How to Design An Interactive System for Data Science: A Literature Review

Ana Sofia Almeida *asa@ipn.pt*
Laboratory for Informatics and Systems, Instituto Pedro Nunes
Coimbra, Portugal

Licínio Gomes Roque *lir@dei.uc.pt*
Department of Informatics Engineering, University of Coimbra
Coimbra, Portugal

Paulo Rupino da Cunha *rupino@dei.uc.pt*
Department of Informatics Engineering, University of Coimbra
Coimbra, Portugal

Abstract

As part of an ongoing design science research project, we present a systematic literature review and the classification of 214 papers scoping the work on Data Science (DS) in the fields of Information Systems and Human-Computer Interaction. The overall search was conducted on Web of Science, Science Direct and ACM Digital Library, for papers about the design of IT artefacts for Data Science, over the period of 1997 until 2017.

Our work confirms a rich interdisciplinary field of inquiry and identifies promising research clusters, with examples. Moreover, we found few studies with concrete guidance on how to design a system for DS when targeting for broader technical and business user profiles and multi-domain application. Being a multidimensional and creative complex process, there is potential in the development of hybrid methods of design theory and practice, for a variety of further work from researchers and practitioners.

Keywords: Data Science, Information Systems Design, Human Computer Interaction, Design Science Research, Design methods, UI/UX design, Systematic Literature Review.

1. Introduction and Motivation

Solving the "big challenges and opportunities" of Big Data [9, 36] has emerged as an important area of study. With the convergence of Big Data and powerful computational capabilities, Machine Learning (ML) and Artificial Intelligence (AI) are becoming effectively viable for widespread use. Enterprises need to adapt and combine domain-expertise with Data Science (DS), to respond with speedier and tailored solutions to demanding markets. This increases the demand for data competences, with a focus on collecting, processing, analysing and using data to create value. DS is the new needed literacy for professionals, regardless of their technical background [37, 39].

This paper is a partial result of an ongoing Design Science Research project [25, 40, 42] with a global leading company using DS for Fraud Fighting, in the financial sector. The main research goal is to improve the design and evaluation of an interactive system to support the DS activities, focusing on the user experience and the continuous improvement of an integrated software for IT-teams and other business stakeholders and professionals. This raises the question of "*How to design such a system, keeping the user/consumer perspective and needs, in i) a competitive and fast-changing environment/market; with ii) scarce access to end-users, while iii) meeting agile development pace deadlines?"*

Currently, there are several viewpoints regarding the definition of *'what is data science'*, but no consensus has emerged [55]. One of the first attempts to a definition of "what is Data Science?" can be found back in 1998, with a partial answer from Hayashi [23], that DS "unify statistics, data analysis and their related methods, but also comprises its results", implying "multidimensional, dynamic and flexible ways of thinking". In 2012, Chen [9] provided a framework for identifying applications and emerging research areas and technologies in "Business Intelligence and Analytics" where DS can be regarded as an evolutionary step, incorporating, among other, disciplines such as statistics, mathematics, computer science, modelling and analytics. Loukides [35] proposes that "data science is a holistic approach" and "data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others". It is not just about using data or retrieve information, but all the activities performed in order to create and withdraw value from the data, occurring within the social and organizational context of an enterprise [1].

As a working definition, we consider DS as a data-driven science and a holistic approach that uses automated methods to gather, enrich and analyse large amounts of data in order to discover and extract knowledge from them to be reported to others, while creating value and new data, with great production system operational performance. The discovery experience, the iterative nature of an intended flexible process and the need to report the results increases the demand upon the interaction with data, the interface and overall experience of the different types of users (e.g. data analysts, data scientists, delivery teams, stakeholders, managers). In order to approach the design of a IT-solution for supporting DS life cycle, we considered a twofold perspective over the underlying process:

i) as the design of an *Information System (IS*) – where data, technology and people are involved to deliver a data product or service – and

ii) as the design of a system for *Human-Computer Interaction (HCI)* – an interactive system, used by different user profiles to perform complex activities.

A third area of *Design* with its actual methods and techniques (e.g. design theories, practices, studies, and approaches) is taken to be transversal to the two fields and thus was also mapped, by including relevant search keywords, as tertiary terms of search.

In the next section, we present our initial systematic literature review effort, detailing the search used to attain a representative list of publications for classification.

In section 3, relevant articles were then grouped by similar concepts into a first set of three categories and ten sub-categories, outlining the scope of retrieved literature and proposed as major (current and) future research clusters.

In section 4, a second list of 80 papers was selected from the previous sample, mentioning "*design*" or "*design methods*" on paper's *Keywords* or *Abstract* and narrowed to a final list of 20 papers, further inspected to inform the aforementioned research question.

In the last section, we discuss the work done and some of the shortcomings of the literature review, proposing found research gaps as opportunities for further work.

2. Research Method and Search Strategy

A systematic literature review is a formal iterative approach of "analyse the past to prepare the future" [52] and locate, select, explore and report [33, 50, 52]. Although challenging, systematic literature review provides a means for practitioners to use the evidence provided by previous research to inform their decisions [50].

For theoretical background, we conducted a systematic literature review for papers in the areas of *Information Systems*, *Human-Computer Interaction* and *Design* studies, relevant to the problem of design and evaluate a system for *Data Science*. An overall search was carried out on Web of Science, Science Direct and ACM Library databases, in November and December of 2017, considering the period from 1997 to 2017.

Table 1 illustrates the following search query for retrieving the articles from the three databases:

1. Boolean combinations of the primary and secondary terms were searched, using the search engines of the databases for cross-reference work in the relevant areas.
2. Tertiary terms were used to refine the retrieved results of 1).
3. Boolean combinations of primary and tertiary terms were considered and additionally combined with other tertiary terms, progressively narrowing the scope of results.
4. Transversal tertiary terms (e.g. 'approach') along with discipline specific terms (e.g. 'user experience') were used to capture part of discourse community specificities.
5. Conceptual terms were mainly used in manual searches, to capture different discourses over the theme (e.g. 'mental model; 'survey', 'theory'; 'case study').

Table 1. Search query was done using combinations of AND and OR between the terms in each row.

Primary Terms Searched in metadata (WOS+SD+ACM)	data science'; 'data scientist', 'big data'; 'data analytics'; 'business intelligence'; 'business analytics';	
Secondary terms Searched in *Topic*, *Title*, *Abstrac*t and *Keywords*, depending of data source (WOS, SD, ACM)	'information systems'; information systems design'; 'IS'	'human computer interaction'; 'HCI';
Tertiary Terms Searched in *Topic* and also in *Title* / *Abstract* / *Keywords* (WOS+SD, ACM)	'design'; 'system design'; 'approach'; 'process'; 'method'; 'tools'; 'solutions'; 'practice'; 'application';	
	'software'; 'system design'; 'artefact'; 'service'; 'lifecycle'	'user interface'; 'usability'; 'user experience'; 'user experience design'; UX design'
Conceptual Terms	'model'; 'mental model'; 'theory'; 'survey'; 'case study'; 'field study'; 'review'	

We searched papers appropriate for review, not considering: i) Papers not written in English and ii) Dissertations, lecture notes, reports on tutorials, posters, demos and workshops. We are aware that some articles may have not been retrieved, partially due to the search method and the different technical capabilities of each database. From the 519 articles that resulted, we selected 214 papers, directly mentioning "*Data Science*" in *Title*, *Abstract* and/or *Keywords*, for a first categorization, revealing research clusters and promising gaps for further work.

3. Paper Categorization and Main Research Clusters

Paper categorization was done iteratively, in a process of clarification and refinement. Considering the two perspectives adopted (IS development and HCI views over the research work studies) and the somewhat blurred frontier of the research in the emerging field of "Data Science", we first organized the papers into three categories, closely related to the secondary terms and thematic research, thus over-lapping:

1. **Data Science and Information Systems**: includes papers about foundational theories and overviews of the area; questions about data production, storage and availability or reporting on tools, algorithms and technologies for DS.
2. **Data Science and Human-Computer Interaction**: papers related to data-driven and data-informed design, implementation and evaluation of systems and services; data visualization and discovery or about the applications of DS on multi-domains.
3. **Emerging perspectives/shared work for Data Science**: articles proposing conceptual models or methods; discussing emerging trends and future challenges; the impact of potential technologies or how to educate for DS and train professionals.

Table 2 presents a rapid overview of the three categories and papers' research problems, leading to the ten sub-categories presented. For each category, we present some examples of articles

and how they relate to the found research gaps and potential opportunities to future work. Due to space limitation, only a sample of the categorized papers appear on the references. The full list is available, by requesting the first author.

3.1. Data Science and Information systems

A total of 102 (48%) articles, classified in this first category, reveal a body of knowledge pushing the confluence of more traditional disciplines (e.g. statistics, mathematics, data analysis or decision support systems). Among other themes, we found main research clusters and favorable avenues of work to be related to:

1. **Theories and Case Studies:** scoping literature reviews that organize research and practice; studies reporting about technologies, tools and techniques borrowed from other areas to be applied in DS and overall work to map the complex phenomena of DS. Some exemplary work can be found in, for example, [10, 32, 43, 48].
2. **Production, Storage and Analysis of [Big] Data:** information systems produce large amounts of data, collecting, storing, managing and distributing it. Organizations struggle to grasp its potential to make decisions, to innovate and to create value for Customers and Citizens. Arguably, a fast-pace growing study area, in need of "tools to deal with variety, velocity and volume of data" [32, 36, 44] so as to decision-making [16, 30, 38].
3. **Data Quality:** a fast-growing problem, with research challenges focusing "on scalability, availability, data integrity, data transformation, data quality" [21] urging the need to manage identified "causes for 'bad big data science', focusing primarily on the quality of the input data" [22], or the "data quality in supply chain research and practice", calling for "interdisciplinary research topics based on complementary theory" [24].

3.2. Data Science and Human-Computer Interaction

The 54 articles (25%) classified in this category show the growing application of DS in multi-domain scenarios. How data can improve the user experience and the service design, adding value for clients and citizens is pushing the work on data visualization (for communication and discovery) and crowdsourced research.

Main challenging clusters, from a user-centred perspective, relate to:

1. **Application of DS in multi-domains:** DS is fast becoming ubiquitous in multi-domain sectors. Examples range from Health [51, 53], Agriculture [45, 54]; Social Studies [8, 12, 34] or Urban Planning [3], urging for the genericity of solutions. There are challenges on structuring the DS practices and how those can inform the design of a system for DS.
2. **Data Visualization and Data Discovery:** Examples come from research addressing the challenge of "making sense out of big data using visual analytics" [18] or by "allow users to directly interact with the visualization to build combination models" [49], promising to guide users to extract value in data. Other samples address data visualization courses for sectors as tech industry [2] or citizen science [47], targeting non-experts in DS.
3. **DS to improve Research and Learning**: ML combined with large datasets can improve research in different fields, "converting data to actionable knowledge" [5] and creating "new opportunities for researchers to achieve high relevance and impact"[8]. DS is also important "in an educational context", a field where studies "related to outcome measurement and prediction, to be linked to specific interventions" [34] are needed. This can be considered yet another domain of DS application, lacking a much-needed framework.

Table 2. Categorization of major problems in the selected publications, for Data Science

214 Papers Selected	Categories (study focus)	Main Problems / Research Questions found	Examples / Quotes
CAT1 **102 papers** **48%**	**Data Science and IS** Theories & Studies; Models, Methods from IS applied to Data Science (e.g. statistics, data analysis, business intelligence, decision making, Data Science lifecycle, technologies)	Evolution/overview of Big Data and related research (challenges, trends, future directions) Production, storage and analysis of [big] data Efficiency of systems through data Call for methods and approaches and systematization (from ECIS) Tools and Technologies (design and development of tools to deal with data problems) Data Quality (integrity, security, availability)	"big data over the past 20 years" [10, 32, 43, 48] "increasing open source tools to deal with variety, velocity and volume of data" [36] "research challenges …, with focus on scalability, availability, data integrity, data transformation, data quality" [21] "how to align [organizations'] decision-making and organizational processes to data that could help them make better-informed decisions"[27] "big data has resulted in the development and applications of technologies and methods aimed at effectively using massive amounts of data to support decision-making and knowledge discovery activities" [48] "apply multiple technologies, carefully select key data for specific investigations, and innovatively tailor large integrated datasets … All these actions will flow from a data value chain" [38] "causes for 'bad big data science', focusing primarily on the data quality of the input data, and suggests methods for minimizing"[22]
CAT2 **57 papers** **25%**	**Data Science and HCI** Work being done in HCI field, related to data-driven research and user experience improvement (mainly, the application of DS in multi-domain scenarios and how data can enable UI/UX or service design)	Context of use: several domains for DS application (e.g. health, heritage, urban cities, social, manufacturing, finance, research and learning) Tools for data understanding, discovery and visualization Using data to enable HCI projects or research with a user-centred perspective (user data collected to improve UI/UX design or UI/UX Studies involving diverse users and profiles are possible)	multi-domain examples of HCI projects with DS application range from Health [51, 53], Agriculture [45, 54]; Social Studies [8, 12, 34] to Urban Planning [3] "making sense out of big data using visual analytics" [18] or "allow users to directly interact with the visualization to build combination models" [49] "interactive visualization over data sets" [13] "converting data to actionable knowledge" [5] "yet to fully harness the potential of visualization when interacting with non-scientists." [20] "new opportunities for researchers to achieve high relevance and impact"[8] DS to improve studies "related to outcome measurement and prediction, to be linked to specific interventions" [34]
CAT3 **97 papers** **45%**	**Emerging research & shared problems** Everything related with cloud computing, machine learning, algorithms, [data, mobile, web, network, text, social media] analytics; privacy; ethics and policy making.	Education for Data Science (skills and competences) Emerging Technologies and their applications (e.g. ML, AI, IoT, Industry 4.0, Cloud Computing) Case studies and Contextual Surveys Legal and regulatory issues. Ethics and public policies concerns	"DS increasingly significant for business strategies, operations, performance, efficiency and prediction … little work on this to provide a detailed guideline."[41] "how we might design effective methods for systematizing such practice and research"[14] "Data science literacy = computational literacy + statistical literacy + machine learning literacy + visualization literacy + ethical literacy." [17] "we identified four types of DS projects, and … some of the sociotechnical challenges" [46]

3.3. Emerging/ Shared Research Problems

Ninety-seven articles (45%) report on emerging challenges and trends [36]. Internet-of-Things or Industry 4.0 with Cloud Computing, ML, AI and other emerging technologies urge for interdisciplinary collaborative efforts to create effective DS solutions for the data-related challenges societies face, beginning in the design phase until production an need to "process large real-time streams of data" [19]. Ongoing research and practice work will grow, namely, in the areas of:

1. **Education and Training for Data Science**: shortage of data-professionals is a reality and "we're missing the boat again" [28]. Calls as "towards data science literacy" [17] effectively organize the body of knowledge while industry urges for fast-forward training pipelines and top short the lifecycle process for demanding markets.
2. **Methods and Systematization of DS lifecycle:** in data-driven societies, DS is "significant for business strategies, operations, performance, efficiency and prediction" [41] but "there is little work on this to provide a detailed guideline". "Executing a data science project is more than just identifying the best algorithm and tool set to use. Additional sociotechnical challenges" [42] and how we might "design effective methods for systematizing such practice and research"[14] eventually automatizing parts of the process or the ML pipeline are worth for further work.
3. **Regulatory, Ethical and Public Policies Issues:** a fast growing topic concerns legal ethical, privacy and security issues, with studies outlining data problems "that could result in invalid conclusions and unsound public health policies"[26], for instance.
4. **Real Life Global Problems with Social Impact:** DS can help solve global problems such as understand "the evolving global economy"[36] or the "open collaboration ecosystems" [4] phenomena, with data-driven designed or data-informed solutions.

Not being exhaustive, these categories outline what we found to be known or at least reported, focusing on the consensus and shared perspectives on both fields. Organizing the papers on these generic categories revealed a pattern of emergent research focus on the process in itself:

- **Data Acquisition and Quality**: how to collect and prepare data to be used with confidence?
- **Data Visualization and Data Discovery**: with questions of user interaction for better understanding by users and scalability issues to deal with larger datasets;
- **Model and Training of ML**: emerges as an iterative, creative and collaborative process, and an IT-solution should account for trial and error. That urges for user experience design and improved interaction of professionals with the ML technology and results;
- **Enhance DS Skills Among Professionals**: shortening the time to produce professionals with the current needed skills OR improve the tools they use so that – at least – part of the DS cycle could be done by professionals with different backgrounds.

We find that few studies cross IS and HCI disciplines (only 7 articles are classified both in categories 1 and 2), revealing a somewhat separate investigation, at least on what relates to DS and data-related challenges and user-centred issues.

The over-lapping is stronger between categories 1 and 3, where 20 papers from IS field are closely related to emerging research trends and gaps, considering the data science context. This is not that surprising, since IS and digital information systems are increasingly responsible for generating data and demanding solutions to extract value from it in order to improve operational, organizational, strategical and societal activities, pushing the boundaries of both related theory and practice.

Surprisingly, we did not found articles reporting user experience studies involving data scientists (as a target group) or about the UI design of tools for both data scientists and non-data scientists. As if, albeit being a complex process and an activity best served by digital tools, usability and a user-centred perspective over the design for DS does not seems to be present or reported, at least in the academic discourse.

4. Second Paper Categorization and Contributions to our Research Question

Being an iterative process of clarification and keeping in mind our research question, we further tried a thematic analysis, resorting to the tertiary terms to select papers for a second paper categorization. The original list of 214 papers, relating to "*data science*" work in the fields of IS and HCI (categorized on section 3), was narrowed to a list of 80 papers explicitly mentioning "*design*" (e.g. 'design methods', 'design approach', 'user-centred design') in *Title*, *Keywords* or *Abstracts*. Doing the best effort to avoid bias, we used the following content analysis criteria to narrow this second list:

- Papers mentioning 'data science' (or related terms, such as 'big data') or 'design' just as a tool or noun within the work/project being reported, not being the main issue or research problem addressed;
- Papers reporting solely on tools, algorithms or mathematics associated with DS, with detail about the programmatic questions of the field;
- Papers not focusing on the design for DS problems primarily (e.g. data is analysed for marketing and business approaches; papers arguing about the need of big data);
- Papers not having explicit HCI design or IS design implications for further studies, specifically in their discussion or conclusion sections;
- Papers not focusing on the design of tools and/or participatory methods targeting data science users (e.g. agronomical studies, health studies or others were data is analysed to solve a different problem and the problem of data scientists or of doing data science is not discussed);
- Papers not including studies with or about the DS process as main research problem or not involving data scientists as users;
- Papers without enough detail and/or impossible to extract data regarding our study;

The list of 80 selected papers was thus narrowed to a final list of 23 papers, inspected for more direct contributions to our research question by either 1) add to the understanding or structure the DS practices and activities (about the *process*) or 2) helping to identify attributes and/or requirements of the IT-solution (about the *artefact*).

Table 3 presents an overview of some of the research questions addressed in these papers quoting those with more relevant contribution to our inquiry. Next, we briefly discuss the findings and impact on the practice of our own work.

4.1. About the Data Science *Process*: A Generic Workflow for the Full Lifecycle

From reviews and evolution studies of the field (an body of knowledge in evolution), such as the DS journey presented by recent Cao's overview of the field [6], we learn that DS lifecycle is in itself an evolving concept, aggregating activities, and a creative process "in need of other methodological research" [29]. Case studies and reports on the application of DS in a particular project or sector, and the problems researchers and practitioners report (or not) having in practice with the process, the toolbox and the technical competences needed are important to learn about DS lifecycle. Based on literature review and the work in practice, we propose a generic workflow of the main DS activities, covering the whole lifecycle, performed in loop:

Data acquisition → Data cleaning & enrichment → Model and Training
→ Experimentation & Evaluation → Deployment → Feedback & Model Tuning (restart)

It was also claimed by [29] that some design methods can be used in data-science, to analyse and make sense of the phenomenology (e.g. studying factors that affect success of a DS project) or to help the data scientist (or other professional, as we are targeting) in structuring the creative exploration process in a controlled way.

We agree and argue that there is potential to use and adapt hybrid design methods to structure the DS activities, and fruitful design inspiration can be found in the tested solutions of existing digital collaborative design studios and of digital tools for creativity.

Table 3. Understanding the *process* and the user perspective over the *artefact*: contributions from previous work to the context of design for DS

Research Questions	Main Problems	Quotes / Contributes	Contribution / New Questions to Our Work
About the Process *Which are the DS phases and activities to support?* *Is there a framework to structure the practice to be used to inform the design for DS?*	"No consensus on DS definition" albeit some shared phases Different DS activities & practices to structure Integration of DS into agile development environments There is a separation between model and deployment phases and cope the DS process in agile environments	"An essential need of the DS state lies in the standardization of its components, methods, tools, data formats, and analytic processes" [7]; Kazakci [29] talks about "controlled creativity" and claims to use design methods to structure the DS activities and its innovation dynamics; Larson [31] presents a most comprehensive overview of DS process, in particular, in agile development contexts. Phases for the complete process are proposed; A "Business Data Science model, focusing on the model and experimental development" allowing "different functions, processes and roles" is proposed by Newman [41] Saltz [46] also provides a framework for DS (not so for designing for DS) and Demchenko [15] states that "the education and training of DS lacks a commonly accepted, … design the whole lifecycle of data handling in modern, data driven research and digital economy."	How to deal with data acquisition and data quality? Which problems to solve when digitalizing each DS phase? How to manage the DS process activities? How to account for the complex and creative nature of the DS process? Can we reduce the technical skills along the phases of the process? Where? How? How much of the process resorts to memory & recall from user? What can be re-used among phases, activities or DS projects?
About the Artefact *Which functionality has to be provided / supported, giving the new technologies (functionality)?* *How to lower the skills & competences needed to do DS (user-side)?*	Skills and competences needed are highly technical to use the available tools; Toolbox instead of an integrated tool; Shortage of Data Scientists urges the call for DS to be accessible to diverse professional backgrounds;	Chuprina [11] proposes an ontology of DS to improve DS skills for CS, since "both industry and academia have met a growing gap" Dichev proposes a set of DS skills & competences, such as "ability to visualise and report summary data and formulate productive questions", calling for data science literacy [17] Grainger [20] explore visualization of data to be communicated and shared (to non-scientists); Carbone [7] argues about the "social dimension of DS" and the problems associated with data misuse", considering an essential need for DS " lies in the standardization of its components, methods, tools, data formats, and analytic processes"	Full lifecycle support, integration of tools and collaborative work support? Can an IT-solution help with standards and integration? How? Improve usability and user experience, while performing the activities? How humans can cope with the increasing complexity? How to evaluate the data products and data science solutions? Ethics and the designer's responsibility in designing for DS?

4.2. About the *Artefact* for Data Science: Guiding Design Principles

Considering that the design of an interactive solution, from a user-centred perspective, ultimately intends to empower the end-user, papers discussing the challenges in educating for data science or improving data-related skills and competences of other professions reveal potential user interaction issues to account for when designing for DS.

However, there are few studies providing concrete guidance on how to design an interactive system for DS and the nature of the multidimensional process it has to support. On the other hand, as argued, design methods can provide, to some extent, a framework for the innovation and creative dynamics of data-related processes and DS activities. We consider a worthy and

pertinent research effort to develop an integrated approach and supportive platform to democratise DS, from a user perspective. That could be of interest to academic research gaps and of practical relevance in relation to the future data-related challenges of several disciplines. For further discussion and work, we propose a set of design principles to account for and guide a design approach for an integrated DS:

- **Abstraction** – use of abstractions and meta-level concepts to let users deal with data and data operations on a higher conceptual level. Some examples come from lazy binding data, that allows perform and propagate transformations on a set or cluster of data by changing one of its elements. Related concepts should guide UI/UX design for DS, particularly, design for non-experts;
- **Automation** – at some extension, at least of repetitive tasks and activities that data scientists need to perform (e.g. features enrichment for model training and evaluation), automation can ease the task and allow non-experts to use the system, hiding some complexity;
- **Collaboration** – increasingly, data science will be a collaborative work and co-creation and co-editing of models, plans and evaluation has to be supported;
- **Step-by-Step to Mastery** – decompose complex activities into smaller tasks and provide a step-by-step workflow, with inline help and recall support for every phase of the process. This allows non-data scientists to learn from doing and be able to accomplish the full lifecycle process, shortening the learning curve, of the tool and of the DS process itself.

These general design principles are shared with digital design and creative tools, lacking further work to operationalize and translate into specific requirements for the system.

5. Conclusions and Future Work

We present a systematic literature review (over 519 articles retrieved), define a search strategy and criteria to iteratively refine and narrow to a selected list of 214 publications regarding '*Data Science'*, for theoretical background to our current real-life DSR problem.

This paper aims at a scoping review of previous work outlining which consensus is shared across the two major themes, of IS development and HCI. The structured review already reveals DS, with ML and sophisticated algorithms, to be a fertile area of interdisciplinary inquiry and work, given the multidimensional phenomena.

With application pitches from researchers and practitioners in different disciplines, a future research agenda should increase research efforts with real life contexts, cross referencing rigor and relevance in future research projects to be carried out [40].

Improving the usability of the available tools or – more relevant and challenging – design new tools and integrated solutions for digitally support the DS lifecycle, focusing on the user needs is a research gap.

A shared framework to structure DS activities, standardizing data formats, for instance, and guidelines to address the design of an IT-solution supporting the innovation and creative dynamics of the process, particularly, when targeting non-expert's users in DS, is lacking and worth further research and practice.

Finally, evidences were found sustaining the potential of using hybrid design methods and shared research and practice perspectives on interaction, experience and service design, as promising avenues for future work.

Acknowledgements

Cofinanciado por:

CENTRO 2020 PORTUGAL 2020 UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional

The research work has been funded by the European Commission, under the Portugal 2020 structural fund (CENTRO2020), for the period of 2014-2020 (Ref. 2016/017728).

References

1. Anya, O., Moore, B., Kieliszewski, C., Maglio, P., Anderson, L.: Understanding the practice of discovery in enterprise big data science:An agent-based approach. Procedia Manuf. 3 882–889 (2015)
2. Bandi, A., Fellah, A.: Crafting a Data Visualization Course for the Tech Industry. J. Comput. Sci. Coll. 33 (2), 46–56 (2017)
3. Bibri, S.E., Krogstie, J.: Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustain. Cities Soc. 31 183–212 (2017)
4. Brunswicker, S., Bertino, E., Matei, S.: Big Data for Open Digital Innovation - A Research Roadmap. Big Data Res. 2 (2), 53–58 (2015)
5. Bumblauskas, D., Nold, H., Bumblauskas, P., Igou, A.: Big data analytics: transforming data to action. Bus. Process Manag. J. 23 (3), 703–720 (2017)
6. Cao, L.: Data Science: A Comprehensive Overview. ACM Comput. Surv. 50 (3), 1–42 (2017)
7. Carbone, A., Jensen, M., Sato, A.-H.: Challenges in data science: a complex systems perspective. Chaos, Solitons & Fractals. 90 1–7 (2016)
8. Chang, R.M., Kauffman, R.J., Kwon, Y.: Understanding the paradigm shift to computational social science in the presence of big data. Decis. Support Syst. 63 (SI), 67–80 (2014)
9. Chen, H., Chiang, R.H.L., Storey, V.C.: Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quartely. 36 (4), 1165–1188 (2012)
10. Cheng, S., Liu, B., Shi, Y., Jin, Y., Li, B.: Evolutionary Computation and Big Data: Key Challenges and Future Directions. In: Tan, Y and Shi, Y. (ed.) Data Mining and Big Data 2016. pp. 3–14. Springer International Publishing, AG, GEWERBESTRASSE 11, CHAM, CH-6330, SWITZERLAND (2016)
11. Chuprina, S., Alexandrov, V., Alexandrov, N.: Using ontology engineering methods to improve computer science and data science skills. In: Procedia Computer Science. pp. 1780–1790. (2016)
12. Conte, R., Giardini, F.: Towards Computational and Behavioral Social Science. Eur. Psychol. 21 (2), 131–140 (2016)
13. Crotty, A., Galakatos, A., Zgraggen, E., Binnig, C., Kraska, T.: The case for interactive data exploration accelerators (IDEAs). In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics - HILDA '16. pp. 1–6. ACM Press, New York, New York, USA (2016)
14. Das, M., Cui, R., Campbell, D.R., Agrawal, G., Ramnath, R.: Towards Methods for Systematic Research On Big Data. In: K, H.H. and O.B. and Z.M. and H.X. and H.L. and K.V. and R.S. and Y.S. and H.M. and L.J. and L.F. and P.S. and O. (ed.) 2015 IEEE International Conference on Big Data. pp. 2072–2081. IEEE, New York, USA (2015)
15. Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M., Brewer, S.: EDISON Data Science Framework: A Foundation for Building Data Science Profession For Research and Industry. In: 2016 8th IEEE International Conference On Cloud Computing Technology and Science (CLOUDCOM 2016). pp. 620–626. IEEE, New York, USA (2016)
16. Demirkan, H., Delen, D.: Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decis. Support Syst. 55 (1), 412–421 (2013)
17. Dichev, C., Dicheva, D., Salem, W., Salem, W., Salem, W., Salem, W.: Towards Data Science Literacy. Procedia Comput. Sci. 108 (June), 2151–2160 (2017)
18. Fischer, F., Fuchs, J., Mansmann, F., Keim, D.A.: BANKSAFE: Visual analytics for big data in large-scale computer networks. Inf. Vis. 14 (1), 51–61 (2015)
19. Fu, X., Asorey, H.: Data-Driven Product Innovation. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2311–2312. ACM, New York, USA (2015)
20. Grainger, S., Mao, F., Buytaert, W.: Environmental data visualisation for non-scientific

contexts: Literature review and design framework, http://www.sciencedirect.com/science/article/pii/S1364815216305990, (2016)
21. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Ullah Khan, S.: The rise of “big data” on cloud computing: Review and open research issues. Inf. Syst. 47 98–115 (2015)
22. Haug, F.S.: Bad Big Data Science. In: Joshi, J and Karypis, G and Liu, L and Hu, X and Ak, R and Xia, Y and Xu, W and Sato, AH and Rachuri, S and Ungar, L and Yu, PS and Govindaraju, R and Suzumura, T. (ed.) 2016 IEEE International Conference on Big Data (BIG DATA). pp. 2863–2871. IEEE, New York, USA (2016)
23. Hayashi, C.: What is Data Science ? Fundamental Concepts and a Heuristic Example. In: Data Science, Classification, and Related Methods. pp. 40–51. Springer, Tokyo (1998)
24. Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farmer, L.A.: Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. Int. J. Prod. Econ. 154 72–80 (2014)
25. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. 28 (1), 75–105 (2004)
26. Hoffman, S., Podgurski, A.: Big Bad Data: Law, Public Health, and Biomedical Databases. Journbal Law, Med. Ethics. 41 (1, SI), 56–60 (2013)
27. Horita, F.E.A.A., de Albuquerque, J.P., Marchezini, V., Mendiondo, E.M.: Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in Brazil. Decis. Support Syst. 97 12–22 (2017)
28. Howe, B., Franklin, M., Haas, L., Kraska, T., Ullman, J.: Data Science Education: We’re Missing the Boat, Again. In: 2017 IEEE 33RD International Conferenece on Data Engineering. pp. 1473–1474. IEEE, New York, USA (2017)
29. Kazakci, A.O.: Data Science as a New Frontier for Design. In: Weber, C and Husung, S and Cantamessa, M and Cascini, G and Marjanovic, D and Venkataraman, S. (ed.) ICED 15, VOL 10: Design Information and Knowledge Management. DESIGN SOC, Glasgow, England (2015)
30. Kowalczyk, M., Buxmann, P.: An ambidextrous perspective on business intelligence and analytics support in decision processes: Insights from a multiple case study. Decis. Support Syst. 80 1–13 (2015)
31. Larson, D., Chang, V.: A review and future direction of agile, business intelligence, analytics and data science. Int. J. Inf. Manage. 36 (5), 700–710 (2016)
32. Lee, I.: Big data: Dimensions, evolution, impacts, and challenges. Bus. Horiz. 60 (3), 293–303 (2017)
33. Levy, Y., Ellis, T.J.: A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. Inf. Sci. J. 9 181–212 (2006)
34. Liu, M.-C., Huang, Y.-M.: The use of data science for education: The case of social-emotional learning. Smart Learn. Environ. 4 (1), 1 (2017)
35. Loukides, M.: What is Data Science ? The future belongs to the companies and people that turn data into products. (2012)
36. Manyika, J., Chui, M., Bughin, J., Dobbs, R., Bisson, P., Marrs, A.: Big data: The next frontier for innovation, competition, and productivity | McKinsey Company. (June), 1–22 (2011)
37. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D.: Big data: the management revolution. Harv. Bus. Rev. 90 (10), 61–67 (2012)
38. Miller, H.G., Mork, P.: From data to decisions: A value chain for big data. IT Prof. 15 (1), 57–59 (2013)
39. Mockus, A.: Operational data are missing, incorrect, and decontextualized. In: Perspectives on Data Science for Software Engineering. pp. 317–322. (2016)
40. Nagle, T., Sammon, D.: The development of a Design Research Canvas for data practitioners. J. Decis. Syst. 25 (sup1), 369–380 (2016)

41. Newman, R., Chang, V., Walters, R.J., Wills, G.B.: Model and experimental development for Business Data Science. Int. J. Inf. Manage. 36 (4), 607–617 (2016)
42. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. J. Manag. Inf. Syst. 24 (3), 45–77 (2008)
43. Philip Chen, C.L., Zhang, C.-Y., Chen, C.L.P., Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf. Sci. (Ny). 275 314–347 (2014)
44. Rabl, T., Sadoghi, M., Jacobsen, H.-A., Gómez-Villamor, S., Muntés-Mulero, V., Mankowskii, S.: Solving Big Data Challenges for Enterprise Application Performance Management. Proc. VLDB Endow. 5 (12), 1724–1735 (2012)
45. Roy, S., Ray, R., Roy, A., Sinha, S., Mukherjee, G., Pyne, S., Mitra, S., Basu, S., Hazra, S.: IoT, Big Data Science & Analytics, Cloud Computing and Mobile App based Hybrid System for Smart Agriculture. In: Chakrabarti, S and Saha, H. (ed.) 8th Anual Industrial Automation and Electromechanical Engineering conference (IEMECON). pp. 303–304. IEEE, New York, USA (2017)
46. Saltz, J., Shamshurin, I., Connors, C.: Predicting data science sociotechnical execution challenges by categorizing data science projects. J. Assoc. Inf. Sci. Technol. 68 (12), 2720–2728 (2017)
47. Snyder, J.: Vernacular Visualization Practices in a Citizen Science Project. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. pp. 2097–2111. ACM, New York, NY, USA (2017)
48. Storey, V.C., Song, I.-Y.: Big data technologies and Management: What conceptual modeling can do. Data Knowl. Eng. 108 50–67 (2017)
49. Talbot, J., Lee, B., Kapoor, A., Tan, D.S.: EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1283–1292. (2009)
50. Tranfield, D., Denyer, D., Smart, P.: Towards a methodology for developing evidence-informed management knowledge by means of systematic review *. Br. J. Manag. 14 207–222 (2003)
51. Vedula, S.S., Hager, G.D.: Surgical data science: the new knowledge domain. Innov. Surg. Sci. 2 (3), 109+ (2017)
52. Webster, J., Watson, R.T.: Analyzing the Past to Prepare for the Future : Writing a Literature Review. MIS Q. 26 (2), xii–xxiii (2002)
53. Westra, B.L., Sylvia, M., Weinfurter, E.F., Pruinelli, L., Park, J.I., Dodd, D., Keenan, G.M., Senk, P., Richesson, R.L., Baukner, V., Cruz, C., Gao, G., Whittenburg, L., Delaney, C.W.: Big data science: A literature review of nursing research exemplars. Nurs. Outlook. 65 (5), 549–561 (2017)
54. Woodard, J.: Big data and Ag-Analytics: An open source, open data platform for agricultural & environmental finance, insurance, and risk. Agric. Financ. Rev. 76 (1), 15–26 (2016)
55. Zhu, Y., Xiong, Y.: Defining Data Science. CoRR. abs/1501.0 8 (2015)