فصل ۱
تعاریف و مفاهیم پایه

۱.۱ مقدمه

در دنیای امروز که رقابت حرف اول را می‌زند، بی‌شک اطلاعات بسیار مهم خواهد بود. داده‌ها نشان‌دهنده واقعیات، معلومات، رویدادها و ... برای پردازش توسط انسان و یا ماشین هستند. با ورود رایانه در این حیطه از سال ۱۹۵۰، پس از ۲۰ سال حجم اطلاعات ذخیره شده تا دو برابر افزایش یافت و با پیشرفت فناوری اطلاعات در هر دو سال باز هم حجم داده‌ها در پایگاه داده دو برابر شد. در حال حاضر نیز با ایجاد شبکه جهانی وب، سیستم‌های یکپارچه بانکی و اطلاعاتی، در هر لحظه حجم داده‌ها در حال رشد و افزایش است و این امر باعث تولید انبارهای بسیار بزرگی از داده‌ها شده است. با وجود رقابت شدید میان کشورهای مختلف در زمینه‌های علمی، اجتماعی، اقتصادی، نظامی و سیاسی، اهمیت کشف و استخراج سریع و دقیق دانش موجود در داده‌ها، کاملا روشن می‌شود. از این رو دو نیاز، حیاتی احساس می‌شود:

اول: طراحی سیستم‌های کشف اطلاعات مورد علاقه با دخالت حداقلی کاربر انسانی

دوم: به‌کارگیری روش‌های تحلیلی که با حجم انبوه داده‌ها سازگارند.

در پاسخ به این دو نیاز، شاخه جدیدی از کاوش داده‌ها گسترش یافته است که به "داده‌کاوی" شهرت دارد. این دانش در کشف اطلاعات با ارزش از مجموعه‌های عظیم داده‌ها، مورد استفاده قرار می‌گیرد.

در آغاز دهه ۹۰، با انجام تحقیقات در رشته‌های آمار[۱]، یادگیری ماشین [۲]و علوم رایانه[۳]، داده‌کاوی پا به عرصه ظهور گذاشت تا با نگرشی نو به مسأله استخراج دانش از داده‌ها بپردازد.

نخستین بار فیاض[۹][۴] در اولین کنفرانس داده‌کاوی و کشف دانش[۵] در سال ۱۹۹۵ اصطلاح داده‌کاوی را مطرح نمود و در همین زمان بود که داده‌کاوی به طور جدی وارد مباحث آماری شد.

با وجود جدید بودن داده‌کاوی باید اذعان کرد که این رشته امروزه نقش‌های گسترده و متنوعی را در رشته‌های بازرگانی، پزشکی، مهندسی، علوم رایانه، صنعت، کنترل کیفیت، ارتباطات، کشاورزی و بسیاری از رشته‌های دیگر ایفا می‌کند. به نظر می‌رسد اولین انگیزه برای کاوش داده‌ها، رشد روز افزون آن است. در حقیقت این رشد سرسام‌آور تا اندازه‌ای است که تنها در صورت در اختیار داشتن ابزار مکانیزه، می‌توان به استخراج دانش مفید از آن امیدوار بود. چرا که در صورت تحلیل دستی اطلاعات، تنها قادر به کار با حجم اندکی از داده‌ها خواهیم بود در حالی که حجم انبوهی از آن‌ها همزمان در حال تولید است و این بسیار ناامید کننده است.

داده‌کاوی تنها یک مفهوم کلی که هدفش ایجاد درک از داده‌ها می‌باشد نیست. اصلی‌ترین ویژگی داده‌کاوی که آن را از بقیه‌ی رویکردها متمایز می‌سازد، داده‌گرا بودن آن است در حالی‌که بقیه‌ی روش‌ها مدل‌گرا هستند. محققین درعلم آمار بیشتر به دنبال یافتن کوچک‌ترین اندازه‌ای از داده‌ها هستند که منجر به تخمین‌هایی با اطمینان مطلوب می‌شود. در داده‌کاوی با وضعیت کاملا متفاوتی مواجه هستیم. اندازه داده‌ها بزرگ بوده و قصد داریم تا مدلی از داده‌ها را بسازیم که پیچیده نبوده و در عین حال داده‌ها را به خوبی توصیف نماید. پیدا کردن یک مدل خوب برای داده‌ها که فهم آن نیز آسان باشد هدف اصلی داده‌کاوی است. می‌بایست دو موضوع را مورد توجه قرار دهید:

اول هیچ مدلی کامل نیست.

دوم ما تقریباً همیشه به دنبال حالت تعادلی میان کامل بودن مدل و پیچیدگی مدل خواهیم بود و این امری عادی است.

---

[۱] Statistics
[۲] Machine Learning
[۳] Computer Siences
[۴] Fayyad
[۵] Knowledge Discovery and Data mining

در واقع داده‌کاوی عبارت است از ایجاد درک از حجم زیادی از داده‌ها، در یک حیطه معین علمی که در اغلب موارد این داده‌ها به صورت هدایت نشده جمع‌آوری شده‌اند. بیشتر افرادی که علم داده‌کاوی را مورد استفاده قرار می‌دهند، اشخاص متخصص دریک زمینه خاص علمی بوده و نه تنها به داده‌ها دسترسی دارند، بلکه خود نیز به جمع‌آوری داده اقدام می‌ورزند. اما فرض می‌کنیم که صاحبان داده درک مختصری از داده و فرآیندی که منجر به تولید آن می‌شود دارند.

۲.۱ مفاهیم اولیه

در اینجا به معرفی چند مفهوم اساسی در مبحث داده‌کاوی می‌پردازیم:

داده[۶] : مجموعه‌ای از اعداد، حروف، علائم و نشانه‌های بی‌مفهوم هستند که بدون انجام پردازش فاقد ارزش هستند. داده‌ها حقایقی هستند که از طریق مشاهده و تحقیق بدست می‌آیند.

اطلاعات[۷] : به مجموعه‌ای از داده‌ها گفته می‌شود که طی عملیات‌های منطقی پردازش می‌گردند و تبدیل به اطلاعاتی می‌گردند که مطلب قابل فهمی را به کاربر منتقل می‌نمایند. بنابراین اطلاعات از داده‌های پردازش شده استخراج می‌شود.

دانش[۸] : بر درک و تجربه دلالت دارد که می‌تواند بین استفاده درست و نادرست از اطلاعات، تفاوت قائل شود. اطلاعات جمع‌آوری می‌شوند، در حالی‌که دانش توسعه و گسترش می‌یابد و افزایش پیدا می‌کند.

خرد[۹] : داشتن دانش و نیز فهم آن به همراه قابلیت به کار بستن آن است. فرآیندی که به وسیله آن، مسائل حل می‌شوند.

مثال زیر مفهوم چهار اصطلاح مطرح شده را بیشتر روشن می‌کند:

مثال ۱.۱: عدد ۱۸ به تنهایی و به خودی خود یک داده محسوب می‌شود زیرا نمی‌دانیم این عدد دقیقا بیان‌گر چه مفهومی است. اگر در کنار این عدد، سن را مطرح کنیم، مجموعه سن ۱۸ سال می‌تواند یک عنصر اطلاعاتی باشد. اما برای روشن شدن مفهوم دانش می‌توانیم از این قانون استفاده کنیم که: "برای اخذ گواهینامه رانندگی، داشتن حداقل سن ۱۸ سال اجباری است." این قانون دانش را به ما معرفی

[۶] Data
[۷] Information
[۸] Knowledge
[۹] Wisdon

می‌کند. همان‌گونه که واضح است مفهوم دانش بسیار فراتر از داده است. در این‌جا ارتباط مشخصی میان اخذ گواهینامه و سن ۱۸ سال برقرار است.

درکل قوانینی که به صورت "اگر-آن‌گاه" مطرح می‌شوند یکی از اشکال نمایش دانش هستند. اما برای روشن شدن مفهوم خرد یک پزشک متخصص و یک دانشجوی پزشکی تازه فارغ‌التحصیل شده را در نظر می‌گیریم. هر دوی آن‌ها مجموعه بسیار زیادی از قوانین اگر-آن‌گاه مربوط به دانش پزشکی را در ذهن خود نگهداری می‌کنند. اما ممکن است در برخورد با یک بیمار مشخص، تشخیص‌های بسیار متفاوتی بدهند. این تفاوت در تشخیص در واقع ریشه در فرادانش یا همان خردی دارد که پزشک متخصص در مدت سال‌های طولانی کار، به آن دست یافته است .

مثال ۲.۱: سوال زیر را در نظر بگیرید.

شما به کیفیت غذای سلف دانشگاه، از یک تا پنج، چه نمره‌ای می‌دهید؟

الف) یک ب) دو ج) سه د) چهار ه) پنج

پاسخ فرد به تنهایی و بدون درنظر گرفتن سایر گزینه‌ها، یک داده می‌باشد. با قرار دادن پاسخ شخص در کنار سایر گزینه‌ها و تعیین نوع داده‌ها (کیفی، کمی) ما به اطلاعات دست یافته‌ایم. با تجزیه و تحلیل اطلاعات می‌توانیم سطح کیفیت غذای دانشگاه را مشخص کنیم. این درک و تجربه همان دانش است. وقتی که با تعیین سطح کیفی، در صدد ارتقاء یا تثبیت این کیفیت برآییم و از دانش حاصله در جای درست استفاده نماییم، به خرد رسیده‌ایم.

۳.۱ تعاریف داده‌کاوی

شاید نگاهی به ترجمه لغوی داده‌کاوی، بتواند ما را در درک بهتر مفهوم این واژه یاری کند. واژه لاتین "$Mine$" به معنای استخراج از منابع پنهان و ارزشمند زمین می‌باشد. ترکیب این واژه با "$Data$" می‌تواند بر جستجوی عمیقی در میان حجم انبوه داده‌ها برای یافتن دانشی مفید، دلالت کند که این دانش قبلا پنهان و نهفته بوده است.

داده‌کاوی دارای تعاریف مختلفی است، این تعاریف به مقدار زیادی به پیش زمینه‌ها و نقطه‌نظرهای افراد بستگی دارد. هر نویسنده و محقق با توجه به دیدگاه و نوع نگرش خود تعریفی برای داده‌کاوی ارائه کرده

است که برخی از آن‌ها به صورت زیر است:

- داده‌کاوی فرآیند شناخت الگوهای معتبر، جدید، مفید و قابل فهم از داده‌ها می‌باشد[۷].
- داده‌کاوی، اکتشاف و تحلیل حجم زیادی از داده‌ها برای کشف الگوها و قواعد معنادار است. فرآیند داده‌کاوی گاهی کشف دانش نیز نامیده می‌شود[۲۰].
- داده‌کاوی، فرآیند به خدمت گرفتن یک روش رایانه‌ای است که با استفاده از تکنیک‌های مختلف مستقیماً از داده‌ها دانش استخراج می‌کند[۱۰] .
- داده‌کاوی، جستجویی است برای اطلاعات جدید و نوین از میان حجم زیادی از داده‌ها و فرآیندی مشارکتی میان انسان و کامپیوتر است [۲].

اما کامل‌ترین تعریف برای داده‌کاوی به‌صورت زیر بیان شده است:

تعریف۱.۱: داده‌کاوی عبارت است از فرآیند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه عظیمی از داده‌ها و استفاده از آن در تصمیم‌گیری در یک حیطه علمی خاص [۲۵]. بیشتر افرادی که علم داده‌کاوی را مورد استفاده قرار می‌دهند، اشخاص متخصص در یک زمینه خاص علمی بوده و نه تنها به داده‌ها دسترسی دارند بلکه خود نیز به جمع‌آوری داده اقدام می‌ورزند. اما فرض می‌کنیم که صاحبان داده درک مختصری از داده و فرآیندی که منجر به تولید آن می‌شود دارند. اولین عبارت مهم ایجاد درک است که با توجه به تجربه کاربر می‌تواند معانی مختلفی داشته باشد. به منظور درک بهتر، تصور می‌کنیم که این دانش جدید بایستی تعدادی از ویژگی‌های اساسی از جمله: قابل فهم بودن، معتبر بودن، جدید و مفید بودن را دارا باشند. با توجه به این توضیحات بهترین نتیجه‌ای که می‌توان انتظار آن را داشت عبارت خواهد بود از دانش یا مدلی که بتوان آن را به کمک عبارات ساده (به عنوان مثال از طریق قوانین) توصیف نمود.

مثال ۳.۱: اگر در عروق کرونری قلب مورد غیرعادی (گرفتگی) مشاهده شود، آنگاه با بیماری عروق کرونری مواجه هستیم. در این مثال داده‌های ورودی می‌تواند تصاویری از قلب و عروق متعلق به آن باشد. اگر تصاویر، توسط متخصص قلب بررسی شود و او این تصاویر را عادی یا غیر عادی (گرفتگی عروق

کرونری) ارزیابی نماید، آن‌گاه چنین داده‌هایی را داده‌های یادگیری یا آموزشی می‌نامند. یعنی بوسیله‌ی معرفی این داده‌ها به سیستم داده‌کاوی، روند بررسی و استخراج اطلاعات مفید از تصاویر ورودی را به سیستم آموزش می‌دهیم و آن را برای استخراج اطلاعات از تصاویر بعدی آماده می‌کنیم.

برخی از تکنیک‌های داده‌کاوی به منظور ایجاد قوانین و با توجه به داده‌های موجود، مدل‌هایی را می‌سازند. متخصصین قلب با بررسی این قوانین، آن‌ها را می‌پذیرند و یا رد می‌نمایند. اعتبار مدل ساخته شده، ویژگی دوم است. اگر متخصصین قلب با تمامی قوانین تولید شده از قبل آشنا باشند، آن‌گاه این قوانین، کم اهمیت به شمار رفته و هیچ جذابیتی نخواهند داشت. اما باید این موضوع را نیز درنظر داشت که تولید قوانینی که از قبل در مورد آن‌ها آگاهی داریم، می‌تواند به نوعی کار اعتبار بخشی را در ارتباط با مدل‌های تولید شده و نحوه‌ی به کارگیری تکنیک‌ها و مفاهیم داده‌کاوی انجام دهد.

سومین ویژگی مرتبط با ایجاد درک، جدید بودن دانش استخراج شده است.

چهارمین ویژگی نیز مفید بودن دانش به دست آمده است. دانش استخراج شده بایستی مفید باشد. این مفید بودن باید از نوع مدلی که مورد استفاده قرار گرفته است، مستقل باشد. یعنی با تغییرمدل میزان فایده‌ی دانش استخراج شده تغییر نکند.

حجم زیاد داده‌ها عبارت کلیدی دیگری است که در تعریف به آن اشاره شد. هدف از داده‌کاوی تحلیل مجموعه‌های کوچکی از داده‌ها نیست، زیرا می‌توان آن‌ها را به کمک بسیاری از تکنیک‌های استاندارد تحلیل کرد یا آن‌که حتی تحلیل آن‌ها را به صورت دستی انجام داد.

موارد زیر مثال‌هایی از حجم زیاد داده‌هاست که در داده‌کاوی مورد استفاده قرار می‌گیرند.

- شرکت مخابراتی ای تی اند تی[۱۰] روزانه بیش از سیصد میلیون تماس تلفنی را دریافت و به حدود ۱۰۰ میلیون نفر از مشتریان خود خدمات‌دهی می‌کند. این شرکت اطلاعات خود را در یک پایگاه داده چندین ترابایتی ذخیره می‌سازد.

- وال مارت[۱۱] روزانه درحدود ۲۱ میلیون تراکنش را در تمامی شعب خود مدیریت کرده و اطلاعات به دست آمده را در یک پایگاه داده چند ده ترابایتی ذخیره می‌کند.

[۱۰] AT&T
[۱۱] Wal-Mart

- ناسا [۱۲] در هر ساعت چند گیگابایت اطلاعات را از سیستم رصد سیاره زمین به دست آورده و ذخیره می‌سازد.
- شرکت های نفتی مانند موبایل اویل[۱۳] صدها ترابایت از اطلاعات مختلف و مرتبط با استخراج نفت را ذخیره می‌کند.

با توجه به مقوله سرعت و حافظه روشن است که هیچ کدام از پایگاه‌های داده‌ای که در بالا به آن اشاره شد را نمی‌توان به کمک نیروی انسانی یا حتی بهترین الگوریتم‌ها تحلیل کرد. پس به منظور کاهش مقدار و بعد این حجم عظیم از اطلاعات به تکنیک‌های داده‌کاوی نیاز داریم.

سومین عبارت مهمی که در تعریف به آن اشاره شد، "هدایت نشده" است. جمع‌آوری داده‌ها به صورت هدایت نشده بسیار سریع‌تر و ارزان‌تر از جمع‌آوری داده‌ها به صورت هدایت شده است. دلیل آن است که برای جمع‌آوری داده‌ها به صورت هدایت شده به ورودی‌های معلومی که با خروجی‌های معلوم متناظر هستند، نیاز داریم. این موارد توسط متخصصین تعیین می‌شوند. در مثالی که قبلا بیان کردیم، "ورودی" یعنی تصاویر با "خروجی" که همان تشخیص بیماری کرونری قلب می‌باشد، متناظر است. بنابراین اگر جمع‌آوری داده‌ها را صرفا به صورت هدایت نشده انجام دهیم چگونه می‌توانیم از آن‌ها استفاده کنیم؟ به منظور حل این مشکل که یکی از دشوارترین مسائل در داده‌کاوی می‌باشد، به الگوریتم‌هایی نیاز داریم که توانایی یافتن گروه‌بندی‌ها، خوشه ها، رابطه‌ها و وابستگی‌های طبیعی و موجود در داده‌ها را داشته باشند. به عنوان مثال، اگر بتوان خوشه‌هایی را پیدا کرد، آنگاه متخصصین فن، می‌توانند این خوشه‌ها را برچسب‌گذاری کنند [۲۳]. وضعیتی نیز وجود دارد که در آن با داده‌های نیمه هدایت شده مواجه هستیم. منظور، داده‌های است که هم شامل تعدادی از جفت‌های داده‌ای آموزشی معلوم بوده و هم شامل هزاران نقطه داده‌ای هدایت نشده هستند. در مثالی که در رابطه با بیماری عروق کرونری مطرح شد، این موقعیت، متناظر با زمانی است که هزاران تصویر داشته باشیم که هیچ تشخیصی در رابطه با آن‌ها صورت نگرفته باشد و تنها چند تصویر داشته باشیم که مورد تحلیل قرار گرفته باشند. حال باید به این پرسش پاسخ دهیم که آیا این تعداد (اندک) از نقاط داده‌ای قادر هستند تا در فرآیند درک تمامی

[۱۲] NASA
[۱۳] Mobile Oil

مجموعه داده به ما یاری رسانند؟ خوشبختانه، تکنیک‌های یادگیری نیمه هدایت شده، این تعداد از نقاط داده‌ای آموزشی را به خوبی مورد استفاده قرار می‌دهند.

ساده‌ترین وضعیتی که در داده‌کاوی با آن روبه‌رو هستیم، زمانی است که تمامی نقاط داده‌ای به طور کامل هدایت شده باشند. تعداد زیادی از تکنیک‌های داده‌کاوی موجود، به منظور کار با چنین داده‌هایی کاملا مناسب هستند؛ اما ممکن است که از نظر مقیاس‌پذیری با یکدیگر متفاوت باشند. یک الگوریتم داده‌کاوی که هم با داده‌های کوچک و هم با داده‌های بزرگ به خوبی کار کند، مقیاس‌پذیر نامیده می‌شود. متأسفانه تعداد کمی از الگوریتم‌ها دارای چنین ویژگی‌هایی هستند.

آخرین واژه مهمی که در تعریف به آن برخورد می‌کنیم، حیطه معین علمی است.کسب موفقیت در یک پروژه‌ی داده‌کاوی، به در دسترس بودن دانش موجود در حیطه علمی مورد نظر، بسیار وابسته است و بنابراین برای متخصصین داده‌کاوی ضروری است تا با افراد متخصص در حیطه علمی مورد نظر همکاری بسیار نزدیکی داشته باشند. استخراج دانش جدید فرآیندی است که نیاز به تعامل متخصصین در حیطه علمی مورد نظر و تکرار بسیار زیاد دارد. نمی‌توانیم یک سیستم داده‌کاوی که با موفقیت در یک حیطه علمی معین پیاده سازی شده است را به سادگی و در ارتباط با یک حیطه علمی دیگر به کار گرفته و انتظار کسب نتایج خوبی را هم داشته باشیم.

در حقیقت هدف اصلی داده‌کاوی کشف دانش است که این دانش همان نظمی است که در داده‌ها وجود دارد.

پس از کشف این دانش دو حالت قابل تصور است:

حالت اول آن که افراد با تجربه در دامنه مورد کاوش، به دانش استخراج شده آگاه بوده باشند که در این صورت آن دانش به عنوان قانونی صحیح تلقی خواهد شد.

حالت دوم افراد خبره از قبل به دانش کشف شده آگاهی نداشته باشند و در نتیجه این دانش کاملا جدید و ناشناخته باشد. در این حالت دانش جدید به عنوان یک فرضیه مطرح شده و درست یا غلط بودن آن با آزمایشات متعدد اثبات می‌شود که در صورت اثبات درستی، فرضیه حاصل از دانش کشف شده به قانون تبدیل می‌گردد.

در حالت کلی داده‌کاوی به معنای "معدن‌کاری" دانش از میان مقدار زیادی داده خام است. البته این نامگذاری تا حد زیادی برای این فرآیند نامناسب است. چرا که عملیات معدن‌کاری برای استخراج طلا از میان ماسه و صخره را طلا کاوی می‌نامیم نه صخره کاوی. لذا شاید نام بهتر برای این فرآیند، عنوان "دانش کاوی" باشد که به جستجو در میان داده‌ها برای کشف دانش می‌پردازد. اما نقص کوچک این نام هم این است که بر جستجو میان حجم عظیم داده‌ها دلالت ندارد حال آنکه عبارت معدن‌کاری بلافاصله انسان را به یاد جستجو در انبوه مواد خام برای کشف قطعه‌ای کوچک اما ارزشمند می‌اندازد. در هر صورت نام داده‌کاوی تا حدی برای چنین فرآیندی ناقص است اما باید توجه داشت که این نام بسیار فراگیر است و کاملا عمومیت پیدا کرده است. ذکر این نکته ضروری به نظر می‌رسد که در حقیقت داده‌کاوی بخشی از فرآیند بزرگ کشف دانش است.

کشف دانش دارای مراحل زیر است:[۲۲]

۱- پاکسازی داده‌ها (از بین بردن نوسان و ناسازگاری داده‌ها).

۲- یکپارچه‌سازی داده‌ها (چندین منبع داده ترکیب می‌شوند).

۳- انتخاب داده‌ها (داده‌های مرتبط با آنالیز از پایگاه داده بازیابی می‌شوند).

۴- تبدیل کردن داده‌ها (تبدیل داده‌ها به فرمی که مناسب برای داده‌کاوی باشد مثل خلاصه سازی).

۵- داده‌کاوی (فرآیند اصلی که روندهای هوشمند برای استخراج الگوها از داده‌ها به کار گرفته می‌شوند).

۶- ارزیابی الگو (برای مشخص کردن الگوهای صحیح و مورد نظر به وسیله معیارهای اندازه‌گیری).

۷- ارائه دانش (یعنی نمایش بصری، تکنیکهای بازنمایی دانش برای ارائه دانش کشف شده به کاربر استفاده می‌شود) .

۴.۱ موارد استفاده از داده‌کاوی

افرادی که در فعالیت‌های تجاری مشارکت دارند، بزرگ‌ترین گروه استفاده کننده ازتکنیک‌ها و مفاهیم داده‌کاوی را تشکیل می‌دهند. زیرا به طور منظم و همیشگی حجم زیادی از داده‌ها را جمع‌آوری کرده و تمایل زیادی دارند که به درک درست و کاملی از آن داده‌ها برسند. هدف این افراد آن است که شرکت‌های تجاری‌شان سود آورتر شده و نیز توانایی رقابت بیشتر و موفق‌تر با شرکت‌های دیگر را به دست آورند. صاحبان داده نه تنها ترجیح می‌دهند که داده خود را بهتر بشناسند، بلکه مایل هستند تا دانش جدیدی (که درون داده‌ها پنهان است) را در رابطه با زمینه فعالیت خویش کسب نمایند. هدف آن‌ها حل مسائل و مشکلات با استفاده از راه‌های جدید و درصورت امکان بهتر می‌باشد.

۵.۱ روند پیشرفت داده‌کاوی

- از ۱۹۶۰
  - ایجاد سیستم‌های جمع‌آوری و مدیریت داده‌ها.
  - ذخیره داده‌ها روی دیسک‌ها و کامپیوترها.
  - بازیابی ایستا (محاسبه کل سود یک فروشگاه در ۵ سال گذشته).
- از ۱۹۸۰
  - ایجاد زبان پرس و جو برای تهیه گزارشات از پایگاه داده.
  - شاخص‌گذاری و سازماندهی داده‌ها.
  - بازیابی پویا در سطح رکورد (میزان فروش یک کالا در یک شعبه بصورت روزانه).
- از ۱۹۹۰
  - ایجاد پایگاه داده‌های چند بعدی.
  - بازیابی پویا در چند سطح.

■ یافتن اطلاعات کاملی از رخدادهای گذشته با این کمبود که نمی‌تواند بگوید چرا اتفاق افتاده و یا پیش‌بینی کند.

- در حال حاضر[۲]

■ ابزارهای پیشرفته مانند SPSS وSAS.

■ کشف الگوهای جدید در پایگاه داده‌ها.

■ بازیابی پویا با نگاه پیشرو به آینده (فروش یک کالا در ماه آینده در یک شعبه خاص چقدر است و چرا؟).

۶.۱ ویژگی‌های داده‌کاوی

۱.۶.۱ قابلیت‌های داده‌کاوی

۱) عملکرد بدون هیچ پیش‌زمینه و شناخت قبلی از برچسب داده‌ها.

۲) کارکرد بدون در نظر گرفتن ساختار داده‌ها یعنی استقلال داده‌کاوی از ساختار داده‌ها.

۳) پیش بینی وقایع آینده بر اساس روند گذشته.

۴) طبقه‌بندی اشیا و افراد برای شناسایی الگو.

۵) دسته‌بندی اشیا و افراد بر اساس صفات و ویژگی‌ها.

۶) شناسایی وقایعی که احتمال دارد همزمان رخ دهند.

۷) شناسایی وقایعی که یکی باعث وقوع دیگری می‌شود.

و مهم‌تر از همه:

۸) سازگاری با حجم انبوه داده‌ها و ابعاد زیاد آن‌ها.

۲.۶.۱ چالش‌های داده‌کاوی

۱) دخالت مستقیم کاربر، به خصوص در مراحل اولیه یک فرآیند داده‌کاوی.

۲) نیاز به پاکسازی داده‌ها. (مدیریت داده‌های مفقود شده، غیر دقیق، ناقص، افزونه و دارای اختلال)

۳) دقیق نبودن داده‌کاوی.

۳.۶.۱ کاربردهای داده‌کاوی

تعیین الگوهای خرید مشتریان

خرده‌فروشی: تجزیه و تحلیل سبد خرید بازار پیشگویی میزان خرید مشتریان از طریق پست (فروش الکترونیکی)

بیمه: تجزیه و تحلیل دعاوی پیشگویی میزان خرید بیمه نامه‌های جدید توسط مشتریان

پزشکی: تعیین نوع رفتار با بیماران و پیشگویی میزان موفقیت اعمال جراحی و تعیین میزان موفقیت روش‌های درمانی در برخورد با بیماری‌های سخت

بانکداری: پیش‌بینی الگوهای کلاهبرداری از طریق کارت‌های اعتباری و تعیین میزان استفاده از کارت‌های اعتباری بر اساس گروه‌های اجتماعی

اعتبارسنجی مشتریان برای ارائه خدمات بیشتر

۷.۱ مراحل داده‌کاوی

۱) تعریف مساله

۲) جستجوی داده

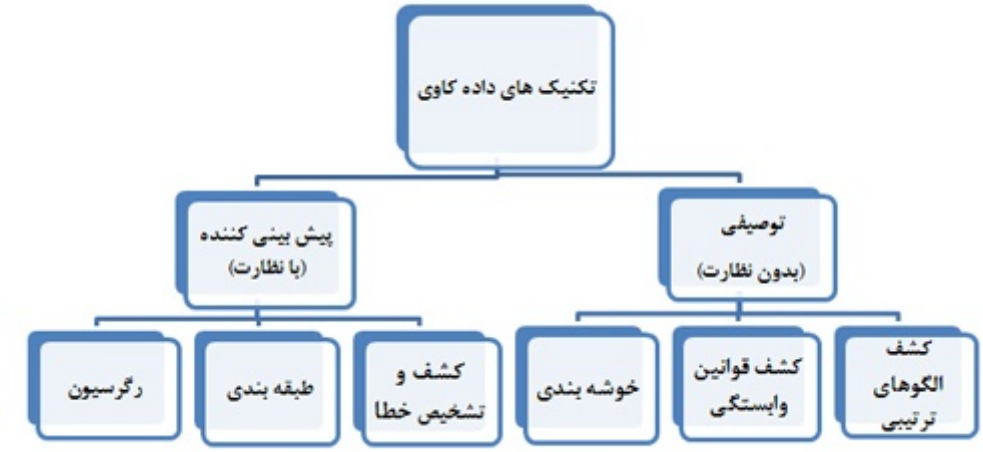
۳) آماده ساختن داده برای مدل سازی

۴) ساختن مدلی بر اساس داده‌ها

۵) پیش‌بینی بر اساس مدل ساخته شده

مثال۴.۱: در یک شرکت بیمه مساله مورد نظر، کشف روابطی میان نحوه انتخاب شرکت توسط مشتریان و مدت زمان استفاده از خدمات آن شرکت بیمه است. در این بررسی از داده‌های موجود از سال‌های قبل می‌توان استفاده کرد و یا بوسیله روش‌های نمونه‌گیری داده‌های جدیدی را ایجاد نمود. سپس می‌بایست به مدیریت داده‌های ناقص، غیر دقیق، افزونه و اختلال‌ها بپردازیم. و در نهایت بوسیله آمار و داده‌کاوی مدلی ساخته و از آن برای پیش‌بینی و نتیجه‌گیری استفاده می‌کنیم.

۸.۱ تکنیک‌های داده‌کاوی



تکنیک توصیفی : در روش‌های توصیفی، به دنبال پیدا کردن الگوریتم‌هایی هستیم که روابط حاکم بر داده‌ها را با متغیر خروجی تبیین می‌نماید.

تکنیک پیش‌بینی کننده : در این روش، الگوریتم می‌کوشد تا مدلی را تولید کند که متغیر پاسخ را پیش‌بینی کند.

۹.۱ کاربردهای آمار در هر بخش از داده‌کاوی

الف) درآماده‌سازی داده‌ها [۲۸]:

انتخاب مؤثرترین متغیرها از طریق:

- تحلیل مؤلفه‌های اصلی

- تفکیک‌کننده خطی فیشر
- تجزیه مقدار منفرد
- تحلیل مؤلفه‌های مستقل
- گشتاورهای زرنیک

ب) در داده‌کاوی با نظارت[۲۸]:

- رگرسیون (جزء جدانشدنی از آمار)
- درخت‌های تصمیم در طبقه‌بندی

ج) در داده‌کاوی بدون نظارت[۲۸]:

- خوشه‌بندی بر اساس الگوریتم K-Means (مبتنی بر کمینه واریانس)
- خوشه‌بندی بر اساس الگوریتم K-medoids (مبتنی بر میانه)
- خوشه‌بندی بر اساس الگوریتم C-Means فازی