Contornos da ciência de dados: uma meta-revisão conceitual

Daniel Brito dos Santos, Annabell Del Real Tamariz

Em 1962, John Tukey previu e conclamou uma nova ciência. Sessenta anos mais tarde, a ciência de dados é parte essencial de muitas indústrias. Além de ser o principal produto de muitas das empresas mais valiosas do mundo como Google e Amazon. Portanto, expandir e aprofundar a ciência de dados se reflete na ampliação das possibilidades e efetividade de negócios, pesquisas, tomadas de decisão e organização social. Entretanto, apesar do seu rápido desenvolvimento, os desafios apontados por Tukey permanecem centrais: formação de mão-de-obra qualificada, padrões de projeto, comunicação entre profissionais de diferentes áreas, e até mesmo consensos de suas próprias delimitações. Nesse sentido, é interessante compreender os esforços atuais na direção de estruturar o campo de modo a construirmos uma estrutura teórica que fundamente tanto essa quanto as próximas pesquisas. Desse modo, o presente trabalho buscou definir o que é ciência de dados e quais são suas principais técnicas, conceitos e ferramentas. Para tanto, efetuamos uma revisão sistemática de revisões. Buscamos artigos do tipo "Review" contendo o termo "data science". Analisamos os títulos e abstracts dos 100 primeiros resultados para selecionarmos aquelas que tinham por objetivo definir ciência de dados. Finalmente efetuamos uma análise comparativa dos 4 artigos selecionados para responder nossas perguntas de pesquisa. Dessa forma, podemos concluir que a Ciência de Dados se cristalizou a partir dos scripts computacionais de softwares numéricos como a linguagem R. A partir desses softwares, rotinas de coleta, processamento e análise de dados puderam ser compartilhadas, reproduzidas e escrutinadas com inédita precisão. Assim, Ciência de Dados é o domínio que se ocupa dos problemas derivados de todo esse processo fundamentalmente interdisciplinar, desde a definição de um problema respondível até a implementação de uma solução computacional, interpretação e comunicação dos resultados. Nesse sentido, Donoho oferece uma estrutura chamada Greater Data Science (GDS) para organizar a ciência de dados em seis diferentes áreas de estudo e aplicação: 1. Coleta, preparação e exploração; 2. Representação e transformação; 3. Computação de dados; 4. Modelagem de dados; 5. Visualização e apresentação; 6. Ciência sobre Ciência de Dados. Analogamente, Crisan, Fiore-Gartland e Tory construíram um modelo de trabalho em quatro macro processos: 1. Preparação; 2. Análise 3. Implantação; 4. Comunicação. Assim, a partir dessa elucidação, foi possível direcionar os próximos aprofundamentos ao evidenciar as frentes de pesquisa em potencial que cada uma dessas sub-áreas representam.

Instituição do Programa de IC: UENF
Fomento da bolsa: PIBIC-UENF

Data science contours: a conceptual meta-review

Daniel Brito dos Santos, Annabell Del Real Tamariz

In 1962, John Tukey predicted and called for a new science. Sixty years later, data science is an essential part of many industries, as well as being the main product of many of the most valuable companies in the world such as Google and Amazon. Therefore, expanding and deepening data science is reflected in expanding the possibilities and effectiveness of business, research, decision-making and social organization. However, despite its rapid development, the challenges identified by Tukey remain central: training of qualified labor, design standards, communication between professionals from different backgrounds, and even consensus on their own boundaries. In this sense, it is interesting to understand the current efforts towards structuring the field in order to build a theoretical framework that supports both this and future research. In this way, the present work sought to define what data science is and what are its main techniques, concepts and tools. To this end, we performed a systematic review of reviews. We searched for articles of the "Review" type containing the term "data science". We analyzed the titles and abstracts of the first 100 results to select those that aimed to define data science. Finally, we performed a comparative analysis of the 4 articles selected to answer our research questions. Thus, we can conclude that Data Science was crystallized from the computational scripts of numerical software such as the R language. From these software, routines of data collection, processing and analysis could be shared, reproduced and scrutinized with unprecedented precision. Thus, Data Science is the domain that deals with the problems derived from this fundamentally interdisciplinary process, from the definition of an answerable problem to the implementation of a computational solution, interpretation and communication of the results. In this sense, Donoho offers a framework called Greater Data Science (GDS) to organize data science into six different areas of study and application: 1. Collection, preparation and exploration; 2. Representation and transformation; 3. Data Computing; 4. Data Modeling; 5. Visualization and presentation; 6. Science about Data Science. Analogously, Crisan, Fiore-Gartland and Tory built a working model in four macro processes: 1. Preparation; 2. Analysis 3. Implementation; 4. Communication. Thus, from this elucidation, it was possible to direct the next deepening by highlighting the potential research fronts that each of these sub-areas represent.

Institution of the Program IC: UENF
Grant promotion: PIBIC-UENF