Data Integration (Chapter 17)

Peter Revesz

CSCE 413/813
Computer Science and Engineering
University of Nebraska-Lincoln

Classification

A classifier is a mapping from a feature space $X=\{x_i\}$ to a set of labels $Y=\{y_i\}$.

$$f : X \rightarrow Y$$

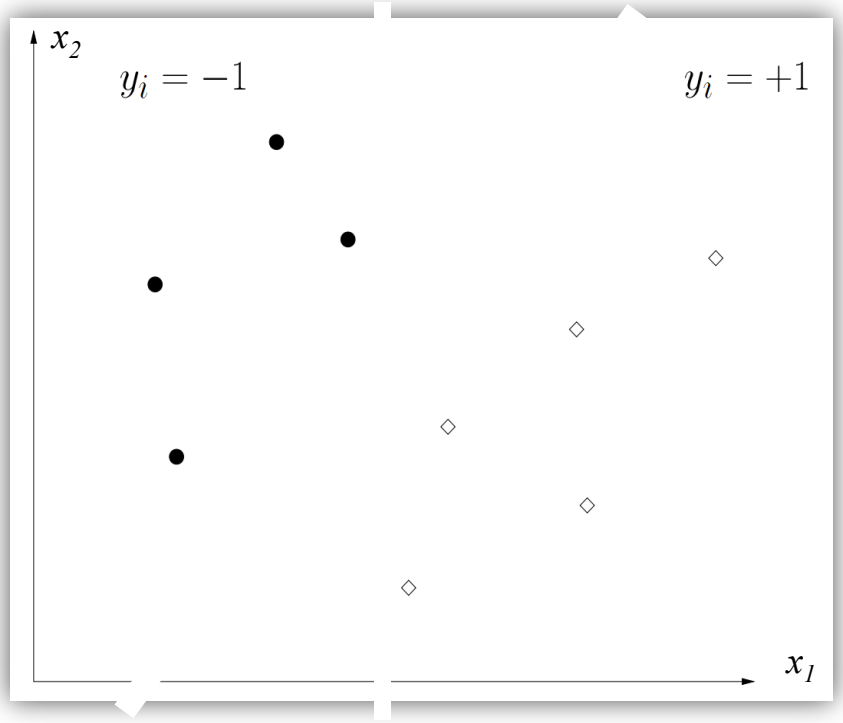
A linear classifier for feature y_i has the form:

$$y_i = f\left(\sum_j w_j x_j\right)$$

where $x_j \in \mathbb{R}$ are the features of the classifier

$w_j \in \mathbb{R}$ are the weights of the classifier

Support Vector Machine (SVM)



- Classification using a hyperplane to split training examples into 2 sets of classes (+1 and -1)
- Several suitable hyperplanes
- SVM: maximum-margin hyperplane

Relational Database → SVM → Constraint Database

Heart Patients in Cottonwood Hospital

Features:

- chest pain *P*
- cholesterol *C*
- gender *G*
- resting blood pressure *B*

Label:

- disease *D*

Cottonwood

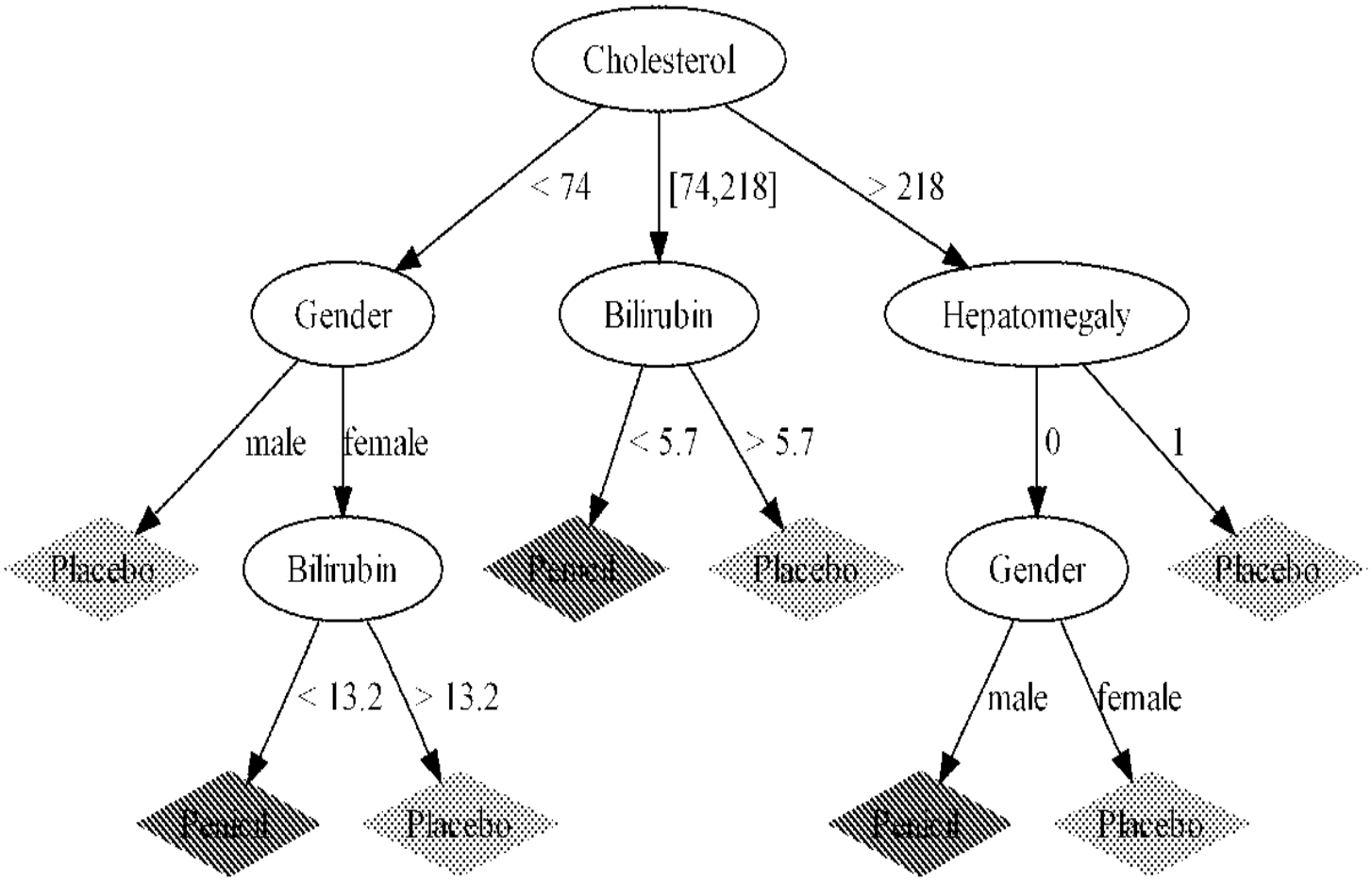
P	C	G	B	D
1	233	1	145	No
3	250	1	130	Yes
3	275	0	110	No
4	230	1	117	Yes
2	198	0	105	No
4	266	1	124	Yes

Cottonwood Classification

P	C	G	B	D_1	
p	c	g	b	d_1	$1.1568p - 0.0172c + 0.4278g + 0.0333b - 3.404 = d_1$

Relational Database → Decision Tree

Features: bilirubin *B*, cholesterol *C*, gender *G*, hepatomegaly *H*
Label: drug prescribed *D*



Decision Tree → Constraint Database

Features: bilirubin *B*, cholesterol *C*, gender *G*, hepatomegaly *H*

Label: drug prescribed *D*

Drug

B	C	G	H	D	
b	c	g	h	d	$c < 74,\ g = 1,\ d =$ 'Placebo'
b	c	g	h	d	$c < 74,\ g = 0,\ b < 13.2,\ d =$ 'Penicil.'
b	c	g	h	d	$c < 74,\ g = 0,\ b > 13.2,\ d =$ 'Placebo'
b	c	g	h	d	$c \geq 74,\ c \leq 218,\ b < 5.7,\ d =$ 'Penicil.'
b	c	g	h	d	$c \geq 74,\ c \leq 218,\ b > 5.7,\ d =$ 'Placebo'
b	c	g	h	d	$c > 218,\ h = 0,\ g = 1,\ d =$ 'Penicil.'
b	c	g	h	d	$c > 218,\ h = 0,\ g = 0,\ d =$ 'Placebo'
b	c	g	h	d	$c > 218,\ h = 1,\ d =$ 'Placebo'

Example: Heart Patients in Two Hospitals

Features: chest pain *P*, cholesterol *C*, gender *G*, resting blood pressure *B*

Label: disease *D*

Cottonwood

P	C	G	B	D
1	233	1	145	No
3	250	1	130	Yes
3	275	0	110	No
4	230	1	117	Yes
2	198	0	105	No
4	266	1	124	Yes

Pineridge

P	C	G	B	D
1	237	0	170	No
2	219	0	100	No
4	270	1	120	Yes
2	198	0	105	No
4	246	0	100	Yes
1	156	1	140	Yes
2	257	1	110	Yes

If we could integrate the raw data into

Cottonwood-Pineridge

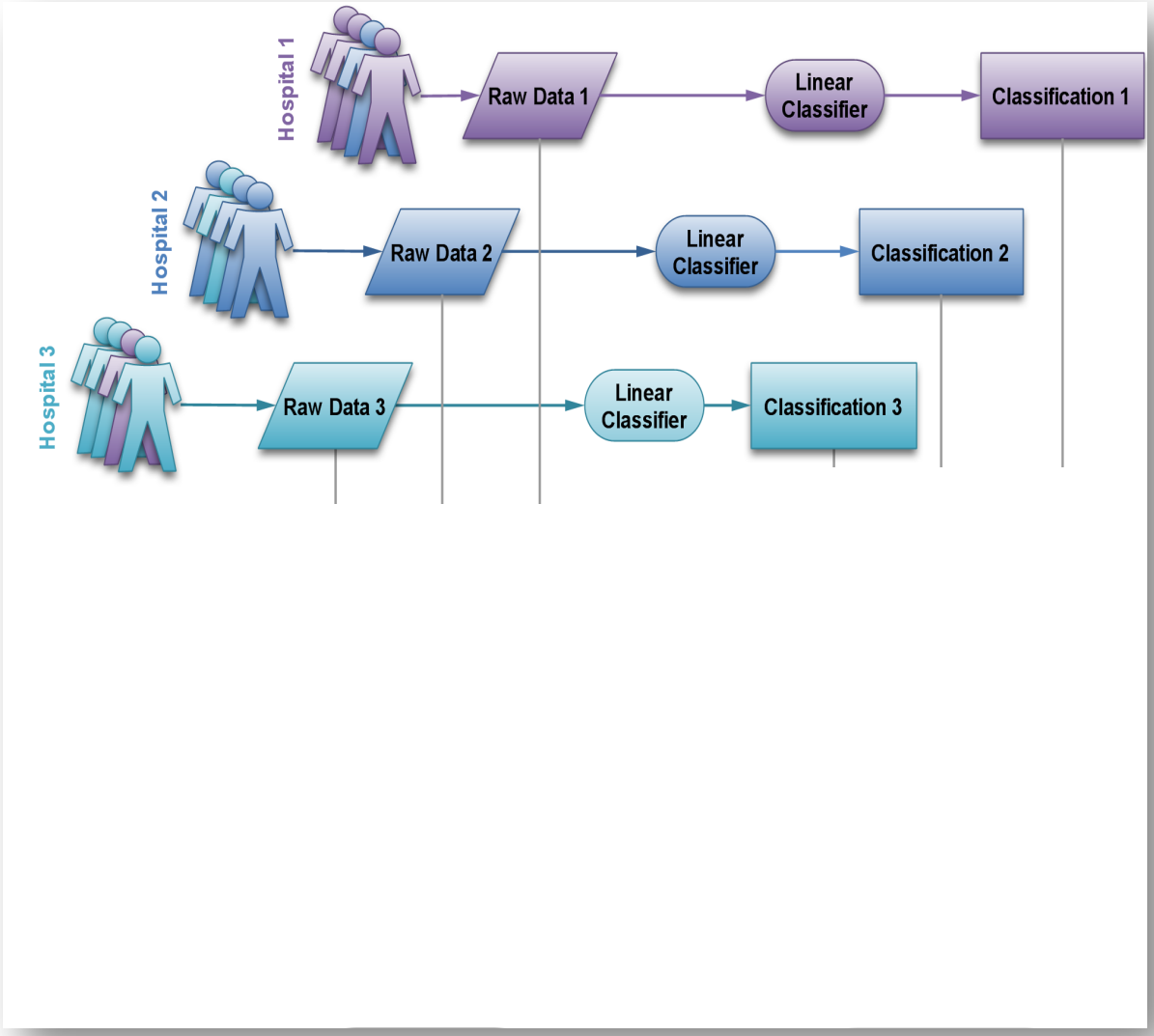
P	C	G	B	D
1	233	1	145	No
3	250	1	130	Yes
3	275	0	110	No
4	230	1	117	Yes
2	198	0	105	No
4	266	1	124	Yes
1	237	0	170	No
2	219	0	100	No
4	270	1	120	Yes
2	198	0	105	No
4	246	0	100	Yes
1	156	1	140	Yes
2	257	1	110	Yes

then we could run an algorithm to build an SVM classifier on it.

Suppose the hospitals do not want to share their raw data because of privacy concerns and legal restrictions.

What can we do?

Data Integration vs. Classification Integration



Classification integration → Constraint Database

Step 1: The hospitals find separate SVM classifications.

Cottonwood Classification

P	C	G	B	D_1	
p	c	g	b	d_1	$1.1568p - 0.0172c + 0.4278g + 0.0333b - 3.404 = d_1$

Pineridge Classification

P	C	G	B	D_2	
p	c	g	b	d_2	$1.0213p - 0.0016c + 1.9091g + 0.015b - 4.2018 = d_2$

Step 2: The hospitals combine their knowledge using an SQL query (MLPQ system).

Cottonwood-Pineridge Classification

P	C	G	B	D	
p	c	g	b	d	$d = 0.5d_1 + 0.5d_2,$ $1.1568p - 0.0172c + 0.4278g + 0.0333b - 3.404 = d_1,$ $1.0213p - 0.0016c + 1.9091g + 0.015b - 4.2018 = d_2$

Simplifying we get:

Cottonwood-Pineridge Classification

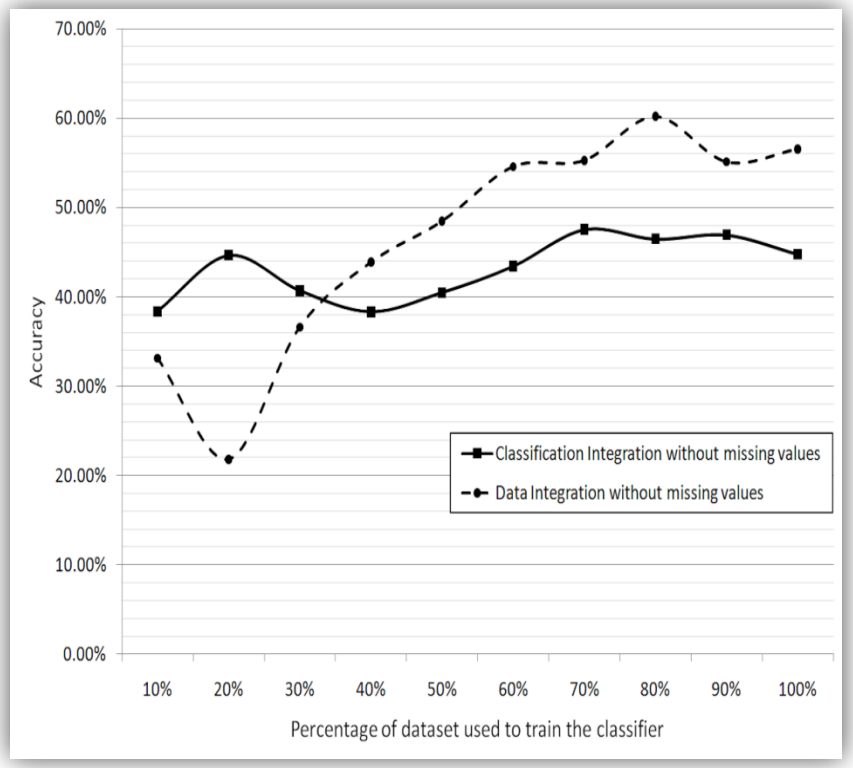
P	C	G	B	D	
p	c	g	b	d	$1.0891p - 0.0094c + 1.1685g + 0.0242b - 3.8029 = d$

Data Integration vs. Classification Integration – Accuracy Comparison

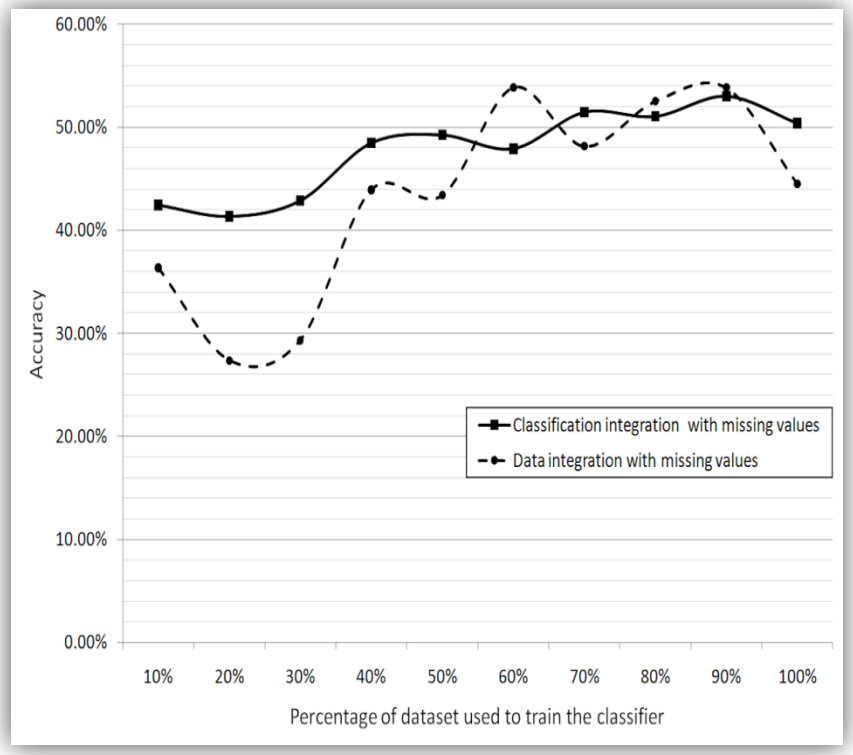
- Dataset: *Heart Disease Diagnostic* for Budapest, Cleveland and Zurich

- Procedure
  1. Randomly select 60 records as independent testing set S_T
  2. Randomly select $n\%$ of remaining records as training set
  3. Build a classifier f_{123} on the union of the training data
  4. Build classifier f_1, f_2, f_3 on training data source
  5. Integrate f_1, f_2, f_3 using classification integration to find f_{class_integ}
  6. Test accuracy of f_{123} and f_{class_integ} using S_T
  7. Iterate [2-6] for $n=5\%, n=15\%, \ldots, n=95\%$

Results using ID3 on *Heart disease*



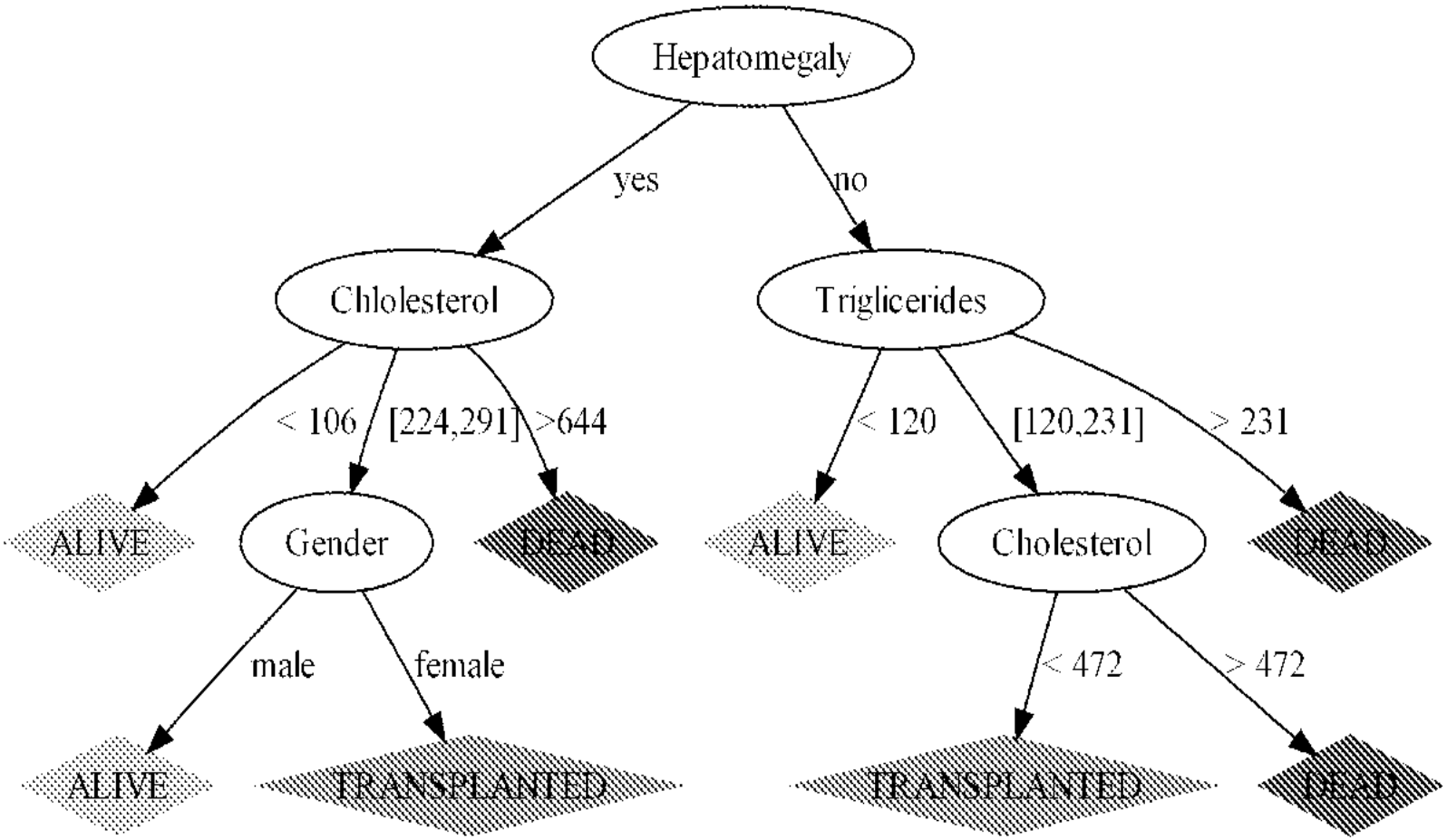
No missing values



With missing values

Features: cholesterol *C*, gender *G*, hepatomegaly *H,* triglicerides *T*

Label: patients status at end of study *S*



Question

- What is the relationship between the drug prescribed to the patient and the patient's status at the end of the study?

- That is difficult to answer because the drug prescribed D and the patient status at the end of the study S are in separate decision trees.

- What can we do?

Reclassification – Let $c_1 : X_1 \to Y_1$ and $c_2 : X_2 \to Y_2$ be classifiers

$c : X_1 \cup X_2 \to Y_1 \times Y_2$ new classifier

different sets of features and labels

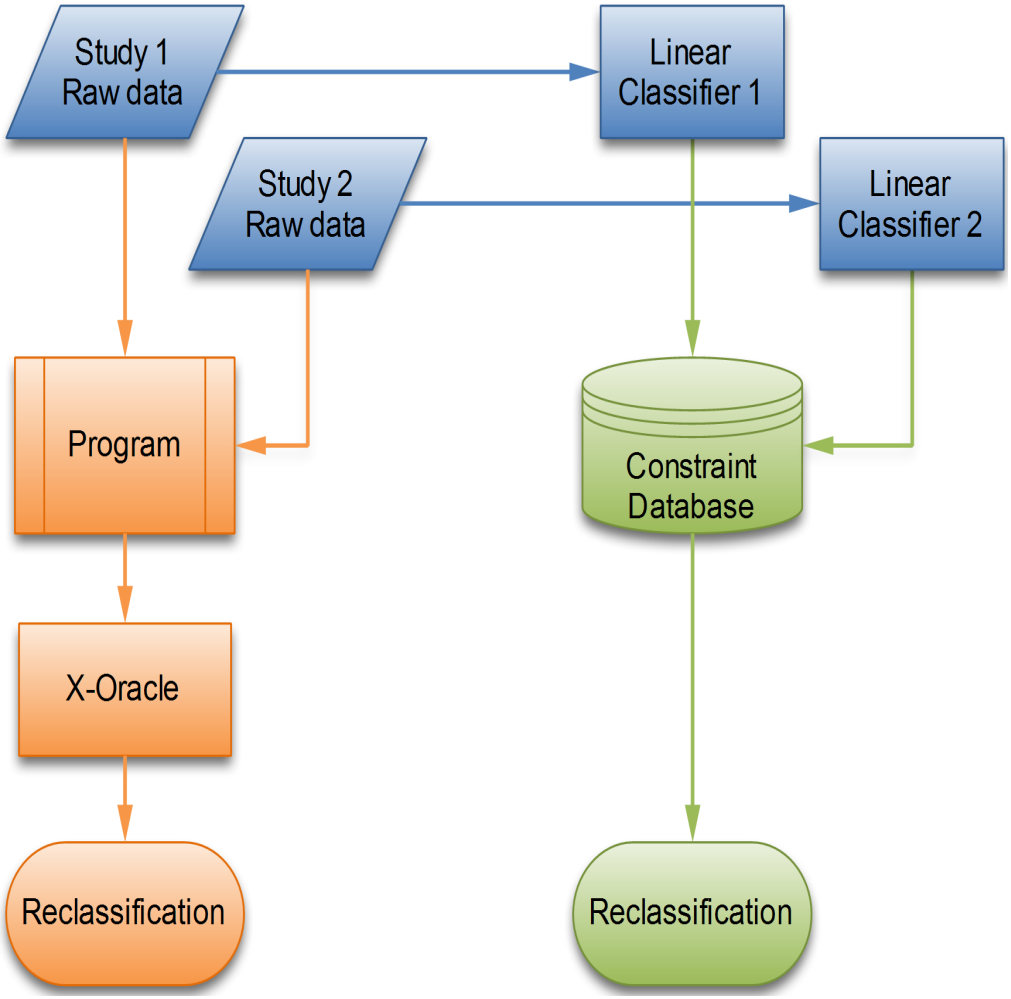
- Original classifications with

$$X_1 = \{B, C, G, H\} \quad \text{and} \quad Y_1 = \{D\}$$
$$X_2 = \{C, G, H, T, S\} \quad \text{and} \quad Y_2 = \{S\}$$

- New classification

$$X = X_1 \cup X_2 = \{B, C, G, H, T\}$$
$$Y = Y_1 \times Y_2 = D \times S$$

Reclassification methods



- **Original classifiers: 2 or more**
- **Standard approach:**
  Build a new linear classifier
  - Requires access to the raw data
  - Complex to operate: Programming skills
  - Requires an oracle: not a practical method
- **Proposed approach:**
  Combine the existing linear classifiers
  - No required access to the raw data, only the resulting classifications
  - Flexible: 1 table per classifier
  - Easy to use: SQL / Datalog queries have been well known for many years

Join the following constraint tables:

Drug

B	C	G	H	D	
b	c	g	h	d	$c < 74,\ g = 1,\ d =$ 'Placebo'
b	c	g	h	d	$c < 74,\ g = 0,\ b < 13.2,\ d =$ 'Penicil.'
b	c	g	h	d	$c < 74,\ g = 0,\ b > 13.2,\ d =$ 'Placebo'
b	c	g	h	d	$c \geq 74,\ c \leq 218,\ b < 5.7,\ d =$ 'Penicil.'
b	c	g	h	d	$c \geq 74,\ c \leq 218,\ b > 5.7,\ d =$ 'Placebo'
b	c	g	h	d	$c > 218,\ h = 0,\ g = 1,\ d =$ 'Penicil.'
b	c	g	h	d	$c > 218,\ h = 0,\ g = 0,\ d =$ 'Placebo'
b	c	g	h	d	$c > 218,\ h = 1,\ d =$ 'Placebo'

Status

C	G	H	T	S	
c	g	h	t	s	$h = 1,\ c < 224,\ s =$'Alive'
c	g	h	t	s	$h = 1,\ c \geq 224,\ c \leq 291,\ g = 1,\ s =$'Alive'
c	g	h	t	s	$h = 1,\ c \geq 224,\ c \leq 291,\ g = 0,\ s =$'Transplanted'
c	g	h	t	s	$h = 1,\ c > 291,\ s =$'Dead'
c	g	h	t	s	$h = 0,\ t < 120,\ s =$'Alive'
c	g	h	t	s	$h = 0,\ t \geq 120,\ t \leq 231,\ c < 472,\ s =$'Transplanted'
c	g	h	t	s	$h = 0,\ t \geq 120,\ t \leq 231,\ c > 472,\ s =$'Dead'
c	g	h	t	s	$h = 0,\ t > 231,\ s =$'Alive'

Join can be done by a simple SQL query in the MLPQ system:

```
CREATE VIEW   Reclass(B,C,G,H,T,D,S)
SELECT        Dr.B, Dr.C, Dr.G, Dr.H, St.T, Dr.D, St.S
FROM          Drug AS Dr, Status as St
WHERE         Dr.C = St.C  AND
              Dr.G = St.G  AND
              Dr.H = St.H
```

Practice

1. A patient has the following record: $id = 1, b = 6.5, c = 326, g = 1, h = 0$ and $t = 130$.

   (a) Find the drug of the patient using the decision tree in Figure 17.1.

   (b) Find the status of the patient using the decision tree in Figure 17.2.