SAND2011-1384C

A Study on the Importance of and Time Spent on Different Modeling Steps

M. Arthur Munson
Sandia National Laboratories
Livermore, CA 94551, USA
mamunso@sandia.gov

ABSTRACT

Applying data mining and machine learning algorithms requires many steps to prepare data and to make use of modeling results. This study investigates two questions: (1) how time consuming are the pre- and post-processing steps? (2) how much research energy is spent on these steps? To answer these questions I (a) surveyed practitioners about their experiences in applying modeling techniques and (b) categorized data mining and machine learning research papers from 2009 according to the modeling step(s) they addressed. Survey results show that model building consumes only 14% of the time spent on a typical project; the remaining time is spent on pre- and post-processing steps. Both survey responses and the categorization of research papers show that data mining and machine learning researchers spend the majority of their energy on algorithms for constructing models and significantly less energy on other steps. These findings collectively suggest that there are research opportunities to simplify the steps that precede and follow model building.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*; I.6.m [**Pattern Recognition**]: Miscellaneous

Keywords

data mining process, machine learning process, practitioner experiences

1. INTRODUCTION

Many steps are involved in applying machine learning and data mining to real problems [1, 2, 3, 6, 12]. The heart of this iterative process is the actual machine learning and/or data mining step, during which the practitioner poses the problem, selects or designs an algorithm, tunes hyperparameters, etc. The output of this step is typically a model that can be used to make predictions about future data (e.g., a decision tree) or that helps summarize and visualize the input data (e.g., a dendrogram produced by hierarchical clustering). Before model learning, however, the data itself must be collected and prepared. Similarly, much work can be required after the model is learned to understand, evaluate, and make use of the modeling results.

Conventional wisdom, accumulated from individual experiences, holds that these steps are both time consuming and crucial to successful applications, but little published data exists to support or disprove this belief. Existing studies of the end-to-end modeling process state that model building comprises only a small portion of project time [1, 3], and occasionally estimate the size of this step (e.g., "15 to 25% of the overall effort" [6, p. 90]), but do not provide quantitative data or citations to support the observations. The two exceptions are the 2nd Annual Data Miner Survey (which found that 20% of project time was spent generating models) [8] and a KDnuggets poll (which found that the majority of data miners spend 60% or more of their time on data cleaning and preparation) [4].

In this paper I investigate how time consuming the various modeling steps are for practitioners and how much research effort is focused on each step. To answer these questions I surveyed practitioners about their experiences in applying data mining (machine learning) techniques and manually categorized the 2009 proceedings from two of the top conferences in the area (ICML 2009[1] and KDD 2009[2]) based on the step(s) they addressed.

There are two main findings in this study. First, in the typical project only 14% of the time is spent building the model. The rest of the time is spent preparing to do model learning and verifying the results after model construction. In contrast, the data mining and machine learning research communities spend the majority of their energy on how to learn a model from data, and moderate energy or less on other modeling steps. This finding is supported both by survey responses and by the distribution of papers at ICML 2009 and KDD 2009. In addition to documenting the current state of practice and research, these findings suggest that there are research opportunities to simplify the steps before and after model building. Such improvements would greatly benefit practitioners and should facilitate further adoption of data mining and machine learning techniques.

Section 2 describes the study methodology; results follow in Section 3. I close with a discussion of the results in the context of previous studies, potential study limitations, and my thoughts on the difference between the focus of researchers and the activities of practitioners (Section 4).

2. METHODOLOGY

2.1 Survey Distribution and Collection

Survey participants were solicited through word of mouth, the machine learning question and answer forum at www.metaoptimize.com, and three email lists:

[1] The International Conference on Machine Learning.

[2] The ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

1. ml-news@googlegroups.com (a news forum for the machine learning community),
2. KDnuggets (an online news letter for the data mining community), and
3. corpora@uib.no (a news forum for the natural language processing community).

Appendix B contains the text of the survey announcement.

Responses were collected over two time spans. I first ran the survey over 1.5 weeks in February 2010 and advertised the survey on ml-news@googlegroups.com and KDnuggets. Twenty-four respondents completed the survey during this period. I discarded two of these that contained dummy values for all questions (e.g., zero percent time spent during all modeling stages), leaving 22 survey responses.[3] The initial results from the first run were posted on my web page in early March 2010 and were published in my dissertation [5]. In mid-March I re-opened the survey but only advertised it with a note on the web page of initial results. Two completed surveys were collected over March and April. The second major data collection spanned September 2010 to 10 January 2011. During this second round, I announced the survey on `www.metaoptimize.com`, ml-news@googlegroups.com, and corpora@uib.no. Thirty-three people completed surveys during the second round, resulting in 57 total responses.

The survey questions are reproduced in Figure 1.

2.2 Data Normalization

Time percentages were normalized to sum to 100 for each survey response to correct nine responses that did not sum to 100. This preprocessing put all the responses on the same scale and facilitated comparing the relative energy spent in each step. Times for eight responses originally summed to values between 90 and 110, and normalization produced minor adjustments in percentage values. The times in the ninth response originally summed to 27%, and the survey participant commented that the survey did not include the steps where the remaining time was spent (e.g., project planning, finding a good learning algorithm, publishing results). Normalization converted this answer to how much relative time was spent among the steps included in the survey.

2.3 Categorizing Papers

In parallel with collecting the initial survey results, I manually categorized the papers published at ICML 2009 and KDD 2009. This categorization was completed before analyzing the survey results. Each paper was labeled as addressing one of eight categories: the six modeling stages from the survey (see Figure 1) plus two extra categories. The *Domain Knowledge* category covered papers reporting new domain knowledge or ways to take advantage of domain knowledge. The final category, *Other*, captured papers that did not fit easily elsewhere.

Conclusions from this categorization are limited by the fact that it is based on one person's subjective judgment. Only big picture trends can be considered reliable, and the exact percentages of papers in each category should be viewed with skepticism. A future study with multiple annotators could repeat this categorization if the exact proportions per category is sufficiently interesting to the community.

Page 1: Your Background.

1. How many *completed* systems have you worked on where data mining or machine learning were important to success?
   A completed system is either deployed or results in significant contributions to a domain outside of computer science (e.g., a publication in non-CS journal).[a]
2. (Optional) Please list key words or phrases that describe your interests, expertise, and/or background. One phrase or key word per line.

P2: Difficulty and Importance of Modeling Steps.

This page asks questions about your experience with the following modeling steps:

- Data Collection (not raw data collection, but any work team did to gather data into hands of analysts)
- Data Preparation (e.g., data integration and fusion, data cleaning, handling missing values)
- Change Data Representation (e.g., rescaling and normalizing features, transforming prediction target, feature selection, dimensionality reduction)
- Learning a Model from Data (e.g., posing the problem, algorithm selection or design, hyper-parameter selection)
- Performance Evaluation (accuracy and confidence in predictions)
- Study Model (e.g., to understand the model, to discover knowledge about domain theory, or to identify regions where model makes risky extrapolations)

1. Choose one system you have worked on with a modeling component (most recent, biggest, most successful, etc.). Estimate the percentage of time spent in each modeling step. Please include both your effort and your collaborators' efforts, but omit computer time. Rough estimates are sufficient.
2. How important was each step to the success of the system in the previous question? [Respondent chose one of following for each step: not important, slightly important, moderately important, important, or critically important.]

P3: Focus of Research Community.

In your opinion, how much energy does the *research community* spend addressing each modeling step? [Respondent chose one of following for each step: negligible energy, a little energy, moderate energy, lots of energy, or enormous energy.]

P4: Thank you!.

Thank your for completing the survey. If you have any extra comments or feedback you may leave them in the box below. [*Added in March 2010 re-posting:* Previous respondents left very interesting comments that were not published because I did not ask for their permission to quote comments. If your comments can be quoted and published, please write 'OKAY TO QUOTE' in the box. Quotes will be anonymous unless you sign your name in the comments.]

[a] In mid-September this text was amended to read: "...or makes a significant domain contribution outside of computer science (e.g., lives saved, corporate policy changes, increased profit margin, a publication in non-CS journal, etc.)." This clarification was a response to comment 1 in Appendix A. Thirty-three surveys were collected with the original wording.

Figure 1: Survey questions. Survey takers could not return to pages they had already completed.

[3] In the comments of one discarded survey, a respondent stated that he/she simply wanted to view all of the survey questions.

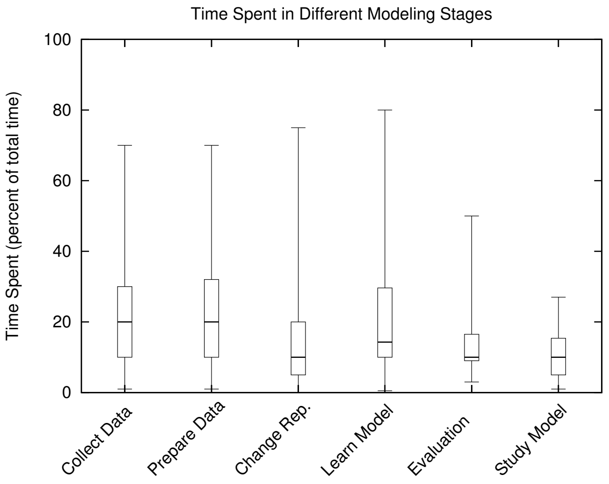

Figure 2: Allocation of time spent building systems with machine learning or data mining components. Time estimates were collected from practitioners with experience deploying systems. Boxes show the 25th and 75th quantiles of time spent per stage; the line within each box marks the median time spent. Whiskers show the minimum and maximum time spent.

3. RESULTS

Fifty-seven completed surveys were collected. Respondents varied greatly in their experience (Table 1). Areas of expertise included medicine, robot control, natural language processing, customer modeling and retention, advertising, finance, computer vision, bioinformatics, semantic audio processing, and child language acquisition.

Table 1: Number of systems completed by survey respondents.

# Systems	Frequency
0	3
1–3	34
4–10	16
11+	4

The relative time spent in each stage also varied greatly by project (Figure 2). Data collection and preparation were the most time consuming stages, based on median values (both 20% of project time). In the typical project, only 14% of the effort was actually spent learning the model. In comparison, practitioners spent 10% of project time on each of the other steps (median values). *In other words, the stages that precede and follow model building are* individually *roughly as time consuming as learning the model.* However, projects with such an even distribution of effort across all steps were relatively rare. Out of 57 responses, only four reported time allocations in which the longest and shortest step were separated by 15 percentage points or less. Instead, individual projects usually required larger time commitments on one or two stages and smaller time commitments for a single step (usually 5% or less of project time).

Respondents generally rated most modeling steps as important to building successful systems; in contrast, they felt that the research community focuses the bulk of its energy on learning algorithms (Figure 3 vs. Figure 4). As with answers about how much time was spent per step, respondents individually attributed varying importances to the different steps. Twenty-two respondents rated at least one step *not important* or *slightly important*. Conversely, 51/57 participants rated four or more steps *important* (or *critically important*), and 30/57 participants rated five or more steps important. Unlike the relative eveness of time spent per step (Figure 2) and step importance (Figure 3), the energy spent by researchers shows a strong concentration on the *Learn Model* step (Figure 4).

The distribution of conference papers at ICML 2009 and KDD 2009 reinforces the results from Figure 4. Figure 5 shows that researchers spend more energy, by a wide margin, on learning algorithms than on other steps. Non-trivial energy is being spent on other steps, however. The majority of respondents felt that all stages except data collection received at least moderate research attention (Figure 4). Similarly, a significant fraction of ICML papers studied how to change data representations, and 5% or more of KDD papers addressed each of the modeling steps. Of note, many KDD papers were case studies of applying machine learning and data mining to solve real problems (included in *Other*). Arguably, some of these papers might also be counted towards other steps since they likely contain lessons that could be generalized to other applications.

Of the twenty-two comments, eleven elaborated on why certain modeling step(s) were the most important for success (often in particular domains), and eight pointed out survey shortcomings. Specifically:

- Question wording implied that modeling is a waterfall process with each stage executed once in a serial pipeline. In reality modeling is an iterative process, and estimating time spent in discrete steps is not straightforward.
- The meaning of the Change Representation step was unclear.
- Steps were missing. Specifically, a) convincing business users of a model's utility; b) integrating modeling results into a larger system; c) project planning; d) publishing results; and e) turning a research prototype into industrial-strength software.
- The survey over emphasized batch-oriented modeling (vs. online learning).

A handful of participants agreed to be quoted; their comments are listed in Appendix A.

4. DISCUSSION

4.1 Results in Context

This study quantifies the time practitioners spend on different modeling steps, the importance of those steps to application success, and how much energy the research community spends per modeling step. The results show that only a small percentage, 14%, of the modeling process is spent building models from data. This is consistent with conventional wisdom and with results from Rexer Analytics (Table 2). The two studies subdivide the process into different stages, but there are other obvious similarities: (a) collecting and preparing data are the most time consuming activities (36% of time vs. 20%+20% median time in Figure 2); (b) evaluation requires about 10% of time. One can

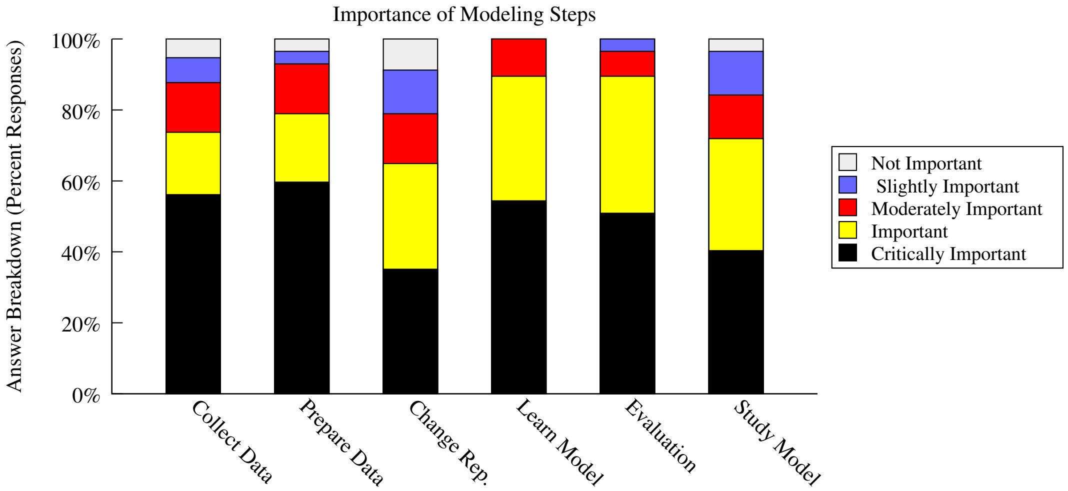

Figure 3: The majority of surveyed practitioners rated all steps as important or critically important to their systems' successes. Chart shows the breakdown of importance ratings for each modeling step. For example, 55% of respondents rated data collection *critically important*, 17% rated it *important*, 14% rated it *moderately important*, 7% rated it *slightly important*, and 5% rated it *not important*. (Percentages do not add to 100% due to rounding.)

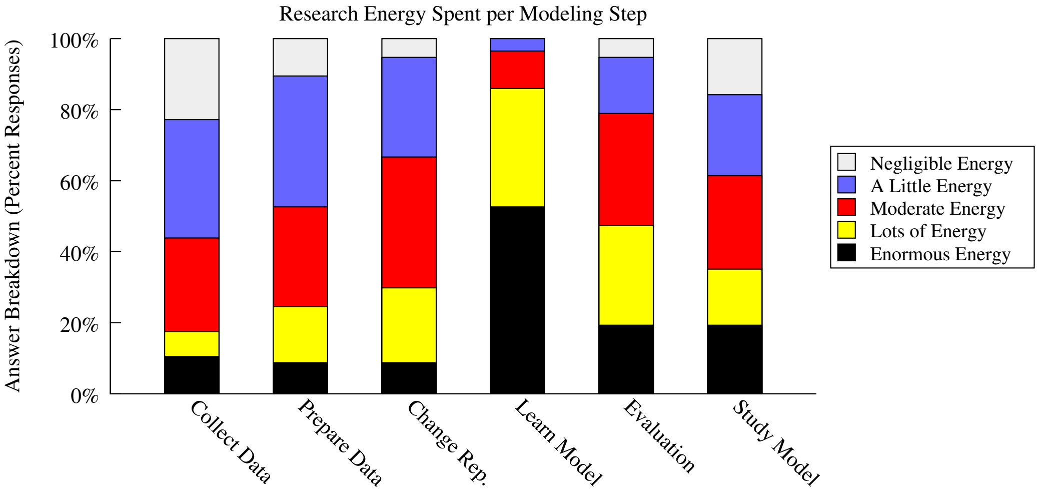

Figure 4: Surveyed practitioners felt that the machine learning and data mining research communities spend the most energy on how to learn a model from data. Most respondents rated the communities as spending *moderate energy* or less on the modeling steps preceding and following learning a model. In contrast, 84% of respondents felt the community spent *lots of energy* or *enormous energy* on how to learn a model.

also compare this study's results to a poll from KDnuggets which found that most data miners spend at least 40% of their time, and often more than 60%, on data cleaning and preparation (Table 3). The times reported here are lower (40% median time in total for the two steps), but this is partially due to having a distinct step for changing data representations (commonly considered part of data preparation).

More surprising is how consistently (and highly) respondents rated the importance of *all* the modeling steps. Access to data and dirty data are consistently reported as the biggest challenges to data miners [7, p. xvii] [8, 9, 10, 11], and data collection and preparation are the most time consuming steps in Figure 2. It is therefore surprising that these two steps are not considered more important than the other steps.

Table 2: Time spent on various data mining tasks. Sample size was 265 data miners. Source: Rexer Analytics [8]. Reproduced with permission.

Task	% Time
Understanding Business Problem	20%
Accessing & Preparing Data	36%
Generating Models	20%
Writing Reports / Presentations	15%
Scoring / Deploying	9%

While the majority of respondents felt that all modeling steps are important to success, the research community is strongly focused on the model building step. This can be seen in the topics of recent conference papers and in the an-

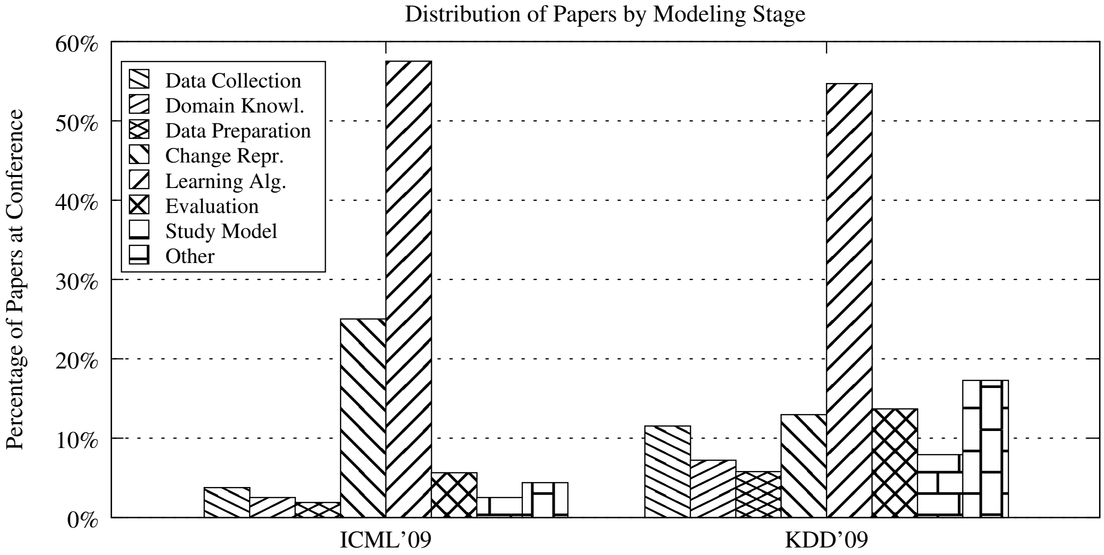

Figure 5: Distribution of papers published at ICML 2009 and KDD 2009. Papers were manually categorized according to which modeling step(s) they addressed (see Section 2.3 for details). Percentages do not add to 100 because some papers were counted in multiple categories.

Table 3: Percent project time spent on data cleaning and preparation. Source: October 2003 KDnuggets poll [4]. Reproduced with permission.

% Time	# Votes	% Responses
0–20%	15	8%
21–40%	7	4%
41–60%	46	25%
61–80%	73	39%
80–100%	46	25%

swers of survey respondents. This supports previous qualitative observations:

> Most previous work on [knowledge discovery in databases] has focused on . . . the data mining. However, the other steps are as important (and probably more so) for the successful application of [knowledge discovery in databases] in practice. [3, p. 42]

> [I]t is fair to say that very little consideration is given in the research literature to the overall process of developing classification applications and, in particular, to problem-specific factors such as domain, data and human factors. . . . [I]t is unfortunate that this is the case since in practical applications it is often the data and human issues which ultimately dictate success or failure of a project rather than algorithmic and model issues. [1, p. 54]

The results in Figures 4 and 5 hint that more research is being done today on other modeling steps than is described in the above quotes from the mid-1990's. One plausible hypothesis is that interest in the practical issues around data mining and machine learning has been growing as learning algorithms have been applied to new tasks. An interesting study would be to track the number of research papers addressing different modeling steps from the 1990's to the present day.

4.2 Study Limitations

It is important to note the potential limitations of the results in Section 3. First and most importantly, the survey is prone to self-selection bias because respondents decided to participate or not based on their personal motivations. As a result, there is no way to know how representative the respondents are. One mitigating factor is the moderate sample size (58 respondents), which reduces the risk of drawing a completely biased sample. Of course, a larger sample would be even better.

Second, practitioners may have a skewed perspective of how the research community spends its energy. Perhaps lots of research energy *is* actually spent on all steps of the modeling process. This is unlikely for three reasons. First, some of the survey participants are known to be researchers as well as practitioners; they presumably are aware of the research community's general activities. Second, members of the research community have, in the past, pointed out researchers' predominant focus on modeling algorithms (see quotes in previous section). Third, the distribution of papers at ICML 2009 and KDD 2009 shows a similar picture.

Third, the distribution of conference papers is based on one person's subjective judgment and cursory reviews of 299 papers. The large patterns agree with the survey results and are probably accurate. The exact percentages of papers per category should be considered unreliable, however, until multiple annotators conduct their own categorization. Note also that the distribution of papers at other data mining and machine learning conferences—as well as more domain specific conferences that feature modeling applications—may be different.

4.3 Difference between Research and Practice

Despite the above study limitations, the differences between where researchers focus their energy and where practitioners spend their energy are so striking that it is hard to completely dismiss the results. This difference in focus is natural given the different goals of each group: researchers aim to discover general algorithms that are applicable to a wide range of modeling tasks, while practitioners strive for concrete domain results. The former requires abstracting away non-essential domain specifics, while the latter requires applying general algorithms to a specific domain through a combination of adaptation, data engineering, and system engineering.

This difference also represents a research opportunity. It seems fair to say that the practical steps preceding and following model building are the limiting factor for data driven analysis and applications. To maximize the impact and adoption of our data mining and machine learning algorithms, we should strive to simplify the other steps as much as possible. While many application obstacles are task specific (making general purpose solutions unrealistic), there remain modeling issues that span applications (e.g., handling missing values, detecting data outliers, estimating prediction reliability). Innovations that remove or mitigate these issues have the potential to change how and where learning algorithms are applied.

Acknowledgments

The participation of dozens of data mining and machine learning practitioners and researchers made this study possible: thank you to everyone who completed the survey. I am grateful to Rich Caruana for suggesting the idea of conducting a survey and for helpful comments throughout the work. Thank you to Yisong Yue, Daria Sorokina, Philip Kegelmeyer, and Alex Niculescu-Mizil for thoughtful comments on an early draft. Finally, thank you to Karl Rexer at Rexer Analytics for sharing the results of their Annual Data Miner Surveys.

This work was started at Cornell University and completed at Sandia National Laboratories. It was funded by the National Science Foundation (grants DUE-0734857 and IIS-0612031) and by an Early Career Award from the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

5. REFERENCES

[1] C. E. Brodley and P. Smyth. Applying classification algorithms in practice. *Statistics and Computing*, 7(1):45–56, 1997.

[2] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. *CRISP-DM 1.0: Step-by-step Data Mining Guide*. CRISP-DM Consortium, 2000.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[4] KDnuggets. What percent of time in your data mining project(s) is spent on data cleaning and preparation?, 2003. `http://www.kdnuggets.com/polls/2003/data_preparation.htm`.

[5] M. A. Munson. *Outside the Machine Learning Blackbox: Supporting Analysts Before and After the Learning Algorithm.* Ph.D. thesis, Cornell University, 2010.

[6] G. Piatetsky-Shapiro, R. Brachman, T. Khabaza, W. Kloesgen, and E. Simoudis. An overview of issues in developing industrial data mining and knowledge discovery applications. In *KDD'96*, pp. 89–95, 1996.

[7] D. Pyle. *Data Preparation for Data Mining.* Morgan Kaufmann, 1999.

[8] K. Rexer. 2nd annual data miner survey: Summary report. Presented at Predictive Analytics World (PAW'08), 2008. Email Rexer Analytics to obtain a copy.

[9] K. Rexer. 3rd annual data miner survey — 2009 survey summary report. Presented at Predictive Analytics World (PAW'09), 2010. Email Rexer Analytics to obtain a copy.

[10] K. Rexer. 4th annual data miner survey — 2010 survey summary report. Presented at Predictive Analytics World (PAW'10), 2010. Email Rexer Analytics to obtain a copy.

[11] K. Rexer, P. Gearan, and H. N. Allen. Surveying the field: Current data mining applications, analytic tools, and practical challenges, 2007. Data Miner Survey Summary Report. Email Rexer Analytics to obtain a copy.

[12] C. Shearer. The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22, 2000.

APPENDIX

A. QUOTED SURVEY COMMENTS

This section lists the quotable survey comments with minor editing (marked with square brackets).

1. *Anonymous:* "I was a little confused / concerned that 'success' in the application of data mining / machine learning was suggested to be related to publications rather than such things as: lives saved, corporate policy changes, increased profit margin, etc. I would be interested in hearing about those situations where publications were significantly related to the success of a data mining effort."
2. *Anonymous:* "I have found that the most important steps in building a real world system are the mundane ones: truly understanding the data and how the model will react to it, fixing the data if necessary, and avoiding common pitfalls like target leaks. Machine Learning advancements (usually from academia) are great, but in the end they can only take you so far."
3. *Yann LeCun:* "Much of the time and energy (and the largest number of people) was devoted to turning the research prototype into an industrial-strength piece of software, which is not an item in your survey. Most people spend a large amount of time massaging the data and finding a good representation, or figuring out a good post-processing scheme. Since we used "end-to-end" learning, including feature learning methods for the front-end, and structured-prediction for the post-processing, we didn't have to spend any effort on that. The application was a bank check reader, which was

deployed commercially in 1996. At some point in the late 90's, the system collectively read 10 to 20% of all the checks in the US."

4. *Anonymous:* "... when you're not working on the benchmark data sets[,] data collection can be a very frustrating process. On my project it has taken a lot of bugging people in the field to send their data from a logging system. They send over an SQL database. A lot of time can be spent setting up the tables for analysis, exporting to csv, creat[ing] matlab data structures. Real project[s] really need someone [whose] full time job is just the data collection and preprocessing steps."

B. SURVEY ANNOUNCEMENT TEXT

The text of the survey announcement is included here for completeness.

> There are many steps required to build and deploy a system with a significant machine learning or data mining component. I am interested in how much time is spent per step in developing real systems and the community's opinion of the importance of the various steps.
>
> To study this question I am conducting a short survey (5–10 minutes) of the community's experience with developing real systems. The survey can be found at:
>
> `http://www.surveymonkey.com/s/39YCRX`
>
> Please note that I posted this same survey in the spring of 2010, and the results from the roughly 20 responses were very interesting. This time I hope to get enough responses to be confident in the results and communicate them to the community. There is no need to complete the survey a second time if you responded in the spring.
>
> Thank you in advance to anyone who can spare a few minutes to take the survey.
>
> Sincerely, Art Munson