Ludwig-Maximilians-Universität München
Lehrstuhl für Datenbanksysteme und Data Mining
Prof. Dr. Thomas Seidl

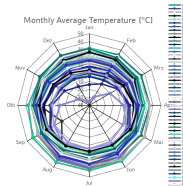
Knowledge Discovery and Data Mining I

Winter Semester 2018/19

Agenda

Data Visualization

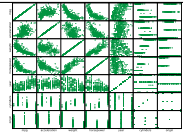
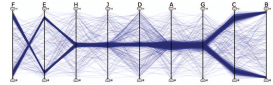
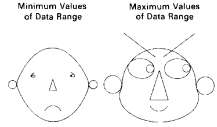
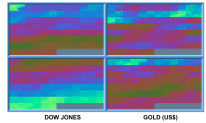
- Patterns in large data sets are hardly perceived from tabular numerical representations
- Data visualization transforms data in visually perceivable representations ("a picture is worth a thousand words")
- Combine capabilities:
  - Computers are good in number crunching (and data visualization by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



Monthly average temperature [°C]

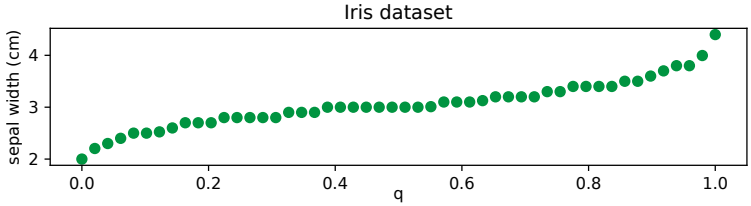
Städte Ø	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
Abu Dhabi	25	27	31	36	40	41	42	43	42	37	31	27
Acapulco	32	31	32	32	33	33	33	33	33	33	32	32
Anchorage	-4	-2	0	6	13	17	18	17	13	5	-3	-5
Antalya	15	16	19	22	27	32	35	36	32	27	21	17
Athen	13	14	17	20	26	30	34	34	29	24	18	14
Atlanta	11	13	18	23	26	30	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
Heraklion (Kreta)	15	16	18	20	24	27	30	30	27	24	20	17
Hongkong	19	20	23	26	30	32	33	33	32	30	25	21
Honolulu	26	26	27	27	28	30	30	31	30	30	28	27
Houston	16	19	23	27	30	33	34	35	32	28	21	17
Irkutsk	-14	-9	1	9	16	23	24	21	16	7	-4	-13
Istanbul	9	9	13	17	23	27	30	30	26	20	15	11
Jakutsk (Nordostsibirien)	-35	-28	-10	3	14	23	26	21	11	-3	-25	-34
Johannesburg	25	25	24	22	20	17	17	20	24	25	25	25
Kairo	19	20	24	27	32	35	35	35	34	30	25	20
Kapstadt	27	27	26	24	21	18	18	18	19	22	24	26
Kathmandu	18	21	25	28	28	29	28	28	28	26	23	20
Larnaka (Zypern)	17	18	20	23	26	31	33	34	31	28	23	19
Las Palmas	21	20	22	23	24	25	27	28	28	27	24	22
Las Vegas	15	16	23	26	31	38	40	39	35	27	20	14
Lhasa	9	10	13	17	21	24	23	22	21	17	13	10
Lima	26	26	27	24	21	20	19	18	19	20	22	24
Lissabon	14	15	18	20	23	27	28	29	27	22	17	15

Data Visualization Techniques

Type	Idea	Examples
Geometric	Visualization of geometric transformations and projections of the data	 Scatterplots  Parallel Coordinates
Icon-Based	Visualization of data as icons	 Chernoff Faces  Stick Figures
Pixel-oriented	Visualize each attribute value of each data object by one coloured pixel	 Recursive Patterns
Other		Hierarchical Techniques, Graph-based Techniques, Hybrid-Techniques, ...

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

Quantile Plot



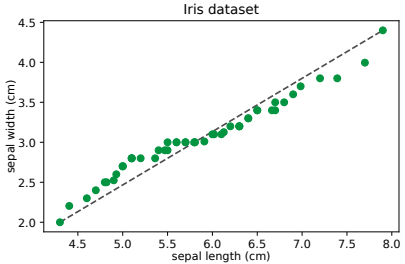
Characteristic

The p-quantile x_p is the value for which the fraction p of all data is less than or equal to x_p.

Benefit

Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)

Quantile-Quantile (Q-Q) Plot



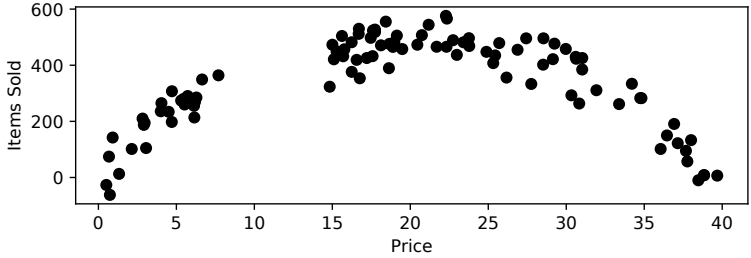
Characteristic

Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

Benefit

Allows the user to compare to distributions against each other.

Scatter Plot



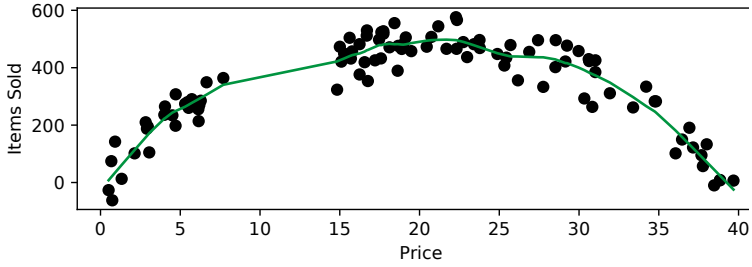
Characteristic

Each pair of values is treated as a pair of coordinates and plotted as points in the plane.

Benefit

Provides a first look at bivariate data to see clusters of points, outliers, etc.

Loess Curve



Characteristic

Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression.

Benefit

Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence.

Scatterplot Matrix

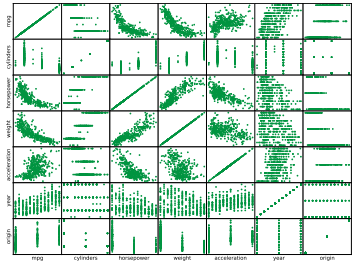
Characteristic

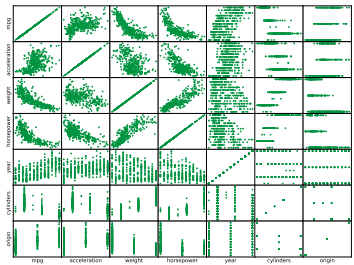
Matrix of scatterplots for pairs of dimensions

Ordering

Ordering of dimensions is important:

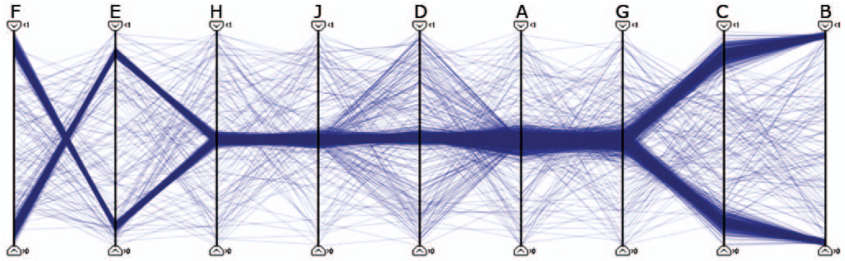
- Reordering improves understanding of structures and reduces clutter
- Interestingness of orderings can be evaluated with quality metrics (e.g. Peng et al.)





Clutter Reduction in Multi-Dimensional Data Visualizazion Using Dimension Reordering, IEEE Symp. on Inf. Vis., 2004.

Parallel Coordinates



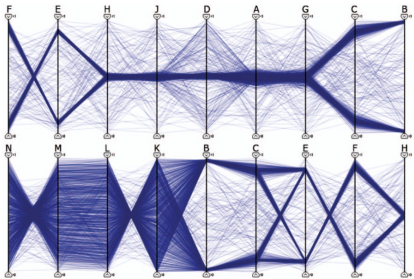
Characteristics

- d-dimensional data space is visualised by d parallel axes
- Each axis is scaled to min-max range
- Object = polygonal line intersecting axis at value in this dimension

Parallel Coordinates

Ordering

- Again, the ordering of the dimensions is important
- Quality metric for interestingness of ordering
- Quality or interestingness of orderings depends on what you want to visualize



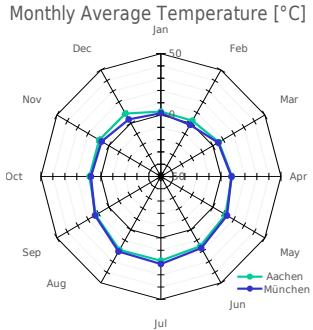
- Visualize clusters
- Visualize correlations between dimensions

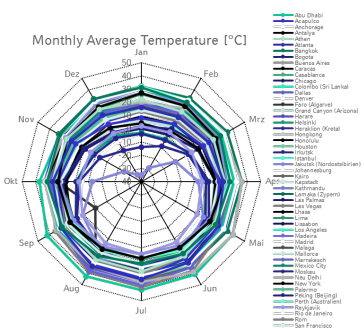
Bertini et al., Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, Trans. on Vis. and Comp. Graph., 2011.

Spiderweb Model

Characteristics

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





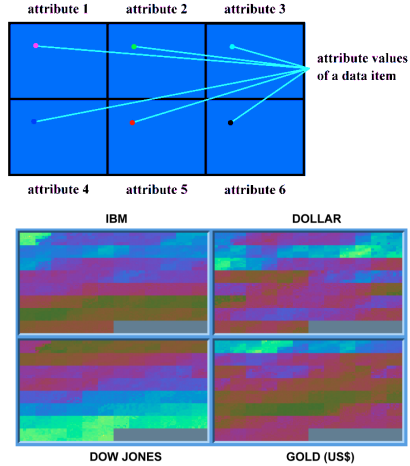
Pixel-Oriented Techniques

Characteristics

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

How to arrange the pixel ordering?

One strategy: Recursive Patterns iterated line and column-based arrangements



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

Chernoff Faces

Characteristics

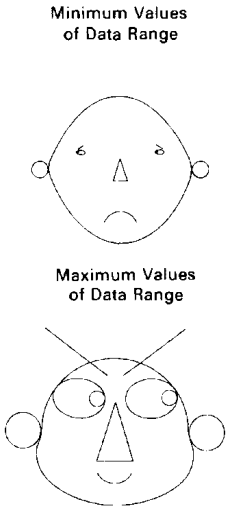
Map d-dimensional space to facial expression, e.g. length of nose = dim 6; curvature of mouth = dim 8

Advantage

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

Disadvantages

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



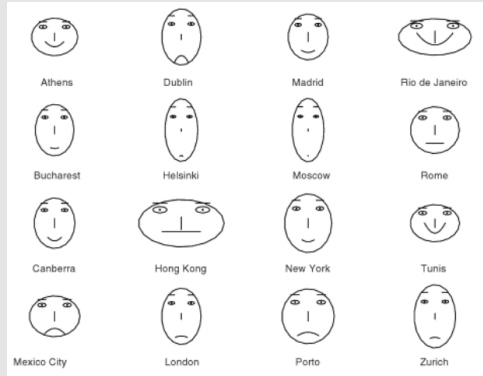
Figures taken from Mazza, Introduction to Information Visualization, Springer, 2009.

Chernoff Faces

Example: 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.

Chernoff Faces

Example: Finance 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.