TEMA 2a

Caracterització i preparació de les dades

Índex

- Representació de les dades
- Representació tabular atribut-valor
- Tipus d'atributs
- Casos no evidents de representació tabular:
  - Atributs relacionals
  - Grafs
  - Series temporals
  - Imatges
- Representació esparsa

Representació de les dades: <atribut-valor>

- Conjunts de dades estan constituïts per objectes (entitats)
- Objectes poden ser transaccions comercials, individus, textos, pagines web, etc.
- Representació més habitual dels objectes és com a parelles <atribut-valor>
- Altres noms: *examples, instances, data points, objects*
- Els individus es defineixen segons un conjunt de característiques (atributs) escollits
- L'aspecte resultant és el de una taula

Representació de les dades: <atribut-valor>

- Exemple:

Acme Investors Incorporated

Customer ID	Account Type	Margin Account	Transaction Method	Trades/ Month	Sex	Age	Favorite Recreation	Annual Income
1005	Joint	No	Online	12.5	F	30–39	Tennis	40–59K
1013	Custodial	No	Broker	0.5	F	50–59	Skiing	80–99K
1245	Joint	No	Online	3.6	M	20–29	Golf	20–39K
2110	Individual	Yes	Broker	22.3	M	30–39	Fishing	40–59K
1001	Individual	Yes	Online	5.0	M	40–49	Golf	60–79K

Representació de les dades: <atribut-valor>

- Aquesta representació ens permet imaginar els objectes descrits a traves de *n* atributs com a punts en un espai *n-dimensional*
- El concepte de representació n-dimensional ens permetrà raonar de forma més intuïtiva sobre els algorismes de Data Mining

Representation of Raw Data

Common data types:

- **Numeric**
- **Categorical**

1. Numeric

A feature with numeric values has two important properties:

a) Order relation (for example, 2 < 5 and 5 < 7) ,

b) Distance relation (for example, d(2.3, 4.2) = 1.9)

Representation of Raw Data

2. Categorical

- Categorical (often called symbolic) values have neither of two relations (order or distance)
- Boolean data is an example of categorical data.
- The two values of a categorical variable can be either equal or not equal: they only support equality relation (Blue = Blue, or Red ≠ Black)
- Coded categorical variables are known as "dummy variables" in statistics.

Feature value	*Code*
Black	1000
Blue	0100
Green	0010
Brown	0001

Black	1
Blue	2
Green	3
Brown	4

Representation of Raw Data

Another dimension of (numerical) **data types:**

continuous vs. Discrete

1. **Continuous**
   (also known as quantitative or metric)

   These variables are measured using either:

   a) **interval scale:**
   The zero point in the interval scale is placed arbitrarily.
   (Temperatures: 40^{o} and 80^{o})

   b) **ratio scale:**
   It has an absolute zero point and the ratio relation holds.
   (Lengths: 2 ft. and 4 ft.)

Representation of Raw Data

2. Discrete

(also called qualitative)

They use one of two kinds of non-metric scales:

a) nominal scale

A nominal scale is an order-less scale.
(A, B, and C values for the variable, or ZIP-code)

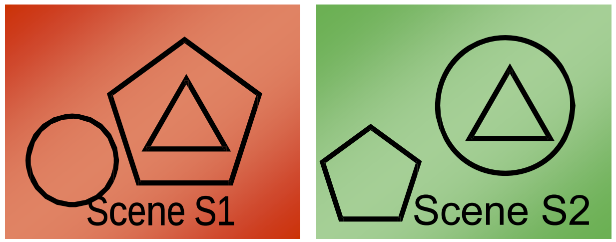
b) ordinal scale

An ordinal scale consists of ordered discrete gradations, e.g. rankings. An order relation is defined but no distance relation. (gold, silver, and bronze medal, or students ranked as 15th and 16th)

---

* c) **Periodic variable is a feature for which the distance relation exists, but there is no order relation. (Days of the week, month, or year) .**

Building Mineable Data Sets: Data Representation = Table



Single table representation

SCENE				
SceneID	Triangle	Square	Circle	Pentagon
S1	+	-	+	+
S2	+	-	+	+

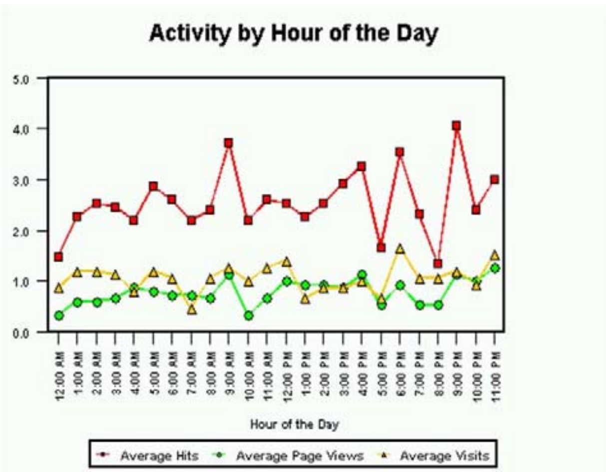
Relational representation

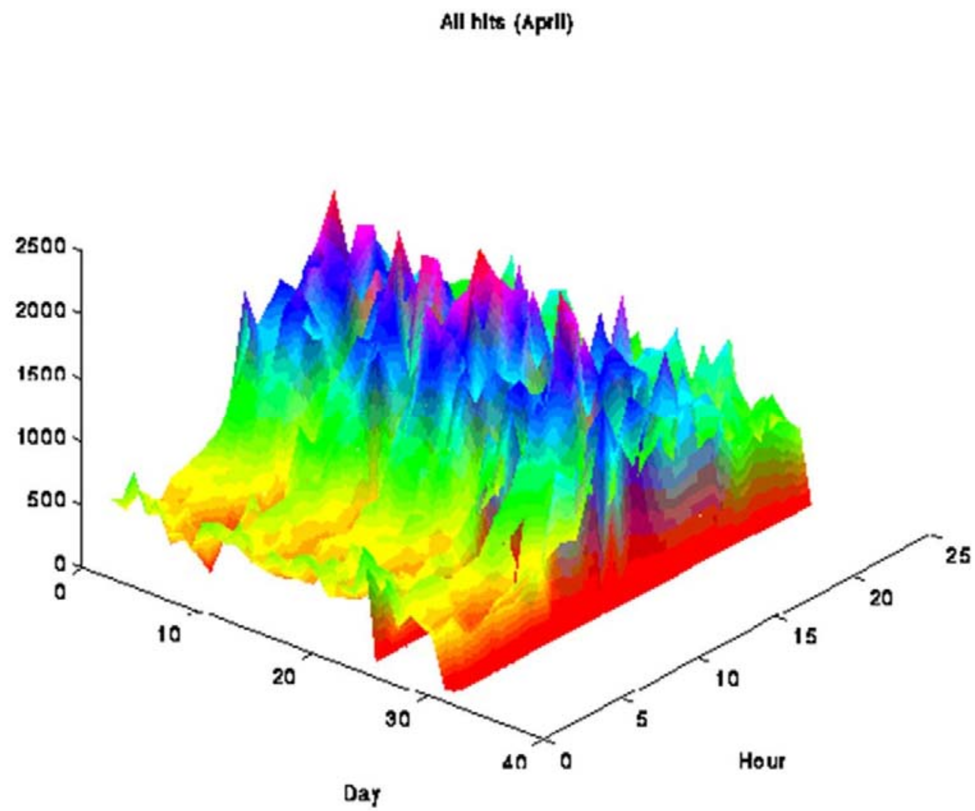
SCENE		
SceneID	ObjectID	Shape
S1	O1	Triangle
S1	O2	Circle
S1	O3	Pentagon
S2	O1	
S2	O2	
S2	O3	

INSIDE		
SceneID	ObjectID	ObjectID
S1	O1	O3
S2	O1	O2

Web Log Data over Time -> Table?

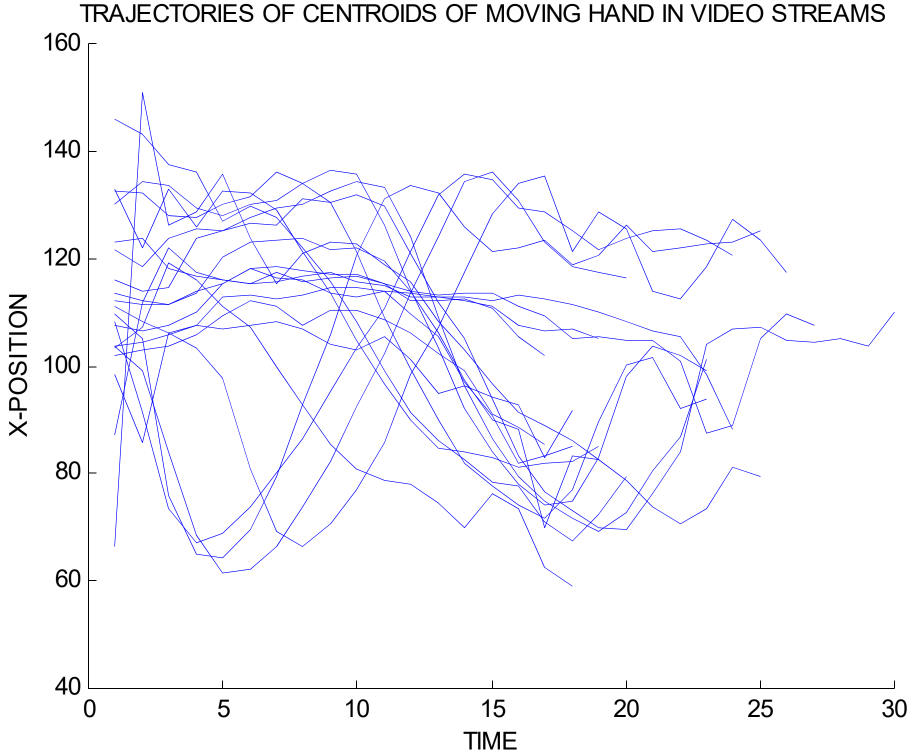
Day	Hour	# of hits
06/06/05	5 a.m.	58
06/07/05	6 a.m.	83
…	…	…





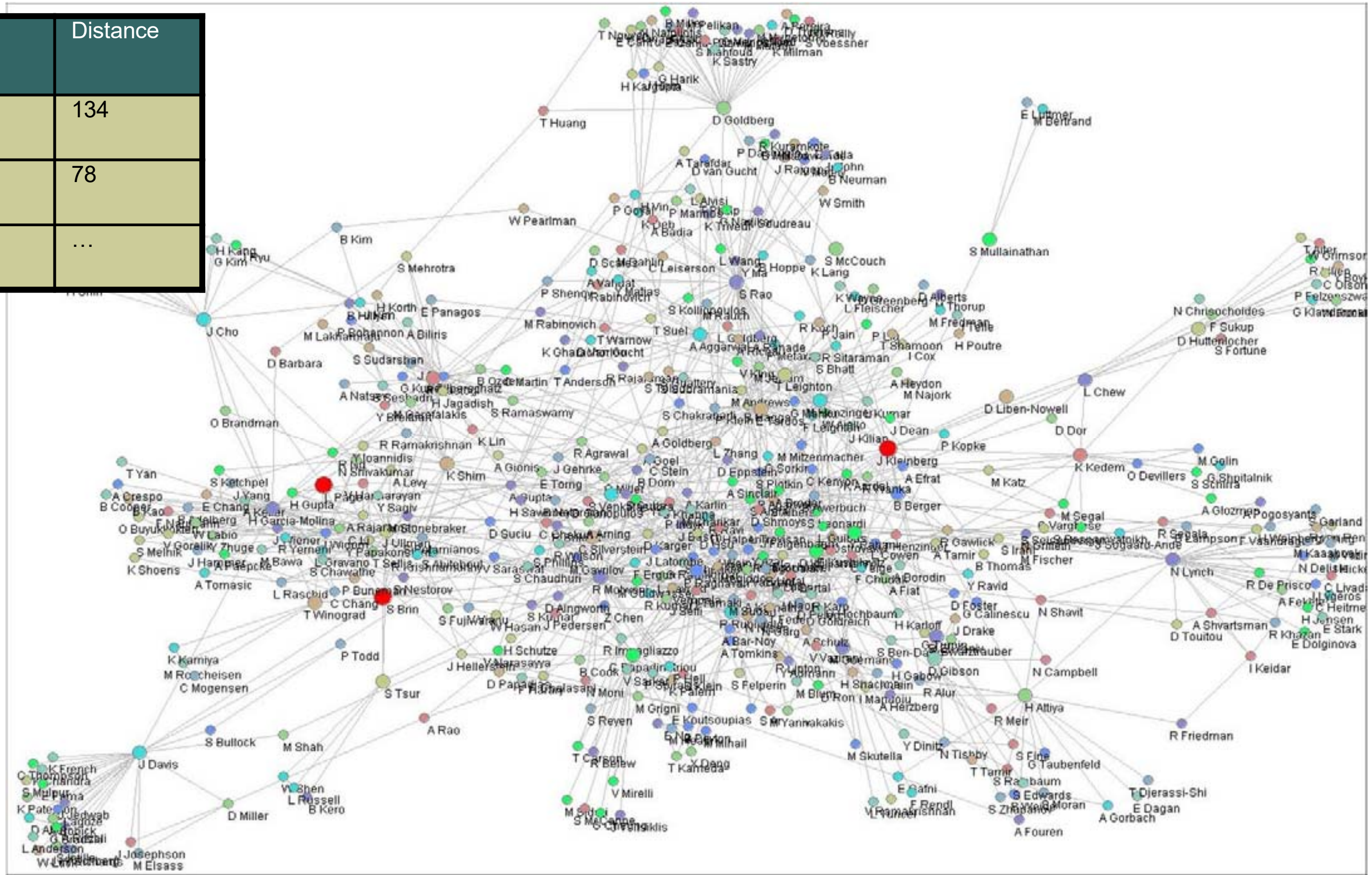
Time Series Data -> Table?

Time	TS1	TS 2		TS n
1	86	74	…	140
2	99	133	…	91
…	…	…	…	…



Relational Data = Graph -> Table ?

Beginning node	Ending node	Distance
Miller	Todd	134
Mile	Rao	78
…	…	…



Sparse representation

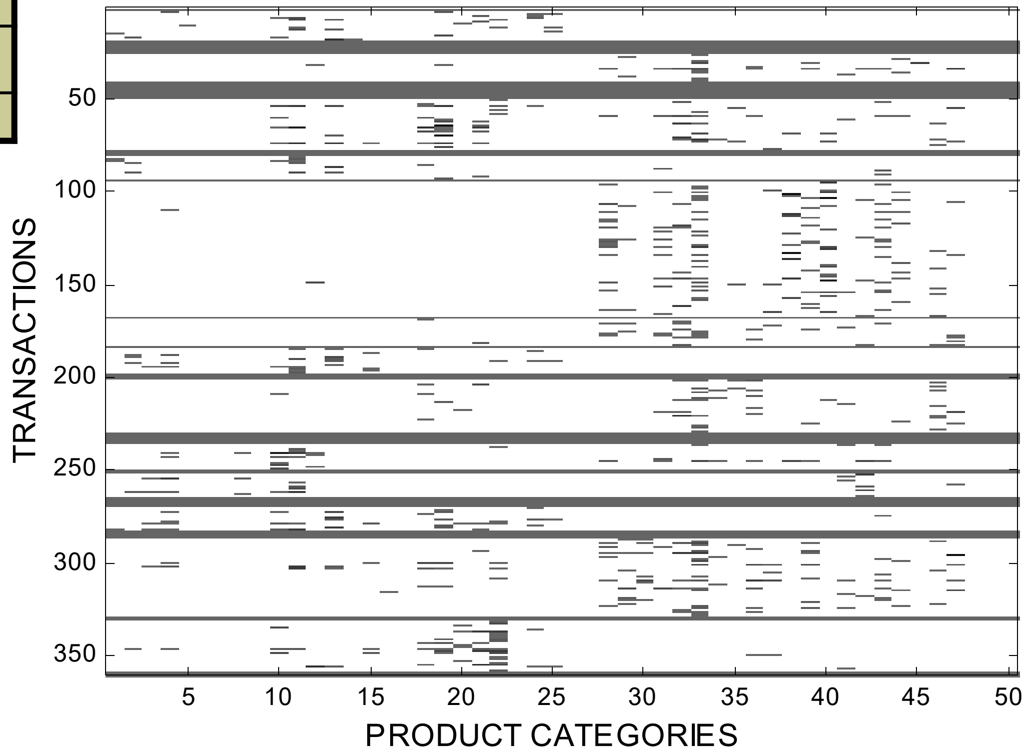
TID	banana	cheese	floppy	pizza	wine
100	yes	no	yes	yes	no
200	no	yes	yes	no	yes
300	yes	yes	yes	no	yes
400	no	yes	no	no	yes

Sparsity:
Eliminate no's
------------------7

TID	items
100	{banana, floppy, pizza}
200	{cheese, floppy, wine}
300	{banana, cheese, floppy, wine}
400	{cheese, wine}

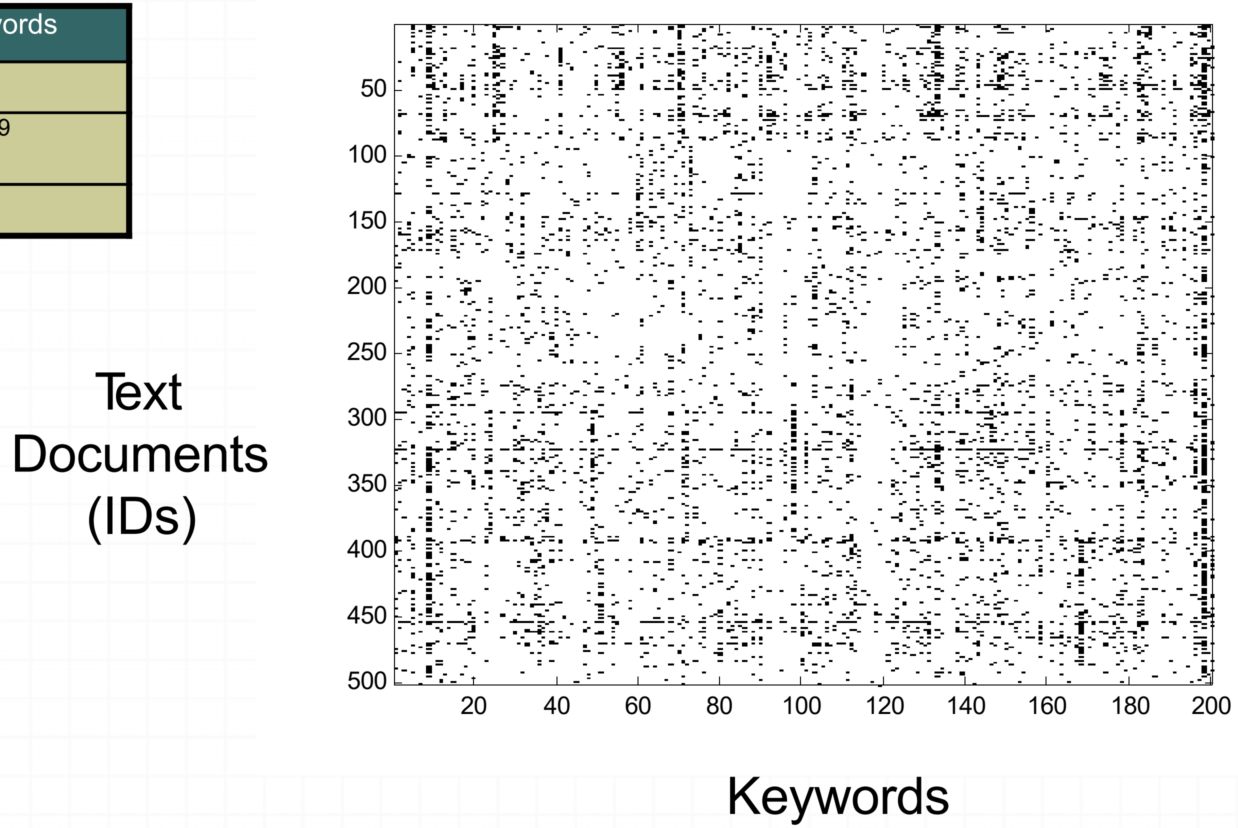
Sparse representation

Trans. ID	Products
01	01, 03, 44, 76
02	22, 37, 76
…	…



Text Representation = Table

Text ID	Keywords
001	56, 34, 79
002	07, 122, 189
…	…



TEMA 2b

Caracterització i preparació de les dades

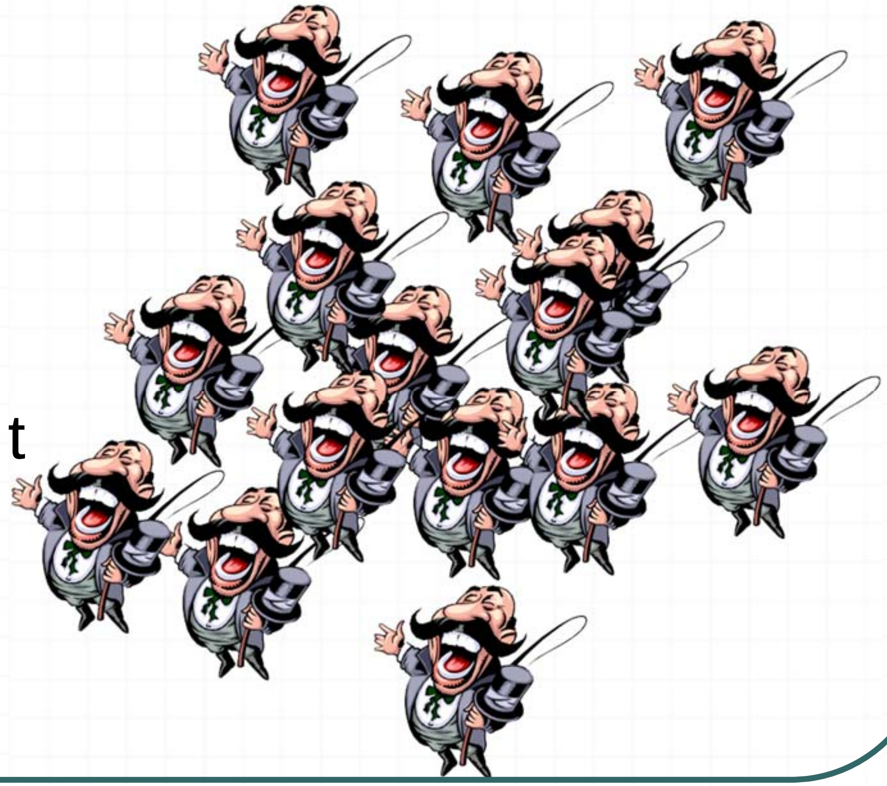
Índex

- Processament de les dades
- Neteja de les dades:
  - Dades perdudes. Solucions
  - Normalització
  - Dades temporals
  - Eliminació del soroll
- Reducció de les dades:
  - Eliminació de *outliers*
  - **Selecció de característiques**
  - **Formació de característiques**

Preparing the Data for Data Mining

Let the data speak...

The data may have quite a lot to say..... but it may just be **noise**!



Preparing the Data for Data Mining

Two central tasks for the preparation of data:

1. To organize data into a standard form:
typically a standard form is a **relational table** (or tables).

2. To prepare data sets by:
- preprocessing and
- dimensionality reduction

that will lead to the best data mining performances.

Preparing the Data for Data Mining

Two central tasks for the preparation of data:

1. To organize data into a standard form (typically a standard form is a relational table).
2. To prepare data sets by **preprocessing** and dimensionality reduction that will lead to the best data mining performances.

Raw Data = Messy Data

* Missing data,
* Misrecorded data,
* Data may be from the other population (heterogeneous),
* Different structures & formats,
* With or without compression,
* Redundant,
* With implicit temporal & spatial components, …

Characteristics Of Raw Data -> Require Preprocessing

Missing Data

Replacement solutions:

1.) Manually examine samples with missing data values.

2.) Automatic replacement:

- Replace all missing values with a single global constant (selection of a global constant is highly application dependent).
- Replace a missing value with its feature mean.
- Replace a missing value with its feature mean for the given class (only for classification problems).

Missing Data

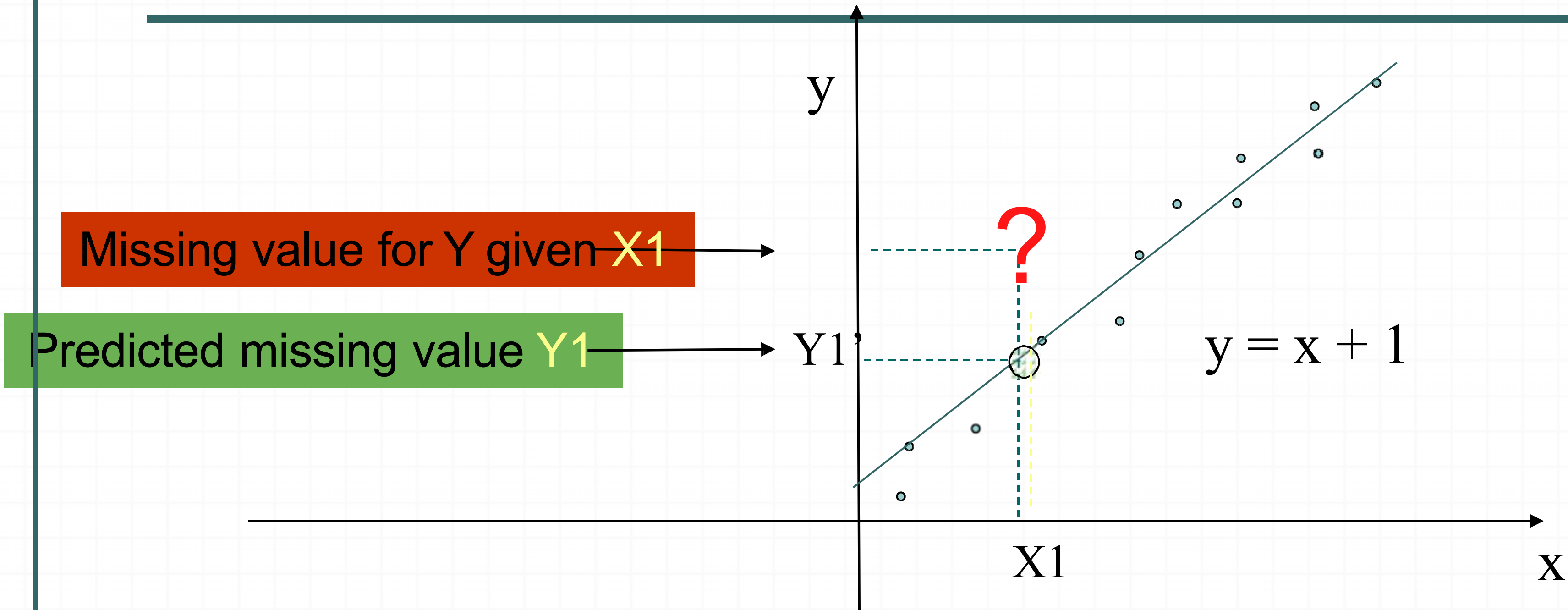
3. One possible interpretation of missing values is that they are ***"don't care" values***:

X = {1, ?, 3} → for the second feature the domain is [0, 1, 2, 3, 4]:

X1 = {1,0,3}, X2 = {1,1,3}, X3 = {1,2,3}, X4 = {1,3,3}, X5 = {1,4,3}

4. *Data miner* can generate model of **correlation between features**. Different techniques may be used such as regression, Bayesian formalism, clustering, or decision tree induction.

Missing Data Replacement (Using Regression Analysis)



- In general, replacement of missing values is speculative and often misleading to replace missing values using a simple, artificial schema of data preparation.

- It is best to generate multiple solutions of data mining with and without features that have missing values, and then make comparison, analysis and interpretation.

Data Preprocessing: Transformation Of Raw Data

1. Normalizations

a) Decimal scaling :

$v'(i) = v(i)/ 10^k$

for the smallest k such that max (| v'(i)|) < 1.

b) Min-max normalization:

$v'(i) = (v(i) - \min(v(i))) / (\max(v(i)) - \min(v(i)))$

for normalized interval [0,1].

c) Standard deviation normalization:

$v'(i) = (v(i) - \text{mean}(v)) / \text{sd}(v)$

Data Preprocessing: Transformation Of Raw Data

2. Data smoothing:

F = {0.93, 1.01, 1.001, 3.02, 2.99, 5.03, 5.01, 4.98},

F smoothed = {1.0, 1.0, 1.0, 3.0, 3.0, 5.0, 5.0, 5.0}.

3. Differences and ratios:

s(t+1)-s(t) s(t+1)/s(t)

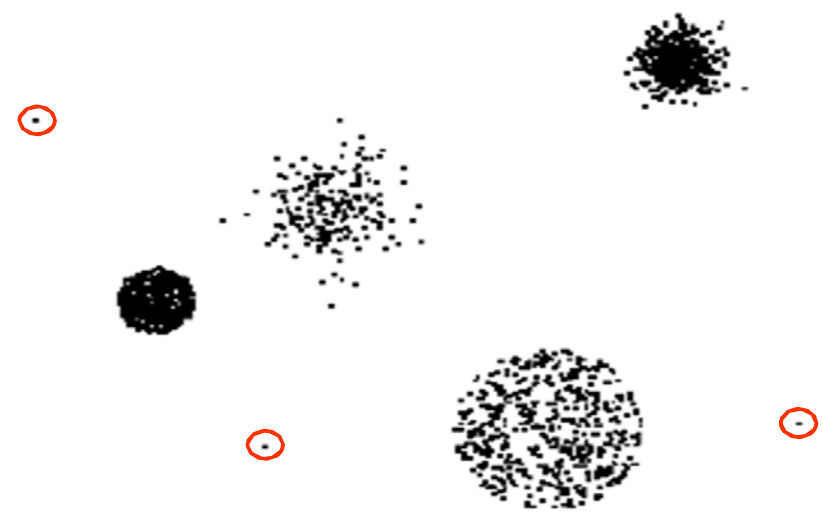
4. Composing new features:

For example: Body mass index **BMI= k F(Weight, Hight)**

Tractament d’outliers

Data Preprocessing: Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.
- Detection + correction/removal?



Anomaly/Outlier Detection

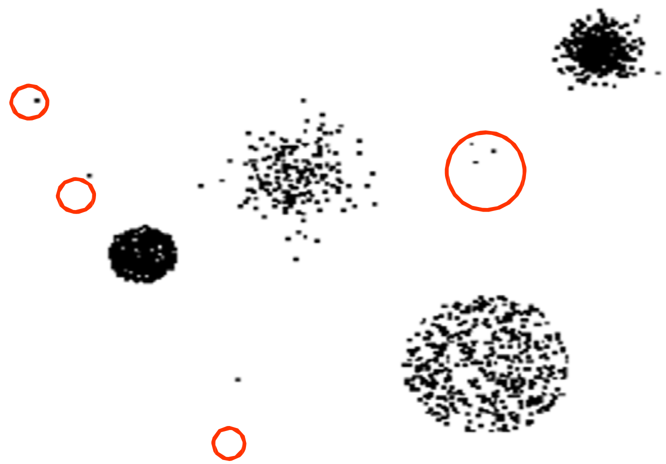
- ## Working assumption
  - There are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data

- ## Challenges
  - How many outliers are there in the data?
  - Finding needle in a haystack

Outlier Detection Schemes

- **General Steps**
  - **Build a profile of the "normal" behavior**
    - Profile can be patterns or summary statistics for the overall population
  - **Use the "normal" profile to detect outliers**
    - Outliers are observations whose characteristics differ significantly from the normal profile

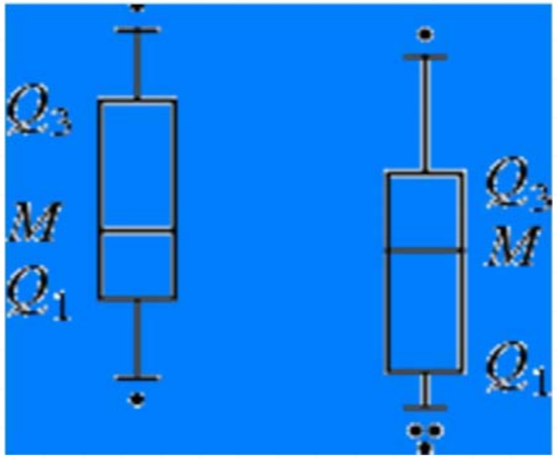
- **Types of outliers detection schemes**
  1. **Graphical**
  2. **Statistical-based**
  3. **Distance-based**
  4. **Model-based**



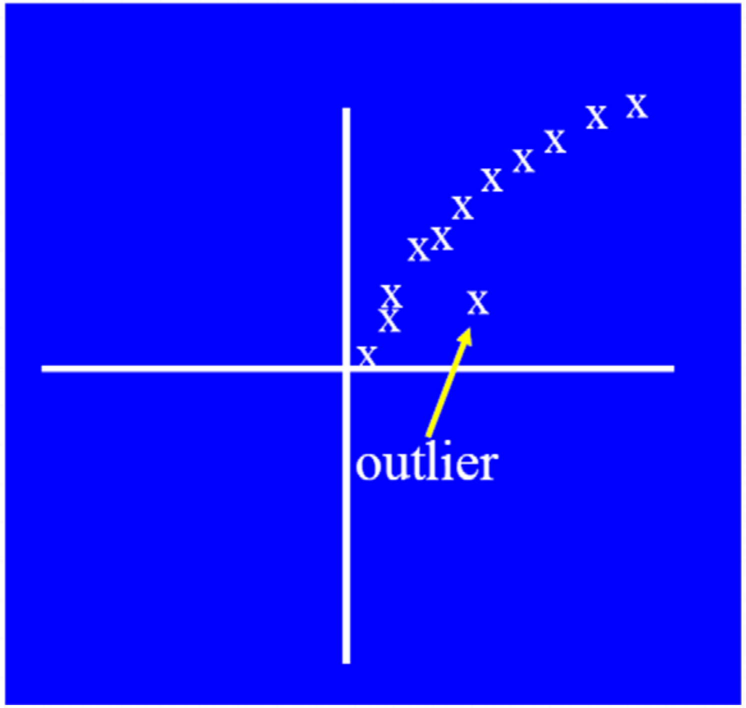
Outliers: Graphical Approaches

- Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D),…

- **Limitations**
  - Time consuming
  - Subjective



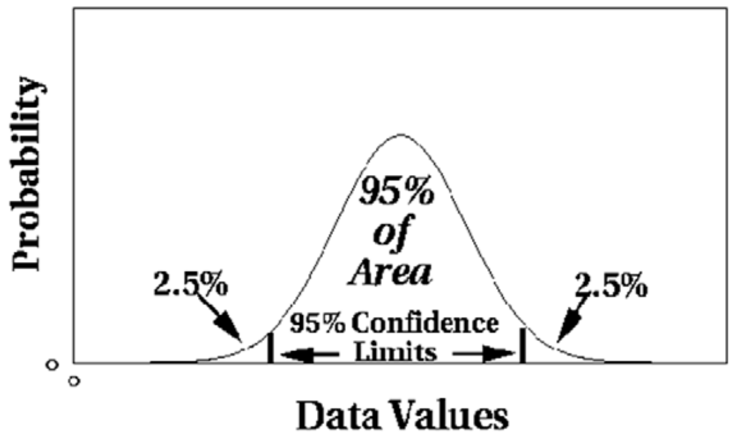
1D



2D

Outliers: Statistical Approaches

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on:
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)



Outliers: Statistical Approaches

EXAMPLE: Outlier detection for one-dimensional samples

Age = {3,56,23,39,156,52,41,22,9,28,139,31,55,20, -67,37,11,55,45,37}

Statistical parameters are:

Mean = 39.9
Standard deviation = 45.65

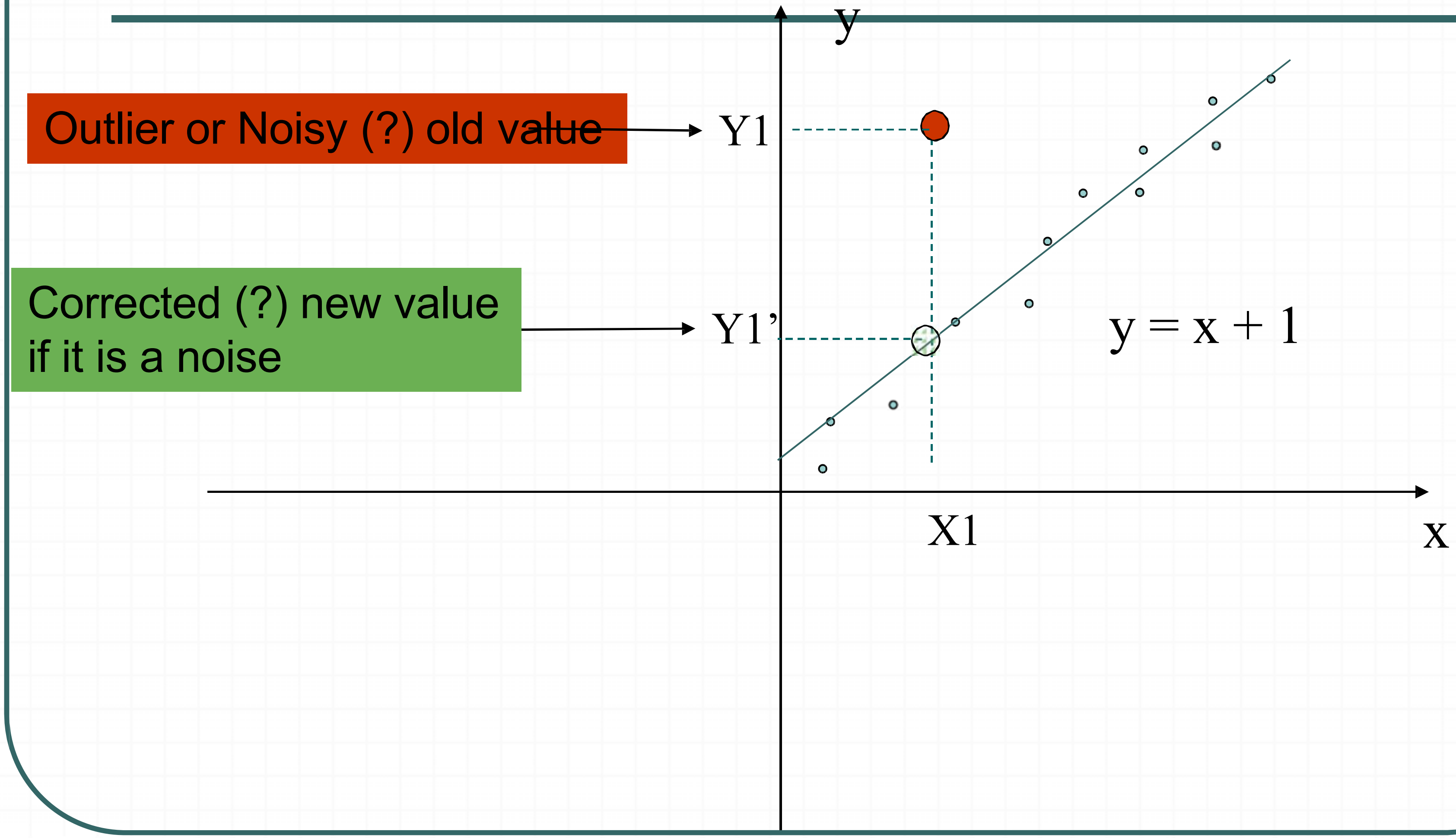
If we select that the threshold value for normal distribution of data is

Threshold = Mean ± 2 × Standard deviation

then all data out of range [-54.1, 131.2] will be potential outliers: {156, 139, -67}

Outliers or Noisy Data?

(Using Regression Analysis)



Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution

Outliers: Distance-based Approaches

- Data is represented as a nD vector of features
- Three major approaches:
  - Nearest-neighbor based
  - Density based
  - Clustering based

Outliers: Distance-based Approaches

Outlier detection for n-dimensional samples – *nearest-neighbor based* approach**:**

a) Evaluate the distance measures between all samples in n-dimensional data set.

b) A sample s_i in a data set S is an outlier if at least a fraction **p** of the samples in S lies at a distance greater than **d.**

Outliers: Distance-based Approaches

Outlier detection for n-dimensional samples - EXAMPLE

Data set: S = { (2,4), (3,2), (1,1), (4,3), (1,6), (5,3), (4,2)}
Requirements: p > 4, d > 3.00

***a)** Table of distances*

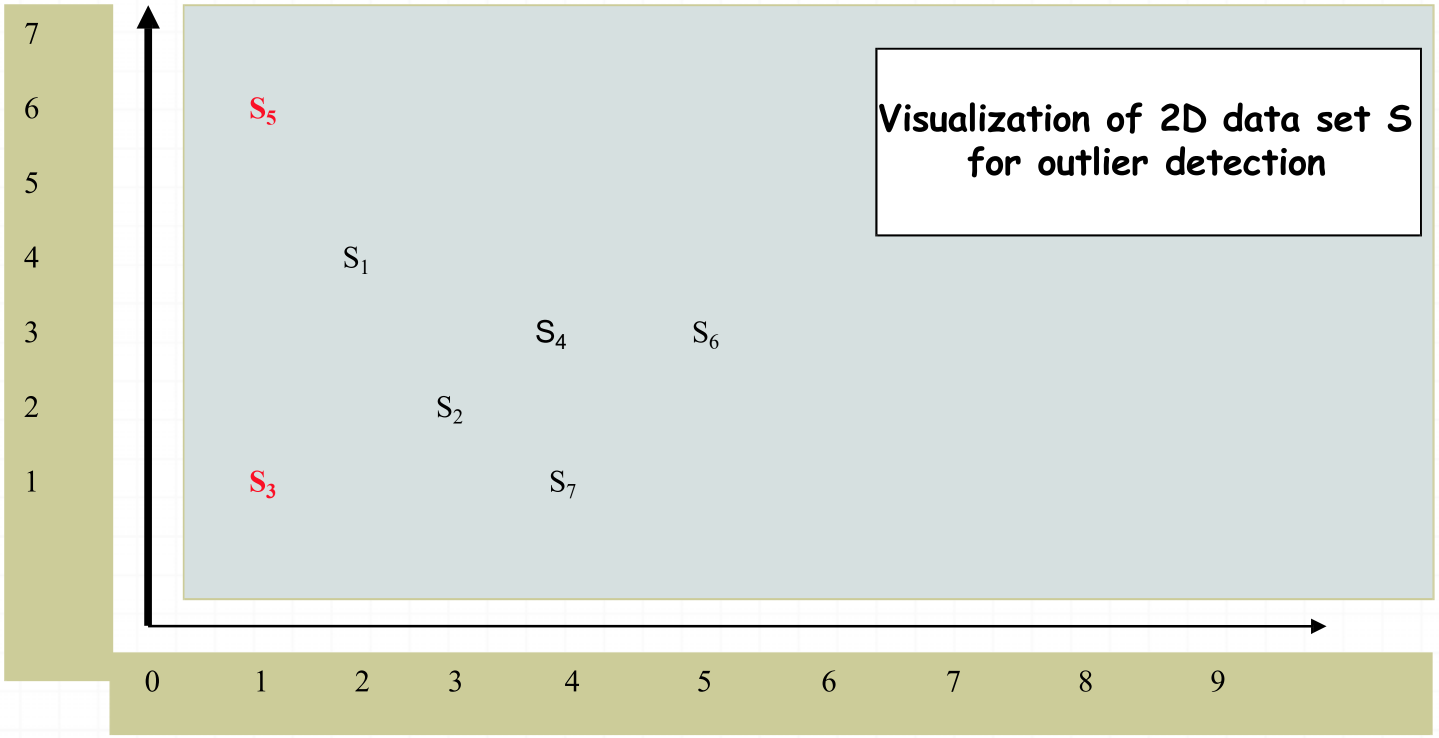
	S2	S3	S4	S5	S6	S7
S1	2.236	3.162	2.236	2.236	3.162	2.828
S2		2.236	1.414	4.472	2.236	1.000
S3			3.605	5.000	4.472	3.162
S4				4.242	1.000	1.000
S5					5.000	5.000
S6						1.414

b) p computation

Sample	p
S1	2
S2	1
S3	5
S4	2
S5	5
S6	3
S7	2

Outliers

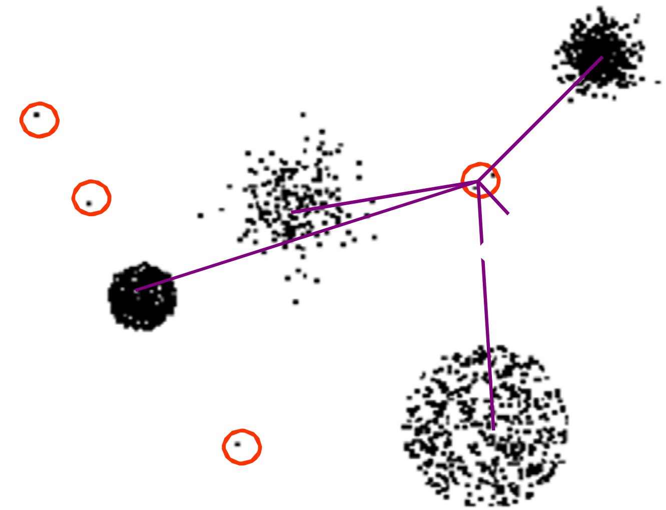
Outliers: Distance-based Approaches



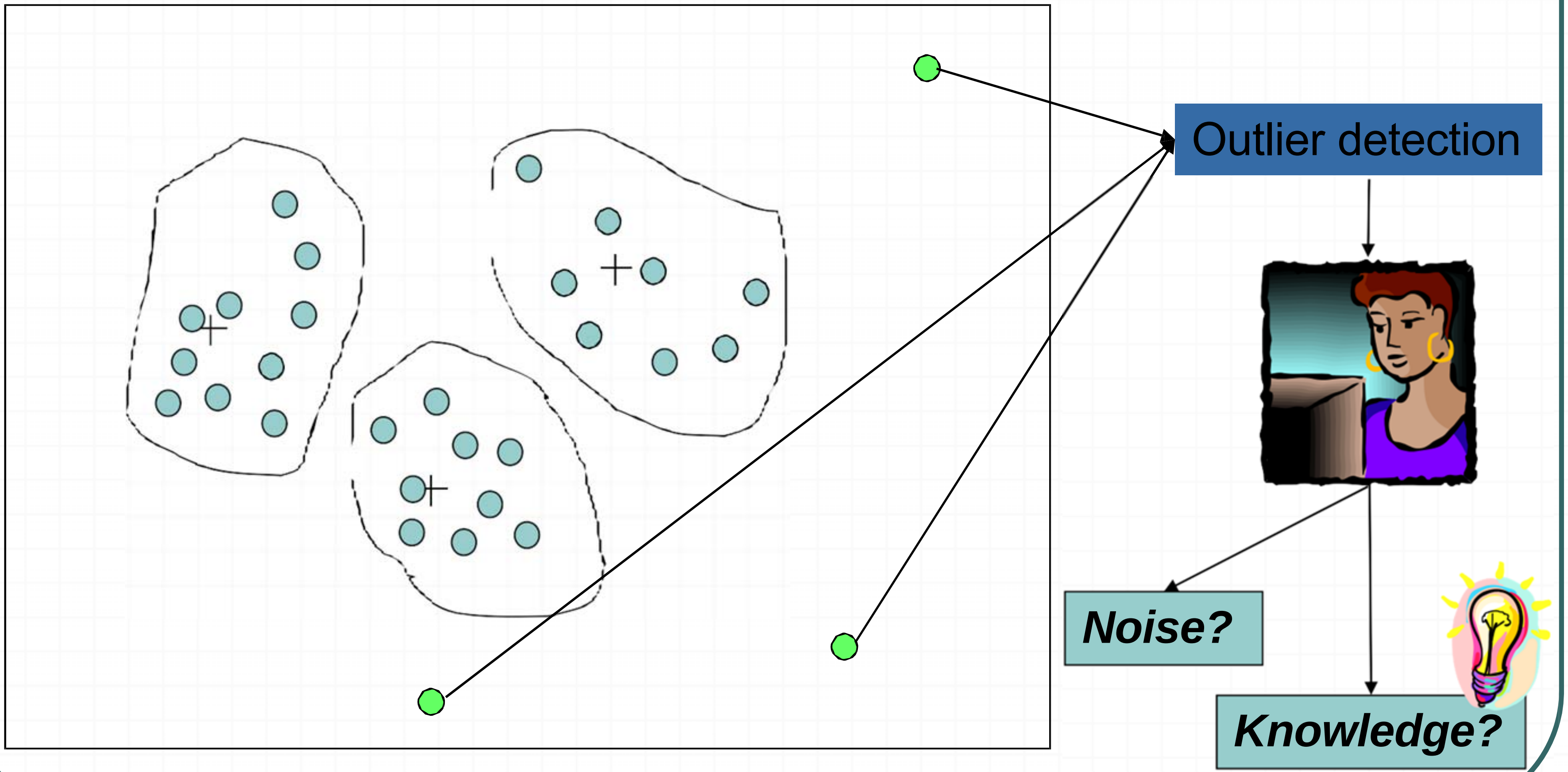
S = { (2,4), (3,2), (1,1), (4,3), (1,6), (5,3), (4,2) }

Outliers: Distance-based Approaches

- Basic idea for large data sets - **clustering based**:
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters:
    - If candidate points are far from all othe
outliers



Outliers or Noisy Data?
(Using Cluster Analysis)



Anomaly/Outlier Detection

- **Variants of Anomaly/Outlier Detection Problems**
  - Given a database D, find all the data points $\mathbf{x} \in D$ with anomaly scores greater than some threshold t
  - Given a database D, find all the data points $\mathbf{x} \in D$ having the top-n largest anomaly scores $f(\mathbf{x})$
  - Given a database D, containing mostly normal (but unlabeled) data points, and a test point **x**, compute the anomaly score of **x** with respect to D

- **Applications:**
  - Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection

Curse of Dimensionality

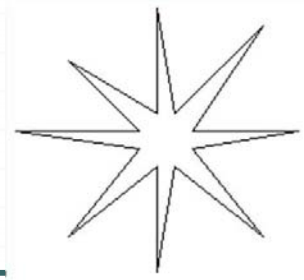
Curse of Dimensionality

- **The "curse of dimensionality" is due to the geometry of high- dimensional spaces.**

- The properties of high-dimensional spaces often appear **counterintuitive** because our experience with the physical world is in low-dimensional space such as space with two or three dimensions.

- Conceptually objects in high-dimensional spaces have a larger amount of surface area for a given volume than objects in low-dimensional spaces.

- For example, a high-dimensional hypercube, if it could be visualized, would look like a porcupine. As the dimensionality grows larger, the edges grow longer relative to the size of a central part of the hypercube.



Curse of Dimensionality

1. A size of a data set yielding the same density of data points in n-dimensional space, increase exponentially with dimensions.

SAME DENSITY OF DATA:

one-dimension	**k dimensions**
k = 1	**k=5**
n = 100 (samples)	**n^k = 100^5 = 10^{10} (samples)!!!**

Curse of Dimensionality

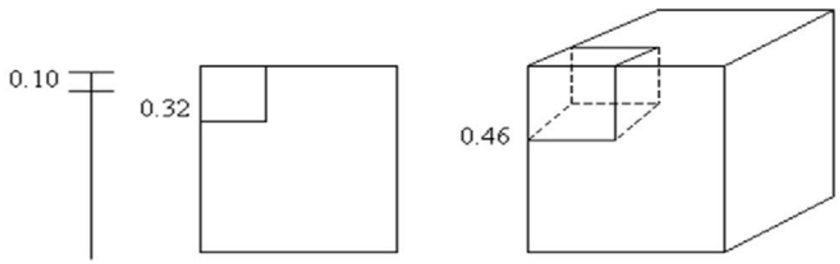
2. A larger radius is needed to enclose the same fraction of data points in a high-dimensional space. The edge length **e** of the hypercube:

$$e(p) = p^{1/d}$$

where p is the pre-specified fraction of samples and d is the number of dimensions.

10% of the samples (p=0.1):

one-dimension	2 dimensions	3 dimensions	10-dimensions
$e_1(0.1) = 0.1$	$e_2(0.1) = 0.32$	$e_3(0.1) = 0.46$	$e_{10}(0.1) = 0.80$



Curse of Dimensionality

3. Almost every point is closer to an edge than to another sample point in a high-dimensional space.

For a sample size n, the expected distance **D** between normalized data points in d-dimensional space is:

$$D(d, n) = \frac{1}{2} (1/n)^{1/d}$$

---

* For a two-dimensional space with 10000 points → D(2,10000) = 0.005

* For a 10-dimensional space with 10000 points → D(10,10000) = 0.4

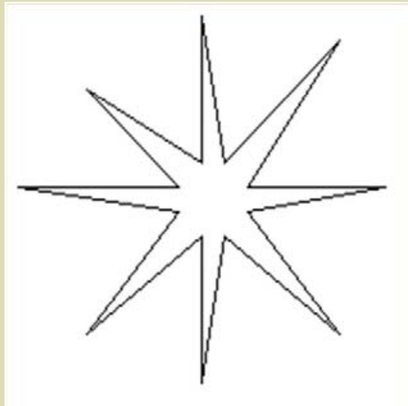
Curse of Dimensionality

4. Almost every point is an outlier in high-dimensional spaces

As the dimension of the input space increases, the distance between the prediction point and the center of data points increases.

When **d=10**, the expected value of the prediction point is **3.1 SD** away from the center of the data.

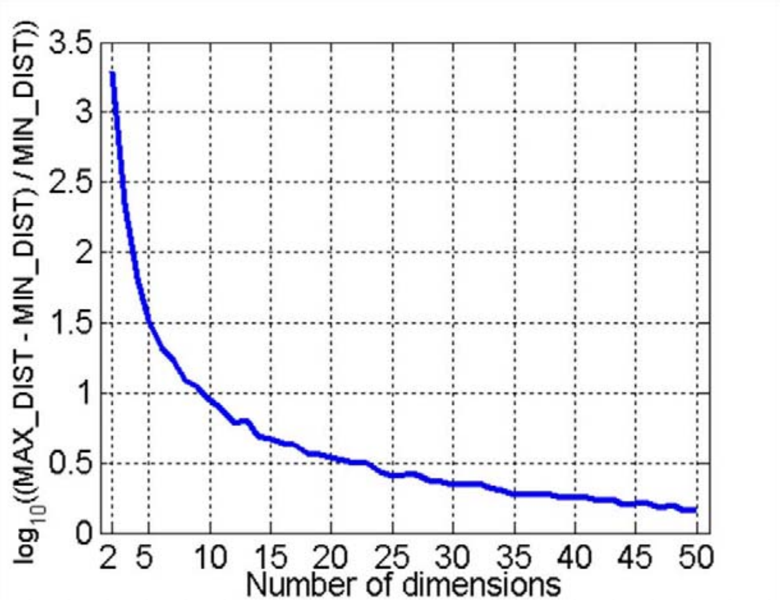
When **d=20**, the distance is **4.4 SD**.



Curse of Dimensionality

Experimental Confirmation:

- When dimensionality of data set increases, data becomes increasingly sparse with mostly outliers in the space that it occupies.
- Definitions of density and distance between points, which is critical for many data mining tasks, change the meaning!!!!!!



- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**