Panorama Setorial da Internet

Número 1
Março, 2023
Ano 15

Ecossistema e produção de dados

Governança de dados e a Datasfera: revisão da literatura[1]

Por Datasphere Initiative[2]

Introdução

Nos últimos anos, o termo "governança de dados" tem atraído cada vez mais atenção. Deixou de ser um tópico de nicho, tratado exclusivamente como um aspecto técnico dos projetos de compartilhamento de dados ou um complemento para a gestão de dados nas áreas de tecnologias de informação e comunicação (TIC) das empresas, tornando-se ponto central para tratar de questões como captura, acesso, uso e proteção de dados de forma a gerar maior inclusão social e econômica.

Desta forma, a governança de dados torna-se um campo mais complexo, emergindo como uma estrutura-chave para abordar as oportunidades e os riscos da coleta, do compartilhamento e do uso de dados. Isto reflete um reconhecimento crescente da importância dos dados dentro de processos mais amplos de governança, assim como o potencial poder que os dados têm, seja enquanto um recurso para o progresso econômico e social, seja enquanto um catalisador de danos, quando mal utilizados.

Entretanto, a convergência relativamente rápida do interesse de formuladores de políticas, tecnólogos, ativistas e profissionais na "governança de dados" traz alguns desafios. Diferentes agendas, vocabulários, preocupações e áreas de ênfase colidem, e não há ainda um campo coerente de pesquisa sobre a governança de dados, considerando essa visão complexa.

Ao fornecer um mapeamento inicial de quem está escrevendo sobre governança de dados e os tipos de temas que estão sendo abordados, este artigo oferece uma base para responder ao chamado dos autores de La Chapelle e Porciuncula (2021, p. 3) por um trabalho sobre governança de dados capaz de "reestruturar a discussão, reunir as abordagens inovadoras emergentes e engajar-se em um debate global, intersetorial e com as múltiplas partes interessadas, que se mostra muito necessário".

Para apoiar tal reformulação, este artigo também analisa o marco referencial emergente da Datasfera, entendida como "o sistema complexo que abrange todos os tipos de dados e suas interações dinâmicas com normas e grupos humanos" (de La Chapelle & Porciuncula, 2022, p. 3). A mudança conceitual que isso introduz convida a deslocar a discussão de noções relativamente planas de "governança de dados" para a "governança da Datasfera": colocando em foco a interação de conjuntos de dados, normas e grupos humanos.

---

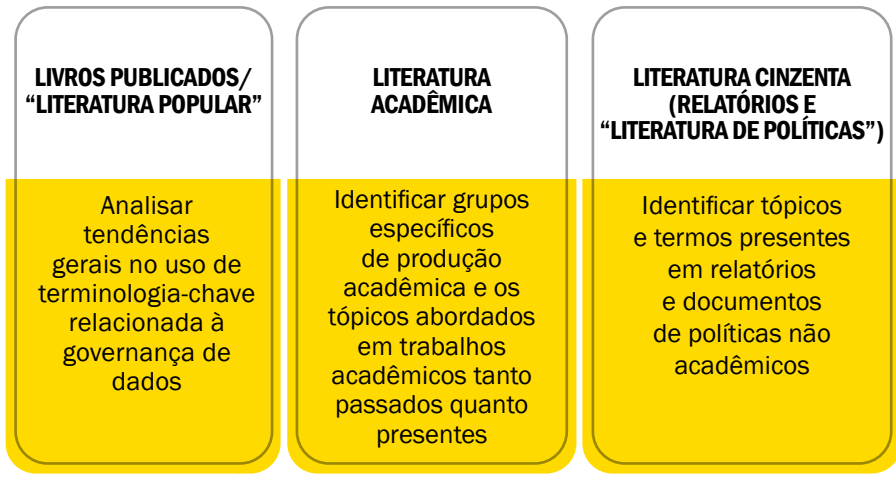
[1] Versão editada do trabalho homônimo publicado pela Datasphere Initiative (DI). Disponível em: https://www.thedatasphere.org/datasphere-publish/data-governance-and-the-datasphere/

[2] Este relatório foi elaborado para a DI por Tim Davies como resultado de uma consultoria e bolsa de estudos, contando com a orientação e a contribuição de Carolina Rossini, diretora de Pesquisa e Parcerias da DI. Tim Davies é diretor de Pesquisa da organização sem fins lucrativos Connected by Data, sediada no Reino Unido. É mestre em Ciências Sociais da Internet pelo Oxford Internet Institute, da Universidade de Oxford (Reino Unido), e foi membro do Berkman Klein Center for Internet and Society, da Universidade Harvard.

Metodologia

A presente revisão bibliográfica sobrepõe várias estratégias para fornecer uma visão geral das produções escritas sobre governança de dados. Embora a análise a seguir baseie-se principalmente na literatura acadêmica, são utilizados livros publicados – via o *corpus* do Google Books – e literatura cinzenta (*gray literature*) – via um *corpus* baseado no *Datasphere Governance Atlas* (Datasphere Initiative, 2022) – para oferecer *insights* complementares.

Figura 1 – TIPO DE MATERIAL REVISADO E PROPÓSITO DA REVISÃO

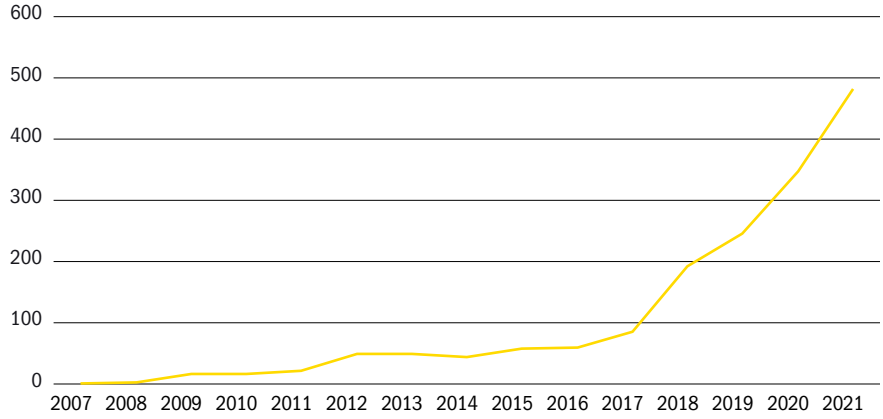


Fonte: Elaboração própria.

A governança de dados é um campo em crescimento que reúne várias áreas do conhecimento

A governança de dados reúne várias áreas do conhecimento até então tratadas separadamente. Dada a proliferação de bibliografia sobre governança de dados – o recente *Datasphere Governance Atlas* (Datasphere Initiative, 2022) contabiliza nada menos que 261 organizações focadas em alguma medida em tópicos de governança de dados –, pode ser surpreendente observar que o termo "governança de dados" entrou no léxico de pesquisas e políticas em escala somente na última década. O uso do termo em títulos e resumos de trabalhos acadêmicos, por exemplo, aumentou quase cinco vezes entre 2015 e 2021, e há indicações de que tenha crescido ainda mais em 2022.

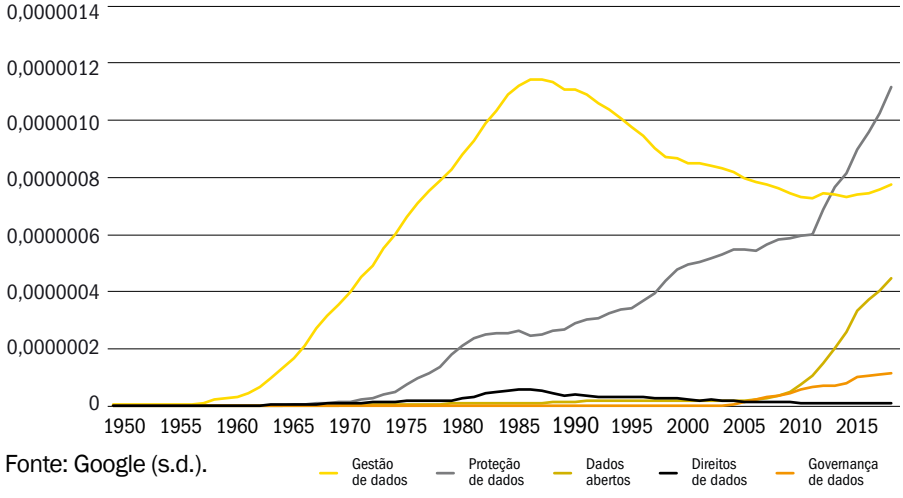
Gráfico 1 – NÚMERO DE PUBLICAÇÕES REGISTRADAS NO CONJUNTO DE DADOS DE DIMENSÕES QUE INCLUEM O TERMO "GOVERNANÇA DE DADOS" EM SEU TÍTULO, RESUMO OU PALAVRAS-CHAVE, POR ANO DE PUBLICAÇÃO (2007-2021)



Fonte: Elaboração própria.

O rápido desenvolvimento do discurso sobre governança de dados não significa que os debates existentes (antes em silos de conhecimento) tenham sido inteiramente subsumidos em um único campo – da governança de dados. Uma análise a respeito da presença de outros termos na literatura popular destaca que há muito mais probabilidade de os leitores encontrarem trabalhos sobre "proteção de dados" ou "gestão de dados" em livros e manuais técnicos do que discussões sobre governança de dados. Mesmo tópicos como dados abertos – que, pode-se argumentar, é apenas uma abordagem particular sobre governança de dados – receberam nos últimos anos uma atenção significativamente mais direta do que a governança de dados.

Gráfico 2 – COMPARATIVO DE MENÇÕES AOS TERMOS (EM INGLÊS) "PROTEÇÃO DE DADOS", "GESTÃO DE DADOS", "DADOS ABERTOS", "GOVERNANÇA DE DADOS" E "DIREITOS DE DADOS" NO *CORPUS* DO GOOGLE BOOKS NGRAM VIEWER (1950-2019)



Fonte: Google (s.d.).

(...) o volume de documentos sobre governança de dados na literatura acadêmica tem apresentado um crescimento percentual anual maior do que aqueles centrados na proteção de dados (...).

Padrões de evolução na literatura popular são amplamente espelhados na produção científica, em que ainda há a cada ano muito mais trabalhos publicados a respeito de proteção de dados, gestão de dados ou dados abertos do que usando explicitamente o termo "governança de dados" em seu título ou resumo. No entanto, nos últimos anos, o volume de documentos sobre governança de dados na literatura acadêmica tem apresentado um crescimento percentual anual maior do que aqueles centrados seja na proteção de dados, seja na gestão de dados.

Entretanto, este aumento pequeno mas contínuo do emprego do termo "governança de dados" tanto em pesquisas quanto em políticas tende a diminuir em função da posterior especialização do conhecimento. Isto porque autores estão discutindo "governança de dados" com base em uma gama de pontos de partida, e isto influenciará o que se enquadra no escopo de suas definições e prescrições do tema em questão. Por exemplo, conforme mencionado, grande parte da literatura no campo da computação e da gestão considera a governança de dados somente dentro dos limites de uma empresa, enquanto os estudos nos campos das ciências sociais e a literatura cinzenta com frequência exploram a governança de dados como uma questão social. Ao mesmo tempo, ainda que a especialização do conhecimento e consequente produção escrita ocorra, a produção em andamento de trabalhos estruturados em termos de proteção de dados, privacidade, gestão de dados e dados abertos (para citar apenas algumas áreas) pode ter muito a contribuir para o desenvolvimento de normas, políticas e práticas de governança de dados mais complexas, e por sua vez, da Datasfera, mesmo que não adotem diretamente uma linguagem de governança de dados.

Revisões anteriores da literatura revelam a diversidade da área

Muitos dos temas que cada vez mais se inserem no amplo quadro da governança de dados foram discutidos anteriormente em termos de proteção de dados (Greenleaf, 2012), gestão de dados (Ladley, 2019; Panian, 2009) ou dados abertos (Davies *et al*., 2019; Verhulst *et al*., 2020), cada um com suas próprias agendas em torno de privacidade, exploração de ativos de dados empresariais e reutilização pública de dados, respectivamente. Uma mudança no enquadramento desses tópicos dentro do invólucro mais abrangente da governança de dados responde ao reconhecimento da complexidade e das negociações envolvidas na decisão de quando e como os dados devem ser coletados, estruturados, compartilhados, transferidos, usados e excluídos.

Os esforços para resolver ou reestruturar essas negociações e tensões também deram origem a uma série de novas agendas em relação ao compartilhamento de dados (Micheli *et al*., 2020) e a novos modelos de propriedade e administração de dados (Delacroix & Lawrence, 2019; Lehtiniemi & Haapoja, 2020; Susha *et al*., 2017), que se enquadram no campo em expansão da governança de dados. Na literatura cinzenta sobre o tema, um forte elemento normativo é cada vez mais evidente: a vinculação desse termo a agendas mais amplas de boa governança e desenvolvimento global. Conforme colocado por Pisa *et al*. (2020, p. 2), o ideal de governança de dados incorpora "regras a respeito de como os dados são coletados, analisados, usados e compartilhados de forma a proteger os cidadãos contra abusos, ao mesmo tempo que apoia a inovação, o desenvolvimento e o crescimento inclusivo".

Uma análise de oito revisões da literatura sobre o tema da governança de dados, publicadas entre 2016 e início de 2022, mostra essa mudança de ênfase. Enquanto os trabalhos iniciais se centravam na governança de dados principalmente em termos de gestão de dados e informações (Alhassan *et al*., 2016; Brous *et al*., 2016), as produções têm abordado cada vez mais a governança de dados como uma questão pública mais ampla, demandando ênfase no compartilhamento de dados entre organizações (Abraham *et al*., 2019; Benfeldt Nielsen, 2017) e em dados abertos (Bozkurt *et al*., 2022). Ainda assim, McCaig e Rezania (2021, p. 5) argumentam que, em última análise, a literatura permanece "indicativa de uma escassa base de conhecimentos teóricos e empíricos" sobre governança de dados.

Uma definição operacional ampla de governança de dados coloca em primeiro plano tanto os benefícios quanto os danos

Dada a abrangência dos contextos em que a governança de dados deve ser aplicada, não é razoável esperar uma única definição capaz de agrupar um só campo de estudo. Entretanto, aspectos comuns da governança de dados ainda podem ser extraídos. Para os fins deste artigo, é oferecida a seguinte definição operacional, formada por dois elementos:

- A governança de dados diz respeito às regras, processos e comportamentos relacionados à coleta, gestão, análise, uso, compartilhamento e descarte de dados – pessoais e/ou não pessoais.
- A boa governança de dados deve tanto promover benefícios quanto minimizar danos em cada estágio dos ciclos relevantes de dados.

Em nível organizacional, isso geralmente se traduz em um foco nas políticas internas e em sua implementação; na conformidade com a regulamentação externa; e na criação de marcos referenciais e responsabilidades interfuncionais para gerir e extrair valor dos dados como um ativo comercial (Abraham *et al*., 2019). Em nível estatal – seja nacional, regional ou internacional –, isso pode se traduzir em um foco no desenvolvimento e na implementação de políticas, normas, leis, regulamentações, acordos e práticas que cobrem a gestão de dados nos países, bem como a transferência de dados entre as fronteiras jurisdicionais (Aaronson, 2021). Entretanto, com frequência a literatura organizacional dedica pouca atenção ao nível estatal e vice-versa.

Diversos autores também destacam que a governança de dados faz parte de um conjunto mais amplo de preocupações práticas e de governança. Wendehorst (2020) descreve a governança de dados como um entre vários marcos referenciais sobrepostos preocupados com a governança em relação à Inteligência Artificial (IA), considerando, por exemplo, de que modo a mesma questão pode ser explorada pela lente da governança de dados (pensando como os conjuntos de dados são criados, gerenciados e usados); pela lente do desenho de sistemas de IA (usando a linguagem de viés ou a adequação de métodos); ou com um foco na governança social mais abrangente (fazendo

A governança de dados diz respeito às regras, processos e comportamentos relacionados à coleta, gestão, análise, uso, compartilhamento e descarte de dados – pessoais e/ou não pessoais.

Isso sublinha a importância de resistir à tendência de tratar os dados de maneira inteiramente generalizada: dados significativos são sempre sobre algo, e com frequência esse "algo" também está sujeito aos seus próprios regimes de governança (...).

perguntas sobre os objetivos e a governança das áreas políticas mais amplas com as quais os conjuntos de dados e os sistemas de IA se relacionam).

Isso sublinha a importância de resistir à tendência de tratar os dados de maneira inteiramente generalizada: dados significativos são sempre sobre algo, e com frequência esse "algo" também está sujeito aos seus próprios regimes de governança, com os quais qualquer governança prática de dados se cruzará. Muitos pesquisadores chegaram ao tema da governança de dados por causa de desafios profundamente fundamentados em torno da proteção, gestão ou compartilhamento de dados em relação a um campo específico de ação.

Ao se voltar à literatura acadêmica, é importante compreender até que ponto diferentes projetos e trabalhos são parte de uma agenda de pesquisa coerente ou – ao contrário – até que ponto cada publicação que utiliza a linguagem de governança de dados pode ter se desenvolvido de forma isolada de outros trabalhos relativos ao tema.

O conceito de governança de dados não só reúne acadêmicos que antes trabalhavam em questões distintas de proteção, gestão e acesso de dados, mas também tem sido invocado em campos acadêmicos díspares, da pesquisa em saúde a trabalhos em comércio internacional. Nesses espaços, a governança de dados ainda pode parecer mais ou menos como uma subárea de nicho, em vez de um campo de investigação transversal por direito próprio.

O desenvolvimento de narrativas da Datasfera pode oferecer uma perspectiva holística para trabalhos futuros sobre governança de dados

Esta seção fornece um breve panorama do conceito de Datasfera e explora o que pode significar olhar para a literatura de governança de dados pela lente da Datasfera. O relatório We Need To Talk About Data (de La Chapelle & Porciuncula, 2021) parte de um trabalho de Bergé *et al*. (2018, p. 2) que trouxe uma descrição conceitualmente expansiva mas digitalmente focada da Datasfera. O artigo explica como:

> A noção de "Datasfera" propõe uma compreensão holística de toda a "informação" existente na Terra, originada tanto em sistemas naturais quanto socioeconômicos, que pode ser capturada em forma digital, flui por meio de redes e é armazenada, processada e transformada por máquinas.

O desejo de se afastar de um cardápio restrito de opções de políticas em parte motivou a adoção de uma terminologia refinada da Datasfera, descrita como: "o complexo sistema que abrange todos os tipos de dados e suas interações dinâmicas com normas e grupos humanos" (Porciuncula & de La Chapelle, 2022, p. 3).

Em essência, essa fórmula chama atenção para as interações mútuas entre artefatos digitais (conjuntos de dados); partes interessadas e relações sociais (grupos humanos); e regras e expectativas sociais (normas) – bem como para a multiplicidade de cada um desses aspectos. Notavelmente, o modelo implica a governança

de uma Datasfera interconectada, e não de muitas instâncias isoladas, e o faz com o propósito de fornecer uma lente holística sobre a complexidade em evolução da governança de dados e seu impacto na criação de valor e bem-estar para todos. Ou seja, a Datasfera é vista como um único sistema complexo (Siegenfeld & Bar-Yam, 2020). Ou, indo além, conforme Porciuncula e de La Chapelle (2022), a Datasfera é um complexo sistema adaptativo com dinâmicas emergentes.

Passar de uma discussão sobre "governar os dados" para "governar a Datasfera" envolve identificar os elementos em foco da Datasfera, bem como reconhecer as relações de governança de dados entre elementos (por exemplo, em relação ao indivíduo ou à empresa) e os níveis em que esses elementos se estruturam (por exemplo, organizacional, industrial, social, nacional ou global). Ao oferecer a tipologia de conjuntos de dados, grupos humanos e normas, o marco referencial da Datasfera convida então a um detalhamento mais claro sobre o campo da governança e dos fatores levados em consideração ao propor ou avaliar regimes de governança particulares.

Conclusão

Este documento oferece um ponto de partida para a reflexão sobre o desenvolvimento da governança de dados como um campo de conhecimento complexo e integrado por várias áreas do conhecimento. Fornece uma visão geral de alto nível dos grupos de pesquisa e temas abordados na literatura, destacando que, em última análise, ainda não há um único campo de governança de dados a se falar, mas sim uma gama de campos de trabalho distintos, cada um respondendo a desafios temáticos ou setoriais. Embora a governança de dados em níveis empresarial e social sejam, em termos amplos, duas faces da mesma moeda, relativamente poucos trabalhos exploraram questões de governança transfronteiriça de dados, deixando uma lacuna significativa a ser preenchida.

Além disso, o presente artigo sugere que o marco referencial da Datasfera tem uma contribuição significativa a fazer para as pesquisas e práticas atuais em governança de dados, principalmente ao trazer a noção de "governança da Datasfera" como uma abordagem sistêmica à governança de dados.

(...) ainda não há um único campo de governança de dados a se falar, mas sim uma gama de campos e trabalhos distintos, cada um respondendo a desafios temáticos ou setoriais.

Referências

Aaronson, S. A. (2021). *Data is disruptive: How data sovereignty is challenging data governance.* Hinrich Foundation. https://www.wita.org/wp-content/uploads/2021/08/Data-is-disruptive-Hinrich-Foundation-white-paper-Susan-Aaronson-August-2021.pdf

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management, 49*.

Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: An analysis of the literature. *Journal of Decision Systems, 25*.

Benfeldt Nielsen, O. (2017). A comprehensive review of data governance literature. *Selected Papers of the IRIS*, (8). https://core.ac.uk/download/pdf/301373908.pdf

Bergé, J. S., Grumbach, S., & Zeno-Zencovich, V. (2018). The 'datasphere', data flows beyond control, and the challenges for law and governance. *European Journal of Comparative Law and Governance, 5*(2). https://papers.ssrn.com/abstract=3185943

Bozkurt, Y., Rossmann, A., & Pervez, Z. (2022). *A literature review of data governance and its applicability to smart cities*. Hawaii International Conference on System Sciences. http://hdl.handle.net/10125/79666

Brous, P., Janssen, M., & Vilminko-Heikkinen, R. (2016). Coordinating decision-making in data management activities: A systematic review of data governance principles. *Lecture Notes in Computer Science, 9820*.

Datasphere Initiative. (2022). *Datasphere governance atlas*. https://www.thedatasphere.org/wp-content/uploads/2022/04/Datasphere-Governance-Atlas-2022-Datasphere-Initiative.pdf

Davies, T., Walker, S. B., Rubinstein, M., & Perini, F. (Eds.). (2019). *The state of open data: Histories and horizons*. https://www.idrc.ca/en/book/state-open-data-histories-and-horizons

de La Chapelle, B., & Porciuncula, L. (2021). We need to talk about data: Framing the debate around the free flow of data and data sovereignty. *Internet & Jurisdiction Policy Network*.

de La Chapelle, B., & Porciuncula, L. (2022). *Hello datasphere: Towards a systems approach to data*. Datasphere Initiative. https://www.thedatasphere.org/datasphere-publish/hello-datasphere/

Delacroix, S., & Lawrence, N. D. (2019). Bottom-up data trusts: Disturbing the "one size fits all" approach to data governance. *International Data Privacy Law*, *9*(4).

Google. (s.d.). *Google Books Ngram Viewer*. https://books.google.com/ngrams/

Greenleaf, G. (2012). Global data privacy laws: 89 countries, and accelerating. *Privacy Laws & Business International Report*, (115). https://papers.ssrn.com/abstract=2000034

Ladley, J. (2019). *Data governance: How to design, deploy, and sustain an effective data governance program*. Academic Press. https://books.google.com?id=AkW9DwAAQBAJ

Lehtiniemi, T., & Haapoja, J. (2020). Data agency at stake: MyData activism and alternative frames of equal participation. *New Media & Society*, *22*(1).

McCaig, M., & Rezania, D. (2021). *A scoping review on data governance*. International Conference on IoT Based Control Networks & Intelligent Systems.

Micheli, M., Ponti, M., Craglia, M., & Suman, A. B. (2020). Emerging models of data governance in the age of datafication. *Big Data & Society, 7*(2).

Panian, Z. (2009). Recent advances in data management. WSEAS Transactions on Computers, *9*(7), 1061–1071.

Pisa, M., Dixon, P., Ndulu, B., & Nwankwo, U. (2020). *Governing data for development: Trends, challenges, and opportunities*. Center for Global Development. https://www.cgdev.org/sites/default/files/governing-data-development-trends-challenges-and-opportunities.pdf

Porciuncula, L., & de La Chapelle, B. (2022). *Hello datasphere: Towards a systems approach to data governance*. The Datasphere Initiative. https://www.thedatasphere.org/news/hello-datasphere-towards-a-systems-approach-to-data-governance/

Siegenfeld, A. F., & Bar-Yam, Y. (2020). *An introduction to complex systems science and its applications*. Complexity.

Susha, I., Janssen, M., & Verhulst, S. (2017). Data collaboratives as a new frontier of cross-sector partnerships in the age of open data: Taxonomy development. *Hawaii International Conference on System Sciences*. https://aisel.aisnet.org/hicss-50/eg/open_data_in_government/4

Verhulst, S., Young, A., Zahuranec, A., Aaronson, S. A., Calderon, A., & Gee, M. (2020). *The emergence of a third wave of open data*. GovLab. https://apo.org.au/node/311570

Wendehorst, C. (2020). *Data governance working group: A framework paper for GPAI's work on data governance*. Global Partnership on Artificial Intelligence. https://gpai.ai/projects/data-governance/gpai-data-governance-work-framework-paper.pdf

Entrevista I

Big Data e a produção de estatísticas

Nesta entrevista, Pedro Luis do Nascimento Silva, secretário da Sociedade para o Desenvolvimento da Pesquisa Científica (SCIENCE), discute as oportunidades e os desafios para a adoção de fontes de *Big Data* na produção de estatísticas de qualidade, os sistemas estatísticos nacionais e a articulação de redes para a governança de dados entre múltiplos atores.

Foto: Álvaro Vasconcellos

Pedro Luis do Nascimento Silva
Secretário da Sociedade para o Desenvolvimento da Pesquisa Científica (SCIENCE).

Panorama Setorial da Internet (P.S.I.)_ Quais são as possibilidades de adoção de fontes de dados orgânicos ou* Big Data *para a produção de estatísticas de qualidade? Há cuidados indispensáveis no uso desse tipo de dado?

Pedro Luis do Nascimento Silva (P.S.)_ A geração e a disponibilidade de dados que vemos atualmente não têm precedentes na história. Ao mesmo tempo, há uma demanda crescente por dados mais frequentes e detalhados sobre as condições de vida, o ambiente, etc. As oportunidades para tirar proveito de fontes de dados orgânicos ou *Big Data* têm levado as organizações produtoras de estatísticas públicas e oficiais a estudar, desenvolver e aplicar métodos e sistemas visando produzir estatísticas de qualidade a partir dessas informações.

Há dois caminhos principais: usar uma ou mais fontes novas para gerar diretamente estatísticas de interesse; ou combinar dados de uma ou mais fontes novas com dados de fontes tradicionais, tais como pesquisas amostrais, censos e registros administrativos. Em ambos os casos, os objetivos podem incluir a cobertura de lacunas sobre temas não explorados, a substituição de estatísticas antes obtidas com base em fontes tradicionais e a ampliação da produção estatística em termos de frequência ou nível de detalhamento. Nesse sentido, o uso de dados orgânicos para gerar estimativas mais frequentes ou mais desagregadas (em pequenas áreas) a partir de pesquisas amostrais já existentes é um dos campos de grande potencial e interesse.

Em todo caso, as novas estatísticas devem satisfazer aos requisitos de qualidade estabelecidos nos Princípios Fundamentais das Estatísticas Oficiais[3] da Organização das Nações Unidas (ONU), de modo que atendam às necessidades dos interessados e sejam adequadas para uso. Há vários sistemas de referência e códigos de boas práticas que podem guiar a produção de estatísticas derivadas dessas novas fontes, com destaque para a proposta da Comissão Econômica das Nações Unidas para a Europa (UNECE)[4]. Entre os principais desafios em torno dessa questão, há a tensão natural entre satisfazer demandas por estatísticas e assegurar a qualidade delas.

[3] Disponível em: https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx
[4] Disponível em: https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf

"(...) é um processo delicado a produção de estatísticas inéditas ou a substituição de estatísticas tradicionais por outras baseadas nas fontes novas, assegurando os requisitos de qualidade. Isso envolve aplicar ou desenvolver novos métodos e sistemas (...)."

Progressos têm sido feitos, mas ainda há poucos exemplos de situações em que as fontes tradicionais foram de fato substituídas pelas novas fontes de dados orgânicos. Um caso de sucesso interessante se deu na produção de estatísticas sobre a mobilidade da população durante o período da pandemia COVID-19[5]. É um exemplo claro de uso que não poderia ser facilmente coberto com dados de fontes tradicionais.

P.S.I._ Quais são os principais desafios enfrentados pelos institutos de estatísticas oficiais para o uso de fontes tipo* Big Data*?

P.S._ Há três desafios principais: o acesso aos dados de muitas das novas fontes orgânicas; a capacitação de pessoal para lidar com dados das novas fontes, bem como da combinação destes com dados de fontes tradicionais; e a necessidade de produzir estatísticas que satisfaçam aos exigentes requisitos de qualidade impostos à produção estatística "tradicional".

Grande parte dos dados de novas fontes orgânicas é produzida e mantida por organizações privadas que os enxergam como ativos de alto valor e, por essa razão, não se mostram dispostas a compartilhá-los com terceiros – nem mesmo com agências de estatísticas oficiais, que servem ao bem público. Além disso, há questões de comparabilidade ao longo do tempo, de escassez de padrões para a captura e a harmonização dos dados, e até mesmo de continuidade na obtenção e no armazenamento das informações.

Nesse sentido, um exemplo marcante na história recente do país foi a recusa das empresas de telefonia em ceder ao Instituto Brasileiro de Geografia e Estatística (IBGE) informações de seus cadastros de clientes de telefonia fixa e móvel. A intenção era que o IBGE pudesse realizar via telefone a coleta de sua principal pesquisa por amostragem domiciliar, impossibilitada de acontecer por meio de entrevistas face a face durante o ano inicial da pandemia COVID-19.

Em benefício do aumento e da atualização de sua produção estatística, tanto o IBGE quanto outras agências no Brasil e no exterior têm investido na capacitação de suas equipes para que possam absorver, desenvolver e aplicar os métodos e processos necessários à exploração das novas fontes de dados orgânicos. Entretanto, o retorno desses esforços requer tempo e maturação, sendo cedo para afirmar que as agências estatísticas estão prontas para tirar o máximo proveito das novas fontes de dados orgânicos disponíveis.

Por fim, é um processo delicado a produção de estatísticas inéditas ou a substituição de estatísticas tradicionais por outras baseadas nas fontes novas, assegurando os requisitos de qualidade. Isso envolve aplicar ou desenvolver novos métodos e sistemas, bem como realizar consultas com usuários especializados, validações externas e diversas etapas de testes até que as estatísticas sejam consideradas aptas para uso e publicação. Além de o tempo necessário ser longo, muitas vezes as melhores propostas testadas não conseguem entregar dados com a qualidade requerida.

[6] Saiba mais: https://www.gstatic.com/covid19/mobility/2022-10-15_BR_Mobility_Report_pt-BR.pdf

P.S.I._ De que forma o sistema estatístico brasileiro pode incorporar fontes alternativas de dados, produzidos e/ou coletados por outras instituições (privadas, não governamentais, etc.)?

P.S._ Um passo importante para acelerar o aproveitamento das fontes alternativas de dados no Brasil seria elaborar um novo marco legal para a produção de estatísticas públicas e oficiais. A lei brasileira não contempla instrumentos fundamentais que permitem às agências estatísticas o acesso a dados produzidos ou mantidos por instituições privadas e não governamentais. Mesmo entre instituições públicas há limitações de acesso. Para citar um exemplo ilustrativo, a Receita Federal nunca permitiu ao IBGE acessar microdados do Imposto de Renda, seja de empresas, seja de pessoas físicas, nem mesmo anonimizados.
O marco legal deveria definir claramente papéis, direitos e deveres das instituições que buscam acesso a dados individuais de pessoas, empresas, transações visando produzir estatísticas públicas e oficiais. Deveria constar entre as obrigações a proteção de confidencialidade de informações individuais, tal como previsto na atual legislação, mas seria preciso prever a possibilidade de aproveitá-las para fins legítimos de produção de estatísticas de interesse público.
Outra área em que um novo marco legal teria papel decisivo seria na criação de instâncias de governança referentes à produção, ao armazenamento e à utilização de dados para fins estatísticos. Há em alguns países um arranjo interessante de *data archive*, isto é, instituições dedicadas ao armazenamento, à curadoria, à descoberta e à disseminação de conjuntos de dados de interesse público. Podemos citar como exemplos o UK Data Archive[6], no Reino Unido, e o Inter-university Consortium for Political and Social Research (ICPSR)[7], sediado nos Estados Unidos. Ainda não temos no Brasil uma instituição similar, com o mandato legal e o aparato institucional necessários para promover atividades dessa natureza.
Por último, mas não menos importante, seria essencial criar e ativar uma instância de coordenação efetiva do sistema estatístico nacional. Hoje, esse papel é delegado ao IBGE, mas este não o exerce por falta de instrumentos efetivos. Um modelo relevante é o da UK Statistics Authority , criada em 2007, quando da publicação do marco legal mais recente para a produção de estatísticas oficiais no Reino Unido.

"Um passo importante para acelerar o aproveitamento das fontes alternativas de dados no Brasil seria elaborar um novo marco legal para a produção de estatísticas públicas e oficiais."

P.S.I._ Dada a natureza privada de grande parte das fontes de* Big Data, *quais os caminhos possíveis para a articulação entre instituições proprietárias e usuárias dos dados visando à gestão e à governança dessas fontes? Quais aspectos devem ser considerados na construção de redes para a governança dos dados entre os múltiplos atores?

P.S._ Um caminho possível é a criação de um "arquivo nacional de dados" para implementar o armazenamento, a curadoria, a descoberta e a disseminação de conjuntos de dados de interesse público. Tal organização poderia desempenhar um papel de intermediação entre os proprietários e os usuários dos dados, executando a gestão e a governança dos dados de que fosse depositária. Uma das vantagens desse arranjo é a garantia de permanência ou de longevidade dos

[7] Disponível em: https://www.icpsr.umich.edu/web/pages/
[8] Disponível em: https://uksa.statisticsauthority.gov.uk/

"(...) os mecanismos de governança precisam considerar modelos como o do Comitê Gestor da Internet no Brasil (CGI.br), formado por representantes de vários segmentos (...)."

dados depositados, bem como a explicitação das regras e condições de acesso para todos os interessados em usá-los. Mas há também limitações, como a provável defasagem entre o momento da produção dos dados e sua disponibilização para acesso por terceiros.

Além disso, é possível estabelecer "contratos de uso" entre as instituições proprietárias dos dados e as agências estatísticas interessadas no seu uso, em moldes similares à atuação de empresas de auditoria externa. Estas requerem acesso não limitado aos dados econômicos, financeiros e contábeis das empresas auditadas, mas se comprometem com a manutenção de sua confidencialidade e a utilização dos dados apenas para fins da prestação dos serviços contratados.

Seguindo esse modelo, as agências estatísticas poderiam receber acesso não limitado aos dados orgânicos de interesse para uma operação estatística específica, assumindo compromissos de preservação da confidencialidade e de uso exclusivo para os fins autorizados nos contratos de uso. Esse tipo de arranjo permite o acesso direto aos dados na fonte, sem intermediários nem defasagem temporal. Por outro lado, existe o risco de que as instituições proprietárias cobrem pelo acesso valores que as agências estatísticas não têm condições de pagar, já que costumam ser financiadas pelo poder público, têm capacidade limitada de captar recursos por iniciativa própria e precisam disponibilizar suas estatísticas de forma gratuita ao público.

Em qualquer um dos casos, devemos sempre buscar preservar: os requisitos de proteção da confidencialidade dos dados individuais de pessoas e organizações; os interesses comerciais legítimos das instituições proprietárias dos dados; e o interesse público na produção de estatísticas. Tais aspectos sugerem que os mecanismos de governança precisam considerar modelos como o do Comitê Gestor da Internet no Brasil (CGI.br)[9], formado por representantes dos vários segmentos envolvidos na questão e que serve de exemplo brasileiro bem-sucedido.

Artigo II

As promessas e os desafios da transformação digital centrada em dados na era da Inteligência Artificial

Por Moinul Zaber[10]

A emergência da Inteligência Artificial (IA) que pode empregar dados de diferentes naturezas na construção de ferramentas eficientes ou na reunião de

[9] Saiba mais: https://cgi.br/

[10] Doutor pela Universidade Carnegie Mellon, é cientista social computacional com foco na aprendizagem aplicada de máquina e na ciência de dados para a política tecnológica. Atualmente, é acadêmico sênior na Universidade das Nações Unidas (UNU), além de trabalhar em colaboração com autoridades reguladoras de telecomunicações e de concorrência no Sul Global. Como professor e pesquisador, atuou em Bangladesh, Japão, Suécia, Sri Lanka e Portugal.

insights tem demonstrado potencial para mudar a maneira como seres humanos e instituições tradicionalmente tomam decisões. O uso de IA e de dados em órgãos públicos tem viabilizado serviços públicos mais proativos e automatizados. Entretanto, uma vez que a IA como ciência está em fase incipiente, a aplicação institucional de várias ferramentas de IA é um desafio. O obstáculo mais significativo vem dos dados – a matéria-prima da IA. Para evitar o risco de falha no uso de IA, as instituições que aspiram a uma tomada de decisão centrada em dados precisam se adaptar às novas formas de transformação digital. Este artigo lança luz sobre diversos aspectos da transformação digital centrada em dados de modo a permitir a automação focada na IA por parte de órgãos públicos.

As promessas dos dados e da IA

Um órgão público é qualquer entidade autônoma (como um departamento, uma comissão ou uma autoridade) estabelecida pelo governo local ou nacional. Seu principal objetivo é executar as tarefas necessárias ordenadas pelo governo e pelos residentes a quem essas organizações servem. Tais deveres podem incluir: a salvaguarda da segurança nacional; o provimento de proteção civil; a regulação de mercados; e a garantia de acesso a necessidades como alimentação, abrigo, energia, comunicação e proteção ambiental. Entretanto, a maioria desses serviços tem duas características distintas. Do lado da demanda, a tomada de decisão em vários níveis de prestação de serviços; do lado da oferta, o envolvimento com as pessoas beneficiárias dos serviços. Como em qualquer processo de tomada de decisão, exige-se uma grande quantidade de dados, às vezes de fontes internas e, muitas vezes, de várias fontes externas.

A transformação digital centrada em dados ajuda a automatizar o processo de movimentação de dados. A meta é alcançar uma tomada de decisão centrada no cidadão e na prestação de serviços. Trata-se de um esforço contínuo que envolve a integração de dados e tecnologias digitais para aperfeiçoar processos comerciais, criar novos modelos de negócios e prestar melhores serviços aos clientes. Por exemplo, os serviços públicos com foco na segurança social proporcionam aos indivíduos certo grau de segurança de renda quando confrontados com contingências de velhice, sobrevivência, incapacidade, invalidez, desemprego ou criação de filhos. Podem oferecer também acesso a cuidados médicos terapêuticos ou preventivos. Em todo o mundo, a transformação digital está permitindo a implementação de sistemas de serviços públicos cada vez mais abrangentes.

A adoção emergente em instituições públicas de ferramentas de IA que têm como matéria-prima várias formas de dados possibilita serviços públicos mais proativos e automatizados. O uso de dados e de IA pode ajudar a melhorar a eficiência, a eficácia e a capacidade de resposta do serviço público. Por meio dessas tecnologias, é possível aos órgãos governamentais atender melhor às necessidades dos cidadãos e oferecer mais valor às comunidades. Muitos órgãos têm usado a análise de dados para identificar áreas onde o serviço público é inexistente ou precisa ser aperfeiçoado. O advento dos campos de IA – tais como aprendizagem de máquina (*machine learning*), reconhecimento de

Foto: Cristina Braga

Moinul Zaber
Universidade das Nações Unidas (UNU).

A análise preditiva pode ser empregada para antecipar tendências e necessidades futuras, de modo que os órgãos governamentais planejem e aloquem recursos com mais eficácia.

padrões, processamento de linguagem natural, visão computadorizada e visualização de dados – está moldando a maneira como os dados em suas várias formas podem ser usados para tornar o serviço público mais eficaz e centrado no usuário.

OS DADOS COMO MATÉRIA-PRIMA DA TRANSFORMAÇÃO DIGITAL NA ERA DA IA

No lado da oferta, ao coletar e analisar dados como taxas de resposta, satisfação do cliente e tempo de espera, é possível melhorar o nível de envolvimento em relação às pessoas beneficiárias dos serviços. *Chatbots* e assistentes virtuais alimentados por IA são capazes de fornecer assistência 24 horas aos cidadãos que procuram informações ou ajuda com serviços do governo. Já do lado da demanda, os órgãos governamentais conseguem identificar as áreas que requerem alocação de mais recursos ou implementação de novas políticas. Ao analisar dados demográficos e socioeconômicos, entre outros, a IA pode auxiliar os funcionários públicos a localizar disparidades na prestação de serviços e tomar medidas para resolvê-las.

Quando usados para identificar padrões nos dados, algoritmos de aprendizagem de máquina são capazes de ajudar órgãos governamentais a detectar fraudes, desperdícios e abusos, poupando dinheiro dos contribuintes e melhorando a eficiência geral dos serviços. A análise preditiva pode ser empregada para antecipar tendências e necessidades futuras, de modo que os órgãos governamentais planejem e aloquem recursos com mais eficácia.

Por meio da análise preditiva, por exemplo, governos locais antecipam picos na demanda por serviços de emergência durante determinados períodos do ano. Os dados e a IA podem ser usados para identificar potenciais riscos e ameaças em tempo real, permitindo aos órgãos de aplicação da lei prevenir crimes e melhorar a segurança pública. Já ferramentas de visualização de dados conseguem apresentar dados complexos de tal forma que seja fácil de entender e interpretar, o que ajuda os órgãos governamentais em uma comunicação mais eficaz com os cidadãos e na tomada de decisão baseada em dados.

Entretanto, é importante assegurar que essas tecnologias sejam implementadas de forma responsável, com as devidas salvaguardas para proteger a privacidade e evitar enviesamentos. Uma vez que a IA depende muito de dados para treinar modelos e fazer previsões, as organizações precisam garantir que seus dados sejam de alta qualidade, confiáveis e acessíveis. Há também questões relacionadas à ética e à legalidade que abrangem privacidade pessoal, transparência e equidade, bem como assuntos de segurança nacional. Além disso, para garantir sua eficácia a IA e as intervenções baseadas em dados devem ser compatíveis com os sistemas e as práticas herdados dos órgãos. As organizações precisarão investir em capital humano para reduzir o fosso de habilidades. Como a automação herdada está sendo substituída pela automação centrada em dados e baseada em IA, habilidades em ciência de dados, aprendizagem de máquina e desenvolvimento de IA se fazem necessárias.

Tradicionalmente, a computação se concentra no código. A IA se concentra nos dados. O desempenho de um sistema baseado em IA depende de um fornecimento contínuo de dados de boa qualidade. A precisão e a confiabilidade dos resultados gerados pelos algoritmos estão muito associadas à qualidade dos dados utilizados para treiná-los. Por exemplo, os algoritmos de aprendizagem de máquina partem dos padrões e relações entre os dados para fazer previsões ou classificar os dados. Se os dados forem imprecisos, incompletos ou contiverem erros, é possível que o algoritmo aprenda padrões incorretos ou faça previsões imprecisas, levando a um desempenho insatisfatório. Além disso, a ausência de dados de qualidade gera padrões tendenciosos que podem resultar em previsões discriminatórias ou injustas. É importante notar que esses sistemas são projetados para que sejam aprendizes contínuos. Portanto, deve haver um fornecimento permanente de dados de boa qualidade, do contrário o resultado dos algoritmos não será compatível com o contexto.

Os fatos: breve introdução sobre IA e suas ramificações

Os processos algorítmicos da IA diferem dos tradicionais e podem ser mais eficientes para diversas tarefas. No caso da IA, em vez de escrever um programa para cada tarefa específica, são coletados muitos exemplos que especificam o resultado (*output*) correto (ou incorreto) para uma determinada entrada (*input*). Os algoritmos de IA então partem desses exemplos para produzir um programa que faça o trabalho e seja escalável para novos casos. Os programas se adaptam às mudanças nos dados, uma vez que é da essência dos programas de IA se retreinarem com base em novos dados. Hoje, quantidades maciças de capacidade computacional estão disponíveis para essas tarefas, razão pela qual seu uso é mais barato do que a escrita de um programa específico para cada tarefa.

Essa capacidade de escalabilidade e de reunião de *insights* a partir de dados tornou a IA uma ferramenta essencial e complementar para formuladores de políticas e prestadores de serviços visando ao bem social. Várias ferramentas de IA estão sendo utilizadas para: responder a crises; promover fortalecimento econômico; atenuar desafios educacionais; mitigar desafios ambientais; garantir igualdade e inclusão; promover saúde; reduzir a fome; verificar e validar informação; gerir infraestrutura; gerir os setores público e social; e até mesmo para segurança e justiça.

Essa capacidade de escalabilidade e de reunião de *insights* a partir de dados tornou a IA uma ferramenta essencial e complementar para formuladores de políticas e prestadores de serviços visando ao bem social.

APRENDIZAGEM DE MÁQUINA E A NECESSIDADE DE DADOS DE MELHOR QUALIDADE

A IA é um campo amplo que engloba muitas ramificações, cada uma focalizando diferentes aspectos da inteligência e do processamento cognitivo. Algumas das principais ramificações da IA incluem aprendizagem de máquina, processamento de linguagem natural, visão por computador, robótica e sistemas especializados. Entre estas, a aprendizagem de máquina lida com processo de aprendizagem, raciocínio, busca por padrões e tomada de

Em termos amplos, categoriza-se a aprendizagem de máquina em três ramos principais: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem de reforço.

decisão. Trata-se de um guarda-chuva de métodos que ajudam a construir ferramentas práticas para outros ramos da IA.

É possível definir um problema de aprendizagem como o problema de aperfeiçoar alguma medida de desempenho ao executar determinadas tarefas, usando para tal algum tipo de experiência de treinamento. Por exemplo, na aprendizagem para detectar elegibilidade em casos de aposentadoria, a tarefa é determinar "elegível" ou "não elegível" no requerimento de qualquer morador. A métrica de desempenho pode ser medir a precisão desse classificador de elegibilidade. O algoritmo pode ser treinado a partir de um conjunto de dados contendo informações históricas de elegibilidade dos requerimentos, cada um rotulado em retrospectiva como elegível ou não. Podem existir muitas outras medidas alternativas de precisão e conjuntos de treinamento misturados com dados rotulados e não rotulados. Em termos amplos, categoriza-se a aprendizagem de máquina em três ramos principais: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem de reforço.

Para diversas aplicações, pode ser muito mais fácil treinar um sistema mostrando-lhe exemplos do comportamento desejado de entrada-resultado do que por meio da programação manual, antecipando a resposta almejada para todas as entradas possíveis. A aprendizagem supervisionada é um tipo de aprendizagem de máquina em que o algoritmo é treinado com dados rotulados. Aqui, a entrada é emparelhada com o resultado. O algoritmo aprende a mapear os dados de entrada para os dados de saída, permitindo que sejam feitas previsões sobre novas instâncias de dados não vistos. Prever se um *email* é ou não *spam* é um exemplo de classificação, enquanto prever o valor de uma casa com base em suas características é um exemplo de regressão – dois tipos de algoritmos de aprendizagem supervisionada.

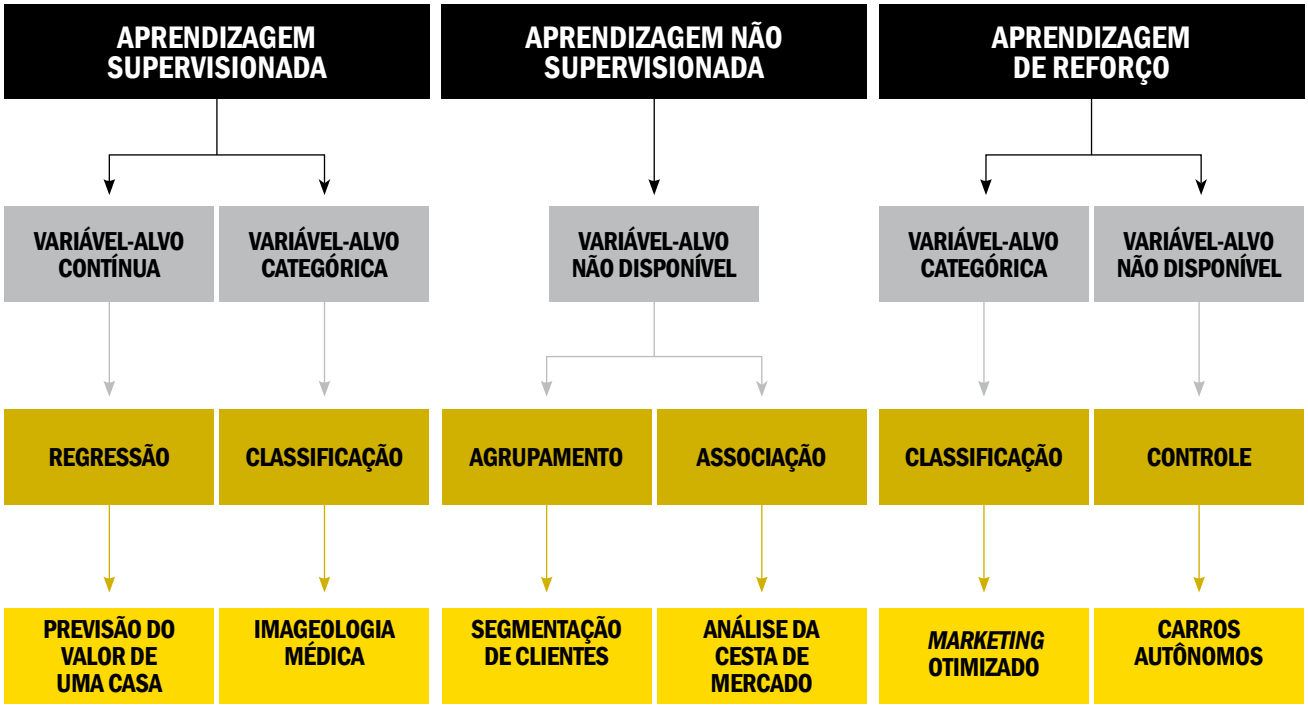
Já na aprendizagem não supervisionada o algoritmo é treinado com dados não rotulados. Aqui, os dados de entrada não são emparelhados com nenhuma variável-alvo. O algoritmo aprende a identificar padrões e estrutura nos dados sem qualquer conhecimento prévio sobre qual deve ser o resultado. Tarefas de agrupamento – como organizar clientes em segmentos com base em seu comportamento de compra – e tarefas de redução de dimensionalidade – como reduzir o número de variáveis em um conjunto de dados – são exemplos de aprendizagem não supervisionada.

A aprendizagem de reforço, por sua vez, é um tipo de aprendizagem de máquina em que um agente é ensinado a interagir com um ambiente, aprende por treinamento e erro, e realiza ações que maximizam um sinal de recompensa. Ao tentar ensinar determinada tarefa a um animal de estimação, por exemplo, podemos presenteá-lo (recompensa) se ele executar a tarefa corretamente. Caso contrário, podemos dizer "NÃO", indicando uma penalidade. Com o tempo, o animal de estimação passa a associar o comportamento correto à recompensa e melhora a execução do truque. A aprendizagem de reforço pode ser usada em *chatbots* ou agentes de conversação para aperfeiçoar seu desempenho na compreensão ou resposta às consultas dos usuários, levando a uma maior satisfação e melhor experiência do usuário.

Uma das áreas de alto impacto dos progressos na aprendizagem supervisionada envolve redes profundas. Sistemas de aprendizagem profunda utilizam algoritmos de otimização baseados em gradientes para ajustar parâmetros ao longo dessas redes multicamadas a partir de erros em seu resultado. Redes profundas baseiam-se no algoritmo de rede neural artificial que tem como modelo a estrutura e função do cérebro humano. A aprendizagem profunda permite que as máquinas aprendam com grandes quantidades de dados e reconheçam padrões que podem ser de difícil ou impossível identificação pelo ser humano.

A aprendizagem de máquina em geral depende muito da disponibilidade de dados de treinamento. A quantidade de dados rotulados necessários para treinar um modelo de aprendizagem de máquina varia a partir do conjunto de dados e do modelo adotado. A exigência aumenta de acordo com a complexidade dos conjuntos de dados e a profundidade dos modelos. Redes profundas permitem uma generalização muito maior do que redes neurais rasas e abordagens tradicionais de aprendizagem de máquina, alcançando, portanto, uma precisão significativamente melhor. Ao aplicar aprendizagem profunda a um problema, o principal desafio é a grande quantidade de dados exigidos para treinar os modelos.

Figura 1 – CLASSIFICAÇÃO AMPLA DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA



Fonte: Elaboração própria, com base em LITSLINK (2019).

APRENDIZAGEM COM DADOS LIMITADOS

Quando a quantidade de dados é limitada, muitas vezes os modelos são construídos com dados tendenciosos. Além disso, é possível que os algoritmos projetados se ajustem tão bem ao conjunto de dados de treinamento que não consigam prover a solução certa no mundo real. Em diversos casos, é difícil acumular grandes quantidades de dados ou talvez estes sequer existam (Defense Science and Technnology Laboratory [DSTL], 2020). Por exemplo, uma instituição deseja criar um modelo para prever certos traços dos usuários, mas possui apenas um histórico de dados de 50 usuários. As abordagens tradicionais de aprendizagem de máquina para esse modelo podem ter um resultado enviesado que vá contra gênero, idade e/ou raça devido à falta de variação. Problemas de aprendizagem de máquina relacionados a um conjunto pequeno de dados requerem técnicas e abordagens distintas em comparação com grandes conjuntos de dados, entre elas: uma engenharia de características que ajude a criar novas características; regularização para evitar a combinação excessiva de resultados de múltiplos modelos; e transferência de aprendizagem usando um modelo que tenha sido previamente treinado em um conjunto de dados maior. A Tabela 1 mostra diferentes tipos de métodos de aprendizagem de máquina com base na quantidade de dados.

Tabela 1 – MÉTODOS DE APRENDIZAGEM DE MÁQUINA COM BASE NA QUANTIDADE DE DADOS

QUANTIDADE DE DADOS	ROTULADOS OU NÃO ROTULADOS	MÉTODO DE APRENDIZAGEM UTILIZADO	COMENTÁRIO
Pequena quantidade	Na maioria das vezes, não rotulados	Aprendizagem com zero tentativa	Utiliza descrição do conceito para treinar o modelo, ontologia do conceito, incorporação da palavra semântica
Pequena quantidade	Na maioria das vezes, não rotulados	Rotulagem manual	Rotula manualmente os dados
Pequena quantidade	Na maioria das vezes, rotulados	Aprendizagem de máquina rasa, meta-aprendizagem, raciocínio do conhecimento	Treina um metamodelo para ser aplicado a tarefas invisíveis, máquinas vetoriais de suporte, árvores de decisão, *perceptron* multicamadas, abordagem ontológica alavancando a descrição dos objetos
Grande quantidade	Na maioria das vezes, rotulados	Aprendizagem profunda	Rede neural convolucional
Grande quantidade	Na maioria das vezes, não rotulados	Aprendizagem ativa, aprendizagem semissupervisionada, aprendizagem autossupervisionada, aprendizagem não supervisionada	Utilizando dados rotulados e não rotulados, consulta visando selecionar exemplos para a rotulagem por um operador humano, agrupamento, detecção de anomalias, variável latente, rotulagem autônoma

Fonte: Elaboração própria.

Os desafios do uso ineficiente de dados para a IA

As ferramentas de IA imitam a maneira como os seres humanos pensam e agem. Isto significa que os algoritmos podem ser imprecisos em muitas ocasiões. Tais imprecisões são capazes de gerar riscos à privacidade pessoal, à segurança nacional, à equidade, à transparência e à *accountability*. É possível que a imprecisão engendre dados, algoritmos e a interação humana com o processo de *design*.

Os sistemas de IA, se treinados em conjuntos de dados tendenciosos, podem perpetuar e até mesmo amplificar vieses existentes nos dados. Se os dados de treinamento não forem representativos ou não tiverem diversidade, o sistema de IA aprenderá a fazer previsões enviesadas. Uma enorme quantidade de dados é introduzida na máquina para reconhecer certos padrões. Dados não estruturados da Web, de mídias sociais, de dispositivos móveis, de sensores e dispositivos inteligentes (isto é, Internet das Coisas) dificultam a absorção, ligação, ordenação e manipulação dos dados. Na falta de uma curadoria cuidadosa dos dados, o conjunto de dados pode estar repleto de dados incompletos ou ausentes, assim como pode ser impreciso ou tendencioso.

Em termos amplos, há quatro tipos de vieses: viés de amostra; viés de medição; viés algorítmico; e viés contra grupos ou classes de objetos e pessoas. Entretanto, o viés algorítmico parece ser o menos discutido. Alguns algoritmos são sistematicamente enviesados em direção a um tipo específico de dados. Vários motivos dificultam a correção de vieses. Em primeiro lugar, a introdução de viés nem sempre é óbvia durante a construção de um modelo. Em segundo lugar, é difícil identificar de forma retroativa de onde o viés surgiu. Em terceiro lugar, a aprendizagem de máquina, um dos campos de IA amplamente utilizados para a análise de dados, precisa treinar, testar e validar seu algoritmo com o conjunto de dados. Para tal, em muitos casos os dados são divididos de maneira aleatória para treinamento, teste e validação, o que pode preservar os mesmos vieses. Em quarto lugar, a falta de contexto devido à incapacidade de compreender os usuários-alvo é capaz de criar vieses. Um sistema projetado no país A não pode ser aplicado no país B, já que comunidades diferentes têm maneiras diferentes de encarar problemas de políticas públicas. Por fim, o contexto não é apenas afetado pelas comunidades, mas é também definido pelas instituições. Por exemplo, a "justiça" no caso do "problema do desemprego" pode diferir de quando se trata de "justiça penal".

Dados pessoais podem ser removidos de um conjunto de dados, enquanto outro conjunto pode dispor de dados revelados pelo sistema de IA. Há o risco de que isso cause uma divulgação inadvertida de dados sensíveis, a menos que se tenha o cuidado de remover dados pessoais de todos os conjuntos de dados. Além disso, o viés depende em grande medida das pessoas que fazem a curadoria dos dados ou desenvolvem os algoritmos, decidindo como estes serão implantados e, em última instância, como serão utilizados. Muito deriva da forma como um problema é enquadrado. Ao definir um problema,

Os sistemas de IA, se treinados em conjuntos de dados tendenciosos, podem perpetuar e até mesmo amplificar vieses existentes nos dados.

Por fim, a interação entre humanos e máquinas tem de ser avaliada. Se os operadores das ferramentas de IA não reconhecem quando os sistemas precisam ser revogados, acidentes e lesões podem ocorrer.

cientistas decidem o que de fato querem alcançar quando criam um modelo de aprendizagem. Mesmo a composição da equipe de engenharia pode ser tendenciosa. O enquadramento do problema depende de quem o projeta, quem decide como ele é implantado, qual o nível aceitável de precisão e se as aplicações da IA são éticas. O fracasso em endereçar essas questões tem proliferado algoritmos que ditam quais propagandas políticas as pessoas veem, como recrutadores filtram os candidatos a vagas de emprego e até o modo como agentes de segurança são destacados para os bairros.

Por fim, a interação entre humanos e máquinas tem de ser avaliada. Se os operadores das ferramentas de IA não reconhecem quando os sistemas precisam ser revogados, acidentes e lesões podem ocorrer. Por exemplo, um voo da companhia aérea Air France sobre o Oceano Atlântico, em junho de 2009, sofreu um acidente devido em parte ao excesso de confiança da tripulação do *cockpit* no piloto automático (o sensor de velocidade confundiu os pilotos). Os julgamentos humanos podem ser falhos quando sobrepostos aos sistemas. Lapsos na gestão de dados, erros de *script* e decisões equivocadas no treinamento de modelos são capazes de comprometer a justiça, a privacidade, a segurança e a conformidade. De modo involuntário, é possível que os coletores de dados induzam a um viés, caso acessem os dados de pessoas de determinada demografia em detrimento de outras.

EXPLICABILIDADE E DADOS

Os algoritmos de aprendizagem de máquina aprendem padrões e relações a partir de grandes quantidades de dados, com frequência sem programação explícita de regras ou critérios de tomada de decisão. Por consequência, é possível que os algoritmos produzam resultados que são precisos, mas não são intuitivos ou facilmente compreendidos pelos seres humanos. Grande parte da IA (em particular, a aprendizagem profunda) é atormentada pelo "problema da caixa-preta" (*black box problem*). Esses modelos podem ser altamente complexos, com diversas camadas e nós interligados. Muitas vezes, conhecemos as entradas e resultados do modelo, mas não sabemos o que acontece no meio. Para garantir confiança e *accountability*, é imperativo determinar como uma máquina inteligente sugere certas decisões. Além disso, se os sistemas de IA se tornam explicáveis, eles são capazes de aumentar os lucros das organizações de modo significativo, aumentar a precisão dos modelos de 15% a 30% e reduzir os esforços de monitoramento em até 50%[11].

Há várias razões para a falta de explicabilidade dos algoritmos de aprendizagem de máquina. A principal delas está relacionada aos dados e ao uso que é feito deles. Os algoritmos de aprendizagem de máquina podem perpetuar vieses presentes nos dados, gerando resultados que reforçam vieses sociais existentes. Por consequência, esses modelos também se tornam propensos a enviesamentos e ataques adversos. De forma mais significativa

[11] Saiba mais: https://www3.technologyevaluation.com/research/brochure/new-technology-the-projected-total-economic-impact-of-explainable-ai-and-model-monitoring-in-ibm-cloud-pak-for-data.html

devido à natureza de caixa-preta dos algoritmos, mostra-se difícil identificar as características que causam tais vieses. Com frequência, algoritmos de aprendizagem de máquina operam em espaços de alta dimensão. Isto resulta em relações não lineares entre características e previsões de resultado, dificultando sua explicação.

Tornar os modelos de aprendizagem de máquina explicáveis é um campo ativo de pesquisa. Alguns dos trabalhos notáveis feitos para entender a relação entre característica e produto são SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME) e Gradient-weighted Class Activation Mapping (Grad-CAM). Uma das respostas populares para o problema da caixa-preta é a IA Explicável (Explainable AI [XAI]) – um conjunto de processos de aprendizagem de máquina que permite aos usuários humanos compreender, confiar e gerenciar a IA. O objetivo da XAI é viabilizar interações entre pessoas e sistemas de IA, fornecendo informações sobre como as decisões e os eventos acontecem (Tjoa & Guan, 2021). Sua adoção foi tão ampla que há menção no Regulamento Geral sobre a Proteção de Dados (RGPD), da União Europeia, e desde 2016 a Agência de Projetos de Pesquisa Avançada de Defesa (DARPA) do governo estadunidense se dedica a investigá-la.

Por causa da natureza orientada por dados dos algoritmos de IA – em oposição à natureza orientada por programas dos algoritmos tradicionais –, os mecanismos convencionais de auditoria de sistemas de *software* são bastante inadequados. Os sistemas de IA baseiam seus resultados em milhões de pontos de dados. Além disso, mudanças nas amostras de treinamento podem induzir a diferentes aprendizagens. Portanto, não há um resultado esperado com muitos algoritmos de IA. Os sistemas aprendem qual é a melhor previsão, o que torna difícil a validação. Tal imprevisibilidade é um desafio. Isto significa que os conjuntos de dados de auditoria e os resultados não são suficientes para avaliar as ferramentas de IA.

DESAFIOS DA GOVERNANÇA ADEQUADA DOS DADOS

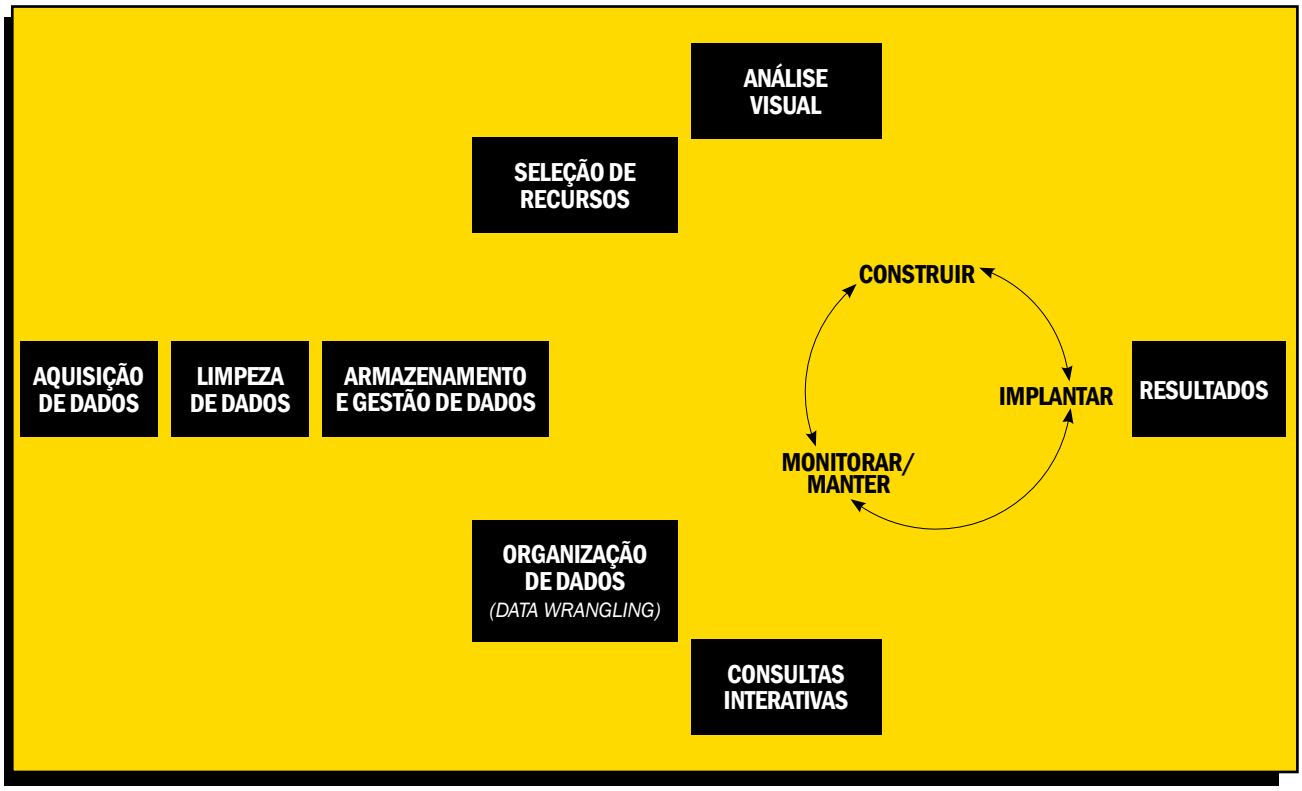
Tradicionalmente, as instituições públicas trabalham como silos que criam barreiras ao acesso e à disponibilidade de dados para outras instituições. A falta de acesso aos dados resulta em análises de dados e ferramentas de IA de qualidade inferior. A implementação de IA centrada em dados pode exigir mudanças significativas nos sistemas e processos de tecnologia de informação (TI) existentes. Isto significa que as organizações precisam assegurar que as soluções de IA se integrem perfeitamente aos seus sistemas e fluxos de trabalho, sem causar perturbações.

A precisão e a confiabilidade dos modelos de IA dependem da qualidade dos dados usados para treiná-los. Para obter o máximo dos dados, é imperativo estabelecer um modelo de governança de dados que se volte ao processo de gestão e garantia da qualidade, precisão, integridade e segurança dos dados utilizados para treinar e desenvolver algoritmos de IA (Abraham *et al*., 2019). Visando converter dados em informação, os dados devem passar por um *pipeline* que consiste em uma série de etapas, sendo os resultados de uma etapa capazes de influenciar a próxima. Há uma ordem específica que pode não ser linear, uma vez que o processamento de dados pode ser um processo iterativo.

(...) é imperativo estabelecer um modelo de governança de dados que se volte ao processo de gestão e garantia da qualidade, precisão, integridade e segurança dos dados utilizados para treinar e desenvolver algoritmos de IA.

As etapas partem da coleta de dados, em que os dados brutos são reunidos, e do pré-processamento, quando a limpeza e a transformação acontecem para garantir a qualidade. Depois, os dados são armazenados de diversas formas em armazéns de dados. A seguir, os dados passam à fase de análise, momento em que vários padrões são identificados; à modelagem, quando modelos matemáticos são empregados para detectar anomalias ou prever resultados; e à visualização, em que os *insights* são resumidos visualmente.

Figura 2 – REPRESENTAÇÃO DE UM PROCESSO SIMPLES DE APRENDIZAGEM DE MÁQUINA, EM QUE DADOS BRUTOS SE TRANSFORMAM EM RESULTADOS



Fonte: Elaboração própria.

Para garantir a devida gestão dessas etapas, é necessário um modelo adequado de governança de dados, o que envolve a definição de políticas e procedimentos para cada uma das fases acima mencionadas. Isto inclui identificar as fontes de dados, estabelecer padrões de qualidade de dados, definir a propriedade e a administração de dados, bem como garantir a conformidade com os regulamentos e padrões relevantes da indústria. É importante assegurar que os dados usados para treinar modelos de IA sejam consistentes, precisos e relevantes.

A governança de dados para IA envolve ainda o estabelecimento de processos de preparação e pré-processamento de dados, incluindo limpeza e

normalização de dados, junto com a engenharia de características. Na esfera das políticas, a governança de dados também trata de preocupações éticas e de privacidade.

DESAFIOS INSTITUCIONAIS DA IMPLEMENTAÇÃO DE IA CENTRADA EM DADOS

Em termos amplos, os desafios institucionais da implementação de IA centrada em dados nas instituições públicas podem ser divididos em três categorias: legal, regulatória e de disponibilidade de recursos humanos. Os desafios legais surgem das questões inerentes à modelagem de dados e aprendizagem de máquina, tais como ética, privacidade de dados, viés e discriminação, transparência e explicabilidade, *accountability* e propriedade intelectual. Muitos países ainda não conseguiram renovar suas políticas em torno dessas questões, e está se tornando extremamente difícil para as instituições legais acompanhar o ritmo da rápida transformação tecnológica em andamento.

Existem vários regulamentos proeminentes, tais como o RGPD, a Lei de Privacidade do Consumidor da Califórnia (California Consumer Privacy Act [CCPA]) ou o marco referencial de privacidade da Cooperação Econômica Ásia-Pacífico (Asia-Pacific Economic Cooperation [APEC]) atualmente em vigor. Países como Brasil, Índia, Austrália e Canadá também adotaram suas próprias leis de proteção de dados. Além da proteção de dados e da privacidade pessoal, há diversas iniciativas regulatórias que giram em torno de considerações éticas sobre o uso da IA; conformidade com regulamentos como a Lei de Portabilidade e Responsabilidade dos Seguros de Saúde (Health Insurance Portability and Accountability Act [HIPAA]), na área da saúde, ou a Lei dos Direitos Educacionais e da Privacidade da Família (Family Educational Rights and Privacy Act [FERPA]), para instituições educacionais nos Estados Unidos; questões relacionadas a quem assume a responsabilidade pelas decisões tomadas por modelos de IA; e regulamentos de aquisição, como o Regulamento Federal de Aquisições (Federal Acquisition Regulation [FAR]), nos Estados Unidos. As instituições públicas precisam enfrentar esses desafios para que possam alavancar o uso dos dados com vistas a tomar melhores decisões, aperfeiçoar os serviços e aprimorar o atendimento às suas partes interessadas, garantindo tanto a conformidade com os regulamentos quanto a proteção à privacidade e à segurança.

Além dos desafios legais e regulatórios na esfera das políticas, há vários desafios de recursos humanos na esfera da implementação. Os órgãos públicos seguem processos que são lentos em comparação com organizações privadas. Isto significa que eles são morosos em sua resposta às mudanças tecnológicas. Para que os funcionários públicos formulem as políticas públicas a serem adotadas, eles têm de compreender e ser capazes de explorar o potencial da IA e dos dados. Precisam conhecer as oportunidades oferecidas pela IA, estando ao mesmo tempo cientes dos riscos e desafios. Trabalhar com IA e dados requer habilidades especializadas, como no caso de cientistas de dados, engenheiros de aprendizagem de máquina, desenvolvedores de *software* e especialistas em políticas de engenharia. As organizações devem planejar

Uma vez que a adoção de IA pode envolver mudanças significativas nos processos, fluxos e funções de trabalho estabelecidos, é preciso que as organizações criem um plano de gestão de mudanças para ajudar os funcionários a navegá-las de modo efetivo.

a contratação e a retenção desses profissionais qualificados. Além disso, precisam treinar e aumentar a qualificação da força de trabalho existente. A implementação da IA pode resultar na substituição de ocupações passíveis de automatização. Cabe às organizações o desenvolvimento de um plano de requalificação e redistribuição de funcionários cujos empregos são afetados pela IA centrada em dados. Uma vez que a adoção de IA pode envolver mudanças significativas nos processos, fluxos e funções de trabalho estabelecidos, é preciso que as organizações criem um plano de gestão de mudanças para ajudar os funcionários a navegá-las de modo efetivo.

Para que a transformação digital baseada em IA e dados seja bem-sucedida, os governos devem mudar a maneira como funcionam. Isto é difícil. Entretanto, é possível começar, caso funcionários públicos em diferentes níveis adquiram competências para entender as transformações que a IA centrada em dados traz. É importante, portanto, expandir a conscientização sobre as habilidades exigidas nos diferentes níveis. Isto pode ser feito por meio da compreensão das necessidades de capacitação na esfera do indivíduo, da equipe, do departamento e do governo. A colaboração e comunicação crescentes entre as instituições ajudariam os departamentos a compartilhar *insights*. E, mais significativamente, deve haver um monitoramento contínuo do impacto das iniciativas de capacitação.

Conclusões

Os dados são o ingrediente mais significativo do progresso nesta era da Inteligência Artificial. A IA está gradualmente se tornando uma tecnologia-chave para as organizações de serviço público, uma vez que aumenta a eficiência administrativa. Sua capacidade de fazer uso do enorme número de dados em seus vários tipos e encontrar *insights* que ajudam a automatizar processos contribui para a tomada de decisão.

No entanto, embora sejam observados desenvolvimentos positivos, vários desafios surgem. A IA tem fome de dados. Portanto, um fluxo contínuo de dados de qualidade deve ser assegurado para que as ferramentas de IA não tomem decisões enviesadas ou incorretas. Entre os fatores críticos, a disponibilidade e a qualidade dos dados são a necessidade mais proeminente para treinar os sistemas de IA de maneira adequada. Tais "necessidades de dados" exigem o estabelecimento de uma estratégia de governança de dados para o uso de dados internos, bem como dados potenciais de outras organizações, e envolvem a avaliação da conformidade com regulações de proteção de dados.

Há diversos algoritmos de IA capazes de funcionar bem com conjuntos de dados pequenos ou grandes. No entanto, a IA é uma ciência nascente. Essas soluções devem ser examinadas antes de demandarem uso na vida real, especialmente em relação às limitações e aos riscos da IA, assim como à contrapartida entre automação de processos *versus* controle humano.

Além disso, as diferenças metodológicas entre a IA e o desenvolvimento de *software* tradicional colocam desafios para as instituições que realizam os projetos. Em particular, a transparência e a "explicabilidade" da aplicação da IA constituem uma questão importante, com destaque para as decisões que impactam as pessoas.

A transformação digital centrada em dados pode acontecer se as instituições estiverem preparadas para as mudanças que os dados e a IA trazem. As autoridades que contemplam tal transformação têm de considerar os desafios legais, regulatórios e institucionais. Os governos devem avaliar as competências de seus funcionários públicos e enfatizar a capacitação para assegurar uma transição suave em direção à transformação digital centrada em dados.

(...) a transparência e a "explicabilidade" da aplicação da IA constituem uma questão importante, com destaque para as decisões que impactam as pessoas.

Referências

Abraham, R., Brocke, J., & Schneider, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management, 49*(S1).

Amodei, D., & Hernandez, D. (2018). *AI and compute*. OpenAI. https://openai.com/research/ai-and-compute

Bienvenido-Huertas, D., Farinha, F., Oliveira, M., Silva, E., & Lança, R. (2020). Comparison of Artificial Intelligence algorithms to estimate sustainability indicators. *Sustainable Cities and Societies, 63*.

Defence Science and Technology Laboratory. (2020). Machine Learning with Limited Data. *Future of AI for Defense project*. https://www.gov.uk/government/publications/machine-learning-with-limited-data

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems,* 1097-1105.

LITSLINK. (2019, September 19). *An introduction to Machine Learning Algorithms*. https://litslink.com/blog/an-introduction-to-machine-learning-algorithms

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. MIT Press.

Stanford Institute for Human-Centered Artificial Intelligence. (2022). *The AI index report: Measuring trends in Artificial Intelligence*. https://aiindex.stanford.edu/report/

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. *MIT Initiative on the Digital Economy Research Brief*, *4*. https://ide.mit.edu/wp-content/uploads/2020/09/RBN.Thompson.pdf

Tjoa, E., & Guan, C. (2021). A survey on Explainable Artificial Intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11).

Entrevista II

Foto: Arquivo pessoal

Anita Lübbe
Juíza do Tribunal Regional do Trabalho da 4ª Região (TRT-4).

Gestão documental e memória no Poder Judiciário

Nesta entrevista, Anita Lübbe – juíza do Tribunal Regional do Trabalho da 4ª Região (TRT-4), presidente do Fórum Nacional Permanente em Defesa da Memória da Justiça do Trabalho (Memojutra) e coordenadora do Subcomitê de Preservação Digital do Comitê do Programa Nacional de Gestão Documental e Memória do Poder Judiciário (Proname) – aborda sua experiência na gestão documental e memória no Poder Judiciário, bem como as políticas implementadas para preservação e o acesso a seus acervos documentais.

Panorama Setorial da Internet (P.S.I.)_ Considerando a digitalização dos processos e acervos do Poder Judiciário, bem como as possibilidades de armazenamento facilitadas pelas tecnologias digitais, qual a importância do estabelecimento de políticas para a governança dos dados e a gestão do ecossistema de informações judiciais?

Anita Lübbe (A.L.)_ A importância está justamente no fato de que esse "armazenamento" só se tornará uma ação efetiva de preservação quando for decorrente de uma política consistente de gestão documental e memória, seguindo a legislação aplicável e as diretrizes técnicas. De forma resumida, uma gestão de dados abrange ações que incluem: a geração dos documentos tanto administrativos quanto judiciais; a classificação, por meio da aplicação de Tabelas Processuais Unificadas aos processos; a aplicação do Plano de Classificação e Tabela de Temporalidade de Documentos da Administração do Poder Judiciário (PCTTDA)[12]; e o recolhimento seguro dos documentos selecionados como de guarda permanente.

O Artigo 16 da Resolução CNJ n. 324/2020[13] (Conselho Nacional de Justiça) indica a classificação dos documentos do Poder Judiciário como correntes, intermediários ou permanentes, aliada à consequente definição de seus prazos de guarda conforme aplicação da respectiva Tabela de Temporalidade. Em caso de descarte, é preciso considerar ainda o Plano para Amostra Estatística Representativa[14].

Temos um conjunto de leis e normativos que precisam ser observados para estabelecer e manter políticas de governança de dados e preservação de documentos e de memória, a iniciar pelo Artigo 216 da Constituição Federal; a Lei de Arquivos (Lei n. 8.159/1991); a Lei de Acesso à Informação (LAI – Lei

[12] Disponível em: https://www.cnj.jus.br/programas-e-acoes/gestao-documental-e-memoria-proname/gestao-documental/tabelas-de-temporalidade-da-area-administrativa/

[13] Disponível em: https://atos.cnj.jus.br/atos/detalhar/3376

[14] Saiba mais: https://www.cnj.jus.br/programas-e-acoes/gestao-documental-e-memoria-proname/gestao-documental/instrumentos-do-proname/

n. 12.527/2011); a Lei Geral de Proteção de Dados Pessoais (LGPD – Lei n. 13.709/2018); resoluções do Conselho Nacional de Arquivos (Conarq) e do CNJ, em especial a já referida Resolução n. 324/2020, que instituiu parâmetros, definições e regras objetivas de gestão documental e memória para o Poder Judiciário. Chamo atenção para os Manuais de Gestão Documental e de Gestão de Memória do Poder Judiciário[15], publicados pelo CNJ em fevereiro de 2021 visando possibilitar o gerenciamento e a operacionalização das disposições contidas na Resolução CNJ n. 324/2020.

De acordo com a legislação e os normativos mencionados, as políticas de gestão documental e memória devem ser instituídas por todos os tribunais, efetivando o dever do Estado de guarda dos documentos e o direito do cidadão de acesso à informação. Para tal, são fundamentais a correta classificação dos documentos, a aplicação das Tabelas de Temporalidade e dos planos amostrais, além da implantação de Repositórios Arquivísticos Digitais Confiáveis (RDC-Arq)[16]. Por meio de uma adequada política de gestão documental e memória, é possível identificar e criar no âmbito do Judiciário variados conjuntos de ecossistemas de informação.

P.S.I._ Quais são os principais aspectos a se considerar tanto no estabelecimento de sistemas de gestão documental e memória quanto no devido armazenamento de dados jurídicos?

A.L._ Destaco a formação de equipes com profissionais habilitados para promover desde o início de seus planos de ação um diagnóstico sobre a extensão do acervo a ser considerado, bem como para interpretar o diagnóstico e realizar ações de seleção, classificação, arquivamento adequado e plano amostral em caso de descarte dos documentos. Imprescindível a participação de profissionais das áreas de arquivologia, história, biblioteconomia, museologia e ciência da informação, fortalecendo em cada instituição uma cultura de preservação em seus variados aspectos. É de extrema importância a inclusão desses profissionais nos quadros permanentes dos tribunais.

No que diz respeito aos requisitos técnicos dos sistemas de informação, cabe salientar que o CNJ está em fase final de revisão do Modelo de Requisitos para Sistemas Informatizados de Gestão de Processos e Documentos do Poder Judiciário (MoReq-Jus)[17], que elenca de forma minuciosa todos os requisitos obrigatórios ou desejáveis a serem considerados. Importante e necessária também é a implantação dos RDC-Arq, repositórios para onde são remetidos e arquivados os documentos, garantindo a sua preservação, a cadeia de custódia e o adequado acesso. Em março de 2023, o CNJ lançou o Manual de Digitalização de Documentação do Poder Judiciário[18], que deverá ser utilizado em conjunto com os manuais de gestão documental e de gestão de memória.

---

[15] Disponíveis em: https://www.gov.br/arquivonacional/pt-br/cnj-lanca-manuais-para-gestao-de-documentos-e-da-memoria-do-judiciario

[16] Saiba mais: https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/conarq_diretrizes_rdc_arq_resolucao_43.pdf

[17] Saiba mais: https://www.cnj.jus.br/poder-judiciario/consultas-publicas/modelo-de-requisitos-para-sistemas-informatizados-de-gestao-de-processos-e-documentos-do-judiciario-brasileiro-moreq-jus/

Nesse aspecto, merece reconhecimento o fato de o Tribunal de Justiça do Distrito Federal e dos Territórios (TJDFT) ter sido o primeiro no país a iniciar a implementação de um RDC-Arq, em 2018, em parceria com o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), criando um paradigma a ser seguido. Desde o início, o TJDFT passou a mensagem clara de que suas soluções estavam e estão à disposição dos demais tribunais. No entanto, o Processo Judicial Eletrônico (PJe) utilizado pelos cinco ramos do Judiciário – Justiça Estadual, Justiça do Trabalho, Justiça Federal, Justiça Eleitoral e Justiça Militar – não é exatamente o mesmo, e cada ramo tem matérias e competências específicas, o que em grande medida dificulta a definição de um único RDC-Arq para todos.
Em meados de 2022, o Conselho Superior da Justiça do Trabalho (CSJT) e o TRT-4 iniciaram seu projeto de implantação do RDC-Arq, também em parceria com o IBICT. O objetivo é aplicar e instalar o RDC-Arq em todos os 24 tribunais regionais do trabalho, tendo como tribunal piloto o TRT-4, no Rio Grande do Sul. No caso da Justiça do Trabalho, a iniciativa se fortalece ao visar à unificação do RDC-Arq aplicado nesses tribunais, uma vez que se tem por base o mesmo modelo de PJe e ramo do Judiciário. O projeto tem previsão de cinco anos, sendo os dois primeiros dedicados à criação das soluções necessárias e outros três anos para o acompanhamento da implantação nos tribunais.

P.S.I._ Poderia nos contar um pouco sobre suas experiências em iniciativas de articulação entre diferentes instituições e atores para a gestão dos dados e da memória no Judiciário? Quais foram os principais aprendizados obtidos nessas iniciativas?

A.L._ A partir do Memorial do TRT-4[19], criado pela Resolução Administrativa TRT-4 n. 22/2003, presidido na época pela Ministra Rosa Weber, atual Presidente do Supremo Tribunal Federal (SFT) e CNJ; iniciou-se o processo de preservação de memória. A seguir integrei a Comissão Coordenadora do Memorial. Importante registrar que já em meados dos anos 1990 começaram em vários tribunais de todos os ramos do Judiciário movimentos de preservação de memória, com o estabelecimento de espaços e projetos iniciais de preservação documental.
Em 2006, foi criado o Fórum Nacional Permanente em Defesa da Memória da Justiça do Trabalho (Memojutra), do qual participo desde então e atualmente presido. É formado por magistrados e servidores do Tribunal Superior do Trabalho (TST), do Conselho Superior da Justiça do Trabalho (CSJT) e dos 24 tribunais regionais do trabalho, com participação permanente de todos os coordenadores dos respectivos Centros de Memória.
No Memojutra, lembro o compartilhamento quanto à importância da seleção e apresentação dos acervos dos tribunais regionais do trabalho no âmbito do Programa Memória do Mundo (MoW), da Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO). A iniciativa ocorreu a partir do exemplo do Tribunal

[18] Disponível em: https://www.cnj.jus.br/wp-content/uploads/2023/03/proname-manual-digitalizacao-15-03-2023.pdf
[19] Disponível em: https://www.trt4.jus.br/portais/memorial

Regional do Trabalho da 6ª Região (TRT-6), em Pernambuco, o primeiro tribunal da Justiça do Trabalho a receber o selo Memória do Mundo como patrimônio documental da humanidade, em 2012.

Em anos seguintes, diversas instituições no Brasil obtiveram esse reconhecimento da UNESCO. Entre elas: o TRT-4, em 2014; o Tribunal Regional do Trabalho da 3ª Região (TRT-3), em Minas Gerais, em 2015; e, em 2016, o Tribunal Superior do Trabalho (TST). Ainda sobre ações importantes para a preservação da memória, cito a extinta Câmara Setorial sobre Arquivos do Judiciário (CSAJ), do Conarq, da qual fiz parte brevemente em 2019.

Atualmente, integro o Comitê do Programa Nacional de Gestão Documental e Memória do Poder Judiciário (Proname), do qual participo desde 2019. Em 2022, passei a coordenar o Subcomitê de Preservação Digital. No Comitê, temos promovido ações de atualização e criação de normativos voltados à gestão documental e de memória, tais como: a Resolução n. 316/2020, que estabeleceu o Dia da Memória do Poder Judiciário e a realização anual do Encontro Nacional de Memória do Poder Judiciário (ENAM); a Resolução CNJ n. 429/2021, que instituiu o Prêmio CNJ Memória do Poder Judiciário (ENAM); além da Resolução n. 324/2020 e dos Manuais de Gestão Documental e de Gestão de Memória do Poder Judiciário, já mencionados.

Penso que o principal aprendizado tem sido a interlocução estabelecida entre todos os integrantes desses espaços: magistrados, servidores, profissionais das áreas de história, arquivologia, museologia, biblioteconomia, tecnologia da informação, ciência da informação, entre outros. Com suas expertises, esses atores nos ensinam e possibilitam o aprimoramento da gestão documental e de memória, garantindo o acesso e a segurança da informação.

P.S.I._ Em sua opinião, qual o grau de maturidade do Judiciário brasileiro no que se refere à gestão de memória e à disponibilização de seus acervos para um público mais amplo?

A.L._ O Judiciário tem feito várias ações de preservação e acesso de seus acervos nas últimas décadas. Passamos gradativamente do processo físico para o processo eletrônico, com a implantação do Sistema de Processo Judicial Eletrônico a partir da Resolução CNJ n. 185/2013. Embora já tenhamos o processo eletrônico em todos os ramos do Judiciário, a tramitação eletrônica não é a totalidade. Temos ainda um passivo de processos físicos que precisam passar por classificação, definição de seus prazos de guarda e digitalização, para que então sejam preservados em repositórios digitais confiáveis. Hoje, há em todo o Judiciário nacional um número expressivo de processos natodigitais (isto é, criados no suporte digital) e outros em formato físico (criados no suporte papel). Entre estes últimos, uma parte já está digitalizada, enquanto outra parte expressiva segue em fase de digitalização, definição de temporalidade e, conforme o caso, eliminação, com a consequente formação de plano amostral estatístico.

Além das muitas iniciativas de preservação de sua memória, o CNJ tem realizado os ENAM, conforme regulamentado na Portaria CNJ n. 80/2022. A primeira edição, em maio de 2021, ocorreu de forma virtual devido à pandemia COVID-19. No ano seguinte, em maio de 2022, o II ENAM aconteceu presencialmente no Tribunal de Justiça de Pernambuco (TJPE), em Recife, com

expressivo número de participantes entre magistrados, servidores, profissionais de áreas diversas – como história, arquivologia, biblioteconomia, museologia, tecnologia da informação e ciência da informação – e estudantes. Também presencial, a terceira edição (III ENAM) será de 10 a 12 de maio de 2023[20], em Porto Alegre (RS), tendo como anfitriões os cinco tribunais ali situados: Tribunal de Justiça do Rio Grande do Sul (TJ-RS), Tribunal Regional Federal da 4ª Região (TRF-4), Tribunal Regional Eleitoral do Rio Grande do Sul (TRE-RS) e Tribunal de Justiça Militar do Rio Grande do Sul (TJM-RS). Com o tema "Estruturando a memória", o evento contará com a presença com a presença da Ministra Rosa Weber no encerramento no dia 12 de maio. O objetivo é fornecer a todos os tribunais ferramentas e alternativas para auxiliar na implementação ou ampliação das atividades de preservação da memória do judiciário nacional, bem como no aperfeiçoamento da gestão documental e de memória de cada instituição. Um dos diferenciais do III ENAM é o pré-encontro *online* nos dias 13 e 14 de abril de 2023, com o objetivo de atualizar os inscritos quanto a conceitos e alterações legislativas recentes (incluindo normativos do CNJ), motivando os participantes para as palestras presenciais do evento em maio. O tema da gestão de memória e de documentos do Judiciário é, sem dúvida, um belo e instigante caminho a percorrer.

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), departamento do Núcleo de Informação e Coordenação do Ponto BR (NIC.br), monitora mensalmente o número de nomes de domínios de topo de código de país (*country code Top-Level Domain* [ccTLD]) registrados entre os países que compõem a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) e o G20[21]. Considerados os membros de ambos os blocos, as 20 nações com maior atividade somam mais 89,80 milhões de registros. Em março de 2023, os domínios registrados sob .de (Alemanha) chegaram a 17,49 milhões. Em seguida, aparecem Reino Unido (.uk), China (.cn) e Países Baixos (.nl), com, respectivamente, 9,65 milhões, 7,19 milhões e 6,29 milhões de registros. O Brasil teve 5,09 milhões de registros sob .br, ocupando a quinta posição na lista, como mostra a Tabela 1[22].

[20] Saiba mais: https://sites.google.com/trt4.jus.br/enam

[21] Grupo composto pelas 19 maiores economias mundiais e a União Europeia. Saiba mais: https://g20.org/

[22] A tabela apresenta a contagem de domínios ccTLD segundo as fontes indicadas. Os valores correspondem ao registro publicado por cada país, tomando como base os membros da OCDE e do G20. Para países que não disponibilizam uma estatística oficial fornecida pela autoridade de registro de nomes de domínios, a contagem foi obtida em: https://research.domaintools.com/statistics/tld-counts. É importante destacar que há variação no período de referência, embora seja sempre o mais atualizado para cada localidade. A análise comparativa de desempenho de nomes de domínios deve considerar ainda os diferentes modelos de gestão de registros ccTLD. Assim, ao observar o *ranking*, é preciso atentar para a diversidade de modelos de negócio existentes.

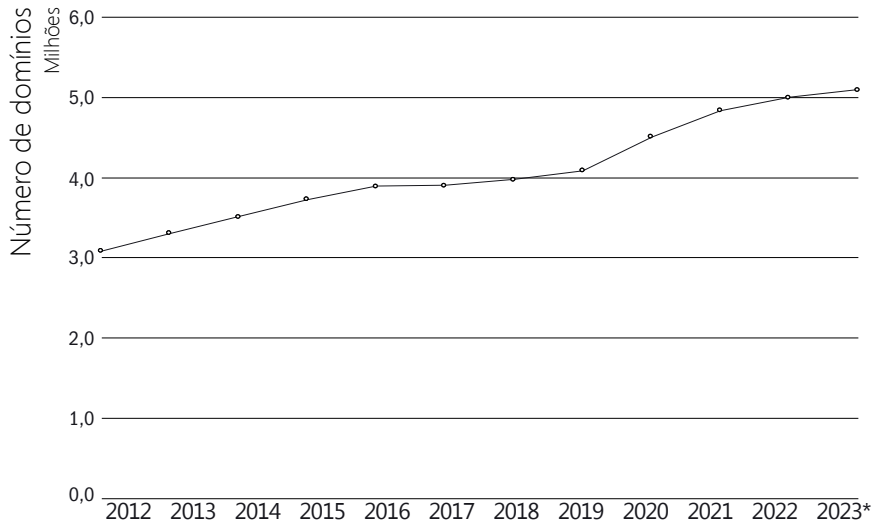
Tabela 1 – TOTAL DE REGISTROS DE NOMES DE DOMÍNIOS ENTRE OS PAÍSES DA OCDE E DO G20

Posição	País	Número de domínios	Data de referência	Fonte (*website*)
1	Alemanha (.de)	17.498.904	31/03/2023	https://www.denic.de
2	Reino Unido (.uk)	9.659.204	28/02/2023	https://www.nominet.uk/news/reports-statistics/uk-register-s-tatistics-2023/
3	China (.cn)	7.193.640	31/03/2023	https://research.domaintools.com/statistics/tld-counts/
4	Países Baixos (.nl)	6.295.609	31/03/2023	https://stats.sidnlabs.nl/en/registration.html
5	**Brasil (.br)**	**5.094.470**	**31/03/2023**	**https://registro.br/dominio/estatisticas/**
6	Rússia (.ru)	4.935.204	31/03/2023	https://cctld.ru
7	Austrália (.au)	4.214.524	31/03/2023	https://www.auda.org.au/
8	França (.fr)	3.975.180	31/03/2023	https://research.domaintools.com/statistics/tld-counts/
9	União Europeia (.eu)	3.669.172	31/03/2023	https://research.domaintools.com/statistics/tld-counts/
10	Itália (.it)	3.493.029	31/03/2023	http://nic.it
11	Colômbia (.co)	3.365.252	31/03/2023	https://research.domaintools.com/statistics/tld-counts/
12	Canadá (.ca)	3.361.681	31/03/2023	https://www.cira.ca
13	Índia (.in)	2.893.157	31/03/2023	https://research.domaintools.com/statistics/tld-counts/
14	Suíça (.ch)	2.535.872	15/03/2023	https://www.nic.ch/statistics/domains/
15	Polônia (.pl)	2.509.765	31/03/2023	https://www.dns.pl/en/
16	Espanha (.es)	2.024.766	20/03/2023	https://www.dominios.es/dominios/en
17	Estados Unidos da América (.us)	1.932.390	31/03/2023	https://research.domaintools.com/statistics/tld-counts/
18	Bélgica (.be)	1.746.750	31/03/2023	https://www.dnsbelgium.be/en
19	Japão (.jp)	1.728.299	01/03/2023	https://jprs.co.jp/en/stat/
20	Portugal (.pt)	1.678.130	31/03/2023	https://www.dns.pt/en/statistics/

Data de coleta: 31 de março de 2023.

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

Gráfico 1 – TOTAL DE REGISTROS DE DOMÍNIOS DO .BR – 2012 a 2023*



*Data de coleta: 31 de março de 2023.
Fonte: Registro.br
Recuperado de: https://registro.br/dominio/estatisticas/

Em março de 2023, os cinco principais domínios genéricos (*generic Top-Level Domain* [gTLD]) totalizaram mais de 190,08 milhões de registros. Com 159,71 milhões de registros, destaca-se o .com, conforme apontado na Tabela 2.

Tabela 2 – TOTAL DE REGISTROS DE DOMÍNIOS DOS PRINCIPAIS gTLD

Posição	gTLD	Número de domínios
1	**.com**	159.717.469
2	**.net**	13.030.989
3	**.org**	10.754.641
4	**.info**	3.738.226
5	**.xyz**	3.649.013

Data de coleta: 31 de março de 2023.
Fonte: DomainTools.com
Recuperado de: research.domaintools.com/statistics/tld-counts

/Tire suas dúvidas

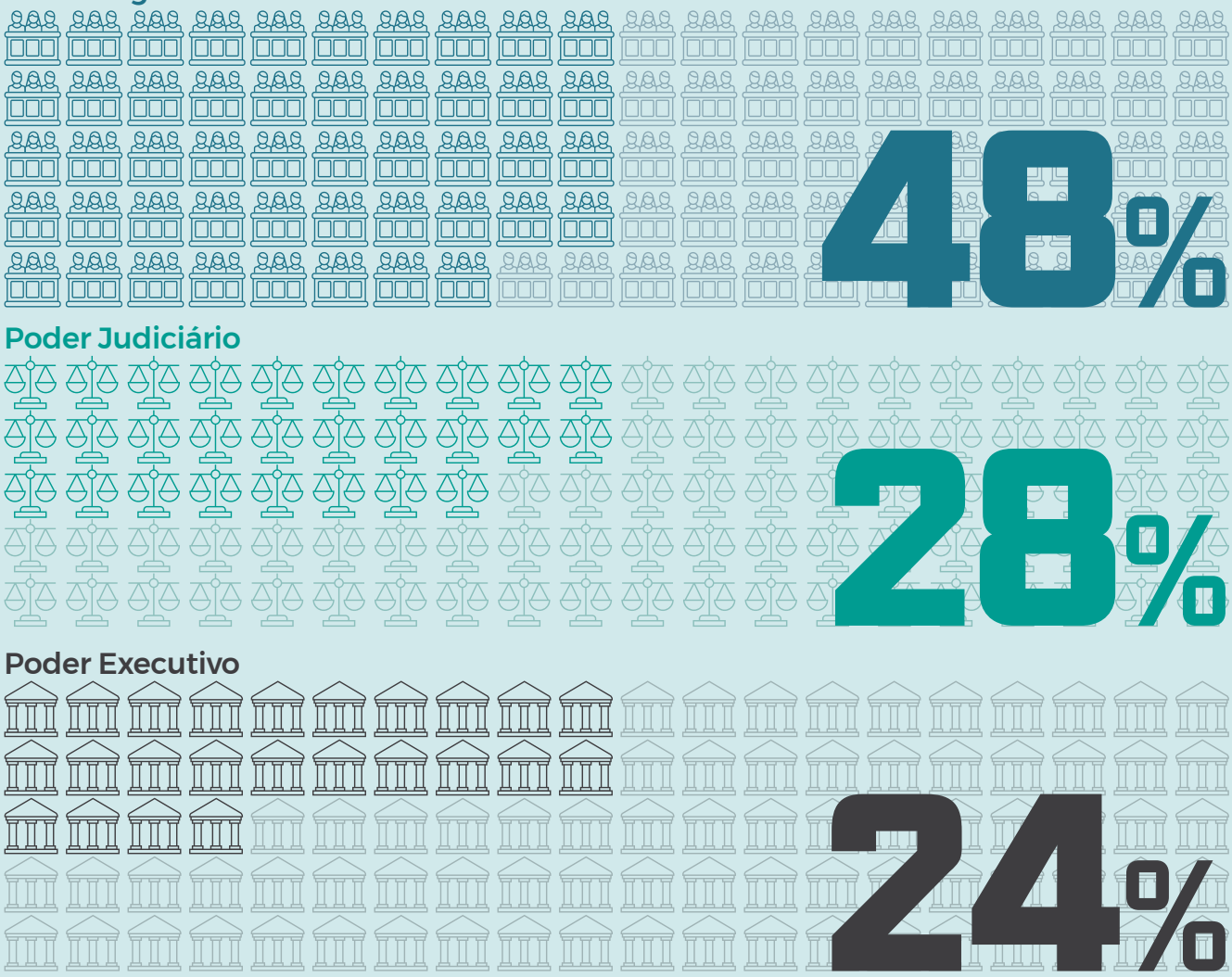
ANÁLISES DE *BIG DATA* NO SETOR PÚBLICO

Com a crescente adoção de tecnologias digitais, uma grande quantidade de dados é gerada por indivíduos, máquinas, sistemas e sensores. Nesse contexto, um novo ecossistema de dados vem se consolidando, no qual fontes de *Big Data* possibilitam análises e informações inéditas para o aprimoramento dos processos de tomada de decisão, inclusive no setor público.

Em 2021, 25% dos órgãos públicos federais e estaduais realizaram análises de *Big Data*. Os indicadores[23] a seguir apresentam o cenário de uso e não uso de *Big Data* pelo setor público brasileiro.

Órgãos públicos federais e estaduais que realizaram análises de *Big Data*, por poder

Total de órgãos públicos federais e estaduais (2021)



[23] Dados da pesquisa TIC Governo Eletrônico 2021, do Cetic.br|NIC.br. Disponível em: https://cetic.br/pt/pesquisa/governo-eletronico

Órgãos públicos federais e estaduais, por motivos para não realizar análises de *Big Data*[24] e poder

Total de órgãos públicos federais e estaduais (2021)

	Falta de pessoas capacitadas no órgão público para realizar análises de *Big Data*	Não é uma prioridade para o órgão público	Falta de necessidade ou de interesse
Total	41%	31%	30%
Poder Legislativo	33%	20%	30%
Poder Judiciário	64%	54%	43%
Poder Executivo	40%	30%	30%

[24] Outros motivos para não realizar análises de *Big Data* coletados pela pesquisa TIC Governo Eletrônico 2021 podem ser conhecidos em: https://cetic.br/pt/tics/governo/2021/orgaos/H1B/

/Créditos

REDAÇÃO

RELATÓRIO DE DOMÍNIOS
Thiago Meireles (Cetic.br | NIC.br)

INFOGRAFIA E DIAGRAMAÇÃO

Giuliano Galves, Larissa Paschoal e Maricy Rabelo (Comunicação | NIC.br)

EDIÇÃO DE TEXTO EM PORTUGUÊS

Mariana Tavares

TRADUÇÃO INGLÊS-PORTUGUÊS

Ana Zuleika Pinheiro Machado e Robert Dinham

COORDENAÇÃO EDITORIAL

Alexandre F. Barbosa, Graziela Castello, Javiera F. M. Macaya e Mariana Galhardo (Cetic.br | NIC.br)

AGRADECIMENTOS

Anita Job Lübbe (TRT-4)

Carolina Rossini (The Datasphere)

Datasphere Initiative

Delfina Soares (UNU EGOV)

Marlova Jovchelovitch Noleto e Adauto Candido Soares (Representação da UNESCO no Brasil)

Moinul Zaber (UNU EGOV)

Pedro Luis Nascimento da Silva (SCIENCE)

SOBRE O CETIC.br

O Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação – Cetic.br (https://www.cetic.br/), departamento do NIC.br, é responsável pela produção de estudos e estatísticas sobre o acesso e o uso da Internet no Brasil, divulgando análises e informações periódicas sobre o desenvolvimento da rede no país. O Cetic.br atua sob os auspícios da UNESCO.

SOBRE O NIC.br

O Núcleo de Informação e Coordenação do Ponto BR – NIC.br (https://nic.br/) é uma entidade civil de direito privado e sem fins de lucro, encarregada da operação do domínio .br, bem como da distribuição de números IP e do registro de Sistemas Autônomos no país. Conduz ações e projetos que trazem benefícios à infraestrutura da Internet no Brasil.

SOBRE O CGI.br

O Comitê Gestor da Internet no Brasil – CGI.br (https://cgi.br/), responsável por estabelecer diretrizes estratégicas relacionadas ao uso e desenvolvimento da Internet no Brasil, coordena e integra todas as iniciativas de serviços Internet no país, promovendo a qualidade técnica, a inovação e a disseminação dos serviços ofertados.

*As ideias e opiniões expressas nos textos dessa publicação são as dos respectivos autores e não refletem necessariamente as do NIC.br e do CGI.br.

unesco
Centro sob os auspícios da UNESCO

cetic.br
Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação

nic.br
Núcleo de Informação e Coordenação do Ponto BR

cgi.br
Comitê Gestor da Internet no Brasil

ISSN - 2965-2642

POR UMA INTERNET CADA VEZ MELHOR NO BRASIL

CGI.BR, MODELO DE GOVERNANÇA MULTISSETORIAL

https://cgi.br

nic.br cgi.br

FOTOS: RICARDO MATSUKAWA E SHUTTERSTOCK