Data Mining with R

John Maindonald (Centre for Mathematics and Its Applications, Australian National University)

and

Yihui Xie (School of Statistics, Renmin University of China)

December 13, 2008

Data Mining Motivations and Emphases

- "Big Data", the challenge of analyzing data sets of unprecedented size, perhaps collected automatically.
  - The term "Big Data" is mildly ironic, tilting a bit at the overblown use of this phrase in Weiss and Indurkhya (1998)[1].
- New types of data
  - Web pages, images
- New algorithms (analysis methods, models?)
  - Note especially trees and tree ensembles, (NB random forests of which many data miners seem unaware), boosting methods, support vector machines (SVMs), and neural nets.
- Automation.
- Machine Learning and Statistical Learning have somewhat similar motivations and emphases to data mining.

[1] Predictive Data Mining, Morgan Kaufmann 1997.

An Example – the Forensic Glass Dataset

We require a rule to predict the type of any new piece of glass.

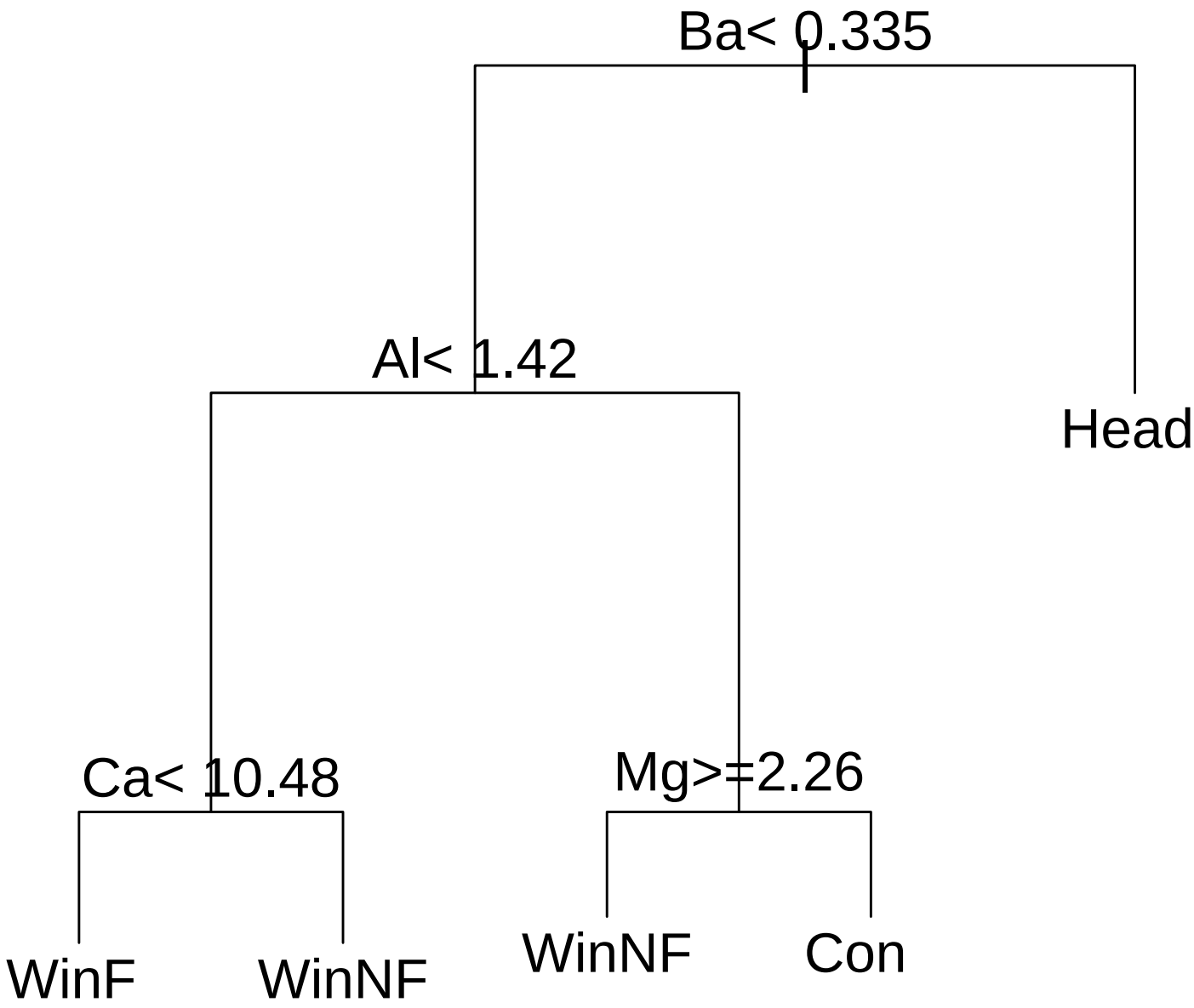
Type of glass	Short name (number of samples)
Window float	WinF (70)
Window non-float	WinNF (76)
Vehicle window	Veh (17)
Containers	Con (13)
Tableware	Tabl (9)
Headlamps	Head (29)

The data consist of 214 rows $\times$ 10 columns.

Variables are

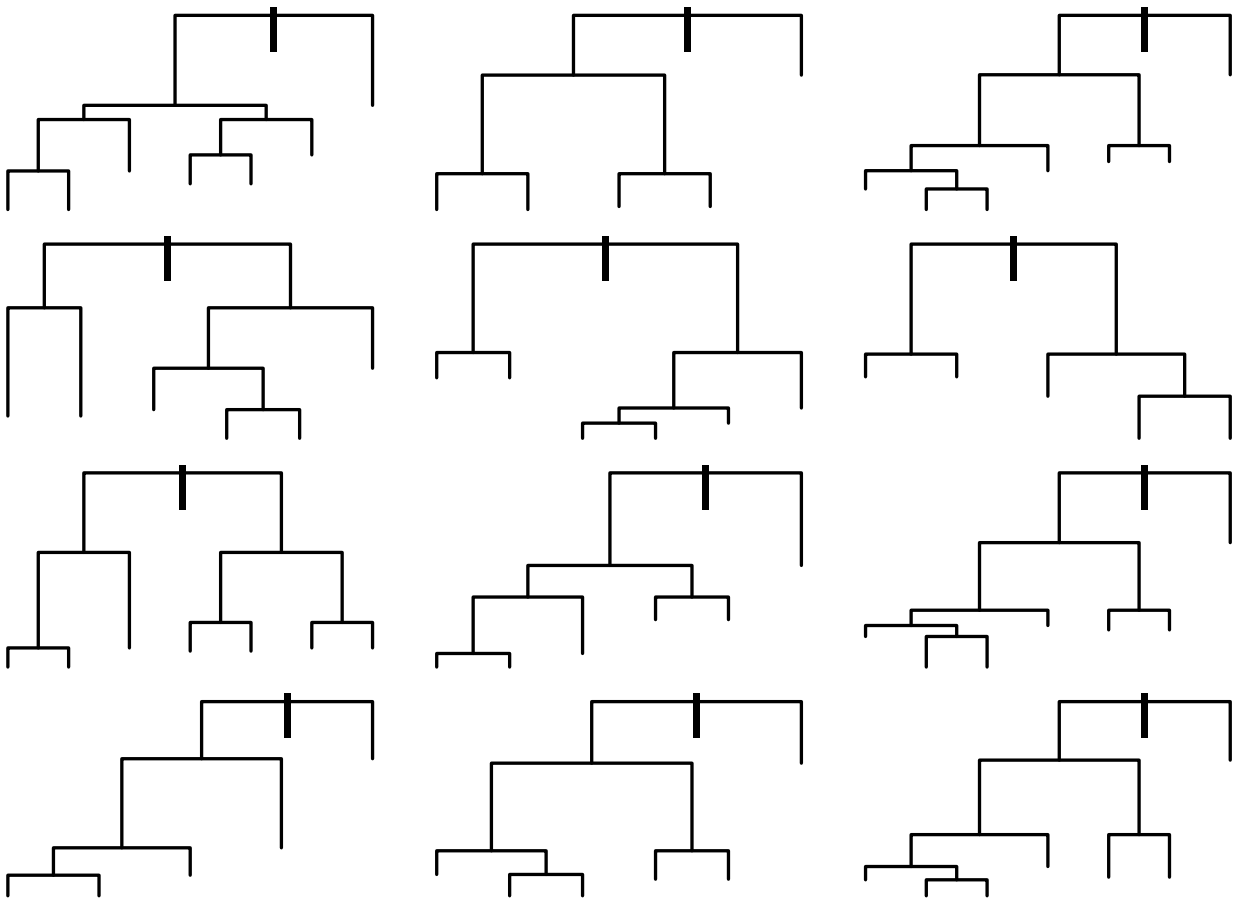
`RI` = Refractive index	`Na` = sodium (%)	`Mg` = manganese
`Al` = aluminium	`Si` = silicon	`K` = potassium
`Ca` = calcium	`Ba` = barium	`Fe` = iron
`type` = type of glass		

This tree is too simple to
give accurate predictions



Random forests – A Forest of Trees!

Each tree is for a different random with replacement sample of the data



Each tree has one vote; the majority wins

Data Mining in Practice

- The data mining tradition is recent, from computer science
- Classification and clustering are the most common problems.
- Favorite methods are trees and other new methods.
- Dependence in time or space is usually ignored.
- Prediction is usually the aim, not interpretation of model coefficients or other parameters
  - Where regression or other coefficients are of interest, data miners may be unaware of the traps.
- Often, there is extensive variable and/or model selection.
  - This brings a risk of finding spurious effects.

Ways to Think about Data Mining

- Data Mining is Exploratory Data Analysis with "muscle"? (Berk , 2006)
- Statistical Learning and Machine Learning are more theoretical versions of data mining?
- *Analytics* has become a popular name for applications in business and commerce.
- Several recent books have catchy invented titles, much like the names "Data Mining" and "Machine Learning"!
  - Ayres, I, 2006: Super Crunchers: Why Thinking-By-Numbers is the New Way to be Smart.
  - Baker, S, 2008: The Numerati.

Homework Exercise: Think of a new catchy title for a new book on data mining.

Data Mining and R

- The R project is the ideal platform for the analysis, graphics and software development activities of data miners and related areas
  - Weka, from the computer science community, is not in the same league as R.
  - Weka, and other such systems, quickly get incorporated into R!
- Note the *rattle* Graphical User Interface (GUI) for data mining applications. (developer: JM's colleague Graham Williams).

Common Methods for Assessing Accuracy

- Training/Test, with a random split of the available data
  - Do not use the test data for tuning or variable selection (this is a form of cheating!)
- Cross-validation – a clever use of the training/test idea
  - NB: Repeat tuning etc at each training/test split.
- Bootstrap approaches (built into random forests)
- Theoretical error estimates
  - Error estimates are rarely available that account for tuning and/or variable selection effects.

Punchlines

- Be clear how error estimates were obtained
- Give error estimates that do not cheat!

Key Problem with All the Above Methods:

Model is developed (for example) on 2008 data

Model will be applied in 2009.

Accuracy Assessment – Methodology

- Example: Use default rates for past loan applicants to predict next year's default rates.
  - Academic papers rarely hint that accuracy will be reduced because conditions will be different.
- There is very little public data that can be used to test how methods perform on target populations that differ somewhat (e.g., later in time) from the source population.

One answer: Update models continuously.

An Example – the Forensic Glass Dataset

The random forest algorithm gave a rule for predicting the type of any new piece of glass. For glass sourced from the same "population", here is how the rule will perform.

	WinF	WinNF	Veh	Con	Tabl	Head	CE[2]
WinF (70)	63	6	1	0	0	0	0.10
WinNF (76)	11	59	1	2	2	1	0.22
Veh (17)	6	4	7	0	0	0	0.59
Con (13)	0	2	0	10	0	1	0.23
Tabl (9)	0	2	0	0	7	0	0.22
Head (29)	1	3	0	0	0	25	0.14

The data consist of 214 rows $\times$ 10 columns.

WinF = Window float WinNF = Window non-float
Veh = Vehicle window Con = Containers
Tabl = Tableware Head = Headlamps

[2] Classification Error Rate (cross-validation)

Questions, Questions, Questions, . . .

- How/when were data generated? (1987)
- Do the samples truly represent the various categories of glass? (To make this judgement, we need to know how data were obtained.)
- Are they relevant to current forensic use? (Glass manufacturing processes and materials have changed since 1987.)
- What are the prior probabilities? (Would you expect to find headlamp glass on the suspect's clothing?)

These 1987 data are not a good basis for judgements about glass fragments found, in 2008, on a suspect's clothing.

The Data Mining "Big Data" Theme – Issues

- Data can be large in bulk, but contain a small number of independent items of information,
  - e.g., a days's temperatures, collected at millisecond intervals.
- Beware of increased risks of detection of spurious effects.
- Graphics often requires care (points overlap too much).

Why plot the data?

- Which are the difficult points?
- Some points may be mislabeled (faulty medical diagnosis?)
- Improvement of classification accuracy is a useful goal only if misclassified points are in principle classifiable.

What if points are not well represented in 2-D?

Cunning is needed!

Methodologies for Low-Dimensional Representations

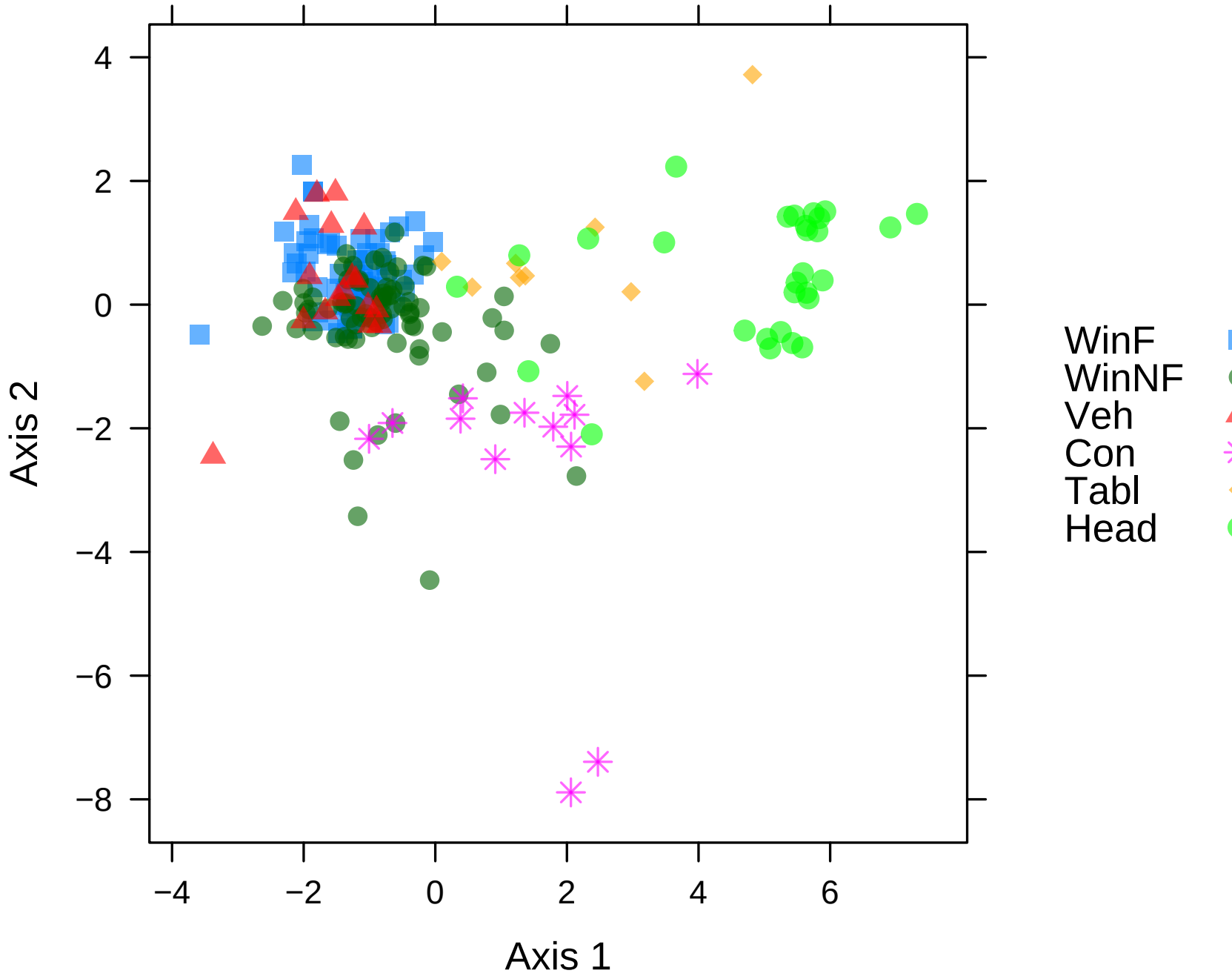
- From linear discriminant analysis, use the first two or three sets of scores
- Random forests yields proximities, from which relative distances can be derived.

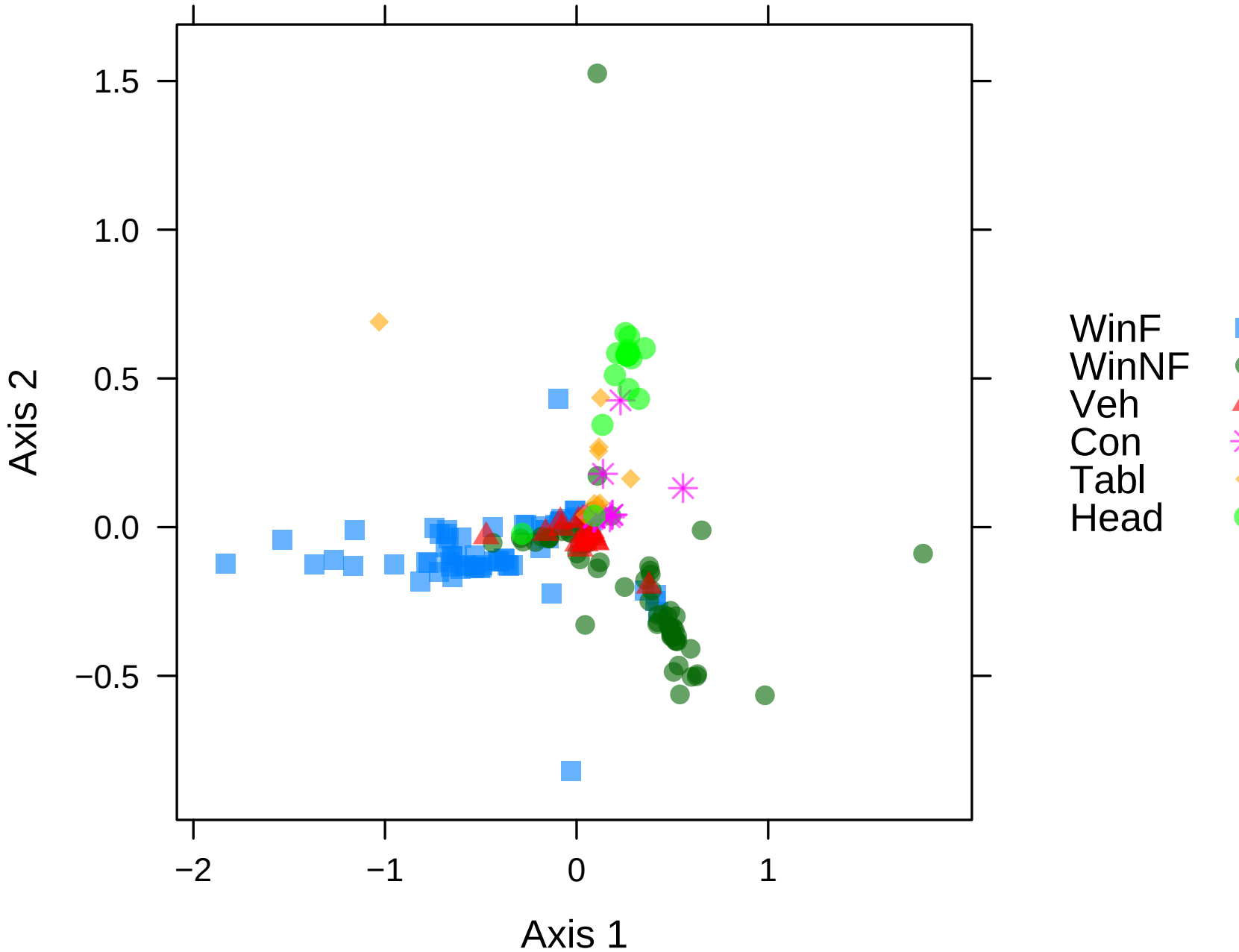
  Use semi-metric or non-metric multidimensional scaling (MDS) to obtain a representation in 2 or 3 dimensions.

  The *MASS* package has `sammon()` (semi-metric) and `isoMDS()` (non-metric) MDS.

The next two slides give alternative two-dimensional views of the forensic glass data, the first using linear discriminant scores, and the second based on the random forest results.[3]

[3] Code for these graphs will be placed on JM's webpage.


Axis 2
Axis 1
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
WinF
WinNF
Veh
Con
Tabl
Head



Distances were from random forest proximities

Advice to Would-be Data Miners – the Technology

Four classification methods may be enough as a start

- Use linear discriminant analysis (`lda()` in the *MASS* package) as a preferred simple method. The first two sets of discriminant scores allow a simple graphical summary.
- Quadratic discriminant analysis (`qda()` in the *MASS* package) can perform excellently, if the pattern of scatter is different in the different groups.
- Random forests (`randomForest()` in the *randomForest* package) can be used in a highly automatic way, does not overfit with respect to the source data, and will often outperform or equal all other common methods.
- Where complicated (but perhaps clearly defined) boundaries separate the groups, SVMs (`svm()` in the *e1071* package) may perform well.

Getting the science right is more important than finding the true and only best algorithm! (There is no such thing!)

But surely all this stuff can be automated?

Advice often credited to Einstein is:
"Simplify as much as possible, but not more!"

In the context of data analysis, good advice is:
"Automate as much as possible, but not more!"

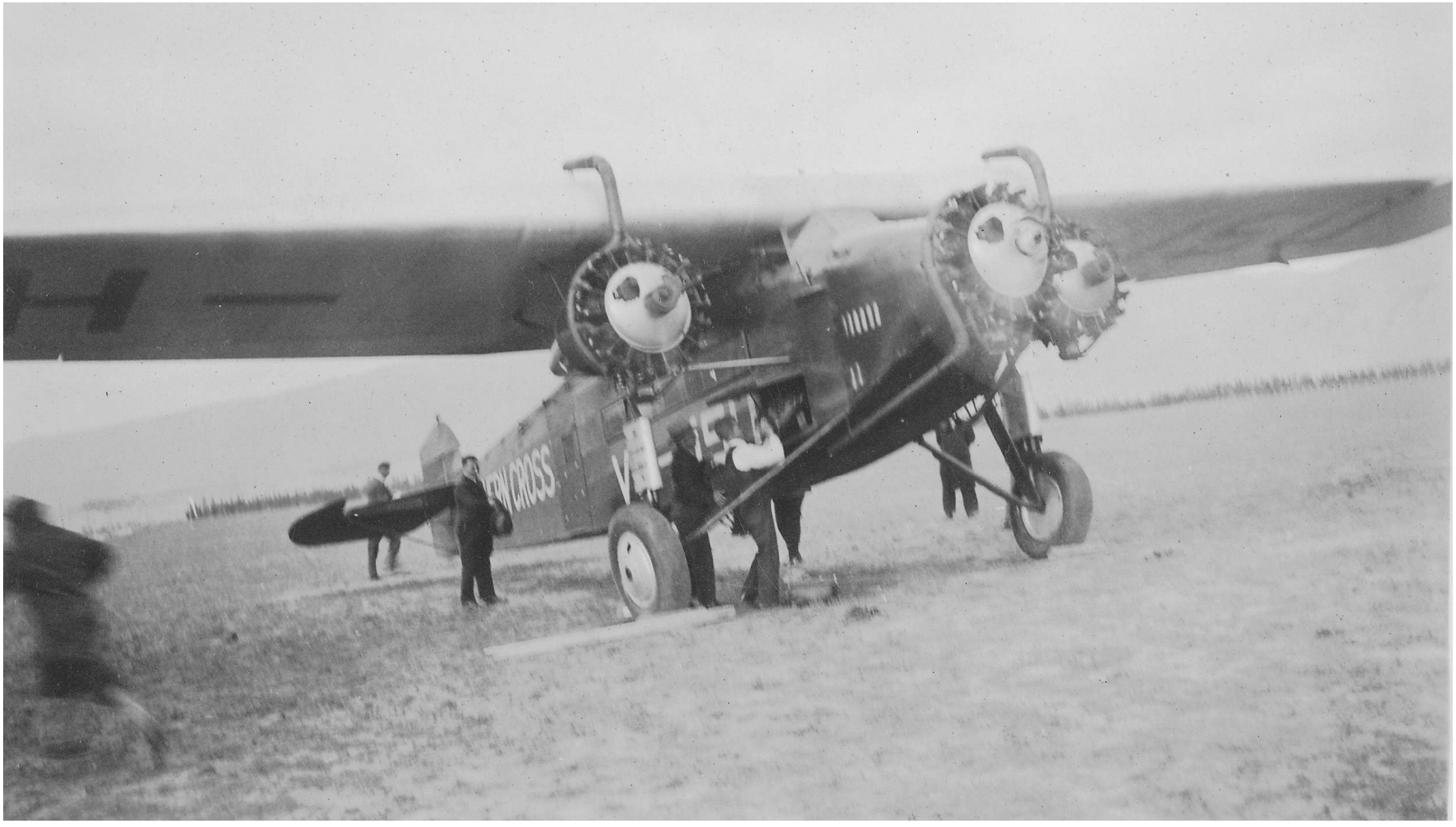
Many researchers are working on automation within R, or based on R.

The reality is that the extravagant promises of the early years of computing are still a long way from fulfilment:

> *1965, H. A. Simon: "Machines will be capable, within twenty years, of doing any work a man can do" !!!*

Think about what automation has achieved in the aircraft industry, and the effort it required!

Analytics on autopilot?



"...analytical urban legends..." (D & H)

Sometimes, autopilot can work and is the only way!

. . . but there is a massive setup and running cost

Computer Systems are Just the Beginning

Even with the best modern software, it is hard work to do data analysis well.

References

Berk, R. 2008. *Statistical Learning from a Regression Perspective.*
[Berk's extensive insightful commentary injects much needed statistical perspectives into the discussion of data mining.]

Maindonald, J.H. 2006. Data Mining Methodological Weaknesses and Suggested Fixes. Proceedings of Australasian Data Mining Conference (Aus06)'[4]

Maindonald, J. H. and Braun, W. J. 2007. *Data Analysis and Graphics Using R – An Example-Based Approach.* 2nd edition, Cambridge University Press.[5]
[Statistics, with a slight data mining flavor.]

[4] `http://www.maths.anu.edu.au/~johnm/dm/ausdm06/ausdm06-jm.pdf` and `http://wwwmaths.anu.edu.au/~johnm/dm/ausdm06/ohp-ausdm06.pdf`

[5] `http://www.maths.anu.edu.au/~johnm/r-book.html`

Web Sites

http://www.sigkdd.org/
[Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining.]

http://www.amstat.org/profession/index.cfm?fuseaction=dataminingfaq
[Comments on many aspects of data mining.]

http://www.cs.ucr.edu/~eamonn/TSDMA/
[UCR Time Series Data Mining Archive]

http://kdd.ics.uci.edu/ [UCI KDD Archive]

http://en.wikipedia.org/wiki/Data_mining
[This (Dec 12 2008) has useful links. Lacking in sharp critical commentary. It emphasizes commercial data mining tools.]

The R package *mlbench* has "a collection of artificial and real-world machine learning benchmark problems, including, e.g., several data sets from the UCI repository."