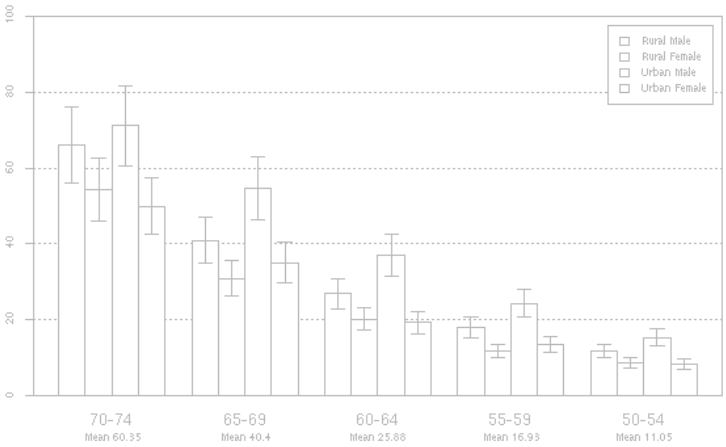
REVOLUTION ANALYTICS



The Coming Revolution in Statistics

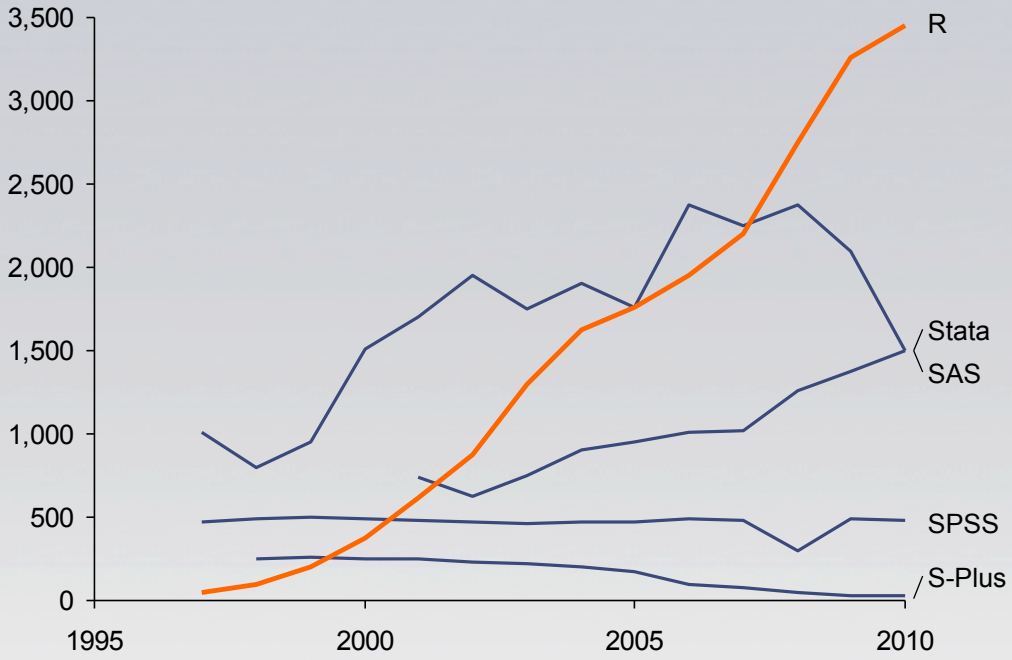
Lee E. Edlefsen, Ph.D.
Chief Scientist

- The leading commercial provider of software and support for the popular open source R statistics language.
- Palo Alto, Seattle, New York.
- www.revolutionanalytics.com/video.php

"R is the most powerful & flexible statistical programming language in the world"…

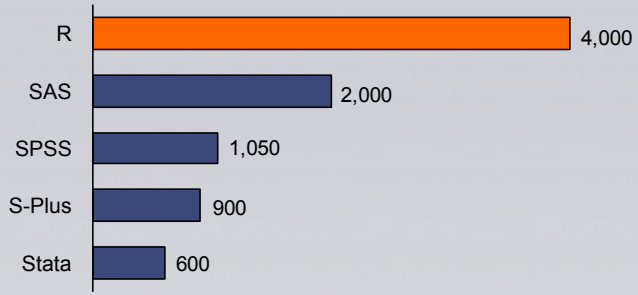
Internet Discussion

Mean monthly traffic on email discussion list



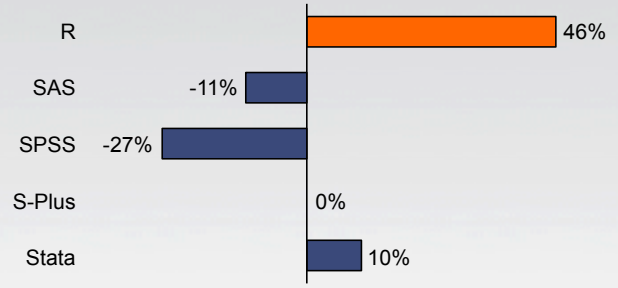
Web Site Popularity

Number of links to main web site



Scholarly Activity

Google Scholar hits ('05-'09 CAGR)



Source: http://r4stats.com/popularity

The coming revolution – due to disruptive technological change

- I believe there is going to be a revolution in both statistical practice and theory over the next several years

- This revolution will be driven by disruptive technological change: our ability to collect and store data is rapidly and greatly outpacing our ability to analyze that data

Huge benefits to huge data

- More information, more to be learned
- Variables and relationships can be visualized and analyzed in much greater detail
- Can allow the data to speak for itself; can relax or eliminate assumptions
- Can get better predictions and better understandings of effects

We are currently incapable of analyzing much of the data we have

- The most commonly-used statistical software tools either fail completely or are too slow to be useful on huge data sets

- In many ways we are back where we were in the ‘70s and ‘80’s in terms of ability to handle common data sizes

Code museums and the end of an era

- The vast majority of the data analysis software in use today is based on algorithms that are 30, 40, 50 or more years old
- Much of the actual code dates back that far
- During that period of time the rising tide of technology allowed the same code to run faster and on bigger data sets
- We are at the end of that era

To keep up with the tsunami of data

- We must:
  - use more cores
  - use more hard drives
  - use more computers
- Existing statistical software can't do this
- We need new software

New statistical software must be

- Scalable – the same code that works on 100 observations should work on a 100 billion
- Fast – need results in a timely manner; good data analysis requires interactivity
- Easy to use – we need to be able to use clusters and clouds as easily as we can use single workstations today

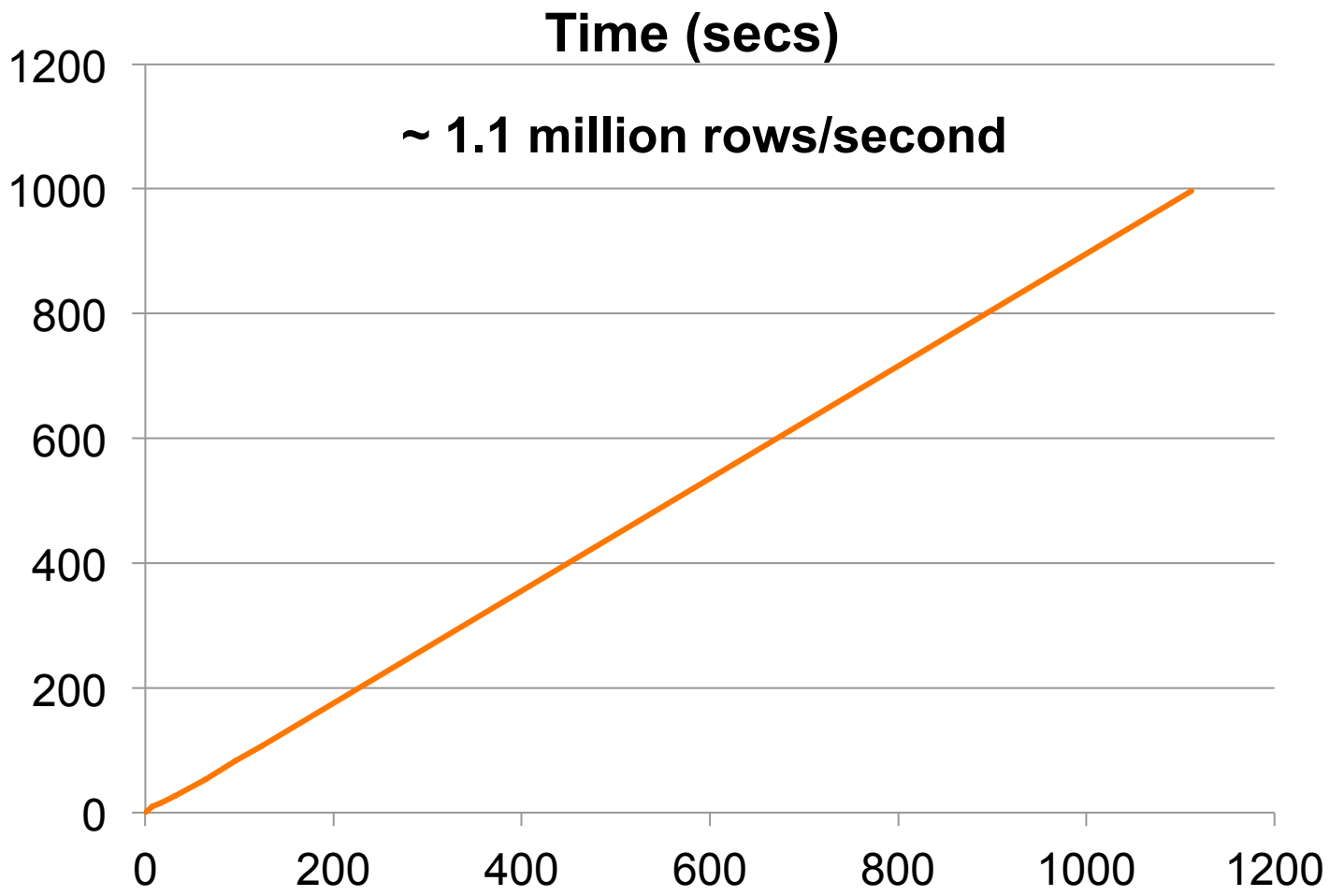
New software should also

- Be easily extendible to new fast, scalable algorithms
- Leverage as much existing software as possible
- Be flexible, forgiving, and familiar to lots of people

Is this possible? Yes!

- **Based on R** – R is not only the statistical language of the present, in my opinion it is the language of the future
- **Based on "parallel external memory algorithms"** -- at Revolution Analytics, we have released a framework for automatically and efficiently parallelizing and distributing a wide class of statistical and data mining algorithms

Scalability of RevoScaleR:
Regression, 1 million - 1.1 billion rows, 443 betas



Please contact me if you have questions

Lee Edlefsen

lee@revolutionanalytics.com