DS4ALL: All you need for democratizing data exploration and analysis

Paolo Bethaz, Khalid Belhajjame, Genoveva Vargas-Solar, Tania Cerquitelli

▶ **To cite this version:**

Paolo Bethaz, Khalid Belhajjame, Genoveva Vargas-Solar, Tania Cerquitelli. DS4ALL: All you need for democratizing data exploration and analysis. 2021 IEEE International Conference on Big Data (Big Data), Dec 2021, Orlando, United States. pp.4235-4242, 10.1109/BigData52589.2021.9671883 . hal-03621357

HAL Id: hal-03621357
https://hal.science/hal-03621357

Submitted on 28 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DS4ALL: All you need for democratizing data exploration and analysis

Paolo Bethaz
Department of Control and Computer Engineerin
Politecnico di Torino
Turin, Italy
paolo.bethaz@polito.it

Khalid Belhajjame
LAMSADE
Université Paris Dauphine
Paris, France
khalid.belhajjame@dauphine.fr

Genoveva Vargas-Solar
LIRIS
CNRS
Lyon, France
genoveva.vargas-solar@cnrs.fr

Tania Cerquitelli
Department of Control and Computer Engineerin
Politecnico di Torino
Turin, Italy
tania.cerquitelli@polito.it

***Abstract*—Today, large amounts of data are collected in various domains, presenting unprecedented economic and societal opportunities. Yet, at present, the exploitation of these data sets through data science methods is primarily dominated by AI-savvy users. From an inclusive perspective, there is a need for solutions that can democratise data science that can guide non-specialists intuitively to explore data collections and extract knowledge out of them. This paper introduces the vision of a new data science engine, called DS4ALL (Data Science for ALL), that empowers users who are neither computer nor AI experts to perform sophisticated data exploration and analysis tasks. Therefore, DS4ALL is based on a conversational and intuitive approach that insulates users from the complexity of AI algorithms. DS4ALL allows a dialogue-based approach that gives the user greater freedom of expression. It will enable them to communicate using natural language without requiring a high level of expertise on data-driven algorithms. User requests are interpreted and handled internally by the system in an automated manner, providing the user with the required output by masking the complexity of the data science workflow. The system can also collect feedback on the displayed results, leveraging these comments to address personalized data analysis sessions. The benefits of the envisioned system are discussed, and a use case is also presented to describe the innovative aspects.**

***Index Terms*—Data science, Data mining, Machine Learning, Data exploration, Industry 4.0, friendly data science, inclusive Data science.**

I. Introduction

The availability of data collections within organisations and almost any audience has introduced the need to provide "friendly" and inclusive ways of exploring and analysing them. The idea is that these tasks must go beyond the expertise of data scientists towards non-technical users with different expertise. That is, engineering, humanities, social sciences and any discipline requiring to experiment on data for answering research questions. Any non-technical user should exploit data collections with the right guide of an "intelligent" assistant. Friendly exploration and analysis systems must provide intuitive and interactive access to data processing operations in an agile and visual step by step manner. They should help a user drive conclusions about the data collection content and identify the potential questions that data can help answer. Through conversational loops and feedback, a friendly exploration and analysis system must calibrate the tasks according to the characteristics of the data and the expertise and expectations of the user. Through metadata collection and user profiling, an exploration and analysis conversation loop should propose actions, insight and results' display (and visualisations) that assist the user in completing a given goal.

This vision paper proposes a methodology to support the domain expert in the data analysis process, making the exploration simple and user-friendly. It showcases our vision using an industrial use case related to predictive maintenance activity. In this scenario, manufacturing machinery is constantly monitored through sensors to collect parameters characterizing the production cycle. After appropriate preprocessing steps and feature engineering, these parameters can be used to build a predictive model able to estimate the degradation of the machinery at the end of each production cycle. In a context like this, the user that interacts with the system has a technical-industrial profile, with often no expertise in the data science field. Therefore, it is crucial that he/she intuitively interacts with the system, without being involved in the complexity of the analysis required. The user may be interested in analyzing the data collected on a specific date before or after preprocessing. He/she may want to check the predicted degradation value after a particular production cycle, compare it with the predicted values for other cycles, or compare it with other predictions related to the same cycle but obtained with different predictive models. All these requests can be made through a conversational approach, giving the user freedom of expression that leads to greater user-friendliness of the system.

Accordingly, the remainder of the paper is organised as follows. Section II describes conversational based approaches that focus on some phase of data processing processes. Section

III enumerates and discusses the main challenges and research directions for tackling conversational based data exploration that provides friendly interfaces adapted to users profiles. Section IV introduces the general lines of a preliminary conversation-based data exploration process that we propose. Section V describes the first experiment results through a use case related to the exploration of manufacture data. Finally, section VI concludes the paper and discusses future work.

II. Related Work

In recent years, the use of artificial intelligence-based chatbots has become increasingly common in many areas. AI chatbots are programs that simulate human-like conversations using natural language processing (NLP). They can understand language outside of a set of pre-programmed commands and continue learning based on the inputs it receives. Instead of a traditional chatbot, which offers pre-set structured dialogues to only answer predefined questions, an AI chatbot provides greater freedom to the user without imposing form or structure in formulating the question. So, if the goal is to have a smarter bot than a traditional bot, handle complex queries or help you make sense of massive datasets, AI bots are the best choice here.

Taking advantage of the recent popularity of these applications, many of the prominent technology leaders have begun offering their solutions to allow the user to get an easily customized chatbot. Dialogflow is a natural language understanding platform developed by Google [1] that makes it easy to design and integrate a conversational user interface into a mobile app or web application. Amazon Lex is an AWS service for building conversational interfaces into applications using voice and text. It was developed by Amazon[2]. It provided the deep functionality and flexibility of natural language understanding (NLU) and automatic speech recognition (ASR) to build lifelike and conversational interactions with the user. The Facebook AI Research (FAIR) team designed ParlAI [1] to create a community-based platform in which researchers can collaborate and reuse significant dialogue tasks, encouraging collaboration in the implementation of NLP systems that includes integration with bots as well as humans. Language Understanding (LUIS) is a cloud-based conversational AI developed by Microsoft[3] to identify valuable information in dialogues. LUIS interprets user goals (intents) and distils useful information from sentences (entities) for a high-quality language model. It integrates with the Azure Bot Service, making it easy to create a sophisticated bot.

Despite the interest and promising role as friendly interfaces, existing chat-bot conversational platforms have been rarely used for guiding users through data collections exploration and analysis. There are a multitude of data exploration and business intelligence systems, e.g. Trifacta[4], Tableau[5], Qlik Sense[6], NADEEF [2], KATARA [3], Tamer [4], VADA [5] and Data Civilizer [6]. However, these systems do not adopt a conversation-led approach. Data exploration with a human in the loop perspective has emerged [7]–[10] as an initiative to include the user requirements within the process. User requirements are modelled and included in the exploration process as input parameters provided within interactive settings. However, this intent can be enhanced by recommendation like techniques in which the system suggests exploration strategies. Suggestions can be calibrated according to the profile of the user.

Under this perspective, Intuitive Data Analytics (IDA[7]), a personal virtual data assistant using Natural Language Processing (NLP), allows to request reports, play with data visualizations and foresee predictions. IDA enables users to play with data, test out "what-if" scenarios, and visualize disruptive events and their effects. IDA assists decision-makers to ask for specific content and analytics tasks. The user does not choose the strategies to be used for solving his/her questions. Besides, her expert profile is not considered for calibrating the process. Our vision and approach aim to guide users with different expertise profiles to choose the possible techniques and tools to perform a data exploration task. The suggested methods and exploration pipeline are driven by the user expertise profile and the characteristics of the data. The process is conversation-based, meaning that the user chats with a data exploration system to acquire hindsight about datasets. The features of a smart chat provide the user with the illusion of a friendly and agile way for exploring data (i.e., dividing data science tasks into short phases that are frequently assessed and adapted according to partial results.).

Some attempt to make the interaction between user and system more conversational has already been made, for example, in the conversion from natural language to query and vice versa. In such a context, stand the works done in [11]–[17] and [18], [19], which try to tackle the SQL-to-NL and NL-to-SQL problem. However, in this case, the conversational aspect between user and application is purely query-oriented. Other approaches like [20] address dialogue-based approaches to supports rich visualizations of data. Our idea is to go beyond query and answer settings and offer a complete dialogue to the user, allowing an interaction that is not exclusively finalised to the definition of a query.

III. Vision

The idea of our vision to make data science inclusive for non-data science-savvy people is that conversation and automated data mining can drive data exploration and analysis intuitively and step by step. For an automated conversational mining system, the main goal of the research is to design an engine that can effectively handle all possible requests that may come from the user. Providing greater freedom of expression to the user implies recognising the various requests

[1] https://cloud.google.com/dialogflow/docs
[2] https://docs.aws.amazon.com/lex/latest/dg/what-is.html
[3] https://docs.microsoft.com/en-us/azure/cognitive-services/luis/what-is-luis
[4] https://www.trifacta.com/
[5] https://www.tableau.com/
[6] https://www.qlik.com/us/products/qlik-sense
[7] https://www.intuitivedataanalytics.com

and managing them correctly. Furthermore, the underlying complexity of data analysis tasks must be hidden from the end-user in the knowledge extraction process.

Figure 1 shows the general architecture of our system. The user provides an input to the system that can be textual or speech. This natural language input is converted into a task to be executed by leveraging an AI-based block that can understand the meaning of the user's request. Depending on the request, a different analytic block within the Data Operations Manager is used. In addition, requests are handled by leveraging logs collected during past explorations performed by the system, looking for similarities and affinities between various analyses to show the most useful and suitable results for the user's request. These results can be shown in textual, tabular, or graphical form. The user can then evaluate the clearness and utility of the results by leaving feedback that will be added to the exploration log and will be used to guide future analysis.

The design and development of the envisioned architecture open up a wide range of research questions, such as:
1. How to design a friendly user interaction to democratize data science?
2. What are the different main types of conversations between the user and DS4ALL and vice-versa allowed by the system?
3. How to translate a user request into a DS4ALL task? What kind of machine learning algorithms could be exploited? How to extend existing NLP trained models?
4. How to structure/define data science tasks (e.g., cluster analysis, predictive task, data exploration)?
5. How should the output be displayed?
6. How to design a personalized data science task? How to leverage logs containing information about past analysis, to improve current analysis?
7. What kind of information should be collected to provide personalized data science workflow? How do we analyze it?

The first question represents an entirely new problem. To the best of our knowledge, other works that aim to make data analysis more friendly approach the problem from a different point of view. IDA[8] for example, allows immediate and straightforward actions to the user, who can interact with a meta-language. Instead, other works such as [11], and [19] focus mainly on converting natural language to SQL, or vice versa, basing interaction exclusively on queries. However, although all of these works provide methods for facilitating interaction, none of them incorporates a dialogue-based approach that promotes a human-like conversation between system and user.

For questions 2-7, extensive research brought many innovative and efficient algorithms customized for a targeted analytics task. The proposed approach can integrate existing methodologies and algorithms. Using speech recognition libraries and pre-trained NLP models like Google's Bert [21] or OpenAI's GPT [22] is it possible to build a conversational chatbot that listens and replies like a human. The data science tasks that DS4ALL can perform are those offered by the leading business intelligence systems (data profiling, data exploration), with some more specific analysis related to the industrial scenario (e.g. predictive maintenance and clustering analysis to label production cycles). The results of these analyses can be displayed using the most popular data visualization libraries such as D3[9] and Highcharts[10]. Then, an ML model can leverage the information contained in the exploration log and provide more custom results to the user.

IV. Conversational data science

This paper proposes a system that promotes an inclusive and friendly conversational data exploration and analysis process.

The originality of the proposed architecture can be summarized in the following points:

1) The **user is assisted** in the exploration and analysis of a data source.
   - *How*: by leveraging a registry of existing data exploration and analysis pipelines.
2) The **user benefits from a layer of abstraction** that insulates him from the complexity linked to the specifics of the libraries and tools for data exploration and analysis.
   - The user does not need to be an IT expert familiar with data exploration and analysis tools or how they are used.
   - The user does not need to specify the tool/method to be used for a given action. The results can, in some cases, be presented with an explanatory interpretation that facilitates the user's understanding of the concepts.
   - *Example*: The user can ask a question like: *which variables are pair-wise correlated?* To do this, s/he will not be asked which method to use (e.g., Spearman, Kendal, etc.). Moreover, besides the matrices produced by these tools, the user will receive a text that interprets the results for him, e.g., *the variables* v_1 *and* v_k *seem to be strongly correlated*. We can imagine a similar scenario for outlier detection.
3) The **interaction is conversational** between the user and the system: The system responds to the user's request, but it can also ask for the user's feedback on a given point to better meet his/her needs for the data exploration and analysis task at hand. An example of conversational interaction (see figure 2) with an initial dialogue is shown in which the user loads the dataset to be analyzed, and the system immediately provides some statistical information about it.

The proposed methodology to support the user in the analysis process of a data source can be integrated into different architectures, with scenarios of varying complexity:

[8] https://www.intuitivedataanalytics.com

[9] https://d3js.org

[10] https://www.highcharts.com

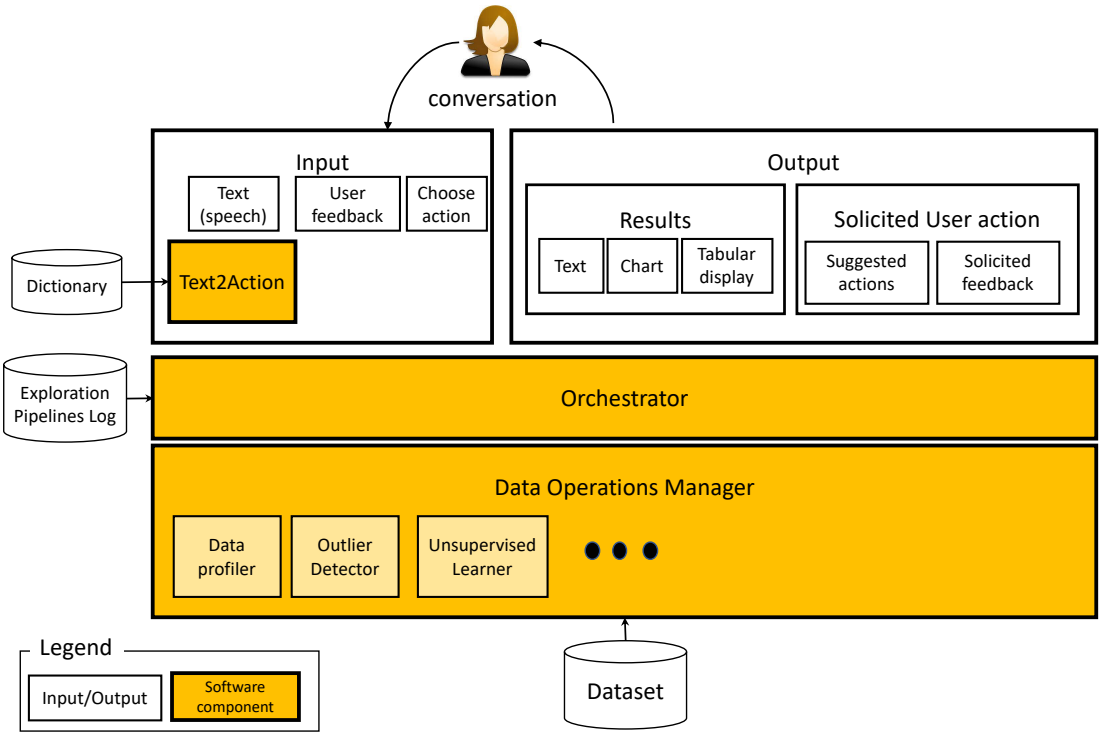

Fig. 1. Envisioned Architecture

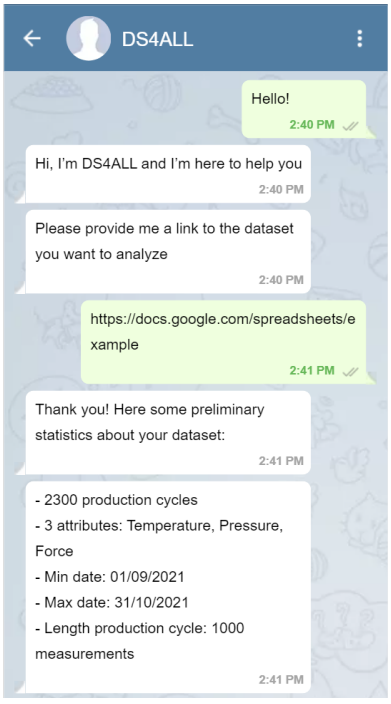

Fig. 2. Conversational example

- In the simplest scenario, the methodology is integrated into an instant messaging app such as Telegram, Slack, or Facebook Messenger. In this scenario, the user can send textual questions, which the app will answer via sentences and charts.
- In the most complex scenario, the methodology is integrated into a custom web app. Here, the user can formulate questions in oral form, which the app will appropriately convert into text before providing a written or graphical response. The user can then provide feedback on the answer received, indicating his/her level of satisfaction. Feedback can be given, for example, by selecting from the various graphics provided the one considered most relevant as a response or by leaving a textual/oral comment that the app will save. The feedback collected is then used to improve the analysis process, adapting the application based on the users' needs.

Figure 3 shows the sequence diagram of the meta-conversation between DS4ALL and the user. The conversational approach exploits a bidirectional interaction (red lines for application and blue ones for the user), starting and ending either from the user or the application. Every question posed by the user receives an answer from the application. However, also the application can ask the user to collect additional information or feedback. In addition to the opening and closing of the conversation, we initially identify five types of interactions (depicted with different lines in Figure 3, e.g. dotted, continuous, etc.). Each interaction is characterized by specific questions and answers: three of them are started by the users, and two by the application. As a first step, the user can begin interacting with the system by uploading a dataset and gathering different statistics, also displayed through plots, adequately selected based on user expertise. Then, the other

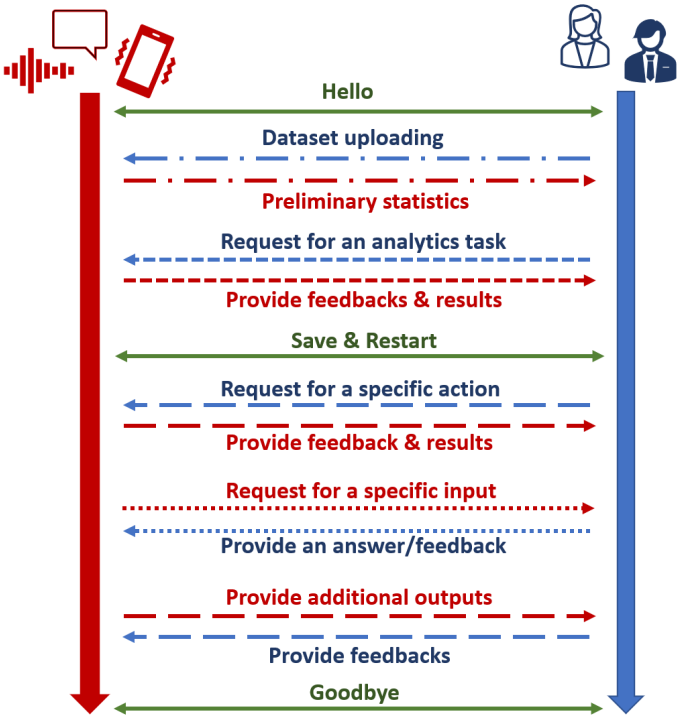

Fig. 3. Sequence diagram of meta-conversation.

four identified types of interactions are all independent of each other. There is no time constraint between them: the only time constraint is between questions/answers within the same type of interaction. The conversation can also be interrupted at any time. In this case, the user can save the current conversation status to continue later without losing the analysis results conducted up to that moment.

The user can request an analytical task (e.g. data modeling, predictive maintenance), delegating the complexity of the required analysis to the system and obtaining only the graphical results. S/he can also ask for a more specific action, such as changing the time interval considered in the analysis or excluding a particular attribute in subsequent analysis.

Regarding the interactions started by the system, they can ask for additional input parameters, which the user can provide thanks to his domain knowledge. This interaction is helpful for the application to capture domain knowledge through data analysis. Moreover, the application may also provide output to the user without being specifically requested to do so. This can be considered the most complex interaction among those proposed because the user does not require anything specific from the application. The system must provide the user with the analysis and the results it considers most suitable and of interest without receiving any initial input.

V. Preliminary development and results

This section illustrates through a use case preliminary results regarding the modelling of conversational pipelines designed to explore data.

A. Case study

Industry 4.0 settings monitor entire production chains through sensors that can constantly collect a large amount of data. This data, processed using scalable data analytics architectures, can be leveraged to optimise the machinery maintenance process. For example, a predictive maintenance strategy for a complex machine can require a sophisticated and non-trivial analytical stage to provide accurate and trusted predictions [23]. Some signals in the collected data can represent fully functioning machine conditions; others have been collected in the proximity of a failure and therefore represent critical production issues. In our Industry 4.0 scenario [23], [24] we consider time-stamped data concerning: (1) the components participating in a first phase of the industrial process: temperatures, pressures and mass flow of the two components; (2) metrics regarding oil pressure of pistons and mechanical vibration recorded from machines; (3) alarms recorded by monitoring systems. A second data source is provided by the past recording of maintenance activities, such as failures, time of failure, and recovery actions. This dataset comprises several sources, including the existing preventive maintenance dataset and additional manually compiled spreadsheets by maintenance operators.

It is necessary to have a good and thorough understanding of the data collections, to determine whether these data help answer prediction questions. The users require an exploration approach to acquire a quantitative vision of these data and their quality. Data exploration can imply performing other data science tasks with which the user might not be familiar. Next, we show how a system can guide the user through data collections in a friendly and intuitive manner through a conversational approach. The user can also adjust the exploration process giving feedback to specify whether partial results are aligned to his/her expectations.

B. Conversational exploration of manufacture data: key tasks

In the industrial domain, engineers in the field and decision-makers must exploit datasets produced from observing processes. Even if their mathematical knowledge guides them through quantitative data analysis, performing more ambitious analytics is not easy to design. We chose a use case in the industry 4.0 domain to show how DS4ALL can promote inclusiveness for engineering experts with few data science expertise. Here we describe the key tasks of conversational data science with experts in Industry 4.0 that DS4ALL provides. Within each task, the conversation approach will leverage the interaction with the user. All the conversational types shown in Figure 3 and described in Section 1 can be exploited. The conversation may include more or less numerous interactions of two-way text/audio messages for each task.

Each task requires input variables (which may be optional or mandatory) and provides a corresponding output. Table I shows the inputs/outputs for each considered task. The first task represents the *dataset upload* phase by the user: in this task, the user must provide a tabular dataset to be used for analysis. Currently, the only formats accepted are CSV or

Task	Input request	Output provided
Dataset uploading	Dataset in csv format	Preliminary statistics
Data quality verification	Domain knowledge	Feedback on quality aspects
Definition of the analytics objective	Information of interest	Analysis results
Data exploration	Optional parameters	Numerical results and plots
Pattern Mining	Optional parameters	Numerical results and plots
Cycle partitioning	Optional parameters	Numerical results and plots
Predictive maintenance	Optional parameters	Numerical results and plots
Cycle Labeling	Optional parameters	Numerical results and plots

TABLE I
INPUT/OUTPUT FOR EACH TASK

XLS, but these limitations will be overcome in the future. The upload of the dataset can be done through an appropriate section of a custom web page or by sharing a link to a data repository using Telegram or other messaging apps. This task ends with some preliminary statistics (both textual and graphical) extracted from the loaded data and shown to the user.

Two other tasks in the table are examples of situations where the app initially asks the user to provide specific input. This happens in the *data quality verification* and in the *definition of the objective*. In the first case, the system asks the user to provide domain information related to the loaded data (e.g. validity range for each variable, unit of measure, sampling rate). After receiving these additional inputs, the system provides the user with a report on the data quality, highlighting possible anomalies and asking the user for advice on how to handle them. The second task asks the user questions to collect preliminary information about the data analytics objectives. If the user does not answer, the app proposes additional questions to help the user understand his/her data and choose a specific data analytics objective.

Two other tasks in the table are examples of situations where the app initially asks the user to provide specific input. This happens in the *data quality verification* and in the *definition of the objective*. In the first task, the system asks the user to provide some domain information related to the loaded data (e.g., validity range for each variable, unit of measure, sampling rate). After receiving this additional input, it provides the user with a report on the quality of the data, highlighting possible anomalies and asking the user for advice on how to handle them. The second task asks the user some questions to collect preliminary information about the data analytics objectives. If there is no answer, the app asks additional questions to help the user understand his/her data and choose a specific data analytics objective.

C. Conversational exploration of manufacture data: an example

Here is an example of interaction between user and app. The purpose is to show how the user can request a particular task/action and display the analysis results. The task addressed in this case is the *definition of the analytics objective*. Here, the app initiates the interaction, asking the user what type of analysis s/he is interested in, providing a list of possibilities. The user can select one of these choices, and the analysis starts. However, the user may not know which option to choose, either because he is not clear about the type of analysis he wants to see or doubts about the semantic definition of the options. In this case, the "I don't know" answer is given, and the app proposes additional questions to direct the user to the most suitable analysis. A different task is addressed depending on whether these questions are answered positively or negatively. Figure 4 offers an example of a conversation between the system and the user during the definition of the analysis objective in which the user does not have a clear idea of the type of analysis he wants to carry out. Here the app offers additional questions to help the user perform a clustering analysis through his answers. It is essential to say that the same kind of analysis would be performed even if the interaction started directly from the user, who explicitly requests this specific analytic task.

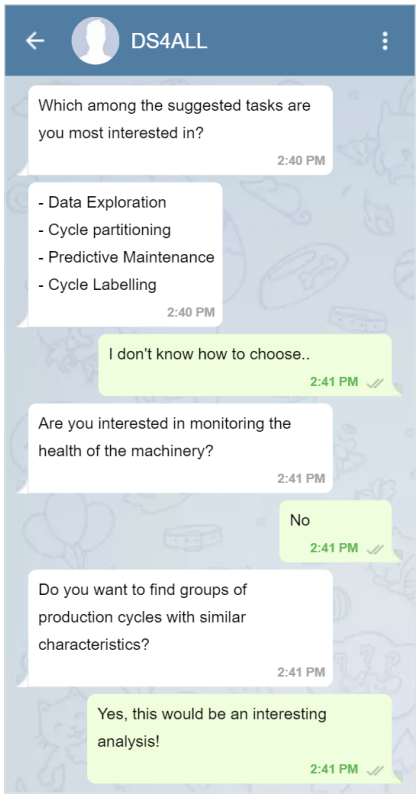

Fig. 4. Conversation about the definition of the analytics objective

Figure 5 represents the output provided to the user if the

selected task is the clustering analysis. The screenshot shows the comparison between two different approaches to perform clustering: in the first case, the distance between signals is calculated by exploiting statistical features extracted from the raw data collected by the sensor; in another case, the distance between the original raw data is used.

The two approaches are compared graphically using a 3d scatterplot showing how the various clusters are distributed. In addition, the performance of each solution is represented by three quality indices (Average Silhouette Index, Global Silhouette Index, Dunn Index). Since these indices presuppose technical knowledge that the user may not possess, each indicates the range of values that can take, specifying which values guarantee a better solution. In this way, even if the user does not know what these indices represent, s/he can still interpret their values and compare them, evaluating the solution that could be better.

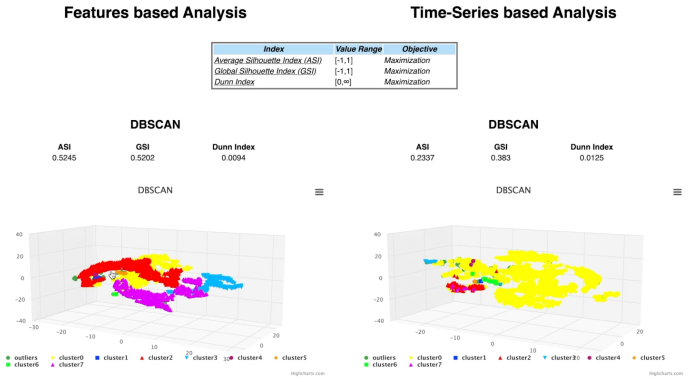

Fig. 5. Comparison between two clustering techniques

VI. Discussion and Challenges

This paper introduces our vision and research challenges related to the design of friendly conversational data exploration calibrated according to the expertise of users. The objective is to propose strategies that can democratise data science processes to non-expert users and thereby promote inclusiveness in the area. Our vision is that interactive, conversational approaches can guide users with different expertise and background through data exploration. The challenges to achieving this objective include designing agile, intuitive and friendly conversations. Conversations must be guided by the characteristics of the data and the users' profiles. They must lead to an understandable description of data and the user familiarised with the possible exploration tools used for an exploration task. We highlight below several issues that need to be addressed to make the vision we have outlined in this paper a reality.

It is by no means an exhaustive enumeration of all possible open problems, which in itself would be impossible, but rather a guide to some of the most relevant future research directions. In particular, the section focuses primarily on the conversational aspect, as our vision poses most of its innovative open problems.

- *Interpreting user requests.* This issue lies at the interface of multiple disciplines: processing user requests issued in natural language, mapping and adapting requirements (user needs) into actions that consider the nature of the data at hand.
- *Providing personalized answers.* Feedback/answers provided by the envisioned engine should become more adaptive and personalized to guide the users in friendly data exploration and in easily detecting data value. The new services should include *data storytelling* capabilities, i.e., provide interactive conversations combined with narrative techniques which deliver data stories and explanations the human user can easily understand that.
- *Prediction of the next-step analytics task* relies on guessing the next step (action) to undertake. Sometimes the user is unable to identify the best next step in the data and exploration task. The underlying challenge is to identify the characteristics of the dataset and the previous actions that could be used to foresee the next steps.
- *Personalized data visualization.* What is the best display (visualization) for a given dataset or the results of an action. The expertise of the user must be taken into account, but also the intrinsic characteristics of the data and the results.
- *Drilling down into data science tasks.* Data science tasks are typically complex and do not follow a well-specified workflow. This complexity generates the need to identify the granularities of the subtasks that make up the overall task, which is not necessarily known a priori. In addition, it is necessary to investigate new methods to navigate through (draw dependencies between) sub-tasks.

References

[1] Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*, 2017.

[2] A. Ebaid, A.d K. Elmagarmid, I. F. Ilyas, et al. NADEEF: A generalized data cleaning system. *PVLDB*, 6(12):1218–1221, 2013.

[3] X. Chu, J. Morcos, I. F. Ilyas, et al. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In *SIGMOD*, pages 1247–1261. ACM, 2015.

[4] M. Stonebraker, D. Bruckner, I. F. Ilyas, et al. Data curation at scale: The data tamer system. In *CIDR*. www.cidrdb.org, 2013.

[5] N. Konstantinou, M. Koehler, E. Abel, et al. The VADA architecture for cost-effective data wrangling. In *SIGMOD Conference*, pages 1599–1602. ACM, 2017.

[6] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. The data civilizer system. In *8th Biennial Conference on Innovative Data Systems Research, CIDR 2017, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*. www.cidrdb.org, 2017.

[7] Paolo Bethaz and Tania Cerquitelli. Enhancing the friendliness of data analytics tasks: an automated methodology. In *EDBT/ICDT Workshops*, 2021.

[8] Genoveva Vargas-Solar, Mehrdad Farokhnejad, and Javier Espinosa-Oviedo. Towards human-in-the-loop based query rewriting for exploring datasets. In *Proceedings of the Workshops of the EDBT/ICDT 2021 Joint Conference*, 2021.

[9] Alexander Kalinin, Ugur Cetintemel, and Stan Zdonik. Interactive data exploration using semantic windows. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 505–516, 2014.

[10] Alfredo Alba, Chad DeLuca, Anna Lisa Gentile, Daniel Gruhl, Linda Kato, Chris Kau, Petar Ristoski, and Steve Welch. Task oriented data exploration with human-in-the-loop. a data center migration use case. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 610–613, 2019.

[11] Stavroula Eleftherakis, Orest Gkini, and Georgia Koutrika. Let the database talk back: Natural language explanations for sql. In *SEA-Data@VLDB*, 2021.

[12] Tao Yu, Rui Zhang, He Yang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, et al. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases. *arXiv preprint arXiv:1909.05378*, 2019.

[13] Anthony Colas, Trung Bui, Franck Dernoncourt, Moumita Sinha, and Doo Soon Kim. Efficient deployment of conversational natural language interfaces over databases. *arXiv preprint arXiv:2006.00591*, 2020.

[14] George Obaido, Abejide Ade-Ibijola, and Hima Vadapalli. Talksql: A tool for the synthesis of sql queries from verbal specifications. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–10. IEEE, 2020.

[15] Jaydeep Sen, Chuan Lei, Abdul Quamar, Fatma Özcan, Vasilis Efthymiou, Ayushi Dalmia, Greg Stager, Ashish Mittal, Diptikalyan Saha, and Karthik Sankaranarayanan. Athena++ natural language querying for complex nested sql queries. *Proceedings of the VLDB Endowment*, 13(12):2747–2759, 2020.

[16] Nathaniel Weir, Andrew Crotty, Alex Galakatos, Amir Ilkhechi, Shekar Ramaswamy, Rohin Bhushan, Ugur Cetintemel, Prasetya Utama, Nadja Geisler, Benjamin Hättasch, et al. Dbpal: Weak supervision for learning a natural language interface to databases. *arXiv preprint arXiv:1909.06182*, 2019.

[17] Christoph Brandt, Nadja Geisler, and Carsten Binnig. Towards robust and transparent natural language interfaces for databases. 2020.

[18] Georgia Koutrika, Alkis Simitsis, and Yannis Ioannidis. Conversational databases: explaining structured queries to users. Technical report, Stanford InfoLab, 2009.

[19] Ursin Brunner and Kurt Stockinger. Valuenet: A natural language-to-sql system that learns from database information, 2021.

[20] Abhinav Kumar, Jillian Aurisano, Barbara Di Eugenio, Andrew Johnson, Alberto Gonzalez, and Jason Leigh. Towards a dialogue system that supports rich visualizations of data. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 304–309, 2016.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[22] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[23] Pierluigi Petrali, Tania Cerquitelli, Paolo Bethaz, Lia Morra, Nikolaos Nikolakis, Claudia De Vizia, and Enrico Macii. Estimating remaining useful life: A data-driven methodology for the white goods industry. In *Predictive Maintenance in Smart Factories*, pages 149–164. Springer, 2021.

[24] Paolo Bethaz, Xanthi Bampoula, Tania Cerquitelli, Nikolaos Nikolakis, Kosmas Alexopoulos, Enrico Macii, and Peter van Wilgen. Predictive maintenance in the production of steel bars: A data-driven approach. In *Predictive Maintenance in Smart Factories*, pages 187–205. Springer, 2021.