Some Reflections on Data Science

John M. Chambers

October 30, 2018

Some Reflections on Data Science

1. Data analysis/data science at Bell Labs
2. R and data science
3. Data science for the future

1. Data analysis/data science at Bell Labs

Bell Labs, Murray Hill, New Jersey

Time frames:

1940s - 1980s The "glory years" of Bell Labs.
From transistors and information theory in the 1940s to Unix and S in the 1980s, with many others in between.

1945-1985 John Tukey at Bell Labs, while also at Princeton.

1964-2005 My own 40 years at Bell Labs.

Today's discussion focuses on the intersection of these 40 year periods.

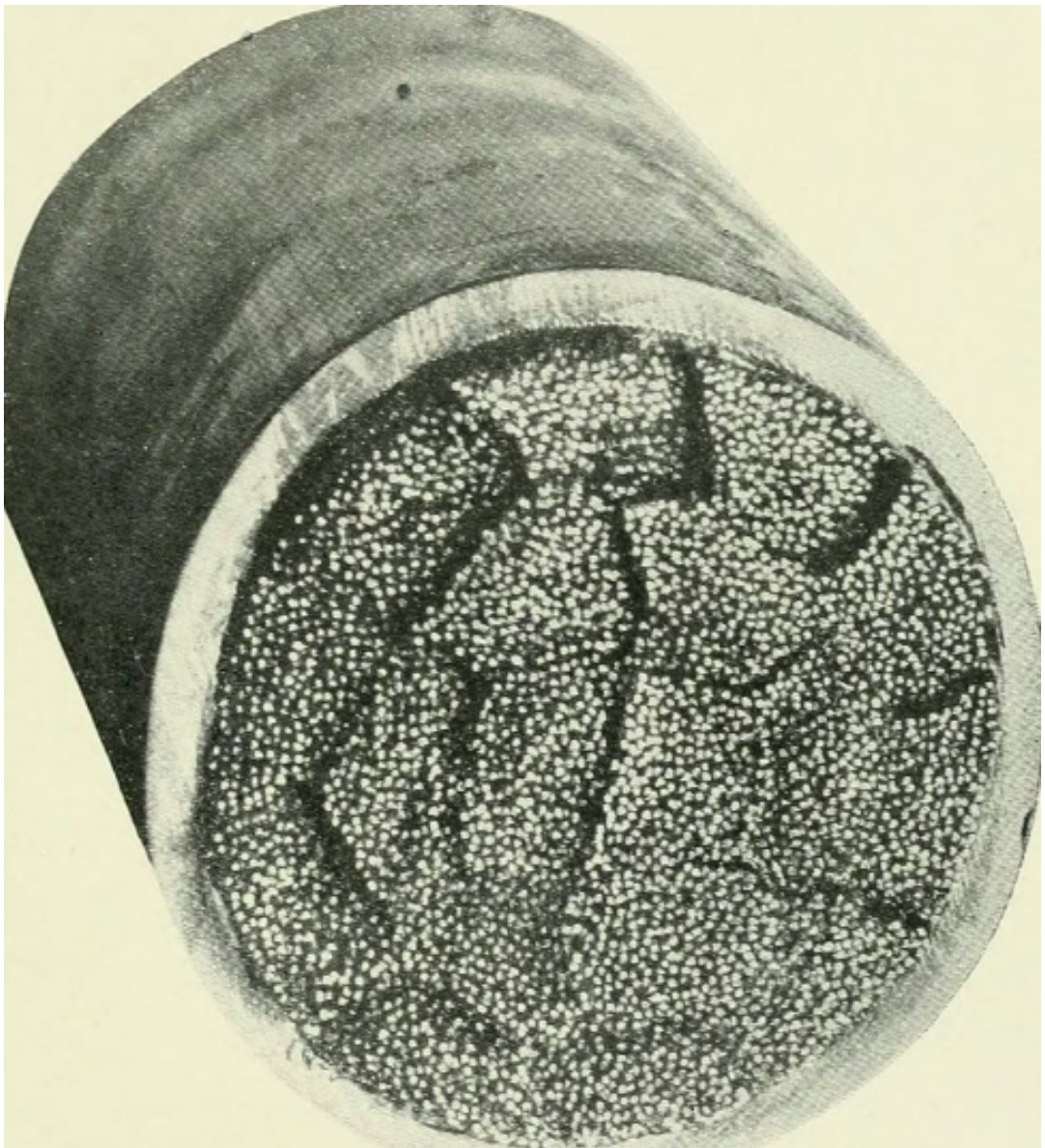
1964/5

Two early anecdotes:

Summer, 1964 Telephone cable sheath study: data taken along a stretch of cable recording quality of the sheath at this point. Goal: quality was occasionally inadequate. Find a pattern.

1964/5

1964/5



1964/5

The data analysis

- Examining the data (probably with spectral analysis) showed some periodic "bad" spots.
- The engineer, after some reflection, thought maybe wrapping the cable after extrusion might be causing problem. So, use the diameter of the drum?
- An analysis of variance with a factor coded to reflect position of that point on the drum did show a substantial effect.
- So the chemists & engineers went off to investigate. End of summer.

1964/5

Two early anecdotes:

Summer, 1964 Telephone cable sheath study: data taken along a stretch of cable recording quality of the sheath at this point. Goal: quality was occasionally inadequate. Find a pattern.

February, 1965 Some Bell Labs statistics researchers (among them John Tukey), got together to discuss future software for data analysis.
The result was a plan to create BLISS, a "Bell Labs Interactive Statistical System".

1964/5

Statistical computing

- Bell Labs was involved in a frontier experiment in computer operating systems (Multics).
- Computing was already seen as important to data analysis, so some researchers got together to plan a new project.
- For reasons nothing to do with BLISS, it was never completed.
- But, 11 years later another similar brainstorming session started work on a system that became S.

Data Analysis, Bell Labs Style

Data-centered i.e., not theory centered. Look at the data; worry about data quality. Deal with big data (for the period).

but ... *Do* use models and other high-tech tools. With emphasis on diagnostics & examining residuals.

Graphics Visualization of data and other graphics to aid analysis.

Computing is important for all of this.

Data Analysis and Data Science

The Bell Labs version of data analysis in the "glory years" evolved to look a lot like current data science (given differences in computing power, analytic techniques and data sources)

During this period, ideas of data analysis had spread; for example, Tukey's EDA book and other writings.

Data visualization and graphical methods became important; Bill Cleveland's contributions; probability plotting; interactive graphics.

But most of all, perhaps, data analysis spread and was ready for data science through the S, and then R, software.

Why Bell Labs?

The management of a very large company—one that depended on technology for its products and that had little competition—came to believe that *science and innovation* were its best road to continued growth.

Management, and funding, were very hierarchical. Research's annual budget was divided up into divisions, those into centers, those into departments (e.g., Statistics and Data Analysis). Managers at all levels were themselves researchers.

The key belief was that you if you did original research that was "important", you would be rewarded by technically informed management.

References:

- David Donoho. "50 years of data science." *Journal of Computational and Graphical Statistics*, 26(4):745-766, 2017
- Jon Gertner. *The Idea Factory: Bell Labs and the Great Age of American Innovation*. Penguin, 2013

2. R and Data Science

"Theorem:" R began as a domain-specific language for data science.

"Lemmas:"

1. R was explicitly designed to reproduce nearly all of version 3 of S.
2. S was designed as an interactive environment and programming language to implement the Bell Labs version of data analysis.
3. Bell Labs' data analysis at this time was a prototype for data science.

R added important new assets, notably the package mechanism.

The most important asset is the R community of user/contributors.

References:

- John M. Chambers. *Extending R*. Chapman and Hall/CRC, 2016 [Chapter 2]
- John M. Chambers. "S, R and data science." *History of Programming Languages, IV* [in preparation, ask me]

3. Data Science and the Future

and why data science is needed to save the world.

What is Data Science?

Data science includes all the techniques needed to make scientifically valid inferences and predictions from data, and to communicate those effectively.

All domains of science (and other areas of enquiry) have crucial new opportunities and challenges to learn from data.

In return, data science needs to collaborate with and learn from the knowledge-seekers in these varied domains.

What is Data Science?

The skills needed:

Analysis: Providing the inferences and predictions. This will draw on statistics, machine learning, optimization and other disciplines; the best current methods and research on better ones.

Computation: Dealing with data sources that may be extremely large, diverse and complexly structured; applying the analysis to such data; and enhancing the environment for people to learn from the data and the results.

Science: Collaborating with every branch of science to transfer these techniques. Also, an understanding of what makes for valid science when creating and using the techniques.

Communication: Conveying the results effectively and honestly to a variety of audiences, particularly those needing to act on these results. Visualization, interactive interfaces and other new techniques will be important.

The challenge to universities is to bring together those with relevant skills and motivation, both as part of the broader scientific contribution and to educate future contributors to data science.

Why it's needed — Some threats to the future:

- The global climate is changing rapidly, largely because of greenhouse gas emissions.
- A growing and more prosperous human population is destroying essential natural areas for its food, housing and industry.
- The oceans are being devastated by commercial exploitation and pollution, threatening species essential to life.
- A combination of human activities has already radically reduced the diversity of living things, and the process is accelerating.

Data Science is needed to respond to each of these

All of these problems have in common that they are complex and global. Local responses, not guided by an understanding of the whole picture, will not be adequate.

Understanding requires scientific analysis. The science is hard, only partly known now. It must involve many domains of science, dealing with the planet, its biological contents and with human behavior.

Data Science is needed to respond to each of these

None of the science can work without complete global data. That will be large, diverse and very challenging to manage well.

And the scientists must be able to communicate what they learn to a variety of audiences, particularly to those who will be making decisions about future actions.

This means data science with capabilities for **analysis**, **computation** and **communication** to support all of **science** .