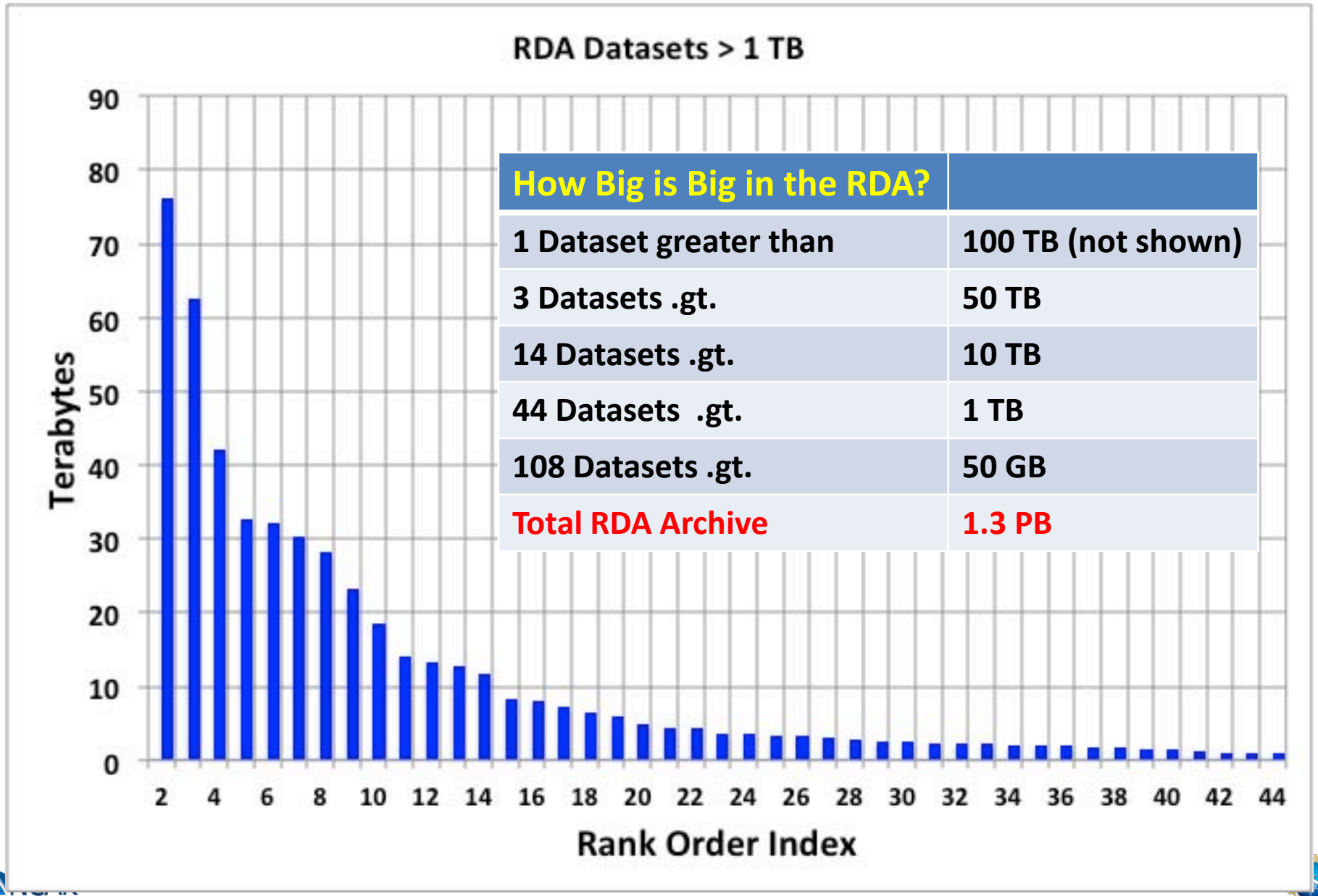
Big Data in the RDA
Datos Grandes

Steven Worley

Bob Dattore

Where we are now



How Big is Big in the RDA?	
1 Dataset greater than	100 TB (not shown)
3 Datasets .gt.	50 TB
14 Datasets .gt.	10 TB
44 Datasets .gt.	1 TB
108 Datasets .gt.	50 GB
Total RDA Archive	1.3 PB

Where we are now

Access

- **Structured file organization (time series & synoptic)**
- **On demand 'wget' scripts**
- **Point and click through a browser**
- **On demand 'cURL' scripts (web service)**

System driven access features	Datasets
Automated HPSS access (re-stage to online)	52
Format conversion (HPSS + online)	14
Subsetting (nearly all include format conversion)	38

Where we plan to be near-term

- Expand subsetting and format conversion
- Reduce subsetting turn around
  - Parallel data processing
- Collaborate across UCAR with distributed discovery and access tools
  - Make discovery more certain
  - Experiment with interoperable access (e.g. OPeNDAP)
- Manage dataset DOIs on a dynamic archive
  - Version control through various scenarios

Challenges for the longer-term

1. Federation
2. Security
3. Advanced web services
    a. Build translation services to meet API standard
4. Server-side data analysis and visualization
5. Data discovery with intended meaning – semantics
    a. “heat flux data sets derived from ICOADS R2.5”

End