SDI and Science

Michael F. Goodchild
University of California
Santa Barbara

Outline

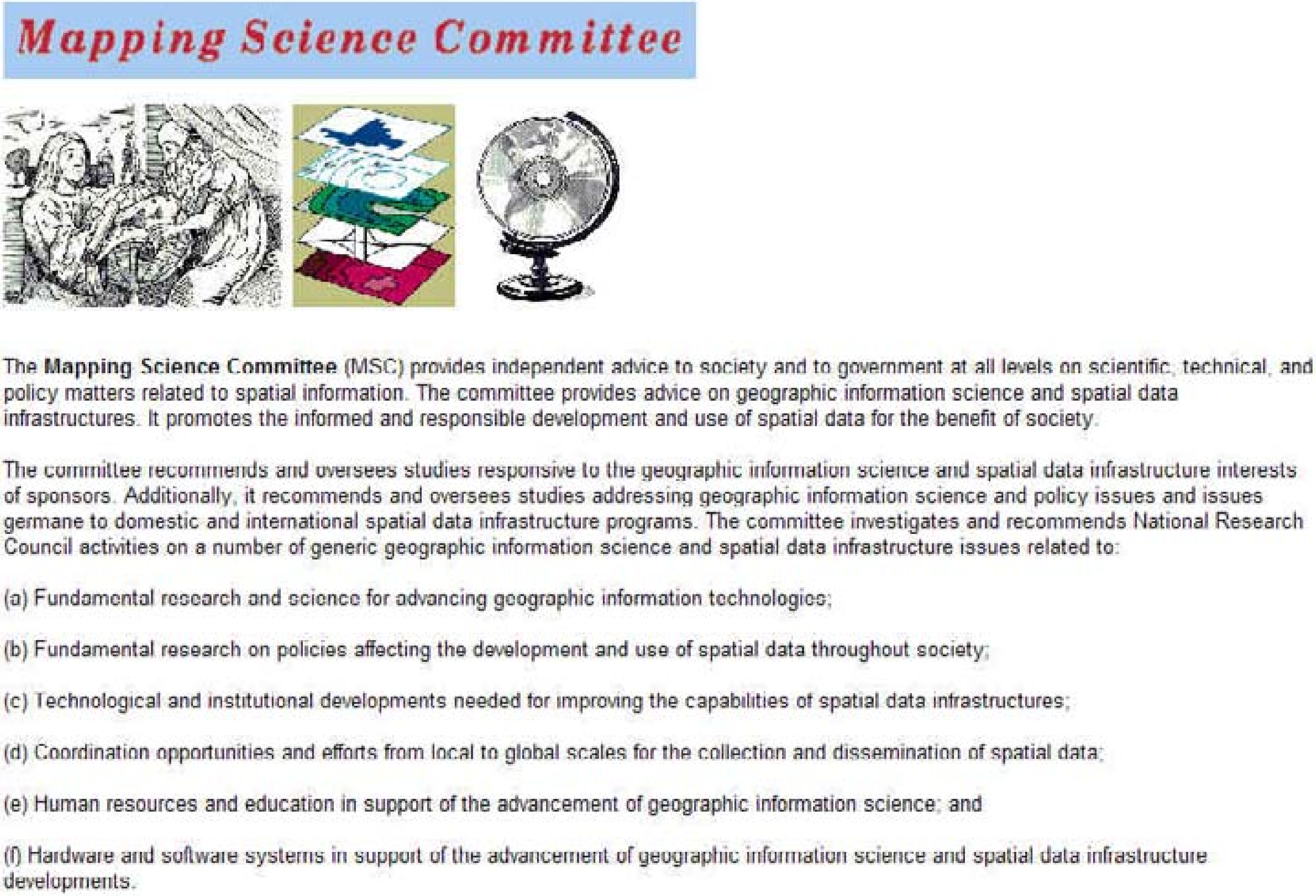
- A US perspective on SDI
- Some outstanding research issues
- SDI and the scientific enterprise

The US context

- National Research Council
  - operational arm of the National Academies
  - advice to government
  - $300 million annual budget
  - Mapping Science Committee
    - advice to US Geological Survey etc.
- Jeffersonian democracy
  - public ownership of federal IP
  - no federal copyright
  - Feist decision
    - copyright of style but not facts

MSC

- Originated NSDI in 1992 as a policy framework for spatial data
  - 6+ reports since 1993
  - www.nas.edu under Earth Science, Board on Earth Science and Resources
- NSDI administered by the Federal Geographic Data Committee
  - by Executive Order of the President, 1994



Mapping Science Committee

The **Mapping Science Committee** (MSC) provides independent advice to society and to government at all levels on scientific, technical, and policy matters related to spatial information. The committee provides advice on geographic information science and spatial data infrastructures. It promotes the informed and responsible development and use of spatial data for the benefit of society.

The committee recommends and oversees studies responsive to the geographic information science and spatial data infrastructure interests of sponsors. Additionally, it recommends and oversees studies addressing geographic information science and policy issues and issues germane to domestic and international spatial data infrastructure programs. The committee investigates and recommends National Research Council activities on a number of generic geographic information science and spatial data infrastructure issues related to:

(a) Fundamental research and science for advancing geographic information technologies;

(b) Fundamental research on policies affecting the development and use of spatial data throughout society;

(c) Technological and institutional developments needed for improving the capabilities of spatial data infrastructures;

(d) Coordination opportunities and efforts from local to global scales for the collection and dissemination of spatial data;

(e) Human resources and education in support of the advancement of geographic information science; and

(f) Hardware and software systems in support of the advancement of geographic information science and spatial data infrastructure developments.

MSC Studies (In Progress):

- * Beyond Mapping: The Challenges of New Technologies in the Geographic Information Sciences
- * National Needs for Coastal Mapping and Charting
- * Licensing Geographic Data and Services

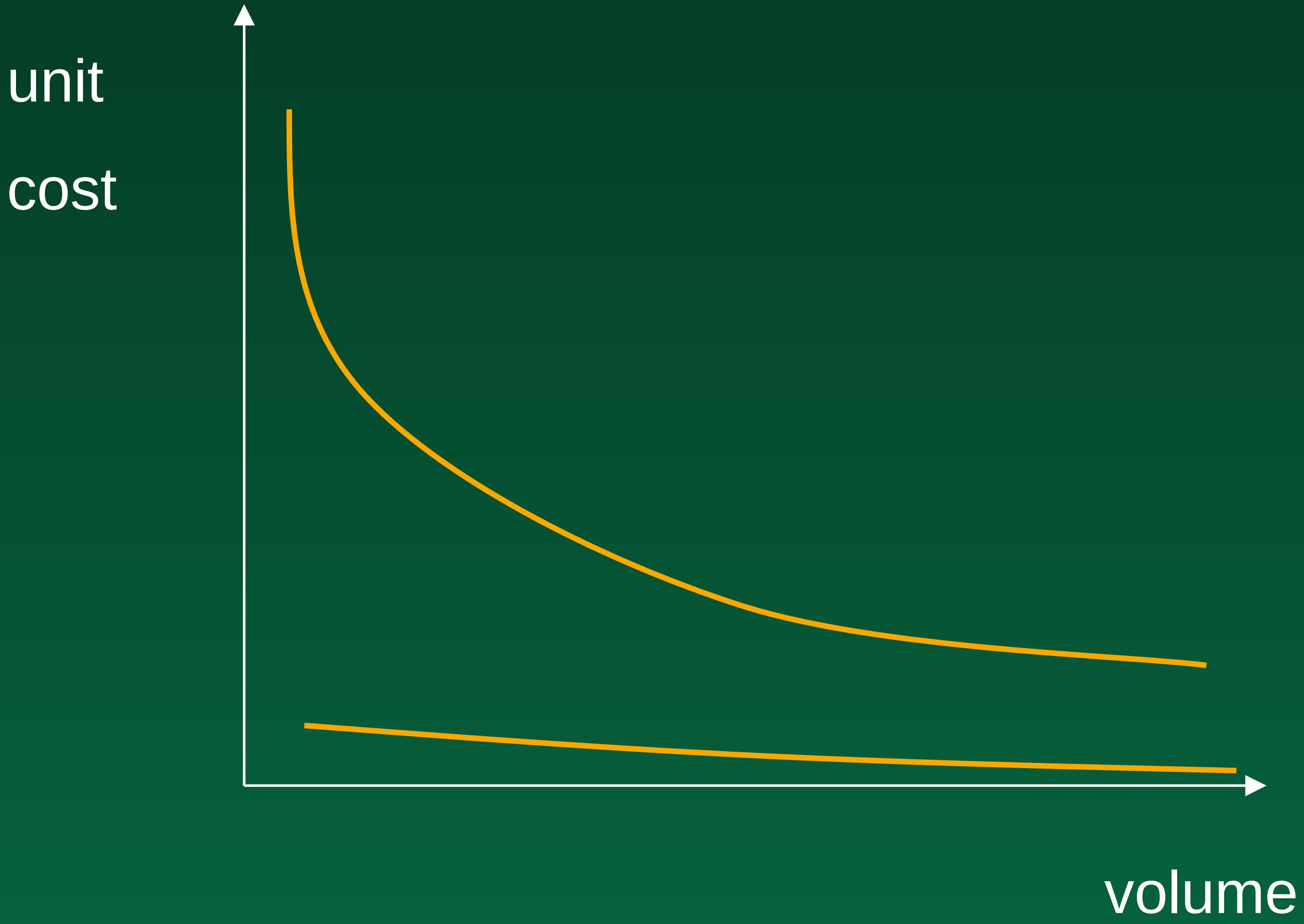
MSC Reports:

2003	*Weaving a National Map: A Review of the U.S. Geological Survey Concept of the National Map *Workshop White Papers*
2002	* Down to Earth: Geographic Information for Sustainable Development in Africa *Summary of document (pdf)*
2001	* National Spatial Data Infrastructure Partnership Programs: Rethinking the Focus
1999	* Distributed Geolibraries: Spatial Information Resources *Workshop White Papers*
1997	* The Future of Spatial Data and Society *Workshop White Papers*
1996	* Technical Issues in NOAA's Nautical Chart Program
1995	* A Data Foundation for the National Spatial Data Infrastructure
1994	* Promoting the National Spatial Data Infrastructure Through Partnership
1994	* Charting a Course into the Digital Era: Guidance for NOAA's Nautical Charting Mission
1993	* Toward a Coordinated Spatial Data Infrastructure for the Nation
1991	* Research and Development in the National Mapping Division, USGS: Trends and Prospects
1990	* Spatial Data Needs: The Future of the National Mapping Program

Senator
Pete Domenici
Senator
Susan Collins

The changing economics of spatial data production

- Declining fixed costs of entry
- Declining willingness of federal government to pay the bills
- Increasing activity in state and local governments
  - and by individuals
- Civilian versus military
- Remote sensing versus ground survey


unit
cost
volume

The concept of patchwork

- Variable scale
  - more detail in areas of high demand
- Distributed production and custodianship
  - distributed to avoid versioning problems
  - integrated by IT to provide a uniform view

The paradox of NSDI

- Why would two jurisdictions want to share geographic information?
  - if they don't overlap
  - if they don't even share an edge
- NSDI is about sharing within the hierarchy
  - jurisdictions that overlap
  - it's about pushing the cost of GI production down the hierarchy
  - local agencies have been empowered by technology to produce GI at higher quality
    - farmers can make better soil maps
- Sharing is good, duplication is bad
  - but duplication is what competition is about

The concept of foundation or framework

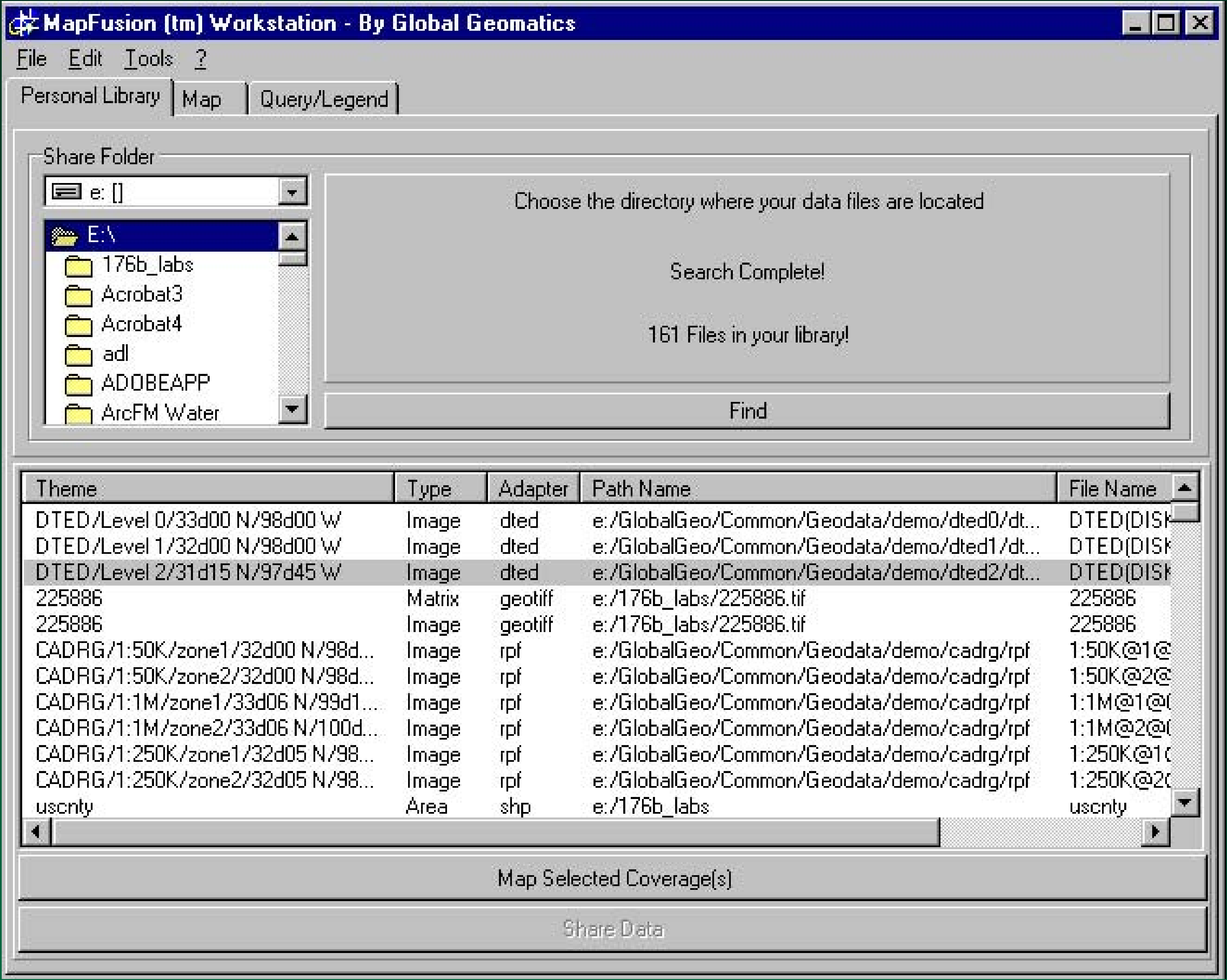
- Certain layers provide infrastructure
  - a basis for other layers
  - geolocation, context
  - the test: would I use it to locate other activities?
- The seven layers
  - topography
  - boundaries
  - transportation
  - hydrography
  - cadaster
  - orthoimagery
  - geodetic control

The placename layer

- The orphan of NSDI
- The institutional context
  - Board on Geographic Names
  - derivative of map production
    - but supports numerous independent applications
- The gazetteer
  - official or vernacular
  - points or footprints
  - space-time
  - language

The components of NSDI

- Standards
  - SDTS, FIPS 173
  - CSDGM
  - OGC specifications
    - the Grid
    - distributed data
    - distributed services
- Clearinghouse
  - 350+ sites integrated through common catalog, search mechanism
    - based on Z39.50
    - the Geospatial One-Stop
- www.fgdc.gov

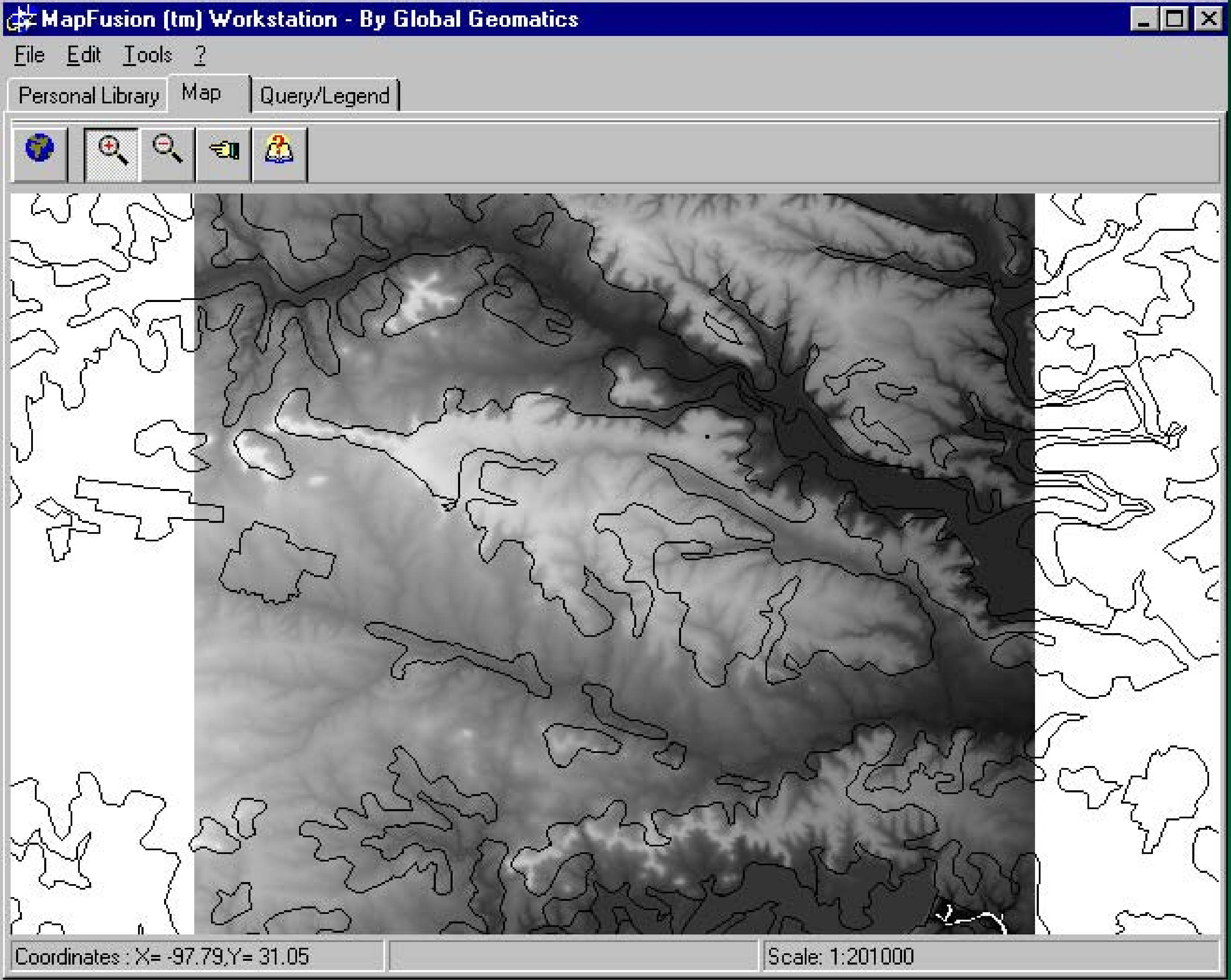


Choose the directory where your data files are located

Search Complete!

161 Files in your library!

Theme	Type	Adapter	Path Name	File Name
DTED/Level 0/33d00 N/98d00 W	Image	dted	e:/GlobalGeo/Common/Geodata/demo/dted0/dt...	DTED(DISK
DTED/Level 1/32d00 N/98d00 W	Image	dted	e:/GlobalGeo/Common/Geodata/demo/dted1/dt...	DTED(DISK
DTED/Level 2/31d15 N/97d45 W	Image	dted	e:/GlobalGeo/Common/Geodata/demo/dted2/dt...	DTED(DISK
225886	Matrix	geotiff	e:/176b_labs/225886.tif	225886
225886	Image	geotiff	e:/176b_labs/225886.tif	225886
CADRG/1:50K/zone1/32d00 N/98d...	Image	rpf	e:/GlobalGeo/Common/Geodata/demo/cadrg/rpf	1:50K@1@
CADRG/1:50K/zone2/32d00 N/98d...	Image	rpf	e:/GlobalGeo/Common/Geodata/demo/cadrg/rpf	1:50K@2@
CADRG/1:1M/zone1/33d06 N/99d1...	Image	rpf	e:/GlobalGeo/Common/Geodata/demo/cadrg/rpf	1:1M@1@(
CADRG/1:1M/zone2/33d06 N/100d...	Image	rpf	e:/GlobalGeo/Common/Geodata/demo/cadrg/rpf	1:1M@2@(
CADRG/1:250K/zone1/32d05 N/98...	Image	rpf	e:/GlobalGeo/Common/Geodata/demo/cadrg/rpf	1:250K@1(
CADRG/1:250K/zone2/32d05 N/98...	Image	rpf	e:/GlobalGeo/Common/Geodata/demo/cadrg/rpf	1:250K@2(
uscnty	Area	shp	e:/176b_labs	uscnty



Coordinates : X= -97.79,Y= 31.05

Scale: 1:201000

Current NRC activities

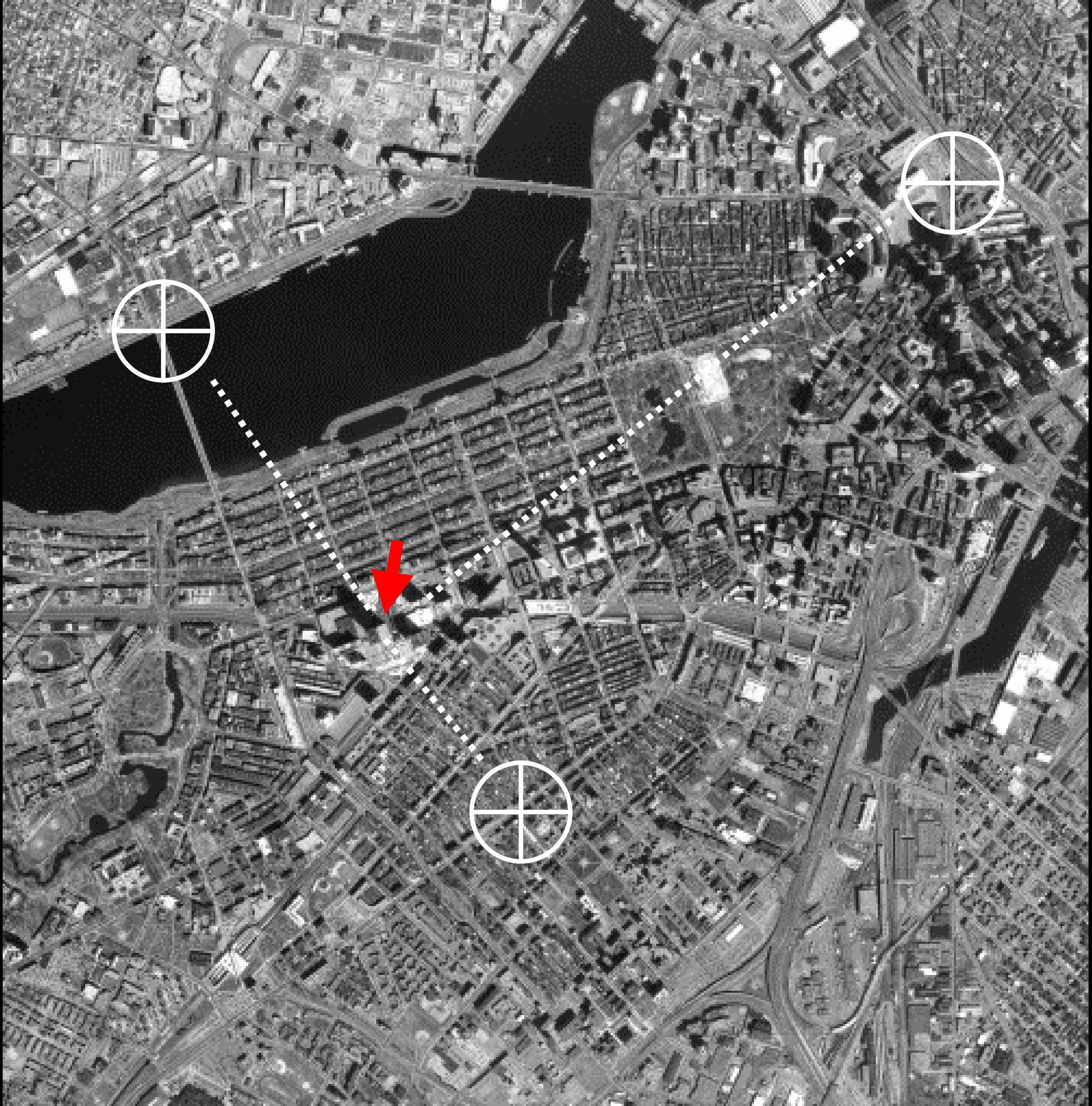
- MSC
  - *Beyond Mapping: The Challenge of New Technologies in the Geographic Information Sciences*
  - a discipline study
    - transformation of traditional mapping sciences
    - strengths, weaknesses of GIScience
    - state of funding, research infrastructure
    - due late 2003
- COG: *Support for Thinking Spatially: GIScience in the K-12 Curriculum*
  - a GIS for K-12
  - due late 2003

Some research issues

- The patchwork
  - edgematching
  - multiple representations
    - scale
    - semantics
  - international integration
- Other spaces
  - geographic, geospatial, spatial
    - bioinformatics
- The value of GI

Measurement of position

- Position measured
  - $\boldsymbol{x} = f(\boldsymbol{m})$
- Position interpolated
  - between measured locations
  - surveyed straight lines
  - registered images
- The inverse function
  - $\boldsymbol{m} = f^{-1}(\boldsymbol{x})$



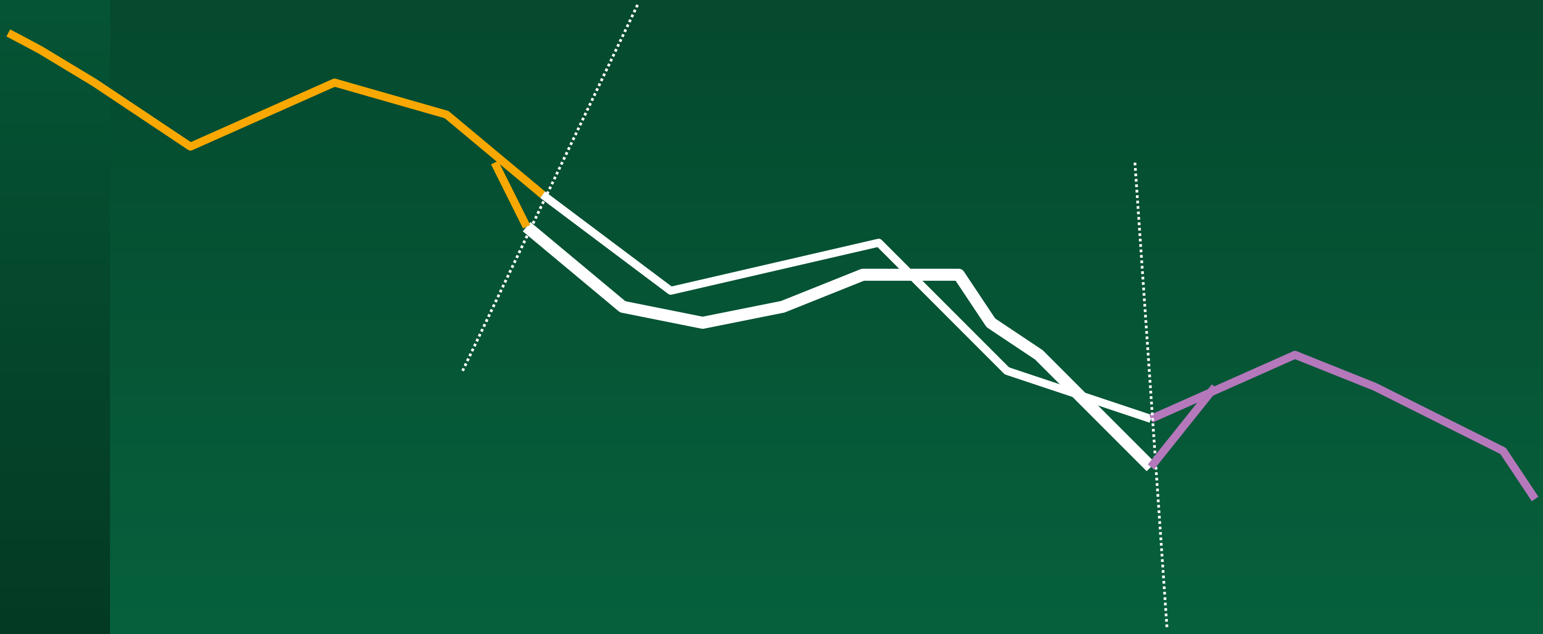
Theory of measurement error

- Measured value = true value + distortion
  - $x' = x + \delta x$
  - some derived value $y = x^2$
  - $y + \delta y = (x + \delta x)^2$
  - expanding and ignoring terms in $(\delta x)^2$
  - $\delta y = 2\, x\, \delta x$
  - more generally if $y = f\,(x)$; $\sigma_y = df/dx\ \sigma_x$
  - generalizes to several variables, variance-covariance matrices

The inverse f^{-1}

- An error is discovered in ***x***
  - error at $\boldsymbol{x}_1$ is correlated with error at $\boldsymbol{x}_2$
  - both errors are attributed to some erroneous measurement ***m***
  - to determine the effects of correcting $\boldsymbol{x}_1$ on the value of $\boldsymbol{x}_2$ it is necessary to know *f* and its inverse f^{-1}

Updating a street database through transactions



Definitions

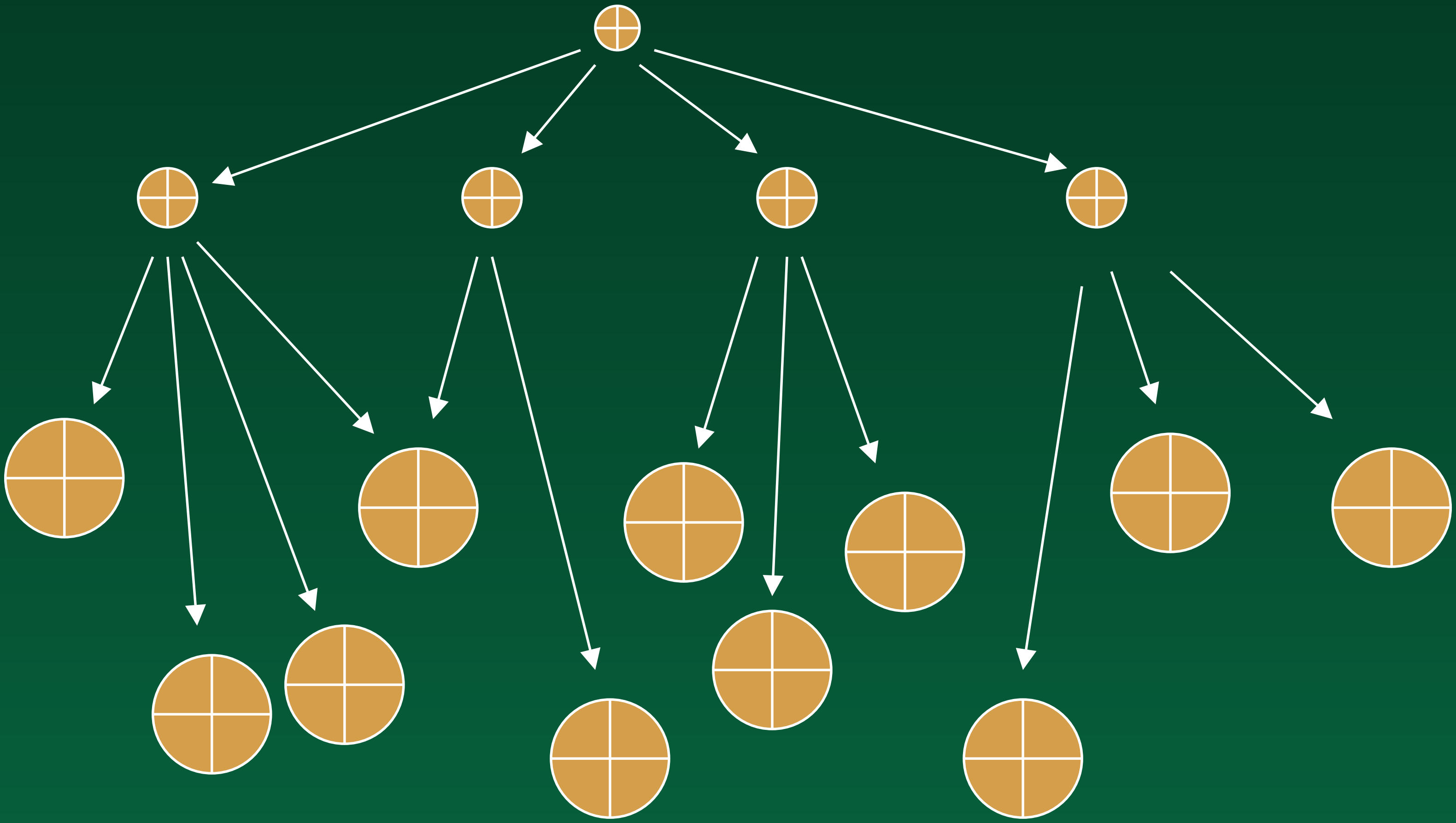
- *Coordinate-based* GIS
  - locations represented by $\boldsymbol{x}$
  - f, f^{-1} and $\boldsymbol{m}$ are lost during database creation
- *Measurement-based* GIS
  - f and $\boldsymbol{m}$ available
  - $\boldsymbol{x}$ may be determined on the fly
  - f^{-1} may be available

Partial correction

- The ability to propagate the effects of correcting one location to others
  - preserving the shapes of buildings and other objects
  - avoiding sharp displacements in roads and other linear features
- Partial correction is impossible in coordinate-based GIS
  - major expense for large databases

The geodetic model

- Equator, Poles, Greenwich
- Sparse, high-accuracy points
  - First-order network
- Dense, lower-accuracy points
  - Second-order network
- Interpolated positions of even lower accuracy
- Locations at each level inherit the errors of their parents



Formalizing measurement-based GIS

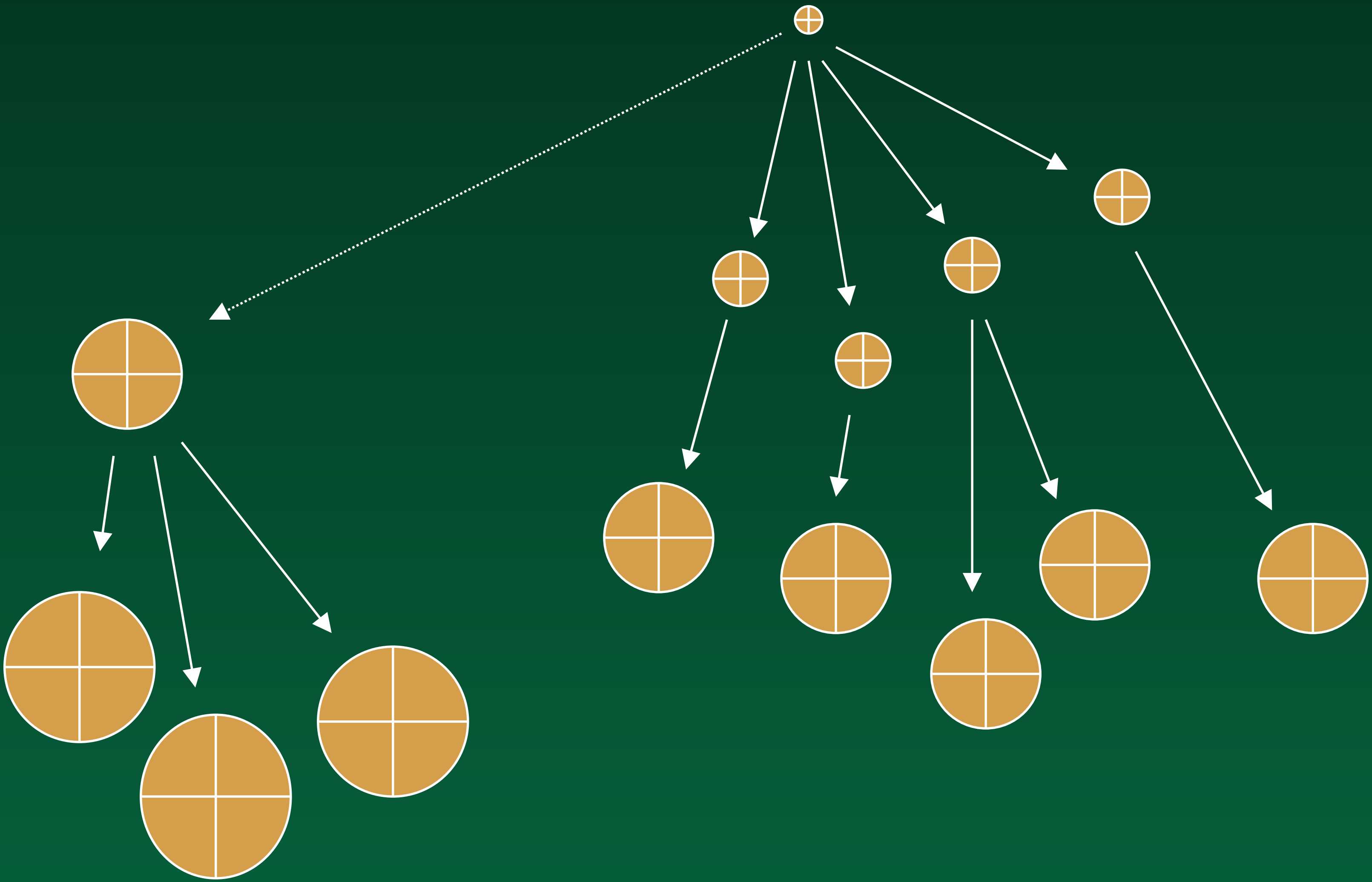
- Structured as a hierarchy
  - levels indexed by *i*
  - locations at level *i* denoted by $\boldsymbol{x}^{(i)}$
  - locations at level (*i*+1) derived through equations of the form $\boldsymbol{x}^{(i+1)} = f(\boldsymbol{m},\boldsymbol{x}^{(i)})$
  - locations at level 0 *anchor* the tree
  - locations established independently (GPS but not DGPS) are at level 0

An example

- A utility database
- Pipe's location is measured at 3 ft from a property boundary
  - ***m*** = {3.0,L}
  - property at level 3, pipe at level 4
- Property location is later revised or resurveyed
  - new ***m*** = {2.9,L}
  - effects are propagated to dependent object

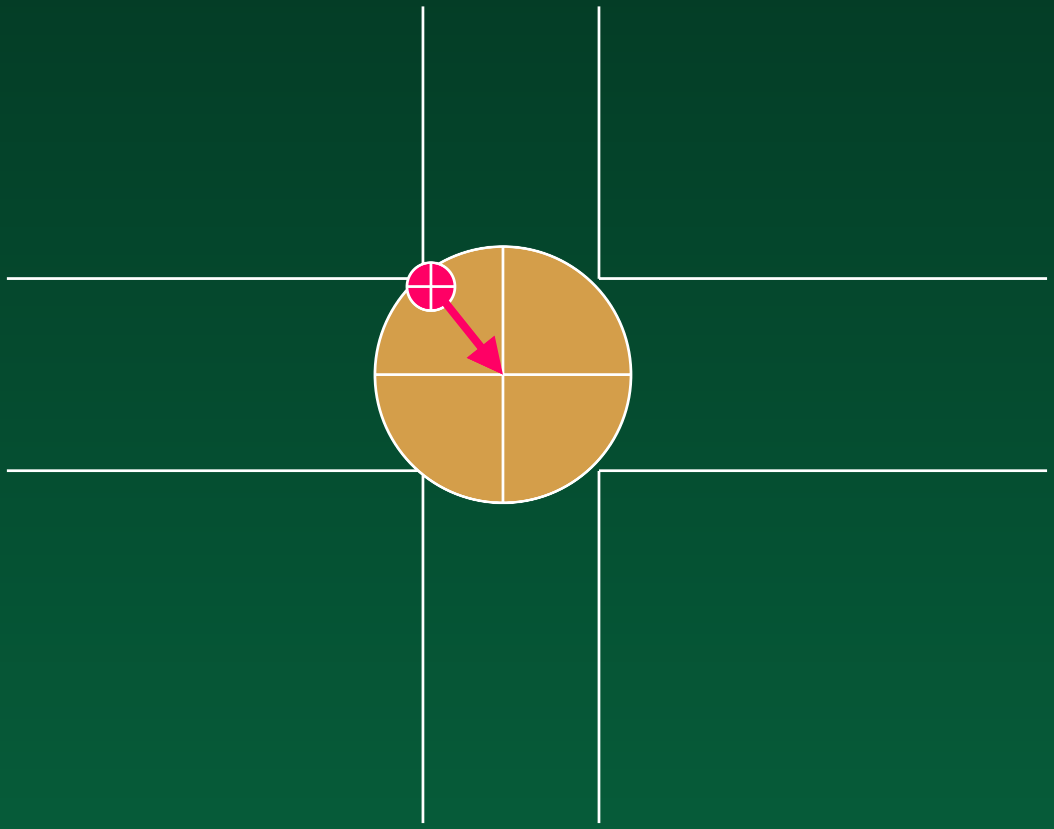
Beyond the geodetic model

- National database of major highways
  - 100m uncertainty in position
    - sufficient for agency
    - relative accuracies likely higher, e.g. highways are comparatively straight, no sudden 100m offsets
- Local agency database
  - 1m accuracy required
  - two trees with different anchors



Merging trees

- Link with a pseudo-measurement
  - displacement of 0
  - standard error of 100m
  - revisions of the more accurate anchor can now be inherited by the less accurate tree
    - but will normally be inconsequential



Summary arguments

- Almost universal adoption of coordinate-based GIS
  - assumes it is possible to know location exactly
  - design precision greatly exceeds actual accuracy
  - in practice exact location is not knowable
  - attempts at partial correction lead to unacceptable topological and geometrical distortions

Measurement-based GIS

- Retains measurements and derivation functions
  - may obtain absolute locations on the fly
- Supports incremental update and correction
- Supports merger of databases with different inheritance hierarchies
- Legacy GIS designs are not optimal

Implementation

- Design from the ground up
- Accept a model that includes necessary features
  - hierarchical databases
  - object-oriented databases
    - but support for complex functions, variance-covariance matrices?

What is the GI in GIScience?

- Is information "stuff"?
  - if it is, then it must be possible to measure its quantity
  - $Q(A+B) = Q(A)+Q(B)$
  - a market in GI requires that such means be agreed
    - otherwise all transactions would be unique and no market could exist
  - conventional metrics of quantity are arbitrary, media-dependent, structure-dependent
    - e.g. per sq km, per quadrangle, per megabyte

Shannon-Weaver information theory

- Measures the information content of a message
  - by comparison to the number of distinct messages that could exist in a given code
    - e.g. one Roman letter resolves among 26 possibilities
    - but not all possibilities are equally likely in English
    - an E conveys less information than an X
- Is code-dependent, media-dependent, structure-dependent
  - is syntactic rather than semantic

The information content of a number

- There are 100 2-digit numbers
  - any one 2-digit number resolves among 100 possibilities
- Consider the infinite series of digits starting 3.14159...
  - resolves among an infinite number of possibilities
  - but can be sent by sending one letter from the Greek alphabet
  - provided the receiver knows the code
  - the value of information depends on knowledge of codes

Towards a semantic theory of GI

- Measuring the meaning conveyed by a message
  - the increment to the receiver’s knowledge
  - in ways that are independent of media, syntax, structure
- Accommodating the ability of GIS to transform
  - information can easily mutate into other forms
  - how do we know if the content of two data sets is the same?
- Why GI?

Atoms of GI

- GI is composed of atomic pairs of the form <**x**,**z**>
  - compare Berry, Sinton, Plewe
  - where **x** is a location in space-time
    - of 2 to 4 dimensions
    - using agreed methods for referring to times, and locations on the Earth's surface (latitude/longitude, WGS 84, GMT, …)
    - methods that are shared between sender and receiver of GI (and are frequently universal)

The nature of **z**

- A vector of properties
  - using definitions that are already agreed between sender and receiver
  - some such definitions are universal, e.g. Celsius
  - some are not, e.g. vegetation cover type
  - the value of an atom sent to a receiver who does not share the same definitions will be uncertain, and may be nil

Domains of GI

- **x**
  - limited to the Earth’s surface and near-surface
  - to the present, near-past, and near-future
  - a rigid Newtonian frame
  - “mappable”
- **z**
  - physical, social, environmental properties associated with locations

Geographic information systems

- Systems that combine GI with expertise
  - to perform transformations and respond to queries
- A geographic query
  - a query to which GI provides the answer
  - satisfied by access to one or more atoms
    - e.g., “What is the temperature at **x**?”
    - e.g., “Where is the temperature equal to *T*?”

Possession of GI

- A GIS is said to possess an item of GI if it is capable of responding successfully to a query to which the item is the answer
  - item = one or more atoms
  - independent of format, structure, medium
  - may imply transformations
  - a message has no value if the information it contains is already possessed

Quantity of information

- A polygon describing the State of California enables an infinite number of queries of the form “Is **x** in California?”
  - does the system possess an infinite amount of information?
- Suppose location is knowable to an accuracy λ (a linear measure)
  - there are only $4\pi R^2/\lambda^2$ distinct locations on the Earth’s surface
  - only that number of distinct queries can be answered

…and in addition

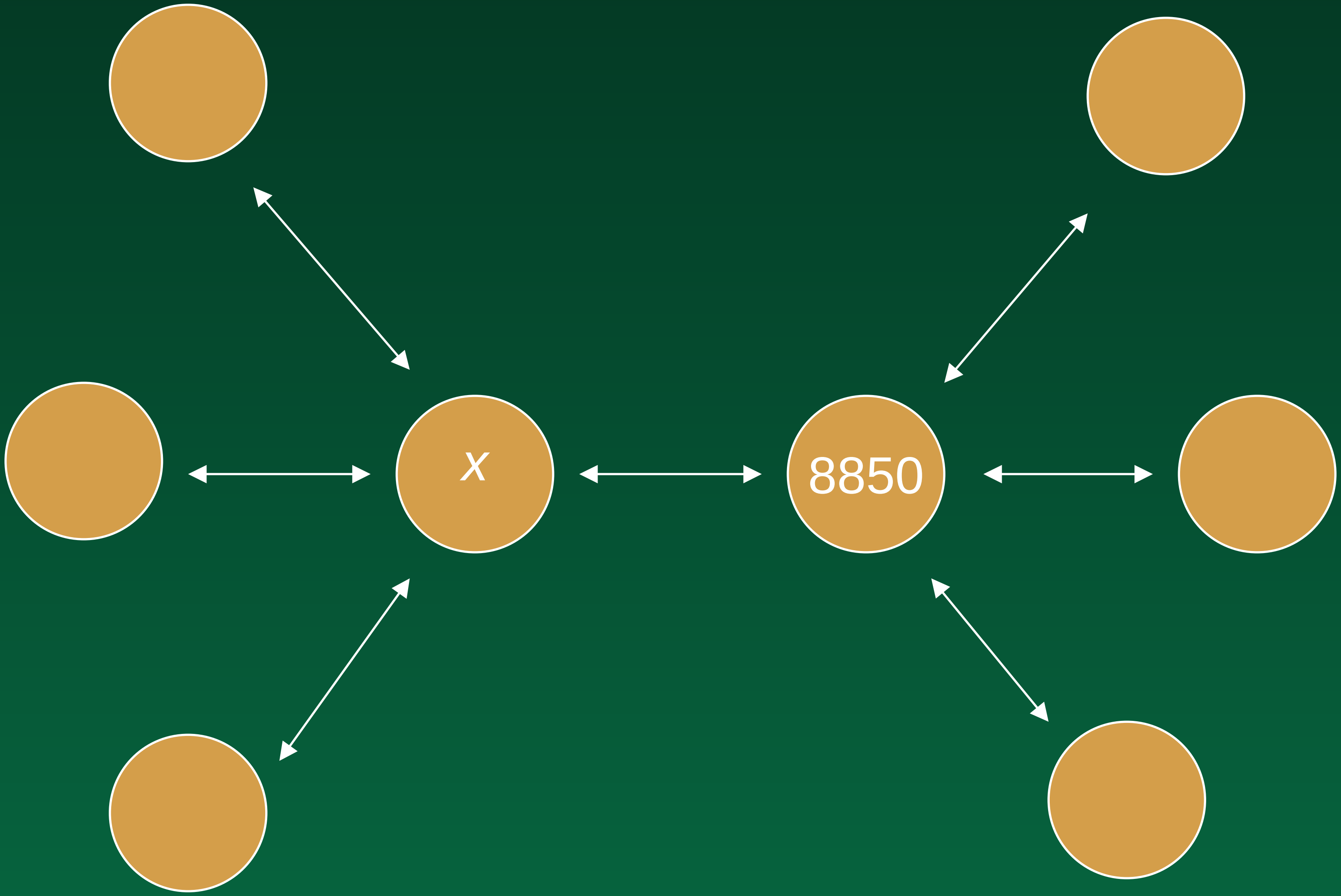
- If $\mathbf{x}_1$ and $\mathbf{x}_2$ are in California, then $\alpha\mathbf{x}_1+(1-\alpha)\mathbf{x}_2$ is also probably in California
  - and certainly so if California is convex
    - the number of convex states in the US?
    - 2 on cylindrical projections, 0 on ellipsoid
- The system actually possesses the coordinates of a polygon, plus a universal rule
  - the volume of information is bounded by the volume of the polygon definition

A semantic theory of GI

- Atomic pairs link understood concepts
  - **x** is universally understood
  - **z** is understood by an information community that includes the receiver
- The value of an atom of GI is related to the level of understanding on the part of the receiver of the concepts that it links
  - linking a concept that is not understood is of no value

<“Mt Everest”,8850m>

- Of no value to a receiver who does not recognize “Mt Everest”, the concept of height, or the metric system
- Given <**x**,”Mt Everest”> the system can deduce <**x**,8850m>
  - other pairs can be deduced from other prior knowledge
- “Understanding”: the number of prior linkages to a concept
  - the higher the understanding, the greater the value of a new linkage


x
8850

Key points

- GI in atomic form
  - almost never exposed except for point data
  - must be compressed in practice
- Pairs linking already-understood concepts
  - value depends on number of linkages
  - and whether tuple is already possessed
- Systems as combinations of information and expertise
  - tuples and rules
- Independent of media, structure, format

NSDI and the scientific community

- G, GS, or S?
  - G, GS redundant
  - S generalizes G
  - Ian's list
    - five G, three S, two GS, one L
- NSDI does not even cover all of G
  - NASA's EOSDIS collects and distributes 3 TB of GI every day
  - NGIA, NSA march to the military and intelligence drummer
  - NASA promotes science standards that do not overlap NSDI
    - hdf, netcdf

Other S communities

- Brain research
  - gazetteers
  - the average brain
  - admire NSDI for its institutional successes
  - some limited interest in AV 3.x
- Astronomy, geophysics, medical imaging
  - NSDI needs to engage with these communities
  - social sciences, ecology, …
    - GI is important but not dominant
    - domain-specific approaches

NSDI and cyberinfrastructure

- An excellent start
- Technology is moving rapidly
  - the scientific community initiated many of the more important breakthroughs
    - the Web and physics
  - Grid computing
  - semantic web
  - sensor networks
  - NSDI must be a player
    - watch what's happening in physics