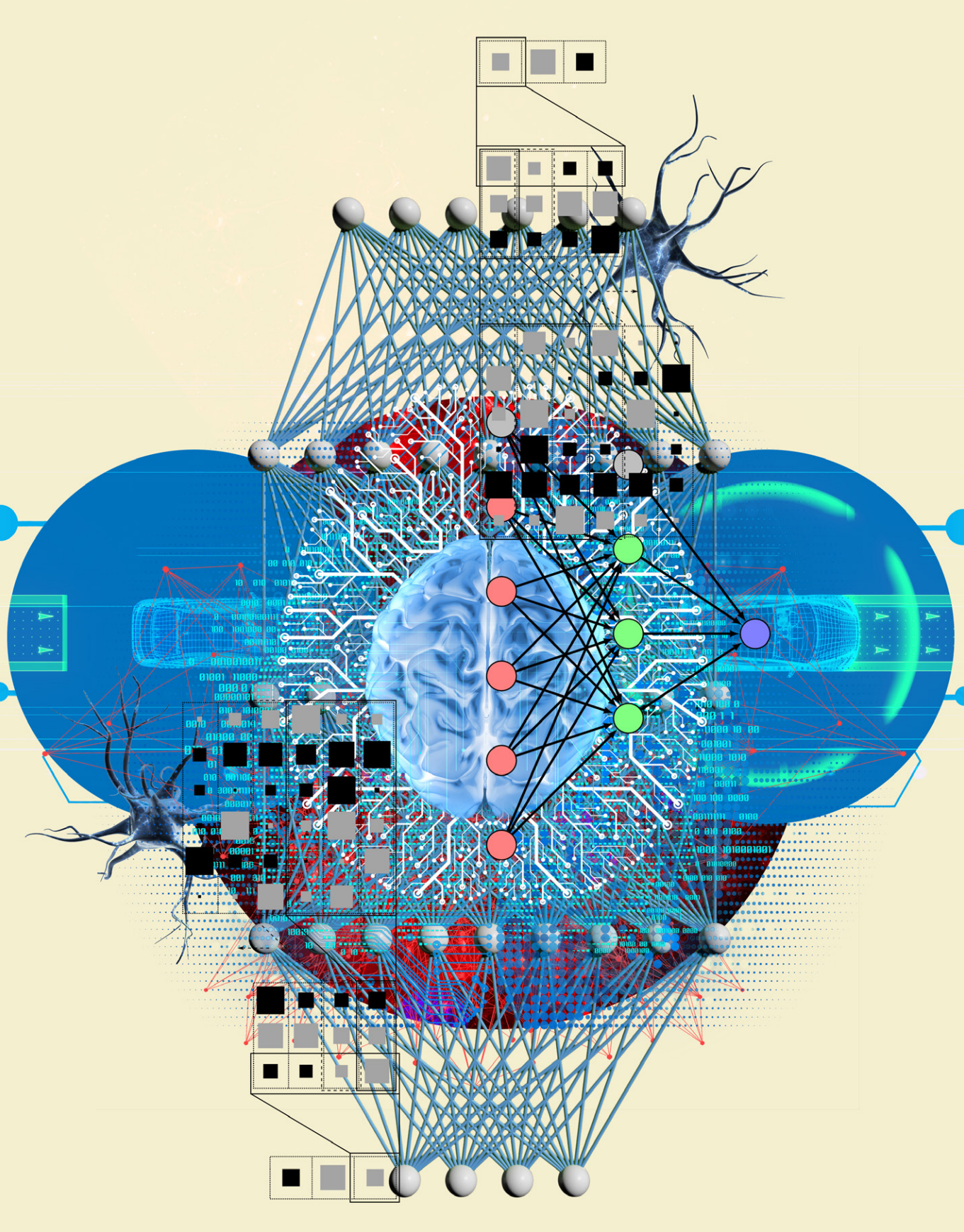
MỤC LỤC



LỜI NÓI ĐẦU

Xiao-Li Meng

"Nếu bạn muốn giải quyết tất cả các vấn đề trên thế giới, hãy học chuyên ngành Khoa học máy tính."

Khi một diễn giả tại một hội nghị AI hiển thị dòng chữ này, cái tôi thống kê của tôi đã bị kích động đến mức gần sáu sigma*. Rất may, dòng tiếp theo xuất hiện trong vòng chưa đầy ba giây: "Nếu bạn muốn giải quyết tất cả các vấn đề do khoa học máy tính tạo ra, hãy đăng kí vào chương trình Thạc sĩ tại khoa Khoa học Nghệ thuật".

Bất cứ ai chọc cười chúng ta với cách ghép nối thông minh này đều là người hiểu sâu sắc về thời kì rực rỡ và hoang mang mà chúng ta đang sống. Những tiến bộ trong khoa học máy tính và công nghệ đã tạo ra thời đại kĩ thuật số, từ đó tạo ra khoa học dữ liệu. Dường như không có gì vượt quá tầm tay khi chúng ta đang có rất nhiều dữ liệu để khám phá những bí mật của tự nhiên và của sinh vật tiến hóa nhất.

Nhưng không có bữa trưa nào miễn phí - đó là qui luật phổ quát của khoa học dữ liệu (và của cuộc sống). Sau đây là một số ví dụ tương phản cho chúng ta suy nghĩ. Y tế cá nhân nghe có vẻ hay ho đấy, nhưng liệu có thể tìm đủ "chuột thí nghiệm" cho tôi không? Chắc chắn là chúng ta cần thu thập càng nhiều dữ liệu càng tốt về con người để phát triển công nghệ AI. Nhưng xin vui lòng chỉ nghiên cứu những người khác, đừng có xâm phạm quyền riêng tư của tôi!

Đối với những bạn có khả năng theo học chương trình sau đại học và có thể học ít nhất 31.536.000 giây, hãy cứ tiếp tục công việc của bạn như thể không có 30 giây tiếp theo. Đối với những bạn không có khả năng, cuốn sách này có thời lượng 50 x 30 giây, tùy theo sáu sigma cá nhân của bạn. Đọc xong quyển sách này sẽ không khiến bạn trở thành nhà khoa học dữ liệu 30 giây. Nhưng nếu không hiểu nội dung của nó, đơn xin cấp "quốc tịch thời đại kĩ thuật số" của bạn sẽ có 99% khả năng bị từ chối. Tất nhiên, đây là cơ hội để cho bạn nắm bắt.

***Sáu sigma (six sigma)**: Một hệ phương pháp cải tiến qui trình kinh doanh và quản lí chất lượng dựa trên thống kê với mục đích tìm ra lối giảm thiểu sai sót. Trong đó 6 là cấp độ ít lỗi nhất, hoàn hảo đến 99,99966%. (ND)*

LỜI GIỚI THIỆU

Liberty Vittert

Chúng ta đã sống theo chủ nghĩa nhân văn

trong một thời gian dài, sử dụng bản năng, suy nghĩ, quan điểm và kinh nghiệm để dẫn dắt các quyết định. Tuy nhiên, chúng ta đang chuyển sang kỉ nguyên của Chủ nghĩa Dữ liệu - để cho dữ liệu hướng dẫn mọi quyết định của mình. Từ biến đổi khí hậu đến khủng hoảng người tị nạn, đến chăm sóc sức khỏe, dữ liệu luôn là động lực, và không chỉ trong các vấn đề lớn lao này mà còn cả những chuyện nhỏ nhặt trong cuộc sống hàng ngày nữa. Thay vì đến hiệu sách, Amazon có thể cho bạn biết bạn muốn đọc gì. Tương tự như vậy, các ứng dụng hẹn hò sẽ cho bạn biết bạn hợp cạ với ai, dựa trên vô số dữ liệu thu thập được.

Chủ nghĩa Nhân văn và Chủ nghĩa Dữ liệu hiện đang kình chống nhau. Một số người muốn hoàn toàn theo hướng dữ liệu, những người khác không muốn từ bỏ sự can thiệp của con người. Khoa học dữ liệu, với tư cách là một ngành học, kết hợp chủ nghĩa nhân văn và chủ nghĩa dữ liệu lại với nhau. Nó kết hợp các cơ sở dữ liệu rộng lớn, các công cụ thống kê mạnh mẽ thúc đẩy các quá trình tính toán và phân tích cùng với suy luận thông thường và định lượng mà con người chúng ta đã và đang phát triển qua hàng nghìn năm. Khoa học dữ liệu không chỉ là điều khiển bởi dữ liệu hay điều khiển bởi con người: Nó là sự kết hợp nghệ thuật của cả hai.

Trước khi chúng ta bắt đầu trình bày chi tiết về cuốn sách này, hãy lùi lại một chút thời gian đến thế kỉ 17 và đến thăm Blaise Pascal, một tu sĩ người Pháp bị khủng hoảng đức tin. Ông ta quyết định suy nghĩ về các lựa chọn trong tương lai của mình với thông tin mà ông ta có - hay nếu bạn muốn, có thể gọi là dữ liệu cũng được:

Nếu Chúa không tồn tại nhưng tôi lại tin vào Ngài, thì tôi có thể có một cuộc sống lãng phí với niềm tin sai lầm, nhưng không có gì xảy ra.

Nếu Chúa không tồn tại và tôi không tin vào Ngài, thì tôi đã không lãng phí cuộc đời mình với niềm tin sai lầm, nhưng cũng như vậy, không có gì xảy ra.

Nếu Chúa tồn tại và tôi tin vào Ngài, thì tôi có một cuộc sống vĩnh hằng tuyệt vời trên Thiên đàng.

Nhưng nếu Chúa có tồn tại mà tôi lại không tin vào Ngài, thì tôi có thể phải đối mặt với lửa địa ngục vĩnh viễn.

Pascal đã sử dụng dữ liệu mà ông có để đưa ra quyết định nhằm tối ưu hóa hạnh phúc trong tương lai và giảm thiểu rủi ro tiềm ẩn. Thật vậy, khoa học dữ liệu là như thế: Lấy thông tin trong quá khứ và hiện tại để dự đoán khả năng xảy ra của các sự kiện trong tương lai, hay nói nôm na, là thứ gần nhất với quả cầu pha lê dự đoán tương lai mà chúng ta có thể sử dụng. Sự khác biệt duy nhất giữa chúng ta và Pascal là chúng ta đang sống trong một thế giới có nhiều hơn bốn bit dữ liệu để phân tích. Chúng ta có vô vàn dữ liệu.

Ước tính có hơn 2,5 exabyte dữ liệu được tạo ra mỗi ngày. Một phép tính nhanh cho ta biết lượng thông tin này tương đương với việc xếp chồng các cuốn sách *Harry Potter* từ Trái đất lên Mặt trăng, thu chúng lại, rồi lại xếp vòng quanh chu vi Trái đất 550 lần. Mà đó chỉ đơn giản là lượng dữ liệu được tạo ra mỗi ngày!

Cách trình bày của quyển sách này

Hai chương đầu tiên chia khoa học dữ liệu thành các yếu tố cơ bản, tiếp theo là khía cạnh quan trọng nhất nhưng lại ít được thảo luận nhất - những gì khoa học dữ liệu không thể cho chúng ta biết. Năm chương tiếp theo liên quan đến cách khoa học dữ liệu ảnh hưởng đến chúng ta trong mọi lĩnh vực - khoa học, xã hội, kinh doanh, giải trí và tương lai của thế giới. Trong mỗi chủ đề riêng lẻ có: Mẫu 3 giây, một cái nhìn thú vị về chủ đề; tiếp theo là phần giải thích Dữ liệu 30 giây chi tiết hơn; và cuối cùng là Phân tích 3 phút, mang đến cho người đọc cơ hội đi sâu hơn vào những vấn đề phức tạp và những sắc thái khác nhau của cuộc thảo luận.

Cuốn sách này được biên soạn tỉ mỉ bởi các chuyên gia trong ngành nhằm giúp chỉ cho chúng ta thấy cách dữ liệu đang thay đổi mọi ngành nghề và mọi khía cạnh trong cuộc sống theo những cách mà chúng ta thậm chí còn chưa nghĩ đến, đồng thời nêu lên các luận cứ kèm theo số liệu và vấn đề đạo đức trong buổi bình minh của bất kì kỉ nguyên mới nào.