MARIANA VITTI RODRIGUES

IMPLICAÇÕES EPISTEMOLÓGICAS DO USO DE MASSIVA QUANTIDADE DE DADOS NO PROCESSO DE DESCOBERTA CIENTÍFICA

Relatório de Pós-doutorado realizado na Universidade Estadual Paulista (UNESP), Faculdade de Filosofia e Ciências, Marília/SP

Supervisora:
Profa. Dra. Maria Eunice Quilici Gonzalez

FAPESP n. 2020/03134-1

Marília
2024

Faculdade de Filosofia e Ciências
UNESP – Campus de Marília

Projeto de pesquisa de pós-doutorado

Implicações epistemológicas do uso de massiva quantidade de dados no processo de descoberta científica

Resumo: O objetivo do presente projeto é investigar possíveis implicações epistemológicas da aplicação de massiva quantidade de dados, denominada *Big Data*, no processo de descoberta científica. O problema central que direciona nossa investigação pode ser assim formulado: Quais são as possíveis implicações epistemológicas do uso de técnicas de análise de *Big Data* no processo de descoberta científica? Dessa questão central decorrem duas questões específicas: (i) Em que medida aspectos semânticos e pragmáticos presentes no processo de descoberta científica podem ser automatizados através do emprego de técnicas de análise de Big Data? (ii) Que critérios estão envolvidos na escolha de relações (causais e/ou correlacionais) presentes em massiva quantidade de dados disponíveis para análise científica? Na investigação dessas questões, apresentaremos o debate contemporâneo sobre implicações epistemológicas do emprego de técnicas de análise de Big Data na investigação científica. Em seguida, elucidaremos os conceitos de dado, informação e significado no contexto da filosofia da informação, com ênfase nos trabalhos de Charles S. Peirce, Fred Dretske e Sabina Leonelli. Analisaremos também o conceito de causalidade indicando proximidades entre relações causais e informacionais na pesquisa científica. Finalmente, propomos uma análise do processo de descoberta científica na era dos Big Data, ressaltando o papel da informação nos raciocínios abdutivo e diagramático. Exemplos ilustrativos de descoberta científica na era do Big Data serão analisados no decorrer da pesquisa.

Palavras-chave: Abdução, Big Data, informação, descoberta científica, semiótica.

Relatório Final

FAPESP Processo n. 2020/03134-1

1. Introdução

O presente relatório, referente ao período de 14 de junho de 2022 a 31 de agosto de 2024, apresenta os resultados finais da investigação sobre possíveis implicações epistemológicas da aplicação de técnicas de *Big Data* no processo de *descoberta científica*. Parte da pesquisa foi desenvolvida na Universidade de Exeter com o projeto de Bolsa Estágio de Pesquisa no Exterior intitulado "Descoberta por serendipidade e abdução no contexto das práticas de ciência aberta", realizado no período de 01 de maio de 2023 a 30 de abril de 2024. As atividades realizadas neste período estão descritas no relatório em anexo.

Com o desenvolvimento de novas técnicas computacionais, a possibilidade de detecção de *padrões estatísticos* em massivas bases de dados motivou alguns estudiosos de Big Data a declararem o fim da necessidade de teorias científicas baseadas em hipóteses explicativas (Anderson, 2008). Entretanto, como explicitado no projeto, o desafio epistemológico envolvido no processo de descoberta científica diz respeito à possibilidade de reconhecimento de padrões que se tornam *relevantes* em um determinado contexto (Floridi, 2012). No decorrer da pesquisa, focalizamos nossa análise no processo dinâmico e complexo de fluxo de dados que possibilita o desenvolvimento e uso de massiva quantidade de dados, inspiradas na noção de jornada de dados (Leonelli 2015). As jornadas de dados envolvem o processo de descontextualização dos dados de sua origem local, de modo que eles se tornem digitais e integrados em conjuntos de dados estruturados ou semiestruturados, possibilitando processos de reutilização, reanálise, recombinação, reaproveitamento e reposicionamento de dados (Leonelli, 2014, 2015; Collmann et al., 2016). Diferentes arquiteturas algorítmicas são elaboradas para lidar com diferentes tipos de dados, às vezes dentro de uma estrutura específica em um domínio de pesquisa limitado. Os modelos algorítmicos para análise de dados, por sua vez, podem ser usados para diferentes fins em diversas áreas de especialização. Deste modo, modelos algorítmicos de análise de dados podem ser caracterizados como *ferramentas de pesquisa* que visam a facilitar a detecção de padrões latentes em estruturas de dados existentes.

Entendidos como instrumentos de pesquisa, modelos algorítmicos devem possuir formas de avaliação de sua eficiência, escopo e limites (Alvarado 2023a). Ademais, o crescente uso de modelos de *deep learning* na prática científica tem gerado discussões sobre o tipo de confiança

que se pode atribuir a resultados obtidos por meio de algoritmos opacos (Humphreys 2016; Alvarado, 2023b). Diferente da opacidade epistêmica dos instrumentos científicos, em que a confiança dos pesquisadores advém do fato de que engenheiros e técnicos conhecem todos os seus recursos relevantes, o uso de algoritmos de *machine* e *deep learning* pode ser inacessível para qualquer pessoa devido à sua complexidade inerente. Como fica a descoberta científica no processo de jornada de dados em que algoritmos opacos vêm sendo cada vez mais empregados como forma de detectar padrões em massivas quantidades de dados?

Em nossa investigação, partimos da noção de ***descoberta científica*** caracterizada como um processo de desvelamento (des-cobrimento) de *padrões disposicionais* que expressem possíveis propriedades do objeto de inquirição (Hanson 1958a, 1958b, 1965; Gonzalez, 1984; Paavola, 2012). Inspiradas pela semiótica peirciana, entendemos *padrões disposicionais* como expressão de hábitos adquiridos (de curta ou longa duração) que são incorporados no objeto de inquirição e expressos através da ação do signo (caracterizado como uma representação parcial do objeto que o determina tendo potencial de geração de signos mais desenvolvidos, ou interpretantes). Compreendemos que padrões disposicionais podem ser desvelados por um esforço autocrítico e autocontrolado de pensamento coletivo através do processo de inquirição constituído na dinâmica dos raciocínios abdutivo, dedutivo e indutivo.

Ênfase foi conferida ao raciocínio abdutivo por ser considerado um tipo de inferência ampliativa que possibilita a emergência de novas ideias através da experimentação com diagramas, e da formulação de hipóteses explicativas de eventos anômalos. Entendemos que a detecção de padrões relevantes por abdução compreende aspectos semânticos e pragmáticos que fornecem graus de direcionalidade à pesquisa científica através da elaboração de cenários hipotéticos que indicam caminhos razoáveis de investigação. Analisamos, também, processos de descoberta que requerem a *criação* de estratégias metodológicas, o desenvolvimento de novos aparatos tecnológicos de processamento de informação, e da elaboração de hipóteses explicativas que, em uma dinâmica coletiva da prática científica, promovem a ampliação do conhecimento. Nesse cenário, questionamos em que medida algoritmos de inteligência artificial, em especial os que envolvem técnicas de análise de Big Data, possibilitam o reconhecimento da relevância de padrões detectados em massivas bases de dados contribuindo para a aquisição de valor epistêmico orientado de acordo com certos objetivos de pesquisa.

Para ilustrar nossa análise, investigamos o desenvolvimento, uso e impacto de softwares projetados para predizer e classificar proteínas. Especial ênfase foi conferida ao desenvolvimento do software SignalP 6.0 projetado para predizer e classificar sinais peptídicos, i.e., uma sequência hidrofóbica no terminal N de uma cadeia polipeptídica responsável por

direcionar a secreção de proteína (Cooper 2000). Neste estudo de caso, analisamos as etapas das jornadas de dados apreciando qual seria o papel da abdução na elaboração e uso do software. Entendemos que os resultados obtidos a partir do uso de modelos algorítmicos de análise de dados devem ser compreendidos como hipóteses que podem ser reconhecidas como frutíferas, mas que, primordialmente, possuem um caráter provisório que exige cautela por parte dos investigadores em sua interpretação.

Como um segundo estudo de caso, analisamos o software denominado AI-Descartes, um sistema de Inteligência Artificial que combina raciocínio lógico com regressão simbólica, projetado para obter descobertas científicas a partir de conhecimento axiomático e dados experimentais (Cornelio et al. 2023). Investigamos a possibilidade de automação do raciocínio abdutivo por meio de estrutura de programação lógica. Concluímos, provisoriamente, que se compreendermos raciocínio abdutivo sob a perspectiva peirciana, se iniciando com o sentimento de surpresa e se desenvolvendo na busca de hipóteses explicativas que, se verdadeiras, dissipariam esse tipo de sentimento, abdução não seria passível – ainda – de automatização.

Nossa investigação foi de encontro com afirmações feitas pelos entusiastas dos chamados ‘big data’ que julgam que técnicas de análise de dados sofisticadas anunciariam o fim da investigação científica. Buscamos mostrar como a complexa e dinâmica prática científica exige processos de escolhas nos quais a incerteza, ambiguidade e dúvida incitam o raciocínio abdutivo o que demanda, por sua vez, o estabelecimento de estratégias heurísticas que são contexto-dependente. Cada estratégia depende da área de atuação e de critérios epistêmicos específicos que possibilitam a emergência de informação significativa em um dado contexto.

Em síntese, destacamos as seguintes implicações epistemológicas da aplicação de massiva quantidade de dados no processo de descoberta científica: (i) possibilidade de novas estratégias heurísticas para desvelamento de relações entre variáveis de interesse; (ii) possibilidade de projetar algoritmos para estimular descoberta por serendipidade; (iii) verticalização da prática científica em que o desenvolvimento de algoritmos é altamente *dependente* não apenas das bases de dados por meio das quais são treinados, mas de outros modelos que também são treinados com outras bases de dados que, por sua vez, *dependem* da coleta, padronização, armazenamento de dados e assim por diante; (iv) crescimento da opacidade epistêmica a partir do uso de modelos complexos que não são transparentes.

As dificuldades técnicas encontradas ao longo da pesquisa apontam para a natureza interdisciplinar do projeto que exigiu o aprofundamento de nosso conhecimento em ferramentas de análise de dados, infraestruturas de dados, e bioinformática. Essas dificuldades

foram parcialmente superadas pelo diálogo com colegas de diferentes áreas, cuja colaboração foi imprescindível para o desenvolvimento do projeto.

Abaixo, esboçamos o desenvolvimento da pesquisa por meio de representação diagramática:

Figura 1 – Diagrama da estrutura da pesquisa desenvolvida

2. Resultados:

2.1 Teóricos: Como resultado teórico, desenvolvemos um arcabouço conceitual em torno dos conceitos de informação, abdução, jornada de dados, serendipidade, e opacidade epistêmica a fim de investigar possíveis implicações epistemológicas da aplicação de massiva quantidade de dados no processo de descoberta científica. Concluímos provisoriamente que o uso de modelos algoritmos de análise de dados pode auxiliar o desenvolvimento de estratégias heurísticas para o desvelamento de relações entre variáveis de interesse. Entretanto, seu uso não garante que essas relações expressem padrões relevantes que expressem relações causais ou correlações estatisticamente significativas. O resultado obtido por meio de modelos mecânicos de análise de dados exige, por parte de investigadores, o estabelecimento de critérios de relevância que são altamente contexto-dependentes. Argumentamos que o estabelecimento de relevância se dá a partir da adoção provisória dos resultados entendidos como promissores que, quando situados na área de especialização das pesquisadoras, possibilitam a emergência de informação significativa. O desenvolvimento desta pesquisa se encerra, provisoriamente, com a indicação de duas questões a serem futuramente investigadas: Quais são as consequências epistemológicas da opacidade algorítmica para a prática das ciências de dados intensivos? Em que medida a crescente automação da prática científica está alterando formas de raciocinar por abdução?

2.2 Práticos: Como resultados práticos, apresentamos nossa pesquisa em eventos, congressos e seminários nacionais e internacionais. Publicamos artigos e capítulos de livros em revistas nacionais e internacionais em colaboração com autores de diferentes disciplinas. Além disso, estamos organizando um livro, para servir como referencial teórico, sobre informação cujo objetivo é reunir estudos, reflexões e pesquisas em estudos sobre informação. Estabelecemos novas colaborações com grupos de estudos e pesquisadores nacionais e internacionais, por exemplo, na Universidade Federal do ABC Paulista, na Universidade Federal do Paraná, Universidade de Exeter, Universidade Tecnológica de Munique, Universidade de Copenhagen, Universidade Tecnológica da Dinamarca, Universidade Tecnológica de Delft, entre outros centros acadêmicos.