— *extract from* ***Flatland: A Romance of Many Dimensions****, by Edwin A. Abbott, 1884.*

Stranger. *What do you know of space. Define space.*

I. *Space, my Lord, is height and breadth indefinitely prolonged.*

Stranger. *Exactly: you see you do not even know what Space is. You think of it as Two Dimensions only; but I have come to announce to you a Third – height, breadth and length.*

I. *Your Lordship is pleased to be merry. We also speak of length and height, or breadth and thickness, thus denoting Two Dimensions by four names.*

Stranger. *But I mean not only three names, but Three Dimensions… I am in no jesting humour. I tell you I come from Space, or, since you will not understand what Space means, from the Land of Three Dimensions whence I but lately looked down upon your Plane which you call Space forsooth. From that position of advantage I discerned all that you speak of as solid (by which you mean "enclosed on four sides"), your houses, your churches, your very chests and safes, yes even your insides and stomachs, all lying open and exposed to my view.*

ITI 2009 Cavtat, Croatia 24 June 2009

DYNAMIC GRAPHICS FOR RESEARCH AND TEACHING, WITH APPLICATIONS IN THE LIFE SCIENCES

Michael Greenacre
Universitat Pompeu Fabra,
Barcelona

www.econ.upf.edu/~michael

www.globalsong.net

The Millennium Song
Pjesma Mileniju

www.globalsong.net

translated into *Hrvatski* by
Jasna Horvat & Sande Katavic

Pjesma Mileniju

Sad je vrijeme promjena
Novom životu u susret
Bez osvrta unatrag
Oprostimo prošlosti.

Primi ruku
Haj'mo složno
U predivan svijet
U novi milenij.

The Millennium Song (original)

Michael Greenacre

Time has come to change
Time to find new life
No more looking back
Let's forgive the past

Take my hand, friend,
We'll go on as one
To a different world
The Third Millennium

The Millennium Song

Srpski
(Tamara Djermanovic)

Песма Миленијума

Долази време,
Време за промене,
Не окрећи се,
Опрости прошло.

Дај ми руку,
Кренимо заједно,
Ка свету новом,
Миленијума.

Pesma Milenijuma

Dolazi vreme,
Vreme za promene,
Ne okreci se,
Oprosti proslo.

Daj mi ruku,
Krenimo zajedno,
Ka svetu novom,
Milenijuma.

Slovensko
(Anuška Ferligoj, Vladimir Batagelj)

Pesem tisočletja

Zdaj prišel je čas,
čas za spremembe,
ne oziraj se,
včeraj za nami je.

Daj mi roko,
stopimo skupaj
v nov in boljši svet,
tretje tisočletje.

Click here to listen to song sung by *Nuška Drašček*, recorded at Radio Ljubljana

Summary

Our interest is in the use of motion in scientific graphics.

The status quo:

- Scientific publications are produced primarily for print media – their online editions are almost always the same as the print version. Even in exclusively online journals the figures are static, as if they were printed.
- Statistical graphics as a medium for communicating numerical data has changed very little in the applied sciences. Usually the standard graphical templates that are available to researchers in software packages are the ones that are used.

The potential:

- Are you taking sufficient advantage of the enormous resources such as increased computing power and increasing bandwidth?
- Online publishing allows new graphical elements to be introduced: color, motion, and sound.

The challenge:

- To demonstrate the added value of animating scientific graphics, especially in the context of high-dimensional (multivariate) data.
- To create tools to allow easy design and construction of dynamic graphical displays.

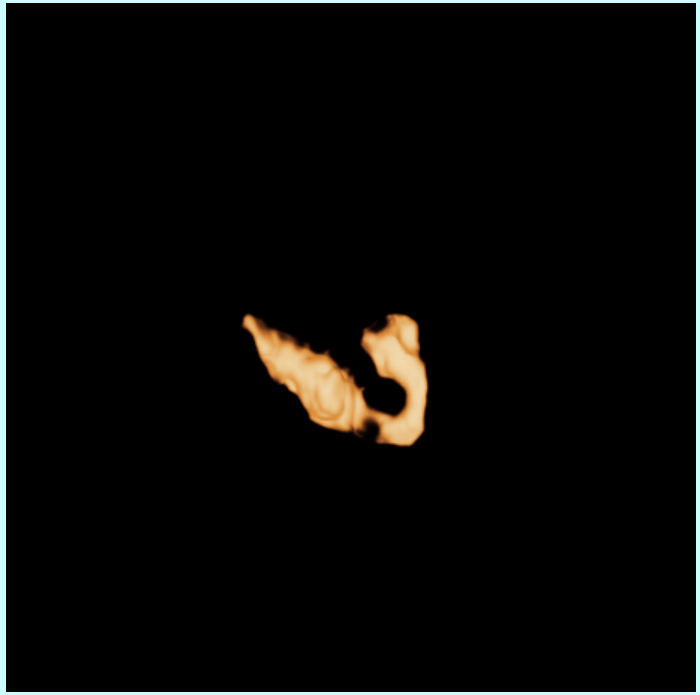
Use of motion in online articles

Here we are <u>not</u> talking about:

- **Showing three-dimensional objects**

For example, the *Journal of Ultrasound in Medecine* regularly publishes videos, such as the rotating embryo opposite.

Rotation of embryo



GIF file, 1.6mb; online version in AVI format, 221mb

- **Interactive graphics**

Several programs for interactive graphics: **`XGobi, GGobi, iPlot`**, but these remain tools of statisticians in the process of data analysis and publications about them only include screenshots of various moments in the interactive analysis.

Here we <u>are</u> talking about:

- **Dynamic data displays**

These show animated graphics for 5–10 seconds to (i) guide the eye in interpreting the data; (2) allow us to see a third dimension; and (iii) to tour through higher-dimensional spaces in the search for structure or predictive models.

Simple initial examples

Animating regular two-dimensional graphics

This is a typical printed page from a journal article, which includes a figure showing the evolution of glucose levels in females and males at different ages, taken from a retrospective epidemiological study of cardiac heart disease (CHD) sufferers in New Zealand.

Admission blood glucose predicts survival following all types of Troponin positive acute coronary syndrome – increased mortality in New Zealand Maori

Adrian R Scott[1],
Anthony Cheng[1]
Michael Greenacre[3]
Gerard Devlin[2]
[1]*Diabetes and* [2]*Cardiology departments, Waikato Hospital, Hamilton, New Zealand.*
[3]*Departments of Economics and Life Sciences, Pompeu Fabra University, Barcelona*

The aim of this retrospective study was to investigate the effect of admission blood glucose and ethnicity on long term as well as short term mortality in patients admitted with ACS – both STEMI and NSTEMI.

Study Design and Methods

Waikato Health Board serves approximately 9% of the total New Zealand population (312,918 in the 1996 Census), in a largely rural setting in North Island. Approximately 20% identify themselves as Maori. An estimated 10000 (New Zealand Ministry of Health data) have diabetes (predominantly Type 2 –T2DM) which is more common amongst Maori and Pacific Islanders. Demographic, admission and discharge data for patients with ICD10 codes for all types of transmural MI, subendocardial MI or unstable angina between Jan 1st 1999 to Dec 31st 2002 was collected using the hospital patient database (HOSPRO). In addition the hospital Pathology laboratory database was interrogated and all patients with raised Trop I (up to 2000) or Trop T (2000 onwards) were identified. The two databases were merged for each patient and the admission blood glucose, maximum Troponin, creatinine, and cholesterol identified. The date of death was determined from data sent every 3 months from the New Zealand Health Information Service up to 1st April 2004. The re-classification of ACS by cardiologists is not reflected in ICD10 codes, so the data have been analysed in the following way: any transmural infarct has been classified as STEMI (N=1091). Admissions coded as subendocardial myocardial infarction have been analysed both separately and in combination with unstable angina or angina, unspecified. NSTEMI therefore includes any of the three diagnoses (N=3317).
Data were analysed using life-tables analysis, taking into account data censoring, to give estimated survival proportions, as well as fitting the Cox proportional-hazards regression model to assess the different effects of the covariates on survival time after ACS (20)

Results

There were 4408 (2846 male) episodes of ACS with a corresponding elevated Troponin. Patient characteristics can be seen in **Table 1**. A graph of glucose levels against age is shown in **Figure 1**, using half-Gaussian smoothing of values less than the age ordinate. The increase in male glucose

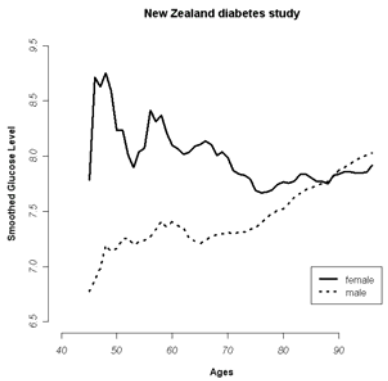
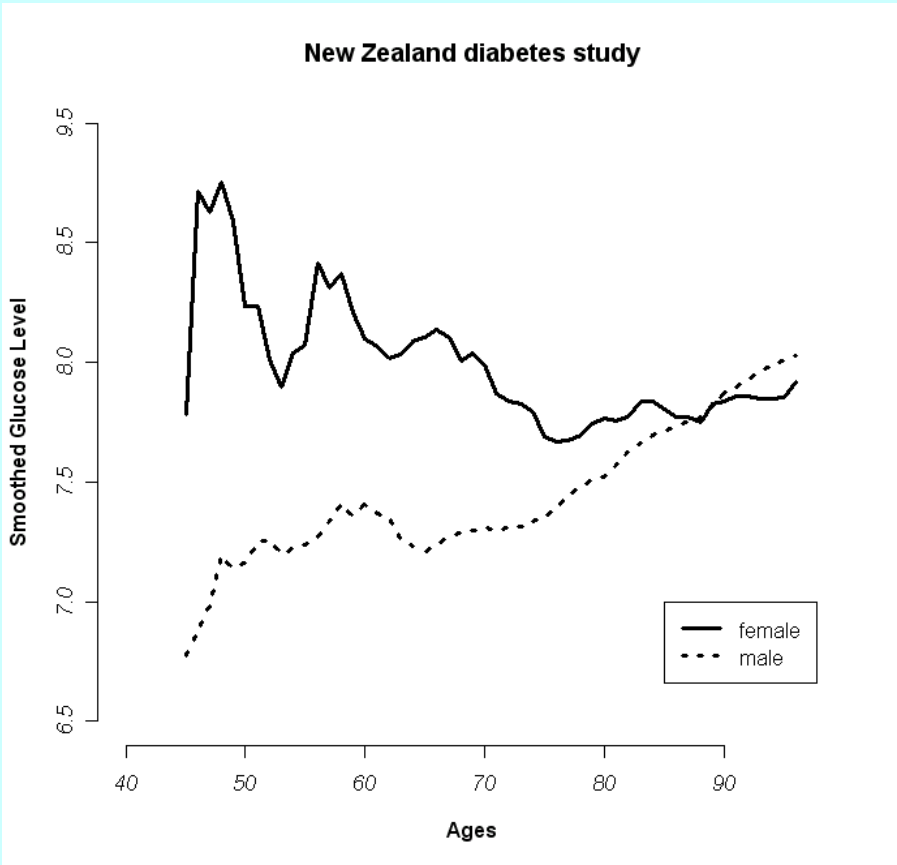
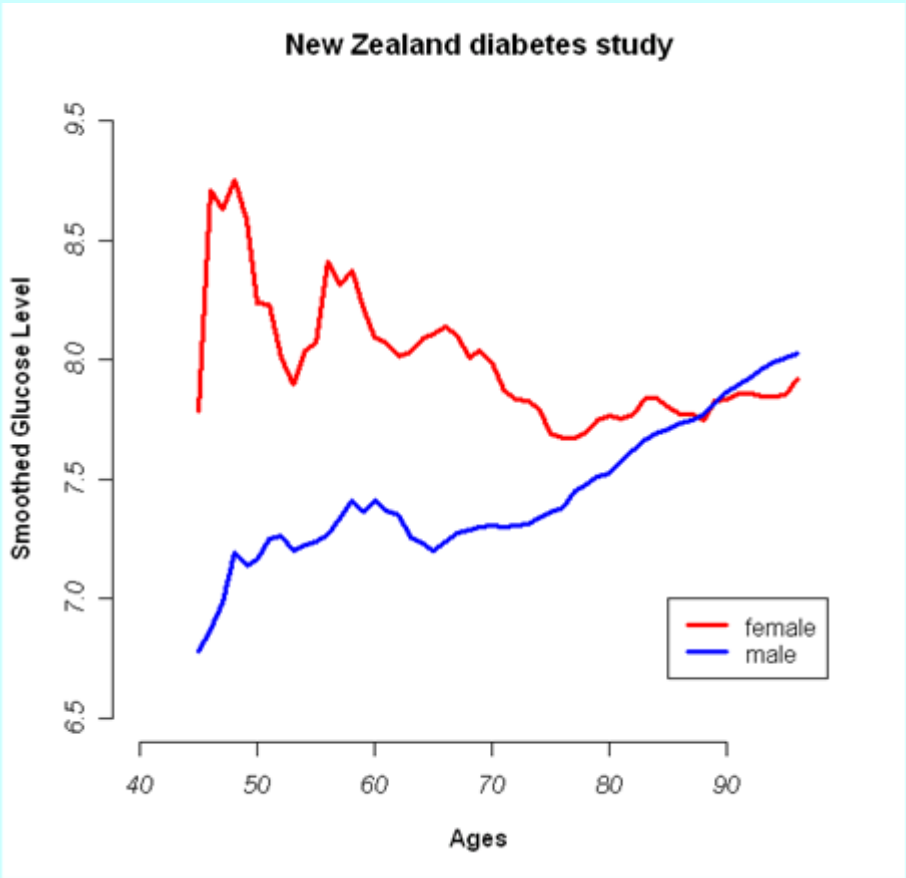


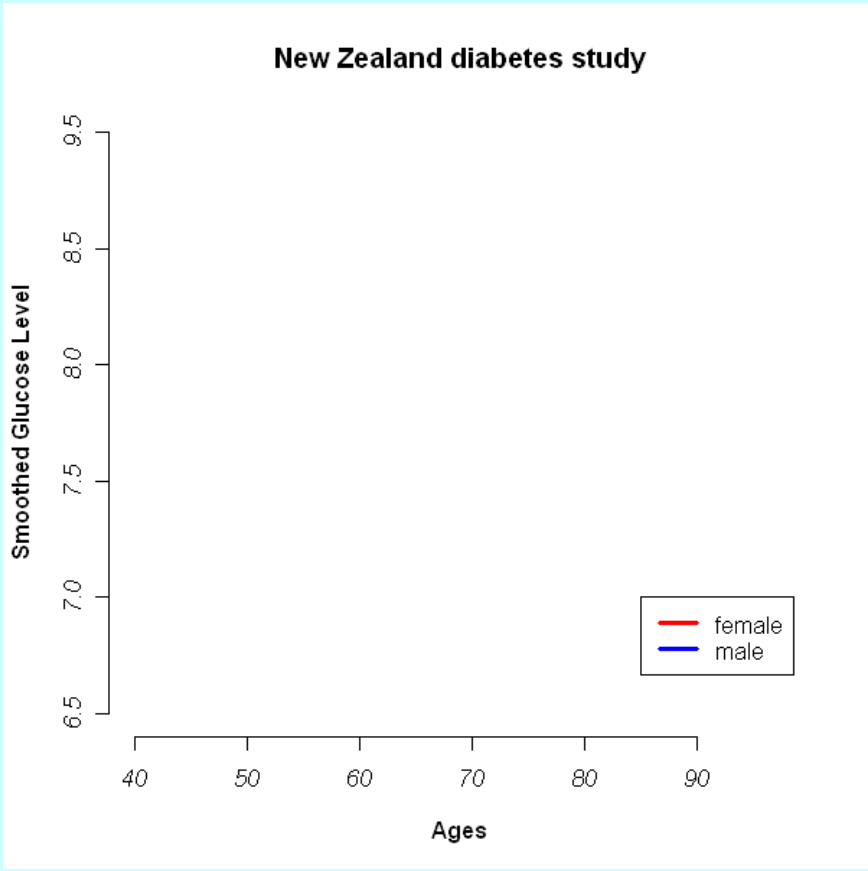
Figure 1: Evolution of blood glucose levels across ages, for female and male subjects



The horizontal axis, age, provides a natural **timeline** for the animation.



final "freeze-frame" of display



dynamic data display

The original printed page…

Admission blood glucose predicts survival following all types of Troponin positive acute coronary syndrome – increased mortality in New Zealand Maori

Adrian R Scott[1],
Anthony Cheng[1]
Michael Greenacre[3]
Gerard Devlin[2]
[1]*Diabetes and* [2]*Cardiology departments, Waikato Hospital, Hamilton, New Zealand.*
[3]*Departments of Economics and Life Sciences, Pompeu Fabra University, Barcelona*

The aim of this retrospective study was to investigate the effect of admission blood glucose and ethnicity on long term as well as short term mortality in patients admitted with ACS – both STEMI and NSTEMI.

Study Design and Methods

Waikato Health Board serves approximately 9% of the total New Zealand population (312,918 in the 1996 Census), in a largely rural setting in North Island. Approximately 20% identify themselves as Maori. An estimated 10000 (New Zealand Ministry of Health data) have diabetes (predominantly Type 2 –T2DM) which is more common amongst Maori and Pacific Islanders. Demographic, admission and discharge data for patients with ICD10 codes for all types of transmural MI, subendocardial MI or unstable angina between Jan 1st 1999 to Dec 31st 2002 was collected using the hospital patient database (HOSPRO). In addition the hospital Pathology laboratory database was interrogated and all patients with raised Trop I (up to 2000) or Trop T (2000 onwards) were identified. The two databases were merged for each patient and the admission blood glucose, maximum Troponin, creatinine, and cholesterol identified. The date of death was determined from data sent every 3 months from the New Zealand Health Information Service up to 1st April 2004. The re-classification of ACS by cardiologists is not reflected in ICD10 codes, so the data have been analysed in the following way: any transmural infarct has been classified as STEMI (N=1091). Admissions coded as subendocardial myocardial infarction have been analysed both separately and in combination with unstable angina or angina, unspecified. NSTEMI therefore includes any of the three diagnoses (N=3317).
Data were analysed using life-tables analysis, taking into account data censoring, to give estimated survival proportions, as well as fitting the Cox proportional-hazards regression model to assess the different effects of the covariates on survival time after ACS (20)

Results

There were 4408 (2846 male) episodes of ACS with a corresponding elevated Troponin. Patient characteristics can be seen in **Table 1**. A graph of glucose levels against age is shown in **Figure 1**, using half-Gaussian smoothing of values less than the age ordinate. The increase in male glucose

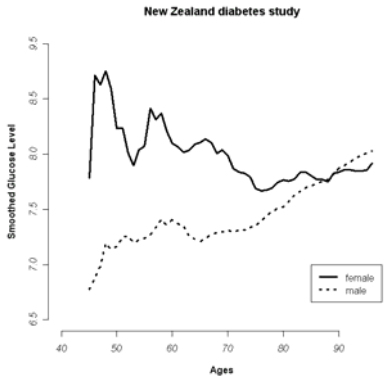


Figure 1: Evolution of blood glucose levels across ages, for female and male subjects

Here's the online version – click to freeze last frame.

Admission blood glucose predicts survival following all types of Troponin positive acute coronary syndrome – increased mortality in New Zealand Maori

Adrian R Scott[1],
Anthony Cheng[1]
Michael Greenacre[3]
Gerard Devlin[2]
[1]*Diabetes and* [2]*Cardiology departments, Waikato Hospital, Hamilton, New Zealand.*
[3]*Departments of Economics and Life Sciences, Pompeu Fabra University, Barcelona*

The aim of this retrospective study was to investigate the effect of admission blood glucose and ethnicity on long term as well as short term mortality in patients admitted with ACS – both STEMI and NSTEMI.

Study Design and Methods

Waikato Health Board serves approximately 9% of the total New Zealand population (312,918 in the 1996 Census), in a largely rural setting in North Island. Approximately 20% identify themselves as Maori. An estimated 10000 (New Zealand Ministry of Health data) have diabetes (predominantly Type 2 –T2DM) which is more common amongst Maori and Pacific Islanders. Demographic, admission and discharge data for patients with ICD10 codes for all types of transmural MI, subendocardial MI or unstable angina between Jan 1st 1999 to Dec 31st 2002 was collected using the hospital patient database (HOSPRO). In addition the hospital Pathology laboratory database was interrogated and all patients with raised Trop I (up to 2000) or Trop T (2000 onwards) were identified. The two databases were merged for each patient and the admission blood glucose, maximum Troponin, creatinine, and cholesterol identified. The date of death was determined from data sent every 3 months from the New Zealand Health Information Service up to 1st April 2004. The re-classification of ACS by cardiologists is not reflected in ICD10 codes, so the data have been analysed in the following way: any transmural infarct has been classified as STEMI (N=1091). Admissions coded as subendocardial myocardial infarction have been analysed both separately and in combination with unstable angina or angina, unspecified. NSTEMI therefore includes any of the three diagnoses (N=3317).
Data were analysed using life-tables analysis, taking into account data censoring, to give estimated survival proportions, as well as fitting the Cox proportional-hazards regression model to assess the different effects of the covariates on survival time after ACS (20)

Results

There were 4408 (2846 male) episodes of ACS with a corresponding elevated Troponin. Patient characteristics can be seen in **Table 1**. A graph of glucose levels against age is shown in **Figure 1**, using half-Gaussian smoothing of values less than the age ordinate. The increase in male glucose

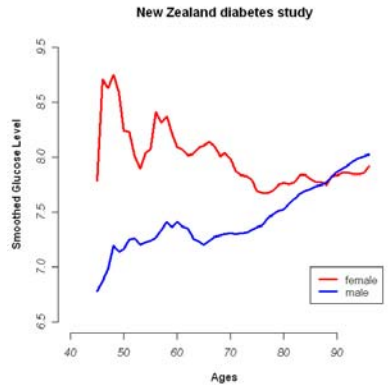


Figure 1: Dynamic data display of the evolution of blood glucose levels across ages, for female and male subjects

Animating some graphics from highly cited papers in the life sciences

SCOPUS top-cited papers

Several illustrations of our approach have been taken from the list of top-cited papers in the abstract and citation database SCOPUS. Of the top 20 cited papers since 2004, with numbers of citations ranging from 1500 to 4453, there are

- **12 in biological and health sciences**
- **4 in chemistry**
- **3 in physics**
- **1 in computer science**

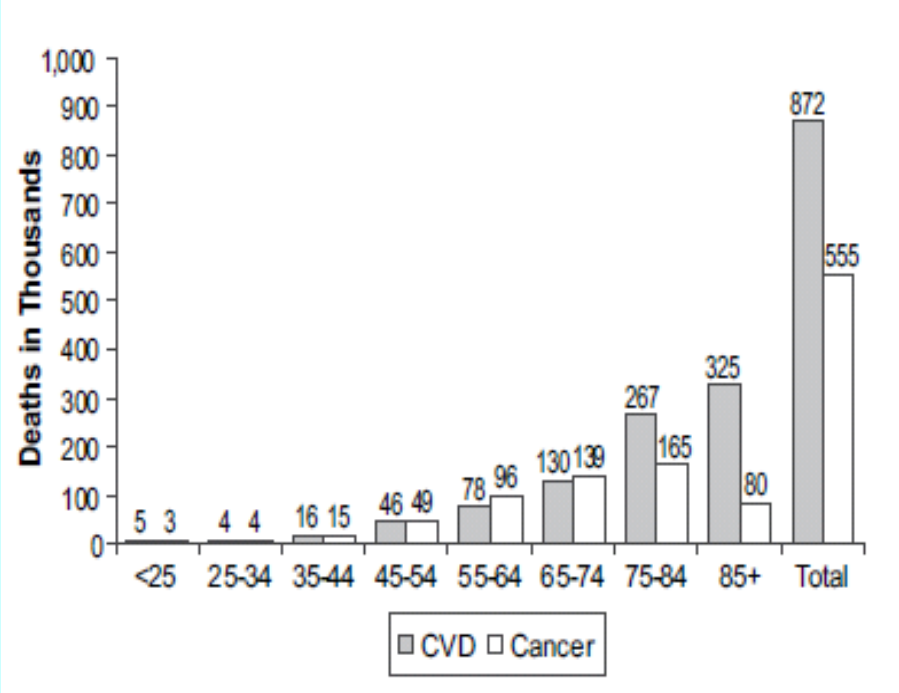
Furthermore, of the 12 articles in the biological and health sciences, there are

- **6 research articles**
- **4 articles with updated statistics on disease**
- **2 articles about software**

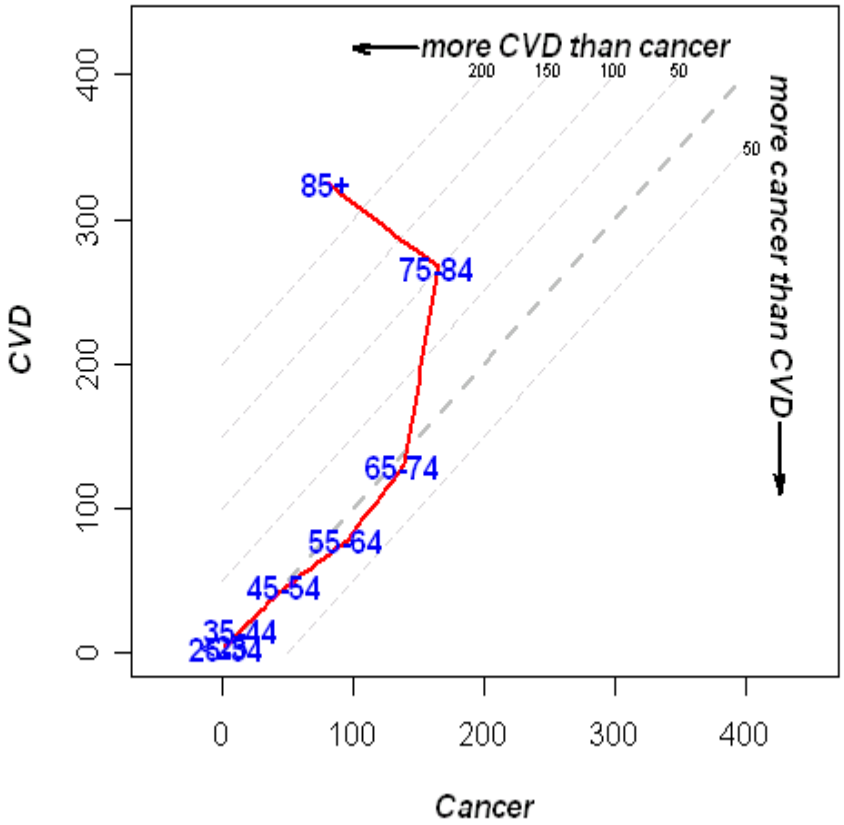
Cardiovascular (CVD) deaths vs cancer deaths by age (deaths in 1000s, USA, 2004)

(Rosamond, et al. (2007). Heart disease and stroke statistics - 2007 Update. *Circulation (Journal of the American Heart Association)* 115:69–171.

original histogram representation



final frame of display



Into the third dimension

Data again from Rosamond et al. (2007), showing age-adjusted death rates for cardiovascular disease (CVD), coronary heart disease (CHD) and stroke, for each state in the USA. This data table occupies over a page of the journal article.

Three variables can be plotted together in three dimensions, but we can only see the third dimension if motion is introduced – the most obvious way is to define the timeline as a rotation of the configuration.

CVD CHD Stroke

U.S. states

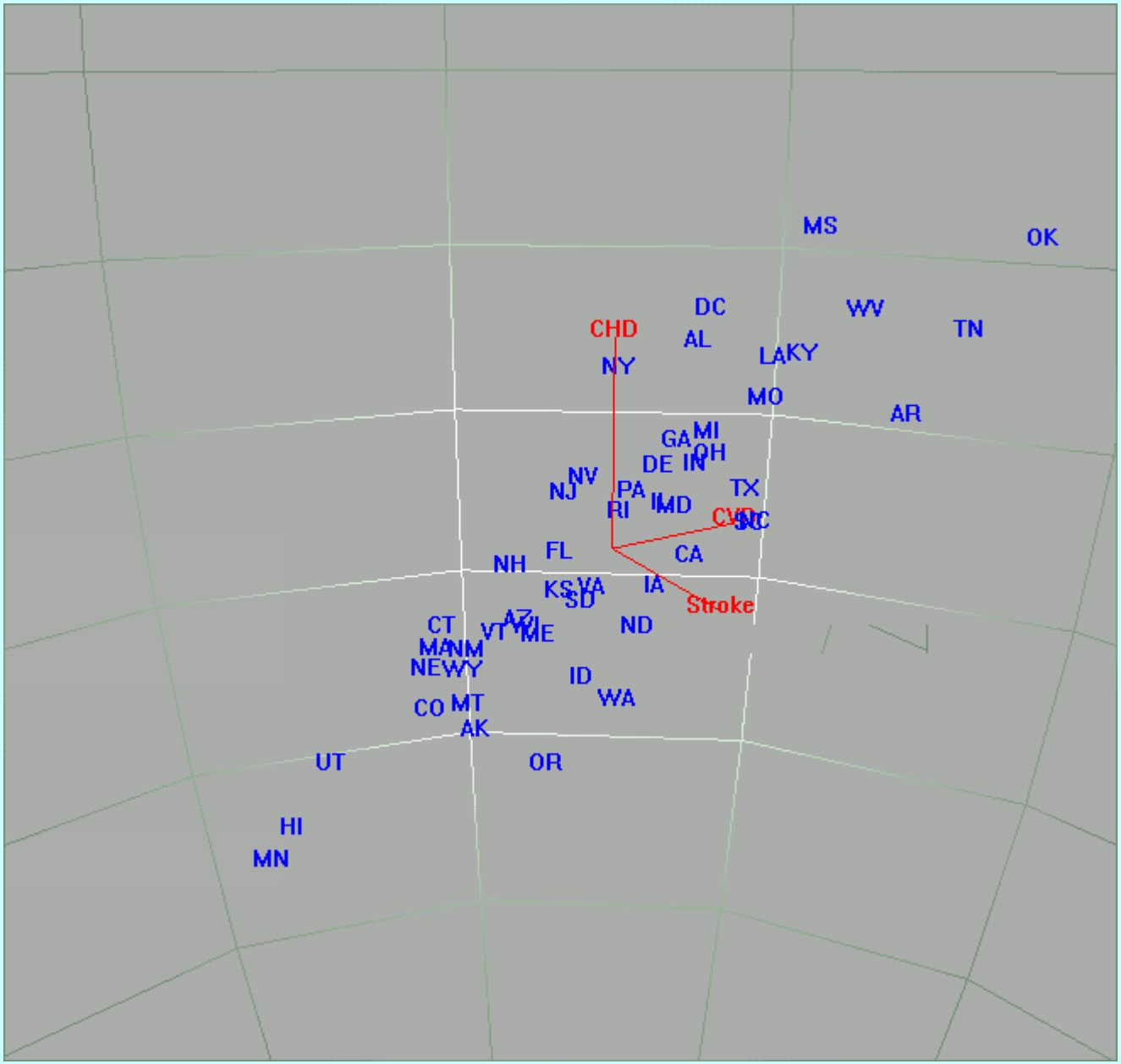
TABLE 2-2. 2003 Age-Adjusted Death Rates for CVD, CHD, and Stroke by State (Includes District of Columbia and Puerto Rico)

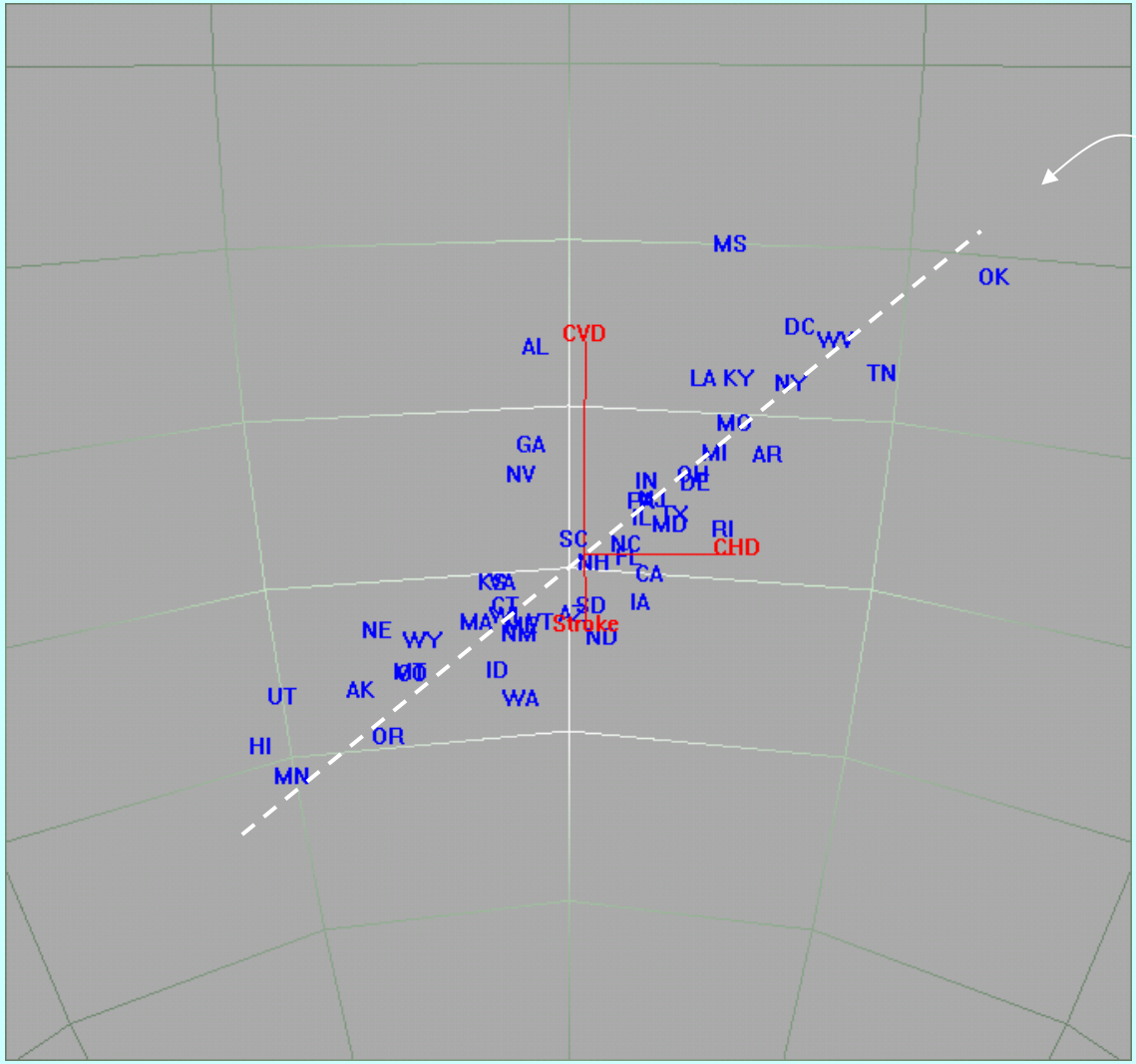
State	CVD* Rank§	CVD* Death Rate	CVD* % Change# 1993 to 2003	CHD† Rank§	CHD† Death Rate	CHD† % Change# 1993 to 2003	Stroke‡ Rank§	Stroke‡ Death Rate	Stroke‡ % Change# 1993 to 2003
Alabama	49	378.4	−11.5	22	143.0	−23.9	47	66.5	−5.8
Alaska	16	273.7	−23.6	4	114.2	−35.5	40	60.9	−19.7
Arizona	9	261.5	−22.7	24	149.2	−27.6	8	44.4	−22.7
Arkansas	43	353.2	−15.6	45	181.9	−22.6	51	70.7	−20.0
California	27	298.0	−20.1	34	164.3	−29.1	30	56.7	−16.4
Colorado	4	258.1	−19.9	6	118.8	−32.2	16	50.9	−16.0
Connecticut	6	260.1	−26.9	11	134.2	−31.8	5	42.7	−21.7
Delaware	32	311.1	−20.9	41	176.7	−25.4	13	48.7	−13.2
District of Columbia	44	357.9	−12.6	48	204.5	+26.7	11	45.4	−39.8
Florida	20	277.4	−20.2	29	162.2	−26.4	6	43.6	−22.4
Georgia	42	348.5	−18.3	20	142.0	−33.0	44	65.2	−17.3
Hawaii	2	241.7	−20.9	1	96.0	−34.5	19	52.6	−9.9
Idaho	17	275.6	−20.0	13	135.3	−28.5	35	59.0	−17.1
Illinois	31	310.0	−25.0	30	162.9	−34.6	24	54.1	−21.5
Indiana	37	326.5	−22.9	32	163.1	−32.0	33	57.4	−23.8
Iowa	24	284.4	−22.4	31	162.9	−30.8	23	53.7	−14.0
Kansas	26	296.0	−19.3	12	134.5	−30.6	32	56.8	−13.6
Kentucky	46	362.6	−16.3	43	179.9	−25.7	42	61.3	−15.6
Louisiana	45	362.4	−18.6	38	172.9	−30.2	41	61.0	−15.0
Maine	15	273.4	−18.6	17	138.9	−35.3	17	51.7	−12.1
Maryland	29	306.3	−17.7	37	169.2	−21.6	21	53.1	−13.4
Massachusetts	8	260.3	−24.9	10	128.7	−34.8	10	45.1	−16.2
Michigan	38	327.2	−21.4	42	179.2	−30.4	20	52.7	−21.8
Minnesota	1	221.2	−31.7	2	97.0	−42.9	12	47.2	−32.4
Mississippi	51	405.9	−16.0	48	176.1	−28.6	43	63.6	−12.3
Missouri	41	344.3	−18.5	44	181.2	−27.0	34	57.7	−16.4
Montana	10	267.5	−21.4	8	119.9	−30.7	28	55.3	−19.1
Nebraska	19	277.1	−26.7	5	114.6	−37.5	25	54.2	−17.7
Nevada	39	327.4	−28.1	18	139.4	−36.4	31	56.7	−13.6
New Hampshire	12	270.8	−25.5	26	154.0	−30.9	3	41.4	−21.3
New Jersey	25	292.4	−22.3	36	168.5	−28.6	4	41.8	−25.1
New Mexico	3	255.0	−28.1	16	137.2	−21.8	7	44.1	−26.1
New York	33	319.1	−27.0	50	212.4	−30.2	1	35.0	−29.1
North Carolina	34	321.9	−21.3	28	158.0	−31.2	46	65.8	−21.6
North Dakota	22	277.6	−23.3	27	154.6	−28.5	29	55.4	−18.8
Ohio	36	324.7	−20.4	39	173.6	−28.5	27	55.0	−11.2
Oklahoma	50	400.7	−8.6	51	226.1	−9.1	48	69.0	−5.9
Oregon	11	270.5	−22.1	7	119.5	−36.4	45	65.4	−16.4
Pennsylvania	30	308.9	−23.8	33	163.3	−32.1	15	50.8	−18.1
Puerto Rico	…	232.7	…	…	116.5	…	…	46.2	…
Rhode Island	23	280.9	−22.7	46	188.5	−22.8	2	41.2	−27.1
South Carolina	40	328.9	−25.2	23	148.8	−35.7	50	69.5	−23.5
South Dakota	18	275.9	−24.1	25	152.8	−33.0	14	49.9	−17.9
Tennessee	48	373.6	−15.5	49	205.0	−21.1	49	69.0	−20.3
Texas	35	322.1	−16.1	35	168.0	−24.7	37	60.6	−13.9
Utah	5	258.7	−18.2	3	100.3	−37.5	26	54.2	−12.8
Vermont	7	260.2	−30.0	21	142.1	−36.1	9	45.9	−28.3
Virginia	28	300.5	−25.5	15	136.4	−31.4	36	59.4	−20.9
Washington	13	271.3	−21.2	19	139.6	−24.6	39	60.7	−15.0
West Virginia	47	373.4	−17.1	47	201.0	−24.7	38	60.6	−9.2
Wisconsin	21	277.4	−23.7	14	135.7	−33.7	18	52.1	−23.5
Wyoming	14	272.5	−21.6	9	121.6	−33.1	22	53.4	−26.2
Total United States		307.7	−22.2		162.9	−29.3		53.5	−19.5

Virginia	28	300.5	−25.5
Washington	13	271.3	−21.2
West Virginia	47	373.4	−17.1
Wisconsin	21	277.4	−23.7
Wyoming	14	272.5	−21.6
Total United States		307.7	−22.2

USA states in three-dimensional space of CVD, CHD and Stroke

Data: log(death rates); map centered at averages (CVD:308, CHD:163, Stroke: 54)





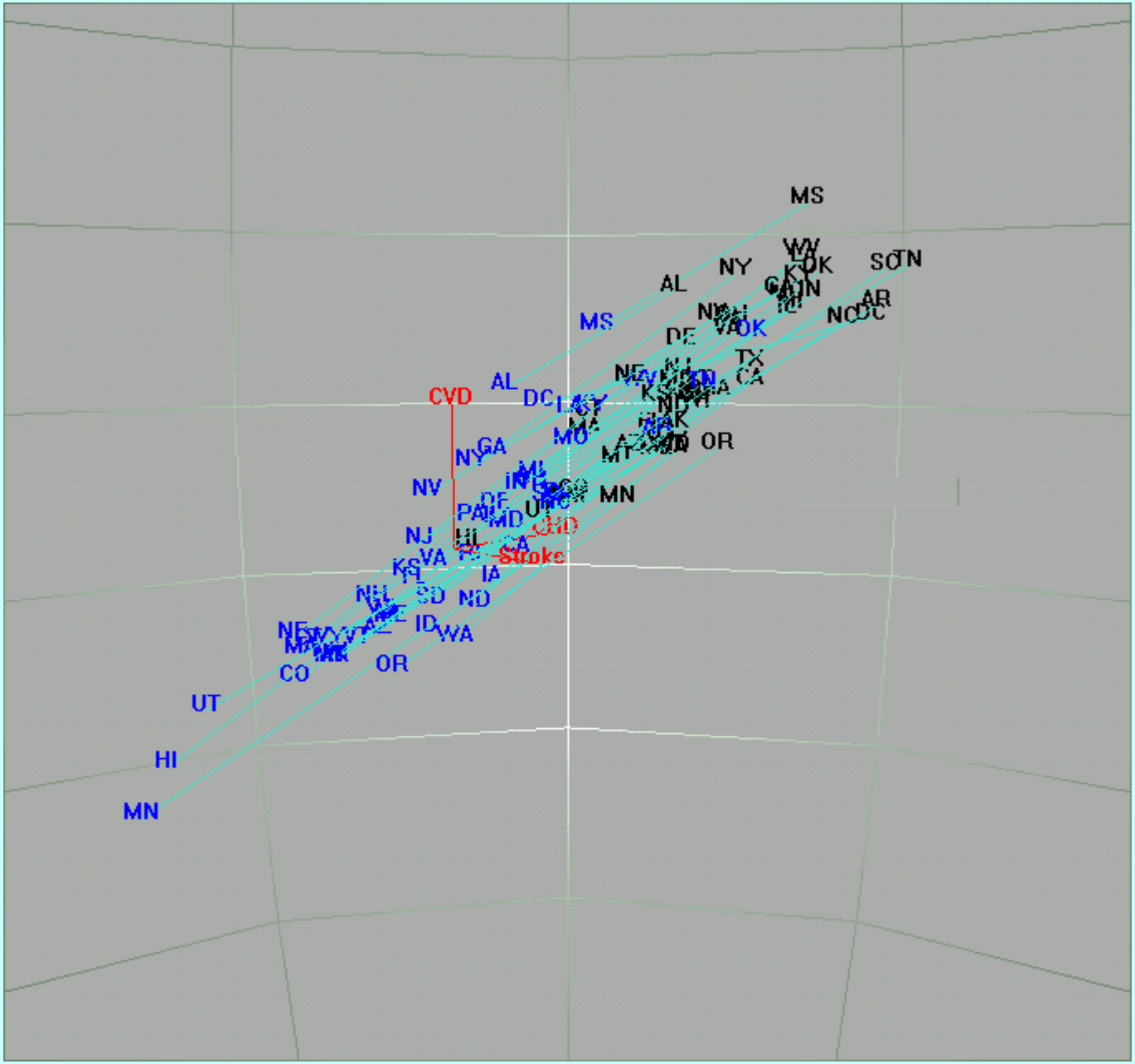
STATISTICAL LEARNING

This is an example of **unsupervised learning**. There is not specific outcome variable that we are trying to predict, and we discovered this close relationship between the variables during our dynamic exploration of the data.

Later we look at **supervised learning**, where there is an outcome variable, for example probability of survival or type of cancer.

USA states in three-dimensional space of CVD, CHD and Stroke

Data: log(death rates) in 1993 (black) and 2003 (blue); map centered at averages for 2003, as before

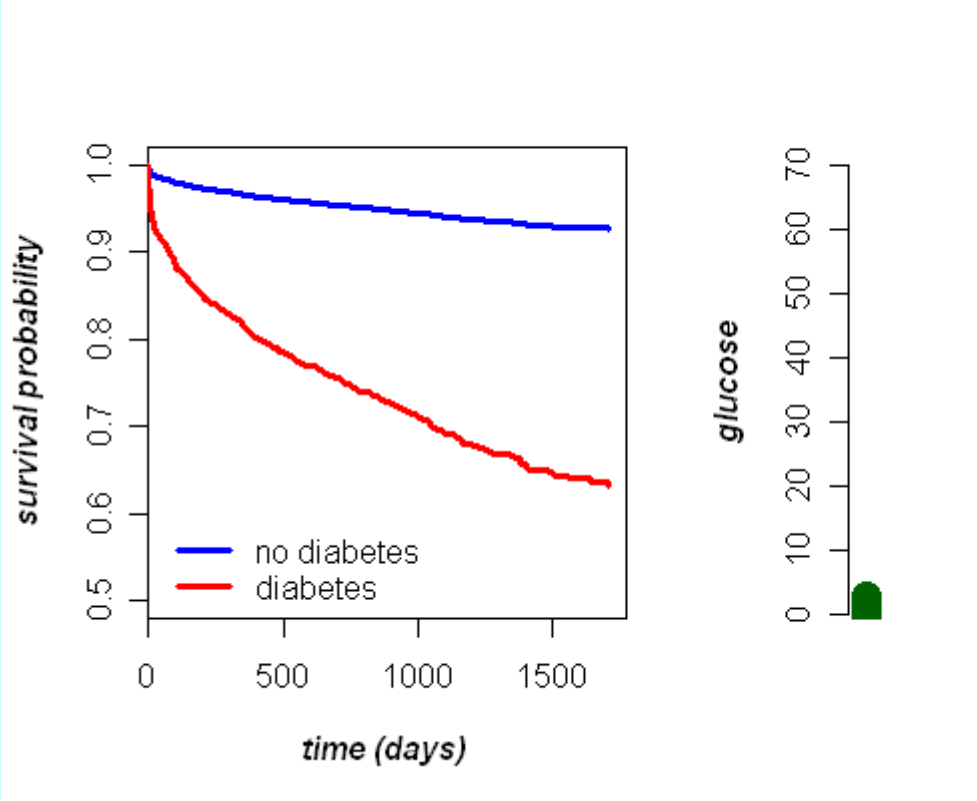


Another example in three dimensions

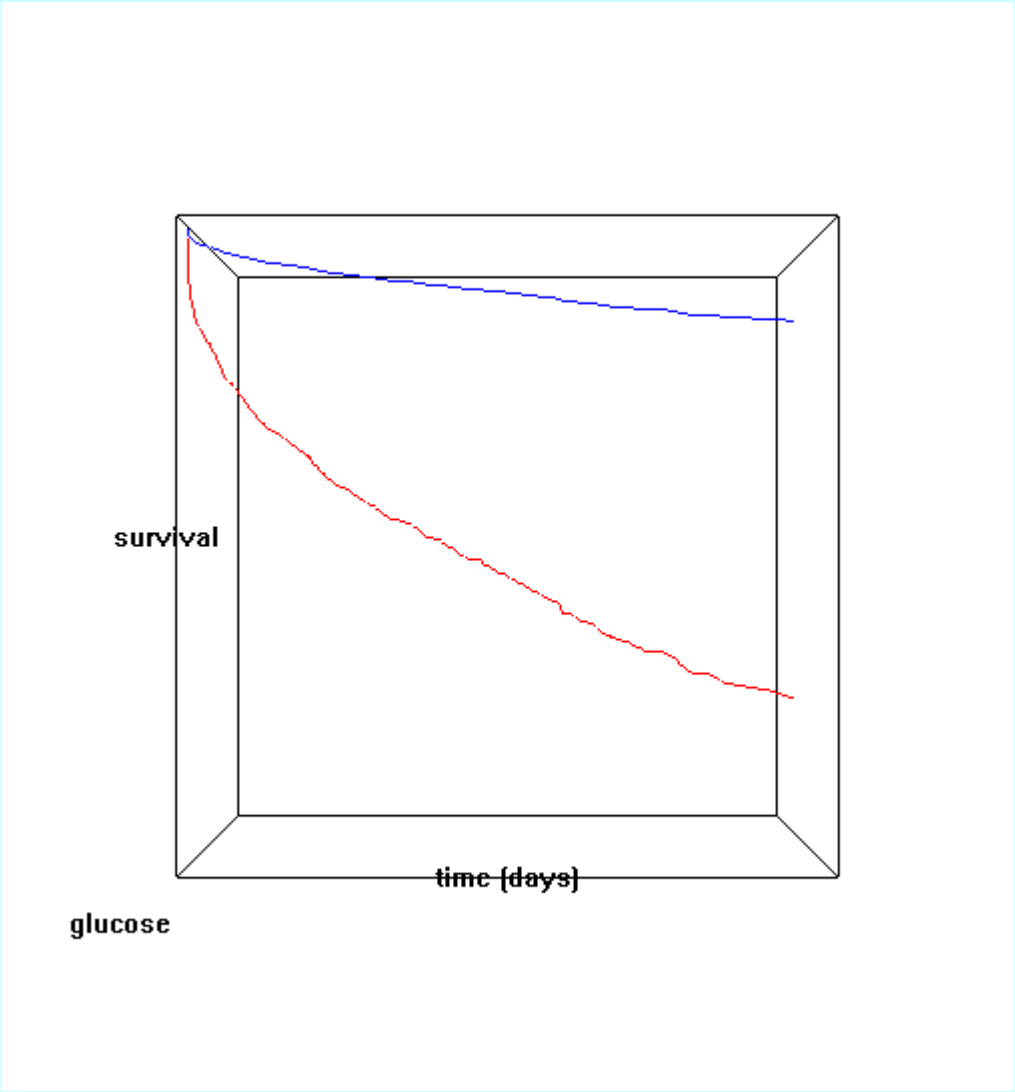
Data source:

Scott, A.R., *et al.*, (2007). Implications of hyperglycaemia and ethnicity in patients with acute coronary syndromes in New Zealand. *Diabetes, Obesity and Metabolism* 9:121–126.

Modeled survival curves for patients after heart attack, separated by diabetic status *and as a dynamic function of glucose level*.

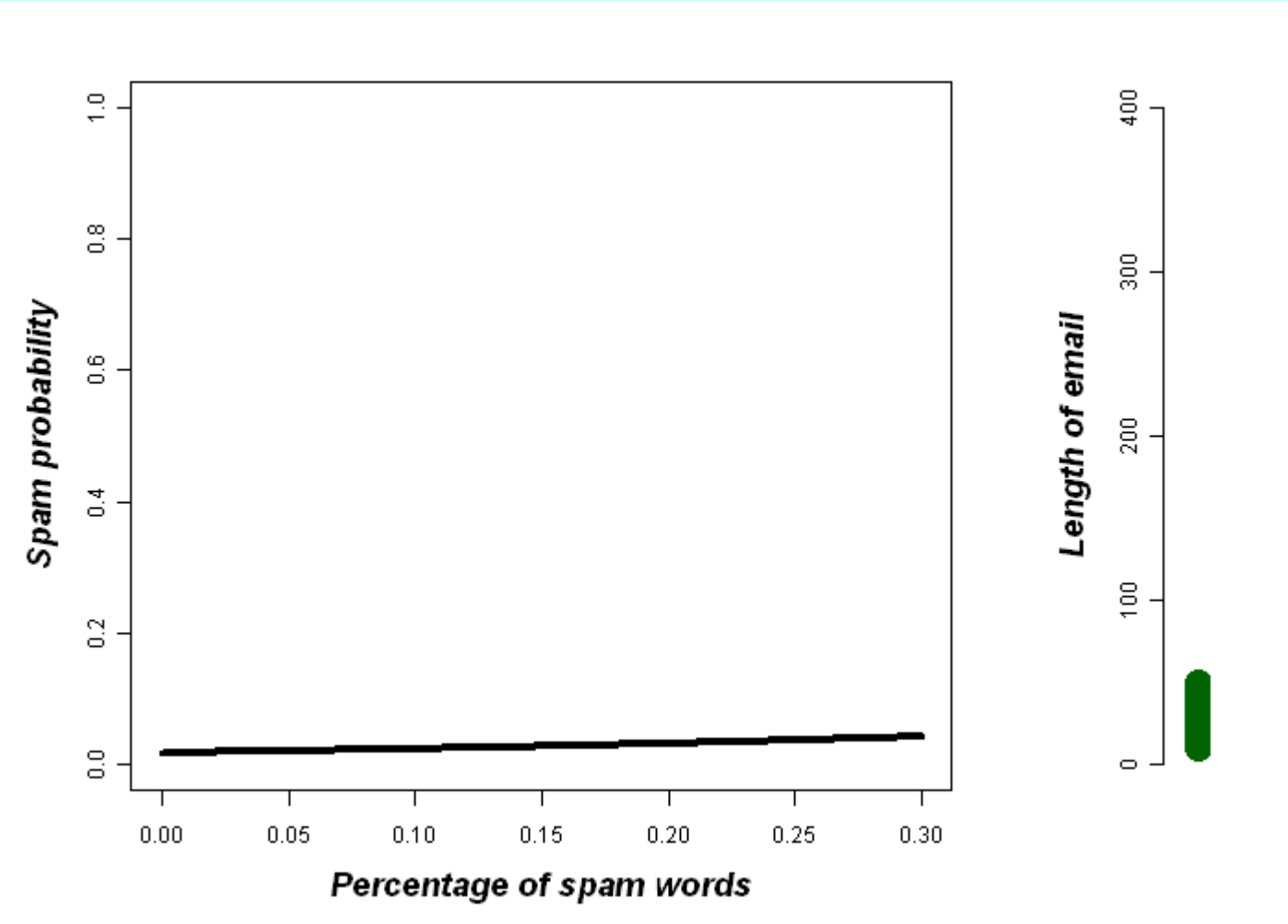


Modeled survival surfaces for patients after heart attack, separated by diabetic status and as a function of glucose level; surface is built up for increasing glucose as image rotates.



Another conditioned plot: the model (a logistic regression) predicts probability p that an email is SPAM, given the proportion S of "spam words" and the length L of the email

$$\log\left(\frac{p}{1-p}\right) = -3.37 - 20.0S + 0.0142L + 0.0000279L^2 + 20.52SL - 0.00114SL^2$$



Into *n*-dimensional space

Using motion for

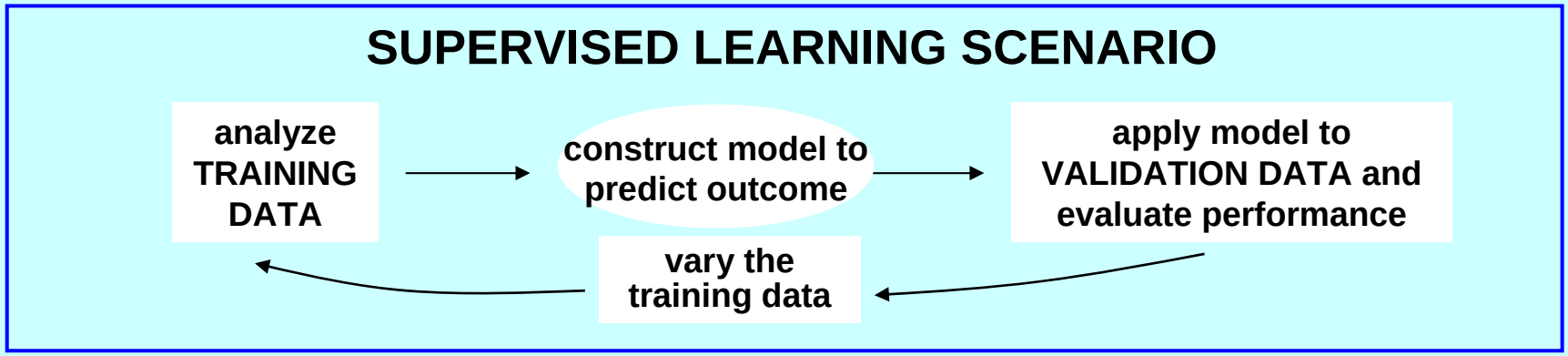
– two-dimensional data: can be useful if there is an obvious “timeline”

– three-dimensional data: is essential for seeing the interrelationships

– high-dimensional data: permits us to take a guided tour through the data space

Large data set (source: Hastie et al. *The Elements of Statistical Learning. Second Edition*, to appear 2009): Gene expression data for 7399 genes in a study of 160 Lymphoma patients.

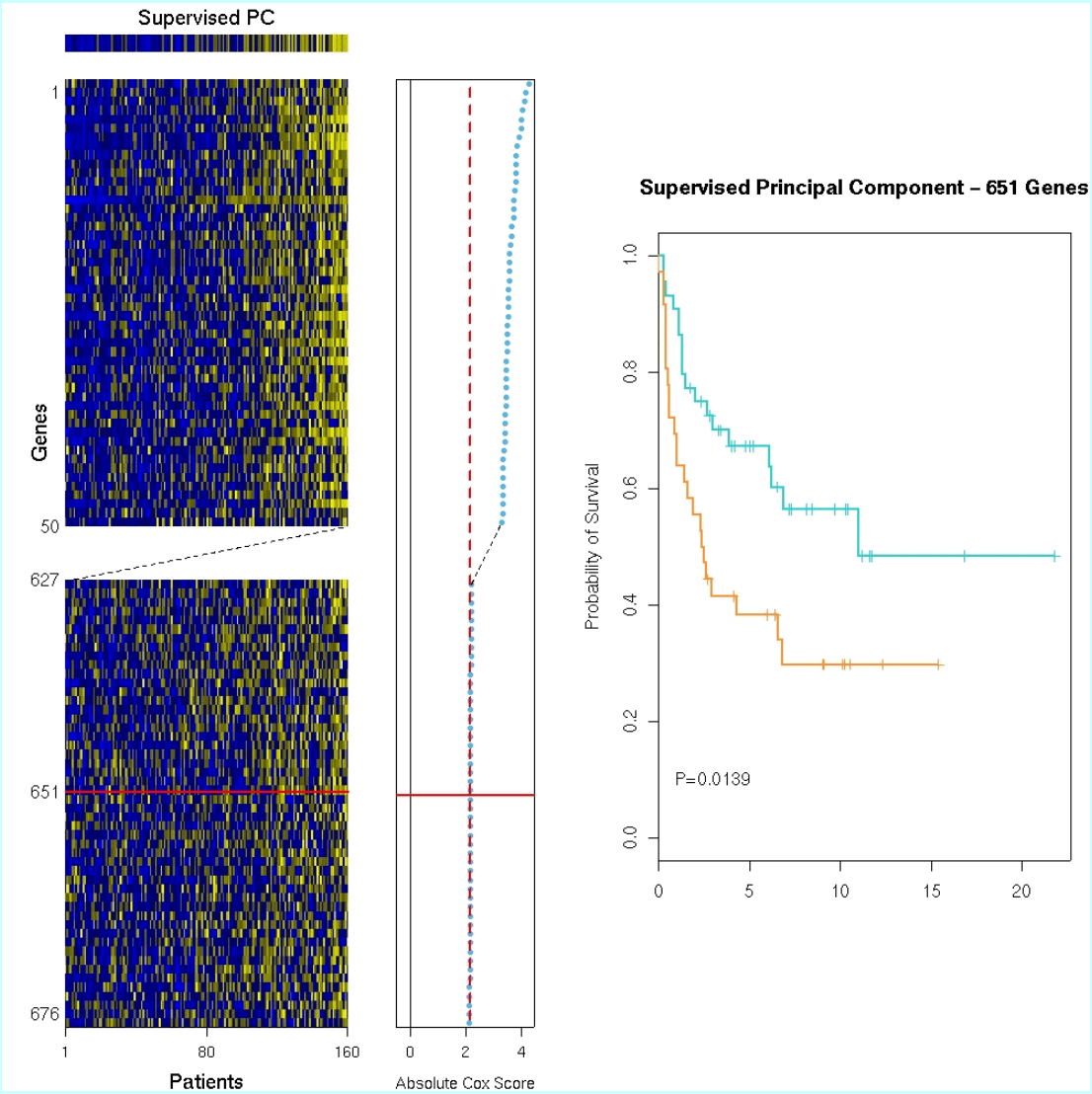
Objective: to look for a small subset of genes that can reliably predict the probability of survival. Motion will allow us to visualize the learning process.



First principal component

The decreasing Cox scores. Red line indicates a threshold on the Cox score for selecting genes in the training set

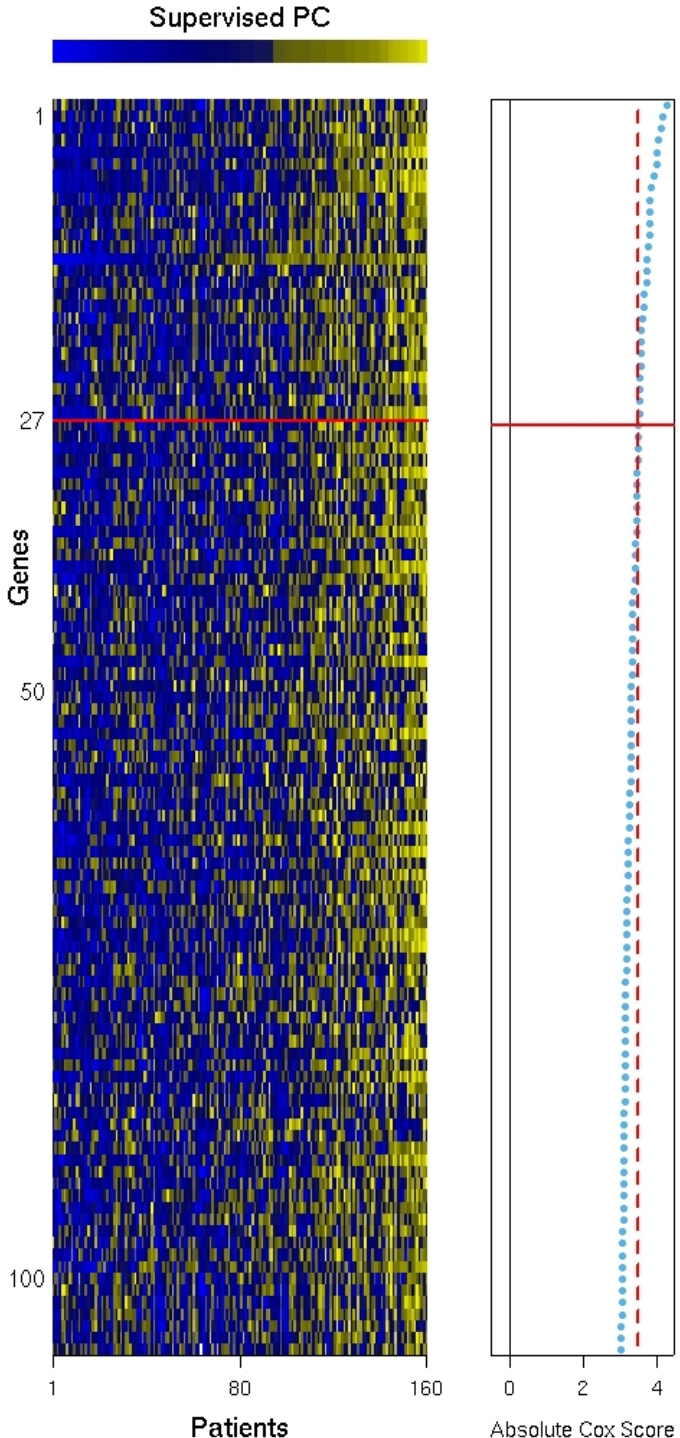
Gene-expression **training data,** where the data are colour coded from blue (low expression) to yellow (high expression) – each row is a **gene** (7399 genes), each column a **Lymphoma patient** (160 patients).

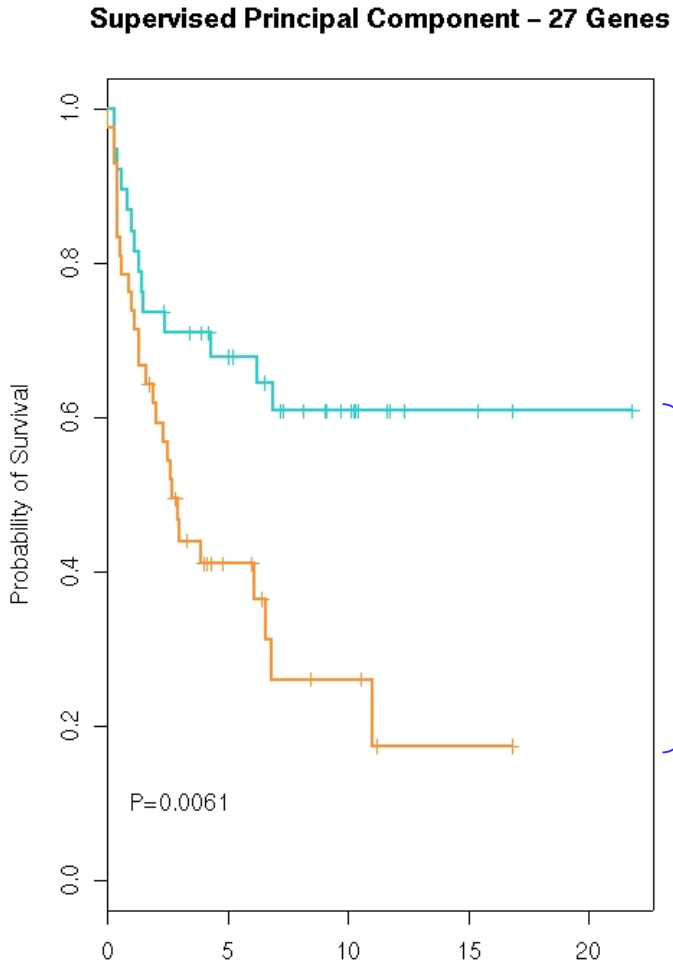


Separate survival curves when sample is divided into two according to principal component.

Objective: to find a subset of genes that gives **maximum separation of the survival curves**.

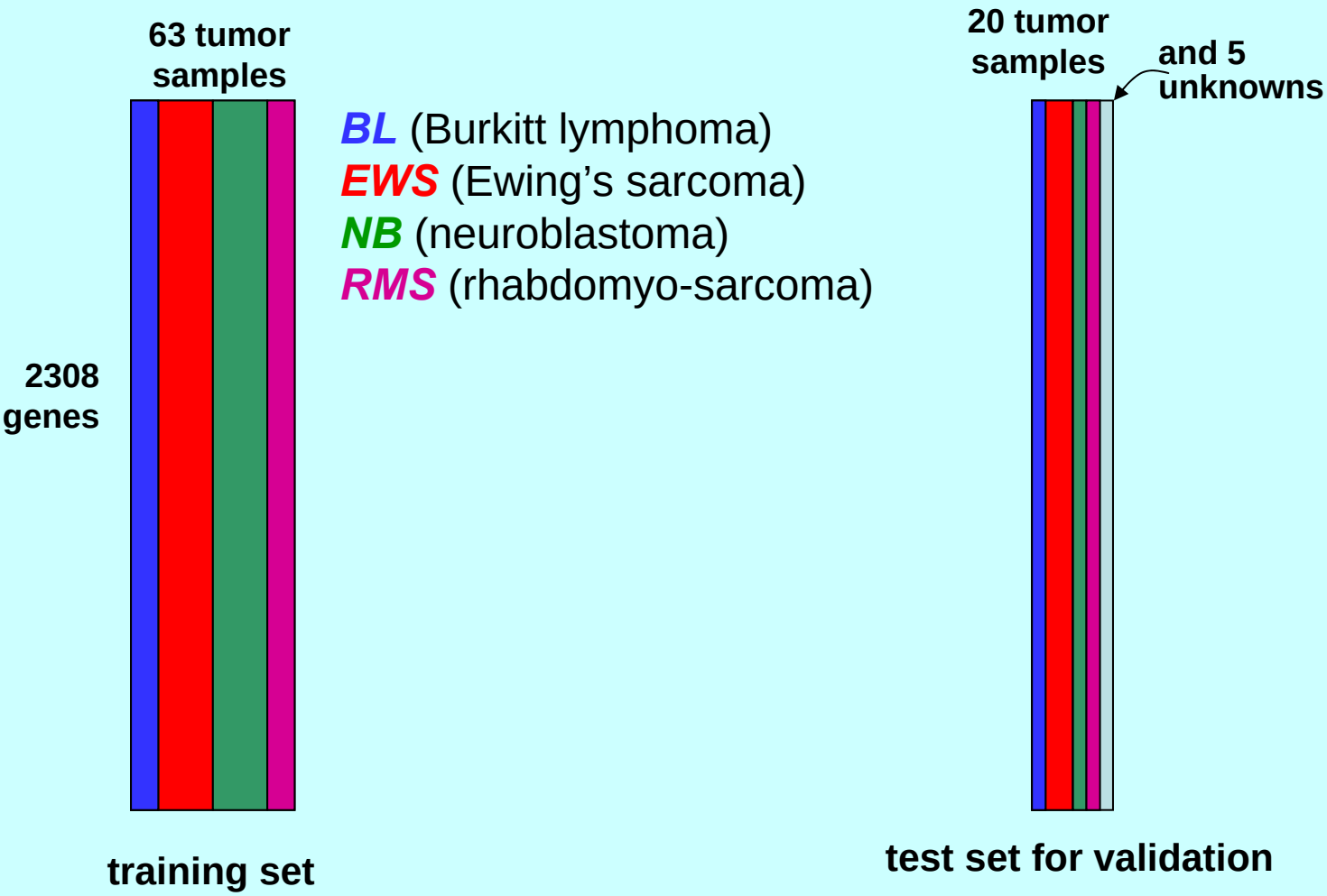
Evolution of supervised learning process to find best set of genes to predict survival; the best predictors are a set of 27 genes.





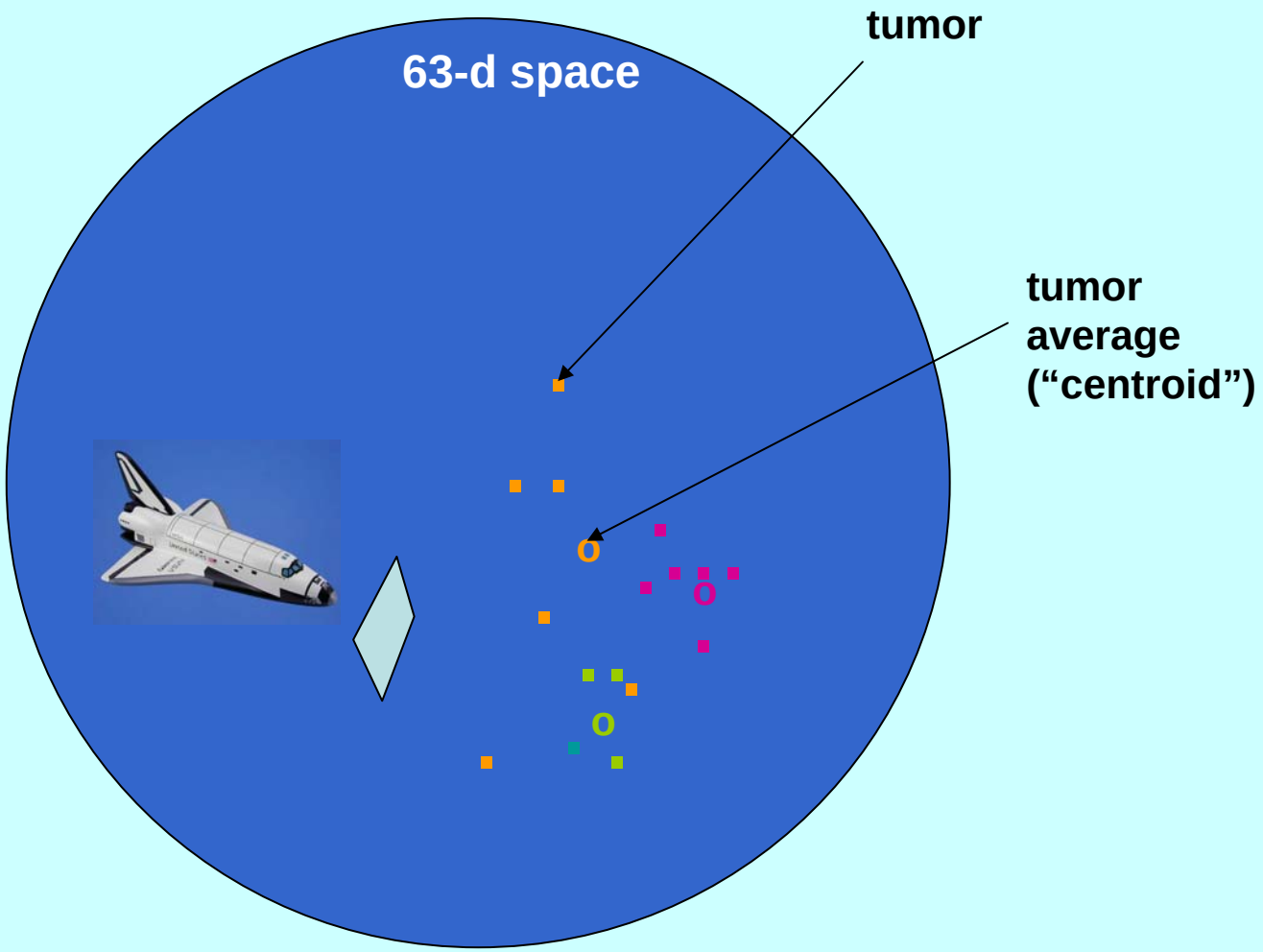
Good distinction between survival curves for indepen-dent test data

Another example from Hastie et al. (2009)
The Elements of Statistical Learning. 2nd Edition.

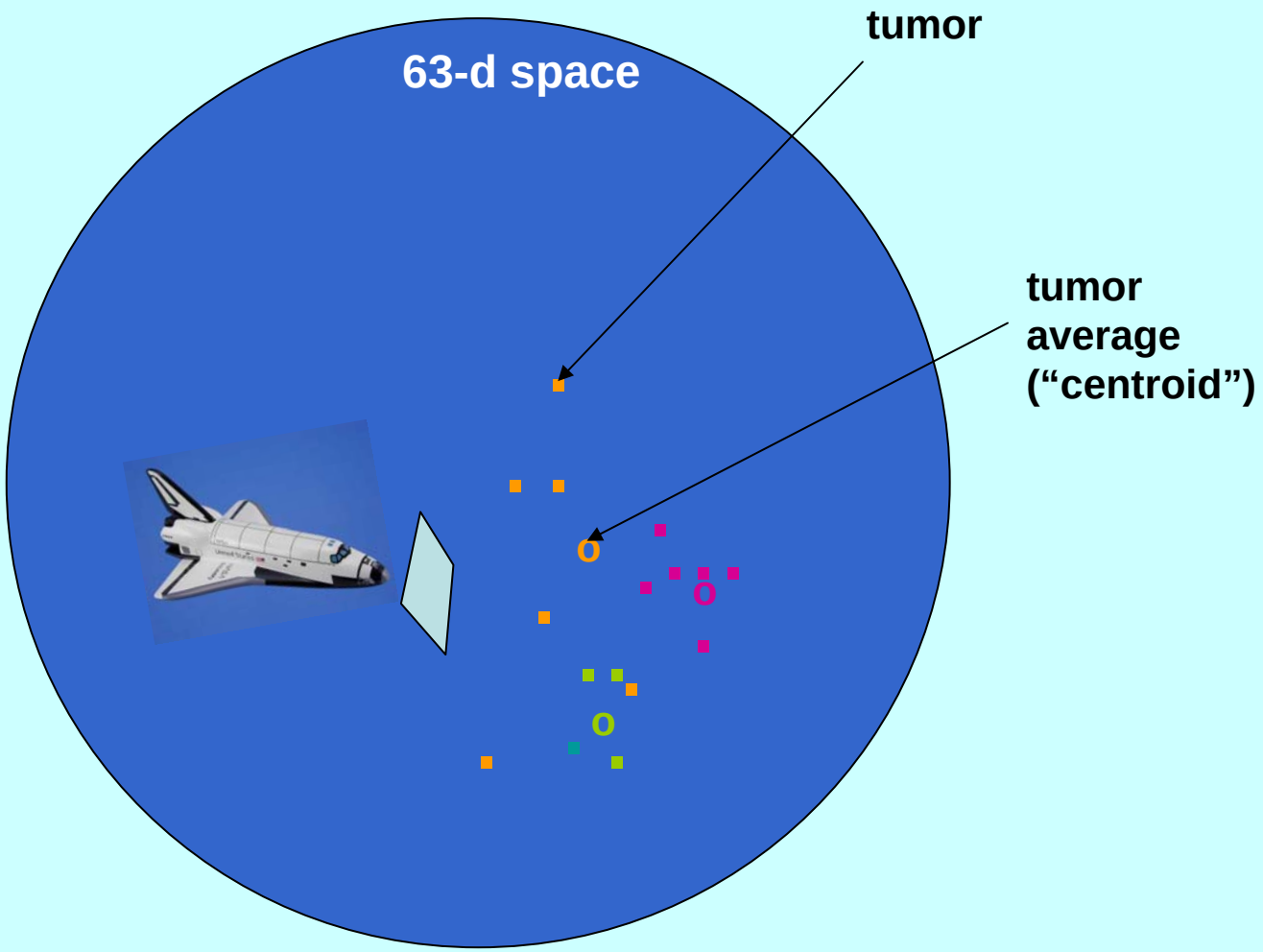


Objective: to identify a small subset of genes that is determinant in classifying the type of tumour

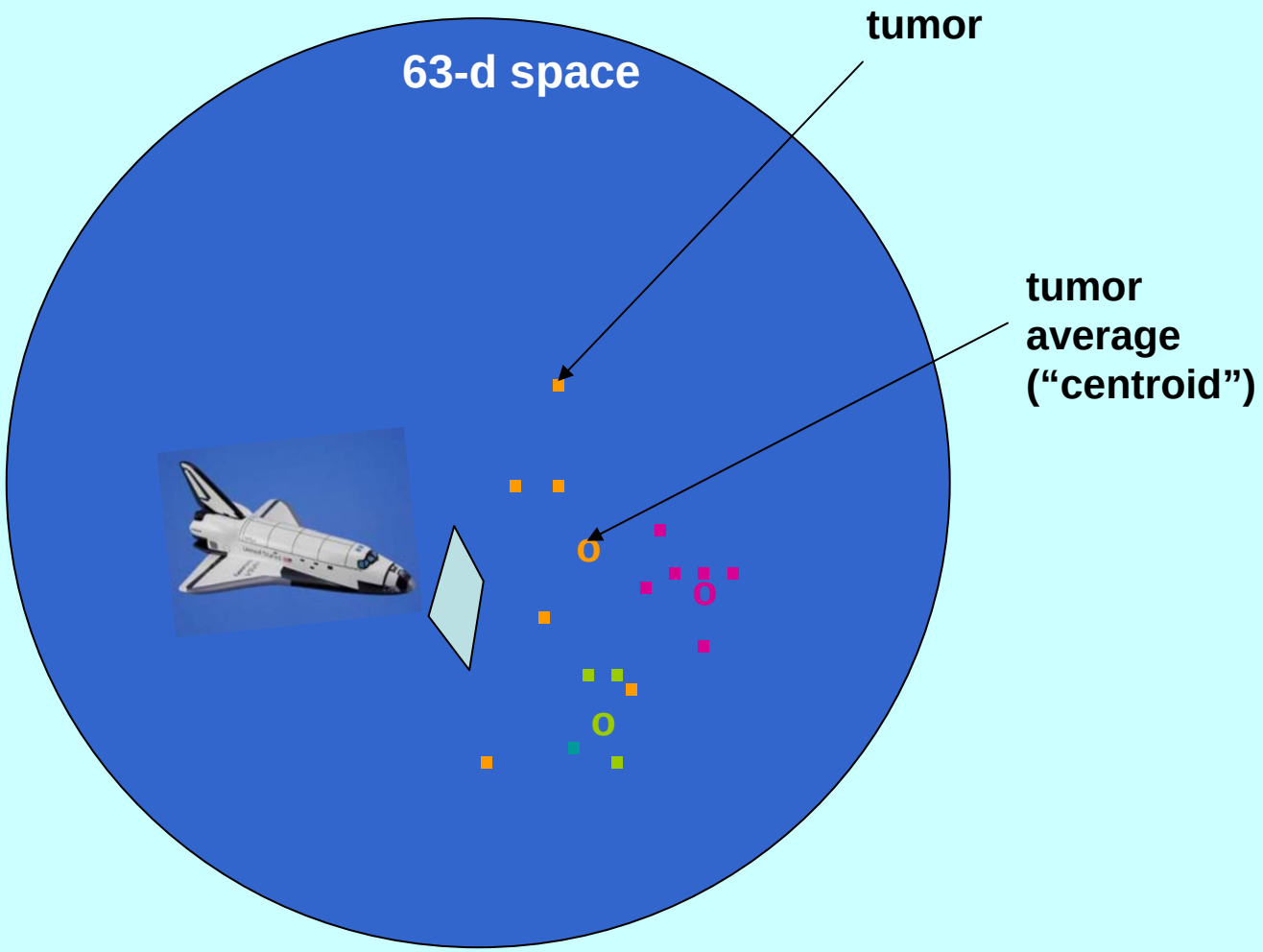
Guided tour through n-dimensional space



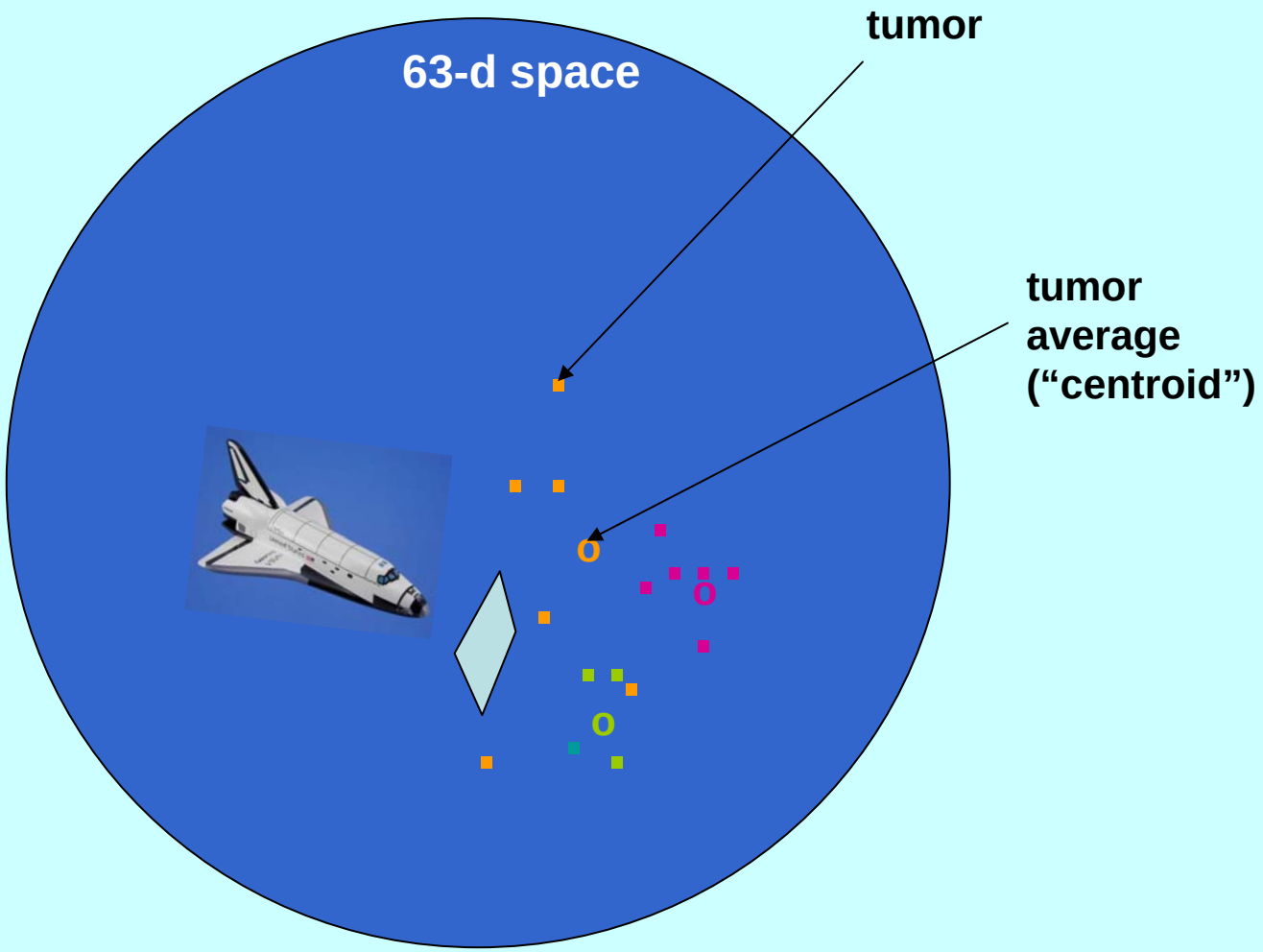
Guided tour through n-dimensional space



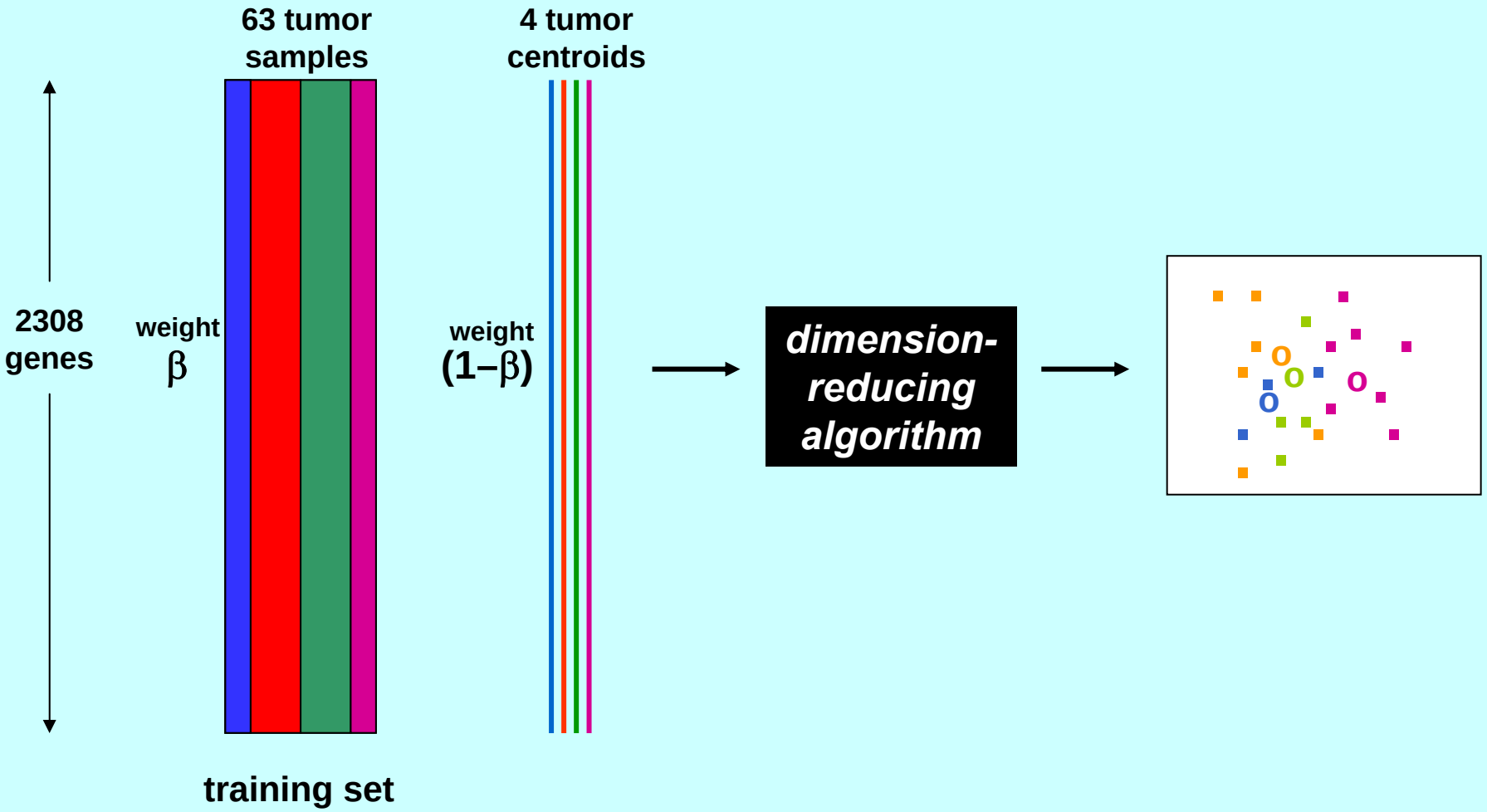
Guided tour through n-dimensional space



Guided tour through n-dimensional space



How do we do the tour?

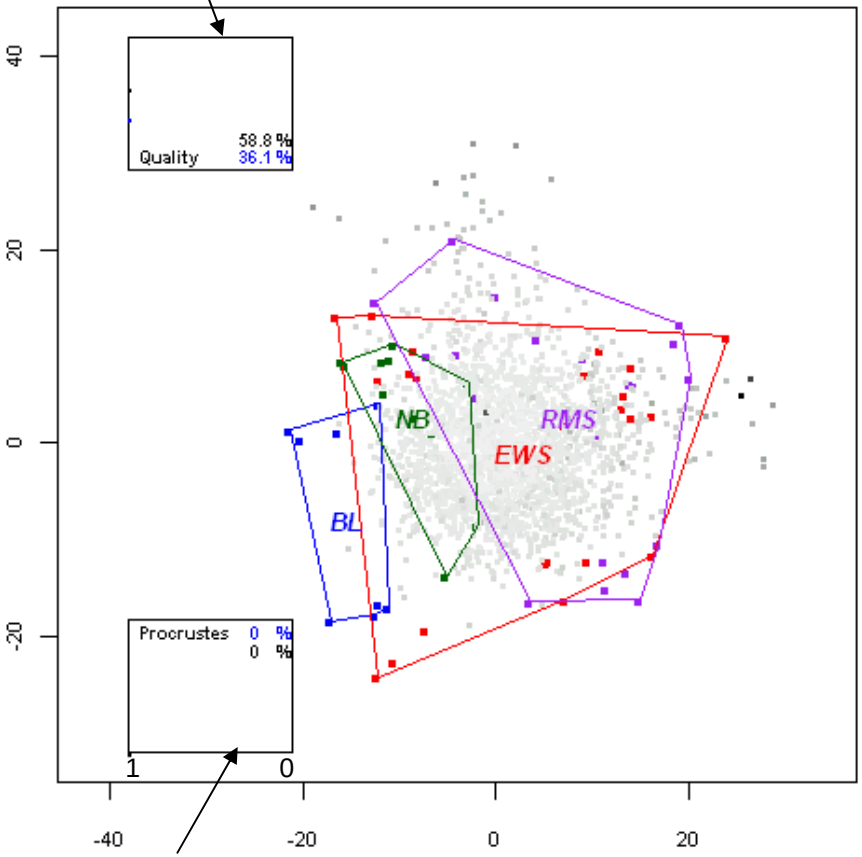


Let weight β vary smoothly from 1 to 0 (typically 1, 0.99, 0.98,…, 0.01, 0) – each time we get a different view as we move into the multidimensional space. β is a tuning parameter.

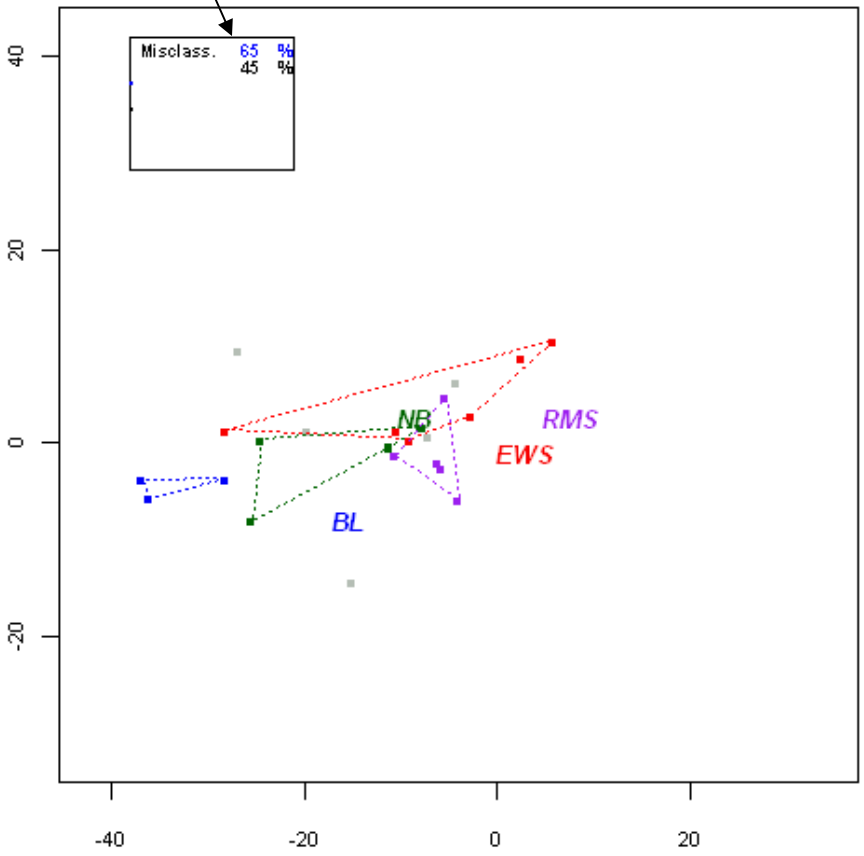
Shows how well the training set centroids are being depicted (blue:2-d,black:3-d)

Shows the misclassification rate for the test data, i.e. the error of our model (blue:2-d,black:3-d)

Training data



Test data



Shows how far we have moved into "deep space" (blue:2-d,black:3-d)

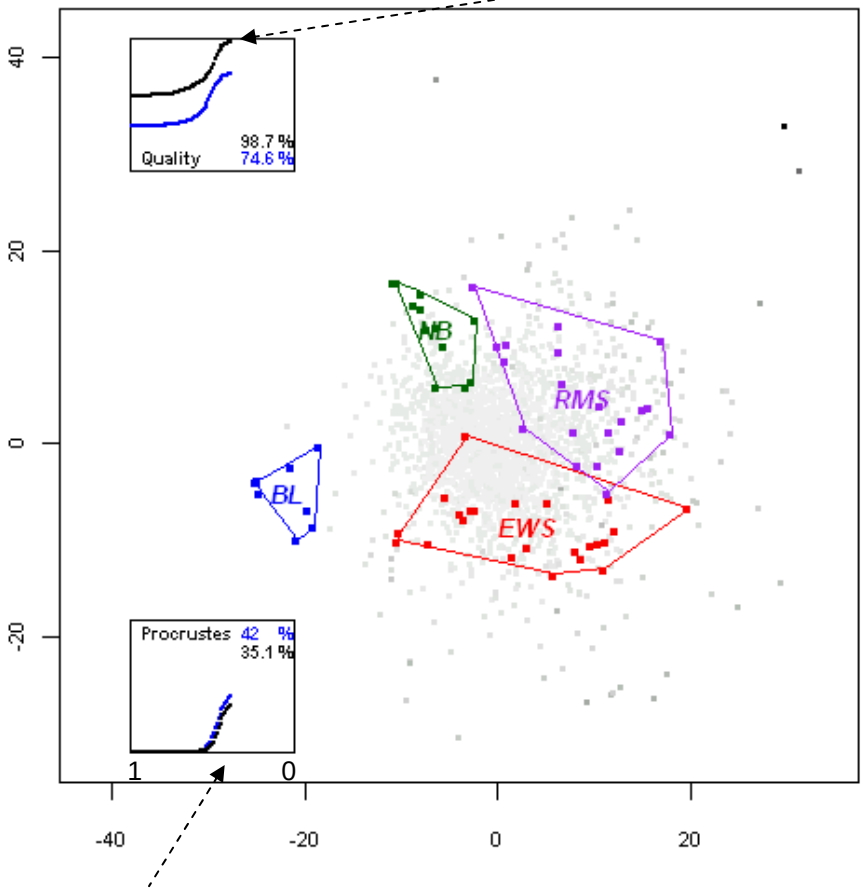
Voyage through 63-dimensional space, looking for best predictor of tumors in test data

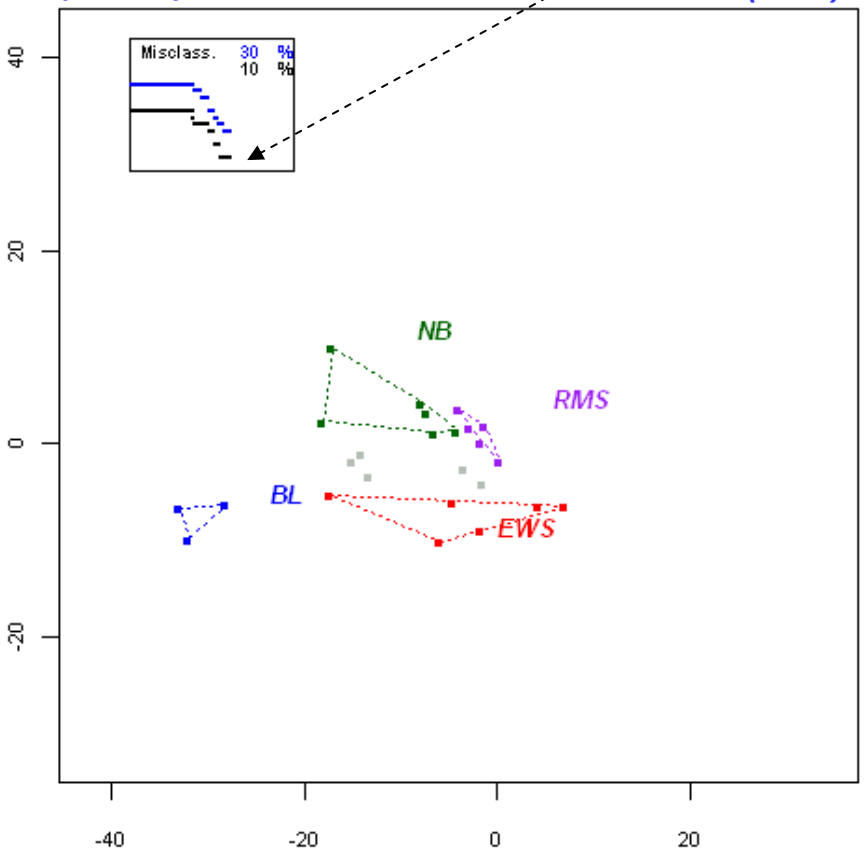
Training data

Quality of representation of tumor centroids almost perfect (98.7%)

Test data

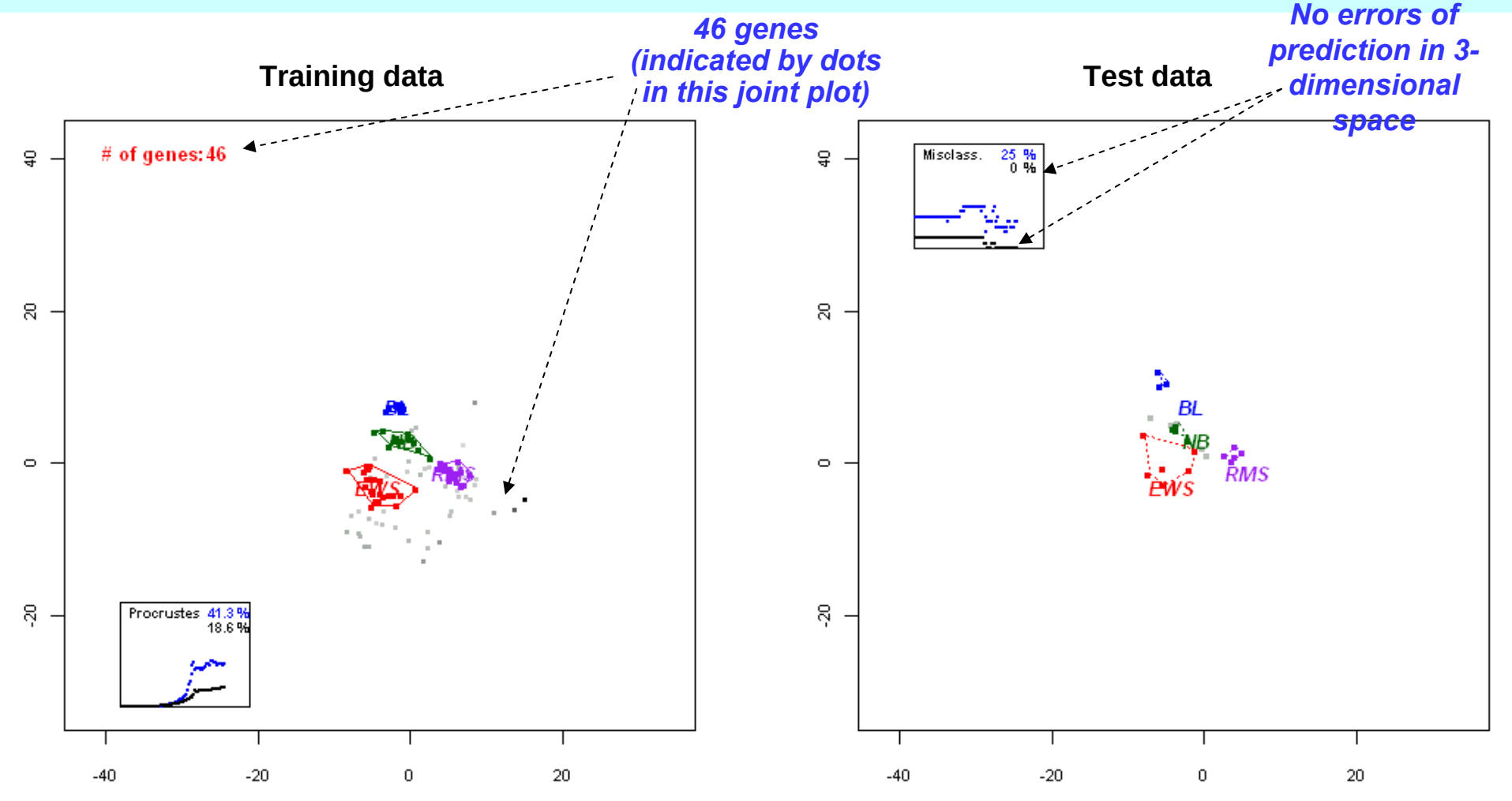
Lowest misclassification achieved (10%)



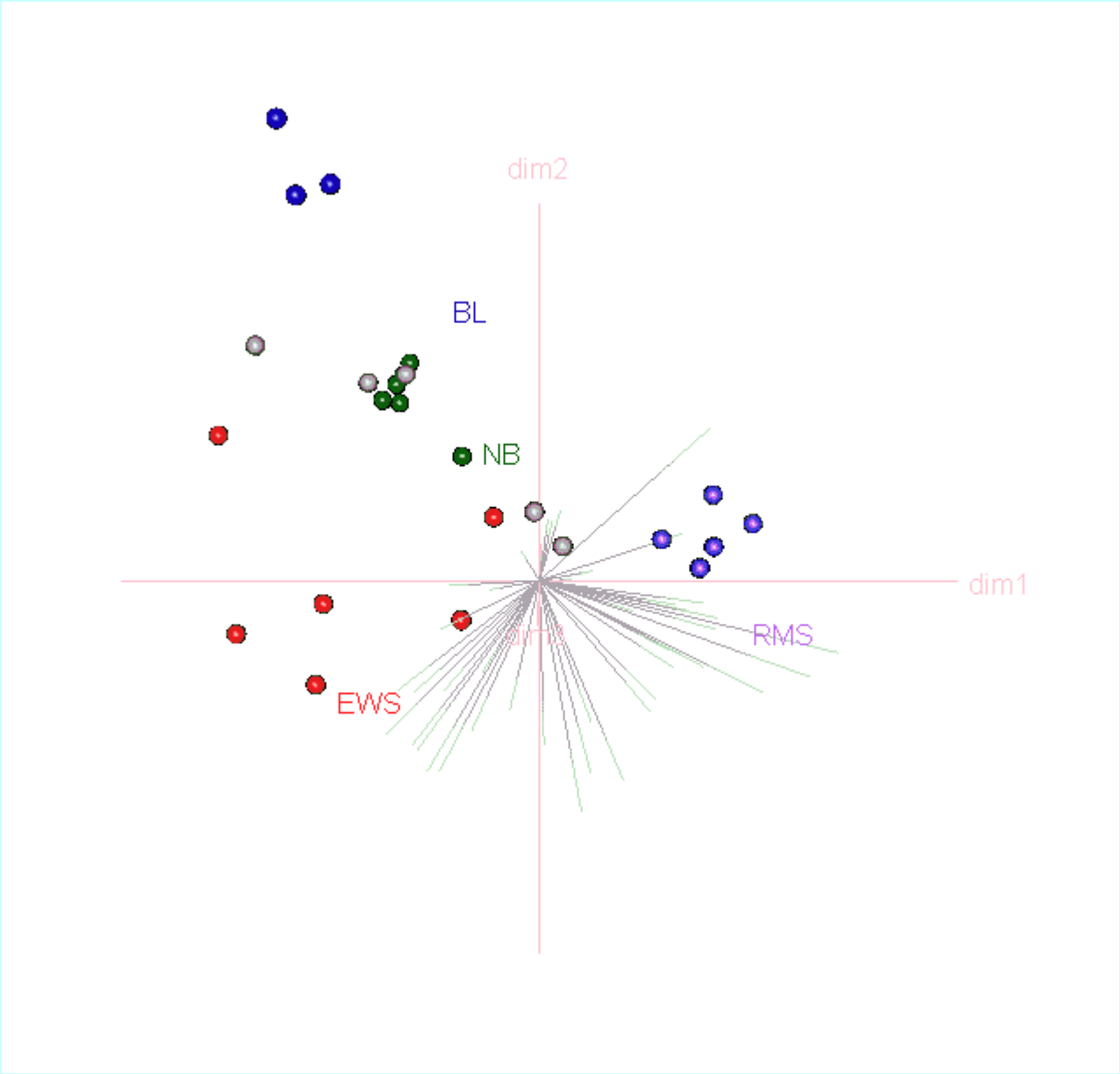


This frame is at approximate **0.4 : 0.6** *distribution of weight between tumors and tumor centroids* (β = 0.6).

Peeling away genes with least predictive power, seeing effect on prediction of test data.



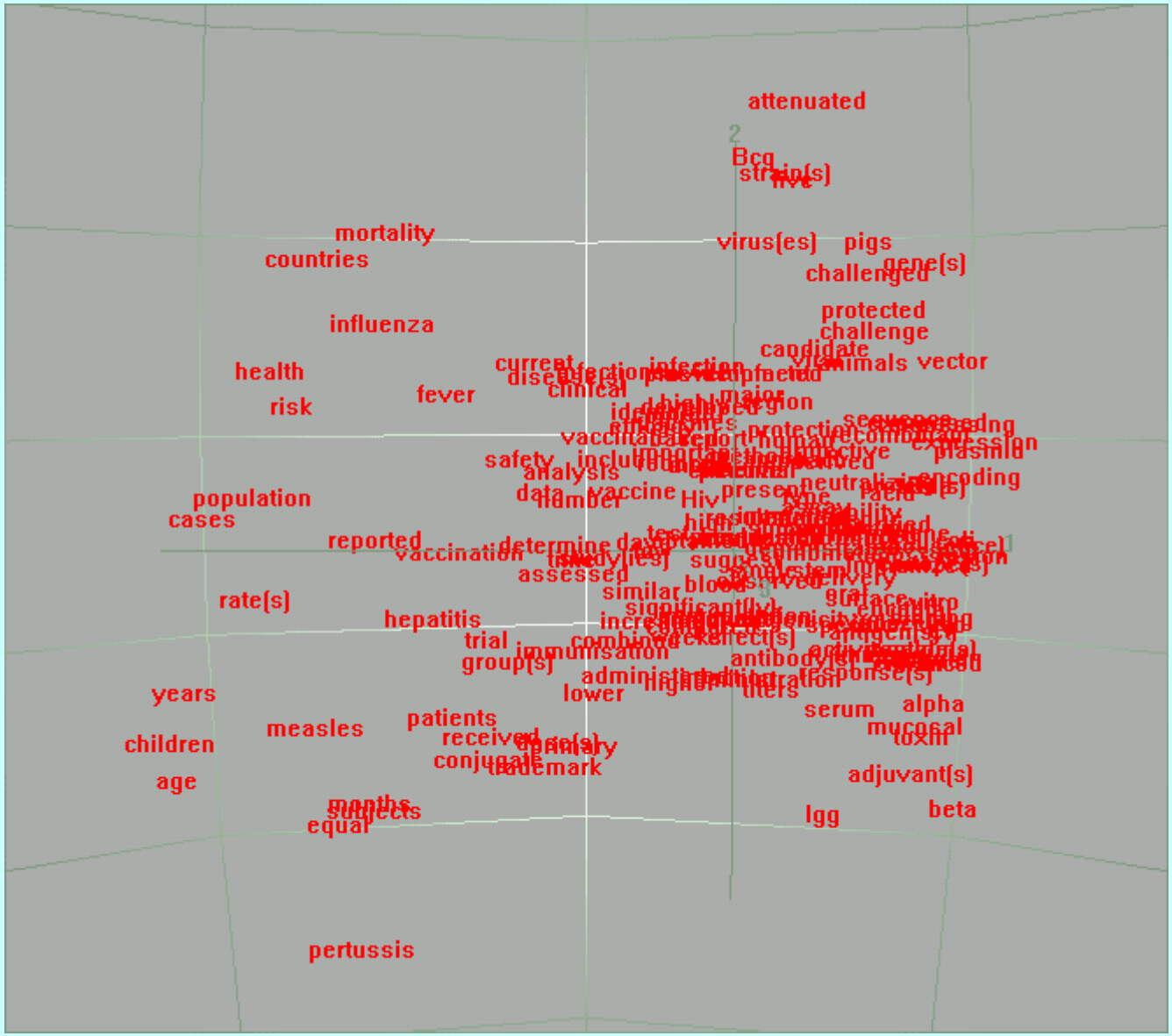
Rotation of 3-dimensional solution space where prediction takes place; test samples are classified to the closest centroid of the training set.



Mining a corpus of text

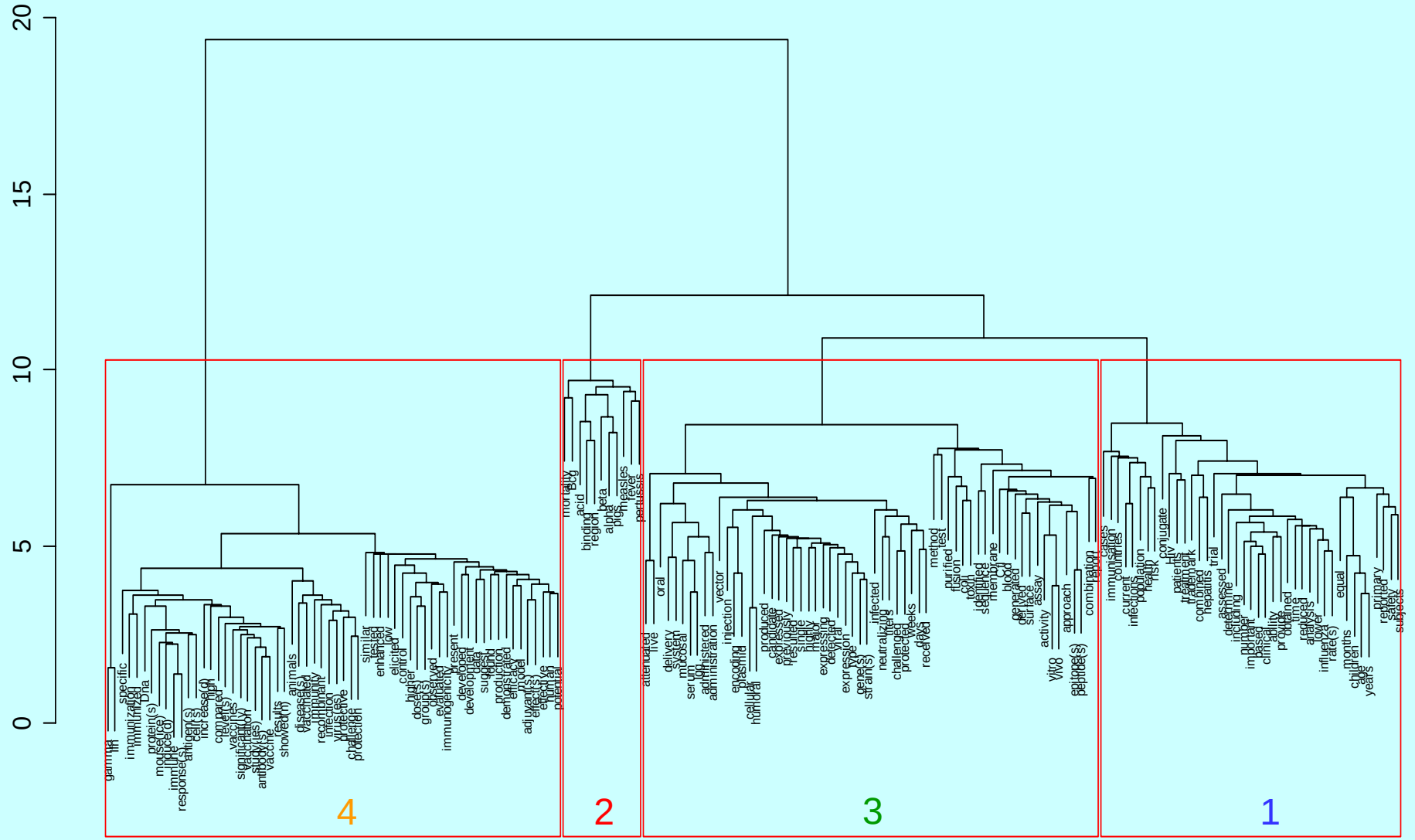
- all 2735 abstracts from articles published in Vaccine in the period 2003–2006.
- 178 most frequent words used (words are the “genes” here)
- data (“word expression”) are the counts of each word in each abstract
- of the 481,360 elements in this 2735x178 matrix, 429,419 elements are 0 (only 11.8% of the table contains some positive count)
- correspondence analysis is perfectly adapted as a dimension-reduction technique for this type of data – see Greenacre, M. (2007), *Correspondence Analysis in Practice. 2nd edition.*
- we will first show an unsupervised analysis, looking at the co-occurrences of the words and then a supervised analysis, using citation count as an outcome variable.

Three-dimensional representation of the 176 most frequent words in the abstracts of the journal Vaccine (2003–2006). The closer words are, the more they co-occur in abstracts.



Another unsupervised learning algorithm: cluster analysis

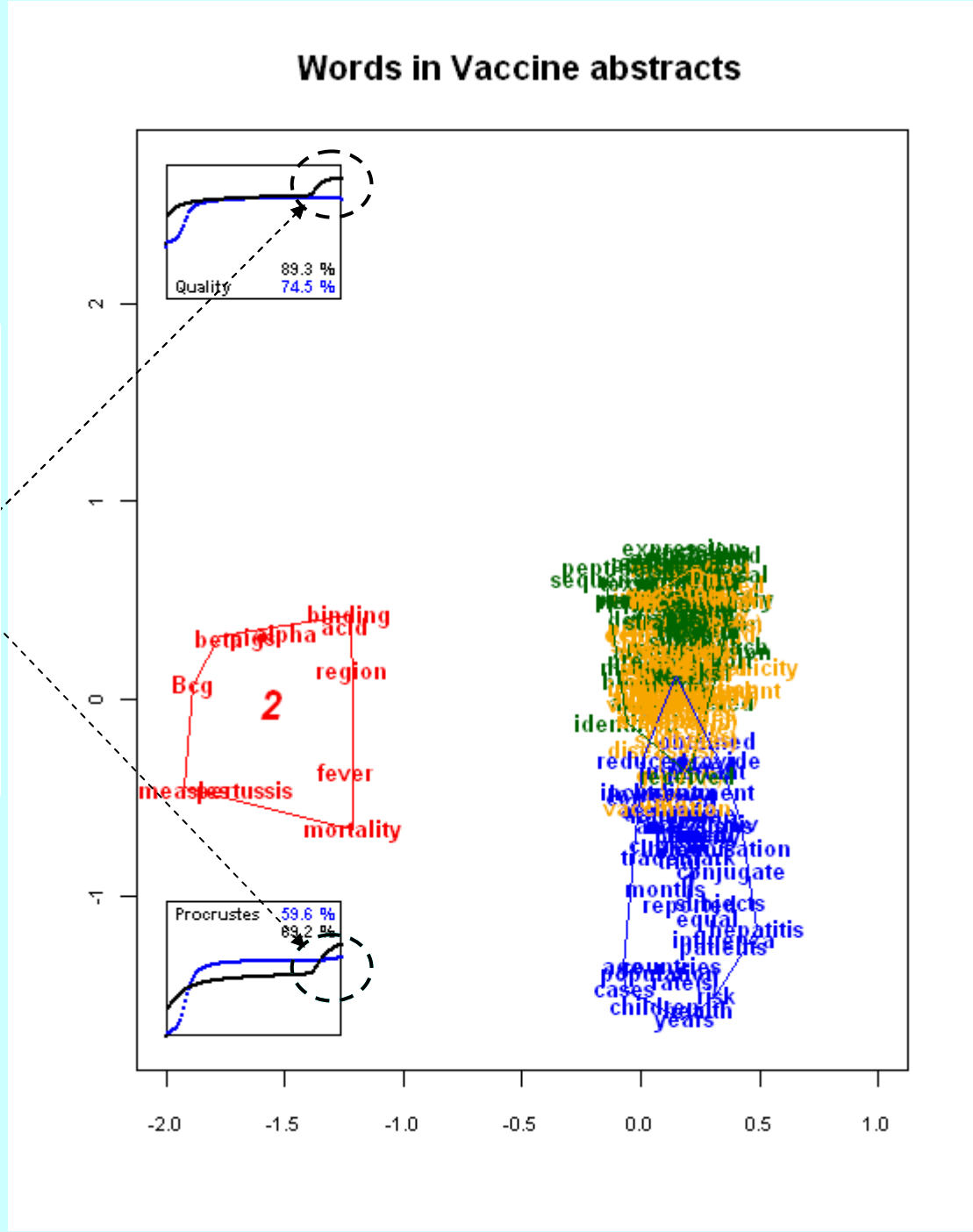
– 4 clusters of words identified

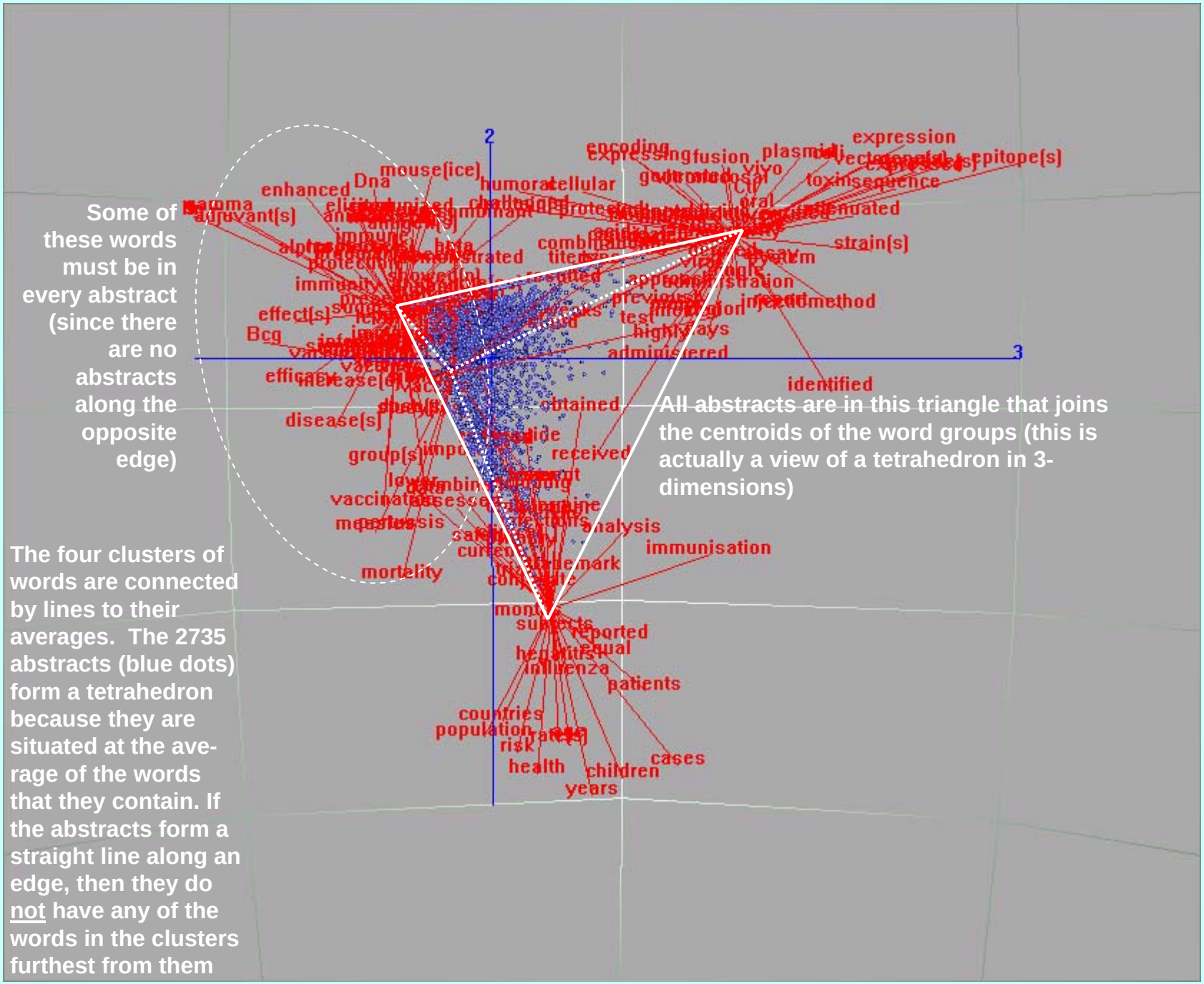


Tour through 176-dimensional space to find the four clusters of words

Final frame of the animation

something is happening at the end of the tour on the third dimension (remember that the blue curve refers to the two-dimensional display we see, while the black curve refers to the three-dimensional display)





Some of these words must be in every abstract (since there are no abstracts along the opposite edge)

All abstracts are in this triangle that joins the centroids of the word groups (this is actually a view of a tetrahedron in 3-dimensions)

The four clusters of words are connected by lines to their averages. The 2735 abstracts (blue dots) form a tetrahedron because they are situated at the average of the words that they contain. If the abstracts form a straight line along an edge, then they do <u>not</u> have any of the words in the clusters furthest from them

Supervised learning on the Vaccine abstracts: can citation count be predicted?

For supervised learning we need an outcome variable.

Elsevier supplied to use the citation counts for each of the 2735 articles in *Vaccine*.

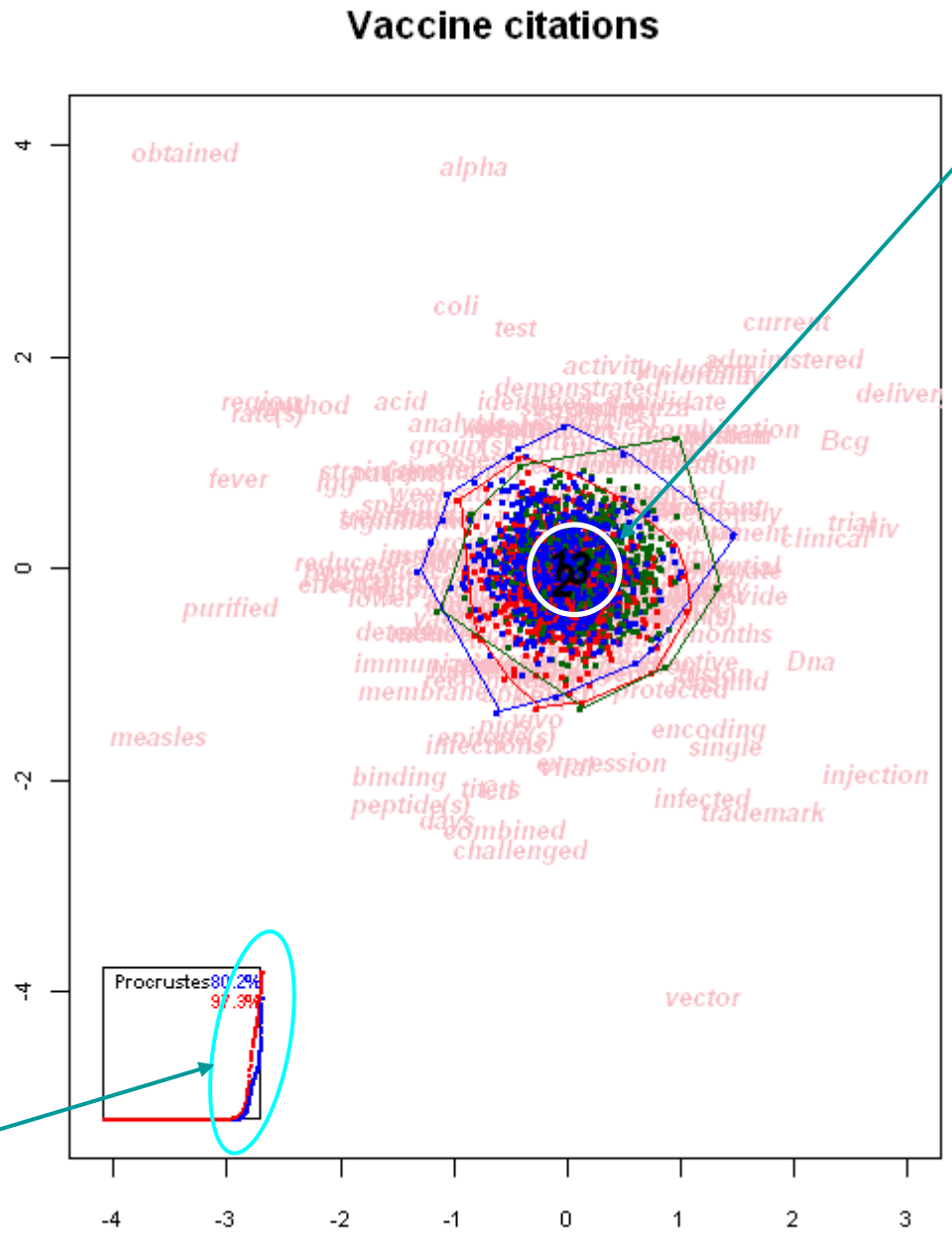
Since the articles are published in different years, we adjusted the overall count according to publication period, so that citations are effectively per unit of time (three-month unit).

The citation count was then classified into one of three groups: low, medium and high citation count. This is the classification we will try to predict by trying to separate the articles with a view to predicting this classification.

We shall tour 178-dimensional space of the words to try to learn the prediction rule. As shown in the next and final animation, we can look all we like in this space, and we will not find a rule that successfully predicts citation count!

Attempting to separate the centroids of the three citation groups in 176-dimensional word space. The black numbers (*1*, *2* and *3*) at the centre are the centroids of the citation groups LOW (blue points), MEDIUM (red) and HIGH (green), respectively.

shows that all the change is taking place right at the end of the search, buried in the last dimensions of the 176-dimensional word space



Slight separation of citation centroids at end of tour, with highest citation group (3) to the right, associated with words to the right:

For example, "delivery" is used on average 66% more in highly cited articles than in low & medium cited ones; "clinical" 46% more, "trial" 52% more, "Hiv" 57% more, "Bcg" 52% more, and "Dna" 40% more. Based on a permutation test, this group of words does occur significantly more often in highly cited papers (P=0.008).

But these differences are <u>not</u> sufficient to give a useful prediction rule, as shown by the poor separation of the three groups. This is a good example of statistical significance not necessarily implying substantive worth.

Tools and technology

Animations created using open source statistical package **R**: www.R-project.org

Frames that were computed in **R** then packaged into animated GIF files using *Animation Shop* : www.corel.com. It is also possible to animate the frames using *Adobe Fireworks*, although the resulting files are far larger: www.adobe.com

R has all the tools for creating a single static image, so what we need to do is to extend these tools to creating a sequence of images.

Creating a tool that is easy to use for life scientists requires the writing of **R** functions that take the major burden off the user, and eventually combining these functions in an R **package** for free distribution on the Internet.

As a prototype we have written an R function which implements the dynamic visualization used the most in this project. The function is called **`tourgroup`** and requires as arguments:

- **the data matrix**
- **an indication whether groups are defined on rows or columns**
- **the groups to which each row (or column) is assigned**
- **the number of frames to calculate**
- **various graphical parameters such as size, font sizes, colours, labels, etc…**

INSTALLATION AND TESTING OF THE TOOL

1. Download the R package for free from www.R-project.org; install the package and start R.
2. Using the pull-down menu **`File>Load workspace`** load the workspace **`Elsetwee.RData`** provided in this portfolio – this will load the two objects **`gene`** (with the data) and **`group`** (with the tumor classes).
3. Copy the function **`tourgroup`** from the script file **`tourgroup.txt`** provided in this portfolio, and paste it into the R command window.
4. Call the function as follows:

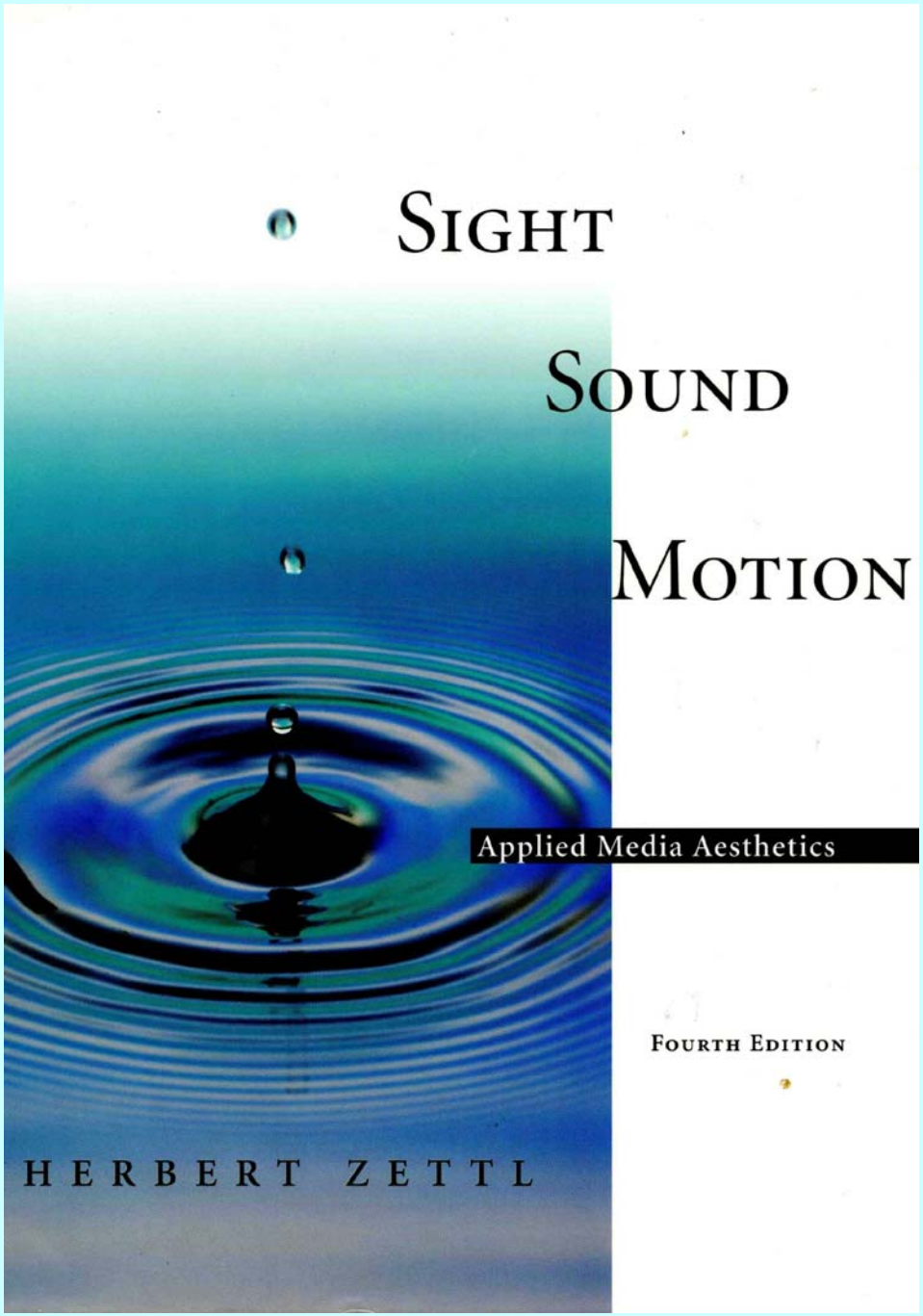
```
tourgroup(genes, group,
+    colortable=c("blue","red","darkgreen","purple"),
+    which="cols", rowscale=300)
```

(this will plot the dynamic graphic in an R graphics window)

5. If you want to save the frames for animating, include the option **`plotpng=T`** in the function call. A directory **`_png_out`** will be created in your working directory, and 101 png files will be created (the default number of frames is 101).

SUMMARY

- **In this project we propose the introduction of dynamic graphics in online publishing, to facilitate the interpretation of data and the understanding of data analysis.**
- **So far we have seen no significant movement by publishers in this direction, nor a trend in the scientific community towards the use of motion in published graphics.**
- **As we have shown, dynamic graphics can be used to animate simple data structures already commonly found in the literature, as well as complex ones that are the subject of present-day research – we believe that understanding is considerably enhanced by introducing motion into scientific graphics.**
- **This project is perfectly feasible technically – the technology is available, bandwidth is increasing daily – we just need more tools to make the creation of such dynamic graphics accessible to researchers**
- **In the case of the life sciences, a concerted didactic effort is needed to win over researchers to the benefits of innovative data visualization.**



…moulding an idea "*so that it fits the medium's technical as well as aesthetic production and reception requirements*." Applied media aesthetics "*places great importance on the influence of the medium on the message – the medium itself acts as an integral structural agent*."

- **1-d field of light & color**
- **2-d field of area**
- **3-d field of space**
- **4-d field of space/time**
- **5-d field of space/time/sound**

Annotations that we had to place on our written presentation could be replaced by audio commentary.

We are more or less stuck in the 2-d field in scientific publishing – when are we going to move towards the 5-d field?