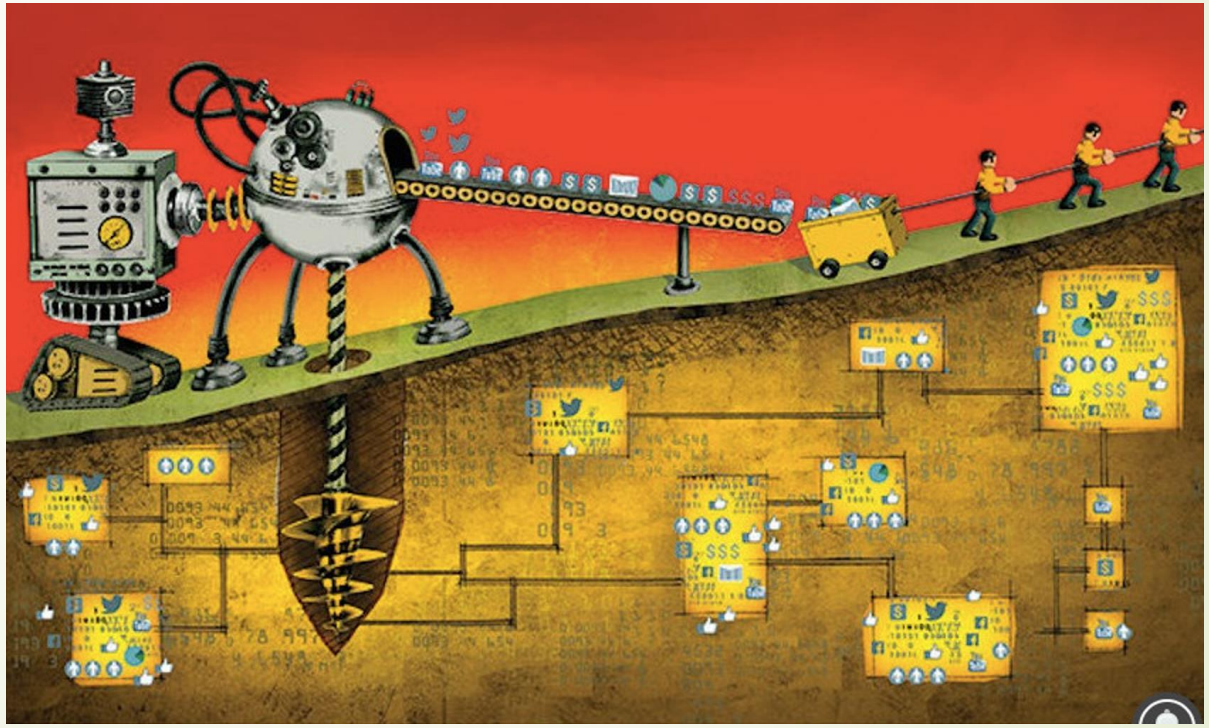
داده کاوی

Data mining

آرزو حبیبی‌راد
گروه آمار
دانشگاه فردوسی مشهد

بخش اول:

معرفی داده کاوی و ارتباط ان با سایر علوم



مقدمه

- چرا داده‌کاوی؟
- مگر در آمار با داده‌ها سرو کار نداریم؟
- چرا برای تحلیل داده ها، داده‌کاوی مطرح می شود؟

- داده کاوی ترجمه‌ی کلمه «Data Mining» بوده که به معنای **«کاویدن معادن داده»** است.
- داده کاوی یعنی استخراج اطلاعات گرانبها از حجم عظیم معادن داده!
- کلمه Mining در معنای تحت الفظی خود یعنی «استخراج از معدن» است و در واقع عبارت **Data Mining** نشان می دهد که حجم انبوه اطلاعات مانند یک معدن عمل می کند و از ظاهر آن مشخص نیست چه عناصر گرانبهایی در

عمق این معدن وجود دارد. تنها با کند و کاو و استخراج این معدن است که می توان به آن عناصر گرانبها دست پیدا کرد.

داده‌کاوی چیست؟

- داده کاوی یک فرآیند محاسباتی است که در واقع الگو یا الگوهایی را در مجموعه از داده‌های عظیم کشف می‌کند. در تمامی تعریف‌های مرتبط به داده کاوی کلمه **کشف کردن** را می‌توان پیدا کرد.
- ✓ داده کاوی شاخه‌ای از علوم کامپیوتر و ریاضی است که در واقع ترکیبی از تکنیک‌های آماری، علوم اطلاعات، یادگیری ماشینی و نظریه پایگاه داده است.
- ✓ در سال ۱۹۸۹ اصطلاح کشف دانش از پایگاه داده **(Knowledge Discovery form Databases)** یا به اختصار **(KDD)** مطرح شد و در سال ۱۹۹۰ اصطلاح داده کاوی به وجود آمد و در همین سال با استفاده از داده کاوی، خرده فروش‌ها و بازارهای مالی به تجزیه تحلیل داده‌ها و پیش بینی نوسانات در نرخ بهره و افزایش مشتری پرداختند.

- در دهه 1930 اولین بار واژه داده کاوی به کار گرفته شده است، و شکوفایی آن همزمان با رشد ساختار پایگاه داده‌ای صورت گرفت.

- داده‌کاوی در پاسخ به نیاز «تحلیل داده های بزرگتر از حد و اندازه معمول برای کشف دانش و الگو » شکل گرفت. جایی که الگوریتم‌ها و روش‌های آماری از لحاظ تئوری پاسخگو نبودند، مفاهیم داده‌کاوی پا به عرصه می‌گذاشتند.

✓ نکته مهم در داده‌کاوی این است که کشف دانش، عموما توسط یک خبره و در واقع یک انسان صورت گرفته و می گیرد.

- حوزه‌ها و محدوده آن شامل همه روش‌های داده محور تحلیلی می‌باشد و زمانی از داده‌کاوی استفاده می‌شود که مسئله‌ای بروز پیدا کرده و کاربر و خبره قصد حل آن را داشته باشد. **بنابراین داده کاوی«مسئله محور» است.**

چرا داده‌کاوی؟

- با رشد و افزایش توجهات به داده‌کاوی، پرسش «چرا داده‌کاوی؟» همواره مطرح می‌شود. در پاسخ به این پرسش باید گفت، داده‌کاوی دارای کاربردهای زیادی است. بدین ترتیب، زمینه‌ای جوان و آینده‌دار برای نسل کنونی محسوب می‌شود.

- این زمینه توانسته توجهات زیادی را به صنایع و جوامع اطلاعاتی جلب کند. با وجود گستره وسیع داده ها، نیاز حتمی به تبدیل چنین داده‌هایی به اطلاعات و دانش وجود دارد.

- بنابراین، بشر از اطلاعات و دانش برای گستره وسیعی از کاربردها، از تحلیل بازار گرفته تا تشخیص بیماری‌ها، کشف کلاهبرداری و و پیش بینی قیمت سهام استفاده می‌کند.
- در مجموع باید گفت، ضرب‌المثل انگلیسی « نیاز مادر همه ابداعات بشر است»، پاسخی کوتاه و گویا به پرسش مطرح شده است.

داده‌کاوی و علم آمار

- **داده کاوی شاخه ی توسعه یافته و پیچیده ی علم آمار است**. علم آمار به تنهایی برای صاحبان صنایع و شرکت‌ها بسیار اهمیت دارد.
- تفاوت اصلی علم آمار با شاخه ی توسعه یافته‌ی خود یعنی داده کاوی، در **حجم داده‌های مورد تحلیل** و **روش مدل‌سازی داده‌هاست**. در بیشتر موارد حجم داده های اولیه‌ی مورد استفاده در داده کاوی آنقدرزیاد است که به یک چالش زمان‌بر و پرهزینه تبدیل می‌شود. اما مدل‌سازی داده‌های ورودی و دستیابی به اطلاعات پنهان و ارزشمند موجود در این حجم عظیم داده، با کمک هوش مصنوعی و شیوه‌های خودکار یادگیری انجام می‌شود.
- در داده‌کاوی به دنبال ارتباطهای پنهان و کشف نشده هستیم به طوری که **داده کاوی پل ارتباطی میان علم آمار، علم کامپیوتر، هوش مصنوعی، یادگیری ماشین داده می باشد**.
- وظیفه ی **داده کاوی**، کاویدن و استخراج دانش از منابع عظیم داده است تا اطلاعات گرانبهایی که در حجم انبوهی از اطلاعات سطحی پنهان شده است را آشکار سازد.

ارتباط داده‌کاوی با یادگیری ماشین، هوش مصنوعی و علم داده

❖ **هوش مصنوعی:**

- ✓ هوش مصنوعی منظور دنیایی است که در آن ربات‌ها بر انسان‌ها چیره می‌شوند و حیات و ممات نسل بشر درمعرض تهدید قرار می‌گیرد.
- ✓ اما در علم کامپیوتر، هوش مصنوعی استفاده از الگوریتم‌هایی است که بتوانند رفتار انسانی را در قالب کدهای صفر و یک ریاضی درآورده و رفتار مناسب با آن رفتار صورت گرفته را ارائه دهند.
- ✓ حال این شبیه سازی رفتار انسانی می تواند در قالب الگوریتم و منطق ریاضی **«اگر آنگاه»**، یا در قالب اینکه خود ماشین و ربات یاد بگیرد، باشد.
- ✓ **هر چه که هست، هوش مصنوعی به ربات‌ها قابلیت تصمیم گیری می‌دهد** و در واقع اینجا خبره به جای استفاده از الگوریتم‌ها برای حل یک مسئله (که در داده کاوی مورد بحث است)، یک بار برای همیشه کلیه حالت‌هایی که به ذهن‌اش می‌رسد را در قالب قواعد و منطق به ماشین تعریف می کند یا کاری می کند که ماشین خودش یاد بگیرد.

یادگیری ماشین:

- **یادگیری ماشین بخشی از هوش مصنوعی است.**

- بخشی که ماشین ها به جای تطبیق دهی میان قواعد ازپیش تعریف شده، خود یاد می گیرند و برای حالت های جدیدی که مشاهده نشده، تصمیم می گیرند. با این نگاه، یادگیری ماشین بخش اصلی هوش مصنوعی است و هر روز این اهمیت بیشتر می شود چرا که دیگر نمی‌توان تمام قواعدِ رفتارهای پیچیده انسانی را کشف کرد و به خورد ربات داد بلکه می توان با داده های زیادی که این روزها تولید می شود، کاری کرد که ربات و ماشین خود یاد بگیرد.

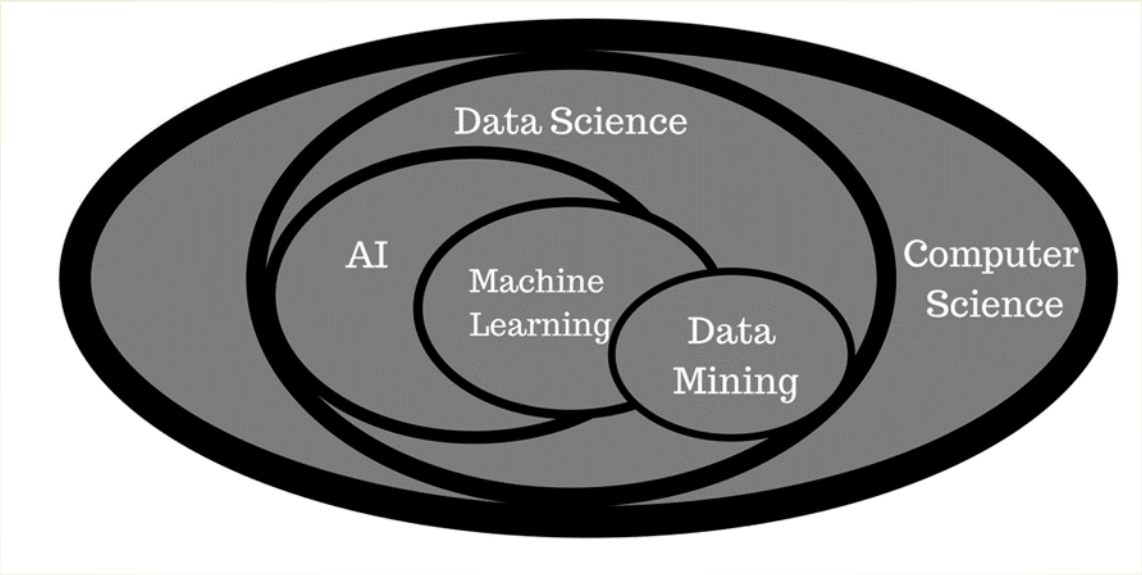
- با این بحث ها، می توان گفت **یادگیری ماشین و هوش مصنوعی قرین هم بوده و بسیار شبیه به هم هستند.**

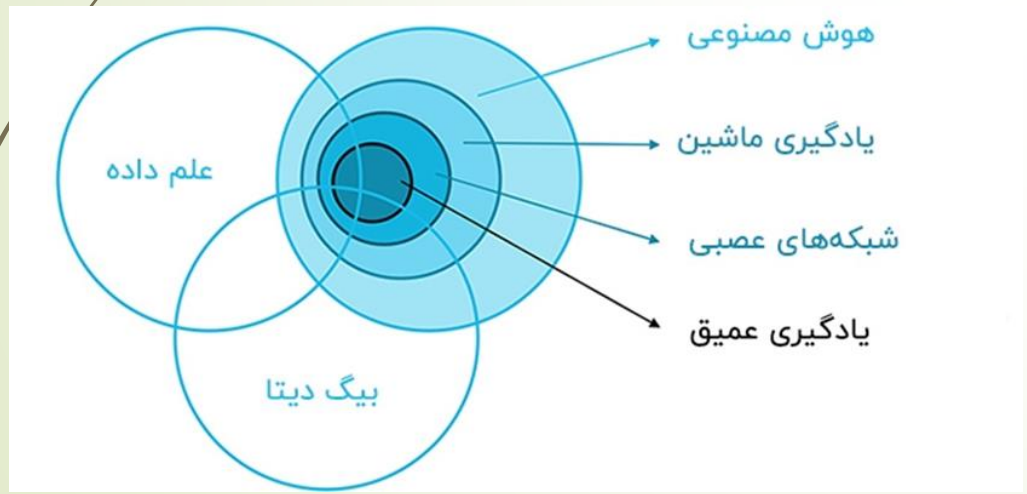
معرفی علم داده:

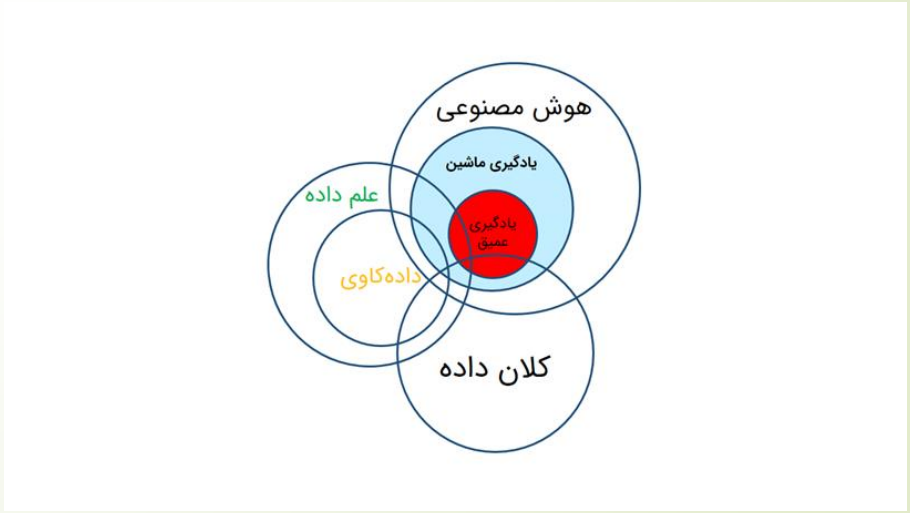
- علم از هر نوع آن به **بررسی و کشف روابط موجود در آن زمینه** می پردازد. برای مثال، علم فیزیک به کشف قواعد دنیای ماده می پردازد. همین طور علم ریاضی به ارائه تعریف‌های گوناگون از منظرهای گوناگون برای پدیده‌های طبیعی از نظر عددی (و نه لزوما داده) می‌پردازد.

- علم داده، نیز فارغ از اینکه کاربردها آن در چه جاهایی مصداق پیدا می‌کند، به دنبال شناخت ماهیت خود داده، الگوریتم‌های موجود و ایجاد مفاهیم جدید برای کارایی بیشتر داده در تئوری و عمل است.

- لذا هر جا قرار بر این باشد که الگوریتم جدیدی در حوزه داده ایجاد شود، ترجیح به کار بردن واژه **«علم داده »** است.


Data Science
AI
Machine Learning
Data Mining
Computer Science


هوش مصنوعی
یادگیری ماشین
شبکه‌های عصبی
یادگیری عمیق
علم داده
بیگ دیتا


هوش مصنوعی
یادگیری ماشین
یادگیری عمیق
علم داده
داده‌کاوی
کلان داده

چه مسئله‌ای علت نیاز به داده کاوی شده است؟

- ✓ می‌توان گفت اصلی‌ترین دلیل توجه به این دانش در اختیار داشتن **حجم وسیعی از داده ها موجود است،** که این نیاز به وجود آمده بود که باید از این داده‌ها اطلاعات و دانش مفیدی استخراج کرد و از آن برای **کنترل، تحلیل** و **پیش‌بینی** استفاده کرد.
- ✓ بنابراین دلایل کلی را می‌توان در سه مورد خلاصه کرد:

1. حجم داده‌ها با سرعت زیادی در حال رشد است.
2. اطلاعات ما در مورد این داده ها کم است.
3. دانش ما نسبت به این اطلاعات صفر است.

- ✓ اما پاسخ تمامی این مشکلات تنها و تنها داده کاوی است.
- ✓ **با داده‌کاوی می‌توان حجم عظیم داده‌ها را اصطلاحاً کاوید و دانش را از دل آن بیرون کشید.**

❖ **ویژگی‌های اصلی داده‌کاوی:**

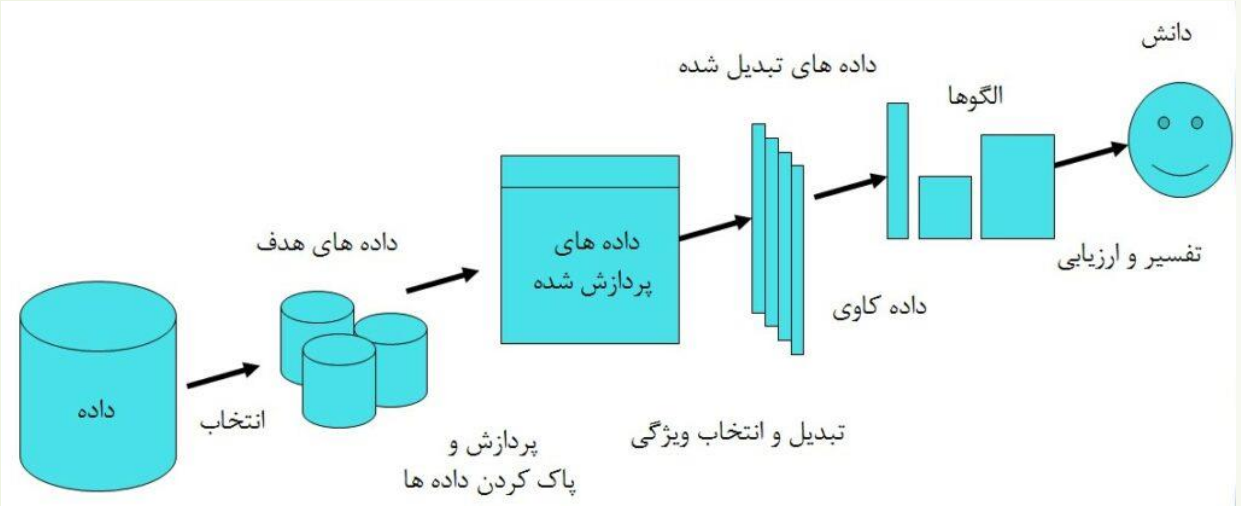
- کشف اتوماتیک الگوها در انجام پروژه‌های داده کاوی پیش بینی احتمالی نتایج وخروجی ها
- تولید اطلاعات اجرایی و مفید متمرکز بر داده‌های بزرگ.

❖ **مزایای داده‌کاوی:**

- ایجاد روابط بصورت اتوماتیک
- استفاده از داده‌های متنوع
- دینامیک بودن
- عدم نیاز به داده‌های صحیح
- ساخت مدل‌های واقعی
- تحلیل کردن داده‌های واقعی
- دوری از اشکالات احتمالی نمونه‌گیری

مراحل داده‌کاوی:

1. پالایش داده‌ها
2. یکپارچه‌سازی داده‌ها
3. انتخاب داده‌ها
4. تبدیل داده‌ها
5. داده‌کاوی
6. ارزیابی و تفسیر الگوها
7. ارائه دانش



چهار گام اصلی داده‌کاوی:

- **گام صفر: تعریف مسئله شناخت داده‌ها**

قبل از شروع فرآیند داده‌کاوی به تعریف دقیق مساله اختصاص پرداخته می‌شود و روش‌های به کارگیری داده‌ها جهت پاسخگویی به مسئله مدنظر قرار می‌گیرد. در این گام باید به خوبی داده‌ها را شناخت.

- **گام اول: آماده‌سازی داده‌ها (پالایش داده‌ها)**

در این گام، تحلیل‌گر ممکن است رو ش‌های مختلف داده‌کاوی را بر روی داده‌های آماده شده امتحان کند تا بتواند به هدف اصلی پروژه برسد.

ساخت مدل یک فرآیند خطی نیست و رفت وبرگشت‌های زیادی وجود دارد. یک مدل بهینه در داده کاوی وجود ندارد و بسته به مسئله‌ای که تحلیل‌گر با آن مواجه است، روش‌های مختلف باید آزمایش شوند و خروجی آن‌ها باهم مقایسه گردند. در این مرحله احتمالاً لازم است به گام قبلی بازگشت و برای برخی از الگوریتم‌ها داده‌ها را به شکل دیگری آماده کرد.

- **گام سوم : یادگیری مدل**

بسته به نیاز کسب وکار، داده‌کاوی ممکن است با هدف **پیش‌بینی** (Prediction) **پیدا کردن روابط** (Association)و یا برای **خوشه بندی** (Clustering) استفاده گردد. درهریک از این دسته‌ها الگوریتم‌های متفاوتی وجود دارند که بسته به شرایط یکی ازآن‌ها یا ترکیبی از آنان استفاده می‌شوند.

- **گام سوم: ارزیابی مدل**

در این فرآیند، الگوهای حاصل شده در گام قبل، از جنبه‌های گوناگونی شامل **دقت**، **صحت** و **قابلیت تعمیم** و دیگر موارد مورد ارزیابی قرار می‌گیرند. مدل‌های ساخته شده تست و ارزیابی می‌شوند و بهترین مدل از نظر مسئله‌ی طرح شده در این مرحله، انتخاب می‌شود. سپس در تبادل نظر با کارفرما، موثر بودن مدل انتخاب شده بررسی می‌شود. در صورتی که مدل انتخاب شده کمکی در حل مسئله نمی کند کل فرآیند از مرحله ی یک دوباره انجام می‌شود.

گام چهارم: تفسیر مدل

- تفسیر نتیجه در این مرحله نتایج و الگوهای ارائه شده توسط ابزار داده‌کاوی مورد بررسی قرار گرفته و نتایج مفید معین می‌شود.

- طرز کار ابزار داده کاوی این گونه است که ابزار به دنبال اثبات این است که وجود چیزی به معنای وجود چیز دیگری است و سعی می کند در درجه اول از توالی ارتباطات برای کشف یک الگو بهره بگیرد و در نهایت اطلاعات بدست آمده را دسته بندی کند تا به الگوی خاصی برسد که بتواند آن را براساس فاکتورهای داخلی به مخاطبش ارائه دهد.

در داده‌کاوی چه داده‌هایی کاوش می‌شوند؟

- در داده‌کاوی داده‌های ثبت شده در **پایگاه داده‌ها یا انبار داده‌ها** مورد تحلیل و تفسیر قرار می گیرند.

- **پایگاه داده‌ها:**

✓ مجموعه ای از داده های ثبت شده توسط سیستم‌های کاربردی در مؤسسات، سازمان‌ها، مراکز تجاری، صنعتی، ... را پایگاه داده می گویند.

✓ داده‌های یک سیستم پایگاه داده شامل مجموعه‌ای از **داده‌های مرتبط** و **نرم‌افزارهایی جهت دستیابی به این داده‌ها** می‌باشد.

✓ پایگاه داده ها توسط متخصصین علم کامپیوتر نوشته و آماده می‌شوند.

✓ **مثال‌هایی از پایگاه داده:** ثبت اطلاعات افراد توسط کارمندان یک فروشگاه زنجیره‌ای.

انبار داده‌ها:

- ✓ مخزنی از داده ها هستند که از چندین پایگاه داده جمع آوری شده اند و به صورت یک شمای یک دست و کلی‌اند و غالباً در سایت ذخیره می‌شوند، را (Data warehousing) گفته نمی‌شود.
- ✓ با کمک انبار داده ها می توان دانش یا اطلاعات جامع تری کسب نمود.
- ✓ در این انبار، اطلاعاتی ذخیره می شود که یک سازمان برای تصمیم گیری های راهبردی و کلان به آن نیاز دارد.
- ✓ به فرایند ساخت و استفاده از انبار داده ها **انبارش داده ها** می گویند.
- ✓ پایگاه داده ها با کمک انبارش داده تبدیل به یک منبع مفید برای تحلیل داده ها می شود.
- ✓ امروزه موفق‌ترین شرکت‌ها آن‌هایی هستند که بتوانند به سرعت و با انعطاف بالا به تغییرات و فرصت‌ها پاسخ دهند. کلید پاسخ به این مسئله کاربرد مؤثر و کارآمد داده‌ها و اطلاعات از سوی تحلیل‌گران و مدیران است.
  یک انبار داده مخزنی از داده‌های تاریخی است که بر اساس موضوع سازمان یافته‌اند تا به تصمیم گیران سازمان کمک کند. زمانی که داده ها در یک **داده‌گاه** یا **انبار داده** ذخیره می شوند، می توان آن ها را مورد ارزیابی قرار داد.

- به عبارتی انبار داده ها مجموعه ای از **داده های موضوع –گرا**، **یکپارچه شده**، **زمان-متغیر** و **غیر فرّار** هستند.

- **موضوع-محور:** داده‌ها در انبار داده بر خلاف سیستم های عملیاتی پیرامون موضوعات سازمان گرد هم می آیند. جهت گیری موضوعی برای تصمیم گیری واقعا مفید است. گردآوری اشیای مورد نیاز، موضوع-محوری نامیده می شود.
- **یکپارچه شده:** داده‌های موجود درون یک انبار داده یکپارچه هستند. از آنجا که این داده‌ها از چند سیستم عملیاتی می‌آیند همه ناسازگاری‌ها می‌بایست حذف شوند.

خصوصیات آن شامل قراردادهای نام گذاری، اندازه‌گیری متغیرها، ساختار داده، خصوصیات فیزیکی داده و مواردی از این دست است.

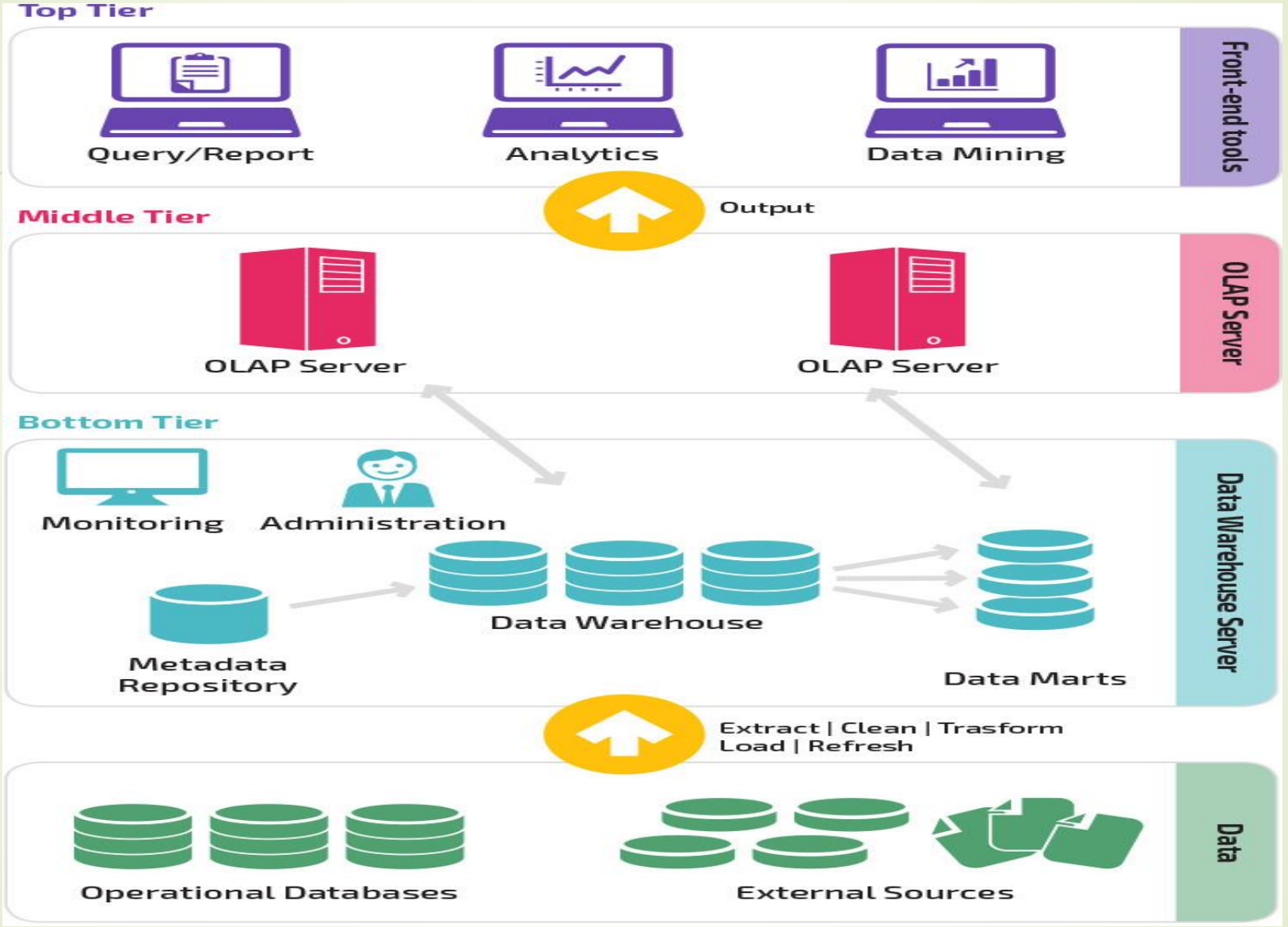
- **زمان-متغیر:** با این که سیستم‌های عملیاتی به دلیل پشتیبانی از عملیات‌های هر روزه مقادیر فعلی را نشان می‌دهند؛ اما انبارهای داده، نشان دهنده داده‌هایی در افق زمانی بلندمدت‌تر هستند. این بدان معنی است که انبار داده حاوی داده‌های تاریخی است. انبار داده به طور عمده برای داده‌کاوی و پیش‌بینی استفاده می‌شود، چون اگر کاربری در جستجوی یک الگوی خرید برای یک مشتری خاص باشد، می‌بایست به داده‌هایی در مورد خریدهای فعلی و گذشته نگاه کند.
- **غیر فرّار:** داده‌های موجود در انبار داده تنها خواندنی هستند، یعنی نمی‌توان آن‌ها را به روزرسانی، ایجاد یا حذف کرد.

✓ **در انبارش داده ها با دو سیستم OLAP و OLTP سروکار داریم.**

- واژه OLTP مخفف عبارت On Line Transition Procedure
- واژه OLAP مخفف عبارت On Line Analytically Procedure
- سیستم OLTP یک سیستم مشتری‌گرا است، کاربران این سیستم معمولا کارمندان سازمان هستند.
- سیستم OLAP یک سیستم بازارگرا است، کاربران این سیستم معمولا متخصصین ومدیران سطح بالای سازمان هستند.

✓ **طراحی انبارش داده و لایه های مختلف معماری انبارش داده:**

انبارش داده ها ساخت یک معماری چند ردیفی بر روی داده ها می باشد. غالبا از معماری سه ردیفی استفاده می شود.

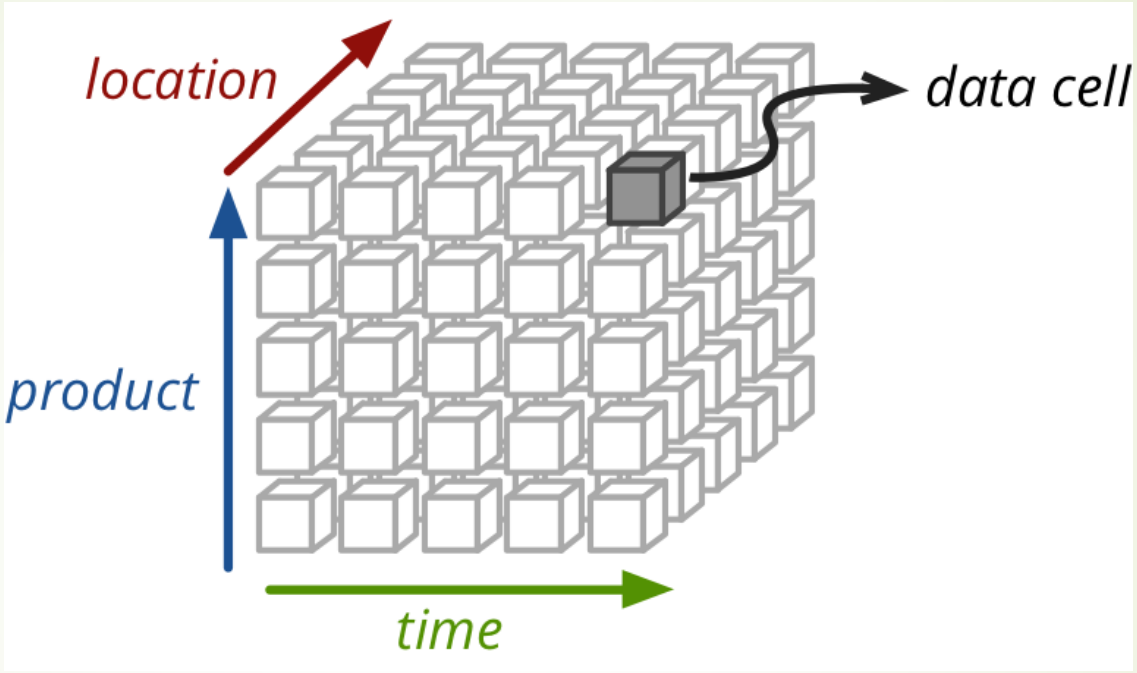

Top Tier
Query/Report
Analytics
Data Mining
Front-end tools
Output
Middle Tier
OLAP Server
OLAP Server
OLAP Server
Bottom Tier
Monitoring
Administration
Data Warehouse
Metadata Repository
Data Marts
Data Warehouse Server
Extract | Clean | Trasform
Load | Refresh
Operational Databases
External Sources
Data

مکعب داده‌ها (Data Cube)

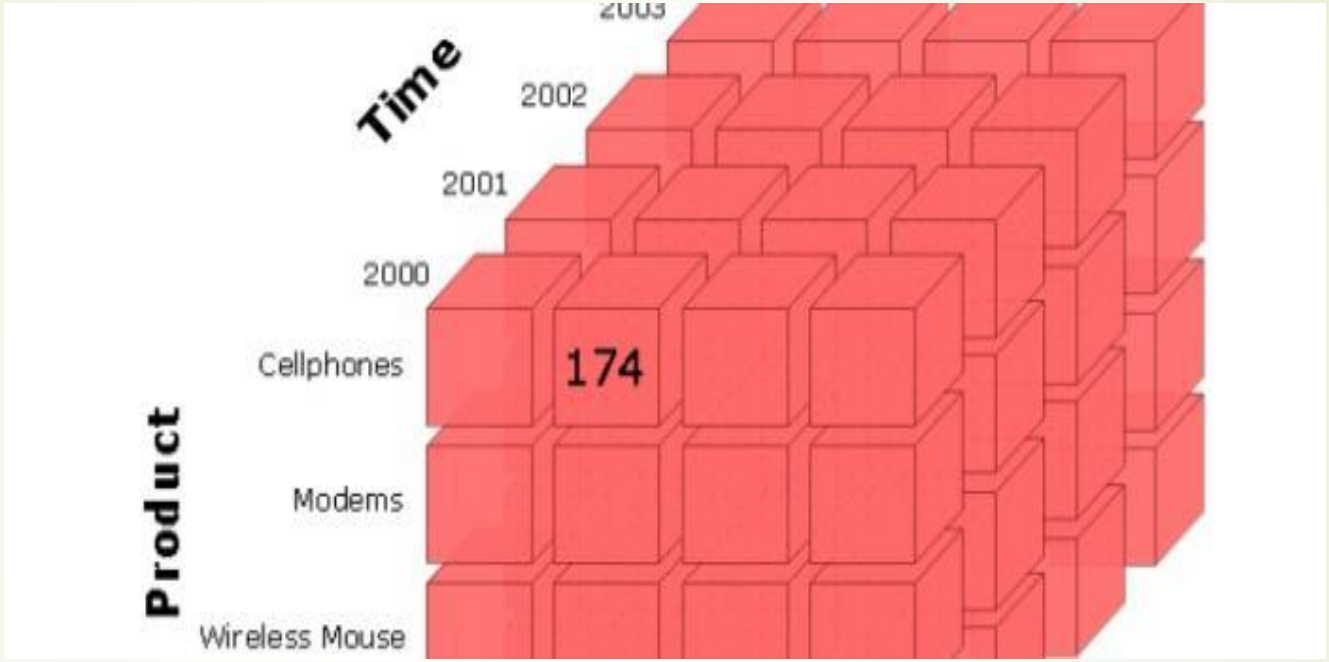
- انبار داده‌ها بر اساس یک مدل داده‌ای چند بعدی به نام مکعب داده نمایش داده می شود.
- مکعب داده، نمایش یا الگو چند بعدی از داده‌ها است به طوری که این امکان را فراهم می‌کند تا داده‌ها در چند بعد نمایش داده شوند.
- **مکعب داده چند بعدی(MOLAP):** اکثر محصولات OLAP بر اساس ساختاری طراحی شده‌اند که داده‌ها را به شکل آرایه‌های مرتب شده براساس ابعاد داده ذخیره می‌کنند.

- MOLAP در مقایسه با رویکردهای دیگر و به دلیل ساختار خاص نگهداری اطلاعات، از سرعت بالایی برخوردار است و کارایی را بهبود می‌بخشد. وقتی که تعداد ابعاد بزرگتر است، مکعب کوچکتر می شود.

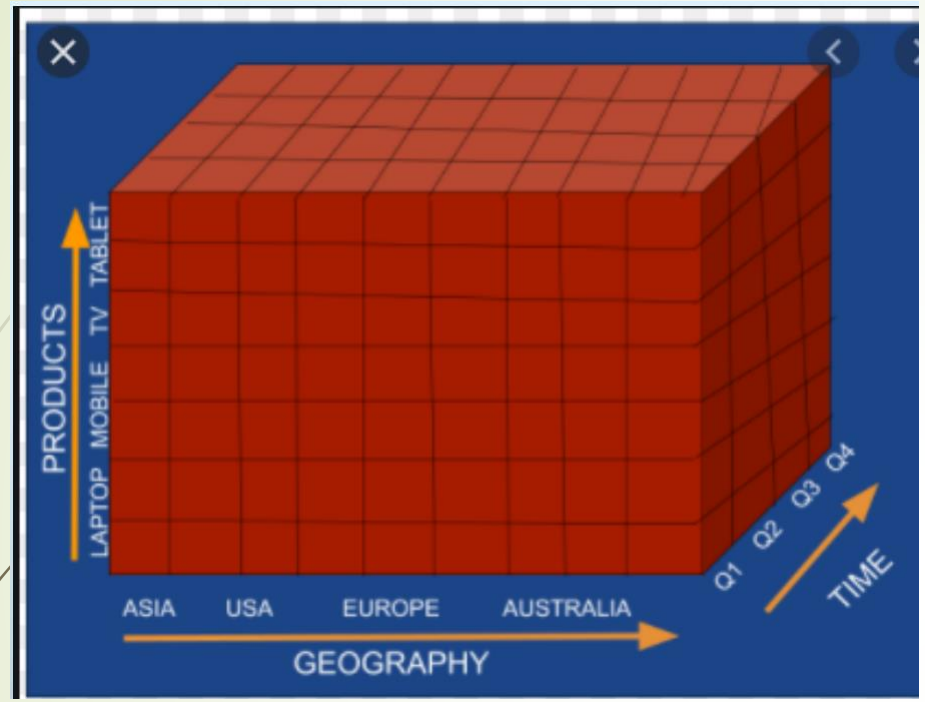
- **مکعب داده رابطه ای (ROLAP):** OLAP ارتباطی یا ROLAP، از مدل پایگاه داده رابطه‌ای استفاده می‌کند و عملیات OLAP را در قالب استفاده از مدل Relational ساماندهی می‌کند و لذا از سیستم‌های MOLAP کندتر عمل می‌کنند و به فضای بیشتری نیز نیاز دارند.

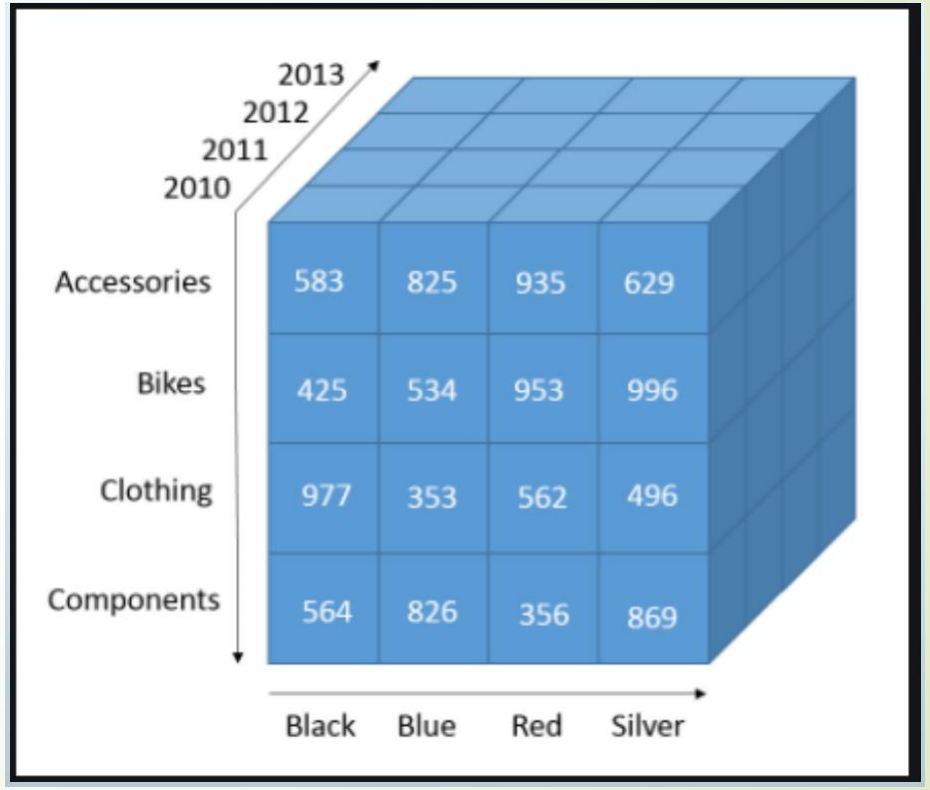


- یک انبار داده معمولا با کمک یک ساختار چند بُعدی با نام مکعب داده مدل سازی می‌شود. به عبارتی یک مکعب داده، یک دید چند بُعدی را به کاربر ارائه می‌کند و با پیش محاسبه‌ی آن می توان دسترسی سریعی به داده های خلاصه شده داشت.



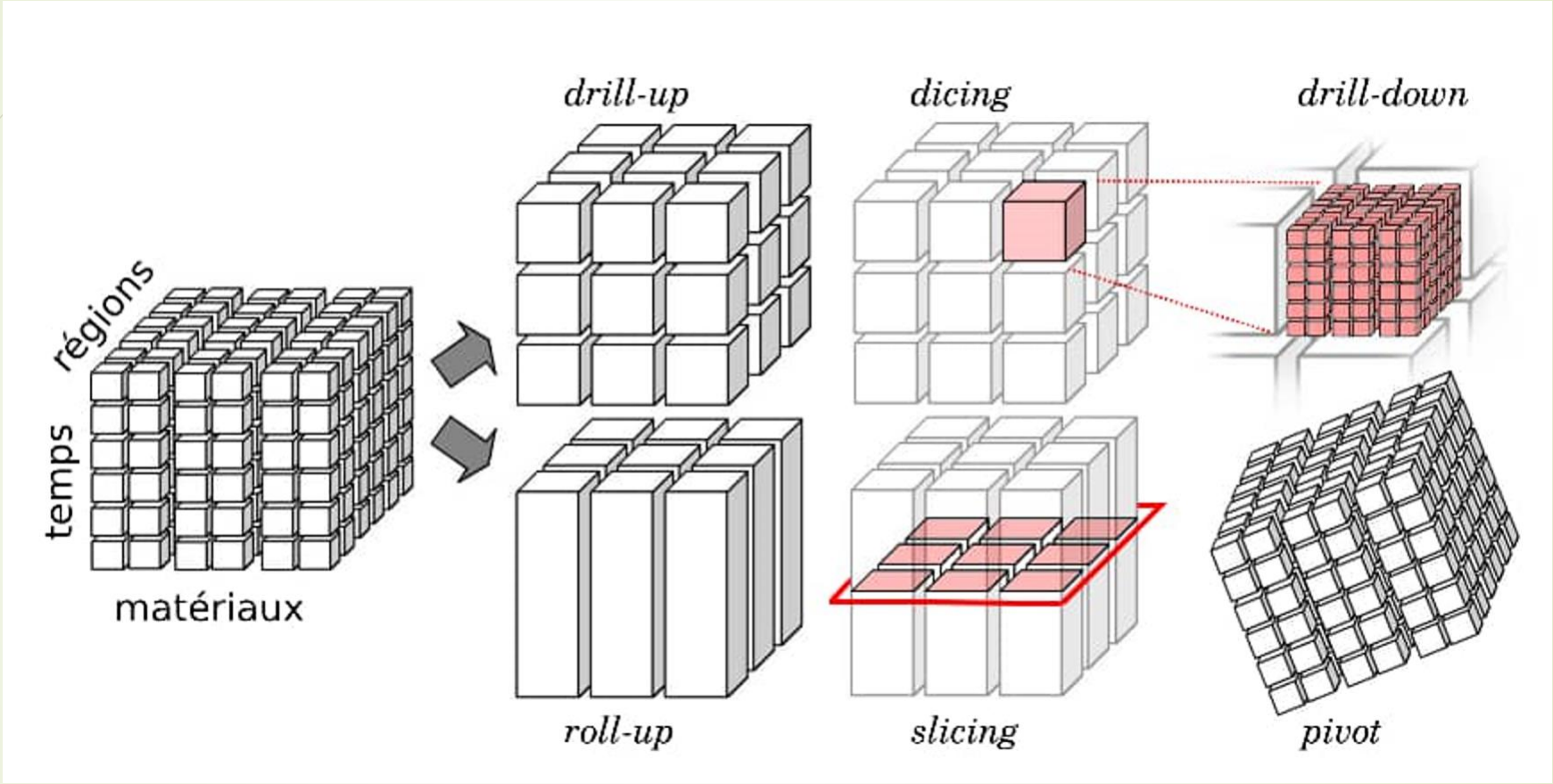
- یک انبار داده معمولا با کمک یک ساختار چند بُعدی با نام مکعب داده مدل سازی می شود. به عبارتی یک مکعب داده، یک دید چند بُعدی را به کاربر ارائه می کند و با پیش محاسبه ی آن می توان دسترسی سریعی به داده های خلاصه شده داشت.

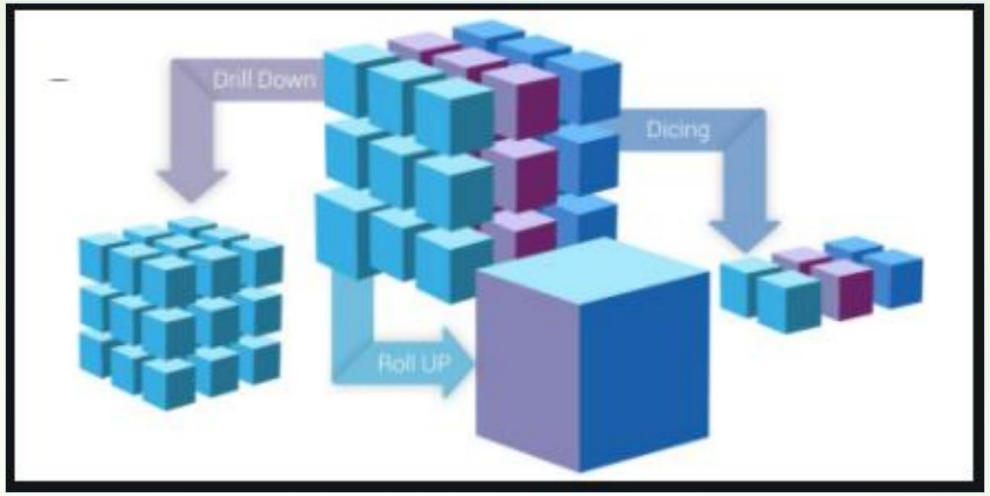

PRODUCTS
LAPTOP
MOBILE
TV
TABLET
ASIA
USA
EUROPE
AUSTRALIA
GEOGRAPHY
Q1
Q2
Q3
Q4
TIME

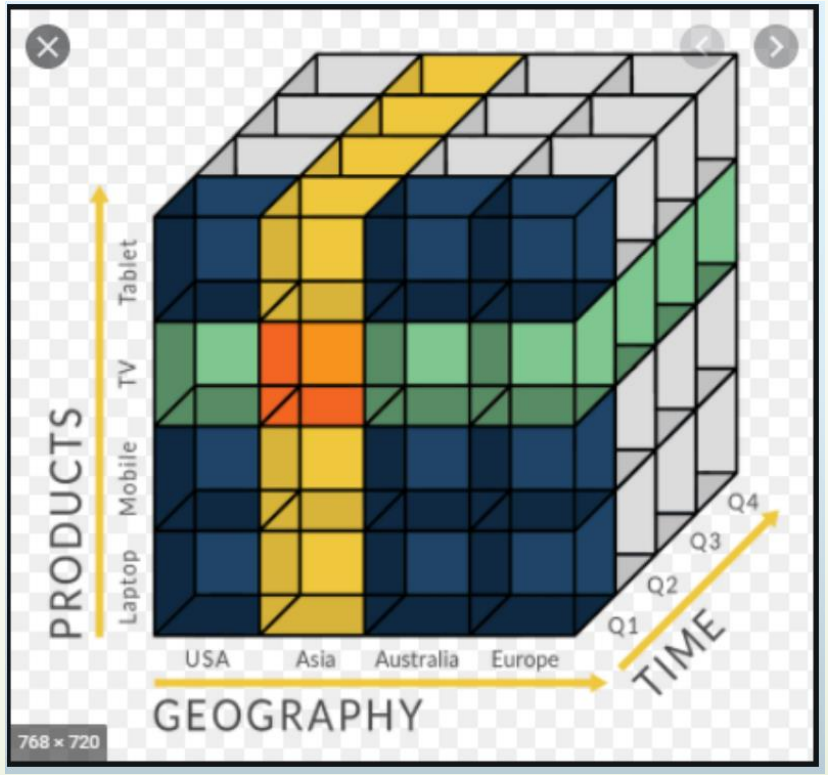

2013
2012
2011
2010
Accessories
Bikes
Clothing
Components
583
825
935
629
425
534
953
996
977
353
562
496
564
826
356
869
Black
Blue
Red
Silver

❖ **پنج عمل اصلی زیر را می‌توان بر روی مکعب داده‌ها انجام داد:**

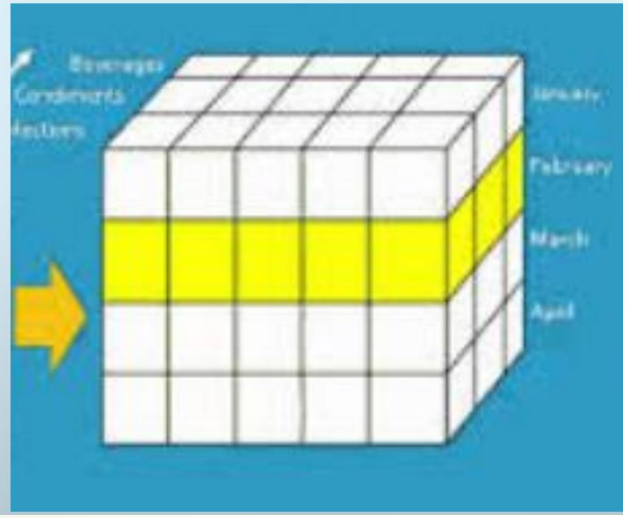
1. ROLL-UP: با کاهش تعداد بعدها ساختار جدیدی را در مکعب‌ها ایجاد می‌کند.
2. DRILL-DOWN: برعکس عمل ROLL-UP است و اطلاعات را به صورت مفصل‌تری نشان می‌دهد. برای مثال زمانی که به جای اطلاعات فروش فصلی، اطلاعات فروش ماهانه مورد نیاز باشد از این عمل استفاده می‌شود.
3. SLICE: انتخابی را بر روی یک بعد انجام می‌دهد.
4. DICE: انتخاب را روی چند بعد انجام می‌دهد به عبارتی مکعب‌ها را به صورت جدا جدا بیرون می‌آورد.
5. PIVOT: با چرخش مکعب از یک بعد دیگر می‌توان به آن نگاه کرد (جای محورها را می‌توان عوض کرد).


drill-up
dicing
drill-down
régions
temps
matériaux
roll-up
slicing
pivot

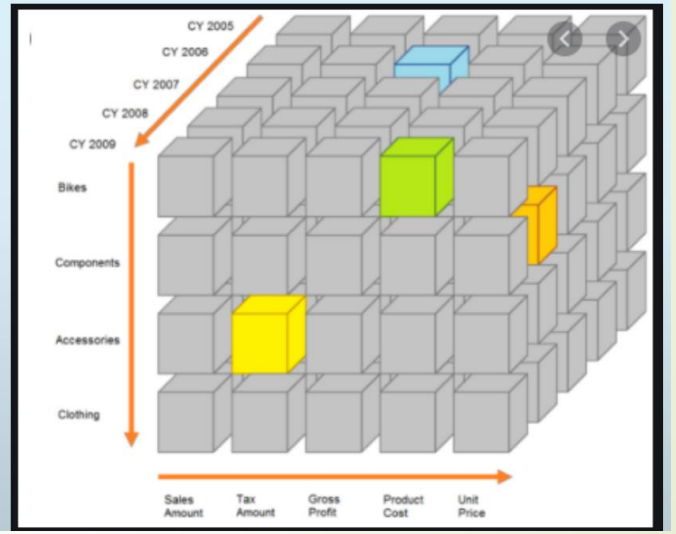

Drill Down
Dicing
Roll UP


PRODUCTS
Laptop
Mobile
TV
Tablet
USA
Asia
Australia
Europe
GEOGRAPHY
Q1
Q2
Q3
Q4
TIME
768 × 720

- Slice


February
March
April

- Dice


CY 2005
CY 2006
CY 2007
CY 2008
CY 2009
Bikes
Components
Accessories
Clothing
Sales Amount
Tax Amount
Gross Profit
Product Cost
Unit Price

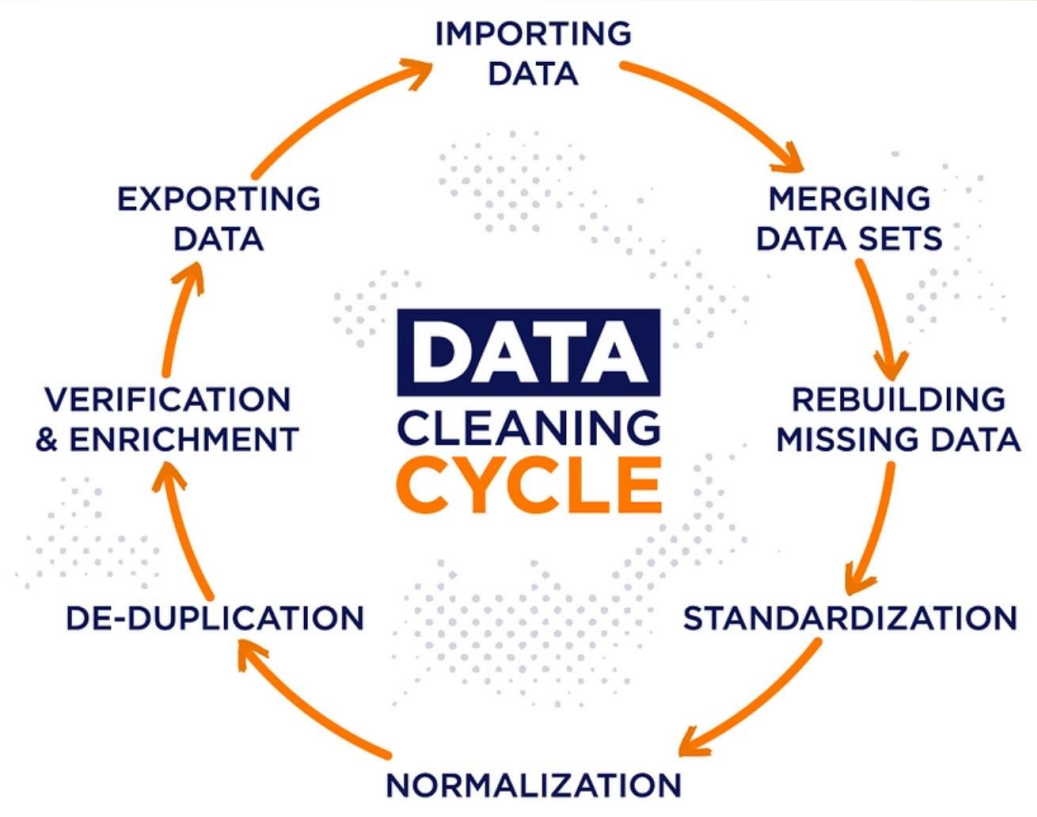
بخش دوم:

آماده سازی و پیش پردازش داده ها

در این مرحله هدف آماده کردن داده‌ها برای قسمت تحلیلی با استفاده از روش های داده کاوی است و معمولا بیشترین زمانی که در انجام پروژه داده کاوی صرف می شود در این مرحله انجام می شود. به عبارتی ۸۰% زمان یک پروژه مرتبط به آماده سازی داده است و دلیلش هم آن است که در دنیای واقعی داده ها آنطوری نیستند که باید باشند! وجود نمادهای نامربوط، داده های پرت، گمشده، ناسازگاری و ... دلیلی برای زمان زیادی است که در این مرحله هدر می رود و زمان هایی نیز پیش می آید که داده ها به صورت الکترونیکی ذخیره نشدند و برگردان این اطلاعات به صورت الکترونیکی کاری بسیار طاقت فرسا و زمان‌بر است.

آماده سازی داده‌ها وقت زیادی را در داده کاوی به خود اختصاص می‌دهد.

آماده سازی داده‌ها برای بهبود کیفیت داده و افزایش دقت تحلیل داده‌ها ضروری می باشد.



آماده‌سازی داده‌ها شامل سه مرحله به شرح زیر است:

1. **شناخت انواع داده‌ها**
2. **پیش‌پردازش داده‌ها**
3. **پویش داده‌ها**

- این مرحله از داده‌کاوی به تجمیع و پاکسازی داده‌هایی اختصاص می‌یابد که در مرحله تعریف مسئله شناسایی شده‌اند.
- ممکن است داده‌ها در سراسر سازمان توزیع شده و در قالب‌های مختلفی ذخیره شوند و یا ممکن است شامل تناقضات و ناسازگاری‌هایی از جمله ورودی‌های نادرست یا گمشده باشد.
- فرآیند پاکسازی داده‌ها تنها به حذف داده‌های نامناسب و یا وارد کردن مقادیر از دست رفته خلاصه نمی‌شود بلکه کشف روابط پنهان شده میان داده‌ها، شناسایی دقیق‌تر منابع داده و تعیین مناسب‌ترین متغیرها برای استفاده در مرحله تحلیل را دربرمی‌گیرد.

❖ داده‌ها:

- ✓ هر چیزی که قابل ثبت شدن باشد داده است.
- ✓ داده می تواند عددی باشد یا غیر عدد باشد.
- ✓ به هر رقم، عدد، کاراکتر، رشته، سیگنال، .... و به طور کلی به هرآنچه که قابل ثبت شدن باشد داده می گویند.

❖ ویژگی داده‌ها:

- ✓ داده های ثبت شده باید با یک مقیاس مناسب اندازه گیری شودند.
- ✓ منظور از اندازه گیری نسبت دادن یک عدد حقیقی به یک ویژگی می‌باشد که این امر طبق یک قائده ی مشخص صورت گیرد.

❖ **مقیاس‌های اندازه‌گیری داده‌ها:**

1. **مقیاس اسمی:** تنها برای تمیز میان اسم افراد به کار می‌رود و نمی‌توان آن‌ها را برای مقایسه 4 عمل اصلی به کار برد.
2. **مقیاس ترتیبی:** داده‌ها خاصیت ترتیبی دارند اما قابلیت انجام عملیات حسابی وجود ندارد.
3. **مقیاس فاصله‌ای**: داده‌ها خاصیت ترتیبی دارند و قابلیت انجام عملیات حسابی به طور محدود وجود دارد. در مقیاس فاصله‌ای صفر، صفر حقیقی نیست.
4. **مقیاس نسبتی:** داده‌ها خاصیت ترتیبی دارند و قابلیت انجام عملیات حسابی به طور کامل وجود دارد.

❖ **متغیرها و ویژگی‌های آن‌ها:**

هر رکورد ثبت شده از داده ها، شامل چند ستون داده است که به این ستون ها فیلد یا متغیر می گویند.

❖ **ویژگی متغیرها:**

- متغیر کمی یا کیفی
- متغیر گسسته یا پیوسته
- متغیر مستغل یا وابسته

پیش‌پردازش داده‌ها:

- داده‌ها زمانی که نیاز های خواسته شده را برآورده سازند دارای کیفیت هستند. عوامل متعددی وجود دارند که با کمک آن‌ها کیفیت داده‌ها توصیف می‌شوند. مانند صحت و درستی، کامل بودن، بدون نویز بودن، قابل تفسیربودن و ...

در دنیای واقعی با توجه به حجم زیاد داده و اینکه داده‌ها از منابع مختلفی جمع‌آوری می‌شوند. این داده‌ها مستعد نویز، ناسازگاری و ناموجود بودن هستند. برای بهبود نتایج کاوش در قسمت اماده سازی داده ها به پیش‌پردازش داده‌ها می‌پردازیم.

روش‌های پیش‌پردازش داده‌ها در 4 گروه زیر سازماندهی می‌شوند:

1. پالایش داده‌ها
2. یکپارچه‌سازی داده‌ها
3. کاهش داده‌ها
4. تبدیل داده‌ها

پالایش داده‌ها:

- قبل از شروع مدل‌سازی و تحلیل داده ها انجام پالایش داده ها ضروری است تا داده‌های مشکل‌دار، غیر مفید، .... مشخص و در مورد آن‌ها تصمیم گیری صورت گیرد.

- **مشکلات موجود در داده‌ها و راه‌های برخورد با آن‌ها:**

  - **داده‌های گمشده:** داده هایی که به دلایل عمدی یا غیرعمدی ثبت نشده اند. در داده کاوی به دلیل حجم زیاد داده، احتمال برخورد با داده های گمشده زیاد است.

- **در صورت مواجهه با داده‌های گمشده می‌توان از روش‌های زیر استفاده کرد:**
- ✓ حذف داده های گمشده
- ✓ جایگزینی مقدار مناسب
- ✓ پیش بینی برای خانه های خالی

- **داده‌های پرت:**

داده‌هایی که نسبت به اکثر داده‌ها دچار انحراف یا تمایز زیادی هستند یا به عبارت دیگر خارج از محدوده منطقی داده‌ها قرار گرفتند.

- **راه‌های تشخیص داده‌های پرت:**
- ✓ روش خوشه بندی داده ها
- ✓ رسم نمودارهای پراکنش یا هیستوگرام

- **روش برخورد با داده‌های پرت:**
- ممکن است با مشورت مدیر پروژه حضور داده های پرت در فایل داده ها لازم باشد.
- در غیر این صورت مشابه داده‌های گمشده برخورد میکنیم.

داده‌های ناسازگار:

- هرگونه تخریب در مقدار داده که به صورت عمدی یا غیرعمدی صورت پذیرد و منجر به این شود که ما به اصل داده دسترسی نداشته باشیم نویز گفته می شود.

- داده هایی که در صورت عمل نویز اتفاق بیافتند داده‌های نویزدار یا ناسازگار می گویند.
- شناخت این نوع داده ها بسیار مشکل و زمان بر هست.
- این مشکل درهنگام یکپارچه سازی پایگاه داده‌ها خیلی رخ می دهد.

➢ روش برخورد با داده‌های ناسازگار:

- **الف) روش هموار سازی یا بسته بندی (Binning):**

در این روش پس از تشخیص متغیر نویز دار، ابتدا مشاهدات مرتب می‌شوند سپس به درون بسته های k تایی قرار می گیرند هر بسته با کمک مقداری براساس میانگین بسته، میانه بسته، کرانه های بسته، ... هموار می‌شود.

❖ روش رگرسیون:

در صورت برخورد با هر نوع اغتشاش در داده ها استفاده از مدل های رگرسیونی برای جایگزینی بهترین راه است. در این روش متغیرهایی که پیش‌بینی‌کننده‌های خوبی برای متغیر وابسته هستند را انتخاب کنیم و معادله رگرسیون را بدست آوریم. سپس با تعیین مقادیر اولیه برای نقاط پرت با استفاده از خط رگرسیون دوباره رگرسیون را تکرار می‌کنیم تا مرحله‌ای که تعداد نقاط مخشوش پیدا شود و با نظر فرد خبره تعیین شود.

یکپارچه‌سازی داده‌ها:

- در پروژه‌های داده کاوی ممکن است داده ها از منابع مختلف گردآوری شوند لذا لازم است قبل از تحلیل پایگاه‌های داده یکپارچه شوند که برای این منظور باید عملیات یکسان‌سازی ساختار داده‌ها صورت گیرد.

- **نکات لازم برای یکپارچه‌سازی داده‌ها**

✓ شناسایی موجودیت داده‌ها (متا داده): شامل جواب سوالاتی درباره داده است که باعث سهولت در فهم متغیرها و چگونگی استفاده از داده می‌شود.

- نسبت متا داده به داده را می‌توان مانند نسبت شناسنامه به فرد در نظر گرفت.

✓ افزونگی و تحلیل همبستگی:

- افزونگی: اگر دو ستون در پایگاه داده‌ها معرف یک موجودیت باشد با افزونگی مواجه هستیم. ممکن است یک متغیر مانند درآمد سالانه افزونه تلقی شود و اگر بتوانیم با کمک متغیرهای دیگر آن را بدست آوریم. برخی افز.نگی‌ها را می‌توان با کمک تحلیل همبستگی تشخیص داد.

- ✓ **متغیرهای تکراری:** رکوردهایی هستند که بار اطلاعاتی جدیدی ندارند و اصلاعات تکراری زیادی در آن‌ها وجود دارد.
- در داده‌کاوی به دنبال این هستیم که رکوردهای تکراری را حذف نموده و رکوردهایی ایجاد کنیم که بار اطلاعاتی زیادی دارند. اثر مثبت این کار این است که هم حجم داده‌ها کم می‌شود و هم اطلاعات اضافی حذف می‌شود.

- ✓ تشخیص تضاد در اندازه گیری مقادیر داده‌ها: ممکن است مقادیر متغیر مربوط به یک موجودیت در دنیای واقعی در منابع مختلف به صورت متفاوتی نگهداری شده باشد. این تفوات می‌تواند ناشی از نمایش مقدار یا مقیاس‌گذاری آن‌ها باشد.

کاهش داده‌ها:

- تحلیل بر روی داده‌های حجیم زمان‌بر است. به این منظور با افزایش سرعت در تحلیل داده‌ها می‌توان با کمک تکنیک‌هایی داده‌ها را کاهش داد که این کاهش می‌تواند در بعد داده‌ها، کاهش رکوردها و یا فشرده‌سازی باشد.

روش‌های کاهش داده‌ها:

- **روش های نمونه گیری**
- **رسم نمودار های هیستوگرام**
- **حذف متغییر های غیر ضروری**
- **روش های خوشه بندی**
- **تحلیل عاملی**
- **تجمیع داده ها**

تبدیل داده‌ها:

- داده‌ها به گونه‌ای تبدیل می‌شوند که فرآیند کاوش ممکن است کارآمدتر اجرا شوند و الگوهای پیدا شده به راحتی قابل فهم باشد.

- **روش‌های تبدیل دادها:**
- ✓ **ساخت یک متغیر جدید**
- ✓ **گسسته سازی داده ها ی پیوسته**
- ✓ **نرمال سازی داده ها**
- ✓ **هموارسازی**
- ✓ **تجمیع داده‌ها**

ساخت یک متغیر جدید:

- در این روش با ترکیب متغیرها، متغیر جدیدی می‌سازیم که برای کاوش مورد نیاز است.
- **روش نرمال‌سازی داده‌ها:**

با نرمال داده ها می توان به گونه ای مناسب داده ها را ر بازه های کوچکتری تبدیل نمود که از جمله آن ها می توان به سه روش زیر اشاره کرد.

- **روش نرمال‌سازی min-max:**

فرض کنید در یک نمونه مینیمم مشاهدات را با $x_{(1)}$ و ماکسیمم مشاهدات را با $x_{(n)}$ نمایش دهیم. برای آنکه هر مشاهده x_i به بازه $[a,b]$ تبدیل کنیم، لازم است تبدیل زیر را انجام دهیم.

$$x_i' = \frac{x_i - x_{(1)}}{x_{(n)} - x_{(1)}}(b-a) + a$$

- **روش نرمال‌سازی Z-score:**

اگر $\bar{X}$ میانگین مشاهدات و S انحراف معیار مشاهدات باشد با تبدیل زیر می توان مقدار مشاهده را به مقدار استانداردشده تبدیل کرد.

$$Z_i = \frac{x_i - \bar{x}}{s}$$

- ✓ در صورت وجود داده پرت بهتر است از این روش استفاده شود.

- **روش مقیاس‌گذاری دهدهی:**

در این روش بزرگترین داده به لحاظ قدرمطلق را در نظر می‌گیریم (y)، سپس j را که کوچکترین عددی است به طوری که $\frac{y}{10^j} < 1$ را پیدا کرده و تمام اعداد را بر 10^j تقسیم می‌کنیم که با این کار داده‌ها در بازه $[-1,1]$ قرار می‌گیرند.

- **گسسته سازی داده‌های پیوسته:**

هدف در گسسته سازی داده‌ها این است که داده‌ها با مقیاس‌های فاصله‌ای و نسبتی را به داده‌هایی با مقیاس‌های اسمی یا ترتیبی تبدیل کنیم.

- **تجمیع داده ها:**

گاهی اوقات با تجمیع متغیرها و ساخت متغیر جدید می توان به داده‌هایی با اطلاعات بیشتر دسترسی داشت، که این امر به آماده سازی

داده ها کمک می کند.

✓ مثال: ساخت متغیر توده بدنی (BMI) با کمک متغیر های وزن و قد

$$BMI = \frac{\text{وزن}}{(\text{قد})^2}$$

پویش داده‌ها:

هدف اصلی در پویش داده‌ها این است که پیش از داده‌کاوی با استفاده از روش‌های پویش داده‌ها، اطلاعات ارزشمندی را از داده‌ها استخراج کنیم.

روش‌های پویش داده‌ها عبارتنداز:

1) **شاخص‌های آماری:** شامل شاخص‌های مرکزی و شاخص‌های پراکندگی است
2) **مصورسازی داده‌ها**
3) **پردازش تحلیلی برخط** (OLAP)**:** پیشتر در قسمت معرفی مکعب داده به آن اشاره شد

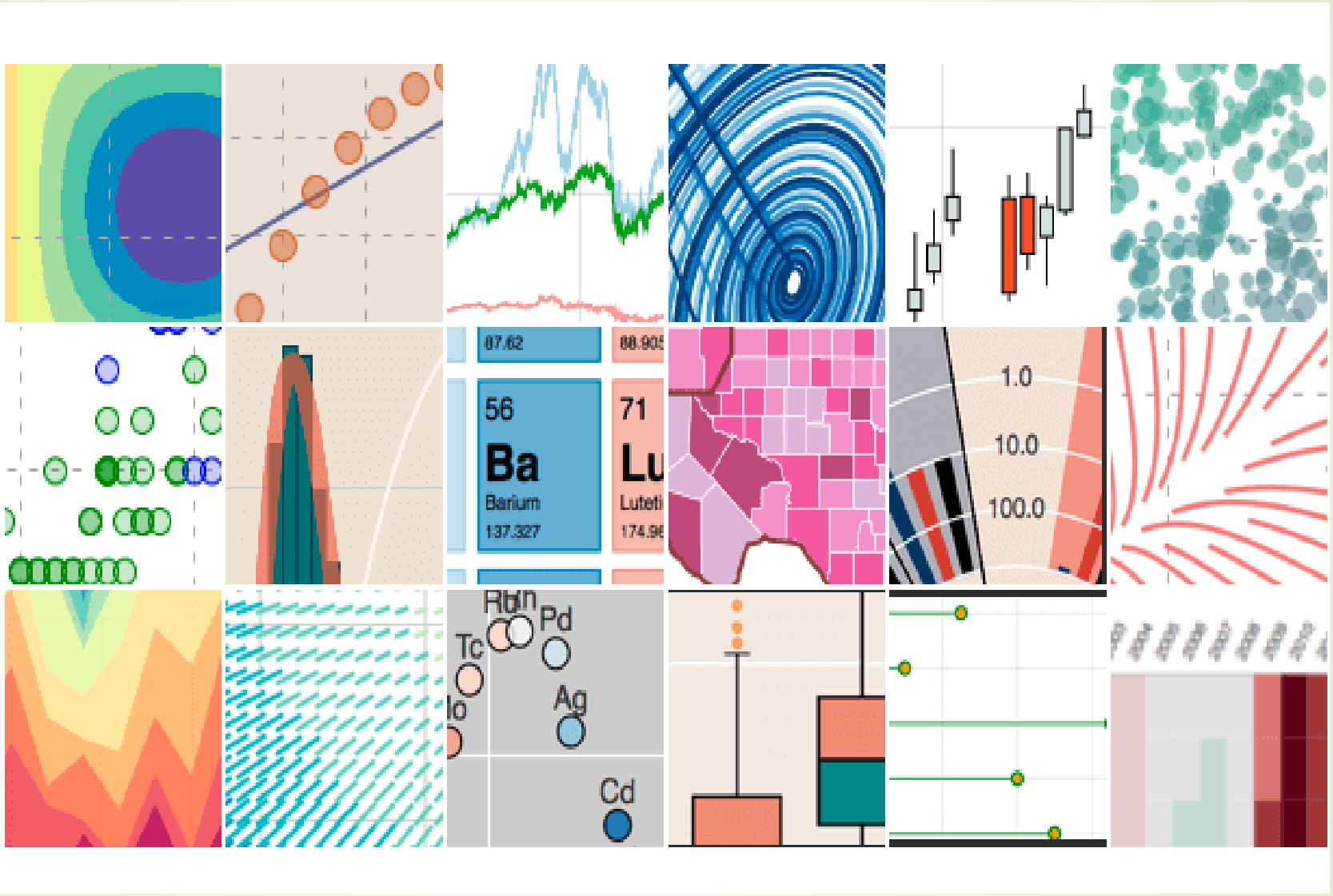
مصورسازی داده‌ها:

- ✓ مصور سازی اطلاعات **یعنی ارائه داده ها در غالب تصویری و گرافیکی** است. این نوع ارائه تصمیم گیران و صاحبین اطلاعات را قادر می‌سازد، تا آنالیز داده‌های خود را به صورت بصری دریافت کنند. همین موضوع باعث می شود تا درک مفاهیم دشوار و شناسایی الگوهای جدید برایشان ساده‌تر گردد.
- ✓ ثابت شده که ما با تماشای تصاویر اطلاعات را بهتر درک می‌کنیم. ما با تجسم همواره یک گام به جلوحرکت خواهیم کرد. **نمودارها همواره سریعتر از جداول هستند** و نمودارهای چند بعدی که حاصل یک هوش تجاری پیشرفته است می تواند چندین صفحه جدول را در آن واحد به ما نشان دهند.

گرافیک به کمک دانش پایگاه داده ها آمده و حجم عظیمی از اطلاعات را می تواند به ما در یک صفحه، نمایش دهد.

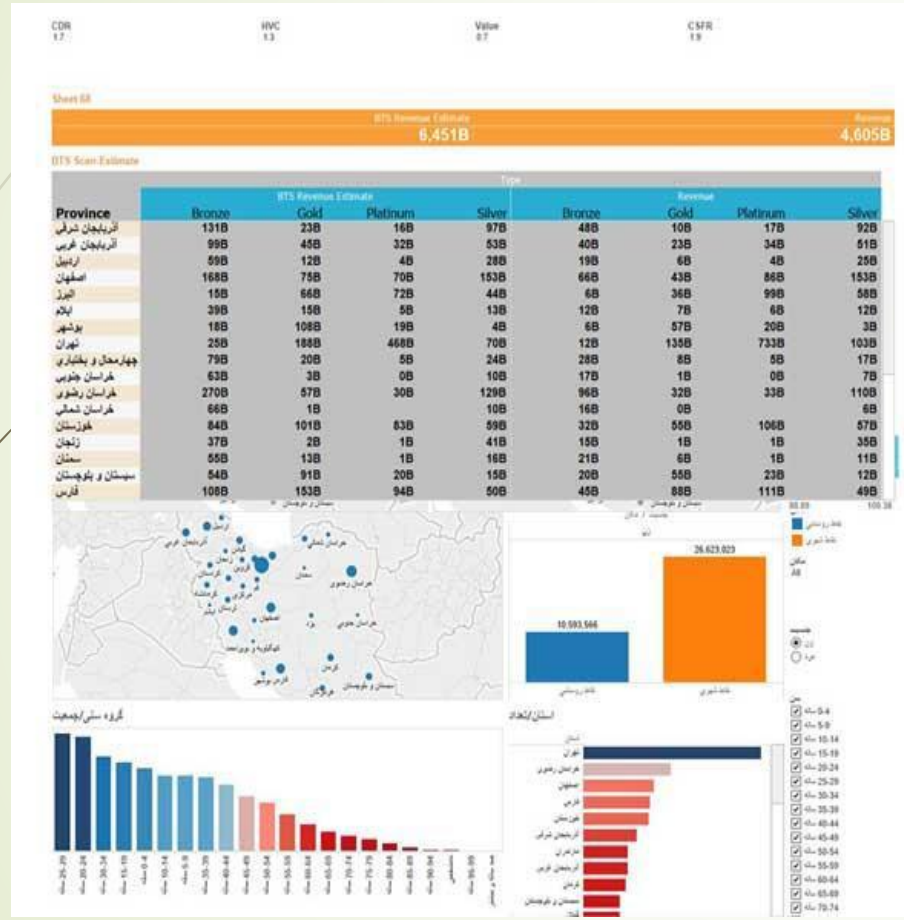
- **با مصور سازی اطلاعات می توان:**

- 1 - مواضعی که نیاز به توجه و بهبود دارند را برای شما شناسایی کند.
- 2 - عواملی که رفتار مشتری را تحت تاثیر قرار می‌دهد، را برایتان نمایان سازد.
- 3 - به شما کمک کند تا بفهمید چه محصولی در کجا بهتر فروخته می شود.
- 4 - پیش بینی (حجم فروش، نقاط فروش و ....) کند.

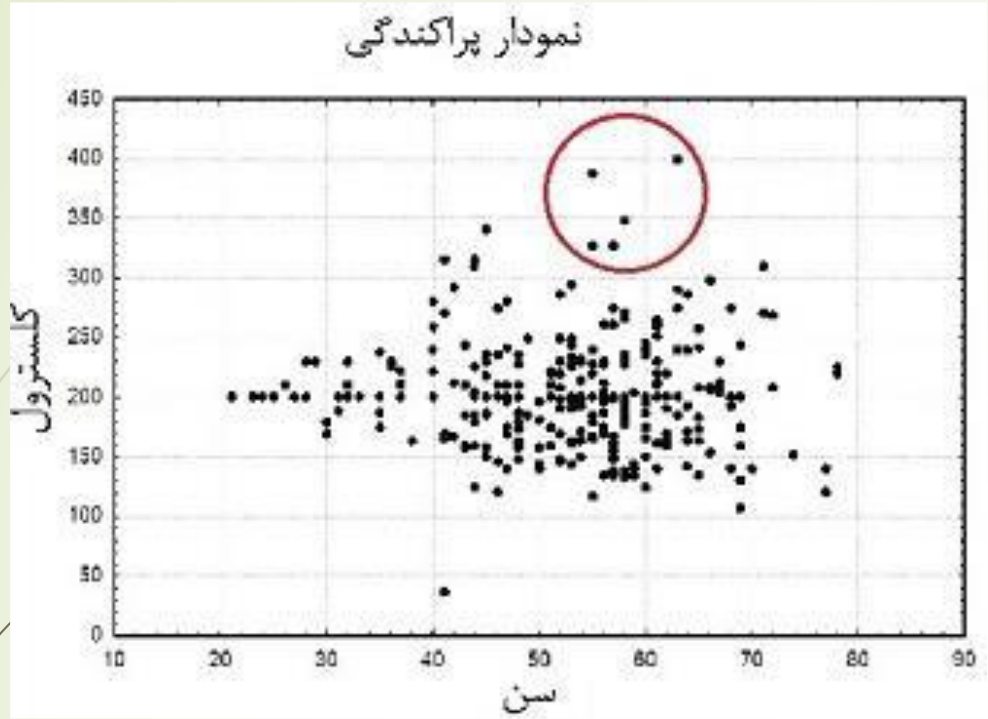

56
Ba
Barium
137.327
71
Lu
1.0
10.0
100.0
Tc
Pd
Ag
Cd

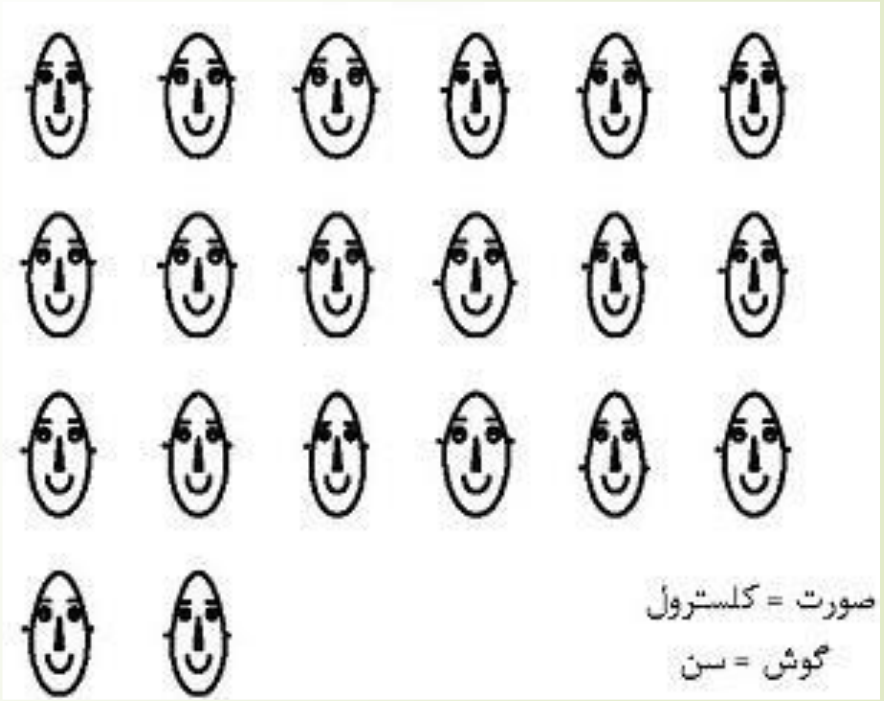
- از روش‌ها مصورسازی داده‌ها می‌توان به موارد زیر اشاره کرد:

- رسم نمودارها
- مصور سازی پیکسل گرا: در این نوع مصور سازی هر مشاهده به صورت یک پیکسل در صفحه نمایش داده می شود.
- ماتریس نمودار پراکنش: نمودار پراکنش شامل نقاط مربوط به مشاهدات X و Y در مقابل هم می باشد.
- نمودار وب یا تار عنکبوتی

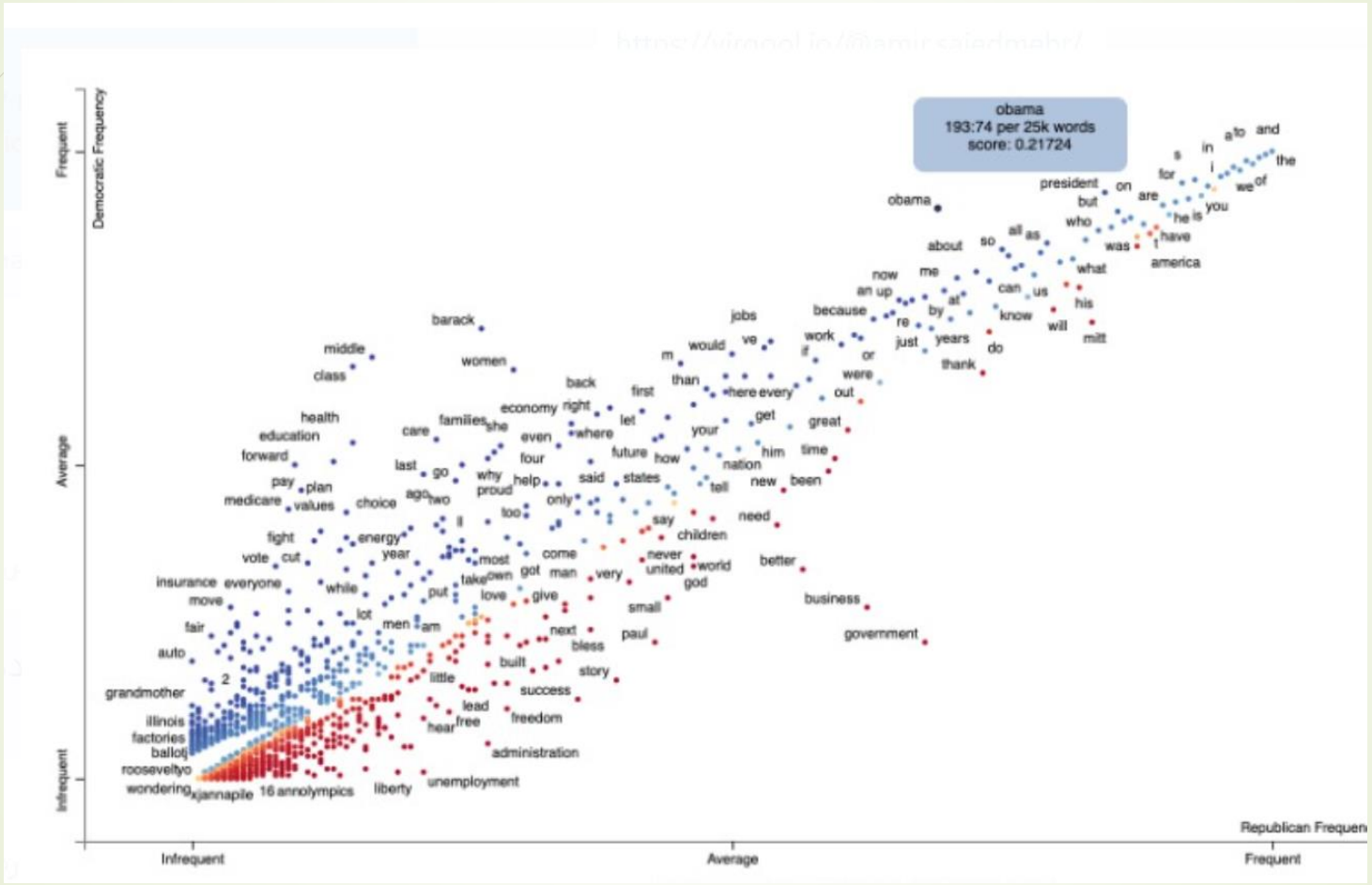

CDR
1.7
HVC
1.3
Value
0.7
CSFR
1.9
Sheet 68
BTS Revenue Estimate
6,451B
Revenue
4,605B
BTS Scan Estimate
Type
BTS Revenue Estimate
Revenue
Province
Bronze
Gold
Platinum
Silver
Bronze
Gold
Platinum
Silver
آذربایجان شرقی 131B 23B 16B 97B 48B 10B 17B 92B
آذربایجان غربی 99B 45B 32B 53B 40B 23B 34B 51B
اردبیل 59B 12B 4B 28B 19B 6B 4B 25B
اصفهان 168B 75B 70B 153B 66B 43B 86B 153B
البرز 15B 66B 72B 44B 6B 36B 99B 58B
ایلام 39B 15B 5B 13B 12B 7B 6B 12B
بوشهر 18B 108B 19B 4B 6B 57B 20B 3B
تهران 25B 188B 468B 70B 12B 135B 733B 103B
چهارمحال و بختیاری 79B 20B 5B 24B 28B 8B 5B 17B
خراسان جنوبی 63B 3B 0B 10B 17B 1B 0B 7B
خراسان رضوی 270B 57B 30B 129B 96B 32B 33B 110B
خراسان شمالی 66B 1B 10B 16B 0B 6B
خوزستان 84B 101B 83B 59B 32B 55B 106B 57B
زنجان 37B 2B 1B 41B 15B 1B 1B 35B
سمنان 55B 13B 1B 16B 21B 6B 1B 11B
سیستان و بلوچستان 54B 91B 20B 15B 20B 55B 23B 12B
فارس 108B 153B 94B 50B 45B 88B 111B 49B
26,623,023
10,593,564
گروه سنی/جمعیت
استان/ابعاد

گروه سنی/جمعیت
مکان/جمعیت
26,623,023
10,593,564
ARPU
Revenue
BTS Scan
Acquistion
1391
1392
GPRS Revenue Per Package
1391
1392
Package Revenue
GPRS Revenue
GPRS ARPU


نمودار پراکندگی
کلسترول
سن
450
400
350
300
250
200
150
100
50
0
10
20
30
40
50
60
70
80
90

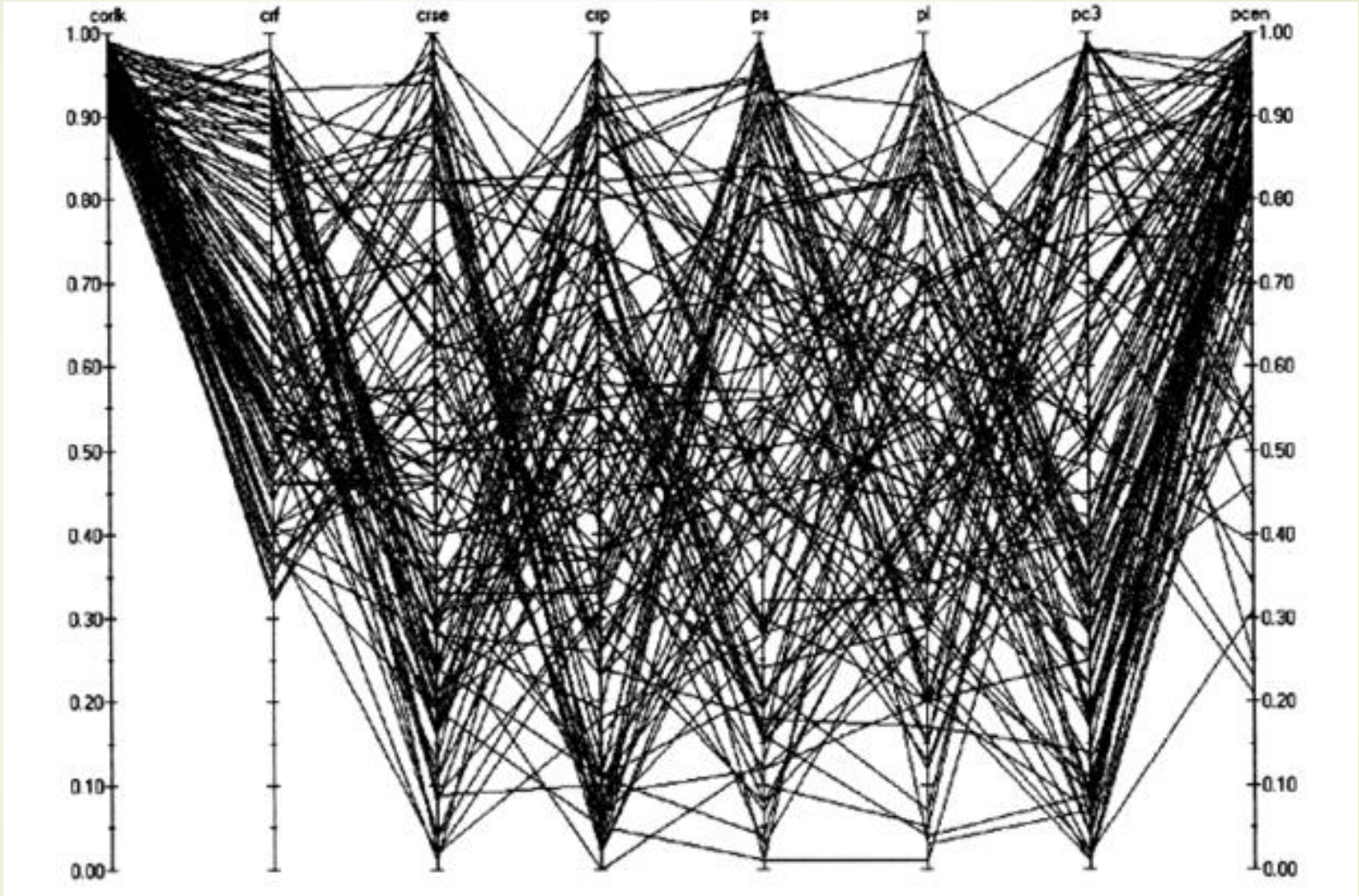

صورت = کلسترول
گوش = سن

p.value
trust
satisfaction
loyalty
WOM
p.value
trust
satisfaction
loyalty
WOM


obama
193:74 per 25k words
score: 0.21724
Democratic Frequency
Frequent
Average
Infrequent
Republican Frequen
Infrequent
Average
Frequent
obama
barack
middle
class
women
health
education
forward
pay
plan
medicare
values
choice
care
families
she
economy
right
back
first
jobs
ve
would
m
than
there
every
out
work
if
or
were
because
re
by
at
up
an
now
me
about
so
all
as
president
but
on
who
was
what
can
us
his
know
will
mitt
just
years
do
thank
are
for
s
in
i
a
to
and
the
we
of
you
he
is
have
america
let
where
even
four
future
how
your
get
great
him
time
nation
new
been
tell
last
go
why
proud
help
said
states
ago
two
only
too
need
say
children
fight
energy
vote
cut
year
most
come
never
united
world
god
better
got
man
very
own
take
put
love
give
small
business
government
insurance
everyone
while
move
lot
men
am
next
paul
bless
fair
auto
built
story
grandmother
little
success
lead
free
freedom
hear
illinois
factories
ballot
administration
rooseveltyo
wondering
xjannapile
16 annolympics
liberty
unemployment

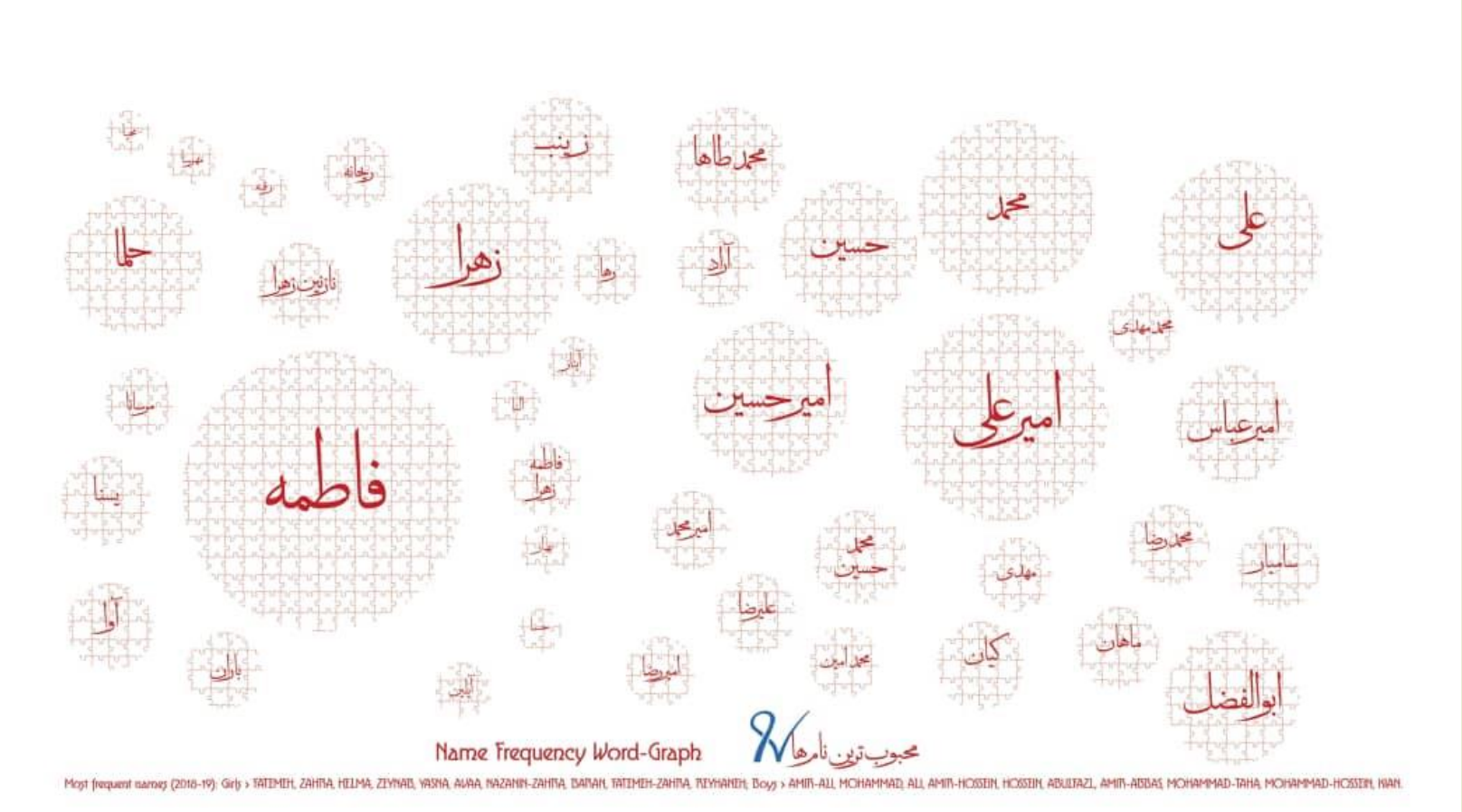
نمودار وب یا تار عنکبوتی:

- با این نمودار می توان میزان یا به عبارتی شدت روابط بین سطوح مختلف متغیرها را نشان داد. در این نمودار هرچه خطوط بین سطوح پررنگ تر باشد نشان دهنده حجم بیشتر داده ها و یا به عبارتی ارتباط قویتر داده ها در این قسمت بوده و خطوط ظریف تر نشان دهنده ارتباط ضعیف تر می باشد.


corlk
crf
crse
crp
ps
pl
pc3
pcen
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00

متن کاوی (Text Mining)

کار
وقت
حس
زندگی
نظرم
خونه
دلم
منم
میخوام
سال
دانشگاه
ساعت
دوست
صبح
بابا
شب
اینه
تمومه
میکنن
دیدم
برم
یکم
اینجوری
داستان
اینا
سرویس
کتاب
میاد
هفته
مدیریت
حرف
تهران
یاد
آدم
شه
دوتا
گوش
میتونه
بحران
قضیه
جواب
منو
صحبت
خاطر


فاطمه
زهرا
حلما
زینب
محمد طاها
حسین
محمد
علی
امیرحسین
امیرعلی
امیرعباس
ابوالفضل
Name Frequency Word-Graph
محبوب‌ترین نام‌ها
Most frequent names (2018-19): Girls › FATEMEH, ZAHRA, HELMA, ZEYNAB, YASNA, AVAA, NAZANIN-ZAHRA, DANAH, FATEMEH-ZAHRA, REYHANEH, Boys › AMIR-ALI, MOHAMMAD ALI, AMIR-HOSSEIN, HOSSEIN, ABOLFAZL, AMIR-ABBAS, MOHAMMAD-TAHA, MOHAMMAD-HOSSEIN, KIAN.