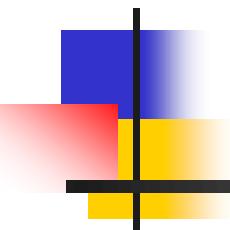


داده کاوی

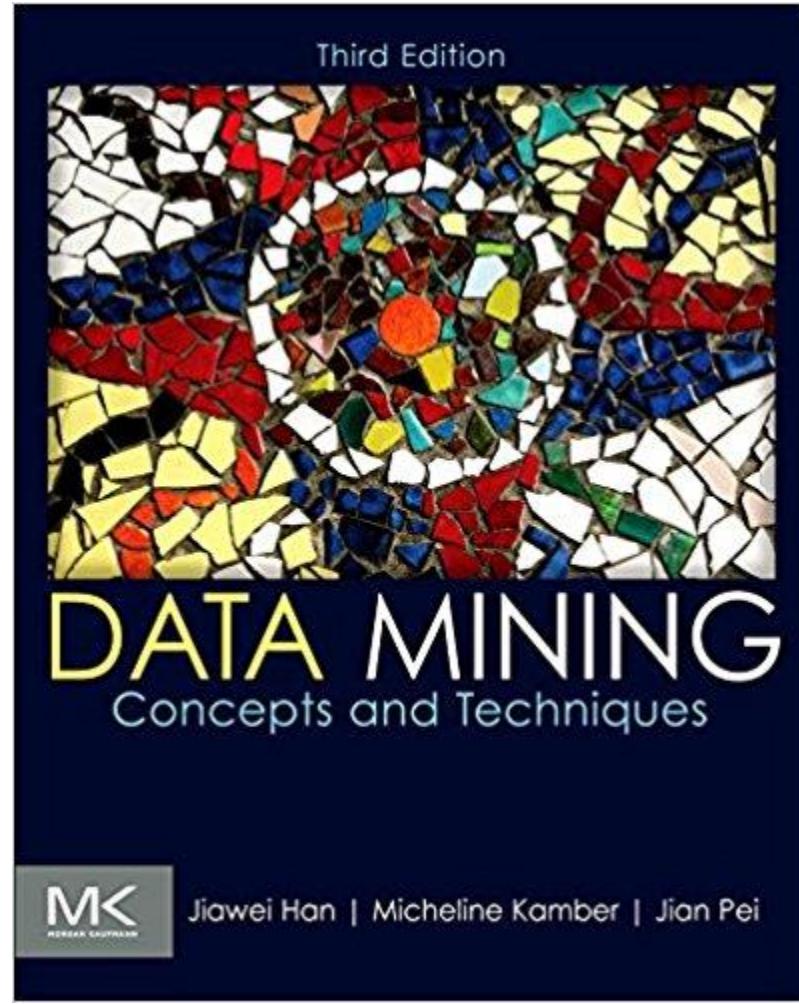


اهداف درس



- آشنایی با مفاهیم و تکنیک های داده کاوی
- آماده کردن دانشجویان برای پژوهش در این زمینه

كتاب مرجع



نحوه تعیین نمره



- امتحان میان ترم ۲۵%
- امتحان پایان ترم ۵۰%
- تمرین ۱۰%
- امتحان عملی ۱۵%
- ارائه ۵% نمره اضافه

فصل ۱ : مقدمه



- چرا داده کاوی؟
- داده کاوی چیست؟
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوها کاوش می شوند؟
- چه تکنولوژی هایی استفاده می شوند؟
- کدام برنامه های کاربردی هدف داده کاوی هستند؟
- موضوعات عمدی در داده کاوی
- تاریخچه ای از داده کاوی و جامعه داده کاوی
- خلاصه

چرا داده کاوی؟

- رشد نمایی حجم اطلاعات
- جمع آوری داده و در دسترس بودن آن
- ابزار اتوماتیک جمع آوری داده ها، سیستم های پایگاه داده ها، وب، جامعه کامپیوتري شده
- منابع اصلی داده ها
- کسب و کار: وب، تجارت الکترونیکی، تراکنش ها، انبار کالا...
- علوم: حسگر های راه دور، بیوانفورماتیک، شبیه سازی...
- جامعه و افراد: اخبار، دوربین های دیجیتال، شبکه های اجتماعی
- دنیای لبریز از اطلاعات اما فاقد دانش!
- داده کاوی: تحلیل اتوماتیک حجم عظیم داده

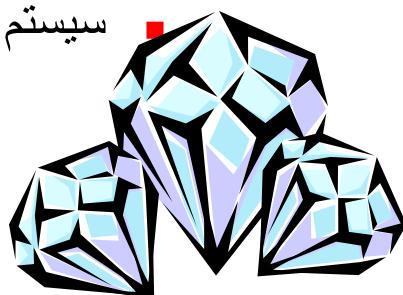
فصل ۱ : مقدمه

- چرا داده کاوی؟
- داده کاوی چیست؟ 
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوها کاوش می شوند؟
- چه تکنولوژی هایی استفاده می شوند؟
- کدام فن آوری ها استفاده می شوند؟
- موضوعات عمدی در داده کاوی
- تاریخچه ای از داده کاوی و جامعه داده کاوی
- خلاصه

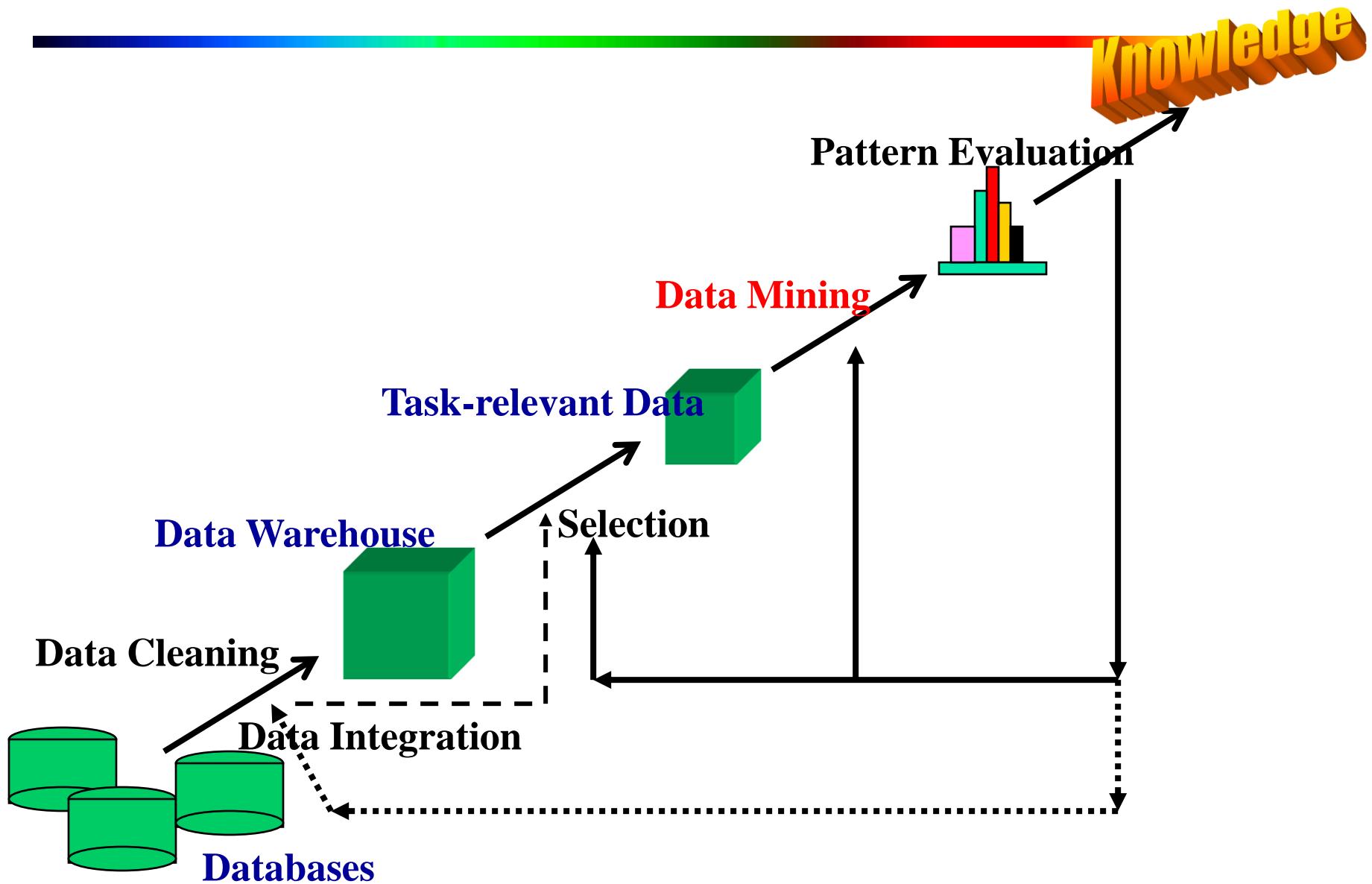


داده کاوی چیست؟

- داده کاوی (کشف دانش از داده ها)
- استخراج الگوهای و دانش جالب (مهم، صریح، جدید، و بالقوه مفید) از حجم عظیم داده ها
- داده کاوی: نام بی مسمی
- اسامی دیگر: دانش کاوی از داده ها، تحلیل داده/الگو، باستان شناسی داده ها، لایروبی داده ها ...
- آیا هر کاوشی در داده داده کاوی است؟
 - جستجو های ساده و پردازش پرسش ها
 - سیستم های خبره (استقراء)



فرایند کشف دانش

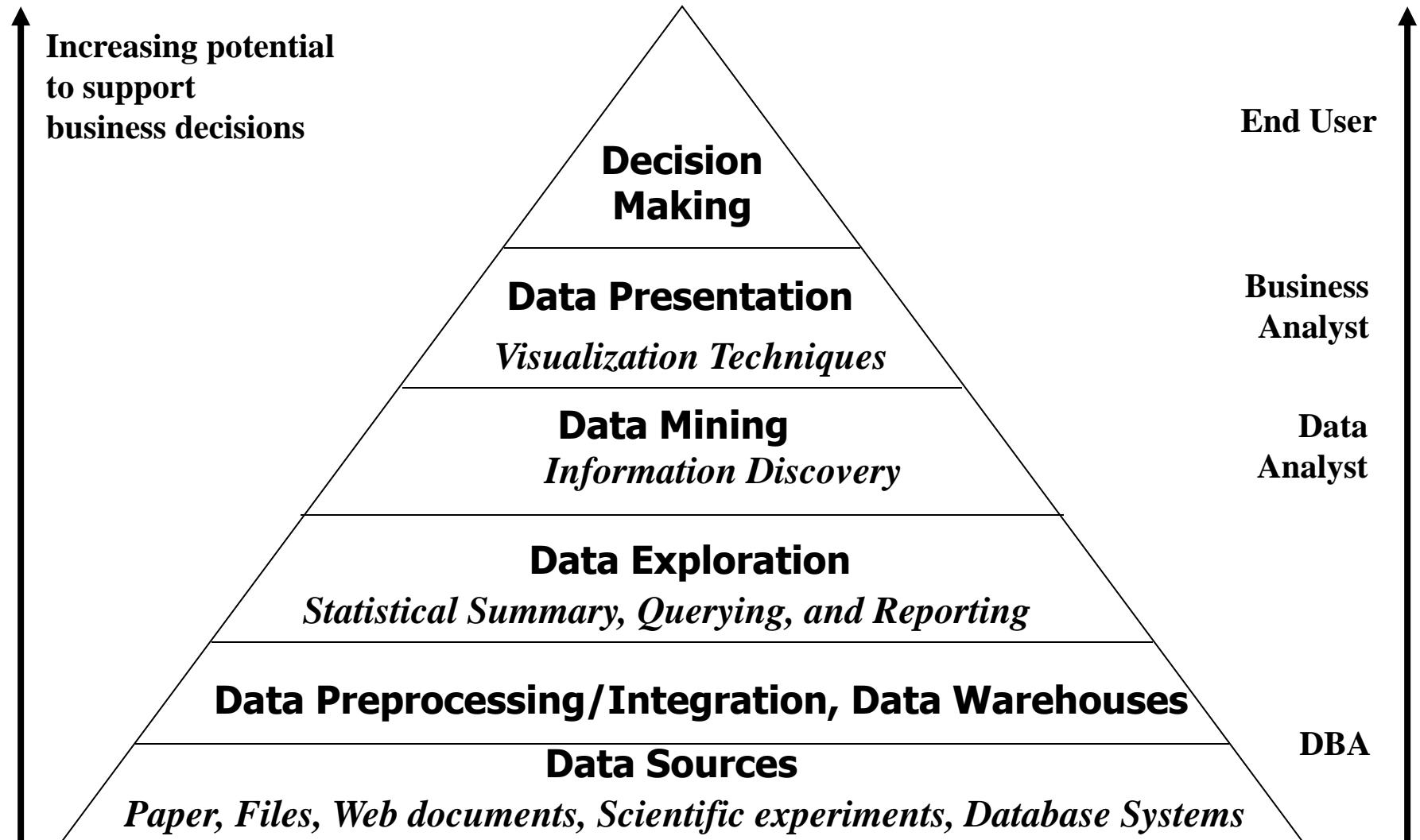


مثال: چارچوبی برای وب کاوی

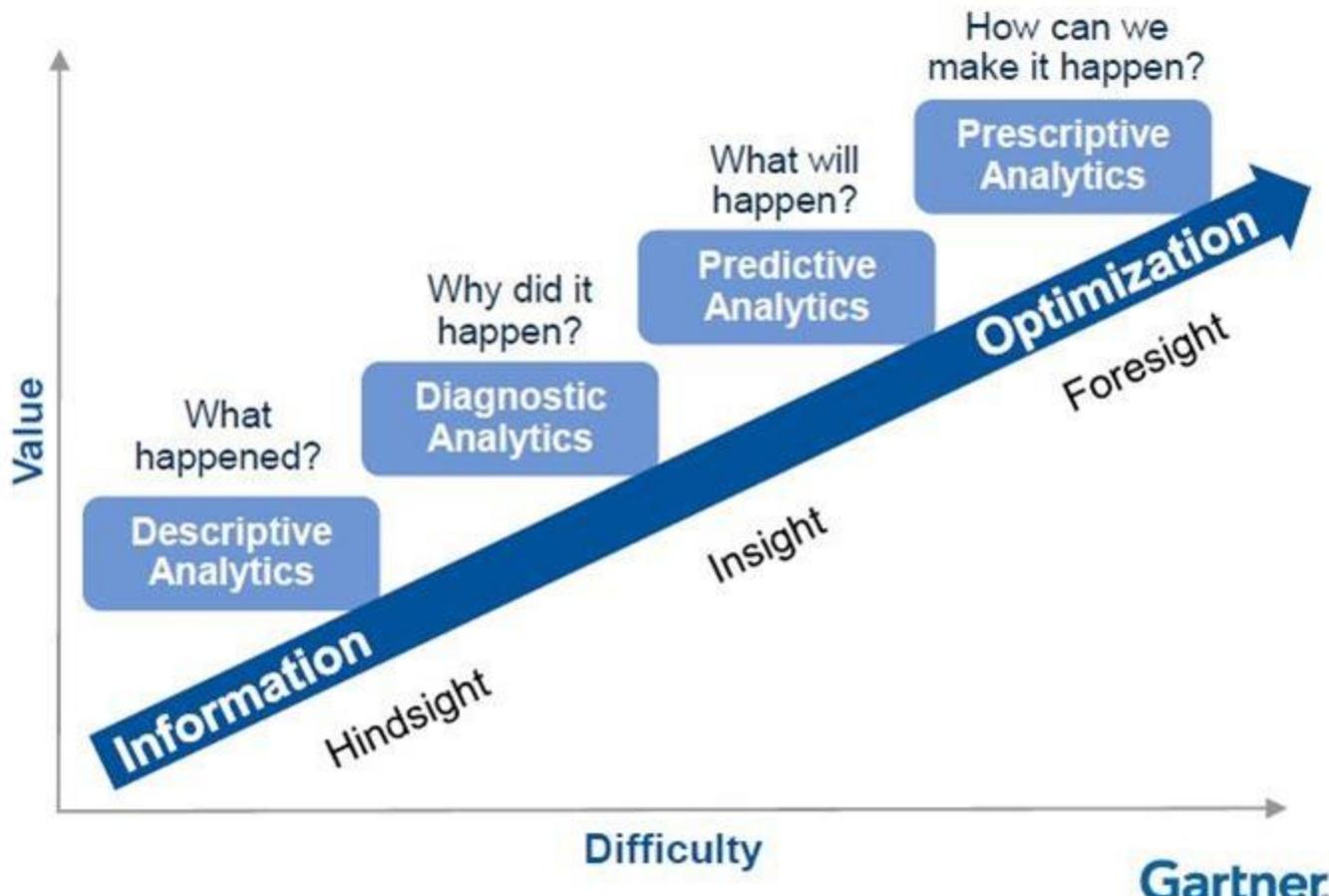
■ وب کاوی معمولاً شامل مراحل زیر است:

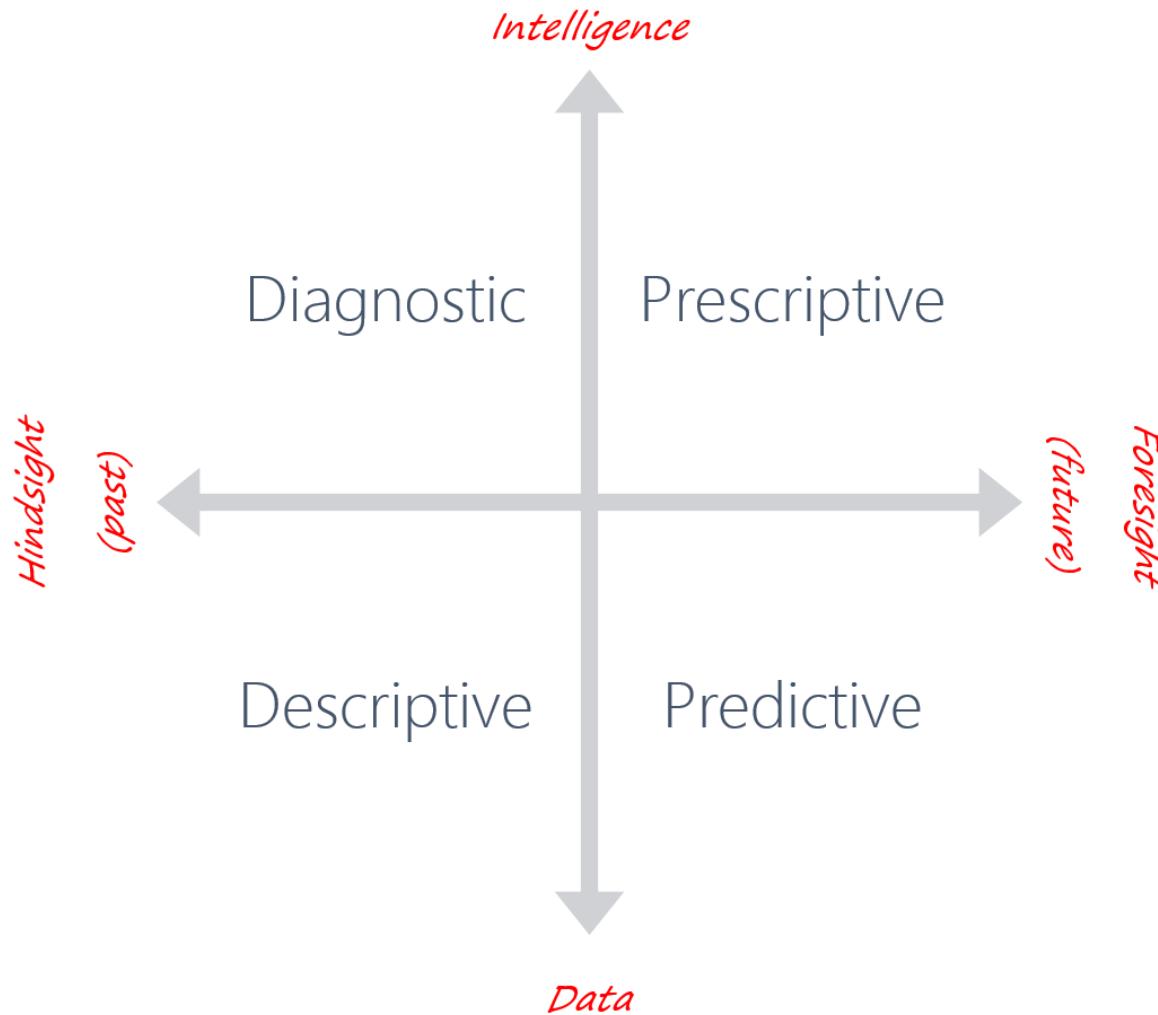
- پالایش داده (برای حذف داده های نویز و ناسازگار)
- یکپارچه سازی داده (از منابع مختلف)
- ساخت انبار داده
- انتخاب داده ها
- ساخت مکعب داده
- داده کاوی
- ارزیابی الگوهای کشف شده
- ارائه دانش

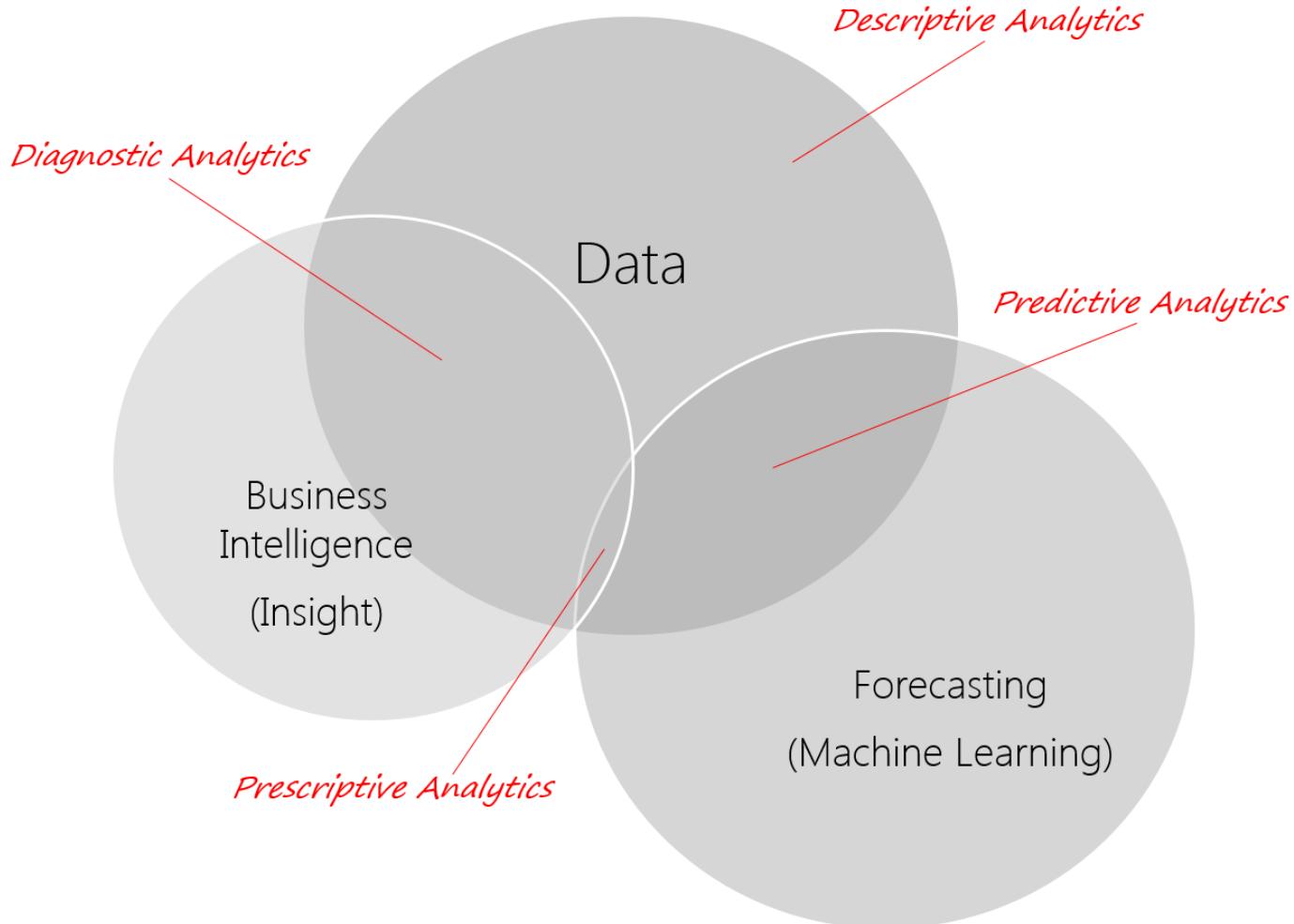
نقش داده کاوی در هوشمندی کسب و کار



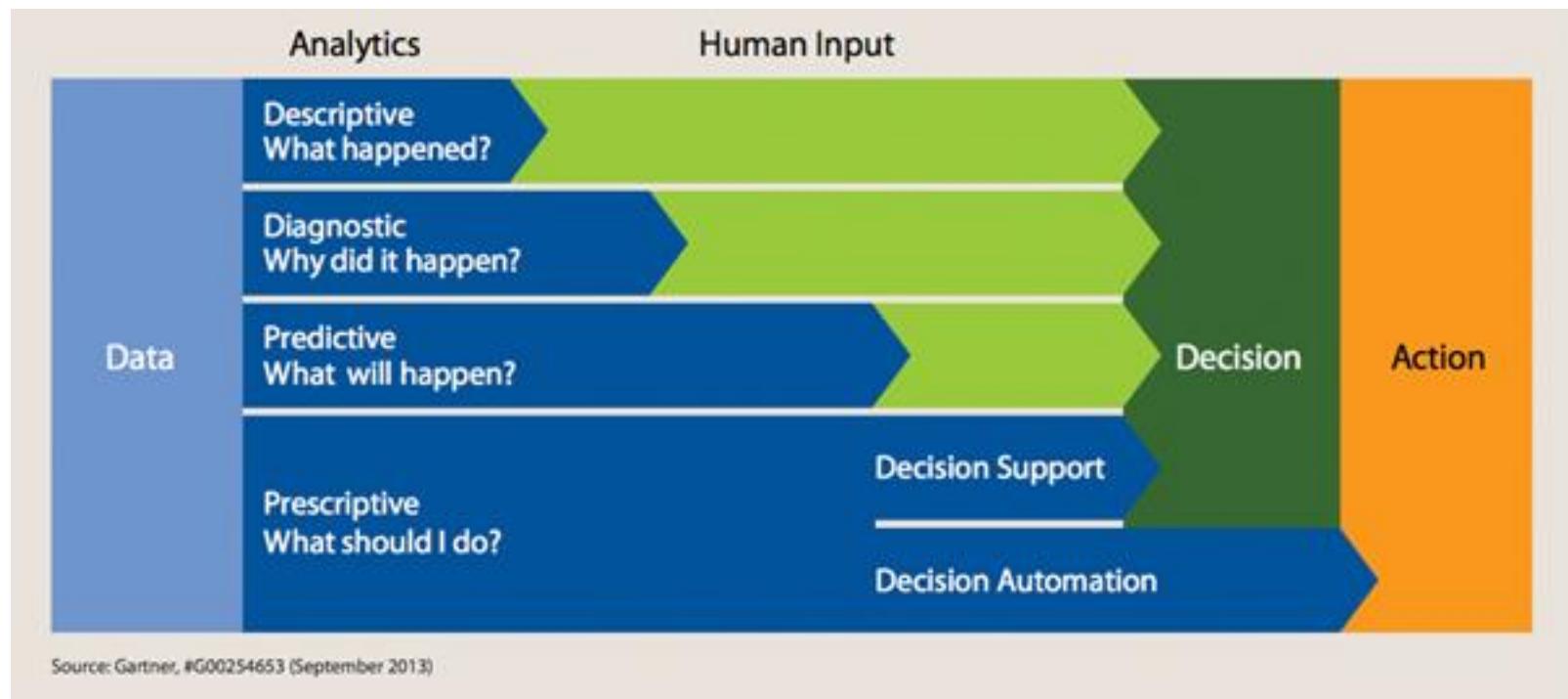
مدل بلوغ تحلیل گارتner







حوزه اثر انسان و ماشین در مراحل مختلف بلوغ



فصل ۱ : مقدمه



- چرا داده کاوی؟
- داده کاوی چیست؟
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوها کاوش می شوند؟
- چه تکنولوژی هایی استفاده می شوند؟
- کدام برنامه های کاربردی هدف داده کاوی هستند؟
- موضوعات عمدی در داده کاوی
- تاریخچه ای از داده کاوی و جامعه داده کاوی
- خلاصه

نمای چند وجهی از داده کاوی

چه داده هایی کاوش می شوند

Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks

چه دانشی کاوش می شود (یا عملیات مختلف داده کاوی)

Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.

Descriptive vs. predictive data mining

Multiple/integrated functions and mining at multiple levels

چه تکنیک هایی استفاده می شود

Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

چه برنامه های کاربردی هدف داده کاوی هستند؟

Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

فصل ۱ : مقدمه



- چرا داده کاوی؟
- داده کاوی چیست؟
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوها کاوش می شوند؟
- چه تکنولوژی هایی استفاده می شوند؟
- کدام برنامه های کاربردی هدف داده کاوی هستند؟
- موضوعات عمدی در داده کاوی
- تاریخچه ای از داده کاوی و جامعه داده کاوی
- خلاصه

داده کاوی: روی چه انواع داده ای

- داده های پایگاه های داده و برنامه های کاربردی
 - پایگاه داده های رابطه ای، انبار های داده، داده های تراکنشی
 - دیتا ست ها و برنامه های کاربردی پیشرفته
 - داده های جریانی (سنسورها...)
- داده های زمانی یا توالی ها (رکوردهای تاریخمند، بورس، توالی های زیستی...)
- داده های ساختمند، گراف ها، شبکه های اجتماعی...
- سیستم های اطلاعاتی شی گرا
- پایگاه داده های متفرقه و قدیمی
- داده های فضایی یا مکان محور (نقشه ها...)
- داده های چند رسانه ای
- داده های متند
- وب

فصل ۱ : مقدمه



- چرا داده کاوی؟
- داده کاوی چیست؟
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوهای کاوش می شوند؟
- چه تکنولوژی هایی استفاده می شوند؟
- کدام برنامه های کاربردی هدف داده کاوی هستند؟
- موضوعات عمدی در داده کاوی
- تاریخچه ای از داده کاوی و جامعه داده کاوی
- خلاصه

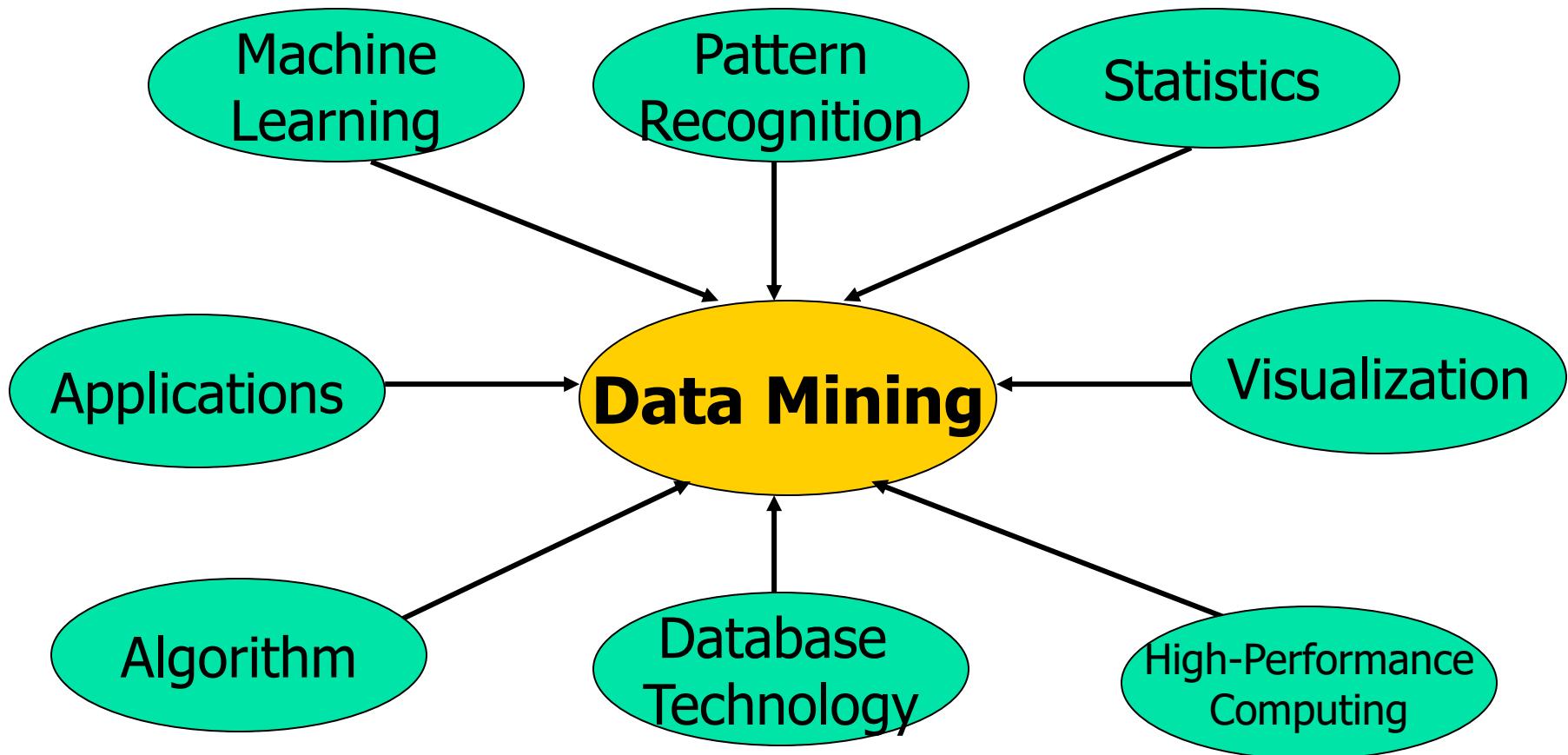
عملیات داده کاوی

- توصیف و تفکیک داده ها
- کاوش الگوهای مکرر، مشارکت ها و وابستگی ها
- دسته بندی و رگرسیون برای تحلیل پیشگویی
- تحلیل خوش
- تحلیل داده های پرت

فصل ۱ : مقدمه

- چرا داده کاوی؟
- داده کاوی چیست؟
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوهای کاوش می شوند؟
- چه تکنولوژی هایی استفاده می شوند؟ 
- کدام برنامه های کاربردی هدف داده کاوی هستند؟
- موضوعات عمدی در داده کاوی
- تاریخچه ای از داده کاوی و جامعه داده کاوی
- خلاصه

داده کاوی: محل تلاقی رشته های مختلف



فصل ۱ : مقدمه



- چرا داده کاوی؟
- داده کاوی چیست؟
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوهای کاوش می شوند؟
- کدام فن آوری های استفاده می شوند؟
- کدام برنامه های کاربردی هدف داده کاوی هستند؟
- موضوعات عمدی در داده کاوی
- تاریخچه ای از داده کاوی و جامعه داده کاوی
- خلاصه

کاربردهای داده کاوی

- تحلیل صفحات وب: دسته بندی صفحات، خوش بندی، رتبه بندی صفحات...
- تحلیل همبستگی و سیستم های پیشنهاد دهنده
- تحلیل سبد خرید و بازاریابی هدفمند
- تحلیل داده های زیستی و پزشکی
- ... ■

فصل ۱ : مقدمه

- چرا داده کاوی؟
- داده کاوی چیست؟
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوهای کاوش می شوند؟
- کدام فن آوری ها استفاده می شوند؟
- کدام برنامه های کاربردی هدف داده کاوی هستند؟
-  موضوعات عمدی در داده کاوی
- تاریخچه ای از داده کاوی و جامعه داده کاوی
- خلاصه

موضوعات عمدہ در داده کاوی

- متدولوزی کاوش
- تعامل کاربر
- کارامد بودن و قابلیت مقیاس پذیری
- تنوع گونه های پایگاه داده
- داده کاوی و جامعه

فصل ۱ : مقدمه



تاریخچه ای از داده کاوی و جامعه داده کاوی

خلاصه

- چرا داده کاوی؟
- داده کاوی چیست؟
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوهای کاوش می شوند؟
- کدام فن آوری ها استفاده می شوند؟
- کدام برنامه های کاربردی هدف داده کاوی هستند؟
- موضوعات عمدی در داده کاوی

تاریخچه مختصر جامعه داده کاوی

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
 - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
 - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
 - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
 - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

کنفرانس ها و ژورنال ها

- KDD Conferences
 - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (**KDD**)
 - SIAM Data Mining Conf. (**SDM**)
 - (IEEE) Int. Conf. on Data Mining (**ICDM**)
 - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (**ECML-PKDD**)
 - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (**PAKDD**)
 - Int. Conf. on Web Search and Data Mining (**WSDM**)
- Other related conferences
 - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ...
 - Web and IR conferences: WWW, SIGIR, WSDM
 - ML conferences: ICML, NIPS
 - PR conferences: CVPR,
- Journals
 - Data Mining and Knowledge Discovery (DAMI or DMKD)
 - IEEE Trans. On Knowledge and Data Eng. (TKDE)
 - KDD Explorations
 - ACM Trans. on KDD

فصل ۱ : مقدمه

- چرا داده کاوی؟
- داده کاوی چیست؟
- یک نمای چند وجهی از داده کاوی
- چه انواعی از داده کاوش می شوند؟
- چه انواعی از الگوهای کاوش می شوند؟
- کدام فن آوری ها استفاده می شوند؟
- کدام برنامه های کاربردی هدف داده کاوی هستند؟
- موضوعات عمدی در داده کاوی
- تاریخچه ای از داده کاوی و جامعه داده کاوی
- خلاصه

خلاصه

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining technologies and applications
- Major issues in data mining

کتب مرجع پیشنهادی

- **S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data. Morgan Kaufmann, 2002**
- **R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000**
- **T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003**
- **U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996**
- **U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001**
- **J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011**
- **D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001**
- **T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009**
- **B. Liu, Web Data Mining, Springer 2006.**
- **T. M. Mitchell, Machine Learning, McGraw Hill, 1997**
- **G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991**
- **P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005**
- **S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998**
- **I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005**



داده کاوی

مفاهیم و تکنیک ها

— فصل ۲ —

فصل ۲: داده های خود را بیشتر بشناسید



- اشیاء داده ای و انواع صفات

- توصیفات آماری پایه از داده ها

- مصور سازی داده ها

- اندازه گیری میزان شباهت و عدم شباهت داده ها

- خلاصه

انواع مجموعه داده ها (Data Sets)

	team	coach	pla	ball	score	game	wi	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

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

- رکوردها
 - رکوردهای رابطه ای
 - ماتریس داده نظیر ماتریس عددی یا crosstab
 - اسناد داده ای مانند متن ها یا term-frequency vector
 - داده تراکنشی
 - گراف و شبکه
- وب
 - شبکه های اجتماعی یا اطلاعاتی
 - ساختار های مولکولی
 - داده های ترتیبی (Ordered)
 - داده ویدئویی (دباله ای از تصاویر)
 - داده های وابسته به زمان (Temporal) مانند سری های زمانی
 - داده های متوالی (sequential): دنباله ای از تراکنش ها
 - داده توالی ژنتیک
 - داده های فضایی، تصویری و چند رسانه ای
 - داده های فضایی: نقشه ها
 - داده تصویری
 - داده ویدیویی

ویژگی های مهم داده های ساخت یافته

- ابعاد
- پراکندگی یا تنک بودن
- وضوح یا تفکیک پذیری
- توزیع

اشیاء داده ای (Data Objects)

- مجموعه های داده ای از اشیاء داده ای تشکیل شده اند
- یک شیء داده ای نمایانگر یک موجودیت است.
- مثال:
- پایگاه داده فروش: مشتریان، اقلام فروشگاه، فروش ها
- پایگاه داده پزشکی: بیماران، معالجات
- پایگاه داده دانشگاه: دانشجویان، اساتید، دروس
- نام های دیگر اشیاء داده ای: *samples, examples, instances, data points, objects, tuples.*
- اشیاء داده ای توسط صفات (Attributes) توصیف می شوند.
- سطر های پایگاه داده: اشیاء داده ای ستون ها: صفات

صفات

- صفت (یا بعد، ویژگی، متغیر) یک فیلد داده ای یک خصوصیت یک شیء داده ای را نشان میدهد.
- مثل: شماره مشتری، نام، آدرس
- انواع:
 - اسمی یا Nominal
 - دودویی یا Binary
 - ترتیبی یا Ordinal
 - عددی یا Numeric
 - مقیاس بازه ای یا Interval-scaled
 - مقیاس نسبتی یا Ratio-scaled

انواع صفات

اسمی

- گروه‌ها، حالت‌ها یا نام چیزها
- مثل رنگ مو یا وضعیت تاہل، شناسه مشتری، کدپستی
- مقادیر ممکن است بصورت عدد بیان شوند اما ارزش عددی ندارند و اعمال جبری روی آن‌ها بی معنی است.

دودویی

- صفت اسمی که فقط دو مقدار یا حالت ممکن دارد. (۰ و ۱)
- دودویی متقارن(Symmetric binary): هر دو طرف اهمیت یکسان دارد مثل جنسیت
- دودویی نامتقارن(Asymmetric binary): اهمیت دو طرف یکسان نیست. مثل نتیجه مثبت یا منفی در مورد یک تست پزشکی (مقدار یک به طرف پراهمیت تر نسبت داده می‌شود مثلاً HIV+)

ترتیبی

- مقادیر ترتیب معنی دار دارند اما مقدار بین مقادیر شناخته شده وجود ندارد.
- مثل سایز(کوچک، متوسط، بزرگ)، نمرات(B-,B,B+,A-,A,A+,...)، درجات نظامی

انواع داده عددی

داده های عددی کمیت (مقادیر صحیح یا حقیقی) هستند و بر دو نوع دند:

فاصله یا Interval

- اندازه گیری در یک مقیاس از واحدهای هم اندازه
- مقادیر دارای ترتیب
- مثل درجه حرارت در واحدهای سانتیگراد یا فارنهایت یا روزهای تقویم
- عدم وجود نقطه صفر واقعی

نسبت یا Ratio

- صفر واقعی دارد.
- مقادیر به معنی چند برابر بزرگتر از واحد اندازه گیری هستند. مثلا ۱۰ کیلوگرم ۲ برابر ۵ کیلوگرم است.
- مثل درجه حرارت کلوین، طول، تعداد، مقدار پول

صفات گسته یا پیوسته

صفات گسته

مجموعه مقادیر متناهی یا نامتناهی قابل شمارش

- مثل کد پستی، مجموعه کلمات یک متن، سن
- گاهی به شکل یک عدد صحیح نمایش داده می شود.
- صفات دودویی شکل خاصی از صفات گسته هستند.

صفات پیوسته

- اعداد حقیقی را به عنوان مقدار می پذیرند.
- مثل دما، قد، وزن
- در عمل اعداد حقیقی در قالب مقادیر متناهی از ارقام اندازه گیری و نمایش داده می شوند.
- معمولًا به شکل متغیر های ممیز شناور نمایش داده می شوند.

فصل ۲: داده های خود را بیشتر بشناسید



- اشیاء داده ای و انواع صفات
- توصیفات آماری پایه از داده ها
- مصور سازی داده ها
- اندازه گیری میزان شباهت و عدم شباهت داده ها
- خلاصه

توصیفات آماری پایه از داده ها



شاخص های مرکزی

میانگین، میانه، مد



شاخص های پراکندگی داده

دامنه تغییرات، چارک ها و دامنه بین چارکی، واریانس و انحراف معیار



نمایش گرافیکی آمار توصیفی

اندازه گیری مرکزیت داده ها

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

میانگین ■

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

■ میانگین وزنی

- میانگین هرس شده (حذف مثلا ۲% مقادیر بالا و پایین برای از بین بردن اثر سوء داده های پرت)

میانه

■ داده وسطی در داده های مرتب شده با تعداد فرد یا میانگین دو داده وسط در داده های با تعداد زوج

age	frequency	
1–5	200	L1 کران پایین دسته میانه
6–15	450	n تعداد کل داده ها
16–20	300	($\sum freq$)l.
21–50	1500	جمع همه فراوانی های قبل از بازه میانه
51–80	700	freq _{median} روانی بازه میانه
81–110	44	عرض بازه میانه Width

$$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}} \right) width$$

نما

- داده ای که بیشترین تکرار را دارد.
- داده های با یک نما Unimodal، دو نما bimodal یا سه نما trimodal
- داده بدون نما یا no mode داده ای که از هر مقدار فقط یکبار تکرار شده باشد مثل شماره مشتری فرمول تجربی

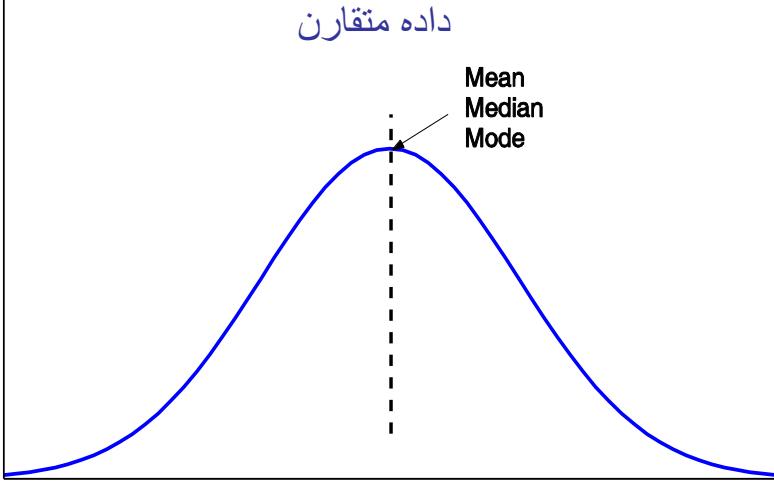
$$mean-mode = 3 \times (mean-median)$$

داده های متقارن یا نامتوازن

۱

داده متقارن

Mean
Median
Mode

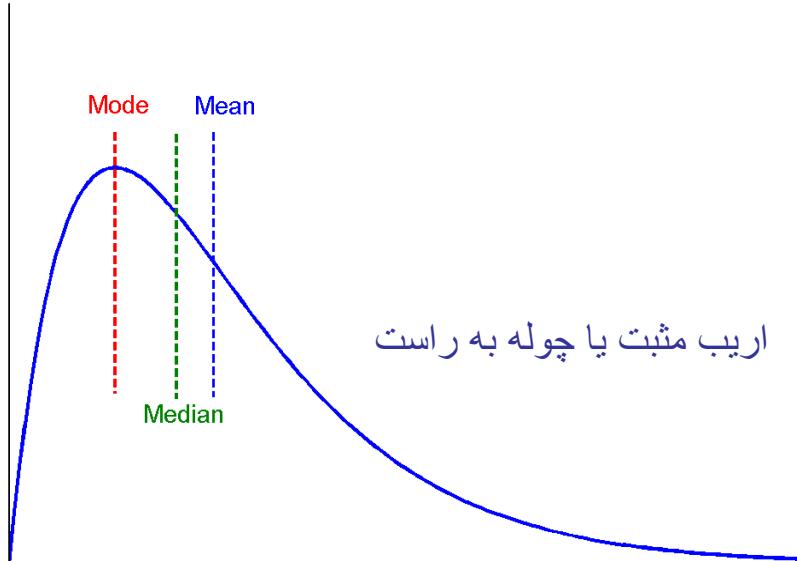


Mode

Mean

Median

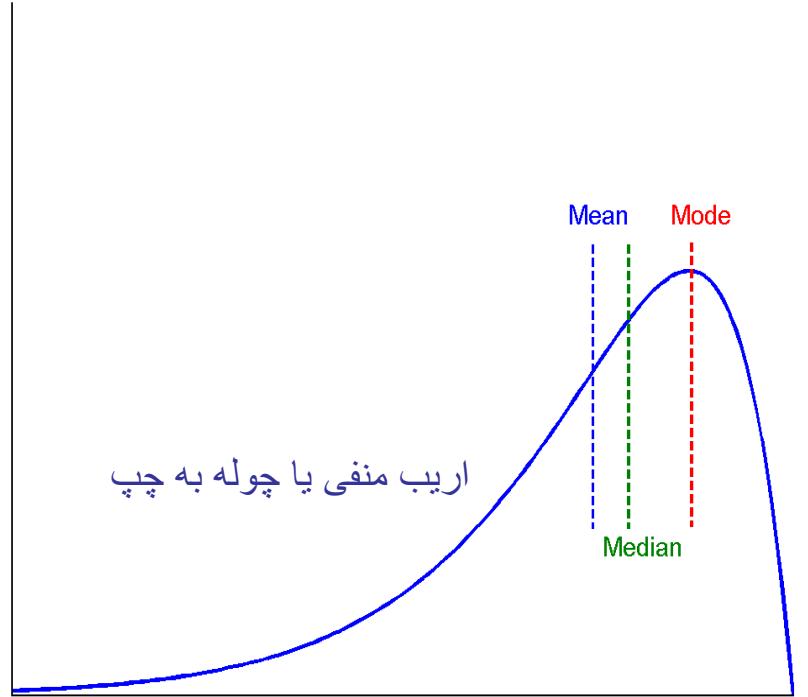
اریب مثبت یا چوله به راست



- میانگین، میانه و نما در داده های متقارن، چوله به راست یا به چپ

Mean
Mode
Median

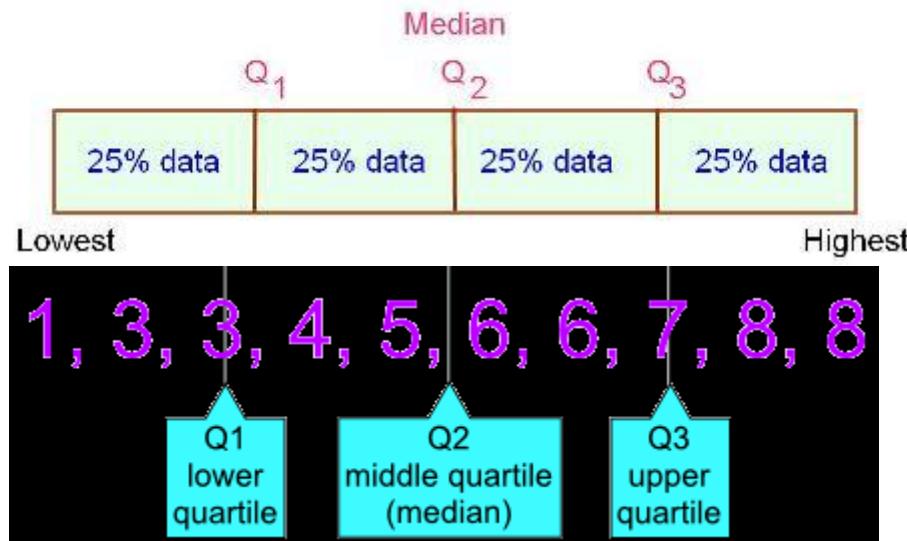
اریب منفی یا چوله به چپ



اندازه گیری پراکندگی داده

چارک ها، داده های پرت و نمودارهای جعبه ای (Quartiles, outliers and boxplots) ■

Quartiles and data distribution

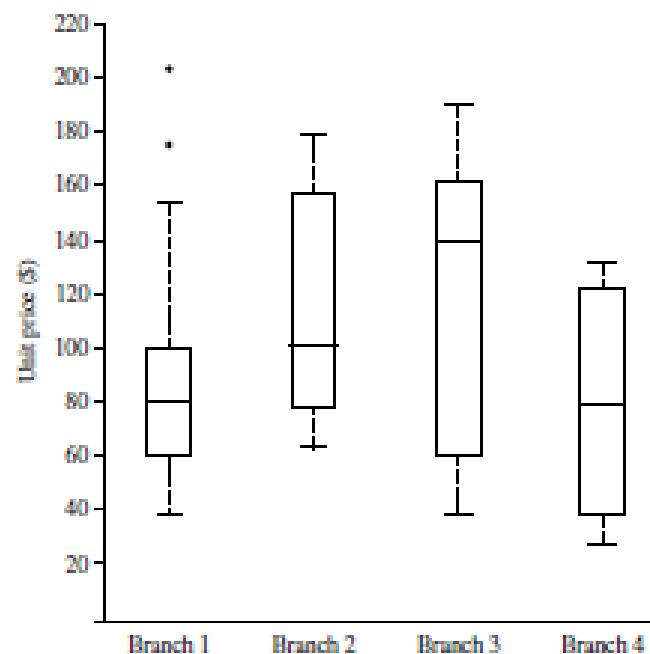
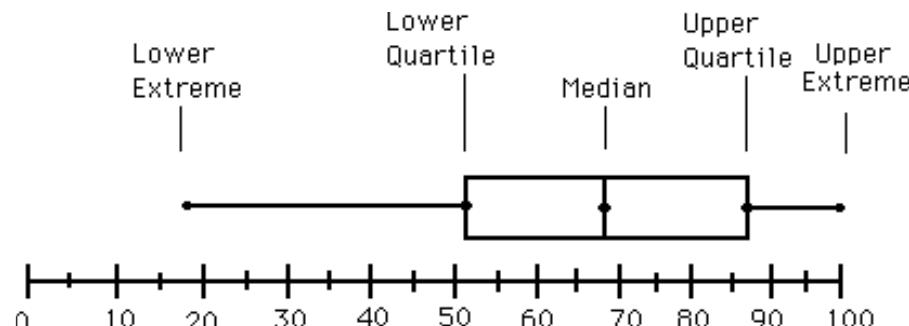


چارک ها: Q1 (۲۵ امین صدک)، Q3 (۷۵ امین صدک)

Inter-quartile range: $IQR = Q_3 - Q_1$: دامنه بین چارکی

Five number summary: \min , Q_1 , median, Q_3 , \max

- نمودار جعبه ای: دو انتهای جعبه چارک ها هستند. میانه با رسم خطی در میان جعبه علامتگذاری می شود. کوچکترین و بزرگترین مقدار نیز با دو خط خارج جعبه که به طرف آنها رسم شده است، نشان داده می شوند.
- داده پرت: معمولاً مقداری که خارج از محدوده $1/5$ برابر IQR قرار دارند.



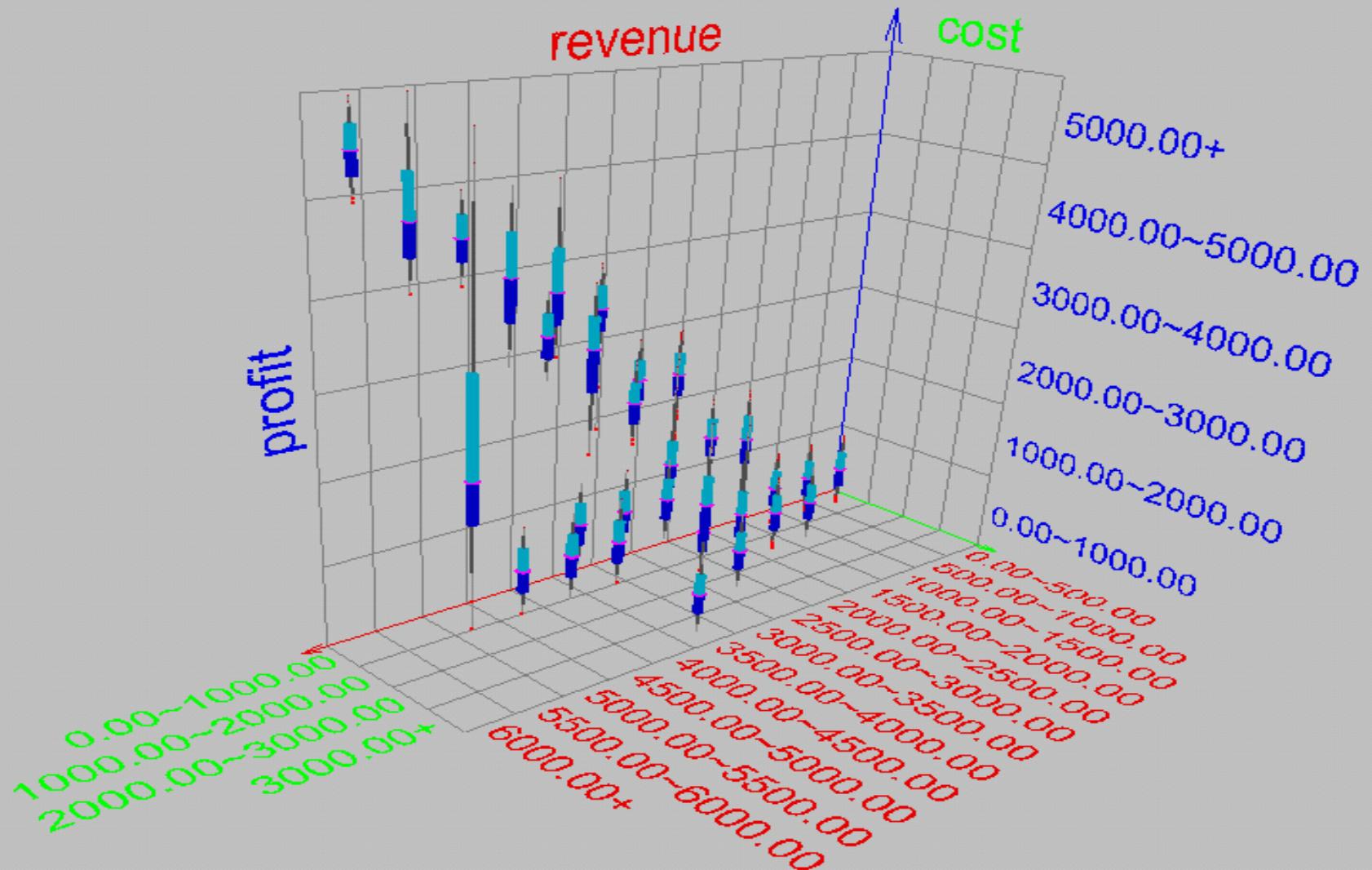
واریانس و انحراف معیار

- برای تعداد کم داده از s و برای تعداد زیاد از σ استفاده می‌شود
- انحراف معیار جذر واریانس است.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 \right]$$

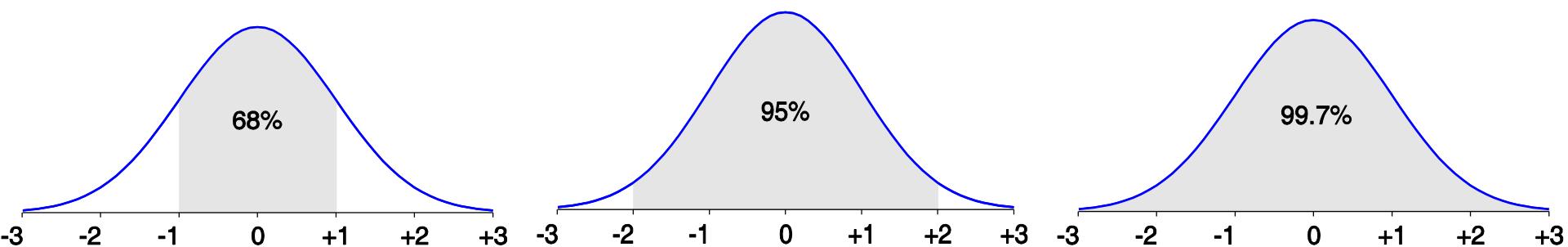
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

تصویر سازی پردازشگر داده ها: نمودار جعبه ای سه بعدی



ویژگی های منحنی توزیع نرمال

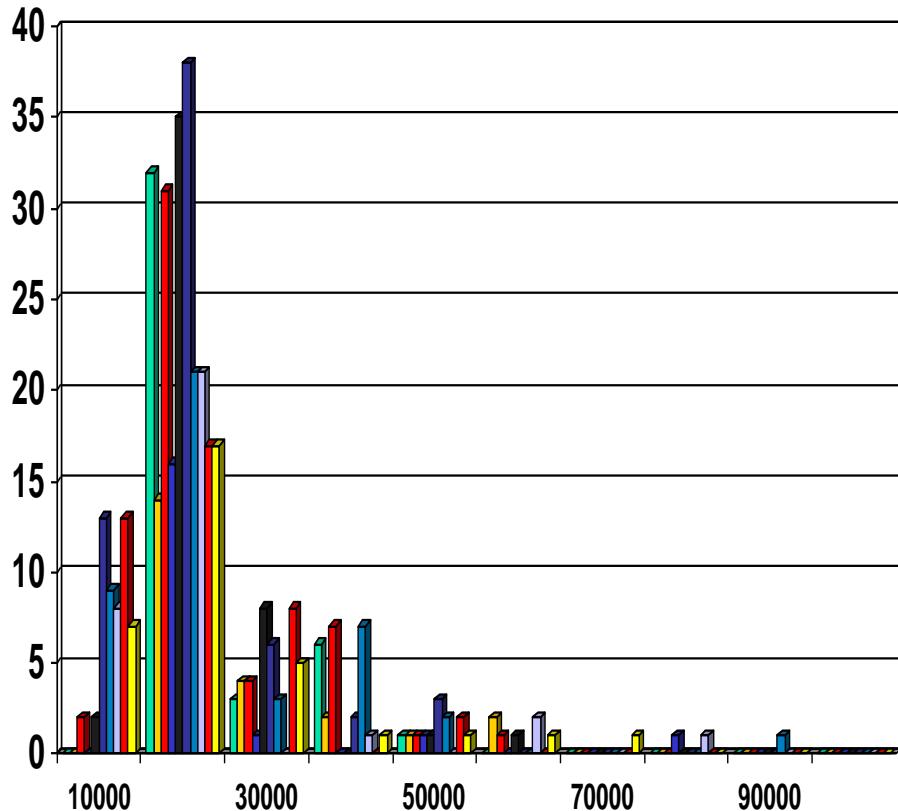
- منحنی توزیع نرمال
- From $\mu-\sigma$ to $\mu+\sigma$: contains about 68% of the measurements (μ : mean, σ : standard deviation)
- From $\mu-2\sigma$ to $\mu+2\sigma$: contains about 95% of it
- From $\mu-3\sigma$ to $\mu+3\sigma$: contains about 99.7% of it



نمایش گرافیکی از توصیفات آماری

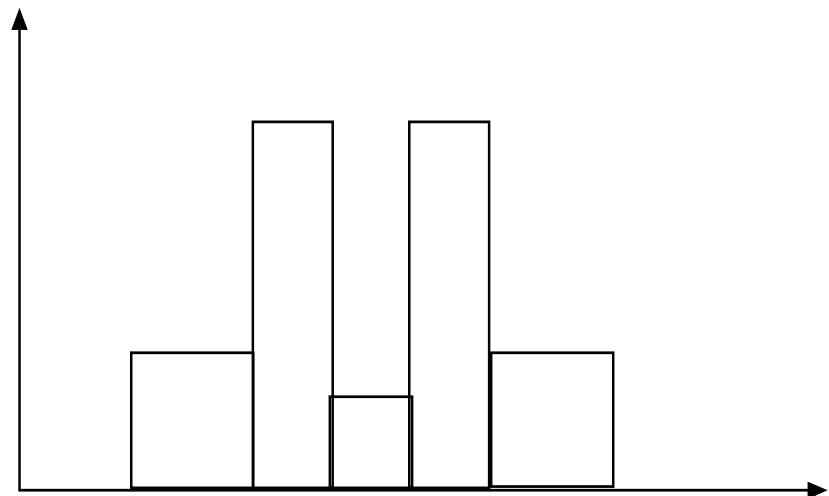
- نمایش گرافیکی از خلاصه پنج عددی : **Boxplot**
- محور X مقادیر و محور Y فرکانس تکرار : **Histogram**
- هر مقدار x_i با زوج f_i نشان میدهد که تقریبا $100 f_i \%$ داده ها کوچکتر مساوی x_i هستند. : **Quantile plot**
- نمودار quantiles : **Quantile-quantile (q-q) plot**
- در مقابل quantiles مربوطه از دیگری univariant
- هر جفت از مقادیر یک جفت مختصات هستند که بصورت Scatter plot
- یک نقطه در صفحه ترسیم می شوند.

تحلیل هیستوگرام



- هیستوگرام: شمای گرافیکی از فرکانس های جدول بندی شده بصورت نمودار میله ای این نمودار نشان می دهد که چه نسبت از موارد در هر یک از دسته بندی های مختلف قرار می گیرند.
- تفاوت این نمودار با نمودار میله ای این است که سطح زیر هر میله ارزش آن را نشان میدهد نه ارتفاع آن. این تفاوت زمانی که دسته ها عرض یکسان نداشته باشند مشخص می شود.
- این دسته ها معمولاً به عنوان فواصل غیر همپوشای یک متغیر تعریف می شوند. دسته ها (میله ها) باید مجاور باشند.

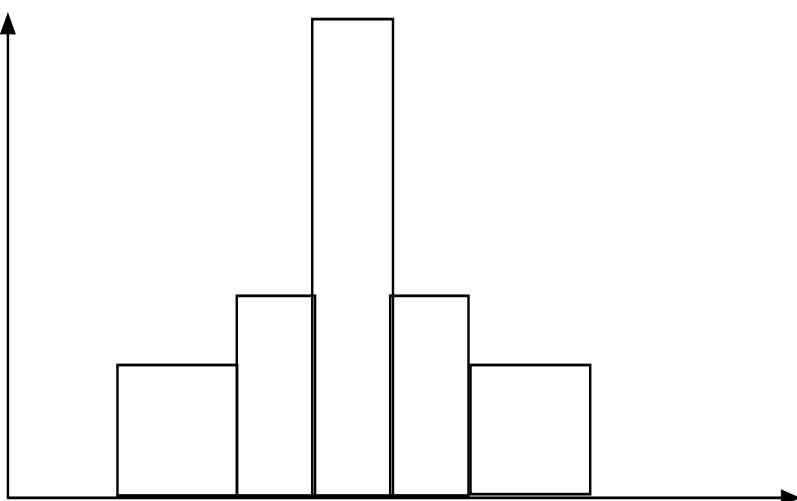
هیستوگرام ها معمولاً بیشتر از نمودار جعبه ای اطلاعات منتقل می کنند.



- دو هیستوگرام روبرو ممکن است نمودار جعبه ای یکسانی داشته باشند.

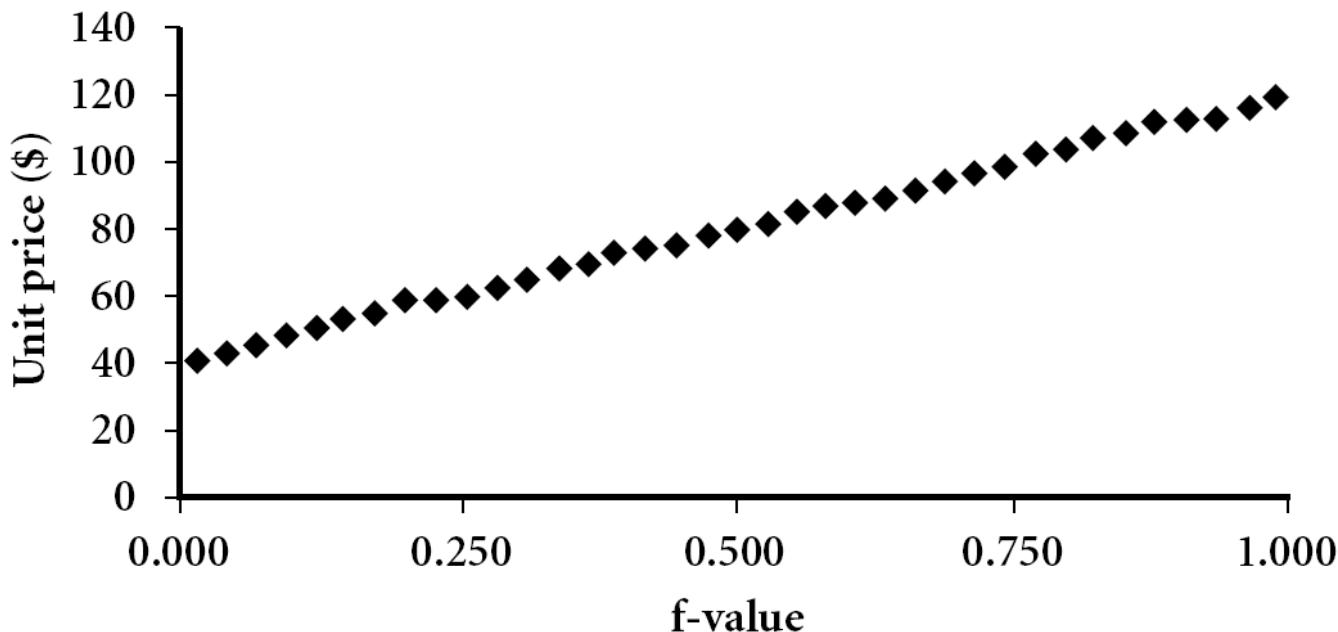
- مقادیر مشترک برای :
 \min , $Q1$, median,
 $Q3$, \max

- اما آنها توزیع داده های متفاوتی دارند.



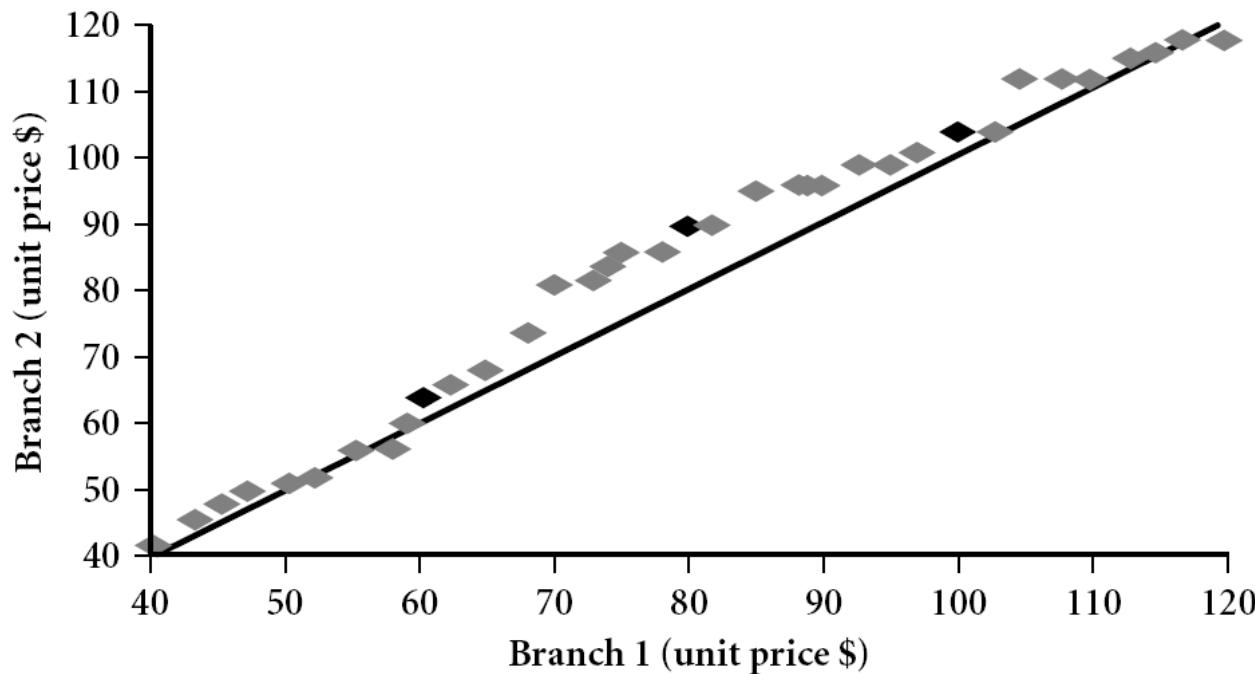
نمودار چندک یا Quantile

- همه داده ها را نشان میدهد و امکان ارزیابی رفتار کلی داده و مشاهده مقادیر غیر عادی را برای کاربر فراهم میکند.
- اطلاعات نمودار چندک
- برای هر داده x_i که بصورت صعودی مرتب شده اند f_i نشان میدهد که تقریباً $f_i \cdot 100$ درصد از داده ها کوچکتر یا مساوی مقدار x_i هستند.



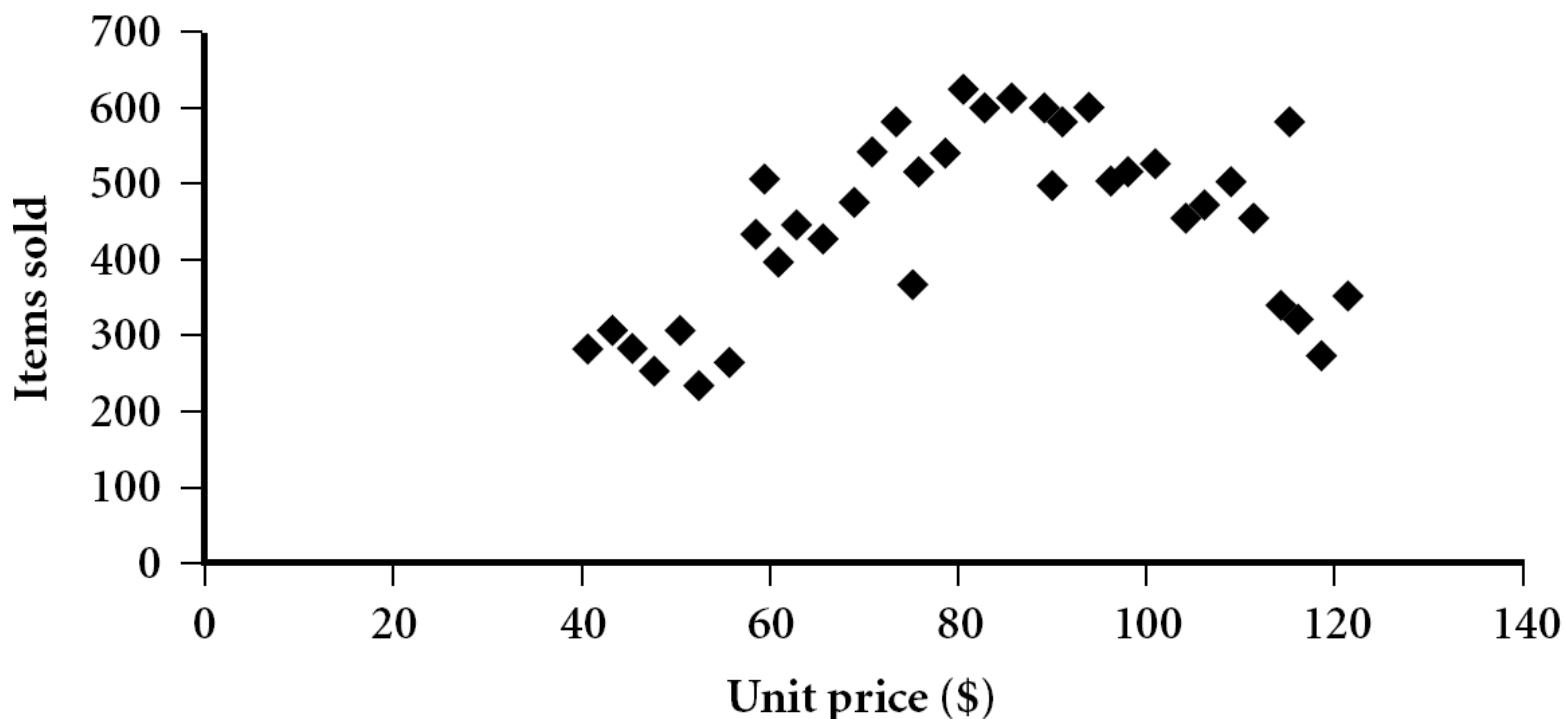
نمودار Quantile-Quantile (Q-Q)

- چندک یک توزیع تک متغیره را در مقابل چندک یک توزیع دیگر نمایش میدهد.
- این مثال قیمت اقلام فروش رفته در شعبه یک در مقابل قیمت اقلام در شعبه ۲ در هر چندک نشان میدهد. نمودار نشان می دهد در شعبه یک اقلام با قیمت کمتر بیشتر از شعبه ۲ به فروش رفته است.

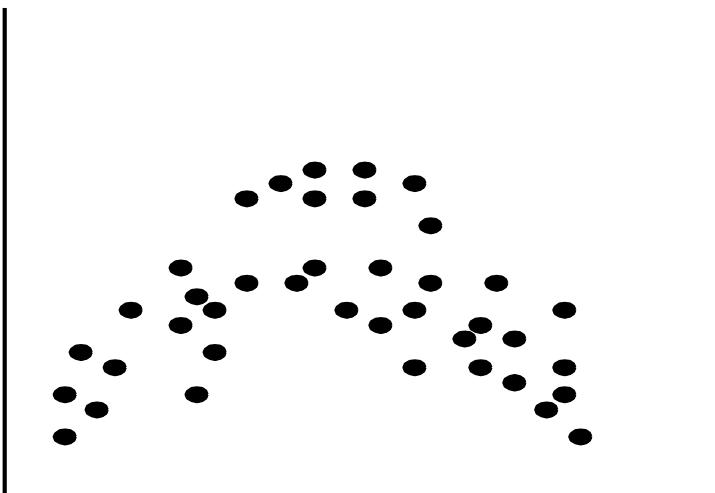
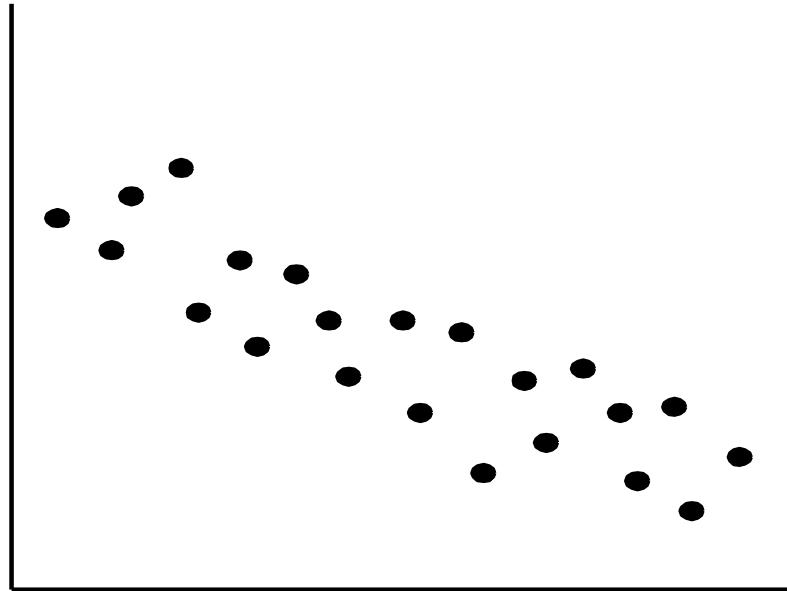
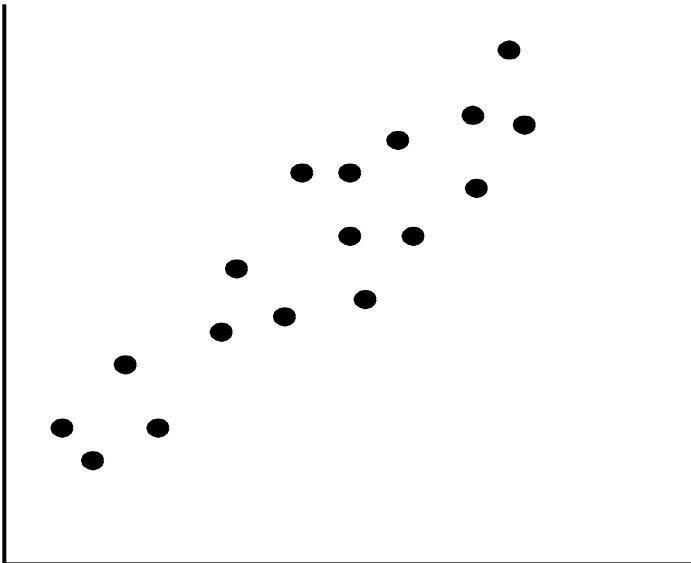


نمودار پراکنشی یا Scatter plot

- یک دید اولیه از یک داده با دو متغیر برای دیدن خوشی ها، داده های پرت و فراهم میکند.
- هر جفت مقدار از دو متغیر بعنوان مختصات یک نقطه در نظر گرفته و رسم می شود.

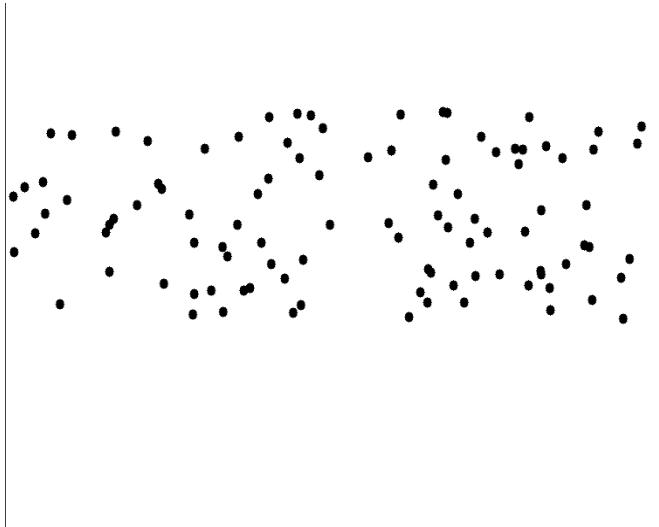
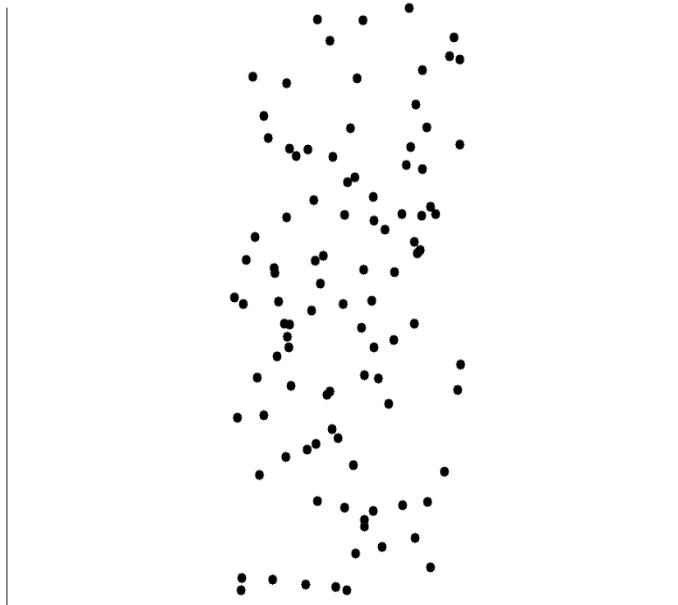


داده های با همبستگی مثبت و منفی



در این نمودار نیمه سمت چپ همبستگی مثبت و نیمه سمت راست همبستگی منفی دارد.

داده های بدون همبستگی



فصل ۲: داده های خود را بیشتر بشناسید

-
- اشیاء داده ای و انواع صفات
 - توصیفات آماری پایه از داده ها
 - مصور سازی داده ها
 - اندازه گیری میزان شباهت و عدم شباهت داده ها
 - خلاصه



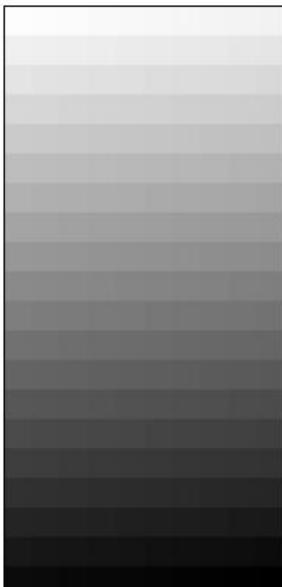
تصویرسازی داده ها

چرا تصویرسازی داده ها؟

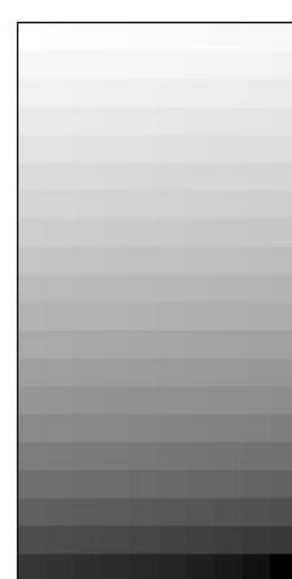
- دستیابی به نگرشی از فضای اطلاعاتی به کمک نگاشت داده ها به اشکال گرافیکی
- ارائه خلاصه کیفی از مجموعه داده های بزرگ
- جستجو برای الگوهای روندها، ساختار، بی نظمی ها، روابط بین داده ها
- کمک به پیدا کردن بخش های جالب و پارامتر های مناسب برای تجزیه و تحلیل کمی بیشتر
- ارائه برهان بصری از نمایش کامپیوترا بست آمده
- دسته بندی روش های تصویر سازی
 - تصویرسازی پیکسل گرا
 - تکنیک های تصویر کردن هندسی
 - تکنیک های تصویرسازی مبتنی بر شمایل
 - تکنیک های سلسله مراتبی تصویرسازی
 - تصویرسازی داده ها و روابط پیچیده

تصویرسازی پیکسل گرا

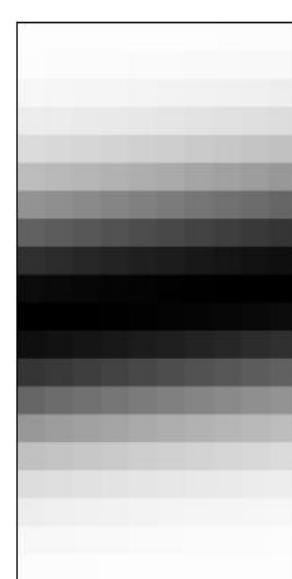
- برای یک مجموعه داده با m بعد m پنجره ایجاد می شود. هر پنجره مربوط به یک بعد است.
- مقادیر m بعد یک رکورد در پیکسل های پنجره متناظر با آن بعد تصویر می شود.
- رنگ پیکسل ها بیانگر مقادیر است.



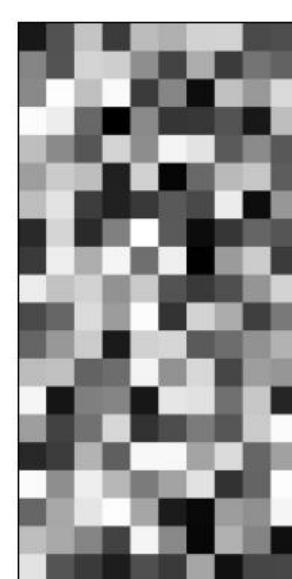
(a) درآمد



(b) حد اعتبار



(c) حجم تراکنش

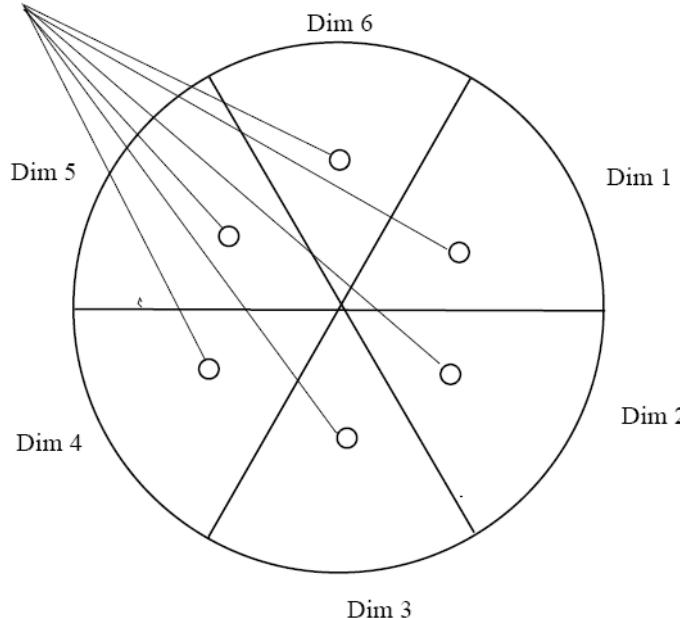


(d) سن

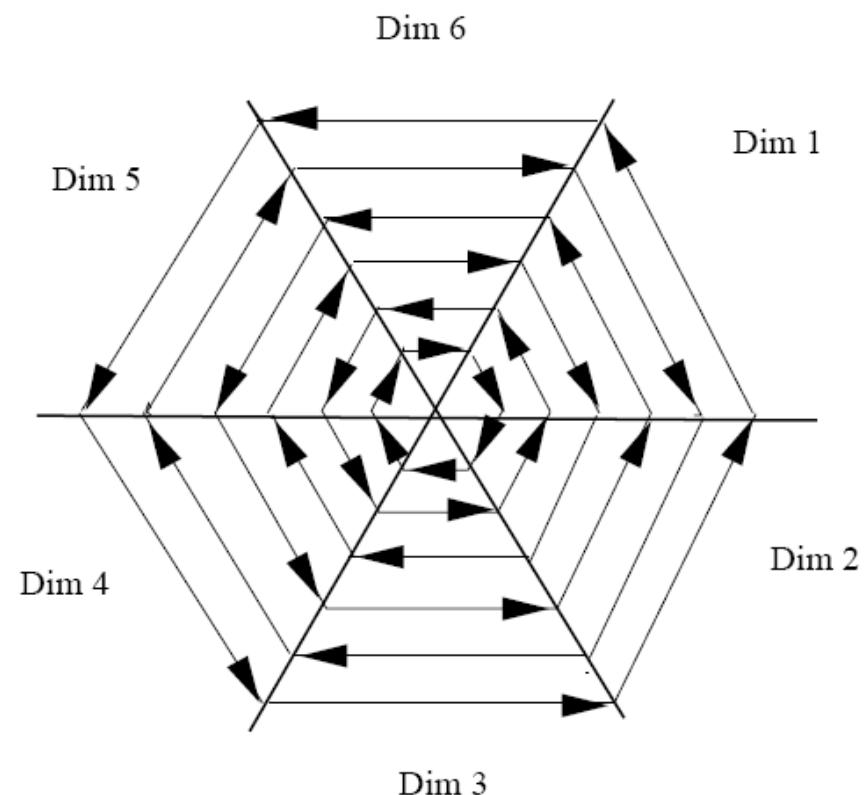
نمایش پیکسل ها در بخش های مختلف یک دایره

برای صرفه جویی در فضا و نشان دادن ارتباط بین ابعاد مختلف این کار معمولاً در بخش های مختلف یک دایره انجام می گیرد.

one data record



(a) نمایش یک رکورد داده در بخش های یک دایره



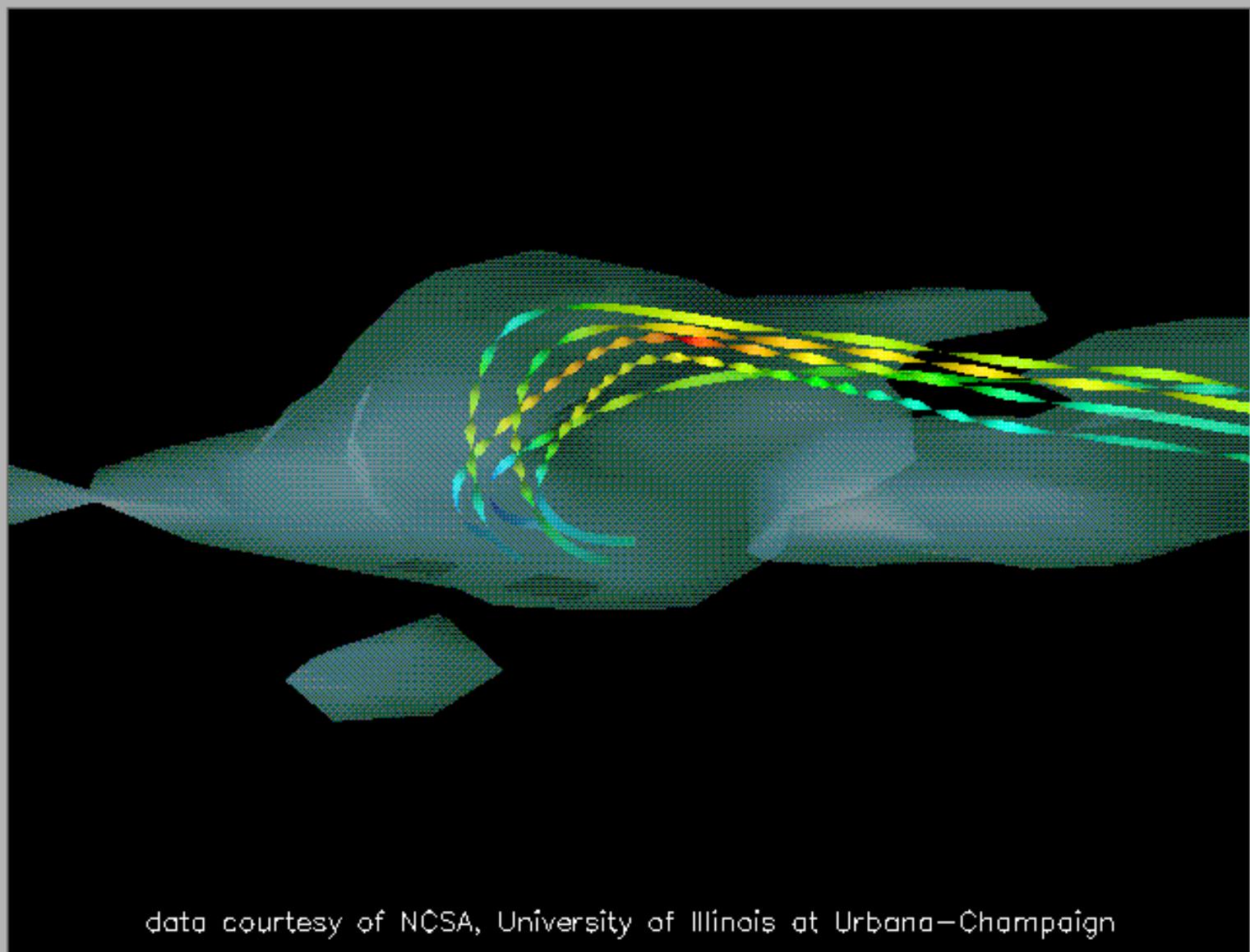
(b) جایدهی پیکسل ها در بخش های یک دایره

تکنیک های تصویر کردن هندسی

- مصورسازی تبدیل هندسی و تجسم داده
- روش ها:
 - Direct visualization
 - Scatterplot and scatterplot matrices
 - Landscapes
 - Projection pursuit technique: Help users find meaningful projections of multidimensional data
 - Prosection views
 - Hyperslice
 - Parallel coordinates

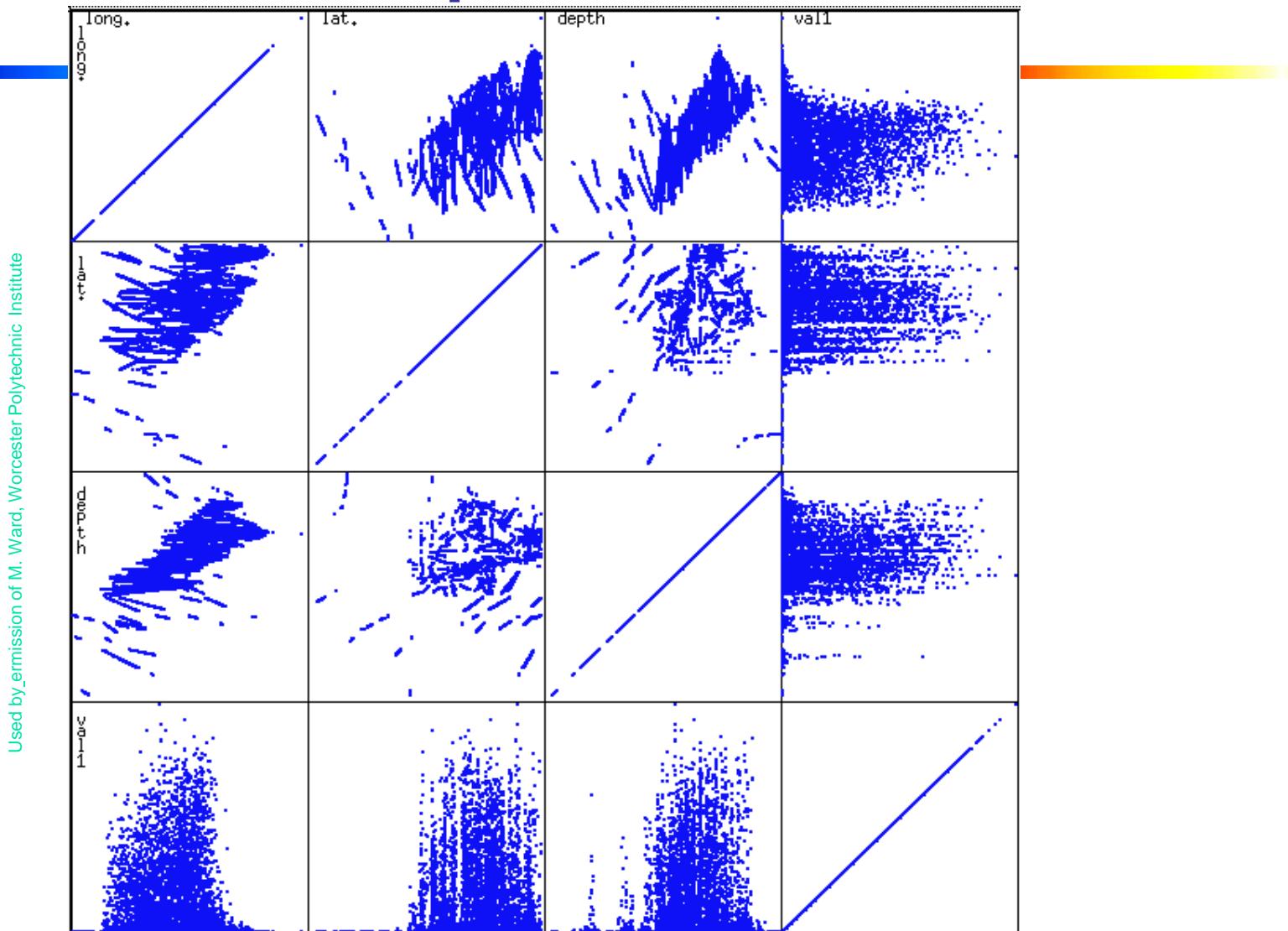
Direct Data Visualization

رویان بازیخواهی و تابعی و اساسی بر این اطمینان



data courtesy of NCSA, University of Illinois at Urbana-Champaign

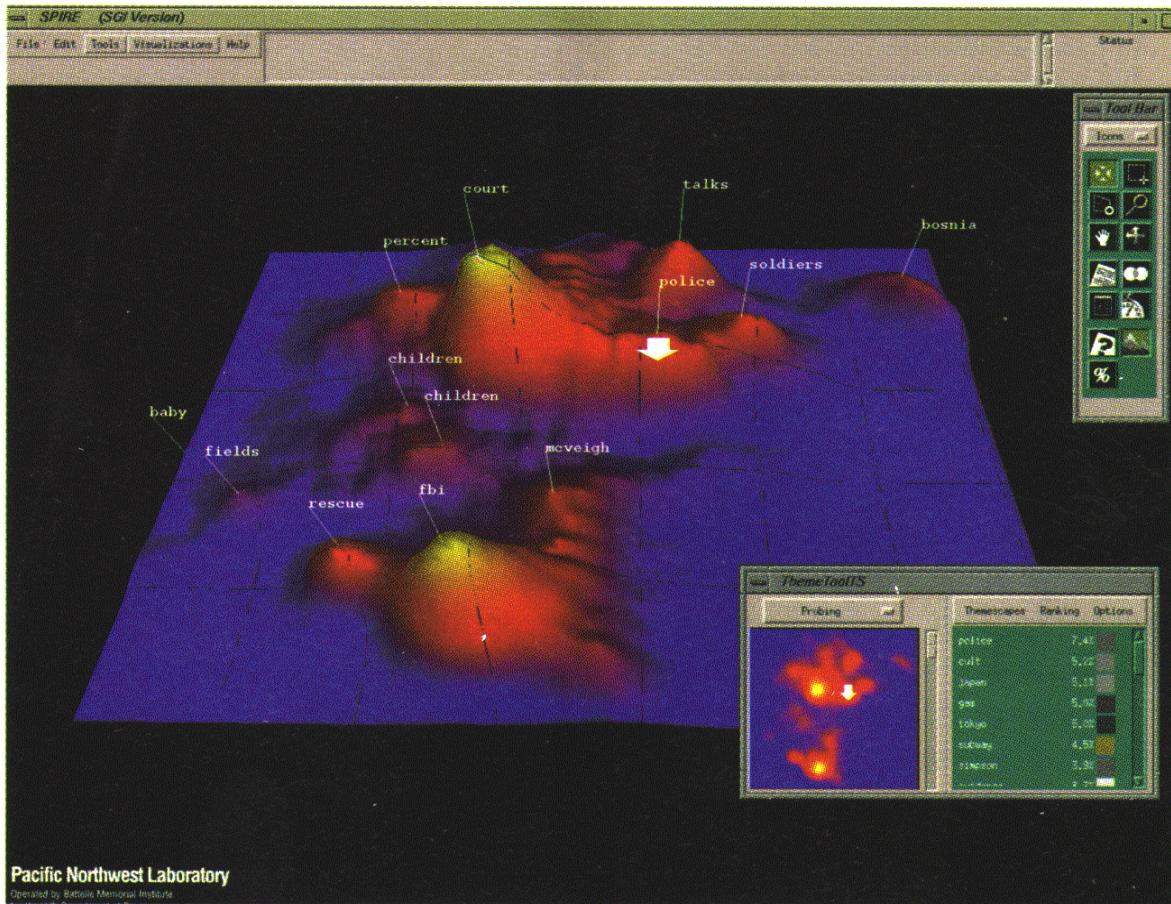
Scatterplot Matrices



Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of $(k^2/2-k)$ scatterplots]

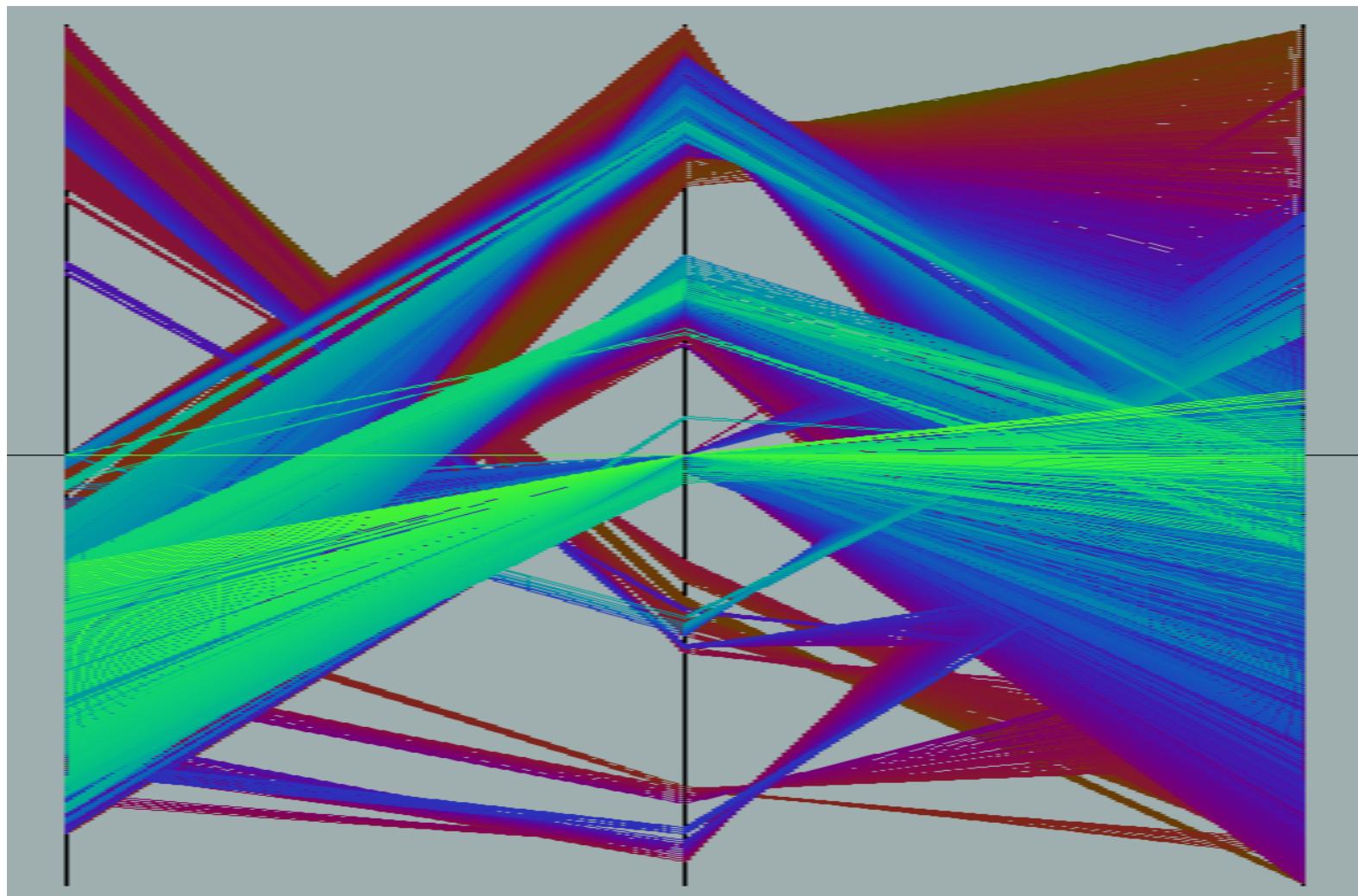
Landscapes

Used by permission of B. Wright, Visible Decisions Inc.



news articles
visualized as
a landscape

Parallel Coordinates of a Data Set



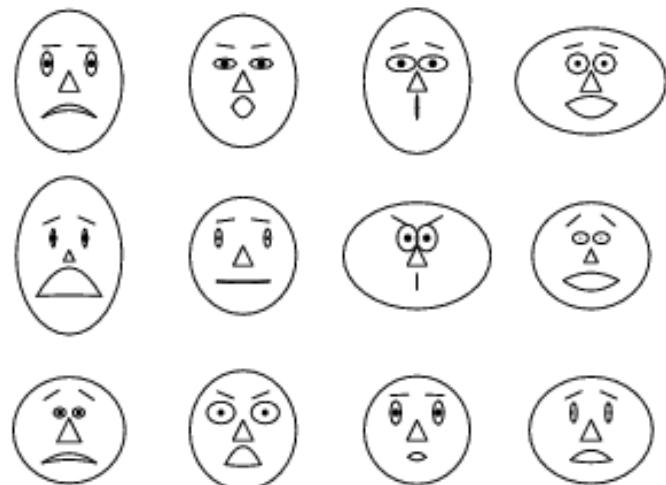
تکنیک های مصورسازی مبتنی بر شمایل

- مصورسازی مقادیر داده با مشخصات چهره ها
- روش های متداول مصورسازی
- Chernoff Faces
- Stick Figures
- روش های عمومی
- Shape coding: Use shape to represent certain information encoding
- Color icons: Use color icons to encode more information
- Tile bars: Use small icons to represent the relevant feature vectors in document retrieval

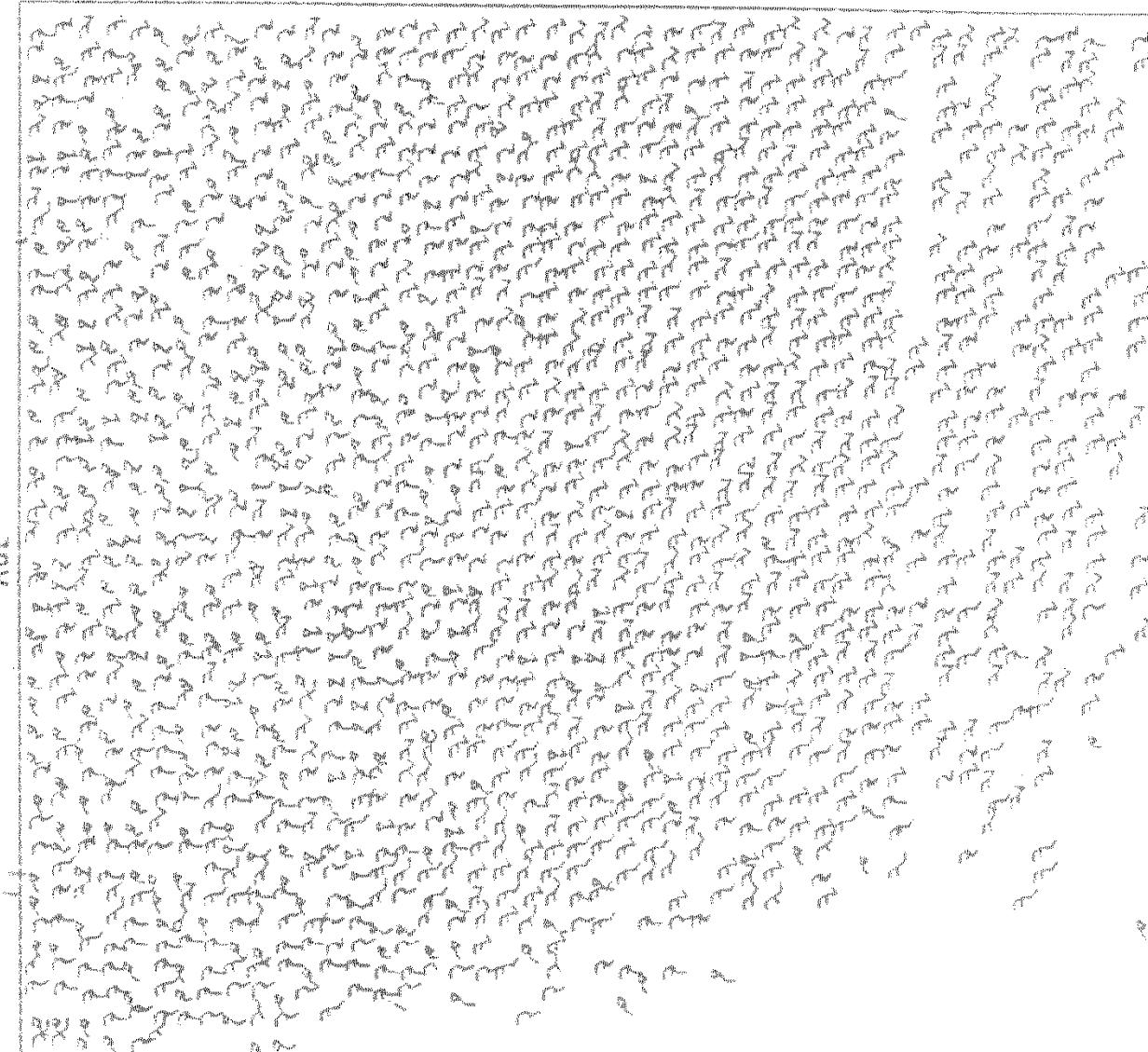
چهره های چرنوف

- راهی برای نمایش متغیرها در یک سطح دو بعدی مثلا فرض کنید X کجی ابرو، y اندازه چشم، z طول بینی و ... باشد.
- شکل صورت هایی را نشان می دهد که با استفاده از ۱۰ صفت خروج از مرکز سر، اندازه چشم، فاصله چشم، خروج از مرکز چشم، اندازه مردمک، کجی ابرو، اندازه بینی، شکل دهان، اندازه دهان و باز و بسته بودن دهان که به هر یک یکی از ۱۰ مقدار ممکن مناسب شده است با استفاده از Mathematica ایجاد شده است.

- REFERENCE: Gonick, L. and Smith, W. The Cartoon Guide to Statistics, New York: Harper Perennial, p. 212, 1993
- Weisstein, Eric W. "Chernoff Face." From *MathWorld--A Wolfram Web Resource*.
mathworld.wolfram.com/ChernoffFace.html



Stick Figure



داده های سرشماری
نشان دهنده سن،
درآمد، جنسیت،
تحصیلات و غیره

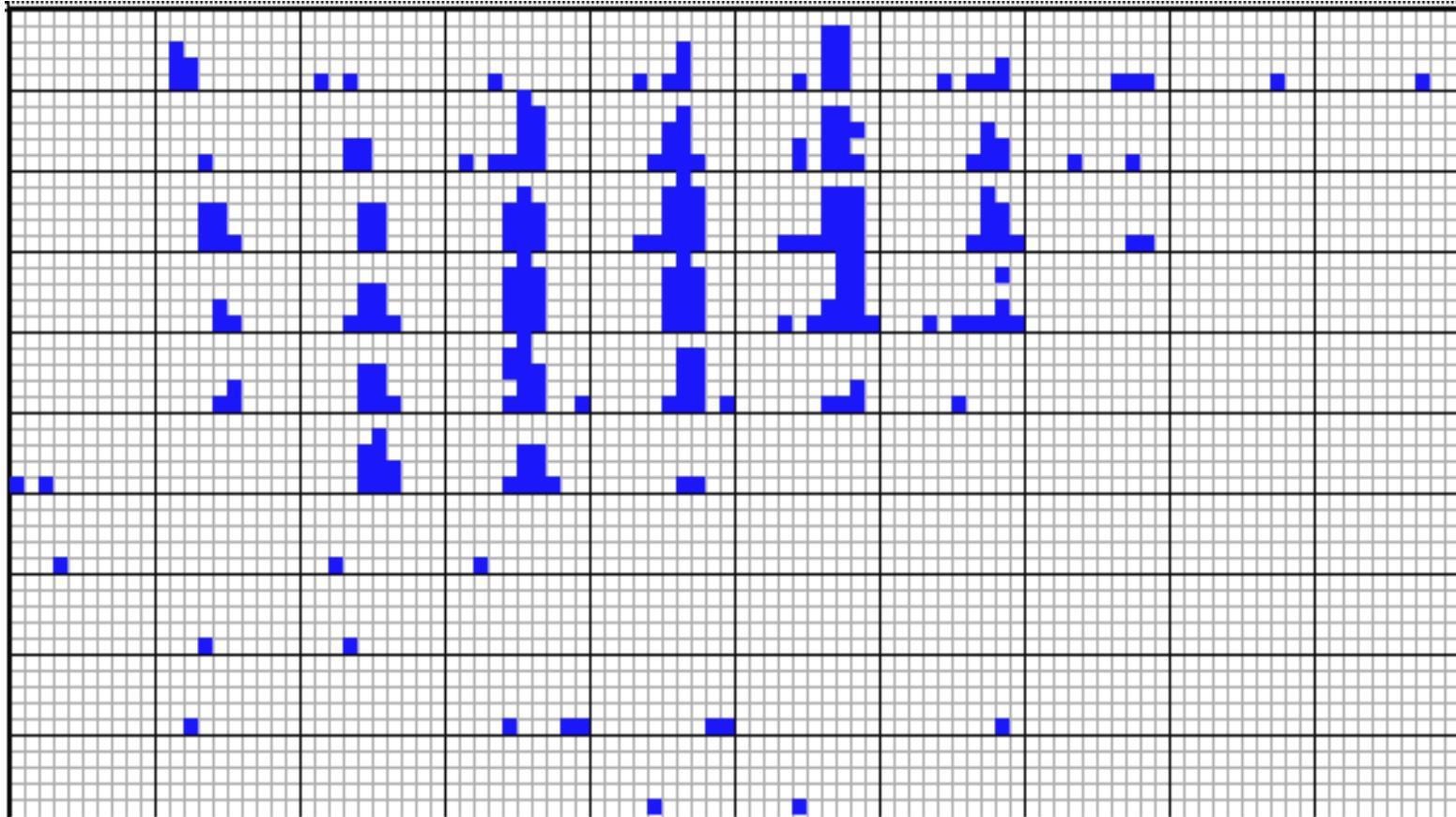
یک شکل ۵ قطعه ای
یک بدن و ۴ اندام با
طول و زاویه متفاوت

تکنیک های سلسله مراتبی مصورسازی

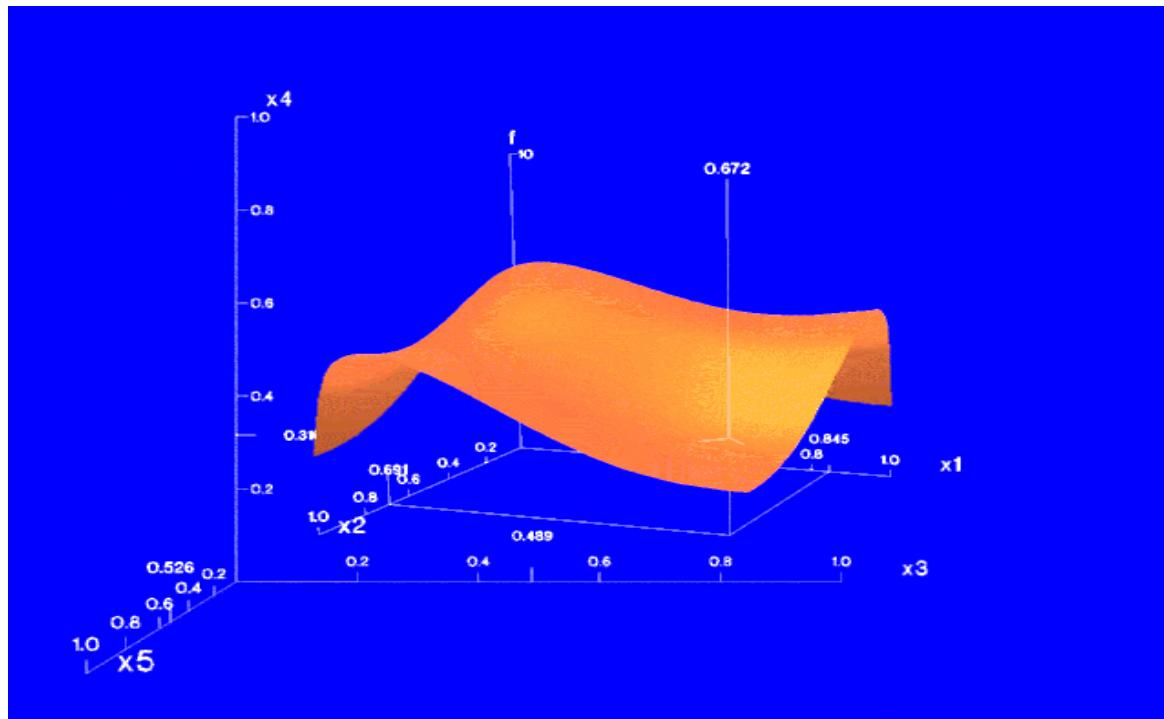
- مصورسازی داده با استفاده از افزار سلسله مراتبی به زیرفضاهها
- روش ها
- Dimensional Stacking
- Worlds-within-Worlds
- Tree-Map
- Cone Trees
- InfoCube

Dimensional Stacking

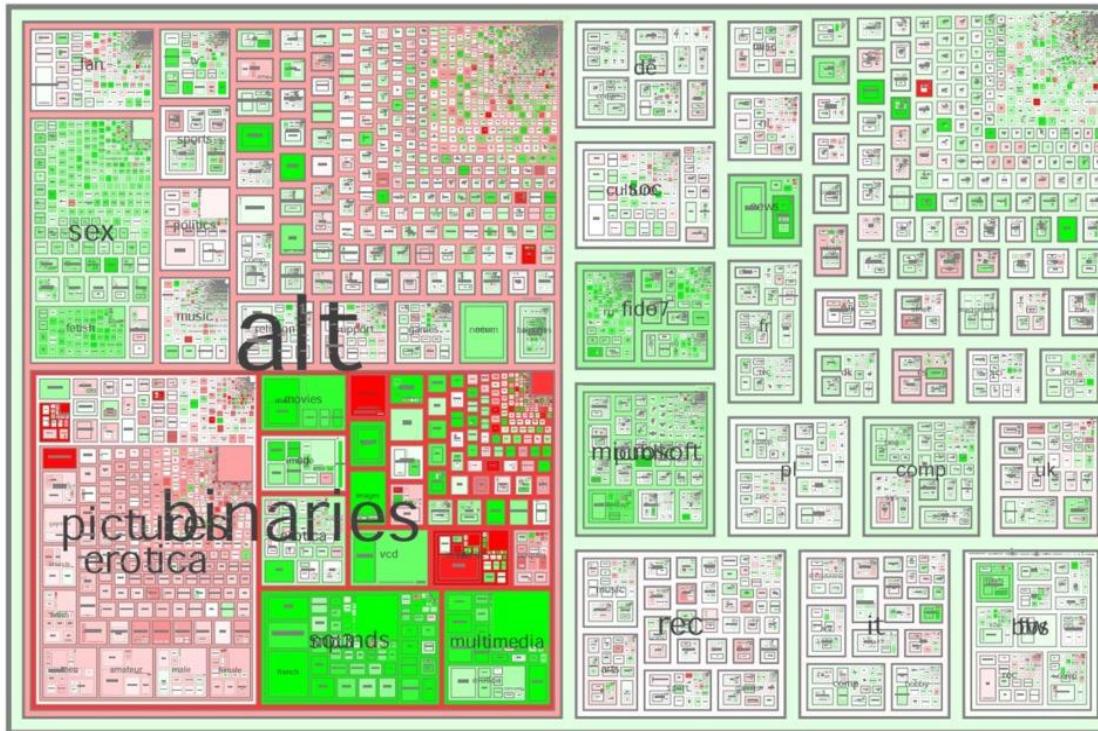
Used by permission of M. Ward, Worcester Polytechnic Institute



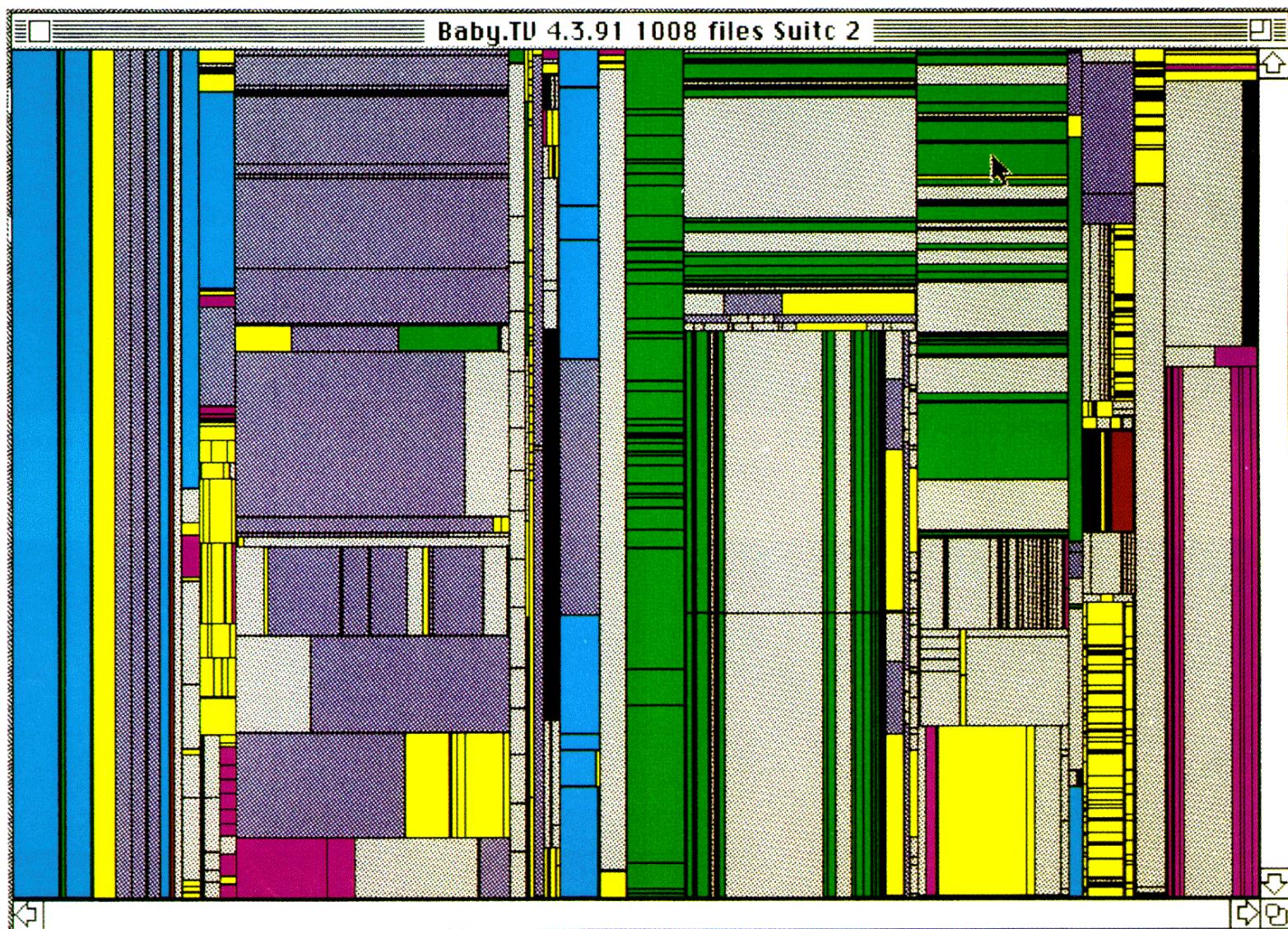
Worlds-within-Worlds



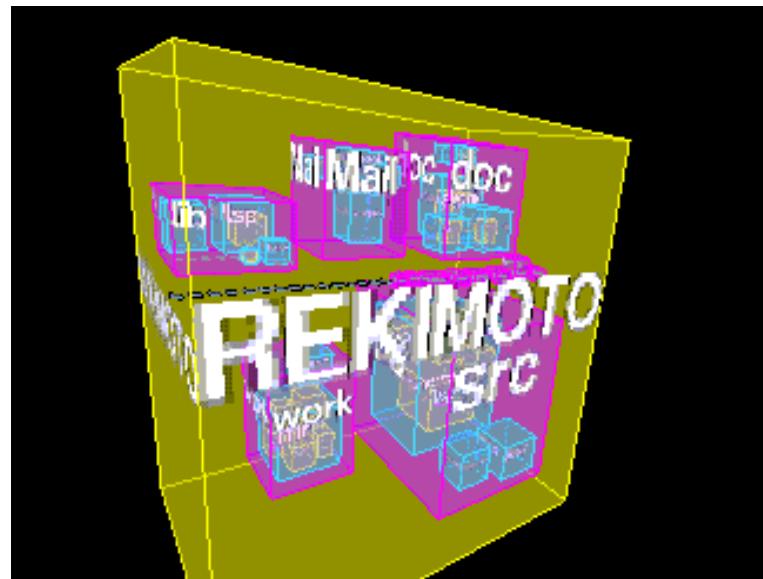
Tree-Map



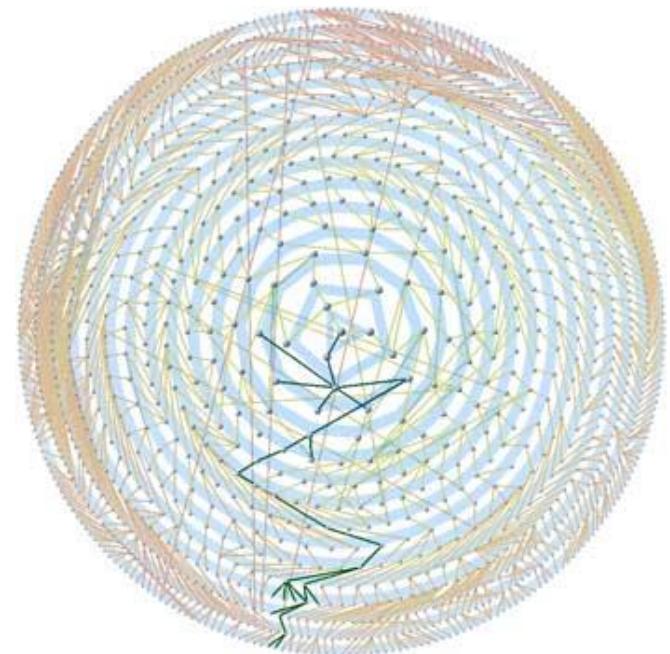
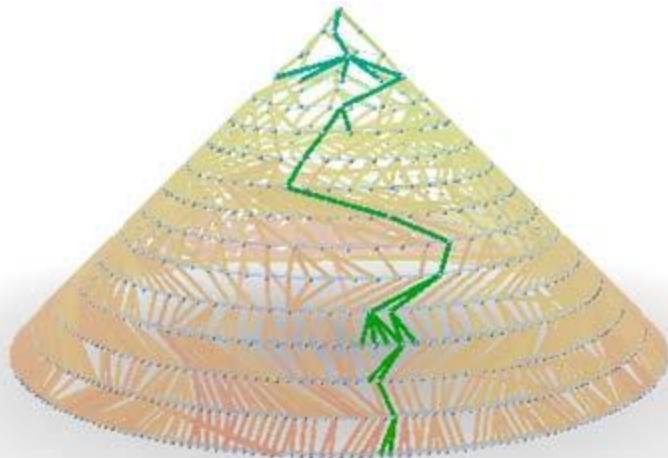
Tree-Map of a File System (Schneiderman)



InfoCube



Three-D Cone Trees



Ack.: <http://nadeausoftware.com/articles/visualization>

تصویرسازی داده ها و روابط پیچیده

■ اهمیت برچسب با سایز و رنگ فونت مشخص می شود.

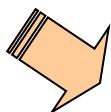
■ علاوه بر داده های متن، روش هایی برای تجسم روابط، مانند تصویرسازی شبکه های اجتماعی وجود دارد

- تصویر سازی داده های غیر عددی: متن و شبکه های اجتماعی
- ابر برچسب: تصویر سازی برچسب های تولید شده توسط کاربران



Newsmap: Google News Stories in 2005

فصل ۲: داده های خود را بیشتر بشناسید



- اشیاء داده ای و انواع صفات
- توصیفات آماری پایه از داده ها
- مصور سازی داده ها
- اندازه گیری میزان شباهت و عدم شباهت داده ها
- خلاصه

شباخت و عدم شباخت

■ شباخت

- اندازه عددی از میزان شباخت دو شی داده ای
- هرچه شباخت بیشتر باشد مقدار بالاتر است.
- معمولا در بازه $[0,1]$ بیان می شود.
- عدم شباخت (مثل فاصله)
 - اندازه عددی از میزان اختلاف دو شی داده ای
 - هرچه اشیا شبیه تر باشند مقدار کمتر می شود.
 - کمترین میزان عدم شباخت صفر است.
 - حد بالا متفاوت است.
- Proximity یا مجاورت به شباخت یا عدم شباخت اشاره می کند.

ماتریس داده و ماتریس عدم شباهت

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

- ماتریس داده
- n نمونه داده هر یک دارای p بعد
- دو وجهی

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

- ماتریس عدم تشابه
- ثبت فاصله n نمونه داده
- ماتریس مثلثی
- تک وجهی

اندازه شباهت برای داده های Nominal

- دو یا چند حالت را بخود می گیرد مثل قرمز، زرد، آبی، سبز، ...
(تعمیمی از صفات دودویی)

روش ۱: تطابق ساده

- m تعداد تطابق ها، p تعداد کل ویژگی ها

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

روش ۲: استفاده از تعداد زیادی از صفات دودویی

- ایجاد یک صفت دودویی به ازای هر حالت ممکن صفت اسمی

مثال: صفت test-1

A Sample Data Table Containing Attributes of Mixed Type

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

$$\begin{bmatrix} 0 \\ 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

اندازه شباهت برای صفات باینری

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

یک جدول حدوث برای موجودیت باینری ■

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

اندازه فاصله برای متغیرهای باینری متقارن: ■

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

اندازه فاصله برای متغیرهای باینری نامتقارن: ■

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

ضریب ژاکارد (اندازه مشابهت برای متغیرهای دودویی نامتقارن): ■

عدم شباهت بین صفات دودویی

مثال

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

جنسیت یک صفت دودویی متقارن است.

بقیه صفات نامتقارنند.

فرض کنید مقادیر Y و P یک و مقدار N صفر باشد.

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

استانداردسازی داده های عددی

$$z = \frac{x - \mu}{\sigma}$$

محاسبه **Z-score**

▪ X : داده خام، μ : میانگین σ : انحراف معیار

▪ فاصله بین داده خام و میانگین در واحد انحراف معیار

▪ اگر داده زیر حد میانگین باشد منفی و اگر بالای میانگین باشد مثبت است.

▪ یک راه دیگر: محاسبه میانگین انحراف مطلق

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

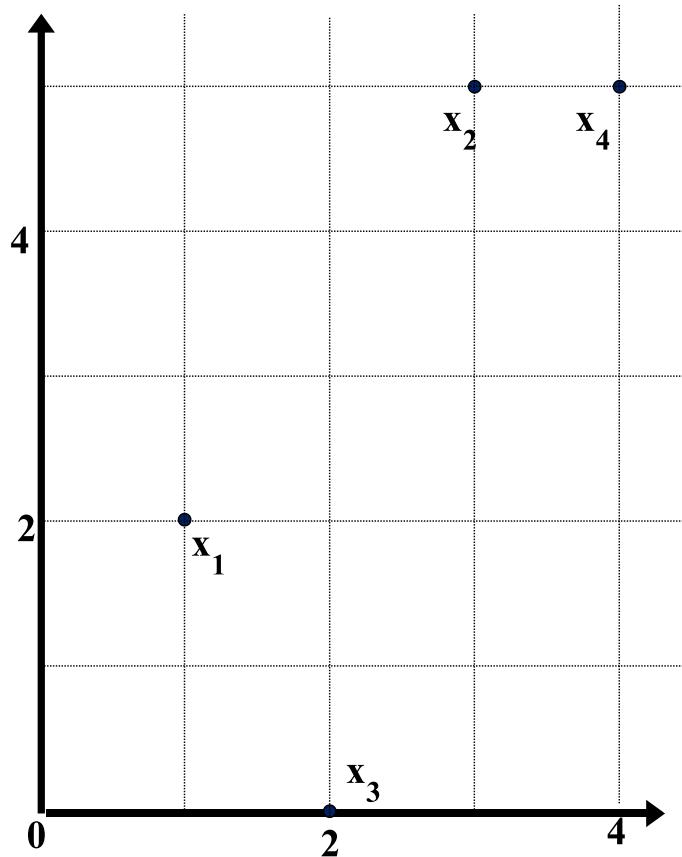
$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf}).$$

$$z_{if} = \frac{x_{if} - m_f}{s_f} \quad \text{اندازه استاندارد شده (z-score)}$$

▪ استفاده از میانگین انحراف مطلق بهتر از انحراف معیار است.

مثال:

ماتریس داده و ماتریس عدم شباهت



ماتریس داده

point	attribute1	attribute2
$x1$	1	2
$x2$	3	5
$x3$	2	0
$x4$	4	5

ماتریس عدم شباهت

(با فاصله اقلیدسی)

	$x1$	$x2$	$x3$	$x4$
$x1$	0			
$x2$	3.61	0		
$x3$	5.1	5.1	0	
$x4$	4.24	1	5.39	0

فاصله داده های عددی (فاصله مینکوفسکی)

■ فاصله مینکوفسکی: یک معیار فاصله رایج

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

و h نرم نامیده می شود و مقداری بزرگتر یا مساوی یک است.

■ $x_{ip}, \dots, x_{i2}, x_{i1}, x_{jp}, \dots, x_{j2}, x_{j1}$ رسمی م بعدی است.

■ خواص

- $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
- $d(i, j) = d(j, i)$ (Symmetry)
- $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)

■ فاصله ای که این خواص را داشته باشد متریک نامیده می شود.

حالت های خاص فاصله مینکوفسکی

اگر $h = 1$: فاصله منهتن ■
یا تعداد بیتهاي متفاوت بین دو رشته بیت مثل Hamming distance

$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$ ■
اگر $h = 2$: فاصله قلیدسی

$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$ ■
اگر $\infty \rightarrow$ فاصله سوپر نیم بیشترین اختلاف بین صفات

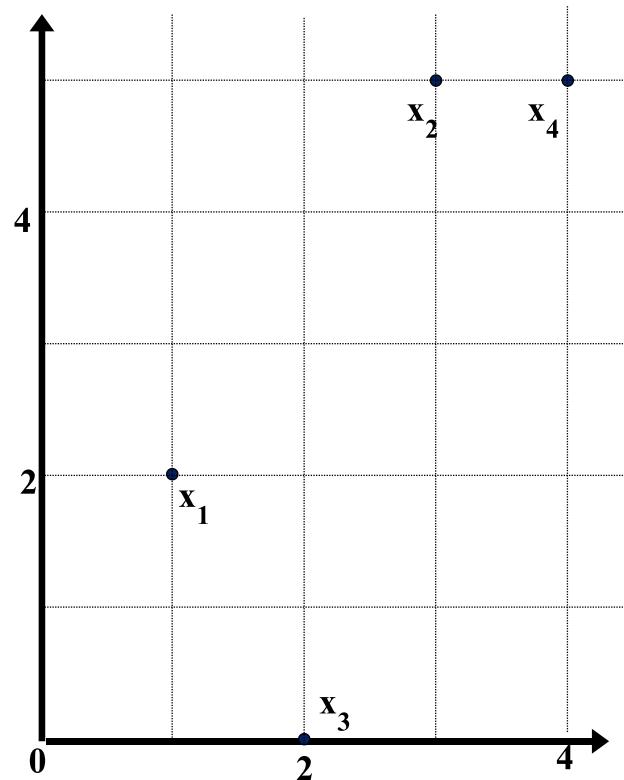
$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f^p |x_{if} - x_{jf}|$$

مثال: فاصله مینکوفسکی

Dissimilarity Matrices

Manhattan (L_1)

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



Euclidean (L_2)

L_2	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

متغیرهای ترتیبی

- یک متغیر ترتیبی می‌تواند گسته یا یک مقدار پیوسته طبقه‌بندی شده باشد.
- ترتیب در این نوع مهم است. مثل رتبه می‌تواند مثل فواصل عددی محاسبه شوند.
- هر x_{if} را با رتبه اش جایگزین کنید.
- با جایگزینی f امین متغیر از α امین شی با استفاده از فرمول زیر، بازه هر متغیر را به بازه $[0, 1]$ نگاشت کنید.

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- با استفاده از روش‌های بخش قبل فاصله دو شی را محاسبه کنید.

مثال

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

$$\begin{bmatrix} 0 & & & \\ d(2, 1) & 0 & & \\ d(3, 1) & d(3, 2) & 0 & \\ d(4, 1) & d(4, 2) & d(4, 3) & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

مخلوطی از انواع صفات

- یک پایگاه داده می تواند شامل ترکیبی از انواع صفات باشد.
- اسمی، دودویی متقارن و نامتقارن، عددی، ترتیبی
- ممکن است یک فرمول وزن دار برای محاسبه تاثیر متغیر ها

$$d(i,j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- در صورتی که یکی از دو مقدار ناموجود باشند یا یک صفت دودویی نامتقارن با دو مقدار صفر داشته باشیم مقدار وزن صفر و در غیر اینصورت یک است.

■ اگر صفت دودویی یا اسمی باشد:

$$d_{ij}^{(f)} = 0 \text{ if } x_{if} = x_{jf}, \text{ or } d_{ij}^{(f)} = 1 \text{ otherwise}$$

■ اگر صفت عددی باشد: از فاصله نرمال استفاده کنید.

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}},$$

■ اگر صفت ترتیبی باشد

■ رتبه r_{if} را محاسبه کنید.

■ با Z_{if} مانند یک فاصله عددی برخورد کنید.

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

مثال:

A Sample Data Table Containing Attributes of Mixed Type

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

$$\begin{bmatrix} 0 \\ 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1.0 & 0 \\ 1.0 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0.55 & 0.45 & 0.40 \\ 0.55 & 0 & 1.00 & 0.14 \\ 0.45 & 1.00 & 0 & 0.86 \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0.85 & 0.65 & 0.13 \\ 0.85 & 0 & 0.83 & 0.71 \\ 0.65 & 0.83 & 0 & 0.79 \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

شباخت کسینوسی

یک سند می تواند بوسیله هزاران صفت که هر یک تعداد تکرار یک کلمه را ثبت می کند نمایش داده شود.

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

بردارهای فراوانی های دیگر نظیر مشخصات ژنتیک ،...

کاربردها: بازیابی اطلاعات، طبقه بندی بیولوژیک، نگاشت ویژگی های ژنتیک،...

اگر d_1 و d_2 دو وکتور باشند، اندازه کسینوسی به شکل زیر تعریف می شود:

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

where \bullet indicates vector dot product, $\|d\|$: the length of vector d

مثال: شباهت کسینوسی

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

where \bullet indicates vector dot product, $\|d\|$: the length of vector d

شباهت بین دو سند ۱ و ۲ با مشخصات زیر را پیدا کنید:

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

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

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

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

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

$$\cos(d_1, d_2) = 0.94$$

فصل ۲: داده های خود را بیشتر بشناسید

-
- اشیاء داده ای و انواع صفات
 - توصیفات آماری پایه از داده ها
 - مصور سازی داده ها
 - اندازه گیری میزان شباهت و عدم شباهت داده ها
 - خلاصه



خلاصه

- انواع صفات داده ای: اسمی، دودویی، ترتیبی، مقیاس بازه ای یا مقیاس نسبتی
- انواع مختلف مجموعه داده نظیر داده های عددی، متن، گراف، وب، تصویر
- دستیابی به نگرش در مورد داده با:

 - توصیفات آماری پایه: گرایش مرکزی، پراکندگی، نمایش های گرافیکی
 - مصورسازی داده ها: ترسیم داده ها به شکل اشکال گرافیکی
 - اندازه گیری میزان شباهت در داده ها

- مراحل بالا مقدمات پیش پردازش داده ها هستند.
- روش های زیادی در این زمینه توسعه یافته اند اما هنوز یک موضوع فعال برای پژوهش تلقی می شود.

مراجع

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu , et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009



داده کاوی

مفاهیم و تکنیک ها

— فصل ۳ —

فصل ۳: پیش پردازش داده ها



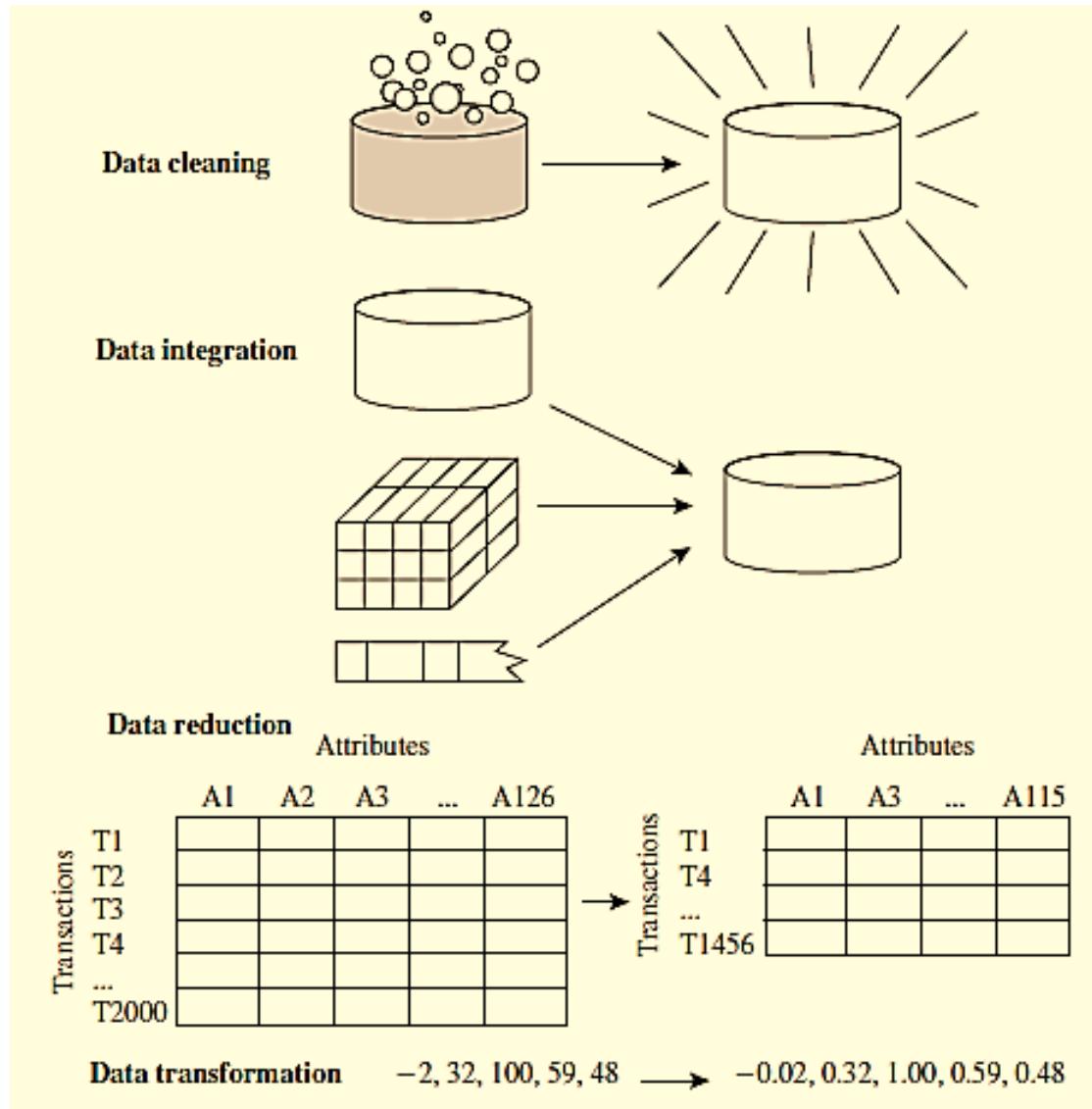
- پیش پردازش داده ها : مرور کلی
- کیفیت داده
- وظایف عمده در پیش پردازش داده
- پاکسازی داده
- تجمعیع داده
- کاهش داده
- تغییر شکل و گسته سازی داده
- خلاصه

کیفیت داده: چرا پیش پردازش داده ها

- معیارهای کیفیت داده: یک نگاه چند بعدی
- دقیق: صحیح یا غلط، دقیق یا نه
- کامل بودن: داده های ثبت نشده، غیر قابل دسترس
- سازگاری: بعضی داده ها تغییر پیدا کردند و بعضی نه، ...
- به هنگام بودن، آیا داده ها به موقع بهنگام شده اند؟
- باورپذیری: اطمینان از درستی داده ها
- قابلیت تفسیر: آیا داده ها به سادگی قابل درک هستند؟

وظایف اصلی در پیش پردازش داده ها

- پاکسازی داده ها
- پر کردن مقادیر ناموجود، اصلاح داده های نویزی، تشخیص یا حذف داده های پرت، از بین بردن تناقضات
- تجمیع داده ها
- تجمیع پایگاه داده های متفرق و چندگانه، مکعب های داده و فایل ها
- کاهش داده ها
- کاهش ابعاد
- کاهش تعداد
- فشرده سازی داده
- تغییر شکل و گسته سازی داده
- نرمالسازی
- تولید سلسله مراتب مفاهیم



فصل ۳: پیش پردازش داده ها

- پیش پردازش داده ها : مرور کلی
- کیفیت داده
- وظایف عمده در پیش پردازش داده
- پاکسازی داده 
- تجمعی داده
- کاهش داده
- تغییر شکل و گسته سازی داده
- خلاصه

پاکسازی داده

- داده ها در دنیای واقعی کثیف هستند : داده های زیاد بالقوه نادرست مثلا در اثر ابزار معیوب و ناقص، خطای انسانی یا کامپیوتر، خطای انتقال
- ناکامل: مقدار نداشتن بعضی صفات، نبودن صفات مطلوب و مورد نیاز، یا تجمیعی بودن داده ها

■ مثلا $\text{Occupation} = " "$ (missing data)

■ نویزی: حاوی نویز، خطأ یا داده های پرت

■ مثلا $\text{Salary} = "-10"$ (an error)

■ ناسازگار: حاوی اختلاف در کدها یا نام ها

■ $\text{Age} = "42"$, $\text{Birthday} = "03/07/2010"$

■ رده بندی قبلی "A, B, C" رده بندی فعلی "1, 2, 3"

■ تفاوت بین رکوردهای تکراری

■ عمدی: مثلا برای پنهان کردن داده های از دست رفته

■ اول ژانویه بعنوان تاریخ تولد همه

داده های ناقص

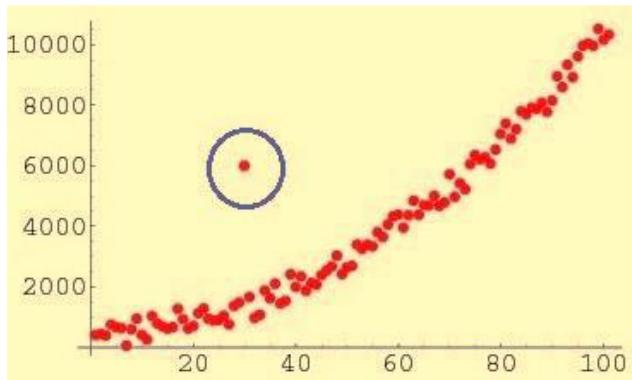
- داده همیشه در دسترس نیست
- مثلا خیلی از تاپل ها برای بعضی صفات مقداری ثبت نشده است. مثلا میزان درآمد مشتری در داده های فروش
- دلایل ایجاد داده های ناقص:
 - نقص تجهیزات
 - داده با بقیه داده های ثبت شده سازگاری نداشته و حذف شده است.
 - داده به دلیل عدم درک درست وارد نشده.
 - داده در زمان ورود اطلاعات مهم تلقی نشده و وارد نشده است.
 - تاریخچه یا تغییرات داده ثبت نشده است.
 - داده های مفقود نیاز به استنباط دارند.

چگونه مشکل داده های مفقود را حل کنیم؟

- نادیده گرفتن رکوردها: معمولاً این کار زمانی که برچسب کلاس مفقود باشد (در دسته بندی)- زمانی که در صد مقادیر از دست رفته برای هر ویژگی بطور قابل توجهی متفاوت باشد موثر نیست.
- پر کردن مقادیر بصورت دستی: خسته کننده و غیر عملی
- پر کردن بصورت اتوماتیک با
- یک مقدار ثابت عمومی مثل "unknown" یا یک کلاس جدید
- میانگین مقادیر صفت
- میانگین مقادیر صفت برای نمونه های کلاس مشترک: هوشمندانه تر
- محتمل ترین مقدار: مبتنی بر استنتاج مثل فرمول بیزین یا درخت تصمیم

داده نویزی

- نویز: خطای اتفاقی یا واریانسی در یک متغیر اندازه گیری شده
- دلایل وجود مقادیر غلط در صفات:
 - نقص در ابزار جمع آوری داده
 - مشکلات در ورود اطلاعات
 - مشکلات در انتقال اطلاعات
 - محدودیت تکنولوژی
 - تناقض در قراردادهای نامگذاری
- سایر مشکلات در داده که نیاز به پاکسازی داده دارند:



- رکوردهای تکراری
- داده ناقص
- داده متناقض

چگونه مشکل داده های نویزی را حل کنیم؟

- بسته بندی (Binning)
- ابتدا داده ها مرتب می شوند و به بین های با فرکانس یکسان تقسیم می شوند.
- سپس هموارسازی با میانگین بین یا میانه بین یا مرزهای بین یا ... انجام شود.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

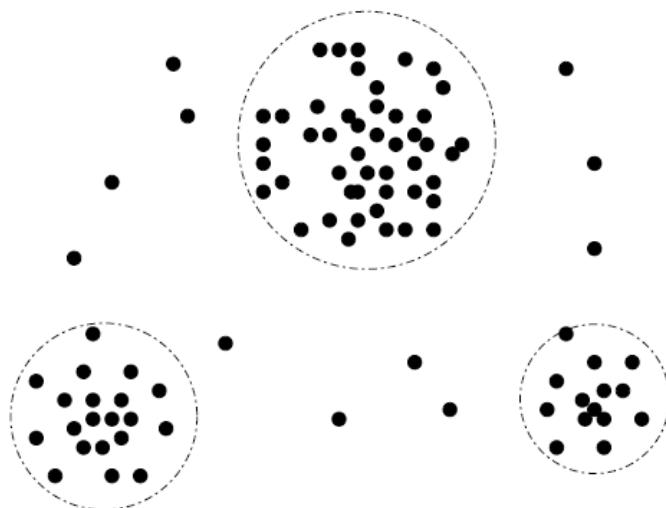
Bin 3: 25, 25, 34

■ رگرسیون

■ هموارسازی با قراردادن داده در توابع رگرسیون

■ خوشه بندی

■ تشخیص و حذف داده های پر



■ ترکیب بازرسی انسان و کامپیوتر

■ تشخیص مقادیر مشکوک توسط کامپیوتر و سپس بررسی

توضیح انسان ها (تعامل داده های دست)

پاکسازی داده بعنوان یک رویه

تشخیص انحراف داده

- استفاده از متادیتا (مثل دامنه، بازه، وابستگی ها، توزیع)
- چندکارگی فیلد
- بررسی قواعد یکتایی، پی در پی بودن و هیچقدار بودن
- استفاده از ابزارهای تجاری
- ابزارهای شستشو و سایش داده ها (Data scrubbing): استفاده از دانش دامنه ساده نظیر کد پستی یا تصحیح املایی برای تشخیص خطاها و تصحیح آن ها
- ابزارهای حسابرسی داده ها: آنالیز داده برای کشف قوانین و وابستگی ها و تشخیص داده هایی که آن ها را نقض می کنند. مثل همبستگی و خوشه بندی برای پیدا کردن داده های پرت

مهاجرت و یکپارچه سازی داده ها

- ابزارهای مهاجرت داده ها: انجام تبدیلات ساده مثل جایگزینی رشته gender به جای sex
- ابزارهای ETL (Extraction/Transformation>Loading): ایجاد این امکان برای کاربر که تبدیلات مورد نظر خود را از طریق یک واسط کاربر گرافیکی مشخص کند.
- تجمیع دو فرایند

(Potter's Wheels) بصورت تعاملی و تکراری (مثل

فصل ۳: پیش پردازش داده ها

- پیش پردازش داده ها : مرور کلی
- کیفیت داده
- وظایف عمده در پیش پردازش داده
- پاکسازی داده
- تجمعی داده
- کاهش داده
- تغییر شکل و گسته سازی داده
- خلاصه



تجمعیع داده

تجمعیع داده

- ترکیب داده ها از منابع متعدد در یک انباره منسجم
▪ تجمعیع شما مثلا $A.cust-id \equiv B.cust\#$
- تجمعیع متادیتا از منابع مختلف
▪ مشکل شناسایی یک موجودیت:
 - Bill Clinton = William
 - Clinton
 - تشخیص و برطرف نمودن مقادیر داده ناسازگار
 - برای یک موجودیت مشترک مقادیر صفات متفاوت در منابع مختلف وجود دارد.
 - دلایل احتمالی: نمایش های متفاوت، مقیاس های متفاوت نظیر metric در مقابل units

حل مشکل افزونگی در تجمعی داده ها

- تجمعی داده ها از منابع مختلف معمولاً منجر به ایجاد داده افزونه می شود.
- شناسایی موجودیت: یک موجودیت یا صفت ممکن است در پایگاه داده های متفاوت اسمی متفاوت داشته باشد.
- داده قابل اشتقاق: یک صفت ممکن است مشتق صفت دیگری در جدول دیگر باشد. مثل درآمد سالانه تشخیص باشند.
- صفات افزونه ممکن است توسط تحلیل همبستگی و تحلیل کوواریانس قابل تشخیص باشند.
- تجمعی دقیق داده از منابع مختلف به تقلیل یا حذف افزونگی ها و ناسازگاری ها و بهبود سرعت و کیفیت داده کاوی کمک کند.

تحلیل همبستگی (داده اسمی)

X² (chi-square) test ■

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

- هر چه مقدار χ^2 بزرگتر باشد احتمال ارتباط دو متغیر بیشتر است.
- سلول هایی که بیشترین اثر را در بزرگ کردن مقدار دارند آن هایی هستند که مقدار واقعی شان با مقدار مورد انتظار تفاوت بسیار دارد.
- همبستگی به معنای علیت نیست.
- تعداد بیمارستان ها و تعداد سرقت ماشین در یک شهر همبستگی دارند.
- هر دو این متغیرها وابسته به متغیر سومی هستند: جمعیت

مثال

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

■ مقادیر داخل پرانتز مقادیر مورد انتظار هستند که بر اساس توزیع دو متغیر محاسبه می شوند:

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

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

■ این نشان میدهد که علاقه به داستان های علمی و شطرنج بازی کردن به هم وابسته است.

تحلیل همبستگی (داده عددی)

ضریب همبستگی (Pearson's product moment coefficient) یا ■

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

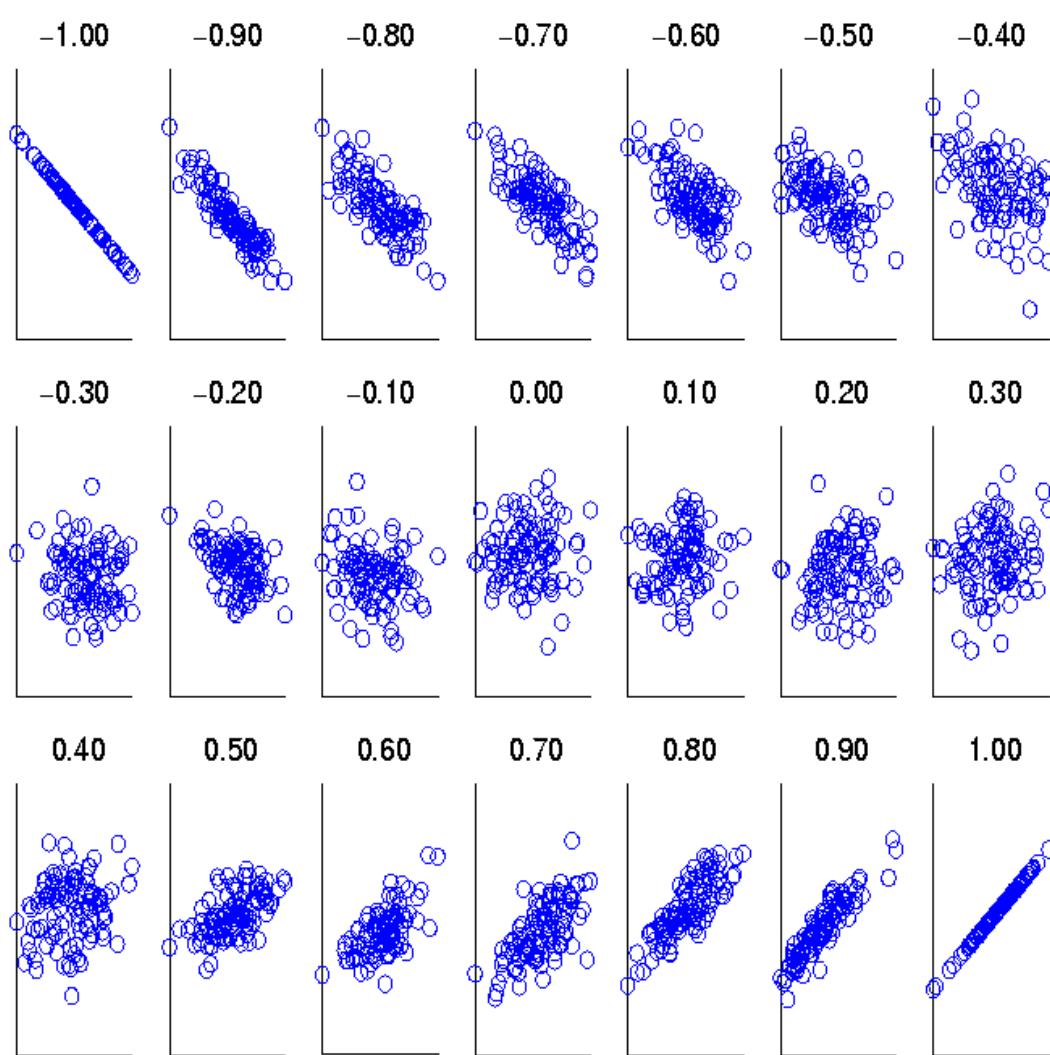
where n is the number of tuples, \bar{A} and \bar{B} are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\sum(a_i b_i)$ is the sum of the AB cross-product.

اگر $r_{A,B} > 0$ بین دو متغیر وابستگی مثبت وجود دارد. هر چه عدد بزرگتر باشد وابستگی بیشتر است. ■

اگر $r_{A,B} = 0$ دو متغیر به هم وابستگی ندارند. ■

اگر $r_{AB} < 0$ وابستگی منفی وجود دارد. ■

ارزیابی تصویری همبستگی



نمودارهای پراکندگی
شباهت بین -1 تا $+1$
را نشان می‌دهد.

Correlation (viewed as linear relationship)

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, A and B, and then take their dot product

$$a'_k = (a_k - \text{mean}(A)) / \text{std}(A)$$

$$b'_k = (b_k - \text{mean}(B)) / \text{std}(B)$$

$$\text{correlation}(A, B) = A' \bullet B'$$

کوواریانس(داده عددی)

کوواریانس شبیه 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:

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

where n is the number of tuples, \bar{A} and \bar{B} are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

اگر $Cov_{A,B} > 0$ دو متغیر میل به مقادیر بزرگتر از مقدار مورد انتظار دارند.

اگر $Cov_{A,B} < 0$ دو متغیر میل به مقادیر کوچکتر از مقدار مورد انتظار دارند.

اگر $Cov_{A,B} = 0$ متغیرها مستقل هستند اما عکس این صادق نیست.

مثال

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

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week:
(2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
 - $E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$
 - $E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$
 - $Cov(A,B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) /5 - 4 \times 9.6 = 4$
- Thus, A and B rise together since $Cov(A, B) > 0$.

فصل ۳: پیش پردازش داده ها

- پیش پردازش داده ها : مرور کلی
- کیفیت داده
- وظایف عمده در پیش پردازش داده
- پاکسازی داده
- تجمعیع داده
- کاهش داده
- تغییر شکل و گسته سازی داده
- خلاصه

راهبردهای کاهش داده ها

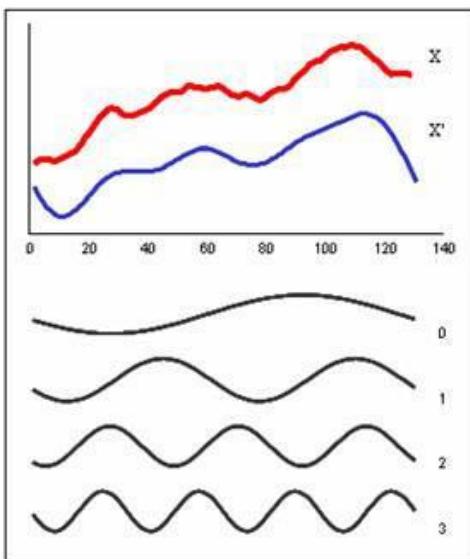
- کاهش داده ها: بدست آوردن یک نمای کاهش یافته از مجموعه داده که حجم بسیار کمتر اما نتایج تحلیلی یکسان با داده اصلی داشته باشد.
- چرا کاهش داده ها؟ یک پایگاه یا انبار داده ممکن است چندین تراپایت داده داشته باشد و تحلیل همه داده ها زمان زیادی لازم داشته باشد.
- راهبردهای کاهش داده ها:
 - کاهش ابعاد: مثلا حذف صفات غیر مهم
 - تبدیل موجک
 - تحلیل مولفه های اصلی
 - انتخاب زیرمجموعه ای از صفات خاصه
 - کاهش تعداد
 - رگرسیون و مدل های لگاریتمی
 - هیستوگرام، خوش بندی، نمونه گیری
 - تجمیع در مکعب داده ها
 - فشرده سازی داده ها

کاهش داده ها ۱: کاهش ابعاد

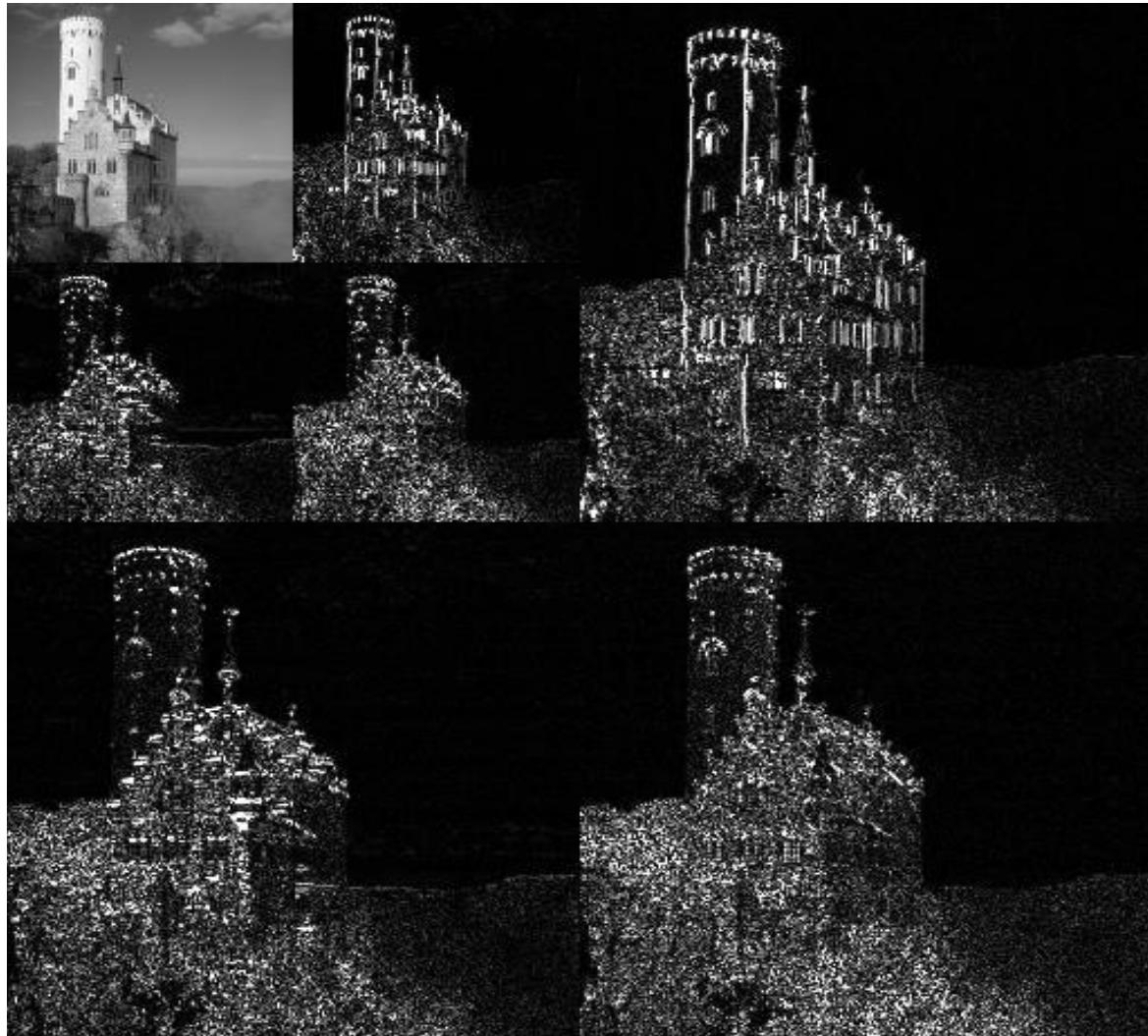
- زیاد بودن تعداد ابعاد باعث تنک شدن داده ها و مشکل شدن محاسبه فاصله ها و شباهت ها و ... در الگوریتم های داده کاوی می شود.
- کاهش ابعاد به منظور اجتناب از مشکلات بالا، کمک به حذف صفات غیرمرتب و کاهش نویز، کاهش فضا و زمان مورد نیاز برای داده کاوی و مصور سازی ساده تر انجام می شود.
- روش های مطرح شده برای کاهش ابعاد:
 - Wavelet transforms
 - Principal Component Analysis
 - Supervised and nonlinear techniques (e.g., feature selection)

تبدیل موجک

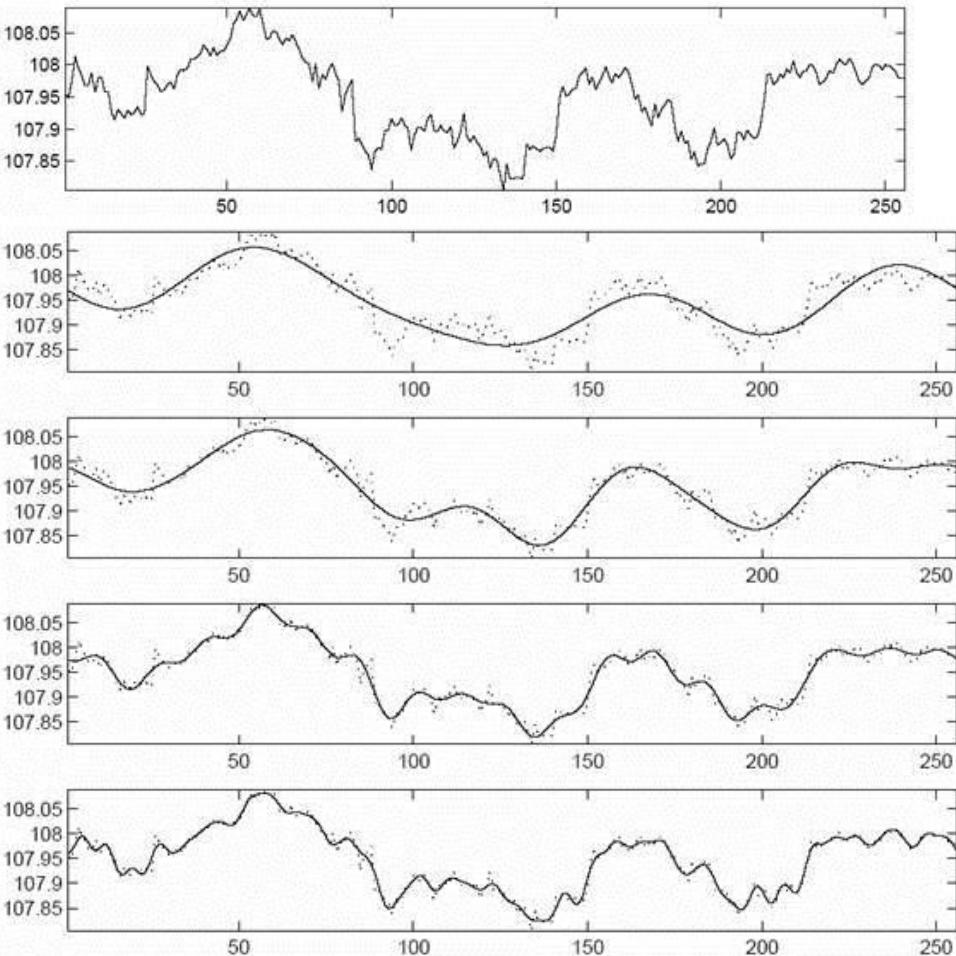
- مشابه تبدیل فوریه است.
- در بسیاری از کاربردها مرسوم است که از ترکیب توابع پایه ای برای تقریب یک تابع استفاده می شود.
- تبدیل فوریه یک تابع را بصورت توابع پایه ای سینوسی که هر کدام در مقادیری ضرب شده اند نشان می دهد.
- تبدیل فوریه یک تبدیل پرگشت پذیر است.
- این تبدیل دو حالت پیوسته و گسته دارد.
- در کامپیوتر و بخصوص در پردازش سیگنال معمولاً از تبدیل فوریه گسته استفاده می شود.
- خوشبختانه الگوریتم های سریعی تحت عنوان **FFT** برای این تبدیل بوجود آمده است.



تبديل موجك دو بعدی در تصاویر



مثال: قیمت روزانه سهام شرکت IBM در سال ۲۰۰۱



- این بردار از ۲۵۶ مولفه تشکیل شده است.
- اگر فقط ۱۰ ضریب پراهمیت‌تر را نگه داشته و بقیه را حذف کنیم نموداری شبیه به اولین نمودار بدست خواهیم آورد.

- تبدیل موجک ابتدا توسط شخصی بنام Alfred Harr بوجود آمد.
 داده ها بصورت یک توالی به طول توانی از ۲ می باشند.
 این اعداد بصورت جفت جفت با هم جمع شده و به مرحله بعد منتقل می شوند.
 اختلاف هر جفت هم محاسبه می شود.
 این مراحل با حاصل جمع های مرحله قبل دوباره تکرار می شود.
 این فرایند بصورت بازگشتی تکرار می شود تا در نهایت یک عدد که حاصل جمع کل اعداد است بدست آید و به همراه کلیه اختلاف جفت ها که در مراحل مختلف الگوریتم محاسبه شده بعنوان خروجی این تبدیل برگردانده می شود.

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	$a_1+a_2+a_3+a_4+a_5+a_6+a_7+a_8$				$(a_1+a_2+a_3+a_4)-(a_5+a_6+a_7+a_8)$			

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	

[..., +, -, +, -, +, -, +, -, +, ...,]

ضرایب کم اهمیت تر در سمت راست واقع شده اند و می توانند تا حد دلخواه حذف شوند.

تحلیل مولفه های اصلی

- پیدا کردن یک زیرمجموعه از ویژگی ها که بیشترین میزان تفاوت را در داده ها ایجاد می کند.

www.models.kvl.dk

	Workload	Distance to work	Salary
Smith	1.0	0.2	1.2
Johnson	2.0	0.0	0.3
Williams	-1.0	0.1	-1.0
Jones	-2.0	0.2	-0.1
Davis	0.0	-0.4	-0.4

انتخاب زیرمجموعه‌ای از صفات خاصه

- راه دیگری برای کاهش ابعاد
- ویژگی‌های افزونه
- صفاتی که همه یا بیشتر اطلاعاتشان در سایر صفات مستتر است
- مثل میزان خرید و مالیات پرداختی
- ویژگی‌های بی‌ربط
- صفاتی که فاقد اطلاعات مفید برای موضوع داده کاوی هستند.
- مثلاً صفت شماره دانشجویی در پیش‌بینی معدل بی تاثیر است.

جستجوی اکتشافی برای انتخاب صفات

- برای d صفت 2^d زیرمجموعه مختلف از صفات وجود دارد. بنابراین باید از روش‌های هیورستیک استفاده کرد.
- انواع روش‌های اکتشافی انتخاب صفات:
 - انتخاب رو به جلو (Forward Selection): شروع از مجموعه تهی و اضافه کردن صفات به ترتیب اهمیت
 - حذف رو به عقب (Backward Elimination): شروع از مجموعه کامل صفات و حذف بی اهمیت ترین صفت در هر مرحله ترکیب دو روش بالا
 - استفاده از درخت تصمیم: انتخاب صفاتی که در درخت تصمیم به کار رفته اند.

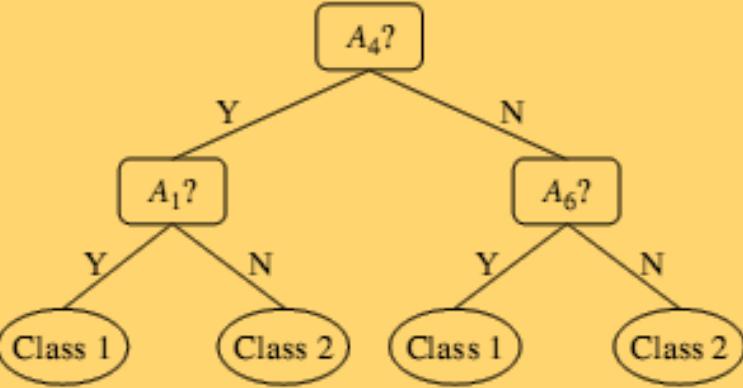
Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <pre> graph TD A4[A4?] -- Y --> A1[A1?] A4 -- N --> A6[A6?] A1 -- Y --> Class1_1((Class 1)) A1 -- N --> Class2_1((Class 2)) A6 -- Y --> Class1_2((Class 1)) A6 -- N --> Class2_2((Class 2)) </pre> <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>

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

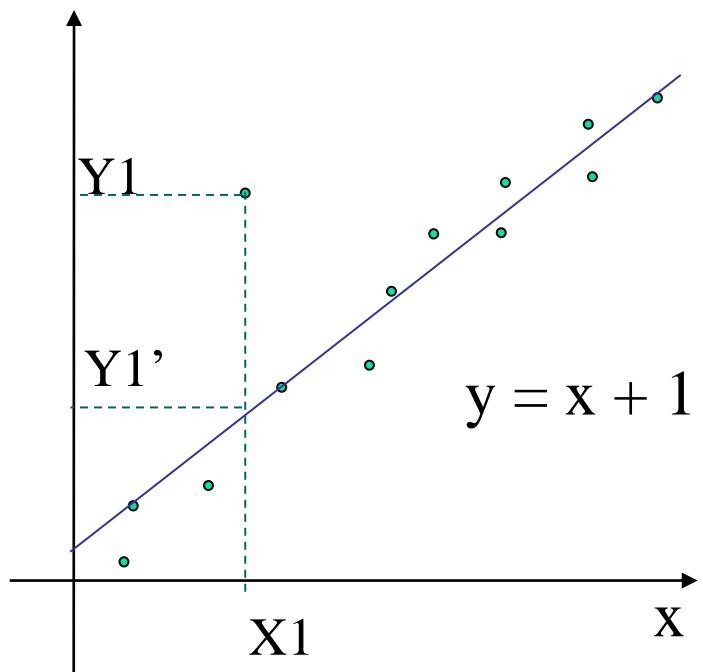
ایجاد ویژگی

- ایجاد ویژگی های جدید می تواند در افزایش دقت و فهم ما از ساختار داده های با ابعاد زیاد کمک کند.
- مثلا به جای طول و عرض می توان مساحت را ایجاد نمود.

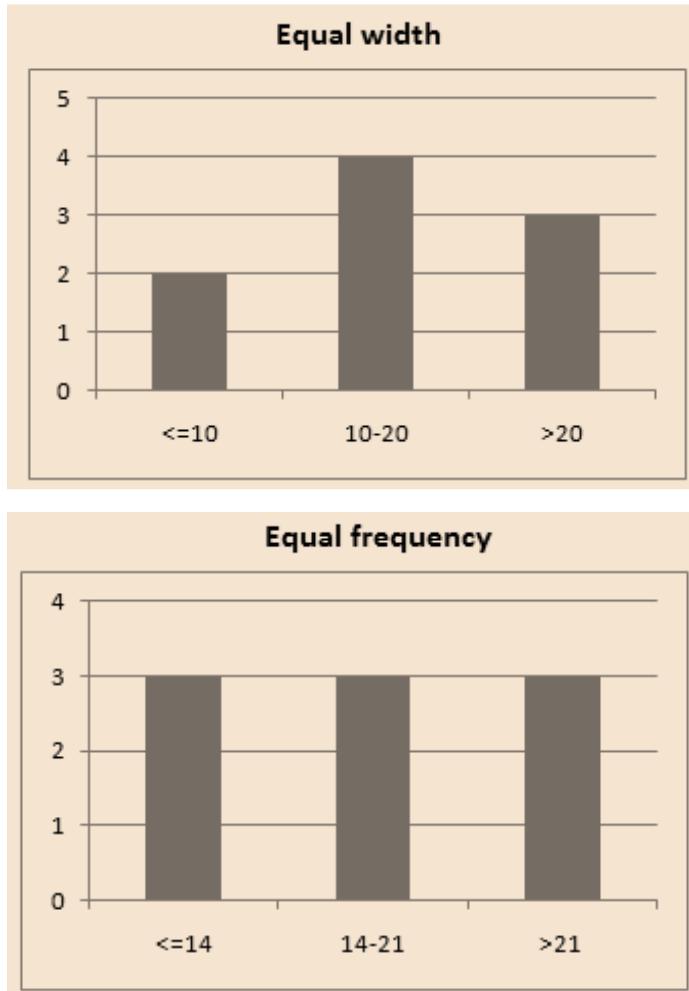
کاهش داده ۲ : کاهش تعداد

- کاهش حجم داده ها از طریق انتخاب شکل کوچکتر و جایگزین برای نمایش داده ها
- روش های پارامتریک:
 - در این روش ها فرض می شود داده ها با یک مدل انطباق دارند.
پارامترهای مدل محاسبه و ذخیره می شوند و اصل داده ها کنار گذاشته می شوند.
 - مثل رگرسیون
- روش های غیرپارامتریک:
 - برای داده ها مدلی فرض نمی شود.
 - مثل هیستوگرام، خوش بندی، نمونه گیری

رگرسیون



تحلیل هیستوگرام



تقسیم داده به تعدادی bin یا bucket و ذخیره نمودن میانگین اعداد

دو روش برای بخش بندی داده ها:

Equal-width: equal bucket range

Equal-frequency (or equal-depth)

- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28

- **Equal width**

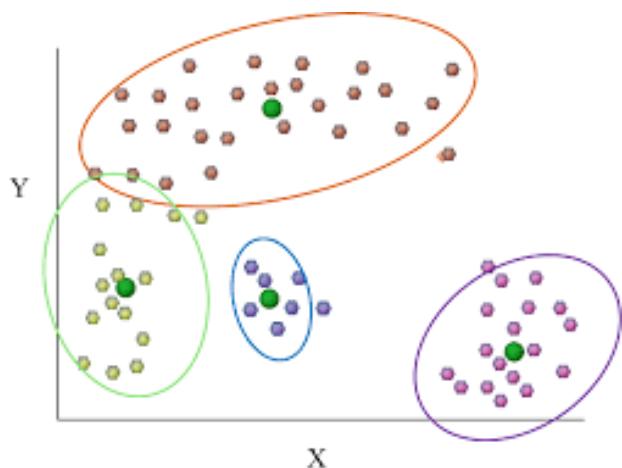
- Bin 1: 0, 4 [-,10)
- Bin 2: 12, 16, 16, 18 [10,20)
- Bin 3: 24, 26, 28 [20,+)

- **Equal frequency**

- Bin 1: 0, 4, 12 [-, 14)
- Bin 2: 16, 16, 18 [14, 21)
- Bin 3: 24, 26, 28 [21,+)

خوشه بندی

- تقسیم داده ها به خوشه ها بر اساس شباهت آن ها و تنها ذخیره و نگهداری نماینده خوشه (مثلا مرکز خوشه)
- اگر داده ها به خوبی خوشه بندی شوند بسیار مفید است.
- خوشه بندی می تواند بصورت سلسله مراتبی انجام و در یک ساختار درختی چند بعدی ذخیره شود.
- روش ها و الگوریتم های متفاوتی برای خوشه بندی وجود دارد که در فصل ۱۰ بررسی می شود.



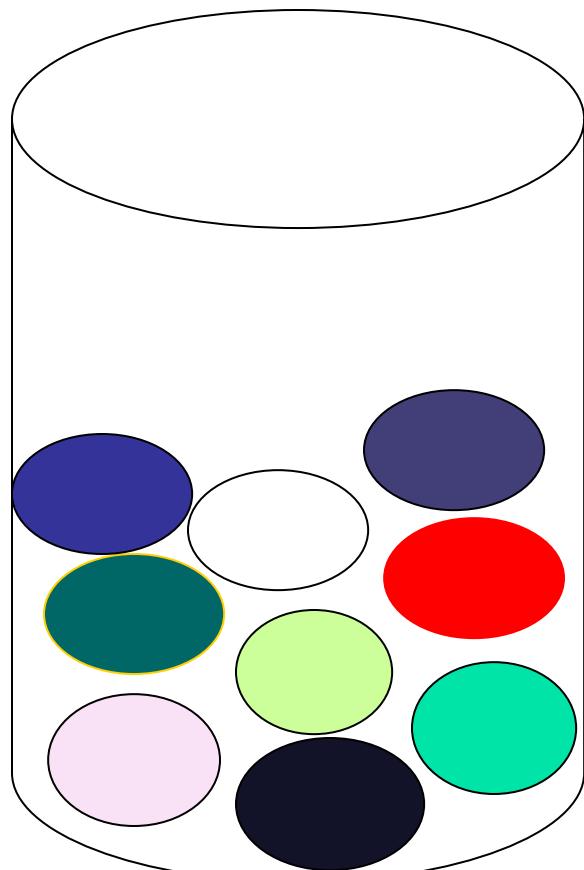
نمونه گیری

- نمونه برداری: در نظر گرفتن مجموعه S برای نمایش همه داده ها N
- اصل کلیدی: انتخاب زیرمجموعه ای از داده ها که نماینده کل مجموعه داده باشند.
- روش نمونه گیری تصادفی ساده ممکن است عملکرد ضعیفی در مجموعه داده های نامتوازن داشته باشد.
- توسعه روش های نمونه گیری تطبیقی مثل stratified sampling

انواع نمونه گیری

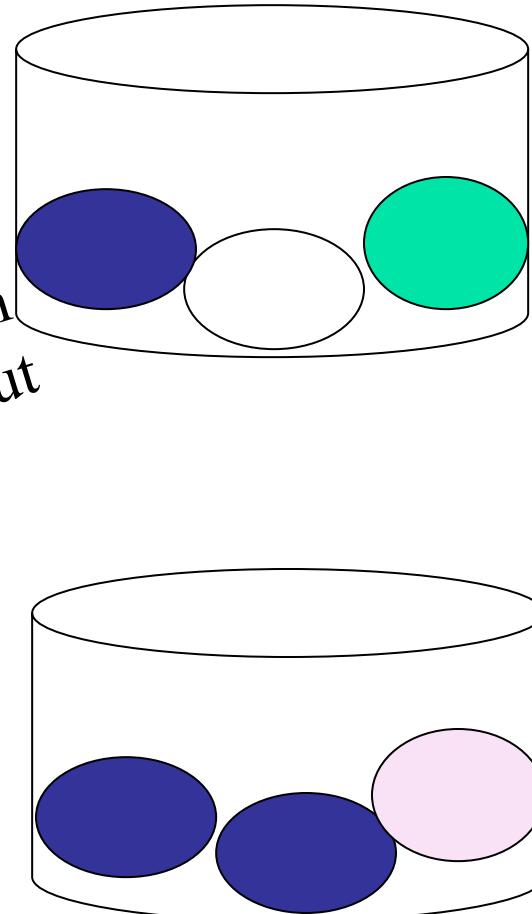
- نمونه گیری تصادفی ساده
 - انتخاب هر شی داده ای احتمال یکسان با بقیه دارد.
- نمونه گیری بدون جایگذاری
 - هنگامی که یک شی انتخاب شد آن را از جمعیت حذف می کنیم.
- نمونه گیری با جایگذاری
 - شی انتخاب شده از جمعیت حذف نمی شود.
- نمونه گیری طبقه بندی شده
 - پارتیشن بندی مجموعه داده ها و برداشتن نمونه از هر پارتیشن (متناوب با همان درصد از داده ها که در این پارتیشن قرار دارند).
 - مناسب برای مجموعه داده های نامتوازن

نمونه گری با یا بدون جایگذاری



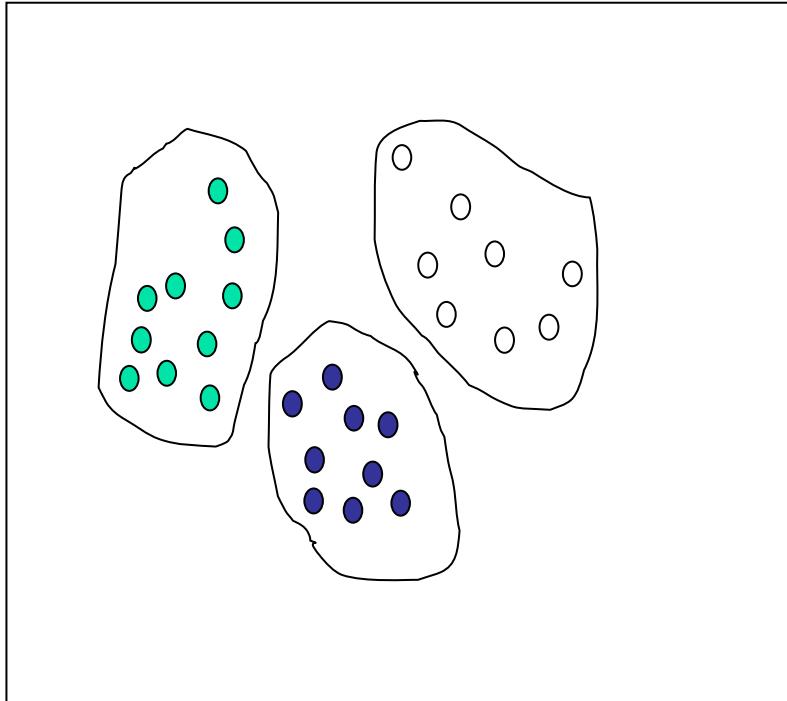
SRSWOR
(simple random
sample without
replacement)

SRSWR

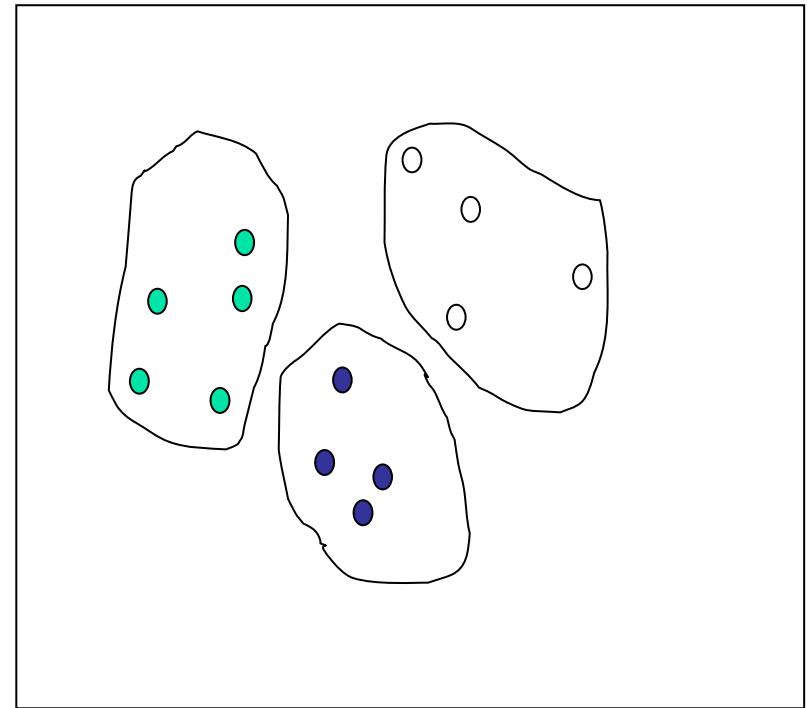


نمونه گیری طبقه ای

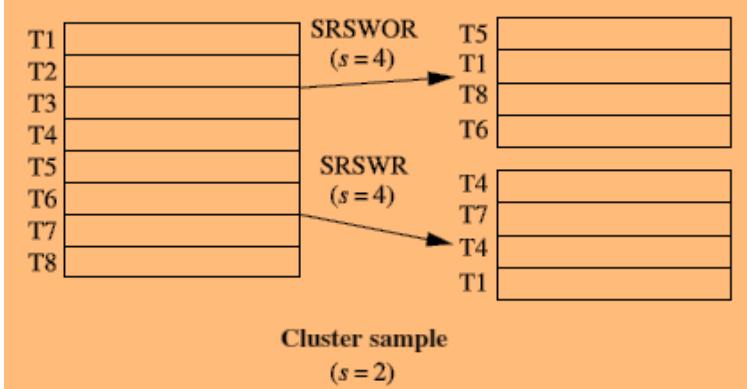
Raw Data



Cluster/Stratified Sample



مثال های نمونه گیری



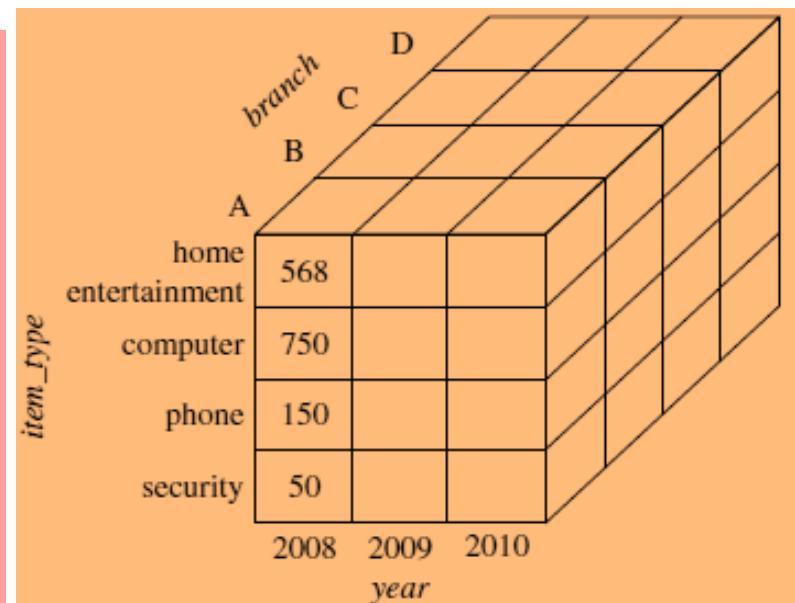
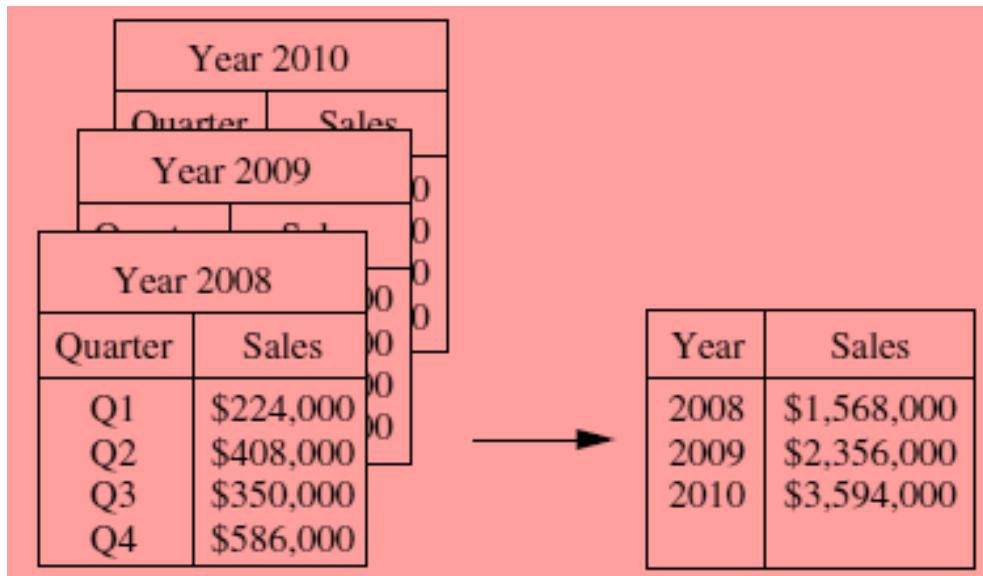
Startified sample
(according to *age*)

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

تجمیع مکعب داده

اطلاعات خلاصه شده و مثلا در یک مکعب تجمیع می شود. ■



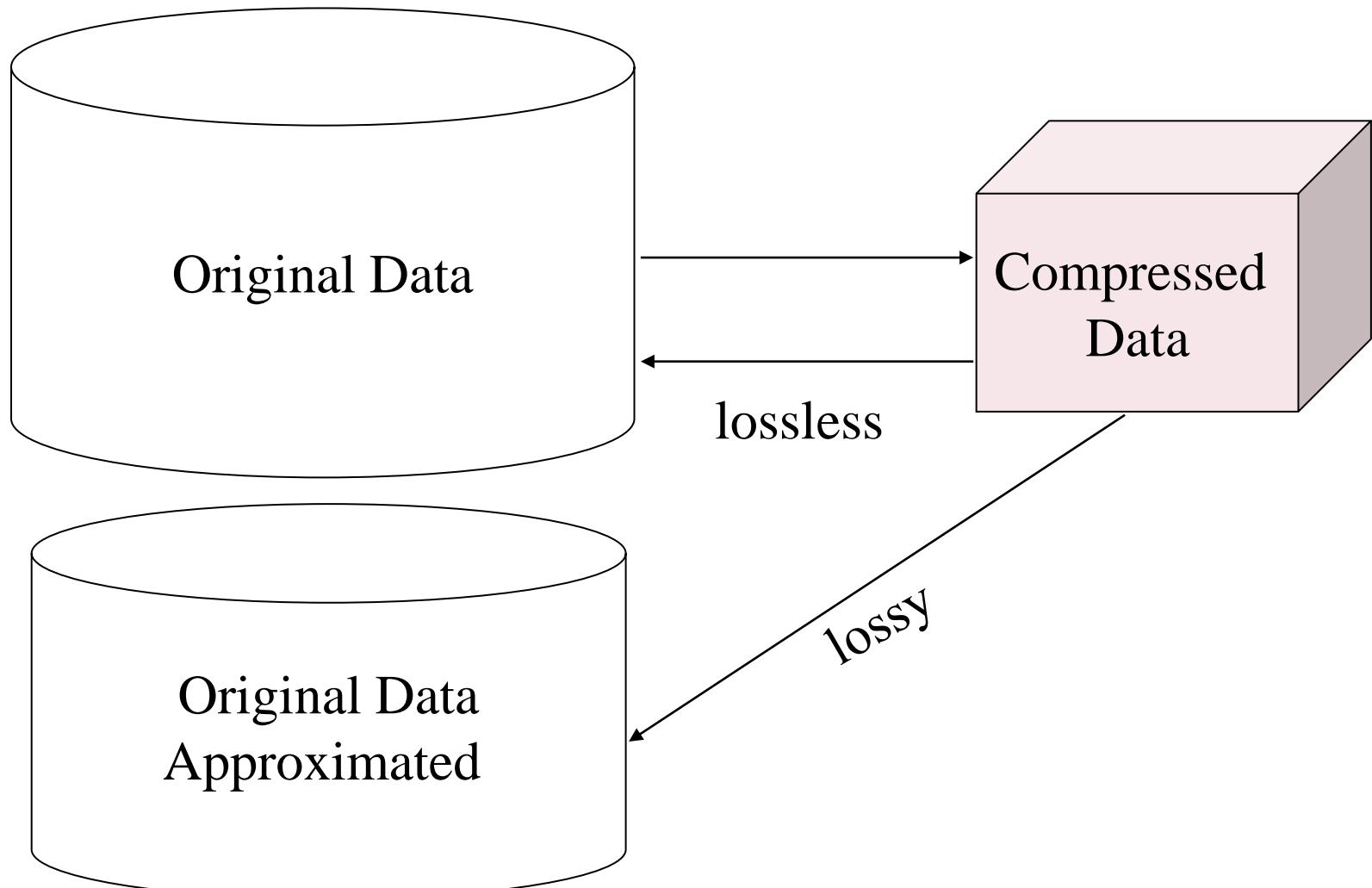
کاهش داده ها ۳: فشرده سازی داده

- فشرده سازی رشته ها

- فشرده سازی صوت و ویدیو

بعضی روش ها بدون از دست دادن بخشی از داده ها هستند و بعضی بخشی از داده را از دست می دهند.

فشرده سازی داده ها



فصل ۳: پیش پردازش داده ها

- پیش پردازش داده ها : مرور کلی
- کیفیت داده
- وظایف عمده در پیش پردازش داده
- پاکسازی داده
- تجمعیع داده
- کاهش داده
- تغییر شکل و گسته سازی داده
- خلاصه



تبديل داده ها

- یک تابع کل مجموعه مقادیر از یک ویژگی داده شده به یک مجموعه جدید از مقادیر نگاشت میدهد. هر مقدار قدیمی با یک مقدار جدید شناخته می شود.
- روش ها:
 - هموارسازی (smoothing): حذف نویز از داده ها
 - ساخت ویژگی
 - ویژگی های جدید از ویژگی های موجود ساخته می شود.
- تجمعی (Aggregation): خلاصه سازی، ساخت مکعب داده
- نرمال سازی(Normalization): مقیاسی برای انتقال داده به محدوده کوچکتر
 - نرمال سازی min-max
 - نرمال سازی z-score
 - نرمال سازی با مقیاس اعشاری
 - Discretization گسته سازی

نرمال سازی

[new_min_A , new_max_A] نرمالسازی **Min-max**: انتقال به بازه جدید ■

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

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

$$\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$$

نرمال سازی

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

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

Ex. Let $\mu = 54,000$, $\sigma = 16,000$. Then

$$\frac{73,600 - 54,000}{16,000} = 1.225$$

نرمال سازی

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

نرمال سازی با مقیاس اعشاری ■

Where j is the smallest integer such that $\text{Max}(|v'|) < 1$

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

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

گسته سازی

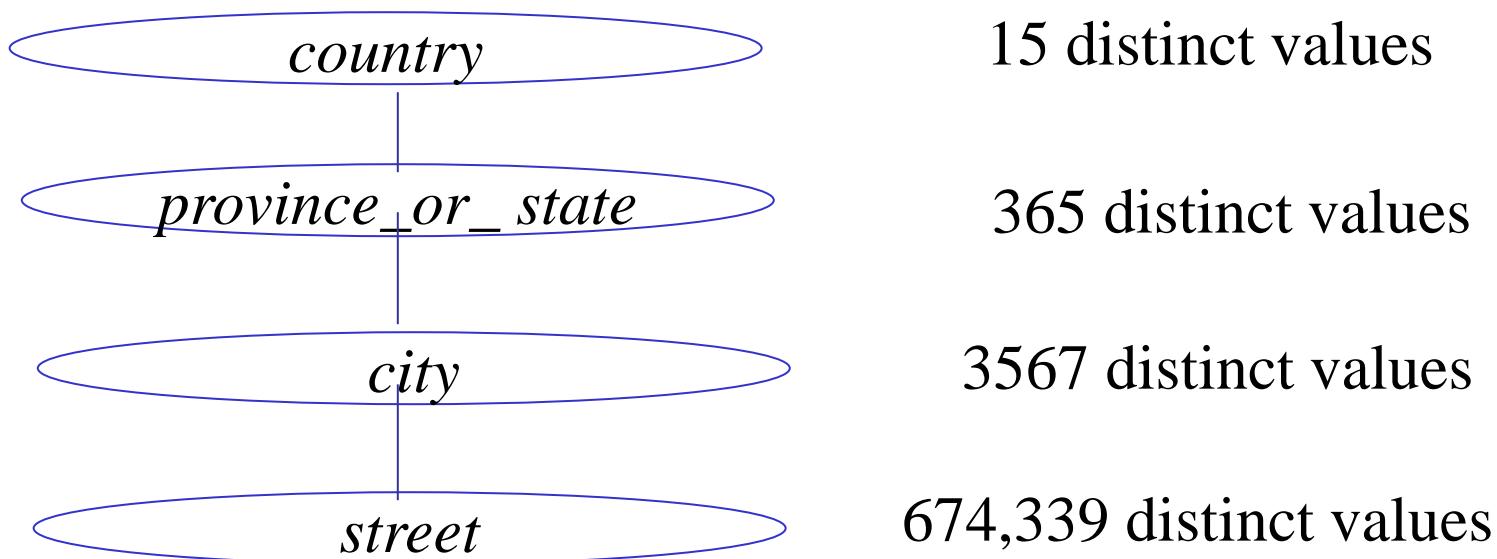
- تقسیم محدوده مقادیر پیوسته یک صفت به بازه های مختلف
- قرار دادن برچسب بازه بجای مقادیر داده ها
- کاهش اندازه داده ها
- روش های بانظارت در مقابل بدون نظارت
- تقسیم (از بالا به پایین) در مقابل ادغام(پایین به بالا)
- گسته سازی در یک صفت می تواند بصورت بازگشتی انجام شود.
- آماده سازی برای تجزیه و تحلیل بیشتر مثل طبقه بندی

روش های گسته سازی

- روش های معمول: (تمام روش ها را میتوان بصورت بازگشتی به کار گرفت.)
 - **Binning:** تقسیم بالا به پایین، بدون نظارت
 - **تحلیل هیستوگرام:** تقسیم بالا به پایین، بدون نظارت
 - **خوشه بندی:** بدون نظارت، بالا به پایین یا بر عکس
 - **درخت تصمیم:** با نظارت، بالا به پایین
 - **تحلیل همبستگی (χ^2 , e.g.):** بدون نظارت، پایین به بالا

تولید مفاهیم سلسله مراتبی

■ تولید مفاهیم سلسله مراتبی معمولاً توسط کارشناسان هر حوزه و گاهی بصورت خودکار انجام می شود و می تواند مفاهیم سطح پایین مثل سن را به مفاهیم سطح بالاتر مثل رده سنی(جوانان، بزرگسالان، سالمندان،...) ارتقاء دهد.



فصل ۳: پیش پردازش داده ها

- پیش پردازش داده ها : مرور کلی
- کیفیت داده
- وظایف عمده در پیش پردازش داده
- پاکسازی داده
- تجمعیع داده
- کاهش داده
- تغییر شکل و گسته سازی داده
- خلاصه



خلاصه

- کیفیت داده: دقت، کامل بودن، ثبات، به موقع بودن، باورپذیری، تفسیرپذیری
- پاکسازی داده: مثل حل مشکل داده های مفقود یا نویزی یا داده های پرت
- تجمعی داده از منابع مختلف:
 - مسئله شناخت یک شی
 - از بین بردن افزونگی ها
 - تشخیص تناقض ها در داده
- کاهش داده ها
 - کاهش ابعاد
 - کاهش تعداد
 - فشرده سازی داده ها
- تبدیل و گسته سازی داده ها
 - نرمال سازی
 - تولید سلسله مراتب مفاهیم

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- A. Bruce, D. Donoho, and H.-Y. Gao. Wavelet analysis. *IEEE Spectrum*, Oct 1996
- T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley, 2003
- J. Devore and R. Peck. *Statistics: The Exploration and Analysis of Data*. Duxbury Press, 1997.
- H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model, and algorithms. *VLDB'01*
- M. Hua and J. Pei. Cleaning disguised missing data: A heuristic approach. *KDD'07*
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. *Bulletin of the Technical Committee on Data Engineering*, 20(4), Dec. 1997
- H. Liu and H. Motoda (eds.). *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Kluwer Academic, 1998
- J. E. Olson. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999
- V. Raman and J. Hellerstein. *Potters Wheel: An Interactive Framework for Data Cleaning and Transformation*, *VLDB'2001*
- T. Redman. *Data Quality: The Field Guide*. Digital Press (Elsevier), 2001
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995



داده کاوی

مفاهیم و تکنیک ها

— فصل ۴ —

فصل ۴: انبار داده و پردازش تحلیلی برخط (OLAP)



- انبار داده: مفاهیم پایه
- مدلسازی انبار داده: مکعب داده و OLAP
- طراحی و کاربرد انبار داده
- پیاده سازی انبار داده
- تعمیم داده ها با کمک استنتاج صفت گرا
- خلاصه

انبار داده چیست؟

- تعاریف مختلف و غیر دقیقی از انبار داده ارائه شده است.
- یک پایگاه داده پشتیبان تصمیم که مجزای از پایگاه داده عملیاتی سازمان نگهداری می شود.
- پشتیبانی کننده پردازش اطلاعات با ارائه یک پایگاه جامع از اطلاعات تلفیقی و تاریخی برای تجزیه و تحلیل.
- انبار داده یک مجموعه داده **موضوع محور**، **مجتمع**، **تاریخمند** و **پایا** برای پشتیبانی تصمیم سازی مدیران است. W. H. Inmon-
- انبارسازی داده:
- فرایند ساخت و استفاده از انبار داده

انبار داده – موضوع محور

- انبار داده حول موضوعات اصلی مورد نیاز مثل **مشتری، محصول، فروش سازماندهی** می شود.
- انبار داده بر مدلسازی و تحلیل داده برای تصمیم سازان متمرکز است نه بر عملیات روزانه یا پردازش تراکنش ها
- با حذف داده هایی که در فرایند حمایت از تصمیم گیری مفید نیست، یک منظر ساده و مختصر در مورد مسائل موضوع خاص ارائه می دهد.

انبار داده – تجمعیع شده

- انبار داده با تجمعیع منابع داده متفرق ساخته می شود.
- پایگاه داده های رابطه ای، فایل های تخت، رکوردهای تراکنشی برخط تکنیک های پاکسازی و تجمعیع داده بر آن اعمال شده است.
- اطمینان از تطابق در قراردادهای نامگذاری، ساختارهای کدگذاری، مقیاس های مربوط به صفات وغیره در میان منابع داده های مختلف داده ای که به انبار داده منتقل می شود تغییریافته است.

انبار داده – تاریخمند

- افق زمانی یک انبار داده به طور قابل ملاحظه‌ای طولانی‌تر از سیستم‌های عملیاتی است.
- پایگاه داده عملیاتی: اطلاعات ارزشمند فعلی
- داده انبار داده: تامین اطلاعات از یک منظر تاریخی (مثلاً ۵ تا ۱۰ سال)
- هر ساختار کلیدی در انبار داده
- بصورت تلویحی یا صریح شامل یک عنصر زمان است.
- اما کلید داده‌های عملیاتی ممکن است شامل عنصر زمان نباشند.

انبار داده - پایا

- یک مخزن فیزیکی **جداگانه** از داده های تبدیل شده از محیط عملیاتی
- بهنگام سازی داده عملیاتی در محیط انبار داده اتفاق نمی افتد.
- نیازی به مکانیزم های پردازش تراکنش ها، ترمیم و کنترل همروندی در انبار داده وجود ندارد.
- فقط دو عملیات **بارگذاری اولیه داده** و **دسترسی به داده ها** مورد نیاز است.

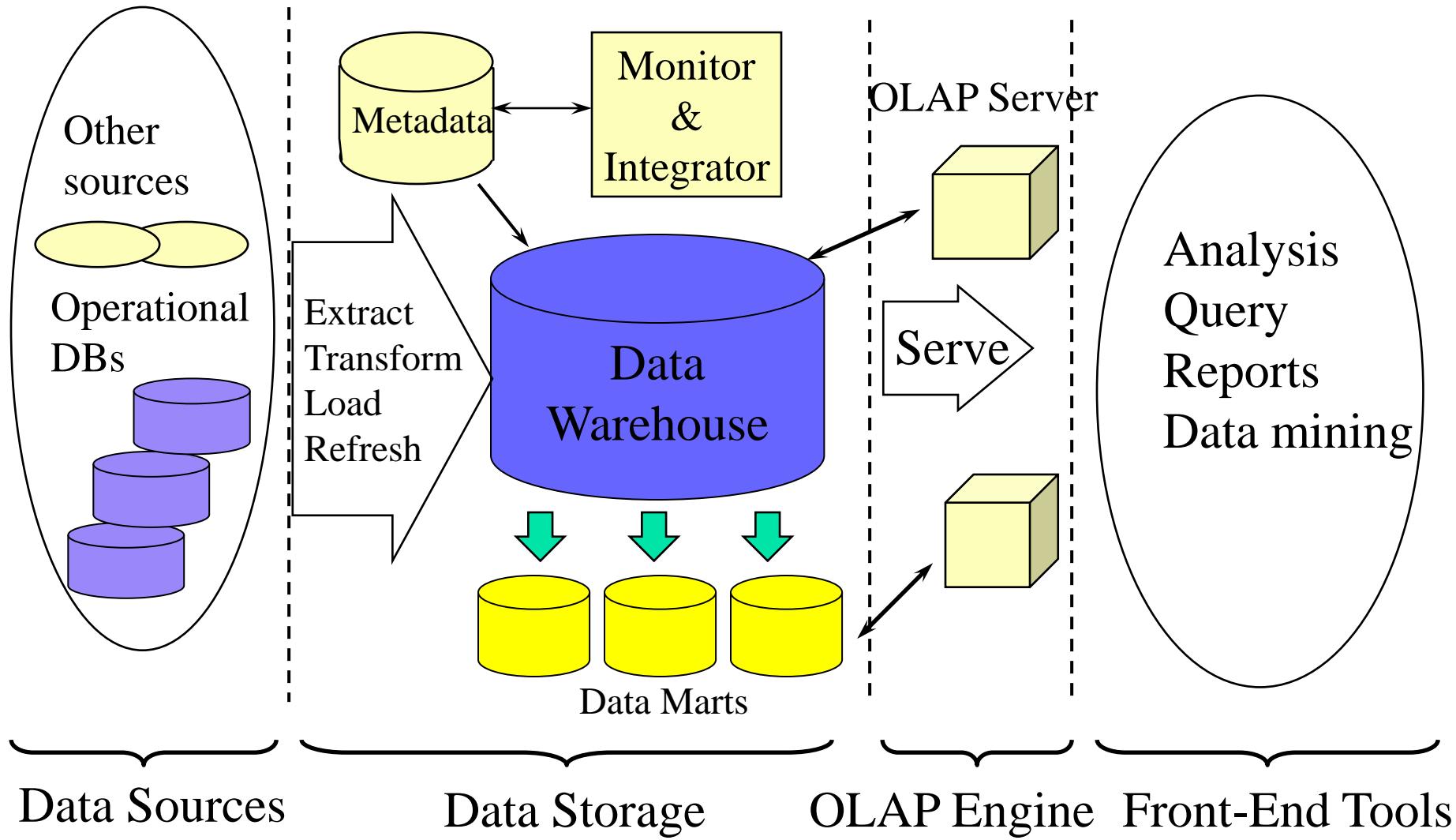
OLTP در مقابل OLAP

OLAP	OLTP	ویژگی
پردازش اطلاعاتی	پردازش عملیاتی	مشخصه
تحلیل	تراکنش	گرایش
کارکنان دانشی(مدیر، مدیر اجرایی، تحلیل گر)	کارمند دفتری، مدیر و متخصص پایگاه داده ها	کاربر
نیازمندی های اطلاعاتی بلندمدت جهت پشتیبانی از تصمیم گیری	عملیات روزمره	کارکرد
ستاره، برفگونه/ موضوع گرا	مبتنی بر ER، کاربردگرای	طراحی پایگاه داده ها
تاریخ دار، خلاصه شده، مجتمع، چند بعدی، تلفیقی	داده های جاری، بهنگام شده	داده ها
خلاصه شده، یکپارچه	ابتدايی، با جزئيات زياد	تلخیص
خلاصه شده، چندبعدی	داراي جزئيات، رابطه اي تخت	دید
پرس و جوی پیچیده	کوتاه، تراکنش ساده	واحد کاری
أغلب خواندن	خواندن/نوشتن	دستیابی
خروج اطلاعات	ورود داده ها	مرکز توجه
پیمایش زیاد داده ها	شاخص بندی یا درهم سازی بر	عملیات

چرا انبار داده جدا از پایگاه داده عملیاتی میسازیم؟

- کارایی بالا برای هر دو سیستم
- برای DBMS برای OLTP تنظیم می شود: روش های دستیابی، شاخص بندی، کنترل همزمانی، ترمیم
- انبار داده برای OLAP تنظیم می شود. پرس و جوهای پیچیده OLAP، نمای چند وجهی کارکردهای متفاوت و داده متفاوت:
- پشتیبانی از تصمیم نیاز به داده تاریخمند دارد که پایگاه داده عملیاتی عموماً فاقد آن است.
- پشتیبانی از تصمیم نیاز به تجمعی و خلاصه سازی داده از منابع مختلف را دارد.
- منابع متفاوت معمولاً نما، کد و فرمت های ناسازگاری از داده ها دارند.
- سیستم های بسیاری تحلیل های OLAP را مستقیماً روی پایگاه داده رابطه ای اجرا می کنند.

انبار داده: معماری چند سطحی



سه مدل انبار داده

انبار داده سازمان

- تمام اطلاعات در مورد موضوعاتی که کل سازمان را پوشش می دهد جمع آوری می کند.

بازار داده

- زیرمجموعه ای از کل داده ها که برای یک گروه خاص از کاربران دارای ارزش است.

انبار داده مجازی

- یک مجموعه از دید ها روی پایگاه داده های رابطه ای
- فقط بعضی از دیدهای تجمعی ممکن می توانند ساخته شوند.

استخراج، تبدیل و بارگذاری (ETL)

- استخراج داده
 - گرفتن داده از منابع متعدد و متفاوت خارجی
 - پاکسازی داده
- تشخیص خطاهای در داده و رفع آن تا حد امکان
- تبدیل داده
 - تبدیل داده از فرمت قبلی به فرمت انبار داده
 - بارگذاری
- مرتب کردن، خلاصه کردن، ثبت، نمایش محاسبات، بررسی یکپارچگی، ساختن شاخص‌ها و افزایش
- تازه کردن (**Refresh**)
 - انتشار تغییرات منابع داده به انبار داده

مخزن متادیتا

- متادیتا داده ایست که اشیاء انبار داده را تعریف می کند و شامل موارد زیر است:
- توصیف ساختار داده در انبار داده شامل شمای انبار، دید، ابعاد، سلسله مراتب ها، تعاریف داده های مشتق و محل و محتویات دیتامارت ها
- متا داده عملیاتی شامل سیر حرکتی داده ها(تاریخچه داده های جابجا شده و تبدیلاتی که بر روی آنها اعمال شده است)، وضعیت کنونی داده ها(فعال، بایگانی شده یا پاکسازی شده) و اطلاعات نظارتی و پایشی(آمارهای استفاده از انبار، گزارشات خطاهای و حسابرسی)
- الگوریتم های استفاده شده برای خلاصه سازی
- داده های مربوط به نگاشت محیط عملیاتی به انبار داده: منابع داده، روش های تبدیل ...
- داده های مربوط به کارایی سیستم: شاخص ها، زمانبندی عملیات تازه سازی و بهنگام آوری ...
- متاداده تجاری شامل تعاریف و اصطلاحات کسب و کار، مالکیت داده، سیاست های مالی ...

شما دیگری از معماری انبار داده

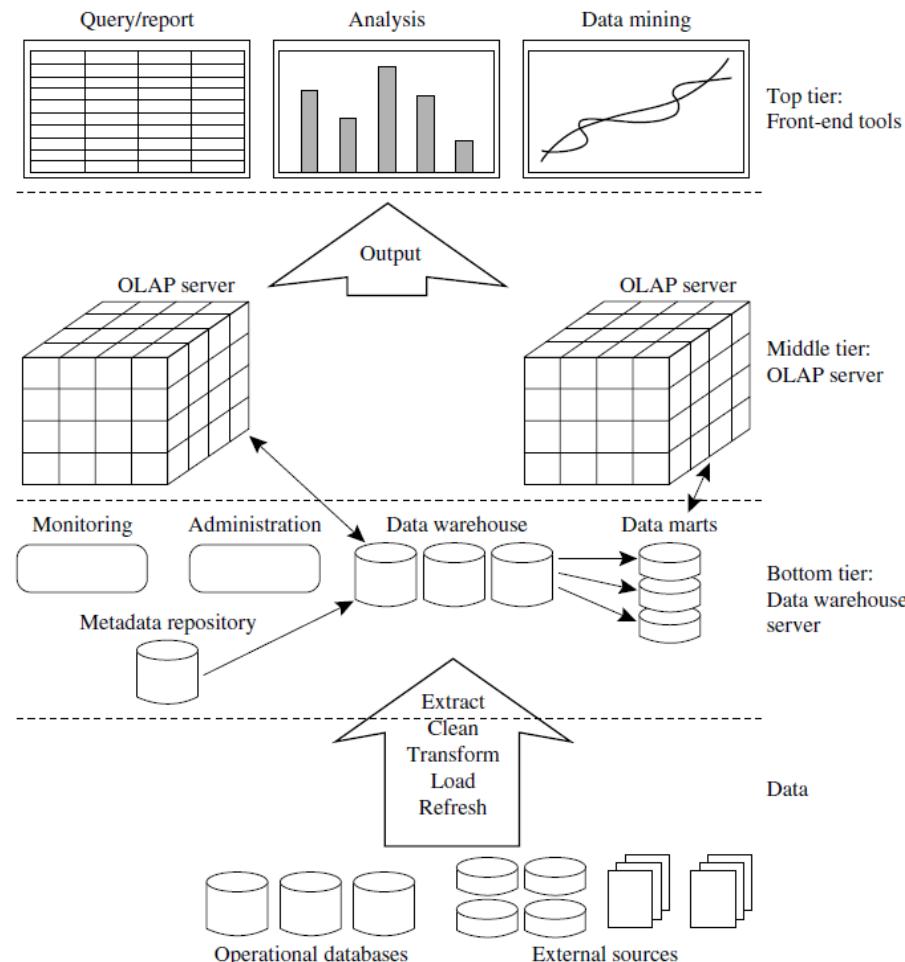


Figure 4.1 A three-tier data warehousing architecture.

فصل ۴: انبار داده و پردازش تحلیلی برخط (OLAP)



- انبار داده: مفاهیم پایه
- مدلسازی انبار داده: مکعب داده و OLAP
- طراحی و کاربرد انبار داده
- پیاده سازی انبار داده
- تعمیم داده ها با کمک استنتاج صفت گرا
- خلاصه

از جداول و صفحات گسترده به مکعب داده

- یک انبار داده بر اساس یک مدل داده ای چند بعدی که داده را در فرم یک مکعب داده نمایش می دهد شکل می گیرد.
- یک مکعب داده مثلا مکعب فروش امکان مدل کردن و نمایش داده به صورت چند بعدی را فراهم می سازد.
- **جدول ابعاد:** مثل item (item_name, brand, type) time(day, week, month, quarter, year) یا
- **جدول Fact:** شامل سنجه های عددی (مثل تعداد فروش) و کلیدهای جداول ابعاد مرتبط در ادبیات انبار داده، یک مکعب n بعدی پایه یک شبه مکعب پایه نامیده می شود. بالاترین شبه مکعب صفر بعدی که بالاترین سطح خلاصه سازی را دارد شبه مکعب نوک نامیده می شود. شبکه شبه مکعب ها مکعب داده را می سازد.

مثال: ساخت مکعب داده از داده های فروش لباس

فرض کنید جدول sales اطلاعات مربوط به لباس ها و فروش را نگهداری می کنند.

Clothes = {item-name, color, size}

Item-name = {skirt, dress, shirt, pant}

Color = {dark, pastel, white}

Size = {small, medium, large}

Sales = (item-name, color, size, number)

نمایی از داده جدول sales

name	color	size	quantity
skirt	dark	small	2
skirt	dark	medium	5
skirt	dark	large	1
skirt	pastel	small	11
skirt	pastel	medium	9
skirt	pastel	large	15
skirt	white	small	2
skirt	white	medium	5
skirt	white	large	3
dress	dark	small	2
dress	dark	medium	6
dress	dark	large	12
dress	pastel	small	4
dress	pastel	medium	3
dress	pastel	large	3
dress	white	small	2
dress	white	medium	3
dress	white	large	0
shirt	dark	small	2
shirt	dark	medium	6
shirt	dark	large	6
shirt	pastel	small	4
shirt	pastel	medium	1
shirt	pastel	large	2
shirt	white	small	17
shirt	white	medium	1
shirt	white	large	10
pants	dark	small	14
pants	dark	medium	6
pants	dark	large	0
pants	pastel	small	1
pants	pastel	medium	0
pants	pastel	large	1
pants	white	small	3
pants	white	medium	0
pants	white	large	2

نمونه یک cross-tab از داده های جدول

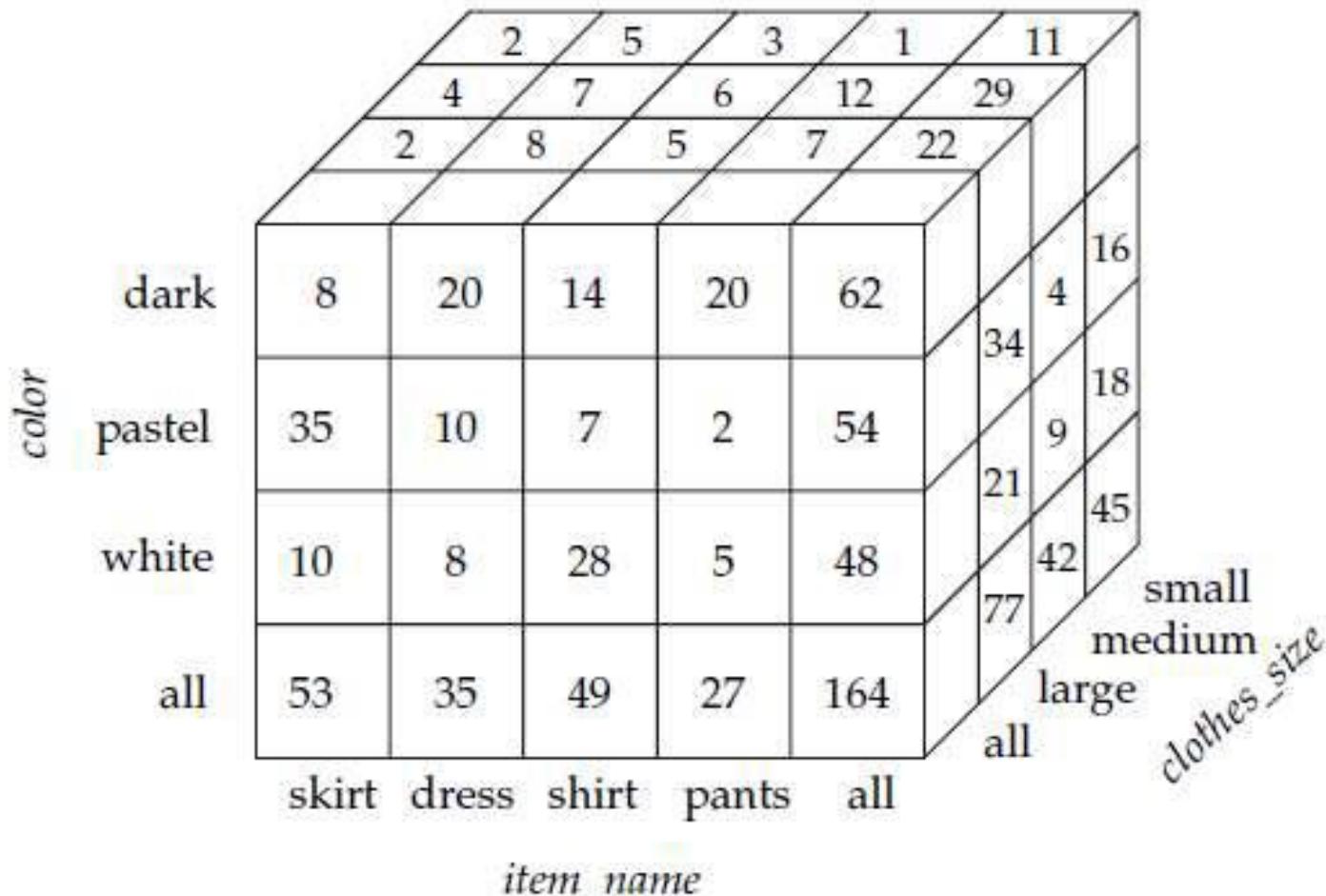
- در این نمونه ابعاد نام و رنگ بعنوان ویژگی های بعد (dimension attributes) و تعداد فروش بعنوان ویژگی معیار (measure attribute) در نظر گرفته شده است.

clothes_size **all**

<i>item_name</i>	<i>color</i>			
	dark	pastel	white	total
skirt	8	35	10	53
dress	20	10	5	35
shirt	14	7	28	49
pants	20	2	5	27
total	62	54	48	164

Cross tabulation of *sales* by *item_name* and *color*.

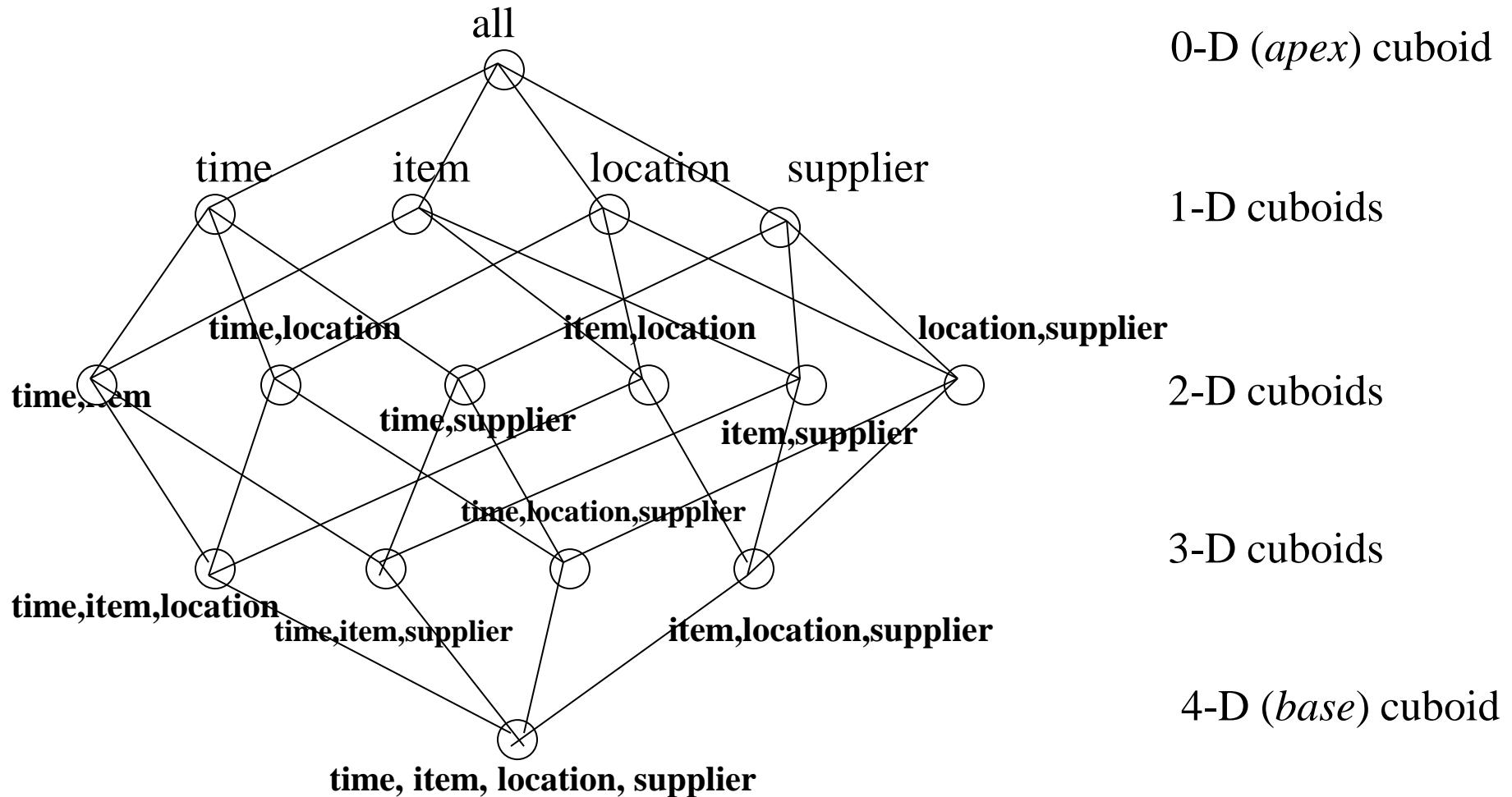
پک مکعب داده با سه بعد



Three-dimensional data cube.

- 
- هر سلول در مکعب داده یک تجمعی از بعضی ابعاد است.
 - برای n بعد 2^n تجمعی متفاوت قابل انجام است.

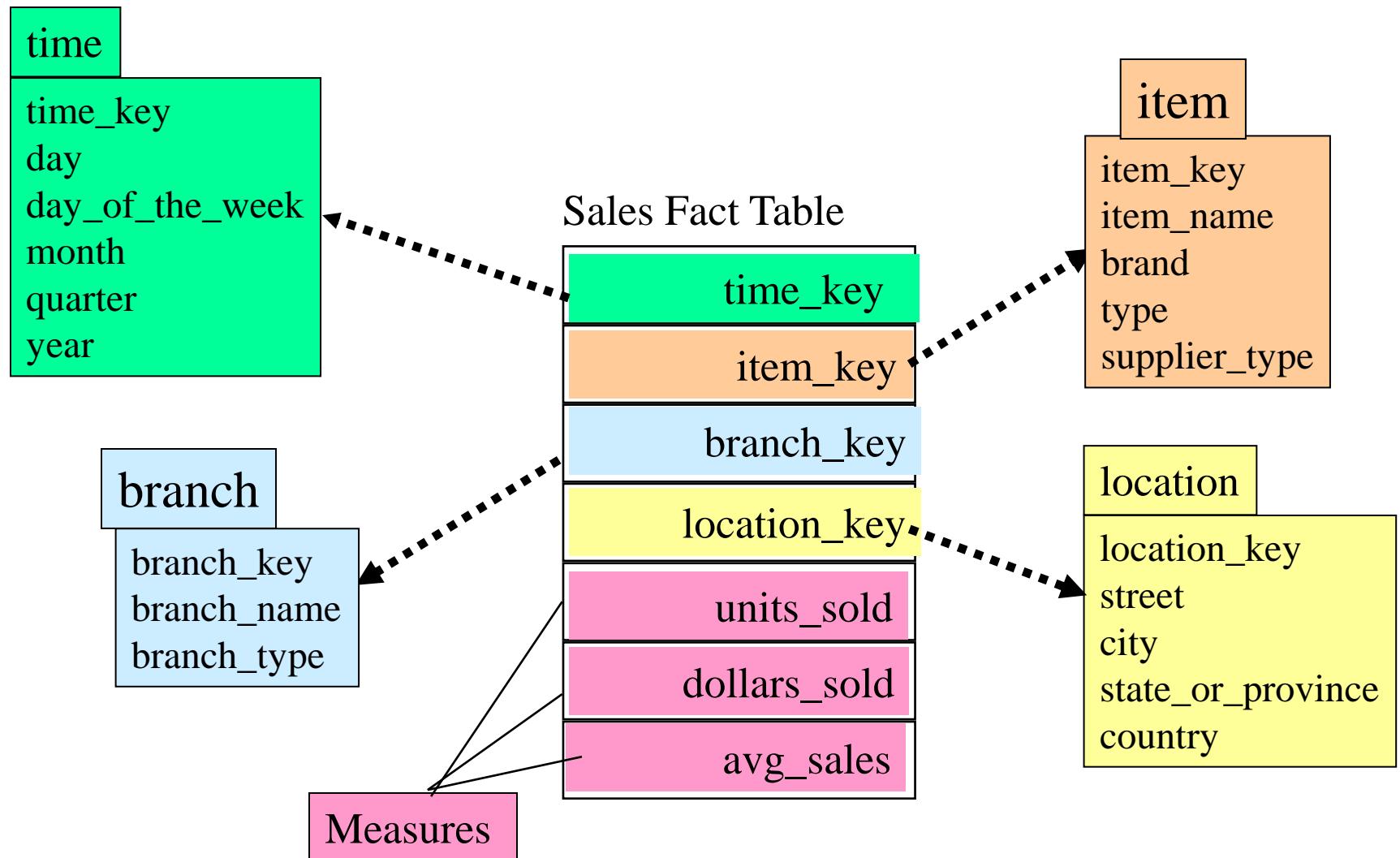
مکعب: شبکه ای از شبکه مکعب ها



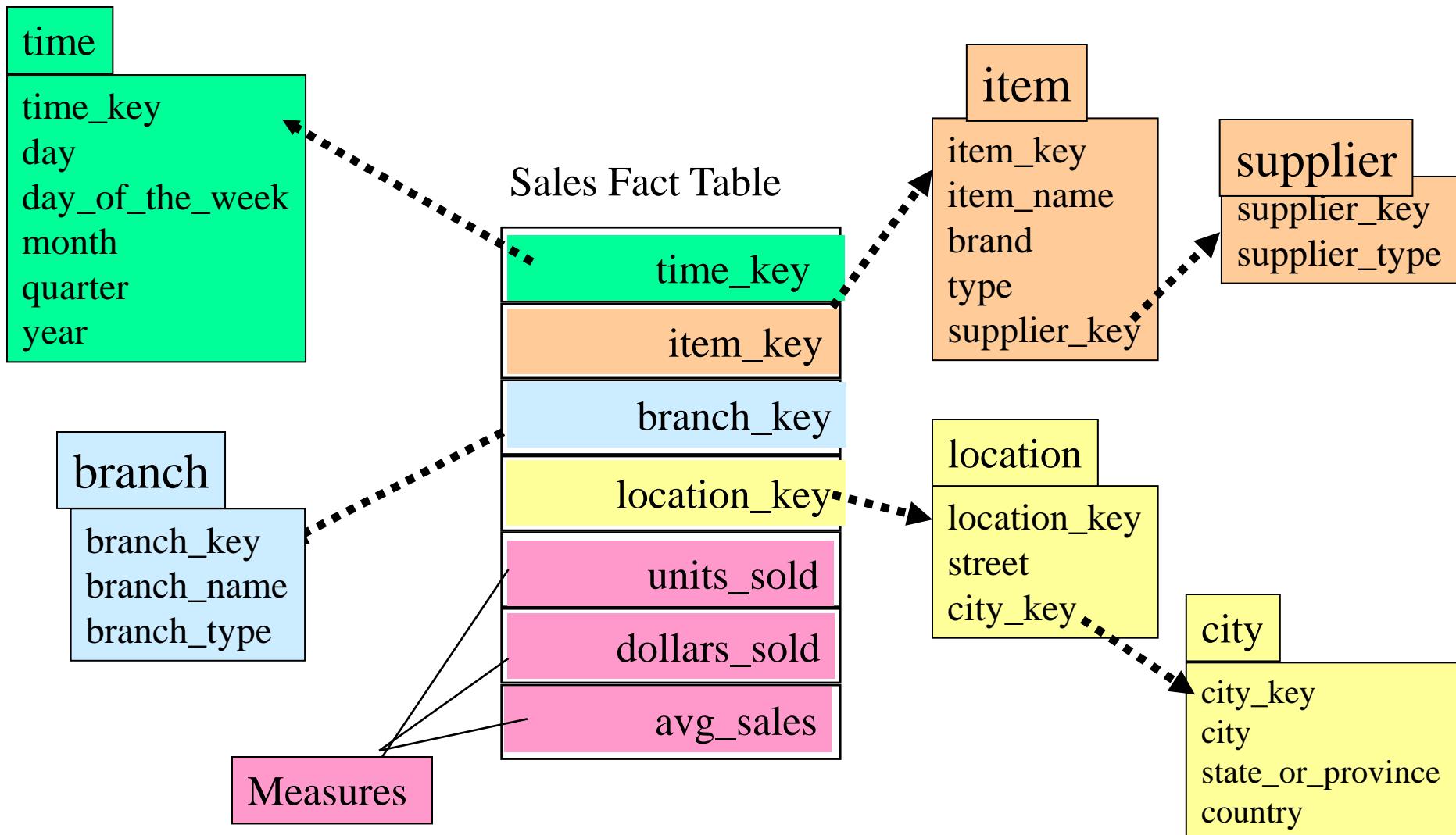
مدل‌سازی مفهومی یک انبار داده

- انواع مدل‌سازی در انبار‌های داده
- شمای ستاره‌ای: یک جدول fact در وسط که به مجموعه‌ای از جداول ابعاد مرتبط شده است.
- شمای برفگونه: اصلاحی در شمای ستاره‌ای که در آن جداول بعضی ابعاد نرمال شده‌اند و مجموعه‌ای از جداول کوچکتر را ساخته‌اند.
- صورت فلکی: چند جدول fact از بعضی جداول ابعاد بصورت مشترک استفاده می‌کنند و شبیه یک مجموعه ستاره بنظر می‌آیند که به آن شمای کهکشان هم می‌گویند.

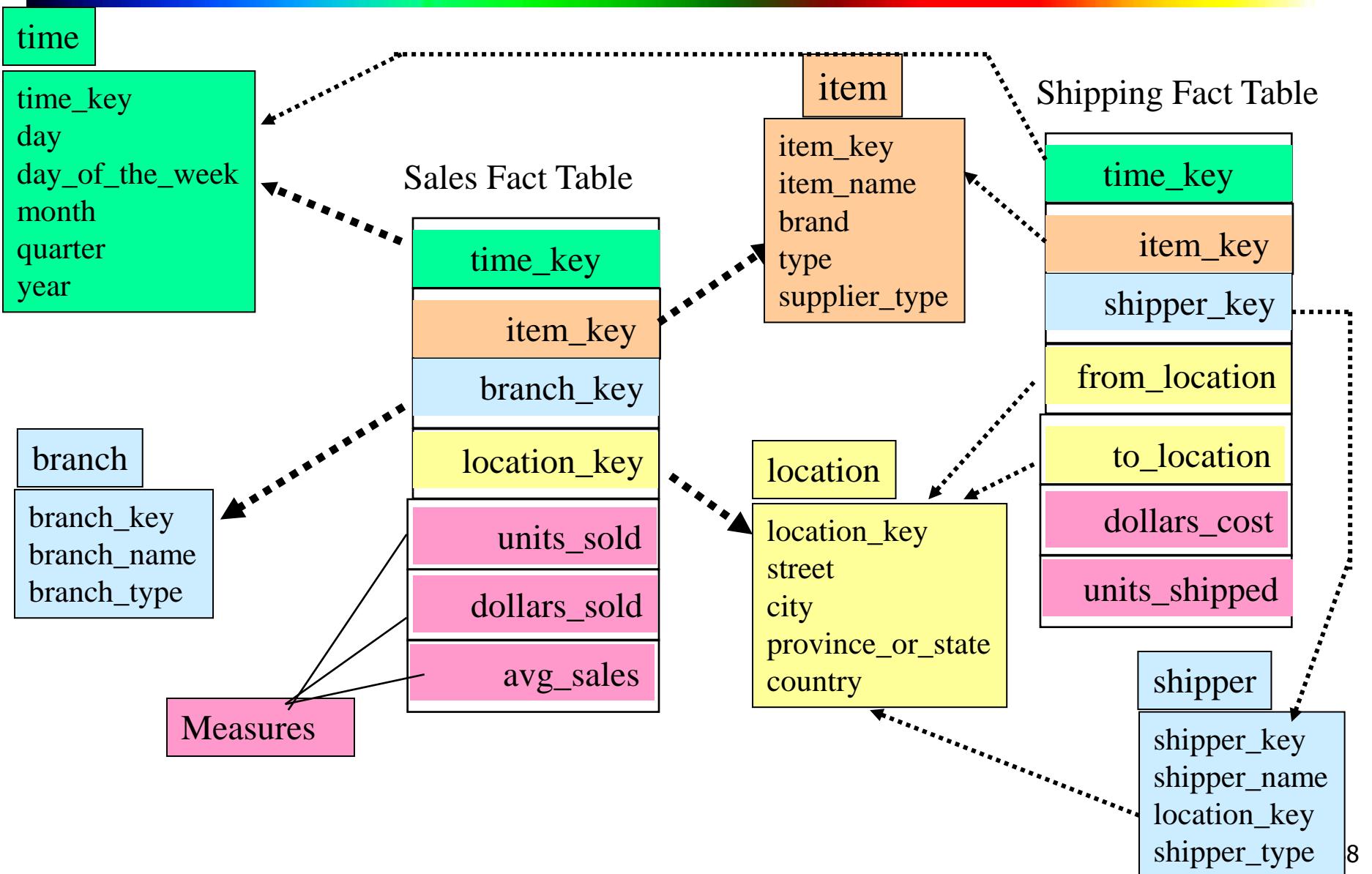
مثالی از شمای ستاره ای



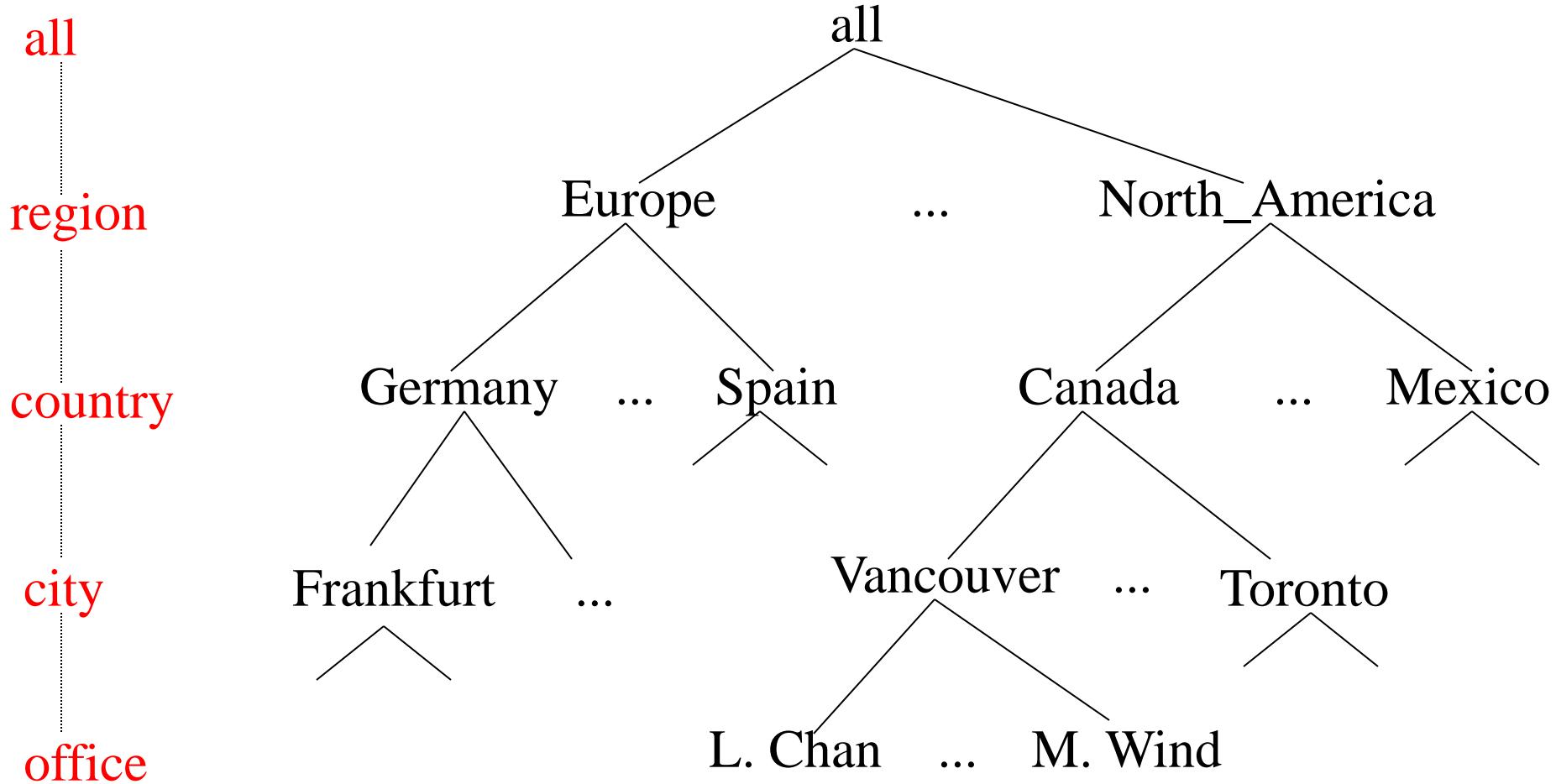
مثالی از شمای برفگونه



مثالی از صورت فلکی

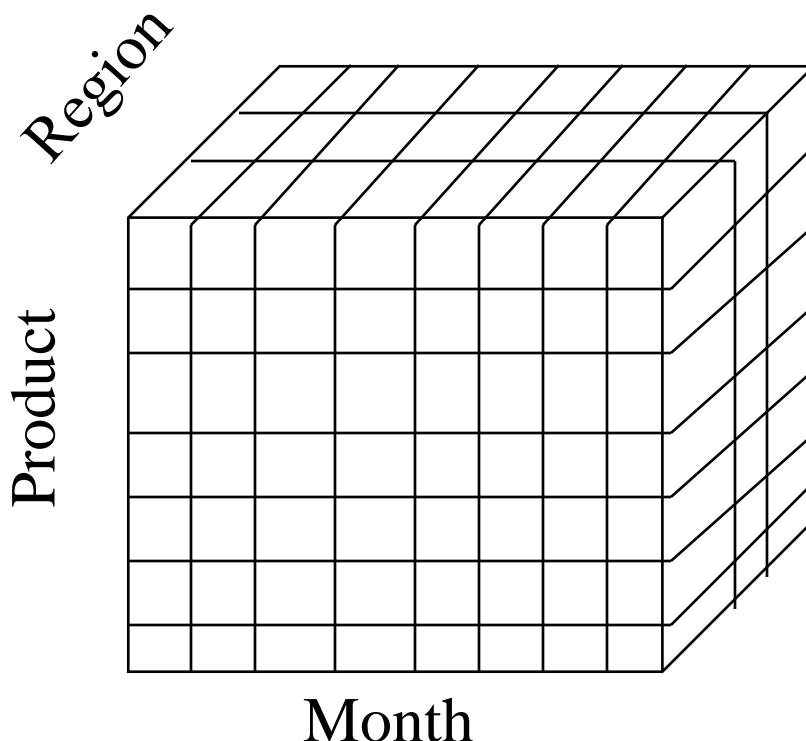


سلسله مراتب مفهومی در ابعاد

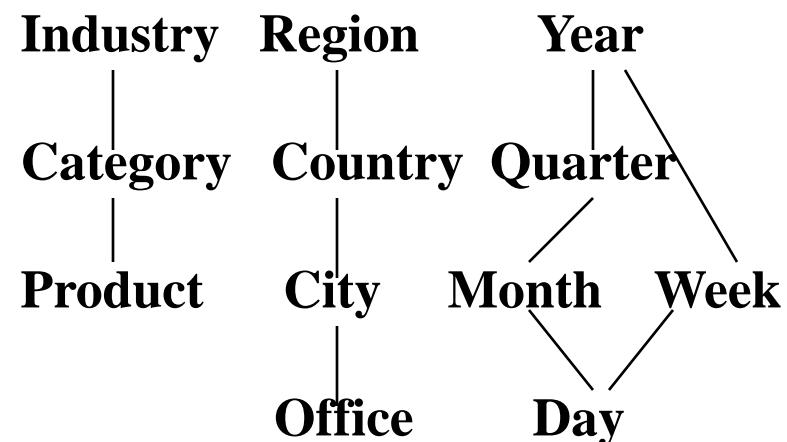


داده چند بعدی

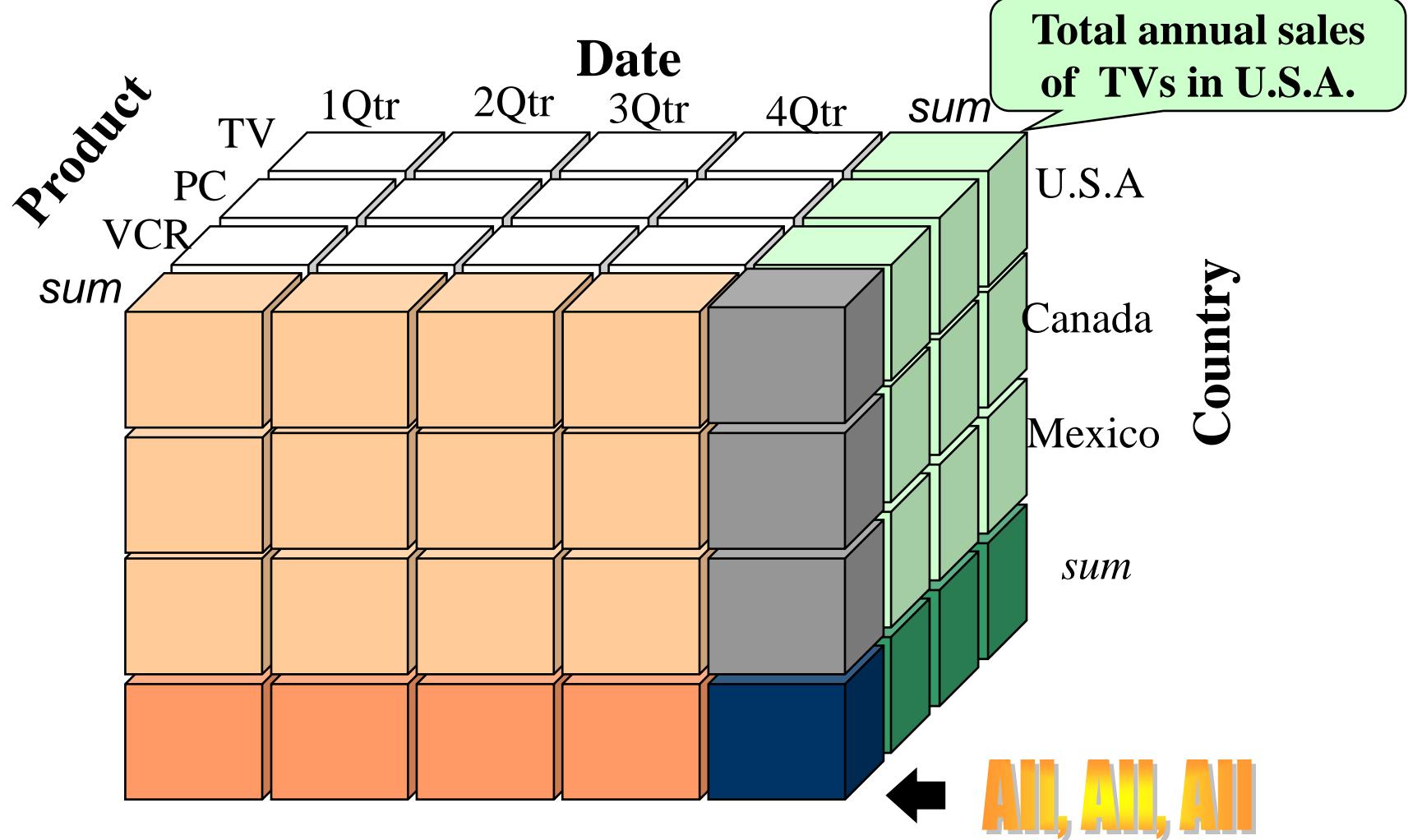
- حجم فروش به عنوان عملکرد محصول، ماه و منطقه



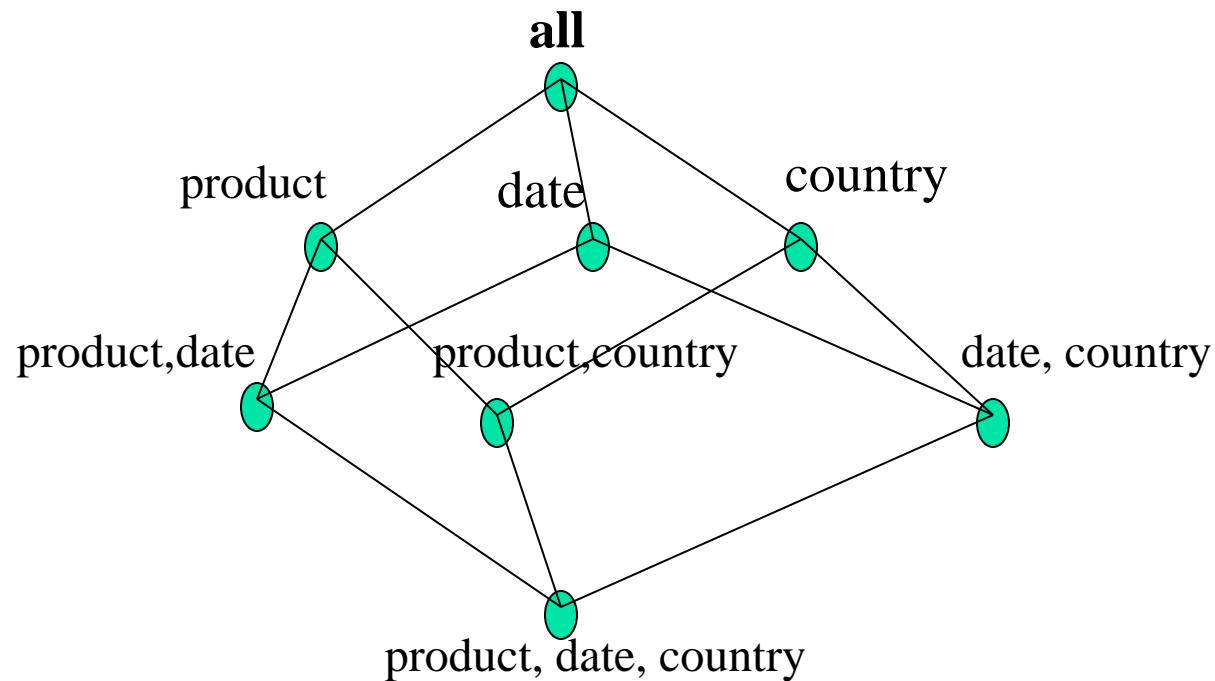
Dimensions: *Product, Location, Time*
Hierarchical summarization paths



یک مکعب داده نمونه



شبه مکعب های وابسته به مکعب داده



0-D (*apex*) cuboid

1-D cuboids

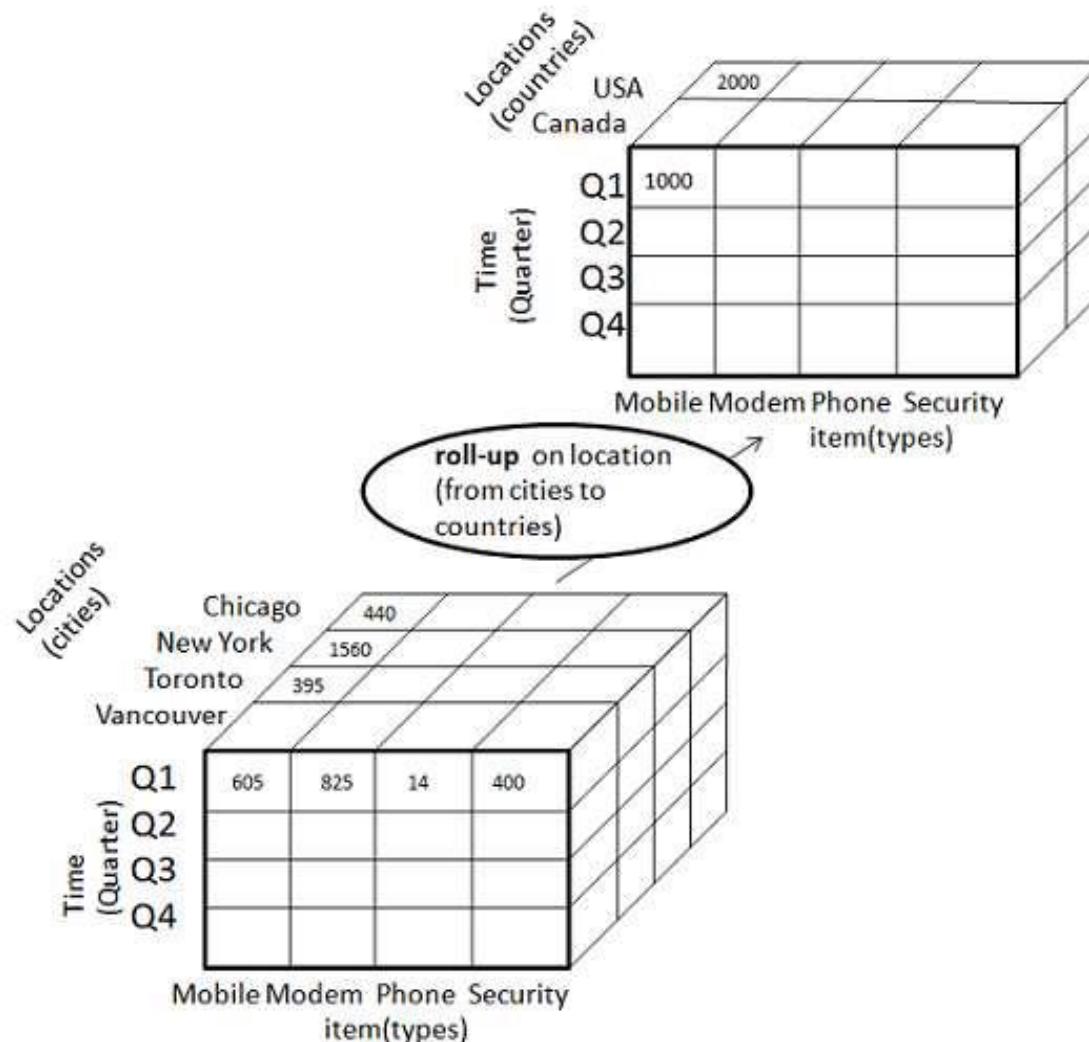
2-D cuboids

3-D (*base*) cuboid

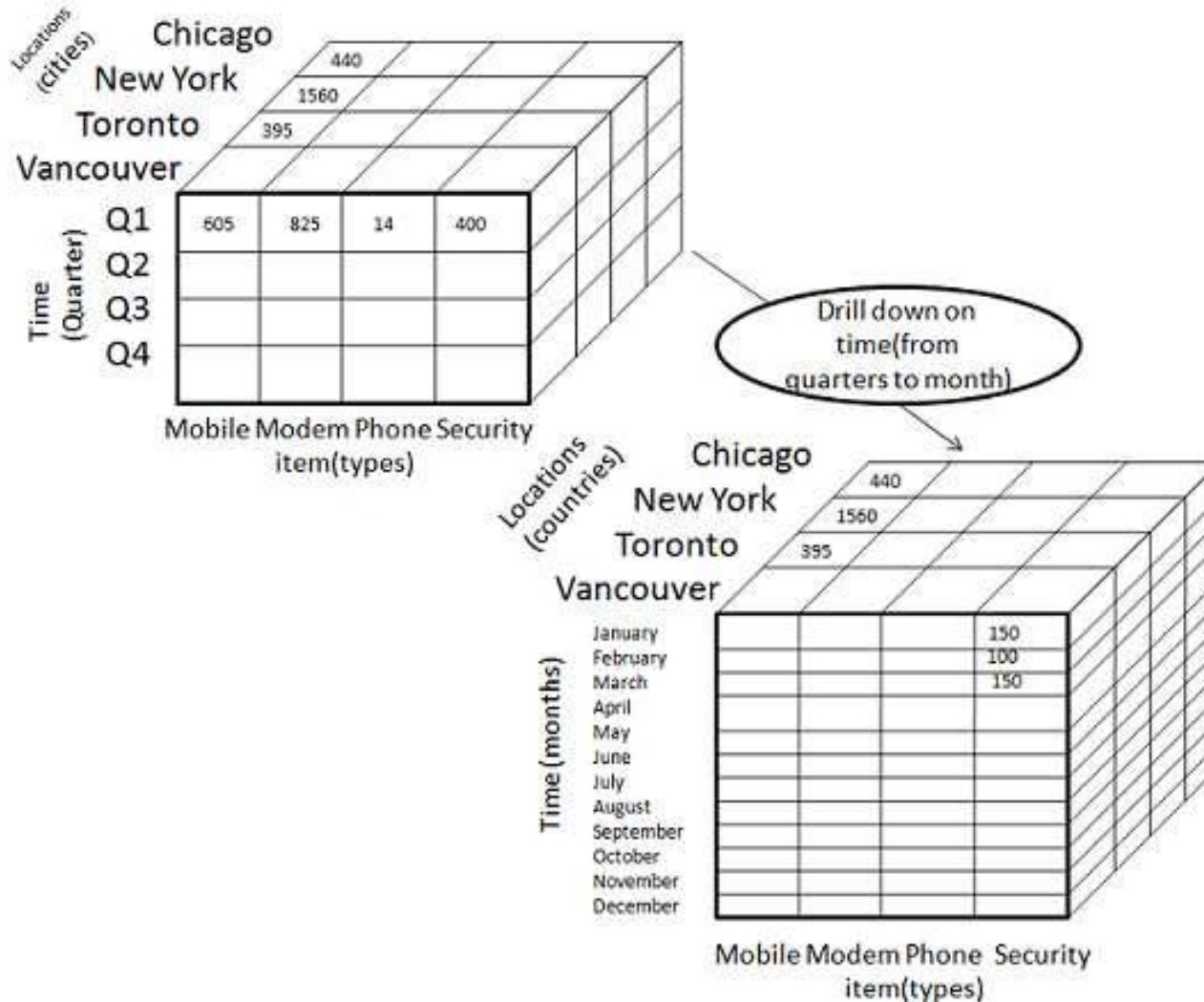
عملیات متداول در OLAP

- **Roll up (drill-up)**: خلاصه کردن داده با بالا رفتن در سلسله مراتب یا حذف ابعاد
- **Drill down (roll down)**: بر عکس Roll up، از خلاصه سازی سطح بالا به خلاصه سازی سطح پایین تر یا داده های جزئی تر، یا معرفی ابعاد جدید
- **Slice**: انجام select روی یک بعد
- **dice**: انجام select روی چند بعد
- **Pivot (rotate)**: چرخش مکعب و دیدن داده ها از منظر دیگر

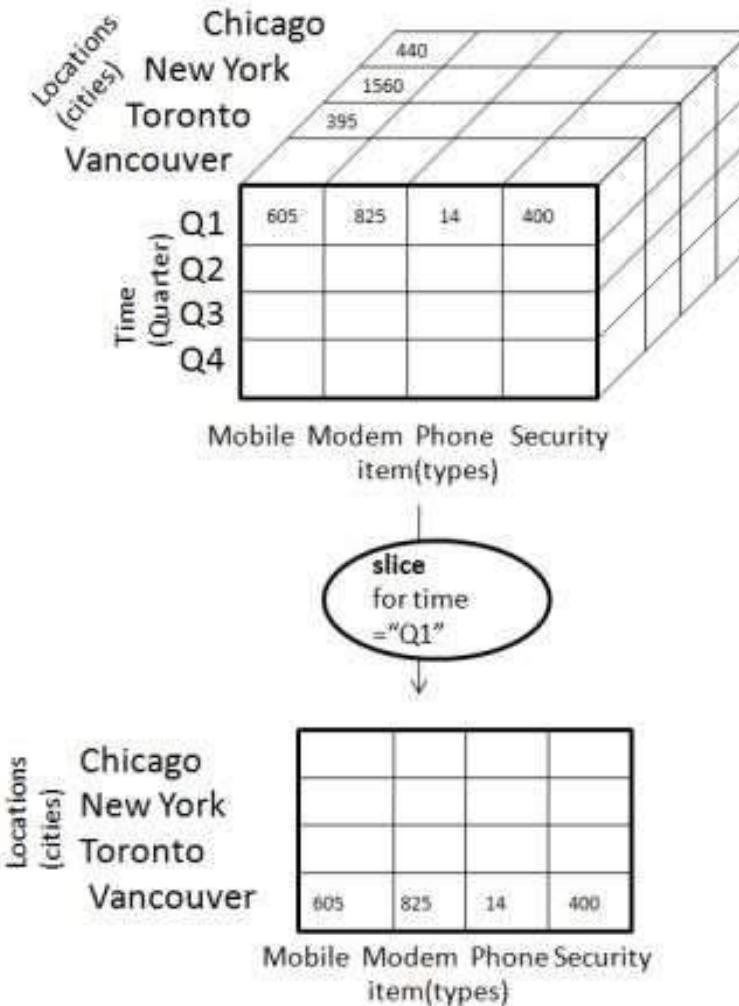
Roll-up



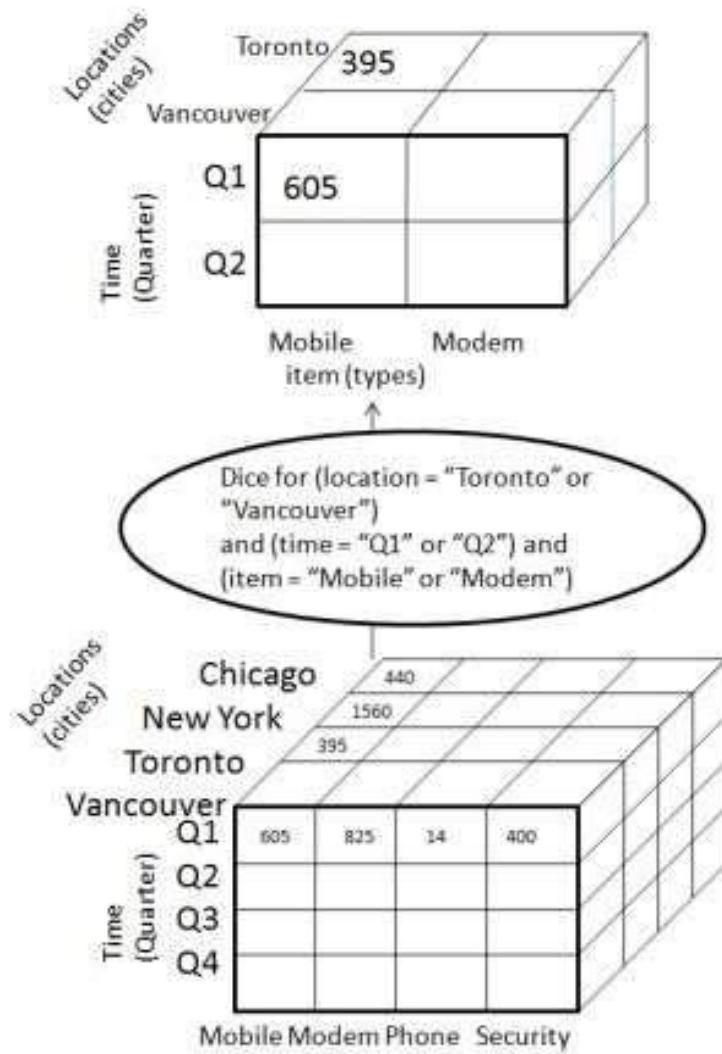
Drill-down



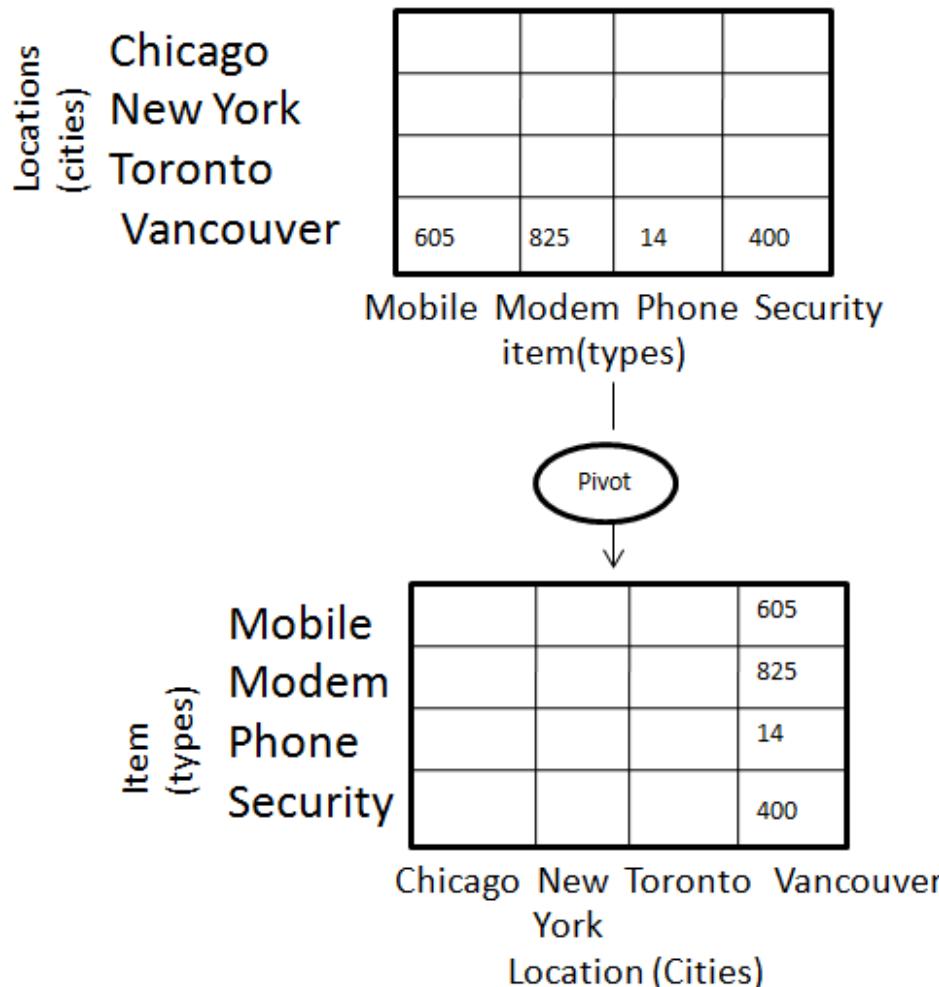
Slice

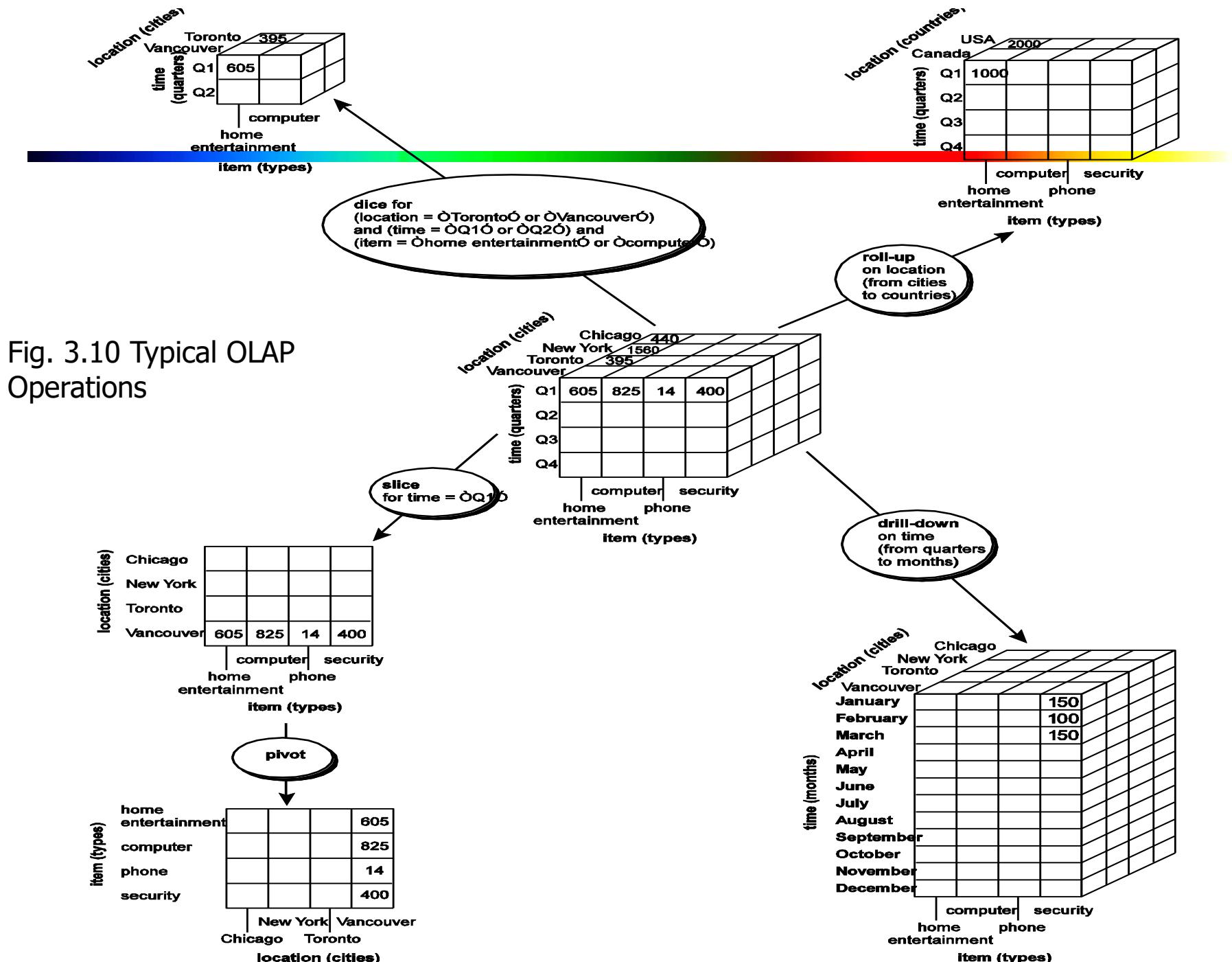


Dice



Pivot







داده کاوی

مفاهیم و تکنیک ها

— فصل ۶ —

کاوش الگوهای مکرر، مشارکت‌ها و همبستگی‌ها: مفاهیم پایه و روش‌ها

■ مفاهیم اولیه 

■ روش‌های کاوش الگوهای مکرر

■ چه الگوهایی جذاب هستند؟ (روش‌های ارزیابی الگوها)

■ خلاصه

تحليل الگو چیست؟

الگوی پر تکرار: یک الگو (یک مجموعه از اقلام، زیرتوالی، زیرساختار و غیره) که بصورت مکرر در یک مجموعه داده رخ میدهد.

مجموعه اقلام: مثل خرید توام نان و شیر در بسیاری از تراکنش ها

زیرتوالی ها: مثلاً معمولاً بعد از دوربین کارت حافظه خریده می شود.

زیرساختار ها: مانند زیر گراف یا زیر درخت

...

نخستین بار توسط Agrawal, Imielinski, and Swami [AIS93] در زمینه مجموعه آیتم های پر تکرار (association rule mining) و (frequent itemsets) پیشنهاد شد.

انگیزه: پیدا کردن نظم ذاتی در داده ها

- چه محصولاتی معمولاً با هم خریده می شوند؟
- خرید بعدی بعد از خریدن یک کامپیوتر چیست؟
- چه انواعی از DNA به این دارو حساس است؟
- آیا می توانیم بصورت اتوماتیک اسناد وب را دسته بندی کنیم؟

کاربردها:

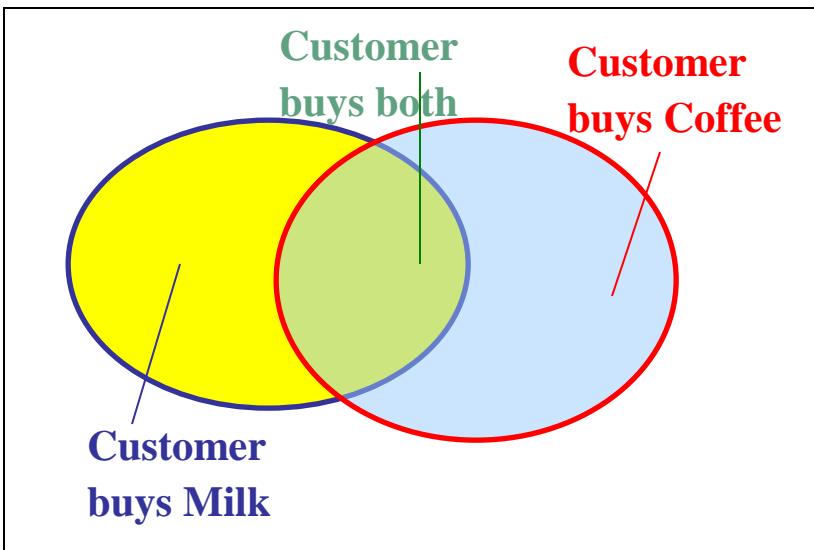
- تحلیل سبد خرید، بازاریابی، طراحی کاتالوگ، آنالیز سلسله عملیات فروش، تحلیل جریان کلیک در وب، تحلیل توالی DNA

چرا کاوش الگو مهم است؟

- الگوهای مکرر یک ویژگی ذاتی و مهم مجموعه داده ها است.
- اساس بسیاری از عملیات مهم در داده کاوی است
- تحلیل قواعد انجمنی، همبستگی و علیت
- الگوهای متوالی، ساختاری (به عنوان مثال، زیر گراف)
- تجزیه و تحلیل الگوهای مکانی، چند رسانه ای، سری زمانی، و جریان داده ها
- دسته بندی: تحلیل الگوی مکرر متمایز کننده
- آنالیز خوشه ای: خوشه بندی بر اساس الگوهای مکرر
- انبارسازی داده: cube-gradient و iceberg cube
- فشرده سازی داده های معنایی
- کاربردهای گسترده

مفاهیم پایه: الگوهای پر تکرار

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- مجموعه اقلام: مجموعه ای از یک یا چند آیتم (مثل کالا)

- مجموعه اقلام شامل K آیتم

$$X = \{x_1, \dots, x_k\}$$

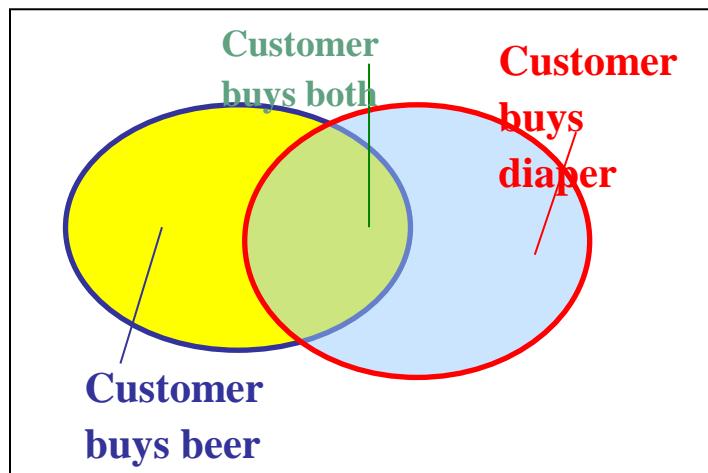
- support یا (absolute) support : تعداد تکرار یک مجموعه آیتم count

- (relative) support : درصد تراکنش های حاوی X یا احتمال اینکه یک تراکنش حاوی X باشد.

- یک itemset پر تکرار است اگر آن از حد minsup پایین تر نباشد.

مفاهیم پایه: قوانین انجمنی

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



پیدا کردن همه قوانین $Y \rightarrow X$ که دو مقدار آستانه (min sup, min conf) را ارضامی کنند.

- به معنی احتمال اینکه تراکنش $Y \cup X$ باشد.

- احتمال شرطی $X \rightarrow Y$ که یک تراکنش حاوی X حاوی Y هم باشد.

- فرض کنید

$$\text{minsup} = 50\%, \text{minconf} = 50\%$$

Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

Association rules: (many more!)

- $\text{Beer} \rightarrow \text{Diaper}$ (60%, 100%)
- $\text{Diaper} \rightarrow \text{Beer}$ (60%, 75%)

Max-Patterns و Closed Patterns

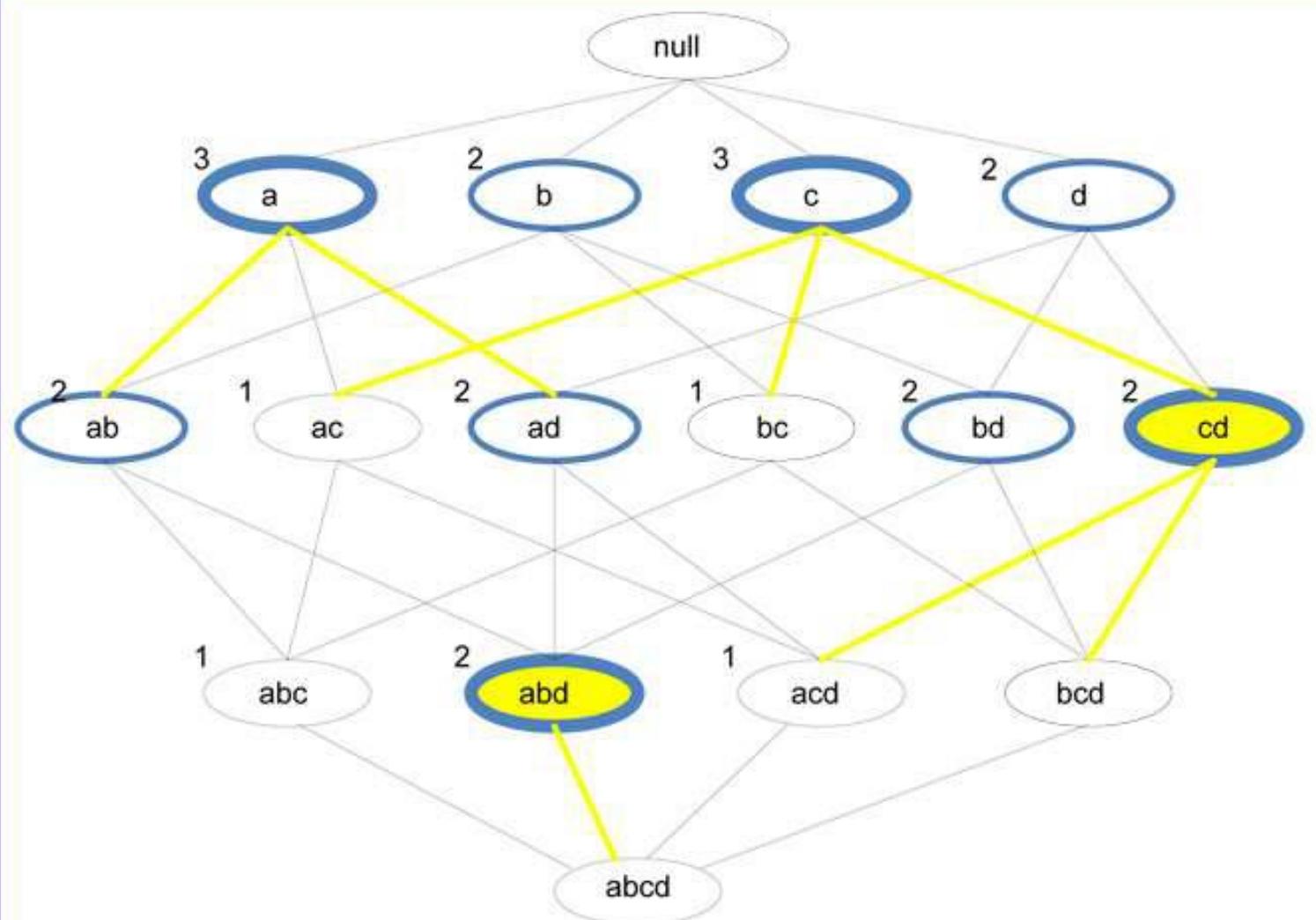
یک الگوی طولانی شامل تعداد زیادی زیر الگو است. مثلا $\{a_1, \dots, a_{100}\}$ شامل $(_{100}^1) + (_{100}^2) + \dots + (_{100}^{100}) = 2^{100} - 1 = 1.27 * 10^{30}$ زیر الگو می باشد.

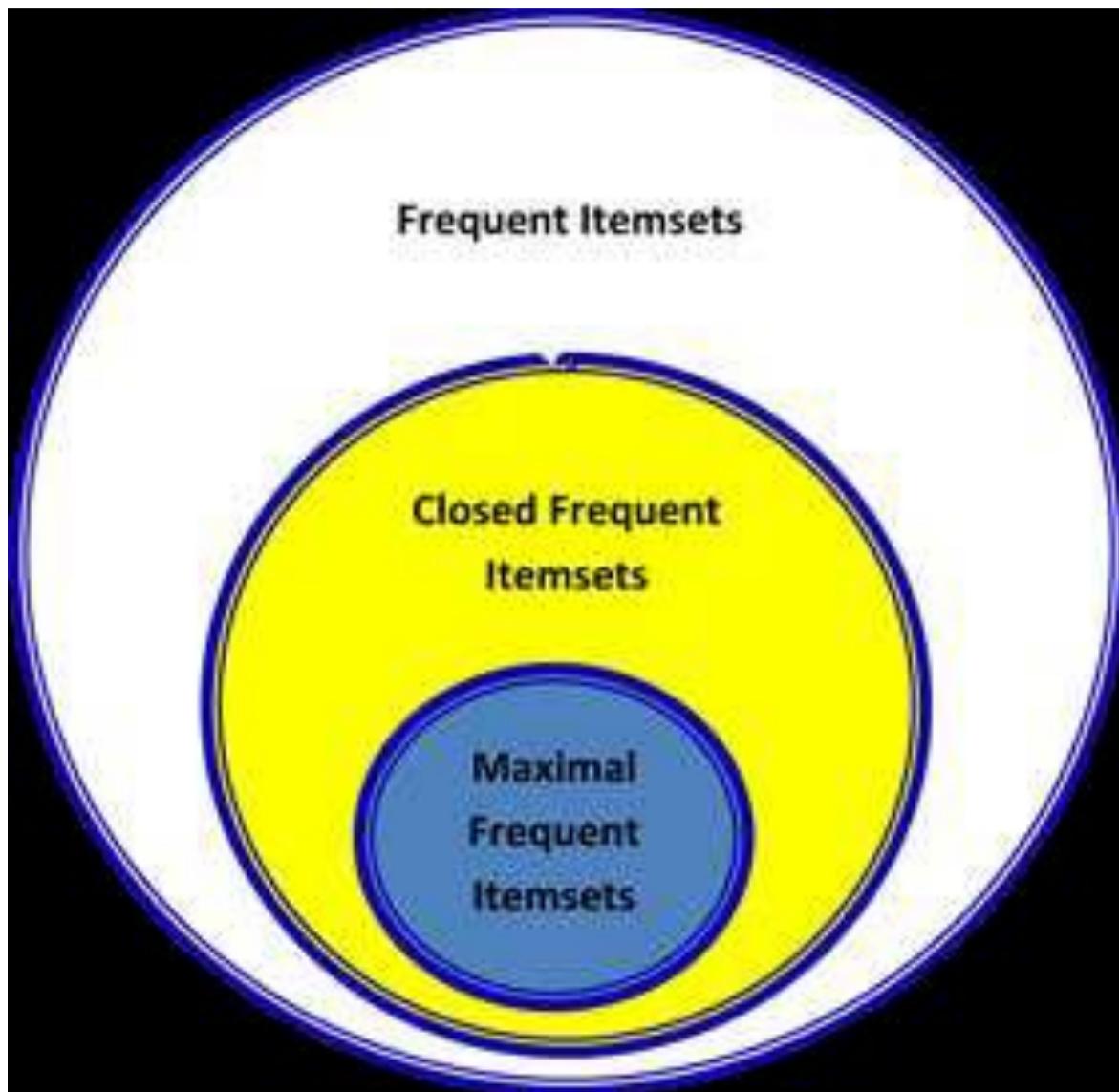
راه حل : کاوش *max-patterns* و *closed patterns* بجای همه الگوها

مجموعه اقلام X را *closed pattern* می نامیم اگر X پر تکرار باشد و هیچیک از *support* های آن *super-pattern* معادل آن را نداشته باشند

مجموعه اقلام X را *max-pattern* می نامیم اگر X پر تکرار باشد و هیچیک از *support* های آن پر تکرار نباشد.

مثال

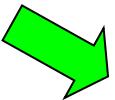




پیچیدگی محاسباتی کاوش الگوهای مکرر

- در بدترین حالت چند itemset ممکن است تولید شوند؟
 - تعداد الگوهای مکرر تولید شده وابسته به $minsup$ است.
 - اگر $minsup$ پایین باشد تعداد الگوهای پرتکرار نمایی خواهد بود.
 - بدترین حالت: اگر M تعداد اقلام و N ماکریم طول تراکنش ها باشد تعداد الگوها M^N خواهد بود.
-
- پیچیدگی در بدترین حالت در مقابل احتمال رخداد یک الگو:
 - فرض کنید فروشگاهی 10^4 نوع محصول متفاوت دارد.
 - شанс انتخاب یکی از محصولات: 10^{-4}
 - شанс انتخاب یک مجموعه حاوی ۱۰ محصول مشخص: $\sim 10^{-40}$
 - شанс اینکه این مجموعه خاص به اندازه 10^3 بار در 10^9 تراکنش تکرار شده باشد چیست؟

کاوش الگوهای مکرر، مشارکت‌ها و همبستگی‌ها: مفاهیم پایه و روش‌ها

- مفاهیم اولیه
- روش‌های کاوش الگوهای مکرر 
- چه الگوهایی جذاب هستند؟ (روش‌های ارزیابی الگوها)
- خلاصه

روش‌های مقیاس پذیر کاوش الگوهای مکرر

- Apriori: یک روش مبتنی بر تولید و آزمایش مجموعه های کاندید
- FP-Growth: ساخت الگوهای مکرر از طریق گسترش آنها
- ECLAT: کاوش الگوهای مکرر بر اساس چیدمان عمودی داده ها

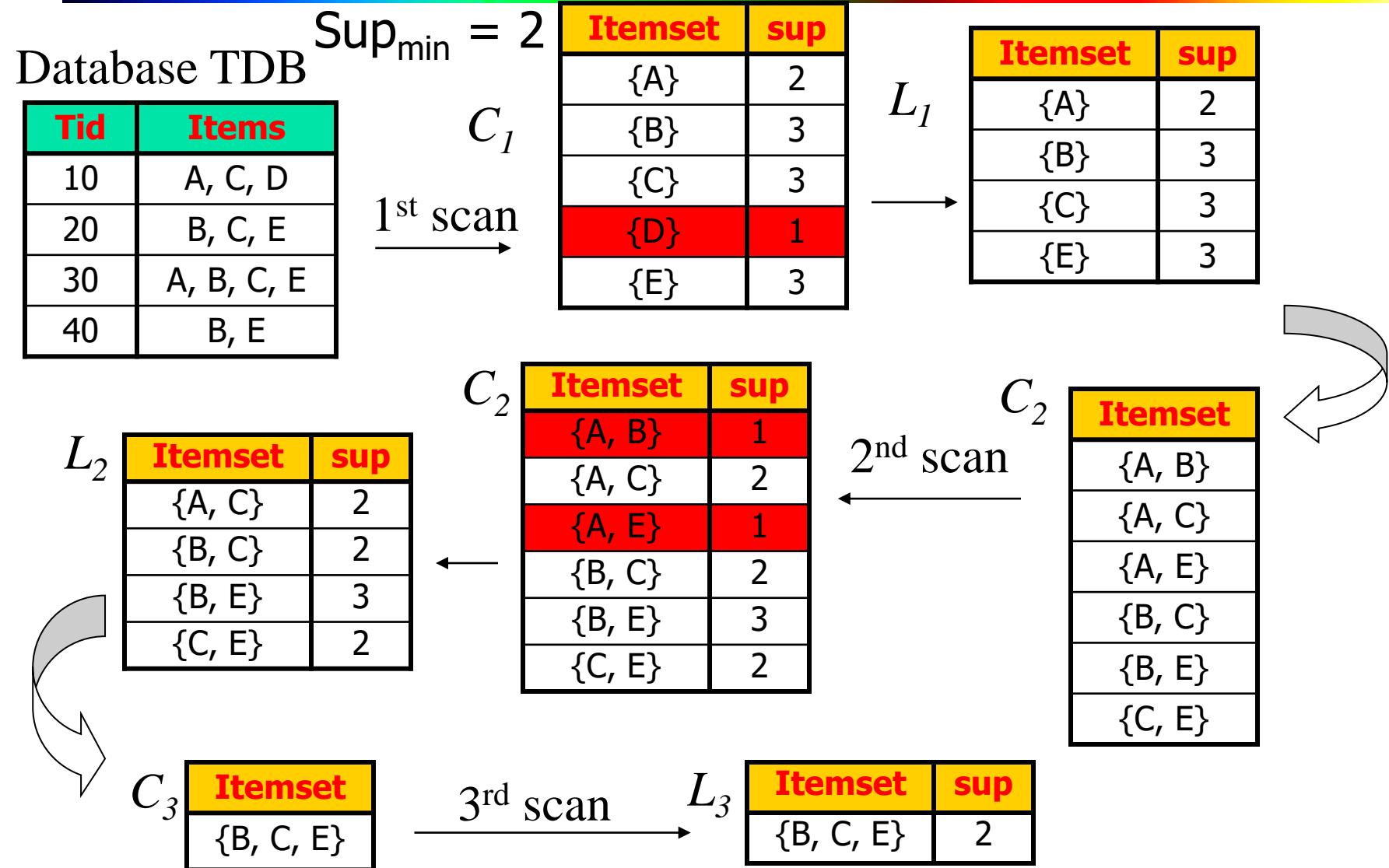
ویژگی Downward Closure و روش های مقیاس پذیر

- ویژگی downward closure الگوهای مکرر
- هر زیرمجموعه یک الگوی یرتکرار لزوماً یرتکرار است.
- اگر **{beer, diaper}** پر تکرار است **{beer, diaper, nuts}** هم پر تکرار است.
- همه تراکنش های حاوی اولی دومی را هم در خود دارد.
- سه روش عمده مقیاس پذیر برای کاوش الگوهای مکرر:
 - Apriori (Agrawal & Srikant@VLDB'94)
 - Freq. pattern growth (FPgrowth—Han, Pei & Yin @SIGMOD'00)
 - Vertical data format approach (Charm—Zaki & Hsiao @SDM'02)

Apriori: یک روش تولید و تست الگوهای کاندید

- اساس هرس کردن روش Apriori: اگر یک مجموعه اقلام پرتکرار نباشد هیچ superset آن نباید تولید و آزمایش شود.
- روش کار:
 - پایگاه داده در ابتدا برای پیدا کردن مجموعه های تک عضوی پرتکرار اسکن می شود.
 - کاندیداهای با طول $K+1$ از الگوهای پرتکرار با طول K تولید می شوند.
 - کاندیداهای تست می شوند.
 - هرگاه مجموعه پرتکرار یا کاندیدای دیگری نبود خاتمه می یابد.

مثال



The Apriori Algorithm (Pseudo-Code)

C_k : Candidate itemset of size k

L_k : frequent itemset of size k

$L_1 = \{\text{frequent items}\};$

for ($k = 1; L_k \neq \emptyset; k++$) **do begin**

C_{k+1} = candidates generated from L_k ;

for each transaction t in database do

 increment the count of all candidates in C_{k+1} that
 are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

end

return $\cup_k L_k$

پیاده سازی Apriori

چگونه کandidاها تولید می شوند؟

- Step 1: self-joining L_k
- Step 2: pruning

مثال برای تولید کandida

- $L_3 = \{abc, abd, acd, ace, bcd\}$
- Self-joining: $L_3 * L_3$
 - $abcd$ from abc and abd
 - $acde$ from acd and ace

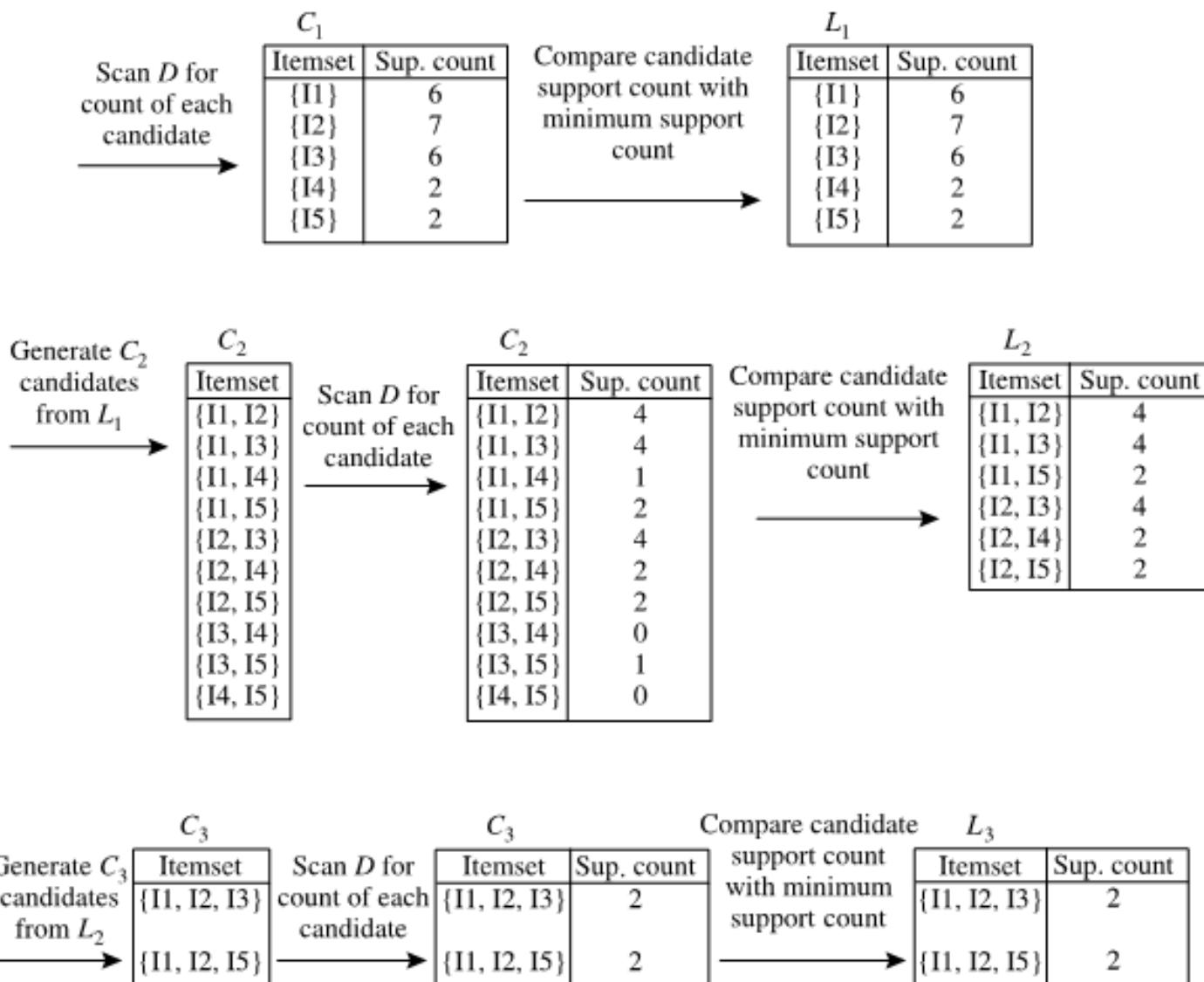
هرس کردن:

- $acde$ is removed because ade is not in L_3
- $C_4 = \{abcd\}$

مثال دیگر

Table 6.1 Transactional Data for an *AllElectronics* Branch

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3



تولید قوانین انجمنی از مجموعه های پر تکرار

- شرط قوی بودن قواعد انجمنی تامین آستانه تعیین شده برای min-confidence و min-support است.
- در الگوهای پر تکرار min-support تامین است.
- قوانین انجمنی به شکل زیر تولید می شوند:
برای هر مجموعه اقلام پر تکرار مانند L تمام زیر مجموعه های غیر تهی آن ایجاد می شوند. برای هر زیر مجموعه غیر تهی S قانون به صورت زیر تولید می شود:

$$s \rightarrow (l - s) \text{ if } \text{support_count}(l)/\text{support_count}(s) \geq \text{min_conf}$$

مثال

- فرض کنید $X = \{I1, I2, I5\}$ پر تکرار باشد.
- قوانین انجمنی تولید شده از X :

$\{I1, I2\} \Rightarrow I5$, $confidence = 2/4 = 50\%$

$\{I1, I5\} \Rightarrow I2$, $confidence = 2/2 = 100\%$

$\{I2, I5\} \Rightarrow I1$, $confidence = 2/2 = 100\%$

$I1 \Rightarrow \{I2, I5\}$, $confidence = 2/6 = 33\%$

$I2 \Rightarrow \{I1, I5\}$, $confidence = 2/7 = 29\%$

$I5 \Rightarrow \{I1, I2\}$, $confidence = 2/2 = 100\%$

در صورتی انتخاب خواهند شد.

بهبود روش Apriori

- مهمترین چالش های محاسباتی
- مرور چندباره پایگاه داده تراکنش ها
- تعداد زیاد کandidاها
- حجم زیاد کار شمارش تعداد تکرار کandidاها
- اصلاح Apriori: ایده های کلی
 - کاهش تعداد مرور های پایگاه داده تراکنش ها
 - کاهش تعداد کandidاها
 - تسهیل شمارش تعداد تکرار کandidاها

پارتیشن بندی: فقط دو بار مرور پایگاه داده

- هر itemset که امکان پر تکرار بودن در کل پایگاه داده را داشته باشد باید حداقل در یکی از پارتیشن ها به نسبت پر تکرار باشد.
- مرور اول: تقسیم پایگاه داده و پیدا کردن الگوهای مکرر محلی
- مرور دوم: شمارش و تعیین الگوهای مکرر سراسری از میان کandidاهای مرحله قبل

A. Savasere, E. Omiecinski and S. Navathe, VLDB'95 ■

$$\boxed{\text{DB}_1} + \boxed{\text{DB}_2} + \dots + \boxed{\text{DB}_k} = \boxed{\text{DB}}$$
$$\text{sup}_1(i) < \sigma \text{DB}_1 \quad \text{sup}_2(i) < \sigma \text{DB}_2 \quad \dots \quad \text{sup}_k(i) < \sigma \text{DB}_k \quad \text{sup}(i) < \sigma \text{DB}$$

تکنیک مبتنی بر درهم سازی: کاہش تعداد کاندیداها

Table 6.1 Transactional Data for an *AllElectronics* Branch

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

$$h(x, y) = ((\text{order of } x) \times 10 + (\text{order of } y)) \bmod 7$$



■ همزمان با شمارش مجموعه اقلام یک عضوی جدول درهم سازی از ترکیب های دوتایی پر می شود. مجموعه اقلام دوتایی که مجموعه تعداد باکت آنها کمتر از $\min-sup$ باشد از کاندیداهای مرحله بعد حذف می شوند.

H_2

bucket address	0	1	2	3	4	5	6
bucket count	2	2	4	2	2	4	4
bucket contents	{I1, I4} (I3, I5)	{I1, I5} (I1, I5)	{I2, I3} (I2, I3) (I2, I4)	{I2, I4} (I2, I5)	{I2, I5} (I1, I2)	{I1, I2} (I1, I3)	{I1, I3}

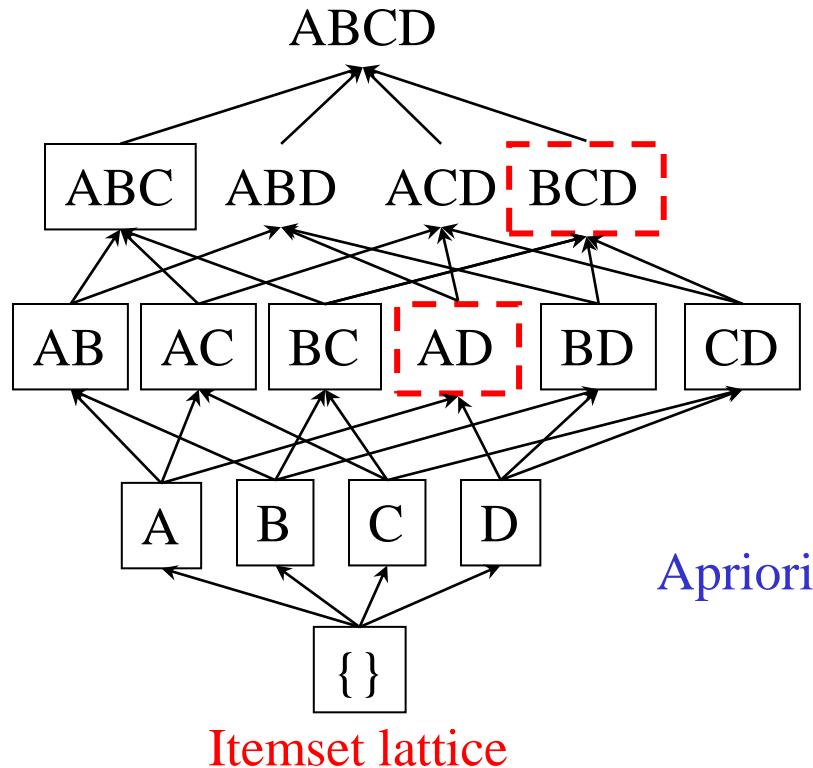
Hash Table

نمونه گیری برای الگوهای پر تکرار

- انتخاب یک نمونه از پایگاه داده اصلی که در حافظه اصلی قرار گیرد، جستجوی الگوهای مکرر در داخل نمونه با استفاده از Apriori
- پیمایش کل پایگاه برای شمارش تعداد کاندیداها در کل (با توجه به تقریبی بودن روش، برای اینکه تعداد کمتری الگوی پر تکرار از دست بروند کاندیداها را از حد آستانه پایین تری انتخاب می کنیم).

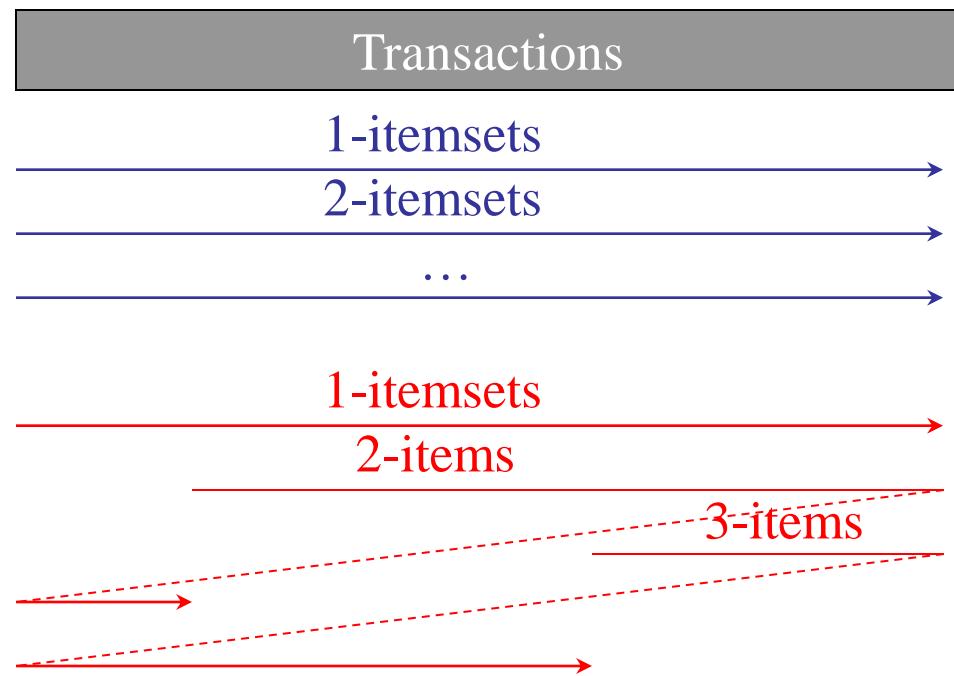
H. Toivonen. Sampling large databases for association rules. In VLDB'96

شمارش پویا (DIC): کاهش تعداد پیمایش ها



S. Brin R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *SIGMOD'97*

- در طی شمارش هر جا پر تکرار بودن A و D محزز شد
- شمارش AD از همان محل شروع می شود
- هر جا پر تکرار بودن همه زیرمجموعه های دوتایی BCD مشخص شد از همان نقطه شمارش تعداد تکرار BCD آغاز می شود.



DIC

مثال

<i>Database</i>			
<i>Items</i>		<i>Transaction ID (tid)</i>	<i>Items bought</i>
Bread	<i>a</i>	1	<i>a b d e</i>
Butter	<i>b</i>	2	<i>b c e</i>
Milk	<i>c</i>	3	<i>a b d e</i>
cheese	<i>d</i>	4	<i>a b c e</i>
coke	<i>e</i>	5	<i>a b c d e</i>
		6	<i>b c d</i>

C_1		
pattern	support	state
a	0/0	PI
b	0/0	PI
c	0/0	PI
d	0/0	PI
e	0/0	PI

reading
objects
1-3



C_1		
pattern	support	state
a	2/3	PF
b	3/3	PF
c	2/3	PF
d	2/3	PF
e	3/3	PF

C_1		
pattern	support	state
a	4/6	CF
b	6/6	CF
c	4/6	CF
d	4/6	CF
e	5/6	CF

C_2		
pattern	support	state
ab	0/0	PI
ac	0/0	PI
ad	0/0	PI
ae	0/0	PI
bc	0/0	PI
bd	0/0	PI
be	0/0	PI
cd	0/0	PI
ce	0/0	PI
de	0/0	PI

C_2		
pattern	support	state
ab	2/3	PF
ac	1/3	PF
ad	1/3	PF
ae	2/3	PF
bc	2/3	PF
bd	2/3	PF
be	2/3	PF
cd	1/3	PF
ce	1/3	PF
de	1/3	PF

C_3		
pattern	support	state
abc	0/0	PI
abd	0/0	PI
abe	0/0	PI
acd	0/0	PI
ace	0/0	PI
ade	0/0	PI
bcd	0/0	PI
bce	0/0	PI
bde	0/0	PI
cde	0/0	PI

C_2		
pattern	support	state
ab	4/6	CF
ac	2/6	CF
ad	3/6	CF
ae	4/6	CF
bc	4/6	CF
bd	4/6	CF
be	5/6	CF
cd	2/6	CF
ce	3/6	CF
de	3/6	CF

reading
objects
4-6



C_3		
pattern	support	state
abc	2/6	CF
abd	3/6	CF
abe	4/6	CF
acd	1/6	CI
ace	2/6	CF
ade	3/6	CF
bcd	2/6	CF
bce	2/6	CF
bde	3/6	CF
cde	1/6	CI

reading
objects
1-3



C_4		
pattern	support	state
abcd	1/6	CI
abce	2/6	CF
abde	3/6	CF
acde	1/6	CI
bcde	1/6	CI

C_3		
pattern	support	state
abc	1/3	PF
abd	2/3	PF
abe	2/3	PF
acd	1/3	PF
ace	1/3	PF
ade	2/3	PF
bcd	1/3	PF
bce	1/3	PF
bde	2/3	PF
cde	1/3	PF

C_4		
pattern	support	state
abcd	0/3	PI
abce	1/3	PF
abde	1/3	PF
acde	0/3	PI
bcde	0/3	PI

C_4		
pattern	support	state
abcd	0/0	PI
abce	0/0	PI
abde	0/0	PI
acde	0/0	PI
bcde	0/0	PI

FPGrowth: روشی برای کاوش الگوهای مکرر بدون تولید کاندیداها

معایب روش Apriori

- جستجوی اول سطح
- تولید مجموعه های کاندید و تست آن ها
- اغلب تعداد کاندیداها بسیار زیاد است.

روش FPGrowth

- جستجوی اول عمق
- پرهیز از تولید صریح کاندیداها

فلسفه اصلی: رشد الگوهای طولانی از الگوهای کوتاه با استفاده از الگوهای مکرر محلی

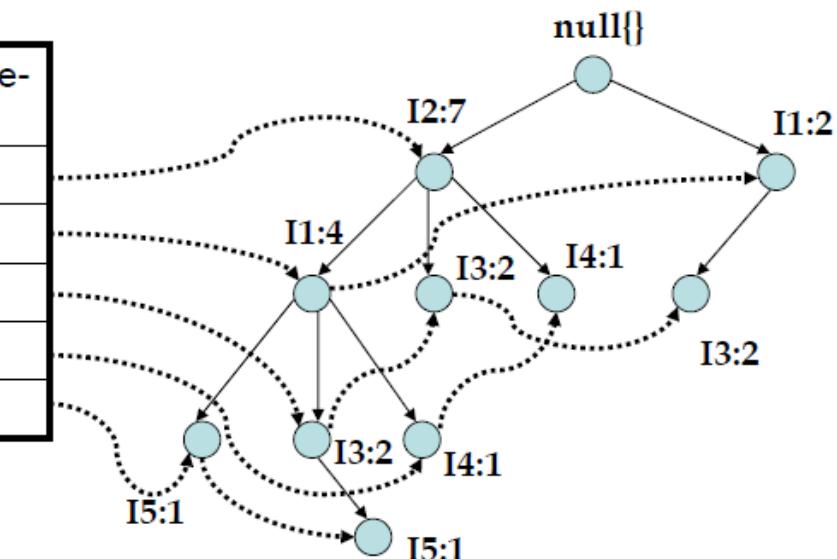
- “abc” is a frequent pattern
- Get all transactions having “abc”, i.e., project DB on abc: DB|abc
- “d” is a local frequent item in DB|abc → abcd is a frequent pattern

فاز اول: ساخت FP-tree از یک پایگاه داده تراکنش

۱. پایگاه داده را یکبار مرور کنید و الگوهای مکرر تک عضوی را بیابید.
۲. الگوهای مکرر یافته شده را بر مبنای تعداد تکرار بصورت نزولی مرتب کنید، f-list
۳. یکبار دیگر پایگاه را مرور کنید اقلام موجود در هر تراکنش را بر اساس تعداد تکرار آنها مرتب کنید و FP-tree را بسازید.

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Item Id	Sup Count	Node-link
I2	7	
I1	6	
I3	6	
I4	2	
I5	2	



فاز دوم: استخراج درخت های شرطی و الگوهای مکرر

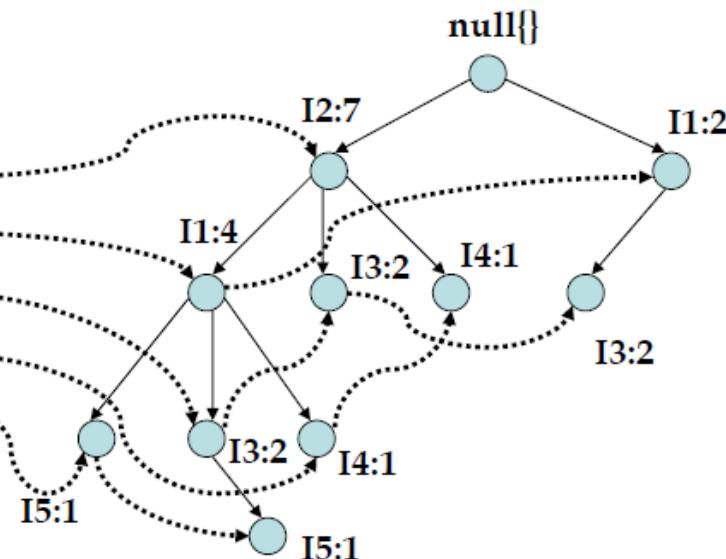
- از آخرین عنصر فهرست یعنی I5 شروع می کنیم. دو مسیر در درخت به I5 ختم می شود:

1 : (I2, I1, I3, I5:1)

2 : (I2, I1, I5:1)

- I5 را پسوند الگو و (I2, I1, I3) (I2,I1) را پایگاه الگوی شرطی یا Conditional pattern base می نامیم.

Item Id	Sup Count	Node-link
I2	7	
I1	6	
I3	6	
I4	2	
I5	2	



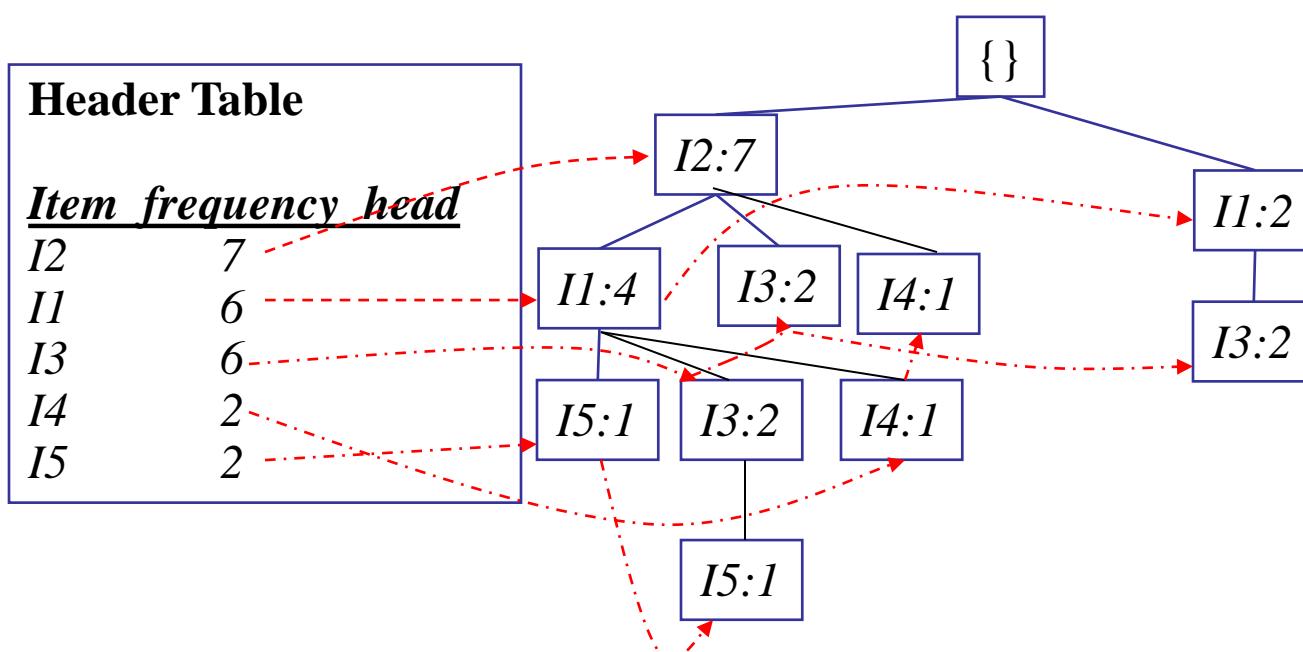
■ دو مقدار $(I_2, I_1, I_3, I_5:1)$ و (I_2, I_1, I_3) را به عنوان تراکنش در نظر می‌گیریم و الگوریتم را ادامه می‌دهیم.

■ با استفاده از این دو تراکنش درخت شرطی I_5 ساخته می‌شود که حاوی تنها یک مسیر $<I_2:2, I_1:2>$ است. I_3 به دلیل اینکه تعداد تکرار آن کمتر از min-sup است حذف می‌شود.

■ ترکیبات الگوهای مکرر با استفاده از این تک مسیر ایجاد می‌شود:

- $\{I_2, I_5 : 2\}$
- $\{I_1, I_5 : 2\}$
- $\{I_2, I_1, I_5 : 2\}$

Item	conditional pattern base	conditional FP-tree	frequent patterns generated
I5	$\{(I2 \ I1: 1), (I2 \ I1 \ I3:1)\}$	$\langle I2:2, I1:2 \rangle$	$I2 \ I5:2, I1 \ I5:2, I2 \ I1 \ I5:2$
I4	$\{(I2 \ I1:1), (I2:1)\}$	$\langle I2: 2 \rangle$	$I2 \ I4:2$
I3	$\{(I2 \ I1: 2), (I2:2), (I1:2)\}$	$\langle I2:4, I1:2 \rangle, \langle I1:2 \rangle$	$I2 \ I3:4, I1 \ I3:4$
I1	$\{(I2: 4)\}$	$\langle I2: 4 \rangle$	$I2 \ I1:4$



ECLAT: کاوش مجموعه اقلام مکرر با استفاده از قالب عمودی داده ها

- در این الگوریتم قالب نگهداری داده ها عوض می شود و بجای نگهداری تراکنش ها اطلاعات به شکل زیر نگهداری می شود:

The Vertical Data Format of the Transaction Data Set D of Table 6.1

itemset	TID_set
I1	{T100, T400, T500, T700, T800, T900}
I2	{T100, T200, T300, T400, T600, T800, T900}
I3	{T300, T500, T600, T700, T800, T900}
I4	{T200, T400}
I5	{T100, T800}

برای پیدا کردن 1-itemset های پر تکرار کافیست تعداد اعضای مجموعه ها شمرده شوند.

برای پیدا کردن 2-itemset های پر تکرار لازم است بین مجموعه ها اشتراک گرفته شود و تعداد اعضای مجموعه حاصل شمرده شود.

...

The Vertical Data Format of the Transaction Data Set D of Table 6.1

itemset	TID_set
I1	{T100, T400, T500, T700, T800, T900}
I2	{T100, T200, T300, T400, T600, T800, T900}
I3	{T300, T500, T600, T700, T800, T900}
I4	{T200, T400}
I5	{T100, T800}

2-Itemsets in Vertical Data Format

itemset	TID_set
{I1, I2}	{T100, T400, T800, T900}
{I1, I3}	{T500, T700, T800, T900}
{I1, I4}	{T400}
{I1, I5}	{T100, T800}
{I2, I3}	{T300, T600, T800, T900}
{I2, I4}	{T200, T400}
{I2, I5}	{T100, T800}
{I3, I5}	{T800}

The Vertical Data Format of the Transaction Data
Set D of Table 6.1

itemset	TID_set
I1	{T100, T400, T500, T700, T800, T900}
I2	{T100, T200, T300, T400, T600, T800, T900}
I3	{T300, T500, T600, T700, T800, T900}
I4	{T200, T400}
I5	{T100, T800}

2-Itemsets in Vertical Data Format

itemset	TID_set
{I1, I2}	{T100, T400, T800, T900}
{I1, I3}	{T500, T700, T800, T900}
{I1, I4}	{T400}
{I1, I5}	{T100, T800}
{I2, I3}	{T300, T600, T800, T900}
{I2, I4}	{T200, T400}
{I2, I5}	{T100, T800}
{I3, I5}	{T800}

3-Itemsets in Vertical Data Format

itemset	TID_set
{I1, I2, I3}	{T800, T900}
{I1, I2, I5}	{T100, T800}



بهدود الگوریتم ECLAT

- برای مجموعه های متراکم و با اشتراکات زیاد از قابلیت diffset استفاده می شود.
- در این روش بجای ذخیره سازی اشتراک دو مجموعه، اختلاف آن ها ذخیره می شود:

$$\{I_1\} = \{T100, T400, T500, T700, T800, T900\}$$
$$\{I_1, I_2\} = \{T100, T400, T800, T900\}$$

$$Diffset(\{I_1\}, \{I_1, I_2\}) = \{T500, T700\}$$

در این روش فضای حافظه کمتری مصرف می شود اما الگوریتم ها متفاوت و پیچیده تر است.

کاوش الگوهای مکرر، مشارکت‌ها و همبستگی‌ها: مفاهیم پایه و روش‌ها

■ مفاهیم اولیه

■ روش‌های کاوش الگوهای مکرر

■ چه الگوهایی جذاب هستند؟ (روش‌های ارزیابی الگوها)

■ خلاصه

قوانین جالب

قوانینی که تابحال با استفاده از معیارهای **confidence** و **support** بدست آورده‌یم قوانین قوی هستند اما لزوماً جالب نیستند.

بنابراین به روش‌های ارزیابی بهتری برای قوانین تولید شده نیاز داریم.

مثال

- به طور مثال فرض کنید مجموعه داده شامل ۱۰۰۰۰ تراکنش است.
- ۶۰۰۰ تراکنش شامل بازی های کامپیوتری
- ۷۵۰۰ تراکنش شامل فیلم ویدیویی
- ۴۰۰۰ تراکنش شامل هر دو مورد
- فرض کنید
- $Min - sup = 30\%$ $Min - conf = 60\%$
- $Buys\ Computer\ games \Rightarrow Buys\ videos$
- $support = 40\%, Confidence = 66\%$

قوانین قوی لزوماً جالب نیستند

- قانون ایجاد شده در مثال قبل یک قانون قوی است ولی جالب نیست.
بلکه یک قانون گمراه کننده است.
- احتمال خرید فیلم ویدیویی ۷۵ درصد است که از ۶۶ درصد بیشتر است.
- در واقع خرید این دو محصول با هم رابطه عکس دارند. یعنی خرید بازی احتمال خرید فیلم را کم می کند.

سایر معیارها

برای بهود چارچوب support-confidence می‌توان از آنالیز همبستگی استفاده کرد:

$A \Rightarrow B$ (*support, confidence, correlation*)

استقلال آماری

جمعیت ۱۰۰۰ دانشجو

۶۰۰ دانش آموز شنا کردن را می دانند. (S)

۷۰۰ دانش آموز دوچرخه سواری را می دانند. (B)

۴۲۰ دانشجو هم دوچرخه سواری و هم شنا می دانند. (S,B)

$$P(S \cap B) = 420/1000 = 0.42$$

$$P(S) * P(B) = 0.6 * 0.7 = 0.42$$

استقلال آماری

ارتباط مثبت

ارتباط منفی

معیار همبستگی lift

- معیار lift با استفاده از رابطه زیر بدست می آید:

$$corr_{A,B} = \frac{P(A \cup B)}{P(A)P(B)}$$

- اگر نتیجه رابطه کمتر از ۱ باشد آنگاه پیشامد A با پیشامد B همبستگی منفی دارد.
- اگر نتیجه بزرگتر از ۱ باشد همبستگی مثبت است و به این معناست که پیشامد یکی بر پیشامد دیگری دلالت دارد.
- اگر نتیجه برابر ۱ باشد آنگاه A و B مستقل هستند و همبستگی بین آنها وجود ندارد.

سایر معیارها

#	Measure	Formula
1	ϕ -coefficient	$\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j, B_k) + \sum_k \max_j P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)}{2 - \max_j P(A_j) - \max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(AB) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(AB) + P(A,\bar{B})P(\bar{A},B)} = \frac{\alpha - 1}{\alpha + 1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(AB)} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(AB)} + \sqrt{P(A,\bar{B})P(\bar{A},B)}} = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1}$
6	Kappa (κ)	$\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$
7	Mutual Information (M)	$\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}$
8	J-Measure (J)	$\min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j))$ $\max \left(P(A, B) \log \left(\frac{P(B A)}{P(B)} \right) + P(\bar{A}\bar{B}) \log \left(\frac{P(\bar{B} A)}{P(\bar{B})} \right), \right.$ $\left. P(A, B) \log \left(\frac{P(A B)}{P(A)} \right) + P(\bar{A}\bar{B}) \log \left(\frac{P(\bar{A} B)}{P(\bar{A})} \right) \right)$
9	Gini index (G)	$\max \left(P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] \right.$ $\left. - P(B)^2 - P(\bar{B})^2, \right.$ $\left. P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] \right.$ $\left. - P(A)^2 - P(\bar{A})^2 \right)$
10	Support (s)	$P(A, B)$
11	Confidence (c)	$\max(P(B A), P(A B))$
12	Laplace (L)	$\max \left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2} \right)$
13	Conviction (V)	$\max \left(\frac{P(A)P(\bar{B})}{P(AB)}, \frac{P(B)P(\bar{A})}{P(BA)} \right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A, B) - P(A)P(B)$
17	Certainty factor (F)	$\max \left(\frac{P(B A) - P(B)}{1 - P(B)}, \frac{P(A B) - P(A)}{1 - P(A)} \right)$
18	Added Value (AV)	$\max(P(B A) - P(B), P(A B) - P(A))$
19	Collective strength (S)	$\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A,B) + P(\bar{A}\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$

کاوش الگوهای مکرر، مشارکت‌ها و همبستگی‌ها: مفاهیم پایه و روش‌ها

- مفاهیم اولیه
- روش‌های کاوش الگوهای مکرر
- چه الگوهایی جذاب هستند؟ (روش‌های ارزیابی الگوها)
- خلاصه 

خلاصه

- مفاهیم پایه: association rules, support-confident framework, closed and max-patterns
- روش های مقیاس پذیر کاوش الگوهای مکرر
- Apriori (Candidate generation & test)
- Projection-based (FPgrowth, CLOSET+, ...)
- Vertical format approach (ECLAT, CHARM, ...)
- کدام الگوها جالب هستند؟
- روش های ارزیابی الگوها

Ref: Basic Concepts of Frequent Pattern Mining

- (**Association Rules**) R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. SIGMOD'93
- (**Max-pattern**) R. J. Bayardo. Efficiently mining long patterns from databases. SIGMOD'98
- (**Closed-pattern**) N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. ICDT'99
- (**Sequential pattern**) R. Agrawal and R. Srikant. Mining sequential patterns. ICDE'95

Ref: Apriori and Its Improvements

- R. Agrawal and R. Srikant. Fast algorithms for mining association rules. VLDB'94
- H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. KDD'94
- A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. VLDB'95
- J. S. Park, M. S. Chen, and P. S. Yu. An effective hash-based algorithm for mining association rules. SIGMOD'95
- H. Toivonen. Sampling large databases for association rules. VLDB'96
- S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket analysis. SIGMOD'97
- S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. SIGMOD'98

Ref: Depth-First, Projection-Based FP Mining

- R. Agarwal, C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent itemsets. *J. Parallel and Distributed Computing*, 2002.
- G. Grahne and J. Zhu, Efficiently Using Prefix-Trees in Mining Frequent Itemsets, Proc. FIMI'03
- B. Goethals and M. Zaki. An introduction to workshop on frequent itemset mining implementations. *Proc. ICDM'03 Int. Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, Melbourne, FL, Nov. 2003
- J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. SIGMOD' 00
- J. Liu, Y. Pan, K. Wang, and J. Han. Mining Frequent Item Sets by Opportunistic Projection. KDD'02
- J. Han, J. Wang, Y. Lu, and P. Tzvetkov. Mining Top-K Frequent Closed Patterns without Minimum Support. ICDM'02
- J. Wang, J. Han, and J. Pei. CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets. KDD'03

Ref: Vertical Format and Row Enumeration Methods

- M. J. Zaki, S. Parthasarathy, M. Ogihsara, and W. Li. Parallel algorithm for discovery of association rules. DAMI:97.
- M. J. Zaki and C. J. Hsiao. CHARM: An Efficient Algorithm for Closed Itemset Mining, SDM'02.
- C. Bucila, J. Gehrke, D. Kifer, and W. White. DualMiner: A Dual-Pruning Algorithm for Itemsets with Constraints. KDD'02.
- F. Pan, G. Cong, A. K. H. Tung, J. Yang, and M. Zaki , CARPENTER: Finding Closed Patterns in Long Biological Datasets. KDD'03.
- H. Liu, J. Han, D. Xin, and Z. Shao, Mining Interesting Patterns from Very High Dimensional Data: A Top-Down Row Enumeration Approach, SDM'06.

Ref: Mining Correlations and Interesting Rules

- S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generalizing association rules to correlations. SIGMOD'97.
- M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. CIKM'94.
- R. J. Hilderman and H. J. Hamilton. *Knowledge Discovery and Measures of Interest*. Kluwer Academic, 2001.
- C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. VLDB'98.
- P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the Right Interestingness Measure for Association Patterns. KDD'02.
- E. Omiecinski. Alternative Interest Measures for Mining Associations. TKDE'03.
- T. Wu, Y. Chen, and J. Han, "Re-Examination of Interestingness Measures in Pattern Mining: A Unified Framework", Data Mining and Knowledge Discovery, 21(3):371-397, 2010



داده کاوی

مفاهیم و تکنیک ها

— فصل ۸ —

فصل ۸: دسته بندی- مفاهیم پایه



- دسته بندی: مفاهیم پایه
- درخت تصمیم
- روش های دسته بندی بیز
- دسته بندی مبتنی بر قانون
- ارزیابی و انتخاب مدل ها
- روش های بهبود دقت در دسته بندی
- خلاصه

یادگیری بانظارت در مقابل یادگیری بدون نظارت

- **یادگیری با نظارت (دسته بندی)**
 - نظارت: داده های آموزشی (مشاهدات، اندازه گیری ها، و غیره) با برچسب هایی که نشان دهنده دسته مشاهدات است همراه است
 - داده جدید بر اساس داده های آموزشی دسته بندی می شود.
- **یادگیری بدون نظارت (خوش بندی)**
 - برچسب کلاس داده های آموزشی نامشخص است.
 - مقادیری از اندازه گیری ها، مشاهدات و غیره داده شده. هدف بررسی وجود کلاس ها یا خوش های متفاوت در داده ها است.

پیش بینی: دسته بندی در مقابل پیش بینی عددی

■ دسته بندی

- پیش بینی بر چسب کلاس های دسته بندی (گسته یا اسمی)
- دسته بندی داده (ساخت مدل) بر اساس مجموعه داده آموزشی و مقادیر موجود در صفتی که دسته بندی بر اساس مقادیر آن صورت می گیرد و سپس استفاده از این مدل در دسته بندی داده های جدید فاقد مقدار کلاس

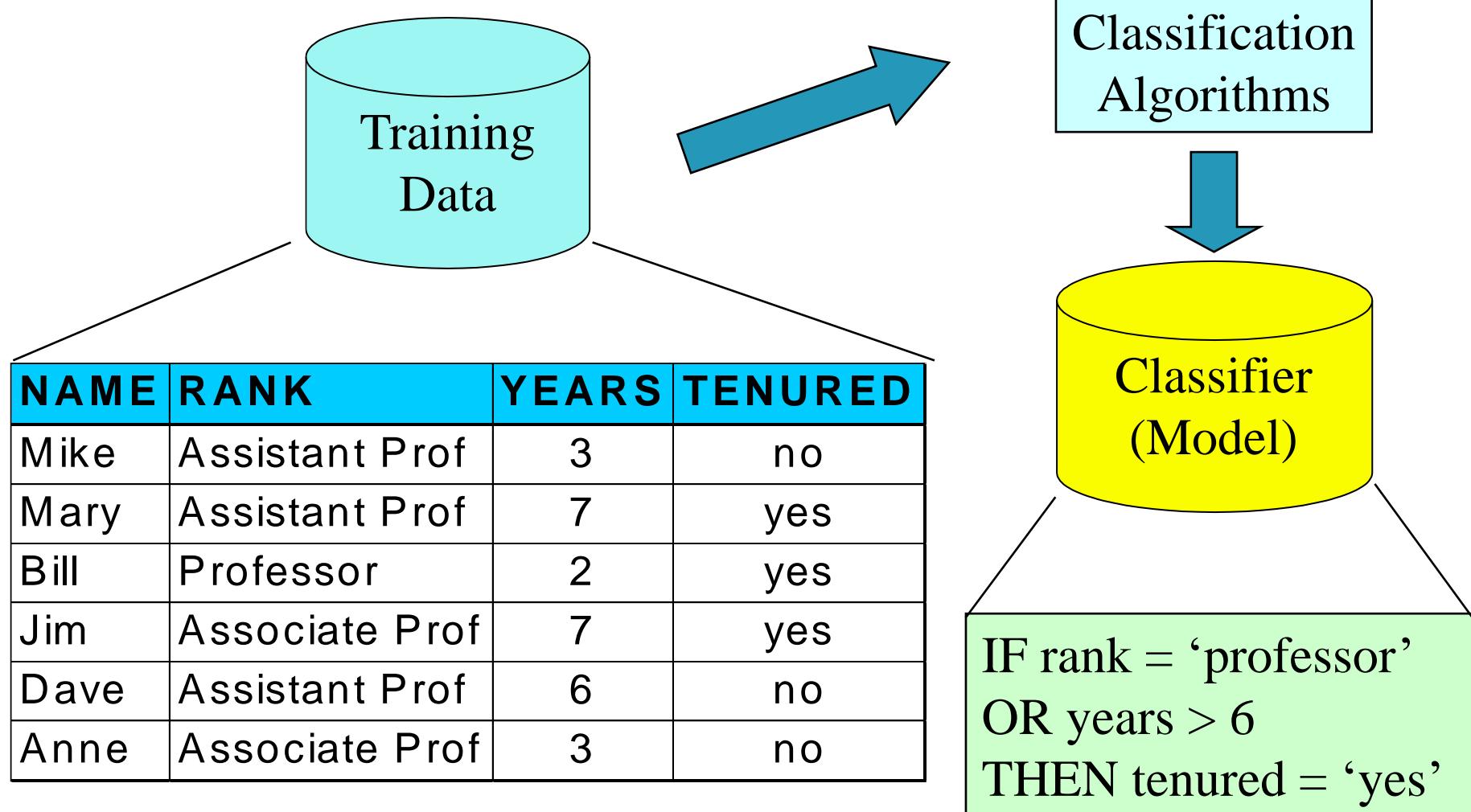
■ پیش بینی عددی

- مدل کردن توابع با مقادیر پیوسته مثلا پیش بینی مقادیر نامعلوم کاربردهای نمونه
- تأیید اعتبار / وام
- تشخیص پزشکی: آیا یک تومور سرطانی است یا خوش خیم؟
- تشخیص تقلب: تراکنش جعلی است یا خیر؟
- طبقه بندی صفحات وب: صفحه در کدام دسته است؟

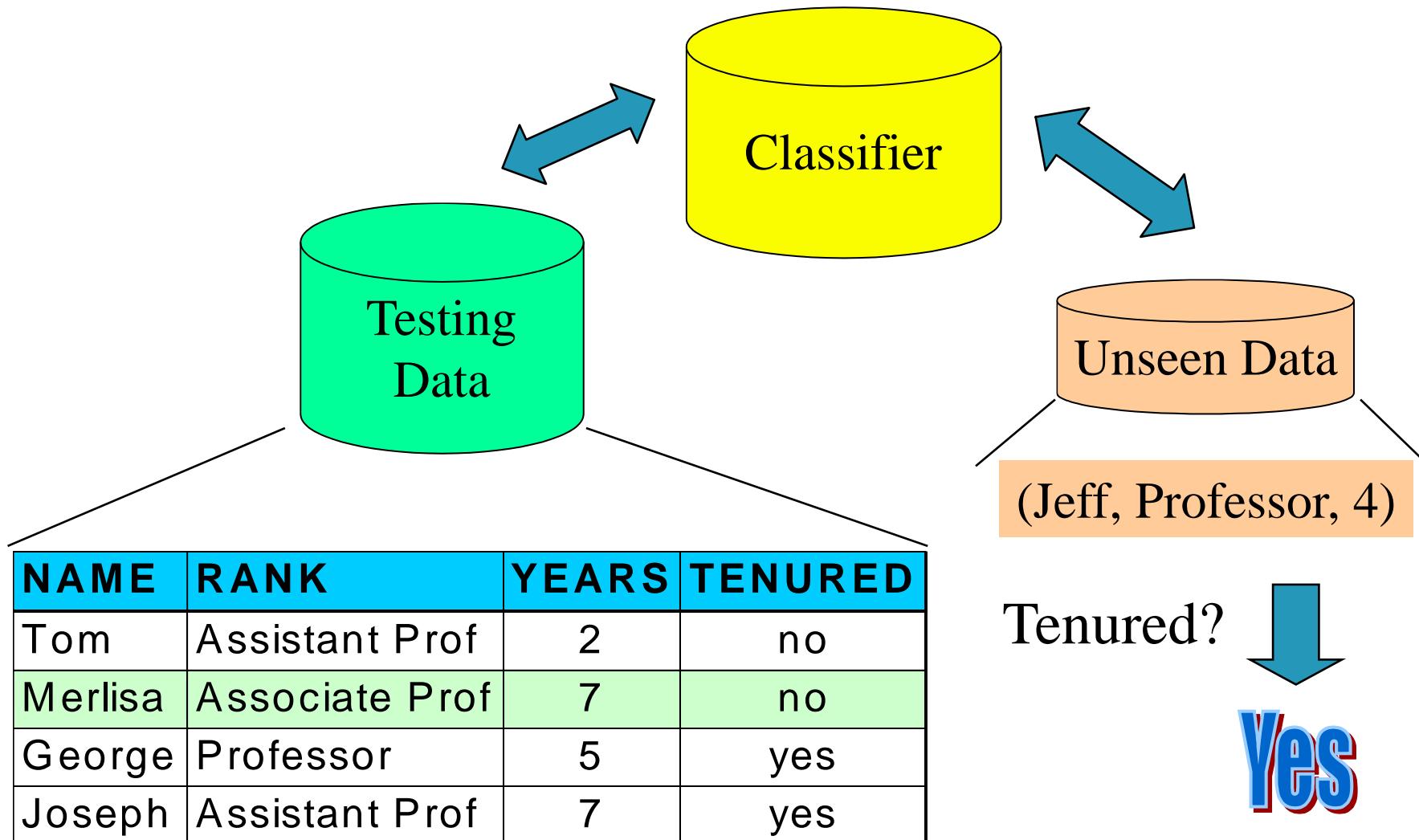
دسته بندی : یک فرایند دو مرحله ای

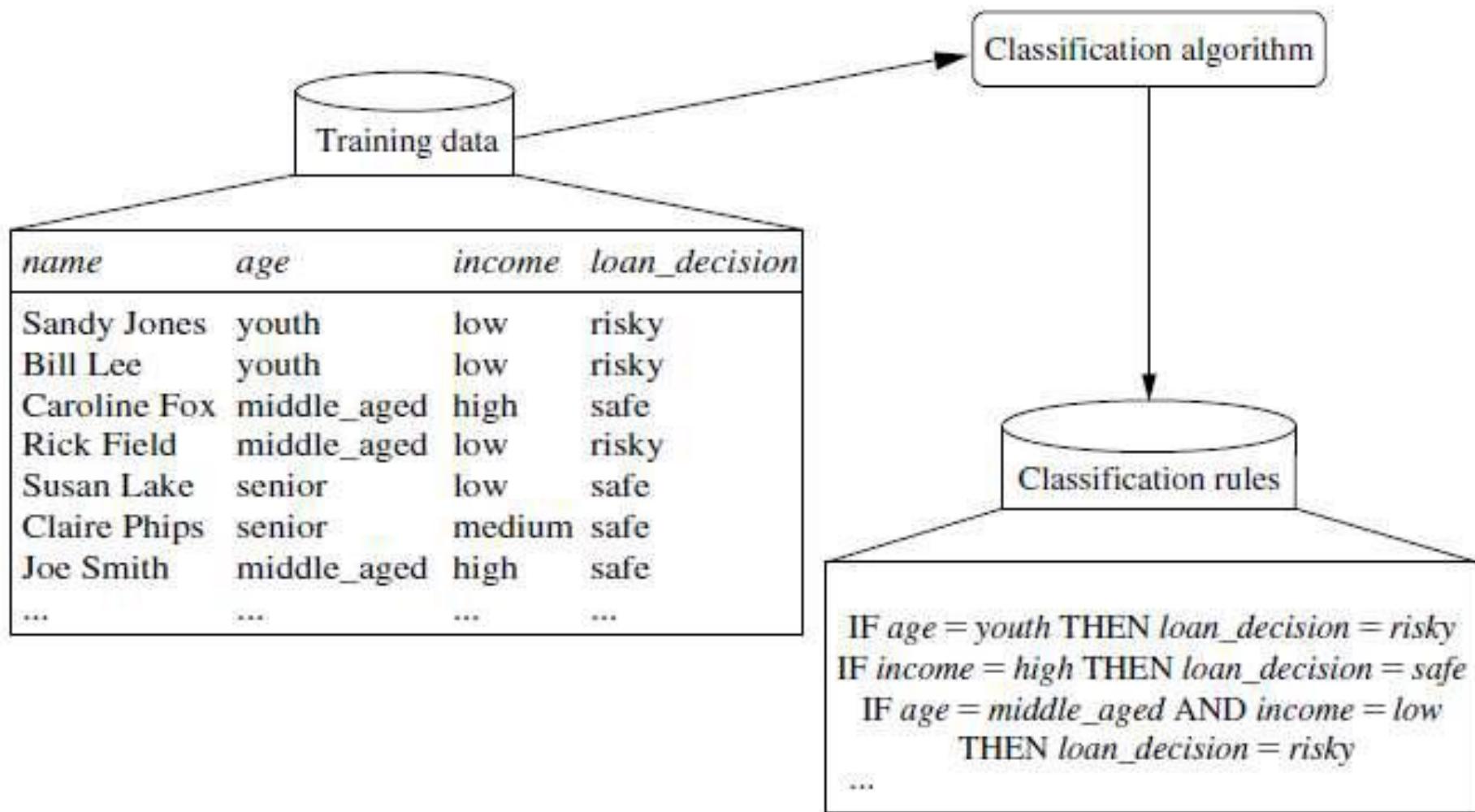
- ساخت مدل: توصیف مجموعه ای از دسته های از پیش تعیین شده
- فرض می شود هر نمونه یا تاپل به یکی از کلاس های از پیش تعریف شده تعلق دارد. (بر اساس صفتی که بعنوان برچسب کلاس تعیین شده است.)
- مجموعه تاپل هایی که برای ساخت مدل بکار می رود مجموعه آموزشی یا **training set** نامیده می شود.
- مدل بصورت مجموعه ای از قوانین دسته بندی، درخت تصمیم یا فرمول های ریاضی نمایش داده می شود.
- استفاده از مدل: برای دسته بندی نمونه های آینده یا با برچسب نامعلوم
- دقت تخمین مدل
 - برچسب معلوم نمونه تست با حاصل کار مدل در مورد آن نمونه مقایسه می شود.
 - دقت درصد نمونه هایی است که برچسب کلاس آن ها به درستی توسط مدل پیش بینی شده است.
 - **مجموعه تست** مستقل از مجموعه آموزشی است.
- اگر دقت مدل قابل قبول بود، مدل برای دسته بندی داده های جدید مورد استفاده قرار می گیرد.
- اگر مجموعه تست برای انتخاب مدل استفاده شود، مجموعه ارزیابی **validation (test) set** نامیده می شود.

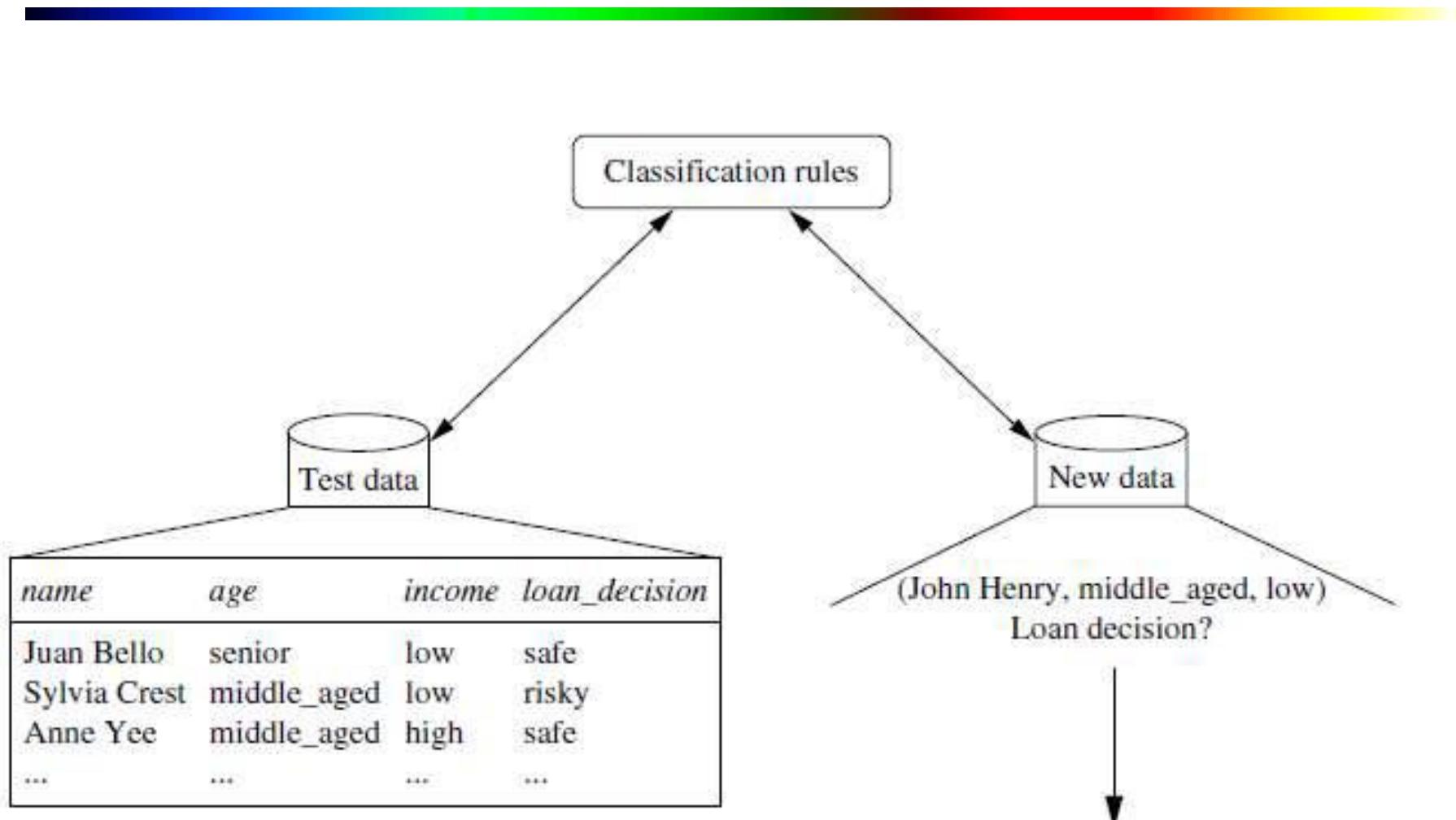
مرحله ۱: ساخت مدل



مرحله ۲: استفاده از مدل در پیش بینی





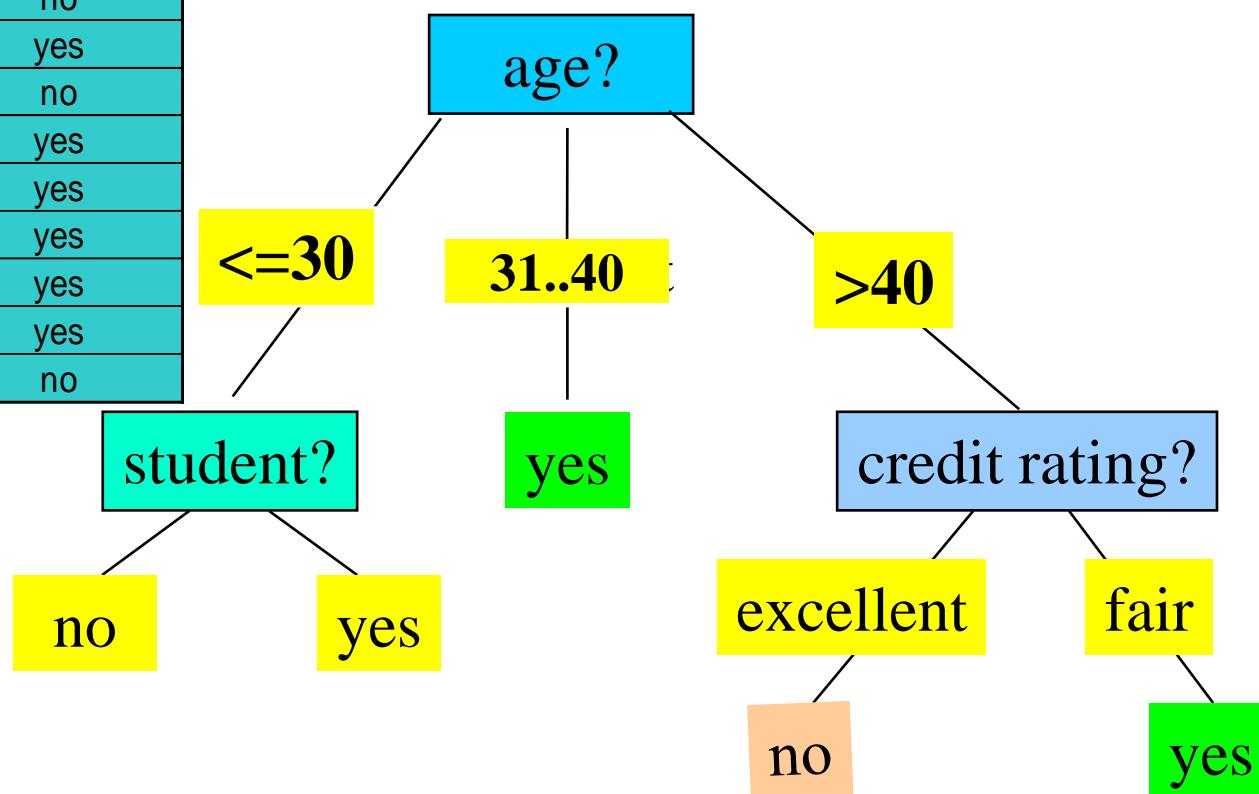


فصل ۸: دسته بندی- مفاهیم پایه

- دسته بندی: مفاهیم پایه
- درخت تصمیم 
- روش های دسته بندی بیز
- دسته بندی مبتنی بر قانون
- ارزیابی و انتخاب مدل ها
- روش های بهبود دقیق در دسته بندی
- خلاصه

درخت تصمیم: یک مثال

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no



الگوریتم استنتاج درخت تصمیم

- الگوریتم پایه (یک الگوریتم حریصانه)
 - درخت بر اساس یک روش بالا به پایین بازگشتی تقسیم و حل بنا شده است.
 - در ابتدا همه نمونه های آموزشی در ریشه هستند.
 - صفات دسته بندی شده هستند. (صفات پیوسته، گستره سازی می شوند.)
 - نمونه ها بصورت بازگشتی بر اساس صفات انتخاب شده تقسیم می شوند.
 - صفات تقسیم بندی در هر سطح بر اساس یک مبنای حریصانه یا اندازه آماری انتخاب می شوند. (مثل **information gain**)
 - شرایط خاتمه تقسیم:
 - همه نمونه های یک نود به یک کلاس مشترک تعلق داشته باشند.
 - صفت باقیمانده دیگری برای تقسیم بندی وجود نداشته باشد: در این حالت برچسب اکثریت نمونه ها، برای دسته بندی بعنوان برچسب برگ استفاده می شود.
 - هیچ نمونه بیشتری وجود نداشته باشد.

مرور مختصری بر آنتروپی

- آنتروپی (تئوری اطلاعات)
 - اندازه عدم قطعیت مناسب به یک متغیر تصادفی
 - محاسبه: برای متغیر تصادفی گسته Σ که m مقدار متمایز (y_1, \dots, y_m) دارد
- $H(Y) = -\sum_{i=1}^m p_i \log(p_i)$, where $p_i = P(Y = y_i)$

■ تفسیر:

آنتروپی بالاتر : عدم قطعیت بیشتر
آنتروپی کمتر : عدم قطعیت کمتر

$m = 2$

معیار انتخاب صفت:

Information Gain (ID3/C4.5)

- انتخاب صفت با بالاترین information gain
- فرض کنید p_i احتمال تعلق یک تاپل در D به کلاس C_i باشد که با $|C_{i,D}|/|D|$ تخمین زده می شود.

■ میزان آنتروپی مورد نیاز برای دسته بندی کردن یک تاپل در D معادل است با:

$$Info(D) = -\sum_{i=1}^m p_i \log_2(p_i)$$

■ میزان Information مورد نیاز بعد از استفاده از صفت A برای تقسیم D به v قسمت برای دسته بندی کردن D معادل است با:

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j)$$

■ با تقسیم بر اساس A Information gained

$$Gain(A) = Info(D) - Info_A(D)$$

Attribute Selection: Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14} \log_2(\frac{9}{14}) - \frac{5}{14} \log_2(\frac{5}{14}) = 0.940$$

age	p_i	n_i	$I(p_i, n_i)$
≤ 30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0)$$

$$+ \frac{5}{14} I(3,2) = 0.694$$

عنی "age ≤ 30 " ۵ نمونه از ۱۴ نمونه را شامل می شود، با دو مقدار yes و no مقدار ۰.۹۷۱ برابر است.

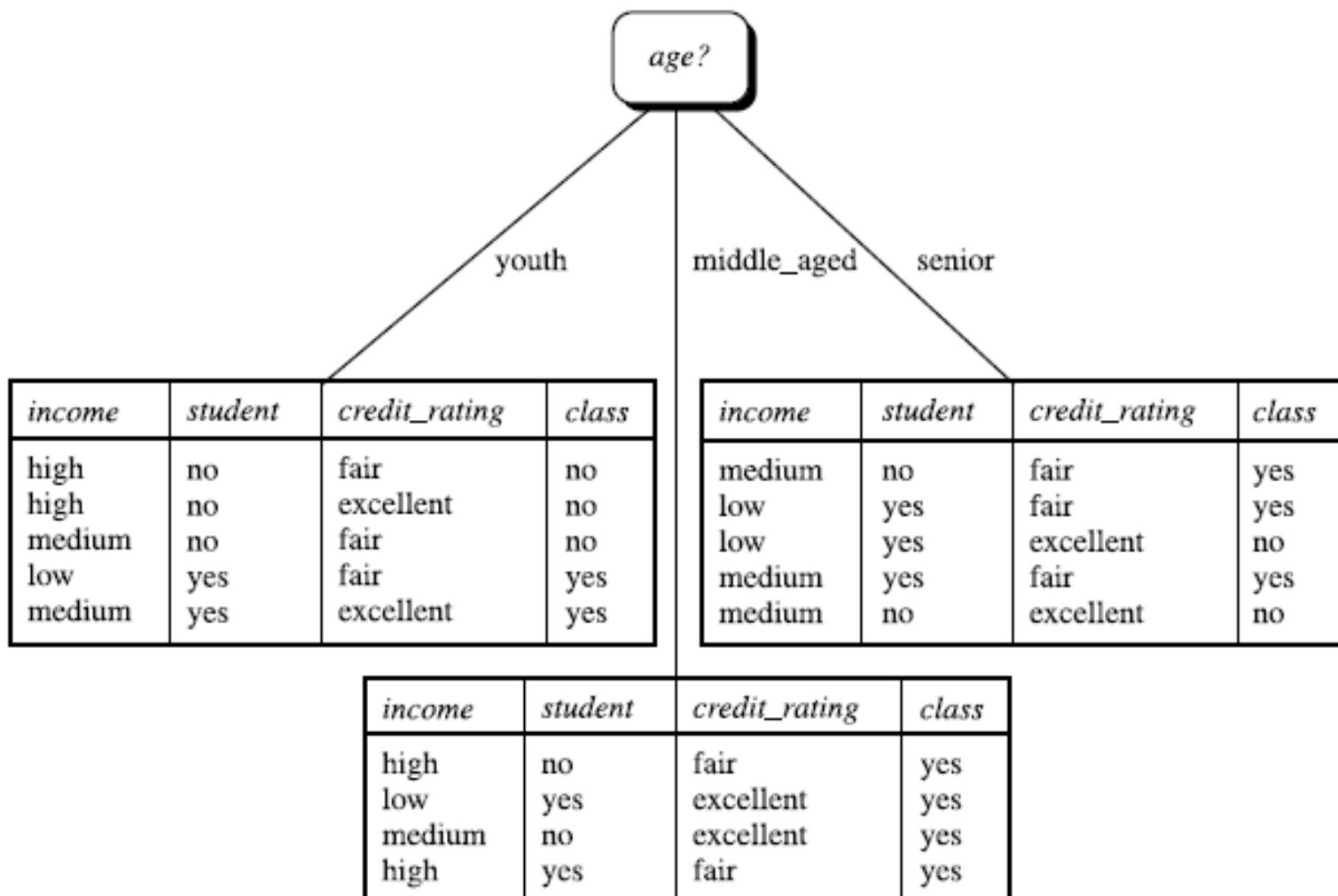
$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

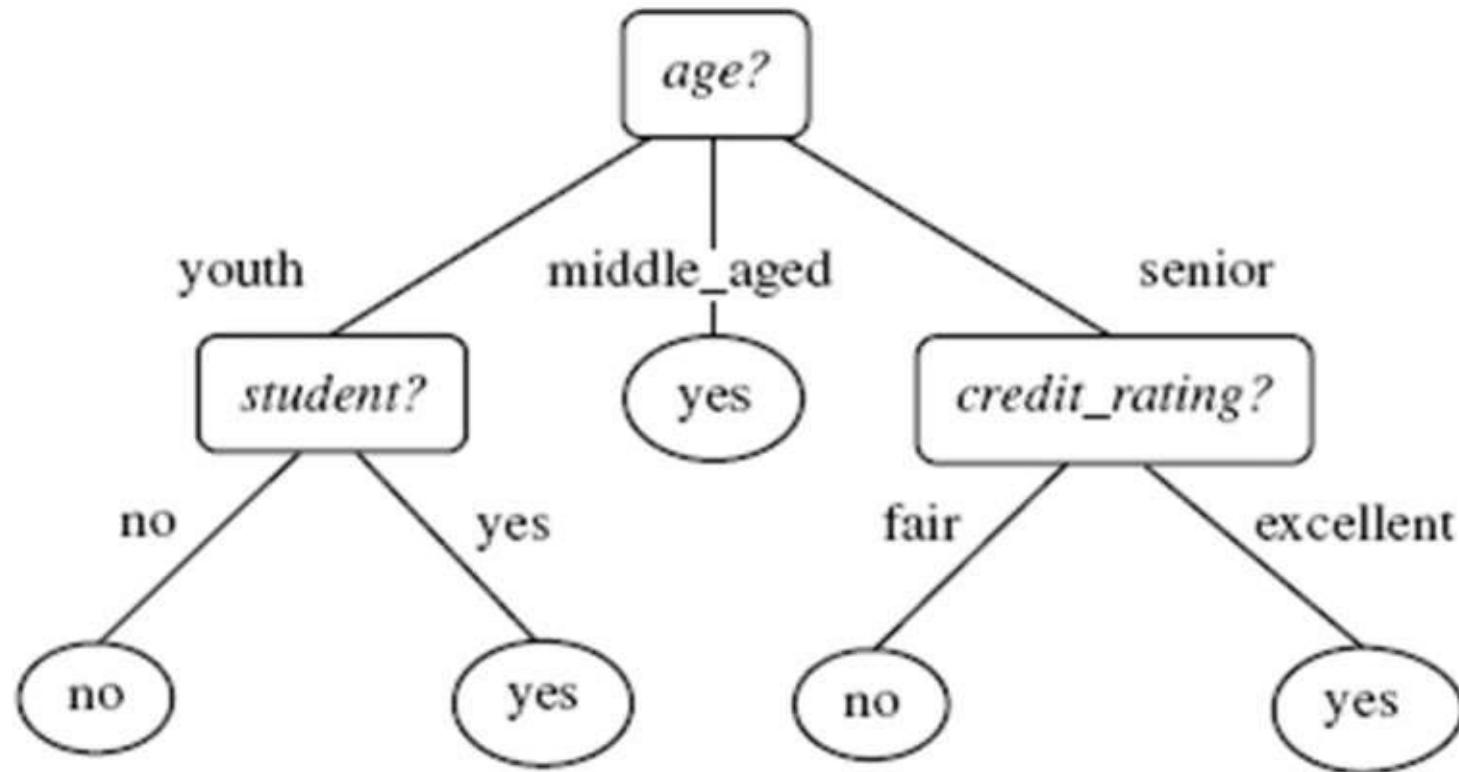
بطور مشابه:

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit_rating) = 0.048$$





محاسبه Information-Gain برای صفات پیوسته

- فرض کنید صفت A یک صفت با مقادیر پیوسته باشد.
- لازم است بهترین نقطه تقسیم یا *best split point* برای A تعیین شود.
 - مقادیر A را به ترتیب صعودی مرتب کنید.
 - هر نقطه بین هر دو مقدار همسایه یک نقطه تقسیم ممکن است.
 - نقطه وسط بین دو مقدار a_i و a_{i+1} است.
 - نقطه ای با کمترین اطلاعات مورد نیاز *minimum expected information requirement* برای A بعنوان نقطه شکست انتخاب می شود.
- تقسیم:
 - $D_1 \leq \text{split-point}$ مجموعه تاپل های با مقدار A $>$ split-point
 - $D_2 \leq \text{split-point}$ مجموعه تاپل های با مقدار A \leq split-point

معیار Gain Ratio برای انتخاب صفت در C4.5

- معیار Information gain به سمت صفاتی با تعداد مقادیر بیشترگرایش دارد.
- الگوریتم C4.5 از معیار gain ratio برای غلبه بر این مشکل استفاده می کند.
- (نرمالسازی information gain)

$$SplitInfo_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

- $\text{GainRatio}(A) = \text{Gain}(A)/\text{SplitInfo}(A)$

: مثال

- $\text{gain_ratio(income)} = 0.029/1.557 = 0.019$

- صفت با بیشترین مقدار gain ratio بعنوان صفت تقسیم استفاده می شود.

$$SplitInfo_{income}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right) = 1.557$$

معیار شاخص Gini (CART, IBM IntelligentMiner)

- اگر یک مجموعه داده D شامل نمونه هایی از n کلاس باشد شاخص gini به صورت زیر تعریف می شود:

$$gini(D) = 1 - \sum_{j=1}^n p_j^2$$

که در آن p_j فرکانس مربوط به کلاس j در D است.

- اگر مجموعه داده D بر اساس صفت A به دو زیرمجموعه D_1 و D_2 تقسیم شود شاخص gini آن به صورت زیر تعریف می شود:

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

: Reduction in Impurity یا کاهش ناخالصی

- صفتی که کمترین $gini_{split}(D)$ (یا بیشترین کاهش ناخالصی) را داشته باشد برای تقسیم انتخاب می شود. (لازم است این محاسبه برای همه نقاط تقسیم برای همه صفات انجام شود.)

$$\Delta gini(A) = gini(D) - gini_A(D)$$

Computation of Gini Index

- Ex. D has 9 tuples in buys_computer = “yes” and 5 in “no”

$$gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

- Suppose the attribute income partitions D into 10 in D_1 : {low, medium} and 4 in D_2

$$gini_{income \in \{low, medium\}}(D) = \left(\frac{10}{14}\right)Gini(D_1) + \left(\frac{4}{14}\right)Gini(D_2)$$

$$= \frac{10}{14} \left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14} \left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$

$$= 0.443$$

$$= Gini_{income \in \{high\}}(D).$$

Gini_{low,high} is 0.458; Gini_{medium,high} is 0.450. Thus, split on the {low,medium} (and {high}) since it has the lowest Gini index

مقایسه معیارهای انتخاب صفت

عموما هر سه معیار نتایج خوبی دارد اما:

Information gain:

به سمت انتخاب صفات با تعداد مقادیر زیاد گرایش دارد.

Gain ratio:

به ایجاد تقسیم های نامتعادل گرایش دارد که در آن یک پارتیشن بسیار کوچکتر از دیگران است.

Gini index:

به سمت انتخاب صفات با تعداد مقادیر زیاد گرایش دارد.

زمانی که تعداد کلاس ها زیاد باشد با مشکل زمانگیر بودن محاسبات مواجه است.

تمایل به حالت هایی دارد که در نتیجه پارتیشن های هم سایز و خلوص در هر دو پارتیشن ایجاد شود.

سایر معیارهای انتخاب صفت:

- CHAID: a popular decision tree algorithm, measure based on χ^2 test for independence
- C-SEP: performs better than info. gain and gini index in certain cases
- G-statistic: has a close approximation to χ^2 distribution
- MDL (Minimal Description Length) principle (i.e., the simplest solution is preferred):
 - The best tree as the one that requires the fewest # of bits to both (1) encode the tree, and (2) encode the exceptions to the tree
- Multivariate splits (partition based on multiple variable combinations)
 - CART: finds multivariate splits based on a linear comb. of attrs.
- Which attribute selection measure is the best?
 - Most give good results, none is significantly superior than others

هرس درخت تصمیم

- Overfitting: درخت حاصل ممکن است نسبت به مجموعه آموزشی ما بزرگ باشد.
- انشعابات زیاد ممکن است در اثر آنومالی های ناشی از نویز یا داده پرت باشد.
- دقیق ضعیف برای دسته بندی داده های جدید
- دو روش برای پرهیز از overfitting
- پیش هرس: توقف در حین ساخت درخت - اگر تقسیم معیار ارزیابی ما را زیر حد آستانه برداشت تقسیم انجام نشود.
- مشکل تعریف حد آستانه خوب
- پس هرس: حذف انشعابات از درختی که زیاد رشد کرده است به نحوی که درخت بهتری حاصل شود.
- برای تعیین اینکه چه هرسی بهتر است از یک مجموعه داده تست مستقل از مجموعه آموزشی استفاده می شود.

نمونه هرس درخت

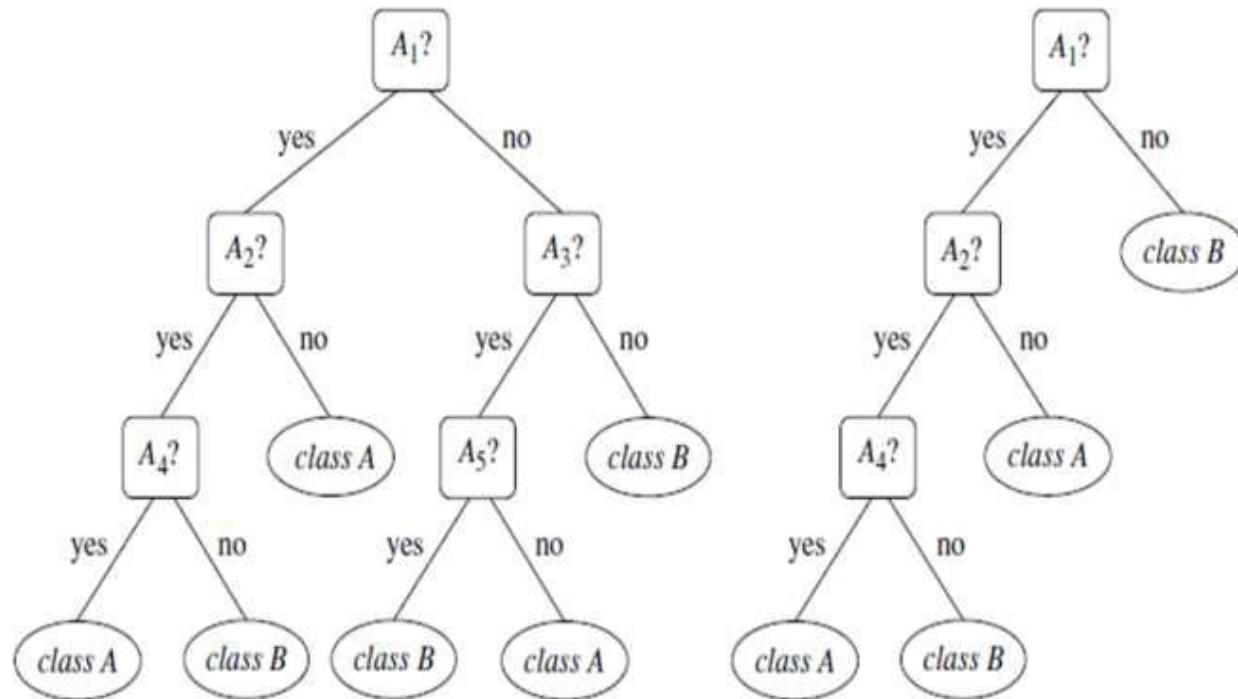


Figure 8.6 An unpruned decision tree and a pruned version of it.

فصل ۸: دسته بندی- مفاهیم پایه



■ دسته بندی: مفاهیم پایه

■ درخت تصمیم



■ روش های دسته بندی بیز

■ دسته بندی مبتنی بر قانون

■ ارزیابی و انتخاب مدل ها

■ روش های بهبود دقت در دسته بندی

■ خلاصه

دسته بندی بیزین : چرا؟

- یک دسته بندی کننده آماری: پیش بینی بر اساس احتمال مثلا پیش بینی احتمال عضویت در یک کلاس
- اساس: تئوری بیز
- کارایی: *naïve Bayesian* که یک دسته بندی کننده ساده بیزین است کارایی قابل مقایسه‌ای با درخت تصمیم و بعضی دسته بندی‌های شبکه عصبی دارد.
- تدریجی: در این روش احتمال درست بودن یک فرضیه با اضافه شدن هر نمونه تجربی به تدریج افزایش یا کاهش می‌یابد. در واقع در هر مرحله دانش قبلی با داده جدید ترکیب می‌شود.
- استاندارد: حتی زمانی که روش‌های بیزین از نظر محاسباتی غیر قابل مدیریت باشند، می‌توانند استانداردی از تصمیم‌گیری بهینه را فراهم کنند که سایر روش‌ها با آن سنجیده شوند.

تئوری بیز: اصول اولیه

$$P(B) = \sum_{i=1}^M P(B|A_i)P(A_i)$$

تئوری مجموع احتمالات :
تئوری بیز :

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H)/P(\mathbf{X})$$

- فرض کنید X یک نمونه داده باشد که کلاس آن ناشناخته است.
- فرض کنید H فرضیه تعلق X به کلاس C باشد.
- دسته بندی باید $P(H|\mathbf{X})$ را تعیین کند : احتمال اینکه فرضیه با وجود داده جدید X درست باشد.
- $(P(H))$ (احتمال پیشین) : احتمال اولیه

- بعنوان مثال، X بدون در نظر گرفتن سن و درآمد و ... کامپیوتر خواهد خرید.
- $P(\mathbf{X})$: احتمال مشاهده نمونه داده
- $P(\mathbf{X}|H)$: احتمال مشاهده نمونه X با توجه به این فرضیه
- مثلاً با فرض اینکه X کامپیوتر خواهد خرید، احتمال اینکه X در بازه سنی ۳۱..۴۰ باشد و درآمد متوسط داشته باشد.

مثال:

سوال: اگر بیماری با علامت خشکی گردن مراجعه کند احتمال ابتلای او به بیماری منژیت چقدر است؟

$P(H | X)$

منژیت در ۵۰٪ مواقع باعث خشکی گردن می‌شود.

$P(X | H)$ (likelihood)

احتمال اینکه یک بیمار منژیت داشته باشد برابر $1/50000$ است.

$P(H)$ (prior probability)

احتمال اینکه یک بیمار خشکی گردن داشته باشد $1/20$ است.

$P(X)$ (evidence)

$$P(H | X) = \frac{P(X|H)P(H)}{P(X)} = P(X|H) \times P(H) / P(X)$$
$$= 0.5 * (1/50000) / (1/20) = 0.0002$$

پیش بینی بر اساس تئوری بیز

- با فرض وجود داده آموزشی X ، احتمال پسین فرضیه H ، $P(H|X)$ از تئوری بیز تبعیت می کند:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} = P(X|H) \times P(H) / P(X)$$

- به صورت غیر رسمی، این موضوع را می توان به شکل زیر دید:
 $\text{posteriori} = \text{likelihood} \times \text{prior/evidence}$
- پیش بینی تعلق X به C_i اگر احتمال $P(C_i|X)$ بین همه موارد $P(C_k|X)$ برای همه K کلاس بیشترین باشد.
- دشواری عملی: این روش نیاز به دانش ابتدایی از بسیاری احتمالات دارد، که شامل هزینه محاسباتی قابل توجهی است.

دسته بندی معادل بیشترین احتمال پسین

- فرض کنید D یک مجموعه آموزشی از تاپل ها و برچسب کلاس مربوطه شان باشد و هر تاپل با یک آرایه n بعدی $(x_1, x_2, \dots, x_n) = \mathbf{X}$ نمایش داده شود.
- فرض کنید m کلاس C_1, C_2, \dots, C_m وجود داشته باشد.
- دسته بندی استخراج ماکزیمم مقدار پسینی است. مثلا بیشترین مقدار $P(C_i | \mathbf{X})$ این می تواند از تئوری بیز استخراج شود:

$$P(C_i | \mathbf{X}) = \frac{P(\mathbf{X} | C_i) P(C_i)}{P(\mathbf{X})}$$

- به دلیل اینکه $P(\mathbf{X})$ برای همه دسته ها ثابت است فقط لازم است مقدار صورت ماکزیمم شود.

$$P(C_i | \mathbf{X}) = P(\mathbf{X} | C_i) P(C_i)$$

دسته بندی کننده Naïve Bayes

- در این نوع دسته بندی یک فرض ساده انجام می شود: صفات مستقل از هم هستند.

$$P(\mathbf{X} | C_i) = \prod_{k=1}^n P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \dots \times P(x_n | C_i)$$

- این روش هزینه محاسباتی را به مقدار زیادی کاهش می دهد.

آموزشی داده مجموعه :Naïve Bayes Classifier

Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

Data to be classified:

X = (age <=30,

Income = medium,

Student = yes

Credit_rating = Fair)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayes Classifier : مثال

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

- $P(C_i)$: $P(\text{buys_computer} = \text{"yes"}) = 9/14 = 0.643$
 $P(\text{buys_computer} = \text{"no"}) = 5/14 = 0.357$
- Compute $P(X|C_i)$ for each class
 - $P(\text{age} = \text{"}<=30\text{"} | \text{buys_computer} = \text{"yes"}) = 2/9 = 0.222$
 - $P(\text{age} = \text{"}<=30\text{"} | \text{buys_computer} = \text{"no"}) = 3/5 = 0.6$
 - $P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"yes"}) = 4/9 = 0.444$
 - $P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.4$
 - $P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$
 - $P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"no"}) = 1/5 = 0.2$
 - $P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$
 - $P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.4$
- X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$$P(X|C_i) : P(X|\text{buys_computer} = \text{"yes"}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$

$$P(X|\text{buys_computer} = \text{"no"}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$$

$$P(X|C_i) * P(C_i) : P(X|\text{buys_computer} = \text{"yes"}) * P(\text{buys_computer} = \text{"yes"}) = 0.028$$

$$P(X|\text{buys_computer} = \text{"no"}) * P(\text{buys_computer} = \text{"no"}) = 0.007$$

Therefore, X belongs to class ("buys_computer = yes")

حل مشکل احتمال صفر

- در الگوریتم دسته بندی Naïve Bayesian لازم است همه احتمالات شرطی غیر صفر باشند در غیر اینصورت احتمال پیش بینی شده صفر خواهد بود.

$$P(X | C_i) = \prod_{k=1}^n P(x_k | C_i)$$

- بعنوان مثال یک مجموعه داده با ۱۰۰۰ تاپل را فرض کنید که در آن

income=low (0), income= medium (990), and income = high (10)

- از تخمین یا تصحیح لاپلاس استفاده کنید:
- اضافه کردن یک رکورد به هر کلاس

$$\text{Prob}(\text{income} = \text{low}) = 1/1003$$

$$\text{Prob}(\text{income} = \text{medium}) = 991/1003$$

$$\text{Prob}(\text{income} = \text{high}) = 11/1003$$

- تخمین های احتمال تصحیح شده نزدیک به حالت تصحیح نشده هستند و مشکل احتمال صفر هم از میان می رود.

نکاتی در مورد Naïve Bayes Classifier

■ مزایا

- پیاده سازی آسان
- نتایج خوب در بیشتر موارد

■ معایب

- فرض مستقل بودن صفات باعث کاهش دقت می شود.
- در عمل بین صفات وابستگی وجود دارد.

■ چگونه وابستگی بین صفات را لاحظ کنیم؟
(Chapter 9)

فصل ۸: دسته بندی- مفاهیم پایه



- دسته بندی: مفاهیم پایه
- درخت تصمیم
- روش های دسته بندی بیز
- دسته بندی مبتنی بر قانون
- ارزیابی و انتخاب مدل ها
- روش های بهبود دقت در دسته بندی
- خلاصه

استفاده از قوانین IF-THEN برای دسته بندی

- دانش را به شکل قوانین IF-THEN نمایش می‌دهد.

R: IF *age* = youth AND *student* = yes THEN
buys_computer = yes

- به بخش اول پیش شرط یا مقدم قانون گفته می‌شود. (Rule antecedent or precondition)
- به بخش دوم نتیجه یا تالی قانون گفته می‌شود. (Rule consequent) ارزیابی یک قانون: پوشش و دقت
- تعداد تاپل هایی که توسط R پوشش داده می‌شوند. n_{covers}
- تعداد تاپل هایی که به درستی توسط R دسته بندی می‌شوند. n_{correct}
- مجموعه داده آموزشی D

$$\text{coverage}(R) = n_{\text{covers}} / |D|$$

$$\text{accuracy}(R) = n_{\text{correct}} / n_{\text{covers}}$$

■ اگر پیش از یک قانون triggered شود نیاز به برطرف کردن تصادم داریم.

■ Size ordering: بیشترین اولویت را به قانونی می دهد که دشوارترین پیش

شرط ها را داشته باشد. (مثلاً بیشترین تعداد پیش شرط)

■ Class-based ordering: قوانین بر اساس کاهش میزان شیوع یا هزینه

اشتباه طبقه بندی در هر کلاس مرتب می شوند.

■ Rule-based ordering (**decision list**): ترتیب دادن قوانین بر

اساس کیفیت قانون یا توسط افراد خبره

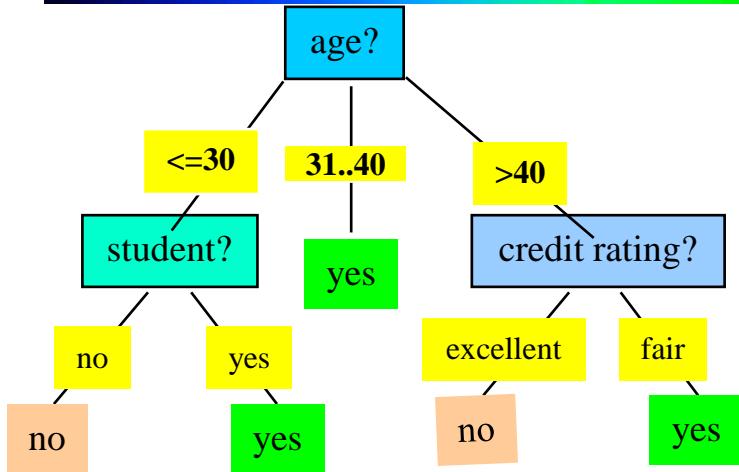
■ اگر هیچ قانون trigger نشود از قانون پیش فرض استفاده می شود. (مثلاً کلاس

ساخر تاپل هایی که هیچ قانونی را ارضانمی کند.)

روش های ایجاد قوانین

- استخراج قوانین از درخت تصمیم
- روش پوشش ترتیبی

استخراج قوانین از درخت تصمیم



قوانين، قابل فهم تر از درختان بزرگ هستند.

برای هر مسیر از ریشه تا برگ یک قانون ساخته می شود.

هر زوج مقدار-خصوصیت در امتداد هر مسیر یک ترکیب عطفی می سازد و برگ نشان دهنده کلاس است.

قوانين متقابلاً منحصر به فرد و جامع هستند. (مشکل تصادم نداریم).

- Example: Rule extraction from our *buys_computer* decision-tree

IF *age* = young AND *student* = no

THEN *buys_computer* = no

IF *age* = young AND *student* = yes

THEN *buys_computer* = yes

IF *age* = mid-age

THEN *buys_computer* = yes

IF *age* = old AND *credit_rating* = excellent

THEN *buys_computer* = no

IF *age* = old AND *credit_rating* = fair

THEN *buys_computer* = yes

استنتاج قوانین: روش پوشش ترتیبی

الگوریتم پوشش ترتیبی: قوانین را بصورت مستقیم از داده آموزشی استخراج می کند.

الگوریتم های پوشش ترتیبی نمونه: FOIL, AQ, CN2, RIPPER

قوانین به ترتیب فراگرفته می شوند. هر قانون برای یک کلاس C به نحوی که بسیاری از تاپل های این کلاس را بپوشاند ولی هیچ تاپلی از کلاس دیگر را پوشش ندهد.

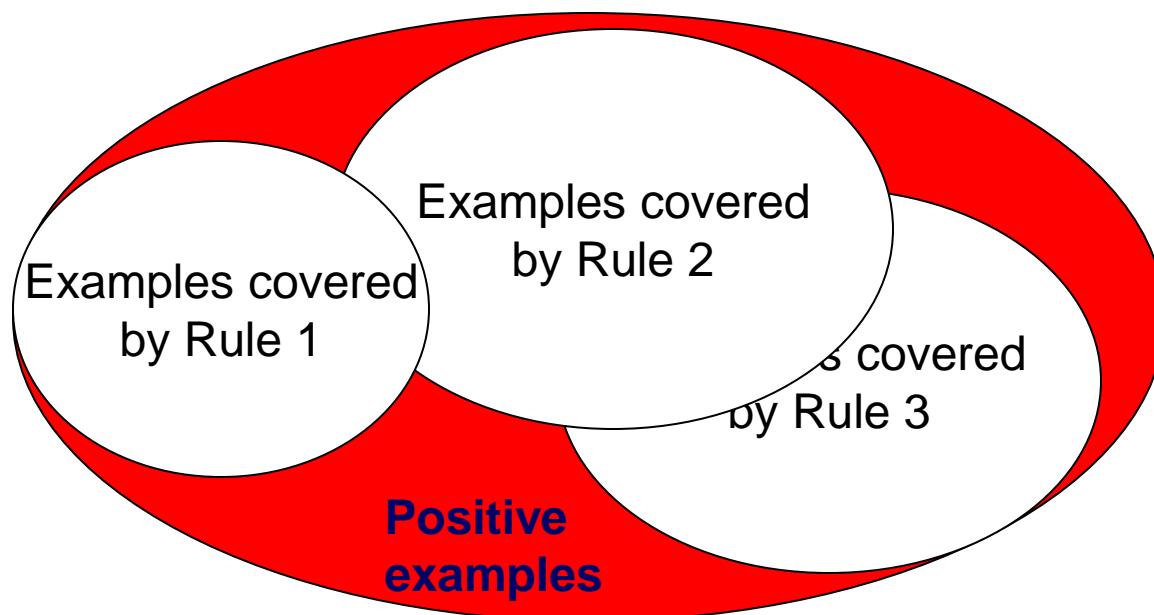
مراحل:

- در هر مرحله یک قانون فرا گرفته می شود.
- هر گاه قانونی فراگرفته شد، داده هایی که توسط آن قانون پوشش داده می شود حذف می شوند.
- فرایند را روی تاپل های باقیمانده تا تحقق شرط خاتمه تکرار کن : مثلا تا زمانی که هیچ تاپلی باقی نماند باشد یا کیفیت قوانین تولید شده زیر آستانه تعریف شده توسط کاربر باشد.

مقایسه با استخراج قوانین از درخت تصمیم: فراگیری یک مجموعه قانون بصورت همزمان

الگوریتم پوشش ترتیبی

while (enough target tuples left)
generate a rule
remove positive target tuples satisfying this rule



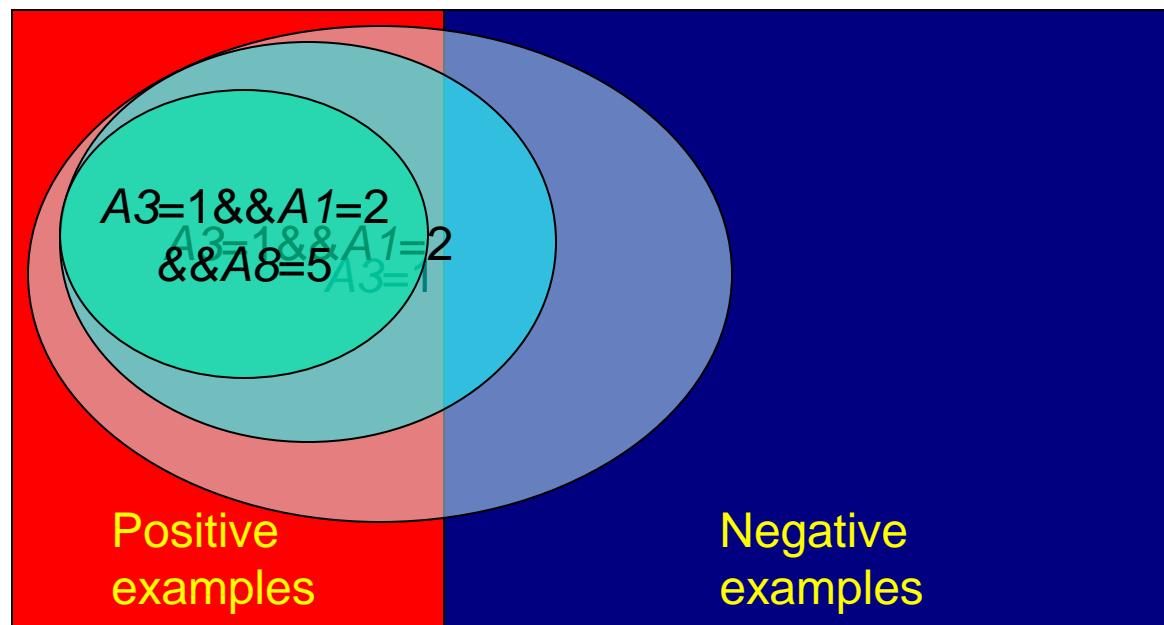
تولید قانون

- To generate a rule

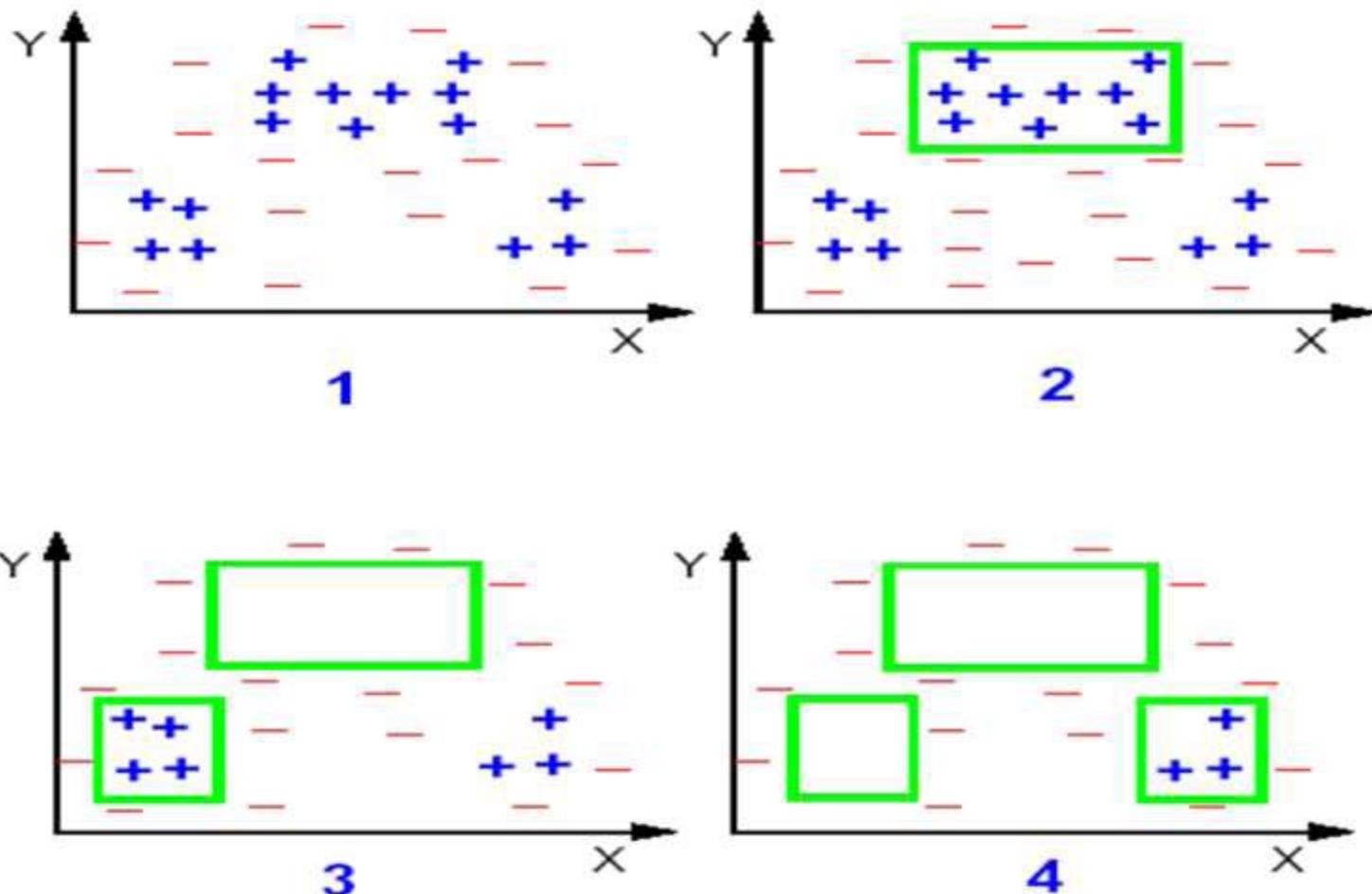
while(true)

find the best predicate p

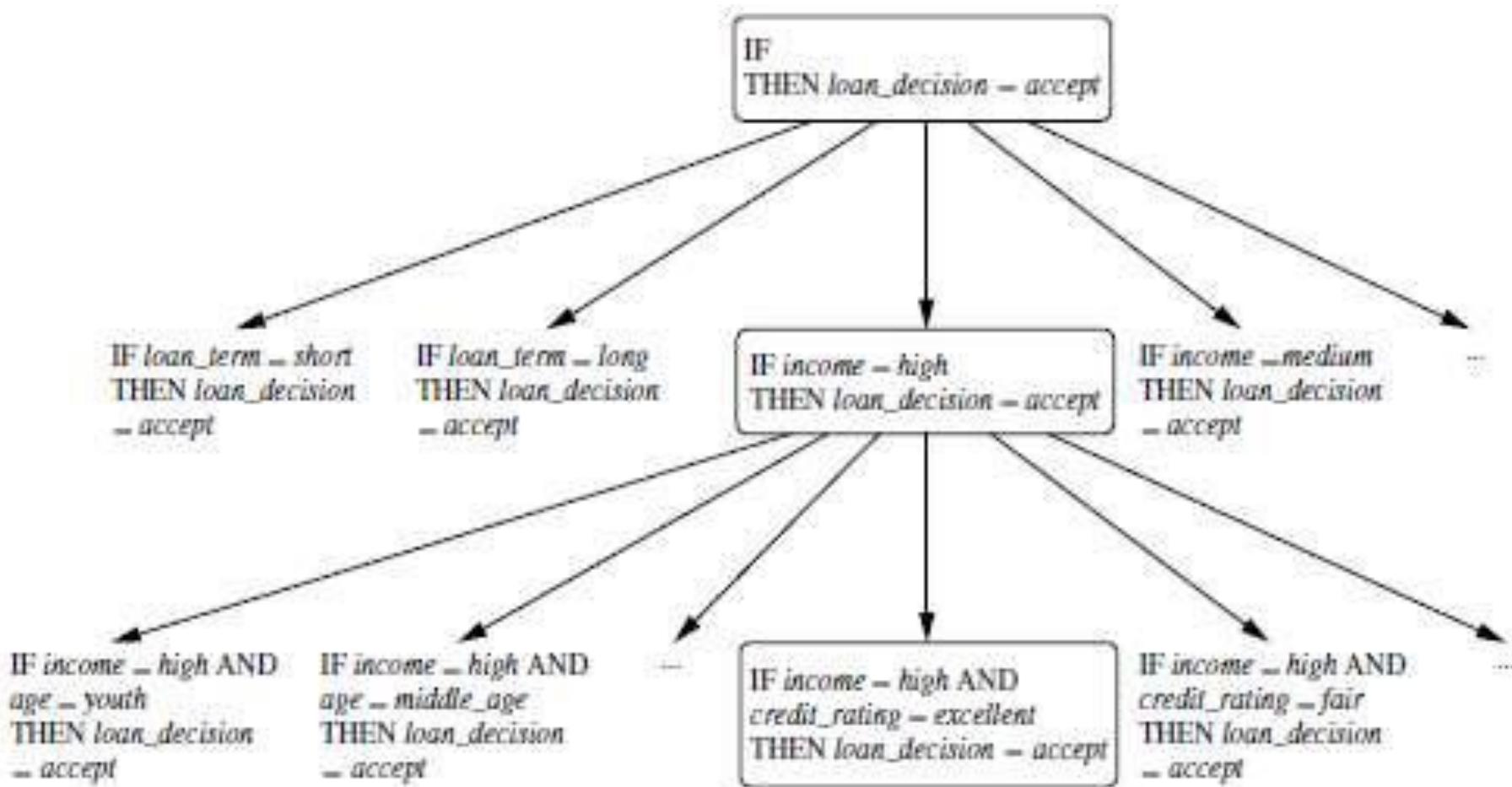
if foil-gain(p) > threshold **then** add p to current rule
else break



الگوریتم پوشش ترتیبی



نمایی از الگوریتم

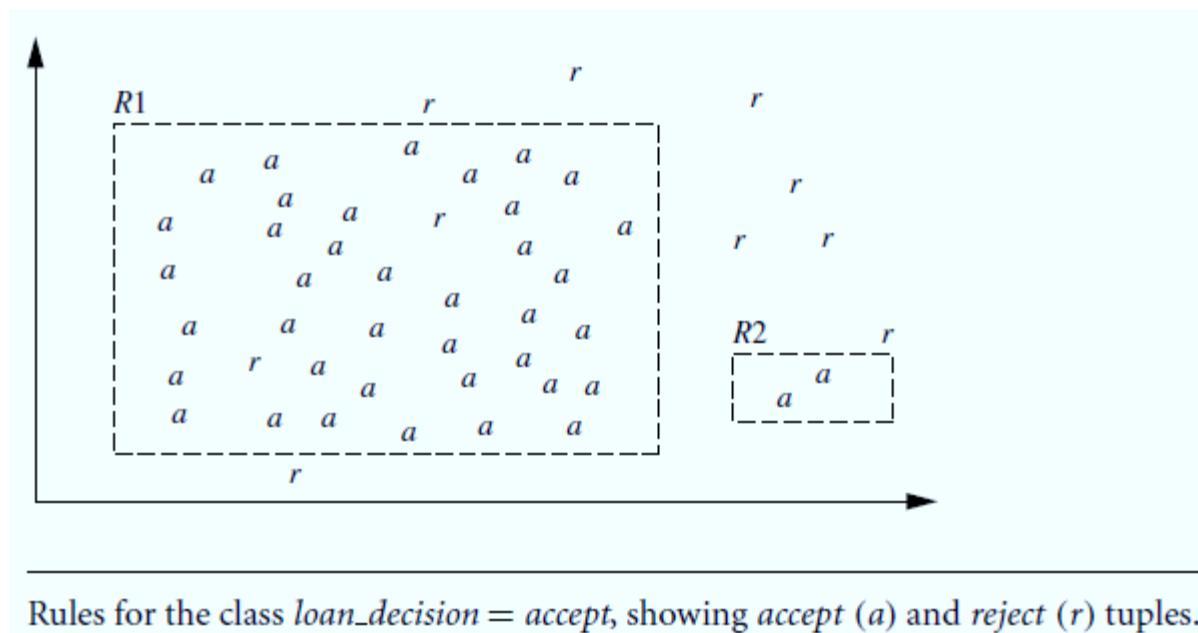


چگونه یک قانون فراگرفته می شود؟

- با کلی ترین قانون شروع می کنیم: condition = empty
- صفات جدید با اتخاذ یک روش حریصانه عمق اول اضافه می شود.
- یکی از مواردی که کیفیت قانون را بهبود می بخشد، انتخاب می شود.
- مقررات کیفیت قانون: هر دو پوشش و دقت را در نظر بگیرید
$$FOIL_Gain = pos' \times (\log_2 \frac{pos'}{pos' + neg'} - \log_2 \frac{pos}{pos + neg})$$
- info_gain : با افزایش شرایط، Foil-gain (in FOIL & RIPPER) ارزیابی می کند
- قوانینی مطلوبند که دقت بالا داشته باشند و تاپل های زیادی را پوشش دهند.

مشکل معیار دقیقت بدون در نظر گرفتن پوشش

- قانون ۱
۴۰ مورد را پوشش داده است، ۳۸ مورد درست است.
- قانون ۲
۲ مورد را پوشش داده است، هر دو مورد درست است.



هرس قوانین

■ هرس قوانین بر اساس یک مجموعه تاپل تست مستقل انجام می شود

$$FOIL_Prune(R) = \frac{pos - neg}{pos + neg}$$

تعداد تاپل های مثبت و منفی پوشش داده شده توسط قانون R Pos/neg هستند.

اگر $FOIL_Prune$ در نسخه هرس شده بیشتر از هرس نشده باشد قوانین هرس می شوند.

فصل ۸: دسته بندی- مفاهیم پایه

- دسته بندی: مفاهیم پایه
- درخت تصمیم
- روش های دسته بندی بیز
- دسته بندی مبتنی بر قانون
- ارزیابی و انتخاب مدل ها
- روش های بهبود دقت در دسته بندی
- خلاصه



ارزیابی و انتخاب مدل ها

- معیارهای ارزیابی: چگونه دقت را اندازه گیری کنیم؟ چه معیارهای دیگری باید وجود دارد؟
- برای اندازه گیری دقت از مجموعه تست ارزیابی بجای مجموعه آموزش استفاده کنید.
- روش های تخمین دقت یک دسته بندی کننده
 - Holdout method, random subsampling
 - Cross-validation
 - Bootstrap
- مقایسه دسته بندی کننده ها:
 - Confidence intervals
 - Cost-benefit analysis and ROC Curves

معیارهای سنجش دسته بندی : Confusion Matrix

Confusion Matrix:

Actual class\Predicted class	C_1	$\neg C_1$
C_1	True Positives (TP)	False Negatives (FN)
$\neg C_1$	False Positives (FP)	True Negatives (TN)

Example of Confusion Matrix:

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	6954	46	7000
buy_computer = no	412	2588	3000
Total	7366	2634	10000

با داشتن m کلاس، سلول $CM_{i,j}$ در confusion matrix تعیین کننده تعداد تاپل های کلاس i که توسط دسته بندی کننده برچسب کلاس j به آن ها داده شده است.

ماتریس می تواند سطروستون های اضافه برای درج مجموع مقادیر داشته باشد.

معیارهای ارزیابی طبقه بندی: دقت، نرخ خطأ، حساسیت و اختصاصی بودن

A\P	C	-C	
C	TP	FN	P
-C	FP	TN	N
	P'	N'	All

درصد تاپل هایی که به درستی دسته بندی شده اند.

$$\text{Accuracy} = (\text{TP} + \text{TN})/\text{All}$$

$$1 - \text{accuracy}, \text{ or: Error rate}$$

$$\text{Error rate} = (\text{FP} + \text{FN})/\text{All}$$

مشکل کلاس های غیر متعادل:

یک کلاس ممکن است نادر باشد مثل قلب یا HIV مثبت

اکثریت قابل توجهی از کلاس منفی و اقلیت از کلاس مثبت هستند.

نرخ تشخیص مثبت واقعی: **Sensitivity**

$$\text{Sensitivity} = \text{TP}/\text{P}$$

نرخ تشخیص منفی واقعی: **Specificity**

$$\text{Specificity} = \text{TN}/\text{N}$$

معیارهای ارزیابی دسته بندی:

Precision, Recall, F-measures

▪ **Precision**: چند درصد از داده هایی که در دسته مثبت دسته بندی شده اند واقعاً مثبت هستند:

$$precision = \frac{TP}{TP + FP}$$

▪ **Recall**: چند درصد از تاپل های مثبت با برچسب مثبت طبقه بندی شده اند.

$$recall = \frac{TP}{TP + FN}$$

▪ امتیاز کامل ۱ است.

▪ **Fmeasure (F₁ or F-score)**: میانگین هارمونیک بین precision و recall

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

▪ **F_β**: معیار وزنی precision و recall که ابعاد و وزن نکا، مه، ند.

$$F_{\beta} = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

▪ اگر β یک باشد معا

معیارهای ارزیابی دسته بندی: مثال

Actual Class\Predicted class	cancer = yes	cancer = no	Total	Recognition(%)
cancer = yes	90	210	300	30.00 (<i>sensitivity</i>)
cancer = no	140	9560	9700	98.56 (<i>specificity</i>)
Total	230	9770	10000	96.40 (<i>accuracy</i>)

- $Precision = 90/230 = 39.13\%$ $Recall = 90/300 = 30.00\%$

محاسبه دقت دسته بندی:

Holdout & Cross-Validation Methods

Holdout method

- داده بصورت تصادفی به دو مجموعه مستقل تقسیم می شود.
 - 2/3 برای داده آموزشی
 - 1/3 برای تست یا تخمین دقت
- تغییر یافته روش بالا: Random sampling
- Holdout را k بار تکرار کنید و میانگین نرخ دقت را محاسبه کنید.

Cross-validation (k -fold, where $k = 10$ is most popular)

- داده ها را بصورت تصادفی به k زیر مجموعه متقابلاً منحصر بفرد با اندازه یکسان تقسیم کنید.

در تکرار ام از بخش D_i به عنوان مجموعه تست و بقیه را بعنوان مجموعه آموزش استفاده کنید.

■ روشن بالا، k برابر تعداد تاپل های مجموعه (برای مجموعه داده های کوچک)

■ **Stratified cross-validation**: ممکن است توزیع کلاس ها نامتوازن باشد. پارتیشن ها به شکلی تقسیم می شوند که نسبت توزیع کلاس ها حفظ شود.

محاسبه دقت دسته بندی: Bootstrap

Bootstrap

- با مجموعه داده های کوچک خوب کار می کند.
- نمونه ها با استفاده از روش نمونه گیری با جایگذاری برای آموزش انتخاب می شوند.
- بنابراین هر داده ممکن است چند بار برای آموزش انتخاب شود.
- روش های مختلفی ارائه شده که bootstrap 632. رایج ترین آنهاست:
- در این روش اگر مجموعه داده دارای d تاپل باشد d بار نمونه گیری می کنیم. ممکن است تاپلی تکرار شود. اما از لحاظ آماری ثابت می شود که ۶۳٪ از داده ها در این پروسه انتخاب خواهند شد. بقیه نمونه ها که انتخاب نشده اند $(since (1 - 1/d)^d \approx e^{-1})$ ۰.۳۶۸ برای تست استفاده می شوند.
- نمونه گیری k مرتبه انجام می شود و در هر بار مجموعه bootstrap متفاوت خواهد بود.
- نهایتاً برای محاسبه ارزیابی از رابطه زیر استفاده می شود.

$$Acc(M) = \frac{1}{k} \sum_{i=1}^k (0.632 \times Acc(M_i)_{test_set} + 0.368 \times Acc(M_i)_{train_set})$$

مقایسه دو مدل

- فرض کنید دو دسته بندی کننده M_1 و M_2 را داریم. کدام بهترند؟
- استفاده از معیار خطا برای مقایسه با استفاده از روشی مثل k-fold می‌تواند غیر دقیق باشد و خطای زیاد در بعضی حالت‌ها را به علت محاسبه میانگین نشان ندهد.
- ممکن است میزان تفاوت خطای بدست آمده بین دو روش کم باشد یا وابسته به شانس باشد.
- روش‌های مورد استفاده:

- فاصله اطمینان (confidence intervals)
- تجزیه و تحلیل هزینه-سود (cost – benefit)
- منحنی ROC

معیارهای انتخاب مدل

دقت

■ پیش بینی درست برچسب کلاس ها

سرعت

■ زمان ساخت مدل (زمان آموزش)

■ زمان استفاده از مدل (زمان لازم برای دسته بندی یا پیش بینی)

■ مقاومت مدل: حل مسئله نویز یا داده های مفقود

■ مقیاس پذیری: کارایی در مجموعه داده های بزرگ که در دیسک قرار دارد.

■ قابلیت تفسیر

■ درک و بینش ارائه شده توسط مدل

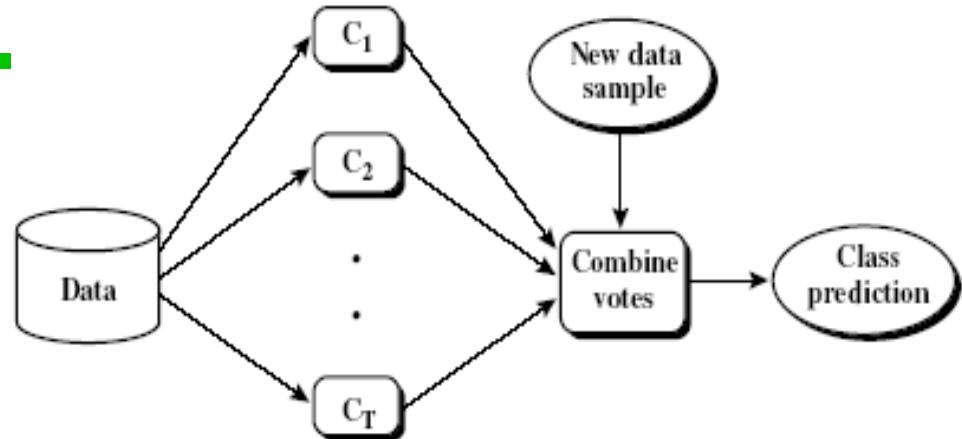
■ سایر معیارها نظیر خوب بودن قوانین، سایز درخت تصمیم یا فشردگی قوانین دسته بندی

فصل ۸: دسته بندی- مفاهیم پایه



- دسته بندی: مفاهیم پایه
- درخت تصمیم
- روش های دسته بندی بیز
- دسته بندی مبتنی بر قانون
- ارزیابی و انتخاب مدل ها
- روش های بهبود دقت در دسته بندی
- خلاصه

روش های گروهی: افزایش دقت



- روش های گروهی
- استفاده از ترکیب چند مدل برای افزایش دقت
- ترکیب k مدل آموزش دیده M_1, M_2, \dots, M_k با هدف ایجاد مدل بهتر M^*
- مدل های ترکیبی رایج
 - میانگین پیش بینی های تعدادی دسته بندی کننده: Bagging
 - آراء وزن دار تعدادی دسته بندی کننده: Boosting
 - ترکیب مجموعه ای از طبقه بندی های ناهمگن: Ensemble

فصل ۸: دسته بندی- مفاهیم پایه

- دسته بندی: مفاهیم پایه
- درخت تصمیم
- روش های دسته بندی بیز
- دسته بندی مبتنی بر قانون
- ارزیابی و انتخاب مدل ها
- روش های بهبود دقیق در دسته بندی
- خلاصه



Summary (I)

- Classification is a form of data analysis that extracts models describing important data classes.
- Effective and scalable methods have been developed for decision tree induction, Naive Bayesian classification, rule-based classification, and many other classification methods.
- Evaluation metrics include: accuracy, sensitivity, specificity, precision, recall, F measure, and F_β measure.
- Stratified k-fold cross-validation is recommended for accuracy estimation. Bagging and boosting can be used to increase overall accuracy by learning and combining a series of individual models.

Summary (II)

- Significance tests and ROC curves are useful for model selection.
- There have been numerous comparisons of the different classification methods; the matter remains a research topic
- No single method has been found to be superior over all others for all data sets
- Issues such as accuracy, training time, robustness, scalability, and interpretability must be considered and can involve trade-offs, further complicating the quest for an overall superior method

References (1)

- C. Apte and S. Weiss. **Data mining with decision trees and decision rules**. Future Generation Computer Systems, 13, 1997
- C. M. Bishop, **Neural Networks for Pattern Recognition**. Oxford University Press, 1995
- L. Breiman, J. Friedman, R. Olshen, and C. Stone. **Classification and Regression Trees**. Wadsworth International Group, 1984
- C. J. C. Burges. **A Tutorial on Support Vector Machines for Pattern Recognition**. *Data Mining and Knowledge Discovery*, 2(2): 121-168, 1998
- P. K. Chan and S. J. Stolfo. **Learning arbiter and combiner trees from partitioned data for scaling machine learning**. KDD'95
- H. Cheng, X. Yan, J. Han, and C.-W. Hsu, **Discriminative Frequent Pattern Analysis for Effective Classification**, ICDE'07
- H. Cheng, X. Yan, J. Han, and P. S. Yu, **Direct Discriminative Pattern Mining for Effective Classification**, ICDE'08
- W. Cohen. **Fast effective rule induction**. ICML'95
- G. Cong, K.-L. Tan, A. K. H. Tung, and X. Xu. **Mining top-k covering rule groups for gene expression data**. SIGMOD'05

References (2)

- A. J. Dobson. **An Introduction to Generalized Linear Models**. Chapman & Hall, 1990.
- G. Dong and J. Li. **Efficient mining of emerging patterns: Discovering trends and differences**. KDD'99.
- R. O. Duda, P. E. Hart, and D. G. Stork. **Pattern Classification**, 2ed. John Wiley, 2001
- U. M. Fayyad. **Branching on attribute values in decision tree generation**. AAAI'94.
- Y. Freund and R. E. Schapire. **A decision-theoretic generalization of on-line learning and an application to boosting**. J. Computer and System Sciences, 1997.
- J. Gehrke, R. Ramakrishnan, and V. Ganti. **Rainforest: A framework for fast decision tree construction of large datasets**. VLDB'98.
- J. Gehrke, V. Gant, R. Ramakrishnan, and W.-Y. Loh, **BOAT -- Optimistic Decision Tree Construction**. SIGMOD'99.
- T. Hastie, R. Tibshirani, and J. Friedman. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer-Verlag, 2001.
- D. Heckerman, D. Geiger, and D. M. Chickering. **Learning Bayesian networks: The combination of knowledge and statistical data**. Machine Learning, 1995.
- W. Li, J. Han, and J. Pei, **CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules**, ICDM'01.

References (3)

- T.-S. Lim, W.-Y. Loh, and Y.-S. Shih. **A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms.** Machine Learning, 2000.
- J. Magidson. **The Chaid approach to segmentation modeling: Chi-squared automatic interaction detection.** In R. P. Bagozzi, editor, Advanced Methods of Marketing Research, Blackwell Business, 1994.
- M. Mehta, R. Agrawal, and J. Rissanen. **SLIQ : A fast scalable classifier for data mining.** EDBT'96.
- T. M. Mitchell. **Machine Learning.** McGraw Hill, 1997.
- S. K. Murthy, **Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey**, Data Mining and Knowledge Discovery 2(4): 345-389, 1998
- J. R. Quinlan. **Induction of decision trees.** *Machine Learning*, 1:81-106, 1986.
- J. R. Quinlan and R. M. Cameron-Jones. **FOIL: A midterm report.** ECML'93.
- J. R. Quinlan. **C4.5: Programs for Machine Learning.** Morgan Kaufmann, 1993.
- J. R. Quinlan. **Bagging, boosting, and c4.5.** AAAI'96.

References (4)

- R. Rastogi and K. Shim. **Public: A decision tree classifier that integrates building and pruning.** VLDB'98.
- J. Shafer, R. Agrawal, and M. Mehta. **SPRINT : A scalable parallel classifier for data mining.** VLDB'96.
- J. W. Shavlik and T. G. Dietterich. **Readings in Machine Learning.** Morgan Kaufmann, 1990.
- P. Tan, M. Steinbach, and V. Kumar. **Introduction to Data Mining.** Addison Wesley, 2005.
- S. M. Weiss and C. A. Kulikowski. **Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.** Morgan Kaufman, 1991.
- S. M. Weiss and N. Indurkha. **Predictive Data Mining.** Morgan Kaufmann, 1997.
- I. H. Witten and E. Frank. **Data Mining: Practical Machine Learning Tools and Techniques**, 2ed. Morgan Kaufmann, 2005.
- X. Yin and J. Han. **CPAR: Classification based on predictive association rules.** SDM'03
- H. Yu, J. Yang, and J. Han. **Classifying large data sets using SVM with hierarchical clusters.** KDD'03.

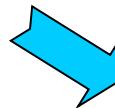


داده کاوی

مفاهیم و تکنیک ها

— فصل ۱۰ —

فصل ۱۰: خوشه بندی - مفاهیم و روش های پایه



- تحلیل خوشه ای : مفاهیم پایه
- روش های بخش بندی: (Partitioning Methods)
- روش های سلسله مراتبی: (Hierarchical Methods)
- روش های مبتنی بر چگالی : (Density-Based Methods)
- روش های مبتنی بر گردید : (Grid-Based Methods)
- ارزیابی روش های خوشه بندی
- خلاصه

تحلیل خوشه ای چیست؟

- خوشه: مجموعه ای از اشیاء داده ای
- مشابه (یا مربوط) به یکدیگر در یک گروه واحد
- متفاوت (یا نامربط) با اشیاء موجود در سایر گروه ها
- تحلیل خوشه (یا خوشه بندی، بخش بندی داده ها,...)
- پیدا کردن مشابهت ها بین داده ها با توجه به ویژگی های آن و گروه بندی داده های مشابه در یک خوشه
- آموزش بدون نظارت: کلاس های از پیش تعریف شده ای وجود ندارد. (یادگیری با مشاهده، در مقابل یادگیری با مثال ها یا یادگیری با نظارت)
 - کاربردهای نمونه
- به عنوان یک ابزار مستقل برای بدست آوردن دیدی نسبت به توزیع داده ها
- به عنوان یک گام پیش پردازش برای سایر الگوریتم ها

خوشه بندی برای فهم داده ها و کاربردها

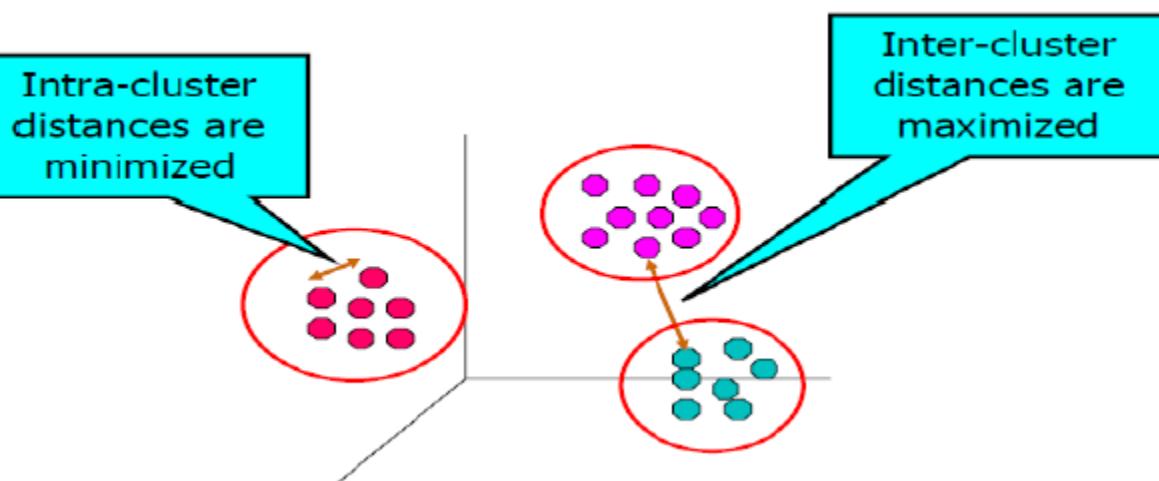
- زیست شناسی: طبقه بندی موجودات زنده: قلمرو، شاخه، طبقه، نظم، خانواده، جنس و گونه
- بازیابی اطلاعات: خوشه بندی اسناد
- کاربردهای مربوط به زمین: شناسایی نواحی با کاربری یکسان در یک پایگاه داده رصد زمین
- بازاریابی: کمک به بازاریابان در شناخت گروه های متمایز در میان مشتریان، و سپس استفاده از این دانش در برنامه های بازاریابی هدفمند
- برنامه ریزی شهری: شناسایی گروه های مسکن بر اساس نوع، ارزش و موقعیت جغرافیایی آن ها
- مطالعات مربوط به زمین لرزه ها: مراکز زمین لرزه های مشاهده شده می توانند بعنوان گسل های قاره ای خوشه بندی شوند.
- آب و هوا: بدست آوردن درک درست از آب و هوا، پیدا کردن الگوهی جوی و اقیانوسی
- علوم اقتصادی: تحقیقات بازار

خوشه بندی بعنوان ابزار پیش پردازش

- خلاصه سازی
- پیش پردازش برای رگرسیون، PCA، دسته بندی، قوانین انجمنی، ...
- فشرده سازی:
- پردازش تصویر
- پیدا کردن K نزدیکترین همسایه
- محلی کردن جستجو در یک یا تعداد کمی از خوشه ها
- شناخت داده های پرت
- داده های پرت عموما به عنوان داده هایی که از سایر خوشه ها خیلی دورند شناخته می شوند.

کیفیت: خوشه بندی خوب چیست؟

- یک روش خوب خوشه بندی خوشه هایی با کیفیت بالا تولید می کند.
- بیشترین شباهت درون کلاسی
- کمترین شباهت بین کلاسی
- کیفیت یک روش خوشه بندی مبتنی است بر:
- معیار های شباهت مورد استفاده در روش



- پیاده سازی آن و
- توانایی روش در کشف
- همه یا برعی از الگوهای

اندازه گیری کیفیت خوشه بندی

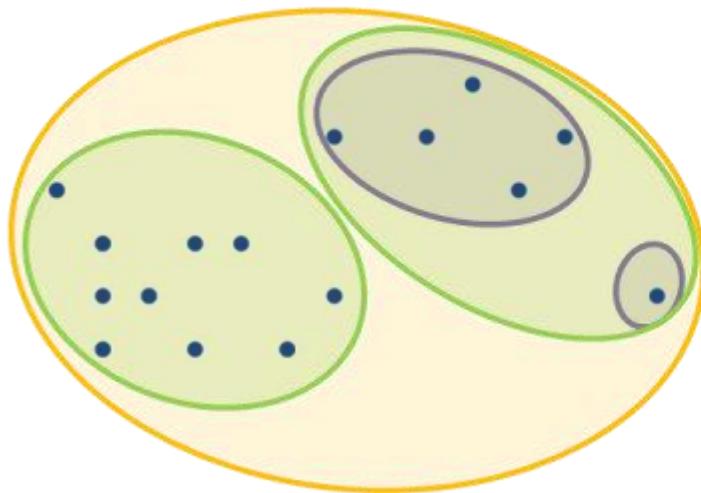
■ معیار شباht/عدم شباht

- شباht معمولا با استفاده از معیار فاصله بیان می شود : $d(i, j)$
- تعاریف توابع فاصله معمولا برای انواع مختلف داده (اسمی، دودویی,...) متفاوت است.
- متغیرهای مختلف می توانند براساس معنا و کاربردشان وزن بگیرند.
- کیفیت خوشه بندی:
 - در خوشه بندی تعریف شباht، کیفیت یا خوبی خوشه بندی کار مشکلی است و معمولا بسیار وابسته به موضوع است.

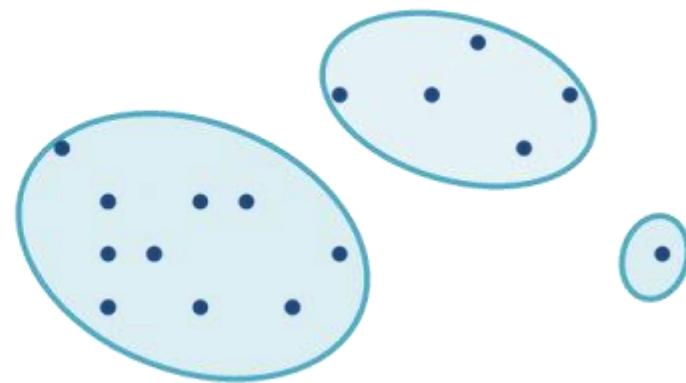
ملاحظات تحلیل خوشه

- معیارهای پارتیشن بندی
- پارتیشن بندی تک سطحی یا چند سطحی (اغلب پارتیشن بندی چند سطحی یا سلسله مراتبی مطلوب تر است.)

Hierarchical Clustering

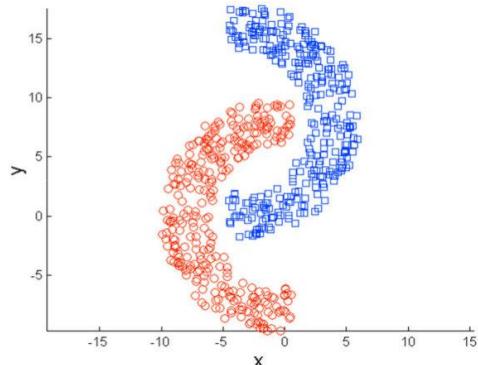


Partitional Clustering



تفکیک خوشه ها

- انحصاری (مثلا یک مشتری فقط به یک ناحیه متعلق است). در مقابل غیرانحصاری (مثلا یک سند به بیش از یک کلاس تعلق دارد).
- معیار شباht
- مبتنی بر فاصله (مثل تحصیلات، شبکه راه ها,...) در مقابل معیار های مبتنی بر چگالی یا پیوستگی)



فضای خوشه بندی

- فضای کامل (غلب زمانی که تعداد ابعاد کم است). در مقابل زیرفضاهای (در حالت خوشه بندی داده های با ابعاد زیاد)

الزامات و چالش ها

- مقیاس پذیری
- توانایی در خوش بندی انواع مختلف داده
- خوش بندی مبتنی بر بعضی محدودیت ها
 - قابلیت تفسیر و استفاده
- سایر
 - کشف خوش های با اشکال دلخواه
 - توانایی مقابله با داده نویز
 - خوش بندی افزایشی و حساس نبودن به ترتیب ورود داده
 - ابعاد بالا

روش های عمدۀ خوشۀ بندی

- Partitioning approach:
 - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
 - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical approach:
 - Create a hierarchical decomposition of the set of data (or objects) using some criterion
 - Typical methods: Diana, Agnes, BIRCH, CAMELEON
- Density-based approach:
 - Based on connectivity and density functions
 - Typical methods: DBSACN, OPTICS, DenClue
- Grid-based approach:
 - based on a multiple-level granularity structure
 - Typical methods: STING, WaveCluster, CLIQUE

- 
- Model-based:
 - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
 - Typical methods: EM, SOM, COBWEB
 - Frequent pattern-based:
 - Based on the analysis of frequent patterns
 - Typical methods: p-Cluster
 - User-guided or constraint-based:
 - Clustering by considering user-specified or application-specific constraints
 - Typical methods: COD (obstacles), constrained clustering
 - Link-based clustering:
 - Objects are often linked together in various ways
 - Massive links can be used to cluster objects: SimRank, LinkClus

فصل ۱۰: خوشه بندی - مفاهیم و روش های پایه



- تحلیل خوشه ای : مفاهیم پایه
- روش های بخش بندی: (Partitioning Methods)
- روش های سلسله مراتبی: (Hierarchical Methods)
- روش های مبتنی بر چگالی : (Density-Based Methods)
- روش های مبتنی بر گردید : (Grid-Based Methods)
- ارزیابی روش های خوشه بندی
- خلاصه

الگوریتم های بخش بندی: مفاهیم پایه

- روش های پارتیشن بندی: پارتیشن بندی یک پایگاه داده D شامل n شی به K خوش به نحوی که مجموع مربعات خطأ حداقل شود:

$$E = \sum_{i=1}^k \sum_{p \in C_i} (p - c_i)^2$$

(where c_i is the centroid or medoid of cluster C_i)

- هدف این است که با توجه به k داده شده، k خوش به ای را پیدا کنیم که معیار بخش بندی را بهینه نماید.

بهینه سراسری

- روشهای حریصانه: k-means and k-medoids algorithms
- : k-means (MacQueen'67, Lloyd'57/'82) t هر خوش به با استفاده از مرکز خوش معرفی می شود.

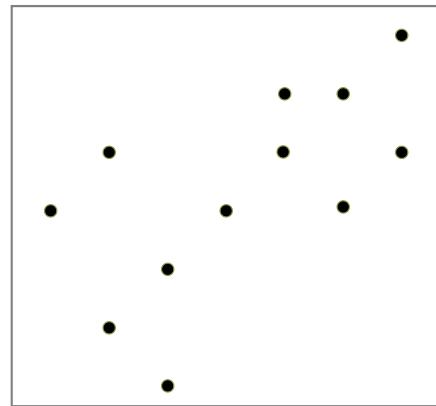
- k-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): هر خوش به با یک نماینده معرفی می شود.

روش خوشه بندی K-Means

با استفاده از K داده شده الگوریتم در ۴ مرحله بیان می شود:

- Partition objects into k nonempty subsets
- Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., *mean point*, of the cluster)
- Assign each object to the cluster with the nearest seed point
- Go back to Step 2, stop when the assignment does not change

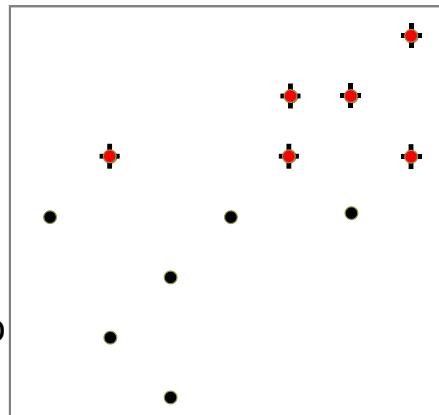
An Example of K-Means Clustering



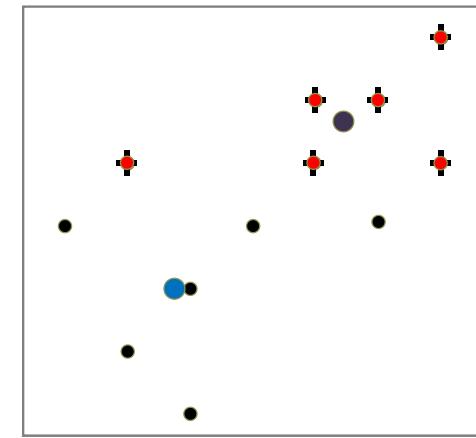
The initial data set

K=2

Arbitrarily partition objects into k groups

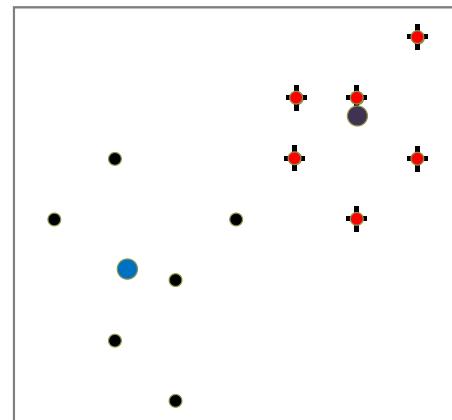


Update the cluster centroids

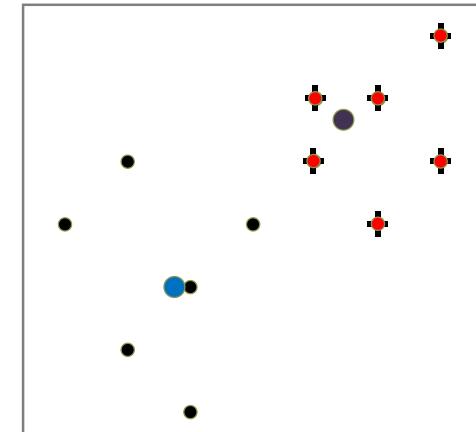


Reassign objects

Loop if needed

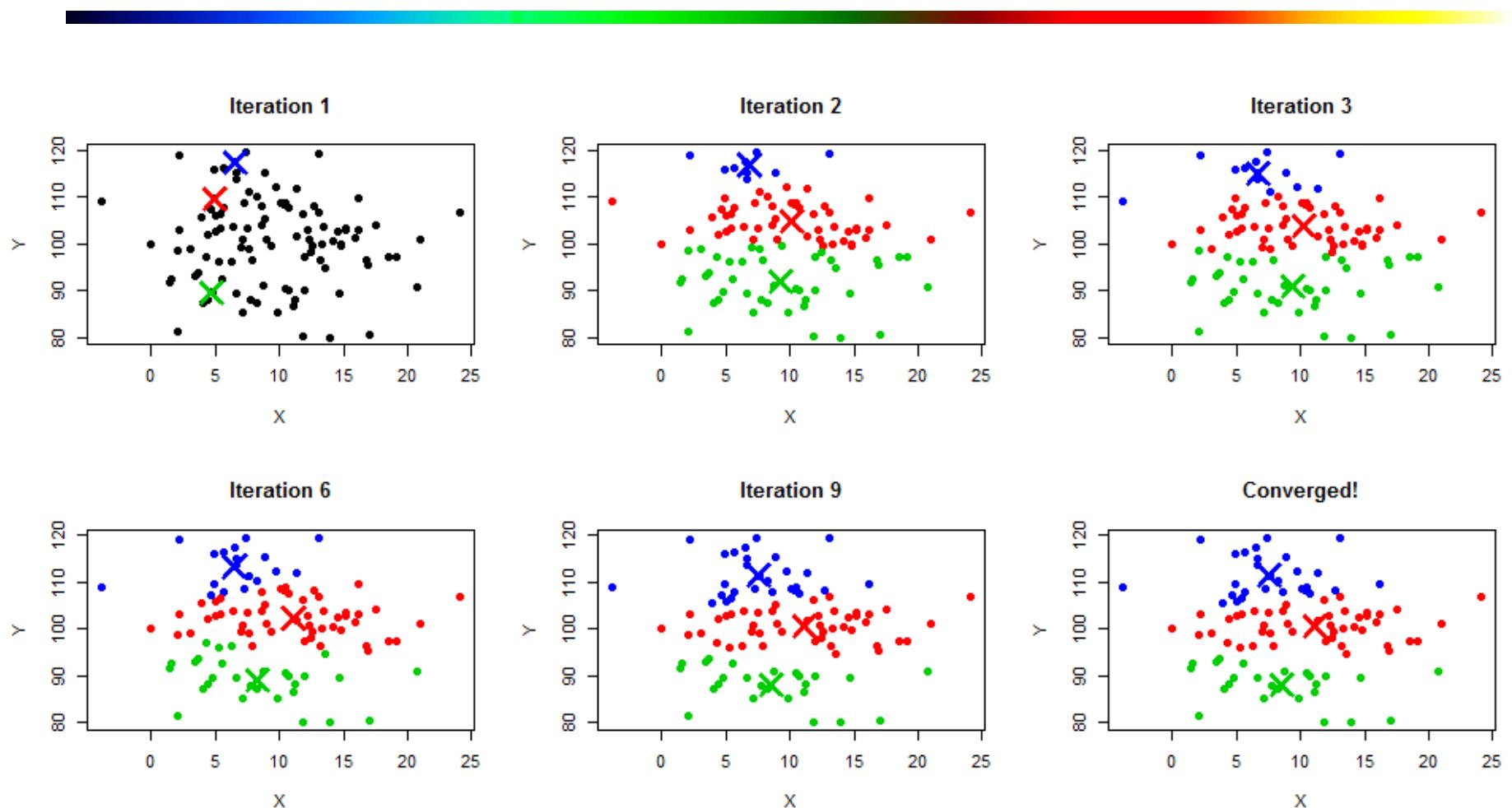


Update the cluster centroids



- Partition objects into k nonempty subsets
- Repeat
 - Compute centroid (i.e., mean point) for each partition
 - Assign each object to the cluster of its nearest centroid
- Until no change

روش مشابه دیگر



نکاتی در مورد الگوریتم K-Means

- ارزیابی معمولاً با مجموع مربعات خطأ (Sum of Squared Error)

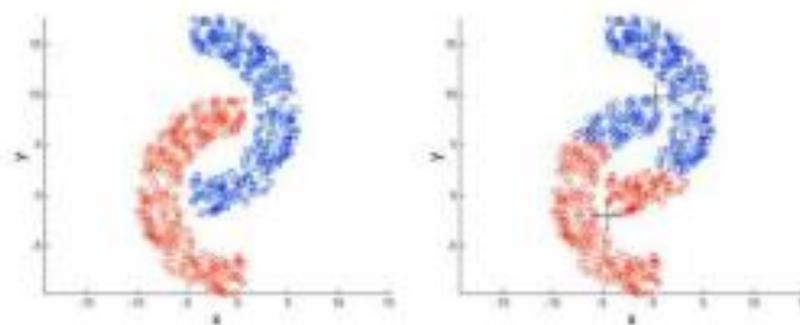
$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(m_i, x)$$

انجام می‌گیرد.

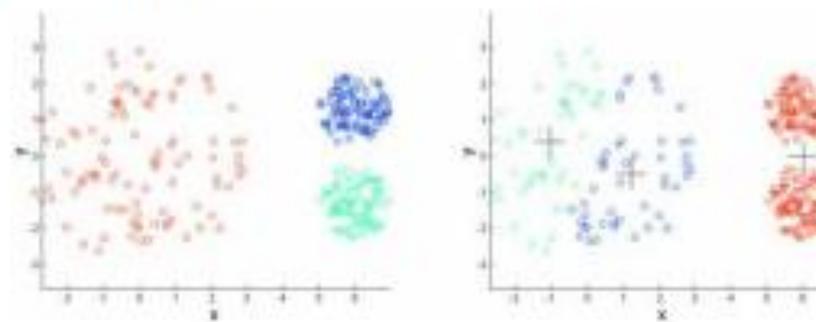
- نقاط ضعف:

- خطأ در زمانی که خوش‌ها در اندازه‌یا چگالی متفاوت باشند، حاوی داده پرت باشند، یا اشکال غیر کروی داشته باشند.
- حساس به چگونگی دسته‌های اولیه که بصورت تصادفی انتخاب می‌شوند.
- با افزایش k ارزیابی به شکل نادرست بهتر می‌شود.

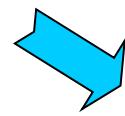
Non-convex/non-round-shaped clusters: Standard K -means fails!



Clusters with different densities



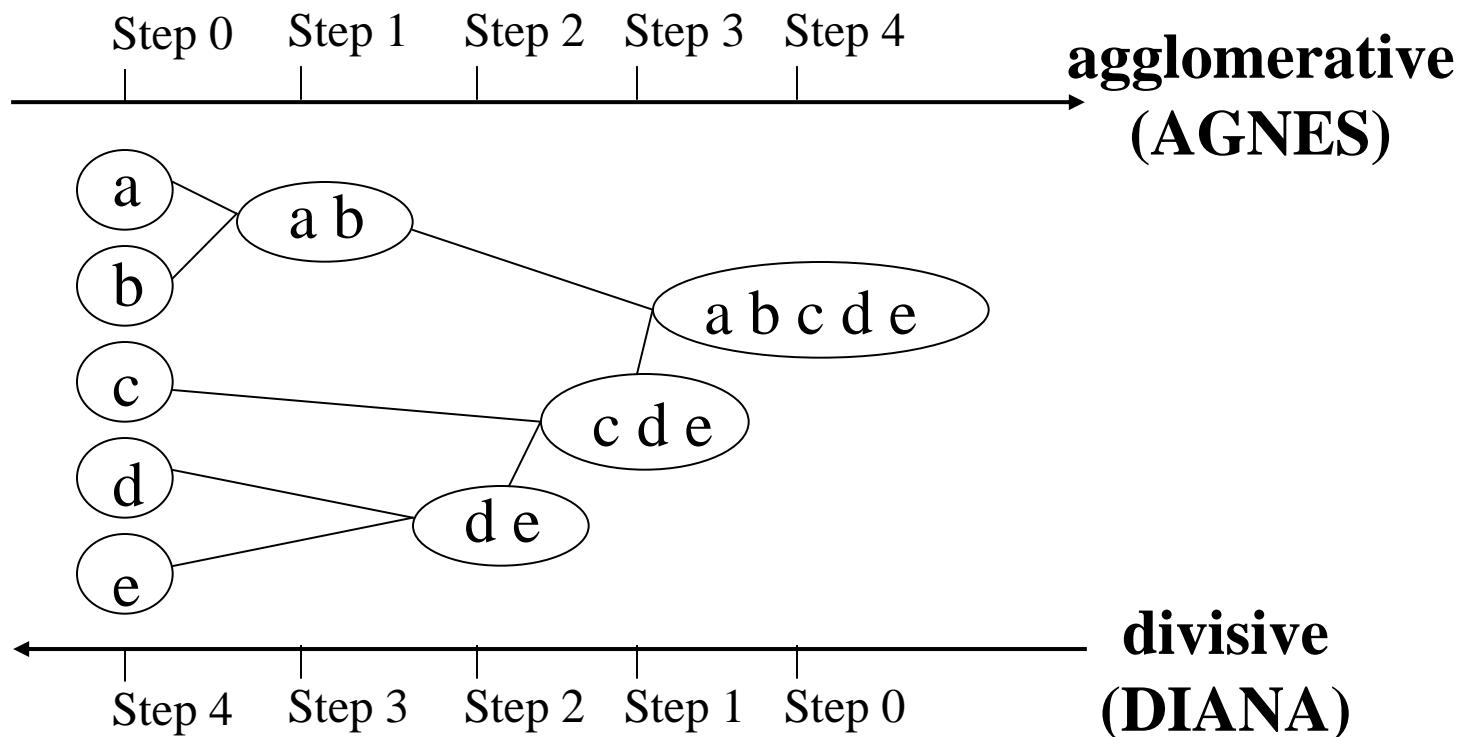
فصل ۱۰: خوشه بندی - مفاهیم و روش های پایه



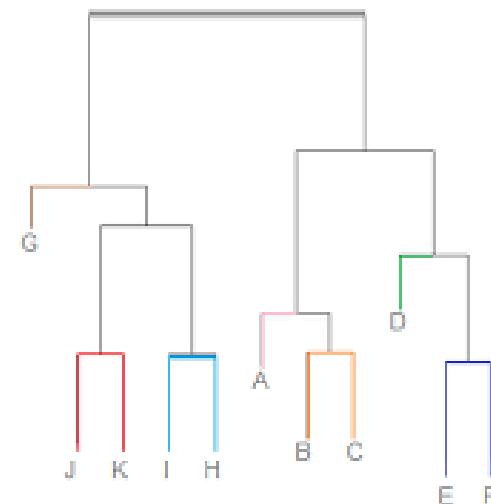
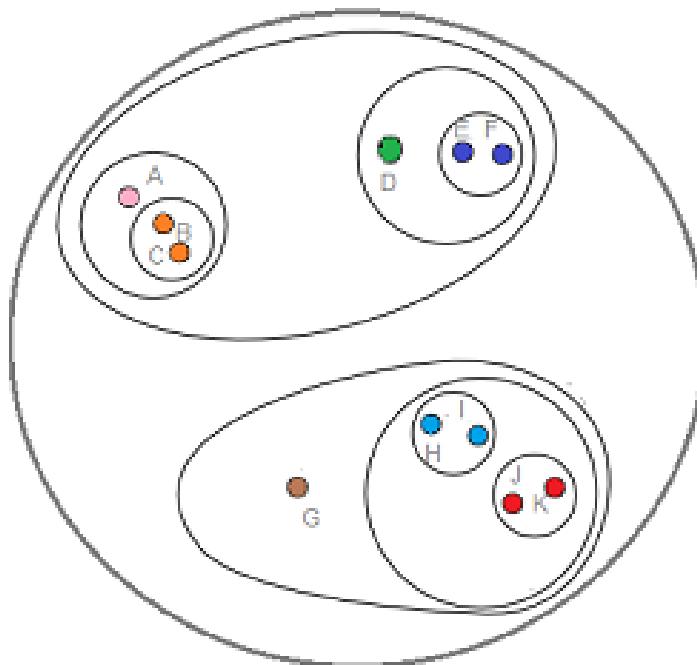
- تحلیل خوشه ای : مفاهیم پایه
- روش های بخش بندی: (Partitioning Methods)
- روش های سلسله مراتبی: (Hierarchical Methods)
- روش های مبتنی بر چگالی : (Density-Based Methods)
- روش های مبتنی بر گردید : (Grid-Based Methods)
- ارزیابی روش های خوشه بندی
- خلاصه

خوشه بندی سلسله مراتبی

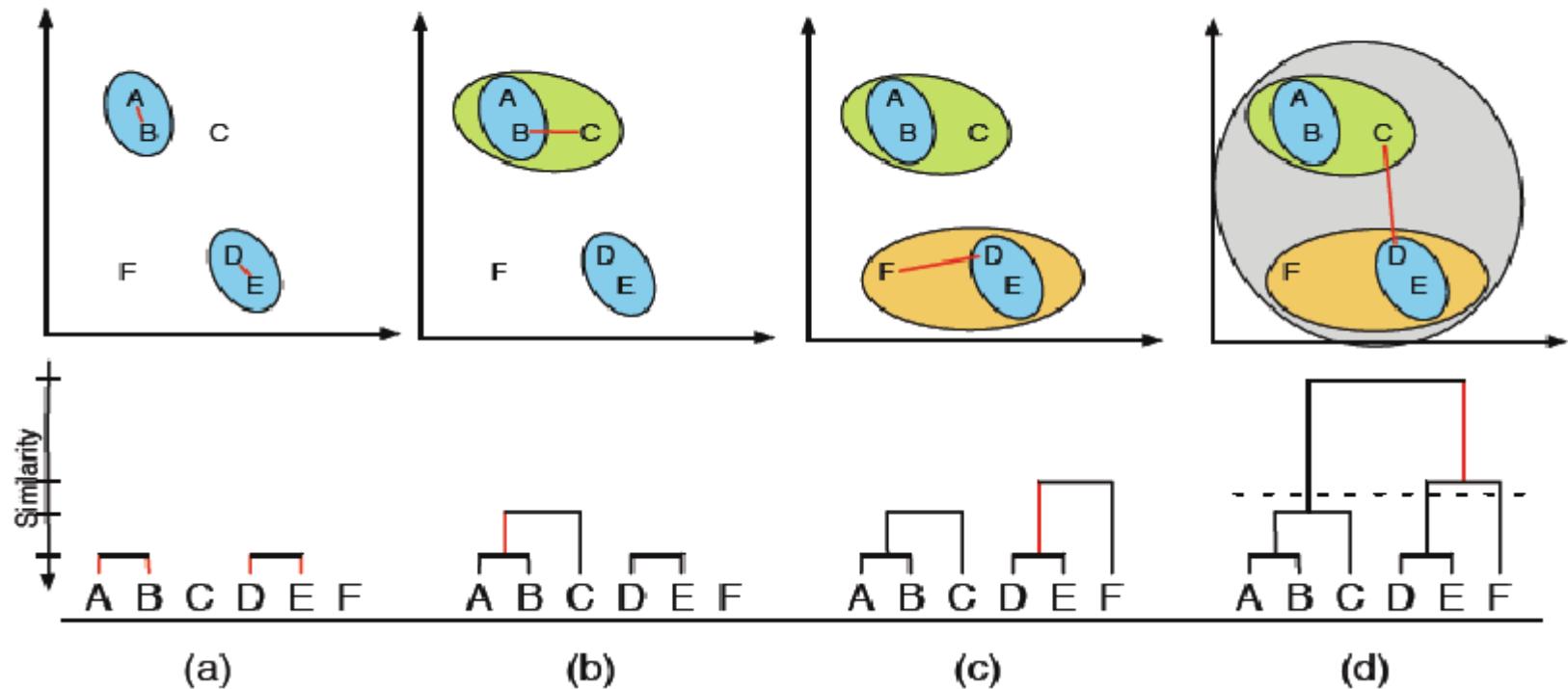
- استفاده از ماتریس فاصله بعنوان شرط خوشه بندی
- این روش تعداد خوشه ها را بعنوان ورودی دریافت نمی کند ولی نیاز به یک شرط پایان دارد.



مصور سازی نتیجه توسط dendogram



Example: Hierarchical Agglomerative Clustering



نقاط قوت خوش بندی سلسله مراتبی

- نیازی به فرض کردن تعداد خوش بندی ها نیست.
- هر تعداد مورد نظر از خوش بندی ها را می توان با برش dendrogram در سطح دلخواه بدست آورد.
- طبقه بندی بدست آمده می تواند معنادار باشد. (مثلا در علوم زیستی)

دو نوع اصلی خوشه بندی سلسله مراتبی

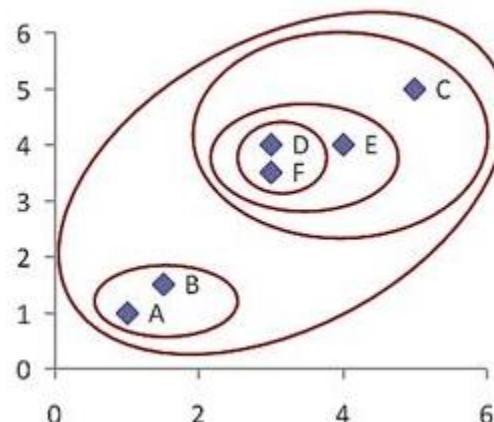
• Agglomerative

شروع با خوشه های تک عنصری و در هر مرحله ادغام نزدیکترین دو خوشه تا در نهایت یک خوشه باقی بماند

• Divisive

شروع با یک خوشه و تقسیم آن بصورت مرحله ای تا k خوشه حاصل گردد.

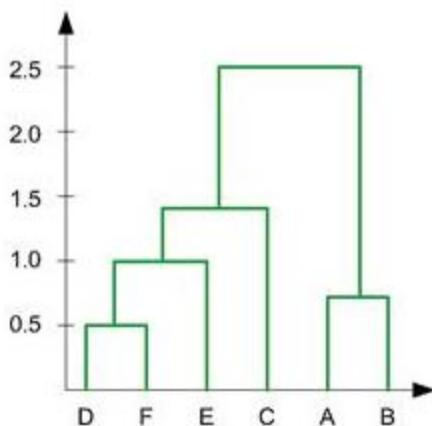
استفاده از ماتریس فاصله



Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	3.20	2.50	2.24	0.00	1.00
E	4.24	3.54	1.41	1.00	0.00



- 
- عملیات کلیدی محاسبه فاصله دو خوش است.
 - الگوریتم های متفاوتی در تعیین فاصله بکار می رود:
 - کمترین فاصله بین عناصر
 - بیشترین فاصله بین عناصر
 - میانگین فاصله عناصر
 - فاصله بین مرکز گروه ها