All Geometries are Wrong but Some are Useful*: the Art of Simplification in GIS

*With appreciation to George Box

Josh Lieberman

Workshop on Data Science in Low-dimensional Spaces

May 14, 2019

https://icerm.brown.edu/video_archive/?play=1905

Center for Geographic Analysis

Harvard University

OGC®

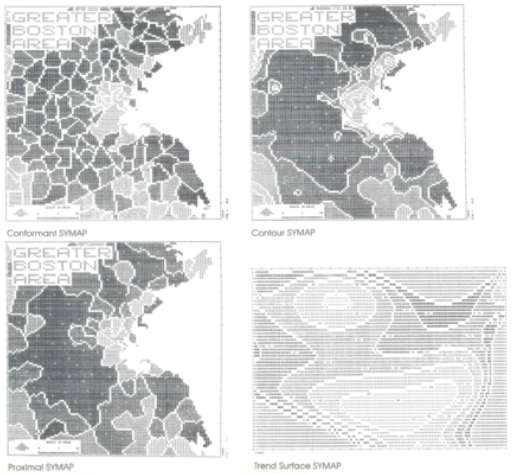
Making location count.

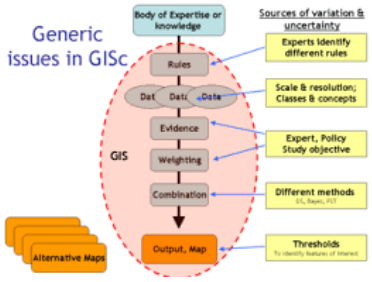
Outline

- Geographic information systems and GIScience
- Discernment, discourse, and features of interest
- Spatial data science, geometries, and processes
- Cases of simple insight
  - The map is not the terrain, but flat stacks coincide
  - Gridding the field: sampling, scale, and resolution
  - Navigating from topography to topology without Euclid
  - Flowing likelihood of water, fish, and heavy metals
  - River miles, coastlines, and paths through space-time
  - Making mereology with feature graphs and surface networks
- Summary and discussion

GIS and GIScience

- Long history of cartography symbolizing spatial data on maps
- Howard Fisher at Harvard Laboratory for Computer Graphics and Spatial Analysis among others hacked type-ball line printers to generate map layers of different spatial data representations (1965-91). Led to SYMAP, Odyssey, Arcinfo and other GIS products.
- GIScience coined by Michael Goodchild "…redefines geographic concepts and their use in the context of geographic information…"*





Center for Geographic Analysis
Harvard University

* David Mark

OGC®
Making location count.

Discernment, discourse, interesting features

- OGC / ISO abstract general feature model (ISO 19109)
  - Interaction of perception, discernment, agreement (discourse) -> phenomena
  - Intention -> feature distinction
  - Phenomenon <-> feature association as properties
  - 1..n coordinate geometric representations as properties

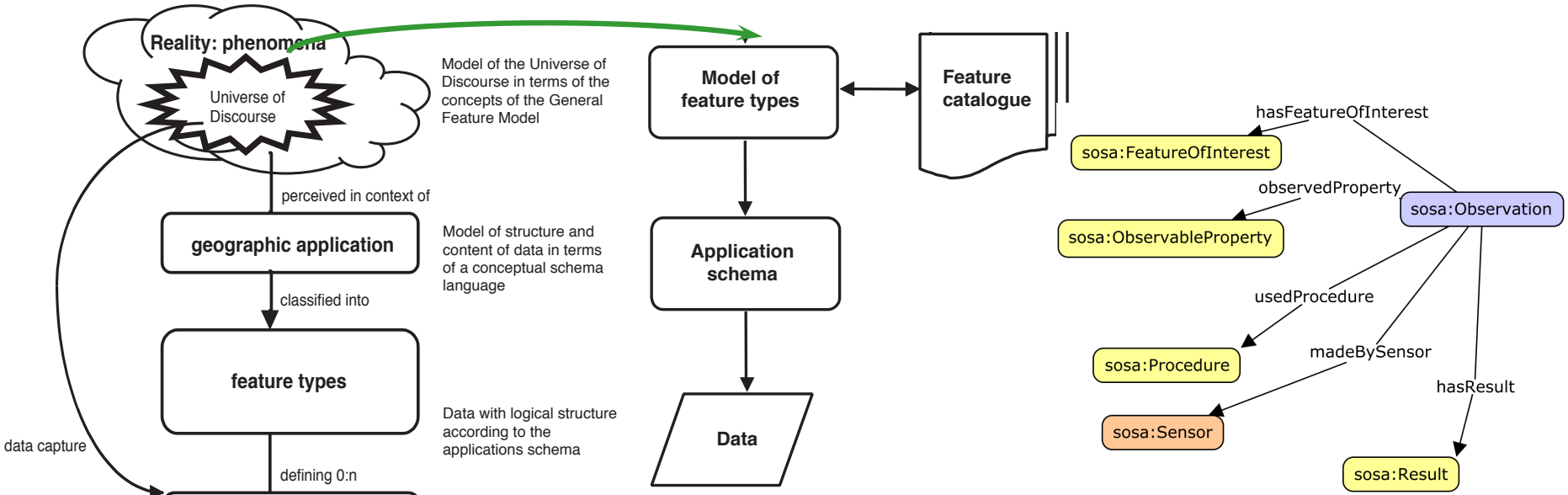


Figure 4 — From reality to geographic data

Center for Geographic Harvard University

OGC® Making location count.

Spatial data science, geometries, processes

- Data science (statistics, machine learning, numerical modeling) techniques that recognize unique relationships of spatial data to reality.
- Double (and more) model challenge:
  - Phenomena -> Geometric model -> Process model
- Intermediate processes obscure primary ones
- Good geometric simplification improves vision.

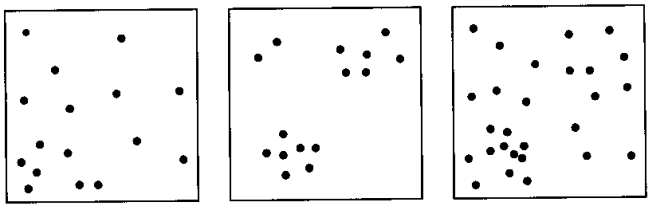
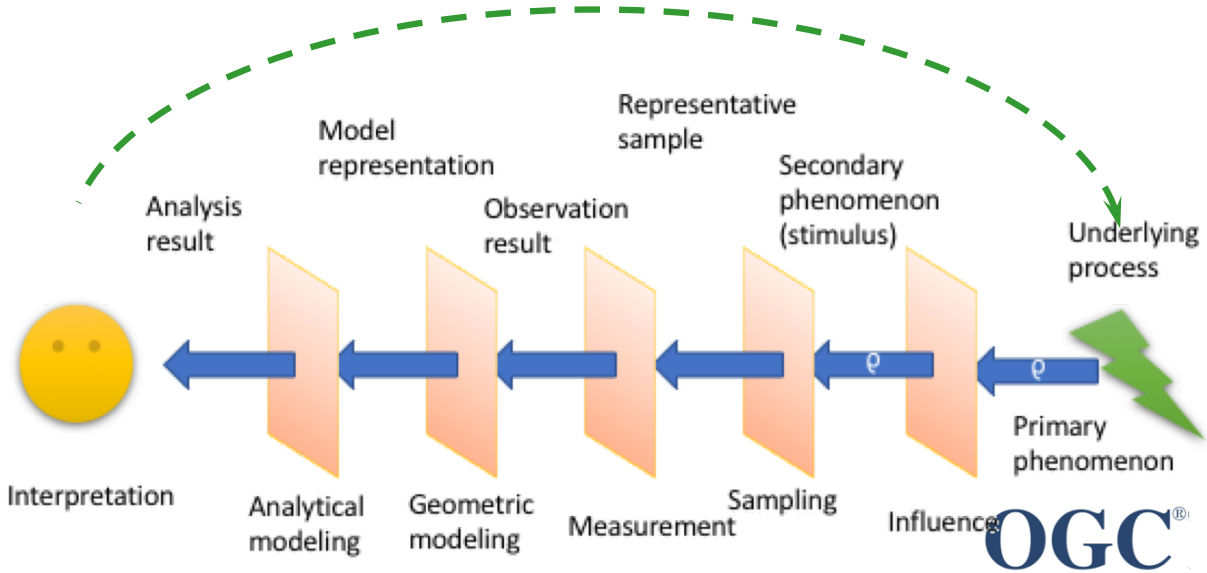
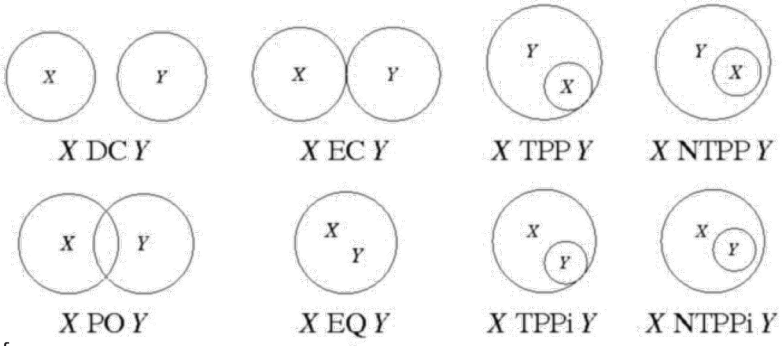


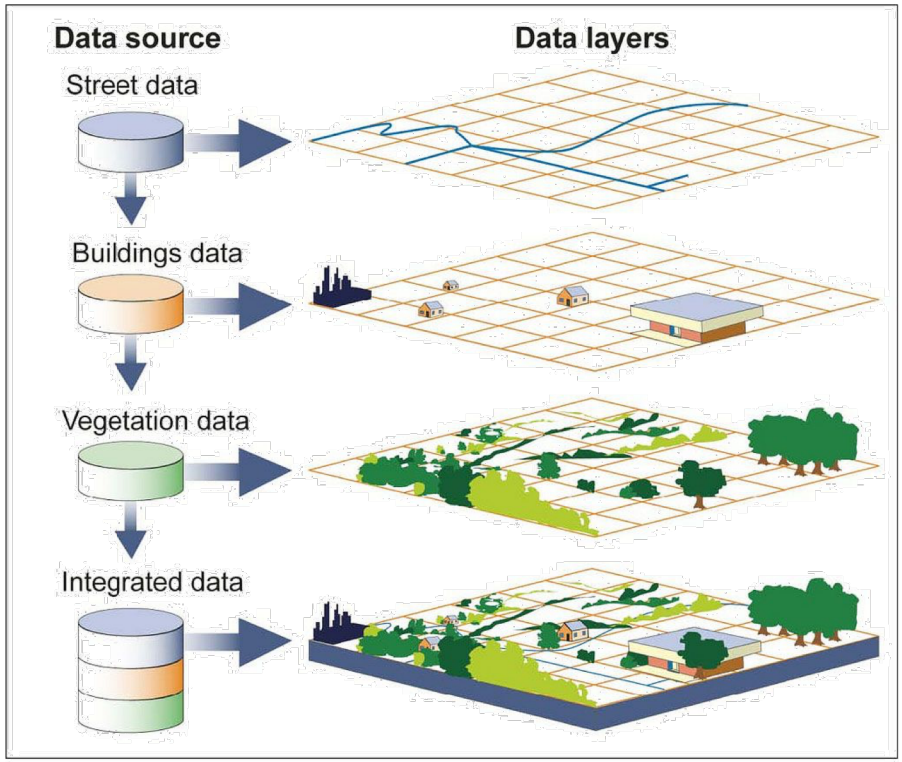
Figure 5.1 The difficulty of distinguishing first- and second-order effects.



Case study 1

- The map is not the terrain, but flat stacks coincide
  - First level of simplification emphasizes horizontal processes within each layer, e.g. Tobler's law
  - Perceptually, technologically effective
  - Geoid, topographic projection issues easily forgotten
  - Conducive to 2D geometries (point, line, polygon) and topological relations
  - Can incorporate 2.5D effects, e.g. radiance, but obscures 3D processes



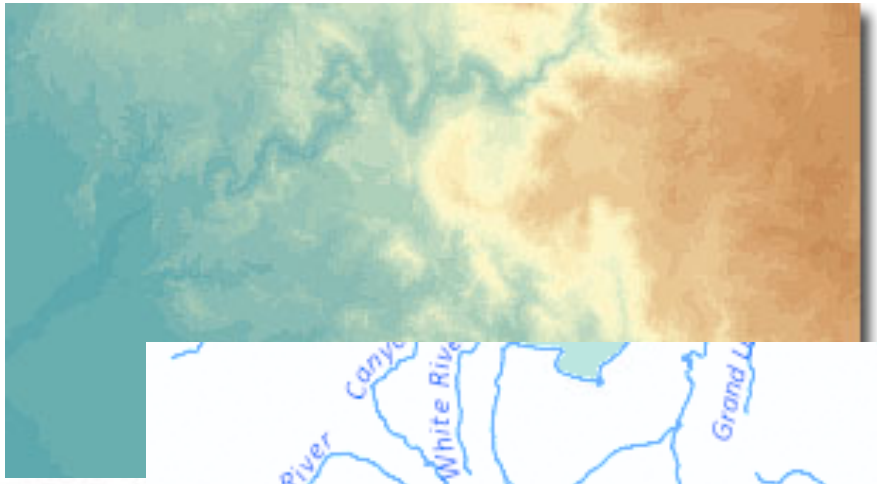


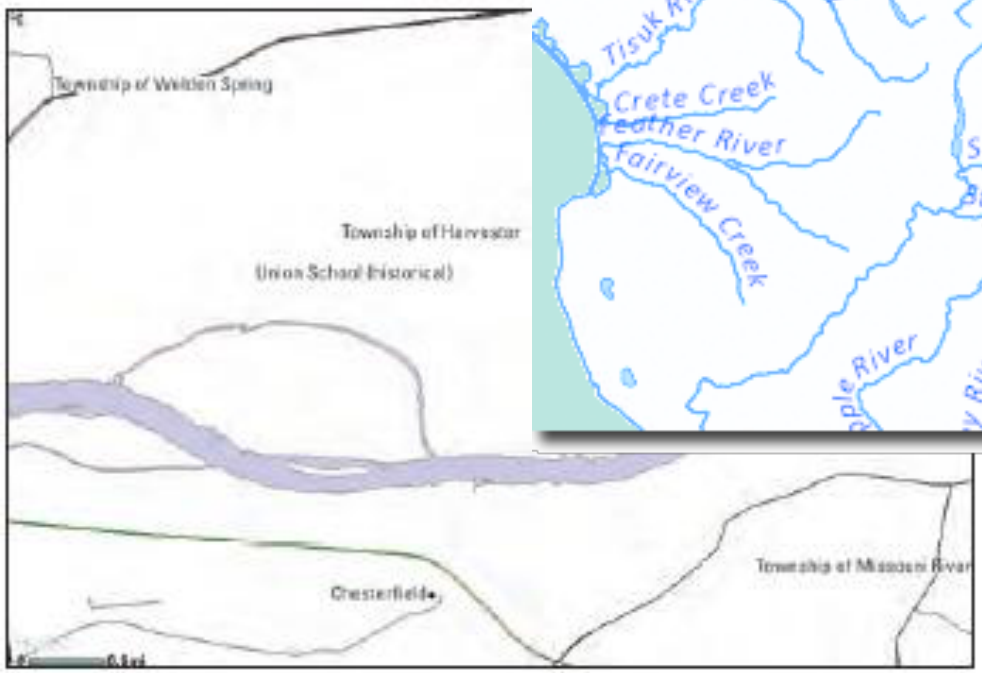
Source: GAO.

Domain: The National Map

Nationally extensive data layers pertaining to USGS Topos

- Elevation (NED, DLG),
- Orthoimagery,
- Hydrography (NHD, WBD),
- Geographic Names (GNIS),
- Boundaries,
- Transportation,
- Structures,
- Land Cover,



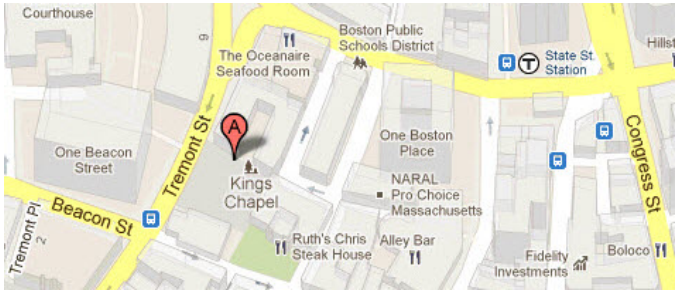


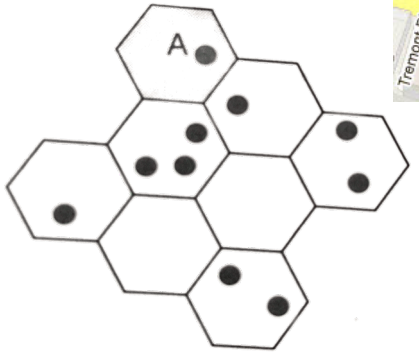
Center for Geographic Analysis
Harvard University

OGC®
Making location count.

Case study 2

- Gridding the field: sampling, scale, and resolution
  - Discrete vector geometries (points, lines, polygons) reduce dimensional complexity of features and interactions (e.g. point patterns, sampling bias, autocorrelation)
  - Hypotheses at one spatial scale may not be valid at another, e.g. polygon aggregations (modifiable areal unit problem), centroid points
  - Continuous phenomena? Polygon / line coverages and point grids (e.g. images) have inherent resolution. Coarse may miss significant spatial variation and patterns, fine can be voluminous and miss larger patterns.
  - Grids reduce sampling error but can add process constraints, e.g. X-Y bias.

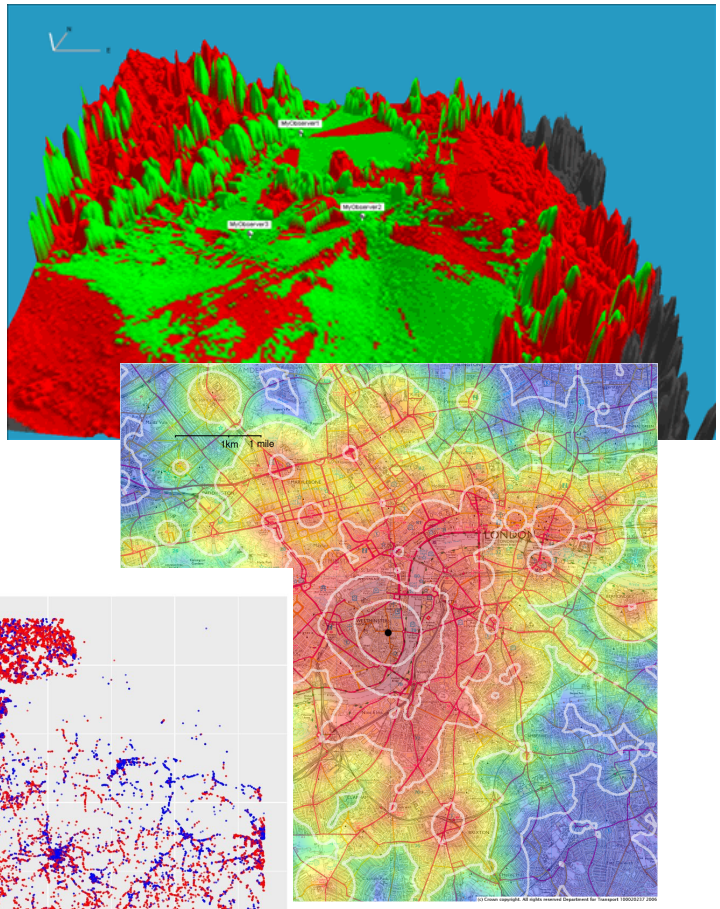




Case study 3

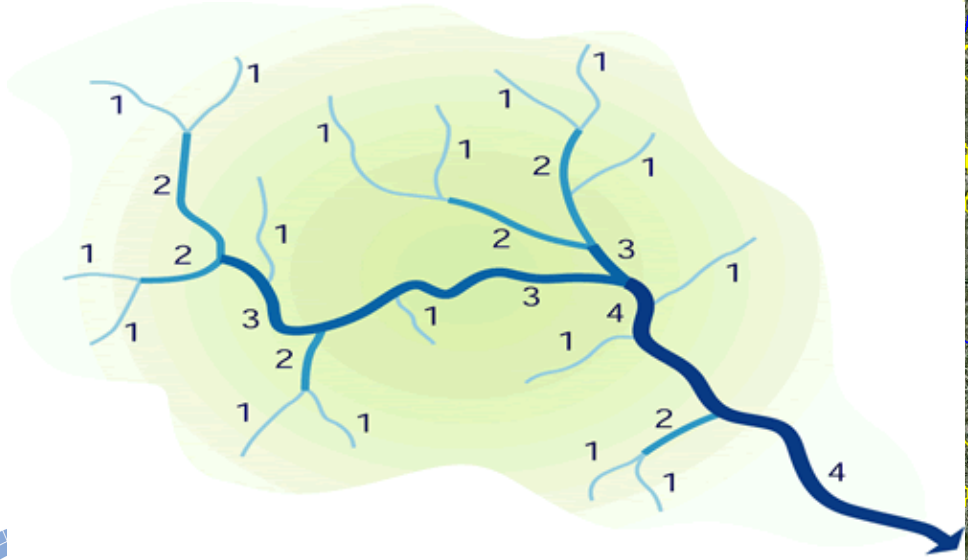
Navigating from topography to topology without Euclid

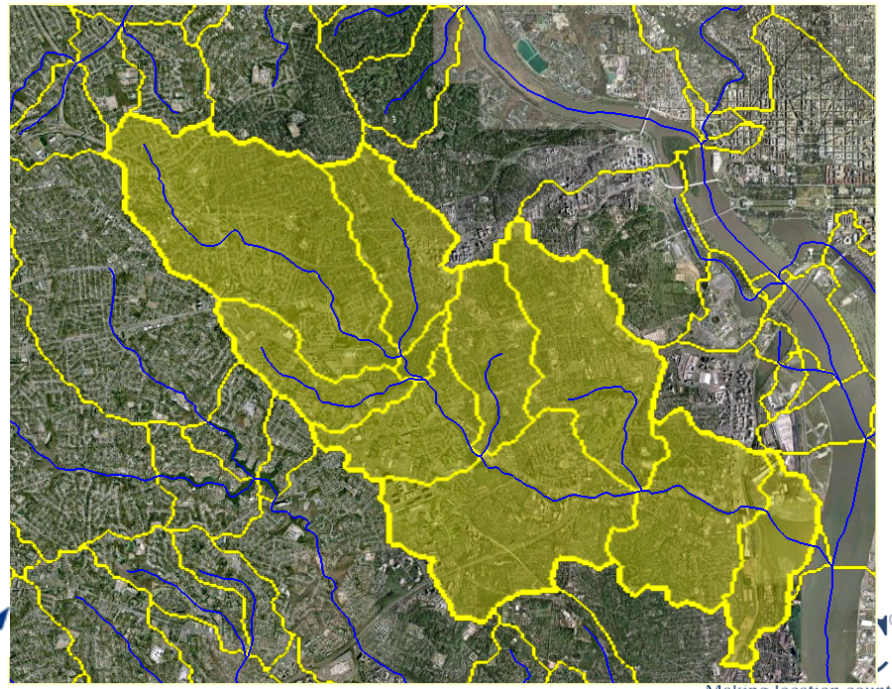
- Spatial processes of interest often involve movement or paths of influence
- Sometimes unbounded and Euclidean, e.g viewsheds
- Sometimes unknown (Enos – partisan influence) or empirical (eigenbehaviors)
- Often clearly constrained to network pathways, e.g. streets, stream channel networks, utility pipe networks.
- Opportunity for lower dimensional representation but may be complex to combine with other geometries



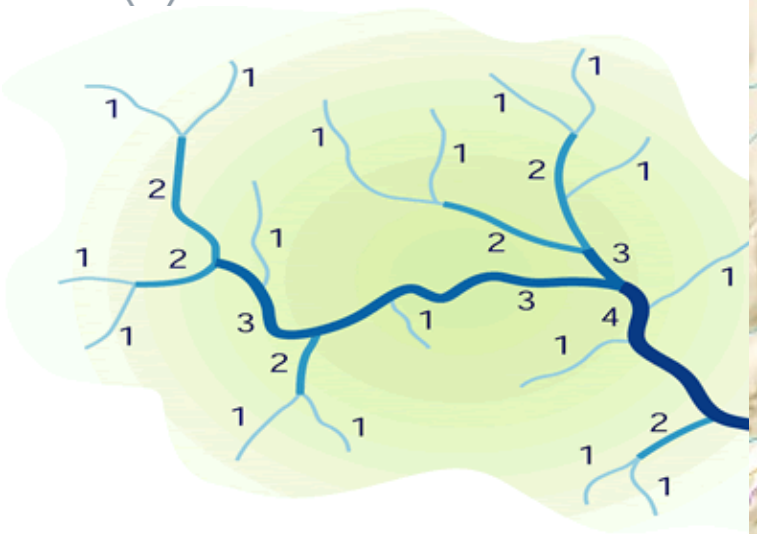
Surface Hydro Network Example

- Surface (near-surface) hydro network basics
  - Water flows into and onto nested *Catchment* areas
  - Convergence into *Basin* filling *Waterbodies* and *Channel* flowing *Streams*
  - Water flows out of *Catchments* at lowest elevation *Outfalls*
  - *Arrangement of* Outfalls *and connecting hydro features on the landscape defines a node-edge topological network*





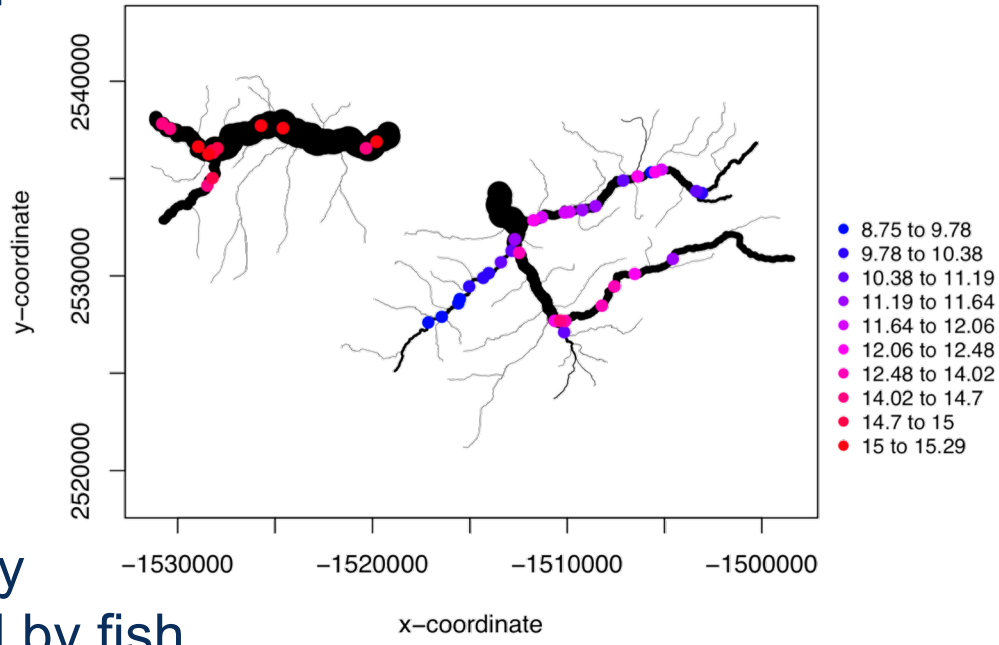
National Hydrography Dataset



- NHD: all streams and lakes at scales of 1:24000, 1:100000 in a network
- WBD: defines the areal extents of surface water drainage to a point
- NHD+ connects each NHD reach to a catchment
- Virtual reaches needed to include water bodies in graph

Case study 4

- Flowing likelihood of water, fish, and heavy metals
  - Network processes may introduce complexity, e.g. spatial correlation analysis constrained by network proximity
  - Study (B. Gonzalez) of dioxin in Maine rivers:
  - Sediment-borne dioxin only moves downstream
  - Fish-borne dioxin moves up and downstream
  - R package SSN_STARS for stream network correlation modeling: dioxin correlated by clay-rich sediment, dispersed by fish…



Case study 5

- Mile posts, river miles, length of a coast, and paths through space-time
  - Linear referencing a common dimensional reduction measure, e.g. road / train rail mileposts
  - Length of natural features more problematic.
  - River miles used for sampling – independent of streamline resolution, more representative of flow process, hard to compare with stream reaches
  - Which coastline length /pertains to which process depends on scale.
  - Interest as well in position along 4D trajectories to index mobility data.
  - Generalized to space-filling curves or geohashes as 1d indices for regions. Tradeoff of easy scanning vs unreliable proximity.

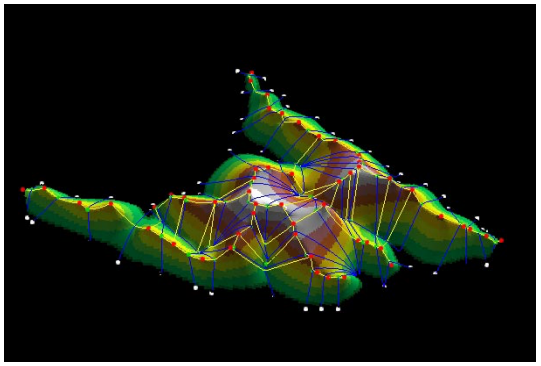
Geohashes

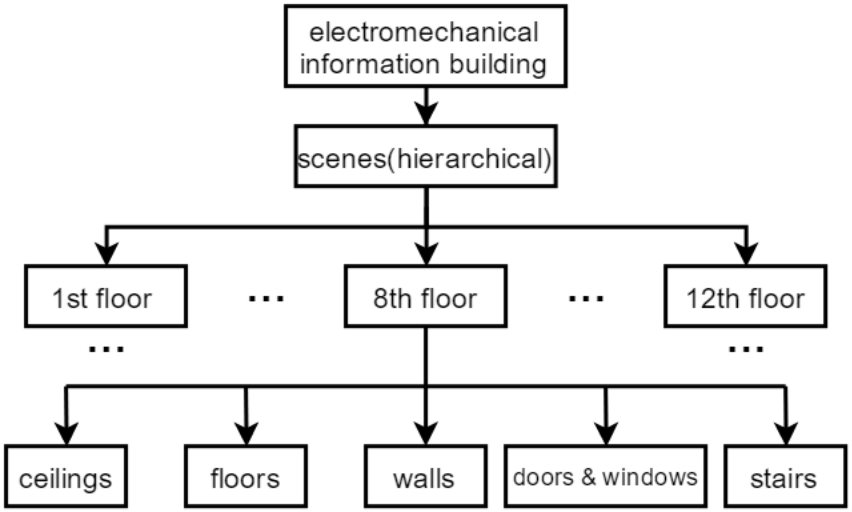
- **Geohash** was originally a 1D spatial index method used by Microsoft Terraserver that interleaved lat and lon digits from a point location.
- It now refers to a ”public domain geocoding system invented by Gustavo Niemeyer[1], which encodes a geographic location into a short string of letters and digits. It is a hierarchical spatial data structure which subdivides space into buckets of grid shape, which is one of the many applications of what is known as a Z-order curve, and generally space-filling curves.”*

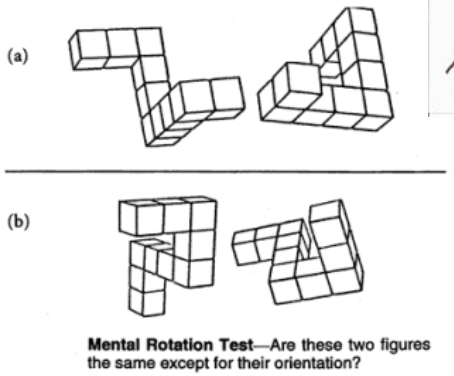
geohash length	lat bits	lng bits	lat error	lng error	km error
1	2	3	±23	±23	±2500
2	5	5	±2.8	±5.6	±630
3	7	8	±0.70	±0.70	±78
4	10	10	±0.087	±0.18	±20
5	12	13	±0.022	±0.022	±2.4
6	15	15	±0.0027	±0.0055	±0.61
7	17	18	±0.00068	±0.00068	±0.076
8	20	20	±0.00008 5	±0.00017	±0.019

Case study 6

- Making mereology with feature graphs and surface networks
  - Sometimes “part-of” is enough of a position, e.g. BIM hierarchies
  - Purpose of coordinate geometry analysis is often to derive feature graphs / world maps consistent with spatial cognition
  - Feature networks have also been applied to continuous landscapes with mixed results





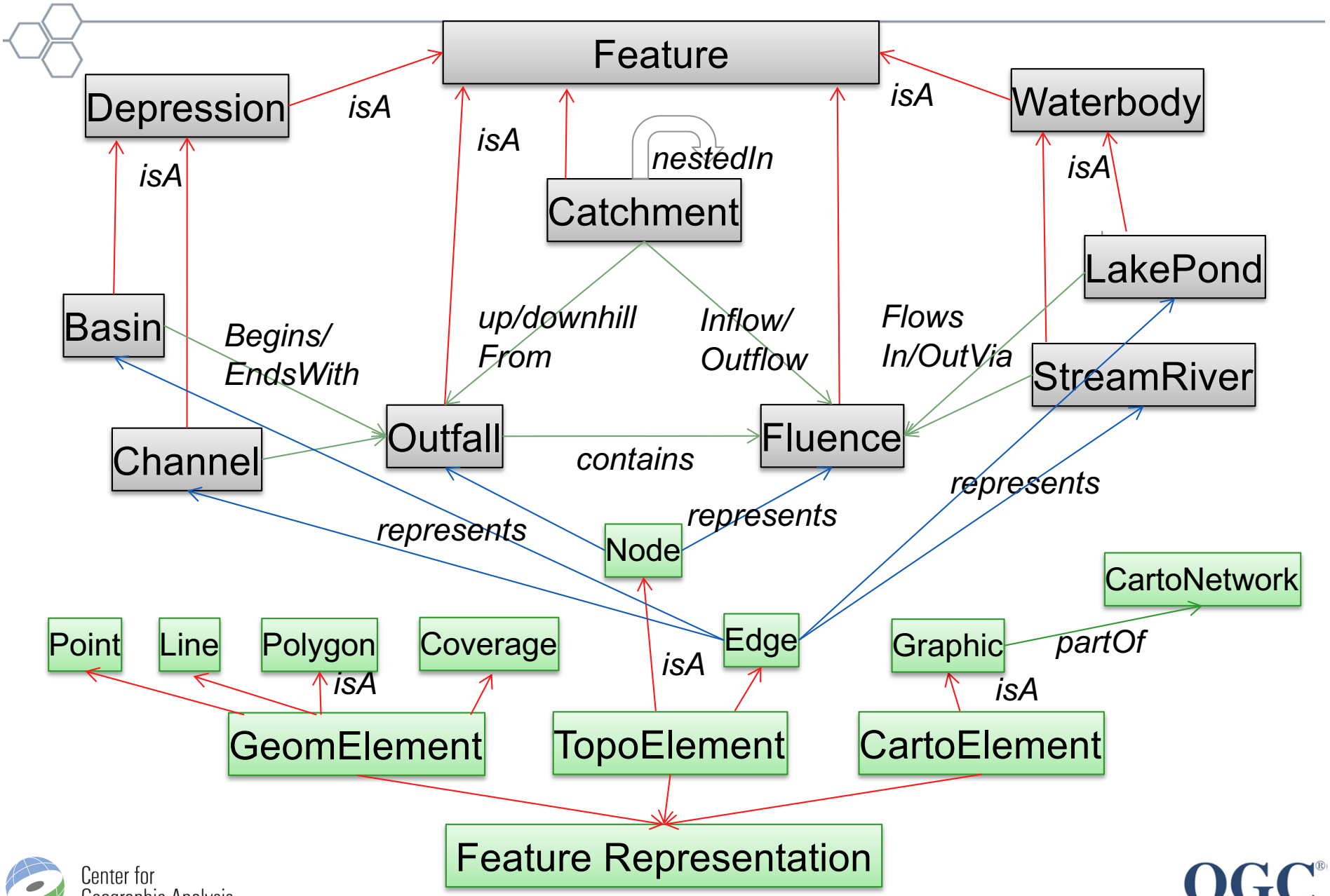


Mental Rotation Test—Are these two figures the same except for their orientation?

Center for Geographic A
Harvard University

OGC®
Making location count.

Ontological Approach to Hydro Networks



Center for Geographic Analysis
Harvard University

OGC®
Making location count.

Surface Network

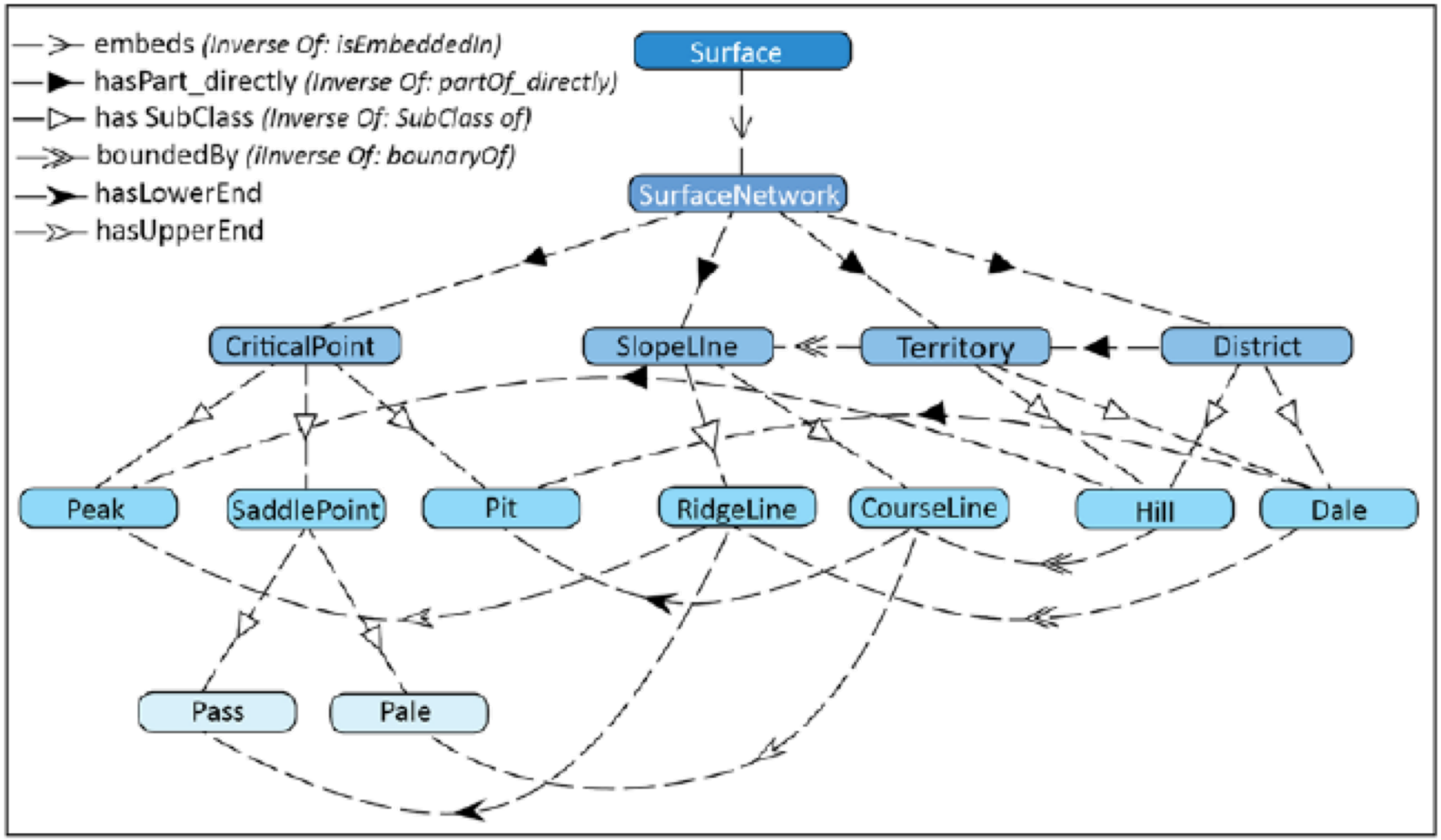
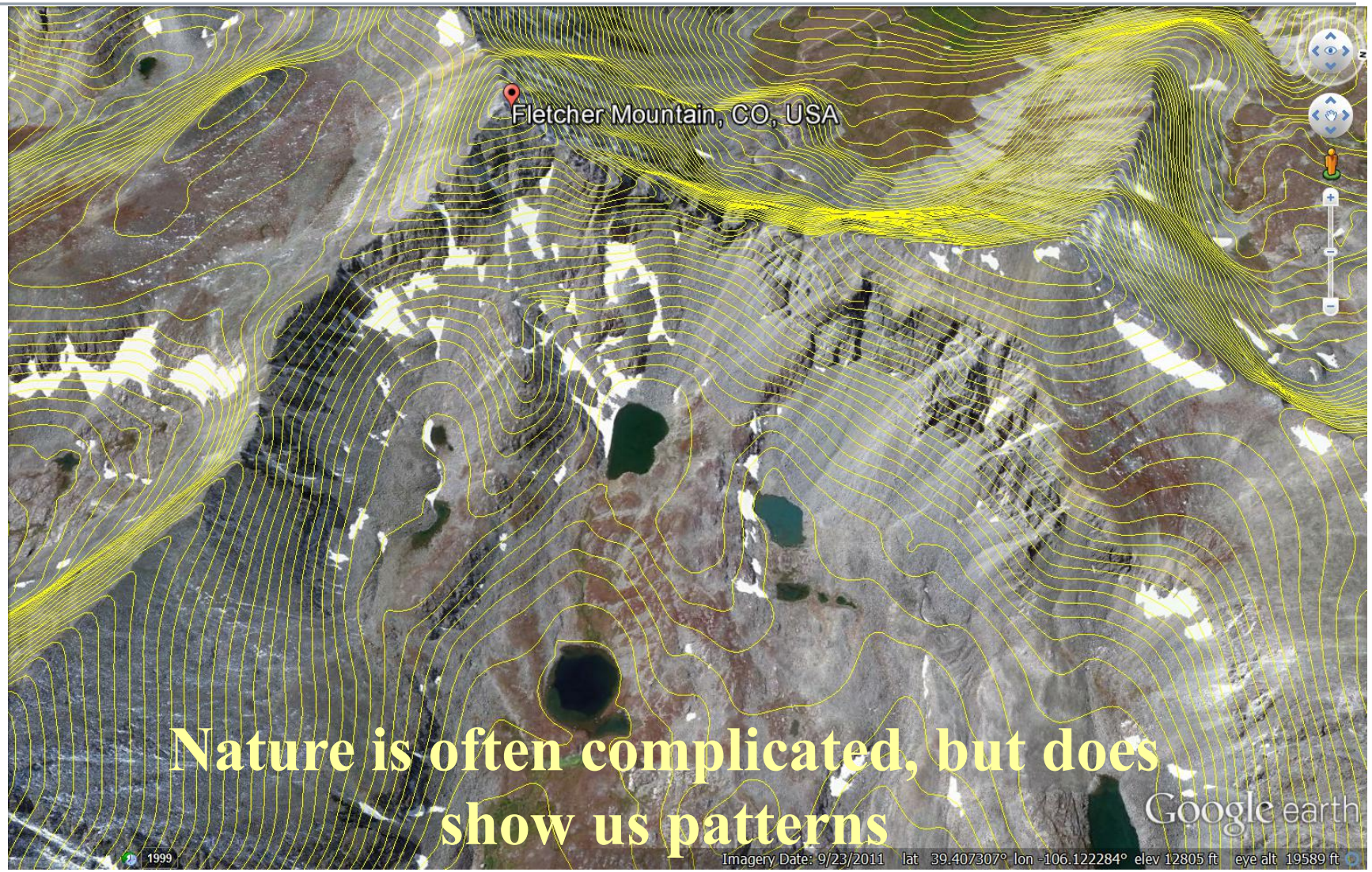


Fig. 1. Schematic representation of (only) the relationships between the classes of the Surface Network ODP. The inverse relationships are not shown in the diagram for maintaining clarity, but are mentioned in the diagram key. For other properties specific to a particular class only, consult the full ontology and the main text.

Spatial Processes Matter



Conclusions

- Simplification in GIS can be cognitive, conceptual, and/or computational
- Lower dimensions can be represented by chosen geometric representations, constraints on positioning, even constraints on scale.
- Dimensionality, just as the feature discernment, depends on the intended application and the target phenomena
- Many such decisions are made assuming certain algorithmic, computational, and/or visual limitations that may not be (any longer) valid.
- Collaboration between computational and geographic information scientists can yield new possibilities for geographic understanding.