MATILDA: Inclusive Data Science Pipelines Design through Computational Creativity

Genoveva Vargas-Solar[1], Khalid Belhajjame[2], Javier A. Espinosa-Oviedo[1,3], Santiago Negrete-Yankelevich[4] and José-Luis Zechinelli-Martini[5]

[1]CNRS, Univ Lyon, INSA Lyon, UCBL, LIRIS, UMR5205, F-69221, France
[2]PSL, Université Paris Dauphine, LAMSADE, UMR7243, France
[3]CPE Lyon, 43 Blvd. du 11 Novembre 1918, 69616 Villeurbanne Cedex, France
[4]Universidad Autónoma Metropolitana (Cuajimalpa). Avenida Vasco de Quiroga 4871, Cuajimalpa de Morelos 05348, Ciudad de México
[5]Fundación Universidad de las Américas-Puebla, Exhacienda Sta. Catarina Mártir s/n 72820 San Andrés Cholula, Mexico

Abstract
This paper argues for developing innovative data science frameworks that render the latest progressions in data engineering and artificial intelligence accessible to non-technical users across diverse fields. Such frameworks would empower these users to leverage advanced data science solutions' capabilities fully. We propose a methodology that merges computational creativity with conversational computing to facilitate an intuitive pathway for non-experts to navigate and derive insights from datasets. We present MATILDA, a platform rooted in creativity-driven data science, and demonstrate its utility in augmenting the data science pipeline's design process through the synergy of human innovation and algorithmic ingenuity.

Keywords
Data science pipelines, graph analytics, knowledge graphs, computational creativity

1. Introduction

Harnessing extensive datasets across numerous sectors via data science techniques offers substantial economic and social benefits. These methods are predominantly within the purview of individuals proficient in AI, with deep knowledge of mathematics, statistics, numerical analysis, and artificial intelligence frameworks. However, there is a growing need for these data science methodologies to be accessible and applicable to inquiries and challenges from a broader range of disciplines, catering to users who may not possess extensive expertise in data science. These methodologies must extend beyond the data science community to embrace users from non-technical backgrounds—such as engineering, humanities, and social sciences—who rely on data analysis to address research questions pertinent to their domains.

This shift necessitates the emergence of novel data science solutions that capitalise on and democratise the latest advancements in data engineering and AI. These solutions should shield non-technical users from the intricacies of the underlying technologies while still allowing them to fully exploit the capabilities of these advanced tools to satisfy the specific analytical requirements of their respective fields.

Addressing complex transdisciplinary issues requires a collaborative effort where experts from diverse fields engage in dialogue and exchange ideas, a process that inherently demands creativity[1] Creating data science solutions calls for a multifaceted approach, incorporating algorithmic, data-centred, information technology, and cross-disciplinary perspectives, particularly from those outside the data science domain. Consequently, there is a pressing need for innovative solutions underpinned by Creativity that can make data science accessible to non-experts, allowing them to intuitively navigate data repositories and distil valuable knowledge.

In this paper, we explore an approach that melds computational creativity, enabling users to venture into new realms of data analysis design with conversational computing, which offers user-friendly abstractions for directing and customising their data analysis endeavours without the necessity of engaging with intricate technical specifics.

Accordingly, the remainder of the paper is organised as follows. Section 2 gives a general overview of ap-

Published in the Proceedings of the Workshops of the EDBT/ICDT 2024 Joint Conference (March 25-28, 2024), Paestum, Italy.

*Genoveva Vargas-Solar.
†The authors' list is alphabetical except for the first author.
genoveva.vargas-solar@cnrs.fr (G. Vargas-Solar);
khalid.belhajjame@dauphine.fr (K. Belhajjame);
javiera.espinosa@liris.cnrs.fr (. J. A. Espinosa-Oviedo);
snegrete@cua.uam.mx (S. Negrete-Yankelevich);
joseluis.zechinelli@udlap.mx (J. Zechinelli-Martini)

[1] Creativity is a process that can combine familiar ideas in new ways, explore the potential within existing conceptual spaces, or transform these spaces to allow for previously inconceivable ideas. Creativity is not considered novelty but the capacity to generate surprising and valuable ideas that push beyond conventional boundaries [1].

proaches contributing to model creative-based processes and friendly design data science pipelines. It discusses how both areas can provide novel ways of designing data science-driven solutions. Section 3 after that describes the challenges of modelling creative-driven data science design processes. It gives the general lines of an approach that can enhance and envision a new way of addressing analytics problems using data and artificial intelligence models. Section 4 introduces a creativity-based data science design platform. It describes the general architecture and functions and shows how it can support the design process of data science pipelines guided by human and computational creativity. Finally, Section 5 concludes the paper and discusses future work.

2. Related work

Addressing the creative design of data science pipelines to make them inclusive for non-experts requires drawing from methods and results from three areas: creative models, friendly data science and provenance. This section gives an overview of the relevant results that provide a scientific background to the project and help to contextualise our objectives.

Creativity-driven systems Artificial intelligence (AI) offers opportunities to reify and transform how we think about human cognitive capabilities. In this context, computational creativity (CC)[2], aims at studying human creativity and building systems that perform in such a way as to be considered creative. Concerning creativity, whether this cognitive capability is individual or collective. CC has moved from the classic individualistic and cognitive model of creativity [1] to a social and collective creativity model [2, 3, 4, 5, 6, 7].

Collective-creativity models are concerned with understanding the roles and tasks that different agents (both human and non-human) play in a process that requires creativity and how creativity can be measured in this context [8, 9]. Co-creativity is an excellent approach to establishing efficient, context-aware interactive systems and setting long-term programmes where solutions to problems requiring human and machine collaboration can be studied.

CC has been widely applied in art with systems that promote "artificial" creation. For example, Disco Diffusion is an online tool that runs on Google Collab to execute Python programs that, using a learning model, result in "creative" artwork. Language models such as GPT-3 are capable of interpreting and generating text. With over 540 billion parameters, the Google Pathways model can explain jokes, follow a chain of reasoning, recognise patterns, perform Q&A sessions on scientific knowledge, and summarise texts. As more parameters are added to language models, the depth of "understanding" they can demonstrate expands. "The painting fool" is a system developed by Simon Colton that draws portraits taking into account emotional information obtained from the subjects being painted through a camera [10]. Negrete-Yankelevich and Morales-Zaragoza [4] propose a model to develop and assess creativity in computational agents embedded in mixed teams. The Apprentice Framework model establishes a series of roles (or levels of responsibility) agents can play within the group over time with the possibility of ascent through the ladder as the system is developed, acquiring thus more responsibility in the creative process. The model also helps identify aspects of the product being produced, at which point the agent is supposed to be creative. By keeping track of both responsibilities and aspects, it is possible to plan and assess the development of the system. Negrete-Yankelevich and Morales-Zaragoza [11] propose a framework to create animatics. These animated storyboards constitute an essential artefact in producing animations by a team of expert animators that made an award-winning series of one-minute shorts for Mexican TV called Imaginantes (Televisa, "Imaginantes* - YouTube."). The system's creativity is measured by how well the overall creativity of the team is affected by the system's performance.

Developing friendly data science solutions. Friendly data science systems must provide intuitive and interactive access to data processing operations in an agile and visual step-by-step manner [12]. They should help a user to derive 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 data's characteristics and the user's expertise and expectations. 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.

Discussion. We believe the collective CC can reproduce the collaborative transdisciplinary conditions in which data science solutions are developed. It can drive the proposal of data science solutions combining data, algorithms, and computing resources to model complex systems and contribute to answering research questions to understand and predict them. While computational creativity and conversational techniques have proven

[2] Computational creativity is the study of building software that exhibits behaviour that would be deemed creative in humans. Such creative software can be used for autonomous creative tasks, such as inventing mathematical theories, writing poems, painting pictures, and composing music (https://computationalcreativity.net

effective, they have not been explored with the design of exploratory data analysis pipelines. Moreover, the two approaches are somewhat opposed because conversational techniques tend to rely on known territories (i.e. previously explored data manipulation and analysis actions). In contrast, computational creativity allows for exploring unknown territories (data manipulation and analysis), which may, in some cases, prove more effective.

Our work addresses two challenges. On the one hand, adapt and leverage both techniques to design an efficient and exploratory data analysis pipeline. On the other hand, strike the right balance when creating data analysis pipelines between 'known' prior data exploration and analysis actions and 'unknown' creative actions.

3. Creative process for designing data science driven solutions

Data science pipelines combining machine learning and deep learning are the new query types with specific needs regarding how data must be structured and managed. The "one all-fits-all" data structure and associated management functions approach are no longer adapted for data science queries. Indeed, every query has a specific objective (modelling, prediction), and its design entirely depends on the input dataset and an initial research question (RQ). The data science query is not based on an explicit knowledge of the data. It includes tasks devoted to mathematically understanding the data; then, the partial results of those tasks determine the design of other studies devoted to the computation of a model representing some hidden knowledge. Given statistical and machine learning methods and a target objective, data scientists rely on libraries that provide methods that they combine to define a data science pipeline. The results obtained by this pipeline are never definite, and they are always, to some degree, close to the target.

To illustrate the design process of a DS pipeline, consider the following scenario. Consider a trendy decision-making group willing to adopt a data-driven approach for designing public policies to enhance citizens' lives in urban spaces and reduce energy and economic costs. Public policies are intended to modify built environments to improve them from financial and well-being perspectives. Decision-makers know that from the urbanism perspective, small changes in the built environment can alter how people use the space. For instance, increasing pedestrian areas in a city downtown close to restaurant zones reduces CO2 footprint. Still, it impacts the influx of restaurant customers in the area and lowers real estate prices. Customers can suddenly start preferring restaurants close to parking slots. People living in the area can have problems accessing it and park their cars close to home. The research question is *to which extent public policies can impact the quality of life of different categories of citizens willing to evolve in a given urban area?*. Decision makers now call for data scientists' creativity to provide studies and mathematical evidence of the kind of urban changes to be considered in public policies.

Designing a data science pipeline. Datasets and research questions drive the design of DS pipelines. Through a simplified creative scenario using the main phases of a DS pipeline: (1) collect or search for datasets that can be used for answering a research question, and then (2) prepare them (explore, clean, engineer) to feed one or several Artificial Intelligence (AI) models. (3) Models are trained and tested with dataset fragments. These tasks are calibrated recurrently until specific performance scores are reached. (4) Results are constantly assessed and eventually considered good enough to be interpreted by experts, and conclusions are drawn on answering the initial research question to some extent.

Sketching a creative process, the elements to consider are: What data is needed to answer the research question and develop a strategy to collect them? How do we transform the initial research question into a quantitative statement that can be addressed by mathematical or AI models? Which model families can be pertinent for answering the question? How do you design a series of tasks where data are processed? How do you connect the results format with the research question statement? How do we determine whether results converge? How do you decide whether results are fair enough for considering an answer?

For example, data scientists can film civilians in the target urban spaces to collect their behavioural patterns on how they occupy and evolve along those spaces before and after implementing public spaces. Extract behavioural patterns that imply designing a DS pipeline for processing videos and detecting civilians, for example, using perceptrons [13] and behaviour patterns within a series of scenes. The patterns can then be classified according to properties that detect changes before and after implementing some change. Other possibilities would be to run other data collection techniques like questionnaires to describe urban civilians' behaviour through quantitative variables that can be correlated for detecting changes produced after applying public policies. The possibilities are numerous, and they rely on data scientists' expertise, on the facilities or not for collecting certain types of data (e.g., video vs questionnaires) and their knowledge of specific AI models' families.

Discussion Data Science and Machine Learning Environments provide all the necessary AI models. They are supported by enactment stacks that deal with the storage, fragmentation, indexing and distribution of the data required and produced by the tasks composing a pipeline. What are the rules and strategies to combine different components that can transform input data into models and predictions that provide quantitative elements to answer initial research questions?

Generative artificial intelligence [3] has started to be consolidated into solutions that give the illusion of creation through interactive and conversational approaches [4]. Systems like chat-GPT, in its various versions, mimic conversational and question-answering experiences intended to perform target tasks or produce “new" content based on existing evidence. The principle of this system is synthesising the creation process as an exercise of wrapping together “content" with specific characteristics and considering some constraints to produce artefacts that look, to some extent, novel.

In the case of DS pipelines, the first challenge is to model the creation process behind them. How does someone (a domain expert) state a research question so that a data-driven quantitative study can be run? How are data collected and selected to answer such questions? Which comes first, data or questions? How do we conclude that given datasets representing observations of an object of study are representative enough to produce a model or predict the behaviour of that object? How is the human integrated into the loop and intervene in the design milestones of a DS pipeline?

Challenges and Open Issues. A computational-creativity-based methodology for designing DS pipelines should consider at least the following scientific challenges and associated open problems:

- Modelling hybrid (human and nonhuman) creativity-driven data science pipelines' design: propose a computational creativity model to represent end-to-end pipeline design. The creativity model can integrate design patterns like the ones presented in [14] (design, mutant shopping, chorus line, simulation and approximating feedback, entertaining evaluations and no blank canvas). Depending on the tasks to be designed within a DS pipeline, different creativity patterns can best be adapted to address the task.
- Define the interaction among humans designing a DS pipeline and artificial system(s) that can take on tasks and propose results. Model the input/output required to feed and expect to/from the system and the type of feedback to be given by humans.
  Intervene the process with an agent by selecting a relevant subprocess where creativity would contribute significantly to the overall solution and assess how it works. Then, try other similar subprocesses and verify again. This bottom-up approach would establish a practice to turn the overall process into a friendly one in a stepwise fashion.
- Collecting provenance and data from DS pipelines design tasks: implement processes for data curation, annotation, identification, and quality control in research.
- Proposing an ad-hoc computational creativity tool for making DS science pipelines design-friendly for non-data scientists.

4. Towards a Human in the loop creative platform for designing data science pipelines

Figure 1 shows the general architecture of the MATILDA platform that assists people with different expertise to follow a creative process for designing DS pipelines given datasets and target research questions. The platform relies on a step-by-step conversational approach based on our previous work [12] and provides interaction entry points to allow humans feedback, validate and guide the creative process. For each phase of a DS pipeline (data exploration and preparation, fragmentation, training, testing and assessing), the platform suggests possible scenarios that are adopted or not. Therefore the platform relies on a knowledge base representing data science pipelines, with research questions and data features modelled that can be used to propose solutions similar as case based reasoning approaches.

1. Data search: given keywords about the topic or a sample of data to be analysed, the platform relies on queries as answers and exploration techniques to propose related data sets. The platform shows the possible questions associated with data through "queries as answers" techniques. Through an interactive process, a data scientist can converge to a sample of data representative of the type of questions she/he wishes to express (e.g., factual, modelling, prediction, etc.).
2. Designing data exploration and cleaning pipeline: given a dataset, the platform performs a quantitative analysis of the attributes, their dependencies and their values' distribution. The platform also suggests cleaning and data engineering strategies, allowing data to have specific mathematical properties. The platform gathers information about

[3] According to the Bing chat-GPT and validated by this paper's authors: Generative AI refers to a category of artificial intelligence (AI) algorithms that generate new outputs based on the data they have been trained on (www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/).

[4] Microsoft Deepspeed https://github.com/microsoft/DeepSpeed/tree/master/blogs/deepspeed-chat

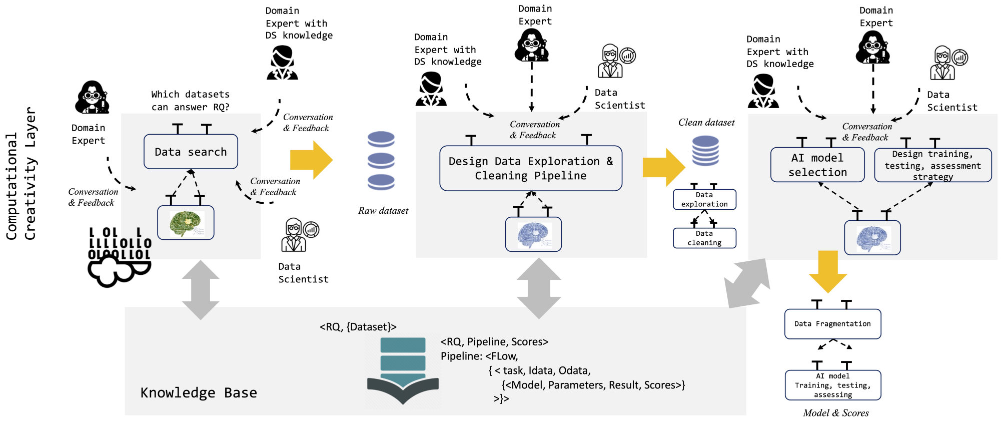

Figure 1: Matilda platform creation pipeline.

their decisions by interacting with the data scientists. This information can be used to keep track of the design process. For now, this is a very quantitative perspective of the creation process, even if, for future work, we will try to approach creativity with other perspectives.

3. DS pipeline creation: the current platform does not rely on existing AI model recommendation systems but on knowledge about the questions previously addressed with AI models; it proposes building blocks that can be combined into pipelines. These building blocks could be used to answer the questions produced in 1). The building blocks include suggestions on the scores that can be used for assessing and calibrating training phases. The platform is also shared for every building block with similar solution contexts in which they have been used.

Our DS creativity platform allows us to study how overall creativity is affected if computer systems take over different roles within the design of data science pipelines. The platform provides a collaborative environment that integrates an artificial actor within in the creative production process of DS pipelines by data scientists.

5. Conclusions and Future Work

The research and development market associated with data science is fuelling the economies of countries in the world. Almost all sectors in the global economies see data science as a promising alternative to develop original solutions to critical societal problems and promote data-driven decision-making processes that can create know-how and value. Yet, despite the availability of datasets and technology for transforming any phenomenon produced in reality into digital data and the variety of algorithms (Mathematical and artificial intelligence models), the design of data science solutions remains artisanal. The impact on person-hours and economic investment is not anecdotic. The time has come to propose methodologies that can formalise the design of data science solutions and model the "know-how" developed by data scientists during the creation process. Besides, data science addresses trans-disciplinary challenges. It is critical to bridge the gap between technical vocabulary, tasks, and the vocabulary of other disciplines and users with different expertise. This strategy will ensure the usability and acceptability of solutions (i.e. pipelines). In summary, data science must become inclusive and accessible to all. Our work addresses this challenge by aiming to adopt computational creativity methods to model the data science design process(es) that combine human and nonhuman creativity.

The platform MATILDA proposed in this paper is based on the original methodologies that we propose. It contributes to creating data science pipelines according to the expectations of knowledge discovery. It is interesting for answering target research questions, the input data's characteristics and the data scientists' models. Creativity-based methodologies applied to data science will make it accessible and inclusive to address increasingly complex problems humanity faces.

6. Acknowledgements

The work reported in this paper is performed in the context of the project FRIENDLY [5] funded by the inter-group program of the laboratory LIRIS, Lyon.

[5] http://vargas-solar.com/friendly/

References

[1] M. A. Boden, et al., The creative mind: Myths and mechanisms, Psychology Press, 2004.

[2] M. L. Maher, Evaluating creativity in humans, computers, and collectively intelligent systems, in: Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design, 2010, pp. 22–28.

[3] M. L. Maher, Computational and collective creativity: Who's being creative?, in: ICCC, 2012, pp. 67–71.

[4] S. Negrete-Yankelevich, N. Morales-Zaragoza, The apprentice framework: planning and assessing creativity, Proceedings of the Fifth International Conference on Computational Creativity, 2014.

[5] R. K. Sawyer, Explaining creativity: The science of human innovation, Oxford university press, 2011.

[6] A. K. Goel, A. G. de Silva Garza, Special issue on artificial intelligence in design, Journal of Computing and Information Science in Engineering 10 (2010).

[7] N. Gu, P. Amini Behbahani, A critical review of computational creativity in built environment design, Buildings 11 (2021) 29.

[8] A. A. Kantosalo, J. M. Toivanen, H. T. T. Toivonen, Interaction evaluation for human-computer co-creativity: A case study, in: Proceedings of the sixth international conference on computational creativity, Brigham Young University, 2015.

[9] A. Kantosalo, S. Riihiaho, Experience evaluations for human–computer co-creative processes–planning and conducting an evaluation in practice, Connection Science 31 (2019) 60–81.

[10] S. Colton, J. W. Charnley, A. Pease, Computational creativity theory: The face and idea descriptive models., in: ICCC, Mexico City, 2011, pp. 90–95.

[11] S. Negrete-Yankelevich, N. Morales-Zaragoza, e-motion: a system for the development of creative animatics, Proceedings of the Fourth International Conference on Computational Creativity, 2013.

[12] P. Bethaz, K. Belhajjame, G. Vargas-Solar, T. Cerquitelli, Ds4all: All you need for democratizing data exploration and analysis, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 4235–4242.

[13] E. Cruz-Esquivel, Z. J. Guzman-Zavaleta, An examination on autoencoder designs for anomaly detection in video surveillance, IEEE Access 10 (2022) 6208–6217.

[14] P. Glines, I. Griffith, P. M. Bodily, Software design patterns of computational creativity: A systematic mapping study., in: ICCC, 2021, pp. 218–221.