NCAR

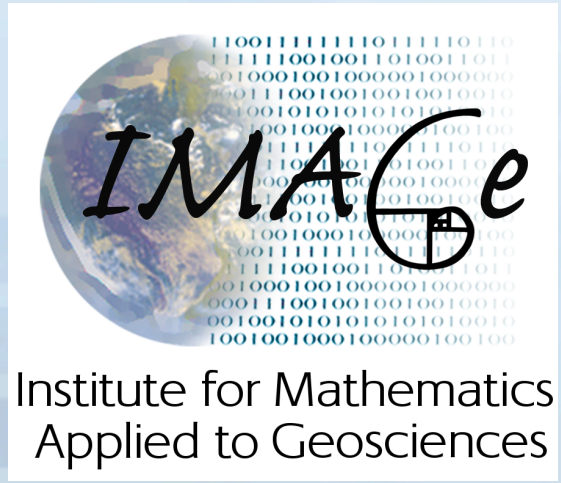
Big Data:
The view from climate science

Seth McGinnis

2016-06-15

National Center for Atmospheric Research

Who I Am and Why I'm Here



Associate Scientist IV

Data Manager

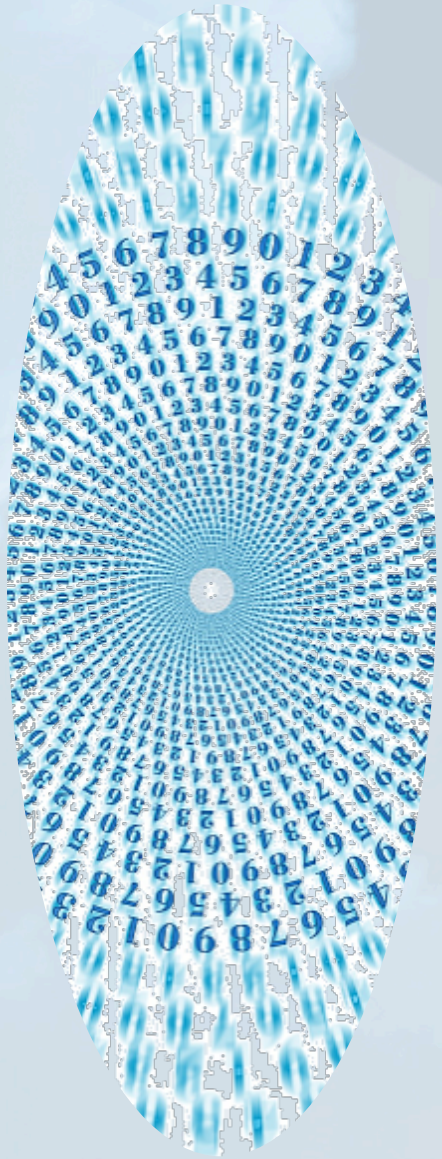
My job is to make climate model outputs useful to other people

Roadmap

- What is Big Data
- The Big V's
  - Volume
  - Variety
  - Velocity
  - …
- Climate data
  - NARCCAP
  - ACADIS
- Dealing with Volume
  - Format
  - Standards
  - Structure
- Dealing with Variety
  - Discovery metadata
- Data Services
  - Capstone

What is Big Data?

- Any dataset that's "too much" to handle using traditional tools and techniques
- Contextual and relative, not absolute
- Enabled by exponential growth of computational capability
- Also gets used to refer to specialized systems used to handle Big Data (machine learning, hadoop, etc.)



The 3 Big V's (+1) (+ N more)

Originate in business & tech sector

- Beware assumptions!

- Big 3 (Laney 2001)
  - Volume
  - Velocity
  - Variety
- Plus 1
  - Value

- Plus many more
  - Veracity
  - Validity
  - Variability
  - Viscosity & Volatility
  - Viability, Venue, Vocabulary, Vagueness, …

Volume

How much storage space the data takes up

- Driven by exponential growth in storage capacity
- Dealt with by technology
  - Parallel processing
  - Better hardware

Velocity

How quickly data must be processed

- Speed of storage / retrieval / analysis
- Aspects:
  - Real-time (acted on immediately)
  - Timeliness (rate of capture/usage)
  - Lifespan (how long it's valuable)
  - Response time

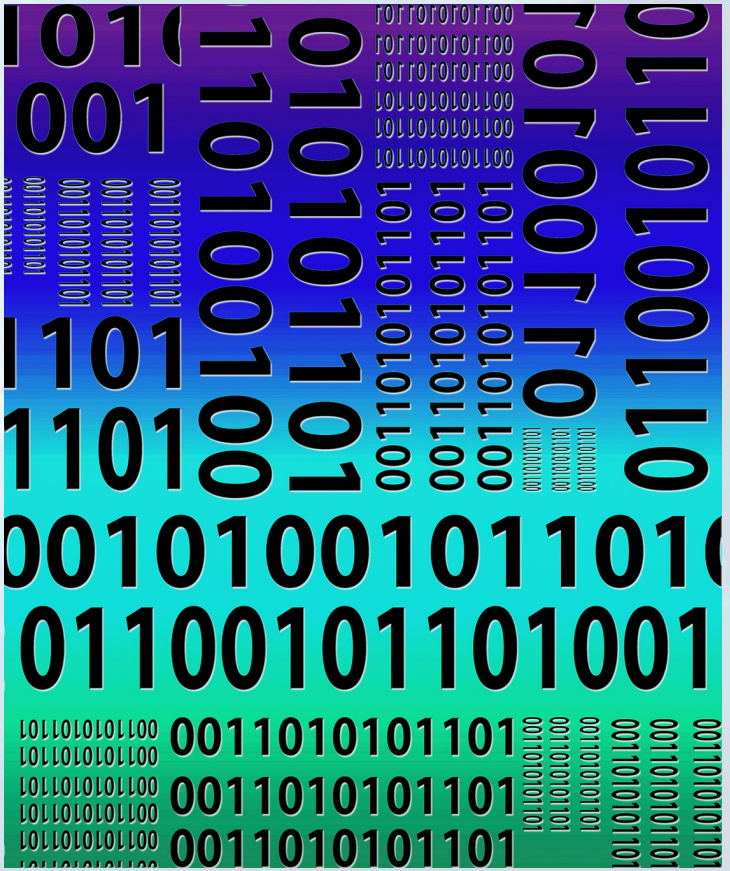
More Velocity

- Strategies:
  - simple ingest & access
  - parallelization
  - better hardware
- “Proper velocity”:
  tied to real-world demand
- “Apparent velocity”:
  trade-off for volume

Variety

How heterogeneous the data is; complexity

- Many features per item
- Irregular structure
- Need to store and retrieve different data types quickly, efficiently, cheaply
- Need to align & integrate different representations
- Dealt with using standards, specs, ontologies



Dimensions of Variety

- Content:
  - Image, spectrum, timeseries
- Form:
  - text, numeric, relational, graphical, geospatial, sensory
- Format:
  - plain-text file, .csv, fixed-width, Excel spreadsheet, HTML table
- Structure:
  - unstructured text, semi-structured email, semantically-marked-up document
- Source:
  - human-generated, automated sensor logging, scientific instruments, simulations
- Meaning:
  - “This dish is hot”
- Representation:
  - Jan. 14, 2016 vs 2016/01/14
- etc.

Value

"Business value" or ROI

Business Big Data is free; Scientific Big Data is not.

- If you're bothering to save a scientific dataset, it has value.

Scientific Big Data is Big because the big V captures value that would otherwise be missed.

- Volume: rare or thin
- Velocity: fast events / brief lifespan
- Variety: embedded in disparate relationships

Veracity

Is the data trustworthy?

- Provenance, reliability, accuracy, completeness, ambiguity
- Importance of Veracity depends on what the Value of the data is

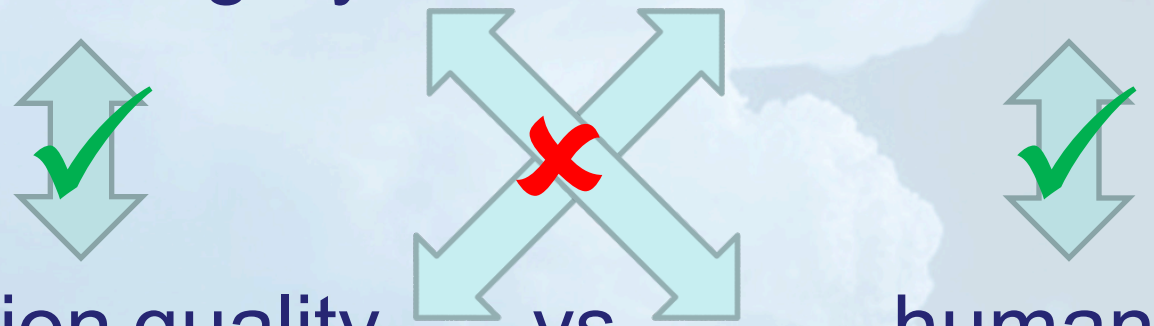
- Strategies:
  - transparent QC
  - provenance tracking
  - data management best practice
  - good governance practices
- Note: provenance and other veracity metadata can itself become Big Data

Validity

Accuracy and correctness of the data relative to a particular use

- Example: Gauging storm intensity

satellite imagery vs social media posts



prediction quality vs human impact

Variability

How the meaning of the data changes over time

- Language evolution
- Data availability
- Sampling processes
- Changes in characteristics of the data source

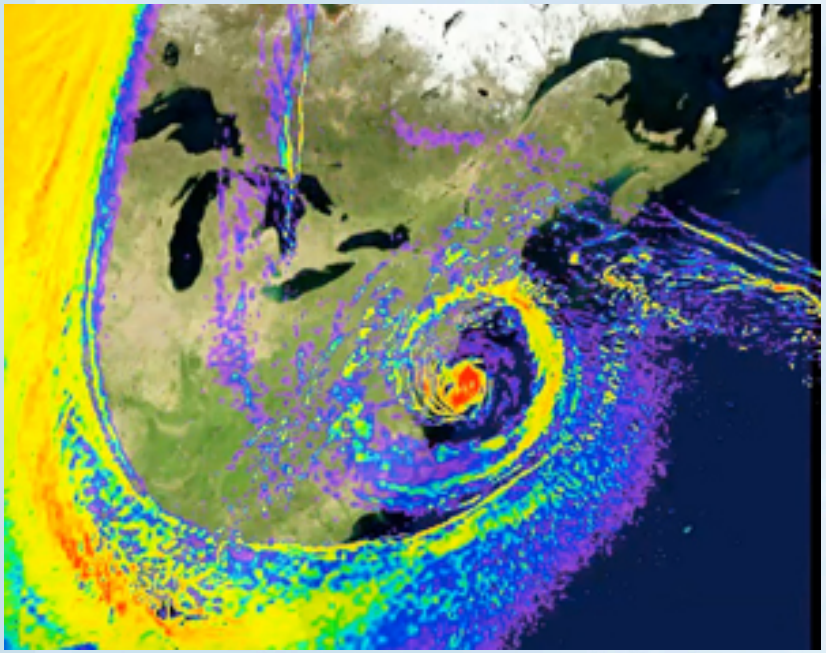
Viscosity & Volatility

- Both related to velocity
- Viscosity: *data velocity relative to timescale of event being studied*
- Volatility: *rate of data loss and stable lifetime of data*
  - (Scientific data often has practically unlimited lifespan, but social / business data may evaporate in finite time)

More V's

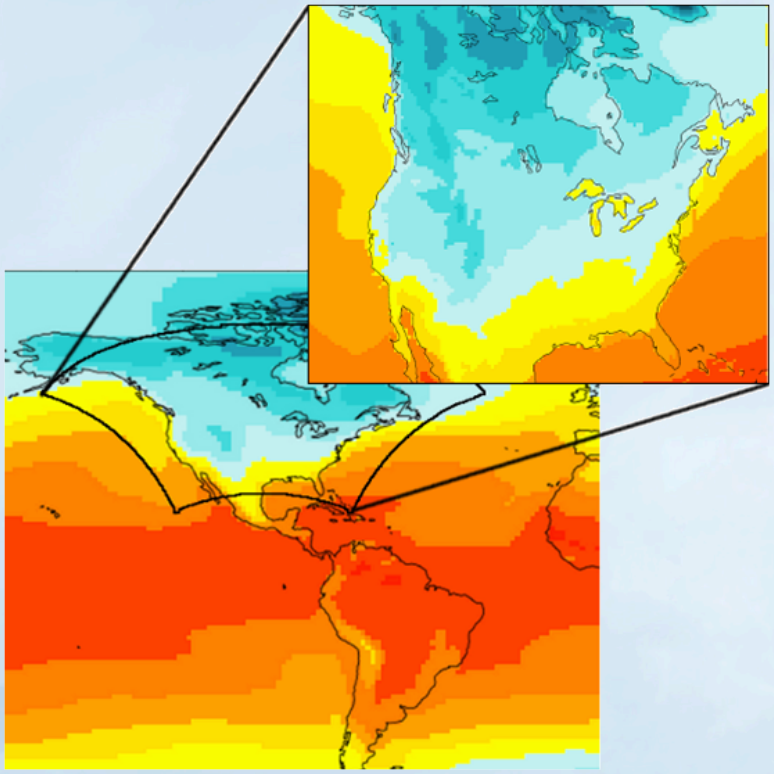
- Viability
  - Another take on value. Which data has meaningful relations to questions of interest?
- Venue
  - Where does the data live and how do you get it?
- Vocabulary
  - Metadata describing structure, content, & provenance
  - Schemas, semantics, ontologies, taxonomies, vocabularies
- Vagueness
  - Confusion about what "Big Data" means

Big Data in Climate Science



- Models:
  Big Volume
- Observations:
  Big Variety
- The Future:
  Data Services

NARCCAP: North American Regional Climate Change Assessment Program



Nest high-resolution regional climate models (RCMs) inside coarser global models (GCMs) over North America

NARCCAP Program Goals

- Generate high-res climate change scenario data for impacts analysis
- Support further dynamical downscaling experiments
- Evaluate model performance and uncertainty

    i.e.: a dynamical downscaling model intercomparison project

Experimental Design

	25 years	Two 30-year runs, current & future			
	NCEP	GFDL	CGCM3	HADCM3	CCSM
CRCM	X	--	X	--	X
ECP2	X	X	--	X	--
HRM3	X	X	--	X	--
MM5I	X	--	--	X	X
RCM3	X	X	X	--	--
WRFG	X	--	X	--	X
Timeslices		X	--	--	X

6 RCMs x 4 GCMs
+ NCEP and Timeslices
= 34 runs total

Simulation Output Archive

- 3-hourly frequency
- 50-km gridcells
- Avg domain size:
  139×112 gridpoints
- 2D variables: 35
- 3D variables: 7
- Vertical levels: 28
- NetCDF format

34 runs × 30 years × 365 days × 8 timesteps ×
139 X × 112 Y × (35 + 7×28 vars) × 4 bytes =

~40 TB total data volume

CMIP Data Volumes

- CMIPs: model comparison programs
- Used in IPCC assessment reports

	CMIP1	CMIP2	CMIP3	CMIP5	CMIP6
Year	1996	1997	2005	2010	2016
# Models	19	24	21	45	TBD
# Runs	19	48	211	841	TBD
Volume	1 GB	500 GB	36 TB	3.3 PB	~90 PB

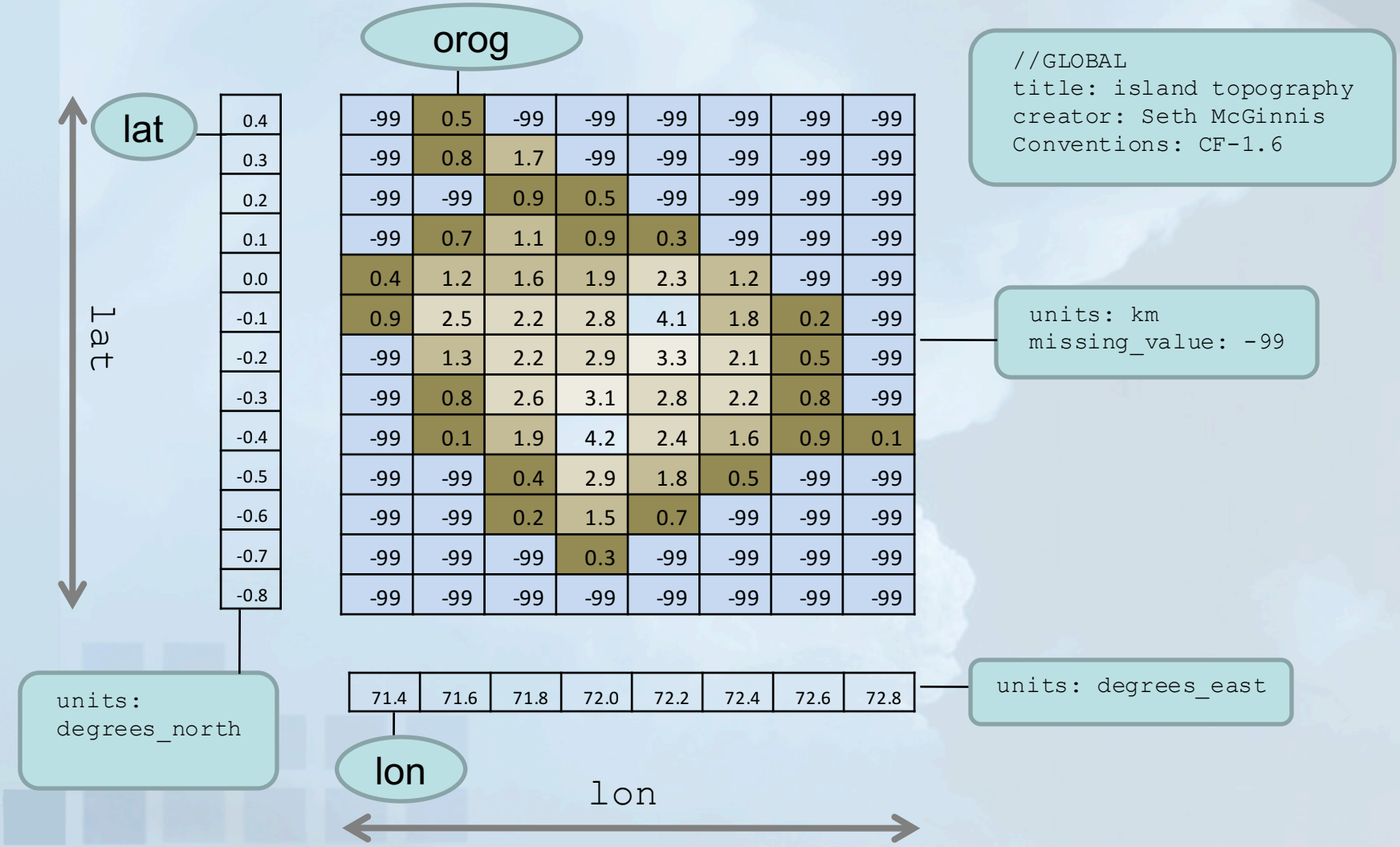
Dealing with Volume

- Distill big data down to small information
- Parallel and automated analysis
- Automation requires standardization
- Standardize by reducing Variety:
  - Format
  - Standards
  - Structure

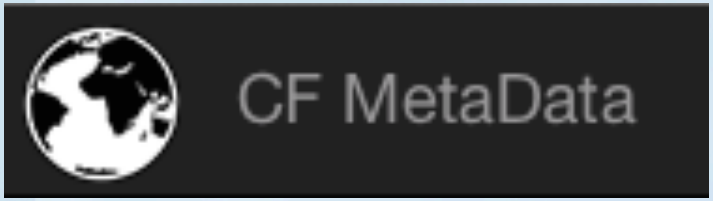
Data Formatting

- NetCDF:
  - Self-describing
  - Machine-independent
  - Integrated metadata
  - Geo-data oriented data model
- Archive Specs:
  - 1 variable per file
  - Filename conventions
  - Metadata requirements
  - Versioning
  - Standard set of outputs

NetCDF Data Model



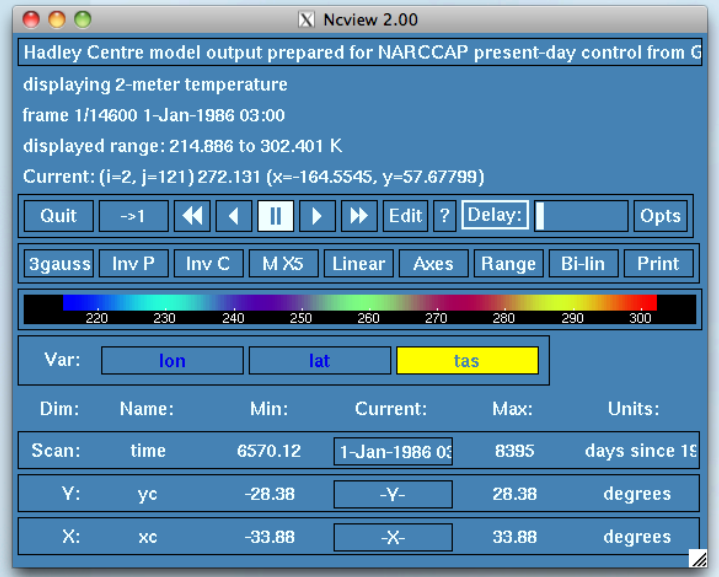
Data Standards

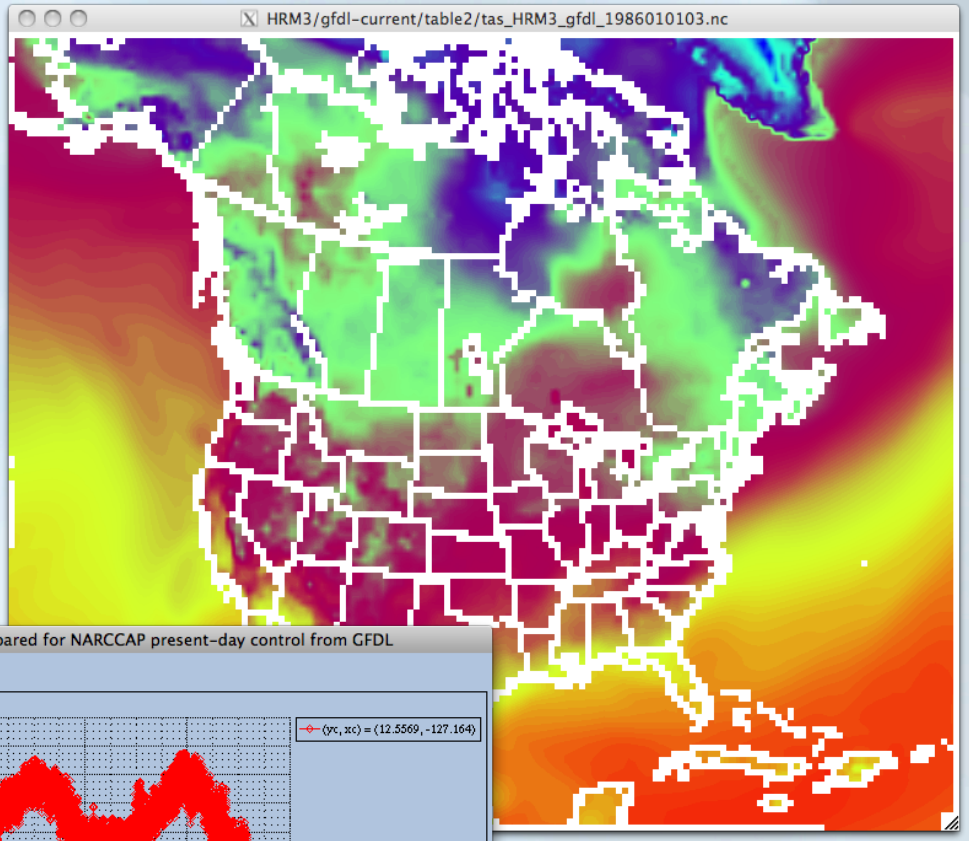


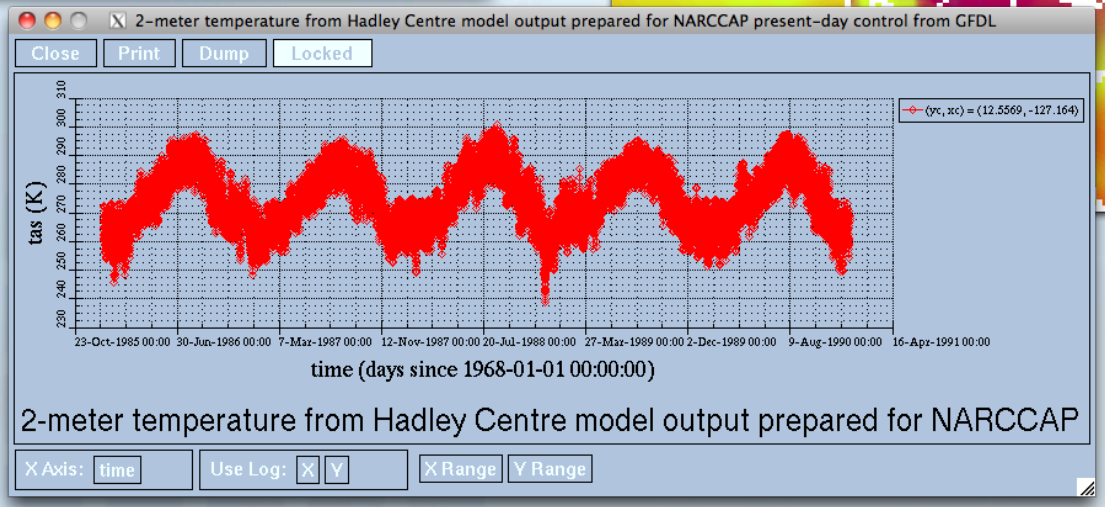
- Community-governed
- Extensive and detailed
- Resolves many representation questions

- Rules governing file element naming conventions and metadata contents for NetCDF
- Controlled vocabulary for semantic attributes (physical quantity, canonical units)

NetCDF + CF enables smart tools



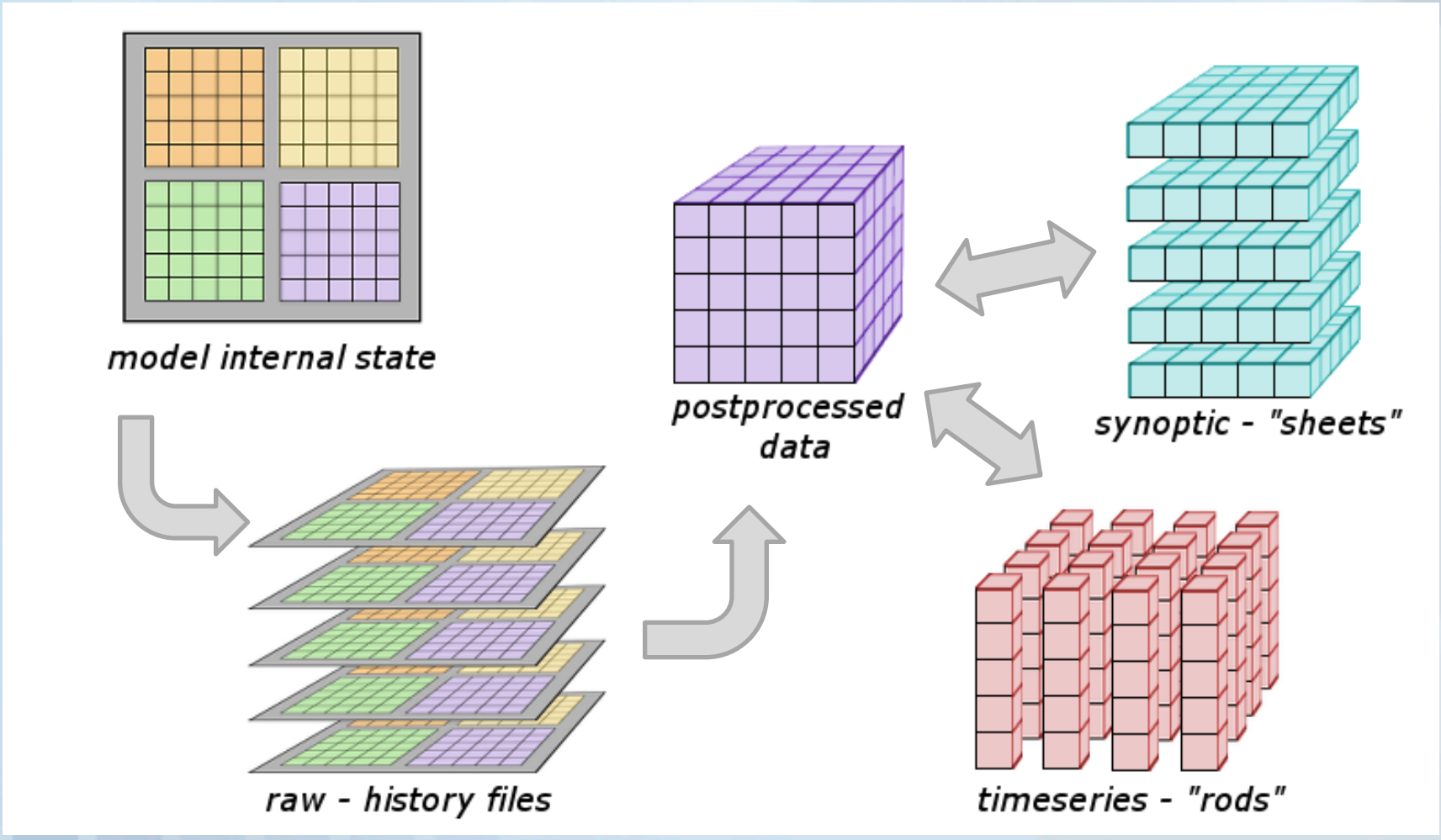




The Data Structure Quandry

- Most natural structure for data generation is not the same as for data usage
- Restructuring ("transposition") requires ALL THE DATA
  - Requires memory, time, or parallelization
  - Big task if done all at once at the end
  - Less so if amortized across data generation

Data Structure



Data Structure Solutions

- Best structure depends on use
  - Sheets vs rods vs blocks vs granules
  - Multiple uses → multiple structures?
- Task-parallel PyReshaper tool
- Better model outputs?

(Wouldn't it be nice to skip postprocessing entirely?)

Archiving Lessons Learned

1. Automate everything you can
2. Be stringent about specs & standards
3. Minimize splitting; maximize segregation
4. Version your data & plan for iteration
5. Structure data to suit end uses
6. Prioritize based on end value
7. Filenames are not metadata

ACADIS: Advanced Cooperative Arctic Data Information Service

- Common repository for Arctic data
- Heterogeneous: images, audio, text, spreadsheets, sensors
- Long-tail: 3500 datasets, 500k files
- Irreducible Variety
- Goal: retain variety, *maximize discoverability*

Maximizing Discoverability

- Use a short "discovery metadata" profile common to EVERYTHING in the archive
  - Then: Dublin Core
  - Now: ISO-19115
- Doesn't need to be comprehensive; 80% coverage nets big improvement
- Controlled vocabulary for semantics
- Well-documented profile is important

Documentation Is Hard

- Data producers know the information but don't need it
- Data consumers need the information but don't know it
- Incentives?
- Unsolved problem

General Strategies

- Have to meet data producers' needs
- Data curation
  - Agile
  - Humans still beat machines
- Be a middleman in the data supply chain

The Future: Data Services

Analyze and transform Big Data before transfer to end user

Network $ > Disk $ > CPU $

∴ Server-side computation makes sense

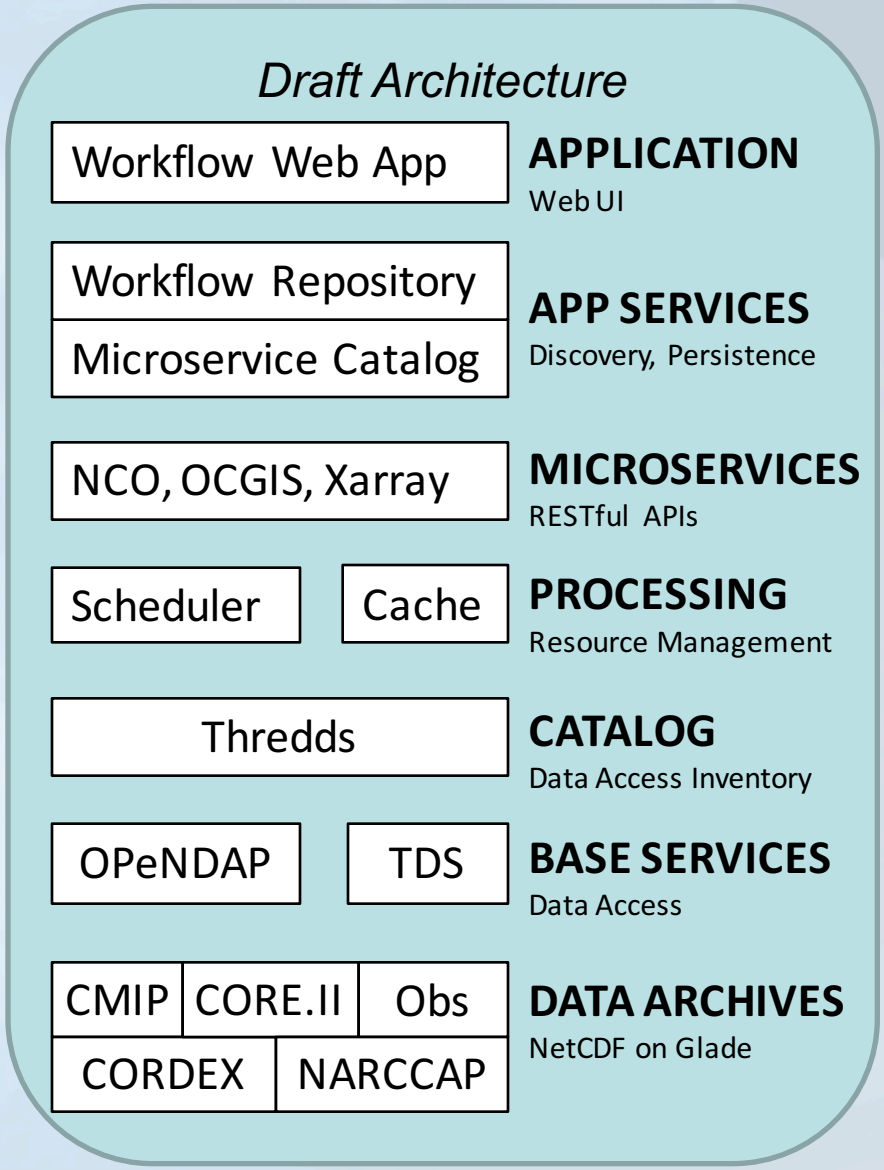
- Reduce the need for large data downloads
- Capture expertise as automated processing
- Improve usability for applications & non-specialists

Types of Data Services

- Data archive hosts raw data
- Data access services
  - don't change data values (can change formats)
- Data transformation services
  - modify data values
  - derived data products
- Visualization and interpretation services
  - non-data outputs
- *Provenance threaded through all services*

CAPSTONE

- Web-based server-side data analysis
- Modular community-governed toolbox architecture
- Chained data services
- Workflow persistence
  - Share, cite, discuss, etc.



Questions?