Information Theory and ID3 Algo.

Sohn Jong-Soo
mis026@korea.ac.kr
Intelligent Information System Lab.
2007.10.09

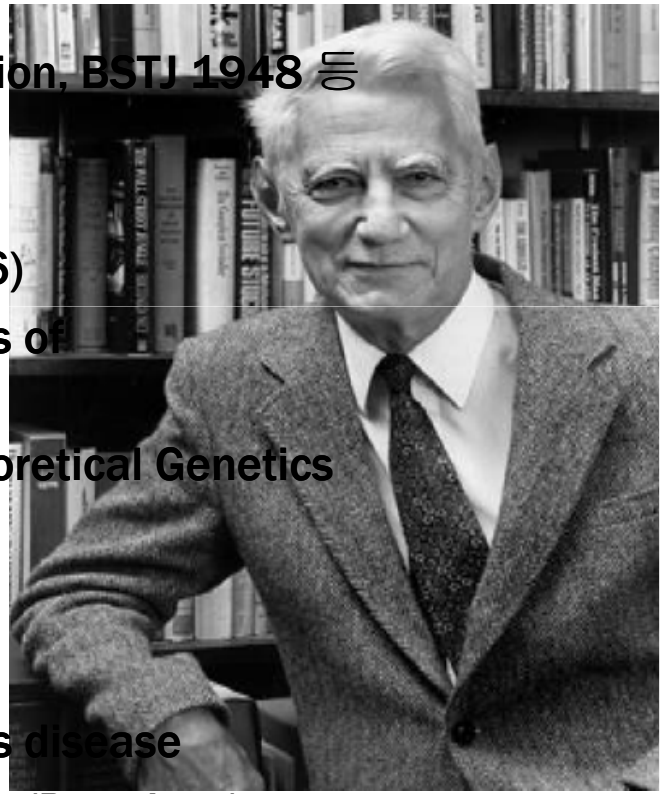
Claude Elwood Shannon

"The father of information theory"

- 본적 : **Petosky, Michigan**
- 논문 : **A Mathematical Theory of Communication, BSTJ 1948 등**
- 경력 : **Researcher at Bell Lab(1941~1972), MIT Professor(1956~2001)**
- 학력 : **B.S. in Univ. of Michigan (EE & AM 1936)**

  M.S. in MIT (1937) : A Symbolic Analysis of Relay and Switching Circuits

  Ph.D in MIT (1940) : An algebra for Theoretical Genetics
- 생년월일 : **1916.4.30**
- 직업 : **Electrical engineer, Mathematician**
- 병역 : **2**차 대전 시 암호 해독가로 활약
- 사망 : **2001.2.24, Medford, Mass. Alzheimer`s disease**
- 가족관계 : **Betty(Wife), Andrew(Son), Margarita(Daughter)**



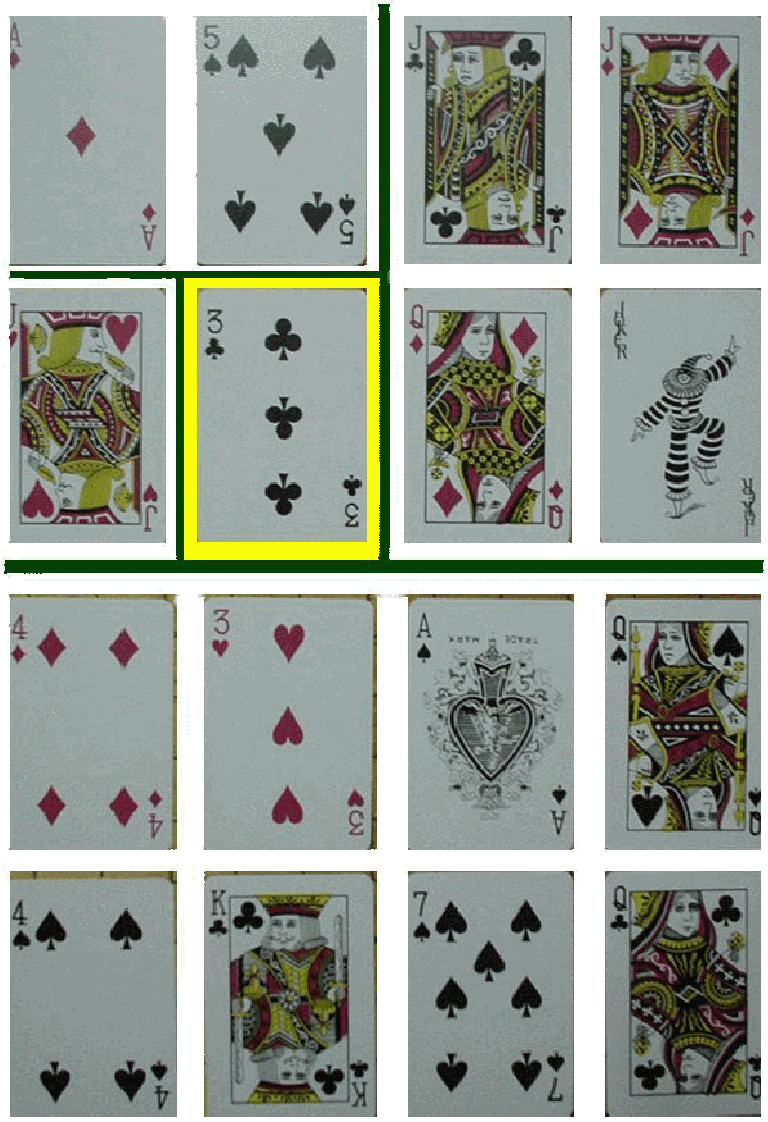
Fundamentals of Information Theory

- Three fundamental problems
  - Data compression
  - Sampling (Digitalizing)
  - Classification
- Q : How to measure information?
  - Concept of Entropy
    - 111111000000 : Low Entropy
    - 110101011010 : High Entropy
- 정보량 : (Bits)

  - 예 : 정상인 동전의 엔트로피 : 1 (Bits)
    - 비정상인 동전(P(앞면)=0.9)의 엔트로피 : 0.274 (Bits)
    - 로또(6/45)의 엔트로피 : 0.0000028185 (Bits)

Fundamentals of Information Theory

"여기 세로 가로 4장씩 16장의 트럼프가 있다. 이중 한 장만 머릿속에 점 찍어 두시오"
"예, 점 찍었습니다"
"그것은 상단에 있는가?"
"그렇습니다"
"그럼 상단의 오른쪽 반에 있는가?"
"아닙니다"
"그럼 왼쪽 반의 상단에 있는가?"
"아닙니다"
"그럼 하단의 오른쪽에 있는가?"
"그렇습니다"
"당신이 점 찍은 것은 크로바3이군요"



첫째,카드의 수.여기서는 16장이므로 아무런 정보도 없는 상황에서 알아맞힐 확률은 1/16

둘째,상대의 대답의 종류.여기서는 예,아니오 두 종류만 허용

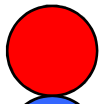
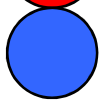
셋째,질문횟수.여기서는 4회.16장의 카드에서 특정 카드를 알아맞히는데 4회의 질문이 필요

Fundamentals of Information Theory

- $2^4 = 16$ (generalizing) → $W = 2^n$
  - W : number of cards
  - n : number of questions
- 16장의 카드에서 1장을 알아맞히는데
  - 4회의 질문이 필요
  - $log_2 16=4$
    - 16장의 카드에서 특정 1장을 고르는 정보량은 4bit
  - $logW = n\ log2 \rightarrow n = logW/log2$
  - $n = log_2 W$
  - 16장에서 1장을 고르는 확률은 1/16이므로 이 확률을 중심으로 정보량을 표현하면 $n = -log_2 1/16$
  - $n = -log_2 P$

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} log_2 \frac{s_i}{s}$$

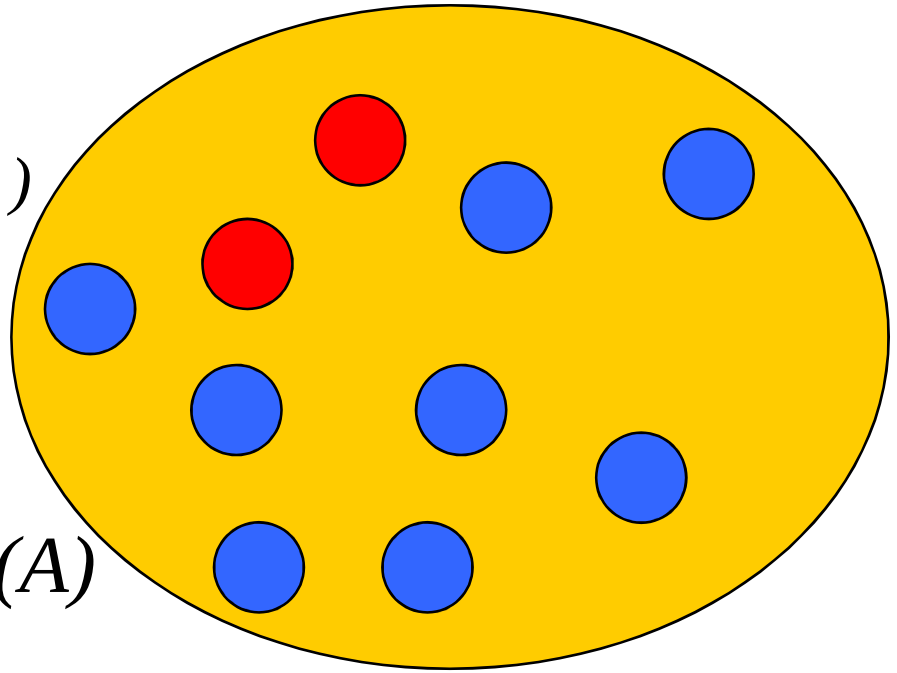
Fundamentals of Information Theory

-  $= - P_1 log_2 P_1$
-  $= - P_2 log_2 P_2$

- $E = - P_1 log_2 P_1 - P_2 log_2 P_2$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$$

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$



Entropy

- X: discrete random variable, p(X)
- Entropy (or self-information)

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- Entropy measures the amount of information in a RV
  - it's the average length of the message needed to transmit an outcome of that variable using the optimal code

Joint Entropy

- The joint entropy of 2 RV X,Y is the amount of the information
  - needed on average to specify both their values

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(X,Y)$$

Conditional Entropy

- The conditional entropy of a RV Y given another X
  - expresses how much extra information one still needs to supply on average to communicate Y given that the other party knows X
  - It's also called the **equivocation** (애매함) of *X* about *Y*

$$H(Y \mid X) = \sum_{x \in X} p(x) H(Y \mid X = x)$$

$$= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x) \log p(y \mid x)$$

$$= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y \mid x) = -E(\log p(Y \mid X))$$

Mutual Information

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X,Y)$$

- I(X,Y) is the mutual information between X and Y.
- It is the reduction of uncertainty of one RV due to knowing about the other,
  or the amount of information one RV contains about the other

Mutual Information (cont)

$$I(X, Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

- I is 0 only when X,Y are independent: H(X|Y)=H(X)
- H(X)=H(X)-H(X|X)=I(X,X) Entropy is the self-information

ID3 algorithm

ID3

- Nonincremental algorithm
  - meaning it derives its classes from a fixed set of training instances.
- The classes created by ID3 are inductive,
  - that is, given a small set of training instances, the specific classes created by ID3 are expected to work for all future instances.
- The distribution of the unknowns must be the same as the test cases. Induction classes cannot be proven to work in every case since they may classify an infinite number of instances.
- Note that ID3 (or any inductive algorithm) may misclassify data.

ID3 algorithm

Data Description

- The sample data used by ID3 has certain requirements, which are:
  - Attribute-value description
    - the same attributes must describe each example and have a fixed number of values.
  - Predefined classes
    - an example's attributes must already be defined, that is, they are not learned by ID3.
  - Discrete classes
    - classes must be sharply delineated. Continuous classes broken up into vague categories such as a metal being "hard, quite hard, flexible, soft, quite soft" are suspect.

ID3 algorithm

- Attribute Selection
  - information gain, is used
  - The one with the highest information (information being the most useful for classification) is selected.
  - Entropy measures the amount of information in an attribute.
- Given a collection S of c outcomes
- Entropy(S) = $\sum$-p(I) $\log_2$ p(I)
  - where p(I) is the proportion of S belonging to class I. S is over c. Log2 is log base 2.
  - Note that S is not an attribute but the entire sample set.

ID3 algorithm - Example

- should we play baseball?
  - Over the course of 2 weeks, data is collected to help ID3 build a decision tree (see following table)
  - The weather attributes are outlook, temperature, humidity, and wind speed. They can have the following values:
    - outlook = { sunny, overcast, rain }
    - temperature = {hot, mild, cool }
    - humidity = { high, normal }
    - wind = {weak, strong }

ID3 algorithm - Example

Day	Outlook	Temperature	Humidity	Wind	Play ball
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

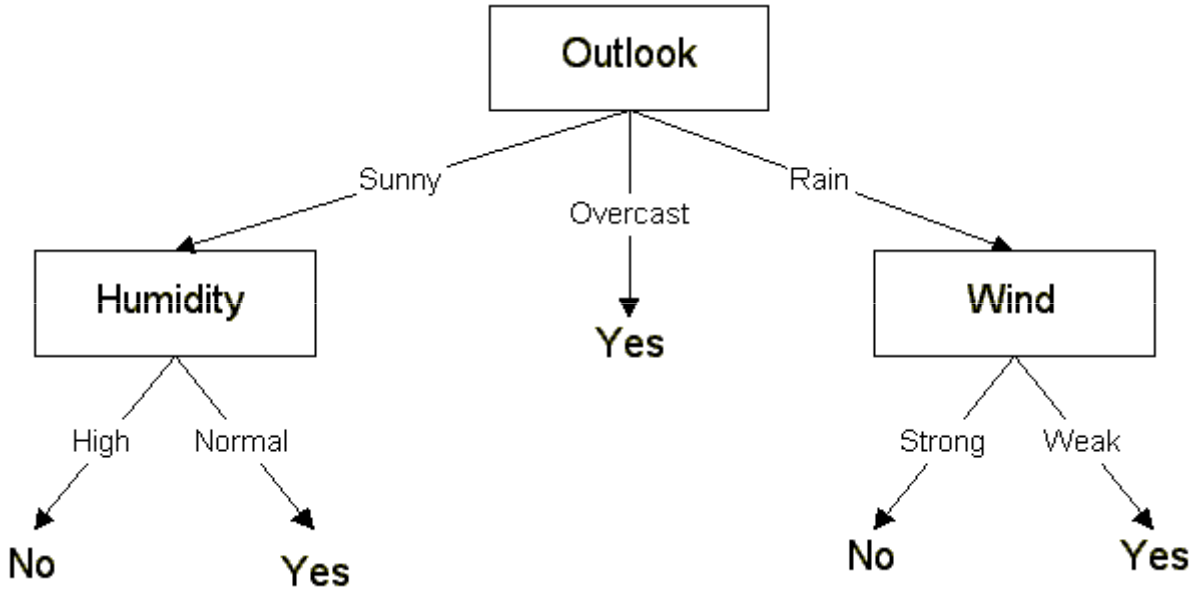
ID3 algorithm - Example

- We need to find which attribute will be the root node in our decision tree.
  - The gain is calculated for all four attributes:
    - Gain(S, Outlook) = 0.246
    - Gain(S, Temperature) = 0.029
    - Gain(S, Humidity) = 0.151
    - Gain(S, Wind) = 0.048
  - Outlook attribute has the highest gain, therefore it is used as the decision attribute in the root node.

ID3 algorithm - Example

- Since Outlook has three possible values, the root node has three branches (sunny, overcast, rain).
- The next question is "what attribute should be tested at the Sunny branch node?“
  - S_{sunny} = {D1, D2, D8, D9, D11} = 5
  - Gain(S_{sunny}, Humidity) = 0.970
  - Gain(S_{sunny}, Temperature) = 0.570
  - Gain(S_{sunny}, Wind) = 0.019
- Humidity has the highest gain
  - it is used as the decision node.

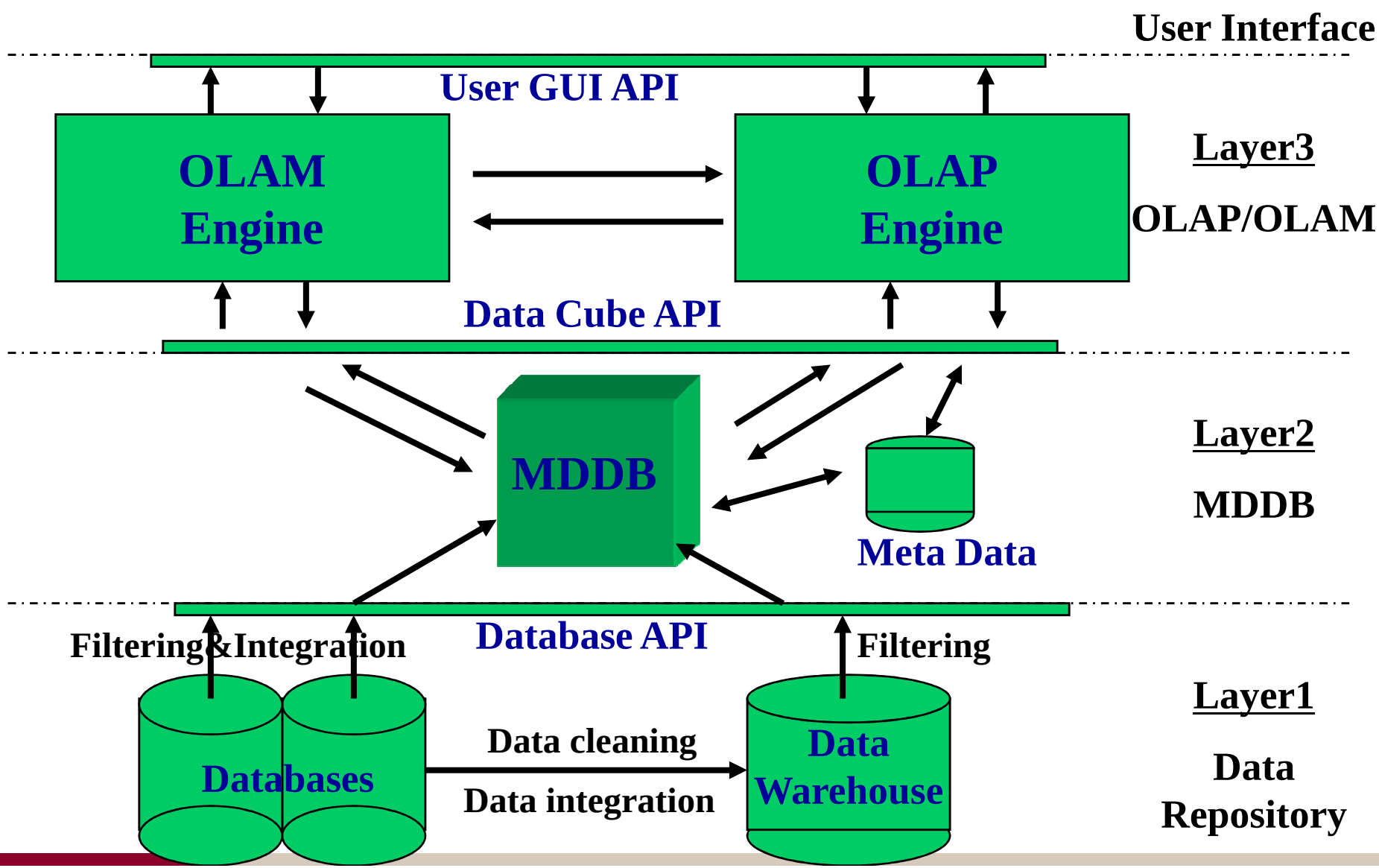
ID3 algorithm - Example



ID3 algorithm - Example

- The decision tree can also be expressed in rule format:
  - IF outlook = sunny AND humidity = high THEN playball = no
  - IF outlook = rain AND humidity = high THEN playball = no
  - IF outlook = rain AND wind = strong THEN playball = yes
  - IF outlook = overcast THEN playball = yes
  - IF outlook = rain AND wind = weak THEN playball = yes

Next presentation



Next presentation

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..