

استخراج دانش (داده کاوی)

در دو دهه گذشته توانایی‌های فنی بشر برای تولید و جمع‌آوری داده‌ها به سرعت افزایش یافته‌است. عواملی نظیر استفاده گسترده از بارکد برای تولیدات تجاری، به خدمت گرفتن کامپیوتر در کسب و کار، علوم، خدمات دولتی و پیشرفت در وسائل جمع‌آوری داده از اسکن کردن متون و تصاویر تا سیستم‌های سنجش از دور ماهواره‌ای، در این تغییرات نقش مهمی داشته‌اند (تولید حجم انبوهی از داده‌ها). بنابراین امروز در روزگاری زندگی می‌کنیم که عصر انفجار اطلاعات نامیده می‌شود. همه روزه با حجم وسیعی از داده‌ها و اطلاعات پیرامون خود مواجه هستیم. اما حلقه‌ی مفقوده‌ی عصر ما دانش است. رشد روز افزون داده‌ها در شاخه‌های مختلف صنعت و علوم باعث شده است تا از کامپیوتر و علوم مربوط به آن جهت پردازش این حجم بالا از داده‌ها استفاده شود. هدف از پردازش داده‌ها، استخراج دانش از آنها به گونه‌ای است که بتوان در کاربردهای دیگر از آنها استفاده نمود.¹

داده کاوی: علمی است که حجم عظیمی از داده‌ها را مورد پردازش عمیق قرار می‌دهد تا نظم‌هایی را که در عمق داده‌ها وجود دارد - همچون طلا در یک معدن طلا - به صورت دانشی با ارزش کشف و جهت استفاده عرضه می‌کند.

انگیزه‌ها و دلایل اصلی ظهور داده کاوی

- رشد روز افزون داده‌ها (انفجار اطلاعات)
- معمولاً دانشی که در داده‌ها وجود دارد، خیلی بدیهی و روشن نیست.
- کشف اطلاعات با ارزش از داده‌ها توسط تحلیل‌گران انسانی ممکن است هفته‌ها یا ماه‌ها زمان نیاز داشته باشد.

داده‌های مورد استفاده در داده کاوی عمدتاً در یکی از دو نوع زیر خواهد بود 1- داده‌های تجاری 2- داده‌های علمی

سه منبع اصلی برای داده‌های تجاری قابل تصور است:

- (1) داده‌های وب یا داده‌های تجارت الکترونیک
- (2) خرید و فروش‌های موجود در فروشگاه‌ها
- (3) تراکنش‌های بانکی و کارت‌های بانکی

نکته: انگیزه اصلی داده کاوی در داده‌های تجاری، انگیزه مالی است که باعث می‌شود شرکت‌ها برای رقابت بهتر از کاوش داده‌ها بهره بگیرند.

به طور کلی برای جمع‌آوری داده‌های علمی چهار منبع وجود دارد که عبارتند از:

- (1) تصاویر ارسالی از طریق ماهواره‌ها
- (2) تصاویر ارسالی از طریق تلسکوپ‌ها
- (3) داده‌های دنباله زنی
- (4) داده‌های حاصل از شبیه‌سازی علمی

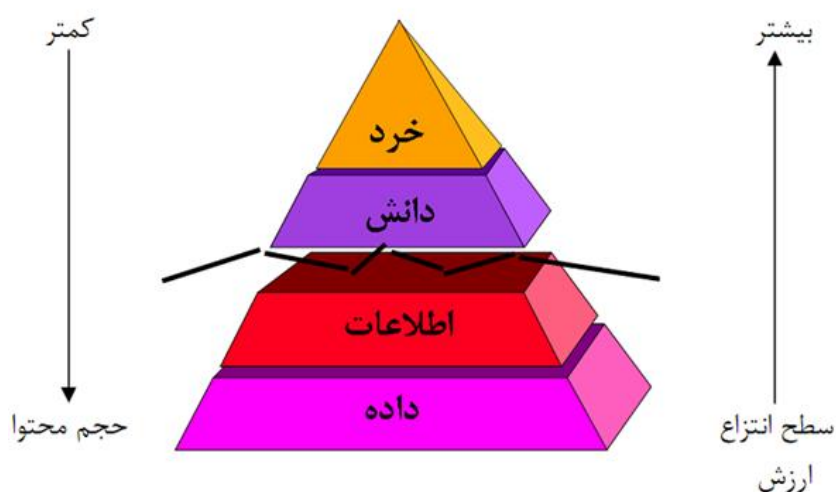
نکته: انگیزه اصلی در پردازش داده‌های علمی، گسترش مرزهای دانش بشری در یک حوزه خاص علمی است.

¹ - بخشی از مطالب سری نهم از وبلاگ مصطفی سبزه کار و بخشی هم از کتاب داده کاوی کاربردی - نویسندگان: صنیعی، محمودی و طاهرپور - برگرفته شده‌است.

شرایط لازم برای داده کاوی 1- وجود داده 2- صحت داده 3- کافی بودن ویژگی‌هاست. برای فهم و درک بهتر داده کاوی لازم است تعاریف داده، اطلاعات و دانش، مرور و یادآوری شود.

- **داده:** هر گونه سمبل، عدد، رقم، کاراکتر، رشته و یا سیگنالی که معنای خاصی نداشته باشد (فاقد معنای خاصی باشد)
- **اطلاعات:** هر گونه سمبل، عدد، رقم، کاراکتر، رشته و یا سیگنالی که معنای خاصی داشته باشد (دارای معنای خاصی باشد)
- **دانش:** وجود رابطه بین دو عنصر اطلاعاتی بیان می‌کند.

در شکل 1-9 که هرم دانش نامیده می‌شود سلسله مراتب دانش، نشان داده شده است. همانطور که از شکل 1-9 پیداست برای رسیدن به دانش موجود در داده‌ها، راه سختی در پیش داریم.



شکل 1-9: هرم دانش (یادآوری)

🔖 **نکته:** با افزایش ارزش معنایی مفاهیم، حجم آنها کاهش می‌یابد که کاملاً طبیعی است. به عبارت دیگر حجم بالایی از داده‌ها را می‌توان تنها با چند قانون توصیف و تبیین کرد که از اهداف اصلی داده کاوی است.

هدف اصلی در داده کاوی، کشف دانش است. دانش نظم موجود داده‌هاست. پس از کشف دانش ممکن است یکی از دو حالت زیر اتفاق بیفتد.

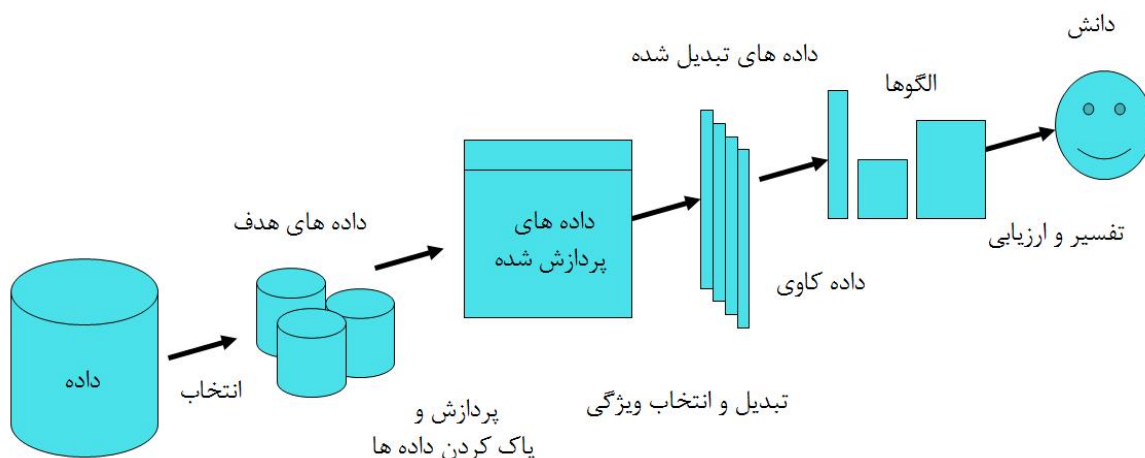
- 1- افراد خبره، آگاه به دانش استخراج شده باشند (دانش به صورت قانون تلقی می‌شود)
- 2- دانش کشف شده دانش جدید باشد. در این حالت، دانش جدید بررسی شده و در صورت منطقی بودن، به فرضیه تبدیل می‌شود. با آزمایش‌ها و بررسی‌های عمیق‌تر و بیشتر، درست یا نادرست بودن فرضیه اثبات می‌شود. اگر فرضیه اثبات شد، به قانون تبدیل خواهد شد.

خرد و دانش‌های پادیده نمی‌شود، خرد تنها در دراهم پیدا می‌گردد.

می‌توان گفت که برای فرایند استخراج دانش، دو رویکرد اصلی قابل تصور است:

- این فرایند به صورت مهندسی انجام شود (یعنی توسط متدولوژی‌های مهندس دانش انجام شود) که معمولاً اکتساب دانش یا دریافت دانش نامیده می‌شود. (فرم‌های خاصی پر شود، یا نمودارهایی ترسیم شود و ...) اکتساب دانش برای هر مسأله‌ای امکان پذیر نیست.
- راه حل دیگر داده کاوی است که به صورت خودکار دانش، استخراج می‌شود که به آن اکتشاف دانش هم گفته می‌شود.

اکتشاف دانش (Knowledge Discovery = KD) عبارتست از پروسه‌ای جهت استخراج اطلاعات مهم و اساسی، ضمنی، قبلاً ناشناخته و سودمند از داده‌های خام در پایگاه داده‌های بزرگ. هدف اصلی اکتشاف دانش، یافتن دانش نهفته در داده‌ها با کمترین (یا عدم) دخالت انسانی است. مراحل یک پروسه اکتشاف دانش در شکل 9-2، نشان داده شده است:



شکل 9-2: پروسه اکتشاف دانش

گردآوری داده: به طور کلی دو روش برای جمع آوری داده‌ها وجود دارد: در حالت اول خود طراح مدل، تولید داده را نیز کنترل می‌کند. این روش، آزمون طراحی شده، نامیده می‌شود. دومین وضعیت هنگامی است که شخص خبره نمی‌تواند تأثیری در فرآیند تولید داده داشته باشد. این وضعیت با نام رویکرد مشاهده‌ای شناخته می‌شود.

پاک‌سازی داده‌ها (Data Cleaning): در این مرحله داده‌های غیرمعتبر از مجموعه داده‌های آموزشی خارج می‌شود. داده‌های دارای خطا یا پرت و اطلاعات ناکامل، نمونه‌هایی از داده‌های زائد هستند که باید پاک‌سازی در مورد آنها انجام شود.

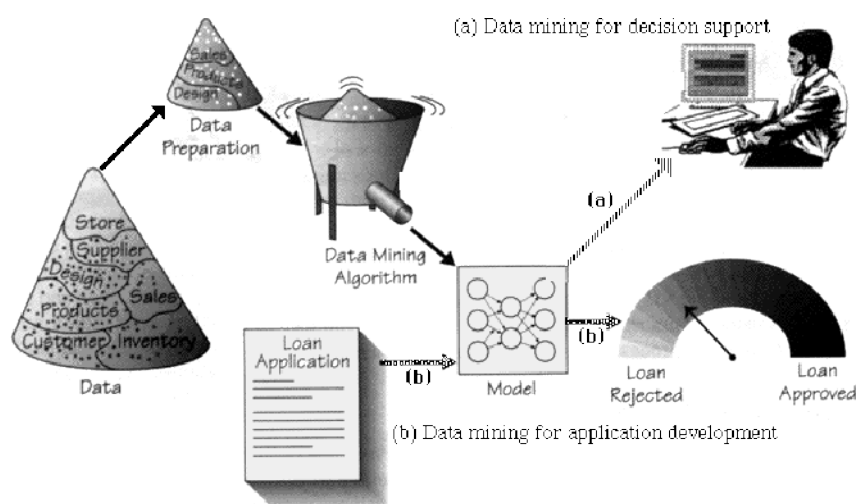
تبدیل داده‌ها (Data Transformation): در این گام داده‌ها به قالبی قابل استفاده برای داده کاوی در می‌آیند.

برآورد مدل یا داده کاوی: بخش اصلی فرآیند داده کاوی این بخش است که در آن با استفاده از روش‌ها و تکنیک‌های خاص، استخراج الگوهای دانش صورت می‌گیرد. به طور کلی روش‌ها و الگوریتم‌های مختلفی جهت یادگیری و تولید یک مدل بر اساس داده‌های ورودی وجود دارد. به نوعی الگوریتم‌های مزبور را می‌توان یک روال جستجو نیز در نظر گرفت. این روال سعی می‌کند مدلی پیدا کند که به بهترین نحو داده‌های ورودی را پوشش دهد. بایستی توجه نمود که الگوریتم‌های داده کاوی که در این مرحله اجرا می‌گردند، با توجه به ماهیت مسأله‌ای که فرآیند داده کاوی سعی در تحلیل داده‌های آن را دارد، طراحی می‌گردند. به عبارت دیگر الگوریتم مزبور با توجه به انواع کاربردهای داده کاوی، پیاده‌سازی می‌گردد.

ارزیابی الگوها (Pattern Evaluation): تشخیص الگوهای صحیح مورد نظر از سایر الگوها در این مرحله انجام می‌شود. صحت الگوها بر اساس یک‌سری از معیارهای سنجیده انجام می‌شود.

نمایش دانش (Knowledge Representation): در این بخش به منظور ارائه دانش استخراج شده به کاربر، از یک سری ابزارهای بصری سازی استفاده می شود. برای استفاده از دانش و مدل استخراج شده، بایستی آن دانش قابل تفسیر باشد. این امر به خاطر آن است که انسان تمایل ندارد که اساس و پایه ی تصمیم های خود را بر مبنای مدل های پیچیده ی جعبه سیاه قرار دهد. موضوع مهمی که اینجا وجود دارد آن است که اهداف دقت مدل و قابلیت درک آسان، معمولاً با هم در تضاد هستند. اغلب مدل های ساده، قابلیت تفسیر بهتری دارند اما دقت آنها پایین تر است. از طرف دیگر مدل های دقیق معمولاً ساختار پیچیده ای دارند.

پرسش: شکل 9-3 را با توجه به مراحل پروسه اکتشاف دانش تشریح کنید.



شکل 9-3: فرایند داده کاوی در یک مثال کاربردی

نکته: واژه های «داده کاوی»، «اکتشاف دانش در پایگاه داده» و «استخراج دانش از پایگاه داده» اغلب به صورت مترادف با یکدیگر استفاده می شوند.

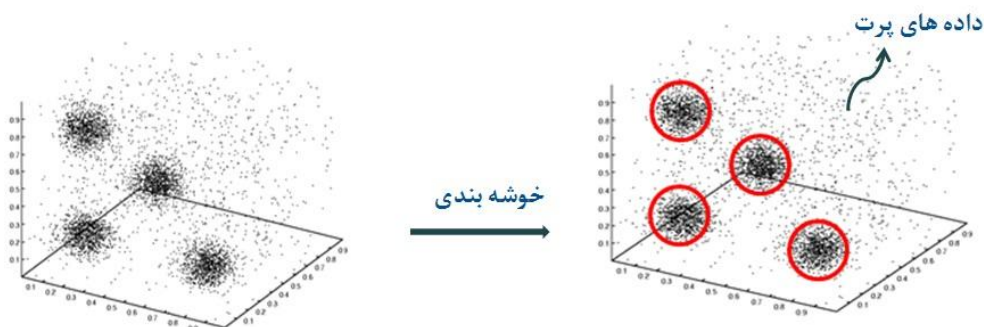
انواع روش های اکتشاف دانش (داده کاوی) از داده ها

- خوشه بندی (Clustering): در خوشه بندی، هدف یافتن مجموعه متناهی از خوشه ها برای توصیف داده ها است.
- دسته بندی (Classification): هدف در دسته بندی داده ها این است که یک مدل پیشگویی کننده بدست آوریم که این مدل اولاً توانایی دسته بندی داده های ورودی را داشته باشد و ثانیاً بتوان از آن جهت پیشگویی برای تعیین دسته ی یک داده که تازه به سیستم اضافه شده، استفاده نمود.
- تخمین (Regression): هدف در رگرسیون ارائه ی یک مدل پیشگویی کننده با توانایی نگاشت یک نمونه ی داده ای به یک متغیر تخمینی است.
- ...

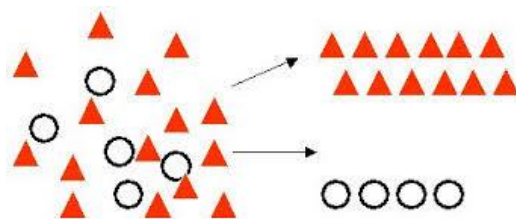
خوشه‌بندی عبارتست از قرار دادن اشیای داده‌ای در چند خوشه به نحوی که:

- داده‌هایی که در یک دسته قرار می‌گیرند، بیشترین میزان شباهت را به هم داشته باشند.
- داده‌هایی که در دسته‌های مجزا هستند، بیشترین عدم شباهت را به هم داشته باشند.

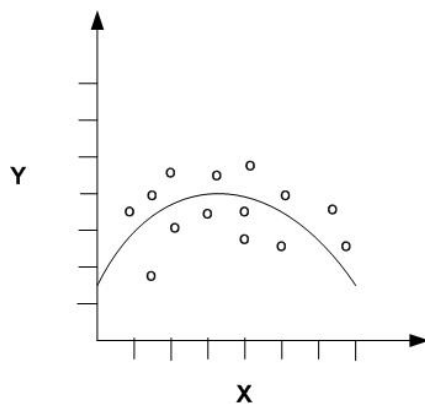
شکل 9-4، نحوه عملکرد روش‌های خوشه‌بندی را شرح می‌دهد. همانطور که در شکل 9-4 مشخص است داده‌ها به چهار خوشه مجزا تقسیم شده‌اند و داده‌های باقی مانده هم به عنوان داده‌های پرت و نویزی تلقی می‌شوند.



شکل 9-4: خوشه‌بندی داده‌ها



شکل 9-5: دسته‌بندی داده‌ها



شکل 9-6: رگرسیون داده‌ها

داشتن؛ تصمیمی برای رفتار خوب نیست، اما نادانی یک تصمیم بالقوه برای رفتار بد است! (هارتنام)

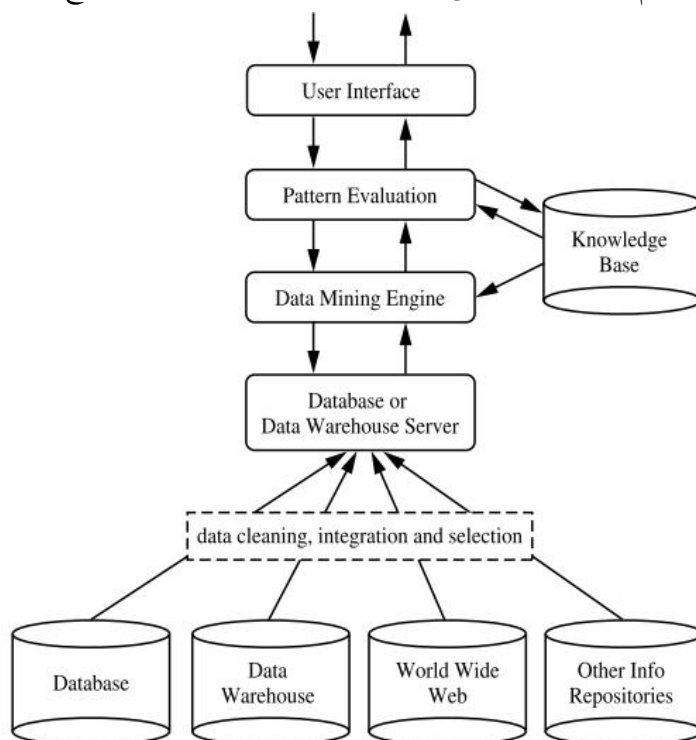
داده کاوی چه کارهایی نمی تواند انجام دهد؟

داده کاوی فقط یک ابزار است و نه یک عصای جادویی. داده کاوی به این معنی نیست که شما راحت به کناری بنشینید و ابزارهای داده کاوی همه کار را انجام دهد. داده کاوی نیاز به شناخت داده ها، ابزارهای تحلیل و افراد خبره در این زمینه ها را از بین نمی برد. داده کاوی فقط به تحلیلگران برای پیدا کردن الگوها و روابط بین داده ها کمک می کند و در این مورد نیز روابطی که یافته می شود باید به وسیله داده های واقعی، دوباره بررسی و تست گردد.

تمرین ها

تمرین 9-1*: چه روش هایی دیگری علاوه بر روش های معرفی شده، در داده کاوی (استخراج دانش) استفاده می شوند؟

تمرین 9-2: شکل 9-7، معماری یک سیستم داده کاوی را نشان می دهد، با مثال کاربردی این معماری را تشریح کنید.



شکل 9-7: معماری یک سیستم داده کاوی (Han & Kamber, 2006)

تمرین 9-3*: روش K نزدیکترین همسایه (KNN = K Nearest Neighbor) چگونه کار می کند؟ با مثال ساده، توضیح دهید. KNN روش دسته بندی محسوب می شود یا خوشه بندی؟ چرا؟

تمرین 9-4*: مراحل الگوریتم K-Means چیست؟ روند کار آن را با مثال توضیح دهید. این روش جزء روش دسته بندی است یا خوشه بندی؟ چرا؟

معلومات گفته را به کاربردن و از روی آن معلومات تازه ای به دست آوردن، از اصول عمو آموزش است. (گنسیوس)