UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Marília

OFICINA UNIVERSITÁRIA
CULTURA ACADÊMICA
Editora

Problemas dos Big Data

Ricardo Peraça Cavassane
Felipe Sobreira Abrahão
Itala Maria Loffredo D'Ottaviano

Como citar: CAVASSANE, Ricardo Peraça; ABRAHÃO, Felipe Sobreira; D'OTTAVIANO, Itala Maria Loffredo. Problemas dos Big Data. *In*: ALMEIDA, Carlos Cândido de; VITTI-RODRIGUES, Mariana (org.). **Estudos pluridisciplinares da informação**: filosofia, tecnologia e semiótica. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2025. p. 187-202. DOI: https://doi.org/10.36311/2025.978-65-5954-633-6.p187-202.

6
PROBLEMAS DOS BIG DATA
BIG DATA PROBLEMS

Ricardo Peraça CAVASSANE
UNICAMP
ricardo.peraca@gmail.com
https://orcid.org/0000-0002-2189-1055

Felipe Sobreira ABRAHÃO
UNICAMP
felipesabrahao@gmail.com
https://orcid.org/0000-0001-7314-6543

Itala Maria Loffredo D'OTTAVIANO
UNICAMP
italaloffdotta@gmail.com
https://orcid.org/0000-0002-0591-6785

https://doi.org/10.36311/2025.978-65-5954-633-6.p187-202

Resumo: Independentemente de trazer uma inovação metodológica ou epistemológica para o fazer científico, o paradigma dos Big Data certamente vem influenciando a dinâmica da sociedade contemporânea, especialmente com a maior atuação de algoritmos de inteligência artificial. Neste trabalho, apresentamos problemas dos Big Data que questionam os alicerces tanto de sua confiabilidade empírica quanto de suas propriedades fundamentais. Primeiramente, discutimos sobre: a prevalência esperada de correlações espúrias na análise de dados, constituindo o chamado paradoxo dos Big Data; e a emergência de profecias autorrealizáveis a partir de ciclos de retroalimentação. Ambas constituem a ocorrência de zemblanidade (o contrário de serendipidade, ou seja, um resultado desafortunado e esperado) dentro desse paradigma, a qual se distingue de maneira diametralmente oposta à serendipidade, até então almejada pelos entusiastas dos Big Data. Por fim, apresentamos uma limitação teórica fundamental dos Big Data que impacta o próprio conhecimento formal obtido pela ciência.

Palavras-chave: Big Data, Bolha de Simplicidade, Correlação Espúria, Profecia Autorrealizável, Zemblanidade.

Abstract: Regardless of whether it brings methodological or epistemological innovation to scientific practice, the Big Data paradigm has certainly influenced the dynamics of contemporary society, especially with the increased use of artificial intelligence algorithms. In this chapter, we present Big Data problems that question the foundations of both its empirical reliability and its fundamental properties. First, we discuss: the expected prevalence of spurious correlations in data analysis, constituting the so-called Big Data paradox; and the emergence of self-fulfilling prophecies from feedback loops. Both constitute the occurrence of zemblanity (the opposite of serendipity, i.e. an unfortunate and expected outcome) within this paradigm, which is diametrically opposed to the serendipity hitherto sought by Big Data enthusiasts. Finally, we present a fundamental theoretical limitation of Big Data that impacts the very formal knowledge obtained from science.

Keywords: Big Data, Simplicity Bubble, Spurious Correlation, Self-Realizing Prophecy, Zemblanity.

1 Introdução

Neste capítulo tratamos de alguns problemas que podem ocorrer em modelos baseados em Big Data, isto é, em dados caracterizados principalmente por seu grande volume, alta velocidade de expansão e ampla variedade de estruturas. Muito embora o termo Big Data já não esteja mais em tanta evidência, predominam nas discussões sobre os avanços recentes de tecnologias informacionais termos que dizem respeito aos métodos de análise dos dados, como Inteligência Artificial, *Machine Learning* e *Deep Learning* e, mais recentemente, Inteligência Artificial Generativa e *Large Language Models*. Todas estas tecnologias se fundamentam na análise de quantidades massivas de dados, coletadas principalmente da atividade de usuários de diversas plataformas *online*, portanto em constante crescimento e atualização, em diferentes formatos e incluindo tanto dados estruturados quanto semi-estruturados e não-estruturados; ou seja, Big Data.

2 Correlações espúrias e o paradoxo dos Big Data

Alguns autores defendem que modelos baseados em Big Data permitem a detecção de correlações inesperadas nos dados, sem a necessidade de estabelecer hipóteses de investigação ou de se fundamentar em teorias científicas (Anderson, 2008; Mayer-Schönberger; Cukier, 2013). Antecipando a advertência de que "correlação não implica em causalidade" (Pearl, 2009; Zenil *et al.*, 2020), tais autores afirmam que, com a análise de dados massivos, encontrar correlações pode ser o suficiente e que saber "o que" ocorre é o bastante, não sendo necessário saber "por que" isso ocorre.

Porém, a análise de Big Data em geral está sujeita a um grande número de correlações espúrias, isto é, padrões estatísticos sem nenhum significado. Uma correlação entre duas variáveis pode ser considerada espúria devido a uma terceira variável causalmente correlacionada às duas primeiras, a chamada variável de confusão; mas também por se tratar de uma mera coincidência. Conforme demonstrado por Calude e Longo (2017), correlações espúrias aparecem necessariamente na maioria das análises realizadas sobre grandes bases de dados. Isso ocorre mesmo que, por construção ou por princípio, estas bases somente contenham dados gerados alea-

toriamente, isto é, que tais dados sejam livres de quaisquer redundâncias e estruturas subjacentes.

> Cientistas e filósofos sempre se preocuparam com as falácias que se comete ao olhar apenas para correlações [...] Nosso trabalho confirma a intuição de que, quanto maior a base de dados em que se busca por correlações, maior é a chance de encontrar regularidades e maior é o risco de cometer tais falácias. [...] a maioria esmagadora de correlações é espúria. Em outras palavras, haverá regularidades, mas, por construção, na maior parte do tempo (quase sempre, no sentido matemático), essas regularidades não podem ser usadas para predições e ações confiáveis (Calude; Longo, 2017, p. 609, tradução nossa).

Dentre as teorias matemáticas utilizadas pelos autores está a Teoria de Ramsey, segundo a qual "A desordem total é uma impossibilidade. Todo conjunto grande de números, pontos ou objetos necessariamente contém um padrão altamente regular" (Graham; Spencer, 1990, p. 112, tradução nossa). Do resultado de sua análise, os autores derivam um fenômeno matemático e científico que desafia a intuição: "Assim, paradoxalmente, quanto mais informação temos, mais difícil é extrair algum sentido delas. Muita informação tende a se comportar como pouca informação" (Calude; Longo, 2017, p. 600, tradução nossa). Ou, como coloca Smith:

> Quanto maior o conjunto de dados, maiores são as chances de encontrarmos padrões que sejam, a priori, altamente improváveis. [...] O mesmo princípio vale para modelos em que um conjunto de variáveis explicativas seja usado para predizer uma variável de resposta, de modo que quanto maior o número de potenciais variáveis explicativas, maior é a probabilidade de que relações coincidentes estatisticamente significativas sejam encontradas. Este é o paradoxo dos big data: a mineração de dados é mais sedutora quando há um grande número de variáveis, mas os inevitáveis padrões coincidentes em grandes conjuntos de dados tornam mais provável que as relações estatísticas descobertas em mineração de dados sejam espúrias (Smith, 2020, p. 2, tradução nossa).

Segundo Smith (2020), o grande número de correlações espúrias encontradas em uma base de dados massiva pode acabar dificultando a

detecção de correlações não-espúrias que possam estar presentes nos dados. Padrões espúrios podem ser enganadores porque podem levar um observador a pensar que um padrão estatisticamente significativo foi encontrado, mesmo que este padrão fosse de qualquer maneira ocorrer numa base de dados desprovida de padrões (isto é, perfeitamente aleatória), e assim tais padrões não fornecem qualquer informação útil acerca de características dos dados ou do processo que os gerou.

3 *Feedback loops* e profecias autorrealizáveis

Um *feedback loop* pode ocorrer em um modelo baseado em Big Data quando este é aplicado a uma determinada população e o seu algoritmo e/ou os seus dados, quer estes últimos sejam verdadeiros ou falsos, suscitam respostas no público, público este que, por sua vez, através de suas ações, retroalimenta o modelo com dados enviesados (Auerbach, 2014). Quando não há nada com o que o conjunto de dados possa ser comparado, tal tipo de má circularidade pode gerar erros flagrantes, como quando o *Google Flu Trends* (GFT) sobre-estimou grosseiramente os casos de gripe mais tarde confirmados pelo *CDC*, o centro de controle e prevenção de doenças estadunidense (Lazer *et al.*, 2014); porém, isso pode levar também a um aparente sucesso do modelo, transformando-o assim em uma profecia autorrealizável.

Ou seja, um modelo baseado em Big Data pode ter sua base de dados alterada pela sua própria aplicação, devido às reações do público a tal aplicação, se o público afetado e o público cujas atividades geram os dados utilizados pelo modelo forem os mesmos, o que ocorre frequentemente no contexto dos Big Data, uma vez que tais bases de dados são continuamente expandidas, e em especial quando o modelo se utiliza de dados coletados em redes sociais. Tal alteração na base de dados do modelo o tornará enviesado, possivelmente de modo a confirmá-lo ou reforçá-lo, perpetuando uma certa propriedade, tendência ou um certo estado de coisas.

No caso do *Google Flu Trends*, a estimativa do modelo, que visava detectar um padrão da realidade por meio de um padrão de atividade na internet, foi inflacionada devido ao impacto do modelo naquela mesma

atividade online, isto é, as buscas por sintomas da gripe no *google*, que deveriam constituir variáveis *proxy* apropriadas para casos de gripe, foram inflacionadas pelo próprio mecanismo de busca do *google*, que sugere algumas opções para completar automaticamente uma determinada sequência de termos de pesquisa. Porém, nesse caso, havia com o que comparar os resultados do *GFT*: os dados obtidos através de exames médicos reunidos e publicados pelo *CDC* algumas semanas depois. Quando não há com o que comparar os resultados do modelo, ele pode se comportar como uma "profecia autorrealizável", cuja descrição por Merton (1948b) se assemelha muito àquela de um *feedback loop* pernicioso:

> A profecia autorrealizável é, no início, uma falsa definição da situação que evoca um novo comportamento, que faz a concepção originalmente falsa tornar-se verdadeira. A validade ilusória da profecia autorrealizável perpetua um reino do erro. Pois o profeta citará o atual curso de eventos como prova de que ele estava correto desde o início (Merton, 1948b, p. 195, tradução nossa).

O fato de que o "profeta" é um agente do sistema que introduz ou propicia um novo comportamento no mesmo (que seria diferente ou mesmo o oposto sem a intervenção deste agente) através da interação entre o algoritmo orientado por dados e os componentes do sistema datificado (por exemplo, a população representada pelo modelo do *GFT*), faz com que a ocorrência de profecias autorrealizáveis seja um efeito que, em geral, é muito difícil de predizer, de tal maneira demandando muito mais atenção de cientistas e filósofos de sistemas complexos. Este é o caso uma vez que tal novo comportamento emergentemente enviesado pode ser caracterizado como um caso de auto-organização secundária (ver Debrun, 2009a; Debrun, 2009b). Além do agente causador do *feedback loop* induzir uma estabilização do comportamento coletivo na forma de um viés, viés este que é reforçado no tempo, esta forma de auto-organização ocorre devido a um processo irredutivelmente emergente, ou criativo, com respeito aos elementos constituintes do sistema (Cavassane; D'Ottaviano; Abrahão, no prelo). Nesse caso, teorias formais podem, por princípio, ser incapazes de predizer a presença de tais "profetas" no sistema (Abrahão; Zenil, 2022).

O surgimento de profecias autorrealizáveis devido à presença de *feedback loops* em modelos baseados em Big Data fica claro em alguns exemplos apresentados por O'Neil (2016). O *PredPol*, hoje *Geolitica*[1], um modelo de policiamento preditivo, por exemplo, "[...] processa dados criminais históricos e calcula, hora a hora, onde há maior probabilidade de que ocorram crimes" (O'Neil, 2016, p. 85, tradução nossa). Seu algoritmo, projetado com base em programas de detecção de terremotos, tem um enfoque geográfico. Por não focalizar nos dados de indivíduos, não seria, supostamente, influenciado por preconceitos étnico-raciais. Porém, é preciso notar que modelos de policiamento preditivo estão situados no contexto de uma reforma no policiamento que visa "[…] tornar a polícia mais proativa e vigilante, ao invés de reativa e focada em emergências" (Benbouzid, 2019, p. 1, tradução nossa). Direcionados pela política de tolerância zero, os departamentos de polícia utilizam o *PredPol* não apenas para prever crimes violentos, mas também infrações menores, como o uso de entorpecentes. Este tipo de infração é mais fácil de prever, pois tende a ocorrer sempre nos mesmos locais, enquanto assaltantes e criminosos violentos tentam evitar as áreas mais visadas pela polícia (O'Neil, 2016).

No entanto, o estudo de Lum e Isaac (2016) mostrou que os dados de uso de drogas ilícitas produzidos pela polícia não são representativos da realidade do uso de drogas, super-representando bairros pobres, onde a maioria da população é afro-americana ou hispânica. Desta forma, alimentado com os dados coletados pela polícia, o algoritmo define como prioritárias tais áreas. O direcionamento do policiamento a estas áreas – e as consequências das agressões e prisões voltadas contra uma população vulnerável – aumenta ainda mais a quantidade de dados criminais acerca delas. Isso gera um *feedback loop*, uma má circularidade que não apenas reforça e reitera os vieses e preconceitos da sociedade, ao invés de evitá-los, como também os justifica, devido à aura de cientificidade dos Big Data (O'Neil, 2016). Ou seja, o modelo do *PredPol* é alimentado com dados da polícia acerca de atividades criminosas, dados estes já enviesados, de modo a priorizar certas populações; a aplicação do modelo, por sua vez, direciona o policiamento a estas mesmas populações, produzindo dados cada vez

[2] https://www.linkedin.com/company/geolitica.

mais enviesados e confirmando o "sucesso" não apenas do modelo, como da própria polícia, que vê seus números de flagrantes e prisões crescerem. E, mesmo que um crime não ocorra, o modelo é considerado bem-sucedido, como nota Dupuy:

> [...] o programa PredPol [...] sempre vence! O PredPol anuncia que um crime irá ocorrer em uma área específica da cidade. O policial vai responder à situação. Uma de duas coisas acontece: ou um crime ocorre como o planejado e o policial impede o infrator, em cujo caso o PredPol recebe sua medalha de ouro; ou nenhuma infração ocorre. Mas isto provavelmente está ligado à presença do policial no local, e então ainda se trata de uma medalha de ouro para o programa (Dupuy, 2018, p. 160, tradução nossa).

Para Benslimane (2014), o *Predpol* ignora os diversos fatores sociológicos que levam à criminalidade, bem como os vieses do policiamento, de modo a criar uma representação simplificadora e aparentemente objetiva de acordo com a qual ocorrem mais crimes em certas áreas de uma cidade. Tal modelo falha em representar acuradamente a presença real de criminalidade em uma cidade porque os dados que o alimentam não representam a totalidade dos crimes ocorridos na cidade (e tampouco representam uma amostra aleatória dos crimes, mas sim uma amostra arbitrária). Outro motivo para tal falha, é que o modelo não é informado por teorias científicas acerca das razões que fazem com que certos crimes ocorram mais em determinadas áreas do que em outras.

Preocupado apenas com "o que" ocorre e não com o "porquê" de isso ocorrer (motivado por uma lógica de mercado, cujo foco está em aumentar a eficiência para diminuir custos), tal modelo gera "previsões" altamente enviesadas (tal modelo certamente não é capaz de prever ocorrências daqueles tipos de crime mais comuns em regiões mais ricas, por exemplo). Além disso, uma vez que seus resultados sejam aplicados em políticas públicas, tais modelos podem não apenas não solucionar os problemas sociais já existentes, como também agravá-los.

4 Serendipidade e zemblanidade

Se a análise de Big Data funcionasse como alguns de seus entusiastas mais otimistas afirmam, de forma a encontrar correlações não-espúrias inesperadas e de valor, sem a necessidade de estabelecer qualquer hipótese de investigação, ela poderia ser caracterizada como uma geradora de serendipidade artificial; porém, conforme argumentaremos, modelos baseados em Big Data frequentemente produzem o oposto da serendipidade, isto é, a zemblanidade.

O termo "zemblanidade" aparece pela primeira vez em um romance, em um excerto do diário mantido por seu personagem principal:

> *Serendipidade*. De *Serendip*, um antigo nome do Ceilão, hoje Sri Lanka. Uma palavra cunhada por Horace Walpole, que a inventou com base em um conto folclórico, cujos heróis estavam sempre fazendo descobertas de coisas que eles não estavam procurando. *Ergo*: serendipidade, a faculdade de fazer descobertas felizes e inesperadas por acidente. Então qual seria o oposto de Serendip, uma terra tropical rica em especiarias, vegetação exuberante e pássaros canoros, banhada pelo mar e pelo sol? Pense em outro mundo no norte longínquo, árido, coberto de gelo, frio, um mundo de rocha. Chame-o *Zembla*. *Ergo*: zemblanidade, o oposto da serendipidade, a faculdade de fazer descobertas infelizes, desafortunadas e esperadas, e não por acidente. Serendipidade e zemblanidade: os dois polos do eixo ao redor do qual giramos (Boyd, 1998, p. 234-235, tradução nossa).

O termo "zemblanidade" é, portanto, o antônimo binário ou complementar do termo "serendipidade", isto é, a zemblanidade é diametralmente oposta ou contraditória em relação à serendipidade: enquanto a serendipidade se refere a uma descoberta afortunada e inesperada, a zemblanidade é um resultado desafortunado e esperado.

A serendipidade é uma descoberta afortunada, tanto no sentido de ser trazida por boa sorte ou por um acidente feliz, quanto no sentido de ser benéfica ou valiosa. Já a zemblanidade é um resultado desafortunado. Note, porém, que sendo "zemblanidade" o antônimo complementar de "serendipidade", ao invés de envolver má sorte, a zemblanidade não en-

volve, de maneira alguma, a sorte ou o acaso; e que ao invés de ter pouco ou nenhum valor ou benefício, um resultado zemblanitoso é maléfico ou indesejado. Enquanto a ocorrência de serendipidade, no contexto da pesquisa científica, resolve um problema, ou mostra o caminho para a sua solução, ou ainda oferece uma solução para um problema que ainda não havia sido colocado (Merton, 1948a), a ocorrência de zemblanidade revela ou traz à tona um problema subjacente, como um pressuposto, uma consequência ou um efeito colateral.

A serendipidade, sendo fruto do acaso, é inesperada; já a zemblanidade é, ou ao menos deveria ser esperada, pois é fruto de uma necessidade. Enquanto a serendipidade é uma descoberta surpreendente e imprevisível, a zemblanidade não é uma descoberta, mas um resultado que foi ou deveria ter sido previsto. Trata-se de um resultado necessário, com o qual tal pesquisa se depararia inevitavelmente. Assim, uma descoberta serendipitosa ocorre "[...] na interseção entre o acaso e a sabedoria [...]" (Copeland, 2019, p. 2386, tradução nossa), ou é realizada por acidente e sagacidade, como o próprio Walpole, que cunhou o termo, coloca (Walpole, 2011). Isto é, ela requer do pesquisador atenção e conhecimento a fim de que ela seja notada e compreendida enquanto tal; já a zemblanidade decorre da negligência e/ou da ignorância do pesquisador que foi incapaz de antecipá-la, apesar de ela estar intrinsecamente ligada ao objeto da pesquisa em questão. Não sendo um fenômeno emergente como a serendipidade, mas redutível ao estado do conhecimento prévio do observador, um resultado zemblanitoso é, em princípio, facilmente previsto.

Do ponto de vista das motivações iniciais subjacentes ao programa de Hilbert[2], as posteriores provas de resultados de incompletude podem ser interpretadas como casos de zemblanidade em teorias formais. Sob o mesmo formalismo e a mesma suposição do poder de sistemas axiomáticos

[3] Em resumo, o programa de Hilbert buscava uma esquematização da matemática para que a produção da mesma fosse garantidamente válida, correta e seguisse "boas práticas" (Abrahão, 2011; Zach, 2023). Por exemplo, em um dos pontos centrais, buscava-se a consistência das teorias fundacionais de forma que se evitasse os princípios básicos, como os axiomas, levassem a conclusões contraditórias. O programa de consistência posteriormente teve suas motivações e objetivos postos em cheque com os resultados de incompletude (e incomputabilidade). Outro ponto central era o programa de conservação em que se buscava que teoria matemática mais abstratas (como números ordinais e cardinalidade) possam levar a conclusões verdadeiras sobre domínios menos abstratos da matemática, como a teoria dos números inteiros.

na matemática, pode-se obter uma prova de um metateorema que afirma a impossibilidade de certas provas matemáticas. Nesse contexto das aspirações mecanicistas e finitistas do programa de Hilbert, a incompletude aparece como um efeito colateral negativo de teorias formais. Porém, na verdade, incompletude na matemática evidencia a emergência de informação algorítmica (e, consequente, também a abertura ilimitada à criatividade) em formalizações matemáticas de sistemas suficientemente complexos (Abrahão *et al.*, 2020; Abrahão; Zenil, 2022).

Um exemplo de zemblanidade em Big Data é a ocorrência de correlações espúrias em grandes bases de dados, das quais tratamos na Seção 2. Isso se dá justamente porque a ocorrência de correlações espúrias é matematicamente demonstrável no contexto de uma quantidade de dados muito grande, embora qual correlação (ou correlações) em particular será essa a ser revelada possa ser imprevisível.

A zemblanidade pode resultar também de um *feedback loop* previsível e que determine o resultado da análise dos dados, como aqueles que fazem o modelo se comportar como uma profecia autorrealizável, a qual já esperamos que aconteça. Conforme o exemplo supracitado, parte dos resultados do *PredPol* podem ser considerados zemblanitosos, uma vez que qualquer morador que conheça a realidade das zonas periféricas e com maior prevalência de populações vulneráveis saberá dizer quais regiões provavelmente já serão apontadas como as mais "perigosas" pelos agentes classificadores, detectores e registradores, os quais são partes de tal modelo de policiamento preditivo.

Outro exemplo de modelo de Big Data que exibe resultados zemblanitosos seria o *COMPAS*, modelo criado para avaliar o potencial risco de reincidência de detentos, que foi utilizado nos Estados Unidos para informar certas decisões de juízes. Assim como o *PredPol*, tal modelo é alimentado com dados produzidos pela polícia e pelos sistemas judiciário e carcerário (Northpointhe, 2015). Ora, qualquer observador que conheça a realidade dos sistemas policial, judiciário e carcerário estadunidenses saberá que um modelo que seja alimentado por tais dados produzirá resultados com vieses raciais. De fato, o estudo de Larson *et al.* (2016) mostrou que no modelo, cuja taxa de sucesso é de aproximadamente sessenta por

cento, detentos afro-americanos são frequentemente erroneamente classificados com um alto risco de reincidência, enquanto detentos caucasianos são frequentemente erroneamente classificados com um baixo risco de reincidência.

5 Bolhas de simplicidade

Esses são exemplos de problemas empiricamente evidenciados dos Big Data, cujos altos impactos práticos na sociedade suscitam uma pervasiva investigação acadêmica, assim como sua decorrente ação política e social. Por outro lado, argumentamos que uma outra faceta dos Big Data também revela limites no contexto da justificação do conhecimento, no entanto, agora aparecendo como um limite fundamental do conhecimento formal (seja este produzido pelo empreendimento científico humano somente ou com a ajuda de implementações de inteligência artificial): o problema das bolhas de simplicidade (Abrahão *et al.*, 2024; Abrahão *et al.*, 2023).

Suponha uma teoria formal e um algoritmo de inteligência artificial (IA) computavelmente implementável arbitrariamente escolhidos. A teoria formal pode ser, por exemplo, aquela considerada capaz de idealmente explicar grande parte (ou mesmo a totalidade) dos dados (ou fenômenos) conhecidos até o momento. O algoritmo pode ser, por exemplo, aquele considerado mais eficiente e/ou mais poderoso que combine da melhor forma disponível os mais avançados métodos de aprendizado de máquina e de processamento lógico-dedutivo simbólico, isto é, que combine tanto IA *bottom-up* quanto a *top-down* (Zenil *et al.*, 2023). O problema que levantamos concerne uma quantidade suficientemente grande de dados externos disponível a esse sistema lógico-matemático-computacional, quantidade a partir da qual a teoria formal não conseguiria mais minimizar a probabilidade dela própria estar errada. Estas condições configuram, então, um limite ou um "teto" epistemológico para aquilo o que conseguimos garantir que a ciência continuará progredindo em direção a minimizar erros ou diminuir o desconhecimento. Em geral, não se pode evitar o aparecimento de modelos ótimos espúrios, ou o aparecimento de teorias tanto válidas

quanto empiricamente verificáveis apesar de, ao mesmo tempo, serem espúrias.

Para atingirmos tal problema, demonstra-se um método matemático de ludibriação do observador (Abrahão *et al.*, 2023). Em particular, quando o processo externo que gera os fenômenos (i.e., os dados) armadilha o observador (i.e., o sistema lógico-matemático-computacional supracitado) dentro de uma bolha de simplicidade. Isso ocorre porque a fonte dos fenômenos disponíveis para o observador adquirir conhecimento é complexa o suficiente para garantir que qualquer teoria ou modelo propostos pelo observador não são suficientemente complexos para realmente conseguir prever fenômenos futuros de forma global, mesmo que o observador considere (ou demonstre) que a teoria ou modelos consiga prever fenômenos futuros de forma global.

6 Considerações Finais

Nesse texto defendemos a proposição de que o estudo dos limites teóricos e empíricos dos Big Data é crucial tanto para o progresso da ciência quanto para a mitigação de problemas sociais. Em particular, no contexto dos Big Data, estudamos correlações espúrias, profecias autorrealizáveis, zemblanidade e bolhas de simplicidade. Os problemas aqui discutidos são pervasivos e impactantes especialmente numa sociedade atual massivamente digitalizada, onde as próprias relações humanas estão cada vez mais tecnológico-intermediadas. Nesse sentido, como pesquisa futura, aqui levantamos a questão de que quanto mais as relações interpessoais ou institucionais são intermediadas por algoritmos de IA, maior a pertinência na sociedade dos efeitos dos limites, outrora teóricos e puramente matemáticos, de nossos métodos científicos e tecnológicos.

Referências

ABRAHÃO, F. S. *Demonstrando a consistência da aritmética.* 2011. Dissertação (Mestrado em história das ciências e das técnicas e epistemologia) – Coppe/IQM/IM, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2011. Disponível em: https://doi.org/10.5281/zenodo.1213459. Acesso em: 10 jul. 2025.

ABRAHÃO, F. S.; CAVASSANE, R. P.; WINTER, M.; VITTI-RODRIGUES, M.; D'OTTAVIANO, I. M. L. A simplicity bubble problem and zemblanity in digitally intermediated societies. In: IPPOLITI, E.; MAGNANI, L.; ARFINI, S. (eds.) Model-Based Reasoning, Abductive Cognition, Creativity. MBR 2023. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 70. Berlim: Springer. Disponível em: https://link.springer.com/chapter/10.1007/978-3-031-69300-7_20. Acesso em: 30 nov. 2024.

ABRAHÃO, F. S.; D'OTTAVIANO, Í. M. L.; WEHMUTH, K.; DÓRIA, F. A.; ZIVIANI, A. Learning the undecidable from networked systems. *In*: WUPPULURI, S.; DORIA, F. A. (ed.) *Unravelling complexity*: the life and work of Gregory Chaitin. Singapore: World Scientific Publishing, 2020. p. 131-179.

ABRAHÃO, F. S.; ZENIL, H. Emergence and algorithmic information dynamics of systems and observers. *Philosophical transactions of the Royal Society A*: mathematical, physical and engineering sciences, London, v. 380, n. 2227, p. 1-20, July 2022. Disponível em: https://doi.org/10.6084/m9.figshare.c.5901204. Acesso em: 01 jul. 2023.

ABRAHÃO, F. S.; ZENIL, H.; PORTO, F.; WINTER, M.; WEHMUTH, K.; D'OTTAVIANO, I. M. L. A simplicity bubble problem in formal-theoretic learning systems. *arXiv*, New York, n. 2112.12275v2, p. 1-38, Apr. 2023. Disponível em: https://arxiv.org/abs/2112.12275. Acesso em: 01 jul. 2023.

ANDERSON, C. End of theory: Will the Data Deluge make the scientific method obsolete? *Wired magazine*, San Francisco, v. 16, n. 7, June 2008. Disponível em: https://www.wired.com/2008/06/pb-theory. Acesso em: 01 jul. 2023.

AUERBACH, D. Big data is overrated: and Ok-Cupid's user experiments prove it. *Slate*, New York, 30 July 2014. Disponível em: http://www.slate.com/articles/technology/bitwise/2014/07/facebook_okcupid_user_experiments_ethics_aside_they_show_us_the_limitations.html. Acesso em: 01 jul. 2023.

BENBOUZID, B. To predict and to manage. Predictive policing in the United States. *Big Data & society*, London, v. 6, n. 1, p. 1-13, Jan./June 2019. Disponível em: https://doi.org/10.1177/2053951719861703. Acesso em: 01 jul. 2023.

BENSLIMANE, I. Étude critique d'un système d'analyse prédictive appliqué à la criminalité: Predpol. *Cortecs*, Grenoble, 10 Dec. 2014. Disponível em: https://cortecs.org/politique-societe/predpol-predire-des-crimes-ou-des-banalites. Acesso em: 01 jul. 2023.

BOYD, W. *Armadillo*. London: Hamish Hamilton, 1998.

CALUDE, C. S.; LONGO, G. The deluge of spurious correlations in Big Data. *Foundations of science*, Amsterdam, v. 22, n. 3, p. 595-612, Sept. 2017.

CAVASSANE, R. P.; D'OTTAVIANO, I. M. L.; ABRAHÃO, F. S. Big data and the emergence of zemblanity and self-fulfilling prophecies. *In*: INTERNATIONAL MEETING ON INFORMATIONAL, KNOWLEDGE AND ACTION, 11., 2021, [*s.l.*]. *Proceedings* [...]. [202-?]. No prelo.

COPELAND, S. On serendipity in science: discovery at the intersection of chance and wisdom. *Synthese*, London, v. 196, n. 6, p. 2385-2406, June 2019.

DEBRUN, M. A ideia de auto-organização. *In*: D'OTTAVIANO, I. M. L.; GONZALEZ, M. E. Q. (org.). *Michel Debrun*: Brazilian national identity and self-organization. Campinas: Coleção CLE, 2009a. p. 53-74.

DEBRUN, M. A dinâmica da auto-organização primária. *In*: D'OTTAVIANO, I. M. L.; GONZALEZ, M. E. Q. (org.) *Michel Debrun*: Brazilian national identity and self-organization. Campinas: Coleção CLE, 2009b. p. 111-146.

DUPUY, J. P. Science without philosophy: the case of Big Data. *Crisis and critique*, Tirana, v. 5, n. 1, p. 147-161, Mar. 2018.

GRAHAM, R.; SPENCER, J. H. Ramsey theory. *Scientific American*, New York, v. 263, n. 1, p. 112-117, July 1990.

LARSON, J.; MATTU, S.; KIRCHNER, L.; ANGWIN, J. How we analyzed the COMPAS recidivism algorithm. *ProPublica*, New York, 23 May 2016. Disponível em: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Acesso em: 01 jul. 2023.

LAZER D.; KENNEDY, R.; KING, G.; VESPIGNANI, A. The parable of Google flu: traps in Big Data analysis. *Science*, Washington, v. 343, n. 6176, p. 1203-1205, Mar. 2014.

LUM, K.; ISAAC, W. To predict and serve? *Significance*, London, v. 13, n. 5, p. 14-19, Oct. 2016. Disponível em: https://doi.org/10.1111/j.1740-9713.2016.00960.x. Acesso em: 01 jul. 2023.

MAYER-SCHÖNBERGER, V.; CUKIER, K. *Big Data*: a revolution that will transform how we live, work and think. New York: Houghton Mifflin Harcourt, 2013.

MERTON, R. K. The bearing of empirical research upon the development of social theory. *American sociological review*, Chicago, v. 13, n. 5, p. 505-515, Oct. 1948a.

MERTON, R. K. The self-fulfilling prophecy. *The Antioch review*, Yellow Springs, v. 8, n. 2, p. 193-210, summer 1948b.

NORTHPOINTE INC. *Practitioner's guide to COMPAS core*. Wheat Ridge, Mar. 2015. Disponível em: https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf. Acesso em: 1 jul. 2023.

O'NEIL, C. *Weapons of math destruction*: how big data increases inequality and threatens democracy. New York: Crown Publishers, 2016.

PEARL, J. *Causality*: models, reasoning and inference. 2nd ed. Cambridge: Cambridge University Press, 2009.

SMITH, G. The paradox of Big Data. *SN Applied sciences*, Bern, v. 2, n. 1041, p. 1-8, May 2020.

WALPOLE, H. *Horace Walpole's correspondence*. New Haven: Yale University Library, 2011. Disponível em: http://images.library.yale.edu/hwcorrespondence. Acesso em: 01 jul. 2023.

ZACH, R. Hilbert's program. *In*: ZALTA, E. N.; NODELMAN, U. (ed.). *Stanford encyclopedia of philosophy*. Stanford, CA: Metaphysics Research Lab, Stanford University, 2023.

ZENIL, H.; KIANI, N. A.; ABRAHÃO, F. S.; TEGNÉR, J. N. Algorithmic information dynamics. *Scholarpedia journal*, v. 15, n. 7, p. 53143, 2020.

ZENIL, H. *et al*. The future of fundamental science led by generative closed-loop artificial intelligence. *arXiv Preprints*, New York, n. arXiv:2307.07522v3, 2023. Disponível em: https://arxiv.org/abs/2307.07522. Acesso em: 10 jul. 2023.