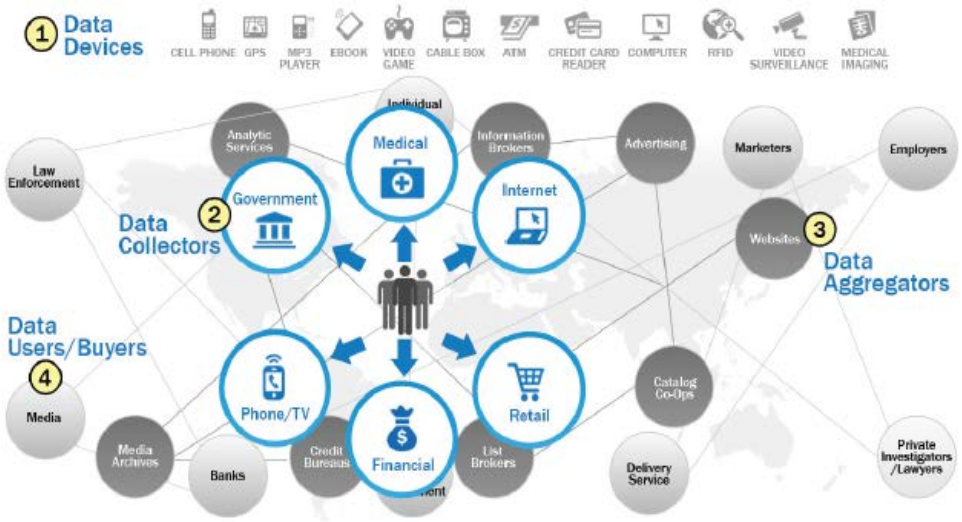
دکتر فرزاد اسکندری- عضو هیات مدیره انجمن آمار ایران و استاد دانشگاه

چرا حکم‌رانی داده‌ها:

در جهان امروز رفتار اجتماعی سیاسی و اقتصادی انسانها وابسته به یک اکوسیستم نوینی شده است که اساس و بنیان این اکوسیستم توسط داده ها تولید گردیده است. در این اکوسیستم تقریبا همه انسانها به دو صورت آگاهانه و ناآگاهانه در مرکزاین اکوسیستم قرار دارند. چنان داده ها در لایه لایه زندگی انسانهای درون این اکوسیستم بوسیله ابزارهای مختلف تولید داده رسوخ پیدا نموده ایت که فضای کسب و کار مردم و آداب زندگی مردم بر اساس آن در حال شکل‌گیری است و بدون آن تقریبا نمی‌توان تحولی در زندگی ایجاد نمود. در واقع داده‌ها دارند بر انسانها حمکرانی می‌کنند. در این اکوسیستم نوین همه داده‌ها به صورت منظم در اختیار همه افراد قرار ندارند و تولید نمی‌شوند. بر اساس بررسیهای انجام شده چیزی در حدود ۱۰ درصد داده‌ها فقط به صورت ساختاریافته ایجاد می‌گردند و در حدود۹۰ درصد داده‌ها به صورت شبه ساختاری و یا ناساختار تولید شده ونقش‌آفرینی می‌کنند. دراین میان انتظار ایجاد تجمیع، مدیریت، مدلسازی، مصورسازی، پردازش، تحلیل، تفسیر و تصمیم‌گیری چیزی است که قرار است از این اکوسیستم استخراج شود. می‌توان گفت یک رویکرد جدید در کشف علم در حال رخ دادن است. این کشف علمی در این اکوسیستم رفتاری همان چیزی است که امروزه ان را علم داده ها می نامند.

علم داده‌ها، مطالعه ساختار و چگونگی تولید داده‌هاست؛ که چگونه می‌تواند به یک منبع ارزشمند در ایجاد استراتژی‌های کسب‌وکار و فناوری اطلاعات تبدیل شود. استخراج مقادیر زیادی داده ساختار یافته و ناساختار برای شناسایی الگوها می‌تواند به یک دستگاه و یا سازمان کمک کند تا هزینه‌ها را مهارکرده، بازده را افزایش داده، فرصت‌های بازار جدید را شناسایی کرده و مزیت رقابتی سازمان را افزایش دهد. رشته علم داده‌ها از ریاضیات، آمار و علوم رایانه استفاده می‌کند و فن‌هایی مانند یادگیری ماشین، تحلیل خوشه‌ای، داده‌کاوی و مصورسازی را در بر می‌گیرد.

علم داده‌ها یکی از علوم بین‌رشته‌ای است که از روش‌های علمی، فرآیندها، الگوریتم‌ها و سیستم‌ها برای استخراج دانش و بینش از داده‌ها در اشکال مختلف، ساختار یافته و ناساختاری شبیه به داده‌کاوی استفاده می-کند. به منظور "درک و تجزیه و تحلیل پدیده‌های واقعی" با داده‌ها، "علم داده‌ها" یک مفهوم برای متحد کردن آمار، تجزیه و تحلیل داده‌ها، یادگیری ماشین و روش های مرتبط با آن است. این فن‌ها و نظریه‌ها را از مفاهیم مختلف در زمینه ریاضیات، آمار، علم اطلاعات و علوم رایانه به کار می‌برد.



از نظر تاریخی جدول زمانی مختلفی وجود دارد که می‌تواند برای ردیابی رشد آهسته علم داده‌ها و تأثیر فعلی آن در ایجاد اکوسیستم جاری مورد توجه قرار بگیرند. برخی از موارد مهمتر عبارتند از:

به عنوان اولین فرد جان توکی (۱۹۶۲) در مورد تغییر در دنیای آمار نوشت:

«... همانطور که مشاهده کردم آمار ریاضی در حال تکامل است و مصور سازی ، من دلیلی برای شگفتی و تعجب داشتم ... من فکرمی‌کردم که علاقه من به مطالعه در داده‌ها است اما تصویر پر دازی در آن ایجاد گردید ...»

توکی به ادغام آمار و رایانه‌ها اشاره دارد، در زمانی که نتایج آماری در ساعات شبانه‌روز، به جای روزها و یا چند هفته که با دست انجام می‌شود، ارائه می‌شود.

در سال ۱۹۷۴، پیتر نائور محاسبات مختصر روش‌های کامپیوتری را با استفاده از عبارت "علم اطلاعات"، به طور مکرر انجام داد. نائور، تعریف پیچیده خود را از مفهوم جدید ارائه داد:

علم رسیدگی به داده‌ها، درزمانی که آن‌ها تاسیس شده‌اند، در حالی که رابطه داده‌ها به آنچه که نماینده آن‌ها هستند به حوزه‌های دیگر و علوم محول شده‌است.

در سال ۱۹۷۷ انجمن بین‌المللی محاسبات آماری شکل گرفت. اولین عبارت بیانیه رسالت آن‌ها این است:

" رسالت سازمان برای پیوند دادن روش‌شناسی آماری سنتی، فن‌آوری کامپیوتر مدرن، و دانش متخصصان حوزه به منظور تبدیل داده‌ها به اطلاعات و دانش است.

در ده سال گذشته، علم داده‌ها به آرامی رشد کرده‌است تا شرکت‌ها و سازمان‌ها را از هم جدا کند. امروزه استفاده از آن توسط دولت‌ها، نسل‌شناس‌ها، مهندسین و حتی ستاره شناسان مورد استفاده قرار می‌گیرد. البته در طول تکامل علم داده ها، استفاده از واژه جدیدی به نام مه داده‌ها ایجاد گردید که به سادگی قبل دارای "مقیاس بندی" مشابه داده‌ها نبود، بلکه شامل تغییر به سیستم‌های جدید برای پردازش داده و روش‌های مورد مطالعه و تحلیل گردید.

علم داده‌ها به بخش مهمی از تحقیقات علمی و دانشگاهی تبدیل شده‌است. از لحاظ فنی این شامل ترجمه ماشینی، روباتیک، بازشناسی گفتار، اقتصاد دیجیتال و موتورهای جستجو است. علم داده از نظر حوزه‌های تحقیقاتی گسترش‌یافته تا علوم زیستی، مراقبت‌های بهداشتی، انفورماتیک پزشکی، علوم انسانی و علوم اجتماعی را در بر گیرد. علم اطلاعات اکنون بر اقتصاد، دولت‌ها، و امور مالی و مالی تاثیر می‌گذارد. به زبانی ساده علم داده ها در حال حکم رانی بر کلیه سامانه‌ها است.

یکی از نتایج انقلاب علم داده‌ها، یک تغییر تدریجی برای نوشتن بیشتر و بیشتر برنامه‌نویسی دقیقتر و محافظه کارانه را ایجاد نموده است. بسیاری از پژوهشگران علم داده‌ها در حال حاضر فکر می‌کنند که بازبینی‌های کامل کل سامانه‌ها خیلی خطرناک است و بهتر است ایده‌ها را به بخش‌های کوچک‌تر تقسیم کنند. هر قسمت مورد آزمایش قرار می‌گیرد و سپس با احتیاط به درون جریان داده‌ها وارد می‌شود.

اگر چه این فلسفه بازی ایمن می‌تواند زمان و پول شرکت‌ها را نجات دهد و از لغزش‌ها و خطاهای بزرگ اجتناب کند، آن‌ها بر قیدهای بسیار محدودی تمرکز می‌کنند و از پی‌گیری دستاوردهای واقعی اجتناب می‌کنند.

اسکات هافمن از گوگل می‌گوید:

چیزی که زمان زیادی را صرف صحبت در مورد آن می‌کنیم این است که چگونه می‌توانیم در برابر تغییرات زمانی که تغییرات بزرگتری مورد نیاز است محافظت کنیم. سخت است، چون این ابزار آزمایش واقعا می‌توانند به تیم مهندسی انگیزه دهند، اما آن‌ها همچنین می‌توانند مشوق‌های زیادی به آن‌ها بدهند تا فقط تغییرات کوچکی را امتحان کنند. ما خواهان این اصلاحات کوچک هستیم، اما همچنین می‌خواهیم که پرش بیرون از جعبه را انجام دهیم.

اما از نگاه دیگر بین آمار و علم داده تفاوت هایی وجود دارد

علم داده‌ها یکی از جدیدترین گرایش‌های در حال ظهور در رایانش است و یک حیطه وسیع چند رشته‌ای است. علم داده‌ها کاربرد مفاهیمی همچون علوم رایانه، مهندسی نرم‌افزار، ریاضی و آمار، برنامه‌ریزی، اقتصاد، و مدیریت کسب‌وکار را ترکیب می‌کند. علم داده‌ها مبتنی بر جمع‌آوری، آماده‌سازی، تحلیل، مدیریت، تجسم سازی و ذخیره حجم زیادی از اطلاعات است. علم داده‌ها در شرایط ساده می‌توان به عنوان داشتن ارتباطات قوی با پایگاه‌های اطلاعاتی، از جمله مه داده‌ها و علوم کامپیوتر، درک کرد. یک پژوهشگر داده‌ها یک فرد با دانش حوزه کافی مرتبط با این مفهوم است.

تفکر مه داده ها با علم داده‌ها به شدت ادغام شده‌اند و در حقیقت، با ادغام علم داده و مه‌داده‌ها در کاربردهای مختلف در اکوسیستم مورد بررسی تکامل‌یافته است. اطلاعات مفید به آسانی در مه‌داده‌ها که از بلاگ‌ها، فایل‌های

صوتی / تصویری، تصاویر، پیغام‌های متنی، شبکه‌های اجتماعی و غیره ساخته شده‌است از نظر پنهان می‌ماند. تمام این داده‌ها فقط نویز هستند مگر اینکه تحلیل شود و اطلاعات مفید از آن‌ها استخراج شود. به علاوه، امروزه کسب و کارهای اینترنتی را به عنوان کانال اطلاعات اولیه خود به دلیل نقش رو به رشد وبهای اجتماعی و پتانسیل کسب‌وکار خود در نظر می‌گیرند. تمام این داده‌ها برای یک پژوهشگر داده‌ها بسیار جالب هستند، زیرا با استفاده از این داده‌ها، بسیاری از مشکلات را می‌توان برای سازمان‌ها و نیز جوامع حل کرد.

علم داده‌ها یک مهارت تخصصی است و می‌تواند به صورتهای مختلفی درک شود:

- به‌کارگیری تکنیک‌های پیشرفته در ریاضیات و آمار برای مدل کردن داده‌ها برای تحلیل داده‌ها
- مهارت برنامه نویسی و توسعه موثر، مهارت‌های پیشرفت الگوریتم
- مهارت‌های استدلال تحلیلی
- مهارت‌های ارتباطی و کسب‌وکار

بنابراین، واضح است که علم داده یک حوزه میان‌رشته‌ای است و به مجموعه‌های مهارتی مختلف برای کسب مهارت در این حوزه نیاز دارد. موارد استفاده در علم داده شبیه تجزیه و تحلیل‌های داده‌ها هستند - آن‌ها با یک دستور مشکل واضح شروع می‌کنند و تصمیم می‌گیرند که در نهایت با معیارهای تعریف‌شده به خوبی به پایان برسند. بنابراین، پژوهشگران داده در نظر گرفته می‌شوند تا با مدل‌های کسب‌وکار و الگوها آشنا باشند. اما آمار یک موضوع گسترده دیگر است که با مطالعه داده‌ها سر و کار دارد و به طور گسترده در زمینه‌های متعدد به کار می‌رود.

اگرچه آمار روش‌های جمع‌آوری و تجزیه و تحلیل داده‌ها را فراهم می‌کند، اما به کسب اطلاعات از داده‌های عددی و رسته ای کمک می‌کند. داده‌های رسته‌ای به داده‌های منحصر به فرد اشاره دارد. در واقع آمار در مطالعات مربوط به داده‌ها بسیار قابل‌توجه است، چرا که به آن‌ها کمک می‌کند تا بتوانیم:

- تصمیم‌گیری درباره نوع داده مورد نیاز برای رسیدگی به یک مشکل مشخص
- سازمان دهی و خلاصه‌سازی اطلاعات
- بدست آوردن تحلیل برای نتیجه‌گیری از داده‌ها
- ارزیابی اثربخشی نتایج و ارزیابی عدم قطعیت
- طراحی برای برنامه‌ریزی و انجام تحقیق
- شرحی که دلالت بر کاوش و خلاصه‌سازی اطلاعات دارد
- ساخت پیش‌بینی‌ها و استنتاج از پدیده ارائه‌شده توسط داده‌ها

انجام پذیرد.

پس درکلام آخر ره آورد علم داده‌ها ایجاد یک نظم نوین علمی در این دنیای بی نظم است که فقط می تواند حاصل از به کارگیری برون دادهای علم داده‌ها در تمام مراحل مختلف زندگی بشری باشد. با این رویکرد می‌توان به دنبال یک شبکه هوشمند جهانی بود که انسانها در آن با هم در ارتباط هستند و در هر لحظه و هر مکان انتظار ایجاد یک تحلیل و تصمیم دقیق و درست را برای هر فرد تصمیم گیر را در هر سازمان و دستگاه فراهم نمود که دارای کمترین مخاطره باشد.