CIS501 – Lecture 2

Woon Wei Lee
Fall 2013, CIS Program, EECS
10:00pm-11:15pm, Sundays and Wednesdays

For today:

- Administrative stuff
    - Presentation list
    - Software
    - E-mail contacts and details
- Course structure
- Core Data Mining concepts
    - Data
        - types and representation
    - Problems with data
        - Noise
        - Curse of dimensionality

Course structure

Data Mining

1. Supervised
   a. Probability based
   b. Discriminant function

2. Unsupervised
   a. Clustering
   b. Visualization, dimensionality reduction
   c. Collaborative filtering

1

1a

(1) Prediction and Classification with k-Nearest Neighbors

(2) Classification and Bayes Rule, Naïve Bayes

1b

(3) Decision Trees

(4) Decision Trees II

(5) Discriminant Analysis

(6) Logistic Regression Case

(7) Neural Networks

(8) Neural Networks II

(9) Regularization and training algorithms.

2

[1c][1b][1a]

(10) k-Means clustering, hierarchical clustering

(11) Visualization, Principal Components

(12) Association rules, recommendation systems: collaborative filtering

Types of Data

- **Numerical**
  - ***Continuous data***
    - Any real numerical number,
    - Technically with “infinite precision”
    - Possibly bounded
  - ***Discrete data***
    - Integers – e.g. population, Number of correct/valid responses, etc
    - Fixed number of increments (0, 0.5, 1.0, 1.5, 2.0..).
    - e.g. histograms, “Digital” data

Masdar INSTITUTE

Types of Data (Cont'd)

- **Categorical data**
  - ***Ordinal***
    - e.g. Small, Quite small, Normal, Big, Very big, etc
  - ***Nominal (unordered)***
    - e.g. Tastes – Salty, Sweet, Peppery, Spicy, Tangy, Sour, etc..
  - ***Binary data*** *(special case)*
    - Yes/No situations
- **However, conversions are often possible:**
  - Discretization
  - Thresholding (*floor/ceiling)*
  - Expansion of nominal data to multiple variables

Data representations: (a) Vector format

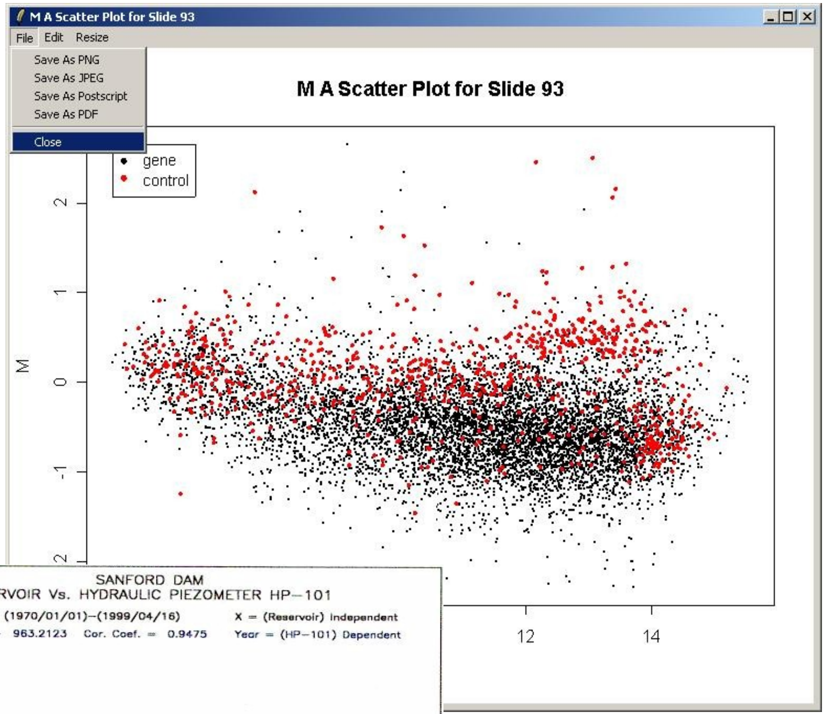
- Data structured as a series of vectors, so, for e.g.:
  - …, (“Wei Lee”, “CIS”, “Associate Professor”),

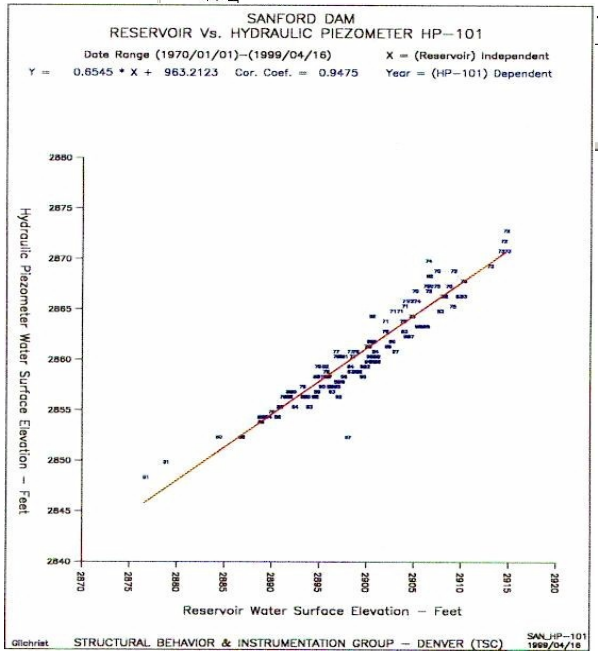
    (“Hatem”, “EPE”,”Associate Professor”),

    (“Andreas”, “CIS”, “Assistant Professor”), ...
- Each vector contains one record and represents an *object*, *transaction* or *observation.*
- Each element within the vector describes one property or attribute of this object.
- Common situation where some elements are unavailable → “missing data”

(b) Geometrical representation

- **For numerical data, common to use a “scatter plot”**
- **Advantages**
    - Allows structure of data to be intuitively represented
    - e.g. for classification – easy to determine separability. Visual clustering, etc.
- **Disadvantages**
    - Only really viable for 2D or 3D data
    - Really only useful for data visualization

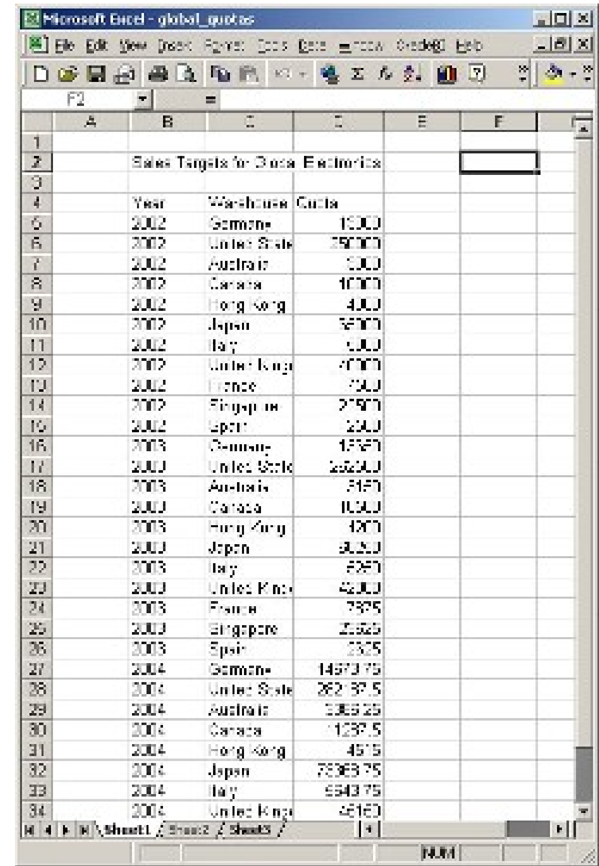




(c) Matrix representation

- **Closely related to spreadsheets**
  - Basically, concatenation of vector-formatted data
  - i.e. $n_{rows} \times n_{columns}$
  - In general, each row represents one record
  - Each column contains one “variable”, “field”, “feature” or “dimension”
- **Characteristics of data**
  - Many rows/Few columns - over-determined
  - Many rows/Many columns
  - Few rows/Many columns – under-determined

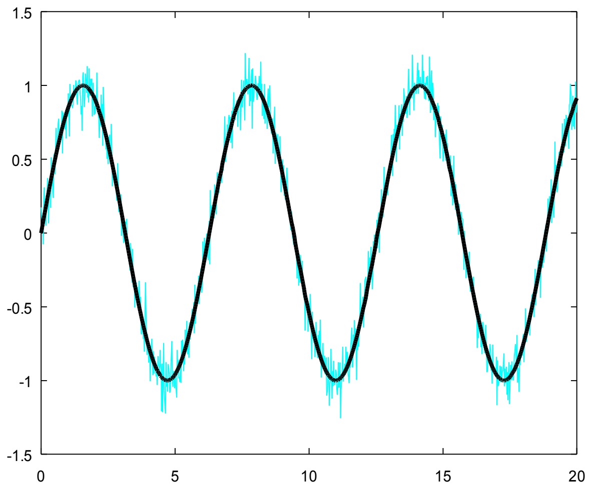
Q: Other data representations?



Masdar INSTITUTE

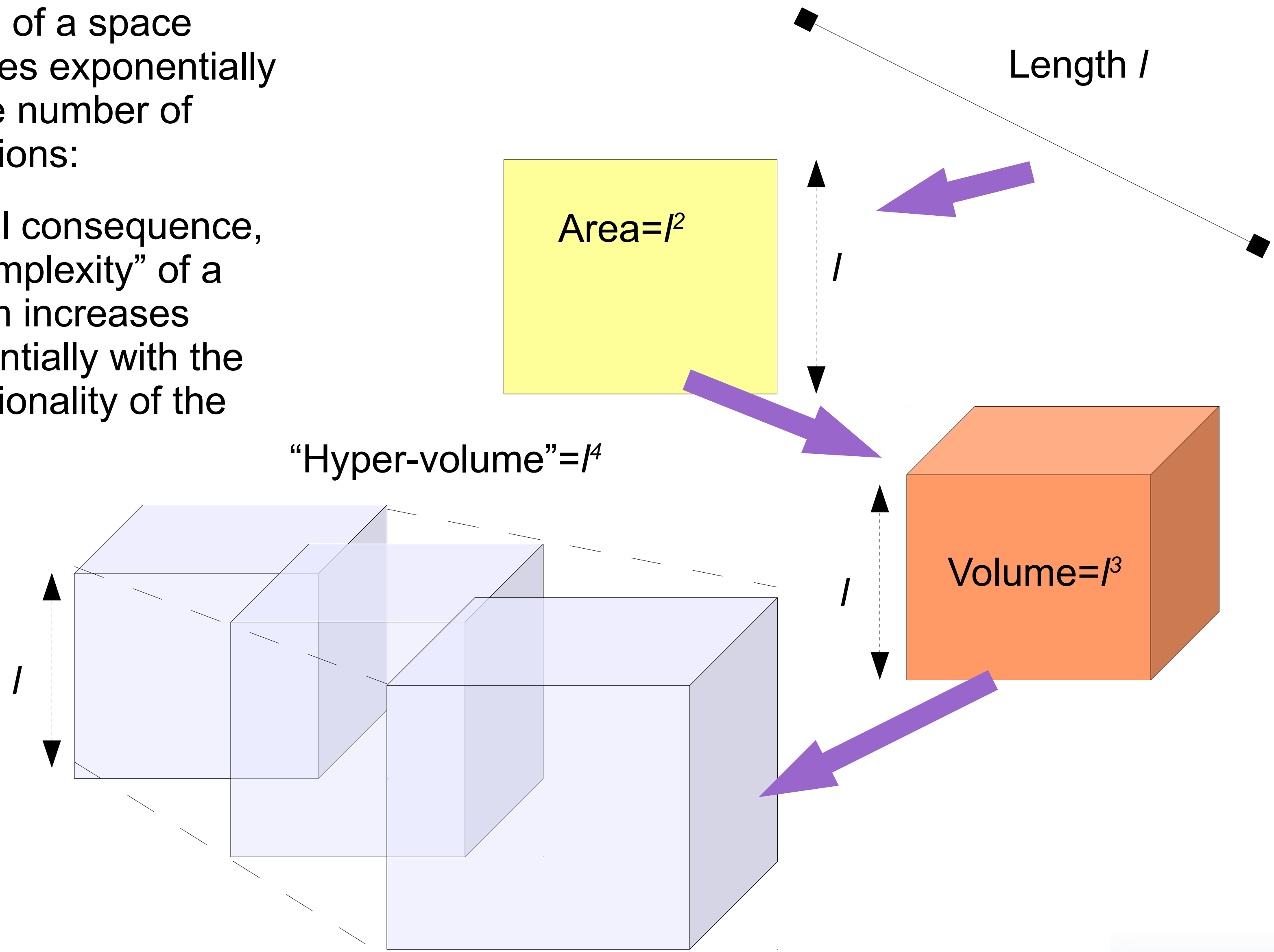
Noise/Data contamination

- **Noise: random error or unpredictability in a variable**
  - Most commonly in the form of an *additive* component, (as shown on right)
  - May also manifest in other forms, e.g. multiplicative noise, and as spoilt or ***missing data*** values.
  - Distinguish between observation noise, process and modelling noise.
- **Causes:**
  - Faulty or damaged instruments
  - Artifacts and faulty data collection (human error)
  - Physical limits in data collection (finite resolution, for e.g.)
- **Solutions – two general approaches**
  - Noise resistant algorithms, based on probabilistic principles
  - Pre-processing/filtering (sometimes manual)



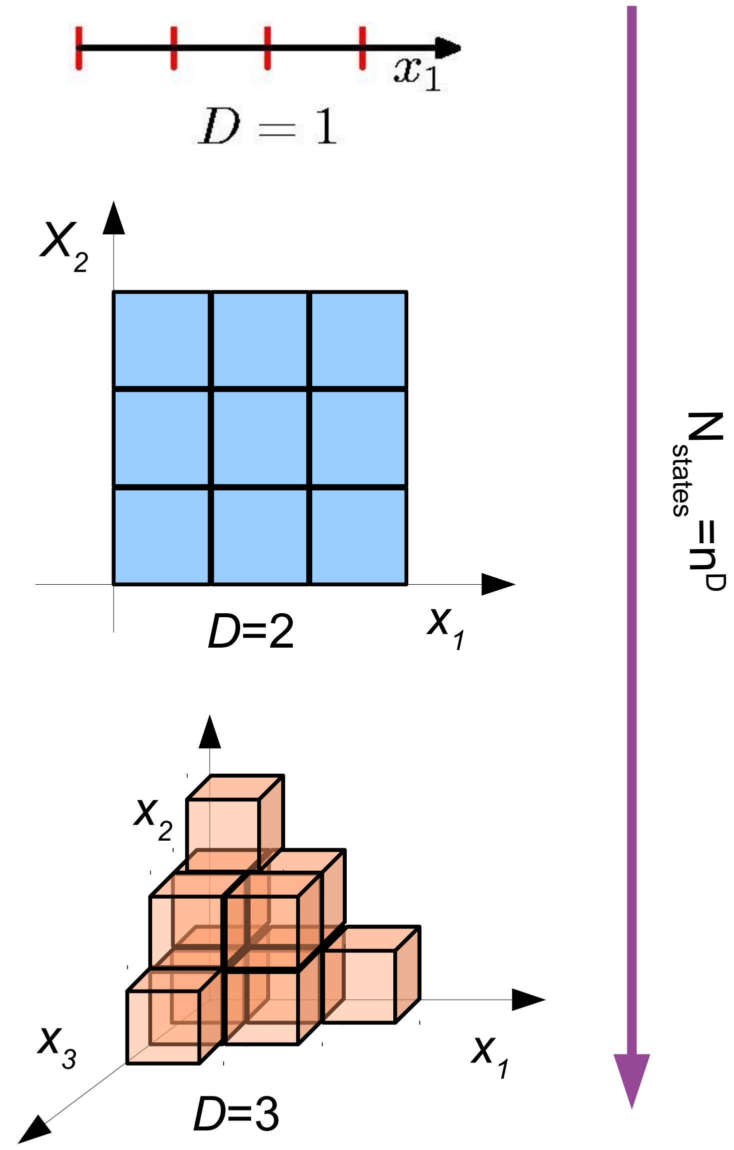
Curse of Dimensionality

- Volume of a space increases exponentially with the number of dimensions:
- General consequence, the “complexity” of a problem increases exponentially with the dimensionality of the data



Effects of the Curse (1/3)

- **Optimization: find minimum of a function**
  - Brute-force solution – search all states!
- **Explosion in the number of states to search**
  - Consider a line in [-5,5]
  - Using a grid search (spacing one), to cover this line we would require 10 points
  - For a square, this would be 100, cube 1000, etc.



Effects of the Curse (2/3)

- **Increase in the number of parameters to tune**
  - Models have more degrees of freedom → larger parameter spaces.
  - For e.g., 3 dimensional polynomial curve fitting:

$$y(x,w)=w_0+\sum_{i=1}^{D} w_i x_i+\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j+\ldots$$

$$\ldots+\sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

- **Difficulty in estimating the distributions of the data**
  - Number of parameters for a gaussian distribution is $n+n(n-1)/2$
  - Corresponding increase in the number of data points required

Effects (3/3)

- **Sparseness in the space of data**
  - Insufficient data →large uncertainty in models
- **Data is singular**
  - Data dimension exceeds the “intrinsic” dimensionality of the data
  - Numerical issues
  - Many of the dimensions are redundant → unnecessary wastage of computational power
  - Addition of dimensions beyond the “intrinsic dimensionality of the data results in addition of noise
- **“Distances” in data become less useful**
  - Almost all points are 'far away” - kills non-parametric methods!
  - Fundamental property of high dimensional spaces.