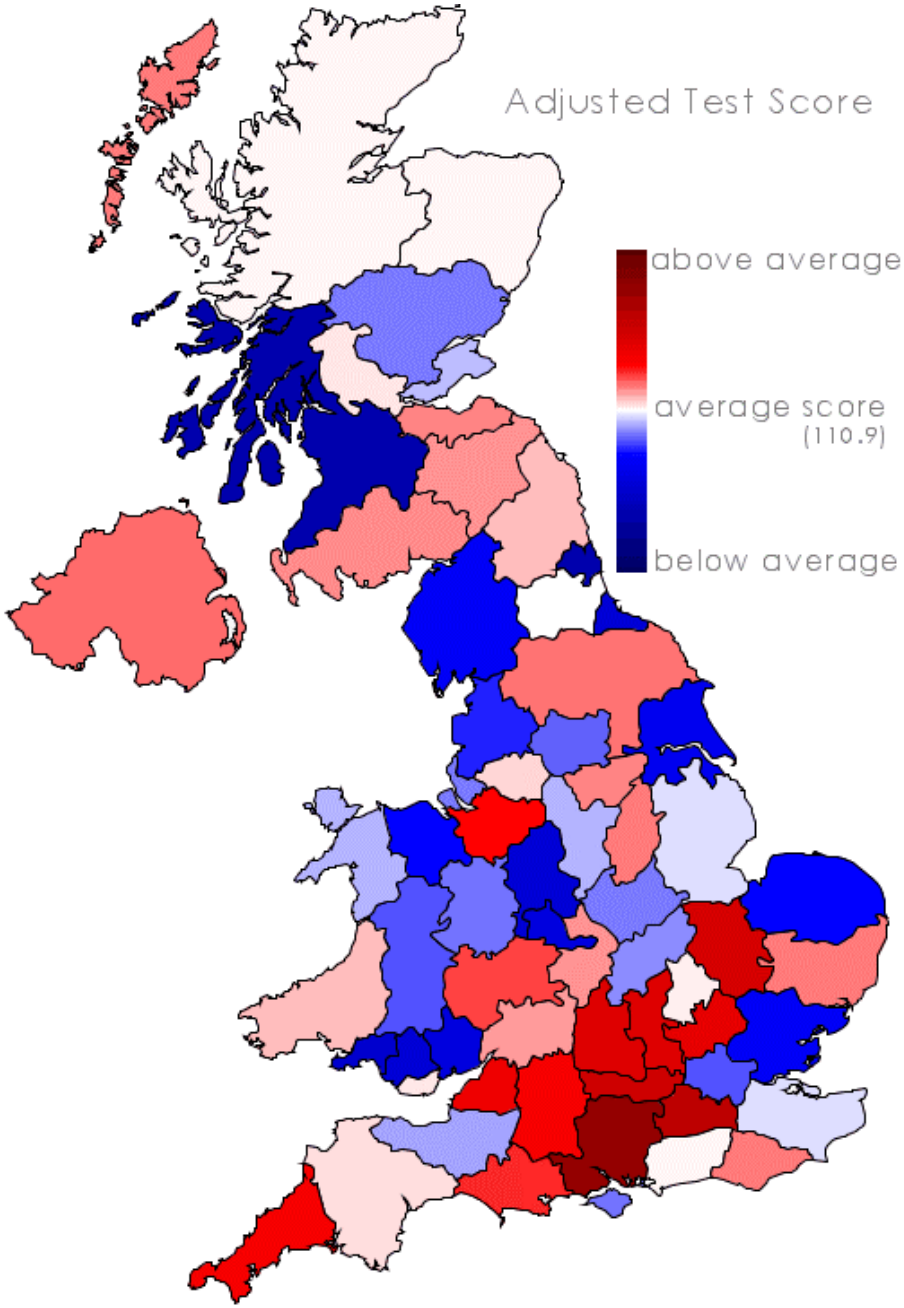
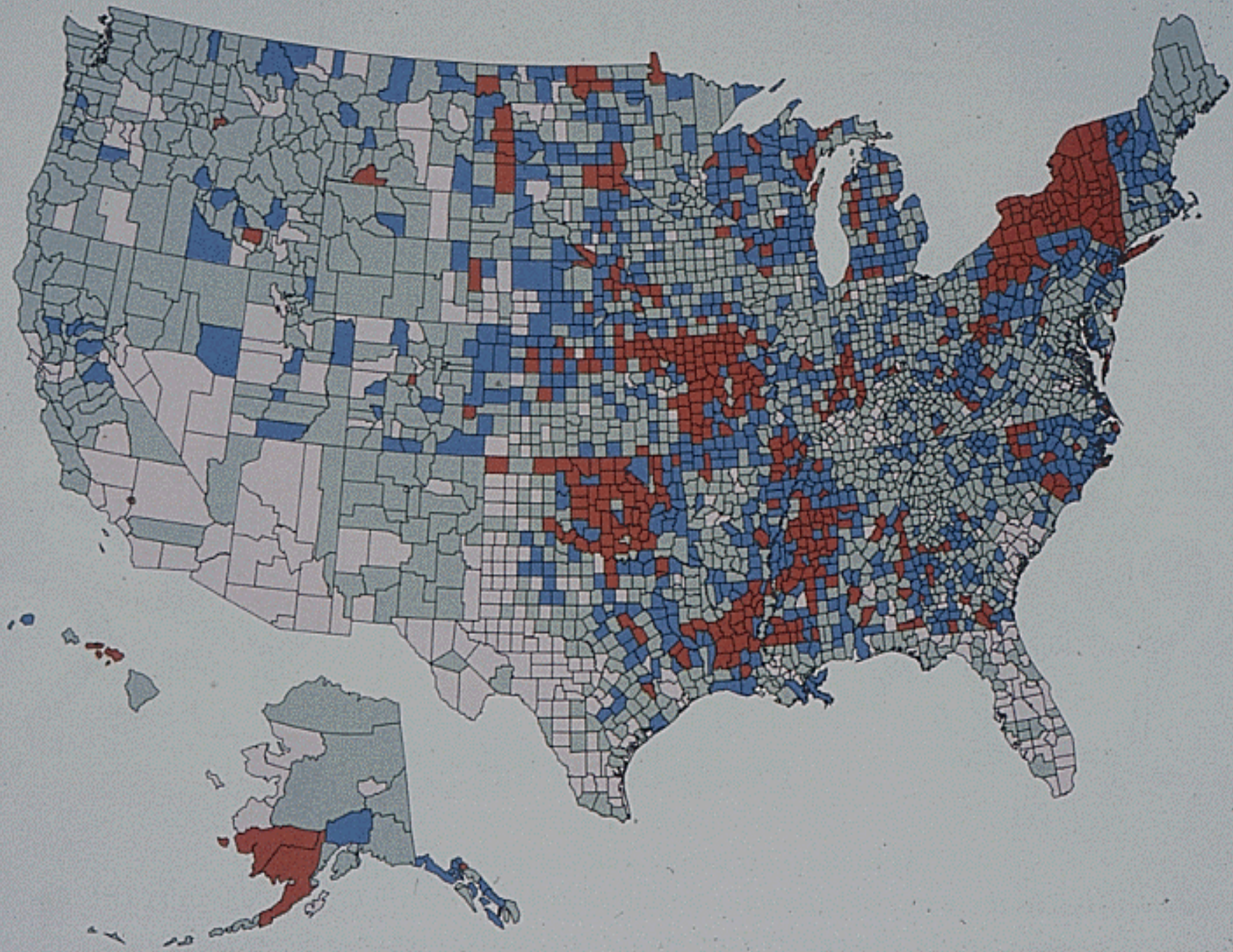
The Nature and Value of Geographic Information

Michael F. Goodchild
University of California
Santa Barbara

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


Adjusted Test Score
above average
average score
(110.9)
below average



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

Continuity of **x**

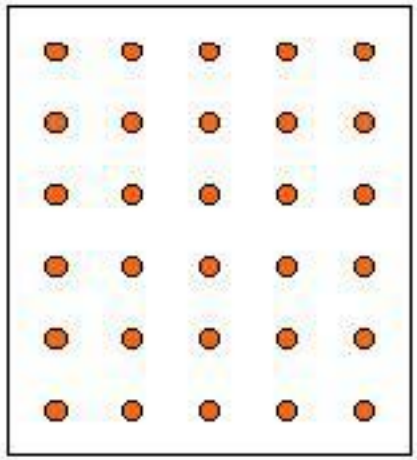
- Description is impossible because **x** is continuous and **z** is infinitely dimensioned
  - we are saved by Tobler's Law
    - all things are related but nearby things are more related than distant things
    - $\langle \mathbf{x}+\delta\mathbf{x},\mathbf{z}\rangle = \langle \mathbf{x},\mathbf{z}\rangle$ for $\delta\mathbf{x}<\lambda$
    - an infinite number of pairs is not required
    - hell is a place with no spatial dependence
  - a potentially infinite number of properties exist, but in practice they are strongly correlated and only a finite number are needed for useful description

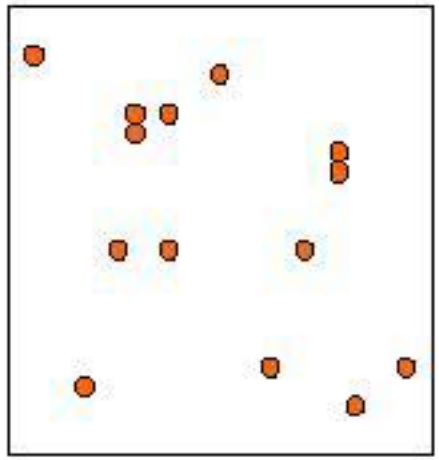
Consequences of Tobler's Law

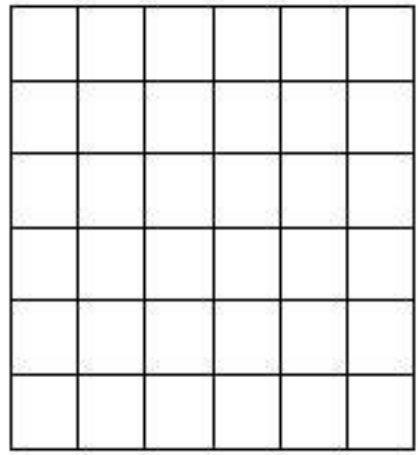
- The explicit atomic form is never needed
  - atoms are inferred from larger structures using appropriate universal rules and transformations
    - e.g. the boundary of California leads to an infinite number of pairs <**x**,*z*> where *z* is binary
    - databases are built using larger structures as well as atoms

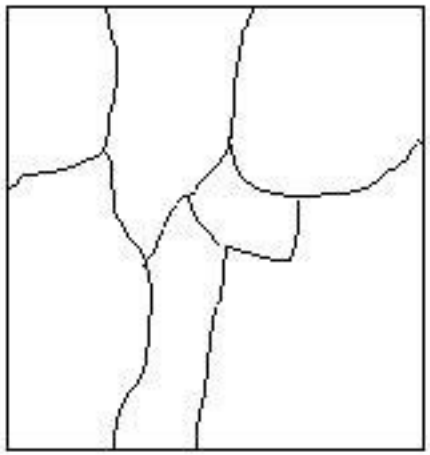
Six field representations

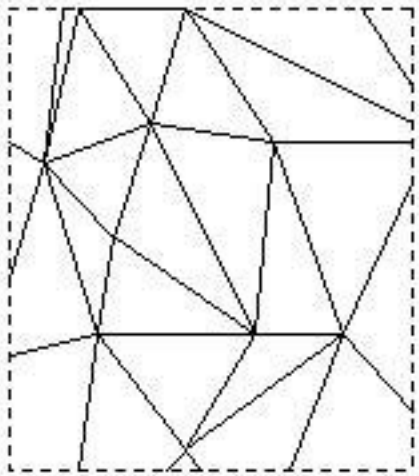
- Representing a single property *z*
- Irregularly spaced sample points
  - a finite number of pairs <**x**,*z*>
  - plus an interpolator, e.g. inverse-distance weighting, Kriging, splines, proximal/Thiessen
- Regularly spaced sample points
  - a single tuple <**G**,O,z_1,z_2,…,z_n> where **G** defines georeferencing, O defines ordering

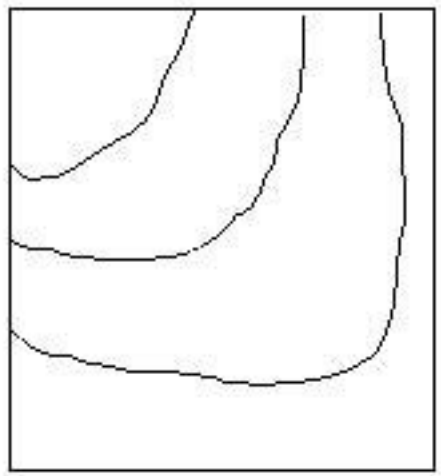












Irregular polygons

- Tuples defining each polygon and its field value
  - $<x_1,y_1,x_2,y_2,\ldots,x_m,y_m,z>$
- Polygons do not overlap, and collectively exhaust the space
  - every point x,y lies in exactly one polygon
- Loss of detail justified by reference either to some λ (pixel size, MMU), or to knowledge of the properties of the phenomenon (land ownership parcels)

Discrete objects

- Points are atomic
- Lines and areas as tuples
  $\langle x_1, y_1, x_2, y_2, \ldots, x_m, y_m, \mathbf{z} \rangle$

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

Derivative queries and spatial analysis

- “What is the distance from A to B?”
- Requires $<\mathbf{x}_1, A>$ and $<\mathbf{x}_2, B>$
- Requires a rule for determining distance (a metric)
- Within the capabilities of a GIS, but beyond those of a human?

Digital Earth

- “I believe we need a 'Digital Earth'. A multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of georeferenced data.” U.S. Vice President Gore, 1/98
- A single (distributed?) repository for all GI
  - a complete description of the planet
- A system that contained DE would be able to respond to all queries about Earth

Bit or it?

- A DE and someone with access to the Earth would be equally successful at answering queries
  - there is no query that could resolve whether Earth is real or digital
  - two Chinese postmen
  - a DE would contain sufficient information to reconstruct Earth
  - sending a DE is equivalent to transporting the planet
  - Siegfried, *The Bit and the Pendulum*

Naïve geography

- Geocentric perspective: Newtonian frame, scientific measurement
- Human-centric perspective: individual differences, perception, uncertainty
  - proliferation of **z**
    - e.g., multiple definitions of wetland
  - “the body of knowledge that people have about the surrounding geographic world” (Egenhofer and Mark 1995)

Consistency with geometric principles

- All points contained within the boundary of California are in California
  - what if someone believes otherwise?
- “Santa Barbara is north of Los Angeles”
  - between 337.5 and 22.5 degrees
  - potential violation of geometric principles
- The rules, transformations on which GIS is based break down
  - information is not necessarily reducible to atomic form
  - queries are not necessarily answerable

Scale and spatial resolution

- In practice the ability to locate precisely on the Earth's surface is limited
  - there are not an infinite number of possible locations
    - e.g., ROSE
- Tobler's Law enables approximately complete description with a finite number of atoms

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

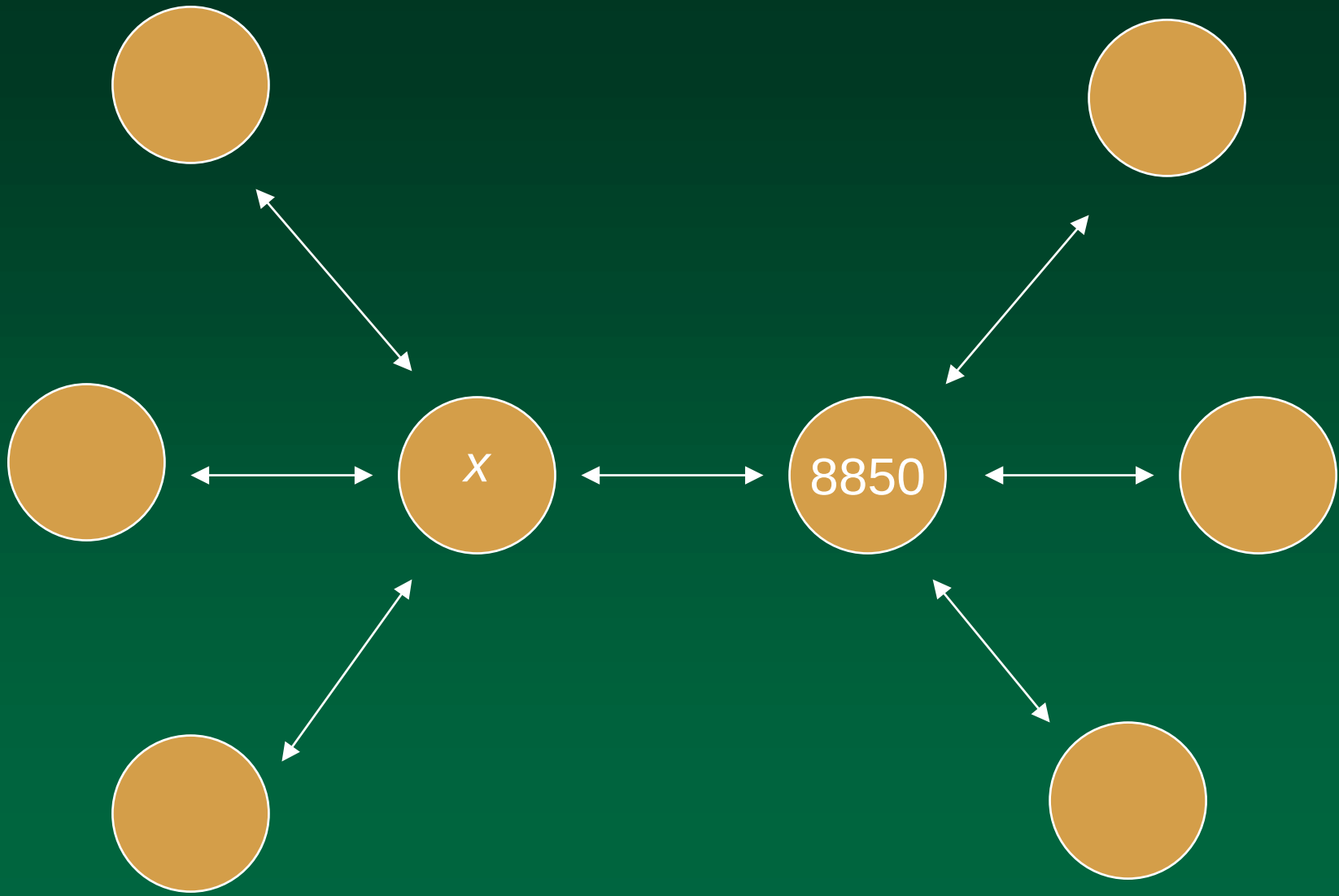
- If $\mathbf{x}_1$ and $\mathbf{x}_2$ are in California, then $\alpha\mathbf{x}_1+(1-\alpha)\mathbf{x}_2$ is also probably in California
  - and certainly so if California is convex
- The system actually possesses the coordinates of a polygon, plus a universal rule
  - the volume of information is bounded 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

Unresolved issues

- Partial resolution of uncertainty
  - incomplete answers to queries
  - what is the relative value of <“Mt Everest”,8848m±2>?
  - what is the value of increased spatial resolution?
- Naïve and inconsistent belief
  - is it possible to build such a GIS?

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