2018 EE448, Big Data Mining, Lecture 2

Fundamentals of Data Science
Know Your Data

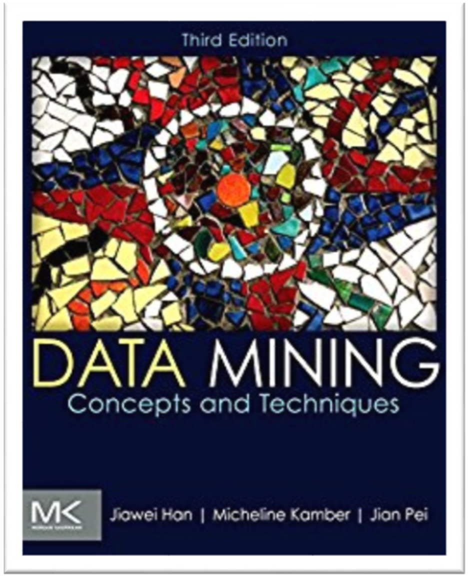
Weinan Zhang

Shanghai Jiao Tong University

http://wnzhang.net

http://wnzhang.net/teaching/ee448/index.html

References and Acknowledgement



- A large part of slides in this lecture are originally from Prof. Jiawei Han's book and lectures
  - http://hanj.cs.illinois.edu/bk3/bk3_slidesindex.htm
  - https://wiki.cites.illinois.edu/wiki/display/cs512/Lectures

Content

- Data Instances, Attributes and Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity

Data Instances

- Data sets are made up of data objects.
- A data object represents an entity.
- 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.

Data Instances

- A data instance represents an entity
  - Also called data points, data object

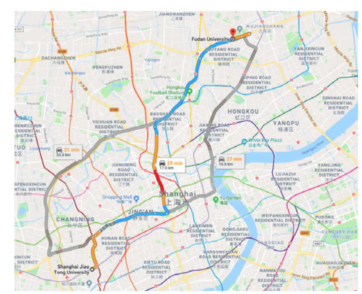
A news article

An image

A song

A Facebook user profile

A transcript of a student



A trajectory of a car from SJTU to FDU

Data Attributes

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
  - E.g., customer_ID, name, address

- Attribute Types
  - Nominal
  - Binary
  - Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- Nominal: categories, states, or "names of things"
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- 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., HIV positive)
- Ordinal
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings

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
  - No true zero-point
- 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., temperature in Kelvin, length, counts, monetary quantities

Discrete vs. Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- Continuous Attribute
  - 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
  - Continuous attributes are typically represented as floating-point variables

Data Attributes

- A data attribute is a particular field of a data instance
  - Also called dimension, feature, variable in difference literatures

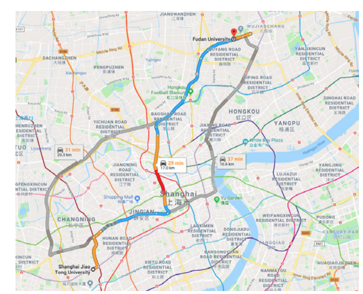
The frequency of ‘USA’ in a news article

The upper left pixel RGB value of an image

The pitch of the 320th frame of a song

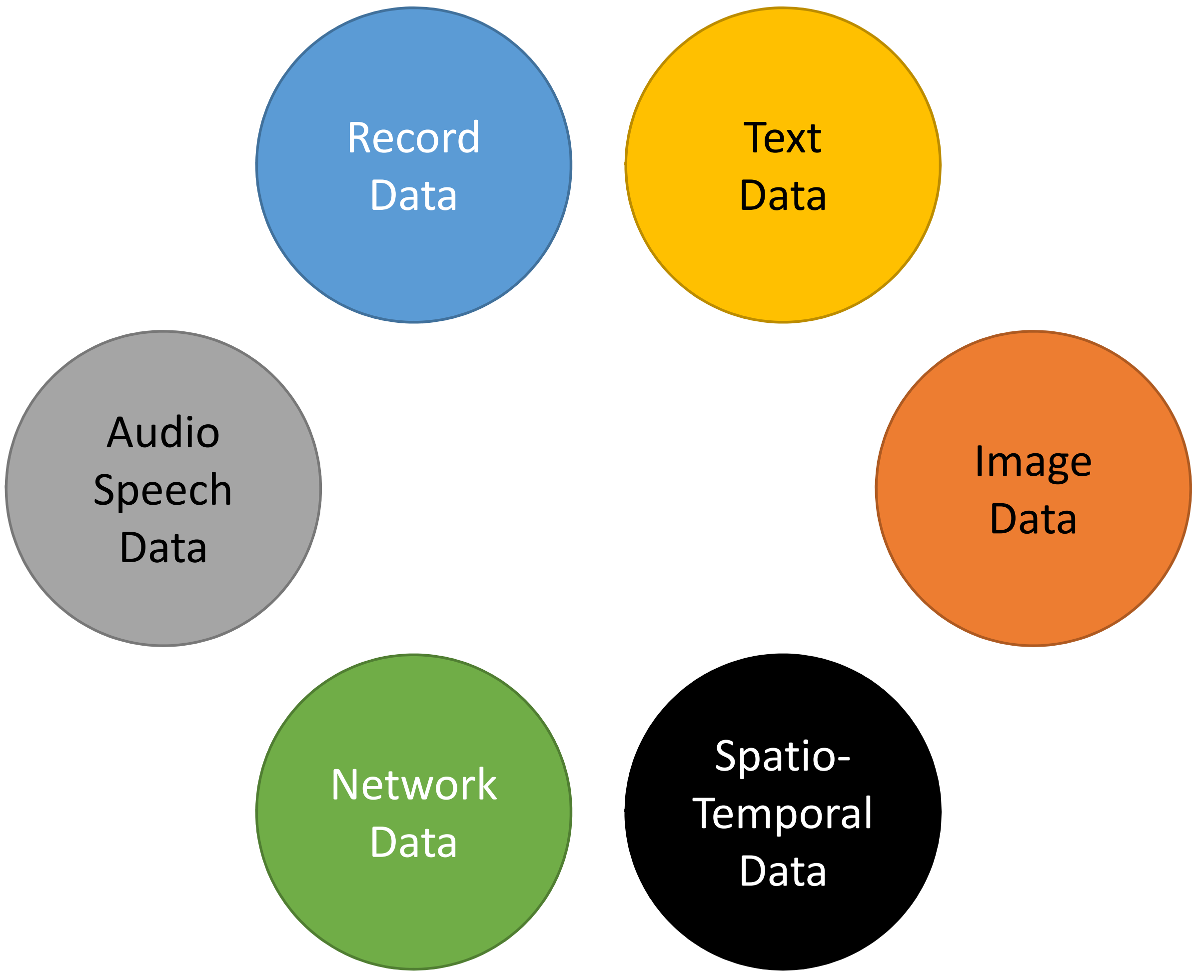
The friend set of a Facebook user

The Algebra score of a student’s transcript



The time-location of the 3rd point of a trajectory

6 Major Data Types



Data Type 1: Record Data

- Very common in relational databases
  - Each row represents a data instance
  - Each column represents a data attribute

WEEKDAY	GENDER	AGE	CITY
Tuesday	Male	28	London
Monday	Female	24	New York
Tuesday	Female	36	Hong Kong
Thursday	Male	17	Tokyo

JSON Format:
```
{
    WEEKDAY: Monday;
    GENDER: Female;
    AGE: 24;
    CITY: New York;
}
```

- Term 'KDD': Knowledge discovery in databases

Data Type 2: Text Data

- A sequence of words/tokens that represents semantic meanings of human

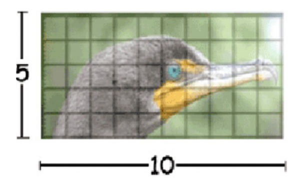
```
Text mining, also referred
to as text data mining,
roughly equivalent to text
analytics, is the process
of deriving high-quality
information from text.
```

Bag-of-Words Format:

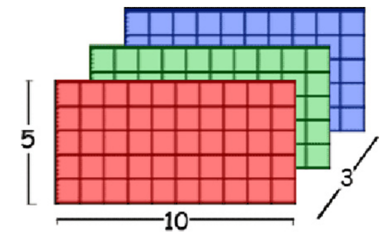
```
{
    text: 4;
    mining: 2;
    also: 1;
    referred: 1;
    to: 2;
    as: 1;
    data: 1;
    roughly: 1;
    equivalent: 1;
    analytics: 1;
    is: 1;
    the: 1;
    process: 1;
    of: 1;
    deriving: 1;
    high-quality: 1;
    information: 1;
    from: 1;
}
```

Data Type 3: Image Data

- A 3-layer matrix (3*height*width) of [0,255] real value

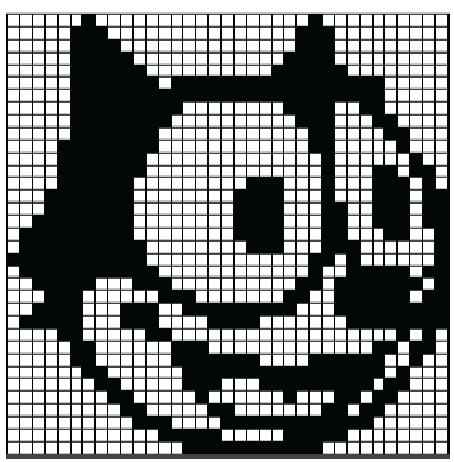


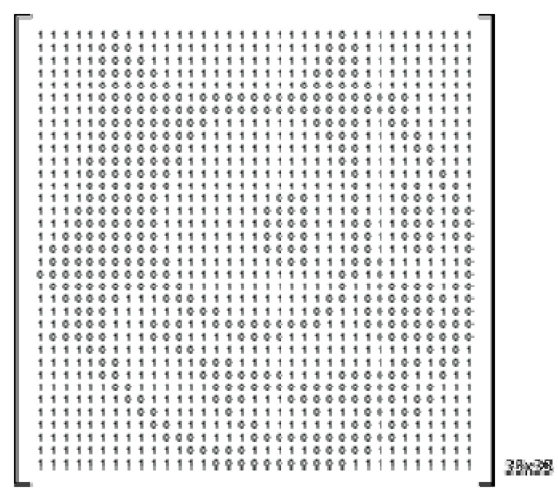
Original Color Image



Matlab RGB Matrix

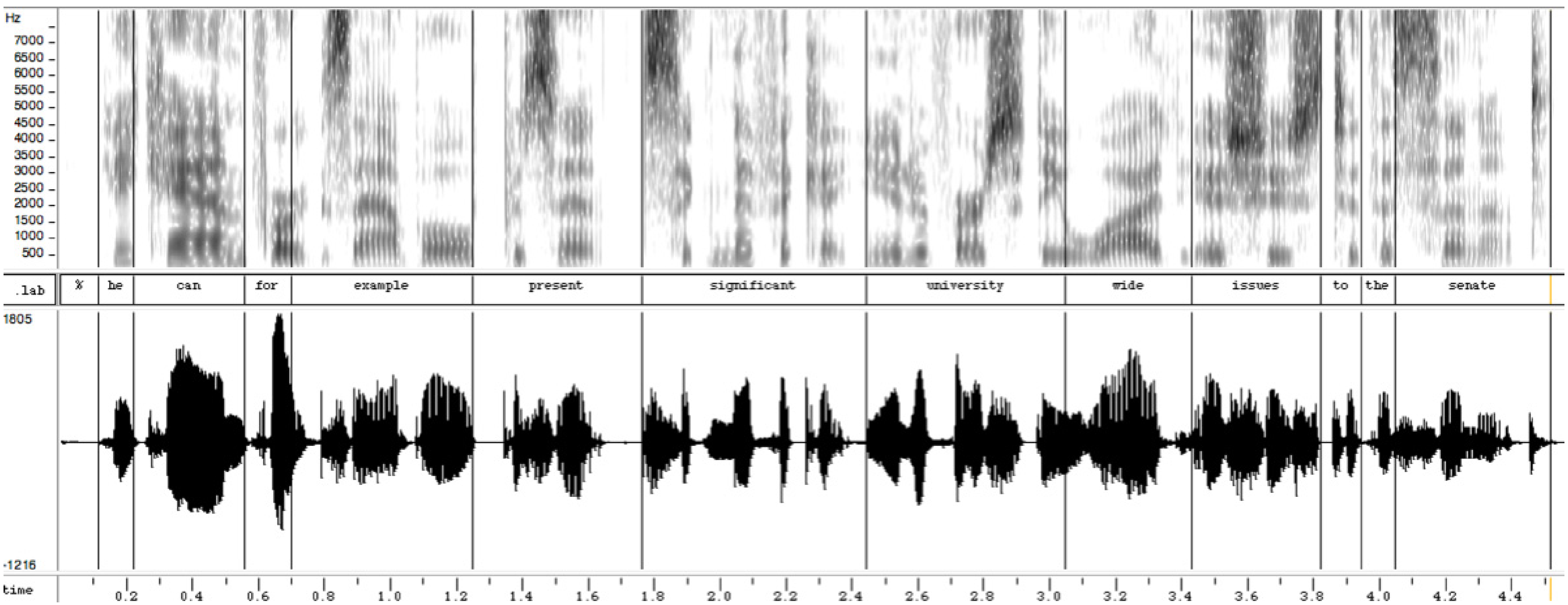
- A simple case: binary image
  - 1-layer matrix (height*width) of {0,1} binary value





Data Type 4: Speech Data

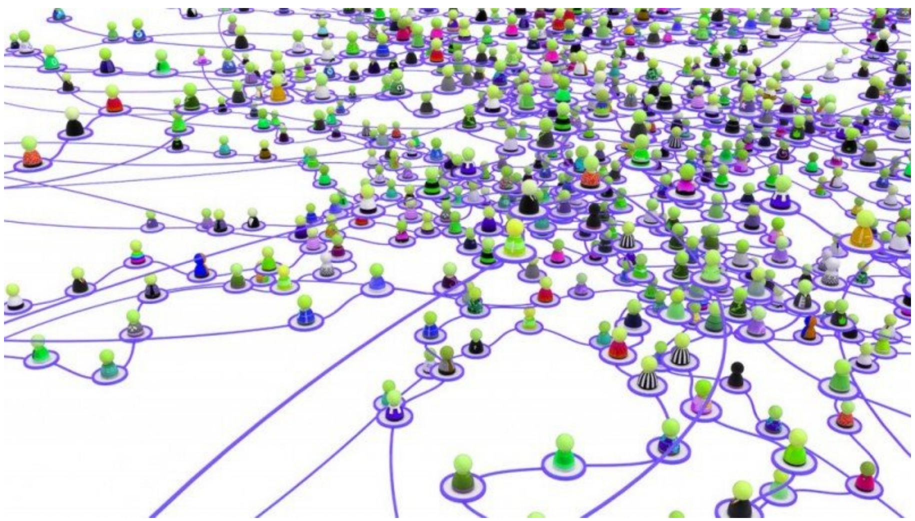
- A sequence of multi-dimensional real vectors
  - Directly decoding from the audio/speech data



http://languagelog.ldc.upenn.edu/nll/?p=8116

Data Type 5: Network Data

- A directed/undirected graph
  - Possibly with additional information for nodes and edges



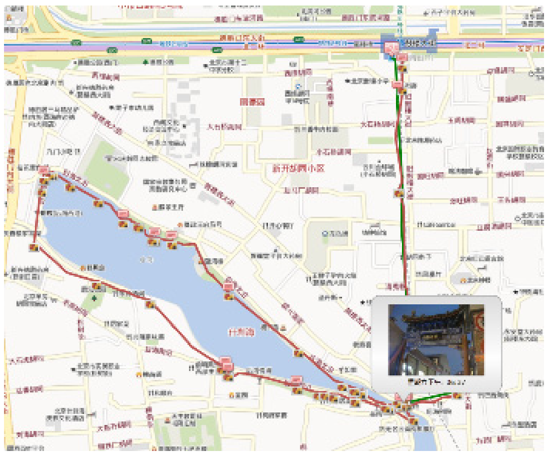
Friendship Format:

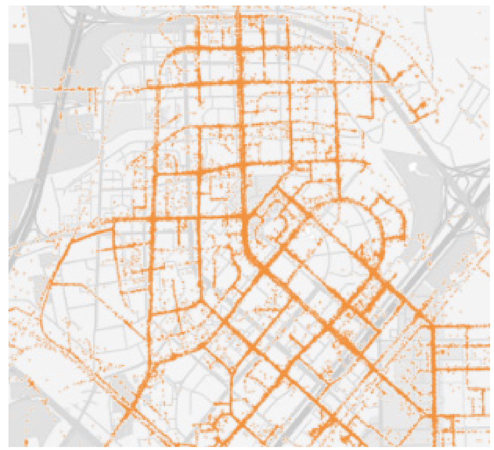
Alice	Bob
Bob	Carl
Carl	Victor
Bob	Victor
Alice	Victor
…	

Stanford network dataset collection: https://snap.stanford.edu/data/

Data Type 6: Spatio-Temporal Data

- A sequence of (time, location, info) tuples

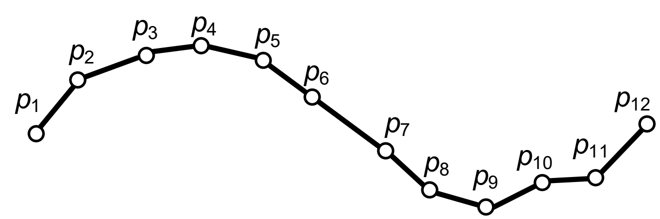




- A spatio-temporal trajectory

$$p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$$

$$p_i = (t, x, y, a)$$



- Time series data is a special case of ST data
  - without location information $p_i = (t, a)$

https://www.microsoft.com/en-us/research/project/trajectory-data-mining/

Slide credit: Yu Zheng

Content

- Data Instances, Attributes and Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - Median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population)

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - Weighted arithmetic mean:

$$\mu = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Trimmed mean: chopping extreme values
- Median
  - Middle value if odd number of values, or average of the middle two values otherwise
- Example
  - Five data points {1.2, 1.4, 1.5, 1.8, 10.2}
  - Mean: 3.22   Median: 1.5

Measuring the Central Tendency

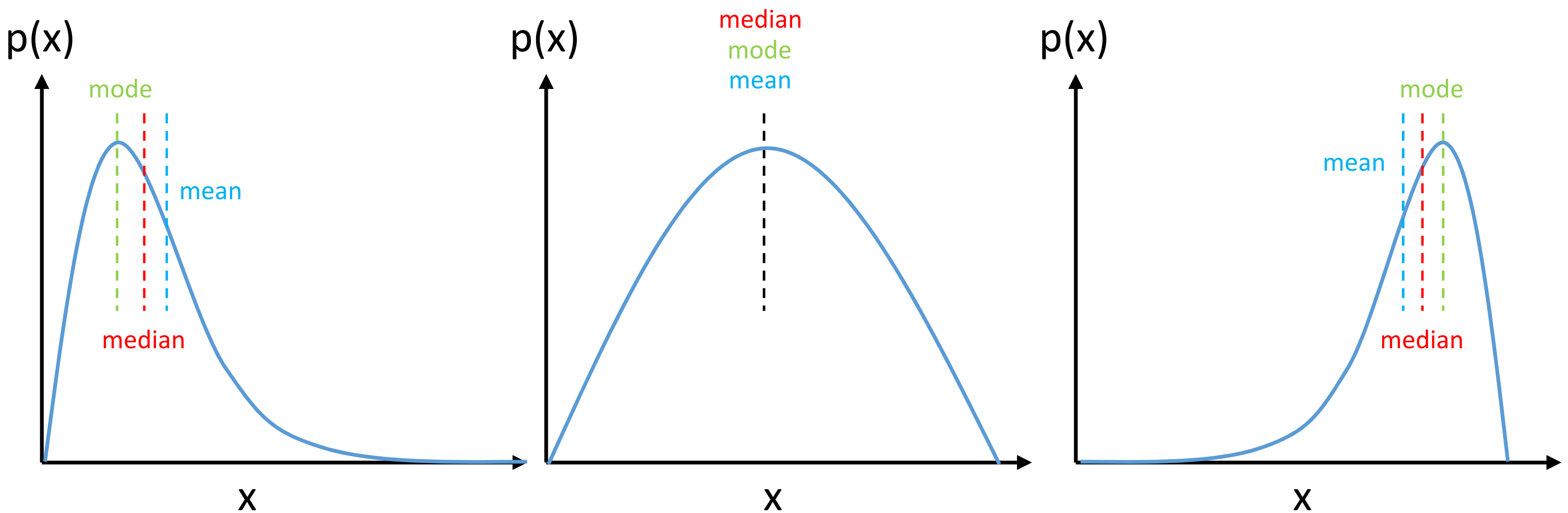
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

$$\text{mean} - \text{mode} \simeq 3 \times (\text{mean} - \text{median})$$

- Example
  - Five data points {1, 1, 1, 1, 1, 2, 2, 2, 3, 3}
  - Mean: 1.7   Median: 1.5   Mode: 1

Symmetric vs. Skewed Data

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



Positively skewed data

mode < median

Symmetric data

mode = median

Negatively skewed data

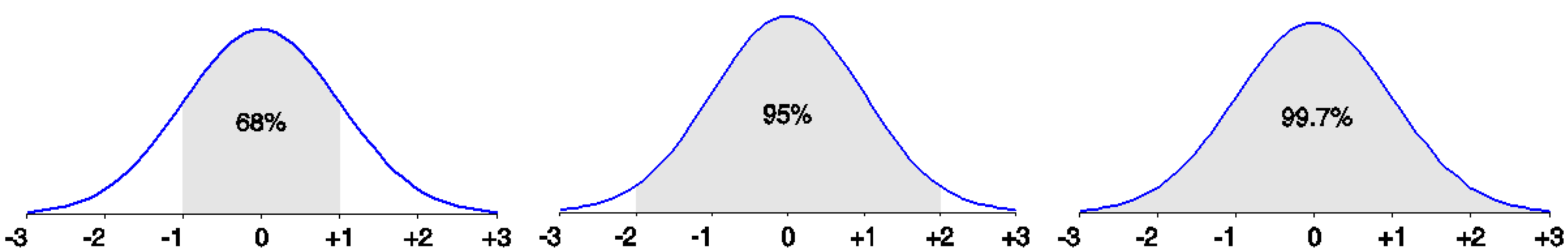
mode > median

Measuring the Dispersion of Data

- Variance and standard deviation
  - Variance

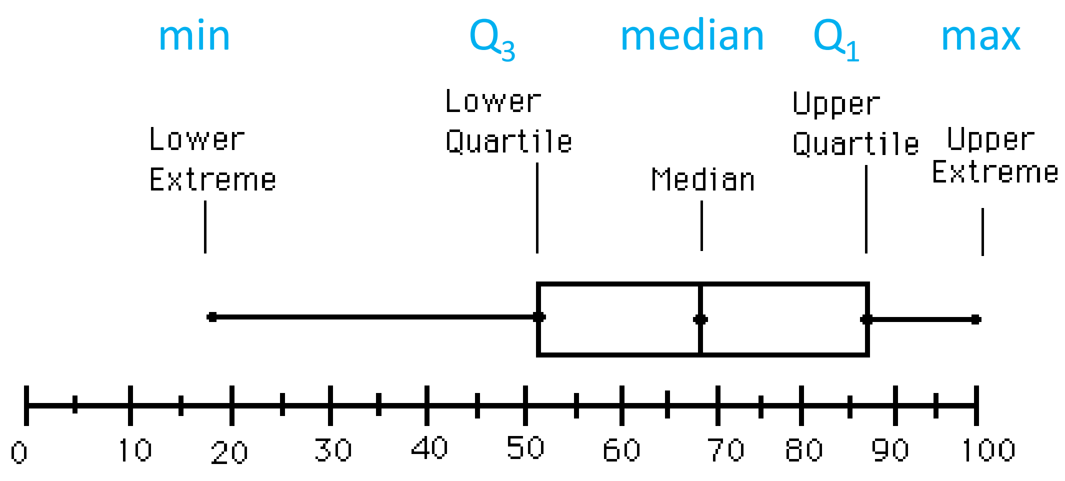
$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \mathbb{E}[x] \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

  - Standard deviation σ is the square root of variance σ^2
- The normal (distribution) curve
  - From $\mu-\sigma$ to $\mu+\sigma$: contains about 68% of the measurements
  - From $\mu-2\sigma$ to $\mu+2\sigma$: contains about 95% of it
  - From $\mu-3\sigma$ to $\mu+3\sigma$: contains about 99.7% of it



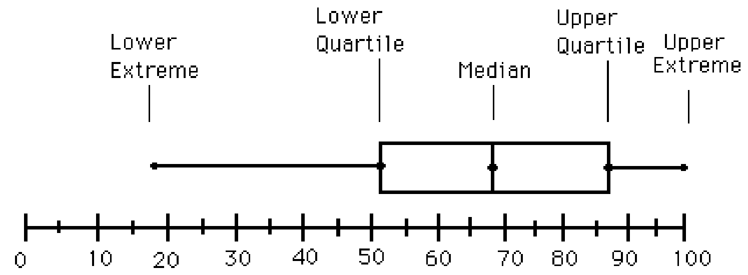
Measuring the Dispersion of Data

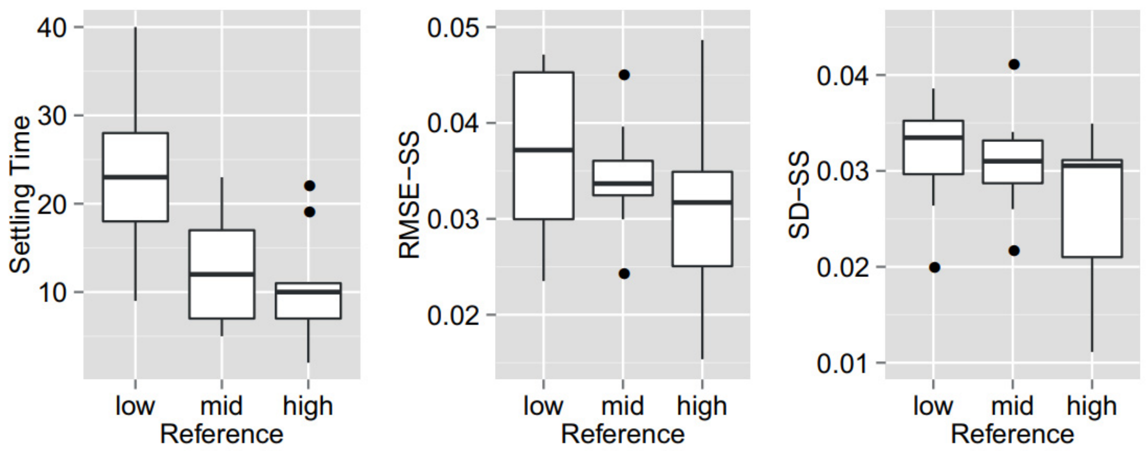
- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR



Boxplot Analysis

- Five-number summary of a distribution
  - Minimum, Q1, Median, 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 IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually





Content

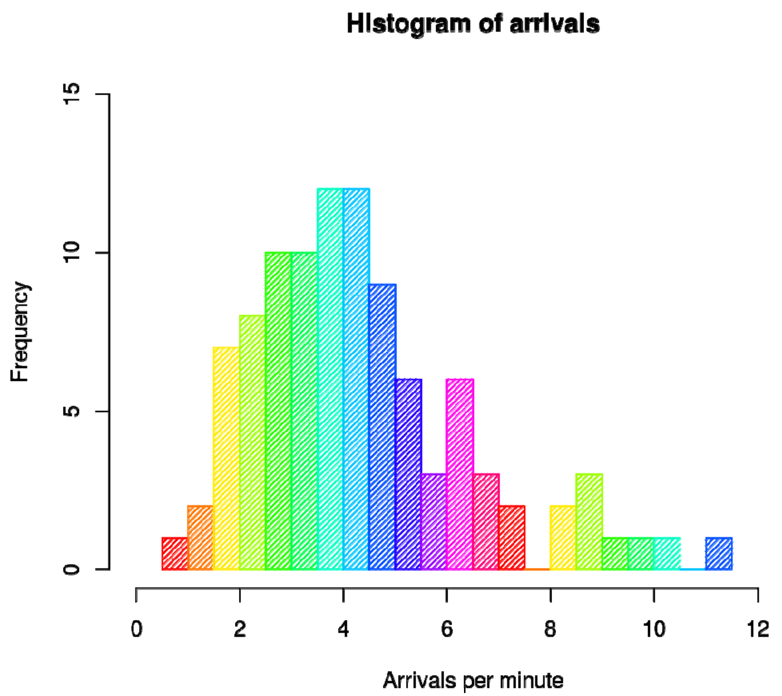
- Data Instances, Attributes and Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity

Graphic Displays of Basic Statistical Descriptions

- Boxplot: graphic display of five-number summary
- Histogram: x-axis are values, y-axis represents frequencies
- Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

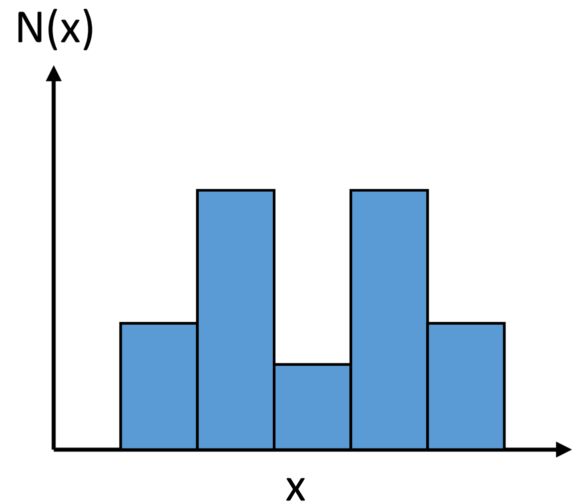
Histogram Analysis

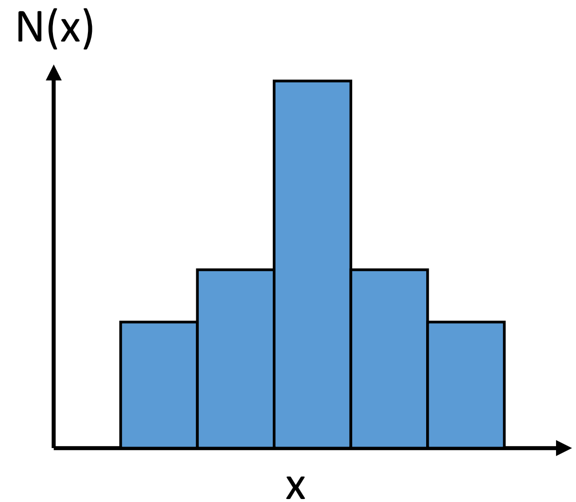
- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Histograms Often Tell More than Boxplots

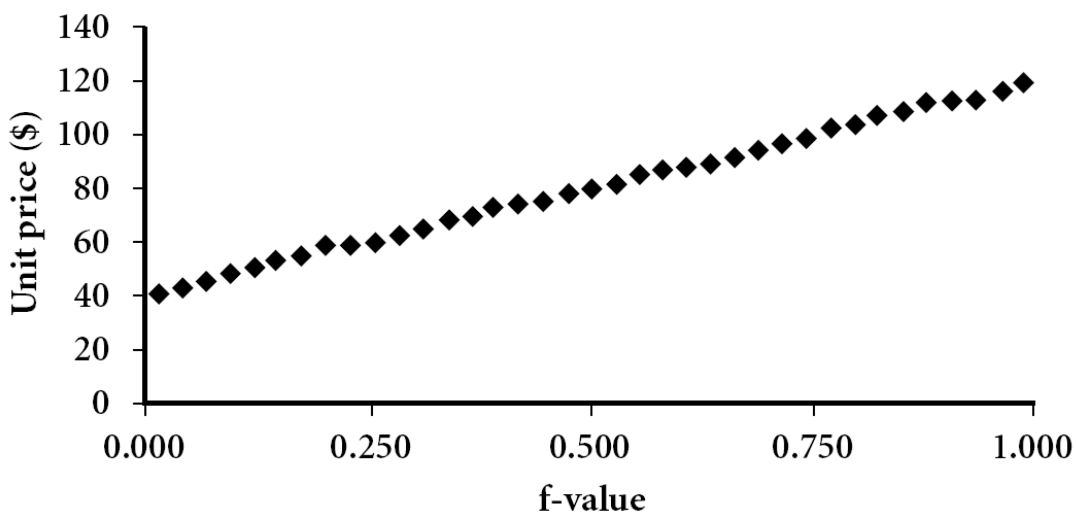
- The two histograms shown on the right may have the same boxplot representation
- The same values for: min, Q_1, median, Q_3, max
- But they have rather different data distributions





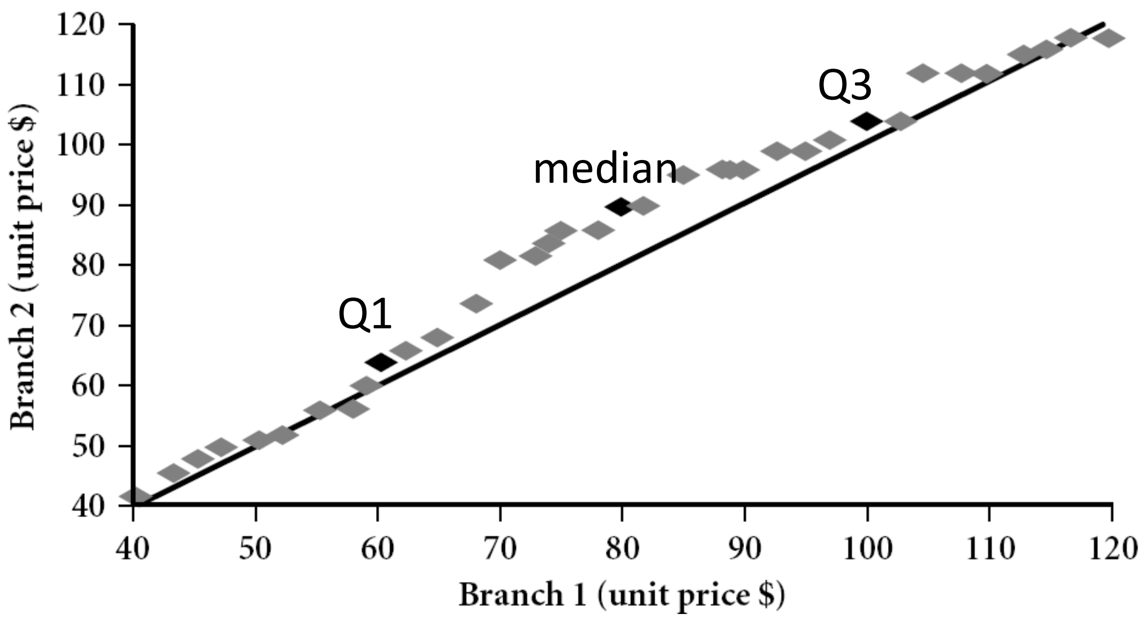
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
- Each value x_i is paired with f_i indicating that approximately 100 f_i% of data $\leq x_i$



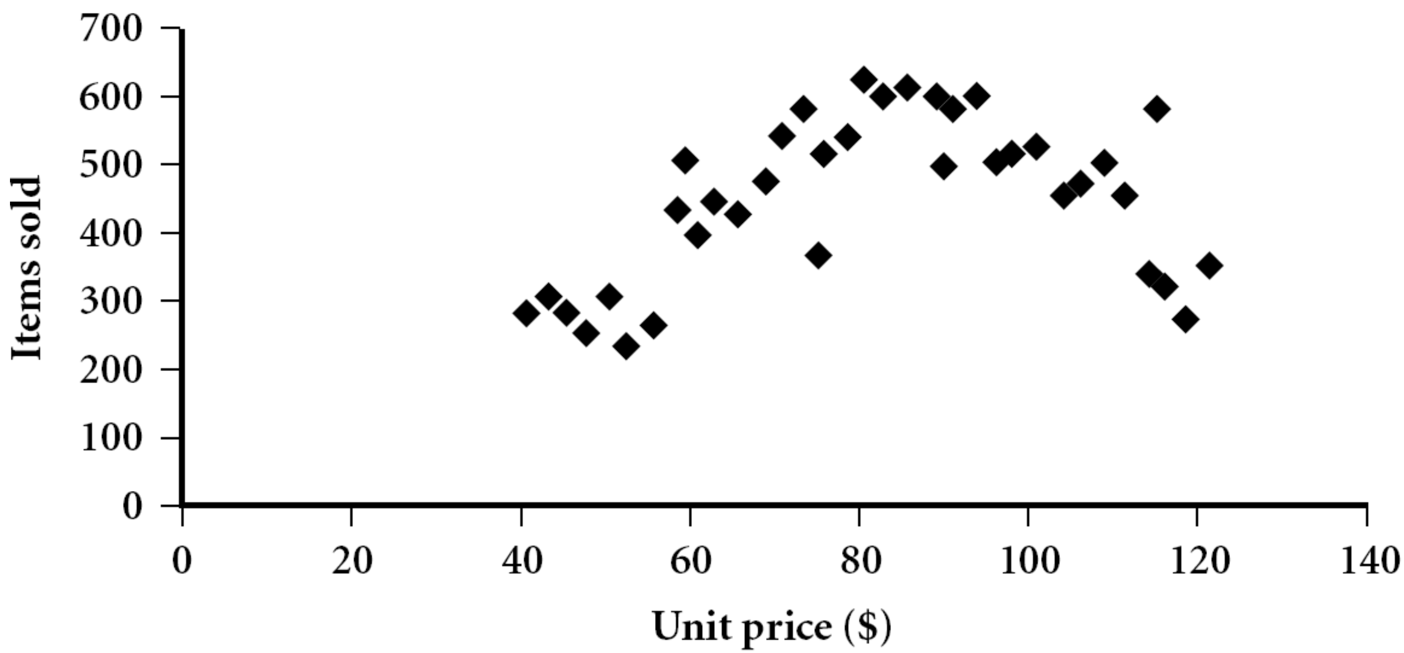
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



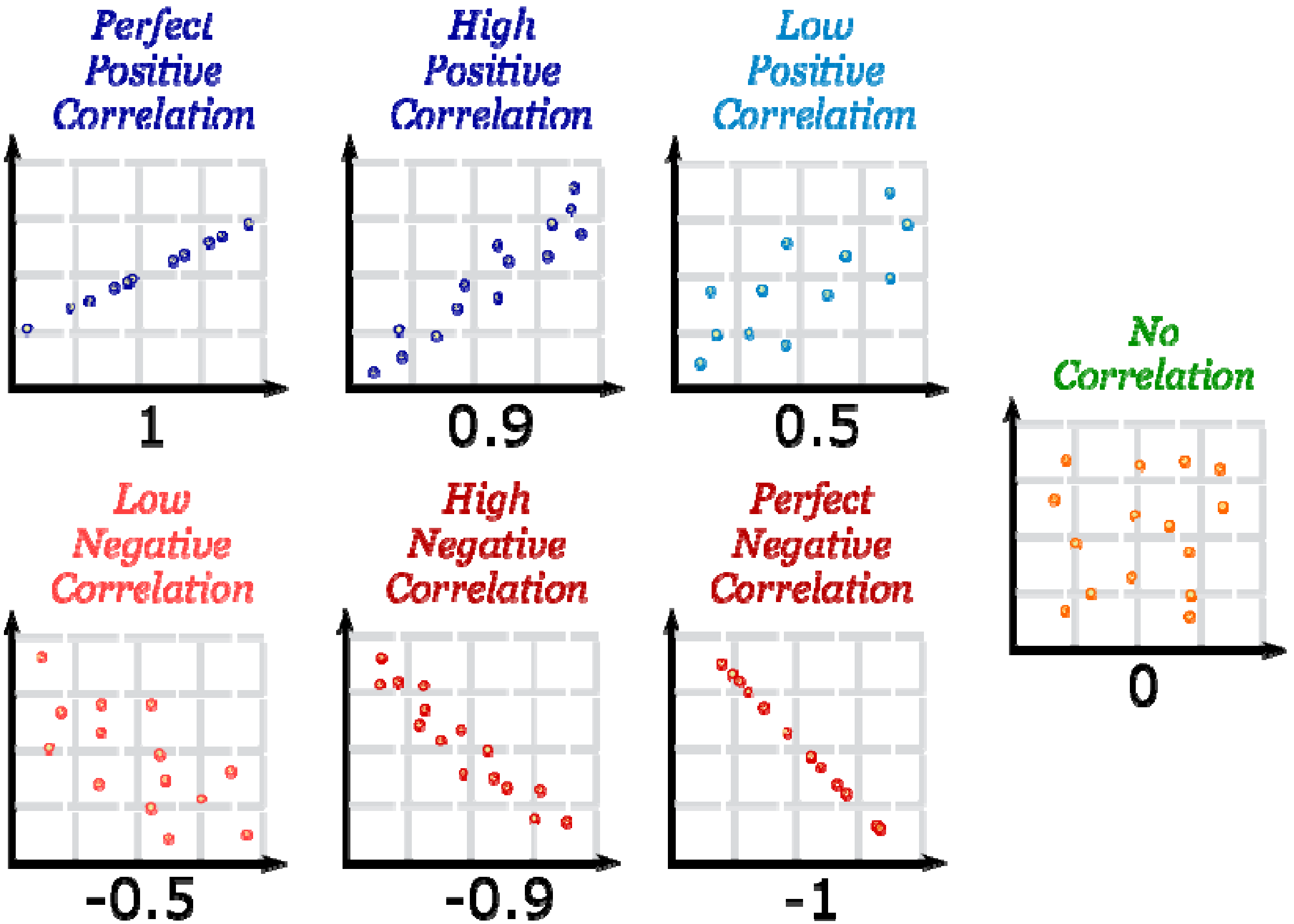
Scatter Plot

- 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



Positively and Negatively Correlated Data

- One can also quickly check the correlation of the two variables by scatter data.



Data Visualization

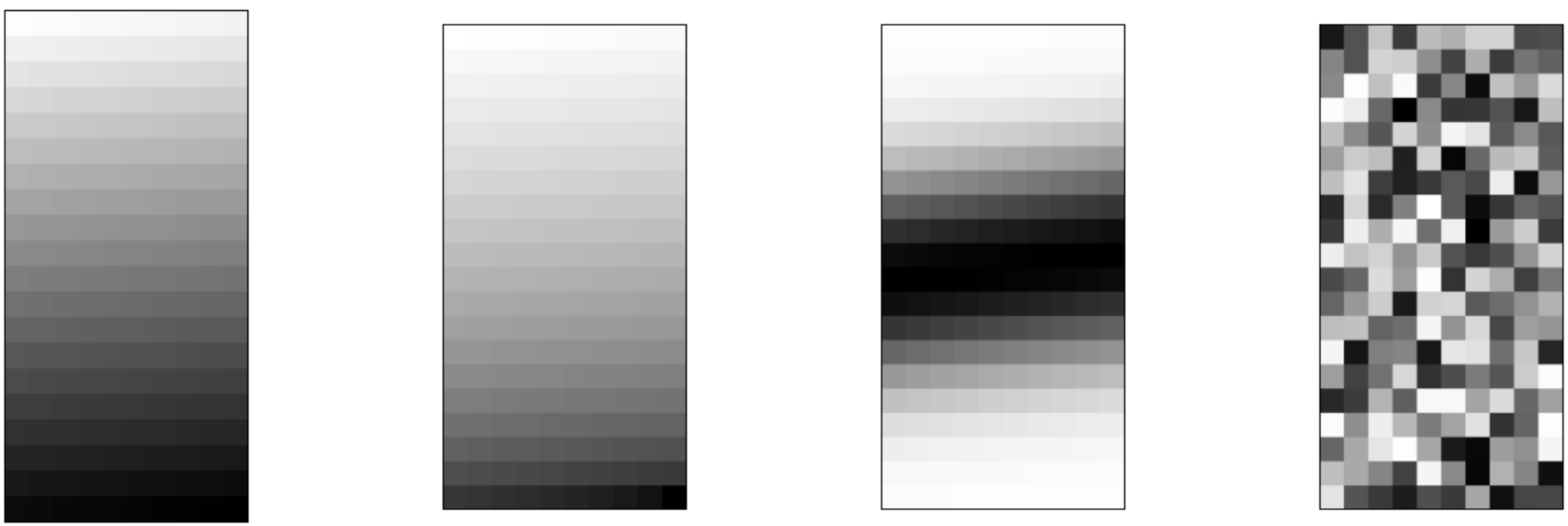
- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived

Data Visualization

- Different of visualization methods include
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations
  - Visualizing decision-making data
  - …

Pixel-Oriented Visualization Techniques

- For a data set of *m* dimensions, create *m* windows on the screen, one for each dimension
- The *m* dimension values of a record are mapped to *m* pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



(a) Income (b) Credit Limit (c) Transaction volume (d) Age

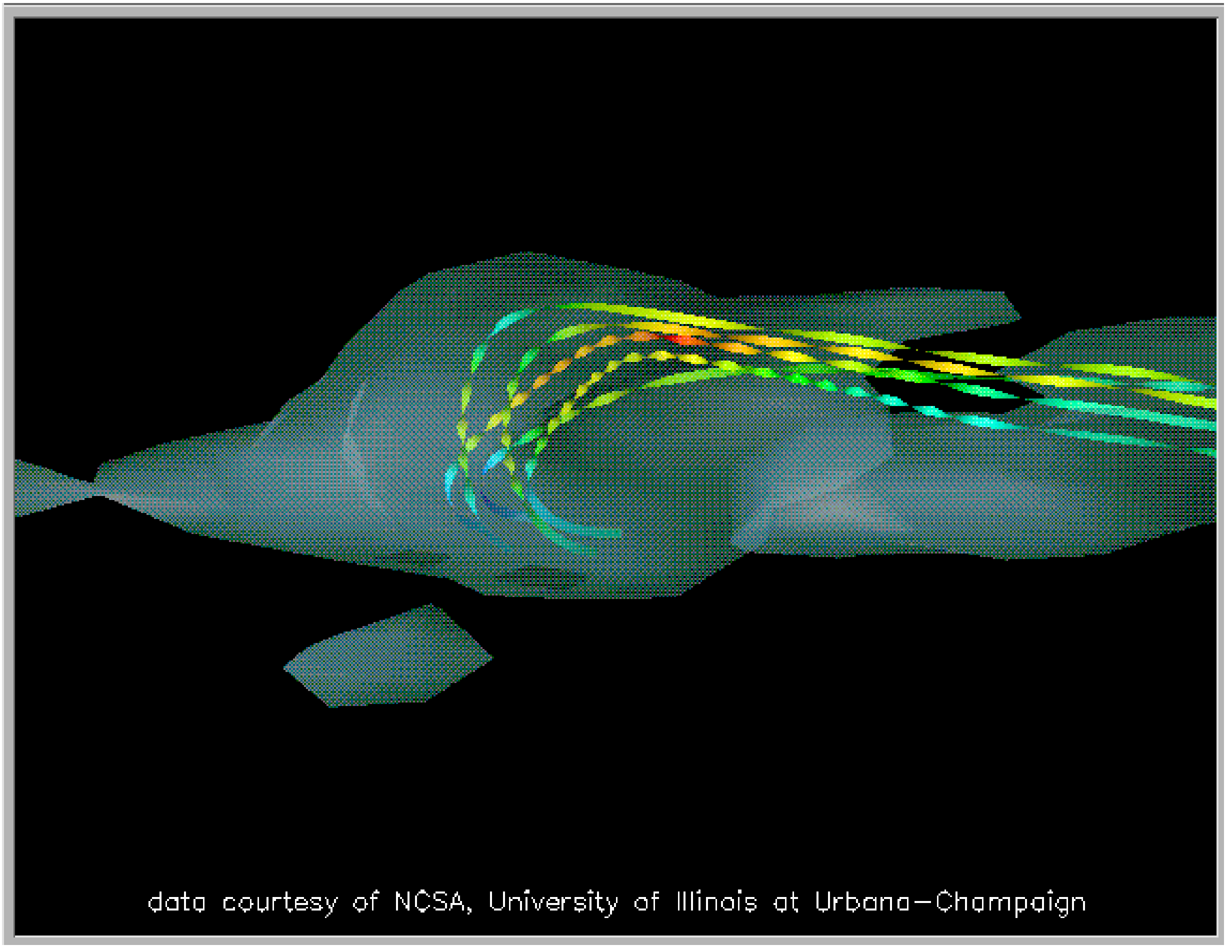
Note: here the m windows are arranged by income. We can check the correlations of other dimension data w.r.t. income.

Geometric Projection Visualization Techniques

- Visualization of geometric transformations and projections of the data
- Methods
  - Direct visualization
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Projection pursuit technique: Help users find meaningful projections of multidimensional data
  - Prosection views
  - Hyperslice
  - Parallel coordinates

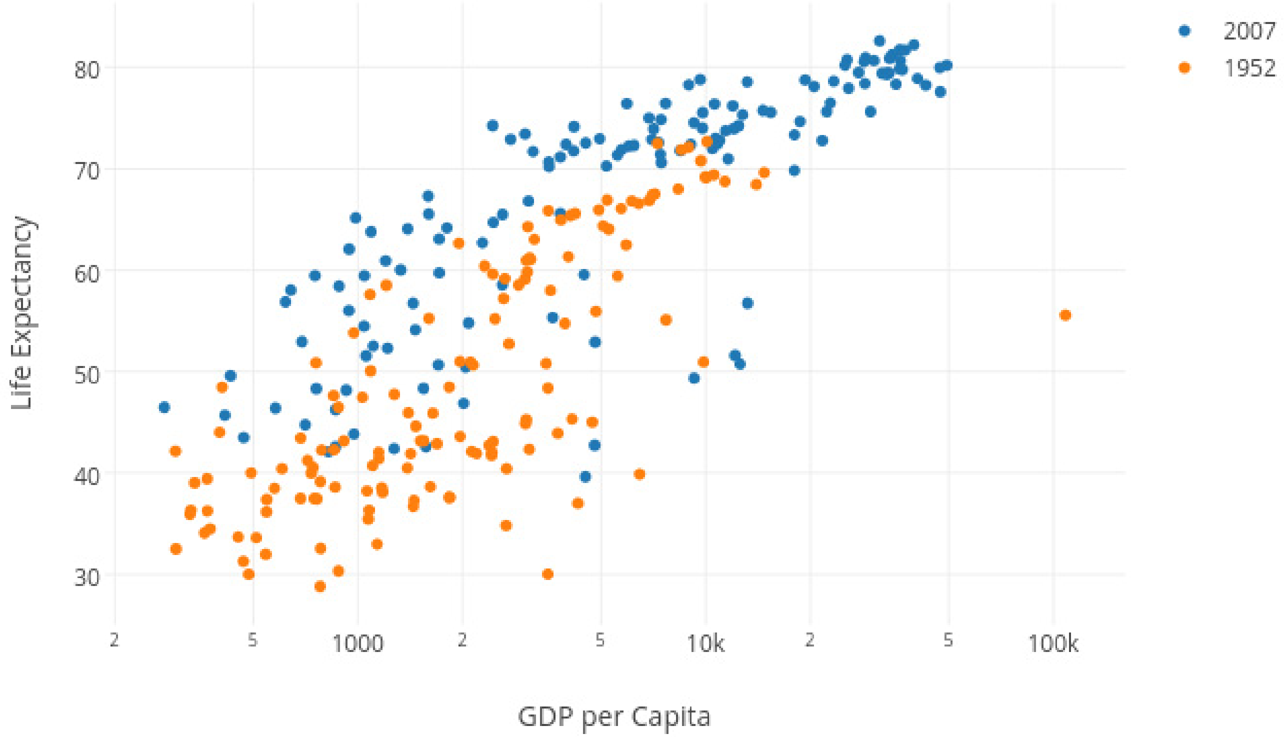
Direct Data Visualization

- Ribbons with Twists Based on Vorticity



Scatter Plots

- Scatter plot with category of data points in colors

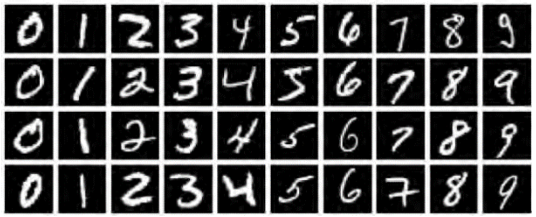


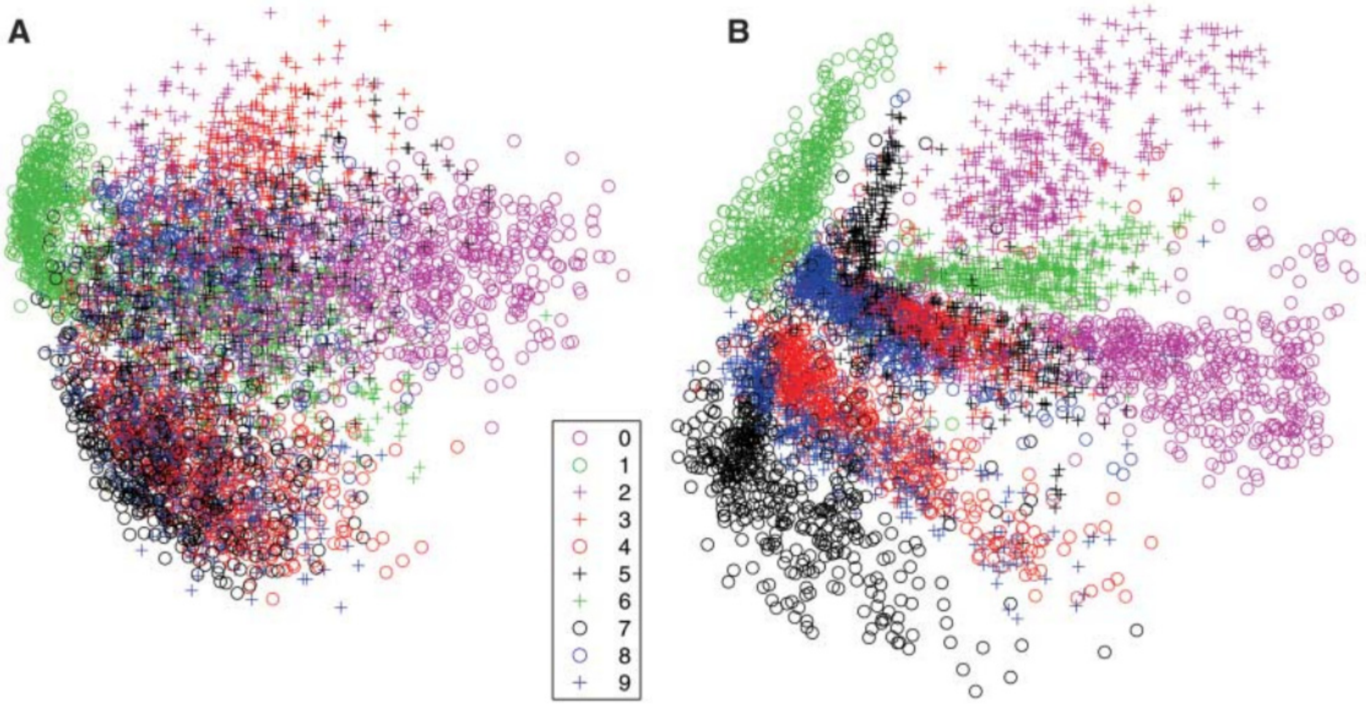
https://plot.ly/pandas/line-and-scatter/

Scatter Plots

MNIST data of hand written numbers

- 60,000 training images
- 28×28 pixels for each image





(A) The two-dimensional codes for 500 digits of each class produced by taking the first two principal components

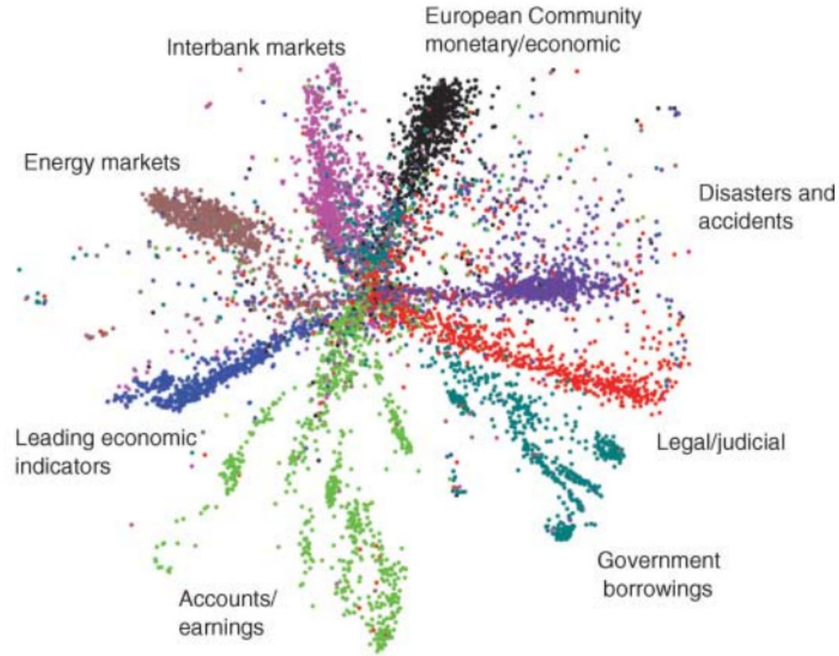
(B) The two-dimensional codes found by a 784-1000-500-250-2 autoencoder (a deep learning model).

Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *science*313.5786 (2006): 504-507.

Scatter Plots

The Reuter Corpus Volume 2

- 804,414 newswire stories
- 2000 commonest word stems

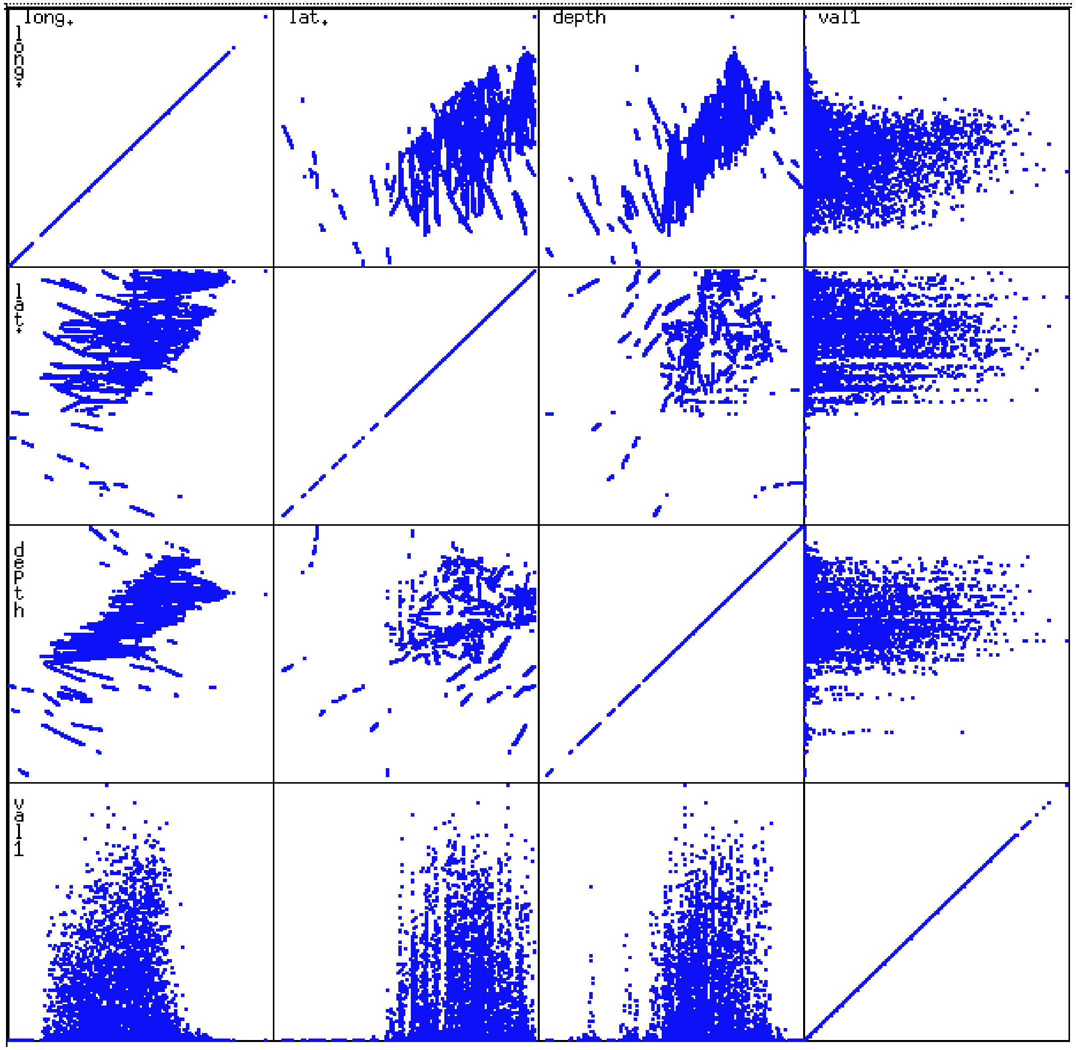


(A) The codes produced by two-dimensional latent semantic analysis (LSA).

(B) The codes produced by a 2000-500-250-125-2 autoencoder. (a deep learning model).

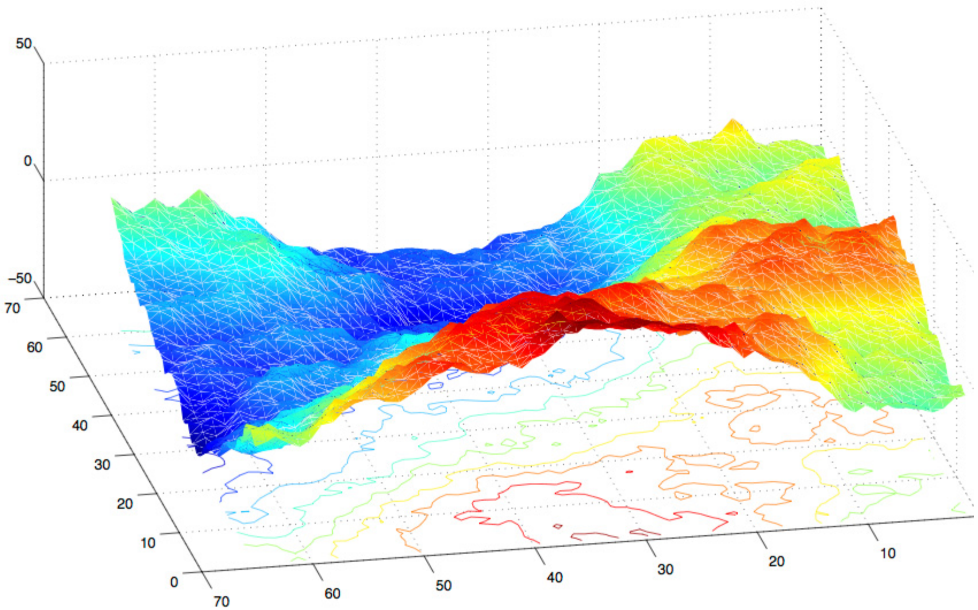
Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *science*313.5786 (2006): 504-507.

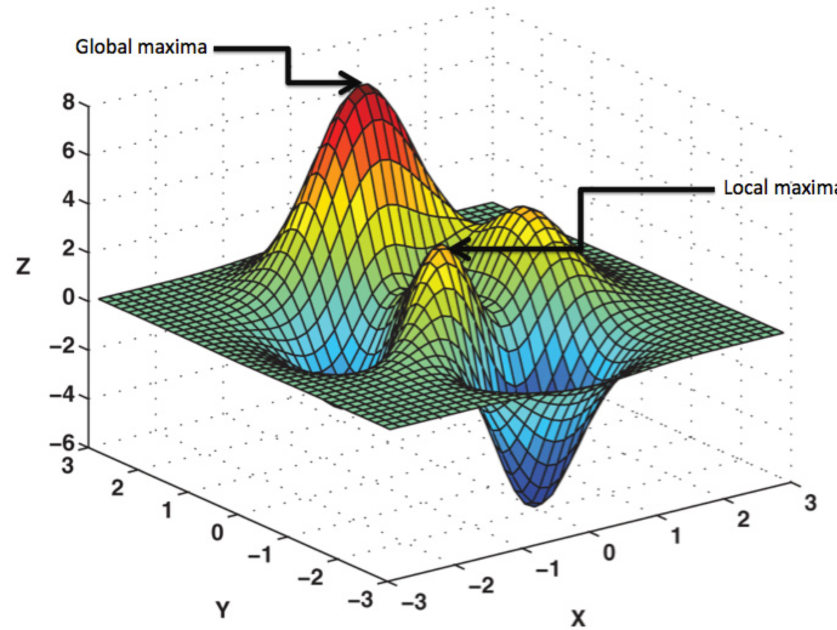
Scatterplot Matrices



Matrix of scatterplots (x-y-diagrams) of the k-dimensional data

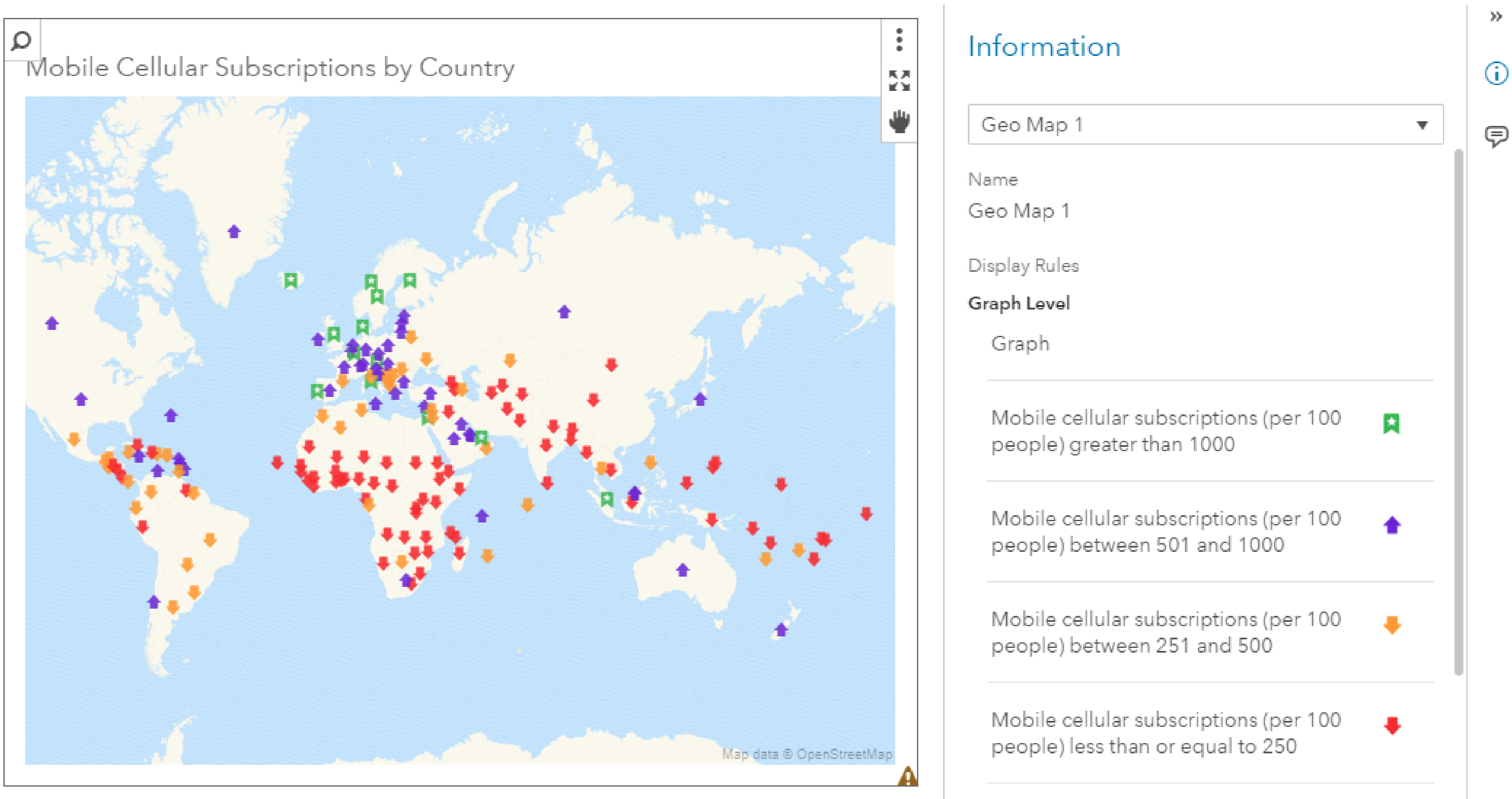
Landscapes





- 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

Icon based Visualization

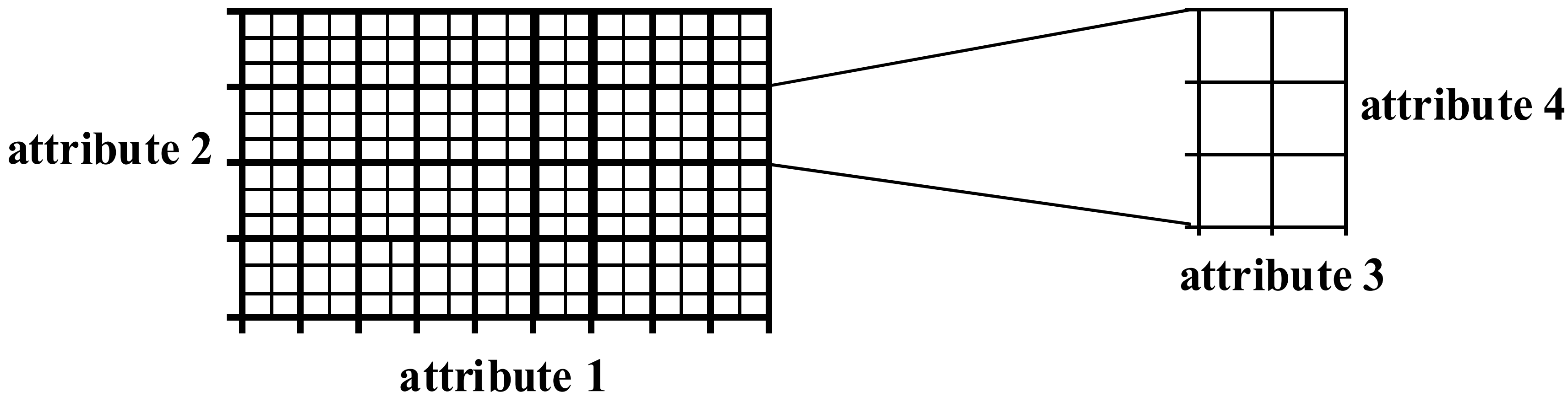


https://blogs.sas.com/content/sgf/2018/02/06/jazz-geo-map-colorful-icon-based-display-rules/

Hierarchical Visualization Techniques

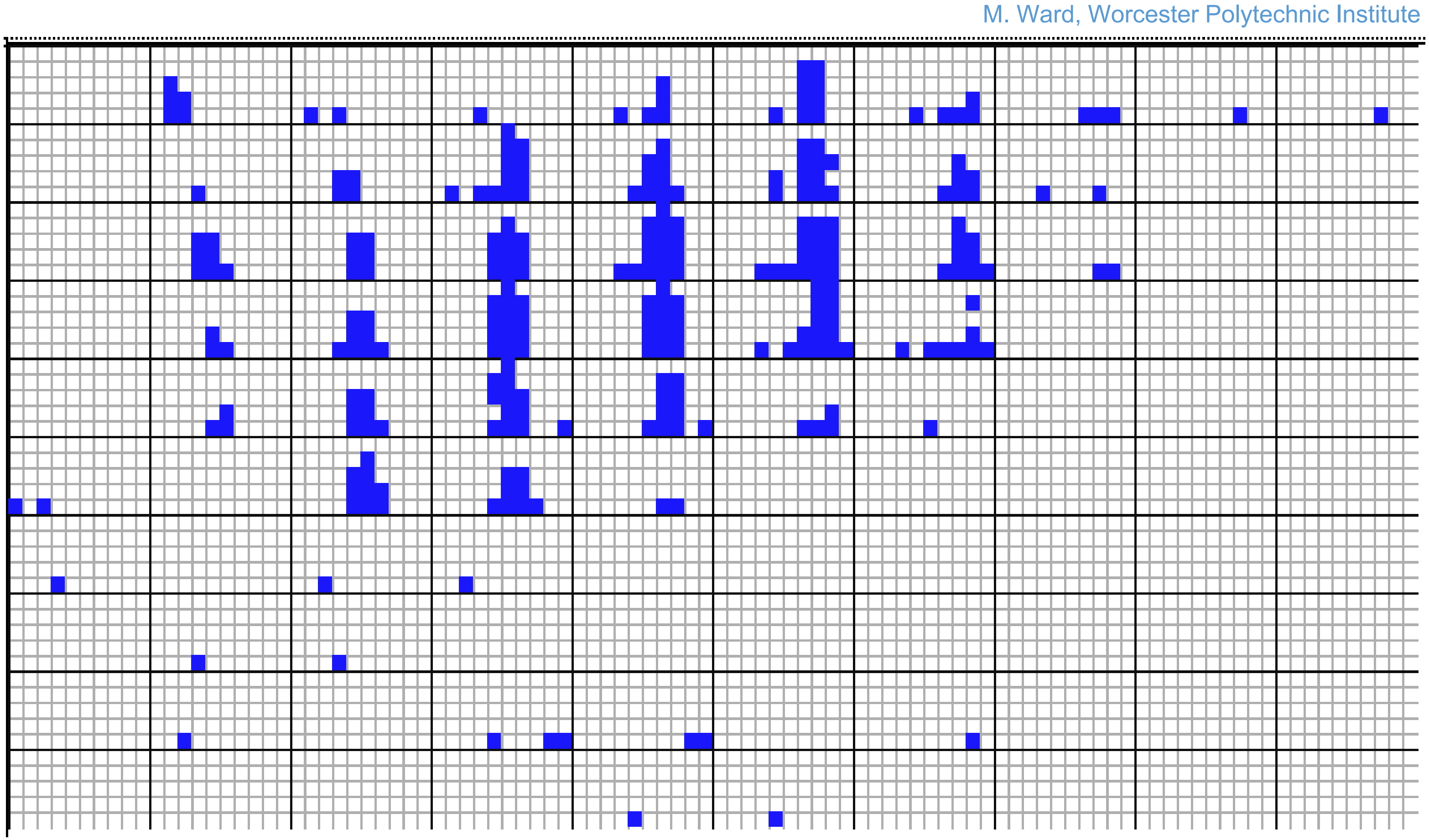
- Visualization of the data using a hierarchical partitioning into subspaces
- Methods
  - Dimensional Stacking
  - Worlds-within-Worlds
  - Tree-Map
  - Cone Trees
  - InfoCube

Dimensional Stacking



- Partitioning of the *n*-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately

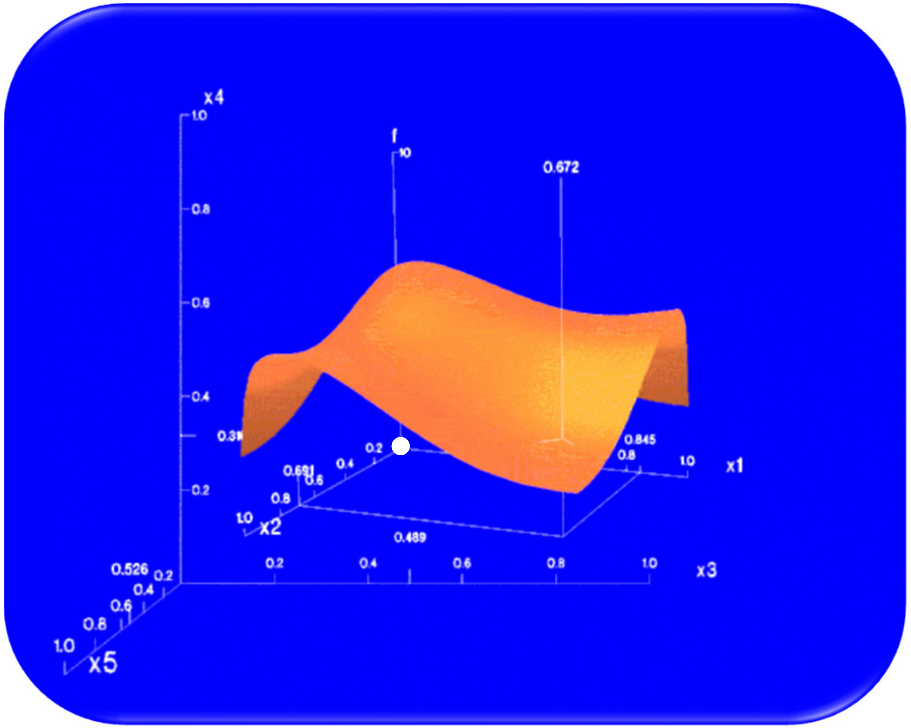
Dimensional Stacking



- Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

Worlds-within-Worlds Visualization

- Assign the function and two most important parameters to innermost world
- Fix all other parameters at constant values - draw other (1 or 2 or 3 dimensional worlds choosing these as the axes)
- Software that uses this paradigm
  - N–vision: Dynamic interaction through data glove and stereo displays, including rotation, scaling (inner) and translation (inner/outer)
  - Auto Visual: Static interaction by means of queries

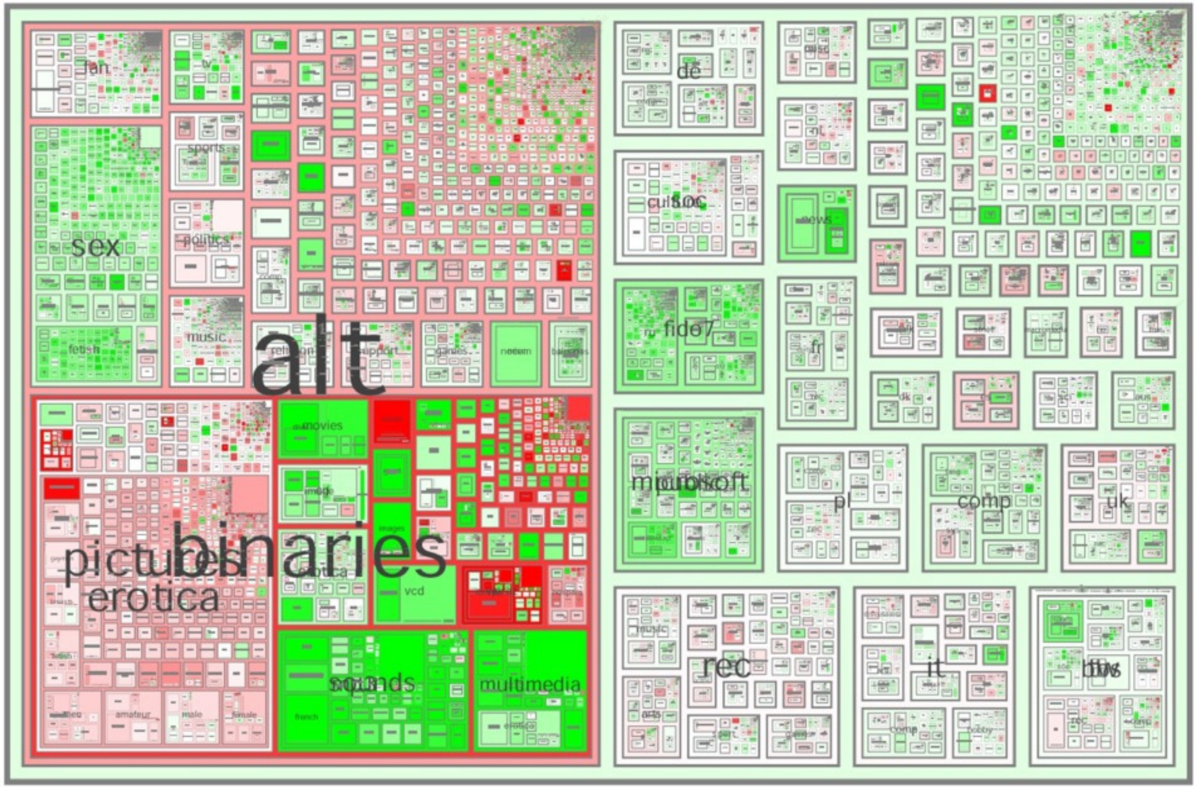


Tree-Map

http://www.cs.umd.edu/hcil/treemap-history/

- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values
- The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)

MSR Netscan Image



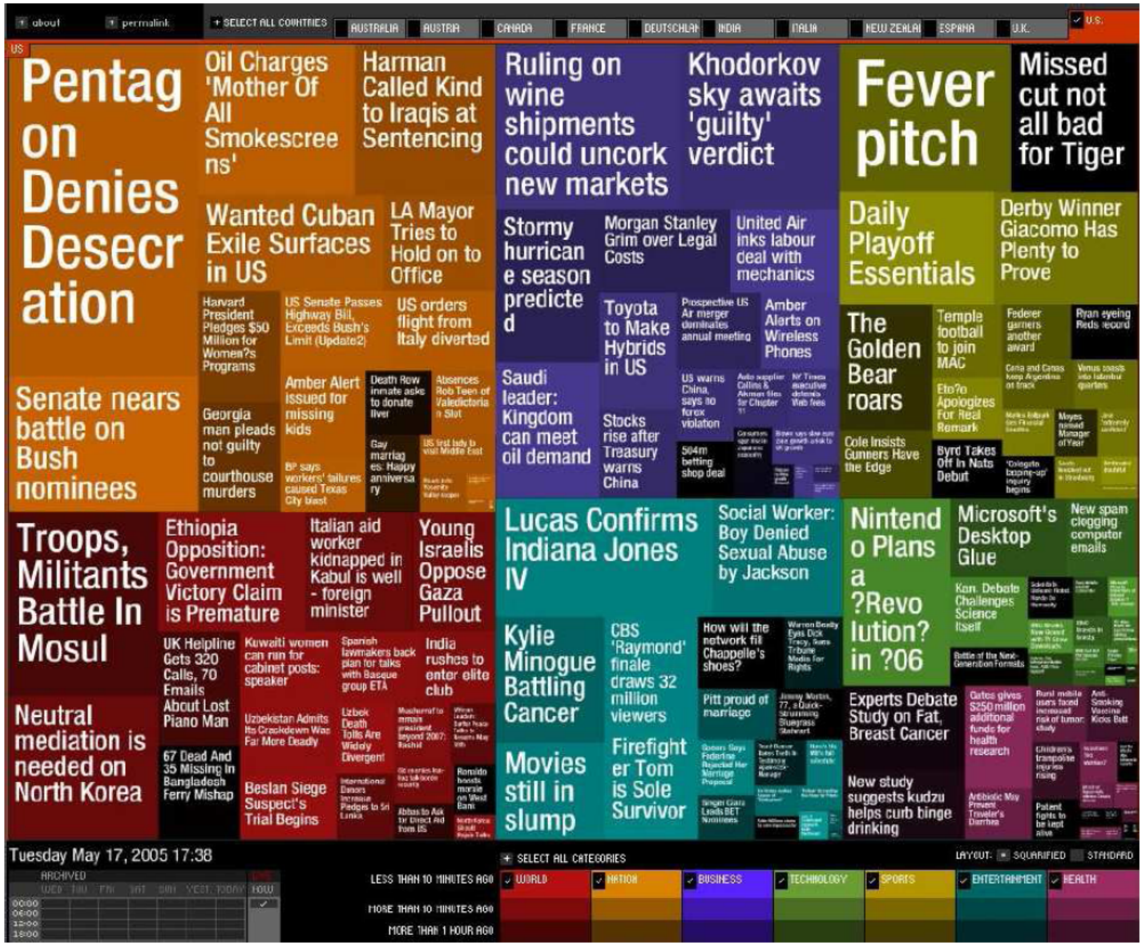
http://www.cs.umd.edu/hcil/treemap-history/all102001.jpg

Visualizing Complex Data and Relations

- Visualizing non-numerical data: text and social networks
- Tag cloud: visualizing user-generated tags

Google News output

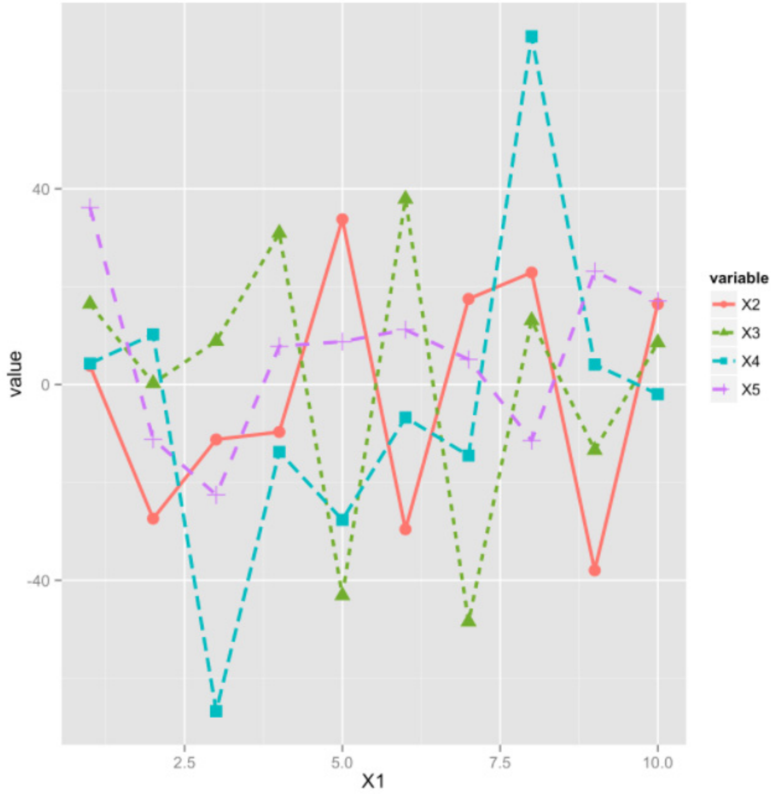
- The importance of tag is represented by font size/color
- Besides text data, there are also methods to visualize relationships, such as visualizing social networks



http://www.industrial-electronics.com/images/dmct_3e_2-20.jpg

ggplot2 Data Visualization Code

```
ggplot(data, aes(x=X1, y=value, color=variable)) +
  geom_line(aes(linetype=variable), size=1) +
  geom_point(aes(shape=variable, size=4))
```



When a data scientist draws a plot, she just needs to differ the lines (color, line type) and points (color, shape) by a certain categorical variable instead of specifying particular style to each line and point.

http://ggplot2.tidyverse.org/

Content

- Data Instances, Attributes and Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity

Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- Dissimilarity (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

- Data matrix
  - *n* data points with *p* dimensions
  - Two modes
    - Row: objects
    - Column: attributes

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
  - *n* data points, but registers only the distance
  - A triangular matrix
  - Single mode

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

- Similarity

$$\mathrm{sim}(i,j) = 1 - d(i,j)$$

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

x_1=[Weekday=Friday, Gender=Male, City=Shanghai]

x_2=[Weekday=Friday, Gender=Female, City=Shanghai]

$$d(1, 2) = \frac{3 - 2}{3} = \frac{1}{3}$$

One-Hot Encoding for Nominal Attributes

- One-hot encoding: creating a new binary attribute for each of the *p* nominal states

$$x_i = [\text{Weekday=Friday}, \text{Gender=Male}, \text{City=Shanghai}]$$

$$x_i = [0,0,0,0,1,0,0\ \ 0,1\ \ 0,0,1,0...0]$$

Whether Weekday=Friday Whether City=Shanghai

- As such, we transform the nominal data instances into binary vectors, which can be fed into various functions
  - High dimensional sparse binary feature vector
  - Usually higher than 1M dimensions, even 1B dimensions
  - Extremely sparse

Proximity Measure for Binary Attributes

- A contingency table for binary data

		Object j		
		1	0	sum
	1	q	r	$q+r$
Object i	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- Distance measure for symmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for asymmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient (similarity measure for asymmetric binary variables):

$$\text{sim}_{\text{Jaccard}}(i,j) = \frac{q}{q+r+s}$$

- Note: Jaccard coefficient is the same as "coherence":

$$\text{coherence}(i,j) = \frac{\text{sup}(i,j)}{\text{sup}(i)+\text{sup}(j)-\text{sup}(i,j)} = \frac{q}{(q+r)+(q+s)-q}$$

Dissimilarity between Binary Variables

- Example data

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

$$d(i,j) = \frac{r+s}{q+r+s}$$

$$d(\text{Jack}, \text{Mary}) = \frac{0+1}{2+0+1} = 0.33$$

$$d(\text{Jack}, \text{Jim}) = \frac{1+1}{1+1+1} = 0.67$$

$$d(\text{Jim}, \text{Mary}) = \frac{1+2}{1+1+2} = 0.75$$

Object j

		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Standardizing Numeric Data

- Numeric data examples

$$x_1=[1.2, 3.5, 1.1, 2.7, 123.9]$$
$$x_2=[2.0, 1.5, 1.3, 3.1, 145.1]$$

This dimension may dominate the proximity calculation

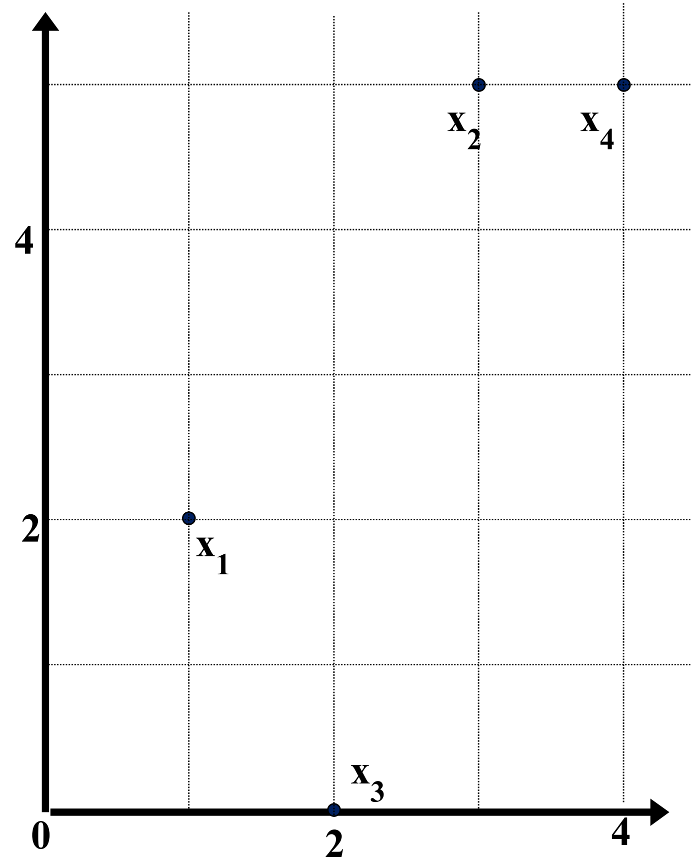
- Z-score: perform normalization for each dimension

$$z = \frac{x - \mu}{\sigma}$$

  - x: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - The distance between the raw score and the population mean in units of the standard deviation
  - Negative when the raw score is below the mean, positive when above

Example:

Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute 1	attribute 2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5

Dissimilarity Matrix

(with Euclidean Distance)

	x_1	x_2	x_3	x_4
x_1	0			
x_2	3.61	0		
x_3	2.24	5.1	0	
x_4	4.24	1	5.39	0

Distance on Numeric Data: Minkowski Distance

- Minkowski distance: A popular distance measure

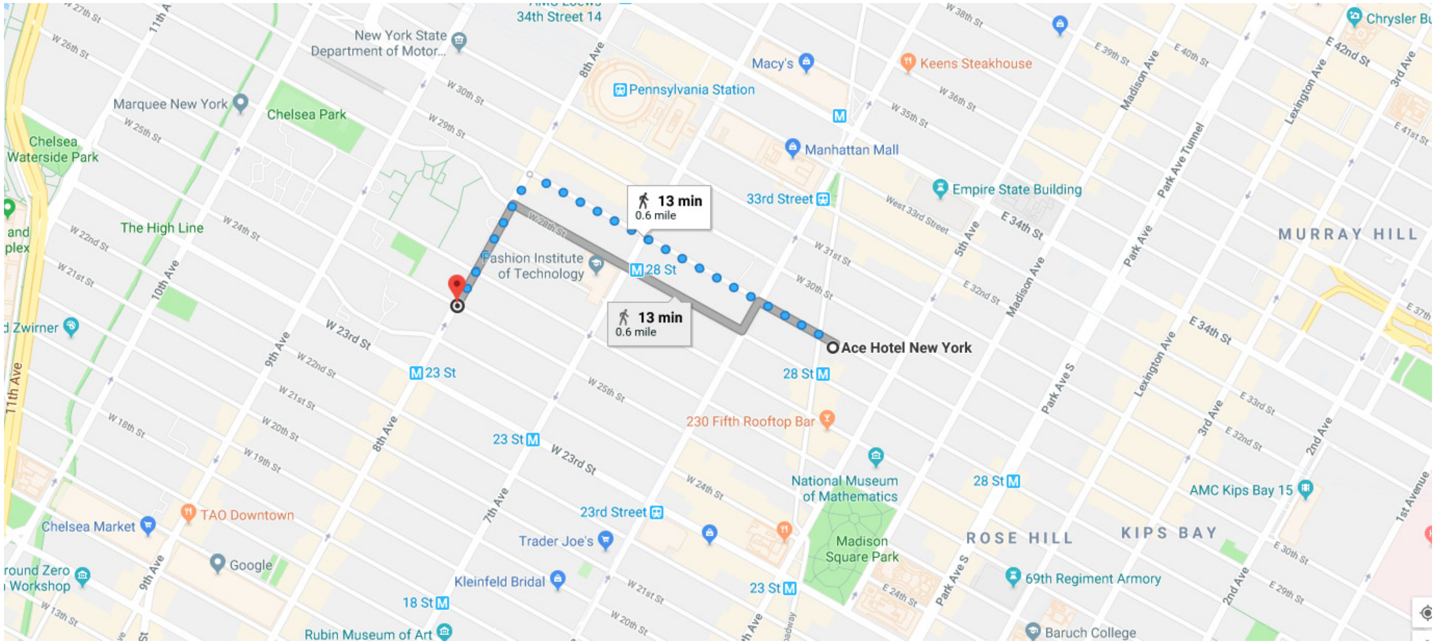
$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$
$$x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$$
$$d(i,j) = \left(|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h\right)^{\frac{1}{h}}$$

  - *h* is the order (the distance so defined is also called L-*h* norm)
- Properties
  - Positive definiteness: d(*i*, *j*) > 0 if *i* ≠ *j*, and d(*i*, *i*) = 0
  - Symmetry: d(*i*, *j*) = d(*j*, *i*)
  - Triangle Inequality: d(*i*, *j*) ≤ d(*i*, *k*) + d(*k*, *j*)
- A distance that satisfies these properties is a metric

Special Cases of Minkowski Distance

- h = 1: Manhattan (city block, L_1 norm) distance
    - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$



Special Cases of Minkowski Distance

- *h* = 2: Euclidean (L_2 norm) distance

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2}$$

- *h* -> ∞ : Supremum (L_{max} norm) distance
  - This is the maximum difference between any component (attribute) of the vectors

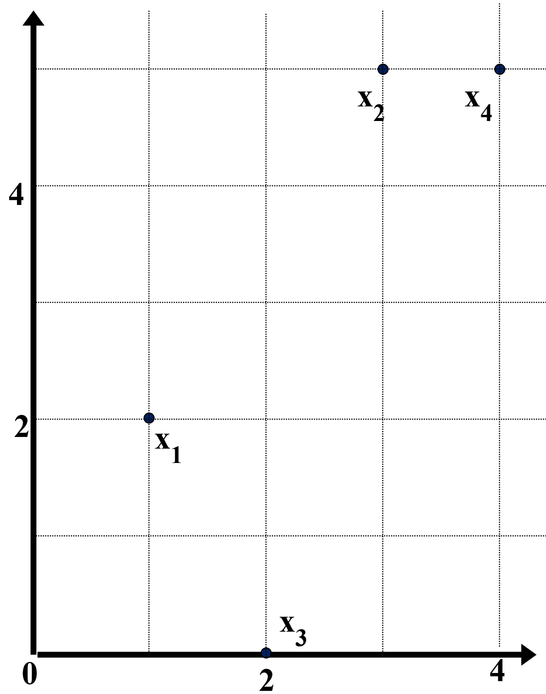
$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

Example:

Minkowski Distances

Data Matrix

point	attribute 1	attribute 2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5



Dissimilarity Matrices

Mahantan (L_1)

	x_1	x_2	x_3	x_4
x_1	0			
x_2	5	0		
x_3	3	6	0	
x_4	6	1	7	0

Euclidean (L_2)

	x_1	x_2	x_3	x_4
x_1	0			
x_2	3.61	0		
x_3	2.24	5.1	0	
x_4	4.24	1	5.39	0

Supremum (L_{max})

	x_1	x_2	x_3	x_4
x_1	0			
x_2	3	0		
x_3	2	5	0	
x_4	3	1	5	0

Cosine Similarity

- A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document.

Document	Team	Coach	Hockey	Baseball	Soccer	Penalty	Score	Win	Loss	Season
d1	5	0	3	0	2	0	0	2	0	0
d2	3	0	2	0	1	1	0	1	0	1
d3	0	7	0	2	1	0	0	3	0	0
d4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: gene features in micro-arrays, …
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \cdot d_2)/(\|d_1\| \cdot \|d_2\|)$$

where $\cdot$ indicates vector dot product, $\|d\|$ is the length of vector d

Example: Cosine Similarity

Document	Team	Coach	Hockey	Baseball	Soccer	Penalty	Score	Win	Loss	Season
d1	5	0	3	0	2	0	0	2	0	0
d2	3	0	2	0	1	1	0	1	0	1
d3	0	7	0	2	1	0	0	3	0	0
d4	0	1	0	0	1	2	2	0	3	0

$$\cos(d_1, d_2) = (d_1 \cdot d_2)/(\|d_1\| \cdot \|d_2\|)$$

- Ex: Find the similarity between documents 1 and 2.

  d1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)

  d2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$$d_1 \cdot d_2 = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 1 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$$

$$\|d_1\| = (5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0)^{0.5} = 42^{0.5} = 6.48$$

$$\|d_2\| = (3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1)^{0.5} = 17^{0.5} = 4.12$$

$$\cos(d_1, d_2) = 0.94$$

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables
  - Note: this is just a trivial solution

Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
  - Different fields may bring different level of importance
- One may use a weighted formula to combine their effects

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

  - f is binary or nominal
    - $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
  - f is numeric: use the normalized distance
  - f is ordinal
    - Compute ranks r_{if} and
    - Treat z_{if} as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
  - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.