دانشگاه سمنان
Semnan University
پردیس فرزانگان

بسمه تعالی

داده کاوی

Data mining

مدرس

فاطمه دارائی

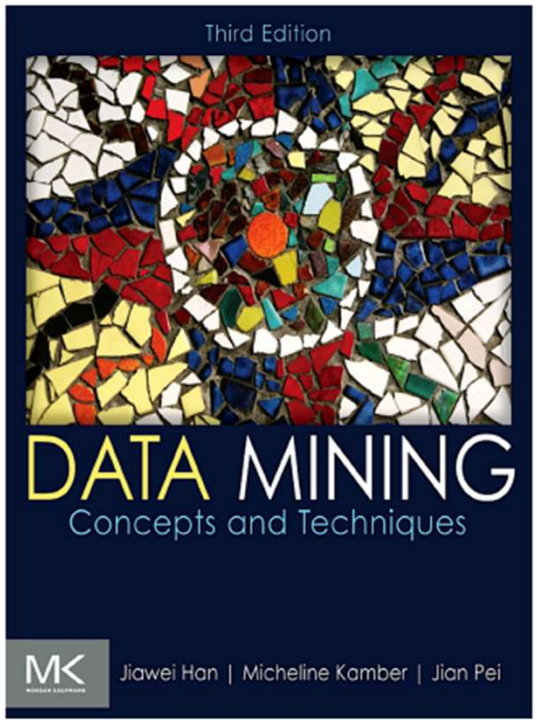
f_daraei@semnan.ac.ir

https://fdaraei.profile.semnan.ac.ir

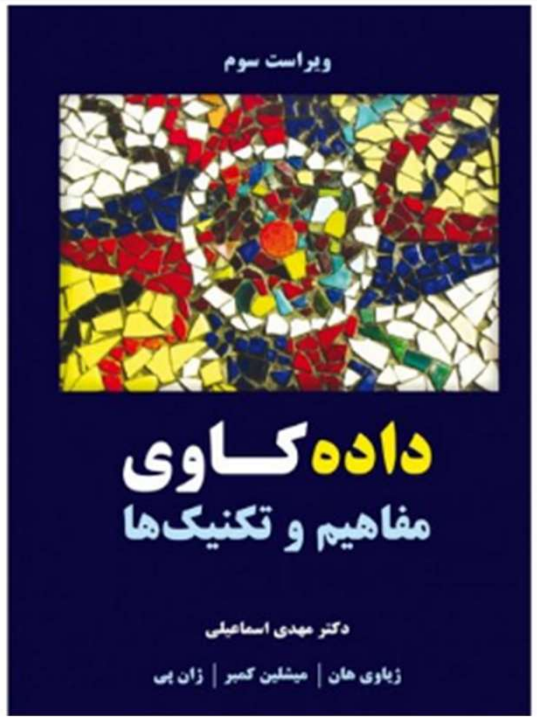
دانشگاه سمنان
Semnan University
پردیس فرزانگان

مراجع اصلی درس

Data Mining: Concepts and Techniques (3rd ed.)
Jiawei Han, Micheline Kamber, and Jian Pei



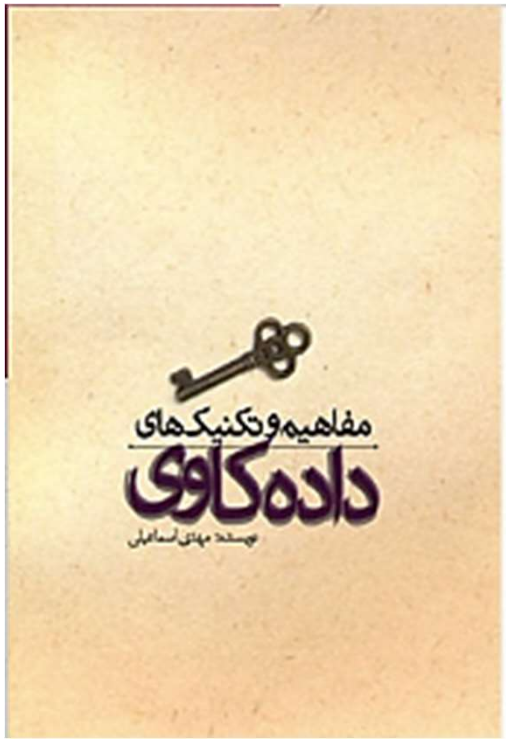
داده کاوی مفاهیم و تکنیک ها
نویسنده ژیاوی هان
مترجم مهدی اسماعیلی



کتاب مفاهیم و تکنیک های داده کاوی

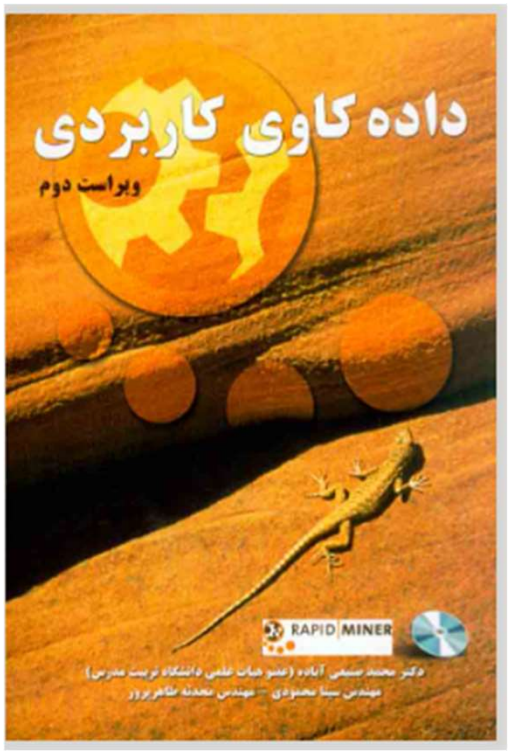
نویسنده: مهدی اسماعیلی

انتشارات دانشگاه آزاد اسلامی واحد کاشان



داده کاوی کاربردی

نویسنده: دکتر محمد صنیعی آباده (عضو هیئت علمی دانشگاه تربیت مدرس)، سینا محمودی و محدثه طاهرپرور



دانشگاه سمنان
Semnan University
پردیس فرزانگان

ارزشیابی


میان ترم ۶ نمره


پایان ترم ۱۰ نمره

تکلیف ۲ نمره

پروژه ۳ نمره

مجموع از ۲۱ نمره

دانشگاه سمنان
Semnan University
پردیس فرزانگان

سرفصل مطالب

✓فصل اول: معرفی داده کاوی

شناخت انواع داده ها و ویژگی ها، توصیف آماری داده ها

✓فصل دوم: پیش پردازش داده‌ها

تجمیع داده، پاک سازی داده، عملیات پیش پردازش، روش های کاهش داده

✓فصل سوم: تحلیل الگوهای مکرر و قواعد انجمنی

قواعد انجمنی، الگوریتم Apriori

✓فصل چهارم: دسته بندی و طبقه بندی

Decision Tree , Classification , Knn

✓فصل پنجم: رده بندی و خوشه بندی

Kmeans, Clustering

✓فصل ششم: شبکه عصبی

معرفی یادگیری عمیق، مروری بر شبکه های عصبی کانولوشن

دانشگاه سمنان
Semnan University
پردیس فرزانگان

چرا داده کاوی؟

مؤسسه جهانی مک‌کنزی MGI گزارش می‌دهد (در سال ۲۰۱۱) که بیشتر شرکت‌های آمریکایی با بیش از ۱۰۰۰ کارمند به طور متوسط حداقل ۲۰۰ ترابایت داده ذخیره شده داشتند.

MGIپیش‌بینی می‌کند که میزان داده‌های تولید شده در سراسر جهان سالانه ۴۰٪ افزایش خواهد یافت.

مقادیر بسیار زیادی از داده‌های خام!!

رشد انفجاری داده‌ها: از ترابایت‌ها به پتابایت‌ها

در عصر دیجیتال، در هر لحظه ترابایت‌ها داده تولید می‌شود. داده ها با سرعت زیاد جمع آوری و ذخیره می شوند (گیگا بایت در ثانیه)

جمع‌آوری داده و دسترسی به داده

ابزارهای خودکار جمع‌آوری داده، سیستم‌های پایگاه‌داده، وب، جامعه‌ی کامپیوتری

منابع اصلی داده‌های فراوان، توسعه نرم افزارهای تجاری آماده بکار

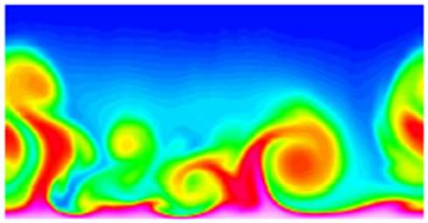
رشد فوق العاده در قدرت محاسباتی و ظرفیت ذخیره سازی

دانشگاه سمنان
Semnan University
پردیس فرزانگان

حسگرهای موجود در ماهواره ها، تلسکوپ های فضایی، دستگاه‌های موبایل، عکس‌های دیجیتال، اسناد وب.

به‌روزرسانی‌های فیسبوک، توییت‌ها، وبلاگ‌ها، محتوای تولید شده توسط کاربر.
تراکنش‌ها، داده‌های سنسورها، داده‌های نظارتی.
پرسش‌ها، کلیک‌ها، مرور صفحات.

علم: سنجش از دور، زیست‌اطلاعاتی، شبیه‌سازی علمی، ...
جامعه و همه: اخبار، دوربین‌های دیجیتال، یوتیوب
تجارت: وب، تجارت الکترونیک، تراکنش‌ها، سهام، ...



دانشگاه سمنان
Semnan University
پردیس فرزانگان

ما در داده غرق شده‌ایم، اما از دانش گرسنه‌ایم!

حجم زیاد داده می تواند قدرتمندتر از الگوریتم ها و مدل های پیچیده باشد

داده قدرت است!!

نیاز به تحلیل داده‌های خام برای استخراج دانش و تحلیل خودکار مجموعه‌های بزرگ داده

داده ها نیز بسیار پیچیده هستند

انواع مختلفی از داده ها:

جداول، سری های زمانی، تصاویر، نمودارها و غیره

جنبه های مکانی و زمانی

داده های به هم پیوسته از انواع مختلف:

از تلفن همراه می‌توانیم، مکان کاربر، اطلاعات دوستی، ورود به مکان‌ها، نظرات از طریق توییتر، تصاویر از طریق دوربین، پرسش‌ها به موتورهای جستجو را جمع‌آوری کنیم.

مثال: داده های تراکنش

میلیاردها مشتری واقعی:

WALMART: دویست میلیون تراکنش در روز

AT&T: سیصد میلیون تماس در روز

شرکت های کارت اعتباری: میلیاردها تراکنش در روز

کارت‌های امتیاز به شرکت‌ها اجازه می‌دهند تا اطلاعات مربوط به کاربران خاص را جمع‌آوری کنند

مثال: داده های سند

وب به عنوان یک مخزن اسناد:

۵۰ میلیارد صفحه وب تخمین زده می شود

ویکی پدیا:

۴ میلیون مقاله (و در حال افزایش)

پورتال های خبری آنلاین:

هر روز یک جریان ثابت از ۱۰۰ مقاله جدید

توییتر:

حدود ۳۰۰ میلیون توییت در روز

مثال: داده های شبکه

وب:

۵۰ میلیارد صفحه که از طریق لینک ها به یکدیگر پیوند داده شده اند

فیس بوک:

۵۰۰ میلیون کاربر

توییتر:

۳۰۰ میلیون کاربر

پیام رسان فوری:

~ ۱ میلیارد کاربر

وبلاگ ها:

۲۵۰ میلیون وبلاگ در سراسر جهان

تکامل علوم

قبل از سال ۱۶۰۰: علم تجربی (Empirical Science):
در این دوره، علم بیشتر بر مشاهده و آزمایش مستقیم استوار بود. دانشمندان با جمع‌آوری داده‌های تجربی از محیط پیرامون خود، به صورت دستی نتایج را تحلیل می‌کردند. این مرحله شامل دوران پیش از توسعه نظریه‌های علمی منسجم است.

۱۶۰۰ تا دهه ۱۹۵۰: علم نظری (Theoretical Science):
با توسعه نظریه‌های علمی، علوم وارد مرحله نظری شدند. در این دوره، مدل‌های نظری توسعه یافتند که برای توضیح پدیده‌های طبیعی استفاده می‌شدند. این مدل‌ها اغلب به‌عنوان انگیزه‌ای برای انجام آزمایش‌ها و عمومی‌سازی نتایج تجربی به کار می‌رفتند.

دهه ۱۹۵۰ تا دهه ۱۹۹۰: علم محاسباتی (Computational Science):
با ظهور کامپیوترها و پیشرفت‌های محاسباتی، علوم وارد مرحله محاسباتی شدند. در این مرحله، بیشتر رشته‌های علمی به یک شاخه محاسباتی مجهز شدند که امکان شبیه‌سازی و تحلیل مدل‌های پیچیده ریاضی را فراهم کرد. این تغییر به ویژه در علومی مانند فیزیک، اکولوژی، و زبان‌شناسی مشهود بود.

۱۹۹۰ تاکنون: علم داده (Data Science):
از دهه ۱۹۹۰ به بعد، با افزایش سریع حجم داده‌ها و پیشرفت در فناوری‌های ذخیره‌سازی و پردازش داده‌ها، علوم به سمت داده‌محوری حرکت کردند.
در این دوره، داده‌ها از منابع مختلف از جمله ابزارهای علمی جدید و شبیه‌سازی‌ها به‌سرعت تولید و جمع‌آوری می‌شوند. علوم داده بر تحلیل و استخراج دانش از این داده‌ها تمرکز دارد.
اینترنت و شبکه‌های محاسباتی جهانی باعث دسترسی عمومی به این داده‌ها و تحلیل‌های مرتبط با آن‌ها شدند.
داده‌کاوی در این مرحله به عنوان یک چالش و همچنین یک ابزار ضروری برای مدیریت، سازماندهی، و تحلیل داده‌های عظیم مطرح شد.

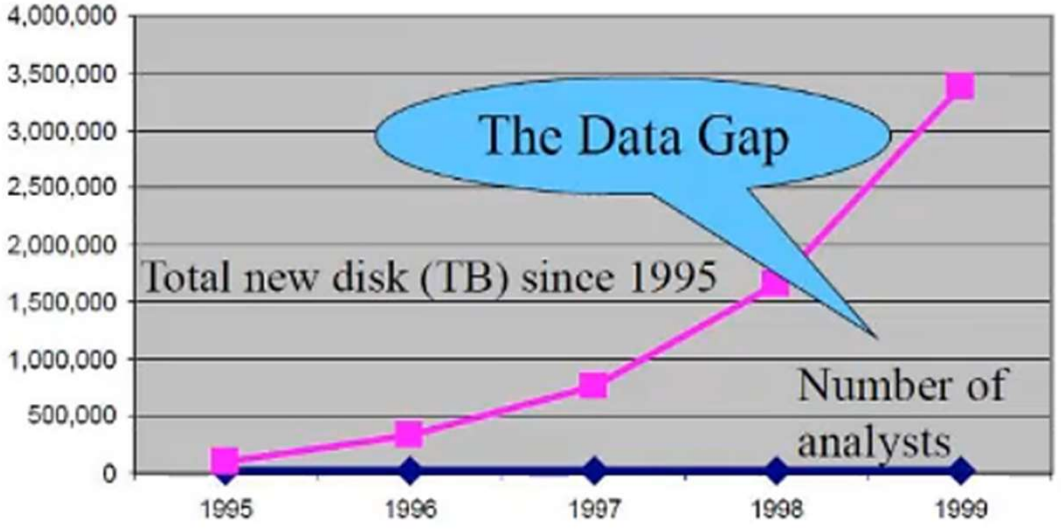
انگیزه

کاوش مجموعه داده های بزرگ

معمولا اطلاعات نهفته ای در داده ها وجود دارند که تاکنون آشکار نشده اند.

برای کشف اطلاعات مفید توسط انسانها هفته ها زمان نیاز است.

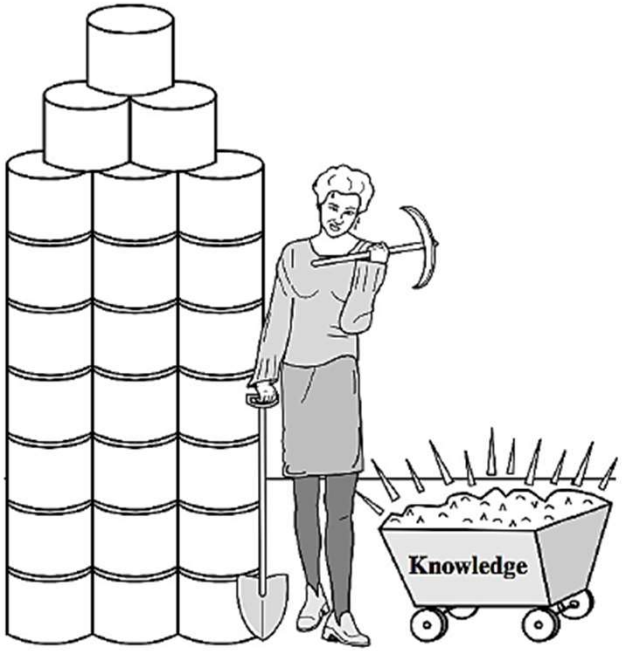
خیلی از داده ها هنوز تحلیل نشده اند.

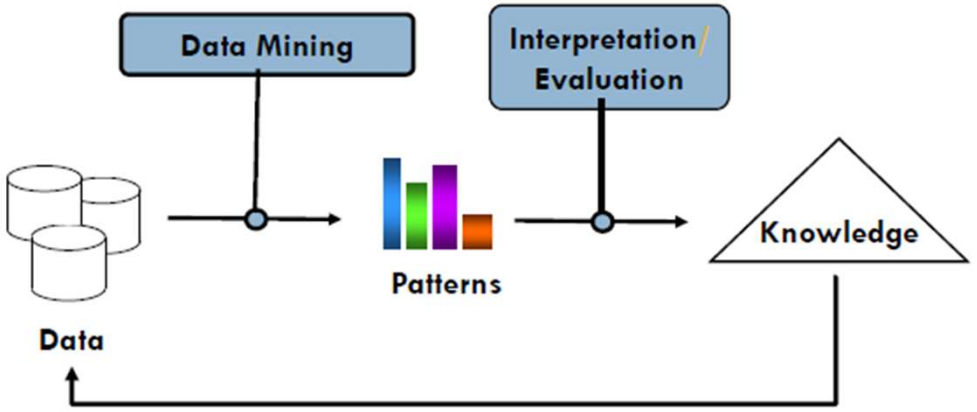


دانشگاه سمنان
Semnan University
پردیس فرزانگان

تعریف داده کاوی

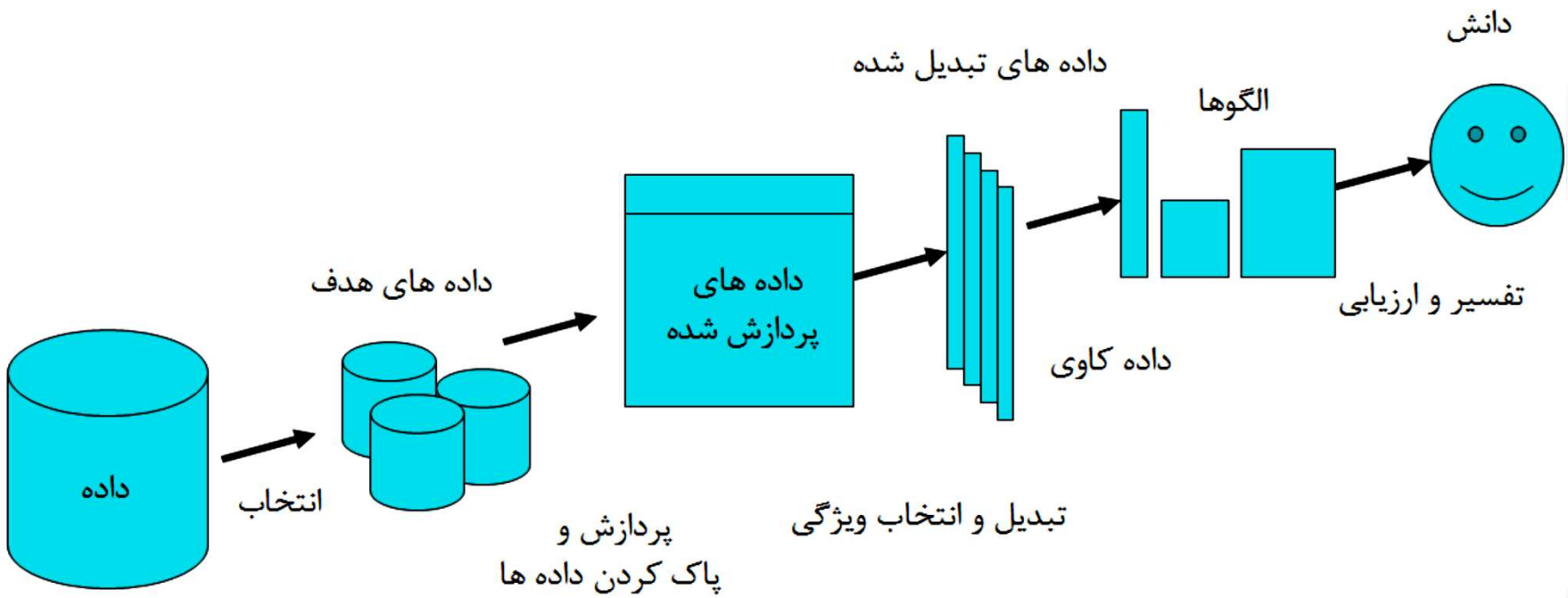
- استخراج اطلاعات مفید و ناشناخته از داده ها
- فرآیند استخراج دانش، الگوها و اطلاعات مفید از مجموعه‌های بزرگ داده به وسیله ابزارهای خودکار و نیمه خودکار
- استفاده از تکنیک‌های کارآمد برای تحلیل مجموعه‌های بسیار بزرگ داده و استخراج الگوهای مفید و احتمالاً غیرمنتظره در داده‌ها





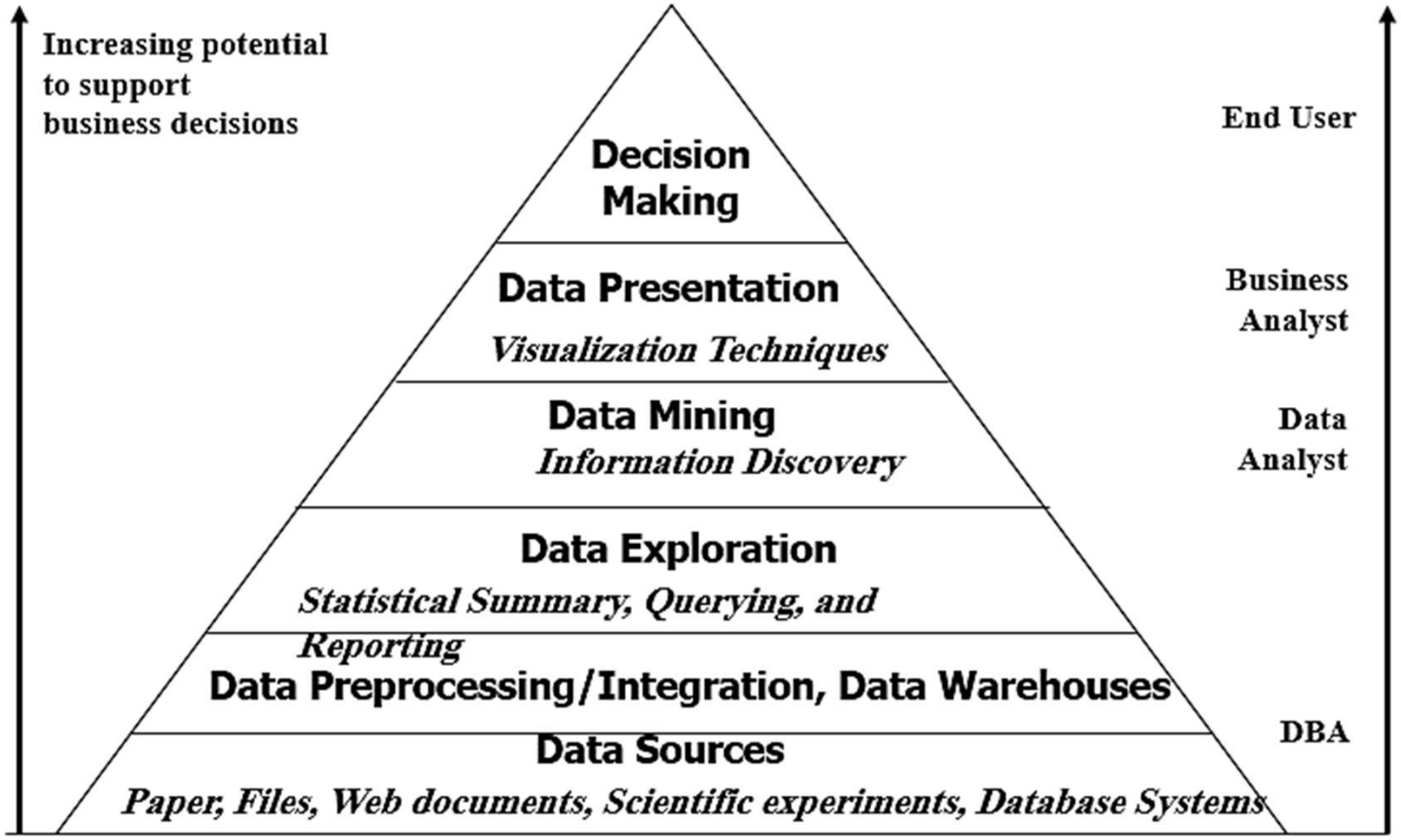
مراحل کشف دانش

داده کاوی نقش اساسی در فرآیند کشف دانش دارد.



Knowledge Discovery from Data(KDD)

داده کاوی در هوش تجاری



یک نگاه کلی به داده کاوی

❖**داده‌هایی که باید استخراج شوند:**

داده‌های پایگاه داده (رابطه‌ای توسعه‌یافته، شیءگرا، ناهمگن، میراثی)، انبار داده‌ها، داده‌های تراکنشی، سری زمانی، توالی، متن و وب، چندرسانه‌ای، گراف‌ها و شبکه‌های اجتماعی و اطلاعاتی

❖**دانش استخراج‌شده (یا: وظایف داده‌کاوی):**

شخصیت‌سازی Characterization، تبعیض discrimination، وابستگی association، طبقه‌بندی classification، خوشه‌بندی clustering، روند/انحراف trend/deviation، تحلیل نقاط پرت outlier analysis و غیره

داده‌کاوی توصیفی در مقابل داده‌کاوی پیش‌بینی‌کننده

توابع چندگانه/یکپارچه و استخراج در سطوح چندگانه

❖**تکنیک‌های استفاده‌شده:**

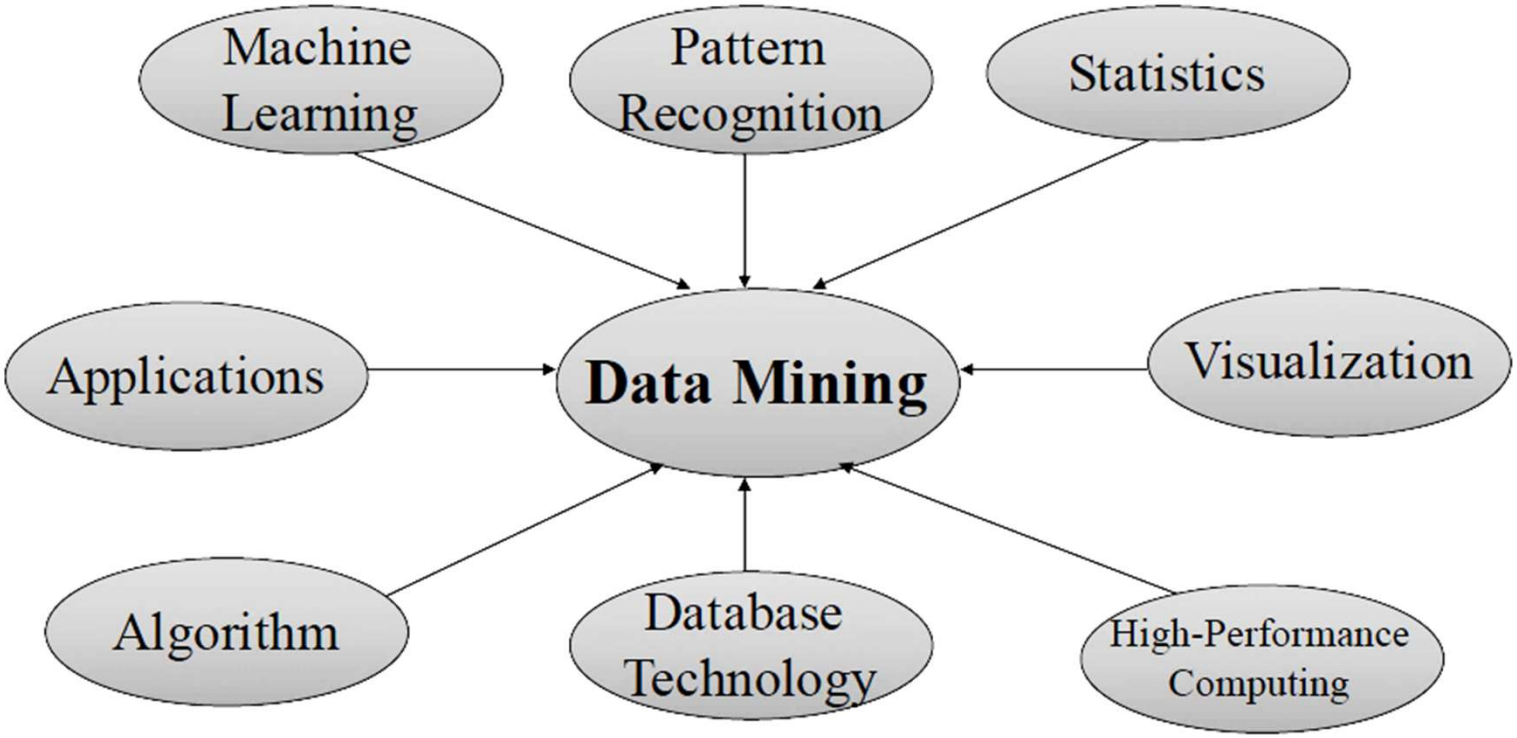
داده‌محور، انبار داده (OLAP)، یادگیری ماشین، آمار، تشخیص الگو، تجسم‌سازی، عملکرد بالا و غیره

❖**کاربردهای تطبیق‌یافته:**

خرده‌فروشی، ارتباطات از راه دور، بانکداری، تحلیل تقلب، داده‌کاوی زیستی، تحلیل بازار سهام، داده‌کاوی متنی، داده‌کاوی وب و غیره

دانشگاه سمنان
Semnan University
پردیس فرزانگان

داده کاوی: تلاقی چندین رشته



دانشگاه سمنان
Semnan University
پردیس فرزانگان

از جمله کاربردهای داده کاوی

- کاربردهای تجاری
- کاربردهای علمی
- کاربردهای امنیتی

دانشگاه سمنان
Semnan University
پردیس فرزانگان

کاربردهای تجاری

- تقریبا در تمام سازمانها و انواع تجارتها، به دلیل وجود اطلاعات، می توان داده کاوی را مورد استفاده قرار داد.
- پیش بینی مربوط به بازار بورس
- تحلیل سبد خرید
- شناسائی طبقات و گروههای اصلی مشتریان
- تعیین میزان تاثیر عوامل مختلفی نظیر تبلیغات، تخفیف، ... بر میزان و الگوهای فروش

دانشگاه سمنان
Semnan University
پردیس فرزانگان

کاربردهای علمی

- اطلاعات جمع آوری شده در حوزه های مختلف: اطلاعات جغرافیائی، اطلاعات اقلیمی، اطلاعات پزشکی
- حجم بسیار بالا و ویژگی های متعدد
- تنوع اطلاعات
- نویز شدید در غالب اطلاعات جمع آوری شده توسط سنسورها
- حوزه پزشکی:
  - تشخیص بیماریها براساس انواع اطلاعات (تصاویر پزشکی، مشخصات بیمار احتمالی)
  - تشخیص ناهنجاریهائی که توسط انسان به سختی قابل تشخیص خواهند بود (لکه ها و نقاط خاص داخل چشم که نشانه شروع کوری ناشی از دیابت می باشد)

دانشگاه سمنان
Semnan University
پردیس فرزانگان

کاربردهای علمی (ادامه)

- حوزه اطلاعات جغرافیائی و اقلیمی
  - کشف پدیده های اقلیمی جدید
  - تکنیکهای بصری سازی و بازنمائی اطلاعات
  - پردازش انواع اطلاعات (تصاویر، اطلاعات به دست آمده از سنجنده ها)
- حوزه کاربردی فضا و سفرهای فضائی
  - حجم بسیار زیادی از اطلاعات
  - نویز بسیار بالا
  - ارزش بسیار زیاد دانش قابل استخراج
  - پردازش اطلاعات جمع آوری شده از فضا
  - پردازش اطلاعات مربوط به سفینه های فضائی
  - ارائه دانش مفید برای اتخاذ تصمیم نهائی جهت پرتاب یا عدم پرتاب یک سفینه به فضا

کاربردهای امنیتی

- سیستمهای تشخیص نفوذ
  - روشهای سنتی، نظیر تشخیص حملات با استفاده از قوانین ارائه شده توسط متخصصان، علاوه بر نیاز به اصلاح دائم، برای مقابله با انواع جدید حملات کافی نیستند.
  - حجم اطلاعات بسیار زیاد و فضای حالت غیرقابل تصور
  - عدم امکان بررسی تمام گزارشهای فعالیت توسط متخصصان شبکه
  - نیاز به شناسائی خودکار الگوهای جدید و مشکوک به تلاش برای نفوذ
  - لزوم همکاری با متخصصان شبکه، از طریق خلاصه سازی وضعیت موجود و درخواست نظر متخصص در موارد مشکوک
  - لزوم اجتناب از سیستمهای بسیار بدبین که موجب بی اعتنایی متخصصان به هشدارهای سیستم خواهد شد.

کاربردهای امنیتی (ادامه)

- مقابله با تروریسم
  - در سالهای اخیر، به خصوص پس از واقعه ۱۱ سپتامبر، به صورت فزاینده ای مطرح شده است.
  - به دلیل عدم امکان انتشار تمامی اطلاعات مفید، پیشرفت کندتری (حداقل از نظر افراد عادی) دارد.
  - در حالت ایده آل، داده کاوی باید بتواند با پردازش اطلاعات از انواع مختلف، نسبت به احتمال وقوع حملات تروریستی، با ذکر جزئیات کافی، هشدار دهد.
  - نتایج حاصل از آن می تواند در صورت عدم وجود دقت کافی، فاجعه آمیز باشد.

دانشگاه سمنان
Semnan University
پردیس فرزانگان

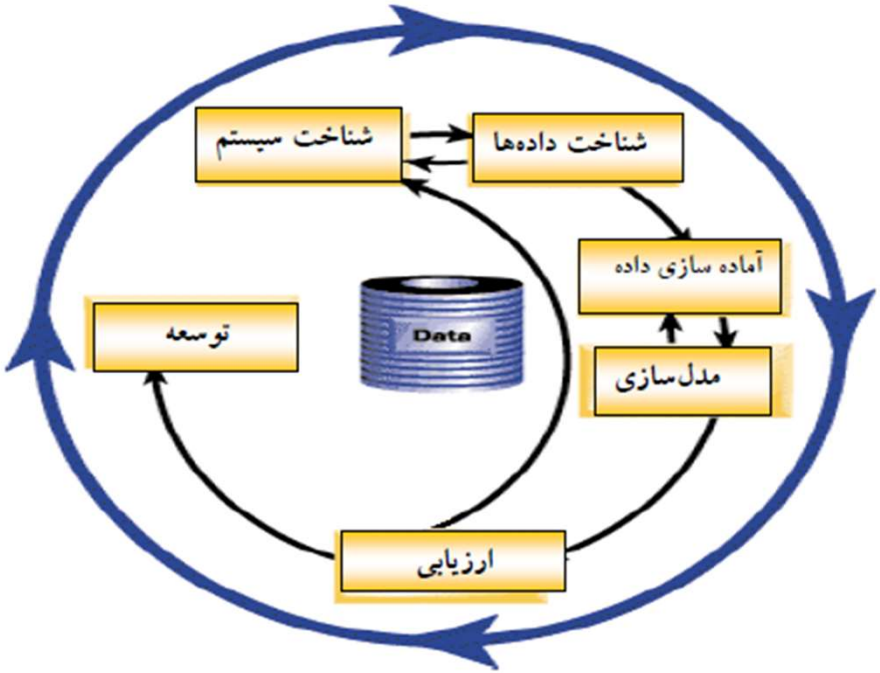
چگونگی انجام داده کاوی

فرایند داده کاوی می‌باستی قابل استفاده به وسیله افراد با پیش زمینه کمی از اطلاعات داده کاوی باشد.

- به سادگی برای افراد و کاربران قابل فهم و استفاده باشد.
- مطلب جدیدی که کاربر قبلا از آن اطلاعی نداشت را ارائه دهد.
- مدل های بدست آمده معتبر و مطمئن باشد.
- روابط کشف شده مفید و قابل اجرا و استفاده باشند.

متدلوژی CRISP

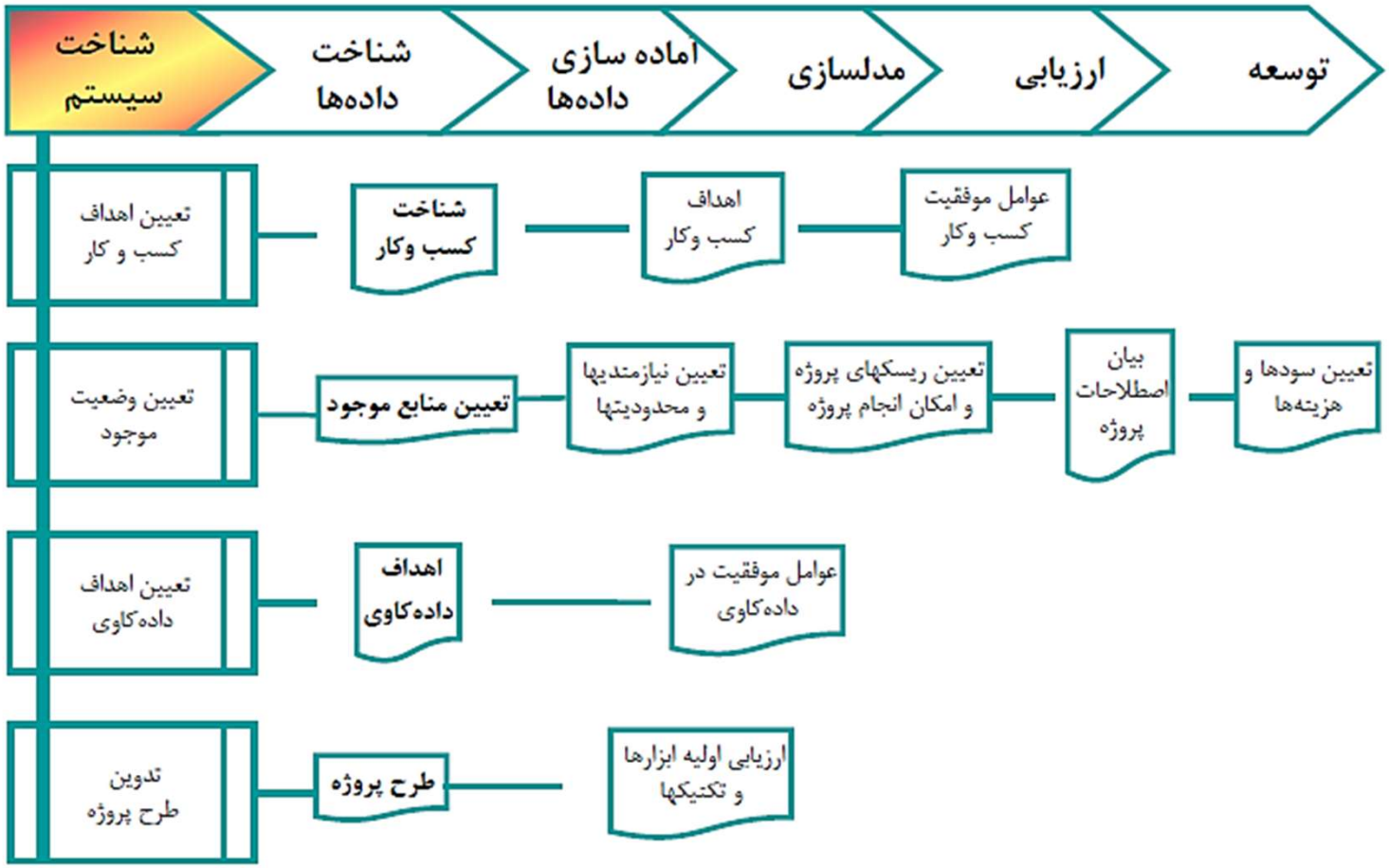
CRoss Industry Standard Process for Data Mining

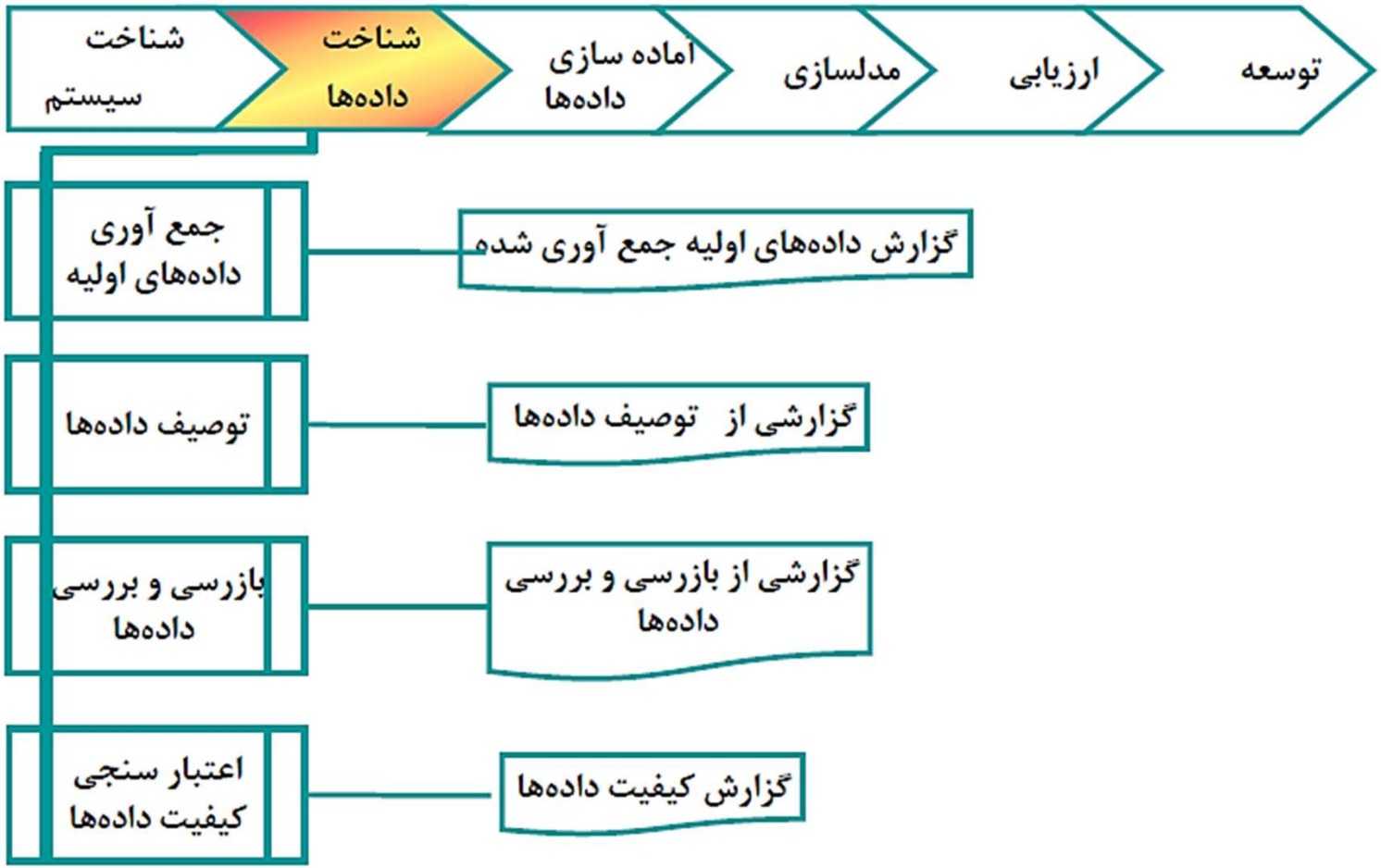


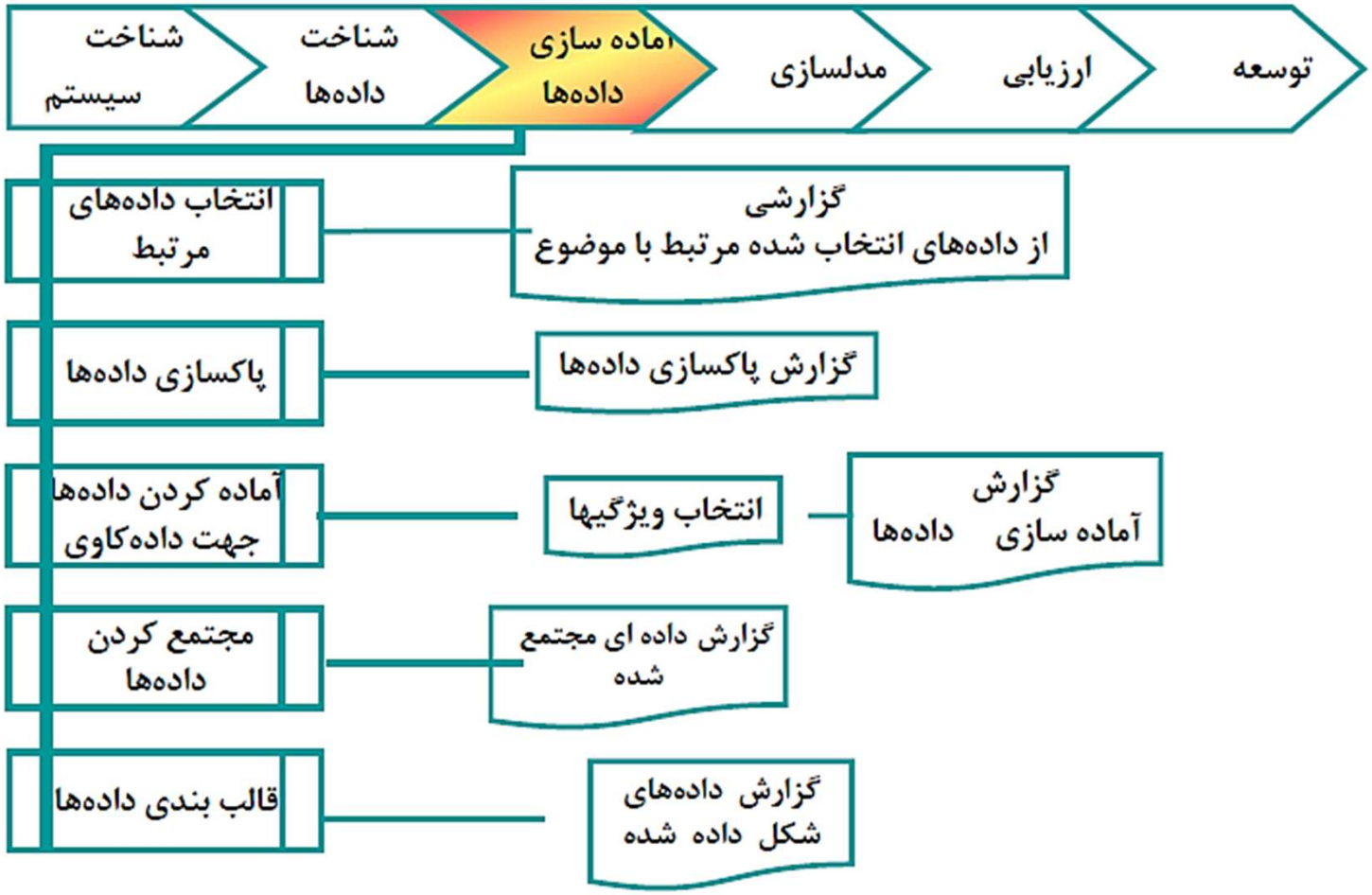
فازها:

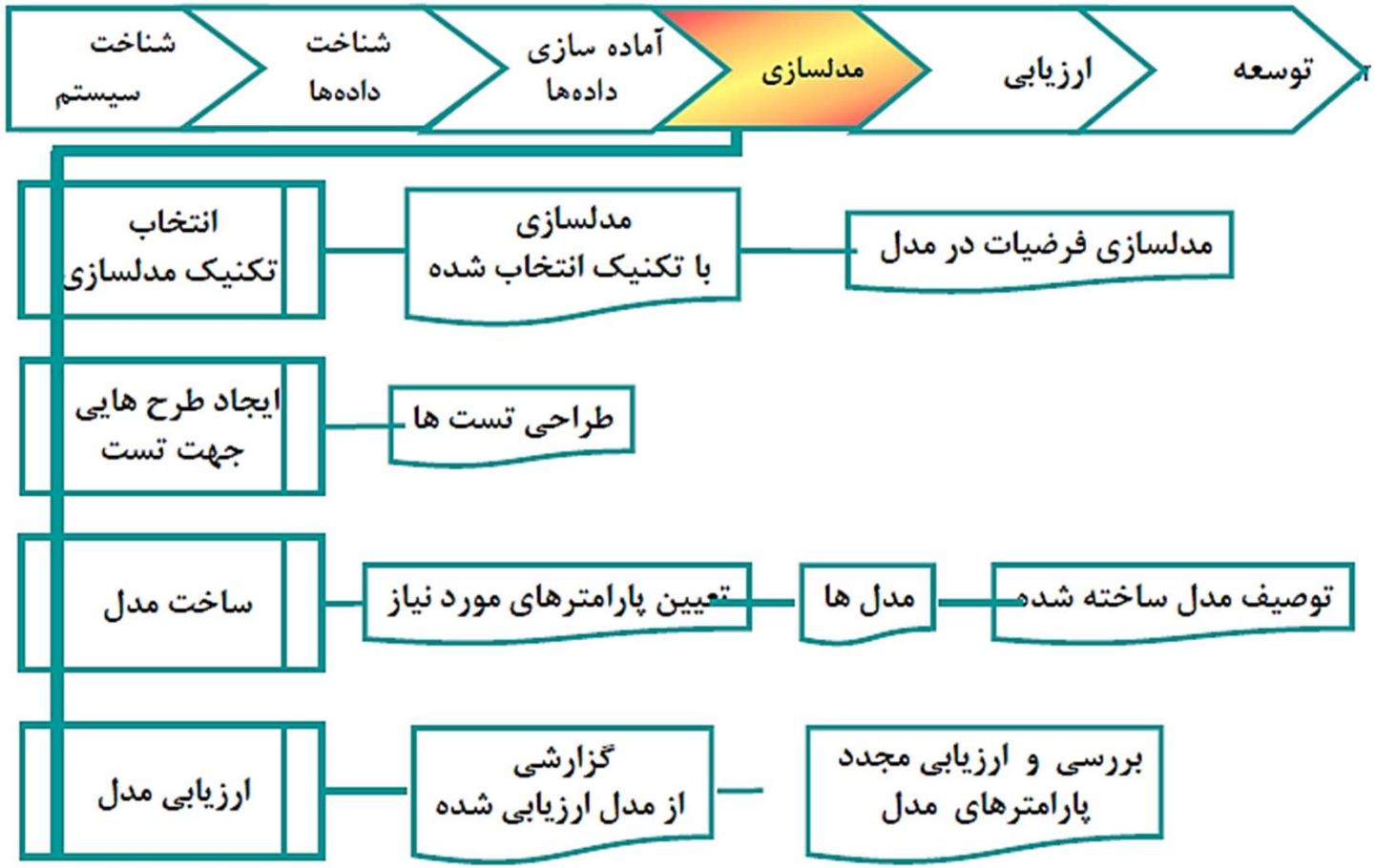
- شناخت سیستم
- شناخت داده ها
- آماده سازی داده ها
- مدلسازی
- ارزیابی
- توسعه

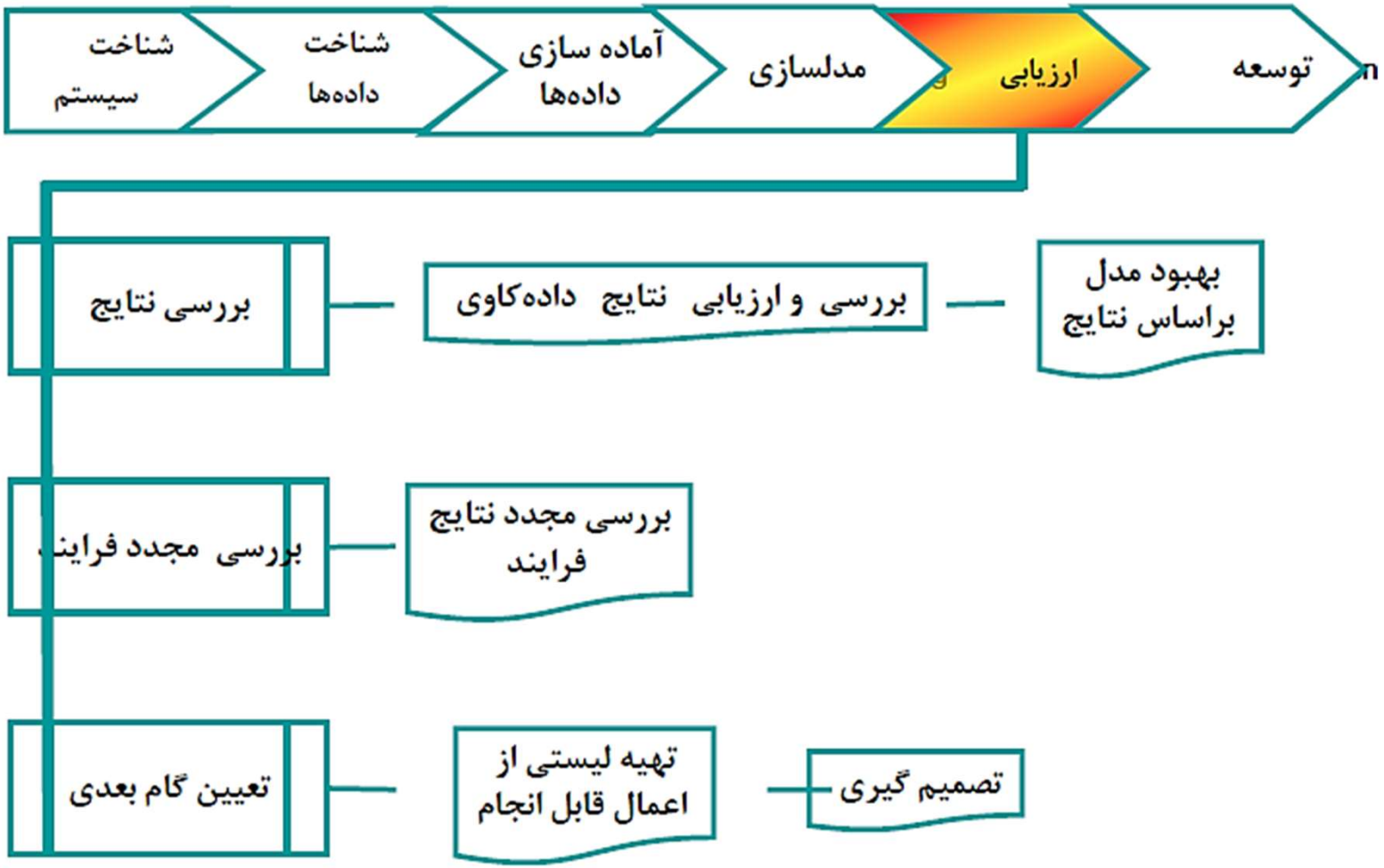
هر کدام از این فازها به زیربخش هایی تقسیم می‌شوند:

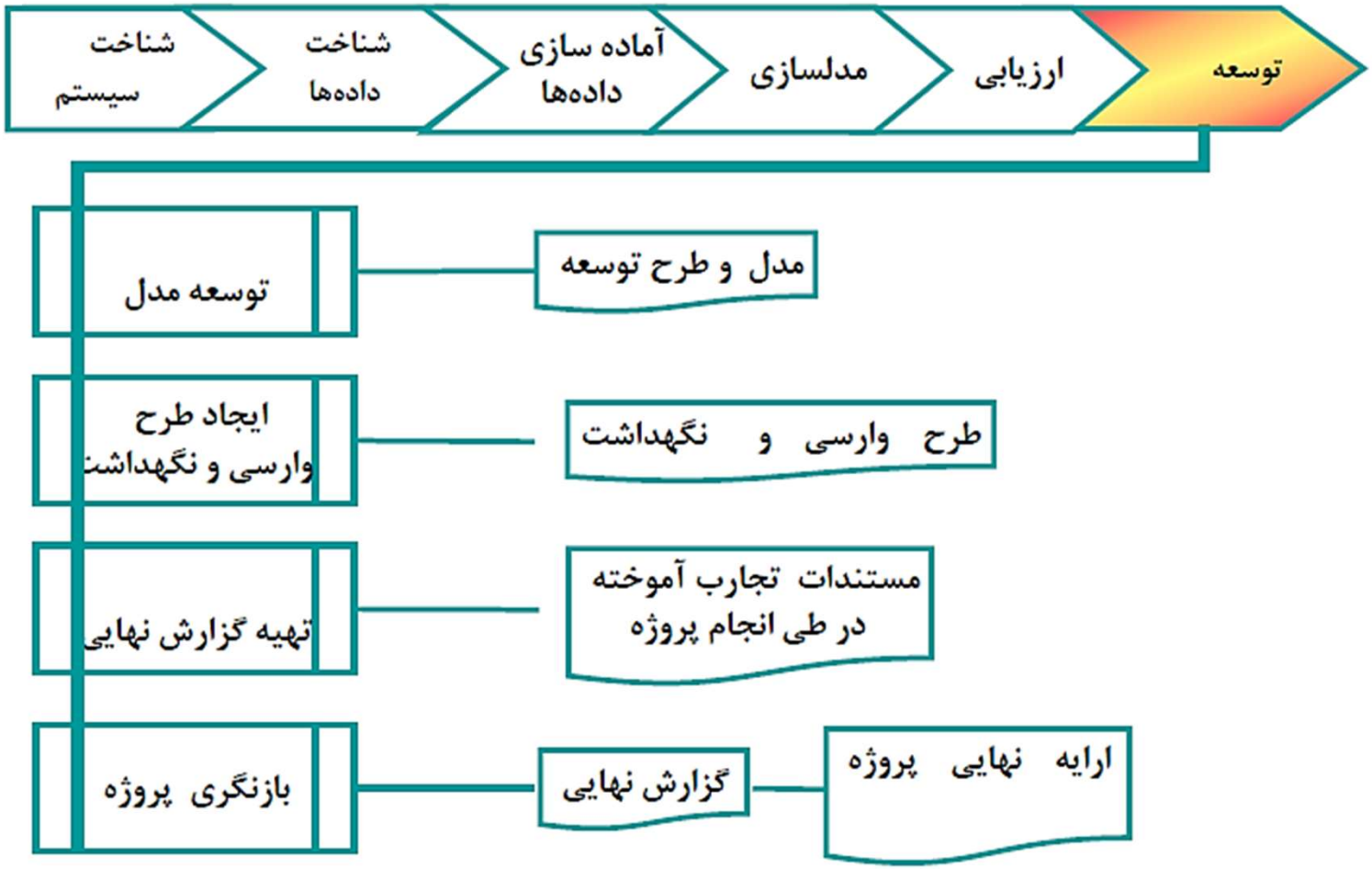

شناخت سیستم
شناخت داده‌ها
اماده سازی داده‌ها
مدلسازی
ارزیابی
توسعه
تعیین اهداف کسب و کار
شناخت کسب وکار
اهداف کسب وکار
عوامل موفقیت کسب وکار
تعیین وضعیت موجود
تعیین منابع موجود
تعیین نیازمندیها و محدودیتها
تعیین ریسکهای پروژه و امکان انجام پروژه
بیان اصطلاحات پروژه
تعیین سودها و هزینه‌ها
تعیین اهداف داده‌کاوی
اهداف داده‌کاوی
عوامل موفقیت در داده‌کاوی
تدوین طرح پروژه
طرح پروژه
ارزیابی اولیه ابزارها و تکنیکها


شناخت سیستم
شناخت داده‌ها
آماده سازی داده‌ها
مدلسازی
ارزیابی
توسعه
جمع آوری داده‌های اولیه
گزارش داده‌های اولیه جمع آوری شده
توصیف داده‌ها
گزارشی از توصیف داده‌ها
بازرسی و بررسی داده‌ها
گزارشی از بازرسی و بررسی داده‌ها
اعتبار سنجی کیفیت داده‌ها
گزارش کیفیت داده‌ها


شناخت سیستم
شناخت داده‌ها
آماده سازی داده‌ها
مدلسازی
ارزیابی
توسعه
انتخاب داده‌های مرتبط
گزارشی از داده‌های انتخاب شده مرتبط با موضوع
پاکسازی داده‌ها
گزارش پاکسازی داده‌ها
آماده کردن داده‌ها جهت داده‌کاوی
انتخاب ویژگیها
گزارش آماده سازی داده‌ها
مجتمع کردن داده‌ها
گزارش داده ای مجتمع شده
قالب بندی داده‌ها
گزارش داده‌های شکل داده شده


شناخت سیستم
شناخت داده‌ها
آماده سازی داده‌ها
مدلسازی
ارزیابی
توسعه
انتخاب تکنیک مدلسازی
مدلسازی با تکنیک انتخاب شده
مدلسازی فرضیات در مدل
ایجاد طرح هایی جهت تست
طراحی تست ها
ساخت مدل
تعیین پارامترهای مورد نیاز
مدل ها
توصیف مدل ساخته شده
ارزیابی مدل
گزارشی از مدل ارزیابی شده
بررسی و ارزیابی مجدد پارامترهای مدل


شناخت سیستم
شناخت داده‌ها
آماده سازی داده‌ها
مدلسازی
ارزیابی
توسعه
بررسی نتایج
بررسی و ارزیابی نتایج داده‌کاوی
بهبود مدل براساس نتایج
بررسی مجدد فرایند
بررسی مجدد نتایج فرایند
تعیین گام بعدی
تهیه لیستی از اعمال قابل انجام
تصمیم گیری


شناخت سیستم
شناخت داده‌ها
آماده سازی داده‌ها
مدلسازی
ارزیابی
توسعه
توسعه مدل
مدل و طرح توسعه
ایجاد طرح وارسی و نگهداشت
طرح وارسی و نگهداشت
تهیه گزارش نهایی
مستندات تجارب آموخته در طی انجام پروژه
بازنگری پروژه
گزارش نهایی
ارایه نهایی پروژه

دانشگاه سمنان
Semnan University
پردیس فرزانگان

ابزارها و زبان های برنامه نویسی داده کاوی

داده چیست؟

مجموعه ای از اشیاء داده‌ای و ویژگی های آنها

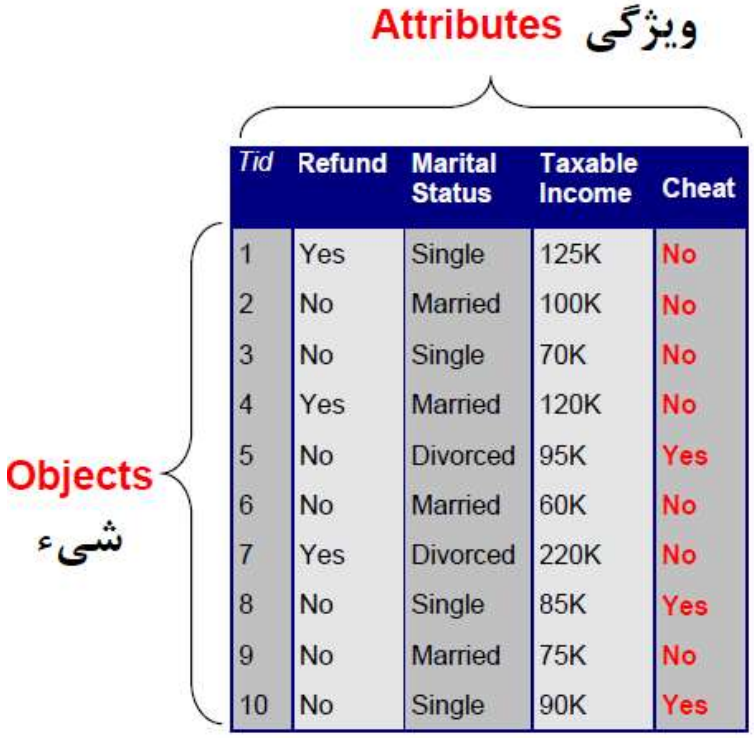
صفت یک ویژگی یا خصیصه یک شی است.

مثال: رنگ چشم یک فرد، درجه حرارت و غیره.

ویژگی به عنوان متغیر variable، میدان field، مشخصه characteristic یا ویژگی feature نیز شناخته می شود.

مجموعه ای از ویژگی ها یک شی را توصیف می‌کند.

شی نیز به عنوان رکورد record، نقطه point، مورد case، نمونه sample، موجودیت entity یا نمونه موردی instance شناخته می شود.



دانشگاه سمنان
Semnan University
پردیس فرزانگان

انواع صفات و ویژگی‌ها

انواع مختلفی از ویژگی‌ها وجود دارند:

✓ویژگی‌های دسته‌ای Categorical (این ویژگی ها کیفی هستند)

اسمیNominal: چنانچه دسته‌هایی که متغیر در آنها قرار می‌گیرد دارای هیچ گونه ترتیب طبیعی نباشد، هر یک از آن متغیرها را متغیر صوری یا اسمی کیفی می‌نامند. مثال: شماره شناسه، رنگ چشم، رنگ مو، وضعیت تاهل، کدهای پستی.

ترتیبی Ordinal: در یک جهت معنادار ارزش دارند (رتبه بندیِ) اما مقدار بین مقادیر متوالی شناخته شده نیست. بین مثال: رتبه‌بندی‌ها (مانند خوب، متوسط، بد یا مثلاً طعم چیپس سیب‌زمینی در مقیاس از ۱-۱۰)، قد در دسته‌بندی‌های {بلند، متوسط، کوتاه}

✓ویژگی‌های عددی Numeric (این ویژگی ها کمیتی هستند)

فاصله زمانی Interval: اندازه گیری در یک مقیاس از واحد های هم اندازه، مثال‌ها: تاریخ‌های تقویم، درجه حرارت بر حسب سانتی‌گراد یا فارنهایت.

نسبت Ratio مثال‌ها: دما بر حسب کلوین، طول، زمان، شمارش

✓مورد خاص: ویژگی‌های باینری Binary (بله/خیر، وجود دارد/وجود ندارد)

مقادیر صفات و ویژگی‌ها

مقادیر صفات، اعداد یا نمادهایی هستند که به یک ویژگی اختصاص داده شده‌اند.

تمایز بین صفات و مقادیر ویژگی

- همان ویژگی را می توان به مقادیر مشخصه‌های مختلف نگاشت کرد.

مثال: ارتفاع را می توان با فوت یا متر اندازه‌گیری کرد

- ویژگی‌های مختلف را می‌توان به یک مجموعه از مقادیر نگاشت کرد.

مثال: مقادیر مشخصه برای ID و age اعداد صحیح هستند. اما خصوصیات مقادیر مشخصه می تواند متفاوت باشد. شناسه محدودیت ندارد اما سن دارای حداکثر و حداقل مقدار است

خصوصیات مقادیر صفات

نوع یک صفت، بستگی به خواص زیر دارد:

تمایز Distinctness: = ≠

سفارش Order: < >

اضافه Addition: +

ضرب Multiplication: * /

- صفت اسمی: تمایز
- صفت ترتیبی: تمایز و نظم
- ویژگی فاصله: تمایز، ترتیب و اضافه
- ویژگی نسبت: هر ۴ ویژگی

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

صفات گسسته و پیوسته

صفت گسسته

- فقط مجموعه ای از مقادیر متناهی یا نامحدود دارد. مثال‌ها: کدهای پستی، شمارش یا مجموعه‌ای از کلمات در مجموعه‌ای از اسناد
- اغلب به عنوان متغیرهای عدد صحیح نشان داده می شود.
- توجه: ویژگی‌های باینری یک مورد خاص از ویژگی‌های گسسته هستند.

صفت پیوسته

- دارای اعداد واقعی به عنوان مقادیر ویژگی. مثال‌ها: دما، قد یا وزن.
- در عمل، مقادیر واقعی را می‌توان با استفاده از تعداد محدودی از ارقام اندازه‌گیری و نمایش داد.
- ویژگی‌های پیوسته معمولاً به عنوان متغیرهای ممیز شناور نشان داده می‌شوند.

دانشگاه سمنان
Semnan University
پردیس فرزانگان

انواع مجموعه های داده datasets

رکوردها Record:

- رکوردهای رابطه ای
- ماتریسی از داده ها مانند ماتریس عددی
- اسناد داده ای مانند متن ها
- داده تراکنش Transaction

گراف یا شبکه Graph:

- شبکه جهانی وب
- شبکه های اجتماعی یا اطلاعاتی
- ساختارهای مولکولی

داده های مرتب (Ordered)

- داده های ویدئویی: دنباله ای از تصاویر
- داده های وابسته به زمان Temporal: سری زمانی
- داده های متوالی Sequential: توالی تراکنش
- داده های توالی ژنتیکی

داده های فضایی (Spatial)، تصویری و چند رسانه ای:

- داده های مکانی: نقشه ها
- داده های تصویری
- داده های ویدیویی

داده رکورد

داده‌ای که شامل مجموعه‌ای از رکوردها است، که هر کدام از آن‌ها شامل یک مجموعه ثابت از ویژگی‌ها می‌باشد.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

ماتریس داده

اگر اشیاء داده‌ای دارای مجموعه ثابتی از ویژگی‌های عددی باشند، آنگاه می‌توان این اشیاء داده‌ای را به عنوان نقاطی در یک فضای چند بعدی در نظر گرفت که هر بعد نشان‌دهنده یک ویژگی متمایز است.

چنین مجموعه داده‌ای می‌تواند توسط یک ماتریس m در n نمایش داده شود، جایی که m تعداد ردیف‌ها (هر ردیف نشان‌دهنده یک شیء) و n تعداد ستون‌ها (هر ستون نشان‌دهنده یک ویژگی) است.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

دانشگاه سمنان
Semnan University
پردیس فرزانگان

داده سند

هر سند به یک بردار ترم term تبدیل می شود، هر عبارت جزء (ویژگی) بردار است، مقدار هر جزء تعداد دفعاتی است که عبارت مربوطه در سند رخ می دهد.

نمایش Bag of word

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

داده تراکنش

نوع خاصی از داده های رکورد، که در آن:

- هر رکورد (معامله) شامل مجموعه ای از آیتم است. مجموعه‌ای از آیتم‌ها را می‌توان به‌عنوان یک بردار باینری نیز نمایش داد که در آن هر ویژگی یک آیتم است.
- به عنوان مثال، یک فروشگاه مواد غذایی را در نظر بگیرید. مجموعه ای از محصولات خریداری شده توسط یک مشتری در طی یک سفر خرید، یک معامله را تشکیل می دهد، در حالی که تک تک محصولات خریداری شده، اقلام هستند.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

دانشگاه سمنان
Semnan University
پردیس فرزانگان

داده مرتب

سری زمانی

دنباله ای از مقادیر عددی مرتب شده (در "زمان").

داده مرتب

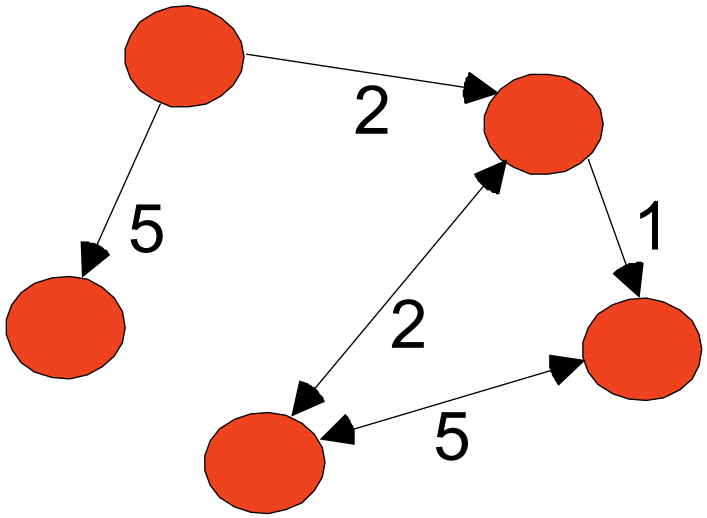
- داده های توالی ژنومی
- داده یک رشته مرتب شده طولانی است

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

دانشگاه سمنان
Semnan University
پردیس فرزانگان

داده گراف یا شبکه

مثال: نمودار وب و پیوندهای HTML

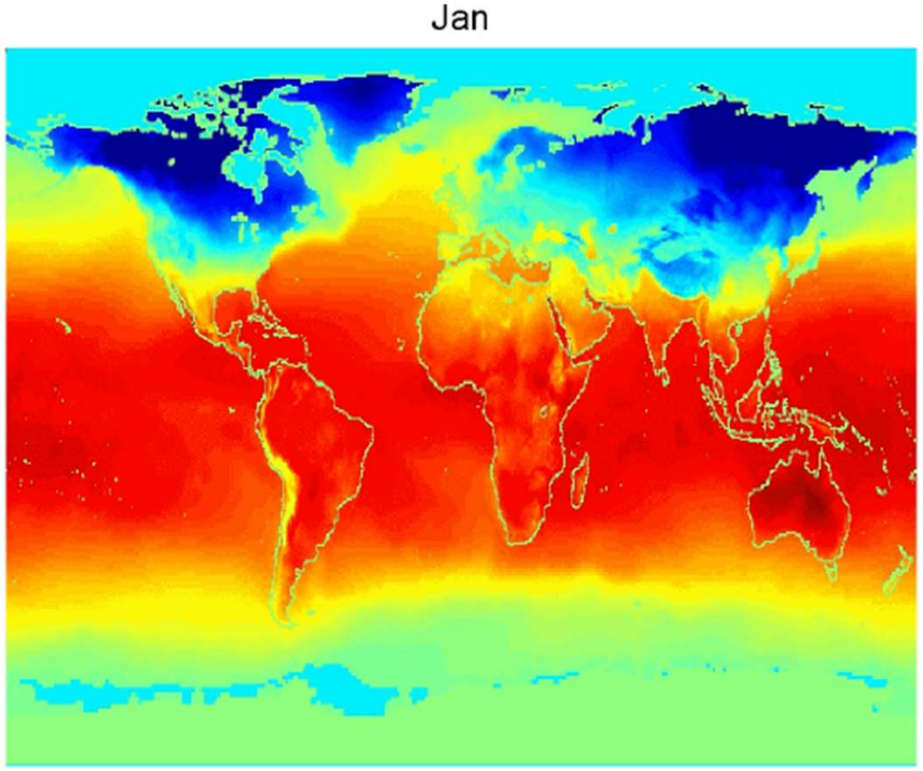



```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

دانشگاه سمنان
Semnan University
پردیس فرزانگان

داده‌های فضایی و چند رسانه ای

میانگین دمای ماهانه خشکی و اقیانوس



دانشگاه سمنان
Semnan University
پردیس فرزانگان

توصیف آماری داده‌ها

برای خلاصه سازی داده ها یا بدست آوردن یک الگوی خاص از روابط آماری استفاده می شود.

دو نوع شاخص آماری وجود دارد:

- شاخص مرکزی

میانگین ، میانه ، مد

- شاخص پراکندگی

دامنه تغییرات و میانگین انحرافات ، انحراف چارکی ، واریانس و انحراف استاندارد

میانگین

یکی از معروف ترین شاخص های مرکزی میانگین است، که انواع آن عبارت است از:
میانگین حسابی ، میانگین وزن دار ، میانگین هندسی و میانگین هارمونیک.

میانگین حسابی

$$\mu_A = \frac{x_1 + x_2 + \ldots + x_n}{n} = 1/n \sum_{i=1}^{n} x_i$$

چنانچه داده‌ها دارای وزن یکسانی (تأثیر یکسانی) نباشند، برای محاسبه‌ی میانگین حسابی آن باید هر یک از آنها را در وزن خود ضرب و حاصل‌جمع نهایی را تقسیم بر مجموع وزن‌ها کنیم. این میانگین به میانگین وزنی نیز شناخته می‌شود.

میانگین وزن دار

$$\mu_W = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i}$$

میانگین

میانگین هندسی

$$\mu_G = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

از میانگین هندسی موقعی

استفاده می‌شود که صحبت از نرخ رشد یک ویژگی مطرح باشد.

برای مثال جهت محاسبه‌ی نرخ رشد جمعیت، نرخ رشد سود و نرخ رشد تولید به این نوع میانگین رجوع می‌شود.

میانگین هارمونیک عبارت است از تقسیم تعداد داده‌ها بر مجموع معکوس مقادیر موجود در داده‌ها

$$\mu_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} 1/x_i}$$

بطور معمول هنگامی که واحد اندازه‌گیری داده‌ها ترکیبی باشد، از این نوع میانگین استفاده می‌شود. برای مثال اگر یک وسیله‌ی نقلیه فاصله‌ی میان دو شهر را در مرحله‌ی رفت با یک سرعت و در برگشت با سرعت دیگری طی کند، برای محاسبه‌ی میانگین سرعت آن از میانگین هارمونیک استفاده می‌کنیم.

$$\mu_H \leq \mu_G \leq \mu_A$$

میانه

میانه عددی است که توزیع داده ها را به دو قسمت مساوی تقسیم می کند، به نحوی که نیمی از داده ها بزرگتر و نیم دیگر کوچکتر از آن هستند.

در مجموعه اعداد مرتب شده داده ی میانی به عنوان میانه لحاظ می گردد. در صورتی که تعداد داده ها زوج باشد، نصف مجموع دو داده ای که در وسط قرار دارند، میانه محسوب می شود.

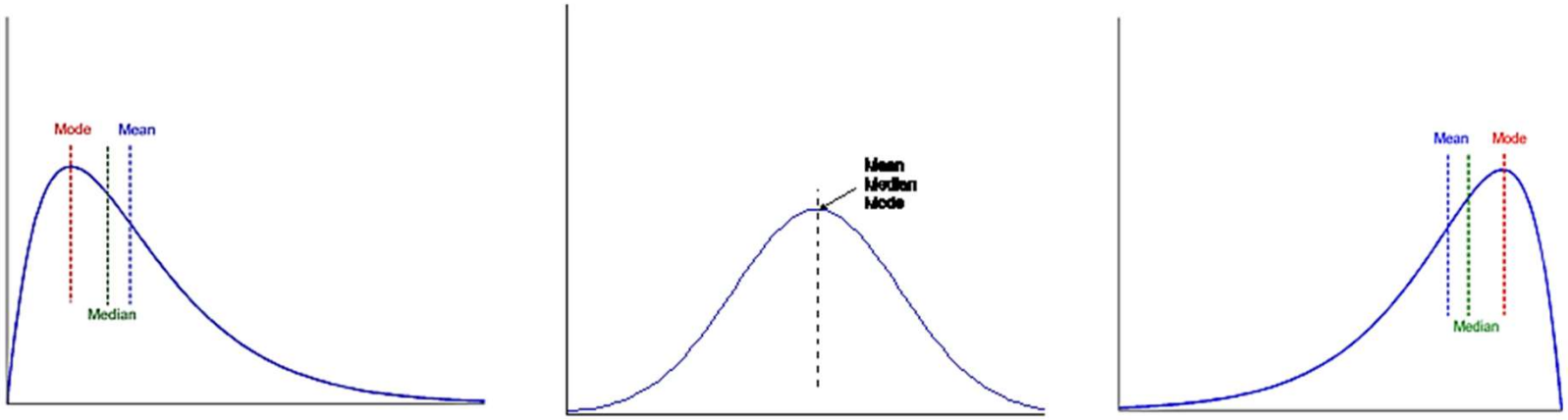
میانه برای داده های پیوسته ی دسته بندی شده از فرمول زیر محاسبه می شود :

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

دانشگاه سمنان
Semnan University
پردیس فرزانگان

مد

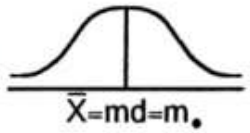
داده ای که فراوانی آن از سایر داده ها بیشتر باشد را نما یا مد می نامیم. اگر دو داده ی مجاور دارای بیشترین فراوانی باشند، نصف مجموع آن ها به عنوان مد انتخاب می شود، در غیر این صورت اگر دو داده مجاور نباشند هر دو به عنوان مد انتخاب می شوند.



کجی مثبت　　توزیع متقارن　　کجی منفی

۱- اگر میانگین و میانه و نما با هم برابر باشد، توزیع متقارن است و کجی آن صفر است.

$$\overline{X} = md = mo$$

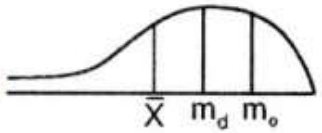


۲- اگر میانگین بزرگتر از میانه و میانه بزرگتر از نما باشد کجی مثبت است، یعنی اکثر نمرات پایین بوده و امتحان سخت است.

$$\overline{X} > md > m_o$$

۳- اگر میانگین کوچکتر از میانه و میانه کوچکتر از نما باشد، کجی منفی است و اکثر نمرات بالاست و امتحان آسان بوده است.

$$\overline{X} < md < mo$$



شاخص های پراکندگی

دامنه تغییرات : تفاوت میان بزرگترین و کوچکترین داده را به عنوان دامنه ی تغییرات داده ها می شناسیم.

$$R=Max(x_1,x_2,\ldots,x_n)-Min(x_1,x_2,\ldots,x_n)$$

میانگین انحرافات : به منظور دخالت تاثیر تمام داده ها می توان فاصله ی کلیه داده ها با میانگین را محاسبه نمود، که به آن انحراف از میانگین داده ها گوییم.

$$D=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\mu_A\right|$$

شاخص های پراکندگی (ادامه)

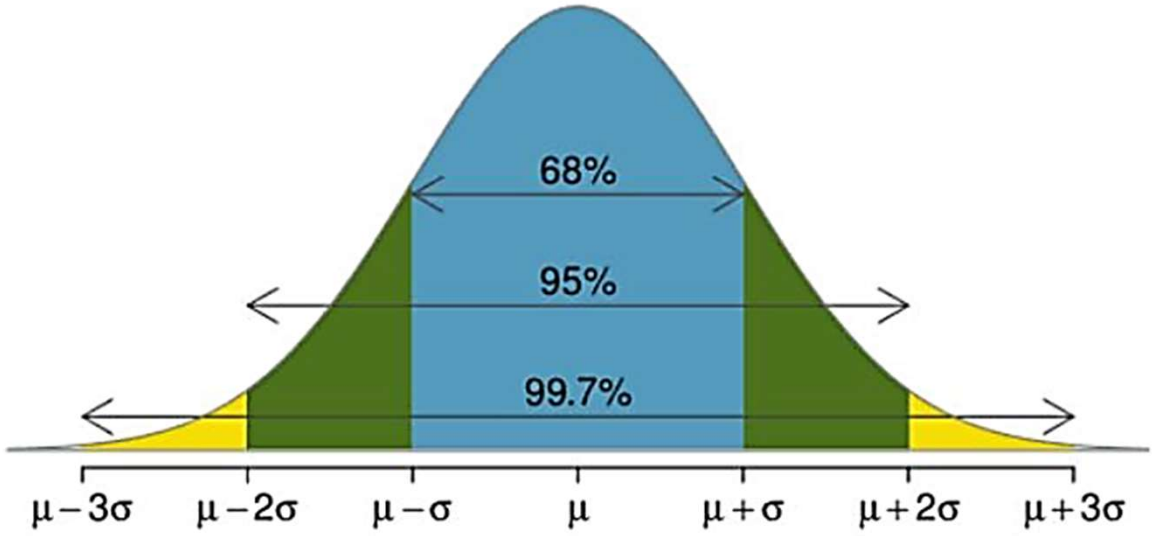
واریانس : میانگین مجذور انحرافات را واریانس گویند.

$$var(x) = \frac{1}{m}\sum_{i=1}^{m}(x - \bar{x})^2$$

انحراف استاندارد : با جذر گرفتن از مقدار واریانس انحراف استاندارد بدست می آید.

$$\sigma(x) = \sqrt{var(x)}$$

توزیع نرمال



$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

منحنی نرمال (توزیع). (μ: میانگین، σ: انحراف معیار)

از μ–σ تا σ+μ: شامل حدود ۶۸ درصد از اندازه گیری‌ها

از μ-2σ تا μ+2σ: شامل حدود ۹۵ درصد از آن است.

از μ-3σ تا μ+3σ: حدود ۹۹.۷ درصد آن را شامل می شود.

کواریانس

اندازه ی تغییرات هماهنگ دو متغیر تصادفی را کواریانس گویند. اگر دو متغیر یکی باشند، کواریانس برابر با واریانس خواهد بود. برای مقادیر ویژگی های x و y مقدار کواریانس برابر است با :

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

اگر این مقدار صفر باشد، میان دو ویژگی همبستگی وجود ندارد و دو ویژگی دارای رابطه خطی نیستند. مقدار مثبت آن نشان می دهد با افزایش یکی، دیگری نیز افزایش میابد و مقدار منفی دلالت بر این دارد که افزایش یکی باعث کاهش دیگری می شود.

برای داده ها با d بعد، یک ماتریس $d \times d$ به نام ماتریس کواریانس ویژگی داریم که:

$$S_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

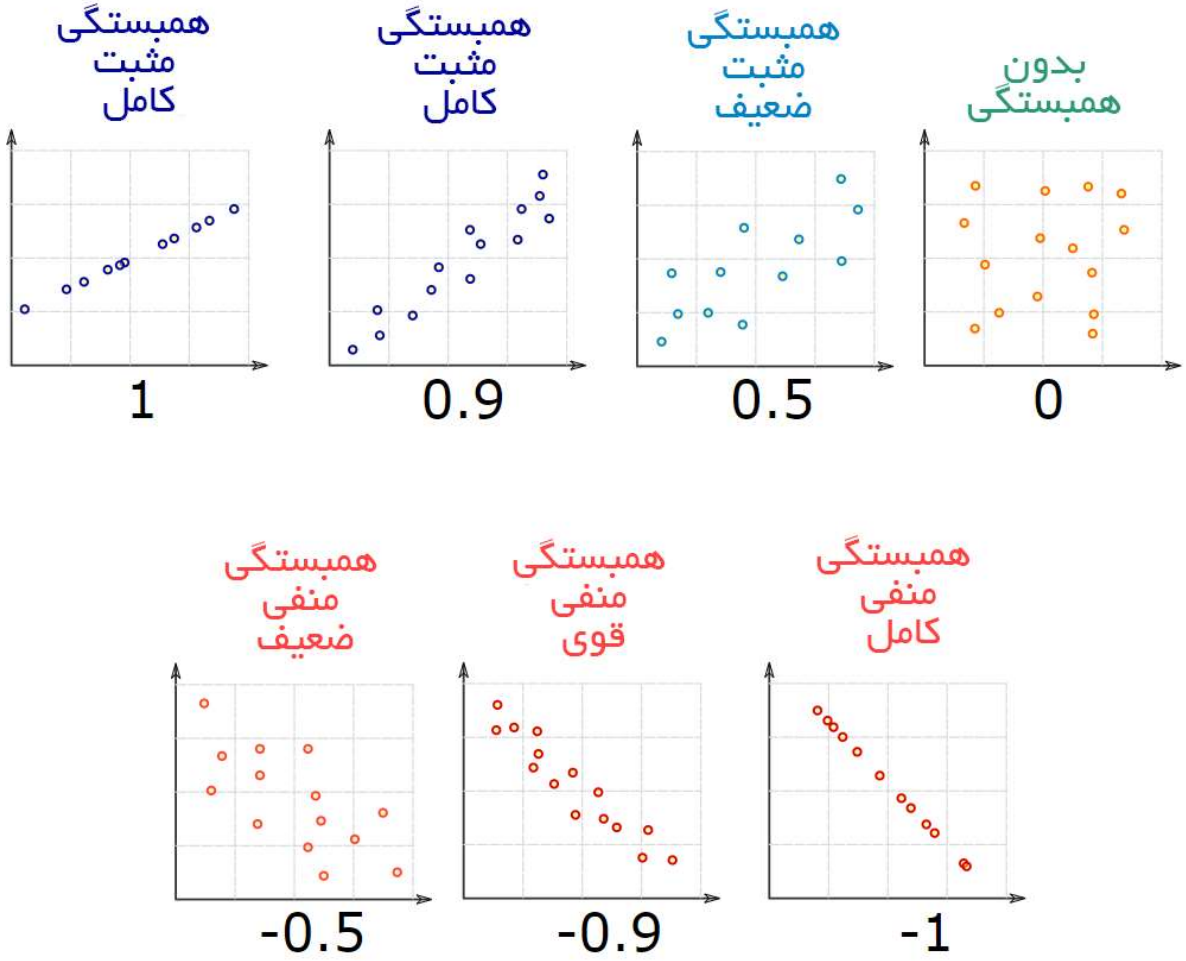
همبستگی

همبستگی به رابطه بین دو متغیر اشاره می کند، که می توان آن را با کمک نمودار پراکندگی نشان داد. ارزش مقداری آن را ضریب همبستگی می نامیم.

$$cor(x,y) = \frac{cov(x,y)}{\sqrt{cov(x,x)cov(y,y)}} = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

ضریب همبستگی دارای دامنه ای بین ۱ و ۱- است.

مقدار صفر عدم همبستگی را نشان می دهد و مقادیر۱ و ۱- به ترتیب دلالت بر همبستگی کامل مثبت و همبستگی کامل منفی برای دو ویژگی فوق را دارد.


همبستگی
مثبت
کامل
1
همبستگی
مثبت
کامل
0.9
همبستگی
مثبت
ضعیف
0.5
بدون
همبستگی
0
همبستگی
منفی
ضعیف
-0.5
همبستگی
منفی
قوی
-0.9
همبستگی
منفی
کامل
-1

شباهت و عدم شباهت

برای بسیاری از مسائل مختلف، نیاز داریم که میزان نزدیکی دو شیء را اندازه‌گیری کنیم.

مثال‌ها:

برای یک آیتم خریداری‌شده توسط مشتری، یافتن آیتم‌های مشابه دیگر.

گروه‌بندی مشتریان یک سایت به‌طوری که مشتریان مشابه، تبلیغات یکسانی را مشاهده کنند.

گروه‌بندی اسناد وب به‌طوری که بتوانید اسنادی که درباره سیاست صحبت می‌کنند را از آن‌هایی که درباره ورزش صحبت می‌کنند، جدا کنید.

یافتن تمام اسناد وب نزدیک به هم که تقریباً تکراری هستند.

برای حل این مسائل، نیاز به یک تعریف از شباهت یا فاصله داریم. این تعریف به نوع داده‌هایی که داریم بستگی دارد.

شباهت

اندازه‌گیری عددی برای بیان میزان مشابهت دو شیء داده‌ای.
مقدار بالاتر نشان‌دهنده شباهت بیشتر بین اشیاء است.
اغلب در محدوده [۰,۱] قرار دارد.

عدم شباهت (برای مثال، فاصله)

اندازه‌گیری عددی برای بیان میزان تفاوت دو شیء داده‌ای.
مقدار کمتر نشان‌دهنده شباهت بیشتر بین اشیاء است.
حداقل عدم شباهت معمولاً ۰ است.

ماتریس داده و ماتریس عدم شباهت

Proximity (مجاورت): به شباهت یا عدم شباهت اشاره دارد.

اندازه‌گیری‌های شباهت اغلب می‌توانند به‌عنوان تابعی از اندازه‌گیری‌های عدم شباهت بیان شوند. $sim(i, j) = 1 - d(i, j)$

برای مثال، برای داده‌های اسمی:

ماتریس داده:

n نقطه داده با ابعاد p

دارای دو حالت است: سطرها برای اشیا و ستون ها برای ویژگی ها

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

ماتریس عدم شباهت:

n نقطه داده اما تنها فاصله درج می شود.

یک ماتریس مثلثی است.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

معیار شباهت برای ویژگی های اسمی

روش تطبیق ساده: عدم شباهت بین دو شی i و j را می توان بر اساس نسبت عدم تطابق محاسبه کرد.

$$d(i, j) = \frac{p - m}{p}$$

m: تعداد موارد منطبق، p: تعداد کل متغیرها

به همین ترتیب، شباهت می تواند به صورت زیر محاسبه شود:

$$sim(i, j) = 1 - d(i, j) = \frac{m}{p}$$

مثال:

فرض کنید که داده های نمونه جدول مقابل را داریم، با این تفاوت که فقط شناسه شی و ویژگی test-1 موجود است و test-1 اسمی است.

از آنجایی که در اینجا یک ویژگی اسمی داریم، test-1، p=1 را تنظیم می کنیم.

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

از ماتریس مقابل می بینیم که همه اشیاء غیر مشابه هستند به جز اشیاء ۱ و ۴

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

معیار شباهت برای ویژگی های باینری

یک ویژگی باینری تنها یکی از دو حالت را دارد: ۰ و ۱، که در آن ۰ به معنای وجود ندارد و ۱ به معنای وجود آن است.

اگر تصور شود که همه صفات دودویی دارای وزن یکسانی هستند، جدول احتمالی را خواهیم داشت که در آن:

جدول احتمال برای ویژگی های باینری

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

q تعداد صفاتی است که برای هر دو شی i و j برابر با ۱ است.

r تعداد صفاتی است که برای شی i برابر با ۱ و برای شی j برابر ۰ است.

s تعداد صفاتی است که برای شی i برابر با ۰ و برای شی j برابر با ۱ است.

t تعداد صفاتی است که برای هر دو شی i و j برابر ۰ است.

تعداد کل صفات p است که $p = q + r + s + t$

اندازه‌گیری فاصله برای متغیرهای دودویی متقارن

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

برای ویژگی های باینری نامتقارن، این دو حالت به یک اندازه مهم نیستند، مانند نتایج مثبت (۱) و منفی (۰) یک آزمایش بیماری.

اندازه‌گیری فاصله برای متغیرهای دودویی نامتقارن

$$d(i, j) = \frac{r + s}{q + r + s}$$

زمانی که هر دو یک باشند مهمتر است از زمانی که هر دو صفر باشند.

بنابراین t بی اهمیت در نظر گرفته می شود.

معیار شباهت برای ویژگی های عددی

اندازه‌گیری‌های فاصله که معمولاً برای محاسبه معیارشباهت اشیاء توصیف‌شده توسط ویژگی‌های عددی استفاده می‌شوند، شامل فواصل اقلیدسی، منهتن و مینکوفسکی می باشند.

محبوب ترین اندازه گیری فاصله، فاصله اقلیدسی است. فرض کنید $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ و $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ دو شیء باشند که با صفات عددی p توصیف می شوند. فاصله اقلیدسی بین اجسام i و j به صورت تعریف می شود:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

یکی دیگر از معیارهای معروف فاصله، منهتن Manhattan است که به صورت زیر تعریف می شود:

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$$

فاصله مینکوفسکی تعمیم فاصله اقلیدسی و منهتن است:

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

به این فاصله نرم L-h نیز گفته می شود. اگر h=2 باشد فاصله اقلیدسی یا نرم L-2 و اگر h=1 باشد فاصله منهتن یا نرم L-1 خواهد بود.

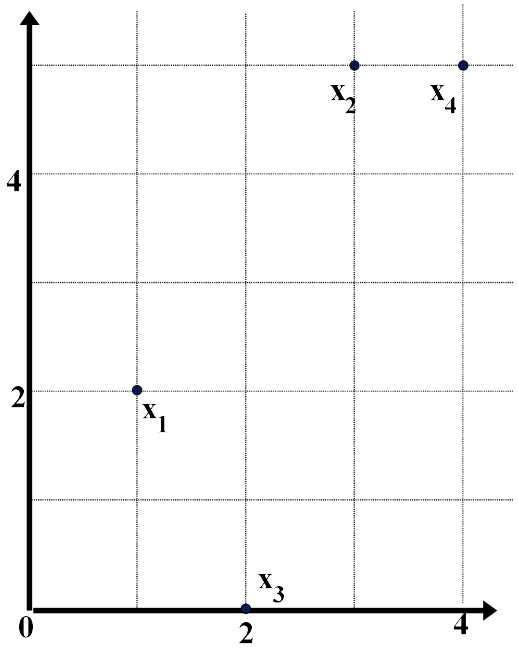
اگر $h \to \infty$ به آن فاصله supremum گفته می شود.

$$d(i, j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

برای محاسبه کافی است که بزرگترین اختلاف در ویژگی ها را در نظر بگیریم.

مثال

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

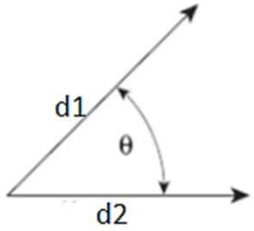
Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

فاصله کسینوسی



اگر d1 و d2 دو بردار باشند، فاصله کسینوسی آنها مانند زیر بدست می آید:

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$$

که در آن • ضرب نقطه ای بین دو بردار و ||d|| طول بردار را نشان می دهد.

شباهت کسینوسی معمولا برای مقایسه اسناد که بردارها بر اساس طول سند نرمال سازی شده اند، استفاده می شود.

مثال:

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

$d_1 =$ (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
$d_2 =$ (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$

$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$

$\cos(d_1, d_2) = 0.94$

فاصله همینگ

فاصله همینگ تعداد موقعیت‌هایی است که بردارهای بیتی در آن‌ها تفاوت دارند.
مثال:

p1 = 10101
p2 = 10011

d(p1, p2) = 2 به دلیل اینکه بردارهای بیتی در موقعیت‌های ۳ و ۴ با هم متفاوت‌اند.

نرم L1 برای بردارهای باینری

فاصله همینگ بین دو بردار از ویژگی‌های دسته‌بندی شده، تعداد موقعیت‌هایی است که در آن‌ها تفاوت دارند.
مثال:

x = (married, low income, cheat),
y = (single, low income, not cheat)
d(x,y) = 2