Details on Datamining

- knowledge discovery is nontrivial extraction of implicit information from data

- includes set of facts F (data), language L and measure of certainty C

- a pattern is a statement S in L describing relations among subset of F with certainty C such that S is simpler than enumeration of all facts in subset of F

- a pattern that is interesting (user-imposed interest measure) and certain enough (according to user's criteria) = knowledge

- the output of a program that monitors facts in a database = patterns = discovered knowledge

Automating Discovery

- example: DB-LEARN: attribute-oriented generalization to discover new and non-trivial information
- idea: repeatedly replace specific attribute values in a relation with more general concepts
- then: certain tuples are redundant and eliminated
- each attribute has an associated concept hierarchy, defined by a domain expert
- thresholds set by users: max of distinct values for attribute permitted in a generalized relation; max of tuples permitted in a generalized relation
- example:: 10 percent Saskatchewan, 20 percent Manitoba and 25 percent Alberta becomes: 55 percent Prairies
- technique: learning algorithm, substituting higher level values