Introduction to Data Mining

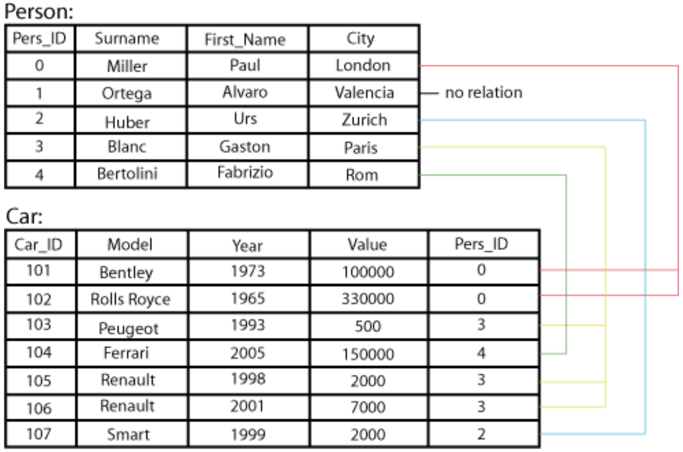
Data and data preprocessing

Agenda

- Know your data
- Preprocess the data
- Some public datasets
- Summary

Data Types

- Relational records
  - Relational tables, highly structured
- Data matrix, e.g., numerical matrix, crosstabs



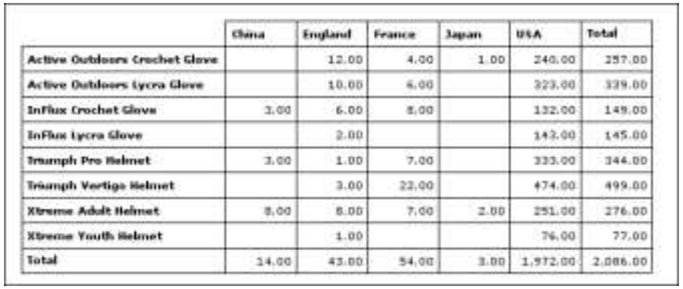
Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

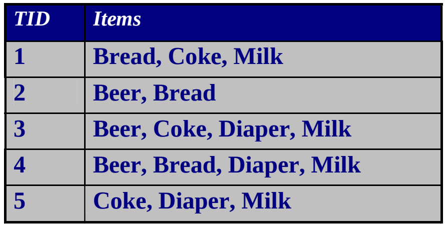
Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

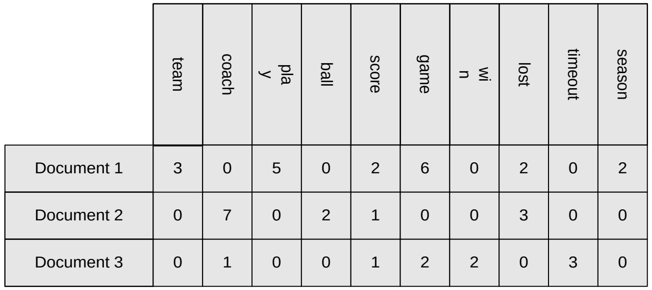


	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

- Transaction data



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

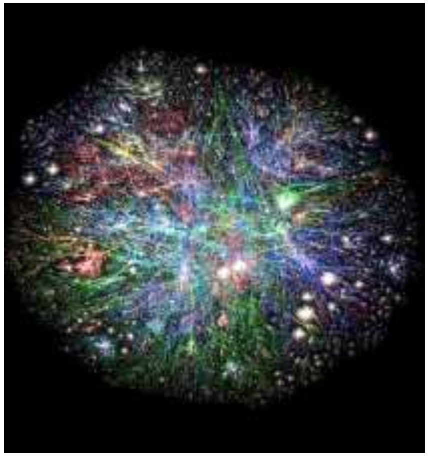


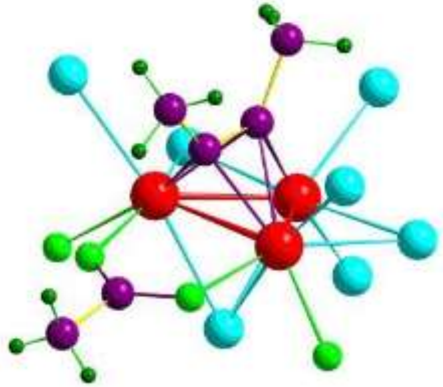
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- Document data: Term-frequency vector (matrix) of text documents

Data Types

- Transportation network
- World Wide Web

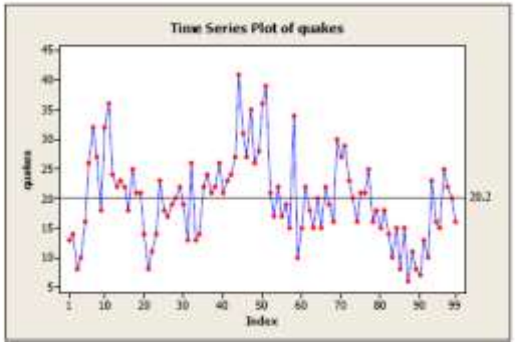




- ❑ Molecular Structures
- ❑ Social or information networks

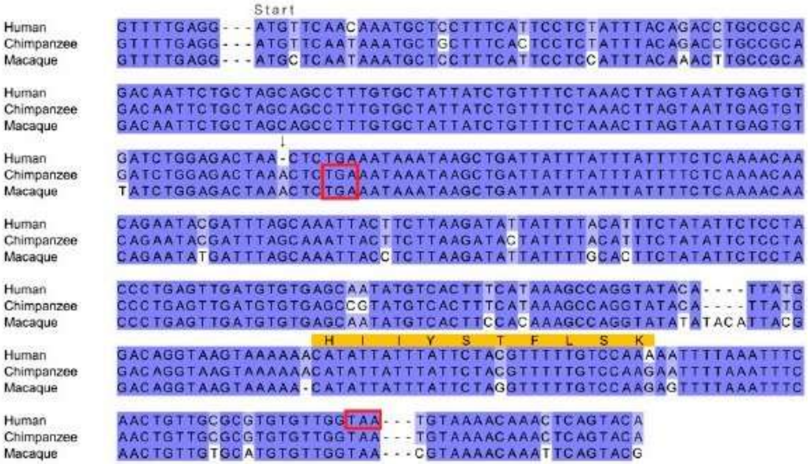
Data Types

- Video data: sequence of images
- Temporal data: time-series



- Sequential Data: transaction sequences

- Genetic sequence data


Start
Human GTTTTGAGG---ATGTTCAACAAATGCTCCTTTCATTCCTCTATTTACAGACCTGCCGCA
Chimpanzee GTTTTGAGG---ATGTTCAATAAATGCTGCTTTCACTCCTCTATTTACAGACCTGCCGCA
Macaque GTTTTGAGG---ATGCTCAATAAATGCTCCTTTCATTCCTCCATTTACAAACTTGCCGCA
Human GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Chimpanzee GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Macaque GACAATTCTGCTAGCAGCCTTTGTGCTATTATCTGTTTTCTAAACTTAGTAATTGAGTGT
Human GATCTGGAGACTAA-CTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Chimpanzee GATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Macaque TATCTGGAGACTAAACTCTGAAATAAATAAGCTGATTATTTATTTATTTTCTCAAAACAA
Human CAGAATACGATTTAGCAAATTACTTCTTAAGATATTATTTTACATTTCTATATTCTCCTA
Chimpanzee CAGAATACGATTTAGCAAATTACTTCTTAAGATACTATTTTACATTTCTATATTCTCCTA
Macaque CAGAATATGATTTAGCAAATTACCTCTTAAGATATTATTTTGCACTTCTATATTCTCCTA
Human CCCTGAGTTGATGTGTGAGCAATATGTCACTTTCATAAAGCCAGGTATACA----TTATG
Chimpanzee CCCTGAGTTGATGTGTGAGCCGTATGTCACTTTCATAAAGCCAGGTATACA----TTATG
Macaque CCCTGAGTTGATGTGTGAGCAATATGTCACTTCCACAAAGCCAGGTATATATACATTACG
H I I Y S T F L S K
Human GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCAAAAATTTTAAATTTC
Chimpanzee GACAGGTAAGTAAAAAACATATTATTTATTCTACGTTTTTGTCCAAGAATTTTAAATTTC
Macaque GACAGGTAAGTAAAAA-CATATTATTTATTCTAGGTTTTTGTCCAAGAGTTTTAAATTTC
Human AACTGTTGCGCGTGTGTTGGTAA---TGTAAAACAAACTCAGTACA
Chimpanzee AACTGTTGCGCGTGTGTTGGTAA---TGTAAAACAAACTCAGTACA
Macaque AACTGTTGTGCATGTGTTGGTAA---CGTAAAACAAATTCAGTACG

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

Numeric Attribute Types

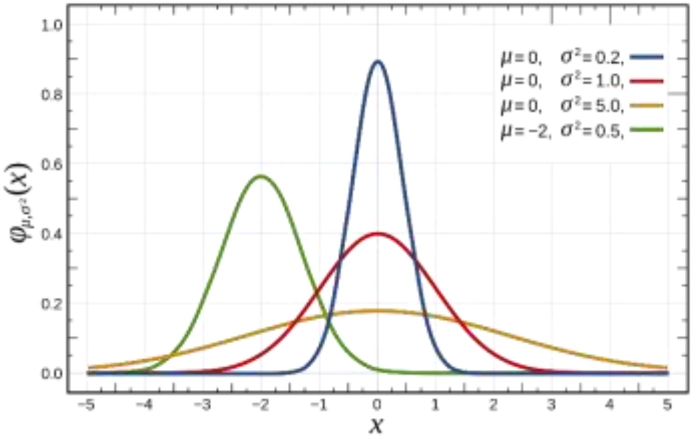
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ 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

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, ...
- Numerical dimensions correspond to sorted intervals
  - Data dispersion:
  - Boxplot or quantile analysis on sorted intervals



Mean

- Mean (algebraic measure) (sample vs. population):
  Note: *n* is sample size and *N* is population size.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

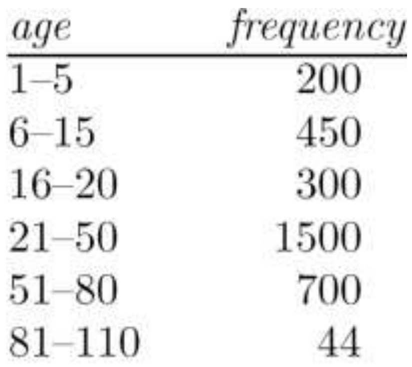
- Weighted arithmetic mean:

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

- Trimmed mean:
  - Chopping extreme values

Median

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for *grouped data*):



age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Sum before the median interval

Approximate median

Interval width ($L_2 - L_1$)

$$median = L_1 + \left(\frac{n/2 - (\sum freq)_l}{freq_{median}}\right) width$$

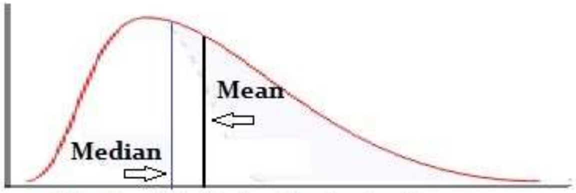
Low interval limit

Mode

- Mode: Value that occurs most frequently in the data

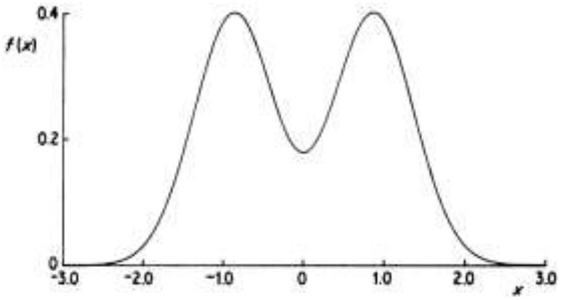
- Unimodal
  - Empirical formula:

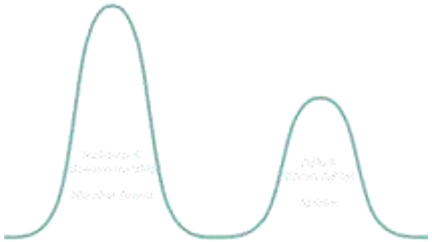
$$mean - mode = 3 \times (mean - median)$$

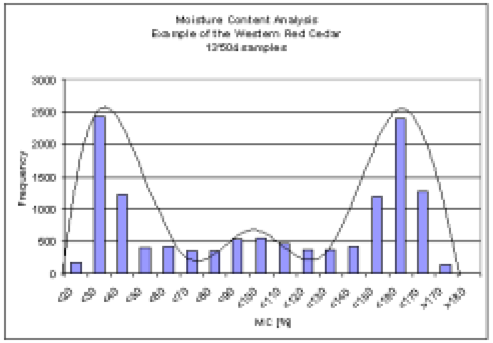


Right skewed distribution: Mean is to the right

- Multi-modal
  - Bimodal
  - Trimodal

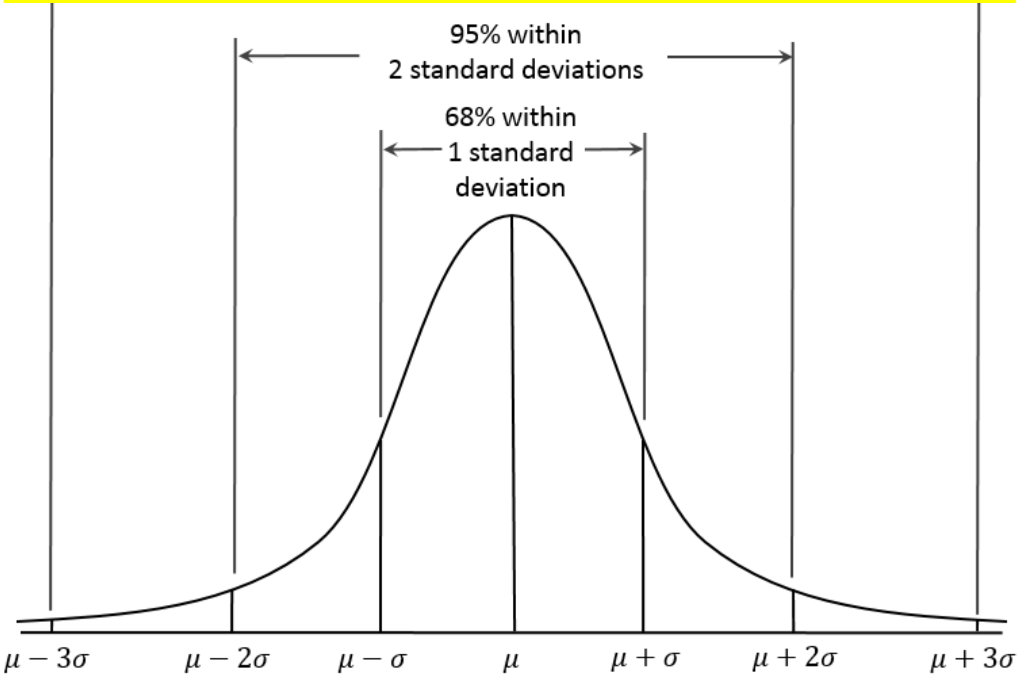






Properties of Normal Distribution Curve

← — ————Represent data dispersion, spread — —————→



Represent central tendency

Variance and Standard Deviation

- Variance and standard deviation (*sample: s, population: σ*)
  - **Variance**: (algebraic, scalable computation)

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

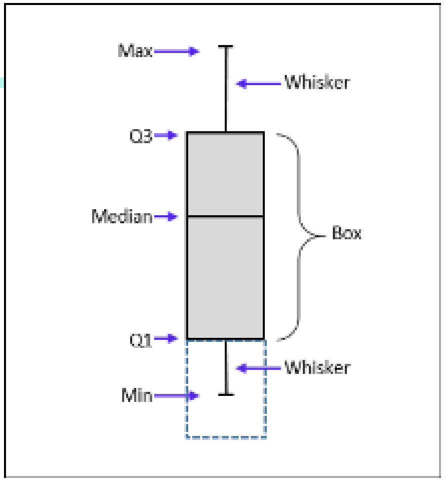
  - **Standard deviation** *s (or σ)* is the square root of variance *s*^{2} *(or σ*$^{2)}$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} - \mu^2$$

Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. 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

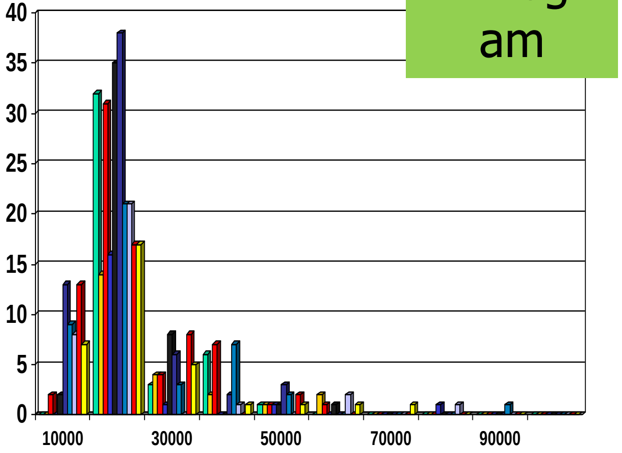
Quartiles & Boxplots



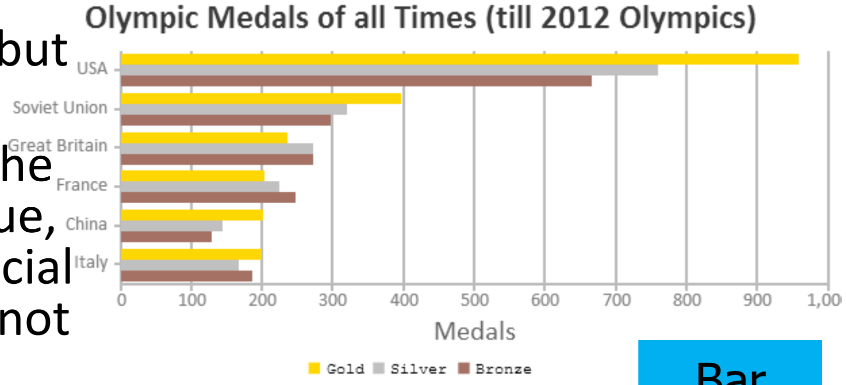
- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range**: IQR = $Q_3 - Q_1$
- **Five number summary**: min, Q_1, median, Q_3, max
- **Boxplot**: Data is represented with a box
  - Q_1, Q_3, IQR: The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - Median (Q_2) 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
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR

Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- Differences between histograms and bar charts
  - Histograms are used to show distributions of variables while bar charts are used to compare variables
  - Histograms plot binned quantitative data while bar charts plot categorical data
  - Bars can be reordered in bar charts but not in histograms
  - Differs from a bar chart in that it is the area of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width

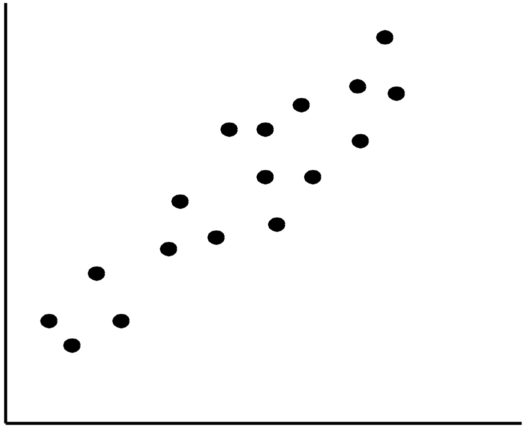


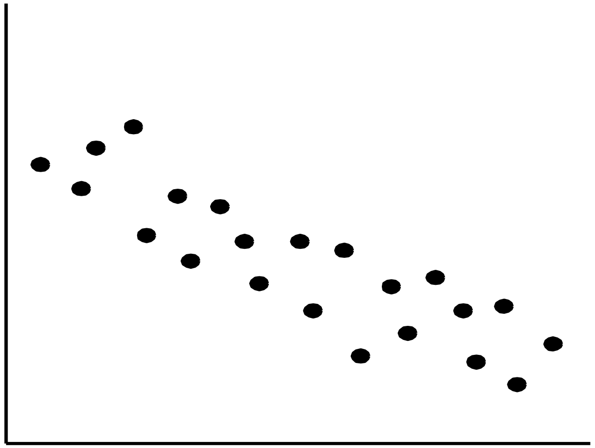
Histogram

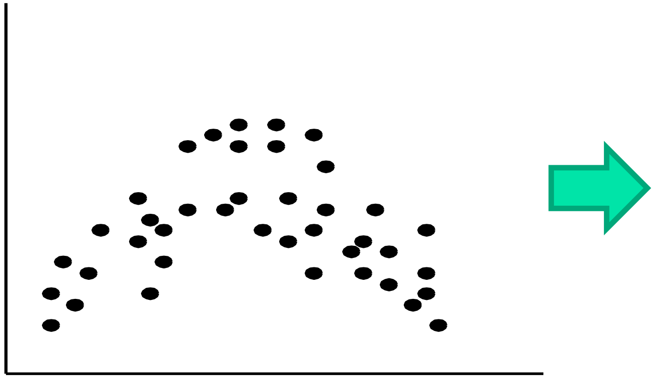


Bar chart

Positively and Negatively Correlated Data







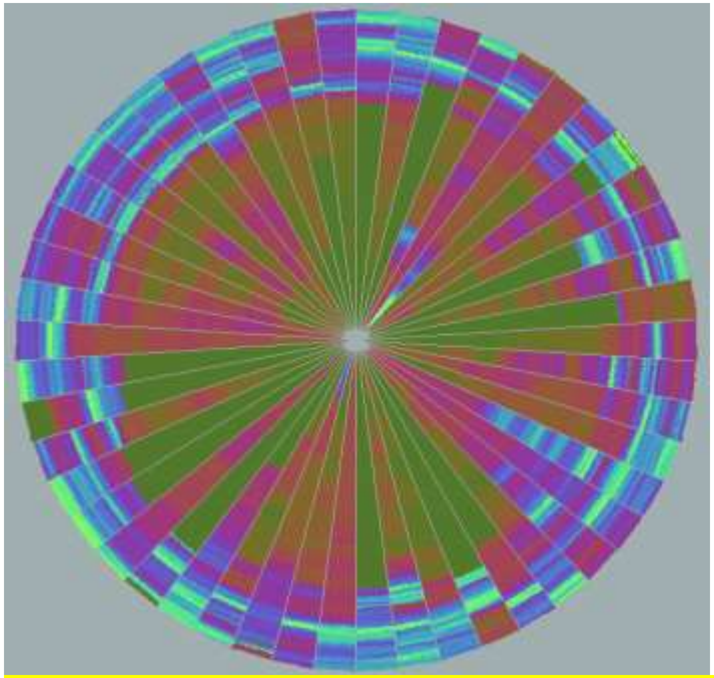
- The left half fragment is positively correlated
- The right half is negative correlated

Data Visualization

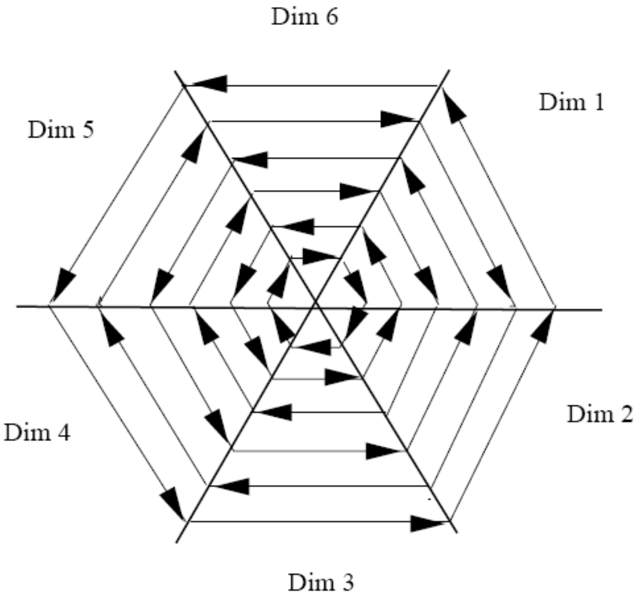
- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters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

Laying Out Pixels in Circle Segments

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segment



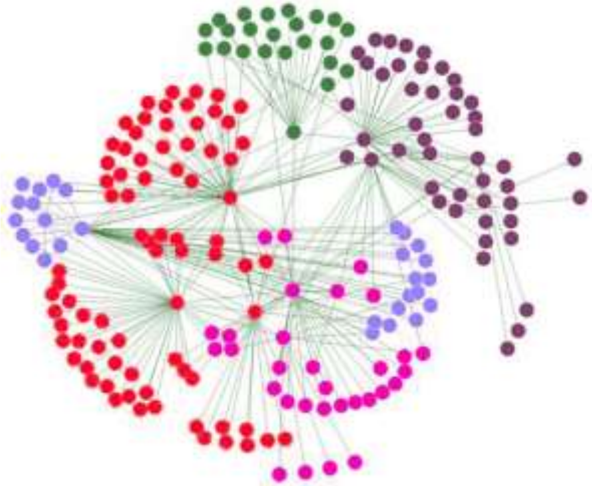
Representing about 265,000 50-dimensional Data Items with the 'Circle Segments' Technique



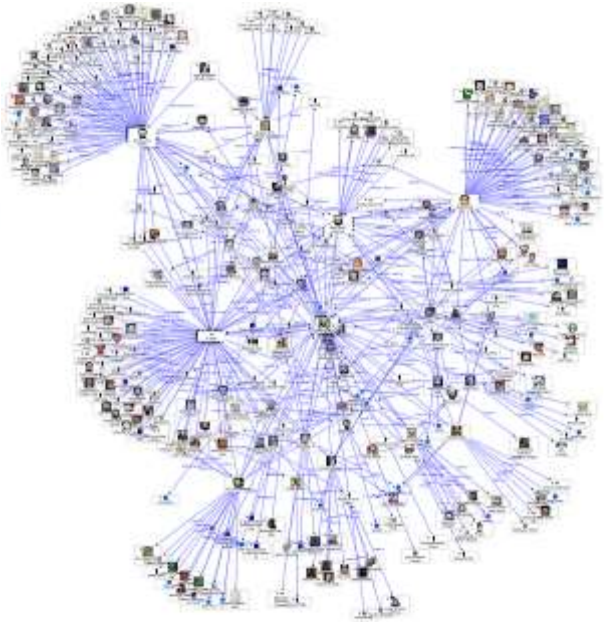
(b) Laying out pixels in circle segment

Social Networks

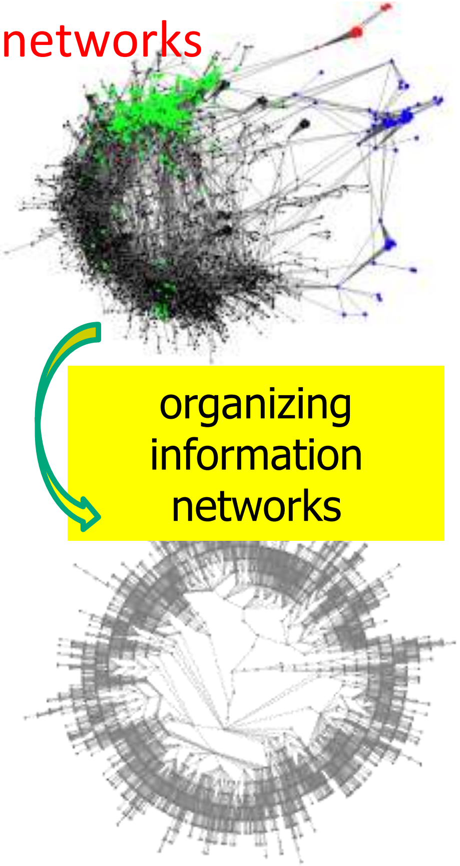
- Visualizing non-numerical data: social and information networks



A typical network structure



A social network



organizing information networks

Similarity, Dissimilarity, and Proximity

- **Similarity measure** or **similarity function**
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike: The higher value, the more alike
  - Often falls in the range [0,1]: 0: no similarity; 1: completely similar
- **Dissimilarity** (or **distance**) measure
  - Numerical measure of how different two data objects are
  - In some sense, the inverse of similarity: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar)
  - Range [0, 1] or [0, ∞) , depending on the definition
- **Proximity** usually refers to either similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

- Data matrix
  - A data matrix of n data points with *l* dimensions
- Dissimilarity (distance) matrix
  - n data points, but registers only the distance *d*(*i, j*) (typically metric)
  - Usually symmetric, thus a triangular matrix
  - Distance functions are usually different for real, boolean, categorical, ordinal, ratio, and vector variables
  - Weights can be associated with different variables based on applications and data semantics

$$D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}$$

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$$

Standardizing Numeric Data

- Z-score: $z = \frac{x - \mu}{\sigma}$
  - 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, "+" when above
- An alternative way: Calculate the mean absolute deviation

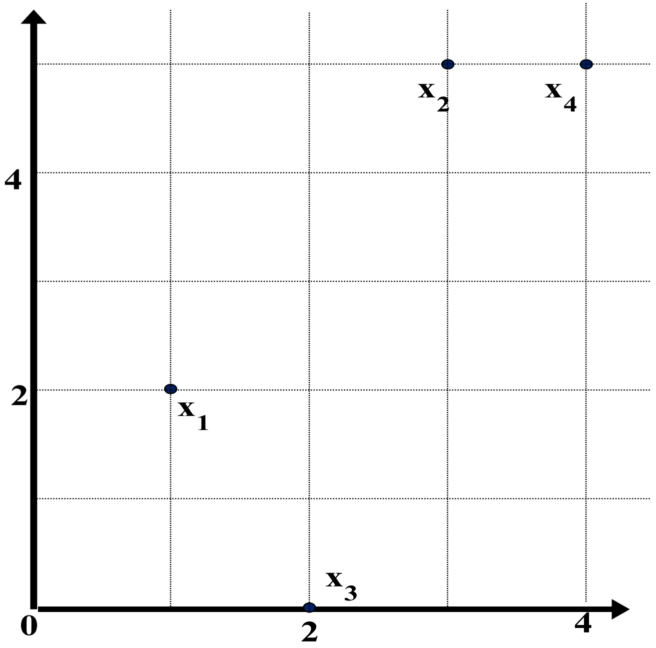
$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf}).$$

  - standardized measure (*z-score*): $z_{if} = \frac{x_{if} - m_f}{s_f}$
- Using mean absolute deviation is more robust than using standard deviation

Example: Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix (by Euclidean Distance)

	x1	***x2***	***x3***	***x4***
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Distance on Numeric Data: Minkowski Distance

- Minkowski distance: A popular distance measure

$$d(i, j) = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jl})$ are two *l*-dimensional data objects, and *p* is the order (the distance so defined is also called L-*p* norm)

- Properties
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positivity)
  - d(i, j) = d(j, i) (Symmetry)
  - d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)
- A distance that satisfies these properties is a metric
- Note: There are nonmetric dissimilarities, e.g., set differences

Special Cases of Minkowski Distance

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

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \cdots + |x_{i_l} - x_{j_l}|$$

- p = 2: (L_2 norm) Euclidean distance

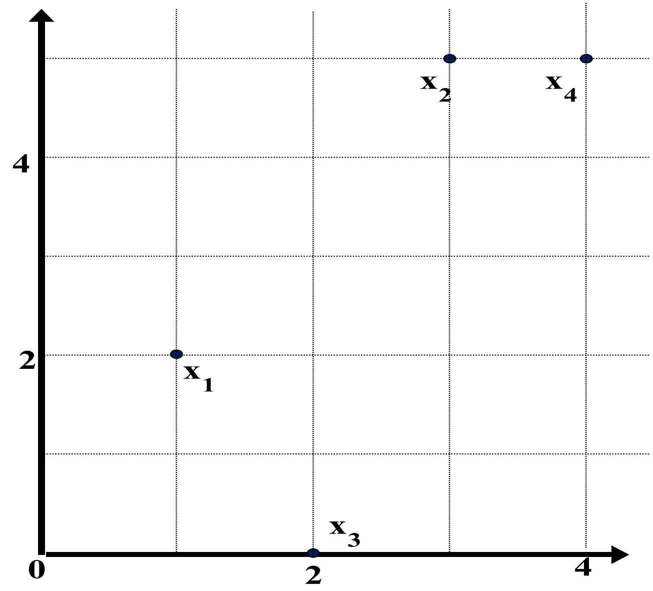
$$d(i,j) = \sqrt{|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \cdots + |x_{i_l} - x_{j_l}|^2}$$

- $p \rightarrow \infty$: (L_{max} norm, L_∞ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{p\to\infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^{l} |x_{if} - x_{jf}|$$

Minkowski Distance at Special Cases

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

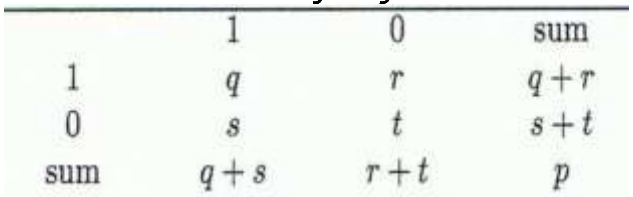
L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum (L_∞)

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Proximity Measure for Binary Attributes

- A contingency table for binary data



		Object *j*		
		1	0	sum
Object *i*	1	q	r	$q+r$
	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): $sim_{Jaccard}(i,\,j) = \frac{q}{q+r+s}$
- Note: Jaccard coefficient is the same as

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

Dissimilarity between Asymmetric Binary Variables

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 (not counted in)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0
- Distance: $d(i, j) = \frac{r + s}{q + r + s}$

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Mary

Jack	1	0	Σ_{row}
1	2	0	2
0	1	3	4
Σ_{col}	3	3	6

Jim

Jack	1	0	Σ_{row}
1	1	1	2
0	1	3	4
Σ_{col}	2	4	6

Mary

Jim	1	0	Σ_{row}
1	1	1	2
0	2	2	4
Σ_{col}	3	3	6

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes
  - Example: Color (red, yellow, blue, green), profession, etc.
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

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

- Method 2: Use a large number of binary attributes
  - Creating a new binary attribute for each of the *M* nominal states

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- Can be treated like interval-scaled
  - Replace *an ordinal variable value* by its 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

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

    - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
      - Then distance: d(freshman, senior) = 1, d(junior, senior) = 1/3
  - Compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

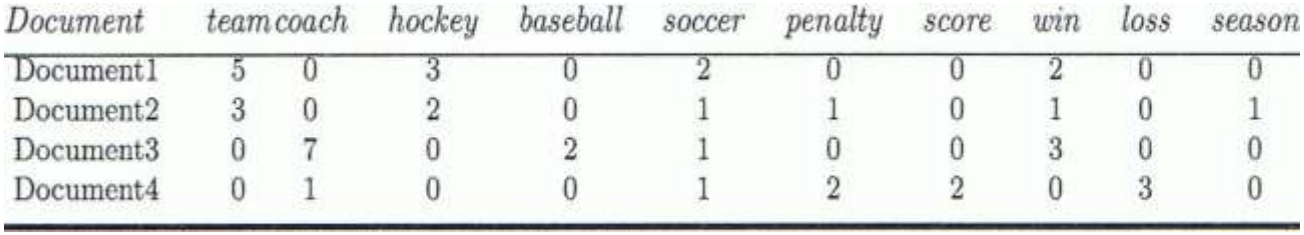
- A dataset may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- One may use a weighted formula to combine their effects:

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

  - If f is numeric: Use the normalized distance
  - If f is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; or $d_{ij}^{(f)} = 1$ otherwise
  - If f is ordinal
    - Compute ranks z_{if} (where $z_{if} = \frac{r_{if} - 1}{M_f - 1}$)
    - Treat z_{if} as interval-scaled

Cosine Similarity of Two Vectors

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document



Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	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, etc.
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

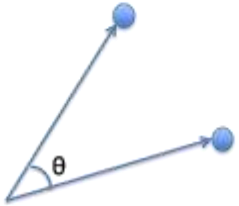
$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

where • indicates vector dot product, $||d||$: the length of vector *d*

Calculating Cosine Similarity

- Calculating Cosine Similarity:

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|} \qquad sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



where • indicates vector dot product, ||*d*||: the length of vector *d*

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

  d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

  - First, calculate vector dot product

    $d_1 \bullet d_2$ = 5 X 3 + 0 X 0 + 3 X 2 + 0 X 0 + 2 X 1 + 0 X 1 + 0 X 1 + 2 X 1 + 0 X 0 + 0 X 1 = 25

  - Then, calculate $||d_1||$ and $||d_2||$

$$\| d_1 \| = \sqrt{5\times5+0\times0+3\times3+0\times0+2\times2+0\times0+0\times0+2\times2+0\times0+0\times0} = 6.481$$

$$\| d_2 \| = \sqrt{3\times3+0\times0+2\times2+0\times0+1\times1+1\times1+0\times0+1\times1+0\times0+1\times1} = 4.12$$

  - Calculate cosine similarity: cos(d_1, d_2) = 25/ (6.481 X 4.12) = 0.94

Agenda

- Know your data
- Preprocess the data
- Some public datasets
- Summary

Major Tasks

- **Data cleaning**
  - Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Data Quality Issues

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error
    - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
        - e.g., *Occupation* = “ ” (missing data)
    - Noisy: containing noise, errors, or outliers
        - e.g., *Salary* = “−10” (an error)
    - Inconsistent: containing discrepancies in codes or names, e.g.,
        - *Age* = “42”, *Birthday* = “03/07/2010”
        - Was rating “1, 2, 3”, now rating “A, B, C”
        - discrepancy between duplicate records
    - Intentional (e.g., *disguised missing* data)
        - Jan. 1 as everyone’s birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data were not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Did not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - **the most probable value: inference-based such as Bayesian formula or decision tree**

Noisy Data

- **Noise:** random error or variance in a measured variable
- **Incorrect attribute values** may be due to
  - Faulty data collection instruments
  - Data entry problems
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention
- **Other data problems**
  - Duplicate records
  - Incomplete data
  - Inconsistent data

How to Handle Noisy Data?

- Binning
  - First sort data and partition into (equal-frequency) bins
  - Then one can **smooth by bin means, smooth by bin median, smooth by bin boundaries**, etc.
- Regression
  - Smooth by fitting the data into regression functions
- Clustering
  - Detect and remove outliers
- Semi-supervised: Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning as a Process

- **Data discrepancy detection**
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- **Data migration and integration**
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter’s Wheels)

Data Integration

- Data integration
  - Combining data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- **Entity identification:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- **Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis***
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis

- **X^2 (chi-square) test:**

$$\chi^2 = \sum_{i}^{n} \frac{(\overset{\text{observed}}{\overset{\downarrow}{O_i}} - E_i)^2}{\underset{\text{expected}}{E_i}}$$

- Null hypothesis: The two distributions are independent
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
  - The larger the X^2 value, the more likely the variables are related
- Note: Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

How to derive 90? 450/1500 * 300 = 90

We can reject the null hypothesis of independence at a confidence level of 0.001

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Variance for Single Variable (Numerical Data)

- The variance of a random variable *X* provides a measure of how much the value of *X* deviates from the mean or expected value of *X:*

$$\sigma^2 = \text{var}(X) = E[(X-\mu)^2] = \begin{cases} \sum_{x} (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

  - where σ^2 is the variance of X, σ is called *standard deviation*

  μ is the mean, and μ = E[X] is the expected value of X

  - That is, variance is the expected value of the square deviation from the mean
  - It can also be written as $\sigma^2 = \text{var}(X) = E[(X-\mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(x)]^2$

- Sample variance is the average squared deviation of the data value x_i from the sample mean $\hat{\mu}$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Covariance for Two Variables

- Covariance between two variables X_1 and X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1\mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

where $\mu_1 = E[X_1]$ is the respective mean or **expected value** of X_1; similarly for μ_2

- Sample covariance between X_1 and X_2: $\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \mu_1)(x_{i2} - \hat{\mu}_2)$
- Sample covariance is a generalization of the sample variance:

$$\hat{\sigma}_{11} = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \mu_1)(x_{i1} - \hat{\mu}_1) = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \mu_1)^2 = \hat{\sigma}_1^{\,2}$$

- **Positive covariance:** If $\sigma_{12} > 0$
- **Negative covariance**: If $\sigma_{12} < 0$
- **Independence**: If X_1 and X_2 are independent, $\sigma_{12} = 0$ but the reverse is not true
  - Some pairs of random variables may have a covariance 0 but are not independent
  - Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:
  - (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
- Covariance formula

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

- Its computation can be simplified as: $\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$
  - $E(X_1) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$
  - $E(X_2) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$
  - $\sigma_{12} = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14)/5 - 4 \times 9.6 = 4$
- Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

Correlation between Two Numerical Variables

- **Correlation** between two variables X_1 and X_2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1{}^2\sigma_2{}^2}}$$

- **Sample correlation** for two attributes X_1 and X_2: $$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\sigma_1\text{溶}\sigma_2} = \frac{\sum_{i=1}^{n}(x_{i1} - \mu\text{溶}_1)(x_{i2} - \mu_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \mu\text{溶}_1)^2 \sum_{i=1}^{n}(x_{i2} - \mu_2)^2}}$$

  where n is the number of tuples, μ_1 and μ_2 are the respective means of X_1 and X_2 , σ_1 and σ_2 are the respective standard deviation of X_1 and X_2

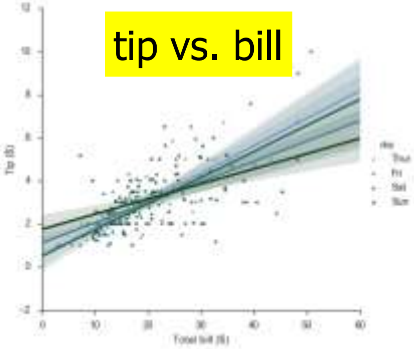
- If $\rho_{12} > 0$: A and B are positively correlated (X_1's values increase as X_2's)
  - The higher, the stronger correlation
- If $\rho_{12} = 0$: independent (under the same assumption as discussed in co-variance)
- If $\rho_{12} < 0$: negatively correlated

Data Reduction

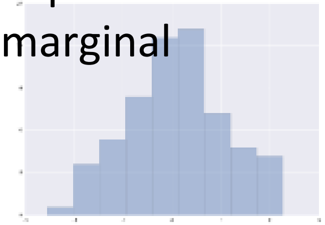
- **Data reduction**:
  - Obtain a reduced representation of the data set
    - much smaller in volume but yet produces *almost* the same analytical results
- Why data reduction?—A database/data warehouse may store terabytes of data
  - Complex analysis may take a very long time to run on the complete data set
- **Methods for data reduction** (also *data size reduction* or *numerosity reduction*)
  - Regression and Log-Linear Models
  - Histograms, clustering, sampling
  - Data cube aggregation
  - Data compression

Data Reduction: Parametric vs. Non-Parametric Methods

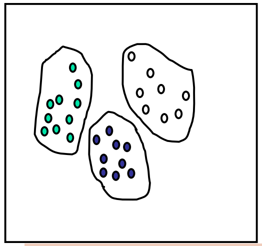
- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Parametric methods** (e.g., regression)
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Ex.: Log-linear models—obtain value at a point in *m*-D space as the product on appropriate marginal subspaces
- **Non-parametric** methods
  - Do not assume models
  - Major families: histograms, clustering, sampling, …



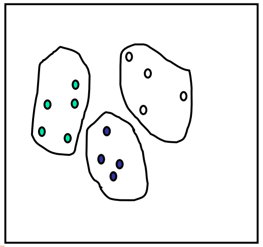
tip vs. bill



Histogra
m



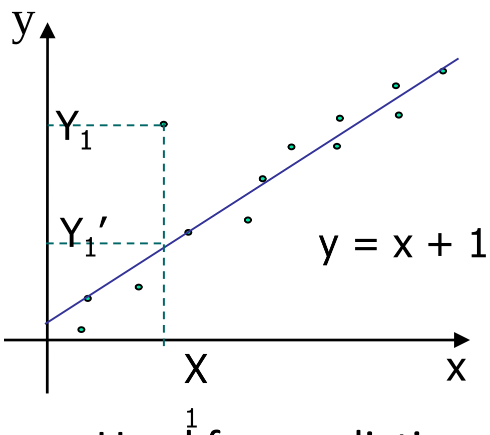
Clustering on the Raw Data



Stratified Sampling

Parametric Data Reduction: Regression Analysis

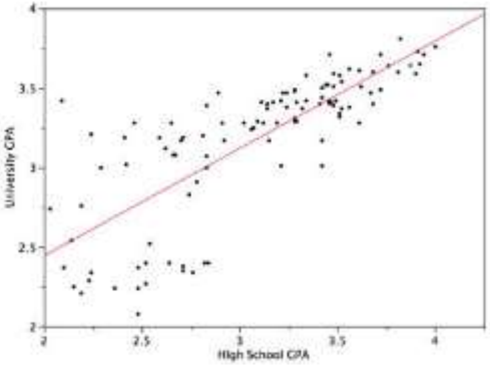
- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable*** or *measurement*) and of one or more *independent variables* (also known as ***explanatory variables*** or ***predictors***)
- The parameters are estimated so as to give a "**best fit**" of the data
- Most commonly the best fit is evaluated by using the ***least squares method***, but other criteria have also been used



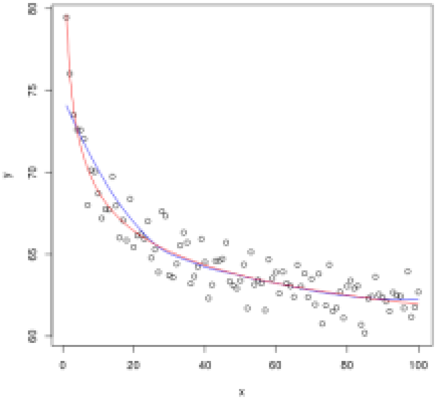
- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Linear and Multiple Regression

- Linear regression: *Y = w X + b*
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
  - Two regression coefficients, *w* and *b,* specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots$.

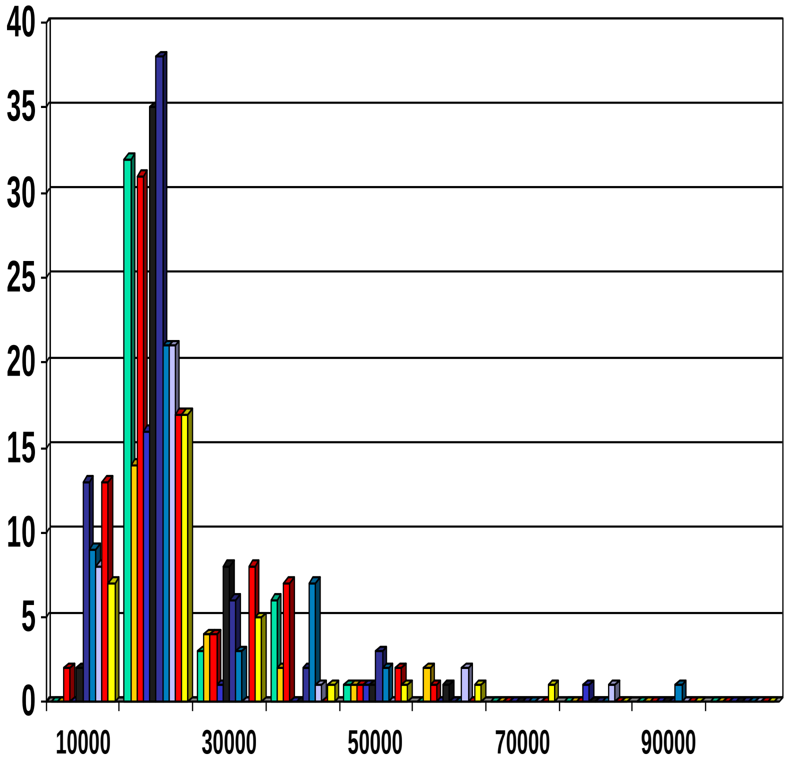


- Nonlinear regression:
  - Data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables
  - The data are fitted by a method of successive approximations



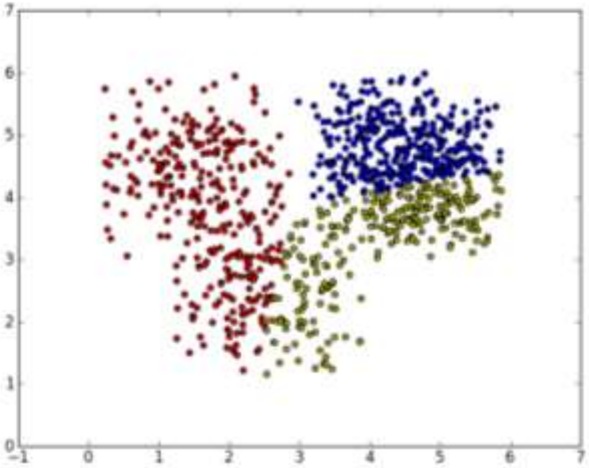
Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied in depth in Chapter 10

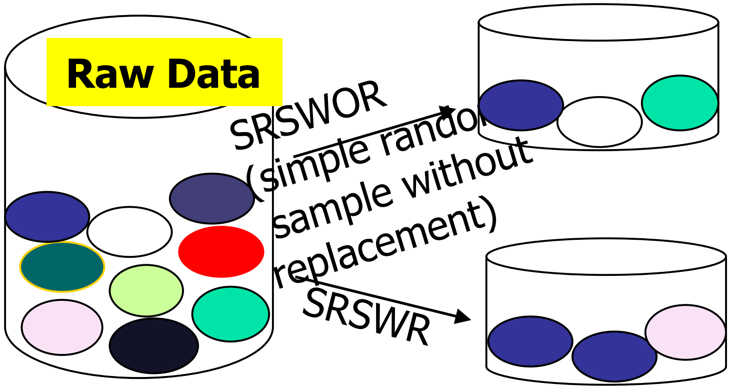


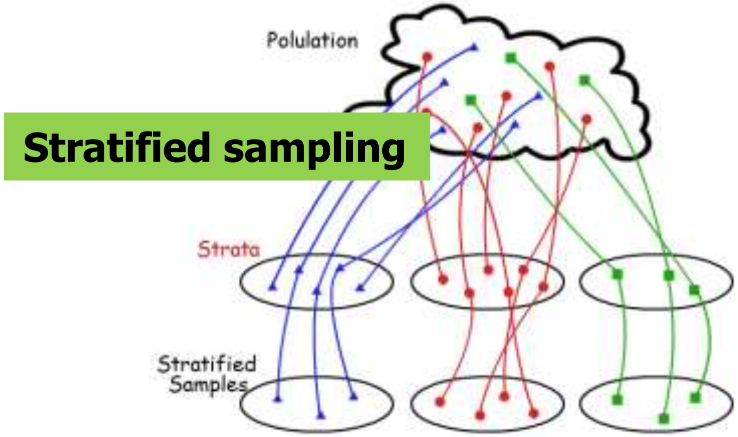
Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a **representative** subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling:
- Note: Sampling may not reduce database I/Os (page at a time)

Types of Sampling

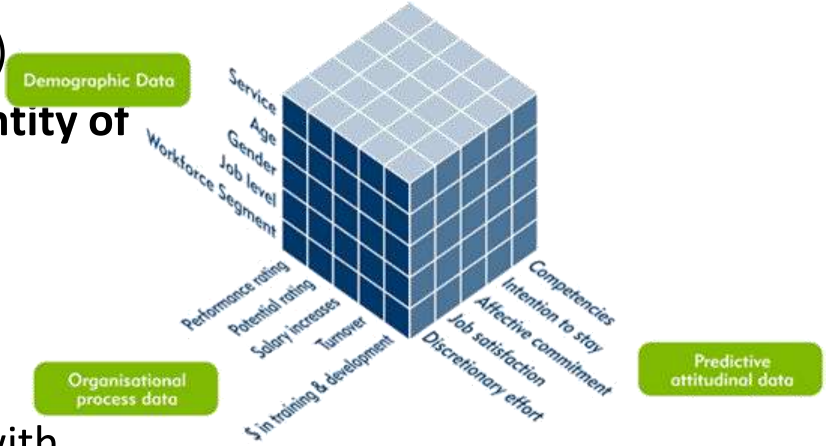
- **Simple random sampling:** equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling**
  - Partition (or cluster) the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)





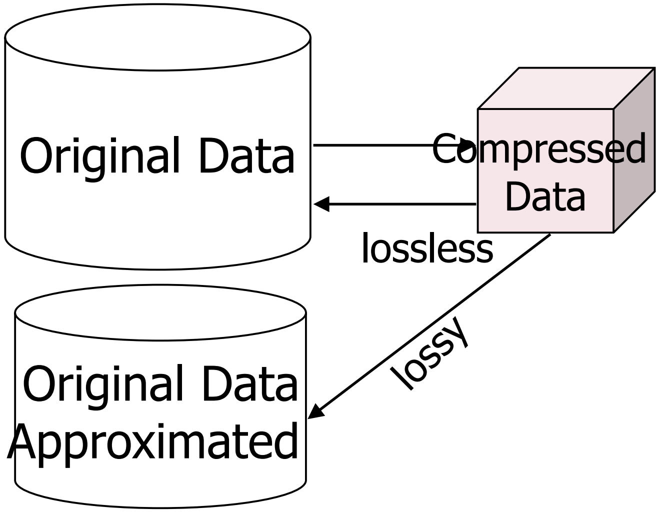
Data Cube Aggregation

- The lowest level of a data cube (base cuboid)
  - The aggregated data for an **individual entity of interest**
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible



Data Compression

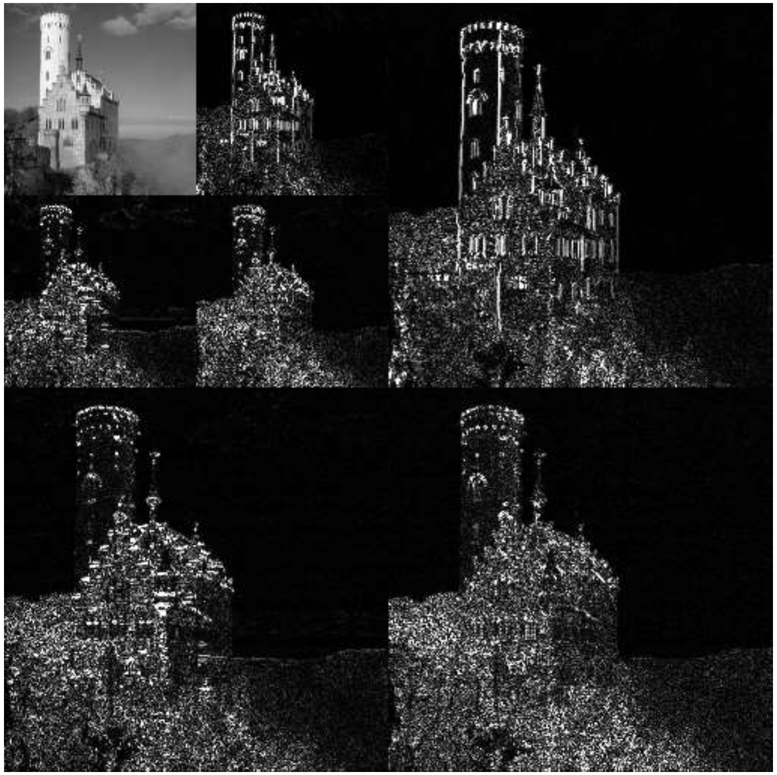
- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless, but only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and vary slowly with time
- Data reduction and dimensionality reduction may also be considered as forms of data compression



Lossy vs. lossless compression

Wavelet Transform: A Data Compression Technique

- Wavelet Transform
  - Decomposes a signal into different frequency subbands
  - Applicable to n-dimensional signals
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable
- Used for image compression



Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: Concept hierarchy climbing

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

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

  - Ex. Let income range $12,000 to $98,000 normalized to [0.0, 1.0]
    - Then $73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation

  - Ex. Let μ = 54,000, σ = 16,000. Then

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|ν'|) < 1

Discretization

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

Data Discretization Methods

- Binning
  - Top-down split, unsupervised
- Histogram analysis
  - Top-down split, unsupervised
- Clustering analysis
  - Unsupervised, top-down split or bottom-up merge
- Decision-tree analysis
  - Supervised, top-down split
- Correlation (e.g., χ^2) analysis
  - Unsupervised, bottom-up merge
- Note: All the methods can be applied recursively

Simple Discretization: Binning

- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Example: Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equi-depth**) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

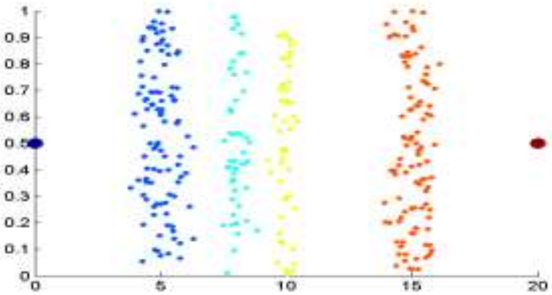
* Smoothing by **bin means**:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

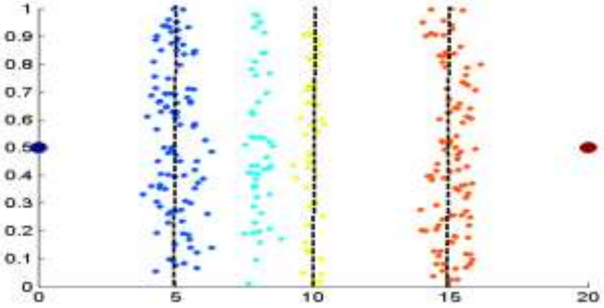
* Smoothing by **bin boundaries**:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

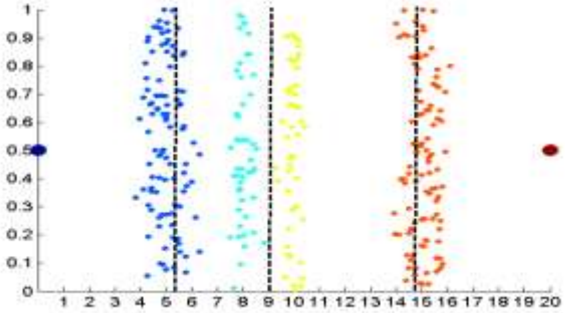
Discretization Without Supervision: Binning vs. Clustering



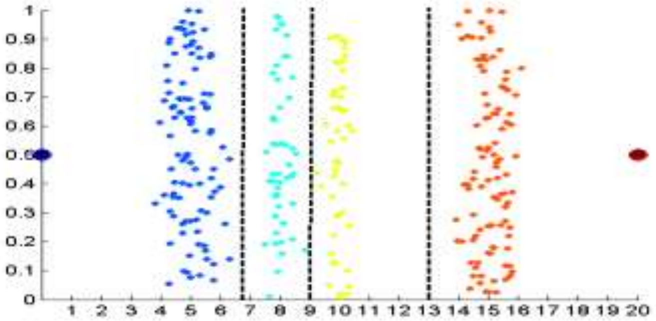
Data



Equal width (distance) binning



Equal depth (frequency) (binning)



K-means clustering leads to better results

Discretization by Classification & Correlation Analysis

- Classification (e.g., decision tree analysis)
  - Supervised: Given class labels, e.g., cancerous vs. benign
  - Using *entropy* to determine split point (discretization point)
  - Top-down, recursive split
  - Details to be covered in Chapter “Classification”
- Correlation analysis (e.g., Chi-merge: χ^2-based discretization)
  - Supervised: use class information
  - Bottom-up merge: Find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
  - Merge performed recursively, until a predefined stopping condition

Concept Hierarchy Generation

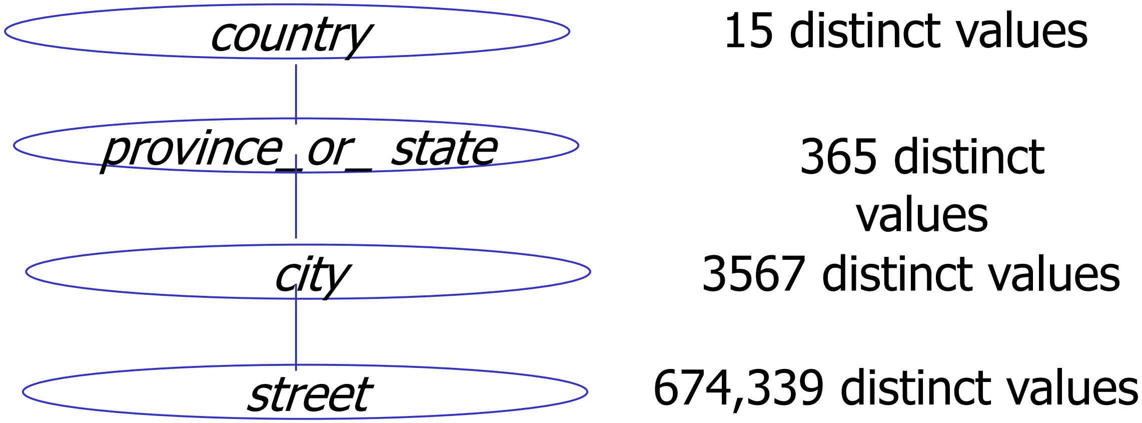
- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth, adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data—For numeric data, use discretization methods shown

Concept Hierarchy Generation for Nominal Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - *street* < *city* < *state* < *country*
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
  - E.g., only *street* < *city*, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {*street, city, state, country*}

Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



Dimensionality Reduction

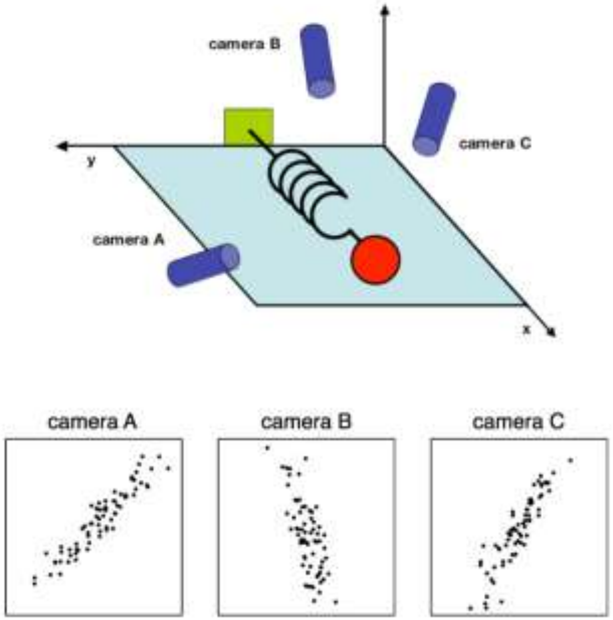
- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Reducing the number of random variables under consideration, via obtaining a set of principal variables
- **Advantages of dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

Dimensionality Reduction Techniques

- Dimensionality reduction methodologies
  - **Feature selection**: Find a subset of the original variables (or features, attributes)
  - **Feature extraction**: Transform the data in the high-dimensional space to a space of fewer dimensions
- Some typical dimensionality methods
  - Principal Component Analysis
  - Supervised and nonlinear techniques
    - Feature subset selection
    - Feature creation

Principal Component Analysis (PCA)

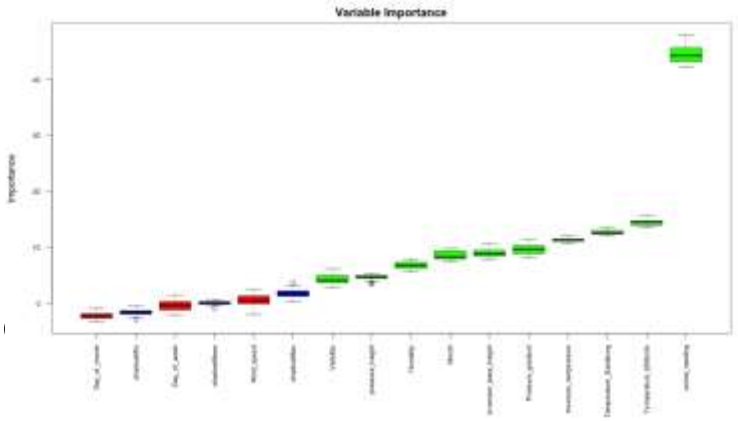
- PCA: A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called ***principal components***
- The original data are projected onto a much smaller space, resulting in dimensionality reduction
- Method: Find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



Ball travels in a straight line. Data from three cameras contain much redundancy

Attribute Subset Selection

- Another way to reduce dimensionality of data
- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
    - E.g., purchase price of a product and th amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
    - Ex. A student’s ID is often irrelevant to the task of predicting his/her GPA



Heuristic Search in Attribute Selection

- There are *2^d* possible attribute combinations of *d* attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - Best combined attribute selection and elimination
  - Optimal branch and bound:
    - Use attribute elimination and backtracking

Attribute Creation (Feature Generation)

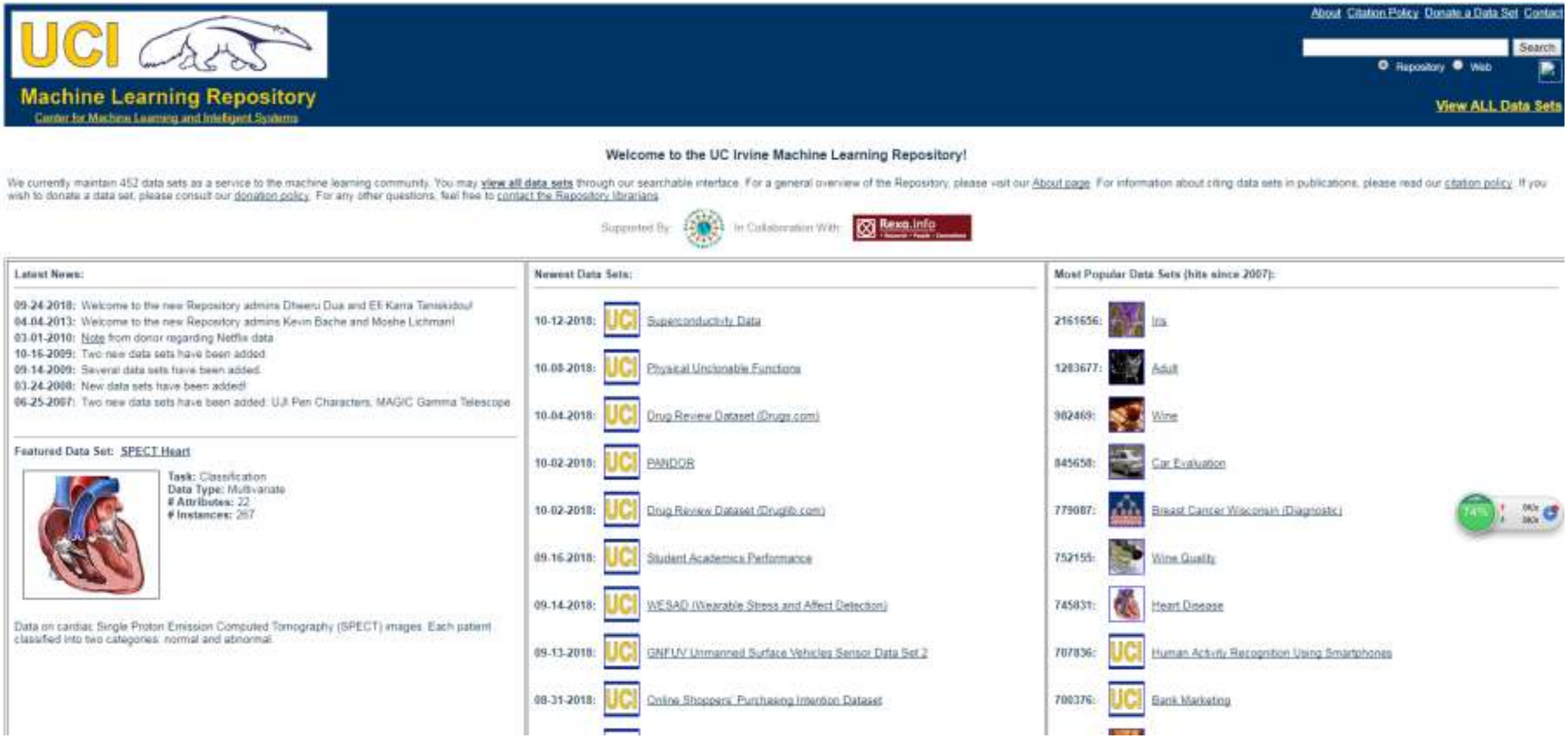
- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - Domain-specific
  - Mapping data to new space (see: data reduction)
    - E.g., Fourier transformation, wavelet transformation, manifold approaches (not covered)
  - Attribute construction
    - Combining features (see: discriminative frequent patterns in Chapter on "Advanced Classification")
    - Data discretization

Agenda

- Know your data
- Preprocess the data
- Some public datasets
- Summary

Public datasets

- UCI machine learning repository
  - http://archive.ics.uci.edu/ml/index.php


UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
About Citation Policy Donate a Data Set Contact
Search
○ Repository ● Web
View ALL Data Sets
Supported By:
In Collaboration With:
Rexa.info
74%

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 452 data sets as a service to the machine learning community. You may **view all data sets** through our searchable interface. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians.

Latest News:

09-24-2018: Welcome to the new Repository admins Dheeru Dua and Efi Karra Taniskidou!
04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
03-01-2010: Note from donor regarding Netflix data
10-16-2009: Two new data sets have been added.
09-14-2009: Several data sets have been added.
03-24-2008: New data sets have been added!
06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Featured Data Set: SPECT Heart

Task: Classification
Data Type: Multivariate
# Attributes: 22
# Instances: 267

Data on cardiac Single Proton Emission Computed Tomography (SPECT) images. Each patient classified into two categories: normal and abnormal.

Newest Data Sets:

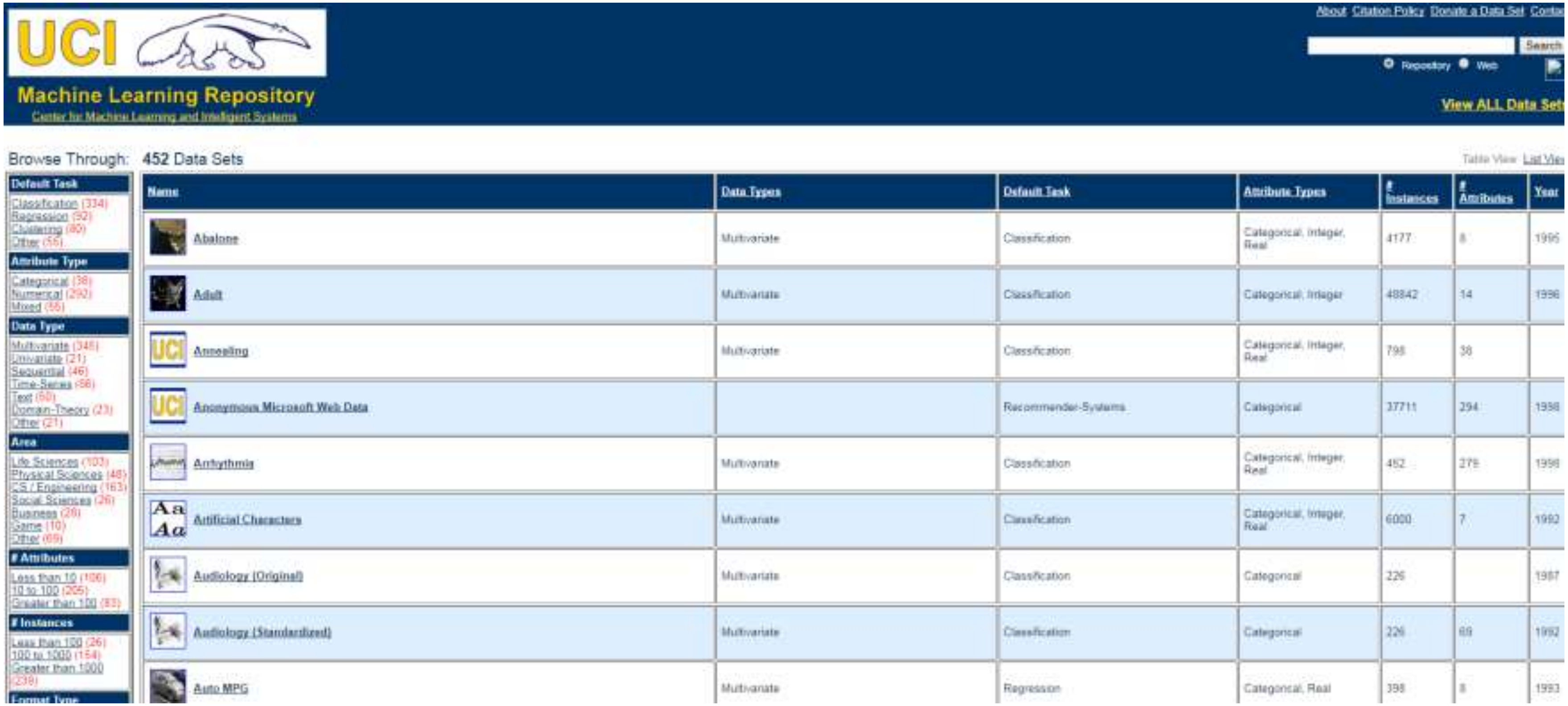
10-12-2018:	Superconductivty Data
10-08-2018:	Physical Unclonable Functions
10-04-2018:	Drug Review Dataset (Drugs.com)
10-02-2018:	PANDOR
10-02-2018:	Drug Review Dataset (Druglib.com)
09-16-2018:	Student Academics Performance
09-14-2018:	WESAD (Wearable Stress and Affect Detection)
09-13-2018:	GNFUV Unmanned Surface Vehicles Sensor Data Set 2
08-31-2018:	Online Shoppers' Purchasing Intention Dataset

Most Popular Data Sets (hits since 2007):

2161656:	Iris
1283677:	Adult
982469:	Wine
845658:	Car Evaluation
779087:	Breast Cancer Wisconsin (Diagnostic)
752155:	Wine Quality
745831:	Heart Disease
707836:	Human Activity Recognition Using Smartphones
700376:	Bank Marketing

Public datasets

- UCI machine learning repository
  - View all data sets



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
About Citation Policy Donate a Data Set Contac
Search
○ Repository ● Web
View ALL Data Sets

Browse Through: 452 Data Sets

Table View List Vie

Default Task
Classification (334)
Regression (92)
Clustering (80)
Other (55)

Attribute Type
Categorical (38)
Numerical (292)
Mixed (55)

Data Type
Multivariate (345)
Univariate (21)
Sequential (46)
Time-Series (56)
Text (60)
Domain-Theory (23)
Other (21)

Area
Life Sciences (103)
Physical Sciences (48)
CS / Engineering (163)
Social Sciences (26)
Business (28)
Game (10)
Other (69)

# Attributes
Less than 10 (106)
10 to 100 (205)
Greater than 100 (83)

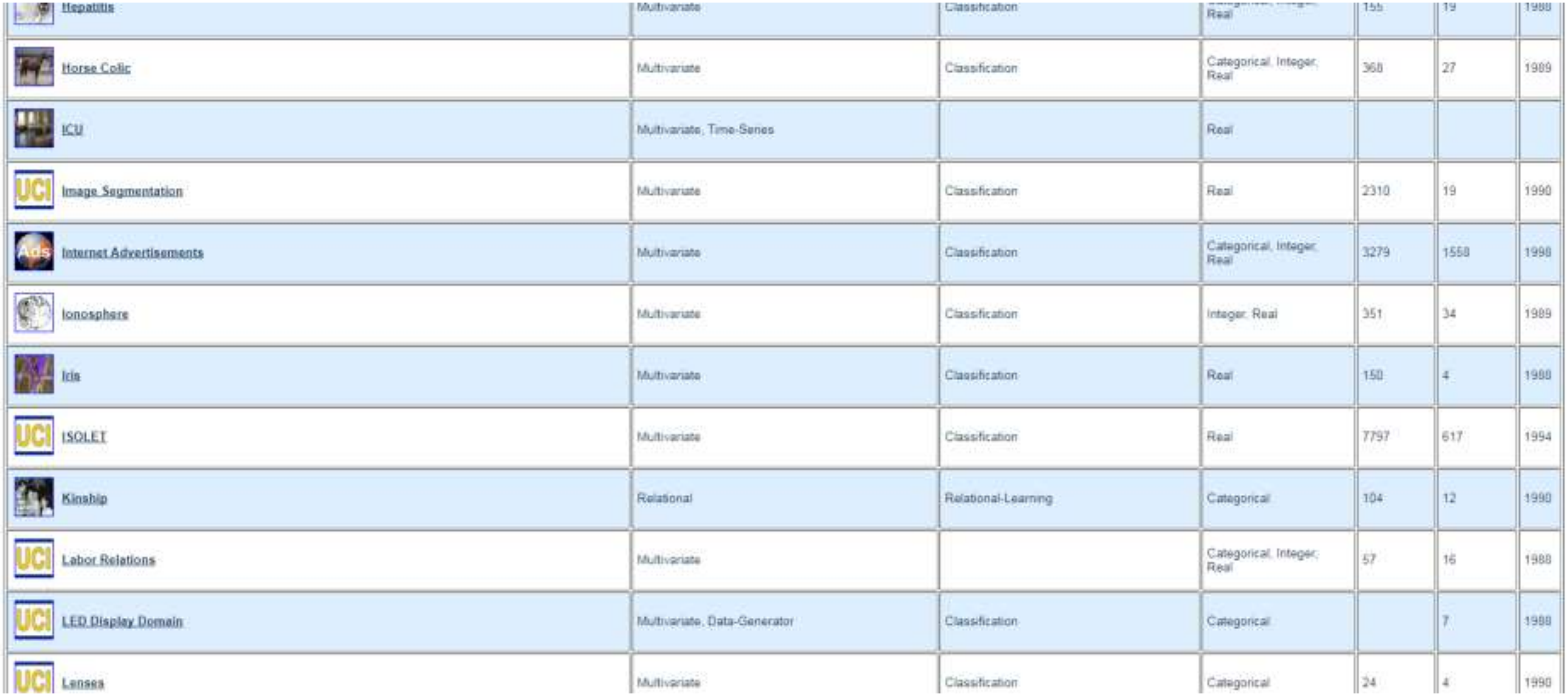
# Instances
Less than 100 (26)
100 to 1000 (154)
Greater than 1000 (239)

Format Type

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Audiology (Original)	Multivariate	Classification	Categorical	226		1987
Audiology (Standardized)	Multivariate	Classification	Categorical	226	69	1992
Auto MPG	Multivariate	Regression	Categorical, Real	398	8	1993

Public datasets

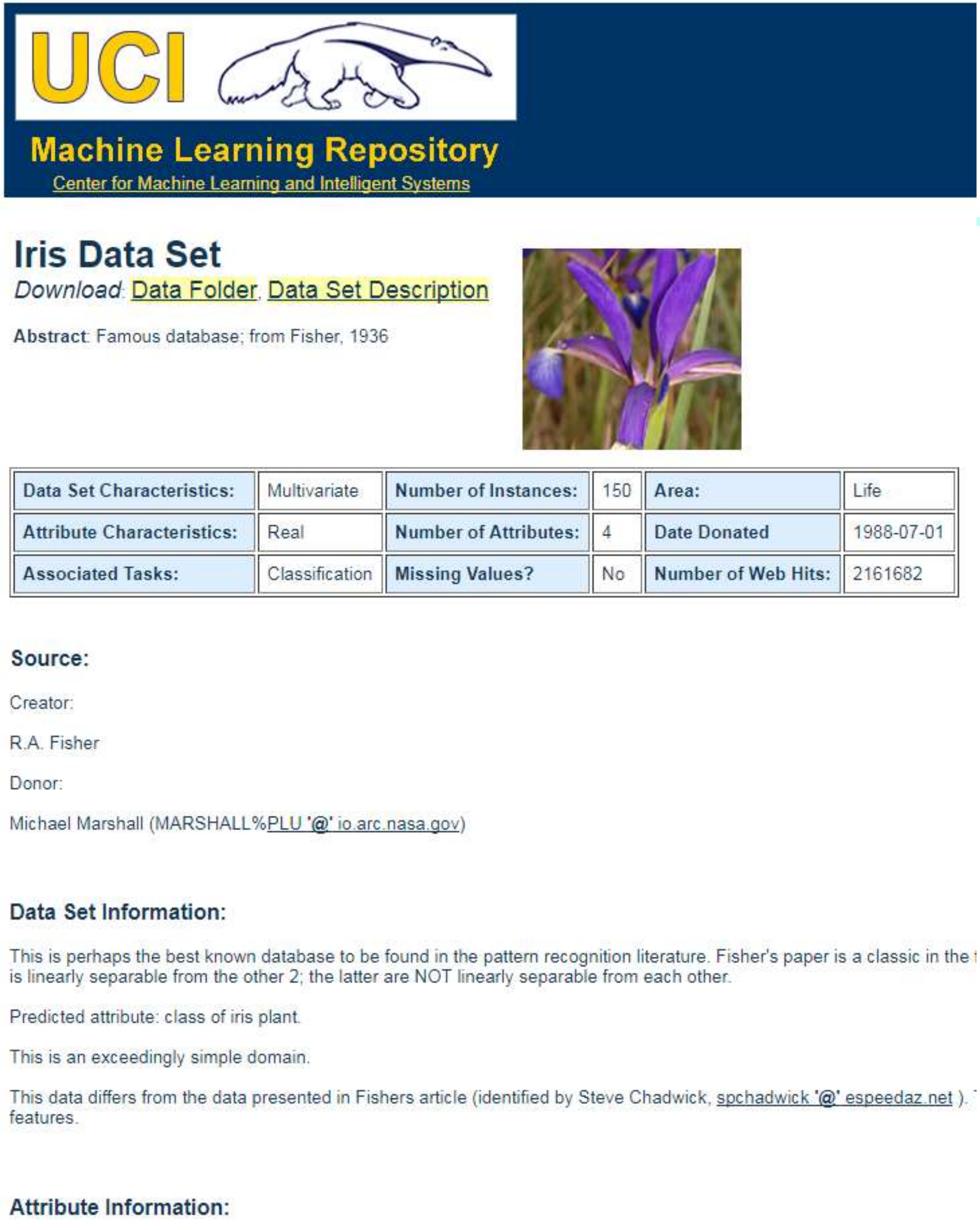
- UCI machine learning repository:
  - Search for a data set, e.g. Iris



Hepatitis	Multivariate	Classification	Categorical, Integer, Real	155	19	1988
Horse Colic	Multivariate	Classification	Categorical, Integer, Real	368	27	1989
ICU	Multivariate, Time-Series		Real			
Image Segmentation	Multivariate	Classification	Real	2310	19	1990
Internet Advertisements	Multivariate	Classification	Categorical, Integer, Real	3279	1558	1998
Ionosphere	Multivariate	Classification	Integer, Real	351	34	1989
Iris	Multivariate	Classification	Real	150	4	1988
ISOLET	Multivariate	Classification	Real	7797	617	1994
Kinship	Relational	Relational-Learning	Categorical	104	12	1990
Labor Relations	Multivariate		Categorical, Integer, Real	57	16	1988
LED Display Domain	Multivariate, Data-Generator	Classification	Categorical		7	1988
Lenses	Multivariate	Classification	Categorical	24	4	1990

P

- UCI machine learnir
  - Data set information



UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	2161682

Source:

Creator:

R.A. Fisher

Donor:

Michael Marshall (MARSHALL%PLU '@' io.arc.nasa.gov)

Data Set Information:

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the
is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Predicted attribute: class of iris plant.

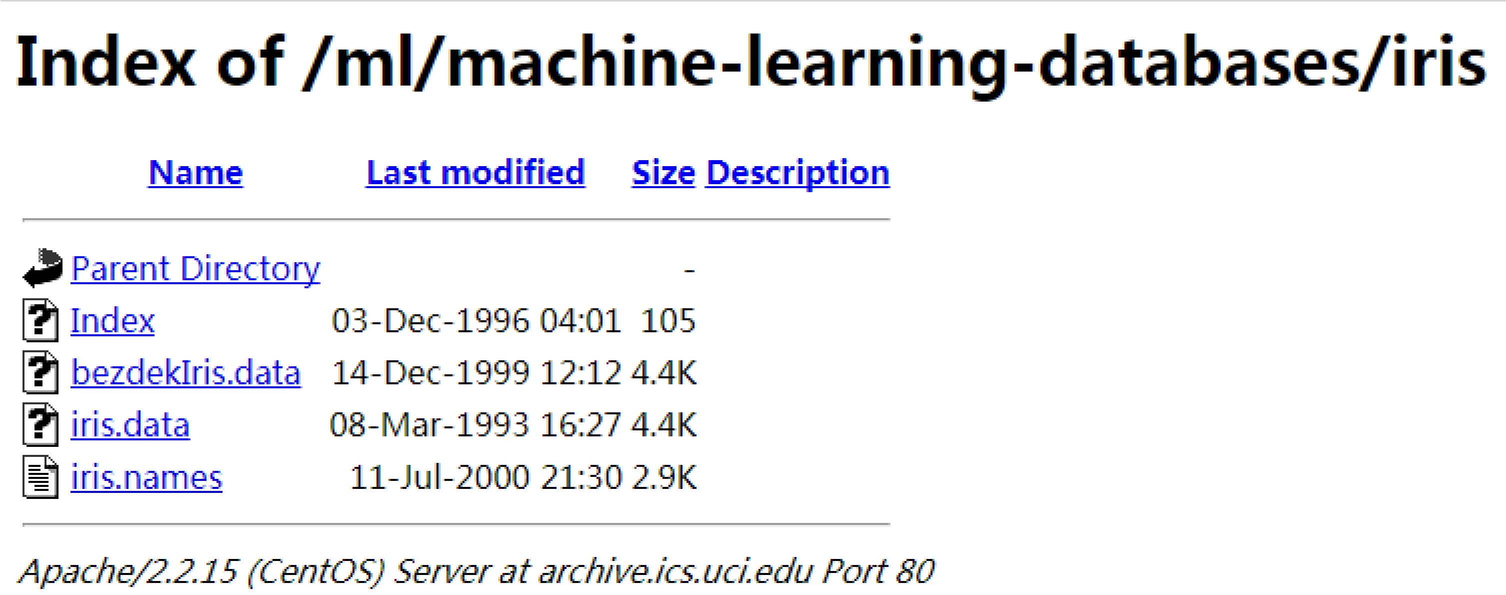
This is an exceedingly simple domain.

This data differs from the data presented in Fishers article (identified by Steve Chadwick, spchadwick '@' espeedaz.net).
features.

Attribute Information:

Public datasets

- UCI machine learning repository
  - Download data



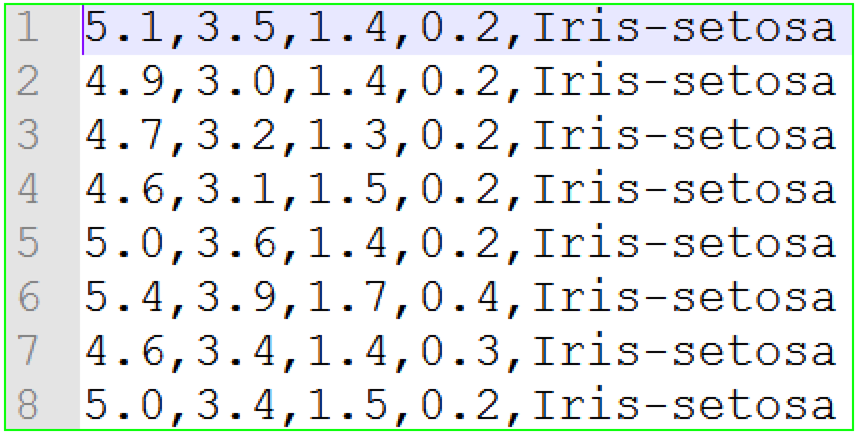
Index of /ml/machine-learning-databases/iris

Name	Last modified	Size	Description
Parent Directory		-	
Index	03-Dec-1996 04:01	105	
bezdekIris.data	14-Dec-1999 12:12	4.4K	
iris.data	08-Mar-1993 16:27	4.4K	
iris.names	11-Jul-2000 21:30	2.9K	

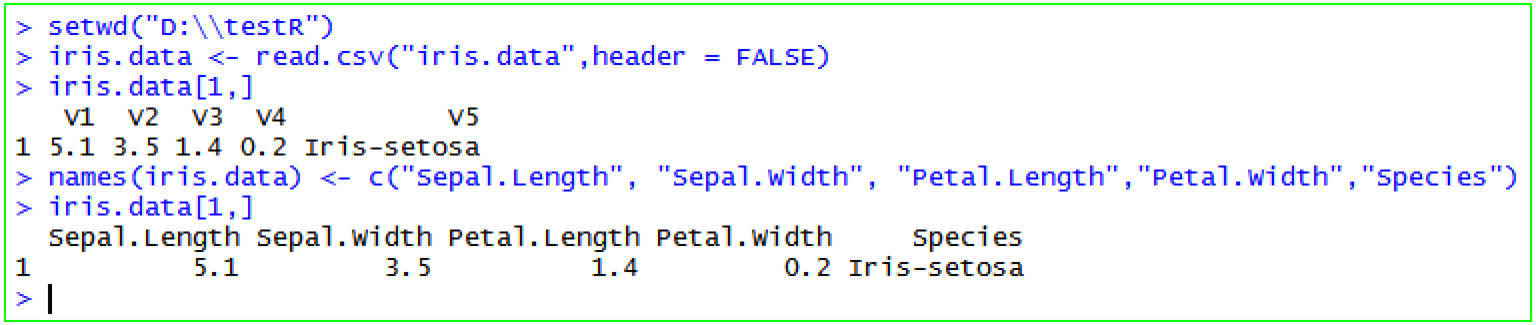
Apache/2.2.15 (CentOS) Server at archive.ics.uci.edu Port 80

Public datasets

- UCI machine learning repository
  - Read the data



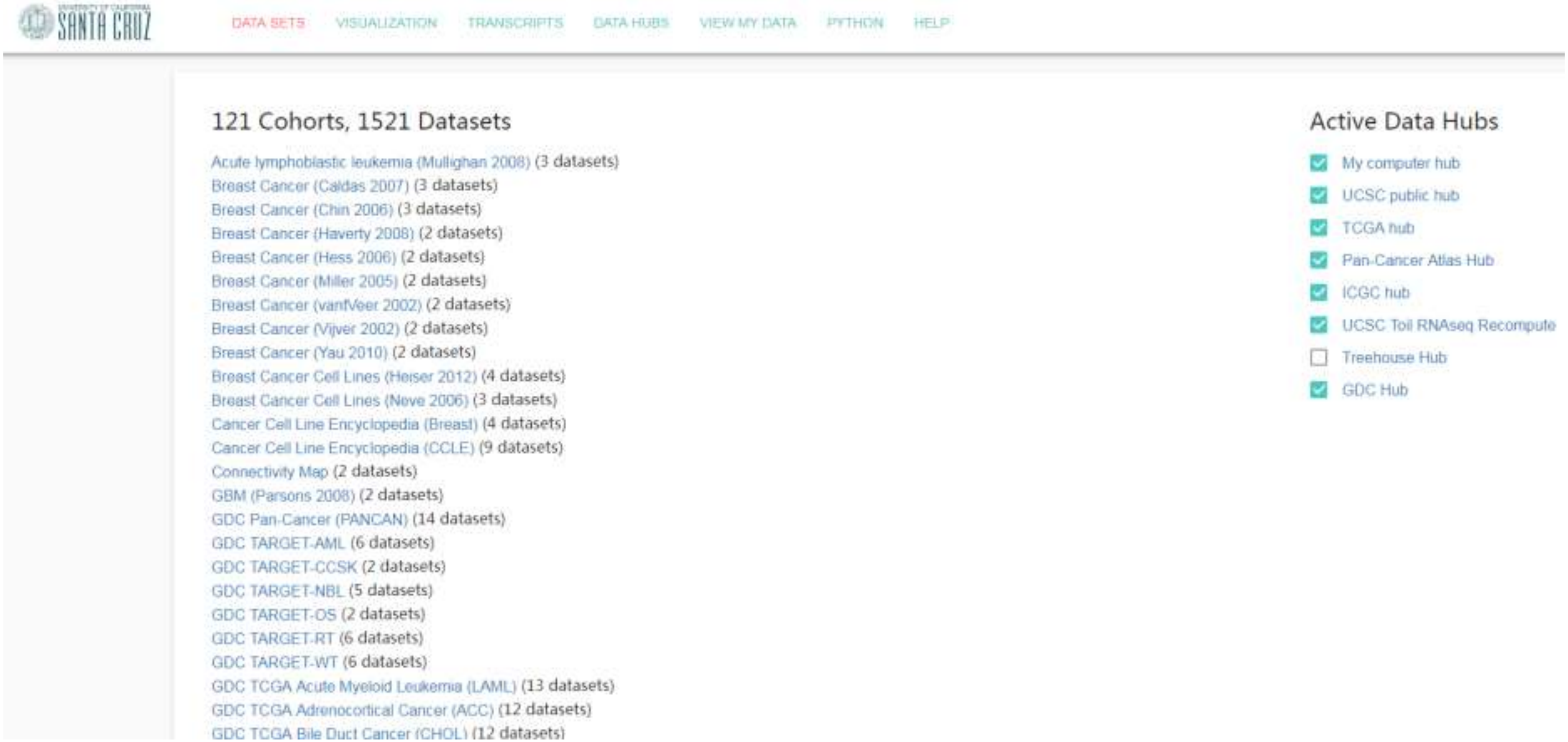
```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
```



```
> setwd("D:\\testR")
> iris.data <- read.csv("iris.data",header = FALSE)
> iris.data[1,]
   V1  V2  V3  V4          V5
1 5.1 3.5 1.4 0.2 Iris-setosa
> names(iris.data) <- c("Sepal.Length", "Sepal.Width", "Petal.Length","Petal.Width","Species")
> iris.data[1,]
  Sepal.Length Sepal.Width Petal.Length Petal.Width     Species
1          5.1         3.5          1.4         0.2 Iris-setosa
> 
```

Public datasets

- TCGA (The Cancer Genome Atlas)
  - https://xenabrowser.net/datapages/



121 Cohorts, 1521 Datasets

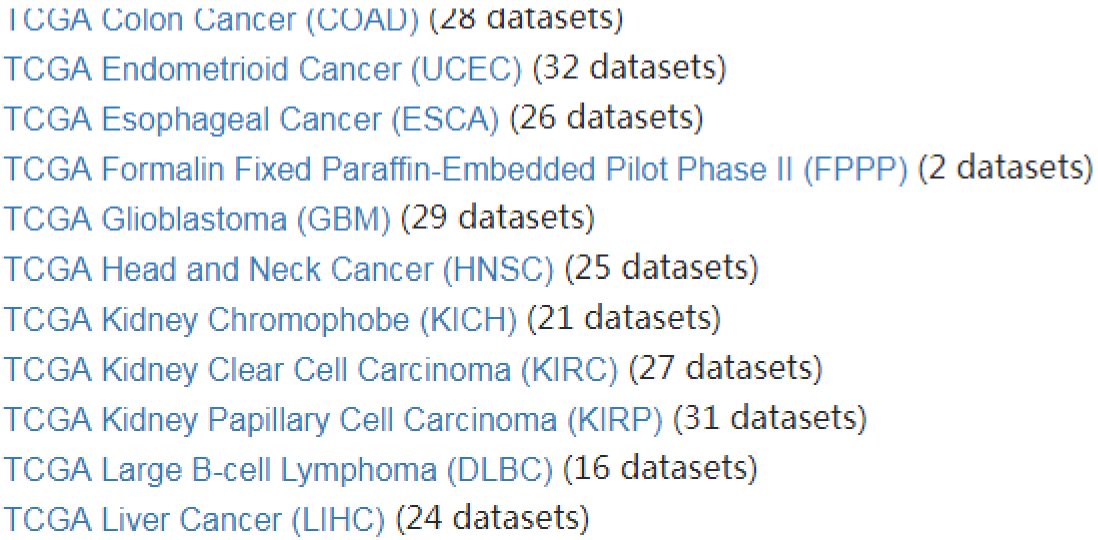
Acute lymphoblastic leukemia (Mullighan 2008) (3 datasets)
Breast Cancer (Caldas 2007) (3 datasets)
Breast Cancer (Chin 2006) (3 datasets)
Breast Cancer (Haverty 2008) (2 datasets)
Breast Cancer (Hess 2006) (2 datasets)
Breast Cancer (Miller 2005) (2 datasets)
Breast Cancer (vantVeer 2002) (2 datasets)
Breast Cancer (Vijver 2002) (2 datasets)
Breast Cancer (Yau 2010) (2 datasets)
Breast Cancer Cell Lines (Heiser 2012) (4 datasets)
Breast Cancer Cell Lines (Neve 2006) (3 datasets)
Cancer Cell Line Encyclopedia (Breast) (4 datasets)
Cancer Cell Line Encyclopedia (CCLE) (9 datasets)
Connectivity Map (2 datasets)
GBM (Parsons 2008) (2 datasets)
GDC Pan-Cancer (PANCAN) (14 datasets)
GDC TARGET-AML (6 datasets)
GDC TARGET-CCSK (2 datasets)
GDC TARGET-NBL (5 datasets)
GDC TARGET-OS (2 datasets)
GDC TARGET-RT (6 datasets)
GDC TARGET-WT (6 datasets)
GDC TCGA Acute Myeloid Leukemia (LAML) (13 datasets)
GDC TCGA Adrenocortical Cancer (ACC) (12 datasets)
GDC TCGA Bile Duct Cancer (CHOL) (12 datasets)

Active Data Hubs

☑ My computer hub
☑ UCSC public hub
☑ TCGA hub
☑ Pan-Cancer Atlas Hub
☑ ICGC hub
☑ UCSC Toil RNAseq Recompute
☐ Treehouse Hub
☑ GDC Hub

Public datasets

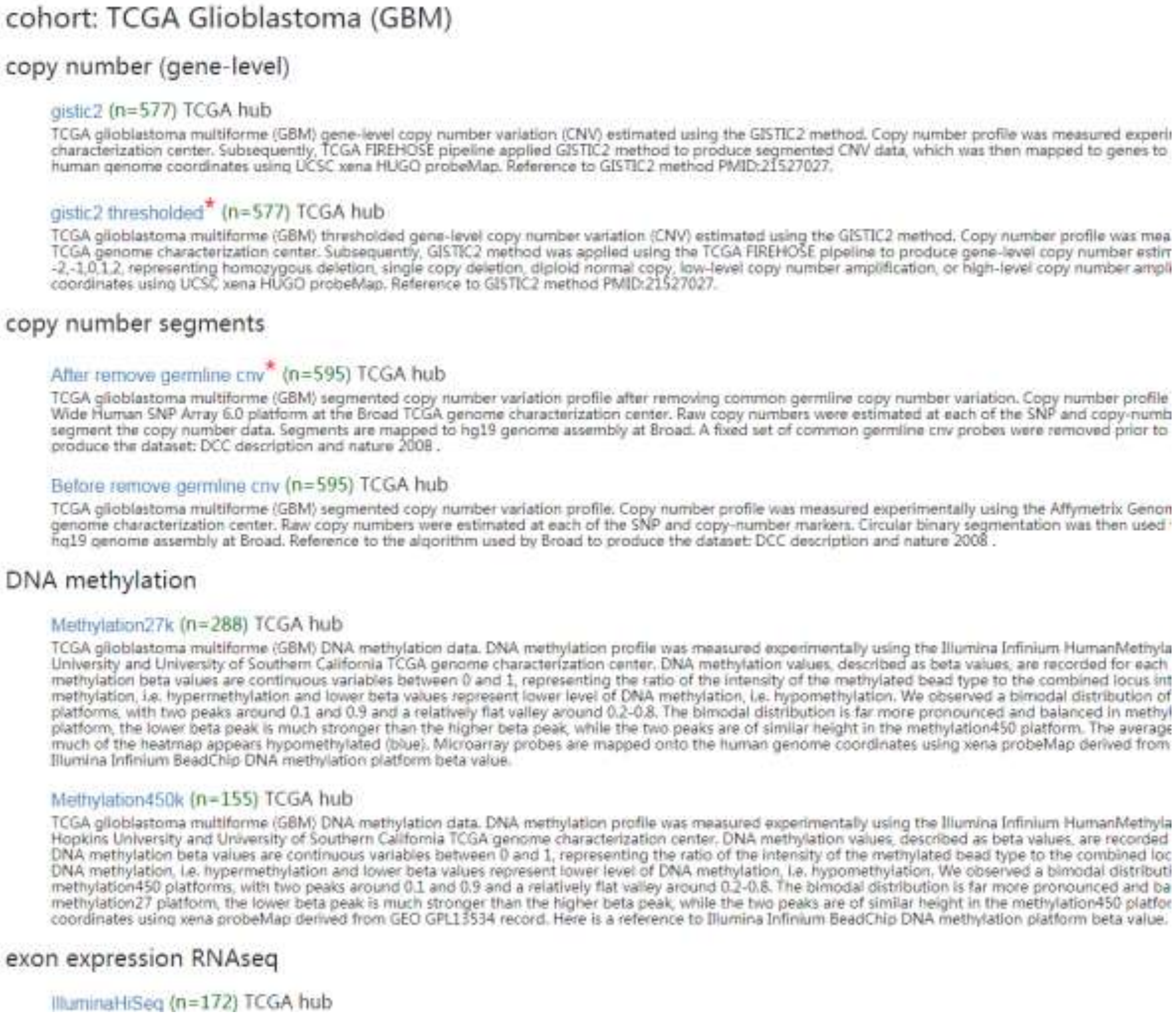
- TCGA (The Cancer Genome Atlas)
  - Select a cancer type



TCGA Colon Cancer (COAD) (28 datasets)
TCGA Endometrioid Cancer (UCEC) (32 datasets)
TCGA Esophageal Cancer (ESCA) (26 datasets)
TCGA Formalin Fixed Paraffin-Embedded Pilot Phase II (FPPP) (2 datasets)
TCGA Glioblastoma (GBM) (29 datasets)
TCGA Head and Neck Cancer (HNSC) (25 datasets)
TCGA Kidney Chromophobe (KICH) (21 datasets)
TCGA Kidney Clear Cell Carcinoma (KIRC) (27 datasets)
TCGA Kidney Papillary Cell Carcinoma (KIRP) (31 datasets)
TCGA Large B-cell Lymphoma (DLBC) (16 datasets)
TCGA Liver Cancer (LIHC) (24 datasets)

Public datasets

- TCGA (The Cancer Genome Atlas)
  - Data available



cohort: TCGA Glioblastoma (GBM)

copy number (gene-level)

gistic2 (n=577) TCGA hub

TCGA glioblastoma multiforme (GBM) gene-level copy number variation (CNV) estimated using the GISTIC2 method. Copy number profile was measured experi
characterization center. Subsequently, TCGA FIREHOSE pipeline applied GISTIC2 method to produce segmented CNV data, which was then mapped to genes to
human genome coordinates using UCSC xena HUGO probeMap. Reference to GISTIC2 method PMID:21527027.

gistic2 thresholded* (n=577) TCGA hub

TCGA glioblastoma multiforme (GBM) thresholded gene-level copy number variation (CNV) estimated using the GISTIC2 method. Copy number profile was mea
TCGA genome characterization center. Subsequently, GISTIC2 method was applied using the TCGA FIREHOSE pipeline to produce gene-level copy number estir
-2,-1,0,1,2, representing homozygous deletion, single copy deletion, diploid normal copy, low-level copy number amplification, or high-level copy number ampli
coordinates using UCSC xena HUGO probeMap. Reference to GISTIC2 method PMID:21527027.

copy number segments

After remove germline cnv* (n=595) TCGA hub

TCGA glioblastoma multiforme (GBM) segmented copy number variation profile after removing common germline copy number variation. Copy number profile
Wide Human SNP Array 6.0 platform at the Broad TCGA genome characterization center. Raw copy numbers were estimated at each of the SNP and copy-numb
segment the copy number data. Segments are mapped to hg19 genome assembly at Broad. A fixed set of common germline cnv probes were removed prior to
produce the dataset: DCC description and nature 2008 .

Before remove germline cnv (n=595) TCGA hub

TCGA glioblastoma multiforme (GBM) segmented copy number variation profile. Copy number profile was measured experimentally using the Affymetrix Genon
genome characterization center. Raw copy numbers were estimated at each of the SNP and copy-number markers. Circular binary segmentation was then used
hg19 genome assembly at Broad. Reference to the algorithm used by Broad to produce the dataset: DCC description and nature 2008 .

DNA methylation

Methylation27k (n=288) TCGA hub

TCGA glioblastoma multiforme (GBM) DNA methylation data. DNA methylation profile was measured experimentally using the Illumina Infinium HumanMethyla
University and University of Southern California TCGA genome characterization center. DNA methylation values, described as beta values, are recorded for each
methylation beta values are continuous variables between 0 and 1, representing the ratio of the intensity of the methylated bead type to the combined locus int
methylation, i.e. hypermethylation and lower beta values represent lower level of DNA methylation, i.e. hypomethylation. We observed a bimodal distribution of
platforms, with two peaks around 0.1 and 0.9 and a relatively flat valley around 0.2-0.8. The bimodal distribution is far more pronounced and balanced in methyl
platform, the lower beta peak is much stronger than the higher beta peak, while the two peaks are of similar height in the methylation450 platform. The average
much of the heatmap appears hypomethylated (blue). Microarray probes are mapped onto the human genome coordinates using xena probeMap derived from
Illumina Infinium BeadChip DNA methylation platform beta value.

Methylation450k (n=155) TCGA hub

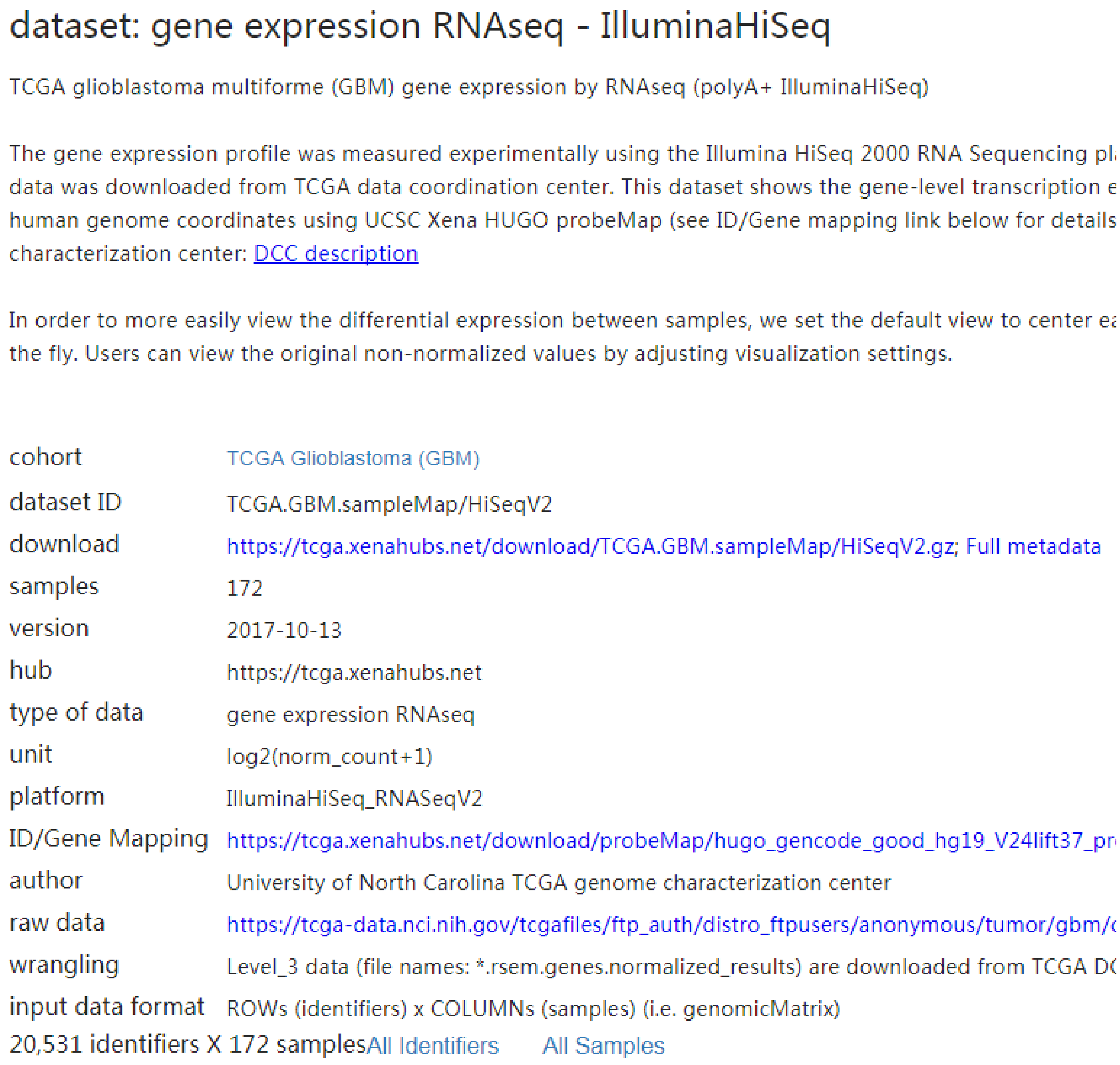
TCGA glioblastoma multiforme (GBM) DNA methylation data. DNA methylation profile was measured experimentally using the Illumina Infinium HumanMethyla
Hopkins University and University of Southern California TCGA genome characterization center. DNA methylation values, described as beta values, are recorded
DNA methylation beta values are continuous variables between 0 and 1, representing the ratio of the intensity of the methylated bead type to the combined loc
DNA methylation, i.e. hypermethylation and lower beta values represent lower level of DNA methylation, i.e. hypomethylation. We observed a bimodal distributi
methylation450 platforms, with two peaks around 0.1 and 0.9 and a relatively flat valley around 0.2-0.8. The bimodal distribution is far more pronounced and ba
methylation27 platform, the lower beta peak is much stronger than the higher beta peak, while the two peaks are of similar height in the methylation450 platfor
coordinates using xena probeMap derived from GEO GPL13534 record. Here is a reference to Illumina Infinium BeadChip DNA methylation platform beta value.

exon expression RNAseq

IlluminaHiSeq (n=172) TCGA hub

Public datasets

- TCGA (The Cancer
    - Download data



dataset: gene expression RNAseq - IlluminaHiSeq

TCGA glioblastoma multiforme (GBM) gene expression by RNAseq (polyA+ IlluminaHiSeq)

The gene expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing pl data was downloaded from TCGA data coordination center. This dataset shows the gene-level transcription e human genome coordinates using UCSC Xena HUGO probeMap (see ID/Gene mapping link below for details characterization center: DCC description

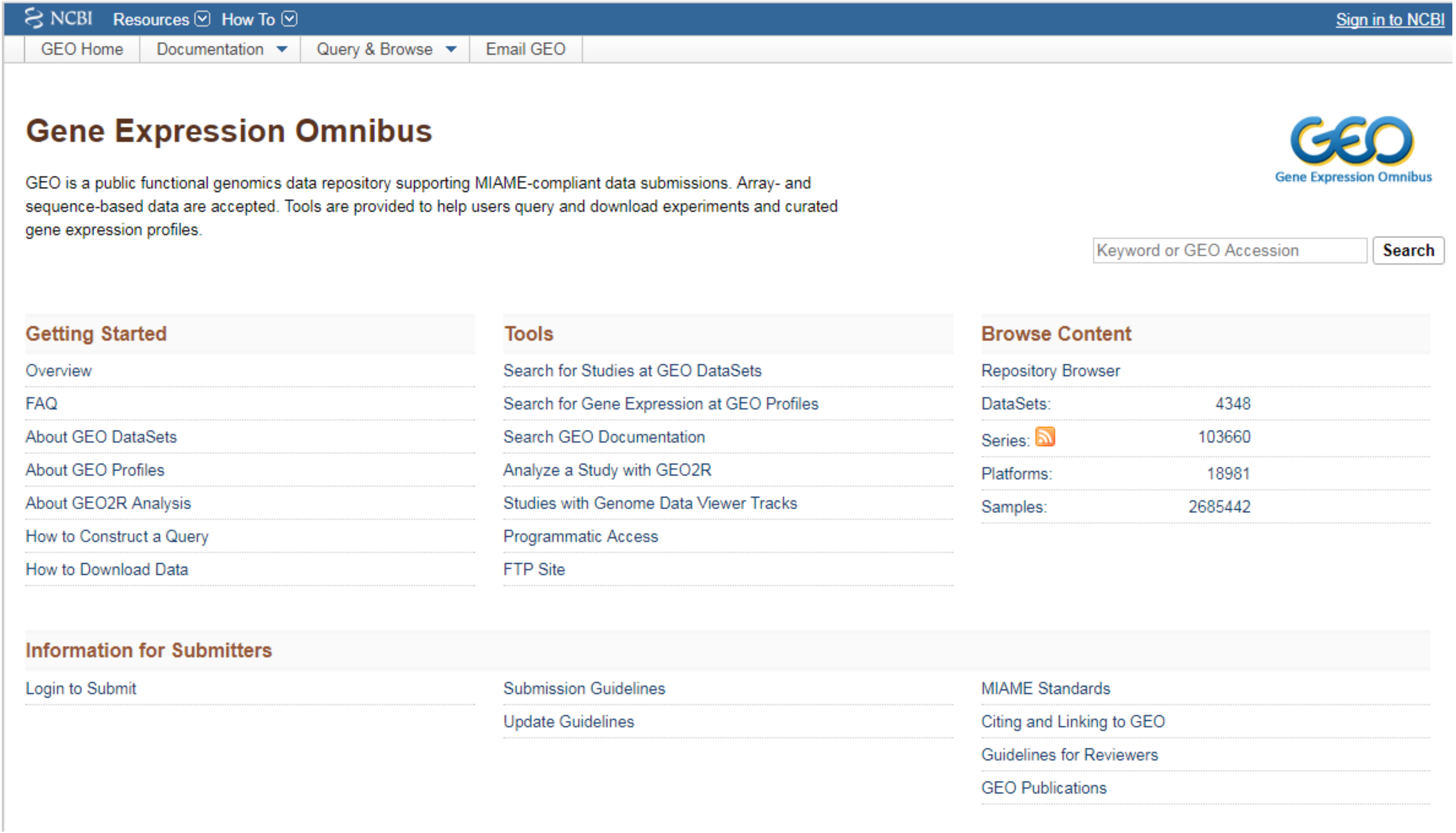
In order to more easily view the differential expression between samples, we set the default view to center e the fly. Users can view the original non-normalized values by adjusting visualization settings.

cohort	TCGA Glioblastoma (GBM)
dataset ID	TCGA.GBM.sampleMap/HiSeqV2
download	https://tcga.xenahubs.net/download/TCGA.GBM.sampleMap/HiSeqV2.gz; Full metadata
samples	172
version	2017-10-13
hub	https://tcga.xenahubs.net
type of data	gene expression RNAseq
unit	log2(norm_count+1)
platform	IlluminaHiSeq_RNASeqV2
ID/Gene Mapping	https://tcga.xenahubs.net/download/probeMap/hugo_gencode_good_hg19_V24lift37_pr
author	University of North Carolina TCGA genome characterization center
raw data	https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/gbm/
wrangling	Level_3 data (file names: *.rsem.genes.normalized_results) are downloaded from TCGA D
input data format	ROWs (identifiers) x COLUMNs (samples) (i.e. genomicMatrix)

20,531 identifiers X 172 samplesAll Identifiers All Samples

Public datasets

- GEO (Gene Expression Omnibus)
  - https://www.ncbi.nlm.nih.gov/geo/



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

Browse Content

Repository Browser

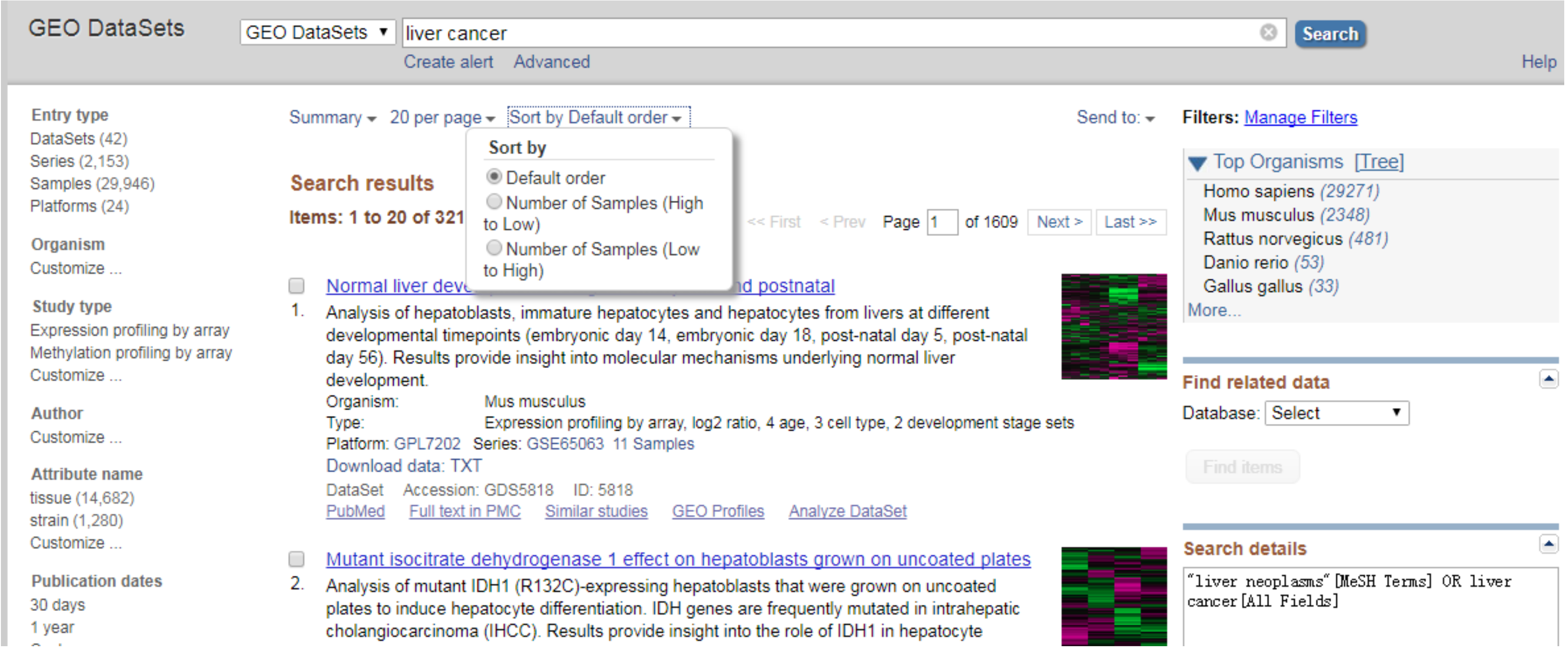
DataSets:	4348
Series:	103660
Platforms:	18981
Samples:	2685442

Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

Public datasets

- GEO (Gene Expression Omnibus)
  - Search for a data set or a cancer type



GEO DataSets | GEO DataSets ▾ | liver cancer | Search

Create alert Advanced Help

Entry type
DataSets (42)
Series (2,153)
Samples (29,946)
Platforms (24)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling by array
Customize ...

Author
Customize ...

Attribute name
tissue (14,682)
strain (1,280)
Customize ...

Publication dates
30 days
1 year

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

Sort by
● Default order
○ Number of Samples (High to Low)
○ Number of Samples (Low to High)

Search results

Items: 1 to 20 of 321

<< First < Prev Page 1 of 1609 Next > Last >>

□ 1. Normal liver deve ... nd postnatal

Analysis of hepatoblasts, immature hepatocytes and hepatocytes from livers at different developmental timepoints (embryonic day 14, embryonic day 18, post-natal day 5, post-natal day 56). Results provide insight into molecular mechanisms underlying normal liver development.
Organism: Mus musculus
Type: Expression profiling by array, log2 ratio, 4 age, 3 cell type, 2 development stage sets
Platform: GPL7202 Series: GSE65063 11 Samples
Download data: TXT
DataSet Accession: GDS5818 ID: 5818
PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

□ 2. Mutant isocitrate dehydrogenase 1 effect on hepatoblasts grown on uncoated plates

Analysis of mutant IDH1 (R132C)-expressing hepatoblasts that were grown on uncoated plates to induce hepatocyte differentiation. IDH genes are frequently mutated in intrahepatic cholangiocarcinoma (IHCC). Results provide insight into the role of IDH1 in hepatocyte

Filters: Manage Filters

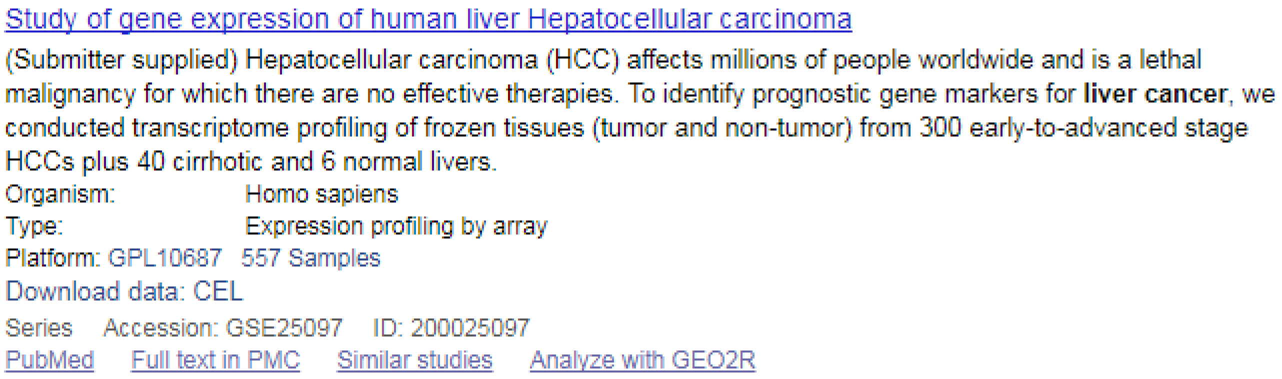
Top Organisms [Tree]
Homo sapiens (29271)
Mus musculus (2348)
Rattus norvegicus (481)
Danio rerio (53)
Gallus gallus (33)
More...

Find related data
Database: Select ▾
Find items

Search details
"liver neoplasms"[MeSH Terms] OR liver cancer[All Fields]

Public datasets

- GEO (Gene Expression Omnibus)
  - Search for a data set or a cancer type



Study of gene expression of human liver Hepatocellular carcinoma

(Submitter supplied) Hepatocellular carcinoma (HCC) affects millions of people worldwide and is a lethal malignancy for which there are no effective therapies. To identify prognostic gene markers for **liver cancer**, we conducted transcriptome profiling of frozen tissues (tumor and non-tumor) from 300 early-to-advanced stage HCCs plus 40 cirrhotic and 6 normal livers.

Organism:	Homo sapiens
Type:	Expression profiling by array

Platform: GPL10687 557 Samples

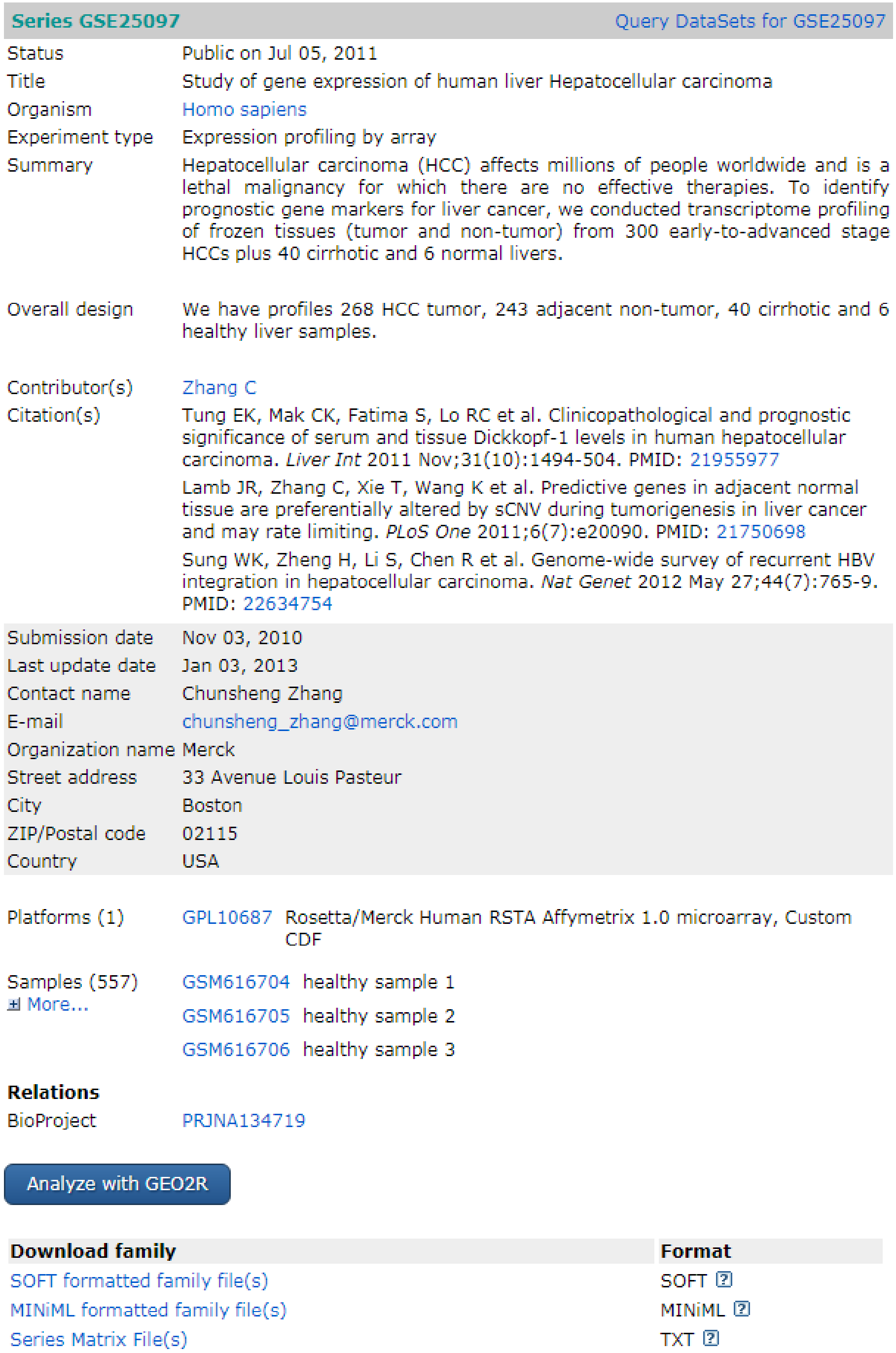
Download data: CEL

Series Accession: GSE25097 ID: 200025097

PubMed Full text in PMC Similar studies Analyze with GEO2R

Publi

- GEO (Gene Expression Om
  - Data information



Series GSE25097 Query DataSets for GSE25097

Status	Public on Jul 05, 2011
Title	Study of gene expression of human liver Hepatocellular carcinoma
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	Hepatocellular carcinoma (HCC) affects millions of people worldwide and is a lethal malignancy for which there are no effective therapies. To identify prognostic gene markers for liver cancer, we conducted transcriptome profiling of frozen tissues (tumor and non-tumor) from 300 early-to-advanced stage HCCs plus 40 cirrhotic and 6 normal livers.
Overall design	We have profiles 268 HCC tumor, 243 adjacent non-tumor, 40 cirrhotic and 6 healthy liver samples.
Contributor(s)	Zhang C
Citation(s)	Tung EK, Mak CK, Fatima S, Lo RC et al. Clinicopathological and prognostic significance of serum and tissue Dickkopf-1 levels in human hepatocellular carcinoma. *Liver Int* 2011 Nov;31(10):1494-504. PMID: 21955977 Lamb JR, Zhang C, Xie T, Wang K et al. Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. *PLoS One* 2011;6(7):e20090. PMID: 21750698 Sung WK, Zheng H, Li S, Chen R et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet* 2012 May 27;44(7):765-9. PMID: 22634754
Submission date	Nov 03, 2010
Last update date	Jan 03, 2013
Contact name	Chunsheng Zhang
E-mail	chunsheng_zhang@merck.com
Organization name	Merck
Street address	33 Avenue Louis Pasteur
City	Boston
ZIP/Postal code	02115
Country	USA
Platforms (1)	GPL10687 Rosetta/Merck Human RSTA Affymetrix 1.0 microarray, Custom CDF
Samples (557) ⊞ More...	GSM616704 healthy sample 1 GSM616705 healthy sample 2 GSM616706 healthy sample 3

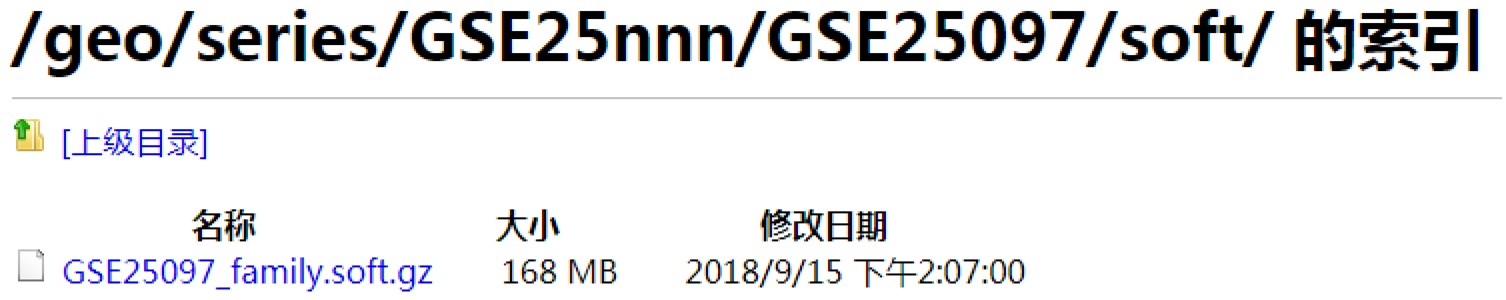
Relations

BioProject	PRJNA134719

Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

Public datasets

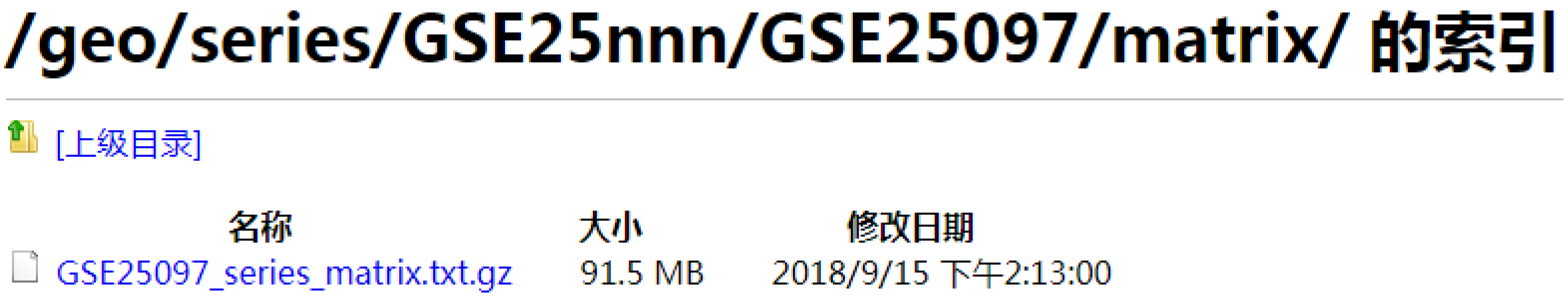
- GEO (Gene Expression Omnibus)
  - Download data



/geo/series/GSE25nnn/GSE25097/soft/ 的索引

[上级目录]

名称	大小	修改日期
GSE25097_family.soft.gz	168 MB	2018/9/15 下午2:07:00



/geo/series/GSE25nnn/GSE25097/matrix/ 的索引

[上级目录]

名称	大小	修改日期
GSE25097_series_matrix.txt.gz	91.5 MB	2018/9/15 下午2:13:00

Agenda

- Know your data
- Preprocess the data
- Some public datasets
- Summary

Summary

- Known your data
  - Data types
  - Statistical description of the data
  - Data visualization
  - Data similarity and dissimilarity
- Preprocess your data
  - Cleaning the data
  - Integration the data
  - Reduction the data
  - Dimensionality reduction
- Some public datasets available

Thank You!