History of Data (and Information), Data Science, Current Challenges

Peter Fox

Data Science – ITEC/CSCI/ERTH-4350/6350

Week 1, August 27, 2013

Admin info (keep/ print this slide)

- Class: ITEC/CSCI/ERTH-4350/6350 (note new number)
- Hours: 9am-11:50am Tuesday
- Location: Lally 104
- Instructor: Peter Fox
- Instructor contact: pfox@cs.rpi.edu, 518.276.4862 (do not leave a msg)
- Contact hours: Monday** 3:00-4:00pm (or by appt)
- Contact location: Winslow 2120 (or Lally 207A)
- TA: Saurabh Sharma; sharms3@rpi.edu
- Web site: http://tw.rpi.edu/web/courses/DataScience/2013
  - Schedule, lectures, syllabus, reading, assignments, etc.

Contents

- Intro – about this course
- Learning objectives
- Outline of the course
- Definitions/ Current Challenges
- ~ History of data and information
- Information Science -> Data “Science”
- Paradigm shifts and a few examples
- What skills are needed
- What is expected

Assessment and Assignments

- Via written assignments with specific percentage of grade allocation provided with each assignment
- Via individual oral presentations with specific percentage of grade allocation provided
- Via group presentations – depending on class size
- Via participation in class (not to exceed 10% of total, start with 10% and lose % by not participating)
- Late submission policy: first time with valid reason – no penalty, otherwise 20% of score deducted each late day

Assessment and Assignments

- Reading assignments
  - Are given almost every week
  - Most are background and informational
  - Some are key to completing assignments
  - Some are relevant to the current week's class (i.e. follow up reading)
  - Others are relevant to following week's class (i.e. pre-reading)
  - Will be discussed in class and participation in these discussions is taken into account
- You will progress from individual work to group work

Project options (examples)

- Data Collection and Management
- Data Models and Metadata
- Data Standards, Policy, Fair use and Rights
- Tool Use and Evaluation
- Data Life-Cycle Studies
- Data and Information Product Generation
- Science with someone else's Data

Objectives

- To instruct future scientists how to sustainably generate/ collect and use data for their research as well as for others: data science
- To instruct future technologists how to understand and support essential data and information needs of a wide variety of producers and consumers
- For both to know tools, and requirements to properly handle data and information
- Will learn and be evaluated on the full life-cycle of data and relevant methods, technologies and best practices

Learning Objectives

- Through class lectures, practical sessions, written and oral presentation assignments and projects, students should:
  - Develop and demonstrate skill in Data Collection and Management
  - Develop Data Models and Generate Metadata
  - Demonstrate Knowledge Application of Data and Metadata Standards
  - Demonstrate Skill in Data Science Tool Use and Evaluation
  - Demonstration the Application of the Data Life-Cycle principles
  - Become Proficient in Data and Information Product Generation

Undergraduates/ Grads

- Graduate students are assessed at:
  - Higher level of demonstration
  - Additional questions or tasks in assignments
- Undergraduates are welcome to complete these higher requirements to extra grade
- Extra points for outstanding/ above and beyond are given**

Academic Integrity

- Student-teacher relationships are built on trust. For example, students must trust that teachers have made appropriate decisions about the structure and content of the courses they teach, and teachers must trust that the assignments that students turn in are their own. Acts, which violate this trust, undermine the educational process. The Rensselaer Handbook of Student Rights and Responsibilities defines various forms of Academic Dishonesty and you should make yourself familiar with these. In this class, all assignments that are turned in for a grade must represent the student's own work. In cases where help was received, or teamwork was allowed, a notation on the assignment should indicate your collaboration. Submission of any assignment that is in violation of this policy will result in a penalty. If found in violation of the academic dishonesty policy, students may be subject to two types of penalties. The instructor administers an academic (grade) penalty, and the student may also enter the Institute judicial process and be subject to such additional sanctions as: warning, probation, suspension, expulsion, and alternative actions as defined in the current Handbook of Student Rights and Responsibilities. **If you have any question concerning this policy before submitting an assignment, please ask for clarification.**

Current Syllabus/Schedule

- Week 1 (Aug. 27): History of Data and Information, Data, Information, Knowledge Concepts and State-of-the-Art
- Week 2 (Sep. 3): Data and information acquisition (curation, preservation) and metadata - management
- Week 3 (Sep. 10): Data formats, metadata standards, conventions, reading and writing data and information
- Week 4 (Sep. 17): Class exercise - collecting data - individual
- Week 5 (Sep. 24): Class Presentations: present your data I
- Week 6 (Oct.1) : Class Presentations: present your data II
- Week 7 (Oct. 8): - Data Analysis and Data Mining/ Class exercise - group project - working with someone else's data
- Oct. 15: no classes (Tuesday follows Monday schedule)
- Week 8 (Oct. 22): Guest lectures in Data Science
- Week 9 (Oct. 29): Week 12 (Nov. 19): Webs of Data and Data on the Web, the Deep Web, Data Discovery, Data Integration, Data Citation
- Week 10 (Nov. 5): Academic basis for Data Science, Data Models, Schema, Markup Languages and Data as Service Paradigms
- Week 11 (Nov. 12): Data Workflow Management and Data Stewardship
- Nov. 19: study week, work on projects.
- Week 13 (Nov. 26): Data Quality, Uncertainty and Bias – exploiting provenance
- Week 14 (Dec. 3): Final Project Presentations

Questions so far?

Introductions

- Who you are, background?
- Why you are here?
- What you expect to learn?

So what are we talking about?

http://images2.fanpop.com/image/photos/9400000/Lt-Commander-Data-star-trek-the-next-generation-9406565-1694-2560.jpg

Definitions (at least for this course)

- Data - are encodings that represent the qualitative or quantitative attributes of a variable or set of variables.
- Data (plural of "datum", which is seldom used) - are typically the results of measurements, computations, or observations and can be the basis of graphs, images of a set of variables.
- Data - are *often* viewed as the lowest level of abstraction from which information and knowledge are derived***

How do we ‘get’ data?

- Transduction: (e.g. for a signal)
  - any process by which a biological cell converts one kind of signal or stimulus into another (wikipedia)
- Example:
  - Thermometer, watch, calculator…

Definitions ctd.

- Information
  - Representations (of facts? data?) in a form that lends itself to human use
- Knowledge
  - Check out Wikipedia…. meaning
- Metadata – data about data
- Metainformation – information about information
- Data documentation – integrated collection of information and metadata intended to support all aspects of data (find, access, use…)

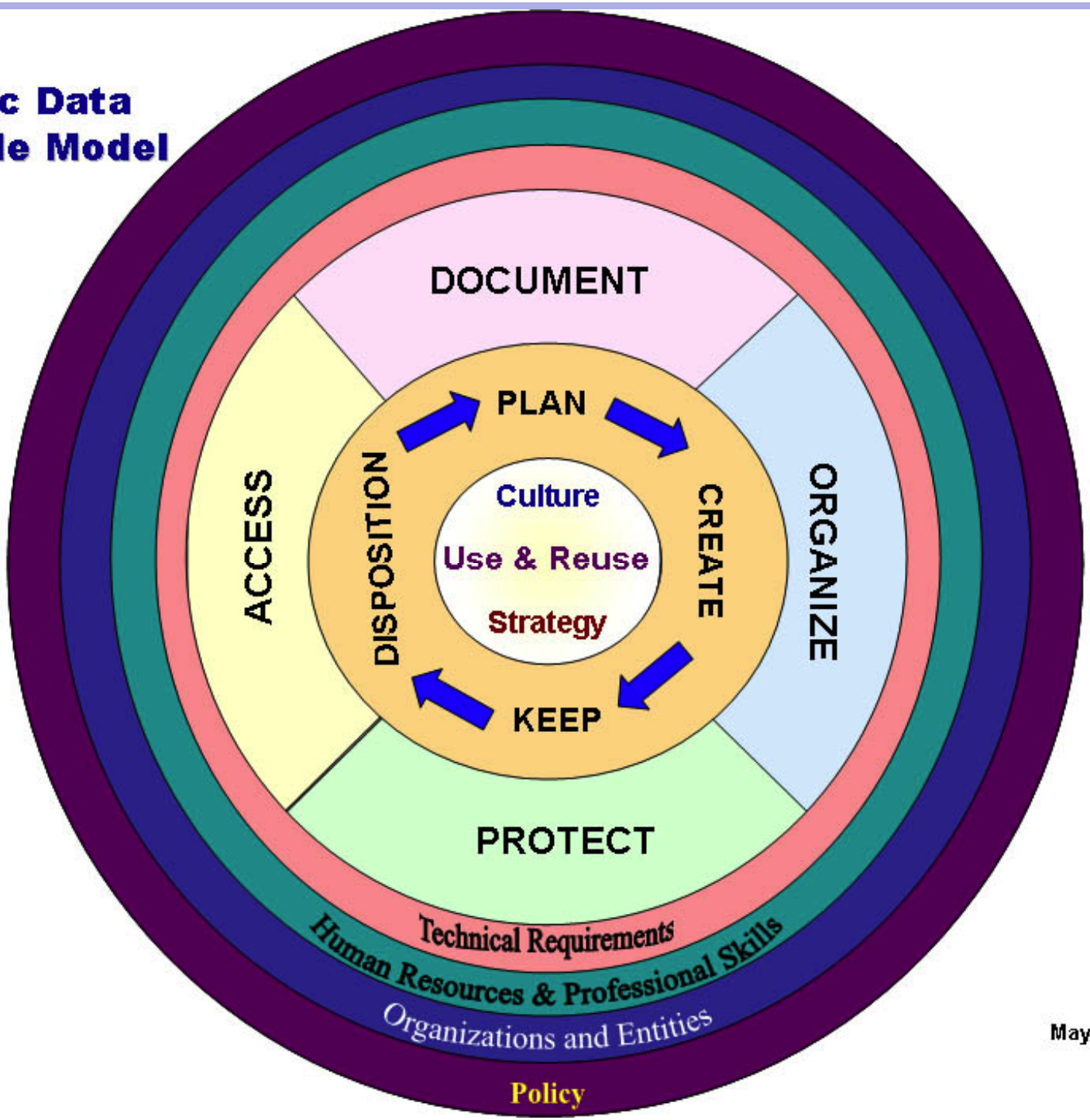
Examples

- Rock sample:
  - Data – weight, composition, shape, size
  - Information – images of the rock as collected
  - Knowledge – evidence of geologic activity
  - Metadata – location and time of collection
  - Documentation – published lab report …
- Weather
  - Data – wind speed and direction, temperature, ..
  - Information – weather map with contours and features
  - Knowledge – high pressure system, stable weather
  - Metadata – type of radar, sensor, use of model

Definitions ctd.

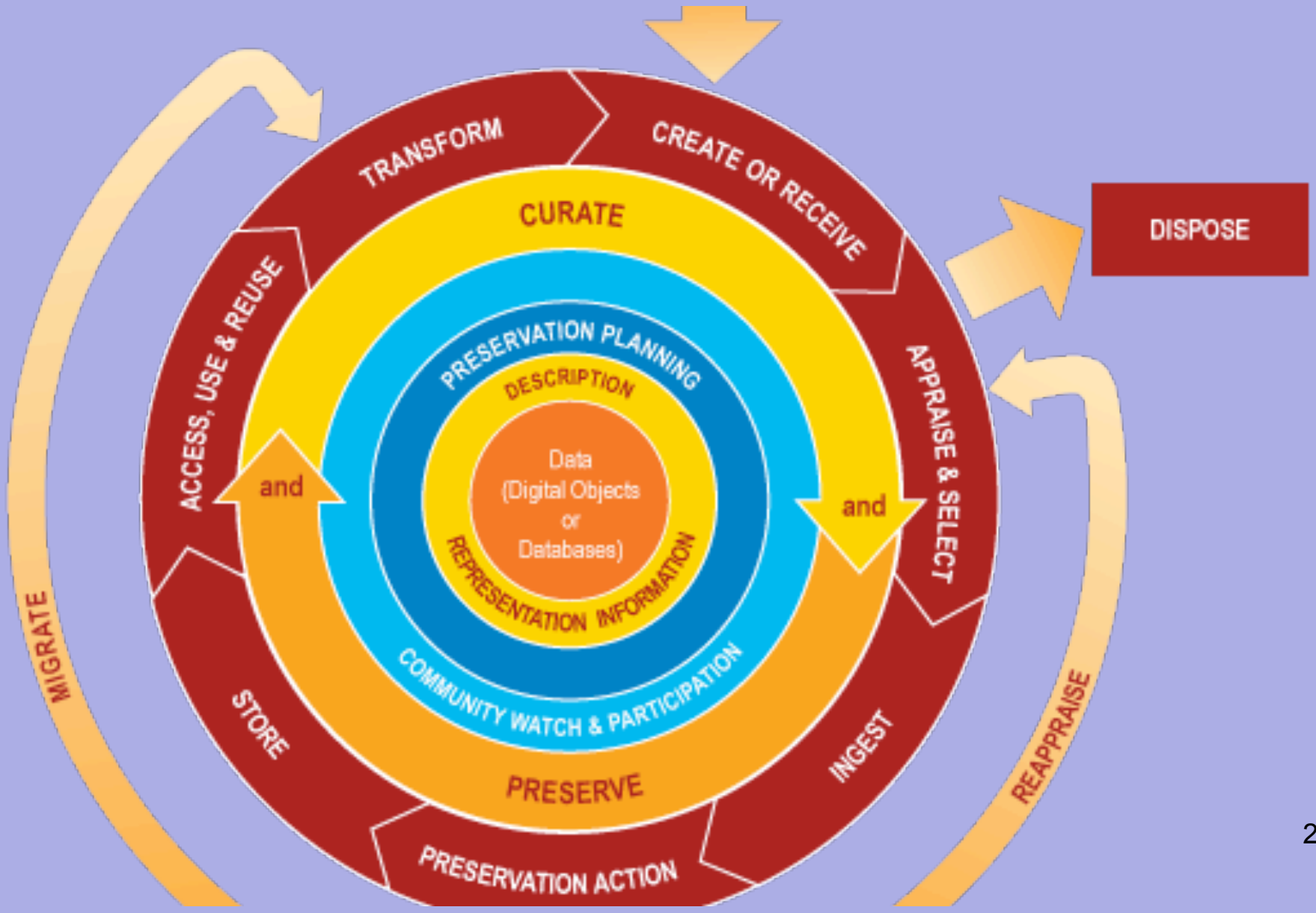
- Data life-cycle elements -
  - Acquisition: Process of recording or generating a concrete artefact from the concept (see transduction)
  - Curation: The activity of managing the use of data from its point of creation to ensure it is available for discovery and re-use in the future (http://www.dcc.ac.uk/FAQs/data-curator)
  - Preservation: Process of retaining usability of data in some source form for intended and unintended use
- Stewardship: Process of maintaining integrity for acquisition, curation and/ or preservation

IWGDD Scientific Data Life Cycle Model

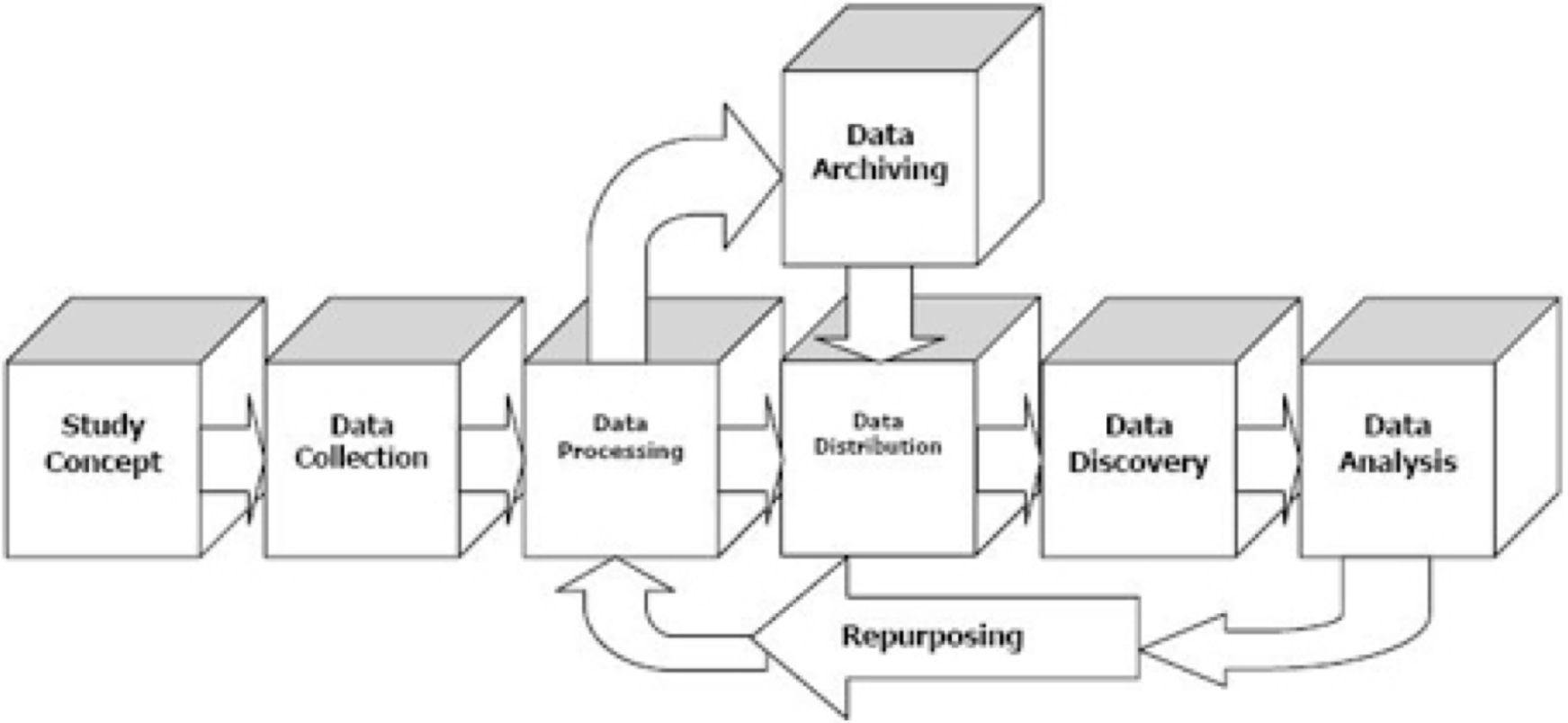


May 2008

Digital Curation Centre



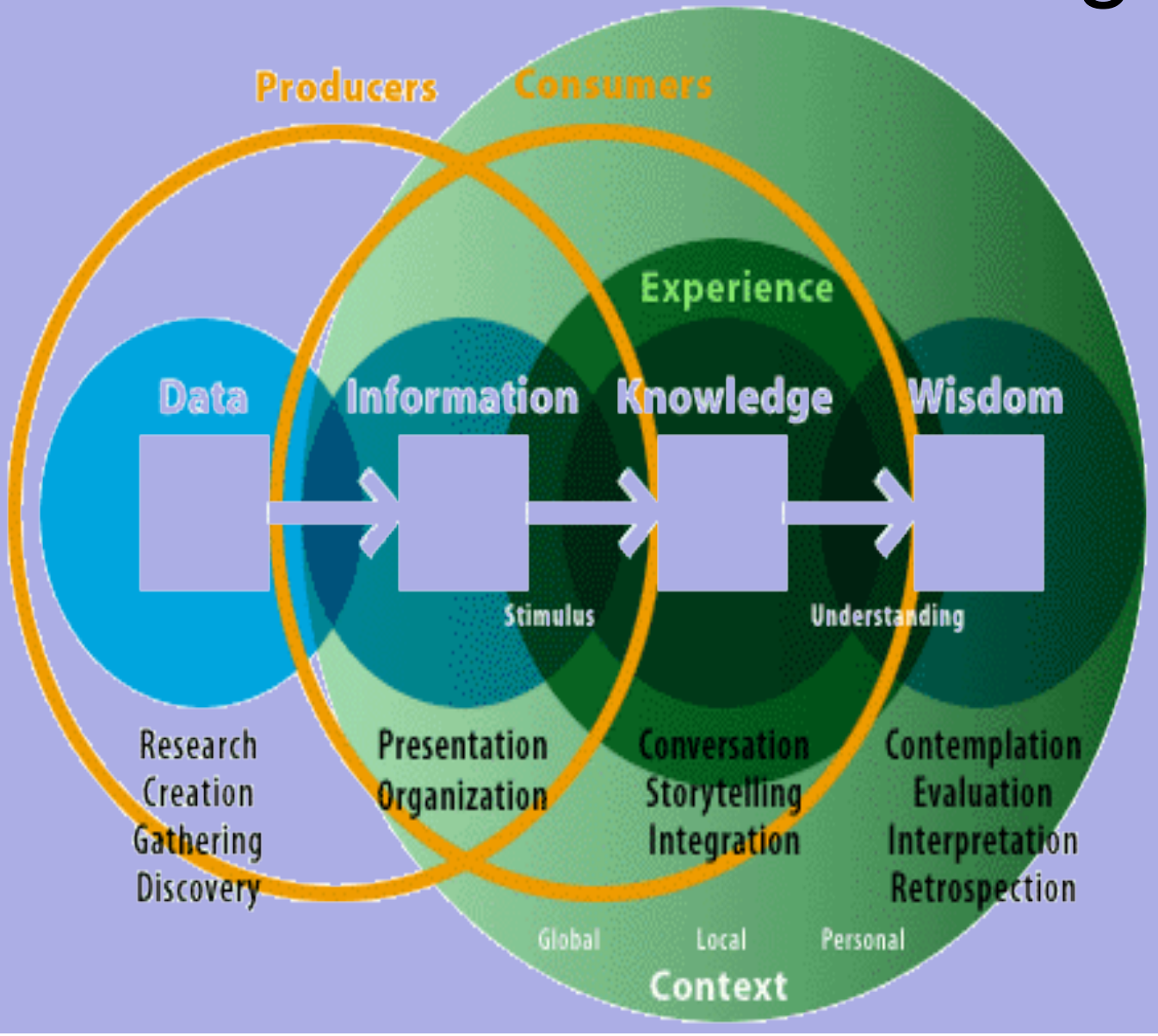
MIT DDI Alliance Life Cycle



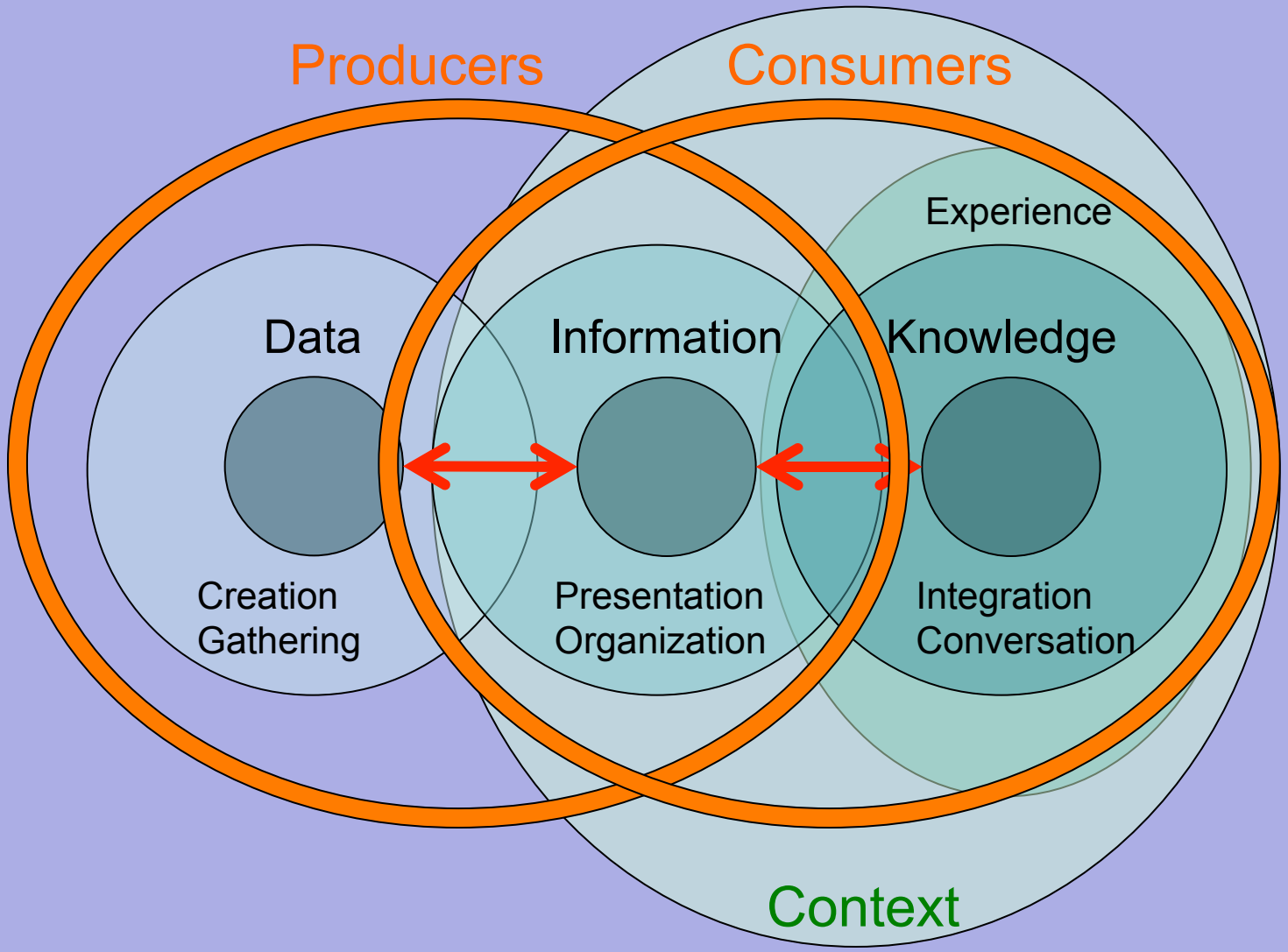
Definitions ctd.

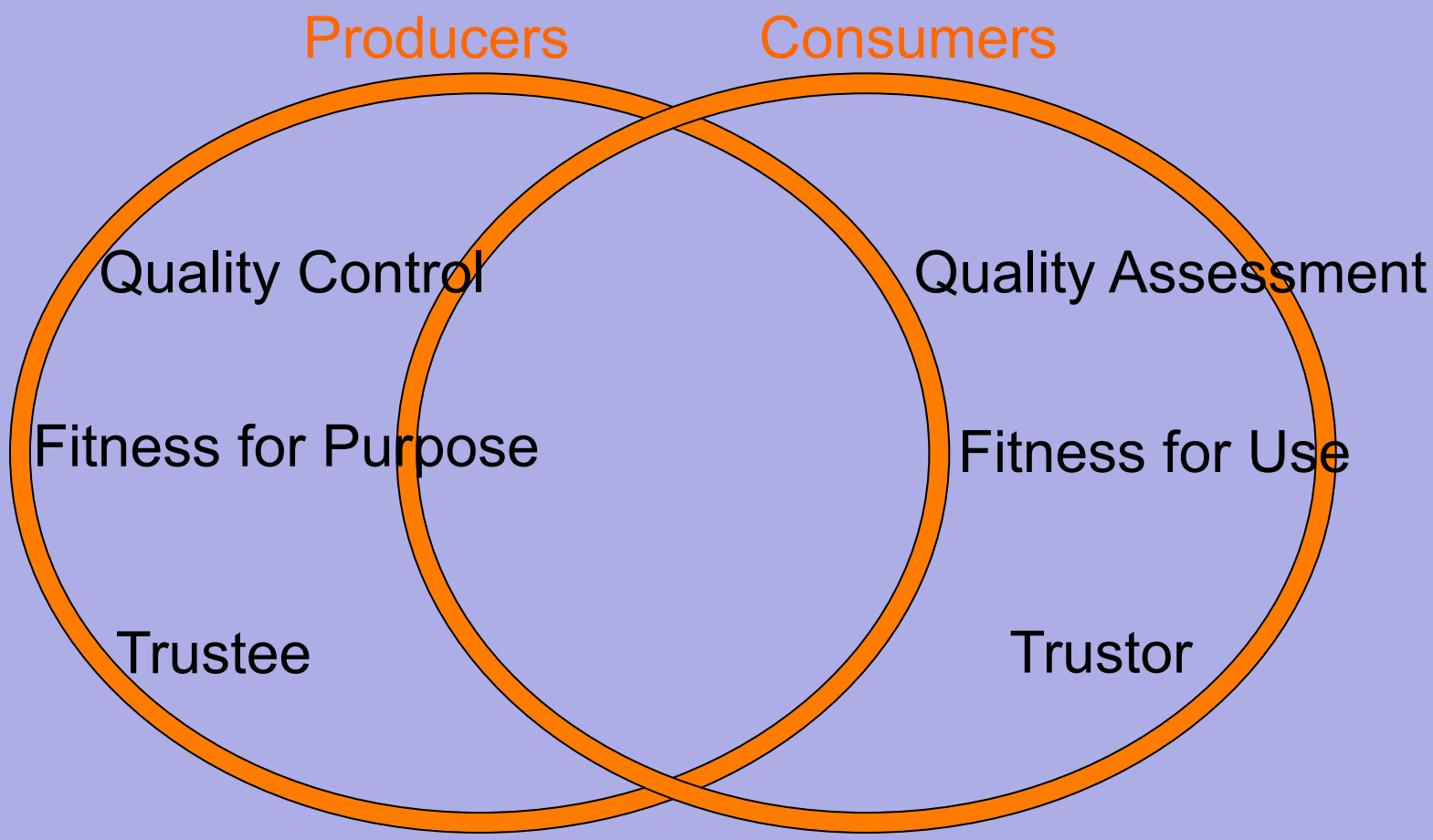
- Management: Process of arranging for discovery, access and use of data, information and all related elements.
- Also oversees or effects control of processes for acquisition, curation, preservation and stewardship.
- Involves fiscal and intellectual responsibility.

An OLD view of data, information and knowledge

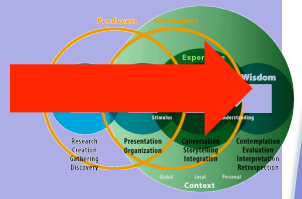


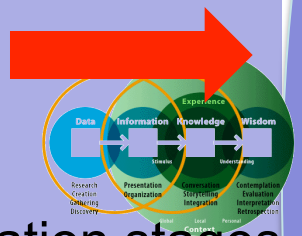
Data-Information-Knowledge Ecosystem

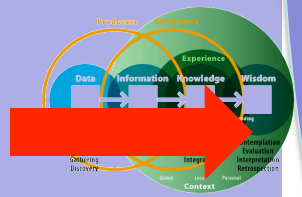


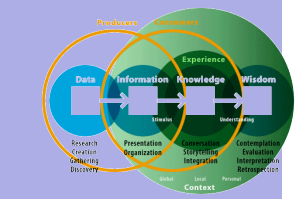

Producers
Consumers
Quality Control
Quality Assessment
Fitness for Purpose
Fitness for Use
Trustee
Trustor

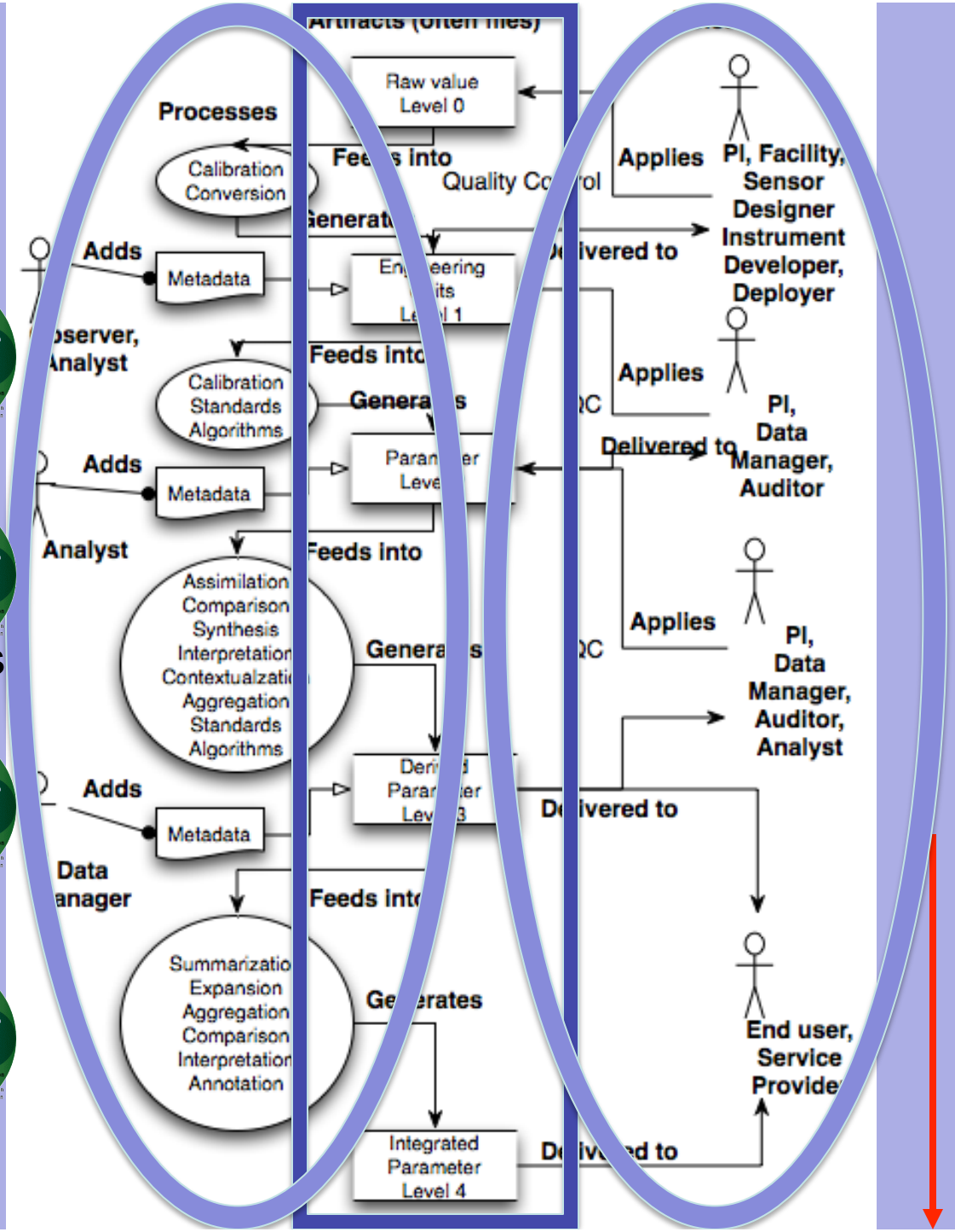
Curation stages











Need to be here

Often ‘get’ the (meta) data here

Too many ‘get’ it here

The nature of the challenge

- To do data science today
  - You may play many roles in different parts of the life cycle, or all of them
  - You may not get all the metadata or information you need even if you get the data
  - You will need skills that you were not taught
- To work with scientists today
  - You may have lots of technical experience
  - You will need new skills in addressing the changing use of data and information
  - One ‘size’ does not fit all
- Some examples follow:

HYDROLOGIC DATA
Data from: http://pubs.usgs.gov/dds/dds-81/#Database

Date	SODIUM, DIS-SOLVED (MG/L AS NA) (00930)	ANC UNFLTRD TIT 4.5 LAB (MG/L AS CACO3) (90410)	ANC WATER UNFLTRD FET FIELD MG/L AS CACO3 (00410)	CHLO-RIDE, DIS-SOLVED (MG/L AS CL) (00940)	FLUO-RIDE, DIS-SOLVED (MG/L AS F) (00950)	SILICA, DIS-SOLVED (MG/L AS SIO2) (00955)
APR 1986						
10...	380	--	628	150	4.9	190
10...	--	--	--	150	--	200

Physical quantity versus measured as quantity

METEOROLOGICAL DATA NEAR LONG VALLEY CALDERA:
http://www.wrh.noaa.gov/sto/getRaws.php?sid=CRVC1&num=48

Date	Time (PDT)	Wind (mph)	Temperature		RH (%)	Precipitation Accumulation (inches)
			Air (°F)	Dewpoint (°F)		
24	12:53	WSW 4 G 10	77	42	29	63.76

Value and units?

SEISMICITY DATA
http://pubs.usgs.gov/dds/dds-81/Intro/MonitoringData/Earthquakes/earthquakes.html

Date	Time	Lat	Lon	Depth	Mag
1/30/1975	22:53.3	37.6117	-118.6425	4.59	3.0
7/4/1975	13:54.8	38.0545	-118.6727	8.34	3.2
8/17/1975	24:26.8	37.5947	-118.7993	6.06	4.2
8/21/1975	36:18.4	37.6193	-118.7910	4.99	3.6

HEAT FLOW DATA:
http://earthquake.usgs.gov/heatflow/Data/all_by_code.html

H.F.	Code	State	Latitude	Longitude	Elev(m)	Depth(m)	H.F
91	ABT	CA	37 51.6	119 04.5	2238	124	3.7

Reference frame?

GRAVITY DATA
http://pubs.usgs.gov/dds/dds-81/Intro/MonitoringData/Gravity/gravity.html

STATION	X (UTM)	Y (UTM)	SEPT 1982	StDev
12DORT5	329558.81	4171897.32	979242.333	0.005
12JCM82	330078.01	4175261.24	979250.95	0.006
15JCM82	332778.6	4174484.55	979234.052	0.006
16JCM82	333520.47	4173659.22	979253.378	0.003

Value and units?

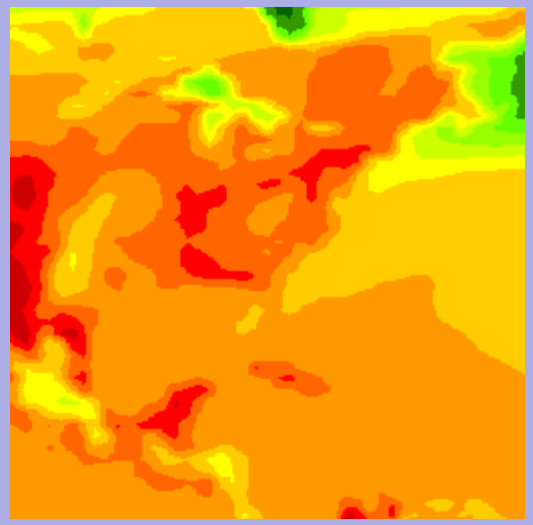
GPS Data
http://pubs.usgs.gov/dds/dds-81/Intro/MonitoringData/Geodetic/GPS/GPS.html

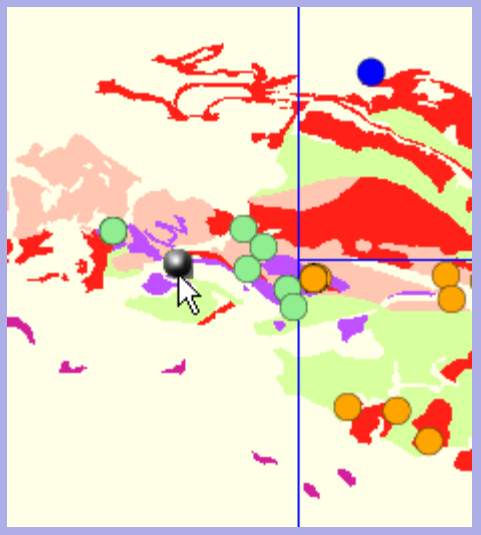
Site Name	Lon DDEG	Lon DEG	Lon MIN	Lon SEC	Lat DDEG	Lat DEG	Lat MIN	Lat SEC	Height
Bald_cont	-118.901	-118.000	54.000	5.004	37.783	37.000	47.000	0.340	2749.505

Reference units?

An example of heterogeneity in databases from Long Valley caldera that requires Semantic (ontologic) registration
Compiled for SESDI project

Fields vs. objects





classic geophysics
"Coverage" viewpoint

- simple data structures
- collated/gridded ready for analysis

netCDF, HDF-EOS

classic geology
"Feature" viewpoint

- complex data
- database insertion
- complete feature interpretations

XML documents

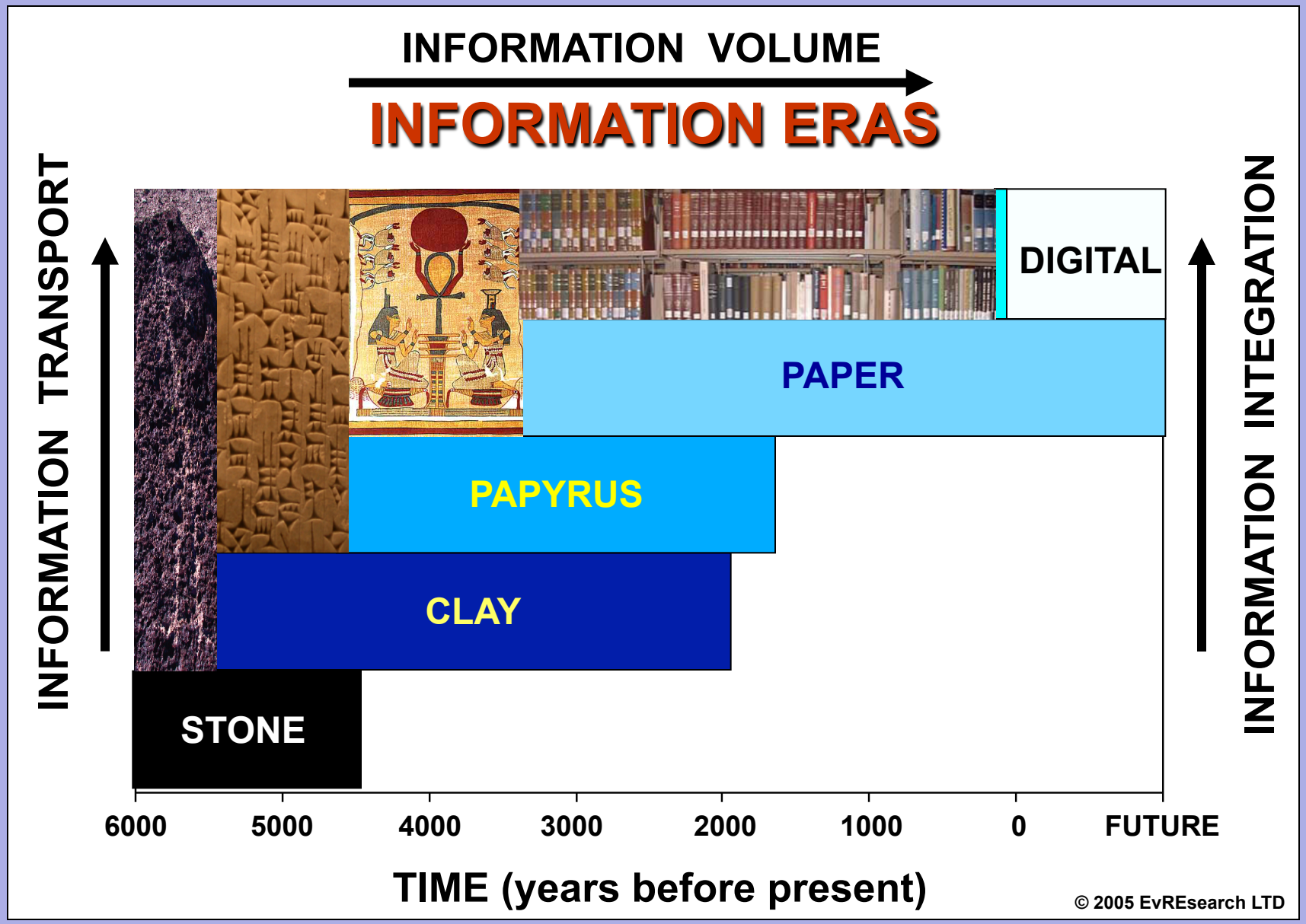
Can you answer this use case?

- I know that the data in Fig. 3.1 (doi: 10.1001/2010z.f3.1) in the paper by Omega et al. 2010 (doi:10.1001/2967q) is published but did it use the corrections to the underlying level 2 data by Alpha (unpublished)?

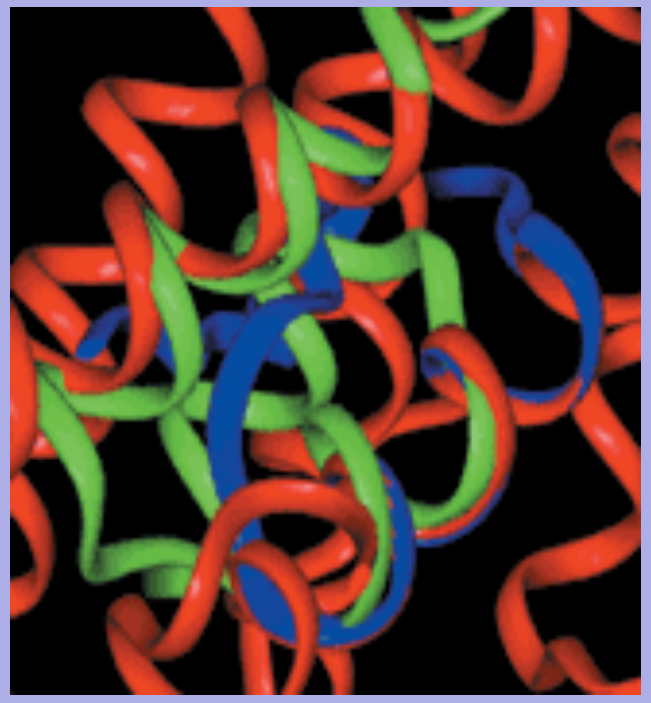
History

- Observation of our natural world
  - Kepler's laws of planetary motion
  - Aurora and sunspots
  - Cosmic microwave background
- Experiment
  - Michelson-Morley and the search for ether
  - Particle physics colliders
- Difference between these?
  - Never repeating versus repeatable
- Recording means is changing 'improving' - digital

HISTORY OF INFORMATION THRESHOLDS



The Information Era: Interoperability



Modern information and communications technologies are creating an "interoperable" information era in which ready access to data and information can be truly universal. Open access to data and services enables us to meet the new challenges of understand the Earth and its space environment as a complex system:

- managing and accessing large data sets
- higher space/time resolution capabilities
- rapid response requirements
- data assimilation into models
- crossing disciplinary boundaries.

History (ctd)

- Theory
  - Leads to need for observations, experiments
- Simulation
  - Leads to simulate observations, experiments
- Data Science is frequently called the FOURTH PARADIGM

- Data assimilation
  - Combines data, theory and models

OBSERVATIONS

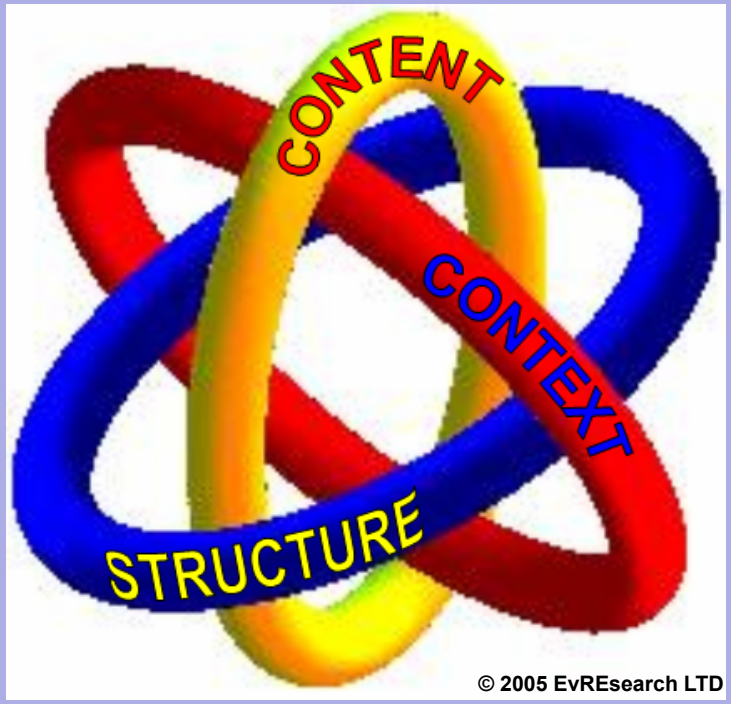
Information has content, context and structure. The notion of "unstructured" really means information that is "unmanaged" with conventional technologies (i.e., metadata, markup and databases).

Databases, metadata and markup provide control for managing digital information, but they are not convenient. This is why less than 20% of the available digital records are managed with these technologies (i.e., conventional technologies are not scalable).

Search engines are extremely convenient, but provide limited control for managing digital information (i.e., long lists of ranked results conceal relationships within and between digital records). The search engine problem is that accessing more information does not equal more knowledge.

We already have effectively infinite and instantaneous access to digital information. The challenge is no longer access, but being able to objectively integrate information based on user-defined criteria independent of scale to discover knowledge.

THE PHYSICS OF *INFORMATION*



BORROMEAN RINGS

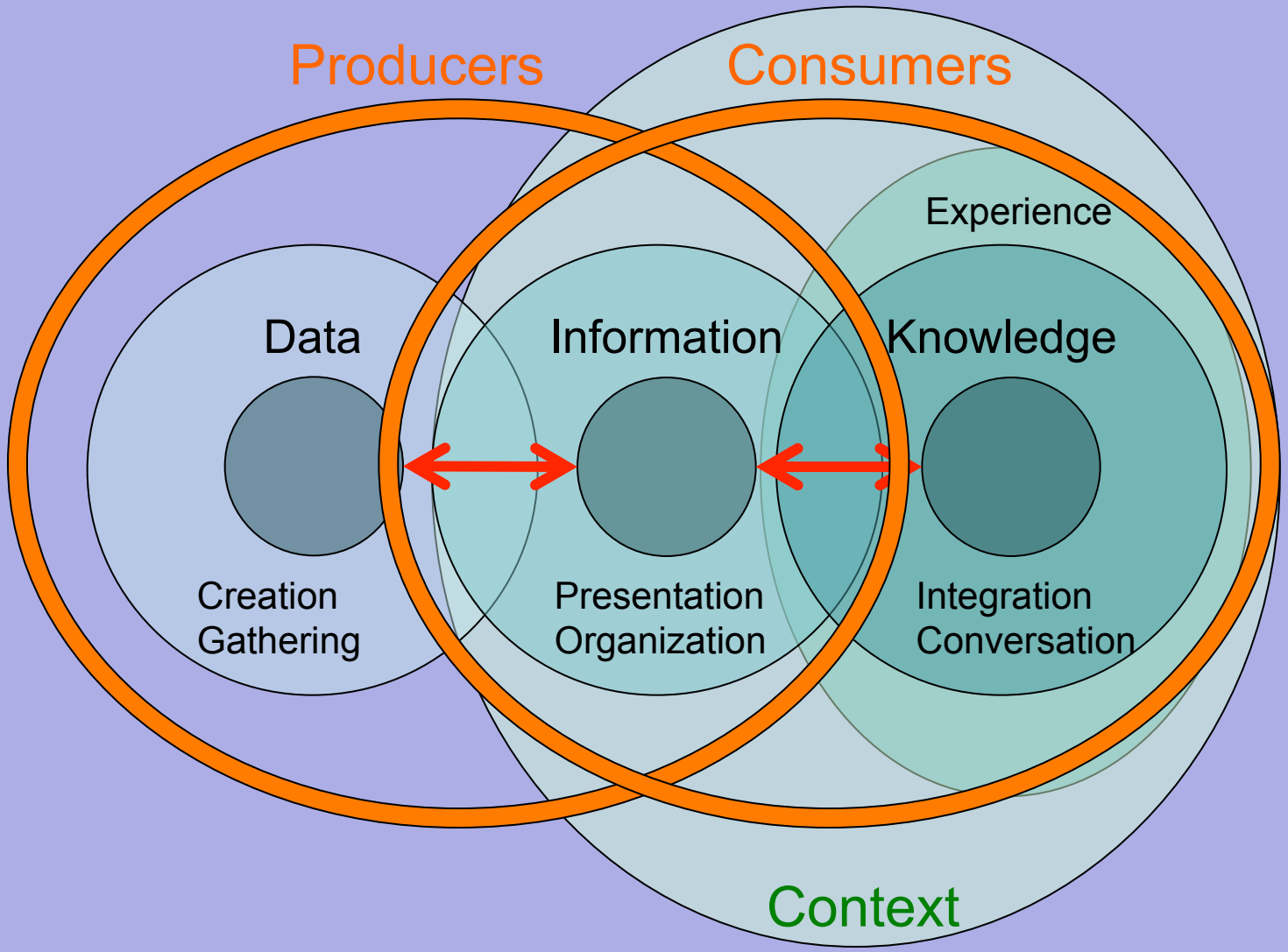
Three interlinked circles that represent inseparable parts of the whole. Remove any one ring and the other two fall apart. Because of this property, Borromean Rings have been used as a symbol of unity in many fields.

- **Information has three indivisible ingredients – *content, context and structure.***
- **The ability to automatically utilize the inherent structure of information is the threshold in information management from hardcopy to digital media.**

Later on in the course we will

- Start to tackle the <pretty small> number of theoretical ideas about data
- If you took Xinformatics, you will see some familiar ideas but also a few new ones
- Until then, there's more "practice" than "theory"

Data-Information-Knowledge Ecosystem



So What's the Fuss?

- International attention
  - Strategic Coordinating Committee for Information and Data (SCCID) formed by the International Council for Science (2008) http://www.icsu.org/1_icsuinscience/DATA_SCCID_1.html
  - Electronic Geophysical Year eGY, http://www.egy.org
  - Committee on Data for Science and Technology: CODATA, http://www.codata.org (~1959)
- National Academy and Research Council
  - Networking Information Technology Research and Development: NITRD
  - Board on Research Data and Information: BRDI

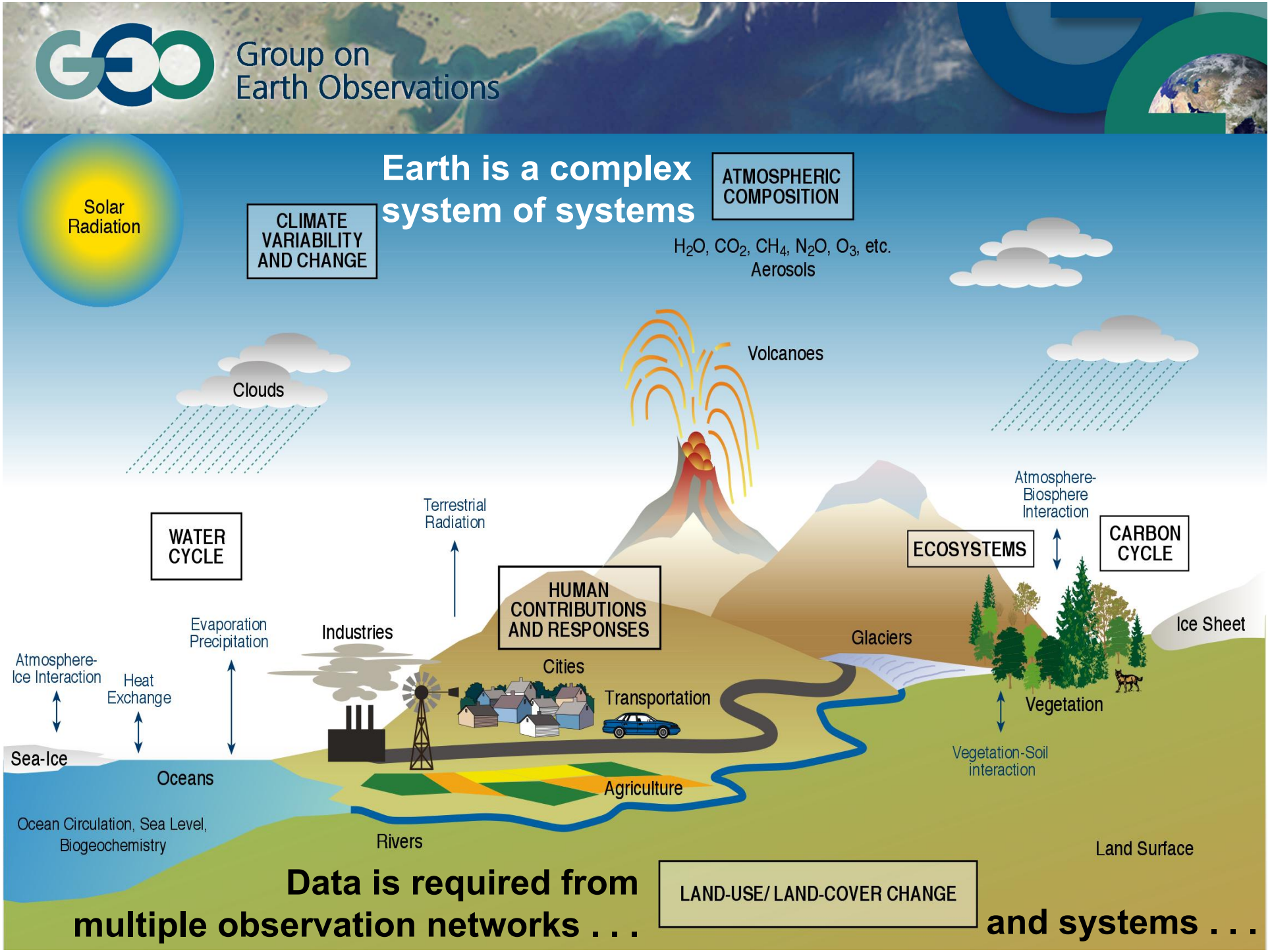
So What's the Fuss?

- U.S. Agencies: all of them have reports on the importance of data, some even have data policies and some (less) are enforcing those policies
  - National Science Foundation: NSF
  - National Aeronautic and Space Administration: NASA
  - National Oceanographic and Atmospheric Administration: NOAA
  - Environmental Protection Agency: EPA
  - U.S. Geological Survey: USGS
  - Department of Energy: DoE
  - National Institutes of Health: NIH
- The public are seeking a return on their 'investment'
- So are **corporations** and private funders

- Data and information has 'ruin' value…

However…

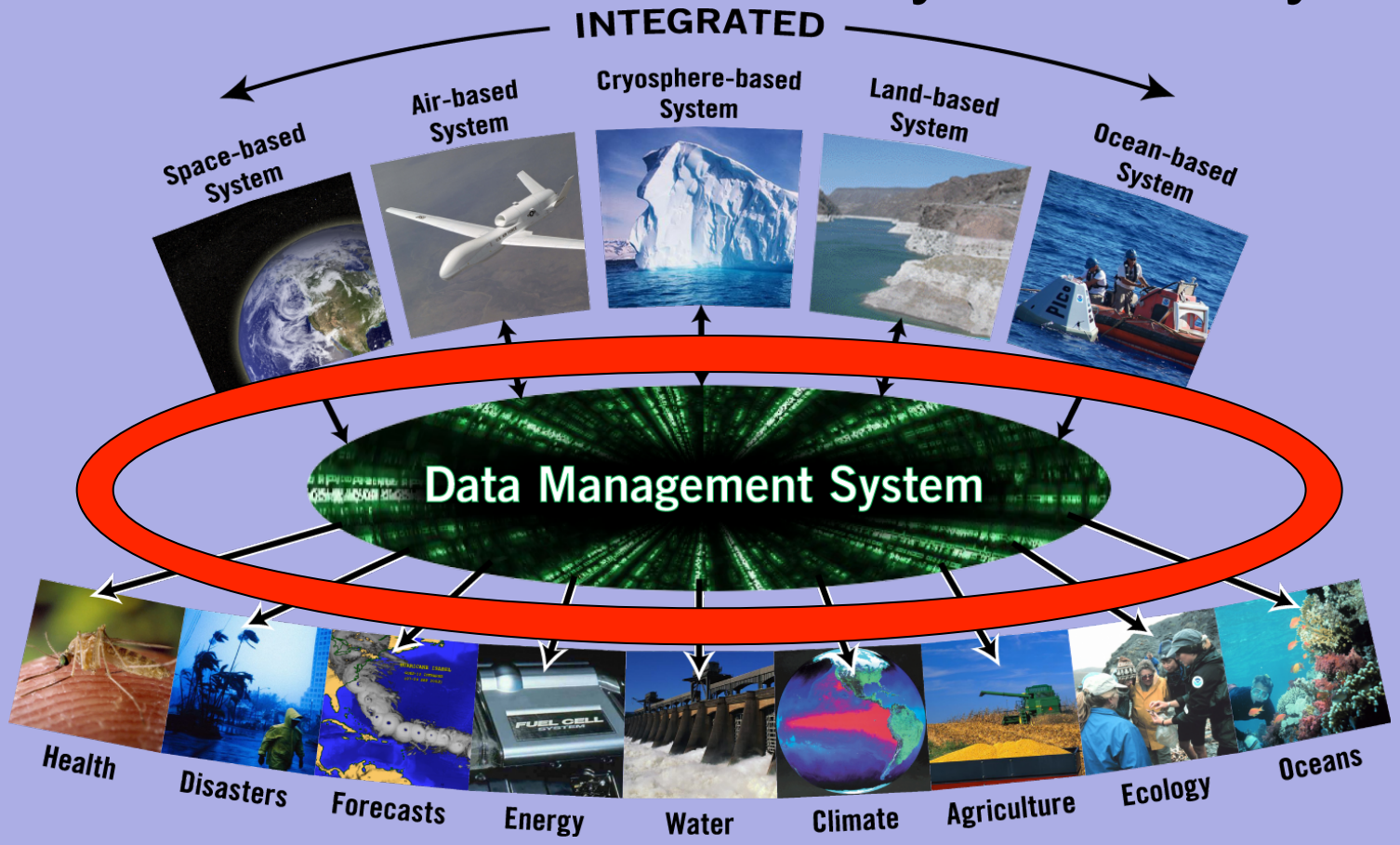
- This is not a new problem…
- The rules have changed…
- The scale is changing…
- And data is handled at all scales … from


Group on Earth Observations
Earth is a complex system of systems
Solar Radiation
CLIMATE VARIABILITY AND CHANGE
ATMOSPHERIC COMPOSITION
H_2O, CO_2, CH_4, N_2O, O_3, etc.
Aerosols
Clouds
Volcanoes
Terrestrial Radiation
WATER CYCLE
HUMAN CONTRIBUTIONS AND RESPONSES
Atmosphere-Biosphere Interaction
ECOSYSTEMS
CARBON CYCLE
Evaporation Precipitation
Industries
Glaciers
Ice Sheet
Atmosphere-Ice Interaction
Heat Exchange
Cities
Transportation
Vegetation
Sea-Ice
Oceans
Agriculture
Vegetation-Soil interaction
Ocean Circulation, Sea Level, Biogeochemistry
Rivers
Land Surface
Data is required from multiple observation networks . . .
LAND-USE/ LAND-COVER CHANGE
. . . and systems . . .

To

Group on Earth Observations

The Future for Earth Science: A Global Earth Observation System of Systems



~ 2009

Millennium Ecosystem Assessment

MILLENNIUM ECOSYSTEM ASSESSMENT

Home | About | Reports | Newsroom | Resources | Contacts | Sitemap

Guide to the Millennium Assessment Reports

Oops! Google Chrome could not find wdc.nbii.gov

Google

Suggestions:

- Go to *nbii.gov*
- Search on Google:

wdc nbii gov | Google Search

Statement of the MA Board

The MA Board of Directors has developed an interpretation of the key messages to emerge from the assessment, entitled *Living Beyond Our Means: Natural Assets and Human Well-Being.*

- Learn more
- Download the Statement
- About the MA Board of Directors

A Framework for Assessment

In late 2003, the MA and Island Press published *Ecosystems and Human Well-being: A Framework for Assessment.* This volume lays out the assumptions, processes and parameters that were used in the MA.

- Learn more

Useful Links

ISLANDPRESS

- Order printed reports from Island Press

GreenFacts.org

- GreenFacts: Popularized synthesis report

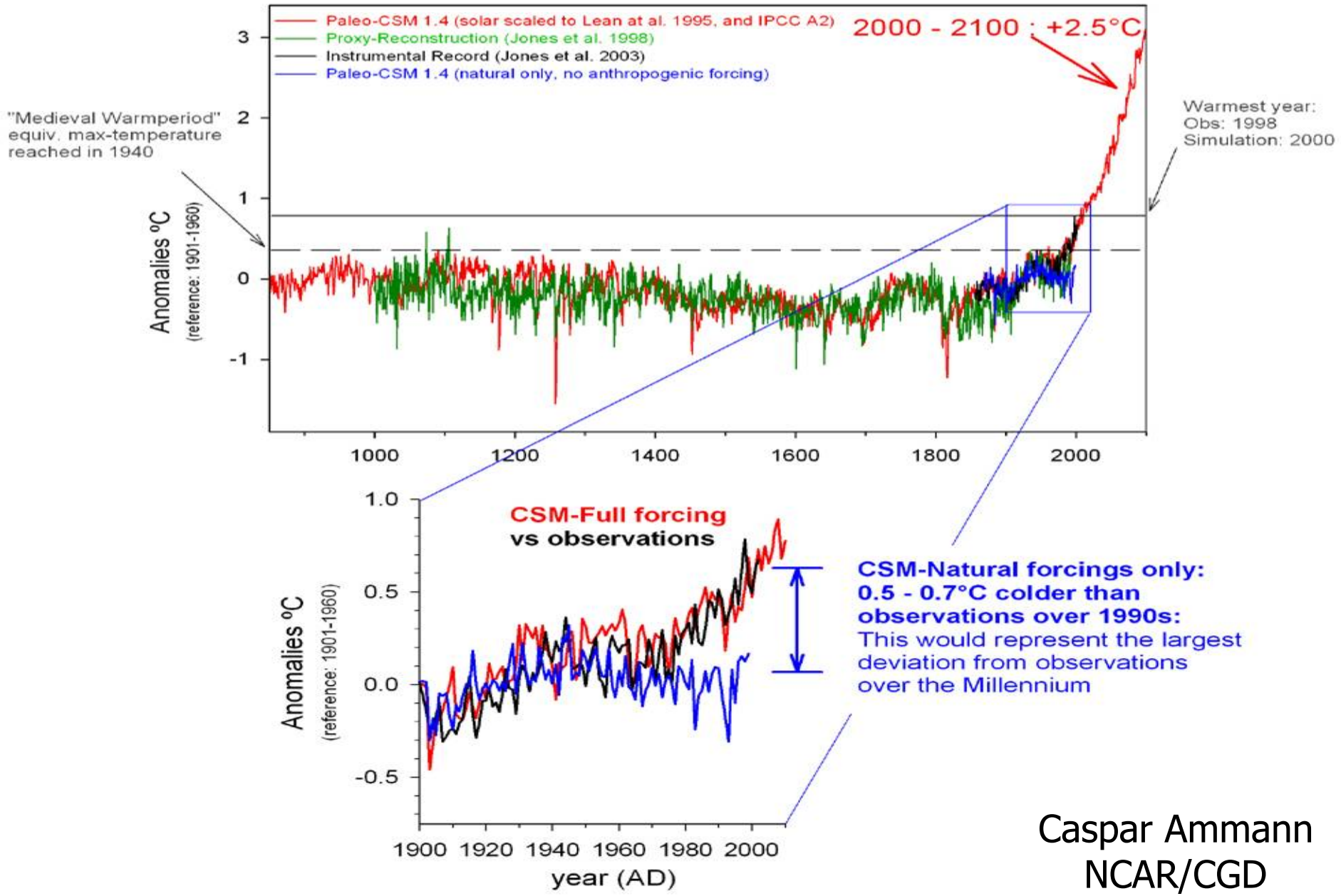
USGS nbii

- MA Data Portal

Also on This Site


CIESIN
Columbia University

Climate of the last Millennium



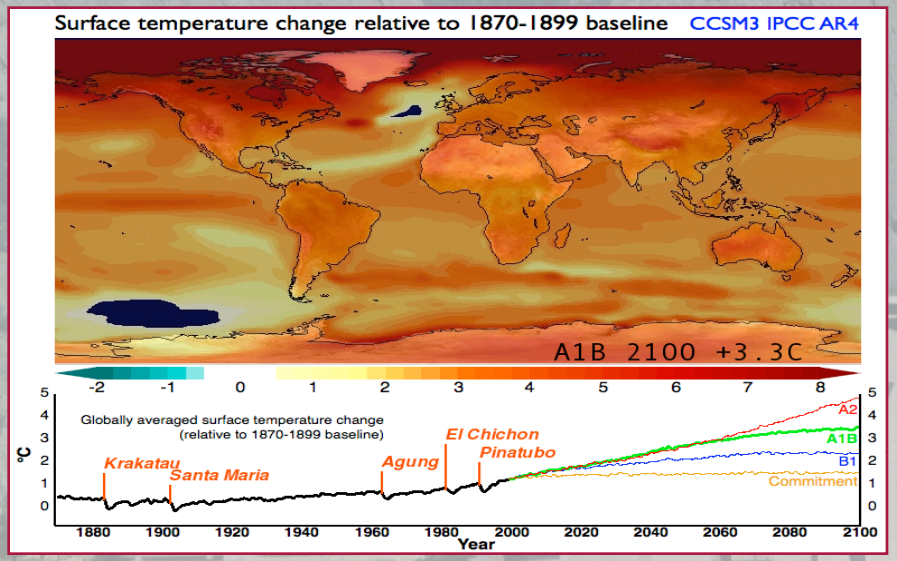
Caspar Ammann
NCAR/CGD

The National Center for Atmospheric Research

In recognition and appreciation of the

PCMDI / LLNL

for its invaluable contribution to the CCSM3 development, production, and data analysis effort for the 2007 IPCC Fourth Assessment Report.



"The Norwegian Nobel Committee has decided that the Nobel Peace Prize for 2007 is to be shared, in two equal parts, between the Intergovernmental Panel on Climate Change (IPCC) and Albert Arnold (Al) Gore Jr. for their efforts to build up and disseminate greater knowledge about man-made climate change, and to lay the foundations for the measures that are needed to counteract such change."

A Vision Fulfilled

Pre-2000
Home Grown
Data Systems

- Initially Cheap
- $$$ in long term
- Limited Scale

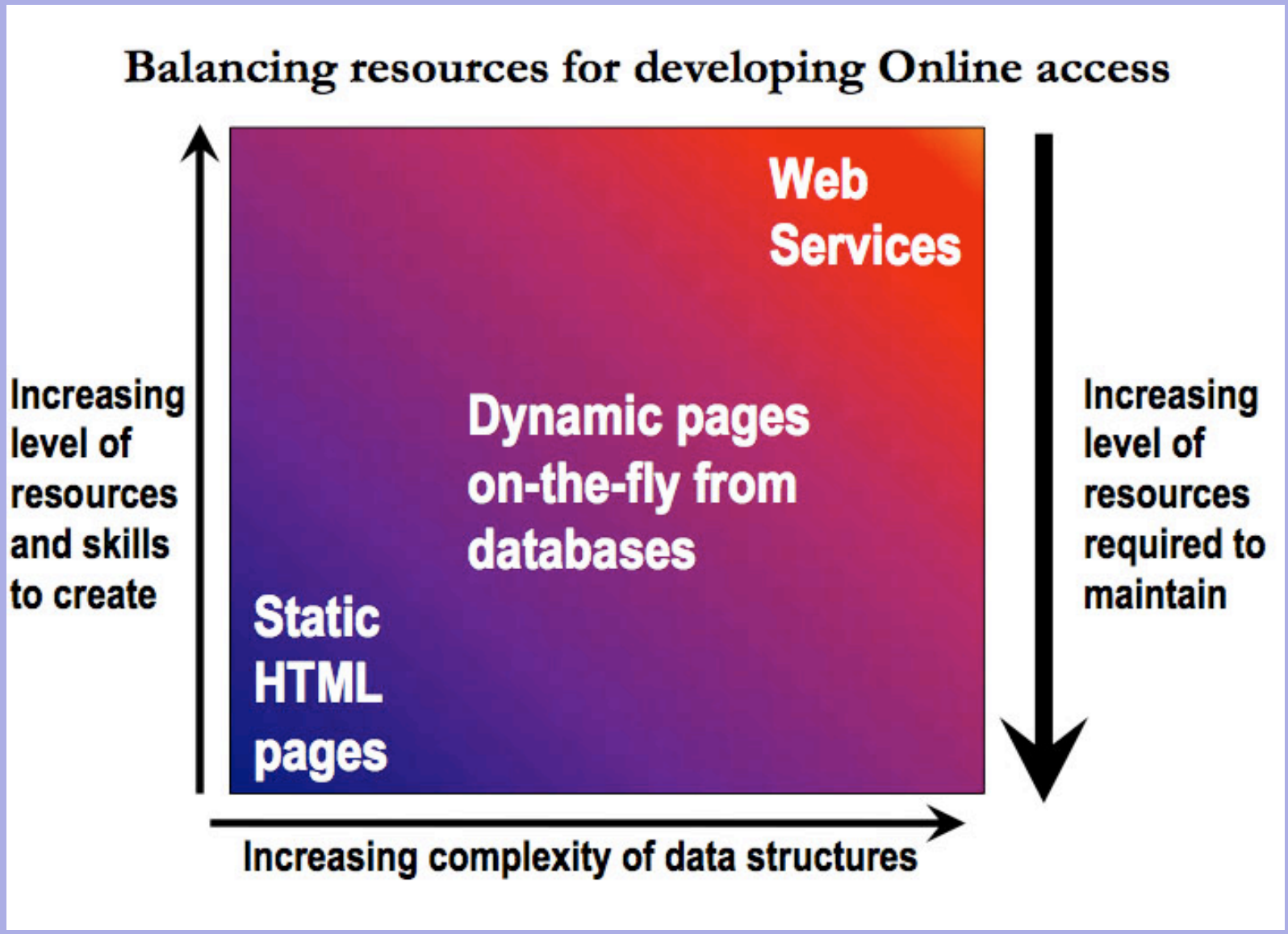
2000-Present
Community Data Portal
dataportal.ucar.edu

- Modest Investment
- Agile and Right-sized for Many Projects
- Institutional Scale

2002-Present
Earth System Grid

- Large Investment
- Infrastructure for Large Projects
- Spans Institutions

Shifting the Burden from the User to the Provider

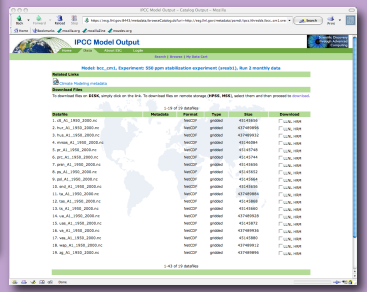


Providing climate scientists with virtual proximity to large simulation results needed for their research

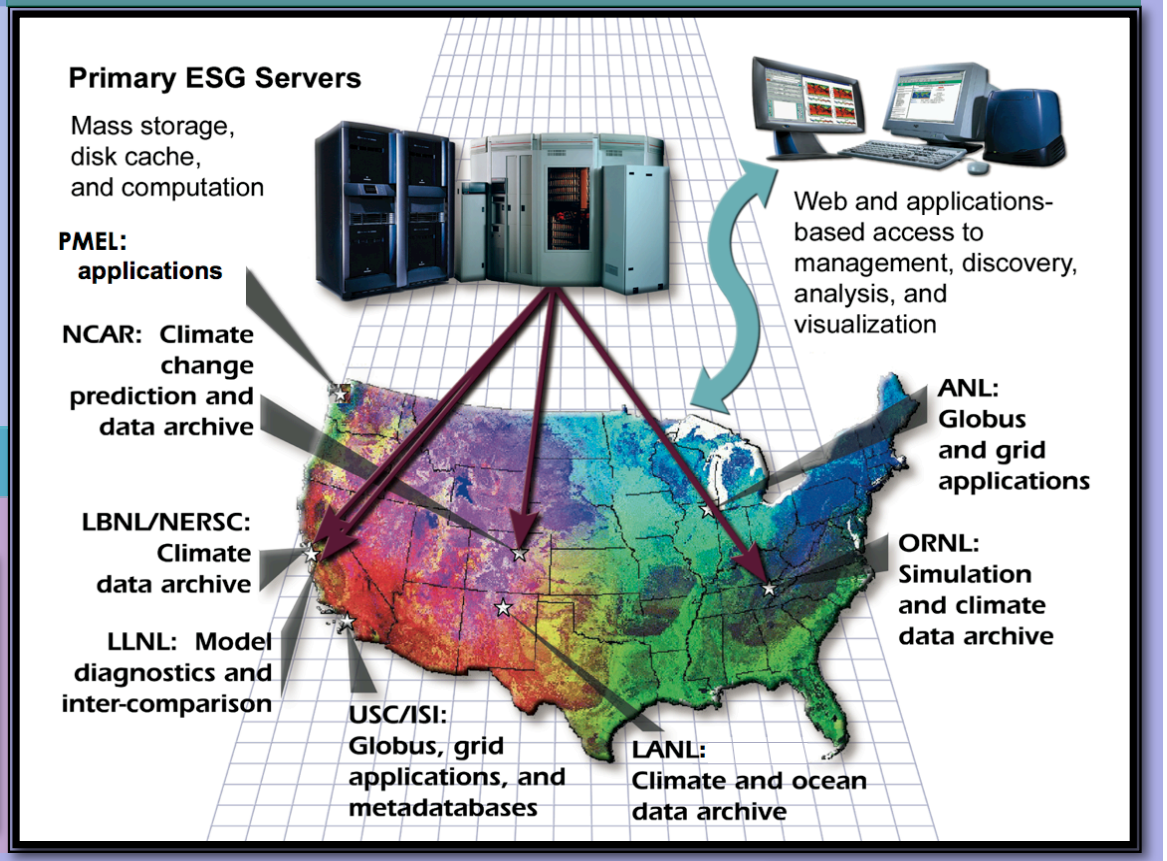
ESG Goal

- Very large distributed data archives
  - ➢Easy federation of sites
  - ➢Across the US and around the world
- "Virtual Datasets" created through subsetting and aggregation
- Metadata-based search and discovery
- Web-based and analysis tool access
- Increased flexibility and robustness
- Server-side analysis

http://www-pcmdi.llnl.gov



Current ESG Sites



Dean Williams, PCMDI, ~ 2008

Earth System Grid

Evolving for the future

ESG Data System Evolution

2008

Central database
Centralized curated data archive
Time aggregation
Distribution by file transport
No ESG analysis
Shopping-cart-style web portal
ESG connection to desktop analysis tools

Early 2009

Testbed data sharing
Federated metadata/portals
Unified user interface
Quick look server-side analysis with CDAT
Location independence
Distributed aggregation
Manual data sharing/publishing

2011

Full data sharing (add to testbed…)
Synchronized federation metadata, data
Full suite of server-side analysis with CDAT
Model/observation integration
ESG embedded into desktop productivity tools with CDAT
GIS integration
Model intercomparison metrics
User support, life cycle maintenance

Terabytes

CCSM AR4

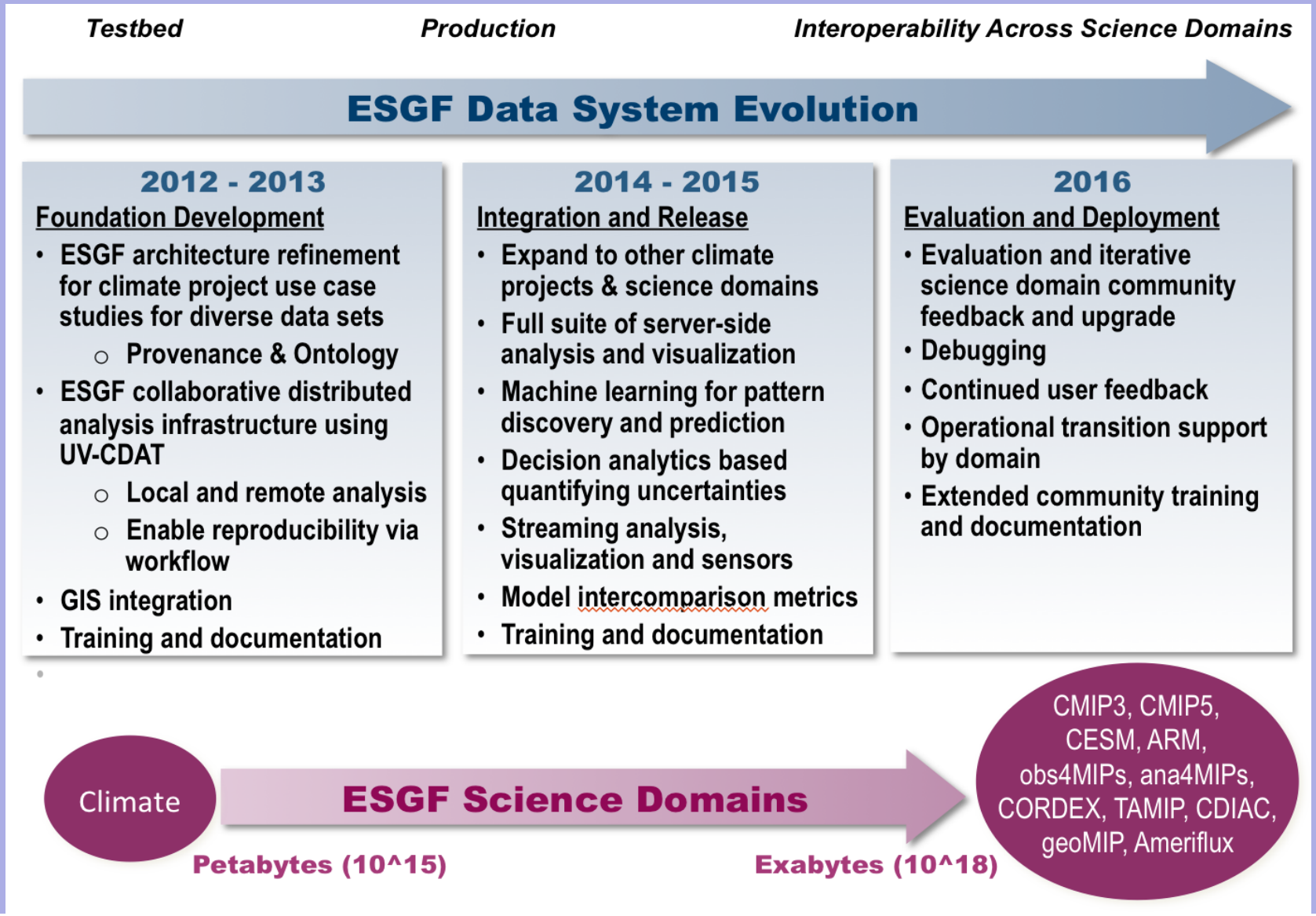
ESG Data Archive

Petabytes

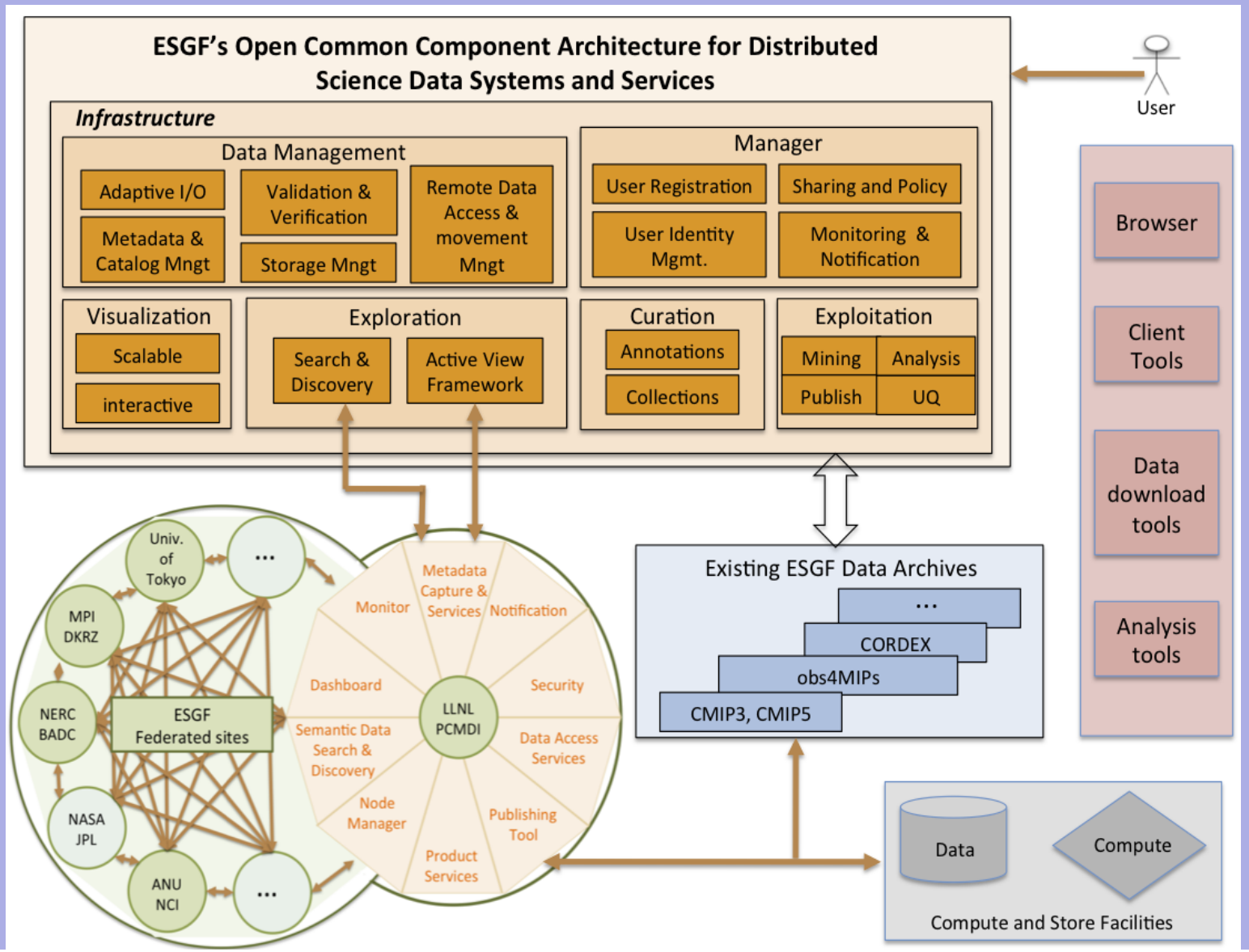
CCSM, AR5, satellite, In situ biogeochemistry, ecosystems

Dean Williams, PCMDI

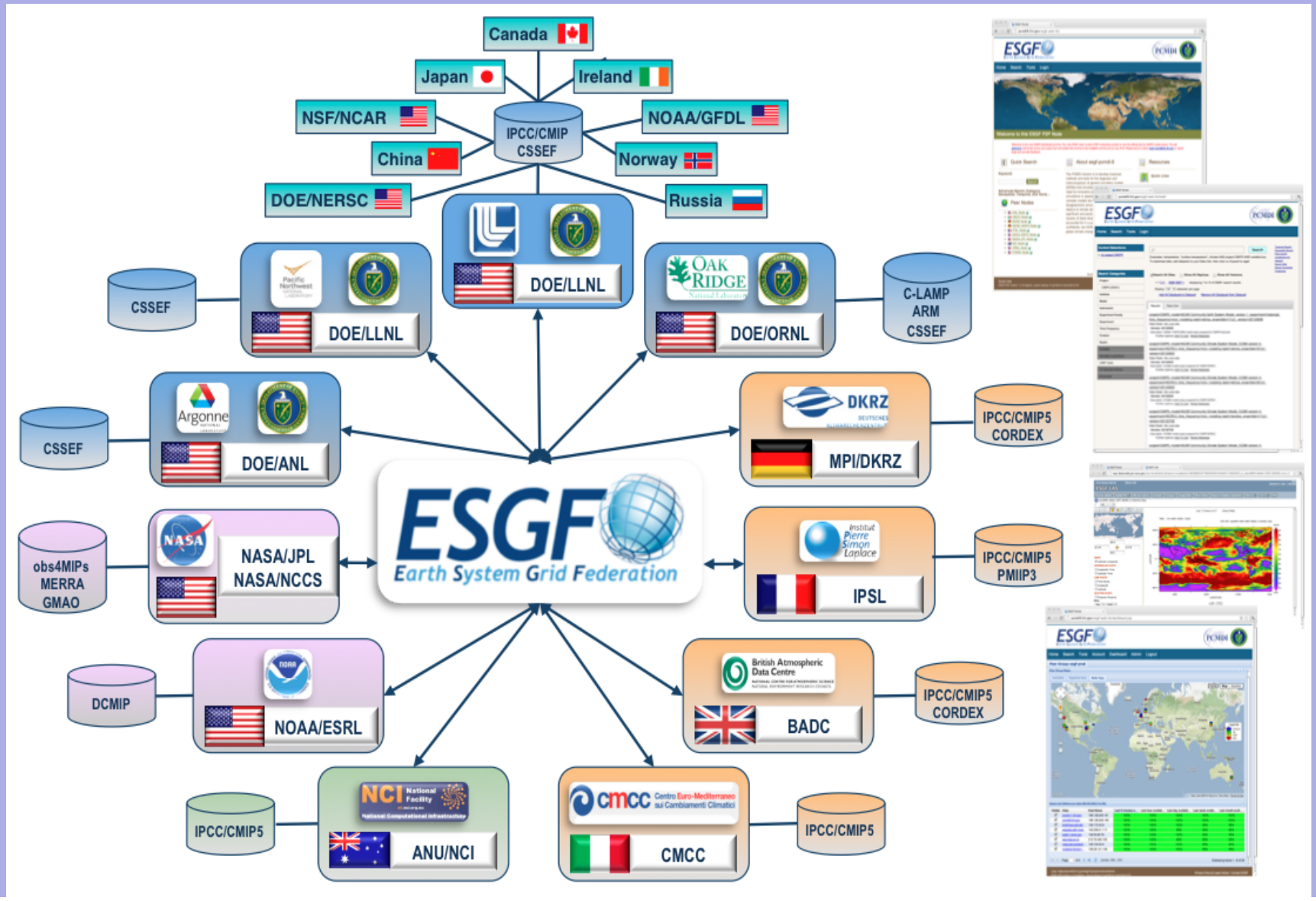
And beyond…


Testbed
Production
Interoperability Across Science Domains
ESGF Data System Evolution
2012 - 2013
Foundation Development
• ESGF architecture refinement for climate project use case studies for diverse data sets
○ Provenance & Ontology
• ESGF collaborative distributed analysis infrastructure using UV-CDAT
○ Local and remote analysis
○ Enable reproducibility via workflow
• GIS integration
• Training and documentation
2014 - 2015
Integration and Release
• Expand to other climate projects & science domains
• Full suite of server-side analysis and visualization
• Machine learning for pattern discovery and prediction
• Decision analytics based quantifying uncertainties
• Streaming analysis, visualization and sensors
• Model intercomparison metrics
• Training and documentation
2016
Evaluation and Deployment
• Evaluation and iterative science domain community feedback and upgrade
• Debugging
• Continued user feedback
• Operational transition support by domain
• Extended community training and documentation
Climate
ESGF Science Domains
Petabytes (10^15)
Exabytes (10^18)
CMIP3, CMIP5, CESM, ARM, obs4MIPs, ana4MIPs, CORDEX, TAMIP, CDIAC, geoMIP, Ameriflux

Scaling a big infrastructure!



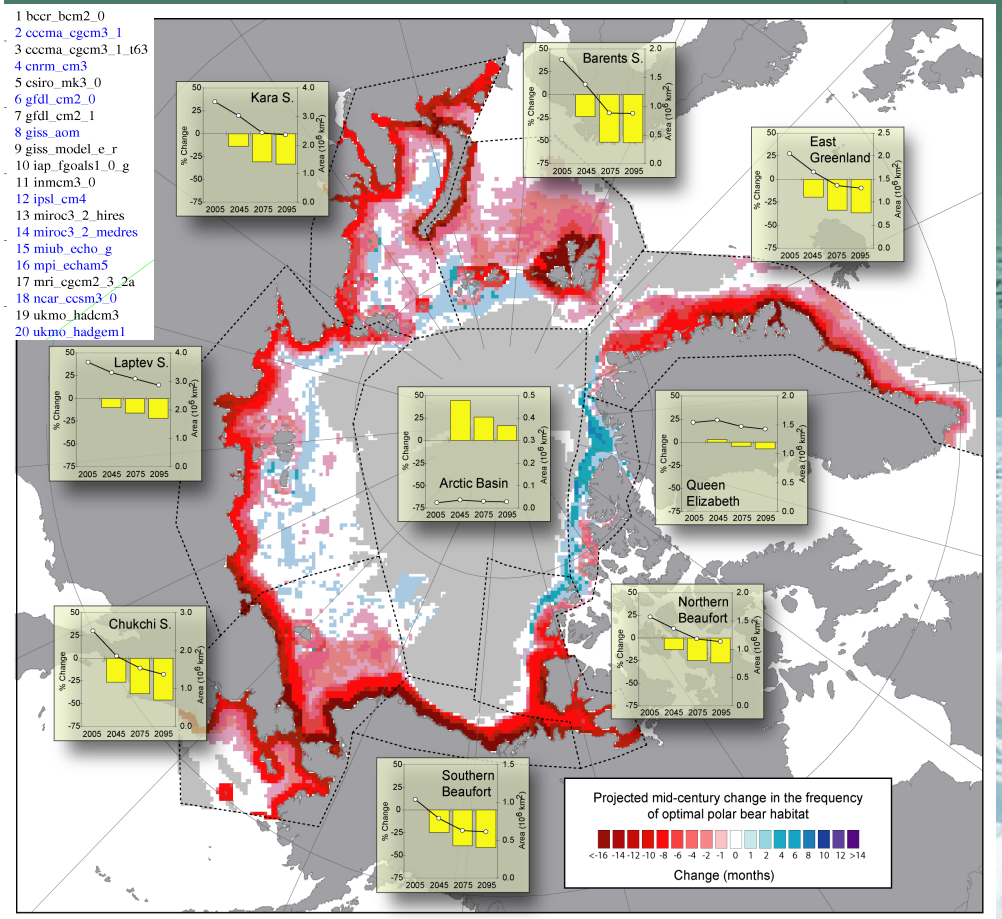
Climate data - worldwide



USGS
science for a changing world

Briefing on Results:

USGS Science Strategy to Support U.S. Fish & Wildlife Service Polar Bear Listing Decision: *a 6 month effort*

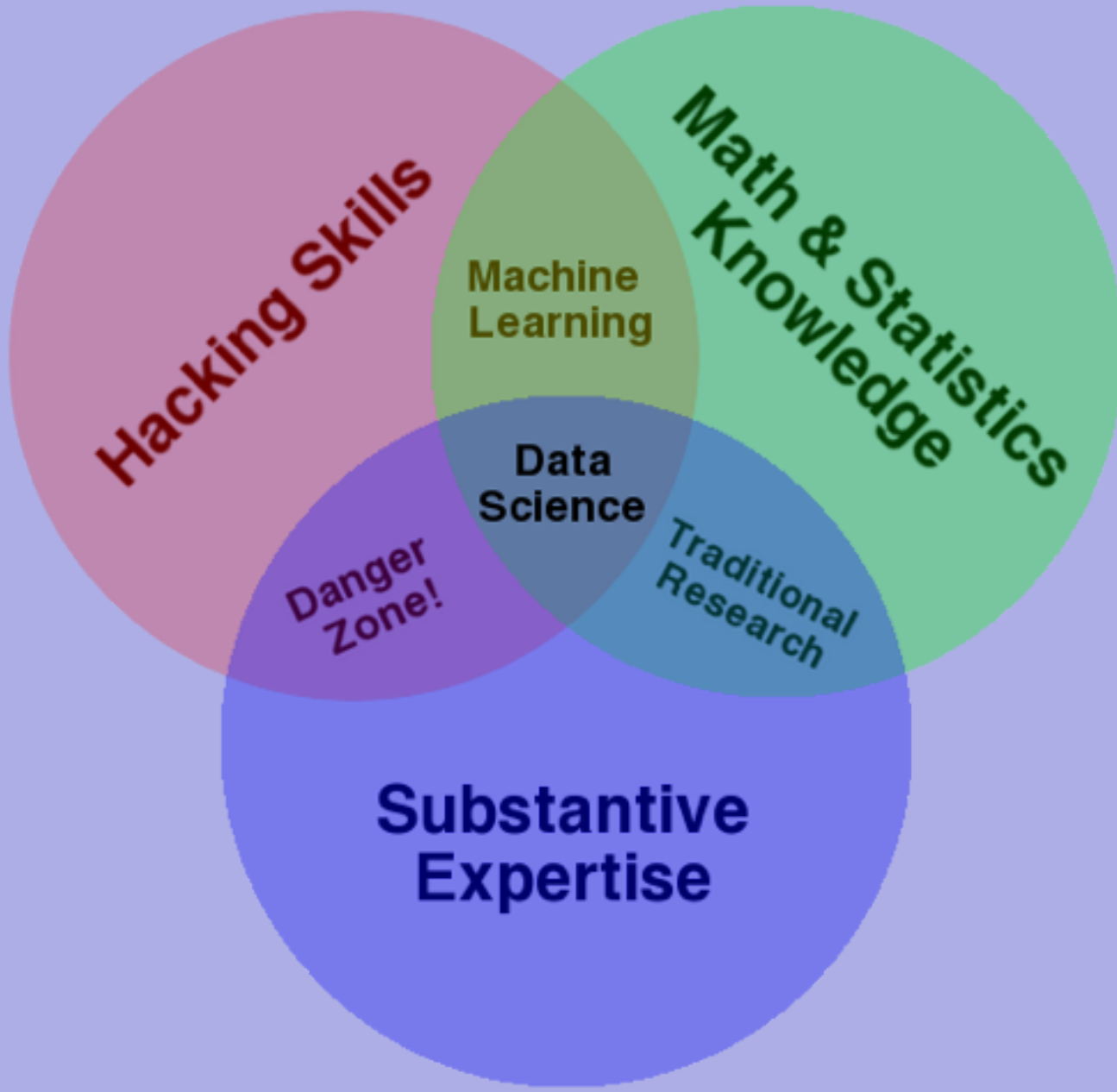


U.S. Department of the Interior
U.S. Geological Survey

Rise of the Data Scientist

- http://flowingdata.com/2009/06/04/rise-of-the-data-scientist/
- http://www.drewconway.com/zia/?p=2167
- http://radar.oreilly.com/2010/06/what-is-data-science.html
- http://www.wired.com/magazine/2010/06/ff_sergeys_search/all/1
- http://research.microsoft.com/en-us/collaboration/fourthparadigm/

You will see many diagrams like



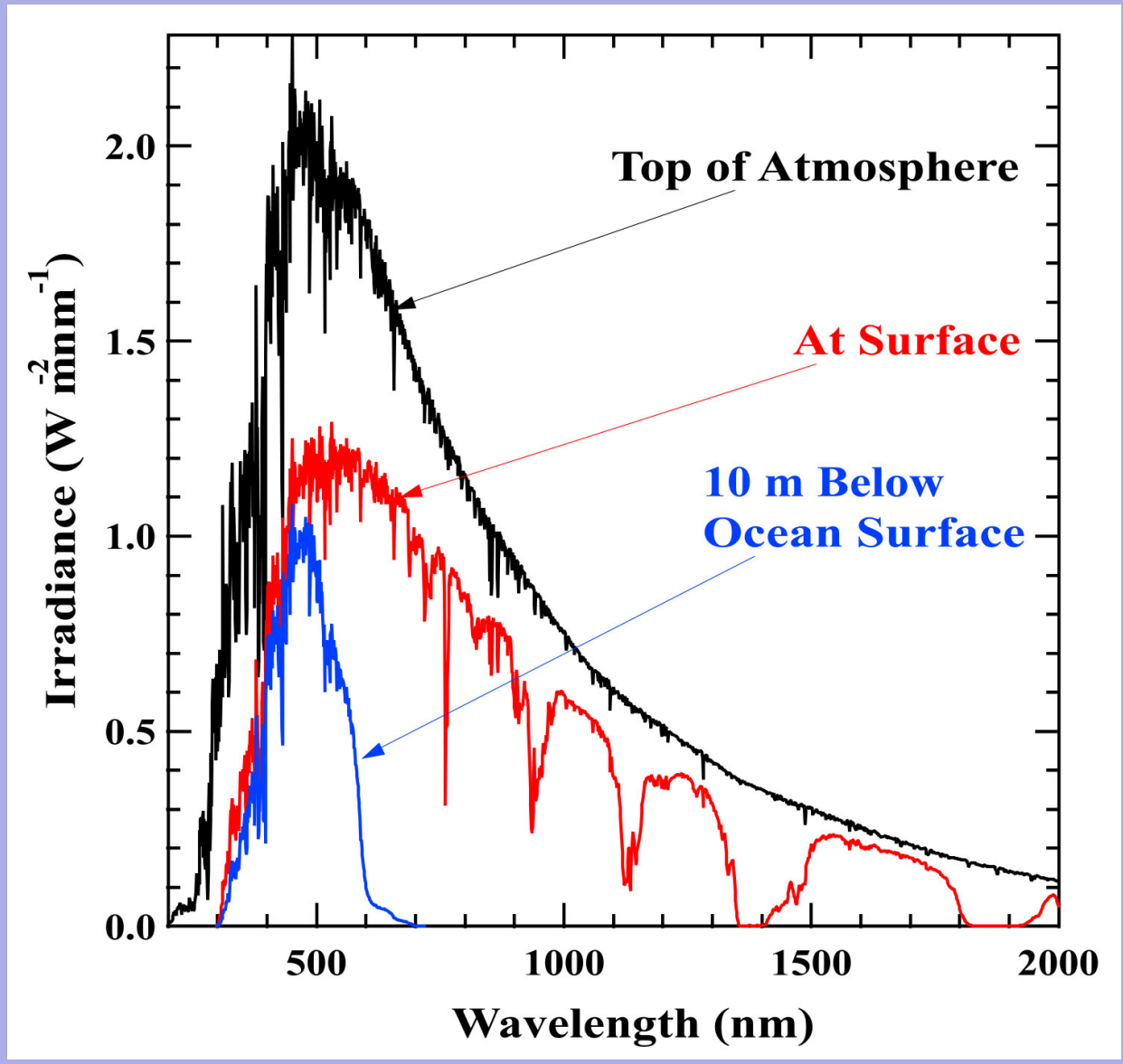
And now a more detailed e.g.

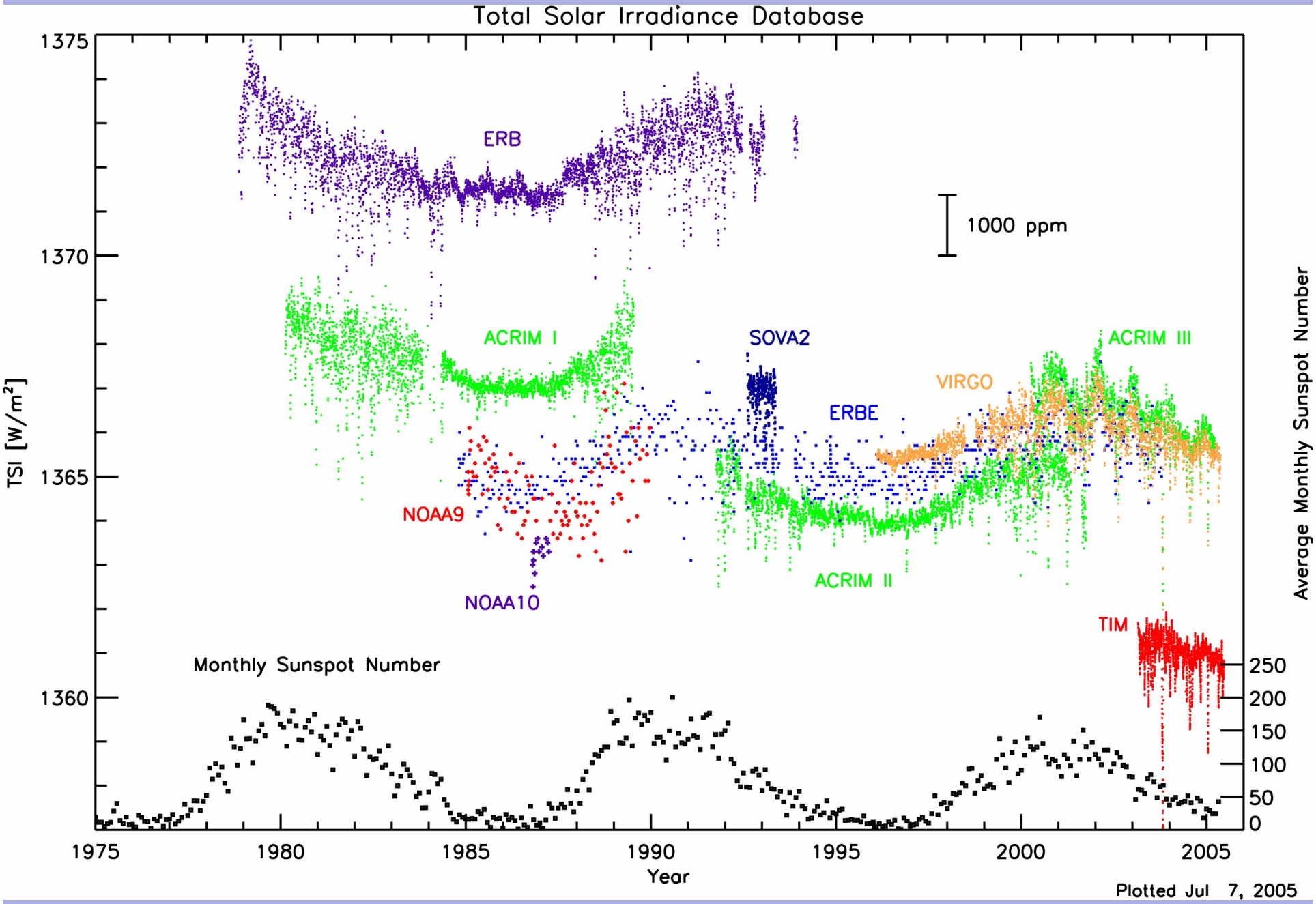
- Or … part of my path to data science

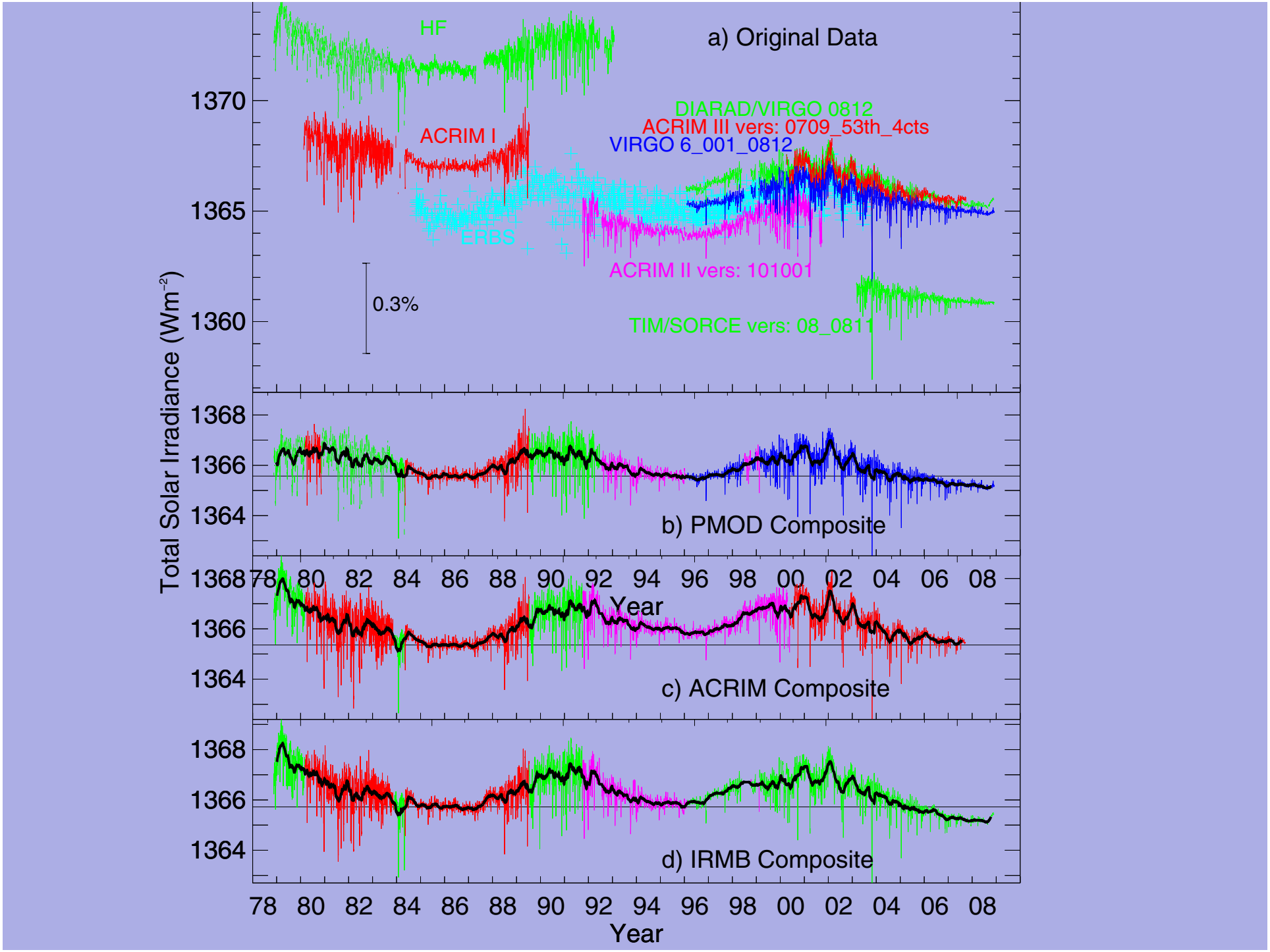
Why do we (I) care about the Sun?

- The Sun's radiation is the single largest external input to the the Earth's atmosphere and thus the Earth system.
- Add, it varies – in time and wavelength
- Also, for a long time – Solar Energetic Particles and the near Earth environment (and more recently the effect on clouds?)
- Observations commenced ~ 1940's, with a resurgence in the late 1970's
- Two quantities of scientific interest
  - Total Solar irradiance – TSI in Wm^{-2} (adjusted to 1AU)
  - Solar Spectral Irradiance – SSI in $Wm^{-2}m^{-1}$or $Wm^{-2}nm^{-1}$
- Measure, model, understand -> construct, predict

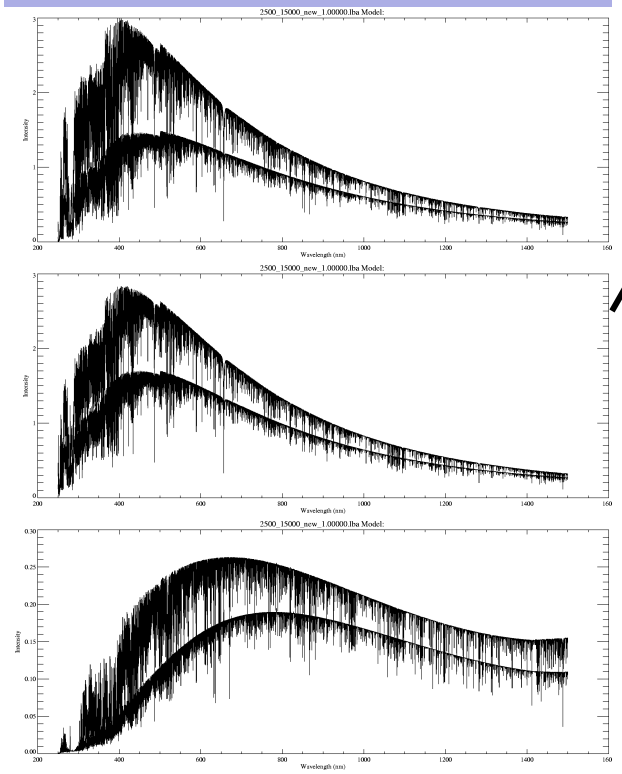
Solar radiation as a function of altitude

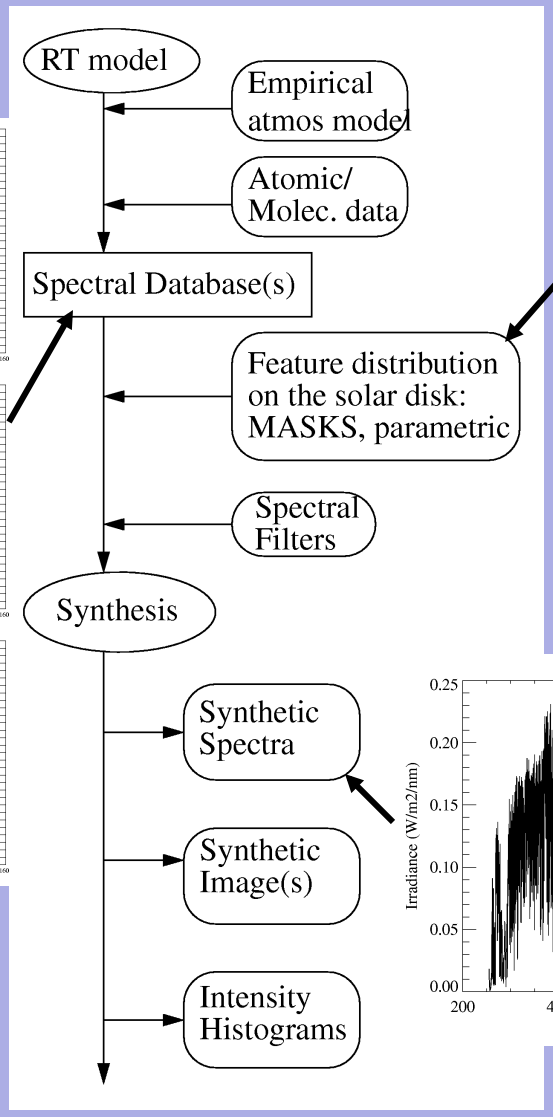


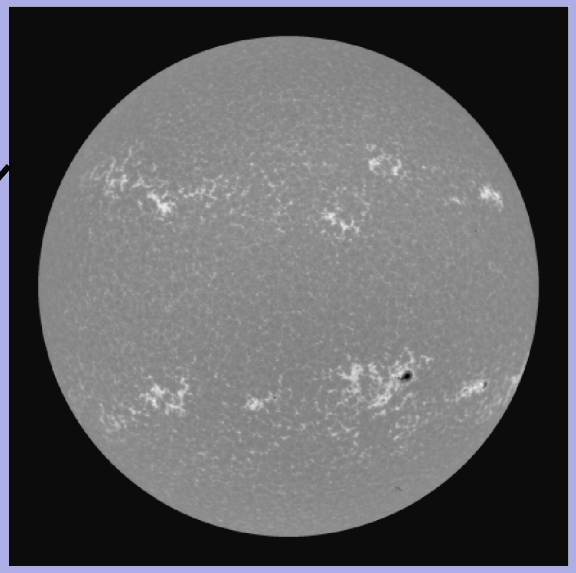

Total Solar Irradiance Database
TSI [W/m²]
Average Monthly Sunspot Number
Year
1375
1370
1365
1360
1975
1980
1985
1990
1995
2000
2005
250
200
150
100
50
0
ERB
1000 ppm
ACRIM I
SOVA2
ACRIM III
VIRGO
ERBE
NOAA9
NOAA10
ACRIM II
TIM
Monthly Sunspot Number
Plotted Jul 7, 2005

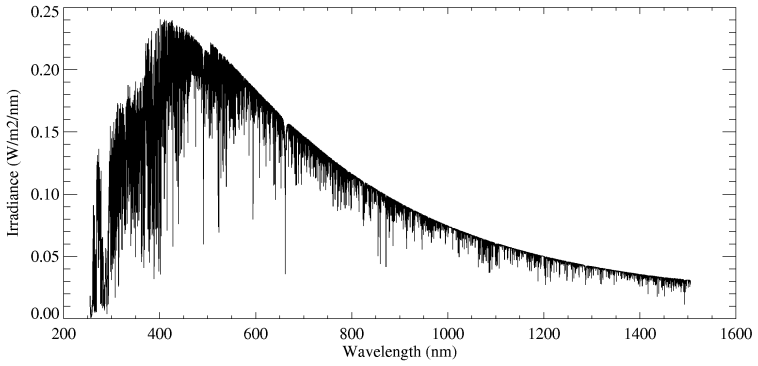

a) Original Data
HF
ACRIM I
ERBS
DIARAD/VIRGO 0812
ACRIM III vers: 0709_53th_4cts
VIRGO 6_001_0812
ACRIM II vers: 101001
TIM/SORCE vers: 08_0811
0.3%
b) PMOD Composite
c) ACRIM Composite
d) IRMB Composite
Total Solar Irradiance (Wm-2)
1370
1365
1360
1368
1366
1364
78 80 82 84 86 88 90 92 94 96 98 00 02 04 06 08
Year

Spectral synthesis components and flow









Summary of Results

- First comprehensive 'database' of:
  - Empirical models of the thermodynamic structure of the solar atmosphere suitable for different solar magnetic activity levels
- First comprehensive (70 component) synthetic spectral irradiance database in absolute units
  - 10 disk angles, 7 models, far ultra- violet to far infrared, multi-resolution
  - ~724 GB
- Strong validation in ultraviolet, visible, lines, infrared
  - Correct center to limb prediction for red-band irradiances
  - Found 30-45% network contribution to Ly-α irradiance
- Several comparisons led to improvements in the atomic parameters
- Led to choice of PICARD (new satellite) filter wavelengths

Which brings us to DATA SCIENCE

- Drum roll…..
- Some dirty secrets
- And some … universal truths…

Needs (this is our mantra)

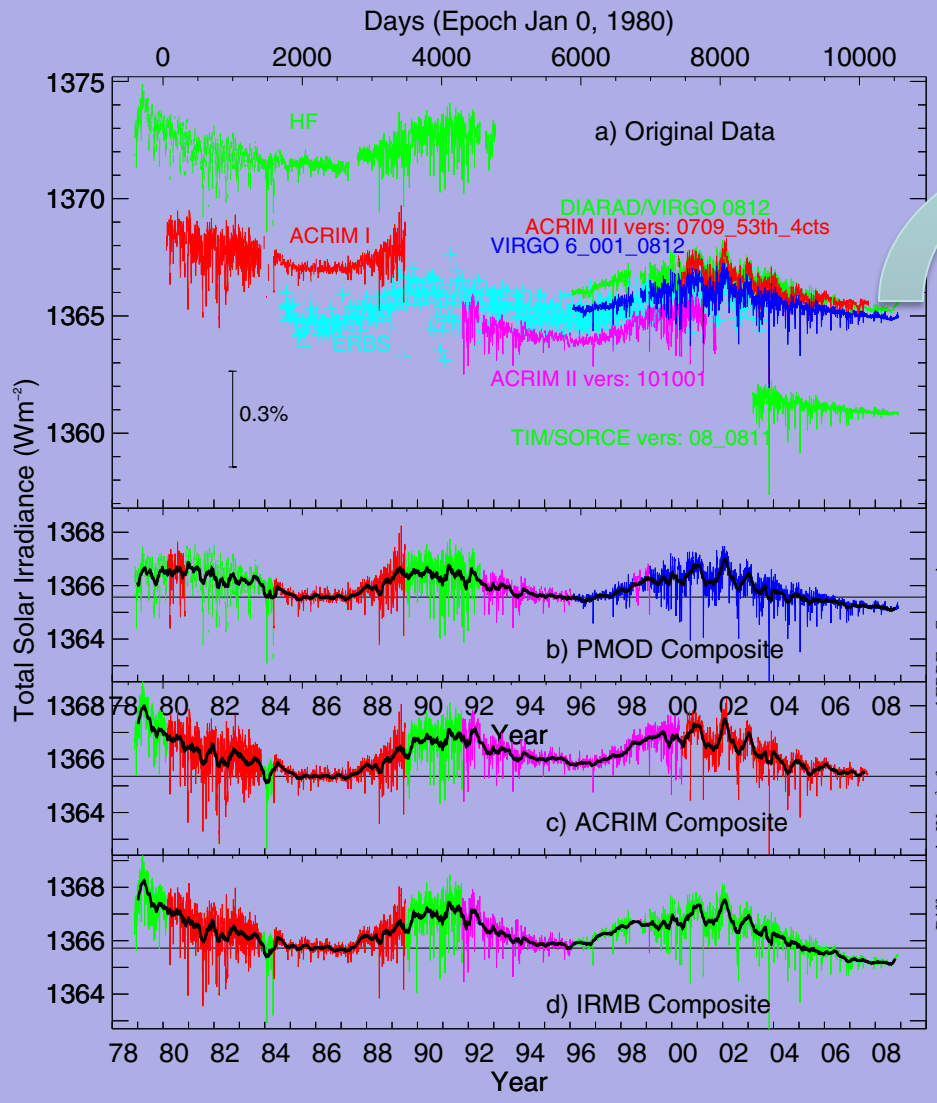
Scientists should be able to access a global, distributed knowledge base of scientific data that:

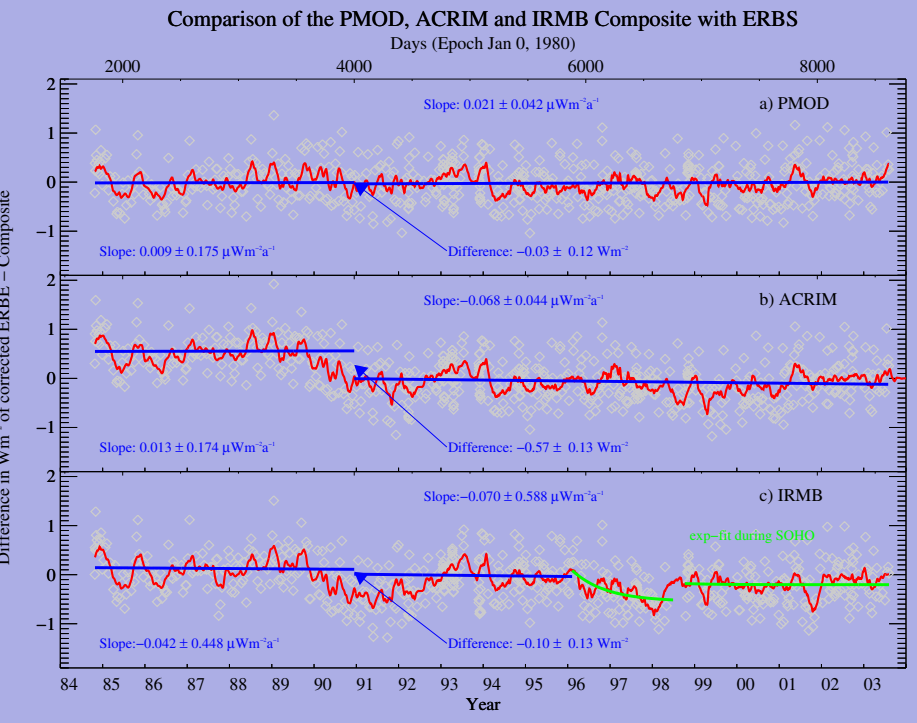
- appears to be integrated
- appears to be locally available

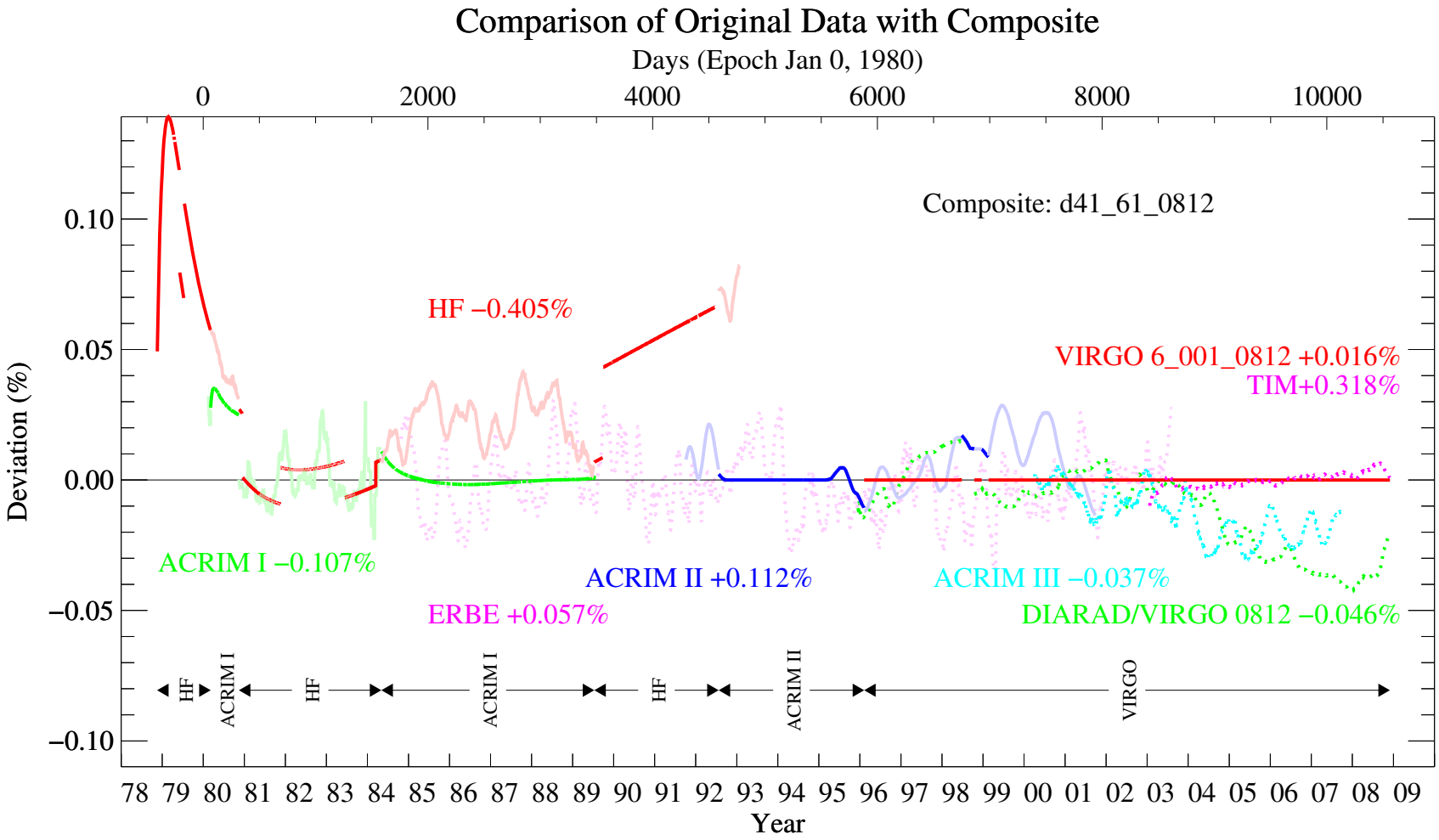
But… data is obtained by multiple means (models and instruments), using various protocols, in differing vocabularies, using (sometimes unstated) assumptions, with inconsistent (or non-existent) meta-data. It may be inconsistent, incomplete, evolving, and distributed. And created in a manner to facilitate its generation NOT its use.

And… there exist(ed) significant levels of semantic heterogeneity, large-scale data, complex data types, legacy systems, inflexible and unsustainable implementation technology

Back to the TSI time series…

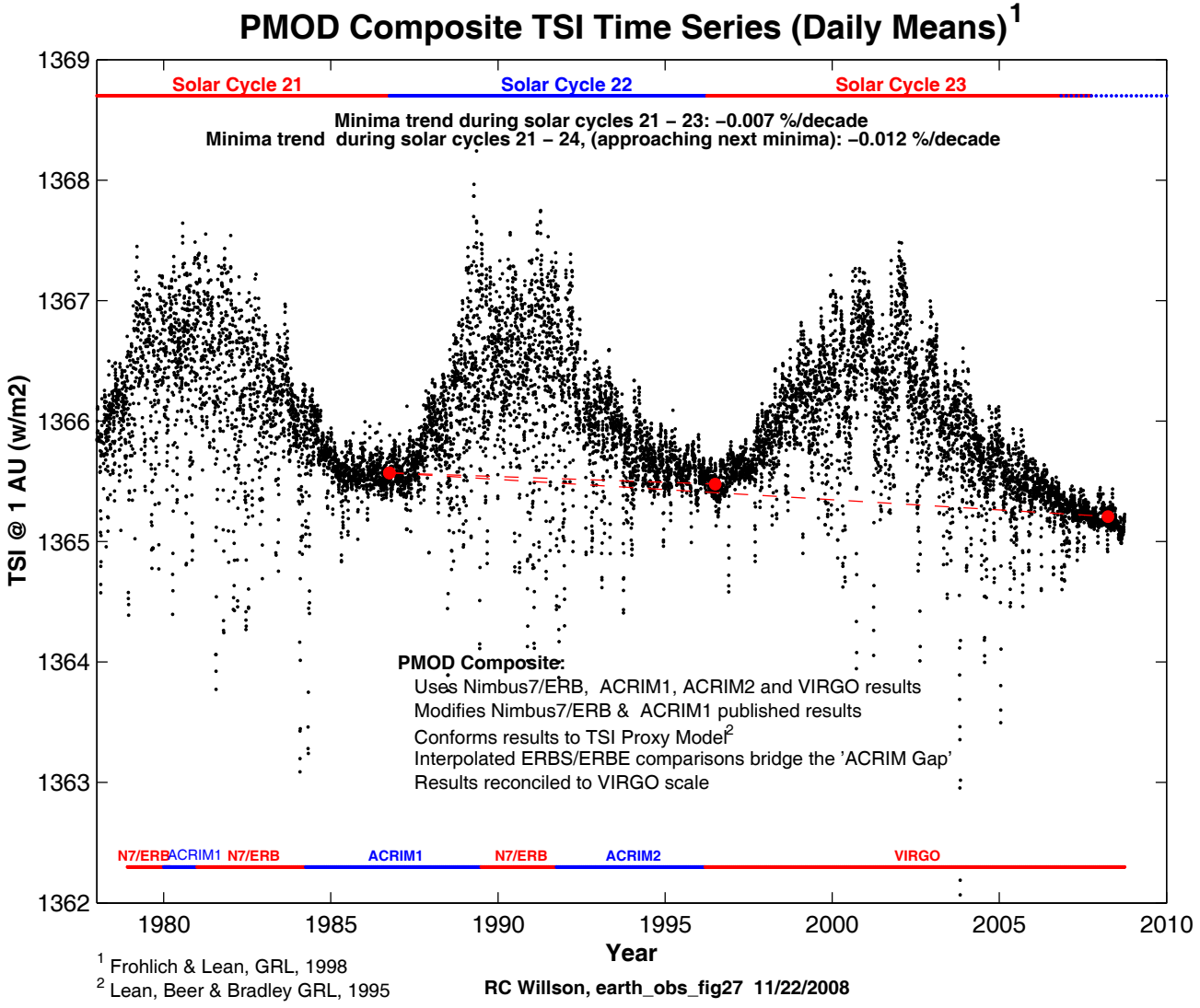




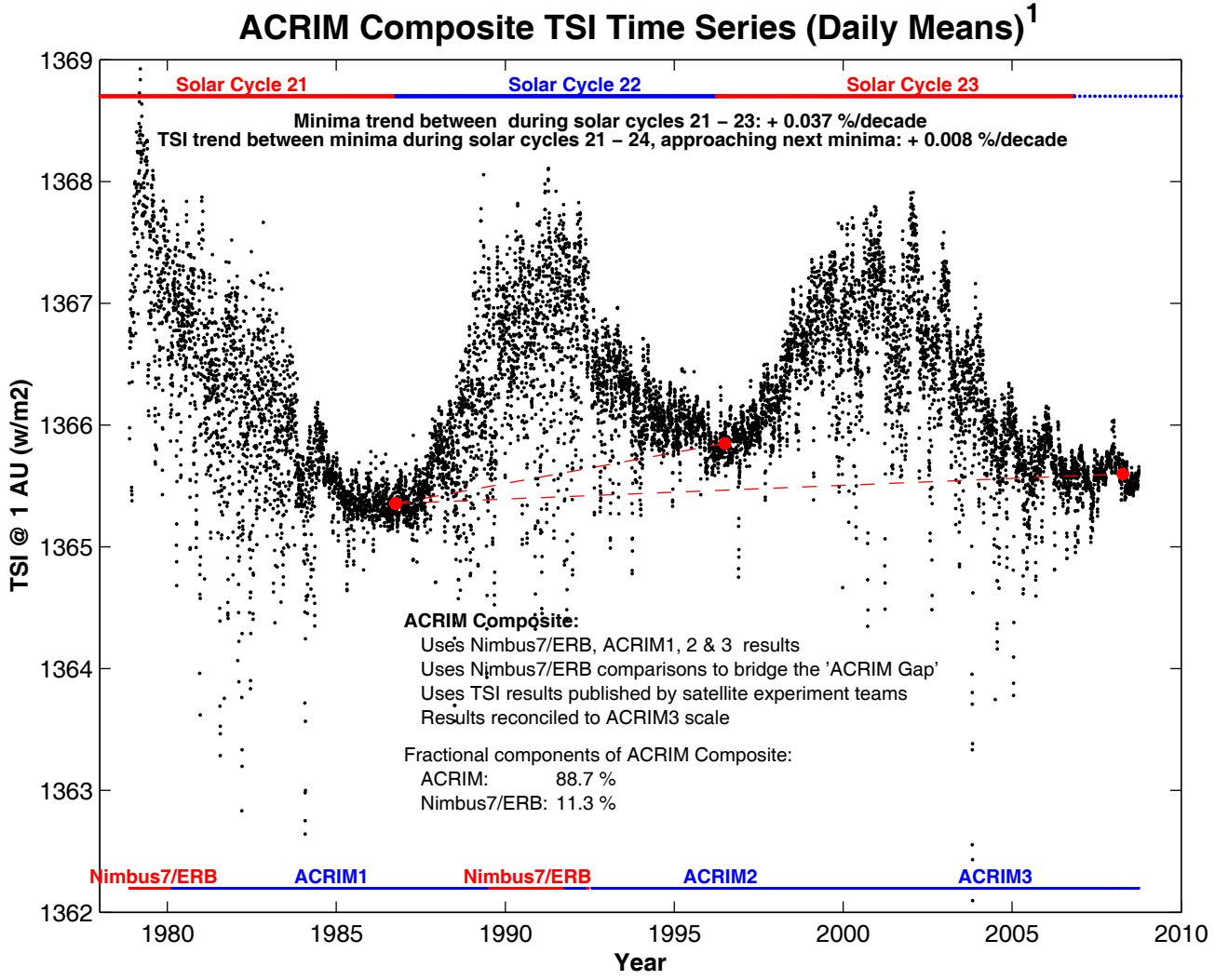


from: C. Fröhlich, Metrologia, *0*, pp.60–65, 2003, with composite (vers d41_61_0812), ACRIM–II/III (vers 101001/0709_53th_4cts) and VIRGO 6_001_0812 data (Dec 05, 2008)

One composite, one assumption



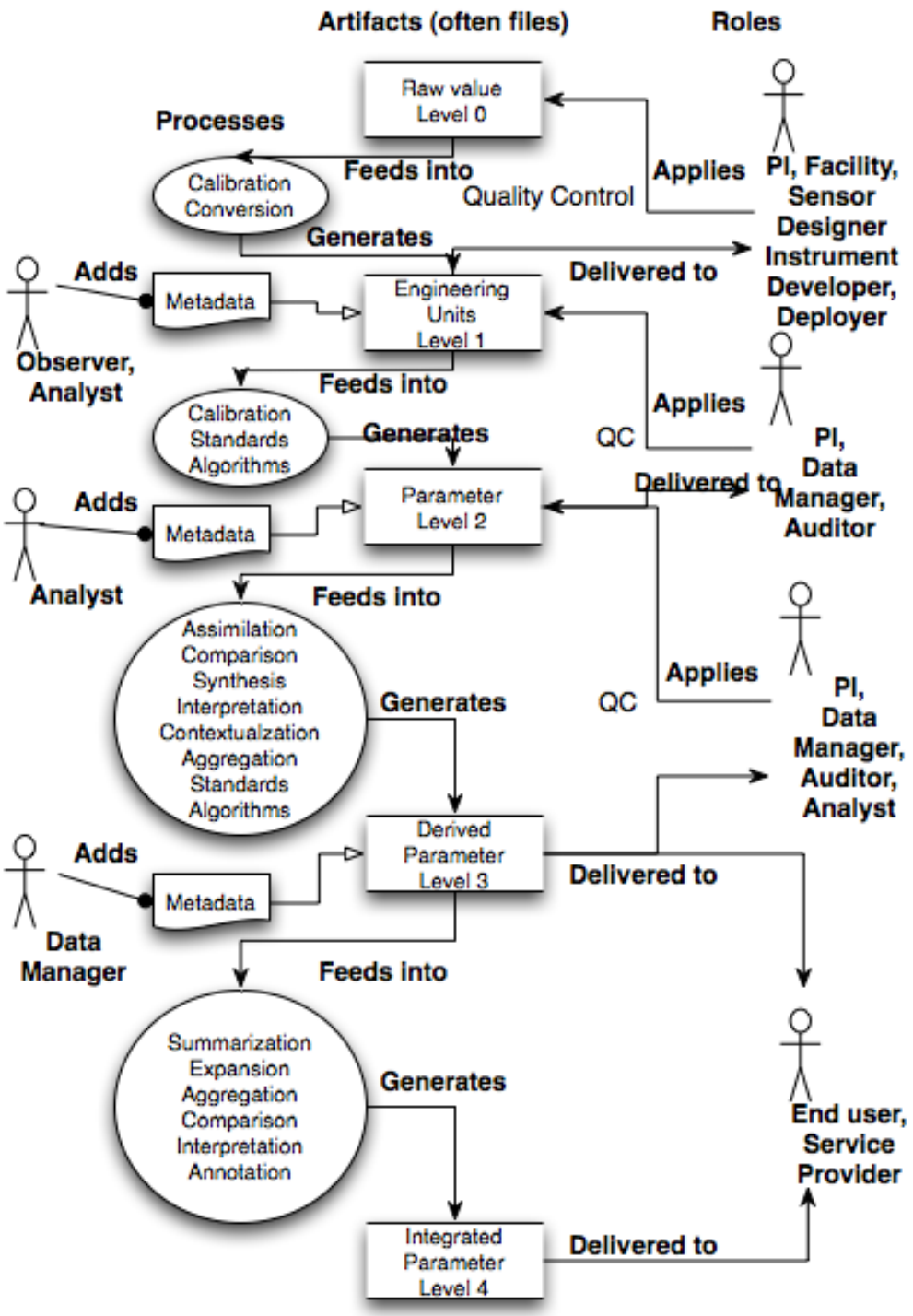
Another composite, different assumption



[1] Willson & Mordvinov, GRL, 2003 **RC Willson, earth_obs_fig26 11/23/2008**

Data pipelines: we have problems

- *Data is coming in faster, in greater volumes and forms and outstripping our ability to perform adequate quality control*
- *Data is being used in new ways and we frequently do not have sufficient information on what happened to the data along the processing stages to determine if it is suitable for a use we did not envision*
- *We often fail to capture, represent and propagate manually generated information that need to go with the data flows*
- *Each time we develop a new instrument, we develop a new data ingest procedure and collect different metadata and organize it differently. It is then hard to use with previous projects*
- *The task of event determination and feature classification is onerous and we don't do it until after we get the data*
- *And now much of the data is on the Internet/Web (good or bad?)*


Artifacts (often files)
Roles
Raw value
Level 0
Processes
Calibration
Conversion
Feeds into
Quality Control
Applies
PI, Facility,
Sensor
Designer
Instrument
Developer,
Deployer
Generates
Delivered to
Adds
Metadata
Engineering
Units
Level 1
Observer,
Analyst
Feeds into
Calibration
Standards
Algorithms
Generates
Applies
QC
PI,
Data
Manager,
Auditor
Delivered to
Adds
Metadata
Parameter
Level 2
Analyst
Feeds into
Assimilation
Comparison
Synthesis
Interpretation
Contextualization
Aggregation
Standards
Algorithms
Generates
Applies
QC
PI,
Data
Manager,
Auditor,
Analyst
Adds
Metadata
Derived
Parameter
Level 3
Delivered to
Data
Manager
Feeds into
Summarization
Expansion
Aggregation
Comparison
Interpretation
Annotation
Generates
End user,
Service
Provider
Integrated
Parameter
Level 4
Delivered to

Provenance

- Origin or source from which something comes, intention for use, who/what generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility

Summary

- Science data (and information) challenges are being identified as increasingly common
- Data (and information) science now accompanies theory, observation/experiment and simulation as a means of doing science
- Scientists and technologist are not well prepared to cope with 21st century data management and use of tools
- Making data available is now a responsibility not a privilege

Skills needed

- Database or data structures?
- Literacy with computers and applications that can handle data
- Ability to access internet and retrieve/ acquire data
- Presentation of assignments
- Working alone and in groups

What is expected

- Attend class, complete assignments (esp. reading)
- Participate
- **Ask questions**
- Work both individually and in a group
- Work constructively in group and class sessions
- Next class Sep. 3 – Data and information acquisition (curation) and metadata/ provenance - management

Reading

- See web page…