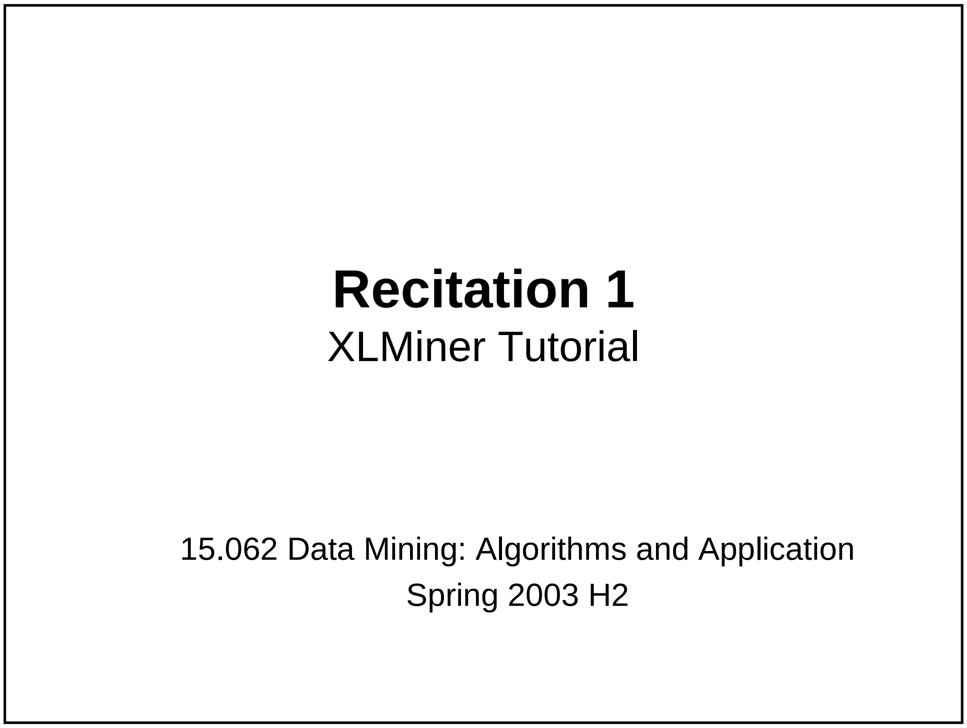
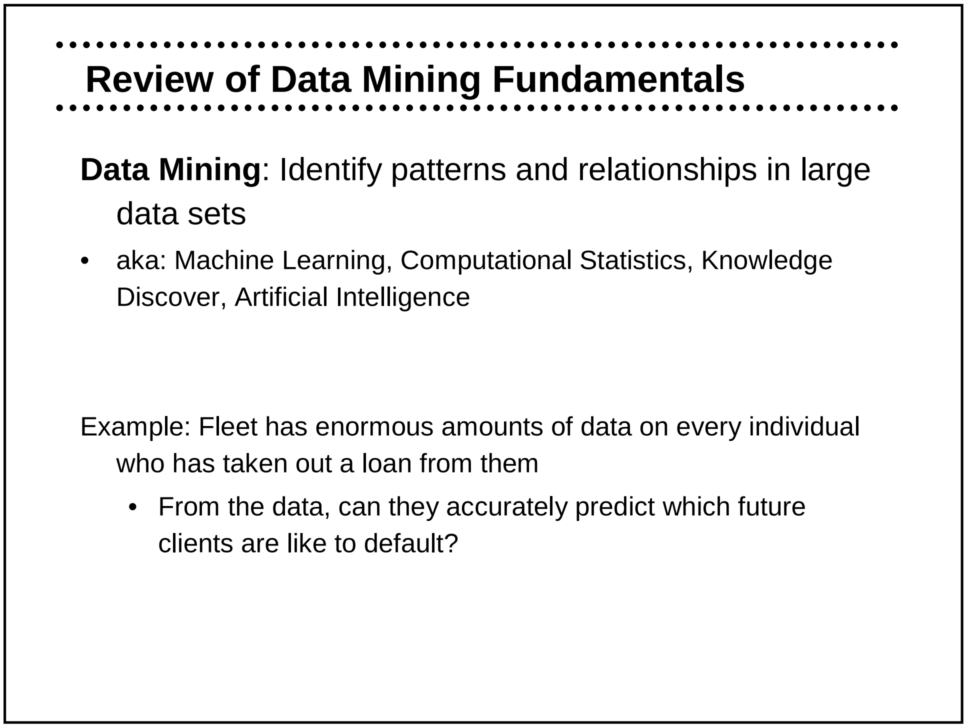


# Recitation 1

XLMiner Tutorial

15.062 Data Mining: Algorithms and Application
Spring 2003 H2



## Review of Data Mining Fundamentals

**Data Mining**: Identify patterns and relationships in large data sets

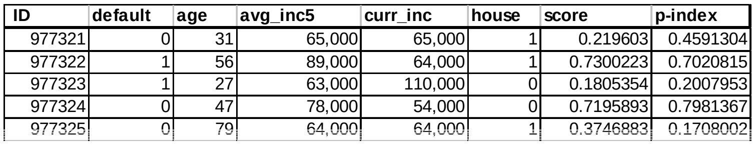
- aka: Machine Learning, Computational Statistics, Knowledge Discover, Artificial Intelligence

Example: Fleet has enormous amounts of data on every individual who has taken out a loan from them

- From the data, can they accurately predict which future clients are like to default?

## Review of Data Mining Fundamentals

### The Data:



| ID | default | age | avg_inc5 | curr_inc | house | score | p-index |
|---|---|---|---|---|---|---|---|
| 977321 | 0 | 31 | 65,000 | 65,000 | 1 | 0.219603 | 0.4591304 |
| 977322 | 1 | 56 | 89,000 | 64,000 | 1 | 0.7300223 | 0.7020815 |
| 977323 | 1 | 27 | 63,000 | 110,000 | 0 | 0.1805354 | 0.2007953 |
| 977324 | 0 | 47 | 78,000 | 54,000 | 0 | 0.7195893 | 0.7981367 |
| 977325 | 0 | 79 | 64,000 | 64,000 | 1 | 0.3746883 | 0.1708002 |

**Columns:** variables, features, attributes,….

**Rows:** observations, data points, cases, records, patterns…

**Inputs:** independent variables, explanatory variables, predictors…

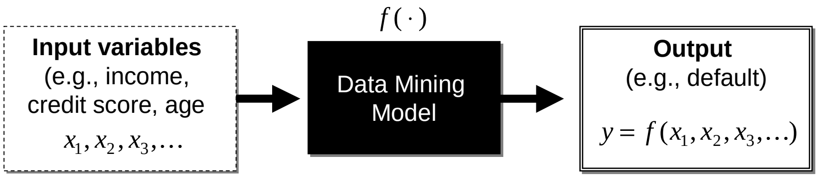
**Output:** dependent variables, predicted variable, target, outcome….

**Goal:** Given these historical records, build a model that can accurately predict defaults (target) given new customers (observations)

## Review of Data Mining Fundamentals

- **Supervised Learning** – Goal is to predict the value of an output based on inputs
  - Classification, Regression
- **Unsupervised Leaning** – No output. Determine/describe patterns in the inputs.
  - Clustering, Association

Challenge in supervised learning is to determine what model to use on the data to get an accurate prediction of the output variable



$f(\cdot)$

**Input variables** (e.g., income, credit score, age $x_1, x_2, x_3, \ldots$

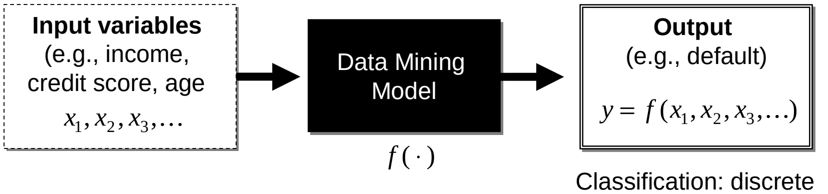
Data Mining Model

**Output** (e.g., default) $y = f(x_1, x_2, x_3, \ldots)$

## Review of Data Mining Fundamentals

**Types of Supervised Learning Problems:**

- **Classification Problems:** Output categorical/discrete
  - E.g., default or no default, fraud or not fraud, cancer or no cancer,
  - Models: k-nearest neighbor, naïve Bayes, classification trees,…
- **Regression Problems:** Output continuous
  - E.g., value of house, price of asset, expected payoff
  - Models: linear regression, regression trees,…



**Input variables** (e.g., income, credit score, age $x_1, x_2, x_3, \ldots$

Data Mining Model

$f(\cdot)$

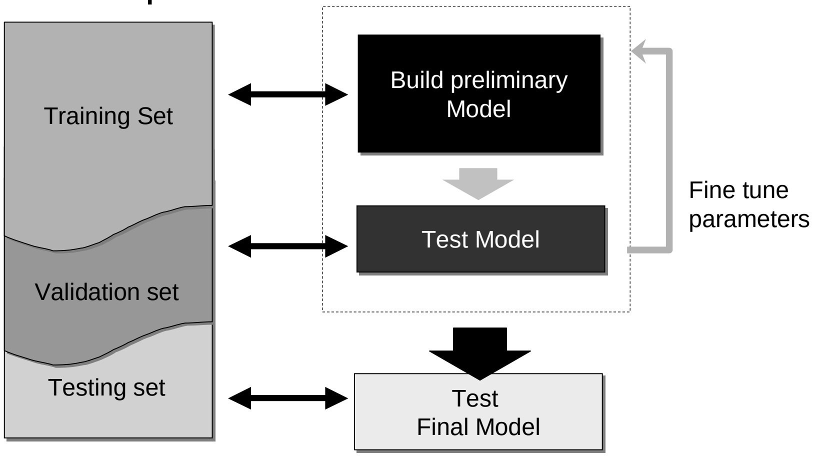
**Output** (e.g., default) $y = f(x_1, x_2, x_3, \ldots)$

Classification: discrete
Regression: continuous

## Model Building Procedures

**Partition data into 2 or 3 parts:**

**Build Model:**



Training Set

Validation set

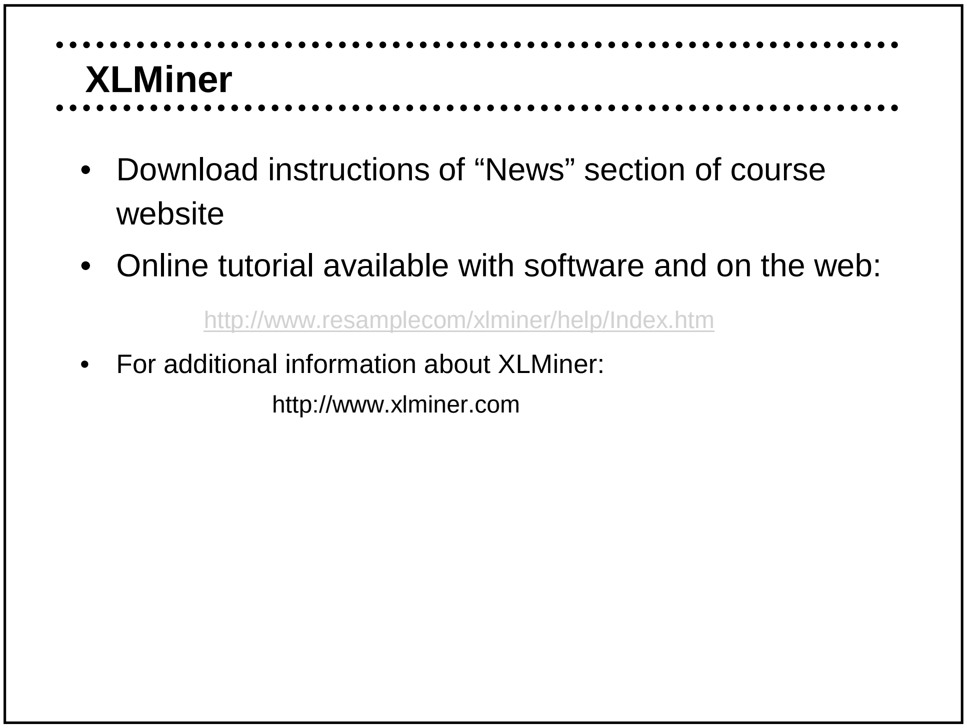
Testing set

Build preliminary Model

Test Model

Fine tune parameters

Test Final Model



## XLMiner

- Download instructions of “News” section of course website
- Online tutorial available with software and on the web:

  http://www.resamplecom/xlminer/help/Index.htm

- For additional information about XLMiner:

  http://www.xlminer.com