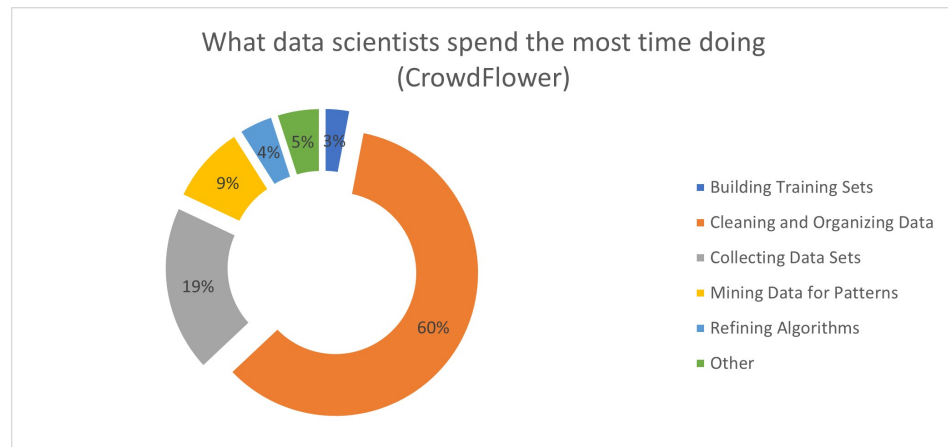


Figure 2. Results of the CrowdFlower Data Science Report regarding the actual time workload of data scientists. Based on [18].

3. Methodology

For this paper, we performed a multi-stage survey of the tasks of data scientists in the domain of manufacturing. The target of the survey was the identification of the most time-consuming steps for data scientists, with a rough indication compared to other steps. Further, personal views, in terms of tasks they liked or disliked performing, were of interest. Lastly, the respondents' view on the data science profession in manufacturing and possible blockers was targeted. To obtain responses for the manufacturing domain, data scientists from that domain were specifically targeted. The steps of the multi-stage survey are described below.

3.1. Pre-Study

In the first step, a literature review and pre-selection of tasks to consider were performed, to build the main survey. The condensed results of the literature review are shown in Section 2 in this paper. Based on the review, the lack of specific surveys in the manufacturing domain and lack of comparable numbers were verified. Using the data pipeline definitions from the review, a short pre-study with a small set of experts was performed to finalize the step selection and description. With the study design, all layers mentioned in the pipeline from [17] should be addressed. The final list of steps is shown in Table 1.

Some steps like the deployment were omitted, while other steps were aggregated based on feedback that they were hard to distinguish between in practice (e.g., interpretation and communication steps). Regarding the amount of time (options per step in the survey), an increasing system, starting with 5% ranges for minor tasks and larger ranges up to 30% for more dominant tasks was used.

Since a variety of steps were named in the field of data cleaning [23,24], pre-selection of data cleaning and organization steps was performed, focusing on the generally most significant steps in manufacturing. The final list of steps was condensed as follows:

- Matching data types (e.g., string to double).
- Data value conversions (e.g., Fahrenheit to Celsius).
- Filling missing data.
- Removing outliers.
- Synchronizing multiple data streams (e.g., matching time-series data from multiple sensors).
- Reduction of data.
- Binning values.
- Duplicate removal.
- Text cleaning (e.g., remove HTML tags or blanks).
- Data normalization.
- Setting up and managing databases.
- Asking people for meaning of variables and data point naming.

Table 1. The final list of data science tasks for the study, resulting from the pre-selection.

Step	Description
Data Acquisition	Acquisition of data from sensors, sensor integration, machine interfaces, manual data
Collecting Datasets	Finding and collecting databases, asking people for sources
Building Data Pipelines	Creating software for transferring data between systems, configuration of industrial Internet of things (IIoT) platforms, setting up middleware software
Cleaning and Organizing Data ¹	Assessment of data quality, cleaning steps, synchronization, mapping
Pattern Searching	Data mining, data exploration, explorative statistical analytics, pattern recognition
Creating and Training Models	Selection of models, setting up and configuration of models, (active time for) model training, programming statistical models
Refining and Retraining Models	Evaluation of model performance, (active time for) retraining, changing and reprogramming models, refining model configuration
Implementing visualizations	Creating dashboards, configuration of visualization software for live data, creating presentations for users, performance monitoring dashboard and reporting

¹ Investigated in further detail in survey.

3.2. Main Survey

Using the results from the pre-selection and review, the main survey was created. The first question dealt with the data science tasks as presented in Table 1 and used the increasing time-indication selection as described above, moving from 0–5% to >70%. This choice mitigated potential issues, including biases in measurements, uncertainties inherent to people, and natural variability in day-to-day user estimations. Since the study did not aim for exact numbers, this reduced the impact of outliers and anomalies potentially skewing the interpretation. Another option would have been the use of fuzzy logic, but the first approach was chosen to also reduce the load on respondents. As a consequence, the actual numbers for workload might not add up to exactly 100% for all respondents, which was not found to be a significant problem statistically. The second major question used the same list of tasks and asked the users to indicate how much they did (not) enjoy each of the tasks on a 5-step rating scheme. In addition, people could add “Other” for significant tasks they performed that were not targeted by the survey.

The third major question dealt with a deep dive into the data cleaning and organizing tasks, as presented in the list above. In this question, the data scientists were asked to give an indication of which are the most/least time-consuming of these tasks on a 5-step scheme ranging from “Least/No Time” to “Most/Much Time”. Again, users were given the opportunity to add other tasks here, including their indication.

In the last segment of the survey, respondents were asked generally about their satisfaction with the 5-star scheme and what they would need for greater satisfaction (free text). The last question dealt with the domain where the users performed data science. This question was used to filter responses that were not aligned with the main target audience of the study, manufacturing domain data scientists. Since people might use data linked from domains, the question allowed multiple-choice answers.

As the aim of this study concerned data scientists’ tasks in the manufacturing domain, questions about personal information (age, gender, etc.) were discussed. Since the target audience is rather specific, there was a high probability of small sample sizes, risking the anonymity of the evaluation. In compliance with GDPR Article 5(1)(c) [25] for data minimization, such questions were omitted, since they were not in line with the main goal of the survey.

4. Main Survey

The survey presented in Section 3.2 was implemented using Microsoft Forms. The survey’s target audience was the set of data scientists in the manufacturing domain. Therefore, the survey was actively promoted throughout international and national conferences, public events, networks like LinkedIn and SurveyCircle [26], and through direct consultation with experts in the industry. All promotions clarified the target audience. Since we started

from a single point and searched a specific group (data scientists in the production domain), the main sampling method was based on snowball sampling, starting from the authors, moving to domain experts, and propagating through their networks. To provide some additional random seeding, independent networks like SurveyCircle were used to expand the reach. Due to the specific audience, the survey was available for 15 months, to ensure a suitable amount of responses. Due to the sampling method and European events, most respondents were likely to come from Germany and Europe, and less likely from other continents.

In the named timeframe, the survey resulted in 89 responses. Of those responses, only 81 were from the targeted audience. The other 8 responses, not in the manufacturing domain, indicated their domain(s) as finance and banking (4 persons), social media and eCommerce (3 persons), education (1 person), enterprise processes (2 persons), government (1 person), and medical (1 person). To align the results with the target audience, these responses were excluded and deleted from the results. All subsequent results present the filtered results only, using the 81 respondents. The detailed results of the survey are shown below, as well in the Supplementary Materials. The following subsections describe the results. A discussion and interpretation is given in the section thereafter.

4.1. Data Science Tasks

The first question of the survey dealt with the time spent on the tasks as a share of the working time of the data scientists, as described in Table 1. The results are shown in Figure 3.

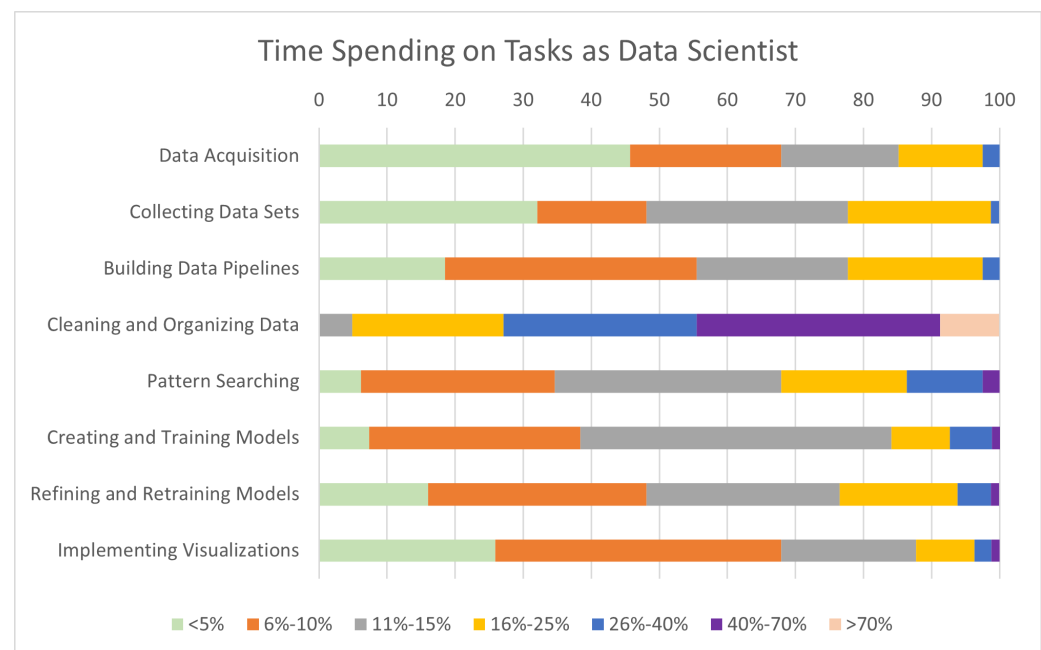


Figure 3. Results for the question regarding the actual time spent as a share of the overall work the data scientists performed in percent. The option “Other” is omitted here and described in the text.

For the task of *Data Acquisition*, 45.7% indicated this task as being from 0 to less than 5 percent of their work share. Two responses indicated a significant share of up to 40 percent of their work. For *Collecting Datasets*, 48.1% named this task as making up to 10 percent of their work, while one respondent named this task as making up to 40 percent of their work. The highest number of responses (37%) for the *Building Data Pipelines* named this task as responsible for 6 to 10 percent of their work time. No one named the *Cleaning and Organizing Data* task as less than 11 percent of their work time, while 44.4% named this task as representing more than 40 percent of their work time. Moreover, 61.7% of the responses named the *Pattern Searching* as taking 6 to 15 percent of their work time. The same work

share was named by 76.69% of the responses for *Creating and Training Models*. *Refining and Retraining Models* had a work share of up to 10 percent according to 48.1% of the responses. A significant 42% of the responses named *Implementation of Visualizations* as responsible for 6 to 10 percent of their work time.

There were three responses in addition to the percentages, as free text. One respondent named *Deployment* as a task with a significant share of work. One respondent wrote that they performed “almost 4 days working on managing, cleaning and converting” per week. The last person named “interviews” as making up a significant share of their work, without further description.

4.2. Joy of Tasks

The second major question of the survey targeted the joy people associated with the tasks described above. A qualitative scheme from “least enjoyable” to “most enjoyable” was used. The results are shown in Figure 4.

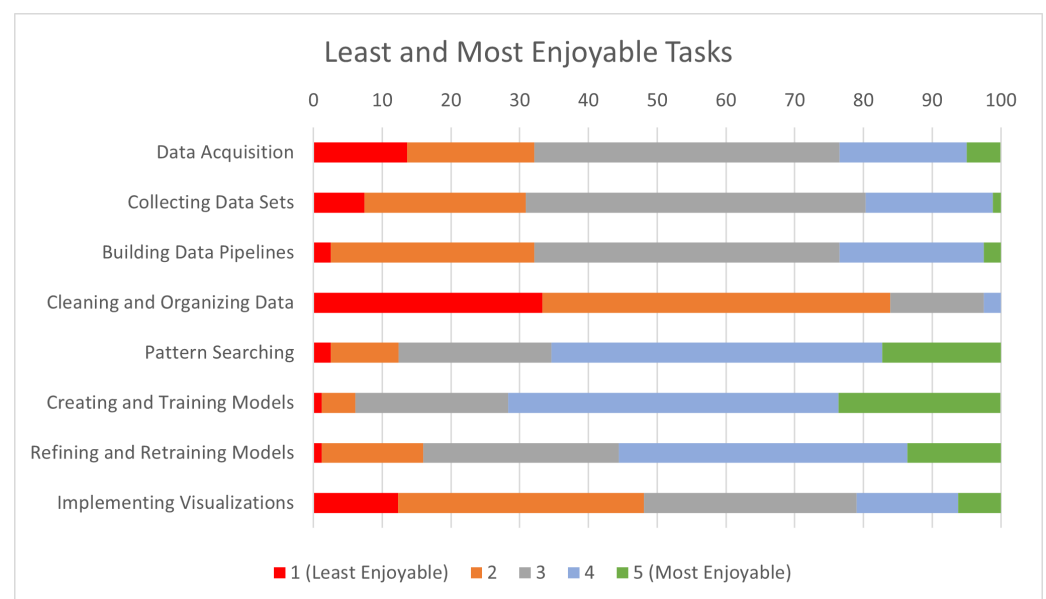


Figure 4. Results for the question regarding how much in comparison the data scientists enjoyed or did not enjoy the tasks they performed. The answer “Other” is omitted in this figure and described in the text.

For the tasks *Data Acquisition*, *Collecting Datasets*, and *Building Data Pipelines*, the most common answer was the neutral response, with almost half of the respondents (44.4%, 49.4%, and 44.4% respectively). The task of *Cleaning and Organizing Data* was named the most enjoyable task by no respondents, while 33.3% of the respondents named it the least enjoyable task. *Pattern Searching*, *Creating and Training Models*, and the task of *Refining and Retraining Models* were the most common answers tending towards the enjoyable task side (rating 4), with shares of 48.1%, 48.1%, and 42.0%, respectively. Meanwhile, 48.1% of the respondents placed the task of *Implementing Visualizations* on the less enjoyable side (ratings 1 and 2) of tasks. In addition, the mean ratings for the joy of the tasks are shown in Figure 5.

There were seven additional responses as free text to this question. One response rated “Deployment” as the least enjoyable task. One response rated “data augmentation with the engineering and guys at the machines” less enjoyable (rating 2). The same rating (2) was given by a respondent for the task “Selection of sensors”. One respondent found the task of “interviews” the most enjoyable task. One person gave “pre-processing also feature calculation” a neutral rating. Another respondent found the task of “Creating Reports” least enjoyable. The last respondent named “Reporting & Bürokratie” (German for “bureaucracy”) as the least enjoyable task.

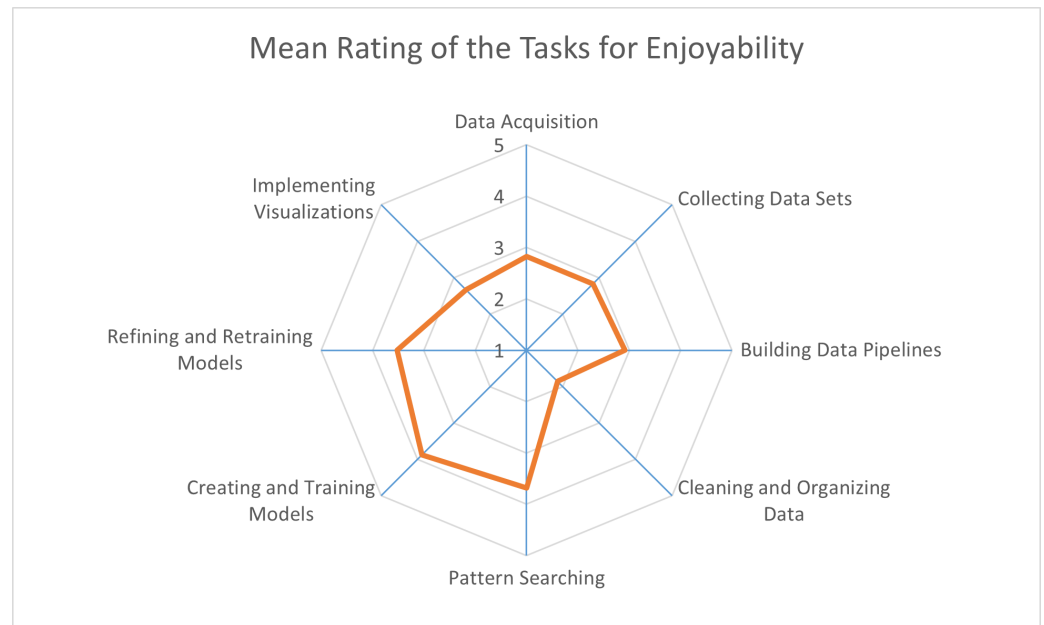


Figure 5. Mean ratings (mean of rating value in the responses) for the question regarding how much in comparison the data scientists enjoyed or did not enjoy the tasks they performed. The answer “Other” is omitted in this figure and described in the text.

4.3. Deep Dive Data Cleaning and Organizing

The third major question in the survey dealt with the detailed steps in the data cleaning and organizing task. The rating for the steps presented above was performed using a qualitative 5-step scheme, from the least to the most time spent on the steps. The results are shown in the Figure 6.

None of the respondents named the step *Matching Data Types* as the most time spent, while most respondents selected neutral or lesser time needed (ratings 3 and 2) with 43.2% and 34.6%. The step of *Data Value Conversion* was most commonly rated less than neutral (rating 2 at 43.2%) or least time (rating 1 at 32.1%) consuming. A total of 56.7% of the respondents named the step *Filling Missing Values* as more or the most time-consuming inside the task of cleaning and organizing. *Removing Outliers* was mostly rated around the center, resulting in only 4.9% for the least and 1.2% for the most time spent on this step by the data scientists. This step had a mean rating of ca. 2.90. The step of *Synchronizing Multiple Data Streams* obtained the largest response for the rating of more time spent (rating 4), at 69.1%. In contrast, the *Reduction of Data* had a majority for less time spent (rating 2) at 53.1%. For the *Binning Values* steps, the neutral response (rating 3) was most commonly selected, at 40.7%. While 1.2% found *Duplicate Removal* the most time-consuming, 18.5% found the same step the least time-consuming. The step of text-cleaning was rated the least time-consuming step by the majority (58%) of the respondents, resulting in the lowest mean rating of ca. 1.53. The most common rating for the step *Data Normalization* was the lesser time option (rating 2), at 38.3%. *Setting up and Managing Databases* was named the least time-consuming step by 35.8% of the respondents. The step *Asking People for the Meaning of data* was named the most time-consuming task by 29.6% and obtained the highest mean rating of 3.89. Figure 7 shows the mean ratings for this question.

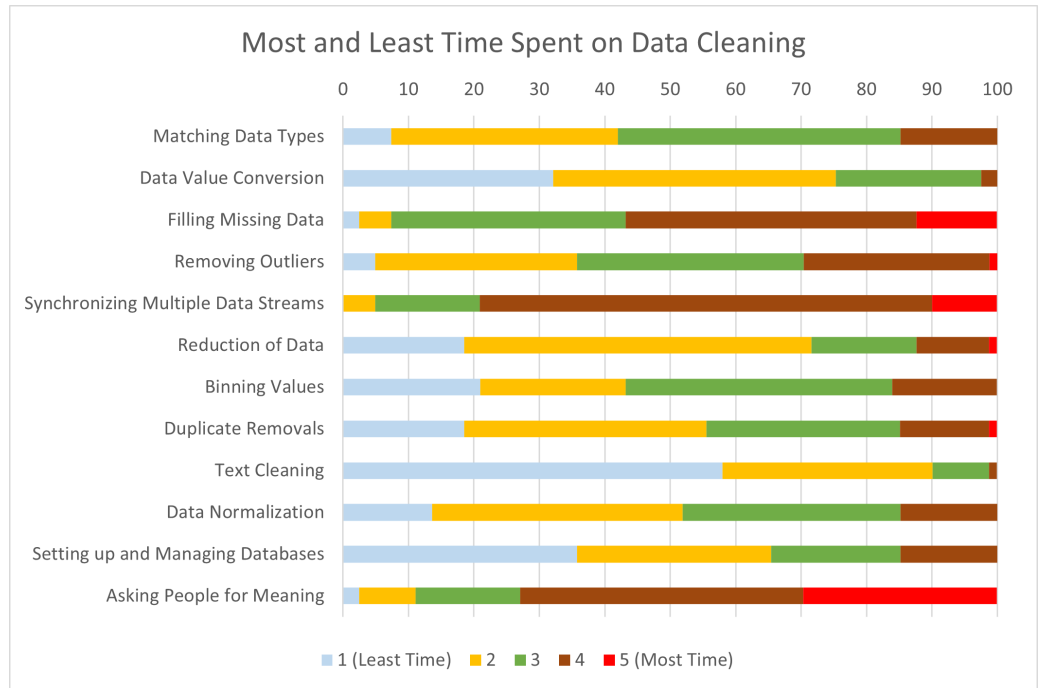


Figure 6. Results for the question regarding how much time in comparison the data scientists spend on the different data cleaning and organizing steps. The answer “Other” is omitted in this figure and described in the text.

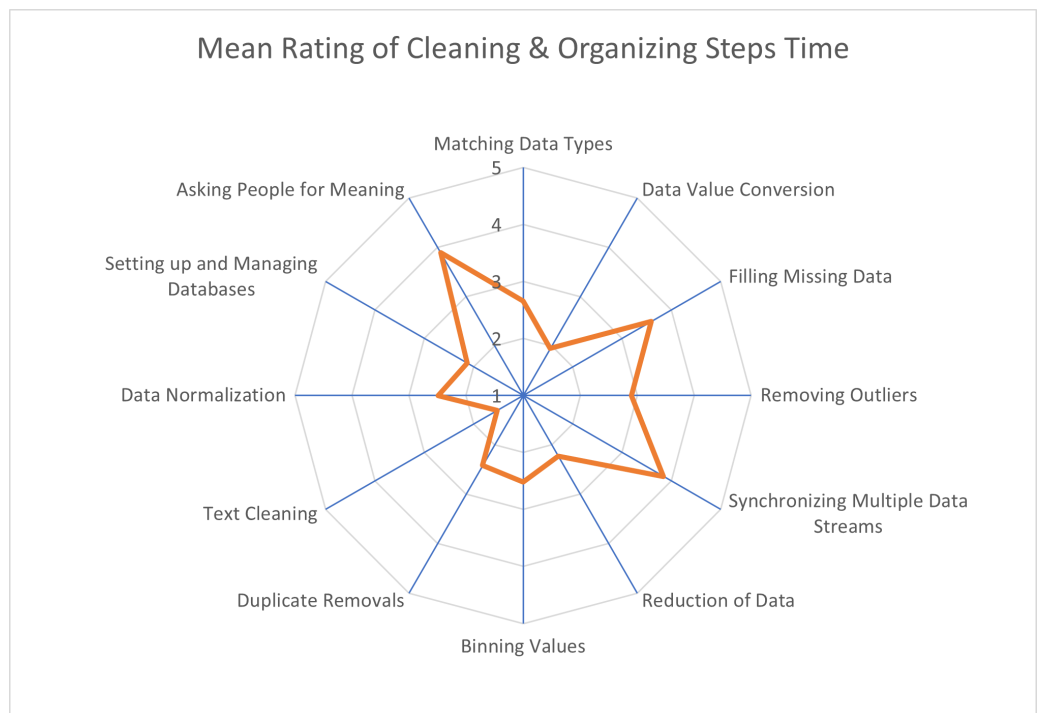


Figure 7. Mean ratings for the question regarding how much in comparison the data scientists spent time on the different data cleaning and organizing steps. The answer “Other” is omitted in this figure and described in the text.

There were six additional responses for steps in the cleaning and organizing task. One respondent wrote “Get to know what is happening in production departments and what the data is referring to”. Another response just named “I had to check other”, without further indication or description. Two answers dealt with the specific image data labeling

step. Another respondent named file format conversion and the last respondent named the importing of CSV as a relevant explicit step.

4.4. Satisfaction

The last major question block dealt with the overall satisfaction of the data scientists with their profession and what they needed to increase satisfaction. Figure 8 shows the overall satisfaction rating of the respondents. In total, with the 5-star rating scheme, the mean value was 4.49 stars.

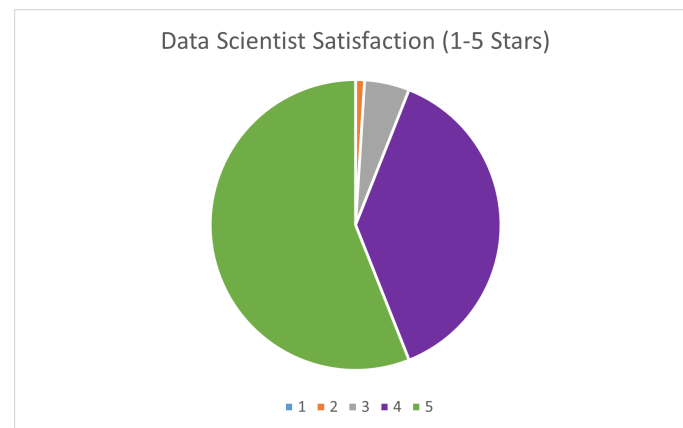


Figure 8. Results for the question regarding the overall satisfaction of the data scientists with their profession.

A total of 19 responses to the free text for requirements to increase satisfaction were given. Need for higher *data quality* (incl. labeling and documentation) was named by eight of the respondents. Three respondents named the need for better *standards and naming* of data. Three responses dealt with the need for *more data* to perform analytics. Two respondents stated a need for *more time for data science and less pre-processing* explicitly. Two responses named a need for *more information or contact with the production teams*. In addition, two responses named the problem of *machine and sensor interfaces*. Further responses described a need for a *relationship database, multilanguage data support, and less reporting*.

5. Discussion

In this section, the results above are interpreted and put into context regarding the survey goals.

5.1. Results Interpretation

Time Spending. The multi-stage survey presented above targeted the actual task loads, the enjoyability of these, a deep dive into the cleaning and managing of data, and general satisfaction with the data science profession. The respondents were filtered for specific manufacturing domain data scientists. For the most time-consuming tasks of the data scientists, the task of cleaning and organizing data was dominant. Almost half of the data scientists spent at least 40% of their time only on that task. Since less than 5% spent less than 16% of their time performing this task, this dominance was crucial to all data scientists.

The tasks of data acquisition, collecting datasets, and visualizations played minor roles, if at all, for most data scientists. The reason for this might be that those tasks are at the edges of the core of data scientists' work and are performed by other professions like data engineers and data analysts [15]. Nevertheless, based on the description of those other professions, the cleaning and organizing of data could also be (partially) performed by them.

Figure 1 shows the layers of the data science pipeline on which the survey was based. The tasks linked to the model-building layer, pattern searching, creating models, and retraining represented a significant part of the working time of the data scientists'

but fell behind the dominance of cleaning and organizing. In total, the aggregated tasks linked to the preprocessing layer dominated the work of the data scientists, while the model-building layer tasks were a significant secondary part of the work time. The time taken for the tasks of the post-processing layer fell behind, even when including the free text responses linked to post-processing in terms of deployment.

Enjoyability. Most of the tasks showed that data scientists enjoyed them, while some others did not. Significantly, the tasks linked to the model-building layer were enjoyed the most by the data scientists, with a mean rating from 3.52 to 3.88. The tasks outside that layer fell behind in the rating by a large margin. The dominant task, in terms of time spent, of cleaning and organizing data was rated by far the least enjoyable task (mean rating 1.85). It also had a large gap to the second least enjoyable task of implementing visualizations (mean rating 2.67). These results show that tasks in the central layer of data science were liked, while tasks outside were less liked by the data scientists. In particular, the task the most time is spent on is the least enjoyed task. This means data scientists are likely to spend most of their time with work they dislike the most. This result was also indicated by the free text responses throughout the survey naming this as a major problem.

Data Cleaning. Inside the task of cleaning and organizing, multiple steps were found. The results showed that most of these steps occurred for the data scientists. The only step that was not or less needed by the majority was the text cleaning step (mean rating 1.53). When dealing with production data, this is likely explained by the focus on machine and sensor data, which usually provide time-series or image data, and less text or speech data. In addition, data value conversion and database management were rated as less time-consuming tasks, but with a significant gap to text cleaning. At the other end, filling missing values and synchronizing data streams took a major amount of the cleaning time (mean ratings of 3.59 and 3.84). The reason for this is likely the same as above when investigating the actual kinds of data. Time-series data from multiple machines, sensors, manufacturing execution systems, and other sources result in an exponential need for synchronization [24]. As such, the missing values also increase and have to be targeted. The most time spent, in fact, was the step of asking people for the meaning of data points and nomenclature (mean rating 3.89). The reason for this might be the lack of standards and proprietary naming schemes from suppliers and vendors.

Satisfaction. In total, the respondents rated their professional satisfaction quite positively. This puts the results about time spent in non-enjoyable tasks into perspective, as not too pessimistic. However, there is room for improvement, which based on the free text responses could be focused on certain requirements. The data quality issue together with the abovementioned time spent on cleaning has major potential for improvement. In addition, descriptions and standards came up multiple times, showing a need for further investigation.

5.2. Putting into Context

The related work presented in Section 2 showed a variety of task loads and their shares of the actual work time. Cleaning and organizing data was named as a time-consuming task in all surveys, ranging from time-consuming in the Anaconda studies to a clear dominant task in the CrowdFlower survey. Our survey focused on the manufacturing domain and showed that the cleaning and organizing task was dominant, which therefore is especially in line with the related studies. At the same time, the visualization task, as part of the post-processing layer, showed a minor role in comparison with the other tasks in our survey. This was similar to the findings of many other surveys, while some, like the Anaconda surveys, found it to be a significant task. For the other tasks, the results of this survey were mainly in line with the findings from other sources. This indicates that the manufacturing data science domain has similarities to the other domains, while the preprocessing layer is at the dominant end and the post-processing layer falls behind in the share of work.

Inside data cleaning, two of the most time-consuming steps, the synchronization and the filling of missing data, are mostly technical challenges. New tools and methods from the

data science domain may be applicable for solving and supporting these challenges in the future. The most time-consuming step, on the other hand, asking people for the meaning of data points, is not solely a technical challenge. To perform this step, communication with experts also has to take place, and an understanding of the data is required by the data scientists. This leads to more domain-specific challenges and solutions. Technical solutions like standards, documentation, and information stores may play a significant role in solving this issue. Nevertheless, domain-specific skills are required for this task. This shows that data cleaning, in the end, can only be performed to a certain degree without domain-specific knowledge.

Using the survey results from above, in the manufacturing domain, data scientists are mainly working in the preprocessing layer, while job positions are mainly advertised for the model-building layer [14]. According to Kaggle [22], the median salary for data scientists in Germany was between USD 70,000 and 79,999, meaning that statistically a data scientist is paid around USD 35,000 per year for preprocessing instead of model building in the manufacturing domain. This implies a significant lever to increase efficiency and reduce costs for data science in companies by tackling the challenges of the preprocessing layer. Since there are no numbers for the currently employed data scientists as a whole, or specifically in the manufacturing domain, the exact potential is hard to estimate.

6. Summary and Outlook

6.1. Summary

In this paper, we showed typical tasks for data scientists identified in the literature and surveys, and challenged this with a domain-specific survey for the manufacturing domain. As indicated by others, the topic of preprocessing tasks is dominant, in contrast to the actual model-building and post-processing tasks in the manufacturing domain. At the same time, these tasks are not enjoyed by the data scientists and are generally not advertised in job offers online. This potentially creates a false image of data scientists in practice. However, the survey did show a high job satisfaction for the data scientists.

For data cleaning, we found that data stream synchronization, the filling of missing values, and asking people for meaning were the most significant steps. These are partially technical cross-domain problems, but partially also problems requiring domain-specific solutions, not solely at a technical level. On the other hand, steps like text cleaning are mostly omitted, due to the kind of data occurring in manufacturing.

Other studies and surveys in the field of data scientists focused on a broad field and range of respondents. In contrast, this survey focused on a specific group, to identify the specific needs of this group. Comparing to the other surveys, we showed a dominance in the field of data cleaning, as well as less clear borders with other professions. These in-depth insights can allow further research and specifically targeted actions to be taken for data scientists in the production domain.

Finally, the results show a major lever for increasing efficiency and reducing costs by tackling preprocessing tasks and challenges in the manufacturing domain for data scientists.

6.2. Outlook

Considering the major lever described above, more research is needed to reduce the workload of the data cleaning and organizing task. This means new methods to ensure data quality and information availability, like a cause-effect or parameter database, should be addressed. Further, specific standards for the industry have a high potential for reducing the actual workload of preprocessing tasks.

For the main drivers of data cleaning tasks, special in-depth research is needed. The lack of understanding (asking people for meaning) about data could be addressed using standards, data mapping approaches, and better onboarding of data sources from the start. For missing data, new solutions detecting missing values and providing recommendations on how to fill these could be beneficial, especially when scaling production. The synchronization of multiple data streams increases in complexity with each data source

added. When dealing with large-scale productions, the need for efficient (semi) automated solutions is crucial for performing this task and should be addressed in further research.

The study also showed the different foci of the data scientists in practice. Additional research into the definition of data science tasks and possible additional roles like data quality officers, data engineers, or others should be performed. A dedicated toolbox for the tasks of preprocessing and methods to quickly identify typical data quality problems have high practical relevance. Metrics in terms of the costs due to lack of data quality or modeling work share could further allow the quantification of the efficiency of new solutions and tools. Finally, as this research indicated that the promotion of the position was different to the actual work share, an adaption of job descriptions and expectations could be beneficial for the happiness of data scientists.

Regarding the survey itself, the survey focused on the manufacturing domain. It would be interesting to know whether these specific requirements and tasks are found the in same way in other domains. In addition, further focusing on specific domains inside manufacturing could be interesting, assuming that domains like milling and battery cell production result in different specific tasks and steps. Due to sampling methods used, most respondents to the survey were expected to come from Germany and Europe, making the statistical analysis rather specific to the European area. Further work could focus on other areas or have a sufficiently global-scale respondent selection.

Finally, the understanding data scientists have of their job and the actual work they do could be interesting to investigate further, regarding whether expectations and reality fit together or whether problems arise from a mismatch.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ai5020043/s1>.

Author Contributions: Conceptualization, Methodology and writing, A.S.; Supervision, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: Funding reference: The project “FoFeBat—Research Fab Battery Cells” is funded by the German Federal Ministry of Education and Research. Grant number: 03XP0256.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The main study was performed using Microsoft Forms. An export of the replies after the filtering step described in Section 4 is available, including all figures shown in this paper.

Acknowledgments: The authors thank all experts taking their time to participate in the interviews, discussions, and questionnaire inputs.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IIoT	Industrial Internet of Things
AI	Artificial Intelligence
ML	Machine Learning
GDPR	General Data Protection Regulation
ETL	Extract Transform Load
CI/CD	Continuous Integration and Deployment
LLM	Large Language Model
IT	Information Technology
HCI	Human Computer Interaction
HTML	HyperText Markup Language

References

1. Tiwari, S.; Bahuguna, P.C.; Srivastava, R. Smart manufacturing and sustainability: A bibliometric analysis. *Benchmarking Int. J.* **2023**, *30*, 3281–3301. [[CrossRef](#)]
2. Yang, L.; Zou, H.; Shang, C.; Ye, X.; Rani, P. Adoption of information and digital technologies for sustainable smart manufacturing systems for industry 4.0 in small, medium, and micro enterprises (SMMEs). *Technol. Forecast. Soc. Chang.* **2023**, *188*, 122308. [[CrossRef](#)]
3. Schmetz, A.; Siegburg R.; Zontar, D.; Brecher, C. Middleware for the IIoT. In *Study of the International Center for Networked, Adaptive Production (ICNAP)*; International Center for Networked, Adaptive Production (ICNAP): Aachen, Germany, 2019.
4. Li, B.H.; Hou, B.C.; Yu, W.T.; Lu, X.B.; Yang, C.W. Applications of artificial intelligence in intelligent manufacturing: A review. *Front. Inf. Technol. Electron. Eng.* **2017**, *18*, 86–96. [[CrossRef](#)]
5. Atlam, H.F.; Walters, R.J.; Wills, G.B. Fog computing and the internet of things: A review. *Big Data Cogn. Comput.* **2018**, *2*, 10. [[CrossRef](#)]
6. Nunes, P.; Santos, J.; Rocha, E. Challenges in predictive maintenance—A review. *CIRP J. Manuf. Sci. Technol.* **2023**, *40*, 53–67. [[CrossRef](#)]
7. Wunderlich, P.; Ehteshami-Flammer, N.; Krauß, J.; Fitzner, A.; Mohring, L.; Dahmen, C. The Power of Digitalization in Battery Cell Manufacturing. Whitepaper, Accenture Industry X. 2024.
8. Escobar, C.A.; McGovern, M.E.; Morales-Menendez, R. Quality 4.0: A review of big data challenges in manufacturing. *J. Intell. Manuf.* **2021**, *32*, 2319–2334. [[CrossRef](#)]
9. Diez-Olivan, A.; Del Ser, J.; Galar, D.; Sierra, B. Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. *Inf. Fusion* **2019**, *50*, 92–111. [[CrossRef](#)]
10. Emmert-Streib, F.; Dehmer, M. Defining data science by a data-driven quantification of the community. *Mach. Learn. Knowl. Extr.* **2018**, *1*, 235–251. [[CrossRef](#)]
11. Simitsis, A.; Skiadopoulos, S.; Vassiliadis, P. The History, Present, and Future of ETL Technology. Invited Talk. *DOLAP*, 2023; pp. 3–12. Available online: https://www.cs.uoi.gr/~pvassil/publications/TALKS/2023_03_dolap_tota/23DOLAP_TestOfTimeAward_CEUR-CR.pdf (accessed on 30 April 2024).
12. Jain, A.; Patel, H.; Nagalapatti, L.; Gupta, N.; Mehta, S.; Guttula, S.; Mujumdar, S.; Afzal, S.; Mittal, R.S.; Munigala, V. Overview and importance of data quality for machine learning tasks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 23–27 August 2020; pp. 3561–3562.
13. Christoforaki, M.; Beyan, O.D. Towards an ELSA Curriculum for Data Scientists. *AI* **2024**, *5*, 504–515. [[CrossRef](#)]
14. LinkedIn: Jobs for “Data Scientist”. Available online: <https://www.linkedin.com/jobs/search/?keywords=data%20scientist> (accessed on 19 April 2024).
15. Ismail, N.A.; Abidin, W.Z. Data scientist skills. *IOSR J. Mob. Comput. Appl.* **2016**, *3*, 52–61. [[CrossRef](#)]
16. Muller, M.; Lange, I.; Wang, D.; Piorkowski, D.; Tsay, J.; Liao, Q.V.; Dugan, C.; Erickson, T. How data science workers work with data: Discovery, capture, curation, design, creation. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland, UK, 4–9 May 2019; pp. 1–15.
17. Biswas, S.; Wardat, M.; Rajan, H. The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. In Proceedings of the 44th International Conference on Software Engineering, Pittsburgh, PA, USA, 22–27 May 2022; pp. 2091–2103.
18. CrowdFlower. *Data Science Report*; Whitepaper CrowdFlower (now: Appen Limited); CrowdFlower: San Francisco, CA, USA, 2016.
19. State of AI 2022. In *Whitepaper Appen*; Appen: Bellevue, WA, USA, 2022.
20. Anaconda Inc. *2023 State of Data Science*; Report Anaconda Inc.; Anaconda, Inc.: Austin, TX, USA, 2023.
21. Anaconda Inc. *2022 State of Data Science*; Report Anaconda Inc.; Anaconda, Inc.: Austin, TX, USA, 2022.
22. Kaggle. *State of Data Science and Machine Learning 2020*; Kaggle: San Francisco, CA, USA, 2020.
23. Frye, M.; Mohren, J.; Schmitt, R.H. Benchmarking of data preprocessing methods for machine learning-applications in production. *Procedia CIRP* **2021**, *104*, 50–55. [[CrossRef](#)]
24. Schmetz, A.; Lee, T.H.; Zontar, D.; Brecher, C. The time synchronization problem in data-intensive manufacturing. *Procedia CIRP* **2022**, *107*, 827–832. [[CrossRef](#)]
25. European Parliament. *Regulation on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC (Data Protection Directive)*; European Parliament: Bruxelles, Belgium, 2016.
26. Research Website SurveyCircle. Available online: <https://www.surveycircle.com> (accessed on 19 April 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.