داده کاوی

فصل سوم

تهیه شده توسط
محمد پورزعفرانی
Pourzaferani.ir
دانشگاه اصفهان

دانشگاه اصفهان

تشکر ویژه

- خانم دهقان
- خانم تیماس
- آقای مفیدی
- آقای حیدری
- خانم طرقی
- آقای تیموری

فصل سوم: پیش‌پردازش داده‌ها

- پیش‌پردازش داده‌ها: خلاصه‌ای جامع
  - کیفیت داده‌ها (Data Quality)
  - عملیات اصلی در پیش‌پردازش داده‌ها
  - پاک‌سازی داده‌ها (Data Cleaning)
  - تجمیع داده‌ها (Data Integration)
  - کاهش داده‌ها (Data Reduction)
  - تغییرشکل (Transformation) و گسسته‌سازی (Discretization) داده‌ها
  - خلاصه فصل

کیفیت داده: چرا داده‌ها را پیش‌پردازش میکنیم؟

- وضعیت پایگاه داده‌های کنونی: حاوی داده‌های نویزی، داده‌های مفقود، داده‌های ناسازگار
- داده‌های بی‌کیفیت منجر به کاوش‌های بی‌کیفیت خواهد شد
- **معیارهای کیفیت داده:**
  - **دقت**: **(Accuracy)** درست یا غلط، دقیق یا نه
  - **کامل بودن**: **(Completeness)** ثبت نشده، غیر قابل دسترسی، ...
  - **سازگاری**: **(Consistency)** برخی از داده‌ها اصلاح شده‌اند اما نه همه، آویزان،...
  - **به‌هنگام بودن**: **(Timeliness)** داده‌ها به موقع به روز رسانی شده‌اند؟
  - **باورپذیری**: **(Believability)** اطمینان از درستی داده‌ها
  - **قابلیت تفسیر**: **(Interpretability)** آیا داده‌ها به سادگی قابل درک هستند؟

وظایف عمده در پیش‌پردازش داده‌ها

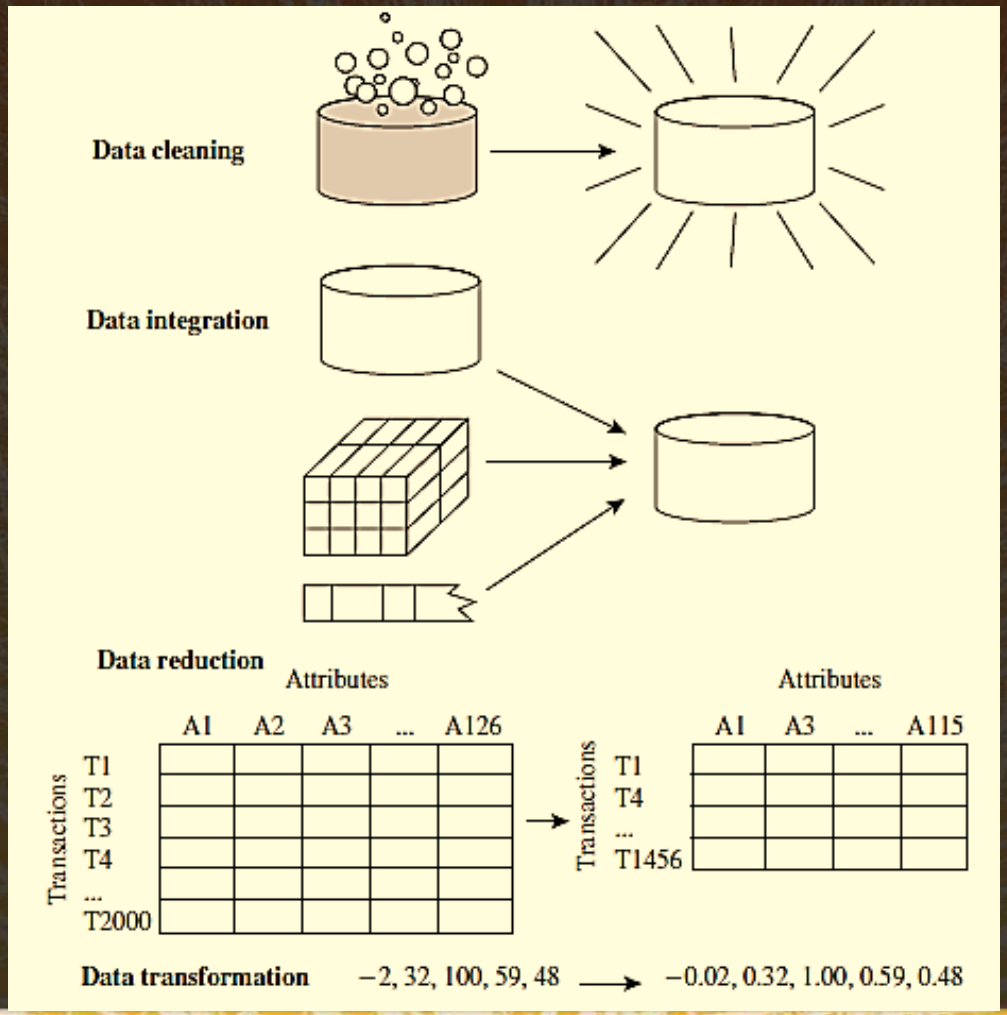
1) پاک‌سازی داده‌ها
- پر کردن مقادیر گم شده
- اصلاح داده‌های نویزی
- شناسایی و حذف داده‌های دورافتاده
- و از بین بردن تناقضات

2) تجمیع داده‌ها
- یک مفهوم ممکن است در پایگاه‌های مختلف نامهای مختلفی داشته باشد که منجر به ناسازگاری و تکرار میشود

3) کاهش داده‌ها
- کاهش ابعاد (Dimensionality reduction)
- کاهش تکثر (Numerosity reduction)
- فشرده‌سازی داده‌ها (compression)

4) تغییرشکل داده‌ها
- نرمال سازی
- گسسته‌سازی
- تولید سلسله مراتب مفاهیم

انواع پیش‌پردازش‌ها



دانشگاه اصفهان

فصل سوم: پیش‌پردازش داده‌ها

- پیش‌پردازش داده‌ها: خلاصه‌ای جامع
  - کیفیت داده‌ها (Data Quality)
- عملیات اصلی در پیش‌پردازش داده‌ها
  - پاک‌سازی داده‌ها (Data Cleaning)
  - تجمیع داده‌ها (Data Integration)
  - کاهش داده‌ها (Data Reduction)
  - تغییرشکل (Transformation) و گسسته‌سازی (Discretization) داده‌ها
  - خلاصه فصل

دانشگاه اصفهان

پاکسازی داده‌ها

- داده‌ها در دنیای واقعی نادرست هستند: بسیاری از داده‌ها به طور بالقوه نادرست اند، دلایل: ابزار معیوب و ناقص، خطای انسانی یا کامپیوتری، خطای انتقال
  - داده‌های ناکامل: (incomplete) ویژگی مقدار ندارد، فقدان ویژگی‌های مورد علاقه یا تنها دارای داده‌های تجمیعی
    - به عنوان مثال : *شغل*= "" (فاقد داده)
  - داده‌های نویز: (noisy) حاوی نویز، اشتباهات یا داده‌های پرت
    - به عنوان مثال : *حقوق*="-10" (خطا)
  - ناسازگاری در داده‌ها: (inconsistent) شامل اختلاف در کد یا نام، به عنوان مثال
    - *سن*= " ۴۲ " , *تاریخ تولد*= "۱۳۸۹/۴/۲۳"
    - دسته‌بندی قبلی "1, 2, 3", دسته‌بندی فعلی "A, B, C"
    - تفاوت بین رکوردهای تکراری
  - عمدی: (Intentional) (پنهان کردن داده‌های از دست رفته)
    - ۱ ژانویه به عنوان روز تولد همه؟

داده‌های ناقص (Missing Values)

- داده‌ها همیشه در دسترس نیستند
  - مثال:بسیاری از رکوردها، مقدار ثبت شده‌ای برای ویژگی‌های مختلف ندارند
- گم‌شدن داده‌ها به دلیل:
  - نقص تجهیزات
  - به دلیل ناسازگاری با دیگر داده‌های ثبت شده حذف شده‌اند
  - داده‌ها به دلیل سوء تفاهم وارد نشده‌اند
  - داده‌های خاصی ممکن است در زمان ورود اطلاعات، مهم قلمداد نشوند
  - نبود تاریخ ثبت یا تغییر داده‌ها
- داده‌های از دست رفته ممکن است به استنباط نیاز داشته باشند

چگونگی رفتار با داده‌های Miss؟

1) نادیده گرفتن رکورد: معمولا زمانیکه برچسب کلاس موجود نباشد، انجام می شود (در زمان دسته‌بندی)—زمانی که درصد مقادیر از دست رفته برای هر ویژگی، بطور قابل توجهی متفاوت باشد، موثر نیست

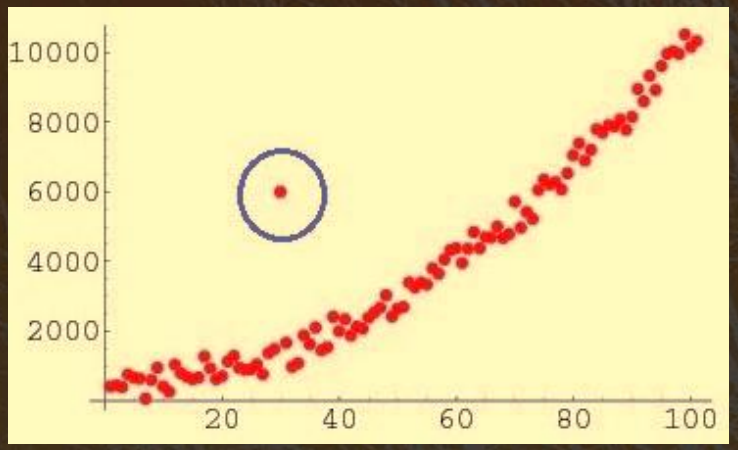
2) پر کردن مقادیر از دست رفته به صورت دستی :خسته کننده +غیرعملی؟

3) پر کردن خودکار با روش‌های زیر:

- یک ثابت جهانی: به عنوان مثال، نا شناخته"unknown"، یک کلاس جدید!
- میانگین ویژگی (به عنوان مثال داده‌های نرمال از میانگین و داده‌های چوله از میانه استفاده شود)
- میانگین ویژگی برای همه نمونه‌های متعلق به همان کلاس: هوشمندانه‌تر
- محتمل ترین مقدار: مبتنی بر استنتاج همچون فرمول های بیزین یا درخت تصمیم

داده‌های نویزی

- نویز :خطای تصادفی یا واریانس در یک متغیر اندازه‌گیری شده
- مقادیر نادرست ویژگی به دلیل
  - نقص در ابزار جمع‌آوری داده‌ها
  - مشکلات ورود داده‌ها
  - مشکلات انتقال داده‌ها
  - محدودیت تکنولوژی
  - تناقض در قراردادهای نامگذاری
- سایر مشکلات داده‌ها، نیاز به پاک‌کردن داده‌ها دارند
  - رکورد تکراری
  - اطلاعات ناقص
  - اطلاعات متناقض



چگونگی برخورد با داده‌های نویزی

1) Binning

- ابتدا داده‌ها مرتب شده و تقسیم به دسته‌های با فرکانس تکرار یکسان(equal-frequency) می شوند.
  - داده‌های طبقه‌بندی شده برای قیمت (به دلار): ۴، ۸، ۹، ۱۵، ۲۱، ۲۱، ۲۴، ۲۵، ۲۶، ۲۸، ۲۹، ۳۴

* **تقسیم‌بندی داده‌ها با فراوانی یکسان bins:**

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 2: 26, 28, 29, 34

* **هموارسازی توسط میانگین:**

- Bin 1 : 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3 : 29, 29, 29, 29

* **هموارسازی توسط میانه مرزها:**

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

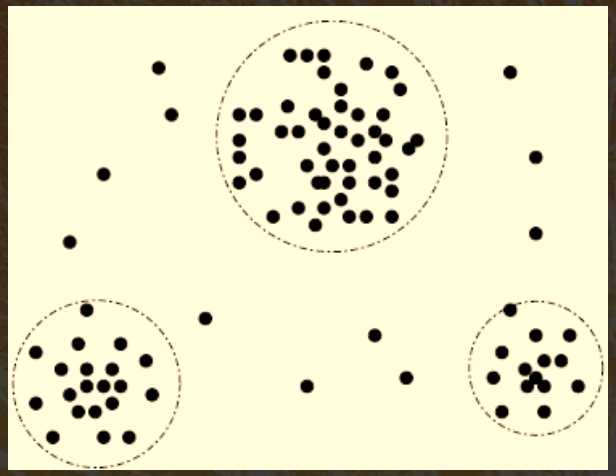
چگونگی برخورد با داده‌های نویزی

2) رگرسیون (Regression)

– هموار کردن داده‌ها به کمک سازگار کردن آن‌ها با تابع رگرسیون

3) خوشه‌بندی

– شناسایی و حذف داده‌های پرت

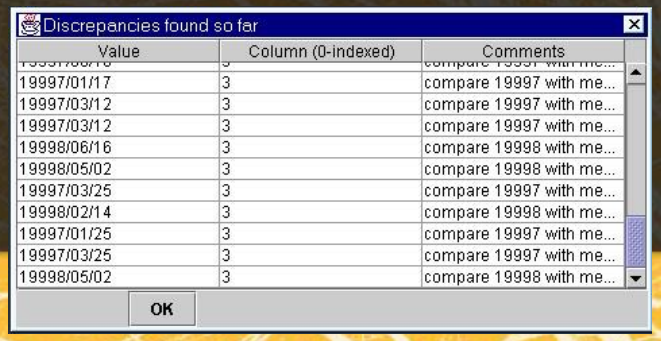


4) ترکیب بازرسی کامپیوتر و انسان

– تشخیص مقادیر مشکوک بصورت خودکار سپس بررسی آن‌ها توسط عامل انسانی (به عنوان مثال تعامل با داده‌های دورافتاده)

پاکسازی داده‌ها به عنوان یک رویه

- تشخیص اختلاف داده‌ها (Data discrepancy detection)
  - استفاده از فراداده Use metadata (به عنوان مثال، دامنه، محدوده، وابستگی، توزیع)
  - بررسی قانون منحصر به فرد بودن، قانون توالی و قانون null
  - استفاده از ابزارهای تجاری
    - تمیزکردن داده: استفاده از دانش زمینه(به عنوان مثال، کد پستی، املای کلمات) جهت تشخیص خطاها و ایجاد اصلاحات
    - حسابرسی داده‌ها: با تجزیه و تحلیل داده‌ها برای کشف قوانین و ارتباط برای شناسایی متخلفان (مثال استفاده از همبستگی و خوشه‌بندی برای پیدا کردن داده‌های پرت)
- مهاجرت و یکپارچه‌سازی داده‌ها (Data migration and integration)
- ابزار مهاجرت داده‌ها: اجازه تحولاتی برای خاص شدن
  - ابزارهای ETL (استخراج/ انتقال/ بارگذاری): اجازه دادن به کاربران برای تعیین تبدیلات از طریق یک رابط کاربر گرافیکی
- تجمیع دو فرایند (Integration of the two processes)
  - تکراری و تعاملی (به عنوان مثال Potter’s wheel)



Value	Column (0-indexed)	Comments
19997/01/17	3	compare 19997 with me...
19997/03/12	3	compare 19997 with me...
19997/03/12	3	compare 19997 with me...
19998/06/16	3	compare 19998 with me...
19998/05/02	3	compare 19998 with me...
19997/03/25	3	compare 19997 with me...
19998/02/14	3	compare 19998 with me...
19997/01/25	3	compare 19997 with me...
19997/03/25	3	compare 19997 with me...
19998/05/02	3	compare 19998 with me...

فصل سوم: پیش‌پردازش داده‌ها

- پیش‌پردازش داده‌ها: خلاصه‌ای جامع
  - کیفیت داده‌ها (Data Quality)
  - عملیات اصلی در پیش‌پردازش داده‌ها
- پاک‌سازی داده‌ها (Data Cleaning)
- تجمیع داده‌ها (Data Integration)
- کاهش داده‌ها (Data Reduction)
- تغییرشکل (Transformation) و گسسته‌سازی (Discretization) داده‌ها
- خلاصه فصل

تجمیع داده‌ها

- **ادغام داده‌ها**
  - ترکیب داده‌ها از منابع متعدد در یک انباره منسجم
- ادغام شماها (Schema integration) : مانند B.cust =A.cust-id
  - ادغام ابرداده‌ها (metadata) از منابع مختلف

1) مشکل شناسایی موجودیت **(Entity identification problem)**
  - باید مشخص شود که یک ویژگی از منبع اول معادل کدام ویژگی از منبع دوم است
    - در یک منبع مقادیر به صورت H و S است در منبع دوم 1 و 0
  - باید مشخص شود دو منبع موجودیت‌هایی را شناسایی کنیم که هر دو یک چیز را توصیف میکنند
    - Bill Clinton = William Clinton
- تشخیص و برطرف نمودن ناسازگاری مقادیر داده
  - برای یک موجودیت دنیای واقعی، مقادیر ویژگی از منابع مختلف، متفاوت است.
  - دلایل احتمالی: نمایش‌های متفاوت، مقیاس های مختلف، به عنوان مثال، متریک در مقابل واحد بریتانیا

بررسی افزونگی در تجمیع داده‌ها(۲)

2) **افزونگی داده‌ها اغلب هنگامی رخ می دهد که پایگاه داده‌های متعدد ادغام می گردند.**

- شناسایی شی: ویژگی‌ها یا اشیا ممکن است نام های مختلف در پایگاه داده‌های مختلف داشته باشند.
- داده قابل اشتقاق: یکی ویژگی ممکن است مشتق از یک ویژگی در جدولی دیگر باشد. به عنوان مثال، درآمد سالانه
- ویژگی‌های افزونه را می‌توان با تحلیل همبستگی (correlation) یا کوواریانس(covariance) تشخیص داد

بررسی افزونگی در تجمیع داده‌ها

- برخی از افزونگی‌ها را می توان با تحلیل همبستگی شناسایی نمود.
- با داشتن دو ویژگی، چنین تحلیلی می تواند میزان وابستگی دو ویژگی را نسبت به هم بسنجد

a) برای داده‌های اسمی، ما از تست (کای-دو) χ^2 استفاده می کنیم.

b) برای ویژگی‌های عددی، ما استفاده می کنیم از correlation coefficient و کواریانس، هر دو نحوه‌ی ارتباط ویژگی‌ها را نمایان می‌سازند

تحلیل همبستگی(داده‌های اسمی)

- ابتدا جدول Contingency را تکمیل می‌کنیم به این صورت که در سطر مقادیر اسمی یک ویژگی و در ستون مقادیر ویژگی دوم قرار می‌گیرند
- خانه‌های جدول تعداد حالات موجود در داده‌ها را نشان می‌دهند
- به طور مثال در جدول زیر ۲۵۰ حالت (fiction,male) داریم که به این اعداد فراوانی مشاهده شده observed frequency می‌گویند و با Oij نشان می‌دهند

	male	*female*	*Total*
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Note: Are *gender* and *preferred_reading* correlated?

- اعداد داخل پرانتز، فراوانی مورد انتظار (expected frequence) هستند که به روش زیر محاسبه می‌شوند

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n},$$

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90,$$

تست کای-دو

- حال بر اساس این جدول **تست** χ^2 محاسبه می‌شود

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

	male	female	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Note: Are gender and preferred_reading correlated?

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

این عدد به چه معناست؟

تست کای-دو

- برای تست کای-دو لازم است دو مورد را تعیین کنیم
  - درجه آزادی (degree of freedom)
  - سطح اطمینان (Significance level)
- درجه آزادی برابر است با (r-1)*(c-1) که c و r تعداد مقادیر متفاوت ویژگیهای هستند
- برای جدول ۲x۲ درجه آزادی بدین ترتیب محاسبه می گردد.
  - (۲-۱)(۲-۱)
- سطح اطمینان معین می‌کند با چه درصدی اطمینان داریم دو ویژگی مرتبطند یا خیر
- نکته: اگر مقدار محاسبه شده از مقدار جدول بیشتر بود نشان میدهد دو ویژگی به یکدیگر مرتبط هستند
- برای درجه آزادی ۱، مقدار χ^2 مورد نیاز برای رد فرضیه در سطح ۰.۰۰۱ برابر ۱۰.۸۲۸ است.
- این نشان می دهد که gender و preferred reading در مجموعه مورد بررسی وابسته هستند.

تست کای-دو

- هر چه مقدار χ^2 بیشتر باشد، احتمال وابستگی آنها بیشتر است.

- وابستگی دلالت بر رابطه علت و معلولی ندارد
  - تعداد بیمارستان ها و تعداد سرقت خودرو در شهرستان ها با هم مرتبطند.
  - هر دو به دلیلی به متغیر سوم مرتبطند: جمعیت

جدول توزیع کای دو

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

تحلیل همبستگی(داده عددی)

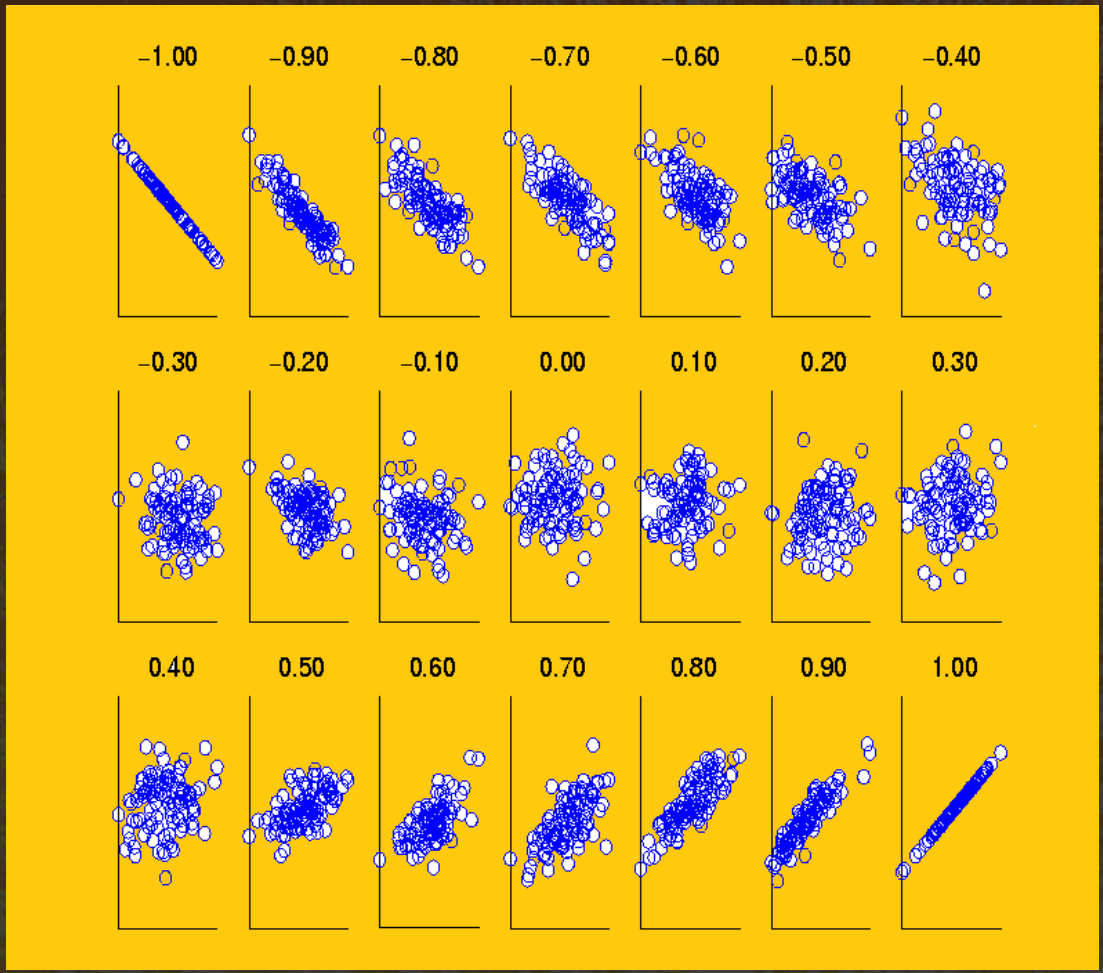
- ضریب همبستگی (Correlation coefficient)

همچنین (Pearson's product moment coefficient) نامیده می شود.

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

- که در آن n تعداد تاپل است
- $\bar{A}$ و $\bar{B}$ به ترتیب میانگین A و B هستند.
- **σA و σB** به ترتیب انحراف معیار استاندارد A و B می باشند
- **Σ(aibi)** مجموع حاصل ضرب خارجی (Cross-product) A و B است.
- -۱< rA,B <+۱
- **اگر $r_{A,B} > 0$ باشد آنگاه همبستگی A و B مثبت است.(مقدار A با افزایش B افزایش می یابد)**
- **اگر rA,B = 0 باشد آنگاه A و B مستقلند.**
- **اگر rA,B < 0 باشد آنگاه همبستگی A و B منفی دارند.**

ارزیابی تصویری همبستگی



نمودارهای پراکندگی از ۱- تا ۱+ نشان داده شده است.

کواریانس(داده عددی)

کواریانس مشابه همبستگی است.

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Correlation coefficient: $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

- کواریانس مثبت: اگر $Cov_{A,B} > 0$ باشد، آنگاه A وB هر دو میل به مقادیر بزرگتر از مقدار مورد انتظار دارند.
- کواریانس منفی: اگر $Cov_{A,B} < 0$ باشد، آنگاه A بزرگتر از مقدار مورد انتظار است و B تمایل به مقادیر کمتر از مقدار مورد انتظار دارد.
- اگر دو ویژگی مستقل باشند آنگاه $Cov_{A,B} = 0$ اما عکس آن صحیح نیست

برخی از جفت متغیرهای تصادفی ممکن است کواریانس۰ بگیرند اما مستقل نیستند. تنها با افزودن فرض های اضافی (به عنوان مثال، داده‌ها دارای توزیع نرمال چند متغیره باشند) کواریانس۰ حاکی از مستقل بودن است.

کواریانس: یک مثال

محاسبه ساده‌تر آن به صورت روبه‌رو است: $Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$

- فرض کنید که دو سهام a و b مقادیر زیر را در یک هفته دارند
- (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- سوال؟اگر روند سهام یک صنعت تحت تاثیر قرار گیرد آیا قیمت ها باهم افزایش یا کاهش می یابند؟
  - E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0

کواریانس:یک مثال

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

$$Cov(AllElectroncis, HighTech) = \frac{6\times 20+5\times 10+4\times 14+3\times 5+2\times 5}{5} - 4\times 10.80$$

$$= 50.2 - 43.2 = 7.$$

تجمیع داده‌ها – افزونگی(۳)

- در بخش قبلی در مورد افزونگی ویژگی‌ها صحبت شد
- افزونگی ممکن است در سطح تاپل‌ها نیز باشد
- تکرار تاپل ها (**Tuple Duplication**)
  - ممکن است مثلا برای یک شخص دو ردیف وجود داشته باشد و در گذر زمان در یکی از این دو مقدار آدرس عوض شود
- تشخیص و حل ناسازگاری مقدار داده (**Data Value Conflict Detection and Resolution**)
  - تجمیع داده‌ها همچنین شامل تشخیص و حل ناسازگاری میان مقادیر داده‌ای می شود
  - مثال: صفت وزن ممکن است در یک سیستم با واحد متریک یا در سیستم دیگر واحد British imperial داشته باشد
  - مثال: تجمیع داده‌های مدارس یا دانشگاه‌ها با سیستم های نمره دهی متفاوت

فصل سوم: پیش‌پردازش داده‌ها

- پیش‌پردازش داده‌ها: خلاصه‌ای جامع
  - کیفیت داده‌ها (Data Quality)
  - عملیات اصلی در پیش‌پردازش داده‌ها
- پاک‌سازی داده‌ها (Data Cleaning)
- تجمیع داده‌ها (Data Integration)
- کاهش داده‌ها (Data Reduction)
- تغییرشکل (Transformation) و گسسته‌سازی (Discretization) داده‌ها
- خلاصه فصل

استراتژی‌های کاهش داده

- کاهش داده: نمایشی کاهش یافته از مجموعه داده فراهم می‌کند که حجم بسیار کوچکتری دارد اما نتایج مشابهی در تحلیل دارد
- چرا کاهش داده؟
- یک پایگاه داده یا انباره داده ممکن است اطلاعات در حد ترابایت را ذخیره کند
- تجزیه و تحلیل پیچیده ممکن است زمان بسیار طولانی را برای اجرای روی مجموعه اطلاعات کامل نیاز داشته باشد

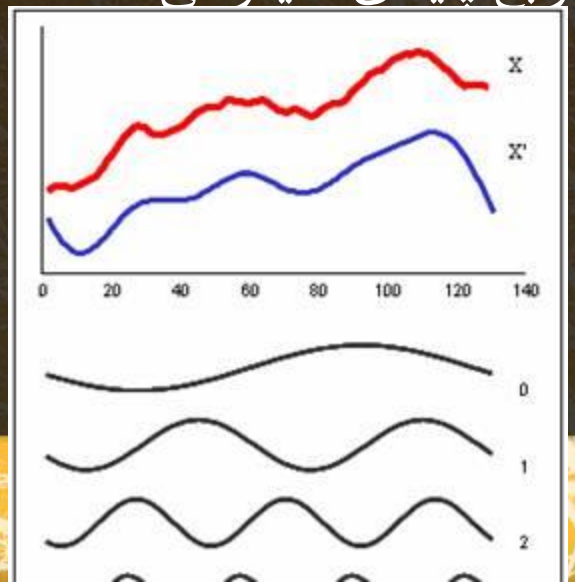
- استراتژی‌های کاهش داده
- کاهش ابعاد، به عنوان مثال، حذف ویژگی‌های بی اهمیت
  - تبدیل موجک (Wavelet transforms)
  - تحلیل مولفه‌های اساسی Principal Components Analysis (PCA)
  - انتخاب زیر مجموعه‌ای از ویژگی‌ها (Feature subset selection)
  - ایجاد ویژگی (feature creation)
- کاهش کثرت داده‌ها (Numerosity)
  - رگرسیون و مدل های خطی لگواریتمی
  - هیستوگرام، خوشه‌بندی، نمونه برداری
  - تجمیع مکعب داده (Data cube aggregation)
- فشرده‌سازی داده‌ها

کاهش داده ۱: کاهش ابعاد

- **ازدیاد ابعاد** (**Curse of dimensionality**)
  - هنگامی که ابعاد افزایش می یابد، اطلاعات به طور فزاینده‌ای تنک (sparse) می شوند
  - تراکم و فاصله بین نقاط ، برای خوشه بندی ، تجزیه و تحلیل کلمات دورافتاده (outlier) بسیار حیاتی است که در این حالت کمتر بامعنی است
  - ترکیبات ممکن از فضاهایی به صورت نمایی رشد کنند
- **کاهش ابعاد**
  - اجتناب از ازدیاد ابعاد
  - کمک به از بین بردن ویژگی‌های بی ربط و کاهش نویز
  - زمان و فضای مورد نیاز در داده کاوی را کاهش می‌دهد
  - مصور سازی ساده‌تر می‌شود
- **تکنیک های کاهش ابعاد**
  - تبدیل موجک (Wavelet transforms)
  - تحلیل مولفه‌های اساسی Principal Components Analysis (PCA)
  - روشهای با نظارت و غیر خطی (به عنوان مثال، feature selection)

تبدیل موجک (Wavelet Transformation)

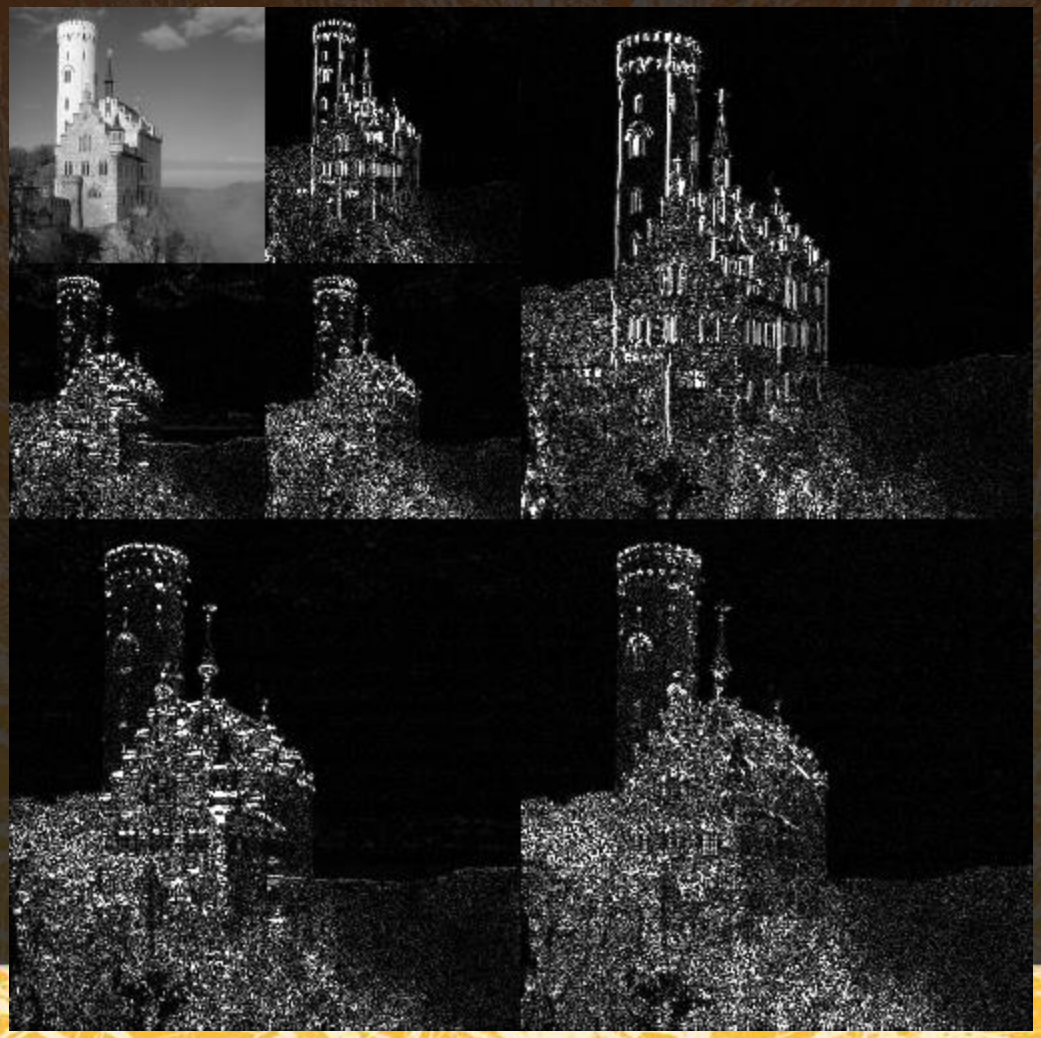
- مشابه تبدیل فوریه است (Fourier Transform)
- در بسیاری از کاربردها مرسوم است که از ترکیب توابع پایه‌ای برای تقریب یک تابع استفاده شود
- به عنوان مثال هر تابع پیوسته را می‌توان توسط مجموعه‌ای از توابع چند جمله-ای نمایش داد
- تبدیل فوریه نوعی تبدیل است که یک تابع را بصورت توابع پایه‌ای سینوسی که هر کدام در مقادیری ضرب شده‌اند نشان می‌دهد



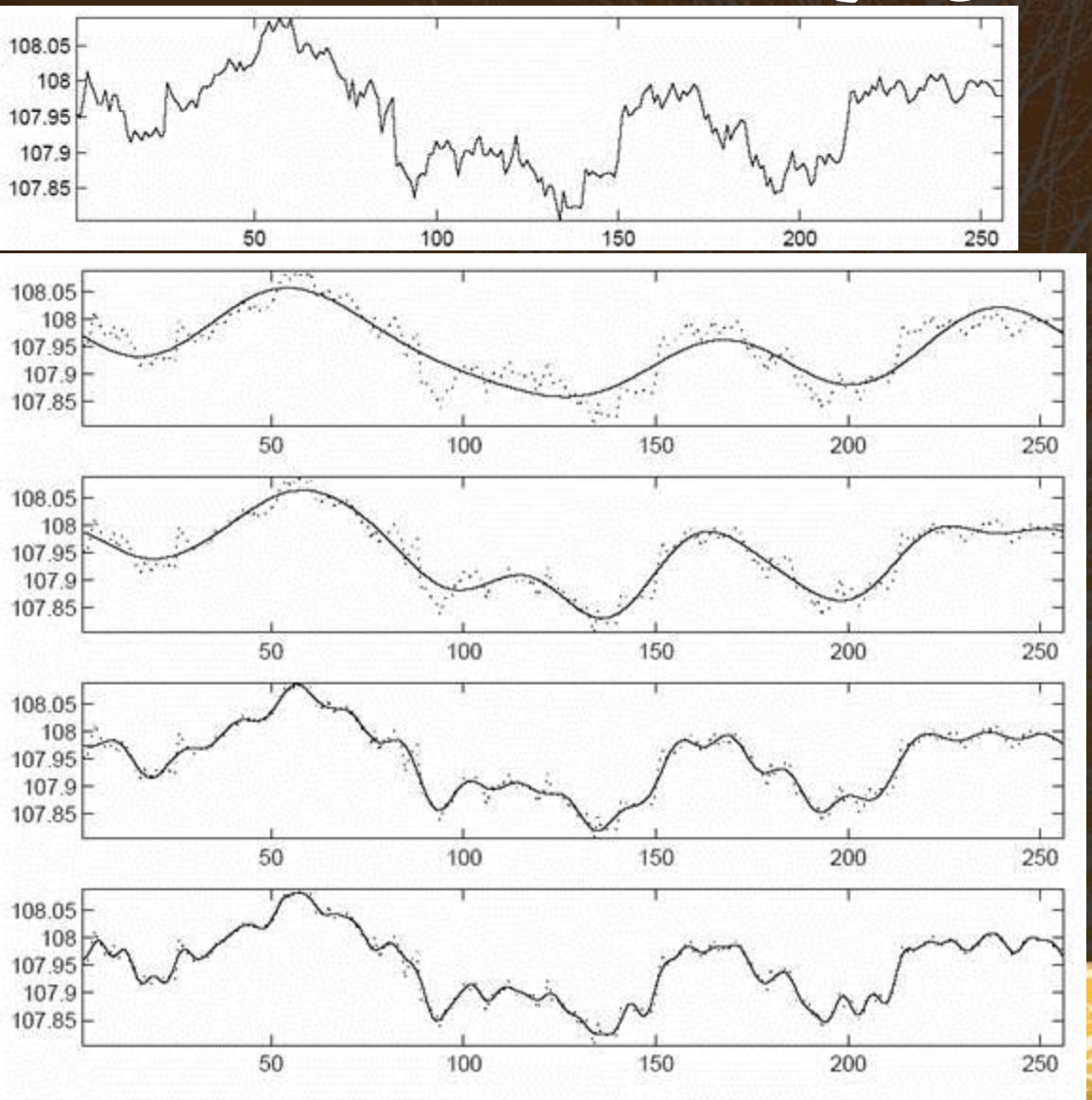
تبدیل فوریه

- تبدیل فوریه یک تبدیل برگشت پذیر است
- این تبدیل می‌تواند به دو صورت پیوسته یا گسسته انجام شود
- در کامپیوتر و بخصوص در پردازش سیگنال معمولاً از تبدیل فوریه‌ی گسسته (DFT) استفاده می‌شود.
- خوشبختانه الگوریتم‌های سریعی تحت عنوان FFT (Fast Fourier Transform) برای تبدیل فوریه‌ی گسسته به وجود آمده است

تبدیل موجک دو-بعدی در تصاویر



تبدیل موجک



- مثال:قیمت روزانه سهام شرکت IBMدر سال ۲۰۰۱
- این بردار از ۲۵۶ مؤلفه تشکیل شده است
- اگر فقط ۱۰ ضریب پر اهمیت‌تر در را نگه داشته و بقیه را حذف نماییم، نموداری شبیه به اولین نمودار در شکل بدست خواهیم آورد

تبدیل موجک

- تبدیل DWT برای اولین بار توسط شخصی به نام Alfred Haar بوجود آمد
- این تبدیل به اینصورت است که یک توالی به طول ۲^n در ورودی داریم
- این اعداد بصورت جفت جفت با هم جمع شده و این حاصل جمع‌ها به مرحله‌ی بعد فرستاده می‌شوند
- . همچنین اختلاف هر جفت نیز محاسبه و ذخیره می‌شود
- دوباره این مرحله تکرار می‌شود با این تفاوت که در ورودی، حاصل جمع جفتهای مرحله‌ی قبل قرار می‌گیرد
- این فرایند بطور بازگشتی تکرار می‌شود تا در نهایت یک عدد که حاصل جمع کل اعداد است بدست آید
- این عدد به همراه ۲^n-1 اختلاف جفت‌ها که در مراحل مختلف الگوریتم محاسبه شده بعنوان خروجی این تبدیل بازگردانده می‌شود

تبدیل Haar : مثال

- بعنوان مثال فرض کنید می‌خواهیم تبدیل Haar Waveletرا برروی رشته S به طول ۸ اعمال نماییم.
- S = (1, 3, 5, 11, 12, 13, 0, 1)

Resolution	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
3	a_1+a_2	a_3+a_4	a_5+a_6	a_7+a_8	a_1-a_2	a_3-a_4	a_5-a_6	a_7-a_8
2	$a_1+a_2+a_3+a_4$		$a_5+a_6+a_7+a_8$		$(a_1+a_2)-(a_3+a_4)$		$(a_5+a_6)-(a_7+a_8)$	
1	$a_1+a_2+a_3+a_4+a_5+a_6+a_7+a_8$				$(a_1+a_2+a_3+a_4)-(a_5+a_6+a_7+a_8)$			

تبدیل Haar : مثال

- S = (1, 3, 5, 11, 12, 13, 0, 1)

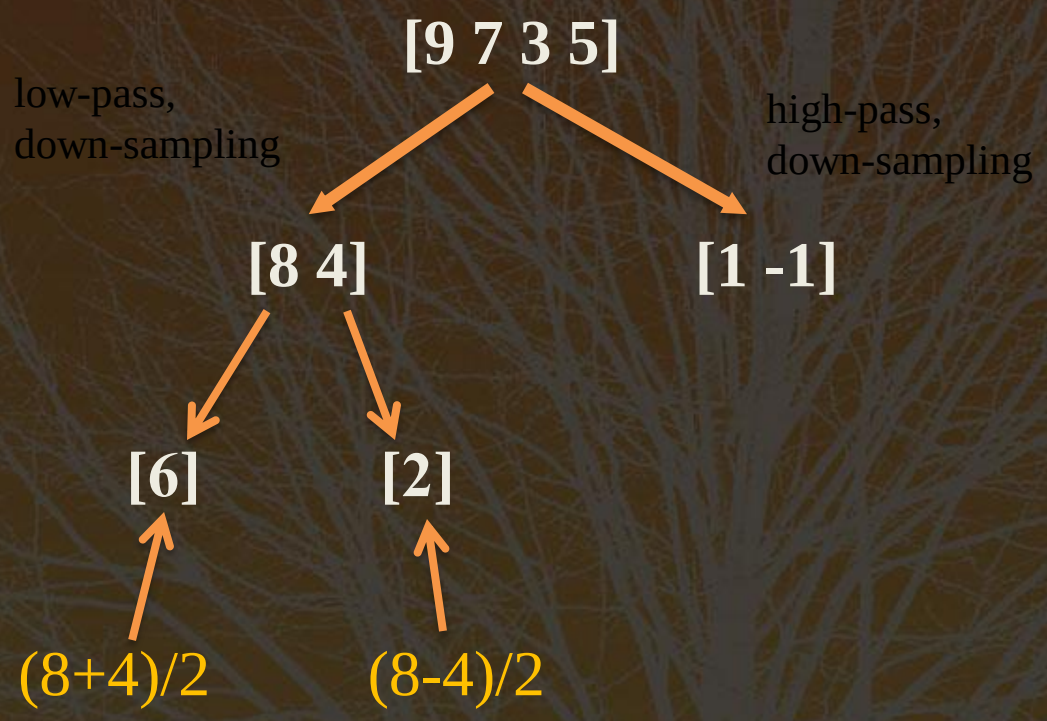
Resolution	Sum				Detail			
4	1	3	5	11	12	13	0	1
3	4	16	25	1	-2	-6	-1	-1
2	20		26		-12		24	
1	46				-6			

DWT(S) = (46, -6, -12, 24, -2, -6, -1, -1)

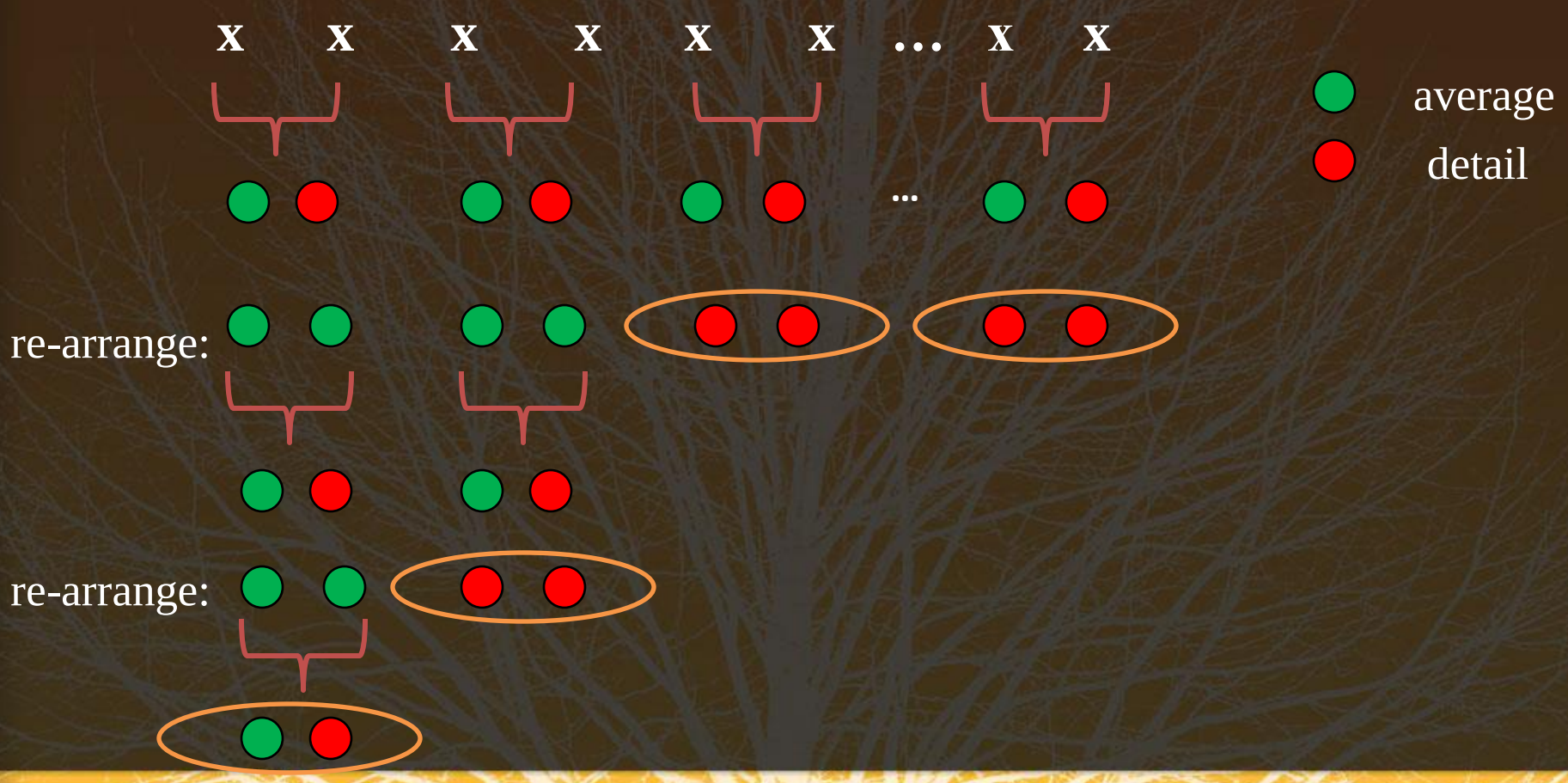
- در اینجا نیز همانند تبدیل فوریه، ضرایب بدست آمده به ترتیب پر اهمیت تا کم اهمیت مرتب شده‌اند
- در واقع ضرایب کم اهمیت همان‌هایی هستند که در مراحل اولیه الگوریتم بدست می‌آیند
- مثلاً در شکل کم اهمیت‌ترین ضرایب مربوط به Resolution=3 هستند، یعنی (۲-، ۶-، ۱- و ۱-). با حذف ضرایب کم اهمیت می‌توان حجم داده‌ها را کاهش داد. البته مقدار کمی از اطلاعات نیز از بین می‌رود

همان تبدیل: به روشی دیگر

همان تبدیل: به روشی دیگر (ادامه)

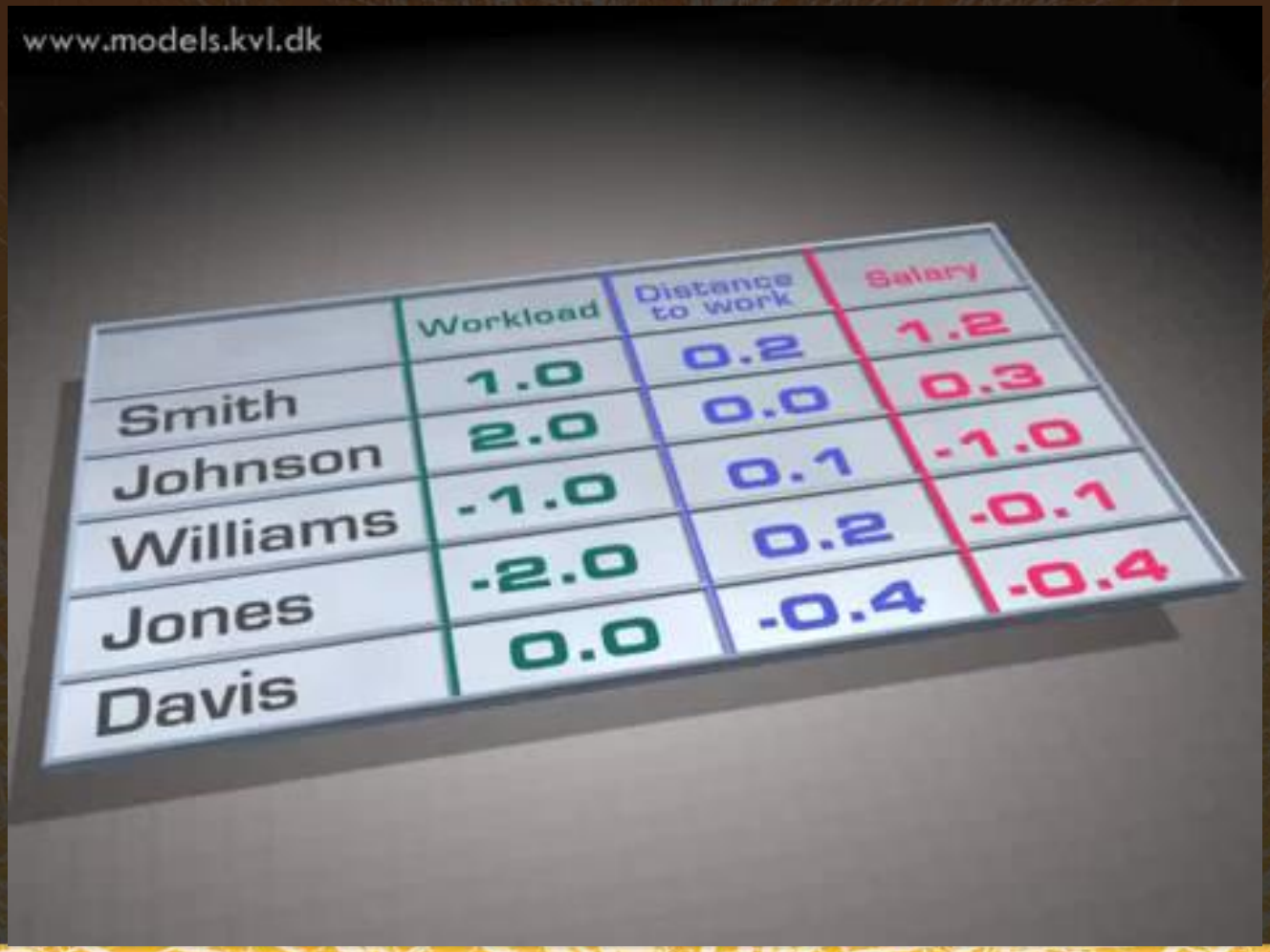


نمایش نحوه تبدیل یک بعدی هاار



حاصل نهایی تبدیل

PCA (Principal Component Analysis)



انتخاب زیرمجموعه ویژگی (Subset Selection)

- راه دیگر برای کاهش ابعاد داده‌ها
- ویژگی‌های اضافی (Redundant)
- ویژگی‌های بی ربط فاقد اطلاعات مفید برای داده کاوی است
- به عنوان مثال id دانش آموزان اغلب برای پیش گویی GPA بی ربط است

جستجو اکتشافی برای انتخاب ویژگی‌ها

اگر n ویژگی وجود داشته باشد می‌توان 2^n زیر مجموعه داشته باشیم

برخی روش‌های هیوریستیک به صورت زیر می‌باشند.

انتخاب روبه‌جلو (Forward Selection): با یک مجموعه تهی شروع کرده و در هر مرحله بهترین ویژگی را انتخاب می‌کنیم

حذف رو به عقب (backward elimination): با مجموعه کامل ویژگی‌ها شروع کرده و در هر مرحله بدترین ویژگی را حذف می‌کنیم

ترکیب دو روش بالا : در هر مرحله بهترین ویژگی را اضافه و بدترین را حذف می‌کنیم

استفاده از درخت تصمیم: با ایجاد درخت تصمیم ویژگی‌هایی که در درخت به کار نرفته‌اند به عنوان نامرتبط در نظر گرفته می‌شوند

جستجو اکتشافی برای انتخاب ویژگی‌ها

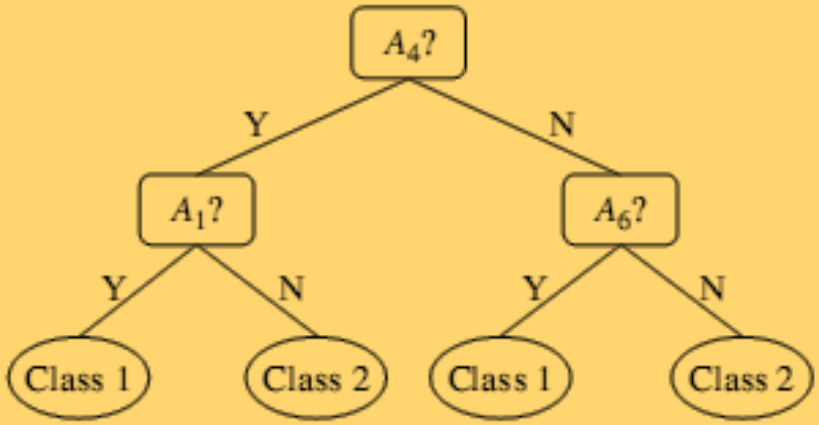
Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Figure 3.6 Greedy (heuristic) methods for attribute subset selection.

ایجاد ویژگی

- ایجاد ویژگی‌های جدید می‌تواند در افزایش دقت و فهم ما از ساختار داده‌های با ابعاد زیاد کمک می‌کند
- به طور مثال: به جای طول و عرض می‌توان مساحت را ایجاد نمود تا در فهم ارتباط بین ویژگی‌ها کمک کننده باشد

روش‌های کاهش داده ۲ : کاهش تعداد

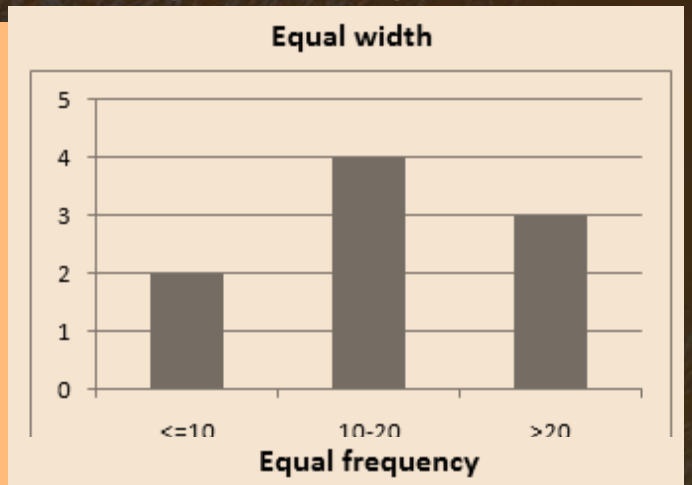
- کاهش حجم داده‌ها از طریق انتخاب شکل کوچکتر و جایگزین برای نمایش داده‌ها
- روش‌های پارامتریک (به عنوان مثال، رگرسیون)
  - در این روش‌ها فرض می‌شود داده‌ها با یک مدل انطباق دارند
  - بنابراین پارامترهای مدل محاسبه گشته
  - پارامترها ذخیره شده و داده‌ها کنار گذاشته می‌شوند
- روش‌های غیر پارامتریک
  - برای داده‌ها مدلی فرض نمی‌شود
  - نمونه روش‌ها:
    - نمودار هیستوگرام
    - خوشه‌بندی
    - نمونه‌برداری، ...

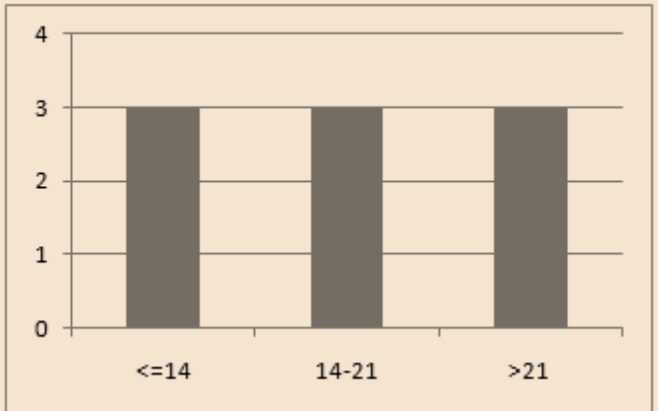
رگرسیون

تحلیل هیستوگرام

- تقسیم داده به تعدادی bucket یا bin و در ادامه ذخیره نمودن میانگین اعداد
- دو روش برای بخش‌بندی داده‌ها وجود دارد.

- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28
- **Equal width**
  - Bin 1: 0, 4 [-,10)
  - Bin 2: 12, 16, 16, 18 [10,20)
  - Bin 3: 24, 26, 28 [20,+)
- **Equal frequency**
  - Bin 1: 0, 4, 12 [-, 14)
  - Bin 2: 16, 16, 18 [14, 21)
  - Bin 3: 24, 26, 28 [21,+)





خوشه‌بندی

- مجزاسازی مجموعه داده‌ها به خوشه‌ها براساس شباهت آنها
- و تنها ذخیره و نگهداری نماینده خوشه (به عنوان مثال مرکز خوشه)
- موجودیت‌های هر خوشه به یکدیگر شبیه‌اند بنابراین می‌توان از یک نماینده برای آن‌ها استفاده نمود
- کیفیت یک خوشه‌بندی می‌تواند توسط عوامل مختلفی سنجیده شود
  - حداکثر فاصله بین هر دو شی در خوشه
  - قطر خوشه
  - میانگین فاصله اشیا از مرکز خوشه‌ی خود (Centroid Distance)

نمونه‌برداری (Sampling)

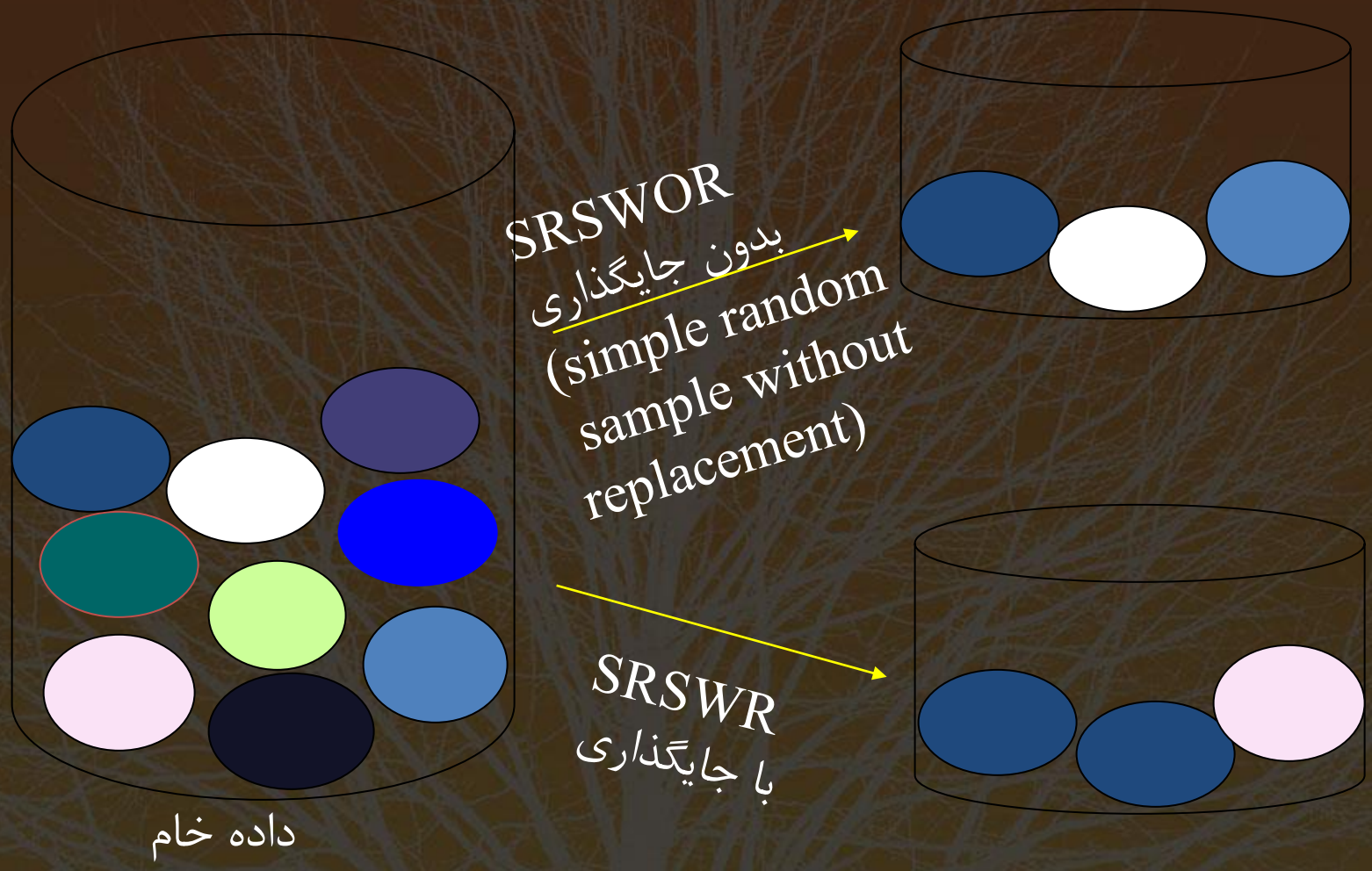
- **نمونه‌برداری: در نظر گرفتن مجموعه s به طوری که نماینده کل مجموعه داده N باشد**
- با این روش پیچیدگی زمان اجرای الگوریتم‌های داده‌کاوی می‌تواند به صورت خطی نسبت به اندازه مجموعه داده افزایش یابد
- **اصل کلیدی : انتخاب زیر مجموعه‌ای از داده‌ها که نماینده کل مجموعه داده باشند**
  - روش نمونه‌گیری تصادفی ساده ممکن است عملکرد بسیار ضعیفی در مجموعه داده‌های نامتوازن (skewed) داشته باشد
  - توسعه روش‌های نمونه‌گیری تطبیقی، به عنوان مثال، stratified sampling

انواع نمونه‌گیری

- **روش نمونه‌گیری تصادفی ساده**
  - یک احتمال برابر انتخاب هر شی داده وجود دارد
- **نمونه‌گیری بدون جایگذاری (Sampling without replacement)**
  - هنگامی که یک شی انتخاب شد، آن را از جمعیت حذف می کنیم
- **نمونه‌گیری با جایگذاری (Sampling with replacement)**
  - شی انتخاب شده از جمعیت اولیه حذف نمی شود
- **نمونه‌گیری خوشه‌ای**

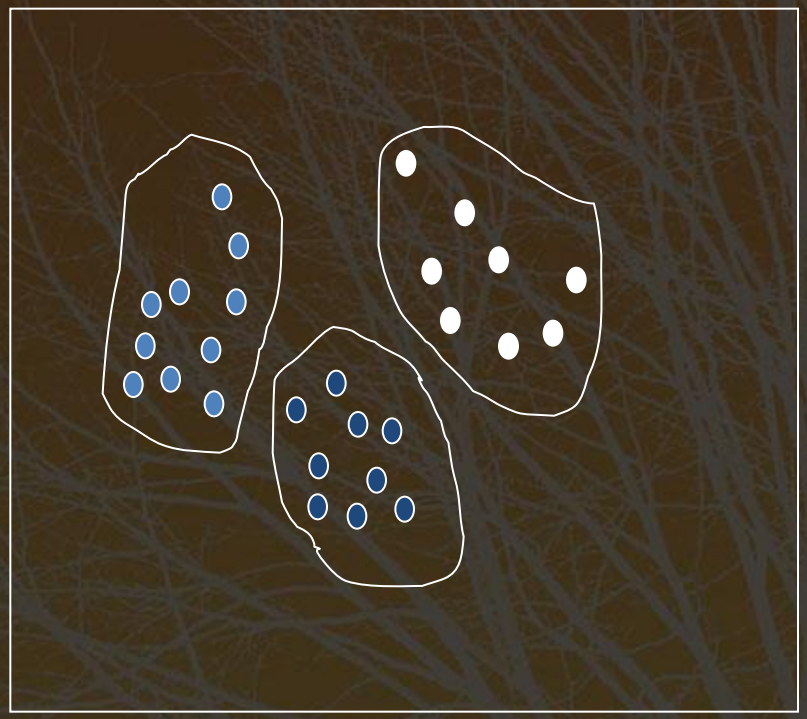
- **نمونه‌گیری Stratified sampling**
  - پارتیشن‌بندی مجموعه داده‌ها و برداشتن نمونه از هر پارتیشن (متناسب با همان درصد از داده‌ها که در این پارتیشن قرار دارند)
  - مناسب برای مجموعه‌داده‌های نامتوازن

نمونه‌گیری: با یا بدون جایگزین

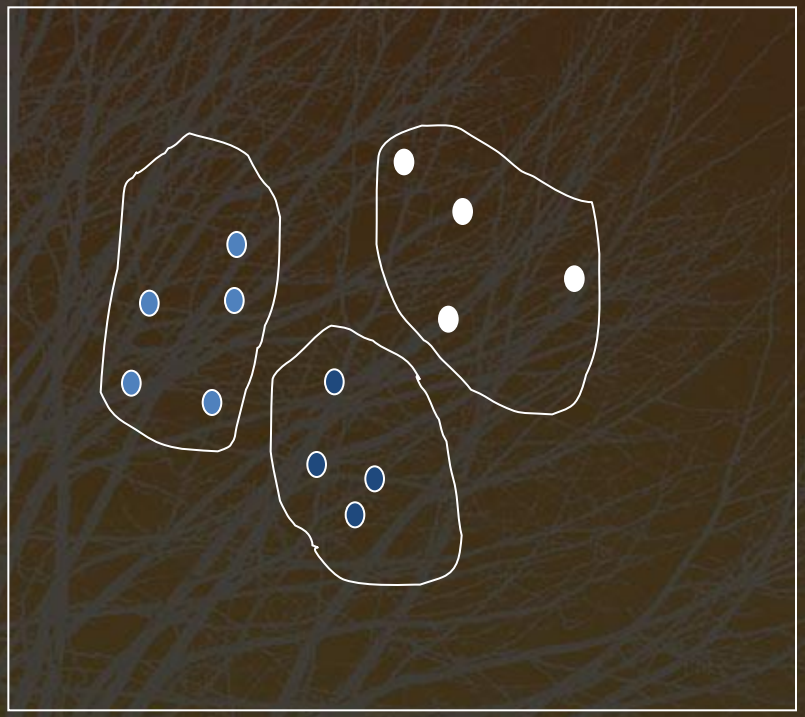


نمونه‌گیری: خوشه یا نمونه‌گیری طبقه ای

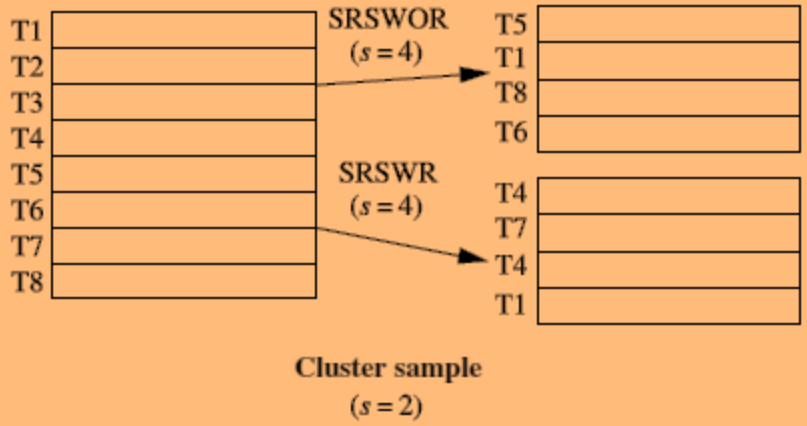
داده خام



نمونه خوشه/طبقه



مثال‌های نمونه‌برداری



Startified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

Data Cube Aggregation تجمیع مکعب داده

به طور مثال حاوی اطلاعات فروش هر سال به تفکیک هر فصل است

می‌توان با تجمیع اطلاعات خلاصه شده را به ازای هر سال بدست آورد

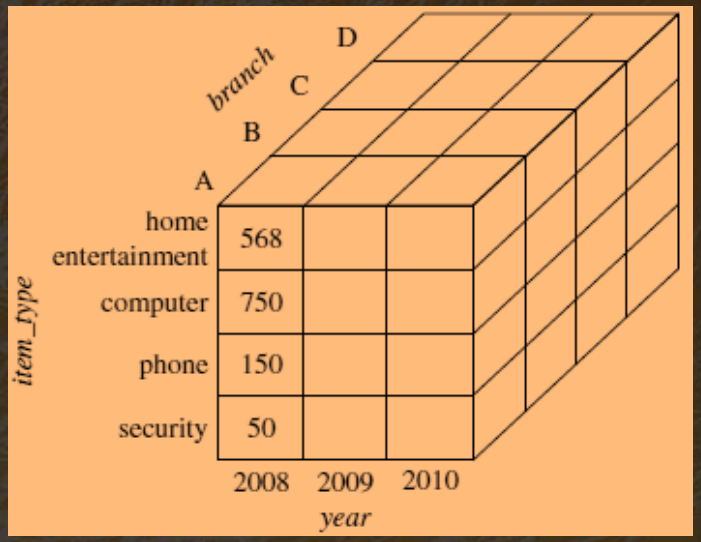
Year 2010

Year 2009

Year 2008

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2008	$1,568,000
2009	$2,356,000
2010	$3,594,000



فصل سوم: پیش‌پردازش داده‌ها

- پیش‌پردازش داده‌ها: خلاصه‌ای جامع
  - کیفیت داده‌ها (Data Quality)
- عملیات اصلی در پیش‌پردازش داده‌ها
  - پاک‌سازی داده‌ها (Data Cleaning)
  - تجمیع داده‌ها (Data Integration)
  - کاهش داده‌ها (Data Reduction)
  - تبدیل داده‌ها (Transformation) و گسسته‌سازی (Discretization) داده‌ها
  - خلاصه فصل

تبدیل داده

- یک تابع، کل مجموعه مقادیر از یک ویژگی داده شده را به یک مجموعه‌ی جدید از مقادیر نگاشت می‌دهد
  - روش‌ها
    - هموارسازی (Smoothing): حذف نویز از داده‌ها
    - ساخت ویژگی
      - ویژگی‌های جدیدی ازویژگی‌های داده شده ساخته می‌شود (ایجاد می‌شود)
    - تجمیع (Aggregation): خلاصه‌سازی، ساخت مکعب داده
    - نرمال‌سازی(Normalization) : مقیاسی است برای انتقال داده‌ها به محدوده کوچکتر
      - **نرمال‌سازی min-max**
      - **نرمال‌سازی Z-score**
      - **نرمال‌سازی با مقیاس اعشاری**
    - گسسته‌سازی(Discretization)

نرمال‌سازی: روش اول

- نرمال‌سازی : انتقال به بازه **[new_min$_A$, new_max$_A$]**

$$v' = \frac{v - min_A}{max_A - min_A}(new_max_A - new_min_A) + new_min_A$$

Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to

$$\frac{73600-12000}{98000-12000}(1.0 - 0) + 0 = 0.716$$

نرمال‌سازی: روش دوم

- نرمال‌سازی Z-score: (**μ: mean, σ : standard deviation**)

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73600 - 54000}{16000} = 1.225$

نرمال‌سازی: روش سوم

- نرمال‌سازی با مقیاس اعشاری

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v'|) < 1

A range from -986 to 917. The maximum absolute value of A is 986.

j=3(1000) so that -0.986 and .917

گسسته‌سازی

- **گسسته‌سازی:** تقسیم محدوده ی وسیعی از ویژگی‌های پیوسته به فواصل زمانی
  - قرار دادن برچسب‌های فاصله به جای مقادیر داده‌های واقعی و حقیقی
  - کاهش اندازه داده شده توسط گسسته‌سازی
  - نظارت در مقابل بدون نظارت
  - تقسیم (از بالا به پایین) در مقابل ادغام (پایین به بالا)
  - گسسته‌سازی در یک ویژگی را می توان به صورت بازگشتی اجرا کرد
  - آماده‌سازی برای تجزیه و تحلیل بیشتر. به عنوان مثال، طبقه‌بندی

روش‌های گسسته‌سازی داده‌ها

- روش معمول: تمام روش‌ها را می توان به صورت بازگشتی به کار گرفت.

Binning

- تقسیم بالا به پایین، بدون نظارت

تجزیه و تحلیل هیستوگرام

- تقسیم بالا به پایین، بدون نظارت

تجزیه و تحلیل خوشه‌بندی (بدون نظارت، تقسیم از بالا به پایین یا ادغام از پایین به بالا)

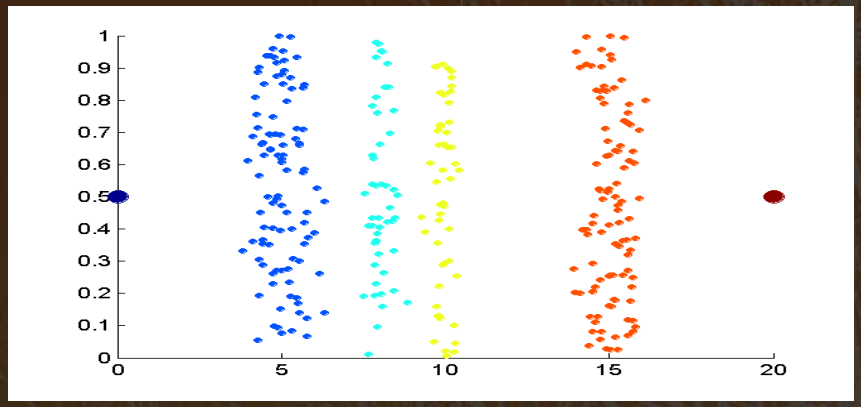
تجزیه و تحلیل درخت تصمیم گیری (تحت نظارت، تقسیم از بالا به پایین)

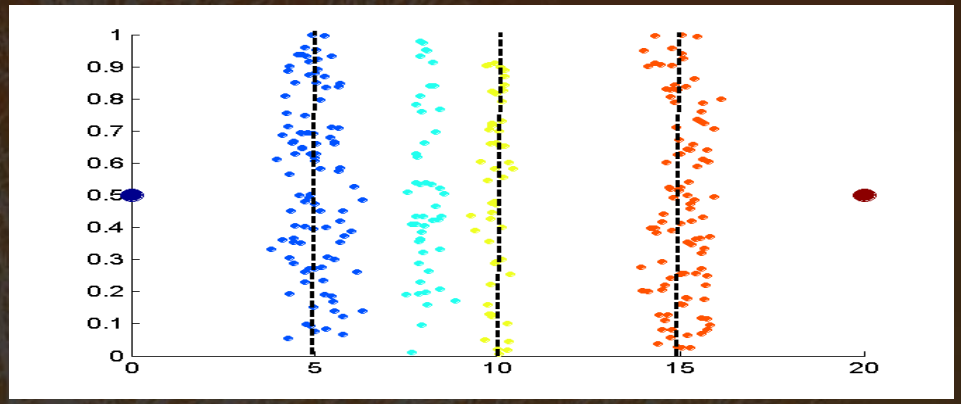
همبستگی (به عنوان مثال، $\chi 2$) تجزیه و تحلیل (بدون نظارت، ادغام از پایین به بالا)

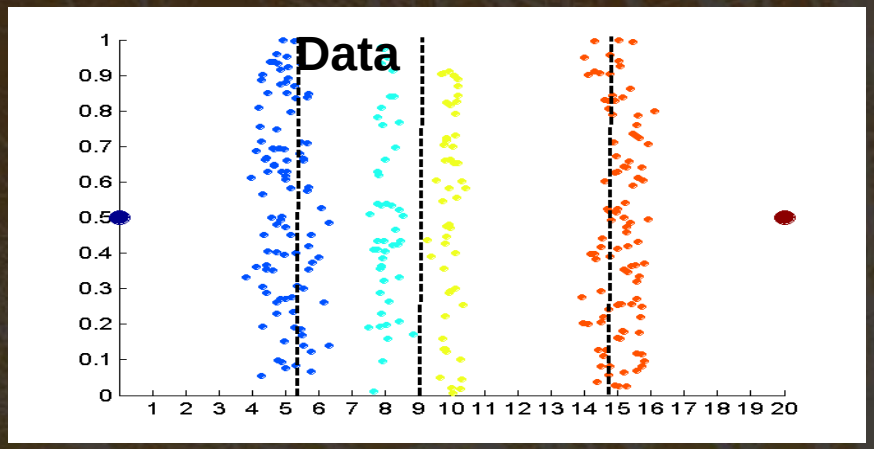
گسسته‌سازی ساده: Binning

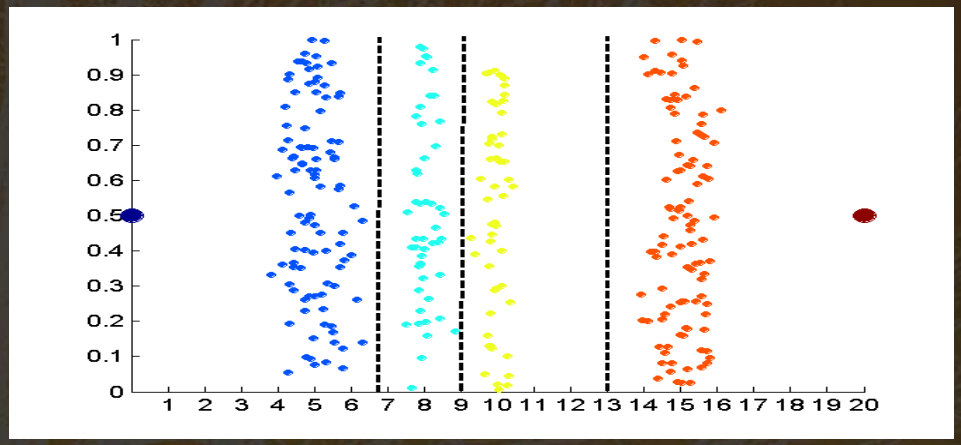
- **پارتیشن‌بندی عرض مساوی** (فاصله)
  - تقسیم دامنه **به فواصل زمانی N با اندازه مساوی**: شبکه یکنواخت
  - **اگر A و B کمترین و بیشترین ارزش از ویژگی باشند، عرض فواصل زمانی برابر است با: W = (B-A) / N**
  - ساده‌ترین است، اما ممکن نقاط دورافتاده تسلط و چیرگی را عرضه کنند.
  - داده‌های تحریف شده، به خوبی به کار گرفته نشده است.
- **پارتیشن‌بندی عمق مساوی** (فرکانس)
  - تقسیم دامنه به فواصل زمانی N، تقریباً هر کدام شامل تعداد نمونه یکسانی می باشند.
  - سنجش(مقیاس) داده‌های خوب
  - ویژگی‌های صریح (بی قید و شرط) می توانند مشکلات و پیچیدگی‌ها را مدیریت کنند.

گسسته‌سازی بدون استفاده از برچسب کلاس (Binning در مقابل خوشه‌بندی)









Equal frequency (binning)

K-means clustering leads to better results

گسسته‌سازی توسط طبقه‌بندی و تجزیه و تحلیل همبستگی

- **طبقه‌بندی (به عنوان مثال، تجزیه و تحلیل درخت تصمیم گیری)**
  - نظارت: دادن برچسب کلاس. به عنوان مثال: مبتلایان به سرطان خوش خیم
  - استفاده از آنتروپی (افت- بی نظمی) برای تعیین نقطه تقسیم (نقطه گسسته)
  - تقسیم بازگشتی از بالا به پایین
  - اطلاعات در فصل ۷ پوشش داده می شود.
- **تجزیه و تحلیل همبستگی (به عنوان مثال، مجذور ادغام: χ^2 مبتنی بر گسسته‌سازی است)**
  - با نظارت: استفاده از اطلاعات کلاس
  - ادغام پایین به بالا : پیدا کردن بهترین فاصله زمانی از همسایه (آنهایی که توزیع های مشابه ای از کلاس ها دارند، به عنوان مثال، ارزش χ^2 کم است) برای ادغام
  - تا زمانی که شرایط توقف از پیش تعریف شده باشد، ادغام به صورت بازگشتی انجام می شود.

مفهوم تولید (نسل) سلسله مراتبی

- مفهوم سازماندهی مفاهیم سلسله مراتبی (به عنوان مثال، ویژگی مقادیر) سلسله مراتبی معمولاً با هر کدام از ابعاد موجود در انبار داده مرتبط می باشد.
- مفهوم سلسله مراتبی به منظور تسهیل حفاری و هموار کردن داده‌ها در انبار داده می باشد، بنابراین می توان داده‌های مجزای متعددی را مشاهده کرد.
- شکل گیری مفهوم سلسله مراتبی: کاهش داده‌های بازگشتی توسط جمع‌آوری و جایگزینی مفاهیم سطح پایین (مانند مقادیر عددی برای سن) و توسط مفاهیم سطح بالاتر (مانند جوانان، بزرگسالان، و یا سالمندان)
- مفهوم سلسله مراتبی را می توان به صراحت توسط کارشناسان حوزه و / یا طراحان انبار داده‌ها تعیین کرد
- مفهوم سلسله مراتبی به صورت خودکار می تواند برای هر دو داده‌های عددی و اسمی تشکیل شود. برای داده‌های عددی، استفاده از روش‌های گسسته‌سازی نشان داده شده است.

مفهوم تولید سلسله مراتبی برای داده‌های اسمی

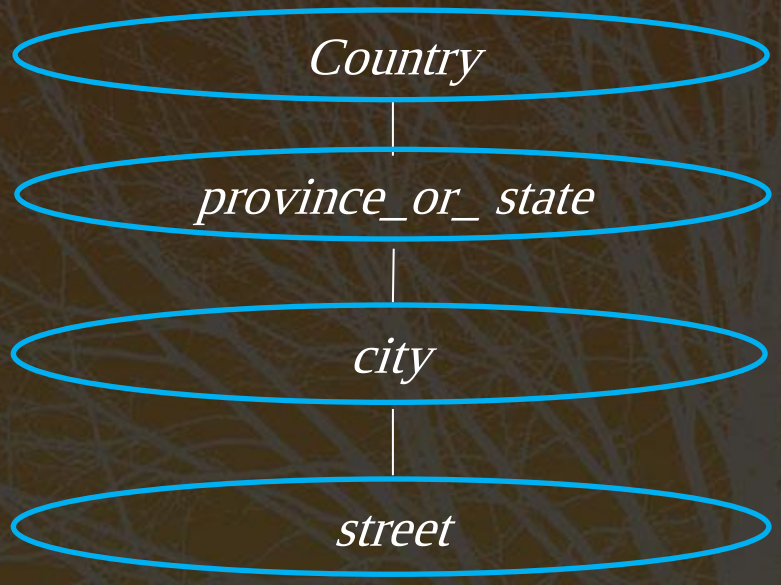
- مشخصات جزئی / سفارش کلی از ویژگی‌ها، صراحتاً توسط کاربران و یا کارشناسان در سطح طرح می باشد.
  - خیابان < شهرستان < استان < کشور
- مشخصات یک سلسله مراتبی توسط دسته‌بندی داده‌های صریح و روشن برای مجموعه ای از مقادیر
  - {اوربانا، شامپاین، شیکاگو} < ایلینوی
- فقط مشخصات از بخشی از ویژگی‌های یک مجموعه
  - به عنوان مثال: فقط خیابان < شهرستان، نه نوع دیگر
- تولید (نسل) خودکار سلسله مراتبی (یا سطح ویژگی‌ها) توسط تجزیه و تحلیل تعدادی از مقادیر مجزا (متمایز)
  - به عنوان مثال: برای یک مجموعه ای از ویژگی: {خیابان، شهرستان، استان، کشور}

مفهوم تولید خودکار سلسله مراتبی

بعضی از سلسله مراتب‌ها را می توان به صورت خودکار، که مبتنی بر تجزیه و تحلیل تعدادی از مقادیر مجزا در هر ویژگی از مجموعه داده‌ها می باشد، را تولید کرد.

ویژگی‌هایی که متمایزترین مقادیر را دارا می باشند، در پایین‌ترین سطح از سلسله مراتب قرار می گیرد

استثنا، به عنوان مثال: هفته، ماه، یک چهارم (فصل)، سال



15 distinct values

365 distinct values

3567 distinct values

674,339 distinct values