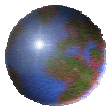
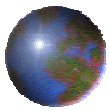
Statistical Data Mining



A Global View and Some Research Potentials

Dennis Lin
Supply Chain & Information Systems

21 February, 2006
OR-Seminar



Practical vs Theoretical: Clustering Example

Example from Jon Kettenring

Knowledge Discovery in Sciences

- Deduction
  - Physics
  - Mathematics
- Induction
  - Biology
  - Chemistry
- Data Mining is more useful in empirical study & experience accumulation, namely "induction" type.

Brief Historical Development

Optimization (SA, GA, *Neural Networks etc*)

Statistics (*classification tree*, *projection pursuit*)

⇒ Data Mining (data-base)

Artificial Intelligent (rule-base)

⇒ Knowledge Discovery

⇒ Machine Learning

Statistical Learning/Math Learning etc.

Data Mining

- **Data**

  (re-design and maintain existing database)

- **Mining**

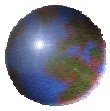
  (Analysis) -- our focus

  Statistical Data Mining

What is Data Mining?

- *Data mining* is a process that uses a variety of data analysis tools to discover patterns and relationships in data.
- Viewed as part of the *Knowledge Discovery* process.
- Uses tools from Computer Science and Artificial Intelligence as well as Statistics.

DeVeaux



Complexity

- Data Complexity
- Algorithm Complexity

Wegman

Complexity

Descriptor	Data Set Size in Bytes	Storage Mode
Tiny	10^2	Piece of Paper
Small	10^4	A Few Pieces of Paper
Medium	10^6	A Floppy Disk
Large	10^8	Hard Disk
Huge	10^{10}	Multiple Hard Disks
Massive	10^{12}	Robotic Magnetic Tape Storage Silos
Supermassive	10^{15}	Distributed Data Archives

The Huber-Wegman Taxonomy of Data Set Sizes

Wegman

Complexity

Algorithmic Complexity

$O(n^{1/2})$	Plot a Scatterplot
$O(n)$	Calculate Means, Variances, Kernel Density Estimates
$O(n \log(n))$	Calculate Fast Fourier Transforms
$O(n\ c)$	Calculate Singular Value Decomposition of an r x c Matrix; Solve a Multiple Linear Regression
$O(n^{2})$	Solve most Clustering Algorithms
$O(a^{n})$	Detect Multivariate Outliers

Wegman

Complexity

Number of Operations for Algorithms of Various Computational Complexities and Various Data Set Sizes

n	$n^{1/2}$	n	$n \log(n)$	$n^{3/2}$	n^{2}
tiny	10	10^{2}	$2x10^{2}$	10^{3}	10^{4}
small	10^{2}	10^{4}	$4x10^{4}$	10^{6}	10^{8}
medium	10^{3}	10^{6}	$6x10^{6}$	10^{9}	10^{12}
large	10^{4}	10^{8}	$8x10^{8}$	10^{12}	10^{16}
huge	10^{5}	10^{10}	10^{11}	10^{15}	10^{20}

Wegman

Complexity

Computational Feasibility on a Pentium PC
10 megaflop performance assumed

n	$n^{1/2}$	*n*	*n log(n)*	$n^{3/2}$	n^2
tiny	10^{-6} seconds	10^{-5} seconds	$2x10^{-5}$ seconds	.0001 seconds	.001 seconds
small	10^{-5} seconds	.001 seconds	.004 seconds	.1 seconds	10 seconds
medium	.0001 seconds	.1 seconds	.6 seconds	1.67 minutes	1.16 days
large	.001 seconds	10 seconds	1.3 minutes	1.16 days	31.7 years
huge	.01 seconds	16.7 minutes	2.78 hours	3.17 years	317,000 years

Wegman

Complexity

Computational Feasibility on a Silicon Graphics Onyx Workstation
300 megaflop performance assumed

n	$n^{1/2}$	*n*	*n log(n)*	$n^{3/2}$	n^2
tiny	$3.3x10^{-8}$ seconds	$3.3x10^{-7}$ seconds	$6.7x10^{-7}$ seconds	$3.3x10^{-6}$ seconds	$3.3x10^{-5}$ seconds
small	$3.3x10^{-7}$ seconds	$3.3x10^{-5}$ seconds	$1.3x10^{-4}$ seconds	$3.3x10^{-3}$ seconds	.33 seconds
medium	$3.3x10^{-6}$ seconds	$3.3x10^{-3}$ seconds	.02 seconds	3.3 seconds	55 minutes
large	$3.3x10^{-5}$ seconds	.33 seconds	2.7 seconds	55 minutes	1.04 years
huge	$3.3x10^{-4}$ seconds	33 seconds	5.5 minutes	38.2 days	10,464 years

Wegman

Statistical Data Mining

- Need methodologies/algorithms that is *computable* (under the constraints of computer memory and complexity).
- So
  - All methodologies need to be labeled its complexity.
  - For powerful $O(n^2)$ methodologies, an approximate $O(n)$ algorithm is needed.

Recent Projects Involved

- Chinese Medicine Study (Fang)
- Teaching Evaluation (Wang)
- Horse Racing Wagering Market (Gu)

Chinese Medicine

川
鳥

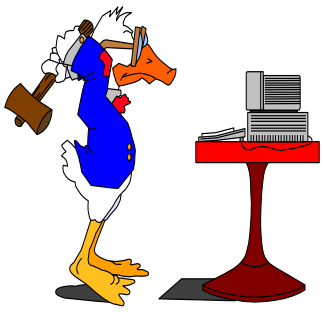
	Excellent	Very Good	Good	Fair	Poor
Instructor is interested in your learning	●	○	○	○	○
Respect and concern for students	●	○	○	○	○
Expresses expectations for student performance	●	○	○	○	○
Communication of ideas	●	●	○	○	○
Organization of the course	○	●	●	○	○
Facilitation of learning	○	●	●	●	●
Use of class time	●	●	○	○	○

Teaching Evaluation Study

Horse Racing Wagering Market

Emphasis on Automatic

- Computer users are domain experts, not computer professionals
  - Too much data
  - Too much technology
  - Not enough useful information
- *When is Automatic Analysis Useful?*



DeVeaux

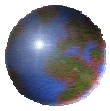
Why Now?

- Computer power
- The price of digital storage
- Data warehouses are being widely implemented
- Individualized marketing strategies

DeVeaux

Problems

- Classification (Supervised Leaning)
- Clustering (Unsupervised Learning)
- Pattern Recognition
- Association (Correlation)
- Modeling
- Estimation
- Prediction
- Description
- Visualization
- Etc.



Data Mining tools

Computer Science Root

- Database
- *ANN: Artificial Neural Network*
- MBA: Market Basket Analysis (Association Rule)
- Genetic Algorithms
- OLAP: On-Line Analytic Processing
- Link Analysis
- High Dimensional Plots
- KDD Process
- Machine Learning
- Text Mining

Statistical Root

- *(Linear & Nonlinear) Regression*
- Logistic Regression
- Classification
- Clustering
- Density Estimation, Bumps and Ridges
- Time Series: Trend & Spectral Estimation
- Dimension Reduction Techniques
- High Dimensional Plots
- Classical Multivariate Methods

Statistical & Computer Science Mixture

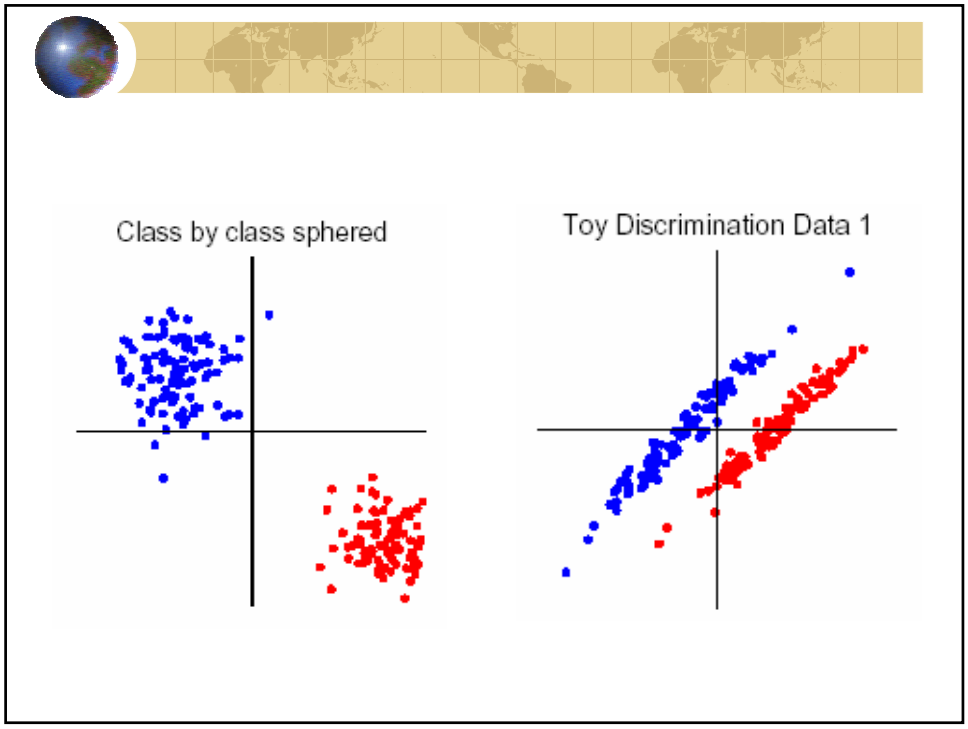
- Data Visualization
- Dimension Reduction
  - Multi-dimensional Scaling
  - Local Linear Embedding
  - Principal Component & Principal Manifolds
  - Local Tangent Space Alignment
- Independent Component Analysis
- K-means & kd-Tree
- Support Vector Machine

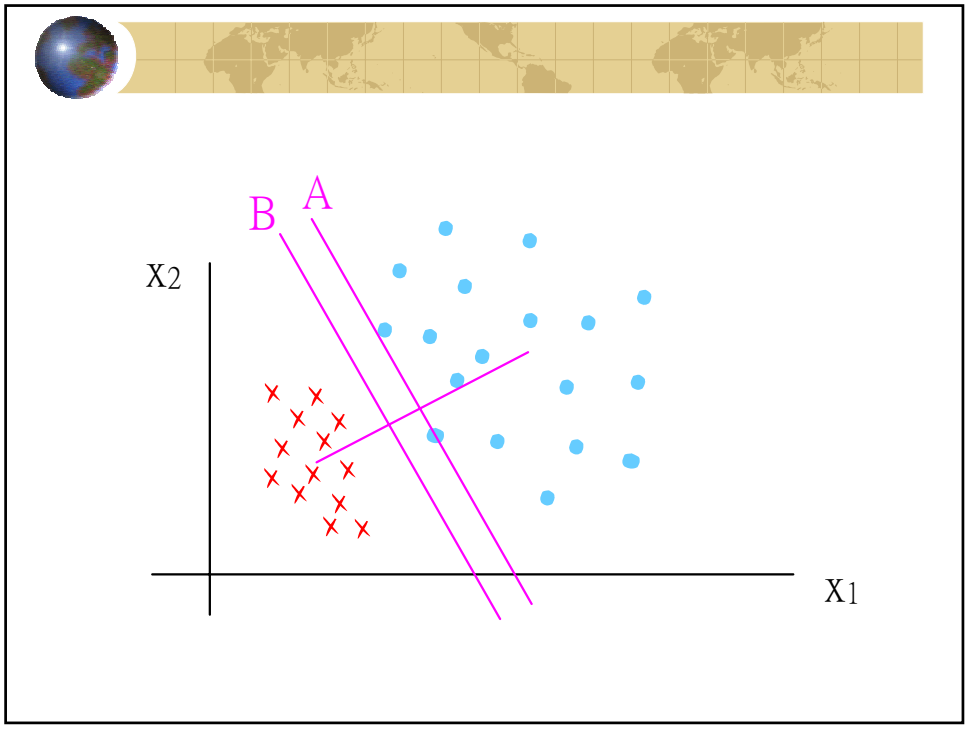
Example:

Methodologies for Classification

Classification Problem

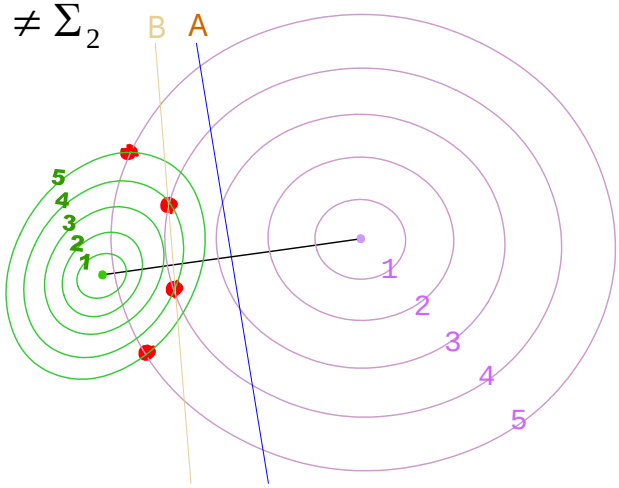
- Regression
- Logistic Regression
- Artificial Neural Networks
- Market Basket Analysis
- Multivariate Classification
- Tree Classification
- Support Vector Machine


Class by class sphered
Toy Discrimination Data 1


B
A
X2
X1

Unequal variance

$\Sigma_1 \neq \Sigma_2$

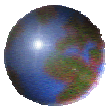


Classification

Classified

Actual		π_1	π_2
	π_1	0	C(2|1)
	π_2	C(1|2)	0

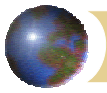
ECM (Expected Cost of Misclassification)
$=C(2|1)P(2|1)P_1+C(1|2)P(1|2)P_2$



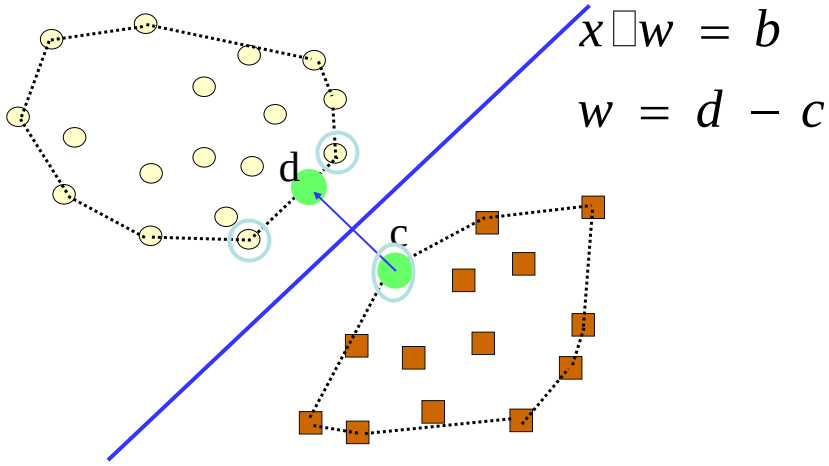
SVM: Support Vector Machine

Dimension Reduction

Dimension Expansion



Plane Bisect Closest Points



$$x \cdot w = b$$

$$w = d - c$$

Maximize margin using quadratic program

$$\min_{w,b} \quad \frac{1}{2}\| w \|^2$$

$$s.t \quad x_i \cdot w \geq b+1 \quad i \in Class\ 1$$

$$x_i \cdot w \leq b-1 \quad i \in Class\ -1$$

Dual of Closest Points Method is Support Plane Method

$$\min_{\alpha} \ \frac{1}{2}\Big\| \sum_{i=1}^{\ell} y_i \alpha_i x_i \Big\|^2 \qquad \min_{w,b} \ \frac{1}{2}\| w \|^2$$

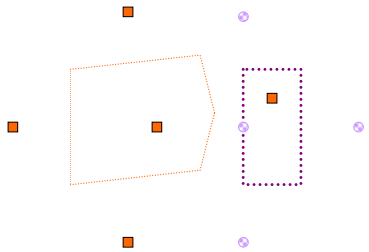
$$s.t. \quad \sum_{i \in 1} \alpha_i = 1 \quad \sum_{i \in -1} \alpha_i = 1 \quad \Leftrightarrow \quad s.t. \quad y_i \left(x_i \cdot w - b \right) \geq 1$$

$$\alpha \geq 0 \qquad\qquad i = 1,..,\ell$$

Solution only depends on support vectors: $\alpha_i > 0$

$$w = \sum_{i=1}^{\ell} y_i \alpha_i x_i \qquad y_i := \begin{cases} 1 & i \in Class\ 1 \\ -1 & i \in Class\ -1 \end{cases}$$

Reduced Convex Hulls-- Don't Intersect



$$D = \frac{1}{2}$$

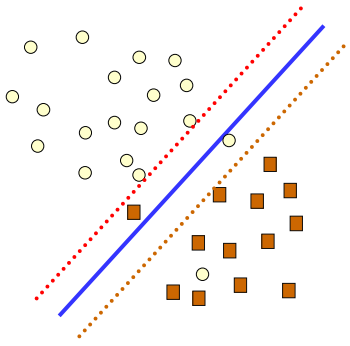
$$d = \sum_{i \in Class1} \alpha_i x_i$$

$$\sum_{i \in Class1} \alpha_i = 1$$

$$0 \le \alpha_i \le D$$

Reduce by adding upper bound D

Linearly Inseparable Case: Supporting Plane Method



Just add non-negative error vector *z*.

$$\min_{w,b,z} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell} z_i$$

$$s.t \quad y_i\left(x_i \cdot w + b\right) + z_i \ge 1$$

$$z_i \ge 0 \quad i = 1,..,\ell$$

Dual of Closest Points Method is Support Plane Method

$$\min_{\alpha} \ \frac{1}{2}\Big\| \sum_{i=1}^{\ell} y_i \alpha_i x_i \Big\|^2 \qquad \min_{w,b} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell} z_i$$

$$s.t. \quad \sum_{i\in 1} \alpha_i = 1 \quad \sum_{i\in -1} \alpha_i = 1 \quad \Leftrightarrow \quad s.t. \quad y_i\left(x_i \cdot w - b\right) + z_i \geq 0$$

$$D \geq \alpha_i \geq 0 \qquad z_i \geq 0 \quad i = 1,..,\ell$$

Solution only depends on support vectors: $\alpha_i > 0$

$$w = \sum_{i=1}^{\ell} y_i \alpha_i x_i$$

Final Classification via Kernels

The Dual SVM becomes:

$$\min_{\alpha} \ \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{\ell} \alpha_i$$

$$s.t. \quad \sum_{i=1}^{\ell} y_i \alpha_i = 0$$

$$\alpha_i \geq 0 \quad i = 1,..,\ell$$

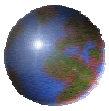
Some Potential Research Projects

- Kernel Method
- Dimension Reduction & Dimension Expansion
- Selection of Tuning Parameters
- Rare Event Classification
- Rare Subject Classification
- More than Two Classes

SVM Software

http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Also check the WEKA website
<http://www.cs.waikato.ac.nz/ml/weka/>



Link Analysis

Recommendation Systems

Process Mining

Process vs Products

What are the issues here?

- How do you “quantify” (measure) your observations (people or dog)?
- How do you “characterize” your observations?
- How do you classify (match) them?
- Others?

Two Examples (Illustration):

Titanic

2004 Taiwan Presidential Election

Titanic Data

(a total of 2208 cases)

Survived	Age	Gender	Class
D	C	M	3
D	C	M	3
D	C	M	3
D	C	M	3
D	C	M	3

The Titanic Data

DeVeaux

Age By Survived

Count Col % Row %	D	S	
A	1438 96.51 68.51	661 92.06 31.49	2099
C	52 3.49 47.71	57 7.94 52.29	109
	1490	718	2208

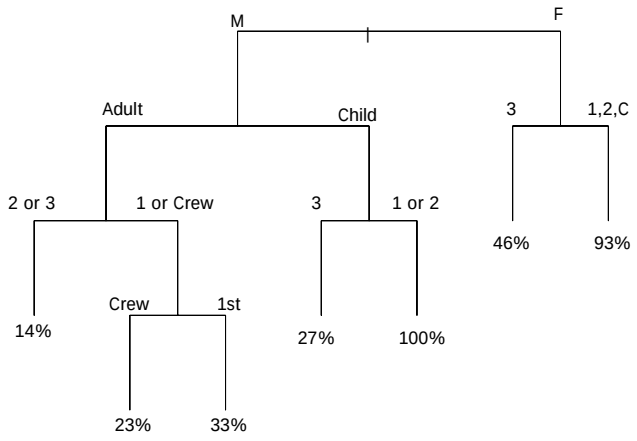
Gender By Survived

Count Col % Row %	D	S	
F	126 8.46 26.81	344 47.91 73.19	470
M	1364 91.54 78.48	374 52.09 21.52	1738
	1490	718	2208

Class By Survived

Count Col % Row %	D	S	
1	122 8.19 37.54	203 28.27 62.46	325
2	167 11.21 58.60	118 16.43 41.40	285
3	528 35.44 74.79	178 24.79 25.21	706
Crew	673 45.17 75.45	219 30.50 24.55	892
	1490	718	2208

Tree model



2004 Taiwan Presidential Election

Theoretical Model

- Suppose the probability for *purely* human error is p.
- The actual count for Candidate A is n_1, among them X_1 was miscounted (to Cand-B).
- The actual count for Candidate B is n_2, among them X_2 was miscounted (to Cand-A).
- X_1 follows Binomial distribution with n_1 & p.
- X_2 follows Binomial distribution with n_2 & p.
- Since there is only human error exists, a common probability for miscount (p) is used.
- True Difference: $D = n_2 - n_1 - 2(X_2 - X_1)$

Theoretical Model

- $E(D) = (N - 2n_1)(1 - 2p)$
- $Var(D) = 2\sqrt{Np(1-p)}$

- More complicated but realistic Model

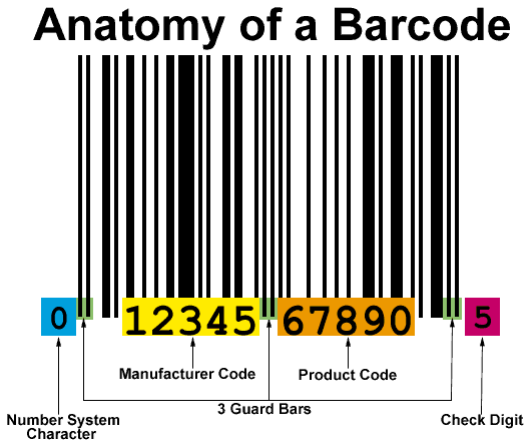
  X_1, X_2, V_1, V_2, J_1 & J_2
- $E(D) = n_2 - n_1 - 2E[X_2 - X_1] - E[V_2 - V_1] + E[J_2 - J_1]$

  $= n_2(1-2p_{21}-p_{23}) - n_1(1-2p_{12}-p_{13}) + n_3(p_{32}-p_{31})$
- $Var(D) = n_1[4p_{12}(1-p_{12}) + p_{13}(1-p_{13}) - 4p_{12}p_{13}] + n_2[4p_{21}(1-p_{21}) + p_{23}(1-p_{23}) - 4p_{21}p_{23}] + n_3[p_{32}(1-p_{32}) + p_{31}(1-p_{31}) + 2p_{31}p_{32}]$

Two More Examples (Ongoing):

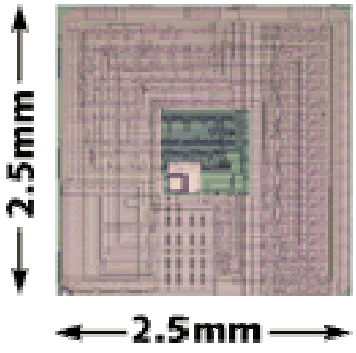
RFID &
Search Engine

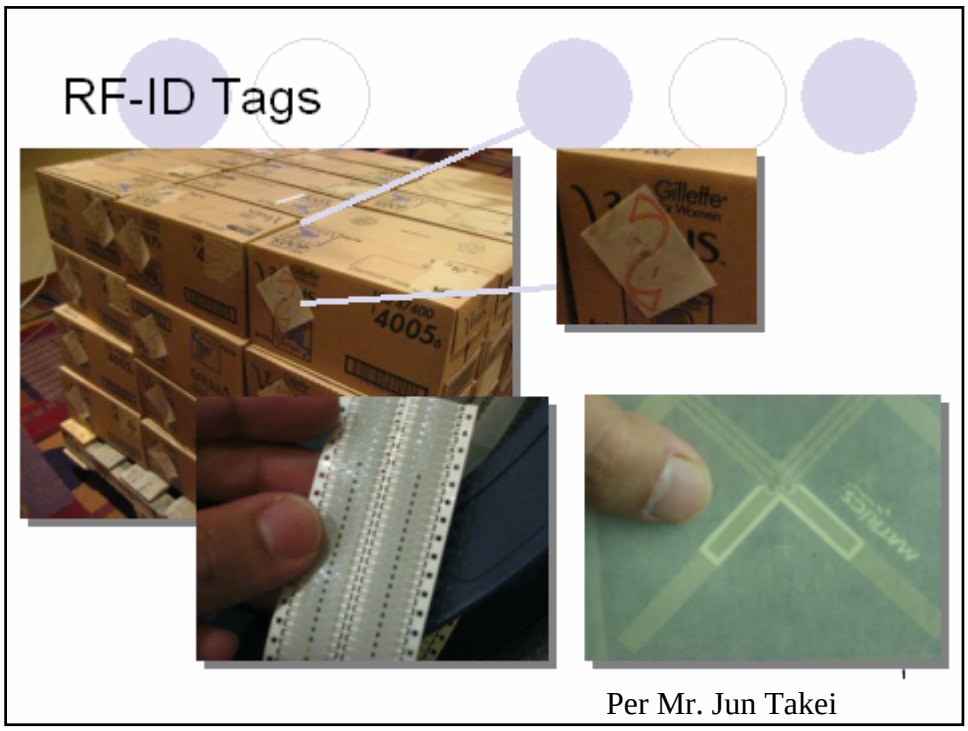
Bar Code



RFID:

Radio Frequency Identification





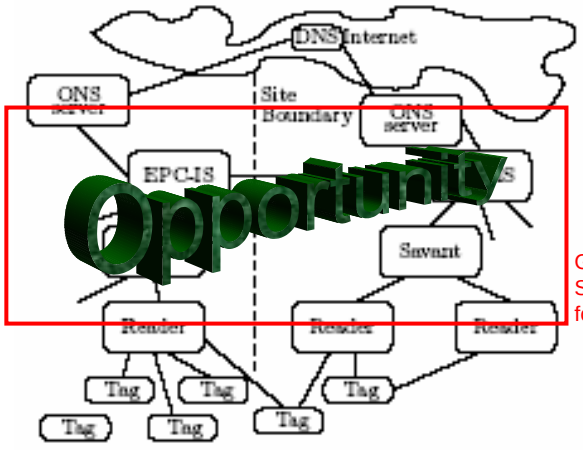
Per Mr. Jun Takei

RF Communication

- Electromagnetic waves modulated to carry data/signals
- Two different ways to generate ways
  - Inductive coupling
    - Close proximity electromagnetic wave
  - Propagating electromagnetic waves
- The fundamental RF communication theories apply—nothing new.
- New: the cost, size, signal processing capability.

J. Shu

Architecture

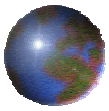


ONS: object name server
Maps EPC→ URL

EPC Information service
Higher level service for apps

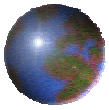
Gather data from readers:
Smoothing, coordination,
forwarding,etc.

Source: Chawathe, et al, VLDB Conference proceedings, 2004



A
Data Explosion
is Coming!

Are You Ready?



Impact to Statistics

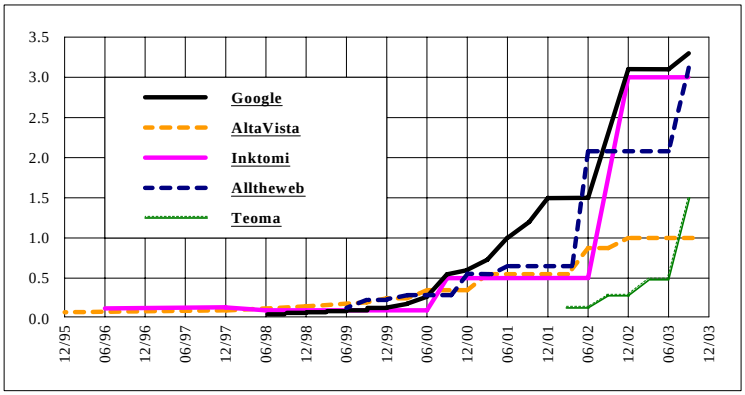
How to analyze the population data?

RFID vs. Barcode

Lin, Dennis K.J. and Wadhwa, Vijay (2005)

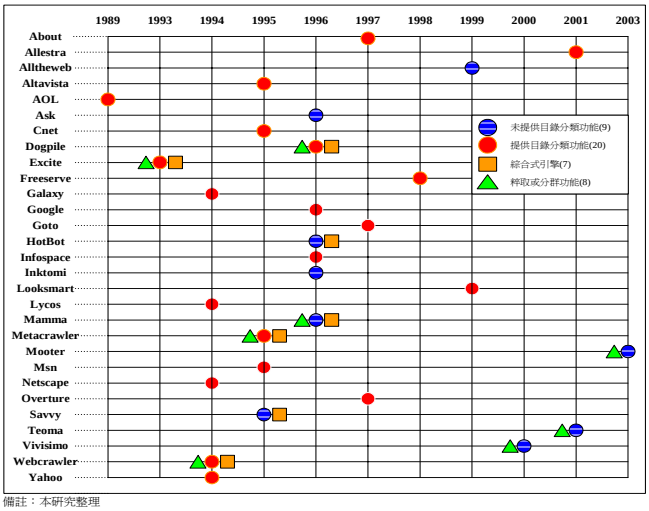
Search Engine & Citation Index Analysis

New Era for Search Engine



Unit: One Billion

History of Search Engine



備註：本研究整理

Google's Page Rank formula

- The PageRank of a page A is given as follows:

$$PR_1(A) = (1-d) + d \times \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right)$$

- PR(A) is the PageRank of page A;
- PR(Ti) is the PageRank of pages Ti which link to page A;
- C(Ti) is the number of outbound links on page Ti;
- d is a damping factor which can be set between 0 and 1; usually set to **0.85**
- n is the total number of all pages which link to page A.

Markov Chains

- Matix A

$$a_{ij} = \frac{(1-d)}{N} + d\frac{g_{ij}}{c_j} \qquad d=0.85$$

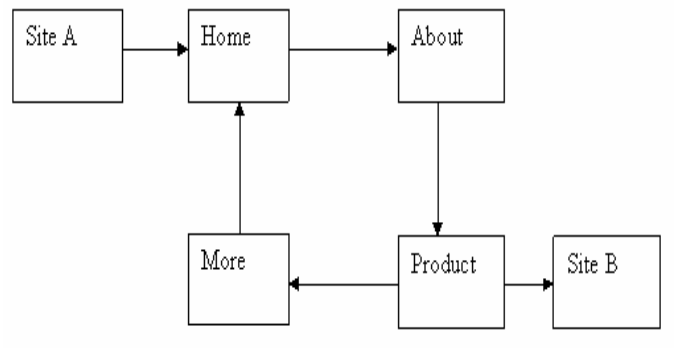
- Matix A max eignvalue =1

$$Ax=x \qquad \sum_i x_i = 1$$

- Matix A eignvector = PageRank(k)

$$x_k = \sum_{j=1}^{N} a_{kj} x_j = \frac{(1-d)}{N} + d \sum_{g_{kj}=1} \frac{x_j}{c_j}$$

Example 5



Example 5

$$H = (1-d) + d(\frac{M}{1} + \frac{SA}{1})$$

$$A = (1-d) + d(\frac{H}{1})$$

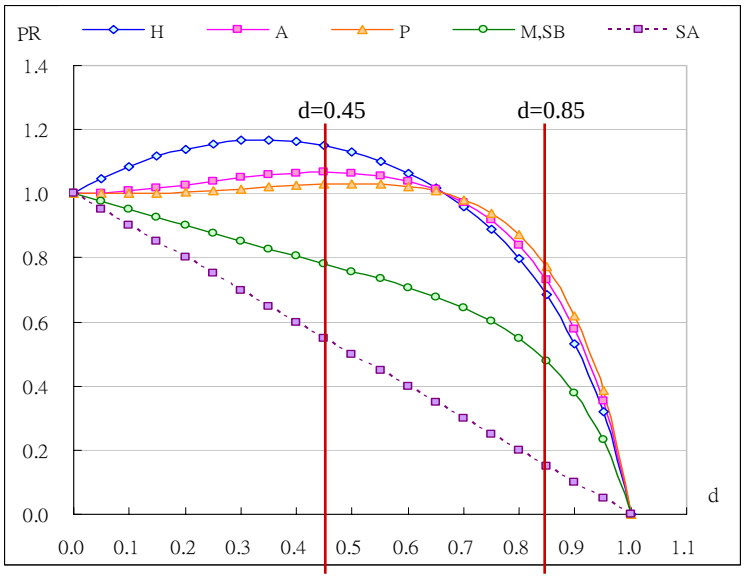
$$P = (1-d) + d(\frac{A}{1})$$

$$M = (1-d) + d(\frac{P}{2})$$

$$SA = (1-d)$$

$$SB = (1-d) + d(\frac{P}{2})$$

Example 5



How to Increase your PageRank?

How to Increase Your Paper Citation?

- Individual article
- Journal
- How is this (Impact Factor) related to PageRank?

Computer Experiment

How to run expensive simulation?

Beattie and Lin (2005)

Statistics vs. Engineering Models

$$y = f(x, \theta) + \varepsilon$$

Statistical Model

$$y = \beta_0 + \Sigma \beta_i x_i + \Sigma \beta_{ij} x_i x_j + \varepsilon$$

Two-Cents on Random Number Generation

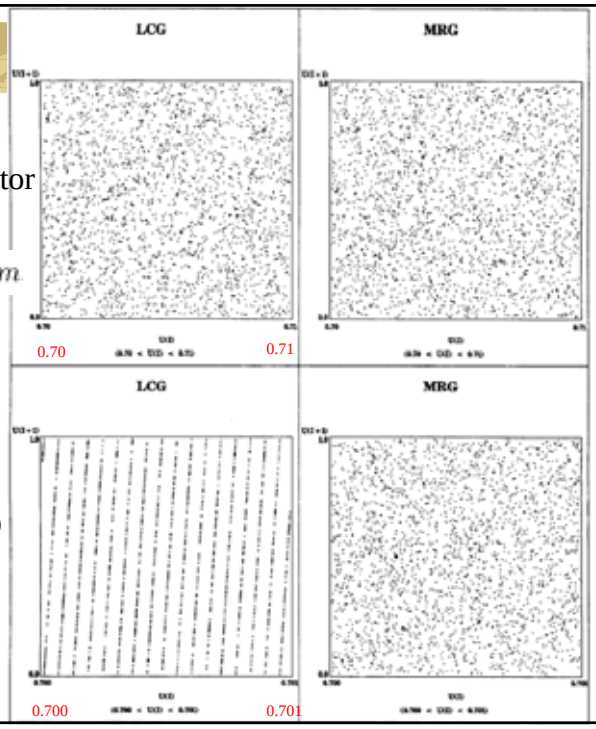
- Random Number Generator
  - Deng and Lin (2000, *The American Statistician*)
  - Deng, Lin, Wang & Yuan (1997, Stat Sinica)
- Transformation to Non-Uniform Distribution
  - Example: From *U(0,1)* to *N(0,1)*

Criteria for a "Good" RNG

- Period Length
- Computing Efficiency
- Randomness (independence)
- Portability
- Empirical Performance

LCG=
Linear Congruent Generator
Length =2^{31}-1 (=2.1*10^9)

$$X_i = (BX_{i-1} + A) \mod m$$



MRG=
Multiple Recursive Generator
Length =2.1* 10^{37} (for k=4)

Deng and Lin (2000)

A Typical Engineering Model (page 1 of 3, in Liao and Wang, 1995)

$$\rho_b A_b \frac{\partial^2 w}{\partial t^2} + E_b I_b \frac{\partial^4 w}{\partial x^4}$$
$$+\{(\rho_t A_t + \rho_c A_c)\frac{\partial^2 w}{\partial t^2} + \rho_t A_t(\frac{t_b + t_t}{2})(\frac{\partial^3 u_b}{\partial x \partial t^2} - \frac{t_b + t_t}{2}\frac{\partial^4 w}{\partial x^2 \partial t^2} + \frac{t_t}{2}\frac{\partial^3 \beta}{\partial x \partial t^2})$$
$$+\rho_c A_c a(\frac{\partial^3 u_b}{\partial x \partial t^2} - a\frac{\partial^4 w}{\partial x^2 \partial t^2} + t_t\frac{\partial^3 \beta}{\partial x \partial t^2}) + C_{11}^0 I_c\frac{\partial^4 w}{\partial x^4} - E_c A_c a(\frac{\partial^3 u_b}{\partial x^3} - a\frac{\partial^4 w}{\partial x^4} + t_t\frac{\partial^3 \beta}{\partial x^3})][H(x-x_1) - H(x-x_2)] \quad (1)$$
$$+\{\rho_t A_t(\frac{t_b + t_t}{2})(\frac{\partial^2 u_b}{\partial t^2} - \frac{t_b + t_t}{2}\frac{\partial^3 w}{\partial x \partial t^2} + \frac{t_t}{2}\frac{\partial^3 \beta}{\partial t^2}) + \rho_c A_c a(\frac{\partial^2 u_b}{\partial t^2} - a\frac{\partial^3 w}{\partial x \partial t^2} + t_t\frac{\partial^2 \beta}{\partial t^2})$$
$$+2C_{11}^0 I_c\frac{\partial^3 w}{\partial x^3} - 2E_c A_c a(\frac{\partial^2 u_b}{\partial x^2} - a\frac{\partial^3 w}{\partial x^3} + t_t\frac{\partial^2 \beta}{\partial x^2})][\delta(x-x_1) - \delta(x-x_2)]$$
$$+\{C_{11}^0 I_c\frac{\partial^2 w}{\partial x^2} - E_c A_c a(\frac{\partial u_b}{\partial x} - a\frac{\partial^2 w}{\partial x^2} + t_t\frac{\partial \beta}{\partial x}) + bd_{31}E_c aV(t)][\delta'(x-x_1) - \delta'(x-x_2)] = f(x,t)$$

$$\rho_b A_b \frac{\partial^2 u_b}{\partial t^2} - E_b A_b \frac{\partial^2 u_b}{\partial x^2}$$
$$+\{\rho_t A_t(\frac{\partial^2 u_b}{\partial t^2} - \frac{t_b + t_t}{2}\frac{\partial^3 w}{\partial x \partial t^2} + \frac{t_t}{2}\frac{\partial^2 \beta}{\partial t^2}) + \rho_c A_c(\frac{\partial^2 u_b}{\partial t^2} - a\frac{\partial^3 w}{\partial x \partial t^2} + t_t\frac{\partial^2 \beta}{\partial t^2})$$
$$-E_c A_c(\frac{\partial^2 u_b}{\partial x^2} - a\frac{\partial^3 w}{\partial x^3} + t_t\frac{\partial^2 \beta}{\partial x^2})][H(x-x_1) - H(x-x_2)] \quad (2)$$
$$+\{-E_c A_c t_c(\frac{\partial u_b}{\partial x} - a\frac{\partial^2 w}{\partial x^2} + t_t\frac{\partial \beta}{\partial x}) + bd_{31}E_c V(t)][\delta(x-x_1) - \delta(x-x_2)] = 0$$

$$[\rho_t A_t\frac{t_t}{2}(\frac{\partial^2 u_b}{\partial t^2} - \frac{t_b + t_t}{2}\frac{\partial^3 w}{\partial x \partial t^2} + \frac{t_t}{2}\frac{\partial^2 \beta}{\partial t^2}) + \rho_c A_c t_t(\frac{\partial^2 u_b}{\partial t^2} - a\frac{\partial^3 w}{\partial x \partial t^2} + t_t\frac{\partial^2 \beta}{\partial t^2})$$
$$+A_t(G*\beta) - E_c A_c t_t(\frac{\partial^2 u_b}{\partial x^2} - a\frac{\partial^3 w}{\partial x^3} + t_t\frac{\partial^2 \beta}{\partial x^2})][H(x-x_1) - H(x-x_2)] \quad (3)$$

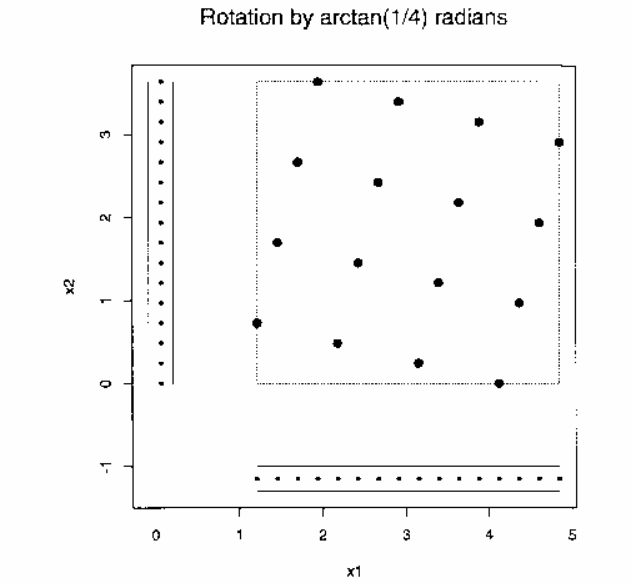
A Structured Roadmap for Verification and Validation--Highlighting the Critical Role of Experiment Design

James J. Filliben

National Institute of Standards and Technology
Information Technology Division
Statistical Engineering Division

2004 Workshop on Verification & Validation of Computer Models of High-Consequence Engineering Systems
NIST Administration Building
Lecture Room D
3:10-3:25, November 8, 2004

Beattie and Lin (1998)



- Rotation Theorem
- Orthogonality Theorem

Statistical Data Mining (Research Potentials)

- Fundamental
  - Population vs Sampling Data

  Using All Data or using some data?
- Data
  - Sampling
  - Data Squashed
- Complexity
  - Algorithm
  - Methodology
- Improved CS Mining Techniques

STILL QUESTION?

Send $500 to

- Dennis Lin
  483 Business Building
  Department of Supply Chain & Information Systems
  Penn State University

- +1 814 865-0377 (phone)
- +1 814 863-7076 (fax)
- DKL5@psu.edu

(Customer Satisfaction or your money back!)