Unidata

Providing data services, tools, & cyberinfrastructure leadership that advance Earth system science, enhance educational opportunities, & broaden participation

Data-Intensive Science and Scientific Data Infrastructure

Russ Rew, UCAR Unidata

ICTP Advanced School on High Performance and Grid Computing

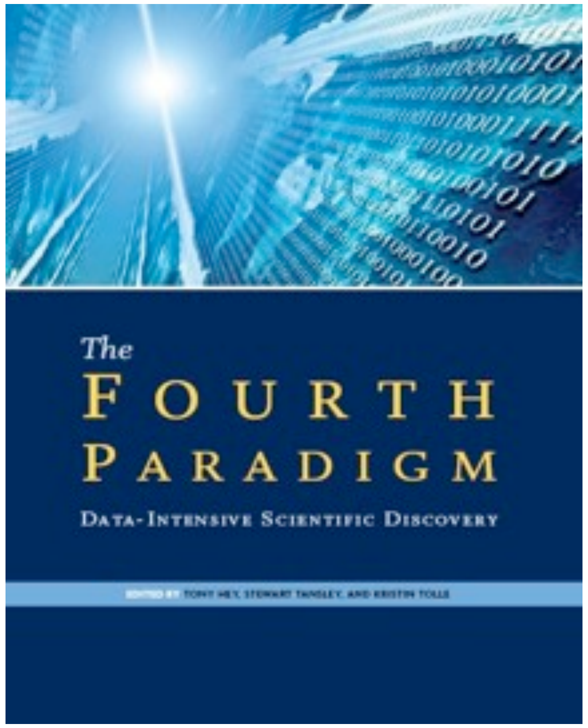
13 April 2011

Overview

- Data-intensive science
- Example: model outputs for IPCC AR5
- Publishing scientific data

Data-intensive science

- A "fourth paradigm" after experiment, theory, and computation
- Involves collecting, exploring, visualizing, combining, subsetting, analyzing, and using huge data collections
- Challenges include
  - Deluge of observational data, "exaflood" of simulation model outputs
  - Need for collaboration among groups, disciplines, communities
  - Finding insights and discoveries in a "Sea of Data"
- Data-intensive science requires
  - New tools, techniques, and infrastructure
  - Standards for interoperability
  - Institutional support for data stewardship, curation



Roles in Data-intensive Science

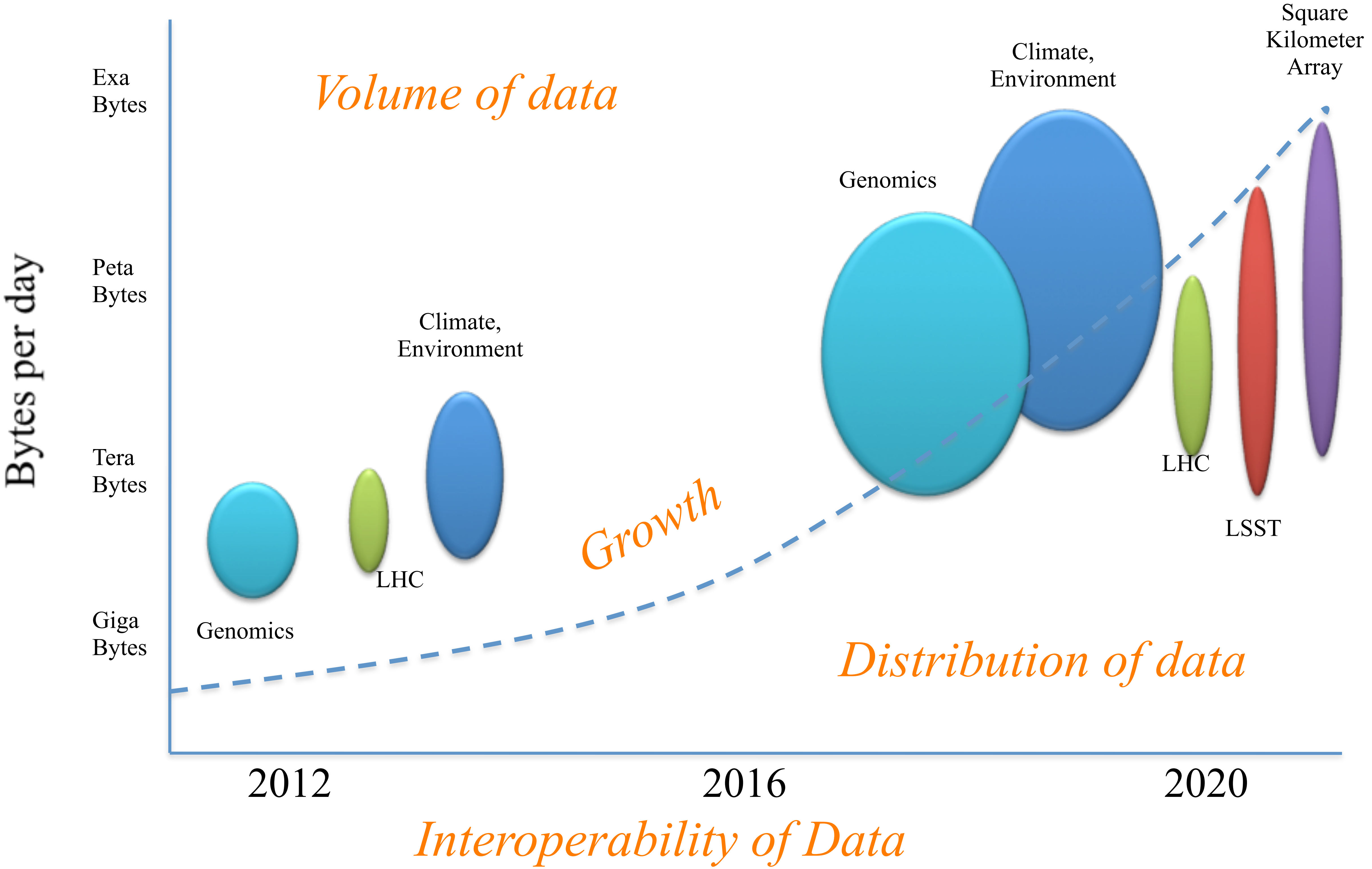
- **Scientists/researchers:** acquire, generate, analyze, check, organize, format, document, share, publish research data
- **Data users:** access, understand, integrate, visualize, analyze, subset, and combine data
- **Data scientists:** develop infrastructure, standards, conventions, frameworks, data models, Web-based technologies
- **Software developers:** develop tools, formats, interfaces, libraries, services
- **Data curators:** preserve data content and integrity of science data and metadata in archives
- **Research funding agencies, professional societies, governments:** encourage free and open access to research data, advocate elimination of most access restrictions

Growth in data from sensor systems

According to Science article *[2011-02-11, Baraniuk]*:

- Majority of data generated each year now comes from sensor systems
- Amount generated passed storage capacity in 2007
  - in 2010 the world generated 1250 billion gigabytes of data
  - generated data growing at 58% per year
  - storage capacity growing at 40% per year
- We generate more scientific sensor data than we can process, communicate, or store (e.g. LHC)

Data challenges: gigabytes to exabytes



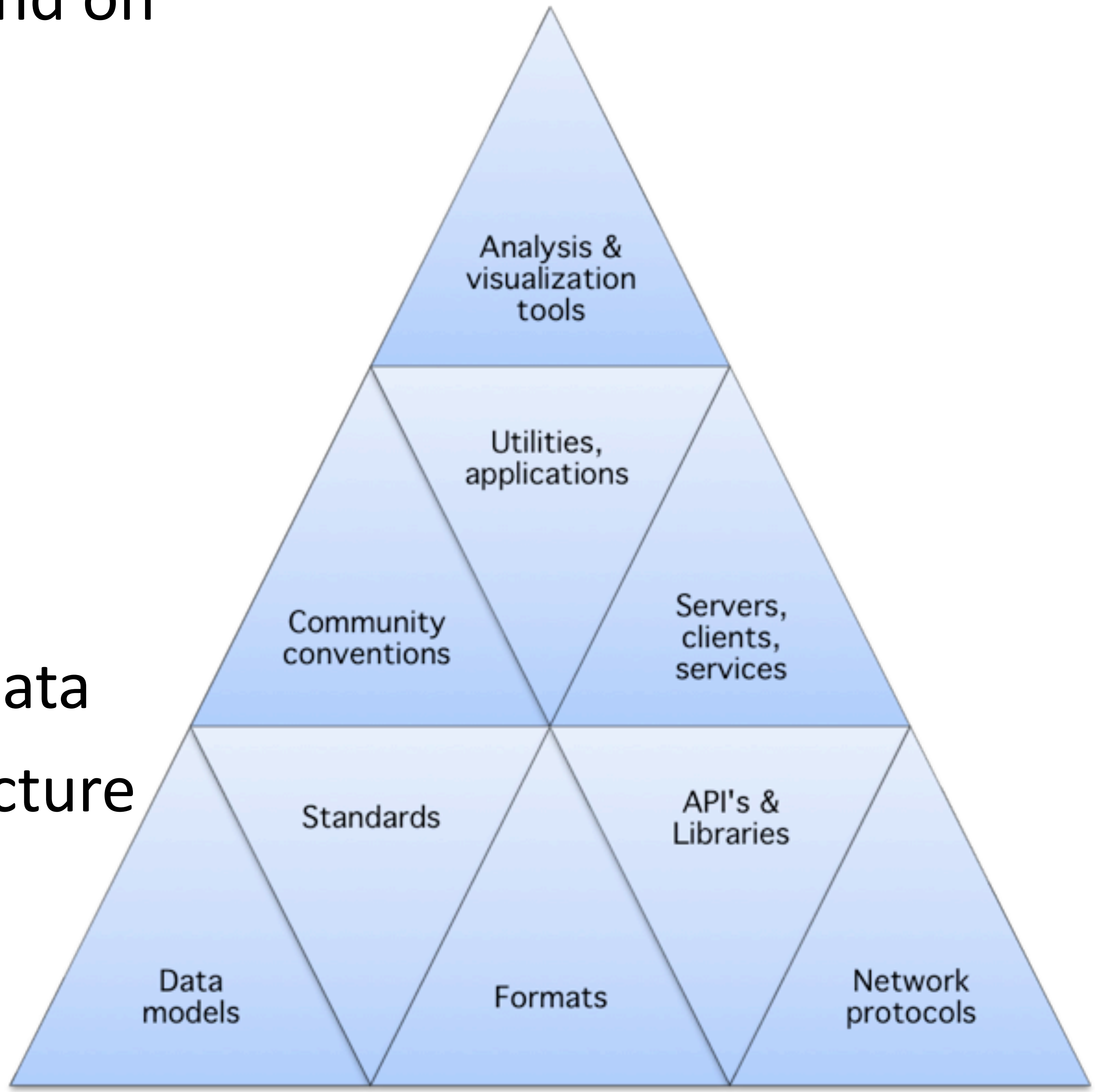
(slide from Tim Killeen, NSF)

Scalability and "Big Data"

- What's the big deal about big data?
  - aren't more and faster computers and larger disks the solution?
- I/O access and bandwidth can't keeping up with computing speed
- Too big to transfer, must move processing to data
- Sensors and models can generate huge datasets easily
- Making huge datasets accessible and useful is difficult
- Other problems: discovery, curation, provenance, organization, integrity, ...

Infrastructure for sharing scientific data

- Applications depend on lower layers
- Sharing requires agreements
  - formats
  - protocols
  - conventions
- Data needs metadata
- Is all this infrastructure really necessary?



Why not use binary I/O?

```
real   :: a(len), b(len)

write (nunit, rec=14) a
read  (nunit, rec=14) b
```

Simple, but …

- Not portable
- Lacks metadata for use, discovery
- Not usable by general analysis and visualization tools
- Inaccessible from other programming languages, for example reading Fortran binary data from Java or C/C++

Why not use formatted I/O?

```
real   :: a(len), b(len)

write (nunit, '(10f10.3') a
read  (nunit, '(10f10.3') b
```

Simple, but …

- Inefficient for large datasets (time *and* space)
- Sequential, not direct ("random") access
- Lacks metadata for use, discovery
- Not usable by general analysis and visualization tools

Why not use relational databases?

- Data model may not be appropriate
  - no direct support for multidimensional arrays
  - tables and tuples are wrong abstractions for model output, coordinate systems
- Tools: lacking for analysis and visualization
- Portability: difficult to share, publish, preserve, cite, database contents
- Performance
  - database row orientation slows access by columns
  - transactions unnecessary for most scienctific use
- But sometimes databases are ideal, e.g. virtual observatories

Other alternatives for scientific data

- XML, YAML, JSON, CSV, other text notations
  - Require parsing
  - Sequential, not direct access
  - Inefficient for huge datasets
  - Conversions between text and binary can lose precision
- Discipline-specific: FITS (astronomy), GRIB (meteorology), XMDF (hydrology, meshes), *foo*ML, …
- General-purpose, for scientific data:
  - CDF: historically one of the first, used in NASA projects
  - netCDF: widely used, simplest data model
  - HDF5: most powerful, most complex data model
  - SciDB: coming soon, multidimensional array-based database

An example: model outputs for IPCC AR5

What is the IPCC?

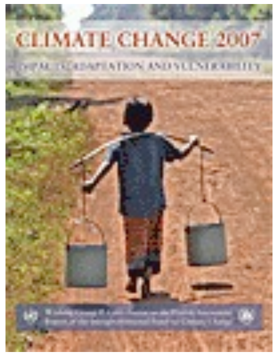
The Intergovernmental Panel on Climate Change

- 1990 - First Assessment Report
- 1995 - Second Assessment Report
- 2001 - Third Assessment Report
- 2007 - Fourth Assessment Report
- **2013 - Fifth Assessment Report**

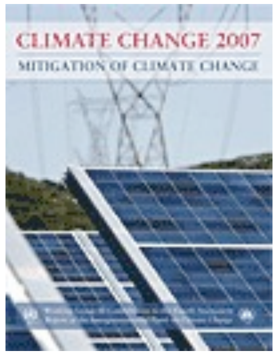
What was the IPCC AR4?

"The 4th Assessment Report of the Intergovernmental Panel on Climate Change"

Working Group I Report:
"The Physical Science Basis"



Working Group II Report:
"Impacts, Adaptation and Vulnerability"



Working Group III Report:
"Mitigation of Climate Change"

(slide from Gary Strand, NCAR)

What was the IPCC AR4?

The first large-scale coordination of climate modeling efforts, data analysis, data management and data dissemination by the global climate modeling community: 24 global coupled climate models from 16 modeling centers located around the world.

Types	**Purpose**	**runs**
"Control"	Assess model internal variability	3
CO2 increase	Determine climate sensitivity	4
20C3M	Simulate 20th century climate	14
SRES	Future scenarios (A1B, B1, A2, "commitment")	36
Other	Sensitivity and "idealized" Earths	6
Totals		**63**

<u>Unprecedented in scale and scope</u>

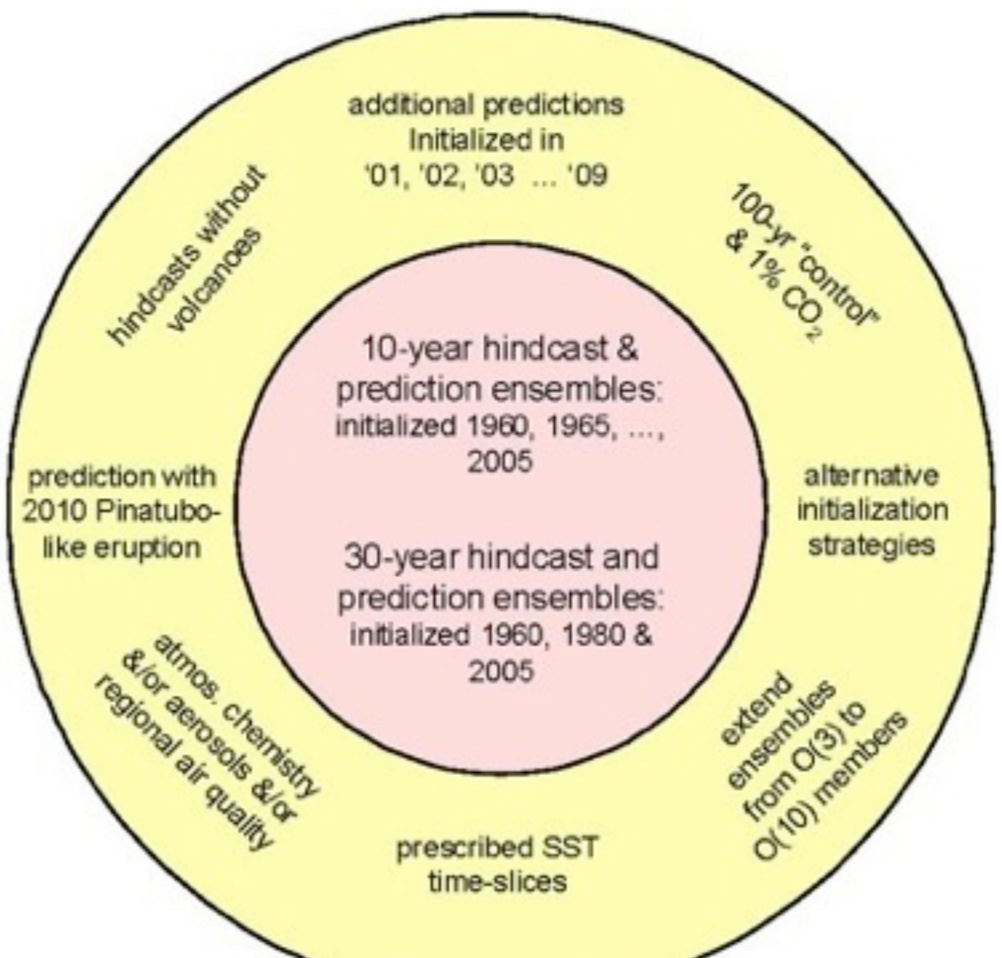
(slide from Gary Strand, NCAR)

What is the IPCC AR5?

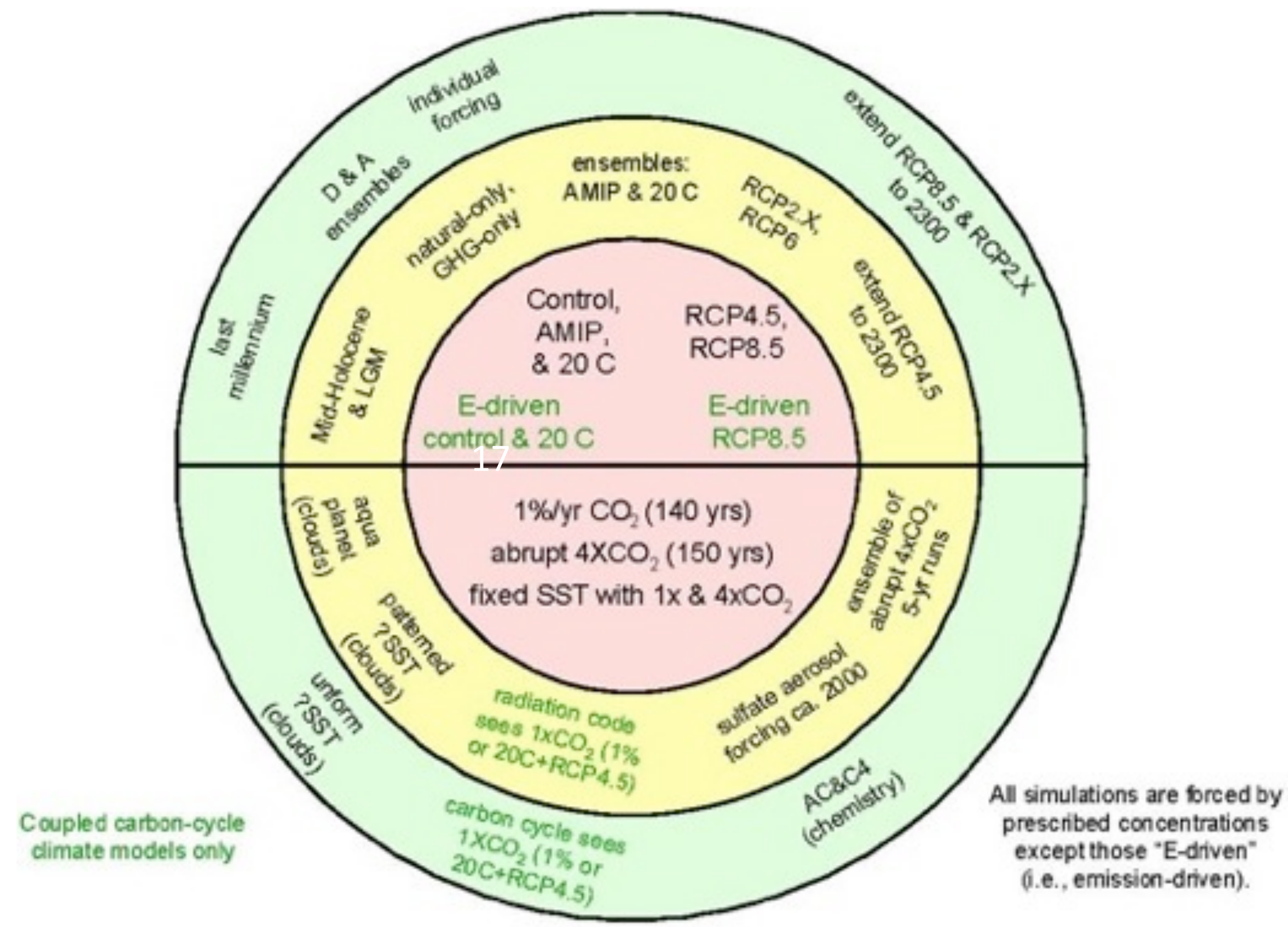
"The 5th Assessment Report of the Intergovernmental Panel on Climate Change"

The second large-scale coordination of climate modeling efforts, data analysis, data management and data dissemination by the global climate modeling community: 20+ global coupled climate models from 15+ modeling centers located around the world

<u>Many more experiments than AR4:</u>



Decadal Prediction Experiments



Long Term Experiments

(slide from Gary Strand, NCAR)

What is the IPCC AR5?

Types	Purpose	runs
"Control"	Assess model internal variability	3
CO2 increase	Determine climate sensitivity	2
20C3M	Simulate 20th century climate and variations	45
RCPs	Future scenarios (2.6, 4.5, 6, 8.5)	28
Paleo	Past climate (LGM, mid-Holocene, past 1000 years)	3
Decadal	Predictions (hindcast and forecast)	150
ESM	Earth System Model (BGC, carbon cycle, &c)	24
Other	Sensitivity and "idealized" Earths	30
Totals		**285**

Unprecedented in scale and scope

(slide from Gary Strand, NCAR)

Really much more data!

Modeling group		AR4 volume (GB)
NCAR	USA	9,200
MIROC3	Japan	4,000
GFDL	USA	3,800
IAP	China	2,900
MPI	Germany	2,700
CSIRO	Australia	2,100
CCCMA	Canada	2,100
INGV	Italy	1,500
GISS	USA	1,100
MRI	Japan	1,000
CNRM	France	1,000
IPSL	France	1,000
UKMO	UK	1,000
BCCR	Norway	900
MIUB	Germany/Korea	500
INMCM3	Russia	400
Totals		**35,200**

Modeling group		AR5 volume (GB)
MPI	Germany	710,000
NCAR	USA	410,000
MRI	Japan	312,000
GFDL	USA	151,000
MIROC3	Japan	115,000
UKMO	UK	89,000
CNRM	France	64,000
IAP	China	63,000
U Reading	UK	63,000
EC	Europe	50,000
GISS	USA	50,000
INGV	Italy	50,000
IPSL	France	45,000
INMCM3	Russia	32,000
NorClim	Norway	30,000
CCCMA	Canada	29,000
CAWCR	Australia	21,000
CSIRO	Australia	20,000
METRI	Korea	13,000
Totals		**2,317,000**

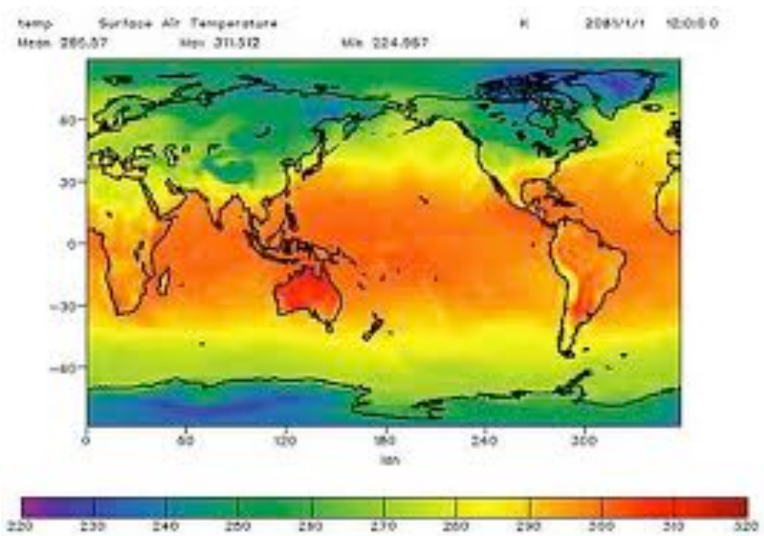
(slide from Gary Strand, NCAR)

Publishing scientific data: advice to data providers

Advice about unsolicited advice

- I could just present advice I think is needed
- I would rather listen, find out what is needed from data management infrastructure
- Consider the following the start of a dialogue

Don’t just provide pictures, provide data



http://www.some-archive.org/id3456/my-results/

- So your research can be reused by others in future research and analyses
- So your plots can be duplicated and integrated with other data
- So users can choose their favorite display and analysis software for your data
- So corrections to data are practical
- So your results have a longer shelf life

Don’t just make data available interactively

- Programs need access to data, not just humans
- Accessing lots of data by mouse clicks or display touching is difficult and slow
- Provide bulk access for large datasets
- Anticipate need for programs to access data remotely

Support efficient access to small subsets of data

- Database queries should return only requested data
- Don't provide only huge files with all the data, that discourages reuse
- Remote access is faster for small subsets
- Interactive visualization integrating data from multiple sources is practical with small subsets
- Some problems require a little data from many places, not a lot of data from one place

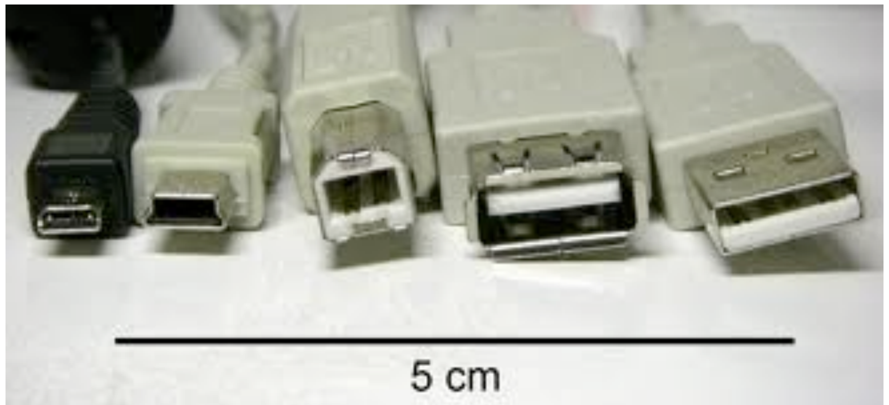
Provide easy access to metadata

- More metadata is usually better
- Make it easy to add more metadata later
- Keep metadata with the data, if practical
- Support *discovery metadata*, so your data can be found
- Support *use metadata*, so your data can be understood
  - coordinate systems
  - units

Strive for interoperability

- Data should be portable now
- Data should be portable to the future
- Don’t optimize packaging or format for specific data or application
- Valuable scientific data is written once, read many times

Support standards



- If available, use them
- If not, help develop them
- If possible, help maintain them

Summary: What Data Producers Should Provide

- Data (not just visualizations)
- Useful metadata (not just data)
- Remote access (not just physical copies or local access)
- Convenient granularities of access (not too large or too small)
- Program access (not just for interactive users)
- Standard formats (not machine-, application-, or language-specific; but what about discipline-specific?)
- Organization for users and readers (not just what's most convenient for provider)

But scientists want to do science ...

- ... not data management
- Valuable scientific data must be acquired, organized, accessed, visualized, distributed, published, and archived
- How can scientists do all this and still have time to do science?
  - graduate students?
  - data managers, curators, stewards, ...?
  - database systems?
  - general purpose scientific data infrastructure?
- Standards supported by open source software may help:

netCDF The HDF Group SciDB

Questions / Discussion