3DSE: Uma Nova Abordagem para Avaliação de Projetos de Ciência de Dados

Rejane C. Oliveira[1], **Bruno Laporais Pereira**[1]

[1]Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte – MG – Brazil

re.ol.2015@gmail.com, laporaisbruno@gmail.com

Abstract. *The remarkable interest in data science projects is due to impacts and recent studies of the area that aims to extract knowledge from data collected to support decision-making. These projects can be characterized by technical aspects of the problem they address or by the way they were structured, based on methodologies and best practices to their development and subsequent communication. In this paper, we introduce the 3DSE: a model for Data Science project evaluation that allows a measurement of their complexity and structure, supported by a graphical representation for the visualization of the results and comparison of projects. The evaluation metrics and the visual representation have potential use in the characterization and evaluation of portfolios of projects.*

Resumo. *O notável interesse em projetos de Ciência de Dados se deve aos impactos e recentes estudos da área que objetiva extrair conhecimento dos dados coletados para subsidiar a tomada de decisão. Esses projetos podem ser caracterizados por aspectos técnicos do problema que abordam ou pela forma com que foram estruturados, baseados em metodologias e boas práticas para seu desenvolvimento e comunicação posterior. Neste trabalho, nós introduzimos o 3DSE, um modelo de avaliação para projetos de Ciência de Dados que permite a mensuração de sua complexidade e estruturação, com apoio de uma representação gráfica para a visualização dos resultados e comparação de projetos. As métricas de avaliação e sua representação visual tem utilização potencial na caracterização e avaliação de portfólios de projetos.*

1. Introdução

O atual sucesso das organizações orientadas a dados[1] exige ser capaz de pensar sobre como alguns conceitos fundamentais do pensamento analítico se aplicam aos problemas de negócios em questão. Com o objetivo de extrair conhecimento dos dados coletados, há uma extensa coleção de técnicas de mineração de dados, algoritmos e ferramentas que vem sendo largamente utilizadas, juntamente com princípios básicos e conceitos que fundamentam essas técnicas e o pensamento sistemático que promove o sucesso na tomada de decisão. Esta área de conhecimento é nomeada como "Ciência de Dados" [Provost and Fawcett 2013].

Em Ciência de Dados, um projeto é descrito como uma atividade multidisciplinar que utiliza técnicas de diversas áreas do conhecimento, como estatística, matemática,

[1]do inglês *data-driven.*

computação, sobre conjuntos de dados estruturados e/ou não-estruturados, originados por diversas fontes como por exemplo sistemas, pessoas e sensores [Saltz et al. 2017]. O sucesso de um projeto de Ciência de Dados envolve a atenção para muitos aspectos durante seu desenvolvimento, com vistas à sua comunicação efetiva e reusabilidade. Entretanto, aprofundar e avaliar tais aspectos em projetos de Ciência de Dados é uma tarefa custosa e muitas vezes subjetiva, por não existir até o momento uma metodologia consolidada que possa servir como referência para isso. A proposta deste trabalho é construir um novo modelo que permita suportar essa avaliação, além de introduzir uma intuitiva representação visual, e demonstrar sua aplicação na avaliação de portfólios de projetos relacionados à área.

Esse artigo está assim organizado: a Seção 2 realiza um levantamento das características comumente observadas em projetos de Ciência de Dados, e de seus desafios para implantação, a partir da revisão da literatura. A Seção 3 descreve o modelo de avaliação *3DSE* proposto para representar as dimensões que caracterizam projetos desta natureza e sua representação visual. A Seção 4 apresenta a aplicação do modelo para um exemplo de portfólio de projetos. A Seção 5 conclui e apresenta futuras direções deste estudo.

2. Projetos de Ciência de Dados: Aspectos e Estrutura

Os projetos de Ciência de Dados podem ser caracterizados por aspectos técnicos do problema que abordam ou pela forma com que foram estruturados, baseados em metodologias e boas práticas para seu desenvolvimento e comunicação.

O trabalho de [Saltz et al. 2017] busca identificar quais são as características-chave que melhor descrevem projetos de Ciência de Dados, e inova em apresentar como tais características podem ser integradas de modo a construir um modelo geral que possibilite escolher a melhor maneira de gerenciar os projetos desta natureza. Segundo estes autores, não é suficiente a tradicional descrição dos projetos por meio dos "4 V's", isto é, com base nas características intrínsecas aos dados: Volume (tamanho dos dados/ necessidade de se utilizar técnicas de Big Data), Variedade (número de fontes e tipo dos dados – estruturado/não-estruturado), Velocidade (velocidade da geração/coleta dos dados que necessitam ser analisados) e Veracidade (confiabilidade dos dados). Os autores argumentam que há ainda outros aspectos que caracterizam os projetos de Ciência de Dados, como a incerteza quanto à entrada/saída da solução (quais dados são relevantes e o que se poderá descobrir com tais dados - isto é, o contexto analítico), bem como os desafios durante seu desenvolvimento. Os autores propõem a inclusão de aspectos como tamanho da organização, tamanho das equipes e sua virtualização, e cultura da organização (se é orientada a dados, se tem foco em retorno do investimento ou P&D, entre outros).

Uma abordagem complementar é a descrita em [Rybicki 2018], que investiga as melhores práticas para estruturação de projetos de Ciência de Dados, utilizadas no meio acadêmico e/ou na indústria. O autor argumenta que a estrutura de um projeto de Ciência de Dados deve ser vista também como mais um dos recursos de comunicação das descobertas decorrentes do projeto, facilitando assim sua reprodutibilidade e uso posterior, o que seria indicativo de sua boa qualidade. Ele descreve como boas práticas em projetos de Ciência de Dados a existência dos seguintes aspectos: o uso de alguma metodologia para seu desenvolvimento como KDD (*Knowledge Discovery in Databases*), CRISP-DM (*Cross-industry Standard Process for Data Mining*), TDSP (*Team Data Science Process*);

controle de versão; um *workflow description* - isto é, uma forma de fazê-lo funcionar de ponta-a-ponta, e a própria hierarquia de diretórios e subdiretórios de arquivos [2].

Grande parte da complexidade dos projetos de Ciência de Dados e Big Data residem sobre os dados e os fluxos de dados (*pipelines*), abordada em vários estudos [Sculley et al. 2014, Garg et al. 2016, Yu et al. 2005, Eichelberger et al. 2017]. [Sculley et al. 2014] destacam que, em projetos de Aprendizado de Máquina, a dependência de dados é fonte importante de débito técnico que não deve ser ignorada e custa mais do que a dependência de código. Ainda segundos os autores, os *pipelines* desenvolvidos para a etapa de preparação de dados também são difíceis de gerenciar, assim como detecção de erros, recuperação de falhas, e testes de usabilidade e automatizados, capazes de cobrir requisitos sobre o código e principalmente sobre os dados. Outro aspecto a ser considerado é a qualidade do código, sendo possível para isto o uso de ferramentas específicas[3]. A atenção para testes sobre os dados e etapas do processamento dos dados em parte atende a outra questão relevante para os projetos de Ciência de Dados: a dificuldade da reprodutibilidade em experimentos em ciência computacional [Baker 2016].

Por fim, os projetos de Ciência de Dados devem atentar para aspectos relativos à segurança dos dados e, em especial, do que determina as recentes iniciativas de legislação de proteção de dados, incluindo dados pessoais e sensíveis, isto é, aqueles relacionados à identificação do indivíduo e que porventura possam ser utilizados para sua discriminação (Lei Geral de Proteção de Dados Pessoais – LGPDP - Lei Nº 13.709, de 14 de Agosto de 2018, Brasil, 2018, e *General Data Protection Regulation* – GDPR, na União Europeia, que entrou em vigor em 25 de maio de 2018).

3. Modelo de avaliação de projetos de Ciência de Dados (*3DSE)*

3.1. Métricas para avaliação de projetos de Ciência de Dados

O modelo aqui proposto - *3DSE: 3-Dimensional Data Science Evaluation* - busca caracterizar um projeto de Ciência de Dados de maneira a contemplar aspectos tanto da complexidade intrínseca a este, da abordagem realizada para seu desenvolvimento e implementação (estruturação), e também da área de conhecimento a que pertence.

Com referência ao trabalho de [Saltz et al. 2017], utilizamos o eixo de complexidade descrito pelas seguintes características: volume, variedade, velocidade, e veracidade dos dados, a intensidade computacional para pré-processamento e/ou modelagem, além do tipo de análise requerida (geração de hipóteses -"escopo aberto", ou teste de hipóteses - "escopo fechado", ou ambas as situações -"escopo aberto com teste de hipóteses"). Nosso modelo acrescenta a segurança dos dados (dados pessoais, sensíveis, etc) como mais um critério associado à complexidade do projeto.

A segunda dimensão contemplada no nosso modelo de avaliação é a estruturação do projeto, que corresponde ao uso de boas práticas para o seu desenvolvimento [Rybicki 2018]. Um bom projeto de Ciência de Dados deve contar com o controle de versão dos dados, *pipeline* para as fases de preparação dos dados, desenvolvimento de testes sobre esta fase, bem como testes sobre os dados, além de garantias para a reprodutibilidade dos experimentos e qualidade do código.

[2]Disponível em https://github.com/drivendata/cookiecutter-data-science

[3]Exemplos em https://bettercodehub.com/; https://www.codacy.com/

Por fim, acrescentamos uma terceira dimensão neste modelo de avaliação de projetos: a área do conhecimento principal a que pertence. De acordo com a nomenclatura do documento *Data Science Framework* proposto pelo projeto EDISON [Demchenko et al. 2016], as áreas do conhecimento podem ser separadas em: *Data Analytics; Data Science Engineering; Data Management; Scientific or Research Methods; Business Process Management*; e *Data Science Domain Knowledge*.

A Tabela 1 resume os critérios de avaliação e propõe uma escala numérica para os níveis de cada critério das duas dimensões principais. Um peso maior foi atribuído aos itens considerados mais críticos para o sucesso do projeto.

Tabela 1. Critérios para Modelo de Avaliação de Projetos de Ciência de Dados

(A) Complexidade

Atributo / Nota	**1**	**2**	**3**
Dados – Volume	GB	TB	-
Dados – Variedade	Estruturados	Não estruturados	Estruturados + Não Estruturados
Dados – Velocidade	Mini Batch	Batch	Near Real-time
Dados – Veracidade	Organizados (Tidy)	Organizados (Tidy(-))	Desorganizados (Messy)
Intensidade computacional	Irrelevante	Relevante em: pré-processamento ou modelagem	Relevante em: pré-processamento e modelagem
Tipo de análise	Escopo fechado	Escopo aberto	Escopo aberto com teste de hipóteses
Segurança dos dados pessoais/sensíveis	-	-	Presente

(B) Estrutura

Atributo/ Nota	**0**	**1**	**2**
Metodologia para desenvolvimento	não	sim	-
Controle de versão de dados	não	-	sim
Pipeline para construção do Data Lake	não	-	sim
Desenvolvimento de testes (sobre dados)	não	sim	-
Métrica de qualidade do código	não	sim	-
Organização dos dados, scripts, resultados, documentação	não	sim	-
Reprodutibilidade de experimentos	não	-	sim

3.2. Representação Gráfica para Avaliação de Projetos de Ciência de Dados

Uma representação gráfica é proposta como recurso para uma visualização direta e adequada dos projetos caracterizados pelo *3DSE*, com objetivo de auxiliar a compreensão dos dados, a partir da substituição de cálculos cognitivos com simples inferências perceptuais [Heer et al. 2010]; FEW, 2006), processo também conhecido como "percepção gráfica" [Cleveland and McGill 1985].

Neste trabalho utilizamos uma representação gráfica posicional, em quem os pontos indicam a correlação entre as duas dimensões que se quer representar: "complexidade" (eixo horizontal) e "estruturação" (eixo vertical), sendo o espaço dividido em quatro quadrantes, possibilitando a leitura das seguintes subdivisões, cada qual representando um tipo de projeto: BREVE, ESTRUTURADO, COMPLEXO e ROBUSTO (Figura 1). As áreas de conhecimento são apresentadas em diferentes cores e símbolos para facilitar a compreensão.

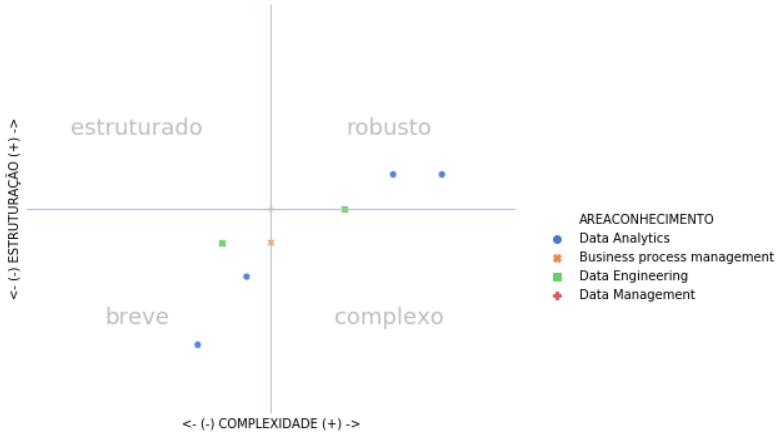

Figura 1. Representação gráfica para Avaliação de Projetos de Ciência de Dados

4. Aplicação do modelo 3DSE

A Figura 1 é um exemplo de aplicação do modelo de avaliação de projetos de Ciência de Dados. As métricas de avaliação segundo a Tabela 1 foram aplicadas para oito projetos (adaptado de [Saltz et al. 2017]) e representadas graficamente. Os oito projetos se posicionam ao longo do espaço que vai do quadrante "BREVE" ao quadrante "ROBUSTO", pelo incremento tanto de complexidade quanto de estruturação. Há um distinto contraste entre o projeto representado no extremo canto inferior esquerdo e o projeto mais distante do canto superior direito: o primeiro representa um projeto com baixa velocidade de geração de dados estruturados e limpos da ordem de Gigabytes, escopo fechado de análise (teste de hipóteses definido) e intensidade computacional pouco relevante nas várias etapas (pré-processamento e modelagem); já o último projeto representa um caso de dados estruturados e não-estruturados, gerados em alta velocidade, de maneira desorganizada (*messy*), em volume de TeraBytes, com análise de escopo aberto (geração e teste de hipóteses), e intensidade computacional importante na fase de pré-processamento. Certamente este último projeto necessita de uma solução mais "robusta" para a confiabilidade dos resultados, mostrando a capacidade do nosso modelo para representação de projetos com características tão dessemelhantes.

5. Conclusões

Neste trabalho foi proposto o modelo *3DSE* para representação das características de projetos de Ciência de Dados. O modelo apresentado possui clareza e profundidade ao introduzir métricas ainda não exploradas pela literatura de forma combinada. Sobreposto a esse modelo, foi criada uma representação gráfica capaz de reproduzir tais características em uma visualização simples e objetiva. A utilização da métrica de avaliação e sua representação visual permitem sua aplicação sobre portfólios para comparação de projetos e conduzem a uma melhor avaliação e compreensão dos desafios das soluções em Ciência de Dados. Como trabalhos futuros, propomos explorar a criação de uma métrica unificada e condução de novos experimentos com 3DSE sobre diversos outros portfólios públicos, a fim de comparar tais perfis e consolidar o modelo aqui proposto.

Referências

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452.

Cleveland, W. S. and McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833.

Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M., and Brewer, S. (2016). Edison data science framework: a foundation for building data science profession for research and industry. In *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 620–626. IEEE.

Eichelberger, H., Qin, C., and Schmid, K. (2017). Experiences with the model-based generation of big data pipelines. *Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband*.

Garg, N., Singla, S., and Jangra, S. (2016). Challenges and techniques for testing of big data. *Procedia Computer Science*, 85:940–948.

Heer, J., Bostock, M., Ogievetsky, V., et al. (2010). A tour through the visualization zoo. *Commun. Acm*, 53(6):59–67.

Provost, F. and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59.

Rybicki, J. (2018). Best practices in structuring data science projects. In *International Conference on Information Systems Architecture and Technology*, pages 348–357. Springer.

Saltz, J., Shamshurin, I., and Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. *Journal of the Association for Information Science and Technology*, 68(12):2720–2728.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., and Young, M. (2014). Machine learning: The high interest credit card of technical debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*.

Yu, L., Wang, S., and Lai, K. K. (2005). An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering*, 18(2):217–230.