کلان‌داده، داده‌کاوی و فراگیریِ ماشین

نویسنده: جرد دین

مترجمین:

دکتر محمدرحیم اسفیدانی (هیات علمی دانشکده مدیریت دانشگاه تهران)

امیرعباس دربانی باسمنج

آبان ۹۴

سرشناسه : دین، جرد، ۱۹۷۸ – م.
Dean, Jared
عنوان و نام پدیدآور : کلان‌داده، داده‌کاوی و فراگیری ماشین: ایجاد ارزش برای رهبران کسب و کار و متخصصان/نویسنده جرد دین ؛ مترجمین محمدرحیم اسفیدانی، امیرعباس دربانی باسمنج، مرکز تحقیقات و نوآوری سازمان اتکا.
مشخصات نشر : تهران: اتکا، انتشارات مرکز تحقیقات و نوآوری سازمان اتکا، ۱۳۹۴.
مشخصات ظاهری : ۳۰۷ ص.: مصور، جدول، نمودار.
شابک : ۲۱۰۰۰۰ ریال ۷-۷۷-۷۶۲۴-۶۰۰-۹۷۸
وضعیت فهرست نویسی : فیپا
یادداشت : عنوان اصلی: Big data, data mining, and machine learning: value creation for business leaders and practitioners [2014]
یادداشت : کتابنامه.
موضوع : مدیریت – داده پردازی
موضوع : داده‌کاوی
موضوع : داده‌های کلان
شناسه افزوده : اسفیدانی، محمدرحیم، ۱۳۵۵ - ، مترجم
شناسه افزوده : ربانی باسمنج، امیرعباس، ۱۳۶۷- مترجم
شناسه افزوده : اتکا، مرکز تحقیقات و توسعه
رده‌بندی کنگره : HD۳۰/۲/د۹۵۰۸ ۱۰۱۴
رده‌بندی دیویی : ۶۵۸/۴۳۱
شماره کتابشناسی ملی : ۶۸۲

انتشارات مرکز تحقیقات و توسعه سازمان اتکا

عنوان: کلان‌داده، داده‌کاوی و فراگیری ماشین: ایجاد ارزش برای رهبران کسب و کار و متخصصان
مؤلف: جرد دین
مترجمین: دکتر محمدرحیم اسفیدانی، امیرعباس دربانی باسمنج
ویراستار فنی: الیاس خواجه‌کریمی
زمان و نوبت چاپ: آبان ۱۳۹٤، اول

شمارگان: ۲۰۰
طراح جلد: نوید یوسفی
شابک: ۷-۷۷-۷٦۲٤-٦۰۰-۹۷۸
بهاء: ۲۱۰۰۰۰ ریال

نشانی: تهران، خ امام خمینی (ره)، خ شهید مرادی، ساختمان شهدای مرکزی اتکا، ط ۲
تلفن: ۶۱۹۱۵۱۲۴-۰۲۱ دورنگار: ۶۱۹۱۵۳۵۵-۰۲۱ وب‌سایت: www.etkaec.ir ایمیل: tec@ERDCenter.com

فهرست مطالب

پیشگفتار ۳

مقدمه ۵

تشکر و قدردانی ۸

مقدمه ۱۰

جدا از زمانبندی کلان داده: ۱۵

چرا اکنون این مبحث مناسب است ۱۹

آیا کلان داده صرفاً یک هیجان زودگذر است؟ ۲۱

وقتی استفاده از کلان داده تفاوتی بزرگ ایجاد میکند ۲۴

بخش اول: محیط محاسبه ۳۵

فصل ۱: سخت‌افزار ۳۸

محل ذخیره (دیسک) ۳۸

واحد پردازش مرکزی ۴۱

حافظه ۴۳

شبکه ۴۵

فصل ۲ : سامانه‌های توزیع شده ۴۷

محاسبه پایگاه داده ۴۹

محاسبه سامانه فایل ۵۰

ملاحظات ۵۲

فصل ۳: ابزارهای آنالیز ۵۴

وکا ۵۵

فهرست مطالب

زبانهای جاوا و JVM ۵۶

آر(R) ۵۹

پایتون ۶۲

SAS ۶۳

بخش دو: تبدیل داده به ارزش کسب‌وکار ۶۷

فصل ۴: مدلهای پیش‌بینی ۷۰

یک متدولوژی برای ساخت مدلها ۷۴

EMMA ۷۷

رده‌بندی باینری ۸۱

رده‌بندی چند سطحی ۸۳

پیش‌بینی بازه‌ای ۸۴

ارزیابی مدلهای پیش‌بینی ۸۵

فصل ۵ :تکنیکهای معمول برای مدل‌سازی پیش‌بینی ۹۰

RFM ۹۱

رگرسیون ۹۵

مدلهای تعمیم یافتهی خطی ۱۰۶

شبکه عصبی ۱۱۲

درختهای رگرسیون و تصمیم ۱۲۶

ماشین بردار پشتیبانی ۱۳۳

رده‌بندی شبکه متدهای بیزین ۱۴۱

متدهای گروهی ۱۵۶

فصل ۶: بخش‌بندی ۱۵۹

فهرست مطالب

آنالیز خوشه ۱۶۶

ارزیابی خوشه‌بندی ۱۶۸

تعداد خوشه‌ها ۱۶۹

الگوریتم K -MEANS ۱۷۲

خوشه‌بندی سلسله مراتبی ۱۷۳

پروفایل کردن خوشه‌ها ۱۷۴

فصل ۷: مدل‌سازی پاسخ انگیزه‌ای ۱۷۵

ساختن مدل پاسخ ۱۷۶

اندازه گیری پاسخ افزایشی ۱۷۸

فصل ۸: داده‌کاوی ۱۸۴

سری‌های زمانی ۱۸۴

کاهش ابعاد ۱۸۶

یافتن الگوها ۱۸۷

فصل ۹: سامانه‌های توصیه ۲۰۰

سامانه‌های توصیه چه هستند؟ ۲۰۱

سامانه‌های توصیه در کجا استفاده میشوند؟ ۲۰۲

سامانه‌های توصیه چگونه کار میکنند؟ ۲۰۳

ارزیابی کیفیت توصیه و پیشنهاد ۲۱۰

توصیه‌ها در عمل: کتابخانهی SAS ۲۱۱

فصل ۱۰:آنالیز متن ۲۱۸

بازیابی اطلاعات ۲۱۹

طبقه بندی محتوا ۲۲۱

فهرست مطالب

متن کاوی ۲۲۲

آنالیز متن در عمل: بیایید "خطر!" بازی کنیم! ۲۲۴

بخش سه:مثال‌های موفق داده‌کاوی در عمل ۲۴۲

فصل ۱۱: بررسی نمونه موردی یک شرکت بزرگ خدمات مالی که دفتر مرکزی آن در ایالات متحده است ۲۴۵

فرایند کمپین بازاریابی سنتی ۲۴۷

راه حل‌های داده‌کاوی با عملکرد بالا ۲۵۱

ایجاد ارزش برای مشتری ۲۵۲

فصل ۱۲: بررسی نمونه موردی: یک ارائه دهنده خدماتی بزرگ در حوزه بهداشت و درمان ... ۲۵۴

CAHPS ۲۵۶

HEDIS ۲۵۷

HOS ۲۵۸

IRE ۲۵۸

فصل ۱۳: بررسی نمونه موردی تولید کننده تکنولوژی ۲۶۶

یافتن دستگاه‌های معیوب ۲۶۷

فصل ۱۴: بررسی نمونه موردی مدیریت آنلاین برند ۲۷۲

فصل ۱۵: بررسی نمونه موردی پیشنهادهای اپلیکیشن تلفن همراه ۲۷۷

فصل ۱۶: بررسی نمونه موردی یک تولید کننده محصول های-تک ۲۸۲

مدیریت داده گم شده ۲۸۳

کاربری ورای تولید ۲۸۵

فصل ۱۷: نگاهی به آینده ۲۸۶

تحقیقات تجدیدپذیر ۲۸۷

فهرست مطالب

حریم خصوصی در مورد مجموعه داده عمومی ۲۸۸

اینترنت اشیاء ۲۹۰

توسعه نرم‌افزار در آینده ۲۹۱

پیشرفت الگوریتمها در آینده..... ۲۹۳

در نتیجه..... ۲۹۶

فهرست منابع ۲۹۸

بسمه‌تعالی

«کار علمی و تحقیقاتی‌تان را جدی بگیرید. دانش خیلی از نیازهای شما در دانشگاه‌ها وجود دارد.»

مقام معظم رهبری

حضرت امام خامنه‌ای (مدظله‌العالی)

امروزه با پیشرفت روزافزون علوم رایانه، آمار و ریاضی و همچنین رشد بی‌وقفه فناوری، شاهد گرایش بیش از پیش استفاده از داده‌ها و اطلاعات در اتخاذ تصمیمات بهینه سازمان‌ها هستیم. با توجه به رشد بی‌سابقه فناوری‌های محاسباتی و بالا رفتن قدرت پردازش رایانه‌ها و همچنین به‌کارگیری فنون و الگوریتم‌های علوم کامپیوتر و آمار، متخصصان حوزه اطلاعات در دهه‌ی گذشته توانسته‌اند تحلیل‌های بسیار اثربخشی را بر روی انباشت داده‌های کلان به منظور استخراج و کشف دانش نهان در آن‌ها انجام دهند.

در یک دهه گذشته رشد نمایی ذخیره داده و بالا رفتن قدرت کشف دانش به گونه‌ای بوده است که بسیاری از کسب‌وکارها به منظور فراهم نمودن زیرساخت‌های مناسب برای ذخیره‌سازی و تحلیل داده، سرمایه‌گذاری‌های کلانی را انجام داده‌اند. امروزه اکثریت شرکت‌ها برای اتخاذ تصمیم مطابق واقعیت، و نه صرفا یک ظن شخصی، به داده کاوی و تحلیل‌های سطح بالای داده روی آورده و بر آن اساس عمل می‌نمایند. در سال‌های اخیر صنعت خرده‌فروشی از جمله صنایعی بوده است که از این موقعیت و فرصت، بیشترین بهره را برده است. کتاب حاضر به عنوان منبعی کامل و جامع برای امر داده‌کاوی، توسط پژوهشگر محترم مرکز تحقیقات و نوآوری سازمان اتکا جناب آقای امیرعباس دربانی ترجمه گردیده است. ضمن سپاسگزاری از مترجم محترم امیدوارم این اثر به بهبود ارتقای سطح دانش مدیران و کارشناسان میهن عزیز اسلامی ایران خصوصاً در سازمان اتکا مفید واقع گردد. انشاءا...

محمدمهدی کربلایی

مدیرعامل سازمان اتکا

پیشگفتار

من شیفته آنالیز پیش‌بینی هستم و تمام عمر کاری خود را در این حوزه گذرانده‌ام. ریاضیات لذت‌بخش است (حداقل برای من)؛ اما تبدیل آنچه الگوریتم‌ها از آن پرده برمی‌دارند به راه‌حلی که یک مجموعه بتواند از آن استفاده و با آن تولید سود کند ریاضیات را ارزشمندتر هم می‌کند. در این رابطه من و جارد دین به نحوی باهم هم‌عقیده هستیم. ما هر دو دوست داریم شاهد چگونگی تأثیر راه‌حل‌ها بر سازمان‌هایی که با آن‌ها کار می‌کنیم باشیم. گرچه چیزی که ما را شگفت‌زده می‌کند این است که حوزه‌ای که ما پیش‌تر و به‌صورت گام‌به‌گام به کار گرفته‌ایم اکنون به یکی از جذاب‌ترین مشاغل قرن بیست و یکم بدل شده. چگونه چنین اتفاقی افتاده است؟

ما در دنیایی زندگی می‌کنیم که داده‌ها در ابعادی همواره رو به رشد در حال جمع‌آوری شدن هستند. این داده‌ها انسان‌ها و ماشین‌ها را خلاصه کرده و دانه‌بندی بهتری از رفتارهای آن‌ها ارائه می‌دهند. این داده‌ها با روش‌های بعضی موارد توسط شاخص‌های حجم، تنوع و سرعت توضیح داده می‌شوند و این یعنی تعریف کلان داده. این سه شاخصه برای درک ارزش در داده‌ای جمع‌آوری می‌شوند که حتی ما هم نمی‌دانیم به چه‌کار می‌آیند. پیش‌ازاین بسیاری از سازمان‌ها داده‌ها را جمع‌آوری کرده و سعی کرده‌اند رویکرد هوش کسب‌وکار که رایج شده است خلاصه‌ای از آن را ارائه می‌دهند.

اما در سال‌های اخیر یک تغییر نمونه‌ای رخ‌داده است. سازمان‌ها دریافته‌اند که آنالیز پیش‌بینی[1] مسیر تصمیم‌گیری آن‌ها را تغییر دگرگون می‌کند. الگوریتم‌ها و رویکردهای مربوط به مدل‌سازی پیش‌بینی که در این کتاب توضیح داده‌شده‌اند در بیشتر موارد چیز جدیدی نیستند. خود جارد هم عنوان کرده است که مسئله کلان داده به‌هیچ‌وجه چیز جدیدی نیست. الگوریتم‌هایی که او در مورد آن‌ها توضیح داده است حداقل ۱۵ سال است که

وجود دارند و این گواهی است بر این‌که اساساً به الگوریتم‌های تازه نیازی نیست. بااین‌وجود مدل‌سازی پیش‌بینی درواقع برای بسیاری از سازمان‌هایی که می‌خواهند تصمیم‌گیری بر اساس داده را گسترش دهند ناشناخته است. این سازمان‌ها نه‌تنها باید به درک درستی از دانش و اصول مدل‌سازی پیش‌بینی دست یابند، بلکه باید بدانند چگونه می‌توانند آن‌ها را در مورد

[1] Predictive Analytics

مسائلی که بر ضد استانداردها و پاسخ‌های این اصول هستند به‌کارگیرند. اما مدل‌سازی پیش‌بینی چیزی بیش از صرفاً ساختن مدل‌های پیش‌بینی است. جنبه‌های عملکردی پروژه-های مدل‌سازی پیش‌بینی عموماً نادیده گرفته می‌شوند و کم‌تر پیش می‌آید که در کتاب‌ها پوشش داده شوند. این امر در ابتدا شامل مشخص کردن سخت‌افزارها و نرم‌افزارهای موردنیاز برای مدل‌سازی پیش‌بینی است. همان‌گونه که جارد توضیح می‌دهد این موضوع بستگی به سازمان، داده و تحلیل‌گرانی دارد که بر روی پروژه کار می‌کنند. اگر منابع مناسب در اختیار تحلیل‌گران قرار نگیرد پروژه با مشکل برخورد مواجه شده و معمولاً شکست می‌خورد. به‌شخصه درزمانی که با در اختیار داشتن سخت‌افزاری نامناسب بر روی پروژه‌ای بود شاهد این مسئله بودم. این مشکل باعث شد مجبور شوم زمان قابل‌توجهی را صرف کاربر روی محدودیت‌های رم و سرعت پردازش کنم.

درنهایت موفقیت پروژه‌های مدل‌سازی پیش‌بینی بر اساس معیارهای سازمانی که از آن استفاده می‌کند سنجیده می‌شود. مسائلی مانند افزایش بهره‌وری، بازگشت سرمایه، ارزش طول عمر مشتری[1] و یا معیارهای نرم‌تر مانند اعتبار شرکت. من به نمونه‌های موردی مطرح‌شده در این کتاب که به این موضوع اشاره دارند بسیار علاقه‌مندم. تعداد زیادی از این نمونه‌ها در اینجا مطرح‌شده‌اند که می‌توانند توجه شما را جلب کنند. این موضوع به‌خصوص برای مدیرانی که سعی در درک چگونگی تأثیر مدل‌سازی پیش‌بینی بر دستاورد نهایی‌شان دارند بسیار حائز اهمیت است.

مدل‌سازی پیش‌بینی یک دانش است؛ اما به‌کارگیری موفق راه‌حل‌های آن نیازمند مربوط کردن مدل‌ها به کسب‌وکار است. تجربه امری ضروری برای تشخیص این ارتباط است و این کتاب گنجی از تجاربی است که می‌توانند شما را در سفر مدل‌سازی پیش‌بینی هدایت کنند.

دین ابوت[2]

شرکت تجزیه‌وتحلیل ابوت[3]

مارس ۲۰۱۴

---

[1] Customer lifetime value
[2] Dean Abbott
[3] Abbott Analytics

مقدمه

پروژه این کتاب برای اولین بار در اولین هفته کاری من در شغل کنونیم به‌عنوان مدیر داده‌کاوی اس ای اس[1] به من پیشنهاد داده شد. نوشتن یک کتاب همواره یکی از آرزوهای من بوده است پس درگیر شدن با آن برای من هیجان زیادی در پی داشت. من متوجه شدم که چرا بسیاری از مردم می‌خواهند که یک کتاب بنویسند اما تنها اندکی از آن‌ها آن‌قدری خوش‌شانس هستند که بتوانند عقاید و افکارشان را منتشر کنند.

من این فرصت را داشته‌ام که در طول تحصیل و فعالیت حرفه‌ایم پیشگام و یا در مرکز پیشرفت‌های مهمی درزمینهٔ داده‌کاوی باشم و پژوهش‌هایم را تحت نظارت ذهن‌های برجسته‌ای به انجام برسانم. این تجارب به من کمک کردند تا مهارت و تجربه لازم برای خلق این اثر را کسب کنم.

داده‌کاوی حوزه‌ای است که به آن عشق می‌ورزم. من از زمان کودکی می‌خواستم درک کنم که چیزها چگونه با بازدهی معمولی و حتی خیلی بالا کار می‌کنند. از دوران دبستان تا دبیرستان فکر می‌کردم مهندس می‌توانم شغلی باشد که کنجکاوی و علاقه من به توضیح دنیای پیرامونم را به‌صورت هم‌زمان ارضا کند بااین‌وجود در دوران کارشناسی آمار را کشف کردم و به دام افتادم. در بخش اول کتاب به تاریخچه سخت‌افزار و معماری سامانه پرداخته‌ام. این علاقه‌ای است که والدینم مرا با مهربانی به آن تشویق کردند، آن‌هم زمانی که کامپیوترها قیمتی خیلی خیلی بیشتر از ۲۹۹ دلار داشتند. اولین کامپیوتر من یک اپل آی سی با دو فلاپی درایو ۵.۲۵ بود و هارد درایو هم نداشت. چند سال بعد با خرید یک کامپیوتر اینتل ۳۸۶ ساختم و به‌خوبی فشردن دکمه توربو را برای بالا بردن سرعت ساعت سی پی یو از ۸ مگاهرتز به ۱۶ مگاهرتز را به یاد دارم. من قانون مور را به‌صورت تجربی اول دیدم و هنوز این مسئله که تلفن هوشمندم قدرت محاسبه‌ای بیش از کامپیوترهای برنامه فضایی مشتری، برنامه فضایی آپولو و برنامه شاتل فضایی مدارگرد به صورتی ترکیبی داشتند من را شگفت‌زده می‌کند. پس از پایان دوره لیسانس در آمار کار خود را در اداره سرشماری دولت فدرال ایالات‌متحده آغاز کردم. آنجا من برای اولین بار با کلان داده مواجه شدم. پیش از پیوستن به اداره سرشماری من هرگز یک برنامه کامپیوتری که اجرای آن به زمانی بیش از یک دقیقه نیاز داشت ننوشته بودم، مگر این‌که هدف نوشتن برنامه‌ای بود که اجرایش به بیش از یک دقیقه زمان احتیاج

[1] SAS

داشت. یکی از اولین پروژه‌های من کار با پرونده اصلی نشانی‌ها بود، یعنی یک فهرست نشانی که توسط اداره سرشماری نگهداری می‌شد. این فهرست نشانی‌ها همچنین شامل نظرسنجی-های اولیه‌ای است که چارچوب نظرسنجی‌های امروزی است که اداره سرشماری انجام می‌دهد و برای این امر در نه سال بعدی کارهای زیاد دیگری صورت گرفت. این فهرست ۳۰۰ میلیون مورد ثبت‌شده دارد که اطلاعات مربوط به آدرس، طول جغرافیایی و اطلاعات جغرافیایی در آن‌ها باهم ترکیب شده‌اند. صدها جنبه مختلف در رابطه با هر مسکن در این فهرست موجود است. من در زمان کار کردن با چنین مجموعه‌ای از اطلاعات بهره وری، مقیاس پذیری و بهینه سازی ساخت‌افزاری در برنامه را آموختم.

قدردان مدیر صبورم "ماری آن"[1] هستم که به من فرصت آموختن داد و پروژه‌های جالب و ارزشمندی را در اختیارم گذاشت که به من تجربه عملی و امکان نوآوری دادند. این شغل بسیار مهم بود چرا که من توانستم تکنیک‌ها و رویکردهای جدیدی را به کار گیرم که پیش از آن در آن دپارتمان در مورد آن تحقیق نشده بود. درست مثل تمام پروژه‌های جدید دیگر بعضی نظرات عملی شدند و بعضی با شکست مواجه گشتند. یکی از پروژه‌های مشخصی که من در آن مشارکت داشتم مشخص کردن این موضوع بود که کدام بلوک در سرشماری سال ۲۰۰۰ بیش شماری و یا کم شماری شده است (اداره سرشماری ایالات‌متحده به مناطق جغرافیایی مخصوصی شامل ایالت، شهرستان، منطقه، مجموعه بلوک‌ها و بلوک تقسیم شده که تقریباً ۸,۲ میلیون بلوک در ایالات‌متحده وجود دارد). ما راهی برای مشخص کردن دقت مدلمان برای پیش‌بینی انحراف تعداد واقعی واحدهای مسکونی نسبت به تعداد واحدهای مسکونی گزارش شده از طریق داده‌های موجود نداشتیم. خوشبختانه پروژه برای انجام تحقیقات میدانی برای تهیه بازخورد و اعتبار مدل از بودجه مجلس بهره مند بود. این ولی، اولین باری بود که من عبارت "داده‌کاوی" را می‌شنیدم و برای بار اول بود که با سامانه SAS™ Enterprise Miner and CARTR by Salford مواجه می‌شدم. بعد از مدتی کار در اداره سرشماری، دریافتم که برای رسیدن به اهداف کاریم نیاز به تحصیلات بیشتری دارم و در نتیجه در دپارتمان آمار دانشگاه جرج میسون در فایرفاکس، وی ای[2] ثبت نام کردم.

در طول دوره کارشناسی ارشد جزئیات بیشتری در مورد الگوریتم‌های معمول در حوزه داده‌کاوی، یادگیری ماشین و آمار آموختم. تجزیه‌وتحلیل بقا، نمونه گیری در پژوهش و آمار محاسباتی بخشی از این موضوع بودند. در زمان تحصیل در دوره کارشناسی ارشد توانستم

---

[1] Maryann

[2] George Mason University in Fairfax, VA.

درس‌هایی که در کلاس آموزش داده می‌شدند را با تجزیه‌وتحلیل داده‌ای که به‌صورت عملی در دفتر کارم انجام می‌شد ترکیب، و نوآوری‌های لازم را به آن اضافه کنم. من توانستم تئوری و قوت و ضعف‌های رویکردهای مختلف تجزیه‌وتحلیل داده و آنالیز پیش‌بینی مرتبط با آن تئوری‌ها را درک کنم.

پس از پایان دوره کارشناسی ارشد مسیر شغلیم را از تجزیه‌وتحلیل داده به تولید نرم افزار تغییر دادم. پس شروع به کار در موسسه اس ای اس کردم که خالق همان نرم افزاری بود که من قبلاً از آن استفاده می‌کردم. من از مرحله استفاده از نرم افزار به تولید آن تغییر مکان داده بودم. این مسئله چالش‌ها و فرصت‌های جدیدی برای رشد در بر داشت چرا که متوجه صحت و دقت جدی اس ای اس در نرم افزارها در کنار مستندات همه جانبه‌شان، تلاش خستگی ناپذیر آن‌ها برای رسیدن به پیشرفت‌های جدید نرم افزاری قابل انطباق با نرم‌افزارهای موجود و ارائه خصوصیات نرم افزاری جدید موردنیاز مشتریان شدم.

در سال‌هایی که در اس ای اس کار می‌کردم کاملاً فهمیدم که نرم افزار چگونه تولید می‌شود و مشتریان چگونه از آن استفاده می‌کنند. من معمولاً بخت آن را دارم که با مشتریان دیدار کنم، چالش‌های شغلی آن‌ها را بشنوم و روش‌ها یا فرآیندهایی را به آن‌ها پیشنهاد بدهم که به موفقیت رهنمونشان کند و برای سازمانشان ایجاد بهره کند.

این کتاب در نتیجه این مجموعه تجارب و در کنار کمک‌های کارمندان و همکاران بی نظیرم هم در موسسه اس ای اس و هم در جاهای دیگر نوشته شده است.

مقدمه

درون این پشته داده دانشی نهفته است که می‌تواند زندگی یک بیمار و یا جهان را دگرگون کند.

اتول بوت، دانشگاه استنفورد[1]

سرطان نامی است که به دسته‌ای از بیماری‌ها داده شده که در آن‌ها سلول‌های غیرمعمول به شکلی غیر قابل واپایش تقسیم می‌شوند و به بافت‌های بدن حمله می‌کنند.

بیش از ۱۰۰ نوع منحصر به فرد و متفاوت از سرطان وجود دارند که بیشترشان بر اساس جایی (معمولاً یک عضو) که بیماری از آن آغاز شده نام گزاری شده‌اند. سرطان از سلول‌های بدن آغاز می‌شود. در شرایط طبیعی بدن میزان تولید سلول‌های جدید جهت جایگزینی با سلول‌های فرسوده یا آسیب دیده را کنترل می‌کند. سرطان طبیعی نیست. در بیماران سرطانی سلول‌ها در زمان مقرر نمی‌میرند و سلول‌های جدید درزمانی که به آن‌ها احتیاجی نیست تولید می‌شوند (درست مثل موقعی که من از بچه‌هایم می‌خواهم برایم چیزی را از فروشگاه روگرفت چاپ کنند و آن‌ها به جای یک نسخه ده نسخه از چیزی که خواسته بودم برایم می‌آورند). سلول‌های اضافی تشکیل یک توده بافت را می‌دهند که به آن تومور گفته می‌شود. تومورها به دو دسته تقسیم می‌شوند: تومورهای خوش خیم که سرطانی نیستند و تومورهای بدخیم که سرطانی هستند. تومورهای بدخیم در سراسر بدن پخش می‌شوند و به بافت حمله می‌کنند. خانواده من، مانند بیشتر خانواده‌هایی که می‌شناسم، اعضای زیادی دارد که این بیماری را دارند. در سال ۲۰۱۳ در ایالات‌متحده ۱٫۶ میلیون مورد سرطانی جدید تخمین زده شده است و بیش از ۵۸۰۰۰۰ مرگ در اثر این بیماری رخداده است.

[1] Atul Butte, Stanford University

در حدود ۲۳۵۰۰۰ نفر در ایالات‌متحده در سال ۲۰۱۴ مبتلا به سرطان پستان تشخیص داده‌شده‌اند و در حدود ۴۰۰۰۰ نفر در سال ۲۰۱۴ در نتیجه ای بیماری در خواهند گذشت. شایع‌ترین نوع سرطان پستان کارسینوم داکتال[1] است که از پوشش داخلی مجاری شیری آغاز می‌شود. دومین گونه شایع سرطان پستان نوع لوبولار[2] است. درمان‌های متفاوتی از جمله عمل جراحی، شیمی درمانی، پرتو درمانی، ایمن درمانی و واکسن درمانی برای سرطان پستان وجود دارند. معمولاً یک یا بیش از یک گزینه درمانی برای اطمینان از دست یابی به بهترین نتیجه برای بیمار استفاده می‌شوند. سازمان غذا و دارو[3] در حدود ۶۰ داروی مختلف را برای درمان سرطان پستان تائید کرده است. تصمیمات درزمینهٔ دوره درمان و انتخاب پروتکل درمانی مورد استفاده در مشاوره‌های بین پزشک و بیمار صورت می‌گیرند و برخی فاکتورها به همین تصمیمات رمی‌گردند.

یکی از داروهای مورد تائید سازمان غذا و دارو برای درمان سرطان تاموکسیفن سیترات[4] است. این دارو که تحت نام تجاری نولوادکس[5] به فروش می‌رسد اولین بار در سال ۱۹۶۹ در انگلستان تجویز و در سال ۱۹۹۸ توسط سازمان غذا و دارو تائید شده است. تاموکسیفن معمولاً به‌صورت قرص‌های روزانه و در دوزهای ۱۰، ۲۰ و ۴۰ میلی گرم مصرف می‌شود. دارو دارای اثراتی جانبی از جمله تهوع، سوء هاضمه و گرگرفتگی (ضلاً ساق پا است. تاموکسیفن برای درمان میلیون‌ها زن و مرد مبتلا به سرطان پستان وابسته به هورمون به کار گرفته شده است. معمولاً تاموکسیفن به دلیل میزان موفقیت بالا، یعنی در حدود ۸۰٪، اولین دارویی است که برای درمان سرطان پستان تجویز می‌شود. علم به این موضوع که یک دارو ۸۰٪ موفق بوده ما را به تأثیر آن بر روی بیماران امیدوار می‌کند. اما یک نکته مهم در مورد دارو تا پیش از دوران کلان داده ناشناخته بود: این‌که تاموکسیفن، ۸۰٪ بر روی بیماران تأثیر (ذا نموده است، بلکه در ۸۰٪ بیماران ۱۰۰٪ تأثیر گذار و در باقی موارد بی تأثیر بوده است. این یافته سالانه زندگی هزاران نفر را تغییرداده است. با استفاده از تکنیک‌ها و ایده‌های مطرح‌شده در این کتاب دانشمندان توانسته‌اند نشانگرهای ژنتیکی را شناسایی کنند که در همان ابتدا مشخص می‌کنند که آیا تاموکسیفن شخص مبتلا به سرطان پستان را به طور موثر درمان خواهد کرد یا

---

1. ductal carcinoma
2. lobular
3. Food and Drug Administration (FDA)
4. tamoxifen citrate
5. Nolvadex

خیر. این گونه از تجزیه‌وتحلیل پیش از دوران کلان داده ممکن نبود. چرا چنین چیزی امکان پذیر نبود؟ چرا که حجم و دانه‌بندی داده وجود نداشت. حجم از ادغام نتایج مربوط به بیماران و دانه‌بندی از ترتیب گذاری دی ان ای به دست می‌آیند. علاوه بر داده، منابع محاسباتی برای حل مسائلی از این دست برای بیشتر دانشمندانی که بیرون از آزمایشگاه محاسبات عالی بودند به سادگی در دسترس نبود. در آخر، سومین عنصر یعنی الگوریتم‌ها و یا تکنیک‌های مدل‌سازی برای درک این رابطه در سال‌های اخیر پیشرفت قابل‌توجهی کرده‌اند.

داستان تاموکسیفن فرصت‌های هیجان انگیزی را که داشتن داده‌های بیشتر و بیشتر در کنار منابع محاسباتی و الگوریتم‌ها برای ما فراهم می‌کنند را مشخص‌تر می‌کند. این موضوع به ما در طبقه بندی و پیش‌بینی نیز یاری می‌رساند. با چنین دانشی که دستاورد دانشمندانی است که بر روی تاموکسیفن تحقیق کرده‌اند ما می‌توانیم تغییر درمان را آغاز کنیم و بسیاری از حوزه‌های دیگر پزشکی را نیز به‌صورت مثبت متحول کنیم. ما می‌توانیم با استفاده از این پیشرفت‌ها از درمان موسوم همه بیماران جلوگیری کنیم و به جای آن مشخص کنیم که کدام درمان خاص برای هر شخص مفید است. دیگر یک دارو ۵٪ تأثیر گذار نخواهد بود، بلکه ما اکنون می‌توانیم بگوییم دارو به کدام ۵٪ از این بیماران کمک خواهد کرد. مفهوم داروی شخصی سال‌ها است که مطرح‌شده است. با پیشرفت‌هایی که در کار با کلان داده و آنالیز پیش‌بینی صورت گرفته است این مفهوم بیش از هر زمان دیگری به حقیقت نزدیک شده است. دارویی با ۲٪ نرخ موفقیت تا زمانی که مشخص نشود برای کدام دسته از بیماران مفید است توسط تولیدکنندگان دارو دنبال و از جانب سازمان غذا و دارو نادیده گرفته خواهد شد. در صورت وجود این اطلاعات می توان زندگی‌ها را نجات داد. تاموکسیفن یکی از مثال‌های بسیاری است که پتانسیل‌هایی که در بهره‌گیری از منابع محاسباتی و صبوری در وقت از بین در داده‌های اطراف ما وجود دارد را به ما نشان می‌دهد.

ما در حال حاضر در دوران کلان داده زندگی می‌کنیم. عبارت "کلان داده" درزمانی نزدیک به شروع دوره کلان داده ایجاد شد. همان قدر که من تاریخ شروع دوره کلان داده را سال ۲۰۰۱ در نظر می‌گیرم، این تاریخ منبع مناظرات و بحث‌های مهیج در وبلاگ‌ها و حتی نیویورک تایمز هم بوده است. به نظر می‌رسد که عبارت "کلان داده" به مفهوم امروزی آن برای اولین بار در اواخر دهه نود به کار گرفته شده است. اولین مقاله در این رابطه با عنوان "مدل‌های کلان داده

با فاکتور دینامیک برای سنجش اقتصاد کلان و پیش‌بینی"[1] در سال ۲۰۰۳ توسط فرانسیس اکس دایبولت[2] منتشر شد، گرچه اعتبار این موضوع بیشتر به جان مشی[3]، دانشمند ارشد اس جی آی[4]، به‌عنوان اولین کسی که عبارت "کلان داده" را به کاربرده است داده می‌شود. در اواخر دهه نود مشی برای گروه کوچکی از افراد در مورد موج جزر و مدی کلان داده که در حال وقوع بود یک مجموعه سخنرانی را ترتیب داد. دوره کلان داده دوره‌ای است که با حجم به سرعت در حال گسترش داده به‌صورتی بزرگ‌تر از آنچه بیشتر افراد حتی قادر به تصور وقوع آن نبودند تعریف می‌شود.

حجم بزرگ داده به تنهایی مشخص کننده دوره کلان داده نیست، چرا که همواره حجم‌هایی از داده وجود داشته‌اند که بیشتر از حد توان ما برای کار کردن به‌صورت موثر بر روی این داده‌های موجود بوده‌اند. چیزی که این دوران داده را از دوره‌های دیگر متمایز می‌کند این است که شرکت‌ها، دولت‌ها و سازمان‌های غیر انتفاعی دستخوش تغییر در رفتار شده‌اند. در این دوره آن‌ها می‌خواهند شروع به جمع‌آوری تمام داده‌های ممکن برای اهداف ناشناخته کنونی و یا آتی کنند تا بتوانند کسب‌وکار خود را پیشرفت دهند. بسیاری بر این باورند که سازمان‌هایی که در تصمیم‌گیری‌هایشان از داده به همراه پشتیبانی پژوهش و نمونه موردی‌های مربوط به آن استفاده می‌کنند در طولانی مدت تصمیمات بهتری می‌گیرند. تصمیماتی که به کسب‌وکاری قدرتمندتر و ماندگارتر منتهی خواهد شد. شرکت‌ها در پاسخ به رشد پر سرعت داده‌های تولید شده شروع به جمع‌آوری داده در بالاترین حد ممکن و ارزش دادن بیش از پیش به پتانسل‌های آن برای موارد آتی کرده‌اند. ما چه مقدار داده شخصی تولید می‌کنیم؟ اولین سوال این است: داده شخصی چیست؟ در سال ۱۶۹۵ اتحادیه اروپایی در حیطه قوانین اشخاص، داده شخصی را هرگونه اطلاعاتی عنوان کرد که می‌توانند شخصی را به‌صورت مستقیم و یا غیر مستقیم از دیگران مشخص کند. شرکت بین‌المللی داده تخمین زده است که ۲,۸ زتابایت داده در سال ۲۰۱۲ ایجاد شده است و میزان داده تولید شده در هر سال تا سال ۲۰۱۵ دو برابر خواهد شد. با چنین عدد بزرگی درک این‌که چه میزان از این داده‌ها درواقع مربوط به شما هستند مشکل است. این عدد به چیزی در حدود ۵ گیگابایت در روز برای هر کارمند

---

[1] Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting

[2] Francis X. Diebolt

[3] John Mashey

[4] SGI

متوسط ایالات متحده تقسیم می‌شود. این داده‌ها شامل ای میل، دانلود فیلم، فایل‌های شنیداری در جریان، صفحات گسترده اکسل و غیره می‌شود. این داده‌ها همچنین شامل داده-هایی می‌شود که به‌عنوان اطلاعات در اینترنت ایجاد می‌شوند. بسیاری از این داده‌های تولید شده مستقیماً توسط شما یا من دیده نمی‌شوند بلکه در مورد ما ذخیره می‌شوند. طول چراغ-های راهنمایی و رانندگی، جی پی اس که از طریق تلفن‌های ما مختصات را مشخص می‌کند یا میزان دریافت عوارضی زمانی که ما به سرعت از میان باجه‌های الکترونیکی دریافت عوارض می‌گذریم مثالی برای این نوع از داده‌ها هستند.

پیش از شروع دوران کلان داده کسب‌وکارها برای داده‌هایی که ارزش آنی نداشتند به نسبت ارزش کمی قائل بودند. وقتی دوران کلان داده آغاز شد سرمایه‌گذاری بر روی داده‌هایی که پتانسیل مفیدی واقع آن در آینده را داشتند تغییر کرد و سازمان‌ها شروع به تلاش آگاهانه برای نگه‌داری داده‌هایی از داده با پتانسیل مربوطه کردند. این تغییر رفتاری دایره بزرگی را ایجاد کرد که در آن داده ذخیره می‌شد و سپس به دلیل ارزشمند بودن داده اشخاص بر روی یافتن ارزش در داده برای سازمان سرمایه‌گذاری می‌کردند. موفقیت در یافتن ارزش در داده به جمع‌آوری بیشتر داده منتهی شد. برخی از این داده‌ها به جایی رهنمون نمی‌گشتند اما در بسیاری از موارد نتیجه ثابت کرد که هر چه داده بیشتری داشته باشیم امکان رسیدن به مطلوب برایمان بیشتر است. تغییر بزرگ دیگری که در ابتدای دوران کلان داده رخ داد رشد و تولید سریع و پختگی در تکنولوژی‌های ذخیره، اداره و آنالیز این داده‌ها به‌صورت موثر بود.

حال که در دوران کلان داده هستیم چالش ما جمع‌آوری داده نیست بلکه استفاده از کامپیوترها برای افزودن به دامنه دانشمان و تشخیص الگوهایی است که پیش از این قادر به دیدن و یا یافتن آن‌ها نبوده‌ایم.

برخی از تکنولوژی‌های کلیدی و افت‌های پیش آمده در بازار منجر به افزایش جمع‌آوری، ذخیره و استفاده از داده در کارهای مربوط به تجزیه‌وتحلیل، به میزان بسیار زیاد شد. این مسئله ناشی از فاکتورهای زیادی از جمله پروتکل اینترنتی نسخه ۶، وسائل پیشرفته ارتباط از راه دور، تکنولوژی‌هایی مانند آر اف آی دی، حس‌گرهای کامپیوتری-مخابراتی، کاهش هزینه تولید هر واحد الکترونیکی، رسانه‌های گروهی و اینترنت است.

در ادامه جدول زمانی‌ای را مشاهده خواهید کرد که در آن برخی از اتفاقات کلیدی که به دوران کلان داده منجر شده‌اند و وقایعی که به شکل‌گیری و استفاده از کلان داده در آینده کمک کرده‌اند، مشخص شده‌اند.

جدول زمان‌بندی کلان داده:

در اینجا تعدادی از مواردی که نشان‌گر اتفاقات تأثیر گذار بر آماده شدن مسیر برای دوران کلان داده و نقاط عطف در طول این دوره هستند آورده شده‌اند.

۱۹۹۱

- اینترنت با معماری شبکه جهانی متولد شد. پروتکل انتقال ابرمتن (HTTP) به وسیله‌ای استاندارد برای به اشتراک گذاشتن اطلاعات در این رسانه نوین تبدیل شد.

۱۹۹۵

- سان[1] پلت فورم جاوا[2] را منتشر کرد. جاوا که در سال ۱۹۹۱ ابداع شد تبدیل به دومین زبان رایج پس از C شد. جاوا در فضای برنامه‌های کاربردی وب غالب می‌شود و استانداردی بالفعل برای برنامه‌های کاربردی رده متوسط است. این برنامه‌ها منبعی برای ضبط و ذخیره شدآمد وب هستند.
- سیستم جهانی تعیین موقعیت (GPS) به‌صورت کامل عملکرد پیدا کرد. جی پی اس در اصل توسط دارپا (آژانس پروژه‌های پژوهشی پیشرفته دفاعی[3]) برای برنامه‌های کاربردی نظامی در ابتدای دهه ۷۰ تولید شد. این تکنولوژی سپس در برنامه برنامه‌های مربوط اتوموبیل، ناوبری هواپیما و یافتن آی فون حضور یافت.

۱۹۹۸

[1] Sun
[2] Java platform
[3] DARPA (Defense Advanced Research Projects Agency)