Data Science e Big Data: Introdução

Dentro do guarda-chuva da Ciência de Dados, encontramos uma série de pilares e áreas de estudo fundamentais, que um profissional da área utiliza todos os dias. Esse organograma deixa isso muito claro.

Data mining. Essa área é famosa há muito mais tempo do que o data science em si, e basicamente se trata de explorar conjuntos normalmente grandes de dados em busca de padrões. Existem uma série de ferramentas capazes de fazer isso, e o machine learning é uma das mais conhecidas, em que tentamos ensinar o computador a extrair padrões e, normalmente, classificar alguma coisa: textos, imagens, áudios, etc.

Big Data Tools. Nesse caso, temos ferramentas muito importantes que servem para armazenar e lidar com a grande quantidade de dados que trabalhamos dentro dessa área. Aqui temos, inclusive, bancos de dados relacionais ou não (SQL e NoSQL), onde cada um tem sua vantagem e sua aplicação específica.

Os métodos estatísticos. Isso é a base matemática que a ciência de dados usa. No fundo, todo o esforço de um cientista de dados é praticamente inútil se não houver alguma ferramenta que consiga analisar corretamente os dados, e é aí que entram os métodos estatísticos.

Data exploration and comunication. Não adianta termos esse trabalho todo se não houver um jeito de enxergar