Graphical Data and Data Graphics

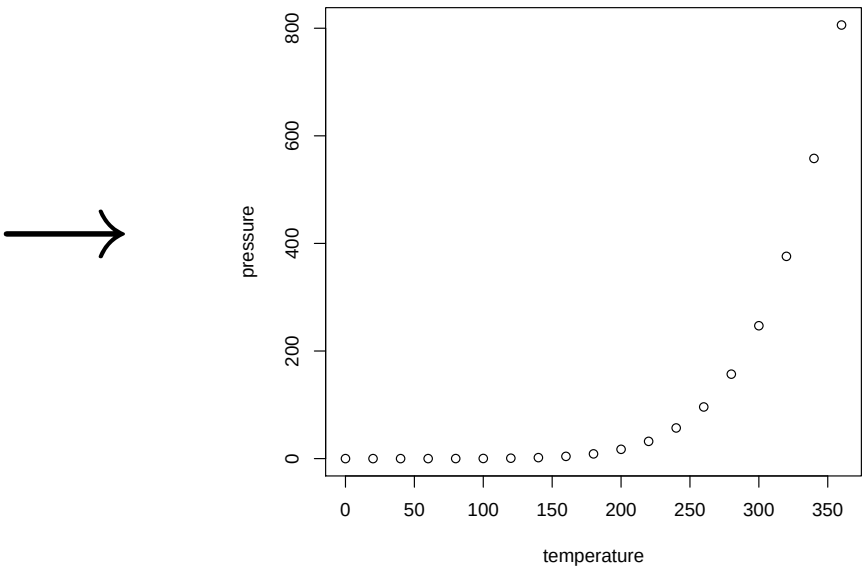
Paul Murrell

The University of Auckland

July 12 2007

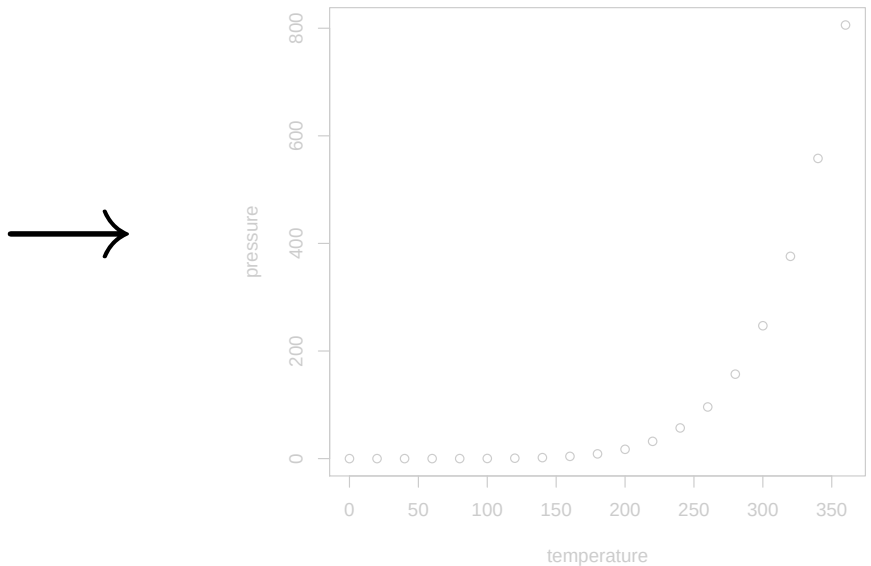
Graphical Statistics

```
> pressure
    temperature pressure
1             0   0.0002
2            20   0.0012
3            40   0.0060
4            60   0.0300
5            80   0.0900
6           100   0.2700
7           120   0.7500
8           140   1.8500
...
```



Statistical Graphics

```
> pressure
   temperature pressure
1            0   0.0002
2           20   0.0012
3           40   0.0060
4           60   0.0300
5           80   0.0900
6          100   0.2700
7          120   0.7500
8          140   1.8500
...
```



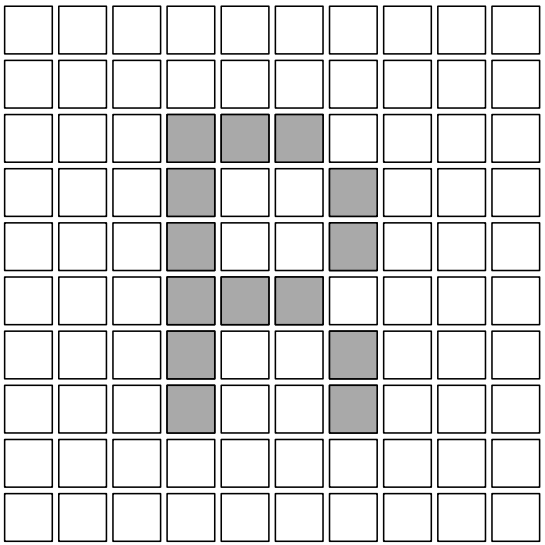
Graphical Data and Data Graphics

- Graphical Statistics: *data* $\longrightarrow$ *plot*
- Statistical Graphics: *data* $\longrightarrow$ *plot*

- Graphical Data: *plot* $\longrightarrow$ *data*
- Data Graphics: *plot* $\longrightarrow$ *data*

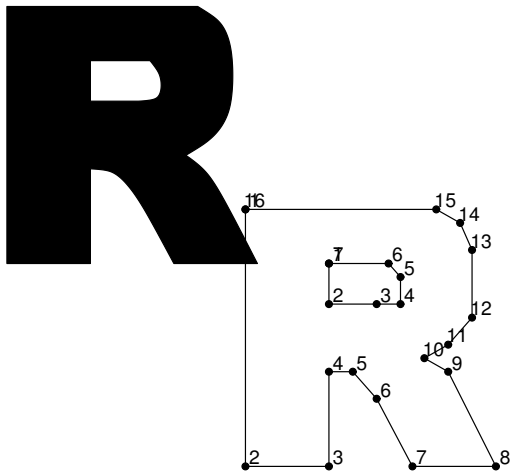
Graphical Formats

Raster



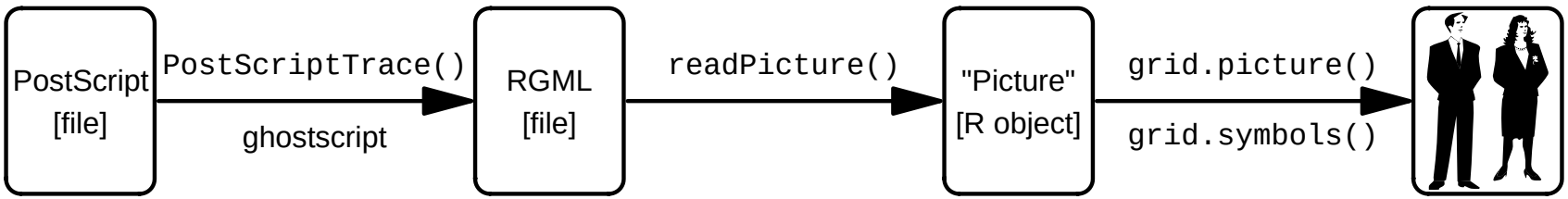
pixmap package
EBimage package

Vector



grImport package

The grImport Package



The PostScript Bezier Tiger

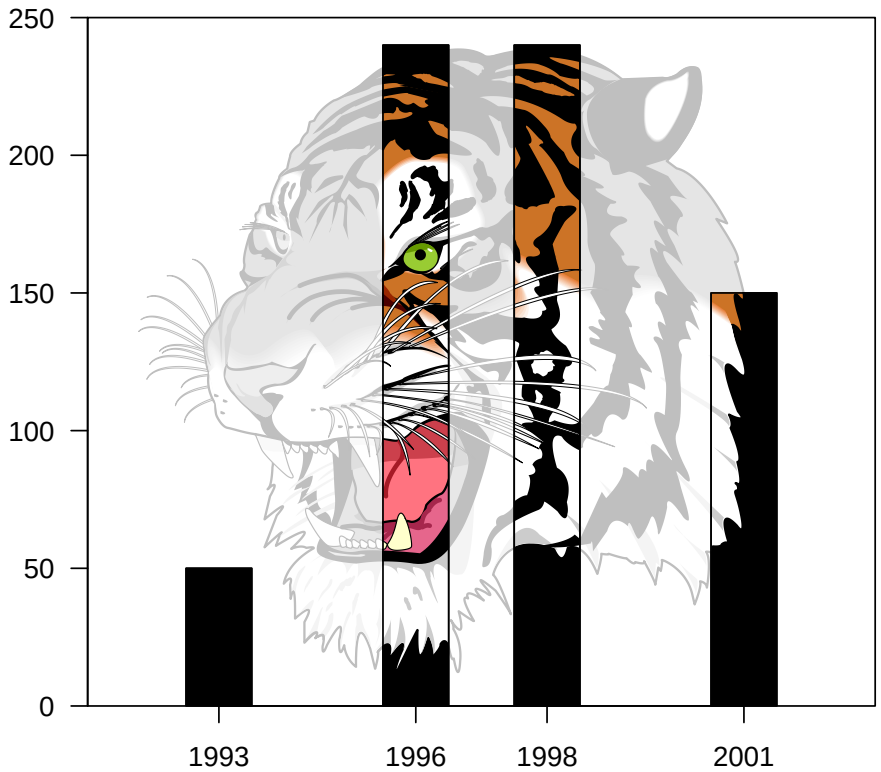
```
%!PS-Adobe-2.0 EPSF-1.2
%%Creator: Adobe Illustrator(TM)
%%For: OpenWindows Version 2
%%Title: tiger.eps
...
.8 setgray
clippath fill
-110 -300 translate
1.1 dup scale

0 g
0 G
0 i
0 J
0 j
0.172 w
10 M
[]0 d
0 0 0 0 k
...
```

Converting the Tiger to Data

```
PostScriptTrace("tiger.ps")

tiger <-
  readPicture("tiger.ps.xml")
```

Using the Tiger in a Plot

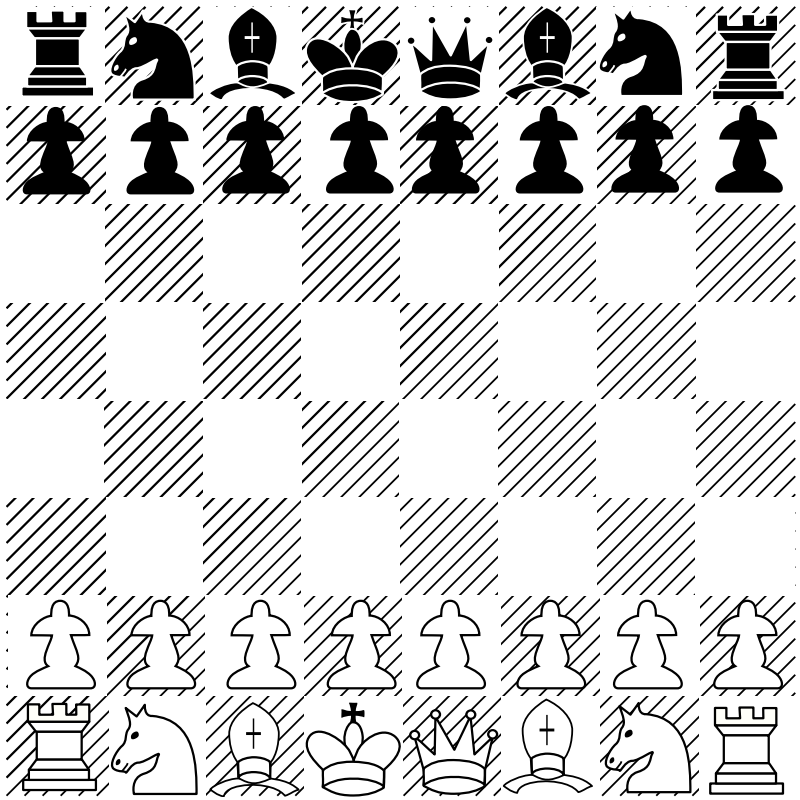
```
grid.picture(tiger)
```



A Chess Board

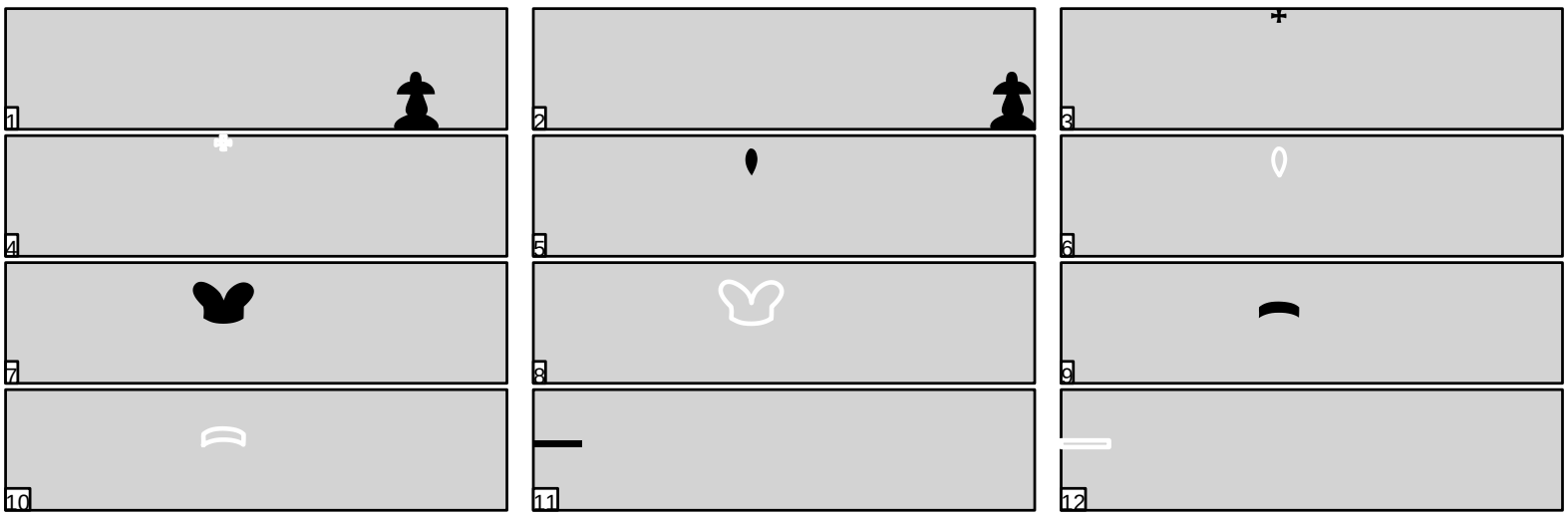
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG"
"http://www.w3.org/TR/2001/REC-SVG...">
<!-- Created with Sodipodi -->
<svg version="1.0">
...
  <g
     style="font-size:12;"
     id="g874">
    <path
       d="M 0 437 L 437 0 "
       style="fill:none;fill-opacity:1"
       id="path616" />
...
```

```
# Convert SVG to PostScript
# using InkScape

PostScriptTrace("chess.ps")

chess <-
  readPicture("chess.ps.xml")
```



The Paths in the Chess Board

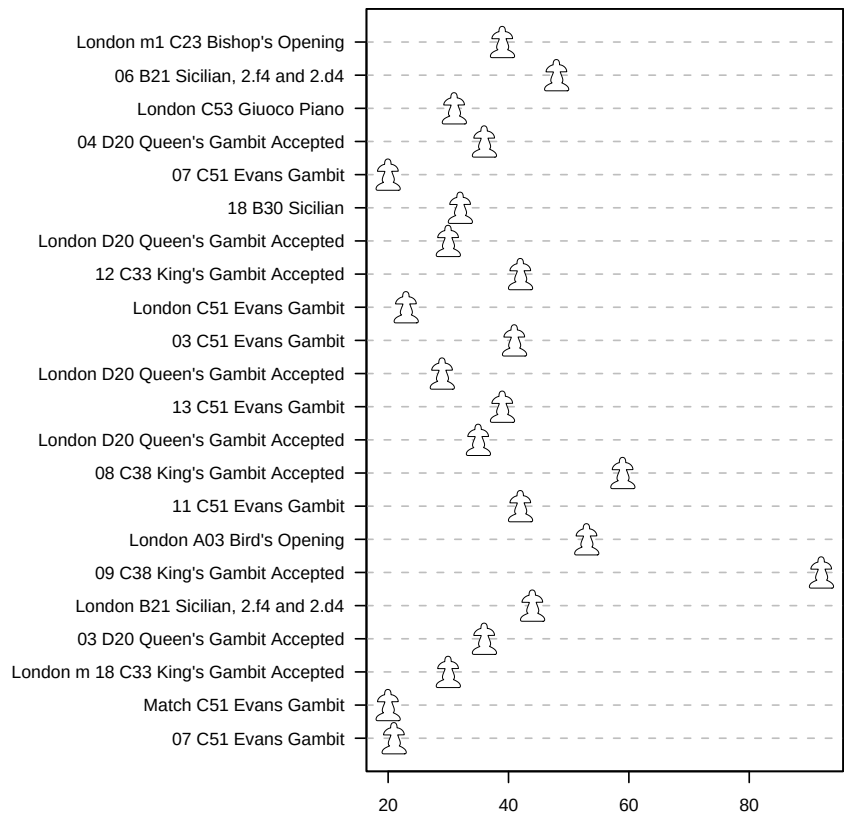
```
picturePaths(chess[125:136])
```



A Chess Piece as a Plotting Symbols

The number of moves required to complete chess games for different opening gambits. From the career of Louis Charles Mahe De La Bourdonnais (circa 1830).

```
grid.symbols(
  chess[205:206],
  x=games$num.moves,
  y=1:ngames,
  "native",
  size=unit(0.5, "cm"))
```



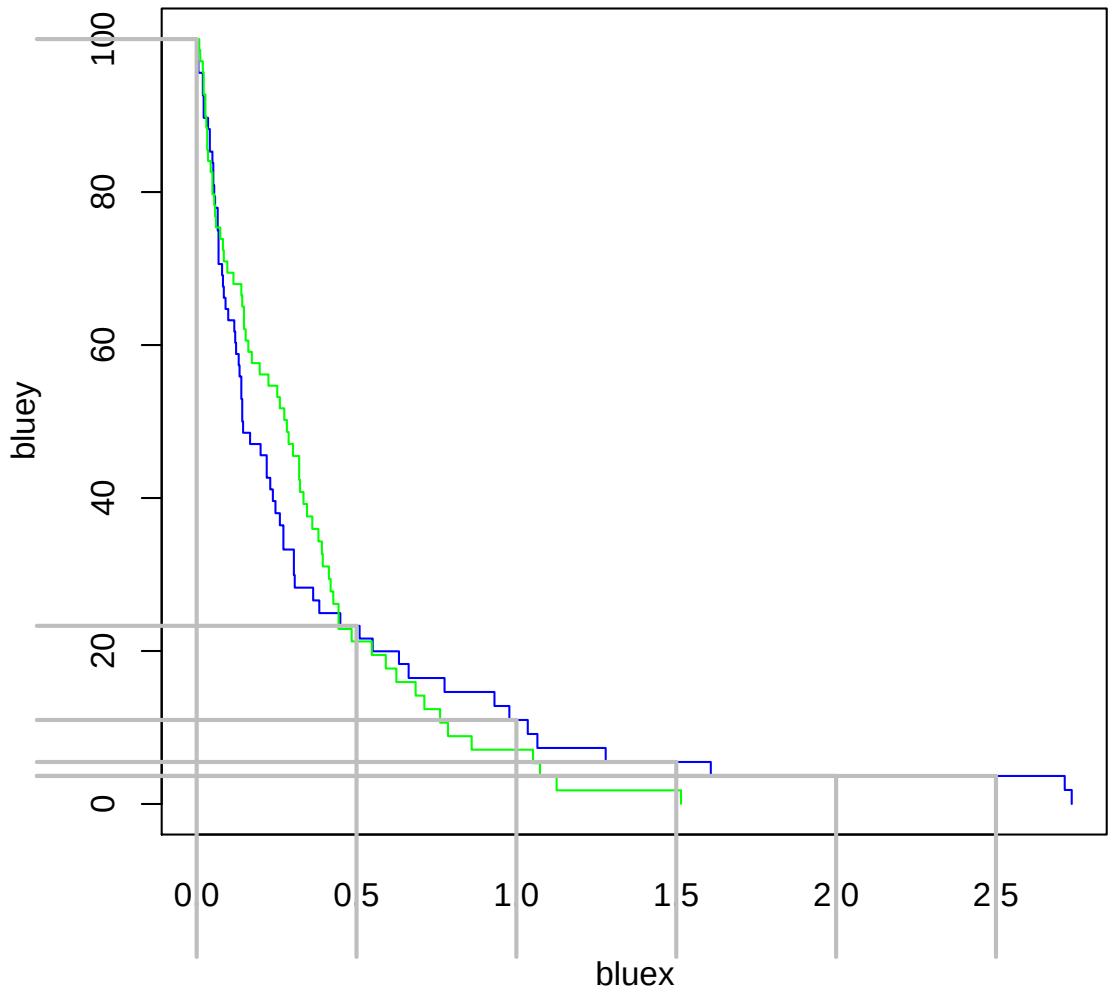
Statistical Data Graphics

- Graphical Statistics: *data* $\longrightarrow$ *plot*
- Statistical Graphics: *data* $\longrightarrow$ *plot*

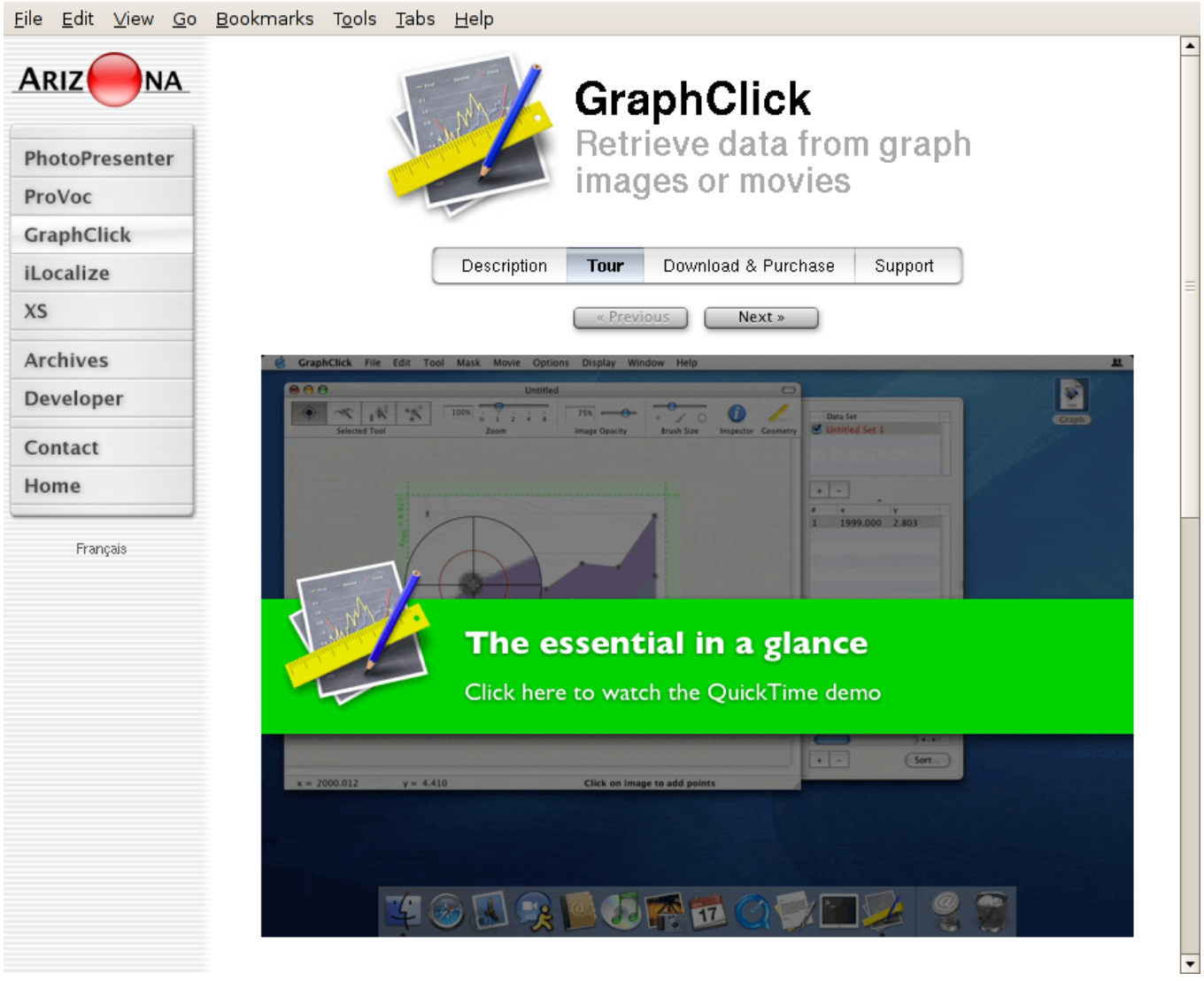
- Graphical Data: *plot* $\longrightarrow$ *data*
- Data Graphics: *plot* $\longrightarrow$ *data*

- Statistical Data Graphics: *data* $\longrightarrow$ *plot* $\longrightarrow$ *data*

Capturing Data By Hand



Capturing Data By Hand



A Published Plot

R News

The Newsletter of the R Project Volume 4/1, June 2004

Editorial

by Thomas Lumley

R has been well accepted in the academic world for some time, but this issue of R News begins with a description of the use of R in a business setting. Marc Schwartz describes how his consulting company, MedAnalytics, came to use R instead of commercial statistical software, and has since provided both code and financial support to the R Project.

The growth of the R community was also evident at the very successful useR! 2004, the first conference for R users. While R aims to blur the distinctions between users and programmers, useR! was definitely different in aims and audience from the previous DSC conferences held in 1999, 2001, and 2003. John Fox provides a summary of the conference in this issue for those who could not make it to Vienna.

A free registration to useR! was the prize in the competition for a new graphic for http://www.r-project.org. You will have seen the winning graphic, from Eric Lecoutre, on the way to downloading this newsletter. The graphic is unedited R output (click on it to see the code), showing the power of R for creating complex graphs that are completely reproducible. I would also encourage R users to visit the Journal of Statistical Software (http:///www.jstatsoft.org), which publishes peer-reviewed code and documentation. R is heavily represented in recent issues of the journal.

The success of R in penetrating so many statistical niches is due in large part to the wide variety of packages produced by the R community. We have articles describing contributed packages for principal component analysis (ade4, used in creating Eric Lecoutre's winning graphic) and statistical process control (qcc) and the recommended survival package. Jianhua Zhang and Robert Gentlement describe tools from the Bioconductor project that make it easier to explore the resulting variety of code and documentation.

The Programmers' Niche and Help Desk columns both come from guest contributors this time. Gabor Grothendieck,known as frequent poster of answers on the mailing list r-help, has been invited to contribute to the R Help Desk. He presents "Date and Time Classes in R". For the Programmers' Niche, I have written a simple example of classes and methods using the old and new class systems in R. This example arose from a discussion at an R programming course.

Programmers will also find useful Doug Bates' article on least squares calculations. Prompted by discussions on the r-devel list, he describes why the simple matrix formulae from so many textbooks are not the best approach either for speed or for accuracy.

Thomas Lumley
Department of Biostatistics
University of Washington, Seattle
thomas.lumley@R-project.org

Contents of this issue:

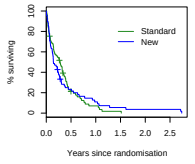


Figure 1: Survival distributions for two lung cancer treatments

Proportional hazards models

The mainstay of survival analysis in the medical world is the Cox proportional hazards model and its extensions. This expresses the hazard (or rate) of events as an unspecified baseline hazard function multiplied by a function of the predictor variables.

Writing $h(t;z)$ for the hazard at time t with predictor variables $Z = z$ the Cox model specifies

$$\log h(t,z) = \log h_0(t) e^{\beta z}.$$

Somewhat unusually for a semiparametric model, there is very little loss of efficiency by leaving $h_0(t)$ unspecified, and computation is, if anything, easier than for parametric models.

A standard example of the Cox model is one constructed at the Mayo Clinic to predict survival in patients with primary biliary cirrhosis, a rare liver disease. This disease is now treated by liver transplantation, but at the same there was no effective treatment. The model is based on data from 312 patients in a randomised trial.

```
> data(pbc)
> mayomodel<-coxph(Surv(time,status)~edtrt+
                     log(bili)+log(protime)+
                     age+platelet,
            data=pbc, subset=trt>0)
> mayomodel
Call:
coxph(formula = Surv(time, status) ~ edtrt +
    log(bili) + log(protime) +
    age + platelet, data = pbc,
    subset = trt > 0)


               coef exp(coef)
edtrt       1.02980     2.800
log(bili)   0.95100     2.588
log(protime) 2.88544    17.911
age          0.03544     1.036
platelet    -0.00128     0.999
            se(coef)     z       p
edtrt       0.300321  3.43 0.00061
log(bili)   0.097771  9.73 0.00000
log(protime) 1.031908  2.80 0.00520
age         0.008489  4.18 0.00003
platelet    0.000927 -1.38 0.17000

Likelihood ratio test=185  on 5 df, p=0  n= 312
```

The survexp function can be used to compare predictions from a proportional hazards model to actual survival. Here the comparison is for 106 patients who did not participate in the randomised trial. They are divided into two groups based on whether they had edema (fluid accumulation in tissues), an important risk factor.

```
> plot(survfit(Surv(time,status)~edtrt,
       data=pbc,subset=trt==-9))
> lines(survexp(~edtrt+
                ratetable(edtrt=edtrt,bili=bili,
                      platelet=platelet,age=age,
                      protime=protime),
                data=pbc,
                subset=trt==-9,
                ratetable=mayomodel,
                cohort=TRUE),
          col="purple")
```

The ratetable function in the model formula wraps the variables that are used to match the new sample to the old model.

Figure 2 shows the comparison of predicted survival (purple) and observed survival (black) in these 106 patients. The fit is quite good, especially as people who do and do not participate in a clinical trial are often quite different in many ways.

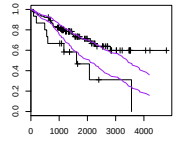
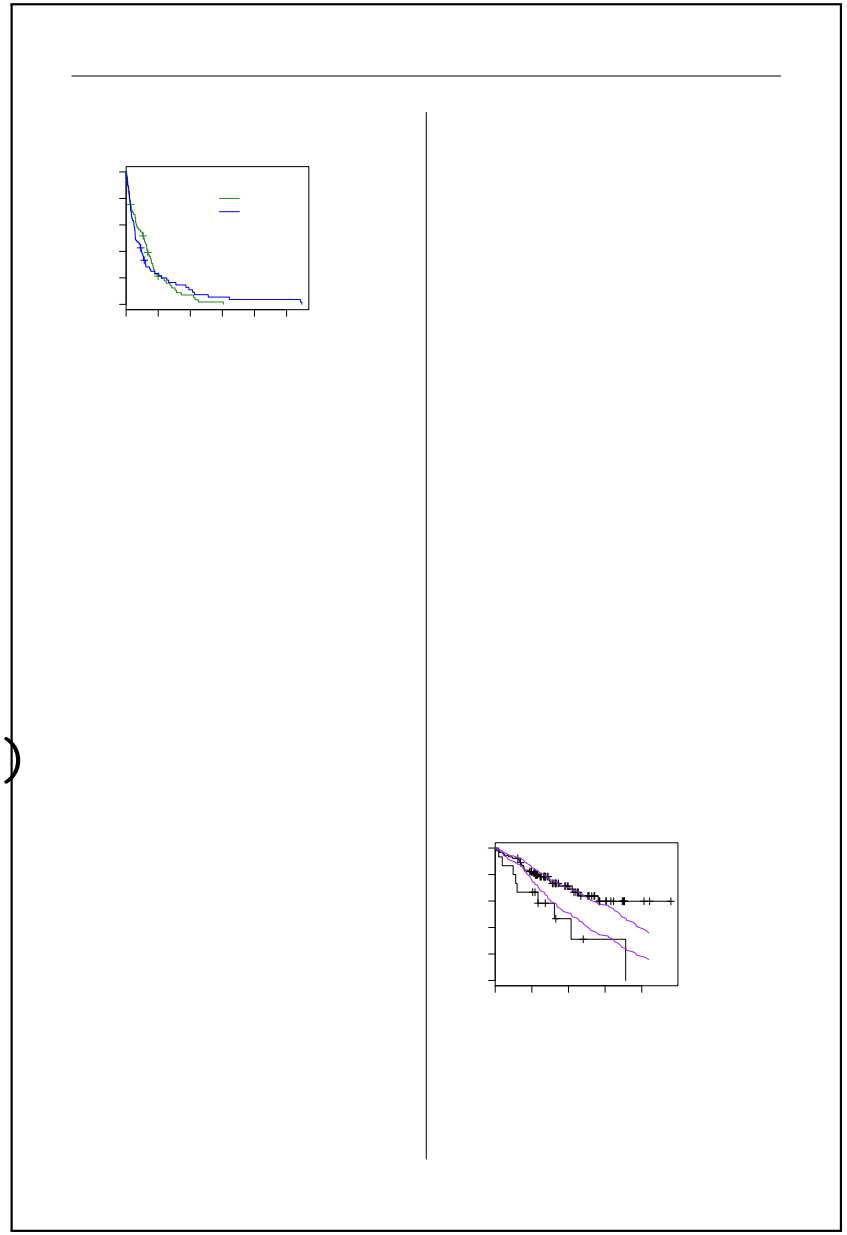


Figure 2: Observed and predicted survival

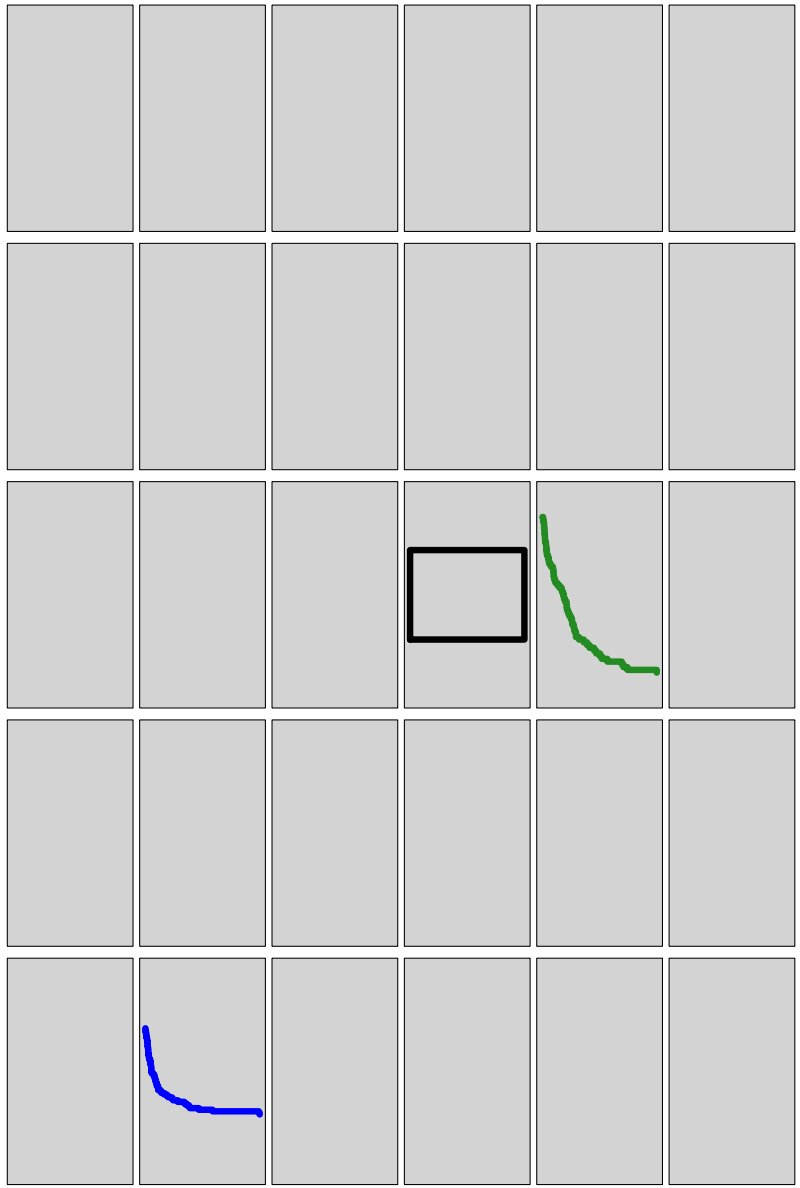
The main assumption of the proportional hazards model is that hazards for different groups are in fact proportional, *i.e.* that β is constant over time. The

 ISSN 1609-3631

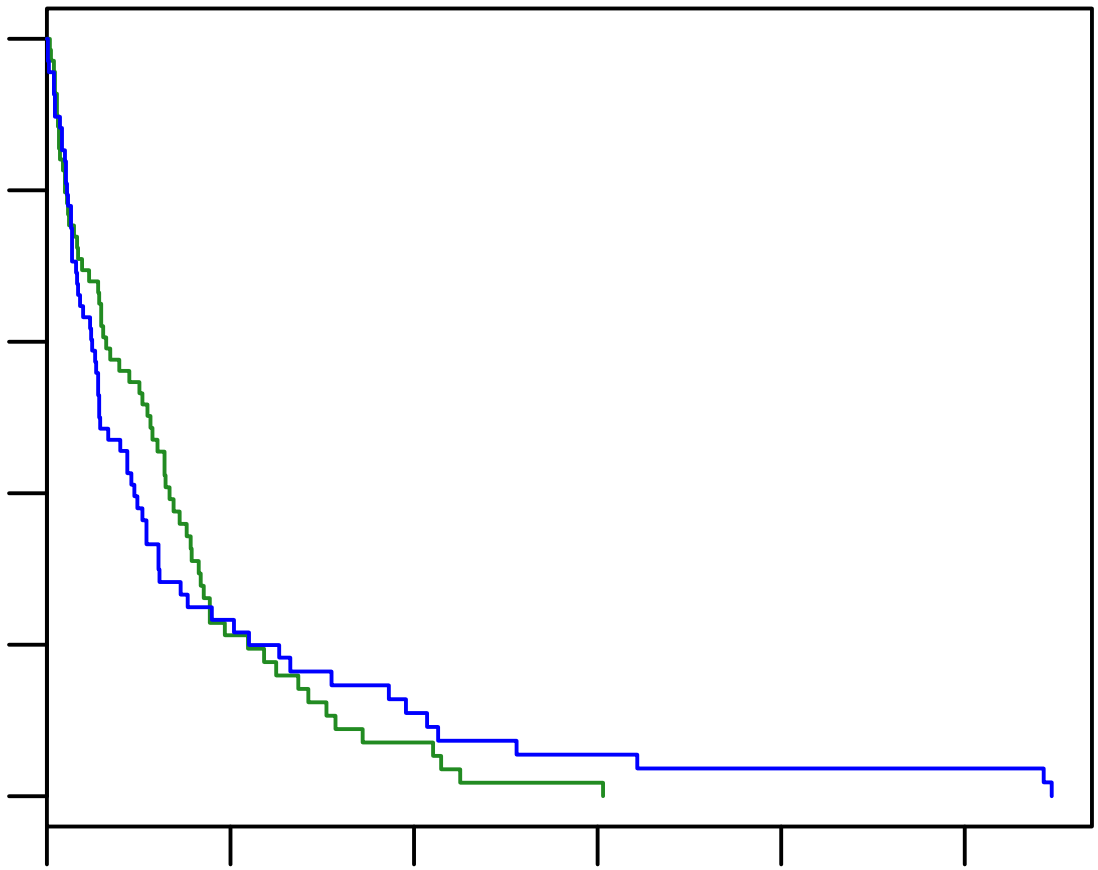
```
# Extract just page 27
# and convert to PostScript

PostScriptTrace("page27.ps")

page27 <-
 readPicture("page27.ps.xml")

grid.picture(page27)
```



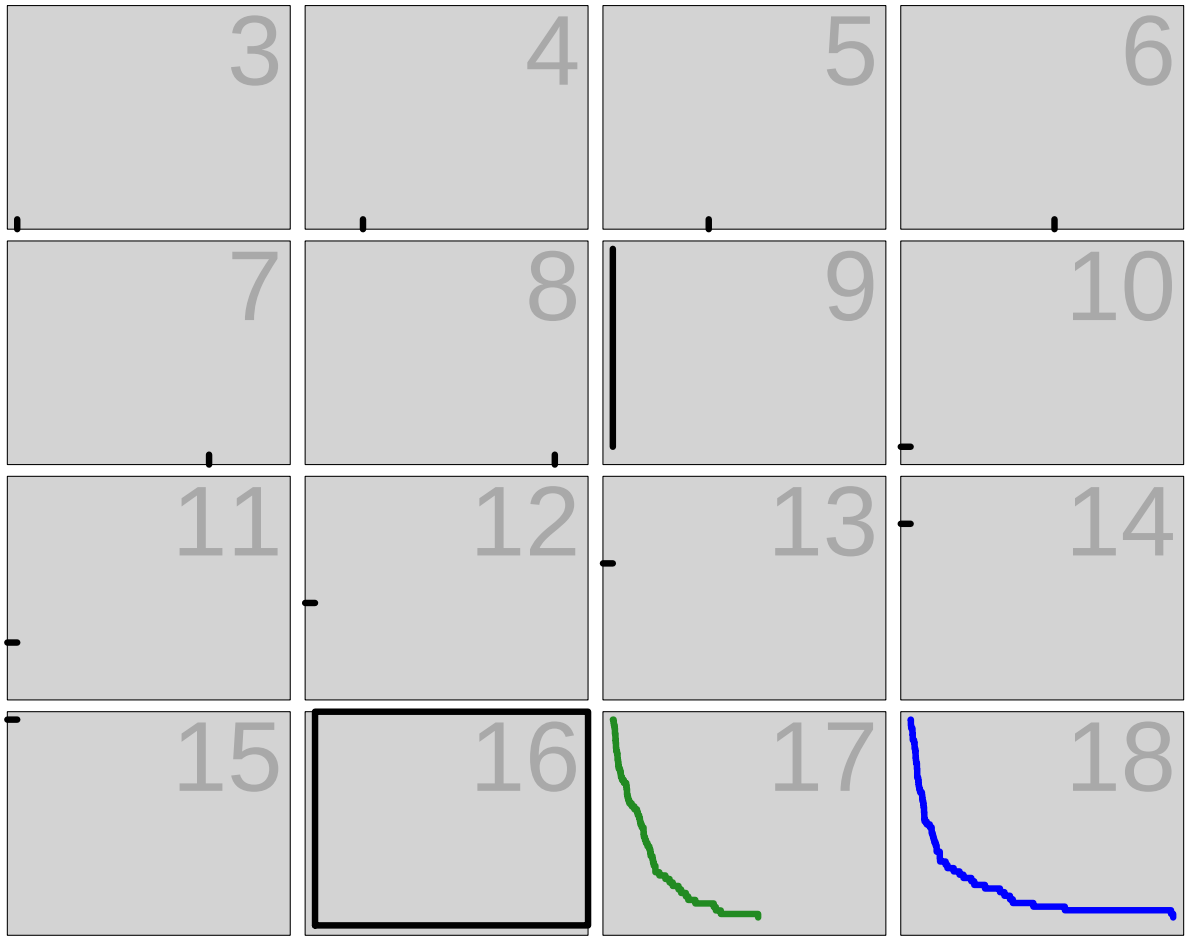
```
picturePaths(page27,
             freeScales=TRUE)
```



```
> grid.picture(page27[c(3:16, 17, 26)], gp=gpar(lex=.3))
```



```
> picturePaths(page27[c(3:16, 17, 26)])
```



```
> barePlot <- page27[c(3:16, 17, 26)]
> slotNames(barePlot)
```

```
[1] "paths"   "summary"
```

```
> barePlot@summary
```

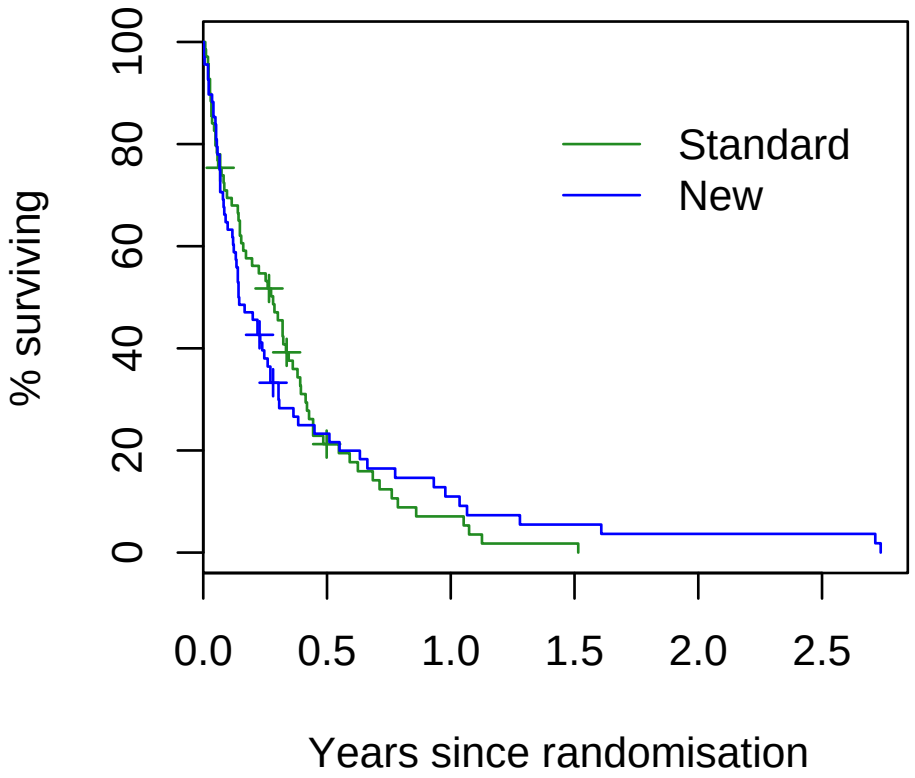
```
An object of class "PictureSummary"
Slot "numPaths":
[1] 16

Slot "xscale":
[1]  874 2161

Slot "yscale":
[1] 6228 7245
```

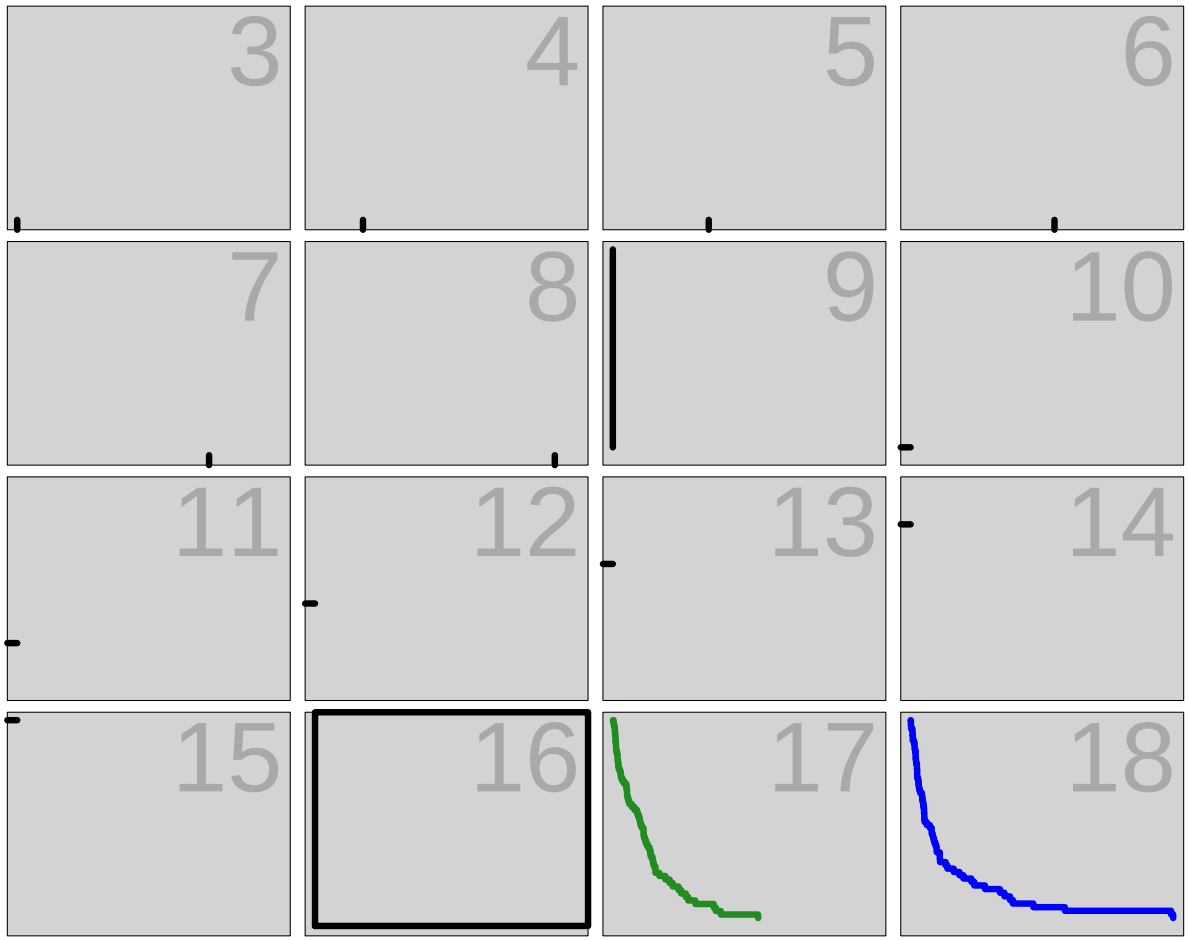
```
> class(barePlot@paths)
```

```
[1] "list"
```

```
> barePlot@paths[[1]]
```

```
An object of class "PictureStroke"
Slot "x":
move line
 919  919

Slot "y":
move line
6273 6228

Slot "rgb":
[1] "#000000"

Slot "lwd":
[1] 6.23
```


% surviving
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
Years since randomisation
Standard
New


3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

```
> xrange <- page27[16]@summary@xscale
```

```
[1]  919 2161
```

```
> yrange <- page27[16]@summary@yscale
```

```
[1] 6273 7245
```

```
> xunit <- (page27@paths[[5]]@x[1] - xrange[1])
```

```
[1] 436
```

```
> yunit <- (page27@paths[[15]]@y[1] -
            page27@paths[[10]]@y[1])/100
```

```
[1] 9
```

```
> greenx <- (page27@paths[[17]]@x -
             xrange[1]) / xunit
> greeny <- (page27@paths[[17]]@y -
             page27@paths[[10]]@y[1]) / yunit
```

```
      greenx greeny
[1,] 0.00000  100.0
[2,] 0.00816  100.0
[3,] 0.00816   98.5
[4,] 0.01102   98.5
[5,] 0.01102   97.1
[6,] 0.01918   97.1
```

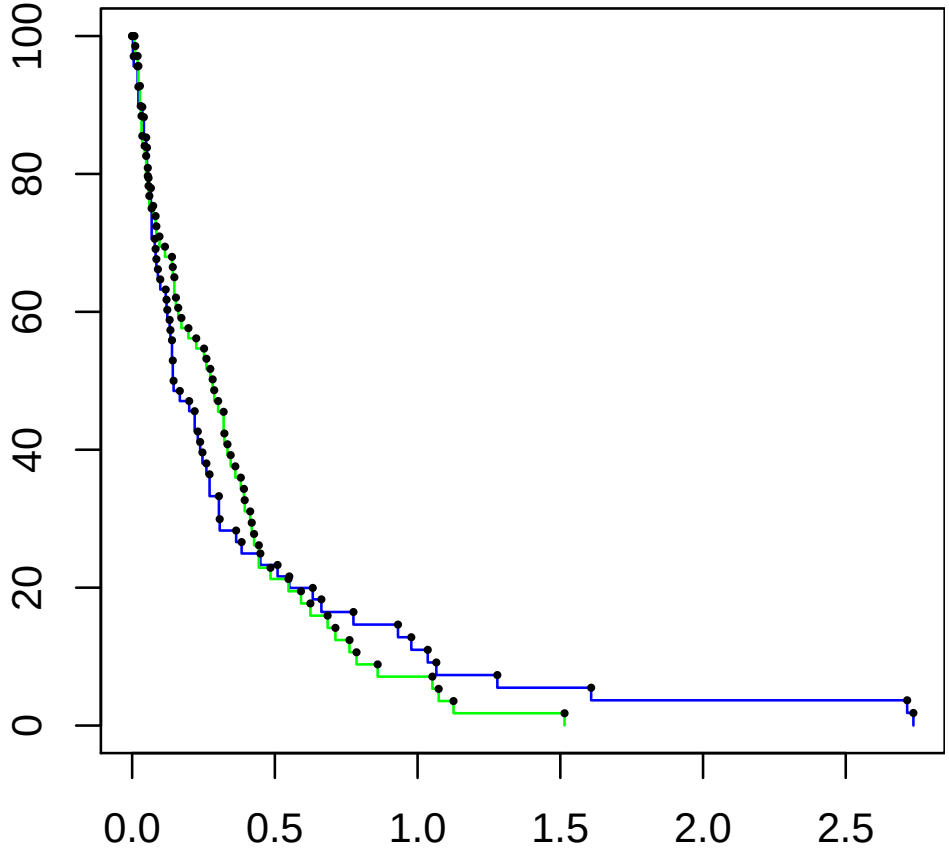
```
> bluex <- (page27@paths[[26]]@x -
            xrange[1]) / xunit
> bluey <- (page27@paths[[26]]@y -
            page27@paths[[10]]@y[1]) / yunit
```

```
> bluedupes <- duplicated(bluex)
> bluecurve <- data.frame(x=bluex,
                          y=bluey)[!bluedupes, ]
```

```
         x     y
1  0.00000 100.0
2  0.00270 100.0
4  0.00543  97.1
6  0.01918  95.6
8  0.02188  92.6
10 0.03563  89.7
```

```
> greendupes <- duplicated(greenx)
> greencurve <- data.frame(x=greenx,
                           y=greeny)[!greendupes, ]
```

```
plot(bluex, bluey,
     type="l",
     col="blue")
lines(greenx, greeny,
      type="l",
      col="green")
points(bluecurve,
       pch=16, cex=.3)
points(greencurve,
       pch=16, cex=.3)
```



Problems

- Many articles, especially old ones where the plot is the only data available, contain **bitmap** images.
- Many articles, and their images, are fiercely guarded by draconian copyright protections.
- Human intervention is still required.

Acknowledgements

- The tiger image is part of the **ghostscript** distribution; the tiger data are from `http://www.globaltiger.org/population.htm`.
- The greyscale version of the tiger used the **colorspace** package by **Ross Ihaka**.
- The chess board image (by **Jose Hevia**) is from the **Open Clip Art Library** `http://openclipart.org/clipart//recreation/games/chess/chess_game_01.svg`
- The chess data are from **chessgames.com** `http://www.chessgames.com/perl/chess.pl?page=1&pid=31596`
- **R News** is a publication of the **R Foundation for Statistical Computing**. The article used is **The survival Package** by **Thomas Lumley** R News, 4(1), pp. 26–28.
- The application for measuring survival curves was suggested by **Dan Jackson**.