DATA MINING

PREPROCESING AND VISUALIZATION

ABOUT ME

Pablo González de Prado Salas

gonzalezdepradosalas.weebly.com

LECTURE OVERVIEW

1. Knowing your data
2. Measuring data
3. Visualizing data
4. Cleaning data
5. Reducing data
6. Transforming data

KNOWING YOUR DATA

DATA OBJECTS

Name	Address	Age
John Doe	Happy Road 2	2
Jane Doe	Happy Road 2	25
Joan Petersen	Spring Way 42	63

- A **data set** is made of **data objects,** also known as:
  - Samples, examples, instances, data points, data tuple…
- Data objects **describe** the entities in a data set
- Each row in a data base is a data object

ATTRIBUTES

Name	Address	Age
John Doe	Happy Road 2	2
Jane Doe	Happy Road 2	25
Joan Petersen	Spring Way 42	63

- An **attribute** is a **data field** and describe a characteristic of a data object
- Known as dimension, feature and variable
- Many types!

ATTRIBUTES

Name	Address	Age
John Doe	Happy Road 2	2
Jane Doe	Happy Road 2	25
Joan Petersen	Spring Way 42	63

Data object

Attribute 1

Attribute 2

Attribute 3

ATTRIBUTE TYPES

- Qualitative
    - Nominal
    - Binary
    - Ordinal
- Quantitative
- Not necessarily exclusive!

NOMINAL ATTRIBUTES

Name	Age	Position
John Doe	2	None
Jane Doe	25	Student
Joan Petersen	63	Professor

- "Of, relating to, or constituting a name"
- No meaningful order between possible values of the attribute
- Known as categorical or enumeration

*Merriam-Webster Dictionary

NOMINAL ATTRIBUTES

Name	**Age**	**Position**
John Doe	2	None
Jane Doe	25	Student
Joan Petersen	63	Professor

- Can be encoded using integers:
  - E.g., none = 0, student = 1, professor = 2
- When encoded as integers, can we use nominal attributes quantitatively?
  - E.g., calculate differences, averages

NOMINAL ATTRIBUTES

- When encoded as integers, can we use nominal attributes quantitatively?
  - E.g., calculate differences, averages
- No!

	Name	Age	Position
	John Doe	2	None (0)
	Jane Doe	25	Student (1)
	Joan Petersen	63	Professor (2)
Average:	**??**	**30**	**Student (1)**

(0 + 1 + 2) / 3

NOMINAL ATTRIBUTES

Nominal attributes should never be used quantitatively

	Name	Age	Position
	John Doe	2	None (0)
	Jane Doe	25	Student (1)
	Joan Petersen	63	Professor (2)
Average:	**??**	**30**	**Student (1)**

(0 + 1 + 2) / 3

BINARY ATTRIBUTES

Name	Likes Coke	Flu-Positive
John Doe	0	1
Jane Doe	0	0
Joan Petersen	1	1

- Nominal attribute that can **only** take **two** possible values
  - 0 usually means attribute absence
  - 1 usually means attribute presence
- Known as **Boolean** when 1/0 correspond to **true/false**

BINARY ATTRIBUTES

Name	Likes Coke	Flu-Positive
John Doe	0	1
Jane Doe	0	0
Joan Petersen	1	1

- Symmetric binary

  Both values equally important

- Asymmetric binary

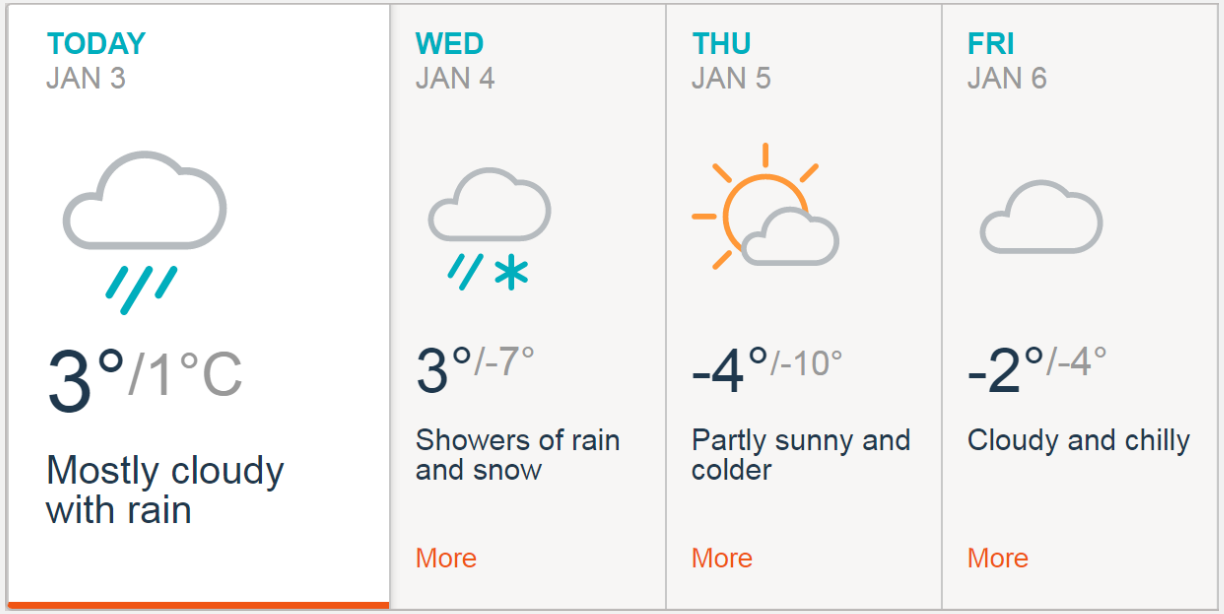
  Convention: **most relevant** outcome takes value 1

ORDINAL ATTRIBUTES

- Similar to nominal, but possible values have a ranking
- Magnitude between elements not known

Drink	**Size**	Price $
Juice	small	1.50
Juice	large	2.50
Smoothie	medium	1.99
Smoothie	large	2.99

NUMERIC: INTERVAL-SCALED



- Numerical attributes measured on an equal-size scale
- Possible values have order (-1 < 2)
- **Differences** in values may be compared and quantified

NUMERIC: INTERVAL-SCALED

Date	**Forecast**	**Temperature (ºC)**
03/01/17	Rain	3º
04/01/17	Snow	3º
05/01/17	Sunny	–4º

6º is not two times 3º!

0º is not “null temperature”

NUMERIC: RATIO-SCALED

Date	Forecast	**Temperature (K)**
03/01/17	Rain	276º
04/01/17	Snow	276º
05/01/17	Sunny	268º

- Numeric attribute with zero-point
- May directly compare values (multiples, ratios)

2.9% decrease in temperature

NUMERIC: INTERVAL VS RATIO

Which example belongs to which category?

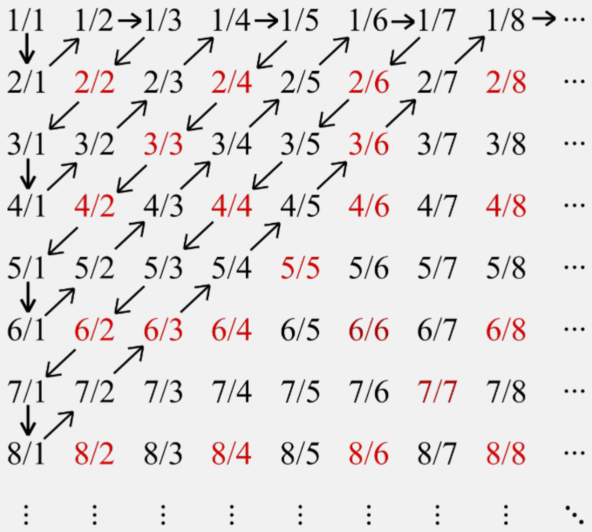
- Height
- X-axis position
- Calendar year
- Speed

DISCRETE VS CONTINUOUS ATTRIBUTES

Classify these examples:

- Drink Size
- Height
- Zip-code
- Speed
- Age

DISCRETE VS CONTINUOUS ATTRIBUTES



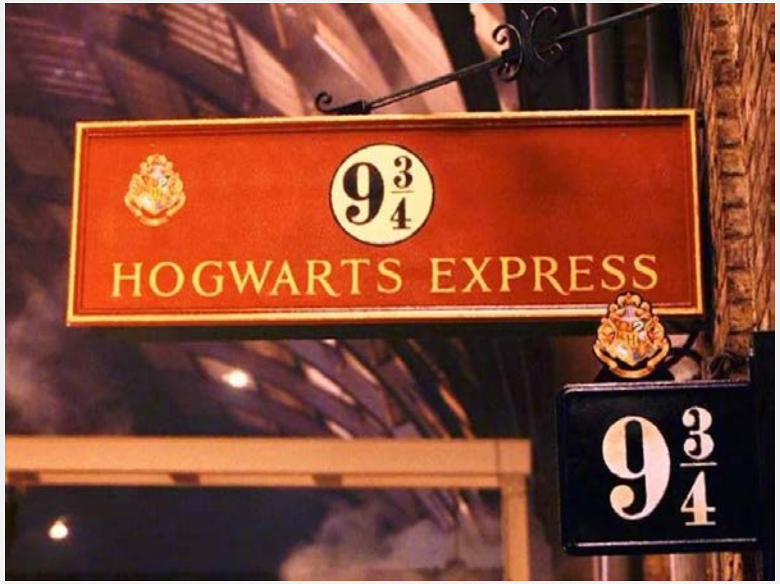
Technical **definition** may be **tricky**!

https://en.wikipedia.org/wiki/Continuous_and_discrete_variables

Are **rational numbers** continuous or discrete?

In practice, memory limitations mean no true continuous!

DISCRETE VS CONTINUOUS ATTRIBUTES



Are any intermediate values valid?

MEASURING DATA

CENTRAL TENDENCY

Where do most values fall?

- Mean
- Median
- Mode

CENTRAL TENDENCY—MEAN

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Most common measure for the “centre” of a data set

CENTRAL TENDENCY—MEAN

	Name	Age	Position
	John Doe	2	None
	Jane Doe	25	Student
	Joan Petersen	63	Professor
	Jerry Perry	53	Janitor
Average:	—	**35.75**	—

(2+25+63+53)/4 = 35.75

Any problems here?

CENTRAL TENDENCY—MEAN

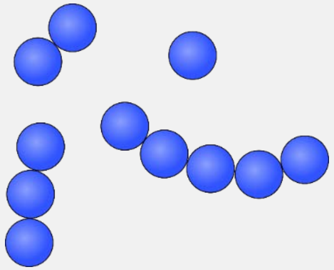
Problem: sensitivity to outliers

- Trimmed mean
  - After removing outliers
  - **Subjective**: careful with data *overcooking!*
- Weighted mean
  - Using weights for each value
  - Weights carry some **meaning**!

	Name	Age
	Johnny Doe	2
	Little Jane	5
	Jerry Small	3
	Old Samuel	93
Average:	—	25.75

CENTRAL TENDENCY—MEAN

Example from my past!



	Filament ID	Length
	Filament 1	2
	Filament 2	1
	Filament 3	5
	Filament 4	3
Average:	—	**2.75**
W. average:	—	**3.54**

Average filament length

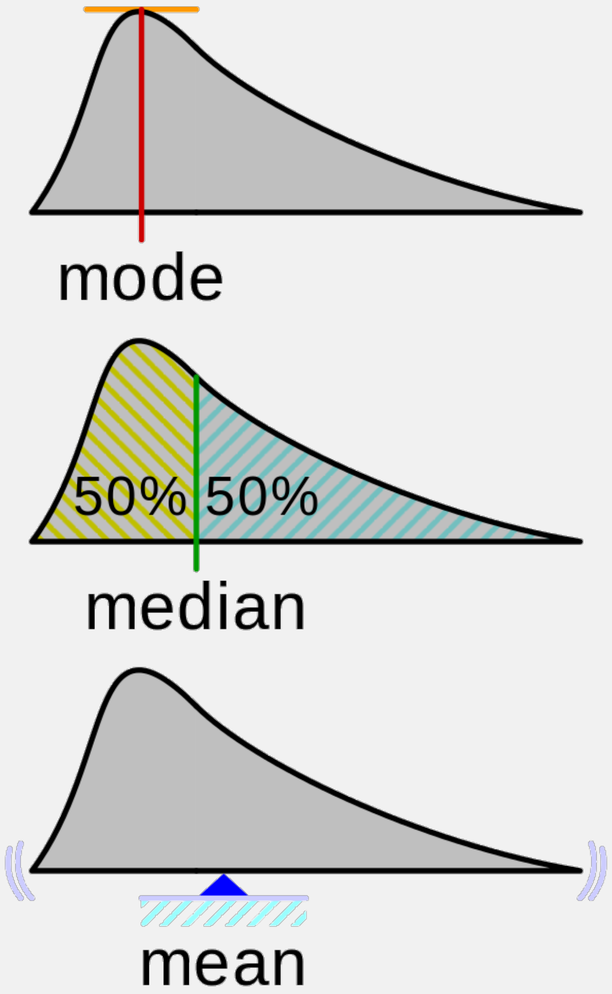
Picking monomers at random: average length of the filaments where they belong

CENTRAL TENDENCY—MODE AND MEDIAN

- **Median**
  At most half values are strictly less/greater than the median
- **Mode**
  Most frequent value

	Name	Age
	Johnny Doe	2
	Little Jane	5
	Jerry Small	3
	Billy Mouse	2
	Patrick Wise	93
Mean:	—	**27.2**
Median:	—	**3**
Mode:	—	**2**

CENTRAL TENDENCY—MODE AND MEDIAN



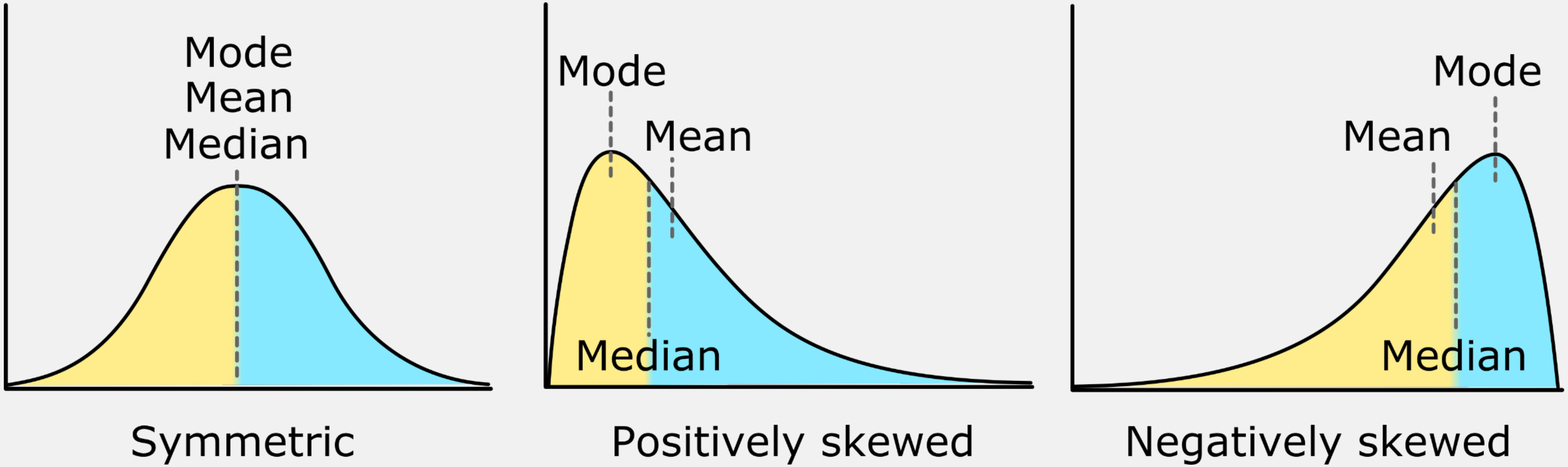
Wikipedia: median

CENTRAL TENDENCY—MIDRANGE

The midrange is the average of the lowest and highest values in the set

	Name	Age
	Johnny Doe	2
	Little Jane	5
	Jerry Small	3
	Billy Mouse	2
	Patrick Wise	93
Mean:	—	**27.2**
Midrange:	—	**47.5**

SYMMETRIC/ASYMMETRIC DATA



DATA DISPERSION

- **Range**

  Difference between largest and smallest value

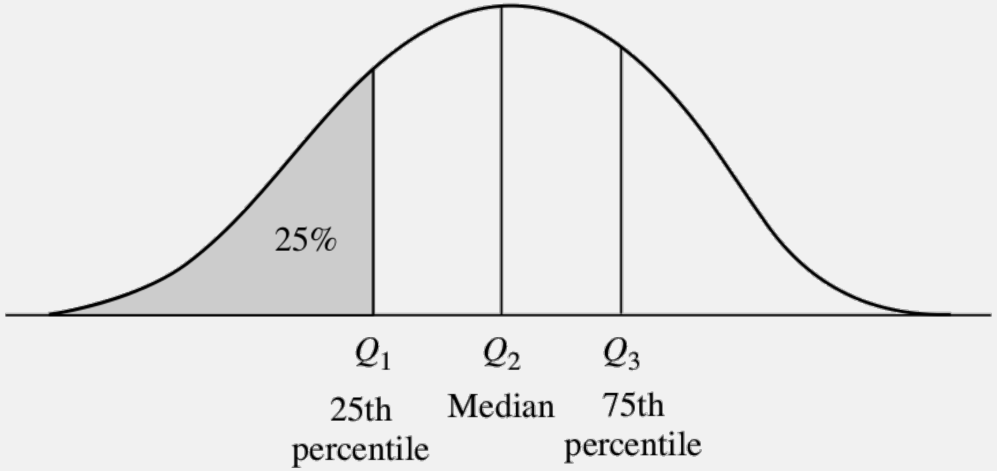
- **Quantiles**

  Divide sorted data into equal-sized sets

  - 4-Quantiles (quartiles)

    Interquartile range, IQR = $Q_3 - Q_1$

  - 100-Quantiles (percentiles)



VARIANCE/STANDARD DEVIATION

- Measurement of how close data values tend to be with respect to the mean
- **Low** standard deviation means values **close** to the mean

VARIANCE/STANDARD DEVIATION

The **variance** of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2, \tag{2.6}$$

where $\bar{x}$ is the mean value of the observations, as defined in Eq. (2.1). The **standard deviation**, σ, of the observations is the square root of the variance, σ^2.

DATA DISPERSION

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

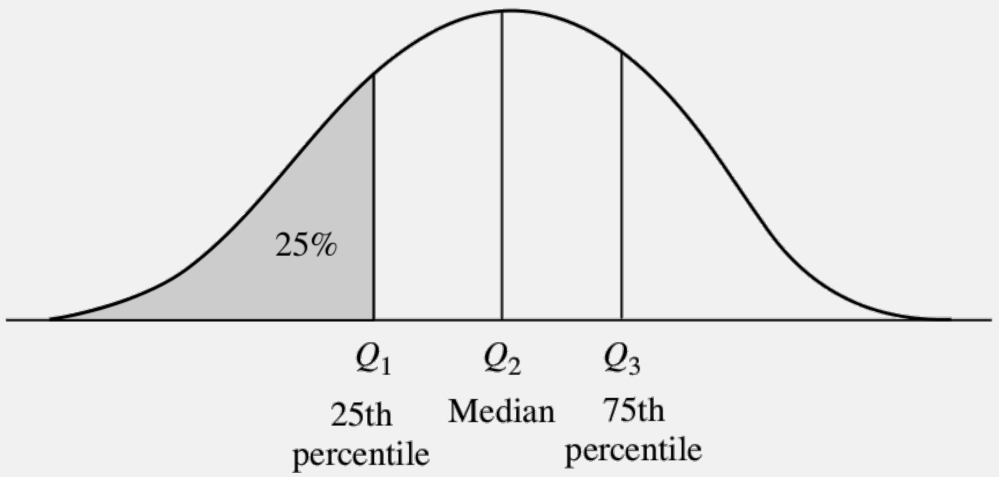
The standard deviation is a **biased** estimator!

Alternate formulas try to correct this

Normally not so important, but be consistent!

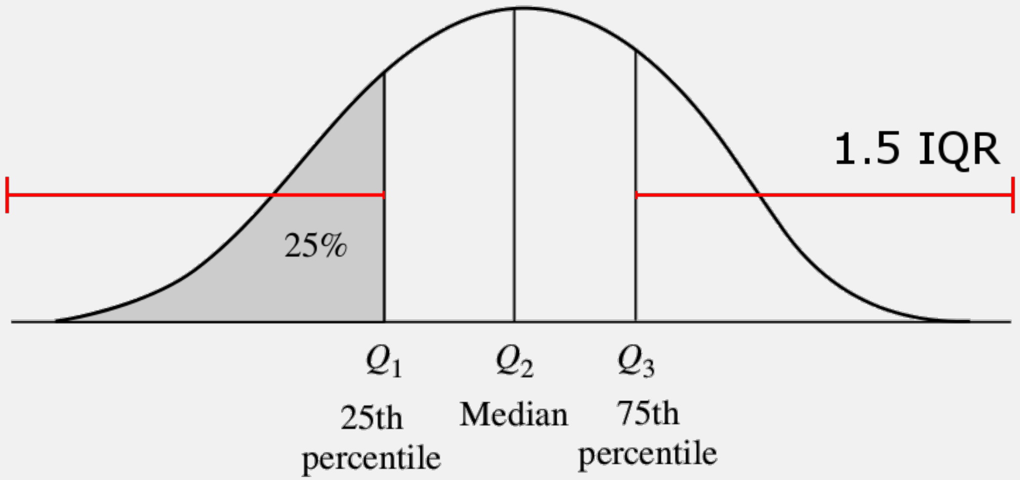
DATA DISPERSION—FIVE NUMBER SUMMARY

- No single measure is enough to describe skewed data
- Five-number summary:
  1. Minimum value
  2. Q_1
  3. Median (Q_2)
  4. Q_3
  5. Maximum value



DATA DISPERSION—FIVE NUMBER SUMMARY

- Its visualization is known as boxplot
- Outlier value:
  - Value that is “distant” from the rest
  - May be errors during data collection or odd behaviours
  - **Rule of thumb**: outliers are over 1.5 IQR below Q_1 or above Q_3

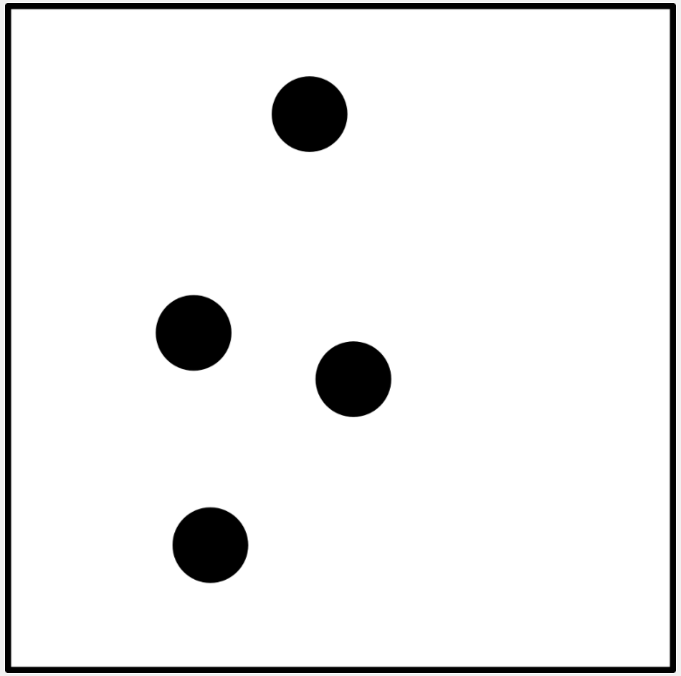


DATA SIMILARITY

- Measures "difference" between two data objects

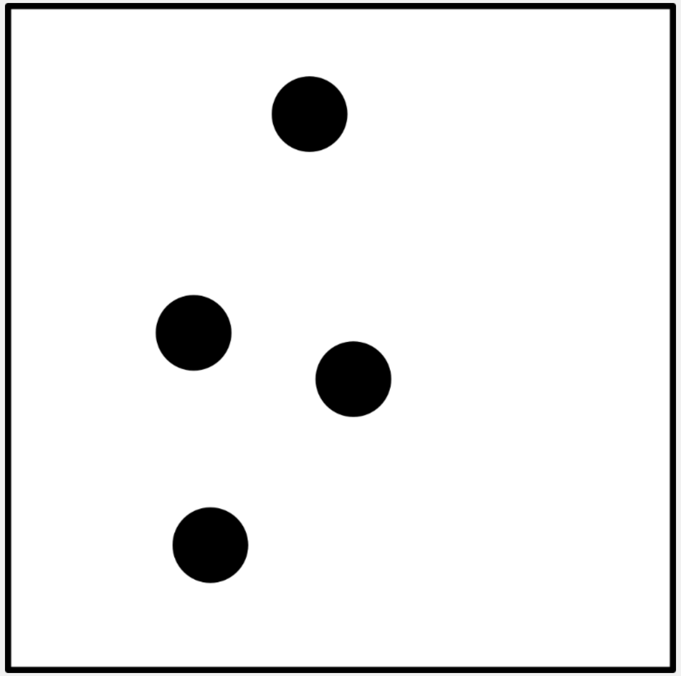
  Used in clustering, outlier analysis, nearest-neighbor classification, ...

- Typically returns **0** if two data objects are completely **unalike**, **1** if they are **the same**
- Dissimilarity is the opposite measure



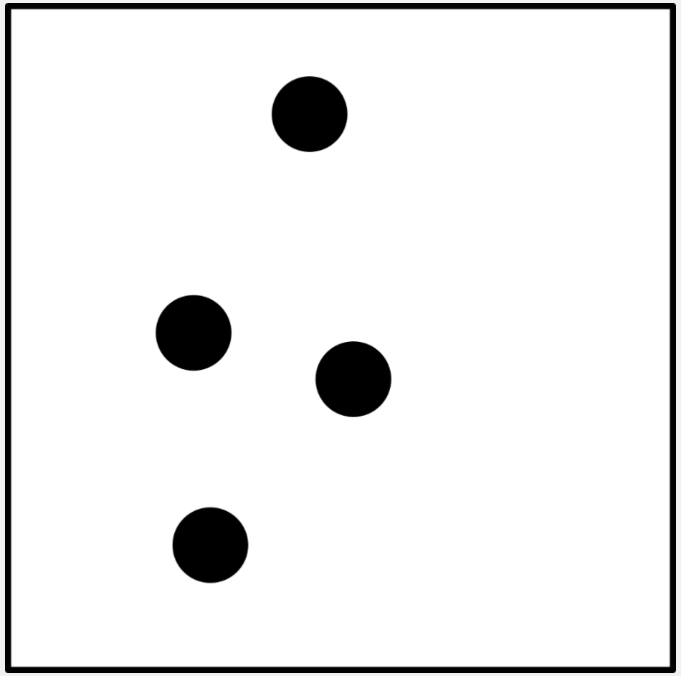
DATA SIMILARITY

- Different measures for each attribute type!
    - See sections 2.4.2–2.4.5 (!) in the book
- Used when the data object has only one kind of attribute
- What to do with mixed attribute types?



DATA SIMILARITY

- Measures “difference” between two data objects
  - Used in clustering, outlier analysis, nearest-neighbor classification, …
- Typically returns **0** if two data objects are completely **unalike**, **1** if they are **the same**
- Dissimilarity is the opposite measure



DATA SIMILARITY

Suppose the data contains *p* attributes:

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

$\delta_{ij}^{(f)}$ equals 0 if either x_i or x_j are absent for variable f *, otherwise it is 1.

*Or $x_i = x_j = 0$ and f is asymmetric binary

DATA SIMILARITY

- If f is interval-based: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$, where h runs over all nonmissing objects for variable f.
- If f is binary or categorical: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise $d_{ij}^{(f)} = 1$.
- If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, and treat z_{if} as interval-scaled.
- If f is ratio-scaled: either perform logarithmic transformation and treat the transformed data as interval-scaled; or treat f as continuous ordinal data, compute r_{if} and z_{if}, and then treat z_{if} as interval-scaled.

DATA SIMILARITY

- If f is interval-based: $d_{ij}^{(f)} = \frac{\boxed{|x_{if} - x_{jf}|}}{max_h x_{hf} - min_h x_{hf}}$, where h runs over all nonmissing objects for variable f.

There are many metrics to calculate distances!

Example: Minkowski distance:

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

DATA SIMILARITY

- If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if}-1}{M_f-1}$, and treat z_{if} as interval-scaled.

Size name	Small	Medium	Large
Rank	1	2	3

If, for object x_i, x_{if} = *medium*, then

$r_{if} = 2 - 1 = 1$

$M_{if} = 3$

$z_{if} = ½ = 0.5$

DATA SIMILARITY

- If f is $\boxed{\text{ratio-scaled:}}$ either perform logarithmic transformation and treat the transformed data as interval-scaled; or treat f as continuous ordinal data, compute r_{if} and z_{if}, and then treat z_{if} as interval-scaled.

Is the scale **linear**? Regardless of the strategy, make sure that your similarity is **normalized**.

DATA SIMILARITY

Any doubts?

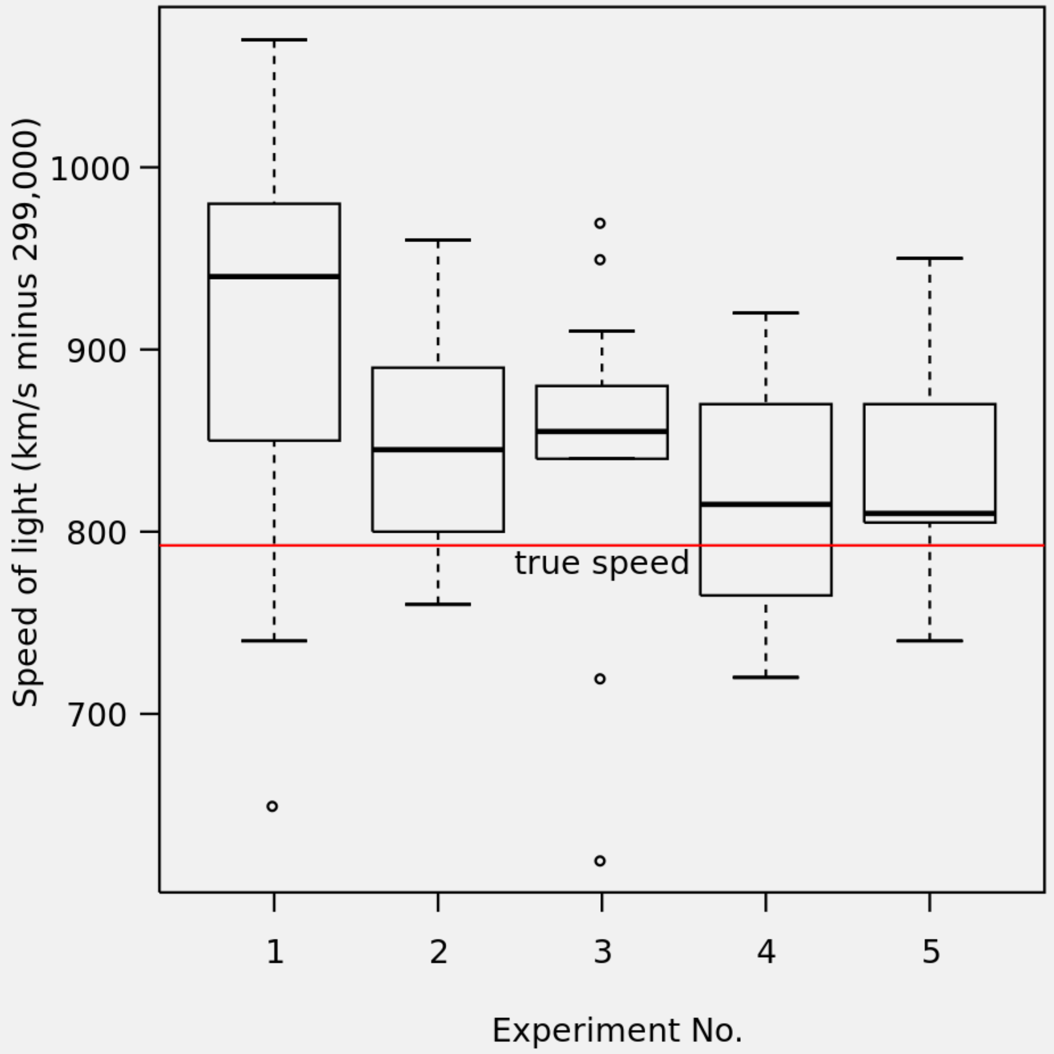
$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

VISUALIZING DATA

BOXPLOT

Visualization of five-number summary

- Ends of box: Q_1 and Q_3
- Median (Q_2) marked by line in box
- "Whiskers": last value within
  $Q_1 - 1.5 \cdot IQR$ and $Q_3 + 1.5 \cdot IQR$*
- Values without whiskers: outliers
- Variations for whiskers exist!



HISTOGRAMS

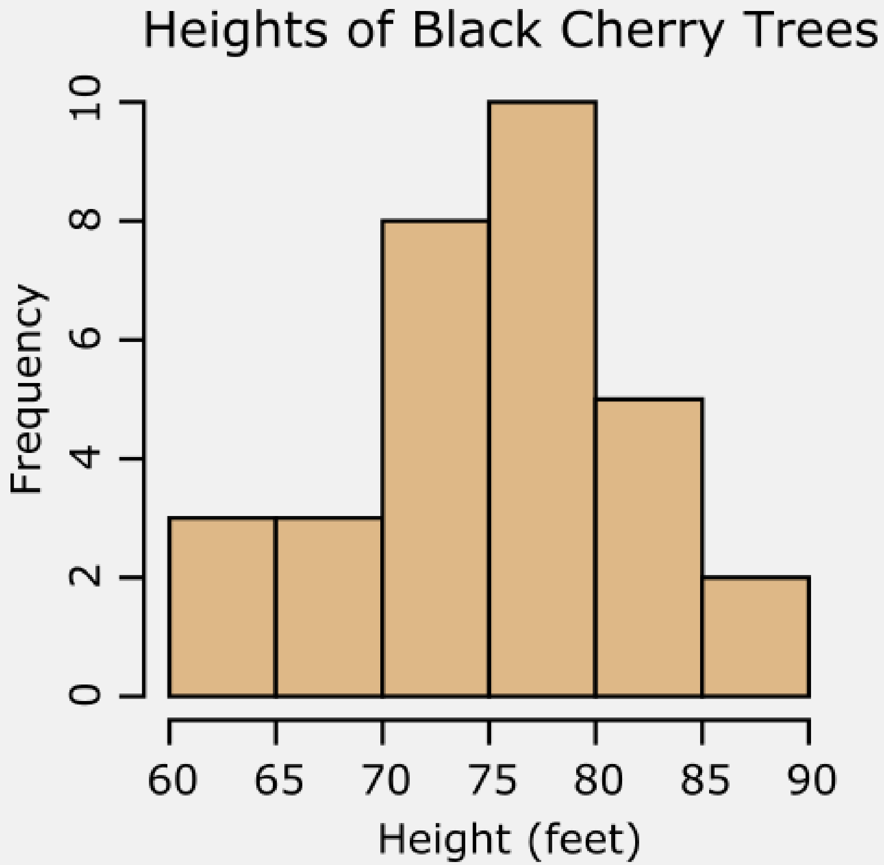
Distribution of attribute values

More informative than boxplots

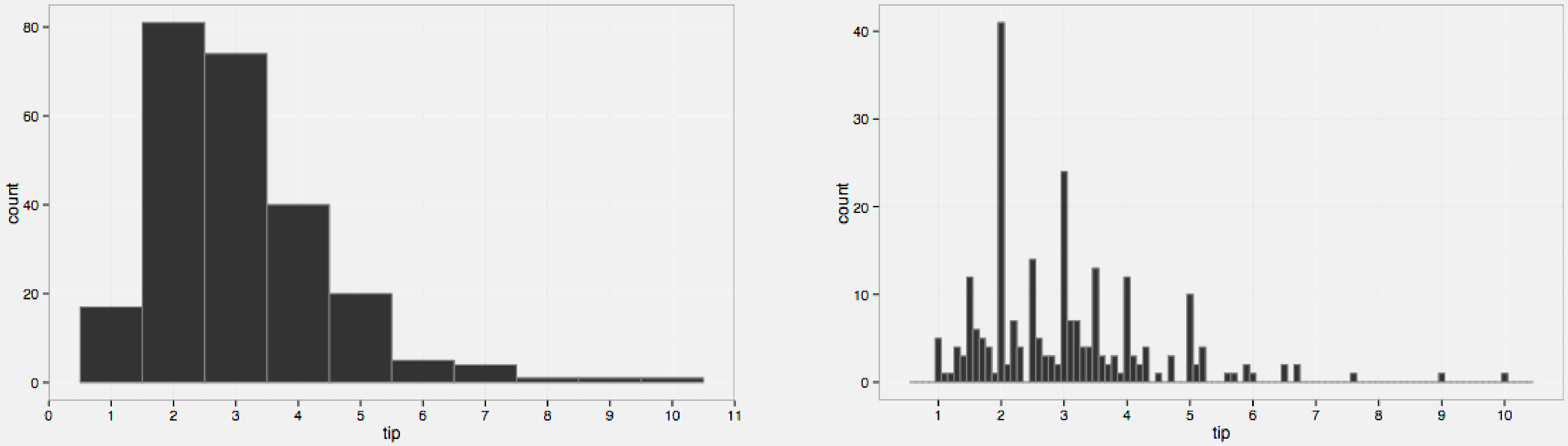
Values divided into buckets/bins

- Bucket range = width
- Typically constant width

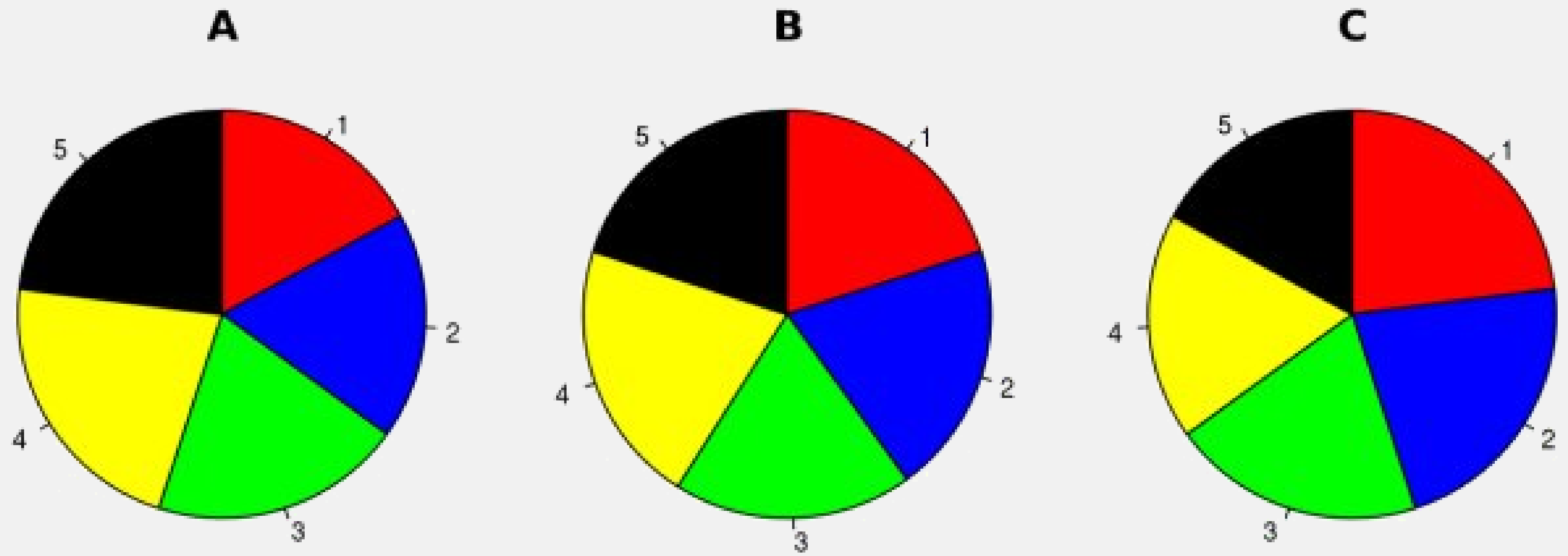
Used in data reduction



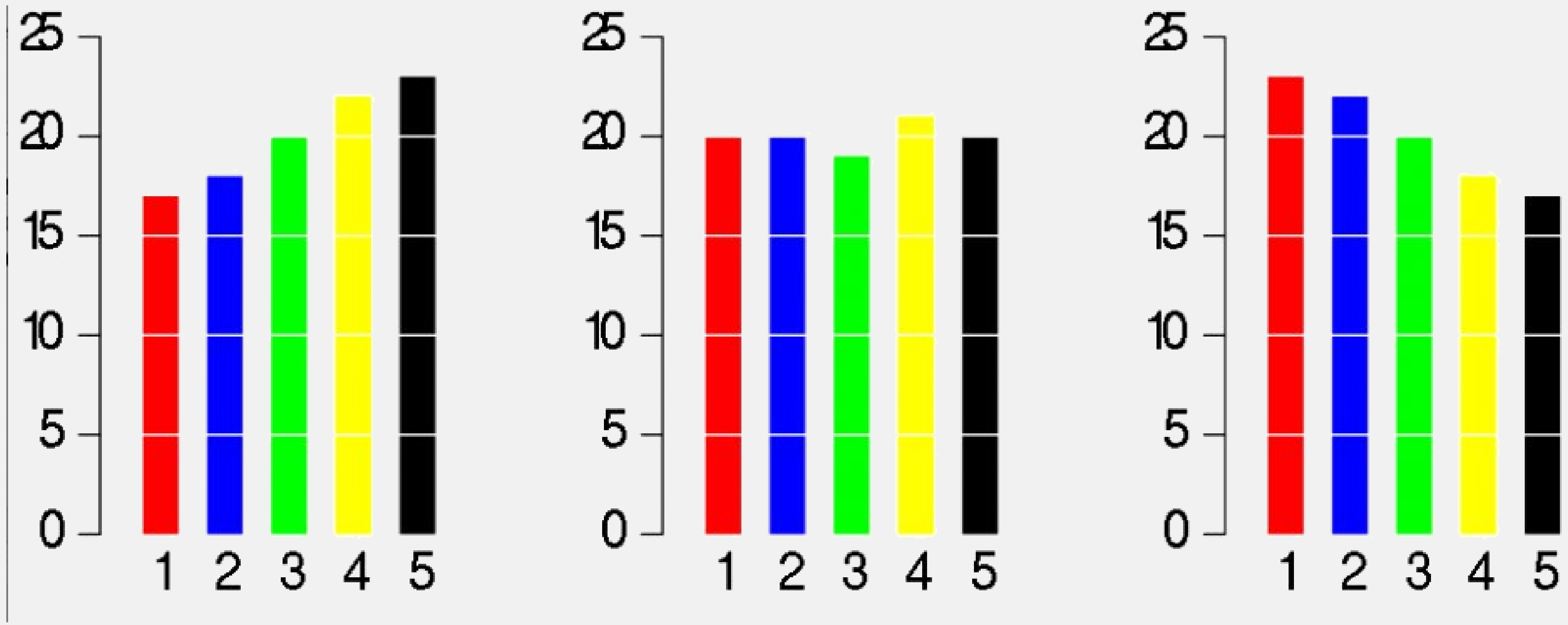
HISTOGRAMS—BIN SIZE



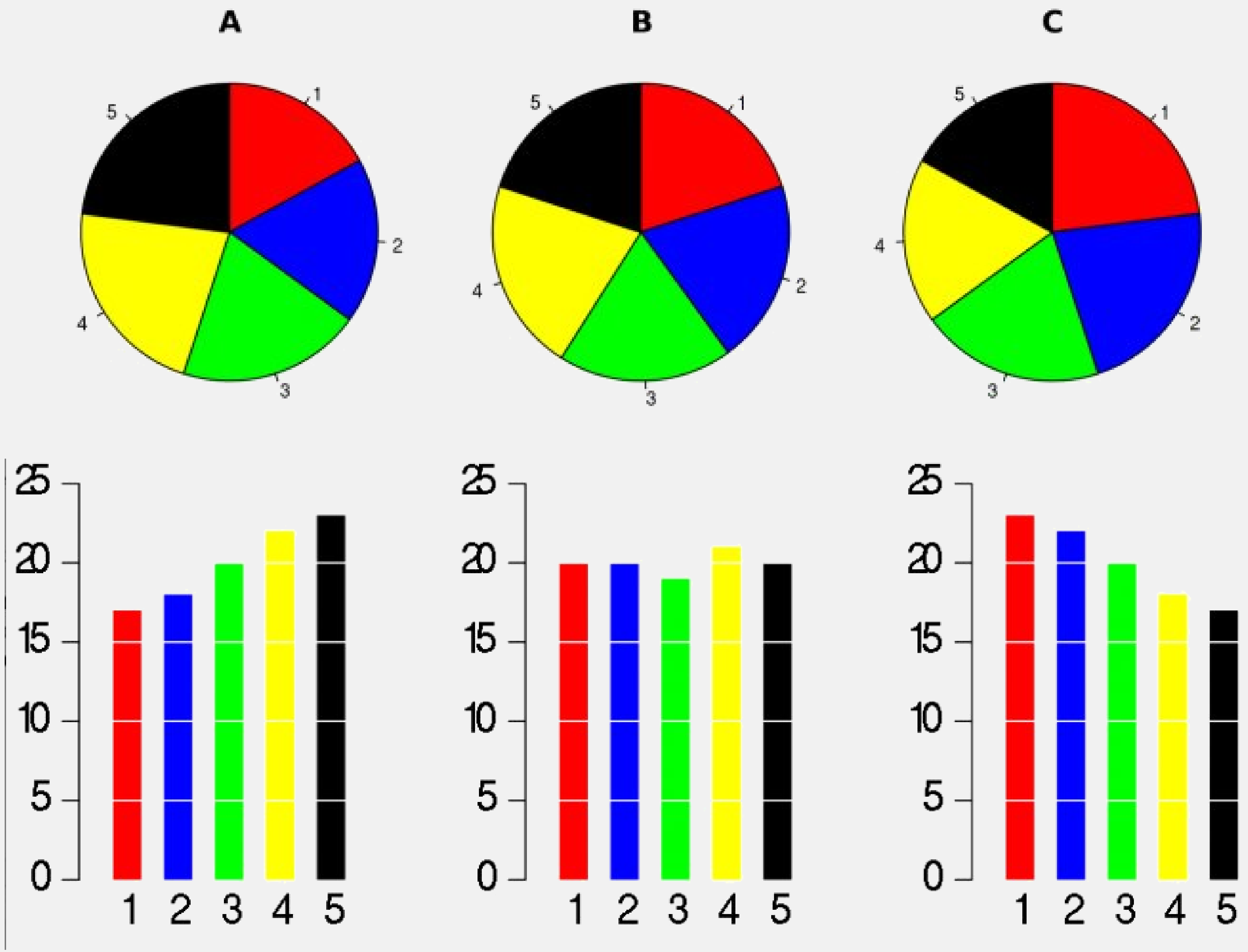
PIE CHARTS: DON'T!



PIE CHARTS: DON'T!

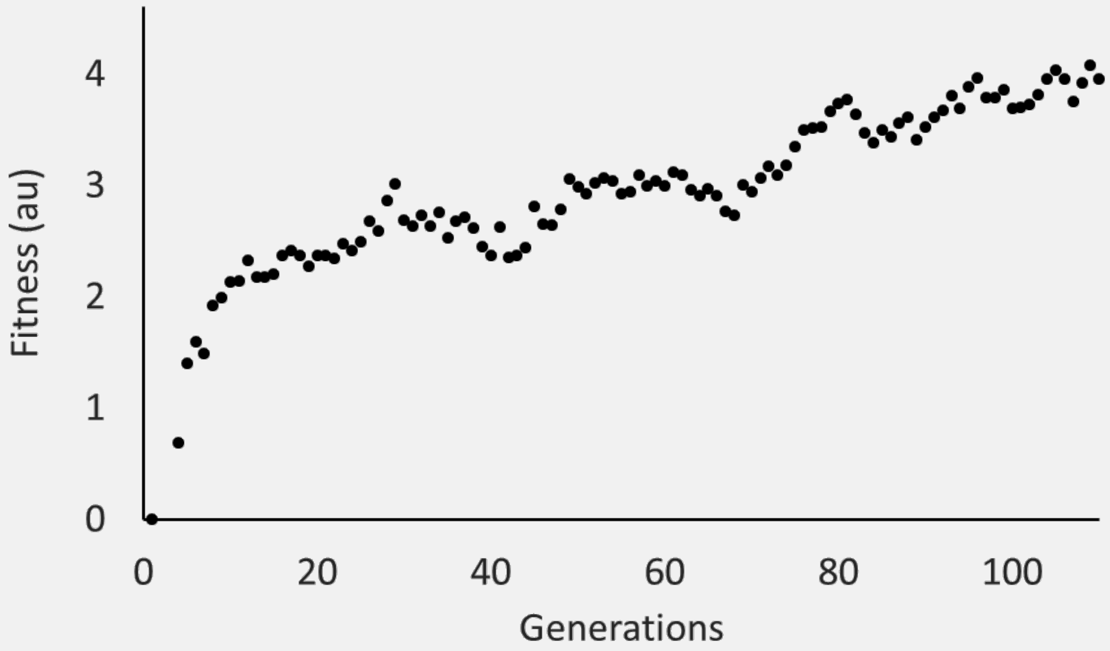


http://www.businessinsider.com/pie-charts-are-the-worst-2013-6?r=US&IR=T&IR=T


A
B
C
1
2
3
4
5
0
5
10
15
20
25

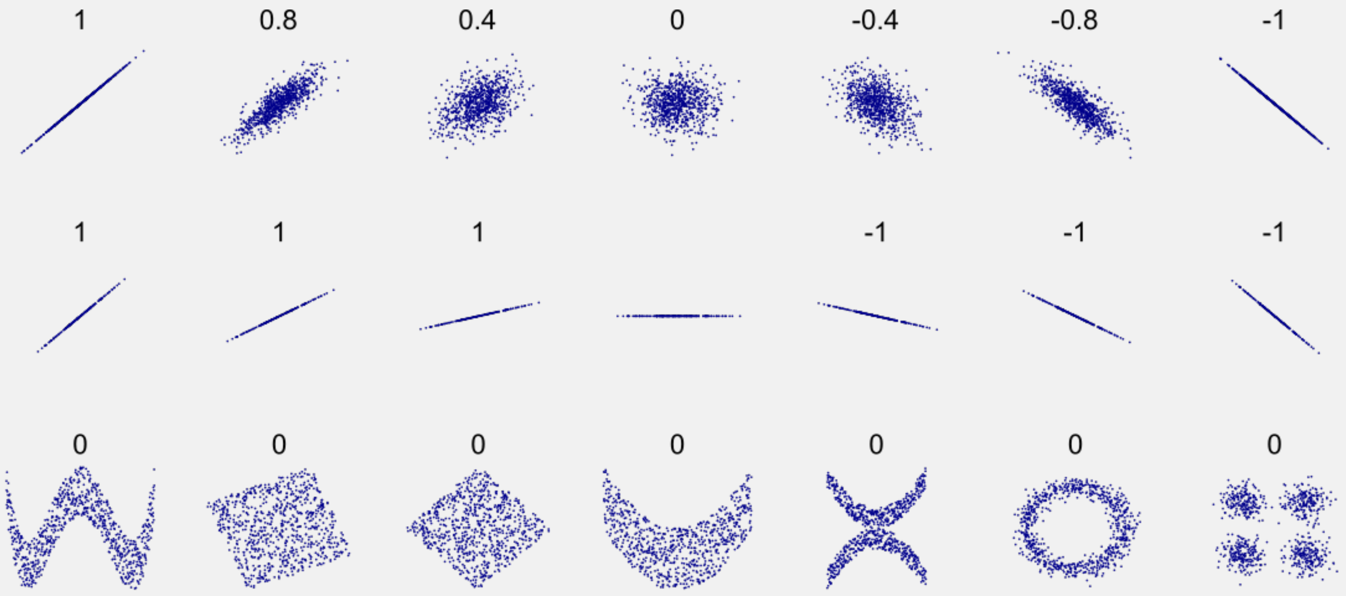
SCATTER PLOTS

- Shows patterns, trends and relationships between attributes
- Attribute values treated as coordinates
- What is correlation?

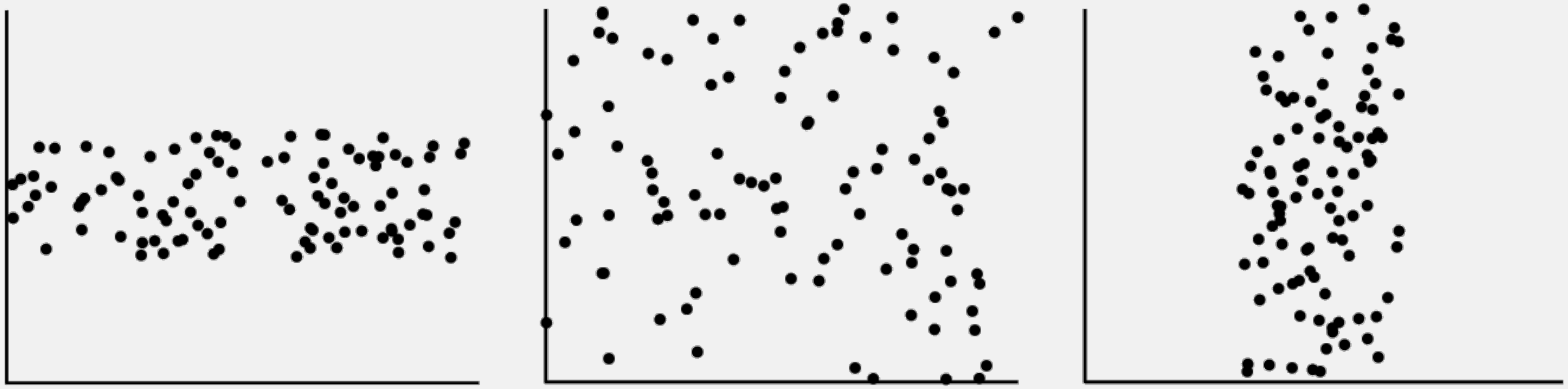


SCATTER PLOTS—CORRELATION

- Positive correlation:
  y increases as x increases
- Negative correlation
  y decreases as x increases
- Complex correlations possible!



SCATTER PLOTS—CORRELATION



Examples of data sets with no correlation between axes

HEAT MAPS

Attributes over a map

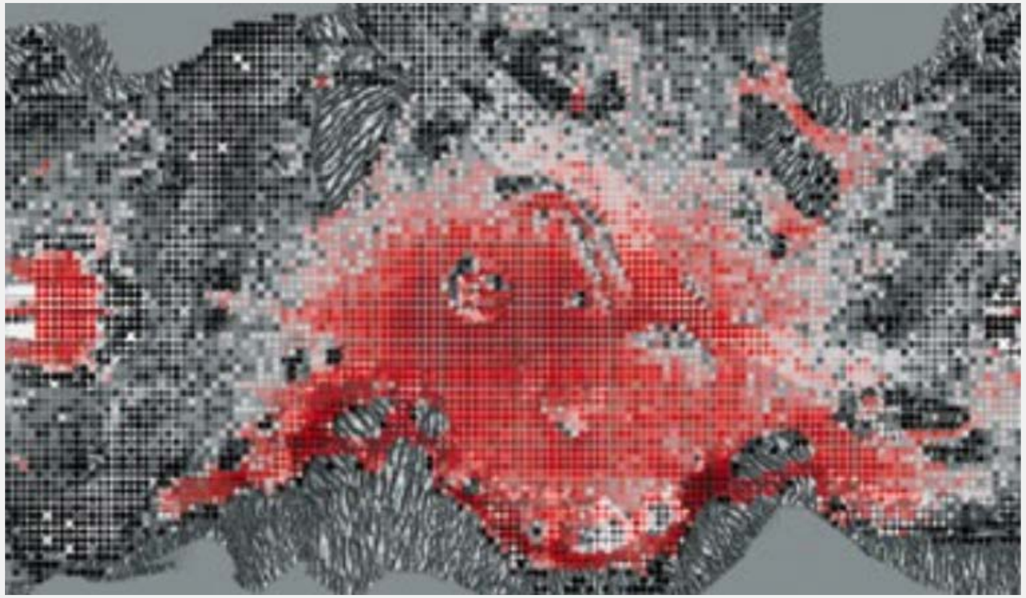
Higher values, higher “temperature”

Attributes are often counts

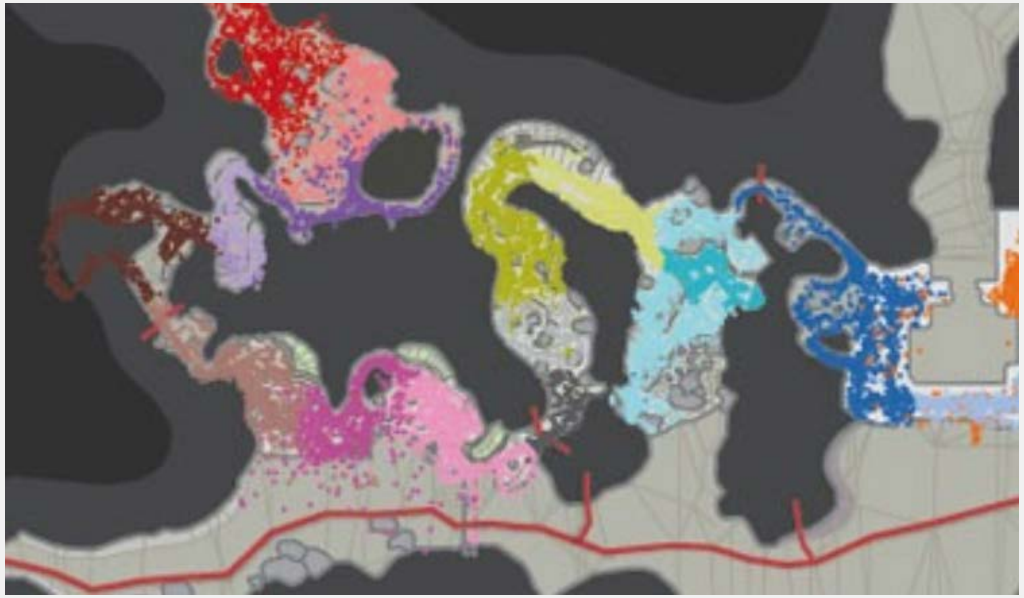
- E.g., number of deaths

HEAT MAPS—HALO 3

Number of deaths

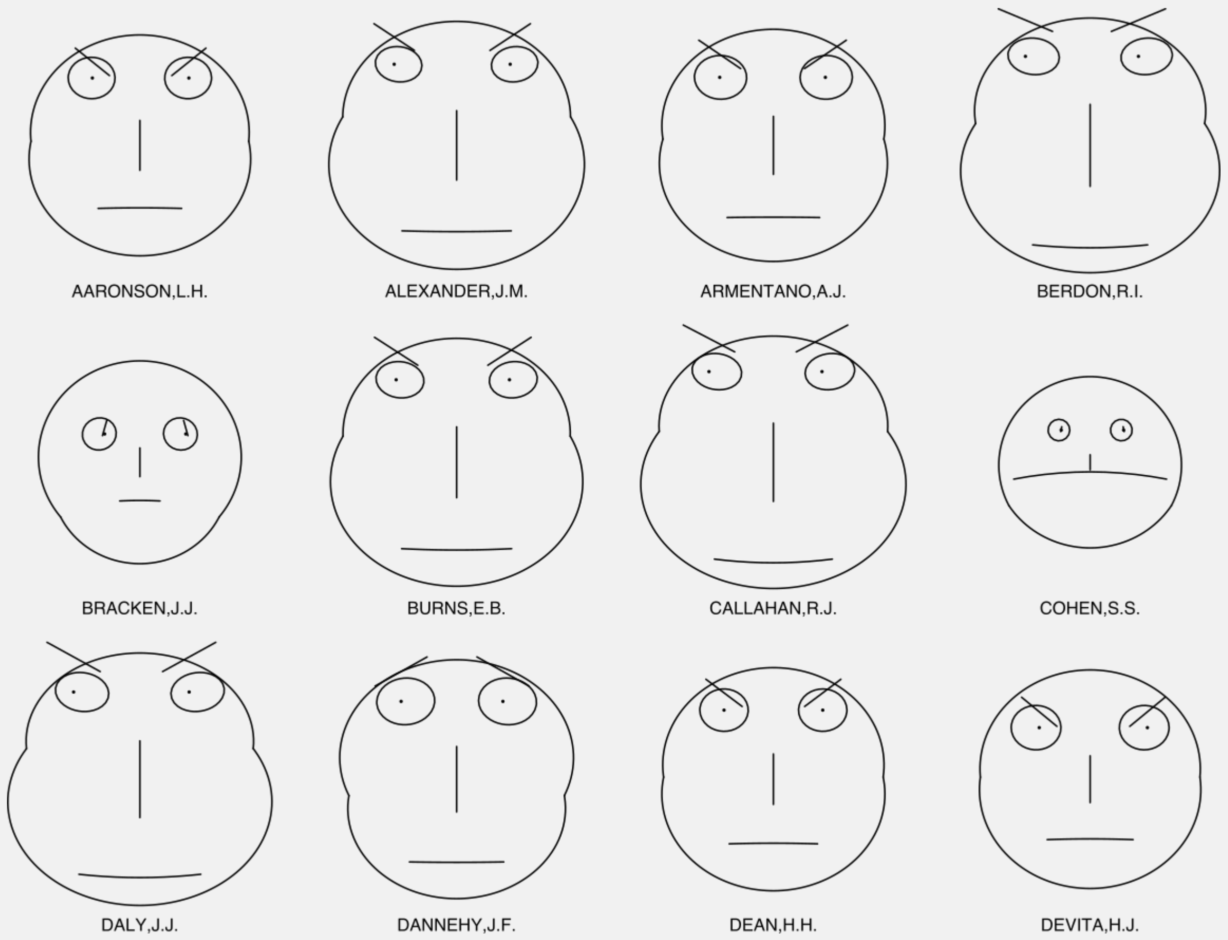


Player navigation



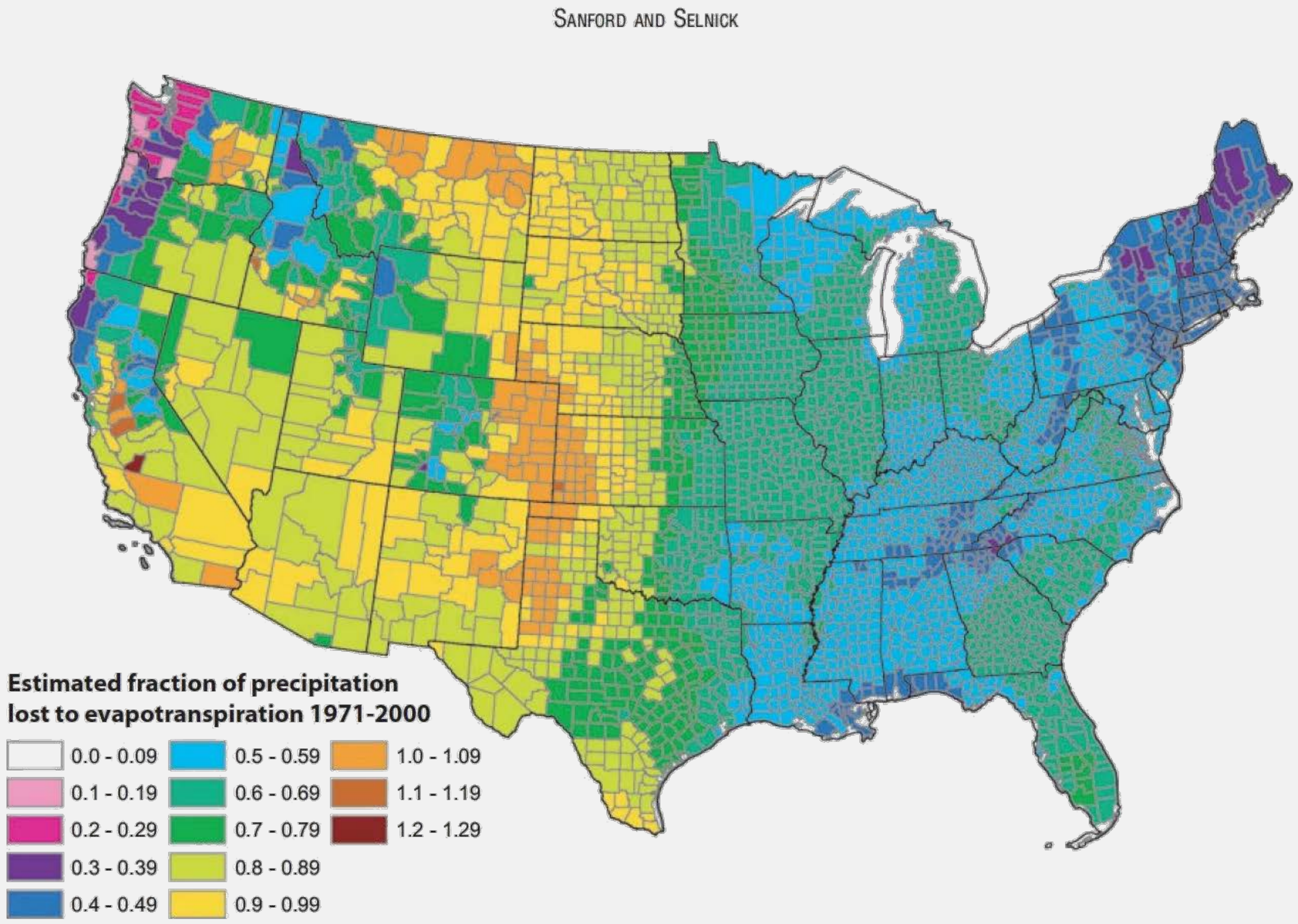
How Microsoft Labs Invented a New Science of Play. Thompson, Wired

OTHER METHODS



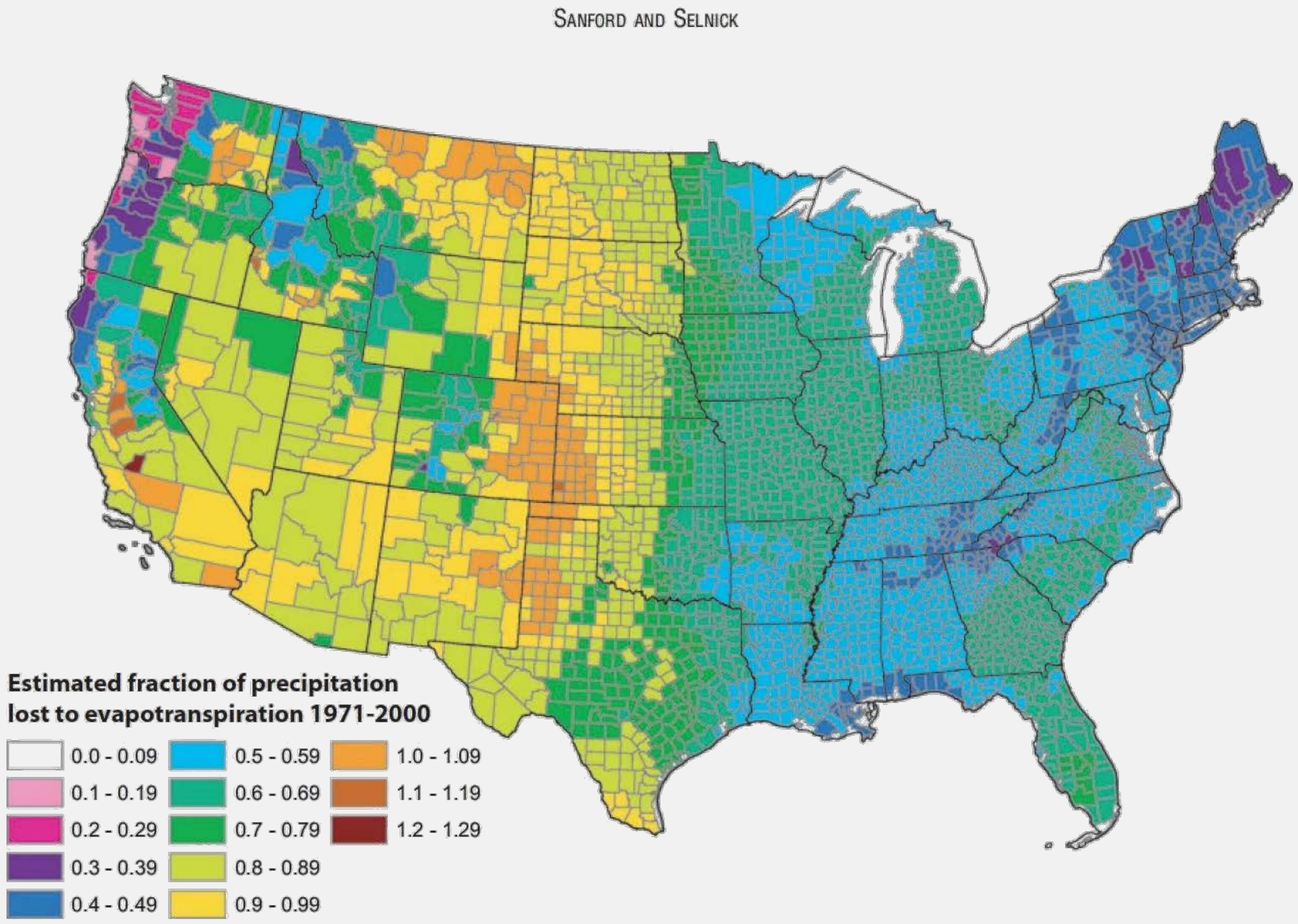
Chernoff faces

QUICK NOTE:



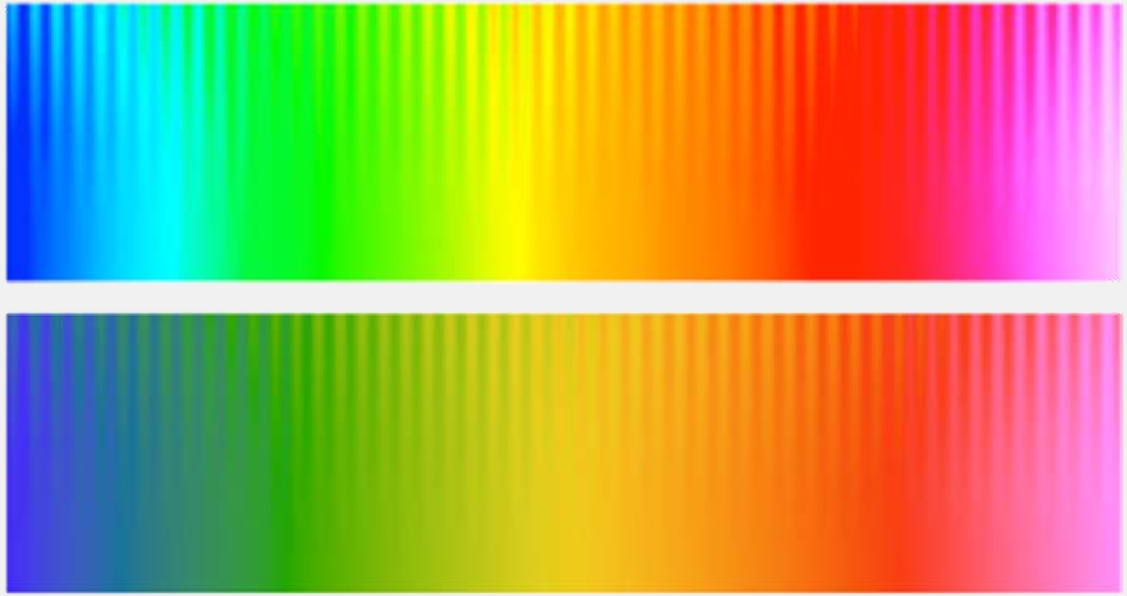
DOI: 10.1111/jawr.12010

QUICK NOTE: AVOID RAINBOW PALETE!



DOI: 10.1111/jawr.12010

QUICK NOTE: AVOID RAINBOW PALETE!



[Peter Kovesi]

[Ten simple rules for better figures]

CLEANING DATA

DATA CLEANING

Missing data

Smoothing

Removal of redundant and inconsistent attributes

MISSING DATA

Ignore object

May be problematic! Usually done when the class label is missing.

Fill in value

How?

MISSING DATA

Manually

Time consuming, often not feasible with big sets

Global constant ("unknown")

May confuse algorithms (why do these objects share the value "unknown"?)

MISSING DATA

Central tendency

Fill in with median (perhaps the mean)

Class tendency

If the object belongs to a known class,
we can use the median/mean for this class

Most probable value

Many inference techniques (regression,
Bayesian formalism, decision trees...)

MISSING DATA

All methods for **filling** in missing attributes may bias the data

NOISY DATA—SMOOTHING

Smoothing is used to **reduce noise** in data

Noise is a random error or variance in a measured variable

SMOOTHING

Binning: smoothing by looking at neighbours

- Sort values and distribute them into equal-sized bins
- Smoothing by means

    Replace values with bin mean

- Smoothing by medians

    Replace values with bin median

- Smoothing by boundaries

    Replace values with closest boundary value in the bin

SMOOTHING

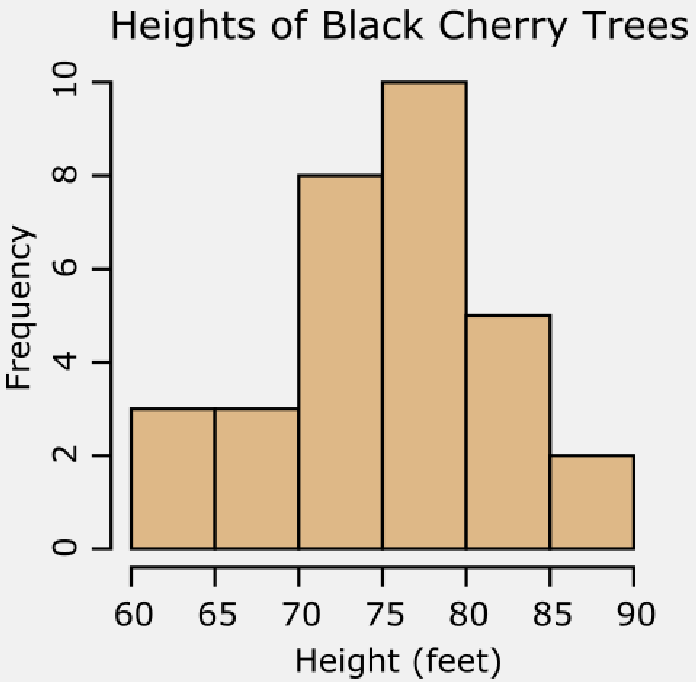
Binning: smoothing by looking at neighbours

- Sort values and distribute them into equal-sized bins
- Smoothing by means

  Replace values with bin mean
- Smoothing by medians

  Replace values with bin median
- Smoothing by boundaries

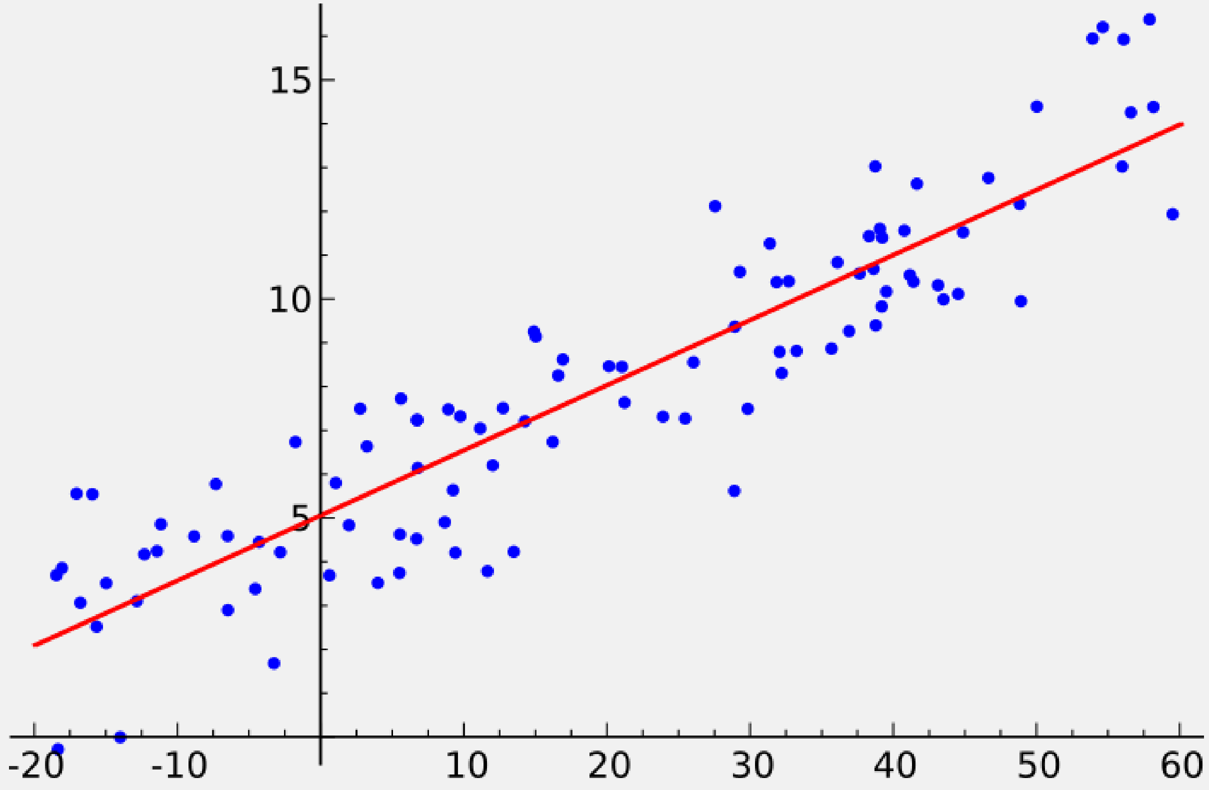
  Replace values with closest boundary value in the bin



SMOOTHING

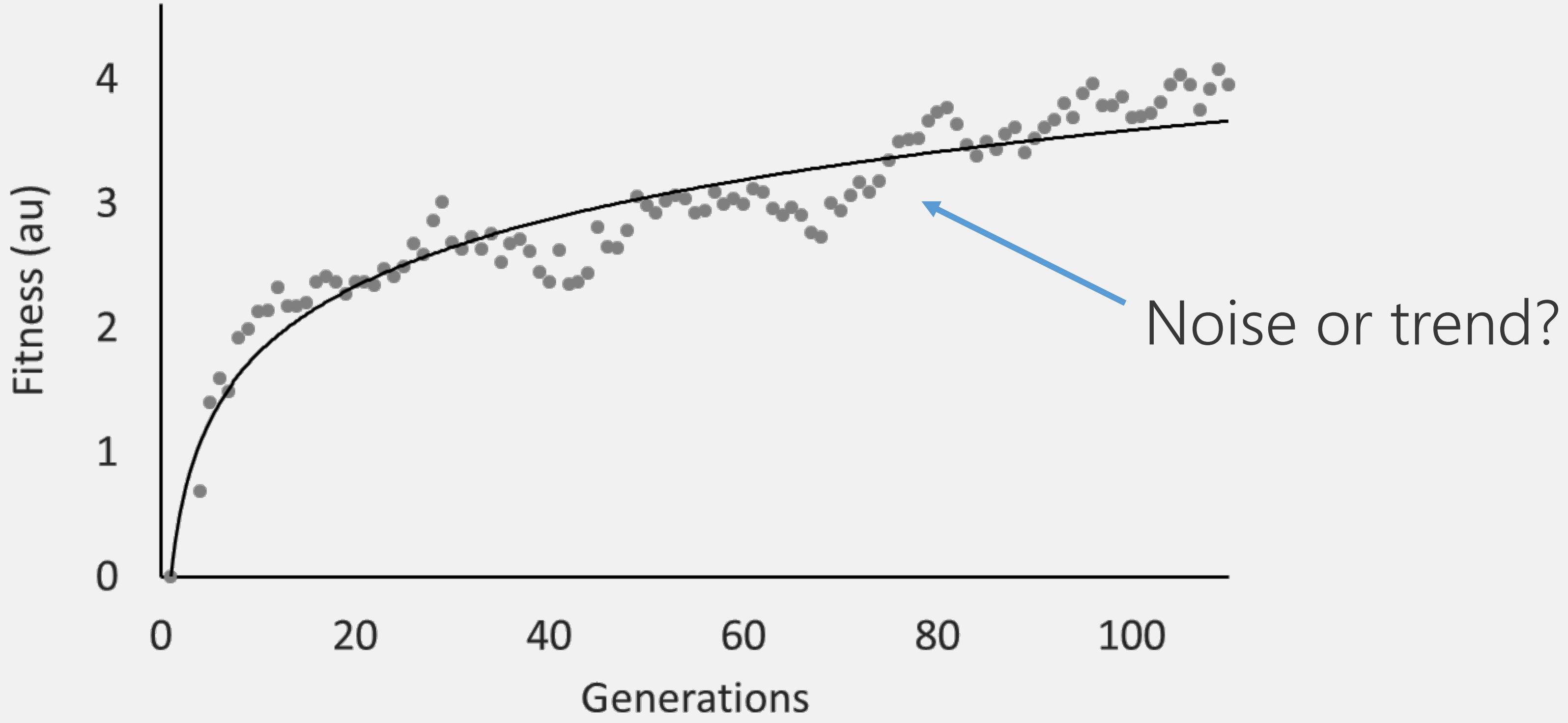
Data	8	9	28	15	21	34	4	21	26	29	25	24
Sorted	4	8	9	15	21	21	24	25	26	28	29	34
By means	9	9	9	9	22.8	22.8	22.8	22.8	29.3	29.3	29.3	29.3
By medians	8.5	8.5	8.5	8.5	22.5	22.5	22.5	22.5	28.5	28.5	28.5	28.5
By boundaries	4	4	4	15	21	21	25	25	26	26	26	34

SMOOTHING



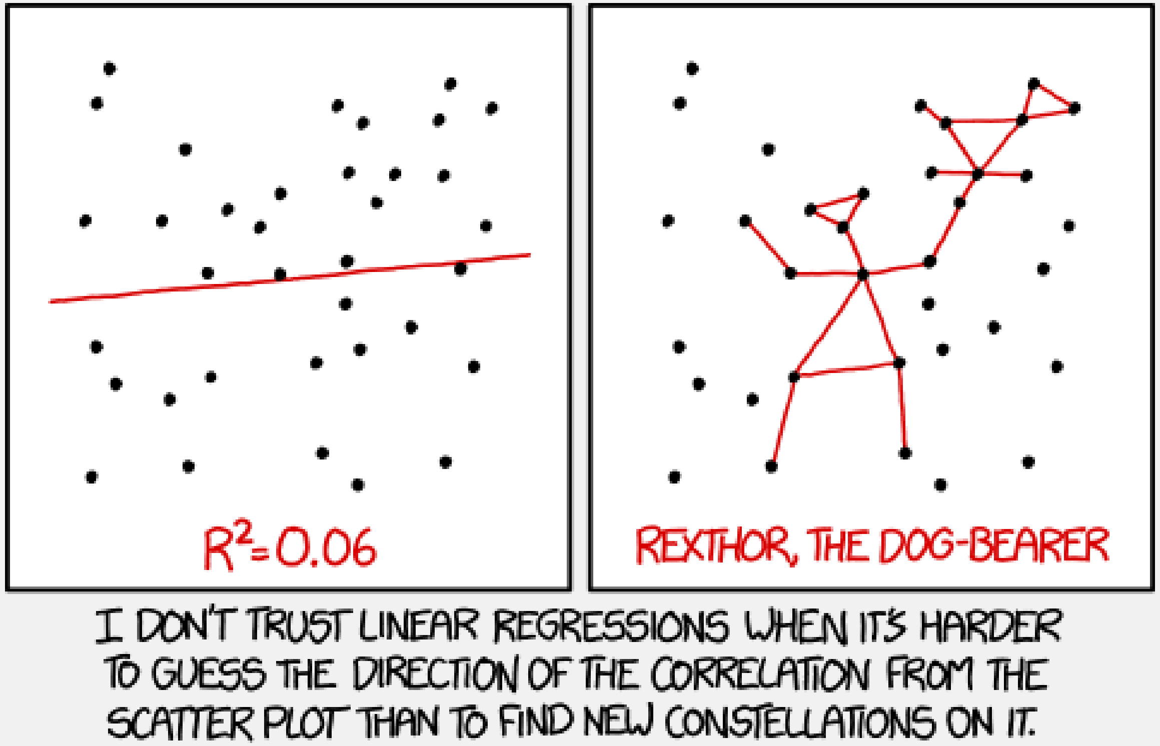
Regression: fit data into a regression function

SMOOTHING



Danger 1: oversimplify underlying phenomena

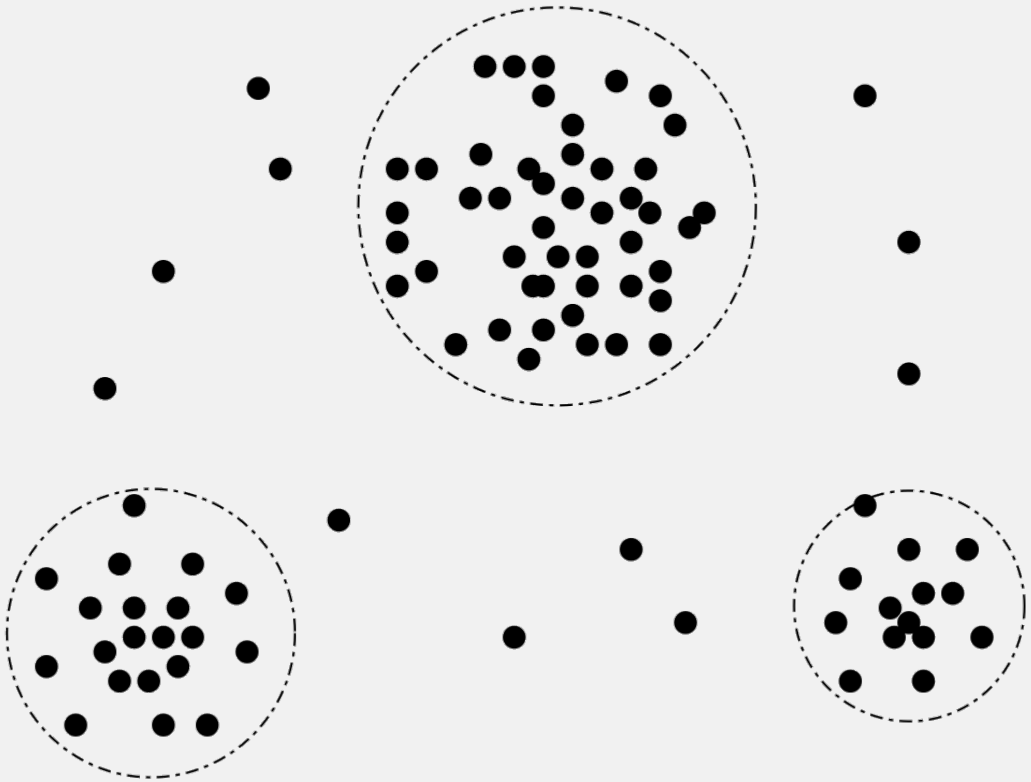
SMOOTHING



Danger 2: wishful fitting (xkcd 1725)

NOISY DATA—CLUSTERING

- Use clustering to **remove outliers**
- Divide data into clusters (for example, k-means)
- Data outside a range considered outliers
- More on clusters ahead on the course!



DATA REDUNDANCY

- An attribute is redundant if it can be derived from other attributes

  Example: area, width, height

- Visual detection (using scatter plots, etc.)
- Correlation analysis (chapter 3.3.2 (!))

  (Chi-square test for nominal data,
  Pearson's correlation coefficient, etc.)

DATA REDUNDANCY

x1	x2	x3
1	2	2.23
2	4	7.82
3	6	11

Both x2 and x3 have positive correlation with x1, but only x2 is redundant!

Correlation does **not** mean **redundancy**!

REDUCING DATA

DATA REDUCTION

- Data analysis using huge data sets can take a long time
- Is it possible to reduce the size while retaining the relevant characteristics of the original set?

DATA REDUCTION

Dimensionality reduction: reduce number of attributes

- Wavelet transform (3.4.2), principal components (3.4.3), attribute subset selection (3.4.4).

DATA REDUCTION

Numerosity reduction: replace data with a smaller-size representation

- Parametric methods create models. Model parameters are stored instead of data. Example: regression.
- Non-parametric methods store a reduced representation of the data. Examples: histograms, clustering, sampling.

DATA REDUCTION

Data **compression**: data is transformed into reduced representation. (Think of mp3.) Lossless (original data can be recreated) or lossy (only an approximation can be recovered).

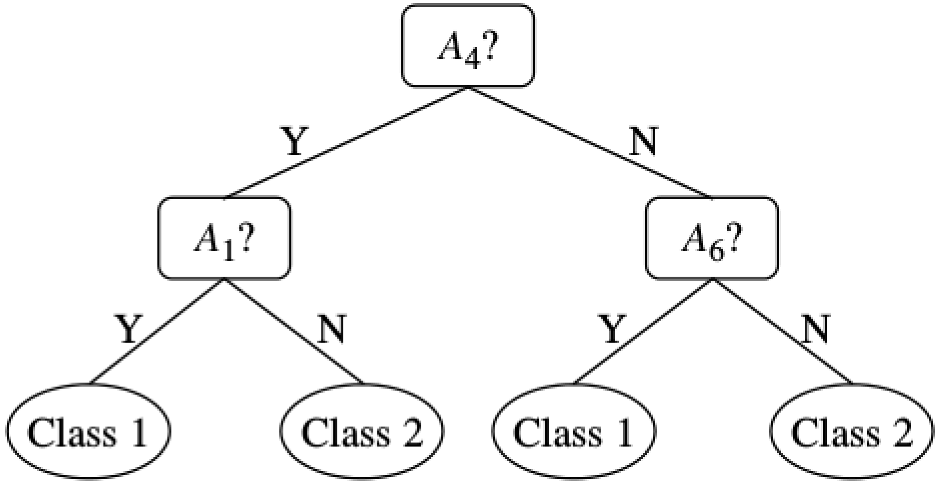
ATTRIBUTE SUBSET SELECTION

- Based on the task at hand we may be able to identify irrelevant attributes
  - Often difficult and time-consuming
  - Danger: accidental removal of relevant attributes
  - Example: student ID for academic results prediction
- Attribute subset selection algorithms

ATTRIBUTE SUBSET SELECTION

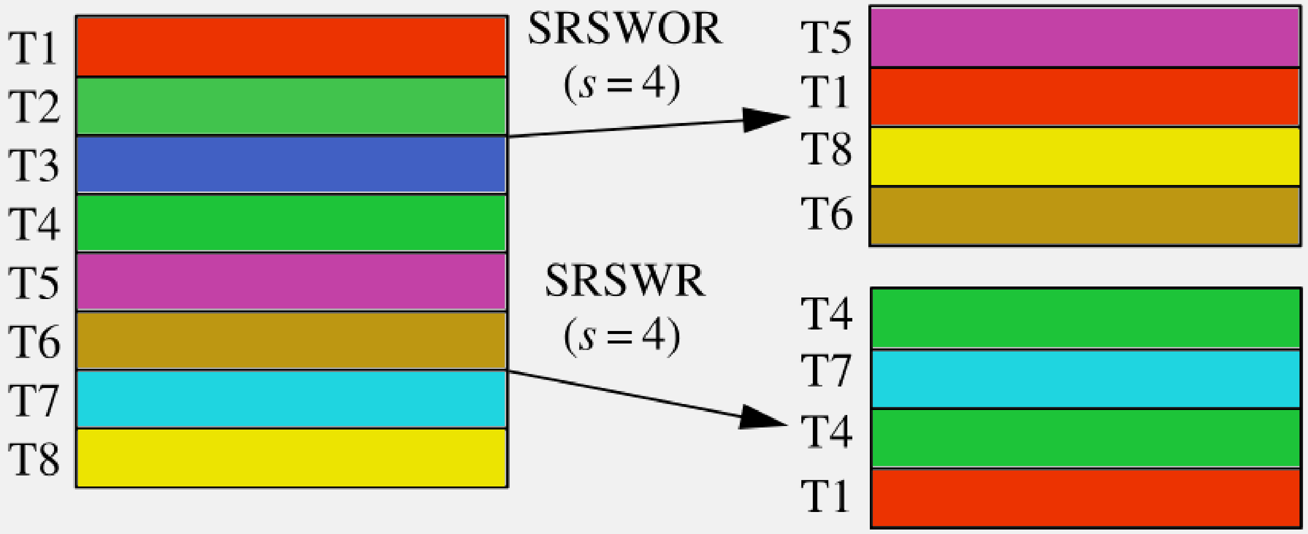
- Algorithms require definition of “good” attribute
- Usually statistical significance or another measure like *information gain* (more on this later!)

ATTRIBUTE SUBSET SELECTION

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

SAMPLING

- Smaller data set by randomly selecting objects in the set
- Different strategies (3.4.8), examples:
  - SRSWOR: simple random sample without replacement
  - SRSWR: simple random sample with replacement



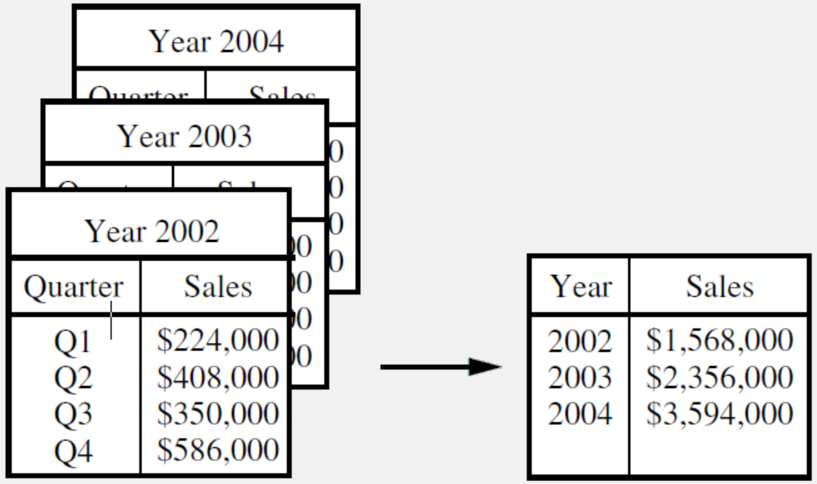
TRANSFORMING DATA

DATA TRANSFORMATION OVERVIEW

- Smoothing
- Attribute construction

    Example: (width, height) → area

- Aggregation
    - Example: daily sales → monthly sales



DATA TRANSFORMATION OVERVIEW

- Normalization

  Data range reduction (typically [0, 1])

- Discretization

  Continuous attributes to discretized or nominal attributes.

  Example: age to "young, old", or age groups: 0–10, 10–20, etc.

DATA TRANSFORMATION OVERVIEW

Concept hierarchy generation

- (street < city < state < country)
- Allow data exploration in different scales
- Different techniques!

NORMALIZATION

- The relative values of numerical attributes may affect results!

  Attribute in centimeters vs meters

- Can make attributes take more weight in results
- Normalization standardizes the values range
- Different techniques
  - Min-max: values based on minimum and maximum values
  - Z-score: using mean and standard deviation of the attribute
  - Decimal scaling: multiplication by a power of 10

NORMALIZATION

Min-max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v, of A to v' in the range $[new\_min_A, new\_max_A]$ by computing

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A. \qquad (2.11)$$

Min-max normalization preserves the relationships among the original data values. It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for A.

CONCLUSION

VISUALIZATION AND DESCRIPTIVE STATISTICS

Data my be too complex to evaluate by looking at it!

Visualization helps us to understand the data

It also helps to identify problems!

PREPROCESSING

Real data is not perfect, we need cleaning and preprocessing!

Good data-collection design avoids many problems

GETTING WHAT YOU ASK FOR

Poor questionnaires yield poor data

Worse: tailoring questions to lead answers

Example: [Yes, Minister](#) (BBC comedy series)

Were any questions in last week's questionnaire framed?

THANKS FOR LISTENING!