IX CONGRESSO PORTUGUÊS DE SOCIOLOGIA
Portugal, território de territórios

ÁREA TEMÁTICA: Teorias e Metodologias [AT]

GRANDES DADOS E OS DESAFIOS ATUAIS PARA A PRODUÇÃO DE CONHECIMENTO

FREITAS, Francisco

Investigador Júnior, Centro de Estudos Sociais da Universidade de Coimbra, franciscofreitas@ces.uc.pt

MENDES, José Manuel

Professor Auxiliar com Agregação, Faculdade de Economia da Universidade de Coimbra, jomendes@fe.uc.pt

Resumo

Os grandes dados (*big data*) constituem uma das mais recentes expressões da sociedade em rede. Ao fenómeno tem sido associado um entusiasmo capaz de perspetivar um considerável conjunto de utilizações e de impactos. Os grandes dados constituem um movimento ainda caracterizado pela incerteza. Contudo, os grandes dados assentam em transações e repositórios válidos com múltiplos pontos de contacto com a realidade. São repositórios que possuem dimensões sociais e que implicam consequências de ordem diversa. De acordo com o exposto, as metodologias e técnicas de investigação em ciências sociais são agora mobilizados para a compreensão deste novo contexto. Estão, por isso, sujeitas às especificidades próprias decorrentes da intervenção e procura de sentido neste domínio do social. De outra forma, é de considerar que a resposta a determinadas questões passará, também, pela utilização destes novos repositórios de dados. Porém, a sua interpretação carece de novas abordagens e estratégias que permitam investigação realmente operante face aos desafios que impõe. Partindo do princípio pelo qual é importante dar resposta a este novo contexto e que o mesmo deve ser compreendido, este artigo visa, como tal, enumerar, de forma não exaustiva, alguns dos desafios suscitados pelos grandes dados em diálogo com as metodologias de investigação tradicionais.

Abstract

Big Data is a new expression inside the network society. There is a growing infatuation surrounding the phenomenon, generating a wide list of prospective uses and impacts for Big Data. Big Data is reviewed as a movement under great uncertainty. Nonetheless, Big Data is made of transactions and valid repositories establishing multiple contact points with social reality. These repositories include social dimensions, generating consequences of different order. According to the latter, social science research methodologies and techniques are thus mobilized for the analysis of this new context. Research methodologies are subjected to the specificities enacted by the search for meaning in this new social realm. In other words, we may consider that the answer to certain questions will be made available using the new data repositories. However, a correct interpretation demands for new approaches and strategies enabling a de facto research due to the existing challenges. Starting with the principle that the new context requires answers and reasoning, this article aims at presenting a few examples of challenges indorsed by Big Data in dialogue with traditional research methodologies.

Palavras-chave: Grande Números; Grandes Dados; Estatística; Amostragem; Privacidade.

Keywords: Large Numbers; Big Data; Statistics; Sampling; Privacy.

[COM0642]

Introdução

Os grandes dados representam uma das manifestações mais recentes da sociedade em rede. A sociedade em rede e o advento da Internet facultaram as condições para o armazenamento e análise de grandes volumes de informação digital sem paralelo, um campo que implica avanços, dificuldades de vária ordem, desafios éticos e legais (Huberman, 2012; Mayer-Schönberger & Cukier, 2013). É um fenómeno acercado ainda por uma enorme incerteza, sobre o qual convergem visões bastante distintas, importando distinguir aplicações e implicações. Sobre as implicações, ainda não é claro de que forma os grandes dados alteram estruturas de poder, processos de decisão, ou processos políticos, elementos fundamentais para a atividade das ciências sociais. Se revistas as suas aplicações, emerge sobretudo uma visão prospetiva, que antecipa avanços na prevenção de doenças, no combate ao crime, na identificação de tendência de negócio, na gestão das cidades, ou até no redesenho do próprio método científico. Este artigo[1] visa, por isso, problematizar, de forma não exaustiva, a tensão estabelecida entre a aplicação dos grandes dados e as implicações que estes acarretam em termos das metodologias de análise em ciências sociais. Serão tidas em conta dimensões importantes como privacidade e ética ou a prática de amostragem.

Sociedades ricas em dados

A fase atual de abundância de dados e de grandes dados corresponderá em boa certeza a mais uma etapa de um processo histórico alargado relativo a uma "sociedade quantificada". Desde a emergência do conceito moderno de probabilidade ao ênfase corrente nos dados e informação, os números representam uma forma de linguagem socialmente aceite que é sublinhada pelo princípio da confiança nos números (Porter, 1996). Os números funcionam como ferramentas para a construção da realidade, assim como para a construção de sujeitos (Hacking, 2000, 2001). Não podendo ser desligados deste movimento, os grandes dados acrescentam outras possibilidades à lógica de permanente quantificação, sobretudo porque reúnem informação de natureza muito diversa que quase sempre pode ser reduzida à quantificação. Se durante séculos se alimentou este importante ênfase na mensuração, na normalização, na burocratização, os grandes dados poderão constituir novas expressões de "tecnologias de poder", novas formas de biopolítica (Foucault, 1982) em que novamente números e cifras estão no centro, coadjuvados agora por novos tipos de informação, concretamente informação temporal e geográfica, eticamente sensível.

Contemporaneamente surgem, assim, referências a sociedades inundadas em dados ou referências como *datification*, referências que procuram descrever o aparecimento de dados de toda a espécie, sejam em circuitos de vídeo, seja em sensores, seja em dados gerados por cidadãos individuais. Qualquer atividade de mensuração corresponderá ou sempre a uma decisão política (Didier & Tasset, 2013). Aos novos dados associa-se, por isso, uma nova governamentalidade no seio de sociedades modernas e pluralistas, nas quais a experiência humana é reduzida, uma vez mais, a números abstratos (Dean, 2009). Aos números, juntam-se agora uma palete muito diversificada de dados, incluindo localização, opiniões, preferência pessoais, atividades realizadas, consumo realizados, entre muitos outros. Pelo que incluem, os dados são agora uma nova mercadoria, sujeita a todo o tipo de transações num mercado amplo, capaz de desencadear inúmeros usos (OECD, 2013). Esta é a época da base de dados, em que desde o nascimento qualquer cidadão é medido, descrito, categorizado, processos e circularidades que remontam já ao Iluminismo e aos primeiros censos (Manovich, 1998).

Em anos recentes, emergiu então uma nova tecnologia de quantificação e de interpretação através de bases de dados cada vez mais elaboradas e complexas. Os grandes dados referem-se ao incremento de escala e de complexidade das bases de dados. Expectavelmente, os grandes dados requerem novos métodos e tecnologias para o processamento da informação quantitativa e qualitativa para produzirem aplicações válidas (Podesta, Pritzker, Moniz, Holdren, & Zients, 2014). Os grandes dados correspondem a uma tradução do acrónimo anglo-saxónico *Big Data* e são aqui definidos como as base de dados cuja dimensão está para lá das possibilidades de captura, armazenamento, gestão e análise por intermédio das aplicações de base de

dados tradicionais (Manyika *et al.*, 2011, p. 1). Comportam quatro características fundamentais, usualmente referidas como 4Vs: volume, variedade, velocidade, veracidade. São dados grandes em volume, de alta velocidade, de tipo diverso, i.e. de natureza estruturada e não estruturada, normalmente incluindo referências temporais e espaciais Este assomo tecnológico tem levado à colocação do foco na significância da escala e âmbito dos rastos digitais deixados por indivíduos, coisas, dinheiro e ideias em movimento (Boyd & Crawford, 2012).

A definição de grandes dados aqui sugerida é alargada, apresentando os grandes dados não apenas em função do volume de dados, mas antes da sua natureza e sobretudo da sua dinâmica. Quantidade não é, de todo, o elemento crucial, ainda que o termo original o possa sugerir, ou porque, como já foi referido, esta é mais uma tecnologia de quantificação, de redução da experiência humana a números. Analiticamente, as atividades associadas aos grandes dados constam de tarefas tais como a análise, a captura, a curadoria, a elaboração de pesquisas e consultas, o armazenamento, a transferência e reutilização, ou a visualização de dados. Estas são tarefas envolvidas na análise de bases de dados tradicionais, contudo aos grandes dados associam-se especificidades ancoradas num novo ecossistema inexistente até há pouco, para o qual não existia sequer a necessária capacidade de processamento ou as tecnologias que o tornam funcional – as primeiras referências surgem apenas nos anos 90 com John Mashey. Sem algumas das ferramentas avançadas e sobretudo sem microcomputação, os dados são ininteligíveis.

Independentemente das controvérsias que se têm gerado sobre o tema, os grandes dados inserem-se numa "revolução dos dados", de um fluxo sem paralelo de informação (Kitchin, 2014, p. 68). De sublinhar, novamente, que os grandes dados não se resumem a elementos de quantificação, abundando informação de tipo qualitativo, desde logo a informação gerada por utilizadores individualmente (e.g. atualização de estado em rede social). Todavia, por razões operacionais, as técnicas de análise de grandes dados assentam sobretudo na identificação de regularidades ou tendências, com uma forte componente matemática e estatística, mesmo quando considerados as abordagens propostas para a análise de informação qualitativa (e.g. *text mining*, *sentiment analysis*, ou *natural language processing*). A procura de sentido ou a interpretação é feita com a mediação de máquinas, modelos e algoritmos.

Novas possibilidades de quantificação e sobretudo novas possibilidades de controlo podem associar-se aos grandes dados. Eventualmente estaremos perante novas incertezas acionadas por tecnologias emergentes e sobretudo por esta fase de afluência de dados, particularmente no tocante à privacidade e ética quando se antecipa eventuais implicações decorrentes dos grandes dados (Mayer-Schönberger & Cukier, 2013). É importante referir que a proposta dos grandes dados surge em paralelo com os avanços na tecnologia e ciência, pela promoção da ligação digital de indivíduos por meio de redes, derivando dos avanços na computação/processamento e no cálculo. Este é, como foi referido, um ecossistema recente. A generalidade das ferramentas que deram origem a muitos dos repositórios atuais não existia até há pouco tempo. A crescente capacidade de computação faculta um ambiente sem paralelo de atividade que é direcionada por algoritmos (Pentland, 2015).

Novos recursos para as ciências sociais, possibilidades e implicações

Cada nova onda de inovação tecnológica acarreta disrupções de ordem diversa. Face aos grandes dados, surgem preocupações sobretudo de ordem ética e quanto à privacidade, algo que será analisado mais detalhadamente num momento ulterior. É necessário ter em conta que face à nova economia dos dados, boa parte das utilizações são estritamente comerciais, assim como o esforço de desenvolvimento é feito por grandes corporações. Este é um ecossistema altamente especializado, que exige recursos avançados. Ora, tal reclama para as ciências sociais um papel fulcral de análise crítica de todo este contexto sociotécnico. Por outras palavras, as diferentes disciplinas poderão recorrer a muitos dos novos repositórios de dados, mas é premente uma análise crítica do ecossistema que é ativado pelos grandes dados. De facto, se forem tidas em conta as aplicações existentes em torno do *slogan* dos grandes dados, se for tido em conta o conjunto de

soluções prometidas por todo o tipo de organizações que operam na área, será importante situar empiricamente os grandes dados, clarificar o recorrente determinismo tecnológico que rodeia todo este movimento, compreender as implicações (Dalton & Thatcher, 2014).

Sobre os novos recursos e sobretudo sobre os novos desafios para as ciências sociais, as metodologias de investigação poderão ser especialmente úteis para a ilustração de algumas das importantes questões de base em causa. Um ponto fundamental consta da dúvida se investigadores conseguirão encontrar formas de aceder, analisar, citar, preservar e proteger a informação, num contexto com especificidades novas (G. King, 2011, p. 719). Não são suficientes, portanto, as visões mais otimistas sobre o fenómeno e as possibilidades existentes. A tecnologia atual oferece algumas ferramentas e alguns métodos aos académicos para analisarem grandes dados, mas faltam inúmeras técnicas e ferramentas para melhor compreender as plataformas nas quais cidadãos comuns comandam vidas sociais rastreáveis, plataformas que contêm padrões de comportamento, crenças, atitudes, atividades, conexões (Salah, Manovich, Salah, & Chow, 2013, p. 411).

A um outro nível, um dos aspetos mais interessantes desta fase de progressão tecnológica é dada pelos múltiplos pontos de contacto que estabelece a cidadania. De facto, muitas das alterações são agora diretamente animadas por cidadãos e pela sua inserção num ecossistema que, como já foi referido, é relativamente complexo. Esta cidadania, independentemente da direção considerada ou de uma avaliação da participação realizada, toma parte neste ecossistema, acede a benefícios, enfrenta impactos e disrupções que esta vaga tecnológica inculca. Exemplificando, é neste contexto que surgem formulações como *data citizen*, i.e. de como o advento do "movimento do eu quantificado", do "genoma pessoal socializado", associado à monitorização de atividades e dados gerados por indivíduos através de um vasto repositório de biossensores atualmente disponíveis, que colocam a proteção de dados pessoais para lá da questão da privacidade, vista aqui como um conceito redutor num momento em que as identidades são apresentadas na nuvem (Gregory & C. Bowker, 2016); a formulação de *digital citizen*, i.e. todos os sujeitos que realizam algum tipo de reivindicação de direitos digitais, não aqueles que detêm a capacidade de ler, escrever, compreender e navegar por informação textual em linha, não aqueles que possuem algum tipo de acesso a banda larga (Isin & Ruppert, 2015).

Ao abrigo da dinâmica desta cidadania digital, informação inúmeras vezes sensível é armazenada em silos que são propriedade de entidades de tipo diferenciado, Estados incluídos. A partilha desta informação é geralmente desaconselhável, em teoria impossível se respeitados preceitos ético-legais. A economia digital alimenta-se de bens e serviços assentes na recolha e processamentos de dados com fins estritamente comerciais. O mercado dos dados inclui transações de vários tipos, revenda de dados incluída para lá dos termos estabelecidos em acordos ou consentimentos iniciais[2]. A combinação de bases de dados é hoje facilitada através de ferramentas várias que superam a ligação por identificadores das bases de dados relacionais. Estes pontos contrastam com um eventual mundo de pequenos dados ou das práticas científicas atuais, em que a curadoria e preservação são mais simples, onde o controlo é relativamente mais simples, em que as tarefas de anonimização não levantam constrangimentos de maior. Repare-se que no seio das ciências sociais, conceitos como "consentimento informado" ou de "proteção de sujeitos de pesquisa" foram já amplamente discutidos, como foi rebatida a aparente neutralidade moral e política de vários desses constructos, sendo prática corrente o ajuste desse dispositivo aos métodos de investigação em uso (Metcalf, 2015).

Especificidades metodológicas dos grandes dados: três exemplos

A partir de uma lista não exaustiva de elementos interrelacionados e por analogia, poderá identificar-se algumas das diferenças que a prática de análise de grandes dados suscita face às aproximações metodológicas tradicionais. A transposição direta de importantes práticas científicas implica, em vários casos, desfasamentos ou inadequações em contextos de grandes dados. Numa fase de grandes dados, os

métodos estatísticos clássicos necessitam de ser revistos e reimaginados em novos contextos (Schutt & O'Neil, 2014, p. 18).

1-) Amostragem: N=All

Umas das características repetidamente associada aos grandes dados assenta na ideia de uma menor importância atribuída às atividades de amostragem. A amostragem continua a ser um recurso vital sempre que existem dificuldades de recolha e processamento de informação por via de imperativos de ordem diversa, custo incluído. A amostragem é um conceito com menos de um século que foi desenvolvido para a resolução de um problema particular num dado momento do tempo sob constrangimentos tecnológicos específicos (Mayer-Schönberger & Cukier, 2013, p. 31). Num momento em que a recolha de dados é executada por inúmeros sensores em dispositivos de todo o tipo ou se baseia na partilha espontânea de dados por utilizadores, o conceito de amostragem está a ser reequacionado. Num contexto digital, o foco fundamental não passa geralmente pela obtenção de amostra aleatória representativa de determinado universo visando uma distribuição normal de casos, um recurso de investigação crucial que tem vindo a ser afinado desde o aparecimento dos primeiros censos nacionais. A tendência atual passa muitas das vezes por operar com todos os casos. Isto porque, sob um ponto de vista técnico, as operações de contagem e sobretudo de tabulação de dados não estão já sujeitas aos constrangimentos usuais, concretamente a necessidade de processamento manual da informação. Contudo, a principal razão advirá do facto de, pelo aumento do número de casos, ser incrementada a precisão dos dados. Acresce que está disponível capacidade de armazenamento, existe capacidade de processamento, e por sua vez o processamento pode assentar em operações de elevado grau de complexidade.

Nas operações com grandes dados, o enfoque está mais em atividade como a quantificação, procura de anomalias, de eliminação de ruído, de *data mining*. de combinação de bases e fontes de informação, de procura de precisão ou correção de todas as fontes de inexatidão ou de erro que se associam aos dados, de obtenção de tendências devidamente sustentadas. Todos os casos são considerados ao mesmo tempo que existe elevada granularidade nos dados, i.e. n=1, em que quantidades assinaláveis de dados de vários tipos estão disponíveis sobre determinado sujeito de pesquisa – tal ativa o problema ético associado aos grandes dados. O recurso a uma amostra aleatória num contexto de grandes dados pode representar, por si, um viés para a análise de um conjunto alargado de dados. Como foi já referido, opera-se inúmeras vezes por saturação de dados na tentativa de produção de representação cabal do fenómeno ou tendências em análise. Por fim, de referir que a amostragem de casos pode seguir outros procedimentos, especificamente a criação de subamostras de dados a partir de uma base inicial para reforço nomeadamente da análise exploratória, uma prática comum nas bases de dados relacionais.

Naturalmente as formulações anteriores têm estado sujeitas a críticas de ordem diversa. Uma interessante contribuição que ao mesmo tempo explica alguns dos constrangimentos associados à amostragem e representatividade dos dados é facultada por Emmanuel Letouzé *et al.* (2013). Para estes autores, a grande mudança é primariamente qualitativa. Efetivamente aos grandes dados associa-se a ideia de dados quantitativos, grandes em volume, de natureza dinâmica, tal como já explicitado. Contudo, é importante ter em conta que que boa parte destes dados são emitidos de forma passiva pelos indivíduos na sua interação com dispositivos digitais e serviços, naquilo a que designam por traduções digitais de ações e interações humanas (2013, p. 10). Ora, tal como sucede na num contexto de pequenos dados quando se reduz a experiência humana a um determinado indicador estatístico, também neste novo contexto se se atender à ideia de utilização de mais informação, i.e. princípio do N=All, provavelmente será ignorado o facto de se estar a utilizar sobretudo informação diferente, de tipo não-amostral, contendo informação sobre os comportamentos e crenças dos indivíduos incluídos no conjunto em análise.

À semelhança do que sucede nas abordagens quantitativas tradicionais das ciências sociais, os dados por si e sua existência não representam um avanço. Os grandes dados não são necessariamente matéria de

quantidade, até porque hoje os dados são crescentemente de acesso facilitado e de custo mais reduzido para a sua obtenção. O foco fundamental passará pelas ferramentas analíticas e dos métodos capazes de extraírem informação e subsequentemente conhecimento relevante dos mesmos, uma área em que os avanços têm sido notáveis (Gary King, 2016). Neste particular, a atividade de amostragem continuará a ser crucial, mas sob novas diretrizes. Repare-se que num contexto de pequenos dados, é usualmente possível extrair um subconjunto de observações (n) relacionável com o número total de observações de uma população (N). A esse subconjunto é dada a designação de amostra. No caso da amostragem probabilística, a ligação ou a modulação entre população traduz-se através de formulações matemáticas, só possível porque existe algum tipo de informação sobre o universo considerado. Para os grandes dados, este tipo de abordagem comporá vieses. Mais, dados, por si, sem informação contextual, induzem geralmente em erro. Nos grandes dados, o problema fundamental passa por compreender como é possível efetuar amostragem de dados e como essa amostragem pode ser utilizada para se efetuar inferências válidas e extrapoláveis a partir dessa amostra. Por outras palavras, é perceber como se pode efetuar amostragem a partir de uma rede preservando a complexa estrutura dessa mesma rede (Schutt & O'Neil, 2014, pp. 22–23).

2-) Enfoque na Correlação

Num mundo orientado por dados, poderá existir menor ênfase nas relações de causa/efeito, no teste de hipóteses, no teste de teorias. Relações de causalidade não poderão deixar de ser consideradas, contudo a correlação pode emergir como nova fonte importante de significado, admitindo-se que, sob um ponto de vista analítico, os grandes dados enfatizam o papel da correlação face à causalidade (Rieder & Simon, 2016, p. 4; Zwitter, 2014, p. 2). Tal é demonstrável pelo crescente interesse por mecanismos de predição. Chris Anderson desencadeou um importante debate por ter proclamado que a vasta quantidade de dados produzida pelos sistemas de software atuais transforma-os em instrumentos científicos, mesmo quando ausente um modelo coerente capaz de informar o desenho de análise (2008).

O pressuposto avançado assenta na alteração da forma como é compreendida a realidade social. Compreensão não se inicia, necessariamente, com uma hipótese, havendo, tal como é sugerido, espaço para modelos sofisticados para a deteção de relações não-lineares ancoradas em dados. Para Anderson, a correlação será muitas das vezes suficiente num contexto de quantidades maciças de dados. A este nível, será diferente ter em conta aplicações utilizadas em ambiente empresarial (e.g. sistemas de recomendações) de análises em contexto académico. Este enfoque na correlação poderá constar em inúmeros casos de relações positivas falsas entre dados, ou seja, padrões nas bases de dados coincidentes de forma estritamente casual, sem poder preditivo, que não são replicáveis, e que podem mascarar tendências significativas ainda que mais fracas – as bases de dados usualmente incluem dados que parecem estar associados mesmo quando a associação é aleatória, o aumento da sua dimensão incrementa o número de falsos positivos (Kitchin, 2014, p. 159).

Da linha de argumentação de Anderson, emergem pelo menos dois importantes constrangimentos que devem ser alvo de uma análise mais aprofundada. Primeiramente, porque os grandes dados, quer na sua génese, quer na sua análise, são fundados em teoria, já que recorrem à teoria da estatística, da matemática, da ciência informática. Os modelos de análise baseados nas diferentes áreas de conhecimento influenciam os métodos e os resultados, assim como os resultados dependem das decisões efetuadas por quem analisa. Como referido, e com modelos fundados em diferentes ramos de conhecimento que se conseguem análises preditivas úteis, mas os modelos implicam sempre alguma forma de reducionismo. Ainda assim, a possibilidade de a análise de grandes dados não estar sobrecarregada pelo pensamento convencional e respetivos viés implícitos às teorias de um campo específico poderá constituir um forma de oferecer uma compreensão mais aprofundadas e nova de determinados fenómenos (Mayer-Schönberger & Cukier, 2013, pp. 70–72). Seguidamente, e voltando à crítica à argumentação de Anderson, porque será erróneo pensar que os dados, por si, são neutros ou objetivos. A investigação sobre dados ricos provindos de redes sociais tem evidenciado que a riqueza dos dados obtidos contrasta com a dificuldade de generalização tendo por base tais dados. Modelos distintos

implicam resultados diferenciados. A informação produzida acarreta limitações de ordem diversa, assim como os algoritmos e processos envolvidos na captura de tais dados (Felt, 2016, p. 6).

A correlação poderá, em todo o caso, revelar-se importante para que se perceba até que ponto os grandes dados estarão a modificar a epistemologia científica. São reconhecidas novas aplicações, novas aproximações à obtenção e análise de dados, que permitem a resposta a questões de forma distinta. A prática de investigação em ciências sociais poderá ser influenciada, contudo tal não pode ser assumido, de forma direta. Uma possibilidade passa pela *data-drive science*, na qual confluem elementos de abdução, indução e dedução com o objetivo de reformular o método científico ao invés de o subverter, numa altura em que os princípios filosóficos nesta matéria estão ainda no seu começo, sendo requeridos reflexão e elaboração de princípios epistemológicos e metodologia adequados (Kitchin, 2014, pp. 147–148).

3-) Privacidade e Ética

A ética é reconhecida como um ramo da filosofia que versa a conduta moral. Sob a influência de Sócrates, a ética passou a constar da procura por um importante bem acima de todos os outros. A nova aproximação substituiu a ideia de viver bem num mundo perigoso, tal como descrito por Homero. Na formulação anterior, um contexto de destino e sorte era utilizado para descrever a vida humana e suas contingências. A vida humana inclui diversos bens, e Sócrates introduz a ideia de moralidade, algo acima de qualquer contingência, para ser mantido por todos os meios. A filosofia grega clássica foi importante na introdução de um conjunto de leis e regras que todos, sem exceção, devem obedecer. Este conjunto particular de valores precedem qualquer outro, uma aproximação mais sofisticada do que a avaliação do que é certo ou errado a cada momento (MacIntyre, 2009).

A ética implica um trabalho constante por forma a responder aos novos arranjos sociais e aos novos dilemas morais. Os grandes dados têm acionado preocupações éticas crescentes no seio de um contexto social em rápida mutação, assente em "dados transacionais" (Beer & Burrows, 2013). Um dos principais efeitos é dado por uma alteração de vontade individual e processo de decisão para decisões por vários implicando consequências não antecipadas para qualquer um (Zwitter, 2014). Tal é confirmado pela falácia da autogestão da privacidade dos dados, ou seja, o "equívoco de que os cidadãos digitais podem autogovernar a informação que disponibilizam num universo digital" (Obar, 2015, p. 2). A paisagem sociotécnica inclui agora relações pouco precisas entre conceções fundamentais, como público e privado. Boas práticas eticamente responsáveis ainda carecem de elaboração, não existindo qualquer garantia de efetivação. Por outras palavras, ao lidar com os grandes dados, é dado como garantia que as tecnologias são meras ferramentas da intenção humana, e que a moralidade (boa ou má), herdada pelos indivíduos e não nas suas ferramentas, terá ainda de ser alcançada (Flanagan, Howe, & Nissenbaum, 2008, p. 347).

Também as questões de ética e privacidade decorrem de processos históricos alargados. As discussões sobre privacidade começaram a ser colocadas na agenda devido a duas tendências interligadas: a revolução pós-industrial da informação e a crescente utilização de dados por governos nacionais no final dos anos 60 (OECD, 2013). São várias as razões que contribuem atualmente para as crescentes preocupações com o respeito por padrões seguros de privacidade e de ética. Estes são conceitos que devem inserir-se numa discussão mais ampla relativa aos próprios direitos humanos. Será necessário ter em conta que, face à originalidade do contexto em causa, é difícil a definição de um padrão, uma referência para gestão de todo um complexo que é animado por uma rápida progressão nos desafios que produz. A questão fundamental passará por conceber limites à utilização eticamente responsável de grandes dados. A centralidade da privacidade é importante para o seu próprio entendimento, mas ainda pelo seu valor instrumental ao mesmo tempo que sublinha as formas diversificadas como outros valores éticos são negativamente afetados e sustentados, especificamente pelas formas como a informação flui ou não. A privacidade é aqui importante porque implica outros valores, depois porque dessa forma possibilita uma melhor formulação de intervenções, regulamentações, ou correções (Barocas & Nissenbaum, 2014, p. 49).

Sob a lente dos direitos humanos, os grandes dados devem, pois, estar sujeitos a um escrutínio aprofundado. Significa isso a participação cidadã nos debates sobre a própria regulação dos grandes dados enquanto ecossistema. Mais, a cidadania associa-se não apenas à reivindicação de direitos como privacidade, mas outras dimensões como a "literacia de dados", i.e. a capacidade de interpretação de forma crítica de dados sobre os mais diversos fenómenos. Isso significa ter em conta elementos como qualidade dos dados, representatividade dos dados, a distinção entre correlação e causalidade, o risco de identificação de correlações espúrias, ou seja, um repositório associado ao conhecimento estatístico básico que se torna ainda mais relevante presentemente. Ora, por aqui se poderá inferir uma limitação importante da governamentalidade atual. Tal como explicitado por Obar (2015), enquanto vão surgindo novas recomendações a encorajar a autogestão da privacidade de dados por parte dos cidadãos, especificamente nos EUA, não só são remetidas para a cidadania, individualmente, questões difíceis de resolver, esperando-se que cada indivíduo atue como um *data miner* e tenha tempo ilimitado à sua disposição, como deixa de ser tido em devida conta o problema macro de grande opacidade de toda a infraestrutura e atores associados aos grandes dados e suas transações.

Haverá, para mais, questões prévias de difícil resolução. Existem limitações quanto à aplicação de metodologias de investigação coerentes em grandes dados, que coexistem com a opacidade deste ecossistema assente numa economia de dados. No entanto, face à atração gerada pelos grandes dados, face ao grande investimento que é feito nas novas tecnologias associadas aos grandes dados, se tidas em contas as aplicações dos grandes dados ou os benefícios que realmente são esperados dos grandes dados, então a privacidade será um constrangimento insustentável – privacidade e grandes dados são incompatíveis (Barocas & Nissenbaum, 2014, p. 63).

Notas finais

O presente artigo consta de uma aproximação a algumas das questões associadas à emergência dos grandes dados através de elementos diretamente ligados às metodologias de investigação, especificamente as questões de amostragem, de relações de causalidade, ou de privacidade e ética. Às ciências sociais caberá, seguramente, o importante papel de melhor destrinçar este contexto de abundância de dados, de enfoque na quantificação, de recurso a informação sensível gerada por utilizadores individualmente. Efetivamente, subsistem problemas de natureza diversa. A resposta a estas questões passará pela inclusão de alguns elementos relacionados: a discussão do papel dos dados e da informação nas sociedades atuais; uma análise crítica do verdadeiro significado dos dados em diversas temáticas, desde logo para dissipar assunções simplistas sobre assuntos relevantes; a discussão sobre a distribuição desigual de dados como fonte *per si* de desigualdades de natureza diversa, coexistindo "buracos" de diferente ordem, seja pela simples ausência de dados para retrato de determinado fenómeno, seja pela má qualidade dos mesmos, que reproduzem lógicas de análise enviesadas (e.g. o recurso exagerado a mecanismos de estimação estatística); a suposta representatividade e/ou totalidade, em contextos para os quais usualmente não existem bases amostrais e como podem ser realizadas inferências em tais contextos (e.g. utilizadores internet); o facto de a própria internet não constituir, sempre, de forma linear, uma representação da realidade (e.g. papel dos algoritmos na disponibilização/invisibilização de resultados); a rede como uma forma reducionista de realidade, dada a não inclusão e não análise de casos relativos a todos os indivíduos que não participam nas plataformas eletrónicas existentes; a modelação em grandes dados e o seu papel na redução da realidade e da diversidade.

Referências

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired Magazine*, Science, Retrieved from https://www.wired.com/2008/06/pb-theory/

Barocas, S., & Nissenbaum, H. (2014). Big Data's end run around anonymity and consent. In Lane, Julia et

al. (edits.), *Privacy, Big Data and the Public Good* (pp. 44–75). New York: Cambridge University Press. http://doi.org/10.1017/CBO9781107590205

Beer, D., & Burrows, R. (2013). Popular Culture, Digital Archives and the New Social Life of Data. *Theory, Culture & Society*, *30*(4), 47–71. http://doi.org/10.1177/0263276413476542

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. http://doi.org/10.1080/1369118X.2012.678878

Dalton, C., & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to "big data." *Society & Space*. Retrieved from http://societyandspace.com/material/commentaries/craig-dalton-and-jim-thatcher-what-does-a-critical-data-studies-look-like-and-why-do-we-care-seven-points-for-a-critical-approach-to-big-data/

Dean, M. (2009). *Governmentality: Power and Rule in Modern Society*. London: SAGE.

Didier, E., & Tasset, C. (2013). Pour un statactivisme. La quantification comme instrument d'ouverture du possible. *Revue de Sciences Humaines*, (24).

Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*, *3*(1), 1–15. http://doi.org/10.1177/2053951716645828

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying Values in Technology Theory and Practice. *Information Technology and Moral Philosophy*, 415. http://doi.org/10.1017/CBO9780511498725.017

Foucault, M. (1982). *The Archaeology of Knowledge*. New York: Vintage.

Gregory, J., & C. Bowker, G. (2016). "The Data Citizen, the Quantified Self, and Personal Genomics". In D. Nafus (Ed.), *Quantified: Biosensing Technologies in Everyday Life* (pp. 211–226). Cambridge, MA: The MIT press.

Hacking, I. (2000). *The Social Construction of What?* Harvard: Harvard University Press.

Hacking, I. (2001). *Probability and Inductive Logic*. Cambridge: Cambridge University Press.

Huberman, B. A. (2012). Sociology of science: big data deserve a bigger audience. *Nature*, *482*(7385), 308. http://doi.org/10.1038/482308d

Isin, E., & Ruppert, E. (2015). *Being Digital Citizens*. London: Rowman & Littlefield.

King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science*, *331*(6018), 719–721. http://doi.org/10.1126/science.1197872

King, G. (2016). Preface: Big Data is Not About the Data! In R. M. Alvarez (Ed.), *Computational Social Science: Discovery and Prediction*. Cambridge: Cambridge University Press.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. London: SAGE.

Letouzé, E., Meier, P., & Vinck, P. (2013). "New Technology and the Prevention of Violence and Conflict". In F. Mancini (Ed.), *New Technology and the Prevention of Violence and Conflict* (pp. 4–27). Report, New York: International Peace Institute.

MacIntyre, A. (2009). *Dependent rational animals: why human beings need the virtues*. London: Duckworth.

Manovich, L. (1998). *Database as a Symbolic Form*. Cambridge, MA: MIT Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: a Revolution that Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

Metcalf, J. (2015). Human-Subjects Protections and Big Data : Open Questions and Changing Landscapes, (2012), *Council for Big Data, Ethics, and Society*. 1–13. Retrieved from http://bdes.datasociety.net/council-output/human-subjects-protections-and-big-data-open-questions-and-changing-landscapes/

Obar, J. A. (2015). Big Data and The Phantom Public: Walter Lippmann and the fallacy of data privacy self-management. *Big Data & Society*, *2*(2), 1–16. http://doi.org/10.1177/2053951715608876

OECD. (2013). *The OECD Privacy Framework*. Paris. Retrieved from www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf

Pentland, A. (2015). *Social Physics: How Social Networks Can Make Us Smarter*. Westminster: Penguin Books.

Podesta, J., Pritzker, P., Moniz, E. J., Holdren, J., & Zients, J. (2014). *Big data: seizing opportunities, preserving values*. Washington.

Porter, T. M. (1996). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Priceton: Princeton University Press.

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, *3*(1), 1–6. http://doi.org/10.1177/2053951716649398

Salah, A. A., Manovich, L., Salah, A. A., & Chow, J. (2013). Combining Cultural Analytics and Networks Analysis: Studying a Social Network Site with User-Generated Content. *Journal of Broadcasting & Electronic Media*, *57*(3), 409–426. http://doi.org/10.1080/08838151.2013.816710

Schutt, R., & O'Neil, C. (2014). *Doing Data Science: Straight Talk from the Frontline*. Sebastopol: O'Reilly Media.

Zwitter, A. (2014). Big Data ethics. *Big Data & Society*, *1*(2). http://doi.org/10.1177/2053951714559253

---

[1] **Projeto de Investigação de Doutoramento** *Is Big Data Shaping Democracy? A Critical Account*. Financiamento: Fundação para a Ciência e Tecnologia (FCT). Referência: SFRH/BD/52258/2013. Supervisão: Prof. Dr. José Manuel Mendes. Instituição de acolhimento: Centro de Estudos Sociais (www.ces.uc.pt).

[2] Conferir, por exemplo, o caso noticiado a partir de 2014 relativo à revenda de dados, especificamente de registos médicos detalhados de pacientes do Sistema Nacional de Saúde britânico (NHC) a grandes corporações.