Early Experience Prototyping a Science Data Server for Environmental Data

Deb Agarwal, LBL (daagarwal@lbl.gov)

Catharine van Ingen, MSFT (vaningen@microsoft.com)

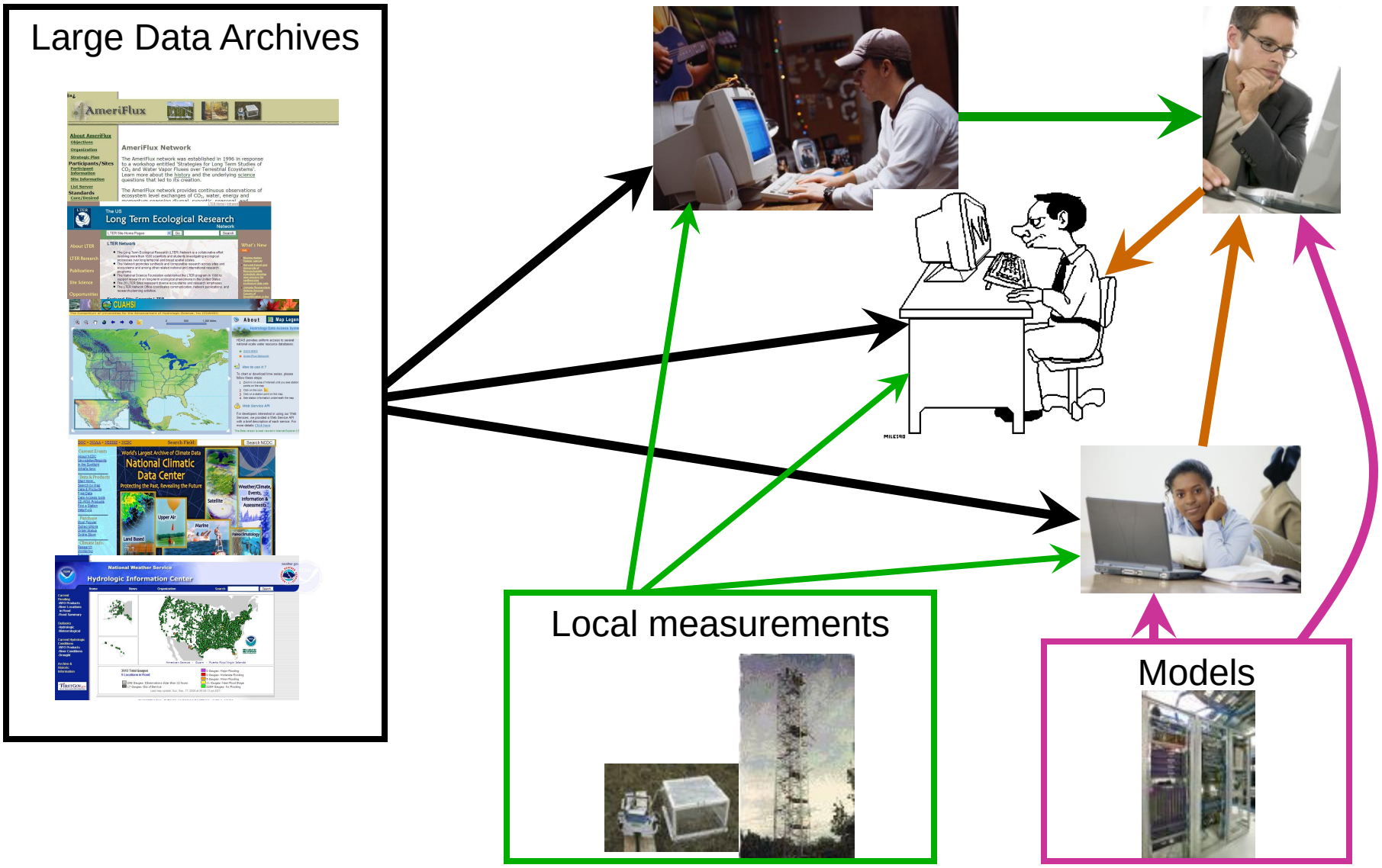
20 September 2006

Outline

- Landscape
  - Data archives and other sources
  - Typical small group collaboration needs
  - Examples using “Ameriflux”
- Science Data Server
  - Goals and ideal capabilities
  - Approach
  - Experiences with the current system
- Next generation
  - Next set of development efforts
  - Research issues
- Conclusion

Unprecedented Data Availability

Typical Data Flow Today



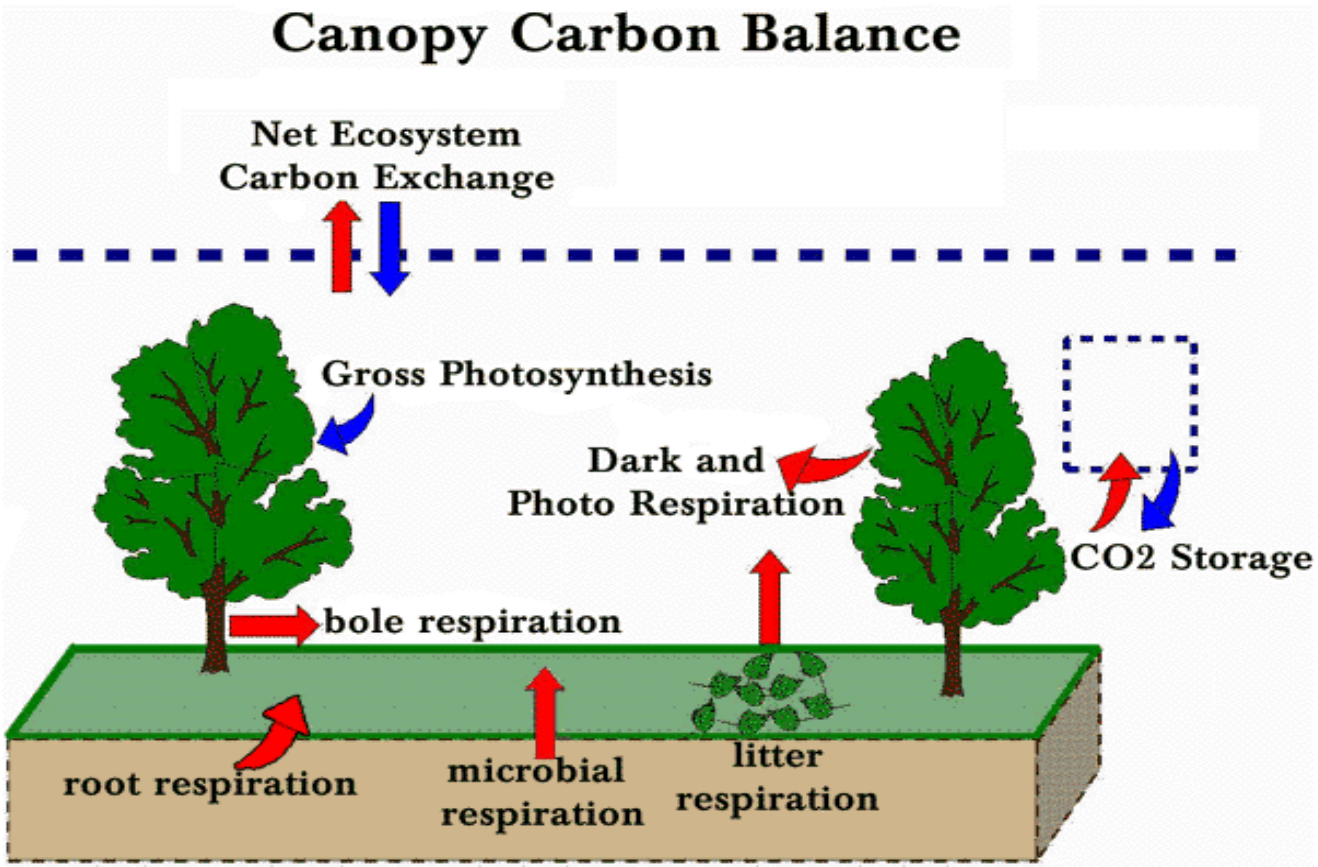
AmeriFlux

Ameriflux Collaboration Overview

- 149 Sites across the Americas
- Each site reports a minimum of 22 common measurements.
- Communal science – each principle investigator acts independently to prepare and publish data.
- Data published to and archived at Oak Ridge.
- Total data reported to date on the order of 150M half-hourly measurements.

- http://public.ornl.gov/ameriflux/

What A Tower Sees



$$F_c + F_{storage} = -NEE = P_{net} + R_{leaf} + R_{wood} + R_{roots} + R_{microbes}$$

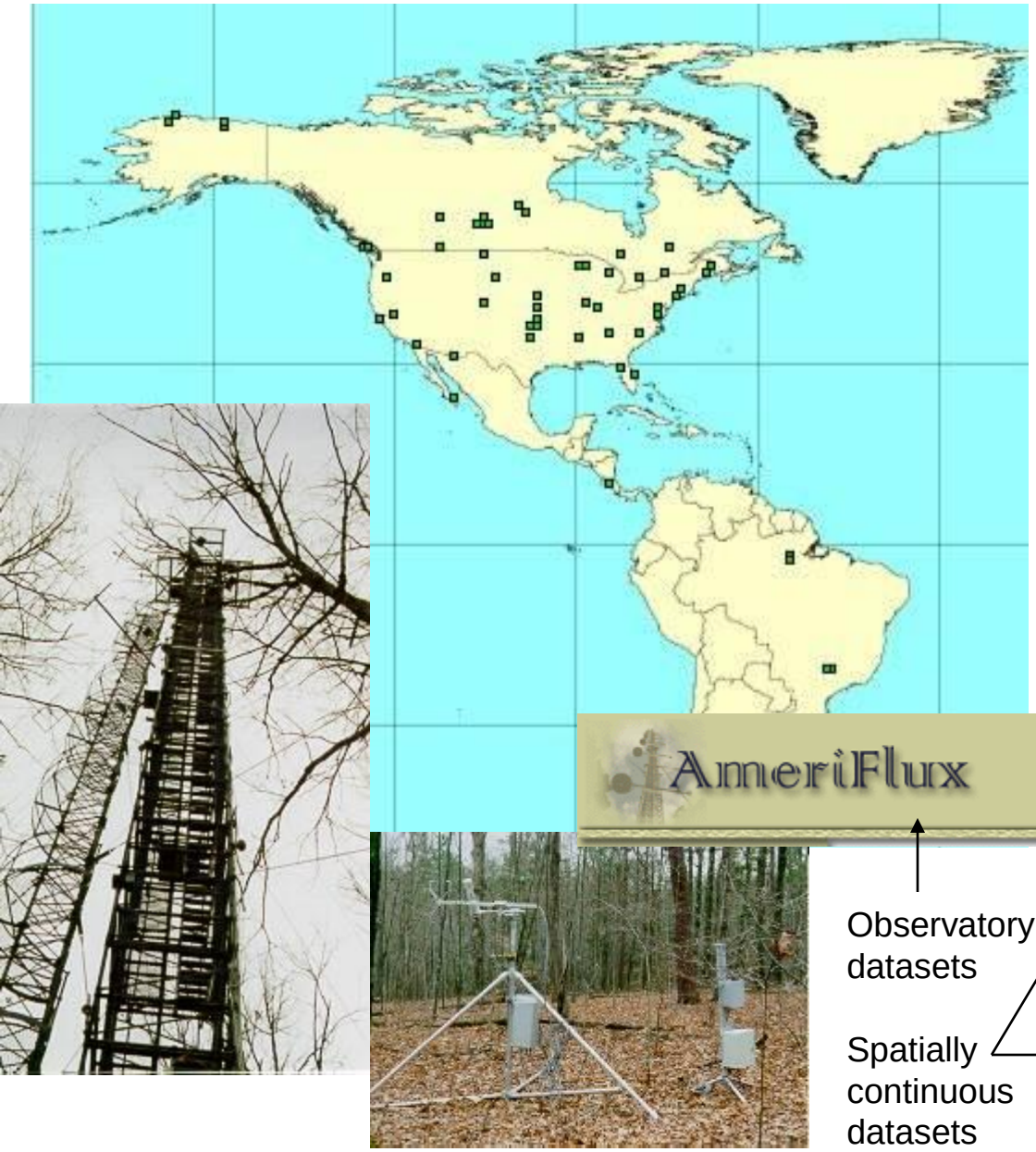
1. Applications of eddy covariance measurements, Part 1: Lecture on Analyzing and Interpreting CO2 Flux Measurements, Dennis Baldocchi, CarboEurope Summer Course, 2006, Namur, Belgium (http://nature.berkeley.edu/biometlab/lectures/)

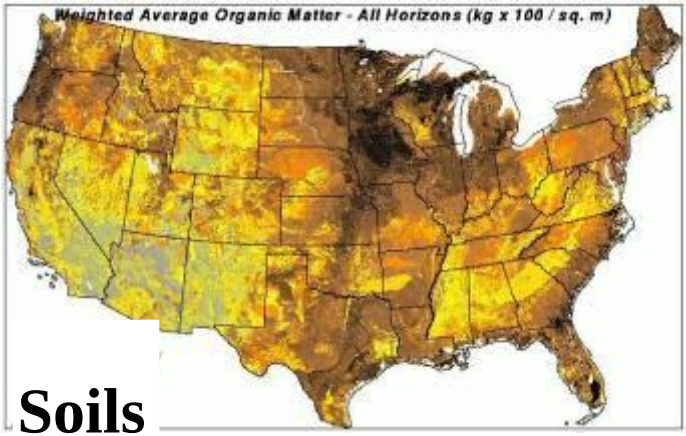
Example Carbon-Climate Investigations

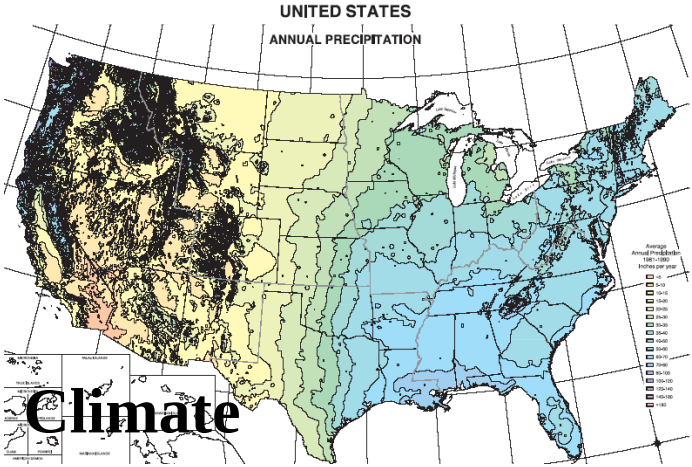
- Net carbon exchange for the ecosystem
- Impact of climate change on the greening of ecosystems
  - Start of leaf growth
  - Duration of photosynthesis
- Effects of early spring on carbon uptake
- Role of ecosystem and latitude on carbon flux
- Effect of various pollution sources on carbon in atmosphere and carbon balance

Measurements Are Not Simple or Complete

- Gaps in the data
  - Quiet nights
  - Bird poop
  - High winds
  - ….
- Difficult to make measurements
  - Leaf area index
  - Wood respiration
  - Soil respiration
  - …
- Localized measurements – tower footprint
- Local investigator knowledge important
- PIs' science goals are not uniform across the towers

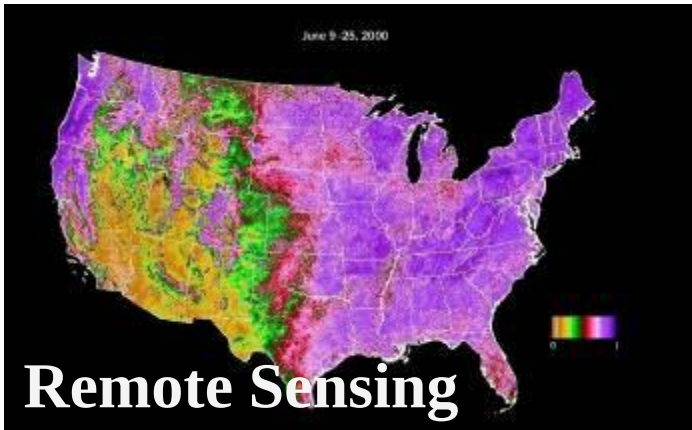






Observatory datasets

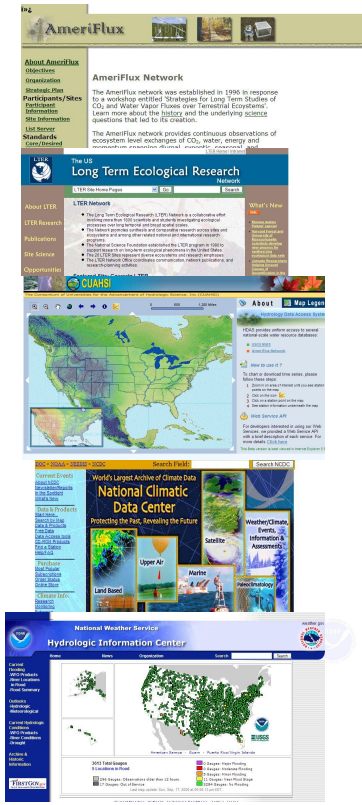
Spatially continuous datasets

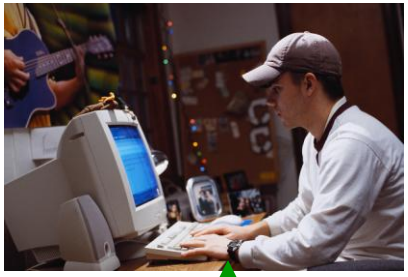


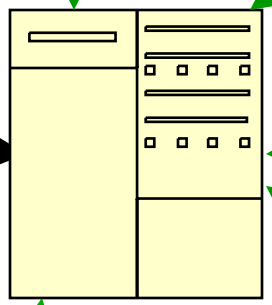
Examples of Carbon-Climate Datasets

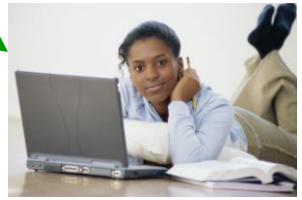
Scientific Data Server

Large Data Archives









Local measurements

Scientific Data Server - Goals

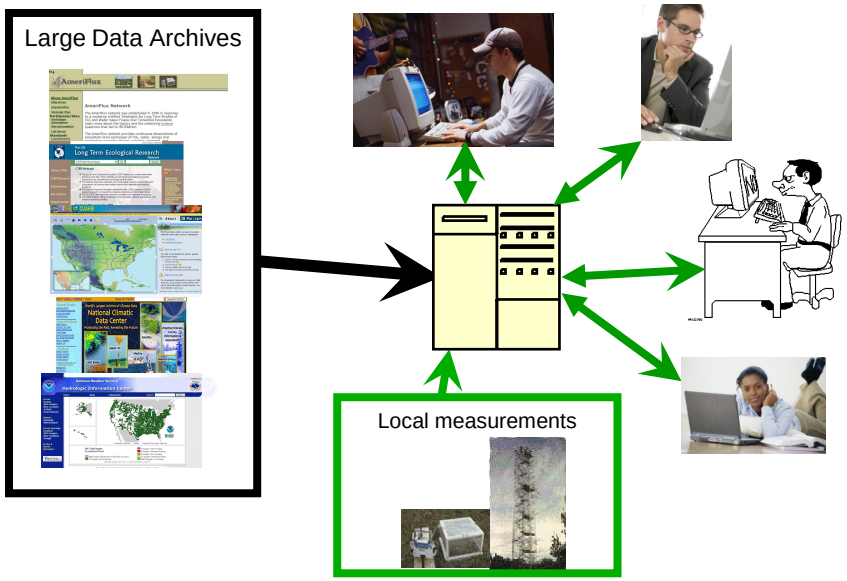
- Act as a local repository for data and metadata assembled by a small group of scientists from a wide variety of sources
  - Simplify provenance by providing a common “safe deposit box” for assembled data
- Interact simply with existing and emerging Internet portals for data and metadata download, and, over time, upload
  - Simplify data assembly by adding automation
  - Simplify name space confusion by adding explicit decode translation
- Support basic analyses across the entire dataset for both data cleaning and science
  - Simplify mundane data handling tasks
  - Simplify quality checking and data selection by enabling data browsing

Scientific Data Server - Non-Goals

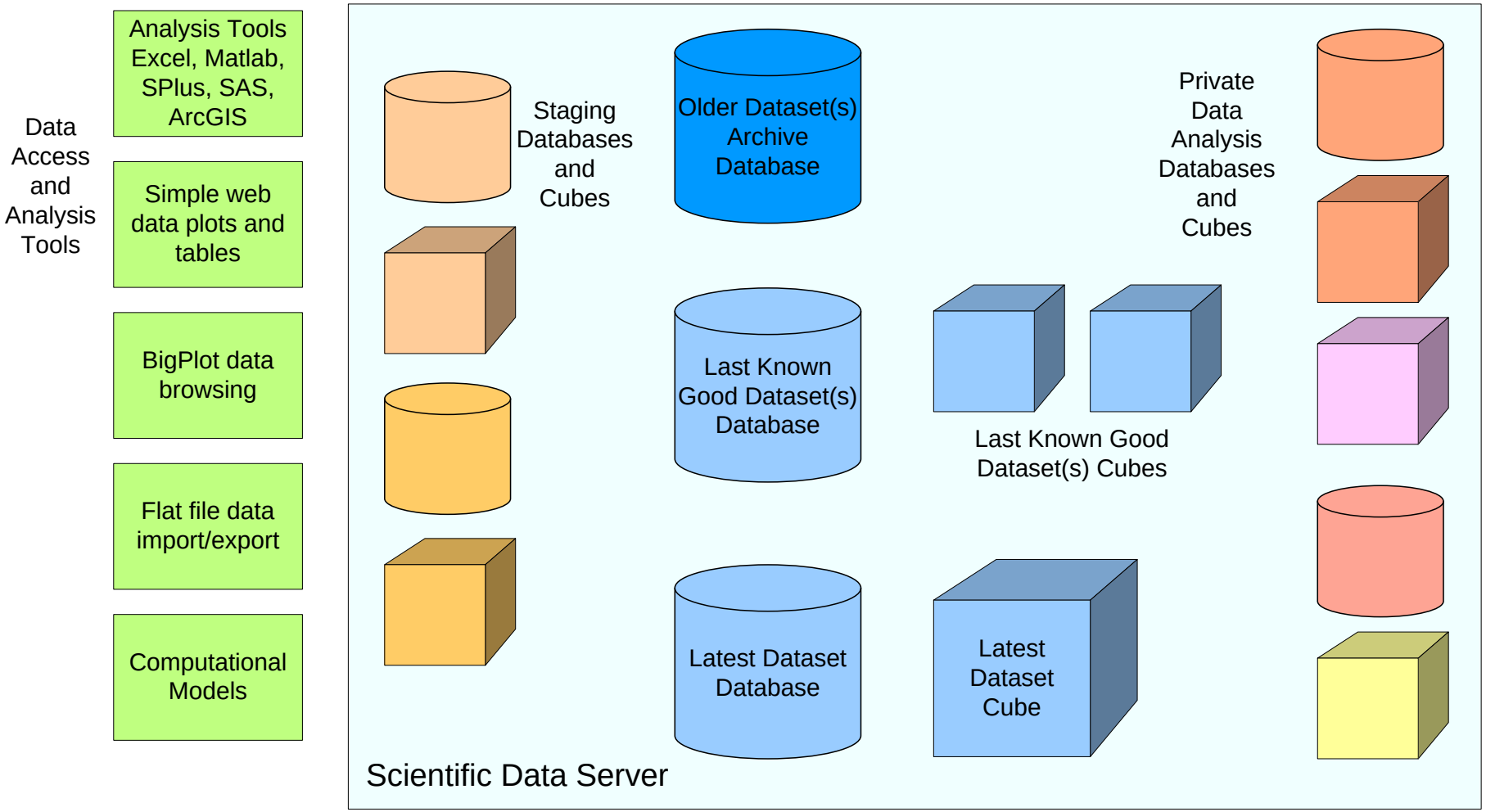
- Replace the large Internet data source sites
  - The technology developed may be applicable, but the focus is on the group collaboration scale and usability
  - Very large datasets require different operational practices
- Perform complex modeling and statistical analyses
  - There are a lot of existing tools with established trust based on long track records
  - Only part of a full LIMS (laboratory information management system)
- Develop a new standard schema or controlled vocabulary
  - Other work on these is progressing independently
  - Due to the heterogeneity of the data, more than one such standard seems likely to be relevant

Scientific Data Server - Workflows

- Staging: adding data or metadata
  - New downloaded or field measurements added
  - New derived measurements added
- Editing: changing data or metadata
  - Existing older measurements re-calibrated or re-derived
  - Data cleaning or other algorithm changes
  - Gap filling
- Sharing: making the latest acquired data available rapidly
  - Even before all the checks have been made
  - Browsing new data before more detailed analyses
- Private Analysis: Supporting individual researchers (MyDB)
  - Stable location for personal calibrations, derivations, and other data transformations
  - Import/Export to analysis tools and models
- Curating: data versioning and provenance
  - Simple parent:child versioning to track collections of data used for specific uses

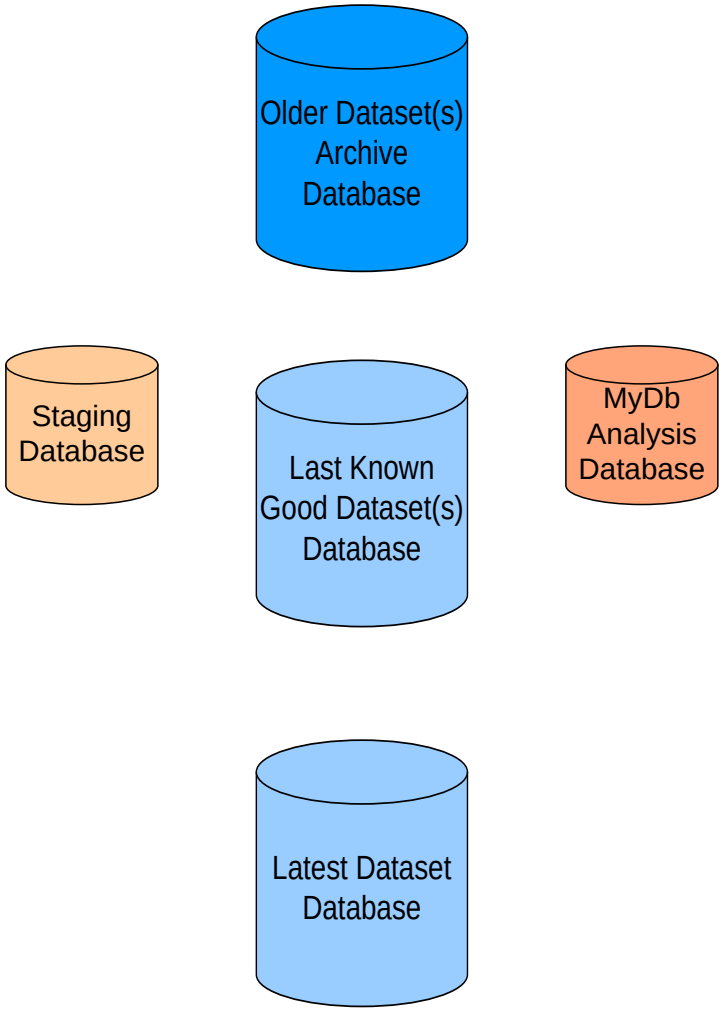


Scientific Data Server - Logical Overview



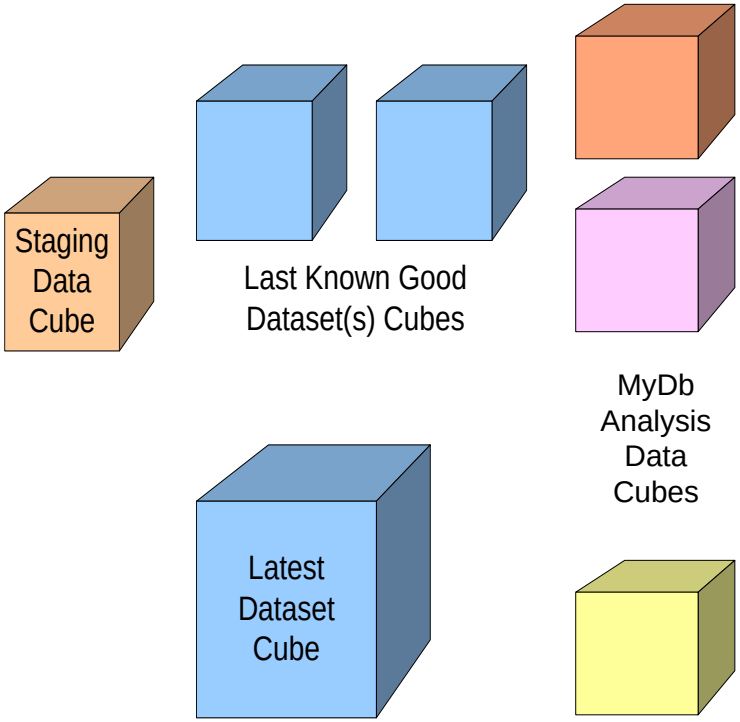
Databases

- All descriptive metadata and data held in relational databases
  - Metadata is important too!
- While separate databases are shown, the datasets may actually reside in a single database
  - Mapping is transparent to the scientist
  - Separate databases used for performance
  - Unified databases used for simplicity
- New metadata and data are staged with a temporary database
  - Minimal quality checks applied
  - All name and unit conversions
- Data may be exported to flat file, copied to a private MyDb database, directly accessed programmatically, or ?



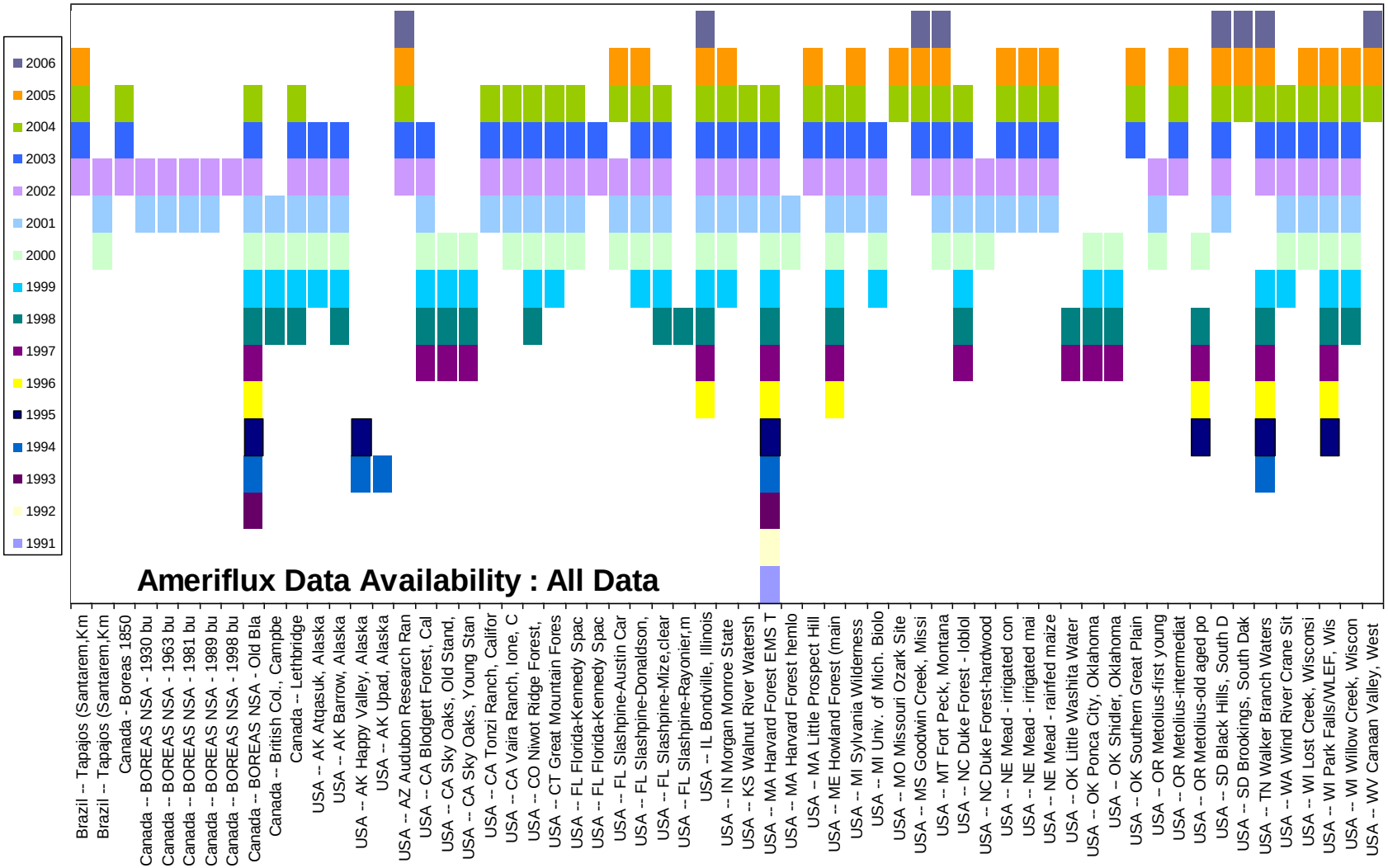
Data Cubes

- A data cube is a database specifically for data mining (OLAP)
  - Initially developed for commercial needs like tracking sales of Oreos and milk
  - Simple *aggregations* (sum, min, or max) can be pre-computed for speed
  - Additional calculations (median) can be computed dynamically
  - Both operate along *dimensions* such as time, site, or datumtype
  - Constructed from a relational database
  - A specialized query language (MDX) is used
- Client tool integrations is evolving
  - Excel PivotTables allow simple data viewing
  - More powerful charting with Tableaux or ProClarity (commercial mining tools)



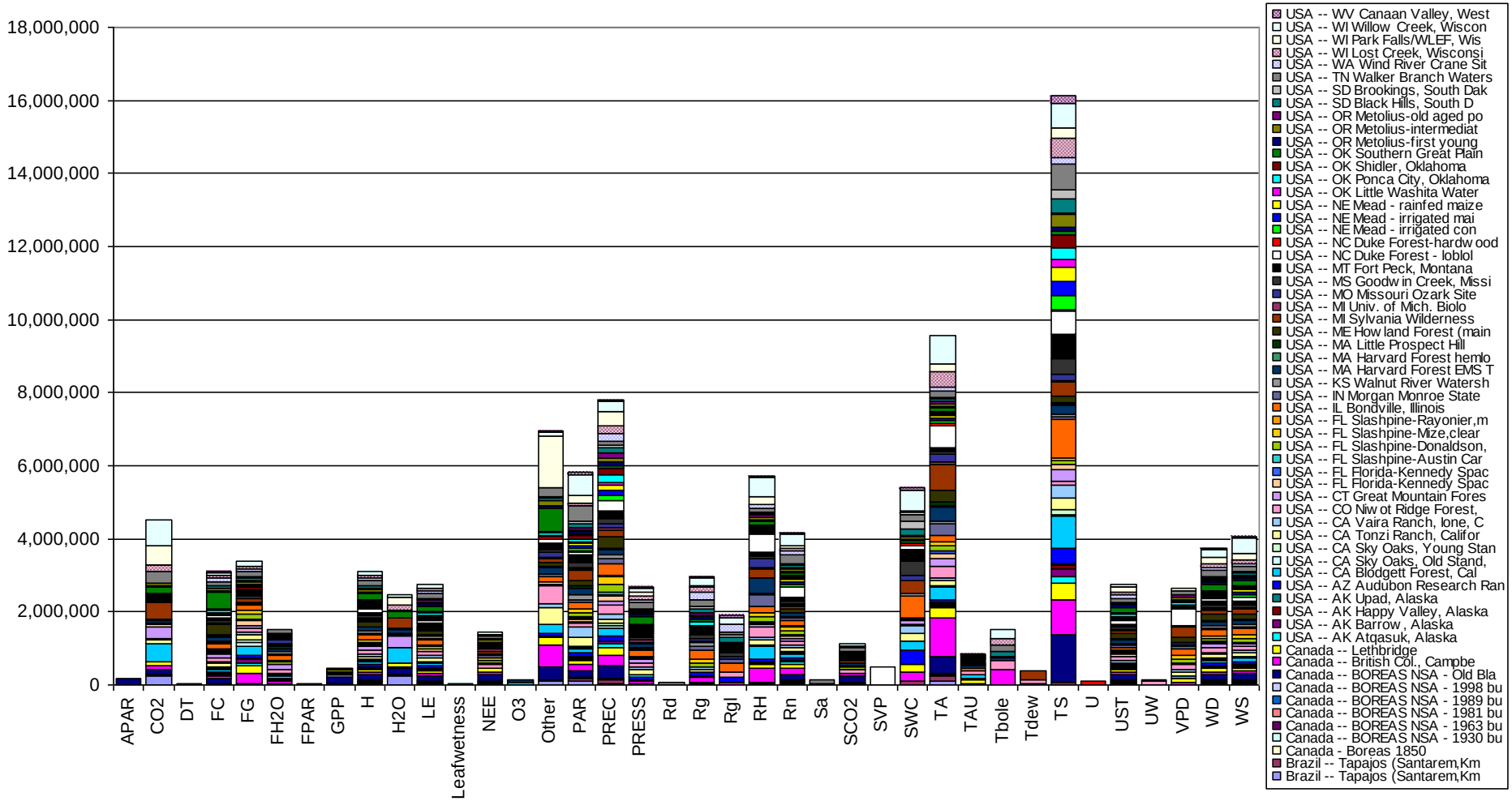
Browsing For Data Availability

Sites Reporting Data Colored by Year



Browsing For Data Availability

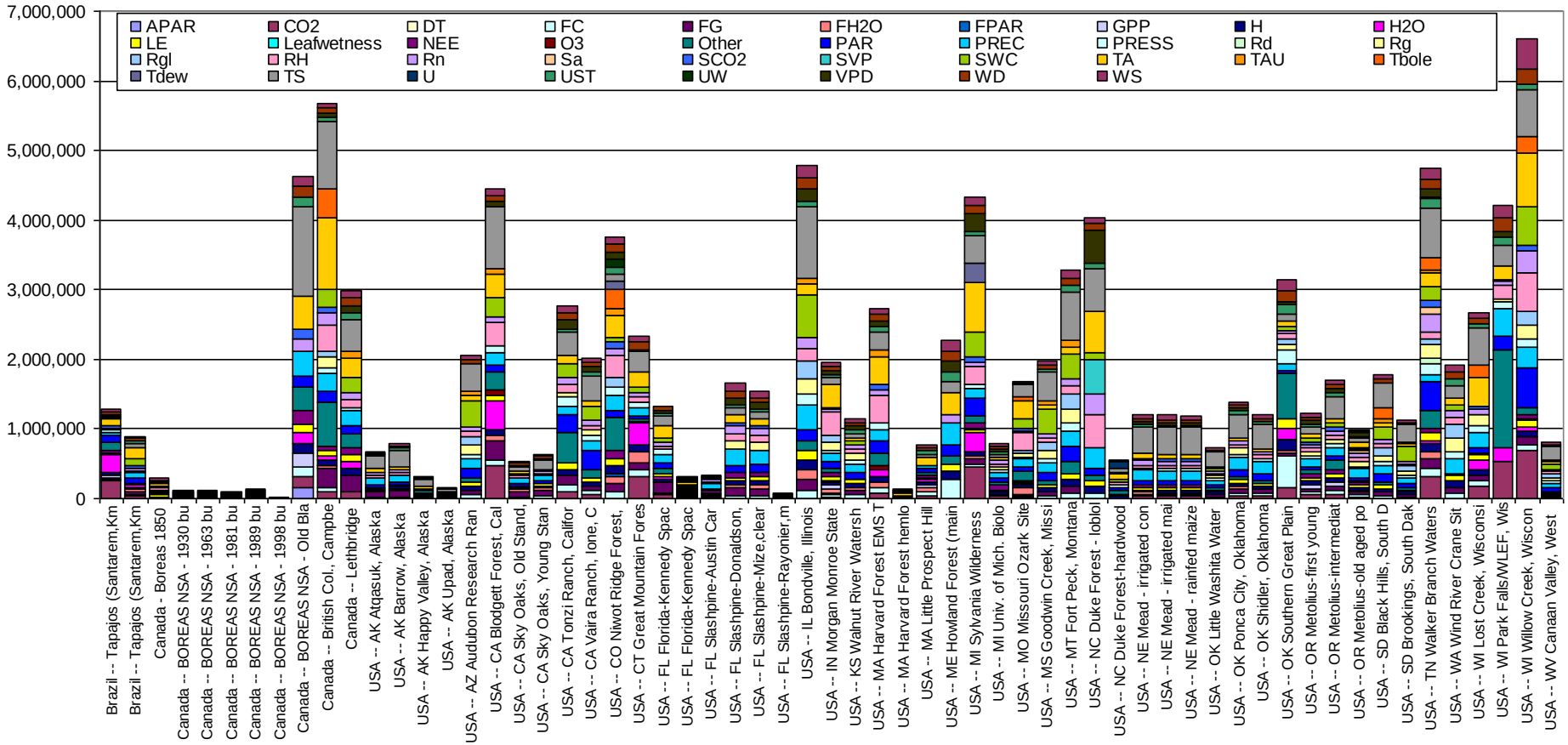
Total Data Availability by Type Colored by Site



Data type reporting is far from uniform across type

Browsing for Data Availability

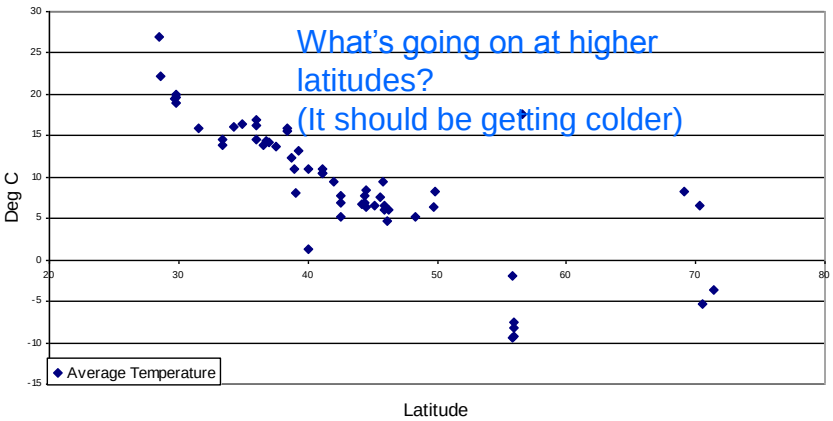
Total Data Availability by Site Colored by Type

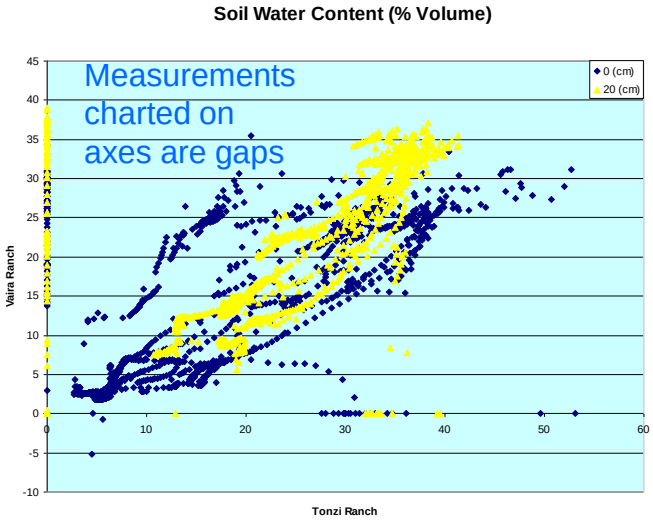


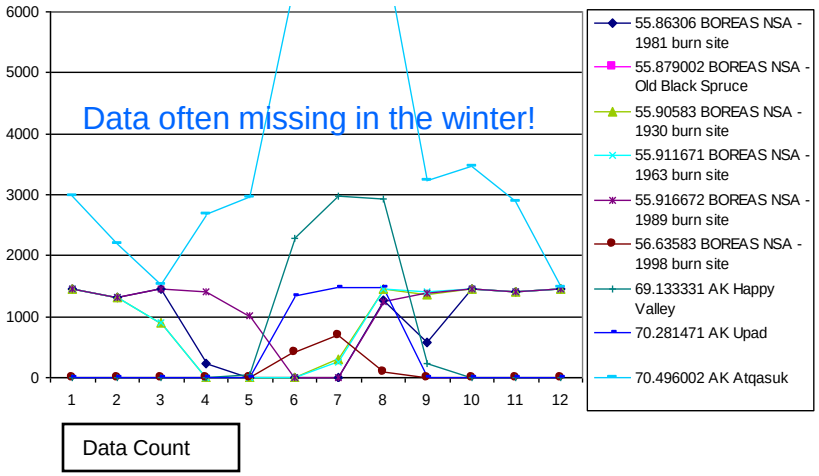
Sites report more data either because of longevity or specific research interests

Browsing for Data Quality

- Real field data has both short term gaps and longer term outages
  - The utility of the data depends on the nature of the science being performed
  - Browsing data counts can give rapid insight into how the data can be used before more complex analyses are performed

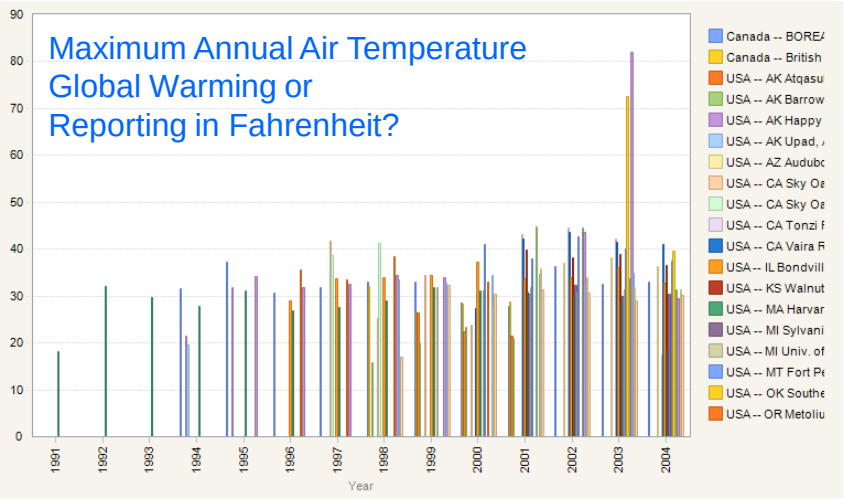


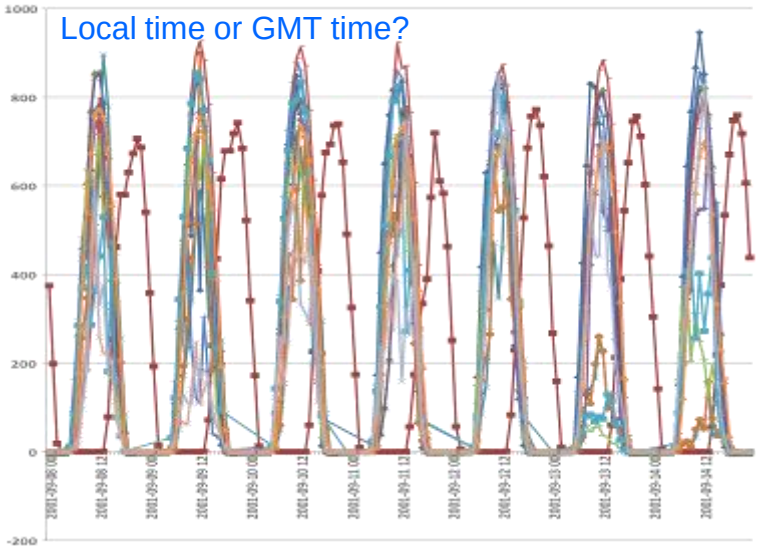


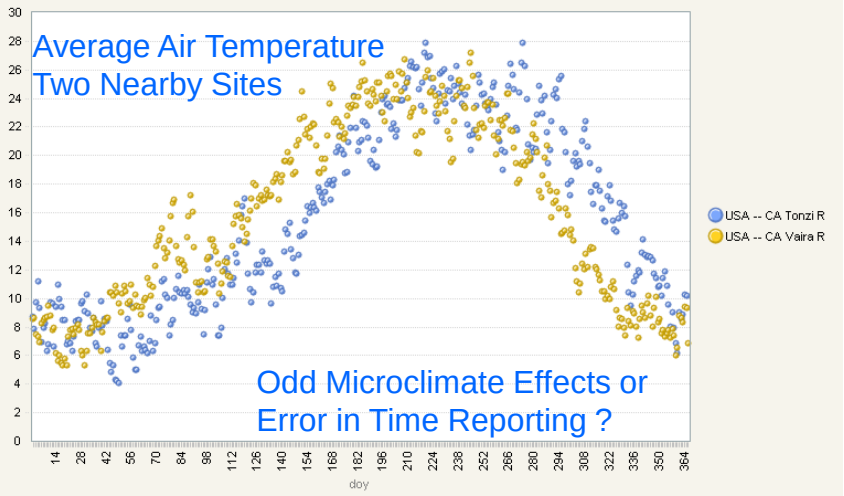


Browsing for Data Quality

- Real field data has unit and time scale conversion problems
  - Sometimes easy to spot in isolation
  - Sometimes easier to spot when comparing to other data
  - Browsing data values can give rapid insight into how the data can be used before more complex analyses are performed







Lessons Learned To Date

- Metadata is as important as data
  - Comparing sites of like vegetation, climate is as important as latitude or other physical quantity
  - *Curate the two together*
- Controlled vocabularies are hard
  - Humans like making up names and have a hard time remembering 100+ names
  - *Assume a decode step in the staging pipeline*
- There are at least three database schema families and two cube construction approaches
  - Everyone has a favorite
  - Each has advantages and disadvantages
  - *Automate the maintenance and use the right one for the right job*
- Visual programming tools are great for prototyping
  - But debugging and maintenance can hit a wall
  - *It's easy to overbuild – use when "good enough"*
- Data analysis and data cleaning are intertwined
  - Data cleaning is always on-going
  - *Share the simple tools and visualizations*

"Another decade or so, and it'll be warm enough for us."

"We have lots of information technology. We just don't have any information."

The saga continues at http://dsd.lbl.gov/BWC/amfluxblog/ and http://research.microsoft.com/~vaningen/BWC/BWC.htm

Near Term Futures

- Improve current capabilities
  - Assemble gap-filled and non-gap filled data sets
  - Implement incremental data staging to enable speedy and simple data editing by an actual scientist (rather than a programmer)
  - Implement expanded metadata handling to enable scientist to add site characteristics and sort sites on those expanded definitions
  - Add basic reporting capabilities for server-side browsing of data availability to speed and simplify locating “interesting” data
- Apply Data Server capabilities to a different set of data with different (but related) science
  - Considering either Russian River or Yosemite Valley hydrological data
  - Will be automating download from multiple different national data sets
  - Spatial (GIS) analyses more important
  - Linkage with imagery data necessary for science

Longer Term Futures

- Handling imagery and other remote sensing information
  - Curating images is different from curating time series data
  - Using both together enables new science and new insights
  - Graphical selection and display of data
- Support for user specified calculations within the database
- Support for direct connections to analysis and statistical packages
- Linkage with models
  - Additional (emerging) data standards such as NetCDF
  - Handling “just in time” data delivery and model result curation
- Data mining subscription services
- Handling of a broader array of data types
- Support for workflow tools

Conclusions

- Large data archives create the opportunity to
  - Do science at the regional and global scale
  - Combine data from multiple disciplines
  - Perform historical trend analysis
- Small scientific collaborations need help to
  - Perform analyses using more data than they can currently manage
  - Enable data handling and versioning
  - Store the currently needed data and metadata
  - Browse the data for science
- It's the science, not the computer science
  - Computer science research can certainly help

URLs

- Berkeley Water Center (BWC)

  http://esd.lbl.gov/BWC/
- Microsoft Project at BWC

  http://esd.lbl.gov/BWC/thrust_areas/mstci.html
- Ameriflux Project

  http://esd.lbl.gov/BWC/thrust_areas/ameriflux.html

  http://dsd.lbl.gov/BWC/amfluxblog/