Chapter 11 Spatial data mining & decision support systems

2. Data mining definition, concepts & techniques

2.1 The origin & nature of data mining

definition - process of extracting interesting & hidden info from DB

several factors for data mining development

- proliferation of DB tech + unprecedented volumes of data
- growing realization that DBs can be used as a basis for knowledge discovery & decision support
- inability of conventional methods of statistical analysis, SQL, OLAP to detect & extract knowledge
- surge of data processing power of computers
- development of DBMS, machine learning, info theory, decision science

DM is an integral part of modern DB tech - critical role in the evolution of DB sys

- combine data management & decision support functionality in a single sys environment
- → referred to as *inductive DB*
- allow collaborative decision support in a distributed network environment

many commercial products support DM - Oracle, IBM DB2, MS SQL Server

DM is differ from conventional SQL & OLAP

- designed for use w/ very large DBs/ data warehouses
- concerned w/ secondary analysis of large datasets to discover unknown knowledge
- follows an inductive strategy of data analysis - gain knowledge progressively w/o any a priori assumption
- focuses on the detection of the characteristics of & correlations among attributes

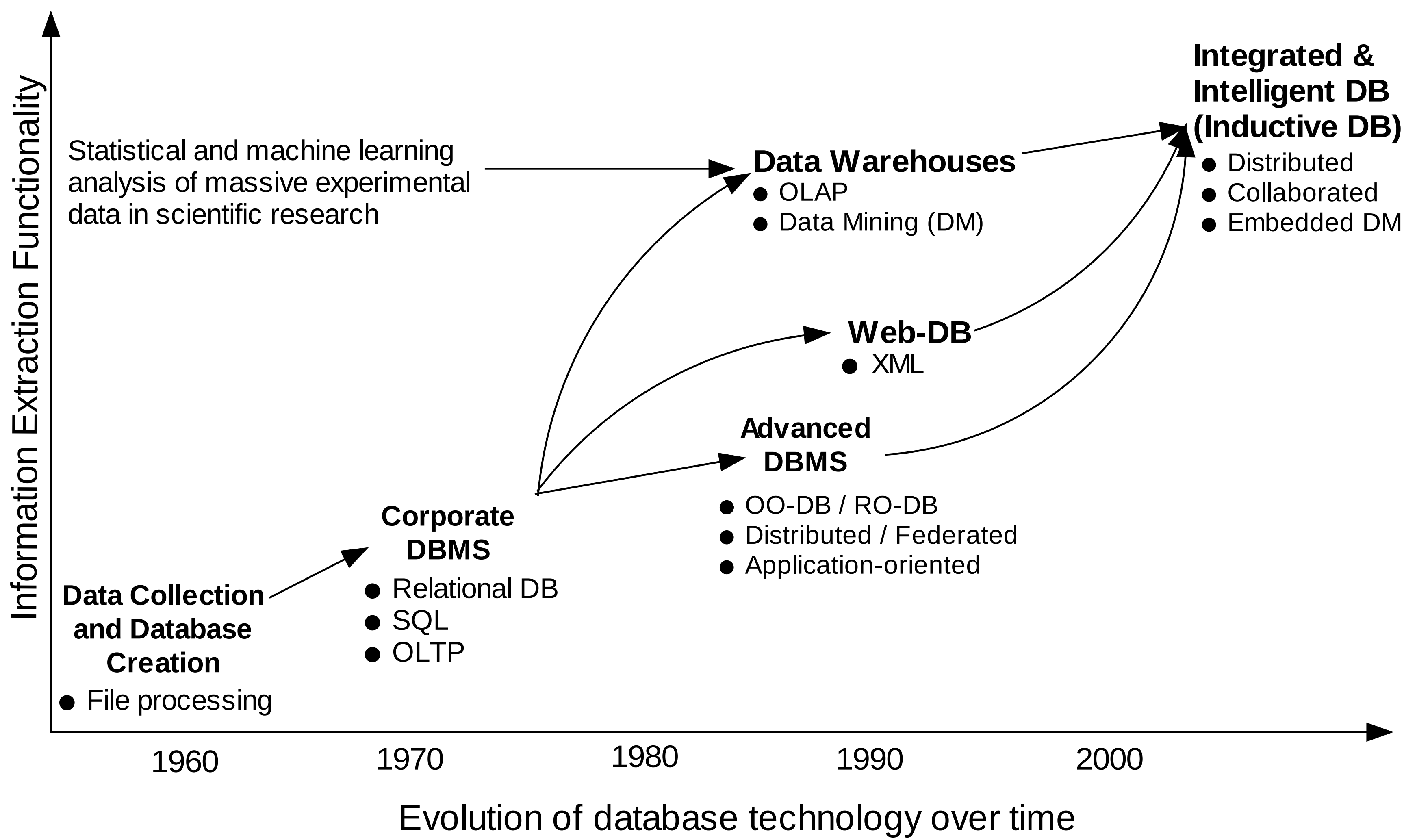


Fig 11-1 The evolution of DB tech from data management to decision support

2.2 Data mining & knowledge discovery in DBs

DM is one of the steps of knowledge discovery in DBs (KDD)

KDD consists of following sequence of steps :

- data integration & cleansing - combine multiple & heterogeneous data sources + rectify
- data selection & transformation - data are retrieved & transformed into a form
- data mining - process of applying machine learning, visualization, statistical analysis
- knowledge discovery & construction - evaluation & interpretation of the extracted info + construction of computerized knowledge base
- deployment - use of results in support of decision making

KDD is an interactive & iterative process

- control DM process by changing input data parameters
- cross-reference knowledge acquired using different mining tech

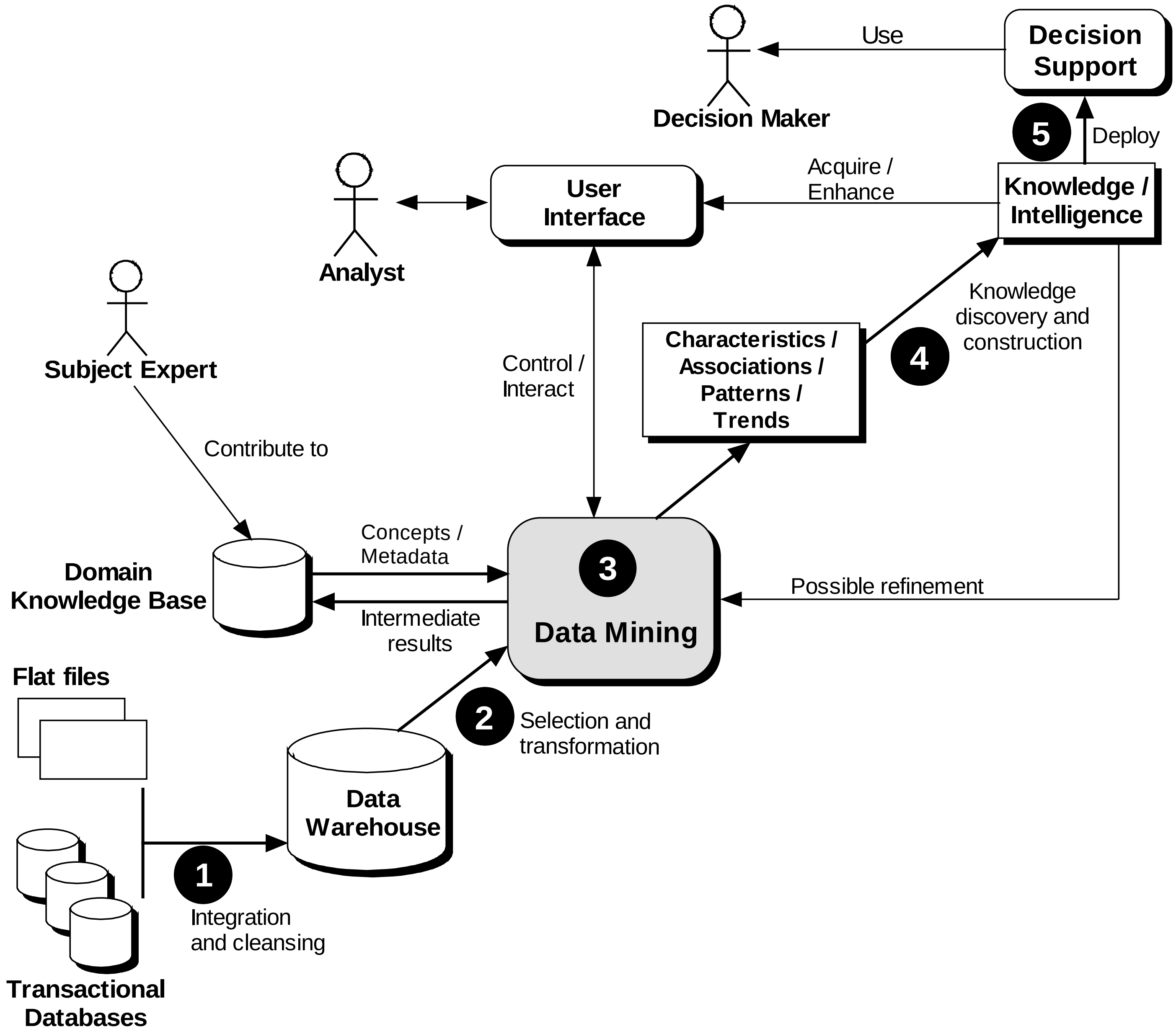


Fig 11-2 The steps of knowledge discovery in DB (KDD)

2.3 Human intelligence in data mining

human intervention is important in the following process :

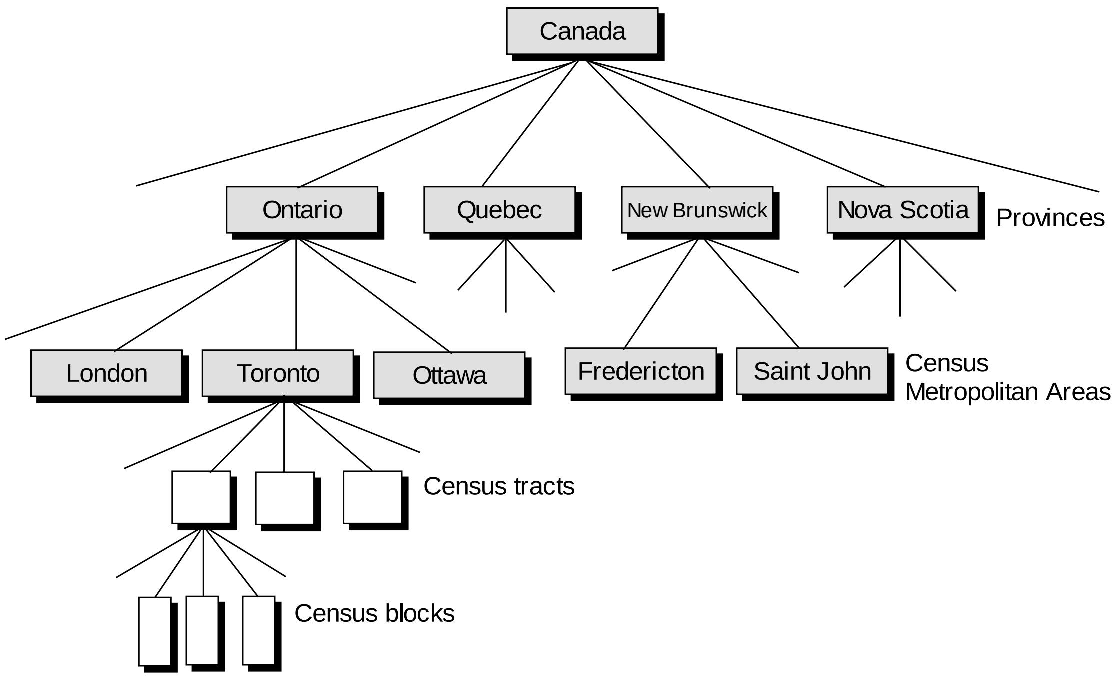
- data preprocessing - determine data usability, clean, transform
- data mining - choice of mining model, mining techniques, underlying algorithms
- knowledge discovery & construction - interface between syntactic knowledge & semantic knowledge
- presenting & visualizing discovered knowledge

concept hierarchy plays an important role in discovering interesting knowledge

- analyst can logically roll up or drill down during the DM process
- also possible to drill across to examine temporal variation within a given level of a concept hierarchy

← CH provides the background knowledge to control the exploration of the dataset at different semantic levels & at different stages of the data mining process

Data_collection_units: census_block<census_tract_CMA<provinces<canada



Annual_family_income: (10,000-19999=A)<(20000-29999=B)<(30000-39999=C)<.....

(a) Schematic concept hierarchy

Annual_family_income: (10,000-19999=A)<(20000-29999=B)<(30000-39999=C)<.....

Years_of_education_head_of_household: (less_than_5=A)<(6-10=B)<(over_10=C)

(b) Sub-grouping hierarchy

Street_address: apt_num<street_num<street_name<city<province<post_code

(c) Operation-driven hierarchy

Below_poverty_line (X): annual_income(X, P1)
and num_people_in_household(X, P2)
and (P1/P2)<CAD$7500

(d) Rule-based hierarchy

Fig 11-3 An example of concept hierarchies in the context of Canada’s census population statistics

2.4 Data mining concepts & techniques

development of DM - ranging from visual interpretation & understanding to algorithmic logic & probability rules

DM techniques - statistical analysis, visualization, machine learning

visualization is made up of 3 elements :

- computation - turn data into graphical images
- cognition - develop mental representation, identify patterns, create order
- graphic design - conceptualization & construction of pictorial displays

machine learning - supervised & unsupervised machine learning

- supervised - predictive data mining, directed toward problem solving, detect patterns & relationships between the independent & dependent variables, build a model of discovered knowledge
- unsupervised - descriptive data mining, exploration oriented, detect aspects of the properties of a dataset

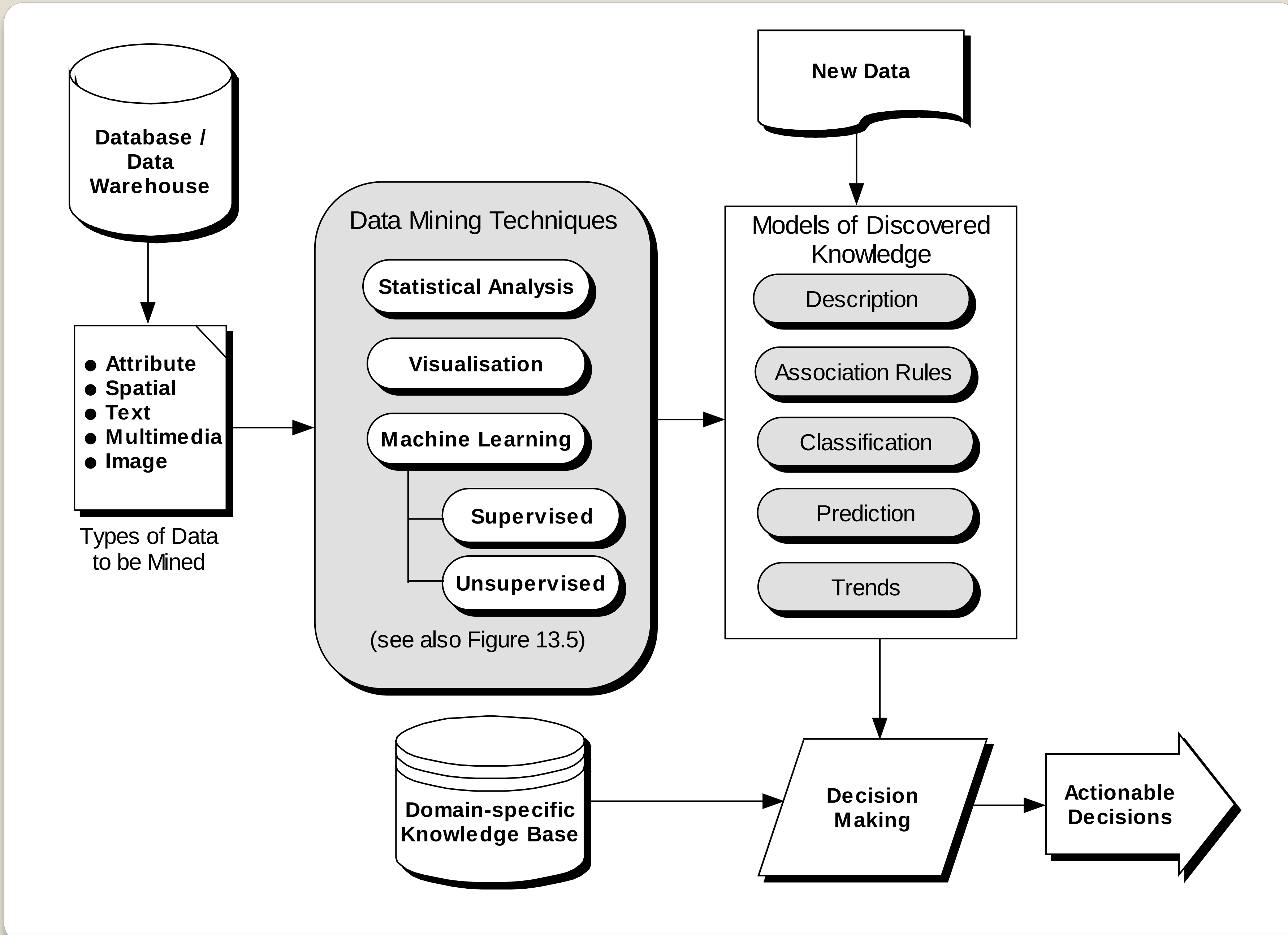


Fig 11-4 The concepts and techniques of data mining

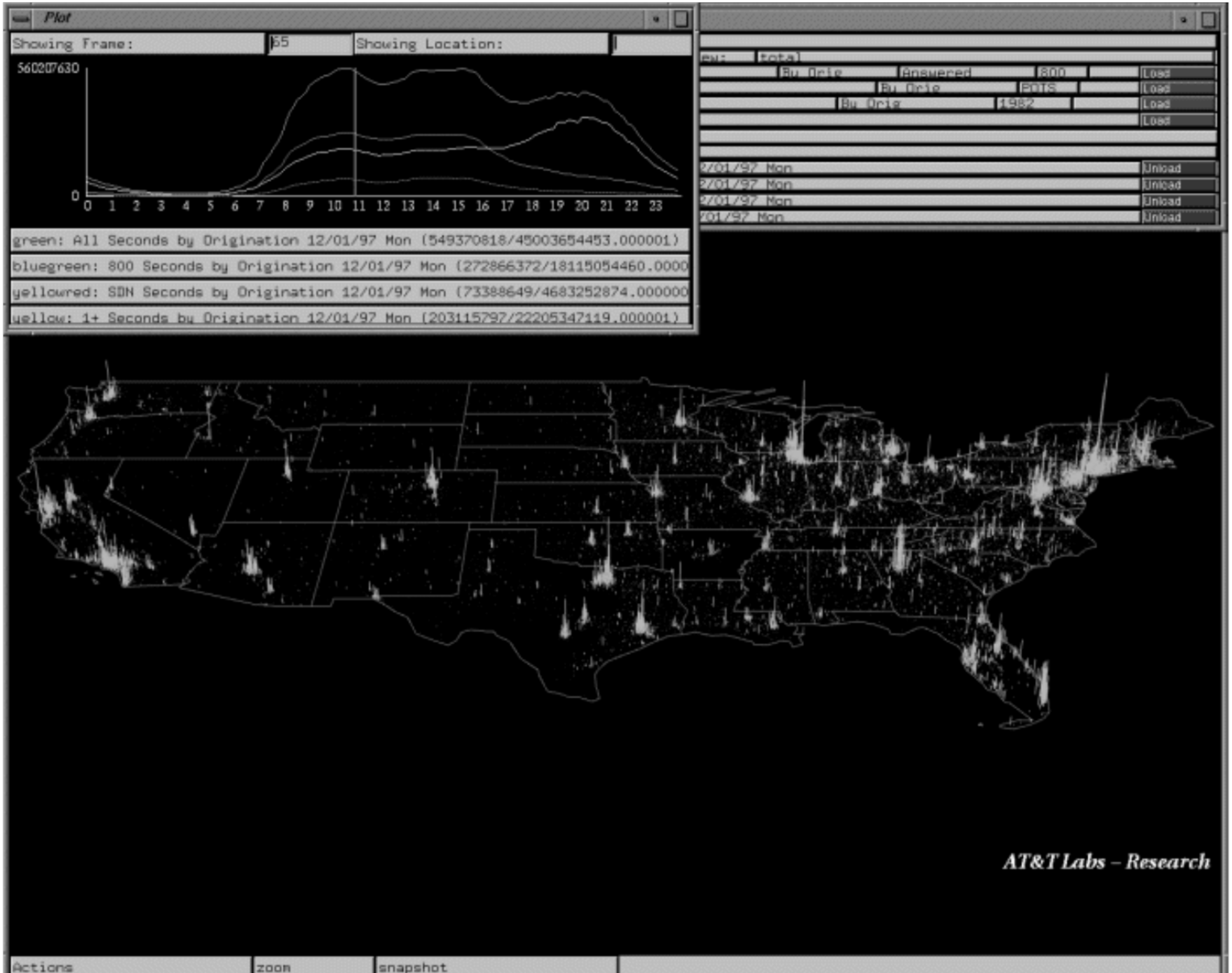


Fig 11-5 A landscape visualization of telephone networks in the US

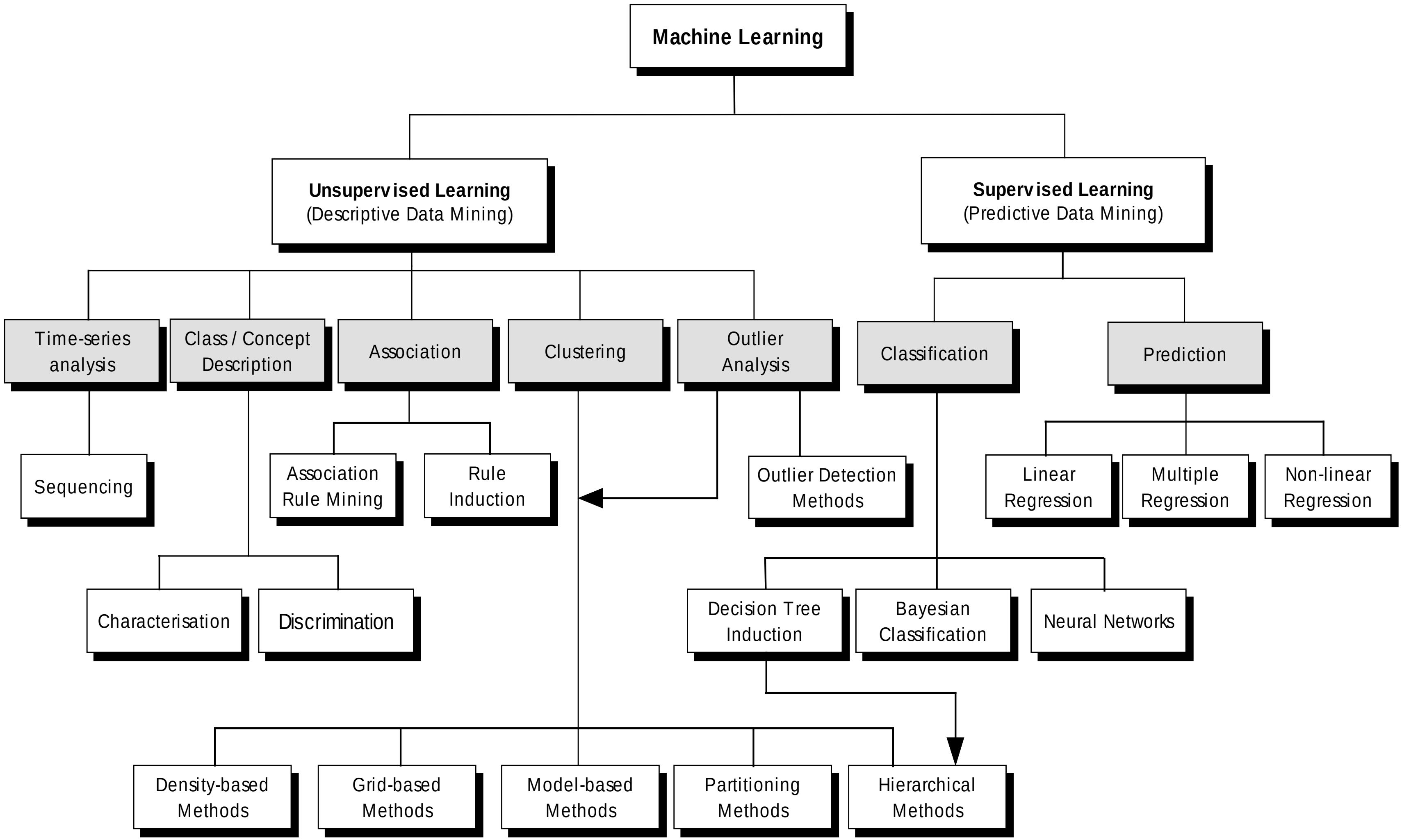


Fig 11-6 Classification of machine learning data mining techniques

2.4.1 Classification

def) grouping of unlabelled data objects into predefined classes

characteristics - a predictive data mining technique

algorithms learn from the training data set & build a classification model

many techniques are proposed :

- decision trees - generated from a training data set in a top down, general to specific direction
- neural networks - analytical tech to predict new observations from known observations
- Bayesian classification - a variety of statistical tech to predict class membership probabilities

2.4.2 Prediction

def) determine possible values of missing data / forecast the values & distribution of attributes

many techniques :

- classification
- ordinary simple linear regression - $Y=\alpha+\beta X+\varepsilon$ (X: independent, Y: dependent, ε: random error)
- ordinary multiple linear regression - more than 1 predictor (X) variable for the response variable (Y)
- non-linear regression - add non linear polynomial / other terms $Y = \alpha + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_3^3 + \varepsilon$

2.4.3 Class/ Concept description

def) a summary of the general properties of individual classes / concepts in a data set

many techniques :

sum, count, average, variance, discrimination, cross-tabulation, chart, graph, maps

2.4.5 Association rule mining (=dependency analysis, linkage analysis)

def) detect correlations among attributes

correlations are expressed by an association rule :

$X \rightarrow Y$ (c%, s%)   X: antecedent, Y: consequence, c%: confidence, s%: support

association can also be expressed as an induction rule :

IF X THEN Y   if event X occurs, then event Y will likely follow

2.4.5 Clustering (=DB segregation)

def) identify clusters / scenarios embeded in a data set

widely used as an unsupervised learning - does not rely on predefined classes

many techniques :

- partitioning methods - develop a partition of the data set under examination
- hierarchical methods - a sequence of partitioning operations, bottom-up / top-down using thresholds
- locality based method - group data objects based on local relationships
  use density/ random distribution statistics
- neural network - clustering using one of the above methods after executing a learning process

2.4.6 Outlier / Deviation analysis

regarded as a special case of clustering, seeks to identify cases of dissimilarity

use regression analysis, other specialized algorithms

2.4.7 Time series analysis (=trend detection)

def) detection of temporal characteristics

detects sequences & subsequences, sequential patterns, periodicities, trends, temporal deviations

3. Spatial data mining (SDM) concepts & techniques

3.1 Characteristics of spatial data mining

challenge of SDM - detect spatial knowledge from the patterns & relationships

SDM is far more complex because of :

- spatial data structure - organized by sophisticated indexing structures & spatial access methods
- spatial data volume - substantial amounts of heterogeneous data
- spatial data collection - by sampling, salient info can be lost due to sample design & interpretation
- spatial dependencies - spatial features are often interrelated / interconnected, hard to discover
- temporality of spatial data - spatial features are often interrelated in time

other factors related to SDM & spatial knowledge :

- SDM techniques - SDM requires geometric computation & spatial operations
- spatial data conceptual models - difficulty to integrate data represented by different models
- different concepts of spatial space & spatial knowledge - Euclidean space vs. non-Euclidean space, interaction is harder

3.2 Spatial concept hierarchies

spatial concept hierarchy provides the knowledge base to drill down & roll up the dataset

an extended spatial data cube that models spatial data warehouse & facilitate OLAP operations on it

much the same as concept hierarchies for attribute-oriented DM w/ additional dimensions :

- attribute dimension - attributes associated w/ locations & geometries
- spatial-to-attribute dimension - primitive level data are spatial but generalization becomes non-spatial
- spatial-to-temporal dimension - generalization becomes non-spatial over time
- spatial-to-spatial dimension - generalization becomes spatial

3 types of measurement of spatial interest :

- numerical measures - apply to numerical data
- classification measures - apply to categorical data
- spatial measures - apply to spatial objects

3.3 Machine learning techniques of spatial data mining

SDM techniques are functional extensions of conventional DM techniques

- using algorithms designed to handle the characteristics & requirements of spatial data
- consists of spatial classification, spatial prediction, spatial class/concept description, spatial association, spatial clustering, spatial outlier analysis, spatial time-series analysis

3.4 Visualization techniques of spatial data mining

provide the most intuitive way of interpreting spatial data & presenting the results

visualization can be used in different phases of SDM :

- pre process - expose extreme / strange attribute values
- DM - display intermediate results, help interpretation & evaluation

2 visualization based approaches to SDM :

- visualization dominant / geography-to-mathematics approach
  - first evaluates the data by visualization → validates results using SDM
- data mining-dominant / mathematics-to-geography approach
  - starts w/ spatial mining methods → uses visualization for an in-depth analysis

recent researches by Guo(2003) - human centered SDM environment, computation + visualization

- interactive feature selection method for identifying interesting, multi dimensional subspaces
- interactive, hierarchical clustering method for searching multiviariate clusters of arbitrary shape
- suite of coordinated visualization & computational components

3.5 Implementation issues of spatial data mining

reference model of implementing SDM by CRISP-DM (Fig 11-7)

** CRISP-DM : Cross-Industry Standard Process for Data Mining*

provides a framework for carrying out DM projects

life cycle of a SDM project contains 6 phases :

business understanding - identify the objectives → convert into SDM problem definition

data understanding - identify data source, quality, usability

data preparation - obtain one / more spatial data sets, data cleaning

modeling - SDM phase, needs expertise & skills in SDM techniques

evaluation - identify knowledge of real interest to the user, examine results against the objectives

deployment - discovered knowledge are organized & presented in a way the user can use

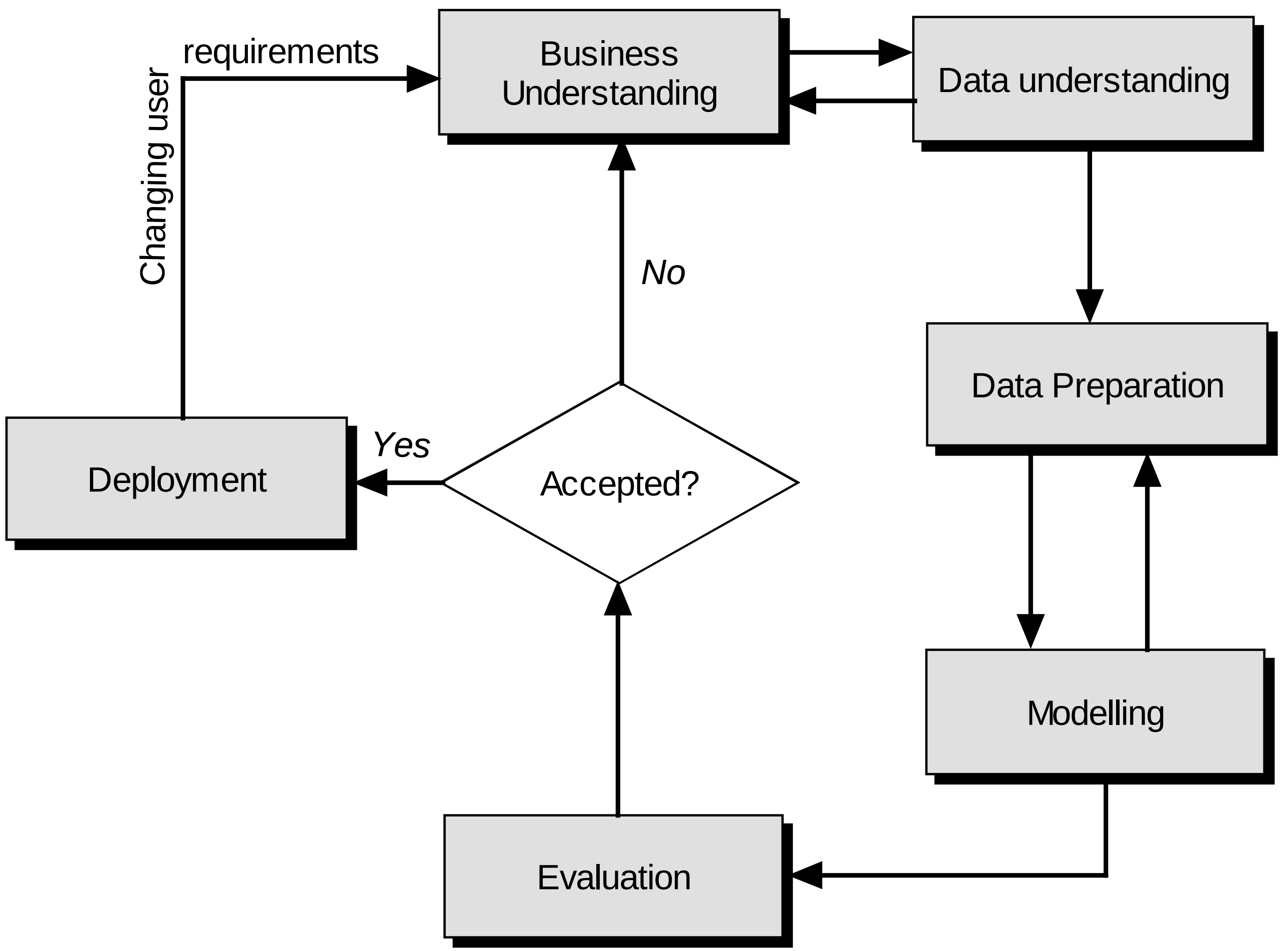


Fig 11-7 The CRISP-DM reference model

4. Spatial decision support concepts, system components & application

4.1 The phases of decision making

structured vs. unstructured decision problems :

- structured - involve routine & repetitive processes
- unstructured - multi faceted & have no clear cut solution

5 sequential phases (Fig 11-8) :

- intelligence - identify & refine a decision problem
- design - create a model of decision problem by refining & constructing relationships between decision
  - components, criteria are set for evaluating alternatives
  - DM, OLAP, ROLAP can be used to structure decision alternatives
- choice - selection of a solution to the model
- implementation
- monitoring

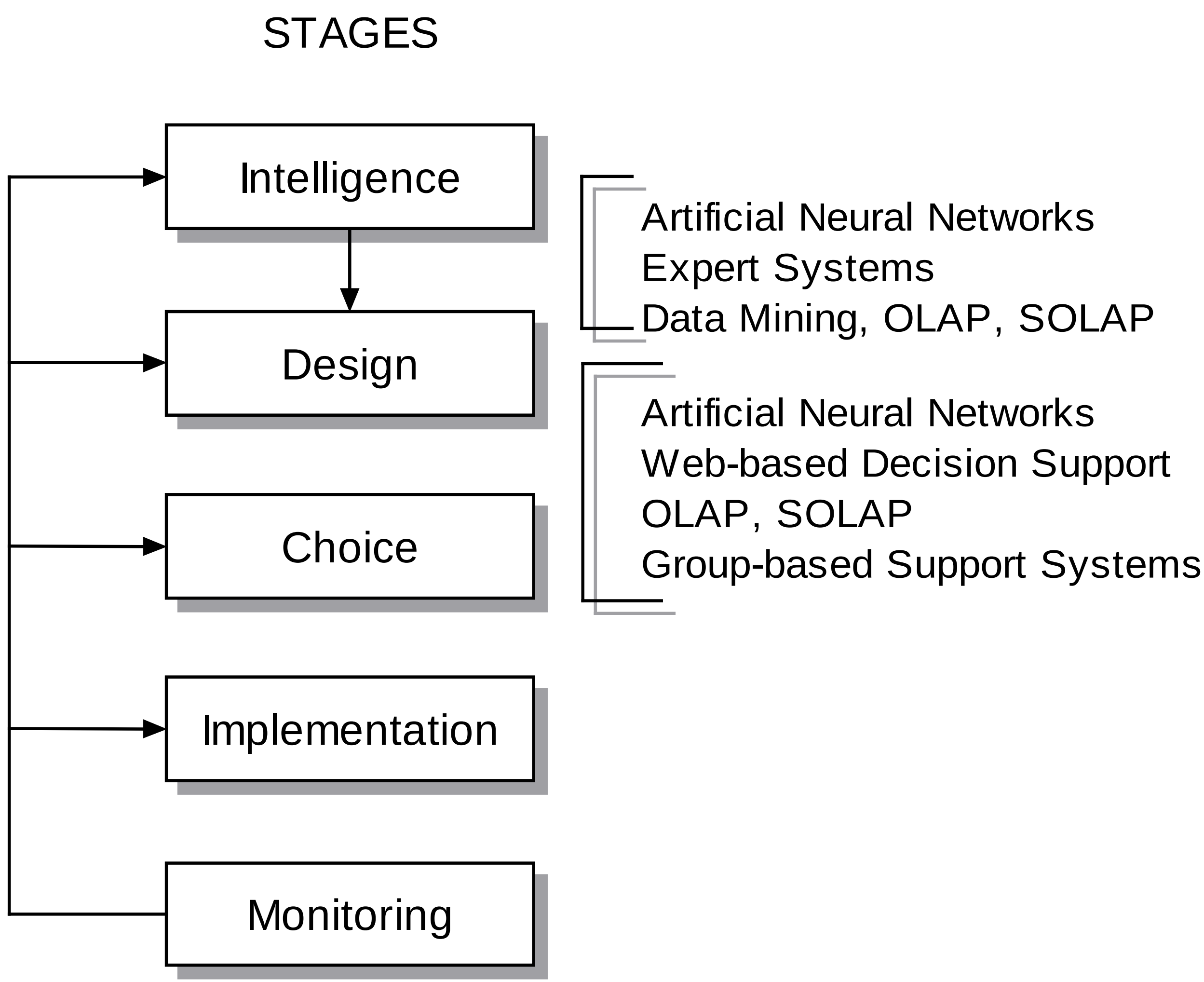


Fig 11-8 Phases of decision making

4.2 Characteristics of decision support systems

a core set of characteristics :

- it is a methodology
- it is computer-based
- it uses data (in DB) that relate to a particular problem domain
- it often includes multiple models & techniques
- it has an easy-to-use graphical user interface
- it must be capable of expressing the decision maker’s own idea
- it is typically iterative & highly interactive
- it supports all phases of the decision making process
- it can be used by a single user in a stand-alone environment / networked to run across the internet

4.3 Decision support system components

core components of DSS (Fig 11-9) :

user interface + knowledge based sub systems + model management sub system

+ external models + DB management sub system

DB management sub system functions :

supports entry, extraction, update, integration, retrieval of data

manages data dictionary & meta data entries

facilitates spatial & other analyses thru interaction w/ a model base / analysis tool box

4 types of models w/ a DSS model base :

strategic models - support mid to long range decision making

tactical decision support models - related to shorter term decision making

operational decision models - support short term / day-to-day issues

analytic models - cut cross aspects of strategic, tactical, operational decision analysis

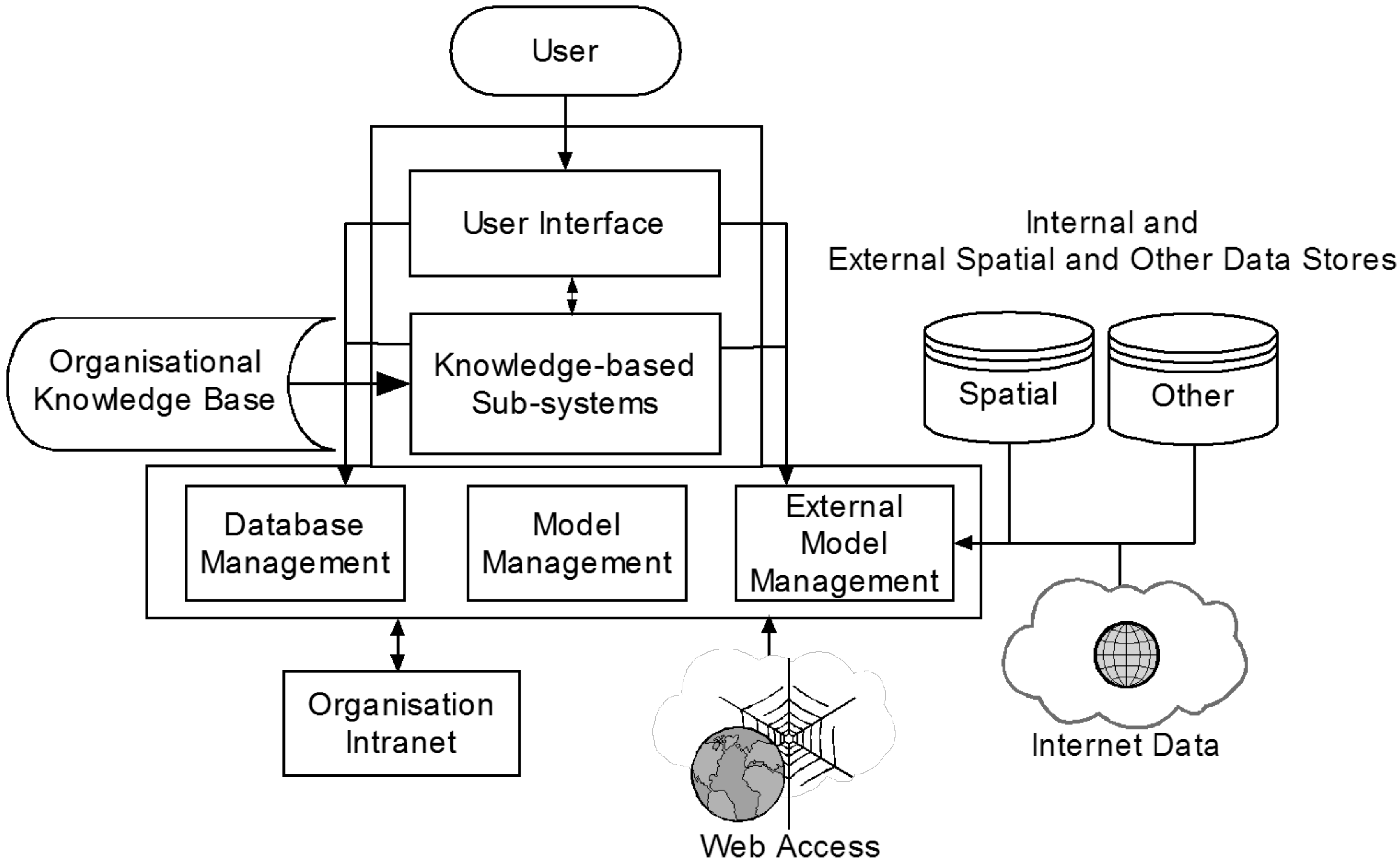


Fig 11-9 High-level decision support system components

4.4 Web-based & web-enabled decision support systems

3 primary forms of DSS architecture characterize web environments :

thin client - client browser simply provides universal access to the info infrastructure

fat client - transfer small parts of the processing load to the client computer thru the use of Java applets, Active X controls, browser plug-ins

distributed approach - manages aspects of the DSS components installed across multiple web, application & DB servers using one/ some combination of CORBA, COM, Java remote method invocation(RMI)

implementation issues

most commercial web-based DSS are designed a 3 / higher tiered architecture (Fig 11-10)

3 tier model can be expanded in the 2nd tier by creating multiple layers - each perform specialized DSS

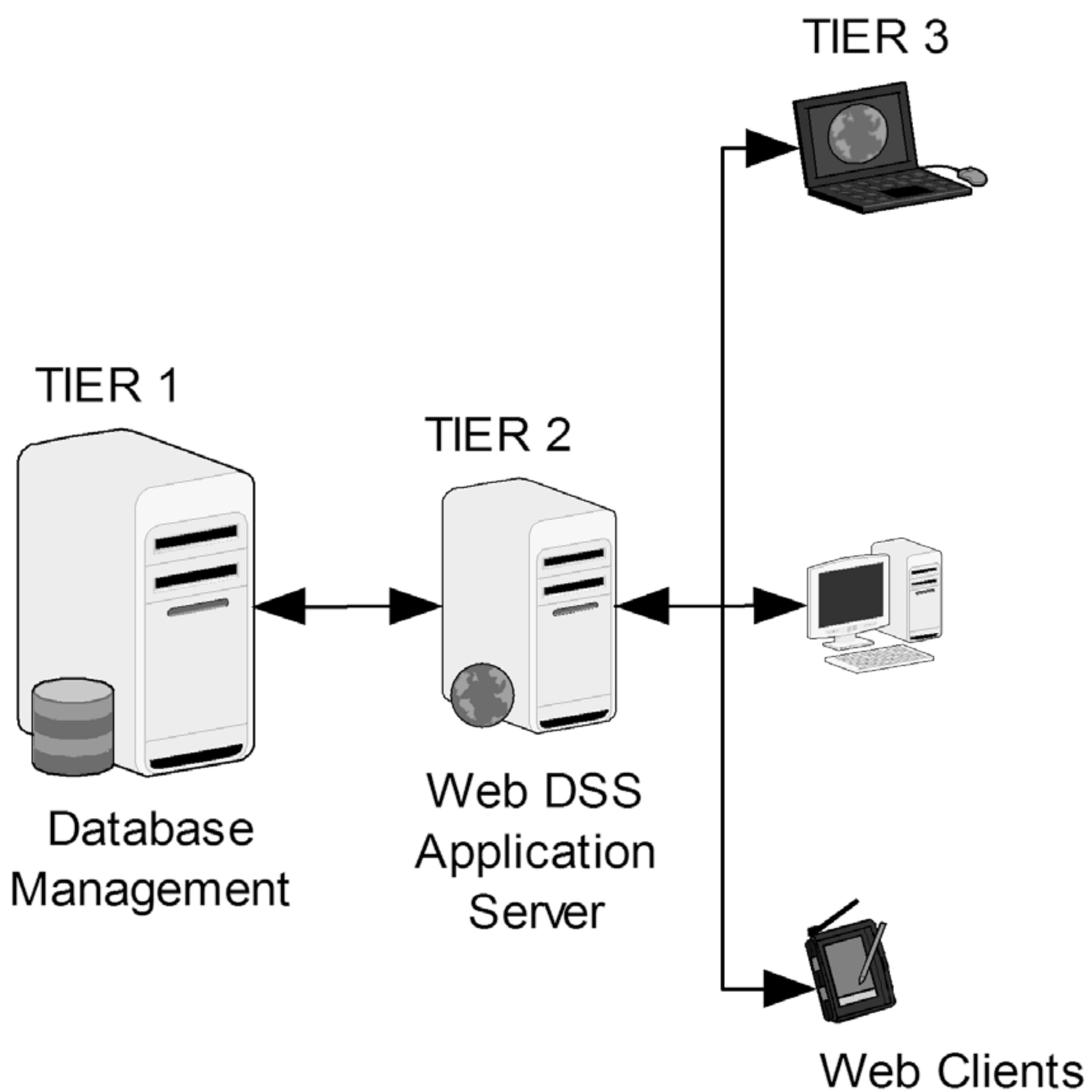


Fig 11-10 Three tier web-based DSS architecture

5. Spatial decision support system applications

5.1 Spatial decision support systems

numerous SDSS have been developed :

web-based SDSS, collaborative SDSS, spatial knowledge-based SDSS, environmental SDSS, group SDSS

5.2 Spatial decision support system applications

5.2.1 Stand-alone spatial decision support for multiple participants

5.2.2 Decision support w/ spatial on-line analytic processing in a web environment

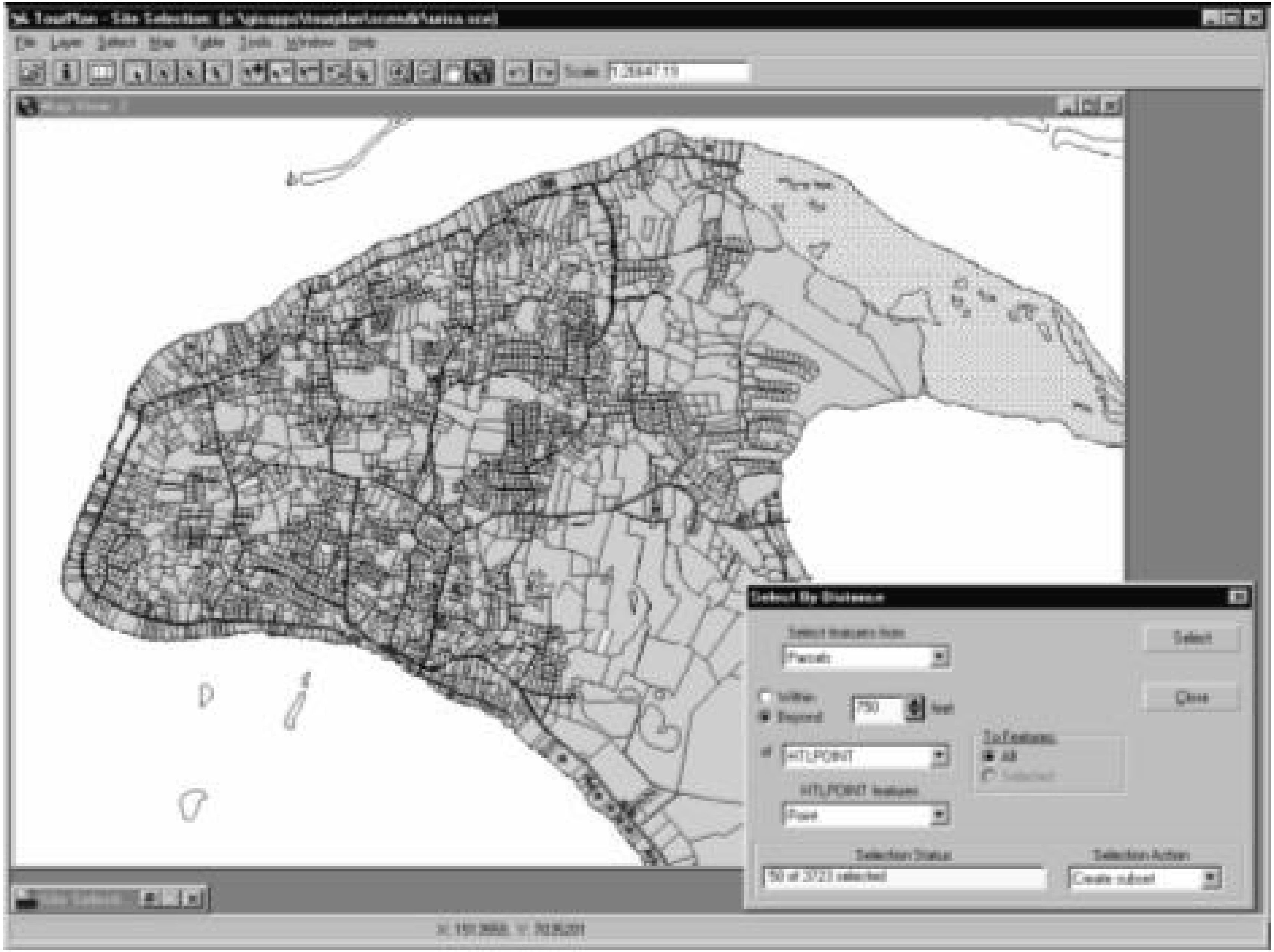


Fig 11-11 Selection of cadastral parcels by distance criteria

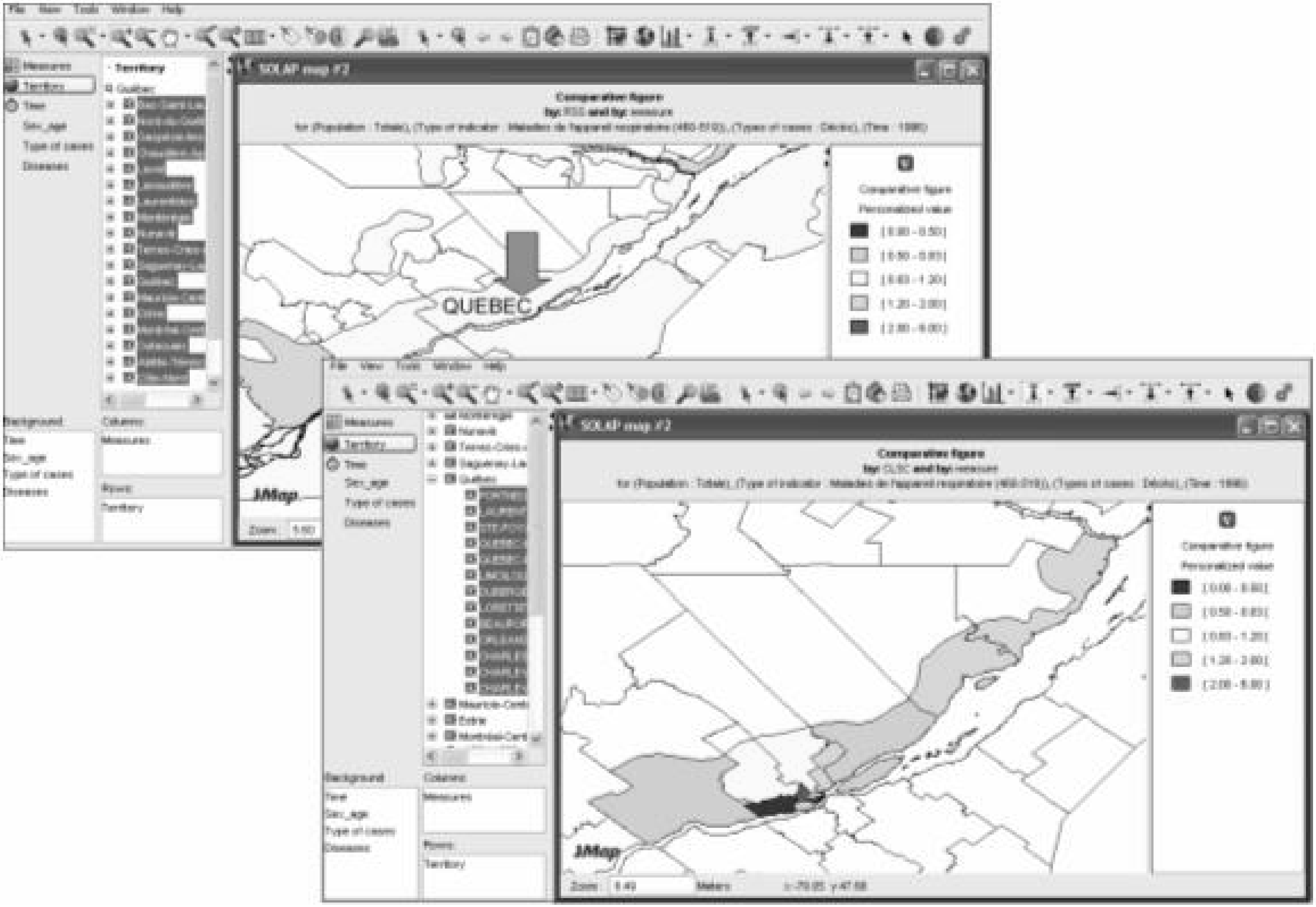


Fig 11-12 Example of a spatial drill down operation from regional to local level

Fig 11-13 Temporal drill across operation with resulting table of all other elements at the same level of detail