Lecture 28

Knowledge Discovery in Databases (KDD) and Data Mining

Tuesday, December 04, 2001

William H. Hsu

Department of Computing and Information Sciences, KSU

http://www.cis.ksu.edu/~bhsu

Readings:

Handout, “Data Mining with MLC++”, Kohavi *et al*

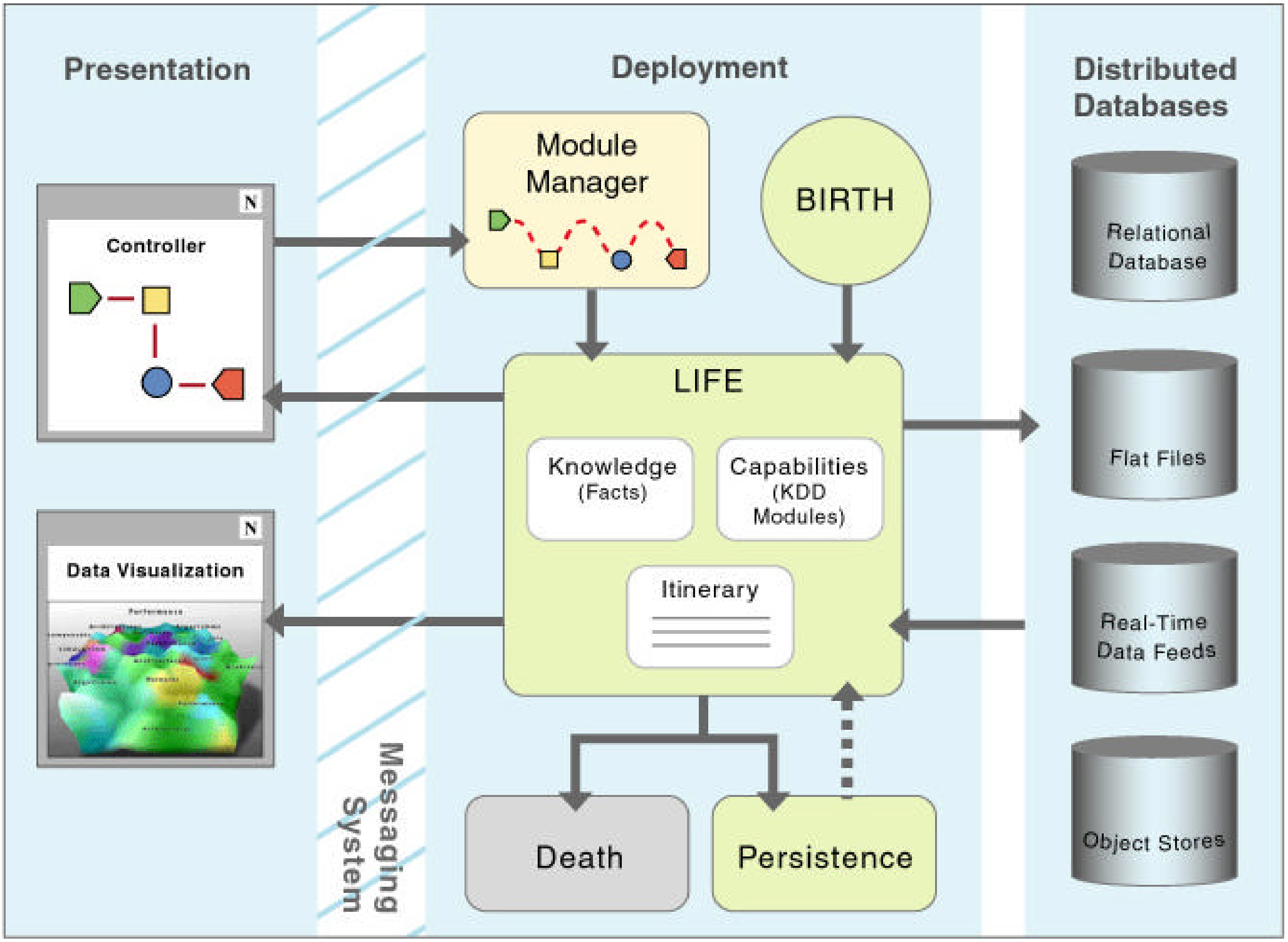
Lecture Outline

- **Readings: “Data Mining with *MLC*++”, Kohavi *et al***
- **Final Exam**
  - **Format**
    - **Open book**
    - **110 minutes**
    - **10 questions (see format online)**
  - **Sample questions online**
- **Knowledge Discovery in Databases (KDD) and Data Mining**
  - **Problem framework (stages)**
  - **Design and implementation issues**
- **Role of Machine Learning and Inference in Data Mining**
  - **Unsupervised learning**
  - **Supervised learning**
  - **Decision support (information retrieval, prediction, policy optimization)**
- **Next Lecture: Final Review Session**

What Is Data Mining?

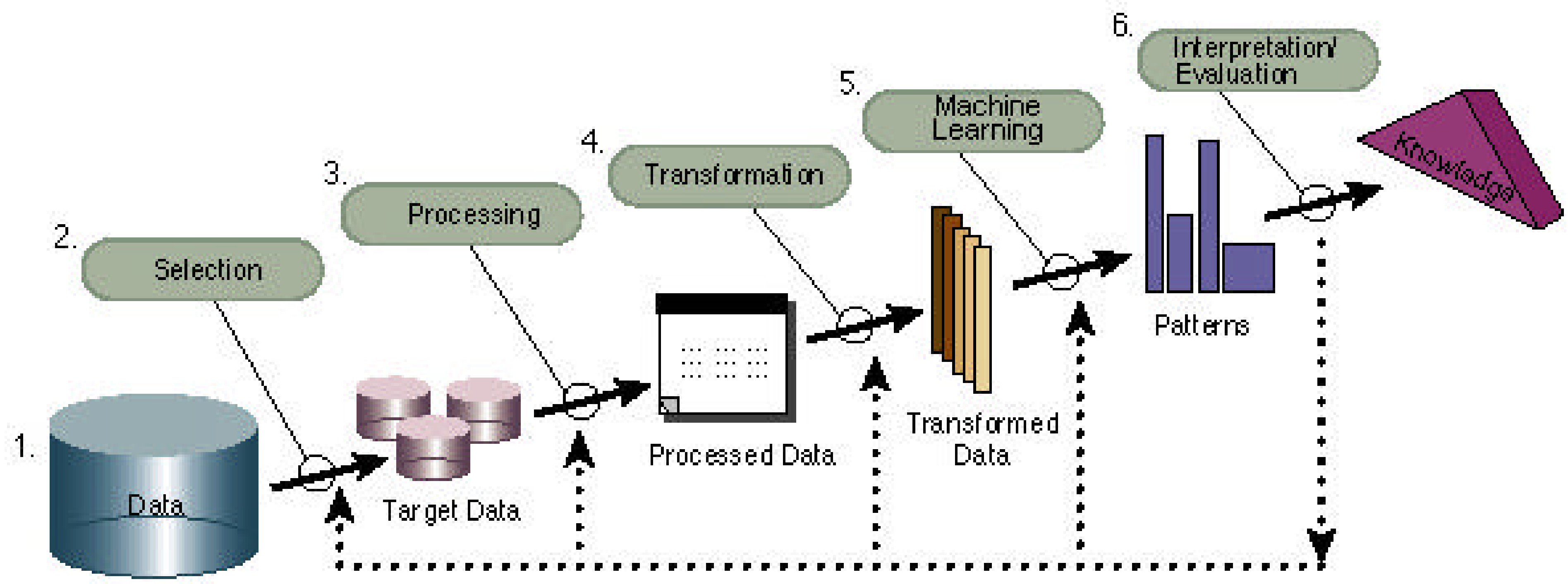
- Two Definitions (FAQ List)
  - The process of automatically extracting valid, useful, previously unknown, and ultimately comprehensible information from large databases and using it to make crucial business decisions
  - *"Torturing the data until they confess"*
- Data Mining: An Application of Machine Learning
  - Guides and integrates learning (model-building) processes
    - Learning methodologies: supervised, unsupervised, reinforcement
    - Includes preprocessing (data cleansing) tasks
    - Extends to pattern recognition (inference or *automated reasoning*) tasks
  - Geared toward such applications as:
    - Anomaly detection (fraud, inappropriate practices, intrusions)
    - Crisis monitoring (drought, fire, resource demand)
    - Decision support
- What Data Mining Is *Not*
  - Data Base Management Systems: *related but not identical field*
  - "Discovering objectives": still need to *understand performance element*

KDD and Software Engineering



Rapid KDD Development Environment

Stages of Data Mining

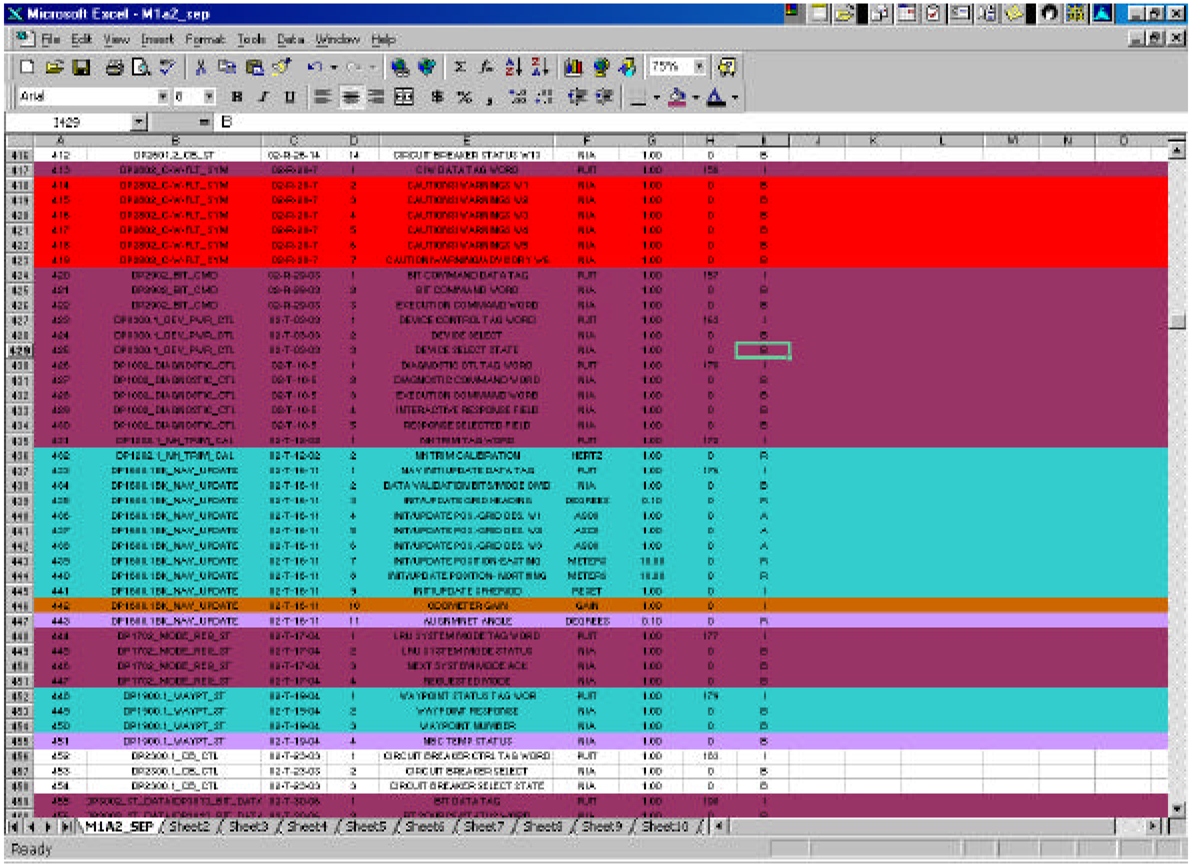


An Overview of the Steps That Compose the KDD Process

Databases and Data Mining

- Database Engineering ? Data Mining!
  - Database design and engineering
    - Data Base Management System (DBMS): computational system that supports efficient *organization*, *retrieval*, and *processing* of data
    - Data warehouse: repository of integrated information for queries, analysis
  - Data mining
    - Often an *application* of DBMS and data warehousing systems
    - Includes inductive model building (learning), pattern recognition, inference
- Selection
  - Guides and integrates learning (model-building) processes
  - Learning methodologies: supervised, unsupervised, reinforcement
  - Includes preprocessing (data cleansing), pattern recognition and inference
- Online Analytical Processing (OLAP)
  - Efficient collection, storage, manipulation, reproduction of multidimensional data
  - Objective: analysis (e.g., for decision support)
  - See: http://perso.wanadoo.fr/bernard.lupin/english/glossary.html

Data Integrity and Data Modeling: Ontologies

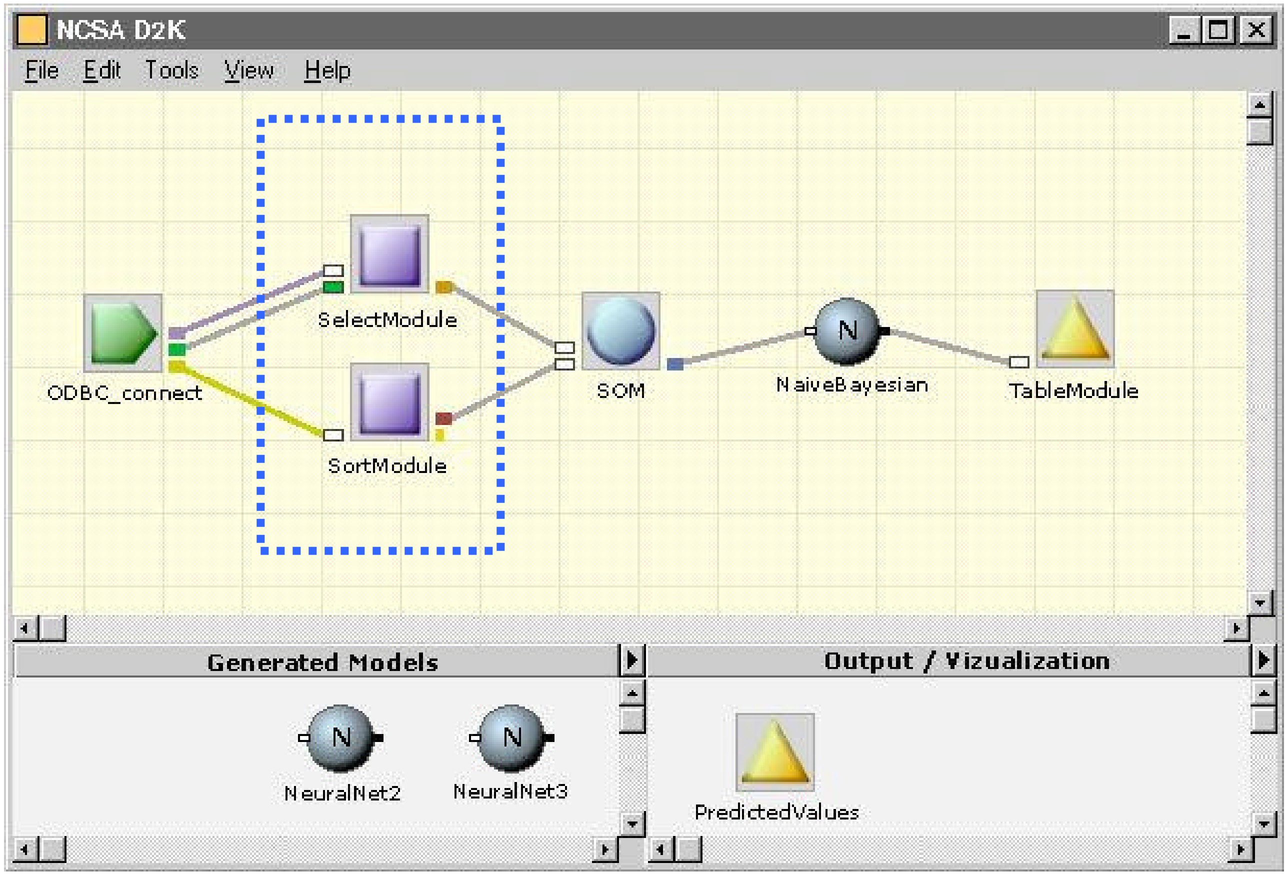

Microsoft Excel - M1a2_sep
File Edit View Insert Format Tools Data Window Help
Arial
75%
Ready

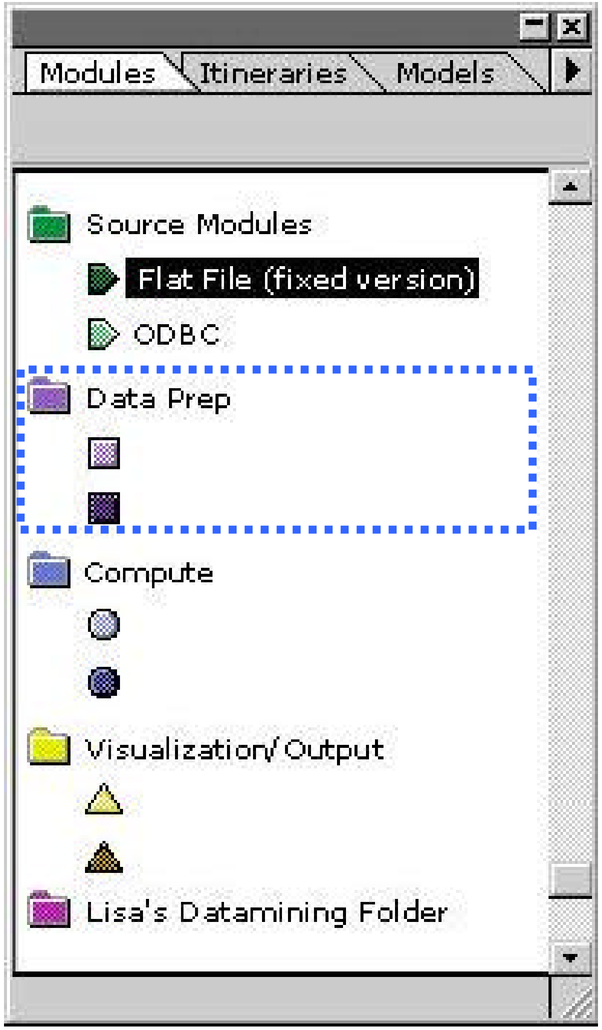
I429 = B

	A	B	C	D	E	F	G	H	I
416	412	DP2801.2_CB_ST	02-R-28-14	14	CIRCUIT BREAKER STATUS W13	N/A	1.00	0	B
417	413	DP2802_C-W-FLT_SYM	02-R-28-7	1	C/W DATA TAG WORD	PUT	1.00	156	I
418	414	DP2802_C-W-FLT_SYM	02-R-28-7	2	CAUTIONS/WARNINGS W1	N/A	1.00	0	B
419	415	DP2802_C-W-FLT_SYM	02-R-28-7	3	CAUTIONS/WARNINGS W2	N/A	1.00	0	B
420	416	DP2802_C-W-FLT_SYM	02-R-28-7	4	CAUTIONS/WARNINGS W3	N/A	1.00	0	B
421	417	DP2802_C-W-FLT_SYM	02-R-28-7	5	CAUTIONS/WARNINGS W4	N/A	1.00	0	B
422	418	DP2802_C-W-FLT_SYM	02-R-28-7	6	CAUTIONS/WARNINGS W5	N/A	1.00	0	B
423	419	DP2802_C-W-FLT_SYM	02-R-28-7	7	CAUTION/WARNING/ADVISORY W6	N/A	1.00	0	B
424	420	DP2902_BIT_CMD	02-R-29-03	1	BIT COMMAND DATA TAG	PUT	1.00	157	I
425	421	DP2902_BIT_CMD	02-R-29-03	2	BIT COMMAND WORD	N/A	1.00	0	B
426	422	DP2902_BIT_CMD	02-R-29-03	3	EXECUTION COMMAND WORD	N/A	1.00	0	B
427	423	DP0300.1_DEV_PWR_CTL	02-T-03-03	1	DEVICE CONTROL TAG WORD	PUT	1.00	163	I
428	424	DP0300.1_DEV_PWR_CTL	02-T-03-03	2	DEVICE SELECT	N/A	1.00	0	B
429	425	DP0300.1_DEV_PWR_CTL	02-T-03-03	3	DEVICE SELECT STATE	N/A	1.00	0	B
430	426	DP1002_DIAGNOSTIC_CTL	02-T-10-5	1	DIAGNOSTIC CTL TAG WORD	PUT	1.00	178	I
431	427	DP1002_DIAGNOSTIC_CTL	02-T-10-5	2	DIAGNOSTIC COMMAND WORD	N/A	1.00	0	B
432	428	DP1002_DIAGNOSTIC_CTL	02-T-10-5	3	EXECUTION COMMAND WORD	N/A	1.00	0	B
433	429	DP1002_DIAGNOSTIC_CTL	02-T-10-5	4	INTERACTIVE RESPONSE FIELD	N/A	1.00	0	B
434	430	DP1002_DIAGNOSTIC_CTL	02-T-10-5	5	RESPONSE SELECTED FIELD	N/A	1.00	0	B
435	431	DP1202.1_NH_TRIM_CAL	02-T-12-02	1	NH TRIM TAG WORD	PUT	1.00	172	I
436	432	DP1202.1_NH_TRIM_CAL	02-T-12-02	2	NH TRIM CALIBRATION	HERTZ	1.00	0	R
437	433	DP1601.1BK_NAV_UPDATE	02-T-16-11	1	NAV INIT/UPDATE DATA TAG	PUT	1.00	176	I
438	434	DP1601.1BK_NAV_UPDATE	02-T-16-11	2	DATA VALIDATION BITS/MODE CMD	N/A	1.00	0	B
439	435	DP1601.1BK_NAV_UPDATE	02-T-16-11	3	INIT/UPDATE GRID HEADING	DEGREES	0.10	0	R
440	436	DP1601.1BK_NAV_UPDATE	02-T-16-11	4	INIT/UPDATE POS.-GRID DES. W1	ASCII	1.00	0	A
441	437	DP1601.1BK_NAV_UPDATE	02-T-16-11	5	INIT/UPDATE POS.-GRID DES. W2	ASCII	1.00	0	A
442	438	DP1601.1BK_NAV_UPDATE	02-T-16-11	6	INIT/UPDATE POS.-GRID DES. W3	ASCII	1.00	0	A
443	439	DP1601.1BK_NAV_UPDATE	02-T-16-11	7	INIT/UPDATE POSITION-EASTING	METERS	10.00	0	R
444	440	DP1601.1BK_NAV_UPDATE	02-T-16-11	8	INIT/UPDATE POSITION-NORTHING	METERS	10.00	0	R
445	441	DP1601.1BK_NAV_UPDATE	02-T-16-11	9	INIT/UPDATE SPHEROID	RESET	1.00	0	I
446	442	DP1601.1BK_NAV_UPDATE	02-T-16-11	10	ODOMETER GAIN	GAIN	1.00	0	I
447	443	DP1601.1BK_NAV_UPDATE	02-T-16-11	11	ALIGNMENT ANGLE	DEGREES	0.10	0	R
448	444	DP1702_MODE_REQ_ST	02-T-17-04	1	LRU SYSTEM MODE TAG WORD	PUT	1.00	177	I
449	445	DP1702_MODE_REQ_ST	02-T-17-04	2	LRU SYSTEM MODE STATUS	N/A	1.00	0	B
450	446	DP1702_MODE_REQ_ST	02-T-17-04	3	NEXT SYSTEM MODE ACK	N/A	1.00	0	B
451	447	DP1702_MODE_REQ_ST	02-T-17-04	4	REQUESTED MODE	N/A	1.00	0	B
452	448	DP1900.1_WAYPT_ST	02-T-19-04	1	WAYPOINT STATUS TAG WOR	PUT	1.00	179	I
453	449	DP1900.1_WAYPT_ST	02-T-19-04	2	WAYPOINT RESPONSE	N/A	1.00	0	B
454	450	DP1900.1_WAYPT_ST	02-T-19-04	3	WAYPOINT NUMBER	N/A	1.00	0	B
455	451	DP1900.1_WAYPT_ST	02-T-19-04	4	NBC TEMP STATUS	N/A	1.00	0	B
456	452	DP2300.1_CB_CTL	02-T-23-03	1	CIRCUIT BREAKER CTRL TAG WORD	PUT	1.00	180	I
457	453	DP2300.1_CB_CTL	02-T-23-03	2	CIRCUIT BREAKER SELECT	N/A	1.00	0	B
458	454	DP2300.1_CB_CTL	02-T-23-03	3	CIRCUIT BREAKER SELECT STATE	N/A	1.00	0	B
459	455	DP3002_ST_DATA/DP3113_BIT_DAT/	02-T-30-06	1	BIT DATA TAG	PUT	1.00	198	I

🟥 Caution/Warning	Profilometer	Fuel Systems	Timing	Spatial/GPS/ Navigation
Hydraulics	Data Bus/Control/ Diagnostics	Ballistics	Electrical	Unused

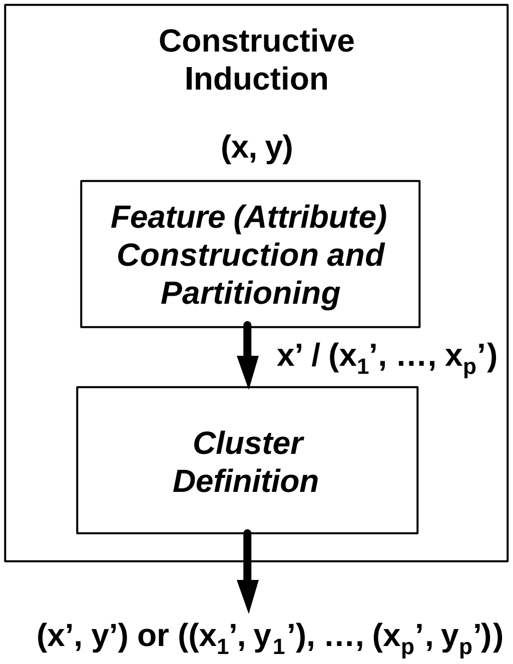
Data Aggregation and Sampling





Unsupervised Learning

- **Unsupervised Learning in Support of Supervised Learning**
  - Given: D ? labeled vectors (x, y)
  - Return: D' ? new training examples (x', y')
  - Constructive induction: transformation step in KDD
    - Feature "construction": generic term
    - Cluster definition
- **Feature Construction: Front End**
  - Synthesizing new attributes
    - Logical: x_1 ? ? x_2, arithmetic: $x_1 + x_5 / x_2$
    - Other synthetic attributes: $f(x_1, x_2, \ldots, x_n)$, etc.
  - Dimensionality-reducing projection, feature extraction
  - Subset selection: finding *relevant attributes* for a given target y
  - Partitioning: finding relevant attributes for given targets $y_1, y_2, \ldots, y_p$
- **Cluster Definition: Back End**
  - Form, segment, and label clusters to get intermediate targets y'
  - Change of representation: find good (x', y') for learning target y



Constructive Induction

(x, y)

Feature (Attribute) Construction and Partitioning

$x' / (x_1', \ldots, x_p')$

Cluster Definition

(x', y') or $((x_1', y_1'), \ldots, (x_p', y_p'))$

Relevance Determination

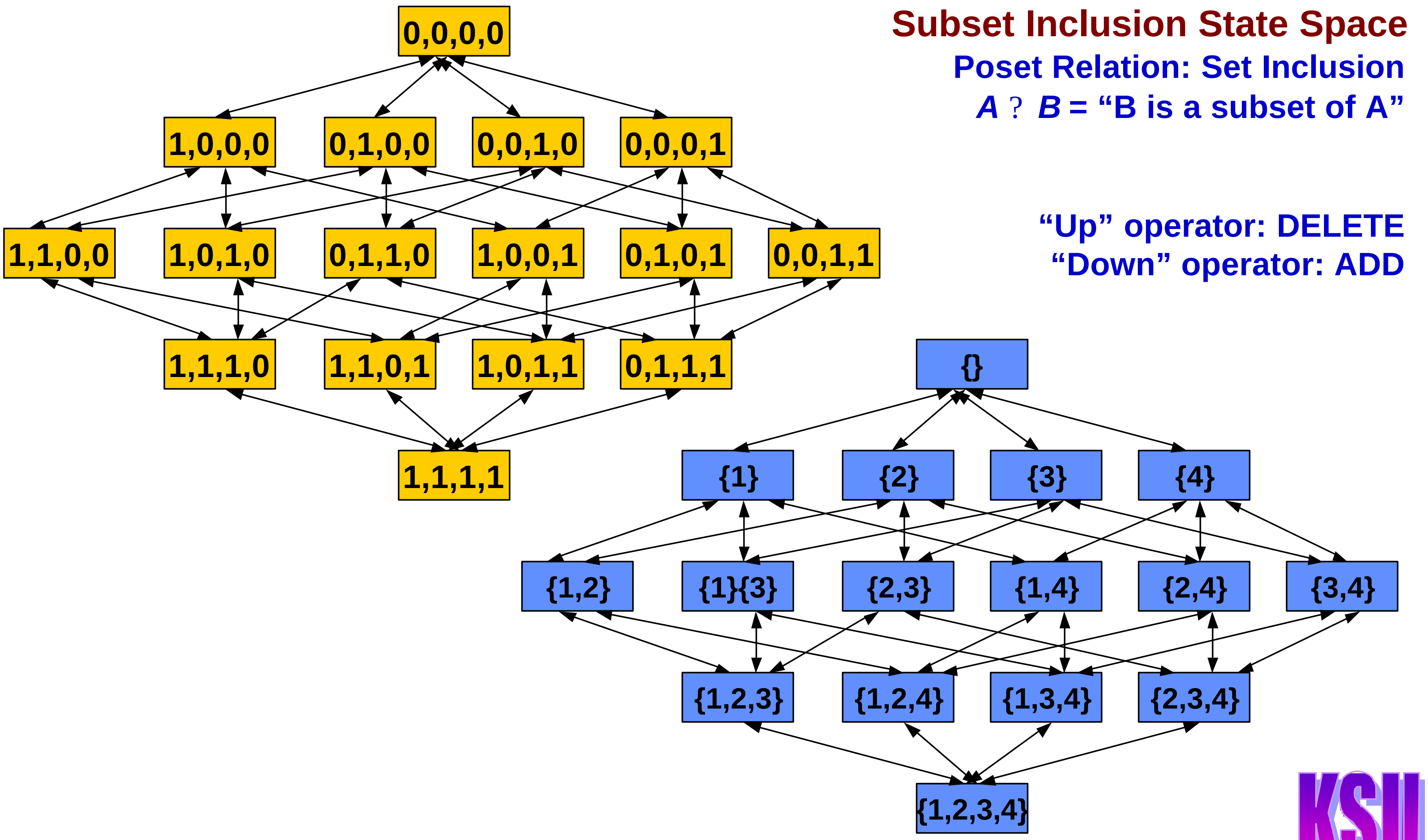
Subset Inclusion State Space

Poset Relation: Set Inclusion

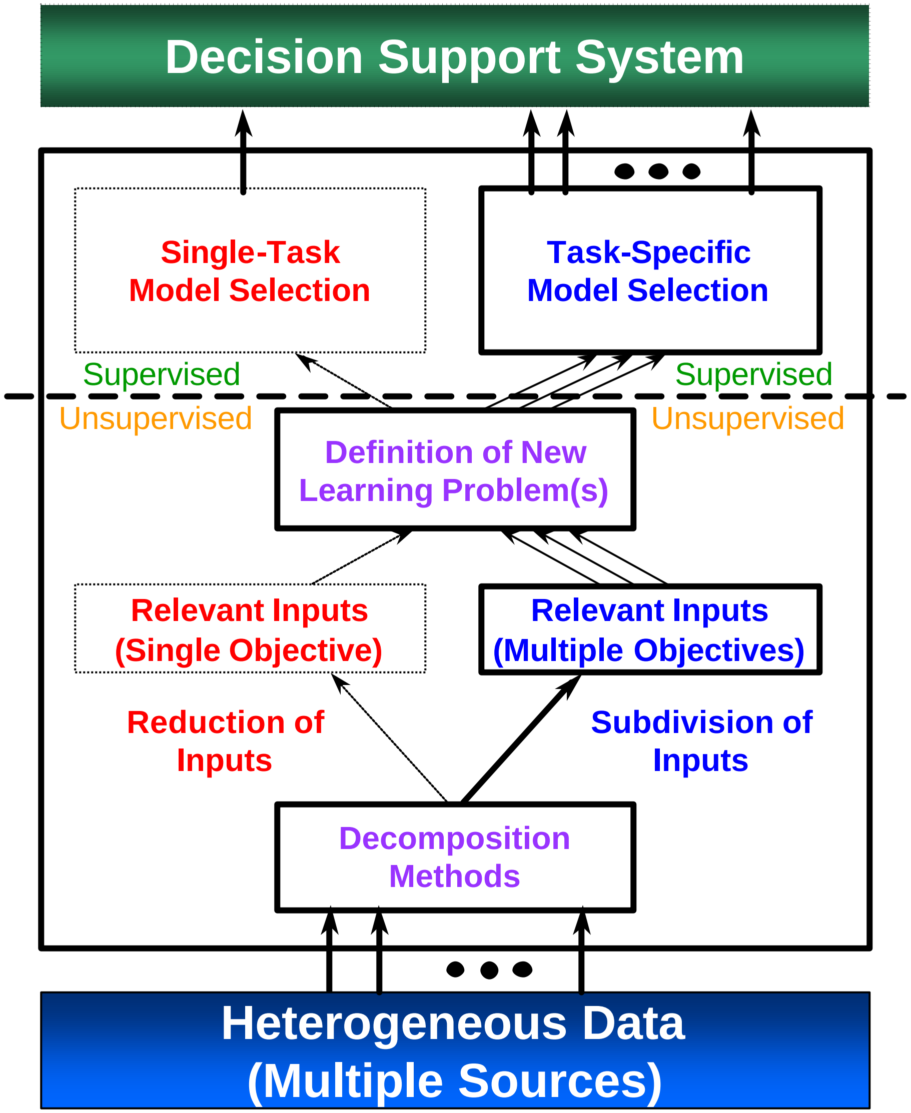
***A* ? *B* = "B is a subset of A"**

"Up" operator: DELETE

"Down" operator: ADD

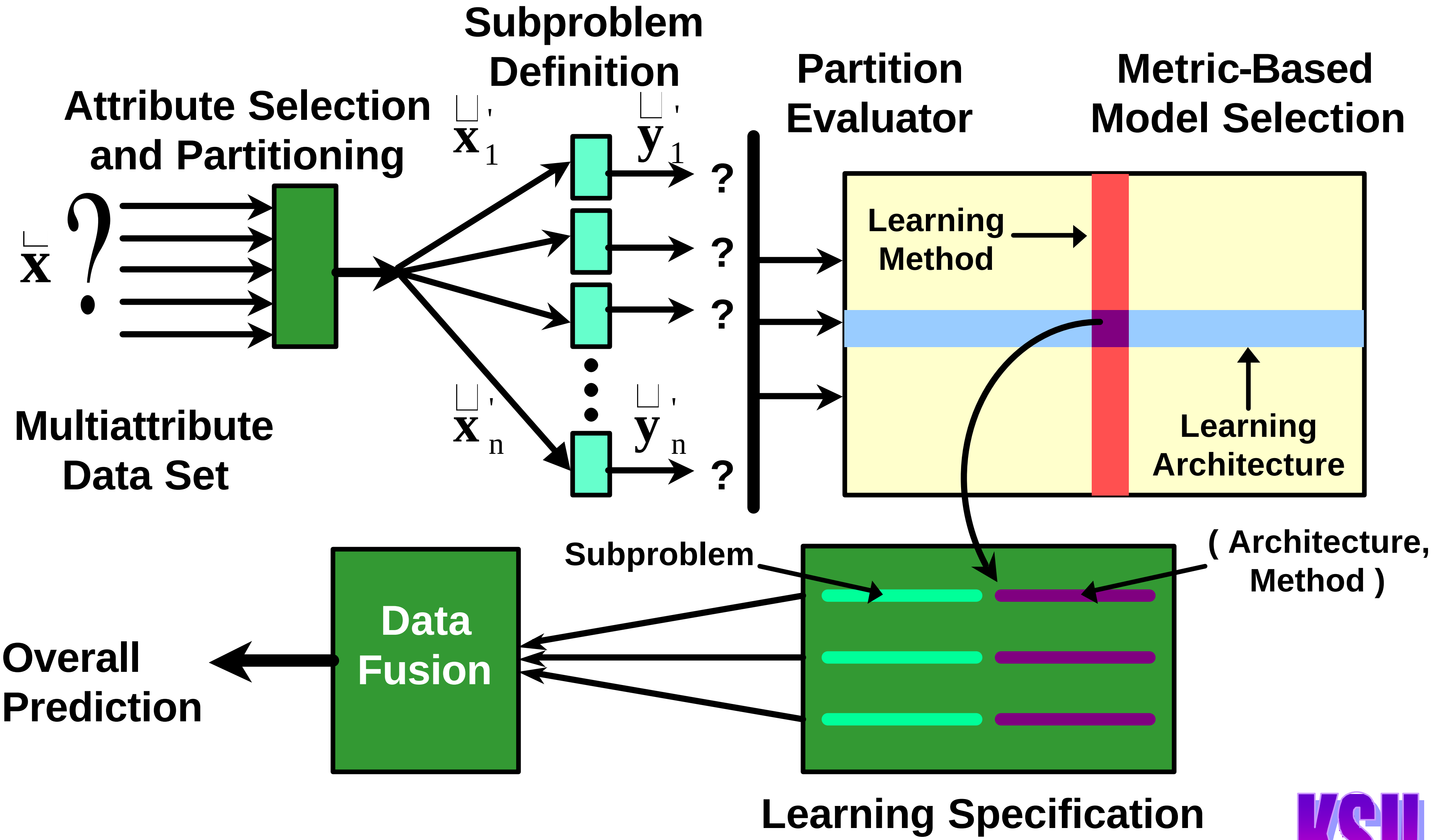


Wrappers for Performance Enhancement



- Wrappers
  - "Outer loops" for improving inducers
  - Use inducer performance to optimize
- Applications of Wrappers
  - Combining knowledge sources
    - Committee machines (static): bagging, stacking, boosting
    - Other sensor and data fusion
  - Tuning hyperparameters
    - Number of ANN hidden units
    - GA control parameters
    - Priors in Bayesian learning
  - Constructive induction
    - Attribute (feature) subset selection
    - Feature construction
- Implementing Wrappers
  - Search [Kohavi, 1995]
  - Genetic algorithm

Supervised Learning Framework



Subproblem Definition

Attribute Selection and Partitioning

Partition Evaluator

Metric-Based Model Selection

$\mathbf{x}$?

$\mathbf{x}'_1$

$\mathbf{y}'_1$

$\mathbf{x}'_n$

$\mathbf{y}'_n$

? ? ? ?

Learning Method

Learning Architecture

Multiattribute Data Set

Subproblem

(Architecture, Method)

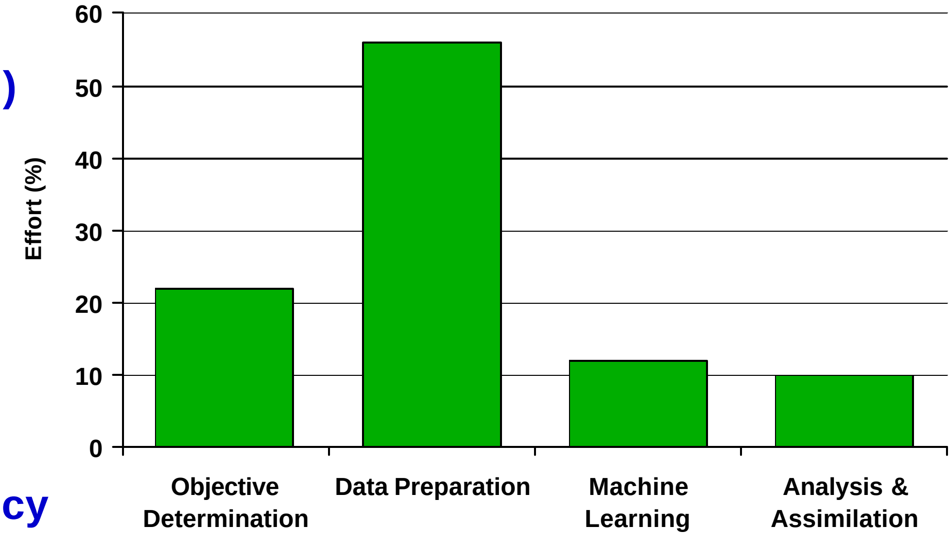
Data Fusion

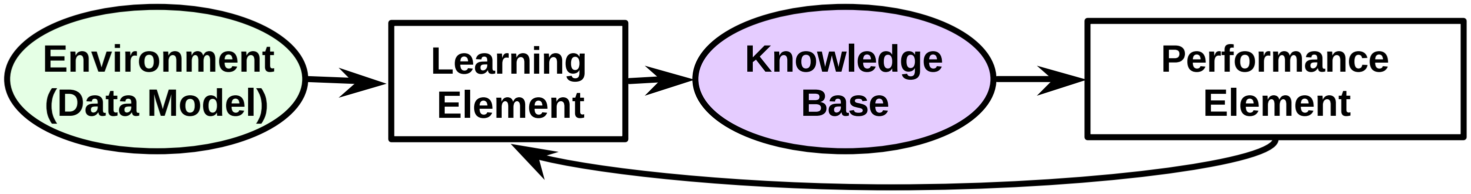
Overall Prediction

Learning Specification

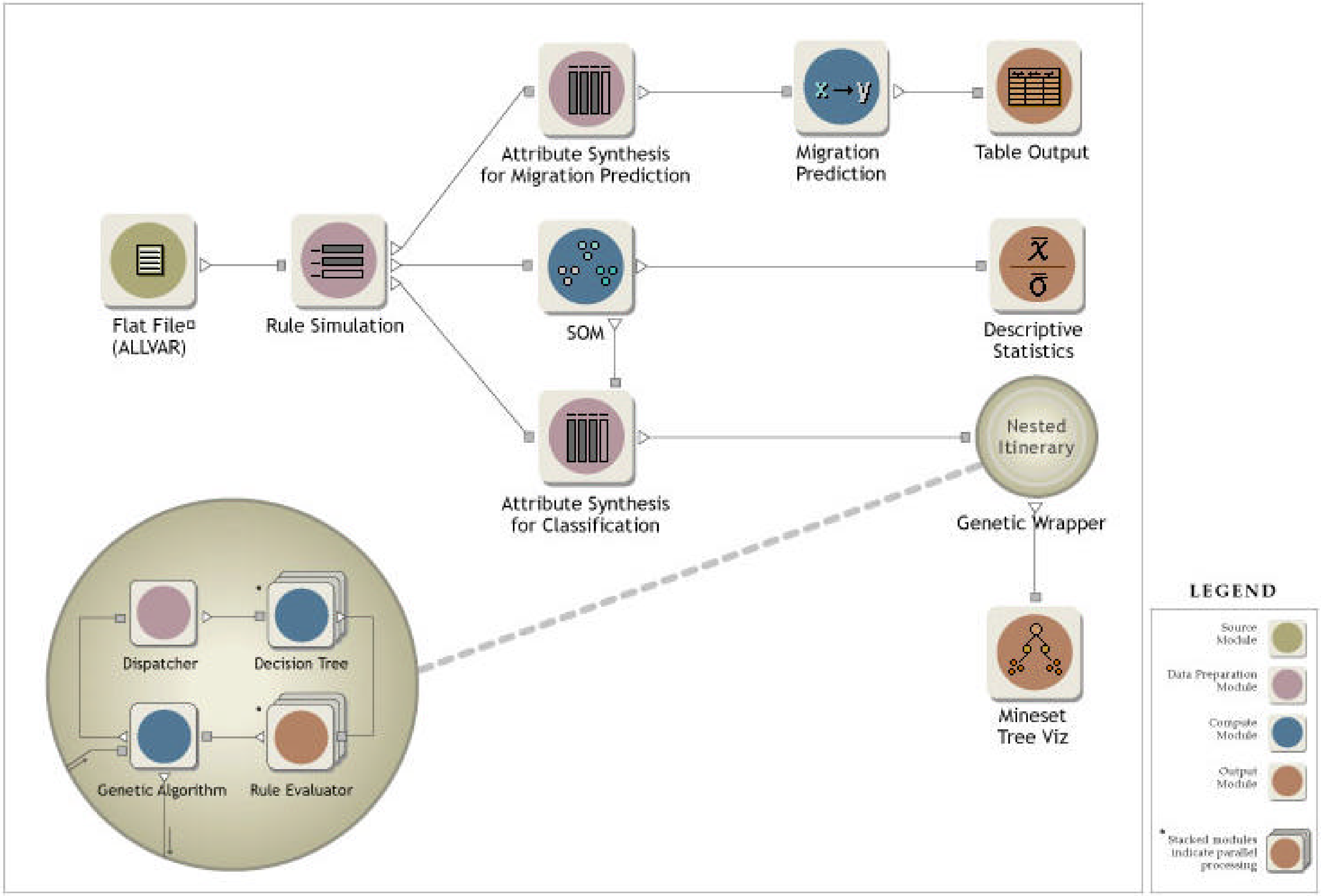
Performance Element: Decision Support Systems (DSS)

- Model Identification (Relational Database)
  - Specify data model
  - Group attributes by type (dimension)
  - Define queries
- Prediction Objective Identification
  - Identify target function
  - Define hypothesis space
- Transformation of Data
  - <u>Reduce</u> data: e.g., decrease frequency
  - <u>Select</u> *relevant* data channels (given prediction objective)
  - <u>Integrate</u> models, sources of data (e.g., *interactively elicited rules*)
- Supervised Learning
- Analysis and Assimilation: Performance Evaluation using DSS

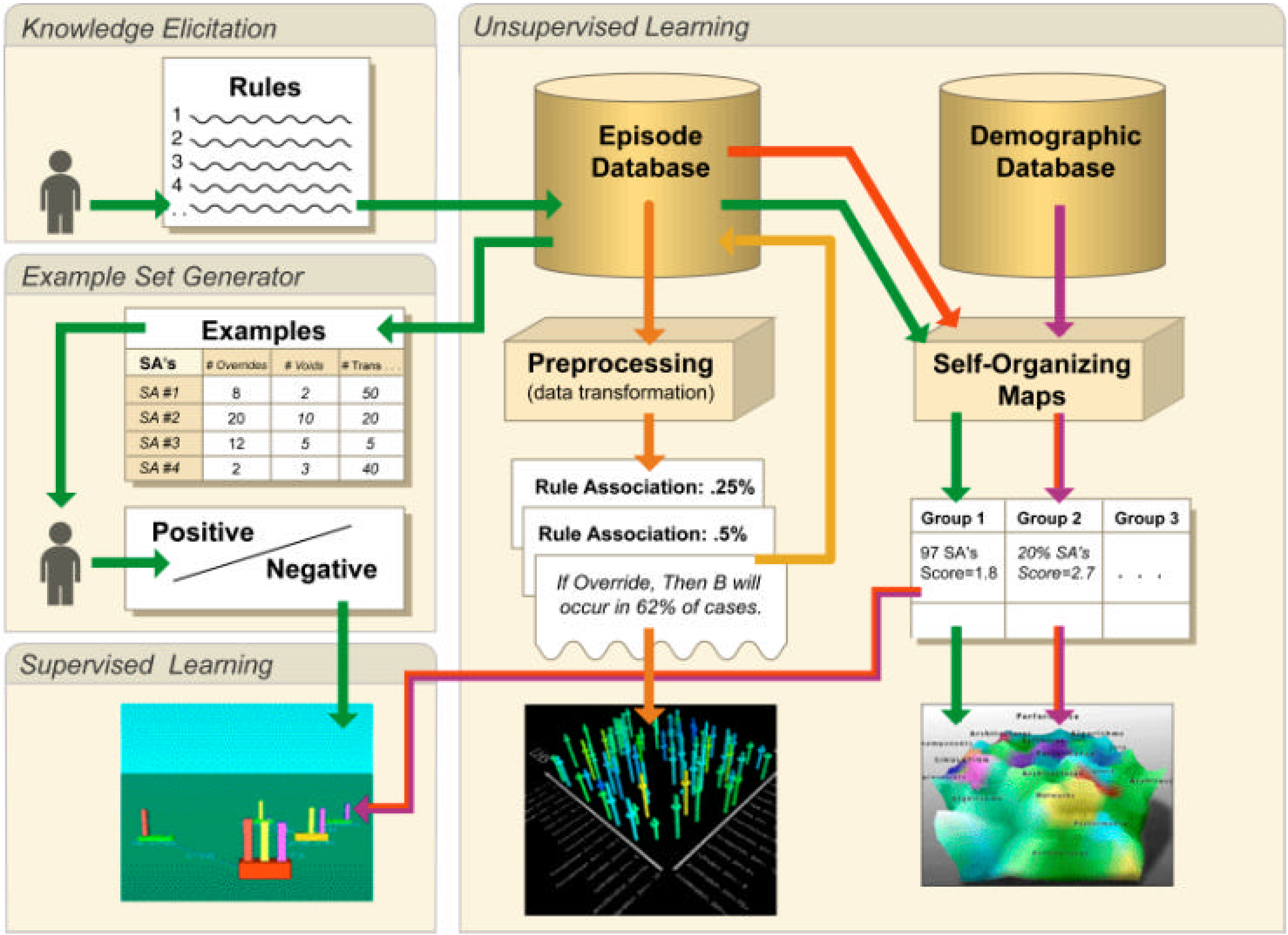




Case Study: Automobile Insurance Risk Analysis



Case Study: Fraud Detection



Knowledge Elicitation

Rules

1
2
3
4
. .

Example Set Generator

Examples

SA's	# Overrides	# Voids	# Trans . . .
SA #1	8	2	50
SA #2	20	10	20
SA #3	12	5	5
SA #4	2	3	40

Positive

Negative

Supervised Learning

Unsupervised Learning

Episode Database

Demographic Database

Preprocessing (data transformation)

Self-Organizing Maps

Rule Association: .25%

Rule Association: .5%

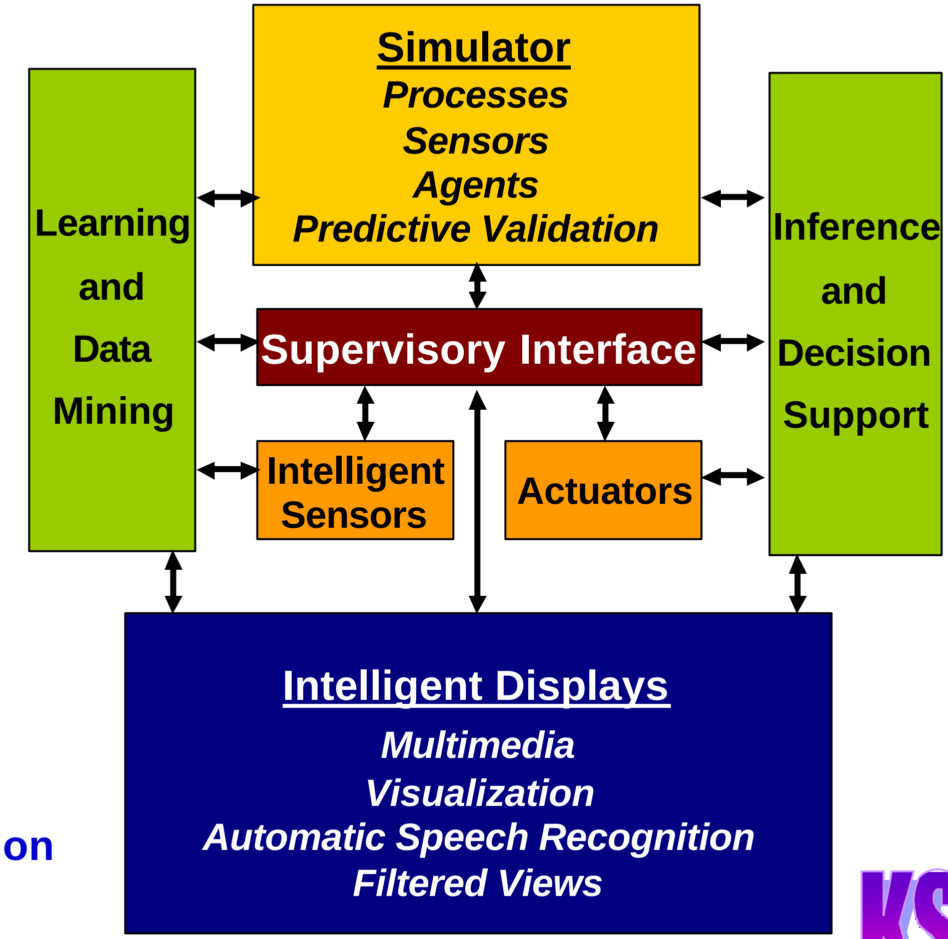
If Override, Then B will occur in 62% of cases.

Group 1	Group 2	Group 3
97 SA's Score=1.8	20% SA's Score=2.7	. . .

NCSA *D2K* - http://www.ncsa.uiuc.edu/STI/ALG

Case Study: Prognostic Monitoring

- Control Interfaces
  - Actuators: fire/smoke suppression, electrical isolation, counterflooding
  - Intelligent sensors
- Simulation Module
  - Process/agent simulation
  - Automation simulation
  - Predictive validation for sensors
- Learning Modules
  - Time series learning
  - Control knowledge acquisition
- Intelligent Reasoning Modules
  - Crisis recognition
  - Casualty response
- Intelligent Displays Module
  - Interactive design and visualization
  - Supervisory interface



Terminology

- **Data Mining**
  - Operational definition: automatically extracting *valid*, *useful*, *novel*, *comprehensible* information from large databases and *using it to make decisions*
  - Constructive definition: expressed in stages of data mining
- **Databases and Data Mining**
  - Data Base Management System (DBMS): data *organization, retrieval, processing*
  - Data warehouse: repository of integrated information for queries, analysis
  - Online Analytical Processing (OLAP): storage/CPU-efficient manipulation of data for summarization (descriptive statistics), inductive learning and inference
- **Stages of Data Mining**
  - Data selection (*aka* filtering): sampling original (raw) data
  - Data preprocessing: sorting, segmenting, aggregating
  - Data transformation: change of representation; feature construction, selection, extraction; quantization (scalar, e.g., histogramming, vector, *aka* clustering)
  - Machine learning: unsupervised, supervised, reinforcement for model building
  - Inference: application of performance element (pattern recognition, *etc.*); evaluation, assimilation of results

Summary Points

- Knowledge Discovery in Databases (KDD) and Data Mining
  - Stages: selection (filtering), processing, transformation, learning, inference
  - Design and implementation issues
- Role of Machine Learning and Inference in Data Mining
  - Roles of unsupervised, supervised learning in KDD
  - Decision support (information retrieval, prediction, policy optimization)
- Case Studies
  - Risk analysis, transaction monitoring (filtering), prognostic monitoring
  - Applications: business decision support (pricing, fraud detection), automation
- Resources Online
  - Microsoft DMX Group (Fayyad): http://research.microsoft.com/research/DMX/
  - KSU KDD Lab (Hsu): http://ringil.cis.ksu.edu/KDD/
  - CMU KDD Lab (Mitchell): http://www.cs.cmu.edu/~cald
  - KD Nuggets (Piatetsky-Shapiro): http://www.kdnuggets.com
  - NCSA Automated Learning Group (Welge)
    - ALG home page: http://www.ncsa.uiuc.edu/STI/ALG
    - NCSA *D2K*: http://chili.ncsa.uiuc.edu