Quão big o seu data deve ser?

Ronaldo Baltar[1] and Claudia Siqueira Baltar[1]

[1]Affiliation not available

April 12, 2024

Resumo

Esse conteúdo digital foi criado como parte da atividade de capacitação "Introdução ao R e RStudio para ciência social computacional". É uma iniciativa de formação e difusão científica do Observatório de Populações e Políticas Púbicas (ObPPP) e do Programa de Formação Complementar de Graduação em Pesquisa Social Computacional (InfoSoc), ambos vinculados ao Dept.º C. Soc. / CLCH, Universidade Estadual de Londrina (UEL), Paraná - Brasil

Com grandes dados, grandes desafios

Bigdata é necessário para garantir cientificidade para uma pesquisa computacional baseada em dados?

A preocupação com o aprimoramento de técnicas para enfrentar o desafio do aumento da capacidade de armazenamento e do processamento de informações não é recente. O conhecimento é cumulativo e a memória das pessoas não é suficiente para garantir a preservação e retransmissão dos conteúdos criados pelas diferentes atividades humanas.

De certa forma, podemos pensar que a invenção da escrita, o papel, a imprensa foram respostas tecnológicas para o problema do crescente acúmulo de registros e a busca de informações.

Com a expansão do uso de computadores, ao final da Segunda Guerra Mundial, os conteúdos da produção de conhecimento, em todas as áreas, foram migrando cada vez mais para o formato digital.

Nos anos 90, fabricantes, desenvolvedores de *software* e pesquisadores, nos Estados Unidos, começaram a projetar a curva de crescimento futuro dos serviços de informática e observaram que o volume de dados tendia a aumentar mais rapidamente do que a capacidade de armazenamento e processamento.

O termo *bigdata* foi criado para expressar a preocupação com o volume de dados que tende a crescer mais rapidamente do que a capacidade de processamento prevista.

Bigdata é um problema, não um método de análise

Os métodos são resultados da tentativa de resolver algum problema. O que se chama de *bigdata* é um problema relativo à desproporção entre a capacidade de processamento e o tamanho dos dados. Esse é um problema relativo: o que foi *bigdata* no passado agora pode ter se tornado uma rotina. Do mesmo modo que, no futuro, poderá não haver dificuldades para o processamento e análise da volumosa quantidade de dados disponíveis atualmente. Mas, certamente, ainda haverá o desafio do *bigdata*, porque a humanidade terá gerado mais informações do que os mais avançados recursos disponíveis conseguirão armazenar e processar.

No século XVI, por exemplo, os registros diários dos movimentos dos planetas e do sol, feitos pela equipe do astrônomo Thyco Brahe, poderiam ser considerados um problema de *bigdata*, que exigiu novos instrumentos e métodos de coleta e registro da informação. A coleção de registros acurados de Brahe ajudaram Kepler a demonstrar, com precisão, a teoria heliocêntrica de Copérnico.

Para a época, o trabalho de execução, correção e revisão dos cálculos do movimento de cada planeta era um desafio que poucas pessoas poderiam realizar. Hoje, estudantes de ensino médio podem ver essas simulações em uma única aula com auxilio da computação.

Os censos demográficos são outro exemplo de problema relativo de *bigdata*. Além da dificuldade logística da coleta de dados, principalmente em um país com a dimensão do Brasil, o censo demográfico apresenta também a dificuldade de processamento das informações.

Em uma época em que não havia nenhum meio computacional ou mesmo mecânico de tabulação dos resultados, como no Recenseamento Geral do Império, quando tudo era feito a bico de pena, o recenseamento dos 10.112.061 habitantes do país foi um grande desafio. Os dez anos de intervalo entre os censos demográficos eram necessários não só pelo custo, mas pelo tempo adequado para que fossem tabulados todos os resultados.

Em 2010, o recenseamento de 190 milhões de habitantes já não pode ser considerado um problema de *bigdata* para o IBGE. Todos os dados do censo 2010 podem ser baixados e rodados em um computador comum.

O Facebook, com seus estimados 500 Tera bytes de dados gerados diariamente, é um exemplo de desafio atual de *bigdata*. Muitos recursos especiais de hardware e software são necessários para garantir as quase 3 bilhões de curtidas e 300 milhões de fotos que são compartilhadas pelo Facebook a cada 24 horas.

O tamanho atual do problema bigdata

Segundo a avaliação da empresa International Data Corporation (IDC) , em 2020, o mundo havia criado e replicado cerca de 64,2 Zettabytes de dados, enquanto a capacidade de armazenamento mundial naquele ano foi prevista como sendo de 6,7 Zettabytes.

Um zettabyte equivale a 1.000.000.000.000.000.000.000 de bytes ou um trilhão de gigabytes.

Como isso é possível? Esses números indicam que, aproximadamente, apenas 10% dos dados criados são armazenados. Os 90% restantes acabam tendo uma existência fugaz. Desaparecem para dar espaço aos novos dados. Essa tendência deve continuar para os próximos anos, o IDC estima que o crescimento da capacidade de armazenamento terá um ritmo bem menor do que o crescimento da capacidade de geração de dados, sobretudo com a popularização da Internet das coisas.

Os desafios para lidar com a disparidade entre o volume de informações produzidas pela humanidade e estratégias de armazenamento, recuperação e análise de todo esse conteúdo, têm sido um dos impulsionadores da inovação em métodos e técnicas desde a antiguidade.

Bigdata: o anúncio da era dos dados sem ciência

No ano de 2008, a revista norte-americana Wired publicou um artigo do seu então prestigiado editor chefe , Chris Anderson . O artigo tinha um título provocativo: "*The end of theory: the data deluge makes the scientific method obsolete*" ("O fim da teoria: o dilúvio de dados que tornou o método científico obsoleto") e alcançou um grande público, que difundiu o termo *bigdata* entre influenciadores, jornalistas, empresários, profissionais de *marketing*, *coachs* e consumidores do mundo digital.

Em seu artigo, Anderson costurou argumentos aparentemente lógicos que deram visibilidade para uma crença que já estava sendo difundida entre usuários e desenvolvedores da emergente ciência dos dados: a ciência está nos dados e não no pensamento. O artigo começa com uma paráfrase do estatístico George Box: "Todos os modelos estão errados, mas alguns são úteis".

Logo, quanto mais dados, mais ciência! Teorias seriam argumentos usados apenas para cobrir, de forma imperfeita, a falta ou a insuficiência de dados. Segundo Anderson, no século XXI, a abundância dos dados tornava desnecessárias hipóteses, modelos e explicações causais.

Muitos profissionais de ciência dos dados, marketing e computação, trabalhando em empresas, viam (e muitos ainda veem) os pré-requisitos da pesquisa científica, incluindo a estatística, como formalismo acadêmico desnecessário ao trabalho prático.

Por exemplo, um analista de dados, trabalhando em uma loja de departamento, para fornecer resultados para o pessoal de *marketing* sobre quais as características de um produto são mais pesquisadas entre grupos etários de usuários, não precisaria "testar hipóteses estatísticas" ou recorrer a nenhum tipo de conceito "teórico" de consumidor. Bastaria rodar todos os dados (supondo que a loja de departamento tivesse o equipamento adequado para atender a capacidade de processamento exigida) que um padrão válido sobre o comportamento dos consumidores se apresentaria como resultado.

A proposição de Anderson não era simplista, mas representava uma tendência que ajudou em disseminar a crença ingênua de que quanto mais dados, mais corretas estão as respostas. Os computadores com alta capacidade de processamento e armazenamento de dados poderiam encontrar padrões de correlação entre as informações. Se os dados forem suficientemente grandes, essas correlações seriam significativas, tornando desnecessários conceitos e testes estatísticos. Difundiu-se a ideia de que o *bigdata* tornava obsoleto o conhecimento científico (e acadêmico) acumulado até então. A nova ciência precisaria apenas da abundância de dados do mundo digitalizado e muita capacidade de processamento.

O volume de dados por si só não é uma garantia de que um estudo sobre qualquer fenômeno possa estar correto. Não é o *bigdata* que establece a cientificidade de um procedimento analítico.

Passado o deslumbre inicial com a avalanche de dados que disseminou o *bigdata*, proposições como a de Anderson passaram a receber sérias críticas entre os analistas e profissionais da área. O volume de dados não é suficiente para garantir uma explicação correta sobre qualquer fenômeno da realidade.

Podemos tomar como exemplo novamente o caso de Thyco Brahe e a revolução que possibilitou que o heliocentrismo de Copérnico e Kepler pudessem suplantar séculos de pensamento centrado na Terra como centro do universo, chamado de modelo geocêntrico.

Thyco Brahe não concordava com a teoria de Copérnico. Era um estudioso da obra de Ptolomeue defendia o geocentrismo. No século III, Ptolomeu expôs, no seu Almagesto, vários elementos de matemática e astronomia, entre os quais um modelo que explicava o movimento dos planetas a partir de uma série de círculos concêntricos. A obra foi uma das principais referências científicas desde o fim do Império Romano até o fim da Idade Média. Foi lida e estudada tanto na Europa cristã, quanto no mundo islâmico (de onde vem o nome Almagesto , corruptela latina da tradução em árabe, cujo título original em grego significa "A grande coleção").

Desde que os povos começaram a olhar para o céu de forma mais sistemática, provavelmente para marcar as mudanças nas estações, muitas explicações foram tentadas para explicar o movimento dos planetas. Visto da Terra, parece que todos os corpos celestes seguem uma direção, como um conjunto coordenado de movimentos.

Exceto pelos planetas (palavra que em grego significa um "errante"), que percorrem trajetórias diferentes. Marte, por exemplo, parece reverter o seu movimento, voltando no céu em relação às demais estrelas, para, em seguida, seguir seu caminho adiante. Aristóteles, no século IV AC, propôs uma teoria para explicar os movimentos dos corpos celestes.

Para o pensador grego, todos os corpos possuíam um tipo de movimento natural. Na Terra, o movimento natural seria retilíneo para cima (leviandade) ou para baixo (gravidade). Qualquer movimento diferente só poderia ser causado por intervenção de uma força externa. No céu seria diferente. Os corpos celestes seriam feitos de uma matéria própria, chamada por ele de éter, cujo movimento natural seria circular. A Terra seria o centro do universo e estaria em repouso, todo o resto estaria se movendo em torno da Terra.

É importante ressaltar que a teoria do geocentrismo não é o mesmo que "terraplanismo". No século IV AC, os gregos já sabiam que a Terra era esférica e já até haviam calculado com certa precisão o seu diâmetro. Aristóteles sabia que a posição das constelações variava quando vistas de regiões diferentes. Mas Aristóteles misturou suas impressões sobre a realidade com valores filosóficos e morais, como a ideia de que o universo é imutável e perfeito.

Sua explicação física do movimento dos corpos celestes demorou quase dois séculos para ser refutada, na medida em que seus argumentos foram incorporados primeiro pelo islamismo e, posteriormente, pelo cristianismo, sustentando o postulado da imutabilidade e perfeição da criação divina. Os planetas desafiavam a ideia de perfeição do universo aristotélico, se pareciam com corpos que vagueavam pelo céu sem uma explicação lógica.

Cláudio Ptolomeu, em seu Almagesto, sistematizou a teoria geocêntrica e propôs o modelo dos movimentos planetários a partir da concepção de órbitas excêntricas, o que contrariava a ideia de perfeição do movimento circular dos corpos no éter, tal como formulada por Aristóteles.

O modelo geocêntrico de Ptolomeu sistematizava grandes avanços nos conhecimentos em matemática e geometria. Conseguia dar explicações parciais ao problema do movimento retrógrado dos planetas e apresentava soluções lógicas para o movimento do sol ao redor da Terra e outros problemas de astronomia da época. No entanto, era um modelo equivocado sobre o movimento dos corpos celestes. Conforme a citação do estatístico George Box, esse seria um exemplo de modelo útil, porém errado.

Mais de mil anos depois do Almagesto ter sido escrito, grandes astrônomos, como Tycho Brahe, ainda utilizavam intensivamente o livro escrito por Ptolomeu como texto de referência e autoridade intelectual para explicar os movimentos dos corpos celestes. A teoria geocêntrica estava de acordo com a doutrina das Igrejas cristãs e mulçumanas, que colocavam o ser humano como centro da criação divina.

Em seu observatório, Brahe criou vários instrumentos para medir, com o máximo rigor possível, a posição das estrelas e planetas, anotando com sua equipe, cuidadosamente todos os movimentos celestes observados.

Um dos objetivos era fornecer dados suficientemente apurados para corrigir as imprecisões do Almagesto de Ptolomeu.

Enquanto Tycho Brahe trabalhava em seu observatório, na Dinamarca, o debate na Europa sobre heliocentrismo e geocentrismo percorria alguns círculos intelectuais da época. A proposição sobre o modelo que tinha o Sol como centro do movimento da Terra e dos planetas era tratada como conjectura filosófica, não como hipótese alternativa, dado que contrariava os postulados bíblicos, que deveriam ser a única fonte de autoridade em disputas argumentativas. Mas, desde meados do século XVI, o modelo heliocêntrico contava com uma fundamentação lógica e matemática, graças às ideias de Copérnico, publicadas em seu livro "As Revoluções dos Orbes Celestes".

Copérnico não foi o primeiro a propor o modelo heliocêntrico. Antes até de Aristóteles, Pitágoras já postulava que a Terra se movia em torno de um "fogo central". Para a seita Pitagórica, o fogo era o elemento primordial do Universo, logo, nada mais natural do que tudo estar ao redor de uma grande fogueira espacial. No modelo de Pitágoras, o Sol era apenas mais um astro a rodear o "fogo central".

Aristarco, também grego, no século III AC, usando os conhecimentos de geometria da época, calculou o diâmetro da Terra e propôs, por suas projeções, que o Sol seria sete vezes maior do que a Terra. Com base nessa informação, Aristarco concluiu que seria mais lógico o objeto menor, no caso a Terra, girar ao redor do objeto maior, o Sol. Assim, a novidade de Copérnico não estava na proposição de que a Terra girasse ao redor do Sol, mas na construção de um modelo que conseguia responder as mesmas indagações que levaram ao modelo ptolomaico, como o movimento retrógrado dos planetas, ao mesmo tempo que apresentava solução para os problemas não respondidos ou criados pelo modelo do Almagesto.

Tycho Brahe concordava com a tese de Ptolomeu e discordava da ideia de Copérnico. O seu trabalho minucioso de coleta de dados, que poderia ser considerado um desfio de *bigdata* para aquele tempo, pretendia confirmar e dar mais precisão ao modelo geocêntrico. Um de seus auxiliares, o alemão Johannes Kepler, após a morte de Tycho, compilou os dados e utilizou-os para construir um modelo heliocêntrico mais apurado do que o de Copérnico, prevendo as órbitas elípticas e não circulares dos planetas ao redor do Sol.

Ao contrário da citação de George Box, o modelo de Kepler foi útil e correto ao mesmo tempo, pois representava de fato o movimento planetário, que, agora no século XXI, pode ser confirmado com várias outras fontes de dados e observações espaciais. Tycho Brahe poderia ter coletado mais dados, caso talvez tivesse mais tempo e recursos, e isso, muito provavelmente, não o faria trocar sua concepção geocentrista pelo modelo heliocêntrico. Copérnico não dispunha dos dados de Tycho Brahe, quando formulou seu modelo de órbitas circulares dos planetas ao redor do Sol.

O oposto ao *bigdata*, a presunção de uma ciência sem dados!

Os dados são desnecessários?

Se Baher trabalhou tanto para construir uma coleção minuciosa de dados sobre o movimento dos planetas e ainda acreditava na teoria incorreta do geocentrismo, e Copérnico, antes dos dados de Tycho Baher, já havia criado o modelo heliocêntrico correto, então pode-se concluir que os dados são desnecessários para o avanço do conhecimento humano?

A resposta certamente é não, os dados não são dispensáveis.

O oposto da proposição de que os dados são suficientes para o entendimento dos fenômenos (e dispensam a ciência) está na visão ingênua de que a genialidade humana provém do pensamento que se isola da realidade.

O conhecimento requer pensamento, observação e fundamentação dos argumentos baseados em evidências. Não há ciência sem dados, assim como os dados, independente do volume, não são explicativos por si só, dependem de modelos (teóricos) que explique as relações causais, os mecanismos e os processos a partir dos quais os fenômenos se originam.

As pessoas não são como os planetas!

As pessoas não são planetas e o movimento da sociedade não pode ser comparado ao movimento sincronizado dos corpos celestes no céu. Mas essa diferença não invalida, para as ciências sociais, a necessidade de se buscar modelos explicativos confrontados com os dados adequados. Embora o método das ciências da natureza seja distinto, em grande parte, do método das ciências sociais, em essência, a construção de modelos teóricos fundamentados em evidências é também a chave explicativa dos processos de mudança social.

Em qualquer campo do conhecimento, do mesmo modo que diante de um fenômeno social, para buscar respostas aos problemas de investigação, primeiro formula-se conjecturas. Quando as conjecturas conseguem ser sistematizadas, passam a se constituir em modelos teóricos (hipotéticos). Mas nenhum modelo teórico se autoevidencia. São necessários dados para tornar evidente um modelo. Por meio das coleções de evidências ou conjuntos de dados é que se pode sustentar ou refutar um modelo teórico explicativo qualquer. Dados sem teoria são apenas descrições. Teoria sem dados são apenas conjecturas, por mais autoridade e respeitabilidade que tenham seus autores.

Um modelo pode ser concebido a partir de dados prévios, da experiência, da refutação de outras teorias ou até da imaginação. O que importa é que o modelo seja demonstrável a partir de dados e evidências. Os dados podem ter erros, muitos dos quais são corrigíveis. O erro incorrigível é a falta de dados.

Do ponto de vista metodológico, a combinação entre as capacidades humanas de reflexão (teoria) e de observação (dados) é o que permite a formulação de conclusões plausíveis e verificáveis sobre os problemas que nos instigam a imaginação ou nos cobram providências práticas.

Se optarmos por fazer ciência, seja qual ramo do conhecimento for, estamos optando por buscar explicações fundamentadas em evidências. As evidências devem ser buscadas em fenômenos, isto é, manifestações concretas da realidade. O entendimento de um fenômeno ocorre quando há uma explicação demonstrável para sua manifestação.

A falta de dados é um problema para a demonstração de evidências. Mas a abundância de dados também não é uma garantia de que se tem uma evidência para algum modelo. A explicação demonstrável ocorre quando se consegue ter os dados necessários para tornar aparentes as interconexões entre os fatos ou eventos que tornam possível a manifestação do fenômeno social que é objeto do estudo.

A compreensão requer mais do que dados volumosos. A compreensão ocorre quando, tendo havido o entendimento de um fenômeno, é possível conhecer as suas implicações e para tal é necessário um modelo, no mínimo, aproximadamente correto. A explicação que se sustenta na autoridade de argumentos ou em fenômenos não demonstráveis, no melhor dos casos, pode ser considerado como uma hipótese a ser investigada. Se não for tomada como hipótese, será um dogma. E um dogma é uma postura anticientífica, que se nutre do negacionismo da realidade, independente do *bigdata* disponível, como o mundo pôde testemunhar com a disputa entre o geocentrismo e o heliocentrismo, e, recentemente, com o negacionismo que aflorou com a pandemia de Covid -19.

Pode-se destacar que:

- O termo *bigdata* não deve ser usado como um argumento de autoridade.

- Não é o volume de dados que vai garantir a cientificidade do seu trabalho, nem tampouco a veracidade ou a credibilidade dos seus resultados.
- Com um modelo inadequado, qualquer quantidade de dados apresentará respostas errôneas ou inconclusas.
- Entre nenhum dado e muitos dados, o que uma pesquisa precisa é de dados representativos, adequados ao problema de pesquisa e metodicamente revistos e escrutinados.
- Os dados devem representar acuradamente um fenômeno que se quer estudar. Para isso, precisam ser construídos dentro de um escopo lógico articulado a um modelo causal metodicamente elaborado.

Sugestões de leitura

Baltar, Ronaldo, e Cláudia Siqueira Baltar. 2013. "As ciências sociais na era do zettabyte". *Mediações* 18 (1): 11–18.

Baltar, Ronaldo e Cláudia Siqueira Baltar. 2021. "Reflexões sobre os percursos metodológicos em ciência social computacional". *Simbiótica. Revista Eletrônica* 8 (4): 17–45.