What Passes for Big Data in Solar Physics at NASA Goddard

Joseph B. Gurman

NASA Goddard Space Flight Center
Heliophysics Division
Laboratory for Solar Physics

Facility Scientist, Solar Data Analysis Center
US Project Scientist for SOHO
STEREO Project Scientist
What Exactly Do We Mean by “Big?”

Big like Google, whose data holdings have been estimated at ~ 15 Ebyte?

- Or put another way, if encoded on Hollerith cards, one New England 4.5 km deep

Even the NSA probably still holds less.

– The Onceler, Dr. Seuss, The Lorax
What Exactly Do We Mean by “Big?” (II)

- The “classic” (2001) definition, which requires:
  - volume
  - variety
  - “velocity” (flux)

– The Onceler, Dr. Seuss, The Lorax
What Exactly Do We Mean by “Big?” (III)

Or do we mean “Big Data Methods,” or something else?

This is what happens when you ask 154 “C-suite” executives:

This is what I’ve been asked to talk about….

And to some extent,
Size of symbol represents nominal data rate

- NISP/SOLIS
- GONG*
- SOHO*
- TRACE*
- Yohkoh*
- STEREO*
- Hinode*
- RHESSI
- IRIS
- SolarG
- DKIST
- SDO
- Label key:
  - NASA missions or partnerships
  - Ground based observatories/networks

* Data served by SDAC

All data sources on this chart are or will be accessible via the VSO.
We (the SDAC) Don’t Yet Hold an Entire “Big Data” Archive

Current largest complete archive holdings are Hinode (20 Tbyte) and STEREO (80 Tbyte): small potatoes.

Largest single-mission holding is SDO (mostly AIA data from periods of interest + last ~ year): 1 Pbyte.

The archive as a whole is characterized by all three V’s (volume, variety, and velocity), but the highest “velocity” (flux) data set has no variety (SDO AIA).

“….as little as we are, when we fall out with each other, the city of Byzantium is not big enough to hold us.” – Leo Byzantius, quoted by Plutarch, Political Precepts
A Story of Days to Come (I)

At some point, entirely TBD, the SDAC is likely to be named the long-term archive for the entire SDO data set

2 Pbyte/year, 6 years so far

Not supportable by current storage architecture (or server room floor)

Comes at a time of planning for change in NASA and data architecture (agency, center, commercial, ? clouds, “consolidation”)

- Likely configuration TBD until sometime in FY18

– Shakespeare, Othello, III, 3

Big wars / That make ambition virtue
A Story of Days to Come (II)

I wish I could tell you how NASA / Goddard / the SDAC will be handling the data from SDO (the last big data solar mission?) three years from now.

Unfortunately, it’s unclear and I have a limited horizon.

Somebody else will be making those decisions.

— Shakespeare, As You Like It, II, 7
Responses to Questions (I)

What are the processes for planning for future (5-10 years) capabilities of your service? How do you gather input for this planning process and where does input typically come from? What new features have highest priority?

SDAC is mission-oriented; meets with mission scientists to plan data service; VSO is user-oriented, using frequent outreach and scientific meetings to identify data providers and services to add. SDAC new features are typically driven by the available technology; VSO features represent what we develop (or adopt off the shelf) in response to user community requests.

– David Packard
Responses to Questions (II)

What feature(s) of your service would you like to stop performing? How do you gather input for making such decisions and where does input typically come from? What is preventing you from doing so?

We would truly, deeply like to get out of the business of specifying new storage on multiple tiers every five to seven years. “Input” is experience-based, by the facility scientist. Wasteful to continue battling numerous, partially conflicting agency directives. Lack of clearly effective alternatives that meet mission risk management criteria and a current period of confusion (“design...
Responses to Questions (III)

What steps are you taking to make your data interoperable with allied data sets from other fields both in and out of NASA? How do you find allied data sets and what criteria make data sets candidates for enabling interoperability?

Our data pretty are currently interoperable with any other modern solar physics data (FITS format; software libraries in IDL and Python). Interoperability with other scientific disciplines, even within heliophysics, requires specialist knowledge, so not clear whether investing in further interoperability buys anything.

Virtual Solar Observatory: all solar physics data, not just space

SPDF + VxOs: access to heliophysics data

Take risks. Ask big questions. Don't be afraid to make mistakes; if you don't make mistakes, you're not reaching far enough.

– David Packard
The real variety: scientific analysis

Given the Virtual Solar Observatory (VSO)’s data dictionary, which recognizes only 16 physical observables, is in a data science sense too rich.

Only three basic organizations of solar data: map (image, data organized as one), spectrum, and light curve.

Real variety is in the solar features analyzed and the methods used to analyze them.

If you want to populate the variety dimension of a NASA solar big data map, look at the work of the SDO Feature Finding Team:
- distributed effort to develop multiple big data tools
- http://solar.physics.montana.edu/sol_phys/fft/static/