Data Science Maturity Model: From Raw Data to Pearl's Causality Hierarchy

Luís Cavique[1](✉), Paulo Pinheiro[2], and Armando Mendes[3]

[1] Universidade Aberta and Lasige-FCUL, Lisbon, Portugal
luis.cavique@uab.pt
[2] Universidade Aberta and Cedis, Lisbon, Portugal
ppinheiro@cedis.pt
[3] Universidade Açores and LIACC, Ponta Delgada, Portugal
armando.b.mendes@uac.pt

Abstract. Data maturity models are an important and current topic since they allow organizations to plan their medium and long-term goals. However, most maturity models do not follow what is done in digital technologies regarding experimentation. Data Science appears in the literature related to Business Intelligence (BI) and Business Analytics (BA). This work presents a new data science maturity model that combines previous ones with the emerging Business Experimentation (BE) and causality concepts. In this work, each level is identified with a specific function. For each level, the techniques are introduced and associated with meaningful wh-questions. We demonstrate the maturity model by presenting two case studies.

Keywords: data science · maturity models · business experimentation · wh-questions · causality

1 Introduction

Data maturity models are a valuable and current topic since they allow organizations to plan their medium and long-term goals (Carvalho et al. 2019). Maturity models are an essential business management tool (Davenport 2018), allowing organizations to improve the planning of actions that should lead to the desired results. This problem is even more relevant as new concepts, keywords, and products are launched yearly in the information technology market, whose impact is rarely known.

In the 2010s, academic journals and companies began recognizing Data Science as an emerging discipline (Chiarello et al. 2021). The emergence of Data Science goes beyond Business Intelligence (BI) and Business Analytics (BA), abbreviated by BI&A. Data Science covers a broad spectrum of data and its derivatives, from initial data engineering (extraction, integration, transforming), exploration (aggregation, visualization), and modeling. Data Engineering (DE) arose in this decade as a synonym of Data Wangling, Feature Engineering, and Data Pre-processing, partially replacing the well-known

A. Rocha et al. (Eds.): WorldCIST 2023, LNNS 801, pp. 326–335, 2024.
https://doi.org/10.1007/978-3-031-45648-0_32

ETL (Extraction, transformation, loading). Data science overlaps multiple data-analytic disciplines, such as databases, statistics, operations research, and machine learning.

Most maturity models do not follow what is done in the technological sector, particularly in the GAFAM (Google, Amazon, Facebook, Apple, Microsoft), regarding Business Experimentation (BE). Experimentation is a simple method to test new ideas systematically. Experimentation can test a theory or hypothesis, help evaluate an existing product, and be valuable beyond the tech sector, making organizations brighter. A new period is rising in companies, the experimental revolution (Luca, Bazerman 2020).

In the requirements elicitation, the right questions lead to good application design. Moreover, questions illustrate the artifact's purpose, like in the well-known Gartner Analytic Ascendancy Model (GAAM) (Gartner 2012).

This work aims to present a comprehensible data science maturity model that includes the well-known business intelligence and analytics areas, the new practices in business experimentation (Thomke 2020), and Pearl's causality hierarchy (Pearl 2019). The proposed pipeline can be scratched as DE → BI → BA → BE. The proposed maturity model named _IABE is the Intelligence, Analytics, and Business Experimentation acronym.

In this paper, the keywords maturity models, hierarchy, and ascendancy models have the same meaning.

Two different contributions to the data science hierarchy are present. The levels of the model are given on the first contribution. Moreover, the techniques and the associated wh-questions are reported in the second contribution. A comprehensive approach to the typical causal questions is also provided.

The remaining paper is organized as follows. Section 2 presents the proposed maturity model _IABE with the levels and wh-questions. Two case studies demonstrate how the proposed model can be applied in Sect. 3. Finally, in Sect. 4, conclusions are drawn.

2 Proposed Maturity Model _IABE

The proposed Data Science Maturity Model comprises four stages: a previous level regarding Data Engineering (DE), followed by Business Intelligence (BI), Business Analytics (BA) at the second level, and finally, Business Experimentation (BE) at the upper level. The model pipeline is DE → BI → BA → BE.

Table 1 shows the rubric framework of the Data Science Maturity Model with four levels and four criteria. The criteria/dimensions include the business information system, the approach to planning, the guidance of the level, and finally, a synthesis function where Tc means the algorithmic running time complexity, represented by the big O notation, and T denotes the treatment of the experiment.

We associate the functions f(X), g(X), and h(X, T) with BI, BA, and BE. Functions f(X) and g(X) use the same argument, X, where X denotes the set of attributes of the system. On the other hand, function h(X, T) has two arguments, where T represents the treatment.

Since $O(N^2)$ is the time complexity between easy and hard problems, we create the threshold of $O(N^2)$ for the time complexity to distinguish between BI and BA, where N is the number of lines of the dataset. The function f(X) can be exemplified by the sum of an attribute in an OLAP system, showing a running time complexity lower than $O(N^2)$,

$Tc(N) \leq O(N^2)$. Moreover, function g(x) can be exemplified as a classification algorithm of a predictive model, with running time complexity usually upper than $O(N^2)$, $Tc(N) > O(N^2)$.

The second criterion in Table 1 is closely associated with how information system planning is developed. The BI behavior is reactive since it only cares about past events. Given the more complex models of BA, it is possible to elaborate on recommendations and take a proactive role in the company. Finally, in the BE stage, the ability to interact with customers by performing controlled trials moves the company to a new level in organizational learning with interactive planning.

Table 1. Data Science Hierarchy

		level 0	level 1	level 2	level 3
criteria	Information system	Data Engineering	Business Intelligence	Business Analytics	Business Experimentation
	approach to planning	inactive	reactive	proactive	interactive
	guidance	data pre-processing	data-driven	model-driven	experiment-driven
	function	n.a	$y = f(X)$, $Tc(N) \leq O(N^2)$	$y = g(X)$, $Tc(N) > O(N^2)$	$y = h(X, T)$

The previous level, named level 0, comprises Data Engineering tasks, where questions or answers are not presented, and there is no planning. This level includes preprocessing data features as an ETL process (Extraction, transformation, and loading) (Cavique et al. 2019).

Business Intelligence (BI) comprehends tools to support data-driven decisions, emphasizing reporting and data visualization. Data warehouse design is essential to provide multidimensional data tables that can be analyzed using OLAP (online analytical processing) systems (Cavique et al. 2020). Based on KPI (key performance indicators), alerts can be triggered in management by exception environments. The approach to planning is reactive based on the current information.

Business Analytics (BA) merges the areas of Data Mining/Machine Learning with Decision-Making tools. In Data Mining, two sub-areas should be mentioned: descriptive and predictive. The descriptive approach looks for relevant patterns in the data (Cavique 2007; Tiple et al. 2016; Cavique et al. 2018b), and the predictive models use supervised algorithms with labeled data to anticipate future events (Cavique et al. 2018a). Decision-making models include the techniques also referred to as Deductive modeling and studied in Operations Research, like decision analysis, simulation, and optimization (Cavique et al. 1999; Santos et al. 2013). These techniques aim to find the best solutions for each decision problem. BA comprehends tools to support model-driven decisions where the approach to planning is proactive.

In Business Experimentation (BE), experimentation interacts with individuals (customers, patients, or users), generating more data and feeding back to the system. A low-cost business experiment can change the way organizations design decision-making. BE comprehends tools to support experiment-driven decisions in interactive planning. Pearl's causality hierarchy refers to association, intervention, and counterfactuals, where the association is closely related to the traditional data mining approach. In the BE level, we add a new sub-level, the explanatory one, reflecting the title of the book 'The book of why' (Pearl, Mackenzie 2018).

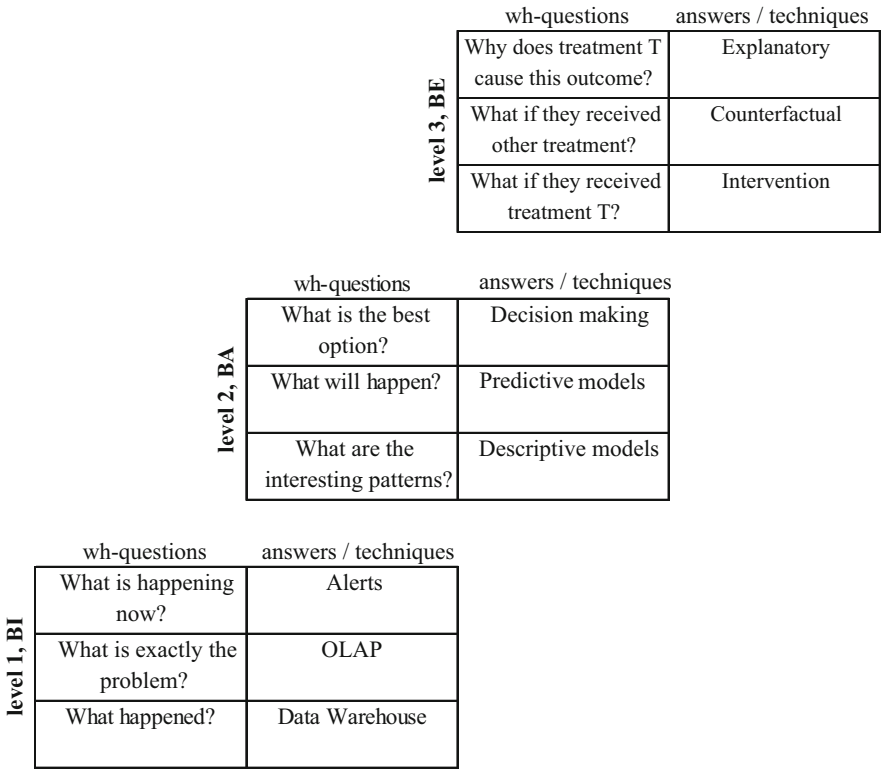

Fig. 1. Sub-levels of the Data Science hierarchy with wh-questions

Figure 1 shows the three stages associated with the standard techniques and the related comprehensive questions that can be asked. BI&A is highly effective at answering questions of 'what'. On the other hand, BE answers questions of 'what if' and 'why', which implies causal relationships.

The sub-levels of BI, in addition to data warehouses, OLAP systems, and Alerts supported by KPI, are included. The BA sub-levels comprehend the descriptive, predictive, and decision-making models. As sub-levels of BE, we included, in addition to the randomized controlled trial, the last rungs of Pearl's causality ladder, the counterfactual, and explanatory reasoning.

3 Case Studies

Given the pipeline of levels, DE → BI → BA → BE, the two different case studies focus on the transitions, BI → BA and BA → BE, are presented. For each level, business wh-questions illustrate the potentialities of the reported techniques.

As stated before, we associate a function to each level: BI with f(x), BA with g(x), and BE with h(x, T). BI and BA use the same data, but different wh-questions can be pointed out. In BE, the concept of intervention/treatment T draws closer to the problem of causality identification.

3.1 From Business Intelligence to Business Analytics

In the first case study, we consider a hospital's information system that includes patient data, medications, surgeries, healthcare professionals, and billing.

Table 2 shows a data extract from the hospital's information system, where the 'complication' attribute indicates the patients who had postoperative difficulties (Dhar 2013).

Table 2. Extract the hospital's information system

patient	age	#medications	payments	complication
1	52	7	3,121 €	Yes
2	57	9	7,113 €	Yes
3	43	6	3,475 €	Yes
4	33	6	520 €	No
5	35	8	789 €	No
6	49	8	4,177 €	Yes
7	58	4	239 €	No
8	62	3	678 €	No
9	48	0	97 €	No
10	37	6	1,690 €	Yes

We present four relevant wh-questions at the BA level. The first three wh-questions are answered by the data mining area (classification, clustering, association), and the last one belongs to the area of management science (inventory management), as follows:

BA.1 – what are the rules that explain postoperative complications? (classification);

BA.2 – which patient groups can be found? (clustering);

BA.3 – which drug D is used with drug C? (association);

BA.4 – how to mitigate breaks in the medication stock? (inventory management).

The first two wh-questions use the data presented in Table 2. The other two wh-questions use the available data from the hospital's information system.

Classification algorithms are supervised methods that make predictions based on a discriminant class. In Table 2, the discriminant class corresponds to the 'complication' attribute. The pattern retrieved by BA.1 can predict whether a new patient 50 years old and not taking medication will have any complications in the postoperative period.

Descriptive methods do not use discriminating attributes; they are also unsupervised since any attribute with unique characteristics does not guide them. Wh-questions BA.2 and BA.3 exemplify this set of methods. For wh-question BA.2, the clustering algorithm divides patients into K groups by age, the number of medications, and payment that best characterizes the dataset. In Table 2, regarding payment, four patients are identified as 'major users' of the hospital, having spent more than the average.

An association rules algorithm is used to answer wh-question BA.3, which finds the most frequent sets in the same attribute. For example, people who use paracetamol also use a nasal decongestant, creating a rule as follows: paracetamol => nasal decongestant.

For wh-question BA.4, different inventory management policies can be applied to determine when and how much to order; deterministic or stochastic models can be applied. The most common determinist models use continuous or periodic reviews.

Given the BA wh-question, in BI, similar wh-questions can be asked. In BI, the wh-questions are more straightforward, as follows:

BI.1 – which patients have postoperative complications?

BI.2 – which patients paid more than the average?

BI.3 – what are the two most used medications?

BI.4 – which medications are in short supply?

Wh-questions BI.1 and BA.1 are similar, considering patients with postoperative complications. However, the first has an easy answer using an SQL query, while the second lacks a classification algorithm.

Likewise, wh-questions BI.2 and BA.2 refer to the amounts paid by patients in the hospital, in which the first wh-question is answered with SQL, and the second to determine the groups of patients is to use a clustering algorithm.

Wh-questions BI.3 and BA.3 also refer to similar issues, considering medications. However, for the first wh-question, it is enough to order the medication consumption, and the second one lacks an association rule algorithm.

Finally, BI.4 and BA.4 refer to medication inventory, but the first is answered using a database query, and the second involves more complex policies and algorithms, like the other BA wh-questions.

A particularity should be highlighted in the case of wh-questions BI.1 and BA.1.

Time complexity distinguishes between BI and BA. BI.1 is solved by a SQL query running time complexity of O(N) in the worst case. On the other hand, BA.1 requires a classification algorithm with a time complexity larger than $O(N^2)$.

In wh-question BI.1, we intend to find the patients who had complications, with patients 1, 2, 3, 6, and 10 retrieved. The SQL query is as follows: Select patient, age, #medications From Table 2 Where complication = 'yes'.

On the other hand, in wh-question BA.1, the data are provided, intended to extract patterns. Thus, we intend to know the attributes that cause postoperative complications.

With a classification algorithm, the following rule is found: (Age $\geq$ 35 and #medications $\geq$ 4) => Complications = 'Yes'. Those 35 years or older who take four or more medications have medical complications.

The two approaches treat the same data from information systems in different ways. Although the wh-questions are similar, BI.1 presents a pattern (e.g., SQL query), and data are retrieved. On the other hand, in BA.1, the data is provided, and the patterns are extracted.

3.2 From Business Analytics to Business Experimentation

The transition from BA to BE is exemplified in the second case study, using the Telco Customer Churn (2018) public dataset that contains information on eighteen covariates potentially related to both the outcomes of interest (churn or no-churn). Telco's churn is around 26%, revealing the importance of customer retention interventions that require concrete and personalized actions.

Traditional data mining studies focused primarily on predictive mining, where the cause-and-effect scenario is described. However, this information alone is insufficient as it does not benefit the final user. What becomes more exciting and critical to organizations is to mine patterns to create actionable knowledge (Cao 2007). Some attributes cannot influence or be changed, such as the attribute 'age' or 'gender', denominated by non-actionable attributes. On the other hand, the attributes that allow operational changes are called actionable attributes. Actionable attributes operationalize actionable knowledge.

Cao (Cao 2007) (Cao 2010) presents a new approach, which opposes data-driven to domain-driven. Data-driven corresponds to traditional data mining, while domain-driven is related to the business domain or business area. The domain-driven data mining, D3M, closes the gap between researchers and practitioners by generating actionable knowledge for real user needs. D3M approach moves away from BA and goes towards BE. Beyond the usual data, treatment T is introduced in the function $y = h(x, T)$.

Each row represents a customer in the Telco dataset, and each column contains the customer's attributes. Those attributes can be grouped in customer demographic information, like gender, age range, and if they have partners and dependents. The second group of attributes describes customers' account information, like how long they have been a customer, contract, payment method, paperless billing, monthly charges, and total charges. Moreover, the third group of attributes presents each customer's services like phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies. There is also an attribute, churn, which indicates whether or not the customer has abandoned services in the last month.

The best actionable attribute is the type of contract since customers with annual or biannual contracts tend to be more loyal than customers with monthly contracts. As shown in Fig. 2, to avoid churn (Y), the actionable attribute 'contract' (T) was chosen from the attributes of the decision tree. We aim to find any actionable attributes in which we can intervene to avoid dropouts and measure their causal effects and impact on the business (Pinheiro, Cavique 2022).

In order to exemplify the transition of information from BA to BE, two wh-questions are formulated. BA.u is a prediction question about the churn variable, and BE.u measures the effect of an intervention T regarding the churn variable, as follows:

BA.u – what are the rules that explain the churn Y?

BE.u – what is the impact of treatment T on the reduction of the churn Y?

As in the previous sub-section, the two approaches treat the same dataset using different information systems.

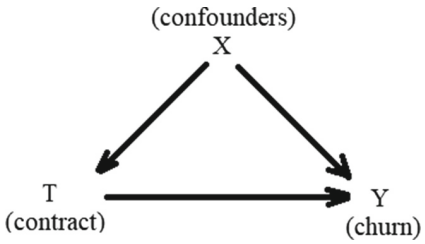

Fig. 2. DAG for Telco customer churn

4 Conclusions

Maturity models (MM) allow organizations to plan their actions to achieve the desired results. The maturity model requires a multi-stage planning tool to identify and control which advances should be made. However, most maturity models do not follow what is done in the digital sector concerning experimentation.

Several maturity models study Business Intelligence (BI) and Analytics (BA) domains (Carvalho et al. 2019). Our first goal aims to find a maturity model in the Data Science domain, including the recent Business Experimentation (BE), approaches (Thomke 2020), and new causality hierarchy (Pearl 2019). Our second objective is to find clear and meaningful maturity levels using a function-based approach and illustrated with wh-questions.

This study proposed the maturity model named _IABE, the acronym for Intelligence, Analytics, and Business Experimentation. Each level can be summarized by a function and a set of techniques associated with meaningful wh-questions.

Two case studies illustrated with wh-questions were presented for the transitions BI-BA and BA-BE. Transition BI-BA showed different questions set side-by-side with the techniques based on a hospital database. The transition from Business Analytics to Experimentation, BA-BE, used the Telco Customer Churn dataset.

This work clarified with comprehensive wh-questions the current levels of a Data Science Hierarchy, where each level was associated with a specific function.

References

Cao, L.: Domain-driven, actionable knowledge discovery. In: IEEE Intelligent Systems, pp. 78–79. IEEE Computer Society, Sydney (2007)

Cao, L.: Domain-driven data mining: challenges and prospects. IEEE Trans. Knowl. Data Eng. **22**(6), 755–769 (2010). https://doi.org/10.1109/TKDE.2010.32

Carvalho, J.V., Rocha, A., Vasconcelos, J., Abreu, A.: A health data analytics maturity model for hospitals information systems. Int. J. Inf. Manage. **46**, 278–285 (2019). https://doi.org/10.1016/j.ijinfomgt.2018.07.001

Cavique, L., Mendes, A.B., Martiniano, H.F.M.C., Correia, L.: A bi-objective feature selection algorithm for large omics datasets. Expert Syst. e12301 (2018a). https://doi.org/10.1111/exsy.12301

Cavique, L.: A scalable algorithm for the market basket analysis. J. Retail. Consum. Serv. Spec. Issue Data Min. Retail. Consum. Serv. **14**(6), 400–407 (2007)

Cavique, L., Rego, C., Themido, I.: Subgraph ejection chains and tabu search for the crew scheduling problem. JORS J. Oper. Res. Soc. **50**(6), 608–616 (1999)

Cavique, L., Cavique, M., Gonçalves, A.: Extraction of fact tables from a relational database: an effort to establish rules in denormalization. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) WorldCIST'19 2019. AISC, vol. 930, pp. 936–945. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-16181-1_88

Cavique, L., Cavique, M., Santos, J.: Supply-demand matrix: a process-oriented approach for data warehouses with constellation schemas. In: Rocha, Á., Adeli, H., Reis, L., Costanzo, S., Orovic, I., Moreira, F. (eds.) WorldCIST 2020. AISC, vol. 1159, pp. 324–332. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45688-7_33

Cavique, L., Marques, N.C., Gonçalves, A.: A data reduction approach using hypergraphs to visualize communities and brokers in social networks. Soc. Netw. Anal. Min. **8**, 60 (2018b). https://doi.org/10.1007/s13278-018-0538-6

Chiarello, F., Belingheri, P., Fantoni, G.: Data science for engineering design: State of the art and future directions. Comput. Ind. **129**, 103447 (2021). https://doi.org/10.1016/j.compind.2021.103447. ISSN 0166-3615

Davenport, T.H.: DELTA plus model & five stages of analytics maturity: a primer, international institute for analytics (2018)

Dhar, V.: Data science and prediction. Commun. ACM **56**(12), 64–73 (2013)

Gartner. Gartner analytic ascendancy model. Gartner.com (2012)

Luca, M., Bazerman, M.H.: The Power of Experiments: Decision Making in a Data-Driven World. MIT Press (2020). ISBN 978-0262043878

Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)

Pearl, J.: The seven tools of causal inference, with reflections on machine learning. Commun. ACM **62**(3), 54–60 (2019)

Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books, New York (2018). ISBN: 978-0-465-09760-9

Pearl, J., Glymour, M.: Causal Inference in Statistics: A Primer. Wiley (2016). ISBN 978-1-119-18684-7

Pfeffer, J., Sutton, R.I.: Knowing 'what' to do is not enough: turning knowledge into action. Calif. Manage. Rev. **42**, 83–108 (1999)

Pinheiro, P., Cavique, L.: Uplift modeling using the transformed outcome approach. In: Marreiros, G., Martins, B., Paiva, A., Ribeiro, B., Sardinha, A. (eds.) EPIA 2022. LNCS, vol. 13566, pp. 623–635. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16474-3_51

Santos, J., Negas, E.R., Santos, L.C.: Introduction to data envelopment analysis. In: Mendes, A., L. D. G. Soares da Silva, E., Azevedo Santos, J. (eds.) Efficiency Measures in the Agricultural Sector, pp. 37–50. Springer, Dordrecht (2013). https://doi.org/10.1007/978-94-007-5739-4_3. ISBN 978-94-007-5738-7

Telco Customer Churn. Dataset (2018). https://www.kaggle.com/blastchar/telco-customer-churn. Accessed 01 Nov 2021

Thomke, S.H.: Experimentation Works: The Surprising Power of Business Experiments. Harvard Business Review Press (2020) ISBN 978-1633697102

Tiple P., Cavique, L., Marques, N.C.: Ramex-forum: a tool for displaying and analyzing complex sequential patterns of financial products. Expert Syst. 1–16 (2016). https://doi.org/10.1111/exsy.12174