Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions

Anastasia Zhdanovskaya, Daria Baidakova, Dmitry Ustalov

Toloka
{ende-neu,dbaidakova,dustalov}@toloka.ai

Abstract

The process of training and evaluating machine learning (ML) models relies on high-quality and timely annotated datasets. While a significant portion of academic and industrial research is focused on creating new ML methods, these communities rely on open datasets and benchmarks. However, practitioners often face issues with unlabeled and unavailable data specific to their domain. We believe that building scalable and sustainable processes for collecting data of high quality for ML is a complex skill that needs focused development. To fill the need for this competency, we created a semester course on Data Collection and Labeling for Machine Learning, integrated into a bachelor program that trains data analysts and ML engineers. The course design and delivery illustrate how to overcome the challenge of putting university students with a theoretical background in mathematics, computer science, and physics through a program that is substantially different from their educational habits. Our goal was to motivate students to focus on practicing and mastering a skill that was considered unnecessary to their work. We created a system of inverse ML competitions that showed the students how high-quality and relevant data affect their work with ML models, and their mindset changed completely in the end. Project-based learning with increasing complexity of conditions at each stage helped to raise the satisfaction index of students accustomed to difficult challenges. During the course, our invited industry practitioners drew on their first-hand experience with data, which helped us avoid overtheorizing and made the course highly applicable to the students' future career paths.

Introduction

Despite a common perception that an average data scientist or machine learning (ML) engineer only designs, evaluates, and iterates machine-learning models, they also spend a disproportionate amount of their time on processing, labeling, and augmenting training data (Sculley et al. 2015). Often they have to build and maintain training data infrastructure themselves. Handling training data efficiently is rapidly becoming an integral part of any ML specialist's work, hence the skills needed to build scalable and sustainable processes for collecting and managing training data are becoming a key competency for ML engineers. This often involves crowdsourcing projects, as human-labeled data is a significant bottleneck for artificial intelligence (AI) practitioners (Sambasivan et al. 2021). In practice, in order to create a training dataset, an ML engineer, a data scientist or another researcher sets up a project on a crowdsourcing platform such as Amazon Mechanical Turk[1] or clickworker[2] and thousands of people complete their task online. To get high-quality data one has to know certain techniques that will let them control the quality in real time and get the best possible results out of different verdicts or labels.

In this paper, we share our experience and insights on introducing a data labeling course centered around various crowdsourcing techniques into a university-level curriculum. With our semester course "Data Collection and Labeling for Machine Learning" we sought to fill the need for this competence among future ML specialists, positioning it as an engineering skill, in direct opposition to the common perception of crowdsourcing as mere worker management. After running two iterations of our course, we decided to enhance it with *project-based learning elements*, which made it more relevant to students' future careers according to the feedback we gathered.

The remainder of the paper is organized as follows. We give the definition of crowdsourcing and an overview of the well-known university-level and vocational courses on crowdsourcing. We discuss the motivation behind creating our course on data labeling with crowdsourcing. Then, we describe the curriculum and course structure and show why and how we added elements of project-based learning and data-centric competitions to our course. After that, we show the results of student surveys and interviews, which demonstrate the greater impact of the course on students' future careers. Next, we analyze the limitations of our course. Finally, we outline the reasons for future work and conclude with final remarks.

Related Work

Estellés-Arolas and González-Ladrón-de Guevara (2012) define crowdsourcing as a type of participative *online activity* in which *the requester* proposes to *a group of individuals* the voluntary undertaking of *a task*. Due to the scalability of

[1] https://www.mturk.com/

[2] https://www.clickworker.com/

Course	Institution	URL
Crowdsourcing & Human Computation	UPenn	https://crowdsourcing-class.org/
Social Computing	Stanford	https://cs278.stanford.edu/
Crowdsourcing Linguistic Datasets	ESSLLI	https://esslli2016.unibz.it/?page_id=346
Crowdsourcing	KAIST	https://www.kixlab.org/courses/cs492-fall-2016/index.html
Crowdsourcing for Computer Vision	UC	https://home.cs.colorado.edu/~DrG/Courses/ CrowdsourcingForCV/AboutCourse.html
Crowdsourcing and Human Computation	Cornell	https://www.cs.cornell.edu/courses/cs5306/

Table 1: A survey of related courses in various institutions across the world in the US (Stanford, UPenn, UC, Cornell), Europe (ESSLLI), and Asia (KAIST).

simple online tasks, crowdsourcing proved itself as a robust methodology for data labeling in machine learning, data science, and other related subfields of AI. Although the core concepts of crowdsourcing are easy to understand for non-experts, achieving consistently high quality in the annotated data often requires sophisticated quality control techniques (Daniel et al. 2018).

Crowdsourcing has been taught at many institutions across the world, including Stanford, KAIST, and others, see Table 1. We noticed that most crowdsourcing classes (UPenn, KAIST, UC, Cornell) include a broad overview of important theoretical results in microtask-based crowdsourcing, such as task design, answer aggregation, and quality control, and then offer a project assignment involving the combination of most of these techniques. Notable exceptions are the classes focused on the specific application domain like computer vision (UC) or language resources (ESSLLI). In these cases, the course curriculum pays additional attention to the practical results in these domains. Some courses also provide a broader context beyond microtasks, discussing social media and Wiki technology (e.g., the course at Stanford).

While designing our course, we wanted the students to master each of the key components of efficient data labeling with crowdsourcing using *focused hands-on exercises based on real-world projects*. So instead of having only one project assignment in our course, we gave a corresponding homework assignment for every topic in the basic course syllabus. Later, we found that the combination of thematic homework assignments and project-based learning increases the usefulness of the course to the future careers of the students.

Context and Course Development

In this section, we will talk about the circumstances of developing our course that contributed to its specifics. A key factor that distinguished our course curriculum from those mentioned above was that the idea sprouted inside the team of the global data labeling platform Toloka[3], which operates as a tool and an expertise center for industrial data labeling. In other words, the course was created by a group of practitioners. Their knowledge of crowdsourcing techniques was scattered internally across various departments, oftentimes held in silos, and was not shared anywhere publicly for years until the spring of 2019. In 2019 we observed a growing demand for specialists who could efficiently build data labeling pipelines for AI on the market and we decided to collect and systematize our knowledge and make it publicly accessible (Chui, Manyika, and Miremadi 2018; Cognilytica 2019).

To design the structure and content of the course, we involved subject matter experts from various domains including self-driving cars, voice assistants, search engines, scientific research in crowdsourcing, etc. who agreed to teach others and share their first-hand experience in creating and running crowdsourcing projects for the algorithms. These subject matter experts created homework assignments based on real tasks they faced during their work. For example, students were asked to collect thousands of hours of speech recordings made by different voices and accents with background noise to train a voice assistant. Another example is to get pairs of matching search queries and web pages judged by the relevance of the match in order to train a recommender system. To support practical case studies, a certain crowdsourcing methodology was introduced for each case (more information on basic course syllabus is available in the corresponding section). Thus, the course content became a combination of theoretical discussions and practical assignments involving data labeling in realistic scenarios.

There was not one particular lecturer who could grasp and teach all of the aspects of crowdsourcing altogether yet. Instead, we chose to give the floor to several industry practitioners who contributed to the course content and held lectures and seminars dedicated to each independent topic. We also invited our research team to lecture on the theory of crowdsourcing. As a result, we had more than ten lecturers ready to teach separate parts of the course. **In fact, we crowdsourced an entire course on crowdsourcing.**

Data scientists and machine learning specialists were the target audience of the course. Hence, we had to find an educational institution eager to collaborate with us and then integrate our course into their curriculum. From 2019 to 2022 we partnered with the Higher School of Economics, a top-tier university that boasts leading data analysis programs in Russia. Its graduates work in the world's largest technology companies such as Google, Apple, and Meta. They also become researchers at top universities worldwide. Taking all of the above into account, "Data Collection and Labeling for Machine Learning" complemented its curriculum for bachelor students from the Computer Science department. Since

[3] https://toloka.ai/

machine learning courses were already packed with content, there was no opportunity to integrate our curriculum into an existing course.

Basic Course Syllabus (2019 and 2020)

In this section, we will describe the structure and syllabus of the initial version of "Data Collection and Labeling for Machine Learning." Each week we covered a certain topic that consisted of a lecture followed by a practicum. Each lecture offered either a deep dive into an aspect of the crowdsourcing methodology or a case study from the work experience of our experts. Most seminars involved practical data labeling tasks when students collected a certain type of dataset (image classification, content moderation, text generation, offline data collection, and others). A detailed syllabus is presented in Table 2.

Based on the fundamental results published in the literature (Bernstein et al. 2010; Dawid and Skene 1979; Bradley and Terry 1952; Zhang, Li, and Feng 2016; Zheng et al. 2017; Daniel et al. 2018, etc.) and our in-house industrial experience (Drutsa et al. 2021), we formulated the specific key components behind building any crowdsourcing pipeline that we saw repeated from task to task:

Task Decomposition: the simpler the task is, the better the annotators will master it, so *the students have to invent a clever way to split their complex task into a pipeline of simpler crowdsourcing tasks*

Worker Interface: a clear task interface will navigate the annotators through the task and prevent them from making mechanical mistakes; *the students have to design a convenient problem-specific task interface using Web technologies*

Task Instructions: explaining all edge cases and grey areas helps prevent noisy and inconsistent labels; *the students have to carefully explain what they want crowd workers to do*

Quality Control: quick monitoring of the crowd's performance and instant feedback are crucial to maintaining stable quality in the final data; *the students have to build a meaningful combination of quality control techniques*

Pricing: a reasonable pricing scheme is a tool for motivating annotators as well as for getting the most value out of a limited budget; *the students have to stay within the fixed budged allocated for each assignment*

Answer Aggregation: since most data labeling projects imply getting multiple responses to each task, choosing the best way to aggregate them into a final label also contributes to final quality of the dataset; *the students need to choose and implement the optimal method to obtain reliable responses from the noisy crowdsourced data*

The final artifact of each homework assignment was a labeled dataset, which was then compared to a ground truth set of labels. Based on this comparison, a grade for each task was calculated. There was no final examination, but the final grade was made up of homework grades that could total a maximum of 100 points. **To sum up, we have introduced one unified methodology for efficient data collection and labeling and applied it to versatile data labeling projects.**

Motivation for Further Development

After teaching this course twice, we collected substantial feedback which stated that although the course seemed rich content-wise and offered diverse applications, overall the course was monotonous and repetitive. The students noted that having to apply similar methods in a new environment over and over again did not leave a sense of progress or cohesive understanding of the subject. Indeed, the tasks were very similar although they were applied in different domains. Also, the students noted that the course seemed easy compared to other courses in their curriculum, and many of them chose it because of its simplicity, spreading the word about data labeling as something that doesn't require much effort. Another concern we had was that the respondents hesitated to answer questions about practical applications of the course, stating that "it may be useful, but I don't know where yet" or "this is useful knowledge about people even though I don't deal with crowdsourcing."

This kind of feedback presented several challenges for us. How do we make the course more engaging? How do we prove that data labeling is a core competence for an ML/AI professional, not just an elective class to attend for the sake of getting a grade? How do we show that this competence can be approached as an engineering task, not as something that requires solely people management? To approach these challenges, we started working on a new version of the course curriculum based on a demonstration of how data works throughout the entire ML cycle. In the next section, we will discuss the key focuses of a new syllabus that were introduced to make the course more engaging.

Project-Based Syllabus (2021)

After diving into the ML model life cycle with a team of experts, we came up with three basic data-related theses that were then transformed into expected learning outcomes:

- **The quality of training data directly influences the quality of the model trained on it.** *Corresponding outcome:* a student can collect or label a dataset that is necessary for training their model so that the model will achieve a certain labeling quality (depending on the underlying ML model).
- **In real life we normally deal with budget restrictions, hence a more conscious approach to data labeling pipelines should be chosen instead of simply overflowing the model with data.** *Corresponding outcome:* a student can collect or label a dataset that is necessary for training their model, meeting quality and budget requirements.
- **Having trained a model that achieves a certain quality, a data engineer does not stop working on it.** Data flow can change over time and a model needs to respond to this change. *Corresponding outcome:* when working with a trained model, a student can track changes in its quality and retrain it, meeting quality and budget requirements.

Week	Lecture	Seminar & Homework
1	Introduction to Crowdsourcing and Task Decomposition	Running a Sample Project on a Platform
2	Task Instruction and Worker Interface Design	Programming a Task Interface
3	Quality Control and Worker Selection	Comparing Techniques for Quality Control
4	Object Classification Tasks	Basic Project: Image Classification
5	Answer Aggregation	Implementing a Probabilistic Aggregation Model
6	Dynamic Overlap and Pricing	Data Transfer between Tasks using the API
7	Computer Vision Tasks	Image Segmentation for Self-Driving Cars
8	Content Creation	Crowdsourcing Actor Biographies
9	Pairwise Comparisons	Ground Truth for Search Relevance Evaluation
10	Spatial Crowdsourcing	Check Street Market Locations
11	Voice Annotation	Audio Transcription and Aggregation
12	User Experience (UX) Testing	UX Testing for Search Engine Results Page

Table 2: Syllabus of our course with 12 lectures and 12 seminars; each seminar results in a hands-on homework assignment.

In a way, this is a single yet complicated learning outcome that is decomposed into a succession of simpler ones, each introducing a new condition. This succession of conditions became the basis for creating an experiential curriculum divided into three project rounds.

Project Work

According to the three outcomes formulated above, "Data Collection and Labeling for Machine Learning" was now divided into three rounds of project work, each lasting approximately 4 weeks.

Although the effectiveness of project-based learning has been widely argued (Boaler 1998; Mioduser and Betzer 2008; Hassan et al. 2008; Fernandes et al. 2014), we believed that such an approach was a natural choice given that we wanted to achieve higher student engagement and create a meaningful experience connected with their daily professional tasks (Sozykin, Koshelev, and Ustalov 2019). Generally, project work was set up as "data-centric competitions,"[4] where students are given a certain model and their task is to collect data to train the model best (as opposed to traditional ML competitions with a fixed data set and a model to be trained). The better labeling quality the model offers in the end, the more points are awarded to the students. Each round was focused on working through a corresponding condition (data quality, budget limitations, and prolonged time span). One round took about a month of group project work along with lectures and seminars dedicated to certain aspects of data labeling.

Another reason to create three project rounds with a similar structure is the challenge of changing the students' mindset towards a more data-conscious attitude. Mindset change is a complicated goal best achieved by experiential learning. As described by Kolb (1984), an experiential learning cycle is a four-stage process, where a learner gains some experiences, reflects on them, conceptualizes these reflections into new approaches, and then tests them to see how the first experience changes. Going through three similar experiences is intended to allow our students to continually modify their approach to data labeling and have an instant opportunity to experiment with newly gained insights (Figure 1).

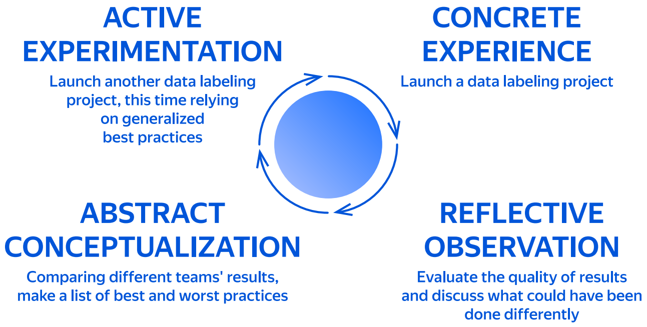

ACTIVE EXPERIMENTATION
Launch another data labeling project, this time relying on generalized best practices
CONCRETE EXPERIENCE
Launch a data labeling project
ABSTRACT CONCEPTUALIZATION
Comparing different teams' results, make a list of best and worst practices
REFLECTIVE OBSERVATION
Evaluate the quality of results and discuss what could have been done differently

Figure 1: Kolb cycle as applied to our course.

Round 1: Quality

The first round of project work addresses the issue of annotation quality. Students are divided into small groups of 3-4, each group working on a collective project. There are two projects to choose from: a computer vision task and a natural language processing task.

Computer Vision Task. The CV task aims at classifying car license plates into different groups (civilian, military, diplomatic etc). It is based on a pre-trained MobileNetV2 model that was to be trained on an additional 8.5k images. The students are not allowed to use external models or models that were not trained on ImageNet; other ways to increase training efficiency are allowed. We provide the students with a public testing data set to evaluate the quality of their model while training it. At the end of the round, the students are given a private testing set to be labeled by their model. The F1 score of the least-represented class was chosen as a quality metric.

[4] https://https-deeplearning-ai.github.io/data-centric-comp/

Natural Language Processing Task. The NLP task aims at recognizing related search queries with a model trained to recognize paraphrases. It can be based on a standard BERT model or the Bilateral Multi-Perspective Matching Model[5] (to be chosen by the students). The students are provided with a set of 15k anonymized search user sessions that can be labeled, or they can collect synonymous queries from the crowd. At the end of the round, their model's estimations are compared with a set of ground truth estimations, using MAE as a metric for quality evaluation.

For both projects, evaluation consists of two aspects. 10 points are given for the data labeling project setup (as evaluated by an internal team of experts), and the other 10 points are distributed according to the competition leaderboard, where a maximum of 10 points is awarded to the team with the best performing model, and the other teams get points proportional to their result.

Round 2: Budget

The second round adds the issue of budget limitations. Again there are two projects, which are exactly the same as in the first round but swapped between the groups. The only novel thing in the task is that now the students need not only to achieve a certain quality, but also stay within a budget limitation which implies experimenting with non-manual labeling or smart data sampling. As before, groups can gain 20 points for the data labeling project itself and their competition result, and up to 5 extra points more if they spend the least amount of money.

Round 3: Time

The third round focuses on the challenge of maintaining stable quality of the ML model throughout a particular period of time and changing contexts. We introduce a new task of content moderation and each student works on it independently. They are given a comment classifier model trained on social media threads that classified comments into positive and negative. The model and its hyper-parameters are fixed and it is prohibited to change it. We prepared an unlabeled text corpus and two labeled test sets for this round. In one the 'positive' class was underrepresented and in another the 'negative' class was underrepresented. The students need to sample texts from the unlabeled corpus and label them using the crowdsourcing platform to increase the accuracy of the classifier on both test sets. As before, there is a competition-based grading system, where students can get up to 10 points depending on their result on the first testing set, and up to another 10 points for their result on the second testing set. Simple accuracy was chosen as a quality metric. The two competitions take place successively within a week, with a session to discuss current results in between. At the end of this round students also prepare a report on their work with this model, which allows them to get up to 10 points. At the end of each round, there are group sessions to reflect on the process and collect insights to be applied next time.

The three projects naturally made up three modules of the course. Apart from the sessions dedicated to the projects, the course consisted of the very same topics that were introduced in its first version: decomposition, project design, quality control, pricing, aggregation, automation, etc. In order to demonstrate the variety of industrial data labeling needs, we preserved the case studies from the latest version of the course, but we left the former practical sessions behind. Lastly, we added a few topics, mostly concerning long-term data management: data quality estimation and monitoring and data pre-processing techniques (active labeling, pre-labeling, pseudo labeling, human-in-the-loop pipelines). All topics were distributed across the modules depending on their relevance to the core module theme: quality, budget, or time (see Figure 2 for a visualization of the curriculum). The final grade for the course consisted mostly of grades for project work (up to 20 + 25 + 30 points) and several extra assignments, all summing up to a maximum of 100 points.

Feedback

In this section, we compare feedback about the two versions of the course and discuss some insights based on the students' impressions. There are two sources of feedback about the course. First, there is a formalized course questionnaire. Second, we held a series of interviews after the new version of the course was launched.

The course questionnaire is launched by our partner university management and is mandatory for all students who finish a class. Thus, it is not course-specific, but covers general aspects of teaching it:

- usefulness of the course for one's future career
- usefulness of the course for broadening one's outlook
- novelty of the knowledge received
- complexity of the course

These questions can be graded from 1 to 5, where 1 is the most negative answer and 5 is the most positive. All questions are mandatory for the students, though they can choose a "Hard to say" answer instead of a score.

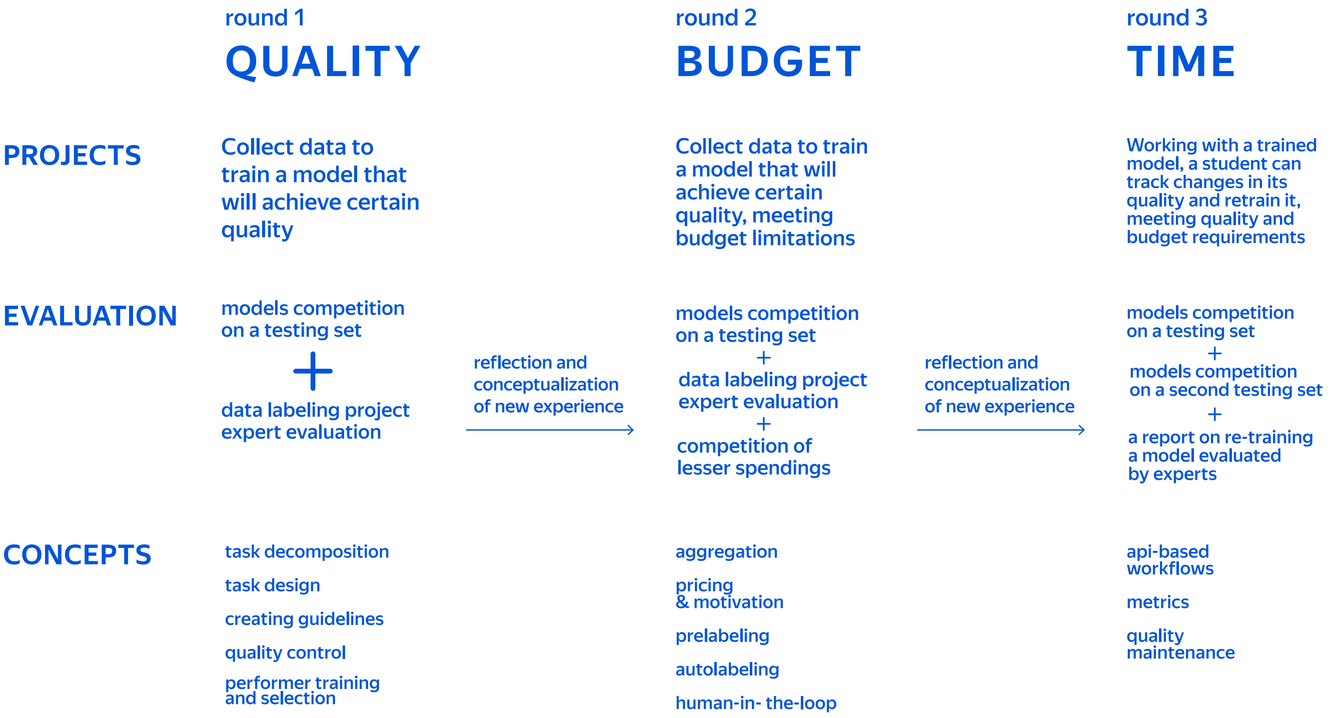
This questionnaire is run after each semester, so we have two sets of measurements, the first from 2020 and the second from 2021, which allows us to compare students' opinions about two versions of "Data Collection and Labeling for Machine Learning." It is important to note that there are two groups of students who participate in our course: third-year and fourth-year students. In this section, we will compare their questionnaire results separately as they show some peculiarities.

Apart from these questionnaires, in 2021, after launching the second version of the course, our team held a series of interviews with students that also offered some insights about their experience during the course.

Results. Third-year students' impressions show a drastic difference between the two versions of the course as related to all four general questions stated above (35 students participated in the survey):

- usefulness of the course for one's future career *increased* from 3.16 in 2020 to 3.8 in 2021
- usefulness of the course for broadening one's outlook *increased* from 3.69 to 4.3

[5] https://github.com/zhiguowang/BiMPM



	round 1 QUALITY		round 2 BUDGET		round 3 TIME
PROJECTS	Collect data to train a model that will achieve certain quality		Collect data to train a model that will achieve certain quality, meeting budget limitations		Working with a trained model, a student can track changes in its quality and retrain it, meeting quality and budget requirements
EVALUATION	models competition on a testing set + data labeling project expert evaluation	reflection and conceptualization of new experience →	models competition on a testing set + data labeling project expert evaluation + competition of lesser spendings	reflection and conceptualization of new experience →	models competition on a testing set + models competition on a second testing set + a report on re-training a model evaluated by experts
CONCEPTS	task decomposition task design creating guidelines quality control performer training and selection		aggregation pricing & motivation prelabeling autolabeling human-in- the-loop		api-based workflows metrics quality maintenance

Figure 2: The project-based syllabus we started using in 2021: Quality, Budget, and Time. It covers the same topics as in the previous years (Table 2), but organizes them in three semantically-coherent topics focused on annotation quality, budget, and time.

- novelty of the knowledge received *increased* from 4.0 to 4.5
- complexity of the course *increased* from 3.0 to 3.8

The fourth-year students' feedback not only shows less contrast but even demonstrates a drop in two aspects (35 students participated in the survey):

- usefulness of the course for one's future career *increased* from 3.37 in 2020 to 3.4 in 2021
- usefulness of the course for broadening one's outlook *decreased* from 3.8 to 3.2
- novelty of the knowledge received *decreased* from 3.8 to 3.4
- complexity of the course *increased* from 2.7 to 3.0

Analysis. The reason for the disparity probably lies in the difference between third-year and fourth-year curriculums. While third-year students have just started courses related to machine learning, deep learning, and such, fourth-year students have already spent a year working with ML models. Thus, a curriculum based on working with algorithms is less novel to them when compared to a curriculum based on data labeling only. This assumption can be indirectly confirmed by the fact that "novelty of the knowledge received" received a lower grade for the second version of the curriculum. Indeed, third-year students are new to training models and feel more optimistic about content novelty and its impact on broadening their outlook.

Even though the third-year students gave high scores for novelty of the course, the in-depth interviews demonstrated some frustration. Indeed, the third-year students often lacked deep learning skills and struggled with training models, and ultimately had to get help from their team members. Most respondents told us that they would recommend taking this course to those who are already acquainted with deep learning: "I would recommend the course for fourth-years, and for third-years I'd say not to attend or you will suffer" (respondent V., third-year student).

The rest of the curriculum critique was related to having to work on the same projects in rounds 1 and 2, lack of using workflow automation tools like Apache Airflow[6], and having various professionals as course lecturers (different lecturers sometimes repeated others), even though the variety of case studies "helped to understand how powerful crowdsourcing is."

The practical focus of the course was perceived as its most valuable point and a strong differentiator from the other courses in the university curriculum. The respondents noted that "the idea of why it is important to invest in data settled better than before," "data labeling appeared in my picture of the world: now I know that if I don't have data – I can label it – here are the tools for that," illustrating the mindset shift that we were aiming at. Other respondents noted that "it was impressive how the full cycle was covered, from raw

[6] https://airflow.apache.org/

data to a trained algorithm," "I added a new skill and tool to my CV," noting that real-life tasks contributed to their professional experience too.

Limitations

Based on the feedback, a project-based curriculum with data-centric algorithm competitions produces various limitations for teaching the course and exporting this approach.

The most serious limitation concerns the place of this course in a curriculum of machine learning specialists or data scientists. As discussed above, students who are relatively new to working with algorithms and training models tend to feel much more optimistic about the course. On the other hand, the very same group of students struggles with tasks that demand machine learning or deep learning skills as they have just started with these disciplines. This contradiction can probably be solved by moving the course to another semester (in our case, spring) and targeting it at those students who are already acquainted with basic deep learning skills, but are not so experienced that they will find the course materials obvious.

The machine learning and deep learning prerequisites also are a limitation on their own. The course itself can not be freely introduced to any program that does not involve advanced programming or machine learning skills.

Another source of course limitations is its focus on industrial and business practices. Not every business case can be easily transformed into a learning task, as some of them contain non-disclosable internal findings. This happened to one of the projects used in this course. We could not reveal the whole algorithm to the students and had to cover various organizational issues such as running the model by ourselves.

Also, the practical focus suggests inviting data labeling experts and engineers to teach the course. Having non-professional lecturers is a strength and a weakness at the same time. Though these lecturers offer a great variety of cases and experience, they often require additional training in public speaking and facilitating learning sessions. This becomes more important because of the project-based approached as curating projects requires more facilitating and supporting skills as compared to traditional lecturing.[7] Also, having multiple lecturers requires an additional person to be a course coordinator (learning specialist) who can create a successive narrative and avoid content duplication.

Future Work

There are several directions of how to develop the existing syllabus and the approach behind it.

First and foremost, this syllabus or certain parts of it could be used in machine learning or data science specializations in other universities or educational projects.

Secondly, since we found that data-centric competitions were a success, the corresponding content about data labeling could be included in courses on machine learning, deep learning, or certain areas of it (i.e., computer vision or natural language processing). This approach can demonstrate the value of proper data labeling pipeline design in direct connection with students' main specialization.

Apart from it, we are planning to create a shorter public version of the main syllabus which will provide a basic introduction to industrial data labeling.

Conclusion

In this paper, we presented two approaches to creating a practical data labeling course for future ML engineers and data analysts. The curricula compared were very similar content-wise but differed in the way the course was designed. The first curriculum followed a straightforward structure, consisting of weekly lectures and seminars that were dedicated to running a certain type of data labeling projects. The second curriculum was organized as three project rounds, designed as data-centric ML competitions. The students went through a whole cycle of collecting training data, which also involved training a model and estimating the quality of its predictions.

The feedback collected allows us to track students' satisfaction separately for two groups: third-year students, who just started with machine learning and related disciplines, and fourth-years, who already have experience in ML. These two groups show different results. Third-years are in all aspects (usefulness, novelty, applicability) more satisfied with the project-based curriculum. Fourth-years tend to be more skeptical about the novelty of the experience, nevertheless evaluating the project-based curriculum as more complex and more applicable to their careers.

References

Bell, S. 2010. Project-Based Learning for the 21st Century: Skills for the Future. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 83(2): 39–43.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, 313–322. New York, NY, USA: ACM. ISBN 978-1-4503-0271-5.

Boaler, J. 1998. Open and Closed Mathematics: Student Experiences and Understandings. *Journal for Research in Mathematics Education JRME*, 29(1): 41–62.

Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4): 324–345.

Chui, M.; Manyika, J.; and Miremadi, M. 2018. What AI can and can't do (yet) for your business. *McKinsey Quarterly*, 1: 96–108.

Cognilytica. 2019. Data Engineering, Preparation, and Labeling for AI 2019. Technical Report CGR-DE100, Cognilytica LLC.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and

[7] On the importance of facilitation in project-based learning see Hmelo-Silver, Duncan, and Chinn (2007); Bell (2010).

Assurance Actions. *ACM Computing Surveys*, 51(1): 7:1–7:40.

Dawid, A. P.; and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society*, 28(1): 20–28.

Drutsa, A.; Ustalov, D.; Fedorova, V.; Megorskaya, O.; and Baidakova, D. 2021. Crowdsourcing Natural Language Data at Scale: A Hands-On Tutorial. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, NAACL-HLT 2021, 25–30. Online: Association for Computational Linguistics.

Estellés-Arolas, E.; and González-Ladrón-de Guevara, F. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2): 189–200.

Fernandes, S.; Mesquita, D.; Flores, M. A. a.; and Lima, R. M. 2014. Engaging students in learning: findings from a study of project-led education. *European Journal of Engineering Education*, 39(1): 55–67.

Hassan, H.; Domínguez, C.; Martínez, J.; Perles, A.; Albaladejo, J.; and Capella, J. 2008. Integrated Multicourse Project-based Learning in Electronic Engineering. *International Journal of Engineering Education*, 24: 581–591.

Hmelo-Silver, C. E.; Duncan, R. G.; and Chinn, C. A. 2007. Scaffolding and Achievement in Problem-Based and Inquiry Learning: A Response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, 42(2): 99–107.

Kolb, D. A. 1984. *Experiential Learning: Experience as the Source of Learning and Development*. Prentice-Hall, 2nd edition. ISBN 9780132952613.

Mioduser, D.; and Betzer, N. 2008. The contribution of Project-based-learning to high-achievers' acquisition of technological knowledge and skills. *International Journal of Technology and Design Education*, 18(1): 59–77.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 1–15. Yokohama, Japan (Online): ACM.

Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; and Dennison, D. 2015. Hidden Technical Debt in Machine Learning Systems. In *Advances in Neural Information Processing Systems 28*, NIPS 2015, 2503–2511. Montréal, QC, Canada: Curran Associates, Inc. ISBN 9781510825024.

Sozykin, A.; Koshelev, A.; and Ustalov, D. 2019. *The Role of Student Projects in Teaching Machine Learning and High Performance Computing*, 653–663. Cham, Switzerland: Springer International Publishing. ISBN 978-3-030-36592-9.

Zhang, X.; Li, G.; and Feng, J. 2016. Crowdsourced Top-k Algorithms: An Experimental Evaluation. *Proceedings of the VLDB Endowment*, 9(8): 612–623.

Zheng, Y.; Li, G.; Li, Y.; Shan, C.; and Cheng, R. 2017. Truth Inference in Crowdsourcing: Is the Problem Solved? *Proceedings of the VLDB Endowment*, 10(5): 541–552.