semeiosis
SEMIÓTICA E TRANSDISCIPLINARIDADE EM REVISTA
TRANSDISCIPLINARY JOURNAL OF SEMIOTICS

Seriam os Big Data capazes de prever o comportamento de Sistemas Complexos?

Ricardo Peraça Cavassane | ricardo.peraca@gmail.com
Doutorando do Programa de Pós-graduação
em Filosofia da Universidade Estadual de Campinas – Unicamp.

resumo

Neste artigo propomos uma investigação acerca da adequação de modelos estatísticos baseados em Big Data à descrição e previsão do comportamento de sistemas complexos auto-organizados, como aqueles que envolvem agentes humanos, uma vez que tais sistemas se caracterizam pela atuação da causalidade circular, e que tais modelos, segundo seus próprios entusiastas, forneceriam explicações baseadas apenas em correlações, sem identificar quaisquer relações de causalidade.

PALAVRAS-CHAVE: Big Data. Sistemas Complexos. Causalidade.Correlação.

abstract

In this paper we propose an investigation about the adequacy of statistical models based on Big Data to the description and prediction of the behavior of self-organized complex systems, like those that involve human agents, since such systems are characterized by the action of circular causality, and that such models, according to their own enthusiasts, would provide explanations based solely on correlations, without identifying any relations of causality.

KEYWORDS: *Big Data. Complex Systems. Causality. Correlation.*

Os entusiastas dos Big Data defendem que modelos estatísticos cujo objetivo é quantificar fenômenos podem gerar informações acerca de seus estados presentes e prever seus estados futuros com muito mais rapidez e precisão do que outros instrumentos da ciência tradicional. Isso seria possível através da análise de grandes quantidades de dados não estruturados, em diferentes formatos, de diversas fontes e em constante crescimento. Estes modelos, ainda segundo seus entusiastas, poderiam explicar os estados de um fenômeno sem a necessidade de hipóteses ou teorias prévias e sem a influência de vieses humanos. Assim, poderiam fazê-lo sem oferecer alguma explicação causal para o fenômeno modelado. Os críticos dos Big Data, por sua vez, acreditam não ser possível quantificar inteiramente um fenômeno ou mesmo selecionar quais seriam os aspectos efetivamente mais relevantes para sua modelagem. Ressaltam também que modelos estatísticos são sempre fundamentados em certas hipóteses, cujas próprias formulações acabam por incorporar, mesmo que tacitamente, vieses humanos. Por fim, declaram que um modelo que pretenda descrever e prever os estados futuros de um fenômeno sem identificar suas causas e possíveis consequências está fadado ao erro. Considerando tais pretensões e problemas dos modelos baseados em Big Data, propomos investigar em que medida tais modelos são ou não adequados à descrição e previsão do comportamento de sistemas complexos auto-organizados, como aqueles que envolvem agentes humanos, uma vez que tais sistemas se caracterizam pela atuação da causalidade circular, e que tais modelos forneceriam explicações baseadas apenas em correlações, deixando de lado relações de causalidade.

Anderson (2008) afirma que os dados, quando em quantidade suficiente, falam por si mesmos. Segundo ele, dados em volumes massivos, na casa dos petabytes (isto é, dos milhões de gigabytes), quando analisados através de algoritmos estatísticos, permitem detectar padrões que a ciência tradicional teria muito mais dificuldades para encontrar. Isso tornaria obsoleto o método científico tradicional e, com ele, suas hipóteses e teorias. O lema deste tipo de visão é "a correlação é suficiente". Ou seja, os padrões estatísticos encontrados nas análises de Big Data seriam suficientes para evidenciar o que ocorre, tornando desnecessária qualquer teoria que pretenda explicar por que isso ocorre. A noção de causalidade poderia assim ser definitivamente abandonada.

Mayer-Schönberger e Cukier discordam da ideia de que grandes quantidades de dados tornam teorias desnecessárias. "Big Data são, eles próprios, fundados em teoria." (2013: 71, tradução nossa). Diferentes teorias, segundo eles, fundamentam tanto os métodos quanto os resultados das análises de dados, desde a seleção dos mesmos. No entanto, eles compartilham da ideia de que os Big Data nos eximiriam de procurar pelas causas dos fenômenos:

> Big Data é sobre o que, não por que. Nós não precisamos sempre saber a causa de um fenômeno; ao contrário, podemos deixar os dados falarem por si mesmos. Antes dos Big Data, nossa análise em geral se limitava a testar um pequeno número de hipóteses que definíamos muito antes mesmo de coletar os dados. Quando deixamos os dados falar, podemos fazer conexões que nunca havíamos pensado que existiam. (MAYER-SCHÖNBERGER; CUKIER, 2013: 14, tradução nossa).

Segundo os autores, a causalidade, assim como a correlação, raramente pode ser dada como certa, mas apenas como estatisticamente provável. Experimentos que visem inferir nexos causais, porém, seriam pouco práticos, consumiriam muito tempo e recursos, e levantariam importantes questões éticas. Por outro lado, análises correlacionais seriam práticas, baratas e rápidas (MAYER-SCHÖNBERGER; CUKIER, 2013: 67). Além disso, com as novas tecnologias de armazenamento e processamento de dados, tornar-se-ia possível analisar a totalidade (ou um número próximo da totalidade) das ocorrências de fenômenos investigados (por exemplo, a totalidade dos itens visualizados e comprados por clientes da Amazon), ao invés de apenas uma pequena amostra das mesmas. A isto Mayer-Schönberger e Cukier chamam de "N=tudo" (2013: 26-31).

Exemplos de aplicações bem-sucedidas dos Big Data vêm, principalmente, do contexto empresarial. A livraria virtual Amazon, acima referida, utiliza a grande quantidade de dados que possui acerca dos livros que seus clientes compram (ou mesmo daqueles que os clientes apenas visualizam) para criar correlações entre os próprios livros, gerando assim um sistema de recomendação muito mais barato, ágil e eficiente que aquele anteriormente realizado por críticos literários. Outro exemplo é o do aplicativo Farecast, hoje MSN Travel, que analisa os preços de bilhões de passagens aéreas e sua variação no tempo para encontrar os preços mais baixos e, o mais importante, determinar se estes devem aumentar ou diminuir no futuro, possibilitando ao comprador adquirir sua passagem no melhor momento.

Os Big Data analisados pela Amazon, ao contrário dos especialistas que ela costumava empregar, não permitem explicar porque um determinado livro está correlacionado com outro; da mesma forma, os dados que a MSN Travel analisa não lançam luz sobre as inúmeras variáveis que as companhias aéreas levam em conta ao determinar o preço de uma passagem. Porém, mesmo sem encontrar as causas, ambos os modelos têm sucesso em seus objetivos: a Amazon é capaz de aumentar suas vendas através da análise de todas as compras e visualizações de seus clientes, e a MSN Travel é capaz de prever com acuidade a flutuação dos preços das passagens aéreas através da análise dos preços de praticamente todas as passagens oferecidas pelas companhias.

Os críticos dos Big Data não são tão rápidos em assumir que, com as novas tecnologias de dados, é possível analisar a totalidade das ocorrências de um fenômeno. Segundo Schutt e O'Neil, praticamente nunca é possível chegar a "N=tudo", e frequentemente os Big Data deixam passar justamente os casos mais relevantes (2014: 25). Um exemplo de fonte de dados bastante utilizada por aqueles que pretendem identificar padrões de comportamentos das pessoas através dos Big Data é o Twitter. Crawford (2013) cita um estudo acerca dos impactos do furacão Sandy que utilizou dados do microblog, e boyd e Crawford (2012) ressaltam a variedade de estudos que utilizam dados desta fonte, pela facilidade com que eles podem ser coletados. Tais estudos tratam de temas tão variados quanto ritmos de humor, engajamento com a mídia, interações conversacionais e até levantes políticos. Porém, como notam as autoras:

> O Twitter não representa 'todas as pessoas', e é um erro assumir que 'pessoas' e 'usuários do Twitter' são sinônimos: eles são um subconjunto bastante particular. A população que usa o Twitter não é representativa da população global. Nem podemos assumir que contas e usuários são equivalentes. Alguns usuários têm múltiplas contas, enquanto que algumas contas são usadas por múltiplas pessoas. [...] Algumas contas são 'bots' que produzem conteúdo automatizado sem envolver diretamente uma pessoa. (BOYD; CRAWFORD, 2012: 669, tradução nossa).

Tomar a totalidade dos tweets acerca de determinado assunto (selecionados através de determinadas palavras-chave ou hashtags) como um substituto para a totalidade das manifestações públicas acerca do tema em questão é um equívoco grave. O estudo acerca dos impactos do furacão Sandy, por exemplo, como ressalta Crawford (2013), sofre de um "erro de sinal": o estudo assume que o furacão teve maior impacto sobre a ilha de Manhattan, pois a maioria dos tweets sobre o evento se originaram desta localidade. Não é o que de fato ocorreu; a região de Manhattan, menos atingida, concentra uma maior quantidade de smartphones e, portanto, de usuários do Twitter. Por outro lado, as regiões mais impactadas, além de concentrarem um número menor de dispositivos móveis, sofreram com a falta de energia elétrica, o que diminui ainda mais o uso destas tecnologias. Neste caso, justamente as áreas mais afetadas foram ignoradas pelo modelo. Outros modelos que se baseiam em dados coletados por tecnologias informacionais inevitavelmente deixam de fora de suas análises populações pobres (e, portanto, mais vulneráveis, por exemplo, aos impactos de desastres naturais), em que a presença e, em especial, o uso especializado de tais tecnologias é menor:

> Conforme nos movemos em direção a uma era em que dispositivos pessoais são vistos como indicadores de

> necessidades públicas, corremos o risco de que injustiças já existentes sejam ainda mais enraizadas. Assim, para cada conjunto de big data, precisamos perguntar que pessoas são excluídas. Que lugares são menos visíveis? O que ocorre se você vive à sombra de conjuntos de big data? (CRAWFORD, 2013: online, tradução nossa).

A impossibilidade de representar a totalidade das ocorrências em um modelo baseado em Big Data não se restringe ao Twitter ou a fontes de dados semelhantes. De acordo com Kitchin, muito embora um tal modelo se pretenda exaustivo, ele é "[...] tanto uma representação quanto uma amostra, moldado pela tecnologia e pela plataforma utilizada, pela ontologia de dados empregada e pelo ambiente regulatório, e está sujeito a viés de amostragem [...]" (2014: 4, tradução nossa). Por trás da crença de que os Big Data nos permitiriam abandonar quaisquer teorias está o equívoco de que o "dado" representa, ipsis litteris, aquilo que estaria dado na natureza. Na realidade, não há "dados brutos", ou, como afirma Bowker: "'Dado bruto' é tanto um oxímoro quanto uma má ideia." (2005: 184, tradução nossa). Ou seja, mesmo que os Big Data pudessem representar de alguma forma a totalidade das ocorrências de um fenômeno, os dados não "falariam por si mesmos", ao menos não no sentido de que os dados pudessem ser neutros ou livres de pressupostos teóricos. Além disso, modelos baseados em Big Data não eliminariam os vieses presentes nas análises realizadas por humanos, como ressalta Crawford:

> Os números realmente podem falar por si mesmos? Infelizmente, não. Dados e conjuntos de dados não são objetivos; eles são criações do desígnio humano. Nós damos voz aos números, extraímos inferências deles, e definimos seu significado através de nossas interpretações. Vieses ocultos tanto nos estágios de coleta quanto de análise apresentam riscos consideráveis, e são tão importantes para a equação dos Big Data quanto os próprios números. (2013: online, tradução nossa).

A ideia de substituir hipóteses e teorias científicas por modelos estatísticos baseados em Big Data, desta forma, é epistemologicamente problemática. Como notam Ibekwe-SanJuan e Bowker, a natureza dos dados e a opacidade dos algoritmos empregados por tais modelos vão contra os "[...] princípios da falseabilidade e do falibilismo [...] que têm guiado a atividade científica até o momento [...]" (2017: 194, tradução nossa). Os dados e os algoritmos empregados para a análise dos mesmos frequentemente são considerados propriedade de empresas privadas, como aquelas que detém o controle de aplicativos, redes sociais e mecanismos de busca, e por esta razão não podemos ter acesso a diversas informações relevantes sobre os mesmos. Desprovidos

destas informações, não podemos nos assegurar da integridade dos dados ou dos modelos.

Tal opacidade dos algoritmos suscita também importantes questões éticas. Empregados para, por exemplo, avaliar o desempenho de professores, selecionar candidatos a uma vaga de emprego ou avaliar a probabilidade de reincidência de um encarcerado, como mostra O'Neil (2016), os modelos tendem a ocultar seus critérios e o peso que cada um deles tem em seus algoritmos. Além dos algoritmos serem propriedade intelectual das empresas de consultoria que prestam estes serviços, tal opacidade se justificaria por buscar impedir aos sujeitos de burlar o sistema. Porém, ela também lhes impede de contestar caso o modelo seja injusto, por exemplo, demitindo um professor porque seus alunos não foram bem em uma única prova padronizada, não selecionando um candidato porque ele mora em um bairro perigoso, ou negando a liberdade condicional a um sentenciado porque ele pertence a uma minoria étnica, ou porque sua família reside em um local considerado de alta criminalidade.

Desta forma, a substituição de teorias por modelos de Big Data não apenas levanta importantes questões epistemológicas, como é também ética e politicamente perigosa. Como afirma Bowker, o ato de incluir dados em um determinado conjunto, com determinados pressupostos teóricos é, simultaneamente, um ato de excluir outros dados que poderiam representar o mesmo fenômeno. Teorias, evidentemente, fazem o mesmo, mas com justificativas das suas formas de exclusão e dos mecanismos que permitem generalizar a partir dos dados escolhidos (2014: 1797). Modelos de Big Data, porém, uma vez que pretendem encontrar correlações entre os dados, sem se preocupar com nexos causais e ignorando teorias existentes acerca do fenômeno modelado, ao invés de eliminar vieses, podem acabar por incorporá-los. Afinal, como lembram Schutt e O'Neil, "Dados são apenas um eco pálido, quantitativo, dos eventos em nossa sociedade." (2014: 26, tradução nossa).

Assim, modelos estatísticos que analisam grandes quantidades de dados e ignoram os resultados, por exemplo, da Ciência Política, podem gerar "previsões" equivocadas. Além disso, uma vez que seus resultados sejam aplicados em políticas públicas, podem não apenas não solucionar os problemas sociais já existentes, como também agravá-los. Isto fica claro em alguns exemplos apresentados por O'Neil, que chama os modelos estatísticos perniciosos de weapons of math destruction, ou "armas de destruição matemática" – um trocadilho com weapons of mass destruction, ou "armas de destruição em massa", a fim de ressaltar o potencial destrutivo dos modelos (2016: 3). Tratemos do exemplo do programa de policiamento preditivo PredPol, já utilizado em algumas cidades norte-americanas.

O PredPol "[...] processa dados criminais históricos e calcula, hora a hora, onde há maior probabilidade de que ocorram crimes." (2016: 85, tradução nossa). Seu algoritmo, projetado com base em programas de detecção de terremotos, tem um enfoque geográfico. Por não focalizar indivíduos, não seria, supostamente, influenciado por preconceitos étnico-raciais. Porém, 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 que assaltantes e criminosos violentos tentam evitar as áreas mais visadas pela polícia (2016: 88). Desta forma, alimentado com os dados previamente coletados pela polícia, o algoritmo define como prioritárias áreas pobres e de população majoritariamente negra e latina. 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 reitera os vieses e preconceitos humanos, ao invés de evitá-los, como também os corrobora, devido à aura de cientificidade dos Big Data (2016: 91).

Uma forma de feedback loop, como ressalta Auerbach (2014a), ocorre quando os dados, verdadeiros ou falsos, suscitam respostas no público, que por sua vez, através de suas ações, retroalimentam os modelos com dados enviesados. Tal tipo de má circularidade não é algo novo ou exclusivo da era dos Big Data, já tendo sido descrita por Merton e denominada "profecia autorrealizável":

> 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. (1948: 195, tradução nossa).

Um viés na coleta e análise dos dados pode direcionar ações que, por sua vez, produzam dados enviesados que corroboram o modelo. Assim, modelos estatísticos baseados em Big Data podem, devido aos feedback loops, comportar-se como profecias autorrealizáveis. 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. (2018: 160, tradução nossa).

As formas como os usuários das redes sociais as alimentam com seus dados também geram um tipo de feedback loop pernicioso, conforme explicitado pelo caso da Cambridge Analytica, envolvendo processos eleitorais. Uma vez que exiba determinada opinião ou posicionamento político em sua atividade no Facebook, por exemplo, o usuário possibilita a uma empresa que tenha acesso aos seus dados (e aos dados dos demais indivíduos na sua rede de relacionamentos) exibir propaganda política a ele direcionada, inclusive fazendo uso de notícias falsas, a serviço da campanha de determinado candidato. Isso pode, por sua vez, reforçar as disposições e crenças do usuário, tal como ocorre nas filter bubbles (bolhas-filtro) e echo chambers (câmaras de eco). Com o uso de Big Data, estas empresas nem mesmo precisam que o usuário explicite sua posição política, por exemplo, à esquerda ou à direita: seus modelos lhes permitem identificar este posicionamento a partir de seus likes e de suas amizades online. Os Big Data têm assim servido à propagação de fake news a serviço, por exemplo, de campanhas políticas, identificando os caminhos para a sua disseminação mais rápida e duradoura.

Tais modelos, portanto, parecem funcionar bem quando visam solucionar problemas de "otimização de seleção", como os descreve Auerbach (2014b). Os dados ajudam a selecionar e priorizar os elementos mais relevantes de um conjunto segundo determinado critério, como os livros de um catálogo que um cliente possa querer comprar ou as máquinas de uma fábrica que possam apresentar mau funcionamento. Em casos como estes, falsos positivos têm pouco impacto: o modelo não será menos bem-sucedido se sugerir um livro que acabe não sendo comprado ou direcionar o funcionário a inspecionar uma máquina que esteja funcionando bem, desde que no panorama geral os números sejam positivos. "Os big data funcionam melhor em problemas que são tolerantes a erros." (AUERBACH, 2014b: online, tradução nossa), e nesses casos podem ajudar a diminuir gastos, economizar tempo, prevenir fraudes e acidentes, aumentar vendas, etc.

Quando se trata de problemas mais complexos, porém, realizar predições acerca do futuro de um fenômeno sem compreender as causas do mesmo, como prometem os Big Data (DUPUY, 2018: 150), não é o suficiente. Como notam Ekbia et al.: "Abandonando explicações mecânicas, algumas perspectivas radicais em Big Data procuram salvar o fenômeno simplesmente salvando a aparência. Ao fazê-lo, eles colapsam a diferença entre os dois: o fenômeno se torna aparência." (2015: 1529, grifo dos autores, tradução nossa). Como vimos no exemplo do PredPol, um modelo baseado em Big Data pode assim resultar em uma representação equivocada de um fenômeno, ao captar

características superficiais do mesmo, ignorando dinâmicas causais (falta de investimentos públicos, desigualdade de acesso a bens culturais, diferenças de classe social, entre outros) já evidenciadas por teorias consagradas a seu respeito.

Uma possível razão para isso seria a pouca adequação dos Big Data à realidade dos fenômenos cujos estados futuros eles pretendem prever, por se tratarem de sistemas complexos auto-organizados. Tal possibilidade é apontada por Auerbach – "Os dados não atuam em sistemas fechados." (2014a: online, tradução nossa) – e por Crawford, Miltner e Gray: "[...] ações individuais agregadas não podem, por si mesmas, ilustrar as complicadas dinâmicas que produzem interações sociais – o todo da sociedade é maior que a soma de suas partes." (2014: 1667, tradução nossa). Ekbia et al. notam, por outro lado, que a defesa do poder preditivo dos Big Data pode estar fundamentada em uma visão nomológica de mundo, pouco compatível com a explicação da conduta de agentes humanos:

> [...] os Big Data [...] parecem se contentar apenas com a predição das aparências. O interesse na predição das aparências, e a coleta das evidências necessárias para lhes dar suporte, ainda se origina do modelo dedutivo-nomológico, bastante estreito, que presume uma realidade governada por leis. Nas ciências sociais, este modelo pode ser rejeitado em favor de outras formas de explicação que permitam os efeitos da intencionalidade e livre-arbítrio humanos, e que requeiram a interpretação dos significados de eventos e de comportamentos humanos de uma maneira sensível ao contexto [...] (2015: 1530, tradução nossa).

Nos termos de Weaver, podemos dizer que os Big Data podem auxiliar na solução de problemas de complexidade desorganizada, isto é, aqueles: "[...] em que o número de variáveis é muito grande, e em que cada uma das muitas variáveis têm um comportamento que é individualmente errático, ou talvez totalmente desconhecido." (1948: 538, tradução nossa), mas cujo todo exibe certas regularidades matematicamente analisáveis. Por outro lado, modelos estatísticos por si mesmos teriam dificuldades para solucionar problemas de complexidade organizada, como aqueles das ciências econômicas, biológicas, sociais, etc., "[...] que envolvem lidar simultaneamente com um considerável número de fatores que estão inter-relacionados em um todo orgânico." (WEAVER, 1948: 539, grifo do autor, tradução nossa).

A noção de sistema complexo auto-organizado é objeto de discussão em diferentes campos, como a filosofia, a sociologia, a biologia e a cibernética, e se apoia em noções da física, em especial da termodinâmica, da biologia e da

lógica. Tal noção é essencial para a compreensão das dinâmicas subjacentes a tais tipos de problema, que incluem aqueles que envolvem a conduta humana, e que os Big Data frequentemente pretendem modelar. Segundo Mitchell, tais sistemas são aqueles em que um comportamento organizado de um todo complexo, de difícil previsão, emerge a partir do comportamento dos elementos simples que o compõem, sem a atuação de um controlador central (2009: 13). Ou seja, o sistema é auto-organizado quando, como ressalta Debrun, sua estrutura e funcionalidade se deve em maior grau às interações entre seus elementos – o que pressupõe algum grau de autonomia dos mesmos –, e em menor grau às suas condições iniciais, ao seu ambiente ou à presença de uma instância supervisora (2018: 8). Bresciani Filho e D'Ottaviano acrescentam ainda que a auto-organização requer um processo recorrente de interação entre os elementos de um sistema, de forma que estes possam se integrar em uma organização com auto-referência (2018: 61).

Os elementos que compõem um sistema interagem ou comunicam-se uns com os outros e, de acordo com Ashby, tais interações restringem, isto é, limitam o espaço de possibilidades de ação dos elementos envolvidos e, consequentemente, do sistema como um todo (1962: 257). Isso faz emergir o que Haken (2000) chama de "parâmetros de ordem" do sistema, padrões emergentes de alto nível, em oposição aos "parâmetros de controle", características de baixo nível que atuam na emergência dos parâmetros de ordem. Conforme explicitam Gonzalez, Broens, Haselager e Bresciani Filho:

> [...] parâmetros de ordem podem ser descritos como padrões de alto nível que resultam da interação entre componentes de baixo nível. Uma vez criados, parâmetros de ordem constrangem (escravizam) o comportamento dos componentes de baixo nível dos quais eles se originam. Estes podem alterar, na forma de feedback circular, os parâmetros de ordem de alto nível [...] (2005: 383, tradução nossa).

Haken chama essa dinâmica em que as partes agem sobre o todo e o todo retroage sobre as partes de "causalidade circular" (2000: 25). É a causalidade circular inerente a um sistema complexo auto-organizado, como é o caso daqueles em que agentes humanos inter(agem), que possibilita a ocorrência, por exemplo, de profecias autorrealizáveis. A causalidade circular pode possibilitar também que a divulgação e/ou aplicação dos resultados de um modelo baseado em Big Data interfira no comportamento do fenômeno modelado. Isso pode alterar os dados que alimentam o modelo e as tendências gerais do todo, podendo inclusive corroborar o modelo, que acaba se comportando, mais uma vez, como uma profecia autorrealizável.

Faz-se necessário, portanto, investigar a seguinte questão: os modelos estatísticos baseados em correlações entre grandes quantidades de dados em constante influxo – os Big Data – são ou não adequados à tarefa de descrever e prever o comportamento de sistemas complexos auto-organizados, em constante alteração devido à dinâmica da causalidade circular ou, ao menos, de descrever e prever o comportamento de elementos particulares – ou conjuntos dos mesmos – destes sistemas?

Uma análise das compatibilidades e incompatibilidades entre os pressupostos teóricos dos Big Data e aqueles da teoria dos sistemas complexos aponta na direção da conclusão de que modelos estatísticos baseados em Big Data são apenas parcialmente adequados à descrição e predição do comportamento de sistemas complexos auto-organizados; ou, mais especificamente, que tais modelos (uma vez que estejam isentos dos muitos problemas de coleta, seleção, análise, modelagem, etc. de dados que frequentemente ocorrem no contexto dos Big Data) podem ser capazes de descrever e prever, com uma margem variável de acuidade, os comportamentos de elementos particulares de um sistema ou de conjuntos dos mesmos, através da análise de padrões estatísticos correlacionais encontrados nos comportamentos de elementos particulares do mesmo sistema (desde que a dinâmica do mesmo não altere drasticamente os comportamentos dos seus elementos particulares, ou desde que a aplicação do modelo não seja um elemento atuante no sistema). Porém, uma vez que a emergência do comportamento geral do sistema depende não apenas de características dos seus elementos particulares ou de outros parâmetros de controle (ambientais, por exemplo), mas de uma dinâmica em que atua uma causalidade circular, na qual as interações entre os elementos particulares, influenciadas por outros parâmetros de controle, são a causa do comportamento geral do sistema que, através da ação dos parâmetros de ordem, é também a causa de certos comportamentos dos elementos particulares, uma análise que não leve em conta a causalidade não poderá descrever ou prever corretamente o comportamento geral do sistema com base apenas em correlações entre elementos particulares.

No entanto, faz-se necessário também evidenciar as técnicas de análise de dados empregadas por diferentes modelos baseados em Big Data – tanto daqueles considerados bem-sucedidos quanto daqueles considerados falhos ou problemáticos –, bem como os possíveis vieses a eles subjacentes, a fim de explicitar as capacidades e os limites dos mesmos. Uma vez que os algoritmos destes modelos em geral não são de acesso público, é preciso identificar os padrões recorrentes presentes nos resultados das aplicações dos modelos.

Assim, acreditamos que tornar claro o funcionamento destes modelos e evidenciar as consequências de sua aplicação é de extrema importância, dado o papel cada vez maior que eles têm exercido na sociedade informatizada em que vivemos.

referências

CHRIS ANDERSON. End of Theory: Will the Data Deluge Make the Scientific Method Obsolete? **Wired Magazine**. [online] Disponível em: <https://www.wired.com/2008/06/pb-theory>. Acesso em 01/08/2018.

ASHBY, William Ross. Principles of the self-organizing system. In: VON FOERSTER, Heinz; ZOPF, George. (eds.), **Principles of Self-Organization**: Transactions of the University of Illinois Symposium. London: Pergamon Press, 1962, p. 255-278.

DAVID AUERBACH. Big Data is Overrated: And Ok-Cupid's User Experiments Prove It. **Slate**. [online] 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 01/08/2018.

_____. The Big Data Paradox: It's never complete, and it's always messy – and if it's not, you can't trust it. In: **Slate**. [online] Disponível em: <http://www.slate.com/articles/technology/bitwise/2014/08/what_is_big_data_good_for_incremental_change_not_big_paradigm_shifts.html>. Acesso em 01/08/2018.

BOWKER, Geoffrey. **Memory Practices in the Sciences**. Cambridge: MIT Press, 2005.

_____. The Theory/Data Thing: Commentary. **International Journal of Communication**, Los Angeles, v. 8, 2014, p. 1795-1799.

BOYD, danah; CRAWFORD, Kate. Critical Questions for Big Data. **Information, Communication & Society**, London, v. 15, n. 5, 2012, p. 662-679.

BRESCIANI FILHO, Ettore; D'OTTAVIANO, Itala Maria Loffredo. Basic Concepts of Systemics. In: PEREIRA JR, Alfredo; PICKERING, William; GUDWIN, Ricardo. (eds.) **Systems, self-organization and information**: An Interdisciplinary Perspective. New York: Routledge, 2018, p. 48-64.

KATE CRAWFORD. The hidden biases in Big Data. In: **Harvard Business Review**. [online] Disponível em: <http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html>. Acesso em 01/08/2018.

_____.; MILTNER, Kate; GRAY, Mary. Critiquing Big Data: Politics, Ethics, Epistemology. **International Journal of Communication**, Los Angeles, v. 8, 2014, p. 1663-1672.

DEBRUN, Michel. The Idea of Self-Organization. In: PEREIRA JR, Alfredo; PICKERING, William; GUDWIN, Ricardo. (eds.) **Systems, self-organization and information**: An Interdisciplinary Perspective. New York: Routledge, 2018, p. 7-22.

DUPUY, Jean-Pierre. Science without Philosophy: The Case of Big Data. **Crisis and Critique**, v. 5, n. 1, 2018, p. 147-161.

EKBIA, Hamid et al. Big Data, Bigger Dilemmas: A Critical Review. **Journal of the Association for Information Science and Technology**, Hoboken, v. 66, n. 8, 2015, p. 1523-1545.

GONZALEZ, Maria Eunice Quilici; BROENS, Mariana Claudia; HASELAGER, Willem; BRESCIANI FILHO, Ettore. Self-organization and Life: A Systemic Approach. **Manuscrito**, Campinas, v. 28, n. 2, 2005, p. 375-390.

HAKEN, Hermann. **Information and Self-Organization**: A Macroscopic Approach to Complex Systems. Berlin: Springer Verlag, 2000.

IBEKWE-SANJUAN, Fidelia; BOWKER, Geoffrey. Implications of Big Data for Knowledge Organization. **Knowledge Organization**, Würzburg, n. 44, v. 3, 2017, p. 187-198.

KITCHIN, Rob. Big Data, new epistemologies and paradigm shifts. **Big Data & Society**, Thousand Oaks, 2014, p.1-12.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big Data**: A Revolution That Will Transform How We Live, Work and Think. New York: Houghton Mifflin Harcourt, 2013.

MERTON, Robert. The Self-Fulfilling Prophecy. **The Antioch Review**, Yellow Springs, v. 8, n. 2, 1948, p. 193-210.

MITCHELL, Melanie. **Complexity**: A Guided Tour. Oxford: Oxford University Press, 2009.

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

SCHUTT, Rachel; O'NEIL, Cathy. **Doing Data Science**. Sebastopol: O'Reilly Media, 2014.

WEAVER, Warren. Science and Complexity. **American Scientist**, Research Triangle Park, v. 36, 1948, p. 536-544.

autores

Ricardo Peraça Cavassane | ricardo.peraca@gmail.com
Doutorando do Programa de Pós-graduação
em Filosofia da Universidade Estadual de Campinas – Unicamp.