Các công ty ngày nay đang phải đối mặt với một lượng dữ liệu lớn từ các phương tiện truyền thông xã hội, tìm kiếm và cảm biến, cũng như từ các nguồn truyền thống. Theo một ước lượng, mỗi ngày trên thế giới tạo ra khoảng 2,5 quintrillion (2 500 000 000 gigabytes) byte dữ liệu. Việc hiểu rõ về "big data" để cải thiện quyết định và hiệu suất kinh doanh đã trở thành một trong những cơ hội chính cho tổ chức có kích thước và hình dạng khác nhau, nhưng cũng đồng thời mang đến những thách thức lớn.

Các doanh nghiệp như Amazon, YouTube và Spotify đã phát triển bằng cách phân tích dữ liệu lớn mà họ thu thập về sở thích và mua sắm của khách hàng để tạo ra hàng triệu đề xuất cá nhân. Một số dịch vụ trực tuyến khác cũng phân tích dữ liệu lớn để hỗ trợ người tiêu dùng, bao gồm các dịch vụ tìm giá thấp nhất cho ô tô, máy tính, kế hoạch điện thoại di động, quần áo, vé máy bay, phòng khách sạn và nhiều loại hàng hóa và dịch vụ khác. Big data cũng mang lại những lợi ích trong lĩnh vực thể thao (xem trường hợp mở đầu chương), giáo dục, khoa học, chăm sóc sức khỏe và lực lượng chức năng.

Việc phân tích hàng tỷ điểm dữ liệu được thu thập về bệnh nhân, nhà cung cấp dịch vụ y tế và hiệu quả của đơn thuốc và phương pháp điều trị đã giúp Dịch vụ Y tế Quốc gia Anh (NHS) tiết kiệm khoảng 581 triệu bảng Anh (784 triệu đô la Mỹ). Dữ liệu được lưu trữ trong một Oracle Exadata Database Machine, có thể nhanh chóng phân tích các lượng dữ liệu rất lớn (hãy xem phần thảo luận về các nền tảng phân tích trong chương này). NHS đã sử dụng kết quả từ phân tích dữ liệu lớn để tạo ra bảng điều khiển nhận biết bệnh nhân đang dùng 10 hoặc hơn loại thuốc cùng một lúc và những bệnh nhân đang sử dụng quá nhiều loại kháng sinh. Tổng hợp một lượng lớn dữ liệu về các loại thuốc và liệu pháp được đưa ra cho bệnh nhân ung thư và liên kết thông tin đó với kết quả của bệnh nhân đã giúp NHS xác định các giao thức điều trị hiệu quả hơn.

Thành phố New York phân tích tất cả dữ liệu liên quan đến tội phạm mà nó thu thập để giảm tỷ lệ tội phạm. Chương trình bản đồ tội phạm CompStat của nó sử dụng một cơ sở dữ liệu toàn thành phố bao gồm tất cả các tội phạm hoặc khiếu nại được báo cáo, các vụ bắt giữ và lệnh triệu tập trong mỗi trong 76 khu phố của thành phố để báo cáo hàng tuần về hoạt động khiếu nại tội phạm và bắt giữ tại cả khu phố, quận tuần tra và cấp toàn thành phố. Dữ liệu CompStat có thể được hiển thị trên bản đồ cho thấy vị trí tội phạm và bắt giữ, các điểm nóng tội phạm và các thông tin khác để giúp các chỉ huy khu phố nhanh chóng xác định mô hình và xu hướng và triển khai nhân viên cảnh sát nơi chúng cần thiết nhất. Dữ liệu lớn về hoạt động tội phạm cũng làm năng động Đội Chiến lược Tội phạm của New York City, mục tiêu là những người phạm tội tồi tệ nhất để truy cứu trách nhiệm một cách mạnh mẽ. Các công ty chăm sóc sức khỏe hiện đang phân tích dữ liệu lớn để xác định các liệu pháp hiệu quả và kinh tế nhất cho các bệnh mãn tính và các bệnh thông thường và đưa ra các đề xuất chăm sóc cá nhân cho bệnh nhân.

Tuy nhiên, có những hạn chế khi sử dụng dữ liệu lớn. Một số công ty đã nhanh chóng bắt đầu các dự án dữ liệu lớn mà không xác định trước một mục tiêu kinh doanh cho thông tin mới này hoặc các chỉ số hiệu suất chính để đánh giá thành công. Việc đối mặt với một lượng lớn số liệu không nhất thiết có nghĩa là thông tin đúng đang được thu thập hoặc rằng người ta sẽ đưa ra các quyết định thông minh hơn. Các chuyên gia trong lĩnh vực phân tích dữ liệu lớn tin rằng quá nhiều công ty, bị quyến rũ bởi hứa hẹn của dữ liệu lớn, nhảy vào các dự án dữ liệu lớn mà không có kết quả gì đáng chú ý. Họ bắt đầu tích lũy một lượng lớn dữ liệu mà không có mục tiêu rõ ràng hoặc hiểu biết về cách phân tích dữ liệu lớn sẽ đạt được mục tiêu của họ hoặc câu hỏi họ đang cố gắng trả lời. Tổ chức cũng sẽ không hưởng lợi từ dữ liệu lớn nếu nó chưa được làm sạch, tổ chức và quản lý đúng cách - nghĩ về chất lượng dữ liệu. Chỉ vì một điều gì đó có thể đo lường không có nghĩa là nó nên được đo lường. Giả sử, ví dụ, rằng một công ty lớn muốn đo lường lưu lượng trang web của mình liên quan đến số lần được đề cập trên Twitter. Nó xây dựng một bảng điều khiển số để liên tục hiển thị kết quả. Trong quá khứ, công ty đã tạo ra phần lớn cơ hội và doanh số bán hàng từ triển lãm và hội nghị. Chuyển sang số lần được đề cập trên Twitter như là chỉ số chính để đo lường thay đổi sự tập trung của bộ phận bán hàng. Bộ phận đó dồn năng lượng và nguồn lực của mình vào theo dõi các lượt nhấp trên trang web và lưu lượng truy cập truyền thông xã hội, tạo ra nhiều cơ hội không đủ chất lượng không dẫn đến doanh số bán hàng.

Mặc dù dữ liệu lớn rất tốt trong việc phát hiện ra sự tương quan, đặc biệt là những tương quan tinh tế mà phân tích của các bộ dữ liệu nhỏ hơn có thể bỏ qua, nhưng phân tích dữ liệu lớn không nhất thiết cho thấy quan hệ nhân quả hoặc xác định những tương quan nào là ý nghĩa. Ví dụ, việc xem xét dữ liệu lớn có thể cho thấy rằng từ năm 2006 đến 2011, tỷ lệ giết người tại Hoa Kỳ có mối quan hệ cao với thị phần của Internet Explorer vì cả hai đều giảm mạnh. Nhưng điều này không nhất thiết có nghĩa là có bất kỳ kết nối có ý nghĩa nào giữa hai hiện tượng này. Người phân tích dữ liệu cần một số kiến thức kinh doanh về vấn đề mà họ đang cố gắng giải quyết bằng dữ liệu lớn.

Các mô hình dự đoán dữ liệu lớn không nhất thiết mang lại ý tưởng tốt hơn về điều gì sẽ xảy ra trong tương lai. Công ty Điện năng Meridian Ltd., một công ty sản xuất và phân phối điện hoạt động tại New Zealand và Úc, đã chuyển sang không sử dụng hệ thống bảo trì thiết bị dự đoán đang lão hóa. Phần mềm này được thiết kế để dự đoán nhu cầu bảo trì của tất cả các thiết bị lớn mà công ty sở hữu và vận hành, bao gồm máy phát điện, turbine gió, biến áp, cầu dao, và pin công nghiệp. Tuy nhiên, hệ thống sử dụng các kỹ thuật mô hình hóa lỗi thời và không thể thực sự dự đoán sự cố của thiết bị. Nó chạy các mô phỏng về các kịch bản khác nhau và dự đoán khi nào tài sản sẽ bị hỏng theo các kiểm tra mô phỏng. Các đề xuất của phần mềm không hữu ích vì chúng không dự đoán chính xác những thiết bị nào thực sự hỏng trong thế giới thực. Cuối cùng, Meridian đã thay thế hệ thống cũ bằng phần mềm Dự đoán Bảo trì và Chất lượng của IBM, dựa trên dự đoán từ dữ liệu thời gian thực hơn từ thiết bị.

Tất cả các bộ dữ liệu và mô hình dự đoán dữ liệu đều phản ánh những đặc tính chủ quan của những người chọn dữ liệu và thực hiện phân tích. Một số năm trước, Google phát triển một thuật toán đầu tiên bằng cách sử dụng dữ liệu thu thập từ tìm kiếm web để xác định chính xác bao nhiêu người mắc cúm và bệnh lan truyền như thế nào. Nó cố gắng tính toán số người mắc cúm ở Hoa Kỳ bằng cách liên kết vị trí của người dùng với các truy vấn tìm kiếm liên quan đến cúm trên Google. Google liên tục đánh giá quá cao tỷ lệ cúm, so với dữ liệu truyền thống được thu thập sau đó bởi Trung tâm Kiểm soát Bệnh tật Hoa Kỳ (CDC). Một số nhà khoa học đề xuất rằng Google đã bị "lừa dối" bởi sự chú ý rộng rãi của truyền thông về mùa cúm nặng nề ở Hoa Kỳ trong năm đó, được tăng cường thêm bởi sự chú ý của truyền thông xã hội. Mô hình được phát triển để dự báo xu hướng cúm dựa trên giả định sai lầm rằng sự xuất hiện của các truy vấn tìm kiếm liên quan đến cúm trên Google là một chỉ số chính xác của số người thực sự mắc cúm. Thuật toán của Google chỉ nhìn vào các con số, không xem xét ngữ cảnh của kết quả tìm kiếm.

Ngoài việc can thiệp vào bầu cử bởi các quốc gia thù địch, sự thiếu chú ý đến ngữ cảnh và các giả định sai lầm có thể đã đóng một vai trò trong việc các chuyên gia chính trị đa số không dự đoán được chiến thắng của Donald Trump trước Hillary Clinton trong cuộc bầu cử tổng thống năm 2016. Chiến thắng của Trump trái ngược với hầu hết các dự báo lớn, dự đoán khả năng chiến thắng của Clinton từ 70 đến 99 phần trăm.