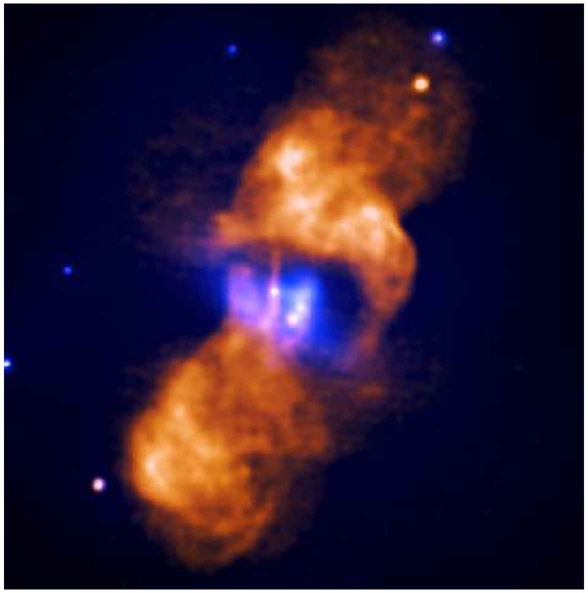
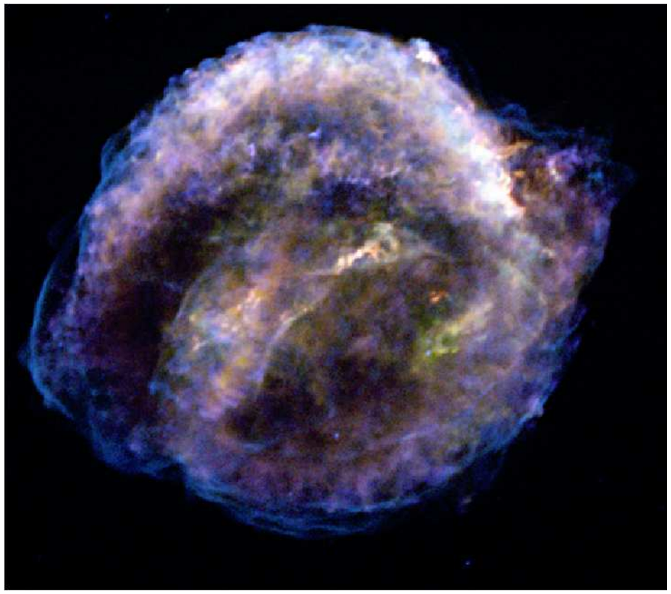
Data-Driven Science: The View of a Statistician

David A. van Dyk

Who I am and what is my perspective:

1. A Statistician (Department of Statistics, University of California, Irvine)
2. The Head of the California-Harvard Astrostatisics Collaboration (CHASC)
3. The Editor of *The Journal of Computational and Graphical Statistics*





Complexity versus Volume of Data

An example from Astronomy:

The Sloan Digital Sky Survey:

- Over five years, SDSS imaged more than 8,000 square degrees of the sky in five bandpasses, detecting nearly 200 million celestial objects, and measured spectra of more than 675,000 galaxies, 90,000 quasars, and 185,000 stars.
- The total quantity of survey information produced is about 15 terabytes.

The Gamma Ray Large Area Space Telescope (GLAST):

- Sophisticated data analysis on each incoming photon is required to characterize uncertainty in the photons sky coordinates and energy.
- Ambitious and time-consuming computation is required for each observation.

Different Tools for Different Jobs

- Currently *very different statistical/computational tools are required* to handle the spectrum of data/scientific challenges within and across disciplines.
  - Complex data typically require highly-specialized models and computational algorithms.
  - Large data mining projects typically have more in common but still require some specialization. (*What are we looking for?*)
- Much work in *Statistical Computation* has focused on handling more complex data and orienting model fitting and computation toward answering specific questions of scientific interest.
  - Since the form of these questions and models can vary enormously, the implementation of the methods also vary.
  - The methods, however, are guided by a common set of underlying mathematical/probabalisitic/statistical principles and use combinations and specifications of tools from a computational library.

How much Automation of Science is Possible?

- *Enormous data sets swamp human capacity.* There is simply too much data to think about in a scientifically meaningful manner.
  - This is true even with relatively small data sets with complex models and/or unrefined scientific questions.
  - Consider summarizing the complex correlations of the posterior distribution of a high-resolution image using noisy data of moderate volume in the context of ill-posed scientific questions. (E.g., A search for "interesting and/or unusual structure".)
- **GOAL:** How can we reduce the dimension to something we can think about without potentially masking the most interesting and unexpected features.

All Models, Methods, and Algorithms bring feature-masking assumptions to the data. Of course, some methods are better at hiding these assumptions!!

A Needed Paradigm Shift

- *Computational/statistical processing* of data is just *an afterthought* for many scientists.
  - Multi-billion dollar space-based observatories may require decades of planning and design before launch.
  - Nonetheless scientists may find computational/statistical methods that require hours or days of computing unacceptable.
- Even when the computational challenges are recognized, *funding, resources, and talent* for computational/methodological development *are often in short supply.*
  - Glitzy Hi-Tech instruments are easier to fund than methodological work despite the enormous asymmetry in cost.
  - There are far too few scientists with the significant interdisciplinary skills necessary for the work.