

برنامہٴ دانشجو

علم دادہ برای کتابداران



یون فی دو - حماد رئوف خان

ترجمہ:

دکتر محسن نوکاریزی

استاد دانشگاه فردوسی مشهد

سرشناسه:	دو، یونفی، ۱۹۷۰-م.	Du, Yunfei, 1970-
عنوان و نام پدیدآور:	علم داده برای کتابداران / یون فی دو، حماد رئوف خان؛ ترجمه محسن نوکاریزی؛ ویراستار ادبی زهره عدالتیان.	علم داده برای کتابداران / یون فی دو، حماد رئوف خان؛ ترجمه محسن نوکاریزی؛ ویراستار ادبی زهره عدالتیان.
مشخصات نشر:	مشهد: دانشگاه فردوسی مشهد، انتشارات، ۱۴۰۴.	مشهد: دانشگاه فردوسی مشهد، انتشارات، ۱۴۰۴.
مشخصات ظاهری:	۸+۲۲۸ ص: جدول، نمودار.	۸+۲۲۸ ص: جدول، نمودار.
فروست:	انتشارات دانشگاه فردوسی مشهد؛ ۹۵۵.	انتشارات دانشگاه فردوسی مشهد؛ ۹۵۵.
شابک:		ISBN: 978-964-386-657-0
وضعیت فهرست نویسی:	فیبا.	فیبا.
یادداشت:	عنوان اصلی:	Data science for librarians, [2020].
یادداشت:	کتاب حاضر با همین عنوان و ترجمه مریم حسن نژاد، کیانوش طافی توسط انتشارات دانشگاه صنعتی جندی شاپور دزفول در سال ۱۴۰۳ (۳۰۰ ص.) منتشر شده است.	کتاب حاضر با همین عنوان و ترجمه مریم حسن نژاد، کیانوش طافی توسط انتشارات دانشگاه صنعتی جندی شاپور دزفول در سال ۱۴۰۳ (۳۰۰ ص.) منتشر شده است.
یادداشت:	واژه نامه. کتابنامه. نمایه.	واژه نامه. کتابنامه. نمایه.
موضوع:	مدیریت پایگاه های اطلاعاتی در کتابخانه ها	Database management in libraries
	گزینش و مدیریت داده ها در کتابخانه ها	Data curation in libraries
	کتابداران داده	Data services librarians
	تحقیق - داده پردازی - مدیریت	Research -- Data processing -- Management
	داده های کلان	Big data
شناسه افزوده:	خان، حماد رئوف	Khan, Hammad Rauf
شناسه افزوده:	نوکاریزی، محسن، ۱۳۴۵ - مترجم	
شناسه افزوده:	دانشگاه فردوسی مشهد، انتشارات.	
رده بندی کنگره:	Z711/97	
رده بندی دیویی:	۰۰۵/۷۴	
شماره کتابشناسی ملی:	۱۰۱۹۵۶۱۲	

علم داده برای کتابداران

پدیدآورندگان: یون فی دو؛ حماد رئوف خان
ترجمه: دکتر محسن نوکاریزی
ویراستار ادبی: زهره عدالتیان
مشخصات: وزیری، ۲۰۰ نسخه، چاپ اول، پاییز ۱۴۰۴
چاپ و صحافی: همیار
بها: ۲/۸۰۰/۰۰۰ ریال
حق چاپ برای انتشارات دانشگاه فردوسی مشهد محفوظ است.



مراکز پخش:

فروشگاه و نمایشگاه کتاب پردیس: مشهد، میدان آزادی، دانشگاه فردوسی مشهد، جنب سلف یاس
تلفن: ۳۸۸۰۲۶۶۶ - ۳۸۸۳۳۷۲۷ (۰۵۱)
مؤسسه کتابیران: تهران، میدان انقلاب، خیابان کارگر جنوبی، بین روانمهر و وحید نظری، بن بست گشتاسب، پلاک ۸
تلفن: ۶۶۴۸۴۷۱۵ (۰۲۱)
مؤسسه دانشیران: تهران، خیابان انقلاب، خیابان منیری جاوید (اردیبهشت) نبش خیابان نظری، شماره ۱۴۲
تلفکس: ۶۶۴۰۰۲۲۰ - ۶۶۴۰۰۱۴۴ (۰۲۱)

<http://press.um.ac.ir>

Email: press@um.ac.ir

فهرست مندرجات

۱.....	سخن مترجم
۳.....	فصل ۱. داده افزون تر، مشکلات بیش تر
۴.....	مفهوم داده
۵.....	داده‌های کمی در برابر کیفی
۶.....	داده‌های رقومی در برابر غیررقومی
۷.....	مفهوم کلان داده
۹.....	چگونگی تأثیر کلان داده
۹.....	مشکلات برخورداری از داده زیاد
۱۰.....	تفاوت داده و اطلاعات
۱۰.....	اشباع داده
۱۰.....	سوگیری تأیید و خطای سیگنال
۱۱.....	اثرات کلان داده بر جامعه
۱۱.....	تأثیر بر خدمات بهداشتی
۱۱.....	برنامه‌ریزی دولت
۱۱.....	استفاده از اخبار و رسانه‌ها
۱۲.....	ورزش‌ها
۱۲.....	کلان داده و سیل داده‌ها
۱۳.....	داده باز
۱۵.....	داده باز دولتی
۱۵.....	وبگاه داده‌های دولت
۱۶.....	اصول داده باز دولتی
۱۶.....	داده‌های پژوهشی در کتابخانه‌های دانشگاهی
۱۷.....	مفاهیم سواد داده
۱۸.....	چرخه حیات داده
۱۹.....	عصر کلان داده

چشم‌انداز آینده.....	۲۱
منابع.....	۲۲

فصل ۲. شاخه جدید کتابداری.....	۲۵
تصمیم‌گیری مبتنی بر داده	۲۵
تاریخچه داده در کتابخانه‌های دانشگاهی.....	۲۸
منظور از کلان‌داده برای کتابخانه.....	۳۰
کتابداری داده	۳۱
خدمات داده‌های پژوهشی.....	۳۴
طرح‌های مدیریت داده	۳۶
مدیریت: سامانه‌های اطلاعات جغرافیایی (جی‌آی‌اس).....	۳۸
نتیجه‌گیری	۴۱
منابع.....	۴۱

فصل ۳. خلق و گردآوری داده.....	۴۵
پیمایش‌ها.....	۴۵
بزار بر خط	۴۷
داده‌های رسانه‌های اجتماعی.....	۴۸
اختلال داده	۴۸
فراهم‌آوری داده.....	۵۱
معایب گردآوری کلان‌داده.....	۵۲
تحلیل نظام‌مند کلان‌داده	۵۴
نتیجه‌گیری	۵۶
منابع.....	۵۷

فصل ۴. داده برای کتابدار دانشگاهی.....	۵۹
علم الکترونیکی و پژوهش الکترونیکی.....	۶۰
مصاحبه مرجع داده	۶۲
ذخیره‌سازی و آرشیو داده‌ها.....	۶۶
مخازن واسپاری.....	۶۸
منابع.....	۷۱

فصل ۵. خدمات داده‌های پژوهشی و زیست‌بوم کتابخانه.....	۷۵
ماهیت خدمات داده‌های پژوهشی.....	۷۶
میزان بازنمایی چرخه حیات داده‌های پژوهشی در خدمات.....	۷۷
داده‌های پژوهشی.....	۷۷
افراد مشغول در بخش خدمات داده‌های پژوهشی.....	۸۱
سواد داده.....	۸۳
منابع.....	۸۷
فصل ۶. منابع داده‌ها.....	۸۹
حرفه‌مند داده و کتابداری.....	۸۹
داده‌های باز دولت.....	۹۰
مخازن واسپاری داده.....	۹۱
ابرداده.....	۹۳
استناد به داده‌ها.....	۹۵
استخراج، تبدیل، و بارگیری داده.....	۹۸
داده کاوی.....	۹۹
پاک‌سازی داده.....	۱۰۰
داده کاوی و تحلیل داده برای کتابداران.....	۱۰۱
داده کاوی: فنون.....	۱۰۳
داده کاوی: مزایا و معایب.....	۱۰۴
تحلیل داده و کتابداران: مروری کلی.....	۱۰۶
نتیجه‌گیری.....	۱۰۷
منابع.....	۱۰۸
فصل ۷. گزینش و مدیریت داده (آرشیو و حفاظت و نگهداری از داده‌ها).....	۱۱۱
فرایند گزینش و مدیریت داده.....	۱۱۱
نظارت بر داده.....	۱۱۳
ابرداده.....	۱۱۴
دسترسی به داده و استفاده مجدد از آن.....	۱۱۷
به اشتراک گذاری داده.....	۱۱۸
کیفیت داده.....	۱۱۹

نتیجه گیری	۱۲۱
منابع	۱۲۱

فصل ۸. ذخیره سازی، مدیریت و بازیابی داده	۱۲۵
راه حل های ذخیره سازی کلان داده	۱۲۵
رایانش کارآمد	۱۲۷
انواع الگوهای ذخیره سازی کلان داده	۱۲۸
داده های شبکه های اجتماعی	۱۳۰
رایانش ابری	۱۳۱
هadoop آپاچی	۱۳۲
راه حل های رایج ذخیره سازی ابری	۱۳۳
دغدغه های حریم خصوصی درباره رایانش ابری	۱۳۴
مدیریت کلان داده	۱۳۵
پاک سازی داده	۱۳۶
امنیت و خط مشی های کلان داده	۱۳۷
مدیریت سرعت کلان داده	۱۳۸
نتیجه گیری	۱۳۹
منابع	۱۴۰

فصل ۹. تحلیل و مصور سازی داده ها	۱۴۵
تحلیل کلان داده	۱۴۵
تحلیل نظام مند توصیفی	۱۴۶
تحلیل نظام مند تشخیصی	۱۴۷
تحلیل نظام مند پیشگویانه	۱۴۷
تحلیل تجویزی	۱۴۸
آمار برای علم داده	۱۴۸
آزمون فرضیه و معناداری آماری	۱۴۸
توزیع های احتمال	۱۴۹
همبستگی	۱۵۲
رگرسیون	۱۵۲
مصور سازی	۱۵۵

تاریخچه مختصر مصورسازی داده	۱۵۵
روش ها و ابزارهای مصورسازی داده	۱۵۶
مصورسازی متن	۱۵۹
برنامه های کاربردی مصورسازی	۱۶۱
نتیجه گیری	۱۶۳
منابع	۱۶۳

فصل ۱۰. اخلاق و خط مشی های داده	۱۶۷
امنیت داده ها	۱۶۸
حریم خصوصی کاربر و حفظ داده ها	۱۷۰
حریم خصوصی	۱۷۲
اخلاق داده ها	۱۷۴
حق تکثیر و مالکیت	۱۷۵
داده های اطلاعات شخصی در کتابخانه ها	۱۷۶
نتیجه گیری	۱۷۸
منابع	۱۷۹

فصل ۱۱. داده برای کتابخانه های عمومی و کتابخانه های تخصصی	۱۸۱
طرح های ابتکاری شهرهای هوشمند	۱۸۲
طرح های ابتکاری دولت باز	۱۸۴
اینترنت اشیا و دغدغه های حریم خصوصی	۱۸۶
اینترنت اشیا	۱۸۶
چالش ها	۱۸۷
داده های سرشماری	۱۸۸
نقش کتابخانه های عمومی در عصر کلان داده	۱۸۹
امکان استفاده کتابخانه های عمومی از کلان داده برای پرداختن به نیازهای محلی	۱۸۹
کتابخانه ها در جایگاه طرفداران حریم خصوصی شهروندان	۱۹۰
کتابداران داده در کتابخانه های عمومی	۱۹۰
کتابخانه های عمومی به منزله مراکز یادگیری برای نوجوانان	۱۹۱
نقش کتابخانه های تخصصی در عصر کلان داده	۱۹۲
کتابداران حقوق	۱۹۳

کتابخانه‌های شرکتی.....	۱۹۴
کتابداران پزشکی.....	۱۹۵
نتیجه‌گیری.....	۱۹۶
منابع.....	۱۹۶

فصل ۱۲. نتیجه‌گیری: علم کتابداری، علم اطلاعات، و علم داده.....	۲۰۱
داده در حکم زیرساختی برای جامعه.....	۲۰۲
داده و اطلاعات.....	۲۰۲
داده به‌منزله کالای عمومی.....	۲۰۲
داده به‌منزله نیروی محرکه اقتصاد.....	۲۰۳
داده برای حاکمیت.....	۲۰۳
کتابداران و چرخه حیات داده.....	۲۰۴
عناوین جدید شغلی برای کتابداران.....	۲۰۵
کتابداران در چرخه حیات داده.....	۲۰۵
مجموعه مهارت‌های تحلیل داده برای کتابداران.....	۲۰۶
درک و جذب داده.....	۲۰۷
گزینش و مدیریت داده.....	۲۰۷
مصورسازی داده.....	۲۰۸
تحلیل نظام‌مند داده.....	۲۰۸
سواد داده برای کاربران کتابخانه.....	۲۰۹
سواد داده در محیط‌های دانشگاهی.....	۲۰۹
سواد داده برای کاربران کتابخانه عمومی.....	۲۰۹
نتیجه‌گیری.....	۲۱۲
منابع.....	۲۱۲
واژه‌نامه.....	۲۱۵
نمایه.....	۲۲۵

سخن مترجم

کتاب علم داده برای کتابداران نخستین کتابی است که برای این حوزه خاص و با زبان قابل درک برای کتابداران و اطلاع‌رسان توسط دو تن از افراد خبره در این زمینه، یعنی یوفن دیو و حمد رؤوف خان، نگاشته شده است و نقطه عطفی برای پرکردن شکاف عمیق موجود در این حوزه به شمار می‌رود. این کتاب دو ویژگی مهم دارد که درک آن را برای کتابداران و اطلاع‌رسانان بسیار ساده ساخته است. یکی اینکه از هیچ گونه الگوریتم پیچیده برنامه‌نویسی برای صحبت در باب علم داده استفاده نشده است، به بیانی دیگر، کتاب فنی و پیچیده و برای مهندسان علوم رایانه نوشته نشده است، دوم اینکه به هیچ وجه از مباحث آماری پیچیده در آن استفاده نشده است. مهم‌ترین بعد مطرح در این کتاب بعد محتوایی مربوط به علم داده است که کار درک ساده و روان آن را برای اطلاع‌رسان بسیار ساده کرده است و مباحثی است که کتابداران از دیرباز با آن آشنا بوده‌اند ولی از منظر کلان داده به منزله موضوعی جدید و بسیار متواتر در این ایام به آن پرداخته شده است.

یون‌فی دو استاد علم کتابداری و معاون دانشکده اطلاعات دانشگاه تگزاس شمالی بود که سال‌ها در مقام کتابداری نظام‌ها در کتابخانه‌های دانشگاهی کار عملی کرده و آثار متعددی در نشریات این حوزه منتشر کرده است. او برنامه درسی مقاطع مختلف علم داده ارائه شده توسط دانشگاه تگزاس شمالی را طراحی و اجرا کرده است. حماد رؤوف خان نیز دانشجوی دکتری و مدرس پاره‌وقت دانشگاه تگزاس شمالی است که سال‌ها تجربه کاری در کتابخانه‌های عمومی و دانشگاهی دارد. وی دارای آثار منتشر و ارائه شده در کنفرانس‌های ملی و بین‌المللی در حوزه علم کتابداری و اطلاع‌رسانی است. او دارای مقالاتی در حوزه سواد داده، خدمات داده‌های پژوهشی، حرفه‌مندان داده، مدیریت داده، و چرخه حیات داده‌های پژوهشی است.

این کتاب برای تمام کتابداران و اطلاع‌رسانان در تمام مقاطع تحصیلی و نیز اساتید و دانشجویان مفید است و به‌ویژه می‌توان به عنوان منبع درسی برای درسی با عنوان «مبانی علم داده» که در برنامه بازنگری شده جدید دوره کارشناسی رشته علم اطلاعات و دانش‌شناسی طراحی شده، استفاده کرد. امید می‌رود ترجمه این درس گامی در راستای تقویت منابع این حوزه باشد و بتواند مورد استفاده همه حرفه‌مندان قرار گیرد. بی‌شک این ترجمه این اثر دارای ایرادتی خواهد بود که در همین جا مترجم از خوانندگان محترم

تقاضا دارد مشکلات آن را به اینجانب منتقل کنند تا در آینده بتوان به رفع مشکلات پرداخت. در پایان از تمام همکاران و دوستان عزیزی که در این کار مرا یاری کردند، صمیمانه تشکر و قدردانی می‌شود. همچنین از زحمات همکاران محترم گروه نشر آثار علمی دانشگاه فردوسی مشهد، آقایان دکتر هادی هراتی، مصطفی قندهاری، حمیدرضا نداف سنگانی و غلامرضا قاسمی بابت چاپ و نشر این کتاب سپاسگزاری و قدردانی می‌کنم.

دکتر محسن نوکاریزی

تیرماه ۱۴۰۴

داده افزون تر، مشکلات بیش تر

پیشرفت‌های اخیر در رشته‌هایی نظیر علم رایانه و آمار، در کنار تکثیر داده‌ها، اهمیت زیست‌بوم حرفه‌ای جدیدی را با عنوان علم داده به‌منزله حوزه بین‌رشته‌ای افزایش داده است. مؤسسه ملی استانداردها و فناوری^۱ اظهار می‌کند، «علم داده فناوری‌ها، فنون، و نظریه‌هایی از حوزه‌های متنوع عمدتاً وابسته به علم رایانه و آمار را با هم ترکیب می‌کند تا دانشی کاربردی از داده‌ها حاصل شود» (ص. هفت). می‌توانیم این حوزه را به شیوه‌های گوناگون تعریف کنیم، اما علم داده اساساً با ایجاد بینشی ارزشمند از داده‌ها به‌منظور اخذ تصمیمات آگاهانه مرتبط است. از این منظر، متخصصان اثرگذار علم داده قادر هستند پرسش‌های مرتبط را شناسایی، داده‌ها را از انواع منابع داده گردآوری، اطلاعات را به‌گونه‌ای قابل درک سازماندهی، و نتایج خود را به راه‌حل‌های موفقیت‌آمیز تبدیل و نیز یافته‌های خویش را به‌گونه‌ای اثربخش مبادله کنند.

فرآورده‌ها و روش‌های علم داده مراقبت‌های بهداشتی، تجارت و دولت را دگرگون ساخته است، و رشد داده‌ها ادامه خواهد یافت تا بسیاری از سایر بخش‌ها را نیز تحت تأثیر قرار دهد. طبیعی است که مهارت‌های وابسته به داده تقریباً در هر صنعتی ضروری است و این خود باعث می‌شود کار متخصصان علم داده در هر سازمانی، از جمله کتابخانه‌ها، ارزش بالایی داشته باشد. کتابداران نباید آماردان، برنامه‌نویس، یا مدیر پایگاه داده بشوند؛ اما لازم است به یادگیری منطق و زبان‌های برنامه‌نویسی رایانه و پایگاه داده و نیز ابزار بازیابی اطلاعات علاقه‌مند باشند. کتابداران، متصدیان داده^۲ و کارشناسان آرشیو متخصصانی هستند که می‌توانند در پاسخ به تقاضای فزاینده برای نواغ تحلیلی ماهر در بخش‌های متنوع نیروی کار ملی و محلی کمک کنند. از آنجایی که جامعه ما مملو از داده است، کتابداران دانشگاهی در دانشگاه‌ها و مؤسسات آموزش عالی نقشی حیاتی در توسعه آموزش سواد

1. National Institute of Standards and Technology (NIST)

2. data curators

داده، خدمات داده‌های پژوهشی^۱، و خط‌مشی داده‌هایی خواهند داشت که در زیست‌بوم علم داده جدید جای دارد.

علم داده به‌منزله رشته جدید قرن بیست‌ویکم تلقی می‌شود، درست همان‌گونه که علم اطلاعات در قرن بیستم چنین جایگاه داشت. هرچند می‌توان در باب شباهت‌ها و تفاوت‌های بین این دو رشته بحث کرد، اما وانگ^۲ (۲۰۱۸) اعتقاد دارد که این دو رشته به هم وابسته هستند و اینکه «این دو با هم مؤلفه‌های پژوهش‌های زنجیره اطلاعات را تشکیل می‌دهند» (ص. ۱۲۴۶). امروزه، شایسته است ادعا کنیم میان علم داده، علم اطلاعات، علم کتابداری و علم رایانه همگرایی وجود دارد. بین‌رشته‌ای بودن این حوزه‌ها فرصت‌های متعددی را برای همکاری‌هایی فراهم می‌آورد که رسالت اصلی این رشته‌ها را تقویت خواهد کرد. رسالت علم داده «تبدیل داده‌های خام و انباشته به دانش کاربردی» است (استتون^۳، ۲۰۱۲، بند ۶). مارکیونینی^۴ (۲۰۱۶) علم اطلاعات را اصطلاح عامی می‌داند «که علم کتابداری و انفورماتیک را شامل می‌شود و تأکیدش بر تمایزها و شباهت‌های میان رشته‌هایی است که هر یک علم داده را پریار می‌سازد» (ص. ۱). سلسله‌مراتب داده-اطلاعات-دانش-خرد که توسط آکوف^۵ (۱۹۸۹) معرفی شد، نشان می‌دهد که داده به اطلاعات، اطلاعات به دانش، و دانش به خرد تبدیل می‌شود.

داده، اطلاعات و دانش به هم مرتبط هستند. باکلند^۶ (۱۹۹۱) سه معنا برای اطلاعات معرفی می‌کند: اطلاعات به‌منزله فرایند، اطلاعات به‌منزله دانش و اطلاعات به‌منزله شیء. در این سه تعریف، می‌توانید ببینید چگونه داده، اطلاعات و دانش در واقع به هم وابسته هستند. اطلاعات به‌منزله فرایند زمانی مطرح است که فردی آگاه می‌شود؛ دانسته‌های فرد تغییر می‌کند. اطلاعات به‌منزله دانش زمانی رخ می‌دهد که اطلاعاتی با کسی مبادله می‌شود تا هوش یا حقایق خاص یک رویداد را فراهم آورد. اطلاعات به‌منزله شیء به‌طور خاص برای اشیایی نظیر داده استفاده می‌شود. داده به‌خودی‌خود حاوی معنا نیست. برای تبدیل داده به اطلاعات، باید آن را تفسیر کرد و از آن معنایی گرفت.

مفهوم داده

انکساغورس^۷، فیلسوف یونانی، باور داشت که ماهیت فیزیکی «سهمی از هر چیز در هر چیز دیگر بود». داده را نیز بسیار شبیه به تعریف انکساغورس از ماهیت فیزیکی، می‌توان در عمل هر چیز و همه‌چیز یافت. هرچند این نوع

1. research data services

2. Wang

3. Stanton

4. Marchionini

5. Ackoff

6. Buckland

7. Anaxagoras

تعریف مبهم است و رضایت بخش نیست، اما دقیقاً همان چیزی است که به هنگام جستجو برای تعریف داده، به آن می‌رسیم.

«داده» اصطلاحی بسیار عام و مورد استفاده در رشته‌ها و سازمان‌های مختلف است. داده مجموعه‌ای از حقایق (نظیر واژه‌ها، اعداد، ابزار اندازه‌گیری، مشاهدات، و نظایر آن) است که به شکلی مناسب تبدیل شده تا سامانه‌های رایانه‌ای بتوانند آن را پردازش کنند. همچنین به خاطر داشته باشید که داده به طور خاص به منظور شناسایی و تفکیک اطلاعات حیاتی از قطعات اطلاعاتی صرف، مورد استفاده قرار می‌گیرد. داده ممکن است به اشکال گوناگون - در قالب متن یا اعداد روی برگه‌های کاغذ، در قالب حقایق در ذهن انسان، یا در قالب بایت و بیت ذخیره شده در حافظه الکترونیکی - وجود داشته باشد. از اواسط دهه ۱۹۹۰، مردم از اصطلاح «داده» برای دلالت بر اطلاعات رایانه‌ای که ذخیره یا منتقل می‌شود^۱، استفاده کرده‌اند.

«داده‌ها^۲» در شکل جمع خود از واژه مفرد Datum لاتین به مفهوم «چیزهایی است که داده شده است» (باکلند، ۱۹۹۱، ص. ۳۵۳). دولت فدرال ایالات متحده در بخشنامه شماره الف-۱۱۰ سازمان مدیریت و بودجه^۳، داده را این گونه تعریف می‌کند: «مطالب حقیقی مضبوط که در جامعه علمی به طور عام برای اعتباریابی یافته‌های پژوهش به منزله چیزی ضروری پذیرفته شده باشد». داده ممکن است از نظر افراد، سازمان‌ها، کسب و کارها، و رشته‌های مختلف معانی متعددی داشته باشد. برای نمونه، در ویرایش سوم واژه‌نامه آمار آکسفورد^۴، داده توسط آپتون و کوک^۵ (۲۰۱۴) به معنای اطلاعات تعریف شده است که معمولاً شامل اطلاعات عددی و کیفی (دسته‌بندی شده) است. هرچند این باز هم تعریفی ساده است، اما برای رشته آمار قابل درک است.

«داده شامل هر گونه و همه مواد رقومی^۶ است که در جستجوی فعالیت‌های علمی گردآوری و تحلیل می‌شود» (بلوم، گنلی، و وینکر^۷، ۲۰۰۴، ص. ۱). هرچند این تعریف ممکن است تا حدودی در قرن ۲۱ مورد قبول باشد، اما داده چه بسا شامل اقلام غیر رقومی نیز باشد که در دفاتر یادداشت، نامه‌ها و پرسشنامه‌های کاغذی می‌توان مشاهده کرد. لازم به یادآوری است که هر نوع داده‌ای الزاماً شیء رقومی نیست. فورنر^۸ (۲۰۱۷) داده را به منزله «نمونه‌های واقعی و ملموس بازنمایی‌های نمادین گزاره‌های توصیفی تعریف می‌کند که از طریق مشاهدات تجربی در باب ویژگی‌های کمی و کیفی پدیده‌های جهان واقعی اطلاعاتی به ما می‌دهد» (ص. ۶۶).

داده‌های کمی در برابر کیفی

بر اساس تعریف داده فورنر (۲۰۱۷)، باید بدانیم که دو نوع داده پژوهشی کمی و کیفی وجود دارد. داده کیفی

1. download
2. data
3. the OMB Circular A-110
4. Oxford Dictionary of Statistics
5. Upton and Cook
6. digital
7. Bloom, Ganley and Winker
8. Furner

عبارت است از گردآوری اطلاعاتی که هدفش بیش تر توصیف موضوعی است نه سنجش آن. نمونه آن آراء، برداشت‌ها و دیدگاه‌هاست. از سوی دیگر، داده کمی، داده‌های عددی است. برای نمونه، اگر می‌گویید ۵۲ درصد گیاهان مورد آزمایش در یک مطالعه ۱۲ اینچ یا بیش تر رشد کردند، حال آنکه ۳۱ درصد تا ۶ اینچ یا کم تر رشد کردند، داده کمی را مورد توجه قرار می‌دهید.

برعکس، جنسیت آزمودنی‌های مطالعه کلینیکی که در آن لامپ‌های برق به دسته‌های متنوعی نظیر «خیلی پرنور»، «کم‌نور»، و «تا حدودی پرنور» تقسیم می‌شوند، یا دسته‌بندی نوع خواسته مشتریان پیتزا، همه نمونه‌هایی از داده‌های کیفی هستند. داده‌های کیفی معمولاً به شکل روایت‌های نوشتاری یا واژه‌های گفتاری، نه اعداد، تفسیر می‌شوند. بنابراین داده کیفی در وهله نخست به داده‌های قابل مشاهده، مثلاً برحسب ظاهر، بو، مزه، بافت، احساس، جنسیت یا ملیت مربوط است.

چند روش گردآوری داده کیفی به شرح زیر است:

- مشاهده
- مصاحبه گروه کانونی
- مصاحبه
- اطلاعات آرشیوی نظیر روزنامه‌ها

داده‌های کمی را می‌توان هم برای رایانش و هم برای آزمون‌های آماری به کار گرفت. این نوع برای اندازه‌گیری‌هایی نظیر وزن، قد، حجم، اندازه، طول، رطوبت، سن و نظایر آن به کار می‌رود. از این گذشته، داده کمی را می‌توان به منزله داده پیوسته و گسسته دسته‌بندی کرد. روش‌های گردآوری داده‌های کمی به شرح زیر هستند:

- آزمایش
- پیمایش
- مشاهده و مصاحبه

هرچند رویکرد یا روش پژوهش در مورد داده‌های کیفی کلی و ذهنی است، اما داده‌های کمی معمولاً مستلزم رویکردی عینی، بدون سوگیری، و متمرکز است.

داده‌های رقومی در برابر غیررقومی

همان گونه که پیش تر گفته شد، همه داده‌ها رقومی نیست. برای نمونه، بسیاری از پژوهشگران اغلب روزنامه‌های دیواری، نتایج آزمایشگاه، و سایر مطالبی را نگهداری می‌کنند که در سامانه رایانه‌ای نگهداری نمی‌شود. این‌ها داده‌های غیررقومی است. برای نمونه، طرح پژوهشی احتمالاً هم شامل داده رقومی و هم غیررقومی یا داده‌هایی است که ممکن است در اصل غیررقومی باشد اما بعدها به رقومی تبدیل شود.

در سطحی عملی تر، تفاوت داده‌های رقومی و غیررقومی در شیوه سنجش/اندازه‌گیری داده یا اطلاعات است. به‌خاطر داشته باشید که داده‌های قیاسی (آنالوگ) [در مقابل دودویی] ماهیتاً پیوسته است و هر تغییر جزئی را در باب آنچه اندازه‌گیری یا ثبت می‌شود، شناسایی می‌کنند؛ ازسوی دیگر، در داده‌های رقومی معمولاً از نمونه‌گیری (فرایند یا عمل اخذ نمونه از چیزی برای تحلیل جامع آن) استفاده می‌شود تا آنچه را سنجیده‌شده کدگذاری کنند.

روش دیگر بررسی داده‌ها، این است که در داده‌های رقومی برای مقاصد عملی «کدگذاری انجام می‌شود» حال آنکه در داده آنالوگ تلاش می‌شود داده «درست شبیه» همان چیزی باشد که مشاهده می‌شود (برویر^۱، ۲۰۰۵). خطر از بین رفتن داده غیررقومی بیش تر است. هرچند داده رقومی را می‌توان رمزنگاری و در فضای ابری ذخیره کرد، داده غیررقومی معمولاً در نوعی رسانه فیزیکی، بدون توجه به اینکه نوار مغناطیسی فیلم ویدئویی، شیارهای سطحی روی صفحه گرامافون، یا هر رسانه غیررقومی دیگری باشد، ذخیره می‌شود.

شیوه‌ای بسیار مناسب جهت تعیین ویژگی‌های داده غیررقومی این است که این نوع داده بدون اندازه‌گیری موجود است. بنابراین، برای تبدیل ساده داده غیررقومی به شکل رقومی، این نوع داده باید هم ثبت شود و هم به کمک فناوری‌های خاص به شکل جدید درآید. ازاین رو، در داده‌های رقومی به‌طور خاص از سامانه دودویی ساده و مناسب به‌منظور ساختن مجموعه‌های داده‌ای استفاده می‌شود تا فیلم، تصویر، یا صوت را بازنمایی کند.

در انواع فناوری امروزی، به‌جای کار با فناوری آنالوگ، اغلب با فناوری رقومی کار می‌شود. برای نمونه، در گوشی‌های تلفن همراه، تماس‌ها با تبدیل اثربخش اصوات صدای یک فرد به اعداد و سپس تبدیل این اعداد از مکانی به مکان دیگر به شکل امواج رادیویی دریافت و منتقل می‌شود. فناوری رقومی با استفاده از این روش چندین مزیت دارد. به یک دلیل، ساده‌تر و مناسب‌تر آن است که داده‌ها به شکل رقومی ذخیره شود چون به‌طور معمول «فضای» کم‌تری را اشغال می‌کند.

یک نمونه ساده و رایج تفاوت اصلی بین داده غیررقومی و رقومی، آب در حال حرکت است. داده غیررقومی سطح واقعی آب در حال حرکت است که حواس انسان آن را تا حد زیادی به‌منزله تغییرات ظریف در رنگ و نیز حرکات فیزیکی، ترکیب و گاهی حتی در بوی آب درک می‌کند. در سوی دیگر، قالب رقومی خواص رنگ یا جنبش فیزیکی، یا هر دو آن را به مجموعه‌های داده مجزایی تبدیل می‌کند که همه این خواص را روی محیط رابطی سخت‌افزاری شبیه‌سازی می‌کند، یا برای اهداف پژوهشی مهمی ذخیره می‌شود.

مفهوم کلان داده

کلان داده را می‌توان به‌منزله حجم عظیمی از داده‌های نیمه‌ساختاریافته، ساختاریافته و ساختارنیافته‌ای تعریف کرد که می‌توان برای رسیدن به اطلاعات و نیز استفاده در طرح‌های متنوع یادگیری ماشین و سایر نرم‌افزارهای

1. Breuer

کاربردی پیشرفته و تحلیل نظام‌مند علمی در آن کاوش کرد. مؤسسه ملی استانداردها و فناوری کلان‌داده را به‌منزله «داده‌هایی تعریف می‌کند که حجم، سرعت فراهم‌آوری، یا بازنمایی آن‌ها ظرفیت استفاده از روش‌های رابطه‌ای سنتی را برای انجام تحلیل مؤثر محدود می‌سازد، یا داده‌هایی که ممکن است به‌طور اثربخشی با فناوری‌های مهم زوم افقی پردازش^۱ شود».

کلان‌داده اغلب با ۳ ویژگی، ۴ ویژگی، و ۵ ویژگی تداعی می‌شود. داگ لنی^۲ در سال ۲۰۰۱ تلاش کرد کلان‌داده را این‌گونه تعریف کند که کلان‌داده نمایانگر ۳ ویژگی حجم، سرعت و تنوع^۳ است. این ۳ ویژگی، خصوصیات اساسی و اصلی کلان‌داده را شکل می‌دهد. منظور از حجم مقدار هنگفتی داده بود، سرعت عبارت بود از فراوانی داده‌هایی که باید در طی زمان ایجاد شود و تنوع به انواع مختلف داده‌های موجود اطلاق می‌شد. این سه ویژگی به‌طور بسیار مناسبی در تعریف کلان‌داده ارائه شده توسط آنالی ساکسنیان^۴ (۲۰۱۴)، رئیس دانشکده اطلاع‌رسانی دانشگاه برکلی در کالیفرنیا^۵، خلاصه شده و در آن او اعتقاد دارد، «کلان‌داده عبارت از داده‌هایی است که نمی‌توان با استفاده از پایگاه‌های داده استاندارد پردازش کرد چون برای ابزار پردازش داده سنتی بسیار بزرگ، بسیار متغیر، یا بسیار پیچیده است».

تعریف کلان‌داده به‌طور پیوسته تغییر می‌کند، و این امر از تعداد خصوصیات که به ویژگی‌های کلان‌داده اضافه می‌شود، مشخص است. پس از سه ویژگی یادشده، ویژگی چهارمی به آن با عنوان صحت و درستی^۶ اضافه شد. صحت به‌دقت داده‌ها اطلاق می‌شود. به صحت کلان‌داده اغلب در پژوهش‌ها استناد می‌شود و مورد استفاده قرار می‌گیرد؛ اما برخی ادعا می‌کنند که تعداد این ویژگی‌ها حتی ممکن است به ۱۰ مورد برسد. ولی پیش از افزودن هر ویژگی دیگری، بهتر است پرسیم، «آیا ویژگی‌های جدید، چیز جدیدی را منعکس می‌کنند یا با سایر ویژگی‌ها همپوشانی دارند؟» بنابراین ویژگی‌های زیر رایج‌ترین ویژگی‌هاست و بهتر است وقتی استفاده شود که کلان‌داده را توصیف می‌کنیم:

- **حجم:** حجم داده به مجموعه وسیعی از داده‌هایی اطلاق می‌شود که براساس پیشینه‌های به‌جای مانده از علم و آموزش و از کسب‌وکار و تعامل انسانی خلق شده است. حجم داده نقش چشمگیری در ذخیره و پردازش ایفا می‌کند.
- **سرعت:** سرعت داده به‌سرعت خلق داده‌ها اطلاق می‌شود.

1. horizontal zoom technologies

2. Doug Laney

3. volume, velocity, and variety

4. AnnaLee Saxenian

5. Dean of University of California Berkeley, School of Information

6. veracity