CSI4142 – Data science

Topic 1

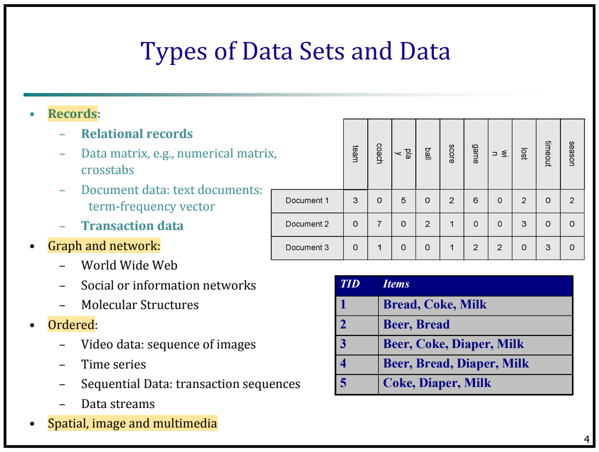
Getting to know your data

Overview of topic

- User Expectations versus Data Reality
- A word about data sets and data types
- Getting to know your data
  - Basic Statistical Descriptions of Data
  - Visualization

User Expectations versus Data Reality

- Decisions
  - Do we have enough data?
  - Do we have enough high quality data?
  - Do we have the ability to get enough high quality data soon?
  - Biggest risk → underestimating the difficulty to source your data
  - List success criteria: specific, measurable

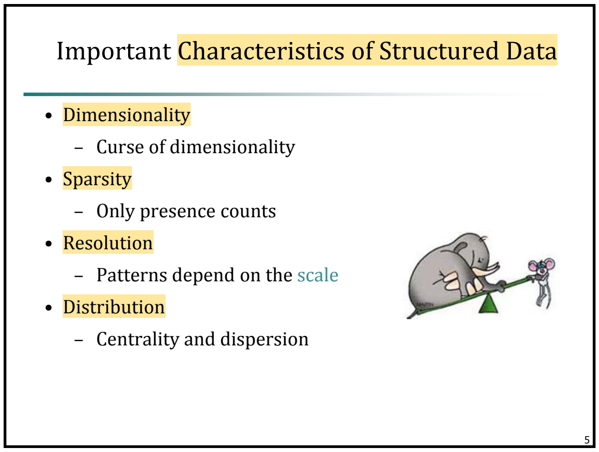


Types of Data Sets and Data

- Records:
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network:
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered:
  - Video data: sequence of images
  - Time series
  - Sequential Data: transaction sequences
  - Data streams
- Spatial, image and multimedia

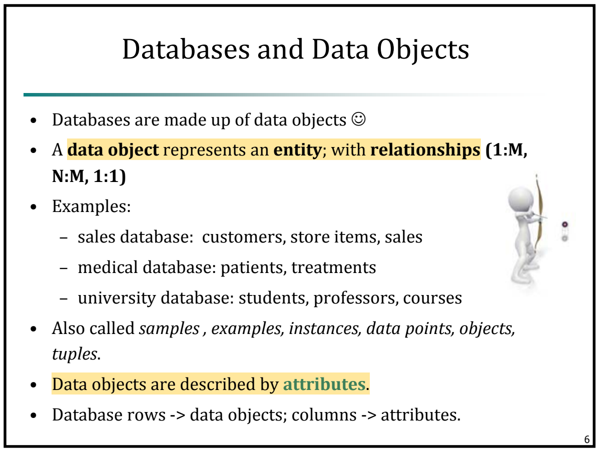
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

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



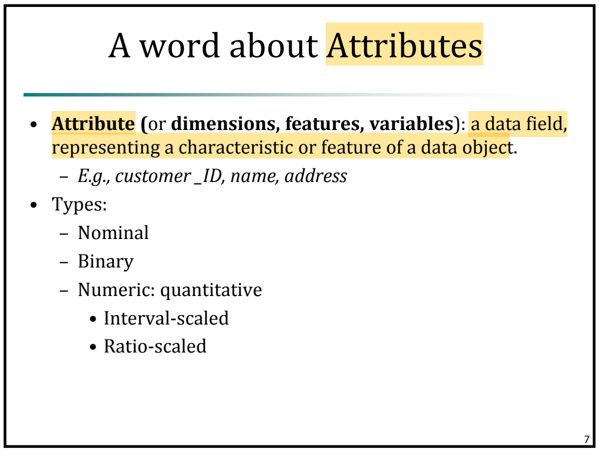
Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion



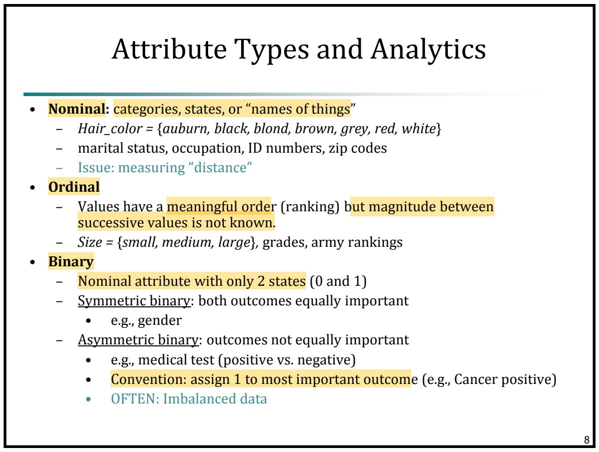
Databases and Data Objects

- Databases are made up of data objects ☺
- A **data object** represents an **entity**; with **relationships** **(1:M, N:M, 1:1)**
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples.*
- Data objects are described by **attributes**.
- Database rows -> data objects; columns -> attributes.



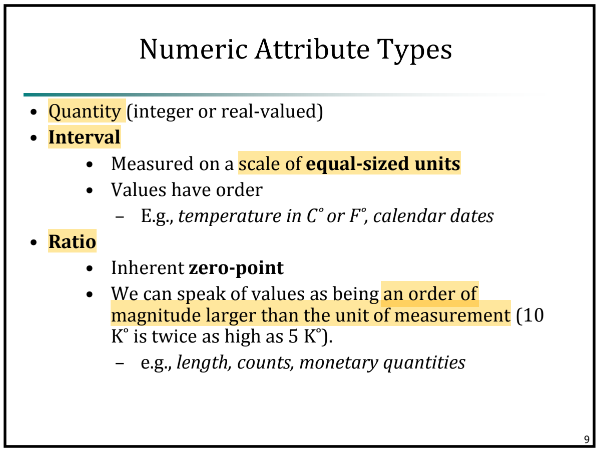
A word about Attributes

- **Attribute** (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled



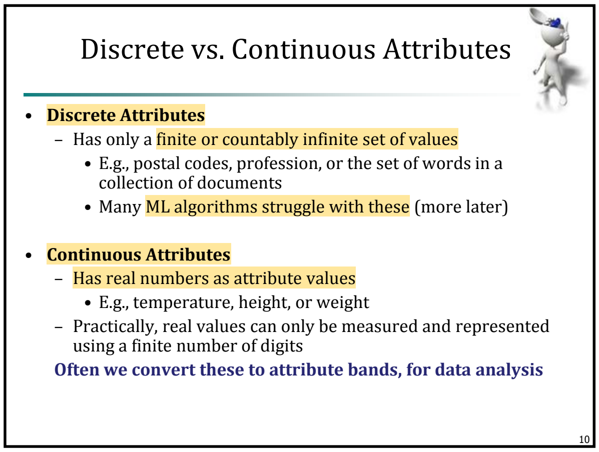
Attribute Types and Analytics

- **Nominal**: categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
  - Issue: measuring "distance"
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., Cancer positive)
    - OFTEN: Imbalanced data



Numeric Attribute Types

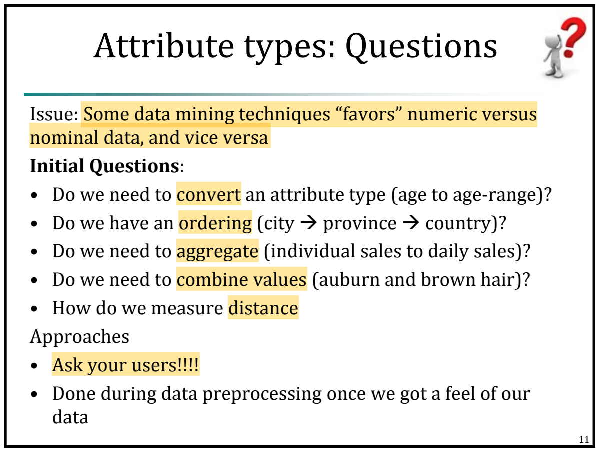
- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚ or F˚, calendar dates*
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *length, counts, monetary quantities*



Discrete vs. Continuous Attributes

- **Discrete Attributes**
  - Has only a finite or countably infinite set of values
    - E.g., postal codes, profession, or the set of words in a collection of documents
    - Many ML algorithms struggle with these (more later)
- **Continuous Attributes**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits

  Often we convert these to attribute bands, for data analysis



Attribute types: Questions

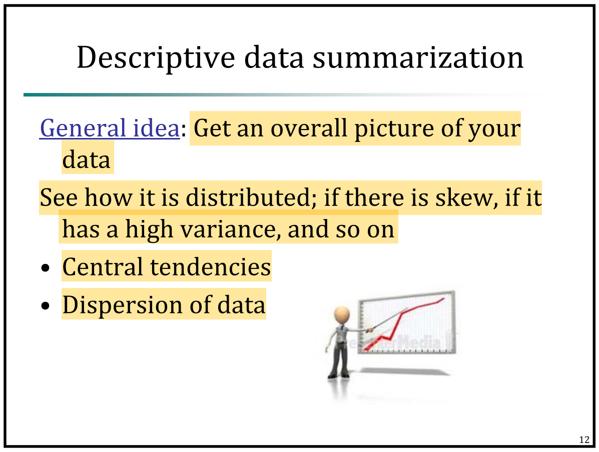
Issue: Some data mining techniques “favors” numeric versus nominal data, and vice versa

Initial Questions:

- Do we need to convert an attribute type (age to age-range)?
- Do we have an ordering (city → province → country)?
- Do we need to aggregate (individual sales to daily sales)?
- Do we need to combine values (auburn and brown hair)?
- How do we measure distance

Approaches

- Ask your users!!!!
- Done during data preprocessing once we got a feel of our data

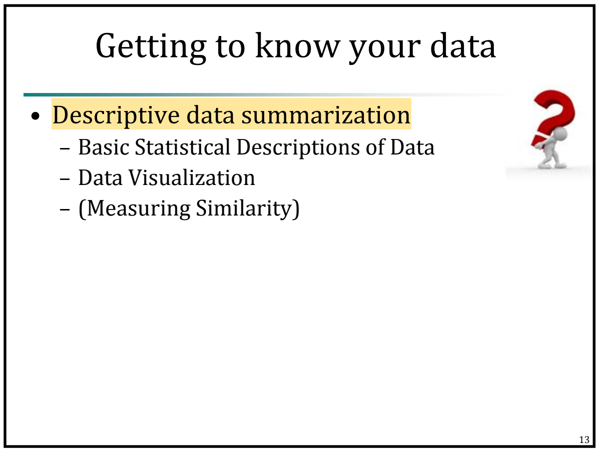


Descriptive data summarization

General idea: Get an overall picture of your data

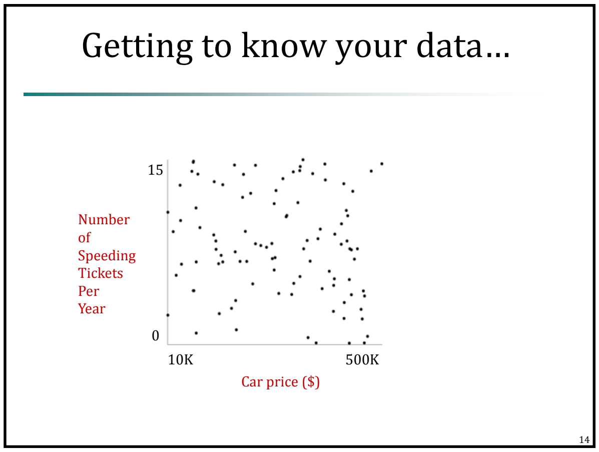
See how it is distributed; if there is skew, if it has a high variance, and so on

- Central tendencies
- Dispersion of data



Getting to know your data

- Descriptive data summarization
  - Basic Statistical Descriptions of Data
  - Data Visualization
  - (Measuring Similarity)



Getting to know your data…

Getting to know your data…

15

Number
of
Speeding
Tickets
Per
Year

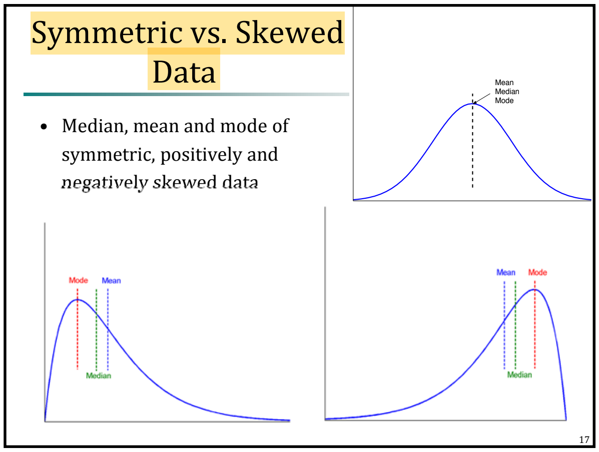
0

10K 500K

Car price ($)

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$
  - Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - Trimmed mean: chopping extreme values
- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*): $median = L_1 + (\frac{n/2 - (\sum f)l}{f_{median}})c$
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$



Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data

Measuring the Dispersion of Data

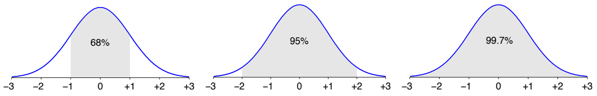
- **Quartiles, outliers and boxplots**
  - **Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)**
  - **Inter-quartile range: IQR = $Q_3 - Q_1$**
  - **Five number summary: min, Q_1, M, Q_3, max**
  - **Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually**
  - **Outlier: usually, a value higher/lower than 1.5 x IQR**
- **Variance and standard deviation (*sample: s, population: σ)***
  - **Variance: (algebraic, scalable computation)**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

  - **Standard deviation *s (or σ)* is the square root of variance *s^2 (or σ^2)***

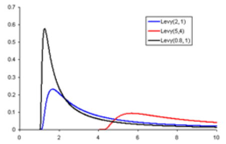
Properties of Normal Distribution Curve

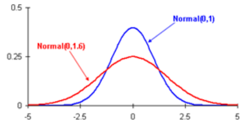
- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it



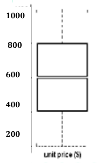
Normal distribution: A strong assumption?

- Very often, we assume a normal distribution
- What if it is not? (e.g. earthquake, financial markets, ketchup sales…)



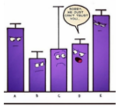


Boxplot Analysis

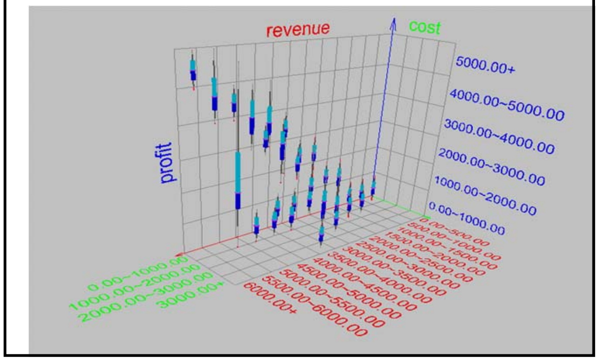


- Five-number summary of a distribution:

  Minimum, Q1, M, Q3, Maximum
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum

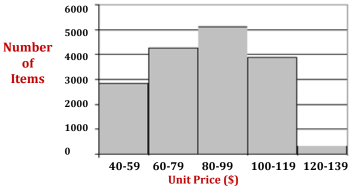


3D Visualization of Data Dispersion: Boxplot Analysis



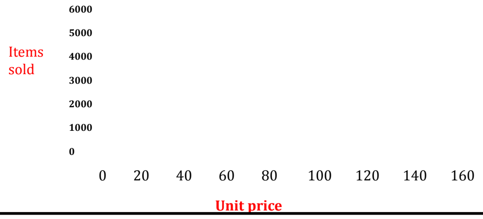
Histogram Analysis

- Graph displays of basic statistical class descriptions
  - Frequency histograms
    - A univariate graphical method
    - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data



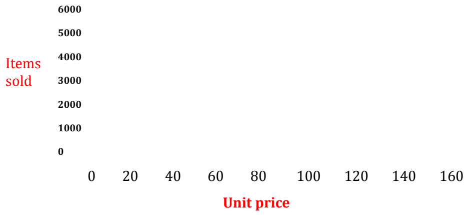
Scatter plot: Often used

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



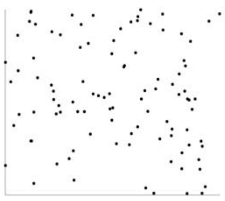
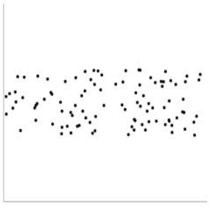
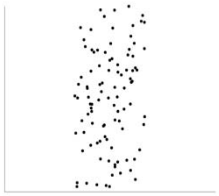
Loess (local regression) Curve

- 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



Positively and Negatively Correlated Data

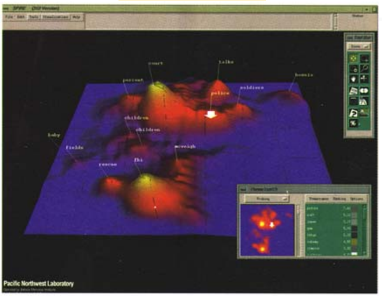
Uncorrelated Data

Data Visualization

- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations

Landscapes

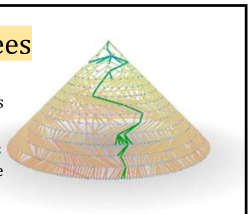


Used by permission of B. Wright, Visible Decisions Inc.

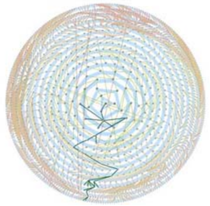
news articles visualized as a landscape

- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

Three-D Cone Trees



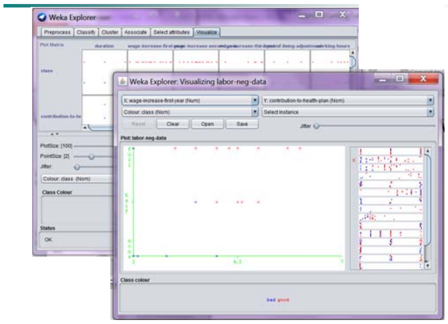
- *3D cone tree* visualization technique works well for up to a thousand nodes or so
- First build a *2D circle tree* that arranges its nodes in concentric circles centered on the root node



- Graph from Nadeau Software Consulting website: Visualize a social network data set that models the way an infection spreads from one person to the next

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

Scatterplots in WEKA (labor data)



PlotSize: [100]

PointSize: [2]

Colour: class (Nom)

OK

X: wage-increase-first-year (Num)

Y: contribution-to-health-plan (Nom)

Colour: class (Nom)

Select Instance

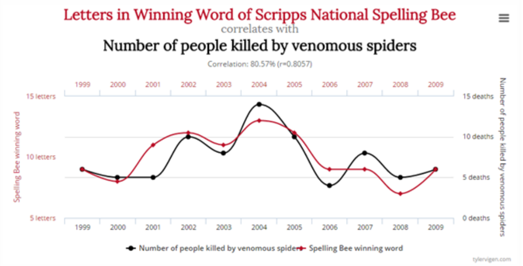
http://www.cs.waikato.ac.nz/ml/weka/

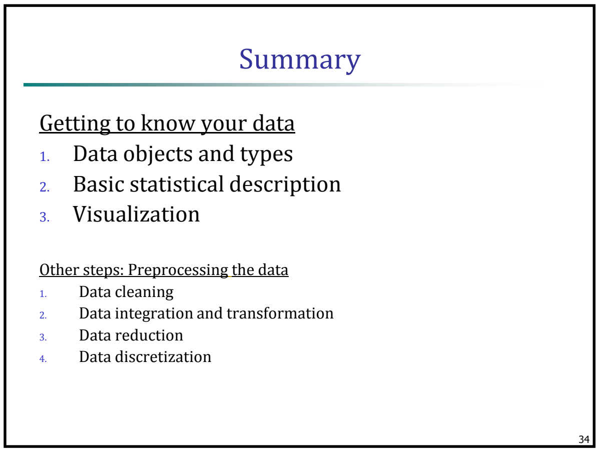
Descriptive Data Summarization

- Crucial initial steps
  - Basic statistical analysis
  - Visualization
  - (Measuring similarity: later)
- Gives us a “feeling” of our data
  - Relationships
  - Patterns
  - Trends

A word of caution…

- http://www.tylervigen.com/spurious-correlations
- We need to use our common sense!!!





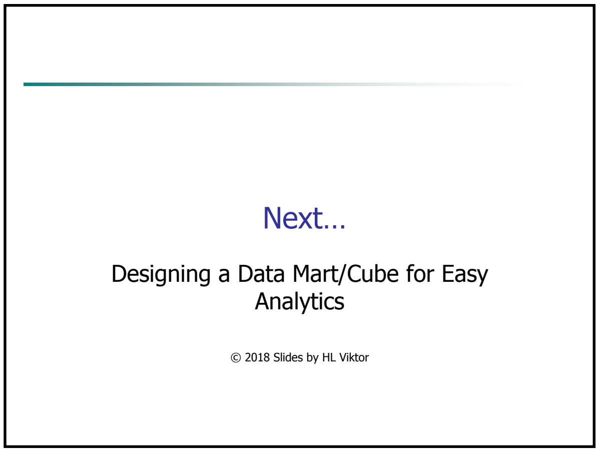
Summary

Getting to know your data

1. Data objects and types
2. Basic statistical description
3. Visualization

Other steps: Preprocessing the data

1. Data cleaning
2. Data integration and transformation
3. Data reduction
4. Data discretization



Next…

Designing a Data Mart/Cube for Easy Analytics