Big Data


CÓPIA NÃO AUTORIZADA É CRIME
ABDR
ASSOCIAÇÃO BRASILEIRA DE DIREITOS REPROGRÁFICOS
RESPEITE O DIREITO AUTORAL

Big Data

Como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana

Viktor Mayer-Schönberger

Kenneth Cukier

Tradução: Paulo Polzonoff Junior

Do original: *Big Data*
Tradução autorizada do idioma inglês da edição publicada por Houghton Mifflin Harcourt Publishing Company

Copidesque: Adriana Kramer
Revisão: Gabriel Augusto Alves Pereira
Editoração Eletrônica: Thomson Digital

Elsevier Editora Ltda.
Conhecimento sem Fronteiras
Rua Sete de Setembro, 111 – 16° andar
20050-006 – Centro – Rio de Janeiro – RJ – Brasil
Rua Quintana, 753 – 8° andar
04569-011 – Brooklin – São Paulo – SP

Serviço de Atendimento ao Cliente
0800-0265340
sac@elsevier.com.br

ISBN original 978-05-440-0269-2
ISBN 978-85-352-7090-7
ISBN digital 978-85-352-7341-0

CIP-BRASIL. CATALOGAÇÃO NA PUBLICAÇÃO
SINDICATO NACIONAL DOS EDITORES DE LIVROS, RJ

M421b

Mayer-Schonberger, Viktor
Big Data : como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana / Viktor Mayer-Schonberger, Kenneth Cukier ; tradução Paulo Polzonoff Junior. - 1. ed. - Rio de Janeiro : Elsevier, 2013.
256 p. ; 23 cm

Tradução de: Big Data
ISBN 978-85-352-7090-7

1. Tecnologia da informação - Administração. 2. Administração de empresas - Inovações tecnológicas. 3. Desenvolvimento organizacional. 4. Administração de pessoal. I. Cukier, Kenneth. II. Título.

13-02352 CDD: 658.4038
CDU: 005.94

Agradecimentos

Ambos tivemos a felicidade de trabalhar e aprender com um gigante das redes de informação e inovação, Lewis M. Branscomb. Seu brilhantismo, eloquência, energia, profissionalismo, esperteza e curiosidade eterna continuam a nos inspirar. E quanto à sua agradável e inteligente parceira, Connie Mullin, pedimos desculpas por não adotarmos sua sugestão de chamar este livro de "Superdata".

Momin Malik foi um excelente assistente de pesquisas, com sua inteligência e pragmatismo. Temos o privilégio de sermos representados por Lisa Adams e David Miller, da Garamond, soberbos em todos os aspectos. Eamon Dolan, nosso editor, foi fenomenal — representante de uma rara safra de editores que têm uma noção quase perfeita de editar o texto e de desafiar nosso raciocínio, por isso o resultado é muito melhor do que jamais esperávamos. Agradecemos a Camille Smith por sua edição. Somos gratos a James Fransham, da *The Economist,* pela ótima verificação dos fatos e críticas ao manuscrito.

Somos especialmente gratos a todos os que lidam com o big data que perderam tempo nos explicando seu trabalho, principalmente Oren Etzioni, Cynthia Rudin, Carolyn McGregor e Mike Flowers.

***

Quanto aos agradecimentos individuais de Viktor: agradeço a Philip Evans, que está sempre dois passos à frente e expressa suas ideias com precisão e eloquência, em conversas que duram mais de uma década.

Também sou grato a meu ex-colega David Lazer, um dos primeiros e mais proeminentes acadêmicos na tecnologia de big data e cujos conselhos busquei várias vezes.

Agradeço aos participantes do 2011 Oxford Digital Data Dialogue (focado em big data), e principalmente ao copresidente Fred Cate, pelas valiosas discussões.

O Oxford Internet Institute, onde trabalho, criou o ambiente certo para este livro, com vários colegas envolvidos em pesquisas de big data. Não podia imaginar um lugar melhor para escrevê-lo. Também agradeço ao apoio do Keble College, onde sou professor convidado. Sem ele, não teria acesso a algumas importantes fontes primárias usadas neste livro.

A família sempre paga caro durante a elaboração de um livro. Não são apenas as horas que gastei diante do computador e no escritório, mas também as muitas, muitas horas em que estive fisicamente presente, mas perdido em pensamentos, pelo que peço perdão à minha esposa Birgit e ao pequeno Viktor. Prometo que me esforçarei mais.

***

Quanto aos agradecimentos pessoais de Kenn: agradeço aos vários cientistas de dados que me ajudaram, especialmente Jeff Hammerbacher, Amr Awadallah, DJ Patil, Michael Driscoll, Michael Freed e vários colegas da Google ao longo dos anos (incluindo Hal Varian, Jeremy Ginsberg, Peter Norvig e Udi Manber, entre outros, e as valiosas e breves conversas com Eric Schmidt e Larry Page).

Meu raciocínio foi enriquecido por Tim O'Reilly, gênio da era da internet, e por Marc Benioff, da Salesforce.com, um professor para mim. Os insights de Matthew Hindman foram imensuráveis, como sempre. James Guszcza, da Deloitte, foi incrivelmente útil, assim como Geoff Hyatt, velho amigo e empreendedor. Agradecimento especial a Pete Warden, filósofo que também trabalha com big data.

Muitos amigos ofereceram ideias e conselhos, entre eles John Turner, Angelika Wolf, Niko Waesche, Katia Verresen, Anna Petherick, Blaine Harden e Jessica Kowal. Entre outros que inspiraram temas para este livro estão Blaise Aguera y Arcas, Eric Horvitz, David Auerbach, Gil Elbaz, Tyler Bell, Andrew Wyckoff e muitos outros do OECD, Stephen Brobst e a equipe da Teradata, Anthony Goldbloom e Jeremy Howard, da Kaggle, Edd Dumbill, Roger Magoulas e a equipe da O'Reilly Media e Edward Lazowska. James Cortada é lendário. Agradeço ainda a Ping Li, da Accel Partners, e Roger Ehrenberg, da IA Ventures.

Na *The Economist,* meus colegas deram ótimas ideias e apoio. Agradeço em especial a meus editores, Tom Standage, Daniel Franklin e John Micklethwait, além de Barbara Beck, que editou a reportagem especial "Data, Data Everywhere", a origem deste livro. Henry Tricks e Dominic Zeigler, meus colegas em Tóquio, foram exemplos de conduta por sempre procurarem as novidades e as expressarem maravilhosamente bem. Oliver Morton nos deu sua sabedoria de sempre quando foi necessária.

O Salzburg Global Seminar, na Áustria, ofereceu a combinação perfeita de repouso idílico e investigação intelectual que me ajudaram a escrever e pensar. Uma rodada de discussão no Aspen Institute, em julho de 2011, me deu várias ideias, pelas quais agradeço aos participantes e ao organizador, Charlie Firestone. Também agradeço a Teri Elniski pelo apoio.

Frances Cairncross, reitor do Exeter College, Oxford, me ofereceu um lugar tranquilo para escrever e também muito encorajamento. É um gesto de humildade fazer perguntas de tecnologia e sociedade para alguém que as evocou há uma década e meia, em *The Death of Distance,* obra que me inspirou como jovem jornalista. Foi bom atravessar o jardim de Exeter todas as manhãs sabendo que eu carregava a mesma tocha que ela carregara, embora a chama brilhasse muito mais nas mãos dela.

Minha mais profunda admiração vai para minha família, que me deu apoio – ou, o que foi mais comum, deu apoio à minha ausência. Meus pais, irmã e outros parentes merecem meu agradecimento, mas reservo a maior parte de minha gratidão a minha esposa, Heather, e nossos filhos, Charlotte e Kaz, sem seu apoio, o encojaramento e as ideias, este livro jamais teria sido possível.

***

Nós dois agradecemos a várias pessoas que discutiram o tema do big data conosco, antes mesmo que o termo se popularizasse. Neste sentido, agradecemos especialmente aos participantes, ao longo dos anos, da Rueschlikon Conference on Information Policy, que Viktor ajudou a organizar e do qual Kenn foi assistente. Agradecemos especialmente a Joseph Alhadeff, John Seely Brown, Peter Cullen, Ed Felten, Jeff Jonas, Douglas Merrill, Cory Ondrejka, Herbert Burkert (que nos apresentou ao comandante Maury), Urs Gasser, Nicklas Lundblad, Paul Schwartz e Joi Ito.

Viktor Mayer-Schönberger
Kenneth Cukier
Oxford/Londres, agosto de 2012

Capítulo 1

Agora

Em 2009, um novo vírus da gripe foi descoberto. Combinando elementos dos vírus que causavam a gripe aviária e suína, este novo tipo, chamado H1N1, se espalhou rapidamente. Em semanas, órgãos de saúde pública em todo o mundo temeram que uma terrível pandemia estivesse a caminho. Algumas pessoas previram um caso como o da gripe espanhola de 1918, que infectou meio bilhão de pessoas e matou dezenas de milhões. Pior, não havia vacina disponível contra o novo vírus. A única esperança que as autoridades de saúde tinham era desacelerar a propagação. Mas, para isso, elas precisavam saber onde estava a doença.

Nos Estados Unidos, os CDCs (Centers for Disease Control - Centros de Controle e Prevenção de Doenças) pediram que os médicos informassem todos os novos casos de gripe. Ainda assim, o cenário obtido da pandemia estava sempre defasado em uma ou duas semanas. As pessoas podiam se sentir mal por dias, mas tinham de esperar para se consultar com um médico. A transferência das informações para as organizações centrais levava tempo, e o CDC só computava os números uma vez por semana. Com a rápida disseminação da doença, o intervalo de duas semanas era uma eternidade. Este atraso impediu que as agências de saúde pública obtivessem o real panorama da pandemia em momentos cruciais.

Poucas semanas antes de o vírus H1N1 aparecer nas manchetes, engenheiros da Google publicaram um notável trabalho na revista científica *Nature* que causou comoção entre autoridades de saúde e cientistas da computação, mas que, de resto, foi ignorado. Os autores explicaram como a Google pôde "prever" a disseminação da gripe de inverno nos Estados Unidos, não apenas nacionalmente, mas em regiões específicas e até mesmo em estados. A empresa obteve essa previsão ao analisar os termos mais pesquisados na internet. Como o Google recebe mais de três bilhões de pesquisas por dia e as salva, a empresa tinha muitos dados com os quais trabalhar.

A Google pegou os 50 milhões de termos de busca mais comuns dos americanos e os comparou a lista com os dados do CDC sobre a disseminação da gripe entre 2003 e 2008. A ideia era identificar pessoas infectadas pelo vírus da gripe de acordo com o que pesquisavam na internet. Outros haviam tentado fazer o mesmo com os termos de busca, mas ninguém dispunha de tantos dados, poder de processamento e conhecimento estatístico como a Google.

Apesar de os funcionários da Google pensarem que as buscas poderiam ser usadas para a obtenção de informações sobre a gripe – ao escrever frases como "remédio para tosse e febre" –, essa não era a questão: eles não sabiam e criaram um sistema sem muita utilidade. Tudo o que ele fazia era procurar correlações entre a frequência de

certas buscas e a disseminação da gripe em relação ao tempo e ao espaço. No total, eles processaram impressionantes 450 milhões de modelos matemáticos diferentes a fim de testar os termos de busca, comparando suas previsões com os casos reais de gripe registrados pelo CDC em 2007 e 2008. E tiveram sorte: o programa descobriu uma combinação de 45 termos de busca que, quando usados juntos num modelo matemático, tinham forte correlação entre a previsão e os números oficiais. Como o CDC, eles podiam ver por onde a gripe havia se espalhado, mas, ao contrário do CDC, podiam apontar a disseminação quase em tempo real, e não com uma ou duas semanas de atraso.

Assim, quando a crise do vírus H1N1 ocorreu, em 2009, o sistema da Google provou ser um indicador mais útil e pontual que as estatísticas do governo, com seu atraso natural. As autoridades de saúde pública estavam munidas de valiosas informações.

Incrivelmente, o método da Google não envolve o uso de amostras de saliva ou o contato com médicos. Ele se baseia em "big data" – a capacidade de uma sociedade de obter informações de maneiras novas a fim de gerar ideias úteis e bens e serviços de valor significativo. Assim, quando a próxima pandemia surgir, o mundo terá um instrumento melhor à disposição para prever e, assim, evitar a disseminação da doença.

A saúde pública é apenas uma área na qual o big data está fazendo a diferença. Setores da economia também estão sendo reformulados. A compra de passagens aéreas é um bom exemplo.

Em 2003, Oren Etzioni precisava viajar de Seattle para Los Angeles, para o casamento de seu irmão caçula. Meses antes da cerimônia, ele entrou na internet e comprou uma passagem aérea, acreditando que, quanto antes o fizesse, mais barato seria o bilhete. Durante o voo, por curiosidade, ele perguntou à pessoa ao seu lado quanto ela pagara pela passagem e quando a comprara. O homem havia pagado consideravelmente menos que Etzioni, mesmo tendo comprado a passagem recentemente. Furioso, Etzioni perguntou aos demais passageiros. A maioria deles havia pagado menos.

Para a maioria de nós, a sensação de traição econômica teria se dissipado assim que tivéssemos recolhido as bandejas e voltado as poltronas à posição vertical. Mas Etzioni é um dos principais cientistas da computação dos Estados Unidos. Ele vê o mundo como uma série de problemas de dados – problemas que ele é capaz de resolver. E Etzione os domina desde que se formou em Harvard, em 1986, ao se tornar mestre em Ciência da Computação.

De sua varanda na University of Washington, ele deu início a várias empresas de "big data" antes de o termo se tornar conhecido. Ajudou a criar um dos primeiros mecanismos de busca da internet, o MetaCrawler, lançado em 1994 e comprado pela InfoSpace, então uma grande empresa da área. Ele cofundou o Netbot, primeiro grande site de comparação de preços, que vendeu para a Excite. Sua *startup* para extrair significado de documentos de texto, chamada ClearForest, mais tarde foi adquirida pela Reuters.

De volta à terra firme, Etzioni estava determinado a descobrir uma maneira de as pessoas saberem se o preço das passagens on-line é ou não um bom negócio. Um assento de avião é uma *commodity:* cada um é basicamente indiferenciável dos demais assentos do mesmo voo. Ainda assim, os preços variam muito, com base em vários fatores, na maior parte conhecidos apenas pelas companhias aéreas.

Etzioni concluiu que não precisava decifrar o enigma ou o motivo para a diferença de preço. Ao contrário, tinha de prever se os preços mostrados aumentariam ou diminuiriam

no futuro. Era possível, ainda que não fosse fácil. Era preciso analisar todos os preços de passagem de uma rota e examinar os preços pagos em relação aos dias anteriores ao voo.

Se o preço médio de uma passagem tendia a diminuir, fazia sentido esperar e comprar a passagem mais tarde. Se o a tendência era aumentar, o sistema recomendaria que se comprasse a passagem imediatamente. Em outras palavras, era preciso uma versão estendida da pesquisa informal que Etzioni fizera a 30 mil pés. Obviamente, tratava-se de outro enorme problema de ciência da computação, mas que ele era capaz de resolver. Então Etzioni se pôs a trabalhar.

Com uma amostragem de 12 mil preços obtidos em um website de viagens durante 41 dias, Etzioni criou um modelo de previsão que proporcionou grande economia aos passageiros simulados. O modelo não entendia o *porquê*, apenas *o quê*. Isto é, ele não conhecia as variáveis por trás da decisão das companhias aéreas, como a quantidade de assentos não vendidos, a temporada ou a mágica possibilidade da redução do preço em uma noite de sábado. O sistema baseava suas previsões no conhecido: probabilidades obtidas a partir de dados de outros voos. "Comprar ou não comprar, eis a questão", pensou Etzioni. Assim, ele chamou o projeto de pesquisa de Hamlet.

O pequeno projeto evoluiu para uma *startup* de capital de risco chamada Farecast. Ao prever se e quanto o preço de uma passagem aérea aumentaria ou diminuiria, a Farecast possibilitou que os escolhessem quando apertar o botão de compra. O sistema os muniu de informações às quais nunca tiveram acesso. De acordo com um princípio de transparência, a Farecast até mesmo anunciava o grau de precisão de suas previsões e também apresentava essas informações ao público.

Para funcionar, o sistema precisava de muitos dados. Para melhorar seu desempenho, Etzioni conseguiu um dos bancos de dados de reservas da indústria aeroviária. Com essa informação, o sistema podia fazer previsões com base em casa assento, em cada voo, na maioria das rotas da aviação comercial americana ao longo de um ano. A Farecast usava agora quase 200 bilhões de registros de preços para fazer previsões. Assim, economizava muito para os usuários.

Com seus cabelos castanhos, sorriso aberto e aparência de querubim, Etzioni não parece a pessoa que negaria à indústria aeroviária milhões de dólares em faturamento potencial. Na verdade, ele tenta fazer mais que isso. Em 2008, ele planejava aplicar o método a outros bens, como quartos de hotel, ingressos de shows e carros usados: qualquer serviço com pouca diferenciação, alta variação de preço e toneladas de dados. Mas antes que pudesse dar continuidade a seus planos, a Microsoft bateu à sua porta, comprou a Farecast por cerca de US$110 milhões e a integrou ao sistema de buscas Bing. Em 2012, o sistema acertava 75% das previsões e os passageiros economizavam, em média, US$50 por passagem.

A Farecast é um ícone de empresas de big data e um exemplo de para onde o mundo está indo. Cinco ou dez anos antes, Etzioni não a poderia ter criado. "Teria sido impossível", diz ele. A quantidade de poder de processamento e de armazenamento de que ele precisava era cara demais. Mas apesar de as mudanças na tecnologia terem sido um fator crítico que possibilitou o negócio, algo mais importante e sutil também mudou: a mentalidade sobre como os dados poderiam ser usados.

Os dados não eram mais considerados estáticos e banais, cuja utilidade terminava depois que o objetivo da coleta era alcançado, como no pouso do avião (ou, no caso

da Google, após o processamento da busca). Em vez disso, os dados se tornaram matéria-prima dos negócios, um recurso econômico vital, usado para criar uma nova forma de valor econômico. Na verdade, com a mentalidade certa, os dados podem ser reutilizados para se tornarem fonte de inovação e novos serviços. Eles podem revelar segredos para aqueles com humildade, disposição e instrumentos para ouvir.

DEIXANDO OS DADOS FALAREM

É fácil ver os frutos da sociedade da informação com um celular em cada bolso, um computador em cada mochila e grandes sistemas de tecnologia da informação em todos os escritórios. Mas discreta, contudo, é a informação em si. Meio século depois de os computadores entrarem no meio social, os dados começaram a se acumular a ponto de algo novo e especial começar a acontecer. O mundo não apenas está mais cheio de informação como também a informação está se acumulando com mais rapidez. A mudança de escala levou a uma mudança de estado. A mudança quantitativa gerou uma mudança qualitativa. Ciências como a astronomia e a genômica, que vivenciaram uma explosão nos anos 2000, cunharam o termo "big data". Hoje o conceito está migrando para todos os campos do conhecimento humano.

Não há uma definição rigorosa para o termo. A princípio, a ideia era a de que o volume de informação crescera tanto que a quantidade examinada já não cabia na memória de processamento dos computadores, por isso os engenheiros tiveram de aprimorar os instrumentos que utilizavam para a análise. Essa é a origem de novas tecnologias de processamento, como a MapReduce da Google e sua equivalente de código aberto, Hadoop, lançada pela Yahoo. Elas permitem que se gerenciem muito mais dados que antes, e os dados – isto é importante – não precisavam ser alocados em fileiras ou nas clássicas tabelas. Outras tecnologias de dados que dispensam as rígidas hierarquias e a homogeneidade também estão no horizonte. Ao mesmo tempo, como as empresas da internet podem coletar imensas quantidades de dados e têm um incentivo financeiro para utilizá-los, elas se tornaram as principais usuárias das mais recentes tecnologias de processamento, superando empresas off-line, em alguns casos com décadas de experiência.

Uma maneira de pensar na questão hoje – a que usamos neste livro – é: big data se refere a trabalhos em grande escala que não podem ser feitos em escala menor, para extrair novas ideias e criar novas formas de valor de maneiras que alterem os mercados, as organizações, a relação entre cidadãos e governos etc.

Mas isto é apenas o começo. A era do big data desafia a maneira como vivemos e interagimos com o mundo. Mais importante, a sociedade precisará conter um pouco da obsessão pela causalidade e trocá-la por correlações simples: sem saber o *porquê,* apenas *o quê*. Essa mudança subverte séculos de práticas consagradas e desafia nossa compreensão mais básica de como tomamos decisões e compreendemos a realidade.

O big data marca o início de uma importante transformação. Como tantas novas tecnologias, o big data com certeza se tornará vítima de um notório ciclo do Vale do Silício: depois de aparecer em capas de revistas e em conferências do ramo, a tendência será ignorada, e muitas das empresas de análise de dados naufragarão. Mas tanto a paixão quanto a danação ignoram profundamente a importância do

que está acontecendo. Assim como o telescópio permitiu que compreendêssemos o universo e o microscópio permitiu que entendêssemos os germes, as novas técnicas de coleta e análise de grandes quantidades de dados nos ajudarão a compreendermos nosso mundo de uma maneira que só estamos começando a admirar. Neste livro, não pretendemos ser evangelizadores do big data, e sim mensageiros. E, mais uma vez, a verdadeira revolução não está nas máquinas que calculam os dados, e sim nos dados em si e na maneira como os usamos.

Para avaliar em que ponto a revolução da informação já se encontra, considere tendências em todo o espectro da sociedade. Nosso universo digital está se expandindo constantemente. Considere a astronomia, por exemplo. Quando a Sloan Digital Sky Survey teve início, em 2000, seu telescópio no Novo México coletou mais dados nas primeiras semanas que em toda a história da astronomia. Em 2010, o arquivo da pesquisa contava com incríveis 140 terabytes de informação. Mas uma sucessora, a Large Synoptic Survey Telescope, no Chile, prevista para entrar em produção em 2016, coletará a mesma quantidade de dados a cada cinco dias.

Essas astronômicas quantidades também são encontradas em outros lugares do mundo. Quando os cientistas desvendaram o genoma humano, em 2003, precisaram de toda uma década para sequenciar três bilhões de pares-base. Hoje, uma década mais tarde, uma única instalação pode sequenciar a mesma quantidade de DNA num só dia. Nas finanças, aproximadamente sete bilhões de ações trocam de mãos todos os dias nos Estados Unidos, das quais cerca de dois terços são comercializados por algoritmos computadorizados com base em modelos matemáticos que calculam grandes quantidades de dados para prever ganhos e reduzir riscos.

Empresas de internet têm sido especialmente afetadas. A Google processa mais de 24 petabyte de dados por dia, volume milhares de vezes maior que todo o material impresso na Biblioteca do Congresso dos Estados Unidos. O Facebook, empresa que não existia há uma década, recebe mais de 10 milhões de fotos a cada hora. Os usuários do Facebook clicam no botão "curtir" ou deixam um comentário cerca de três bilhões de vezes por dia, criando uma trilha digital que a empresa pode usar para descobrir mais sobre as preferências dos usuários.

Enquanto isso, 800 milhões de usuários mensais do YouTube, da Google, enviam mais de uma hora de vídeo por segundo. A quantidade de mensagens no Twitter cresce a uma taxa de 200% ao ano e, em 2012, excedeu 400 milhões de tweets por dia.

Da ciência à saúde, das finanças à internet, os setores podem ser diferentes, mas juntos contam uma história semelhante: a quantidade de dados no mundo cresce rapidamente e supera não apenas nossas máquinas como nossa imaginação.

Muitas pessoas tentaram desvendar a quantidade de informações que nos cercam e calcular seu crescimento, com graus de sucesso variados porque medem elementos diferentes. Um dos estudos mais abrangentes foi feito por Martin Hilbert, da Annenberg School for Communication and Journalism, da University of California. Ele tentou quantificar tudo o que já foi produzido, armazenado e comunicado, o que inclui não apenas livros, imagens, e-mails, fotografias, música e vídeo (analógico e digital), como também videogames, ligações telefônicas e até mesmo cartas e sistemas de navegação para carros. Ele também incluiu transmissões de televisão e rádio, com base na audiência.

De acordo com Hilbert, existiam em 2007 mais de 300 exabytes de dados armazenados. Para entender o que isso significa em termos mais humanos, pense que um filme digital pode ser comprimido num arquivo de um gigabyte. Um exabyte é um bilhão de gigabytes. Em resumo, é muito. O interessante é que, em 2007, apenas 7% dos dados eram analógicos (papel, livro, fotografia e assim por diante). O restante era digital. Mas há pouco tempo, o cenário parecia bem diferente. Apesar de as ideias de "revolução da informação" e "era digital" existirem desde os anos 1960, elas só se tornaram em parte realidade. No ano 2000, apenas um quarto da informação armazenada no mundo era digital. Os outros três quartos estavam em papel, filme, vinil, fitas magnéticas e dispositivos do gênero.

A quantidade de informação digital na época não era muita – insignificante para aqueles que navegam na internet e compram livros digitais há tempos. (Na verdade, em 1986, cerca de 40% de toda a potencia computacional do mundo existia na forma de calculadoras de bolso, o que representava um poder de processamento maior que todos os computadores pessoais da época). Mas como os dados digitais se expandem rapidamente – dobram em pouco mais de três anos, de acordo com Hilbert –, a situação se inverteu. A informação analógica, por oposição, mal cresce. Assim, em 2013, a quantidade de informações armazenadas no mundo é estimada em 1200 exabytes, dos quais menos de 2% são analógicos.

Não há uma boa maneira de pensar o que significa essa quantidade de dados. Se fossem impressas, as páginas cobririam toda a superfície dos Estados Unidos em 52 camadas. Se fosse armazenada em CD-ROMs empilhados, se estenderiam em cinco pilhas até a lua. No século III a.C., enquanto Ptomoleu II tentava guardar uma cópia de todos os livros escritos, a Biblioteca de Alexandria representava a soma de todo o conhecimento do mundo. O dilúvio digital que agora assola o mundo é o equivalente a dar a todas as pessoas que vivem na Terra hoje 320 mais informações do que se estima que havia na Biblioteca de Alexandria.

O mundo muda rapidamente. A quantidade de informação armazenada cresce quatro vezes mais rápido que a economia mundial, enquanto a capacidade de processamento dos computadores cresce nove vezes mais rápido. Todos são afetados pelas mudanças.

Compare a atual torrente de dados com a primeira revolução da informação, a da impressora de Gutenberg, inventada em 1439. Nos 50 anos entre 1453 e 1503, cerca de oito milhões de livros foram impressos, de acordo com a historiadora Elizabeth Eisenstein, número maior do que todos os escribas europeus produziram desde a fundação de Constantinopla, 1200 anos antes. Em outras palavras, foram necessários 50 anos para toda a informação dobrar na Europa, em comparação com cerca de três anos atualmente.

O que esse aumento significa? Peter Norvig, especialista em inteligência artificial da Google, costuma fazer uma analogia de imagens. Primeiro, ele pede que pensemos nas tradicionais imagens de cavalo das pinturas rupestres de Lascaux, na França, que datam do Paleolítico, com aproximadamente 17 mil anos. Depois, pense na imagem de um cavalo – ou melhor, nas pinceladas de Pablo Picasso, que não parecem muito diferentes das pinturas rupestres. Na verdade, quando Picasso viu as imagens de Lascaux, disse que, desde então, "não inventamos nada".

As palavras de Picasso eram verdadeiras num nível, mas não em outro. Lembre-se da fotografia do cavalo. Antes era necessário muito tempo para desenhar a imagem de um cavalo, mas agora a representação poderia ser muito mais rápida. Essa é uma mudança, mas talvez não seja a mais essencial, já que se trata basicamente da mesma imagem: um cavalo. Hoje, contudo, pede Norvig, pense na captação da imagem do cavalo a uma velocidade de 24 frames por segundo. Agora a mudança quantitativa gerou uma mudança qualitativa. Um filme é fundamentalmente diferente de uma imagem congelada. O mesmo acontece com o big data: ao alterarmos a quantidade, mudamos a essência.

Pense numa analogia da nanotecnologia – na qual tudo fica menor, não maior. O princípio por trás da nanotecnologia é o de que, quando você chega ao nível molecular, as propriedades físicas podem se alterar. Ao saber o que significam essas novas características, você pode criar materiais para construir o que não podia ser feito antes. Na escala nanométrica, por exemplo, é possível obter metais e cerâmicas mais flexíveis. Do mesmo modo, quando aumentamos a escala de dados com a qual trabalhamos, ganhamos margem para inovar, o que não era possível quando trabalhávamos com quantidades menores.

Às vezes, os limites com os quais convivemos, supondo que sejam os mesmos para tudo, são apenas funções da escala na qual operamos. Use uma nova analogia, mais uma vez da ciência. Para os humanos, a lei física mais importante é a gravidade: ela reina sobre tudo o que fazemos. Mas, para minúsculos insetos, a gravidade é quase imaterial. Para alguns, como os que vivem sobre a lâmina d́ agua, a lei mais importante do universo físico é a tensão superficial, que lhes permite cruzar uma poça sem afundar.

Com a informação, assim como na física, o tamanho importa. Assim, ao combinar centenas de bilhões de termos de busca, a Google mostrou ser capaz de identificar o surgimento de um surto de gripe quase tão bem quanto os dados oficiais com base nos pacientes que visitam o médico – e pôde gerar uma resposta quase em tempo real, muito mais rápido que as fontes oficiais. Do mesmo modo, a Lei de Etzioni pode prever a volatilidade do preço de uma passagem de avião e, assim, dar um poder econômico significativo para os consumidores. Mas ambos só conseguem isso pela análise de centenas de bilhões de dados.

Esses dois exemplos mostram o valor científico e social do big data, assim como em que medida eles podem se tornar fonte de valor econômico. Os exemplos marcam duas maneiras pelas quais o mundo do big data está fadado a abalar tudo, dos negócios às ciências e saúde, governo, educação, economia, ciências humanas e todos os demais aspectos da sociedade.

Apesar de estarmos apenas nos primórdios do big data, nós o usamos diariamente. Filtros antispam são projetados para automaticamente se adaptarem às mudanças dos tipos de lixo eletrônico: o software não podia ser programado para bloquear o spam cujo assunto seja "via6ra" ou suas variantes. Sites de namoro formam pares com base em como suas várias características se correspondem às de relacionamentos anteriores. O corretor automático dos *smartphones* analisa nossas ações e acrescenta novas palavras a seus dicionários com base no que escrevemos. Mas é apenas o começo. De carros que podem detectar como guiamos ou freamos até o computador Watson da IBM que vence humanos no programa *Jeopardy!,* a abordagem alterará muitos aspectos do mundo em que vivemos.

Em essência, big data relaciona-se com previsões. Apesar de ser descrito como um ramo da ciência da computação chamado inteligência artificial e, mais especificamente, uma área chamada "aprendizado de máquina", esta ideia é enganosa. Big data não tem a ver com tentar "ensinar" um computador a "pensar" como ser humano. Ao contrário, trata-se de aplicar a matemática a enormes quantidades de dados a fim de prever probabilidades: a chance de um e-mail ser um spam; de as letras "msa" na verdade significarem "mas"; de a trajetória e velocidade de uma pessoa que atravesse a rua significarem que ela a atravessará a tempo – o carro com piloto automático precisa reduzir apenas um pouco a velocidade. O segredo é que esses sistemas funcionam porque são alimentados por enormes quantidades de dados, que formam a base das previsões. Além disso, os sistemas são criados para se aperfeiçoarem com o tempo, ao continuamente analisar os melhores sinais e padrões a fim de encontrar mais dados para uso.

No futuro – e mais cedo do que pensamos –, muitos aspectos do nosso mundo, hoje sujeitos apenas à visão humana, serão complementados ou substituídos por sistemas computadorizados. Não apenas a direção ou a formação de pares, mas também tarefas mais complexas. Afinal, a Amazon pode recomendar o livro ideal, o Google pode ranquear o site mais relevante, o Facebook conhece nossos gostos e o LinkedIn exalta quem conhecemos. As mesmas tecnologias serão aplicadas para o diagnóstico de doenças, a recomendação de tratamentos e talvez até para a identificação de "criminosos" antes mesmo que eles cometam um crime. Assim como a internet mudou radicalmente o mundo ao acrescentar a comunicação aos computadores, o big data mudará aspectos fundamentais da vida ao lhe dar uma dimensão quantitativa que ela jamais teve.

MAIS, CONFUSO, BOM O SUFICIENTE

O big data será uma nova fonte de valor econômico e inovação. Mas há mais em jogo. O predomínio do big data representa três mudanças na forma como analisamos informações que transformam a maneira como entendemos e organizamos a sociedade.

A primeira mudança é descrita no Capítulo 2. Neste novo mundo, podemos analisar muito mais dados. Em alguns casos, podemos até mesmo processar *tudo* o que está relacionado com determinado fenômeno. Desde o século XIX, a sociedade depende do uso de amostragens quando se trata de grandes quantidades. Mas a necessidade dessas amostragens remonta a um período de escassez de informações, produto dos limites naturais de se interagir com as informações numa era analógica. Antes do domínio das tecnologias digitais de alto desempenho, não percebíamos a amostragem como algo artificial, mas como algo comum. Usar todos os dados permite que vejamos detalhes nunca antes vistos quando estávamos limitados a quantidades menores. O big data nos dá uma visão clara do que é granular: subcategorias e submercados que as amostragens não alcançam.

A análise de maior quantidade de dados também permite abrandar nosso rigor, a segunda mudança, que identificamos no Capítulo 3. É uma troca: com menos erros de amostragens, podemos aceitar mais erros de medição. Quando a capacidade de medição é limitada, contamos apenas o mais importante. O esforço para conseguir

o número exato é apropriado. Não há sentido em vender o gado se o comprador não tem certeza se são 100 ou apenas 80 cabeças no rebanho. Até recentemente, todas as ferramentas digitais se baseavam na exatidão: presumíamos que os sistemas de dados recuperariam registros perfeitamente adaptáveis à nossa busca do mesmo modo que as tabelas computam os números em uma coluna.

Esse tipo de raciocínio é uma influência do ambiente de "pequenos dados": com tão pouco para medir, o rigor pela contagem precisa dos dados era altíssimo. De certa maneira, é óbvio: uma lojinha pode contar o dinheiro na caixa registradora até o último centavo, mas não faríamos o mesmo para o produto interno bruto de um país – na verdade, não poderíamos fazê-lo. À medida que a escala aumenta, as imprecisões quantitativas também aumentam.

A exatidão necessita de dados precisos. Pode funcionar para pequenas quantias e, claro, algumas situações ainda a requerem: uma pessoa tem ou não dinheiro no banco para passar um cheque. Mas ao usarmos bancos de dados mais amplos, podemos lançar um pouco dessa rigidez para o mundo do big data.

Em geral, o big data é confuso, varia em qualidade e está distribuído em incontáveis servidores pelo mundo. Com o big data, frequentemente nos satisfazemos com uma sensação aproximada de direção, sem a necessidade de um milimétrico conhecimento do fenômeno. Mas não abdicamos completamente da exatidão; apenas de nossa devoção a ela. O que perdemos em precisão microscópica ganhamos em visão macroscópica.

Essas duas mudanças levam a uma terceira, que explicamos no Capítulo 4: um afastamento da antiga busca pela causalidade. Fomos condicionados a procurar causas, mesmo quando a busca pela causalidade é difícil e nos leva a caminhos errados. No mundo do big data, por sua vez, não temos de nos fixar na causalidade; podemos descobrir padrões e correlações nos dados que nos propiciem novas e valiosas ideias. As correlações podem não nos dizer com exatidão *por que* algo está acontecendo, mas nos alertam que *algo* está acontecendo.

Em muitas situações, isso é bom o suficiente. Se milhões de registros médicos eletrônicos revelam que pessoas com câncer que tomam certa combinação de aspirina e suco de laranja apresentam remissão da doença, a causa exata da melhora na saúde talvez seja menos importante que o fato de os doentes sobreviverem. Do mesmo modo, se podemos economizar dinheiro ao saber a melhor hora para comprar uma passagem de avião sem entender o método da precificação, isso é bom o suficiente. Big data tem a ver com *"o quê",* e não com o *porquê.* Nem sempre precisamos saber a causa de um fenômeno; em vez disso, podemos deixar que os dados falem por si.

Antes do big data, nossa análise geralmente se limitava a uma pequena quantidade de hipóteses que definíamos bem antes de coletarmos os dados. Quando deixamos que os dados falem por si, podemos gerar conexões que nem sabíamos que existiam. Assim, alguns fundos hedge usam o Twitter para prever o desempenho do mercado de ações. A Amazon e a Netflix baseiam suas recomendações de produtos nas diversas interações em seus sites. Twitter, LinkedIn e Facebook mapeiam o "gráfico social" das relações entre os usuários para aprender mais sobre suas preferências.

Claro que os seres humanos analisam dados há milênios. A escrita foi desenvolvida na antiga Mesopotâmia porque os burocratas queriam um instrumento eficiente para registrar e manter o controle de informações. Desde os tempos bíblicos, os governos

organizam censos para reunir dados sobre os cidadãos, e há 200 anos atuários do mesmo modo coletam dados que tratam dos riscos que tentam compreender – ou pelo menos evitar.

Mas a era analógica de coleta e análise de dados era muito cara e demorada. Novas questões geralmente exigiam novas coletas de dados e novas análises.

O grande passo na direção de um gerenciamento mais eficiente dos dados aconteceu com o advento da digitalização, que tornou as informações analógicas compreensíveis a computadores, o que também facilitou e barateou o armazenamento e processamento. Esse avanço aumentou drasticamente a eficiência. A coleta e análise de informações, que antes levavam anos, agora podiam ser feitas em dias. Mas pouco mudou. As pessoas que analisavam os dados ainda tinham como base o paradigma analógico de presumir que os bancos de dados tinham objetivos únicos com os quais os valores estavam relacionados. Nossos próprios processos perpetuaram esse preconceito. Por mais importante que a digitalização tenha sido ao permitir a coleta de grande quantidade de dados, a mera existência dos computadores não fazia os grandes dados surgirem.

Não existe um bom termo para explicar o que está acontecendo agora, mas um que ajuda a definir as mudanças seria *datafi cação,* conceito que apresentamos no Capítulo 5. Ele se refere à coleta de informações de tudo o que existe – inclusive informações que nunca foram pensadas como tal, como a localização de uma pessoa, as vibrações de um motor ou o estresse estrutural numa ponte – e à transformação disso em dados que possam ser quantificados. Esse conceito nos permite usarmos as informações de novas maneiras, como na análise previsiva: detectar que um motor está prestes a quebrar com base no calor ou vibrações que ele gera. Como resultado, revelamos o valor latente e implícito das informações.

Há uma caça ao tesouro em andamento, motivada pelas ideias a serem extraídas dos dados e pelo valor adormecido que pode ser despertado por uma mudança de causalidade para correlação. Mas não há apenas um tesouro. Cada banco de dados provavelmente tem um valor intrínseco oculto, e há uma corrida para descobri-lo e captá-lo.

O big data altera a natureza dos negócios, dos mercados e da sociedade, como descrevemos nos Capítulos 6 e 7. No século XX, o valor mudou da infraestrutura física, como terras e fábricas, para valores intangíveis, como marcas e propriedade intelectual. Essa mudança agora se expande para os dados, que estão se tornando um importante bem corporativo, recurso econômico essencial e a base para novos modelos de negócios. É o petróleo da economia da informação. Apesar de os dados ainda não serem registrados em balanços corporativos, provavelmente é apenas uma questão de tempo.

Apesar de algumas técnicas de coleta de dados existirem há algum tempo, no passado, elas só estavam disponíveis para agências de espionagem, laboratórios de pesquisa e as maiores empresas do mundo. Afinal, o Walmart e a Capital One foram pioneiras no uso do big data no varejo e, assim, mudaram seus ramos de atuação. Hoje muitas dessas ferramentas foram democratizadas (mas não os dados em si).

O efeito nas pessoas talvez seja a maior surpresa. O domínio de áreas específicas importa menos num mundo em que a probabilidade e a correlação são soberanas. No filme *O homem que mudou o jogo,* olheiros de beisebol foram superados por estatísticos

quando o instinto deu lugar à análise sofisticada. Do mesmo modo, os especialistas não desaparecerão, mas terão de se contentar com o que diz a análise do big data. Será necessário um ajuste nas ideias tradicionais de gerenciamento, tomada de decisões, recursos humanos e educação.

A maioria das nossas instituições foi criada sob a suposição de que as decisões humanas são tomadas com base em informações pequenas, exatas e causais. Mas a situação muda quando os dados são grandes, podem ser rapidamente processados e toleram a imprecisão. Mais do que isso, e por conta da quantidade de dados, as decisões serão tomadas mais por máquinas, não por seres humanos. Analisamos o lado negro do big data no Capítulo 8. A sociedade tem milênios de experiência na compreensão e no exame do comportamento humano. Mas como regulamos um algoritmo? Nos primórdios da computação, os legisladores perceberam como a tecnologia podia ser usada para acabar com a privacidade. Desde então, a sociedade conta com leis que protegem informações pessoais, mas que, na era do big data, são inúteis Linhas Maginot. Pessoas estão dispostas a compartilhar informações on-line – característica essencial dos serviços virtuais, não uma vulnerabilidade a ser prevenida. Enquanto isso, o perigo para nós, como pessoas, passou da privacidade para a probabilidade: algoritmos preverão a probabilidade de que pessoas tenham um ataque cardíaco (e paguem mais por planos de saúde), de não conseguirem pagar a hipoteca (e não obterem um empréstimo) ou cometerem um crime (e talvez serem presas preventivamente). Isso leva a questões éticas quanto ao papel do livre-arbítrio em comparação com a ditadura dos dados. A pessoa deve se sobrepor ao big data, mesmo que as estatísticas argumentem o contrário? Assim como as prensas abriram caminho para as leis que garantiam a liberdade de expressão – que não existiam antes porque havia pouca expressão para se garantir –, a era do big data exigirá novas regras para salvaguardar a santidade das pessoas.

De certo modo, a maneira como controlamos e lidamos com os dados terá de mudar. Entramos num mundo de constantes previsões com base em dados, no qual talvez não sejamos capazes de explicar os motivos de nossas decisões. O que significaria um médico não poder justificar uma intervenção sem o consentimento do paciente, uma vez que ele pode contar com um diagnóstico com base no big data? O padrão jurídico da "causa provável" precisará mudar para "causa *probabilística*"? Em caso afirmativo, quais são as implicações para a liberdade e a dignidade humanas?

São necessários novos princípios na era do big data, sobre os quais falaremos no Capítulo 9. Eles se baseiam em valores desenvolvidos e sacralizados para o mundo dos pequenos dados. Não se trata apenas de renovar antigas regras para novas circunstâncias, e sim de reconhecer a necessidade também de novos princípios.

Os benefícios da sociedade serão muitos, à medida que o big data se torne parte da solução de problemas mundiais, como a mudança climática, a erradicação de doenças e o estímulo ao bom governo e desenvolvimento econômico. Mas a era do big data também nos desafia a nos prepararmos melhor para a forma como o aproveitamento da tecnologia alterará nossas instituições e a nós mesmos.

O big data marca um importante passo na busca da humanidade por quantificar e compreender o mundo. Vários elementos que não podiam ser medidos, armazenados, analisados e compartilhados antes agora fazem parte de bancos de dados. O uso de grandes quantidades de dados, em vez de poucos, e o privilégio de mais dados e menos

exatidão abrem as portas para novas formas de compreensão, o que leva a sociedade a abandonar sua preferência pela causalidade e, de várias maneiras, a se aproveitar da correlação.

O ideal da identificação de mecanismos causais é uma ilusão autocongratulatória; o big data reverte isso. Mas novamente estamos num impasse histórico no qual "Deus está morto". Isto é, as certezas nas quais acreditávamos estão mudando mais uma vez. Desta vez, porém, estão sendo ironicamente substituídas por provas melhores. Que papel resta para a intuição, a fé, a incerteza, a ação contrária à prova e o aprendizado com a experiência? À medida que o mundo muda de causalidade para correlação, como pragmaticamente podemos avançar sem minar as fundações da sociedade, humanidade e progresso com base na razão? Este livro pretende explicar em que ponto estamos, analisar como chegamos até aqui e oferecer um guia urgentemente necessário quanto aos benefícios e perigos que nos aguardam.

Capítulo 2

Mais

Big data tem a ver com a percepção e compreensão de relações entre informações que, até recentemente, tínhamos dificuldade para entender. O especialista em big data da IBM, Jeff Jonas, diz que você precisa deixar que os dados "falem com você". De certo modo, parece trivial. Os seres humanos há muito tempo usam dados para aprender mais sobre o mundo, seja no sentido informal das observações diárias ou, principalmente nos últimos séculos, no sentido formal de unidades quantificadas que podem ser manipuladas por potentes algoritmos.

A era digital pode ter facilitado e acelerado o processamento de dados ao calcular milhões de números num segundo. Mas quando nos referimos a dados que falam, estamos nos referindo a algo maior – e diferente. Como mencionamos no Capítulo 1, o big data relaciona-se com três importantes mudanças de mentalidade interligadas que, portanto, se reforçam. A primeira é a capacidade de analisar grandes quantidades de dados sobre um tema em vez de ser obrigado a contar com conjuntos menores. A segunda é a disposição de aceitar a real confusão dos dados em vez de privilegiar a exatidão. A terceira é o maior respeito por correlações do que pela contínua busca pela causalidade elusiva. Este capítulo analisa a primeira destas mudanças: o uso de todos os dados em vez de apenas alguns.

O desafio de processar grandes quantidades de dados com precisão nos acompanha há algum tempo. Ao longo de boa parte da história, coletamos poucos dados porque os instrumentos para coletar, organizar, armazenar e analisar eram ruins. Peneirávamos as informações ao mínimo, de modo que pudéssemos analisá-las com mais facilidade. Era uma forma inconsciente de autocensura: tratávamos a dificuldade de interagir com os dados como uma infeliz realidade em vez de vê-los como eram, um limite artificial imposto à tecnologia da época. Hoje o ambiente técnico girou 179 graus. Ainda há, e sempre haverá, um limite de quantos dados podemos gerenciar, mas este limite é bem menor do que já foi e diminuirá ainda mais com o tempo.

De certo modo, ainda não podemos valorizar totalmente essa nova liberdade de coletar e usar grandes quantidades de dados. A maioria de nossas experiências e o projeto de nossas instituições presumem que a disponibilidade de informações é limitada. Achávamos que podíamos coletar só um pouco de informação e geralmente foi o que fizemos. Tornou-se autorrealizável. Até mesmo desenvolvemos técnicas elaboradas para usar o menor número de dados possível. Um objetivo da estatística, afinal, é confirmar as melhores descobertas com a menor quantidade de dados possível. Na verdade, codificamos a prática de reduzir a quantidade de informações que usamos em leis, processos e estruturas de incentivo. Para entender o que a mudança do big data significa, a história começa com uma olhada no passado.

Só recentemente, empresas privadas, e hoje em dia até mesmo pessoas, foram capazes de coletar e analisar informações numa escala massiva. No passado, essa tarefa cabia a instituições mais poderosas, como a igreja e o Estado, que em vários lugares eram uma só instituição. O registro de dados mais antigo remonta a 8000 a.C., quando comerciantes sumérios usaram continhas de barro para denotar os bens comercializados. A contagem em larga escala, contudo, era jurisdição do Estado. Ao longo de milênios, os governos tentaram manter o controle da população coletando informações.

Pense no censo. Dizem que os antigos egípcios realizaram censos, assim como os chineses. Eles são mencionados no Velho Testamento, e o Novo Testamento fala de um censo imposto por César Augusto – "para que todo mundo se alistasse" (Lucas 2:1) – e que levou José e Maria a Belém, onde Jesus nasceu. O livro *Domesday*, de 1086, um dos mais venerados tesouros britânicos, foi uma contagem abrangente - e sem precedentes em seu tempo - dos ingleses, suas terras e propriedades. Funcionários reais se espalharam pelo interior compilando informações para inscrevê-las no livro – que mais tarde ganhou o nome de "Domesday" ou "Doomsday" (fim dos dias), porque o processo era como o Juízo Final bíblico, no qual todas as vidas eram contadas.

Conduzir um censo é caro e demorado; o Rei Guilherme I, que encomendou o livro *Domesday*, não viveu para vê-lo completado. Mas a única alternativa a esta tarefa era ignorar a coleta de informações. E mesmo depois de todo o tempo e dinheiro investidos, as informações eram apenas aproximadas, já que os funcionários não podiam contar todos com perfeição. A própria palavra *censo* vem do termo latim *censere,* que significa "estimar".

Há mais de 300 anos, um comerciante britânico chamado John Graunt teve uma ideia inovadora. Graunt queria saber a população de Londres na época da peste negra. Em vez de contar todas as pessoas, ele inventou uma abordagem – que hoje chamamos de "estatística" – que lhe permitiu *estimar* a população. Sua abordagem era simplória, mas estabeleceu a ideia de que uma pessoa podia extrair conhecimento da população em geral a partir de uma pequena amostra. Mas a abordagem é importante. Graunt apenas aumentou a escala da amostragem.

Seu sistema foi celebrado, ainda que mais tarde tivéssemos descoberto que seu razoável resultado foi obtido por sorte. Durante gerações, a amostragem permaneceu cheia de falhas. Assim, para censos e tarefas semelhantes com a utilização de big data, a abordagem com base na força bruta de tentar contar tudo era a norma.

Como os censos eram complexos, caros e demorados, eram realizados raramente. Os antigos romanos, que se orgulhavam de uma população de centenas de milhares, faziam um censo a cada cinco anos. A Constituição dos Estados Unidos manda que se faça um censo a cada década, já que o crescimento do país se mede em milhões. Mas, no fim do século XIX, até mesmo isto se tornara problemático. Os dados suplantaram a capacidade do Census Bureau.

O censo de 1880 demorou extraordinários oito anos para ser completado. A informação ficava obsoleta antes mesmo de se tornar disponível. Pior, as autoridades previam que o censo de 1890 levaria 13 anos para ser tabulado – algo ridículo, sem falar que era uma violação da Constituição. Mas como os impostos e a representação legislativa tinham por base a população, uma contagem correta e no tempo certo era essencial.

O problema que o Census Bureau dos Estados Unidos enfrentava era semelhante ao de cientistas e empresários do novo milênio, quando ficou claro que estávamos nos afogando em dados: a quantidade de informações coletadas emperrava os instrumentos usados para processá-la, e eram necessárias novas técnicas. Nos anos 1880, a situação era tão crítica que o Census Bureau contratou Herman Hollerith, inventor americano, para usar sua ideia de cartões perfurados e máquinas de tabulação para o censo de 1890.

Com muito esforço, ele conseguiu diminuir o tempo de tabulação de oito anos para menos de um. Foi um feito incrível, que marcou o início do processamento automatizado de dados (e serviu de base para o que mais tarde se tornou a IBM), mas, como método de coleta e análise de grandes dados, ainda era muito caro. Afinal, todas as pessoas nos Estados Unidos tinham de preencher um formulário, e as informações tinham de ser transferidas para um cartão com furos, que por sua vez era usado na tabulação. Com métodos tão custosos, era difícil imaginar qualquer censo feito em menos de uma década, intervalo inútil para uma nação que crescia exponencialmente.

Ainda há uma tensão: usar todos os dados ou apenas um pouco? Conseguir todos os dados sobre o que estiver sendo medido é, com certeza, o caminho mais sensato, apenas nem sempre é prático em vasta escala. Mas como escolher a amostragem? Alguns argumentavam que criar uma amostragem que representasse o todo seria a melhor maneira de seguir adiante. Mas em 1934, Jerzy Neyman, estatístico polonês, demonstrou que essa abordagem levava a erros enormes. O segredo para evitá-los era contar com a aleatoriedade na escolha a amostra.

Os estatísticos demonstraram que a precisão da amostragem melhora dramaticamente com a aleatoriedade, e não com o aumento do tamanho da amostra. Na verdade, e apesar de surpreendente, uma amostra aleatória de 1.100 pessoas com uma pergunta binária (sim ou não, com a mesma probabilidade) é incrivelmente representativa da população em geral. Em 19 entre 20 casos, há uma margem de erro de 3%, independentemente se a população for de 100 mil ou 100 milhões. A explicação é uma questão matemática complexa, mas, em resumo: depois de certo ponto, à medida que os números aumentam, a quantidade marginal de novas informações que obtemos da observação diminui.

O fato de a aleatoriedade superar o tamanho da amostra foi um insight incrível. Ela abriu caminho para uma nova abordagem na coleta de informações. Dados que usavam amostragens aleatórias podiam ser coletados a um preço baixo e mesmo assim extrapolavam o todo com alta precisão. Como resultado, governos podiam realizar versões menores de censos usando amostragens aleatórias anuais em vez de uma a cada década. E foi o que fizeram. O Census Bureau dos Estados Unidos, por exemplo, conduz mais de 200 pesquisas econômicas e demográficas por ano com base em amostragens, além do censo decenal, que tenta contar todos. A amostragem era a solução para o problema da profusão de informações, quando era difícil coletar e analisar dados.

As aplicações desse novo método rapidamente foram além do setor público e dos censos. Em essência, a amostragem aleatória reduz problemas de grandes dados a problemas de dados mais gerenciáveis. No mundo dos negócios, ele foi usado para medir a quantidade de produção – tornando as melhorias mais fáceis e baratas. O controle de qualidade anteriormente exigia que a análise de todos os produtos que saíam da esteira; agora, uma amostragem aleatória de um lote bastava. Do mesmo modo, o novo

método antecipou as pesquisas de consumo e políticas e transformou grande parte do que costumávamos chamar de disciplinas humanas em *ciências* sociais.

A amostragem aleatória foi um sucesso e é a base da medição moderna em escala. Mas é apenas um atalho, a segunda melhor alternativa para coletar e analisar todo um banco de dados. Mas há vários pontos fracos. Sua precisão depende de garantir a aleatoriedade na coleta dos dados, mas essa aleatoriedade é difícil. Tendências sistemáticas na maneira como os dados são coletados podem levar a graves erros.

Esses problemas se repetem em pesquisas eleitorais por telefones fixos. A amostragem é tendenciosa contra pessoas que só usam celulares (mais jovens e liberais), como apontou o estatístico Nate Silver, o que resultou em previsões eleitorais incorretas. Na eleição presidencial de 2008 entre Barack Obama e John McCain, os grandes institutos de pesquisa Gallup, Pew e ABC/Washington Post descobriram diferenças de um a três pontos percentuais quando incluíam ou ignoravam usuários de celulares – uma margem grande considerando a acirrada eleição.

Outro problema é que as amostragens aleatórias não são boas para incluir subcategorias, já que dividir os resultados em grupos menores aumenta a possibilidade de previsões erradas. É fácil entender por quê. Vamos supor que você pesquise uma amostra aleatória de mil pessoas sobre suas intenções de voto. Se sua amostragem for suficientemente aleatória, a probabilidade é que a intenção de voto de toda a população esteja dentro de uma margem de erro de 3%. Mas e se os 3% não forem precisos o bastante? Ou e se você quiser dividir o grupo em grupos menores, por gênero, localização ou renda?

E se você quiser combinar esses subgrupos para obter um nicho da população? Numa amostra de mil pessoas, um subgrupo chamado "eleitoras influentes da região nordeste" teria bem menos de 100 pessoas. Usar algumas dezenas de observações para prever as intenções de voto de *todas* as eleitoras influentes do nordeste seria imperfeito mesmo com a melhor das aleatoriedades. E pequenas tendências em amostras gerais tornarão os erros mais evidentes em subgrupos.

Assim, a amostragem rapidamente deixa de ser útil quando você quer analisar com mais profundidade alguma intrigante subcategoria de dados. O que funciona no nível macro não funciona no micro. A amostragem é como uma fotografia analógica. Parece boa a distância, mas, quando você se aproxima, ao ampliar algum detalhe, ela fica embaçada.

A amostragem também requer planejamento e execução cuidadosos. Geralmente não se pode "fazer" novas perguntas referentes a dados de amostragem se elas não tiverem sido analisadas a princípio. Então, embora o atalho seja útil, o problema é que, bem, trata-se de um atalho. Por ser uma amostra, em vez de um dado generalizado, faltam aos dados certa extensão e maleabilidade, portanto os mesmos dados podem ser reanalisados de uma forma completamente nova, diferente do objetivo para o qual foram originalmente coletados.

Pense no caso da análise de DNA. O custo de sequenciar o genoma de uma pessoa era de aproximadamente US$1 mil em 2012, próximo de uma técnica de massa que pode ser realizada em larga escala. Como resultado, um novo ramo de sequenciamento de genes está prosperando. Desde 2007, a empresa 23andMe, do Vale do Silício, analisa o DNA das pessoas por apenas US$200. Sua técnica pode revelar características no

código genético que podem tornar as pessoas mais suscetíveis a doenças como câncer de mama ou problemas cardíacos. Ao agregar o DNA de seus clientes e informações de saúde, a 23andMe espera obter novos dados.

Mas há um problema. A empresa sequencia apenas uma pequena porção do código genético da pessoa: marcadores que indicam pontos fracos genéticos em particular. Enquanto isso, bilhões de pares de DNA permanecem sem sequenciamento. A 23andMe, portanto, só pode responder a perguntas sobre os marcadores que considera. Sempre que um novo marcador é descoberto, o DNA da pessoa (ou melhor, a parte relevante dele) tem de ser ressequenciado. Trabalhar com um subgrupo em vez do todo é o problema: a empresa pode encontrar o que procura de forma mais rápida e barata, mas não pode responder a perguntas que não considerou antes.

O lendário executivo-chefe da Apple, Steve Jobs, assumiu uma abordagem totalmente diferente em sua luta contra o câncer. Ele se tornou uma das primeiras pessoas no mundo a ter todo o DNA sequenciado, bem como o DNA do tumor. Para isso, ele pagou quase US$1 milhão – centenas de vezes mais do que a 23andMe cobra. Ele, por sua vez, não recebeu uma amostra, um mero grupo de marcadores, e sim um arquivo com todo seu código genético.

Ao escolher o medicamento para um paciente comum de câncer, os médicos têm de esperar que o DNA do paciente seja suficientemente semelhante ao DNA dos pacientes que participaram das experiências com o remédio. Mas a equipe de médicos de Steve Jobs podia selecionar terapias de acordo com seu bom funcionamento em relação à composição genética do paciente. Sempre que um tratamento perdia a eficiência porque o câncer mudava ou se adaptava a ele, os médicos podiam trocar o medicamento – "pulando de uma vitória-régia para outra", como Jobs mencionou. "Serei o primeiro a superar um câncer deste tipo ou um dos últimos a morrer por causa dele", disse. Apesar de sua previsão infelizmente não se concretizar, o método – a obtenção de todos os dados, e não apenas parte deles – lhe rendeu anos de vida a mais.

DE ALGUNS PARA TODOS

A amostragem é resultado de uma era de limites no processamento de informações, quando as pessoas mediam o número, mas faltavam as ferramentas para analisar os dados coletados. Como resultado, a amostragem é também um vestígio desta era. Os problemas de contagem e tabulação não existem mais da mesma forma. Sensores, GPSs, internet e Twitter coletam dados passivamente; os computadores podem lidar com os números com uma facilidade cada vez maior.

O conceito de amostragem não faz mais tanto sentido quando podemos analisar grandes quantidades de dados. As ferramentas técnicas para lidar com os dados já mudaram drasticamente, mas nossos métodos e mentalidades têm se adaptado mais lentamente.

A amostragem vem com um custo há muito tempo conhecido, mas deixado de lado, e que ignora os detalhes. Em alguns casos, não há outra maneira a não ser fazer uma amostragem. Em muitas áreas, contudo, está havendo uma mudança - em vez de se coletarem alguns dados, coleta-se o máximo de dados e, se possível, todos os dados: $N1 = 1$*todo.*

Como vimos, usar N1 = 1todo significa que podemos nos aprofundar nos dados; as amostragens não conseguem fazê-lo tão bem. Em segundo lugar, lembramos que, no exemplo anterior de amostragem, tínhamos uma margem de erro de apenas 3% quando extrapolamos para toda a população. Para algumas situações, não há problema nesta margem de erro. Mas você ignora os detalhes, a granularidade, a capacidade de analisar mais de perto certos subgrupos. Uma distribuição normal é, bem, normal. Em geral, o que há de mais interessante na vida é encontrado em lugares onde as amostragens falham.

Assim, a Google Flu Trends não conta com apenas uma pequena amostragem, e sim bilhões de buscas na internet nos Estados Unidos. Com todos esses dados, em vez de uma amostra, a empresa melhora a análise a ponto de prever a disseminação da gripe numa cidade em particular em vez de num estado ou no país todo. Oren Etzioni, da Farecast, inicialmente usou 12 mil pontos de dados, uma amostragem, e deu certo. Mas à medida que Etzioni acrescentou mais dados, a qualidade das previsões aumentou. Por fim, a Farecast usou os registros de voos domésticos da maioria das rotas por um ano inteiro. "São dados temporais que você reúne ao longo do tempo e, assim, compreende mais e mais os padrões", diz Etzioni.

Então frequentemente consideraremos bom deixar de lado o atalho da amostragem aleatória em nome de dados mais amplos, algo que requer muito poder de processamento e armazenagem e ferramentas de ponta para analisar os dados, além de maneiras fáceis e baratas de coletar dados. No passado, cada um desses elementos era um problema. Mas agora o custo e complexidade de todas as peças do quebra-cabeça diminuíram drasticamente. O que antes era uma previsão apenas das grandes empresas hoje está disponível para muitos.

O uso de todos os dados possibilita a localização de conexões e detalhes que, de outro modo, se perdem na vastidão das informações. A detecção de uma fraude de cartão de crédito, por exemplo, funciona pela busca de irregularidades, e a melhor maneira de encontrá-las é analisando todos os dados em vez de uma amostragem. As irregularidades são as informações mais interessantes, e você só consegue identificá-las em comparação com a massa de transações normais. É um problema do big data. Como as transações com cartão de crédito são instantâneas, a análise geralmente tem de ser feita em tempo real também.

A Xoom é uma empresa especializada em transferências financeiras internacionais e está associada a grandes empresas de big data. Ela analisa todos os dados relacionados com as transações com as quais lida. O sistema disparou um alarme em 2011 ao notar uma quantidade de transações de Discover Card ligeiramente maior que a média em New Jersey. "Havia um padrão onde não deveria haver", explicou John Kunze, executivo-chefe da Xoom. Por si só, todas as transações pareciam legítimas, mas vinham de um grupo criminoso. A única maneira de localizar a irregularidade era examinando todos os dados – a amostragem a teria ignorado.

Usar todos os dados não precisa ser uma grande tarefa. O big data não é necessariamente grande em termos absolutos, apesar de geralmente ser. A Google Flu Trends faz sua previsão com centenas de milhões de modelos matemáticos que usam bilhões de pontos de dados. A sequência completa do genoma humano tem três bilhões de pares-base. Mas o número absoluto de pontos de dados, o tamanho do banco de dados,

não é o que torna esses exemplos "big data". O que os classifica como "big data" é o fato de, em vez de usarem um atalho de uma amostragem aleatória, tanto a Flu Trends quanto os médicos de Steve Jobs usaram todos os dados disponíveis.

A descoberta de partidas arranjadas no esporte nacional do Japão, o sumô, é um bom exemplo de por que o uso do N1 = 1todo não precisa ser algo grandioso. Partidas arranjadas eram uma acusação constante contra o esporte dos imperadores e sempre negadas. Steven Levitt, economista da University of Chicago, procurou por sinais de corrupção nos registros de mais de uma década de lutas – todas. Num maravilhoso trabalho publicado na *American Economic Review* e no livro *Freakonomics,* ele e seu colega descreveram a utilidade da avaliação de tantos dados.

Eles analisaram 11 anos de embates de sumô, mais de 64 mil lutas, tentando encontrar irregularidades. E encontraram o que procuravam. Havia mesmo lutas arranjadas, mas não as que as pessoas suspeitavam. Em vez das finais de campeonatos, que podiam ou não ser armadas, os dados mostraram que algo estranho estava acontecendo durante partidas simples de fim de torneios. Parece que havia pouco em jogo, uma vez que os lutadores não tinham chance de se sagrarem campeões.

Mas uma peculiaridade do sumô é que os lutadores precisam de vitórias em torneios de 15 lutas para manter o ranking e o faturamento, o que às vezes gera uma assimetria de interesses, quando um lutador com uma marca de 7-7 enfrenta um oponente de 8-6 ou melhor. O resultado significa mais para o primeiro lutador e quase nada para o segundo. Nesses casos, os dados revelaram que o lutador que mais precisa da vitória provavelmente vence.

Os lutadores que precisavam da vitória lutavam com mais determinação? Talvez. Mas os dados sugeriam que algo mais estava acontecendo. Os lutadores que tinham mais em jogo ganhavam cerca de 25% a mais que o normal. É difícil atribuir tamanha discrepância somente à adrenalina. Ao analisar melhor os dados, eles mostraram que da vez seguinte que os mesmos dois lutadores disputavam, o perdedor da luta anterior tinha maior probabilidade de ganhar do que quando disputavam lutas posteriores. Assim, a primeira vitória parecia um "presente" de um competidor para o outro, já que o que vai sempre volta no restrito mundo do sumô.

Essa informação sempre foi aparente e clara. Mas a amostragem aleatória das lutas talvez tenha fracassado em revelá-la. Mesmo contando com a estatística básica, sem saber pelo que procurar, ninguém teria ideia do tamanho da amostragem a ser usada. Por outro lado, Levitt e seus colegas descobriram esse fato ao usar um grupo maior de dados – tentando examinar todo o universo das lutas. Uma investigação com big data é quase como uma pescaria: a princípio, não se sabe ao certo nem *se* alguém pegará algo nem *o que* alguém pode pegar.

Os dados não precisam ser terabytes. No caso do sumô, todos os dados tinham menos bits que uma fotografia digital típica de hoje em dia. Mas, como análise de big data, parecia mais que uma amostragem aleatória. Quando falamos em big data, mencionamos "grande" em termos relativos, não absolutos: relativamente ao amplo conjunto de dados.

Durante muito tempo, a amostragem aleatória foi um bom atalho, que possibilitou a análise de problemas de big data numa era pré-digital. Mas tanto quanto na conversão de uma imagem digital ou de uma música num arquivo menor, a informação se perde

na amostragem. Um banco de dados completo (ou perto disso) dá muito mais liberdade para explorar, analisar os dados de ângulos distintos e examinar mais de perto certos aspectos.

Uma boa analogia pode ser a câmera Lytro, que capta não apenas um plano de luz, como as câmeras convencionais, e sim raios de todo o campo de luz, cerca de 11 milhões deles. Mais tarde, os fotógrafos podem decidir quais elementos da imagem focar no arquivo digital. Não há necessidade de foco original, uma vez que a coleta de todas as informações possibilita que isso seja feito depois. Como os raios de todo o campo de luz estão incluídos, os dados são completos, N1 = 1tudo. Como resultado, a informação é mais "reutilizável" que imagens comuns, nas quais o fotógrafo tem de decidir o que focar antes de apertar o botão.

Do mesmo modo, como se baseia em toda a informação, ou pelo menos no máximo possível, o big data permite que analisemos detalhes ou exploremos novas análises sem o risco sair do foco. Podemos testar novas hipóteses em vários graus de granularidade. É essa característica que nos permite ver o arranjo nas lutas de sumô, acompanhar a disseminação do vírus da gripe por região e combater o câncer que atinge uma parte precisa do DNA da pessoa. Ela permite que trabalhemos com um incrível nível de clareza.

Para deixar claro, nem sempre é necessário usar todos os dados em vez de uma amostragem. Ainda vivemos num mundo de recursos limitados. Mas faz sentido usar todos os dados disponíveis em um número de casos cada vez maior, e, hoje em dia, isso é possível em situações em que antes não era.

Uma das áreas mais atingidas pela N1 = 1tudo é a ciência social, que perdeu o monopólio na análise de dados sociais empíricos, já que a análise do big data substitui os especialistas do passado. As disciplinas das ciências sociais contam imensamente com estudos por amostragem e questionários. Mas quando os dados são coletados passivamente, enquanto as pessoas continuam agindo da mesma forma, a velha tendência associada às amostragens e questionários desaparece. Podemos agora coletar informações que não tínhamos antes, seja sobre relações reveladas por meio de ligações de telefones celulares ou sentimentos expressos por tweets. E o mais importante: a necessidade da amostragem desaparece.

Albert-László Barabási, uma das principais autoridades mundiais da ciência da teoria de rede, queria estudar as interações entre as pessoas em escala planetária. Assim, ele e seus colegas examinaram logs anônimos de ligações celulares de uma operadora que servia cerca de um quinto da população de determinado país europeu – todos os logs de um período de quatro meses. Foi a primeira análise de rede em nível social, com um banco de dados que tinha por espírito o N1 = 1tudo. O trabalho numa escala tão grande e a análise de todas as chamadas entre milhões de pessoas durante certo período geraram novas descobertas que provavelmente não seriam reveladas de outro modo.

Curiosamente, e em contraste com estudos menores, a equipe descobriu que, ao remover uma pessoa de uma rede com várias associações na comunidade, a rede social restante se degrada, mas não desaparece. Quando, por outro lado, pessoas com ligações externas, fora de sua comunidade imediata, são excluídas da rede, a rede social repentinamente se desintegra, como se sua estrutura se dobrasse. Foi um resultado importante, mas, de certo modo, inesperado. Quem pensaria que pessoas com muitos

amigos são menos importantes para a estabilidade da estrutura da rede que aqueles com laços com pessoas mais distantes? Isso sugere que há uma recompensa na diversidade dentro de um grupo e numa sociedade.

Tendemos a pensar na amostragem estatística como uma rocha imutável, como os princípios da geometria ou as leis da gravidade. Mas o conceito tem menos de um século e foi desenvolvido para resolver um problema específico, num momento específico e sob limites tecnológicos específicos. Esses limites já não existem na mesma medida. Usar uma amostragem aleatória da era do big data é como andar a cavalo na era dos carros. Ainda podemos usar amostragens em certos contextos, mas não precisa ser – nem será – a maneira predominante de análise dos grandes bancos de dados. Cada vez mais seremos capazes de recorrer a tudo.

Capítulo 3

Confusão

É possível usar todos os dados disponíveis num número cada vez maior de contextos. Mas há custos. O aumento da quantidade de dados abre as portas para a inexatidão. Para ser claro, números errados e informações corrompidas sempre ocorreram em bancos de dados. Mas o ponto sempre foi tratá-los como um problema e resolvê-lo, em parte porque era possível. O que nunca quisemos fazer foi considerá-los inevitáveis e aprender a conviver com eles. Esta é uma das mudanças fundamentais da passagem para o big data.

Num mundo de dados pequenos, reduzir os erros e garantir a alta qualidade dos dados era um impulso natural e essencial. Como coletávamos poucas informações, garantíamos que os números registrados fossem os mais precisos possíveis. Gerações de cientistas aperfeiçoaram seus instrumentos para tornar suas medições mais e mais exatas, fosse para determinar a posição de um corpo celeste ou o tamanho de objetos sob um microscópio. Num mundo de amostragem, a obsessão pela exatidão era ainda mais crítica. Analisar apenas uma limitada quantidade de pontos de dados significa que os erros podem ser ampliados, o que potencialmente reduz a precisão do resultado final.

Durante boa parte da história, os maiores feitos da humanidade surgiram da conquista do mundo por meio da medição. A busca pela exatidão começou na Europa em meados do século XIII, quando astrônomos e eruditos elaboraram uma quantificação mais precisa do tempo e do espaço – "a medida da realidade", nas palavras do historiador Alfred Crosby.

Se alguém fosse capaz de medir um fenômeno, a crença implícita era a de que se podia compreendê-lo. Mais tarde, a medição foi relacionada com o método científico da observação e explicação: a capacidade de quantificar, registrar e apresentar resultados que podiam ser reproduzidos. "Medir é saber", anunciou Lord Kelvin. Essa máxima se tornou a base do poder. "Conhecimento é poder", ensinou Francis Bacon. Ao mesmo tempo, matemáticos, e o que mais tarde se tornaram atuários e contadores, desenvolveram métodos que possibilitaram a coleta, armazenamento e gerenciamento precisos dos dados.

A França do século XIX – na época a nação cientificamente mais evoluída do mundo – criou um sistema que definia com precisão unidades que mediam o espaço, o tempo e outros fatores e começou a convencer os demais países a adotarem o mesmo padrão. Chegou-se ao ponto de se criarem unidades internacionalmente aceitas para serem comparadas em tratados internacionais. Foi o auge da era das medições. Meio século mais tarde, nos anos 1920, as descobertas da mecânica quântica abalaram para sempre o sonho das medições abrangentes e perfeitas. Ainda assim, fora de um

círculo relativamente pequeno de físicos, a mentalidade da vocação para a medição perfeita da humanidade continuou entre engenheiros e outros cientistas. No mundo dos negócios, essa mentalidade até mesmo se expandiu, quando as ciências exatas, como a matemática e estatística, começaram a influenciar todos os ramos do comércio.

Mas em várias novas situações atuais a permissão da imprecisão – da confusão – pode ser uma característica positiva, e não uma falha. É uma permuta. O benefício de afrouxar os padrões de erros permitidos é a possibilidade de coletar mais dados. Não se trata apenas de "mais ser melhor que um pouco", e sim de "mais ser melhor que o melhor".

Há vários tipos de confusões com as quais é preciso lidar. O termo pode se referir apenas ao fato de a probabilidade de erros aumentar à medida que você acrescenta mais pontos de dados. Assim, aumentar a leitura de tensão de uma ponte por um fator de mil aumenta a chance de que algo dê errado. Mas você também pode aumentar a confusão ao combinar diferentes tipos de informações de diversas fontes, que nem sempre se alinham com perfeição. Usar, por exemplo, um software de reconhecimento de voz para caracterizar reclamações num call center e comparar os dados com o tempo que os operadores levam para lidar com as ligações podem gerar uma amostragem imperfeita, embora útil, da situação. Confusão também pode se referir à inconsistência da formatação, de acordo com a qual os dados precisam ser "limpos" antes de processados. Há várias maneiras de se referir à IBM, nota o especialista em big data, DJ Patil, de I.B.M. a T.J. Watson Labs e International Business Machines. E a confusão pode surgir quando extraímos ou processamos os dados, já que, desta forma, os alteramos e transformamos em outra coisa, como quando realizamos análises de sentimentos em mensagens do Twitter para prever o faturamento de um filme de Hollywood. A confusão em si é confusa.

Vamos supor que precisamos medir a temperatura num vinhedo. Se tivermos apenas um sensor de temperatura para todo o vinhedo, temos de garantir que a medição seja precisa e funcione o tempo todo: não se permite confusão. Por outro lado, se tivermos um sensor a cada 100 parreiras, podemos usar sensores mais baratos e menos sofisticados (desde que não gerem uma tendência sistemática). É provável que em alguns pontos uns sensores possam registrar dados incorretos, criando um banco de dados menos exato, ou "mais confuso" que os dados de um único e preciso sensor. Qualquer leitura pode ser incorreta, mas a soma de várias leituras gerará um resultado mais abrangente. Como este conjunto de dados é composto por mais pontos de dado, ele agrega maior valor, que supera a confusão.

Agora vamos supor que aumentemos a frequência de leitura dos sensores. Se fizermos uma medição por minuto, podemos ter certeza de que a sequência da chegada dos dados será perfeitamente cronológica. Mas se a mudarmos para 10 ou 100 leituras por segundo, a sequência talvez não seja tão precisa. À medida que a informação viaja pela rede, um registro pode se atrasar ou chegar fora da sequência ou pode se perder na enxurrada de dados. A informação será um pouquinho menos precisa, mas seu grande volume compensa a perda de exatidão.

No primeiro exemplo, sacrificamos a precisão de cada ponto de dado por abrangência e, como resultado, recebemos detalhes que, de outro modo, podíamos não ter visto. No segundo caso, abrimos mão da exatidão pela frequência e, como resultado,

vimos uma alteração que, de outro modo, teríamos ignorado. Apesar de talvez sermos capazes de suplantar os erros se usarmos o máximo de recursos – afinal, 30 mil transações por segundo ocorrem na Bolsa de Valores de Nova York, onde a sequência correta importa muito –, em muitos casos, é mais produtivo tolerar o erro que trabalhar no sentido de evitá-lo.

Por exemplo, podemos aceitar certo nível de confusão em troca da escala. Como afirma Forrester, consultor de tecnologia: "Às vezes, dois mais dois é igual a 3,9, e isto basta." Claro que os dados não podem ser totalmente incorretos, mas estamos dispostos a sacrificar um pouco da precisão em troca de conhecermos a tendência geral. O big data transforma os números em algo mais "probabilístico" que exato. É preciso muito para se acostumar com essa alteração, que tem seus próprios problemas e dos quais trataremos mais adiante. Por enquanto, vale a pena notar que geralmente precisamos aceitar a confusão quando aumentamos a escala.

Há quem veja uma mudança semelhante em termos da importância dos dados relativos a outras melhorias na computação. Todos sabem como o poder de processamento aumentou ao longo dos anos, de acordo com a Lei de Moore, que afirma que a quantidade de transistores num chip dobra a cada dois anos. Esta contínua melhoria tornou os computadores mais rápidos, e a memória, mais profusa. Poucos sabem que o desempenho dos algoritmos que impulsionam muitos de nossos sistemas também aumentou – em muitas áreas, um aumento maior que seus antecessores sujeitos à Lei de Moore, segundo com o Conselho de Ciência e Tecnologia da Presidência dos Estados Unidos. Muitos dos ganhos da sociedade com o big data, contudo, acontecem não tanto por causa de chips mais rápidos ou de melhores algoritmos, e sim pela existência de mais dados.

Algoritmos de xadrez, por exemplo, pouco mudaram nas últimas décadas, já que as regras do xadrez são conhecidas e bem limitadas. O motivo de os softwares de xadrez jogarem melhor hoje que no passado está no fato de jogarem melhor a etapa final. Isso acontece porque os sistemas são alimentados de mais dados. Na verdade, jogadas com seis ou menos peças no tabuleiro foram totalmente analisadas, e todos os movimentos possíveis (N1 = 1all) foram representados numa grande tabela que, quando descomprimida, tem mais de um terabyte de dados. Isso permite que os computadores joguem sem falhas nas importantes jogadas finais do xadrez. Nenhum humano jamais será capaz de vencer o sistema.

O nível em que mais dados vencem melhores algoritmos foi demonstrado na área do processamento da linguagem natural: a maneira como os computadores aprendem a analisar as palavras como as usamos no nosso discurso cotidiano. Por volta de 2000, os pesquisadores da Microsoft Michele Banko e Eric Brill estavam à procura de um método para melhorar a análise gramatical do programa Word. Eles não sabiam ao certo se seria mais útil dirigir seus esforços para a melhoria dos algoritmos existentes e encontrar novas técnicas ou acrescentar características mais sofisticadas. Antes de optar por um desses caminhos, eles decidiram ver o que acontecia quando acrescentavam mais dados aos métodos existentes. Muitos algoritmos de aprendizado contam com *corpus* de texto de um milhão de palavras ou menos. Banko e Brill pegaram quatro algoritmos comuns e os alimentaram com muitos mais dados: 10 milhões de palavras, depois 100 milhões e, por fim, 1 bilhão.

Os resultados foram incríveis. À medida que mais dados foram acrescentados, o desempenho dos quatro algoritmos aumentou drasticamente. Na verdade, um simples algoritmo, o de pior desempenho, com 500 mil palavras, teve um desempenho melhor que os demais, com 1 bilhão de palavras. Sua taxa de precisão passou de 75% para mais de 95%. Por outro lado, o algoritmo que trabalhava melhor com poucos dados teve um desempenho razoável com quantidades maiores e, como os demais, melhoraram muito, de uma precisão de 86% para 94%. "Esses resultados sugerem que talvez queiramos pensar na troca do tempo e dinheiro investidos no desenvolvimento de algoritmos pelo investimento no desenvolvimento de bancos de dados", escreveram Banko e Brill num de seus trabalhos sobre o assunto.

É assim que mais é melhor que menos. E às vezes mais é melhor que "mais inteligente". E quanto à confusão? Alguns anos depois que Banko e Brill exploraram os dados, pesquisadores da empresa rival, a Google, pensavam de forma semelhante – mas numa escala ainda maior. Em vez de testar algoritmos com um bilhão de palavras, usaram um trilhão. O objetivo da Google não era desenvolver um analisador de gramática, mas algo mais complexo: um tradutor.

As chamadas máquinas de tradução eram um sonho dos pioneiros da computação desde o início da indústria, nos anos 1940, quando os aparelhos eram compostos por tubos a vácuo e enchiam toda uma sala. A ideia ganhou urgência durante a Guerra Fria, quando os Estados Unidos captaram grandes quantidades de material escrito e falado dos russos, mas não tinha homens para traduzi-lo rapidamente.

A princípio, os cientistas optaram por uma combinação de regras gramaticais e dicionário bilíngue. Um computador da IBM traduziu 60 frases em russo para o inglês, em 1954, usando 250 pares de palavras do vocabulário da máquina e seis regras gramaticais. Os resultados foram promissores. "Mi pyeryedayem mislyi posryedstvom ryechyi" foi inserido no IBM 701 por meio de cartões furados e deles se extraiu "Transmitimos pensamentos por meio da fala". As 60 frases foram "bem traduzidas", de acordo com um *press release* da IBM, lançado para celebrar o acontecimento. O diretor do programa de pesquisa, Leon Dostert, da Georgetown University, previu que uma máquina de tradução seria "um fato" dentro de "cinco, talvez três anos".

Mas o sucesso inicial se revelou enganoso. Em 1966, um comitê de estudiosos das máquinas de tradução teve de admitir o fracasso. O problema era mais difícil do que pensavam. Ensinar os computadores a traduzir inclui não apenas regras, mas também exceções. A tradução não tem a ver apenas com memorização; mas com a escolha das palavras corretas diante de diversas alternativas. "Bonjour" se traduz como "bom dia"? Ou "passe bem" ou "olá" ou "oi"? A resposta é: depende.

No fim dos anos 1980, os pesquisadores da IBM tiveram outra ideia. Em vez de tentar alimentar um computador com regras linguísticas explícitas, juntamente com um dicionário, eles decidiram deixar que o computador usasse a probabilidade estatística para calcular que palavra ou expressão num idioma era mais apropriada em outro. Em 1990, o projeto Candide da IBM usou 10 anos de transcrições parlamentares canadenses publicadas em francês e inglês – cerca de três milhões de pares. Como eram documentos oficiais, as traduções eram de alta qualidade. E, para os padrões da época, a quantidade de dados era enorme. A tradução mecânica estatística, como a técnica veio a ser conhecida, transformou o desafio da tradução num grande problema

matemático. Parecia ter dado certo. De repente, a tradução computacional melhorou. Depois do sucesso desse salto conceitual, contudo, a IBM obteve apenas pequenas melhoras investindo muito dinheiro. Por fim, a empresa desistiu do projeto.

Menos de uma década mais tarde, em 2006, a Google entrou no ramo da tradução como parte da missão de "organizar as informações do mundo e torná-las universalmente acessíveis e úteis". Em vez de páginas bem traduzidas de texto em dois idiomas, a Google usou um banco de dados maior, mas também mais confuso: toda a internet global e mais. O sistema usava todas as traduções que encontrava a fim de treinar o computador. Ele utilizava dados de sites corporativos em vários idiomas, traduções idênticas de documentos oficiais e relatórios de corpos intergovernamentais, como as Nações Unidas e a União Europeia. Até mesmo traduções de livros do projeto de escaneamento de livros da Google foram incluídas. Enquanto o projeto Candide usava três milhões de sentenças cuidadosamente traduzidas, o sistema da Google utilizava bilhões de páginas de traduções de qualidade variável, de acordo com o chefe da Google Translate, Franz Josef Och, uma das principais autoridades do ramo. Seu banco de dados de um trilhão de palavras compreende 95 bilhões de frases em inglês, a despeito da duvidosa qualidade.

Embora os dados sejam confusos, o serviço da Google funciona bem. Suas traduções são mais precisas que as dos outros sistemas (ainda que extremamente imperfeitas). E são muito, muito mais profusas. Em meados de 2012, seu banco de dados cobria mais de 60 idiomas e podia até mesmo aceitar dados de voz em 14 idiomas para traduções fluidas. E como trata a linguagem como dados confusos sujeitos a probabilidades, ele pode até mesmo traduzir do hindu para o catalão, idiomas nos quais há poucas traduções diretas para desenvolver o sistema. Nesses casos, ele usa o inglês como ponte. O sistema é muito mais flexível que outras abordagens, já que pode acrescentar ou subtrair palavras de acordo com a frequência de uso.

O sistema de tradução da Google funciona bem não porque seja um algoritmo mais inteligente, mas porque seus criadores, assim como Banko e Brill, da Microsoft, o alimentaram com mais dados – e não apenas dados de alta qualidade. A Google foi capaz de usar um banco de dados *dezenas de milhares* de vezes maior que o do projeto Candide da IBM porque aceitava a confusão. O *corpus* de um trilhão de palavras lançado pela Google em 2006 foi compilado a partir dos despojos do conteúdo virtual – "dados selvagens", por assim dizer. Foi a partir desse "sistema de treinamento" que o projeto pôde calcular a probabilidade de que, por exemplo, uma palavra em inglês se siga a outra. É bem diferente do avô do ramo, o famoso Brown Corpus dos anos 1960, que totalizava um milhão de palavras em inglês. O uso de um conjunto de dados maior permitiu grandes avanços no processamento de idiomas naturais, nos quais programas de reconhecimento de voz e tradução computacional se baseiam. "Modelos simples com muitos dados são melhores que modelos mais elaborados com menos dados", escreveu o guru da inteligência artificial da Google, Peter Norvig, e seus colegas num trabalho intitulado "The Unreasonable Effectiveness of Data".

Como Norvig e seus coautores explicaram, a confusão era o segredo: "De certo modo, este banco de dados é um passo para trás em relação ao Brown Corpus: é tirado de páginas da internet sem filtro e, assim, contém sentenças incompletas, erros de ortografia, de gramática e todos os outros tipos de erros. Ele não é composto por

discursos cuidadosamente corretos. Mas o fato de ele ser um milhão de vezes maior que o Brown Corpus supera as desvantagens".

MAIS É MELHOR QUE MELHOR

Os analistas convencionais têm dificuldade para entender a confusão, uma vez que passaram toda a vida focados em evitá-la e eliminá-la. Eles trabalham duro para reduzir as taxas de erro ao coletar amostras e para testar as amostras à procura de tendências em potencial antes de anunciar os resultados. Eles usam diferentes métodos de redução de erro e se certificam de que as amostras sejam coletadas de acordo com um protocolo exato e por especialistas treinados. A implementação dessas estratégias é cara mesmo para pontos de dados limitados e impossível no caso do big data. Além de serem caras demais, seria improvável consistentemente alcançar os padrões de exatidão em grande escala. Nem mesmo a exclusão da interação humana resolveria o problema.

Entrar num mundo de big data exigirá que mudemos nossa mentalidade quanto ao mérito da exatidão. Aplicar a mentalidade de medição convencional ao mundo digital e conectado do século XXI é ignorar um ponto crucial. Como já mencionado, a obsessão pela exatidão remonta à era analógica e escassa de informações. Quando os dados são esparsos, todos os pontos de dados são essenciais e, assim, toma-se um cuidado maior para evitar que qualquer ponto influencie a análise.

Hoje não temos essa ânsia pela informação. Ao lidar com dados ainda mais abrangentes, que captam não apenas uma porção menor do fenômeno em questão e sim uma parte maior ou até mesmo sua totalidade, não precisamos mais nos preocupar tanto com a influência dos pontos de dados na análise. Em vez de buscar acabar com toda a inexatidão a um custo maior, consideramos a falta de precisão ao fazermos os cálculos.

Veja, por exemplo, a maneira como os sensores estão sendo usados em fábricas. Na refinaria Cherry Point, da BP, em Blaine, Washington, sensores sem fio estão instalados por toda parte, criando uma rede invisível que gera muitos dados em tempo real. O ambiente de calor intenso e máquinas elétricas às vezes distorce a leitura e gera dados confusos. Mas a enorme quantidade de informação gerada por sensores com ou sem fio compensam essas falhas. Apenas aumentar a frequência e quantidade de lugares de leitura com sensores pode valer a pena. Ao medir a tensão nos canos o tempo todo em vez de em certos intervalos, a BP descobriu que alguns tipos de óleo cru são mais corrosivos que outros – característica que não se podia vislumbrar e, portanto, não se podia neutralizar quando o banco de dados era menor.

Quando a quantidade de dados é enorme e de um tipo novo, em alguns casos a exatidão já não é o objetivo, desde que possamos descobrir a tendência geral. Passar para a larga escala altera não apenas as expectativas de precisão como também a habilidade prática de se alcançar a exatidão. Apesar de parecer contraproducente a princípio, tratar os dados como imperfeitos e imprecisos permite que façamos melhores previsões e entendamos melhor o mundo.

Vale a pena notar que a confusão não é inerente ao big data, mas uma função da imperfeição dos instrumentos que usamos para medir, registrar e analisar as informações. Se a tecnologia de algum modo se tornar perfeita, o problema da inexatidão desaparecerá. Mas enquanto a tecnologia for imperfeita, a confusão será uma realidade

prática com a qual deveremos lidar. E provavelmente ela nos acompanhará por muito tempo ainda. Grandes esforços para aumentar a precisão geralmente não fazem sentido em termos econômicos, já que o valor de ter quantidades maiores de dados é mais interessante. Assim como os estatísticos dos primórdios deixaram de lado o interesse pelas grandes amostragens em favor da aleatoriedade, podemos viver com um pouco de imprecisão em troca de mais dados.

O Billion Prices Project é um caso intrigante. Todos os meses, o U.S. Bureau of Labor Statistics (Departamento Americano de Estatísticas Trabalhistas) publica o índice de preço ao consumidor, usado para o cálculo da inflação. O número é fundamental para investidores e empresários. O Federal Reserve o leva em consideração quando eleva ou diminui as taxas de juros. As empresas baseiam os aumentos salariais na inflação. O governo federal o usa para indexar pagamentos, como os benefícios do seguro social e os juros que paga a certos títulos do tesouro.

Para obter esse número, o Bureau of Labor Statistics emprega centenas de funcionários que ligam, enviam faxes e visitam lojas e escritórios em 90 cidades do país, criando um relatório com 80 mil preços de tudo, desde tomates até tarifas de táxi. Esse cálculo custa ao país cerca de US$250 milhões por ano. Por essa quantia, os dados chegam de forma clara e organizada. Mas assim que são divulgados, os números já estão uma semana atrasados. Como ficou claro com a crise financeira de 2008, poucas semanas podem significar um atraso e tanto. Os tomadores de decisões precisam de acesso mais rápido aos números da inflação a fim de reagir melhor a eles, mas não conseguem obter os números com métodos convencionais com base em amostragens e na precisão dos preços.

Em reação a isso, dois economistas do MIT (Massachusetts Institute of Technology), Alberto Cavallo e Roberto Rigobon, inventaram uma alternativa por meio de um caminho muito mais confuso. Com o auxílio de um programa para vasculhar a internet, eles coletaram meio milhão de preços de produtos vendidos nos Estados Unidos todos os dias. A informação é confusa, e nem todos os pontos de dados coletados são facilmente comparáveis. Mas, ao combinar o big data com a análise inteligente, o projeto foi capaz de detectar um movimento deflacionário nos preços imediatamente depois que o Lehman Brothers pediu falência, em setembro de 2008, enquanto os que contavam com o dado oficial do governo tiveram de esperar até novembro.

O projeto do MIT deu origem a uma empresa chamada PriceStats, que os bancos e outros negócios utilizam para tomar decisões econômicas. Ela compila milhões de produtos vendidos por centenas de varejistas em mais de 70 países todos os dias. Claro que o número exige cuidadosa interpretação, mas é melhor que a estatística oficial ao indicar a tendência inflacionária. Como há mais preços e os números são obtidos em tempo real, esse número dá significativa vantagem aos tomadores de decisões. (O método também serve como avaliação externa da estatística governamental. A revista *The Economist,* por exemplo, não confia no cálculo da inflação na Argentina, por isso usa o número da PriceStats.)

A CONFUSÃO EM AÇÃO

Em muitas áreas da tecnologia e da sociedade, tendemos a mais dados e confusão que a menos dados e precisão. Pense no caso do conteúdo categorizado. Durante

séculos, os homens desenvolveram taxonomias e indexações a fim de armazenar e recuperar materiais. Esses sistemas hierárquicos sempre eram imperfeitos, como qualquer pessoa acostumada com a catalogação por meio de cartões de uma biblioteca é capaz de dolorosamente se lembrar, mas num universo de dados menores, eles funcionavam bem. Aumente a escala em várias ordens de magnitude, porém, e esses sistemas, que supõem a localização perfeita, se quebram. Em 2011, por exemplo, o site de compartilhamento de imagens Flickr tinha mais de seis bilhões de fotos de mais de 75 milhões de usuários. Tentar rotular cada foto de acordo com categorias pré-estabelecidas teria sido inútil. Ou será que haveria mesmo uma categoria chamada "gatos que se parecem com Hitler"?

Em vez disso, taxonomias limpas foram substituídas por mecanismos mais confusos, mas também mais flexíveis e adaptáveis num mundo que evolui e muda. Quando enviamos uma foto para o Flickr, nós a "categorizamos", isto é, determinamos algumas "etiquetas" e as usamos para organizar e vasculhar o material. As etiquetas são criadas e afixadas por pessoas de modo aleatório: não há categorias padronizadas e predefinidas, não há uma taxonomia existente à qual se submeter. Em vez disso, qualquer pessoa pode acrescentar novos rótulos. A "etiquetação" surgiu como padrão de fato para a classificação do conteúdo na internet, usada em sites de mídias sociais, como Twitter, blogs e assim por diante. Isso torna a vastidão do conteúdo da internet mais navegável – especialmente para conteúdos como imagens, vídeos e músicas, que não se baseiam em textos, de modo que os buscadores de palavras não funcionam.

Claro que alguns rótulos podem ser escritos de maneira errada, e esses erros geram imprecisão – não aos dados em si, mas à sua organização, o que afeta a mentalidade tradicional acostumada à exatidão. Mas em troca da confusão na organização de nossas fotos, ganhamos um universo muito mais rico de rótulos e, por extensão, um acesso mais vasto e profundo às nossas imagens. Podemos combinar categorias para filtrar imagens de uma maneira que não era possível antes. A imprecisão inerente à categorização tem a ver com aceitar a confusão natural do mundo. É um antídoto a sistemas mais precisos que tentam impor uma falta de esterilidade sobre a entropia da realidade, fingindo que tudo pode ser organizado em linhas e colunas. Há mais mistérios entre o céu e a terra do que supõe nossa vã filosofia.

Muitos dos sites mais populares da internet se gabam de sua admiração pela imprecisão em detrimento da precisão. Quando alguém vê um ícone do Twitter ou um botão de "curtir" do Facebook numa página da internet, eles mostram a quantidade de pessoas que neles clicaram. Quando os números são baixos, cada clique é mostrado, como em "63". Mas, à medida que os números aumentam, a quantidade mostrada é uma aproximação, como "4K". Não que o sistema não saiba o total; mas, à medida que a escala aumenta, o número exato perde importância. Além disso, a quantidade pode se alterar com tanta rapidez que um número específico estaria desatualizado assim que aparecesse na tela. Do mesmo modo, o Gmail da Google apresenta o momento das mensagens recentes com exatidão, como em "há 11 minutos", mas trata prazos mais longos com um simples "há 2 horas", assim como fazem o Facebook e alguns outros.

A indústria da inteligência corporativa e dos softwares analíticos há muito se baseia na promessa de dar aos clientes "uma versão única da verdade" – palavras comuns dos vendedores de tecnologia nestes ramos dos anos 2000. Os executivos usavam a frase

sem ironia. Alguns ainda a usam. Com isso, eles querem dizer que todos os que acessam os sistemas de tecnologia e informação de uma empresa podem usar os mesmos dados; que as equipes de marketing e de vendas não têm de disputar quem é o cliente certo ou a quantidade de vendas antes mesmo do início da reunião. Os interesses talvez fiquem mais alinhados se os fatos forem consistentes, pensava-se.

Mas a ideia de "uma versão única da verdade" está dando reviravolta. Estamos começando a perceber não só que talvez uma única versão da verdade seja impossível como também que sua busca é uma dispersão. Para aproveitar os benefícios dos dados em escala, temos de aceitar a confusão como parte da jornada, e não como algo que deveríamos tentar eliminar.

Vemos a inexatidão invadir até mesmo uma das áreas mais intolerantes à imprecisão: o projeto de bancos de dados. Sistemas de bancos de dados tradicionais requerem que eles sejam extraordinariamente estruturados e precisos. Os dados não são apenas armazenados; são divididos em "registros" que contêm campos. Cada campo abriga uma informação de determinado tipo e extensão. Se um campo numérico, por exemplo, tivesse sete dígitos, uma quantidade de 10 milhões ou mais não poderia ser registrada. Se alguém quisesse informar "não disponível" num campo de números de telefones, não era possível. A estrutura dos bancos de dados teria de ser alterada para acomodar tais entradas. Ainda enfrentamos essas restrições em computadores e *smartphones*, quando o software não aceita os dados que queremos acrescentar.

Indexações tradicionais também eram predefinidas e limitavam o que alguém podia procurar. O acréscimo de um novo índex tinha de ser feito do zero, o que tomava tempo. Bancos de dados convencionais, também chamados de relacionais, são criados para um mundo no qual os dados são esparsos e, assim, podem e serão tratados com precisão. É um mundo onde perguntas respondidas por meio de dados têm de ser simples desde o princípio, de modo que o banco de dados é projetado para respondê-las – e só a elas – com eficiência.

Mas essa visão do armazenamento e análise cada vez mais se opõe à realidade. Hoje temos enormes quantidades de dados de vários tipos e características que raramente se adaptam a categorias definidas e conhecidas desde o princípio. E as perguntas que queremos fazer geralmente surgem somente depois que coletamos e analisamos os dados de que dispomos.

Essas realidades têm gerado nossos projetos de bancos de dados, que rompem com os velhos princípios de gravação e campos predefinidos que refletem hierarquias claras de informação. A linguagem mais comum para acessar bancos de dados há muito tempo é a SQL, ou "linguagem de consulta estruturada". O próprio nome evoca sua rigidez. Mas a grande mudança nos anos recentes tem acontecido em direção a algo chamado noSQL, que não requer uma estrutura pré-determinada para funcionar. Ele aceita dados de vários tipos e tamanhos e permite que sejam vasculhados com sucesso. Em troca da permissão pela confusão estrutural, esses bancos de dados requerem mais recursos de processamento e armazenagem. Mas é uma troca que pode ser feita por conta dos custos cada vez menores de armazenagem e processamento.

Pat Helland, uma das principais autoridades mundiais em projeto de bancos de dados, descreve essa mudança fundamental num trabalho intitulado "If You Have Too Much Data, Then 'Good Enough' Is Good Enough" (Se você tem dados demais, então "bom

o bastante" basta). Depois de identificar alguns dos princípios fundamentais do projeto tradicional, desgastados pelos confusos dados de várias fontes e precisão, ele disserta sobre as consequências: "Não podemos mais fingir que vivemos num mundo limpo." O processamento do big data se reflete numa inevitável perda de informação - Helland chama isso de "compressão com perda". Mas o resultado mais rápido compensa. "Não há qualquer problema em obter respostas com perdas – frequentemente é disso que se precisa", conclui Helland.

O projeto de bancos de dados tradicionais promete consistentes resultados com o tempo. Se você perguntar qual é seu saldo bancário, por exemplo, você espera obter uma quantia exata. E se fizer a mesma pergunta segundos mais tarde, espera que o sistema lhe dê o mesmo resultado, supondo que nada tenha sido alterado. Mas à medida que a quantidade de dados cresce e o número de usuários que acessam o sistema aumenta, manter essa consistência se torna mais difícil.

Grandes bancos de dados não existem num só lugar; eles geralmente se dividem em vários hard drives e computadores. Para garantir a confiabilidade e a velocidade, um registro pode ser gravado em dois ou três lugares distintos. Se você atualiza o registro num lugar, os dados nos outros lugares não estarão mais corretos, até que você os atualize também. Apesar de os sistemas tradicionais terem um atraso até que as atualizações sejam realizadas, a praticidade se perde quando os dados estão mais distribuídos e o servidor recebe dezenas de milhares de perguntas por segundo. Aceitar a confusão é um tipo de solução.

Essa mudança é exemplificada pela popularidade do Hadoop, um rival *opensource* de fonte aberta do sistema MapReduce da Google, muito bom no processamento de grandes quantidades de dados. Ele divide os dados em quantidades menores e os distribui por outras máquinas. Ele espera que o hardware falhe, por isso cria redundância, e presume que os dados não sejam claros e ordenados – na verdade, presume que os dados sejam grandes demais para serem "limpos" antes do processamento. Enquanto a análise de dados típica exige uma operação chamada ETL (extração, transformação e carga) para transferir os dados para onde serão analisados, o Hadoop dispensa tudo isso. Ao contrário, ele despreza a extraordinária quantidade de dados que não podem ser transferidos e que devem ser analisados onde se encontram.

O resultado do Hadoop não é tão preciso quanto o dos bancos de dados relacionais: ele não pode ser usado para se lançar um foguete no espaço ou para dar detalhes de uma conta bancária. Mas para tarefas menos críticas, nas quais uma resposta ultraprecisa não é necessária, ele é muito mais rápido que as alternativas. Pense em tarefas como a segmentação de uma lista de consumidores para envio de uma campanha especial de marketing. Com o Hadoop, a empresa de cartões de crédito Visa foi capaz de reduzir o tempo de processamento de dois anos de registros, equivalentes a 73 bilhões de transações, de um mês para meros 13 minutos. Esse tipo de aceleração de processamento está transformando a indústria.

A experiência da ZestFinance, empresa fundada pelo ex-chefe do departamento de informações da Google, Douglas Merrill, ressalta o argumento. Sua tecnologia ajuda os financiadores a decidir conceder ou não empréstimos pequenos e de curto prazo a clientes com um registro de crédito supostamente ruim. Mas enquanto o registro de crédito tradicional se baseia em alguns fatores, como pagamentos anteriores em atraso,

a ZestFinance analisa outras variáveis "mais fracas". Em 2012, ela teve uma taxa de empréstimo um terço menor que a média do ramo. Mas a única maneira de fazer o sistema funcionar é aceitar a confusão.

"Um dos pontos interessantes", diz Merrill, "é que não há clientes que preencham todos os campos – sempre há uma porção de dados perdidos." A matriz das informações que a ZestFinance reúne é inacreditavelmente esparsa, um arquivo de dados cheio de células vazias. Assim, a empresa "preenche" os dados que faltam. Por exemplo, cerca de 10% dos clientes da ZestFinance são listados como mortos – mas o fato é que isso não afeta os pagamentos. "Então, é claro que, quando estiver se preparando para o apocalipse zumbi, a maioria das pessoas supõe que os débitos não serão pagos. Mas, de acordo com nossos dados, parece que os zumbis pagam, sim, seus empréstimos", brinca Merrill.

Em troca do convívio com a confusão, conseguimos serviços extremamente valiosos, impossíveis em escopo e escala com os métodos e instrumentos tradicionais. De acordo com algumas estimativas, apenas 5% de todos os dados digitais são "estruturados" — isto é, arquivados num formato que se encaixa perfeitamente num banco de dados tradicional. Sem aceitar a confusão, os demais 95% de dados não estruturados, como sites e vídeos, continuam obscuros. Ao permitir a imprecisão, abrimos uma janela para o desaproveitado universo das ideias.

A sociedade fez duas trocas tão engendradas na maneira como agimos que nem mesmo as vemos mais como trocas, e sim como o estado natural das coisas. Primeiro, supomos que não podemos usar muito mais dados, então não os usamos. Mas os limites são cada vez menos relevantes, e há muito a ser ganho por algo que se aproxime de N1 = 1tudo.

A segunda troca trata da qualidade da informação. Era racional privilegiar a exatidão numa época de dados menores, quando, pelo fato de coletarmos pouca informação, a precisão tinha de ser a maior possível. Em muitos casos, isso ainda pode importar. Mas, em outros, a precisão rigorosa é menos importante que o entendimento de sua vastidão ou do progresso em relação ao tempo.

A maneira como enxergamos o uso da totalidade das informações em comparação com fatias menores delas, a maneira como podemos vir a admirar a confusão em vez da exatidão, terão um impacto profundo na nossa interação com o mundo. Na medida em que as técnicas de big data se tornam parte regular da vida cotidiana, nós, como sociedade, podemos começar a ter dificuldade para entender um mundo de uma perspectiva maior e mais abrangente que antes, uma espécie de N1 = 1tudo da mente. E podemos tolerar a imprecisão e a ambiguidade em áreas nas quais estávamos acostumados a exigir clareza e certeza, mesmo que falhas. Podemos aceitar isso desde que, em troca, tenhamos uma noção mais completa da realidade – o equivalente a uma pintura impressionista, na qual cada pincelada é confusa se examinada de perto, mas de longe se vê uma majestosa imagem.

O big data, com sua ênfase em bancos de dados abrangentes e confusos, nos ajuda a nos aproximarmos mais da realidade que nossa dependência dos pequenos dados e da precisão. O apelo de "um pouco" e "certo" é inegável. Nossa compreensão do mundo podia ser incompleta, e às vezes errada, ao nos limitarmos ao que podíamos analisar, mas há uma confortável exatidão em relação a isso, uma estabilidade que nos

dá segurança. Além disso, e pelo fato de estamos atolados em dados que podíamos coletar e examinar, não enfrentamos a mesma compulsão por abranger tudo, por ver tudo de todos os ângulos possíveis. No confinamento mais estreito dos dados menores, podíamos nos orgulhar de nossa precisão – mesmo se, ao medirmos o detalhe a um grau infinito, ignorássemos a imagem completa.

Por fim, o big data pode exigir que mudemos, que nos acostumemos com a desordem e incerteza. As estruturas de exatidão que parecem nos dar limites na vida – o fato de o pino redondo se encaixar no buraco redondo; de haver apenas uma resposta a uma pergunta – são mais maleávamos do que podemos admitir; e, ao admitirmos isso, ao até mesmo aceitarmos, essa maleabilidade nos aproxima da realidade.

Por mais radical que as transformações dessa mentalidade sejam, levam a uma terceira alteração que tem o potencial de acabar com uma convenção ainda mais fundamental sobre a qual a sociedade se baseia: a ideia de compreender os motivos por trás de tudo o que acontece. Em vez disso, e como o próximo capítulo explicará, encontrar as associações nos dados e agir de acordo geralmente é o que basta.

Capítulo 4

Correlação

Greg Linden tinha 24 anos em 1997, quando interrompeu sua pesquisa de PhD em inteligência artificial na University of Washington para trabalhar numa empresa de internet vendendo livros on-line. A empresa fora aberta havia apenas dois anos, mas crescia rapidamente. "Adorei a ideia de vender livros e conhecimento – e de ajudar as pessoas a encontrarem a próxima porção de conhecimento de que gostariam", lembra-se. A loja era a Amazon.com, que contratou Linden como engenheiro de software para se certificar de que o site funcionasse corretamente.

A Amazon não tinha apenas técnicos na equipe. Na época, ela também empregava uma dúzia ou mais de críticos literários e editores para escrever resenhas e sugerir novos títulos. Apesar de a história da Amazon ser conhecida por muitas pessoas, poucos se lembram de que seu conteúdo era de fato elaborado por mãos humanas. Os editores e críticos avaliavam e escolhiam os títulos que apareciam nas páginas da Amazon. Eles eram responsáveis pelo que foi chamado de "a voz da Amazon" — considerada uma das joias da empresa e fonte de sua vantagem competitiva. Por esta época, um artigo no *Wall Street Journal* os considerou os críticos literários mais influentes do país, já que motivavam muitas vendas.

Foi então que Jeff Bezos, fundador e CEO da Amazon, começou a fazer experiências com uma ideia em potencial. E se a empresa pudesse recomendar livros específicos para seus consumidores com base em suas preferências de compra? Desde o início, a Amazon coletou toneladas de dados dos consumidores: o que eles compravam, que livros apenas olhavam, mas não compravam, durante quanto tempo visitavam cada página e que livros compravam juntos.

A quantidade de dados era tamanha que, a princípio, a Amazon os processava do modo convencional: pegava uma amostragem e a analisava a fim de encontrar semelhanças entre os consumidores. As recomendações resultantes eram falhas. Se você comprasse um livro sobre a Polônia seria bombardeado com títulos sobre o Leste Europeu. Se você comprasse um livro sobre bebês seria inundado por livros semelhantes. "A tendência era a de oferecer poucas variações com base nas compras anteriores, *ad infinitum*", lembra James Marcus, resenhista da Amazon de 1996 a 2001, em sua biografia, *Amazonia.* "Parecia que você ia às compras com o idiota da aldeia."

Greg Linden vislumbrou uma solução. Ele percebeu que o sistema de recomendação não precisava comparar as pessoas, tarefa tecnicamente complexa. Ele precisava apenas encontrar associações entre os produtos. Em 1998, Linden e seus colegas pediram a patente da "filtragem colaborativa item a item", como a técnica é conhecida hoje. A mudança de abordagem fez enorme diferença.

Como os cálculos podiam ser feitos antecipadamente, as recomendações eram muito mais rápidas. O método também era versátil, capaz de funcionar em várias categorias de produtos. Assim, quando a Amazon cresceu e passou a vender outros itens além de livros, o sistema podia sugerir filmes e torradeiras também. As recomendações eram muito melhores que antes porque o sistema usava todos os dados. "A piada no grupo era a de que, se o sistema estivesse funcionando perfeitamente, a Amazon deveria lhe mostrar apenas um livro – o próximo que você iria comprar", lembra Linden.

Agora a empresa tinha de decidir o que apareceria no site: Conteúdo automático como recomendações pessoais e listas dos mais vendidos ou resenhas escritas pela equipe editorial da Amazon? O que os cliques diziam ou o que os críticos diziam? Era uma decisão difícil.

Depois que a Amazon fez um teste comparando as vendas geradas pelos editores com as geradas pelo conteúdo automático, os resultados foram bem diferentes. O conteúdo automatizado de dados gerou muito mais vendas. O computador pode não saber por que um leitor que lê Ernest Hemingway pode também gostar de F. Scott Fitzgerald, mas não importava. A caixa registradora estava trabalhando. Por fim, os editores foram informados da porcentagem exata das vendas da Amazon quando a empresa apresentou suas resenhas e o grupo foi demitido. "Fiquei muito triste quando a equipe editorial foi vencida", lembra-se Linden, "mas os dados não mentem, e o custo era muito alto".

Hoje em dia, acredita-se que um terço de todas as vendas da Amazon seja resultante dos sistemas de recomendação e personalização. Com esses sistemas, a Amazon tirou muitos concorrentes do mercado: não apenas grandes lojas de discos e livros, mas também livreiros menores, que achavam que seu toque pessoal os podia isolar dos ventos da mudança. Na verdade, o trabalho de Linden revolucionou o *e-commerce*, depois que o método passou a ser empregado por quase todos. Para o Netflix, empresa de locação de filmes, três quatros de seus pedidos vêm de recomendações. Depois da Amazon, milhares de sites passaram a recomendar produtos, conteúdos, amigos e grupos sem saber por que as pessoas provavelmente se interessarão por eles.

Saber por que talvez seja agradável, mas não é importante no estímulo às vendas. Saber o quê, no entanto, motiva cliques. Este conceito tem o poder de reformar muitos ramos da economia, não apenas o *e-commerce.* Há muito tempo se diz aos vendedores de todos os setores que eles precisam entender o que faz o consumidor comprar, que precisam compreender as razões por trás das decisões. Habilidades profissionais e anos de experiência são extremamente valorizados. Mas o big data mostra que há outra abordagem, de certo modo mais pragmática. Os sistemas inovadores de recomendação da Amazon geram valiosas correlações sem saber as causas em jogo. Saber *o quê,* e não o *porquê,* basta.

PREVISÕES E PREDILEÇÕES

As correlações são úteis num mundo de poucos dados, mas é no contexto do big data que elas realmente se destacam. Por meio delas, podemos obter ideias mais de maneira mais fácil, rápida e clara que antes.

Em essência, uma correlação quantifica a relação estatística entre dois dados. Uma forte correlação significa que, quando um dos dados se altera, o outro provavelmente

se alterará também. Vimos esta forte correlação com o Google Flu Trends: quanto mais pessoas numa localização geográfica procuram por termos específicos no Google, mais pessoas neste lugar têm gripe. Do mesmo modo, uma correlação fraca significa que, quando um dado se altera, pouco acontece ao outro dado. Podemos, por exemplo, calcular a correlação entre o comprimento do cabelo de uma pessoa e sua felicidade, e então descobrir que o comprimento do cabelo não é muito útil em nos fornecer dados a respeito da felicidade.

As correlações permitem que analisemos um fenômeno não pelo foco nas suas engrenagens, e sim pela identificação de um substituto para elas. Claro que nem mesmo as mais fortes correlações são perfeitas. É bem possível que dois elementos sejam semelhantes por mera coincidência. Podemos ser "enganados pela aleatoriedade", para citar uma frase do empírico Nassim Nicholas Taleb. Com correlações, não há certeza, só probabilidade. Mas, se uma correlação é firme, a probabilidade de uma conexão é alta. Muitos consumidores da Amazon podem atestar isso ao apontarem para uma estante inteira composta por recomendações da empresa.

Ao permitir que identifiquemos um bom substituto para o fenômeno, as correlações nos ajudam a captar o presente e a prever o futuro: se A geralmente acontece em conjunto com B, precisamos procurar por B a fim de prever que A acontecerá. Usar B como substituto nos ajuda a entendermos o que provavelmente está acontecendo com A, mesmo que não possamos mensurar ou observar A diretamente. Mais importante, isso também nos ajuda a prevermos o que talvez aconteça a A no futuro. Claro que as correlações não podem prever o futuro; elas apenas podem prevê-lo dentro de certa probabilidade. Mas essa capacidade é extremamente importante.

Pense no caso do Walmart, a maior empresa varejista do mundo, com mais de dois milhões de funcionários e vendas anuais em torno de US$450 bilhões — soma maior que o PIB de três quartos dos países do mundo. Antes de a internet fomentar tantos dados, a empresa talvez tivesse o maior banco de dados do mundo corporativo dos Estados Unidos. Nos anos 1990, ela revolucionou o varejo ao registrar todos os produtos como dados, por meio de um sistema chamado Retail Link, o que levou os fornecedores a monitorar a taxa e volume de vendas e o estoque. A criação dessa transparência permitiu que a empresa obrigasse os fornecedores a cuidarem eles mesmos do fornecimento. Em muitos casos, o Walmart não se "apropria" de um produto até instantes antes da venda, reduzindo, assim, o risco e os custos de estoque. O Walmart usou dados para se tornar, de fato, a maior loja de produtos consignados do mundo.

O que todos esses dados históricos poderiam revelar se analisados apropriadamente? A rede varejista trabalhava com especialistas em dados da Teradata, anteriormente a respeitada National Cash Register Company, a fim de descobrir interessantes correlações. Em 2004, o Walmart vasculhou a enormidade de dados de transações passadas: que item cada consumidor comprou e o custo total, o que mais havia em sua cesta, horário e até mesmo o clima. Ao fazer isso, a empresa notou que, antes de um aviso de furacão, não apenas a venda de lanternas aumentava como também a de Pop-Tarts, típico doce americano. Assim, à medida que uma tempestade se aproximava, o Walmart estocava as caixas de Pop-Tarts na parte da frente das lojas, perto dos artigos para furacões, a fim de facilitar a vida dos consumidores que entravam e saíam – e, assim, aumentando as vendas.

No passado, alguém na sede da empresa teria de intuir a fim de conseguir os dados e testar a ideia. Hoje, com tantos dados e melhores instrumentos, as correlações surgem com mais rapidez e a custos menores. (Portanto, é preciso ter cuidado: à medida que a quantidade de dados aumenta em magnitude, também encontramos mais correlações ilegítimas — fenômenos que parecem correlacionados, mas não estão. É preciso um cuidado extra, que só agora estamos começando a valorizar.)

Muito antes do big data, a análise de correlação já se provava importante. O conceito foi estabelecido em 1888 por Sir Francis Galton, primo de Charles Darwin, depois que ele notou uma relação entre a altura dos homens e o comprimento de seus antebraços. A matemática por trás disso é relativamente simples e sólida, duas de suas características essenciais, que a tornaram uma das medições estatísticas mais usadas. Antes do big data, contudo, sua utilidade era limitada. Como os dados eram escassos, e a coleta, cara, os estatísticos geralmente escolhiam um substituto, depois coletavam os dados relevantes e analisavam a correlação a fim de descobrir quão bom era o substituto. Mas como escolher o substituto certo?

Para se guiar, os especialistas usavam hipóteses nascidas de teorias – ideias abstratas sobre como algo funcionava. Com base em tais hipóteses, eles coletavam dados e usavam a análise das correlações para verificar se os substitutos eram adequados. Se não fossem, os pesquisadores tentavam de novo, antes de finalmente aceitarem que a hipótese da qual partiram ou até mesmo a teoria em que se baseavam era falha e precisava de emendas. O conhecimento avançou lentamente por meio dessa abordagem de tentativa e erro com base em hipóteses, enquanto nossas tendências individuais e coletivas ofuscavam as hipóteses desenvolvidas, a forma como as aplicávamos e os substitutos escolhidos. Era um processo trabalhoso, mas funcional num mundo de poucos dados.

Na era do big data, a tomada de decisões sobre quais variáveis examinar com base apenas em hipóteses não é mais eficiente. Os bancos de dados são grandes demais, e a área sob investigação, provavelmente complexa demais. Por sorte, muitas das limitações que nos obrigaram a usar a abordagem com base em hipótese já não existem da mesma forma. Hoje, temos tantos dados disponíveis e tanto poder de processamento que não é trabalhoso escolher um ou alguns substitutos e examiná-los um a um. Hoje, a análise computacional sofisticada pode identificar o melhor substituto – como aconteceu com o Google Flu Trends, depois da análise de quase meio bilhão de modelos matemáticos.

Já não precisamos necessariamente de uma hipótese válida de um fenômeno para compreendermos o mundo. Assim, não precisamos desenvolver uma ideia a respeito de quais termos as pessoas buscam em relação a quando e onde a gripe se dissemina. Não precisamos saber como as empresas aéreas determinam preços. Não precisamos nos preocupar com os gostos culinários dos consumidores do Walmart. Em vez disso, podemos sujeitar o big data à análise de correlações e deixar que eles nos digam quais buscas são as melhores para a previsão da gripe, se os preços das passagem aéreas vai aumentar ou o que famílias ansiosas querem comer durante uma tempestade. Em vez da abordagem com base em hipóteses, podemos usar uma abordagem com base em dados. Os resultados podem ser menos tendenciosos e mais precisos e quase sempre mais rápidos.

Previsões com base em correlações estão na essência do big data. A análise de correlações hoje é usada com tanta frequência que às vezes somos incapazes de valorizar o caminho que ela construiu. E a utilização só aumenta.

Por exemplo, a pontuação de crédito é usada para prever o comportamento pessoal. A Fair Isaac Corporation, hoje conhecida como FICO, inventou a pontuação de crédito no fim dos anos 1950. Em 2011, a FICO criou a "Pontuação de Fidelidade Médica". A fim de calcular a probabilidade de as pessoas tomarem seus medicamentos, a FICO analisa muitas variáveis – incluindo as que parecem irrelevantes, como há quanto tempo a pessoa mora no mesmo endereço, se é casada, há quanto tempo está no mesmo emprego e se possui carro. A pontuação pretende ajudar os planos de saúde a economizarem dinheiro ao informá-los quais pacientes precisam ser lembrados de tomar os medicamentos. Não há qualquer causalidade em ter um carro e tomar antibióticos como prescrito; a conexão entre os dois fatores é pura correlação. Mas descobertas assim bastaram para inspirar o executivo-chefe da FICO a se vangloriar em 2011: "Sabemos o que vamos fazer amanhã."

Outros corretores de dados também estão entrando no jogo da correlação, como foi documentado pela pioneira série "What Do They Know", do *Wall Street Journal*. A Experian tem um produto chamado Income Insight, que estima o rendimento das pessoas com base, em parte, em seu histórico de crédito. Ela desenvolveu uma pontuação ao analisar seu enorme banco de dados de histórico de crédito em comparação com dados anônimos de impostos do U.S. Internal Revenue Service. Para uma empresa, a confirmação do rendimento de uma pessoa por meio de impostos pagos custa cerca de US$10, enquanto a Experian vende a estimativa dessa mesma informação por menos de US$1. Nesses casos, usar um substituto é mais eficiente do que passar por todo o ritual a fim de conseguir o dado preciso. Do mesmo modo, outra empresa de crédito, a Equifax, vende um "Índice da Capacidade de Pagamento" e um "Índice de Gastos Arbitrários", que prometem prever a capacidade de adimplência de uma pessoa.

As correlações são cada vez mais usadas. A Aviva, gigante do ramo de seguros, estudou a ideia de usar relatórios de crédito e dados de marketing direto ao consumidor como substitutos para a análise de sangue e urina de algumas cobaias. A ideia era identificar aqueles com maior risco de desenvolver problemas como hipertensão, diabetes ou depressão. O método usa dados de estilo de vida que incluem centenas de variáveis, como passatempos, sites visitados e a frequência com que a pessoa assiste à televisão, além da estimativa de rendimento.

O modelo de previsão da Aviva, desenvolvido pela Deloitte Consulting, foi considerado um sucesso na identificação dos riscos de saúde. Outras empresas de seguros, como a Prudential e a AIG, têm iniciativas semelhantes. O benefício é permitir que as pessoas que solicitam um seguro de saúde não precisem fazer testes de urina e sangue, já que são situações desagradáveis para a maioria das pessoas e pelas quais as empresas de seguro têm de pagar. Os testes laboratoriais custam por volta de US$125 por pessoa, enquanto a abordagem com base somente em dados custa cerca de US$5.

Para alguns, o método pode soar assustador, pois se baseia em comportamentos aparentemente sem relação. É como se as empresas pudessem ter um aparelho que espiona cada clique do seu mouse. As pessoas podem pensar duas vezes antes de visitar sites de esportes radicais ou de assistir a comédias que glorificam a preguiça se

acharem que terão de pagar seguros mais caros por isso. É bem verdade que intervir na liberdade das pessoas de interagir com as informações seria terrível. Por outro lado, o benefício está no fato de que tornar o seguro mais fácil e barato pode resultar em mais pessoas asseguradas, o que é bom para a sociedade e para as empresas de seguro.

Mas a vitrine das correlações do big data é mesmo a varejista americana de descontos Target, que há anos utiliza previsões baseadas em correlações de big data. Num extraordinário esforço de reportagem, Charles Duhigg, correspondente de economia do *The New York Times,* contou como a Target sabe que uma mulher está grávida sem que a futura mamãe lhes diga. Basicamente, seu método é usar todos os dados que puder e deixar que as correlações compreendam as situações por si mesmas.

Saber se uma consumidora está grávida pode ser importante para os varejistas, já que a gravidez é um momento de gastos para os casais, hora em que seu comportamento de consumidor está aberto a mudanças. Eles podem começar a ir a outras lojas e se tornar leais a outras marcas. A Target se voltou para o departamento de análise para ver se havia uma maneira de descobrir a gravidez de uma pessoa por meio de seus padrões de consumo.

O departamento de análise reviu o histórico de consumo das mulheres que fizeram uma lista de chá de bebê na loja e notou que essas mulheres compravam muito hidratante sem cheiro por volta do terceiro mês de gestação e, que alguns meses mais tarde, tendiam a comprar suplementos como magnésio, cálcio e zinco. A equipe por fim identificou duas dúzias de produtos que, quando usados como "pontes", permitiram à empresa calcular uma pontuação de "previsão de gravidez" para todas as clientes que pagavam com cartão de crédito ou usavam o cartão fidelidade ou cupons enviados pelos correios. As correlações permitiram até mesmo que os vendedores estimassem a data do parto com uma estreita margem de erro, de modo que pudessem enviar cupons relevantes em cada estágio da gestação. "No alvo" (Target) mesmo.

Em seu livro *O poder do hábito* (Objetiva, 2012)*,* Duhigg conta o que aconteceu a seguir. Um dia, um sujeito com raiva entrou numa loja da Target perto de Minneapolis e exigiu falar com o gerente. "Minha filha recebeu esta carta", gritou ele. "Ela ainda está na escola e vocês estão lhe enviado cupons para roupas de bebê e berços? Vocês estão tentando encorajá-la a engravidar?" Dias mais tarde, contudo, quando o gerente ligou para pedir desculpas ao homem, a voz do outro lado da linha era conciliadora. "Tive uma conversa com a minha filha", disse ele. "O fato é que houve algumas atividades na minha casa que eu desconhecia. Ela dará à luz em agosto. Eu lhe devo um pedido de desculpas."

Encontrar substitutos em contexto sociais é apenas uma maneira pela qual as técnicas de big data podem ser empregadas. Igualmente potentes são as correlações com novos tipos de dados para resolver necessidades diárias.

Uma dessas correlações é a chamada "análise de previsão", usada no mundo dos negócios para prever eventos. O termo pode se referir a um algoritmo capaz de encontrar uma música de sucesso, comumente usado na indústria fonográfica para que as gravadoras tenham melhor ideia de em quais discos apostar. A técnica também está sendo usada para prever grandes falhas estruturais ou mecânicas: instalar sensores em máquinas, motores e infraestruturas como pontes permite a monitoração de dados, como calor, vibração, estresse e som, a fim de detectar alterações que possam indicar um problema mais adiante.

O conceito subjacente é o de que, quando algo apresenta um defeito, em geral não é de uma só vez, e sim gradualmente, com o tempo. Munida de dados dos sensores, a análise correlacional e métodos semelhantes podem identificar padrões específicos, sinais que normalmente aparecem antes que algo se quebre – o zumbido de um motor, o calor em excesso de uma máquina e assim por diante. A partir daí, é preciso apenas analisar o padrão para saber que há algo de errado. Encontrar um defeito com antecedência permite que o sistema dispare um alarme de modo que um novo componente possa ser instalado ou que o problema seja resolvido antes de surgir de fato. O objetivo é identificar e observar uma boa "ponte" e, assim, prever eventos futuros.

A empresa de entregas UPS usava análise preventiva desde o fim dos anos 2000 para monitorar a frota de 60 mil veículos nos Estados Unidos e para saber quando realizar manutenção. Uma quebra na estrada pode causar atrasos em entregas e coletas. Assim, para se prevenir, a UPS costumava trocar certas peças a cada dois ou três anos. Mas a prática era ineficiente, já que algumas peças estavam em bom estado. Desde que passaram a usar a análise preventiva, a empresa economizou milhões de dólares ao medir e monitorar peças individuais e as substituir apenas quando necessário. Em determinado caso, os dados até revelaram que todo um lote de novos veículos tinha uma peça defeituosa que podia ter dado problema, a menos que ele tivesse sido detectado antes.

Da mesma maneira, sensores são instalados em pontes e prédios para encontrar sinais de desgaste e também são usados em grandes instalações químicas e refinarias, onde uma peça quebrada pode interromper a produção. O custo da coleta e análise de dados que indiquem o momento de uma correção antecipada é menor. Note que a análise de previsão talvez não explique a causa do problema, mas apenas sua existência. Ela o alertará que um motor está superaquecendo, mas talvez não diga se o superaquecimento se deve a uma correia ou a um parafuso. As correlações mostram *o quê,* não o *porquê,* mas, como vemos, *o quê* geralmente basta.

O mesmo tipo de metodologia vem sendo aplicada na saúde para evitar "defeitos" na máquina humana. Quando um hospital aplica tubos, fios e instrumentos em um paciente, muitos dados são gerados. Só um eletrocardiograma registra mil leituras por segundo. Ainda assim, o incrível é que apenas uma fração dos dados é usada ou mantida. A maioria é eliminada, ainda que possa indicar importantes aspectos sobre a condição de um paciente ou sua reação ao tratamento. Se mantidos e comparados com os de outros pacientes, esses dados poderiam revelar extraordinários sinais sobre quais tratamentos tendem a funcionar ou não.

Desprezar dados talvez tenha sido apropriado quando o custo e complexidade de coletar, armazenar e analisar eram altos, mas não é mais o caso. A Dra. Carolyn McGregor e uma equipe de pesquisadores do Institute of Technology da University of Ontário e da IBM estavam trabalhando com vários hospitais num software para ajudar os médicos a tomarem melhores decisões em se tratando de bebês prematuros (chamados de "preemies"). O software capta e processa dados dos pacientes em tempo real e analisa 16 fluxos de dados diferentes, como frequência cardíaca e respiratória, pressão e nível de oxigênio no sangue, que juntos reúnem cerca de 1.260 pontos de dados por segundo.

O sistema consegue detectar alterações sutis nas condições dos prematuros, que podem indicar o início de uma infecção 24 horas antes de os sintomas aparecerem.

"Não se pode ver a olho nu, mas o computador pode", explica a Dra. McGregor. O sistema não conta com relações causais ou correlações. Ele lhe diz o quê, não o porquê, mas cumpre seu objetivo. O alerta antecipado permite que os médicos tratem a infecção precocemente e com intervenções médicas mais leves e os alerta com antecedência se o tratamento parecer ineficiente, o que melhora o resultado geral dos pacientes. É difícil pensar que essa técnica não será replicada em muitos outros pacientes e condições no futuro. O algoritmo em si pode não tomar as decisões, mas as máquinas cumprem seu principal objetivo: ajudar os seres humanos a fazer o que fazem de melhor.

O surpreendente é que a análise de big data da Dra. McGregor pôde identificar correlações que de certo modo escapavam à sabedoria convencional dos médicos. Ela descobriu, por exemplo, que sinais vitais constantes são detectados antes de uma infecção séria. Parece estranho, já que costumamos pensar que uma deterioração dos sinais vitais precede uma infecção geral. Podemos imaginar gerações de médicos terminando o dia de trabalho, analisando um prontuário ao lado de um berço, vendo os sinais vitais se estabilizarem e pensando ser seguro ir para casa – somente para receber uma ligação frenética das enfermeiras à meia-noite dizendo que algo saiu tragicamente errado e que sua intuição estava equivocada.

Os dados de McGregor sugerem que a estabilidade dos prematuros, em vez de um sinal de melhora, é mais como a bonança antes da tempestade – como se o corpo do bebê dissesse aos minúsculos órgãos para aguentarem firme a tempestade que está por vir. Não sabemos ao certo: os dados indicam uma correlação, não uma causalidade. Mas sabemos que são necessários métodos estatísticos aplicados a uma enormidade de dados para revelar a associação oculta. Que não haja dúvidas: o big data salva vidas.

ILUSÕES E ESCLARECIMENTOS

Num mundo de pequenos dados, em que havia poucos disponíveis, tanto investigações causais quanto análises correlacionais começavam com uma hipótese, que tinha de ser comprovada ou rejeitada. Mas como ambos os métodos exigiam uma hipótese, ambos eram igualmente suscetíveis ao preconceito ou intuição equivocada. E os dados necessários geralmente não estavam disponíveis. Hoje, com tantos dados por perto e ainda mais por vir, esses já não são obstáculos tão grandes.

Há outra diferença que só agora começa a ganhar importância. Antes do big data, em parte por causa da inadequada potência dos computadores, a maioria das análises correlacionais que usava grandes bancos de dados estava limitada à procura por relações lineares. Na verdade, claro, muitas relações são bem mais complexas. Com análises mais sofisticadas, podemos identificar relações não lineares entre os dados.

Como exemplo, durante anos, economistas e cientistas políticos acreditaram que a felicidade e a renda estavam diretamente relacionadas: aumente a renda, e em geral a pessoa ficará mais feliz. A análise de dados num gráfico, contudo, mostra que há uma dinâmica mais complexa envolvida. Para níveis de renda abaixo de certo patamar, qualquer acréscimo no salário se traduz num aumento substancial da felicidade, mas acima deste nível salários maiores não refletem aumento na felicidade da pessoa. Se fôssemos desenhar um gráfico disso, a linha se pareceria mais com uma curva que com uma reta, como se presumia com a análise linear.

A descoberta foi importante para os políticos. Se houvesse uma relação linear, faria sentido aumentar a renda de todos a fim de aumentar a felicidade geral. Mas depois que a associação não linear foi identificada, a orientação mudou para que se focasse o aumento de renda dos pobres, já que os dados mostraram que o retorno seria maior.

No caso de relações correlacionais multifacetadas, a situação é mais complexa. Por exemplo, pesquisadores de Harvard e do MIT examinaram a disparidade da imunização contra o sarampo entre a população: alguns grupos foram vacinados e outros não. A princípio, a disparidade parecia estar relacionada com a quantia que a pessoa gastava com saúde. Mas uma análise melhor revelou que a correlação não é uma linha reta; é uma curva de forma estranha. Quanto mais as pessoas gastavam com saúde, menor era a disparidade (como era de se esperar), mas, ao aumentar o gasto, a disparidade também aumentava – algumas das pessoas mais influentes pareciam temer a vacina contra o sarampo. Para as autoridades de saúde pública, essa informação é crucial, mas a análise correlacional linear simples não a teria captado.

Só agora os especialistas estão desenvolvendo os instrumentos necessários para identificar e comparar correlações não lineares. Ao mesmo tempo, as técnicas de análise correlacional são apoiadas e aperfeiçoadas por um conjunto cada vez maior de novas abordagens e softwares que identificam relações não causais em dados de diferentes ângulos — mais ou menos como os pintores cubistas tentavam captar a imagem do rosto de uma mulher de diferentes pontos de vista ao mesmo tempo. Um dos mais novos e vibrantes métodos pode ser encontrado no próspero campo da análise de rede, que possibilita mapear, medir e calcular os nós e as conexões em tudo, desde amigos no Facebook, que a justiça pode usar como precedentes, até ligações de celular para celular. Juntas, essas ferramentas ajudam a responder perguntas empíricas e não causais.

Em última análise, na era do big data, esses novos tipos de análises levarão a uma onda de ideias e previsões úteis. Vamos encontrar conexões inéditas. Vamos perceber dinâmicas técnicas e sociais complexas que há muito tempo escapavam à nossa compreensão, apesar de nossos esforços. E o mais importante: as análises não causais nos ajudarão a compreendermos o mundo ao perguntarmos primeiramente *o quê,* e não *porquê.*

A princípio, pode soar contraproducente. Afinal, como seres humanos, desejamos entender o mundo por meio de conexões causais; queremos acreditar que todo efeito tem uma causa, se analisarmos profundamente. Não deveria ser esta nossa maior aspiração, a de conhecer as razões que movem o mundo?

Para ser claro, há um debate filosófico que remonta a séculos sobre se a causalidade sequer existe. Se tudo é causado por algum fator, a lógica diz que não há liberdade de decisão. A vontade humana não existiria, já que todas as decisões e todos os pensamentos seriam causados por algo que, por sua vez, é o efeito de outra causa, e assim por diante. A trajetória de toda a vida seria determinada por causas que geram efeitos. Assim, os filósofos têm debatido o papel da causalidade no mundo e, às vezes, a confrontado com o livre-arbítrio. Esse abstrato debate, contudo, não é o que procuramos aqui.

Em vez disso, quando dizemos que os seres humanos veem o mundo por meio das causalidades, nos referimos a duas maneiras fundamentais de acordo com as quais os humanos explicam e entendem o mundo: por meio da rápida e ilusória

causalidade e de lentos e metódicos experimentos causais. O big data transformará os papéis de ambos.

Primeiro, há o desejo intuitivo de vermos conexões causais. Temos a tendência de presumir causas mesmo que elas inexistam. Isso não se deve à cultura, à criação ou ao nível de educação. Ao contrário, as pesquisas sugerem que se trata de como a cognição humana funciona. Ao vermos dois eventos acontecendo um após o outro, nossas mentes sentem um enorme desejo de vê-los em termos causais.

Veja o exemplo das três frases seguintes: "Os pais de Fred chegaram tarde. O serviço de bufê era esperado em breve. Fred estava furioso." Ao lê-las, instantemente intuímos por que Fred estava furioso – não porque o serviço de bufê chegaria logo, e sim porque seus pais estavam atrasados. Na verdade, não temos como inferir isso a partir da informação que nos foi passada. Ainda assim, nossas mentes não se contêm e criam o que presumimos serem histórias coerentes e causais a partir dos fatos que nos foram dados.

Daniel Kahneman, professor de psicologia de Princeton e ganhador do Prêmio Nobel de Economia de 2002, usa esse exemplo para sugerir que temos dois modos de raciocínio. Um é rápido, exige pouco esforço e nos faz tirar conclusões em segundos. O outro é lento e difícil, e exige que pensemos num tema específico.

O modo de pensamento rápido tende a "enxergar" conexões causais mesmo onde não existem. Ele está condicionado a confirmar nosso conhecimento e crenças pré-existentes. Na história antiga, este modo rápido de pensar nos ajudou a sobrevivermos num ambiente perigoso, no qual geralmente precisávamos decidir rapidamente e com informações limitadas. Mas esse tipo de pensamento dificilmente consegue estabelecer uma relação clara de causa e efeito.

Infelizmente, argumenta Kahneman, em geral, nosso cérebro é preguiçoso no que diz respeito ao pensamento lento e metódico. Em vez disso, deixamos que o pensamento rápido assuma as rédeas. Como consequência, em geral "enxergamos" causalidades imaginárias, e, em essência, compreendemos o mundo de forma equivocada.

Os pais geralmente dizem aos filhos que eles ficaram gripados porque não usaram gorros ou luvas no frio. Mas não há uma conexão causal direta entre as vestimentas e a gripe. Se adoecemos após a ida a um restaurante, intuitivamente culpamos a comida (e talvez evitemos o restaurante no futuro), mesmo que a comida nada tenha a ver com a doença. Podemos ter adoecido do estômago de várias maneiras, como ao cumprimentar outra pessoa doente. O lado rápido do nosso cérebro está programado para chegar apressadamente a quaisquer conclusões causais que possamos inventar, o que geralmente nos leva a decisões erradas.

Ao contrário do que diz a sabedoria convencional, a intuição humana da causalidade não aprofunda a compreensão do mundo. De certo modo, trata-se de pouco mais que um atalho cognitivo que nos dá a ilusão de uma descoberta, mas que, na realidade, não nos deixa enxergar o mundo que nos rodeia. Assim como a amostragem era um atalho quando não podíamos processar todos os dados, a percepção da causalidade é um atalho que nosso cérebro utiliza para evitar pensar muito e lentamente.

Num mundo de poucos dados, a comprovação de que as intuições causais estavam erradas levava tempo. Isso vai mudar. No futuro, as correlações do big data serão rotineiramente usadas para discordar de nossa intuição causal, mostrando que geralmente

há pouco ou nenhuma conexão estatística entre o efeito e o que se supõe ser a causa. Nosso "raciocínio rápido" passará por um longo e duradouro teste de realidade.

Talvez essa lição nos faça pensar mais (e mais lentamente), à medida que buscamos entender o mundo. Mas mesmo nosso raciocínio lento – a segunda maneira com a qual investigamos causalidades – verá seu papel transformado pelas correlações do big data.

Em nosso cotidiano, pensamos tanto em termos causais que podemos acreditar que a causalidade é facilmente demonstrável. A verdade é bem menos consoladora. Ao contrário das correlações, nas quais a matemática é relativamente simples, não há maneira matemática óbvia para se "provar" a causalidade. Não podemos nem mesmo expressar relações causais facilmente com equações simples. Assim, mesmo que pensemos lentamente, a descoberta conclusiva de relações causais é difícil. Como nossas mentes estão acostumadas a um mundo pobre de informações, tendemos a raciocinar com dados limitados, mesmo que frequentemente muitos fatores estejam em jogo apenas para reduzir um efeito a uma causa específica.

Veja o exemplo da vacina antirrábica. Em 6 de julho de 1885, o químico francês Louis Pasteur conheceu Joseph Meister, de 9 anos, que fora mordido por um cão com raiva. Pasteur inventou a vacinação e trabalhava numa vacina experimental antirrábica. Os pais de Meister imploraram a Pasteur que usasse a vacina para tratar o filho. Ele o fez, e Joseph Meister sobreviveu. Na imprensa, Pasteur foi celebrado por ter poupado o menino de uma morte certa e dolorosa.

Mas foi isso mesmo o que aconteceu? A verdade é que apenas uma em sete pessoas mordidas por cães raivosos contraem a doença. Mesmo presumindo que a vacina experimental de Pasteur tenha sido eficiente, ela faria diferença em um entre sete casos. Havia uma probabilidade de 85% de que o menino teria sobrevivido de qualquer maneira.

Neste exemplo, a administração da vacina foi vista como a cura de Joseph Meister. Mas há duas conexões causais em questão: uma entre a vacina e o vírus da raiva e outra entre ter sido mordido por um cachorro infectado e desenvolver a doença. Mesmo que a primeira seja verdadeira, a segunda só se aplica a uma minoria de casos.

Os cientistas superam o desafio de demonstrar causalidade por meio de experimentos, nos quais uma suposta causa pode ser cuidadosamente aplicada ou rejeitada. Se os efeitos correspondem à causa, a experiência sugere uma conexão causal. Quanto mais cuidadosamente controladas as circunstâncias, maior a probabilidade de a conexão causal identificada ser correta.

Portanto, assim como as correlações, a causalidade raramente pode ser provada, apenas demonstrada com alto grau de probabilidade. Mas, ao contrário das correlações, experimentos realizados para confirmar conexões causais geralmente despertam questões éticas ou não são práticos. Como podemos realizar um experimento causal para identificar por que certos termos de busca preveem melhor a gripe? E quanto à vacina antirrábica, teríamos de sujeitar dúzias ou centenas de pacientes a uma morte dolorosa – como parte do "grupo de controle" que não recebeu a vacina –, apesar de termos uma vacina para todos? Além do mais, mesmo quando os experimentos são práticos, ainda são caros e demorados.

Em comparação, análises não causais, como correlações, são geralmente mais rápidas e baratas. Ao contrário das conexões causais, dispomos de métodos matemáticos

e estatísticos para analisar relações e os instrumentos digitais necessários para demonstram sua força com confiança.

Além disso, as correlações não são apenas valiosas por si só; elas também apontam o caminho para investigações causais. Ao nos dizer como duas questões estão potencialmente conectadas, as correlações nos permitem investigar mais para vermos se há uma relação causal presente e, neste caso, por quê. Este barato e rápido mecanismo de filtragem diminui o custo da análise causal por meio de experimentos controlados. Com o uso de correlações, podemos vislumbrar importantes variáveis, que então usamos em experimentos para estudarmos a causalidade.

Mas precisamos ter cuidado. As correlações são potentes não apenas porque nos dão ideias, mas também porque essas ideias são relativamente claras e geralmente se ofuscam quando voltamos à causalidade. A Kaggle, por exemplo, empresa que organiza competições de captação de dados para empresas abertas, organizou em 2012 um concurso sobre a qualidade dos carros usados. Uma revenda de carros forneceu dados para que os estatísticos participantes criassem um algoritmo a fim de prever quais carros disponíveis num leilão provavelmente apresentariam problemas. Uma análise correlacional mostrou que os carros pintados de laranja tinham menor tendência de apresentar problemas – cerca de metade da probabilidade média dos demais carros.

Ao lermos isso, já pensamos no motivo: Será que as pessoas que possuem carros de cor laranja são entusiastas e cuidam melhor dos veículos? Será que, por ser uma cor feita por encomenda, significa que o carro foi feito com mais cuidado em outros aspectos também? Ou talvez os carros de cor laranja sejam mais facilmente notados nas rodovias e, assim, têm menor tendência a se envolverem em acidentes, de modo que têm em melhores condições de revenda?

Rapidamente somos surpreendidos por uma rede de hipóteses causais. Mas nossa tentativa de explicar os eventos deste modo apenas os tornam mais obscuros. As correlações existem; podemos demonstrá-las matematicamente. Mas não podemos fazer o mesmo para as conexões causais. Então seria melhor se parássemos de tentar explicar os motivos por trás das correlações: o *porquê* em vez de *o quê.* De outro modo, talvez aconselhemos os proprietários a pintar suas latas-velhas de laranja a fim de torná-las menos propensas a defeitos – um pensamento tolo.

Ao levar esses fatos em conta, é compreensível que a análise correlacional e métodos não causais semelhantes, com base em big data, sejam superiores às conexões causais mais intuitivas, resultado do "raciocínio rápido". Mas, numa quantidade maior de contextos, essa análise é também mais útil e eficiente que o lento raciocínio causal exemplificado por experimentos cuidadosamente controlados (e, portanto, caros e demorados).

Em anos recentes, os cientistas tentaram diminuir os custos dos experimentos realizados para investigar causas, combinando, por exemplo, pesquisas adequadas a fim de criar "quasi-experimentos". Essa prática talvez facilitasse algumas investigações causais, mas a eficiente vantagem dos métodos não causais dificilmente é superada. Mais que isso, o big data em si ajuda as investigações causais ao guiar os especialistas para as prováveis causas a serem estudadas. Em muitos casos, uma pesquisa mais profunda da causalidade será realizada depois que o big data fizer seu trabalho, momento em que queremos especificamente investigar o *porquê,* não apenas admirar o *o quê.*

A causalidade não será descartada, mas está sendo retirada de seu pedestal como a fonte primária de conhecimento. O big data impulsiona as análises não causais, geralmente substituindo investigações causais. O enigma dos bueiros explodidos em Manhattan é um dos casos.

HOMEM *VERSUS* BUEIRO

Todos os anos, centenas de bueiros em Manhattan começam a fumegar e pegam fogo. Às vezes, os bueiros explodem e a tampa, que pode pesar até 150 quilos, explode e é lançada a vários metros de altura antes de cair no chão. Não é nada agradável.

A Con Edison, empresa pública que fornece eletricidade à cidade, faz inspeções e manutenção regulares dos bueiros, todos os anos. No passado, ela se contentava com o acaso, esperando que um bueiro a ser visitado fosse o destinado a explodir. Era pouco mais que um passeio aleatório pela Wall Street. Em 2007, a Con Edison pediu auxílio aos estatísticos da Columbia University, na esperança de que pudesse usar seus dados históricos sobre o problema, como as explosões anteriores e as interconexões das estruturas, a fim de prever quais bueiros provavelmente apresentariam problemas, de modo que a empresa soubesse em quê concentrar recursos.

É um problema complexo de big data. Há mais de 150 mil quilômetros de cabos subterrâneos em Nova York, o suficiente para dar a volta na Terra três vezes e meia. Só Manhattan tem 51 mil bueiros e caixas de serviço. Algumas dessas infraestruturas remondam ao tempo de Thomas Edison, xará da empresa. Um em cada 20 cabos foi instalado antes de 1930. Apesar de serem mantidos registros desde os anos 1880, eles eram confusos – e nunca foram concebidos para uma análise de dados. Eram dados que vinham do departamento de contabilidade ou de emergência, que faziam anotações manuais sobre os problemas. Dizer que os dados eram uma bagunça é um eufemismo. Só como exemplo, os estatísticos reportaram que o termo "caixa de serviço" (*service box*), infraestrutura comum, tinha pelo menos 38 variantes, incluindo SB, S, S/B, S.B, S?B, S.B., SBX, S/BX, SB/X, S/XB, /SBX, S.BX, S &BX, S?BX, S BX, S/B/X, S BOX, SVBX, SERV BX, SERV-BOX, SERV/BOX e SERVICE BOX. Um algoritmo de computador tinha de compreender tudo isso.

"Os dados eram inacreditavelmente rudimentares", lembra Cynthia Rudin, estatística e mineradora de dados que liderou o projeto, hoje no MIT. "Recebi uma impressão de todas as mesas de cabos diferentes. Se você a desenrolasse, podia mantê-la no alto que ela não chegava ao chão. E você tinha de compreender tudo aquilo – garimpar o ouro, fazer o necessário para conseguir um bom modelo de previsão."

Para trabalhar, Rudin e sua equipe tiveram de usar todos os dados disponíveis, não apenas uma amostragem, já que qualquer um das dezenas de milhares de bueiros podia ser uma bomba-relógio. O trabalho implorava por um N1 = 1tudo. Apesar de ser legal inventar razões causais, mesmo que levasse um século, a conclusão ainda seria errada ou incompleta. A melhor maneira de concluir o projeto era encontrar as correlações. Rudin se preocupou menos com o *porquê* que com o *quê* — apesar de saber que, quando a equipe se sentasse diante dos executivos da Edison, os estatísticos teriam de justificar as bases das descobertas. As previsões podiam ser feitas por uma máquina, mas os consumidores eram humanos, e as pessoas tendem a querer razões compreensíveis.

A mineração de dados descobriu as pepitas de ouro que Rudin esperava encontrar. Depois de fazer uma limpeza nos dados a fim de que as máquinas pudessem processá-lo, a equipe começou com 106 indícios de um grande desastre envolvendo um bueiro. Eles, então, reduziram a lista a alguns indícios mais fortes. Num teste que envolvia o sistema de energia do Bronx, eles analisaram todos os dados de que dispunham, até meados de 2008. Depois usaram os dados para prever problemas em 2009. Deu certo. Os 10% de bueiros no alto da lista continham assustadores 44% de bueiros que acabaram apresentando incidentes graves.

Por fim, os maiores fatores eram a idade dos cabos e o fato de os bueiros já terem apresentado problemas. Essa informação era útil, já que a Con Edison podia facilmente organizar um ranking. Mas espere! Idade e problemas anteriores? Não soa bem óbvio? Bem, sim e não. Por um lado, como o teórico de rede Duncan Watts gosta de dizer: "Tudo é óbvio depois que você sabe a resposta." (Título de um de seus livros.) Por outro lado, é importante lembrar que havia 106 indícios no modelo inicial. Não era tão claro que peso dar a eles e depois priorizar dezenas de milhares de bueiros, cada qual com muitas variáveis que compunham milhões de pontos de dados – e os dados em si não estavam sequer em condições de serem analisados.

O caso das explosões de bueiros enfatiza a ideia de que os dados estão sendo usados para resolver difíceis problemas da vida real. Para isso, contudo, precisávamos mudar o modo como operávamos. Tínhamos de usar todos os dados, o máximo que podíamos coletar, e não uma pequena porção. Precisávamos aceitar a confusão em vez de priorizar a exatidão e tínhamos de confiar nas correlações sem saber por completo qual a base causal das previsões.

O FIM DA TEORIA?

O big data transforma a maneira como entendemos e exploramos o mundo. Na era dos pequenos dados, éramos motivados por hipóteses sobre como o mundo funcionava, as quais depois tentávamos validar coletando e analisando dados. No futuro, nossa compreensão será motivada mais pela profusão de dados que por hipóteses.

Essas hipóteses geralmente surgem de teorias das ciências naturais ou sociais, que por sua vez ajudam a explicar e/ou prever o mundo. À medida que passamos de um mundo movido por hipóteses para um mundo movido por dados, talvez nos sintamos tentados a pensar que já não precisamos de teorias.

Em 2008, o editor-chefe da revista *Wired*, Chris Anderson, anunciou que "o dilúvio de dados torna o método científico obsoleto". Numa história de capa intitulada "The Petabyte Age", ele proclamou nada menos que "o fim da teoria". O processo tradicional da descoberta científica –hipótese testada em oposição à realidade que usa um modelo com base em causalidades – está de saída, argumentou Anderson, substituído pela análise estatística de correlações puras que desprezam as teorias.

Para apoiar seu argumento, Anderson descreveu como a física quântica se tornou um campo quase puramente teórico, uma vez que os experimentos são caros, complexos e grandes demais. Há teorias, sugeria ele, que nada mais tinham a ver com a realidade. Como exemplos do novo método, Anderson se referiu ao sistema de buscas da Google e o sequenciamento genético. "Este é um mundo onde enormes quantidades de dados e

a matemática aplicada substituem todos os outros instrumentos de que dispúnhamos", escreveu. "Com dados suficientes, os números falam por si. Os petabytes permitem que digamos: 'A correlação basta."

O artigo despertou um furioso e importante debate, apesar de Anderson rapidamente recuar em relação às suas ousadas afirmações. Mas vale a pena examinar seu argumento. Em essência, Anderson afirma que, até recentemente, quando buscávamos analisar e entender o mundo, precisávamos de teoria para testar. Em contrapartida, na era do big data, não precisamos de teorias: podemos consultar os dados. Se verdadeira, essa afirmação sugeriria que todas as regras gerais sobre como o mundo funciona, como os homens se comportam, o que os consumidores compram, que componentes se quebram e assim por diante se tornam irrelevantes na medida em que o big data ganha importância.

O "fim da teoria" parece significar que, apesar de as teorias terem existido em importantes ramos, como a física e química, a análise do big data não precisa de modelos conceituais. Isso é um absurdo.

O próprio big data se baseia em teorias. Por exemplo, eles empregam teorias estatísticas e matemáticas e às vezes também usam teorias da ciência da computação. Sim, não são teorias sobre a dinâmica causal de um fenômeno específico como a gravidade, mas ainda assim são teorias. E, como demonstramos, os modelos baseados no big data têm um poder de previsão bastante útil. Na verdade, o big data pode oferecer um novo olhar e novas ideias exatamente porque está livre do raciocínio tradicional e das tendências inerentes às teorias de um campo específico.

Além disso, e como a análise do big data se baseia em teorias, não podemos fugir delas. Elas moldam nossos métodos e resultados. Tudo começa com a forma como selecionamos os dados. Nossas decisões podem ser motivadas por conveniência: Os dados estão disponíveis? Ou por uma questão econômica: Os dados podem ser coletados a um custo baixo? Nossas escolhas são influenciadas por teorias. O que escolhemos influencia o que encontramos, como disseram Danah Boyd e Kate Crawford. Afinal, a Google usou termos de busca como ponte para a gripe, não para o comprimento dos cabelos das pessoas. Do mesmo modo, quando analisamos os dados, escolhemos instrumentos que se baseiam em teorias. Ao interpretarmos os resultados, mais uma vez aplicamos teorias. A era do big data claramente não vive sem teorias – elas estão sempre presentes, com tudo o que isso significa.

Anderson merece crédito por ter levantado as questões certas – como sempre, antes dos outros. O big data talvez não represente o "fim da teoria", mas fundamentalmente transforma a maneira como entendemos o mundo. Levaremos tempo para nos acostumar a essa transformação, que desafia muitas de nossas instituições. Mas o enorme valor que ela desperta tornará a troca não apenas válida, como também inevitável. Porém, antes de isso acontecer, vale a pena notarmos como chegamos a esse ponto. Muitas pessoas do ramo da tecnologia gostam de dar crédito a novos instrumentos digitais, de chips mais rápidos a softwares mais eficientes, porque são eles que constroem os instrumentos. A sabedoria técnica é importante, mas não tanto quanto se pensa. A razão mais profunda para essa tendência é o fato de termos muito mais dados. E isso acontece porque expressamos mais aspectos da realidade na forma de dados, tema do próximo capítulo.

Capítulo 5

Dataficação

Matthew Fontaine Maury era um promissor oficial da Marinha americana prestes a assumir um novo posto no brigue *Consort*, em 1839, quando sua carruagem de repente derrapou, capotou e o jogou para o ar. Com o impacto, ele fraturou o fêmur e deslocou o joelho. A junção foi recolocada no lugar por um médico local, mas o fêmur estava mal e precisou ser reconstruído alguns dias mais tarde. Os ferimentos deixaram Maury, aos 33 anos, parcialmente incapacitado e inapto para o trabalho no mar. Depois de quase três anos de recuperação, a Marinha o colocou num cargo administrativo, como chefe do nada inspirador Departamento de Mapas e Instrumentos.

Aquele acabou se revelando o lugar perfeito para Maury. Como jovem navegante, ele sempre se sentiu assoberbado pelo fato de os navios ziguezaguearem pela água em vez de realizarem rotas mais diretas. Ao perguntar aos capitães a respeito disso, eles lhe responderam que era muito melhor navegar por um curso familiar do que arriscar um caminho novo que pudesse conter perigos ocultos. Eles viam o oceano como um reino imprevisível, onde os marinheiros enfrentavam o inesperado a cada vento e onda.

Por conta de suas viagens, Maury sabia que não era totalmente verdade. Ele via padrões em todos os ambientes. Durante uma pausa maior em Valparaiso, no Chile, ele observou as ondas, que funcionavam como um relógio. Uma tempestade de fim de tarde terminava abruptamente ao pôr do sol e se tornava uma brisa calma, como se alguém tivesse desligado a torneira. Em outra viagem, ele cruzou as quentes águas azuis da Corrente do Golfo ao flutuar por entre as paredes do Atlântico, imóveis, como se estivesse no Rio Mississippi. Na verdade, os portugueses navegaram pelo Atlântico há séculos e contavam com os ventos regulares do Ocidente, chamados “trade” (que, em inglês antigo, significava “caminho” ou “rumo” e que só depois passou a ser associado ao comércio).

Sempre que o aspirante da Marinha, Maury, chegava a um novo porto, procurava os velhos capitães do mar para extrair conhecimento deles, com base em experiências que passavam de uma geração a outra. Ele aprendeu tudo sobre marés, ventos e correntes marinhas que agiam com regularidade – informações não encontradas nos livros e mapas que a Marinha publicava para os marinheiros. Ao contrário, eles contavam com mapas que às vezes tinham mais de 100 anos, muitos com várias omissões ou imprecisões. No novo posto como superintendente do Departamento de Mapas e Instrumentos, teve por objetivo corrigir essa falha.

Ao assumir o posto, ele inventariou os barômetros, bússolas, sextantes e cronômetros do departamento. Também anotou os muitos livros náuticos, mapas e cartas náuticas que o departamento abrigava. Encontrou caixas cheias de velhos livros de

registros de todas as viagens passadas dos capitães da Marinha. Seus antecessores no cargo os consideravam lixo. Com versos satíricos ou esquisitos desenhos nas margens, pareciam às vezes mais uma fuga do tédio das viagens que um registro efetivo dos destinos dos navios.

Mas à medida que Maury tirava a poeira salgada dos livros cheios de manchas e os analisava, ficava muito empolgado. Lá estavam as informações de que precisava: registros sobre os ventos, as águas e o clima em lugares e datas específicos. Apesar de alguns registros terem pouco valor, muitos continham importantes informações. Ao reunir tudo, percebeu Maury, uma nova forma de carta náutica seria possível. Maury e sua dúzia de "computadores" – cargo daqueles que calculavam os dados – começaram o laborioso processo de extrair e tabular as informações contidas nos registros que se deterioravam.

Maury agregou os dados e dividiu todo o Atlântico em blocos de cinco graus de longitude e latitude. Para cada segmento, anotou a temperatura, velocidade e direção dos ventos e ondas, e também o mês, já que as condições são diferentes dependendo do período do ano. Quando combinados, os dados revelaram padrões e apontaram rotas mais eficientes.

Gerações de conselhos dos homens do mar acabaram por enviar os navios diretamente para calmarias ou os colocaram em oposição a ventos ou correntes. Em uma rota comum, de Nova York para o Rio de Janeiro, os marinheiros havia muito tendiam a lutar contra a natureza em vez de utilizá-la a seu favor. Os comandantes americanos aprendiam a evitar os perigos de uma rota direta para o Rio. Assim, os navios perfaziam uma rota sudeste antes de virar para sudoeste, logo depois de cruzarem a linha do Equador. A distância percorrida geralmente equivalia a cruzar o Atlântico três vezes. A confusa rota não fazia sentido. Nada havia de errado com um caminho direto para o sul.

Para melhorar a exatidão, Maury precisava de mais informações. Então, criou um formulário-padrão de registros náuticos e obrigou que toda a frota americana o usasse e o submetesse a seu departamento depois do aportamento. Os navios mercantes queriam desesperadamente manter as cartas náuticas; Maury insistiu que eles também entregassem os registros (numa versão primitiva de uma rede social viral). "Todos os navios que navegarem pelos oceanos", disse, "devem agora ser considerados um observatório flutuante, um templo da ciência." Para aperfeiçoar as cartas náuticas, ele procurou outros pontos de dados (assim como a Google contou com o algoritmo PageRank para incluir mais sinais). Ele fez os capitães jogarem ao mar garrafas com bilhetes que indicavam o dia, posição, vento e corrente em intervalos regulares e recuperarem as garrafas que avistassem. Muitos navios usavam uma bandeira especial para mostrar que estavam cooperando com a troca de informações (num presságio dos ícones de compartilhamento de links que aparecem nos sites).

A partir dos dados, surgiram naturalmente rotas marítimas nas quais os ventos e as correntes eram particularmente favoráveis. As cartas náuticas de Maury puseram fim a longas viagens, diminuindo-as em média em um terço e gerando economia para os mercadores. "Antes de conhecer seu trabalho eu atravessava o oceano às cegas", escreveu um capitão admirador. Até mesmo velhos lobos do mar, que rejeitavam os novos mapas e se baseavam na maneira tradicional ou na intuição, serviam a um propósito: se a jornada demorava mais ou acabava em desastre, eles provavam a utilidade do sistema de Maury.

Em 1855, quando publicou sua grande obra, *The Physical Geography of the Sea,* Maury havia acumulado 1,2 milhão de pontos de dados. "Assim, o jovem marinheiro, em vez de contar com a intuição até que a luz da experiência se abatesse sobre ele [...], encontraria aqui, de uma só vez, a experiência de milhares de navegadores para guiá-lo", escreveu.

Sua obra foi essencial para a construção do primeiro cabo telegráfico transatlântico. Depois de uma trágica colisão no oceano, ele rapidamente inventou um sistema de rotas, comum hoje. Até mesmo aplicou seu método à astronomia: quando o planeta Netuno foi descoberto, em 1846, Maury teve a brilhante ideia de combinar arquivos que erroneamente se referiam ao planeta como estrela, o que lhe permitiu desenhar sua órbita.

Maury tem sido ignorado pelos livros americanos de História, talvez porque o nativo da Virgínia se separou da Marinha da União durante a Guerra Civil e serviu como espião para os Confederados na Inglaterra. Mas, anos antes, ao chegar à Europa para conseguir apoio internacional às cartas náuticas, quatro países lhe concederam títulos de bravura, e ele recebeu medalhas de ouro de mais oito países, incluindo o Vaticano. No fim do século XXI, as cartas náuticas americanas ainda usam seu nome.

O comandante Maury, "explorador dos mares", estava entre os primeiros a perceber que havia um valor especial num amontoado de dados que faltava em quantidades menores – essencial aos grandes dados. Fundamentalmente, ele compreendeu que os antigos registros da Marinha realmente constituíam "dados" que podiam ser extraídos e tabulados. Assim, ele foi um dos pioneiros da dataficação, ou seja, da extração de dados de um material que ninguém achava que tinha valor. Assim como Oren Etzioni, da Farecast, que usou as informações dos antigos preços da indústria aérea para criar um negócio lucrativo, ou como os engenheiros da Google, que aplicaram velhas buscas na internet para entender surtos de gripe, Maury usou informações geradas com um objetivo e as utilizou para outro.

Seu método, em geral semelhante às atuais técnicas de big data, era incrível, considerando que utilizava lápis e papel. Sua história enfatiza a que grau o uso de dados precede a digitalização. Hoje em dia, é comum uni-los, mas é importante mantê-los separados. Para entender melhor como os dados estão sendo extraídos dos lugares mais improváveis, pense num exemplo mais moderno.

A arte e a ciência de Shigeomi Koshimizu, professor do Instituto Avançado de Tecnologia do Japão, em Tóquio, é admirar o traseiro das pessoas. Poucos pensariam que a maneira como uma pessoa se senta constitui informação, mas é possível. Quando uma pessoa está sentada, os contornos do corpo, a postura e a distribuição do peso podem ser quantificados e tabulados. Koshimizu e sua equipe de engenheiros converteram espaldares em dados ao medir a pressão em 360 pontos diferentes com sensores numa poltrona de automóvel e ao indexar cada ponto numa escala de 0 a 256. O resultado é um código digital único para cada pessoa. Num teste, o sistema foi capaz de distinguir algumas pessoas com uma precisão de 98%.

A pesquisa não é tola. A tecnologia está sendo desenvolvida como sistema antifurto para carros. Um veículo equipado com ele reconheceria quando alguém além do motorista designado estivesse ao volante e exigiria uma senha para continuar funcionando ou o motor seria desligado. A transformação da posição sentada em dados cria um serviço viável e um empreendimento potencialmente lucrativo. Sua utilidade pode ir muito além de deter o furto de veículos. Os dados agregados podem, por exemplo,

revelar a relação entre a postura dos motoristas e a segurança nas estradas, como alterações na posição antes de um acidente. O sistema também pode detectar quando um motorista muda de posição por fadiga e automaticamente acionar os freios, além de não apenas evitar que um carro seja roubado como também identificar o ladrão pelas costas (por assim dizer).

O professor Koshimizu pegou um elemento que nunca fora tratado como dado – e nem mesmo se imaginava que contivesse informações de qualidade – e o transformou num formato numérico. Da mesma forma, o comandante Maury pegou materiais que pareciam de pouca utilidade e extraiu informações, transformando-os em dados úteis. Ao fazer isso, eles permitiram que as informações fossem usadas de uma nova maneira e gerassem um único valor.

A palavra "dado" tem origem latina e pode ter o sentido de "fato". Ela se tornou título de um trabalho clássico de Euclides, no qual ele explica a geometria como ela é conhecida e pode ser demonstrada. Hoje, dados se referem a algo que permite ser registrado, analisado e reorganizado. Não há ainda um bom termo para o tipo de transformação empreendida pelo comandante Maury e o professor Koshimizu. Então vamos chamá-las de *datificação.* Datificar um fenômeno é colocá-lo num formato quantificado de modo que possa ser tabulado e analisado.

É bem diferente da digitalização, processo de converter informações analógicas nos zeros e uns do código binário, de modo que os computadores possam usá-las. A digitalização não foi o primeiro recurso dos computadores a ser utilizado. A era inicial da revolução digital foi computacional, como sugere a etimologia da palavra computador. Usamos máquinas para realizar cálculos que levariam muito tempo pelos métodos antigos: a trajetória de um míssil, os censos e a previsão do tempo. Só mais tarde é que veio o conteúdo analógico e sua digitalização. Portanto, quando Nicholas Negroponte, do MIT Media Lab, publicou seu revolucionário livro *A vida digital* (Companhia das Letras, 1995), um de seus grandes temas era a mudança de átomos para bits. Digitalizamos muitos textos nos anos 1990. Recentemente, à medida que a capacidade de armazenagem, processamento e banda aumentou, digitalizamos outras formas de conteúdo também, como imagens, vídeos e música.

Hoje há uma crença implícita entre os tecnólogos de que a linhagem do big data remonta à revolução do silício. Mas não é assim. Sistemas modernos de TI (tecnologia da informação) certamente possibilitam o big data, mas, em essência, o avanço rumo ao big data é uma continuação da antiga busca da humanidade por medir, registrar e analisar o mundo. A revolução da TI é evidente, mas a ênfase estava mais no *T,* na tecnologia. É hora de voltarmos o olhar para o *I,* a informação.

A fim de captar informações quantificáveis, de datificar, precisamos saber como medir e registrar essa medição, o que exige um conjunto certo de instrumentos, além do desejo de quantificar e registrar. Ambos são pré-requisitos da datificação, e desenvolvemos os fundamentos dessa prática séculos antes do nascimento da era digital.

QUANTIFICANDO O MUNDO

A capacidade de registrar informações é uma das linhas de demarcação entre as sociedades primitivas e avançadas. A contagem básica e a medida de comprimento e

peso estão entre os instrumentos conceituais mais antigos da civilização. No terceiro milênio antes de Cristo, a ideia da informação registrada avançara significativamente no Vale do Rio Indo, no Egito e Mesopotâmia. A precisão aumentou, assim como o uso de medições na vida cotidiana. A evolução da escrita na Mesopotâmia serviu como método preciso de registro da produção e das transações. A linguagem escrita permitiu que as primeiras civilizações medissem a realidade, a gravassem e, mais tarde, a evocassem. Juntas, a medição e a gravação facilitaram a criação dos dados. Elas são as bases da dataficação.

Passou a ser possível replicar a atividade humana. Os prédios, por exemplo, podiam ser reproduzidos a partir dos registros de suas dimensões e materiais. A dataficação também possibilitou a experimentação: um arquiteto ou um construtor podia alterar certas dimensões e manter outras e criar um novo projeto – que, por sua vez, também podia ser registrado. As transações comerciais podiam ser registradas, de modo que se soubesse quanto foi produzido numa colheita ou num campo (e quanto seria entregue ao Estado na forma de imposto). A quantificação permitiu a previsão e o planejamento, mesmo que simples, como prever se a colheita do ano seguinte seria tão farta quanto a dos anos anteriores. Possibilitou ainda que os parceiros de uma transação registrassem quanto deviam um ao outro. Sem medição e registro, não haveria dinheiro, porque não haveria dados de suporte.

Ao longo dos séculos, as medidas passaram do comprimento e peso para área, volume e tempo. No início do primeiro milênio antes de Cristo, as principais características da medição já existiam no Ocidente. Mas havia um significativo obstáculo na maneira como as primeiras civilizações faziam medições. Elas não foram criadas para cálculos, nem mesmo os relativamente simples. O sistema de contagem dos numerais romanos era ruim para a análise numérica. Sem um sistema decimal, a multiplicação e a divisão de grandes números eram difíceis mesmo para os especialistas, e à simples adição e subtração faltavam transparência.

Um sistema numérico alternativo foi desenvolvido na Índia por volta do século I depois de Cristo. Foi parar na Pérsia, onde foi aperfeiçoado, e depois repassado aos árabes, que o melhoraram ainda mais. Esse sistema é a base dos algarismos que usamos hoje. As cruzadas podem ter levado a destruição para as terras que a Europa invadia, mas o conhecimento migrou do Oriente para o Ocidente, e talvez a transferência mais importante tenha sido a dos algarismos arábicos. O papa Silvestre II, que os estudara, defendera seu uso no fim do primeiro milênio. No século XII, textos árabes que descreviam o sistema foram traduzidos para o latim e disseminados pela Europa. Como resultado, a matemática alçou voo.

Mesmo antes de os algarismos arábicos chegarem à Europa, o cálculo foi aperfeiçoado pelo uso de ábacos, tabuleiros com fichas que denotavam quantias. Ao deslocar as fichas, era possível adicionar ou subtrair. Mas o método tinha graves limitações. Era difícil calcular números grandes e menores ao mesmo tempo. E o mais importante: os números no ábaco eram imprecisos. Bastava um movimento errado ou um tranco para mudar um dígito de lugar, levando a resultados incorretos. Os ábacos podiam ser razoáveis para cálculos, mas eram ruins para registros, e a única maneira de registrar e armazenar os números do ábaco era traduzi-los para os ineficientes algarismos romanos. (Os europeus jamais foram expostos aos ábacos orientais – hoje, uma

vantagem, já que esses instrumentos talvez tivessem prolongado o uso dos algarismos romanos no Ocidente).

A matemática deu novo significado aos dados – eles agora podiam ser *analisados,* e não apenas armazenados e recuperados. A adoção disseminada dos números arábicos na Europa levou centenas de anos, desde sua implementação, no século XII, até o século XVI. Nessa época, os matemáticos se vangloriavam de que podiam calcular seis vezes mais rápido com os números algébricos que com os ábacos. O que ajudou a tornar os algarismos árabes um sucesso foi a evolução de outro instrumento de datificação: o registro de lucro e prejuízo.

Os contadores inventaram os balanços no terceiro milênio antes de Cristo. Apesar de a contabilidade ter se desenvolvido nos anos seguintes, em essência, ela continua sendo um sistema para registrar uma transação em particular em determinado lugar. Mas ela fracassava em mostrar facilmente e a qualquer momento o que os contadores e os patrões mercadores mais valorizavam: se uma transação específica ou todo um empreendimento era ou não lucrativo. O cenário começou a mudar no século XIV, quando contadores italianos começaram a registrar transações usando dois tipos de dados, um para o crédito e outro para o débito, de modo que as contas se equilibrassem. A beleza desse sistema estava no fato de ele facilitar o reconhecimento de lucros e prejuízos. De repente, todos os dados começaram a se manifestar.

Hoje em dia, o balanço de lucro e prejuízo é levado em conta apenas por suas consequências para a contabilidade e as finanças. Mas ele também representa um marco na evolução do uso dos dados, pois permitiu que as informações fossem registradas na forma de "categorias" que conectavam as contas. O sistema funcionava com o uso de regras sobre como registrar os dados – um dos primeiros exemplos de registro padronizado de informações. Um contador podia olhar os livros de outro e compreendê-los. O sistema era organizado para tornar um tipo específico de dado – o lucro e prejuízo de cada conta – rápido e direto e deixava um rastro das transações, de modo que se podia facilmente consultar os dados. Os viciados em tecnologia valorizam esse sistema hoje em dia: ele tinha um "sistema de correção de erro" embutido. Se um lado do balanço parecesse díspar, era possível consultar a entrada correspondente.

Mas, assim como os algarismos árabes, o balanço de lucros e prejuízos não foi um sucesso instantâneo. Duzentos anos depois de o método ter sido inventado, foi necessário que um matemático e uma família de mercadores mudassem a história da datificação.

O matemático foi um monge franciscano chamado Luca Pacioli. Em 1494, ele publicou um livro escrito para os leigos sobre a matemática e sua aplicação comercial, que fez enorme sucesso e se tornou o mais importante texto matemático do seu tempo. Também foi o primeiro livro a usar apenas algarismos algébricos; sua popularidade, portanto, facilitou o uso desses algarismos na Europa. Mas sua contribuição mais duradoura foi a parte dedicada aos balanços, na qual Pacioli explicou bem o sistema de lucros e prejuízos da contabilidade. Nas décadas seguintes, o material sobre os balanços foi publicado separadamente em seis idiomas e permaneceu como maior referência do tema por séculos.

Quanto à família de mercadores, foram os famosos comerciantes venezianos e patronos das artes, os Médici. No século XVI, eles se tornaram os banqueiros mais

influentes da Europa, em parte porque tinham um método mais avançado de registro de dados: o de lucro e prejuízo. Juntos, o livro de Pacioli e o sucesso dos Médici selaram a vitória da contabilidade – e, por extensão, estabeleceram o uso dos números árabes no Ocidente.

Paralelamente aos avanços no registro de dados, formas de medir o mundo – de denotar tempo, distância, área, volume e peso – continuavam a ganhar precisão cada vez maior. O empenho em compreender a natureza por meio da quantificação definiu a ciência no século XIX, quando eruditos inventaram novos instrumentos e unidades para medir as correntes elétricas, a pressão do ar, a temperatura, a frequência do som e assim por diante. Foi uma era em que absolutamente tudo tinha de ser definido, demarcado e denotado. A fascinação chegou a tal ponto que até o crânio das pessoas era medido para se calcular sua habilidade mental. Por sorte, a pseudociência da frenologia praticamente desapareceu, mas o desejo de quantificar só se intensificou.

A medição da realidade e o registro de dados prosperaram por conta de uma combinação de instrumentos e de uma mentalidade receptiva. Essa combinação é o solo fértil no qual a datificação moderna cresceu, cujos ingredientes já existiam, apesar de caros e demorados, num mundo ainda analógico. Em muitos casos, a datificação exigia paciência infinita ou no mínimo a dedicação de uma vida inteira, como as cansativas observações das estrelas e planetas de Tycho Brahe nos anos 1500. Nos poucos casos em que a datificação teve sucesso na era analógica, como no das cartas náuticas do comandante Maury, isso só acontecia por causa de uma confluência de coincidências: Maury, por exemplo, ficou confinado a um trabalho administrativo, mas com acesso a um tesouro de registros. Mas sempre que a datificação foi bem-sucedida, obteve-se muito valor agregado pelas informações subjacentes, e grandes ideias foram reveladas.

O surgimento dos computadores gerou aparelhos de medição e armazenagem digitais que tornaram a datificação muito mais eficiente. Eles também permitiram que a análise matemática dos dados revelasse seu valor oculto. Em resumo, a digitalização impulsiona a datificação, mas não é uma substituta. O ato de digitalização – a transformação de informações analógicas no formato que os computadores possam ler – em si não datifica.

QUANDO AS PALAVRAS SE TORNAM DADOS

A diferença entre digitalização e datificação se torna óbvia quando analisamos um caso no qual as duas aconteceram e comparamos suas consequências. Pense nos livros. Em 2004, a Google anunciou um plano incrivelmente arrojado: ela pegaria todas as páginas de todos os livros disponíveis e (até onde as leis de direito autoral deixassem) permitiria que todos pesquisassem e acessassem os livros pela internet gratuitamente. Para tanto, a empresa se juntou com algumas das maiores e mais prestigiadas bibliotecas acadêmicas do mundo e desenvolveu scanners que viravam automaticamente as páginas, de modo que o escaneamento de milhões de livros fosse possível e financeiramente viável.

Primeiro, a Google *digitalizou* os textos: todas as páginas foram escaneadas e transformadas numa imagem de alta resolução armazenada nos servidores da empresa. Cada página foi transformada numa cópia digital que podia facilmente ser consultada na

internet. Para recuperá-la, contudo, era preciso saber em que livro estava a informação ou ler muitas páginas até encontrá-la. Não era possível procurar determinadas palavras no texto, nem analisá-lo, porque o texto não fora dataficado. Tudo o que a Google tinha eram imagens que apenas os seres humanos podiam transformar em informações úteis – isto é, lendo.

Ainda que fosse um ótimo instrumento – uma versão moderna e digital da Biblioteca de Alexandria, a mais abrangente da História –, a Google queria mais. A empresa entendia que aquelas informações tinham um valor agregado que só podia ser utilizado quando dataficado. Assim, a Google passou a usar programas de reconhecimento de caracteres que pegavam a imagem digital e reconheciam as letras, palavras, sentenças e parágrafos. O resultado foi um texto dataficado em vez de uma imagem digitalizada da página.

Agora, a informação na página era útil não apenas para os leitores, mas também para que os computadores a processassem e os algoritmos a analisassem. A datafi cação permitiu a indexação e a busca nos textos, além de um fluxo infinito de análise textual. Hoje podemos descobrir quando certas palavras foram usadas pela primeira vez, ou se tornaram populares, conhecimento que lançou luz sobre a disseminação de ideias e a evolução do raciocínio humano ao longo dos séculos e em vários idiomas.

Você mesmo pode experimentar. O Ngram Viewer, da Google (http://books.google.com/ngrams), gerará um gráfico com o uso de palavras e frases ao longo do tempo, utilizando todo o índice do Google Books como fonte de dados. Em segundos, descobrimos que, até 1900, o termo "causalidade" era usado com mais frequência que "correlação", mas depois a taxa se inverteu. Podemos comparar estilos de escrita e termos uma ideia melhor das disputas autorais. A datafi cação também facilitou a descoberta do plágio no mundo acadêmico; como resultado, vários políticos europeus, entre eles um ministro da defesa alemão, foram obrigados a pedir demissão.

Estima-se que 130 milhões de livros tenham sido publicados desde a invenção da imprensa, em meados do século XV. Em 2012, sete anos depois de a Google começar o projeto, a empresa havia escaneado mais de 20 milhões de títulos, mais de 15% de toda a herança escrita do mundo – uma porção substancial. Foi criada uma nova disciplina acadêmica chamada "Culturomics": lexicologia computacional que tenta compreender o comportamento humano e tendências culturais por meio da análise quantitativa de textos.

Em outro estudo, pesquisadores de Harvard analisaram milhões de livros (que equivaliam a mais de 500 bilhões de palavras), a fim de revelar que menos da metade das palavras em inglês que aparecem nos livros estão incluídas em dicionários. Ou seja, escreveram eles, a riqueza de palavras "consiste em uma 'matéria oculta' sem documentação em referências padronizadas". Além disso, ao analisar algoritmicamente referências ao artista Marc Chagall, cujas obras foram banidas na Alemanha nazista porque ele era judeu, os pesquisadores mostraram que a supressão ou censura de uma ideia ou pessoa deixa "pegadas quantificáveis". As palavras são como fósseis contidos nas páginas em vez de rochas sedimentares. Os praticantes da culturomics podem estudá-las como arqueólogos. Claro que o conjunto de dados expressa milhões de tendências – os livros de uma biblioteca são um reflexo verdadeiro do mundo real ou um reflexo que autores e bibliotecários estimam? A despeito disso, a culturomics nos fornece novas lentes com as quais podemos nos entender.

A transformação de palavras em dados gera inúmeros usos. Sim, os dados podem ser usados pelos seres humanos para a leitura e pelas máquinas para a análise. Mas, como paradigma de uma empresa de big data, a Google sabe que a informação tem vários propósitos em potencial que justificam a coleta e a datificação. Assim, a empresa inteligentemente usou o texto datificado do projeto de escaneamento dos livros para melhorar seu serviço de tradução automática. Como explicado no Capítulo 3, o sistema identificava livros traduzidos e analisava quais palavras e frases os tradutores usavam como alternativas de um idioma para outro. Sabendo disso, a Google podia agora tratar a tradução como um gigantesco problema matemático, com o auxílio do computador, que calculava probabilidades a fim de determinar qual palavra substituiria melhor a outra entre os idiomas.

Claro que a Google não era a única empresa a sonhar em levar a exuberância da cultura escrita para a era dos computadores, e não foi a primeira a tentar. O projeto Gutenberg, iniciativa voluntária de disponibilizar obras em domínio público on-line, já em 1971, tratava de tornar os textos disponíveis para as pessoas, mas não pensava na utilização das palavras como dados. Tratava-se de leitura, não de reutilização. Do mesmo modo, há anos os editores têm feito experiências com livros eletrônicos. Eles também perceberam o valor essencial dos livros como conteúdo, não como dados – seu modelo de negócios se baseia nisso. Assim, eles nunca usaram nem permitiram que se usassem os dados inerentes ao texto de um livro. Nunca viram a necessidade nem valorizaram o potencial.

Muitas empresas agora tendem a explorar o mercado de livros eletrônicos. A Amazon, com o Kindle, parece ter disparado na frente. Mas, neste ramo, as estratégias da Amazon e da Google são bem diferentes.

A Amazon também tem datificado livros – mas, ao contrário da Google, não tem conseguido explorar novas utilidades do texto como dado. Jeff Bezos, fundador e executivo-chefe da empresa, convenceu centenas de editoras a lançar seus livros no formado Kindle. Os livros eletrônicos não são compostos por imagens. Se fossem, não seria possível alterar o tamanho da fonte ou exibir a página em telas coloridas ou em preto e branco. O texto é datificado, não apenas digital. Na verdade, a Amazon tem feito com milhões de livros novos o que a Google está sofrendo para conseguir fazer com muitos livros mais antigos.

Mas além do brilhante serviço de "palavras estatisticamente significantes" – que usa algoritmos para encontrar conexões entre temas de livros, que, de outro modo, podem não ser aparentes –, a Amazon não usou a riqueza de palavras para análises de big data. A empresa acredita que a base do negócio é o conteúdo que os seres humanos leem, não a análise dos textos datificados. E, para ser franco, ela provavelmente enfrenta restrições de editores conservadores quanto a como usar as informações contidas nos livros. A Google, como o menino travesso do big data disposto a ir ao limite, não tem os mesmos impedimentos: seu conteúdo é obtido pelos cliques dos usuários, não pelo acesso aos títulos dos editores. Talvez não seja injusto dizer que, pelo menos por enquanto, a Amazon entende o valor do conteúdo digitalizado, enquanto a Google entende o valor de datificá-lo.

QUANDO O LUGAR SE TORNA UM DADO

Uma das informações mais básicas do mundo é, bem, o próprio mundo. Mas, na maioria das vezes, a área nunca foi quantificada ou usada na forma de dados. A

geolocalização da natureza, de objetos e pessoas claramente constitui informação. A montanha está lá; a pessoa está aqui. Mas, para ser mais útil, essa informação precisa ser transformada em dados. Dataficar localizações exige alguns pré-requisitos: um método para medir todos os centímetros quadrados de área da Terra, uma maneira padronizada de fazer essa medição e um instrumento capaz de monitorar e gravar os dados. Quantificação, padronização, coleta. Só então poderemos armazenar e analisar a localização não como um lugar em si, mas como dado.

No Ocidente, a quantificação dos lugares começou com os gregos. Por volta de 2000 a.C., Eratóstenes inventou um sistema de linhas para demarcar a localização, algo semelhante à latitude e longitude. Mas, como várias boas ideias da Antiguidade, a prática acabou com o tempo. Um milênio e meio mais tarde, por volta de 1400 d.C., uma cópia da obra *The Geography*, de Ptolomeu, chegou a Florença vinda de Constantinopla, justamente quando a Renascença e o comércio despertavam interesse pela ciência e extraíam conhecimento dos antigos. O tratado de Ptolomeu foi uma sensação, e suas antigas lições foram aplicadas para resolver modernos desafios de navegação. A partir de então, os mapas passaram a trazer longitude, latitude e escala. O sistema mais tarde foi aperfeiçoado pelo cartógrafo flamengo Gerardus Mercator, em 1570, que permitiu que os navegantes viajassem em linha reta num mundo esférico.

Apesar de nesta época haver meios de registrar localizações, não havia um formado genericamente aceito de compartilhar a informação. Era preciso um sistema comum de identificação, exatamente como a internet tirou proveito dos domínios para fazer os e-mails, por exemplo, funcionarem universalmente. A padronização da longitude e latitude demorou, mas finalmente foi consagrada em 1884, na Conferência do Meridiano Internacional, em Washington, D.C., no qual 25 países escolheram Greenwich, na Inglaterra, como meridiano de referência e ponto zero de longitude (os franceses, que se consideravam os líderes mundiais dos padrões, se abstiveram). Nos anos 1940, foi criado o sistema de coordenada UTM (Universal Transverse Mercator), que dividiu o mundo em 60 zonas para maior precisão.

A localização geoespacial podia agora ser identificada, registrada, computada, analisada e comunicada num formato numérico e padronizado. A posição podia ser dataficada. Mas, por conta do alto custo da medição e registro de informações no cenário analógico, raramente se fazia isso. Para que a dataficação acontecesse, instrumentos mais baratos para medir a localização tinham de ser inventados. Até os anos 1970, a única maneira de determinar a localização física era por meio de marcos na paisagem, constelações, navegação com cálculos e uma tecnologia limitada de radioposicionamento.

Uma grande mudança ocorreu em 1978, quando o primeiro dos 24 satélites que compõem o GPS foi lançado. Receptores no chão podiam triangular a posição ao perceber a diferença de tempo que levava para receber um sinal de satélites a 20 mil quilômetros de altitude. Desenvolvido pelo Departamento de Defesa dos Estados Unidos, o sistema foi aberto para o uso não militar nos anos 1980 e se tornou totalmente operacional nos anos 1990; sua precisão aumentou para aplicações comerciais na década seguinte. Com precisão de um metro, o GPS marcou o momento em que um método de medir a localização, sonho dos navegantes, mapeadores e matemáticos desde a Antiguidade,

finalmente se uniu aos meios técnicos para realizar a medição rapidamente, a um custo relativamente baixo e sem a necessidade de conhecimento especializado.

Mas a informação deve, na verdade, ser gerada. Nada impedia Eratóstenes e Mercator de estimar sua localização em qualquer momento do dia, caso se dessem ao trabalho para tanto. Apesar de possível, não era prático. Do mesmo modo, os primeiros receptores GPS eram complexos e caros, adequados a submarinos, mas não às pessoas em geral nem a todo instante. O cenário mudaria, graças à onipresença dos baratos chips contidos nos aparelhos digitais. O custo de um módulo GPS caiu de centenas de dólares nos anos 1990 para cerca de US$1 hoje, em grandes volumes. Geralmente, demora uns poucos segundos para o GPS apontar uma localização, e as coordenadas foram padronizadas. Assim, 37°14'06″ N, 115°48'40″ W significa somente que alguém está na base militar americana supersecreta, local remoto de Nevada conhecido como "Area 51", onde os alienígenas (talvez) estejam.

Hoje em dia, o GPS é apenas um entre vários sistemas de captação de localização. Sistemas concorrentes existem na China e Europa. E uma precisão ainda maior pode ser alcançada com a triangularização de torres de celulares e roteadores a fim de determinar a posição com base na força do sinal, já que o GPS não funciona em lugares fechados ou em meio a prédios altos. Isso ajuda a explicar por que empresas como Google, Apple e Microsoft estabeleceram seus próprios sistemas de geolocalização para complementar o GPS. Os carros do Google Street View coletam informações de roteadores ao tirar fotografias, e o iPhone tem um sistema de localização "spyPhone" que envia os dados de volta para a Apple sem que os usuários saibam. (Os celulares Android, da Google, e o sistema operacional móvel da Microsoft também coletam esse tipo de dado.)

Além de pessoas, objetos também podem ser localizados. Com módulos sem fio, colocados dentro de veículos, a dataficação transformará a ideia dos seguros. Os dados oferecem um olhar granular em relação a horários, localizações e distâncias da direção a um preço melhor. Na Grã-Bretanha, os motoristas podem pagar por um seguro de carro de acordo com a área e o momento do dia que mais dirigem, sem taxa anual com base na idade, gênero e histórico. Essa abordagem dos seguros gera incentivos ao bom comportamento e altera o caráter do seguro, antes com base no risco coletivo calculado, agora na ação individual. A localização de pessoas por seus veículos também altera a natureza dos custos fixos, como os de estradas e outras infraestruturas, ao unir o uso desses recursos aos motoristas e outros que os "consomem". Isso era impossível antes de a geolocalização ser continuamente convertida para a forma de dados – mas esse será o futuro.

A UPS, por exemplo, utiliza dados de geolocalização de várias maneiras. Os veículos contêm sensores, módulos sem fio e GPS, de modo que a sede da empresa possa prever problemas com o motor, como vimos no capítulo anterior. Além disso, os dados permitem que a empresa saiba a localização das vans, no caso de atrasos, monitore os funcionários e avalie os itinerários para otimizar as rotas. O caminho mais eficiente é determinado em parte por dados de entregas anteriores, praticamente como as cartas náuticas de Maury se baseavam em antigas viagens marítimas.

Os programas de análises têm efeitos extraordinários. Em 2011, a UPS reduziu em quase 50 milhões de quilômetros as rotas dos motoristas e economizou 11 milhões de litros de combustível e 30 mil toneladas de emissão de carbono, de acordo com

Jack Levis, diretor de gerenciamento de processo da UPS. A segurança e a eficiência também melhoraram: o algoritmo reúne rotas com menos tráfego em cruzamentos, que tendem a gerar acidentes, desperdício de tempo e aumento do consumo de combustível, já que a van geralmente precisa diminuir a velocidade antes de fazer uma curva.

"A previsão nos trouxe conhecimento", diz Levis, da UPS. "Mas, depois do conhecimento, há algo mais: sabedoria e perspicácia. Em determinado momento, o sistema será tão inteligente que preverá problemas e os corrigirá antes que o usuário perceba que há algo de errado."

A localização dataficada em relação ao tempo é notadamente aplicada às pessoas. Durante anos, as operadoras sem fio coletaram e analisaram informações para melhorar o serviço de redes. Mas os dados são cada vez mais usados e coletados por agentes terceirizados para outros propósitos e serviços. Alguns aplicativos de *smartphones*, por exemplo, reúnem informações de localização independentemente de o aplicativo em si ter um sistema de informação de localização. Em outros casos, o objetivo do aplicativo é gerar um negócio em torno da localização dos usuários. Um exemplo disso é o Foursquare, que permite que as pessoas façam "check in" em seus lugares preferidos. Ele gera receita de programas de fidelidade, recomendações de restaurantes e outros serviços relacionados com a localização.

A capacidade de coletar dados de geolocalização está se tornando muito valiosa. Num nível individual, ela permite a publicidade direcionada com base no local onde a pessoa está ou aonde ela vai. Além disso, as informações podem ser agregadas para revelar tendências. Dados de localização permitem, por exemplo, que as empresas detectem engarrafamentos sem precisar ver os carros: a quantidade e velocidade dos telefones que se deslocam numa estrada revelam a informação. A empresa AirSage reúne 15 bilhões de registros de geolocalização por dia, a partir dos deslocamentos de milhões de celulares, a fim de criar registros de tráfego em tempo real em mais de 100 cidades dos Estados Unidos. Duas outras empresas de geolocalização, a Sense Networks e a Skyhook, podem usar dados de localização para dizer quais áreas da cidade têm vida noturna agitada ou para estimar quantos protestantes estiveram em determinada manifestação.

Mas os usos não comerciais da geolocalização podem ser os mais relevantes. Sandy Pentland, diretora do Human Dynamics Laboratory, do MIT, e Nathan Eagle foram pioneiros no que chamam de "mineração da realidade", que se refere ao processamento de enormes quantidades de dados de celulares para compreender e prever o comportamento humano. Num estudo, a análise de movimentos e padrões de ligações permitiu que eles identificassem com sucesso pessoas que contraíram a gripe antes mesmo de saberem que estavam doentes. No caso de um surto fatal de gripe, essa capacidade pode salvar milhões de vidas ao permitir que as autoridades de saúde saibam quais as áreas mais afetadas em qualquer momento. Mas, em mãos erradas, o poder da mineração da realidade pode ter consequências terríveis, como veremos mais à frente.

Eagle, fundador da *startup* de dados sem fio Jana, usa dados de celulares agregados de mais de 200 operadoras em mais de 100 países – algo em torno de 3,5 bilhões de pessoas na América Latina, África e Europa –, a fim de responder a perguntas dos executivos de marketing, como quantas vezes por semana as pessoas lavam roupa. Mas ele também usa o big data para examinar questões como a prosperidade das cidades.

Ele e um colega combinaram a localização de celulares pré-pagos na África com a quantia que gastavam ao recarregar créditos. O valor está firmemente relacionado com a renda: pessoas com maior poder aquisitivo compram mais minutos. Mas uma das descobertas contraditórias de Eagle é que as favelas não são apenas comunidades pobres, mas também agem como trampolim econômico. A questão é que o uso indireto da localização nada tem a ver com o deslocamento da comunicação móvel, propósito para o qual a informação foi inicialmente gerada. Em vez disso, depois que a localização é datificada, novas utilizações surgem e novos valores podem ser gerados.

QUANDO AS INTERAÇÕES SE TORNAM DADOS

O próximo âmbito da datificação é mais pessoal: nossas relações, experiências e estado de humor. A ideia da datificação é a espinha dorsal de várias empresas de mídias sociais. As plataformas de redes sociais não apenas nos oferecem uma maneira de encontrar e manter contato com amigos e colegas, mas usam elementos intangíveis do cotidiano e os transformam em dados que podem ser usados para outros fins. O Facebook datifica relacionamentos; eles sempre existiram e constituíram informações, mas só foram formalmente definidos como dados depois do "gráfico social" do Facebook. O Twitter permitiu a datificação de sentimentos ao criar uma maneira simples de as pessoas registrarem e compartilharem pensamentos, algo que antes se perdia no tempo. O LinkedIn datifica nossa experiência profissional, assim como Maury transformou os antigos registros náuticos, e transforma essas informações em previsões sobre nosso presente e futuro: quem conhecemos ou a que trabalhos podemos almejar.

Essas utilizações de dados ainda são embrionárias. No caso do Facebook, a empresa tem sido astutamente paciente, pois sabe que novos propósitos em excesso para os dados dos usuários os assustariam. Além disso, a empresa ainda está ajustando seu modelo de negócios (e política de privacidade) de acordo com a quantidade e tipo de dados que deseja coletar. Assim, boa parte da crítica que a empresa recebe é mais relacionada com a informação que pode coletar do que realmente com o que faz com os dados. O Facebook tinha por volta de um bilhão de usuários em 2012, conectados por meio de mais de 100 bilhões de amizades. O gráfico social resultante representa mais de 10% de toda a população mundial, datificada e disponível para uma única empresa.

Os usos em potencial são extraordinários. Várias empresas novas têm buscado adaptar o gráfico social para usá-los como sinais capazes de estabelecer a pontuação de crédito. A ideia é que os pássaros voam em bando: pessoas prudentes se tornam amigas de pessoas semelhantes, enquanto os gastadores também se reúnem. Se der certo, o Facebook pode se tornar a próxima FICO, a agência de crédito. Os exuberantes bancos de dados das empresas de mídia social formam a base de novos empreendimentos que vão muito além do superficial compartilhamento de fotos, atualizações de status e botões de "curtir".

O Twitter também tem usado dados de modo interessante. Para alguns, os 400 milhões de tweets enviados todos os dias em 2012, por mais de 140 milhões de usuários mensais, parecem pouco além de tolices aleatórias. Na verdade, geralmente são apenas isso. Mas a empresa permite a datificação dos pensamentos, do estado de espírito e das interações das pessoas, que antes não podiam ser captados. O Twitter

fez negócios com duas empresas, a DataSift e a Gnip, para vender acesso aos dados. (Apesar de todos os tweets serem públicos, o acesso à "casa das máquinas" tem um custo.) Muitas empresas analisam os tweets e às vezes usam uma técnica chamada "análise de sentimentos", a fim de reunir opiniões de consumidores ou de julgar o impacto de campanhas de marketing.

Dois fundos de hedge, a Derwent Capital, de Londres, e a MarketPsych, da Califórnia, começaram a analisar os textos dataficados dos tweets como sinais para investimentos no mercado de ações. (As estratégias reais são mantidas em segredo: em vez de investir em empresas valorizadas, elas talvez podem ter apostado contra elas.) Ambas as empresas hoje em dia vendem informações para os corretores. No caso da MarketPsych, ela se uniu à Thomson Reuters para oferecer nada menos que 18.864 índices em 119 países, atualizados minuto a minuto, sobre estados emocionais, como otimismo, melancolia, alegria, medo, raiva e até mesmo elementos como inovação, litígio e conflito. Os dados são usados por seres humanos e computadores: a matemática de Wall Street, chamada de "quants", associa os dados a algoritmos a fim de procurar correlações não vistas que podem gerar lucros. A própria frequência de tweets sobre um tema pode prever várias situações, como o rendimento de um filme de Hollywood, de acordo com um dos pais da análise das mídias sociais, Bernardo Huberman. Ele e seu colega na HP desenvolveram um modelo que analisava o índice de postagens. Assim, foram capazes de prever o sucesso de um filme com mais sucesso que outros índices de previsão.

Mas há diversas outras possibilidades. As mensagens do Twitter se limitam a 140 caracteres, mas os metadados – isto é, a informação sobre a informação – associados a cada tweet são ricos: contêm 33 itens. Alguns não parecem muito úteis, como o "papel de parede" na página do usuário ou o software para acessar o serviço. Mas outros são extremamente interessantes, como a linguagem dos usuários, sua geolocalização, quantidade e nomes das pessoas que seguem e pelas quais são seguidos. Num estudo sobre metadados, citado na revista *Science,* em 2011, a análise de 509 milhões de tweets ao longo de dois anos, escritos por 2,4 milhões de pessoas em 84 países, mostrou que o humor das pessoas corresponde a padrões diários e semanais semelhantes em várias culturas do mundo – algo impossível de se vislumbrar antes. O humor fora dataficado.

A datificação não tem apenas a ver com transformar atitudes e estado de ânimo em formato analisável, mas também com o comportamento humano. É difícil de ser registrado, principalmente no contexto de uma comunidade maior e de seus subgrupos. O biólogo Marcel Salathé, da Penn State University, e o engenheiro de software Shashank Khandelwal analisaram tweets e descobriram que o comportamento das pessoas quanto à vacinação equivalia à probabilidade de elas de fato serem vacinadas. O mais importante é que o estudo usava metadados sobre quem estava conectado a quem no Twitter para ir além. Eles notaram que talvez existissem subgrupos de pessoas sem vacinação. O que torna a pesquisa especial é o fato de que, enquanto outros estudos, como o Google Flu Trend, usavam dados agregados para considerar o estado da saúde das pessoas, a análise feita por Salathé na verdade previa o *comportamento* de saúde.

Essas primeiras descobertas indicam para onde vai a datificação. Como a Google, as várias mídias sociais, como Facebook, Twitter, LinkedIn, Foursquare e outros,

estão sobre um enorme tesouro de informações dataficadas que, uma vez analisadas, lançarão luz sobre dinâmicas sociais em todos os níveis, no âmbito individual e em sociedade.

A DATAFICAÇÃO DE TUDO

Com um pouco de imaginação, vários elementos podem ser transformados em dados – e nos surpreender. Com o mesmo espírito do trabalho sobre os espaldares e a posição sentada das pessoas, do professor Koshimizu, em Tóquio, a IBM registrou uma patente nos Estados Unidos, em 2012, sobre "Segurança por meio de tecnologia computacional de superfície", termo usado pelos advogados de propriedade intelectual para um tapete sensível ao toque, uma espécie de tela gigantesca de *smartphone*. Os usos em potencial são vários. Ela seria capaz de identificar os objetos sobre ela. Em sua forma básica, ela poderia acender as luzes num ambiente ou abrir as portas quando uma pessoa entrasse. Mais importante, contudo, é a capacidade de identificar pessoas de acordo com o peso ou postura e caminhar. A tela poderia dizer se alguém caiu e não conseguiu se levantar, importante recurso para os idosos. Os varejistas poderiam aprender mais sobre o fluxo e tráfego em suas lojas. Quando o chão é dataficado, não há teto para sua utilização.

A datificação ao máximo não é tão difícil quanto parece. Pense no movimento "eu quantificado". Ele se refere a um eclético grupo de maníacos por saúde, por medicina e por tecnologia, que medem cada elemento dos próprios corpos e vidas a fim de viver melhor – ou pelo menos de aprender novidades que jamais conheceriam de outra forma. A quantidade de "autorrastreadores" (*self-trackers*) é pequena, mas o movimento está crescendo.

Por causa dos *smartphones* e da tecnologia barata da computação, a datificação dos gestos mais essenciais da vida nunca foi tão fácil. Várias empresas permitem que as pessoas verifiquem seus padrões de sono ao medir as ondas cerebrais durante a noite. A empresa Zeo já criou a maior base de dados de atividade de sono do mundo e descobriu diferenças na quantidade de sono REM entre homens e mulheres. A Asthmapolis instalou um sensor a um inalador usado durante ataques de asma que calcula a localização via GPS; a informação agregada permite que a empresa descubra gatilhos ambientais para os ataques de asma, como a proximidade de certas plantações.

As empresas Fitbit e Jawbone permitem que as pessoas meçam a atividade física e o sono. Outra empresa, a Basis, permite que as pessoas que usam seus braceletes monitorem sinais vitais, incluindo o batimento cardíaco e a condução da pele, medições do estresse. A obtenção de dados está se tornando mais fácil e menos intrusiva que nunca. Em 2009, a Apple registrou uma patente para a coleta de dados sobre a oxigenação do sangue, batimentos cardíacos e temperatura corporal por meio de fones de ouvido.

Há muito que aprender pela datificação do funcionamento do corpo de uma pessoa. Os pesquisadores da Gjøvik University, na Noruega, e da Derawi Biometrics desenvolveram um aplicativo para *smartphones* que analisa o caminhar da pessoa e usa a informação como um sistema de segurança para destravar o telefone. Enquanto isso, dois professores do Tec Research Institute, da Georgia, Robert Delano e Brian Parise, estão desenvolvendo um aplicativo chamado iTrem, que usa o acelerômetro

embutido no aparelho para monitorar os tremores de uma pessoa com Parkinson e outros transtornos neurológicos. O aplicativo é bom para médicos e pacientes, pois permite que os pacientes substituam exames caros feitos no consultório e que os médicos monitorem remotamente os problemas dos pacientes e suas reações ao tratamento. De acordo com pesquisadores de Kioto, um *smartphone* é só um pouco menos eficiente na medição dos tremores que os acelerômetros de três eixos usados nos equipamentos médicos especializados, de modo que os telefones podem ser usados com segurança. Novamente, um pouquinho de confusão é melhor que a exatidão.

Na maioria dos casos, estamos captando informações e as colocando no formato de dados que permitem sua reutilização, que se aplica a quase tudo. A GreenGoose, empresa de San Francisco, vende sensores minúsculos que podem ser colocados em objetos para analisar a frequência de uso. Se colocados numa embalagem de fio dental, num regador ou numa caixa de areia para gatos, é possível dataficar a higiene bucal e o cuidado com as plantas e animais de estimação. O entusiasmo quanto à "rede das coisas" – a instalação de chips, sensores e módulos de comunicação em objetos cotidianos – tem a ver, em parte, com a criação de uma rede, mas até o ponto em que a datafição nos cerca.

Depois que o mundo foi dataficado, a utilização em potencial das informações é basicamente limitada apenas pela criatividade. Maury dataficou as antigas viagens dos marinheiros por meio de uma cansativa tabulação manual e, assim, revelou ideias e valores extraordinários. Hoje temos os instrumentos (estatísticas e algoritmos) e equipamentos necessários (processadores digitais e armazenagem) para realizar tarefas semelhantes com muito mais rapidez, em escala e em vários contextos diferentes. Na era do big data, até mesmo os traseiros têm importância.

Estamos em meio a um grande projeto de infraestrutura que, de certo modo, rivaliza com os do passado, dos aquedutos romanos à *Enciclopédia* do Iluminismo. Somos incapazes de valorizar esse fato porque, ao contrário da água que flui pelos aquedutos, o produto de nossos trabalho é intangível. O projeto é a datafição. Como aqueles outros avanços infraestruturais, ele gerará alterações fundamentais na sociedade.

Os aquedutos permitiram o crescimento das cidades; a imprensa facilitou o Iluminismo, e os jornais permitiram a ascensão do Estado. Mas essas infraestruturas estavam voltadas para o fluxo – de água e de conhecimento, assim como o telefone e a internet. Em contrapartida, a datafição representa um essencial enriquecimento da compreensão humana. Com a ajuda do big data, não mais veremos o mundo como uma sequência de acontecimentos explicados como fenômenos naturais ou sociais, e sim como um Universo composto essencialmente por informações.

Por mais de um século, os físicos têm sugerido que é o caso – que a informação, não os átomos, é a base de tudo. É bem verdade que soa esotérico. Por meio da datafição, contudo, em várias situações podemos agora captar e calcular, numa escala muito mais abrangente, os aspectos físicos e intangíveis da existência e agir de acordo com eles.

A visão do mundo como informação, como oceanos de dados que podem ser explorados numa dimensão maior, nos dá uma perspectiva da realidade que não tínhamos. Trata-se de uma mentalidade que pode permear todos os aspectos da vida. Hoje somos uma sociedade numérica porque presumimos que o mundo pode ser compreendido por números e pela matemática. Não damos valor ao conhecimento transmitido através do

tempo e espaço porque a ideia da palavra escrita já está enraizada. Amanhã, as gerações subsequentes talvez tenham uma "mentalidade de big data" – a suposição de que há um componente quantitativo em tudo que fazemos e de que os dados são indispensáveis ao aprendizado da sociedade. A ideia de transformar as várias dimensões da realidade em dados provavelmente parece novidade à maioria das pessoas no presente. Mas, no futuro, certamente trataremos isso como um dado (que agradavelmente remonta à própria origem do termo).

Com o tempo, o impacto da dataficação pode obscurecer o dos aquedutos e jornais, competindo, talvez, com o da imprensa e da internet e fornecendo instrumentos para mapearmos o mundo com dados. Por enquanto, contudo, a utilização mais avançada da dataficação é no mundo dos negócios, onde o big data é usado para gerar novas formas de valor – tema do próximo capítulo.

Capítulo 6

Valor

No fim dos anos 1990, a internet rapidamente se tornava um lugar sem regras e inóspito. Os "spambots" (programas que ajudavam a criar e enviar spams) inundavam as caixas de entrada dos e-mails e assolavam os fóruns on-line. Em 2000, Luis von Ahn, um jovem de 22 anos que acabara de se graduar, teve uma ideia para resolver o problema: obrigar os novos usuários a provar que eram pessoas. Assim, ele criou algo fácil para as pessoas, mas difícil para as máquinas.

Ele teve a ideia de apresentar letras difíceis de ler durante o processo de registro. As pessoas poderiam decifrar e digitar o texto correto em segundos, mas os computadores não. O Yahoo implementou o método e reduziu os spambots da noite para o dia. Von Ahn chamou sua criação de Captcha (acrônico em inglês para Teste Completamente Automatizado para Separar Computadores e Humanos). Cinco anos mais tarde, por volta de 200 milhões de captchas eram digitados diariamente.

O sistema Captcha rendeu a Von Ahn considerável notoriedade e um emprego para lecionar Ciência da Computação na Carnegie Mellon University, depois de concluir seu PhD. O sistema também contribuiu para que ele recebesse, aos 27 anos, um dos mais prestigiosos prêmios "para gênios" da MacArthur Foundation, no valor de US$500 mil. Mas quando percebeu que era o responsável pelo fato de milhões de pessoas perderem tempo digitando as confusas letras – uma quantidade enorme de informação que depois era descartada –, ele não se sentiu tão inteligente.

Procurando maneiras de dar a todo aquele poder computacional humano um fim mais produtivo, ele inventou outro sistema, chamado ReCaptcha. Em vez de escrever letras ao acaso, as pessoas escrevem duas palavras de um texto escaneado que um programa de reconhecimento de caracteres não pode entender. Uma palavra é designada para confirmar o que outras pessoas escreveram, sendo, portanto, um sinal de que a pessoa é um ser humano; a outra é uma palavra para a desambiguação. Para garantir a exatidão, o sistema apresenta a mesma palavra para que cinco pessoas diferentes a escrevam corretamente antes de confirmar sua veracidade. O dado tinha uma utilidade primária – provar que o usuário era humano – e também um propósito secundário: decifrar palavras não claras em textos digitalizados.

O valor do sistema é imenso, considerando o que custaria para contratar pessoas para o mesmo fim. Com cerca de 10 segundos por uso, 200 milhões de ReCaptchas diários compõem meio milhão de horas por dia. O salário-mínimo nos Estados Unidos era de US$7,25 por hora em 2012. Se alguém se candidatasse a escrever palavras ininteligíveis para um computador, custaria US$3,5 milhões por dia, ou mais de US$1 bilhão por ano. Em vez disso, Von Ahn inventou um sistema que trabalhava

gratuitamente, e a Google o disponibilizou sem custo para todos os sites. Hoje o sistema está incorporado em cerca de 200 mil sites, incluindo Facebook, Twitter e Craigslist.

A história do ReCaptcha reforça a importância da reutilização de dados, cujo valor está sendo modificado pela chegada do big data. Na era digital, o papel dos dados é dar apoio às transações, e às vezes se tornam o próprio bem comercializado. No mundo do big data, a situação muda novamente. O valor dos dados passa do uso primário para o uso potencial no futuro, o que gera profundas consequências: afeta a forma como os negócios valorizam seus dados e a quem conferem acesso a eles. Isso permite, e talvez até obrigue, as empresas a mudar seus modelos de negócios e altera a forma como elas enxergam e utilizam os dados.

A informação sempre foi essencial para o comércio. Os dados permitem a descoberta dos preços, por exemplo, indício sobre o quanto se deve produzir. Essa dimensão dos dados é bem compreendida. Certos tipos de informações são comercializados há muito tempo. O conteúdo de livros, artigos, música e filmes é um exemplo, assim como informações financeiras, como o mercado de ações. Nas últimas décadas, a elas se juntaram os dados pessoais. Corretores especializados em dados nos Estados Unidos, como a Acxiom, Experian e Equifax, cobram caro por dossiês abrangentes que contenham informações pessoais de centenas de milhões de consumidores. Com o Facebook, Twitter, LinkedIn e outras plataformas de mídia social, nossas conexões pessoais, opiniões, preferências e padrões de vida se juntaram ao conjunto de informações pessoais já disponíveis a nosso respeito.

Em resumo: apesar de os dados serem valiosos há muito tempo, eram vistos como um subproduto das operações essenciais de um empreendimento ou estavam limitados a categorias relativamente estreitas, como a propriedade intelectual ou a informação pessoal. Em contrapartida, na era do big data, *todos* os dados serão considerados valiosos.

Quando dizemos "todos os dados", nos referimos a até as mais cruas e aparentemente mundanas informações. Pense nas leituras de um sensor de calor de uma máquina ou nas coordenadas em tempo real de um GPS, leituras de acelerômetros e do indicador de combustível de um veículo de entrega – ou de uma frota com 60 mil veículos. Pense ainda nas bilhões de pesquisas ou no preço de quase todos os assentos em todas as companhias aéreas dos Estados Unidos há tempos.

Até recentemente, não havia como coletar, armazenar e analisar tais dados, o que limitava gravemente as oportunidades de extrair valor em potencial. No famoso exemplo do marcador de Adam Smith, que discutia a divisão do trabalho no século XVIII, seria necessário que observadores assistissem aos trabalhadores o tempo todo, medissem e contassem a produção em grossos papéis e com canetas de pena. Até mesmo a medição do tempo era difícil na época, já que os relógios confiáveis não eram comuns. As limitações do ambiente técnico coloriam as visões dos economistas clássicos quanto ao que constituía a economia, o que eles mal faziam ideia, assim como um peixe não sabe que está molhado. Por isso, quando consideravam os fatores da produção (terra, trabalho e capital), o papel da informação lhes faltava. Apesar de o custo de captação, armazenagem e uso de dados ter caído nos últimos dois séculos, até pouco tempo era relativamente caro.

O que diferencia nossa época é que muitas das limitações inerentes à coleta de dados já não existem. A tecnologia chegou a um ponto no qual grandes quantidades

de informação podem ser captadas e armazenadas a preços baratos. Os dados podem frequentemente ser coletados com facilidade, sem muito esforço ou sequer a consciência do que está sendo registrado. Como o custo de armazenagem caiu muito, é mais fácil justificar a manutenção dos dados que seu descarte. Todos esses fatores possibilitam a existência de muito mais dados disponíveis a um custo menor. Ao longo dos últimos 50 anos, o custo do armazenamento digital caiu pela metade a cada dois anos, aproximadamente, enquanto a densidade de armazenamento aumentou 50 milhões de vezes. À luz de empresas de informação, como a Farecast ou a Google – nas quais os fatos brutos entram num lado da linha de produção e saem do outro como informação processada —, os dados estão começando a parecer um novo elemento de produção.

O valor imediato da maioria dos dados é evidente para os que o coletam. Na verdade, eles provavelmente os reúnem com um propósito específico em mente. As lojas registram dados de vendas para realizar um adequado balanço financeiro. As fábricas monitoram a produção para garantir que estejam adequadas a seus padrões de qualidade. Os sites registram cada clique dos usuários – às vezes até mesmo o movimento do cursor do mouse – para analisar e otimizar o conteúdo que fornecem aos visitantes. Esses usos primários dos dados justificam sua coleta e processamento. Quando a Amazon registra não apenas os livros que as pessoas compram, como também as páginas que visitam, ela sabe que usará os dados para oferecer recomendações personalizadas. Da mesma forma, o Facebook registra as atualizações e o botão "curtir" dos usuários para determinar os anúncios mais adequados a serem exibidos no site.

Ao contrário das coisas materiais – os alimentos que comemos, a vela que acende –, o valor dos dados não diminui com o uso; ele pode ser reprocessado. Os dados são o que os economistas chamam de bem "sem concorrente": seu uso por uma pessoa não impede o uso por outra, e a informação não se desgasta como as coisas materiais. Assim, a Amazon pode usar dados de transações passadas quando faz as recomendações para os usuários – e os usa repetidamente, não apenas para o consumidor que gerou o dado mas também para muitos outros.

Assim como os dados podem ser usados várias vezes com o mesmo objetivo, o mais importante é que podem ser utilizados com vários objetivos também. Essa característica é relevante, na medida em que tentamos entender quanto a informação valerá na era do big data. Já vimos um pouco desse potencial concretizado, como quando o Walmart pesquisa os dados dos recibos antigos e percebe a correlação entre os furacões e as vendas de Pop-Tarts.

Tudo isso sugere que o valor total dos dados é muito maior que o extraído de seu primeiro uso, e que as empresas podem explorar os dados com eficiência até mesmo quando o primeiro ou cada uso subsequente agrega pouco valor, desde que utilizem os dados várias vezes.

A "ANÁLISE DE CUSTO E BENEFÍCIO" DOS DADOS

Para entender o que a reutilização dos dados significa para seu valor, pense nos carros elétricos. Seu sucesso como meio de transporte depende de muitas variáveis de logística, todas relacionadas com a duração da bateria. Os motoristas precisam recarregar a bateria dos carros com rapidez e conveniência, e as empresas de energia

precisam garantir que a energia usada por esses veículos não desestabilize o sistema. Hoje temos estações de gás muito eficientes, mas ainda não sabemos quais serão as necessidades de recarregamento e localização das estações para os carros elétricos.

Surpreendentemente, não se trata de um problema de infraestrutura, e sim de informação, e o big data é parte importante da solução. Num teste realizado em 2012, a IBM trabalhou em conjunto com a Pacific Gas and Electric Company, da Califórnia, e a indústria automotiva Honda para coletar muitas informações a fim de responder a questões fundamentais sobre quando e onde os carros elétricos sugarão a energia e o que isso significará para o fornecimento. A IBM desenvolveu um elaborado modelo de previsão com base em vários dados: nível de bateria do carro, localização, hora do dia e pontos disponíveis perto das estações de recarga. A empresa comparou os dados com o consumo atual de energia, bem como com o padrão histórico de consumo. A análise de vários fluxos de dados em tempo real e de históricos de várias fontes permitiu que a IBM previsse os melhores momentos e lugar para que os motoristas recarregassem os carros. O projeto também revelou os melhores locais para a construção das estações de recarga. Por fim, o sistema precisará levar em conta diferenças de preços em estações de recarga próximas. Até mesmo a previsão do tempo terá de ser analisada: se está fazendo sol, por exemplo, e uma estação solar próxima está produzindo muita eletricidade, mas a previsão anuncia uma semana de chuva durante a qual os painéis solares ficarão inutilizados.

O sistema usa informações geradas com um propósito e as reutiliza – em outras palavras, os dados passam de um uso primário para um secundário, o que agrega muito mais valor com o tempo. O indicador de bateria dos carros diz aos motoristas quando recarregá-los. Os dados do sistema elétrico são coletados para gerenciamento da estabilidade da rede. Esses são os usos primários. Ambos os conjuntos de dados encontram novos usos – e novos valores – quando aplicados a um objetivo bem diferente: determinar quando e onde recarregar e onde construir estações de serviços para carros elétricos. Além disso, outras informações subsidiárias são agregadas, como a localização do carro e o consumo histórico da rede. A IBM não processa os dados uma vez, e sim repetidas vezes, ao continuamente atualizar o consumo de energia dos carros elétricos e seu impacto na rede.

O valor real dos dados é como um iceberg que flutua no oceano: apenas parte é visível a princípio, enquanto boa parte permanece oculta sob a água. As empresas inovadoras capazes de entender isso extraem o valor oculto e aproveitam os enormes benefícios. Em resumo, o valor dos dados precisa ser considerado de todas as maneiras possíveis de uso no futuro, não apenas como são usados no presente. Vimos isso em muitos dos exemplos que já citamos. A Farecast utiliza dados de preços de passagens aéreas para prever o preço de passagens futuras. A Google reutilizou os termos de busca para descobrir a ocorrência da gripe. A Dra. McGregor analisou os sinais vitais de um bebê para prever o surgimento de infecções. Maury reutilizou registros de velhos capitães para revelar as correntes oceânicas.

Mas a importância da reutilização dos dados não é totalmente valorizada nos negócios e na sociedade. Poucos executivos da Con Edison, em Nova York, podiam imaginar que informações de cabos com um século de idade e registros de manutenção poderiam ser usados para prever acidentes futuros. Foi necessária uma nova geração de

estatísticos e uma nova onda de métodos e instrumentos para revelar o valor dos dados. Mesmo muitas empresas de tecnologia e internet até recentemente desconheciam o valor da reutilização dos dados.

Pode ser útil analisar os dados do modo como os físicos veem a energia. Eles se referem à energia "armazenada" ou "potencial" que existe num objeto, mas que permanece adormecida. Pense numa mola comprimida ou numa bola no alto de uma colina. A energia desses objetos permanece latente – potencial – até ser liberada, quando, digamos, a mola é solta ou a bola é empurrada colina abaixo. Agora a energia desses objetos se torna "cinética" porque estão se movendo e exercendo força em outros objetos. Depois de seu uso primário, o valor dos dados ainda existe, mas permanece latente, armazenando seu potencial como a mola ou bola, até que os dados são aplicados a um segundo uso e seu poder é novamente liberado. Na era do big data, finalmente temos mentalidade, inteligência e instrumentos para utilizar o valor oculto dos dados.

Em última análise, o valor dos dados está no que uma pessoa pode extrair de todas as maneiras possíveis de uso. Essas utilizações em potencial, aparentemente infinitas, são como opções – não no sentido financeiro do termo, e sim no sentido prático de "escolhas". O valor dos dados é a soma de todas as escolhas: o "custo/benefício" dos dados, por assim dizer. No passado, quando a principal utilização dos dados era realizada, geralmente pensávamos que os dados cumpriram seu propósito, e que estávamos prontos para apagá-los, para descartá-los. Afinal, parecia que o fundamental fora extraído. Na era do big data, os dados são como uma mina mágica de diamantes que continua fornecendo pedras preciosas depois que seu valor primário foi extraído. Há três modos principais de se liberar o valor dos dados: a reutilização básica, a fusão de bancos de dados e a descoberta de "dois pelo preço de um".

A REUTILIZAÇÃO DOS DADOS

Um exemplo clássico da reutilização criativa dos dados está nos termos de busca. À primeira vista, a informação parece simples depois de cumprir seu propósito primário. A interação momentânea entre consumidor e serviço de busca gera uma lista de sites e anúncios que servem a uma função específica e única para aquele momento. Mas antigas buscas podem ser extraordinariamente valiosas. Empresas como a Hitwise, medidora de tráfego na internet e de propriedade da corretora de dados Experian, permite que os clientes garimpem o tráfego a fim de aprender mais sobre as preferências dos clientes. Os publicitários podem usar a Hitwise para entender qual será a cor da moda na primavera. A Google faz uma versão disso com a análise de termos de busca aberta à avaliação de qualquer pessoa. A empresa lançou um serviço de previsão de negócios junto com o segundo maior banco espanhol, o BBVA, a fim de analisar o setor de turismo e também vender indicadores econômicos em tempo real com base nos termos de busca. O Bank of England usa termos de busca relacionados com propriedades para entender melhor se o preço dos imóveis está subindo ou caindo.

Empresas incapazes de valorizar a importância da reutilização dos dados aprenderam a lição da maneira mais difícil. No início, por exemplo, a Amazon fez um acordo com a AOL para administrar a tecnologia por detrás do site de *e-commerce* da AOL.

Para a maioria das pessoas, parecia um serviço de terceirização comum. Mas o que realmente interessava a Amazon, explica Andreas Weigend, ex-cientista-chefe da empresa, era possuir os dados do que os usuários da AOL procuravam e compravam, o que aumentaria o desempenho do seu sistema de recomendações. A pobre AOL nunca percebeu isso. Ela só via seus dados em termos de valor primário – vendas. A inteligente Amazon sabia que podia tirar proveito reutilizando os mesmos dados.

Veja o exemplo da entrada da Google no reconhecimento de voz com o GOOG-411 para buscas locais, que funcionou de 2007 a 2010. O gigante das buscas não tinha tecnologia própria de reconhecimento de voz, por isso precisava licenciá-la. Ela chegou a um acordo com a Nuance, líder no segmento, que ficou encantada por ter um cliente tão precioso. Mas a Nuance era ingênua no mundo do big data: o contrato não especificava quem ficaria com os registros de voz, e a Google os guardou. A análise dos dados permite indicar a probabilidade de que um fragmento de voz corresponda a uma palavra específica, essencial no aperfeiçoamento da tecnologia de reconhecimento de voz ou na criação de um novo serviço. Na época, a Nuance se via em meio às empresas de licenciamento de software, não às de garimpo de dados. Assim que percebeu o erro, ela começou a fazer acordos com operadoras de celulares e fabricantes de aparelhos para usar o serviço de reconhecimento de voz – a fim de poder coletar todos os dados.

O valor da reutilização dos dados é uma boa notícia para empresas que coletam e controlam grandes bancos de dados mas que atualmente os usam pouco, como negócios convencionais que operam off-line, que podem estar sobre gêiseres não aproveitados de informações. Algumas empresas podem ter coletado dados, os usado uma vez (se tanto) e os guardado por conta do custo barato de armazenagem – em "túmulos de dados", como os cientistas chamam os lugares onde informações antigas ficam guardadas.

As empresas de internet e tecnologia estão na linha de frente do uso do dilúvio de dados, já que coletam dados só por estarem on-line e porque estão à frente do restante da indústria que os analisa. Mas todas as empresas podem sair ganhando. Os consultores da McKinsey & Company citam o caso de uma empresa de logística, cujo nome é mantido em segredo, que notou que, no processo de entrega dos bens, havia uma enormidade de dados sobre o envio de produtos pelo mundo. Sentindo o cheiro de oportunidade, ela criou uma divisão especial para vender os dados agregados na forma de previsões empresariais e econômicas. Em outras palavras, ela criou uma versão off-line do antigo sistema de buscas da Google. Pense na SWIFT, sistema de transferências bancárias internacionais, que descobriu que os pagamentos estavam correlacionados com a atividade econômica global. Assim, a SWIFT oferece previsões de PIB com base nos dados de transferências transmitidos por sua rede.

Algumas empresas, graças à sua posição na cadeia de valor das informações, talvez sejam capazes de coletar vastas quantidades de dados, mesmo que tenham pouca serventia para elas ou que não sejam adeptas da reutilização. Operadoras de telefones celulares, por exemplo, coletam informações sobre a localização dos assinantes de modo que possam deslocar as ligações. Para essas empresas, os dados têm apenas aspectos técnicos, mas agregam valor quando reutilizados por empresas que distribuem publicidade personalizada e com base na localização. Às vezes, o valor não vem dos dados individuais e sim do que expressam em conjunto. É assim que as empresas de geolocalização AirSage e Sense Networks, que vimos no capítulo anterior, podem

vender informações sobre onde as pessoas se reúnem na sexta-feira à noite e a que velocidade os carros se deslocam no tráfego. Essas informações podem ser usadas para determinar o valor de imóveis ou os preços de outdoors.

Até mesmo as informações mais banais podem ter valor especial, se utilizadas da forma correta. Considere novamente as operadoras de telefonia celular: elas têm registros de onde e quando os telefones se conectam às bases, incluindo a intensidade do sinal. Há muito tempo, as operadoras usam esses dados para aperfeiçoar o desempenho das redes e decidir onde adicionar ou atualizar a infraestrutura. Mas os dados têm vários outros usos em potencial. Os fabricantes de aparelhos podem usá-los para descobrir o que influencia o sinal, por exemplo, a fim de melhorar a recepção dos produtos. As operadoras de celular há tempos relutam em ganhar dinheiro com essas informações, com medo de violar leis de privacidade. Mas, aos poucos, estão começando a afrouxar as diretrizes à medida que consideram seus dados fonte potencial de renda. Em 2012, a gigantesca operadora espanhola e multinacional Telefonica chegou ao ponto de criar outra empresa, a Telefonica Digital Insights, a fim de vender dados de localização anônimos e agregados para varejistas e outras empresas.

DADOS RECOMBINANTES

Às vezes, o valor latente só pode ser extraído pela combinação de dois bancos de dados, talvez muito diferentes um do outro. Podemos inovar ao combinar dados de novas maneiras. Um exemplo é um inteligente estudo publicado em 2011 sobre a possibilidade de os celulares aumentarem os riscos de câncer. Com cerca de seis bilhões de celulares no mundo, quase um para cada pessoa, a questão é fundamental. Muitos estudos têm procurado uma ligação entre as duas questões, mas esbarraram em obstáculos: amostragens pequenas demais ou período de pesquisa curto demais ou ainda a base em dados errados. Mas uma equipe de pesquisadores da Sociedade Dinamarquesa para o Estudo do Câncer elaborou uma interessante abordagem com base em dados anteriormente coletados.

Dados de todos os assinantes de celulares desde que os aparelhos surgiram na Dinamarca foram obtidos com as operadoras. O estudo analisou quem possuía um celular de 1987 a 1995, com a exceção de clientes corporativos e outros cujos dados socioeconômicos não estavam disponíveis. Ele chegou a 358.403 pessoas. O país também mantinha um registro nacional de todos os pacientes com câncer, que continha os dados de 10.729 pessoas que tiveram tumores no sistema nervoso central entre 1990 e 2007. Por fim, o estudou usou um registro nacional com informações sobre a educação e a renda dos dinamarqueses. Depois de combinar os três bancos de dados, os pesquisadores procuraram saber se os usuários de celular tinham mais câncer que os não usuários. E, entre os usuários, os que possuíam celular há mais tempo tinham maior probabilidade de desenvolver câncer?

A despeito da escala do estudo, os dados não eram nada confusos ou imprecisos: os bancos de dados exigiam o cumprimento de altos padrões de qualidade para objetivos médicos, comerciais e demográficos. A informação não foi coletada de modo a incluir tendências relacionadas com o tema do estudo. Na verdade, os dados foram gerados anos antes, e por motivos que nada tinham a ver com a pesquisa. O mais importante:

os dados não se baseavam numa amostragem, e sim em algo próximo ao N1 = 1tudo: quase todos os casos de câncer e quase todos os usuários de celular, que chegavam a 3,8 milhões de pessoas por ano. O fato de conter todos os casos significava que os pesquisadores podiam controlar subgrupos, como aqueles com maior renda.

Por fim, o grupo não detectou qualquer aumento no risco de câncer associado ao uso de telefones celulares. Por isso, a descoberta não apareceu com destaque na mídia, ao ser publicada em outubro de 2001 no jornal médico britânico *BMJ*. Mas, se uma conexão tivesse sido descoberta, o estudo teria sido estampado nas primeiras páginas dos jornais de todo o mundo, e a metodologia dos "dados recombinantes" teria sido celebrada.

Com o big data, a soma é mais valiosa que as partes e, quando recombinamos as somas de diversos bancos de dados, ela vale mais que os fatores individuais. Hoje os internautas estão acostumados a "misturas" básicas que combinam duas ou mais fontes de dados de uma nova maneira. O site imobiliário Zillow, por exemplo, sobrepõe informações dos imóveis e os preços num mapa dos bairros das cidades americanas. Ele também coleta muitos dados, como as transações recentes na vizinhança e as características das propriedades, a fim de prever o valor de casas específicas na área. A apresentação visual torna os dados mais acessíveis. Mas, com o big data, podemos ir além. O estudo dinamarquês sobre o câncer nos dá uma ideia do que é possível.

DADOS EXPANSÍVEIS

Uma maneira de permitir a reutilização dos dados é projetar expansibilidade desde o início, de modo que sejam adequados a vários usos. Apesar de não ser sempre possível – já que só podemos perceber possíveis usos depois que os dados foram coletados -, há meios de encorajar o uso diversificado de um mesmo banco de dados. Alguns varejistas, por exemplo, estão posicionando as câmeras de vigilância não apenas para pegar furtos, mas também para manter um registro do fluxo de pessoas na loja e os objetos de maior interesse. Os varejistas podem usar a informação para criar melhores projetos de lojas e para julgar a eficiência de campanhas de marketing. Antes disso, as câmeras visavam apenas à segurança. Agora, são vistas como um investimento capaz de aumentar o faturamento.

Uma das melhores empresas coletoras de dados que mantém essa possibilidade em mente é, claro, a Google. Os controversos carros Street View andam por aí tirando fotos de casas e ruas, mas também coletando dados de GPS, verificando informações de mapas e até mesmo o nome das redes wifi (e, talvez ilegalmente, o conteúdo que flui pelas redes desprotegidas). Um único carro Google Street View reúne uma vasta quantidade de dados a todo instante. A extensibilidade está incluída porque a Google aplicou os dados não apenas para o uso primário, como também para muitos usos secundários. Os dados de GPS, por exemplo, melhoraram o serviço de mapeamento da empresa e se tornaram indispensáveis ao funcionamento do carro autopilotado.

O custo extra de coletar diversas fontes e ainda mais pontos de dados em cada uma é geralmente baixo. Então, faz sentido coletar o máximo de dados possível, assim como torná-los reutilizáveis considerando os usos secundários desde o princípio, o que aumenta o valor dos dados. A questão é procurar por "dois pelo preço de um" – casos

em que um único banco de dados pode ser usado em várias situações se for coletado da maneira certa. Assim, os dados podem trabalhar dobrado.

DEPRECIANDO O VALOR DOS DADOS

Como o custo de armazenagem dos dados digitais despencou, as empresas têm forte motivação econômica para manter os dados a fim de que sejam usados para os mesmos ou semelhantes objetivos. Mas há um limite para a utilização.

Empresas como Netflix e Amazon, por exemplo, que transformam as compras dos consumidores, pesquisas e análises em recomendações de novos produtos, podem se sentir tentadas a usar os registros várias vezes ao longo dos anos. Com isso em mente, pode-se argumentar que, desde que a empresa não esteja limitada por leis e regulações como as regras de privacidade, ela pode usar os registros digitais para sempre ou no mínimo até quando for economicamente possível. Mas a realidade não é tão simples assim.

A maior parte dos dados perde parte da utilidade com o tempo. Nessas circunstâncias, continuar contando com os velhos dados não apenas deixa de agregar valor como na verdade destrói o valor dos dados novos. Pense num livro que você comprou há 10 anos na Amazon e que já não reflete seus interesses. Se a Amazon usar a compra de uma década para recomendar outros livros, é menos provável que você os compre – ou até mesmo que considere as recomendações do site. Uma vez que as recomendações da Amazon se baseiam tanto em informações velhas quanto nos dados mais recentes e ainda valiosos, a presença dos antigos dados deprecia o valor dos novos.

Assim, a empresa tem um enorme incentivo para usar dados apenas enquanto continuarem produtivos. Ela precisa incessantemente cuidar dos dados e descartar as informações que perderam valor. O desafio é saber quais dados não são mais úteis. Basear essa decisão no tempo raramente é o mais adequado. Assim, a Amazon e outras empresas criaram sofisticados modelos que as ajudam a separar os dados úteis dos irrelevantes. Se um consumidor, por exemplo, olha ou compra um livro recomendado com base numa compra anterior, as empresas de *e-commerce* podem supor que a compra antiga ainda representa as atuais preferências do consumidor. Assim, elas são capazes de avaliar a utilidade dos velhos dados e elaborar "taxas de depreciação" mais precisas para a informação.

Nem todos os dados perdem valor na mesma velocidade ou da mesma maneira. Esse fato explica por que algumas empresas acreditam que precisam guardar os dados durante o máximo de tempo possível, mesmo se agências reguladoras ou o público quiserem que sejam apagados ou mantidos como anônimos depois de um período. Por isso, a Google tem resistido aos pedidos para apagar completamente o endereço de IP das antigas buscas dos usuários. (Em vez disso, a empresa apaga apenas os quatro últimos dígitos depois de nove meses para tornar a busca quase anônima. Assim, ainda é capaz de comparar dados ano a ano, como buscas de compras de Natal – mas somente em nível regional, não individualmente.) Além disso, saber a localização de quem faz as buscas pode ajudar a aumentar a relevância dos resultados. Se muitas pessoas em Nova York, por exemplo, buscam Turquia e clicam em sites direcionados

ao país, o algoritmo destacará mais essas páginas para outras pessoas de Nova York. Mesmo que o valor dos dados diminua para alguns objetivos, o custo/benefício pode se manter firme.

O VALOR DA UTILIZAÇÃO EXAUSTIVA DOS DADOS

A reutilização dos dados às vezes pode assumir uma forma oculta e inteligente. Empresas de internet podem captar dados sobre tudo o que os usuários fazem e depois tratar cada interação como referência para personalizar o site, melhorar um serviço ou criar um produto totalmente novo. Vemos um belo exemplo disso na história de dois corretores ortográficos.

Ao longo de 20 anos, a Microsoft desenvolveu um sólido corretor ortográfico para o software Word. Ele funcionava pela comparação de um dicionário de termos corretos atualizado com frequência com o fluxo de caracteres que o usuário escrevia. O dicionário estabelecia quais eram as palavras conhecidas; o sistema tratava variações próximas que não estavam no dicionário como erros, que então corrigia. Por causa do esforço necessário para compilar e atualizar o dicionário, o corretor ortográfico do Word só estava disponível nos idiomas mais comuns e custava à empresa milhões de dólares em criação e manutenção.

Agora pense no Google que, supostamente, tem o mais completo corretor ortográfico do planeta, em praticamente todos os idiomas. O sistema é constantemente aperfeiçoado e agrega novas palavras – produto acidental do uso diário do sistema de buscas. Escreveu "iPad" errado? Está lá. "Obamacare"? Também está.

Mais que isso, o Google aparentemente obteve seu corretor ortográfico gratuitamente, reutilizando os erros escritos no sistema de buscas da empresa, em meio às três bilhões de buscas diárias. Um sistema de resposta inteligente diz ao sistema qual palavra o usuário está tentando escrever. Às vezes, os usuários explicitamente "dizem" ao Google a resposta quando o sistema lhes propõe uma pergunta no alto da página — "Você quis dizer *epidemiologia?*" —, clicando no link para dar início a uma nova busca pelo termo correto. Ou o site para o qual os usuários vão implicitamente sinaliza a ortografia correta, já que provavelmente está relacionado com a palavra certa. (Essa característica é mais importante do que pode parecer: como o corretor ortográfico do Google se aperfeiçoa continuamente, as pessoas pararam de se importar em usar os termos corretos, uma vez que o site era capaz de processá-los de qualquer forma.)

O sistema de correção ortográfica do Google mostra que dados "ruins", "incorretos" ou "deficientes" ainda podem ser muito úteis. O interessante é que o Google não foi o primeiro a ter a ideia. Por volta de 2000, o Yahoo percebeu a possibilidade de criar um corretor ortográfico a partir dos erros dos usuários, mas a ideia não foi adiante. Velhos dados de buscas eram tratados como lixo. Do mesmo modo, a Infoseek e a Alta Vista, antigamente populares sistemas de busca, tinham o maior banco de dados de palavras incorretas de seu tempo, mas nunca apreciaram seu valor. Seus sistemas, num processo invisível aos usuários, tratavam os erros como "termos relacionados" e realizavam a busca, mas se baseavam em dicionários que diziam explicitamente o que era o correto e não na reunião das interações reais dos usuários.

Somente o Google reconheceu que o lixo das interações era na verdade ouro em pó, que podia ser reunido e fundido num reluzente lingote. Um dos principais engenheiros da Google estimou que seu corretor ortográfico era melhor que o da Microsoft pelo menos em termos de magnitude (se bem que, quando pressionado, ele confirmou que não tinha como de fato validar essa conclusão). E ele negava a ideia de que o sistema fora desenvolvido "gratuitamente". A matéria-prima – erros ortográficos – pode vir sem custo direto, mas a Google provavelmente gastou mais que a Microsoft para desenvolver o sistema, confessou o engenheiro com um sorriso.

As diferentes abordagens das duas empresas são extremamente reveladoras. A Microsoft só via o valor do corretor ortográfico com um objetivo: o processamento de textos. A Google, por outro lado, compreendeu sua utilidade com mais profundidade. A empresa não usou apenas os erros para desenvolver o melhor e mais atualizado corretor ortográfico a fim de aperfeiçoar as buscas, mas supôs que o sistema tivesse outras utilidades, como a ferramenta de autocompletar os termos da busca, o Gmail, Google Docs e até mesmo o sistema de tradução.

Um termo surgiu para descrever a trilha digital que as pessoas deixam: "exaustão dos dados", que se refere a dados colhidos como subprodutos das ações e dos movimentos das pessoas. Quanto à internet, ele descreve as interações dos usuários on-line: onde clicam, por quanto tempo permanecem numa página, para onde movem o cursor, o que escrevem e muito mais. Muitas empresas criam sistemas de modo que possam exaurir os dados extraídos e reciclá-los, a fim de aperfeiçoar um serviço existente ou desenvolver novos. A Google é a líder indiscutível. A empresa aplica o serviço de "aprendizado recursivo de dados" a vários serviços. Toda ação que um usuário realiza é considerada um sinal para ser analisado e alimentado no sistema.

A Google, por exemplo, sabe muito bem quantas vezes as pessoas buscam por um tema ou temas relacionados e com que frequência clicam num link mas voltam à página de busca sem se impressionar com o que encontraram, apenas para retomar a buscar. Ela sabe se as pessoas clicaram no oitavo link da primeira página ou no primeiro link da oitava página – ou se abandonaram completamente a busca. A empresa pode não ter sido a primeira a ter a ideia, mas a implementou com extraordinária eficiência.

Essa informação é incrivelmente valiosa. Se muitos usuários tendem a clicar num resultado de busca no fim da página de resultados, provavelmente o link é mais relevante que os acima, e o algoritmo do Google sabe que deve automaticamente colocar o link mais acima nas buscas subsequentes. (E ela faz isso para anúncios também.) "Gostamos de aprender com bancos de dados enormes e 'barulhentos', brinca um funcionário da empresa.

A utilização exaustiva dos dados é o mecanismo por trás de vários serviços, como o reconhecimento de voz, filtros de spam, tradução e muitos outros. Quando os usuários indicam a um programa de reconhecimento de voz que ele não entendeu o que foi falado, eles na verdade estão "treinando" o sistema para que ele melhore.

Muitas empresas estão começando a criar sistemas para coletar e usar informações desta maneira. Nos primórdios do Facebook, o primeiro "cientista de dados", Jeff Hammerbacher (supostamente um dos cunhadores do termo) examinou a rica coleção de dados não estruturados. Ele e sua equipe descobriram que um grande indício de que as pessoas se manifestariam (publicariam um comentário, clicariam num ícone e

assim por diante) seria o fato de ver os amigos fazendo o mesmo. Assim, o Facebook remodelou seu sistema para dar mais ênfase às atividades dos amigos, o que gerou um ciclo virtuoso de novas contribuições para o site.

A ideia é a disseminação para além da internet e para qualquer empresa que busca coletar opiniões de usuários. Leitores de e-books, por exemplo, coletam vastas quantidades de dados sobre as preferências literárias e os hábitos dos usuários: quanto tempo demoram para ler uma página ou capítulo, onde leem, se viram a página com um toque leve ou se abandonam o livro. As máquinas registram sempre que os usuários sublinham uma passagem ou fazem anotações nas margens. A capacidade de coletar esse tipo de informação transforma a leitura, antes um hábito solitário, numa espécie de experiência comunal.

Uma vez agregado, os dados não estruturados podem dizer aos editores e escritores o que eles jamais saberiam de forma quantificada: os gostos, desgostos e padrões de leitura das pessoas, informação comercialmente valiosa: podemos imaginar as empresas de e-books a vendendo para os editores a fim de que melhorem o conteúdo e estrutura dos livros. A análise de dados que a Barnes & Noble faz de seu e-reader, Nook, por exemplo, revelou que as pessoas tendiam mais a abandonar livros de não ficção no meio. Essa descoberta inspirou a empresa a criar uma série chamada "Nook Snaps": obras curtas sobre temas específicos, como saúde e atualidades.

Pense em programas educativos on-line, como Udacity, Coursera e edX, que mantêm registros das interações dos alunos para entender o que funciona melhor em termos pedagógicos. As turmas têm milhares de alunos, o que gera uma quantidade extraordinária de dados. Hoje os professores podem saber se uma porcentagem dos alunos assistiu novamente a determinada parte de uma aula, o que talvez sugira que eles não foram claros em certo ponto. Ao lecionar aprendizado mecânico no Coursera, o professor de Stanford, Andrew Ng, notou que cerca de dois mil alunos erraram uma questão específica da lição de casa – mas chegaram exatamente à mesma resposta errada. Claro que estavam cometendo o mesmo erro. Mas qual?

Com um pouco de investigação, ele descobriu que os alunos estavam invertendo duas equações algébricas num algoritmo. Agora, quando outros alunos cometem o mesmo erro, o sistema não diz apenas que estão errados, mas os aconselha a revisar os cálculos. O sistema também aplica o big data ao analisar todas as publicações lidas pelos alunos no fórum de discussão e verificar se completaram a lição de casa corretamente a fim de determinar se um aluno que leu uma publicação específica no fórum alcançará o resultado correto. Essa análise visa estabelecer quais publicações dos fóruns são mais úteis aos alunos. Esse recursos, antes impossíveis, podem mudar o ensino para sempre.

A utilização exaustiva dos dados pode ser uma enorme vantagem competitiva para as empresas e se tornar uma poderosa barreira contra o surgimento de concorrentes. Pense: se uma empresa recém-lançada inventou um site de *e-commerce*, rede social ou sistema de busca melhor que os líderes de hoje, como a Amazon, Google ou Facebook, ela terá problemas para competir não apenas por causa da economia de escala, efeitos em cascata ou marca, e sim porque boa parte do desempenho dessas empresas se deve aos dados não estruturados das interações dos usuários incorporados ao serviço. Seria

possível que um site educacional tivesse o *know-how* para competir com outro que já possui uma enorme quantidade de dados com os quais pode aprender a funcionar melhor?

O VALOR DOS DADOS ABERTOS

Hoje provavelmente somos levados a considerar sites como o Google e a Amazon pioneiros do big data, mas claro que os governos foram os coletores originais de informações em larga escala e ainda concorrem com qualquer empresa privada pelo volume de dados que controlam. A grande diferença entre os detentores de dados no setor privado e público é que os governos geralmente podem obrigar as pessoas a dar informações, sem ter de convencê-las ou oferecer algo em troca. Como consequência, os governos continuarão a ter grandes quantidades de dados.

As lições do big data se aplicam tanto ao setor público quanto às empresas privadas: o valor dos dados governamentais é latente e requer inovadora análise para serem aproveitados. Contudo, a despeito de sua posição especial na coleta de informações, os governos têm sido ineficientes no seu uso. Recentemente ganhou destaque a ideia de que a melhor forma de extrair valor de dados governamentais é dar ao setor privado e à sociedade em geral acesso para que tentem. Há também um princípio por trás disso. Quando um país reúne dados, o faz em nome dos cidadãos e, portanto, tem de dar acesso à sociedade (exceto em alguns casos, como quando põe em risco a segurança nacional ou a privacidade alheia).

Essa ideia gerou inúmeras iniciativas de "abertura dos dados governamentais" no mundo. Com o argumento de que os governos são apenas guardiões das informações que coletam e de que o setor privado e a sociedade serão mais criativos, os defensores dos dados abertos conclamam as autoridades a liberar publicamente os dados para objetivos civis e comerciais. Para funcionar, claro, os dados precisam ser padronizados a fim de que possam ser processados com facilidade. De outra forma, as informações podem ser consideradas públicas somente no nome.

A ideia dos dados governamentais abertos ganhou impulso depois que o presidente Barack Obama, no primeiro dia no cargo, em 21 de janeiro de 2008, assinou um decreto presidencial que ordenava que as agências federais liberassem o máximo de dados possível. "Diante da dúvida, prevalece a abertura", instruiu ele. Foi uma declaração incrível, particularmente se comparada com a atitude de seu antecessor, que instruiu as agências federais a fazer exatamente o oposto. A orientação de Obama gerou a criação de um site chamado data.gov, repositório de informações abertas do governo federal americano. O site passou de 47 bancos de dados em 2009 para quase 450 mil em 172 agências em seu terceiro aniversário, em julho de 2012.

Mesmo na reticente Grã-Bretanha, onde boa parte das informações governamentais pode ser protegida pela Lei de Direito Autoral da Coroa e onde o licenciamento tem sido difícil e caro (como no caso dos códigos postais para empresas de mapas on-line), tem havido substancial progresso. O governo do Reino Unido criou regras para encorajar a abertura das informações e apoiou a criação do Open Data Institute, codirigido por Tim Berners-Lee, inventor da internet, a fim de promover novos usos dos dados abertos e novas maneiras de liberá-lo do controle estatal.

A União Europeia também anunciou iniciativas de abertura de dados que logo se tornarão continentais. Outros países, como Austrália, Brasil, Chile e Quênia, lançaram e implementaram estratégias de abertura de dados. Abaixo do nível nacional, várias cidades e municípios do mundo também acalentaram a ideia dos dados abertos, assim como organizações multinacionais, como o Banco Mundial, que disponibilizou centenas de bancos de dados econômicos e indicadores sociais, antes restritos.

Paralelamente, comunidades de desenvolvedores web e pensadores visionários se formaram em torno dos dados para descobrir meios de extrair o máximo deles, como a Code for America e a Sunlight Foundation, nos Estados Unidos, e a Open Knowledge Foundation, na Grã Bretanha.

Um dos primeiros exemplos das possibilidades dos dados abertos vem de um site chamado FlyOnTime.us. Os visitantes do site podem descobrir interativamente (entre várias outras correlações) com que probabilidade o inclemente clima atrasará voos em determinado aeroporto. O site combina informações do voo e do clima de fontes oficiais disponíveis na internet gratuitamente. Ele foi desenvolvido por defensores dos dados abertos para mostrar a utilidade das informações recolhidas pelo governo federal. Até mesmo o código do site é aberto, de modo que outras pessoas possam aprender com ele e reutilizá-lo.

O FlyOnTime.us permite que os dados interajam, e eles geralmente revelam informações surpreendentes. Podemos ver que, no caso de voos de Boston para o aeroporto LaGuardia, de Nova York, os viajantes devem estar preparados para o dobro de atrasos causados por neblina que por neve. Provavelmente não era o que pensava a maioria das pessoas na sala de embarque; a neve parecia um motivo mais forte para o atraso. Mas é esse tipo de descoberta que o big data possibilita quando combinamos dados históricos de atrasos do Departamento de Transporte com informações atuais dos aeroportos do Departamento Federal de Aviação, relatórios de clima do Departamento Nacional de Oceanos e Atmosfera e as condições climáticas em tempo real do Serviço Nacional de Clima. O FlyOnTime.us destaca como uma entidade que não coleta ou controla informações, como um sistema de busca ou grande varejista, ainda pode obter e usar dados para agregar valor.

AVALIANDO O QUE NÃO TEM PREÇO

Aberto ao público ou trancado em cofres corporativos, o valor dos dados é difícil de mensurar. Considere os eventos da sexta-feira, 18 de maio de 2012. Neste dia, o fundador do Facebook, Mark Zuckerberg, de 28 anos, simbolicamente tocou o sino de abertura do pregão da NASDAQ da sede da empresa, em Menlo Park, Califórnia. A maior rede social do mundo – que se orgulha de ter 1 em cada 10 pessoas do planeta como membro – começou vida nova como empresa de capital aberto. As ações imediatamente aumentaram 11%, como acontece com muitas ações de tecnologia no primeiro dia de pregão. Mas depois algo estranho aconteceu. As ações do Facebook começaram a cair. Tampouco ajudou o fato de ter havido um problema técnico com os computadores da NASDAQ, que temporariamente interrompeu as atividades. Havia um problema maior à frente. Pressentindo-o, os signatários das ações, liderados pelo banco Morgan Stanley, entraram no mercado de modo que as ações permanecessem acima do valor de lançamento.

Na noite anterior, os bancos do Facebook estabeleceram um valor de US$38 por ação, o que significa um valor de mercado de US$104 bilhões. (Para fins de comparação, este é o valor de marcado da Boeing, General Motors e Dell Computers juntas.) Qual era o valor real do Facebook? Em seu balanço financeiro auditado de 2011, a partir do qual os investidores avaliaram a empresa, o Facebook relatou bens no valor de US$6,3 bilhões, que representavam o valor dos computadores, equipamentos de escritório e outros bens físicos. E quanto ao valor das informações que o Facebook mantinha em cofres? Basicamente zero. Elas não estavam incluídas no balanço, ainda que a empresa não fosse praticamente nada *além de* dados.

A situação se torna ainda mais estranha. Doug Laney, vice-presidente de pesquisa da Gartner, empresa de pesquisa em mercado, analisou os números do período anterior à IPO e descobriu que o Facebook havia coletado 2,1 trilhões de "conteúdos monetários" entre 2009 e 2011, como cliques no botão "curtir", material publicado e comentários. Em comparação com a avaliação da IPO, isso significa que cada item, considerado como ponto de dado único, tinha um valor aproximado de US$0,05. Outra análise é considerar que cada usuário do Facebook valia em torno de US$100, uma vez que os usuários são a fonte da informação que o Facebook coleta.

Como explicar a enorme diferença entre o valor contábil do Facebook (US$6,3 bilhões) e o valor inicial de mercado (US$104 bilhões)? Não há uma boa maneira para explicar isso. Geralmente, concorda-se que o método atual de determinar o valor de uma empresa pela análise de seu "valor real" (isto é, o valor de seus bens físicos) já não reflete o valor verdadeiro. Na verdade, a lacuna entre o valor "real" e o de mercado – quanto a empresa arrecadaria com as ações caso se tornasse empresa de capital aberto – aumenta há décadas. O Senado dos Estados Unidos até mesmo promoveu audiências nos anos 2000 quanto à modernização das regras financeiras que surgiram nos anos 1930, quando empresas com base em informações mal existiam. O tema afeta mais que apenas o balanço de uma empresa: a inabilidade de avaliar adequadamente o valor da empresa gera riscos e volatilidade nos mercados.

A diferença entre o valor contábil e o de mercado de uma empresa é o que chamamos de "bens intangíveis", que passou de cerca de 40% do valor de empresas de capital aberto nos Estados Unidos em meados dos anos 1980 para 75% no início do novo milênio. É uma diferença e tanto. Entre os bens intangíveis estão a marca, o talento, a estratégia – qualquer fator que não seja físico nem faça parte do sistema de contabilidade formal. E cada vez mais os bens intangíveis tratam também dos dados que as empresas detêm e usam.

Em última análise, atualmente não há uma maneira óbvia de avaliar os dados. No dia em que as ações do Facebook foram lançadas, a lacuna entre os bens formais e o valor intangível beirou US$100 bilhões. É ridículo. Mas lacunas como essa devem e vão se estreitar à medida que empresas encontrarem uma maneira de registrar seus bens de dados nos balanços financeiros.

Os passos nessa direção ainda são pequenos. Um executivo sênior de uma das maiores operadoras de celulares dos Estados Unidos confidenciou que a empresa reconhecia o imenso valor de seus dados e que estudava se deveria tratá-los como bem corporativo em termos contábeis formais. Mas assim que os advogados da empresa ouviram falar da iniciativa, a impediram. Colocar os dados nos balanços faz a empresa

se tornar legalmente responsável por eles, argumentaram os advogados, o que não era uma boa ideia.

Enquanto isso, investidores também começaram a perceber o valor dos dados. Os preços das ações podem inflar para empresas que detêm dados e que os coletam facilmente, enquanto outras, em posições menos afortunadas, podem testemunhar uma queda no valor das ações. Os dados não têm de aparecer formalmente nos livros contábeis para que isso aconteça. Os mercados e investidores levaram em conta os bens intangíveis em suas avaliações – a despeito da dificuldade, como atesta a volatilidade do preço das ações do Facebook nos primeiros meses. Mas depois que as preocupações e responsabilidades contábeis forem suavizadas, é quase certo que o valor dos dados aparecerá no balanço corporativo e surgirá como uma nova espécie de bem.

Como os dados serão avaliados? Calcular seu valor não significa mais apenas agregar o que foi ganho com o uso primário. Mas se o maior valor dos dados é latente e deriva de desconhecidos usos secundários futuros, não fica claro como se pode estimá-lo, semelhante à dificuldade de precificar os derivativos financeiros antes do desenvolvimento da equação Black-Scholes, nos anos 1970, ou de avaliar patentes, ramo no qual leilões, trocas, vendas privadas, licenciamento e muito litígio estão aos poucos criando um mercado do conhecimento. De qualquer modo, precificar os dados de fato representa ótima oportunidade para o setor financeiro.

Uma maneira de começar é analisar as estratégias diferentes que os detentores de dados aplicam para extrair valor deles. A possibilidade mais óbvia é o uso pela própria empresa. Mas é improvável que uma empresa seja capaz de extrair todo o valor latente dos dados. De forma mais ambiciosa, é possível licenciar os dados a terceiros. Na era do big data, muitos detentores de dados talvez queiram optar por um acordo que lhes dê uma porcentagem do valor extraído dos dados em vez de uma taxa fixa, semelhante à maneira como o pagamento de royalties de direitos autorais pelas vendas de livros, músicas ou filmes para autores e artistas ou aos acordos de propriedade intelectual na biotecnologia, na qual os licenciadores podem exigir royalties sobre quaisquer invenções subsequentes que surjam da tecnologia. Assim, todos os envolvidos têm um incentivo para maximizar o valor extraído da utilização dos dados.

Mas como o licenciado talvez não consiga extrair todo o valor dos dados, os detentores talvez não queiram dar acesso exclusivo a seus bens. Em vez disso, a "promiscuidade de dados" talvez se torne a norma. Assim, os detentores de dados podem diminuir o risco de suas apostas.

Vários mercados têm tentado experimentar a avaliação de dados. O DataMarket, fundado na Islândia, em 2008, dá acesso a bancos de dados gratuitos de outras fontes, como as Nações Unidas, o Banco Mundial e o Eurostat, e ganha dinheiro revendendo dados para provedores comerciais, como empresas de pesquisa de mercado. Outras *startups* tentaram ser intermediários das informações, plataformas para terceiro que compartilham os dados gratuitamente ou pelo pagamento de uma taxa. A ideia é permitir que todos vendam os dados disponíveis em seus bancos, assim como o eBay é uma plataforma para que as pessoas vendam objetos sem uso que guardam no sótão. A Import.io encoraja as empresas a licenciar dados que, de outra forma, seriam "eliminados" da rede ou usados gratuitamente. E a Factual, fundada por um ex-funcionário

da Google, Gil Elbaz, está disponibilizando bancos de dados que demoram para ser compilados.

A Microsoft entrou na arena com o Windows Azure Marketplace, que busca se voltar para dados de alta qualidade e supervisiona o que oferece, do mesmo modo que a Apple supervisiona as ofertas de sua loja de aplicativos. Na visão da Microsoft, um executivo de marketing que trabalha numa tabela do Excel talvez queira comparar os dados internos da empresa com a previsão do crescimento do PIB de uma consultoria econômica. Assim, ele clica em vários links para comprar dados, que instantaneamente fluem para a coluna na tela.

Até agora não é possível prever qual será o desempenho dos modelos de avaliação. Mas certamente há economias girando em torno de dados – e muitos outros envolvidos tirarão proveito deles, enquanto vários antigos envolvidos provavelmente encontrarão outro sentido para a vida. "Os dados são uma plataforma", nas palavras de Tim O'Reilly, editor de tecnologia e profundo conhecedor do Vale do Silício, uma vez que é a base de novos e bons modelos de negócio.

O segredo do valor dos dados está na aparente ilimitada reutilização em potencial: o custo/benefício. Coletar dados é essencial, mas não o bastante, já que a maior parte do valor dos dados está no uso, não na detenção. No capítulo seguinte, examinamos como os dados estão sendo usados e os empreendimentos de big data que estão surgindo.

Capítulo 7

Implicações

Em 2001, uma inteligente *startup* de Seattle, chamada Decide.com, abriu suas portas virtuais com ousadas ambições: ser o sistema de previsão de preços para milhões de produtos. Mas o plano era começar modestamente: com todos os aparelhos tecnológicos possíveis, de celulares e televisores a câmeras digitais. Os computadores obtinham dados de sites de *e-commerce* e vasculhavam a internet para obter qualquer outro preço e informações sobre os produtos que pudessem encontrar.

Os preços na internet mudam constantemente ao longo do dia e são atualizados dinamicamente com base em infinitos e complicados fatores. Assim, a empresa precisava coletar preços o tempo todo. Não se trata somente de "grandes dados", mas também de "grandes textos", uma vez que o sistema tinha de analisar as palavras para reconhecer quando um produto saía de linha ou um novo modelo estava prestes a ser lançado, informação que os consumidores tinham de saber e que afetava os preços.

Um ano mais tarde, a Decide.com analisava 4 milhões de produtos e usava mais de 25 bilhões de observações de preços. O site identificava elementos estranhos que as pessoas nunca "perceberam" antes, como o fato de que o preço de antigos modelos pode temporariamente aumentar quando novos surgem no mercado. A maioria das pessoas compraria o modelo mais antigo por considerá-lo mais barato, mas, dependendo do site no qual compravam, poderiam pagar mais. À medida que as lojas on-line aumentam o uso de sistemas automatizados de preços, a Decide.com é capaz de encontrar preços algorítmicos incomuns, e alertar os consumidores a esperar. As previsões da empresa, de acordo com medições internas, são precisas em 77% das vezes e rendem ao consumidor uma economia média de US$100 por produto.

Superficialmente, a Decide.com parece mais uma entre várias novas empresas que buscam coletar informações de maneira nova e ganhar um dinheiro honesto pelo trabalho. O que diferencia a Decide.com não são os dados: a empresa se baseia em informações que licencia de sites de *e-commerce* e de sobras gratuitas da internet. Tampouco trata-se de conhecimento técnico: o que a empresa faz não é tão complexo a ponto de os engenheiros que trabalham lá serem os únicos do mundo que possam entender. Ao contrário, apesar de a coleta de dados e as habilidades técnicas serem importantes, a essência do que torna a Decide.com especial é a ideia: a empresa tem uma mentalidade de "big data". Ela vislumbrou uma oportunidade e reconheceu que certos dados podiam ser garimpados para revelar valiosos segredos. E, se parece haver semelhanças entre a Decide.com e a Farecast, é por um bom motivo: as duas foram criadas por Oren Etzioni.

No capítulo anterior, notamos que os dados estão se tornando uma nova fonte de valor à medida que são usados para novos objetivos, em grande parte por causa do custo/benefício. A ênfase estava em empresas que coletavam dados. Agora nos ocupamos das que os usam e como elas se encaixam na cadeia de valor da informação. Vamos analisar o que isso significa para as empresas e pessoas, tanto na carreira como na vida cotidiana.

Foram analisados três tipos de empresas de big data, distinguíveis pelo valor que oferecem, como dados, habilidades e ideias.

Empresas de dados são as que os detêm ou pelo menos obtêm acesso a eles, mas talvez não sejam seu principal negócio. Ou melhor, elas não necessariamente têm a habilidade de extrair valor ou gerar ideias criativas. O melhor exemplo é o Twitter, que obviamente conta com uma quantidade enorme de fluxo de dados por seus servidores, mas que recorreu a duas empresas independentes para licenciá-los.

Em segundo lugar estão as habilidades. São, em geral, consultorias de tecnologia, provedores analíticos com conhecimento especializado e que realizam o trabalho, mas que provavelmente não detêm os dados, tampouco a capacidade de elaborar usos mais inovadores para eles, como o Walmart e as Pop-Tarts, varejistas que recorrem a especialistas da Teradata, empresa de análise de dados, para ajudá-los a ter ideias.

Em terceiro lugar está a mentalidade de big data. Para certas empresas, os dados e o *know-how* não são o principal motivo de sucesso. O que as destaca é que seus fundadores e funcionários têm ideias únicas sobre como usar os dados a fim de agregar novas formas de valor. Um exemplo é Pete Warden, o cofundador nerd da Jetpac, que faz recomendações de viagens com base em fotos que os usuários enviam para o site.

Até aqui, os dois primeiros elementos receberam a maior parte da atenção: as habilidades, hoje escassas, e os dados, que parecem profusos. Uma nova profissão surgiu recentemente, a de "cientista dos dados", que combina as habilidades de estatístico, programador, designer de infográfico e contador de histórias. Em vez de se curvar diante de um microscópio para descobrir os mistérios do universo, o cientista de dados analisa bancos de dados a fim de fazer uma descoberta. O McKinsey Global Institute faz sombrias previsões sobre a escassez de cientistas de dados hoje e no futuro (previsões que os cientistas de dados de hoje gostam de citar para se sentir especiais e para aumentar seus salários).

Hal Varian, economista-chefe da Google, ficou famoso por dizer que a estatística é o trabalho mais "atraente" que existe. "Se você quiser fazer sucesso, deverá se sobressair e se tornar imprescindível em uma área considerada onipresente e barata", diz. "Os dados são tão disponíveis e importantes estrategicamente que o mais raro é obter o conhecimento capaz de extrair sabedoria deles. Por isso, os estatísticos, gerentes de bancos de dados e as pessoas da área de tecnologia estarão numa posição realmente fantástica."

Mas todo o foco nas habilidades e o menosprezo à importância dos dados talvez tenham vida curta. À medida que a indústria evolui, a falta de pessoas será superada quando as habilidades se tornarem lugar-comum. Além disso, há uma crença equivocada de que, só porque há muitos dados disponíveis, eles são gratuitos e baratos. Na verdade, os dados são o ingrediente fundamental. Para avaliar o porquê, pense nas várias partes da cadeia de valor do big data e como elas provavelmente mudarão com

o tempo. Para começar, vamos examinar cada categoria individualmente: detentor de dados, especialista em dados e mentalidade de big data.

A CADEIA DE VALOR DO BIG DATA

A matéria-prima do big data é a informação. Assim, faz sentido analisar primeiramente os detentores dos dados. Eles podem não fazer a coleta inicial, mas controlam o acesso à informação e a usam ou licenciam para que outros extraiam valor dela. Por exemplo, a ITA Software, grande rede de reservas de passagens aéreas (depois da Amadeus, Travelport e Sabre), cedeu dados para que a Farecast fizesse sua previsão de preços, mas não fez a análise ela mesma. Por que não? O negócio da ITA usando é o uso dos dados com o objetivo para o qual foram gerados – vender passagens aéreas – e não para usos secundários. Assim, as funções são diferentes. Além disso, ela teria de encontrar uma alternativa para a patente de Etzioni.

A empresa também optou por não explorar os dados por sua localização na cadeia de valor. "A ITA evitou projetos que envolvessem o uso comercial de dados relacionados com a venda de passagens", afirma Carl de Marcken, cofundador da ITA Software e executivo de tecnologia da empresa. "A ITA teve acesso especial a esses dados, necessários para que a empresa fizesse o que se propõe, e não poderia pôr isso em risco." Assim, ela delicadamente se distanciou ao licenciar os dados, mas sem usá-los. Como resultado, a ITA ganhou pouco. A maior parte do valor secundário dos dados foi para a Farecast: para os clientes na forma de passagens mais baratas e para os funcionários e proprietários com o rendimento com anúncios, comissões e finalmente a venda da empresa.

Algumas empresas inteligentemente se posicionaram no centro do fluxo da informação de modo que pudessem alcançar escala e captar o valor dos dados. É o caso da indústria de cartões de crédito nos Estados Unidos. Durante anos, o alto custo da batalha contra as fraudes levou vários bancos, de pequeno e médio portes, a evitar lançar cartões de crédito e a voltar as operações para instituições financeiras maiores, que tinham o tamanho e a escala para investir na tecnologia. Empresas como o Capital One e o MBNA, do Bank of America, aceitaram a incumbência. Mas hoje os pequenos bancos se arrependem da decisão, porque a terceirização das operações os priva de dados sobre padrões de gastos, que lhes permitiriam saber mais sobre clientes, de modo que pudessem lhes vender serviços customizados.

Em vez disso, os grandes bancos e operadoras de cartão, como Visa e MasterCard, parecem ocupar um bom lugar na cadeia de valor da informação. Ao atender vários bancos e varejistas, eles obtêm mais transações em sua rede e podem usá-las para prever o comportamento do consumidor. O modelo de negócios passa do simples processamento para a coleta de dados. A questão está no que fazer com isso.

A MasterCard podia licenciar os dados para terceiros que extrairiam o valor deles, como fez a ITA, mas a empresa prefere fazer a análise sozinha. Um divisão chamada MasterCard Advisors agrega e analisa 65 bilhões de transações dos 1,5 bilhão de clientes em 210 países a fim de prever tendências corporativas e de consumo. Depois, a empresa vende a informação. Ela descobriu, entre outros aspectos, que, se as pessoas enchem o tanque de gasolina por volta das 16 horas, é provável que gastem de

US$35 a US$50 na hora seguinte em uma mercearia ou restaurante. Um publicitário pode usar a informação para criar cupons de estabelecimentos próximos no verso dos recibos de postos de gasolina, a essa hora do dia.

Como intermediária do fluxo de informação, a MasterCard está em ótima posição para coletar dados e extrair valor deles. Pode-se imaginar um futuro no qual as empresas de cartão de crédito desprezem a comissão das transações e as processem gratuitamente em troca de mais acesso a dados, ganhando dinheiro com a venda de sofisticadas análises.

A segunda categoria consiste em especialistas em dados: empresas com o conhecimento ou a tecnologia para fazer complexas análises. A MasterCard optou por fazer isso internamente, e algumas empresas migram entre categorias. Mas muitas outras recorrem a especialistas. A consultoria Accenture, por exemplo, trabalha com empresas de vários segmentos para empregar avançadas tecnologias de sensores sem fio e analisar os dados coletados. Num projeto-piloto com a cidade de St. Louis, no Missouri, a Accenture instalou sensores sem fio em vários ônibus públicos para monitorar os motores e prever quebras ou determinar a hora mais adequada para realizar a manutenção. Esse procedimento diminuiu o custo em até 10%. Só uma descoberta – a de que a cidade podia atrasar a troca de determinada peça a cada 320 mil ou 400 mil quilômetros para 450 mil quilômetros – economizou mais de US$1 mil por veículo. O cliente, não a consultoria, extraiu valor dos dados.

No setor de dados médicos, vemos outro claro exemplo de como empresas fora do ramo tecnológico podem prover serviços úteis. O MedStar Washington Hospital Center, em Washington, D.C., junto com a Microsoft Research e com o uso do software Amalgama, da Microsoft, analisou vários anos de registros médicos anônimos – dados demográficos dos pacientes, exames, diagnósticos, tratamentos e outros – a fim de reduzir as taxas de readmissão e infecções, alguns dos elementos mais caros do tratamento de saúde, de modo que qualquer fator capaz de diminuir essas taxas se traduz em enorme economia.

A técnica revelou correlações surpreendentes. Um resultado foi uma lista de todas as doenças que criavam oportunidades para que pacientes liberados voltassem a ser internados depois de um mês. Algumas são conhecidas: é um quadro difícil de tratar. Mas o sistema também encontrou outro inesperado sinal: o estado mental do paciente. A probabilidade de uma pessoa voltar um mês depois de receber alta aumentava se a reclamação inicial contivesse palavras que sugerissem um transtorno mental como "depressão".

Apesar de essa correlação não estabelecer relação de causalidade, sugere que uma intervenção pós-alta voltada para a saúde mental do paciente possa melhorar também sua saúde física, reduzindo readmissões e diminuindo custos. A descoberta foi feita por uma máquina, após a análise de grande quantidade de dados. Uma pessoa que incansavelmente estudasse os dados talvez nunca fizesse a mesma descoberta. A Microsoft não controlava os dados, que pertenciam ao hospital, tampouco teve uma ideia brilhante; não era o fundamental aqui. Em vez disso, ela ofereceu o software Amalga a fim de revelar a informação.

Empresas detentoras de big data contam com especialistas para extrair valor dos dados. Mas apesar dos altos preços e de funções com nomes curiosos, como "ninja dos dados", a vida dos especialistas técnicos nem sempre é tão glamorosa quanto

parece. Eles garimpam os dados e levam para a casa um belo contracheque, mas entregam as pepitas que encontram para aqueles que detêm os dados.

O terceiro grupo é composto por empresas e pessoas com mentalidade de big data, que veem as oportunidades antes dos outros – mesmo que não possuam dados ou habilidades para agir em relação às oportunidades. Na verdade, precisamente por serem observadores externos, suas mentes são livres: eles enxergam o possível em vez de se limitarem à sensação do que é exequível.

Bradford Cross personifica o significado de uma mentalidade de big data. Em agosto de 2009, com pouco mais de 20 anos, ele e alguns amigos criaram a FlightCaster.com. Assim como a FlyOnTime.us, a FlightCaster previa se determinado voo nos Estados Unidos provavelmente se atrasaria. Para fazer as previsões, eles analisavam todos os voos dos 10 anos anteriores e os comparavam com dados antigos e atuais sobre o clima.

O interessante é que os próprios detentores dos dados não tinham essa possibilidade. Nenhum tinha incentivo – ou a obrigação regulatória – para usar os dados dessa forma. Na verdade, se as fontes dos dados – o U.S. Bureau of Transportation Statistics (Departamento de Estatísticas de Transporte dos Estados Unidos), o Federal Aviation Administration (Departamento Federal de Aviação) e o National Weather Service (Serviço de Clima) – ousassem prever os atrasos dos voos comerciais, o Congresso provavelmente promoveria audiências, e cabeças de burocratas rolariam. As companhias aéreas não podiam – ou não iriam - fazer isso. Elas se aproveitam de manter seu desempenho o mais obscuro possível. Para tanto, foram necessários alguns engenheiros encapuzados. Na verdade, as previsões da FlightCaster eram tão precisas que até mesmo os funcionários das companhias aéreas começaram a usá-las: as empresas só queriam anunciar atrasos no último minuto, de modo que, se eram a principal fonte de informação, por outro lado, não eram a melhor fonte.

Por causa da mentalidade de big data – a inspirada percepção de que dados disponíveis ao público podiam ser processados de forma a oferecer respostas pelas quais milhões de pessoas ansiavam –, a FlightCaster de Cross foi pioneira, mas por pouco. No mês de inauguração da empresa, os nerds por trás da FlyOnTime.us começaram a reunir os dados abertos para construir seu site. A vantagem de que a FlightCaster dispunha em pouco tempo acabaria. Em janeiro de 2011, Cross e seus sócios venderam a empresa para a Next Jump, companhia que gerencia programas de descontos corporativos com o uso de técnicas de big data.

Então Cross se voltou para outro ramo no qual vira um nicho que alguém de fora poderia explorar: a mídia. Sua empresa, Prismatic, agrega e hierarquiza conteúdo da internet com base em análises de textos, preferências dos usuários, popularidade associada às redes sociais e análise de big data. Mais importante, o sistema não distingue um texto de um blog adolescente, um site corporativo e um artigo no *Washington Post:* se o conteúdo for considerado relevante e popular (de acordo com acessos e compartilhamento), aparece no alto da tela.

Como serviço, a Prismatic é um reconhecimento de como a geração mais nova está interagindo com a mídia. Para eles, a fonte de informação perdeu importância. Trata-se de um mero lembrete para os papas da mídia de que o público agrega mais conteúdo que eles e de que jornalistas bem vestidos têm de disputar lugar contra blogueiros de

roupões. Mas o ponto central é que é difícil imaginar que a Prismatic teria surgido da própria indústria da mídia, mesmo que colete muita informação. Os frequentadores do National Press Club nunca pensaram em reutilizar dados on-line sobre o consumo midiático. Nem mesmo os analistas de Armonk, Nova York, ou Bangalore, Índia, têm usado a informação dessa maneira. Foi necessário que Cross, observador externo, com cabelos embaraçados e fala lenta, presumisse que, ao usar dados, ele podia dizer ao mundo - melhor que os editores do *The New York Times* - no que as pessoas deviam prestar atenção.

A noção de mentalidade de big data e o papel de um observador externo criativo, com uma ideia brilhante, não são muito diferentes do que aconteceu no início do *e-commerce,* em meados dos anos 1990, quando os pioneiros estavam livres da rígida mentalidade ou dos limites institucionais de empresas de setores mais antigos. Foi assim que um fundo de hedge, e não a Barnes & Noble, fundou uma livraria on-line (a Amazon, de Jeff Bezos). Um desenvolvedor de software, e não a Sotheby, criou um site de leilões (o eBay, de Pierre Omidyar). Hoje os empreendedores com mentalidade de big data nem sempre os têm quando começam. Mas, por causa disso, tampouco têm os interesses ocultos ou os desincentivos financeiros que podem impedi-los de pôr suas ideias em prática.

Como vimos, há casos nos quais uma empresa combina várias dessas características. Etzioni e Cross talvez tenham tido ideias antes dos outros, mas também possuíam as habilidades. Os operários da Teradata e da Accenture não apenas batem o ponto; também são conhecidos por contribuírem com ideias de tempos em tempos. Mas os arquétipos são úteis como forma de apreciar os papéis que as diferentes empresas exercem. Os pioneiros de hoje geralmente têm origens distintas e exercem suas habilidades em várias áreas. Uma nova geração de investidores-anjo e empreendedores está surgindo, principalmente entre ex-funcionários da Google e da chamada PayPal Mafia (os ex-líderes da empresa Peter Thiel, Reid Hoffman e Max Levchin). Junto com um alguns cientistas da computação acadêmicos, eles são alguns dos principais apoiadores por trás das empresas de dados de hoje em dia.

A visão criativa de pessoas e empresa na cadeia alimentar do big data nos ajuda a reavaliar o valor das empresas. A Salesforce.com, por exemplo, talvez não seja apenas uma plataforma útil para que as empresas hospedem os aplicativos corporativos: ela também existe para extrair valor dos dados que fluem por sua infraestrutura. Empresas de celulares, como vimos no capítulo anterior, coletam quantidades assombrosas de dados, mas geralmente fecham os olhos para seu valor. Elas poderiam, contudo, licenciá-los a outras empresas capazes de extrair valor – assim como o Twitter decidiu ceder os direitos para licenciar seus dados a duas empresas.

Alguns empreendimentos de sucesso caminham por diferentes territórios como uma estratégia consciente. A Google coleta dados, como erros em buscas, tem a brilhante ideia de usá-los para criar provavelmente o melhor corretor ortográfico do mundo e tem as habilidades para executar a tarefa extraordinariamente bem. A Google também tira proveito, com muitas outras atividades, da integração vertical na cadeia de valor do big data, na qual ocupa as três posições de uma só vez. Ao mesmo tempo, a empresa também disponibiliza alguns dados para outras por meio de APIs (Application Programming Interfaces), de modo que possam ser reutilizados e agregar ainda mais

valor. Um exemplo são os mapas da Google, usados de graça na internet por todos os tipos de empresas, desde imobiliárias até sites governamentais, ainda que sites com muito tráfego tenham de pagar.

A Amazon também tem a mentalidade, o conhecimento e os dados. Na verdade, a empresa construiu seu modelo de negócios nessa ordem, contrário à norma. Inicialmente, ela teve apenas a ideia do celebrado sistema de recomendações. A prospecção no mercado de ações de 1997 descrevia a "filtragem colaborativa" antes que a Amazon soubesse como funcionaria na prática e tivesse dados suficientes para torná-la útil.

Tanto a Google quanto a Amazon expandem as categorias, mas as estratégias são diferentes. Quando a Google coleta qualquer tipo de dado, visa as utilidades secundárias. Os carros Street View, como vimos, coleta informação de GPS não apenas para o serviço de mapas mas também para treinar carros automatizados. Por outro lado, a Amazon está mais focada no uso primário dos dados e só os reaproveita marginalmente. Seu sistema de recomendação, por exemplo, conta com um fluxo de dados como indício, mas a empresa não tem usado a informação para feitos extraordinários, como prever o estado da economia ou epidemias de gripe.

Apesar de o e-reader Kindle da Amazon ser capaz de mostrar uma página cheia de anotações e sublinhada pelos usuários, a empresa não vende as informações para autores e editores. Os publicitários adorariam descobrir quais passagens são mais populares e usar o conhecimento para vender melhor os livros. Os autores talvez gostem de saber em que ponto os leitores pararam a leitura e poderiam usar a informação para melhorar o trabalho. Os editores poderiam encontrar os temas que dariam origem ao próximo livro de sucesso. Mas a Amazon parece deixar os dados inaproveitados.

Usado com inteligência, o big data pode transformar os modelos de negócios das empresas e a maneira como os sócios interagem. Num caso marcante, uma grande montadora automotiva europeia mudou sua relação comercial com os fornecedores de peças pelo uso de dados de que o fornecedor não dispunha. (Como o caso nos foi relatado anonimamente por uma das principais empresas envolvidas, infelizmente não podemos revelar nomes.)

Os carros de hoje estão repletos de chips, sensores e softwares que enviam dados de desempenho para os computadores dos fabricantes de carros quando o veículo passa por revisão. Um veículo médio tem hoje cerca de 40 microprocessadores; todos os itens eletrônicos de um carro correspondem a um terço do custo, o que os torna os sucessores dos navios que Maury chamou de "observatórios flutuantes". A capacidade de reunir dados sobre o uso dos componentes na estrada – e para reincorporá-los a fim de melhorar os veículos – está se tornando grande vantagem competitiva para as empresas que detêm a informação.

Ao trabalhar com uma empresa de análise terceirizada, a montadora foi capaz de ver que o sensor no tanque de combustível, criado por um fornecedor alemão, estava ruim e gerara vários alarmes falsos para cada um válido. A empresa podia ter cedido a informação ao fornecedor e pedido um ajuste. Numa era mais cortês dos negócios, talvez ela tivesse feito justamente isso. Mas a empresa vinha gastando uma fortuna com o programa de análise. Ela queria usar a informação para recuperar um pouco do investimento.

A empresa pensou nas alternativas. Deveria vender os dados? Como a informação seria avaliada? E se o fornecedor se sentisse traído e a empresa ficasse com um componente ruim? Ela sabia que, se fornecesse a informação, peças semelhantes dos carros da concorrência também seriam aperfeiçoadas. Garantir que apenas os próprios carros fossem aperfeiçoados parecia uma estratégia inteligente. Por fim, a empresa teve uma ideia: encontrou uma maneira de melhorar a peça sem modificar o software, recebeu uma patente pela técnica e depois vendeu-a ao fornecedor e ganhou um bom dinheiro com isso.

OS INTERMEDIÁRIOS DOS NOVOS DADOS

Quem detém o maior valor na cadeia de valor do big data? Hoje a resposta parece ser aqueles que têm a mentalidade, as ideias inovadoras. Como aprendemos com a era pontocom, aqueles que saem na frente podem mesmo prosperar. Mas a vantagem pode não durar muito. À medida que a era do big data avança, outros adotarão a mentalidade, e a vantagem dos pioneiros diminuirá, relativamente falando.

Então, talvez a essência esteja na habilidade técnica? Afinal, uma mina de ouro nada vale se não formos capazes de extrair o ouro. Mas a história da computação sugere o contrário. Hoje em dia, há grande demanda por conhecimento em gerenciamento de banco de dados, ciência de dados, análise, algoritmos e áreas do gênero. Mas, com o tempo, à medida que o big data se tornar parte do cotidiano, os instrumentos melhorarem e se tornarem mais fáceis de usar e as pessoas adquirirem conhecimento, o valor da técnica também diminuirá em termos relativos. Da mesma forma, a capacidade de programação se tornou muito comum entre os anos 1960 e 1980. Hoje, empresas terceirizadas reduziram ainda mais o valor da programação; o que já foi um paradigma da perspicácia intelectual é hoje motor do desenvolvimento nos países pobres, o que não quer dizer que o conhecimento de big data não seja importante. Mas não é a fonte mais essencial de valor, já que uma pessoa pode trazê-lo de fora.

Hoje, no estágio inicial do big data, as ideias e habilidades parecem ter maior valor. Mas, por fim, o maior valor de fato recai sobre os dados em si, porque seremos capazes de fazer mais com as informações e também porque os detentores dos dados apreciarão mais o valor potencial do bem que possuem. Como resultado, eles provavelmente o protegerão mais do que nunca e cobrarão dos de fora um preço maior para lhes conferir acesso. Para dar continuidade à metáfora da mina de ouro: o ouro em si importará mais.

Mas há um importante aspecto da ascensão de longo prazo dos detentores dos dados que merece ser notado. Em alguns casos, surgirão "intermediários de dados", capazes de coletar dados de diversas fontes, agregá-los e inovar com eles. Os detentores dos dados permitirão que os intermediários exerçam esse papel porque parte do valor dos dados só pode ser extraído por meio deles.

Um exemplo é a Inrix, empresa de análise de tráfego com sede perto de Seattle. Ela reúne localização em tempo real de 100 milhões de veículos Nos Estados Unidos e Europa. Os dados vêm de carros da BMW, Ford e Toyota, entre outras, além de frotas comerciais, como táxis e vans de entrega. A empresa também obtém dados dos celulares dos motoristas (seus aplicativos gratuitos para *smartphones* são importantes aqui: os usuários recebem informações sobre o tráfego, a Inrix recebe as coordenadas). A empresa combina essa

informação com dados históricos, clima e outros fatores, como eventos locais, para prever como o tráfego fluirá. O produto da linha de montagem de dados é transmitido para os sistemas de navegação dos carros e usado por frotas oficiais e comerciais.

A Inrix é o exemplo perfeito do intermediário independente de dados. Ela coleta informações de várias empresas automobilísticas rivais e gera um produto mais valioso que qualquer uma delas seria capaz de gerar por si mesma. Cada fabricante pode ter alguns milhões de pontos de dados de seus veículos na estrada. Apesar de poderem ser usadas para prever o fluxo do tráfego, as previsões não seriam muito precisas ou completas. A qualidade aumenta à medida que aumentam os dados. Além disso, as empresas talvez não tenham o conhecimento técnico: são mais competentes em dobrar metal, não em refletir sobre distribuições de Poisson.* Assim, todas têm um incentivo para ceder a função a um terceiro. Além do mais, apesar de o tráfego ser importante para os motoristas, ele não influencia na compra de determinado carro. Assim, a concorrência não se importa em juntar forças dessa maneira.

Claro que empresas de vários ramos já compartilhavam informações, principalmente laboratórios de subscritores de seguros e setores em rede, como bancos, energia e telecomunicações, nos quais a troca de informações é essencial para evitar problemas e as agências regulares às vezes a exigem. Empresas de pesquisas de mercado têm agregado dados há décadas, assim como empresas voltadas para funções específicas, como a auditoria de circulação de jornais. Para algumas associações, os dados agregados são fundamentais para sua atuação. A diferença hoje é que os dados são agora matéria-prima no mercado; um bem independente do objetivo inicial de mensuração. As informações da Inrix, por exemplo, são mais úteis do que pode parecer superficialmente. Sua análise de tráfego é usada para medir a saúde das economias locais porque oferece indícios sobre desemprego, vendas e lazer. Quando a economia dos Estados Unidos começou a entrar em ebulição, em 2011, os indícios foram detectados por análises de tráfego, a despeito das negativas políticas do que estava acontecendo: os horários de rush estavam menos intensos, o que sugeria maior taxa de desemprego. A Inrix também vendeu dados para um fundo de investimentos que usava os padrões de tráfego perto de uma importante loja varejista como indicativo das vendas, algo que o fundo usa para comercializar as ações da empresa antes dos anúncios trimestrais de faturamento. Mais carros na área significam mais vendas.

Outros intermediários estão surgindo dentro da cadeia de valor do big data. Um dos primeiros *players* foi a Hitwise, mais tarde comprada pela Experian, que faz acordos com provedores de internet para usar dados em troca de receita extra. Os dados são licenciados por uma taxa fixa em vez de uma porcentagem do valor gerado. A Hitwise captava boa parte do valor como intermediária. Outro exemplo é a Quantcast, que ajuda a medir o tráfego on-line para que os sites saibam mais sobre seu público-alvo e padrões de utilização. Ela fornece uma ferramenta on-line, de modo que os sites possam registrar visitas; em troca, a empresa pode ver os dados, o que permite que melhorem os anúncios.

**Nota do Tradutor*: Distribuição de probabilidade discreta que expressa a probabilidade de certo número de eventos ocorrer num dado período, caso ocorra com uma taxa média conhecida e caso cada evento independa do tempo decorrido desde o último.

Os novos intermediários identificaram um lucrativo nicho de mercado sem ameaçar os modelos de negócios dos detentores dos quais obtêm os dados. Por enquanto, os anúncios virtuais são um desses nichos, já que é onde se encontra a maioria dos dados e onde há maior necessidade de garimpar o público-alvo. Mas, à medida que os dados se tornam mais dataficados, e mais ramos percebem que seus negócios aprendem com os dados, os intermediários de informação independentes surgirão em todos os lugares.

Alguns intermediários podem não ser empreendimentos comerciais, e sim instituições sem fins lucrativos. O Health Care Cost Institute, por exemplo, foi criado em 2012 por algumas das maiores seguradoras de saúde dos Estados Unidos. Os dados combinados totalizavam cinco bilhões de reclamações (anônimas) envolvendo 33 milhões de pessoas. Compartilhar os registros permitiu que as empresas encontrassem tendências não identificáveis em bancos de dados menores. Entre as primeiras descobertas estava a de que os custos médicos nos Estados Unidos aumentaram três vezes mais rapidamente que a inflação entre 2009 e 2010, mas com significativas diferenças: os preços das salas de emergência subiram 11%, enquanto as instalações de maternidade tiveram uma deflação. Claro que as seguradoras não entregariam seus valiosos dados para outra empresa que não uma instituição sem fins lucrativos. As motivações de uma instituição desse tipo são menos suspeitas, e as empresas podem ser projetadas com foco em transparência e responsabilidade.

A variedade de empresas de big data mostra que o valor da informação está mudando. No caso da Decide.com, os dados são fornecidos por sites parceiros com base na divisão de receita. A Decide.com ganha comissões quando as pessoas compram bens pelo site, mas as empresas que fornecem os dados também recebem parte do valor. Essa prática sugere maturidade na maneira como a indústria trabalha com dados: a ITA não recebeu comissão pelos dados que forneceu à Farecast, somente um valor pelo licenciamento. Hoje, os provedores de dados são capazes de propor termos mais interessantes. Para a próxima *startup* de Etzioni, podemos presumir que ele tentará obter os dados sozinho, já que o valor migrou do conhecimento para a ideia e agora se transfere para os dados.

Modelos de negócios estão sendo destruídos à medida que o valor passa para aqueles que controlam os dados. A montadora europeia que fez o acordo de propriedade intelectual com o fornecedor dispunha de forte análise interna de dados, mas precisava trabalhar com uma tecnologia externa para extrair ideias dos dados. A empresa de tecnologia recebeu pelo trabalho, mas a montadora manteve a maior parte dos lucros. Ao perceber a oportunidade, contudo, a empresa de tecnologia mudou seu modelo de negócios para compartilhar um pouco do risco e da recompensa com os clientes. Ela experimentou trabalhar por uma taxa menor em troca da divisão de parte da quantia que a análise gerava. (Quanto aos fornecedores de peças, é provavelmente mais seguro dizer que todos quererão acrescentar sensores aos produtos ou insistir num acesso a dados de desempenho como padrão do contrato de fornecimento, a fim de continuamente melhorar os componentes.)

Quanto aos intermediários, trabalham arduamente porque precisam convencer as outras empresas do valor do compartilhamento. A Inrix, por exemplo, começou a coletar mais que apenas informações de localização. Em 2012, ela fez uma experiência ao analisar onde e quando o ABS (sistema automatizado de freio) era acionado, para

uma montadora que criara um sistema de telemetria para coletar as informações em tempo real. A ideia era a de que o acionamento frequente do ABS num trecho específico da estrada significaria que as condições ali eram perigosas e que os motoristas deveriam pensar em rotas alternativas. Assim, com esses dados, a Inrix podia recomendar não apenas a rota mais curta, como também a mais segura.

Mas a montadora não planeja compartilhar os dados com outras. Ao contrário, ela insiste que a Inrix empregue o sistema somente nos carros que produz. Acredita-se que o valor do sistema supere o ganho de agregar os dados com outras montadoras a fim de melhorar a precisão geral do sistema. Portanto, a Inrix acredita que, com o tempo, todas as montadoras verão a utilidade de agregar os dados. Como intermediária, a Inrix tem incentivo para esse otimismo: seu negócio se baseia totalmente no acesso a várias fontes de dados.

As empresas também estão experimentando diferentes formas de organização no negócio do big data. A Inrix não se deparou repentinamente com seu modelo de negócios, como muitas *startups* – ela foi planejada para funcionar como intermediária. A Microsoft, dona das patentes essenciais à tecnologia, descobriu que uma pequena empresa independente – e não uma grande empresa, conhecida por suas táticas agressivas – podia ser vista como um parceiro mais neutro e capaz de unir empresas rivais e extrair o máximo de sua propriedade intelectual. Da mesma forma, o MedStar Washington Hospital Center, que usou o software Amalgama, da Microsoft, para analisar as readmissões sabia exatamente o que estava fazendo com seus dados: o sistema Amalga era originalmente o software da emergência do hospital e se chamava Azyxxi, vendido em 2006 para a Microsoft para ser mais bem desenvolvido.

Em 2010, a UPS vendeu sua unidade interna de análise de dados, chamada de UPS Logistics Technologies, para a empresa Thoma Bravo. Hoje chamada Roadnet Technologies, a unidade é mais livre para fazer análises de rotas para mais de uma empresa. A Roadnet coleta dados de vários clientes para fornecer um serviço usado pela UPS e concorrentes. Como UPS Logistics, ela jamais teria persuadido as rivais da empresa a entregar seus bancos de dados, explica o executivo-chefe da Roadnet, Len Kennedy. Mas, depois de se tornar independente, os concorrentes da UPS se sentiram mais à vontade para lhe dar os dados, e, por fim, todos tiram proveito da exatidão dos dados agregados.

Provas de que os dados propriamente, não habilidades e mentalidade, se tornarão mais valiosos podem ser encontradas nas muitas aquisições no setor do big data. Em 2006, por exemplo, a Microsoft recompensou a mentalidade de big data de Etzioni comprando a Farecast por cerca de US$110 milhões. Dois anos mais tarde, a Google pagou US$700 milhões para comprar o fornecedor de dados da Farecast, a ITA Software.

A MORTE DO ESPECIALISTA

No filme *O homem que mudou o jogo,* sobre como o Oakland A's se tornou um time de beisebol vencedor ao aplicar a análise e novos tipos de medição ao jogo, há uma bela cena na qual velhos olheiros discutem a respeito dos jogadores. A plateia não consegue deixar de se irritar, não apenas porque a cena expõe a maneira como as decisões são

tomadas, sem considerar os dados, mas também porque todos já estivemos em situações nas quais a "certeza" se baseava em sentimentos, não na ciência.

"Ele tem um corpo de jogador de beisebol... Um bom rosto", diz um olheiro.

"Ele tem um belo swing. Quando ele bate, arrasa", entusiasma-se um frágil senhor grisalho que usa um aparelho auditivo. "Arrasa mesmo no bastão", concorda outro olheiro.

Um terceiro homem interrompe a conversa e declara: "A namorada dele é feia."

"O que isso significa?", pergunta o olheiro que comanda a reunião.

"Uma namorada feia significa falta de confiança", explica o pessimista.

"OK", diz o líder, satisfeito e pronto para seguir adiante.

Depois de uma brincadeira, um olheiro que estava quieto se manifesta: "Este cara tem atitude. Atitude é bom. Quero dizer, aí está o cara, ele entra na sala e a macheza dele já estava lá havia dois minutos." Outro acrescenta: "Ele passou no teste do olho. Ele tem o biótipo, está pronto para o papel. Só precisa de mais um pouco de tempo de jogo."

"Só estou dizendo", reitera o pessimista, "que a namorada dele é nota 6, no máximo".

A cena retrata com perfeição os problemas do julgamento humano. O que parece um sensato debate não se baseia em nada de concreto. Decisões sobre milhões de dólares em contratos com jogadores são tomadas com base na intuição, sem objetividade. Sim, é apenas um filme, mas na vida real não é muito diferente. O mesmo raciocínio é empregado das salas de reunião de Manhattan ao Salão Oval e cafeterias e mesas de cozinha em todos os lugares.

O homem que mudou o jogo, com base no livro *Moneyball*, de Michael Lewis, conta a história real de Billy Beane, gerente-geral do Oakland A's que desprezou o livro azul de regras de como avaliar os jogadores em favor de métodos matemáticos que analisavam o jogo a partir de um novo conjunto de medidas. Já foi o tempo de estatísticas como "média de rebatidas" e chegou a hora de um raciocínio aparentemente estranho sobre o jogo como "porcentagem de tempo em bases". A abordagem com foco nos dados revelou uma dimensão ao esporte que sempre esteve presente, mas oculta em meio aos amendoins e salgadinhos. Não importava como um jogador chegava a uma base, se com uma rebatida rasteira ou uma caminhada, desde que ele a conquistasse. Quando as estatísticas mostravam que o roubo de bases era ineficiente, lá se ia um dos mais interessantes, mas menos "produtivos", elementos do jogo.

Em meio a considerável controvérsia, Beane sacralizou na sede do time um método conhecido como "sabermetria", termo cunhado pelo escritor Bill James numa referência à Society for American Baseball Research, que até então era território da cultura nerd. Beane estava desafiando o dogma do banco de reservas, assim como a teoria heliocêntrica de Galileu afrontara a autoridade da Igreja Católica. Por fim, ele liderou um time sofrível e chegou à primeira final na American League West na temporada de 2002, incluindo uma série de 20 vitórias. A partir daí, estatísticos substituíram os olheiros como os mais inteligentes do esporte, e vários outros times se puseram a adotar o método.

No mesmo espírito, o maior impacto do big data será o fato de que decisões tomadas com base em dados disputarão ou superarão o julgamento humano. O especialista no

tema vai perder um pouco de seu brilho em comparação com o estatístico ou analista de dados, que não se impressionam com as velhas práticas e deixam os dados se manifestarem. Essa nova casta vai se basear em correlações sem preconceitos, assim como Maury não confiou no que velhos lobos do mar tinham a dizer sobre certa passagem enquanto bebiam cerveja no bar, mas nos dados agregados para revelar verdades práticas.

Estamos testemunhando o fim da influência dos especialistas em muitas áreas. Na mídia, o conteúdo criado e vendido em sites como Huffington Post, Gawker e Forbes é regularmente determinado por dados, não pela opinião de editores. Os dados podem revelar melhor o que as pessoas querem ler que o instinto de um experiente jornalista. A empresa de educação on-line Coursera usa dados sobre que partes de uma videoaula os alunos reproduzem para saber que material pode não estar claro e envia as informações aos professores a fim de que eles possam melhorar. Como já foi mencionado, Jeff Bezos se livrou dos críticos internos da Amazon quando os dados mostraram que as recomendações algorítmicas geravam mais vendas.

Isso significa que as habilidades necessárias para ter sucesso no ambiente de trabalho estão mudando. Essa constatação altera o tipo de contribuição que as empresas esperam dos funcionários. A Dra. McGregor, que cuida de bebês prematuros em Ontário, não precisa ser a médica mais sábia do hospital ou a principal autoridade em cuidados neonatais para gerar os melhores resultados para os pacientes. Na verdade, ela nem mesmo é médica, mas PhD em Ciência da Computação. No entanto, ela tira proveito de dados há mais de uma década, aqueles que os computadores analisam e que ela transforma em recomendações de tratamento.

Como vimos, em geral os pioneiros do big data vêm de ramos externos ao domínio no qual deixaram sua marca. São especialistas em análise de dados, inteligência artificial, matemática ou estatística e aplicam suas habilidades em ramos específicos. Os vencedores do concurso Kaggle, plataforma virtual de projetos de big data, são em geral novos no setor nos quais geram os melhores resultados, explica o executivo-chefe da Kaggle, Anthony Goldbloom. Um médico britânico desenvolveu algoritmos para prever pedidos de seguro e identificar carros usados com problemas. Um atuário de Cingapura liderou uma competição para prever reações biológicas a compostos químicos. Enquanto isso, no grupo de tradução da Google, os engenheiros celebram traduções em idiomas que ninguém na empresa fala. Do mesmo modo, estatísticos na unidade de tradução da Microsoft saboreiam o surgimento de uma nova piada: a qualidade das traduções aumenta sempre que um linguista deixa a equipe.

Para deixar claro, os especialistas não desaparecerão, mas sua supremacia diminuirá. De agora em diante, eles devem dividir o pódio com os nerds do big data, assim como a causalidade deve dividir os holofotes com a correlação. Esse fato transforma a maneira como valorizamos o conhecimento, porque tendemos a pensar que as pessoas especializadas valem mais que os generalistas – que a riqueza favorece a profundidade. Mas a especialidade é como a exatidão: adequada para um mundo de pequenos dados, onde não há informações bastantes ou a informação certa, e assim é preciso confiar na intuição ou na experiência para se orientar. Nesse mundo, a experiência exerce papel fundamental, já que é o acúmulo de conhecimento latente – que não se pode transmitir ou aprender em um livro, talvez nem mesmo ter consciência dele – que permite a melhor tomada de decisão.

Mas quando dispomos de muitos dados, podemos usá-los melhor. Assim, aqueles capazes de analisar o big data podem ultrapassar superstições ou o raciocínio convencional, não porque são mais inteligentes, e sim porque têm os dados. (Os recursos externos são imparciais quanto a rixas dentro do campo de atuação, que podem estreitar a visão de um especialista para um dos lados da disputa.) Isso sugere que a concepção do valor de um funcionário para uma empresa muda. O que você precisa saber muda, quem você precisa conhecer muda e também muda o que você precisa estudar para se preparar para a vida profissional.

A matemática e a estatística, talvez com uma pitada de programação e ciência de rede, serão tão fundamentais para o ambiente de trabalho moderno quanto o conhecimento dos números foi há um século, e a alfabetização, antes disso. No passado, para ser um excelente biólogo, uma pessoa precisava conhecer vários outros biólogos, o que, em parte, ainda procede. Mas hoje o conhecimento de big data também importa, e não apenas o conhecimento profundo de um tema. É provável que a solução de um problema biológico seja encontrada por meio de uma parceria com um astrofísico ou designer de visualização de dados.

Os videogames são um ramo no qual os soldados do big data já abriram caminho para se colocar ao lado dos generais do conhecimento específico, transformando a indústria ao longo do processo. O setor de videogames é grande, fatura anualmente mais que Hollywood no mundo todo. No passado, as empresas criavam o jogo, o lançavam e esperavam que se tornasse um sucesso. Com base nas vendas, as empresas preparavam uma sequência ou davam início a um novo projeto. Decisões quanto ao ritmo de jogo e elementos como personagens, trama, objetos e eventos se baseavam na criatividade dos projetistas, que levavam o trabalho tão a sério quanto Michelangelo ao pintar a Capela Sistina. Era arte, não ciência; um mundo de instintos, como o dos olheiros de beisebol em *O homem que mudou o jogo.*

Mas esses dias ficaram para trás. FarmVille, FrontierVille, FishVille e outros jogos da Zynga são virtuais e interativos. Aparentemente, os jogos on-line permitem que a Zynga analise os dados e modifique os jogos com base em como são jogados. Assim, se os jogadores estão com dificuldade para passar de um nível para outro, ou tendem a abandonar o jogo em determinado momento porque a ação diminui o ritmo, a Zynga é capaz de detectar os problemas nos dados e remediá-los. Mas o menos evidente é que a empresa pode confeccionar jogos de acordo com os jogadores. Não há apenas uma versão de FarmVille — há centenas.

Os analistas de big data da Zynga estudam se as vendas de bens virtuais são afetadas pela cor ou pelo fato de os jogadores verem os amigos os usar. Por exemplo, depois que os dados mostraram que jogadores de FishVille compravam um peixe translúcido seis vezes mais que outras criaturas, a Zynga ofereceu mais espécies translúcidas e teve bons lucros. No jogo Mafia Wars, os dados revelaram que os jogadores compravam mais armas douradas e tigres de estimação brancos.

Não se trata de informações que um designer num estúdio poderia obter, mas os dados se manifestaram. “Somos uma empresa de análise disfarçada de empresa de jogos. Tudo é determinado pelos números”, explicou Ken Rudin, então chefe de análise da Zynga, antes de virar chefe de análise do Facebook. Obter dados não é garantia de sucesso, mas mostra o que é possível.

A mudança para decisões motivadas por dados é profunda. A maioria das pessoas baseia suas decisões numa combinação entre fatos e reflexão, além de uma boa dose de adivinhação. "Uma confusão de visões subjetivas — sensações no plexo solar", nas belas palavras do poeta W.H. Auden. Thomas Davenport, professor de negócios da Babson College, em Massachusetts, e autor de vários livros sobre análise, chama isso de "instinto de ouro". Os executivos confiam em si mesmos por instinto e agem de acordo. Mas isso está começando a mudar à medida que as decisões gerenciais são tomadas ou, pelo menos, confirmadas por modelos de previsão e análise de big data.

A The-Numbers.com, por exemplo, usa dados e matemática para dizer a produtores independentes de Hollywood qual a probabilidade de faturamento de um filme antes mesmo de a primeira cena ser rodada. O banco de dados da empresa analisa cerca de 30 milhões de registros, que cobrem décadas de filmes comerciais americanos e incluem o orçamento do filme, o gênero, elenco, equipe técnica e prêmios, além do faturamento (da bilheteria americana e internacional, direitos internacionais, vendas e aluguel de vídeos e assim por diante) e muito mais. O banco de dados também contém várias conexões de profissionais, do tipo "este roteirista trabalhou com este diretor; este diretor trabalhou com este ator", explica o fundador e presidente da empresa, Bruce Nash.

A The-Numbers.com é capaz de encontrar complexas correlações que preveem o faturamento de filmes. Os produtores levam a informação para os estúdios ou investidores para conseguir financiamento. A empresa pode até mesmo usar as variáveis para dizer aos clientes como aumentar o faturamento (ou diminuir o risco das perdas). Num dos casos, a análise descobriu que um projeto teria muito mais chance de sucesso se o protagonista fosse um ator de primeira linha: na verdade, um indicado ao Oscar com cachê na faixa de US$5 milhões. Em outro caso, Nash informou o estúdio IMAX que um documentário sobre navegação só seria lucrativo se o orçamento de US$12 milhões caísse para US$8 milhões. "A informação deixou o produtor feliz – o diretor, nem tanto", lembra Nash.

Desde a decisão de rodar um filme até a contratação dos profissionais, a mudança nas decisões corporativas começa a aparecer. Erik Brynjolfsson, professor de negócios na Sloan School of Management, do MIT, e seus colegas estudaram o desempenho de empresas que se destacaram em decisões tomadas com base em dados e o compararam com o desempenho de outras empresas. Eles descobriram que os níveis de produtividade eram até 6% maiores nessas empresas que nas que não enfatizavam o uso de dados na tomada de decisões. A descoberta dá às empresas orientadas por dados significativa vantagem, que – com a vantagem da mentalidade e da técnica – pode não durar muito, à medida que mais empresas usarem a abordagem do big data em seus negócios.

UMA QUESTÃO DE UTILIDADE

Enquanto o big data se torna fonte de vantagem competitiva para muitas empresas, a estrutura de ramos inteiros de atuação é refeita. As recompensas, contudo, se acumularão, e os vencedores serão encontrados em meio a empresas grandes e pequenas, espremendo a massa intermediária.

Os maiores *players*, como a Amazon e a Google, continuarão em alta. Ao contrário da situação da era industrial, contudo, sua vantagem competitiva não estará na escala física. A enorme infraestrutura técnica de centros de dados que elas comandam é importante, mas não a característica mais essencial. Com armazenagem e processamento digital em excesso para ceder e agregar em minutos, as empresas podem ajustar sua potência computacional e de armazenagem para se adequar às demandas. Ao transformar o que já foi custo fixo em custo variável, a mudança acaba com a vantagem de escala com base na infraestrutura técnica de que as empresas há muito dispunham.

A escala ainda importa, mas mudou. A que conta agora é a escala dos dados, ou seja, a detenção de enormes quantidades de dados e a capacidade de captar ainda mais com facilidade. Assim, os grandes detentores de dados prosperarão à medida que reunirem e armazenarem mais matéria-prima de seus empreendimentos, que podem reutilizar para agregar valor.

O desafio para os vencedores no mundo dos poucos dados e para os campeões off-line — empresas como Walmart, Proctor & Gamble, GE, Nestlé e Boeing — é valorizar o poder do big data, coletando e usando os dados de forma mais estratégica. A fabricante de turbinas de avião Rolls-Royce transformou completamente seu negócio na última década ao analisar os dados de seus produtos, não apenas ao desenvolvê-los. Do centro de operações, em Derby, no Reino Unido, a empresa continuamente monitora o desempenho de mais de 3.700 turbinas no mundo para encontrar problemas antes que ocorram defeitos. Ela usou os dados para transformar um negócio de manufatura num empreendimento de primeira linha: a Rolls-Royce vende as turbinas e também se oferece para monitorá-las, cobrando os clientes com base no tempo gasto (e reparos ou trocas no caso de problemas). Os serviços hoje correspondem a cerca de 70% do faturamento da divisão de aviação civil da empresa.

Startups e velhos participantes em novas áreas de negócios estão se posicionando para captar vastos bancos de dados. A investida da Apple no mundo dos celulares é um caso. Antes do iPhone, os operadores reuniam dados potencialmente valiosos dos clientes, mas não conseguiam tirar proveito deles. A Apple, por sua vez, exigiu, por contrato com as operadoras, receber boa parte da informação útil. Ao coletar dados de várias operadoras do mundo, a Apple obtém um panorama muito mais abrangente do uso dos celulares que qualquer empresa sozinha poderia conseguir.

O big data também oferece estimulantes oportunidades para o outro extremo do espectro. *Players* bem menores podem ter "escala sem massa", para citar a famosa expressão do professor Brynjolfsson. Isto é, podem ter enorme presença virtual sem recursos físicos e difundir inovações a baixo custo. Mais importante, e pelo fato de alguns dos melhores serviços de big data se basearem principalmente em ideias inovadoras, elas não necessariamente exigem grandes investimentos iniciais. Pequenas empresas podem licenciar os dados em vez de detê-los, fazer a análise em baratas plataformas de nuvem e pagar as taxas de licenciamento com uma porcentagem do faturamento.

Há grande probabilidade de que essas vantagens, nos dois extremos do espectro, não se limitem aos usuários dos dados, mas que abranjam também os detentores. Grandes detentores de dados têm fortes incentivos para aumentar ainda mais os bancos, já que o aumento rende muitos benefícios a um custo irrisório. Primeiro, eles já têm a infraestrutura instalada, em termos de armazenagem e processamento. Depois, há

valor na combinação de bancos de dados. E, por último, um único modo de obter dados simplifica a vida dos usuários. Porém, o mais intrigante é que um novo tipo de detentores de dados pode também surgir no outro extremo: as pessoas. À medida que o valor dos dados se torna cada vez mais aparente, as pessoas podem querer guardar as informações que lhes pertencem – por exemplo, suas preferências de compras, hábitos de consumo de mídia e talvez até informações de saúde.

A propriedade pessoal dos dados pode conferir poder aos consumidores de maneira inédita. As pessoas podem querer decidir sozinhas para quem licenciar seus dados e a que preço. Claro que nem todos vão querer vender seus dados para quem se dispuser a pagar mais; muitos ficarão contentes em vê-los reutilizados gratuitamente em troca de um serviço mais preciso, como as recomendações de livros da Amazon e melhores experiências no Pinterest, serviço de imagens e compartilhamento de conteúdo. Mas, para alguns consumidores hábeis, a ideia de propagar e vender as informações pessoais pode se tornar tão natural quanto blogar, tuitar ou editar um verbete da Wikipédia.

Para que a ideia funcione, contudo, é preciso mais que apenas uma mudança na sofisticação e preferências dos consumidores. Hoje em dia, seria complicado e caro demais para as pessoas licenciarem seus dados pessoais e para as empresas barganharem com cada uma para obtê-los. O mais provável é que vejamos o surgimento de novas empresas que extraiam dados de vários consumidores, criem uma maneira fácil de licenciá-los e de automatizar as transações. Se os custos forem baixos, e as pessoas confiarem, é possível que um mercado de dados pessoais seja criado. Empresas como a Mydex, no Reino Unido, e grupos como o ID3, cofundado por Sandy Pentland, guru da análise de dados pessoais do MIT, já trabalham para tornar realidade esse cenário.

Até que esses intermediários estejam funcionando e os usuários comecem a usá-los, contudo, as pessoas que desejam ser detentoras dos próprios dados têm opções limitadas à sua disposição. Neste ínterim, para reter as opções para o momento em que a infraestrutura e os intermediários estejam funcionando, as pessoas podem considerar ceder menos dados.

Para empresas de porte médio, porém, o big data é menos útil. Há vantagens de escala para as grandes e de custo e inovação para as pequenas, argumenta Philip Evans, do Boston Consulting Group, pensador da tecnologia e do mundo dos negócios. Nos setores tradicionais, empresas de porte médio existem porque combinam um mínimo de proveito da escala com certa flexibilidade, que falta às grandes. Mas, no mundo do big data, não há uma escala mínima que uma empresa tenha de alcançar para pagar por investimentos em infraestrutura de produção. Os usuários do big data que quiserem se manter flexíveis e bem-sucedidos descobrirão que já não precisam alcançar um patamar com relação ao tamanho. Ao contrário, podem continuar pequenos e ainda assim prosperar (ou ser adquirido por um gigante do setor).

O big data sufoca as empresas de porte médio, pressionando-as a se tornarem muito grandes ou pequenas, rápidas ou mortas. Muitos setores tradicionais serão considerados de big data, dos serviços financeiros e indústria farmacêutica à manufatura. O big data não eliminará as empresas de médio porte em todos os setores, mas certamente pressionará empresas de ramos vulneráveis a se renderem ao poder dos grandes dados.

O big data está destinado a abalar também as vantagens competitivas dos estados. Numa época em que boa parte da indústria foi perdida para os países em desenvol-

vimento, e na qual as inovações parecem aleatórias, os países industrializados têm a vantagem de deter os dados e saber como usá-los. A má notícia é que essa vantagem não é sustentável. Como aconteceu com a computação e a internet, a liderança do Ocidente diminuirá à medida que outras partes do mundo adotarem a tecnologia. A boa notícia para as atuais empresas dos países desenvolvidos é que o big data provavelmente superará os pontos fortes e fracos corporativos. Assim, se uma empresa domina o big data, é bem provável que ela não só supere a concorrência como também passe a liderar o mercado.

Foi dada a largada. Assim como o algoritmo de buscas da Google precisa de dados dos usuários para funcionar bem, e como o fornecedor de peças alemão percebeu a importância dos dados para melhorar seus componentes, todas as empresas têm a ganhar ao usar os dados de maneira inteligente.

A despeito dos benefícios, há também motivos para preocupação. À medida que o big data faz previsões cada vez mais precisas sobre o mundo e nosso lugar nele, talvez não estejamos preparados para o impacto na nossa vida pessoal e na nossa sensação de liberdade. Nossa percepção e intuição foram construídas para um mundo de escassez de informações, não de fartura. Exploraremos o lado negro do big data no próximo capítulo.

Capítulo 8

Riscos

Durante quase 40 anos, até a queda do Muro de Berlim, em 1989, a polícia secreta da Alemanha Oriental, conhecida como Stasi, espionava milhões de pessoas. Ao empregar cerca de 100 mil pessoas em horário integral, a Stasi observava carros e ruas, abria cartas e espionava contas bancárias, instalava escutas em apartamentos, grampeava linhas telefônicas e induzia namorados e casais, pais e filhos, a se espionarem, traindo a mais básica confiança que os seres humanos nutrem uns pelos outros. Os arquivos resultantes – incluindo pelo menos 39 milhões de fichas e 100 quilômetros de documentos – gravavam e detalhavam os aspectos mais íntimos da vida de pessoas comuns. A Alemanha Oriental era um dos maiores Estados policiais jamais vistos.

Vinte anos depois da derrocada da Alemanha Oriental, mais dados a respeito de nós são coletados e armazenados. Estamos sob constante vigilância: quando usamos cartões de crédito, celulares, documento de identidade. Em 2007, a imprensa britânica revelou a ironia do fato de haver mais de 30 câmeras de segurança num raio de cerca de 90 metros do apartamento londrino onde George Orwell escreveu *1984* (Companhia das Letras, 2009). Muito antes da internet, empresas especializadas, como a Equifax, Experian e Acxiom, coletavam, tabulavam e forneciam acesso a informações pessoais de centenas de milhões de pessoas no mundo todo. A internet facilitou, barateou e tornou o rastreamento mais útil. Agências governamentais clandestinas não são as únicas a nos espionarem. A Amazon monitora nossas preferências de compra, e a Google, nossos hábitos de navegação, enquanto o Twitter sabe o que se passa em nossas mentes. O Facebook também parece absorver toda a informação, junto com nossas relações sociais. Operadoras de celular sabem não só com quem conversamos, mas também quem está próximo.

Com o big data prometendo valiosas ideias para aqueles que os analisam, todos os sinais parecem apontar para uma nova onda de coleta, armazenamento e reutilização de nossos dados pessoais. O tamanho e escala dos bancos de dados aumentarão à medida que os custos de armazenamento continuarem a diminuir e as ferramentas analíticas se tornarem ainda mais potentes. Se a era da internet pôs em risco a privacidade, será que o big data a ameaça ainda mais? Será este o lado negro do big data?

Sim, e não é o único. Aqui também a questão essencial sobre o big data é que a mudança de escala leva à mudança de estado. Como explicaremos, essa transformação não só dificulta a proteção da privacidade como também apresenta uma ameaça totalmente nova: castigos com base em propensões, isto é, a possibilidade de usar previsões de big data sobre pessoas para julgá-las e puni-las antes mesmo que elas ajam, o que renega a ideia de justiça e livre-arbítrio.

Além da privacidade e da propensão, há um terceiro perigo. Corremos o risco de sermos vítimas da ditadura dos dados, na qual adoramos as informações e os resultados de nossas análises e acabamos usando-os de forma equivocada. Com responsabilidade, o big data é um instrumento útil de tomada de decisão. Se usados sem sabedoria, eles se tornam um instrumento de poderosos, que podem transformá-los numa fonte de repressão, seja ao frustrar clientes e funcionários ou, pior, ao atacar cidadãos.

Há mais em jogo do que normalmente se reconhece. Os perigos de não ser capaz de controlar o big data em termos de respeito à privacidade e previsão, ou de se enganar quanto a seu significado, vão além do trivial, como anúncios publicitários específicos. A história do século XX está repleta de situações nas quais os dados resultaram em trágicos fins. In 1943, o U.S. Census Bureau (Departamento de Estatística dos Estados Unidos) forneceu endereços dos quarteirões (mas não os nomes das ruas ou os números das casas, para manter uma suposta privacidade) de nipo-americanos para facilitar sua prisão. Os famosos e abrangentes registros civis holandeses foram usados pelos invasores nazistas na perseguição de judeus. Os números de cinco dígitos tatuados nos braços dos presos nos campos de concentração nazistas inicialmente correspondiam a números de cartões pontilhados IBM Hollerith; o processamento de dados facilitou o assassinato em escala industrial.

Apesar de sua habilidade, havia muito que a Stasi não conseguia fazer: não conseguia saber onde as pessoas estavam o tempo todo ou com quem conversavam. Hoje, porém, boa parte dessa informação é coletada por operadoras de celular. O Estado alemão oriental não podia prever quando as pessoas se tornariam dissidentes, nem nós – mas forças policiais estão começando a usar algoritmos para decidir onde e quando patrulhar, indício do rumo que estamos tomando. Essas tendências tornam os riscos inerentes ao big data tão grandes quanto os bancos de dados em si.

PRIVACIDADE PARALISANTE

É tentador extrapolar o perigo à privacidade resultante do crescimento dos dados e ver semelhança com a fiscalização distópica de Orwell em *1984*. Mas a situação é mais complexa. Para começar, nem todos os big data contêm informações pessoais, como os dados de sensores das refinarias, das máquinas das fábricas ou dados sobre a explosão de bueiros ou o clima nos aeroportos. A BP e a Con Edison não precisavam (ou queriam) de informações pessoais a fim de agregar valor na análise que realizavam. A análise de big data desse tipo de informação não oferece praticamente qualquer risco à privacidade.

Mas boa parte dos dados gerados hoje inclui, sim, informações pessoais. E as empresas têm vários incentivos para captar ainda mais dados, mantê-los por mais tempo e reutilizá-los com mais frequência. Os dados podem não parecer informações pessoais explícitas, mas, com os processos de análise, podem facilmente dizer a quem se referem ou facilitar a dedução de detalhes íntimos da vida de uma pessoa.

Por exemplo, nos Estados Unidos e na Europa, aparelhos domésticos contêm medidores elétricos "inteligentes", que coletam dados ao longo do dia, talvez a cada seis segundos - muito mais que as informações que os medidores tradicionais de consumo de energia coletam. O mais importante é que o modo como os aparelhos

elétricos consomem energia gera uma "assinatura de carga" única para cada aparelho. Assim, um aquecedor de água é diferente de um computador, que é diferente de luzes usadas para o cultivo de maconha. Por isso, o uso de energia de uma casa revela informações pessoais, seja o comportamento diário dos residentes, suas condições de saúde e atividades ilegais.

A questão mais importante, contudo, não é se o big data aumenta os riscos para a privacidade (sim, aumenta), mas se muda as características do risco. Se a ameaça for maior, as leis e regras que protegem a privacidade talvez ainda funcionem na era do big data; tudo o que precisamos fazer é redobrar esforços. Por outro lado, se o problema muda, precisamos de novas soluções.

Infelizmente o problema mudou. Com o big data, o valor da informação não está mais somente no propósito primário. Como discutimos, ele está agora nos usos secundários.

Essa mudança prejudica o papel central de cada pessoa nas atuais leis de privacidade. Hoje, lhe dizem em que momento a informação está sendo coletada e com qual objetivo; cada um, então, tem a oportunidade de concordar, de modo que seja dado início à coleta. Apesar de esse conceito de "saber e consentir" não ser a única maneira legal de coletar e processar dados pessoais, de acordo com Fred Cate, especialista em privacidade da Indiana University, ele foi totalmente transformado na base dos princípios de privacidade em todo o mundo. (Na prática, esse conceito gerou contratos gigantescos, raramente lidos e muito menos compreendidos – mas essa é outra história.)

O impressionante na era do big data é que não se pensou nos usos secundários mais inovadores quando os dados foram coletados. Como as empresas podem informar sobre um objetivo que não existe? Como as pessoas podem consentir em dar informações para o desconhecido? Ainda que, na falta de consentimento, toda análise de big data que contenha informações pessoais talvez requeira um novo contato com a pessoa para nova permissão para a reutilização. Você imagina a Google tentando contatar centenas de milhões de usuários para pedir a aprovação para usar suas antigas buscas na internet a fim de prever um surto de gripe? Nenhuma empresa suportaria o custo, nem mesmo se fosse tecnicamente possível.

A alternativa, isto é, pedir aos usuários que concordem com qualquer uso futuro dos dados coletados, tampouco é útil. A permissão ampla castra a própria noção de consentimento informado. No contexto do big data, o testado e confiável conceito de conhecimento e consentimento é geralmente restritivo demais para se extrair o valor latente dos dados ou vazio demais para proteger a privacidade das pessoas.

Outras maneiras de proteger a privacidade também falham. Se a informação de todos estiver num banco de dados, até mesmo a opção de não consentimento pode deixar um rastro. Veja o caso do Street View, da Google. Os carros coletam imagens de estradas e casas em muitos países. Na Alemanha, a Google enfrentou protestos do público e da mídia. As pessoas temiam que as imagens de suas casas e jardins pudessem ajudar quadrilhas de ladrões na escolha de alvos lucrativos. Sob pressão regulatória, a Google concordou em permitir que as imagens das casas fossem embaçadas. Mas a opção de exclusão é visível no Street View – você nota as casas ofuscadas –, e os ladrões podem interpretar o sinal como indício de alvos especialmente bons.

Uma abordagem técnica de proteção da privacidade – a anonimização – também falha em vários casos. A anonimização se refere à exclusão de quaisquer dados pessoais de um banco de dados, como nome, endereço, número do cartão de crédito, data de nascimento ou identidade. Os dados resultantes podem, então, ser analisados e compartilhados sem comprometer a privacidade de ninguém. A prática funciona num mundo de pequenos dados. Mas o big data, com o aumento de quantidade e variedade de informações, facilitam a reidentificação. Pense nos casos das aparentemente anônimas buscas na internet e classificação de filmes.

Em agosto de 2006, a AOL disponibilizou publicamente uma enorme quantidade de antigas buscas, sob o argumento de que os pesquisadores podiam analisá-las para obter ideias interessantes. O banco de dados, com 20 milhões de buscas de 657 mil usuários, entre 1° de março e 31 de maio daquele ano, foi cuidadosamente anonimizado. Informações pessoais, como o nome do usuário e endereço de IP, foram apagadas e substituídas por identificadores numéricos únicos. A ideia era a de que os pesquisadores pudessem conectar buscas da mesma pessoa, mas sem informações que a identificassem.

Em poucos dias, o *New York Times* reuniu buscas como "60 homens solteiros", "chá bom para a saúde" e "paisagistas em Lilburn, Ga", para identificar com sucesso o usuário 4417749 como Thelma Arnold, viúva de 62 anos, de Lilburn, Geórgia. "Meus Deus, é toda a minha vida pessoal", disse ela quando o repórter do *Times* bateu à sua porta. "Não fazia ideia de que havia alguém me vigiando." O clamor público levou à dispensa do chefe de tecnologia do AOL e de outros dois funcionários.

Ainda assim, dois meses mais tarde, em outubro de 2006, a locadora de filmes Netflix fez o mesmo ao lançar o "Prêmio Netflix". A empresa disponibilizou 100 milhões de registros de locação de quase 500 mil usuários — e ofereceu um prêmio de US$1 milhão para qualquer equipe capaz de melhorar o sistema de recomendação de filmes em pelo menos 10%. Novamente, identificadores pessoais foram cuidadosamente removidos dos dados. Novamente, um usuário foi mais uma vez identificado: uma mãe e lésbica não assumida na conservadora região do Meio-Oeste dos Estados Unidos, que por causa disso processou a Netflix sob o pseudônimo "Jane Doe".

Pesquisadores da University of Texas, em Austin, compararam os dados da Netflix com outras informações públicas. Eles rapidamente descobriram que a classificação do usuário anonimizado combinava com as de um colaborador identificado do site IMDB (Internet Movie Database). De forma mais ampla, a pesquisa mostrou que a classificação de apenas seis filmes obscuros (dentre os 500 mais vistos) podia identificar um cliente da Netflix em 84% dos casos. Se alguém soubesse a data da classificação, podia identificar a pessoa entre os quase 500 mil usuários no banco de dados com precisão de 99%.

No caso da AOL, as identidades dos usuários foram expostas pelo conteúdo das buscas. No caso da Netflix, a identidade foi revelada por uma comparação com dados de outras fontes. Nos dois casos, as empresas não valorizaram a relevância do big data na descoberta da identidade. Há dois motivos: coletamos e combinamos mais dados.

Paul Ohm, professor de Direito na University of Colorado, em Boulder, e especialista nos danos causados pela revelação da identidade, explica que não há solução fácil. Com bastante dados, a anonimização perfeita é impossível, por mais que se tente.

Pior, os pesquisadores recentemente mostraram que não apenas dados convencionais como também o gráfico social – as conexões entre as pessoas – estão vulneráveis à revelação da identidade.

Na era do big data, as três principais estratégias usadas para garantir a privacidade — consentimento individual, opção de exclusão e anonimização — perderam a eficiência. Hoje, muitos usuários já sentem que sua privacidade está sendo violada. Espere até que as práticas de big data se tornem mais comuns.

Comparada com a Alemanha Oriental de 25 anos atrás, a vigilância só ficou mais fácil, barata e potente. A capacidade de coletar dados pessoais geralmente está incluída em ferramentas que usamos diariamente, de sites a aplicativos para *smartphones*. Os registradores de dados, instalados na maioria dos carros para captar as ações de um veículo poucos segundos antes da ativação de um airbag, são conhecidos por "testemunhar" em tribunais contra proprietários de carros em ações sobre acidentes.

Claro que quando as empresas coletam dados para melhorar a produção, não precisamos temer que a vigilância terá as mesmas consequências dos grampos realizados pela Stasi. Não vamos para a prisão se a Amazon descobrir que gostamos de ler o *Livro vermelho*, do Camarada Mao (Martin Claret, 2002). A Google não nos exilará porque buscamos "Bing". As empresas podem ser poderosas, mas não têm o poder de coerção dos Estados.

Assim, apesar de não nos prenderem no meio da noite, empresas de todos os tipos coletam e armazenam enorme volume de informações pessoais sobre todos os aspectos de nossas vidas, as compartilham com outros sem nosso conhecimento e as usam de maneiras que mal poderíamos imaginar.

O setor privado não é o único que está começando a trabalhar com o big data. Os governos também. Acredita-se, por exemplo, que a NSA (U.S. National Security Agency) intercepte e armazene 1,7 bilhão de e-mails, ligações telefônicas e outras comunicações todos os dias, de acordo com uma investigação de 2010 do *Washington Post*. William Binney, ex-oficial da NSA, estima que o governo tenha compilado "20 trilhões de transações" entre cidadãos americanos e outros – quem liga para quem, quem manda e-mail para quem, quem transfere dinheiro para quem e assim por diante.

Para compreender todos os dados, os Estados Unidos estão construindo gigantescos centros de dados, como a instalação da NSA de US$1,2 bilhão, em Fort Williams, no estado de Utah. Todos os ramos do governo estão exigindo mais informações que antes, não apenas agências secretas envolvidas com o contraterrorismo. Quando os dados se expandem para informações como transações financeiras, registros de saúde e status do Facebook, a quantidade estudada é inconcebivelmente alta. O governo não pode processar tantos dados assim. Então por que os coleta?

A resposta aponta para a maneira como a vigilância mudou na era do big data. No passado, os investigadores instalavam grampos aos cabos telefônicos para descobrir o máximo sobre um suspeito. O importante era investigar e conhecer melhor a pessoa. A abordagem moderna é diferente. No espírito da Google ou Facebook, o novo raciocínio é que as pessoas são a soma de suas relações sociais, interações on-line e conexões com conteúdo. A fim de investigar totalmente alguém, os analistas têm de estudar todos os dados possíveis que cercam a pessoa – não apenas quem conhecem, mas quem esses amigos também conhecem e assim por diante, tecnicamente muito difícil de fazer no

passado. Hoje, é mais fácil que nunca. E, como o governo nunca sabe quem será o próximo investigado, ele coleta, armazena e garante acesso a informações não necessariamente para monitorar todos o tempo todo, e sim para que, quando alguém for suspeito, as autoridades possam imediatamente investigar em vez de começar a coletar as informações apenas a partir daquele momento.

Os Estados Unidos não são o único país que reúne dados pessoais, e talvez nem seja o melhor nesta prática. Mas, por mais problemática que seja a capacidade de as empresas e governos obterem nossas informações pessoais, um novo problema surge com o big data: o uso de previsões para nos julgar.

PROBABILIDADE E PUNIÇÃO

John Anderton é chefe de uma unidade policial especial em Washington, D.C. Nesta manhã em particular, ele entra numa casa de subúrbio momentos antes de Howard Marks, enraivecido, cravar uma tesoura no peito de sua mulher, que ele encontrou na cama com outro homem. Para Anderton, é apenas mais um dia de prevenção de crimes contra a vida. "Em nome da Divisão de Pré-crimes do Distrito de Colúmbia", anuncia ele, "eu o prendo pelo futuro assassinato de Sarah Marks, que aconteceria hoje..."

Outros policiais começam a prender Marks, que grita: "Não *fiz* nada!"

A cena de abertura de *Minority Report: a nova lei* retrata uma sociedade na qual as previsões se tornaram tão precisas que a polícia prende as pessoas por crimes antes mesmo de serem cometidos. As pessoas não são presas pelo que fizeram, e sim pelo que pretendiam fazer, mesmo que nunca tenham de fato cometido o crime. O filme atribui essa lei preventiva à visão de três clarividentes, e não à análise de dados. Mas o filme retrata um perturbador futuro, no qual a descuidada análise do big data ameaça se tornar realidade, um futuro no qual os julgamentos de culpa se baseiam em previsões individualizadas de comportamento.

Já vemos as bases disso. Os departamentos de liberdade condicional em mais da metade dos estados americanos usam previsões baseadas em análises de dados para decidir se deve ou não libertar alguém da prisão. Cada vez mais lugares nos Estados Unidos – de delegacias em Los Angeles a cidades como Richmond, Virginia — empregam o "policiamento preventivo": usando análises de big data para selecionar que ruas, grupos e pessoas devem ter vigilância redobrada, porque um algoritmo identificou maior probabilidade de ocorrer um crime.

Na cidade de Memphis, no Tennessee, um programa chamado Blue CRUSH (acrônimo em inglês para Redução de Crimes pelo Uso de Estatísticas Históricas) fornece aos policiais, com relativa precisão, áreas de interesse em termos de localização (alguns quarteirões) e tempo (algumas horas durante um dia específico da semana). O sistema ostensivamente ajuda a polícia a direcionar melhor os escassos recursos. Desde sua implantação, em 2006, crimes violentos e contra a propriedade diminuíram 25%, de acordo com uma medição (mas claro que o fato não revela causalidade e nada indica que a diminuição se deva ao Blue CRUSH).

Em Richmond, na Virginia, a polícia relaciona dados de crimes com outros bancos de dados, como informações sobre quando as grandes empresas da cidade pagam os funcionários ou datas de shows ou eventos esportivos. Essa ação confirmou e às vezes

aperfeiçoou a suspeita dos policiais quanto às tendências criminais. Há tempos, por exemplo, a polícia de Richmond sentia que havia um surto de crimes violentos depois de feiras de armas; a análise dos dados provou que ela estava certa, com um detalhe: o surto ocorria duas semanas depois da festa, e não imediatamente após o evento.

O sistema busca prevenir crimes ao fazer previsões, às vezes no âmbito dos criminosos em potencial. A prática aponta para o uso do big data com novo objetivo: prevenção de crimes.

A arte da ficção científica está se aproximando da banalidade do cotidiano. Um projeto de pesquisa do U.S. Department of Homeland Security (Departamento de Segurança Interna dos Estados Unidos), chamado FAST (Future Attribute Screening Technology – Tecnologia de Investigação de Atributos Futuros) tenta identificar terroristas em potencial ao monitorar sinais vitais, linguagem corporal e outros padrões fisiológicos. A ideia é a de que a vigilância sobre o comportamento das pessoas talvez detecte sua intenção de cometer um ato. Em testes, o sistema mostrou precisão de 70%, de acordo com o departamento. (O que isso significa não está claro; os pesquisados foram instruídos a se passar por terroristas para ver se sua "má intenção" seria identificada?) Apesar de os sistemas parecerem embrionários, a questão é que a polícia os leva bem a sério.

Impedir o acontecimento de crimes parece uma atraente perspectiva. Evitar que as infrações aconteçam não é bem melhor que punir os perpetradores posteriormente? Prever crimes não traria proveitos para as vítimas e também para a sociedade em geral?

Mas esse é um caminho conturbado. Se, por meio do big data, podemos prever quem pode cometer um crime no futuro, talvez não nos contentemos apenas em evitar que o crime aconteça; provavelmente também vamos querer punir o provável perpetrador. Faz sentido. Se apenas interviermos para impedir que um ato ilícito aconteça, o perpetrador pode tentar de novo e impunimente. Por outro lado, ao usarmos o big data para responsabilizá-lo por seus atos (futuros), poderemos detê-lo.

Punições com base em previsão parecem um aprimoramento das práticas já aceitas. A prevenção de comportamentos não saudáveis, perigosos ou ilegais é uma das bases da sociedade moderna. Tentamos reduzir a liberdade de fumar para evitar o câncer; exigimos o uso de cinto de segurança para evitar fatalidades em acidentes de carros; não permitimos que as pessoas andem armadas nos aviões para evitar sequestros. Medidas preventivas diminuem nossa liberdade, mas muitos as veem como um pequeno preço a se pagar para evitar um mal maior.

Em muitos contextos, a análise de dados já é empregada em nome da prevenção, usada, por exemplo, para nos categorizar em grupos, geralmente com sucesso. Tabelas atuariais notam que homens com mais de 50 anos estão sujeitos a câncer de próstata, então os membros deste grupo aceitam pagar mais por planos de saúde, mesmo que nunca desenvolvam a doença. Bons alunos do ensino médio, como um grupo, têm menos tendência a se envolverem em acidentes de carro — assim, alguns colegas com notas mais baixa têm de pagar seguros mais caros. Pessoas com certas características são sujeitas a uma investigação mais minuciosa ao passarem pela segurança no aeroporto.

Esta é a ideia por trás do "perfilamento" no mundo atual de pequenos dados: encontrar uma associação comum nos dados, definir um grupo de pessoa à qual ela se aplique e depois passar a investigá-las melhor. É uma regra de generalização que se aplica a

todos no grupo. “Perfilamento”, claro, é uma palavra forte, e o método apresenta sérios problemas. Se usado de forma errada, pode não só levar à discriminação contra certos grupos como também à “culpa por associação”.

Por outro lado, as previsões de big data quanto às pessoas são diferentes. Enquanto as previsões comportamentais de hoje – encontradas em seguros e pontuação de crédito – geralmente contam com vários fatores que se baseiam num modelo mental do tema em questão (isto é, problemas anteriores de saúde e histórico de pagamento de empréstimos), na análise não causal do big data, geralmente identificamos os dados mais adequados de previsão em meio ao oceano de informação.

Mais importante, com o uso do big data, esperamos identificar pessoas específicas em vez de grupos, o que nos liberta do problema da “culpa por associação”, no caso do “perfilamento”. Num mundo de big data, alguém com um nome árabe, que paga em dinheiro por uma passagem só de ida de primeira classe, talvez não esteja mais sujeito a uma minuciosa investigação no aeroporto se outros dados específicos determinarem improvável se tratar de um terrorista. Com o big data, podemos escapar da camisa de força das identidades grupais e substituí-las por previsões mais granuladas para cada pessoa.

A promessa do big data é continuar com a mesma prática – “perfilando” –, mas aperfeiçoada, de forma menos discriminatória e mais individualizada. Parece aceitável se o objetivo for apenas evitar ações indesejadas, mas se torna muito perigosa se usamos o big data para prever se alguém é culpado e deve ser punido por um comportamento que ainda não aconteceu.

A própria ideia de punição com base em propensão é repugnante. Acusar uma pessoa de um comportamento possível futuro é negar a própria base da justiça: a de que alguém tem de fazer algo antes de ser responsabilizado. Afinal, maus pensamentos não são ilegais; más ações, sim. É um pilar fundamental da nossa sociedade que a responsabilidade individual esteja ligada ao ato individual. Se alguém é coagido, diante de uma arma, a abrir o cofre da empresa, a pessoa não tem escolha e, portanto, não tem responsabilidade sobre o ato.

Se as previsões de big data fossem perfeitas, se os algoritmos pudessem prever o futuro com clareza e sem falhas, não teríamos mais escolhas sobre o futuro. Nós nos comportaríamos exatamente como o previsto. Se as previsões perfeitas fossem possíveis, elas negariam o livre-arbítrio, nossa capacidade de viver com liberdade. Ironicamente, ao nos privar da escolha nos isentaria de qualquer responsabilidade.

Claro que previsões perfeitas são impossíveis. A análise de big data preverá a probabilidade de alguém ter determinado comportamento futuro. Pense, por exemplo, na pesquisa conduzida por Richard Berk, professor de Estatística e Criminologia da University of Pensilvânia. Ele afirma que seu método pode prever se uma pessoa em liberdade condicional se envolverá num homicídio (matar ou ser morta). Ele usa muitas variáveis específicas do caso como dados, incluindo o motivo da prisão e a data do primeiro crime, mas também dados demográficos, como idade e gênero. Berk sugere que podemos prever um futuro assassino entre os criminosos em liberdade condicional com probabilidade de pelo menos 75%. Nada mal. Mas também significa que, se os departamentos de liberdade condicional confiarem nas análises de Berk, eles se equivocarão em 25% dos casos.

O problema central em contar com essas previsões não é o fato de exporem a sociedade ao risco, mas de essencialmente punirmos as pessoas *antes* do ato criminoso. Ao intervir antecipadamente (negando-lhes, por exemplo, a liberdade condicional se houver alta probabilidade de envolvimento num homicídio), jamais saberemos se elas teriam de fato cometido ou não o crime previsto. Não deixamos que o destino exerça seu papel e mesmo assim consideramos as pessoas responsáveis pelo que as previsões nos dizem que poderiam ter feito. Essas previsões nunca podem ser contestadas.

Essa prática renega a própria ideia de presunção da inocência, princípio sobre o qual o sistema jurídico e o senso de justiça se baseiam. Se responsabilizarmos as pessoas por atos futuros que talvez nunca sejam cometidos, também negamos que os seres humanos sejam capazes de fazer uma escolha moral.

A questão importante aqui *não* é apenas de policiamento. O perigo é muito maior que a justiça criminal; ele cobre todas as áreas da sociedade, todas as instâncias do julgamento humano nas quais as previsões de big data são usadas para decidir se as pessoas são culpadas ou não por atos futuros, desde a decisão de uma empresa de demitir um funcionário, um médico que se negue a operar um paciente e uma esposa que peça o divórcio.

Talvez, com um sistema assim, a sociedade fosse mais segura e eficiente, mas uma parte essencial do que nos torna humanos – a capacidade de escolher como agimos e assumirmos a responsabilidade por nossas escolhas – seria destruída. O big data se tornaria instrumento de coletivização da escolha humana e de abdicação do livre-arbítrio na nossa sociedade.

Claro que o big data oferece vários benefícios. O que o transforma numa arma de desumanização não é um problema do big data em si, mas da forma como usamos as previsões. O essencial é que a responsabilização das pessoas por atos previstos antes que elas os cometam usa as previsões do big data com base em correlações para tomar decisões causais sobre a responsabilidade individual.

O big data é útil para entender o presente e os riscos futuros e para ajustar nossas ações adequadamente. As previsões ajudam pacientes e segurados, banqueiros e consumidores. Mas o big data nada nos diz sobre causalidades. Por outro lado, determinar a "culpa" – a culpabilidade individual – exige que as pessoas que julgamos tenham optado por uma ação específica. Sua escolha deve ter relação causal com a ação seguinte. Justamente porque se baseia em correlações, o big data é intrinsecamente inadequado para nos ajudar a julgar causalidade e, assim, determinar a culpabilidade individual.

O problema é que os seres humanos são propensos a ver o mundo através das lentes de causa e efeito. Assim, o big data está sob a constante ameaça do uso excessivo para objetivos causais, ligados a visões turvas sobre a eficiência de nosso julgamento e tomada de decisão de determinar culpabilidade, se estivéssemos munidos apenas com previsões de big data.

É o tradicional beco sem saída – que leva diretamente para a sociedade retratada em *Minority Report,* na qual a escolha individual e o livre-arbítrio foram eliminados, na qual nossa bússola moral foi substituída por algoritmos, e as pessoas estão expostas ao peso das decisões coletivas. Se assim empregado, o big data ameaça nos aprisionar – talvez literalmente – a probabilidades.

A DITADURA DOS DADOS

O big data corrói a privacidade e ameaça a liberdade, mas também exacerba um antigo problema: o de contar com os números quando estes são muito mais falíveis do que pensamos. Nada ressalta as consequências da análise equivocada de dados mais que a história de Robert McNamara.

McNamara era um cara dos números. Nomeado secretário de defesa dos Estados Unidos quando as tensões com o Vietnã se intensificaram, no início dos anos 1960, ele insistia em obter dados de tudo o que fosse possível. Somente com o rigor estatístico, acreditava ele, os tomadores de decisões podiam entender uma situação complexa e assim fazer as escolhas certas. Em sua visão, o mundo era uma massa de informações confusas que, se depuradas, demarcadas e quantificadas, podiam ser domadas e controladas. McNamara buscava a Verdade, e essa Verdade seria encontrada nos dados. Entre os números que ele obtinha, estava a "contagem de corpos".

McNamara desenvolveu seu amor por números como aluno da Harvard Business School e depois como o professor assistente mais jovem da universidade, com apenas 24 anos. Ele aplicou seu rigor durante a Segunda Guerra Mundial, como parte de uma equipe de elite do Pentágono chamada Controle Estatístico, que levou as tomadas de decisões com base em dados para o maior corpo burocrático do mundo. Antes disso, o Exército era cego. Ele não sabia, por exemplo, o tipo, a quantidade e a localização de peças de reposição dos aviões. Os dados vieram em seu socorro. A aquisição mais eficiente de armamentos rendeu uma economia de US$3,6 bilhões em 1943. A guerra moderna tratava de alocação mais eficiente de recursos; o trabalho da equipe foi um incrível sucesso.

Ao fim da guerra, o grupo decidiu permanecer junto e oferecer suas habilidades para o mundo corporativo. A Ford Motor Company estava em apuros, e um desesperado Henry Ford II lhes deu as rédeas da empresa. Assim como não sabiam nada sobre o Exército ao ajudá-lo a vencer a guerra, eles também não tinham qualquer conhecimento sobre a fabricação de carros. Mesmo assim os "Whiz Kids" salvaram a empresa.

McNamara subia na hierarquia à medida que obtinha um dado para cada situação. Os gerentes das fábricas geravam os números que ele exigia – corretos ou não. Quando foi dada a ordem de que todo o estoque de um carro deveria ser usado antes do início da produção de um novo modelo, gerentes desesperados jogaram componentes sobressalentes num rio próximo. Os chefões da sede anuíram em sinal de aprovação depois que os gerentes devolveram números que confirmavam que a ordem foi obedecida. Mas a piada na fábrica era a de que se podia andar sobre as águas — sobre peças enferrujadas de carros modelos 1950 e 1951.

McNamara é o exemplo perfeito do gerente de meados do século XX, o executivo hiper-racional que confiava mais em números que em intuição e que podia aplicar suas habilidades quantitativas a qualquer indústria à qual se dedicasse. Em 1960, ele foi nomeado presidente da Ford, cargo que ocupou por apenas algumas semanas antes que o presidente Kennedy o nomeasse secretário de defesa.

À medida que o conflito no Vietnã se intensificava e os Estados Unidos enviavam mais tropas, ficou claro que aquela era uma guerra de decisões, não de tomada de território. A estratégia dos Estados Unidos era trazer o *Viet Cong* à mesa de negociações.

A maneira de medir o progresso, portanto, era pelo número de inimigos mortos. A contagem de corpos era publicada diariamente nos jornais. Para os que apoiavam a guerra, a contagem era prova do progresso; para os críticos, prova de sua imoralidade. A contagem de corpos foi um dado que definiu uma era.

Em 1977, dois anos depois que o último helicóptero decolou da laje da embaixada americana em Saigon, um general aposentado, Douglas Kinnard, publicou um memorável compêndio sobre a visão dos generais. Chamado *The War Managers,* o livro revelava o atoleiro da quantificação. Apenas 2% dos generais americanos consideravam a contagem de corpos uma maneira válida de medir o progresso. Cerca de dois terços disseram que a contagem era geralmente inflada. "Uma mentira – totalmente inútil", escreveu um general em seus comentários. "Mentiras descaradas", escreveu outro. "Elas eram exageradas por muitas unidades, principalmente por causa do interesse demonstrado por pessoas como McNamara", disse um terceiro.

Assim como os operários da Ford, que jogaram partes do motor no rio, os oficiais menos graduados às vezes davam aos superiores números impressionantes para manter o comando ou fomentar suas carreiras – dizendo aos mais graduados o que queriam ouvir. McNamara e os homens seus parceiros confiavam nos números, exaltavam-nos. Com o cabelo perfeitamente penteado e a gravata bem firme, McNamara sentia que só podia compreender o que estava acontecendo na guerra pela análise de uma tabela – todas aquelas fileiras e colunas em ordem, cálculos e gráficos, cujo domínio parecia aproximá-lo de Deus.

O uso, excesso e utilização imprópria dos dados pelo Exército americano durante a Guerra do Vietnã é uma complexa lição sobre os limites da informação numa era de pequenos dados, lição à qual é preciso prestar a atenção à medida que o mundo ruma para uma era de big data. A qualidade dos dados pode ser baixa, tendenciosa, mal analisada e usada equivocadamente. Pior, os dados podem não captar o que se pretende quantificar.

Estamos mais suscetíveis do que imaginamos à "ditadura dos dados" – isto é, deixar que os dados nos governem de maneiras que podem tanto nos prejudicar quanto nos beneficiar. O problema é nos deixarmos guiar pelo resultado das análises mesmo quando tivermos motivos para suspeitar que haja algo de errado, ou que nos tornemos obcecados em coletar fatos e números para o bem dos dados. Ou ainda que atribuamos um grau de verdade que os dados não mereçam.

À medida que mais aspectos da vida se tornam dataficados, a solução que os legisladores e empresários estão começando a buscar é a obtenção de mais dados. "Em Deus confiamos – todos os outros trazem dados" é o mantra do gerente moderno, que ecoa no Vale do Silício, nas fábricas e nos corredores dos departamentos estatais. O sentimento é claro, mas é possível que alguém facilmente se engane com os dados.

A educação parece ter problemas? Use testes padronizados para medir o desempenho e penalizar professores ou escolas que, de acordo com essa medida, não estejam no padrão. Se os testes de fato captam as habilidades dos alunos, a qualidade do ensino ou as necessidades de uma força de trabalho mais criativa e adaptável são questionáveis – mas trata-se de um questionamento que os dados não admitem.

Você quer prever o terrorismo? Crie listas de investigação e de proibição de voo a fim de policiar os céus. Mas o fato de esses bancos de dados realmente oferecerem a

proteção que prometem é questionável. Num famoso incidente, o falecido senador Ted Kennedy, de Massachusetts, foi incluído numa lista de passageiros proibidos, detido e questionado, apenas porque havia uma pessoa com o mesmo nome no banco de dados.

As pessoas que trabalham com dados têm uma expressão para alguns desses problemas: "lixo que entra, lixo que sai". Em certos casos, o motivo é a qualidade da informação, mas em geral, é o uso equivocado da análise produzida. Com o big data, os problemas podem surgir com mais frequência ou ter maiores consequências.

A Google, como mostramos em vários exemplos, administra tudo de acordo com dados. Essa estratégia é claramente responsável por boa parte de seu sucesso, mas às vezes também atrapalha a empresa. Os cofundadores, Larry Page e Sergey Brin, insistiam em saber a pontuação SAT (teste de avaliação de conhecimento exigido para entrar em um curso superior nos Estados Unidos) e a nota média da faculdade de todos os candidatos a empregos. De acordo com o raciocínio deles, o primeiro número media o potencial, e o segundo, as realizações. Bons gerentes por volta dos 40 anos que vinham sendo recrutados eram perseguidos pelas pontuações. A empresa até mesmo continuou a exigir números depois que os estudos internos mostraram que não havia correlação entre as pontuações e o desempenho no emprego.

A Google deveria ter sido mais perspicaz para saber que era preciso resistir à sedução do encanto dos dados. A medida deixa pouco espaço para a mudança na vida de uma pessoa, não leva em conta o conhecimento, apenas boas notas, e não necessariamente reflete as qualificações das pessoas do setor de humanas, onde o *know-how* talvez seja menos quantificado que na ciência e na engenharia. A obsessão da Google por dados com objetivo de recrutamento é especialmente exagerada se considerarmos que os fundadores da empresa são produtos da escola montessoriana, que dá ênfase ao aprendizado, não às notas. E a empresa repete os erros de antigas potências da tecnologia, que valorizavam mais os currículos que as habilidades de fato. Será que Larry e Sergey, dois desistentes do curso de doutorado, desprezariam a oportunidade de trabalhar como gerentes nos lendários laboratórios Bell? Pelos padrões da Google, Bill Gates, Mark Zuckerberg ou Steve Jobs jamais teriam sido contratados, já que lhes faltavam diplomas universitários.

A dependência da empresa quanto aos dados às vezes parece demasiada. Marissa Mayer, quando era uma das principais executivas, certa vez ordenou à equipe que testasse 41 gradações de azul para ver quais as pessoas usavam mais, a fim de determinar a cor de uma barra de ferramentas no site. A deferência da Google por dados foi levada ao extremo e até despertou revolta.

Em 2009, o principal designer da Google, Douglas Bowman, se demitiu por não suportar a constante quantificação de tudo. "Tive um recente debate sobre se uma borda deveria ter 3, 4 ou 5 pixels de largura, e me pediram para defender meu argumento. Não posso trabalhar num ambiente assim", escreveu ele num blog ao anunciar o pedido de demissão. "Quando uma empresa é repleta de engenheiros, ela recorre à engenharia para resolver problemas. Reduz todas as decisões a um simples problema de lógica. Os dados acabarão por se tornar essenciais para todas as decisões, paralisando a empresa."

O brilhantismo não depende dos dados. Steve Jobs talvez tenha continuamente melhorado o laptop Mac ao longo dos anos com base em relatórios, mas usou a intuição, não os dados, para lançar o iPod, iPhone e iPad. Ele confiava em seu sexto sentido.

"Não é função dos consumidores saber o que querem", disse ao contar para um repórter que a Apple não fez pesquisa de mercado antes de lançar o iPad.

No livro *Seeing like a State,* o antropólogo James Scott, da Yale University documenta as maneiras como os governos, por conta da adoração pela quantificação e dados, acabam piorando a vida das pessoas em vez de melhorá-la. Eles usam mapas para determinar como reorganizar comunidades em vez descobrir algo sobre as pessoas; usam tabelas sobre colheitas para tomar decisões sobre a agricultura sem nada saber sobre a atividade; usam todas as maneiras orgânicas e imperfeitas por meio das quais as pessoas vêm interagindo ao longo do tempo e ajustam às suas necessidades, às vezes apenas para satisfazer um desejo de ordem quantificada. O uso dos dados, na visão de Scott, às vezes serve para dar mais poder aos poderosos.

Essa é a ditadura dos dados, e foi algo semelhante que levou os Estados Unidos a prolongar a Guerra do Vietnã, em parte com base na contagem de corpos, em vez de em decisões mais importantes. "A verdade é que nem toda situação humana complexa pode ser reduzida a linhas num gráfico ou a porcentagens ou ainda a números numa tabela", disse McNamara, num discurso em 1967, à medida que os protestos contra a guerra aumentavam. "Mas é possível compreender todas as realidades. Não quantificar o que pode ser quantificado é somente se satisfazer com menos que toda a vastidão da razão." Se ao menos os dados certos fossem usados da maneira certa, e não respeitados apenas pelo bem dos próprios dados...

Robert Strange McNamara administrou o Banco Mundial nos anos 1970, depois se fez de pacifista nos anos 1980. Ele se tornou crítico declarado das armas nucleares e uma das principais figuras na proteção do meio ambiente. Mais tarde, passou por uma conversão intelectual e produziu um livro de memórias, *In Retrospect,* que criticava o raciocínio por trás da guerra e suas próprias decisões como secretário de defesa. "Estávamos errados, terrivelmente errados", escreveu ele. Mas ele estava se referindo à estratégia da guerra. No que diz respeito aos dados, e principalmente à contagem de corpos, ele permaneceu firme. McNamara admitiu que as estatísticas eram "equivocadas e erradas". "Mas o que você pode contar, você deve contar, inclusive a perda de vidas..." McNamara morreu em 2009, aos 93 anos, um homem inteligente, mas não sábio.

O big data pode nos levar a cometer o pecado de McNamara: nos fixarmos tanto em dados e ficarmos tão obcecados com o poder e a promessa que eles despertam, que não conseguiremos apreciar suas limitações. Para entender o equivalente do big data da contagem de corpos, precisamos recorrer novamente ao Google Flu Trend. Pense numa situação, não totalmente implausível, na qual um surto de gripe se abate sobre o país. Os médicos adorariam a possibilidade de prever, em tempo real, os lugares mais atingidos por meio de buscas na internet. Eles saberiam onde intervir para ajudar.

Mas suponha que, num momento de crise, os líderes políticos argumentem que o simples conhecimento dos pontos mais atingidos e a tentativa de impedi-la não bastam. Assim, eles determinam uma quarentena – não para todas as pessoas da região, o que seria desnecessário e exagerado. O big data permite que sejamos mais específicos. Assim, a quarentena se aplica somente a usuários da internet cujas buscam tenham mais correlação com a gripe. Aqui, temos os dados sobre quem sujeitar à quarentena. Agentes federais munidos de listas de IPs e informações de GPS levam os usuários para centros de quarentena.

Por mais sensato que o cenário pareça para alguns, é equivocado. As correlações não implicam relações de causa e efeito. As pessoas selecionadas podem ou não ter a gripe. Elas não foram examinadas; são prisioneiras de uma previsão e, mais importante, vítimas de uma visão dos dados que não aprecia o que a informação realmente significa. A questão do Google Flu Trend é a de que certos termos de busca estão *correlacionados* com o surto da doença – mas a correlação pode existir por causa de circunstâncias como colegas de trabalho saudáveis que ouvem espirros no escritório e procuram na internet formas de se proteger; não significa que estejam doentes.

O LADO NEGRO DO BIG DATA

Como vimos, o big data permite maior vigilância de nossas vidas ao mesmo tempo que torna obsoletos alguns meios legais de proteção à privacidade. Eles também inutilizam o método técnico essencial de preservação do anonimato. Tão perturbador quanto isso, as previsões do big data em relação às pessoas podem ser usadas para, na verdade, puni-las por suas propensões, não por suas ações, o que renega o livre-arbítrio e corrói a dignidade humana.

Ao mesmo tempo, há um risco real de que os benefícios do big data levarão as pessoas a aplicar as técnicas em situações nas quais eles não se enquadram, ou as farão se sentir demasiadamente confiantes nos resultados que analisam. À medida que as previsões do big data melhorarem, seu uso se tornará mais atraente e alimentará uma obsessão pelos dados, uma vez que estes podem realizar muito. Essa foi a maldição de McNamara e a lição de sua história.

Temos de nos proteger contra a exagerada dependência dos dados em vez de repetirmos o erro de Ícaro, que adorou seu poder técnico do voo, mas o usou inadequadamente e caiu no mar. No próximo capítulo, vamos analisar maneiras de controlar o big data, para que não sejamos controlados por ele.

Capítulo 9

Controle

As mudanças na maneira como produzimos e interagimos com a informação levam a mudanças nas regras que usamos para nos governar e nos valores que a sociedade precisa proteger. Pense num exemplo anterior à enxurrada dos dados, era fomentada pela imprensa.

Antes que Johannes Gutenberg inventasse a imprensa com tipos móveis, por volta de 1450, a disseminação de ideias estava limitada às conexões pessoais. Os livros eram praticamente confinados às bibliotecas monásticas, guardados por monges que agiam em nome da Igreja Católica para proteger e preservar seu domínio. Fora da Igreja, os livros eram extremamente raros. Algumas universidades reuniam algumas dúzias ou talvez centenas de livros. A Cambridge University começou o século XV com apenas 122 tomos.

Poucas décadas depois da invenção de Gutenberg, a imprensa se replicara por toda a Europa, possibilitando a impressão em massa de livros e panfletos. Quando Martinho Lutero traduziu a Bíblia do latim para o alemão coloquial, as pessoas de repente tinham um motivo para se alfabetizar: ao ler a Bíblia sozinhas, elas podiam driblar os padres para aprender a palavra de Deus. A Bíblia se tornou um livro de sucesso, e, uma vez alfabetizadas, as pessoas continuavam a ler. Algumas até mesmo decidiram escrever. Em alguns anos, o fluxo de informações passou de um riacho para um rio torrencial.

A drástica mudança também adubou o terreno para novas regras que governavam a explosão de informação gerada pela imprensa de tipos móveis. Quando o Estado laico consolidou seu poder, estabeleceu a censura e o licenciamento para deter e controlar o mundo impresso. A lei de direitos autorais foi criada para dar aos autores incentivos jurídicos e econômicos para criar. Mais tarde, os intelectuais pressionaram por regras que protegessem as palavras da opressão governamental. No século XIX, em vários países, a liberdade de expressão foi transformada em garantia constitucional, mas com os direitos vieram responsabilidades. À medida que os jornais vasculhavam a intimidade e destruíam reputações, regras surgiam para proteger a privacidade das pessoas e permitir que elas processassem casos de injúria.

Porém, essas mudanças também refletem uma fundamental e profunda transformação dos valores. À sombra de Gutenberg, começamos a perceber o poder da palavra escrita – e, por fim, a importância das informações que se espalham pela sociedade. Com a passagem dos séculos, optamos por mais informações e por nos protegermos de seus excessos não por meio da censura, e sim de regras que limitassem os usos arbitrários das informações.

À medida que o mundo se move rumo a uma era de big data, a sociedade também passa por semelhante transformação tectônica. O big data já está transformando muitos aspectos de nossas vidas e o modo como pensamos, e nos obriga a reconsiderar princípios básicos sobre como estimular seu crescimento e diminuir o potencial destrutivo. Mas, ao contrário de nossos antecessores durante e antes da revolução impressa, não temos séculos para nos ajustar; talvez apenas uns poucos anos.

Simples alterações das regras existentes não serão suficientes para controlar a era do big data e abrandar seu lado negro. Em vez de novos parâmetros, a situação pede uma mudança pragmática. Proteger a privacidade exige que os usuários de big data tenham mais responsabilidade sobre suas ações. Ao mesmo tempo, a sociedade terá de redefinir a própria noção de justiça para garantir a liberdade de agir (e, assim, de ser responsável por suas ações). Por fim, novas instituições e profissionais serão necessários para interpretar os complexos algoritmos que perfazem as descobertas do big data e para defender as pessoas que podem ser atingidas por ele.

DA PRIVACIDADE À RESPONSABILIDADE

Durante décadas, um princípio essencial das leis de privacidade em todo o mundo foi colocar as pessoas no controle, deixado que elas decidissem se, como e por quem suas informações pessoais podiam ser processadas. Na era da internet, este louvável ideal foi transformado no imediatista sistema de "leitura e consentimento". Na era do big data, quando boa parte do valor dos dados está no uso secundário, desconsiderado no momento da coleta dos dados, esses mecanismos para garantir a privacidade já não se aplicam.

Imaginamos uma mentalidade de privacidade bem diferente para a era do big data, menos focada no consentimento individual na hora da coleta e mais na responsabilização dos usuários por seus atos. Neste mundo, as empresas irão formalmente avaliar a reutilização dos dados com base no seu impacto sobre as pessoas cujas informações pessoais estão sendo processadas. O processo não tem de ser onerosamente detalhado em todos os casos, já que as futuras leis de privacidade definirão amplas categorias de uso, incluindo com ou sem proteções padronizadas e limitadas. Para iniciativas mais arriscadas, os reguladores estabelecerão regras básicas sobre como os usuários deveriam avaliar os perigos de usos específicos e determinar a melhor maneira de evitar ou mitigar problemas em potencial. Isso fomenta o uso criativo dos dados e, ao mesmo tempo, garante que medidas foram tomadas para que não haja prejudicados.

Administrar uma avaliação formal e correta do uso do big data e implementar as descobertas com precisão trazem benefícios palpáveis para os usuários dos dados: eles terão liberdade para procurar usos secundários de dados pessoas em vários casos sem ter de recorrer à pessoa para obter consentimento. Por outro lado, avaliações ruins ou más implementações da proteção deixarão os usuários expostos a responsabilidades legais e ações regulatórias, como mandatos, multas e até mesmo processos criminais. A responsabilidade dos usuários de dados só funciona se tiver força.

Para ver como isso pode ocorrer na prática, veja o exemplo da dataficação das costas, mencionado no Capítulo 5. Imagine que uma empresa vendeu um serviço antirroubo, pelo qual a postura do motorista é o único fator que o identifica. Mais

tarde, o sistema reanalisa as informações para prever "estados de alerta" do motorista, tais como se está sonolento, desperto ou irritado, a fim de enviar alertas para os outros motoristas próximos, para evitar acidentes. De acordo com as leis atuais, a empresa pode acreditar que precisa de uma nova rodada de consentimento porque não recebeu permissão para usar as informações para esse fim. Mas sob um sistema de responsabilização dos usuários de dados, a empresa avaliará os perigos do uso pretendido e, se descobrir que são mínimos, pode ir adiante – e, assim, melhorar a segurança nas ruas.

Transferir o ônus da responsabilidade do público para os usuários de dados faz sentido por vários motivos. Eles, melhor que ninguém, incluindo consumidores e reguladores, sabem como pretendem usar os dados. Ao realizar a avaliação por si só (ou ao contratar especialistas para isso), eles evitarão o problema de revelar estratégias empresariais confidenciais para terceiros. Talvez o mais importante: os usuários de dados aproveitam ao máximo seu uso secundário, de modo que é apenas justo torná-los responsáveis por suas ações e lhes transferir o ônus da análise.

Com essa mentalidade alternativa de privacidade, os usuários dos dados já não terão a obrigação legal de apagar informações pessoais que serviram ao propósito inicial, como a maioria das leis atualmente exige. É uma mudança importante, já que, como vimos, somente aproveitando o valor latente dos dados, os maurys posteriores prosperarão ao extrair o valor do dado para seu próprio benefício – e da sociedade. Em vez disso, os usuários dos dados terão permissão para manter informações pessoais por mais tempo, mas não para sempre. A sociedade precisa analisar cuidadosamente as recompensas da reutilização contra os riscos da exposição demasiada dos dados.

Para conseguir o equilíbrio adequado, os reguladores talvez optem por cronogramas diferentes para a reutilização, dependendo do risco inerente dos dados, assim como por diferentes valores sociais. Algumas nações podem tomar mais cuidados que outras, assim como alguns dados podem ser considerados mais sensíveis que outros. Essa abordagem também acaba com o espectro da "memória permanente" – o risco de que uma pessoa não possa escapar do passado porque os registros digitais sempre podem ser vasculhados. De outro modo, nossos dados pessoais pairam sobre nós como a Espada de Dâmocles, ameaçando nos punir com algum detalhe particular ou alguma compra de que nos arrependemos. Os limites de tempo também criam um incentivo para que os detentores dos dados os usem antes de perdê-los, o que enfatiza o que acreditamos ser um melhor equilíbrio para a era do big data: as empresas têm o direito de usar dados pessoais por mais tempo, mas, em troca, têm de assumir a responsabilidade por sua utilização além de ter a obrigação de apagar os dados pessoais depois de um tempo.

Além de uma mudança regulatória da "privacidade por consentimento" para "privacidade por meio de responsabilidade", imaginamos uma inovação técnica para ajudar a proteger a privacidade em certos casos. Uma abordagem inovadora é o conceito de "privacidade diferencial": o ofuscamento deliberado dos dados de modo que uma busca num grande banco de dados não revele resultados exatos, apenas aproximados, o que dificulta e encarece a associação de dados específicos às pessoas.

A mistura de informações soa como destruição de valiosas ideias, mas não é necessariamente verdade. A mudança pode ser favorável. Por exemplo, especialistas em diretrizes tecnológicas notam que o Facebook conta com uma forma de privacidade

diferencial ao fornecer informações sobre os usuários para anunciantes em potencial: os números são aproximados, de modo que não revelem as identidades individuais. Procurar por mulheres asiáticas em Atlanta interessadas em ioga Ashtanga gerará um resultado de "aproximadamente 400", não um número exato, tornando impossível que se use esta informação para encontrar um alvo específico.

A mudança nos controles, do consentimento individual para a responsabilidade dos usuários dos dados, é uma alteração fundamental para o uso eficiente do big data, mas não é a única.

PESSOAS *VERSUS* PREVISÕES

Os tribunais consideram as pessoas responsáveis por suas ações. Depois de um julgamento justo, quando os juízes tomam decisões imparciais, fez-se a justiça. Mas, na era do big data, a noção de justiça precisa ser redefinida para preservar a ideia do agente humano: o livre-arbítrio, por meio do qual as pessoas escolhem como agir. É a simples ideia de que cada um pode e deve ser responsabilizado por seu comportamento, não por sua propensão.

Antes do big data, a liberdade fundamental era óbvia, tanto que raramente precisava ser articulada. Afinal, é assim que o sistema jurídico funciona: responsabilizamos as pessoas por seus atos ao avaliarmos o que fizeram. Por outro lado, com o big data, podemos prever as ações humanas com precisão cada vez maior, o que nos leva a julgar as pessoas não pelo que fizeram, mas pelo que prevemos que fariam.

Na era do big data, teremos de expandir nossa compreensão da justiça e exigir que ela inclua proteções para a ação das pessoas, tanto quanto atualmente protegemos os procedimentos jurídicos. Sem essas garantias, a própria noção de justiça estará comprometida.

Ao garantir a ação humana, nos certificamos de que a opinião governamental sobre nosso comportamento se baseie em ações reais, não em análises de big data. Assim, o governo deve nos considerar responsáveis por nossas ações passadas, não por previsões estatísticas de ações futuras. E, quando o governo julga ações passadas, deve evitar contar apenas com o big data. Pense, por exemplo, no caso de nove empresas acusadas de cartel. É totalmente aceitável usar análises de big data para identificar possíveis cartéis de modo que os reguladores possam investigar e criar um caso usando métodos tradicionais. Mas essas empresas não podem ser consideradas culpadas somente porque os dados sugerem que provavelmente cometeram um crime.

Um princípio semelhante deveria ser aplicado fora do governo, quando empresas tomam importantes decisões a nosso respeito: contratar e demitir, oferecer uma hipoteca ou negar um cartão de crédito. Às empresas que baseiam decisões principalmente em previsões de big data, recomendamos algumas salvaguardas. A primeira é a transparência: disponibilizar os dados e o algoritmo que fundamenta a previsão que afeta uma pessoa. Depois, a certificação: ter um algoritmo certificado para certos usos por um terceiro especialista. A terceira é a refutação: especificar maneiras concretas pelas quais as pessoas possam refutar uma previsão a respeito de si mesmas - semelhante à tradição científica de refutar quaisquer fatores que possam minar as descobertas de um estudo.

Mais importante: uma garantia da interferência humana protege contra a ameaça de uma ditadura dos dados, na qual conferimos aos dados significado e importância maiores do que merecem.

É igualmente crucial que protejamos a responsabilidade individual. A sociedade enfrentará a tentação de deixar de responsabilizar as pessoas e pode tender a gerenciar riscos, isto é, basear as decisões quanto às pessoas em avaliações de possibilidades e probabilidades de resultados em potencial. Com tantos dados aparentemente objetivos disponíveis, pode parecer tentador eliminar os caráteres sentimental e individual da tomada de decisão, contar com algoritmos em vez de avaliações subjetivas de juízes e analistas e estruturar as decisões não na responsabilidade pessoal, e sim em termos de riscos mais "objetivos" e de como evitá-los.

Por exemplo, o big data nos tenta a prever quais pessoas têm maior probabilidade de cometer crimes e a sujeitá-las a tratamento especial, investigando-as exaustivamente em nome da redução de riscos. As pessoas assim categorizadas podem sentir, com toda razão, que estão sendo punidas sem sequer ser confrontadas e responsabilizadas por seu real comportamento. Imagine que um algoritmo identifique que determinado adolescente tem alta probabilidade de cometer um crime nos próximos três anos. Como resultado, as autoridades designam um assistente social para visitá-lo mensalmente, mantê-lo sob vigilância e tentar ajudá-lo a ficar longe de problemas.

Se o adolescente e seus parentes, amigos, professores ou chefes virem as visitas como um estigma, o que é provável, a intervenção terá o efeito de uma punição, um castigo por uma ação que não aconteceu. A situação não é muito melhor se as visitas não forem vistas como castigo, mas como simples tentativa de reduzir a probabilidade de problemas futuros – como uma maneira de diminuir o risco (neste caso, o risco de um crime que afetará a segurança pública). Quanto mais deixarmos de responsabilizar as pessoas por seus atos em nome de intervenções com base em dados para reduzir os riscos na sociedade, mais desvalorizaremos o ideal de responsabilidade individual. O Estado previdente é o que toma conta da sociedade. Negar a responsabilidade das pessoas por suas ações destrói sua liberdade fundamental de escolher seu comportamento.

Se o Estado baseia muitas decisões em previsões e num desejo de mitigar os riscos, nossas escolhas individuais – e, assim, nossa liberdade individual de agir – já não importam. Sem culpa, não pode haver inocência. Ceder a essa abordagem não melhorará a sociedade, ao contrário, a piorará.

Um pilar fundamental da governança com big data deve ser uma garantia de que continuaremos a julgar as pessoas pela responsabilidade pessoal e pelo comportamento de fato, não pela análise "objetiva" dos dados para determinar se são criminosas em potencial. Somente assim as trataremos como seres humanos: como pessoas que têm liberdade de escolha sobre suas ações e o direito de serem julgadas por elas.

QUEBRANDO A CAIXA-PRETA

Os sistemas computacionais de hoje baseiam suas decisões em regras às quais foram programados para seguir. Assim, quando uma decisão é equivocada, como é inevitável de tempos em tempos, podemos voltar e descobrir por que o computador cometeu aquele erro. Podemos, por exemplo, investigar questões como "Por que o

sistema de piloto automático inclinou o avião cinco graus a mais quando o sensor externo detectou um repentino aumento de umidade?" O código computacional de hoje pode ser aberto e inspecionado, e aqueles que sabem interpretá-lo podem investigar e compreender a base de suas decisões, por mais complexas que sejam.

Porém, com a análise de big data, essa investigação será muito mais difícil. A base de uma previsão de algoritmo pode ser complexa demais para que a maioria das pessoas entenda.

Quando os computadores foram explicitamente programados para seguir instruções, como o primitivo programa de tradução do russo para o inglês da IBM, de 1954, um ser humano podia entender por que o programa substituiu uma palavra por outra, Mas o Google Translate incorpora bilhões de páginas de traduções em suas escolhas sobre se a palavra "light" deve ser "lumière" ou "léger" em francês (isto é, se a palavra se refere à luminosidade ou à leveza). É impossível para um ser humano investigar os motivos precisos para as escolhas de palavras do programa porque se baseiam em enormes bancos de dados e cálculos estatísticos.

O big data opera numa escala que transcende nossa compreensão comum. Por exemplo, a correlação que a Google identificou entre alguns termos de busca e a gripe foi o resultado de testes com 450 milhões de modelos matemáticos. Por outro lado, Cynthia Rudin inicialmente criou 106 índices de previsão para descobrir se um bueiro pegaria fogo, e ela podia explicar para os gerentes da Con Edison por que o programa priorizava certos lugares. A "explicabilidade", como é chamada nos círculos de inteligência artificial, é importante para nós, mortais, que tendemos a querer saber o por quê, não apenas o quê. Mas e se, em vez de 106 índices, o sistema automaticamente gerasse 601 índices, a maioria dos quais com pouco peso, mas que, quando reunidos, aumentavam a precisão do modelo? A base de qualquer previsão pode ser incrivelmente complexa. O que ela diria, então, aos gerentes para convencê-los a realocar os orçamentos limitados?

Nesses cenários, podemos ver o risco de que as previsões de big data, e os algoritmos e bancos de dados por trás deles, se tornem caixas-pretas que não nos ofereçam responsabilidade, rastreabilidade ou confiança. Para evitar isso, o big data exigirá monitoramento e transparência, que por sua vez exigirão novos tipos de conhecimento e instituições. Esses novos *players* apoiarão áreas nas quais a sociedade precisa investigar previsões de big data e permitirá que pessoas que se sentirem atingidas por ele busquem reparações.

Como sociedade, geralmente vemos essas novas entidades surgirem quando um drástico aumento de complexidade e especialização de determinado campo gera uma urgente necessidade de especialistas para gerenciar as novas técnicas. Áreas como direto, medicina, contabilidade e engenharia passaram por essa mesma mudança há mais de um século. Recentemente, especialistas em segurança e privacidade da informação passaram a certificar que as empresas estejam de acordo com as melhores práticas determinadas por departamentos como a International Organization for Standards, criada para resolver uma nova necessidade de orientações neste campo.

O big data exigirá um novo grupo de pessoas para assumir esse papel. Talvez sejam chamadas de "algoritmistas". Assim como as empresas têm contadores internos e

auditores externos que reveem as finanças, os algoritmistas também podem assumir duas formas: entidades independentes para monitorar empresas externamente ou funcionários e departamentos para monitorá-las internamente.

A ASCENSÃO DO ALGORITMISTA

Esses novos profissionais serão especialistas nas áreas de ciência da computação, matemática e estatística e agirão como revisores das análises e previsões do big data. Os algoritmistas farão um voto de imparcialidade e confidencialidade, como contadores e outros profissionais fazem hoje. Eles avaliarão a seleção de fontes de dados, a escolha de instrumentos analíticos e de previsão, incluindo algoritmos e modelos, e a interpretação dos resultados. No caso de uma contestação, eles teriam acesso aos algoritmos, abordagens estatísticas e bancos de dados que geraram determinada decisão.

Se houvesse um algoritmista na equipe do Department of Homeland Security (Departamento de Segurança Interna dos Estados Unidos) em 2004, talvez tivesse evitado que a agência criasse uma lista de passageiros indesejáveis tão falha a ponto de incluir o senador Kennedy. Casos mais recentes, nos quais os algoritmistas poderiam ser úteis, aconteceram no Japão, França, Alemanha e Itália, onde as pessoas reclamaram que a característica de "autocompletar" da Google, que gera uma lista de termos comuns associados a um nome digitado, as difamavam. A lista se baseia na frequência de buscas anteriores: termos classificados pela probabilidade matemática. Mas quem não ficaria com raiva se a palavra "culpado" ou "prostituta" aparecesse ao lado do nosso nome quando parceiros comerciais ou alguém em quem estejamos interessados afetivamente nos procurassem na internet?

Vemos os algoritmistas como pessoas que poderão conferir uma abordagem mercadológica para problemas como esse, que podem levar a formas mais invasivas de regulação. Eles preencherão uma necessidade semelhante à preenchida por contadores e auditores quando surgiram, no começo do século XX, para lidar com uma enxurrada de informações financeiras. As pessoas tinham dificuldade para entender a avalanche numérica, que exigia especialistas organizados de uma maneira ágil e autorregulatória. O mercado respondeu dando origem a um novo setor de empresas especializadas na vigilância financeira. Ao oferecerem o serviço, um novo tipo de profissional devolveu a confiança da sociedade na economia. O big data poderia e deveria tirar proveito de semelhante confiança que os algoritmistas seriam capazes de prover.

Algoritmistas externos

Antevemos algoritmistas externos agindo como auditores imparciais para rever a precisão ou validade das previsões de big data sempre que o governo exigir, como sob mandado judicial ou regulação. Eles também podem ter empresas de big data como clientes e realizar auditorias para as companhias que precisam de apoio de especialistas. Podem ainda certificar a relevância da aplicação do big data em técnicas antifraude ou sistemas de comercialização de ações. Por fim, algoritmistas externos estão preparados para dar consultoria a agências governamentais quanto ao melhor uso do big data no setor público.

Como na medicina, direito e outras áreas, imaginamos que essa nova profissão se autorregule com um código de conduta. A imparcialidade, confiabilidade, competência e profissionalismo dos algoritmistas serão governados por duras regras; se eles não conseguirem cumprir os padrões, estarão sujeitos a processos. Eles também podem ser convocados para depor como especialistas em julgamentos ou a agir como especialistas nomeados por juízes para ajudá-los com questões técnicas ou casos especialmente complexos.

Além disso, pessoas que acreditam ter sido prejudicadas por previsões de big data — um paciente que teve a cirurgia rejeitada, um prisioneiro que teve a condicional negada ou um cliente que não recebeu a hipoteca – podem recorrer aos algoritmistas como hoje recorrem a advogados para ajudá-las a entender e recorrer das decisões.

Algoritmistas internos

Algoritmistas internos trabalham dentro de uma organização para monitorar as atividades de big data. Eles cuidam não apenas dos interesses da empresa como também dos das pessoas afetadas pelas análises de big data. Eles supervisionam operações de big data e são o primeiro contato de qualquer pessoa afetada pelas previsões da empresa. Também examinam análises de big data à procura de integridade e precisão antes de permitir que ganhem vida. A fim de realizar o primeiro desses papéis, os algoritmistas devem ter certo nível de liberdade e imparcialidade dentro da empresa para a qual trabalham.

A ideia de uma pessoa que trabalha para uma empresa e se mantenha imparcial quanto às operações pode parecer contraditória, mas essas situações são bastante comuns. As divisões de monitoramento nas grandes instituições financeiras são um exemplo; assim como os conselhos diretores de várias empresas, cujas responsabilidades são para com os acionistas, e não de gerenciamento. Muitas empresas de mídia, incluindo o *New York Times* e o *Washington Post,* empregam *ombudsmen*, cuja responsabilidade principal é defender a confiança pública. Esses funcionários lidam com reclamações dos leitores e geralmente criticam publicamente o empregador quando concluem que cometeu um erro.

E há um cargo ainda mais próximo do algoritmista interno – profissional encarregado de garantir que a informação pessoal não seja usada de forma equivocada no ambiente corporativo. A Alemanha, por exemplo, exige que empresas de determinado porte (geralmente com 10 ou mais pessoas empregadas no processamento de informações pessoais) contratem um representante de proteção dos dados. Desde os anos 1970, esses representantes internos desenvolveram uma ética profissional própria. Eles se reúnem regularmente para compartilhar as melhores práticas e treinamentos e têm publicação especializada própria e conferências. Além disso, conseguiram manter a lealdade para com os empregadores *e* funções como analistas imparciais, agindo como *ombudsmen* da proteção dos dados e também conferindo valores de privacidade da informação em todas as operações da empresa. Acreditamos que algoritmistas internos possam fazer o mesmo.

GOVERNANDO OS BARÕES DOS DADOS

Os dados são para a sociedade da informação o que o combustível é para a economia industrial: o recurso crítico que possibilitou inovações com as quais as pessoas podem

contar. Sem um vigoroso suprimento de dados e um mercado robusto de serviços, a criatividade e a produtividade, hoje possíveis, estariam reprimidas.

Neste capítulo, dissertamos sobre três novas estratégias fundamentais para o controle do big data no que se refere à privacidade, propensão e auditoria algorítmica. Temos certeza de que, com a aplicação dessas estratégias, o lado negro do big data será domado. Mas à medida que a incipiente indústria de big data se desenvolve, um desafio crítico adicional será proteger os mercados competitivos do big data. Devemos prevenir o surgimento de barões do big data do século XXI, o equivalente aos barões da borracha do século XIX, que dominaram as ferrovias, metalúrgicas e redes telegráficas dos Estados Unidos.

Para controlar esses industriais, os Estados Unidos estabeleceram regras antitrustes extremamente adaptáveis. Originalmente criadas para as ferrovias dos anos 1800, elas foram mais tarde aplicadas a empresas que dominavam fluxos de informações dos quais outras empresas dependiam, como a National Cash Register nos anos 1910, a IBM nos anos 1960 e, mais tarde, a Xerox nos anos 1970, AT&T nos anos 1980, a Microsoft nos anos 1990 e a Google hoje em dia. As tecnologias nas quais essas empresas foram pioneiras se tornaram componentes essências da "infraestrutura de informação" da economia e exigiram a força da lei para evitar o domínio prejudicial.

A fim de garantir as condições de um mercado agitado por big data, precisaremos de medidas comparáveis às que estabeleceram a competição e a supervisão nos primórdios da tecnologia. Devemos permitir as transações de dados por meio do licenciamento ou interoperabilidade. Essa possibilidade dá origem à questão sobre se a sociedade pode se beneficiar de um cuidadoso e equilibrado "direito de exclusão" para os dados (algo semelhante ao direito autoral, por mais provocativo que pareça!). Será difícil para os legisladores – e um risco para nós.

É obviamente impossível prever como uma tecnologia se desenvolverá; nem mesmo o big data pode prever como ele mesmo evoluirá. Os reguladores precisarão encontrar um equilíbrio entre a ação cautelosa e ousada – e a questão das leis antitrustes aponta uma maneira de se conseguir isso.

A lei antitruste combatia o poder abusivo. Seus princípios se traduziam impressionantemente bem de um setor a outro e a ramos de atuação diferentes. É o tipo de regulação forte – que não favorece uma tecnologia em detrimento de outra –, o que é útil, uma vez que protege a competição sem pressupor que fará muito além disso. Portanto, a lei antitruste pode ajudar o big data assim como ajudou as ferrovias. Da mesma forma que alguns dos maiores detentores de dados do mundo, os governos devem divulgar publicamente seus dados. O encorajador é que alguns já fazem isso – até certo ponto.

A lição da lei antitruste é que, uma vez identificados os princípios gerais, os reguladores podem implementá-los para garantir salvaguardas e apoios. Do mesmo modo, as três estratégicas que propomos – a mudança da proteção da privacidade, do consentimento individual para a responsabilidade dos usuários de dados; a proteção da ação humana em meio a previsões; a invenção de uma nova casta de auditores de big data, que chamamos de algoritmistas — podem servir de base para o controle eficiente e justo das informações na era do big data.

Em muitos campos, da tecnologia nuclear à bioengenharia, primeiro criamos instrumentos que possam ser prejudiciais e só mais tarde inventamos mecanismos

para nos proteger deles. Neste sentido, o big data se posiciona como outras áreas da sociedade, que apresentam desafios sem soluções absolutas, só com questões contínuas sobre como organizamos o mundo. Todas as gerações devem resolver essas novas questões. Nossa função é reconhecer os perigos dessa potente tecnologia, apoiar seu desenvolvimento e avaliar as recompensas.

O big data, assim como a imprensa, gera mudanças na maneira como a sociedade se autocontrola. Eles nos obrigam a resolver desafios antigos de novas maneiras e a confrontar novas preocupações pelo uso de princípios de honra. Para garantir a proteção das pessoas e, ao mesmo tempo, o desenvolvimento da tecnologia, não devemos deixar que o big data se desenvolva para além da capacidade humana de moldar a tecnologia.

Capítulo 10

Depois

Mike Flowers era advogado da procuradoria distrital de Manhattan no início dos anos 2000 e tinha várias incumbências: cobrir desde homicídios a crimes de Wall Street, até entrar para um renomado escritório. Depois de um ano entediante de trabalho burocrático, ele decidiu pedir demissão desse emprego também. Procurando algo mais significativo, pensou em ajudar a reconstruir o Iraque. Um sócio do escritório fez algumas ligações para as pessoas certas. Quando Flowers percebeu, se dirigia rumo à Zona Verde, a área segura das tropas americanas no centro de Bagdá, como parte da banca do julgamento de Saddam Hussein.

A maior parte do trabalho era logístico, não jurídico. Ele precisava identificar áreas de supostas covas coletivas para saber onde os investigadores deveriam cavar. Ele precisava transportar testemunhas para a Zona Verde sem que elas explodissem com os muitos ataques a bomba que faziam parte da triste realidade. Flowers notou que o Exército lidava com essas tarefas como problemas de informação. Os dados vieram ao resgate. Os analistas da inteligência combinavam relatórios de campo com detalhes sobre a localização, hora e baixas de ataques à bomba a fim de prever a rota mais segura naquele dia.

No retorno a Nova York, anos mais tarde, Flowers percebeu que os métodos eram uma importante forma de combate ao crime, que ele jamais tivera à disposição como procurador, e encontrou no prefeito da cidade, Michael Bloomberg, que fizera fortuna com os dados ao fornecer informações financeiras para bancos, um verdadeiro parceiro. Flowers foi nomeado para uma força-tarefa especial criada para analisar os números que podiam revelar os vilões do escândalo das hipotecas de 2009. A unidade teve tanto sucesso que, no ano seguinte, o prefeito Bloomberg pediu que ela expandisse sua área de atuação. Flowers se tornou o primeiro "diretor de análise" da cidade. Sua missão: criar uma equipe com os melhores cientistas que pudesse encontrar e usar os enormes bancos de dados da cidade para alcançar a eficiência em todas as áreas.

Flowers lançou mão de sua rede para encontrar as pessoas certas. "Não estava interessado em estatísticos experientes", diz. "Estava um pouco preocupado com o fato de relutarem em usar a nova abordagem para a solução do problema." Antes, quando entrevistara estatísticos tradicionais para o projeto da fraude financeira, eles tendiam a suscitar incompreensíveis preocupações sobre os métodos matemáticos. "Sequer pensava no modelo matemático que usaria. Queria uma ideia prática, era só o que importava", diz. Por fim, escolheu um grupo de cinco pessoas, que chama de "os garotos". Todos, exceto um, eram economistas recém-formados e sem muita experiência em viver numa cidade grande; todos eram um tanto criativos.

Entre os primeiros desafios que a equipe enfrentou estavam as "conversões ilegais" – prática de dividir habitações em unidades menores de modo que possam abrigar até 10 vezes mais pessoas que o planejado. Um emaranhado de extensões serpenteia pelas paredes; fogareiros ficam perigosamente sobre as camas. As pessoas que vivem assim geralmente morrem em incêndios. Em 2005, dois bombeiros morreram tentando resgatar moradores. A cidade de Nova York recebe cerca de 25 mil reclamações de conversões ilegais por ano, mas tem apenas 200 inspetores para lidar com elas. Parecia não haver uma boa maneira de distinguir casos de simples reclamações dos prontos para explodir em chamas. Para Flowers e seus garotos, porém, tudo parecia um problema que podia ser resolvido por dados.

Eles começaram com uma lista de todas as propriedades da cidade – 900 mil. Depois, usaram bancos de dados de 19 departamentos diferentes, indicando, por exemplo, se o proprietário do edifício pagava os impostos prediais, se houvera processo de retomada do imóvel e se irregularidades no uso de aparelhos ou atrasos de pagamentos levaram a cortes nos serviços. Eles também usaram informações sobre o tipo e data da construção, além de visitas de ambulâncias, taxas de criminalidade, reclamações por conta de ratos e outros. Depois, compararam todas as informações com cinco anos de dados de incêndios classificados por gravidade e procuraram correlações a fim de gerar um sistema que pudesse prever quais reclamações deveriam ser investigadas com mais urgência.

Inicialmente, boa parte dos dados não estava num formato utilizável. Os arquivistas da cidade não usavam uma forma padronizada para descrever os lugares; toda agência e departamento pareciam ter uma forma própria. O departamento de construções confere um número a cada estrutura. O departamento de preservação tem um sistema numérico diferente. A fazenda municipal confere a cada propriedade um identificador com base no distrito, quarteirão e terreno. A polícia usa coordenadas cartesianas. O departamento anti-incêndio conta com um sistema de proximidade a "cabines telefônicas", associadas à localização dos quartéis, mesmo que as cabines não sejam mais usadas. Os meninos de Flowers organizaram a bagunça ao criar um sistema que identifica, com base em coordenadas cartesianas, os prédios por meio de uma pequena área na frente da propriedade e depois usa dados de geolocalização dos bancos de dados de outras agências. O método era intrinsecamente inexato, mas a quantidade de dados que eles eram capazes de usar compensava as limitações.

Os membros da equipe não estavam satisfeitos apenas com a reunião dos números. Eles saíram com os inspetores para observar seu trabalho. Os meninos tomaram notas e questionaram todas as vantagens. Quando um inspetor reclamou que o prédio que estavam prestes a examinar não seria um problema, os nerds quiseram saber por que ele estava tão seguro disso. O inspetor não sabia dizer, mas os meninos aos poucos determinaram que a intuição se baseava no novo revestimento do edifício, o que sugeria que o proprietário cuidava do lugar.

Os meninos voltaram às baias e se perguntaram como podiam incluir "revestimento recente" em seu modelo. Afinal, os revestimentos não eram (ainda) dataficados. Mas certamente era necessária a permissão da cidade para a realização de qualquer reforma na fachada. O acréscimo da permissão melhorou o desempenho de previsão

do sistema, ao indicar que algumas propriedades suspeitas provavelmente *não* eram de grande risco.

A análise acabou por mostrar que algumas maneiras consagradas de ação não eram as melhores, assim como os olheiros em *O homem que mudou o jogo* tiveram de aceitar os problemas da escolha com base na intuição. A quantidade de ligações para o 311, número para reclamações da cidade, por exemplo, era levada em conta para indicar quais prédios precisavam de mais atenção. Mais ligações significam mais problemas sérios. Mas essa medida se revelou equivocada. Um rato visto no imponente Upper East Side podia gerar 30 ligações em uma hora, mas talvez fosse necessário um batalhão de roedores antes que os residentes no Bronx se dessem o trabalho de ligar para o 311. Do mesmo modo, a maioria das reclamações por conversão ilegal podia ser sobre barulho, não sobre condições perigosas.

Em junho de 2011, Flowers e os rapazes acionaram o sistema. Toda reclamação incluída na categoria de conversão ilegal era processada em questão de semanas. Eles reuniam as que tinham mais risco de incêndio e as passavam para os inspetores para imediata investigação. Quando obtiveram os resultados, todos ficaram impressionados.

Antes das análises de big data, os inspetores investigavam as reclamações que pareciam mais desesperadoras, mas em apenas 13% dos casos encontravam condições graves o suficiente para um mandado de evacuação. Agora eles emitiam mandados em mais de 70% dos prédios inspecionados. Ao indicar quais prédios precisavam de mais atenção, o big data quintuplicou sua eficiência. O trabalho se tornou mais satisfatório: o grupo se concentrava em problemas maiores. A nova eficiência dos inspetores teve outros benefícios também. Os incêndios em conversões ilegais apresentam probabilidade 15 vezes maior de resultar em ferimento ou morte dos bombeiros, por isso o departamento anti-incêndio adorou. Flowers e os meninos pareciam magos, com uma bola de cristal que lhes permitia ver o futuro e prever quais lugares apresentavam mais risco. Eles usaram enormes quantidades de dados, disponíveis há anos, em geral sem uso depois de coletados, e os usaram de uma nova maneira para extrair valor. Usar um corpo maior de informação lhes permitiu ver conexões indetectáveis em pequenas quantidades – a essência do big data.

A experiência dos alquimistas analíticos de Nova York enfatiza muitos dos temas deste livro. Eles usaram uma enorme quantidade de dados; a lista de prédios na cidade representava nada menos que o N1 = 1tudo. Os dados eram confusos, como a informação sobre a localização ou registros de ambulâncias, mas isso não os deteve. Na verdade, os benefícios de usar mais dados superaram os problemas de uma informação menos precisa. Eles foram capazes de alcançar os resultados porque várias características da cidade foram dataficadas (ainda que inconsistentemente), o que permitiu que eles processassem a informação.

Os especialistas tiveram de se render à abordagem do big data. Ao mesmo tempo, Flowers e os meninos testaram continuamente o sistema com inspetores veteranos, usando sua experiência para extrair melhor desempenho do sistema. Mas o motivo mais importante para o sucesso do programa foi o fato de ele desprezar a causalidade em favor da correlação.

"Não estou interessado na causalidade, exceto quanto ela leva à ação", explica Flowers. "A causalidade é para outras pessoas e, francamente, é bem arriscado falar

em causalidade. Não acho que exista qualquer relação de causa entre o dia em que alguém pede a devolução judicial de uma propriedade e o fato de o lugar ter ou não um histórico risco de incêndio estrutural. Acho que seria obtuso pensar assim. E ninguém sairia por aí dizendo isso. Eles pensariam: 'Não, são os fatores subjacentes.' Mas não quero sequer entrar neste assunto. Precisaria de um ponto de dado específico ao qual tenho acesso e que me diz sua importância. Se for importante, então agiremos sobre ele. Caso contrário, não. Sabe, temos problemas de verdade a serem solucionados. Não posso sair por ai, sinceramente, pensando em questões como causalidade neste momento."

QUANDO OS DADOS FALAM

Os efeitos do big data são muitos em termos práticos, à medida que a tecnologia é aplicada para encontrar soluções para problemas cotidianos. Mas é apenas o começo. O big data está destinado a reformular a maneira como vivemos, trabalhamos e pensamos. A mudança que enfrentamos é, de certo modo, maior que as geradas por inovações notáveis que drasticamente expandiram o escopo e a escala das informações na sociedade. As convicções estão mudando. Antigas certezas estão sendo questionadas. O big data requer nova discussão quanto ao caráter das tomadas de decisões, destino e justiça. Uma visão de mundo que pensávamos estar relacionada com causas é desafiada pela preponderância das correlações. A posse do conhecimento, que já significou o entendimento do passado, se transforma na capacidade de prever o futuro.

Essas questões são muito mais significativas do que as que se apresentaram quando nos preparamos para explorar o *e-commerce,* conviver com a internet, entrar na era dos computadores ou usar o ábaco. A ideia de que nossa busca por entender a causalidade é supervalorizada – que em muitos casos pode ser mais vantajoso desprezar o *por quê* em favor do *quê* — sugere que o assunto seja fundamental para nossa sociedade e existência. Os desafios propostos pelo big data talvez não nos deem respostas; são parte de um eterno debate quanto ao lugar do homem no universo e sua busca por um sentido em meio à confusão de um mundo caótico e incompreensível.

Por fim, o big data marca o momento em que a "sociedade da informação" finalmente realiza a promessa implícita em seu nome. Os dados ocupam o centro do palco. Todos aqueles bits que reunimos podem agora ser usados de formas inéditas para servir aos nossos propósitos e revelar novas formas de valor. Mas isso requer uma nova maneira de pensar que desafiará as instituições e até mesmo nosso senso de identidade. A única certeza é que a quantidade de dados vai continuar a aumentar, assim como o poder de processamento. Mas, enquanto as pessoas consideram o big data uma questão tecnológica, com foco no hardware ou software, acreditamos que a ênfase precisa mudar para o que acontece quando os dados se manifestam.

Podemos captar e analisar mais informações que nunca. A escassez dos dados já não é a característica que define nossos esforços para interpretar o mundo. Podemos utilizar muito mais dados e, em alguns casos, chegar próximo à totalidade, o que nos obriga a operar de formas não tradicionais e, em específico, alterar a ideia do que significa informações úteis.

Em vez de ficarmos obcecados pela precisão, exatidão, limpeza e rigor dos dados, podemos relaxar um pouco. Não deveríamos aceitar dados que sejam explicitamente

errados ou falsos, mas alguma confusão pode ser aceitável em troca da reunião de um conjunto mais abrangente de dados. Na verdade, em alguns casos, a grandiosidade e confusão podem até mesmo ser benéficas, uma vez que, quando tentamos usar apenas uma porção pequena e exata de dados, não conseguimos captar a abrangência dos detalhes, na qual boa parte do conhecimento se encontra.

Como as correlações podem ser encontradas com mais rapidez e a um custo mais baixo que a causalidade, tornam-se mais desejáveis. Ainda precisaremos de estudos causais e experimentos controlados com dados precisos em certos casos, como no teste dos efeitos colaterais de um medicamento ou no projeto de um componente de aviação, mas, para muitas necessidades cotidianas, saber *o quê* e não o *por quê* basta. As correlações de big data podem apontar para áreas promissoras nas quais é possível explorar relações causais.

As rápidas correlações nos permitem economizar dinheiro em passagens de avião, prever surtos de gripe e saber que bueiros e prédios superpopulosos inspecionar num mundo de recursos escassos. Elas podem permitir que planos de saúde deem cobertura sem exames físicos e diminuam o custo de lembrar os doentes de tomar seus medicamentos. Idiomas são traduzidos e carros são dirigidos com base em previsões feitas por meio de correlações de big data. O Walmart pode descobrir que sabor de Pop-Tarts deve estocar na loja antes de um furacão. (A resposta: morango.) Claro que a causalidade é útil quando você a tem. O problema é que é difícil obtê-la e, quando pensamos que conseguimos, geralmente estamos enganados.

Novas ferramentas, de processadores mais rápidos e mais memória a programas mais inteligentes e algoritmos, são apenas parte do motivo por que podemos fazer tudo isso. Apesar de os instrumentos serem importantes, um motivo mais fundamental é que temos mais dados, uma vez que mais aspectos do mundo estão sendo dataficados. Para deixar claro, a ambição humana de quantificar o mundo é bem anterior à revolução dos computadores. Mas as ferramentas digitais facilitam imensamente a dataficação. Os celulares não apenas podem rastrear para quem ligamos e para onde vamos, como também os dados que coletam podem ser usados para detectar se estamos doentes. Em pouco tempo, o big data poderá nos dizer se estamos apaixonados.

Nossa capacidade de inovar e fazer mais, melhor e mais rápido tem o potencial de gerar muito valor, criando novos vencedores e perdedores. Boa parte do valor dos dados virá dos usos secundários, do custo/benefício, não apenas do uso primário, como estávamos acostumados a pensar. Como resultado, para muitos tipos de dados, parece sensato coletar o máximo possível, retê-los até que agreguem valor e deixar que outros os analisem, se for mais adequado para extrair seu valor (desde que haja compartilhamento do lucro gerado pela análise).

As empresas que se colocarem em meio ao fluxo de informação e que coletarem dados prosperarão. Para usar o big data com eficiência, são necessárias habilidades técnicas e muita imaginação – uma mentalidade de big data. Mas o valor essencial talvez vá para os detentores. Às vezes, um bem importante não é apenas a informação visível, mas os dados criados pela interação das pessoas com a informação, que uma empresa inteligente pode usar para melhorar um serviço já existente ou para lançar um completamente novo.

Ao mesmo tempo, o big data apresenta enormes riscos. Ele torna ineficientes os mecanismos técnicos e legais por meio dos quais atualmente tentamos proteger a

privacidade. No passado, o que constituía informações pessoais era bem conhecido - nomes, número de identidade, imposto de renda e assim por diante – e, portanto, facilmente protegido. Hoje em dia, até mesmo os dados mais inócuos podem revelar a identidade de alguém, se quem os coleta tiver como processá-los. A anonimização ou ocultação não funciona mais. Além disso, vigiar uma pessoa hoje expressa uma invasão mais ostensiva da privacidade, já que as autoridades não querem mais obter o máximo de informações possíveis da pessoa, mas também suas relações, conexões e interações.

Além dos problemas de privacidade, esses usos do big data geram outra preocupação: o risco de que possamos julgar as pessoas não por seu comportamento, mas pela tendência sugerida pelos dados. À medida que as previsões do big data se tornam mais precisas, a sociedade pode usá-las para punir pessoas por comportamentos previstos – atos ainda não cometidos. Essas previsões são axiomaticamente irrefutáveis. Assim, as pessoas acusadas não podem se defender. Esse tipo de punição renega o conceito de livre-arbítrio e a possibilidade, ainda que pequena, de que uma pessoa escolha um caminho diferente. À medida que a sociedade determina a responsabilidade individual (e dá a punição), a escolha deve ser considerada inviolável. O futuro deve continuar moldável de acordo com nossa vontade. Se não, o big data comprometerá a própria essência da humanidade: o pensamento racional e a livre escolha.

Não há um meio seguro de se preparar totalmente para o mundo do big data; ele exigirá que estabeleçamos novos princípios de governança. Uma série de mudanças importantes em nossas práticas pode ajudar a sociedade a se acostumar com as características e problemas do big data. Temos de proteger a privacidade ao tirar a responsabilidade das pessoas e passá-la para os usuários dos dados. Num mundo de previsões, é essencial garantir que a vontade seja protegida e que preservemos não apenas a capacidade pessoal de escolha moral como também a responsabilidade individual pelos atos. A sociedade deve criar salvaguardas para permitir que a nova profissão de "algoritmistas" tenha acesso às análises de big data – de modo que o mundo, que se tornou mais aleatório por meio do big data, não se transforme numa caixa-preta, ao substituir uma forma de desconhecimento por outra.

O big data se tornará fundamental para entender e resolver muitos problemas mundiais. Para lidar com a mudança climática, é preciso analisar dados de poluição a fim de entender para onde voltar nossos esforços e encontrar maneiras de mitigar os problemas. Sensores colocados em todo o mundo, incluindo os contidos em *smartphones,* oferecem vários dados que nos permitirão analisar o aquecimento global com mais detalhes. Enquanto isso, a melhoria e a diminuição do custo da saúde, especialmente nos países mais pobres, estarão relacionadas, em grande parte, com a automação de tarefas que atualmente precisam da opinião humana, mas que poderiam ser feitas por um computador, como biópsias de células cancerígenas ou a detecção de infecções antes que os sintomas apareçam.

O big data já está sendo usado na economia e na prevenção de conflitos. Eles revelaram áreas de favelas africanas, vibrantes comunidades de atividade econômica, ao analisar o movimento dos usuários de celulares. Eles têm revelado áreas propensas a conflitos étnicos e indicado como as crises de refugiados podem ocorrer. Os usos do big data se multiplicarão à medida que a tecnologia for aplicada a mais aspectos da vida.

O big data nos ajuda a continuar fazendo o que já fazemos bem e nos permite inovar. Mas não se trata de uma varinha mágica. Ele não trará a paz mundial, não erradicará a pobreza nem gerarão o novo Picasso. Os novos dados não podem gerar um bebê – mas podem salvar os prematuros. Com o tempo, esperaremos que sejam usados em quase todos os aspectos da vida e talvez até nos assustaremos um pouco quando estiverem ausentes, do mesmo modo que esperamos que um médico peça um exame de raio X para revelar problemas imperceptíveis num exame físico.

À medida que se tornar mais comum, o big data poderá influenciar a maneira como pensamos no futuro. Há cerca de 500 anos, a humanidade passou por uma profunda mudança na percepção do tempo, como parte de uma mudança para uma Europa mais secular, com base na ciência e esclarecida. Antes disso, o tempo era cíclico, assim como a vida. Todos os dias (e anos) eram parecidos, e até mesmo o fim da vida parecia com o começo, com os adultos virando crianças de novo. Mais tarde, o tempo passou a ser visto como linear – uma sequência de dias na qual o mundo podia ser moldado, e a trajetória da vida, influenciada. Se antes o passado, o presente e o futuro podiam se fundir, agora a humanidade tinha um passado para consultar e um futuro para ansiar, à medida que moldava o presente.

Apesar de o presente poder ser moldado, o futuro passou de perfeitamente previsível para aberto, inexplorado – uma grande tela em branco que as pessoas podem preencher com seus valores e esforços. Uma das principais características dos tempos modernos é nossa impressão de nós mesmos como mestres do nosso destino; esse comportamento nos diferencia de nossos ancestrais, para os quais a regra era uma forma de determinismo. Mas as previsões do big data tornam o futuro menos aberto e intocado. Em vez de uma tela em branco, o futuro parece já rascunhado com leves traços, visíveis por aqueles que detêm a tecnologia. Esse cenário parece diminuir nossa capacidade de moldar o destino. A potencialidade é sacrificada no altar da probabilidade.

Ao mesmo tempo, o big data pode significar que somos para sempre prisioneiros de nossas ações passadas, que podem ser usadas contra nós em sistemas que pretendem prever nosso comportamento futuro: nunca podemos escapar do que aconteceu. "O passado é um prólogo", escreveu Shakespeare. O big data sacraliza isso de forma algorítmica, para o bem e para o mal. Será que um mundo de previsões acabará com o entusiasmo diante do nascer do sol, o desejo de imprimir a marca humana no mundo?

O oposto é mais provável. Saber como as ações se darão no futuro nos permitirá agir para prevenir problemas ou melhorar os resultados. Vamos encontrar alunos que começam a perder o rumo antes dos exames finais. Identificaremos pequenos cânceres e os trataremos antes que a doença tenha chance de se desenvolver. Veremos a probabilidade de uma gravidez adolescente indesejada ou de uma vida de crime e interviremos para mudar o cenário. Evitaremos que incêndios fatais consumam apartamentos superlotados de Nova York ao sabermos quais prédios inspecionar.

Nada é pré-ordenado, porque sempre podemos reagir à informação que recebemos. As previsões do big data não são imutáveis – são apenas resultados prováveis, o que significa que podemos alterá-los se quisermos. Podemos identificar como receber melhor o futuro e ser seu mestre, assim como Maury encontrou rotas naturais dentro

da vastidão dos ventos e ondas. E, para isso, não precisaremos compreender a natureza do cosmos ou provar a existência dos deuses — o big data bastará.

DADOS AINDA MAIORES

À medida que o big data transforma nossa vida – ao otimizá-la, melhorá-la, tornando-a mais eficiente e trazendo benefícios –, que papel resta para a intuição, a fé, a incerteza e a originalidade?

A grande lição do big data é que apenas o ato de agir melhor basta – sem a necessidade de qualquer compreensão mais profunda. Fazê-lo continuamente é virtuoso. Mesmo que não saibamos os resultados que nossos esforços trarão, serão melhores do que se não nos esforçássemos. Flowers e os meninos de Nova York talvez não personifiquem o esclarecimento dos sábios, mas salvam vidas.

O big data não é um mundo frio, de algoritmos e robôs. Há um papel essencial para as pessoas, com todos os seus defeitos, percepções equivocadas e erros, já que essas características estão associadas à criatividade humana, ao instinto e à genialidade. Os mesmos processos mentais confusos, que levam à ocasional humilhação ou equívocos, também geram sucessos e se deparam com nossa grandeza. Isso sugere que, ao aprendermos a aceitar os dados confusos por servirem a um propósito maior, devemos aceitar a inexatidão como parte do que significa ser humano. Afinal, a confusão é propriedade essencial do mundo e de nossas mentes; em ambos os casos, só nos beneficiamos por aceitá-la e aplicá-la.

Num mundo no qual os dados informam as decisões, o que resta para as pessoas, a intuição e a contrariedade dos fatos? Se todos apelarem para os dados e seus instrumentos, talvez o que reste como ponto central de diferenciação seja a imprevisibilidade: o elemento humano do instinto, do risco, do acidente e do erro.

Se assim fosse, haveria uma necessidade especial de encontrar um lugar para o humano: reservar um espaço para a intuição, o bom senso e a serendipidade, a fim de garantir que não sejamos dominados pelos dados e por respostas mecânicas. O interessante dos seres humanos é justamente o que os algoritmos e os chips de silício não revelam, o que não podem revelar pela impossibilidade de ser captado em dados. Não se trata de "o que é", e sim de "o que não é": o espaço vazio, as fissuras na calçada, o não falado e o impensado.

Essa visão tem importantes implicações para a ideia de progresso na sociedade. O big data nos permite experimentar com mais rapidez e explorar mais pistas. Essas vantagens deveriam gerar mais inovação. Mas a invenção se torna o que os dados não dizem, algo que nenhum dado pode confirmar ou corroborar, já que ainda não existe. Se Henry Ford tivesse consultado algoritmos para saber o que os consumidores queriam, teria inventado um "cavalo mais rápido" (para reformular sua famosa citação). Num mundo de big data, são as características mais humanas que precisaremos fomentar – a criatividade, a intuição e a ambição intelectual —, já que a inteligência é a fonte do progresso.

O big data é um recurso e um instrumento. Ele existe para informar, não explicar; ele aponta para a compreensão, mas também pode apontar para o equívoco, dependendo da boa ou má utilização. Por mais deslumbrante que consideremos o

poder do big data, não devemos deixar que seu brilho sedutor nos cegue em relação às inerentes imperfeições.

A totalidade de informações no mundo — o N1 = 1tudo — nunca pode ser reunida, armazenada ou processada pela tecnologia. O laboratório de partículas CERN, na Suíça, por exemplo, coleta menos de 0,1% da informação gerada durante seus experimentos — o restante, aparentemente inútil, se dissipa no éter. Mas isso não é novidade. A sociedade sempre foi assolada pelas limitações dos instrumentos que usamos para medir e conhecer a realidade, da bússola e sextante ao telescópio, radar e GPS atuais. Nossos instrumentos podem ser duas, dez ou milhares de vezes mais potentes amanhã do que são hoje, tornando o conhecimento atual minúsculo. O mundo atual do big data vai, em pouco tempo, parecer tão ridículo quanto os 4 kilobytes de memória do computador de controle da Apollo 11.

O volume do que somos capaz de coletar e processar sempre será uma fração da informação existente no mundo. Ele só pode ser um simulacro da realidade, como as sombras na parede da caverna de Platão. Como jamais podemos ter a informação perfeita, as previsões são intrinsecamente falíveis, o que não significa que sejam erradas, somente que serão sempre incompletas. Essa constatação não renega as ideias que o big data oferece, mas os coloca em seu devido lugar – como instrumento incapaz de dar respostas definitivas, apenas respostas boas o suficiente para nos ajudar por enquanto, até que métodos e respostas melhores surjam. Isso também sugere que devemos usar este instrumento com bastante humildade... E humanidade.

Referências

1. ALTER, Alexandra. "Your E-Book Is Reading You". *The Wall Street Journal*. Nova York: Dow Jones & Company, 29 de junho de 2012. Também disponível em http://online.wsj.com/article/SB10001424052702304870304577490950051438304.html..
2. ANDERSON, Benedict. *Imagined Communities*. New Edition Londres: Verso, 2006.
3. ANDERSON, Chris. "The End of Theory". *Wired.* San Francisco: Condé Nast Publications, vol.16, n.7, julho de 2008. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory).
4. ASUR, Sitaram e HUBERMAN, Bernardo A. "Predicting the Future with Social Media". Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, p. 492-499. Também disponível em http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf.
5. BABBIE, Earl. *Practice of Social Research*. 12ª ed. Connecticut: Wadsworth Publishing, 2010.
6. BACKSTROM, Lars, DWORK, Cynthia e KLEINBERG, Jon. "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography". *Communications of the ACM*. Nova York: Association for Computing Machinery, dezembro de 2011 p. 133-141.
7. BAKOS, Yannis e BRYNJOLFSSON, Erik. "Bundling Information Goods: Pricing, Profits, and Efficiency". *Management Science*. Hanover, MD: Institute for Operations Research and the Management Sciences n. 45, pp. 1613–30, dezembro de 1999.
8. BANKO, Michele e BRILL, Eric. "Scaling to Very Very Large Corpora for Natural Language Disambiguation". Microsoft Research, 2001, p. 3. Também disponível em http://acl.ldc.upenn.edu/P/P01/P01-1005.pdf.
9. BARBARO, Michael e ZELLER, JR., Tom. "A Face Is Exposed for AOL Searcher N. 4417749". *The New York Times*. Nova York: The New York Times Company, 9 de agosto de 2006. Também disponível em http://www.nytimes.com/2006/08/09/technology/09aol.html.
10. BARNES, Brooks. "A Year of Disappointment at the Movie Box Office". *The New York Times*. Nova York: The New York Times Company, 25 de dezembro de 2011. Também disponível em http://www.nytimes.com/2011/12/26/business/media/a-year-ofdisappointment-for-hollywood.html.
11. BEATY, Janice. *Seeker of Seaways: A Life of Matthew Fontaine Maury, Pioneer Oceanographer*. Nova York: Pantheon Books, 1966.
12. BERGER, Adam L. *et al*. "The Candide System for Machine Translation". Proceedings of the 1994 ARPA Workshop on Human Language Technology, 1994. Também disponível em http://aclweb.org/anthologynew/H/H94/H94-1100.pdf.
13. BERK, Richard. "The Role of Race in Forecasts of Violent Crime". *Race and Social Problems* Nova York: Springer, n.1, 231-242, 2009.
14. BLACK, Edwin. *IBM e o holocausto*. Rio de Janeiro: Elsevier, 2001.
15. BOYD, Danah e CRAWFORD, Kate. "Six Provocations for Big Data". Trabalho apresentado no simpósio "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society". Oxford Internet Institute. Oxford, 21 de setembro de 2011. Também disponível em http://ssrn.com/abstract=1926431.
16. BROWN, Brad. *et al.* "Are You Ready for the Era of 'Big Data'?" *McKinsey Quarterly*. Nova York: McKinsey&Company n. 4, p. 10, outubro de 2011.
17. BRYNJOLFSSON, Erik; MCAFEE, Andrew; SORELL, Michael e ZHU, Feng. "Scale Without Mass: Business Process Replication and Industry Dynamics". Trabalho apresentado para a Harvard Business

School, setembro de 2006. Também disponível em http://www.hbs.edu/research/pdf/07-016.pdf e em http://hbswk.hbs.edu/item/5532.html.

18. BRYNJOLFSSON, Erik; HITT, Lorin e HEEKYUNG Kim. "Strength in Numbers: How Does Data-Driven Decision making Affect Firm Performance?" ICIS 2011 Proceedings, Paper 13. Também disponível em http://aisel.aisnet.org/icis2011/proceedings/economicvalueIS/13 e em http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486.
19. BYRNE, John. *The Whiz Kids*. Nova York: Doubleday, 1993.
20. Cate, Fred H. "The Failure of Fair Information Practice Principles". In Jane K. Winn, (org.). *Consumer Protection in the Age of the "Information Economy"*. Surrey: Ashgate, 2006, p. 341.
21. CHIN, A. e KLINEFELTER, A. "Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study". 90 North Carolina Law Review 1417 (2012).
22. CROSBY, Alfred. *The Measure of Reality: Quantification and Western Society, 1250-1600*. Cambridge: Cambridge University Press, 1997.
23. CUKIER, Kenneth. "Data, Data Everywhere". *The Economist. Londres*: The Economist Newspaper Ltd, Special Report, p. 1–14 27 de fevereiro de 2010.
24. CUKIER, Kenneth. "Tracking Social Media: The Mood of the Market". *The Economist* on-line. 28 de junho de 2012. Também disponível em http://www.economist.com/blogs/graphicdetail/2012/06/tracking-social-media.
25. DAVENPORT, Thomas H.; BARTH, Paul e BEAN, Randy. "How 'Big Data' Is Different". *Sloan Review.* Massachusetts: MIT, 30 de julho de 2012. Também disponível em http://sloanreview.mit.edu/the-magazine/2012-fall/54104/howbig-data-is-different/.
26. DI QUINZIO, Melanie. e McCARTHY, Anne. "Rabies Risk Among Travellers". *CMAJ*. Ottawa: Canadian Medical Association vol. 178, n. 5, p. 567, 2008.
27. DROGIN, Marc. *Anathema! Medieval Scribes and the History of Book Curses*. Montclair, NJ: Allanheld and Schram, 1983.
28. DUGAS, A.F. *et al*. "Google Flu Trends: Correlation with Emergency Department Influenza Rates and Crowding Metrics". CID Advance Access. 8 de janeiro de 2012. DOI 10.1093/cid/cir883.
29. DUGGAN, Mark e LEVITT, Steven D. "Winning Isn't Everything: Corruption in Sumo Wrestling". *American Economic Review*. Nova York: American Economic Association. N. 92, p. 1594-1605, 2002. Também disponível em http://pricetheory.uchicago.edu/levitt/Papers/DugganLevitt2002.pdf.
30. DUHIGG, Charles. *O poder do hábito*. Rio de Janeiro: Objetiva, 2012.
31. DUHIGG, Charles. "How Companies Learn Your Secrets". *The New York Times*. Nova York: The New York Times Company, 16 de fevereiro de 2012. Também disponível em http://www.nytimes.com/2012/02/19/magazine/shoppinghabits.html.
32. DWORK, Cynthia. "A Firm Foundation for Private Data Analysis". *Communications of the ACM*. Nova York: Association for Computing Machinery, p. 86-95, janeiro de 2011. Também disponível em http://dl.acm.org/citation.cfm?id=1866739.1866758.
33. "ROLLS-ROYCE: Britain's Lonely High-Flier". *The Economist,* January 8, 2009. Também disponível em http://www.economist.com/node/12887368.
34. "BUILDING with Big Data: The Data Revolution Is Changing the Landscape of Business." *The Economist.* Londres: The Economist Newspaper Ltd, 26 de maio de 2011. Também disponível em http://www.economist.com/node/18741392/.
35. "OFFICIAL Statistics: Don't Lie to Me, Argentina". *The Economist*, Londres: The Economist Newspaper Ltd, 25 de fevereiro de 2012. Também disponível em http://www.economist.com/node/21548242.
36. "COUNTING Every Moment". *The Economist.* Londres: The Economist Newspaper Ltd, 3 de março de 2012. Também disponível em http://www.economist.com/node/21548493.
37. "VEHICLE Data Recorders: Watching Your Driving". *The Economist.* Londres: The Economist Newspaper Ltd, 23 de junho de 2012. Também disponível em http://www.economist.com/node/21557309.
38. EDWARDS, Douglas. *I'm Feeling Lucky: The Confessions of Google Employee Number 59*. Boston: Houghton Mifflin Harcourt, 2011.

39. EHRENBERG, Rachel. "Predicting the Next Deadly Manhole Explosion". *WIRED*. San Francisco: Condé Nast Publications, 7 de julho de 2010. Também disponível em http://www.wired.com/wiredscience/2010/07/manholeexplosions.
40. EISENSTEIN, Elizabeth L. *The Printing Revolution in Early Modern Europe*. Cambridge: Canto/Cambridge University Press, 1993.
41. ETZIONI, Oren; KNOBLOCK, C.A.; TUCHINDA, R. e YATES A. "To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price". Proceedings of KDD'03. Nova York: SIGKDD, p. *119-128,* 24 a 27 de agosto de 2003. Também disponível em http://knight.cis.temple.edu/~yates//papers/hamletkdd03.pdf.
42. FREI, Patrizia *et al*. "Use of Mobile Phones and Risk of Brain Tumours: Update of Danish Cohort Study". *BMJ.* Londres: BMJ Group, n. 343, 2011. Também disponível em http://www.bmj.com/content/343/bmj.d6387.
43. FURNAS, Alexander. "Homeland Security's 'Pre-Crime' Screening Will Never Work". The Atlantic On-line, 17 de abril de 2012. Também disponível em http://www.theatlantic.com/technology/archive/2012/04/homeland-securitys-pre-crime-screening-will-never-work/255971/.
44. GARTON ASH, Timothy. *The File*. Londres: Atlantic Books, 2008.
45. GERON, Tomio. "Twitter's Dick Costolo: Twitter Mobile Ad Revenue Beats Desktop on Some Days". *Forbes.* Nova York: Forbes publishing, 6 de junho de 2012. Também disponível em http://www.forbes.com/sites/tomiogeron/2012/06/06/twittersdick-costolo-mobile-ad-revenue-beats-desktop-on-some-days/.
46. GINSBURG, Jeremy, *et al.* "Detecting Influenza Epidemics Using Search Engine Query Data". *Nature*. Londres: Nature Publishing Group, n. 457, p. 1012-14, 2009. Também disponível em http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html.
47. GOLDER, Scott A. e MACY, Michael W. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures". Science. Washington, DC: AAAS, n. 333, p. 1878-81, 30 de setembro de 2011.
48. GOLLE, Philippe. "Revisiting the Uniqueness of Simple Demographics in the US Population". Proceedings of the Association for Computing Machinery Workshop on Privacy in Electronic Society 5, p. 77-80, 2006.
49. GOO, Sara Kehaulani. "Sen. Kennedy Flagged by No-Fly List". *Washington Post.* Washington, DC: 20 de agosto de 2004, p. A01. Também disponível em http://www.washingtonpost.com/wp-dyn/articles/A17073-2004Aug19.html.
50. HAEBERLEN, A. *et al*. "Differential Privacy Under Fire". In: *SEC'11: Proceedings of the 20th USENIX conference on Security.* San Francisco: USENIX, 2011 p. 33. Também disponível em http://www.cis.upenn.edu/~ahae/papers/fuzzsec2011.pdf.
51. HALBERSTAM, David. *The Reckoning*. Nova York: William Morrow, 1986.
52. HALDANE, J. B. S. "On Being the Right Size". *Harper's Magazine*. Nova York: HarperCollins, março de 1926. Também disponível em http://harpers.org/archive/1926/03/on-being-the-right-size/.
53. HALEVY, Alon. e NORVIG, Peter.; PEREIRA, Fernando. "The Unreasonable Effectiveness of Data". *IEEE Intelligent Systems*. Los Alamitos, CA: IEEE Computer Society n. 24, p. 8-12, março/abril de 2009.
54. HARCOURT, Bernard E. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: University of Chicago Press, 2006.
55. HARDY, Quentin. "Bizarre Insights from Big Data". NYTimes.com, 28 de março de 2012. Também disponível em http://bits.blogs.nytimes.com/2012/03/28/bizarre-insightsfrom-big-data/.
56. Hays, Constance L. "What Wal-Mart Knows About Customers' Habits". *The New York Times*. Nova York: The New York Times Company, 14 de novembro de 2004. Também disponível em http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html.
57. HEARN, Chester G. *Tracks in the Sea: Matthew Fontaine Maury and the Mapping of the Oceans*. Nova York: International Marine/McGraw-Hill, 2002.

58. HELLAND, Pat. "If You Have Too Much Data then 'Good Enough' Is Good Enough". *Communications of the ACM*. Nova York: Association for Computing Machinery junho de 2011, p. 40.
59. HILBERT, Martin e LÓPEZ, Priscilla. "The World's Technological Capacity to Store, Communicate, and Compute Information". *Science*. Hanover, MD: Institute for Operations Research and the Management Sciences vol. 1, p. 60–65, abril de 2011.
60. HILBERT, Martin "How to Measure the World's Technological Capacity to Communicate, Store and Compute Information?" *International Journal of Communication*. Washington, DC: USC Annenberg Press, p. 1042-55, 2012. Também disponível em ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742.
61. HOLSON, Laura M. "Putting a Bolder Face on Google". *The New York Times*. Nova York: The New York Times Company, 1 de março de 2009 p. BU 1. Também disponível em http://www.nytimes.com/2009/03/01/business/01marissa.html.
62. HOPKINS, Brian e EVELSON, Boris. "Expand Your Digital Horizon with Big Data". Cambridge, MA: Forrester Research, 30 de setembro de 2011.
63. Hotz, Robert Lee. "The Really Smart Phone". *The Wall Street Journal*. Nova York: Dow Jones & Company, 22 de abril de 2011. Também disponível em http://online.wsj.com/article/SB10001424052748704547604576263261679848814.html.
64. HUTCHINS, John. "The First Public Demonstration of Machine Translation: The Georgetown-IBM System, 7th January 1954". Novembro de 2005. Também disponível em http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf.
65. INGLEHART, R e KLINGEMANN, H. D. *Genes, Culture and Happiness*. Cambridge, MA: MIT Press, 2000.
66. ISAACSON, Walter. *Steve Jobs*. São Paulo: Companhia das Letras, 2011.
67. KAHNEMAN, Daniel. *Rápido e devagar*. Rio de Janeiro: Objetiva, 2012.
68. KAPLAN, Robert S e NORTON, David P. *Mapas estratégicos*. Rio de Janeiro: Elsevier, 2004.
69. KARNITSCHNIG, Matthew.; MANGALINDAN, Mylene. "AOL Fires Technology Chief After Web-Search Data Scandal". *The Wall Street Journal*. Nova York: Dow Jones & Company, 21 de agosto de 2006.
70. KEEFE, Patrick Radden. "Can Network Theory Thwart Terrorists?" *The New York Times*. Nova York: The New York Times Company, 12 de março de 2006. Também disponível em http://www.nytimes.com/2006/03/12/magazine/312wwln_essay.html.
71. KINNARD, Douglas. *The War Managers*. Lebanon, NH: University Press of New England, 1977.
72. KIRWAN, Peter. "This Car Drives Itself". *Wired UK*. Londres: Condé Nast Publications, janeiro de 2012. Também disponível em http://www.wired.co.uk/magazine/archive/2012/01/features/this-car-drives-itself.
73. KLIFF, Sarah. "A Database That Could Revolutionize Health Care". *Washington Post*. Washington, DC: 21 de maio de 2012.
74. KRUSKAL, William e MOSTELLER, Frederick. "Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939". *International Statistical Review*. Haia: International Statistical Institute, v. 48, p. 169-195, 1980.
75. LANEY, Doug. "To Facebook You're Worth $80.95". *The Wall Street Journal*. Nova York: Dow Jones & Company, 3 de maio de 2012. Também disponível em http://blogs.wsj.com/cio/2012/05/03/tofacebook-youre-worth-80-95/.
76. LATOUR, Bruno, *et al. The Pasteurization of France*. Cambridge, MA: Harvard University Press, 1993.
77. LEVITT, Steven D e DUBNER, Stephen J. *Freakonomics*. Rio de Janeiro: Elsevier, 2007.
78. LEVY, Steven. *Google: a biografia*. São Paulo: Universo dos Livros, 2011.
79. LEWIS, Charles Lee. *Matthew Fontaine Maury: The Pathfinder of the Seas*. U.S. Naval Institute, 1927.
80. LOHR, Steve. "Can Apple Find More Hits Without Its Tastemaker?". *The New York Times*. Nova York: The New York Times Company, 19 de janeiro de 2011. p. B1. Também disponível em http://www.nytimes.com/2011/01/19/technology/companies/19innovate.html.

81. LOWREY, Annie. "Economists' Programs Are Beating U.S. at Tracking Inflation." *Washington Post.* Washington, DC: 25 de dezembro de 2010. Também disponível em http://www.washingtonpost.com/wpdyn/content/article/2010/12/25/AR2010122502600.html.
82. MACRAKIS, Kristie. *Seduced by Secrets: Inside the Stasi's Spy-Tech World*. Cambridge: Cambridge University Press, 2008.
83. MANYIKA, James *et al*. "Big Data: The Next Frontier for Innovation, Competition, and Productivity". Nova York: McKinsey Global Institute, maio de 2011. Também disponível em http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.
84. MARCUS, James. *Amazonia: Five Years at the Epicenter of the Dot.Com Juggernaut*. Nova York: The New Press, 2004.
85. MARGOLIS, Joel M. "When Smart Grids Grow Smart Enough to Solve Crimes". Neustar, 18 de março de 2010. Também disponível em http://energy.gov/sites/prod/files/gcprod/documents/Neustar_Comments_DataExhibitA.pdf.
86. MAURY, Matthew Fontaine. *The Physical Geography of the Sea*. Nova York: Harper, 1855.
87. MAYER-SCHÖNBERGER, Viktor. "Beyond Privacy, Beyond Rights: Towards a "Systems" Theory of Information Governance". 98 California Law Review 1853, 2010.
88. MAYER-SCHÖNBERGER, Viktor. *Delete: The Virtue of Forgetting in the Digital Age*. 2ª ed. Princeton, NJ: Princeton University Press, 2011.
89. McGREGOR, Carolyn; CATLEY, Christina, JAMES, Catley; JAMES, Andrew e PADBURY, James. "Next Generation Neonatal Health Informatics with Artemis". In: *User Centred Networked Health Care.* A. Moen *et al*. (orgs.). European Federation for Medical Informatics. Lansdale, PA: IOS Press, 2011, p. 117 *et seq*.
90. MCNAMARA, Robert S e VANDEMARK, Brian. *In Retrospect: The Tragedy and Lessons of Vietnam.* Nova York: Random House, 1995.
91. MEHTA, Abhishek. "Big Data: Powering the Next Industrial Revolution". Tableau Software White Paper, 2011.
92. MICHEL, Jean-Baptiste *et al*. "Quantitative Analysis of Culture Using Millions of Digitized Books". *Science.* Hanover, MD: Institute for Operations Research and the Management Sciences, n. 331, p. 176-182, 14 de janeiro de 2011. Também disponível em http://www.sciencemag.org/content/331/6014/176.abstract.
93. MILLER, Claire Cain. "U.S. Clears Google Acquisition of Travel Software". *The New York Times.* Nova York: The New York Times Company, 8 de abril de 2011. Também disponível em http://www.nytimes.com/2011/04/09/technology/09google.html?_r=0.
94. MILLS, Howard. "Analytics: Turning Data into Dollars". *Forward Focus*. Nova Zelândia: Deloitte, dezembro de 2011 Também disponível em http://www.deloitte.com/assets/DcomUnitedStates/Local%20Assets/Documents/FSI/US_FSI_Forward%20Focus_Analytics_Turning%20data%20into%20dollars_120711.pdf.
95. MINDELL, David A. *Digital Apollo: Human and Machine in Spaceflight*. Cambridge, MA: MIT Press, 2008.
96. MINKEL, J.R. "The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II". *Scientific American.* Nova York, Nature Publishing Group, 30 de março de 2007. Também disponível em http://www.scientificamerican.com/article.cfm?id=confirmedthe-us-census-b.
97. MURRAY, Alexander. *Reason and Society in the Middle Ages*. Oxford: Oxford University Press, 1978.
98. NALIMOV, E.V., HAWORTH, G. e HEINZ, E.A. "Space-Efficient Indexing of Chess Endgame Tables". *ICGA Journal.* Maastricht: International Computer Games Association, vol. 23, n. 3, p. 148-162, 2000.
99. NARAYANAN, Arvind e SHMATIKOV, Vitaly. "How to Break the Anonymity of the Netflix Prize Dataset"- 18 de outubro de 2006. Também disponível em http://arxiv.org/abs/cs/0610105.

100. NARAYANAN, Arvind. “Robust De-Anonymization of Large Sparse Datasets”. Proceedings of the 2008 IEEE Symposium on Security and Privacy, p. 111. Também disponível em http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf.
101. NAZARETH, Rita e LEITE, Julia. “Stock Trading in U.S. Falls to Lowest Level Since 2008”. Bloomberg, 13 de agosto de 2012. Também disponível em http://www.bloomberg.com/news/2012-08-13/stock-trading-in-us-hits-lowest-level-since-2008-as-vix-falls.html.
102. NEGROPONTE, Nicholas. *A vida digital*. São Paulo: Companhia das Letras, 1995.
103. NEYMAN, Jerzy. “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection”. *Journal of the Royal Statistical Society.* Londres: Wiley-Blackwell, vol. 97, n. 4, p. 558-625, 1934.
104. OHM, Paul. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization”. 57 *UCLA Law Review* 1701 (2010).
105. ONNELA, J.P. *et al*. “Structure and Tie Strengths in Mobile Communication Networks”. Proceedings of the National Academy of Sciences of the United States of America (PNAS) 104, maio de 2007 2007, pp. 7332-36. Também disponível em http://nd.edu/~dddas/Papers/PNAS0610245104v1.pdf.
106. PALFREY, John e GASSER, Urs. *Interop: The Promise and Perils of Highly Interconnected Systems*. Nova York: Basic Books, 2012.
107. PEARL, Judea. *Models, Reasoning and Inference*. Cambridge: Cambridge University Press, 2009.
108. POLLACK, Andrew. “DNA Sequencing Caught in the Data Deluge” *The New York Times.* Nova York: The New York Times Company, 30 de novembro de 2011. Também disponível em http://www.nytimes.com/2011/12/01/business/dna-sequencingcaught-in-deluge-of-data.html?pagewanted=all.
109. “REPORT to the President and Congress Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology”. Conselho Presidencial de Assessores para Ciência e Tecnologia, dezembro de 2010. Também disponível em http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf.
110. PRIEST, Dana e ARKIN, William. “A Hidden World, Growing Beyond Control”. *Washington Post.* Washington, DC: 19 de julho de 2010. Também disponível em http://projects.washingtonpost.com/top-secretamerica/articles/a-hidden-world-growing-beyondcontrol/print/.
111. QUERY, Tim. “Grade Inflation and the Good-Student Discount”. *Contingencies Magazine*. Washington, DC: American Academy of Actuaries, maio/junho de 2007. Também disponível em http://www.contingencies.org/mayjun07/tradecraft.pdf.
112. QUINN, Elias Leake. “Smart Metering and Privacy: Existing Law and Competing Policies; A Report for the Colorado Public Utility Commission”. Primavera de 2009. Também disponível em http://www.w4ar.com/Danger_of_Smart_Meters_Colorado_Report.pdf.
113. RESHEF, David, *et al*. “Detecting Novel Associations in Large Data Sets”. *Science.* Hanover, MD: Institute for Operations Research and the Management Sciences, p. 1518-24, 2011.
114. ROSENTHAL, Jonathan. “Banking Special Report”. *The Economist.* Londres: The Economist Newspaper Ltd, p. 7-8, 19 de maio de 2012.
115. ROSENZWEIG, Phil. “Robert S. McNamara and the Evolution of Modern Management”. *Harvard Business Review.* Allston, MA: Harvard Business Publishing, p. 87-93, dezembro de 2010. Também disponível em http://hbr.org/2010/12/robert-s-mcnamara-and-theevolution-of-modern-management/ar/pr.
116. RUDIN, Cynthia *et al*. “21st-Century Data Miners Meet 19th-Century Electrical Cables”. *Computer,* Los Alamitos, CA: IEEE Computer Society, p. 103-105, junho de 2011.
117. RUDIN, Cynthia. “Machine Learning for the New York City Power Grid”. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Los Alamitos, CA: IEEE Computer Society, n. 34.2, p. 328-345, 2012. Também disponível em http://hdl.handle.net/1721.1/68634.
118. RYS, Michael. “Scalable SQL”. *Communications of the ACM.* Nova York: Association for Computing Machinery, n. 48, p. 48-53, junho de 2011.

119. SALATHÉ, Marcel e KHANDELWAL, Shashank. "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control". *PloS Computational Biology.* Nova York: Public Library of Science, vol. 7, n. 10, outubro de 2011.
120. SAVAGE, Mike e BURROWS, Roger. "The Coming Crisis of Empirical Sociology". *Sociology.* Thousand Oaks, CA: Sage Publications, n. 41 p. 885-899, janeiro de 2007.
121. SCHLIE, Erik *et al. Simply Seven: Seven Ways to Create a Sustainable Internet Business*. Londres: Palgrave Macmillan, 2011.
122. SCANLON, Jessie. "Luis von Ahn: The Pioneer of 'Human Computation'. *Businessweek.* Nova York: McGraw-Hill, 3 de novembro de 2008. Também disponível em http://www.businessweek.com/stories/2008-11-03/luis-von-ahnthe-pioneer-of-human-computation-businessweek-business-news-stock-market-and-financial-advice.
123. SCISM, Leslie e MAREMONT, Mark. "Inside Deloitte's Life-Insurance Assessment Technology". *The Wall Street Journal,* November 19, 2010. Também disponível em http://online.wsj.com/article/SB10001424052748704104104575622531084755588.html.
124. SCISM, Leslie. "Insurers Test Data Profiles to Identify Risky Clients." *The Wall Street Journal.* Nova York: Dow Jones & Company 19 de novembro de 2010. Também disponível em http://online.wsj.com/article/SB10001424052748704648604575620750998072986.html.
125. SCOTT, James. *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, CT: Yale University Press, 1998.
126. SELTZER, William e ANDERSON, Margo. "The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses". *Social Research.* Nova York: The New School for Social Research, n. 68 p. 481-513, 2001.
127. SILVER, Nate. *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't*. Nova York: Penguin, 2012.
128. SINGEL, Ryan. "Netflix Spilled Your *Brokeback Mountain* Secret, Lawsuit Claims". *Wired*. San Francisco: Condé Nast Publications, 17 de dezembro de 2009. Também disponível em http://www.wired.com/threatlevel/2009/12/netflix-privacylawsuit/.
129. SMITH, Adam. *The Wealth of Nations* (1776). Reimpressão. Nova York: Bantam Classics, 2003. Versão eletrônica gratuita disponível em http://www2.hn.psu.edu/faculty/jmanis/adam-smith/Wealth-Nations.pdf.
130. SOLOVE, Daniel J. *The Digital Person: Technology and Privacy in the Information Age*. Nova York: NYU Press, 2004.
131. SUROWIECKI, James. "A Billion Prices Now". *The New Yorker*. San Francisco: Condé Nast Publications, 30 de maio de 2011. Também disponível em http://www.newyorker.com/talk/financial/2011/05/30/110530ta_talk_surowiecki.
132. TALEB, Nassim Nicholas. *Iludido pelo acaso: a influência oculta da sorte nos mercados e na vida*. Rio de Janeiro: Record, 2010.
133. TALEB, Nassim Nicholas. *A lógica do cisne negro*. São Paulo: Best Seller, 2008.
134. THOMPSON, Clive. "For Certain Tasks, the Cortex Still Beats the CPU". *Wired.* San Francisco: Condé Nast Publications, 25 de junho de 2007. Também disponível em http://www.wired.com/techbiz/it/magazine/15-07/ff_humancomp?currentPage=all.
135. THURM, Scott. "Next Frontier in Credit Scores: Predicting Personal Behavior". *The Wall Street Journal*. Nova York: Dow Jones & Company, 27 de outubro de 2011. Também disponível em http://online.wsj.com/article/SB10001424052970203687504576655182086300912.html.
136. TSOTSIS, Alexia. "Twitter Is at 250 Million Tweets per Day, iOS 5 Integration Made Signups Increase 3x". *TechCrunch* (periódico on-line), 17 de outubro de 2011. Também disponível em http://techcrunch.com/2011/10/17/twitter-is-at-250-million-tweets-per-day/.
137. VALERY, Nick. "Tech.View: Cars and Software Bugs". *The Economist*. Londres: The Economist Newspaper Ltd, 16 de maio de 2010. Também disponível em http://www.economist.com/blogs/babbage/2010/05/techview_cars_and_software_bugs.

138. VLAHOS, James. “The Department Of Pre-Crime.” *Scientific American.* Nova York, Nature Publishing Group, n. 306, p. 62-67, janeiro de 2012.
139. VON BAEYER, Hans Christian. *Information: The New Language of Science*. Cambridge, MA: Harvard University Press, 2004.
140. VON AHN, Luis *et al*. “reCAPTCHA: Human-Based Character Recognition via Web Security Measures”. *Science.* Hanover, MD: Institute for Operations Research and the Management Sciences, n. 321, p. 1465-68, 12 de setembro de 2008. Também disponível em http://www.sciencemag.org/content/321/5895/1465.abstract.
141. WATTS, Duncan. *Tudo é óbvio desde que você saiba a resposta*. São Paulo: Paz e Terra, 2011.
142. WEINBERGER, David. *A nova desordem digital*. Rio de Janeiro: Elsevier, 2007.
143. WEINBERGER, Sharon. “Intent to Deceive”. *Nature.* Londres: Nature Publishing Group, n. 465, p. 412-415, maio de 2010. (http://www.nature.com/news/2010/100526/full/465412a.html).
144. WEINBERGER, Sharon. “Terrorist ‘Pre-crime’ Detector Field Tested in United States”. *Nature.* Londres: Nature Publishing Group*,* 27 de maio de 2011. Também disponível em http://www.nature.com/news/2011/110527/full/news.2011.323.html.
145. WHITEHOUSE, David. “UK Science Shows Cave Art Developed Early”. BBC News On-line, 3 de outubro de 2001. Também disponível em http://news.bbc.co.uk/1/hi/sci/tech/1577421.stm.
146. WIGNER, Eugene. “The Unreasonable Effectiveness of Mathematics in the Natural Sciences”. *Communications on Pure and Applied Mathematics.* Nova York: Courant Institute of Mathematical Sciences, vol. 13, n. 1, p. 1-14, 1960.
147. WILKS, Yorick. *Machine Translation: Its Scope and Limits*. Nova York: Springer, 2009.
148. WINGFIELD, Nick. “Virtual Products, Real Profits: Players Spend on Zynga's Games, but Quality Turns Some Off”. *The Wall Street Journal*. Nova York: Dow Jones & Company, 9 de setembro de 2011. Também disponível em http://online.wsj.com/article/SB10001424053111904823804576502442835413446.html.

Notas

CAPÍTULO 1

Trabalho sobre a tendência da gripe — Jeremy Ginsburg *et al*. "Detecting Influenza Epidemics Using Search Engine Query Data". *Nature* 457 (2009), pp. 1012–14. Também disponível em http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html.

Estudo secundário da Google Flu Trends por pesquisadores de Johns Hopkins — A.F. Dugas *et al*., "Google Flu Trends: Correlation with Emergency Department Influenza Rates and Crowding Metrics". CID Advanced Access (8 de janeiro de 2012); Também disponível em DOI 10.1093/cid/cir883.

Comprar passagens aéreas, Farecast — A informação vem de Kenneth Cukier, "Data, Data Everywhere". *The Economist,* reportagem especial, 27 de fevereiro de 2010, pp. 1–14, e de entrevistas com Etzioni entre 2010 e 2012.

Projeto Hamlet de Etzioni — Oren Etzioni, C.A. Knoblock, R. Tuchinda e A. Yates, "To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price", SIGKDD '03, 24–27 de agosto de 2003. Também disponível em http://knight.cis.temple.edu/~yates//papers/hamletkdd03.pdf.

Preço que a Microsoft pagou para a Farecast — Reportagens, principalmente "Secret Farecast Buyer Is Microsoft", Seattlepi.com, 17 de abril de 2008. Também disponível em http://blog.seattlepi.com/venture/2008/04/17/secret-farecastbuyer-is-microsoft/?source=mypi.

Uma maneira de pensar sobre o big data – Há um produtivo debate sobre a origem do termo "big data" e como defini-lo com perfeição. As duas palavras parecem ter aparecido em uníssimo há décadas. Uma pesquisa de 2001, de Doug Laney, da Gartner, estabeleceu os "três Vs" dos grandes dados (volume, velocidade e variedade), útil em seu tempo, mas imperfeito.

Astronomia e sequenciamento do DNA — Cukier, "Data, Data Everywhere".

Sequenciamento do DNA — Andrew Pollack, "DNA Sequencing Caught in the Data Deluge", *The New York Times,* 30 de novembro de 2011. Também disponível em http://www.nytimes.com/2011/12/01/business/dna-sequencingcaught-in-deluge-of-data.html?pagewanted=all.

Bilhões de ações comercializadas — Rita Nazareth e Julia Leite, "Stock Trading in U.S. Falls to Lowest Level Since 2008". Bloomberg, 13 de agosto de 2012. Também disponível em http://www.bloomberg.com/news/2012-08-13/stock-trading-in-u-s-hits-lowest-level-since-2008-as-vixfalls.html.

Vinte e quatro petabytes da Google — Thomas H. Davenport, Paul Barth e Randy Bean, "How 'Big Data' Is Different". *Sloan Review,* 30 de julho de 2012, pp. 43–46. Também disponível em http://sloanreview.mit.edu/the-magazine/2012-fall/54104/how-big-data-is-different/.

Estatísticas do Facebook — Prospect de IPO do Facebook, Facebook, Form S-1 Registration Statement, U.S. Securities And Exchange Commission, 1° de fe-

vereiro de 2012. Também disponível em http://sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm.

Estatísticas do YouTube — Larry Page, "Update from the CEO". Google, abril de 2012. Também disponível em http://investor.google.com/corporate/2012/ceoletter.html.

Quantidade de tweets — Tomio Geron, "Twitter's Dick Costolo: Twitter Mobile Ad Revenue Beats Desktop on Some Days". Forbes, 6 de junho de 2012. Também disponível em http://www.forbes.com/sites/tomiogeron/2012/06/06/twitters-dickcostolo-mobile-ad-revenue-beats-desktop-on-some-days/.

Informação sobre a quantidade de dados — Martin Hilbert e Priscilla López. "The World's Technological Capacity to Store, Communicate, and Compute Information". *Science,* 1° de abril de 2011, pp. 60–65; Martin Hilbert e Priscilla López, "How to Measure the World's Technological Capacity to Communicate, Store and Compute Information?" *International Journal of Communication* 2012, pp. 1042–55. Também disponível em http://www.ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742.

Estimativa da quantidade de informação armazenada em 2013 — Cukier entrevista Hilbert.

A imprensa e oito milhões de livros; maior produção desde a fundação de Constantinopla — Elizabeth L. Eisenstein, *The Printing Revolution in Early Modern Europe.* (Cambridge: Canto/Cambridge University Press, 1993), pp. 13–14.

A analogia de Peter Norvig — Palestras de Norvig relacionadas com um trabalho do qual ele foi coautor. O trabalho é A. Halevy, P. Norvig e F. Pereira. "The Unreasonable Effectiveness of Data". *IEEE Intelligent Systems,* Março/abril 2009, pp. 8–12. (Note que o título faz uma brincadeira com o título do famoso artigo de Eugene Wigner, "The Unreasonable Effectiveness of Mathematics in the Natural Sciences", no qual ele explica por que a física pode ser bem expressa com matemática básica, enquanto as ciências humanas resistem às fórmulas elegantes. Leia E. Wigner. "The Unreasonable Effectiveness of Mathematics in the Natural Sciences". *Communications on Pure and Applied Mathematics* 13, n. 1 (1960), pp. 1–14). Entre as palestras de Norvig sobre o trabalho está "Peter Norvig—The Unreasonable Effectiveness of Data", aula na University of British Columbia Britânica, vídeo do YouTube, 23 de setembro de 2010. Também disponível em http://www.youtube.com/watch?v=yvDCzhbjYWs.

Sobre o tamanho físico que afeta a lei da física – Apesar de não totalmente correto, a referência citada com frequência é J.B.S. Haldane. "On Being the Right Size". *Harper's Magazine,* março de 1926. Também disponível em http://harpers.org/archive/1926/03/on-being-theright-size/.

Picasso: sobre as imagens de Lascaux — David Whitehouse, "UK Science Shows Cave Art Developed Early". BBC News On-line, 3 de outubro de 2001. Também disponível em http://news.bbc.co.uk/1/hi/sci/tech/1577421.stm.

CAPÍTULO 2

Sobre Jeff Jonas e como o big data se expressa — Conversa com Jeff Jonas, dezembro de 2010, Paris.

História do censo americano — U.S. Census Bureau, "The Hollerith Machine" Online history. Também disponível em http://www.census.gov/history/www/innovations/technology/the_hollerith_tabulator.html.

A contribuição de Neyman — William Kruskal e Frederick Mosteller. "Representative Sampling, IV: The History of the Concept in Statistics, 1895–1939". *International Statistical Review* 48 (1980), pp. 169–195, pp. 187–188. O famoso trabalho de Neyman é Jerzy Neyman. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection". *Journal of the Royal Statistical Society* 97, no. 4 (1934), pp. 558–625.

Uma amostragem de 1.100 observações basta — Earl Babbie, *Practice of Social Research* (12th ed. 2010), pp. 204–207.

O efeito do telefone celular — "Estimating the Cellphone Effect". 20 de setembro de 2008. Também disponível em http://www.fivethirtyeight.com/2008/09/estimating-cellphoneeffect-22-points.html; para mais informações sobre pesquisas tendenciosas e ideias estatísticas, ver Nate Silver. *The Signal and the Noise: Why So Many Predictions Fail But Some Don't.* (Penguin, 2012).

O sequenciamento genético de Steve Jobs — Walter Isaacson. *Steve Jobs* (2011), pp. 550–551.

Previsão do Google Flu Trends em nível municipal — Dugas *et al*. "Google Flu Trends".

Etzioni sobre dados temporais — Entrevista com Cukier, outubro de 2011.

Executivo-chefe do Xoom — Jonathan Rosenthal. "Special Report: International Banking". *The Economist,* 19 de maio de 2012, pp. 7–8.

Lutas arranjadas de sumô — Mark Duggan e Steven D. Levitt. "Winning Isn't Everything: Corruption in Sumo Wrestling". *American Economic Review* 92 (2002), pp. 1594–1605. Também disponível em http://pricetheory.uchicago.edu/levitt/Papers/DugganLevitt2002.pdf

11 milhões de raios de luz de Lytro — Do site de Lytro: http://www.lytro.com.

Substituindo amostras em ciências sociais — Mike Savage e Roger Burrows. "The Coming Crisis of Empirical Sociology". *Sociology* 41 (2007), pp. 885–899.

Sobre a análise de dados de uma operadora de telefonia móvel — J.P. Onnela *et al*. "Structure and Tie Strengths in Mobile Communication Networks". *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 104 (maio 2007), pp. 7332–36. Também disponível em http://nd.edu/~dddas/Papers/PNAS0610245104v1.pdf.

CAPÍTULO 3

Crosby — Alfred W. Crosby. The Measure of Reality: Quantification and Western Society, 1997.

Sobre as citações de Kelvin e Bacon — Esses aforismas são atribuídos a dois homens, apesar de a expressão em seus trabalhos escritos ser ligeiramente diferente. No trabalho de Kelvin, faz parte de uma citação mais ampla sobre medição, tirada da palestra "Electrical Units of Measurement" (1883). No caso de Bacon, é considerada uma tradução livre do latim em Meditationes Sacrae (1597).

Muitas maneiras de se referir à IBM — DJ Patil. "Data Jujitsu: The Art of Turning Data into Product". O'Reilly Media, julho de 2012. Também disponível em http://oreillynet.com/oreilly/data/radarreports/datajujitsu.csp?cmp=tw-strata-books-data-products.

NYSE: 30 mil transações por segundo — Colin Clark. "Improving Speed and Transparency of Market Data". Post do blog NYSE EURONEXT, 9 de janeiro de 2011. Também disponível em http://exchanges.nyx.com/cclark/improving-speedand-transparency-market-data.

Ideia de que "21 + 121 = 13.9" — Brian Hopkins e Boris Evelson. "Expand Your Digital Horizon with Big Data". Forrester, 30 de setembro de 2011.

Melhoria dos algoritmos — Conselho Presidencial de Assessores para Ciência e Tecnologia. "Report to the President and Congress, Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology". Dezembro de 2010, p. 71. Também disponível em http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf.

Mesas finais de xadrez — As mesas finais mais abrangentes disponíveis, o movimento Nalimov (batizado em homenagem a seu criador), abrange todos os jogos com seis ou menos peças. Seu tamanho excede sete terabytes, e a compressão da informação é o maior desafio. Veja E.V. Nalimov, G. McC. Haworth e E.A. Heinz. "Space-efficient Indexing of Chess Endgame Tables". ICGA Journal 23, n. 3 (2000), pp. 148–162.

Desempenho do algoritmo — Michele Banko e Eric Brill. "Scaling to Very Very Large Corpora for Natural Language Disambiguation". Microsoft Research, 2001, p. 3. Também disponível em http://acl.ldc.upenn.edu/P/P01/P01-1005.pdf.

Banco de 95 bilhões de sentenças da Google — Alex Franz e Thorsten Brants. "All Our N-gram are Belong to You". Post no blog da Google, 3 de agosto de 2006. Também disponível em http://googleresearch.blogspot.co.uk/2006/08/all-our-n-gram-arebelong-to-you.html.

Compilação e uso de CPI — Annie Lowrey. "Economists' Programs Are Beating U.S. at Tracking Inflation". Washington Post, 25 de dezembro de 2010. Também disponível em http://www.washingtonpost.com/wpdyn/content/article/2010/12/25/AR2010122502600.html.

IBM demo, palavras e citações — IBM. "701 Translator". Press release, arquivos da IBM, 8 de janeiro de 1954. (http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html). Veja também John Hutchins. "The First Public Demonstration of Machine Translation: The Georgetown-IBM System, 7 de janeiro de 1954". Novembro de 2005. Também disponível em http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf.

IBM Candide — Adam L. Berger et al. "The Candide System for Machine Translation". Proceedings of the 1994 ARPA Workshop on Human Language Technology, 1994. Também disponível em http://aclweb.org/anthologynew/H/H94/H94-1100.pdf.

História da máquina de tradução — Yorick Wilks. Machine Translation: Its Scope and Limits, 2008, p. 107.

Milhões de textos do Candide contra os bilhões de textos do Google — Och entrevista Cukier, dezembro de 2009.

64 idiomas — "Inside Google Translate". Google: http://translate.google.com/about

Banco de dados Brown e um trilhão de palavras da Google — Halevy, Norvig e Pereira. "The Unreasonable Effectiveness of Data": http://www.computer.org/portal/cms_docs_intelligent/intelligent/homepage/2009/x2exp.pdf.

Citação do trabalho de Norvig — ibid.

Corrosão dos canos da BP e ambiente sem fio hostil — Jaclyn Clarabut. "Operations Making Sense of Corrosion". BP Magazine, issue 2, 2011. Também disponível em http://www.bp.com/liveassets/bp_internet/globalbp/globalbp_uk_english/reports_and_publications/bp_magazine/STAGING/local_assets/pdf/BP_Magazine_2011_issue2_text.pdf. A dificuldade de leitura de dados wireless — de Cukier, "Data, Data, Everywhere." Claro que o sistema não é infalível: um incêndio na refinaria Cherry Point da BP em fevereiro de 2012 foi culpa de um cano corroído.

Billion Prices Project — James Surowiecki. "A Billion Prices Now". The New Yorker, 30 de maio de 2011; dados e detalhes podem ser encontrados no site do projeto: http://bpp.mit.edu/; Annie Lowrey. "Economists' Programs Are Beating U.S. at Tracking Inflation". Washington Post, 25 de dezembro de 2010. Também disponível em http://www.washingtonpost.com/wpdyn/content/article/2010/12/25/AR2010122502600.html.

Sobre o PriceStats e uma verificação nas estatísticas nacionais — "Official Statistics: Don't Lie to Me, Argentina". The Economist, 25 de fevereiro de 2012. Também disponível em http://www.economist.com/node/21548242.

Quantidade de fotos no Flickr — Informação do site Flickr: http://www.flick.com.

Sobre o desafio de categorizar informações — Veja David Weinberger. A nova desordem digital. Elsevier, 2007.

Pat Helland — Pat Helland. "If You Have Too Much Data Then 'Good Enough' Is Good Enough". Comunicações do ACM, junho de 2011, pp. 40, 41. Há um vigoroso debate dentro da comunidade de dados sobre os modelos e conceitos mais aptos a satisfazer as necessidades do big data. Helland representa o time que defende um rompimento total com os instrumentos do passado. Michael Rys, da Microsoft, em "Scalable SQL". Comunicações do ACM, junho de 2011, p. 48, argumenta que versões adaptadas das ferramentas atuais funcionam bem.

Visa usando Hadoop — Cukier, "Data, data everywhere".

Apenas 5% dos bancos de dados bem estruturados — Abhishek Mehta. "Big Data: Powering the Next Industrial Revolution". Tableau Software White Paper, 2011.

CAPÍTULO 4

História de Linden assim como "Amazon voice" — Entrevista de Linden com Cukier, Março de 2012.

WSJ sobre os críticos da Amazon — Citado em James Marcus. *Amazonia: Five Years at the Epicenter of the Dot.Com Juggernaut.* New Press, junho de 2004, p. 128.

Citação de Marcus — Marcus. *Amazonia,* p. 199.

Recomendações são um terço da receita da Amazon — O número nunca foi confirmado pela empresa, mas foi publicado em vários relatórios analíticos e artigos na mídia, incluindo "Building with Big Data: The Data Revolution Is Changing the

Landscape of Business". *The Economist,* 26 de maio de 2011. (http://www.economist.com/node/18741392/). O número também foi citado por dois ex-executivos da Amazon em entrevistas com Cukier.

Informação de preços do Netflix — Xavier Amatriain e Justin Basilico. "Netflix Recommendations: Beyond the 5 stars (Part 1)". Blog do Netflix, 6 de abril de 2012.

"Enganado pela aleatoriedade" — Nassim Nicholas Taleb. *Iludido pelo acaso: a influência oculta da sorte nos mercados e na vida.* Record: 2010. Para mais, leia Nassim Nicholas Taleb. *A lógica do cisne negro*. Best Seller, 2008.

Walmart e Pop-Tarts — Constance L. Hays. "What Wal-Mart Knows About Customers' Habits". *The New York Times,* 14 de novembro de 2004. Também disponível em http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html.

Exemplos de modelos de previsão da FICO, Experian e Equifax — Scott Thurm. "Next Frontier in Credit Scores: Predicting Personal Behavior". *The Wall Street Journal,* 27 de outubro de 2011. Também disponível em http://online.wsj.com/article/SB10001424052970203687504576655182086300912.html.

Modelos de previsão Aviva — Leslie Scism e Mark Maremont. "Insurers Test Data Profiles to Identify Risky Clients". *The Wall Street Journal,* 19 de novembro de 2010. Também disponível em http://online.wsj.com/article/SB10001424052748704648604575620750998072986.html. Veja também Leslie Scism e Mark Maremont. "Inside Deloitte's Life-Insurance Assessment Technology". *The Wall Street Journal,* 19 de novembro de 2010. Também disponível em http://online.wsj.com/article/SB10001424052748704104104575622531084755588.html. Veja ainda Howard Mills. "Analytics: Turning Data into Dollars". *Forward Focus,* Dezembro de 2011. Também disponível em http://www.deloitte.com/assets/Dcom-UnitedStates/Local%20Assets/Documents/FSI/US_FSI_Forward%20Focus_Analytics_Turning%20data%20into%20dollars_120711.pdf.

Exemplo da Target e da adolescente grávida — Charles Duhigg. "How Companies Learn Your Secrets". *The New York Times,* 16 de fevereiro de 2012. Também disponível em http://www.nytimes.com/2012/02/19/magazine/shoppinghabits.html. O artigo é adaptado do livro de Duhigg. *O poder do hábito.* Objetiva, 2012; a Target afirma que há imprecisões nos relatos da mídia, mas se recusa a dizer quais são. Quando perguntado da questão neste livro, o porta-voz da Target respondeu: "O objetivo é usar dados para melhorar o relacionamento entre os consumidores e a Target. Nossos clientes querem receber mais valor, ofertas relevantes e uma experiência superior. Como muitas empresas, usamos ferramentas de pesquisa que nos ajudam a entender as tendências de compras e preferências dos consumidores, de modo que possamos lhes oferecer promoções relevantes. Levamos a responsabilidade de proteger a confiança dos consumidores muito a sério. Uma maneira de fazer isso é ter uma abrangente diretriz, que discutimos abertamente na Target.com, e sempre ensinar a equipe a assegurar as informações dos consumidores."

Análise da UPS — Cukier entrevista Jack Levis, março, abril e junho de 2012.

Prematuros — Com base em entrevistas com McGregor em janeiro de 2010 e abril e julho de 2012. Veja também Carolyn McGregor, Christina Catley, Andrew James e James Padbury. "Next Generation Neonatal Health Informatics with Artemis".

In: European Federation for Medical Informatics, *User Centred Networked Health Care,* A. Moen *et al*. (orgs.) IOS Press, 2011, p. 117. Alguns materiais vêm de Cukier. "Data, Data, Everywhere".

Sobre a correlação entre felicidade e renda — R. Inglehart e H.-D. Klingemann, *Genes, Culture and Happiness.* MIT Press, 2000.

Sobre sarampo e gastos com saúde e sobre novas ferramentas não lineares de análises correlacionais — David Reshef *et al*. "Detecting Novel Associations in Large Data Sets". *Science* 334 (2011), pp. 1518–24.

Kahneman — Daniel Kahneman, *Thinking, Fast and Slow* (2011), pp. 74–75.

Pasteur — Para leitores interessados na influência de Pasteur sobre como percebemos os fatos, sugerimos Bruno Latour *et al*. *The Pasteurization of France.* 1993.

Risco de pegar raiva — Melanie Di Quinzio e Anne McCarthy. "Rabies Risk Among Travellers". *CMAJ* 178, n. 5 (2008), p. 567.

A causalidade raramente pode ser provada — A cientista ganhadora do prêmio Turing, Judea Pearl, desenvolveu uma maneira de representar formalmente a dinâmica causal; apesar de não ser uma prova formal, é uma abordagem pragmática para analisar possíveis conexões causais; veja Judea Pearl. *Models, Reasoning and Inference.* 2009.

Exemplo do carro alaranjado — Quentin Hardy. "Bizarre Insights from Big Data". nytimes.com, 28 de março de 2012. Também disponível em http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-frombig-data/; e Kaggle. "Momchil Georgiev Shares His Chromatic Insight from Don't Get Kicked". Post de blog, 2 de fevereiro de 2012. Também disponível em http://blog.kaggle.com/2012/02/02/momchil-georgiev-shares-hischromatic-insight-from-dont-get-kicked/.

Peso dos bueiros, números de explosões e altura da explosão — Rachel Ehrenberg. "Predicting the Next Deadly Manhole Explosion". *WIRED,* 7 de julho de 2010. Também disponível em http://www.wired.com/wiredscience/2010/07/manhole-explosions.

Con Edison no trabalho com estatísticos da Columbia University — Este caso é descrito numa palestra em Cynthia Rudin *et al*. "21st Century Data Miners Meet 19th -Century Electrical Cables". *Computer,* junho de 2011, pp. 103–105. Descrições técnicas do trabalho estão disponíveis nos artigos acadêmicos de Rudin e colaboradores em seus websites, principalmente Cynthia Rudin *et al*. "Machine Learning for the New York City Power Grid". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, n. 2 (2012), pp. 328–345. Também disponível em http://hdl.handle.net/1721.1/68634.

Confusão do termo "caixa de serviço" — Esta lista vem de Rudin *et al*. "21st -Century Data Miners Meet 19th-Century Electrical Cables".

Citação de Rudin — De uma entrevista com Cukier, março de 2012.

Visões de Anderson — Chris Anderson. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *Wired,* junho de 2008. Também disponível em http://www.wired.com/science/discoveries/magazine/16-07/pb_theory/.

O recuo de Anderson — National Public Radio. "Search and Destroy". 18 de julho de 2008. Também disponível em http://www.onthemedia.org/2008/jul/18/search-anddestroy/transcript/.

O que influencia nossa análise - Danah Boyd e Kate Crawford. "Six Provocations for Big Data". Artigo apresentado no Oxford Internet Institute's. "A Decade in Internet

Time: Symposium on the Dynamics of the Internet and Society", 21 de setembro de 2011. Também disponível em http://ssrn.com/abstract=1926431.

CAPÍTULO 5

Detalhes da vida de Maury foram compilados de várias obras dele próprio e sobre ele - Chester G. Hearn. Tracks in the Sea: Matthew Fontaine Maury and the Mapping of the Oceans. International Marine/McGraw-Hill, junho de 2002. Também disponível em http://books.google.co.uk/books?id(XasPAQAAIAAJ; Janice Beaty. Seeker of Seaways: A Life of Matthew Fontaine Maury, Pioneer Oceanographer. Pantheon Books, 1966. Também disponível em http://www.amazon.com/Seaways-Matthew-Fontaine-Pioneer-Oceanographer/dp/B0012NUKEU; Charles Lee Lewis. Matthew Fontaine Maury: The Pathfinder of the Seas. U.S. Naval Institute, 1927. Também disponível em http://archive.org/details/matthewfontainem00lewi; e Matthew Fontaine Maury. The Physical Geography of the Sea. Harper, 1855. Também disponível em http://books.google.co.uk/books?id=hlxDAAAAIAAJ.

Citações de Maury — Maury. Physical Geography of the Sea. "Introduction," pp. xii, vi.

Dados dos assentos dos carros — Nikkei. "Car Seat of Near Future IDs Driver's Backside". 14 de dezembro de 2011.

A medida da realidade — Boa parte do raciocínio do autor sobre a história da dataficação foi inspirada em Alfred W. Crosby. The Measure of Reality: Quantification and Western Society, 1250–1600. Cambridge University Press, 1997. Citações das pp. 111–113.

Europeus nunca foram expostos ao ábaco — Ibid., 112.

Cálculo seis vezes mais rápido com o uso de algarismos arábicos que tábuas de calcular — Alexander Murray. Reason and Society in the Middle Ages. Oxford University Press, 1978, p. 166.

Número total de livros publicados e estudo de Harvard sobre o projeto de digitalização de livros — Jean-Baptiste Michel et al. "Quantitative Analysis of Culture Using Millions of Digitized Books". Science 331 (14 de janeiro de 2011), pp. 176–182. Também disponível em http://www.sciencemag.org/content/331/6014/176.abstract. Para a palestra de um vídeo sobre o trabalho, veja Erez Lieberman Aiden e Jean-Baptiste Michel. "What We Learned from 5 Million Books". TEDx, Cambridge, MA, 2011. Também disponível em http://www.ted.com/talks/what_we_learned_from_5_million_books.html.

Sobre módulos sem fio em carros e seguradoras — Veja Cukier. "Data, Data Everywhere". The Economist, 27 de fevereiro de 2010.

Jack Levis da UPS — Entrevista com Cukier, abril de 2012.

Dados sobre a economia da UPS — INFORMS (Institute for Operations Research and the Management Sciences). "UPS Wins Gartner BI Excellence Award". 2011. Também disponível em

http://www.informs.org/Announcements/UPS-wins-Gartner-BIExcellence-Award.

Pesquisa Pentland — Robert Lee Hotz. "The Really Smart Phone". The Wall Street Journal, 22 de abril de 2011. Também disponível em http://online.wsj.com/article/SB10001424052748704547604576263261679848814.html.

Estudo Eagle das favelas — Nathan Eagle. "Big Data, Global Development, and Complex Systems". Santa Fe Institute, 5 de maio de 2010. Também disponível em http://www.youtube.com/watch?v=yaivtqlu7iM. Informações adicionais da entrevista com Cukier, outubro de 2012.

Dados do Facebook — Do prospect de IPO, 2012: http://sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm.

Dados do Twitter — Alexia Tsotsis. "Twitter Is at 250 Million Tweets per Day, iOS 5 Integration Made Signups Increase 3x". TechCrunch, 17 de outubro de 2011. Também disponível em http://techcrunch.com/2011/10/17/twitter-is-at-250-million-tweets-per-day/.

Fundos hedge com o uso do Twitter — Kenneth Cukier. "Tracking Social Media: The Mood of the Market". The Economist on-line, 28 de junho de 2012. Também disponível em http://www.economist.com/blogs/graphicdetail/2012/06/trackingsocial-media.

Twitter e a previsão de renda de um filme de Hollywood — Sitaram Asur e Bernardo A. Huberman. "Predicting the Future with Social Media". Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 492–499; uma versão on-line está disponível em http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf.

Twitter e humor global — Scott A. Golder e Michael W. Macy. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures". Science 333 (30 de setembro de 2011), pp. 1878–81.

Twitter e vacinas contra a gripe — Marcel Salathé e Shashank Khandelwal. "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control". PLoS Computational Biology, outubro de 2011.

Patente da IBM — Patente americana n. 8.138.882. Data de aplicação: 5 de fevereiro de 2009.

Autodado quantificado — "Counting Every Moment". The Economist, 3 de março de 2012.

Fones de biomedição da Apple — Jesse Lee Dorogusker, Anthony Fadell, Donald J. Novotney e Nicholas R Kalayjian. "Integrated Sensors for Tracking Performance Metrics". U.S. Patent Application 20090287067. Aplicante: Apple. Data da aplicação: 23 de julho de 2009. Data de publicação: 19 de novembro de 2009.

Derawi Biometrics – "Your Walk Is Your PIN-Code". Press release, 21 de fevereiro de 2011. Também disponível em http://biometrics.derawi.com/?p=175.

Informação iTrem — Veja a página do projeto iTrem do Landmarc Research Center, da Georgia Tech: http://eosl.gtri.gatech.edu/Capabilities/LandmarcResearchCenter/LandmarcProjects/iTrem/tabid/798/Default.aspx e as trocas de e-mail.

Pesquisadores de Kioto sobre acelerômetros de três eixos — iMedicalApps Team. "Gait Analysis Accuracy: Android App Comparable to Standard Accelerometer Methodology". mHealth, 23 de março de 2012.

Os jornais deram origem ao Estado — Benedict Anderson. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, 2006.

Os físicos sugerem que a informação é a base de tudo — Hans Christian von Baeyer. Information: The New Language of Science. Harvard University Press, 2005.

CAPÍTULO 6

História de Luis von Ahn — Com base nas entrevistas de Cukier com von Ahn em 2010 e 2011. Veja também Luis von Ahn. "Luis von Ahn: Expert Q&A". *NOVA scienceNOW,* 6 de julho de 2009. Também disponível em http://www.pbs.org/wgbh/nova/tech/von-ahn-captcha.html; Clive Thompson. "For Certain Tasks, the Cortex Still Beats the CPU". *Wired,* 25 de junho de 2007. Também disponível em http://www.wired.com/techbiz/it/magazine/15-07/ff_humancomp?currentPage=all; Byron Spice. "Brilliant Young Scientist Luis von Ahn Earns $500,000 MacArthur Foundation 'Genius Grant'. *Carnegie Mellon Today,* 18 de setembro de 2006. Também disponível em http://www.cmu.edu/cmnews/extra/060918_ahn.html; Jessie Scanlon. "Luis von Ahn: The Pioneer of 'Human Computation'. *Businessweek,* 3 de novembro de 2008. Também disponível em http://www.businessweek.com/stories/2008-11-03/luis-von-ahn-thepioneer-of-human-computation-businessweek-business-news-stockmarket-and-financial-advice. Sua descrição técnica do reCaptchas está em Luis von Ahn *et al*. "reCAPTCHA: Human-Based Character Recognition via Web Security Measures". *Science* 321(12 de setembro de 2008), pp. 1465–68. Também disponível em http://www.sciencemag.org/content/321/5895/1465.abstract.

Fábrica de Smith — Adam Smith. *The Wealth of Nations* (reimpressão, Bantam Classics, 2003), livro I, Cap. 1. (Uma versão eletrônica gratuita está disponível em http://www2.hn.psu.edu/faculty/jmanis/adam-smith/Wealth-Nations.pdf.

Armazenamento — Viktor Mayer-Schönberger. *Delete: The Virtue of Forgetting in the Digital Age.* Princeton University Press, 2011, 2nd ed., p. 63.

Sobre o consumo dos carros elétricos — IBM. "IBM, Honda, and PG&E Enable Smarter Charging for Electric Vehicles". press release, 12 de abril de 2012. (http://www-03.ibm.com/press/us/en/pressrelease/37398.wss). Veja também Clay Luthy. "Guest Perspective: IBM Working with PG&E to Maximize the EV Potential". *PGE Currents* Magazine, 13 de abril de 2012. Também disponível em http://www.pgecurrents.com/2012/04/13/ibm-working-with-pge-tomaximize-the-ev-potential.

Dados da Amazon e da AOL — Entrevista de Cukier com Andreas Weigend, 2010.

Software Nuance e Google — Cukier, "Data, Data Everywhere".

Empresa de logística — Brad Brown, Michael Chui e James Manyika. "Are You Ready for the Era of 'Big Data'?" *McKinsey Quarterly,* outubro de 2011, p. 10.

Telefônica ganha dinheiro com informações de celulares — "Telefonica Hopes 'Big Data' Arm Will Revive Fortunes". BBC On-line, 9 de outubro de 2012. Também disponível em http://www.bbc.co.uk/news/technology-19882647.

Estudo da Sociedade Dinamarquesa para o Câncer — Patrizia Frei *et al*. "Use of Mobile Phones and Risk of Brain Tumours: Update of Danish Cohort Study". *BMJ* 343 (2011). Também disponível em http://www.bmj.com/content/343/bmj.d6387, e entrevista com Cukier, outubro de 2012.

Registros de GPS do Google's Street View e carros autoguiados — Peter Kirwan. "This Car Drives Itself". *Wired UK,* janeiro de 2012. Também disponível em http://www.wired.co.uk/magazine/archive/2012/01/features/thiscar-drives-itself?page=all.

Corretor ortográfico da Googler e citação — Entrevista com Cukier no Googleplex, em Mountain View, Califórnia, dezembro de 2009; alguns materiais também apareceram em Cukier. "Data, Data Everywhere".

Ideia de Hammerbacher — Entrevista com Cukier, outubro de 2012.

Barnes & Noble analisou dados de seu e-reader, Nook – Alexandra Alter. "Your E-Book Is Reading You". *The Wall Street Journal,* 29 de junho de 2012. Também disponível em http://online.wsj.com/article/SB10001424052702304870304577490950051438304.html.

Aula e dados de Andrew Ng's Coursera — Entrevista com Cukier, junho de 2012.

Diretriz de transparência de dados de Obama — Barack Obama. "Presidential memorandum". Casa Branca, 21 de janeiro de 2009.

Sobre o valor dos dados do Facebook — Para uma excelente análise da discrepância entre o mercado e o valor do IPO do Facebook, veja Doug Laney. "To Facebook You're Worth $80.95". *The Wall Street Journal,* 3 de maio de 2012. Também disponível em http://blogs.wsj.com/cio/2012/05/03/tofacebook-youre-worth-80-95/. Para avaliar os itens discricionários do Facebook, Laney extrapolou o crescimento do Facebook para estimar 2,1 trilhões de peças de conteúdo. Em seu artigo no *WSJ,* ele calculou os itens a $0,03 cada, usando o valor de mercado do Facebook de US$75 bilhões. Por fim, o valor foi de mais de US$100 bilhões, ou US$0,05, ao extrapolarmos o valor com base em seu cálculo.

Abismo entre o valor dos bens físicos e intangíveis — Steve M. Samek. "Prepared Testimony: Hearing on Adapting a 1930's Financial Reporting Model to the 21st Century". U.S. Senate Committee on Banking, Housing and Urban Affairs, Subcommittee on Securities, 19 de julho de 2000.

Valor do intangível — Robert S. Kaplan e David P. Norton. *Strategy Maps: Converting Intangible Assets into Tangible Outcomes.* Harvard Business Review Press, 2004, pp. 4–5.

Dados como bem corporativo das operadoras sem fio americanas — De uma entrevista com uma autoridade sênior de uma agência intergovernamental. Entrevista com Cukier, novembro de 2011.

Citação de Tim O'Reilly — entrevista com Cukier, fevereiro de 2011.

CAPÍTULO 7

Informações sobre o Decide.com — Troca de e-mails entre Cukier e Etzioni, maio de 2012.

Relatório McKinsey — James Manyika *et al*. "Big Data: The Next Frontier for Innovation, Competition, and Productivity". McKinsey Global Institute, maio de 2011. Também disponível em http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation, p. 10.

Citação de Hal Varian — Entrevista com Cukier, dezembro de 2009.

Citação de Carl de Marcken, da ITA — Troca de e-mails com Cukier, maio de 2012.

Sobre MasterCard Advisors — Entrevista de Cukier com Gary Kearns, executivo da MasterCard Advisors, na conferência "The Ideas Economy: Information", da *The Economist*, Santa Clara, Califórnia, 8 de junho de 2011.

Accenture e a cidade de St. Louis, Missouri — Entrevista de Cukier com funcionários municipais, fevereiro de 2007.

Microsoft Amalga Unified Intelligence System — "Microsoft Expands Presence in Healthcare IT Industry with Acquisition of Health Intelligence Software Azyxxi",

Microsoft press release, 26 de julho de 2006, Também disponível em http://www.microsoft.com/enus/news/press/2006/jul06/07-26azyxxiacquisitionpr.aspx. O serviço Amalgama agora faz parte de uma *joint-venture* entre a Microsoft e a General Electric, chamada Caradigm.

Bradford Cross — Entrevistas com Cukier, março a outubro de 2012.

Amazon e a "filtragem colaborativa" — Prospect de IPO, maio de 1997. Também disponível em http://www.sec.gov/Archives/edgar/data/1018724/0000891020-97-000868.txt.

Microprocessadores em carros — Nick Valery. "Tech.View: Cars and Software Bugs". *The Economist,* 16 de maio de 2010. Também disponível em http://www.economist.com/blogs/babbage/2010/05/techview_cars_and_software_bugs.

Maury chamava os navios de "observatórios flutuantes" — Maury. *The Physical Geography of the Sea.*

Inrix — Entrevista de Cukier com executivos, maio e setembro de 2012.

Sobre o Health Care Cost Institute — Sarah Kliff. "A Database That Could Revolutionize Health Care". *Washington Post,* 21 de maio de 2012.

Sobre o Google Street View e os carros autoguiados — Kirwan. "This Car Drives Itself".

Acordo de uso da Decide.com — Troca de e-mail entre Cukier e Etzioni, maio de 2012.

Acordo entre a Google e a ITA — Claire Cain Miller. "U.S. Clears Google Acquisition of Travel Software". *The New York Times*, 8 de abril de 2011. Também disponível em http://www.nytimes.com/2011/04/09/technology/09google.html?_r=0.

Inrix e ABS — Entrevista de Cukier com executivos da Inrix, maio de 2012.

História da Roadnet e citação de Len Kennedy — Entrevista de Cukier, maio de 2012.

Diálogo do filme *O homem que mudou o jogo,* dirigido por Bennett Miller, Columbia Pictures, 2011.

Dados de McGregor acumulados ao longo de uma década — Entrevista com Cukier, maio de 2012.

Citação de Goldbloom — Entrevista com Cukier, março de 2012.

Sobre o faturamento de Hollywood em comparação com a venda de videogames — Para filmes, ver Brooks Barnes. "A Year of Disappointment at the Movie Box Office". *The New York Times,* 25 de dezembro de 2011. Também disponível em http://www.nytimes.com/2011/12/26/business/media/a-year-ofdisappointment-for-hollywood.html. Para videogames, ver "Factbox: A Look at the $65 billion Video Games Industry". Reuters, 6 de junho de 2011. Também disponível em http://uk.reuters.com/article/2011/06/06/us-videogames-factboxidUKTRE7555ZI20110606.

Análise de dados da Zynga — Nick Wingfield. "Virtual Products, Real Profits: Players Spend on Zynga's Games, but Quality Turns Some Off". *The Wall Street Journal,* 9 de setembro de 2011. Também disponível em http://online.wsj.com/article/SB10001424053111904823804576502442835413446.html.

Citação de Ken Rudin — Entrevista de Rudin com Niko Waesche, citada em Erik Schlie, Jörg Rheinboldt e Niko Waesche. *Simply Seven: Seven Ways to Create a Sustainable Internet Business.* Palgrave Macmillan, 2011.

Citação de Thomas Davenport — Entrevista de Cukier com Davenport, dezembro de 2009.

The-Numbers.com — Entrevistas de Cukier com Bruce Nash, outubro de 2011 e julho de 2012.

Estudo Brynjolfsson — Erik Brynjolfsson, Lorin Hitt e Heekyung Kim. "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?". Artigo, abril de 2011. Também disponível em http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486.

Sobre a Rolls-Royce — Ver "Rolls-Royce: Britain's Lonely High-Flier". *The Economist,* 8 de janeiro de 2009. Também disponível em http://www.economist.com/node/12887368. Números atualizados, novembro de 2012.

Erik Brynjolfsson, Andrew McAfee, Michael Sorell e Feng Zhu. "Scale Without Mass: Business Process Replication and Industry Dynamics". Harvard Business School, artigo, setembro de 2006. Também disponível em http://www.hbs.edu/research/pdf/07-016.pdf e também http://hbswk.hbs.edu/item/5532.html.

Sobre o movimento em direção a detentores de dados cada vez maiores — Ver também Yannis Bakos e Erik Brynjolfsson. "Bundling Information Goods: Pricing, Profits, and Efficiency". *Management Science* 45 (dezembro de 1999), pp. 1613–30.

Philip Evans — Entrevista com os autores, 2011 e 2012.

CAPÍTULO 8

Sobre a Stasi — Boa parte da literatura está infelizmente em alemão, mas a exceção é o ótimo trabalho de pesquisa de Kristie Macrakis. *Seduced by Secrets: Inside the Stasi's Spy-Tech World.* Cambridge University Press, 2008; uma história pessoal é contada em Timothy Garton Ash. *The File.* Atlantic Books, 2008. Também recomendamos o filme ganhador do Oscar *A vida dos outros.*

Câmeras de vigilância perto da casa de Orwell — "George Orwell, Big Brother Is Watching Your House". *The Evening Standard,* 31 de março de 2007. Também disponível em http://www.thisislondon.co.uk/news/george-orwell-bigbrother-is-watching-your-house-7086271.html.

Sobre Equifax e Experian — Daniel J. Solove. *The Digital Person: Technology and Privacy in the Information Age.* NYU Press, 2004, pp. 20–21.

Sobre os endereços dos japoneses em Washington entregues pelas autoridades americanas — J.R. Minkel. "The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II". *Scientific American,* 20 de março de 2007. Também disponível em http://www.scientificamerican.com/article.cfm?id=confirmed-theus-census-b.

Informações sobre dados usados pelos nazistas na Holanda — William Seltzer e Margo Anderson. "The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses". *Social Research* 68 (2001), pp. 481–513.

Informação sobre a IBM e o Holocausto — Edwin Black. *IBM e o holocausto.* Rio de Janeiro: Campus/Elsevier, 2001.

Sobre a quantidade de dados que os medidores inteligentes coletam — Elias Leake Quinn. "Smart Metering and Privacy: Existing Law and Competing Policies; A Report for the Colorado Public Utility Commission". Primavera de 2009. Também disponível em http://www.w4ar.com/Danger_of_Smart_Meters_Colorado_Report.pdf. Ver também Joel M. Margolis. "When Smart Grids Grow Smart Enough to Solve

Crimes". Neustar, 18 de março de 2010. Também disponível em http://energy.gov/sites/prod/files/gcprod/documents/Neustar_Comments_DataExhibitA.pdf.

Fred Cate sobre o consentimento — Fred H. Cate. "The Failure of Fair Information Practice Principles". In: Jane K. Winn (org.). *Consumer Protection in the Age of the "Information Economy"*. Ashgate, 2006, p. 341 *et seq*.

Sobre o lançamento de dados da AOL — Michael Barbaro e Tom Zeller Jr. "A Face Is Exposed for AOL Searcher No. 4417749". *The New York Times,* 9 de agosto de 2006. Ver também Matthew Karnitschnig e Mylene Mangalindan. "AOL Fires Technology Chief After Web-Search Data Scandal". *The Wall Street Journal,* 21 de agosto de 2006.

Pessoa identificada pelo Netflix — Ryan Singel. "Netflix Spilled Your *Brokeback Mountain* Secret, Lawsuit Claims". *Wired,* 17 de dezembro de 2009. Também disponível em http://www.wired.com/threatlevel/2009/12/netflixprivacy-lawsuit/.

Sobre o lançamento de dados do Netflix — Arvind Narayanan e Vitaly Shmatikov. "Robust De-Anonymization of Large Sparse Datasets". *Proceedings of the 2008 IEEE Symposium on Security and Privacy,* p. 111. Também disponível em http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf; Arvind Narayanan e Vitaly Shmatikov. "How to Break the Anonymity of the Netflix Prize Dataset". 18 de outubro de 2006; arXiv:cs/0610105 [cs.CR]. Também disponível em http://arxiv.org/abs/cs/0610105.

Sobre identificar pessoas com três características — Philippe Golle. "Revisiting the Uniqueness of Simple Demographics in the US Population". *Association for Computing Machinery Workshop on Privacy in Electronic Society* 5 (2006), p. 77.

Sobre a fraqueza estrutural da anonimização — Paul Ohm. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization". 57 *UCLA Law Review* 1701 (2010).

Sobre o anonimato do gráfico social — Lars Backstrom, Cynthia Dwork e Jon Kleinberg. "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography". *Communications of the Association of Computing Machinery,* dezembro de 2011, p. 133.

"Caixas-pretas" dos carros — "Vehicle Data Recorders: Watching Your Driving". *The Economist,* 23 de junho de 2012. Também disponível em http://www.economist.com/node/21557309.

Coleta de dados da NSA — Dana Priest e William Arkin. "A Hidden World, Growing Beyond Control". *Washington Post,* 19 de julho de 2010. Também disponível em http://projects.washingtonpost.com/top-secretamerica/articles/a-hidden-world-growing-beyond-control/print/. Juan Gonzalez. "Whistleblower: The NSA Is Lying—U.S. Government Has Copies of Most of Your Emails". *Democracy Now,* 20 de abril de 2012. Também disponível em http://www.democracynow.org/2012/4/20/whistleblower_the_nsa_is_lying_us. William Binney. "Sworn Declaration in the Case of Jewel v. NSA". 2 de julho de 2012. Também disponível em http://publicintelligence.net/binney-nsa-declaration/.

Como a vigilância mudou com o big data — Patrick Radden Keefe. "Can Network Theory Thwart Terrorists?" *The New York Times,* 12 de março de 2006. Também disponível em http://www.nytimes.com/2006/03/12/magazine/312wwln_essay.html.

Diálogo de *Minority Report,* dirigido por Steven Spielberg, DreamWorks/20th Century Fox, 2002. O diálogo que citamos é reduzido. O filme se baseia num conto de 1958 de Philip K. Dick, mas há importantes diferenças entre as versões. A cena de abertura do marido traído não aparece no livro, e o enigma filosófico do pré-crime é apresentado com mais ênfase no filme de Spielberg que no conto. Por isso, optamos por usar o filme como comparativo.

Exemplos de policiamento preventivo — James Vlahos. "The Department Of Pre-Crime". *Scientific American* 306 (janeiro de 2012), pp. 62–67.

Sobre a FAST — Sharon Weinberger. "Terrorist 'Pre-crime' Detector Field Tested in United States". *Nature,* 27 de maio de 2011. Também disponível em http://www.nature.com/news/2011/110527/full/news.2011.323.html; Sharon Weinberger. "Intent to Deceive". *Nature* 465 (maio de 2010), pp. 412–415. Sobre o problema dos falsos positivos, ver Alexander Furnas. "Homeland Security's 'Pre-Crime' Screening Will Never Work". The Atlantic On-line, 17 de abril de 2012. Também disponível em http://www.theatlantic.com/technology/archive/2012/04/homelandsecuritys-pre-crime-screening-will-never-work/255971/.

Sobre as notas dos alunos e o preço dos seguros — Tim Query. "Grade Inflation and the Good-Student Discount". *Contingencies Magazine,* American Academy of Actuaries, Maio-Junho de 2007. Também disponível em http://www.contingencies.org/mayjun07/tradecraft.pdf.

Sobre os perigos do perfilamento — Bernard E. Harcourt. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* University of Chicago Press, 2006.

Sobre a obra de Richard Berk — Richard Berk. "The Role of Race in Forecasts of Violent Crime". *Race and Social Problems* 1 (2009), pp. 231–242, e entrevista por e-mail com Cukier, novembro de 2012.

Sobre a adoração de McNamara pelos dados — Phil Rosenzweig. "Robert S. McNamara and the Evolution of Modern Management". *Harvard Business Review,* dezembro de 2010. Também disponível em http://hbr.org/2010/12/robert-smcnamara-and-the-evolution-of-modern-management/ar/pr.

Sobre o sucesso dos "Meninos Sábios" na Segunda Guerra Mundial — John Byrne. *The Whiz Kids.* Doubleday, 1993.

Sobre McNamara na Ford — David Halberstam. *The Reckoning.* William Morrow, setembro de 1986, pp. 222–245.

Livro de Kinnard — Douglas Kinnard. *The War Managers.* University Press of New England, 1977, pp. 71–25. Esta seção se beneficiou de uma entrevista por e-amil com o Dr. Kinnard, graças a seu assistente, pelo qual os autores expressam sua gratidão.

Sobre a citação "Em Deus confiamos..." — Geralmente atribuída a W. Edwards Deming.

Sobre Ted Kenney e a lista de passageiros proibidos — Sara Kehaulani Goo. "Sen. Kennedy Flagged by No-Fly List". *Washington Post,* 20 de agosto de 2004, p. A01. Também disponível em http://www.washingtonpost.com/wpdyn/articles/A17073-2004Aug19.html.

Práticas de contratação da Google — Douglas Edwards. *I'm Feeling Lucky: The Confessions of Google Employee Number 59.* Houghton Mifflin Harcourt, 2011, p. 9. Veja também Steven Levy. *In the Plex.* Simon and Schuster, 2011, pp. 140–141.

Ironicamente, os fundadores da Google quiseram contratar Steve Jobs como CEO (apesar da falta de diploma); Levy, p. 80.

Testando 41 tons de azul — Laura M. Holson. "Putting a Bolder Face on Google". *The New York Times,* 1° de março de 2009. Também disponível em http://www.nytimes.com/2009/03/01/business/01marissa.html.

Demissão do designer-chefe da Google — Citação tirada (na íntegra) de Doug Bowman. "Goodbye, Google". Post de blog, 20 de março de 2009. Também disponível em http://stopdesign.com/archive/2009/03/20/goodbye-google.html.

Citação de Jobs — Steve Lohr. "Can Apple Find More Hits Without Its Tastemaker?" *The New York Times,* 8 de janeiro de 2011, p. B1. Também disponível em http://www.nytimes.com/2011/01/19/technology/companies/19innovate.html.

Livro de Scott — James Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* Yale University Press, 1998.

Citação de McNamara — De um discurso na Millsaps College, em Jackson, Mississippi, citado em *Harvard Business Review,* dezembro de 2010.

Sobre a desculpa de McNamara — Robert S. McNamara com Brian VanDeMark. *In Retrospect: The Tragedy and Lessons of Vietnam.* Random House, 1995, pp. 48, 270.

CAPÍTULO 9

Sobre a biblioteca da Cambridge University — Marc Drogin. Anathema! Medieval Scribes and the History of Book Curses. Allanheld and Schram, 1983, p. 37.

Sobre responsabilidade e privacidade — O Center for Information Policy Leadership está envolvido num projeto multianual sobre interface de responsabilidade e privacidade; veja http://www.informationpolicycentre.com/accountabilitybased_privacy_governance/.

Sobre a data de validade dos dados — Mayer-Schönberger. Delete.

"Privacidade diferencial" — Cynthia Dwork. "A Firm Foundation for Private Data Analysis". Communications of the ACM, janeiro de 2011, pp. 86–95.

Facebook e privacidade diferencial — A. Chin e A. Klinefelter. "Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study". 90 North Carolina Law Review 1417 (2012); A. Haeberlen et al. "Differential Privacy Under Fire". Também disponível em http://www.cis.upenn.edu/~ahae/papers/fuzz-sec2011.pdf.

Sobre os representantes alemães de proteção a dados corporativos — Viktor Mayer-Schönberger. "Beyond Privacy, Beyond Rights: Towards a 'Systems' Theory of Information Governance". 98 California Law Review 1853 (2010).

Sobre interoperabilidade — John Palfrey e Urs Gasser. Interop: The Promise and Perils of Highly Interconnected Systems. Basic Books, 2012.

CAPÍTULO 10

Mike Flowers e os analistas de Nova York — Baseado em entrevista com Cukier, julho de 2012. Para descrição mais abrangente, veja: Alex Howard. "Predictive data analytics is saving lives and taxpayer dollars in New York City". O'Reilly Media, 26

de junho de 2012. Também disponível em http://strata.oreilly.com/2012/06/predictive-data-analytics-big-datanyc.html.

Walmart e Pop-Tarts — Constance L. Hays. "What Wal-Mart Knows About Customers' Habits". *The New York Times,* 14 de novembro de 2004. Também disponível em http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html.

Uso do big data em favelas e para modelar movimentos de refugiados — Nathan Eagle. "Big Data, Global Development, and Complex Systems": http://www.youtube.com/watch?v(yaivtqlu7iM.

Percepção do tempo — Benedict Anderson. *Imagined Communities,* New edition. Verso, 2006.

Experimento CERN e armazenamento de dados — Troca de e-mails entre Cukier e pesquisadores do CERN, novembro de 2012.

Sistema computacional da Apollo 11 — David A. Mindell. *Digital Apollo: Human and Machine in Spaceflight.* MIT Press, 2008.