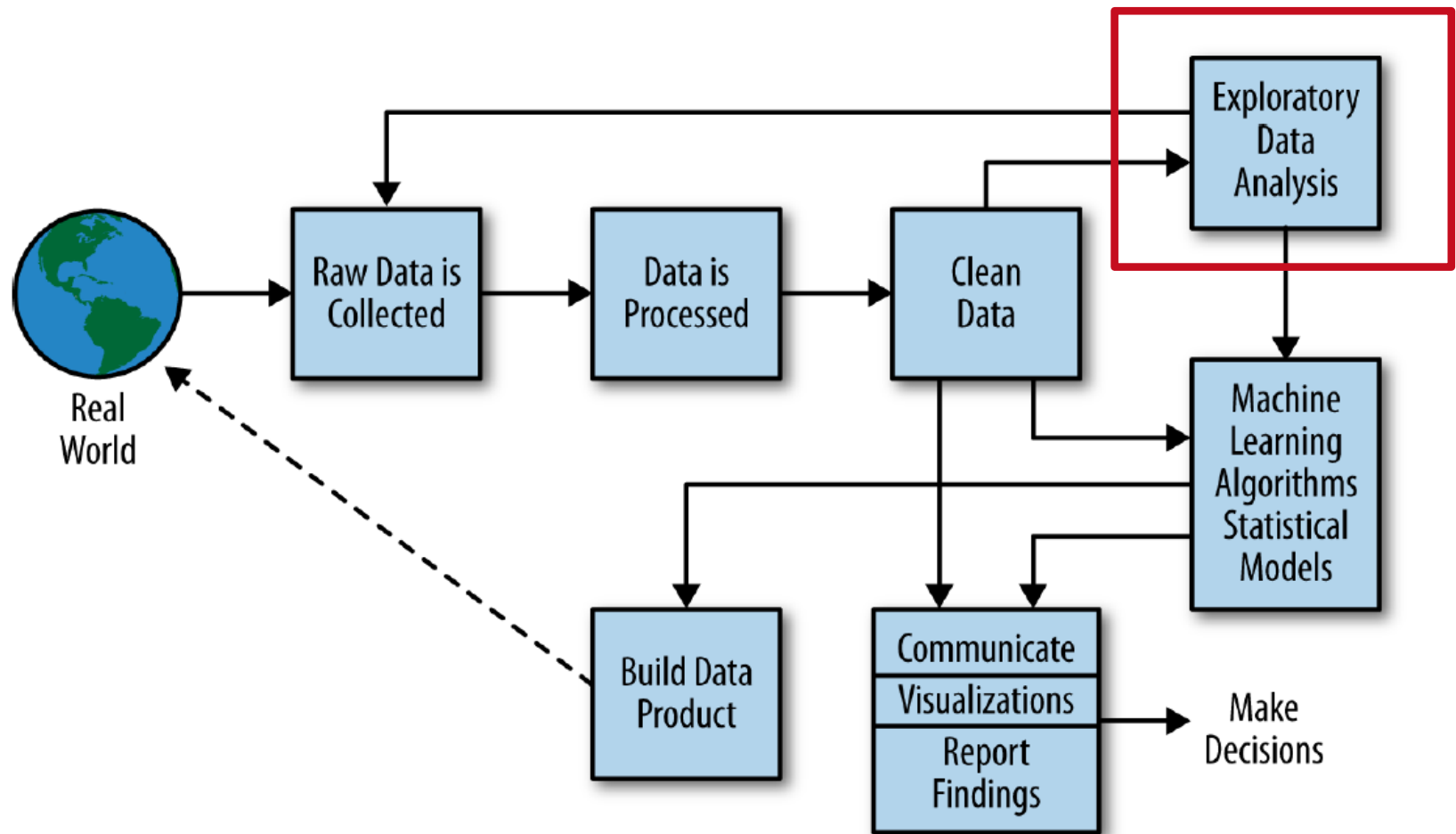




Exploratory Data Analysis

The Data Science Process



Source: Schutt & O'Neil, 2014

Exploratory (Spatial) Data Analysis

„Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there“

John Tuckey, mathematician at Bell Labs and founder of EDA

- The first step towards a data science project / building a machine learning model
- This critical part in the data science process, often involves more than 70% of the time invested in the project and includes:
 - **data manipulation:**
 - solve various origin and scale problems in space and time
 - data filtering, cleaning, outlier detection
 - data transformations (reshaping)
 - import/export to common formats

Exploratory (Spatial) Data Analysis

- data plotting:
 - explore spatial data using (carto-)graphic (or other visual) representations
- descriptive statistics:
 - summarises the main properties of a set of values
 - main measures are those of central tendency and spread
 - it can also quantify the (linear) dependency of two related data sets, e.g., temperature and solar radiation at the same time in one location

Ways to Analyze Data

Univariate:

- each variable is analyzed separately:
data distribution, central value, and data spread/uncertainty

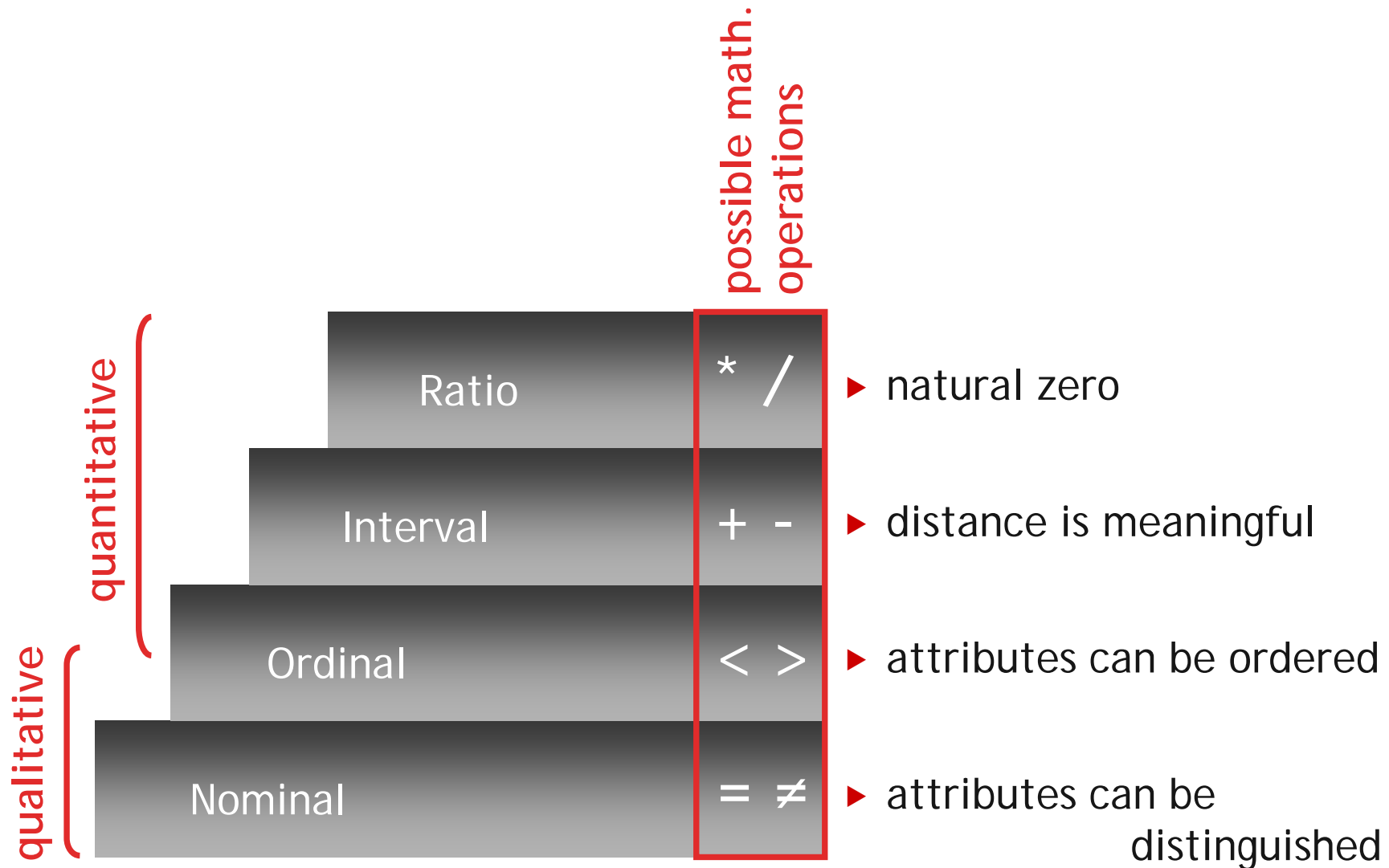
Bivariate:

- two variables are analyzed together to look for correlation or separation of data – regression

Multivariate:













- more than 2 variables are analyzed together.
Generally difficult to visualize the data and results

Types of Observation Variables



Nominal data \neq

- Classification according to type or quality
- Often labeled with numbers or letters, but no ranking implied!



























Point	airport 	town 	mine 	capital 
Line	river 	road 	boundary 	pipeline 
Area	orchard 	desert 	forest 	water 

Representation of nominal data

<http://www.geog.okstate.edu/users/Larson/home.htm>

Ordinal data < >

- Information about rank or hierarchy
- Possible to describe one item as smaller or larger than another
- Not possible to measure differences, because there are no specific values attached

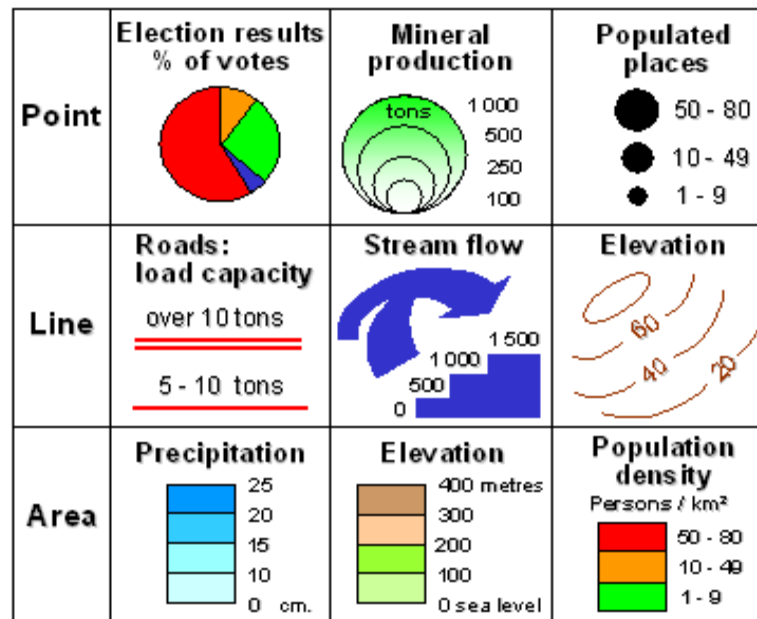
Point	Airports  international  national  regional	Oil well production  high  medium  low	Populated places  large  medium  small
Line	Roads  expressway  major  local	Drainage  river  stream  creek	Boundaries  international  provincial  county
Area	Soil quality  good  fair  poor	Cost of living  high  medium  low	Industrial regions  major  minor

Representation of ordinal data

<http://www.geog.okstate.edu/users/Larson/home.htm>

Interval data + -

- Include numerical values
- Information can be arranged along a scale
→ distance/difference can be calculated

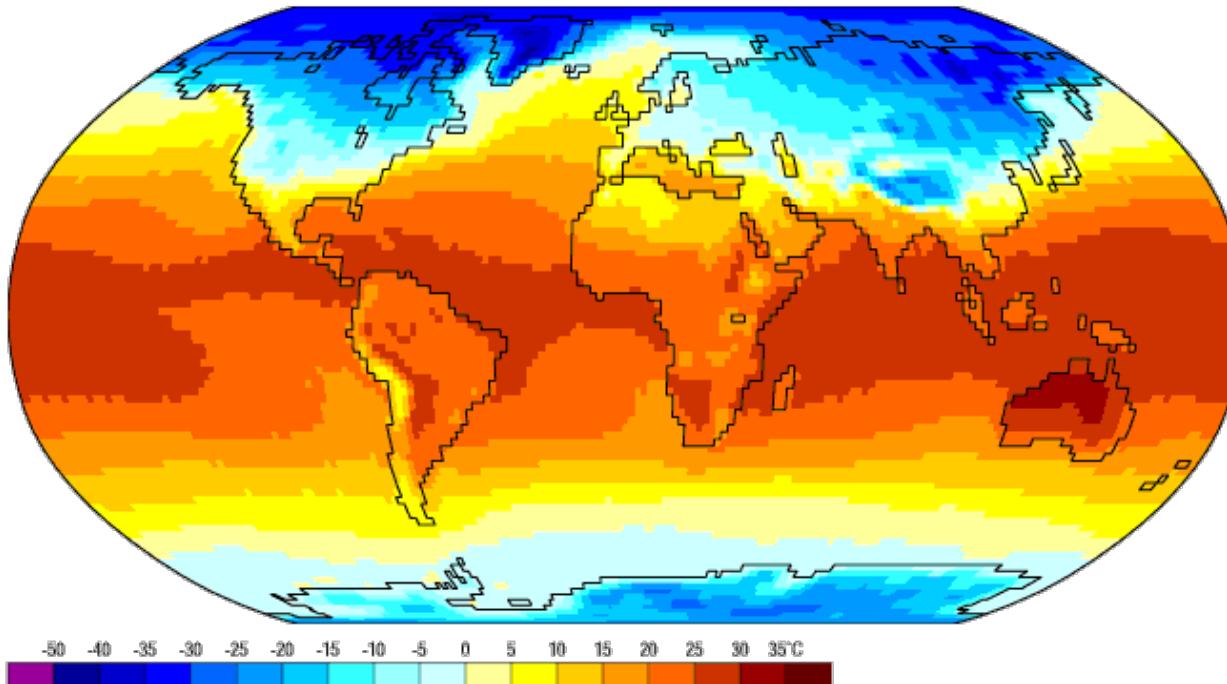


Representation of interval and ratio data

Interval data represented by area symbols

Air Temperature

Dec



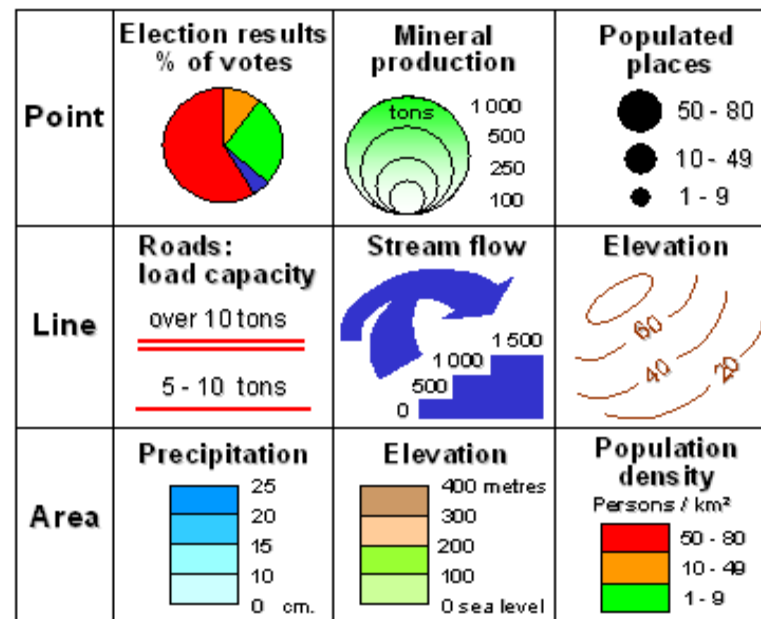
Data: NCEP/NCAR Reanalysis Project, 1959-1997 Climatologies
Animation: Department of Geography, University of Oregon, March 2000

Annual worldwide temperature distribution

<http://www.learner.org/jnorth/mclass/spring2006/Peek033106.html>

Ratio data * /

- Like interval data, but:
- There is a natural zero
→ data can be expressed as ratios



Representation of interval and ratio data

Types of Observation Variables

Categorical data (qualitative data):

- **Nominal**: can be named, e.g. soil types
 - can only be **separated**, but not (uniquely) ranked (no intrinsic order): colors, names
 - can be **coded** as numbers (0,1, 2,...), but many numerical operations do not make sense
- **Ordinal**: can be ordered (and named) e.g. seismic scale, grades, sizes

Numerical data (quantitative data):

- **Interval**: can be subtracted (and ordered and named) → difference. e.g. integers in equally spaced interval
- **Ratio** : can be divided (and subtracted, ordered and named), e.g. amount of money you have in your pocket right now
 - the most informative scale

Categorization by Dimension

- One-Dimensional (18, 20, 43, 32, ...)
- Multi-dimensional ((1,4), (5,3), (6.2, 10.4), ...)
- High-dimensional (e.g. Time Series Data)
- No-dimensional (e.g. Protein Folding Structures)

- 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:

$$\text{count}(D_1 \cup D_2) = \text{count}(D_1) + \text{count}(D_2)$$

$$\text{sum}(D_1 \cup D_2) = \text{sum}(D_1) + \text{sum}(D_2)$$

$$\text{max}(D_1 \cup D_2) = \text{max}(\text{max}(D_1), \text{max}(D_2))$$

- 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.

Examples:

$$avg() = sum() / count()$$

$$standard_deviation()$$

- if 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)
$$\text{median}(D1 \cup D2) \neq \text{simple_function}(\text{median}(D1) + \text{median}(D2))$$
- mode: value that appears most often in a set of values
- rank: k-smallest / k-largest value (cf. quantiles, percentiles)

Statistical Plots: Sample Histogram

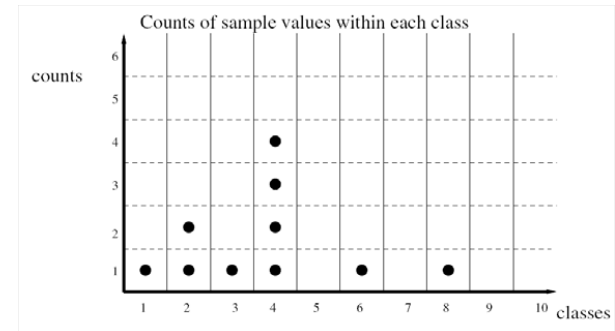
Setting: Consider 10 hypothetical sample values:

4	1	3	8	4	4	2	4	6	2
---	---	---	---	---	---	---	---	---	---

Estimated relative frequency table:

$$\hat{f}_k = (\# \text{ of data in } k\text{-th class}) / (\text{total } \# \text{ of data})$$

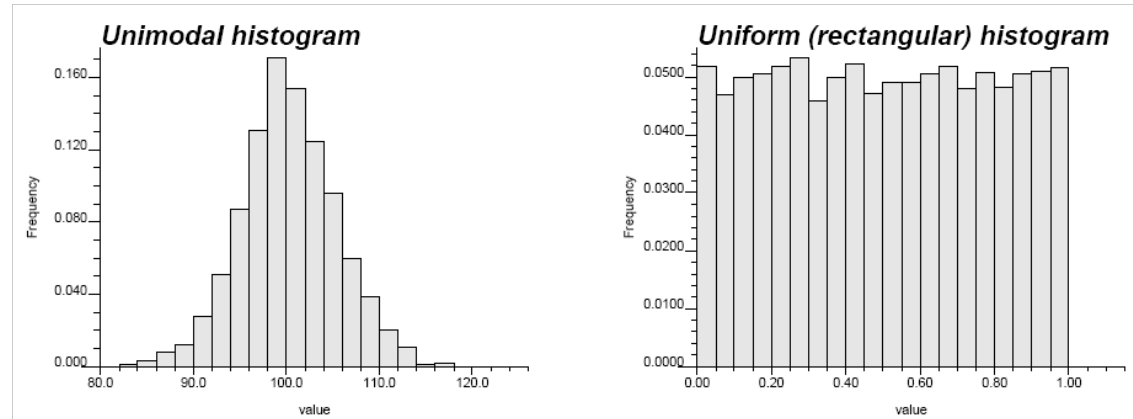
x_k	1	2	3	4	5	6	7	8	9
\hat{f}_k	0.1	0.2	0.1	0.4	0.0	0.1	0.0	0.1	0.0



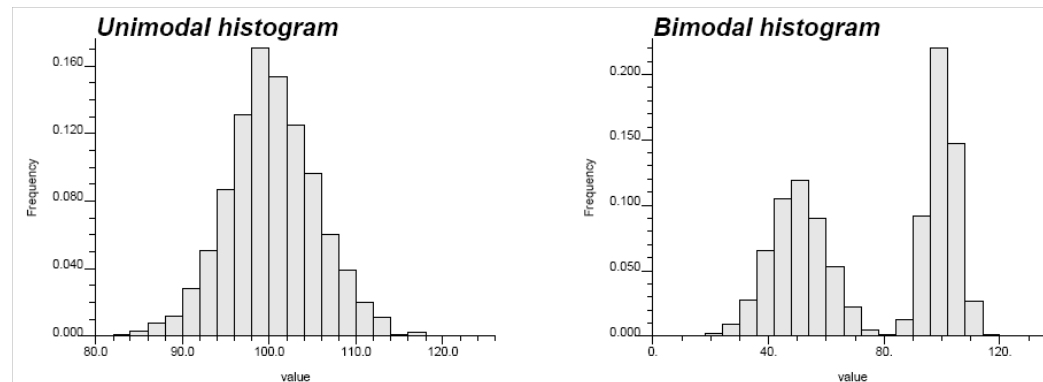
- histogram **shape** depends on **number and width of bins** (“classification”):
 - use non-overlapping equal intervals with simple bounds
 - Rule of thumb for number of classes: $5 \times \log_{10}(\# \text{ of data})$
 - For a density histogram, total area of bars = 1

Histogram Shape Characteristics

Peaked or not:

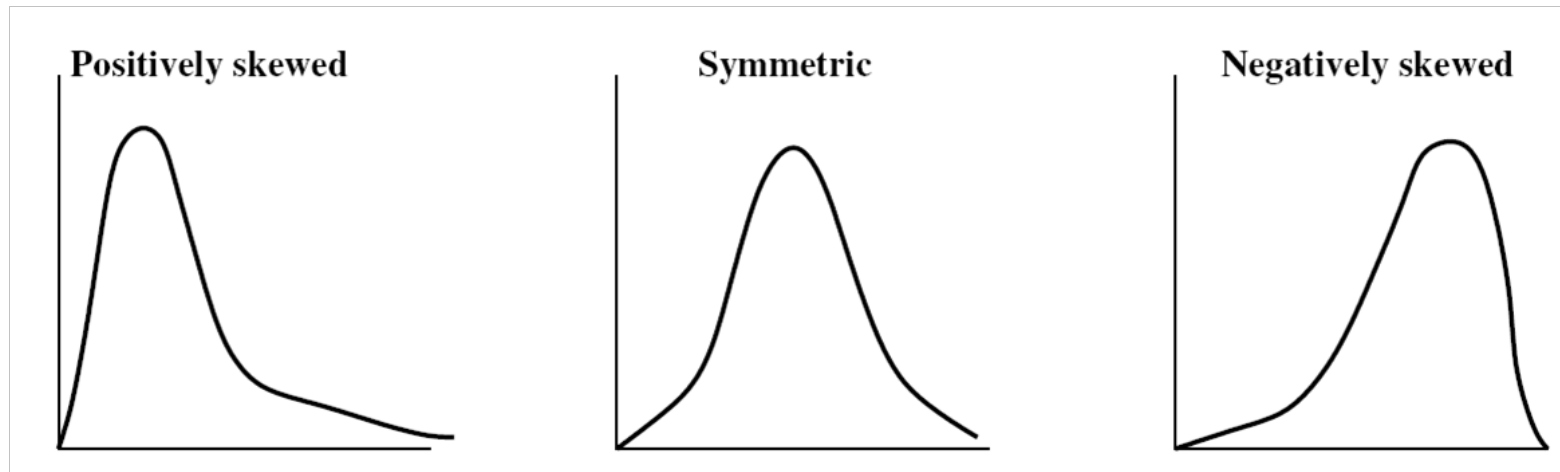


Number of peaks:



Histogram Shape Characteristics

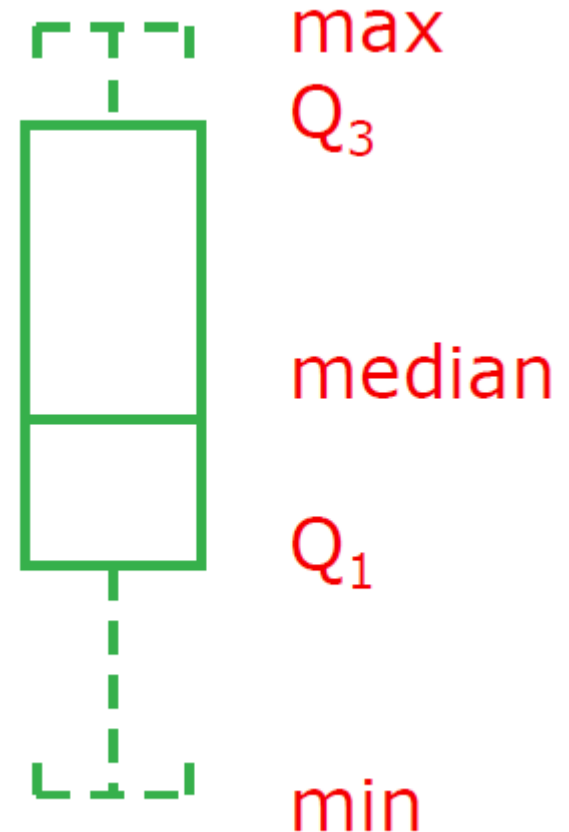
Symmetric or skewed:



$$\text{coefficient of skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - m)^3}{\sigma^3}$$

Plots for Numerical Data

- **Post plot (point cloud)** – for a first overview of the data
- **Boxplot** - shows **characteristic values** of the data:
 - the box covers the **central half** of the values (1st quartile to 3rd quartile),
 - the thick line is the **median**.
 - the whiskers reach out to maximum and minimum unless these are very extreme then they are shown as outliers.
- delete outliers to get a more robust result, but you may lose important information.



Univariate statistics

- describe the distributions of **individual** variables (or one variable at a time in a set of multivariate data)
- Is not sufficient to describe spatial pattern, because the spatial arrangement of attribute values matters

So ...

- the *relationships* and *dependencies* between variables are very important in most (earth) science data sets
- for **comparing the distributions** of paired data (we have two measurements per observation, e.g., porosity and location)
 - Histograms + summary statistics → reveals only gross differences
 - Two very similar distributions → not helpful
- ⇒ suitable visual comparisons + **bivariate analysis**

Scatterplots

- offer qualitative information for how **two variables are related**
- useful also for error checking:
make aberrant data obvious

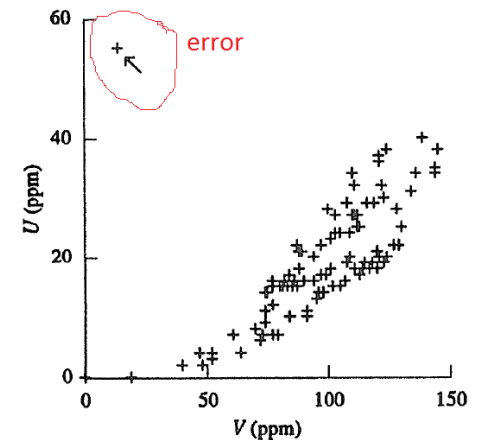
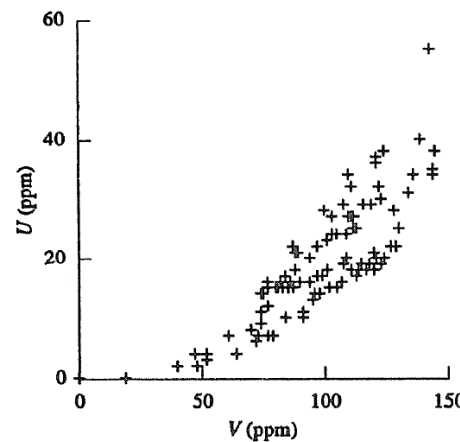
Setting:

Data pairs of two variables U & V ,
measured at N sampling units
there are N pairs of attribute values
 $\{(x_n, y_n), n = 1, \dots, N\}$

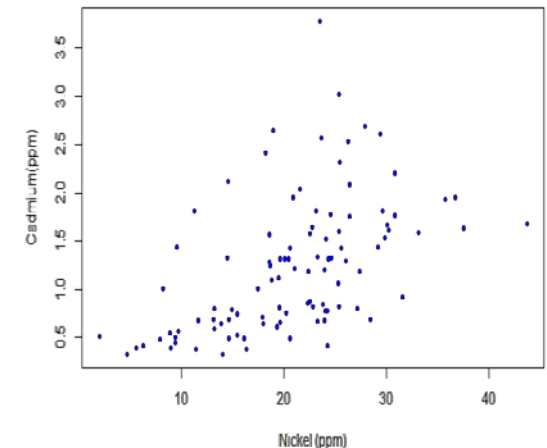
Scatterplot:

graph of V - versus U -values in the
bivariate attribute space:

- x -values as coordinates in horizontal axis
- y -values serve as coordinates in vertical axis
- n -th point in scatter-plot has coordinates (x_n, y_n)



Scatterplot of a 2D feature space



Key feature in a scatterplot → correlation (association or trend) between X and Y

There are three patterns which can be observed on a scatterplot:

– positive correlation

- Higher (lower) X values are associated with higher (lower) Y values
- E.g.: porous rocks → porosity and permeability

多孔性 浸透性

– negative correlation

- Higher (lower) X values are associated with lower (higher) Y values
- E.g.: geological data sets → concentration of major elements are often negatively correlated (dolomitic limestone – increase in amount of calcium results in a decrease in the amount of magnesium)

– no correlation (uncorrelated)

- An increase in one variable has no apparent effect on the other

Correlation Coefficient

- most frequently used to summarise the relationship between two attributes
- quantifies numerically the trend in a bivariate scatterplot
- provides a measure of the **LINEAR** relationship of two variables*

$$\rho = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y}$$

Diagram annotations:

- Red arrows point from the word "mean" to m_x and m_y in the numerator.
- A blue box surrounds the numerator $\frac{1}{n} \sum_{i=1}^n (x_i - m_x)(y_i - m_y)$, with the word "covariance" written in blue to its right.
- A green arrow points from the words "standard deviations" to $\sigma_x \sigma_y$ in the denominator.

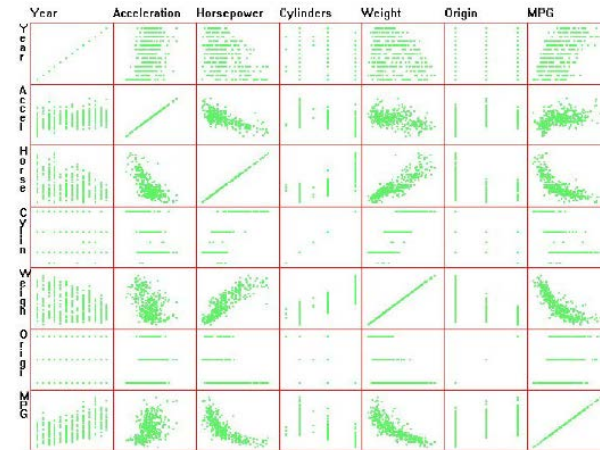
- **Covariance:** $C_{XY} = \frac{1}{n} \sum_{i=1}^n (x_i - m_x)(y_i - m_y)$

- Average of the data deviations from their means
- Depends on the magnitude of the data values
- Division by standard deviation $\rightarrow \rho$ values are between -1 and 1
- Strongly influenced by a few aberrant pairs

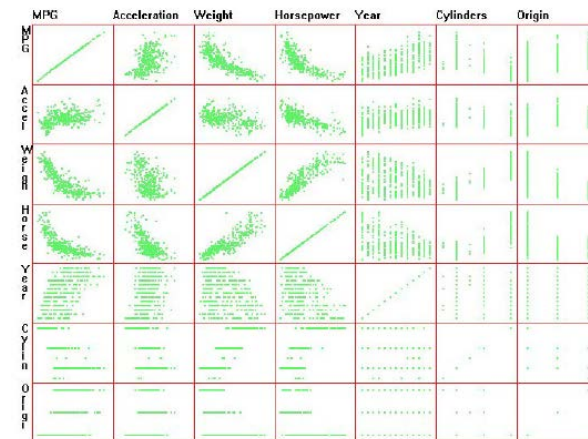
*no linear relationship
 \Rightarrow poor summary statistics

Scatterplot Matrix

- Matrix of scatterplots (x-y diagrams) of 2d-dimensional data
- Ordering of dimensions is important
- Dimension re-ordering
 - The interestingness of different orderings can be evaluated with quality metrics (e.g. Peng et al.)
 - Reduces clutter
 - Better visualization and understanding of data



(a)



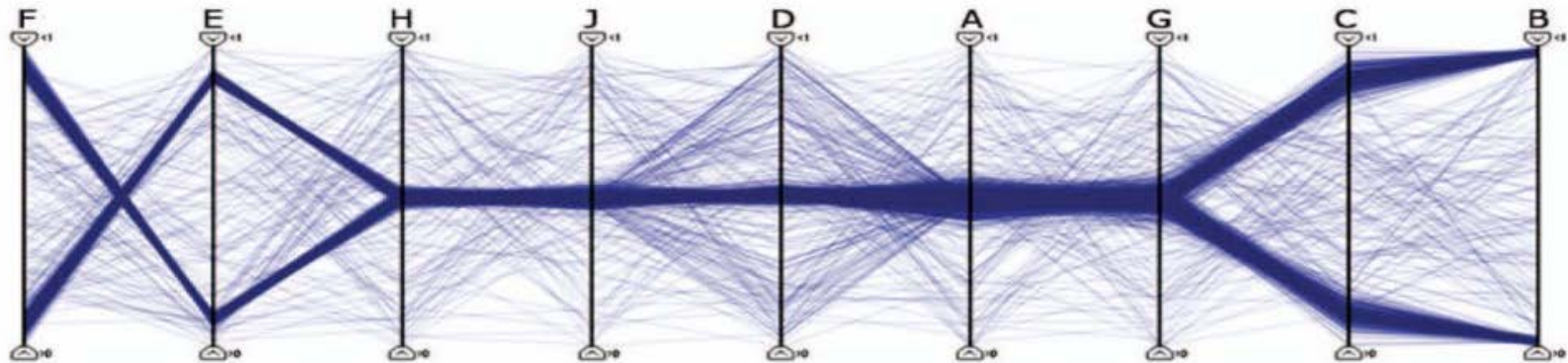
From Peng et al., Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering, IEEE Symp. on Inf. Vis., 2004

Parallel Coordinates

A d -dimensional data space is visualized by d parallel axes

Each axis is scaled to the min-max range in the corresponding dimension

A data point 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



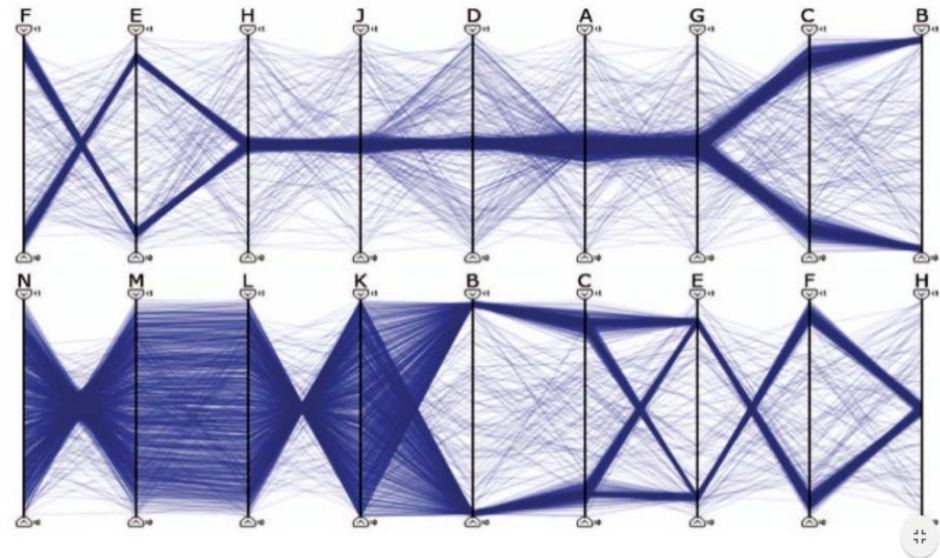
From Bertini et al., Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, Trans. on Vis. and Comp. Graph., 2011

Parallel Coordinates

- the ordering of the dimensions matters!
- 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



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