# بررسی روش‌های داده‌کاوی و کاربرد آن در آمارهای رسمی

حسین هشیارمنش (مجری)

کاوه کیانی

پژوهشکده‌ی آمار

گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی

تابستان ۱۳۹۴

به نام خداوند جان و خرد

# پیش‌گفتار

امروزه با تولید انبوه داده‌ها، به‌ویژه از طریق وسایل الکترونیکی، دیگر نمی‌توان از روش‌های سنّتی تحلیل داده‌ها استفاده کرد. از این رو علم آمار در پیوند با علوم رایانه، به‌ویژه دادگان‌ها و هوش مصنوعی و یادگیری خودکار، ابزار جدیدی به‌نام «داده‌کاوی» را پدید آورده است. به این ترتیب، داده‌کاوی را باید امتداد همان علم آمار دانست.

نظام ملی آمار ایران که تولید کننده‌ی آمارهای رسمی کشور است، روزبه‌روز با دادگان‌های بیش‌تر و بزرگ‌تری درگیر می‌شود. با توجه به قابلیت‌های روش‌های داده‌کاوی در کشف الگوها و ارتباط‌های پنهان در دادگان‌های عظیم، به نظر می‌رسد که می‌توان این روش‌ها را بر روی دادگان‌های تولیدکنندگان آمار رسمی نیز به کار گرفت و الگوها و ارتباط‌های مفید و آگاهی‌بخشی را استخراج کرد.

پژوهشکده‌ی آمار با توجه به رسالت خود در زمینه‌ی اجرای طرح‌های پژوهشی برای پاسخ‌گویی به نیازهای پژوهشی نظام ملی آمار ایران، اجرای طرح پژوهشی «بررسی روش‌های داده‌کاوی و کاربرد آن در آمارهای رسمی» را در دستور کار خود قرار داد. هدف این طرح پژوهشی، معرفی و بررسی روش‌های داده‌کاوی و کاربرد آن‌ها در نظام آماری است.

این پژوهش در گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی پژوهشکده‌ی آمار، توسط آقای حسین هشیارمنش و با همکاری آقای دکتر کاوه کیانی به انجام رسیده است، که از آنان صمیمانه تشکّر و قدردانی می‌شود.

گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی

پژوهشکده‌ی آمار

# فهرست مطالب

# ۱

# مقدمه‌ای بر داده‌کاوی

در دو دهه قبل توانایی‌های فنی بشر در تولید و جمع آوری داده‌ها به سرعت افزایش یافته است. عواملی نظیر استفاده گسترده از بارکد برای تولیدات تجاری، به خدمت گرفتن کامپیوتر در کسب و کار، علوم، خدمات دولتی و پیشرفت در وسائل جمع آوری داده، از اسکن کردن متون و تصاویر تا سیستم‌های سنجش از راه دور ماهواره‌ای، در این تغییرات نقش مهمی دارند.

به طور مثال استفاده همگانی از وب و اینترنت به عنوان یک سیستم اطلاع رسانی جهانی ما را با حجم زیادی از داده و اطلاعات مواجه می‌کند. این رشد انفجاری در داده‌های ذخیره شده، نیاز مبرم به وجود تکنولوژی‌های جدید و ابزارهای خودکاری را ایجاد کرده است که به صورت هوشمند به انسان یاری رسانند تا بتواند این حجم زیاد داده را به اطلاعات و دانش تبدیل کند. داده‌کاوی به عنوان یک راه حل برای این مسائل مطرح می‌باشد. در یک تعریف غیر رسمی، داده‌کاوی فرآیندی است، خودکار برای استخراج الگوهایی که دانش را بازنمایی می‌کنند، که این دانش به صورت ضمنی در پایگاه داده‌های عظیم، انباره داده[۱] و دیگر مخازن بزرگ اطلاعات، ذخیره شده است. داده‌کاوی به طور هم‌زمان از چندین رشته علمی بهره می‌برد نظیر: تکنولوژی پایگاه داده، هوش مصنوعی، یادگیری ماشین، شبکه‌های عصبی، آمار، شناسایی الگو، سیستم‌های مبتنی بر دانش[۲]، حصول دانش[۳]، بازیابی اطلاعات[۴]، محاسبات سرعت بالا[۵] و دیداری سازی داده‌ها[۶] . داده‌کاوی در اواخر دهه ۱۹۸۰ پدیدار گشته است. پژوهشگران در دهه ۱۹۹۰ موفق شدند گام‌های بلندی در این شاخه از علم برداشته و انتظار رشد و پیشرفت آن‌را در این قرن را ایجاد کرده‌اند.

---

[۱] Data warehouses

[۲] Knowledge-based system

[۳] Knowledge-acquisition

[۴] Information retrieval

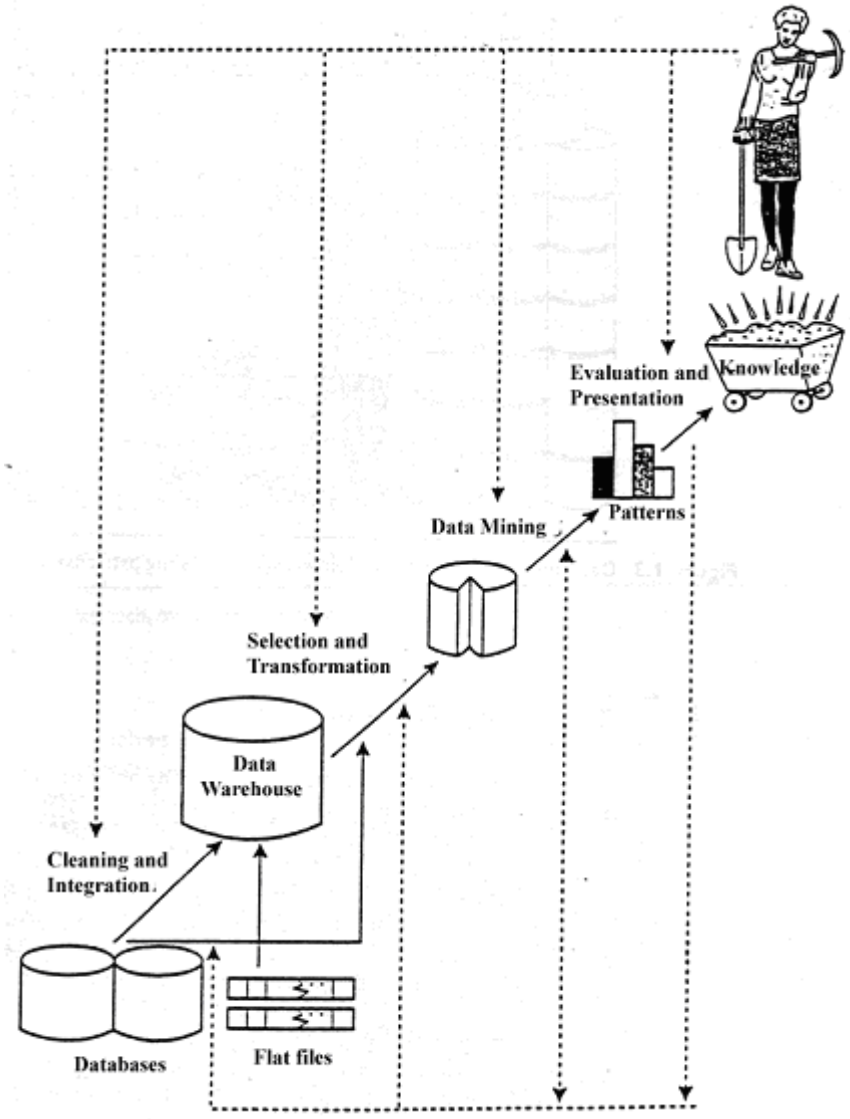
[۵] High-performance computing

[۶] Data visualization

در سال ۱۹۸۹ و ۱۹۹۱ کارگاه‌های کشف دانش از پایگاه داده‌ها توسط پیاتتسکی و همکارانش برگزار شد. در فواصل سال‌های ۱۹۹۱ تا ۱۹۹٤ کارگاه‌های کشف دانش از پایگاه داده‌ها توسط فییاد و پیاتتسکی و دیگران برگزار شد. به طور رسمی اصطلاح داده‌کاوی برای اولین بار توسط فییاد در اولین کنفرانس بین‌المللی «کشف دانش و داده‌کاوی» در سال ۱۹۹۵ مطرح شد. امروزه کنفرانس‌های مختلفی در این زمینه در سراسر دنیا برگزار می‌شود. از سال ۱۹۹۵ داده‌کاوی به طور جدی وارد مباحث آمار شد و در سال ۱۹۹٦ اولین شماره مجله کشف دانش از پایگاه داده‌ها منتشر شد. محققانی نظیر براچمن و آناند کلیه مراحل واقع‌گرایانه و رو به جلو کشف دانش از پایگاه داده‌ها را تشخیص دادند.

واژه‌های «داده‌کاوی» و «کشف دانش در پایگاه داده»[۷] اغلب به صورت مترادف یکدیگر مورد استفاده قرار می‌گیرند. کشف دانش به عنوان یک فرآیند در شکل ۱-۱ نشان داده شده است.

کشف دانش در پایگاه داده فرایند شناسایی درست، ساده، مفید، و نهایتاً الگوها و مدل‌های قابل فهم در




شکل ۱-۱: داده‌کاوی به عنوان یک مرحله از فرآیند کشف دانش

[۷] Knowledge discovery in database

داده‌ها می‌باشد. داده‌کاوی، مرحله‌ای از فرایند کشف دانش می‌باشد و شامل الگوریتم‌های مخصوص داده‌کاوی است، بطوریکه، تحت محدودیت‌های مؤثر محاسباتی قابل قبول، الگوها و یا مدل‌ها را در داده کشف می‌کند. به بیان ساده‌تر، داده‌کاوی به فرایند استخراج دانش ناشناخته، درست، و بالقوه مفید از داده اطلاق می‌شود. تعریف دیگر اینست که، داده‌کاوی گونه‌ای از تکنیک‌ها برای شناسایی اطلاعات و یا دانش تصمیم گیری از قطعات داده می‌باشد، به نحوی که با استخراج آن‌ها، در حوزه‌های تصمیم گیری، پیش بینی، پیشگویی، و تخمین مورد استفاده قرار می‌گیرند. داده‌ها اغلب حجیم ، اما بدون ارزش می‌باشند، داده به تنهایی قابل استفاده نیست، بلکه دانش نهفته در داده‌ها قابل استفاده می‌باشد. به این دلیل اغلب به داده‌کاوی، تحلیل داده‌ای ثانویه[۸] گفته می‌شود.

## ۱-۱- داده‌کاوی چیست؟

یکی از دلایلی که باعث شد داده‌کاوی در کانون توجهات صنعت اطلاعات قرار بگیرد، مسأله در دسترس بودن حجم وسیعی از داده‌ها و نیاز شدید به استخراج اطلاعات و دانش سودمند از این داده‌ها است. اطلاعات و دانش بدست آمده در کاربردهای وسیعی از مدیریت کسب و کار و کنترل تولید و تحلیل بازار تا طراحی مهندسی و تحقیقات علمی مورد استفاده قرار می‌گیرد.

داده‌کاوی را می‌توان حاصل سیر تکامل طبیعی تکنولوژی اطلاعات دانست، که این سیر تکاملی ناشی از یک سیر تکاملی در صنعت پایگاه داده می‌باشد، نظیر عملیات جمع آوری داده‌ها و ایجاد پایگاه داده، مدیریت داده و تحلیل و فهم داده‌ها. در شکل ۱-۲ این روند تکاملی در پایگاه‌های داده نشان داده شده است.

تکامل تکنولوژی پایگاه داده و استفاده فراوان آن در کاربردهای مختلف سبب جمع آوری حجم فراوانی داده شده است. این حجم عظیم داده‌ها، نیاز به وجود ابزارهای قدرتمند به منظور تحلیل داده‌ها را سبب شده است. زیرا در حال حاضر به لحاظ داده ثروتمند هستیم ولی دچار کمبود اطلاعات می‌باشیم.

ابزارهای داده‌کاوی داده‌ها را آنالیز می‌کنند و الگوهای دادها را کشف می‌کنند که می‌توان از آن در کاربردهایی نظیر: تعیین استراتژی برای کسب و کار، پایگاه دانش[۹] و تحقیقات علمی و پزشکی، استفاده کرد. شکاف موجود بین داده‌ها و اطلاعات سبب ایجاد نیاز برای ابزارهای داده‌کاوی شده است تا داده‌های بی ارزش را به دانشی ارزشمند تبدیل کنیم.

به طور ساده داده‌کاوی به معنای استخراج یا «معدن کاری[۱۰]» دانش از مقدار زیادی داده خام است. البته این نام‌گذاری برای این فرآیند تا حدی نامناسب است، زیرا به طور مثال عملیات معدن کاری برای استخراج طلا از صخره و ماسه را طلا کاوی می‌نامیم، نه ماسه کاوی یا صخره کاوی، بنابراین بهتر بود به این فرآیند نامی شبیه به «استخراج دانش از داده» می‌دادیم که متأسفانه بسیار طولانی است. «دانش کاوی» به عنوان یک عبارت کوتاه‌تر به عنوان جایگزین، نمی‌تواند بیانگر تاکید و اهمیت بر داده‌کاوی باشد. معدن کاری عبارتی است که بلافاصله انسان را به یاد فرآیندی می‌اندازد که در آن، ما به دنبال یافتن مجموعه کوچکی از قطعات ارزشمند از حجم بسیار زیادی از مواد خام هستیم.

---

[۸] Secondary data analysis

[۹] Knowledge base

[۱۰] Mining

تعاریف متنوعی از داده‌کاوی در مراجع مختلف و توسط افراد مختلف ارائه شده است. از جمله:

۱- داده‌کاوی عبارت است از فرآیند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های بزرگ و استفاده از آن در تصمیم‌گیری در فعالیت‌های تجاری مهم.

۲- اصطلاح داده‌کاوی به فرآیند نیمه خودکار تجزیه و تحلیل پایگاه داده‌های بزرگ به منظور یافتن الگوی مفید اطلاق می‌شود.

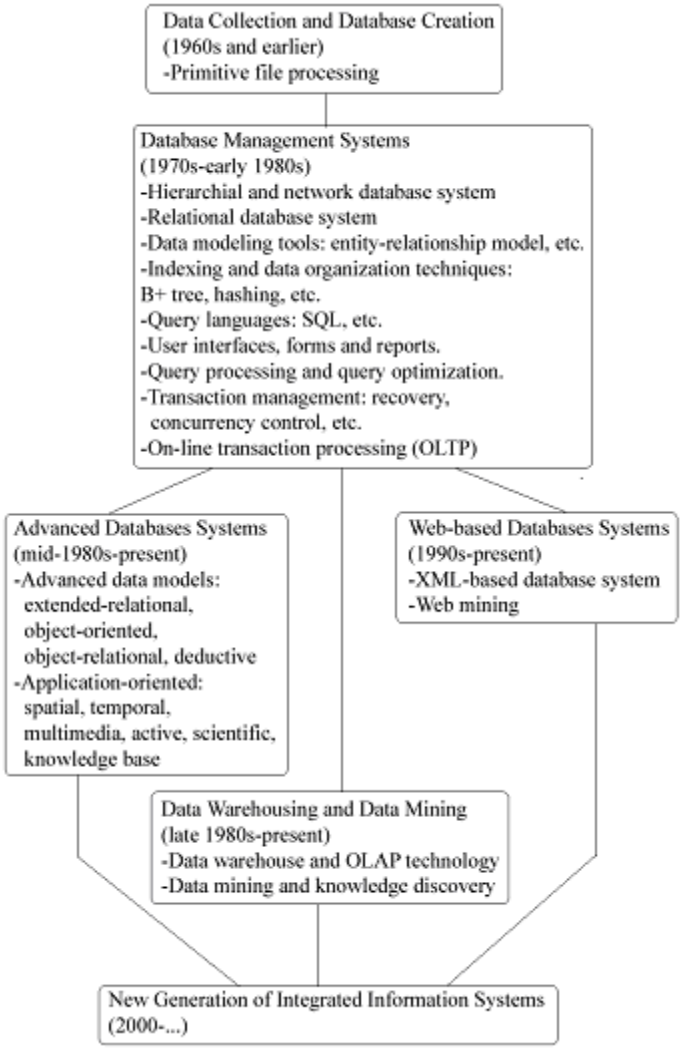
۳- داده‌کاوی یعنی جستجو در پایگاه داده‌ها برای یافتن الگوهای میان داده‌ها.

٤- داده‌کاوی یعنی تجزیه و تحلیل مجموعه داده‌های قابل مشاهده برای یافتن روابط مطمئن بین داده‌ها.

۵- عبارت داده‌کاوی مترادف با یکی از عبارت‌های استخراج دانش، برداشت اطلاعات، وارسی داده‌ها و حتی لای روبی کردن داده‌ها است، که در حقیقت کشف دانش در پایگاه داده‌ها را توصیف می‌کند.

اما تعریفی که در اکثر مراجع به اشتراک ذکر شده عبارت است از «استخراج اطلاعات و دانش و کشف الگوی پنهان از پایگاه داده‌های بسیار بزرگ و پیچیده». داده‌کاوی یک متدولوژی بسیار قوی و با پتانسیل بالا می‌باشد که به سازمان‌ها کمک می‌کند که بر روی مهم‌ترین اطلاعات از مخزن داده‌های خود تمرکز کنند.

داده‌کاوی فرآیندی است که از ابزارهای تحلیلی گوناگونی برای کشف الگوها و روابط بین داده‌ها استفاده




شکل۱-۲: سیر تکاملی صنعت پایگاه داده

می‌کند که ممکن است برای اعتبار بخشیدن به پیش‌بینی استفاده شود.

داده‌کاوی کمک می‌کند تا سازمان‌ها با کاوش بر روی داده‌های یک سیستم، الگوها و رفتارهای آینده را کشف و پیش‌بینی کرده و بهتر تصمیم بگیرند. داده‌کاوی با استفاده از تحلیل وقایع گذشته یک تحلیل خودکار و پیش‌بینانه ارائه می‌کند و به سوالاتی جواب می‌دهد که پاسخ آن‌ها در گذشته ممکن نبوده و یا پاسخ‌گویی به آن‌ها به زمان زیادی نیاز داشته است.

همان‌گونه که در تعاریف گوناگون داده‌کاوی مشاهده می‌شود، تقریباً در تمامی تعاریف به مفاهیمی چون استخراج دانش، تحلیل و یافتن الگوی بین داده‌ها، اشاره شده است.

با توجه به مطالب عنوان شده، با اینکه این فرآیند تا حدی دارای نام‌گذاری ناقص است ولی این نام‌گذاری یعنی داده‌کاوی بسیار عمومیت پیدا کرده است. البته اسامی دیگری نیز برای این فرآیند پیشنهاد شده که بعضاً بسیاری متفاوت با واژه داده‌کاوی است، نظیر: استخراج دانش از پایگاه داده، استخراج دانش[۱۱] ، آنالیز داده / الگو، باستان شناسی داده[۱۲]، و لای روبی داده‌ها[۱۳].

## ۱-۲- مراحل کشف دانش

داده‌کاوی فقط یک ابزار است و نه یک عصای جادویی. داده‌کاوی به این معنی نیست که شما راحت به کناری بنشینید و ابزارهای داده‌کاوی همه کار را انجام دهند.

داده‌کاوی نیاز به شناخت داده‌ها و ابزارهای تحلیل و افراد خبره در این زمینه‌ها را از بین نمی‌برد.

داده‌کاوی فقط به تحلیلگران برای پیدا کردن الگوها و روابط بین داده‌ها کمک می‌کند و در این مورد نیز روابطی که یافت می‌شوند باید به وسیله داده‌های واقعی دوباره بررسی و تست گردد.

کشف دانش دارای مراحل تکراری زیر است:

۱. پاک‌سازی داده‌ها[۱٤] (از بین بردن نویز و ناسازگاری داده‌ها).
۲. یکپارچه سازی داده‌ها[۱۵] (چندین منبع داده ترکیب می‌شوند).
۳. انتخاب داده‌ها[۱٦] (داده‌های مرتبط با آنالیز از پایگاه داده بازیابی می‌شوند).
٤. تبدیل کردن داده‌ها[۱۷] (تبدیل داده‌ها به فرم مناسب برای داده‌کاوی مثل خلاصه سازی[۱۸] و انبوهش[۱۹])
۵. داده‌کاوی (فرایند اصلی که روال‌های هوشمند برای استخراج الگوها از داده‌ها به کار گرفته می‌شوند.)
٦. ارزیابی الگو[۲۰] (برای مشخص کردن الگوهای صحیح و مورد نظر به وسیله معیارهای اندازه گیری)

---

[۱۱] Knowledge extraction
[۱۲] Data archaeology
[۱۳] Data dredging
[۱٤] Data cleaning
[۱۵] Data integration
[۱٦] Data selection
[۱۷] Data transformation
[۱۸] Summary
[۱۹] Aggregation
[۲۰] Pattern evaluation

۷. ارائه دانش[۲۱] (یعنی نمایش بصری، تکنیک‌های بازنمایی دانش برای ارائه دانش کشف شده به کاربر استفاده می‌شود).

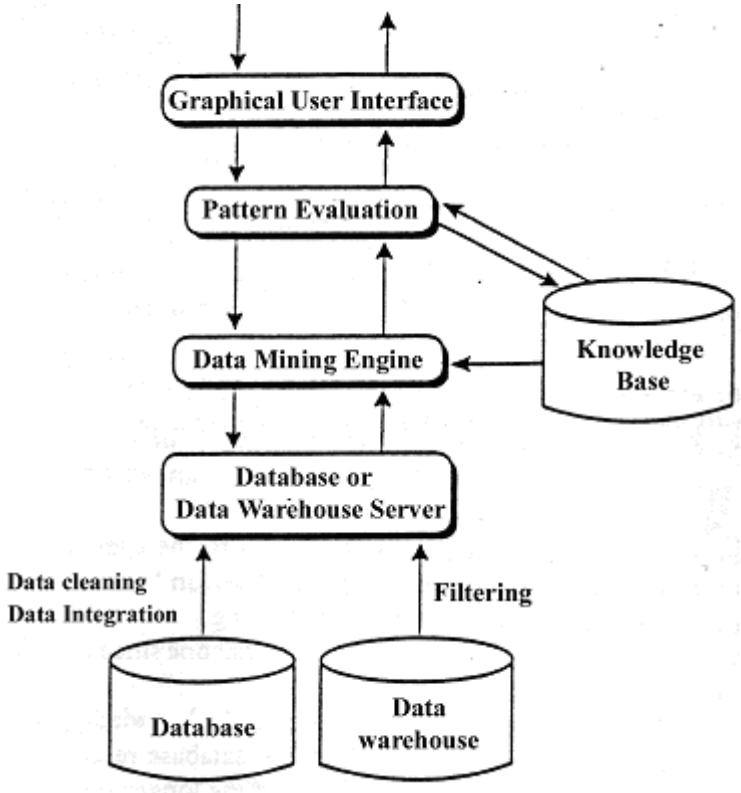
هر مرحله داده‌کاوی باید با کاربر یا پایگاه دانش تعامل داشته باشد. الگوهای کشف شده به کاربر ارائه می‌شوند و در صورت خواست او به عنوان دانش به پایگاه دانش اضافه می‌شوند. توجه شود که بر طبق این دیدگاه داده‌کاوی تنها یک مرحله از کل فرآیند است، البته به عنوان یک مرحله اساسی که الگوهای مخفی را آشکار می‌سازد. با توجه به مطالب عنوان شده، در اینجا تعریفی از داده‌کاوی ارائه می‌دهیم:

"داده‌کاوی عبارت است از فرآیند یافتن دانش از مقادیر عظیم داده‌های ذخیره شده در پایگاه داده، انباره داده و یا دیگر مخازن اطلاعات".

بر اساس این دیدگاه یک سیستم داده‌کاوی به طور نمونه دارای اجزاء اصلی زیر است که شکل۱-۳ بیانگر معماری سیستم است.

۱- پایگاه داده، انباره داده یا دیگر مخازن اطلاعات: که شامل مجموعه‌ای از پایگاه داده‌ها، انباره داده‌ها، صفحه‌های گسترده[۲۲]، یا دیگر انواع مخازن اطلاعات را شامل می‌شود. پاک‌سازی داده‌ها و تکنیک‌های یکپارچه سازی روی این داده‌ها انجام می‌شود.

۲- سرویس دهنده پایگاه داده یا انباره داده: که مسئول بازیابی داده‌های مرتبط بر اساس نوع درخواست داده‌کاوی کاربر می‌باشد.




شکل۱-۳: معماری یک نمونه سیستم داده‌کاوی

[۲۱] Knowledge presentation

[۲۲] Spread sheets

۳- پایگاه دانش: این پایگاه از دانش زمینه[۲۳] تشکیل شده تا به جستجو کمک کند، یا برای ارزیابی الگوهای یافته شده از آن استفاده می‌شود.

٤- موتور داده‌کاوی[۲٤] : این موتور جزء اصلی از سیستم داده‌کاوی است و به طور ایده‌آل شامل مجموعه‌ای از پیمانه[۲۵] هایی نظیر توصیف[۲٦]، تداعی[۲۷]، کلاس‌بندی[۲۸]، آنالیز خوشه‌ها[۲۹]، و آنالیز تکامل و انحراف[۳۰]، است.

۵- پیمانه ارزیابی الگو[۳۱] : این جزء معیارهای جذابیت[۳۲] را به کار می‌بندد و با پیمانه داده‌کاوی تعامل می‌کند بدین‌صورت که تمرکز آن بر جستجو بین الگوهای جذاب می‌باشد، و از یک حد آستانه جذابیت استفاده می‌کند تا الگوهای کشف شده را ارزیابی کند.

٦- واسط کاربر گرافیکی[۳۳] : این پیمانه بین کاربر و سیستم داده‌کاوی ارتباط برقرار می‌کند، به کاربر اجازه می‌دهد تا با سیستم داده‌کاوی از طریق پرس و جو[۳٤] ارتباط برقرار کند، این جزء به کاربر اجازه می‌دهد تا شمای پایگاه داده یا انباره داده را مرور کرده، الگوهای یافته شده را ارزیابی کرده و الگوها را در فرم‌های بصری گوناگون بازنمایی کند.

با انجام فرآیند داده‌کاوی، دانش، ارتباط یا اطلاعات سطح بالا از پایگاه داده استخراج می‌شود و قابل مرور از دیدگاه‌های مختلف خواهد بود. دانش کشف شده در سیستم‌های تصمیم یار، کنترل فرآیند، مدیریت اطلاعات و پردازش پرس و جو[۳۵] قابل استفاده خواهد بود.

بنابراین داده‌کاوی به عنوان یکی از شاخه‌های پیشرو در صنعت اطلاعات مورد توجه قرار گرفته و به عنوان یکی از نوید بخش ترین زمینه‌های توسعه بین رشته‌ای در صنعت اطلاعات است.

## ۱-۳- جایگاه داده‌کاوی در میان علوم مختلف

ریشه‌های داده‌کاوی در میان سه خانواده از علوم، قابل پیگیری می‌باشد. مهم‌ترین این خانواده‌ها، آمار کلاسیک[۳٦] می‌باشد. بدون آمار، هیچ داده‌کاوی وجود نخواهد داشت، بطوریکه آمار، اساس اغلب تکنولوژی‌هایی می‌باشد که داده‌کاوی بر روی آن‌ها بنا شده است. آمار کلاسیک مفاهیمی مانند تحلیل رگرسیون، توزیع استاندارد، انحراف استاندارد، واریانس، تحلیل خوشه، و فاصله‌های اطمینان را که همه این موارد برای مطالعه داده و ارتباط بین

---

[۲۳] Domain knowledge
[۲٤] Data mining engine
[۲۵] Module
[۲٦] Characterization
[۲۷] Association
[۲۸] Classification
[۲۹] Cluster analysis
[۳۰] Evolution and deviation analysis
[۳۱] Pattern evaluation module
[۳۲] Interesting measures
[۳۳] Graphical user interface (GUI)
[۳٤] Query
[۳۵] Query processing
[۳٦] Classic statistics

داده‌ها می‌باشد، را در بر می‌گیرد. مطمئناً تحلیل آماری کلاسیک نقش اساسی در تکنیک‌های داده‌کاوی ایفا می‌کند.

دومین خانواده‌ای که داده‌کاوی به آن تعلق دارد هوش مصنوعی[۳۷] است. هوش مصنوعی که بر پایه روش‌های ابتکاری می‌باشد و با آمار ضدیت دارد. هوش مصنوعی تلاش دارد تا فرایندی مانند تفکر انسان را برای حل مسائل آماری بکار بندد. چون این رویکرد نیاز به توان محاسباتی بالایی دارد، تا اوایل دهه ۱۹۸۰ عملی نشد. هوش مصنوعی کاربردهای کمی را در حوزه‌های علمی و حکومتی پیدا کرد، اما نیاز به استفاده از کامپیوترهای بزرگ باعث شد همه افراد نتوانند از تکنیک‌های ارائه شده استفاده کنند.

سومین خانواده داده‌کاوی، یادگیری ماشین[۳۸] است، که به مفهوم دقیق‌تر، اجتماع آمار و هوش مصنوعی است. در حالی که هوش مصنوعی نتوانست موفقیت تجاری کسب کند، یادگیری ماشین در بسیاری از موارد جایگزین آن شد. از یادگیری ماشین به عنوان تحول هوش مصنوعی یاد شد، چون مخلوطی از روش‌های ابتکاری هوش مصنوعی به همراه تحلیل آماری پیشرفته می‌باشد. یادگیری ماشین اجازه می‌دهد تا برنامه‌های کامپیوتری در مورد داده‌ای که آن‌ها مطالعه می‌کنند، مانند برنامه‌هایی که تصمیم‌های متفاوتی بر مبنای کیفیت داده مطالعه شده می‌گیرند، یادگیری داشته باشند و برای مفاهیم پایه‌ای آن از آمار استفاده می‌کنند و از الگوریتم‌ها و روش‌های ابتکاری هوش مصنوعی را برای رسیدن به هدف بهره می‌گیرند.

داده‌کاوی در بسیاری از جهات، سازگاری تکنیک‌های یادگیری ماشین با کاربردهای تجاری است. بهترین توصیف از داده‌کاوی به وسیله اجتماع آمار، هوش مصنوعی و یادگیری ماشین بدست می‌آید. این تکنیک‌ها سپس با کمک یکدیگر، برای مطالعه داده و پیدا کردن الگوهای نهفته در آن‌ها استفاده می‌شوند. بعضی از کاربردهای داده‌کاوی به شرح زیر است:

- **کاربردهای معمول تجاری:** از قبیل تحلیل و مدیریت بازار، تحلیل سبد بازار، بازاریابی هدف، فهم رفتار مشتری، تحلیل و مدیریت ریسک؛
- **مدیریت و کشف فریب:** کشف فریب تلفنی، کشف فریب‌های بیمه‌ای و اتومبیل، کشف حقه‌های کارت اعتباری، کشف تراکنش‌های مشکوک مالی (پول‌شویی)؛
- **متن کاوی**[۳۹]: پالایش متن (نامه‌های الکترونیکی، گروه‌های خبری و غیره)؛
- **پزشکی:** کشف ارتباط علامت و بیماری، تحلیل آرایه‌های DNA، تصاویر پزشکی؛
- **ورزش:** آمارهای ورزشی؛
- **وب کاوی**[۴۰]: پیشنهاد صفحات مرتبط، بهبود ماشین‌های جستجوگر یا شخصی سازی حرکت در وب سایت؛

---

[۳۷] Artificial intelligence

[۳۸] Machine learning

[۳۹] Text mining

[۴۰] Web mining

## ۱-٤- داده‌کاوی و انبار داده‌ها [٤١]

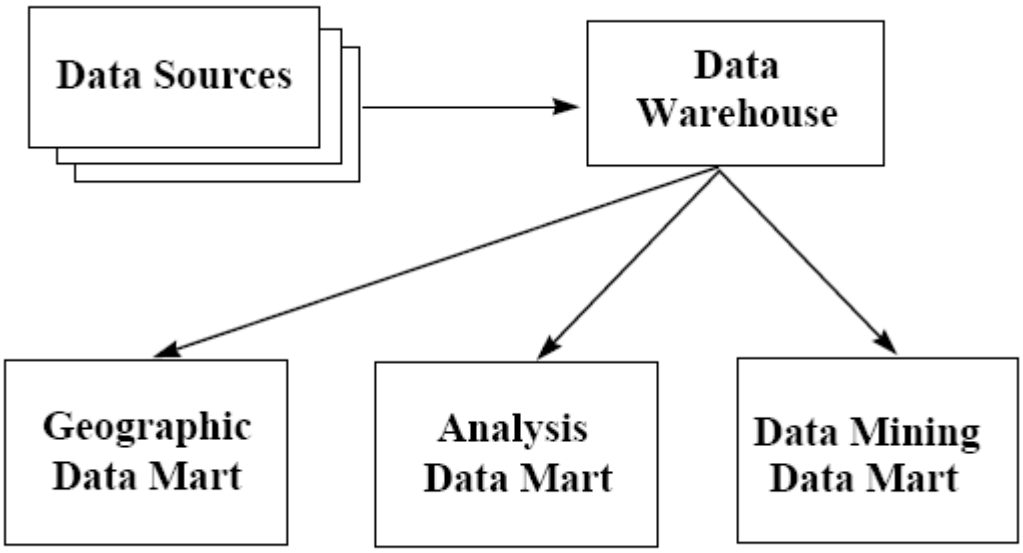
در حال حاضر، داده‌کاوی مهم‌ترین فناوری جهت بهره‌برداری موثر از داده‌های حجیم است و اهمیت آن رو به فزونی است. به طوری که تخمین زده شده است که مقدار داده‌ها در جهان هر ۲۰ ماه به حدود دو برابر می‌رسد. در یک تحقیق که بر روی گروه‌های تجاری بسیار بزرگ در جمع‌آوری داده‌ها صورت گرفت مشخص گردید که ۱۹ درصد از این گروه‌ها دارای پایگاه داده‌هایی با سطح بیش از ۵۰ گیگابایت می‌باشند و ۵۹ درصد از آن‌ها انتظار دارند که در آینده‌ای نزدیک در چنین سطحی قرار گیرند.

معمولاً داده‌هایی که در داده‌کاوی مورد استفاده قرار می‌گیرند از یک انبار داده استخراج می‌شوند، و در یک پایگاه داده[٤٢] یا مرکز داده [٤٣]ای ویژه، برای داده‌کاوی قرار می‌گیرند.

اگر داده‌های انتخابی جزیی از انبار داده‌ها باشند بسیار مفید خواهند بود، چون بسیاری از اعمالی که برای ساختن انباره داده‌ها انجام می‌گیرد با اعمال مقدماتی داده‌کاوی مشترک است و در نتیجه نیاز به انجام مجدد این اعمال وجود ندارد، از جمله این اعمال، پاک‌سازی داده‌ها می‌باشد.

پایگاه داده مربوط به داده‌کاوی می‌تواند جزیی از سیستم انبار داده‌ها باشد و یا می‌تواند یک پایگاه داده جدا را تشکیل دهد.

ولی با این حال وجود انباره داده‌ها برای انجام داده‌کاوی شرط لازم نیست و بدون آن هم اگر داده‌ها در یک یا چندین پایگاه داده باشند می‌توان داده‌کاوی را انجام دهیم و بدین منظور فقط کافی ست داده‌ها را در یک پایگاه داده جمع آوری کنیم و اعمال جامعیت داده‌ها و پاک‌سازی داده‌ها را روی آن انجام دهیم. این پایگاه داده جدید مثل یک مرکز داده‌ای عمل خواهد کرد.

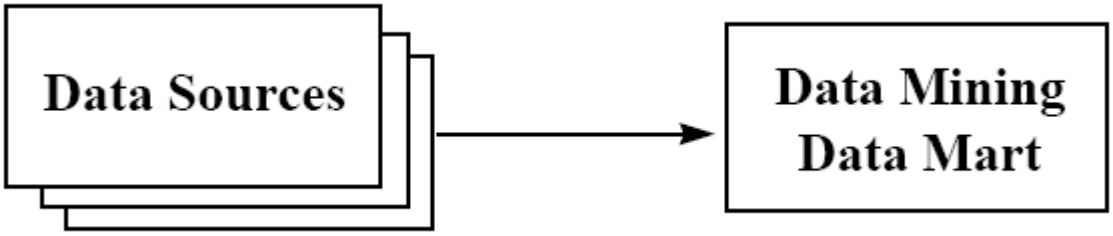



شکل ٤ . داده‌ها از انباره داده‌ها استخراج می‌گردند

[٤١] Data warehouse

[٤٢] Database

[٤٣] Data mart




شکل ۵ .داده‌ها از چند پایگاه داده استخراج شده‌اند

## ۱-۵- کاربرد یادگیری ماشین و آمار در داده‌کاوی

داده‌کاوی از پیشرفت‌هایی که در زمینه هوش مصنوعی و آمار رخ می‌دهد بهره می‌گیرد. هر دو این زمینه‌ها در مسائل شناسایی الگو و طبقه بندی داده‌ها فعالیت می‌کنند و به این واسطه در داده‌کاوی استفاده مستقیم خواهند داشت. و این دو گروه در شناخت و استفاده از شبکه‌های عصبی و درخت‌های تصمیم گیری فعال می‌باشند.

داده‌کاوی جانشین تکنیک‌های آماری سابق نمی‌باشد بلکه وارث آن‌ها بوده و در واقع تغییر و گسترش تکنیک‌های سابق برای متناسب سازی آن‌ها با حجم داده‌ها و مسائل امروزی می‌باشد. تکنیک‌های کلاسیک برای داده‌های محدود و مسائل ساده مناسب بوده‌اند حال آنکه با پیچیده شدن مسائل و رشد روزافزون داده‌ها نیاز به تغییر آن‌ها کاملاً طبیعی است.به عبارت دیگر داده‌کاوی ترکیب تکنیک‌های کلاسیک با الگوریتم‌های جدید مثل شبکه‌های عصبی و درخت تصمیم گیری می‌باشد.

مهم‌ترین نکته این است که داده‌کاوی راهکاری است برای مسائل تجاری امروز به کمک تکنیک‌های آماری و هوش مصنوعی برای افراد حرفه‌ای که قصد دارند یک مدل پیش بینی ایجاد کنند.

## ۱-۶- جایگاه داده‌کاوی

با اختراع کامپیوتر و حجم انبوه اطلاعات و داده‌های مختلف و افزایش دانش بشر به خصوص در زمینه‌ی ریاضیات، مشخص کننده‌ی کمبودی در نظم دهی و آرایش مرتب اطلاعات انبوه، برای سهولت به کارگیری داده‌ها در استخراج دانش نهفته شده در انبوه اطلاعات موجود شدند.

برای رسیدن به این منظور نیاز به علمی که بتواند توانایی ما را در پیش بینی عملکرد اطلاعات بالا ببرد کم کم احساس می‌شود، علم آمار در این زمینه به ما کمک می‌کند.

علمی که توانست از ۳ دانش، ریاضی، آمار، و کامپیوتر، بیشترین بهره‌وری را ایجاد کند، علم داده‌کاوی است. داده‌کاوی با کاربرد ابزارهای متفاوت و مناسب با شرایط موجود توانست از حجم انبوه اطلاعات ذخیره شده در کامپیوتر و معادلات قطعی و محکم ریاضیات، و احتمالات آماری در زمینه‌های مختلف به ما کمک کند.

پژوهشگران با استفاده از کامپیوتر حجم انبوهی از اطلاعات را جمع آوری می‌کنند وبا استفاده از دانش ریاضیات فرمول‌های لازم را کسب کرده، در همین زمان نیاز به وجود ابزاری برای انتخاب داده‌های مناسب شرایط، دسته بندی آن‌ها ونیز رسیدن به شناخت لازم و کافی وهم چنین ارائه‌ی مدلی مناسب احساس می‌شود.

این ابزارها در علم داده‌کاوی موجودند و خدمات ارزنده‌ای به پژوهشگران ارائه می‌کنند، از جمله کاهش زمان و هزینه‌ها. این علم در خدمت صاحبان صنایع و بنگاه‌ها نیز بوده و صرفه جویی در زمان و هزینه‌ها را

برای آن‌ها ممکن می‌کنند. ارائه‌ی مقالات داده‌کاوی در طول سال‌های ۲۰۱۱-۲۰۰۰ رشد صعودی و قابل ملاحظه‌ای داشته است. اما با شناخت بیشتر فواید و کارایی ابزار مورد استفاده در این علم و هم چنین نقش عمده‌ی آن در دسته بندی داده‌ها و ارائه‌ی داده‌های مورد نیاز از میان انبوه داده رشد این گونه مقالات از سال ۲۰۱۱ تا کنون کاملاً صعودی بوده است. در این میان قابل ذکر است، درآمد شغل داده‌کاوی نیز در این مهم بی تأثیر نبوده است و داده کاوان از نظر مالی به خوبی تأمین شده‌اند، چنان که درآمد این شغل به طور میانگین ۲۱% بیشتر از سایر شغل‌ها است، و از سال ۲۰۰٦ نیز روند صعودی داشته است. از سال ۲۰۱۰ این نرخ صعودی با روند افزایشی همراه بوده است و رشد درآمدی قابل توجهی را در این فرجه‌ی زمانی نشان می‌دهد. این روند افزایش درآمدی در مورد همه‌ی شغل‌ها صدق نمی‌کند، چنانچه شغل‌های مربوط به فناوری اطلاعات که از شغل‌های پرکاربرد و جدید است رشد درآمدی به نسبت کمتر، و حتی در فرجه‌ی زمانی ۲۰۰۹-۲۰۰۸ رشد منفی داشته است.

درآمد این شغل‌ها ٤% زیر میانگین درآمد سایر شغل‌ها است، در صورتی که سختی کار قابل ملاحظه در این شغل‌ها نهفته است، اما متأسفانه این شغل‌ها از نظر درآمدی نسبت به داده کاوان که درآمد آن‌ها ۲۵% بیشتر از میانگین درآمدی شغل‌های مربوط به فناوری اطلاعات است، تأمین مالی می‌شوند.

داده‌کاوی علمی است که با استفاده از ابزارهای مختلف به پژوهشگران در شناخت مشکلات و ارائه‌ی راه حل‌های آن کمک می‌کند. این علم شامل بخش‌هایی از علوم ریاضی، کامپیوتر، آمار است. ابزارهای آماری در داده‌کاوی استفاده می‌شوند. اصلی‌ترین این ابزارها، خوشه‌بندی، دسته‌بندی، و مدل‌بندی است. داده کاوان با استفاده از این ابزارها به شناخت مشکلات در علوم مختلف و ارائه‌ی راه حل برای آن‌ها اقدام می‌کنند.

## ۱-۷- درخت داده‌کاوی

استفاده از کامپیوتر در مدت زمان کوتاه سبب شد تا پژوهشگران با اطلاعات متنوع و حجیمی روبرو شوند.

پژوهشگران با وجود اطلاعات متنوع و حجیم، توانایی انجام تحلیل مناسب مسائل را بر اساس اطلاعات موجود نداشتند.

برای بررسی موضوعات مختلف زمان و هزینه زیادی صرف می‌شد، که این مسأله خود موجب کاهش سرعت تحقیقات و دور شدن از نتیجه‌ی مطلوب با گذر زمان می‌شد.

به نحوی شاید بتوان گفت اطلاعات در زمان مناسب در اختیار کاربران قرار نمی‌گرفت و همین مسأله نتایج کار تحقیقاتی را از ارزش به هنگام بودن و در موقع نیاز در دسترس بودن خارج می‌کرد.

پژوهشگران سعی کردند با کاهش حجم اطلاعات موجود در کامپیوتر در یک حد مطلوب و مورد نیاز، و استفاده از معادلات و استدلال‌های محکم علم ریاضیات این مشکل را حل کنند.

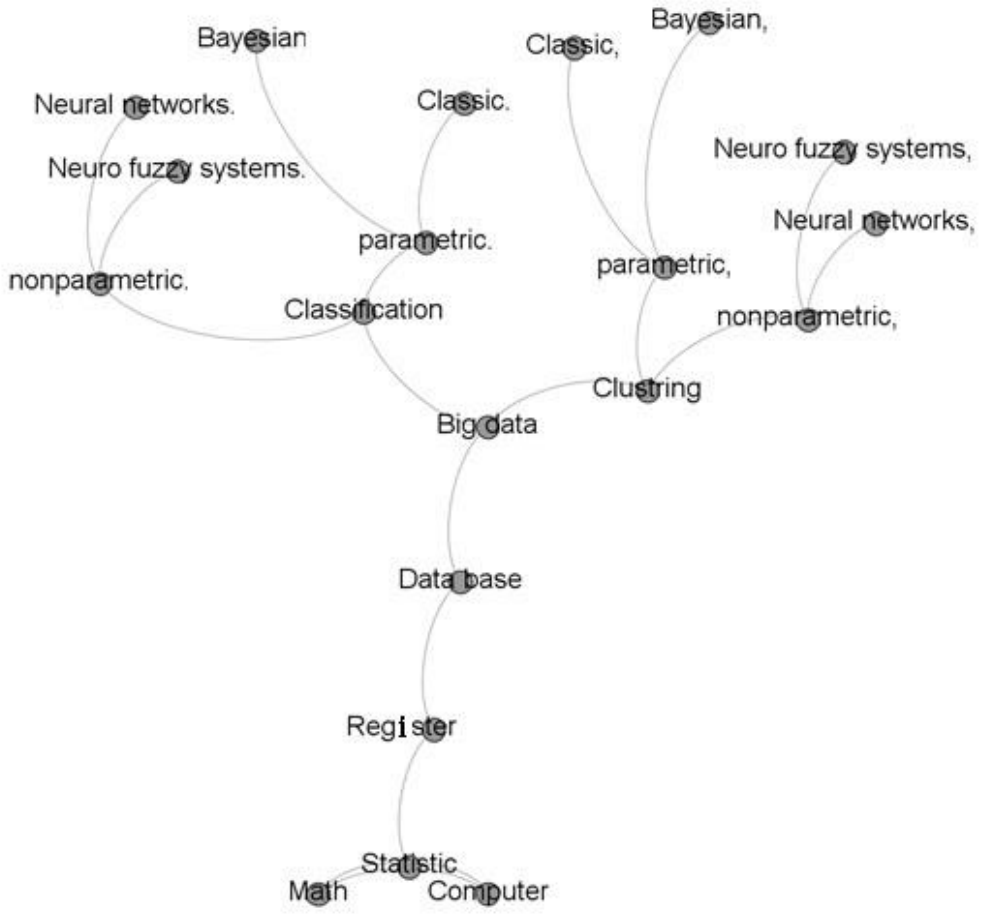
در عمل نتوانستند به موفقیتی برسند، زیرا در بسیاری از موارد نیاز به احتمالات و پیش بینی‌ها به شدت احساس می‌شد، و مشکل زمان بری حل این مسأله‌ها به عنوان چالشی بزرگ هنوز پابرجا بود.

پژوهشگران با استفاده از علم آمار و قوانین آن که شامل احتمال، پیش بینی، و مدل بندی است، در حل مسائل ذکر شده توانستند مشکل را تا حد زیادی حل کنند. علم کامپیوتر نیز در کاهش زمان تحلیل مسائل به کاربران کمک زیادی کرد. ۳ علم ریاضی، کامپیوتر، و آمار پایه‌های علم داده‌کاوی را تشکیل دادند.

باید توجه داشته باشیم که در گذشته نیز از علم داده‌کاوی استفاده‌های بسیاری شده است، البته نه با نام صرف داده‌کاوی، بلکه پژوهشگران مسائلی را با استفاده از علوم ریاضی، کامپیوتر، و آمار پیش بینی می‌کرده اند، اما خود نمی‌دانستند که این ۳ علم در آینده پایه‌های علم جدیدی به نام داده‌کاوی خواهند بود.

به منظور توصیف علم داده‌کاوی، می‌توان آن را به صورت درختی تصور کرد، که علوم ریاضی، کامپیوتر، و آمار ریشه‌های این درخت هستند و در واقع پایه‌های علم داده‌کاوی را تشکیل می‌دهند. منابع داده‌ای مختلف مانند Statistical Registers، Adminstrative data bases و Big data، می‌توانند داده‌های مورد نیاز داده کاوان را در اختیار آن‌ها قرار دهند. در واقع این ۳ پایگاه داده، تنه‌ی درخت داده‌کاوی را تشکیل می‌دهند. داده کاوان به منظور انتخاب داده‌های مناسب از هر یک از پایگاه‌های داده می‌توانند، از یکی از ۲ روش clustring ، classification استفاده کنند.

آن‌ها با استفاده از روش‌های دسته بندی و خوشه بندی، داده‌ها را در دسته‌ها و خوشه‌هایی با ویژگی‌های تقریباً یکسان مرتب می‌کنند. این کار در انتخاب داده‌های مناسب و کاهش زمان مورد نیاز برای حل مسئله مؤثر است.



داده کاوان برای تحلیل داده‌ها از دو روش parametric، nonparametric، استفاده می‌کنند. آن‌ها در صورت وجود فرضیه از روشparametric استفاده کرده، و در غیر این صورت از روش nonparametric استفاده خواهند کرد. داده کاوان به منظور انجام تحلیل در روش parametric، ابزارهای مورد نیازشان را از بین ۲ خانواده‌ی bayesian،classic انتخاب می‌کنند.

اگر آن‌ها برای تحلیل داده‌ها از روش nonparametric استفاده کنند، ابزارهای مورد نیاز خود را از بین ۲ خانواده‌ی شبکه عصبی و شبکه‌ی فازی انتخاب می‌کنند.

داده کاوان با انجام تمامی این مراحل، دانش نهفته را استخراج کرده، و به مدل، طرح، الگو، و یا استراتژی مناسب، در حل مسأله می‌رسند.

## ۱-۸- اسرار موفقیت در داده‌کاوی

برای رسیدن به موفقیت باید برنامه ریزی داشته باشیم، داده‌کاوی از این قاعده مستثنی نیست. نتیجه‌ی خوب در داده‌کاوی با تنظیم درست اهداف میسر می‌شود. در اولین قدم باید تیمی متشکل از پژوهشگران در عرصه‌های مختلف علوم داشته باشیم.

در گام نخست باید از صحت داده‌ها اطمینان حاصل کنیم و پایه‌های فناوری اطلاعات مطمئن داشته باشیم. استفاده از راه حل صحیح داده‌کاوی در این کار بسیار مؤثر است. در این راه می‌توان از داده‌کاوی سایر داده‌ها نیز استفاده کرد. با گسترش حوزه‌ی داده‌کاوی می‌توان به نتیجه‌ی مطلوب‌تر رسید و از تمامی راه‌ها باید حداکثر استفاده را کرد و مدیریت مدل‌ها را افزایش داد.

## ۱-۹- معرفی شرکت‌های پیشرو در داده‌کاوی

در سال ۱۹۹۵ شرکت‌های GTE، JPL ، NASA، AT&T، به عنوان پیشرو در زمینه‌ی داده‌کاوی شناخته شدند. سال ۲۰۰۰ شرکت‌های SAS، Microsoft Research، SALFORD System، به آن‌ها اضافه شد. با پیشرفت تکنولوژی و استفاده‌ی گسترده‌ی مردم از کامپیوتر و اینترنت شرکت‌هایHP، Google، Microsoft، Yahoo، Oracle، HP، SPSS به آن‌ها ملحق شدند. در سال ۲۰۱۰ شناخته شده‌ترین شرکت‌ها در زمینه‌ی داده‌کاوی شرکت‌های SAS، Yahoo، Google، IBM Reserch، Microsoft، Facebook بودند.

## ۱-۱۰- معرفی نرم‌افزارهای مطرح در حوزه داده‌کاوی

در این بخش به طور خلاصه به نام برخی از سیستم‌های تجاری داده‌کاوی اشاره‌ای می‌کنیم تا خواننده‌ی این طرح بداند در زمان نگارش آن چه نرم افزارهایی بیشترین کاربرد را داشته‌اند. البته بسیاری از این نرم افزارها هر ساله نسخه‌ها و ویرایش‌های جدیدتری را به بازار عرضه می‌کنند که دارای امکانات بهتری نسبت به نسخه‌های قبلی آن‌ها است. می‌توان این نرم افزارها را با دیدگاه‌های متفاوتی دسته‌بندی نمود. برای مثال بعضی از آن‌ها نرم‌افزارهای آماری شناخته می‌شوند و بعضی دیگر از زاویه‌ی دید پایگاه داده‌ها طراحی شده‌اند.

اما با توجه به اینکه موضوع این طرح بررسی جزئیات نرم‌افزارهای داده‌کاوی نیست لذا قصد نداریم به جزئیات آن‌ها بپردازیم، و این کار را به عهده‌ی خواننده گذاشته تا اگر مایل باشد با رفتن به سایت محصول مورد نظر اطلاعات بیشتری کسب کند.

- نرم‌افزارWEKA یک بسته‌ی نرم‌افزاری منبع باز است که دانشگاهی در نیوزلند آن‌را با جاوا طراحی و پیاده سازی نموده است .این نرم‌افزار شامل مجموعه‌ای از الگوریتم‌های داده‌کاوی مانند طبقه‌بندی، خوشه‌بندی، رگرسیون، عملیاتی جهت پیش پردازش داده‌ها، یافتن وابستگی‌ها و ابزاری برای بصری سازی است.
- کمپانیIBM محصول داده‌کاوی Intelligent Miner را همراه با طیف وسیعی از تکنیک‌های داده‌کاوی شامل طبقه‌بندی، رگرسیون، مدل‌سازی جهت تخمین، خوشه‌بندی، کاوش وابستگی‌ها و تحلیل الگوهای مکرر ارائه کرده است. جعبه ابزارهایی برای الگوریتم‌های شبکه‌های عصبی، روش‌های آماری، ابزار آماده سازی داده‌ها و همچنین بصری سازی داده‌ها را می‌توانید به این نرم‌افزار اضافه کنید. ویژگی شاخص این نرم‌افزار قابلیت مقیاس پذیری الگوریتم‌های داده‌کاوی موجود در آن و نزدیکی آن با سیستم پایگاه داده‌ی رابطه‌ای DB ۲ است.
- نرم‌افزار Enterprise Miner توسط شرکت SAS توسعه داده شده است و همانند دیگر محصولات شامل تکنیک‌های طبقه‌بندی، رگرسیون، خوشه‌بندی، کاوش وابستگی‌ها، تحلیل داده‌های نوع سری‌های زمانی و همچنین بسته‌ی نرم‌افزاری جهت تحلیل‌های آماری است. تنوع ابزارهایی که برای تحلیل‌های آماری در این نرم‌افزار وجود دارد، مشخصه‌ی بارز آن است که البته قدمت شرکت SAS در این حوزه آن‌را باعث شده است.
- SQL Server ۲۰۰۵ از مایکروسافت یک سیستم مدیریت پایگاه داده است که چندین تابع داده‌کاوی را در سیستم پایگاه داده‌ی رابطه‌ای خود و محیط‌های سیستم انبار داده‌ها گنجانده است. کاوش وابستگی‌ها، طبقه‌بندی(درخت تصمیم، بیز و الگوریتم‌های شبکه عصبی)، درختان رگرسیون، خوشه بندی و تحلیل سری‌های زمانی را می‌توان در آن یافت. با توجه به اینکه این بانک اطلاعاتی به عنوان یکی از قوی‌ترین سیستم‌های مدیریت پایگاه داده‌ی رابطه‌ای مخصوصاً با حجم بالای داده شناخته می‌شود، استفاده از آن در داده‌کاوی جهت کار با حجم وسیع داده‌ها می‌تواند انتخاب خوبی باشد.
- شرکت SPSS نرم‌افزار Clementine را برای کاربران آمار و داده‌کاوی تهیه نموده است. در این نرم‌افزار می‌توانید مجموعه‌ای از توابع را برای طبقه‌بندی، خوشه‌بندی، تخمین و کاوش وابستگی‌ها پیدا کنید. یکی از ویژگی‌های مهم این نرم‌افزار واسط کاربری است که به صورت شیء گرا توسعه یافته و به کاربران الگوریتم‌ها اجازه می‌دهد تا به محیط برنامه نویسی بصری اضافه شوند.
- نرم‌افزار MineSet در سال ۱۹۹۹ توسط SGI معرفی شد و شامل کاوش قوانین وابستگی، طبقه‌بندی و همچنین ابزارهای پیشرفته‌ی آماری و قدرتمند بصری سازی است. ویژگی ممیزه‌ی آن ابزار قدرتمند گرافیکی آن است که این نرم‌افزار را به نحوی از دیگر محصولات جدا می‌سازد.
- نرم‌افزار Insightful Miner از شرکت Insightful هم دارای چندین روش داده‌کاوی شامل پالایش داده‌ها، طبقه‌بندی، تخمین، خوشه‌بندی، بسته‌ی تحلیل آماری همراه با ابزار بصری سازی است. ویژگی

خاصی که در آن می‌توان مشاهده کرد، واسط بصری آن است که کاربر را قادر می‌سازد با اتصال اجزاء و بخش‌ها مستند سازی کند.

- نرم‌افزار منبع باز RapidMiner از گروه Rapid-I محصولی است که دارای امکاناتی جهت پیش پردازش داده‌ها و تکنیک‌های متنوع داده‌کاوی از جمله طبقه‌بندی، خوشه‌بندی، تخمین و کاوش وابستگی‌ها است. این نرم‌افزار با مجموعه بسته‌هایی که به آن افزوده می‌شود، می‌تواند با داده‌های حجیم نیز به خوبی عمل کند. این بسته‌ها به قدرت نرم‌افزار می‌افزایند. برای مثال بسته‌ی متن کاوی باعث می‌شود تا کاربر بتواند الگوریتم‌ها را بر روی مجموعه داده‌های نیمه ساخت یافته اجرا کند. در ضمن می‌توان الگوریتم‌های WEKA را وارد این نرم‌افزار نمود و از آن‌ها نیز استفاده کرد.
- بدون شک بسته‌های نرم‌افزاری دیگری را نیز می‌توان یافت که شاید به صورت خاص برای تکنیک‌های مشخص داده‌کاوی پیاده سازی شده‌اند. برای مثال CART نرم‌افزاری است که برای ساخت درختان تصمیم و رگرسیون استفاده می‌شود.