No texto de Foster Provost e Tom Fawcett, os autores discutem o escopo da área de data science e suas competências, bem como as inter-relações do profissional de ciência de dados com outras áreas, como Big Data e data-driven decision making (tomada de decisão guiada por dados). A definição traduzida de data science é apresentada como “no escopo geral, data science é um grupo de princípios fundamentais que dão suporte e guiam a extração de informação e conhecimento dos dados”. Desse modo, como sugere a definição, a data science auxilia diretamente a mineração de dados (data mining), orientando a forma adequada de extrair e transformar os dados em conhecimento. Porém, essa definição não se restringe apenas à mineração de dados, podendo atuar em diversos campos, como marketing, CRM, publicidade e finanças.

Após a introdução, Provost e Fawcett expõem dois casos de aplicação da data science, concretizando a atuação do profissional. O primeiro é sobre a rede de supermercados Walmart, que prevê o aumento súbito de compras em situações de catástrofes naturais; e o segundo é sobre uma operadora de telefonia que busca reter clientes e atrair novos consumidores, dividindo o orçamento de marketing entre essas duas funções, mas seguindo as diretrizes definidas pela divisão de dados, ou seja, pela data-driven decision making (DDD).

De acordo com os autores, é a partir de dados estruturados que se podem tomar decisões estratégicas, sem confiar cegamente nem apenas nos dados nem somente na intuição, mas considerando ambos. Nos exemplos apresentados, observam-se dois tipos de decisões: as que demandam previsão e as que demandam repetição. Assim, o DDD deve estar alinhado aos objetivos da empresa e, sempre que possível, automatizando o processo decisório.

Além disso, o armazenamento dessas informações deve ser assegurado, introduzindo-se então o conceito de Big Data: grandes conjuntos de dados que superam a capacidade de processamento de computadores comuns e, por isso necessitam de tratamento para extrair significado. Atualmente, segundo Provost e Fawcett, as empresas vivenciam a fase 1 do Big Data, similar à fase 1 da Web, em que coletam e armazenam informações. Na evolução do Big Data, essas organizações passariam a usar estrategicamente os dados armazenados para prever cenários e gerar valor para si mesmas e para os clientes. Algumas Big Techs, como Google e Amazon, já estariam na fase 2 do Big Data, mas essa não é a realidade da maioria das empresas.

Nesse contexto, os autores afirmam que “*Data-science projects require close interaction between the scientists and the business people responsible for the decision making*.” Dessa forma, as empresas que armazenam e utilizam esses dados devem contar em suas equipes estratégicas com pessoas capazes de compreender a análise e o contexto desses dados. Ou seja, espera-se que a cultura organizacional de liderança e decisão inclua profissionais capazes de dialogar com os cientistas de dados, garantindo convenções e alinhamento entre esses setores.

Seguindo este raciocínio de análise de dados, Provost e Fawcett apresentam os conceitos fundamentais da data science: a metodologia CRISP-DM; a consideração do contexto de aplicação dos resultados; a relação entre o problema de negócio e a solução analítica; a avaliação das correlações tecnológicas; a atualização contínua de métricas de similaridade; a prevenção de sobre ajustes; e a atenção a fatores de confusão em inferências causais.

Em seguida, os autores destacam que existe uma diferença conceitual entre data science e o trabalho cotidiano do cientista de dados, o que se deve, em grande parte, à escassez de cursos consolidados na área acadêmica. Assim, embora os próprios profissionais tentem definir data science, cada um faz a partir de sua formação e perspectiva, o que impede um consenso. Além disso, há a presunção de que atividades rotineiras, como o processamento de dados, constituem o escopo da disciplina, mas na verdade, tratam-se apenas de tarefas frequentes e repetitivas, e não de elementos distintivos do campo. Tal analogia é exemplificada ao se comparar o cientista de dados iniciante a um químico de laboratório em início de carreira, ambos dedicados majoritariamente a trabalhos técnicos.

Por fim, Provost e Fawcett ressaltam que os conceitos centrais da data science permaneceram inalterados ao longo de vinte anos, apesar da evolução exponencial das tecnologias. Dessa forma, explicitam que, mais importante do que dominar ferramentas específicas, é compreender os princípios fundamentais que sustentam a área. Nesse sentido, “pensar data-analiticamente”, ou seja, aplicar os conceitos básicos a problemas de negócio, é a competência chave para o sucesso em ambientes empresariais orientados a dados.

**REFERÊNCIAS BIBLIOGRÁFICAS**

**PROVOST**, Foster; **FAWCETT**, Tom. *Data science and its relationship to big data and data-driven decision making*. **Big data**, v. 1, n. 1, p. 51-59, 2013.