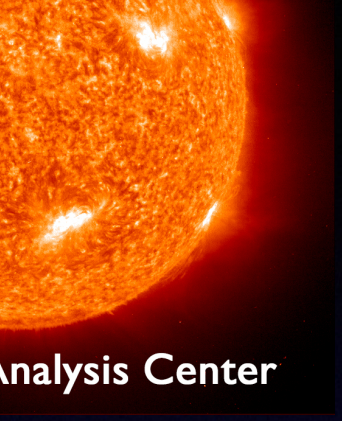
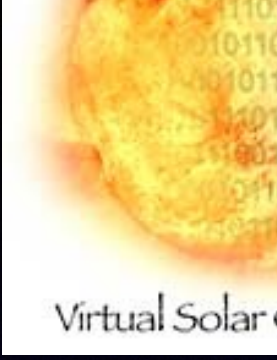


I never never use a Big, Big D.

– *The First Lord's song, W.S. Gilbert,* HMS Pinafore



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

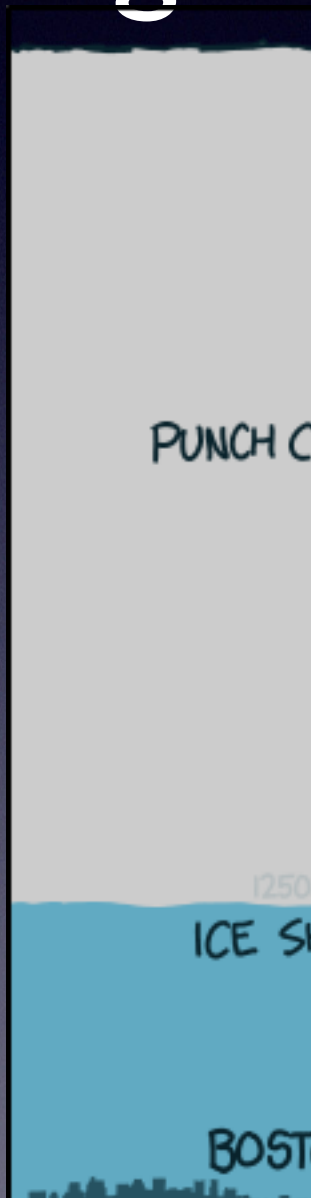
– *The Onceler, Dr. Seuss,* The Lorax

hat Exactly Do We Mean by "Big?"

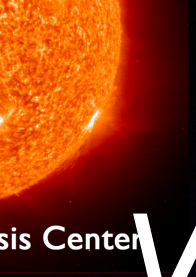
Big like Google, whose data holdings have been estimated at ~ 15 Ebyte?

- 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

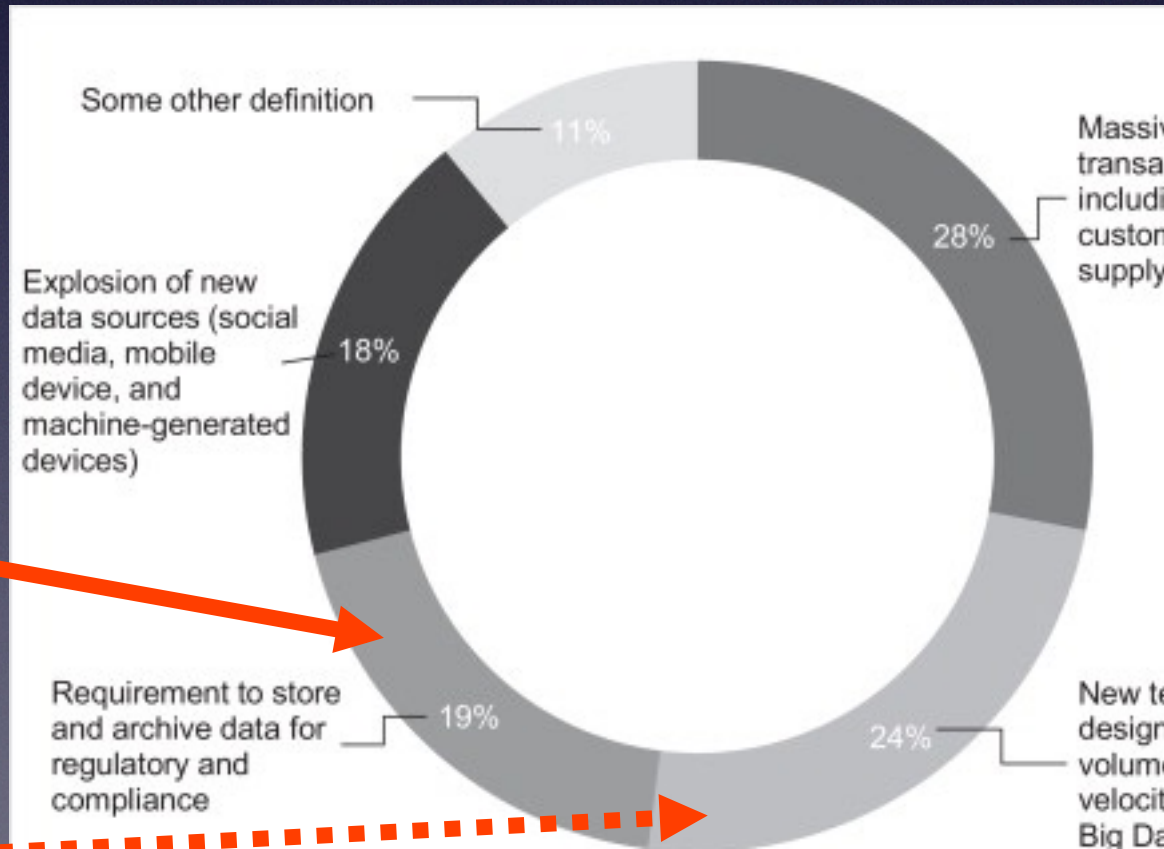
What Exactly Do We Mean by "Big?" (III)

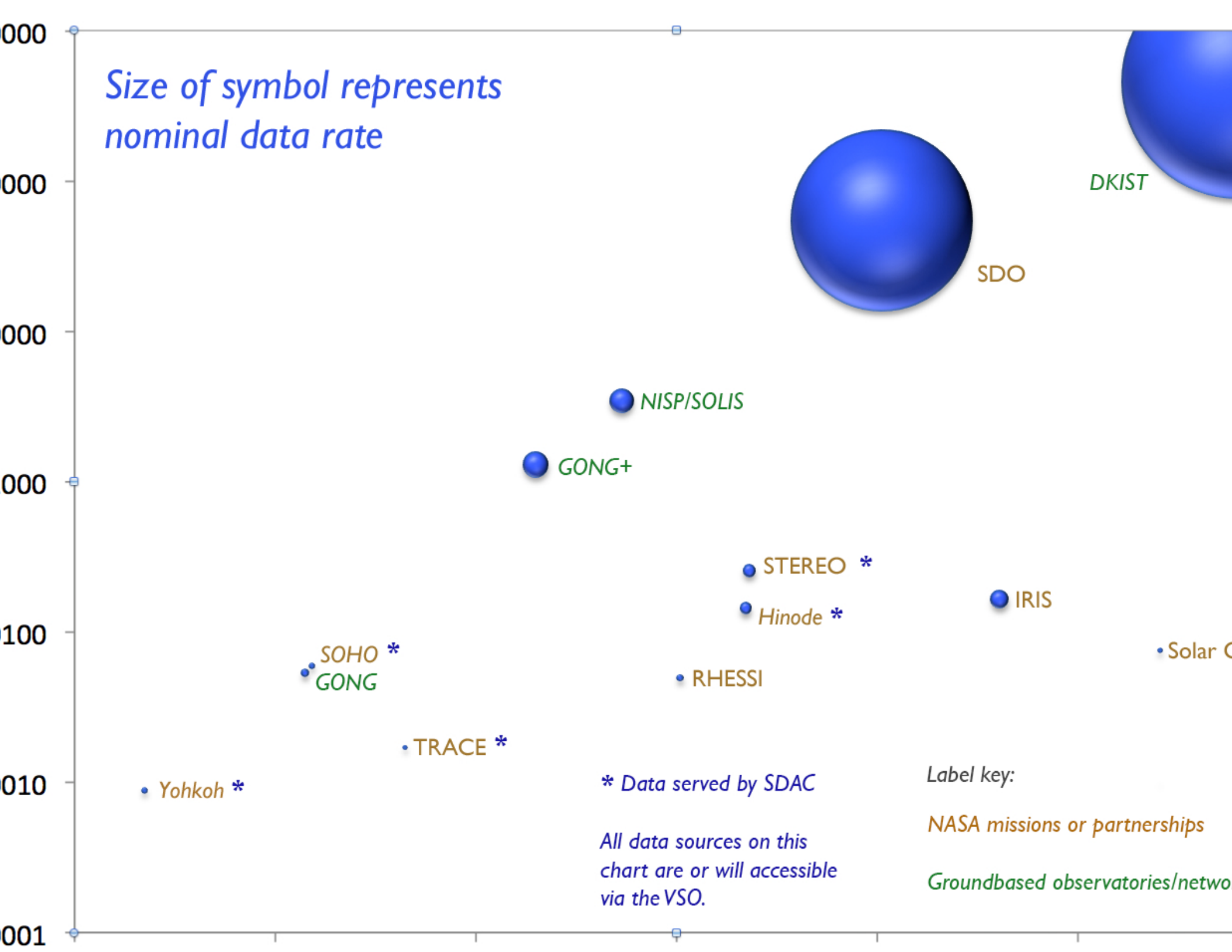
r do we mean "Big Data Methods," or
mething else?

s what happens when
sk 154 "C-suite"
tives:

s what I've been
to talk about….

d to some extent,




Size of symbol represents nominal data rate
DKIST
SDO
NISP/SOLIS
GONG+
STEREO *
IRIS
Hinode *
SOHO *
GONG
RHESSI
TRACE *
Yohkoh *
* Data served by SDAC
All data sources on this chart are or will accessible via the VSO.
Label key:
NASA missions or partnerships
Groundbased observatories/netwo

– *Leo Byzantius, quoted by Plutarch,* Political Precepts

We (the SDAC) Don't Yet Hold an Entire "Big Data" Archive

Current largest complete archive holdings are *H*
20 Tbyte) and STEREO (80 Tbyte): small potat

argest single-mission holding is SDO (mostly A
ata from periods of interest + last ~ year): 1 P

The archive as a whole is characterized by all th
"s (volume, variety, and velocity), but the highe
velocity" (flux) data set has no variety (SDO A

– *Shakespeare*, Othello, *III, 3*

A Story of Days to Come (I)

At some point, entirely TBD, the SDAC is likely
e named the long-term archive for the entire S
ata set

- 2 Pbyte/year, 6 years so far
- Not supportable by current storage architecture (or
room floor)
- Comes at a time of planning for change in NASA and
data architecture (agency, center, commercial, ? cloud
"consolidation")
- Likely configuration TBD until sometime in FY18

And whistles in his sound."

– *Shakespeare,* As You Like It, *II, 7*

A Story of Days to Come (II)

wish I could tell you how NASA / oddard / the SDAC will be andling the data from SDO (the st big data solar mission?) three ears from now

nfortunately, it's unclear and I ave a limited horizon

omebody else will be making those

– David Packard

Responses to Questions (I)

the processes for planning for future (5-10 years) capabilities of your service? H
om do you gather input for this planning process and where does input typically co
v features have highest priority?

C is mission-oriented; meets with mission scientists to
ervice; VSO is user-oriented, using frequent outreach a
fic meetings to identify data provides and services to a

C new features are typically driven by the available
ology; VSO features represent what we develop (or ad
elf) in response to user community requests.

– David Packard

Responses to Questions (II)

ture(s) of your service would you like to stop performing? How do you gather in
ch decisions and where does input typically come from? What is preventing you

would truly, deeply like to get out of the business of
ing new storage on multiple tiers every five to seven y

t" is experience-based, by the facility scientist. Wastef
ue battling numerous, partially conflicting agency direc

of clearly effective alternatives that meet mission risk
ement criteria and a current period of confusion ("des

– David Packard

Responses to Questions (III)

ps you are taking to make your data interoperable with allied data sets from other
d out of NASA? How do you find allied data sets and what criteria make data sets
s for enabling interoperability?

data pretty are currently interoperable with any other
rn solar physics data (FITS format; software libraries in
thon). Interoperability with other scientific disciplines,
heliophysics, requires specialist knowledge, so not clea
er investing in further interoperability buys anything.

al Solar Observatory: all solar physics data, not just spa
SPDF + VxOs: access to heliophysics data

scientific analysis

ven the Virtual Solar Observatory (VSO)'s data diction
hich recognizes only 16 physical observables, is in a da
cience sense too rich

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

eal variety is in the solar features analyzed and the met sed to analyze them

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