Precog

Requirements Engineering toward Safe Perception for Autonomous Mobility

Public report

Authors: Markus Borg, Hans-Martin Heyn, Jennifer Horkoff, Khan Mohammad Habibullah, Alessia Knauss, Eric Knauss, Polly Jing Li

Date: 2022-05-30

Prestudy within ”Trafiksäkerhet och automatiserade fordon”

FFI Fordonsstrategisk Forskning och Innovation

VINNOVA Energimyndigheten TRAFIKVERKET FKG VOLVO SCANIA VOLVO

Innehållsförteckning

1 Sammanfattning

Framtidens autonoma mobilitet kräver tillförlitlig fordonsperception. Perceptionssystem i fordon använder huvudsakligen datorseende med hjälp av kameror som kompletteras med andra sensorer baserade på exempelvis LiDAR, radar, och ultraljud. De bästa resultaten uppnås med hjälp av sensorfusion och en kombination av traditionell signalbehandling och maskininlärningsbaserat datorseende.

Maskininlärning är idag den enda metoden vi känner till för att uppnå framgångsrikt datorseende i trafikmiljön. Klassiska regelbaserade algoritmer för datorseende lyckas inte generalisera till den mångfald av situationer som förekommer i den operativa miljön. I stället tränas olika modeller av typen djupa faltningsnätverk (Eng: convolutional networks) med övervakad (eller semi-övervakad) maskininlärning. Maskininlärningsmodeller kan betraktas som opålitliga översättningsfunktioner från indata till utdata – ibland blir resultatet inte det förväntade. Detta innebär nya utmaningar för verifiering och validering inom fordonsindustrin och stora summor investeras i forskning och standardiseringsinitiativ.

Övervakad träning av maskininlärningsmodeller kräver enorma mängder annoterade träningsdata. Med hjälp av denna data fångas mönster i maskininlärningsmodeller. Perceptionssystemen använder sedan modellerna för att generalisera när nya kameradata tas emot. Beroende på vilken typ av utdata modellerna tränades för kan detta användas för att t.ex. identifiera objekt i trafikmiljön eller för att hitta fritt utrymme på vägbanan.

Användning av maskininlärning leder till en stark koppling mellan träningsdata och kvalitet på den slutgiltiga fordonsfunktionen. Denna koppling har inte tidigare undersökts tillräckligt med systematiska metoder från kravhantering. I denna förstudie använde vi en konceptuell modell (se Figure 3) för att driva gruppintervjuer med syfte att utreda beroenden i teknologistacken från träningsdata, via annotering och modellträning, till perceptionssystem och fordonsfunktion. Intervjuerna genomfördes i kontexten av den säkerhetsargumentation som krävs för driftsättning av maskininlärningsbaserad mjukvara i fordon som ska framföras på allmän väg.

Tematisk analys av den insamlade kvalitativa datan resulterade i utmaningar och framtida forskningsfrågor inom åtta huvudsakliga teman: 1) Data, 2) Perception, 3) AI/ML-aspekter, 4) System- och mjukvaruuveckling, 5) Kvalitet, 6) Affärsekosystem, 7) Kravhantering och 8) Annotering. Dessa åtta teman är inte oberoende utan påverkar varandra i olika grad. I denna rapport presenterar vi en djupare analys av det först temat, dvs. ”Data”, vilket vi identifierade som mest intressant inom förstudien. Vi beskriver bl.a. utmaningar inom 1) urval av data för träning och modellvalidering, 2) processer och specifikationer för datakrav och verifiering, och 3) kopplingen mellan informationsvärde i data och maskininlärningsmodellers korrekthet.

Vi planerar att fortsätta vårt arbete genom att publicera två vetenskapliga artiklar baserat på våra förstudieresultat samt genom att etablera ett större konsortium för en större forskningsstudie. Inom ramen för doktoranden Khan Mohammad Habibullah kommer vi att slutföra analysen av resterande sju teman och sammanställa insikterna i en artikel om nya perspektiv på kvalitetsaspekter för perceptionssystem och en lösningsorienterad artikel om kravhantering för funktionssäkerhet och maskininlärning. Målsättningen är att slutföra artiklar under 2022 och att ansöka om ett fullskaligt FFI-projekt vid nästa utlysning som beräknas öppna i januari 2023.

2 Executive summary in English

In this prestudy, we have explored challenges in key areas relating to data quality for ML-based atuomotive perception. Using our experience in automotive systems, data annotation, machine learning (ML), software and requirements engineering, we have designed and conducted a series of group interviews with autonomous driving (AD), safety, data and ML experts. Questions focused on understanding the key components of autonomous systems and their quality interdependencies, the effects of the driving ecosystem, safety concerns, the role of varying context, and trade-offs between competing non-functional requirements (qualities).

Using qualitative coding methods driven by our research team, we grouped our diverse and rich findings into eight themes. Each of these themes could be a focus of future research studies, with results and publications contributing back to the AD community. An overview of our results, when presented in a final workshop with relevant industrial players, generated much discussion and interest.

We plan to submit a number of publications describing pre-study results, including an overview of our findings, a submission focusing on perception system qualities and quality trade-offs and a solution-oriented submission focusing on new ways of working which better balances the complexity of ML problems with the need for safety and organizational-memory related documentation. We will use the prioritized results to narrow our focus and prepare an application for the Vinnova FFI Safe automated driving call opening in January 2023.

This prestudy has explored the complex space ML-driven AD, safety and data quality, and has enumerated a number of industry-supported topics and issues in this area. We believe our results will significantly contribute to useful knowledge bases, theories, and methods which support perception systems and AD, maintaining innovation while prioritizing safe operation.

3 Background

Autonomous mobility relies on safe perception. State-of-the-art solutions combine traditional signal processing with perception models based on supervised ML as an area with substantial challenges related to verification and validation (V&V) in the automotive industry.

ML for perception relies on massive amounts of annotated training data to learn patterns in the input. Consequently, there is a strong connection between the quality of the dataset on which a perception system's ML model is trained and its resulting correctness. Academic research has focused on how annotation noise influences the output accuracy of ML models. Unfortunately, the connection between dataset annotation precision, ML model accuracy, perception system correctness, and functional safety remains largely unclear.

Requirements engineering is a cornerstone in quality assurance and safety engineering. However, there is not yet an established body of knowledge covering requirements engineering for the ML paradigm. While there is ongoing research, we see great potential in new exploratory research projects dedicated to requirements engineering for ML-based perception systems.

4 Purpose, research questions, and method

Human error is by far the most common cause of traffic accidents, suggesting that increasing the level of driving automation is an important way to reach a substantially reduced traffic mortality rate. The automotive industry has successfully harnessed the disruptive potential of ML over the last decade, enabling new levels of vehicular perception. However, software engineering using ML represents a paradigm shift. No longer do human engineers explicitly express all logic in source code. Instead, ML models are trained using enormous amounts of annotated data.

ISO 26262, the quintessential automotive functional safety standard, is a poor fit for the ML paradigm as established engineering practices such as code reviews and code coverage testing are less effective. New initiatives such as UL4600, ISO 21448, ISO/IEC AWI TR 5469 evolve as complementary resources. As the ML paradigm interweaves source code and training data, novel quality assurance approaches are critically needed.

Several factors influence the performance of an ML-based perception software system. In Precog, we argue that the two most important factors are:

1. **ML model capabilities** - Contemporary ML algorithms are essentially loss-minimization functions that parameterize a model to minimize errors. These algorithms are typically formulated as gradient descent-based optimization problems where the difference between an input and the desired output for a large set of examples is to be minimized. The performance of an ML-based algorithm will be limited by the model's ability to find existing patterns in a dataset. State-of-the-art deep learning model architectures for perception have a vast number of parameters, increasingly making them able to approximate a dataset very precisely.
2. **Dataset size and quality** - If we assume that an ML model is a perfect approximator, we find that the training dataset limits the output accuracy. The main challenge then turns into encoding the expectations on the system into the dataset, i.e., programming by example. The engineers must have a clear idea of what performance you expect from your perception system, or they will be unable to prepare the necessary training dataset.

Before designing solutions in response to the factors above, we need to have a rough agreement on what accuracy is needed from the ML-based perception systems, i.e., a high-level alignment of the corresponding quality requirements. Precog initiated a synthesis of expertise from industry and academia in relation to these two factors. While many research studies focus on ML model capabilities, Precog was inspired by the trend of moving *from model-centric AI to data-centric AI*. The overall gist of the pre-study is captured in Figure 1, i.e., the blurb used when inviting participants to the final workshop.

Autonomous driving requires successful perception. Successful perception relies on machine learning. Machine learning builds on high-quality data. What does that mean for automotive perception? How can we specify the expectations? With what precision shall the training data be annotated? *How accurate does the perception system need to be for use in safety-critical automotive functions?*

Figure 1: The blurb used in the concluding workshop invitation showing the gist of the Precog prestudy.

The backbone of the Precog study was a series of group interviews with key stakeholders in the Swedish AD industry. Our prestudy was guided by a series of research questions, and we used these research questions to help us design interview questions. Our research questions were as follows:

RQ1: What are the main conceptual components and steps in the development of an ML-based perception system, and how do these components and steps depend on each other?
RQ1.1: Are these dependencies documented and how?

RQ2: How are safety and safety cases captured in ML-based perception systems?
RQ 2.1: How do safety requirements relate or depend on data and data quality?

RQ 2.2: How are these dependencies captured?

RQ3: Does the context of a perception system affect how data quality is defined and captured? If so, how?

RQ4: How does the automotive ecosystem play a role in the negotiation and capture of data quality and safety requirements for perception systems?

RQ5: How does data quality for perception systems trade-off against other quality requirements as found in the whole perception system?

We describe our methodology in more detail in the following, including sampling, data collection and data analysis.

Sampling: We choose participants who had experience with ML, perception systems, data science, and who were working in the AD industry. The sampling method was a mix of purposive, convenience, and snowball sampling, and included project partners when possible. We sent open calls to the AD industry, and our known contacts who are working in AD industry, then we asked the interviewees if they knew other qualified people with similar experience we could contact. We interviewed 19 participants from five different companies from the AD industry, and we believe that the selected interviewees are representative of practitioners working with ML, perception systems, and/or data science in the AD industry. At least three people from the researcher side were presented in each of the interview sessions. Two researchers were presented in all the interview sessions for maintaining consistency.

The interviews were conducted between December 2021 and April 2022 via Microsoft Teams, and lasted between 1 hour 30 minutes to 2 hours. We recorded all the interview sessions with the permission of all participants; then transcribed, and anonymized, the recordings for further analysis.

Data Collection: We sent the participants an invitation email to join in the interview session that contained information about the research project (e.g., financial support, study plan, duration of the research study), description of the purpose of the study (research goal), and formal information (e.g., participants' voluntarily participation, permission for recording, data anonymizing, data storing and analyzing process). We also shared our conceptual model of a perception system in an autonomous vehicle (Figure 2) with our participants so that they could get familiar with the idea of the interview topic and if it matches their context.

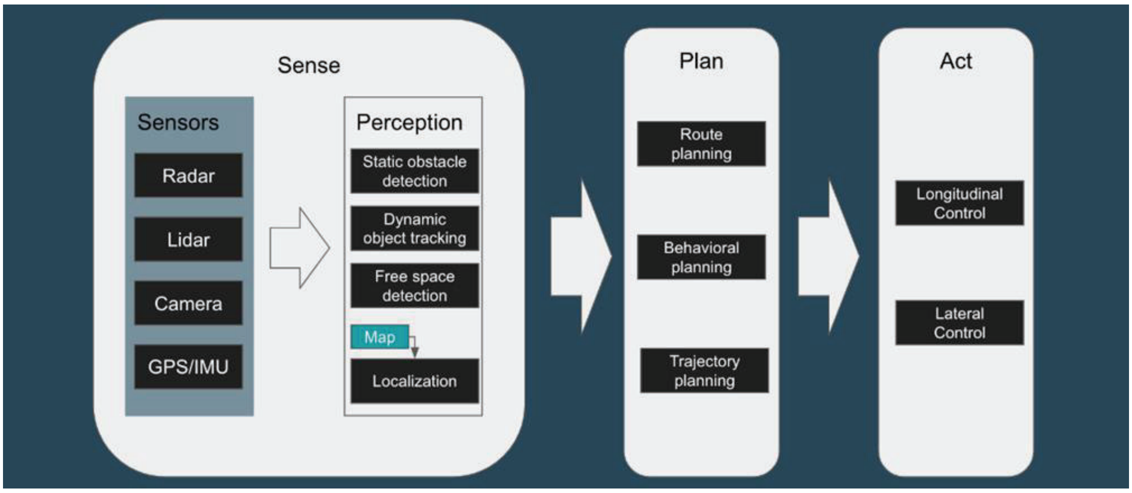

Figure 2: An adapted version of the Sense-Plan-Act model used to align participants in the group interviews.

We used semi-structured interviews with a set of predetermined open-ended questions to keep enough freedom to add follow-up questions and discussions to collect in-depth information. We followed an interview guide as a framework to facilitate the interview process. The interview guide followed an elaborated version of the research questions above. The interview started with an informal introduction of the participants and with a brief description of the research project and the background of the study to make the interviewees comfortable with the context of the interview.

We started by asking for demographic information about the participants, e.g., the participants' role in the company, years of experience, their working domain, and ML/AI experience. Then we moved to the architecture/foundational questions and asked the participants' positions by showing them the conceptual model of the established Sense-Plan-Act robot control procedure customized as an architecture view for automated driving, showed in Figure 2. We also asked them whether the architecture is valid in their context. We then showed them Figure 3, asking for their feedback and using the figure to ground further discussions. We continued with questions about safety and safety cases, context, ecosystems, quality trade-offs, and SAE level data. We closed the interview session by asking them if we forgot anything related to ask or if they wanted to add anything, and who else we should be interviewing as part of this study.

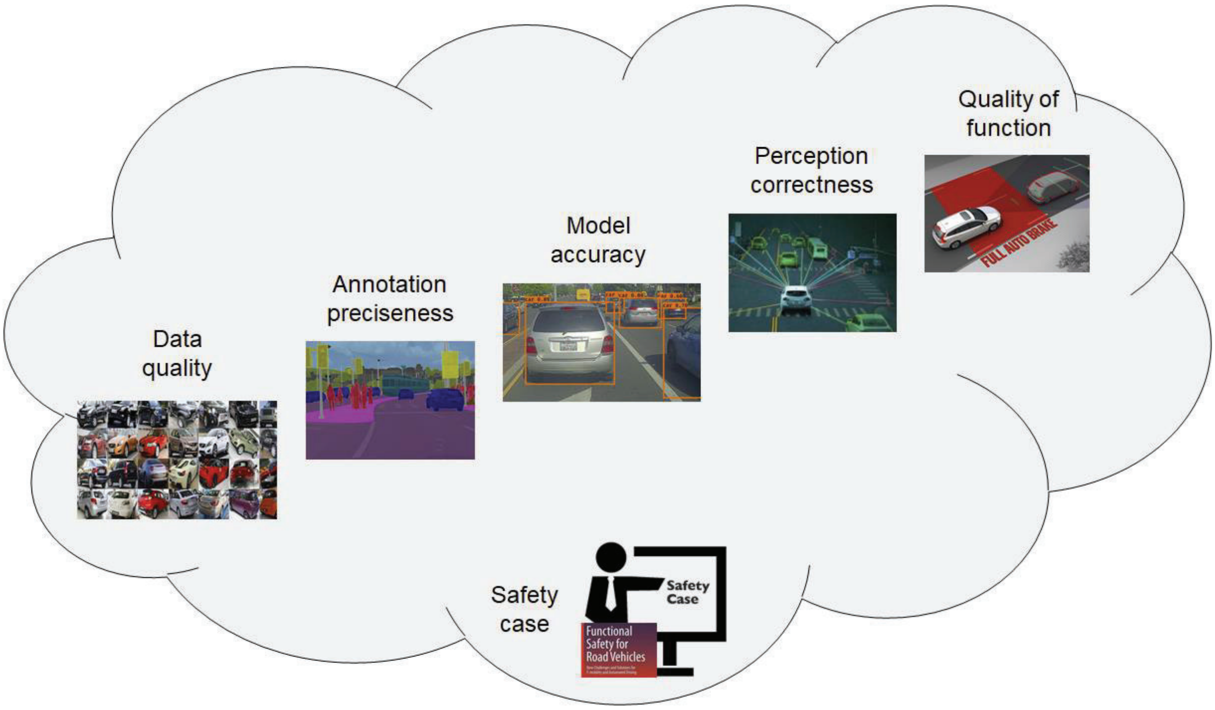

Figure 3: Conceptual model of quality transitions from data collection to the quality of the automotive function. The cloud represents automotive engineering in the context of a holistic safety case.

Data Analysis: The collected data were qualitative, and we used thematic analysis as a data analysis method. We used a mixed form of coding, where we started with a number of high-level deductive codes based on the research questions, interview questions, and the researchers' experience, then we refined and adapted those codes and started inductive coding when going through the transcripts. We developed different codes based on the themes of the transcribed data and mapped the interview data accordingly. At least three of the researchers coded each of the transcribed interviews, and afterward reviewed, validated and added more inductive codes, if necessary, by one researcher who, if possible, did not participate in the first-round coding session of the specific interview. The resulting codes are summarized in Section 6.

Data Validation: We ended the prestudy with a workshop inviting company partners' participants in the interviews, and other interested parties. In a 2.5-hour workshop with roughly 20 participants online and in person, we presented our findings in terms of themes and sub-codes with illustrative quotes. Participates voted on which codes were the most important or relevant to them, and some discussions over the findings occurred. These prioritizations will help us to direct future research and focus follow-up applications.

5 Goals

The goal of this prestudy was to work toward an alignment of high-level expectations on ML-based perception systems in autonomous mobility. We refer to this as establishing an initial requirements framework. Before the study, our initial discussions with a diverse set of stakeholders revealed the need to find common ground prior to starting a full-scale project. Furthermore, based on discussions with international partners, we have encountered substantial discrepancies concerning perception expectations. The discrepancies indicate significant uncertainty within the automotive industry. As an example, experience from a project partner has revealed that dataset annotation accuracy can differ by order of magnitude. We argue that the industrial development of ML-based perception systems will remain hampered until we align expectations throughout the automotive value chain.

As a pre-study, we wanted Precog to pave the way for a larger future research study. The ambition was to enable this by 1) identifying potential partners for a larger Swedish research consortium and to 2) perform an initial challenge elicitation. The latter should be used as a foundation to specify new research questions and provide input for a work package structure for a future project.

6 Results

Precog initiated work on a requirements framework to help align high-level expectations on ML-based perception systems in autonomous mobility. The major contribution toward this goal was a challenge situation based on a series of group interviews with engineers active in the Swedish AD industry. In the final week of the project, we validated and extended the findings in an open hybrid workshop hosted by SAFER in Gothenburg. The interviewees, and the workshop participants, all represent companies that could be partners in a larger Swedish research consortium.

In total eight major themes were identified as results of this study:

1. Data
2. Perception
3. Artificial intelligence and machine learning concerns
4. System and Software Engineering
5. Quality
6. Ecosystem and Business
7. Requirements Engineering
8. Annotation

The figure below shows an overview of the eight major themes identified during the challenge elicitation.

Figure 4: Overview of the eight themes discovered in the interview study.

In this prestudy report, we share an analysis of the Data theme – since this is at the heart of the Precog project. However, we stress that the remaining seven themes are important too, and in many ways interrelated to the challenges within the Data theme. As we continue analyzing the rich qualitative data collected during the project, we will provide similar analyses for the remaining themes. We aim to publish an academic paper when the entire analysis is done and the resulting work will be publicly available as open access.

6.1 General Challenges Related to Data

Unlike conventional software systems in which rules specify the desired behavior, an ML system infers these rules from data. Therefore, data plays a prominent role when trying to ensure correct behavior of an ML-based perception system. If the data provided for training is biased, the resulting ML system will learn that bias, and consequently the decisions of the system will be biased.

During the interviews, we asked about the processes and ways of working used in relation to specify and collect data for the development of ML models. Furthermore, we wanted to learn what the interviewees understand under "data quality".

After the interviews we defined themes based on the given answers and sorted them into the categories *"Data collection"*, *"Processes and ways of working"*, and *"Data quality"*. In a workshop visited by both interviewees and additional practitioners, we asked interviewees to rank the themes according to their opinion of severity towards ensuring safety and correctness of the machine learning model's behavior. We also encouraged to add additional themes, or mark themes as unnecessary.

The results are illustrated in an Ishikawa diagram in Figure 5.

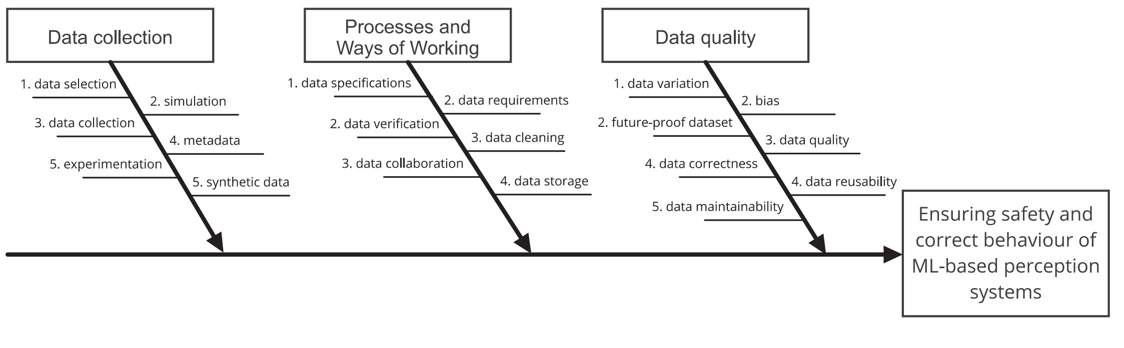

Figure 5: Ishikawa diagram illustrating data aspects contributing to safety and correct behaviour of ML-based perception systems

6.2 Challenges Related to Data collection

Ensuring the correct behavior of an ML model requires a highly data-driven development. The participants of our study ranked correct data selection as a key ability towards ensuring safety and correct behavior. Often, the right data set for training is found through a set of iterations between the AI experts and the data scientists. For example, uncertainty measures can be used to decide which data additional data needs to be selected to reduce uncertainty, as illustrated by a quote from one of our group interviews:

> ***"To ensure some form of safety measures from the model, we produce uncertainty estimations from outputs. Those are used in the data selection of course to look for what type of data are we uncertain? What do we need to learn more from?"***

Because finding the right data can be expensive, development teams try to use simulations to create data for training and validation. The study participants even consider data from simulation more important than data originating from planned experiments, such as test drives. There can be two reasons for using data from simulation: First, it is often significantly cheaper to obtain these kinds of data, and second it is possible to obtain data from rare case scenarios, that could be impossible to obtain in real world experiments:

> ***“Especially like for the rare case scenario that is not really easy to replicate in real world, so we cannot.”***

> ***“Furthermore, simulations are very often an integral part of test strategies for machine learning based systems.”***

To summarize, the key aspects in data collection towards ensuring safety and correct behavior of ML-based perception system is the mechanism for selecting the data used for training and validation of the models. Often, simulations are used to create additional data, especially for rare case scenarios, and to provide a test environment of the system. Today, companies already employ iterative methods to find the "right" data selection, but further research is needed to define standardized ways to specify and find the right data selection.

6.3 Challenges Related to Processes and Ways of Working for Data

We asked during the interviews which processes and ways of working in regard to data for ML models are currently being used. Most safety standards, such as ISO 26262, rely on the correctness of processes to build up a safety case for a product. Therefore, it is important to define processes that ensure the safety of ML-based systems.

The most important capability, defined formally through a process, or informally through a way of working, is the creation of data specifications. We saw earlier that data selection is the most important activity in the data driven development process for ML-based systems. Data specifications are a logical prerequisite to data selection. But it is often quite unclear what a data specification should entail:

> ***"It's very different how you write a data specification [...] it's hard to know what the future expects and what type of classes we want and how we do want to combine certain objects."***

As mentioned before, iterative processes can be used to find the final data specification of the system:

> ***“It's more of a sort of a data driven and then statistical analysis of the data in a continuous way. So we start with logging data, annotating it, training or models, and so on. And then we can also draw some statistics. OK, how is the class balance in this data set? How does that affect the per class accuracy? Do we need to look for more? And then we can of course feedback that to the data selection team and they can start looking for certain classes, for example.”***

A specification is a closed set of requirements. However, the exact meaning of a data requirement is not entirely clear. Data requirements can for example describe desired probability distributions and quantity of the data:

> ***"We write documents, word documents, basically where we describe the distribution of the data and the quantity of the data that needs to be collected."***

Data requirements can also entail specific data quality aspects, such as pixel density, brightness, size of bounding boxes, etc. In both understandings of data requirements, they allow for data verification. Data verification means checking that the data is representative for the desired "real-world" scenario.

Data specifications, as well as data requirements, are key enablers for many companies to collaborate in data collection and processing for example with supplier companies:

> ***"We have a 3rd party company driving around all this mileage and collecting data. They want you to send over that data to them for doing the simulations. And then they will put their requirements on what sort of data we are collecting."***

To summarize, the most important processes aim towards finding data specifications, requirements, and verification strategies. Processes for finding data specification and requirements most likely include iterative steps that reflect the data driven nature of ML model development. Therefore, a final data specification that describes the utilized data can be a key input towards a safety case of a ML-based perception system. Furthermore, data requirements are a key enabler towards data collaboration between companies.

6.4 Challenges Related to Data Quality

Data requirements often entail some desired data quality aspect. Interestingly, the most important data quality aspects mentioned by the interviewees do not describe physical properties of data, such as pixel density, contrast, resolution, brightness, etc., but instead describe the represented information in the data.

Because data selection was identified as the most important aspect for data collection, it comes to no surprise that data variation has been chosen by the study participants as the most important data quality characteristic. Data variation is also directly causally related to bias in data. A lack of data variation will have the effect of a bias, which can propagate into the ML model. Both data quality aspects require however an a-priori understanding of the environment in which the system will be deployed. If you do not know your operational domain, it will be difficult to describe what variety entails:

> ***“You need to understand the distribution of where to collect data and that requires an understanding of where the function in the end will be used.”***

A challenge regarding data variation is the definition of KPIs, or in general measures of variety. How do you measure variety, and when do you know that your data has enough variety?

> ***“How would you divide that space and define it in a way that allows a measure of have I covered not only enough children, but also enough variety of children?“***

> ***“Does my data cover at least enough children to statistically say something about how good my quality to take children is.”***

Collecting and processing data is often a costly part of the development process. Therefore, the re-usability, and future-proofness of data are considered important data quality aspects.

> ***“What do we need to ensure to make use of the data we've collected up to now? I mean, how do we make sure we don't have to start from scratch?”***

To summarize, the most important aspects of data quality for ensuring safety and correctness of ML-based perception systems are not the physical properties of the data themselves, but related to the information content of the data. Specifically, data variation is considered the most important data quality aspect, even before data correctness. This might stem from the aim of preventing bias in ML models because a lack of data variation most likely is a cause of bias. It remains a challenge to define KPIs and useful measures of data variation.

7 Dissemination and publications

This short section briefly positions the knowledge sharing from this prestudy and our future publication plan.

7.1 Knowledge sharing

How has / is the project result to be used and disseminated?	Applicable?	Comments
Increase knowledge in a specific area	X	Based on interviews and a joint workshop on data requirements toward safe autonomous driving, Precog initiated knowledge development in the area.
Be passed on to other advanced technological development projects	X	The results will further evolve as part of Khan Mohammad Habibullah PhD studies at GU. Moreover, we have initiated discussions aiming at a larger project application within a year.
Be passed on to product development projects		
Introduced to the market		
Used in investigations, regulations, permit matters/political decisions.		

7.2 Publications

No publications during this short pre-study. However, as explained in Section 6, we have started working toward a joint academic publication based on the interview study. Khan Mohammad Habibullah is the lead author and the plan is to include the work in his upcoming PhD thesis.

8 Conclusions and future research

In this prestudy, we have explored challenges in key areas relating to data quality for ML-based automotive perception. Using our experience in automotive systems, data annotation, ML, software and requirements engineering, we have designed and conducted a series of group interviews with autonomous driving, safety, data and ML experts. Questions focused on understanding the key components of autonomous systems and their quality interdependencies, the effects of the driving ecosystem, safety concerns, the role of varying context, and trade-offs between competing non-functional requirements (qualities).

Using qualitative coding methods driven by our research team, we grouped our diverse and rich findings into eight themes. In this report, as an illustration, we have focused on exploring the data theme, reporting on general challenges, data collection challenges, process-related data challenges, and data quality challenges. Similar results are available for the other seven themes. Each of these themes could be a focus of future research studies, with results and publications contributing back to the AD community. An overview of our results, when presented in a final workshop with relevant industrial players, generated much discussion and interest.

We can describe our intended future work in terms of both projects and publications. From a publication perspective, we will write an overview paper as part of the PhD work of Khan Mohammad Habibullah, summarizing key findings in the themes. This PhD topic focuses particularly on non-functional requirements for ML, so we will focus a further publication on the quality theme describing emerging qualities and trade-offs related to ML and perception systems. Further work can move towards the solution space. Findings collected as part of the requirements engineering and system and software engineering themes can lead us to develop new ways of working which better balances the complexity of ML problems with the need for safety and organizational-memory related documentation.

In terms of projects, we will use the prioritized results to narrow our focus and prepare an application for a larger project application. We are currently planning to submit an application to the Vinnova FFI Safe automated driving call opening in January 2023. The findings in this project will also contribute to ongoing projects such as the VR project supporting Khan Mohammad Habibullah. Because our prestudy results are both rich and broad, further project applications can potentially use our results as preliminary results (e.g., a project focusing on ML processes, a project focusing on safety, another focusing on data quality, etc.).

To summarize, this prestudy has explored the complex space ML-driven AD, safety and data quality, and has enumerated a number of industry-supported topics and issues in this area. We believe our results will significantly contribute to useful knowledge bases, theories, and methods which support perception systems and autonomous driving, maintaining innovation while prioritizing safe operation.

9 Participating parties and contact persons

Organization	Contact person	Logo
RISE Research Institutes of Sweden AB Scheelevägen 17 223 63 Lund	Markus Borg	
Annotell AB Lindholmspiren 7 417 56 Gothenburg	Polly Jing Li	
Zenseact AB Lindholmspiren 2 417 56 Gothenburg	Alessia Knauss	
University of Gothenburg Jupiter-building, Hörselgången 5, 417 56 Gothenburg	Jennifer Horkoff, Eric Knauss	