Real-time processing and analysis of data streams

(with hand-waving)

Jay Emerson
Department of Statistics, Yale University

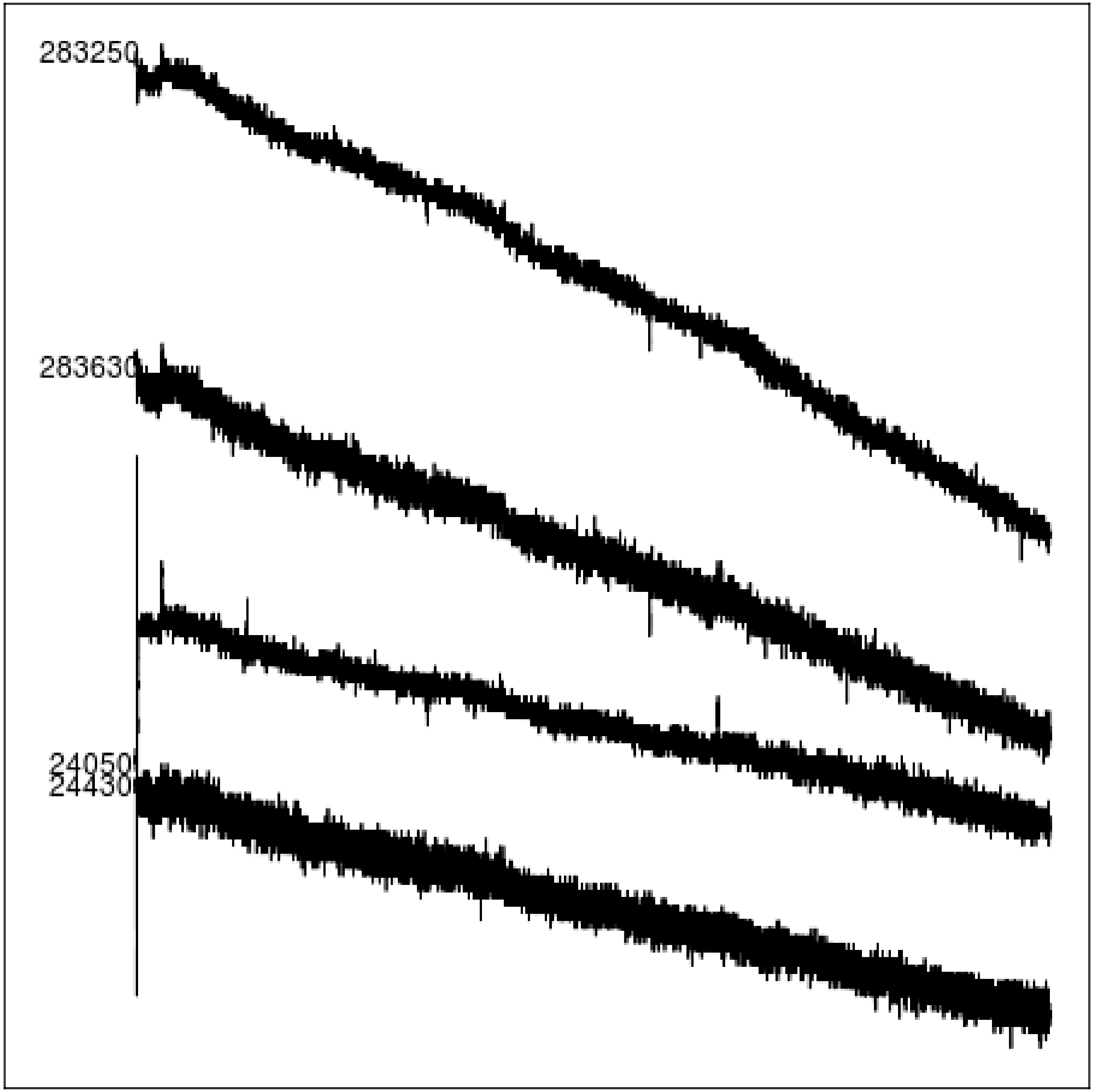
Mike Kane
GRD September 2010

Taylor Arnold
GRD 2013

Bryan Lewis
Independent

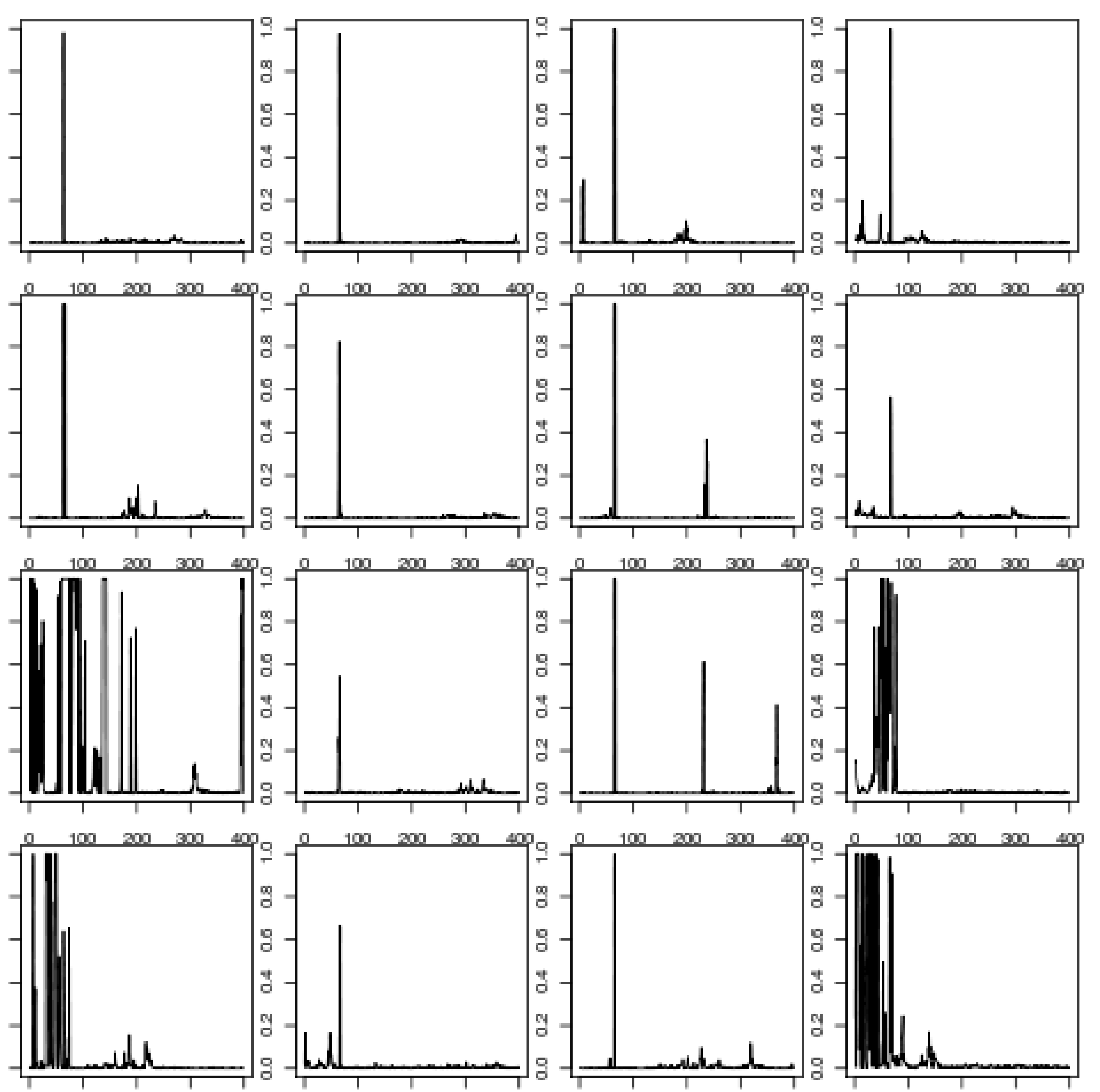
- Case study for real-time data streaming: toy package **bigvideo**
  - On R-Forge; only tested in Linux with one Logitech video cam
  - Yes, a toy package at this point, but… why not?
  - Uses the very cool OpenCV library
  - Entirely my fault if something doesn’t work
  - Introductory demo here: live from UseR! 2010
- Streaming (more or less):
  - Data source (the “feed”)
  - Data analysis (or transformation or reduction or… the “filter”)
  - Output (a plot, a summary, processed data, some end result… the “sink”)
- Speed: slow filter/sink → you fall behind and/or lose data
- Potentially cumbersome amounts of data
- Second demo: recorded last night @ dusk, Crowne Plaza, Rockville, MD. ~13 minutes, ~ 13 GB, ~25 frames/sec.

4 pixels at dusk: the terrace of the Crowne Plaza Rockville



- Speed: slow filter/sink → you fall behind and/or lose data
  - Parallel processing (filtering); difficult synchronization
  - Package **synchronicity** for a shared memory mutex structure (SMP locking) and a basis for simple SMP synchronization
    - Alternatives: NetWorkSpaces infrastructure (powerful but not a simple CRAN package); Norm Matloff's **Rdsm** (may be ideal for distributed signaling beyond simple SMP work); MPI, etc…
- Potentially cumbersome amounts of data
  - Shared memory (avoid copies, perhaps filebacked): mmap
  - Package **bigmemory** (http://www.bigmemory.org/ and on CRAN) and sister packages on CRAN and R-Forge. Provides big matrices. Fast, easy to use, extensible (e.g. BLAS-compatible, with a C++ accessor framework for general algorithm development).
    - Alternatives (particularly if you really want more than matrices: Dan Adler et.al.'s **ff**, Jeff Ryan's **mmap** for data.frame/database designs).
- Third demo: a crude pipeline with signaling between processes, illustrating the challenges.

The plot produced by the crude pipeline (the “sink”)



Sort of a conclusion, or an aside…

- Efron (2005):

  “We have entered an era of massive scientific data collection, with a demand for answers to large-scale inference problems that lie beyond the scope of **classical statistics**.”

- Kane, Emerson, and Weston (in preparation), with reference to Efron’s quote:

  “**classical statistics**” should include “**mainstream computational statistics**.”

- Two interrelated challenges (requirements) of working with large data sets:

  Faster computation

  Ease of access, manipulation and analysis of larger data sets

- So we’re back where we started… facing exactly the issues identified by Luke this morning. Packages are (or may not but can be) great (record video of lots of mutual back-patting in the audience). However:

- **Thank you, R Core!**

Not really streaming… but… it's late… just for fun…

Interpolated Flight Locations, minute-by-minute, for flights taking off after 12:01 AM, January, ending mid-morning, January 2. 1995 I think. I couldn't get the movie into the slides, sorry; email me if interested.