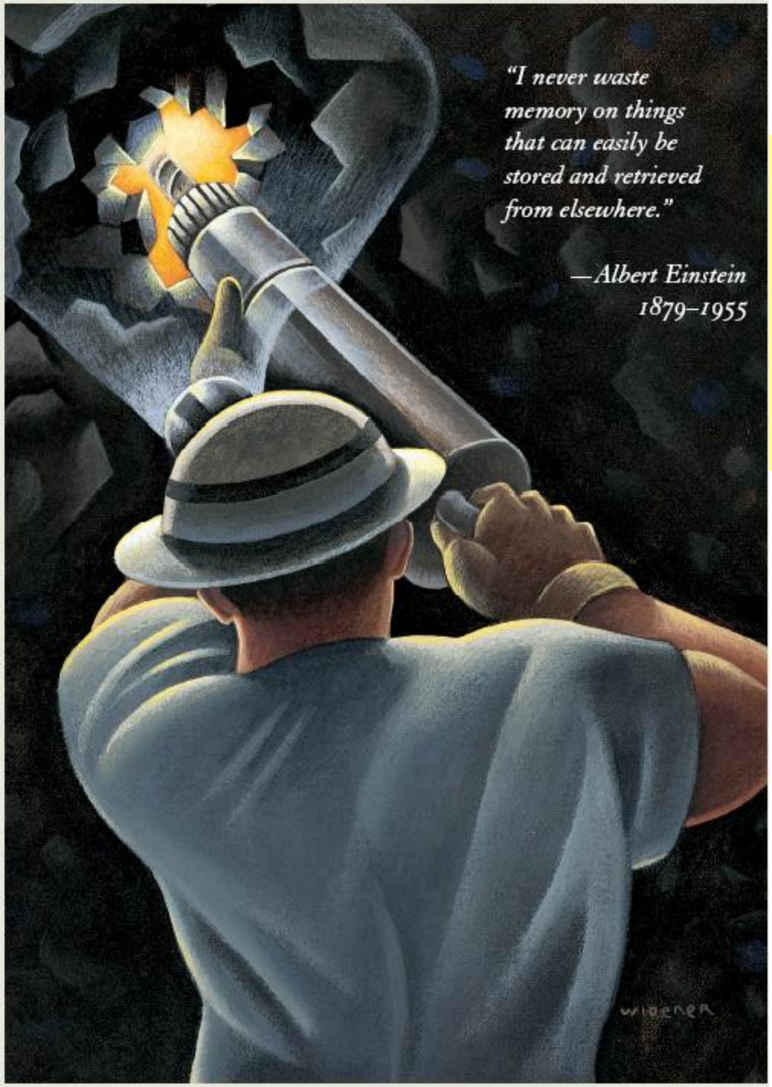
The University of Reading


"I never waste memory on things that can easily be stored and retrieved from elsewhere."
—Albert Einstein
1879–1955
WIDENER

MSc Advanced Computer Science

CSMDM16 - Data Analytics and Mining

Dr. Giuseppe Di Fatta
Associate Professor
Department of Computer Science
G.DiFatta@reading.ac.uk

Data Analytics and Mining

❑ Lectures:

Mon-Fri 10:00 – 13:00 (3 hours) (see timetable for location)

❑ Practical sessions:

Mon-Fri 14:00 – 17:00 (3 hours) W038 G43

❑ Textbooks:

- See Blackboard (Bb)

❑ Course Assessment:

- **100%** coursework (1 major assignment)
  - Released: during the last day of the module
  - Deadline: 1 week from the last day of the module:
    - electronic submission in Bb: archive containing your report (pdf) and workflow/code

Aims & Objectives

- **Aims**:

  Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories. Automated data analytics and mining techniques are becoming essential components to any information system. In the Knowledge Discovery process large data sets have to be cleaned, pre-processed, selected, merged, etc., and finally processed for the automatic extraction of interesting knowledge, such as descriptive and predictive models. The techniques span from statistics to machine learning and information science.

  This module focuses on concepts, methodologies, algorithms and tools for the design, management and deployment of the Knowledge Discovery process.

  In particular, tools for data analytics (R) and workflow management (KNIME) will be adopted for hands-on activities on several test cases. Student will learn general Data Mining principles and techniques and will apply them in different applicative domains.

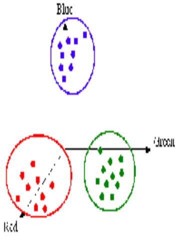
Assessable learning outcomes:

Students are expected to understand the general Knowledge Discovery process, the various Data Mining algorithms and techniques and to be able to apply them in different contexts. During practical activities the students will adopt state-of-the-art tools and languages for implementing data analytics and mining solutions for different applicative domains.

Outline

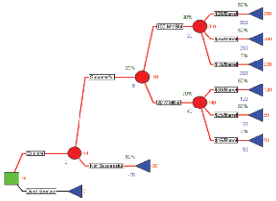
- Introduction to Data Mining
  - Data Mining tasks: Classification, Regression, Clustering, etc.

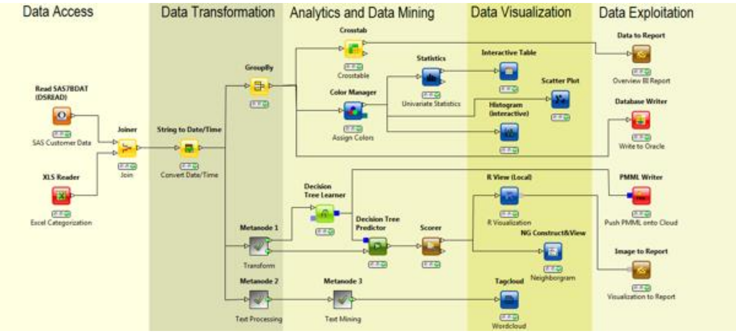
- KNIME: A Data Mining Workflow Management System

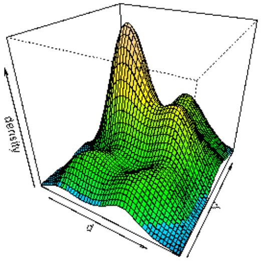


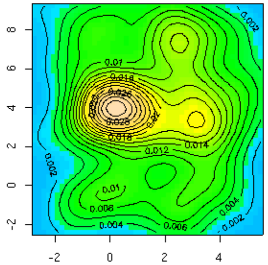
- R: A programming language for computational statistics, data analytics and mining, visualisation

- Integration of R and KNIME









The University of Reading

An Overview of Data Mining Tasks

Dr. Giuseppe Di Fatta

Associate Professor

Department of Computer Science

G.DiFatta@reading.ac.uk

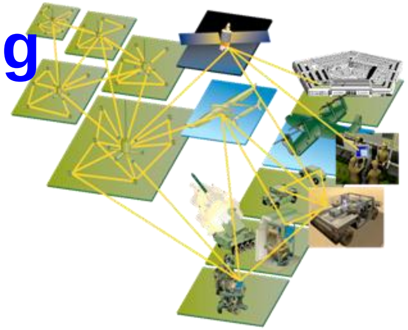
Why Mine Data? Commercial Viewpoint

Drowning in Data but Starving for Knowledge!

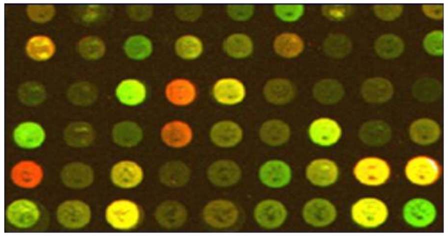
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

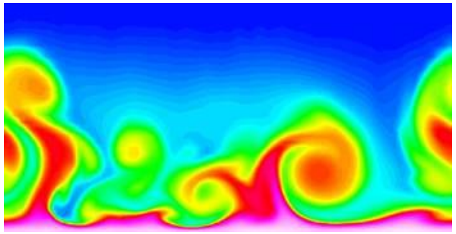
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

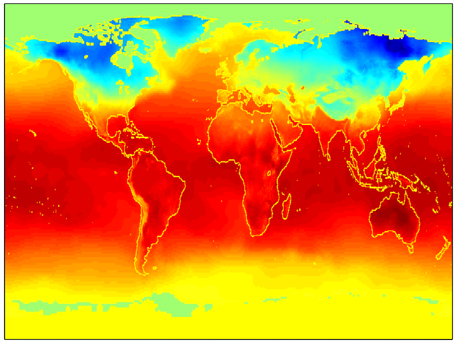
Why Mine Data? Scientific Viewpoint



Drowning in Data but Starving for Knowledge!







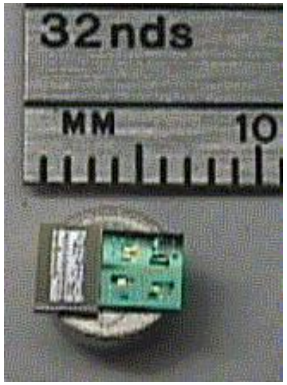
- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

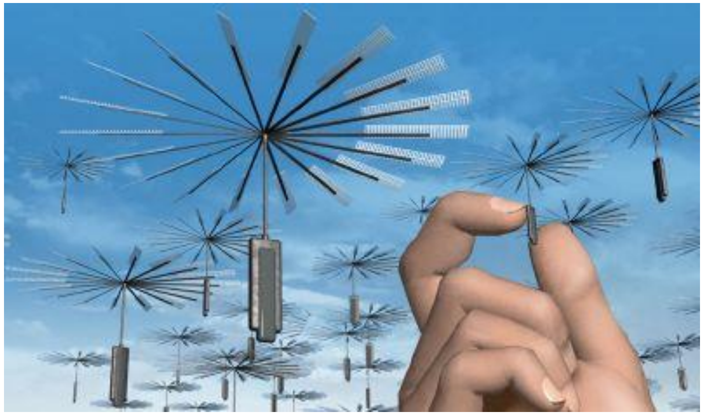
Why Mine Data? Looking ahead…

More and more data…

- RF Tag
  - Radiofrequency tags require no battery to read and operate, are cost-effective, and will read and write multiple tags in an RF field.

- Smart Dust
  - Miniature machines, each the size of a dust mote, may eventually saturate the environment, invisibly performing countless tasks.

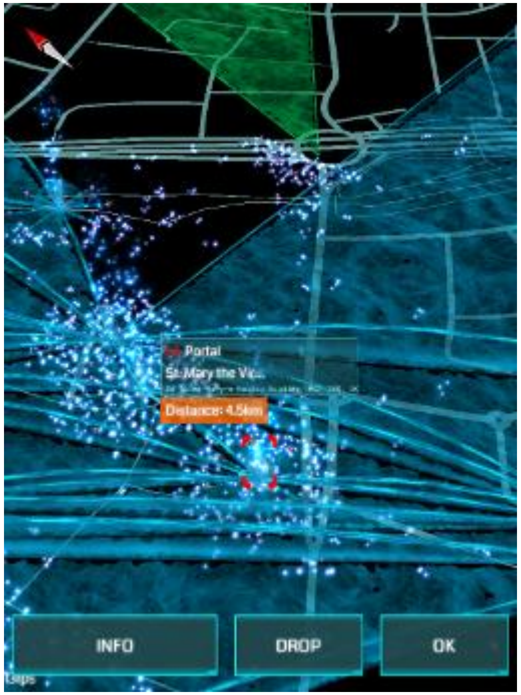


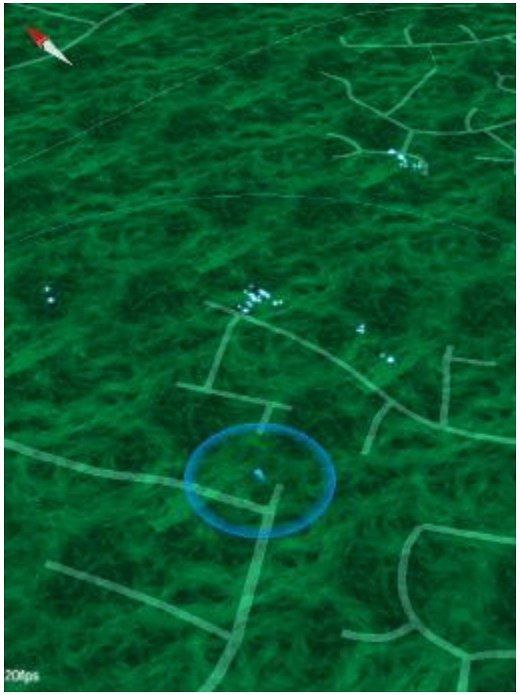


Why Mine Data? Looking ahead…

More and more data…

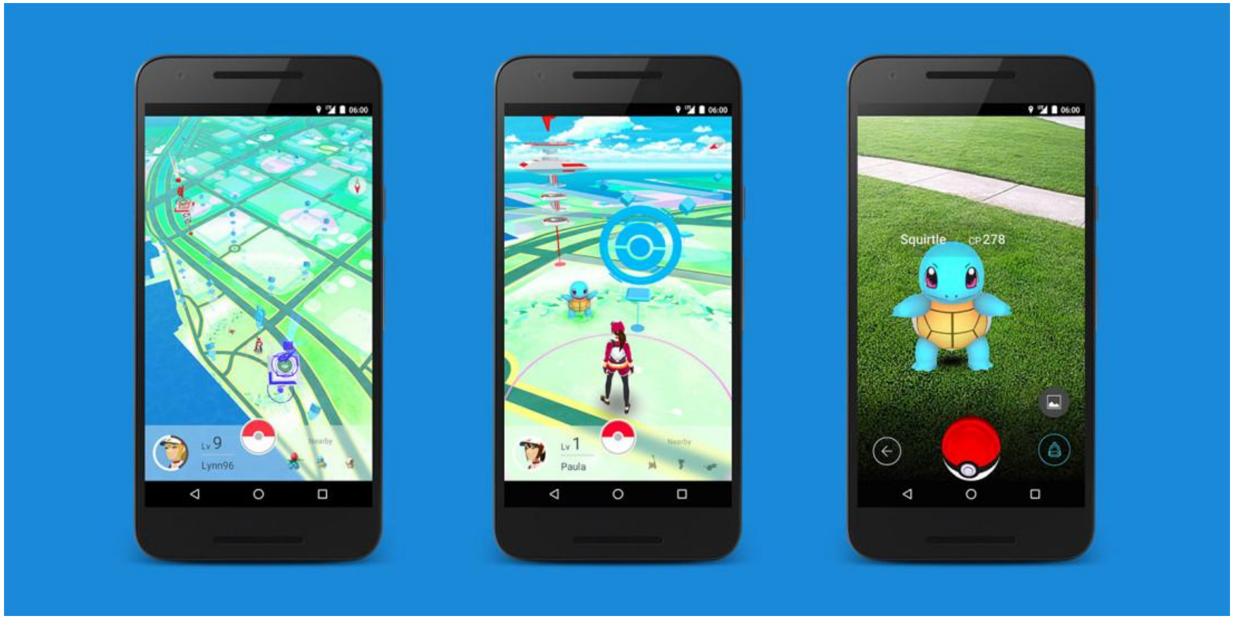
- **Ingress by NianticLabs@Google**
  - an augmented reality massively multiplayer online video game (MMOG), released for Android devices. The game has been called "**a data gold mine**" for Google.





Why Mine Data? Looking ahead…

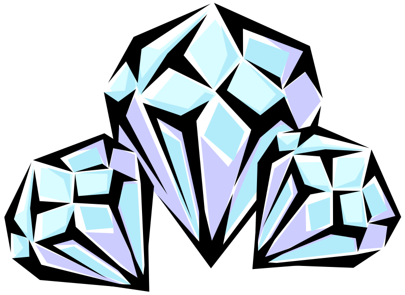
- Pokemon Go
  - an augmented reality massively multiplayer online video game created from previous Ingress, released for Android devices. The game reached a very high popularity in a few weeks.



What Is Data Mining?

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases

- Learning and describing concepts from data
- Alternative names:
  - Data mining: a misnomer?
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc.



Examples: What is (not) Data Mining?

- **What is not Data Mining?**
  - – Look up phone number in phone directory
  - – Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - – Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - – Group together similar documents returned by search engines according to their context
    - ➢ try searching for ambiguous words at www.kartoo.com

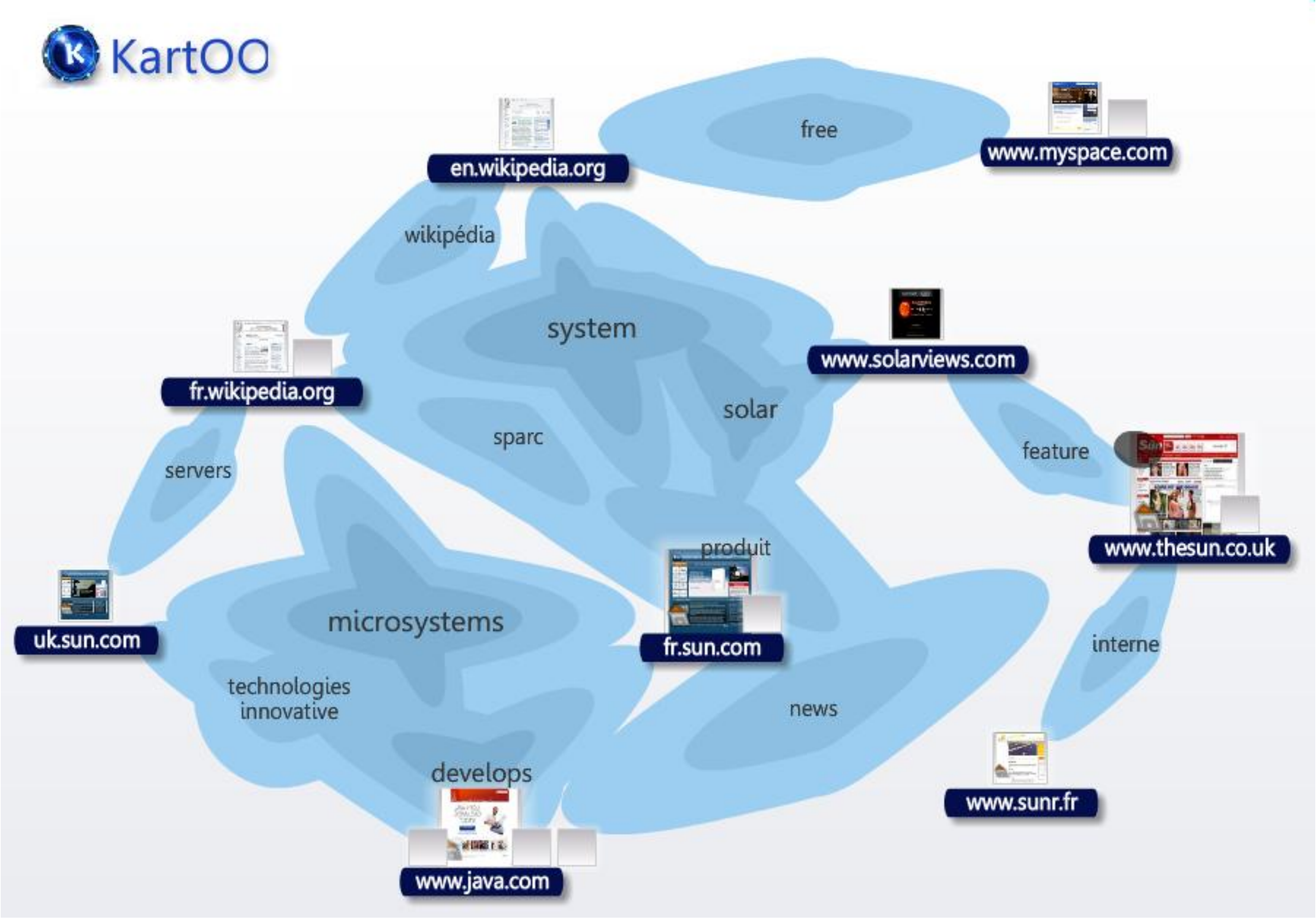
Data Mining is not…

- Data warehousing
- SQL / Ad Hoc Queries / Reporting
- Software Agents
- Online Analytical Processing (OLAP)
- Data Visualization

Data Mining is… about

- describing and understanding data in order to extract "hidden" knowledge, i.e. concepts.
  - analyze the data
  - find relations in the data
  - describe the relations

Searching for “sun”



Data Mining: classification schemes

- Decisions in data mining
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications
- Data mining tasks
  - Descriptive data mining
  - Predictive data mining

Data Mining Tasks

- **Predictive Tasks**
  - Use some variables to predict unknown or future values of other variables
- **Descriptive Tasks**
  - Find human-interpretable patterns that describe the data.

Common data mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification: Definition

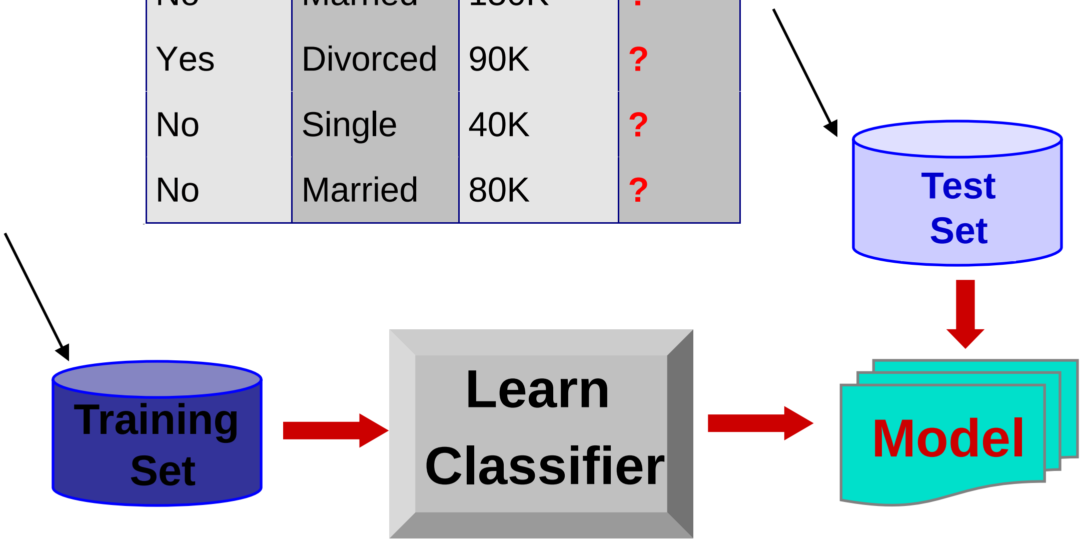
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for the class attribute as a function of the values of other attributes.
- Goal: <u>previously unseen</u> records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	**?**
Yes	Married	50K	**?**
No	Married	150K	**?**
Yes	Divorced	90K	**?**
No	Single	40K	**?**
No	Married	80K	**?**



Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Classification: Application 4

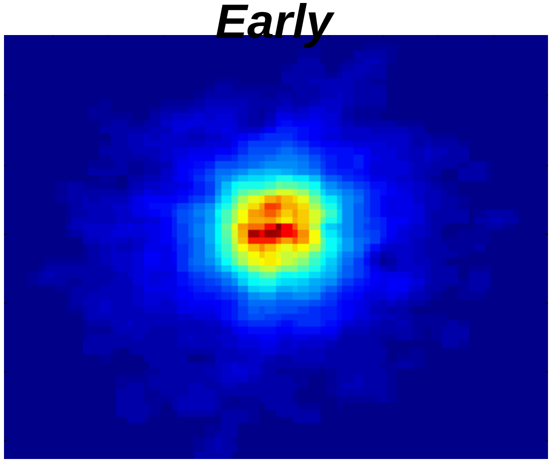
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

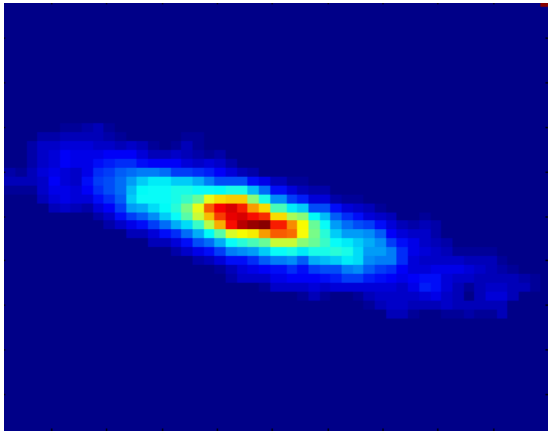
Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

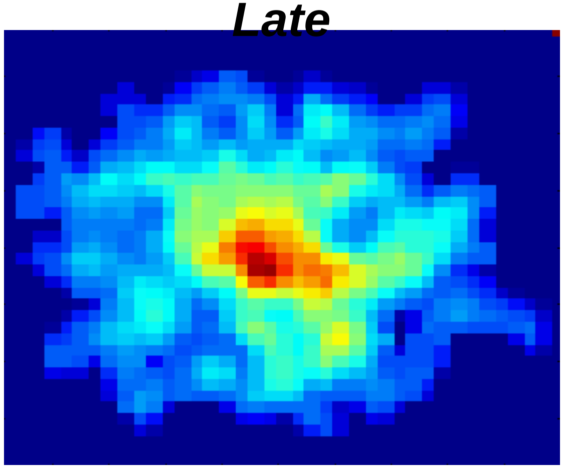
Early



Intermediate



Late



Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Clustering Definition

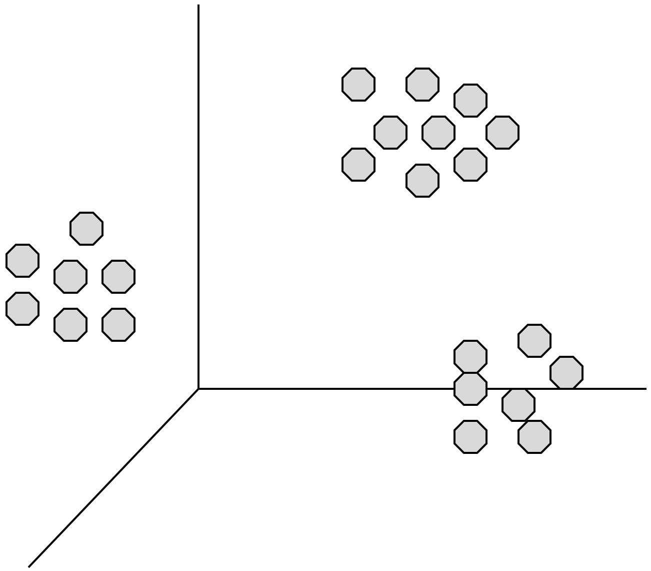
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean distance if attributes are continuous.
  - Other problem-specific measures.

Illustrating Clustering

- Problem: given the 3D data objects below, find clusters of similar objects, where similarity is defined in terms of the Euclidean distance.

Intracluster distances are minimized

Intercluster distances are maximized

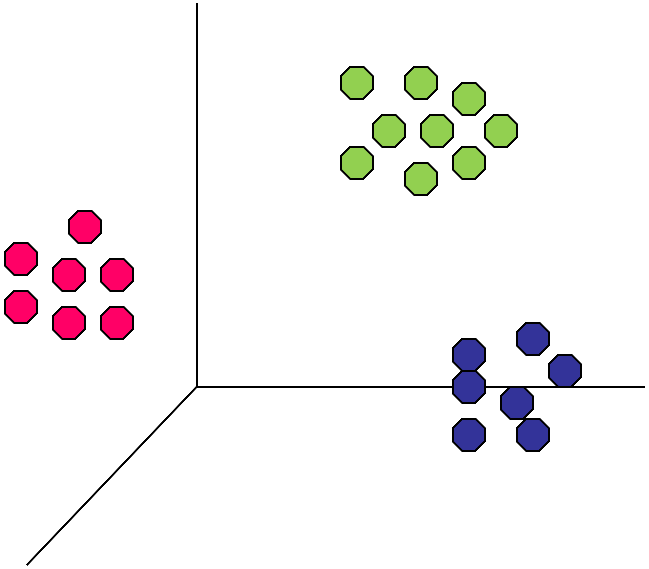


Illustrating Clustering

- Solution: Euclidean distance based Clustering in 3-D space.

Intraclustere distances are minimized

Intercluster distances are maximized



Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

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

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule
    - If a customer buys diaper and milk, then he is very likely to buy beer:

$$\boxed{Diapers \rightarrow Beer,\ \mathit{support} = 20\%,\ \mathit{confidence} = 85\%}$$

Sequential Pattern Discovery: Definition

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events:

  - In telecommunications alarm logs,
    - (Inverter_Problem Excessive_Line_Current)
      (Rectifier_Alarm) --> (Fire_Alarm)
  - In point-of-sale transaction sequences,
    - Computer Bookstore:
      (Intro_To_Visual_C) (C++_Primer) -->
      (Perl_for_dummies,Tcl_Tk)
    - Athletic Apparel Store:
      (Shoes) (Racket, Racketball) --> (Sports_Jacket)

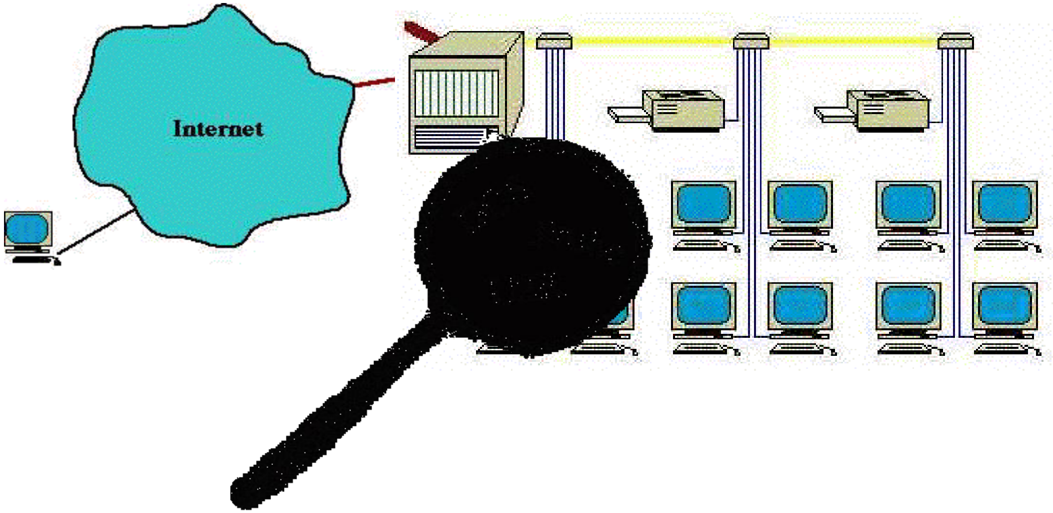
Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation/Anomaly Detection

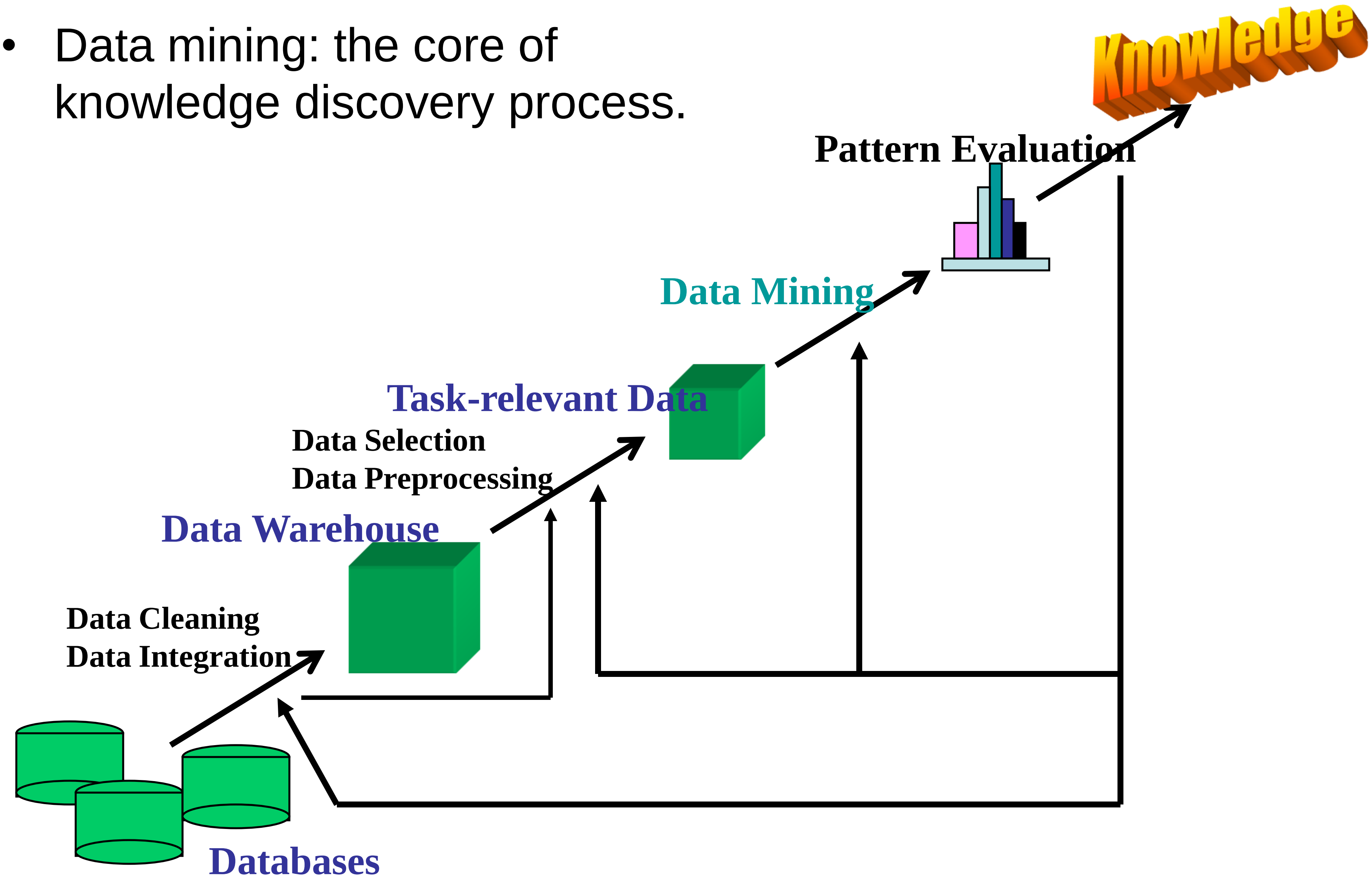
- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection



Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



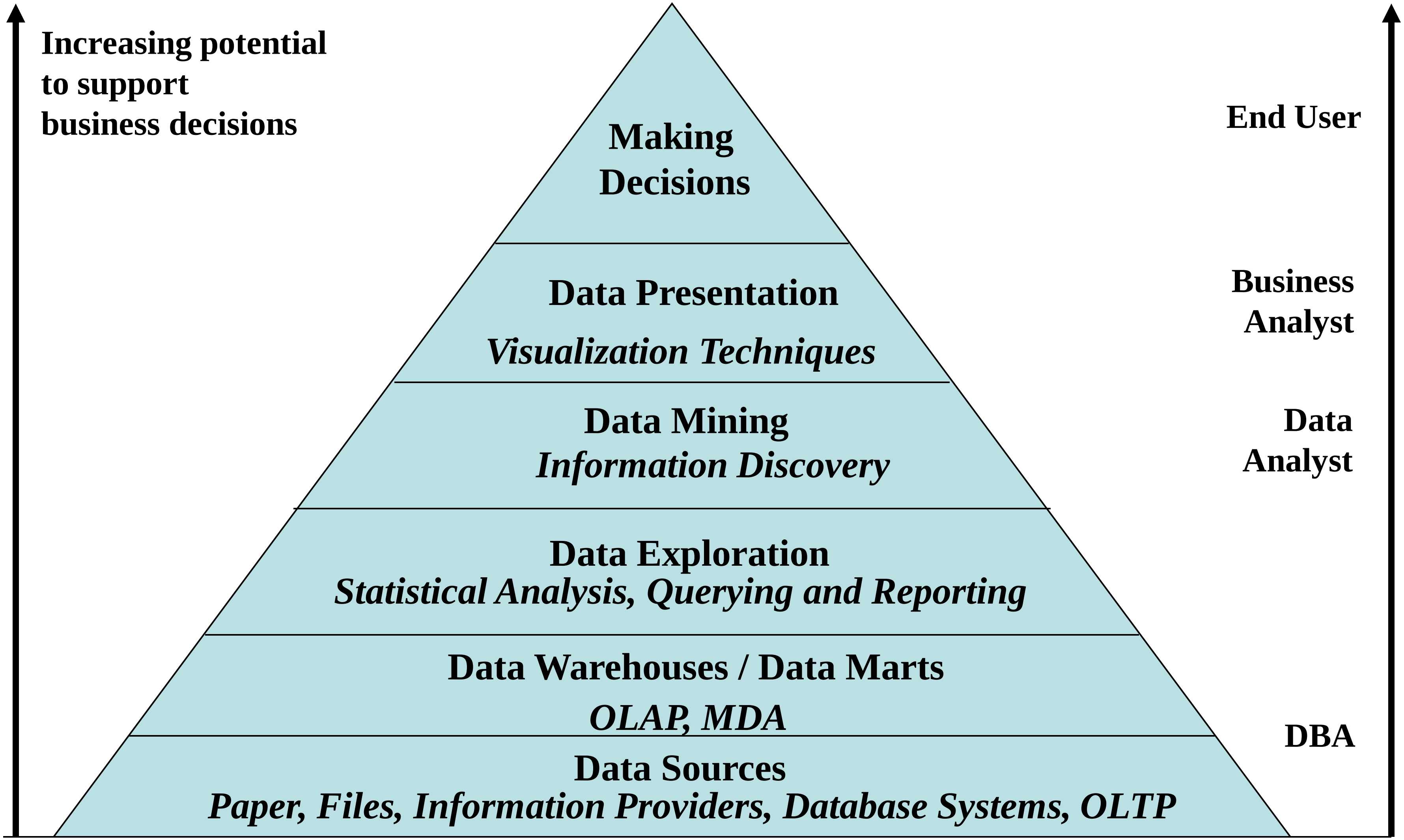
Steps of a KDD Process

1. data gathering
2. data cleansing
3. data transformation
4. selecting techniques
5. applying data mining
6. processing results

Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data gathering, data selection
- **Data cleaning** and preprocessing: (may take 60% of the effort!)
- **Data reduction and transformation**:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering, etc.
- Choosing the specific data mining algorithm(s)
- **Data mining**: search for patterns of interest, models, etc.
- **Pattern/model evaluation and knowledge presentation**
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Data Mining and Business Intelligence



Increasing potential
to support
business decisions

Making Decisions — End User

Data Presentation
Visualization Techniques — Business Analyst

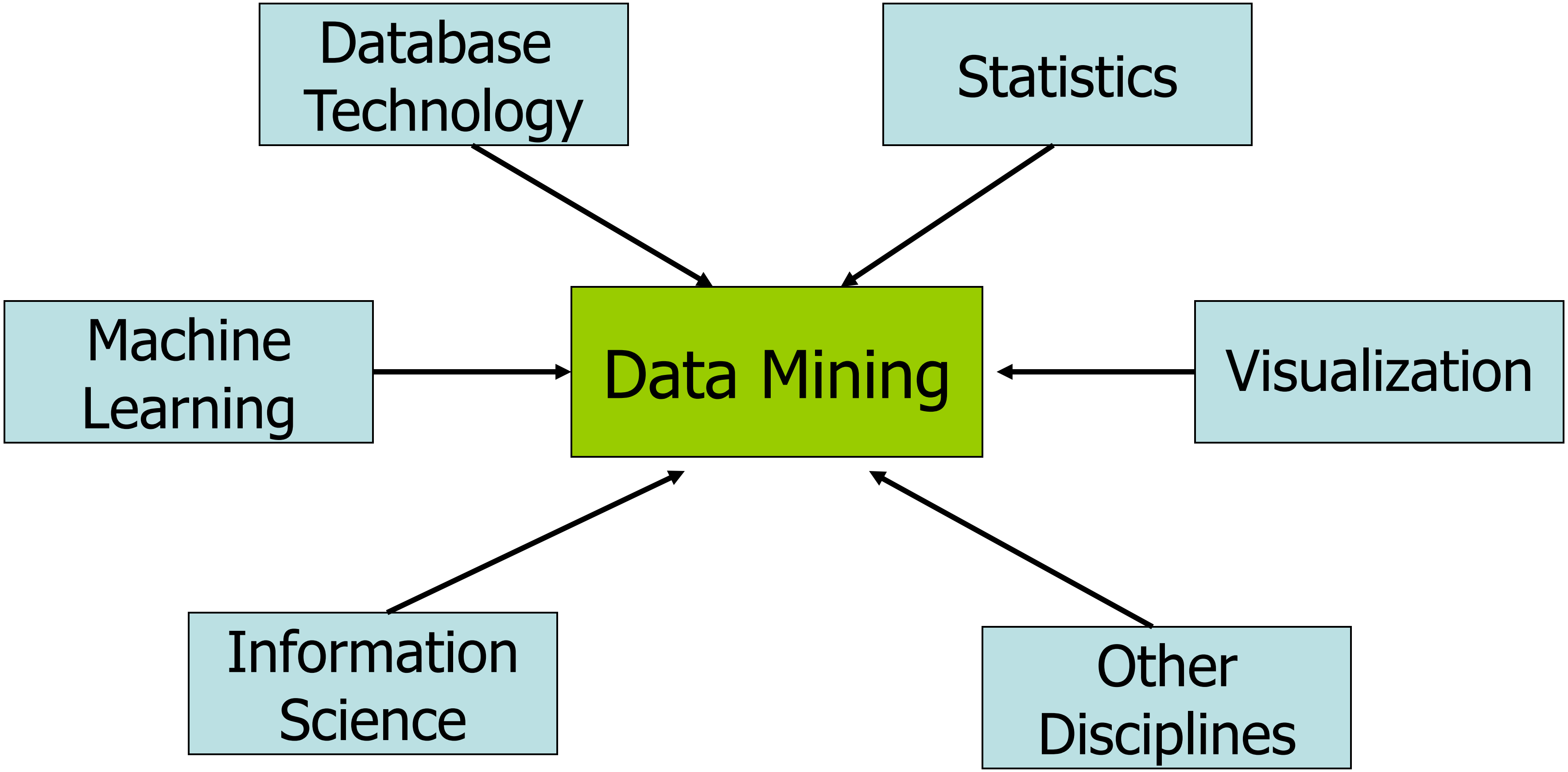
Data Mining
Information Discovery — Data Analyst

Data Exploration
Statistical Analysis, Querying and Reporting

Data Warehouses / Data Marts
OLAP, MDA — DBA

Data Sources
Paper, Files, Information Providers, Database Systems, OLTP

Data Mining: Confluence of Multiple Disciplines



The University of Reading

Input Data

Dr. Giuseppe Di Fatta

Associate Professor

Department of Computer Science

G.DiFatta@reading.ac.uk

What is “Data”?

Input data: a set of instances

- instances, aka: records, objects, points, cases, samples, entities, etc.
- Individual, independent examples of the concept to be learned.
- Single relation DB, flat file.
- A collection of data objects and their attributes

An **attribute** is a property or characteristic of an object

- Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic, or feature

Attributes

Objects

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Attribute Values

- Attribute **values** are numbers or symbols assigned to an attribute

- Distinction between **attributes** and **attribute values**
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet, meters, or categorical values (tall, medium, short)

  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of Attributes

There are different types of attributes:

- Nominal
  - Examples: ID numbers, eye color, zip codes
- Ordinal
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}, ranking in a marathon
- Interval
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit, level of happiness (e.g. rated from 1 to 10)
- Ratio
  - Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

The type of an attribute depends on which of the following properties it possesses:

Distinctness:	= ≠
Order:	< >
Addition:	+ -
Multiplication:	* /

Nominal attribute:	distinctness
Ordinal attribute:	distinctness & order
Interval attribute:	distinctness, order & addition
Ratio attribute:	all 4 properties

Attribute Type Description

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Value Transformation

Attribute Type	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: 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.

Types of Data Sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Multi-Relational
  - Star or snowflake schema
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data

Important Characteristics of Structured Data

- **Dimensionality**
  - Number of attributes each object is described with
  - Challenge: high dimensionality (curse of dimensionality)

- **Sparsity**
  - Sparse data: values of most attributes are zero
  - Challenge: sparse data call for special handling

- **Resolution**
  - Data properties often could be measured with different resolutions
  - Challenge: decide on the most appropriate resolution
    (e.g. “Can’t See the Forest for the Trees”)

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	**Projection of y load**	**Distance**	**Load**	**Thickness**
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a ‘term’ vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	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

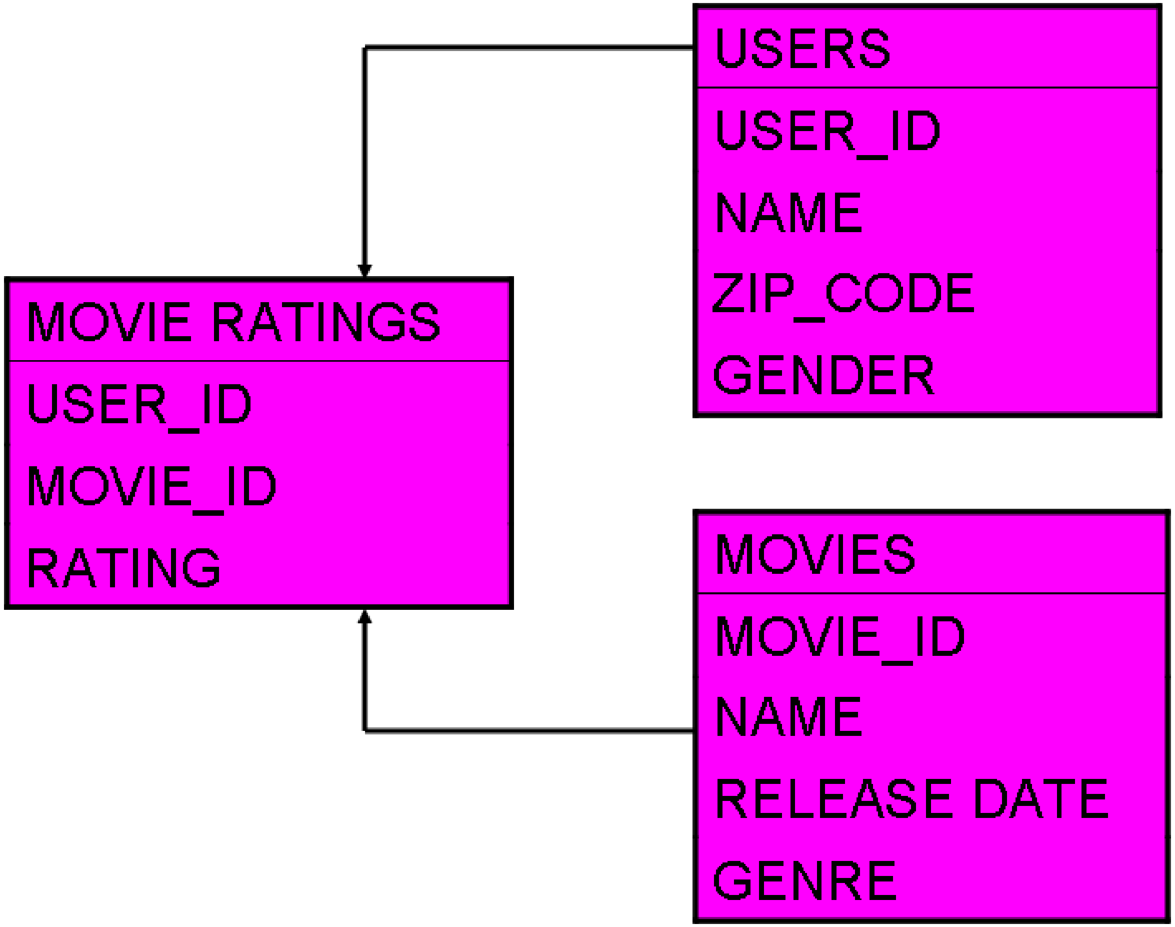
Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - E.g., consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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

Multi-Relational Data

- ❑ Attributes are objects themselves



USERS

USER_ID

NAME

ZIP_CODE

GENDER

MOVIE RATINGS

USER_ID

MOVIE_ID

RATING

MOVIES

MOVIE_ID

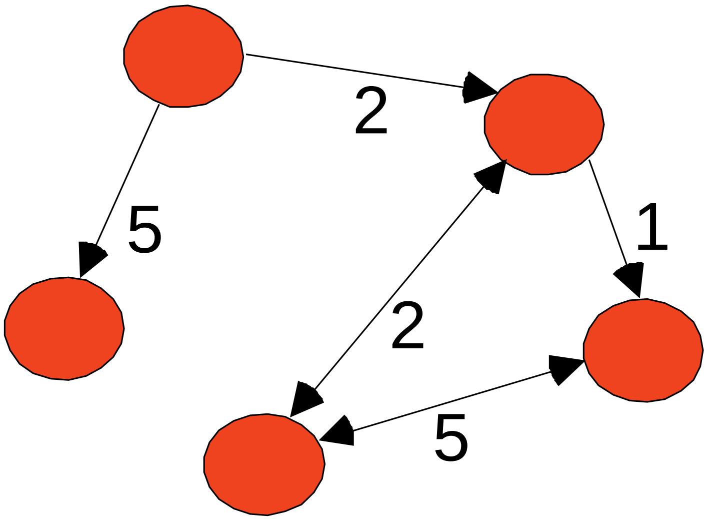
NAME

RELEASE DATE

GENRE

Graph Data

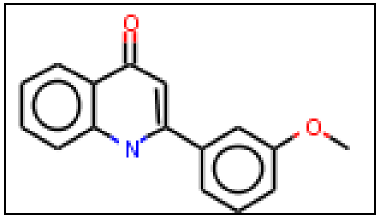
- Examples: Generic graph and HTML Links

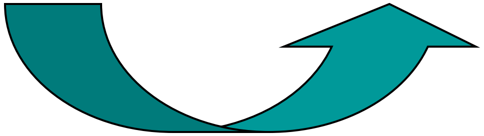


```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

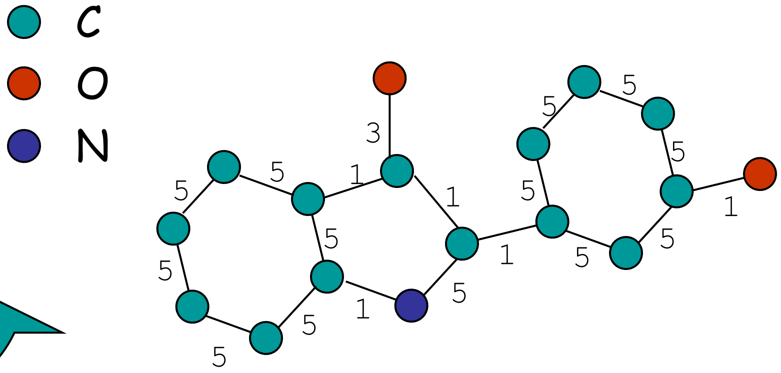
Chemical Data

Chemical compound





Graph



Graph representations:

- ➢ adjacency matrix

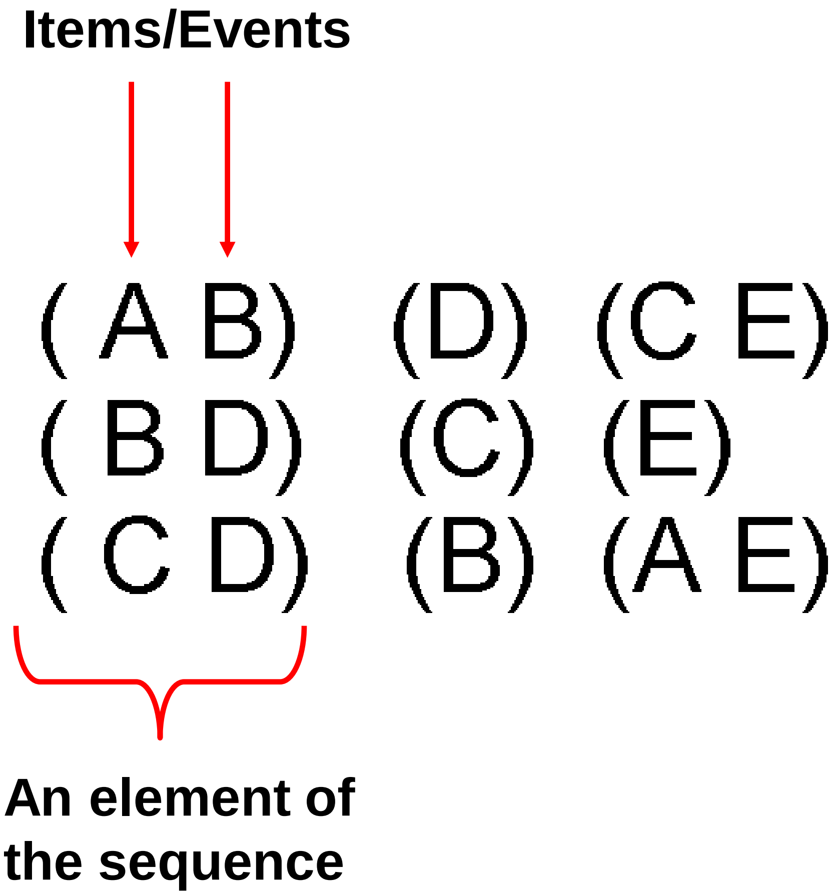
	0	1	2	3	4	5	6	7
0		1			1			1
1	1		3				3	
2		3		5				
3			5		5			
4	1			5		5		1
5					5		5	
6		3				5		1
7	1				1		1	

- ➢ edges list
  - – Source atom type
  - – Source atom
  - – Bond type
  - – Dest. Atom type
  - – Dest. atom

```
(C,0,1,C,1)
(C,0,5,N,4)
(C,0,1,O,7)
...
```

Ordered Data

- ❑ Sequences of transactions



Items/Events

(A B) (D) (C E)

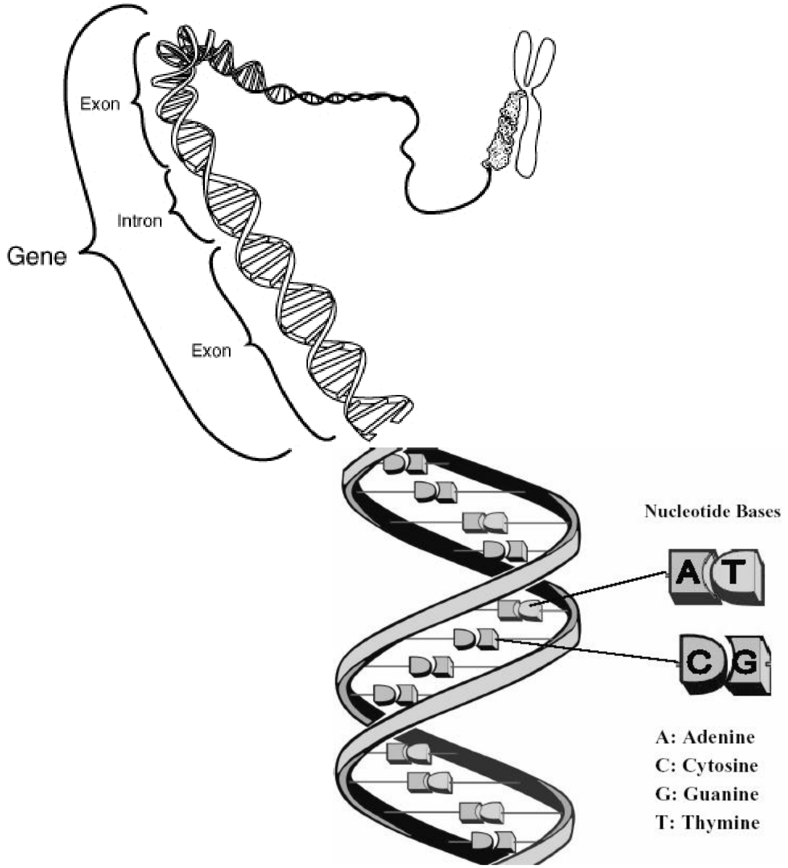
(B D) (C) (E)

(C D) (B) (A E)

An element of the sequence

Ordered Data

- Genomic sequence data

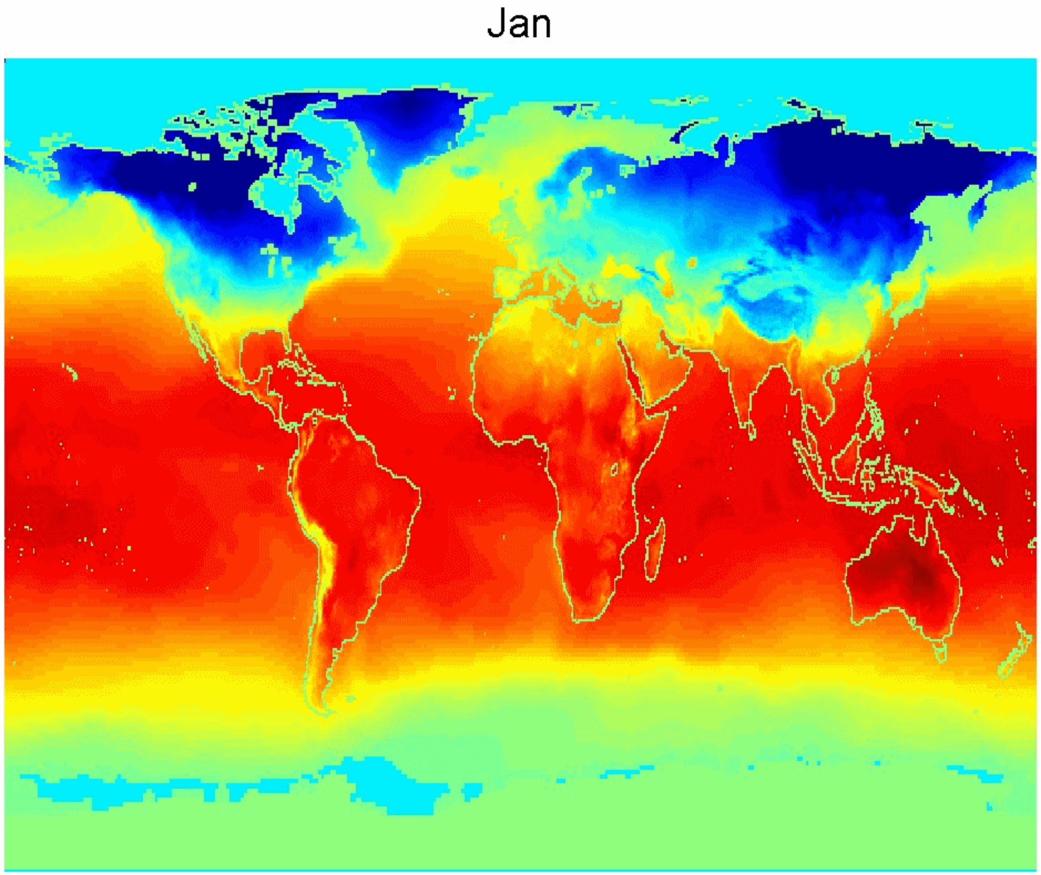


GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Ordered Data

- Spatial-Temporal Data

Average Monthly Temperature of land and ocean



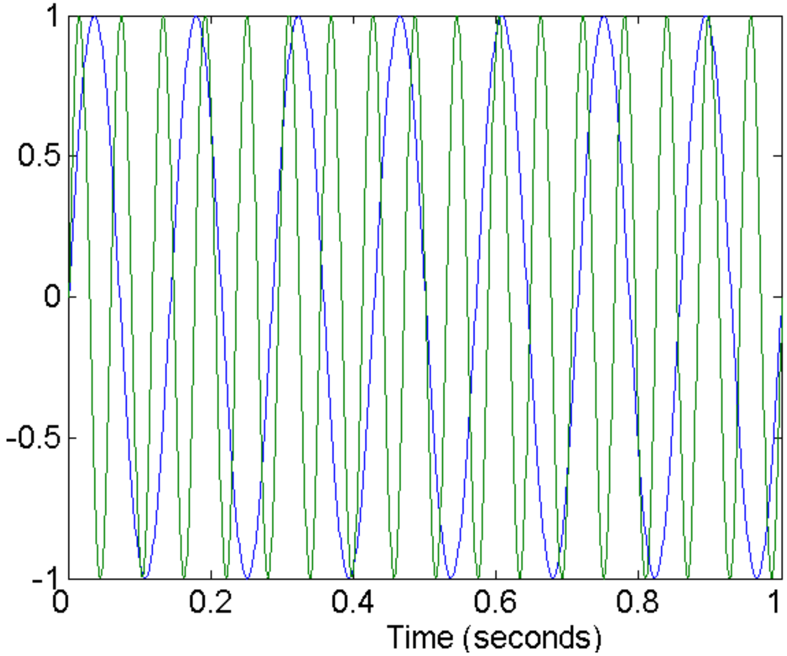
Data Quality

- ✓ What kinds of data-quality problems?
- ✓ How can we detect problems with the data?
- ✓ What can we do about these problems?

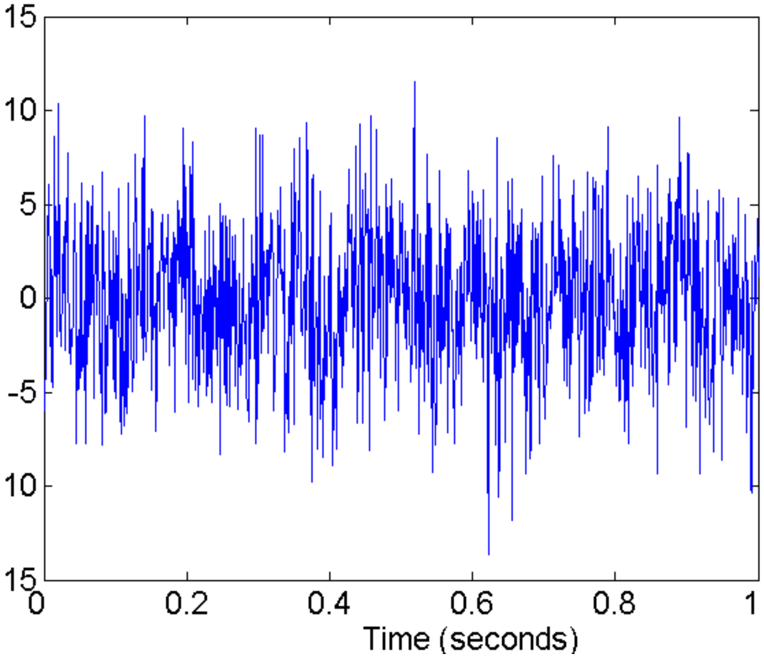
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

Noise

- Noise refers to modification of original values
  - Examples: distortion of a person’s voice when talking on a poor phone and “snow” on television screen



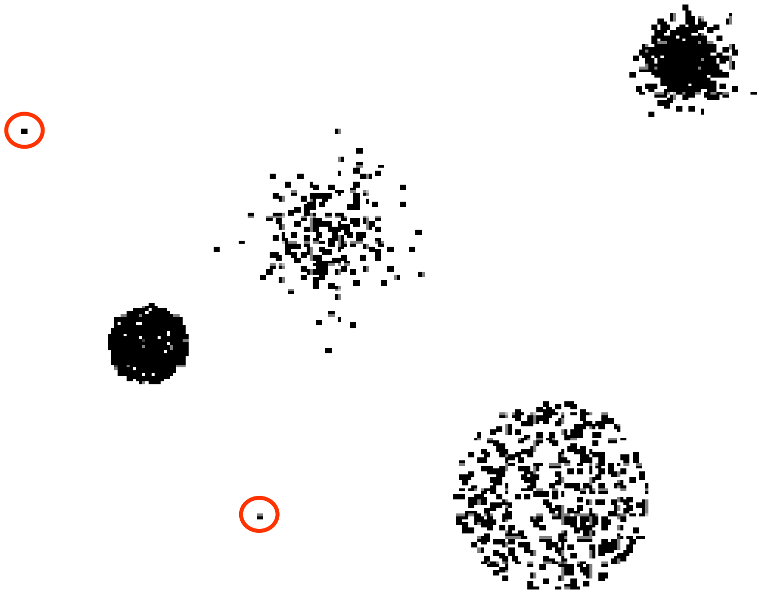
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources

- Examples:
  - Same person with multiple email addresses

- Data cleaning
  - Process of dealing with duplicate data issues

The University of Reading

Data Preprocessing

Dr. Giuseppe Di Fatta
Associate Professor
Department of Computer Science
G.DiFatta@reading.ac.uk

Data Preprocessing:

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

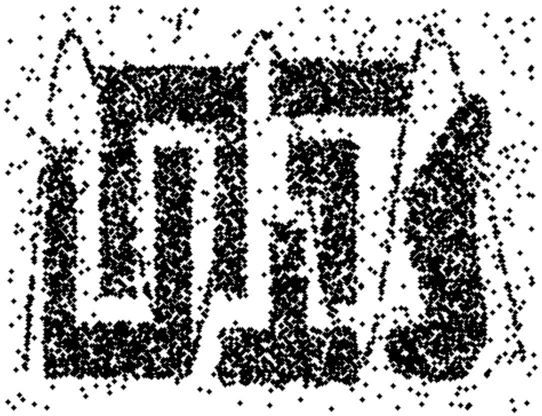
Sampling

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative.
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

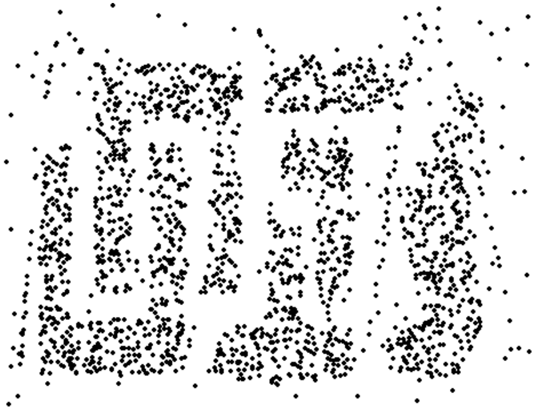
Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - As each item is selected, it is removed from the population
  - Each outcome depends on all previous outcomes
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
    - One outcome does not affect the other outcomes
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

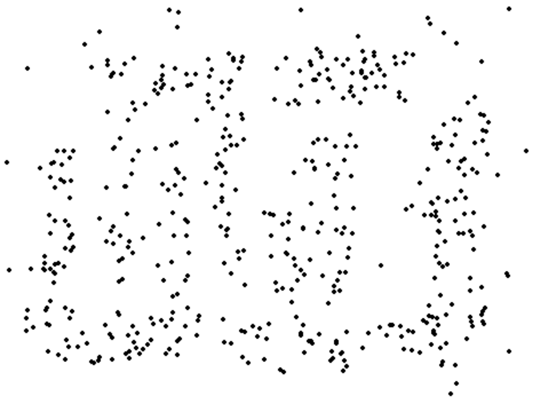
Sample Size



8000 points



2000 Points



500 Points

Curse of Dimensionality

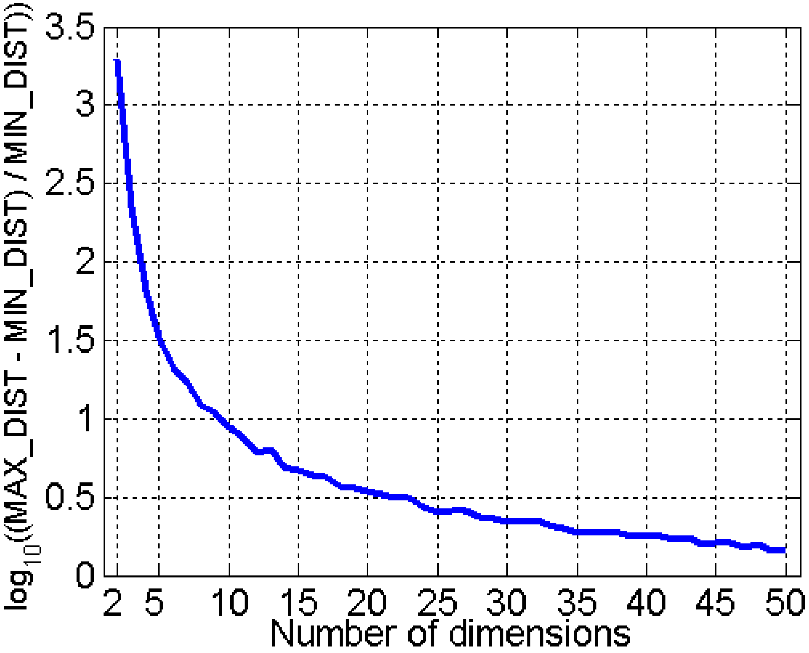
- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.

- Definitions of *density* and *distance* between points, which is critical for clustering and outlier detection, become less meaningful.

Example

- **Randomly generate 500 points in** $\Re^n$
- **Compute difference between max and min distance between any pair of points**

$$f(n) = \log_{10}\left(\frac{\max\left(dist\left(v_i, v_j\right)\right) - \min\left(dist\left(v_i, v_j\right)\right)}{\min\left(dist\left(v_i, v_j\right)\right)}\right)$$



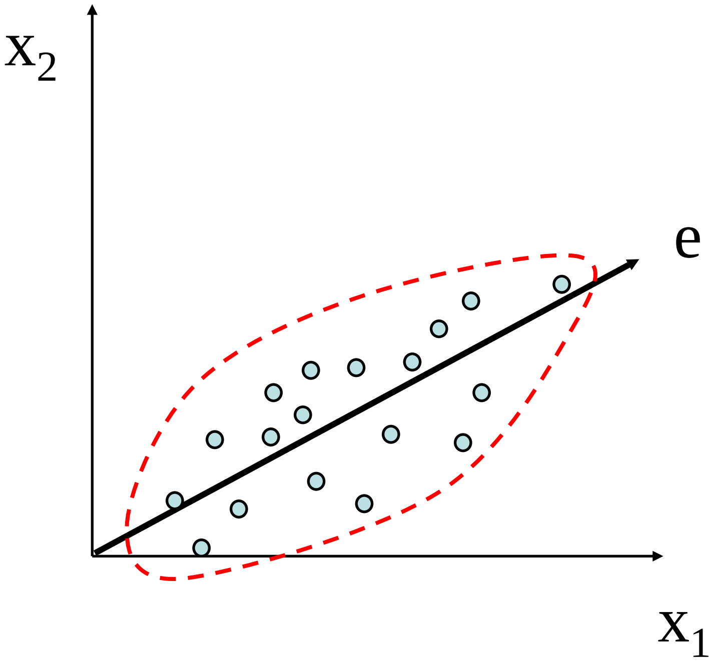
Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Dimensionality Reduction: PCA

- Goal is to find a projection that captures the largest amount of variation in data



Feature Subset Selection

- Another way to reduce dimensionality of data

- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid

- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' Grade Point Average (GPA)

Feature Subset Selection

- Techniques:
  - Brute-force approach:
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches:
    - Features are selected before data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes

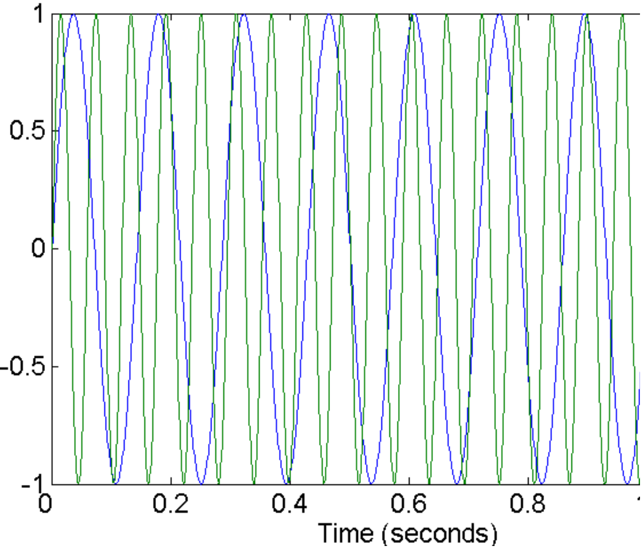
Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

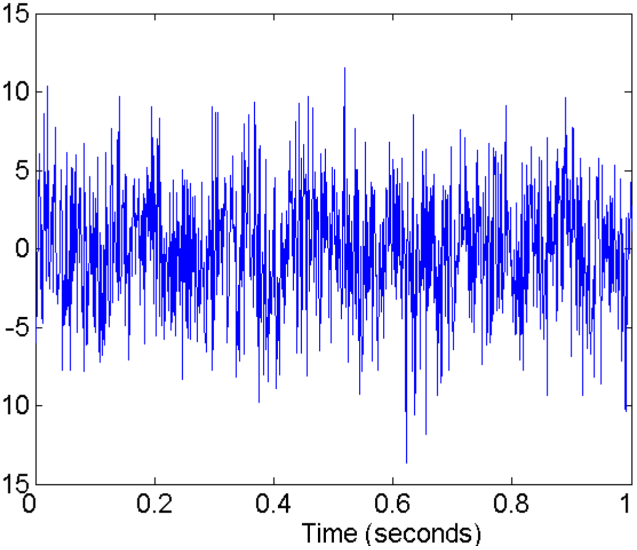
- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features

Example: Mapping Data to a New Space

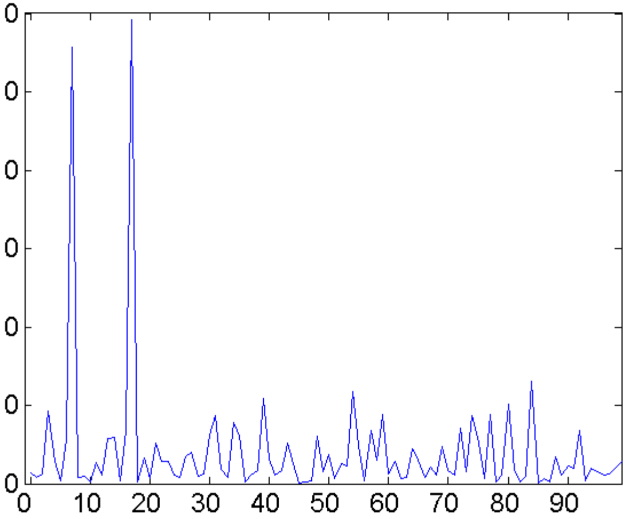
- **Fourier transform**
- **Wavelet transform**



Two Sine Waves



Two Sine Waves + Noise



Frequency

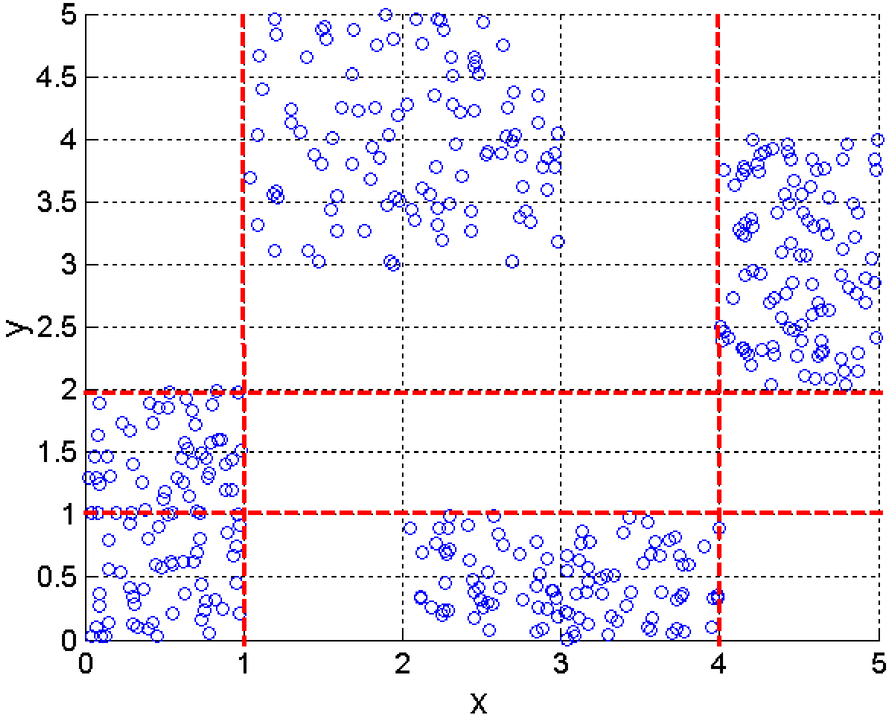
Discretization and Binarization

- Different data mining applications require specific data formats
  - Categorical only (discretization)
  - Binary only (binarization)
  - Interval/Ratio only (binarization)

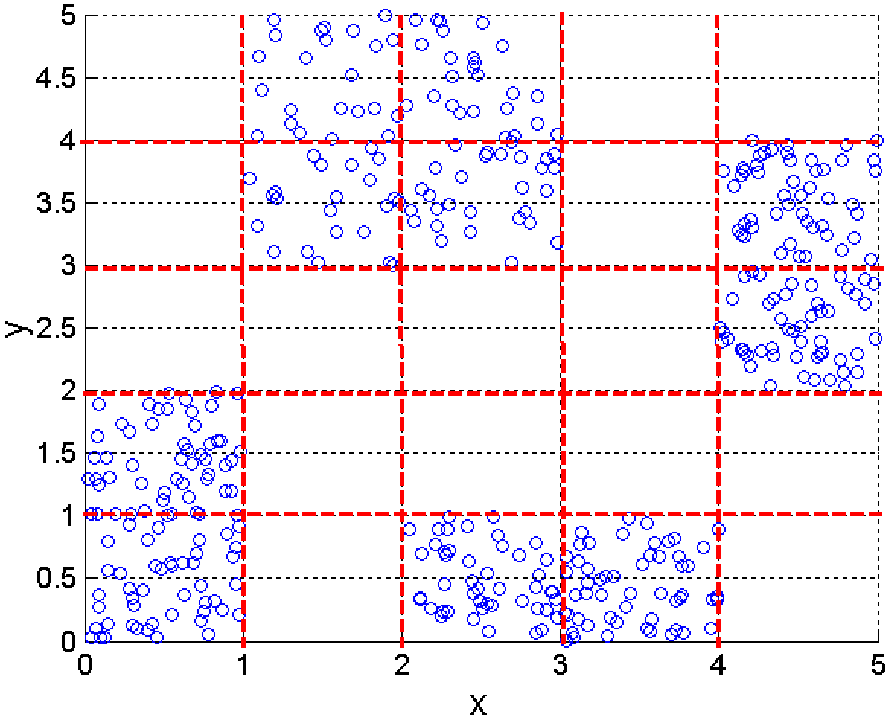
- Discretization: transforming interval attribute into categorical

- Binarization: transforming non-binary attribute into a set of binary attributes

Discretization Using Class Labels



3 categories for both x and y



5 categories for both x and y

Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.
  - E.g., simple functions: x^k, log(x), e^x, |x|
  - Normalization and Standardization
    - *Normalization* is the transformation of the variable vectors into vectors of unit length.
    - *Standardization* transforms the variable vector into a vector of unit length, with a mean of zero, and a standard deviation of one.

$$x^{'} = \frac{(x - \mu_x)}{\sigma_x}$$