https://doi.org/10.1109/CEEC.2018.8674234

A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects

Bakhtiyar Ahmed
Digital Media for Healthcare
Kingston University London
London, UK
b.ahmed@kingston.ac.uk

Thomas Dannhauser
Mental Health Sciences Unit
University College London
London, UK
t.dannhauser@ucl.ac.uk

Nada Philip
Digital Media for Healthcare
Kingston University London
London, UK
n.philip@kingston.ac.uk

***Abstract*—As data projects become more conventional, increase in the use of information has surpassed the knowledge of how to support individuals/teams that undertake such projects. Leading data mining methodology, CRISP-DM has become limited in managing the requirements of working with recent technologies such as Machine Learning. Resultantly, many have either created their own methods or adopted alternative approaches such as the Design Thinking and Lean Startup innovation strategies. Consequently, this paper proposes a novel software development methodology entitled Lean Design Thinking Methodology (LDTM) to guide the development of modern data projects. LDTM combines the strengths of CRISP-DM with the more innovative Design Thinking and Lean Startup strategies to introduce an approach divided into three stages, comprising of seven steps. This paper concludes on how there is no one correct method, nor is one single approach enough, but together, elements of each approach can unite to help guide data projects forward.**

***Index Terms*—Data Mining, Framework, Life Cycle, Machine Learning, Methodology, Software Development**

I. Introduction

A Software Development Methodology (SDM) refers to the framework that is applied to improve the management and control of a Software Engineering or Information System process [1]. Over the years, many methodologies have emerged and gone out of fashion. Today, several different frameworks exist, each with its own set of advantages and disadvantages and each best suited to specific kinds of projects. As data continues to be produced in massive amounts, Big Data, Data Science and Data Analytics projects are growing in frequency and importance. However, the growth in the use of information has outstripped the knowledge of how to assist development teams that take on these projects [2]. While much literature is available on the use of algorithms/models that help produce insightful analysis, much less is available on the methodologies/frameworks and processes that could allow teams to complete such projects more resourcefully and successfully [2].

Despite being the de facto SDM for data mining/predictive analytics projects, the CRoss-Industry Standard Process for Data Mining (CRISP-DM) has in recent times failed to meet the challenges of working with present-day technologies such as Machine Learning [3]. Resultantly, many practitioners have resorted to creating their own methodologies, which they believe will provide a suitable framework for the implementation of data projects [3]. However, as these methodologies are recent concepts, very little work is available on the effectiveness and impact of these approaches. Moreover, many practitioners have also recently shifted from using traditional SDMs onto realising solutions using 'innovation strategies' such as Design Thinking and Lean Startup [4]. Consequently, the main aim of this paper is on how best to combine the well-established CRISP-DM methodology with the more recent Design Thinking and Lean Startup strategies to conceive a novel SDM that will guide the development of modern data projects.

Despite their recent popularity, neither Design Thinking nor Lean Startup clearly define what skills are needed to arrive at innovation through design and implementation. Thus, leaving a knowledge gap, which this paper aims to address by answering the following research questions:

1. What attributes of Design Thinking, Lean Startup, and CRISP-DM can we use to support the design and implementation process of Machine Learning and other data projects?
2. How can we effectively combine Design Thinking and Lean Startup with CRISP-DM to help development teams incorporate user-driven innovations into the development of modern data projects?

By answering these questions, the goal is to introduce a novel SDM for data projects entitled Lean Design Thinking Methodology (LDTM). A methodology that arrives at ideas for possible solutions using Design Thinking, realises the development and testing of algorithms/models using Lean Startup and builds upon the fundamentals and lessons learnt from CRISP-DM.

II. State of the Art

To identify relevant studies, The ACM Guide to Computing Literature, IEEE Explorer, and CiteSeerX bibliographic databases were examined in the search for relevant academic published articles. The inclusion criteria used the keyword "machine learning" and individually combined this with the keywords "methodology", "framework" and variations of "life cycle" including "life-cycle" and "lifecycle". The search results were limited by identifying the search terms in paper

titles, abstract and keywords. In relation to the exclusion criteria, studies that were not in the English language, not between 01/01/2010 to 31/12/2017 along with studies that were not considered as conference proceedings or journal articles were all excluded. From the 4,506 citations screened, 216 papers met the inclusion criteria. However, after reviewing the abstracts of these papers, this analysis revealed that not one of these papers was written specifically on a methodology/framework that supports the design and implementation process of data projects.

Despite there being no single paper within the inclusion criteria focusing specifically upon on SDMs, it is worth noting that a small number of papers did acknowledge and briefly discuss a few key methodological elements that the authors applied within their studies, albeit not in detail. Nevertheless, as the literature abstraction provided no significant results, it was deemed necessary to perform a wider search for SDMs related to data projects. Accordingly, a Google Search was performed using the above inclusion criteria where the first 100 results of each keyword combination were examined. The results of this Google Search produced a very large number of results, again in relation to the machine learning algorithms/models used but only went on to identify two development methodologies that were relevant to this study, IBM's Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM) and Microsoft's Team Data Science Process (TDSP).

ASUM-DM is a new implementation method for analytics projects released in 2015 by IBM. It is a systematic guide to conducting a complete implementation life cycle for analytics solutions [5]. TDSP, in contrast, provides a life cycle to structure the development of a Data Science project. Microsoft claims that development teams should find it relatively easy to map steps from their own processes to the TDSP [6]. Both methodologies encouragingly contain best practices and structures from industry. However, the level of detail both IBM and Microsoft have put into their methodologies can be quite comprehensive and this could inadvertently alienate developers from adopting them, opting in for something more straightforward [7]. This further supports the need for a simple SDM, which not only builds upon established and proven models but also takes into consideration crucial industry practices, which are essential for the success of any modern data project.

III. Conceptual Framework

To reiterate, LDTM works by combining:

- Design Thinking (to understand the customer/user and discover the business need), with
- Lean Startup (to evolve the model/solution), and
- CRISP-DM (to develop the algorithmic/technical elements of the model/solution).

By bringing together these three approaches, the intention is to allow development teams the flexibility and opportunity to routinely improve a data model by iteratively and incrementally acting upon accuracy statistics and user feedback. LDTM can, therefore, be used to advance the design and implementation of new and existing data projects by uniting the advantages of experimentation and iterative working along with a greater understanding of the customer/user requirements into one approach. This section provides an overview of these three approaches, and on the key attributes of each, which plays a role in the LDTM.

Design Thinking is an approach to problem-solving that results in relevant solutions through ideation [8]. The principle here is that development teams should always start by building an understanding of the people that they are building a solution for [9]. Techniques such as ethnographies are used to gain insights into human behaviour, to enable the development team to come to a clear understanding of who the user is and what their needs are. After empathising with the user, developers define and prioritise the users' most imperative problems (using techniques such as MoSCoW prioritisation) and consequently come up with relevant ideas to solve them. Ideas are then converted into prototypes before being tested and evaluated. The goal here for the LDTM is to benefit from and utilise the practices associated with the 'Empathize', 'Define' and 'Ideate' phases of Design Thinking during the earlier stages of the LDTM approach.

Lean Startup coined by Eric Ries is best defined as a blueprint for how to run a start-up [10]. The aim is to find a product-market fit by moving a 'Minimum Viable Product' (MVP) through the 'Build-Measure-Learn' feedback loop [10]. When building solutions, development teams start by building a minimum set of features that satisfy early users. They then test the hypotheses made about these features early on and measure the information/feedback they obtain from experiments. Thus, evidence-based decisions can be made regarding the direction the solution should be progressing into for subsequent iterations. This process is repeated in continuous loops until the product-market fit is achieved. Like Lean Startup, the goal here for the LDTM is to reach a position where the results of all the experiments prove that the development team have built a solution that accurately addresses the needs/demands of the customer/user.

CRISP-DM is an open standard, developed by a consortium of over 200 interested organisations, with funding from the European Union (EU) [11]. CRISP-DM is considered a comprehensive data mining methodology and process model that provides a complete blueprint for conducting a data mining project [11]. CRISP-DM breaks down the life cycle of a data mining project into six phases, their respective tasks, and the relationships between these tasks [11]. Due to its limitations, efforts to update the model started in the late 2000s. However, to date, no updates have been presented [12]. The original model is no longer actively maintained and at the time of writing, the official CRISP-DM.org website is also no longer being maintained. The goal of the LDTM here is to use the lessons learnt from CRISP-DM and use this as a foundation to conceive a methodology that can produce better, quicker results

By extending the attributes of Product Thinking's product definition, Table I compares some of the main distinctions between each approach.

TABLE I
DISTINCTIONS BETWEEN
DESIGN THINKING, LEAN STARTUP AND CRISP-DM

	Design Thinking	Lean Startup	CRISP-DM
Purpose	Create innovative ideas and solve ill-defined problems	Create new businesses and products under uncertain conditions	Create a reliable and repeatable process for delivering value
Problem	Undefined, investigated by the team	Unclear, hypothesised by the team	Defined by the client
Solution	Unknown, based on the team's need	Unknown, based on the team's experiments	Defined by the client
Target Audience	Designers	Entrepreneurs	Engineers
Vision	Being wonderful and providing innovative experiences	Solving an early adopters problem	Delivering what the customer wants
Strategy	Engaging with people (user-centred)	Making learning sustainable and efficient (customer-oriented)	Making software development lightweight (valued-based prioritisation)
Goals	Defined concept for products or services	Working product with simplified features	Working software with finished features
Features	Customer interviews, empathy mapping, low-fit prototypes	Experiments, build-measure-learn feedback loop, innovation accounting	Data analysis and feature selection, prediction tasks, evaluation measures
Measure of Success	Finding potential solutions from exploring the problem	Getting the right solution from continuous testing	Building the solution correctly

IV. RESULTS

LDTM (Fig. 1) is divided into three distinct stages: Business, Data and Product, which further comprises of seven steps (listed and elaborated upon below). The Business stage refers to the creative strategies used in Design Thinking, which developers can use during the process of identifying problems and proposing solutions for these problems. The Data stage is the most vital stage of any data-related project and refers to the understanding and preparation of data that will be passed onto the Product stage. The Product stage refers to how the MVP, like in Lean Startup, is rapidly tested with the customers/users to gain their feedback so to learn and iterate towards the perfect solution. Overall, learning from the lessons of CRISP-DM, the LDTM is a way of working, characterised by ongoing reassessment and adaptation of the business plans along with frequent and incremental delivery of the final solution (algorithm/model).

The seven steps of the LDTM are as follows:

1. Work Discovery: The development team get to know the customer/user and understand their problems/requirements better. The development team then define the problem, the project aims and objectives, along with the functional and non-functional requirements from the business, user, and system perspectives.
2. Analytical Approach: The development team then articulates the work discovery in the context of statistical/machine learning techniques. This allows the development team to identify the most appropriate methods that they would like to implement for the desired output e.g. classification, regression, etc.
3. Data Resources: The selected analytical approach will have an influence on the data requirements i.e. data content, formats, representations, etc. Consequently, the development team need to identify and collect all the relevant data resources associated with the problem domain.
4. Data Preparation: The development team subsequently construct the data set i.e. perform data cleansing, combine data from multiple sources, transform data into more useful variables, etc. They also use descriptive statistics and visualisation techniques to understand the data content.
5. Build MVP: The development team's goal here is to use this data set and build a Minimum Viable Product (MVP). The MVP is the simplest version of the intended data solution, which can be implemented quickly and tested by the customers/users.
6. Measure Value: This MVP is then implemented to understand what works and what does not. Feedback is collected, focusing on a few reliable and valid metrics that allow the development team to take an objective approach to measuring the solution.
7. Learn & Update: By implementing quickly and collecting valid feedback from customers/users early and often, development teams can learn from what they have implemented, make informed decisions (if necessary) on what needs to be changed and make these changes accordingly.

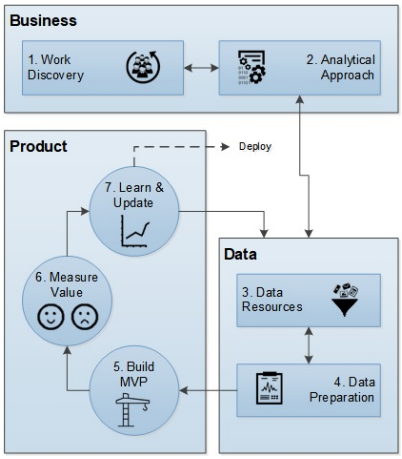

Fig. 1. Lean Design Thinking Methodology (LDTM)

The final step then feedbacks into the Data stage where the development team can collect more and/or amend existing data to improve the outcome of the intended solution. This process is repeated until the final solution is deployed to the customer/user.

V. DISCUSSION

A significant feature of both Design Thinking and Lean Startup is how they both guide a prototype through a 'test-and-learn' cycle, assuming that applying such a practice is the best way to return knowledge about the customer/user. Here, the Lean Startup delivers benefit over Design Thinking by offering

development teams the opportunity to come up with actionable metrics that unite specific and repeatable actions to observed results. Design Thinking, instead, starts by understanding the customers/users, outlining their most critical concerns, and based on this, comes up with a range of appropriate ideas. All these steps are taken before considering the development of a prototype, thus lessening the dangers associated with bringing new solutions to existence by exploring the customer's/user's interests before investing in the development stage. In this respect, Design Thinking is more effective during the earlier stages of the LDTM and Lean Startup during the latter stages.

As mentioned above, during the development stage, LDTM gives more focus on Lean Startup over Design Thinking. Despite Design Thinking having similar principles in its 'Prototype' and 'Test' phases, Lean Startup's 'Build-Measure-Learn' phases offer a more comprehensive approach to the development of a solution in regards to the collecting of customer/user feedback and learning from this. LDTM generates feedback on the prototype model in two ways, the first is direct observation and discussion with the customers/users, and the second is by accuracy statistics gathered from the machine learning algorithm itself. Using these two forms of feedback, the development team can determine whether they should continue in the same direction or rethink the core idea behind the solution, change the dataset, or update/create a new solution. Several iterations of the solution are developed, with features added/removed or amended each time based on the feedback, until a final solution is achieved.

CRISP-DM's involvement in LDTM is through its strengths in focusing upon the technical elements of a solution, which makes CRISP-DM a great methodology for delivering solutions that are often technically excellent. However, this does not necessarily mean that the solution in question delivers significant value to its customers/users, thus highlighting one of CRISP-DM's major concerns being lack of clarity of the business problem. Therefore, exploring opportunities that could lead to innovative solutions that could maximise value to the customer/user is not a common procedure within CRISP-DM. Henceforth, to advance modern data projects it is necessary to update/incorporate the strengths of CRISP-DM with the elements of innovative working found in the Design Thinking and Lean Startup approaches. By doing so, results in a methodology that not only gives importance to the technical aspects of a project but also gives equal importance to the customer/user and their business.

VI. Conclusion

A methodology/framework combining the strengths of Design Thinking, Lean Startup and CRISP-DM does not exist and its existence could, therefore, be beneficial for the development of innovative, modern data solutions. The motivations behind these methodologies all have different origins, helping development teams to think in different ways and to solve different problems. This paper explored how these approaches come together, complement, and overlap one another. Design Thinking brings discipline to how problems are framed, customers/users are considered and creative solutions are explored. Lean Startup brings discipline to learning, making decisions, and coordinating efforts to achieve the business goals. CRISP-DM shows how technology solutions are adapted/evolved by learning and responding to changing needs that emerge over time. From this, it is evident that there is no one correct way, nor is one single approach enough, but together, elements of each methodology can unite to help to guide the way forward.

As development teams learn more about the data they work with, they frequently return to a previous step to make adjustments. Models are not created once, deployed, and left in place as is; instead, through frequent feedback, refinement and redeployment, models are continually improved and adapted to evolving conditions, thus providing continuous value to its customers/users. By applying LDTM, the intention is to speed up the delivery of the solution and to regularly engage customers/users to ensure that the solution is more likely to meet their requirements. However, we do also acknowledge that knowing when to stop tweaking the solution and finishing the development is a limitation of the LDTM and is, therefore, something that needs to be addressed in further enhancements of the methodology. Going forward, the next step would be to test the LDTM in a real-world context/scenario and from this determine the impact of the LDTM.

References

[1] D. Bourgeois, Information Systems for Business and Beyond. Saylor Academy, 2014.

[2] J. Saltz, "The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness", in 2015 IEEE International Conference on Big Data (IEEE BigData 2015), Santa Clara, California, 2015, pp. 2066-2071.

[3] G. Piatetsky, "CRISP-DM, still the top methodology for analytics, data mining, or data science projects", kdnuggets.com, 2014. [Online]. Available: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[4] K. Thoring and R. Mueller, "Design Thinking vs. Lean Startup: A comparison of two user-driven innovation strategies", in 2012 International Design Management Research Conference, Boston, Massachusetts, 2012, pp. 151-161.

[5] J. Haffar, "Have you seen ASUM-DM?", SPSS Predictive Analytics, 2015. [Online]. Available: https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm.

[6] B. Severtson, L. Franks and G. Ericson, "What is the Team Data Science Process?", Microsoft Azure, 2017. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview.

[7] J. Määttä, "Data Science Is Like a Team Sport - You Need the Team, Strategy, Execution, Process and Collaboration to Be Successful", Linkedin, 2017. [Online]. Available: https://www.linkedin.com/pulse/data-science-like-team-sport-you-need-strategy-execution-määttä.

[8] T. Brown, "Design Thinking", Harvard Business Review, vol. 86, no. 6, pp. 84-92, 2008.

[9] S. Bell, "Design Thinking", American Libraries, vol. 39, no. 1&2, pp. 44-49, 2008.

[10] E. Ries, The Lean Startup How Constant Innovation Creates Radically Successful Businesses. London: Portfolio Penguin, 2011.

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

[12] Ö. Marbán, G. Mariscal and J. Segovia, "A Data Mining & Knowledge Discovery Process Model", in Data Mining and Knowledge Discovery in Real Life Applications, J. Ponce and A. Karahoca, Ed. InTech, 2009, pp. 1-16.