Spatial Data Mining for Customer Segmentation

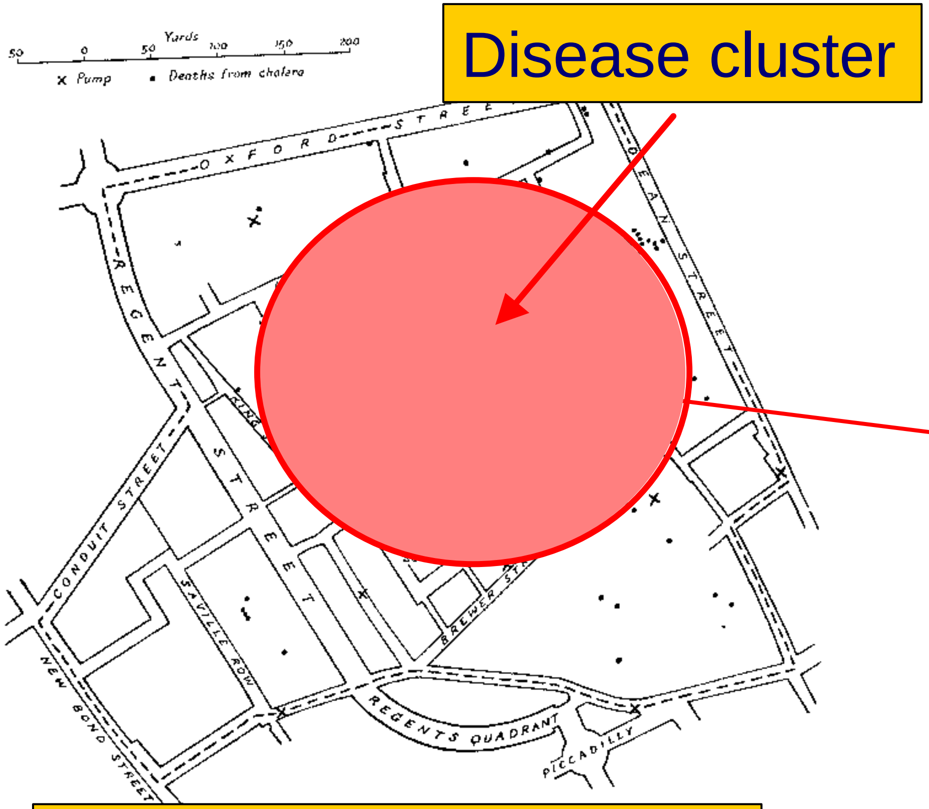
AIS

Fraunhofer Institut
Autonome Intelligente
Systeme

Data Mining in Practice Seminar, Dortmund, 2003

Dr. Michael May
Fraunhofer Institut Autonome Intelligente Systeme

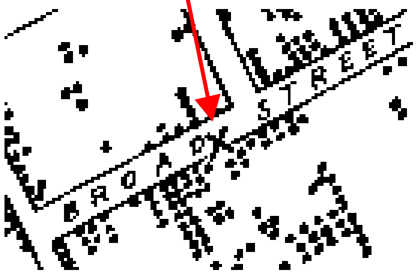
Introduction: a classic example for spatial analysis



Disease cluster

Dr. John Snow
Deaths of cholera epidemia
London, September 1854

Infected water pump?

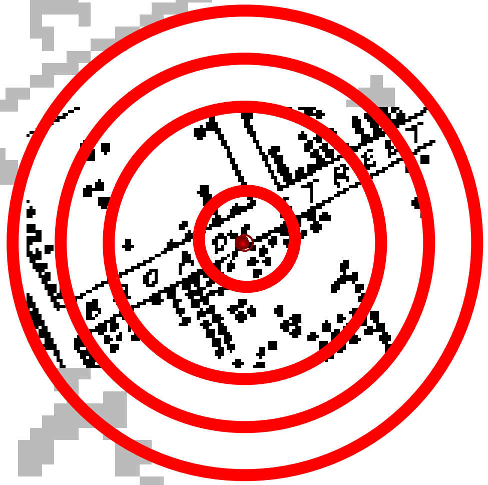


A good representation is the key to solving a problem

Good representation because...

Represents spatial relation of objects of the same type

Represents spatial relation of objects to *other* objects

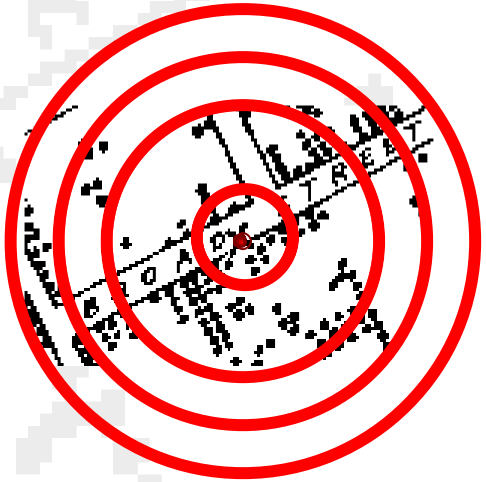


Shows only relevant aspects and hides irrelevant

It is not only important where a cluster is but also, what else is there (e.g. a water-pump)!

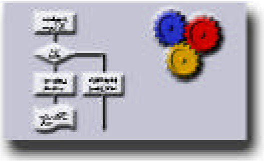
Goals of Spatial Data Mining

- Identifying spatial patterns
- Identifying spatial objects that are potential generators of patterns
- Identifying information relevant for explaining the spatial pattern (and hiding irrelevant information)
- Presenting the information in a way that is intuitive and supports further analysis



Approach to Spatial Knowledge Discovery

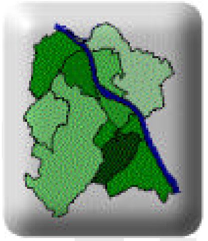
Data Mining



$$\sqrt{\frac{n}{p_0\cdot(1-p_0)}}(p-p_0)$$

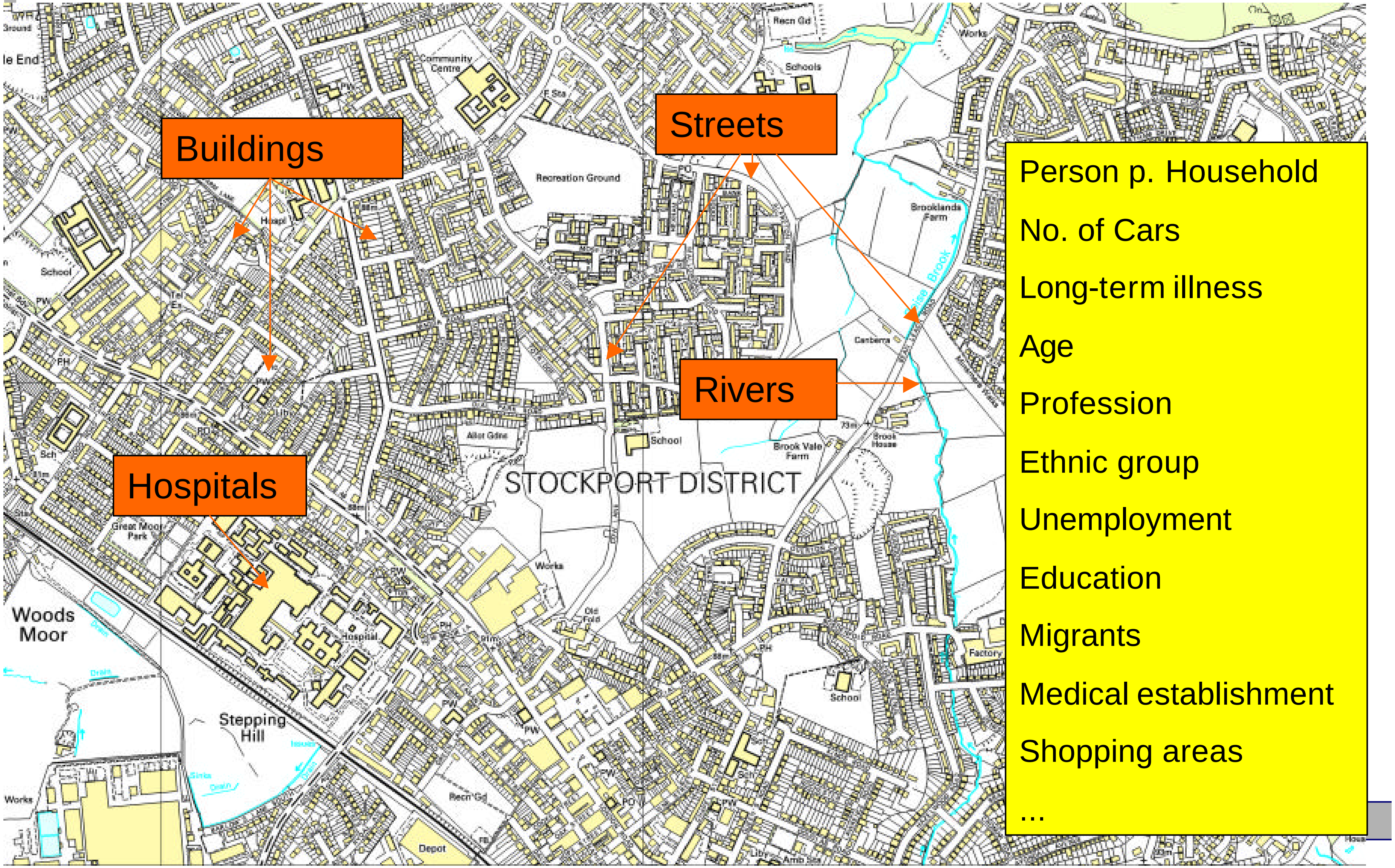
+

Geographic Information Systems



= SPIN!

UK, Greater Manchester, Stockport



Buildings

Streets

Rivers

Hospitals

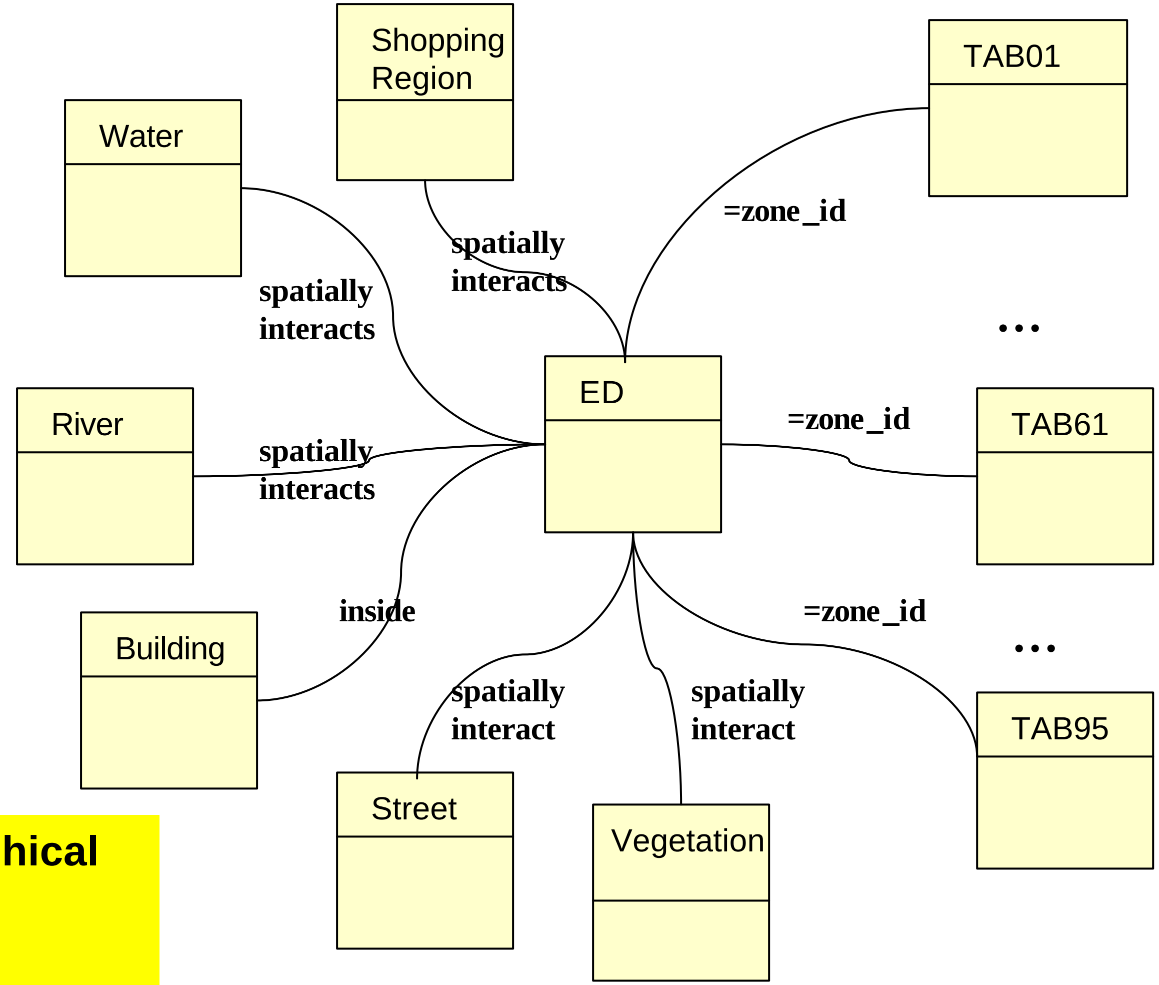
- Person p. Household
- No. of Cars
- Long-term illness
- Age
- Profession
- Ethnic group
- Unemployment
- Education
- Migrants
- Medical establishment
- Shopping areas
- ...

Representation of spatial data in Oracle Spatial

A set of relations $R_1,...,R_n$ such that each relation R_i has a geometry attribute G_i or an identifier A_i such that R_i can be linked (joined) to a relation R_k having a geometry attribute G_k

- Geometry attributes G_i consist of ordered sets of x,y-pairs defining points, lines, or polygons
- Different types of spatial objects are organized in different relations R_i (geographic layers), e.g. streets, rivers, enumeration districts, buidlings, and
- each layer can have its own set of attributes $A_1,..., A_n$ and at most one geometry attribute G

Stockport Database Schema



Attribute data

95 tables with census data,

~8000 attributes

Spatial Hierarchy

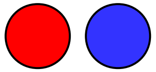
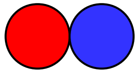
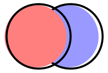
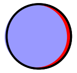
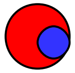
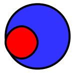
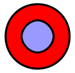
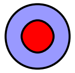
- County
- District
- Wards
- Enumeration district

Geographical Layers

85 tables

Spatial Predicates in Oracle Spatial

Topological relation (Egenhofer 1991)

A disjoint B, B disjoint A	
A meets B, B meets A	
A overlaps B, B overlaps A	
A equals B, B equals A	
A covers B, B covered by A	
A covered-by B, B covers A	
A contains B, B inside A	
A inside B, B contains A	

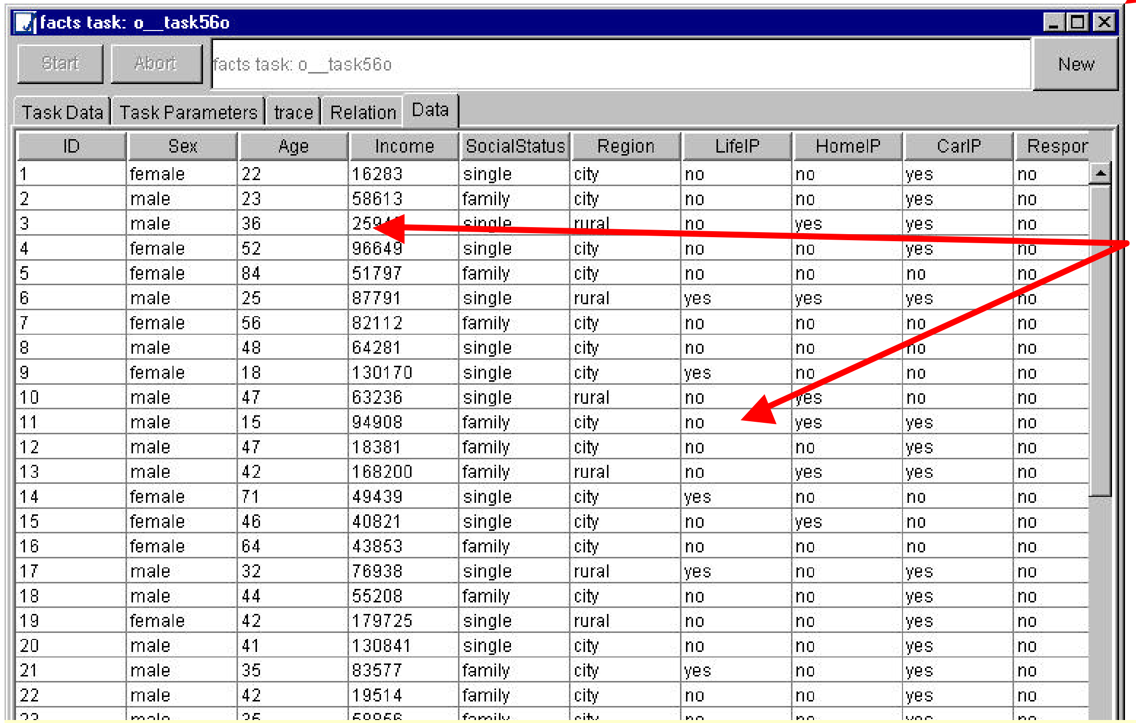
Distance relation: Minimum distance between 2 points

Typical Data Mining representation

‘spreadsheet data’

exactly 1 table

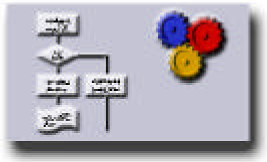
atomic values



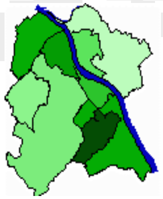
ID	Sex	Age	Income	SocialStatus	Region	LifeIP	HomeIP	CarIP	Respor
1	female	22	16283	single	city	no	no	yes	no
2	male	23	58613	family	city	no	no	yes	no
3	male	36	259[illegible]	single	rural	no	yes	yes	no
4	female	52	96649	single	city	no	no	yes	no
5	female	84	51797	family	city	no	no	no	no
6	male	25	87791	single	rural	yes	yes	yes	no
7	female	56	82112	family	city	no	no	no	no
8	male	48	64281	single	city	no	no	no	no
9	female	18	130170	single	city	yes	no	no	no
10	male	47	63236	single	rural	no	yes	no	no
11	male	15	94908	family	city	no	yes	yes	no
12	male	47	18381	family	city	no	no	yes	no
13	male	42	168200	family	rural	no	yes	yes	no
14	female	71	49439	single	city	yes	no	no	no
15	female	46	40821	single	city	no	yes	no	no
16	female	64	43853	family	city	no	no	no	no
17	male	32	76938	single	rural	yes	no	yes	no
18	male	44	55208	family	city	no	no	yes	no
19	female	42	179725	single	rural	no	no	yes	no
20	male	41	130841	single	city	no	no	yes	no
21	male	35	83577	family	city	yes	no	yes	no
22	male	42	19514	family	city	no	no	yes	no
23	male	35	[illegible]	family	city	no	no	yes	no

Data Mining for spatial data: strong discrepancy between usual and adequate problem representation

SPIN! – The Elements



$$\sqrt{\frac{n}{p_0 \cdot (1-p_0)}}(p-p_0)$$

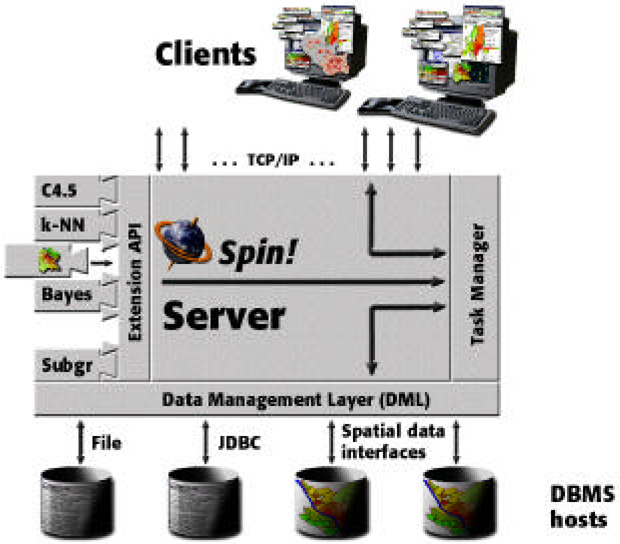


1. Spatial Data Mining Platform

Providing an integrated data mining platform

- Data access to heterogeneous and distributed data sources (Oracle RDBMS, flat file, spatial data)
- Organizing and documenting analysis tasks
- Launching analysis tasks
- Visualizing results

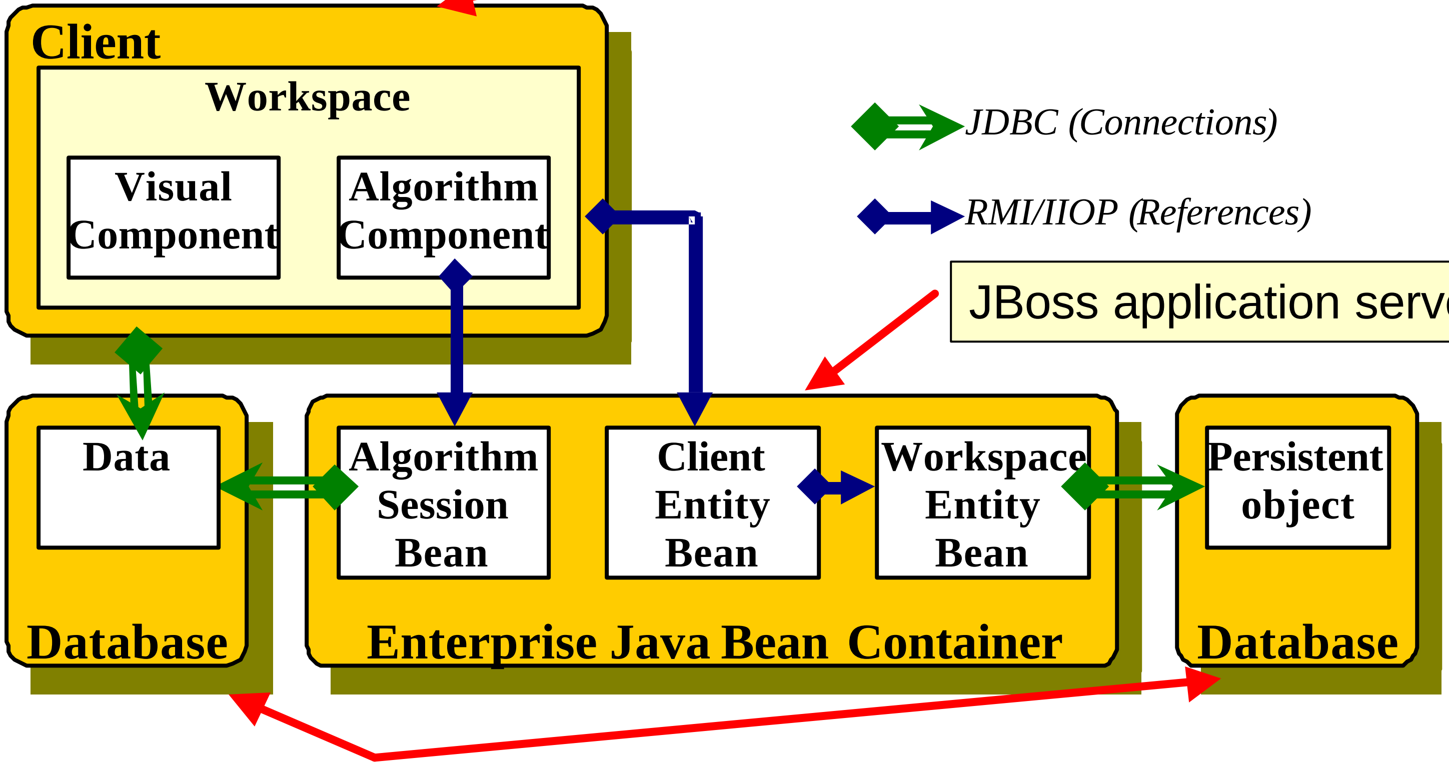
Note: Same software basis as MiningMart!



SPIN! Architecture: Enterprise Java Bean-based

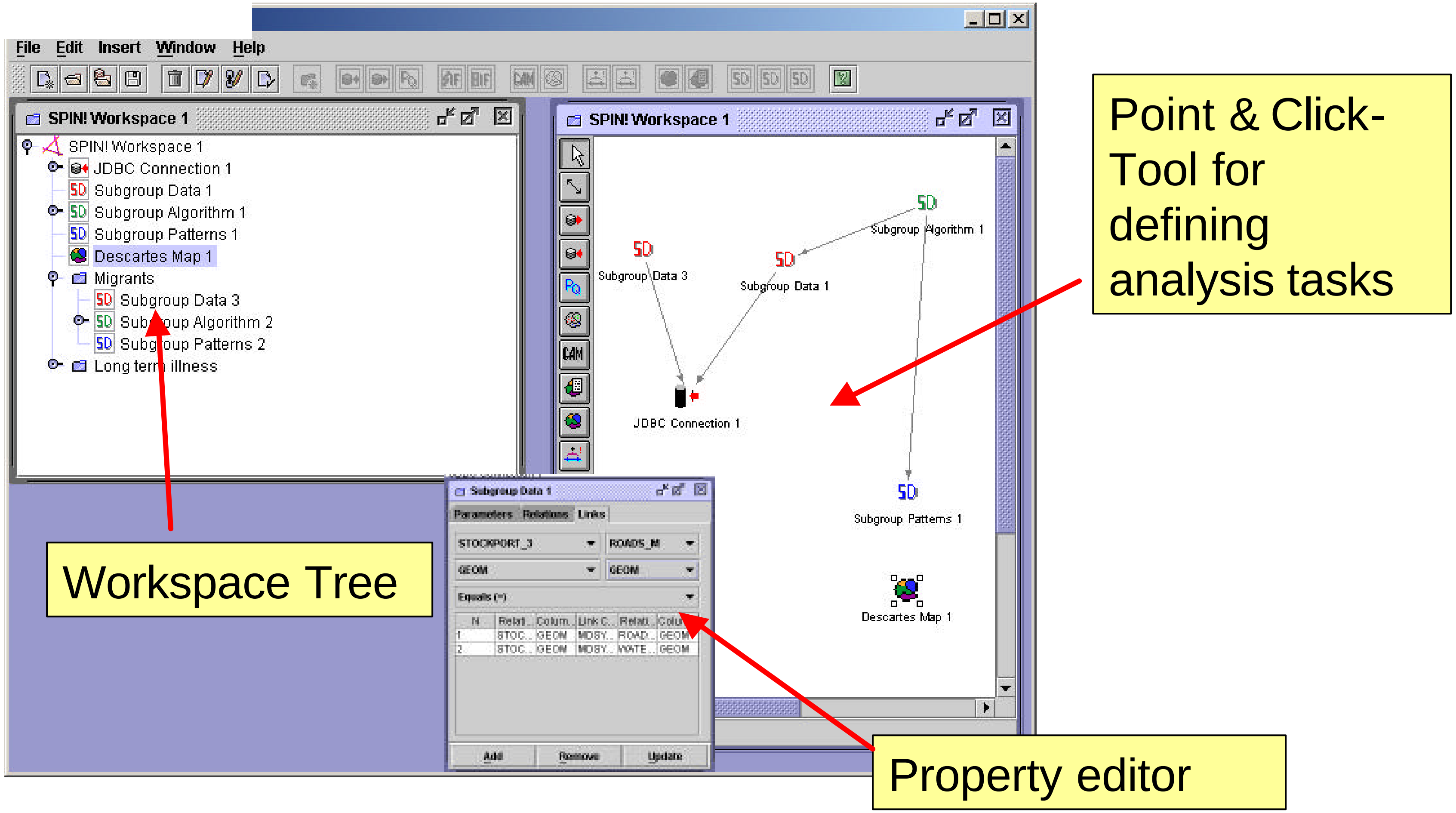
Java Swing based Client

JBoss application server



Object-relational spatial database (Oracle9i)

SPIN! User Interface



STOCKPORT_3 | ROADS_M

GEOM | GEOM

Equals (=)

N	Relati...	Colum...	Link C...	Relati...	Colu...
1	STOC...	GEOM	MDSY...	ROAD...	GEOM
2	STOC...	GEOM	MDSY...	WATE...	GEOM

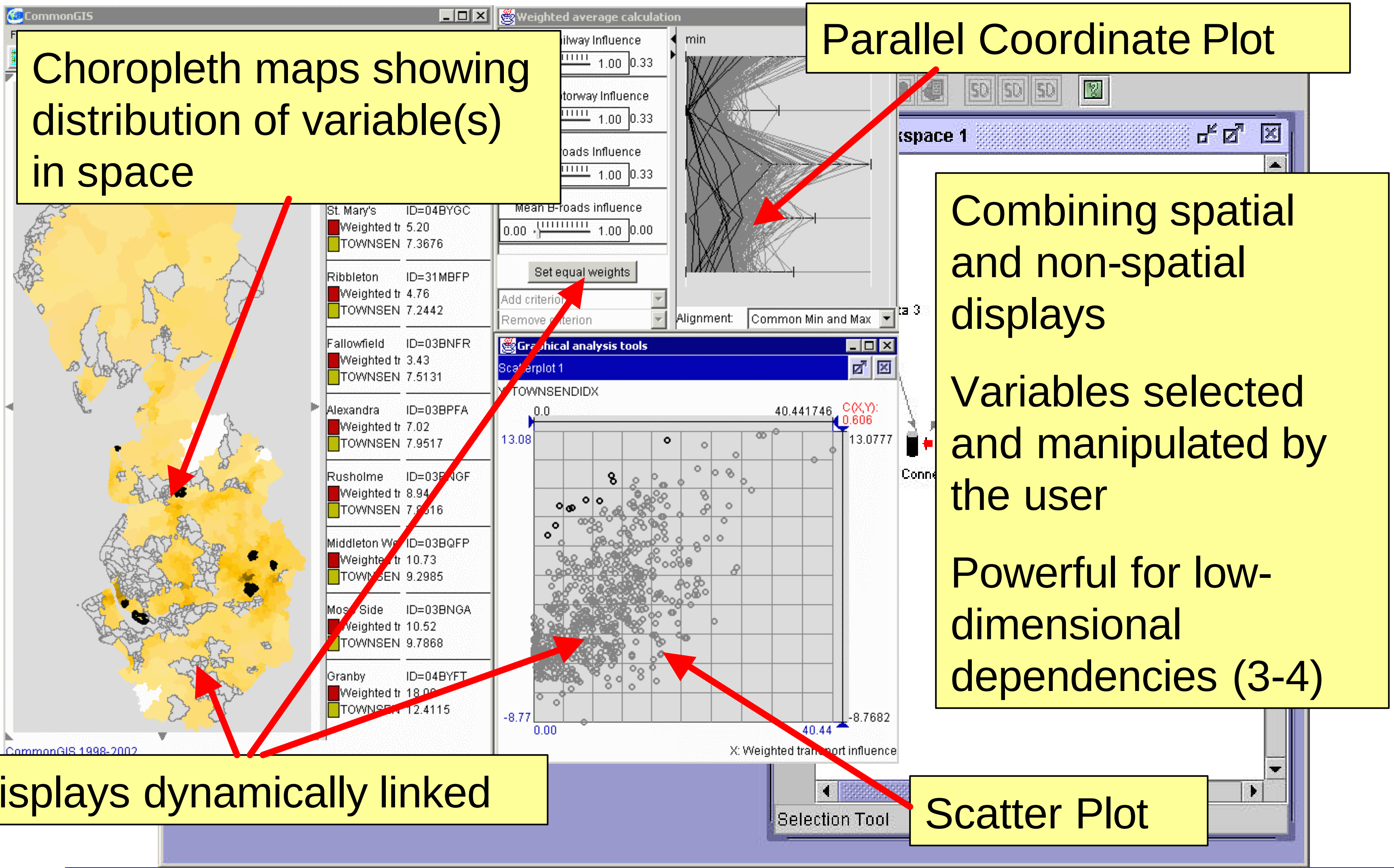
Point & Click-Tool for defining analysis tasks

Workspace Tree

Property editor

2. Visual Exploratory Analyis

Interactive Exploratory Analysis



Choropleth maps showing distribution of variable(s) in space

Parallel Coordinate Plot

Combining spatial and non-spatial displays

Variables selected and manipulated by the user

Powerful for low-dimensional dependencies (3-4)

Displays dynamically linked

Scatter Plot

St. Mary's ID=04BYGC
Weighted tr 5.20
TOWNSEN 7.3676

Ribbleton ID=31MBFP
Weighted tr 4.76
TOWNSEN 7.2442

Fallowfield ID=03BNFR
Weighted tr 3.43
TOWNSEN 7.5131

Alexandra ID=03BPFA
Weighted tr 7.02
TOWNSEN 7.9517

Rusholme ID=03BNGF
Weighted tr 8.94
TOWNSEN 7.9516

Middleton We ID=03BQFP
Weighted tr 10.73
TOWNSEN 9.2985

Moss Side ID=03BNGA
Weighted tr 10.52
TOWNSEN 9.7868

Granby ID=04BYFT
Weighted tr 18.00
TOWNSEN 12.4115

3. Searching for Explanatory

Data Mining Tasks in SPIN!

- Looking for associations between subsets of spatial and non-spatial attributes

  ⇨ **Spatial Association Rules**
- A phenomenon of interest (e.g. death rate) is given but it is not clear which of a large number of spatial and non-spatial attributes is relevant for explaining it

  ⇨ **Spatial Subgroup Discovery**
- A quantitative variable of interest is given and we ask how much this variable changes when one of the relevant independent variables is changed

  ⇨ **Bayesian Local regression**

Subgroup Discovery Search

- Subgroup discovery is a multi-relational approach that searches for probabilistically defined deviation patterns (Klösgen 1996, Wrobel 1997)
- Top-down search search from most general to most specific subgroups, exploiting partial ordering of subgroups ($S_1 \geq S_2$ S_1 *more general* than $S_{2)}$
- Beam search expanding only the *n* best ones at each level of search
- Evaluating hypothesis according to quality function:

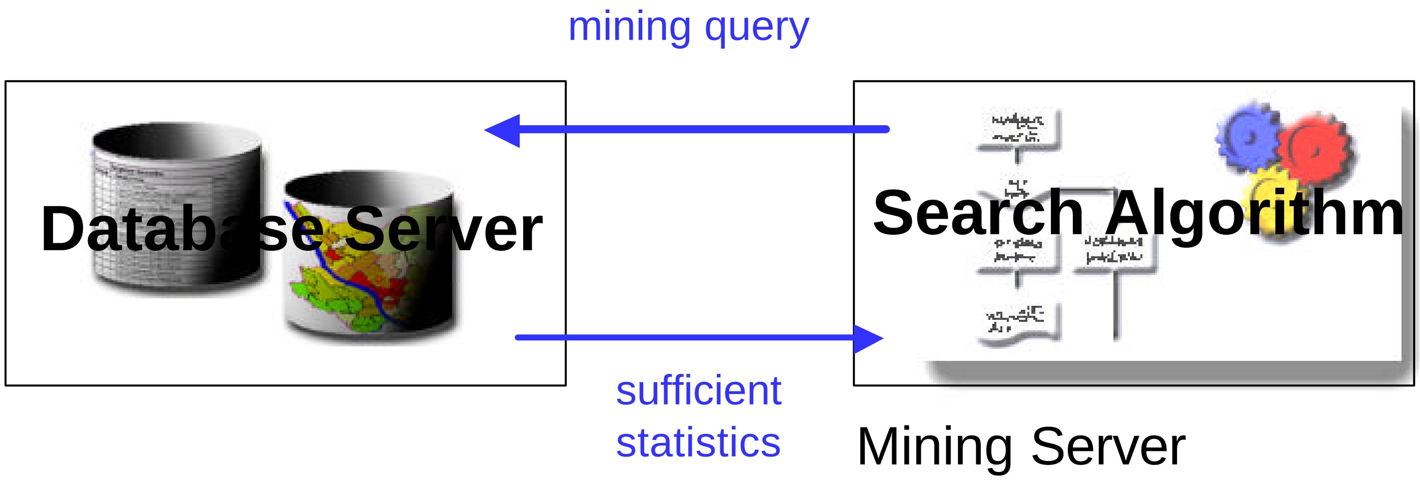
T= target group
C= concept

$$\frac{p(T|C)-p(T)}{\sqrt{p(T)(1-p(T))}}\sqrt{n}\sqrt{\frac{N}{N-n}}$$

T = **`long-term illness=high`**

`C = unemployment=high`

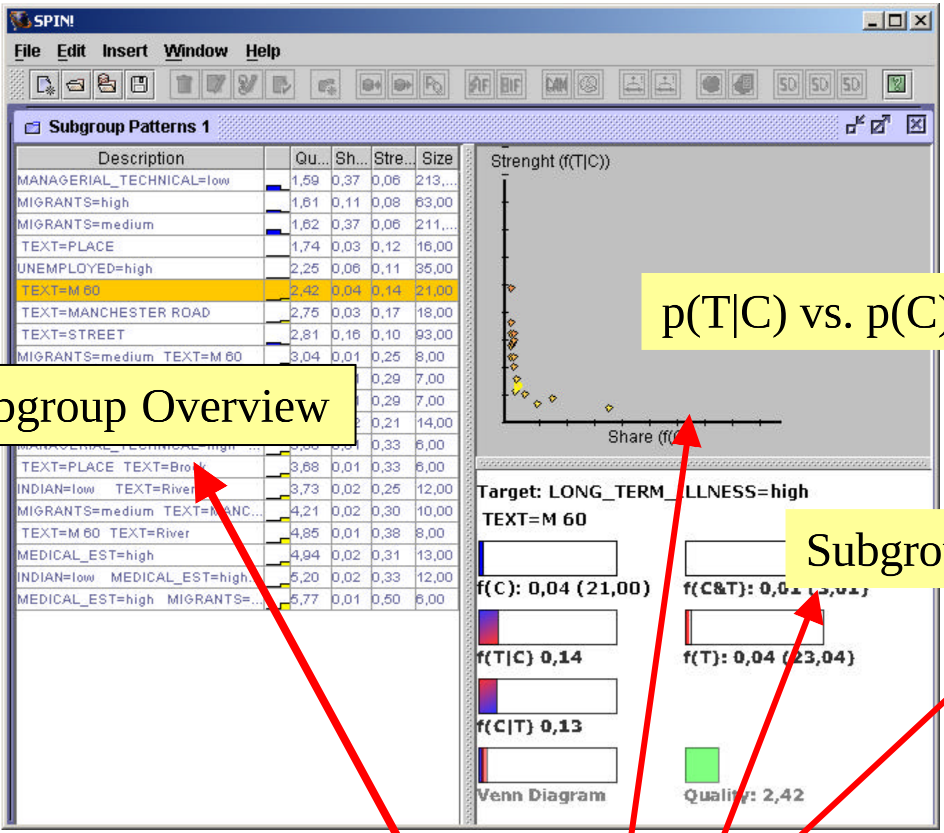
Division of labour between Oracle RDBMS and Search Manager



- Database integration: efficiently organize mining queries
- Mining query delivers statistics (aggregations) sufficient for evaluating **many** hypotheses

- search in hypothesis space
- generation and evaluation of hypotheses (subgroup patterns)

Data Mining visualization



Description		Qu...	Sh...	Stre...	Size
MANAGERIAL_TECHNICAL=low		1,59	0,37	0,06	213,...
MIGRANTS=high		1,61	0,11	0,08	63,00
MIGRANTS=medium		1,62	0,37	0,06	211,...
TEXT=PLACE		1,74	0,03	0,12	16,00
UNEMPLOYED=high		2,25	0,06	0,11	35,00
TEXT=M 60		2,42	0,04	0,14	21,00
TEXT=MANCHESTER ROAD		2,75	0,03	0,17	18,00
TEXT=STREET		2,81	0,16	0,10	93,00
MIGRANTS=medium TEXT=M 60		3,04	0,01	0,25	8,00
[illegible]		[illegible]	[illegible]	0,29	7,00
[illegible]		[illegible]	[illegible]	0,29	7,00
[illegible]		[illegible]	[illegible]	0,21	14,00
MANAGERIAL_TECHNICAL=high ...		[illegible]	[illegible]	0,33	6,00
TEXT=PLACE TEXT=Brook		3,68	0,01	0,33	6,00
INDIAN=low TEXT=River		3,73	0,02	0,25	12,00
MIGRANTS=medium TEXT=MANC...		4,21	0,02	0,30	10,00
TEXT=M 60 TEXT=River		4,85	0,01	0,38	8,00
MEDICAL_EST=high		4,94	0,02	0,31	13,00
INDIAN=low MEDICAL_EST=high...		5,20	0,02	0,33	12,00
MEDICAL_EST=high MIGRANTS=...		5,77	0,01	0,50	6,00

Target: LONG_TERM_ILLNESS=high

TEXT=M 60

f(C): 0,04 (21,00)

f(C&T): 0,01 (3,01)

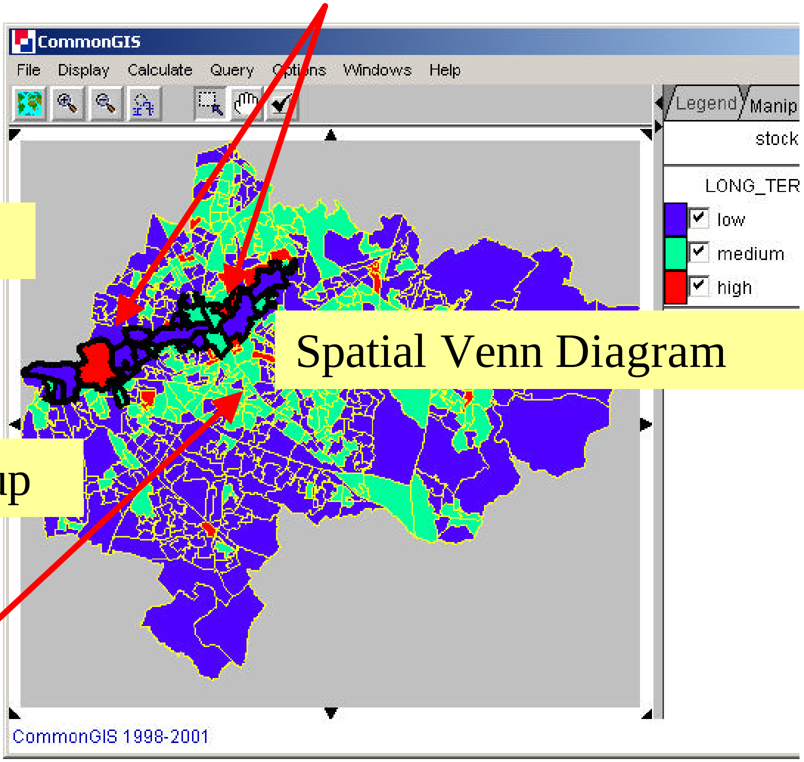
f(T|C) 0,14

f(T): 0,04 (23,04)

f(C|T) 0,13

Venn Diagram

Quality: 2,42



High long-term illness in districts crossed by M60

p(T|C) vs. p(C)

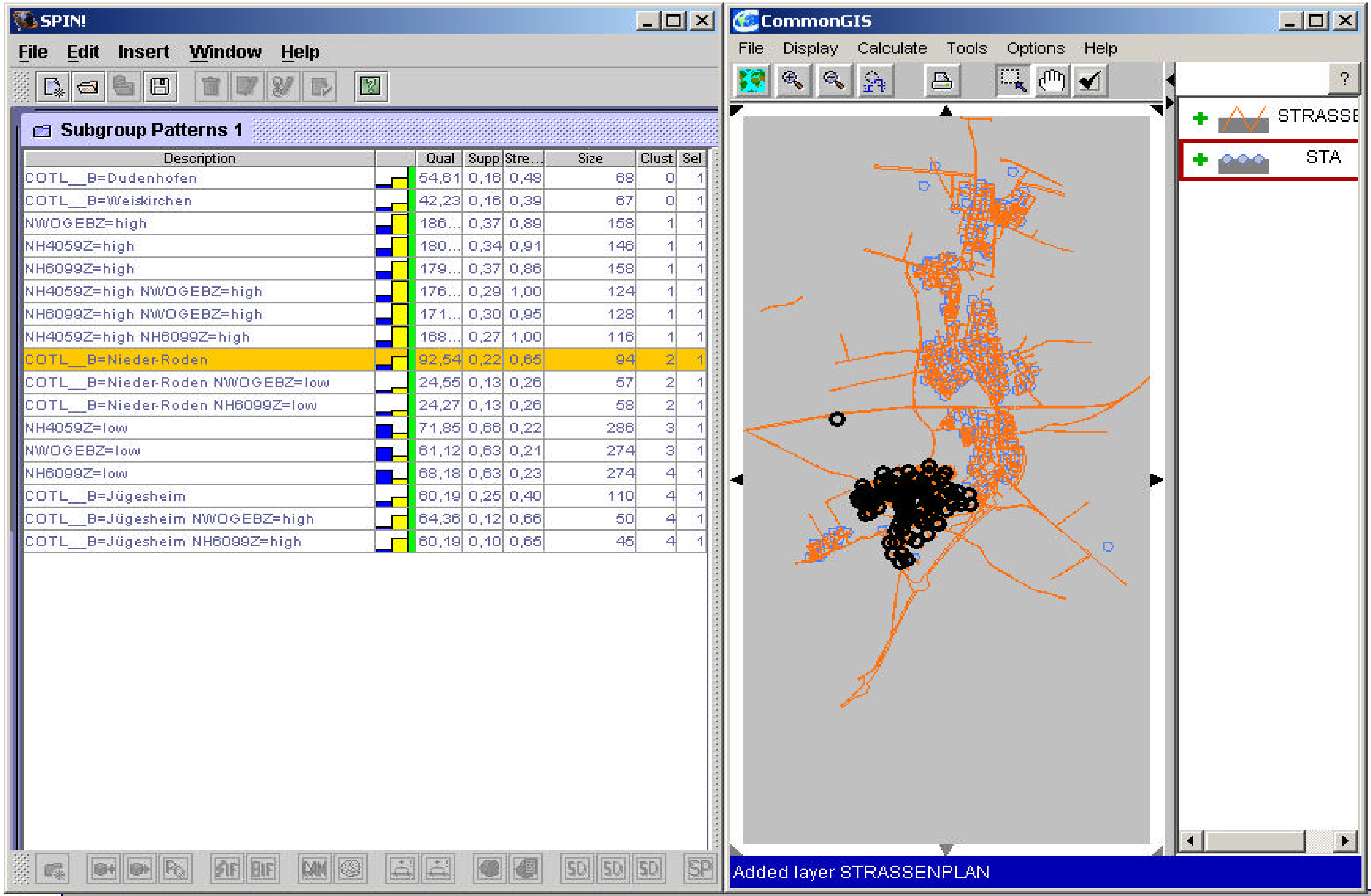
Subgroup Overview

Spatial Venn Diagram

Subgroup

Linked Display

Customer Analysis Rodgau, Germany



SPIN!
File Edit Insert Window Help
CommonGIS
File Display Calculate Tools Options Help
STRASSE
STA

Subgroup Patterns 1

Description		Qual	Supp	Stre...	Size	Clust	Sel
COTL__B=Dudenhofen		54,61	0,16	0,48	68	0	1
COTL__B=Weiskirchen		42,23	0,16	0,39	67	0	1
NWOGEBZ=high		186...	0,37	0,89	158	1	1
NH4059Z=high		180...	0,34	0,91	146	1	1
NH6099Z=high		179...	0,37	0,86	158	1	1
NH4059Z=high NWOGEBZ=high		176...	0,29	1,00	124	1	1
NH6099Z=high NWOGEBZ=high		171...	0,30	0,95	128	1	1
NH4059Z=high NH6099Z=high		168...	0,27	1,00	116	1	1
COTL__B=Nieder-Roden		92,54	0,22	0,65	94	2	1
COTL__B=Nieder-Roden NWOGEBZ=low		24,55	0,13	0,26	57	2	1
COTL__B=Nieder-Roden NH6099Z=low		24,27	0,13	0,26	58	2	1
NH4059Z=low		71,85	0,66	0,22	286	3	1
NWOGEBZ=low		61,12	0,63	0,21	274	3	1
NH6099Z=low		68,18	0,63	0,23	274	4	1
COTL__B=Jügesheim		60,19	0,25	0,40	110	4	1
COTL__B=Jügesheim NWOGEBZ=high		64,36	0,12	0,66	50	4	1
COTL__B=Jügesheim NH6099Z=high		60,19	0,10	0,65	45	4	1

Added layer STRASSENPLAN

System Demo:
Customer Analysis
using
MiningMart and SPIN!

Summary & Outlook

- SPIN! tightly integrates Data Mining analysis and GIS-based visualization
- Main features:
  - A spatial data mining platform
  - New spatial data mining algortihms for subgroup discovery, association rules, Baysian MCMC
  - New visualization methods
- Integration of Spatial Data allows to get results that could not be achieved otherwise
- MiningMart can usefully applied for some pre-processing tasks
- Future tasks: Integrating spatial preprocessing in MiningMart