Knowledge Discovery in Databases
SS 2016

Chapter 2: Data Representation and Data Reduction

Lecture: Prof. Dr. Thomas Seidl

Tutorials: Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

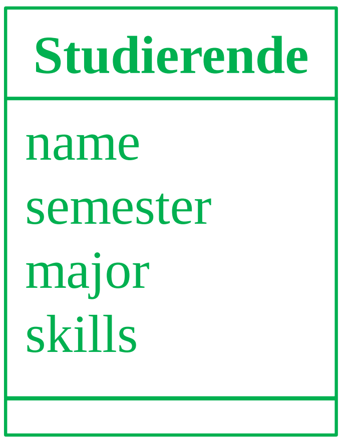
Overview

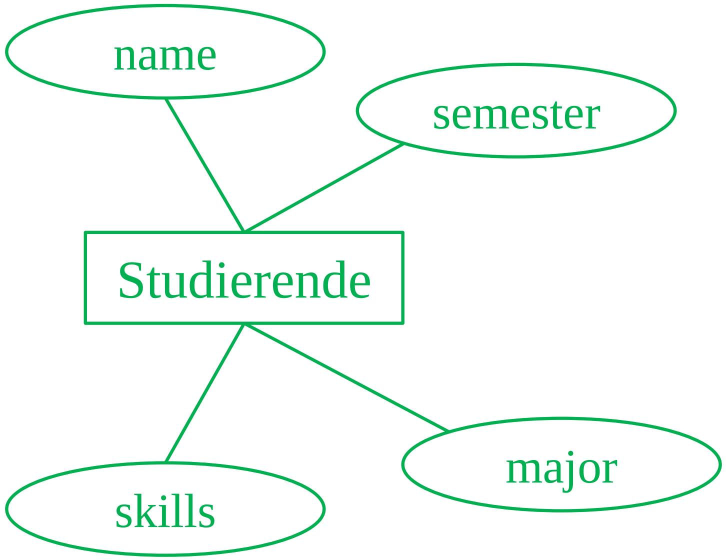
- Data representation
  - Data types
  - Comparison and similarity
  - Data visualization

- Data reduction
  - Aggregation
  - Generalization

Objects carry information

- Objects and attributes
  - Entity-Relationship diagram (ER)
  - UML class diagram
  - Data tables (relational model)





name	sem.	major	skills
Ann	3	CS	Java, C, R
Bob	1	CS	Java, PHP
Charly	4	History	Piano, …
Debra	2	Arts	Painting, …

Overview of (attribute) data types

- Simple data types
  - Numerical, categorical, ordinal
- Composed data types
  - Sets, sequences, vectors
- Complex data types
  - Multimedia: images, videos, audio, text, documents, web pages, etc.
  - Spatial, geometric: shapes, molecules, geography, etc.
  - Structures: graphs, networks, trees, etc.
- Examples for complex objects
  - Molecules: shape + structure + physical-chemical properties + …
  - City maps: shapes + traffic networks + points of interests + …
  - Mechanical parts: shape + physical properties + production process descr.

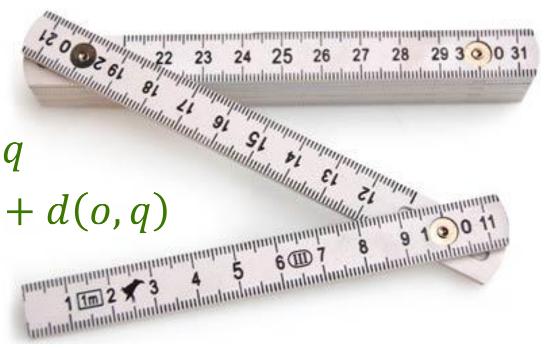
Simple data types and comparisons

- Numeric data
  - Numbers: natural, integer, rational, real numbers
  - Examples: age, income, shoe size, height, weight

  - Comparison: difference
  - Example: 3 is more similar to 30 than to 3,000

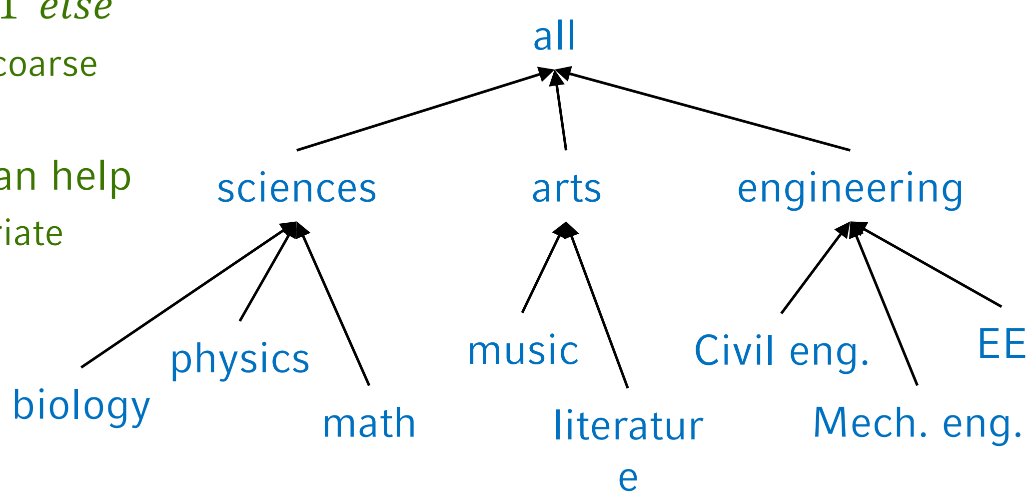
- Generalization: metric data
  - Metric space (O, d) consists of object set O and metric distance d
  - Comparison by (metric) distance $d: O \times O \to \mathbb{R}_0^+$
    - Symmetry: $\forall p, q \in O: d(p,q) = d(q,p)$
    - Identity of indiscernibles $\forall p, q \in O: d(p,q) = 0 \Leftrightarrow p = q$
    - Triangle inequality $\forall p, q, o \in O: d(p,q) \leq d(p,o) + d(o,q)$

  - Example: points in 2D space – Euclidean distance



Simple data types and comparisons

- Numeric data, metric data
- Categorical data
  - „Just identifiers“
    - Example occupation = {butcher, hairdresser, physicist, physician, … }
    - Example subjects = {physics, biology, math, music, literature, history, EE, … }
  - Comparison: how to compare values ???
    - Trivial metric: $d(p,q) = \begin{cases} 0 & if\ p = q \\ 1 & else \end{cases}$
      - Always works but is quite coarse
    - Generalization hierarchies can help
      - Path length seems appropriate
      - $d(music, literature) = 2$
      - $d(music, biology) = 4$
      - $d(music, music) = 0$



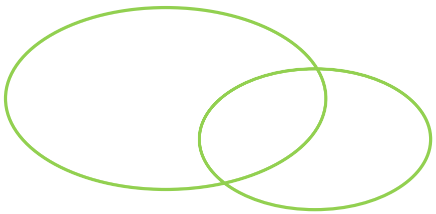
Simple data types and comparisons

- Numeric data, metric data
- Categorical data, hierarchical types
- Ordinal data
  - Some data carry a (total) order $(O, \leq)$
    - Transitivity $\forall p, q, o \in O\colon\ p \leq q \wedge q \leq o \Rightarrow p \leq o$
    - Antisymmetry $\forall p, q \in O\colon\ p \leq q \wedge q \leq p \Rightarrow p = q$
    - Totality $\forall p, q \in O\colon\ p \leq q \vee q \leq p$
  - Examples
    - Numbers $3 \leq 30 \leq 3{,}000$
    - Words high ≤ highschool ≤ highscore (i.e., lexicographical order)
    - Frequencies „How often did you sleep bad last year?“
      never ≤ seldom ≤ rarely ≤ occasionally ≤ sometimes ≤ often ≤ frequently ≤ regularly ≤ usually ≤ always
    - (Vague) sizes „How big was that problem?“
      tiny ≤ small ≤ medium ≤ big ≤ huge

Composed data types

- Sets
  - Put individual values together
  - Example: $skills = \wp(\{Java, C, Python, R, \dots\})$
  - Comparison
    - Symmetric set difference: $d(R,S) = (R - S) \cup (S - R) = (R \cup S) - (R \cap S)$
    - Jaccard distance:

$$d(R,S) = \frac{(R \cup S) - (R \cap S)}{R \cup S}$$



  - Bitvector representation of a set on a given, ordered base set
    - Sample base $B = \langle math, physics, chemistry, biology, music, arts, english \rangle$
    - Example sets $S = \{math, music, english\} = \langle 1,0,0,0,1,0,1 \rangle$
      $R = \{math, physics, arts, english\} = \langle 1,1,0,0,0,1,1 \rangle$
    - Hamming distance = sum of different entries: $d(R,S) = 3$
      - Equals the symmetric set difference

Composed data types

- Sequences, vectors
  - Put n values of a domain D together
  - Order does matter: $\mathrm{I}_n \to D$ for an index set $I_n = \{1, \dots, n\}$

- Comparison of vectors: two steps
  - Determine individual differences $|o_i - q_i|$, or distances $d(o_i, q_i)$
  - Combine individual distances to overall distance $d(o, q)$

- Examples
  - (Simple) sum: $d_1(o,q) = \sum_{i=1}^{n} |o_i - q_i|$ (Manhattan)
  - Root of sum of squares: $d_2(o,q) = \sqrt{\sum_{i=1}^{n} (o_i - q_i)^2}$ (Euclidean)
  - Maximum: $d_\infty(o,q) = \max_{i=1,\dots,n} \{|o_i - q_i|\}$ (Maximum)
  - General formula: $d_p(o,q) = \sqrt[p]{\sum_{i=1}^{n} |o_i - q_i|^p}$ (Minkowski)
  - Weighted Minkowski dist.: $d_{p,w}(o,q) = \sqrt[p]{\sum_{i=1}^{n} w_i \cdot |o_i - q_i|^p}$

Complex data types

- Components
  - Structure: graphs, networks, trees
  - Geometry: shapes / contoures, routes / trajectories
  - Multimedia: images, audio, text, etc.

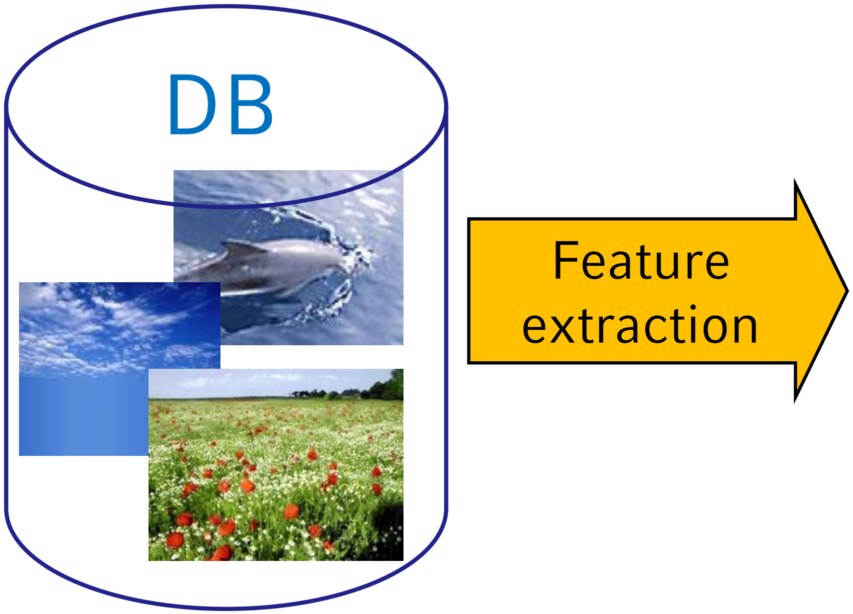
- Similarity models: approaches
  - Direct measures – highly data type dependent
  - Feature extraction – explicit vector space embedding
  - Kernel trick – implicit vector space embedding

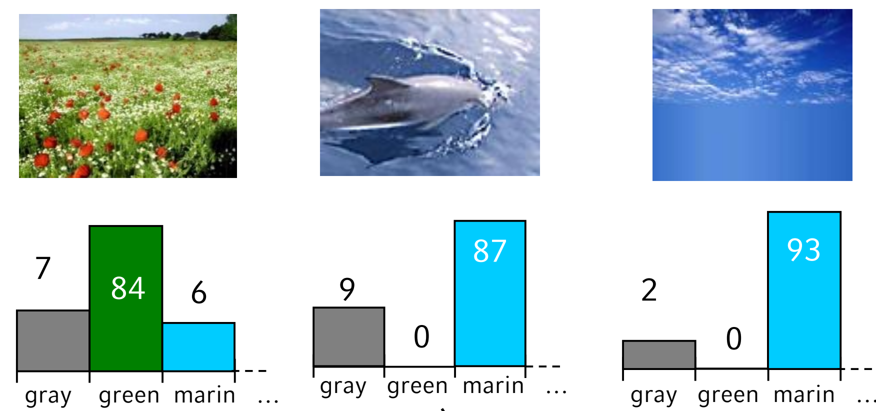
- Examples for similarity models

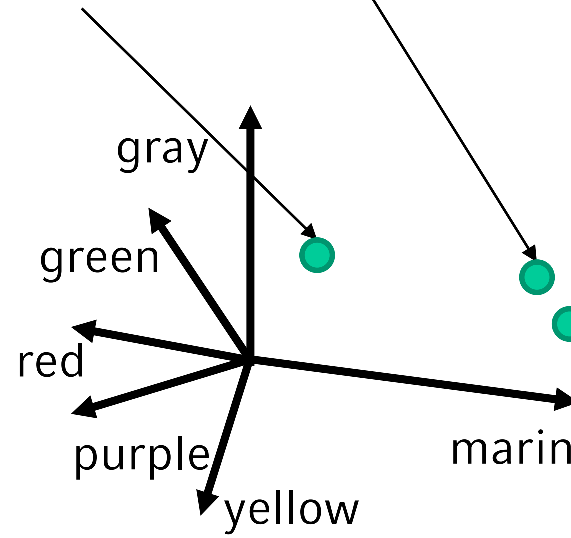
Examples	Direct distance	Feature-based	Kernel-based
Graphs	Structural alignment	Degree histograms	Label sequence kernel
Geometry	Hausdorff distance	Shape histograms	Spatial pyramid kernel
Sequences	Edit distance	Symbol histograms	Cosine distance

Feature extraction

- Objects from database DB are mapped to feature vectors







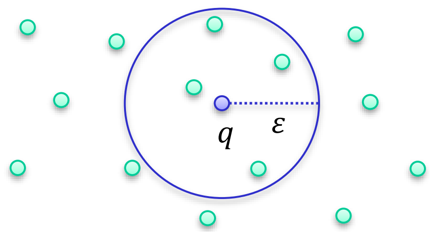
- Feature vector space
  - Points represent objects
  - Distance corresponds to (dis-)similarity

Similarity queries

- Similarity queries are basic operations in (multimedia) databases
- Given a universe O, database DB, distance function d, and query object q:

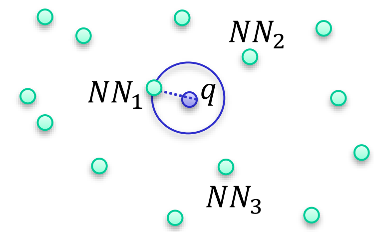
- **Range query** for range parameter $\varepsilon \in \mathbb{R}_0^+$:

  $$range(DB, q, d, \varepsilon) = \{o \in DB \mid d(o,q) \leq \varepsilon\}$$



- **Nearest neighbor query**:

  $$NN(DB, q, d) = \{o \in DB \mid \forall o' \in DB: \; d(o,q) \leq d(o',q)\}$$



- ***k*-nearest neighbor query** for parameter $k \in \mathbb{N}$:

  $$NN(DB, q, d, k) \subset DB \text{ with } |NN(DB, q, d, k)| = k \text{ and}$$
  $$\forall o \in NN(DB, q, d, k), o' \in DB - NN(DB, q, d, k): d(o,q) \leq d(o',q)$$

- **Ranking query** (partial sorting query): „get next“ functionality for picking database objects in an increasing order wrt. to their distance to q:

  $$\forall i \leq j: \; d\left(q, rank_{DB,q,d}(i)\right) \leq d\left(q, rank_{DB,q,d}(j)\right)$$

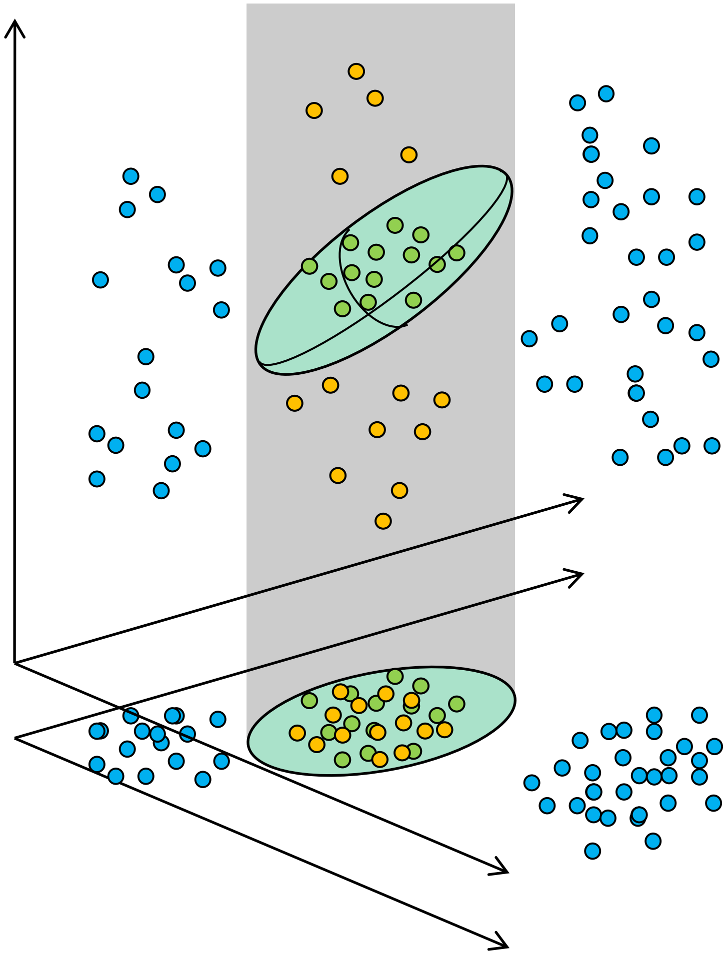
Similarity Search

- Example range query $range(DB, q, d, \varepsilon) = \{o \in DB | d(o, q) \le \varepsilon\}$

- Naive search by sequential scan
  - Fetch database objects from secondary storage (disk, e.g.): $O(n)$
  - Check distances individually: $O(n)$

- Fast search by applying database techniques
  - Filter-refine architecture
    - Filter: Boil database DB down to (small) candidate set $C \subseteq DB$.
    - Refine: Apply exact distance calculation to candidates from C only.
  - Indexing structures
    - Avoid sequential scans by (hierarchical or other) indexing techniques
    - Data access in (fast) $O(n)$, $O(\log n)$ or even $O(1)$

Filter-refine architecture



- Principle of multi-step search
  - Fast filter step produces candidate set C (by approximate distance function d')
  - Exact distance function d is calculated on candidate set $C \subseteq DB$ only.
- Example: dimensionality reduction

  [GEMINI: Faloutsos 1996; KNOP: Seidl&Kriegel 1998]

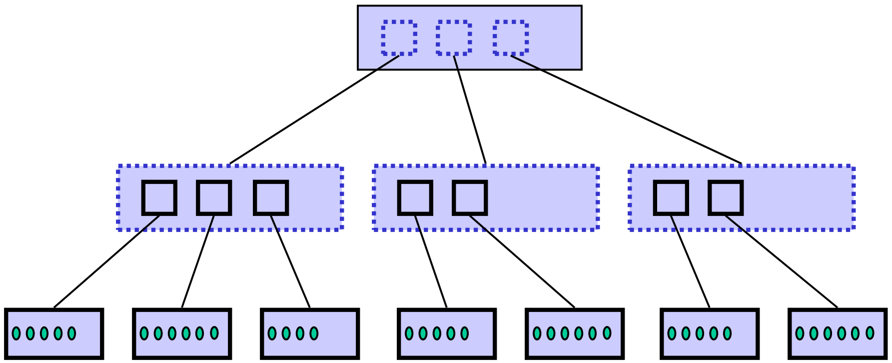
- *ICES* criteria for filter quality

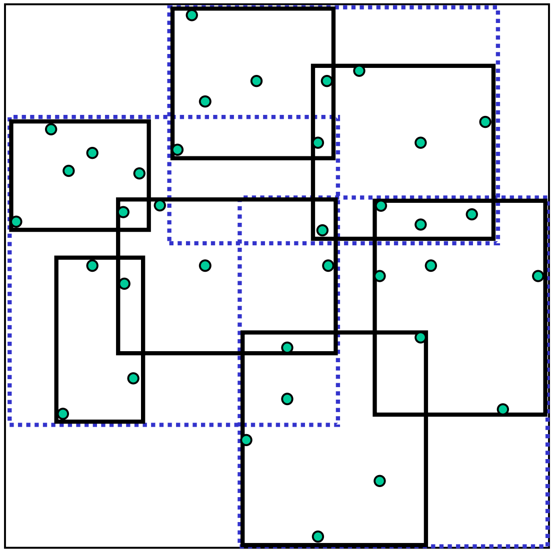
I ndexable	–	Index enabled
C omplete	–	No false dismissals
E fficient	–	Fast individual calculation
S elective	–	Small candidate set

[Assent, Wenning, Seidl: ICDE 2006]

Indexing

- Organize data in a way that allows for fast access to relevant objects, e.g. by heavy pruning.





- R-Tree as an example for spatial index structure
  - Hierarchy of minimum bounding rectangles
  - Disregard subtrees which are not relevant for the current query region

Overview

- Data representation
  - Data types
  - Comparison and similarity
  - Data visualization

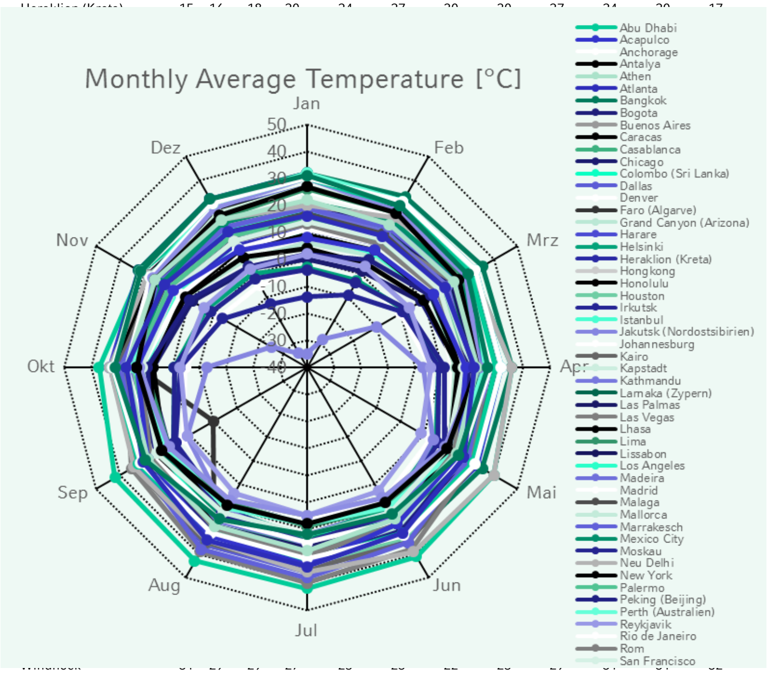
- Data reduction
  - Aggregation
  - Generalization

Data visualization

- Patterns in large data sets are hardly perceived from tabular numerical representations
- Data visualization transforms data in visually perceivable representations
  („a picture is worth a thousand words“)
- Combine capabilities
  - *Computers* are good in number crunching (and data visualization by means of computer graphics)
  - *Humans* are good in visual pattern recognition

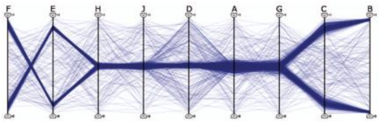
Monthly average temperature [°C]

Städte Ø	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
Abu Dhabi	25	27	31	36	40	41	42	43	42	37	31	27
Acapulco	32	31	32	32	33	33	33	33	33	33	32	32
Anchorage	-4	-2	0	6	13	17	18	17	13	5	-3	-5
Antalya	15	16	19	22	27	32	35	36	32	27	21	17
Athen	13	14	17	20	26	30	34	34	29	24	18	14
Atlanta	11	13	18	23	26	30	31	31	28	23	17	12
Bangkok	32	33	35	36	35	34	33	33	33	32	32	32
Bogota	20	19	19	19	19	18	18	18	19	19	19	20
Buenos Aires	30	28	26	23	19	16	15	17	19	21	26	29
Caracas	30	28	30	30	31	32	32	32	33	32	31	30
Casablanca	18	18	20	21	22	25	26	27	26	24	21	19
Chicago	0	1	9	16	21	26	29	28	24	17	9	2
Colombo (Sri Lanka)	31	31	32	32	32	31	31	31	31	31	31	31
Dallas	13	16	21	25	29	33	36	36	32	26	19	14
Denver	7	8	14	14	21	28	32	30	25	18	12	6
Faro (Algarve)	16	16	19	21	23	27	29	29	26	23	19	17
Grand Canyon (Arizona)	6	8	13	15	21	27	29	27	25	18	12	6
Harare	27	26	27	26	24	21	22	24	28	29	28	27
Helsinki	-3	-3	2	9	15	20	23	21	17	9	3	0

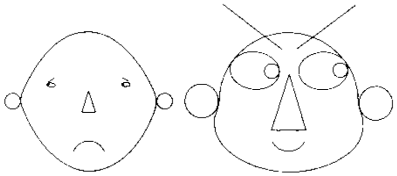


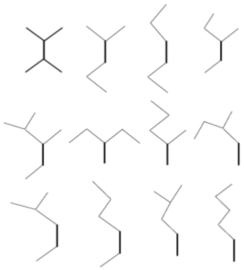
Data Visualization: Techniques

- Geometric Techniques
  - Idea: Visualization of geometric transformations and projections of the data
  - Examples: Scatterplots, Parallel Coordinates

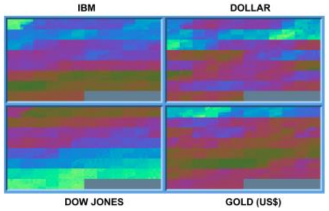


- Icon-based Techniques
  - Idea: Visualization of data as icons
  - Examples: Chernoff Faces, Stick Figures





- Pixel-oriented Techniques
  - Idea: Visualize each attribute value of each data object by one colored pixel
  - Example: Recursive Patterns

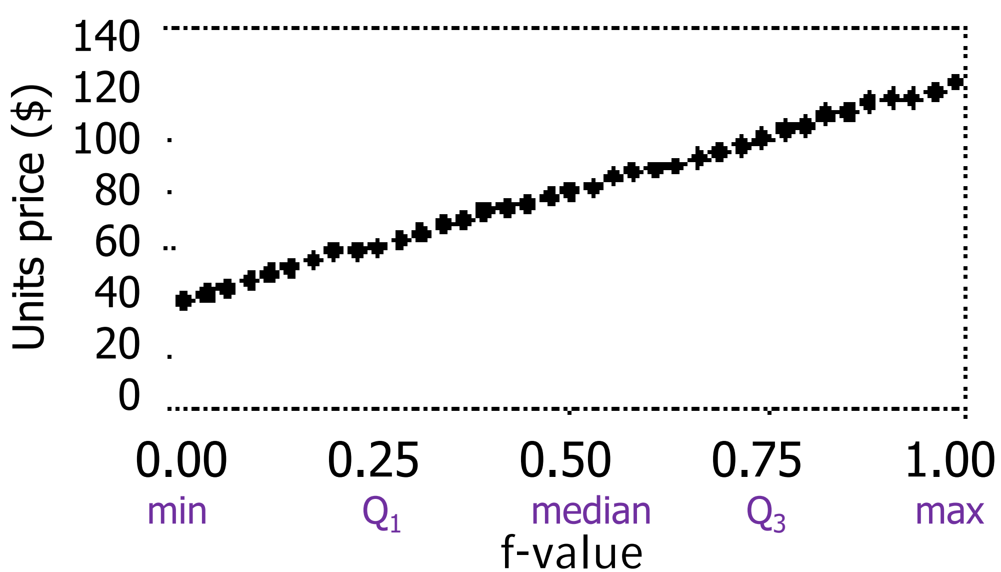


- Other Techniques:
  - Hierarchical Techniques, Graph-based Techniques, Hybrid-Techniques, …

Slide credit: Keim, Visual Techniques for Exploring Databases, Tutorial Slides, KDD 1997.

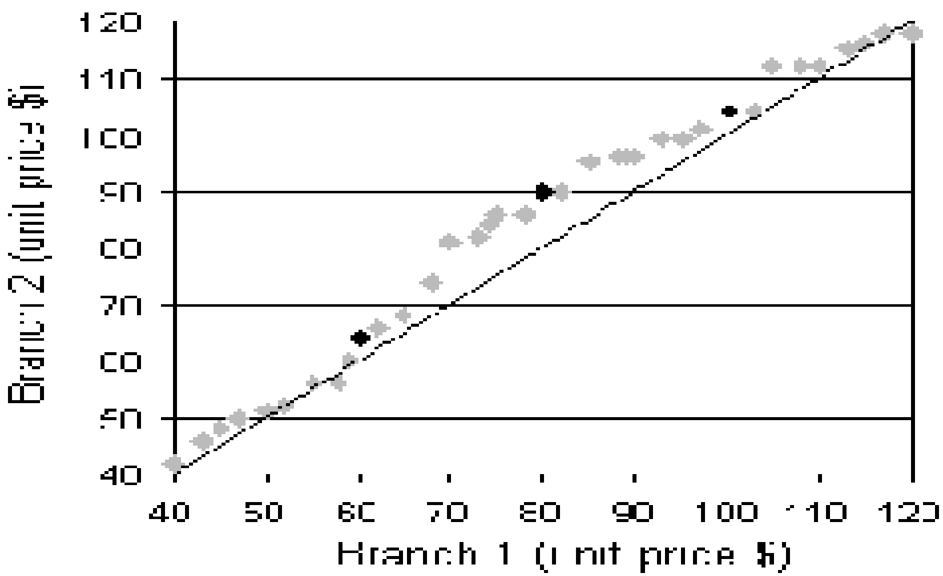
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - The q-quantile x_q indicates the value x_q for which the fraction q of all data is less than or equal to x_q (called percentile if q is a percentage); e.g., median = 50%-quantile or 50th percentile.



Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another

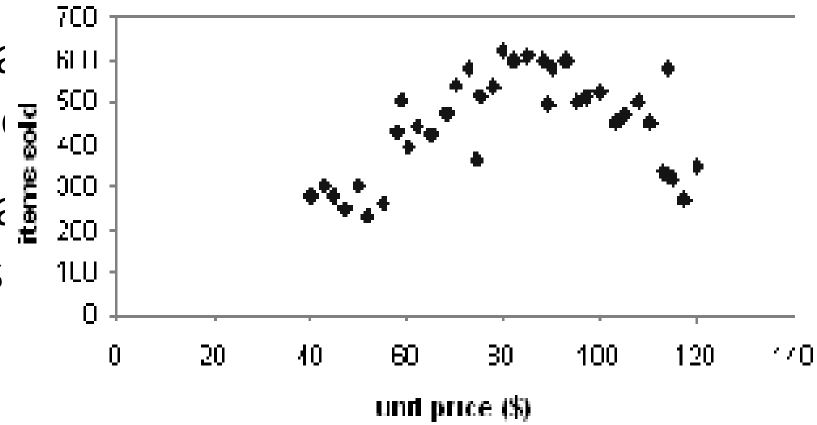


DATABASE SYSTEMS GROUP

Scatter plot & Loess Curve

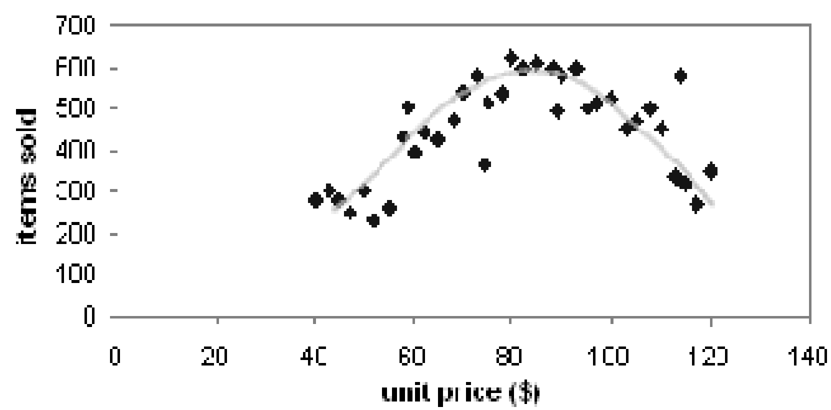
LMU

- Provides a first look at bivariate dat to see clusters of points, outliers, et
- Each pair of values is treated as a pa of coordinates and plotted as points in the plane



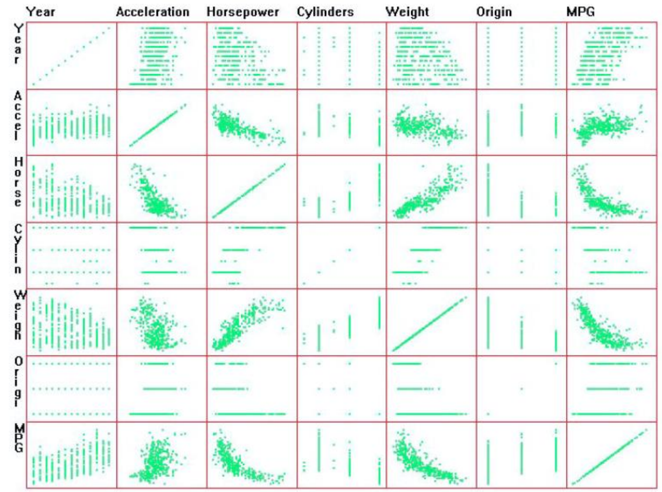
Loess Curve (local regression)

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression

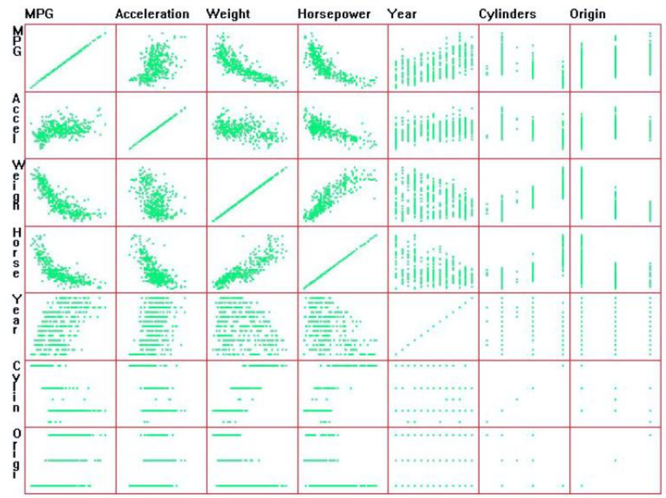


Scatterplot Matrix

- Matrix of scatterplots (x-y diagrams) of d-dimensional data
  - Indicates correlations in pairs of dimensions



- Ordering of dimensions is important
  - Dimension reordering helps to better understand the structures in the data and reduces clutter
  - The interestingness of different orderings can be evaluated with quality metrics (e.g. Peng et al.)

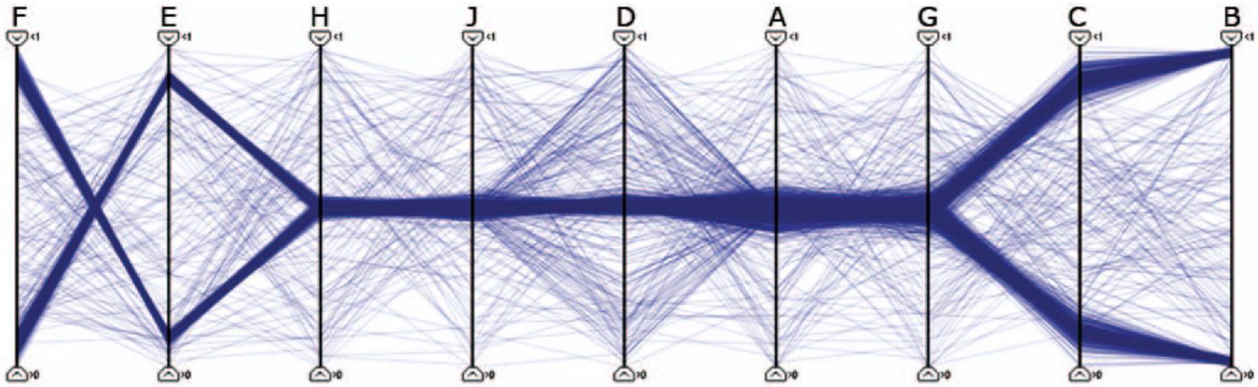


Figures from Peng et al., Clutter Reduction in Multi-Dimensional Data Visualizazion Using Dimension Reordering, IEEE Symp. on Inf. Vis., 2004.

Parallel Coordinates

[Ins 85] Inselberg, A.: The Plane with Parallel Coordinates, Special Issue on Computational Geometry. The Visual Computer, Vol. 1, pp. 69-97, 1985.

- A d-dimensional data space is visualized by d parallel axes
- Each axis is scaled to the min-max range in the corresponding dimension
- Every data object is visualized as a polygonal line which intersects each of the axes at the point that corresponds to the value of the object in the respective dimension



Slide credit: Keim, Visual Techniques for Exploring Databases, Tutorial Slides, KDD 1997.

Figure from Bertini et al., Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, Trans. on Vis. and Comp. Graph., 2011.

Parallel Coordinates

- Again, the ordering of the dimensions is important
- Interestingness of an ordering can be measured with a quality metric
- Quality or interestingness of orderings depends on what you want to visualize

- Example:
  - The first ordering is well-suited to visualize clusters in the data
  - The second ordering is well-suited to visualize correlation between the dimensions

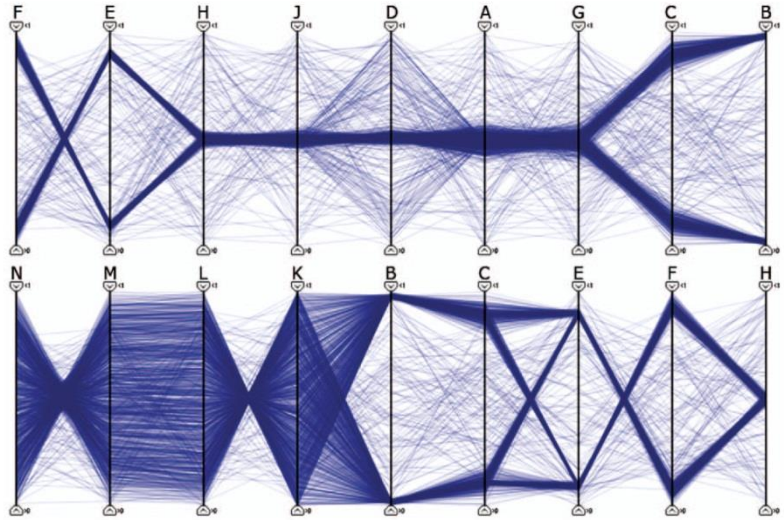
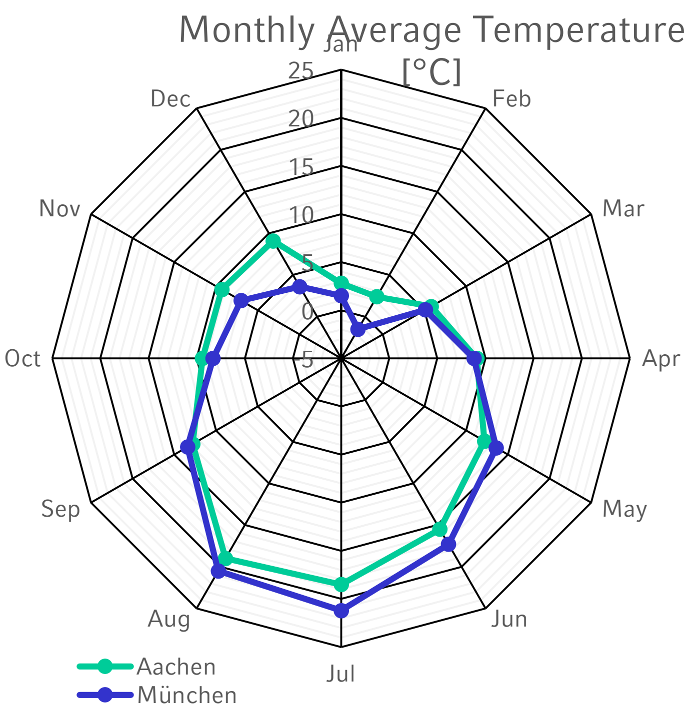
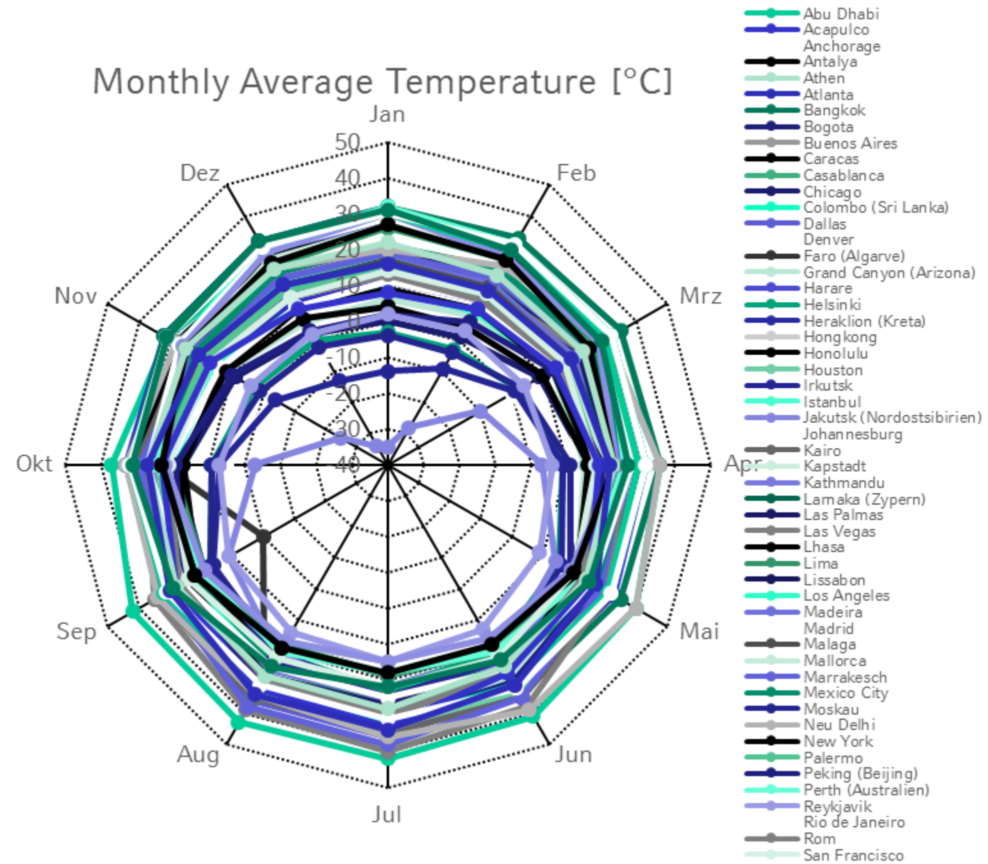


Figure from Bertini et al., Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, Trans. on Vis. and Comp. Graph., 2011.

Spiderweb model

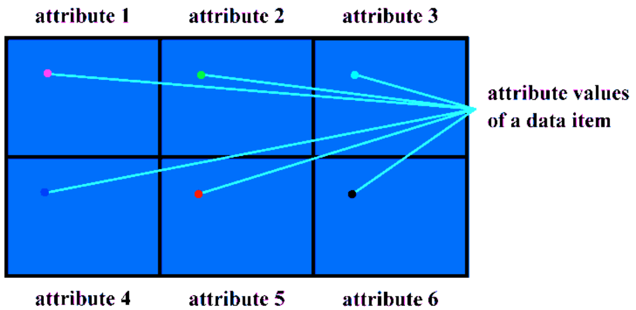
- Illustrate any single object by a polyline
- Contract origins of all axes to a global origin point
- Works well for few objects only



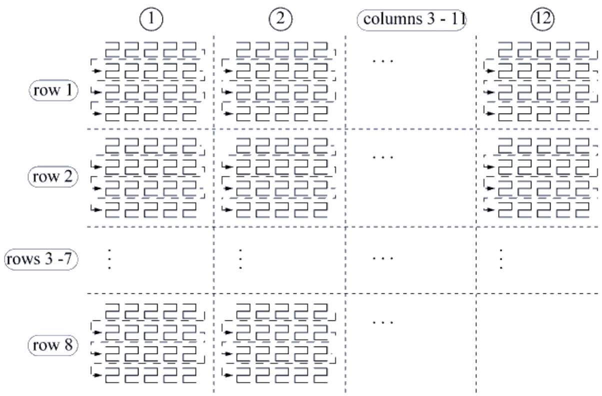


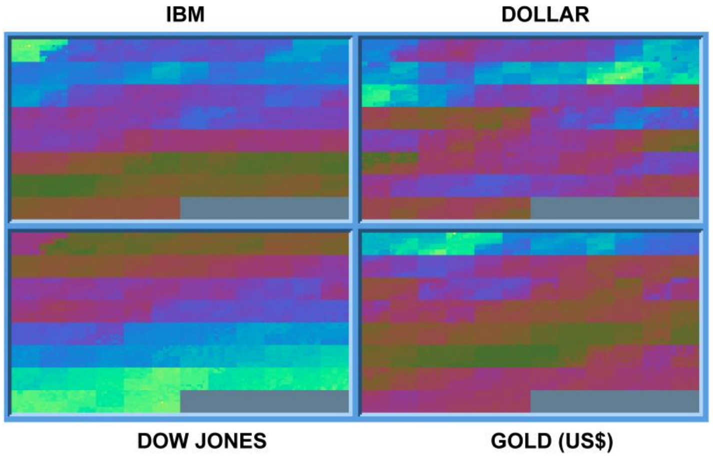
Pixel-oriented Techniques

- Each data value is mapped onto a colored pixel
- Each dimension is shown in a separate window



- Question: How to arrange the pixel ordering?
- One strategy: Recursive Patterns
  - iterated line and column-based arrangements
  - Example: time series of financial data



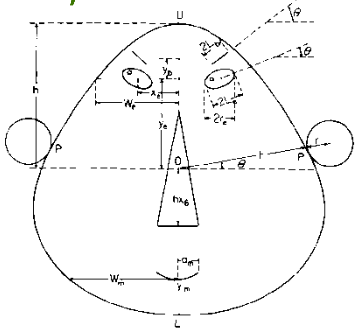


Figures from Keim, Visual Techniques for Exploring Databases, Tutorial Slides, KDD 1997.

Chernoff Faces

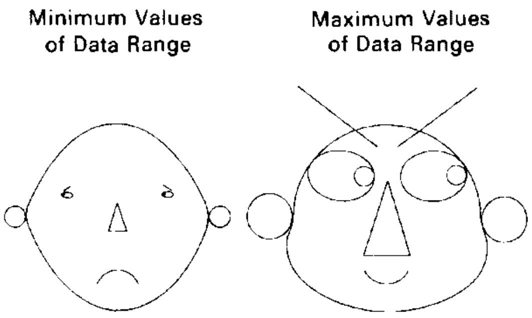
[Che 85] Chernoff, H.: The Use of Faces to Represent Points in k-Dimensional Space Graphically. Journal of the American Statistical Association, Vol. 68, pp. 361-368, 1973.

- Idea:
  - Represent data points from a d-dimensional space by facial expression
  - e.g.: dim 6 represents length of nose, dim 8 curvature of mouth, etc.



- Advantage:
  - Humans can evaluate similarity between faces much more intuitively than between high-dimensional vectors

- Shortcomings:
  - Without dimensionality reduction only applicable to data spaces with up to 18 dimensions
  - Which dimension represents what part?

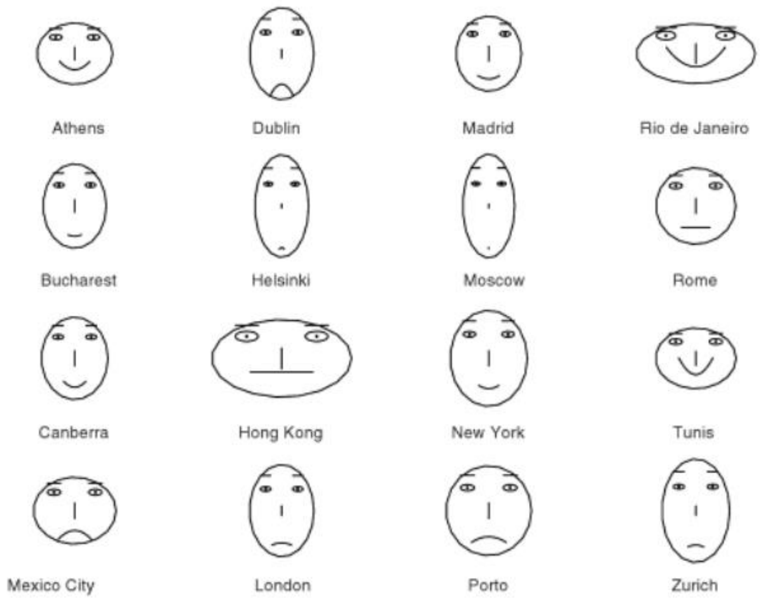


Figures taken from Mazza, Introduction to Information Visualization, Springer, 2009.

Example 1: Weather Data

City	Precip. average	Temp. average	Temp. max average	Temp. min average	Record max	Record min
Athens	37	17	21	13	42	−3
Bucharest	58	11	16	5	49	−23
Canberra	62	12	19	6	42	−10
Dublin	74	10	12	6	28	−7
Helsinki	63	5	8	1	31	−36
Hong Kong	218	23	25	21	37	2
London	75	10	13	5	35	−13
Madrid	45	13	20	7	40	−10
Mexico City	63	17	23	11	32	−3
Moscow	59	4	8	1	35	−42
New York	118	12	17	8	40	−18
Porto	126	14	18	10	34	−2
Rio de Janeiro	109	25	30	20	43	7
Rome	80	15	20	11	37	−7
Tunis	44	18	23	13	46	−1
Zurich	107	9	12	6	35	−20

Table 4.1 Annual climatic values in Celsius of some world cities. Values from http://www.weatherbase.com.



Figures from Riccardo Mazza, Introduction to Information Visualization, Springer, 2009.

Example 2: Financial Data

FIGURE 3
Facial Representation of Financial Performance (1 to 5 Years Prior to Failure)

FEDERAL					
Date Dimensions	Year to Failure 5	4	3	2	1
1. Return on Assets	0.10	0.11	0.06	0.03	−0.16
2. Debt Service	3.66	3.79	1.55	0.78	−14.11
3. Cash Flows	1.53	1.48	1.39	1.35	0.94
4. Capitalization	0.22	0.20	0.18	0.16	−0.02
5. Current Ratio	71.40	89.10	97.85	96.80	58.21
6. Cash Turnover	24.03	25.92	25.62	27.40	71.26
7. Receivables Turnover	5.25	4.46	4.26	4.36	9.56
8. Inventory Turnover	5.38	4.77	4.57	4.44	5.34
9. Sales per Dollar Working Capital	6.74	6.33	7.02	7.61	−45.77
10. Retained Earning/ Total Assets	0.32	0.30	0.01	−0.01	−0.26
11. Total Assets	0.94	.76	0.39	0.45	0.43

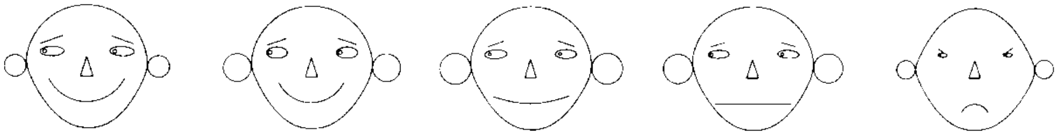


Figure from Huff et al., Facial Representation of Multivariate Data, Journal of Marketing, Vol. 45, 1981, pp. 53-59.

Overview

- Data representation
  - Data types
  - Comparison and similarity
  - Data visualization

- Data reduction
  - Aggregation
  - Generalization

Data reduction

- Why data reduction?
  - Better perception of patterns
    - Raw (tabular) data is hard to understand
    - Visualization is limited to (hundreds of) thousands of objects
    - Reduction of data may help to identify patterns
  - Computational complexity
    - Big data sets cause prohibitively long runtimes for data mining algorithms
    - Reduced data sets are useful …
    - … the more the algorithms produce (almost) the same analytical results

- How to approach data reduction?
  - Data generalization (abstraction to higher levels)
  - Data aggregation (basic statistics)

Data Reduction Strategies

ID	A1	A2	A3
1	54	56	75
2	87	124	65
3	3.4	6.8	76
4	86	23	4

16 bit
0000000001010110
→
4 bit
0110

Numerosity reduction

Reduce number of objects

Dimensionality reduction

Reduce number of attributes

Quantization, discretization

Reduce number of values per domain

ID	A1	A3
1	L	75
3	XS	76
4	XL	4

Examples for reduction techniques

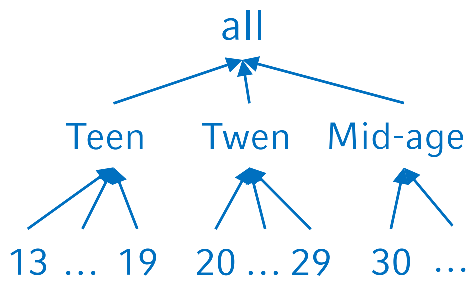
- Numerosity reduction
  - Sampling (loss of data)
  - Aggregation (model parameters, e.g., center / spread)

- Dimensionality reduction
  - Linear methods: feature subselection; principal components analysis; random projections; Fourier transform; wavelet transform
  - Non-linear methods: Multidimensional scaling (force model)

- Quantization
  - Binning (various types of histograms)
  - Generalization along hierarchies (OLAP; attribute-oriented induction)

Aggregate representation

- Connection of quantization, dimensionality reduction and generalization
  - Quantization is a special case of generalization
    - E.g., group *age* (7 bits) to *age_range* (4 bits)
  - Dimensionality reduction is degenerate quantization
    - Dropping *age* reduces 7 bits to zero bits
    - Corresponds to generalization of *age* to „all“ = „any age“ = no information



- Generalization yields duplicates
  - Merge duplicate tuples and introduce (additional) counter attribute
  - Aggregation is numerosity reduction (= less tuples)

Name	Age	Major
Ann	27	CS
Bob	26	CS
Eve	19	CS

Generali-zation

Name	Age	Major
(any)	Twen	CS
(any)	Twen	CS
(any)	Teen	CS

Aggre-gation

Age	Major	Count
Twen	CS	2
Teen	CS	1

Basic aggregates

- Basic descriptive statistics
  - Central tendency: Where is the data located? Where is it centered?
    - Examples: mean, median, mode, etc. (see below)
  - Variation, spread: How much do the data deviate from the center?
    - Examples: variance / standard deviation, min-max-range, …

- Examples
  - Age of students is around 20+
  - Shoe size is centered around 40
  - Recent dates are around 2020±
  - Average income is in the thousands

Distributive aggregate measures

Distributive | Algebraic | Holistic

- *Distributive*
  - The result derived by applying the function to *n* aggregate values is the same as that derived by applying the function on all the data without partitioning.

- Examples:

$$count(D_1 \cup D_2) = count(D_1) + count(D_2)$$
$$sum(D_1 \cup D_2) = sum(D_1) + sum(D_2)$$
$$min(D_1 \cup D_2) = \min(\min(D_1), \min(D_2))$$
$$max(D_1 \cup D_2) = \max(\max(D_1), \max(D_2))$$

Algebraic aggregate measures

Distributive | **Algebraic** | Holistic

- *Algebraic*
  - Can be computed by an algebraic function with *M* arguments (where *M* is a bounded integer), each of which is obtained by applying a distributive aggregate function.
  - E.g.: $avg() = sum() \,/\, count()$; $standard\_deviation()$

- Examples:

$$avg(D_1 \cup D_2) = \frac{sum(D_1 \cup D_2)}{count(D_1 \cup D_2)} = \frac{sum(D_1) + sum(D_2)}{count(D_1) + count(D_2)}$$

$$\neq avg\big(avg(D_1), avg(D_2)\big)$$

Holistic aggregate measures

Distributive | Algebraic | **Holistic**

- *Holistic*
  - There is no constant bound on the storage size which is needed to determine / describe a sub-aggregate

- Examples:
  - $median$: value in the middle of a sorted series of values (= 50% quantile)
  - $median(D_1 \cup D_2) \neq simple\_function(median(D_1), median(D_2))$
  - $mode$: value that appears most often in a set of values
  - $rank$: k-smallest / k-largest value (cf. quantiles, percentiles)

Measuring the Central Tendency (1)

- *Mean* — (weighted) arithmetic mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\quad \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - Well-known measure for central tendency ("average")
  - Algebraic measure
  - However, only applicable to numerical data (sum, scalar multiplication)
- *Mid-range*
  - Average of the largest and the smallest values in a data set:
    $$(\max + \min) \,/\, 2$$
  - Applicable to numerical data only

→ What about categorical data?

Measuring the Central Tendency (2)

- *Median*
  - Middle value if odd number of values
  - For even number of values: average of the middle two values (numeric case), or one of the two middle values (non-numeric case)
  - Examples
    - never, never, never, rarely, rarely, often, usually, usually, always
    - tiny, small, big, big, big, big, big, big, huge, huge
    - tiny, tiny, small, medium, big, big, large, huge
  - Holistic measure
  - Applicable to ordinal data only (an ordering is required)

→ What if there is no ordering?

Measuring the Central Tendency (3)

- *Mode*
  - Value that occurs most frequently in the data
  - Example: blue, red, blue, yellow, green, blue, red
  - Well suited for categorical (i.e., non-numeric) data
  - Unimodal, bimodal, trimodal, …: there are 1, 2, 3, … modes in the data (multimodal in general), cf. mixture models
  - There is no mode if each data value occurs only once
  - Empirical formula for unimodal frequency curves that are moderately skewed:

$$mean - mode \;\approx\; 3 \cdot (mean - median)$$

  - Not directly applicable to continuous data
    - Approach: train (multimodal) model and consider modes of that model
    - Example: means in Gaussian mixture models are modes

Measuring the Dispersion of Data

- Variance
  - Applicable to numeric data
  - Algebraic measure, scalable computation

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

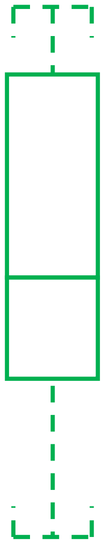
Calculation by two passes, is numerically much more stable

Single pass: calculate sum of squares and square of sum in parallel

- Standard deviation: Square root of the variance
  - Measures the spread around the mean
  - It is zero if and only if all the values are equal
  - Both the deviation and the variance are algebraic

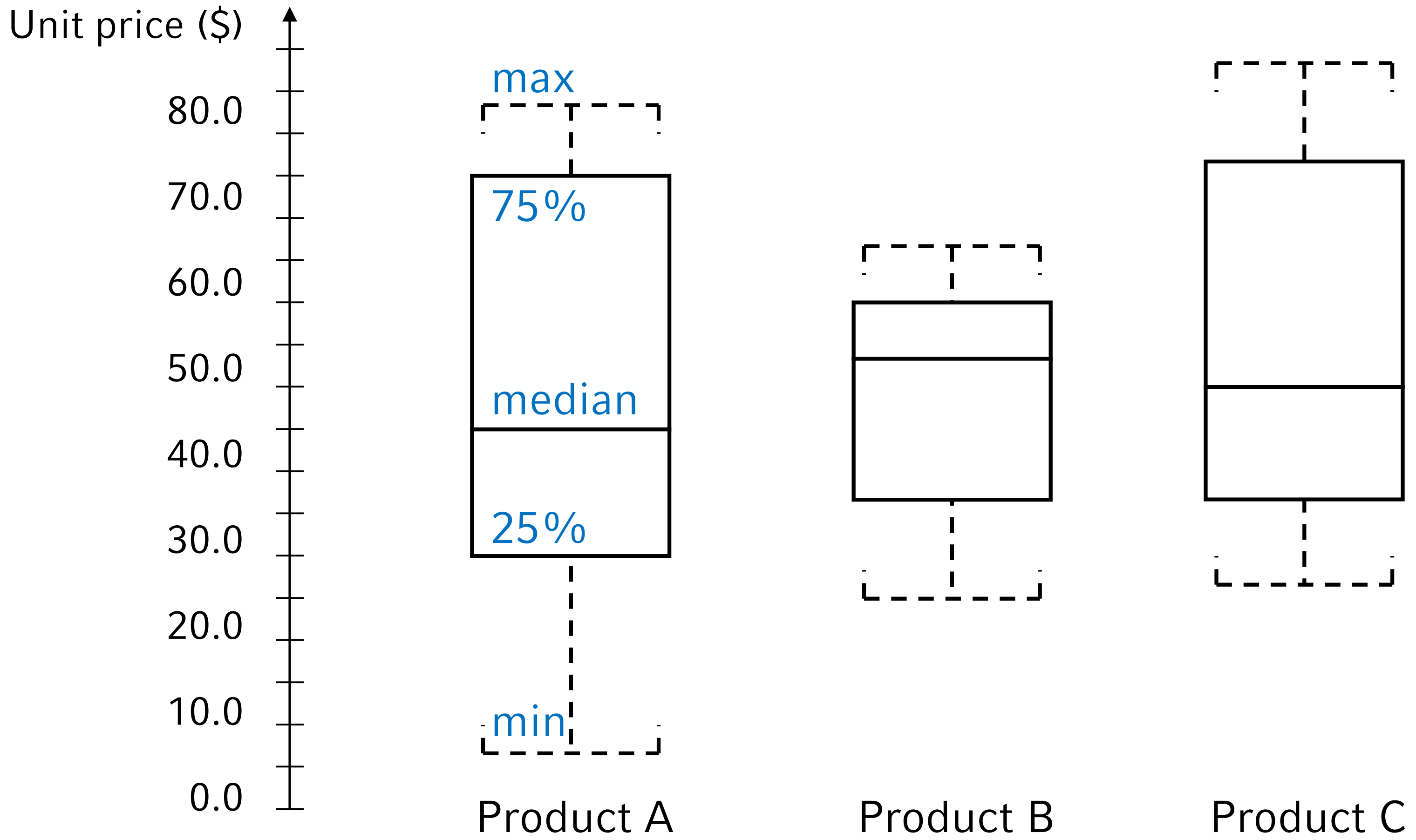
Boxplot Analysis

- Five-number summary of a distribution
  - Minimum, Q1, Median, Q3, Maximum
  - Represents 0%, 25%, 50%, 75%, 100%-quantile of the data
  - Also called “25-percentile”, etc.

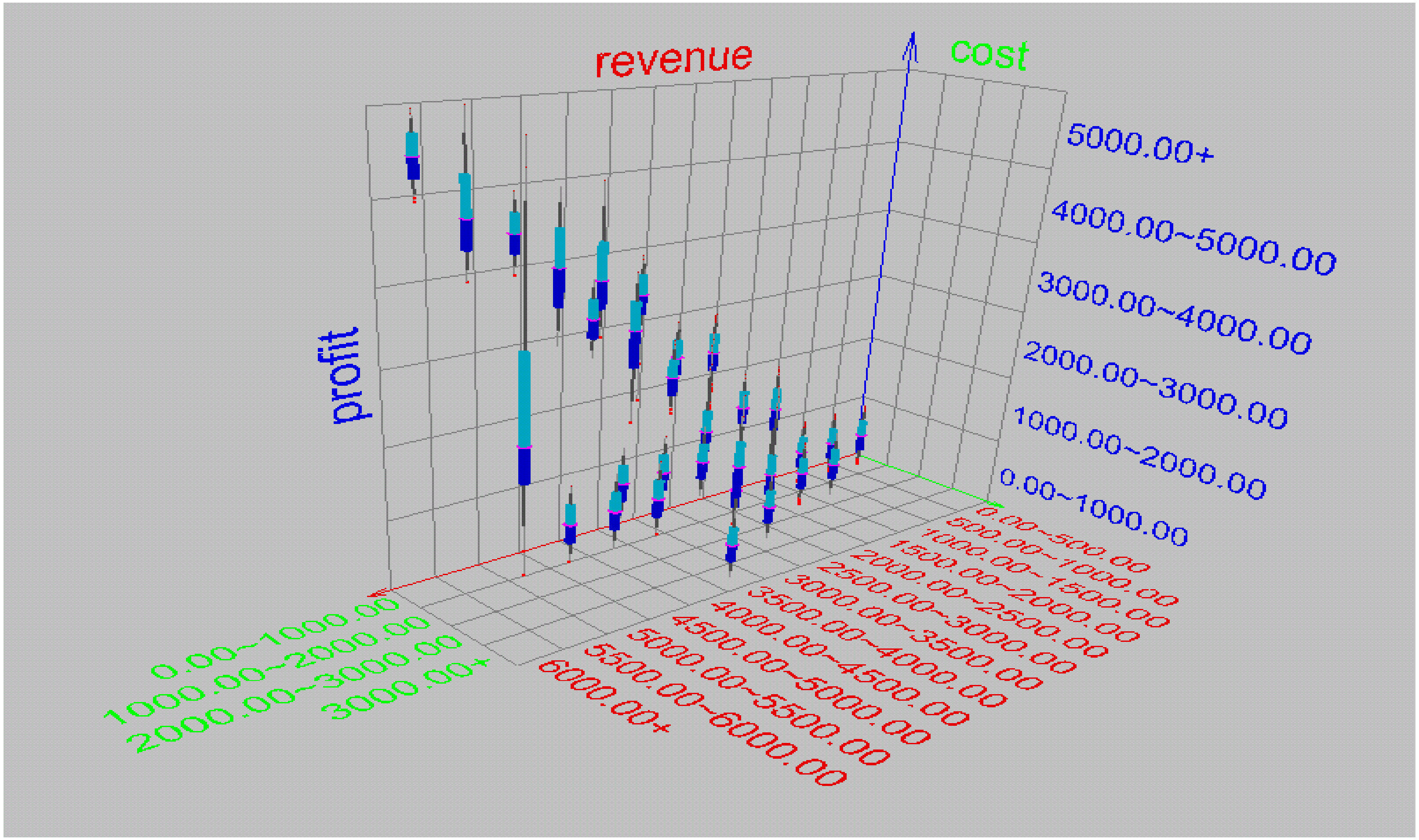


- Boxplot
  - Represent data set by a box
  - The box boundaries indicate the first and third quartiles
  - Its height is the inter-quartile range $IQR = Q_3 - Q_1$
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box, extend to minimum and maximum
  - Outliers: usually, values that are more than 1.5 x IQR below Q_1 or above Q_3

Boxplot Examples



Multidimensional Boxplot Analysis



Overview

- Data representation
  - Data types
  - Comparison and similarity
  - Data visualization
- Data reduction
  - Aggregation
  - Generalization

Generalization

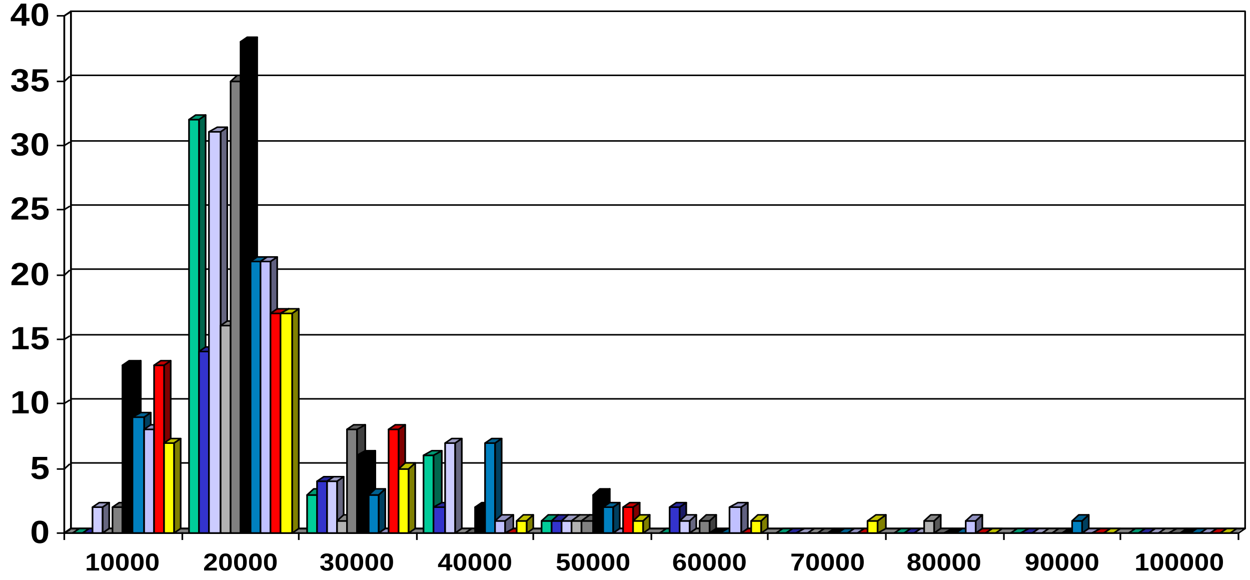
- Which partitions of the data to aggregate?

- All data
  - overall mean, overall variance → too coarse (overgeneralized)

- Different techniques to form groups for aggregation
  - Binning – histograms, based on value ranges
  - Generalization – abstraction based on generalization hierarchies
  - Clustering (see later) – based on object similarity

Binning techniques: Histograms

- Histograms use binning to approximate data distributions
- Divide data into bins and store a representative (sum, average, median) for each bin
- Popular data reduction and analysis method
- Related to quantization problems



Equi-width histograms

- Divide the range into N intervals of equal size: uniform grid
- If A and B are the lowest and highest values of the attribute, the width of intervals will be $W = (B - A)/N$

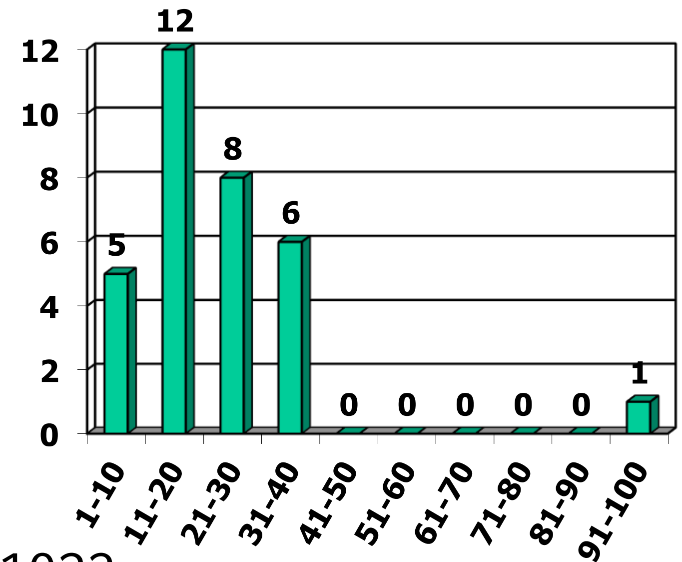
\+ Most straightforward

– Outliers may dominate presentation

– Skewed data is not handled well

- Example (data sorted, here: 10 bins):

  5, 7, 8, 8, 9, 11, 13, 13, 14, 14,
  14, 15, 17, 17, 17, 18, 19, 23, 24,
  25, 26, 26, 26, 27, 28, 32, 34, 36,
  37, 38, 39, 97



- Second example: same data set, insert 1023

Equi-height histograms

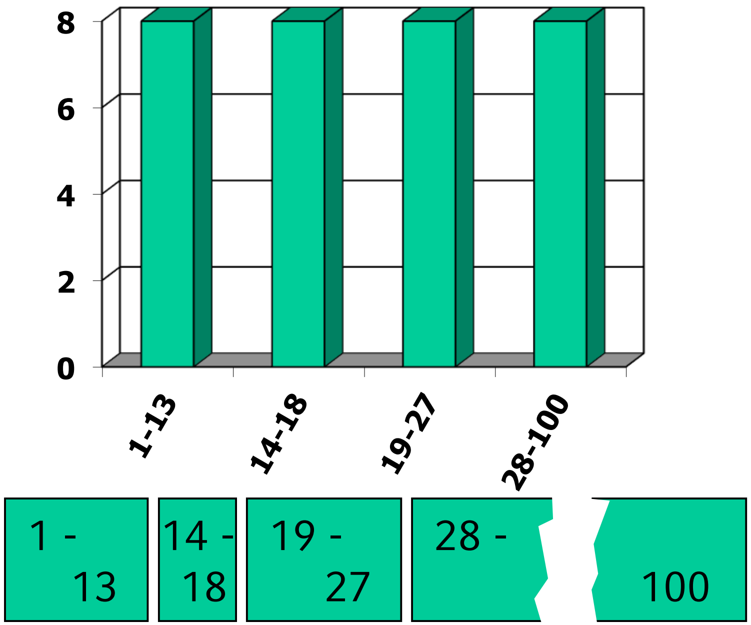
- It divides the range into N intervals, each containing approx. the same number of samples (*quantile-based approach*)

\+ Good data scaling

– If any value occurs often, the equal frequency criterion might not be met (intervals have to be disjoint!)

- Same Example (here: 4 bins):

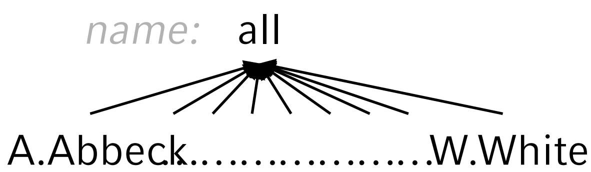
  5, 7, 8, 8, 9, 11, 13, 13, 14, 14, 14, 15, 17, 17, 17, 18, 19, 23, 24, 25, 26, 26, 26, 27, 28, 32, 34, 36, 37, 37, 38, 97

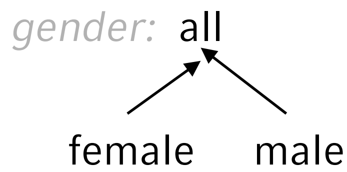


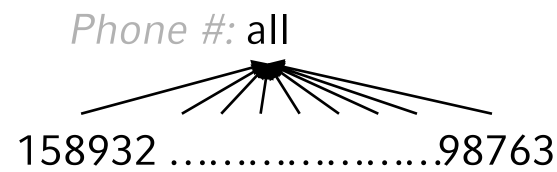
- Median = 50%-quantile
  - is more robust against outliers (cf. value 1023 from above)
  - Four bin example is strongly related to boxplot

Concept Hierarchies: *Examples*

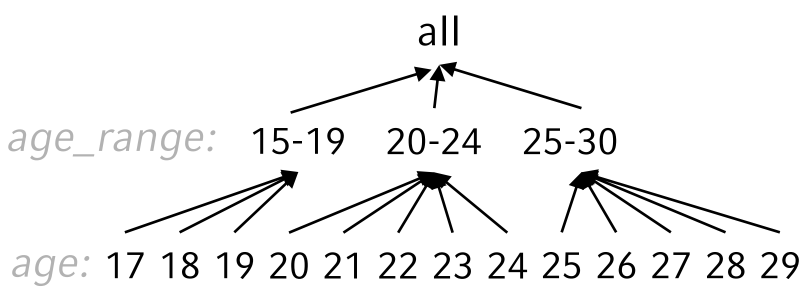
no (real) hierarchies

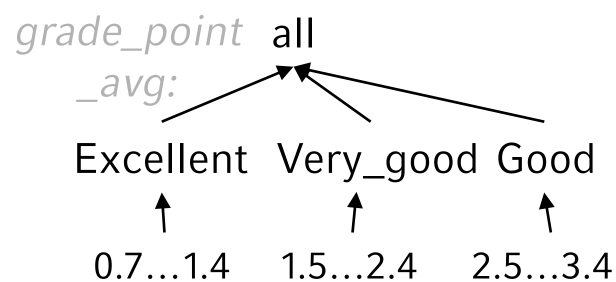




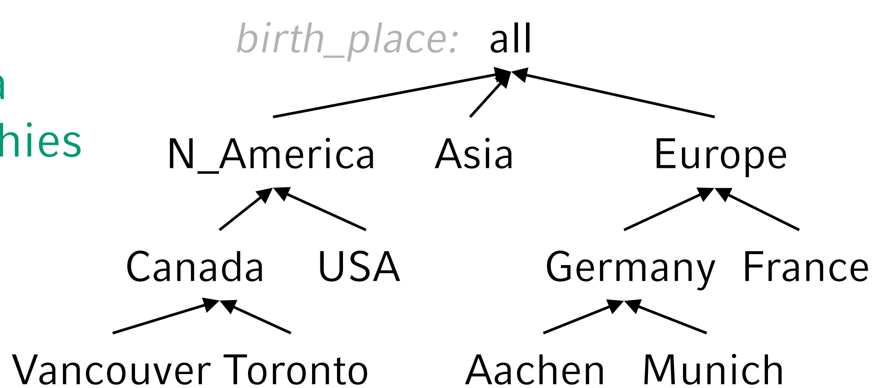


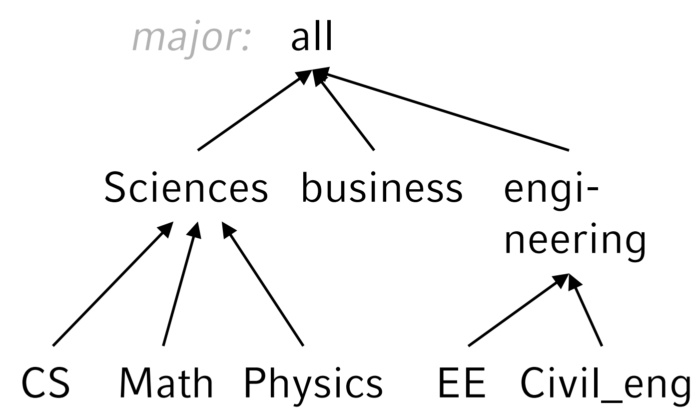
set grouping hierarchies





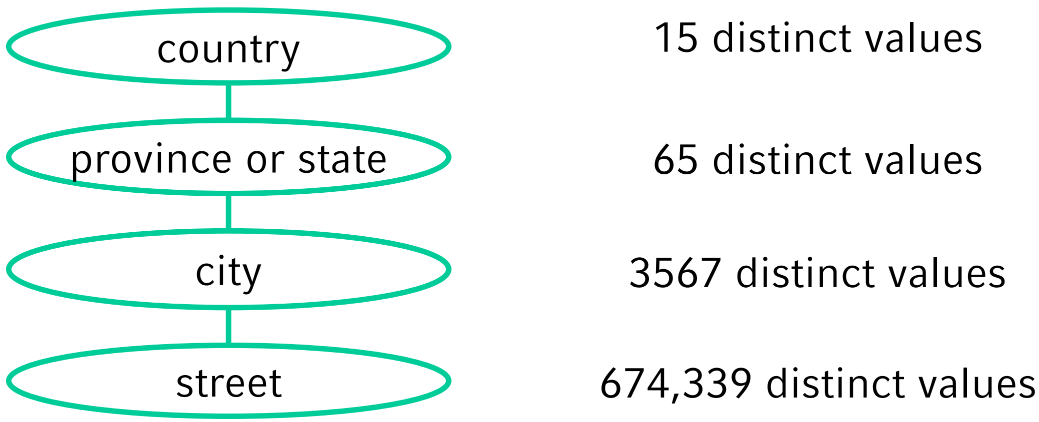
schema hierarchies





Concept hierarchy for categorical data

- Let concept hierarchies be specified by experts or just by users
- Heuristically generate a hierarchy for a set of (related) attributes
  - based on the number of distinct values per attribute in the attribute set
  - the attribute with the most distinct values is placed at the lowest level of the hierarchy



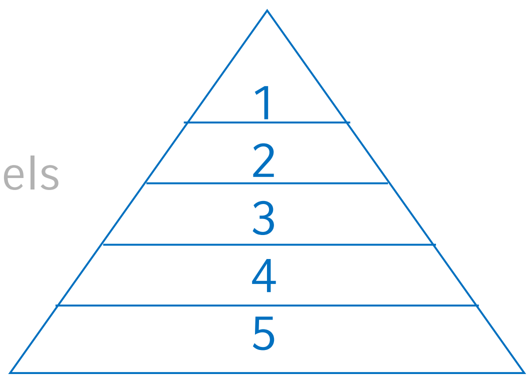
  - Fails for counter examples: 20 distinct years, 12 months, 7 days_of_the_week but not „year < month < days_of_the_week" with the latter on top

Summarization-based Aggregation

Data generalization:

- A process which abstracts a large set of task-relevant data in a database from low conceptual levels to higher ones.

Conceptual levels



<u>example:</u>
all
federal states
states
counties
cities

- Approaches:
  – Data-cube approach (OLAP / Roll-up) → manual
  – Attribute-oriented induction (AOI) → automated

Basic OLAP Operations

- *Roll up:* summarize data
  - by climbing up hierarchy or by dimension reduction
- *Drill down:* reverse of roll-up
  - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- *Slice and dice:*
  - selection on one (slice) or more (dice) dimensions
- *Pivot (rotate):*
  - reorient the cube, visualization, 3D to series of 2D planes
- Other operations
  - *drill across:* involving (across) more than one fact table
  - *drill through:* through the bottom level of the cube to its back-end relational tables (using SQL)

Example: Roll up / Drill down

Query:

SELECT *
FROM facts
WHERE condition
(table: results)

aggregating values

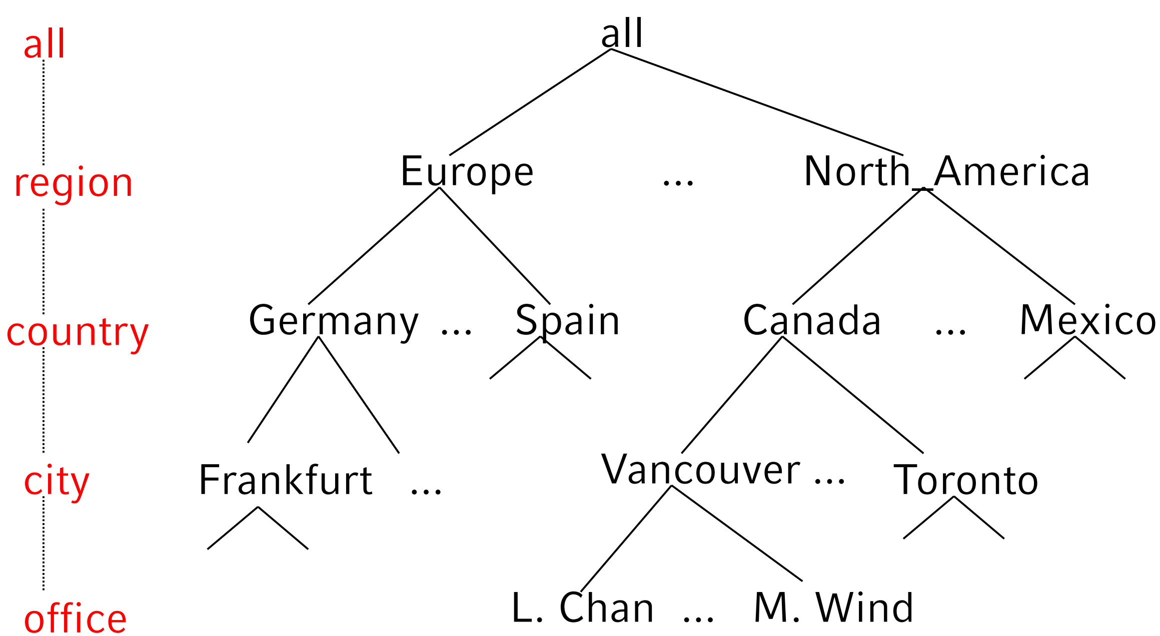
Roll-up:

SELECT …
FROM results
GROUP BY country, year

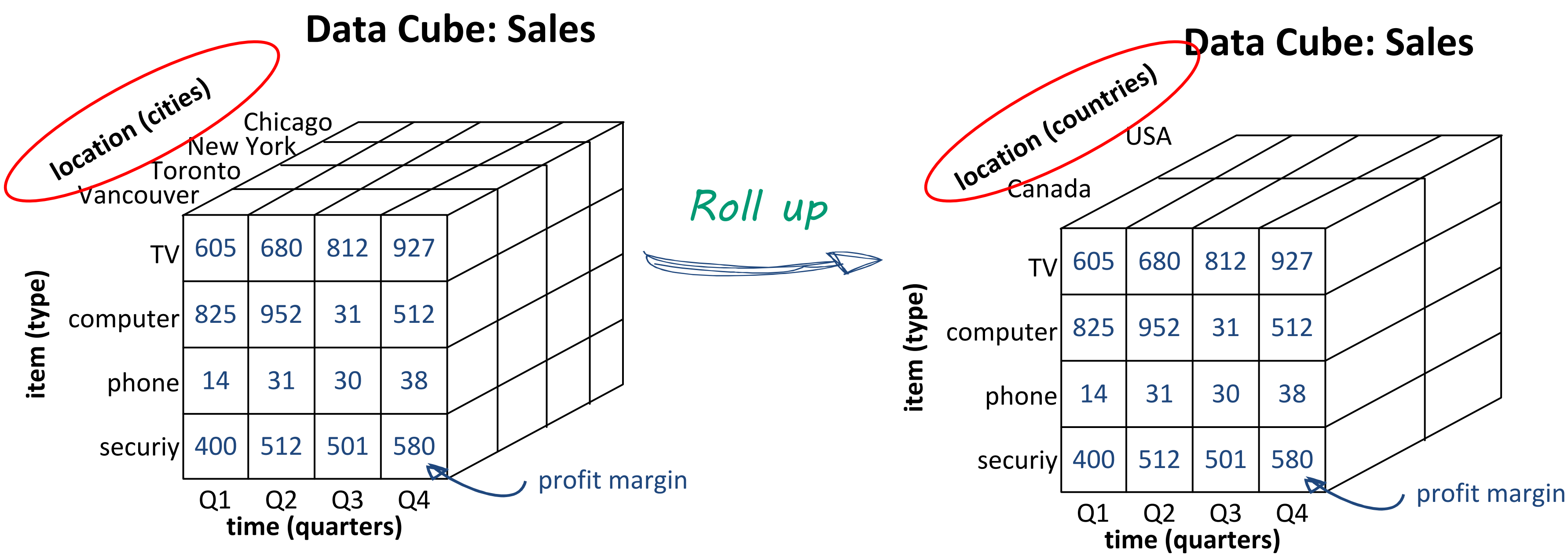
expand results

Drill down:

GROUP BY city, year

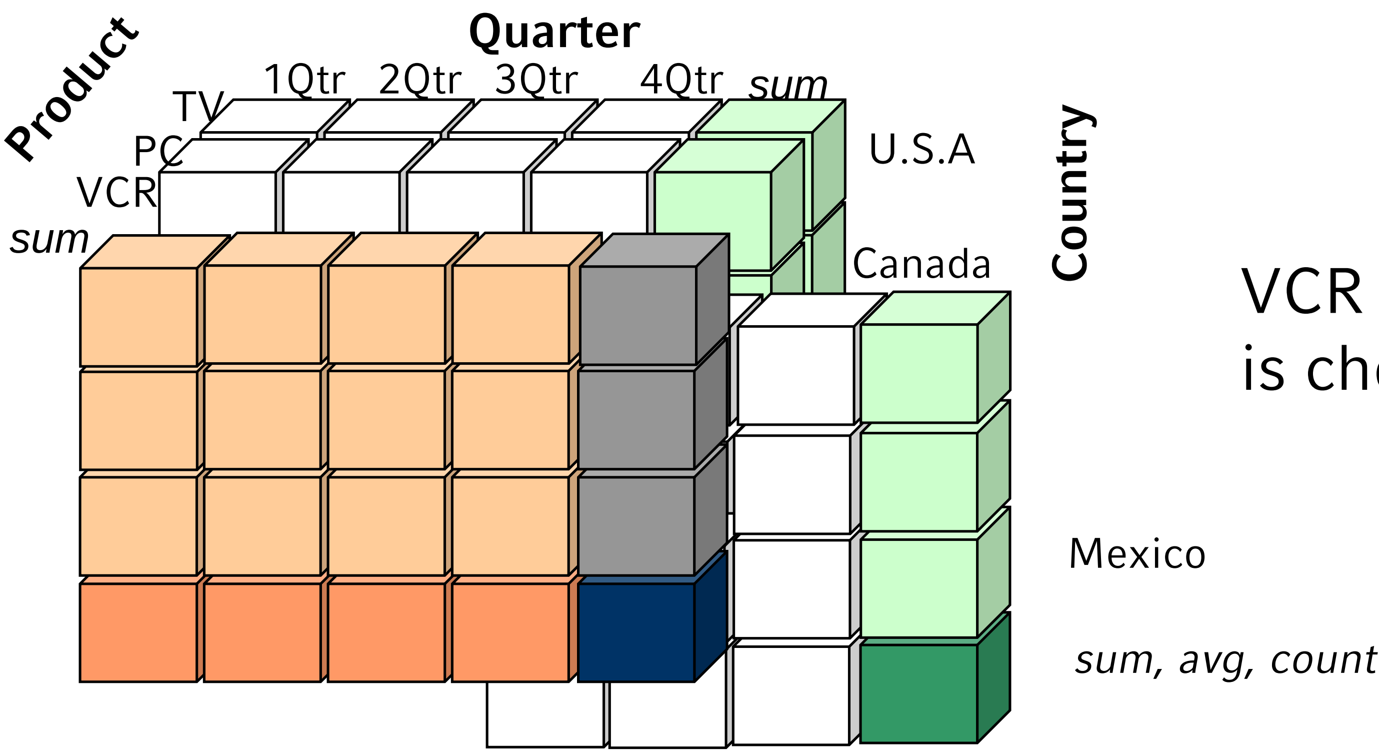


Example: Roll up in a data cube



Example: Slice operation

SELECT income
FROM time t, product p, country c
WHERE p.name = 'VCR'

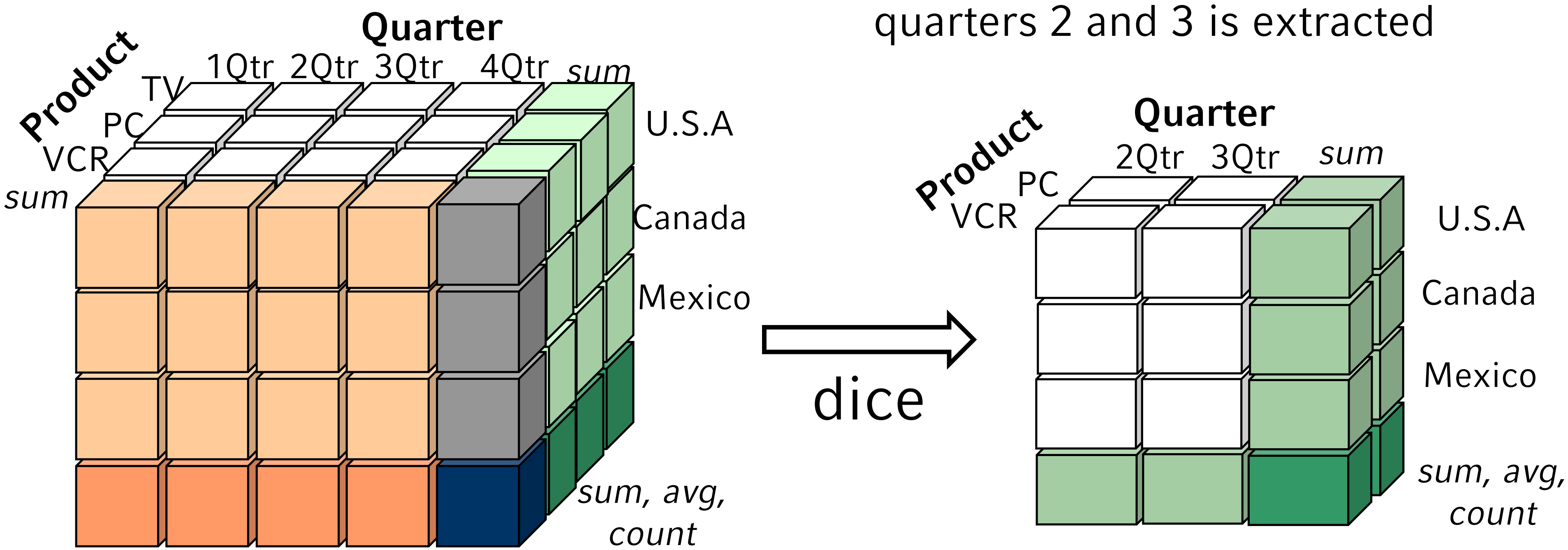


VCR dimension is chosen

Example: Dice operation

SELECT income
FROM time t, product p, country c
WHERE p.name = 'VCR' OR p.name = 'PC'
AND t.quarter BETWEEN 2 AND 3

sub-data cube over PC, VCR and quarters 2 and 3 is extracted



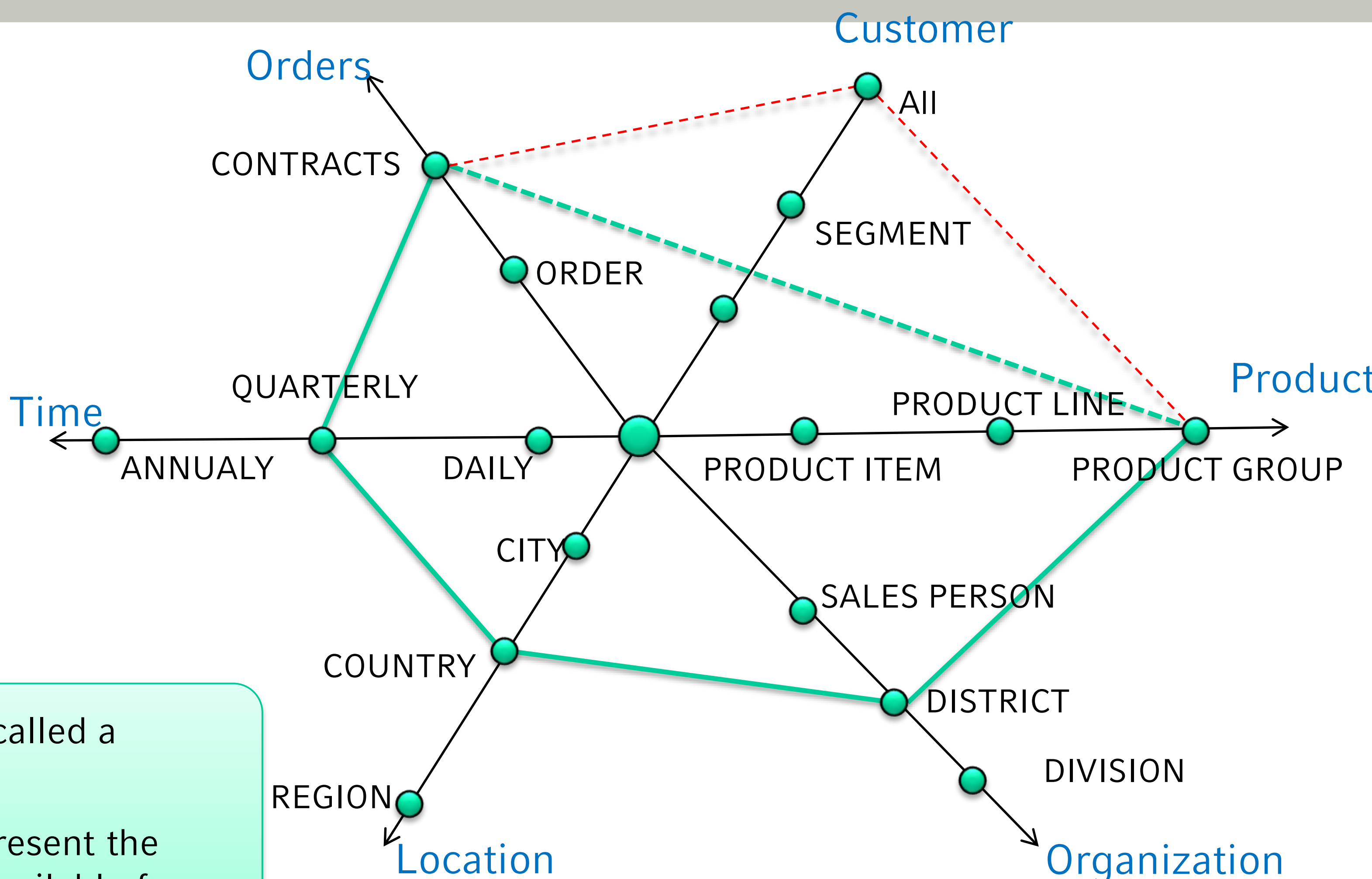
Example: Pivot (rotate)

Quarter 1			Quarter 2			Quarter 3		
TV	PC	VCR	TV	PC	VCR	TV	PC	VCR
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

Pivot (rotate)

TV			PC			VCR		
Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

Specify generalization by a star-net



Each circle is called a <u>footprint</u>

Footprints represent the granularities available for OLAP operations

Discussion of OLAP-based generalization

- Strength
  - Efficient implementation of data generalization
  - Computation of various kinds of measures
    - e.g., count, sum, average, max
  - Generalization (and specialization) can be performed on a data cube by *roll-up* (and *drill-down*)
- Limitations
  - handle only dimensions of *simple nonnumeric data* and measures of *simple aggregated numeric values*.
  - Lack of intelligent analysis, can't tell which dimensions should be used and what levels should the generalization reach

Attribute-Oriented Induction (AOI)

- Apply aggregation by merging identical, generalized tuples and accumulating their respective counts.
- Data focusing: task-relevant data, including dimensions, and the result is the *initial relation*.
- Generalization Plan: Perform generalization by either attribute removal or attribute generalization:

a) Attribute-removal: remove attribute *A* if:

1) there is a large set of distinct values for *A* but there is no generalization operator (concept hierarchy) on *A*, or

2) *A*'s higher level concepts are expressed in terms of other attributes (e.g. *street* is covered by *city*, *province_or_state*, *country*).

b) Attribute-generalization: if there is a large set of distinct values for *A*, and there exists a set of generalization operators (i.e., a concept hierarchy) on *A*, then select an operator and generalize *A*.

Attribute Oriented Induction: Example

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-81	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-80	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-75	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci, Eng, Biz	Country	Age range	City	Removed	Excl, VG,..

- *Name:* large number of distinct values, no hierarchy—removed.
- *Gender:* only two distinct values—retained.
- *Major:* many values, hierarchy exists—generalized to Sci.,Eng.,Biz.
- *Birth_place:* many values, hierarchy—generalized, e.g., to country.
- *Birth_date:* many values—generalized to age (or age_range).
- *Residence:* many streets and numbers—generalized to city.
- *Phone number:* many values, no hierarchy—removed.
- *Grade_point_avg* (GPA): hierarchy exists—generalized to good….
- *Count*: additional attribute to aggregate base tuples

Generalized Relation: Example

Initial Relation / **Plan**

Name	Gender	Major	BirthPlace	BirthDate	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver, BC, Canada	8-12-81	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-80	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-75	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Remove	Retain	Sci,Eng,Biz	Country	Age range	City	Remove	Excel, VG, …

Prime Generalized Relation

Gender	Major	BirthRegion	AgeRange	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

Crosstab for Generalized Relation

Gender \ Birth_Region	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

Attribute Generalization Control

- Problem: How many distinct values for an attribute?
  - *overgeneralization* (values are too high-level) or
  - *undergeneralization* (level not sufficiently high)
  - both yield tuples of poor usefulness.

- Two common approaches
  - Attribute-threshold control:
    default or user-specified, typically 2 - 8 values
  - Generalized relation threshold control:
    control the size of the final relation/rule, e.g., 10 - 30

Strategies to select the next attribute for generalization

- Aiming at minimal degree of generalization
  - Choose attribute that reduces the number of tuples the most.
  - Useful heuristics: choose attribute a_i with highest number m_i of distinct values.
- Aiming at similar degree of generalization for all attributes
  - Choose the attribute currently having the least degree of generalization
- User-controlled
  - Domain experts may specify appropriate priorities for the selection of attributes

Overview

- Data representation
  - Data types
  - Comparison and similarity
  - Data visualization

- Data reduction
  - Aggregation
  - Generalization