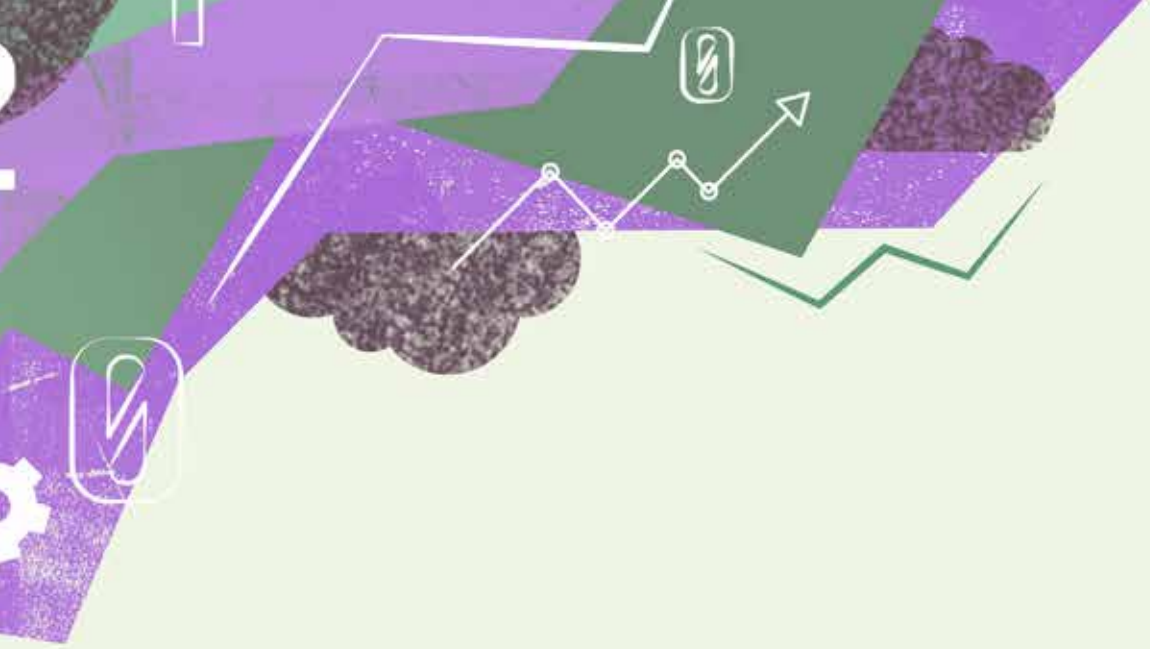
**Ilustrações** Ágatha Kretli



# DESMISTIFICANDO O APRENDIZADO DE MÁQUINA

POR ***FELIPE CRUZ, AGOSTINHO VILLELA E HUGO TADEU***

Durante uma palestra sobre aprendizado de máquina e inteligência artificial, aberta a seção de perguntas e respostas, uma das pessoas da plateia se levanta e afirma: "Isso tudo é muito interessante, mas é como magia negra: ainda bastante inacessível e longe da realidade". O interessante é que esse pensamento não está muito longe do ponto de vista de diversas pessoas ao redor do mundo, quando os dois temas entram em discussão.

Muita gente, quando ouve falar de inteligência artificial, é automaticamente remetida à visão clássica do cinema – os androides de "Star Wars" ou os exterminadores de "O Exterminador do Futuro". O que poucos sabem é que, muito provavelmente, não passam um dia sequer sem interagir com uma inteligência artificial. Logo que você acorda, por exemplo, provavelmente sua caixa de e-mails já terá armazenado várias mensagens do comércio *on-line*, recomendando produtos que podem ser do seu interesse. Todas essas recomendações são geradas automaticamente, com base em seu perfil de compras, feitas naquela mesma loja, ou em buscas que você fez recentemente e nos sites que visitou. Ao mesmo tempo, um algoritmo de aprendizado de máquina filtra várias dessas mensagens e as coloca em sua caixa de *spam*, com base no conteúdo do texto, remetente e outras variáveis. Quando você sai para almoçar e passa seu cartão de crédito na máquina do restaurante, um modelo preditivo da operadora analisa a transação para saber se não é fraudulenta. E, à noite, quando chega em casa e decide assistir a um filme, sua plataforma de *streaming* utiliza outro modelo de aprendizado para sugerir temas que podem ser do seu gosto, baseando-se no que você assistiu anteriormente. Isso só para mencionar algumas situações. Ou seja: antes mesmo de acordar, você já interagiu com pelo menos dois modelos de *machine learning* e ainda vai interagir com muitos outros, ao longo do dia.


**HISTÓRIA DO APRENDIZADO DE MÁQUINA** Engana-se quem pensa que essas recomendações e inteligência são frutos de magia negra. Na verdade, a maior parte desses modelos é amplamente estudada há bastante tempo. Por exemplo, o filtro de *spam* em e-mails, que comentamos antes, é tradicionalmente aplicado, utilizando um modelo de aprendizado de máquina conhecido como Naive Bayes. Esse algoritmo vem sendo estudado desde a década de 1950, com ampla aplicabilidade no campo de classificação textual e até mesmo em diagnósticos médicos, destacando-se por sua simplicidade e eficiência. A base do Naive Bayes é a aplicação do Teorema de Bayes, publicado inicialmente em 1763, que utiliza a frequência de palavras e a probabilidade de eventos, desenvolvida neste teorema para construir um classificador.

Um dos métodos de aprendizado de máquina mais conhecidos, e largamente estudado em cursos de estatística e econometria, é a regressão linear para previsão de valores contínuos, como, por exemplo, o número de unidades vendidas no futuro com base em diversas variáveis. O método dos mínimos quadrados, o mais tradicional na construção de modelos de regressão linear, foi desenvolvido e publicado dois séculos atrás, por Legendre (1805) e Gauss (1809).

Métodos hoje bastante em voga, mesmo que mais modernos – como as redes neurais artificiais –, também tiveram sua origem há algum tempo. O estudo dessas redes começou na década de 1940, na tentativa de construção de modelos matemáticos

para simular o funcionamento de neurônios, e evoluiu até a criação do algoritmo do *perceptron*, em 1958, que permitiria a simulação de um neurônio com capacidade de aprendizado. Esse algoritmo, considerado na época uma revolução no campo da inteligência artificial, foi estudado e aperfeiçoado extensivamente até os dias atuais, sendo suas versões aprimoradas à base das redes neurais modernas e do tão comentado *deep learning*.

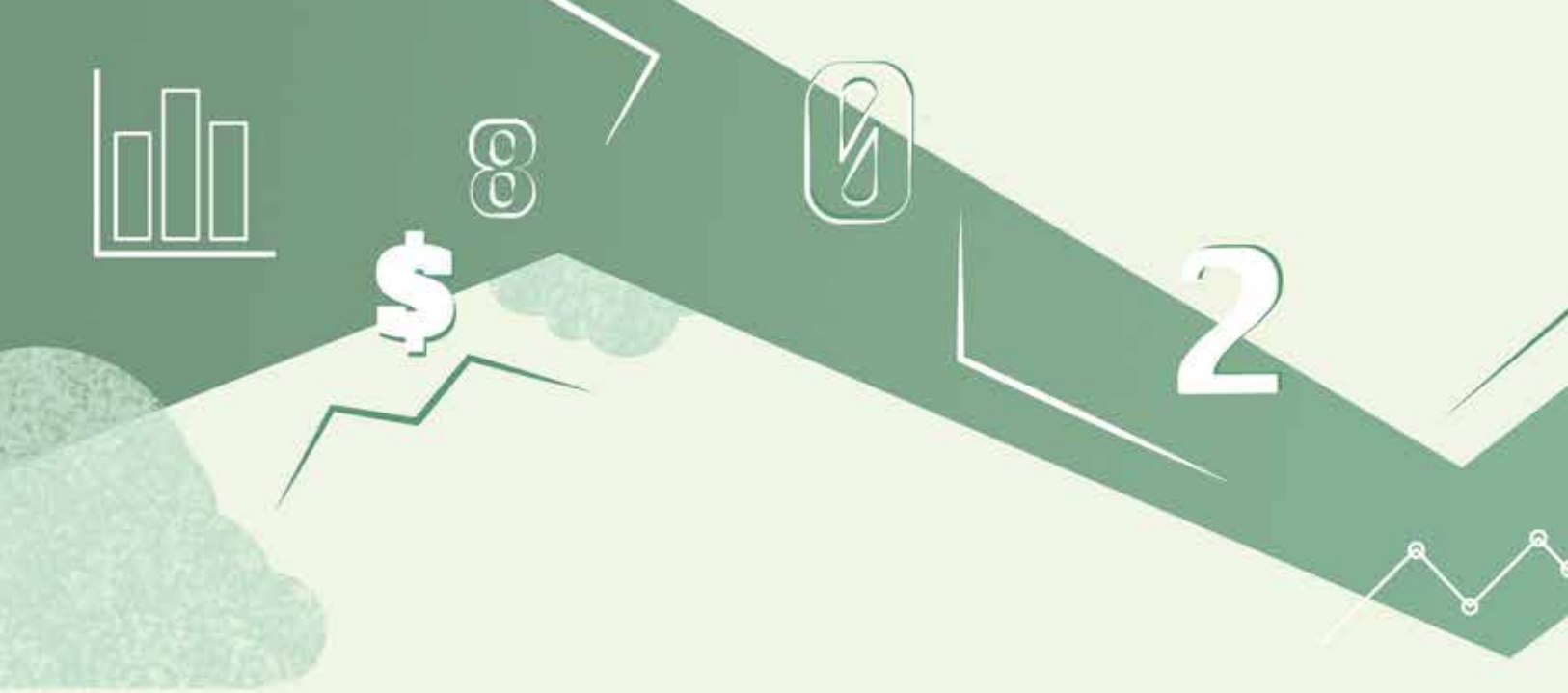
**APRENDIZADO DE MÁQUINA E *BIG DATA*** Se os algoritmos e a base matemática do aprendizado de máquina são tão antigos, porque só recentemente, em especial na última década, estamos vendo o tema da inteligência artificial e *machine learning* retomando uma posição tão importante? A resposta vem, principalmente, do avanço do poder computacional, o aumento da capacidade de armazenamento de dados e a constante evolução nas técnicas de aprendizado de máquina.

Anteriormente, o armazenamento de dados e a capacidade de processamento eram muito caros. Isso limitava bastante a aplicabilidade das técnicas de aprendizado de máquina, pois, com uma massa pequena de dados ou pouco poder de processamento, não era possível ou financeiramente viável treinar modelos para uso comercial. Ainda que os modelos fossem continuamente estudados em instituições acadêmicas, sua aplicação prática era restrita.

Na década de 1990, os primeiros sistemas de armazenamento de grandes volumes de dados e processamento paralelo começaram a ser vendidos. Mas, a enorme popularidade do *big data* veio mesmo com o surgimento do Google File System e sua posterior evolução para o Hadoop Distributed File System (HDFS). Esses sistemas permitiam a utilização de computadores normais para a formação de um grande *cluster* de armazenamento, com baixíssimo custo. Tudo isso, associado à posterior implementação do MapReduce, que permitia o processamento em paralelo dos dados, na mesma máquina onde estavam armazenados, trouxe o grande *boom* do *big data*.

Os algoritmos de *machine learning*, adaptados para processamento paralelo, podiam, então – através do MapReduce –, utilizar as grandes massas de dados disponíveis nos sistemas Hadoop para treinar modelos grandes e complexos. A posterior criação do Apache Spark, que utiliza processamento em memória de forma distribuída, permitiu que o treino desses modelos fosse ainda mais rápido e colaborou na evolução da adoção do *machine learning* com *big data*.

Esse avanço na rapidez de processamento permitiu também que novas arquiteturas de redes neurais, até então reclusas ao ambiente acadêmico, fossem criadas e exploradas comercialmente, surgindo então o *deep learning*. Basicamente, as redes neurais são formadas por camadas de *perceptrons* interconectadas. Porém, o treino dessas redes neurais requer o processamento de toda a massa de dados pela rede, calcular os erros e propagá-los



**TODAS ESSAS FACILIDADES PERMITEM O CONTATO E AJUDAM A DESMISTIFICAR A IMAGEM INALCANÇÁVEL DA INTELIGÊNCIA ARTIFICIAL E DO *MACHINE LEARNING***

de volta pela rede. É possível imaginar como isso se torna dispendioso computacionalmente. Com a evolução do processamento de grandes massas de dados e a inovação com o uso de GPUs (unidades de processamento gráfico), que conseguem realizar cálculos matriciais em alta velocidade, foi possível criar arquiteturas de redes neurais com uma grande quantidade de camadas de *perceptrons* intermediárias escondidas, a que se deu o nome de *deep learning*. Hoje, essas camadas encontram aplicabilidade em uma diversidade de áreas, como visão computacional, reconhecimento de fala, identificação de padrões, entre outras.

**UTILIZANDO O APRENDIZADO DE MÁQUINA** Com toda essa carga matemática, pode parecer complicado utilizar o aprendizado de máquina nos negócios. No entanto, juntamente com os algoritmos houve também uma grande evolução na forma de aplicar *machine learning* no dia a dia. Hoje, diversas bibliotecas, como o *scikit-learn* na linguagem Python, permitem implementar algoritmos de aprendizado de máquina de forma simples e com uma documentação bastante extensiva. Mesmo para computação distribuída, o Apache Spark conta com a biblioteca *SparkML*, que pode ser usada para facilmente implementar mecanismos de aprendizado de máquina em grandes massas de dados, de forma distribuída. Especificamente para as redes neurais, bibliotecas como Tensorflow e Keras possuem métodos prontos para execução rápida de diversas arquiteturas, podendo até fazer uso de GPUs para processamento.

Para os não programadores, também já existem opções no mercado que permitem criar modelos de aprendizado de máquina sem precisar escrever código. A ferramenta IBM SPSS Modeler, por exemplo, permite que fluxos de tratamento de dados e modelos de aprendizado de máquina sejam criados de forma gráfica, acelerando o processo de

desenvolvimento. Ferramentas como o IBM Data Science Experience Local e IBM Watson Studio já possuem um assistente guiado para a criação e prototipagem rápida de modelos. No caso do IBM Watson Studio, existe até uma interface gráfica, capaz de criar graficamente arquiteturas de redes neurais, gerando automaticamente código em Tensorflow ou Keras.

Todas essas facilidades permitem o contato e ajudam a desmistificar a imagem inalcançável da inteligência artificial e do *machine learning*, mas, nem por isso, o cientista de dados se torna dispensável. Trata-se de um profissional que não apenas conhece os algoritmos de aprendizado de máquina e como implementá-los, mas também domina as técnicas estatísticas por trás dos modelos e como tratar os dados para que seu uso no modelo seja otimizado. Esse conhecimento é essencial para que, ao final, seu modelo não seja somente mais uma pilha mágica de álgebra linear, como citado no início deste artigo.

**Felipe Cruz** é cientista de dados na IBM e mestrando em ciência de dados pela Universidade de Berkeley.

**Agostinho Vilela** é Líder do Processo de Inovação para a América Latina e membro da Academia de Tecnologia da IBM.

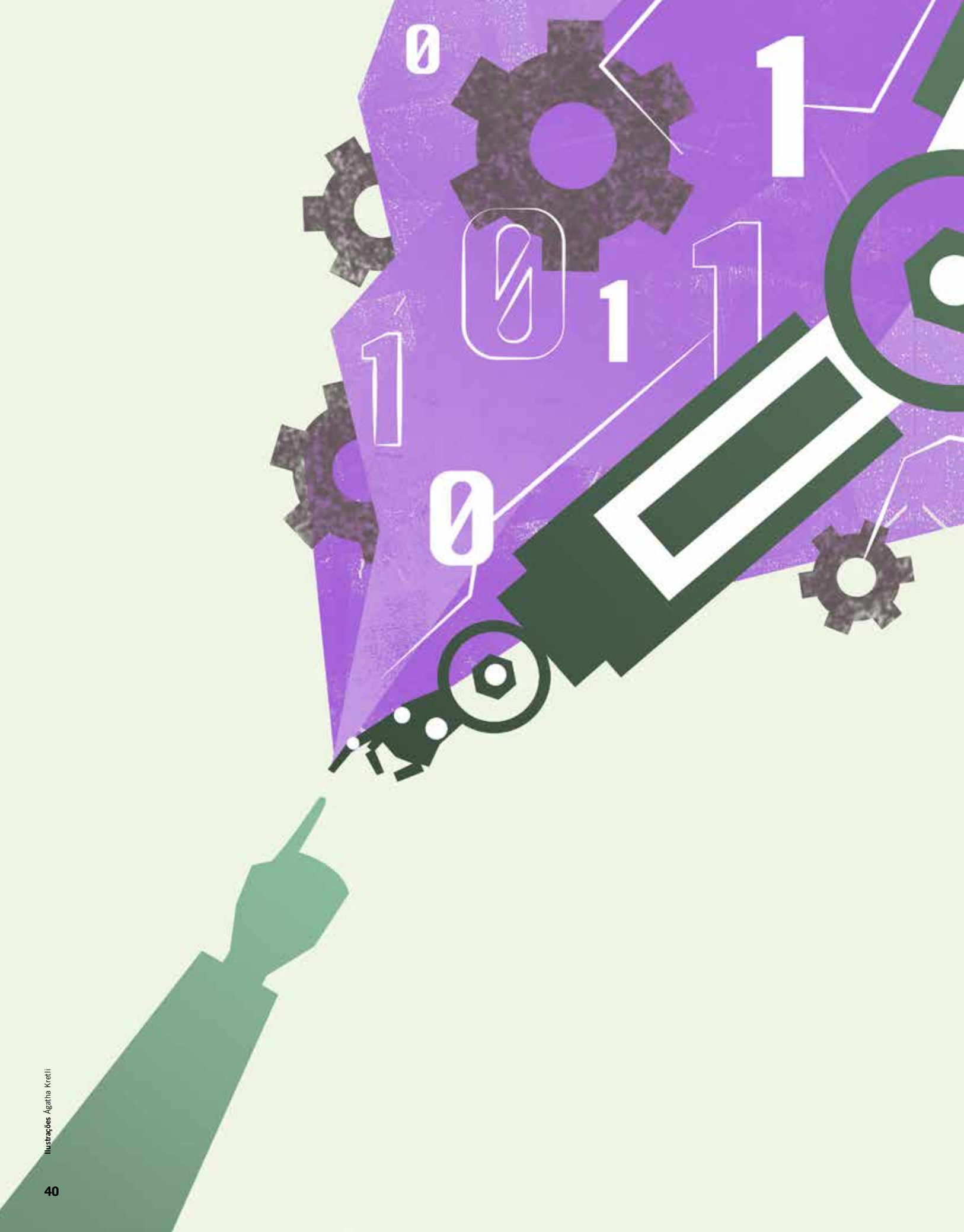
**Hugo Tadeu** é pesquisador do Núcleo de Inovação e Empreendedorismo da Fundação Dom Cabral.

**PARA SE APROFUNDAR NO TEMA**

COGNITIVE CLASS. **Data science, blockchain and cloud computing courses**. Disponível em: <https://cognitiveclass.ai/>. Acesso em: 28 jan 2019.

COURSERA. **Aprendizagem automática**. Disponível em: <https://pt.coursera.org/learn/machine-learning>. Acesso em: 28 jan. 2019.

KHAN ACADEMY. **Matemática**. 2019. Disponível em: <https://pt.Khanacademy.org/math>. Acesso em: 28 jan. 2019.




**Ilustrações** Agatha Kretli