http://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Articles in Advance, pp. 1–3
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

INFORMS Journal on Data Science (*IJDS*) Editorial #1: What Is an *IJDS* Paper?

Galit Shmueli,[a] Editor-in-Chief

[a] Institute of Service Science, College of Technology Management, National Tsing Hua University, Hsinchu 30013, Taiwan
Contact: galit.shmueli@iss.nthu.edu.tw, https://orcid.org/0000-0002-0820-0301 (GS)

Published Online in Articles in Advance: January 11, 2021

https://doi.org/10.1287/ijds.2020.003

I am delighted to share with you my first editorial, which is aimed at answering the multiple excellent questions I have received on editor panels and by email, Zoom calls, and more since the *INFORMS Journal on Data Science* (*IJDS*) opened its doors last month. Our diverse *IJDS* community is asking "What would make for a good *IJDS* submission? How is it different from a submission to a statistics, machine learning (ML), or other data science journal? What are data and code requirements? What about reproducibility and benchmarking?" In short, my challenging task in this editorial is to answer, "*what will an IJDS paper look like?*" I start with the key points from the Editorial Statement, describe how we aim to achieve them, and then give examples of papers to make things more concrete.

To start, the *IJDS* aims to publish important *methodological* data science contributions to *decision making* in business and engineering. This includes novel methodology, or the use of existing methodology in a completely new way, or when applied to a completely new type of problem ("methodological discovery").

Beauty Is in the Eye of the Beholder

How does this translate into practice? In a journal, the translation happens through the editorial board's understanding and taste. We have recruited a fantastic editorial board, reflecting the three communities the *IJDS* aims to serve: data scientists in business schools, in industrial engineering departments, and in industry research groups. Our editorial board is far from a random sample: it is a carefully selected team of top data science researchers in the three communities who are excited to pioneer the well-needed new *IJDS*. My first recommendation is therefore to look at the editorial board members: if you are familiar with some of our work, areas of expertise, and perhaps opinions as expressed on panels or otherwise, you will already have an idea of what the *IJDS* considers valuable "data science research."

Let me add a disclaimer at this point: Defining "an *IJDS* paper" through the editorial board's eyes is limited by our knowledge and familiarity of existing work we have seen and by our imagination. To avoid a "discovery bias," we are making a conscious effort to be open to new types of contributions, and novelties will be discussed in editorial board meetings. This means authors should be very clear on what is the paper's *contribution* to data science research. In addition, to reduce personal biases, we have created detailed review guidelines (for reviewers and for associate editors) that ask specific questions. These questions aim to tease out the contribution of a submission in a form that enables fair, transparent, and less personally tainted decisions as well as generate a clear path forward for authors when a revision is invited.

A Generic Recipe

An *IJDS* paper will have four ingredients, *data* + *methodology* + *decisions* + *implications*, but the key is the *synthesis of the ingredients*, anchored to methodology. Here is one potential approach for such synthesis that will be familiar to many in our community:

> Fired by a managerial/industrial decision-making motivation and potential/actual impact, the paper introduces an innovative data science methodology (model/algorithm/approach), applies it to data (real-world and/or simulated) to illustrate its usefulness and behavior, and considers practical (e.g., computation, implementation) and ethical (e.g., societal, environmental) implications.

The Proof of the Pudding Is in the Eating (Example Papers that Would Have Made for Great *IJDS* Publications)

To give a taste of the above recipe or of other types of formulas that would make for great *IJDS* papers, we collected several exemplars. Although these papers might be structured or formatted differently than they would have been as *IJDS* publications (e.g., the new

algorithm would not be in the appendix, we would see an extended data description or more emphasis on ethical implications), they nevertheless introduce new data science methodology for decision making:

1. In the context of *search and advertising*, in "Content-Based Model of Web Search Behavior: An Application to TV Show Search" (Liu et al. 2019), the authors were motivated by the need for improving search engine recommendations. They develop a flexible and interpretable *model* linking the user's content preferences to query search volume and click-through rates, as well as consider the context of the search (in the form of events taking place during the time of search). Their approach combines graphical models, latent factor models, and content analysis based on probabilistic topic modeling into a unified probabilistic nonnegative matrix factorization model. The authors apply their proposed methodology to *real search data* from Microsoft Bing, showing *potential impact*, and discuss the generalization of their methodology to other search engines and to other *decision-making* contexts (e.g., advertising). They also benchmark against existing methods, showing how their approach can simultaneously capture multiple types of information and investigate multiple aspects of behavioral dynamics in a single, robust modeling framework.

2. In the context of *document classification*, in "Explaining Data-Driven Document Classifications" (Martens and Provost 2014), the authors were motivated by the need of *decision makers* to understand decisions made by document classifiers (e.g., keeping ads off of objectionable web content). The authors define a new type of explanation as "a minimal set of words (terms, generally), such that removing all words within this set from the document changes the predicted class from the class of interest" (p. 73). They then develop an *algorithm* to find such explanations, as well as introduce a *framework* to assess such an algorithm's performance. The authors demonstrate the value of the new approach on a real *data set* from a company classifying web pages as adult content or not. They further illustrate their method empirically on a second data set for news-story topic classification. The authors describe the *implications* of using their method (p. 96): "(1) the manager using it understands how decisions are being made, (2) the customers affected by the decisions can be advised why a certain action regarding them is taken, and (3) the data science/development team can improve the model iteratively. Further, (4) document classification explanations can provide better understanding of the business domain."

3. In the context of decision making in *healthcare*, "Machine Learning Approaches for Early DRG Classification and Resource Allocation" by Gartner et al. (2015) incorporates machine learning classification into a mixed-integer programming (MIP)-based resource allocation model for allocating scarce hospital resources. The authors apply their combined *ML–MIP methodology* to *real data* from a hospital, showing improved diagnosis-related group (DRG) classification accuracy and *practical impact* measures (e.g., utilization rate of operating rooms) as compared with the hospital's current approach. The authors highlight the importance of selecting a concise set of relevant attributes at all stages of care, and also emphasize the generalizability of their approach to similar DRG systems (e.g., in Germany).

4. In the context of *transportation and sharing economy* decision making, in "Predicting Urban Dispersal Events: A Two-Stage Framework Through Deep Survival Analysis on Mobility Data" (Vahedian et al. 2019), the authors were motivated by the need for early prediction of urban dispersal events—processes where an unusually large number of people leave the same area in a short period—for mitigating congestion and safety risks and making better dispatching decisions for taxi and ride-sharing fleets. Asserting that dispersal events follow a complex pattern of past trips and other related features, the authors formulate the prediction problem as a survival analysis problem and propose a *two-stage framework* (DILSA; DIspersaL event prediction using Survival Analysis) that combines deep learning with survival analysis to predict the probability of a dispersal event and its demand volume. They then evaluate their framework by applying it to a real data set on New York City yellow taxis between 2014 and 2016, and find that it outperforms multiple benchmark models in terms of prediction accuracy.

5. In the context of *complex systems and manufacturing*, in "Ensemble-Bayesian SPC: Multi-mode Process Monitoring for Novelty Detection," Bacher and Bengal (2017) introduce a statistical process monitoring method for detecting novelties in systems with a large number of sensors or attributes ("multi-mode systems"). In such systems, new data patterns might emerge that have not been seen before ("novelties"), which are crucial for *decision making*. The authors propose an ensemble-Bayesian statistical process control (SPC) framework for modeling the known operating modes by categorizing their corresponding observations into data classes that are detected during the training stage and exploiting the joint information of the trained classes. The authors compare their approach to anomaly detection methods both conceptually and empirically on a set of 20 public *data sets* from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml). They discuss implications of algorithm choices, computational limitations, and generalization to metrology and the semiconductor industry.

6. In the context of *energy and manufacturing*, in "End of Performance Prediction of Lithium-Ion Batteries" (Wang et al. 2019), the authors were motivated

by the need to improve prediction of rechargeable battery lifetime, a challenge to manufacturers of portable electronics and electric vehicles. The authors propose an accelerated testing version of the trend-renewal process *model*, a type of a time-transformed renewal process that accounts for temporal structure. They apply their proposed model to a *real data set* on accelerated testing of nine lithium-ion batteries and to the NASA battery data set used in prior studies, comparing it against existing prediction and time series methods. They show their model outperforms the other methods for predicting battery end-of-performance.

7. In the context of *behavioral big data*, in "A Tree-Based Approach for Addressing Self-Selection in Impact Studies with Big Data" (Yahav et al. 2016), the authors were motivated by the challenge of business and policy *decision making* in impact studies that rely on large-scale experiments that are not randomized (quasi experiments). The authors introduce a tree-based *approach* suitable for quasi experiments, observational studies, and postanalysis of randomized experimental data. Their approach provides a stand-alone, automated, data-driven methodology that outperforms traditional propensity score methods. The authors illustrate the method and the insights it yields in the context of three real *data sets* from impact studies, comparing it against propensity score approaches. They also use a large simulation to evaluate performance under different scenarios, and discuss computational *implications* and new insights the method can generate.

The editorial board has identified several additional exemplars:

- "Unique in the Crowd: The Privacy Bounds of Human Mobility" (De Montjoye et al. 2013)
- "Explaining Data-Driven Decisions Made by AI Systems: The Counterfactual Approach" (Fernandez et al. 2020)
- "Personalized Market Basket Prediction with Temporal Annotated Recurring Sequences" (Guidotti et al. 2018)
- "Power Curve Estimation with Multivariate Environmental Factors for Inland and Offshore Wind Farms" (Lee et al. 2015)
- "Scalable and Accurate Deep Learning with Electronic Health Records" (Rajkomar et al. 2018)
- "Active Feature-Value Acquisition" (Saar-Tsechansky et al. 2009)
- "Copycats vs. Original Mobile Apps: A Machine Learning Copycat-Detection Method and Empirical Analysis" (Wang et al. 2018)
- "Real-Time Monitoring of High-Dimensional Functional Data Streams via Spatio-temporal Smooth Sparse Decomposition" (Yan et al. 2018)
- "Mind the Gap: Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining" (Yang et al. 2018)
- "Accurate Estimation of Influenza Epidemics Using Google Search Data via ARGO" (Yang et al. 2015)

I hope these examples provide you (our author, reviewer, and reader) a better understanding of the excellent type of research that we seek at the *IJDS*. I emphasize our openness to and welcoming of diverse data science methods, ideas, and approaches for tackling problems in a variety of decision-making contexts. We anticipate such state-of-the-art publications will be highly relevant and valuable to the data science research and industry communities, and will therefore make the *IJDS* your top choice.

References

Bacher M, Ben-Gal I (2017) Ensemble-Bayesian SPC: Multi-mode process monitoring for novelty detection. *IISE Trans.* 49(11):1014–1030.

De Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: The privacy bounds of human mobility. *Sci. Reports* 3:1376.

Fernandez C, Provost F, Han X (2020) Explaining data-driven decisions made by AI systems: The counterfactual approach. Preprint, submitted January 21, https://arxiv.org/abs/2001.07417.

Gartner D, Kolisch R, Neill DB, Padman R (2015) Machine learning approaches for early DRG classification and resource allocation. *INFORMS J. Comput.* 27(4):718–734.

Guidotti R, Rossetti G, Pappalardo L, Giannotti F, Pedreschi D (2018) Personalized market basket prediction with temporal annotated recurring sequences. *IEEE Trans. Knowledge Data Engrg.* 31(11):2151–2163.

Lee G, Ding Y, Genton MG, Xie L (2015) Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *J. Amer. Statist. Assoc.* 110(509):56–67.

Liu J, Toubia O, Hill S (2019) Content-based model of web search behavior: An application to TV show search. Preprint, submitted May 7, http://dx.doi.org/10.2139/ssrn.3372751.

Martens D, Provost F (2014) Explaining data-driven document classifications. *MIS Quart.* 38(1):73–100.

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, et al. (2018) Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 1(1):18.

Saar-Tsechansky M, Melville P, Provost F (2009) Active feature-value acquisition. *Management Sci.* 55(4):664–684.

Vahedian A, Zhou X, Tong L, Street WN, Li Y (2019) Predicting urban dispersal events: A two-stage framework through deep survival analysis on mobility data. *Proc. AAAI Conf. Artificial Intelligence*, vol. 33 (Association for the Advancement of Artificial Intelligence, Menlo Park, CA), 5199–5206.

Wang Q, Li B, Singh PV (2018) Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis. *Inform. Systems Res.* 29(2):273–291.

Wang YF, Tseng ST, Lindqvist BH, Tsui KL (2019) End of performance prediction of lithium-ion batteries. *J. Quality Tech.* 51(2):198–213.

Yahav I, Shmueli G, Mani D (2016) A tree-based approach for addressing self-selection in impact studies with big data. *MIS Quart.* 40(4):819–848.

Yan H, Paynabar K, Shi J (2018) Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition. *Technometrics* 60(2):181–197.

Yang M, Adomavicius G, Burtch G, Ren Y (2018) Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Inform. Systems Res.* 29(1):4–24.

Yang S, Santillana M, Kou SC (2015) Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc. Natl. Acad. Sci. USA* 112(47):14473–14478.