Maio | 2018

Big Data para o desenvolvimento
Ano 10 – Número 1

Panorama setorial da Internet

Big Data e desenvolvimento: uma visão geral

Emmanuel Letouzé [2] *(Data-Pop Alliance)* [3]

Fatos, Números e Contexto Geral

Apesar do entusiasmo existente em torno do "*Big Data* para o desenvolvimento" – expressão que abrange o campo de pesquisa e prática sobre as aplicações e implicações do *Big Data* para a elaboração de políticas e para o desenvolvimento – o seu uso ainda é muito incipiente em termos intelectual e operacional. Seria ele o "novo petróleo" que poderá se tornar uma bênção ou uma maldição para o desenvolvimento humano e o progresso social? Os otimistas o consideram uma revolução que mudará "a forma como vivemos, pensamos e trabalhamos", tendo até expressado a esperança de que "a tragédia estatística da África" possa ser parcialmente solucionada pelo *Big Data*. Entretanto, os céticos e os críticos têm sido mais circunspectos, ou claramente antagônicos, referindo-se ao *Big Data* como um grande truque, um grande exagero ou um grande risco, assim como, é claro, o *Big Brother*. Esse entusiasmo em torno do *Big Data* poderia ser em vão, apontam alguns observadores, afinal a análise automatizada de grandes conjuntos de dados não é novidade. Então, o que há de novo?

[1] Esta é uma versão editada e traduzida. Para ler o artigo original (em inglês), acesse o *website* da Data-Pop Alliance. Recuperado em 07 maio, 2018, de http://datapopalliance.org/wp-content/uploads/2015/12/Big-Data-Dev-Overview.pdf

[2] Emmanuel Letouzé é diretor e co-fundador da Data-Pop Alliance e pesquisador visitante no MIT Media Lab, pesquisador afiliado no Harvard Humanitarian Initiative e pesquisador associado no Overseas Development Institute. É autor do White Paper "Big Data for Development" da UN Global Pulse e dos relatórios Fragiles States de 2013 e 2014 da OCDE. Sua pesquisa e trabalho se centram nas aplicações e implicações do *Big Data* para as estatísticas oficiais, pobreza e desigualdade, conflito, crime, mudança climática, vulnerabilidade e resiliência, e direitos humanos, ética e política.

[3] A Data-Pop Alliance é um grupo de reflexão (*think tank*) que se dedica ao estudo de *Big Data* e desenvolvimento, criado pela Harvard Humanitarian Initiative, pelo MIT Media Lab e pelo Overseas Development Institute, para promover uma revolução do *Big Data* centrada no indivíduo.

Embora não exista uma única definição acordada sobre o *Big Data*, ele deve ser considerado um novo ecossistema que é parte integrante de um fenômeno social mais amplo, impulsionado pelas tecnologias digitais.

Dos 3 Vs aos 3 Cs do *Big Data*

Embora não exista uma única definição acordada sobre o *Big Data*, ele deve ser considerado um novo ecossistema que é parte integrante de um fenômeno social mais amplo, impulsionado pelas tecnologias digitais. Esse ecossistema pode ser caracterizado como a união de 3 Cs: *crumbs* (migalhas), *capacities* (capacidades) e *communities* (comunidades). Essa caracterização é mais precisa e menos confusa do que os 3 Vs (volume, velocidade e variedade), usados nos primeiros anos do *Big Data*, entre 2010 e 2012. A principal limitação dos 3 Vs é focar exclusivamente no *Big Data* como sendo apenas uma "grande quantidade de dados". Outra limitação é enfatizar que o *Big Data* seria essencialmente uma mudança quantitativa, e não qualitativa. Nunca é demais repetir: não se pode restringir o *Big Data* a uma grande quantidade de dados, por isso, os 3 Vs devem ser relegados ao passado.

O que são os 3 Cs? O primeiro C, *crumbs*, refere-se às "migalhas digitais" ou "migalhas de dados", conforme definição de Alex "Sandy" Pentland, professor do Massachusetts Institute of Technology (MIT) e diretor acadêmico da Data-Pop Alliance. Ao contrário dos dados de pesquisas amostrais tradicionais, os *crumbs* não são produzidos para fins de inferência estatística; em vez disso, eles são, em sua maior parte, deixados para trás de maneira passiva por humanos que usam dispositivos e serviços digitais, muitos dos quais não estavam disponíveis há cinco ou dez anos. Cada uma dessas ações deixa um rastro digital, que, somados, compõem a maior parte do *Big Data* como fonte de dados. A proporção de dados digitais produzidos recentemente está crescendo muito rapidamente – frequentemente diz-se que até 90% dos dados mundiais foram gerados apenas no último ano, embora a fonte exata da metodologia dessa estimativa não seja clara.

Esses dados estão disponíveis em três tipos principais. Um deles é o dado pequeno, "concreto", estruturado, que pode ser facilmente quantificado e organizado (em colunas e linhas, por exemplo) para análise sistemática, e que não pode ser editado por seus produtores. Exemplos incluem registros de detalhes de chamadas (do inglês, *Call Detail Records* – CDR) produzidos pelas operadoras de telecomunicações, de transações com cartão de crédito, bem como registros de bilhetes de metrô ou de pagamento automático de pedágio. Há quem argumente que esse tipo de dado constitui a verdadeira novidade e promessa do *Big Data*. De acordo com Alex "Sandy" Pentland, "o poder do *Big Data* reside no fato de se tratar de informações sobre o comportamento das pessoas em vez de informações sobre suas crenças".

Um segundo tipo de dados inclui vídeos, documentos *on-line*, postagens em *blogs* e outros conteúdos de mídia social. São dados "não estruturados", mais difíceis de analisar de maneira automatizada. Eles também estão mais sujeitos às escolhas editoriais dos seus autores: uma pessoa poderia declarar boicotar um determinado produto em um *blog* e, por outro lado, o extrato de seu cartão de crédito revelar uma história diferente. Um terceiro tipo refere-se aos dados que são coletados por sensores digitais que captam ações humanas, como medidores elétricos ou imagens de satélite que podem captar imagens que revelam, por exemplo, o desmatamento de uma floresta. Alguns consideram o universo do *Big Data* muito mais amplo, de modo a incluir, por exemplo, registros administrativos, dados sobre preços ou clima, ou livros que foram previamente digitalizados – o que, coletivamente, pode constituir um quarto tipo. Significativamente, conforme mencionado, o que esses dados têm em comum é o fato de não terem sido coletados ou amos-

trados com a intenção explícita de que se tirem conclusões a partir deles. Portanto, o termo *Big Data* é essencialmente impreciso: o tamanho não é uma característica determinante, é apenas um corolário de sua natureza. Mesmo um pequeno conjunto desses dados pode ser considerado *Big Data*, se não for proveniente de processos totalmente controlados, como pesquisas e imputações estatísticas realizadas por órgãos oficiais.

REGISTROS DE DETALHES DE CHAMADAS (CDR)

Os registros de detalhes de chamadas (CDR) são metadados (dados sobre dados) que captam o uso de telefones pelos assinantes – incluindo um código de identificação e, no mínimo, a localização da torre telefônica que roteou a chamada, além da hora e duração da chamada. Grandes operadoras coletam mais de seis bilhões de CDR por dia.

IDENTIFICAÇÃO DO CHAMADOR	LOCALIZAÇÃO DA TORRE ERB* DA CÉLULA DO CHAMADOR	NÚMERO DE TELEFONE DO RECEPTOR	LOCALIZAÇÃO DA TORRE ERB* DA CÉLULA DO RECEPTOR	HORA DA CHAMADA	DURAÇÃO DA CHAMADA
X76VG588RLPQ	2°24' 22.14" 35°49' 56.54"	A81UTC93KK52	3°26' 30.47" 31° 12' 18.01"	2013-11-07T15:15:00	01:12:02

www.unglobalpulse.org/Mobile_Phone_Network_Data-for-Dev
*Estação Rádio Base

Se esses dados formam o núcleo do *Big Data* como um ecossistema, eles não constituem a sua totalidade. O segundo C de *Big Data* se refere a capacidades – ferramentas, métodos, *software* e *hardware*: nas palavras de Gary King, professor da Universidade de Harvard, "o *Big Data* não é uma questão que diz respeito apenas aos dados". Essas capacidades incluem computadores poderosos, sistemas de computação paralela, bem como técnicas estatísticas de *machine learning* (aprendizado de máquina) e algoritmos capazes de procurar e desvendar padrões e tendências em grandes quantidades de dados complexos.

O terceiro C é de comunidades. O *Big Data* também é composto pelo "movimento" de atores individuais e institucionais que operam em grande parte fora das esferas tradicionais da política e da pesquisa; são equipes multidisciplinares de cientistas sociais e da computação com uma "mentalidade para ordenar a desordem, transformando-a em significado", como diz o cientista de dados Andreas Weigend. Isso também inclui pessoas comuns que usam o Google Maps para decidir se irão a uma reunião de carro ou de metrô. Na era do *Big Data*, mais do que nunca, todos passam a ser tomadores de decisão baseados em dados disponíveis.

Esse conceito sugere que o *Big Data* é um sistema complexo, com ciclos de *feedback*. Novos métodos resultarão em novos dados; novos dados estimularão a ideia de criar uma *startup* de ciência de dados, etc. Isso também aponta que a expressão "usar *Big Data*" não enfoca a questão principal, a menos que o intuito seja explicitamente o uso do ecossistema *Big Data* para atingir

A empolgação com o *Big Data* se originou de dois conjuntos de fatores: da crescente oferta de dados e da capacidade de analisá-los, e da demanda por informações melhores, mais rápidas e mais baratas.

determinados objetivos. Em vez disso, a questão relevante e o desafio que se colocam são por que e como se envolver com o *Big Data* – como tentar fazer parte dele, influenciar sua evolução e/ou se beneficiar das suas inovações?

A promessa: fatores de oferta e demanda

A empolgação com o *Big Data* se originou de dois conjuntos de fatores: da crescente oferta de dados e da capacidade de analisá-los, e da demanda por informações melhores, mais rápidas e mais baratas – em outras palavras, houve e continua havendo um ímpeto e uma atração em direção ao *Big Data*.

A disponibilidade de dados confiáveis e atualizados tem melhorado consideravelmente ao longo do tempo; no entanto, em muitos casos, as lacunas permanecem. Por exemplo, um bom indicador da pobreza ou do subdesenvolvimento de uma região é justamente a falta de dados sobre pobreza ou desenvolvimento. Alguns países (especialmente aqueles com uma história recente de conflito) não realizam censo demográfico há quatro décadas ou mais. É comum que o tamanho, a estrutura e a distribuição da população só possam ser obtidos com a triangulação de informações de diferentes fontes. Mesmo que existam números oficiais, eles geralmente baseiam-se em dados incompletos. Dados de má qualidade também fazem com que os números oficiais do Produto Interno Bruto (PIB) de alguns países aumentem de maneira repentina – 40% para Gana, em 2010, ou 60% para a Nigéria, em 2014 – quando mudanças na estrutura de suas economias, como a ascensão do setor de tecnologia, são finalmente levados em conta.

Essa falta de dados confiáveis tem sido responsável pelo apelo por uma "Revolução de Dados", o que levou à publicação de um relatório por parte de um grupo de especialistas indicados pela Organização das Nações Unidas (ONU). O raciocínio básico e um tanto simplista é que, na era do *Big Data*, as economias deveriam ser dirigidas por formuladores de políticas que dispõem de melhores instrumentos de navegação e indicadores que lhes permitam projetar e implementar políticas e programas mais ágeis e melhor direcionados. Diz-se que o *Big Data* tem o potencial de fazer com que sistemas estatísticos nacionais em áreas com poucos dados pulem etapas, da mesma forma como muitos países em desenvolvimento ignoraram a era do telefone fixo para ir direto para a era do telefone celular.

O apelo para potencialmente saltar à frente também é moldado pelo "lado da oferta" de fontes de *Big Data*. Há evidências em práticas iniciais, além de um crescente conjunto de trabalhos sobre o novo potencial do *Big Data*, que ajudam a entender o seu funcionamento e a maneira como ele afeta as populações e os processos humanos. Por exemplo, o *Big Data* tem sido usado para acompanhar a inflação *on-line*, estimar e prever mudanças no PIB quase em tempo real, monitorar o tráfego ou o surto de epidemias. O monitoramento de dados de mídias sociais para analisar os sentimentos das pessoas está abrindo novos caminhos para medir o bem-estar, enquanto dados de *e-mail* e do Twitter podem ser usados para estudar a migração interna e internacional. Além disso, há uma abundante literatura acadêmica que usa os CDR para estudar padrões de migração, níveis

socioeconômicos e disseminação de doenças, entre outros temas. Da mesma forma que os *smartphones* em breve ultrapassarão os telefones celulares comuns em todo o mundo, o uso de CDR para fins analíticos diminuirá e novas "migalhas de dados" se tornarão as próximas fronteiras.

Enquanto isso, taxonomias são propostas para esclarecer como o *Big Data* pode favorecer o desenvolvimento. Uma delas faz a distinção entre os usos de *Big Data* para "aviso antecipado" e para a "conscientização em tempo real", ou "monitoramento em tempo real" do impacto provocado por uma política ou programa. Outra contrasta a função descritiva do uso de *Big Data*, como um mapa em tempo real, com a de aplicações preditivas e prescritivas.

O uso preditivo de *Big Data* pode ser entendido em dois sentidos do termo, seja como inferência ou *nowcasting*, predizendo o que está acontecendo agora (como quando a atividade do telefone celular é usada para prever níveis socioeconômicos), ou então como previsão (de uma maneira muito semelhante ao que os meteorologistas fazem). O uso prescritivo requer inferências causais, isto é, o estabelecimento da existência, direção e magnitude de um nexo causal entre algum tratamento ou variável X e algum efeito ou variável Y. A tabela "Taxonomias de usos reais e potenciais de *Big Data* para o desenvolvimento", no Anexo, fornece exemplos de aplicações que se enquadram em cada categoria de uso.

O lado cinzento do *Big Data*: riscos e desafios

A promessa da aplicação de *Big Data* para resolver problemas do mundo real foi recebida com advertências sobre seus perigos e com discussões mais amplamente ativas sobre suas implicações sociais. Talvez os riscos mais graves sejam aqueles que dizem respeito aos direitos individuais e de grupo, e também à privacidade, à identidade e à segurança. Além da intrusão óbvia de atividades de vigilância, com temas que envolvem sua legalidade e legitimidade, há questões importantes sobre a "anonimização de dados": o que isso significa e quais são seus limites. Um estudo preliminar, tendo por base o aluguel de filmes, mostrou que mesmo dados "anônimos" poderiam ser "desanonimizados" – e vinculados a um determinado indivíduo, por meio da correlação das datas de aluguel de apenas três filmes com as datas das postagens em uma plataforma de filmes *on-line*. Outras pesquisas descobriram que os CDR que registram a localização e horário, mesmo quando livres de qualquer identificador individual, podem ser reindividualizados – o que é chamado de reidentificação. Nesse caso, quatro pontos de dados seriam teoricamente suficientes para singularizar, com 95% de precisão, os indivíduos em um conjunto de dados. Pesquisas recentes usando transações com cartões de crédito geraram conclusões muito semelhantes: nossos comportamentos são únicos e suficientemente previsíveis para tornar difícil para qualquer indivíduo se esconder na multidão digital.

Os críticos também apontam os riscos básicos associados à tomada de decisões com base em análises sem validade externa ou interna.

Talvez os riscos mais graves sejam aqueles que dizem respeito aos direitos individuais e de grupo, e também à privacidade, à identidade e à segurança.

Big Data – riscos para conclusões válidas

Um dos principais desafios do *Big Data* é que as pessoas que o geram se selecionaram a si mesmas como geradoras de dados por meio de sua atividade. Em termos técnicos, trata-se de um "viés de seleção", o que significa que a análise de *Big Data* provavelmente produzirá um resultado diferente de uma pesquisa tradicional, a qual buscaria um recorte representativo de uma determinada população. Por exemplo, ao tentar responder à pergunta "as pessoas do país A preferem arroz ou batatas fritas?" por meio da mineração de dados do Twitter, a resposta seria tendenciosa em favor da preferência dos jovens, uma vez que eles compõem o maior número de usuários da plataforma. Portanto, análises baseadas em *Big Data* podem não ter "validade externa", embora seja possível que indivíduos muito diferentes entre si possam ter preferências semelhantes e exibir comportamentos idênticos (jovens podem ter as mesmas preferências que idosos). Outro risco advém de análises, que são falhas, porque carecem de "validade interna". Por exemplo, uma queda acentuada no volume de CDR de uma região poderia ser interpretada, com base em ocorrências anteriores, como o anúncio de um conflito iminente – quando, na verdade, essa flutuação poderia ter sido causada por algo como a queda de uma torre de telefonia móvel naquela localidade.

Outro risco deriva do fato de as análises baseadas em *Big Data* se concentrarem demais na correlação e na previsão – relegando a um segundo plano causa, diagnóstico ou inferência causal, sem as quais a política é essencialmente cega. Um bom exemplo é o "policiamento preditivo". A polícia e as autoridades responsáveis pela aplicação da lei em algumas cidades dos Estados Unidos e do Reino Unido, durante anos, divulgaram dados para avaliar a probabilidade do aumento da criminalidade em certas áreas, prevendo aumentos baseados em padrões históricos. As autoridades policiais despenderam seus recursos de acordo com a necessidade e, com isso, na maioria dos casos, houve uma diminuição da criminalidade. No entanto, a menos que se saiba o porquê do aumento da criminalidade, é difícil implementar uma política preventiva que enfrente as causas básicas ou os fatores contribuintes. Ao mesmo tempo, os defensores argumentam que a repressão ao crime em uma determinada área pode ter um efeito estrutural cumulativo.

Outro grande risco que está recebendo crescente atenção é o potencial do *Big Data* para a criação de um novo "hiato digital", o qual pode ampliar, em vez de eliminar, as lacunas existentes em termos de renda e poder em todo o mundo. Um dos "três paradoxos"[4] do *Big Data* ocorre porque, ao exigir capacidades analíticas e de acesso a dados que somente uma fração de instituições, corporações e indivíduos tem, isso pode tirar o poder das comunidades e dos países aos quais ele promete servir. As pessoas com mais dados e capacitação estão em melhor posição para desenvolver sistemas de *Big Data* para tirar suas próprias vantagens econômicas e políticas, mesmo quando afirmam usá-las para beneficiar outros indivíduos.

Um último desafio básico é o da utilização dos dados – fundamentalmente para entender como os dados afetam historicamente as sociedades. A maioria das dis-

4 Para aprofundar sobre os três paradoxos, veja: Richards, N.M. e King, J.H. (2013). Three paradoxes of Big Data. Stanford Law Review.

cussões sobre a "revolução dos dados" pressupõe que "dados são importantes" e que os de má qualidade resultam em políticas de má qualidade. Mas, historicamente, a falta de dados tem desempenhado apenas um papel marginal nas decisões que levam a políticas ruins e a resultados insatisfatórios. E um futuro "algoritmicamente" cego pode minar os próprios processos que visam a garantir que a forma como os dados são convertidos em decisões está sujeita à fiscalização democrática. Ao mesmo tempo, existe um enorme potencial para que as sociedades entendam e influenciem processos com os quais se deparam há séculos. Cumprir essa promessa exigirá uma profunda reformulação e reconfiguração dos nossos sistemas políticos, éticos, tecnológicos e legais.

O futuro do *Big Data* ou o *Big Data* do futuro?

Como é altamente improvável que o crescimento da geração de dados diminua e, uma vez que a criatividade e a curiosidade humanas são quase ilimitadas, é improvável que a "bolha" do *Big Data* estoure no futuro próximo. O mundo pode esperar mais discussões e controvérsias sobre o potencial e os perigos do *Big Data* para o desenvolvimento e para as sociedades em geral. O futuro do *Big Data* provavelmente será moldado por três vertentes principais: pesquisa acadêmica, estruturas legais e técnicas para o uso ético de dados e uma maior demanda da sociedade por maior responsabilidade e participação.

As pesquisas continuarão a examinar se e como as fronteiras metodológicas e científicas podem ser expandidas, especialmente em duas áreas: pela extração de inferências causais mais fortes e pela medição e correção do viés da amostra.

O debate sobre políticas desenvolverá *frameworks* e padrões – normativos, legais, e técnicos – para coletar, armazenar e compartilhar fluxos e conjuntos de *Big Data*. Esses desdobramentos se enquadram no termo abrangente "ética do *Big Data*". Os avanços técnicos ajudarão, por exemplo, introduzindo "ruído" nos conjuntos de dados para dificultar a reidentificação dos indivíduos neles representados – embora tais avanços provavelmente nunca venham a tornar isso impossível. Mas um enfoque abrangente da "ética do *Big Data*" idealmente englobaria outras considerações humanísticas, como privacidade e igualdade, bem como alfabetização em dados e *design* centrado nas pessoas.

Uma terceira influência relacionada ao futuro do *Big Data* será a maneira como ele se desenvolve e evolui paralelamente ao movimento de "dados abertos" e seus impulsores sociais subjacentes – sendo que "dados abertos" referem-se a dados facilmente acessíveis, legíveis por máquinas, acessíveis gratuitamente ou a um custo insignificante e com limitações mínimas em seu uso, transformação e distribuição.

No futuro previsível, o *Big Data* e movimentos de "dados abertos" serão os dois principais pilares de uma "revolução de dados" mais ampla. Ambos surgem em um contexto de crescente demanda pública por mais abertura, agilidade, transparência, *accountability* e participação. As implicações políticas – tão facilmente esquecidas – são visíveis. Uma verdadeira revolução de *Big Data* deve ser aquela em que os dados podem ser aproveitados para mudar as estruturas de poder e os processos de tomada de decisão, e não apenas para criar *insights*.

O futuro do *Big Data* provavelmente será moldado por três vertentes principais: pesquisa acadêmica, estruturas legais e técnicas para o uso ético de dados e uma maior demanda da sociedade por maior responsabilidade e participação.

REFERÊNCIAS

As referências e a bibliografia selecionada podem ser acessadas no documento original, disponível em http://datapopalliance.org/wp-content/uploads/2015/12/Big-Data-Dev-Overview.pdf.

Tabela 1 – TAXONOMIAS DE USOS REAIS E POTENCIAIS DE *BIG DATA* PARA O DESENVOLVIMENTO

	APLICAÇÕES	EXPLICAÇÃO
TAXONOMIA DO RELATÓRIO GLOBAL PULSE DA ONU (Letouzé, 2012)	1- AVISO ANTECIPADO	A detecção precoce de anomalias na forma como as populações usam dispositivos e serviços digitais pode permitir uma resposta mais rápida em tempos de crise.
	2- CONSCIENTIZAÇÃO EM TEMPO REAL	O *Big Data* pode fornecer uma representação atual e precisa da realidade, a qual pode inspirar o *design* e o direcionamento de programas e políticas.
	3- MONITORAMENTO EM TEMPO REAL	A capacidade de monitorar uma população em tempo real permite entender onde as políticas e os programas estão falhando e fazer os ajustes necessários.
TAXONOMIA ALTERNATIVA (Letouzé et al., 2013)	1- DESCRITIVO	O *Big Data* pode documentar e transmitir o que está acontecendo.
	2- PREDITIVO	O *Big Data* poderia dar uma ideia daquilo que pode vir a acontecer, independentemente do motivo.
	3- PRESCRITIVO	O *Big Data* pode esclarecer por que determinados fatos podem acontecer e o que poderia ser feito sobre isso.

EXEMPLOS	COMENTÁRIOS
O policiamento preditivo baseia-se na noção de que a análise de dados históricos pode revelar certas combinações de fatores associados a uma maior probabilidade de que venham a ocorrer crimes em uma área; pode ser usado para alocar recursos policiais. As tendências apontadas pelo Google Flu são outro exemplo, por meio do qual pesquisas por termos específicos ("coriza", "coceira nos olhos") são analisadas para detectar o início da temporada de gripe - embora sua precisão ainda esteja sob análise.	Esta aplicação pressupõe que certas regularidades no comportamento humano podem ser observadas e modeladas. Um dos principais desafios enfrentados pelas políticas é a tendência da maioria dos sistemas de detecção de mau funcionamento e dos modelos de previsão de superestimar - ou seja, ter uma maior prevalência de "falsos positivos".
Usando dados divulgados pela Orange, pesquisadores descobriram um alto grau de associação entre as redes de telefonia móvel e a distribuição de idiomas na Costa do Marfim - sugerindo que tais dados podem fornecer informações sobre as comunidades linguísticas em países onde esses dados não estão disponíveis.	O apelo desta aplicação reside na noção de que o *Big Data* pode ser um substituto para dados ruins ou escassos; mas os modelos que mostram altas correlações entre os indicadores 'baseados em *Big Data*' e os indicadores 'tradicionais' muitas vezes exigem que a disponibilidade destes últimos seja desenvolvida. Neste contexto, 'tempo real' significa usar dados digitais de alta frequência para obter uma imagem da realidade a qualquer momento.
Empresas privadas já usam a análise de *Big Data* voltada para o desenvolvimento, o que inclui a análise em tempo real do impacto de uma ação de política - por exemplo, a introdução de novas regras de trânsito.	Embora atraentes, existem poucos (se houver) exemplos reais dessa aplicação; um dos desafios é garantir que qualquer mudança observada possa ser atribuída à intervenção ou ao "tratamento". No entanto, dados de alta frequência também podem conter "experimentos naturais" - como uma queda repentina nos preços *on-line* de um determinado bem - que podem ser alavancados para inferir causalidade.
Esta aplicação é bastante semelhante à aplicação 'Conscientização em tempo real' - embora seja menos ambiciosa em seus objetivos. Qualquer infográfico, incluindo mapas, que torne grandes quantidades de dados legíveis para o leitor é um exemplo de uma aplicação descritiva.	A descrição de dados sempre implica fazer escolhas e suposições - sobre o que e como os dados são exibidos - que precisam ser explicitadas e compreendidas; é sabido que até mesmo gráficos de barras e mapas podem ser enganosos.
Um tipo de "previsão" refere-se àquilo que pode vir a acontecer em seguida - como no caso do policiamento preditivo. Outro tipo refere-se ao uso de *proxies* de condições predominantes por meio do *Big Data* - como nos casos de níveis socioeconômicos usando CDRs na América Latina e na Costa do Marfim.	Comentários semelhantes aos feitos para os aplicativos de 'alerta antecipado' e 'conscientização em tempo real' se aplicam.
Até agora, houve poucos exemplos dessa aplicação em contextos de desenvolvimento.	A maioria dos comentários sobre '*feedback* em tempo real' se aplica a este caso. Um exemplo exigiria a atribuição de causalidade. A aplicação prescritiva funciona melhor em teoria quando suportada por sistemas e ciclos de *feedback* sobre o efeito das ações de políticas.

Big Data na prática: Projetos do Cetic.br

INDICADORES SOBRE ECONOMIA DIGITAL: USO DE *WEB SCRAPING*[5] PARA A PRODUÇÃO DE INDICADORES DE TIC PARA EMPRESAS	
TEMA	Comércio eletrônico.
PARCEIROS	O projeto se enquadra dentro de um contexto de colaboração entre a Comissão Econômica para América Latina e o Caribe (Cepal) e o Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br) para estimular e desenvolver metodologias para a medição da economia digital na América Latina e o Caribe a partir de ferramentas de *Big Data* e *data analytics*.
CONTEXTO	Conduzida anualmente pelo Cetic.br desde 2005, a pesquisa TIC Empresas mede a presença das tecnologias de informação e comunicação (TIC) em empresas com dez ou mais pessoas ocupadas. O objetivo do estudo é investigar o acesso à infraestrutura, bem como o uso e a apropriação que o setor privado faz das novas tecnologias. Dentre os diferentes temas investigados, a pesquisa aplica um módulo sobre comércio eletrônico, resultando em indicadores sobre compra e venda *on-line*.
USO DE *BIG DATA*	O projeto de *Big Data* visa produzir indicadores sobre comércio eletrônico nas empresas a partir da coleta automatizada de dados, fazendo uso do *scraping* (raspagem) de dados nas páginas Web das empresas contatadas para a pesquisa TIC Empresas. Alguns dos dados coletados são: proporção de empresas que oferecem no seu *website* catálogos de produtos, lista de preços, sistema de pedidos, pagamento *on-line*, serviço de atendimento ao consumidor (SAC); proporção de empresas que compram pela Internet; e proporção de empresas que vendem pela Internet e qual o canal utilizado para tal – *e-mail*, redes sociais, *sites* de compra em grupo. Dentre os resultados esperados, se destacam: a avaliação da precisão da modelagem de dados de fontes de *Big Data* para estimar alguns indicadores de TIC nas empresas; o desenvolvimento de uma ferramenta para coleta automatizada de dados na Web; e o desenvolvimento de um dicionário de palavras-chave (contexto semântico).
STATUS	Projeto piloto em andamento.

5 *Web scraping* é uma técnica de extração de dados utilizada para coletar dados/conteúdos de *websites*. Por meio de processos automatizados, esse tipo de "raspagem" de informações é uma forma de realizar cópias de dados de *websites*, convertendo-os em informação estruturada para posterior análise.

INDICADORES DE BANDA LARGA – BASES DE DADOS SISTEMA DE MEDIÇÃO DE TRÁFEGO INTERNET (SIMET)	
TEMA	Qualidade da banda larga.
PARCEIROS	O projeto decorre de uma parceria entre o Centro de Estudos e Pesquisas em Tecnologia de Redes e Operações (Ceptro.br), do Núcleo de Informação e Coordenação do Ponto BR (NIC.br), responsável pela coleta dos dados por meio da ferramenta Simet[6], e o Cetic.br, responsável pela análise dos dados.
CONTEXTO	Diferentes fontes de dados podem ser utilizadas para medir a qualidade da Internet banda larga, desde registros administrativos até a percepção do usuário sobre a qualidade dos serviços prestados pelas operadoras. O Cetic.br produz dados sobre a conexão da Internet a partir da autodeclaração dos respondentes de suas pesquisas amostrais. No entanto, com vistas a uma avaliação mais completa sobre o tema, buscou-se a análise de outras fontes de dados sobre a qualidade da banda larga.
USO DE *BIG DATA*	Há dez anos o Simet coleta dados sobre a qualidade da Internet brasileira de forma periódica, em tempo real, a partir de testes realizados pelos usuários. As medições, feitas em um intervalo de 11 segundos, ocorrem pelo envio e recebimento de pacotes de informação de forma a produzir diversos indicadores de qualidade quanto à operação em curso. A base de dados resultante de tais medições possui características de *Big Data*, pois se trata de um grande volume de dados, com atualização muito rápida (novas medidas a cada instante), gerada em eventos definidos pelo próprio usuário, isto é, informações que são coletadas se o usuário desejar, e não a partir de seleção de amostra. A partir da análise dos dados dessa base, e considerando as suas peculiaridades e limitações, torna-se possível complementar os dados coletados por meio das pesquisas amostrais. Assim, é possível obter uma visão mais ampla da evolução da banda larga no Brasil.
STATUS	Projeto em andamento.

6 O Sistema de Medição de Tráfego Internet (Simet) é um conjunto de sistemas que testa a qualidade da Internet. Ele realiza medições independentes, de forma automatizada ou manualmente, acionadas pelo usuário, integralmente apoiado na infraestrutura do NIC.br. Atualmente, essas medições são tomadas por meio de três diferentes dispositivos/aplicativos: Simet Box, Simet Mobile e Simet Applet. Para mais informações, acesse: https://simet.nic.br/index.php

Entrevista

Roberto Olinto
é Presidente do Instituto Brasileiro de Geografia e Estatísticas (IBGE)

Big Data e estatísticas oficiais: desafios e oportunidades

Roberto Olinto, Presidente do Instituto Brasileiro de Geografia e Estatística (IBGE), comenta as potenciais oportunidades e os desafios que as mudanças tecnológicas, especialmente o *Big Data*, apresentam para a produção de estatísticas oficiais.

P.S_ Quais ações o IBGE está implementando em relação ao [uso de] Big Data?

R.O_ Antes de implementar ações, um instituto de estatística oficial tem que colocar em sua agenda a reflexão sobre o impacto do *Big Data* no instituto, ou seja, o que as informações geradas pelo *Big Data*, que não é um fenômeno muito claro, causam. De que maneira um instituto de estatística se posicionaria em relação a essa quantidade de informações? Nossa posição no IBGE é voltada para a área de comunicação, que é como nos apresentarmos de forma cada vez mais abrangente, utilizando diferentes mídias, mas também sendo cada vez mais transparentes, explicando o que nós produzimos, dado a quantidade de informações que existem hoje e que podem se contrapor ao nosso trabalho.
No IBGE, nós criamos uma comissão interna. Tentou-se criar um representante de todas as áreas para discutir exatamente como o IBGE, internamente, se posicionaria em relação ao *Big Data*. Isso vai definir uma política maior, inclusive se incorporando as discussões internacionais sobre o uso de *Big Data* para informação oficial, que é um grupo de trabalho que existe. Mas já existem iniciativas em determinadas áreas do IBGE para lidar com o *Big Data*.
A primeira questão fundamental é: de que maneira essa tecnologia vai nos permitir levantar dados mais rápido, com mais eficiência e, fundamentalmente, de acordo com a metodologia que nós adotamos, e não uma busca caótica de dados. Em seguida, teremos que refletir se a geração de dados, de alguma forma, vai nos ajudar e de que forma vai ajudar. Mas esse é um passo posterior, porque, inicialmente, temos que olhar a quantidade de informações que está sendo gerada e de que forma ela pode impactar o instituto de estatística, no sentido de comunicação, e, em segundo lugar, a implementação de uso dessas ferramentas, métodos de busca e outras coisas desse tipo, para melhorar a nossa captação de informação.

P.S._ Quais são as principais oportunidades e vantagens trazidas pelo [uso de] Big Data para as estatísticas oficias?

R.O_ O que nós encaramos é, fundamentalmente, a velocidade na obtenção de informação. Um outro ponto, mais complicado e que tem sido discutido internamente, é se nós poderíamos usar bases de dados privadas para a realização de estatísticas. O maior exemplo é a questão da telefonia celular: se pudessemos ter acesso à base dos telefones celulares, obviamente, sem identificação [do usuário], pois não precisaríamos ter a identificação do informante, o que também não é um problema,

porque somos protegidos pelo sigilo, e, com isso, desenvolveríamos estatísticas de mobilidade, de fluxo, influência das cidades e assim por diante.
A grande questão é que essas informações não são cedidas gratuitamente. E o instituto de estatística não tem dinheiro, ou melhor, o orçamento para isso não é ilimitado. E também não se pode ficar refém, em algum momento, do produtor da base de dados. Essa questão, particularmente, vai ter que ser discutida – e alguns países já estão enfrentando isso, se é possível ter uma legislação que avance nessa ideia de compartilhamento de base de dados.
Existem experiências internacionais de uso da base de dados de grandes supermercados para se ter um acompanhamento de preços. Agora, o *Big Data* pode ser extremante útil exatamente na ideia de captação e operação da informação, que é a nota fiscal eletrônica no Brasil.
A nota fiscal eletrônica tem uma quantidade enorme de informações que podem ser claramente trabalhadas com os novos processos, sejam novos sistemas ou nova capacidade de armazenamento, que é uma das características do *Big Data*. Isso é uma informação, hoje, pública, que nós estamos discutindo com as secretarias de Fazenda. Nos estados, vamos discutir no Confaz [Conselho Nacional de Política Fazendária] e seria, na minha opinião, a primeira e grande operação estatística usando *Big Data*: usar toda a nota fiscal eletrônica para melhorar as estatísticas do IBGE.

P.S_ Quais são os principais desafios a serem enfrentados para poder aproveitar os potenciais benefícios do [uso de] Big Data?

R.O_ Desafio é ter acesso à tecnologia, ter orçamento para se atualizar. O desafio é ter uma equipe técnica em número suficiente, dado que no IBGE as equipes são concursadas – e, para se ter um perfil em que você introduza a figura do analista de dados, é [preciso] capacitar a própria equipe. Isso é um desafio, dada a velocidade com que a tecnologia avança.
O desafio, como disse anteriormente, é ter acesso à base de dados privada, de forma permanente; é você criar bases de dados públicas, por exemplo, a da nota fiscal eletrônica, ou seja, ter uma legislação que olha a massa de dados gerada hoje, através de registros administrativos, e poder operá-la com ferramentas de *Big Data* e ter acesso à informação. Talvez um desafio complementar seja mudar a cultura dentro do instituto de estatística, introduzindo a lógica de *Big Data*.

P.S_ Como o IBGE se posiciona diante de um cenário de integração entre os diversos produtores de dados estatísticos?

R.O_ O IBGE é forte e profundamente defensor dessa ideia. Hoje, a questão da interoperabilidade, do compartilhamento de bancos de dados, é a chave para o avanço do sistema estatístico. A integração dos produtores de informação, a integração das bases de dados, e, obviamente, da transparência das metodologias, é o caminho para se avançar no sistema de informação de um país. Nós defendemos [essa ideia] e atuamos para que isso ocorra o mais rápido possível. Obviamente, os desafios técnicos são fáceis de responder. Agora, os desafios institucionais são o nosso maior desafio. O maior problema é exatamente você ter uma questão de legislação, de cultura, para que a lógica do compartilhamento de bases de dados seja aceita sem maiores problemas, ou como um passo natural na evolução do sistema de informação, ou seja, no sistema de estatística e geoinformação.

"A primeira questão fundamental é: de que maneira essa tecnologia vai nos permitir levantar dados mais rápido, com mais eficiência e, fundamentalmente, de acordo com a metodologia que nós adotamos, e não uma busca caótica de dados".

P.S_ Como o uso de Big Data pode ajudar os institutos nacionais de estatística na medição dos Objetivos de Desenvolvimento Sustentável (ODS)?

R.O_ Ele vai ajudar em todas as informações, dado que os ODS hoje são mais uma demanda e nos exigem aperfeiçoamento e ampliação da nossa produção estatística. Quer dizer, os instrumentos de *Big Data* são chave para atender não só aos ODS, mas todas as novas demandas, que não são restritas aos ODS, mas que vem de uma série de outras fontes.

Entrevista II

Ronald Jansen é Diretor Assistente e Chefe do Departamento de Inovação de Dados e Desenvolvimento de Capacidade da Divisão de Estatísticas das Nações Unidas (UNSD)

Comunidades e *Big Data*: o papel dos *data collaboratives*

Para comentar o papel dos modelos de colaboração multisetoriais que vêm sendo desenvolvidos para o uso do *Big Data*, entrevistamos Ronald Jansen, diretor assistente e chefe do Departamento de Inovação de Dados e Desenvolvimento de Capacidade da Divisão de Estatísticas das Nações Unidas (UNSD).

P.S_ O que são os data collaboratives e por que são importantes no uso do Big Data para as estatísticas oficiais?

R.J_ Os *data collaboratives*, dentro do contexto da comunidade estatística global, são projetos nos quais os institutos nacionais de estatística, o setor privado e outros parceiros trabalham juntos, em um ambiente seguro, com *Big Data* e outras fontes de dados, na produção de estatísticas e indicadores passíveis de serem usados para políticas públicas.

Um exemplo de *data collaborative* seria a colaboração de vários institutos de estatística, da Divisão de Estatísticas das Nações Unidas (UNSD) e da Nielsen, uma empresa privada especializada na coleta de dados de preços, a fim de testar novos métodos estatísticos por meio de dados de *scanner* (*scanner data*), fornecidos pela Nielsen, para o cálculo de índices de preços, principalmente o índice de preços ao consumidor. Nesse *data collaborative*, nem todos os parceiros têm os mesmos direitos de acesso a todos os dados, mas todos terão acesso aos algoritmos e códigos desenvolvidos nos métodos testados.

O objetivo do *data collaborative* é desenvolver métodos e aplicativos com dados e parceiros confiáveis, cujos resultados podem ser transferidos para a comunidade em geral por meio de um aprendizado seguro. Em outras palavras, o *data collaborative* resulta em métodos testados, novas ferramentas, algoritmos e aplicações, que são disponibilizados para toda a comunidade estatística.

O objetivo mais amplo de tal colaboração no uso de *Big Data* é permanecer relevante como uma comunidade estatística. Os institutos de estatística precisam inovar e acompanhar as mudanças da tecnologia, das fontes de dados disponíveis e das expectativas da sociedade. Hoje em dia, o público espera contar com informações de boa qualidade de modo rápido e detalhado. As empresas estão produzindo uma grande quantidade de dados muito rapidamente, e os institutos de estatística precisam competir para continuar sendo relevantes.

P.S_ Como os institutos nacionais de estatística, as empresas de tecnologia e os proprietários de dados podem colaborar de forma mutuamente benéfica em um mundo em transformação, no qual os dados são vistos como a fonte mais importante para criar riqueza e desenvolvimento para todos?

R.J_ Eu dei um exemplo sobre o uso de *scanner data*. Outro exemplo de *data collaborative* é o uso de dados de satélites para a medição da extensão de água doce (indicador 6.6.1 dos Objetivos do Desenvolvimento Sustentável – ODS)[7]. Nesse caso, vários institutos de estatística se uniram à ONU Meio Ambiente, à UNSD e à Google. A Google trouxe para o projeto o uso do mecanismo Google Earth, bem como o espaço no servidor em nuvem e sua capacidade computacional. Os estatísticos trouxeram conhecimento metodológico para desenvolver e testar algoritmos. Em princípio, o resultado desse *data collaborative* poderia ajudar a cada país com o cálculo do indicador 6.6.1 dos ODS.

P.S_ Quais são as experiências e lições aprendidas a partir dos data collaboratives existentes em relação à cobertura, inclusão (e exclusão) de parceiros, atividades, gerenciamento e financiamento?

R.J_ Os dois *data collaboratives* que mencionei envolvem principalmente institutos de países desenvolvidos. Enquanto a governança geral recai sobre o UN Global Working Group (GWG) on *Big Data* (Grupo de Trabalho Global da ONU sobre *Big Data*), os parceiros participantes do projeto irão determinar a gestão direta dos recursos. Por ora, as implicações financeiras foram reduzidas ao mínimo.
O GWG definiu como meta desenvolver uma validação do conceito (*proof of concept*) até março de 2020. Essa validação do conceito do trabalho em *data collaboratives* junto com o setor privado deve incluir um modelo de negócios viável e sustentável (ou vários modelos de negócios), que poderia estar trabalhando sem fins lucrativos, ou numa base comercial com créditos e descontos para aqueles que podem contribuir menos. Por conseguinte, deve também incluir evidências de que os países em desenvolvimento podem participar e beneficiar-se das novas fontes de dados, métodos, ferramentas e aplicações.

P.S_ O que é o Cape Town Global Action Plan for Sustainable Development Data e como ele se relaciona com a chamada "revolução de dados"?

R.J_ A "revolução de dados" foi uma chamada à ação para energizar a comunidade estatística (e a comunidade de dados em um sentido mais amplo) de modo a inovar e usar todos os dados que estavam sendo gerados automaticamente no mundo digitalizado.
O Cape Town Global Action Plan é uma chamada mais abrangente para fortalecer os institutos nacionais de estatística. O uso de novas fontes de dados é apenas uma parte disso. Alguns dos pontos do plano de ação são: modernizar a governança, estruturas institucionais e padrões estatísticos; facilitar a aplicação de novas tecnologias e fontes de dados nas atividades estatísticas gerais; expandir o uso de registros administrativos; e fortalecer as parcerias com governos, academia, sociedade civil, setor privado e outras partes interessadas envolvidas na produção e no uso de dados para o desenvolvimento sustentável.

"Hoje em dia, o público espera contar com informações de boa qualidade de modo rápido e detalhado. As empresas estão produzindo uma grande quantidade de dados muito rapidamente, e os institutos de estatística precisam competir para continuar sendo relevantes".

7 O indicador 6.6.1. corresponde à "mudança na extensão dos ecossistemas relacionados à água ao longo do tempo" e está inserido dentro do ODS 6 de "assegurar a disponibilidade e gestão sustentável da água e saneamento para todos".

Relatório de Domínios

A dinâmica dos registros de domínios no Brasil e no mundo

O Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br) monitora mensalmente a quantidade de nomes de domínios registrados entre os 16 maiores domínios de topo de código de país (do inglês, *country code top-level domain* – ccTLD) no mundo. Somados, eles ultrapassam 97,6 milhões de nomes de domínios registrados. Em maio de 2018, os domínios registrados sob o .tk (arquipélago de Tokelau, no Oceano Pacífico) chegaram a 20,71 milhões. Em seguida, aparecem Alemanha (.de), China (.cn) e Reino Unido (.uk) com, respectivamente, 16,31 milhões, 11,06 milhões e 9,93 milhões de registros.[8] O Brasil continua ocupando a sétima posição no ranking, com 3,95 milhões de registros sob o .br. Na 16ª posição, com 1,91 milhão de registros, está a Espanha (.es), como observado na Tabela 2.

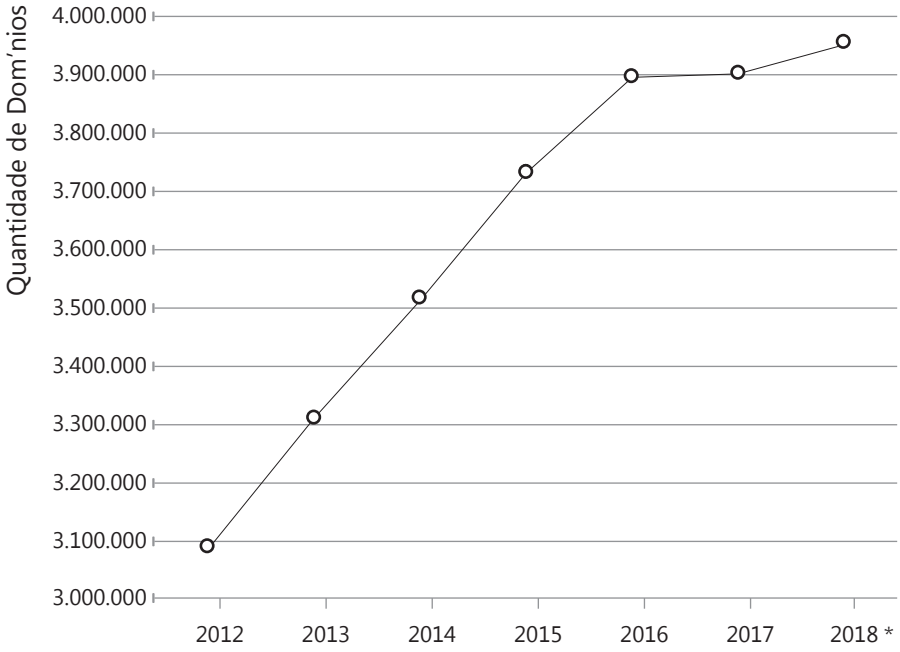
Tabela 2 – REGISTRO DE NOMES DE DOMÍNIOS NO MUNDO – MAIO/2018

Posição	ccTLD	Domínios	Ref.	Fonte
1	Tokelau (.tk)	20.715.276	mai/18	research.domaintools.com/statistics/tld-counts
2	Alemanha (.de)	16.308.345	mai/18	www.denic.de
3	China (.cn)	11.059.726	mai/18	research.domaintools.com/statistics/tld-counts
4	Reino Unido (.uk)	9.926.573	jan/18	www.nominet.uk/uk-register-statistics-2018/
5	Países Baixos (.nl)	5.806.100	mai/18	www.sidn.nl
6	Rússia (.ru)	5.239.190	dez/17	cctld.ru
7	**Brasil (.br)**	**3.952.679**	**mai/18**	**registro.br/estatisticas.html**
8	União Europeia (.eu)	3.717.165	mai/18	research.domaintools.com/statistics/tld-counts
9	França (.fr)	3.212.900	mai/18	www.afnic.fr/en/resources/statistics/detailed-data-on-domain-names
10	Austrália (.au)	3.152.285	mai/18	www.auda.org.au
11	Itália (.it)	3.133.248	mai/18	www.nic.it
12	Canadá (.ca)	2.748.296	mai/18	www.cira.ca
13	Polônia (.pl)	2.564.636	mai/18	www.dns.pl/english/zonestats.html
14	Suíça (.ch)	2.147.841	mar/18	www.nic.ch/reg/cm/wcm-page/statistics/index.html?lid=em*
15	Estados Unidos (.us)	1.966.460	mai/18	research.domaintools.com/statistics/tld-counts/
16	Espanha (.es)	1.911.574	mai/18	dominios.es

8 É importante destacar que o período de referência de cada ccTLD não é o mesmo em todos os casos, embora seja o mais atualizado. Conforme mostra a Tabela 1, os dados de Reino Unido e Suíça correspondem aos meses de janeiro e março de 2018, respectivamente.

O Gráfico 1 apresenta o desempenho do .br desde o ano de 2012.

Gráfico 1 – TOTAL DE REGISTROS DE DOMÍNIOS AO ANO DO .BR – (2012 – 2018)



**Dado referente ao mês de maio de 2018.*
Fonte: Registro.br

No mês de maio de 2018, os cinco principais domínios genéricos (do inglês, *generic top-level domain* – gTLD) totalizavam mais de 167 milhões de registros. O .com se destaca com 134,41 milhões de registros, conforme se observa na Tabela 3.

Tabela 3 – PRINCIPAIS GTLDS – MAIO/2018

Posição	gTLD	Domínios
1	.com	134.407.747
2	.net	14.268.162
3	.org	10.396.758
4	.info	6.105.279
5	.biz	1.983.870

Fonte: http://research.domaintools.com/statistics/tld-counts/
Acesso em: 04/05/2018

BIG DATA? "REVOLUÇÃO DE DADOS?" O QUE É O QUÊ?

Veja a seguir as definições dos principais termos e conceitos[9]:

Big Data

termo guarda-chuva que abriga três tendências: o crescente volume de dados digitais gerados como subproduto do uso diário de dispositivos digitais por pessoas; as novas tecnologias, ferramentas e métodos disponíveis para analisar grandes conjuntos de dados que não foram originalmente planejados para análise; e a intenção de extrair desses dados e ferramentas *insights* que possam ser usados para a elaboração de políticas.

"Revolução de Dados"

expressão comum presente no discurso do desenvolvimento, desde que, durante as discussões da Agenda de Desenvolvimento Pós-2015, participantes eminentes do painel de alto nível chamaram uma "revolução de dados" para "fortalecer dados e estatísticas para fins de *accountability* e tomada de decisões". Refere-se a um fenômeno maior do que o *Big Data*, também sendo denominado de "revolução de dados sociais", definido como a mudança nos padrões de comunicação humana em direção a um compartilhamento maior de informações pessoais.

Algoritmos

na ciência da computação, um algoritmo é uma série de instruções ou regras predefinidas, escritas em uma linguagem de programação destinada para dizer a um computador como resolver sequencialmente um problema recorrente, especialmente ao envolver a realização de cálculos e o processamento de dados. Há um crescente uso de algoritmos para fins de tomada de decisão em uma progressiva gama de atividades e indústrias, tais como policiamento e serviços bancários.

9 Extraído e adaptado de Letouzé, E. (2015). Big Data and development: General overview primer". Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative. Recuperado em 10 maio, 2018, de http://www.un.org/en/development/desa/population/commission/pdf/49/E_CN_9_2016_3_Supplement.pdf

Aprendizado estatístico de máquina

refere-se à construção e estudo de algoritmos computacionais – procedimentos, passo a passo, para cálculos e/ou classificação – que podem aprender a crescer e mudar quando expostos a novos dados. Representa a capacidade das máquinas de "aprender" a fazer previsões e tomar decisões melhores com base em experiências passadas, como na filtragem de *spam*, por exemplo. A inclusão da palavra "estatístico" reflete a ênfase na análise e na metodologia estatística, que é a abordagem predominante para o aprendizado de máquina atualmente.

Registro de detalhes de chamada

(do inglês, *call detail records* – CDR):

nome técnico para os dados de telefones celulares registrados pelas operadoras de telecomunicação para todas as chamadas realizadas. Os CDR contêm informação relativa à hora, duração e à localização dos remetentes e receptores das chamadas ou mensagens de texto transmitidas por meio de suas redes.

(Novo)

Hiato digital

diferença de acesso e capacidade para usar as novas tecnologias de informação e comunicação (TIC) entre indivíduos, comunidades e países, que revela a consequente desigualdade de oportunidades e resultados socioeconômicos e políticos. As habilidades e ferramentas necessárias para absorver e analisar grandes quantidades de dados resultantes do crescente uso de tecnologia podem fundamentar um "novo hiato digital".

Créditos

REDAÇÃO
ARTIGO PRINCIPAL
Emmanuel Letouzé (Data-Pop Alliance)
RELATÓRIO DE DOMÍNIOS
José Márcio Martins Júnior (Cetic.br)

COORDENAÇÃO EDITORIAL
Alexandre Barbosa (Cetic.br)
Tatiana Jereissati (Cetic.br)

AGRADECIMENTOS
Emmanuel Letouzé (Data-Pop Alliance)
Roberto Olinto (Instituto Brasileiro de Geografia e Estatística – IBGE)
Ronald Jansen (Divisão de Estatísticas das Nações Unidas – UNSD)

PROJETO GRÁFICO E DIAGRAMAÇÃO
Comunicação NIC.br

TRADUÇÃO DO ARTIGO PRINCIPAL
Ana Zuleika Pinheiro Machado

REVISÃO EM PORTUGUÊS
Aloisio Milani (Magma Editorial Ltda.)
Alexandre Pavan (Magma Editorial Ltda.)

Organização das Nações Unidas para a Educação, a Ciência e a Cultura

cetic.br • Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação sob os auspícios da UNESCO

nic.br Núcleo de Informação e Coordenação do Ponto BR

cgi.br Comitê Gestor da Internet no Brasil

POR UMA INTERNET CADA VEZ MELHOR NO BRASIL

CGI.BR, MODELO DE GOVERNANÇA MULTISSETORIAL

www.cgi.br

nic.br cgi.br

FOTOS: RICARDO MATSUKAWA E SHUTTERSTOCK