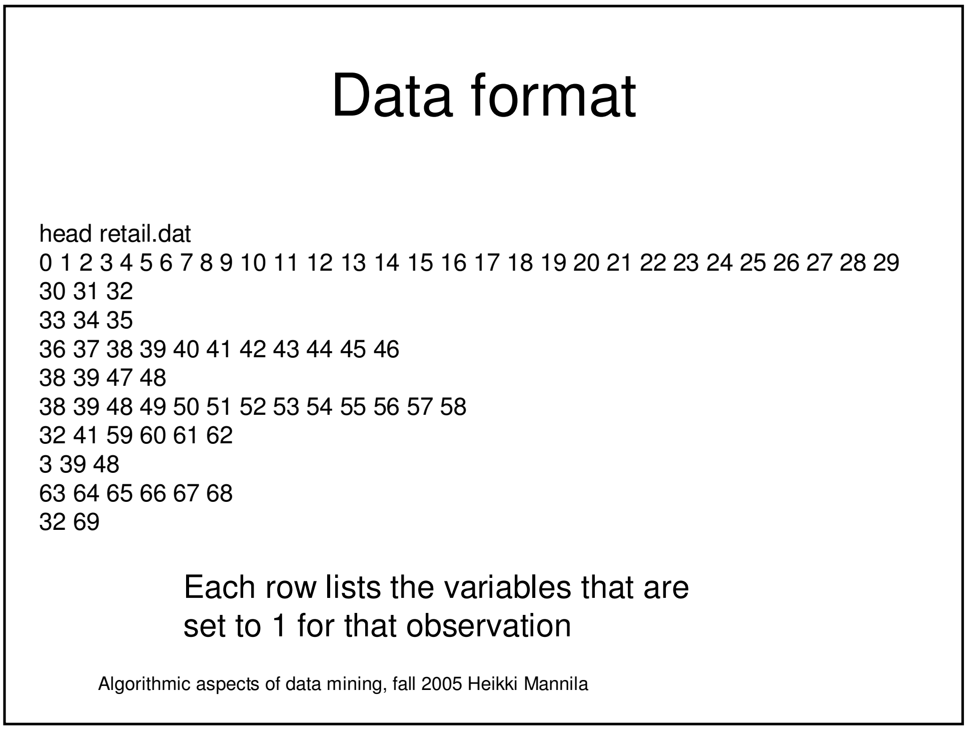
An example of a sparse 0-1 dataset

Example dataset

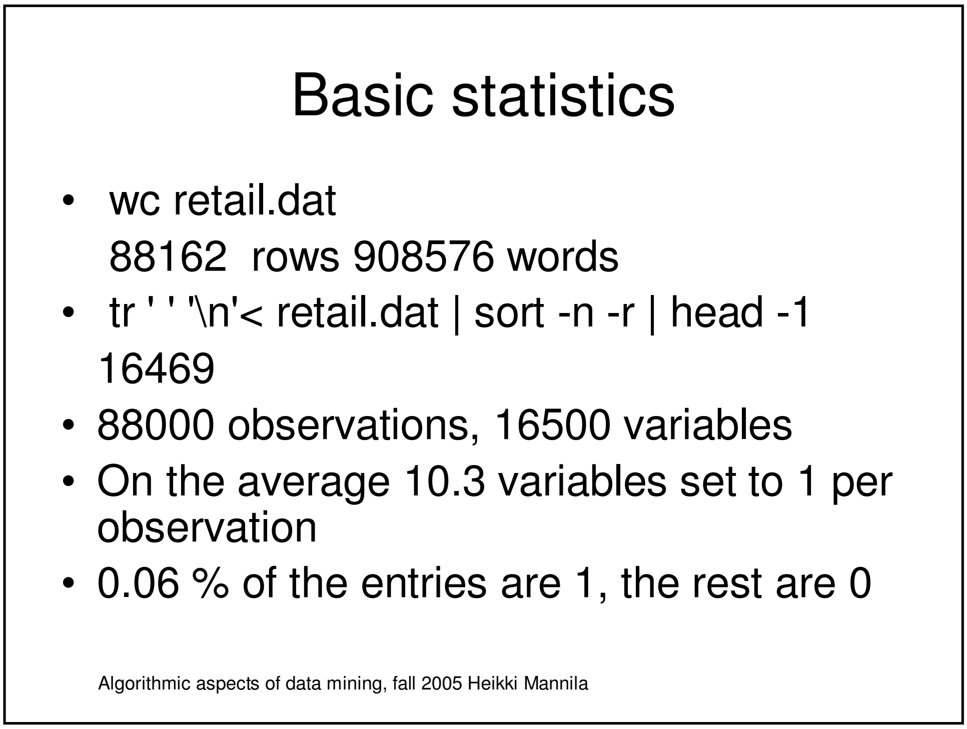
- http://fimi.cs.helsinki.fi/data/
- Retail: dataset was donated by Tom Brijs and contains the (anonymized) retail market basket data from an anonymous Belgian retail store.
- The data are provided ’as is’. Basically, any use of the data is allowed as long as the proper acknowledgment is provided and a copy of the work is provided to Tom Brijs.
  More details can be found here.
- http://fimi.cs.helsinki.fi/data/retail.pdf



Data format

```
head retail.dat
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
30 31 32
33 34 35
36 37 38 39 40 41 42 43 44 45 46
38 39 47 48
38 39 48 49 50 51 52 53 54 55 56 57 58
32 41 59 60 61 62
3 39 48
63 64 65 66 67 68
32 69
```

Each row lists the variables that are set to 1 for that observation



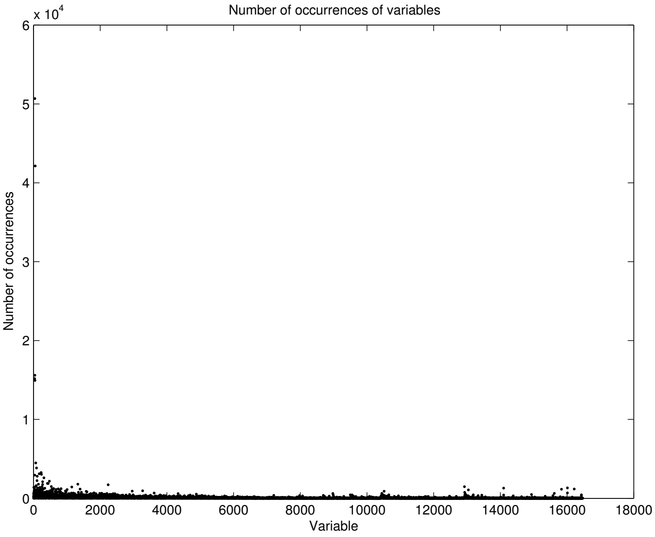
Basic statistics

- wc retail.dat
  88162 rows 908576 words
- tr ' ' '\n'< retail.dat | sort -n -r | head -1
  16469
- 88000 observations, 16500 variables
- On the average 10.3 variables set to 1 per observation
- 0.06 % of the entries are 1, the rest are 0

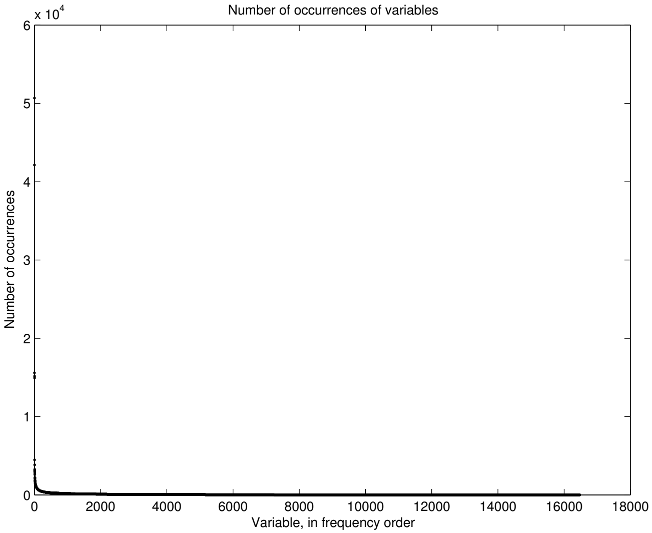
What is there in the data?

- Descriptive statistics
- Association rules
- Decomposition approaches
  - Principal components etc.

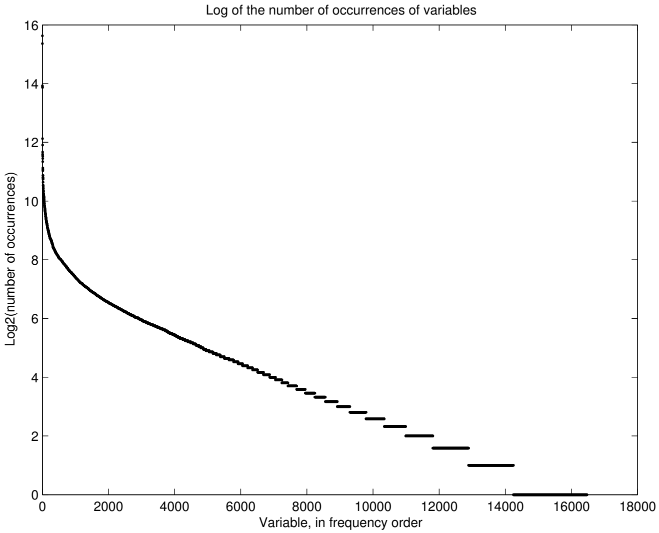
How often a variable is 1?



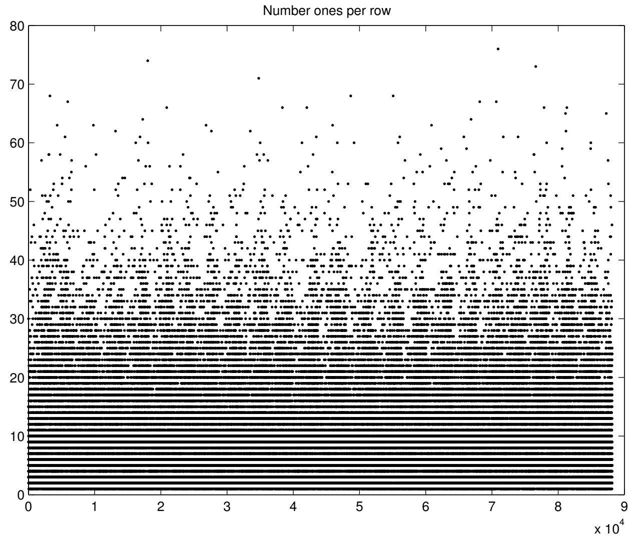
How often a variable is 1?



How often a variable is 1?

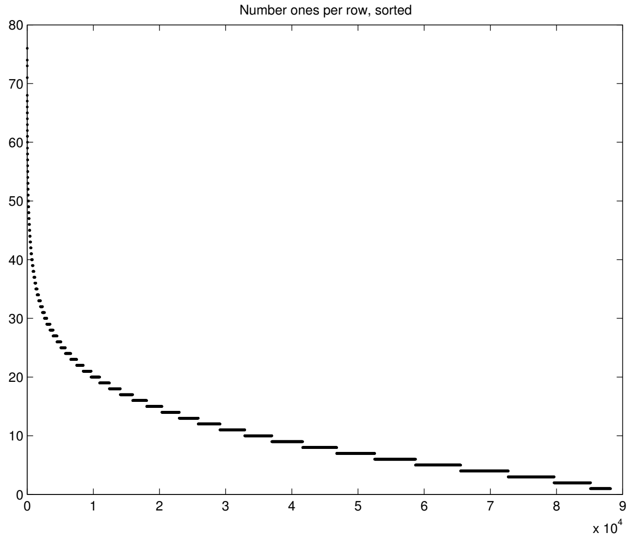


Number of ones per row



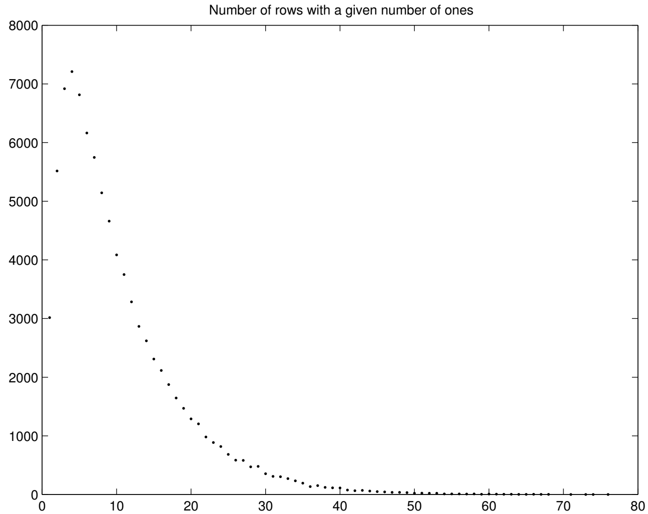
Algorithmic aspects of data mining, fall 2005 Heikki Mannila

Number of ones per row

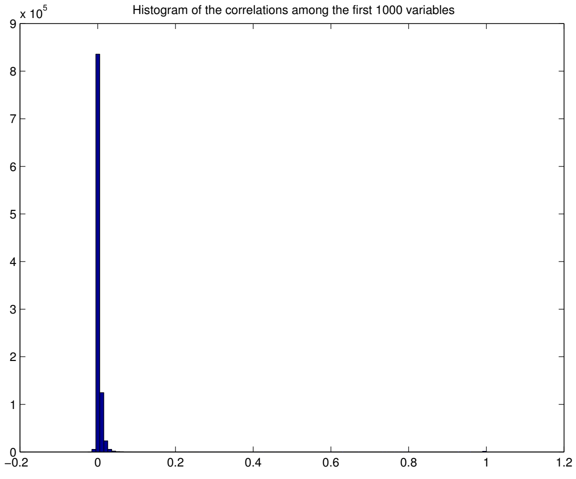


Algorithmic aspects of data mining, fall 2005 Heikki Mannila

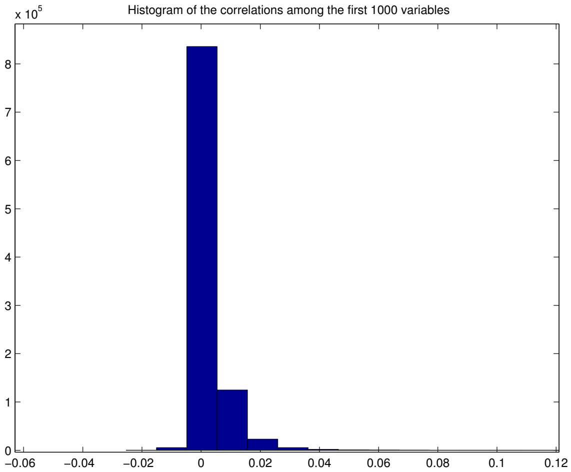
Number of rows with a given number of ones



Correlations

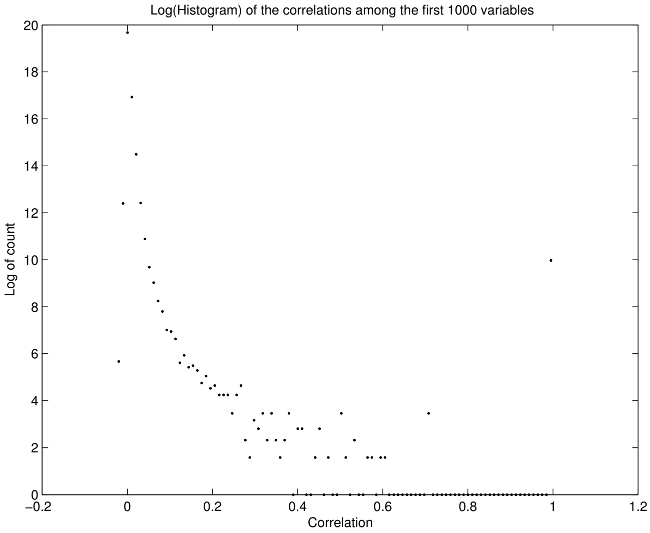


Correlations



Algorithmic aspects of data mining, fall 2005 Heikki Mannila

Histogram of correlations



Algori

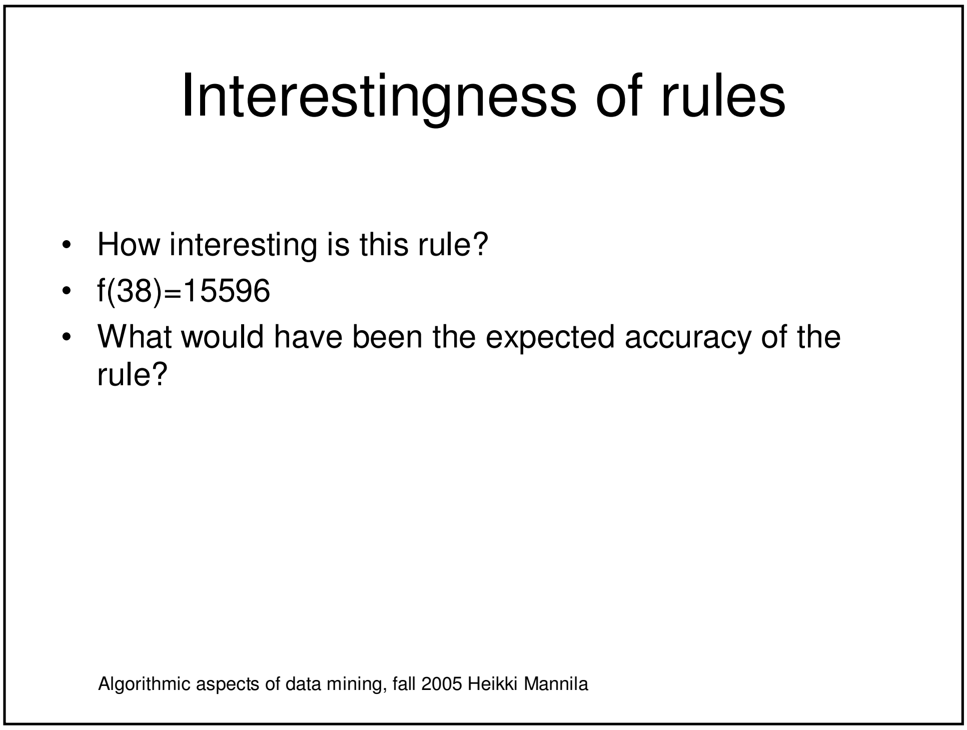
Frequent sets and rules from the retail data set

- Look at occurrence thresholds 5000, 2000, 1000, 500, 400, 300, 200, 100
- Rules with accuracy at least 0.9
- Bart Goethal’s implementation
- http://www.adrem.ua.ac.be/~goethals/software/index.html

Rules with threshold 1000

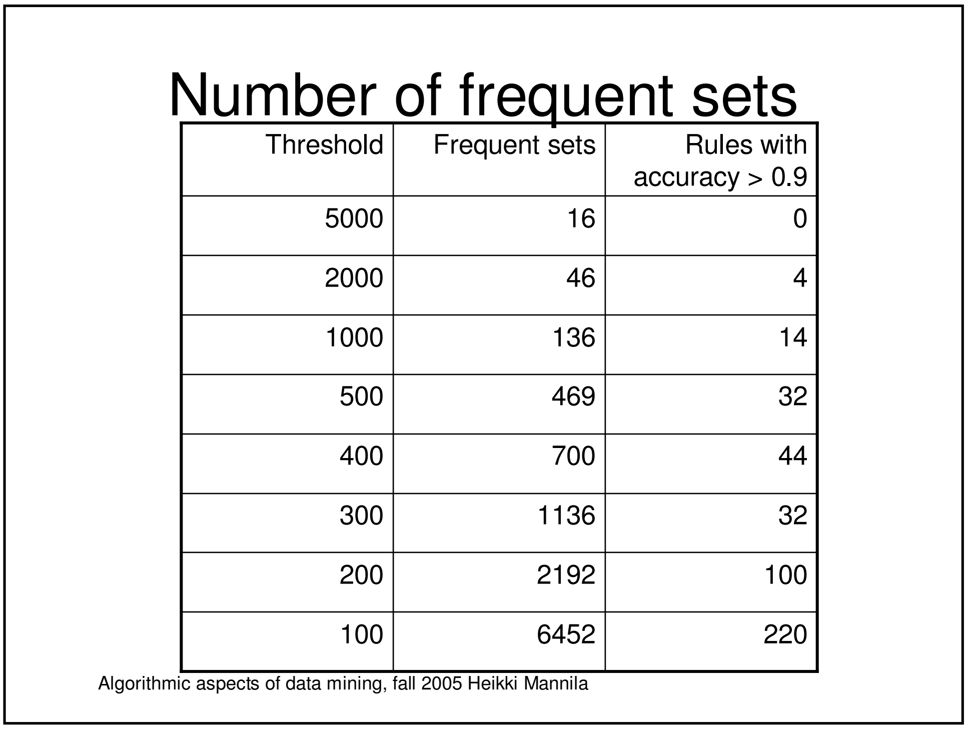
36 => 38	(2790, 0.950273)	
36 39 => 38	(1945, 0.954836)	
36 39 48 => 38	(1080, 0.967742)	
36 48 => 38	(1360, 0.960452)	
37 => 38	(1046, 0.973929)	
39 48 110 => 38	(1031, 0.994214)	
39 48 170 => 38	(1193, 0.989221)	
39 110 => 38	(1740, 0.989198)	
39 170 => 38	(2019, 0.980573)	
48 110 => 38	(1361, 0.986232)	
48 170 => 38	(1538, 0.987797)	
110 => 38	(2725, 0.975304)	
170 => 38	(3031, 0.978057)	
286 => 38	(1116, 0.943364)	= f(38 286)/ f(286)

= f(38 286)



Interestingness of rules

- How interesting is this rule?
- f(38)=15596
- What would have been the expected accuracy of the rule?



Number of frequent sets

Threshold	Frequent sets	Rules with accuracy > 0.9
5000	16	0
2000	46	4
1000	136	14
500	469	32
400	700	44
300	1136	32
200	2192	100
100	6452	220