Statistical Machine Learning

(An Overview)

Ying-Chao Hung

Department of Statistics
National Chengchi University
Email: hungy@nccu.edu.tw

Some slides adapted from George Michailidis,
Informatics Institute, University of Florida

Logistics

- Course Website: http://wm3.nccu.edu.tw
  Note that the slides, computer code, data, and other course information will be available on the website.

- Prerequisite: Basic knowledge in linear algebra (matrix theory), mathematical statistics (such as MLE, bootstrap, Bayesian and regression analysis) and probability (random variables, distributions).

- There is a Lab session after a topic is introduced.
  All the methods introduced in class will be executed by the software package R.

Course Material

No textbooks required.
All course material will be available online.

Some reference books:

- James, G., Witten, D. Hastie, T. and Tibshirani, R. An Introduction to Statistical Learning. Springer, 2014. Available online at: http://www-bcf.usc.edu/gareth/ISL/

- Murphy, K. Machine Learning: a Probabilistic Perspective. MIT Press, 2012.

- Hastie, T., Tibshirani, R. and Friedman, J. The Elementsof Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition. Springer, 2009. Available online at: http://statweb.stanford.edu/tibs/ElemStatLearn/

What is Statistical Machine Learning?

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.

Machine learning focuses on the development of computer programs (or algorithms) that can learn from data and make predictions.

More on Statistical Machine Learning

• Statistical machine learning merges statistics with the computational sciences -- computer science, systems science and optimization.

• Much of the agenda is driven by applied problems in science and technology, where data streams are increasingly large-scale, dynamical and heterogeneous, and where mathematical and algorithmic creativity are required to bring statistical methodology to bear.

• Fields such as bioinformatics, artificial intelligence, signal processing, communications, networking, information management, finance, game theory and control theory are all being heavily influenced by developments in statistical machine learning.

(Adapted from University of California, Berkeley)

Learning From Data

Fact: The amount of data and information collected and retained by organizations and businesses is constantly increasing, due to advances in data collection, computerization of transactions and breakthroughs in storage technology.

Consequence: Statistical problems have exploded both in size and complexity.

Objective: The data analyst's job is to make sense of such vast amounts of data. More specifically, identify patterns and trends, and uncover "interesting" relationships among the variables and/or the observations.

Technology Helps

- Faster computers (offer possibility for more flexible – hence
  powerful - and more assumptions-free techniques)

- Excellent graphic capabilities (explosion in visualization methodology)

A word of caution: Faster computers does not mean that we are home free. Several problems are inherently computationally intractable (EXP-HARD or NP-complete) and algorithms that currently solve them are suboptimal.

Let us look next at some examples:

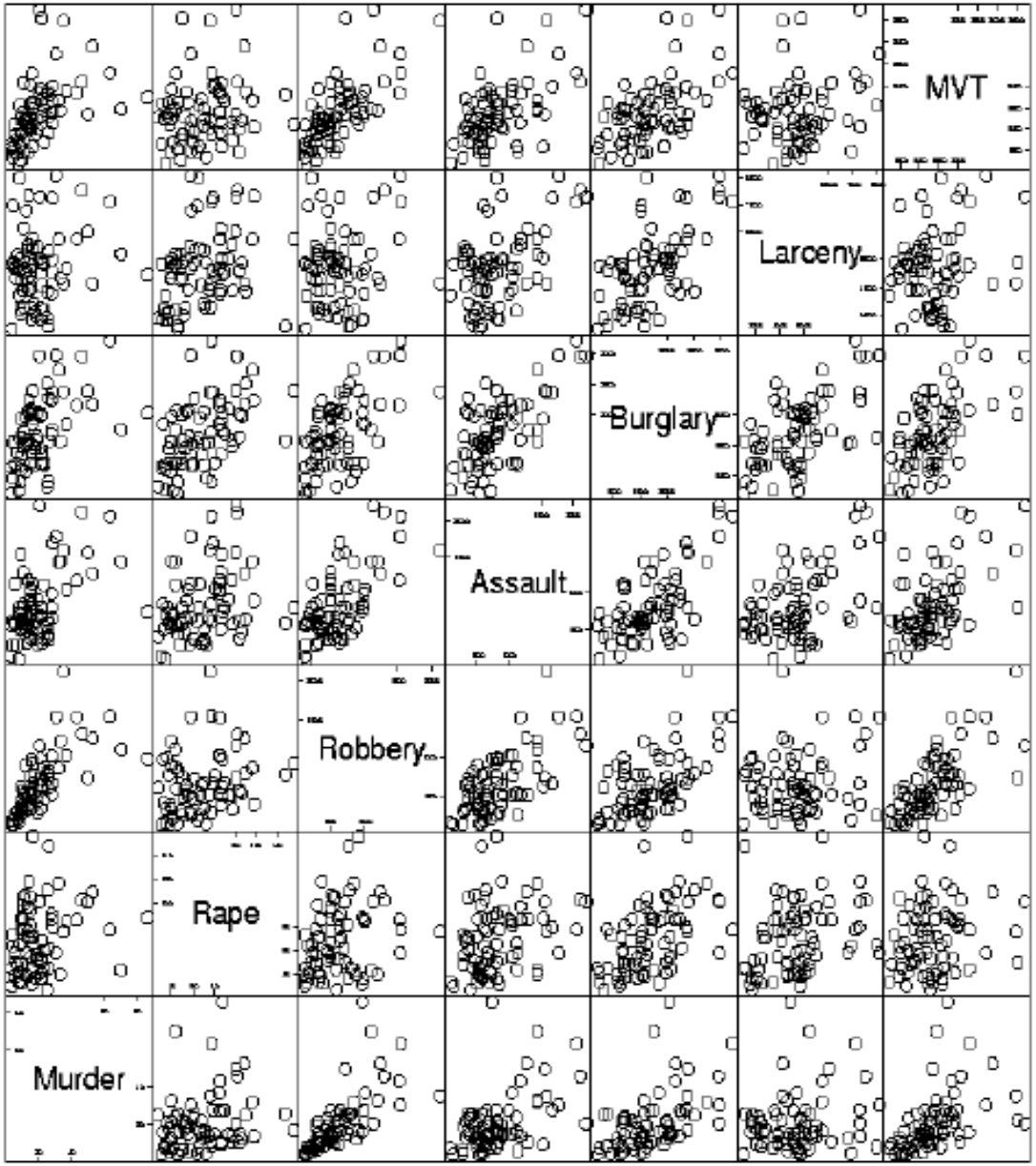
U.S. Cities Crime Data

The data give crime rates per 100,000 people for the largest U.S. cities in 1994. The variables are:

1. Murder
2. Rape
3. Robbery
4. Assault
5. Burglary
6. Larceny
7. Motor Vehicle Thefts

The goal is to identify low/high crime cities.
This is an example of *summarizing* multivariate data.
Let us take a look at the data.

Figure 1: Scatterplot matrix



Comments

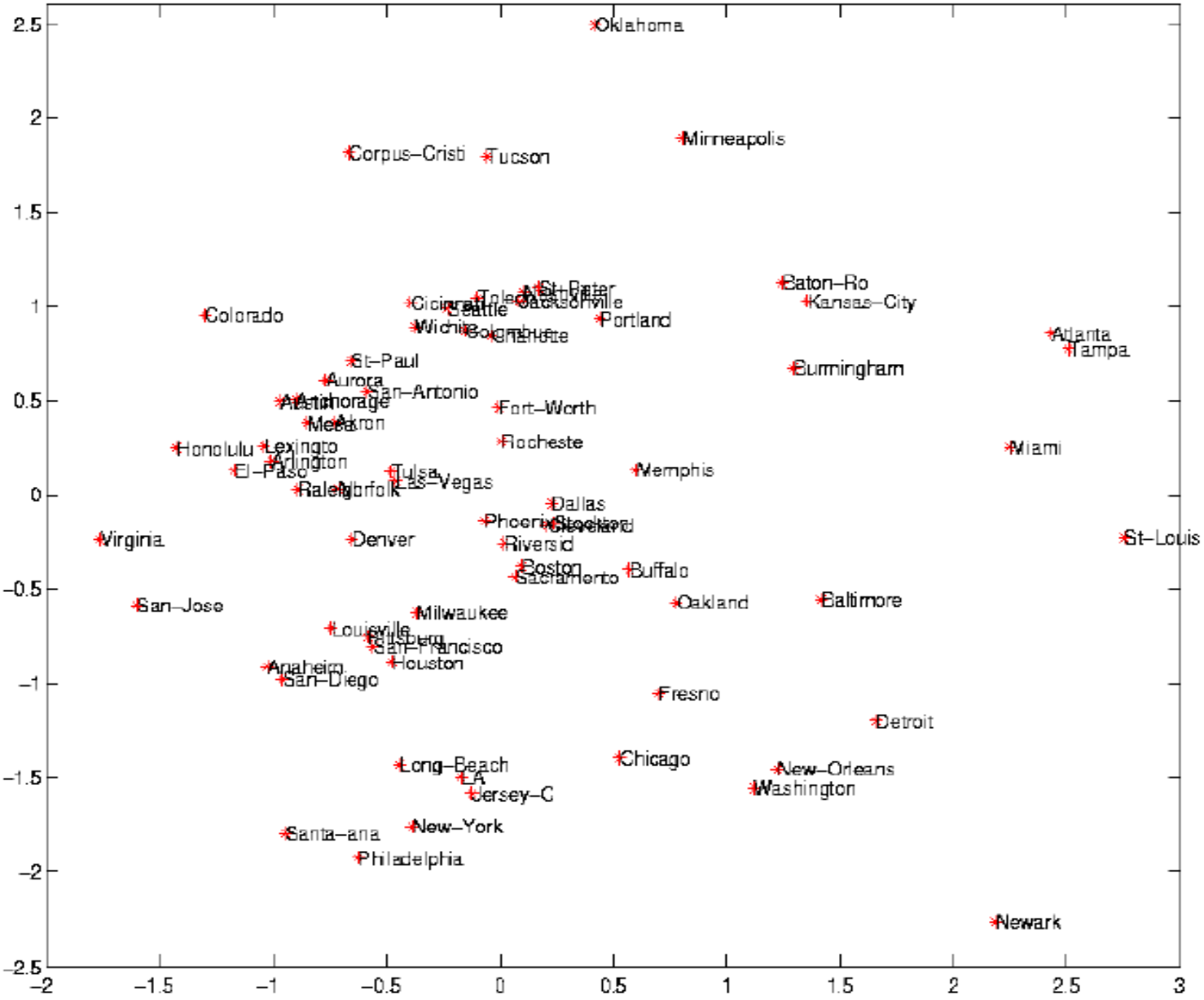
Interesting but NOT particularly informative!!

Q: What if instead I had a procedure that combined the crime variables and produced an “index” for how “safe” a city is.

Remark: The quality of the index variable heavily depends on the procedure used to combine the original crime measurements. A bad procedure will result in an non-informative representation, while a good one will achieve our goal.

A picture of such an index variable is given next (ignore the y-axis) with low overall crime cities located to the left.

Figure 2: Plot of the crime "index" (x-axis)



Sports

In the book/movie “Moneyball” (魔球): The general manager Billy Beane of Oakland Athletics led the baseball team to have the most wins in 2002 MLB with a low salary.

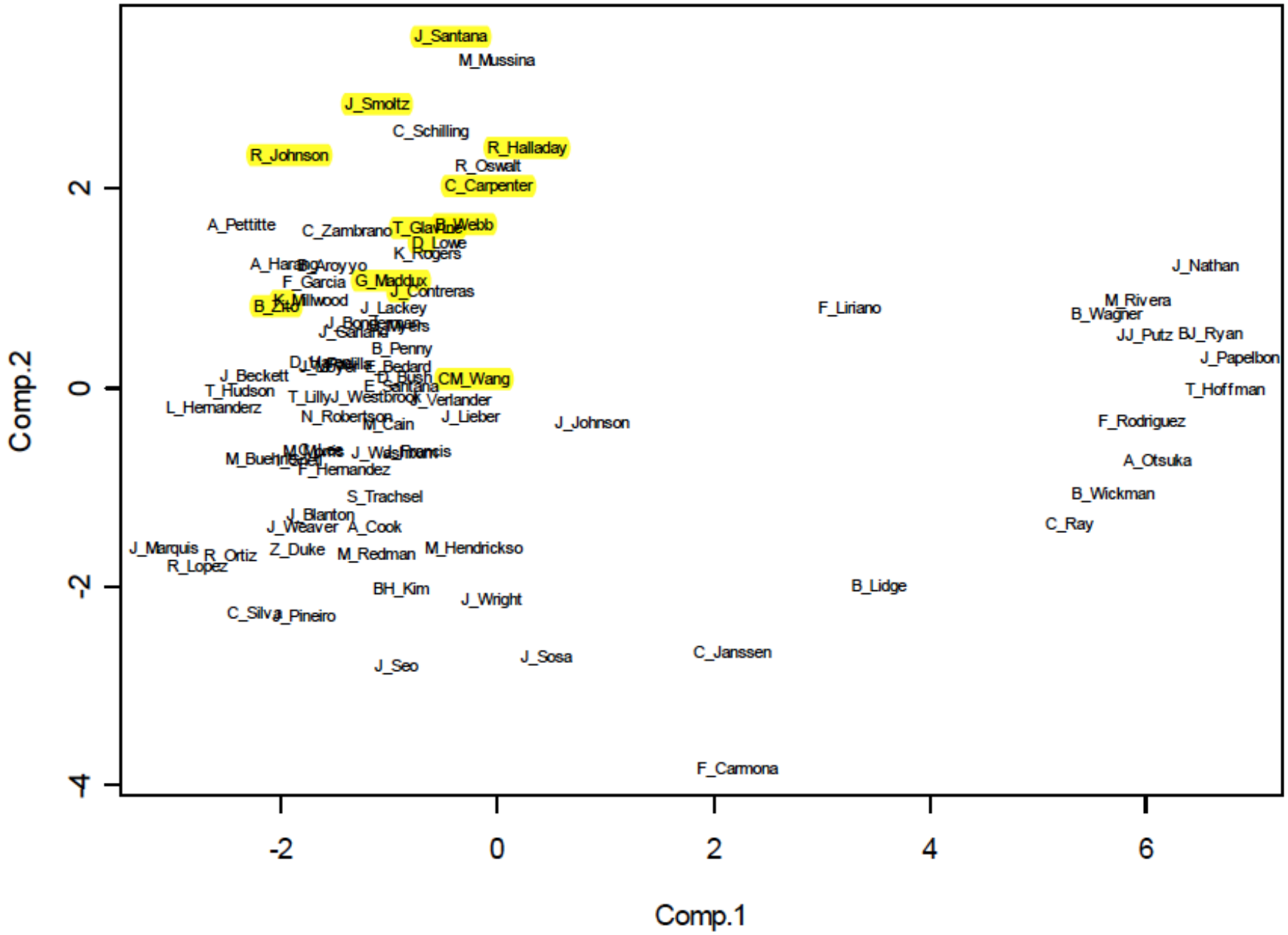
➔A successful example of Big Data Analysis by using MLB players’ statistics!!

The following data are the records of pitchers from 2006 Major League Baseball, which are available at ESPN.com.

2006 MLB Pitcher's Statistics

	Level	Game	IP	Hit	Run	HR	BB	SO	Win	Loss	SV	WHIP
J_Santana	1	34	233.2	186	79	24	47	245	19	6	0	1
CM_Wang	1	34	218	233	92	12	52	76	19	6	1	1.31
J_Garland	1	33	211.1	247	112	26	41	112	18	7	0	1.36
F_Garcia	1	33	216.1	228	116	32	48	135	17	9	0	1.28
R_Johnson	1	33	205	194	125	28	60	172	17	11	0	1.24
K_Rogers	1	34	204	195	97	23	62	99	17	8	0	1.26
J_Verlander	1	30	186	187	78	21	60	124	17	9	0	1.33
J_Beckett	2	33	204.2	191	120	36	74	158	16	11	0	1.29
J_Blanton	2	32	194.1	241	111	17	58	107	16	12	0	1.54
R_Halladay	1	32	220	208	82	19	34	132	16	5	0	1.1
A_Harang	2	36	234.1	242	109	28	56	216	16	11	0	1.27
D_Lowe	1	35	218	221	97	14	55	123	16	8	0	1.27
K_Millwood	1	34	215	228	114	23	53	157	16	12	0	1.31
B_Penny	1	34	189	206	94	19	54	148	16	9	0	1.38
E_Santana	2	33	204	181	106	21	70	141	16	8	0	1.23
J_Smoltz	1	35	232	221	93	23	55	211	15	9	0	1.19
B_Webb	1	33	235	216	91	15	50	178	16	7	0	1.13
C_Zambrano	1	33	214	162	91	20	115	210	16	7	0	1.29
B_Zito	2	34	221	211	99	27	99	151	16	10	0	1.4
E_Bedard	2	33	196.1	196	92	16	69	171	15	11	0	1.35
C_Carpenter	1	32	221.2	194	81	21	43	184	15	8	0	1.07
T_Glavine	2	32	198	202	94	22	62	131	15	7	0	1.33
T_Lilly	2	32	181.2	179	98	28	81	160	15	13	0	1.43
G_Maddux	2	34	210	219	109	20	37	117	15	14	0	1.22
M_Mussina	1	32	197.1	184	88	22	35	172	15	7	0	1.11
R_Oswalt	2	33	220.2	220	76	18	38	166	15	8	0	1.17
V_Padilla	2	33	200	206	108	21	70	156	15	10	0	1.38
C_Schilling	2	31	204	220	90	28	28	183	15	7	0	1.22

A 2D Projection Based on PCA



Sleeping Bags

This dataset contains information on price, fiber and quality for 21 sleeping bags (data taken from Prediger (1997)).

Remark: Notice that all the variables are categorical.

Data

	cheap	not expensive	expensive	down fibers	synthetic fibers	good	acceptable	bad
Sleeping Bag	**Price**			**Fiber**		**Quality**		
One Kilo Bag	1	0	0	0	1	1	0	0
Sund	1	0	0	0	1	0	0	1
Kompakt Basic	1	0	0	0	1	1	0	0
Finmark Tour	1	0	0	0	1	0	0	1
Interlight Lyx	1	0	0	0	1	0	0	1
Kompakt	0	1	0	0	1	0	1	0
Touch the Cloud	0	1	0	0	1	0	1	0
Cat's Meow	0	1	0	0	1	1	0	0
Igloo Super	0	1	0	0	1	0	0	1
Donna	0	1	0	0	1	0	1	0
Tyin	0	1	0	0	1	0	1	0
Travellers Dream	0	1	0	1	0	1	0	0
Yeti Light	0	1	0	1	0	1	0	0
Climber	0	1	0	1	0	0	1	0
Viking	0	1	0	1	0	1	0	0
Eiger	0	0	1	1	0	0	1	0
Climber light	0	1	0	1	0	1	0	0
Cobra	0	0	1	1	0	1	0	0
Cobra Comfort	0	1	0	1	0	0	1	0
Foxfire	0	0	1	1	0	1	0	0
Mont Blanc	0	0	1	1	0	1	0	0

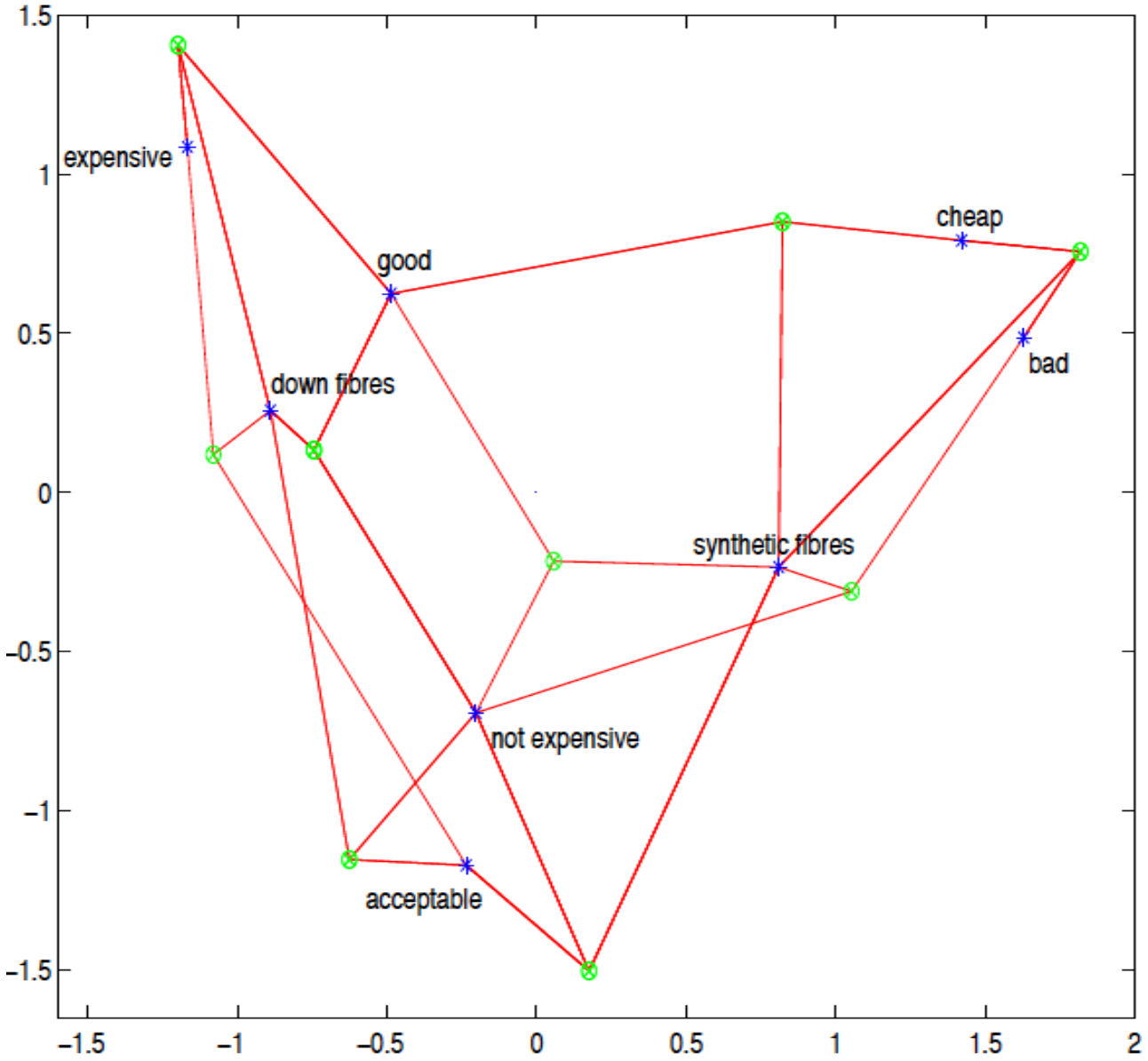
Sleeping Bags

The goal is to learn something about the choices of sleeping bags available to the consumer; in more general terms, understand the structure of this data set.

Remark: The categorical nature of the data renders a scatterplot matrix useless.

However, a ”good” picture would be informative, since *a picture is worth a lot of numbers, especially when these numbers are just zeros and ones*.

Figure 3: A good picture of the sleeping bag data (green points represent the sleeping bags)



Comments based on the Picture

A careful examination of the picture shows that:

- there are good, expensive sleeping bags filled with down fibers
- there are also cheap, bad quality sleeping bags filled with synthetic fibers
- there are also some intermediate sleeping bags in terms of quality and price filled either with down or synthetic fibers.
- there are some expensive ones of acceptable quality and some cheap ones of good quality
- however, there are no bad expensive sleeping bags

Handwritten Letter and Digit Recognition

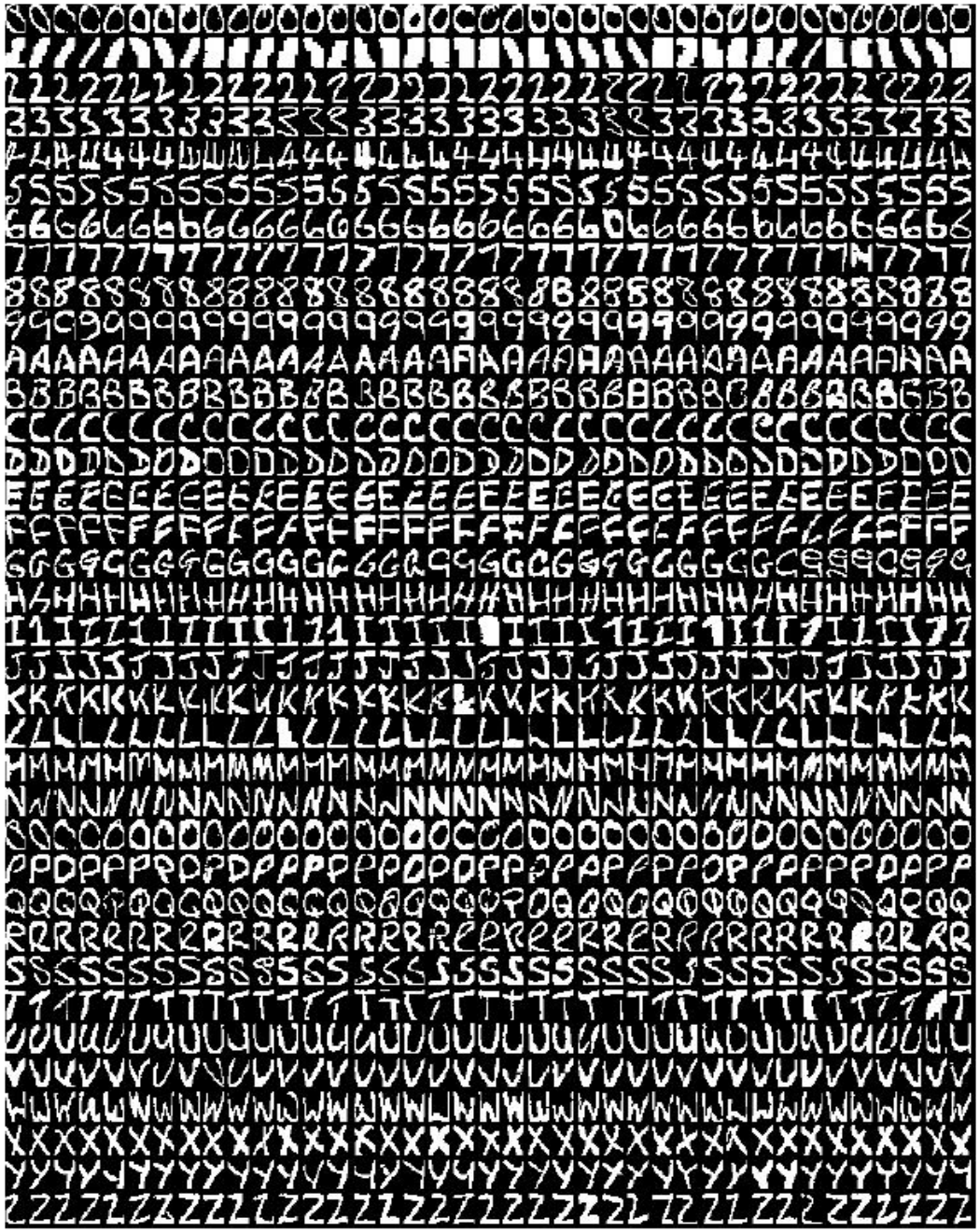
The data from this example correspond to handwritten letters and digits. Each image is a segment, isolating a single letter/digit. The images are 20x16 eight-bit grayscale maps, with each pixel ranging in intensity from 0 to 255. Some sample images are shown in the next figure.

The task is to *predict*, from the 20x16 matrix of pixel intensities, the identity of each image {*A*, *B*,…, *Z*, 0, 1,…, 9}.

If the resulting *algorithm* exhibits a high degree of accuracy, then it could be used as part of an automated hand-writing recognition system.

This is an example of a *supervised learning* problem.

Figure 4: Examples of handwritten digits and letters



Handwritten Letter and Digit Recognition

In this type of problem, we have a categorical outcome measurement (the identity of the image), that we wish to predict based on a set of *features*/variables (the 320 pixel intensities).

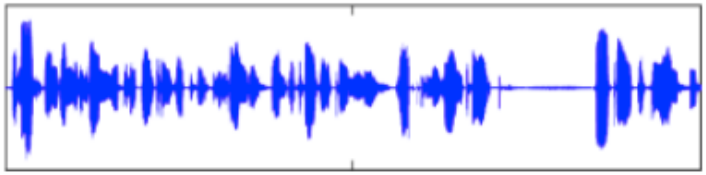
We also have a *training dataset* in which we observe both the outcome and feature measurements for a set of objects. Based on such data we build a model or in machine learning parlance a *learner*.

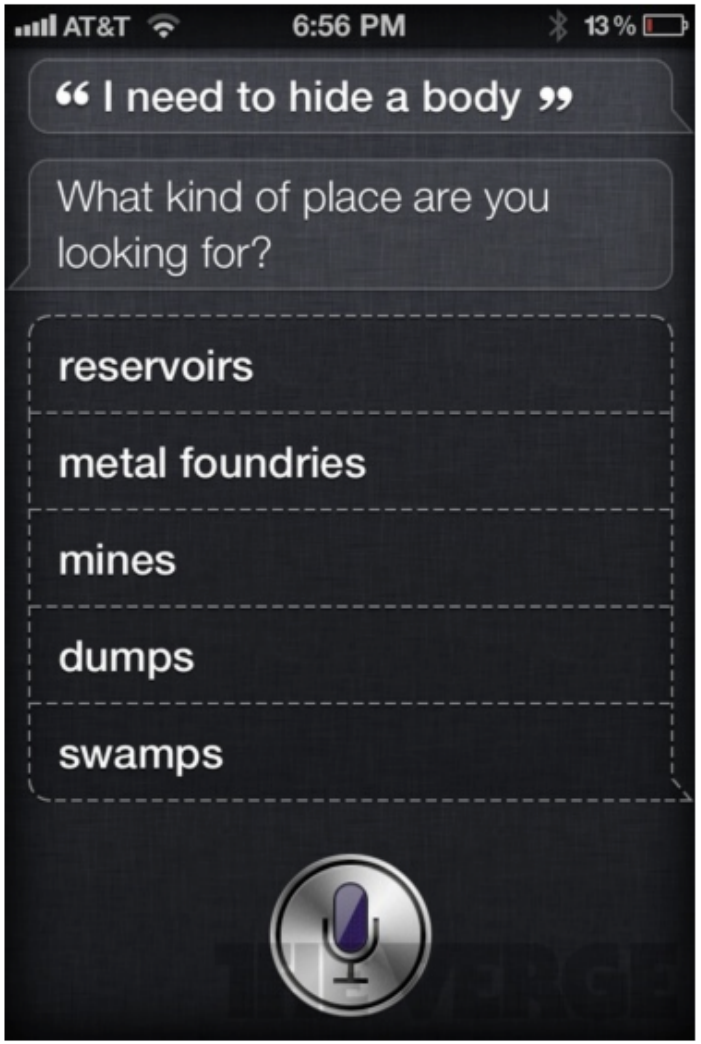
The goal is to build a good learner that accurately predicts the identities not only of the images in the training dataset but also of new never seen before letter and digit images.

Weather Prediction

As IBM points out “An annual economic impact of nearly **5×10^8 USD** in the US alone”.
(The Weather Company links to IBM clouds and Amazon Web Services to obtain the online data information)

Speech Recognition





Face Recognition (e.g. SVD)

Example training images
for each orientation

DNA Expression Data

DNA is the basic material that makes up human chromosomes. DNA microarrays and genechips measure the expression of a gene in a cell by measuring the amount of mRNA present for that gene. Both are considered to be breakthrough technologies, facilitating the quantitative study of thousands of genes simultaneously from a single sample of cells.

Here is a tiny sample of DNA expression data (3 genes (variables) and 5 samples (observations)).

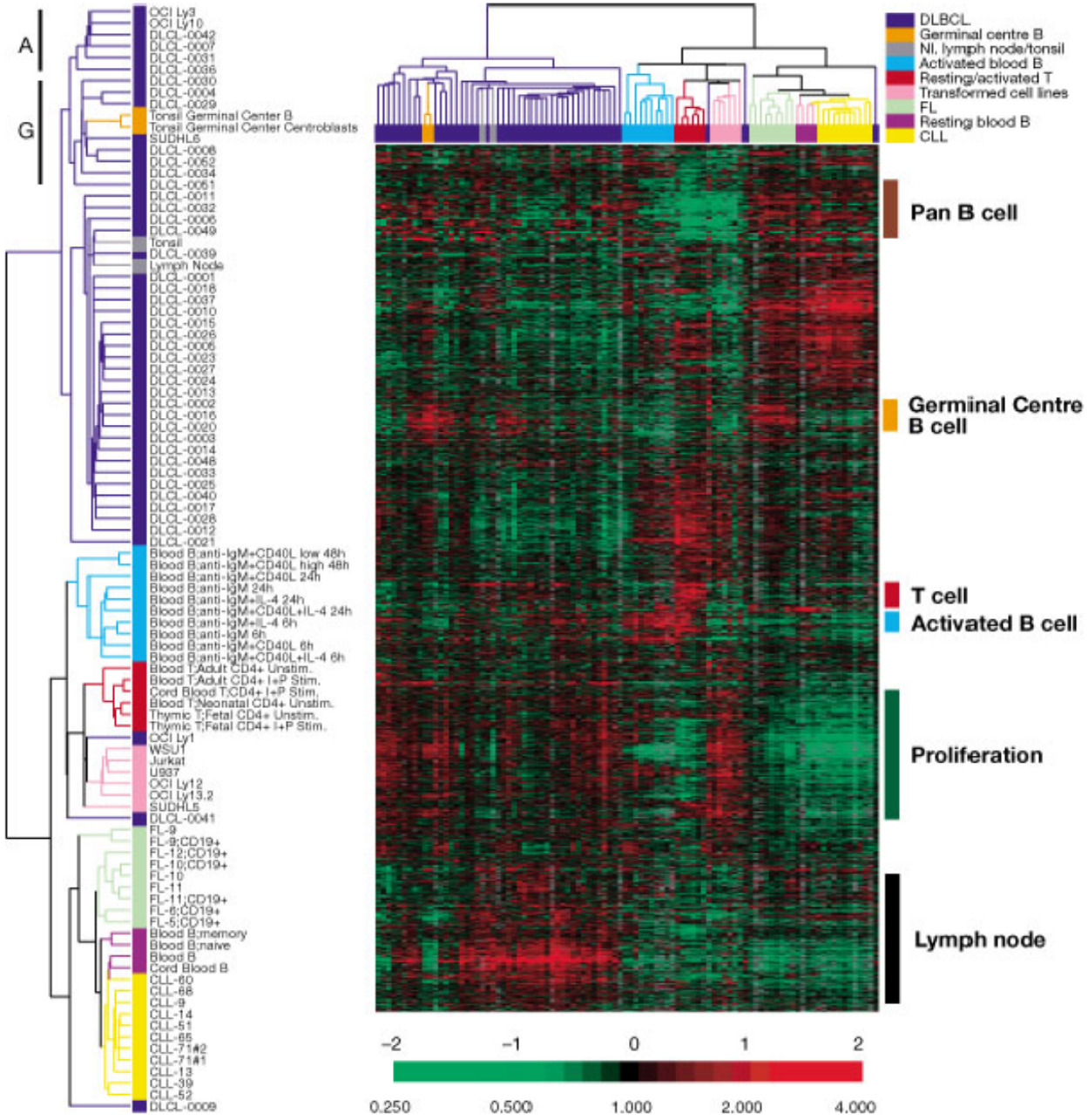
21652	3.2025	1.6547	3.2779	1.0060
25725	0.0681	0.0710	0.1160	0.1906
22260	0.1243	0.0520	0.1014	0.1035

DNA Expression Data

The complete dataset contains approximately 7000 genes (rows) and around 100 samples (columns), where the samples correspond to different cancer tumors.

In the next figure the data are displayed as a heat map, ranging from green (negative values) to red (positive values).

Figure 5: Heat Map of DNA Microarray Data



DNA Expression Data

The challenge with such data is to understand how the genes and samples are organized. Typical questions include the following (Hastie et al. (2001)):

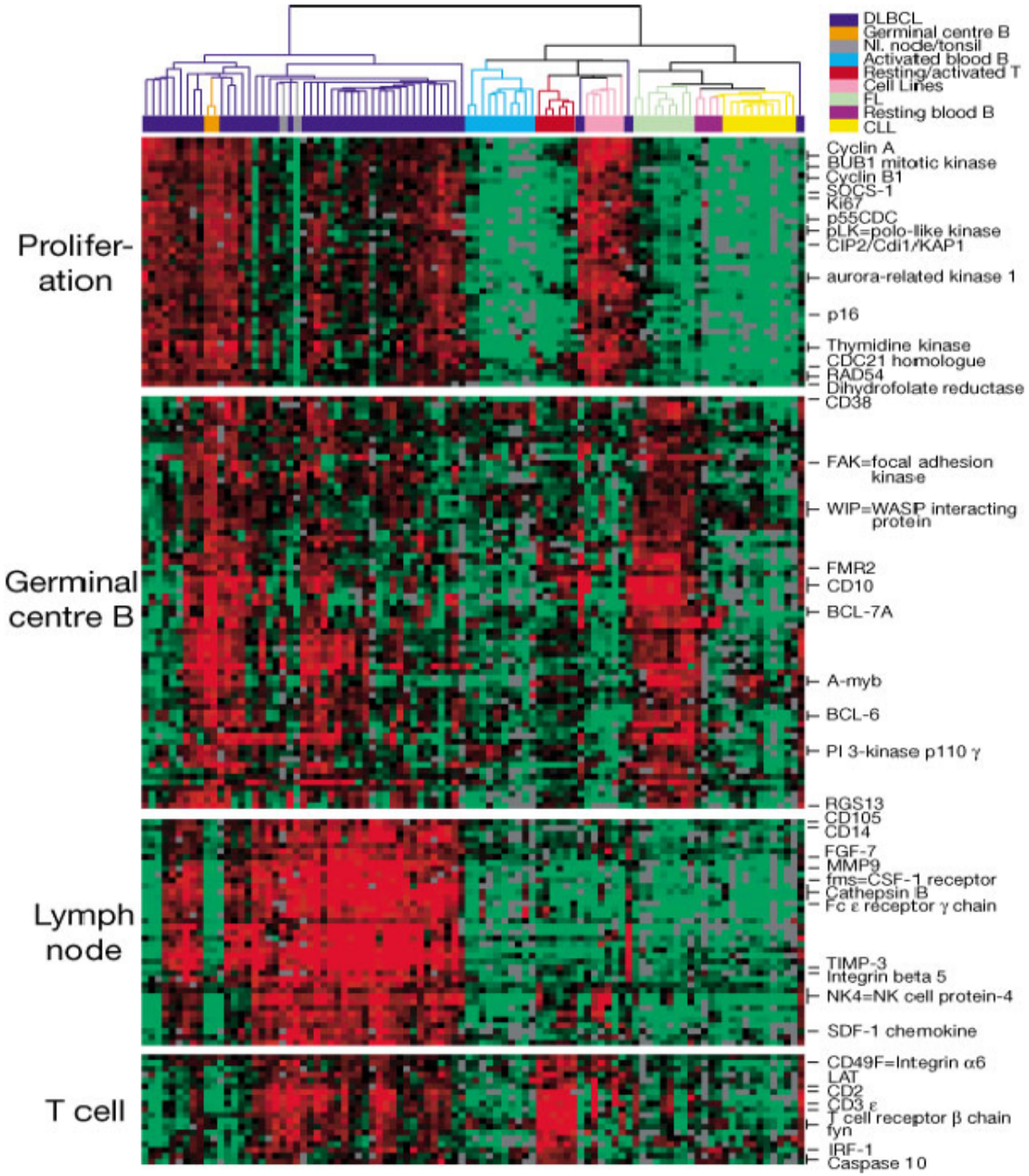
- which samples are most similar to each other, in terms of their expression profiles across genes
- which genes are most similar to each other, in terms of their expression profiles across samples
- do “interesting” patterns exist between subsets of genes and samples (e.g. very high/low expression levels)

Unsupervised Learning Problem

These questions can best be viewed as an *unsupervised learning* problem, where the goal is to *discover* new classes/groups in the data and organize the databases accordingly.

The next picture shows such a grouping for the tumor samples.

Figure 6: Grouping of B-cell lymphomas samples



Lung Tumors Correlation Matrix

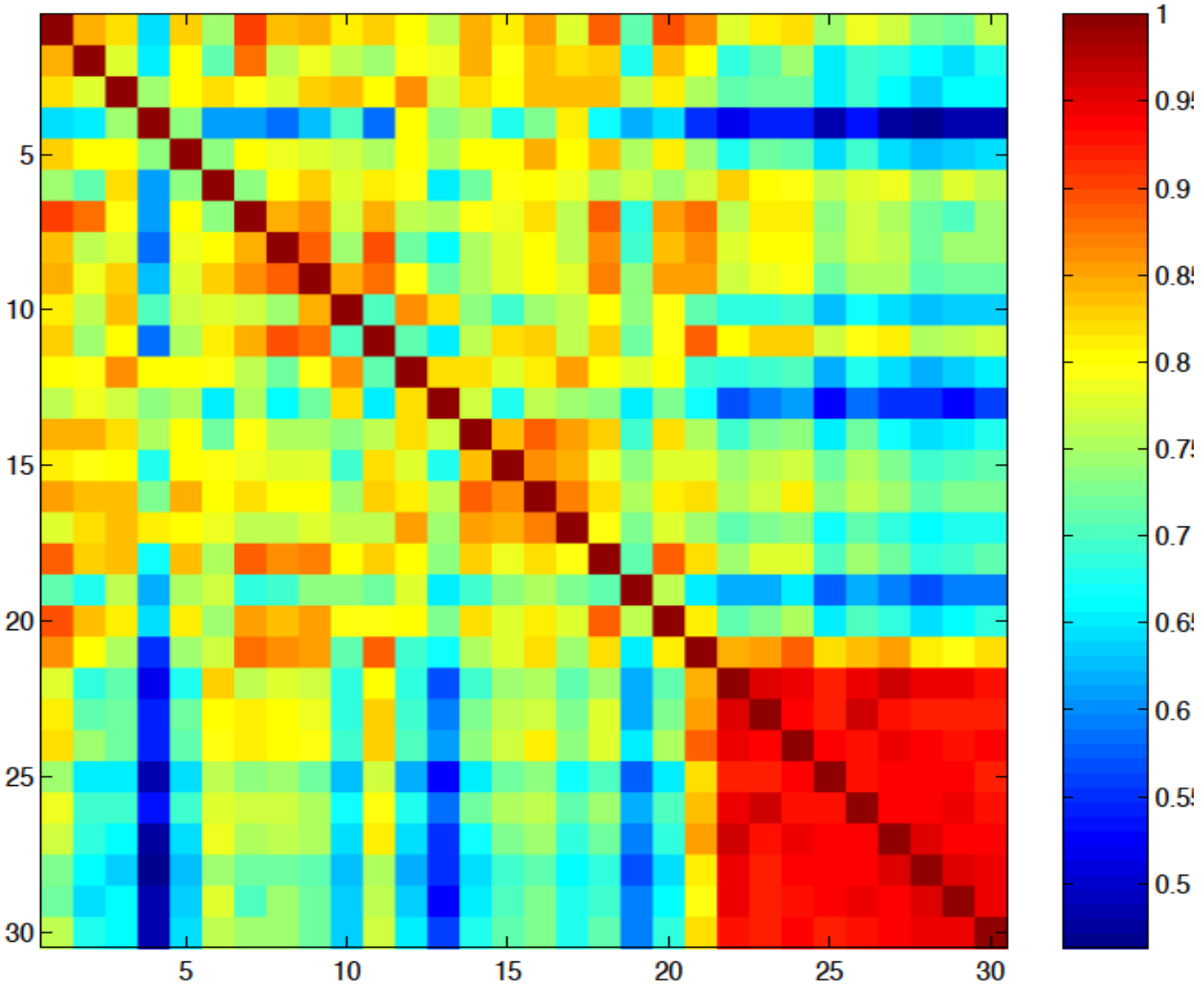
The data is a correlation matrix of 30 lung tumor samples. These correlations have been calculated by using DNA expression measurements of 7000+ genes.

How are these lung tumors related to each other?

Examining all the entries of the correlation matrix is not a good way of doing data analysis. A better way is to try to visualize this matrix. A heat map proves useful to a certain extent (however, it does not work well with larger correlation matrices).

This is a problem of visualizing complex data structures.

Figure 7: Heat Map of Correlation Matrix

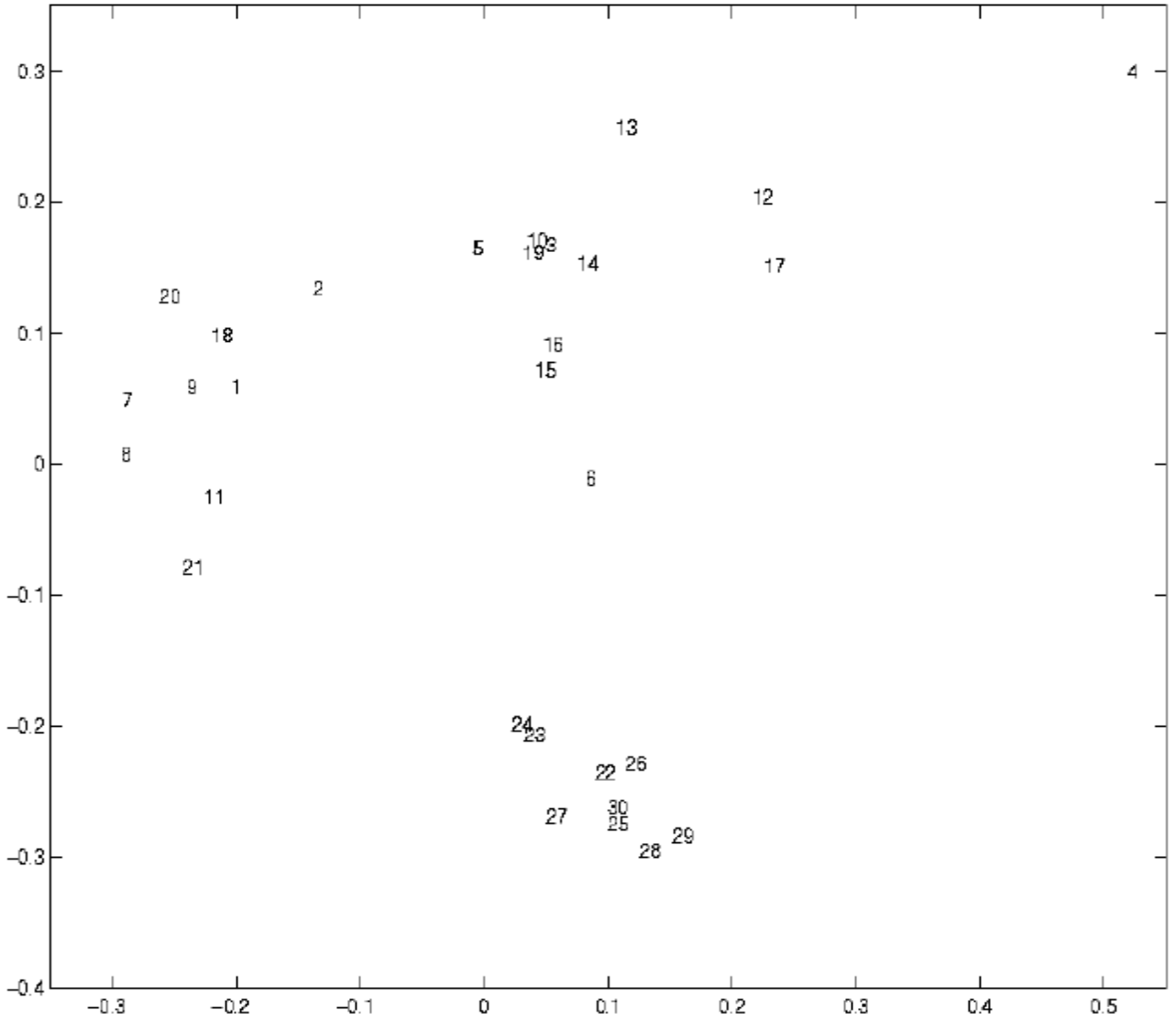


Distance-Based Approaches

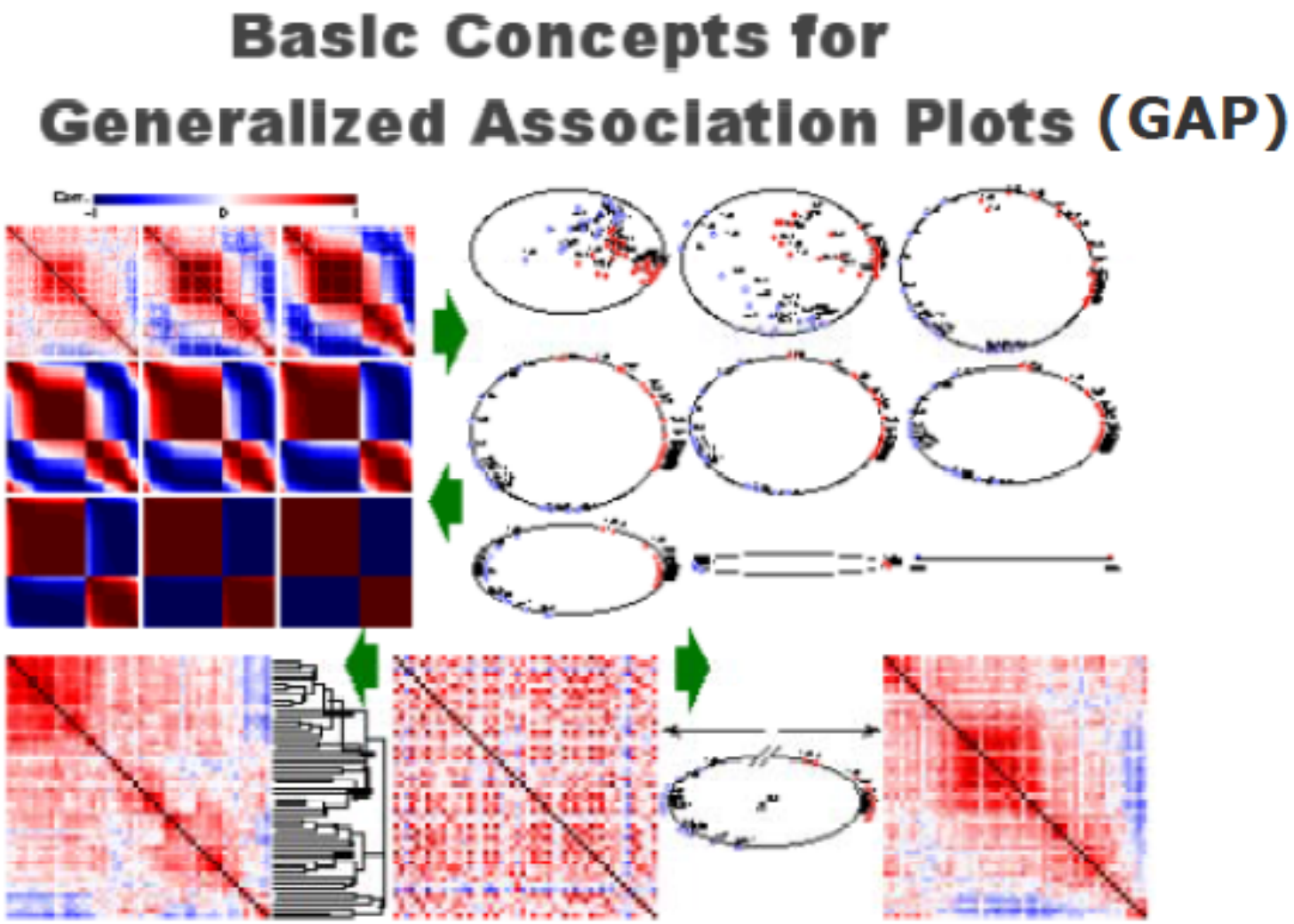
A better representation, which exhibits strong grouping patterns for some of the samples is given next.

In the following picture, small distances imply high correlations and large distances low ones.
The magnitude of the correlation is proportional to the Euclidean distance.

Figure 8: A Distance-based Map of Correlation Matrix

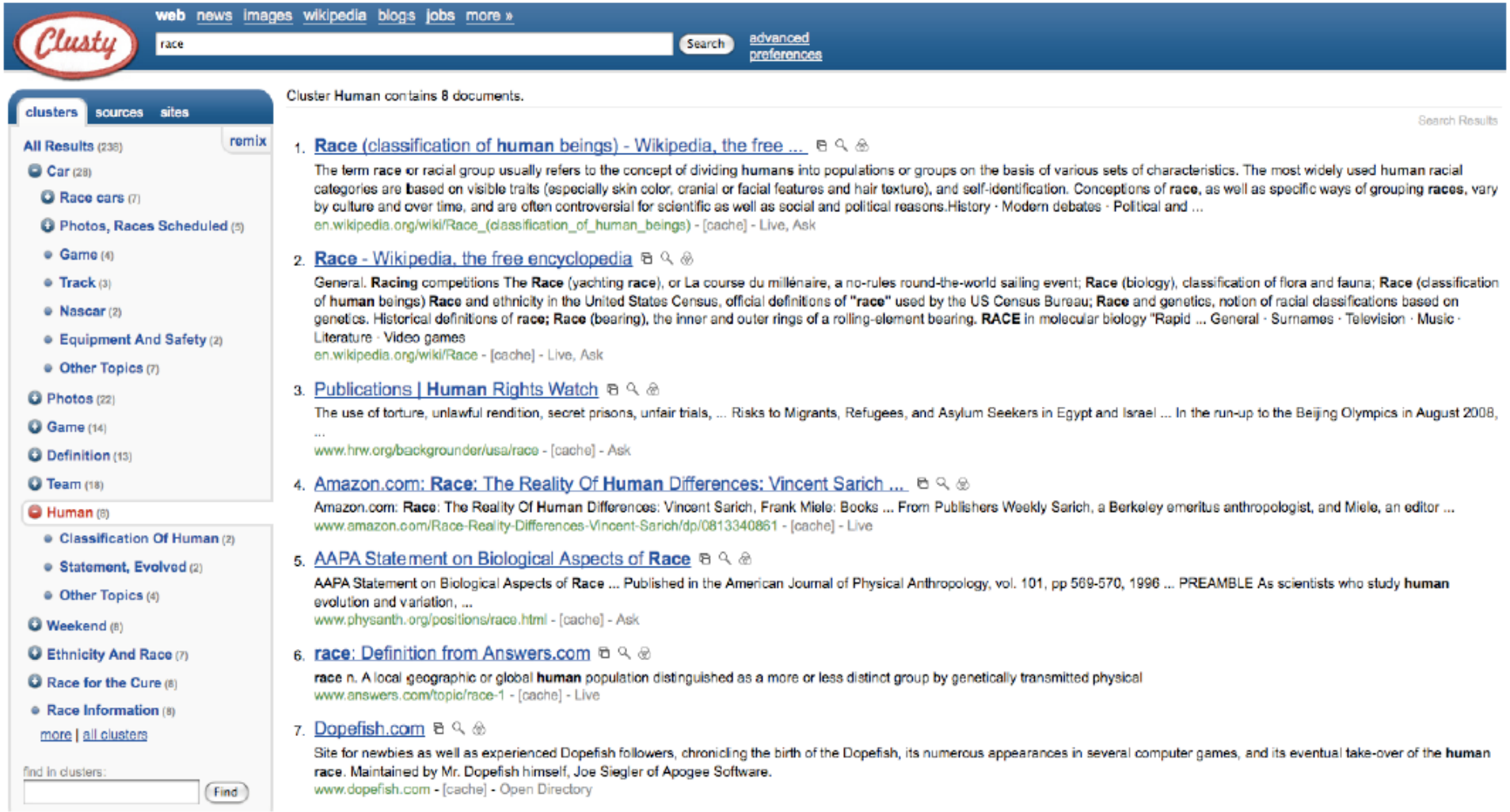


Another Presentation of Object Associations

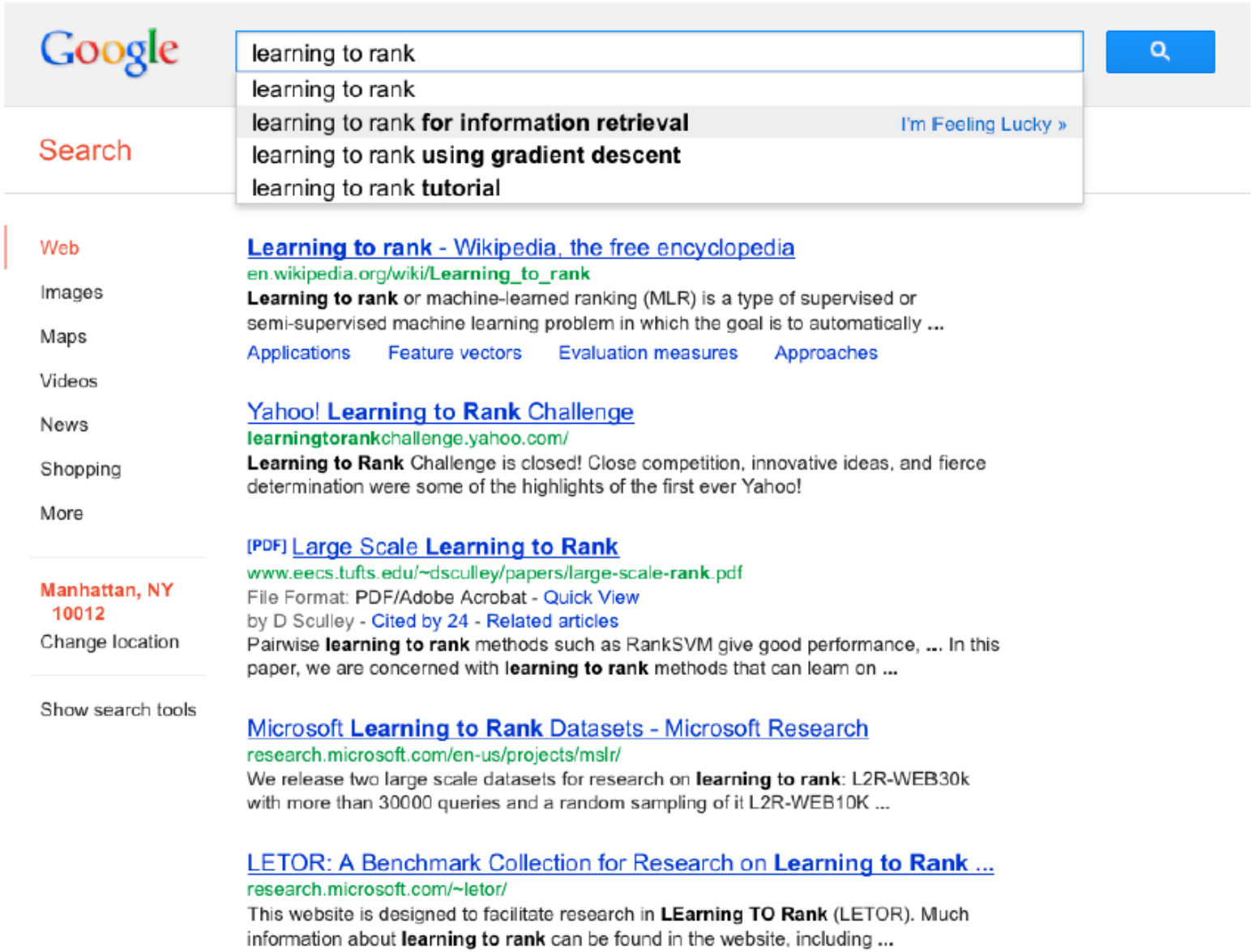


By Chun-houh Chen, Institute of Statistical Science, Academia Sinica, Taiwan.

Web Search (by Clustering)



Web Search (by Ranking)



Recommendation System (Ranking Matching)

Growth of Machine Learning

Machine Learning has been widely used in:

- Speech recognition, natural language processing
- Consumer behavior (online shopping) analysis
- Computer vision
- Medical outcome analysis
- Robot control
- Computational biology
- Sensor networks
- Social network analysis, …

The trend is accelerating:

- Big data
- Improved algorithms
- Faster computers
- Good open-source software
- Better statistical modeling techniques

Some Important Issues

- Underlying probability models – classical inference
- Computational inference (e.g. bootstrapping, permutation tests)
- Algorithm consideration
- Visualization
- The role of multivariate normal distribution

There is no true interpretation of anything; interpretation is a vehicle in the service of human comprehension. The value of interpretation is in enabling others to fruitfully think about an idea.

Andreas Buja

(taken from Hastie et al. (2001))

Course Roadmap

1) Grand Tour
2) Introduction to a Data Visualization System "GGobi" and R
3) Review of Matrix Algebra
4) Supervised Learning: Classification (Loss, Risk, Bayesian inference)
5) Supervised Learning: Classification (Linear Classifier, Quadratic Classifier, Logistic Regression, Nearest Neighbor Methods)
6) Supervised Learning: Classification (Support Vector Machine)
7) Supervised Learning: Classification (Decision Trees)
8) Ensemble Methods (Boosting and Bagging)
9) Midterm Exam Week
10) Unsupervised Learning: Clustering (Hierarchical Trees)
11) Unsupervised Learning: Clustering (K-means, Self-Organizing Maps)
12) Dimension Reduction: Principal Component Analysis
13) Dimension Reduction: Multidimensional Scaling
14) Dimension Reduction: Correspondence Analysis
15) Dimension Reduction: Canonical Correlation Analysis
16) Dimension Reduction: Factor Analysis
17) Information Theory and Independent Component Analysis
18) Final Exam Week