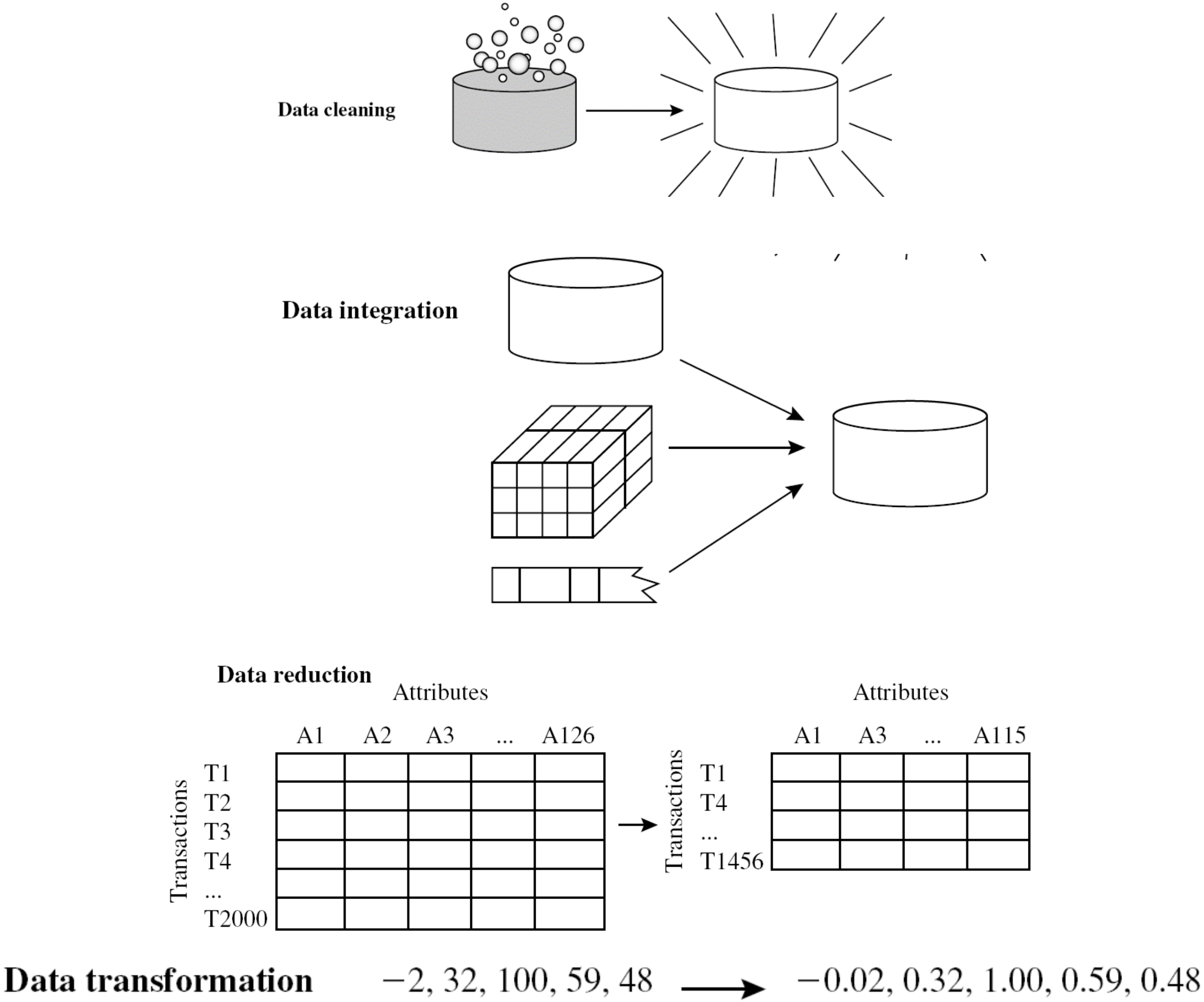
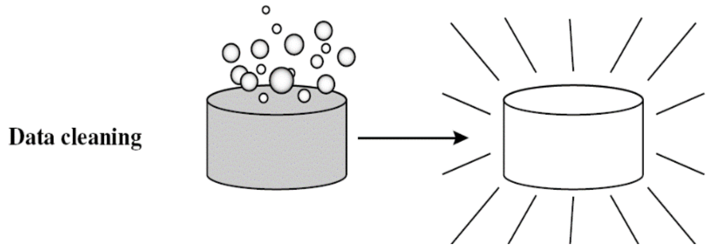
Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking *attribute values*, lacking certain *attributes of interest*,
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data



Forms of data preprocessing

Data Cleaning

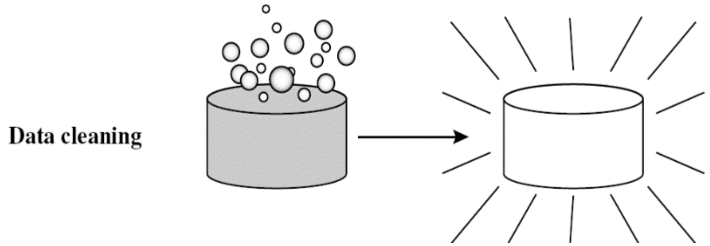


Missing Values

Noisy Data

- Ignore the tuple
- Fill in the missing value manually
- Use a global constant to fill in the missing value
- Use a measure of central tendency for the attribute
- Use the attribute mean or median for all samples belonging to the same class as the given tuple
- Use the most probable value to fill in the missing value

Data Cleaning



Missing Values

Noisy Data

Binning

Sorted data for price : 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

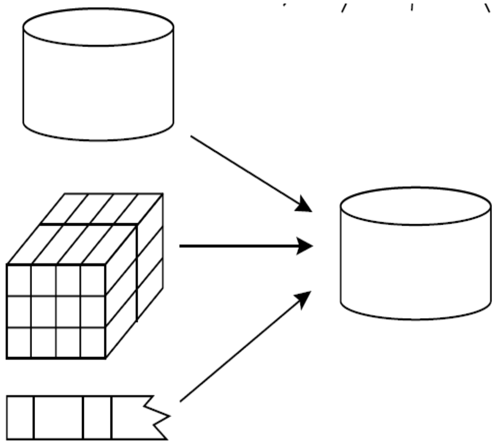
* Partition into (equi-depth) bins:
- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

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

Data Integration

Data integration



How can equivalent real-world entities from multiple data sources be matched up?

customer_id

customer number

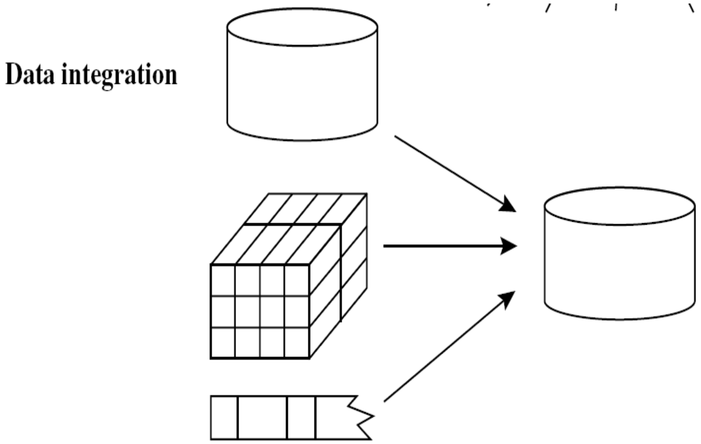
Entity Identification Problem

Redundancy and Correlation Analysis

Tuple Duplication

Data Value Conflict Detection and Resolution

Data Integration



Entity Identification Problem

Redundancy and Correlation Analysis

Tuple Duplication

Data Value Conflict Detection and Resolution

χ^2 Correlation Test for Nominal Data

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

	French	Russian
Male	39	16
Female	21	14

	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955	24.352	26.124
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32.000	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.790
18	6.265	8.231	22.760	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.900	27.204	30.144	32.852	33.687	36.191	38.582	41.610	43.820
20	7.434	9.591	25.038	28.412	31.410	34.170	35.020	37.566	39.997	43.072	45.315

$\mathcal{X}^2$ Correlation Test for Nominal Data

Step 1:
Null Hypothesis
H_o : there is no relationship between choice of language and gender.

Alternative Hypothesis
H_1 : the choice of language is dependent on gender.

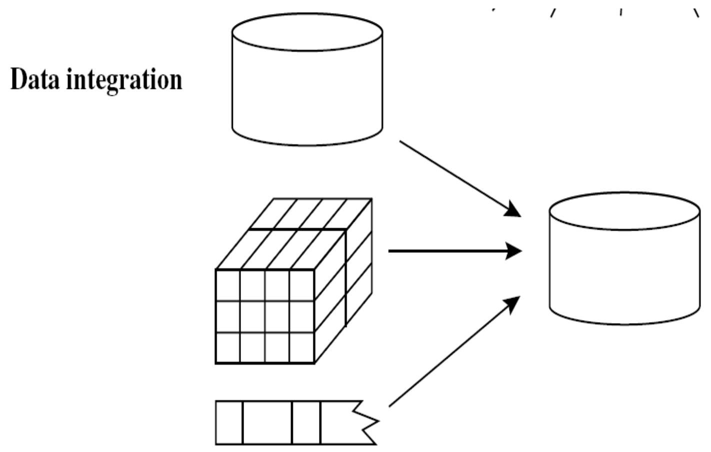
	French	Russian	Total
Male	39	16	55
Female	21	14	35
Total	60	30	90

	French	Russian	Total
Male	36.67	18.33	55
Female	23.33	11.67	35
Total	60	30	90

$\boldsymbol{From\ table, the\ critical\ \mathcal{X}^2\ at\ 5\%\ level\ is\ given\ by\ 3.84}$

H_0 is rejected if $\mathcal{X}^2$>3.84

Data Integration



Entity Identification Problem

Redundancy and Correlation Analysis

Tuple Duplication

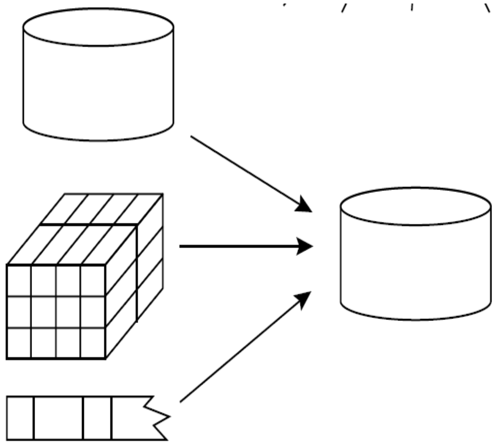
Data Value Conflict Detection and Resolution

Pearson’s product moment coefficient

$$\boxed{r = \frac{\frac{1}{n}\Sigma xy - \overline{xy}}{s_x s_y}}$$

where $s_x = \sqrt{\frac{1}{n}\Sigma x^2 - \overline{x}^2}$ and $s_y = \sqrt{\frac{1}{n}\Sigma y^2 - \overline{y}^2}$.

Data Integration

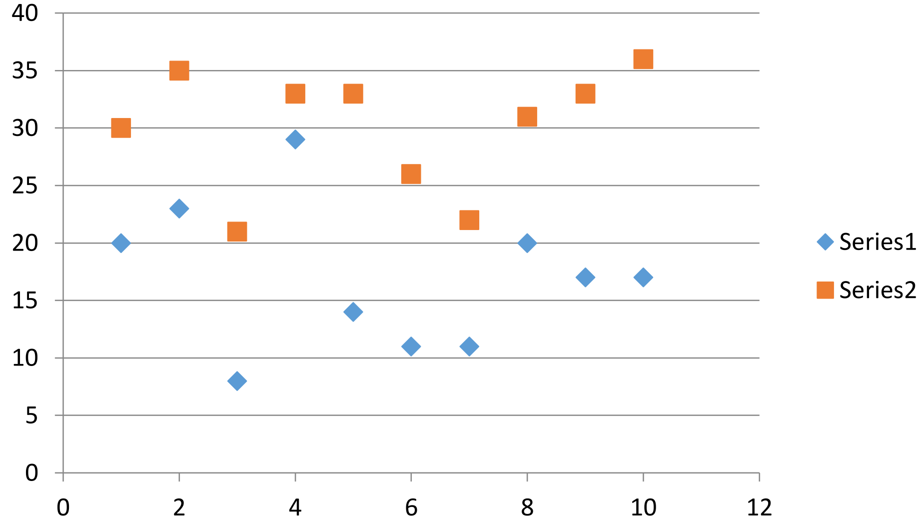


Entity Identification Problem

Redundancy and Correlation Analysis

Tuple Duplication

Data Value Conflict Detection and Resolution



Pearson's product moment coefficient

Pupil	A	B	C	D	E	F	G	H	I	J
Maths mark (out of 30) x	20	23	8	29	14	11	11	20	17	17
Physics mark (out of 40) y	30	35	21	33	33	26	22	31	33	36

the value of r gives how close the points are to lying on a straight line

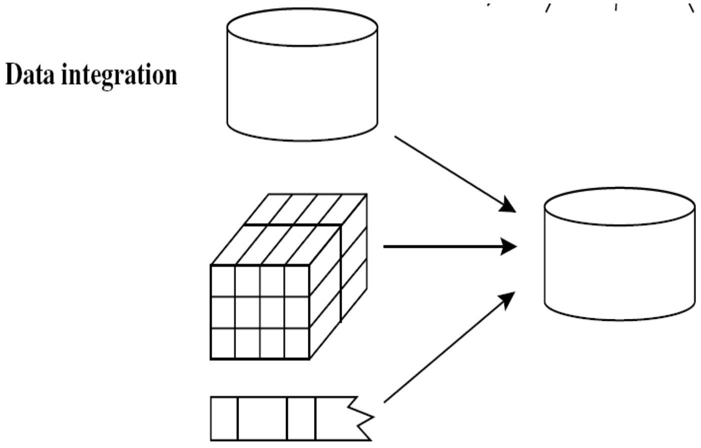
-1<=r<=1

r>0 positively correlated

r<0 negatively correlated

r=0 independent

Data Integration



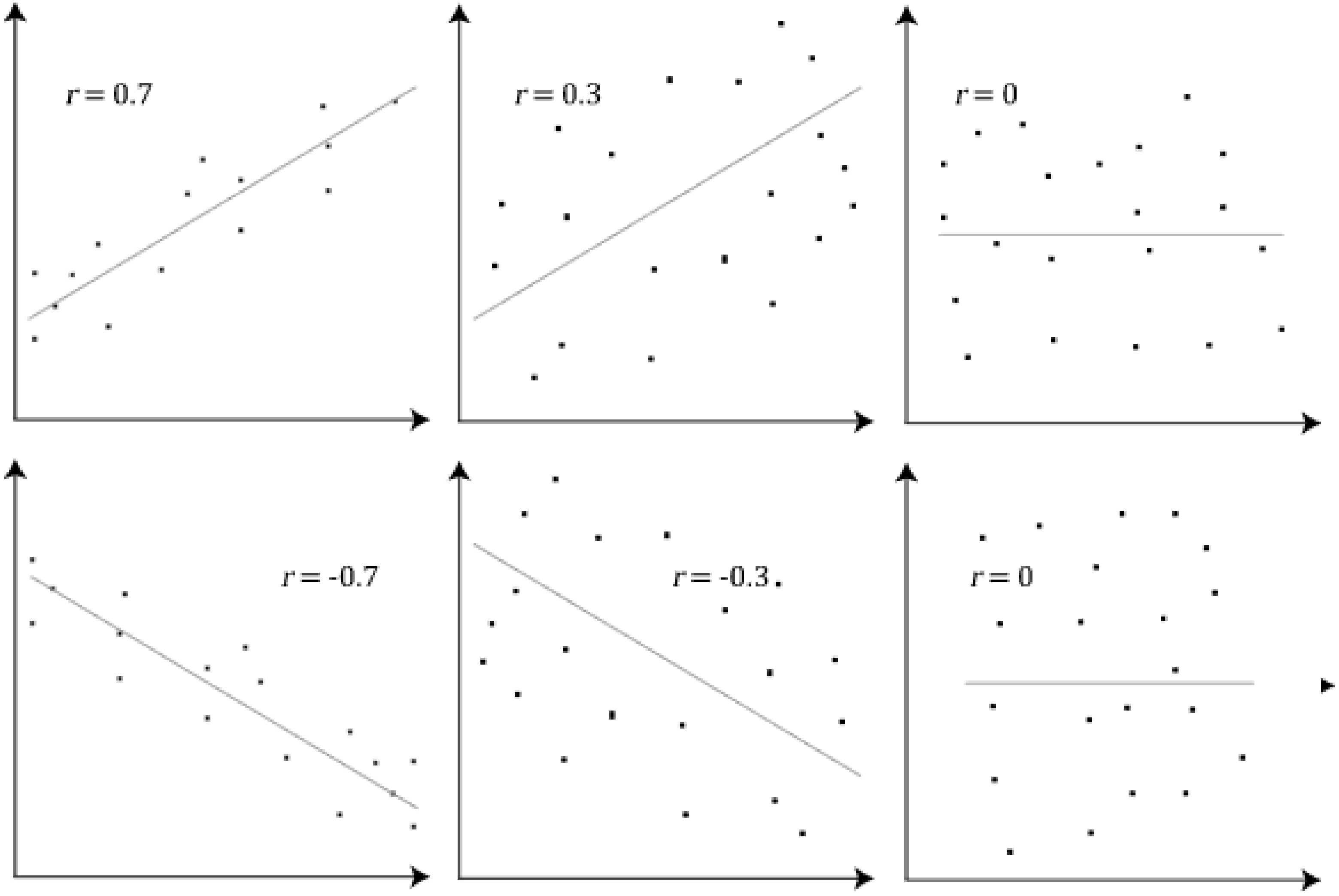
Entity Identification Problem

Redundancy and Correlation Analysis

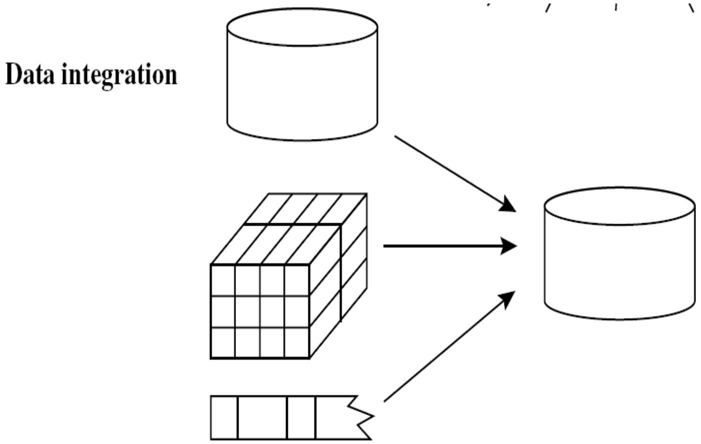
Tuple Duplication

Data Value Conflict Detection and Resolution

Pearson's product moment coefficient



Data Integration



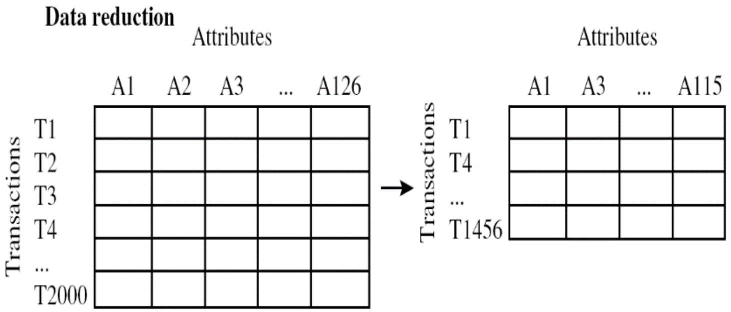
Entity Identification Problem

Redundancy and Correlation Analysis

Tuple Duplication

Data Value Conflict Detection and Resolution

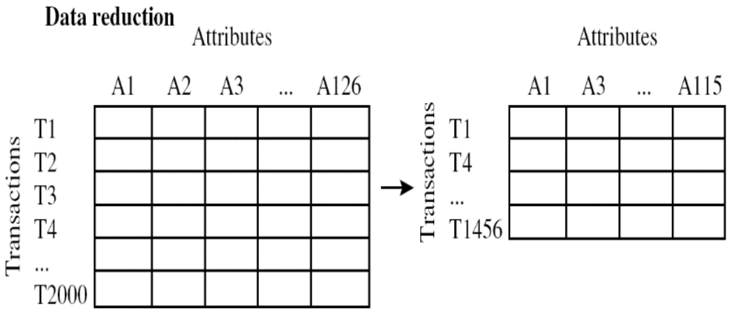
Data Reduction



- It applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data

- **Dimensionality reduction**
- **Numerosity reduction**
- **data compression**

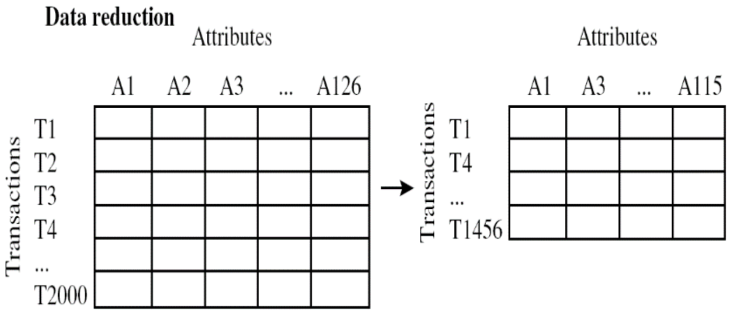
Data Reduction



Discrete wavelet transformation

- It is a signal processing techniques.
- When applied to a vector, convert the vector into a numerically different vector.
- Length of the vectors will be same

Data Reduction



A hierarchical pyramid algorithm

1. The length, *L, of the input data vector must be an integer power of 2. This condition* can be met by padding the data vector with zeros as necessary (*L n).*

2. Each transform involves applying two functions. The first applies some data smoothing, such as a sum or weighted average. The second performs a weighted difference, which acts to bring out the detailed features of the data.

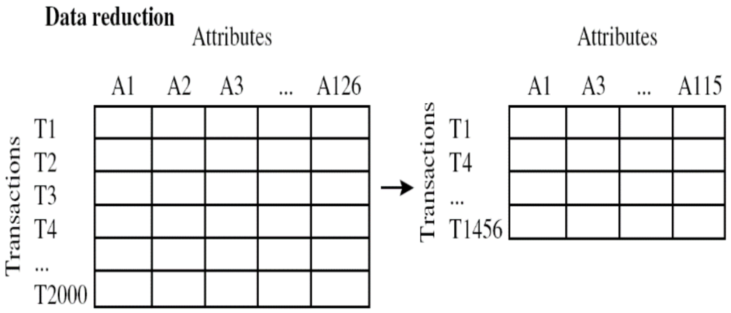
3. The two functions are applied to pairs of data points in *X, that is, to all pairs of* measurements (*x2i ,x2iC1). This results in two data sets of length L=2. In general,* these represent a smoothed or low-frequency version of the input data and the high frequency content of it, respectively.

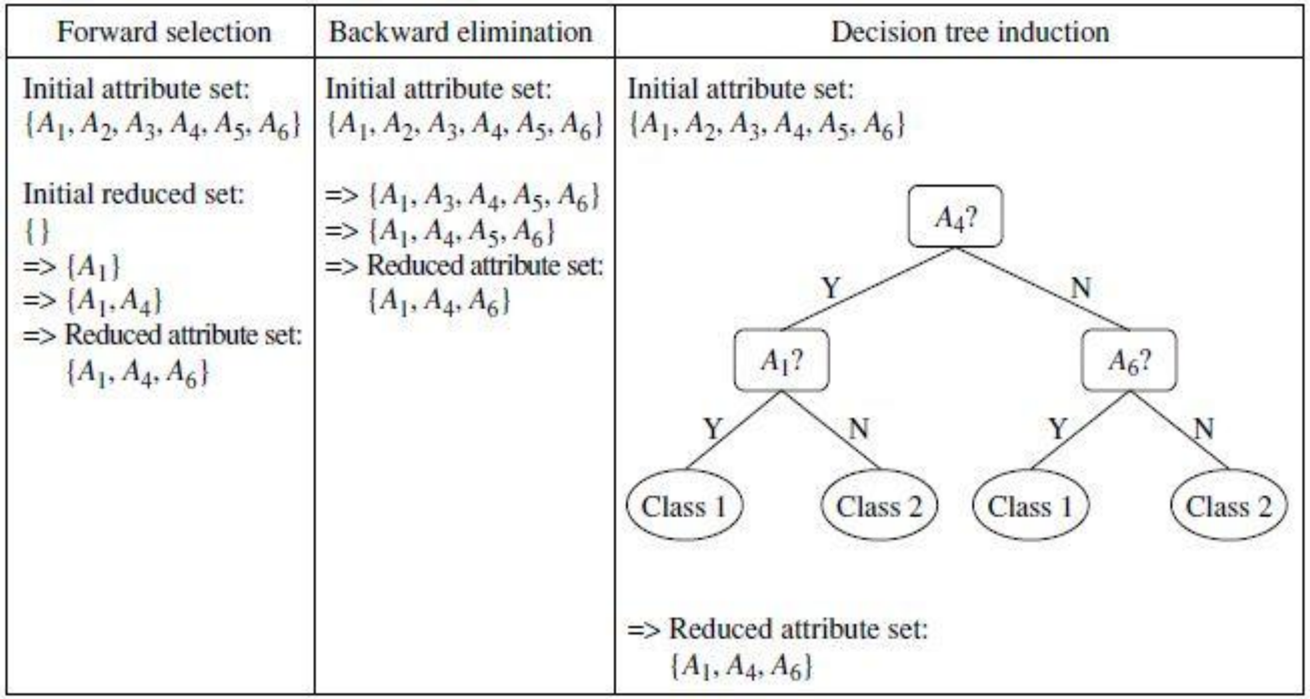
4. The two functions are recursively applied to the data sets obtained in the previous loop, until the resulting data sets obtained are of length 2.

5. Selected values from the data sets obtained in the previous iterations are designated the wavelet coefficients of the transformed data.

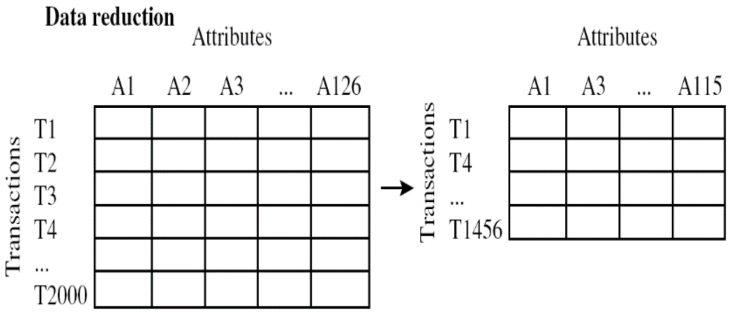
Data Reduction

Attribute subset selection



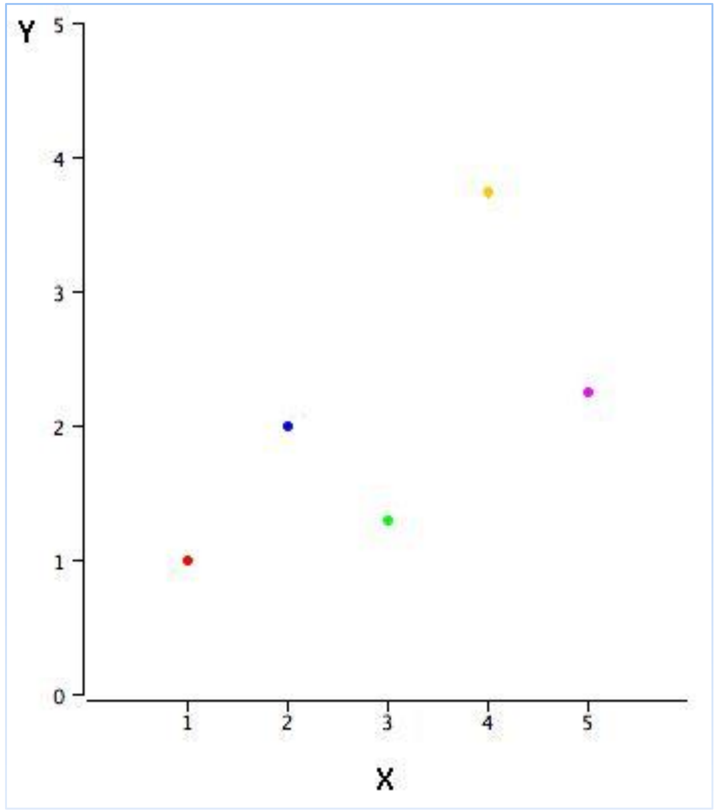


Data Reduction



Linear regression

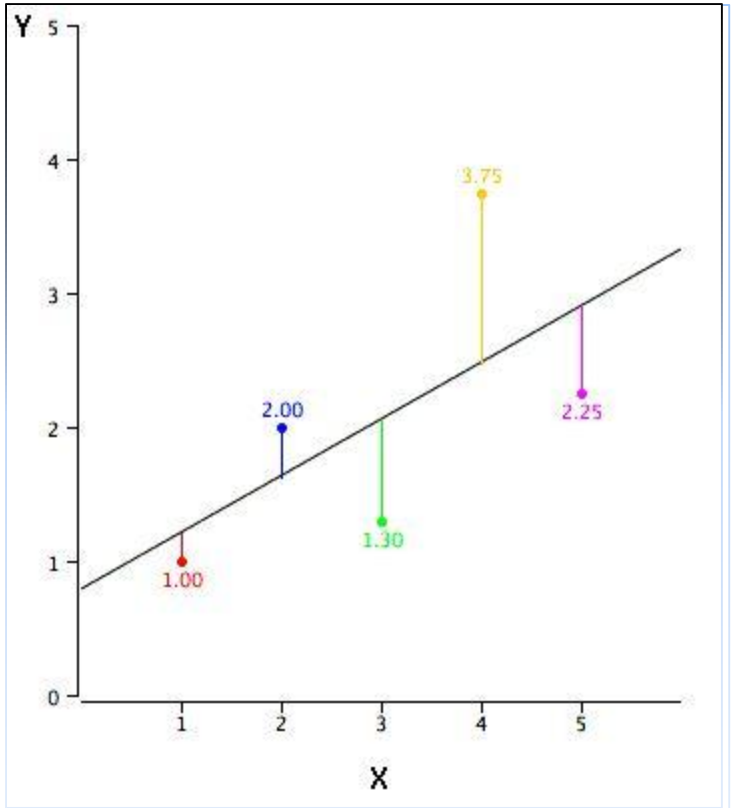
Y=mx+c



Log-linear models

It can be used to estimate the probability of each point in a multidimensional space for a set of discretized-attributes, based on a smaller subset of dimensional combinations

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25



X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

The slope (b) can be calculated as follows:

$b = r \ s_Y/s_X$

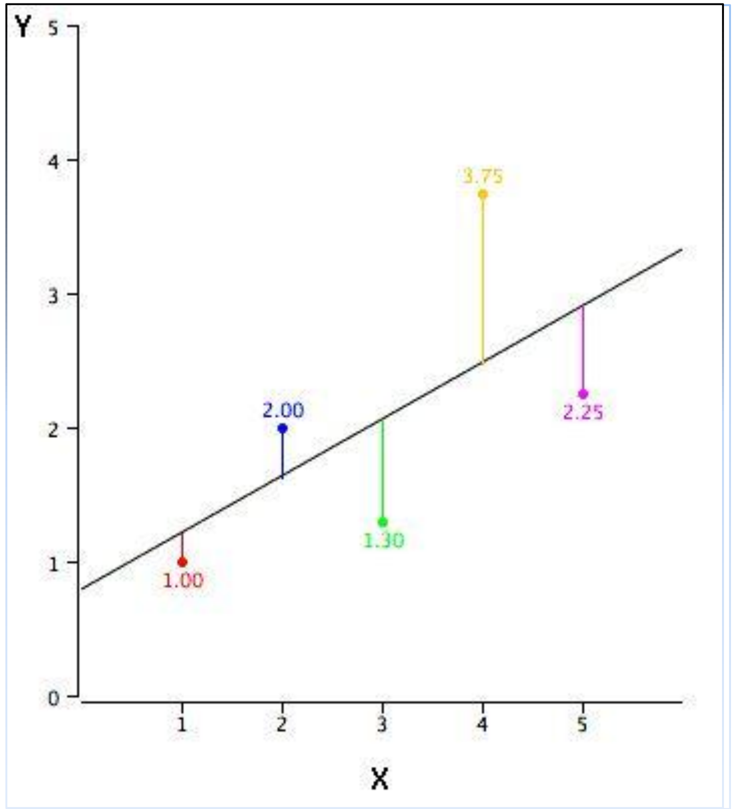
and the intercept (A) can be calculated as

$A = M_Y - bM_X$.

For these data,

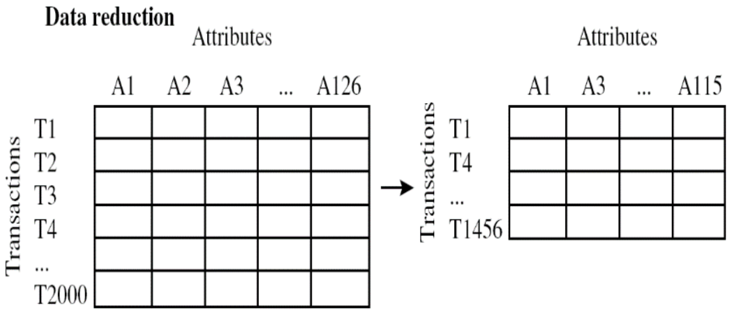
$b = (0.627)(1.072)/1.581 = 0.425$

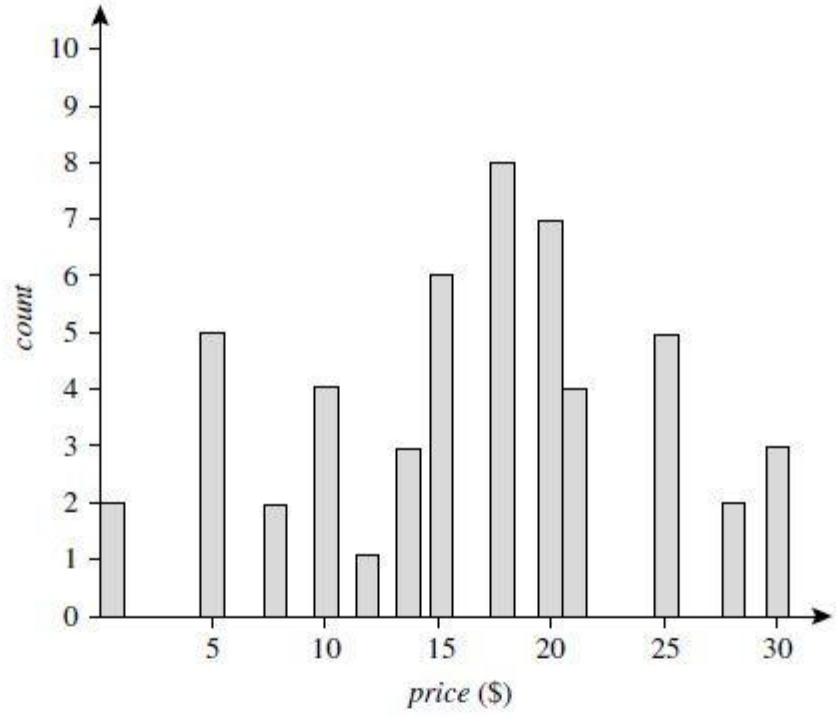
$A = 2.06 - (0.425)(3) = 0.785$



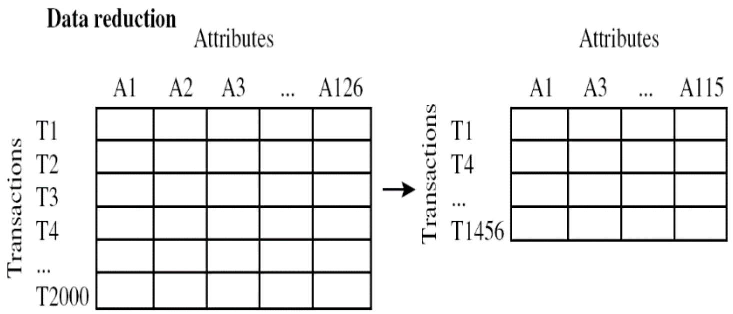
•**Histograms**

Data Reduction



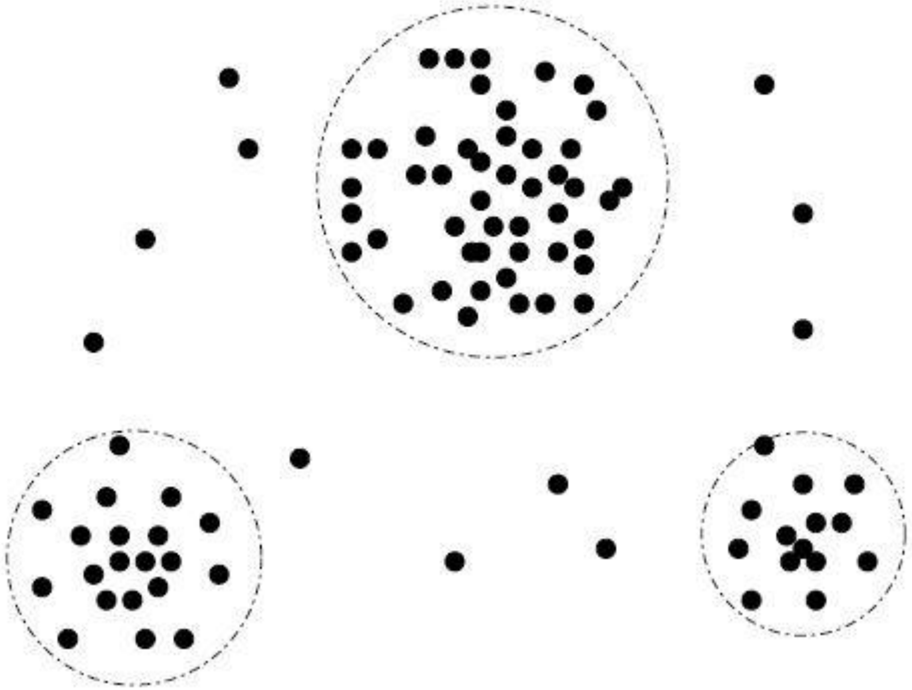


Data Reduction

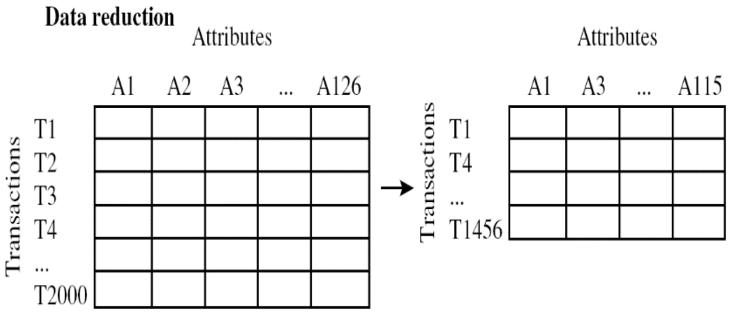


Clustering

• Partition the objects(data tuples) into groups, or clusters based on there similarity.



Data Reduction



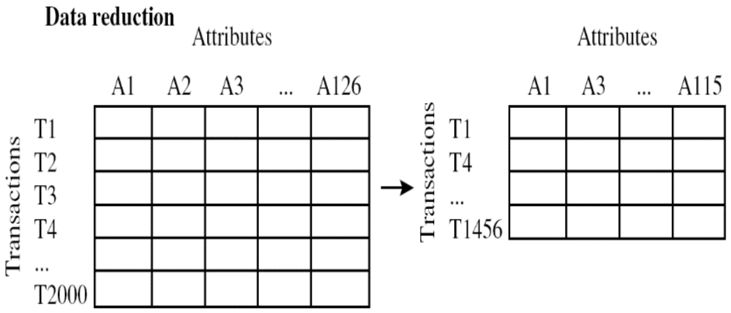
Sampling

- Simple random sample without replacement (SRSWOR) of size s
- Simple random sample with replacement (SRSWR) of size *s*
- Cluster sample

Skewed Data??

Skewed Data

Data Reduction



Startified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

Data Transformation

Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Min-max normalization

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

z-score normalization

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A},$$

Normalization by decimal scaling

56,   40,   8,   24,   48,   48,   40,   16

First row is the original signal. The second row in the table is generated by taking the mean of the samples pair-wise, put them in the first four places, and then the difference between the the first member of the pair and the computed mean. Computations are repeated on the *means.* Differences are kept in each step.

$\frac{56+40}{2}$

56	40	8	24	48	48	40	16
48	16	48	28	8	−8	0	12
32	38	16	10	8	−8	0	12
35	−3	16	10	8	−8	0	12

$56 - 48$

The transform is invertible. We start from the bottom row. We add and subtract the difference to the mean, and repeat the process up to the first row.

56	40	8	24	48	48	40	16
48	16	48	28	8	−8	0	12
32	38	16	10	8	−8	0	12
35	−3	16	10	8	−8	0	12

We replace samples in the transformed signal below 4 by zero (thresholding) and then repeat the reconstruction procedure:

59	43	11	27	45	45	37	13
51	19	45	25	8	−8	0	12
35	35	16	10	8	−8	0	12
35	0	16	10	8	−8	0	12

We now replace samples in the transformed signal below 9 by zero (threshold) and then repeat the reconstruction procedure.

51	51	19	19	45	45	37	13
51	19	45	25	0	0	0	12
35	35	16	10	0	0	0	12
35	0	16	10	0	0	0	12

Full line original signal, and dashed line for thresholding.