em questão
E-ISSN 1808-5245

Big Data: fatores potencialmente discriminatórios em análise de dados

Caio Saraiva Coneglian
Mestrando; Universidade Estadual Paulista "Júlio de Mesquita Filho", São Paulo, SP, Brasil;
caio.coneglian@gmail.com

José Eduardo Santarem Segundo
Doutor; Universidade de São Paulo, São Paulo, SP, Brasil;
santarem@usp.br

Ricardo Cesar Gonçalves Sant'ana
Doutor; Universidade Estadual Paulista "Júlio de Mesquita Filho", São Paulo, SP, Brasil;
ricardosantana@marilia.unesp.br

Resumo: As mudanças tecnológicas vividas a partir da virada do século causaram uma revolução na sociedade, chamada de Big Data, em que as análises de dados para determinar padrões e comportamentos puderam utilizar grandes quantidades de dados. Verifica-se que algumas análises, no contexto do Big Data, estão sendo conduzidas a gerar resultados discriminatórios. O estudo tem como objetivo identificar fatores que, potencialmente, possam gerar discriminação durante o processo de análise de dados. Para tal, a metodologia utilizada foi de natureza qualitativa, exploratória e bibliográfica, enumerando em um quadro os casos de discriminação. Como resultado, identificam-se fatores possivelmente discriminatórios, além de ser feita uma explanação desses fatores. Por meio da pesquisa, verifica-se uma necessidade de existir reflexões profundas dos resultados que são obtidos a partir de análises de dados, ficando clara a necessidade da Ciência da Informação retratar tais questões, a fim de apontar os caminhos a serem tomados.

Palavras-chave*:* Big data. Análises de dados. Discriminação. Fatores potencialmente discriminatórios.

1 Introdução

Um fenômeno chamado Big Data está causando uma revolução em como empresas, governos e organizações coletam e analisam os dados para a tomada de decisão, tanto no âmbito governamental quanto no empresarial. Verificam-se, também, transformações profundas, no campo científico, que estão ocorrendo

pela utilização de altos volumes de dados, permitindo a compreensão de diversos eventos que, até então, não haviam sido identificados.

O conceito de Big Data relaciona-se à heterogeneidade, à rápida geração e processamento e à grande quantidade de dados disponíveis digitalmente. Esse fenômeno despontou mais, significativamente, a partir do início do século XXI, quando a quantidade de dispositivos e de usuários conectados apresentou um crescimento exponencial. Outro fator marcante, deste século, que contribui para a explosão do Big Data, foi o uso das tecnologias de informação e de comunicação para a maioria das tarefas cotidianas, em que, por exemplo, notas fiscais são geradas e enviadas em formatos digitais; pedidos de comidas são feitos pela Web; o acompanhamento de um caso judicial ocorre por meios de sistemas eletrônicos; prontuários médicos são em sua maioria eletrônicos; entre milhares de outros exemplos que estão presentes durante as mais diversas tarefas realizadas por um indivíduo (MAYER-SCHÖNBERGER; CUKIER, 2013)

Uma das consequências desse fenômeno é que todas as ações realizadas por uma pessoa deixam rastros a respeito de quem ela é e do que ela faz, permitindo com que as características de um determinado indivíduo estejam armazenadas em alguma base de dados espalhada pelo planeta, e que grandes organizações e governos possam acessar tais dados para os mais diversos fins. Nesse contexto, o principal uso de todos esses dados são análises na busca a fim de definir padrões e comportamentos de usuários, pacientes, empregados, doenças, produtos, entre outros. Assim, as informações que são encontradas são utilizadas para a tomada de decisão dos gestores.

O uso desses dados nas análises, em esferas acadêmicas, governamentais e organizacionais, tem proporcionado grandes mudanças, pois se verifica padrões inéditos a respeito dos comportamentos e das informações. Tais mudanças só são visíveis por meio de análises automatizadas, utilizando diversas bases de dados, que permitem comparações de inúmeros elementos e fornecendo valores que são ocultos à primeira vista para um analista humano (TAURION, 2013).

Contudo, diversas questões necessitam ser discutidas, quando se trata de análises de dados dentro do contexto do Big Data; para assim, esclarecer as

consequências na sociedade, quando tais processos são realizados sem haver uma reflexão a respeito dos resultados que são obtidos. Questionam-se, principalmente, as organizações, que com o intuito de aumentar seus lucros, buscam quaisquer meios para melhorar sua posição no mercado.

Dessa forma, surgem casos de algoritmos construídos para realizar análises de dados que fornecem conclusões de caráter discriminatório. Alguns dados apontam, por exemplo, que gêneros distintos possuem rendimentos diferentes em suas funções empregatícias, e que, assim, deve ser priorizada a contratação de um determinado gênero, ou ainda, que a oferta de crédito deve ser prioritária para pessoas que nunca tiveram processos judicias, frente àqueles que tiveram algum processo. Tais exemplos demonstram que uma análise de dados pode levar à ações discriminatórias das organizações, quando não ocorre uma reflexão profunda dos resultados que são apresentados (BAROCAS; SELBST, 2016).

Como consequência, a análise de dados, no contexto do Big Data, torna-se uma ferramenta muito poderosa, porém cercada de riscos, caso não considere determinadas parcelas da população, ou caso utilize critérios de análise que possuam uma carga de preconceito oriunda de análises precedentes, realizadas pelos gestores e pelos tomadores de decisão. Diversos são os fatores que podem causar erros intencionais ou não durante o processo de análise dos dados; sendo necessário que tais fatores sejam pontuados e discutidos, para que os analistas de dados, com tais informações, encontrem meios de evitar que a análise de dados, no contexto do Big Data, não seja mais uma ferramenta ocasionadora de discriminação e preconceito (BAROCAS; SELBST, 2016). Como meio de auxiliar nesse processo, o presente estudo discute acerca de fatores potencialmente discriminatórios, compreendendo tais fatores por elementos constituintes das análises de dados que, caso não sejam observados e evitados, tornam os resultados provenientes das análises, em discriminação a determinados grupos de indivíduos.

Assim, essa pesquisa tem como objetivo identificar fatores, que dentro do processo de análise de dados no cenário de Big Data, podem causar discriminação ao fornecer informações e conclusões a respeito da massa de

dados analisada que afete a determinados grupos de indivíduos. Busca-se também, identificar em qual etapa do processo de análise de dados tais fatores ocorrem.

Nesse sentido, o texto é dividido em sete seções, sendo a primeira uma introdução, contendo uma explanação inicial dos conceitos apresentados no texto, além da inserção da problemática, dos objetivos e da metodologia utilizada. A segunda seção apresenta os procedimentos metodológicos da pesquisa. A terceira e a quarta seção abordam sobre Big Data e discriminação, respectivamente. A quinta seção, por sua vez, apresenta casos da literatura que levaram a análises discriminatórias. Os resultados da pesquisa são apresentados na sexta seção. E, por fim, a última seção apresenta as considerações finais, seguidas das referências bibliográficas.

2 Procedimentos metodológicos

A metodologia para a realização desse trabalho caracteriza-se como descritiva, em que a partir de uma análise bibliográfica e documental, identificaram-se casos em que ocorreu discriminação devido ao uso de técnicas de Big Data. A partir destes casos foram enumerados alguns fatores potencialmente discriminatórios no contexto do Big Data.

Estes casos de discriminação foram obtidos baseados em fontes bibliográficas e documentais, em que foi delimitado um número de dez casos, para que pudesse ser realizada uma análise detalhada de cada caso, permitindo a identificação dos fatores que levaram a discriminação. Os dez casos obtidos estão disponíveis em fontes internacionais, sendo que todos foram discutidos e abordados por veículos de mídia e artigos de grande visibilidade.

Vale ressaltar que para a escolha dos casos, buscou-se em artigos científicos e em artigos publicados pela mídia internacional, em que foram expostos problemas referentes à discriminação devido ao uso de técnicas de Big Data. Para identificar esses casos, realizaram-se buscas por "Big Data discrimination" em buscadores da Web e em bases bibliográficas internacionais, sendo localizados aqueles que tratavam da temática desta pesquisa, além disso, a

partir da leitura dos textos encontrados, localizaram-se novas fontes com base nas referências dos textos, em que também ocorriam casos de discriminação.

Justifica-se o uso de veículos de mídias e de uma pesquisa exploratória que busca nas referências dos artigos encontrados outras fontes, pois o presente trabalho visou encontrar os fatores discriminatórios, não tendo como propósito apontar todos os casos que ocorreram discriminação dentro da literatura. Diante do exposto, o uso de procedimentos sistemáticos para busca das fontes poderia não se mostrar eficiente, uma vez que não possibilitaria uma abertura maior aos autores em localizar fontes que refletissem diversas problemáticas, podendo diminuir os fatores discriminatórios identificados.

Desta forma, os dez casos localizados apresentam uma diversidade dos problemas expostos, o que possibilitou uma pesquisa ampla, apontando um espectro de fatores discriminatórios conciso com uma literatura clara que embasa os resultados obtidos.

A realização da pesquisa foi dividida em três fases: na primeira, foi feita uma análise bibliográfica de Big Data e de discriminação; no segundo momento construiu-se um quadro com diversos casos de discriminação em análises de dados encontrados a partir da pesquisa exploratória e documental exposta anteriormente e; a partir do quadro construído identificaram-se os fatores potencialmente discriminatórios.

3 Big Data

O início do século XXI é marcado pelo alto volume de dados gerado na rede de computadores. Todas as informações referentes às esferas organizacionais, institucionais e governamentais encontram-se disponíveis em formatos digitais. Nesse sentido, análises são executadas com o objetivo de verificar padrões, e informações podem ser extraídas com o intuito de melhorar os processos de tomada de decisão.

Além das informações geradas por meios tradicionais, como bibliotecas, formulários, arquivos e banco de dados de empresas, existe uma geração de dados que extrapola tais meios, como as informações produzidos em redes sociais, as transações online, o GPS, os dados produzidos por dispositivos

conhecidos como Internet das Coisas (Internet of Things - IoT), os smartphones, entre outros.

Esse contexto atual foi classificado como Big Data. Sintaticamente, Big Data pode ser aplicado às informações que não podem ser processadas ou analisadas com ferramentas e métodos tradicionais, sendo que esse fenômeno possibilitou que pessoas e programas comunicassem-se não apenas durante um intervalo de tempo, mas também em tempo integral (ZIKOPOULOS et al., 2011)

O termo Big Data não trata apenas de questões relacionadas ao armazenamento dos dados, mas também perpassa por questões como a velocidade que os dados devem ser tratados e processados, além da variedade das fontes e dos formatos com que os dados são gerados (MCAFFE; BRYNJOLFSSON, 2012). Nessa perspectiva, McAffe e Brynjolfsson (2012) elenca três características fundamentais relacionadas ao Big Data: volume, variedade e velocidade. Tais características serão melhores explorados na sequência:

a) volume - a Web interativa, a conexão de um número cada vez maior de dispositivos na rede e o uso mais intenso de redes sociais têm provocado um aumento exponencial na quantidade de dados que são gerados diariamente. Todas essas informações, além de outros tipos de dados como o comportamento de consumidores, os dados financeiros, os relatórios médicos, as conversas realizadas em aplicativos de troca de mensagens, são armazenados em bases de dados, que gera um volume extremamente denso de dados;
b) velocidade - em suma, os dados gerados ficam disponíveis em servidores em tempo real. Tal característica permite que o processamento e as análises dos dados ocorram, simultaneamente, a criação dos mesmos, possibilitando a tomadas de decisões instantâneas. Dessa maneira, a velocidade à qual se refere McAffe trata não apenas da entrada, mas também do fluxo dos dados, em que é necessário ter velocidade para acompanhar a geração e a demanda das requisições realizadas;

c) variedade - o modo em que os dados estão disponíveis na Web apresenta uma diversidade crescente, onde usuários podem inserir textos, músicas, hipertextos, vídeos, conteúdos interativos, entre outros. Kakhani, Kakhani e Biradar (2013) complementam tal questão, afirmando que os dados podem ser estruturados, semiestruturados ou não estruturados, sendo de natureza heterogênea, vindo das mais diversas fontes, como mídias sociais, blogs, páginas empresariais e governamentais.

Dentro do contexto das características apresentadas, inicia-se uma reflexão em torno das questões que perpassam pelas análises dos dados e encontram-se no cenário de Big Data.

Mayer-Schönberger e Cukier (2013) trazem um primeiro ponto de vista, relatando que a mudança quantitativa das informações - grande crescimento na quantidade dos dados - trouxe uma melhora qualitativa das informações ao fornecer dados cada vez mais precisos. Os autores complementam que tais mudanças possibilitaram que as análises realizadas sobre as informações disponíveis permitem que seja considerado toda a base de dados existente, ocorrendo uma mudança no panorama de como sãs feitas as análises dos dados. O que justifica essa mudança são as capacidades de armazenamento e processamento das tecnologias que apresentaram uma grande evolução e oportunizam que as análises não precisassem considerar somente uma amostra dos dados, procedimento que era o mais utilizado até então.

Considerando esse aspecto, compreende a necessidade de identificar as fases das análises de dados no contexto do Big Data, para que essa melhora qualitativa possa ser compreendida, e se torne algo aplicável na prática. Cezar Taurion (2013) arrisca nessa tarefa, ao afirmar que existem quatro atividades dentro do processo de análise de dados em cenário de Big Data: a coleta dos dados, em que são coletadas informações de distintos cenários, como mídias sociais, sistemas transacionais, câmeras de vigilância, entre outros; a segunda etapa trata de um processo de limpeza, formatação e validação dos dados obtidos, para que sejam armazenados e estruturados dados que não sejam incompletos ou inconsistentes; posteriormente, vem a etapa de integração e

agregação dos dados obtidos nas diversas fontes, pois são coletados e cruzados dados de diversos contextos, sendo necessário existir uma integração entre os mesmos; por fim, ocorre a fase analítica, quando cumpre-se a análise e interpretação dos resultados obtidos.

Estando claros os conceitos e os processos envolvidos dentro desse fenômeno chamado de Big Data, surgem questionamentos em torno de possível discriminação em análises de dados. Para compreender melhor a consequência desses fatos, torna-se necessário descrever o conceito de discriminação utilizado nesse trabalho. Dito isso, na próxima seção será discutido sobre esse conceito.

4 Discriminação

O conceito de discriminação relaciona-se às condições históricas vividas por determinados grupos sociais, existindo diversos debates e ações, a fim de diminuir as diferenças existentes entre indivíduos e, consequentemente, diminuir a discriminação existente.

O dicionário Priberam apresenta três definições para o termo discriminação: (i) ato ou efeito de discriminar = DISTINÇÃO; (ii) ato de colocar algo ou alguém de parte e; (iii) tratamento desigual ou injusto dado a uma pessoa ou grupo, com base em preconceitos de alguma ordem, notadamente sexual, religioso, étnico, etc. (DISCRIMINAÇÃO, 2011). Por meio dessas definições, é possível contextualizar o sentido principal que esse termo possui: descrever uma ação que determinados indivíduos sofrem, por qualquer motivo que seja, em que, há distinção entre pessoas, ou elas são colocadas a parte de uma situação.

Uma conceitualização mais aprofundada da questão da discriminação pode ser vista na obra de Pierre Bourdieu. Em seu livro Poder Simbólico (BOURDIEU, 1989), Bourdieu discorre sobre as relações de estruturas de poder existentes na sociedade. Essa obra trata dos sistemas simbólicos, trazendo que tais sistemas possuem a função de integrar socialmente para um determinado consenso de hegemonia ou de dominação.

Nessa perspectiva, Bourdieu afirma que o poder simbólico é uma configuração transformada e legitimada de outras formas de poder, que busca

construir um conformismo lógico. A partir desses aspectos, surge a questão da distinção, pois os símbolos são utilizados para criar uma cultura dominante, que integra os membros da classe dominante, ao mesmo tempo que, as distingue das outras classes, em que o poder simbólico provê meios para a continuidade de tal classe, buscando a legitimidade da sua dominação por meio da reprodução sistemática dos símbolos.

Dessa forma, a distinção é uma consequência da cultura dominante. Em síntese, Bourdieu traz que a distinção é um mecanismo de reprodução da dominação, pois é tal característica que difere uma classe da outra, e permite a dominação entre classes.

Na primeira definição dada pelo dicionário Priberam, verifica-se a utilização do termo distinção, que se relaciona diretamente a obra de Pierre Bourdieu. No contexto desse trabalho, será utilizado o conceito de distinção de Bordieu como sinônimo do termo discriminação.

5 Dados coletados

Diversas questões são capazes de gerar problemas durante o processo de análise de dados. Esses problemas podem causar exclusão e discriminação por não considerarem fatores que influenciam diretamente nos resultados, que como consequência poderá, inevitavelmente, afetar determinados grupos de indivíduos, em virtude das tomadas de decisões baseadas nas análises dos dados. Essa pesquisa não trata, especificamente, da intenção de discriminar dos casos relatados, visto que tal tarefa deve ser tratada em âmbito judicial, onde será analisada cada situação, e assim, será possível haver um julgamento mais específico da culpa das organizações.

Com o intuito de identificar fatores discriminatórios e suas características, elaborou-se o quadro 1, em que é relacionado casos citados na literatura, onde ocorreu discriminação, com uma descrição que demonstra o que levou a análise a apresentar um comportamento discriminatório. A primeira coluna do quadro 1 apresenta um número para identificar o caso apresentado; a segunda coluna apresenta a citação de onde pode ser encontrado o caso descrito;

a coluna do resumo insere uma condensação das características do caso; e a quarta coluna apresenta uma demonstração da ocorrência de discriminação naquele caso.

Quadro 1 - Casos de discriminação em análises de dados

#	Biblio-grafia	Resumo	Demonstração da ocorrência da discriminação
1	Frank et al. (2013)	Estudo analisando o comportamento de pessoas na rede social Twitter conclui que os estadunidenses são mais felizes quando estão longe de suas casas.	A fatia da população estadunidense que utiliza o Twitter corresponde a 23% da população adulta (PEW RESEARCH CENTER, 2014). Além de ser estimada a existência de 20 milhões de contas falsas no Twitter (PERLROTH, 2013). Dessa forma, a análise de dados tira conclusões a respeito dos dados, utilizando uma amostra que não é igualitária. Assim, a amostra de dados apresenta problemas ao não considerar pessoas que não são adeptas a mídia social utilizada.
2	Golds-tein e Winke-lmayer (2015)	Relata que estudos de Big Data devem ser realizados utilizando os dados de sistemas de saúde. Apontando como exemplo, caso de Taiwan, onde há um sistema universal de saúde, que abrange a 99% da população. E assim, ser possível definir políticas públicas e comportamentos de doenças.	Os pesquisadores não consideraram determinados aspectos, que poderiam enviesar a pesquisa. Dentre esses aspectos, pode-se destacar questões como: caso o paciente seja hospitalizado em um hospital particular, ou ainda, se os dados dos formulários forem descritos incorretamente, estes dados serão considerados? Tais questões se não forem adequadamente examinadas, poderão traçar panoramas que não refletem a situação da população, podendo levar a criação de políticas públicas inapropriadas. Isto porque, o conjunto de dados pode estar incompleto, faltando informações essenciais para a tomada de decisão, que terá como possível consequência, a indução a políticas discriminatórias, pois poderá não considerar a situação de algumas pessoas.
3	Street Bump (2015)	Esse projeto trata de um aplicativo que coleta informações de buracos de ruas da cidade de Boston, EUA por meio do uso de acelerômetro e GPS, tecnologias presentes apenas em smartphones. Além disso, o aplicativo contém apenas suporte para dispositivos iPhone, da fabricante Apple. Os dados que são coletados são enviados ao governo municipal, para que possa ser realizado, futuramente, o conserto das ruas que apresentam	Crawford (2013) contrapõem os benefícios desse aplicativo, ao questionar a real abrangência de usuários que utilizam o aplicativo, pois se o município confiar, ou der prioridade aos dados fornecidos pelos cidadãos que têm um smartphone, mais especificamente um iPhone, ocorrerá uma parcela da população que será subamostrada. Sendo que tal parcela será, em sua maioria, moradores de bairros de menor renda, que não possuem proporcionalmente a mesma quantidade de iPhones que parcelas de maior renda. Como consequência, os bairros das populações mais pobres estarão representados nos mapas gerados pelo aplicativo, como bairros que apresentam melhor qualidade nas ruas, que provavelmente não será uma verdade, e não haverá um esforço em consertar tais ruas. Enquanto, que nas ruas dos bairros de maior renda, haverá uma manutenção mais frequente, pois os mapas representarão mais

		buracos.	fielmente a situação daquelas ruas. Dessa forma, a amostra que foi considerada é pouco significativa, e exclui determinados grupos populacionais que não tem um determinado tipo de dispositivo.
4	Chow-White; Green Jr. (2013)	A utilização de genomas em pesquisas de Big Data tem se tornado muito comum, entretanto, vem ocorrendo pesquisas que estão comparando os genomas de pessoas de raças diferentes.	O estudo mostra que pesquisas com esse cunho preconceituoso vêm ocorrendo com maior frequência, onde, após décadas que a ciência provou que não existiam diferenças entre pessoas de raças distintas, pesquisas estão sendo feitas mostrando que há diferenças. Contudo, o olhar que os analistas têm dos dados surge como grande culpado, pois ocorre uma interpretação incorreta da realidade, ao considerar apenas poucos fatores, sem que existisse uma reflexão maior de outras variáveis que poderiam ser consideradas. O olhar do analista induziu a buscar provas que a raça influenciaria na questão dos genomas.
5	Butler (2013)	Um sistema chamado Google Flu, desenvolvido pela empresa Google, busca identificar os casos de gripe que estão ocorrendo em cada estado e região dos Estados Unidos, em que, por meio do comportamento que um usuário possui, é possível traçar em tempo real o quanto uma epidemia de gripe está avançando.	O autor descreve que quando ocorrem epidemias e temporadas de gripe que apresentam alguma diferença de um padrão geral, o sistema Google Flu costuma superestimar ou subestimar a quantidade de casos de gripes, quando verificados pelos sistemas de monitoramento estadunidense. Pesquisadores e o Google acreditam que o sistema Google Flu pode ser utilizado para os governos averiguarem a evolução da epidemia. No entanto, se o sistema estiver sub ou superestimando, as políticas públicas tomadas podem não levar em consideração análises corretas. Os analistas dos dados deveriam ter considerado questões como o ciclo das doenças, as diferenças que ocorrem entre as temporadas das gripes e outros fatores determinantes. E não somente considerar dados anteriores, que podem conter erros, ou uma carga discriminatória ao não considerar outros conjuntos de indivíduos ou de dados.
6	Croll (2012)	A empresa financeira American Express começou a utilizar os históricos de compras de seus clientes, em análises de crédito, baixando o limite de crédito de pessoas que compraram em estabelecimentos, que apresentaram casos de clientes que fizeram baixo reembolso de suas compras.	A realização de inferências sobre a possibilidade de um cliente ser inadimplente, pelos estabelecimentos onde se costuma comprar, pode gerar discriminação de diversos contextos, como caso um bairro apresente populações mais carentes, que pode causar uma maior inadimplência. Mas é um problema, quando se conclui que todas as pessoas que vão ou moram nesse local, devem ter menos crédito ou ter crédito negado, simplesmente pelos locais onde ele frequenta ou habita. Assim, a classificação dos grupos sociais aparece como grande responsável pela discriminação, pois são criados grupos onde existem maior chances de ser feito o pagamento, pelos locais onde as pessoas convivem, existindo assim, locais de pessoas melhores pagadoras e locais de maus pagadores.
7	Lohr (2013)	O processo de contratação de funcionários vem tendo uma série de análises	Conclusões como funcionários que apresentam um maior senso de trabalho serão mais inovadores, são concluídas a partir de um conjunto de dados que são analisados, identificando históricos de diversos

		estatísticas. O autor dá um exemplo da aplicação de um teste de honestidade e aplicação de questionários para verificar se os funcionários apresentam um forte senso de trabalho, pois a partir da análise de massas de dados, é possível verificar se um funcionário permanecerá por mais tempo na empresa e se um funcionário será mais ou menos inovador.	anos ou décadas. Contudo, tais análises apresenta uma grande subjetividade, onde, por exemplo, dependendo do contexto de vida ou da religião de uma pessoa, ela possuirá mais ou menos senso de trabalho, mas isso pode não ser uma verdade quanto ao nível de inovação que uma pessoa possui. Dessa forma, pode ocorrer uma discriminação de determinadas pessoas, devido a aplicação de questionários ou testes subjetivos, por não apresentarem um nível nas varáveis desejadas, que também apresentam características subjetivas. A classificação de bons funcionários e maus funcionários apresenta grandes polêmicas, pela subjetividade existente, e a utilização desses dados, para realizar a tomada de decisão, promove discriminação, pois esses critérios serão utilizados futuramente, para a classificação das pessoas no momento de seleção de novos funcionários. Além disso, dados e tomadas de decisões anteriores influenciam tais mecanismos, embora as informações possam conter uma carga implícita de discriminação
8	Swee-ney (2013)	A autora relata uma pesquisa que analisa os resultados dos anúncios apresentados em buscas feitos no motor de busca Google, onde utilizando listas de registros de nascimentos dos estadunidenses, foram verificados os nomes que eram majoritariamente de pessoas negras e de pessoas brancas, após, foram feitas buscas com esses nomes, para verificar quais tipos de anúncios eram retornados, sendo que foi comprovado que nomes de incidência majoritária de negros, eram vinculados a anúncios de fichas criminais, e no caso dos nomes de maioria branca, a ocorrência de tais tipos de anúncios era significativamente menor.	No artigo, a autora questiona quais foram os métodos utilizados pelo buscador Google para ocorrer essa associação entre nomes predominantes de pessoas negras com anúncios de fichas criminais. Entretanto tais questões, em suma, partem de análises estatísticas feitas em cima de grandes conjuntos de dados. Tais análises apresentam um problema sério de discriminação, pois a associação entre a cor da pele e a criminalidade se configura como um crime de racismo. A análise dos dados, mesmo que supostamente, não seja intencionalmente discriminatória, segue parâmetros classificatórios, que analisam variáveis como a cor da pele para determinar o melhor tipo de anúncio para aquele tipo de registro, tornando os resultados preconceituosos. Também verifica-se que informações anteriores, como uma relação da cor da pele com a criminalidade, e nomes mais comuns em pessoas de uma determinada cor de pele, influenciam nas análises de dados, dessa forma, conjuntos de dados anteriores com uma carga de discriminação, acabam influenciando nos resultados apresentados aos usuários do motor de busca.
9	VALZ - US 200801 54798 A1 Google	Uma patente de nome Modelos de Definição Dinâmica de Preços para Conteúdo Digital (Dynamic Pricing Models for Digital	A utilização da tecnologia que essa patente possui, permite com que caso um usuário tenha um histórico de compras de produtos com preços altos, o preço oferecido ao mesmo por um produto seja superior, ao oferecido a um outro usuário que tenha um histórico diferente. Dessa forma, um

	(2008)	Content) busca definir os preços dos produtos dinamicamente, conforme o comportamento que os usuários possuem.	mesmo produto seria comprado por preços distintos conforme o histórico de compras realizado. Esse processo é, claramente, uma discriminação a grupos de consumidores que apresentam determinados comportamentos, onde uma análise de dados do histórico de um usuário, é utilizado para fazê-lo pagar mais pelos produtos, classificando usuários em grupos propensos a pagar mais ou menos por determinados objetos.
10	Casady (2011)	O departamento de polícia de New Castle Country - EUA busca trabalhar com policiamento preditivo, no qual são feitas análises de dados para que a polícia tenha apoio em desvendar crimes não resolvidos, e possa ocorrer prevenções de futuros crimes	Tais análises de dados levam em considerações diversos aspectos, que podem levar a um problema grave de existir policiamentos preditivos focados em raças ou gêneros. Por mais que, possa existir o cuidado em não utilizar tais variáveis, existem questões históricas e comportamentos preconceituosos, que levarão a padrões e comportamentos discriminatórios por parte das ações da polícia que utilizarão o sistema de tomada de decisões baseado nas análises realizadas. Isso porque, os algoritmos para realizar a mineração de dados, terão como base exemplos anteriores da polícia, que podem apresentar um comportamento preconceituoso já existente antes da utilização desse sistema de predição, que leva critérios classificatórios de pessoas para buscarem prever o acontecimento de algum crime.

Fonte: Elaborado pelos autores.

Por meio do quadro 1, é possível identificar casos de análises de dados que apresentam cunho discriminatório, em que fatores identificados são a causa dessa discriminação. Baseados no quadro construído, na próxima seção, são identificados e detalhados os fatores potencialmente causadores de discriminação em análises de dados em cenários de Big Data.

6 Resultados: fatores potencialmente discriminatórios

A análise bibliográfica, demonstrada no quadro 1, conduziu a definição de fatores que caso não sejam considerados, podem comprometer uma análise de dados, levando a resultados com caráter discriminatório. Tal definição ocorreu verificando que, em suma, a questão discriminatória se conduzia dessa forma por problemas encontrados em um desses cenários. O processo de identificação dos fatores ocorreu por meio da comparação das características discriminatórias apresentadas pelos casos representados no quadro 1, em que foi possível enumerar tais fatores, pelas semelhanças e diferenças dos cenários apresentados.

O primeiro fator foi identificado nos casos 1 e 3 do quadro, em que as amostras utilizadas são parciais ou não utilizam determinadas parcelas da população igualitariamente. Nesse cenário, verifica-se uma exclusão não casual, como por exemplo, quando o conjunto analisado abrange apenas pessoas que utilizam uma determinada tecnologia, excluindo, assim, pessoas que não são adeptas a mesma, seja por seu estilo de vida, por não possuir determinado aparelho, ou por preferência a outros sistemas. Observa-se também, casos em que determinados grupos não são considerados, simplesmente porque as fontes informacionais eram limitadas, não demonstrando a realidade vivida por toda a população (LERMAN, 2013).

No caso 2 do quadro, verificou-se outro fator, que diz respeito à utilização de dados incompletos ou incorretos como amostra de dados. Diferencia-se esse fator do primeiro caso, pois, no primeiro, os dados podem ser corretos e completos, embora, apresentem um problema quanto à abrangência da amostra. Já no caso dos dados incompletos e incorretos, o problema encontra-se nos dados que são parte dessa amostra. Taurion (2013) acredita que dados incompletos ou inconsistentes podem contaminar as análises realizadas, devendo ser eliminado tais dados. Uma consequência dessa contaminação relatada pelo autor são resultados incorretos, que podem conduzir a procedimentos discriminatórios, independentemente, se tais dados estiverem inexatos propositalmente ou não, pois os resultados poderão afetar a determinados grupos de indivíduos.

Um terceiro fator foi identificado por meio dos casos 5, 7, 8 e 10, que traz uma questão do uso de dados anteriores para o treinamento de modelos de mineração e análises de dados em ambientes de Big Data. Barocas e Selbst (2016) discorrem a respeito desse fator, ao afirmar que, quando são utilizados dados frutos de processos de tomada de decisão anteriores, é possível que uma carga de discriminação implícita nesses dados crie resultados preconceituosos. Um exemplo são empresas que decidem automatizar o processo de contratação de funcionários, tendo como alicerce os processos de contratação anteriores, onde, a visão de mundo dos gestores decidia qual era o perfil exigido de um candidato para a contratação. O problema desse processo, é que essa visão dos

gestores pode estar carregada de preconceitos, tornando o algoritmo que decidirá, automaticamente, as pessoas que serão contratadas, a reproduzir tais preconceitos, gerando novas discriminações.

Os três primeiros fatores identificados retratam problemas quanto à amostra de dados, pois todos os fatores apresentam uma possível discriminação relacionada à amostra de dados utilizada para realizar a análise.

Os próximos fatores, que serão descritos na sequência, apresentam características distintas dos primeiros, ocorrendo problemas no momento da análise dos dados em si.

Dessa forma, os casos 4, 5 e 7 indicam um fator potencialmente discriminatório quando há um olhar tendencioso ou equivocado por parte dos analistas de dados, ao direcionarem as análises dos dados. Tal problema acontece quando as hipóteses, os questionamentos e as variáveis escolhidas pelos analistas norteiam os resultados a um caminho que não são considerados determinados grupos sociais ou conjuntos específicos de indivíduos.

Barocas e Selbst (2016) trazem questões fundamentais que podem afetar diretamente os resultados, no momento em que são selecionadas as variáveis e as diretrizes das análises dos dados. Os autores relatam que em questões binárias como detecção de fraudes ou a verificação de spams, as definições das variáveis utilizadas na análise é mais simples, pois não existem muitas controvérsias quanto a quais características são apresentadas nesses casos, entretanto, em questões mais complexas, que afetam pessoas, como por exemplo, a definição da escolha dos funcionários que serão contratados por uma empresa, por meio de uma análise de dados, apresentam questões bastante subjetivas, que torna complexo o processo de seleção das variáveis utilizadas. Dessa forma, as escolhas feitas pelos analistas de dados poderão ocasionar em resultados que apresentem discriminação.

O quinto e último fator apresentado é identificado nos 6, 7, 8, 9 e 10, em que se verifica a ocorrência de classificações. No âmbito dos cenários de Big Data, os processos de classificação ocorrem com frequência, para que os sistemas de informação atinjam com mais eficiência aos indivíduos, sejam consumidores ou usuários, por exemplo. Mayer-Schönberger e Cukier (2013)

relatam que tal processo mostra-se positivo, para que, por exemplo, mecanismos de buscas classifiquem seus usuários em determinados grupos, e os anúncios e os resultados possam ser mais bem explorados naquele tipo de usuário. Afirmação que pode ser contestada pelo caso relatado por Sweeney (2013), em que se verifica que as classificações feitas por um buscador conduziram a discriminação racial.

Existem diversos exemplos que realizam classificação nas análises dos dados, como a escolha dos clientes com maior propensão a comprar um determinado produto, ou a negativa de crédito a um indivíduo pelo mesmo apresentar um perfil de não pagador, mesmo sem haver nenhuma informação de inadimplência em seu nome, ou ainda a contratação somente de pessoas que apresentam uma maior probabilidade de permanecerem por mais tempo no quadro de funcionários da empresa contratante, entre outros exemplos em que percebe-se tal fator.

Tais práticas vêm sendo utilizadas, mais constantemente, pelas organizações, na busca de aumentar o volume de vendas, obter crescimento nas taxas de lucros, diminuir a inadimplência e evitar um fluxo grande de empregadores contratados. Entretanto, todas essas atividades buscam realizar uma classificação entre aqueles aptos e não aptos a uma determinada questão, onde em um conjunto total, são separados somente poucos indivíduos ou dados que atendem a uma condição específica. Esse comportamento por si só, apresenta característica discriminatória, e pode afetar mais intensamente a determinados indivíduos ou grupos sociais, quando pessoas são privadas de ter acesso a algo por meio da utilização desses mecanismos.

Barocas e Selbst (2016) observam, no âmbito de relações trabalhistas, que os resultados obtidos por meio de uma análise classificatória fornecem informações preconceituosas, que considera somente alguns pontos de vistas, para justificar determinadas tomadas de decisões carregadas de discriminação, sem permitir com que as pessoas tenham possibilidades iguais, independentes de sua origem racial ou social, não considerando aspectos individuais de cada ser.

Destarte, resumindo os grupos e os fatores potencialmente discriminatórios, construiu-se o quadro 2 contendo o grupo dos fatores, os fatores e os casos identificados no quadro 1.

Quadro 2 - Fatores discriminatórios

GRUPO	FATOR	CASOS
Amostra dos dados	Amostra com dados parciais	1 e 3
	Amostra com dados incorretos ou incompletos	2
	Amostra de dados anteriores utilizadas para novas análises contendo uma carga de discriminação	5, 7, 8 e 10
Análise dos dados	Ótica do analista de dados	4, 5 e 7
	Classificação	6, 7, 8, 9 e 10

Fonte: Elaborado pelos autores.

Visualizando as informações descritas no quadro 2, identifica-se que, em alguns casos, foram identificados mais que um fator, como nos casos 5, 7, 8 e 10. Tal fato justifica-se, pois, a ocorrência de um fator discriminatório, pode levar a ocorrência de outros fatores discriminatórios, bem como, ocorrer em uma análise dois ou mais fatores de discriminação, que não necessariamente são causas uns dos outros fatores.

5.1 Relação entre as atividades do processo de análise de dados com os fatores discriminatórios

Os fatores identificados ocorrem em determinadas fases das análises de dados, sendo necessário identificar o momento que cada fator discriminatório ocorre, para que possam ser evitados. Para tal, utilizou-se as quatro atividades listas Taurion (2013): coleta de dados, limpeza, formatação e validação dos dados obtidos, integração e agregação dos dados e análise dos dados, descritas mais detalhadamente no capítulo 3. Nessa perspectiva, identificou-se que as quatro atividades relatadas, enquadram-se na Fase de Coleta do Ciclo de Vida dos

Dados, proposto por Santana (2013), em que o autor enumera quatro fases do Ciclo de Vida dentro do contexto da Ciência da Informação, sendo elas: Coleta, Armazenamento, Recuperação e Descarte. A partir disso, construiu-se o quadro 3 em que são relacionadas atividades de análise de dados no contexto do Big Data, com os fatores potencialmente discriminatórios. Destaca-se que a fase de coleta relatada por Taurion (2013), foi substituída por Obtenção, pois na perspectiva do Ciclo de Vida dos Dados, coleta corresponde a todas as fases destacadas pelo autor.

Quadro 3 - Relação entre os fatores potencialmente discriminatórios com as atividades de análises de dados no contexto do Big Data

FASE DE COLETA DO CICLO DE VIDA DOS DADOS				
Atividade	**Obtenção**	**Limpeza, formatação e validação**	**Integração e Agregação**	**Análise**
Fatores	Amostra com dados parciais	Amostra com dados incorretos ou incompletos	Amostra de dados anteriores utilizados para novas análises, contendo uma carga de discriminação	Classificação
	Amostra com dados incorretos ou incompletos		Classificação	Ótica do analista de dados

Fonte: Elaborado pelos autores.

O quadro 3 demonstra em qual etapa do processo de análise de dados no contexto de Big Data há possibilidades dos fatores discriminatórios estarem presentes, e assim, conduzir a análise para resultados passíveis de contestação. Com tal informação, torna-se possível que os analistas de dados possam precaver-se para evitar os fatores descritos.

A etapa de obtenção dos dados pode causar a discriminação por meio de dois fatores, com o uso de amostra de parciais, que considera o cenário imparcialmente, e com a utilização de dados incorretos ou incompletos. Nessa etapa, tais fatores podem ser evitados ao realizar análises que não restrinjam os dados utilizados a determinadas tecnologias ou mídias sociais, como nos casos 1 e 3 do quadro 1 pois isso excluirá pessoas que por qualquer motivo não tenha acesso a tais tecnologias. Outra ação, para evitar tais problemas, seria o cuidado que deve existir para verificar se os dados analisados estão corretos e completos,

porque caso uma análise, que tenha o intuito de encontrar padrões e soluções ou ter uma função preditiva, os dados utilizados devem ser muito fidedignos, e a atenção as fontes utilizadas se torna fundamental, para não existir resultados que poderão determinar, por exemplo, políticas públicas ou comportamento de doenças incorretamente, por estar baseadas em dados incompletos.

A etapa de limpeza, formatação e validação dos dados entra como um complemento da etapa de coleta de dados, entretanto, tendo um olhar sobre conjuntos específicos de dados dentro da massa total analisada. Assim, os cuidados quanto à amostra com dados incompletos ou incorretos se repete, devendo-se verificar com bastante cuidado possíveis falhas na massa de dados, pois nessa etapa erros pontuais, podem conduzir os resultados para conclusões equivocadas.

Na integração e agregação dos dados, quando são fundidos diversos conjuntos de dados, em uma etapa pré-análise, a primeira questão que o analista deve considerar, diz respeito às amostras de dados de análises anteriores que são utilizadas como padrão para análise atual, pois alguns conjuntos de dados que contém preconceitos anteriores irão conduzir as análises a replicarem esse comportamento. Dessa forma, o analista a fim de evitar que as análises sejam discriminatórias, deve refletir profundamente sobre conteúdo e resultados de análises anteriores, sejam elas feitas exclusivamente por pessoas, sejam elas realizadas em conjunto de pessoas com sistemas de informação tomadores de decisão. Casos como o número 10 do quadro 1, esclarecem como bases anteriores de análises estão conduzindo a resultados discriminatórios. Outro fator que influencia nessa etapa, são as classificações que podem ser executadas. O analista de dados deve ter um olhar bastante criterioso ao definir os cruzamentos de dados, e as variáveis condutoras do processo de integração de massas de dados, pois caso isso seja utilizado para realizar classificações de indivíduos e assim, realizar a agregação dos dados baseados nessas classificações, consequências discriminatórias poderão ocorrer, e influenciar diretamente nos resultados, como no caso 8 do quadro 1.

Na análise dos dados em si, novamente o fator discriminatório de classificação está presente, entretanto, tendo uma influência mais nos resultados

que são concluídos após as análises, em que as análises são realizadas buscando apresentar resultados classificatórios, entre indivíduos ou grupos sociais que atendem a determinados critérios e outros que não atendem, ou atendem apenas parcialmente. Os analistas de dados devem evitar a realização de classificação seguindo tais critérios, pois, inevitavelmente, ocorrerá discriminação para determinados grupos, seja uma discriminação racial, ou seja comportamental de consumo que um usuário apresente, como no caso 9 do quadro 1.

O segundo fator observado nessa etapa, é a questão da ótica do analista de dados, quando o modo como a análise é construída, conduz a resultados discriminatórios. Para evitar com que isso ocorra, o analista deve verificar se as variáveis utilizadas durante a construção das análises estão corretas, e se as questões centrais que baseiam a pesquisa, não estão norteando para respostas discriminatórias. Como no caso 4 do quaro 1, em que pesquisas com genomas estão sendo conduzidas a provar diferenças entre os genes de pessoas de raças distintas. Ficando claro que o analista de dados, está sendo diretamente responsável por tais resultados, pois não houve reflexão suficiente nas questões e nas variáveis utilizadas para a realização do estudo.

A partir das análises realizadas, verifica-se a necessidade de pesquisas que busquem discutir o impacto que Big Data vem tendo sobre a sociedade e a discriminação que está sendo causada a partir das análises de dados. Nesse sentido, Gordon (2015, p.28) destaca "[...] a importância de enfrentar as crescentes assimetrias de poder entre aqueles que criam os dados e as organizações que têm a capacidade de coletar, armazenar e analisar dados digitais."

Com essa afirmação, o autor insere uma discussão fundamental no que tange a discriminação nos cenários de Big Data, que são os criadores dos dados não tendo controle deles, perdendo autonomia sobre suas informações pessoais, que serão cada vez mais utilizadas para traçar perfis que possivelmente inserirá mais discriminação entre as pessoas.

7 Considerações finais

A chamada era do Big Data trouxe grandes revoluções na maneira como os dados são analisados, tanto dentro da ciência, quanto nas empresas e até no governo. A possibilidade de ter em mãos uma quantidade incontável de dados podendo inferir padrões e comportamentos, que servirão de base para a tomada de decisão, mostra-se uma ideia muito atrativa, utilizada por diversas instituições, e tendo como principal variável para os processos decisórios.

Nessa perspectiva, cientistas e analistas de Big Data relatam que as análises realizadas é um meio de não existir preconceitos e discriminação que ocorriam, anteriormente, com outros métodos de análises e pesquisa. Contudo, diversos pesquisadores vêm questionando tais afirmações, demonstrando que muitos resultados obtidos por meio de análises de dados no contexto do Big Data mostram sim, comportamentos discriminatórios, que mantém e ampliam os históricos de preconceitos propagados pelos séculos.

Questões mais graves como a indução de resultados de pesquisas científicas e modelos discriminatórios de buscadores de buscas, provam que o Big Data não só mantém comportamentos discriminatórios, como também os amplia. Torna-se necessário colocar a ciência para debater meios de utilizar tais dados, na busca de não se enviesar os resultados com a utilização de perguntas que induzem à respostas equivocadas.

Outro ponto de destaque, diz respeito aos questionamentos sobre a base de dados utilizados para realizar a inferência de ações a serem tomadas. Pesquisadores devem ter pensamento crítico ao analisar o significado, os problemas e as validades dos dados que estão sendo utilizados, para que a pesquisa possa demonstrar resultados reais e válidos.

Olhando para todas essas questões, as análises realizadas demonstraram que os meios utilizados, atualmente, para a extração dos valores dos dados, que são aqueles que irão cruzar e descobrir novas informações, apresentam uma grande relação com os fatores encontrados que permitem a discriminação de determinados grupos populacionais. A identificação desses fatores torna-se fundamental para que as análises de dados em cenários de Big Data possam considerar tais fatores, não permitindo que sejam gerados resultados discriminatórios frutos de análises no contexto do Big Data.

Nessa ótica, os fatores potencialmente discriminatórios encontrados nesse artigo, a partir da análise de diversos casos e referências relatadas, podem auxiliar nas precauções que devem ser tomadas no momento em que são utilizadas grandes massas de dados, para realização de análises. Sendo que, é de fundamental importância, que a Ciência da Informação seja incluída nesse debate, para que os vieses dos trabalhos e algoritmos construídos a partir de análises de dados possam ser entendidos e debatidos, buscando evitar uma ampliação de comportamentos discriminatórios por meio das tecnologias, para que, assim, os estudos com dados possam auxiliar aos pesquisadores e as organizações, sem trazer problemas a grupos sociais menos favorecidos.

8 Agradecimentos

Pesquisa financiada pela Fundação de Amparo à Pesquisa do Estado de São Paulo, processo nº 2015/01517-2 e pela Coordenação de Aperfeiçoamento de Pessoal do Nível Superior.

Referências

BAROCAS, Solon; SELBST, Andrew D. Big Data's Disparate Impact. **California Law Review**, Berkeley, v. 104, p. 671-732, 2016. Disponível em: <http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf > Acesso em: 20 out. 2015.

BOURDIEU, Pierre. **O poder simbólico**. Lisboa: DIFEL; Rio De Janeiro: Bertrand Brasil, 1989.

BUTLER, Declan. When Google got flu wrong. **Nature**, London, v. 494, n. 7436, p. 155-156, Feb. 2013. Disponível em: <http://www.nature.com/news/when-google-got-flu-wrong-1.12413> Acesso em: 26 jan. 2016.

CASADY, Tom. Police Legitimacy and Predictive Policing. **Geography & Public Safety**, Washington, v. 2. n. 4, p.1-16, mar. 2011. Disponível em: <http://www.nij.gov/topics/technology/maps/Documents/gps-bulletin-v2i4.pdf?Redirected=true>. Acesso em: 27 nov. 2015.

CHOW-WHITE, Peter A.; GREEN JR., Sandy E. Data Mining Difference in the Age of Big Data: Communication and the social shaping of genome technologies from 1998 to 2007. **International Journal of Communication**, Los Angeles, v. 7, p. 556-583, 2013. Disponível em: <http://ijoc.org/index.php/ijoc/article/view/1459/869>. Acesso em: 22 set. 2016.

CRAWFORD, Kate. Think again: big data. **Foreign Policy**, Washington, v. 9, 2013. Disponível em: <http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data>. Acesso em: 26 jan. 2016.

CROLL, Alistair. Big data is our generation's civil rights issue, and we don't know it. **Big data now**, Atlanta, p. 55-59, 2012. Disponível em: <http://solveforinteresting.com/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it/>. Acesso em: 26 jan. 2016.

DISCRIMINAÇÃO. In: DICIONÁRIO Priberam da Língua Portuguesa**.** Lisboa: Priberam Informática, 2011. Disponível em: <http://www.priberam.pt/DLPO/discriminação>. Acesso em: 22 set. 2016.

FRANK, Morgan R. et al. Happiness and the patterns of life: A study of geolocated tweets. **Scientific reports**, London, v. 3, Set. 2013. Disponível em: <http://www.nature.com/articles/srep02625?WT.ec_id=SREP-20130917?message-global=remove&WT.ec_id=SREP-20130917> Acesso em: 26 jan. 2016.

GOLDSTEIN, Benjamin A.; WINKELMAYER, Wolfgang C. Comparative health services research across populations: the unused opportunities in big data. **Kidney International**, Bruxelas, v. 87, n. 6, p. 1094-1096, Jun. 2015.

GORDON, Charly. **Big Data exclusions and disparate impact**: investigating the exclusionary dynamics of the Big Data phenomenon. 2015. 37 f. Dissertação (Mestrado em Mídia, Comunicação e Desenvolvimento) - London School of Economics and Political Science, Londres. 2015. Disponível em: <http://www.lse.ac.uk/media@lse/research/mediaWorkingPapers/MScDissertationSeries/2014/Charly-Gordon-MSc-Dissertation-Series-AF.pdf>. Acesso em: 19 jul. 2016.

KAKHANI, Manish Kumar; KAKHANI, Sweeti; BIRADAR, S. R. Research Issues in Big Data Analytics. **International Journal of Application or Innovation in Engineering & Management**, Etmadpur, v. 2, n. 8, Aug. 2013. Disponível em: <http://www.ijaiem.org/volume2issue8/IJAIEM-2013-08-29-070.pdf>. Acesso em: 22 set. 2016.

LERMAN, Jonas. Big data and its exclusions. **Stanford law review online**, Stanford, v. 66, Sep. 2013. Disponível em:

<https://www.stanfordlawreview.org/online/privacy-and-big-data-big-data-and-its-exclusions/>. Acesso em: 22 set. 2016.

LOHR, Steve. Big data, trying to build better workers. **The New York Times**, New York, Apr. 21th 2013. Tecnologia, p. 4. Disponível em: <http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html> Acesso em: 26 jan. 2016.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big data**: A revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt, 2013.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big Data**:** the management revolution. **Harvard Business Review**, Brighton, v. 90, n. 10, p. 61-67, oct. 2012. Disponível em: <https://hbr.org/2012/10/big-data-the-management-revolution#>. Acesso em: 22 set. 2016.

PERLROTH, Nicole. Fake twitter followers become multimillion-dollar business. **The New York Times**, Nova Iorque, 5 Abr. 2013. Bits. Disponível em: <http://bits.blogs.nytimes.com/2013/04/05/fake-twitter-followers-becomes-multimillion-dollar-business/?_r=0>. Acesso em: 25 nov. 2015.

PEW RESEARCH CENTER. **Social Networking Fact Sheet**. Washington, Pew Research Center, 2014. Disponível em: <http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/> Acesso em: 25 nov. 2015.

SANTANA, Ricardo Cesar Gonçalves. Ciclo de vida dos dados e o papel da Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15, Florianópolis, SC, 2013. **Anais eletrônicos...** Florianópolis, SC: ANCIB, 2013. Disponível em <http://enancib2013.ufsc.br/index.php/enancib2013/XIVenancib/paper/view/284/319> Acesso em: 2 fev. 2016.

STREET BUMP. **About street bump**. Boston, 2015. Disponível em: <http://www.streetbump.org/about>. Acesso em: 27 nov. 2015.

SWEENEY, Latanya. Discrimination in online ad delivery. **Ad Delivery**, Nova Iorque, v. 11, n. 3, p. 1-19, Apr. 2013. Disponível em: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240>. Acesso em: 22 set. 2016.

TAURION, Cezar. **Big data**. Rio de Janeiro: Brasport, 2013.

VALZ, Duane R. **Dynamic pricing models for digital content**. US 20080154798 A1. 26 jun. 2008. Disponível em: <http://www.google.com/patents/US20080154798>. Acesso em: 27 nov. 2015.

ZIKOPOULOS, Paul et al. **Understanding big data:** Analytics for enterprise class hadoop and streaming data. New York: McGraw-Hill, 2011. Disponível em: <http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=SWGE_IM_DD_USEN&htmlfid=IML14297USEN&attachment=IML14297USEN.PDF>. Acesso em: 22 set. 2016.

Big Data: potentially discriminatory factors in data analysis

Abstract: The experienced technological changes from the turn of the century caused a revolution in the Big Data society, in which the data analysis to determine patterns and behaviors could use large amounts of data. It is possible to notice that some analyses in the context of the Big Data are being conducted to generate discriminatory results. This study aims to identify factors that can potentially lead to discrimination in the process of data analysis. The methodology used was qualitative, exploratory and bibliographical, enumerating the discrimination cases. As the result, we identified possibly discriminatory factors and we provided an explanation of these factors. Through research, we noticed the need of showing deep reflection about the results that are obtained from the data analysis and the need of Information Science approaching such questions, in order to point out the paths to be taken.

Keywords: Big data. Data analysis. Discrimination. Potentially discriminatory factors.

Recebido: 10/02/2016

Aceito: 28/07/2016