History of S and R

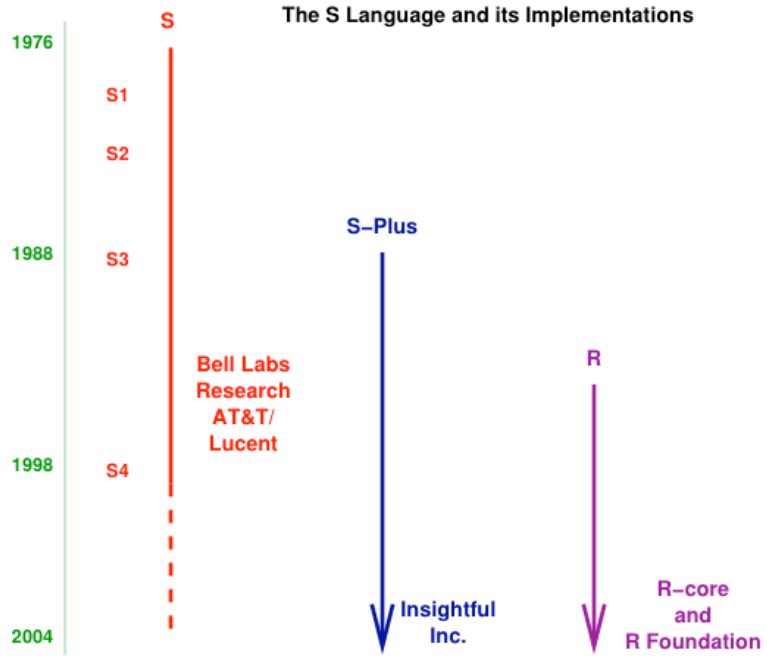
(with some thoughts for the future)

John M. Chambers

June 15, 2006

Statistical Software Today

- More software is available then ever before for data analysis, & much of it is good.
- The S software was written by and for Bell Labs statistics research.
- The open-source R system, based on the S language , dominates new work.
- This talk looks at the history & current state of S and R.



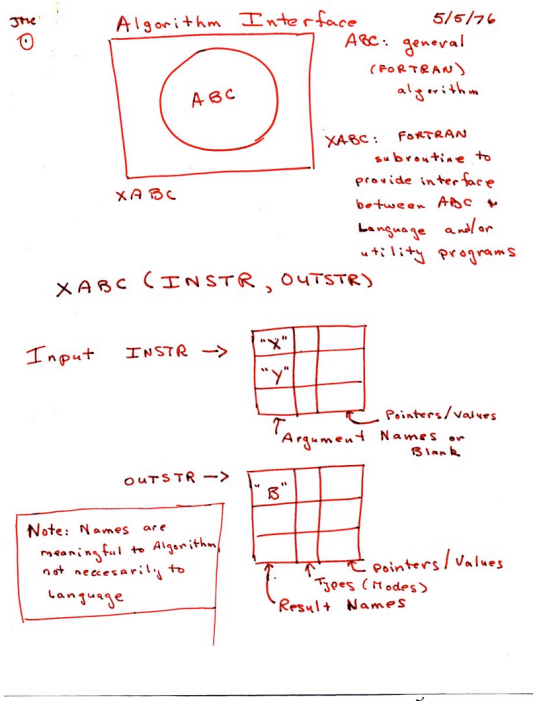
First Discussions, May 1976

- Rick Becker (graphics, NBER systems)
- John Chambers (graphics, data, algorithms)
- Douglas Dunn (time series)
- Paul Tukey (APL, other graphics)
- Graham Wilkinson (GENSTAT)

May 5, 1976

Sketch proposing an interface between S functions and Fortran routines.

And (below) the structure of function arguments and values as lists of named elements.



S Version 1 (1976-1978)

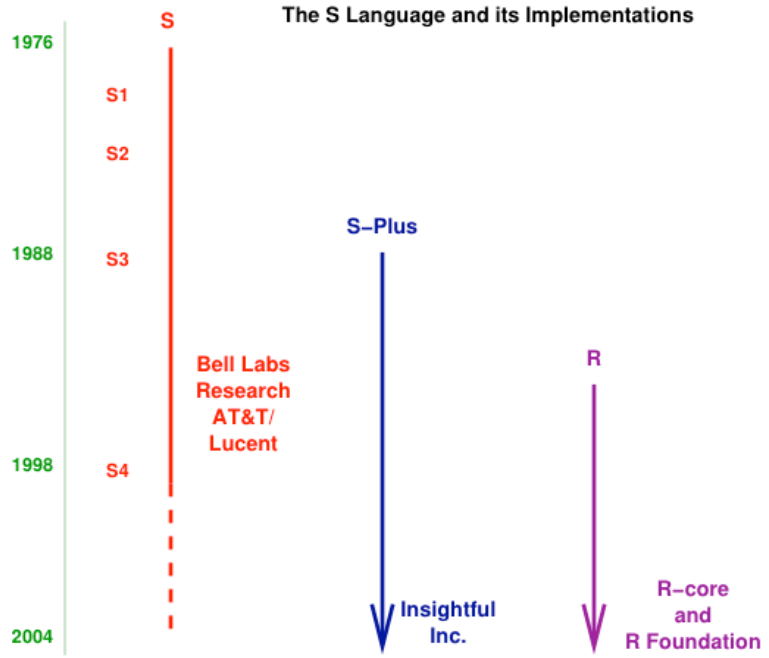
- Implementation nearly all Fortran based, via preprocessing tools.
- Only for our (bizarre) operating system.
- Adopted our existing graphics & data structure software.
- Interfaces to many algorithms (random numbers, linear algebra, some models).

Meanwhile, Unix & Licensing

- Unix developed roughly in parallel to us, also in a local form.
- Portable Unix designed ~1978 (32 bit!).
- We decided to port S to Unix.
- AT&T adopted a licensing policy (very cheap for universities).
- S rode along with Unix & a few others.

S Version 2

- Portability via a Unix implementation:
  - Unix ports most features for us
  - Device-independent graphics
  - Model for machine numerical properties
- Most features carried over from V. 1.
- Licensed to the outside from ~1981; books in 1984/5.



S Version 3 (1983-1992)

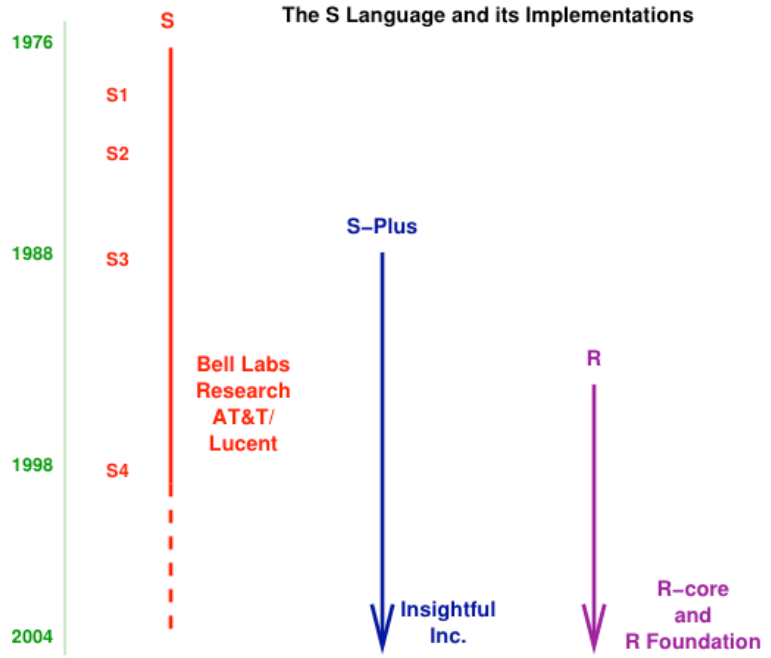
(the `blue book')

- Merged some new ideas with S.
- "Everything is an object" (including functions).
- Functional evaluation model.
- .C(), .Fortran(), *no* Interface Language.
- No direct back compatibility with S2.

Statistical Models in S (S3)

(the `white book')

- An object-based approach.
- Model formulas (& terms objects).
- Data Frames (& model frames, ...).
- S3 methods
  - Give the user a simple call for plot, summary, predict, etc.
  - Minimal additions to S engine & API



Events from 1995 to present

- S Version 4
- S software licensed exclusively (1993), eventually sold to Insightful (2004).
- ACM `Software System' award
- Along came

S Version 4 (1995-1998)

(the `green book')

- `Computing with data' distinguished from statistical computing.
- Extensions to the S programming model:
  - Classes and methods with metadata
  - Connections, documentation objects, ...
- Today we have the S *language*, implemented in R and S-Plus *software*.

What & Who is R?

- Ross Ihaka & Robert Gentleman wrote an experimental R, "not unlike S" (ca 1995).
- R-core (17 people), R Foundation (5 directors) control the design & evolution.
- Contributors from many countries, mostly academics, provide packages & tools.
- Users; number unknown: ~100K? Important concentration among students, researchers.

R–A real success story

- Software for statistics, data management, programming, etc. exists in quantity & variety unimaginable 15 years ago.
- Quality varies, but on average is impressive.
- And, most of this is in an open environment that encourages improvements.
- Wide participation from the statistics profession is also a healthy sign.

The Future

Challenges for statistical software:

- Data *processes* in real-time
- Embed our software in *their* software
- Very large scale applications

Will an open-source system like R respond to these challenges?

Can R Meet the Challenges?

The responses require new software that does more than just add to current R and its packages. The computing research needed is risky: to use the results will require basic changes.

Where are the resources and the organization to take such steps?

Will Fundamental Change Be Possible?

- At two major change points (S3 and start of R), researchers had freedom and support for change.
- Future changes will have to face the popularity of current R (resistance to breaking anything).
- Researchers at the level of expertise needed are scattered , and scarce.
- Needed: support for risky, fundamental change, and a plan to use the results.

Statistical Software Today

- Who would have imagined it all, in 1976?
- Current software is good for statistics, and gratifying for the originators of S.
- But the resources of 1976 are not available now, as we look to meet new challenges.
- Let's hope that new people and new resources will take up the challenges.