Valid statistics with amounts in geographic information

Simon Scheider

Department of Human Geography and Spatial Planning, Utrecht University

(Geospatial Big Data and Societal Transformations lecture series, Bayreuth)

Universiteit Utrecht

03-05-2022

What is and why do we need validity?

Validity makes science “scientific” (e.g. Geographic Information Science)

Yet, compared to its importance, definition in statistics is **very vague**:

~“a method is valid if it measures what it is supposed to measure”

- Validity (statistics), the application of the principles of statistics to arrive at valid conclusions
  - Statistical conclusion validity, establishes the existence and strength of the co-variation between the cause and effect variables
  - Test validity, validity in educational and psychological testing
  - Face validity, the property of a test intended to measure something

Face validity is the extent to which a test is subjectively viewed as covering the concept it purports to measure.

Is there a theory that could let us determine validity for geographic information methods?

Outline

- GIS analysis = valid transformation of geographic information
- Validity of methods
- Case study: invalid measurement of exposure of bike riders
- Concepts:
  - Core concepts
  - Amounts
  - Extensivity
  - Homeomericity
- So why was the student’s solution invalid?
- Outlook: valid transformations for GIS automation and QA

GIS analysis = valid transformation

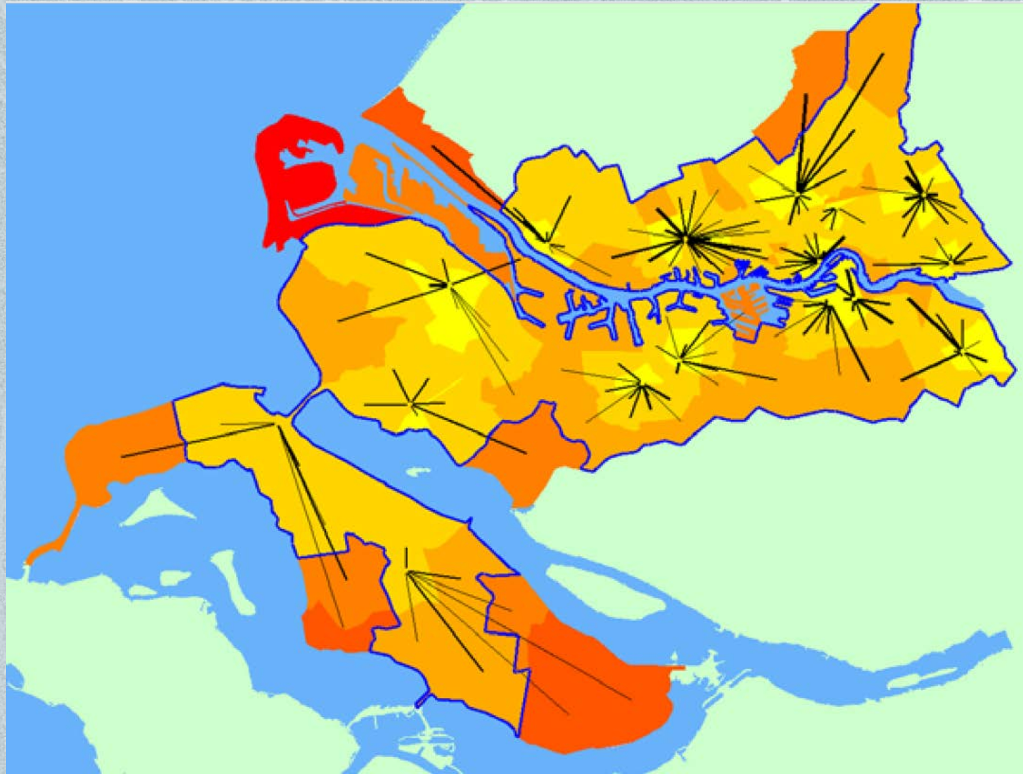
A pragmatic model of GIS know-how

GIS analysis in Geography

For example, Human Geographers ask:

"What is the accessibility of postcode areas for ambulances in Rotterdam?"

- Answer is a *map generated for a purpose*:
  red: low accessibility
  yellow: high accessibity
- *requires design of a valid workflow to transform information for this purpose*
- This is not a problem *of retrieval*
- This is not a problem of *statistics*
- Requires *procedural know-how/practice*



Procedural vs. declarative knowledge

Procedural/pragmatic

Knowing HOW something can be done

- Gilbert Ryle (1949): *knowledge is a disposition*
- Terry Winograd (1972): *Every word is a program* (SHRDLU)
- Helen Couclelis (2009): From *spatial reasoning* to *purpose and design*
- Peter Janich (2006): *A pragmatic view on information science*

Declarative

Knowing THAT something is the case

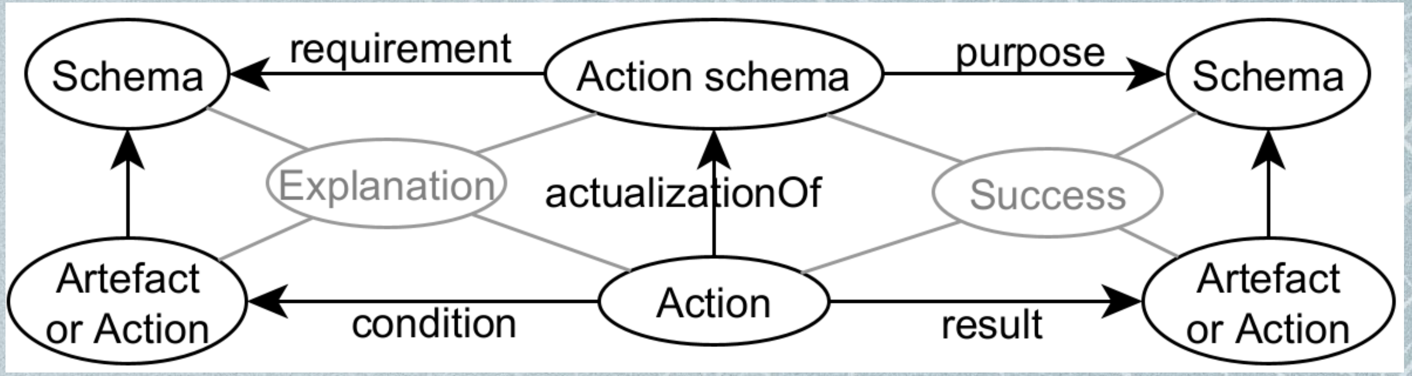
- Inductive: Data-driven research (starting from facts)
- Deductive: Axiomatic reasoning (starting from axioms and facts)
- Abductive: Explanation of facts (starting from axioms and facts)

A pragmatic model of know-how

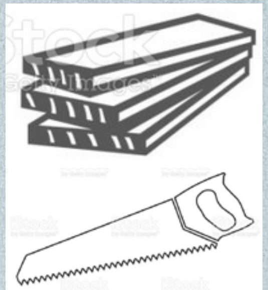
Janich (2001):

- Schemas can be actualized (e.g. in artefacts or actions)
- Action schemas can have *requirements* and *purposes*
- Actions can have *conditions* and *results*
- To *succeed* = realize purposes by actualizing results
- To *explain* = realize requirements by conditions

Explanation of success of actions, based on actualization (satisfaction) of schemas

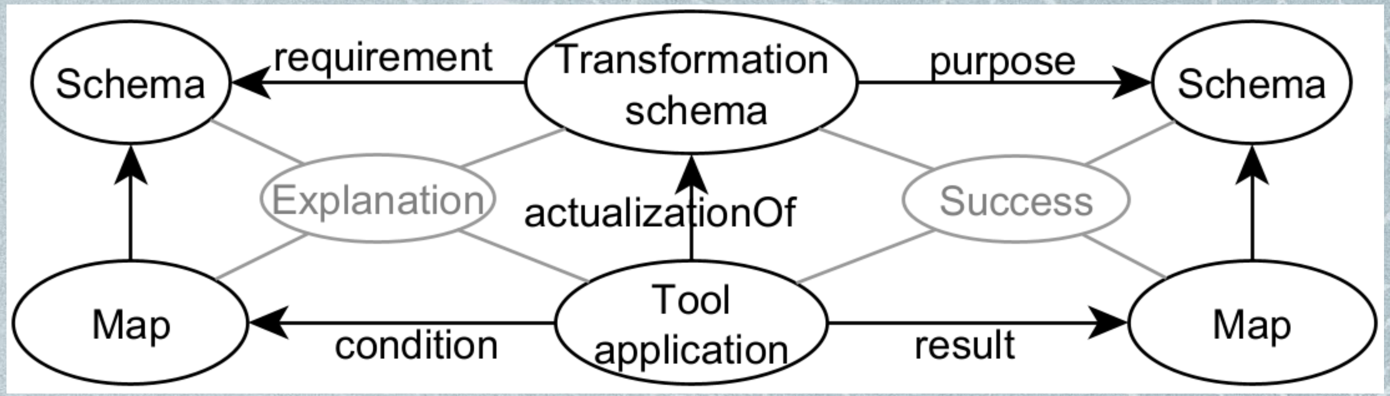


Know-how of a carpenter for making a table:



Know-how in (geographic) information ~ know how to transform information

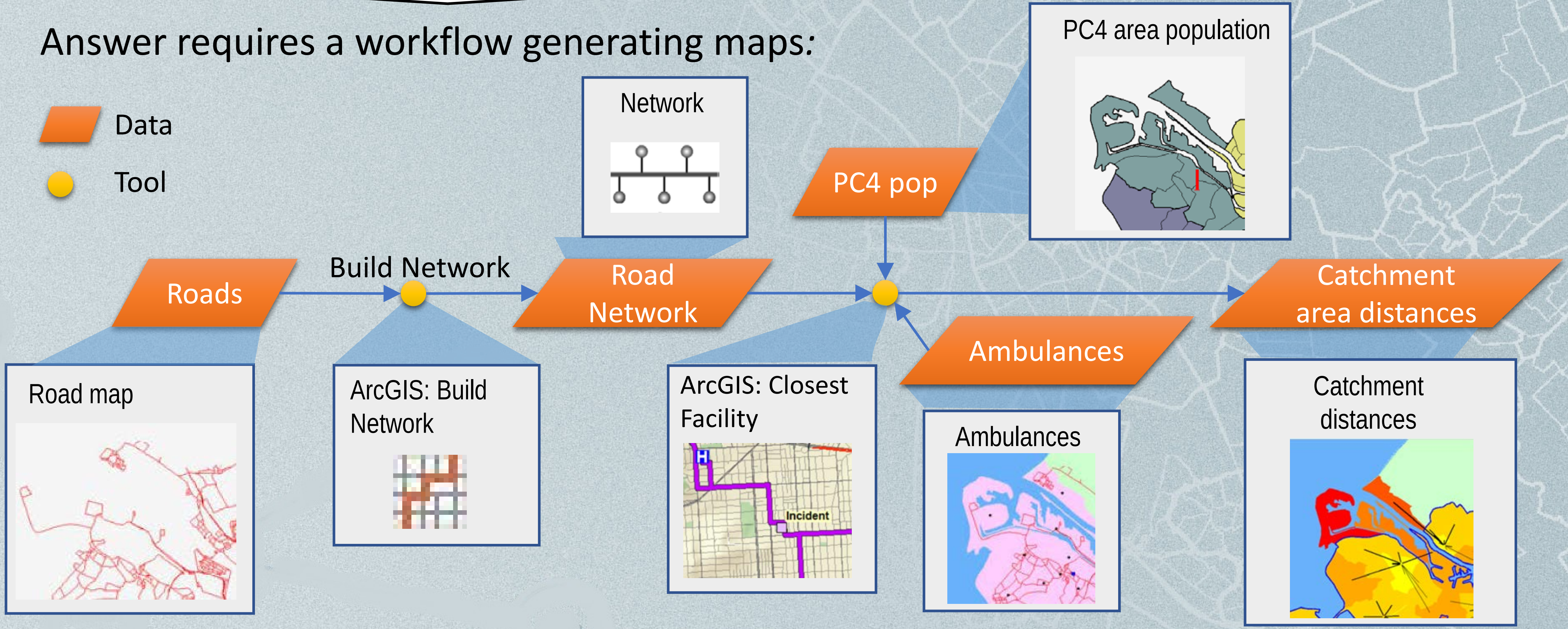
- Action schemas = *transformation schemas*
- Artefacts = *maps*
- Actions = *tool applications*
- Schemas = *map interpretations*
- =“geoinformation”
- succeed = resulting map is actualized purpose of transformation schema



GIS analysis = (valid) data transformation

“What is the accessibility of postcode areas for ambulances in Rotterdam?”

Answer requires a workflow generating maps:



Validity of methods

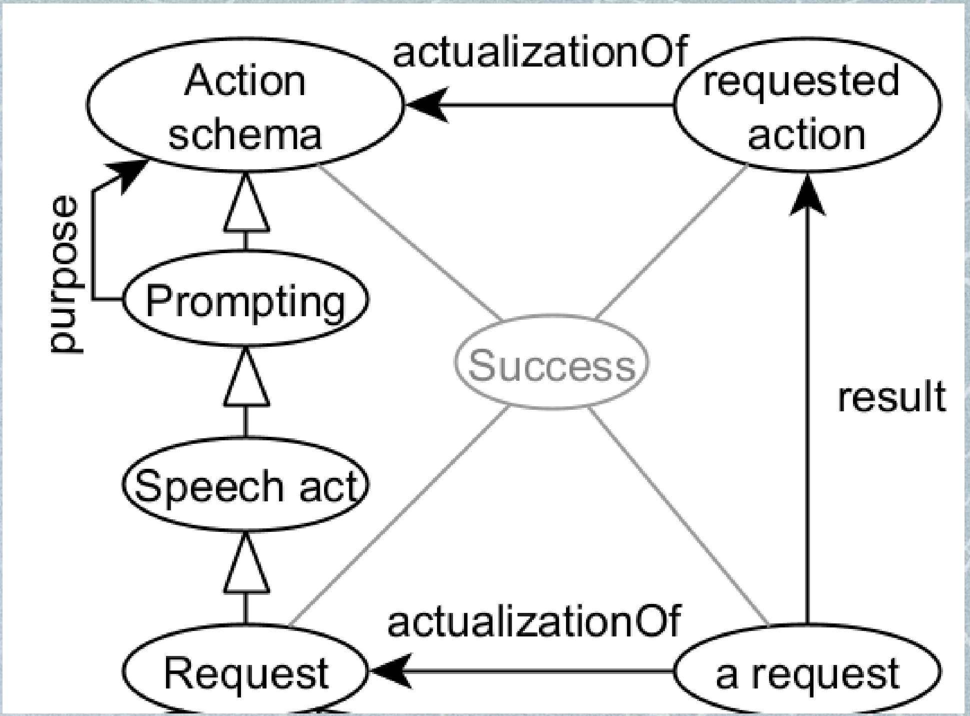
A pragmatic model

What is validity (from a pragmatic viewpoint)?

Validity (“Geltung”) defined based on success (Janich 2001):

- **Requests** are speech acts (purpose are other actions)
- successful if they **prompt actions** of the **purpose schema**
- **Valid request =**
  a request that is expected to be successful (prompts intended action)
- For example: legal validity (following action rules)

  e.g. a Covid-19 rule is valid if rule is expected to be followed



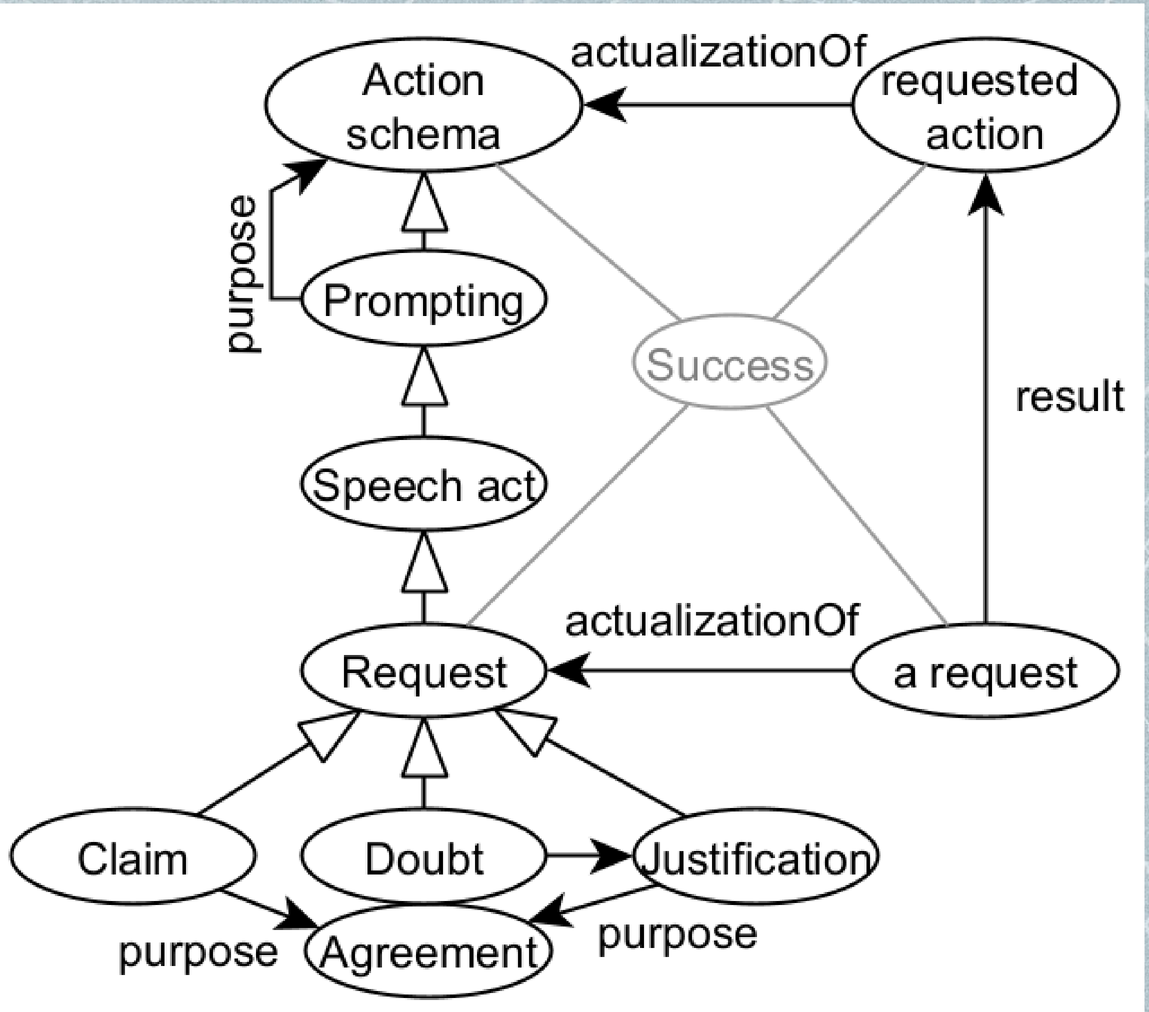
What is validity (from a pragmatic viewpoint)?

Validity of claims (Janich 2001):

(focusing on truth of statements)

- Claim (purpose: agreement)
- Doubt (=unsuccessful agreement) (purpose: justification)
- Justification (purpose: agreement)

Valid justification = justification that succeeds in prompting agreement
Valid claim = claim that is successfully justified (agreeable)



Valid justifications

Successful justifications (Janich 2001) require **trans-subjective truth criteria** for statements:

1. conceptually sharp ("Begriffsschaerfe")
   (known how concepts apply) (defining concepts)
2. implied by inference rules (deriving statements)
3. testable by method/experiment (testing/explaining statements)
4. non-perturbed ("stoerungsfrei") (implicit know-how)

(criteria independent of **who** applies them and **in which context**)

Valid methods

= justifiably successful action schema (= with a valid justification for success)

⇒Method valid only w.r.t. some **purpose**

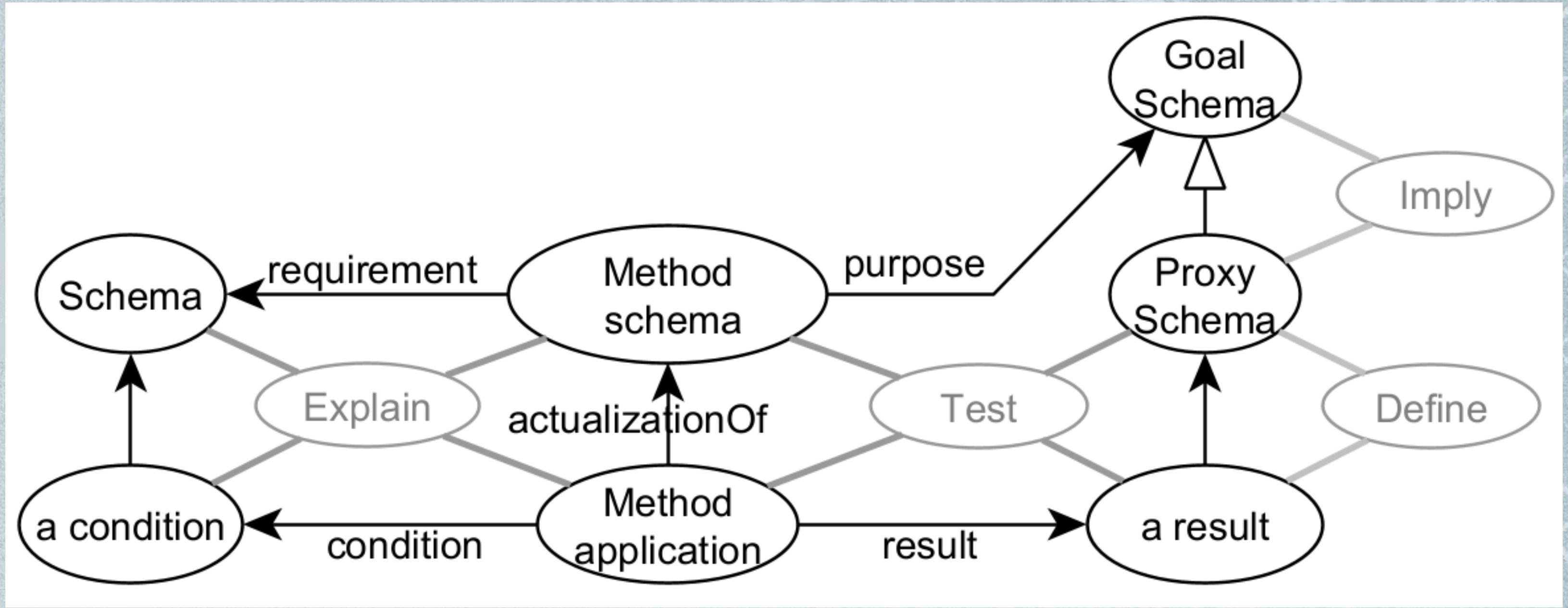
⇒Purpose is a **goal schema,** not necessarily satisfied by result, but

⇒**Implied by proxy:** goal schema is implied by proxy schema

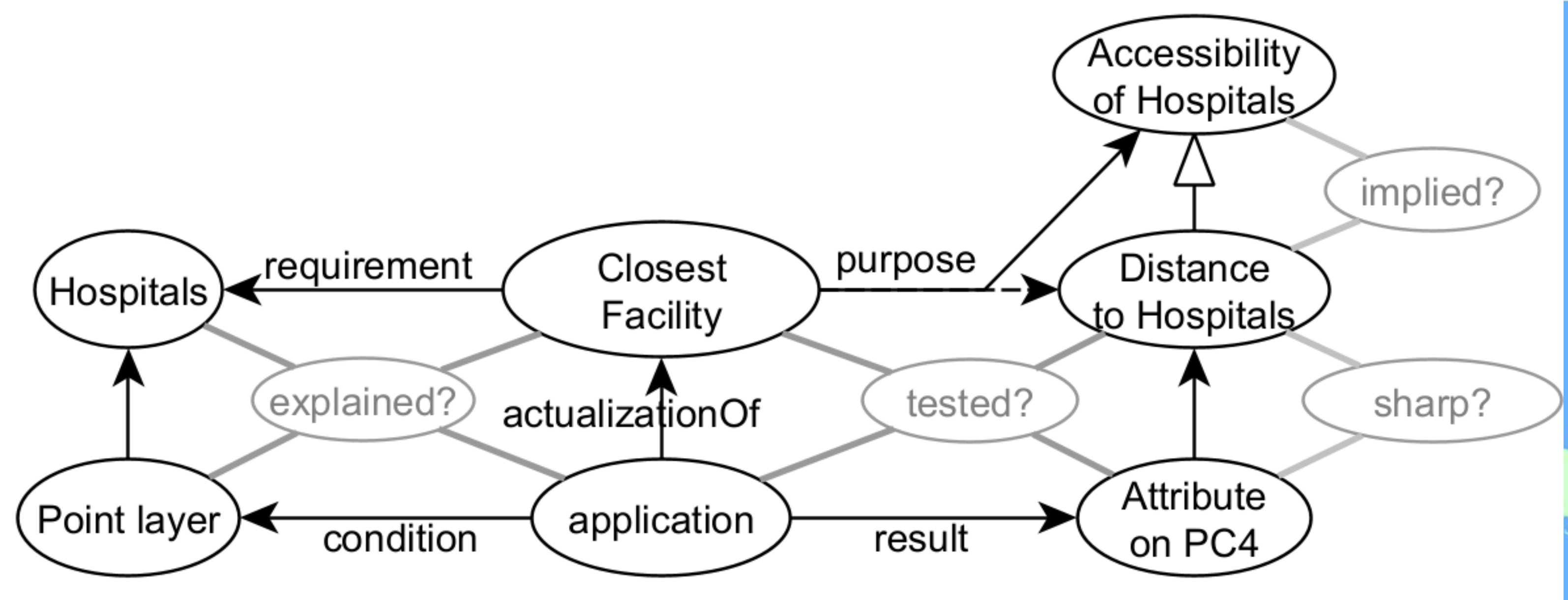
⇒Proxy schema is **defined** (it is known how to it applies to results)

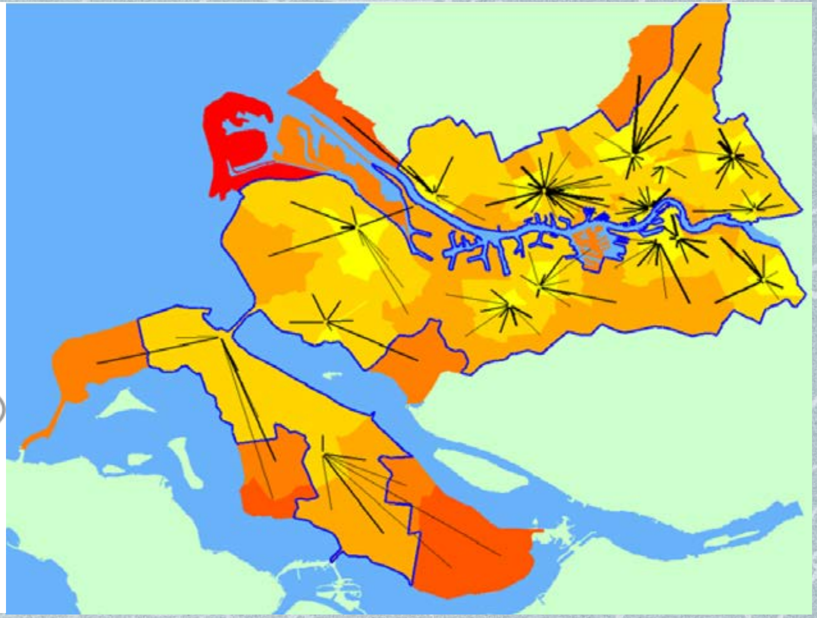
⇒Success is **tested (proxy)**

⇒Success is **explained (requirements)**



Example:
Is this a valid method for accessibility analysis?





In which ways can the method become invalid?

- can success be explained? (*are hospitals used?*)
- is the proxy schema sharp? (*can we determine the result is a distance?*)
- was it tested? (*does the tool function, i.e., does it deliver a distance?)*
- is the goal implied by the proxy? (*does distance to hospitals imply accessibility of hospitals?*)

Why is this important?

- The current tendency to regard evaluation as statistical tests/fitness measures (ML) on data **oversimplifies** the problem of validity
- … because it reduces validity to **testing**, which is **insufficient** for justifying the validity of GIS methods
- A useful theory of validity would require in addition:
    1. Concepts for schemas (if possible, defined): Which concepts are needed?
    2. Purposes: Based on concepts
    3. Inference rules for goal schemas (which rules)?
    4. Explanations of results: Based on provenance

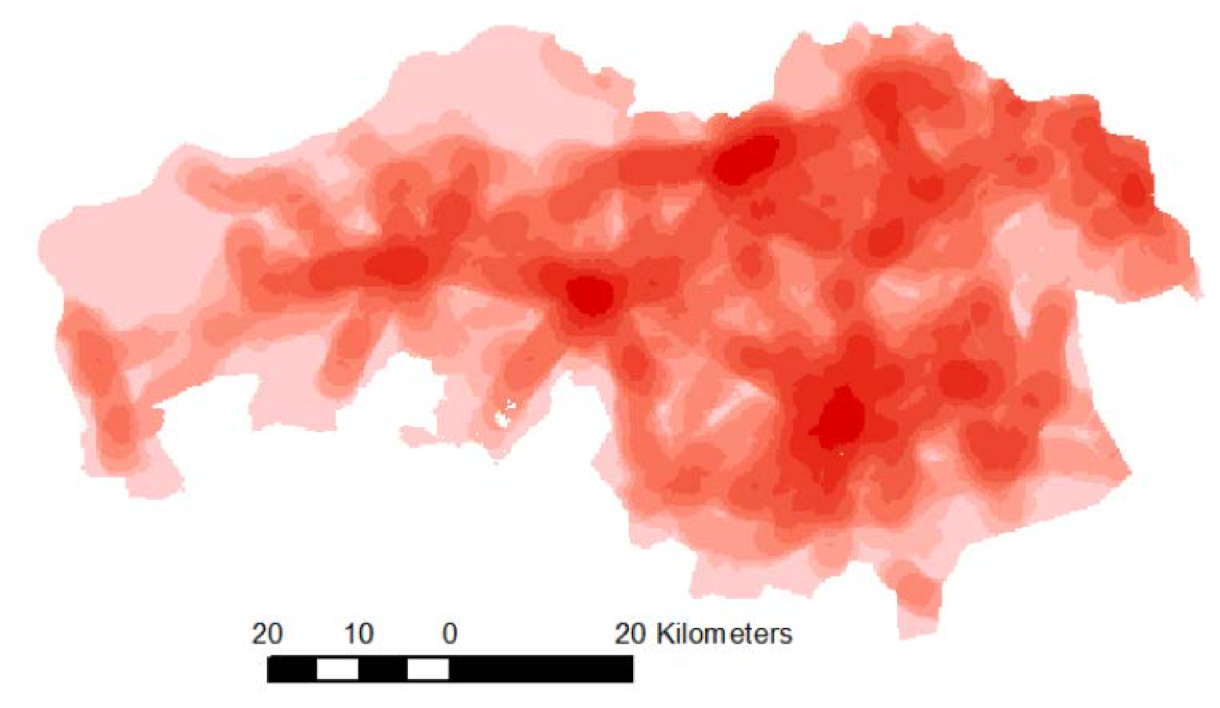
Case study

validity of exposure measurement for bike rides

Case study: Explaining detouring behavior of bike tracks

How does the environment influence route choice?

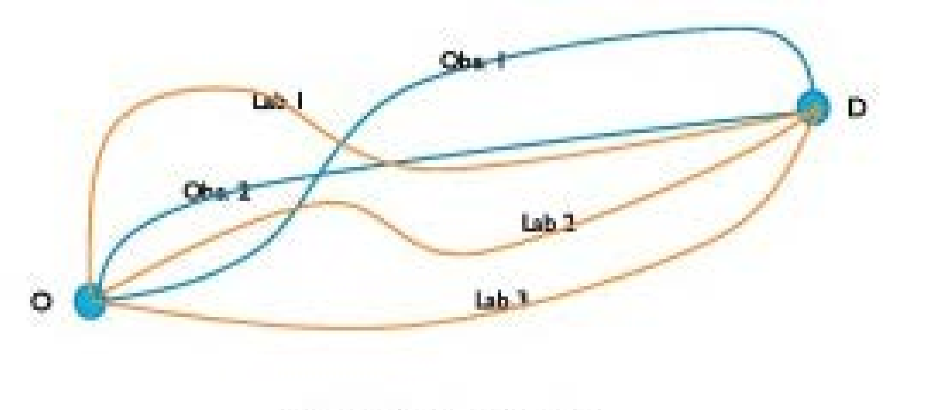
- GIMA MSc course exercise
- 40.000 tracks from a cycling stimulation program in Noord-Brabant
- In the following we study a model proposed by one group of students



Track density map of Noord-Brabant, data source: B-riders

Case study:
Explaining detouring behavior of bike tracks

- Measure deviation of observed track from shortest path in terms of *length difference*
- Regress length difference against difference in *exposure to environmental factors*
- The more different exposure, the more it may explain route deviation from shortest path
- More generally: route choice modeling



d) Labelling choice set

Formation of choice sets
(Broach et al. 2012 "Where do cyclists ride? A route choice model developed with revealed preference GPS data”)
(Ton et al. 2018 "Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam”)

Case study:
How to measure environmental exposure?

Focus on two environmental factors:

1. Exposure to green land cover. Measured by land use coverage (polygon data, source: BBG)

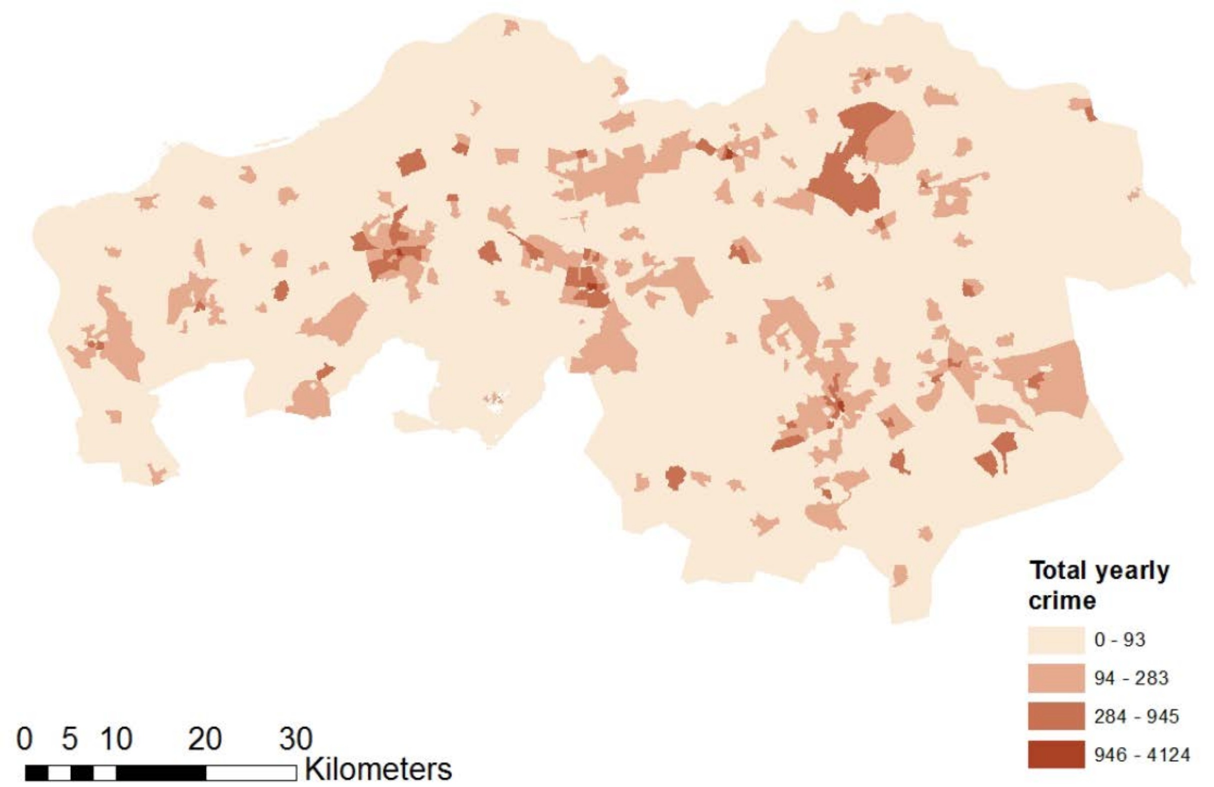
BBG 2012 (red/orange colors are built areas) (Source: CBS)

Case study:
How to measure environmental exposure?

Focus on two environmental factors:

1. Exposure to green land cover. Measured by land use coverage (polygon data, source: CBS)
2. Exposure to route safety. Measured by crime statistics (CBS) on neighborhood level (polygon data, source: CBS)

   …

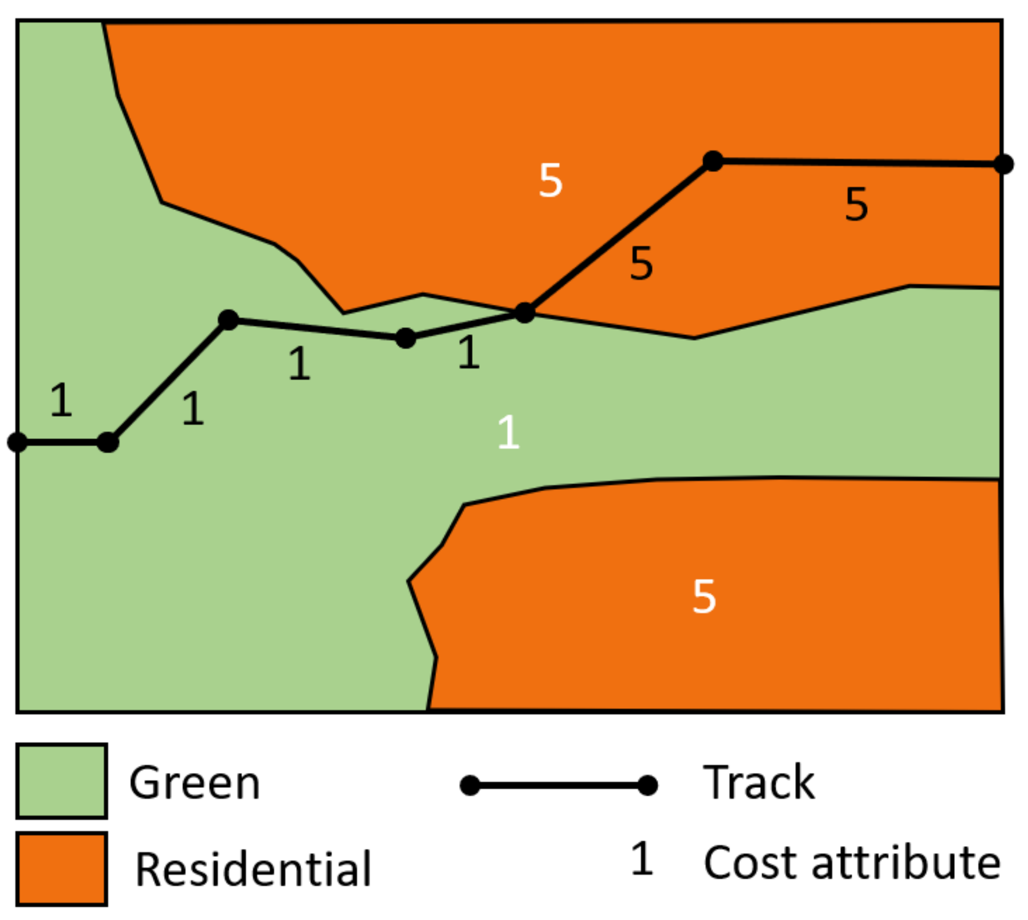


Crime numbers per statistical neighborhood, Noord-Brabant (Source: CBS)

Case study:
Method of exposure measurement I

- Interpret environmental factors as costs for biking
- Measure exposure by
  1. Overlay tracks with polygons
  2. Assign polygon costs to segments
  3. Measure lengths of segments
  4. Multiply lengths with costs and sum over each track t

$$\mathrm{cost(t)} = \sum_{n=1}^{|t|} length(s_n) * cost(s_n) \,|\, \bigcup_{n=1}^{|t|} s_n = t$$



What's wrong with measuring in this way?

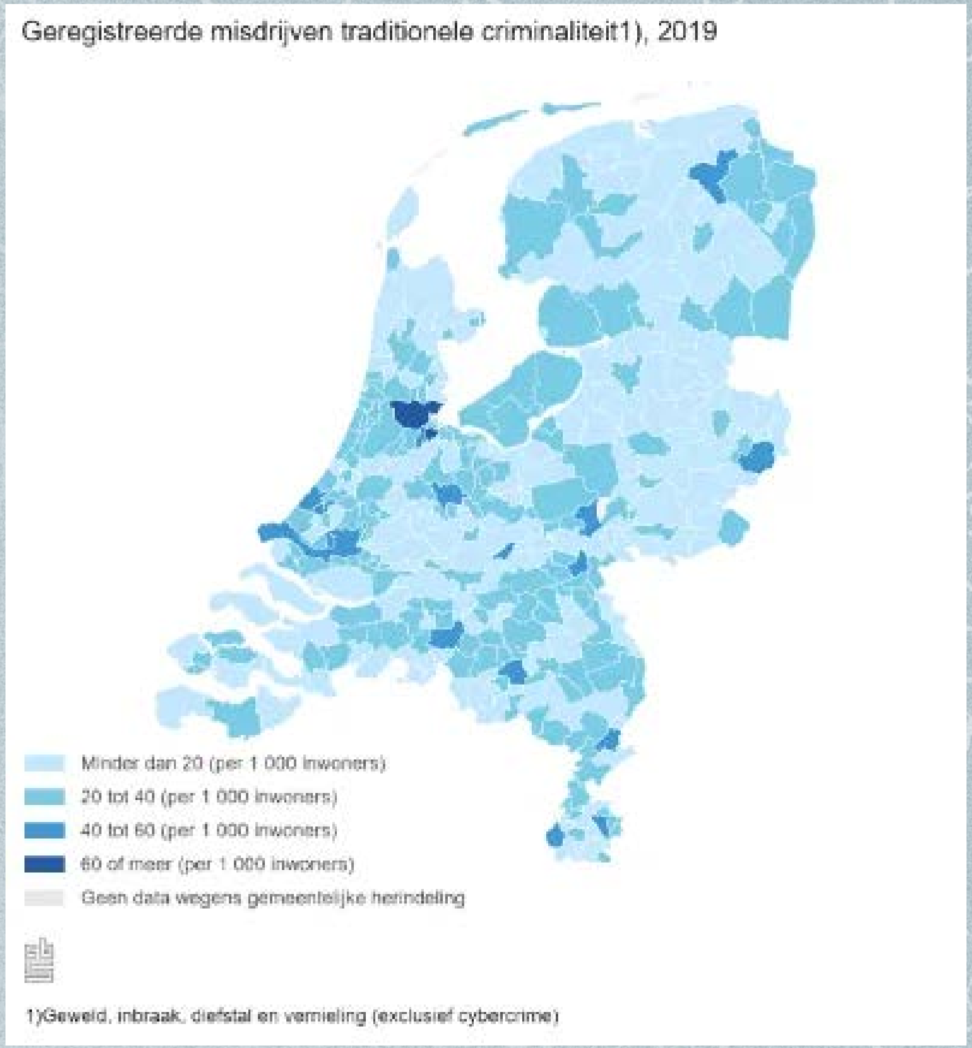
To see the problem, think about what happens if we decrease the resolution of polygons?

Then crime exposure will increase **just because we used larger statistical units (e.g. municipalities).**

Note that this is not the case if we decrease the resolution of landuse!

⇒There must be a fundamental conceptual difference between the two datasets which the method does not account for

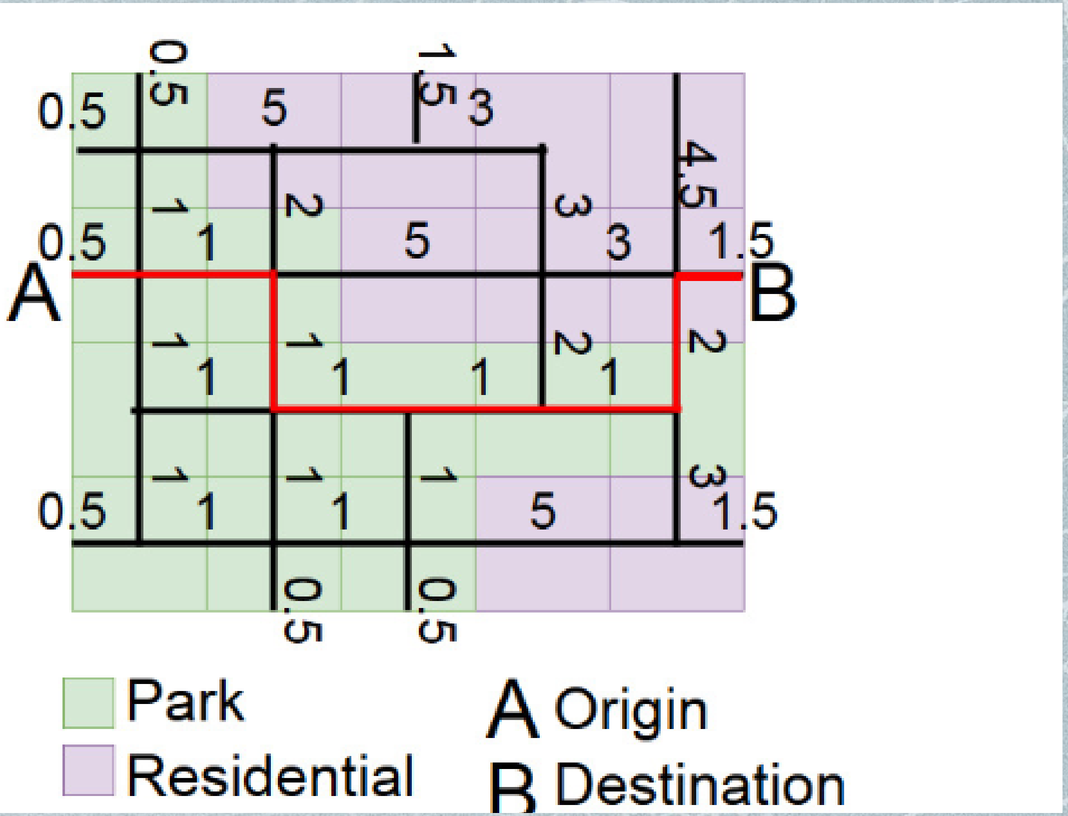
⇒This is causing an invalid method



Case study:
Method of exposure measurement II

- Interpret environmental factors as costs for biking
- Measure exposure by
    1. Generating a cost raster
    2. Rasterizing track
    3. Summing track cell costs over each rasterized track t

$$\mathrm{cost(t)} = \sum_{n=1}^{|rasterize(t)|} cost(c_n) \Big| \bigcup_{n=1}^{|rasterize(t)|} c_n = rasterize(t)$$



Why is this method not valid?

- Valid methods must satisfy the purpose (*goal schema) regardless of context (*purpose: *amount measured over a track)*
- Yet results do not satisfy this goal in the case of crime:
  - Crime statistics: the amount of crime measured within a neighborhood (≠ within a track segment/cell)
  - Landuse: the amount of green measured within a track

⇒This problem is sometimes called **ecological fallacy** (similar to MAUP)

⇒ **but why** is case a fallacy, and not the other?

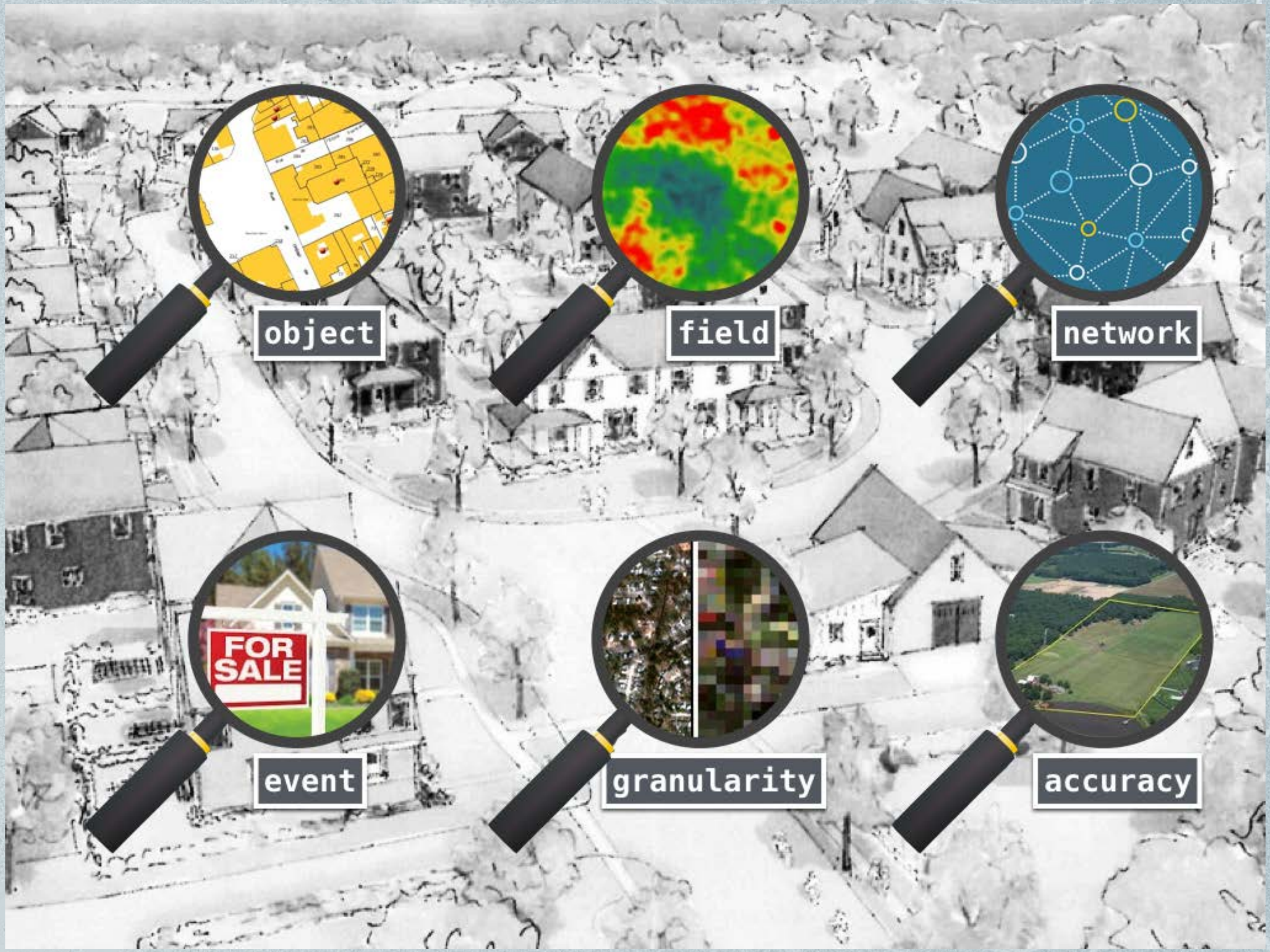
⇒Thus: what precisely is the **conceptual difference** between these cases?

⇒ We currently lack any theory that explains this

Concepts

Core concepts of spatial information (Kuhn 2012)

- Objects of study in Geographic Information Science (… like “cell” in biology or “value” in economics)
- Lenses for studying the environment
- Content concepts, data quality concepts



Amounts and magnitudes (examples)

Examples of amounts:
Amounts of matter,

collections of objects,

amounts of space,

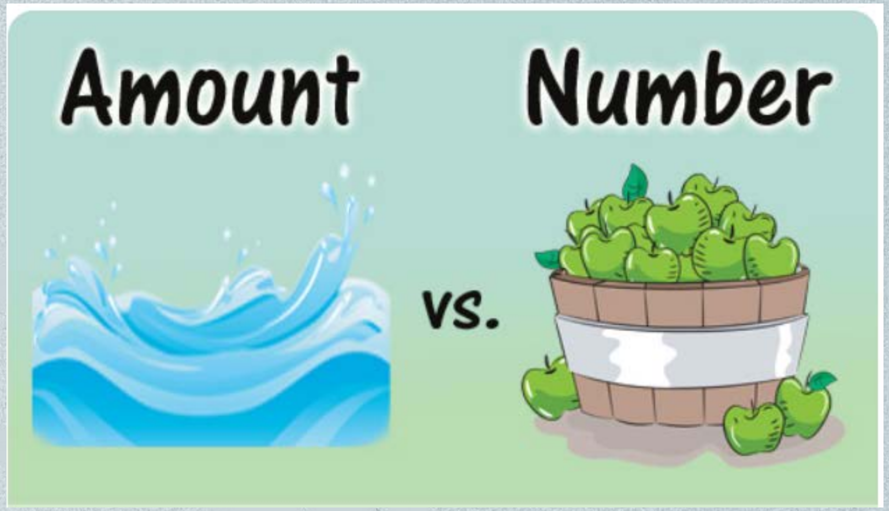
amounts of time, …

Examples of magnitudes:
2 kg,

15 people,

20 km^2,

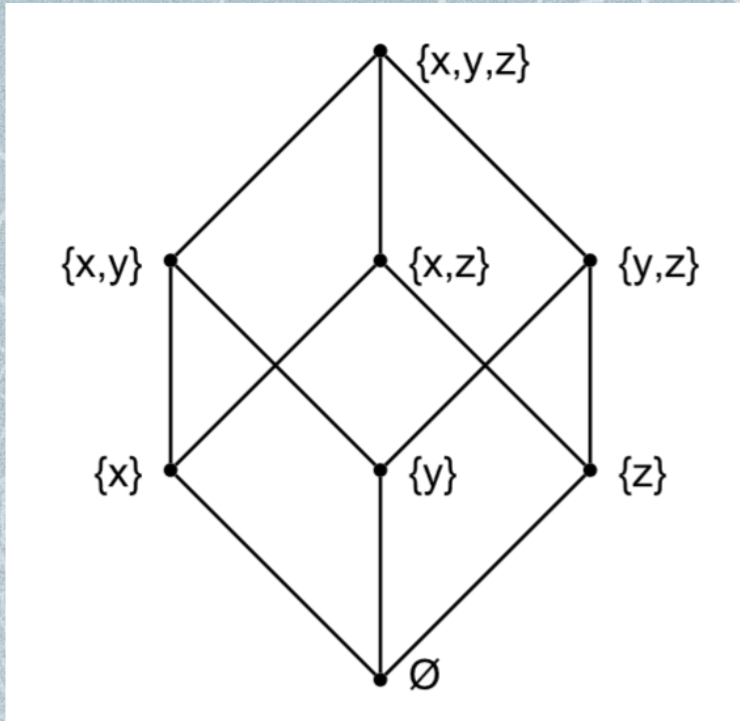
25 hours, …

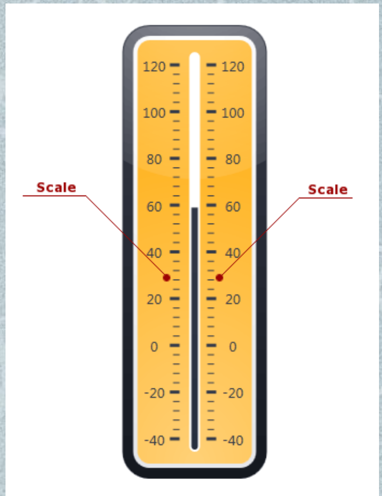


Amounts and magnitudes (formal)

- Amounts = mereological quantities forming a *boolean lattice*
  - Can be 'summed' (+) and 'subtracted' (\) and intersected (*) similar to sets
  - Parts (≤) form a lattice (≠ magnitude)
- Magnitudes = *linearly ordered* monotonic quantities, used to quantify amounts

(Top et al: forthcoming)





Amount algebra is different from magnitude algebra:

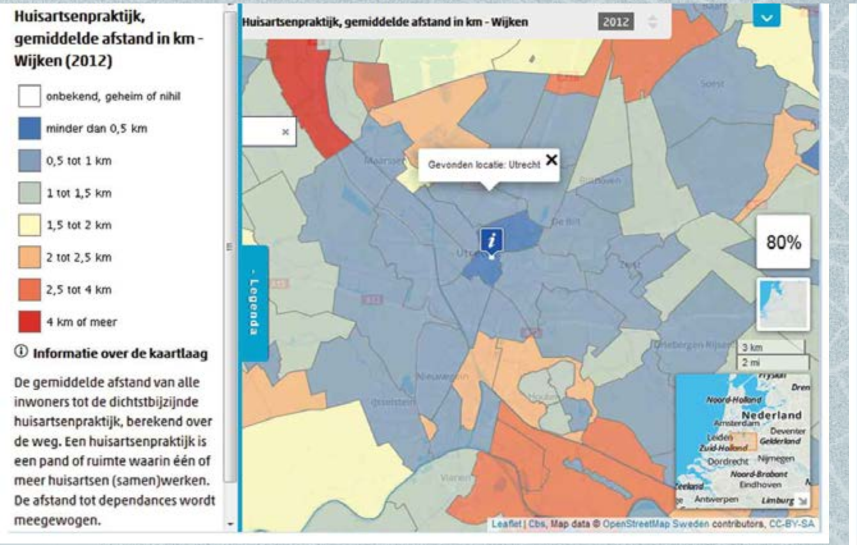
$(x \subseteq y) \implies x + y = y$	*Reflexivity of sums*
$x \subseteq y \implies x * y = x$	*Reflexivity of products*

Extensive/intensive measurement of amounts

- Measurement of amounts
  Control -> Measure
  (Sinton 1978)
- *Extensive* = measures add up over controls
- *Intensive* = this is not the case

(Top et al: forthcoming)

Control: Neighborhood regions
Measure: Mean distance to practitioner



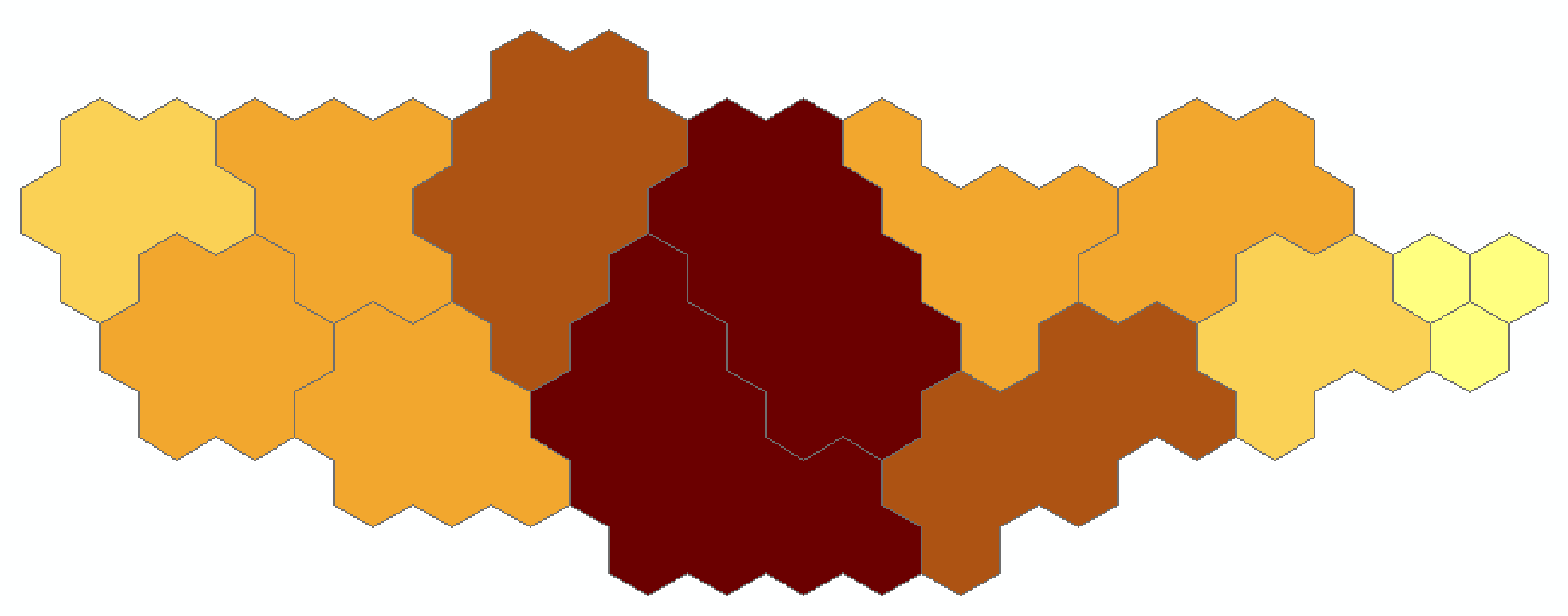
Control: Neighborhood regions
Measure: Number of cars

Definition 3. *Additivity and subtractivity of m measurements in quantity domain X*

$$\forall x, x' \in X(\neg O(x, x') \implies m(x) + m(x') = m(x + x')) \qquad \textit{Additivity}$$

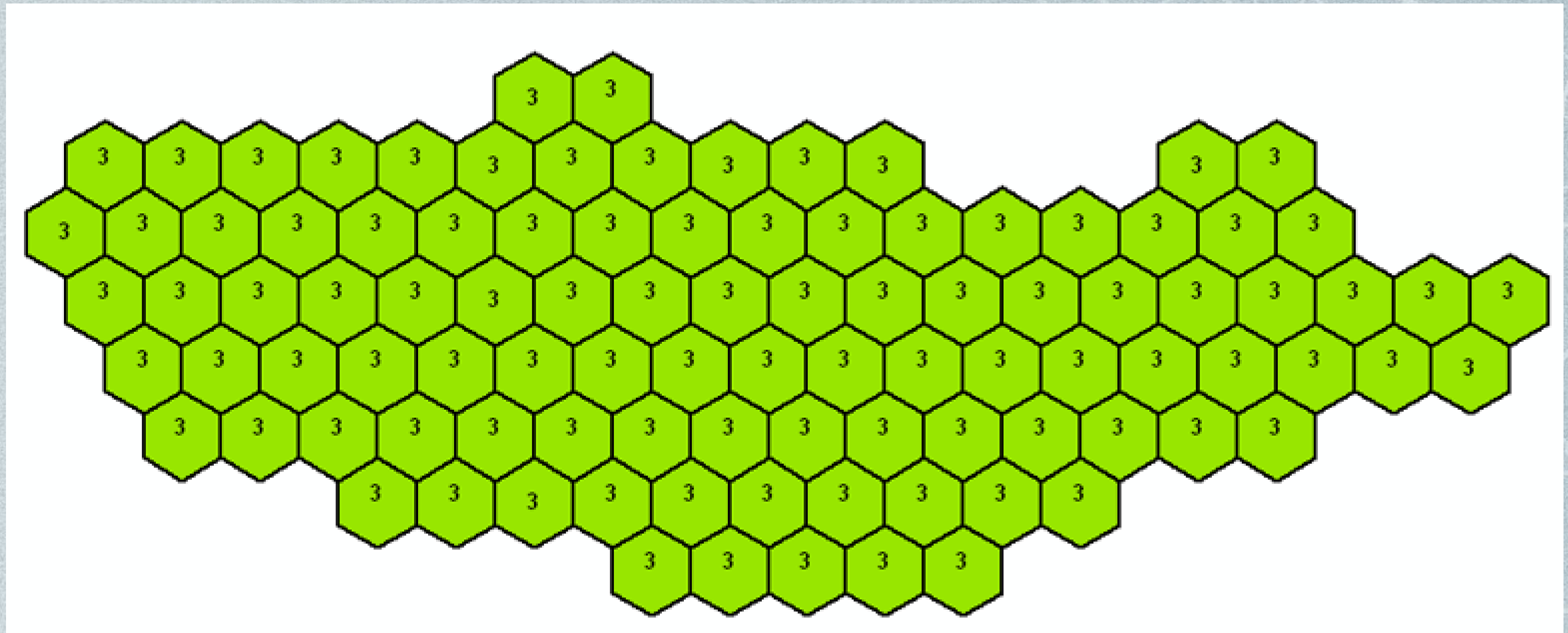
$$\forall x, x' \in X(x \preccurlyeq y \implies m(y) \setminus m(x) = m(y \setminus x) \qquad \textit{Subtractivity}$$

Attribute normalization and map types: Extensive vs intensive attributes



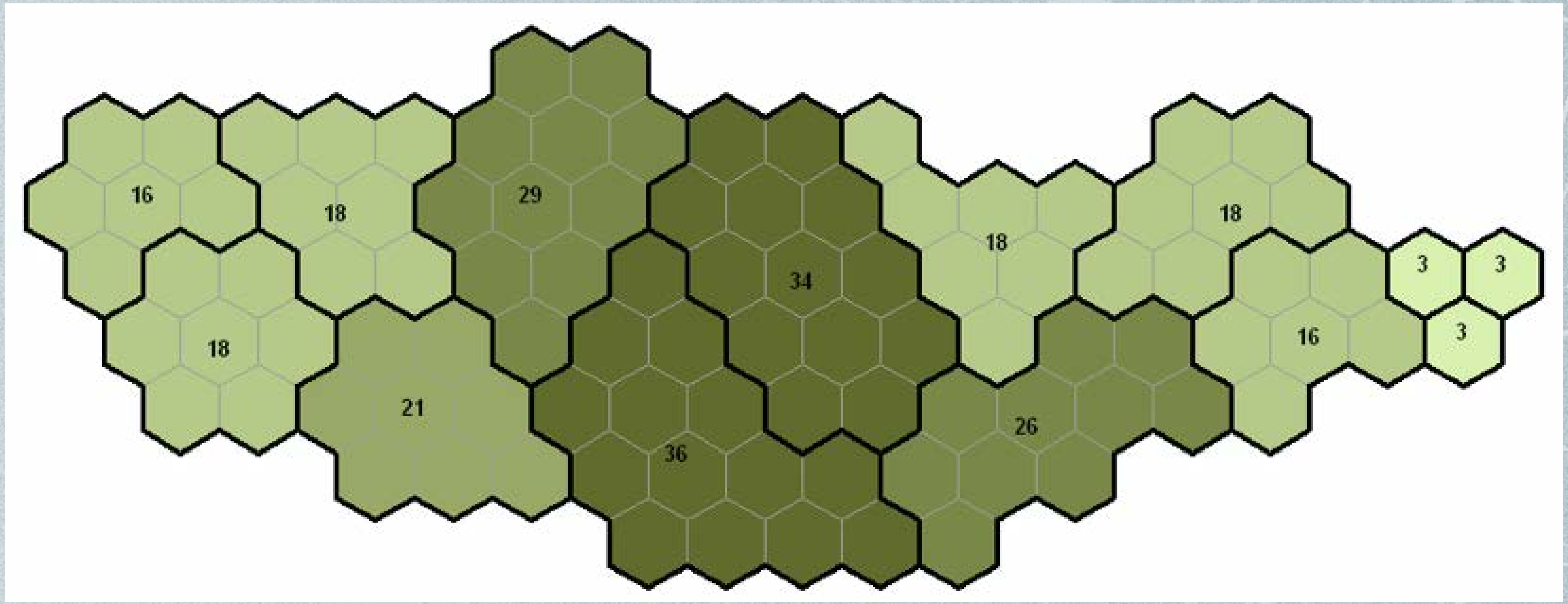
Choropleth map of camels in Mongolia:
where do you think they are concentrated?

Attribute normalization and map type: Extensive vs intensive attributes



Answer: nowhere, because we used a uniform distribution!

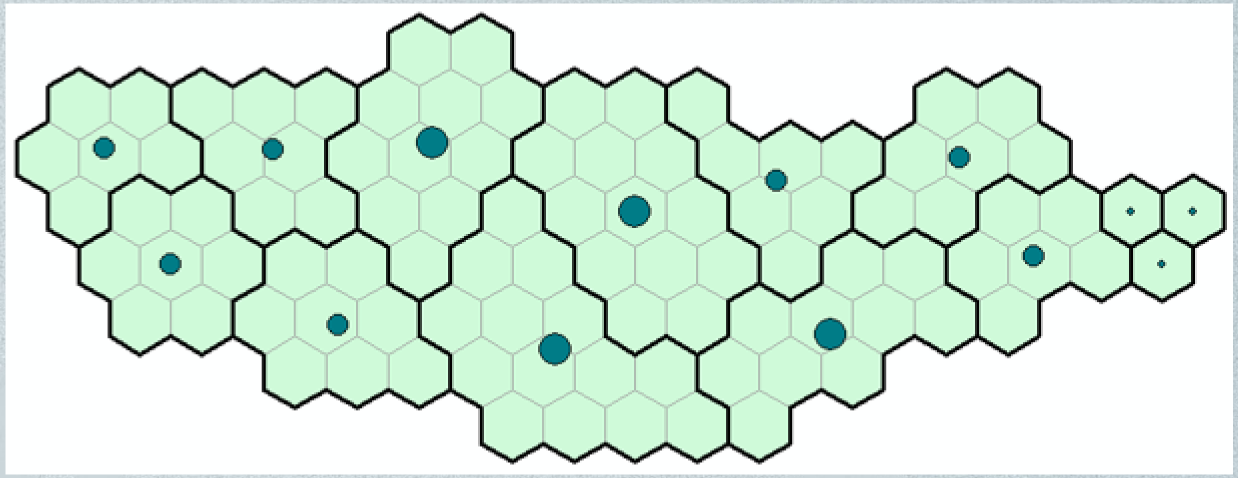
Attribute normalization and map type: Extensive vs intensive attributes

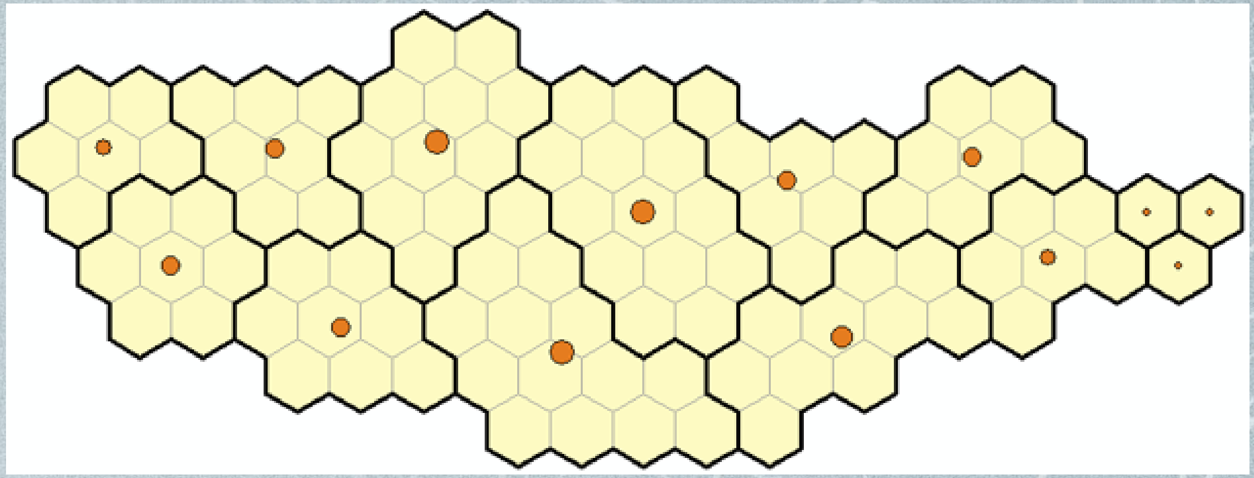


Choropleth map was produced by summing up camels without normalization

Note: Never use non-normalized (extensive) attributes with choropleth maps

Attribute normalization and map type: Extensive vs intensive attributes



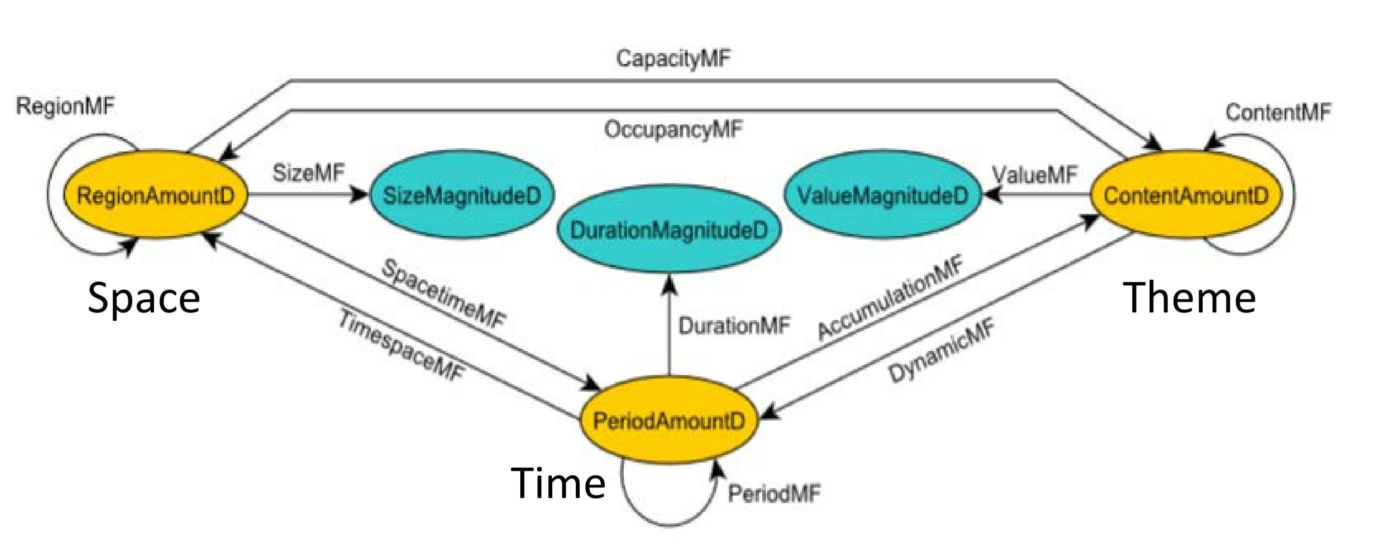


Better use proportional/graduated symbol maps for extensive attributes!
Using Bertin variable: size

Extensive measurement

- can be classified by triangle corners (space, time, theme)
- Capacity: measure theme controlled by space
- Occupancy: measure space controlled by theme
- Accumulation: measure theme controlled by time
- …

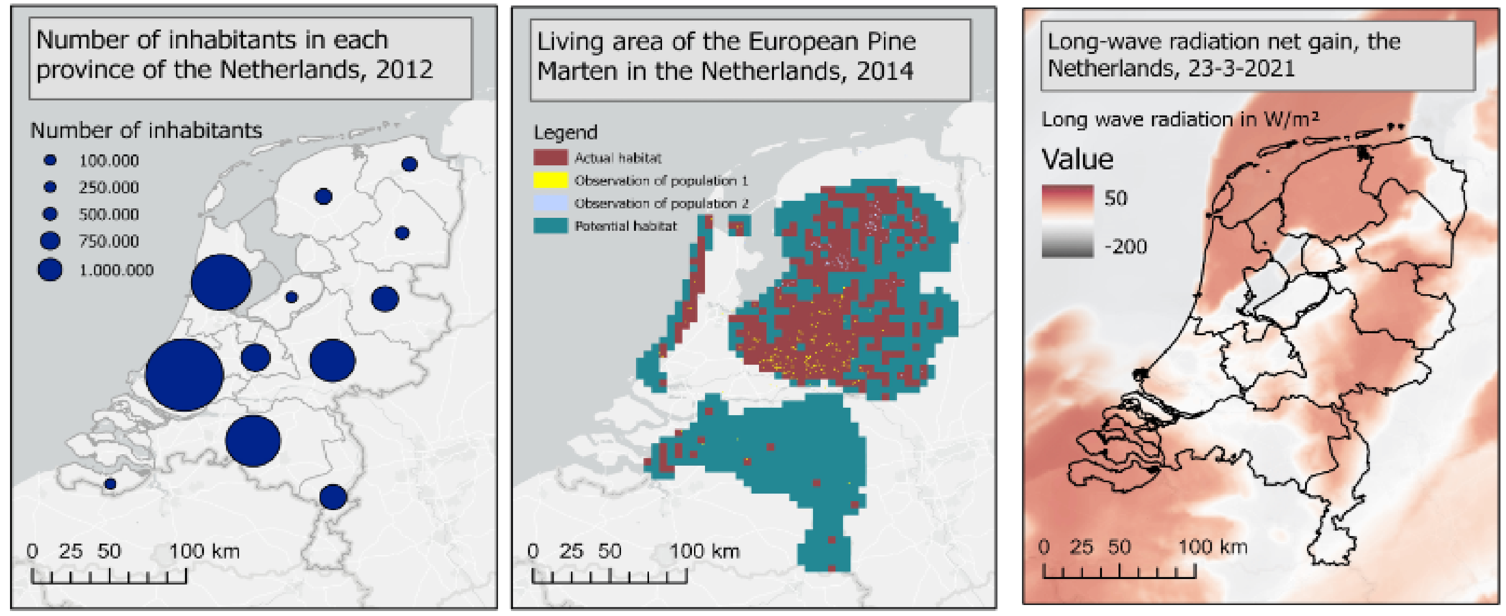
(Top et al: forthcoming)



Capacity

Occupancy

Accumulation



(d) Capacity measurement

(e) Occupancy measurement

(f) Accumulation measurement

Space amount
-> Content amount

Content amount
-> Space amount

Time amount
-> Content amount

Homeomericity

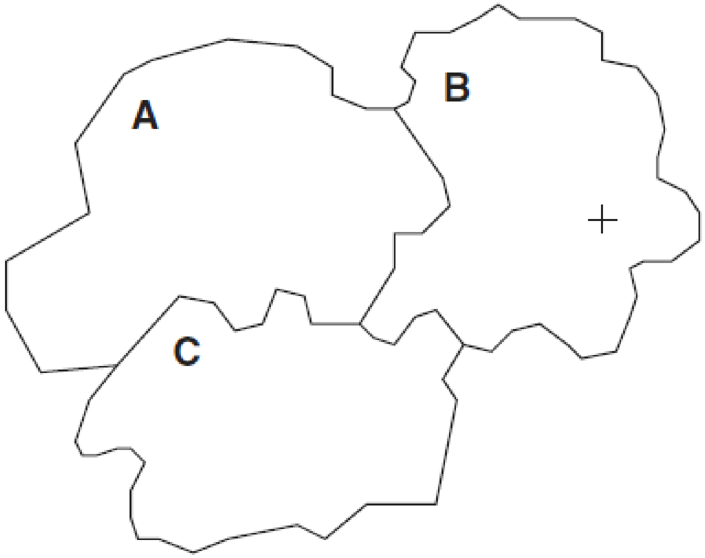
Homeomeric attribute = applies also to parts (Guizzardi 2010)

Example: land cover type

Non-homeomeric: Average elevation, number of inhabitants

⇒Extensive measurements are *never* homeomeric

⇒Intensive: homeomeric *only* in case of homogeneous distribution



	land cover type	average elevation (m)
A	Forest	631
B	Urban	220
C	Water	42
+	Urban	

So…

Why was then the student’s solution invalid?

Validity criteria analysed: Crime

Seems a problem of inference…

Goal schema?
Magnitude of crime controlled by track length on track segment

Requirement?
Magnitude of crime controlled by neighborhood

Explain: … of crime

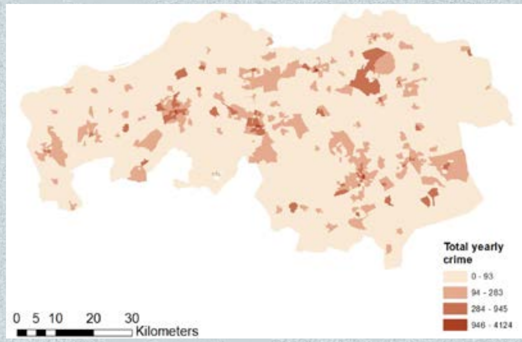
Method?
Simple Overlay

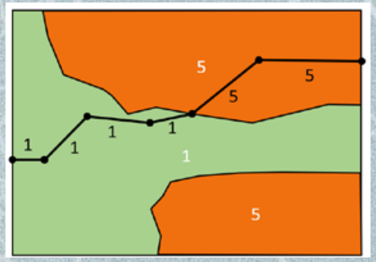
Proxy schema?
Magnitude of crime controlled by neighborhood on track segment in neighborhood

Infer?

Test: … on track segment

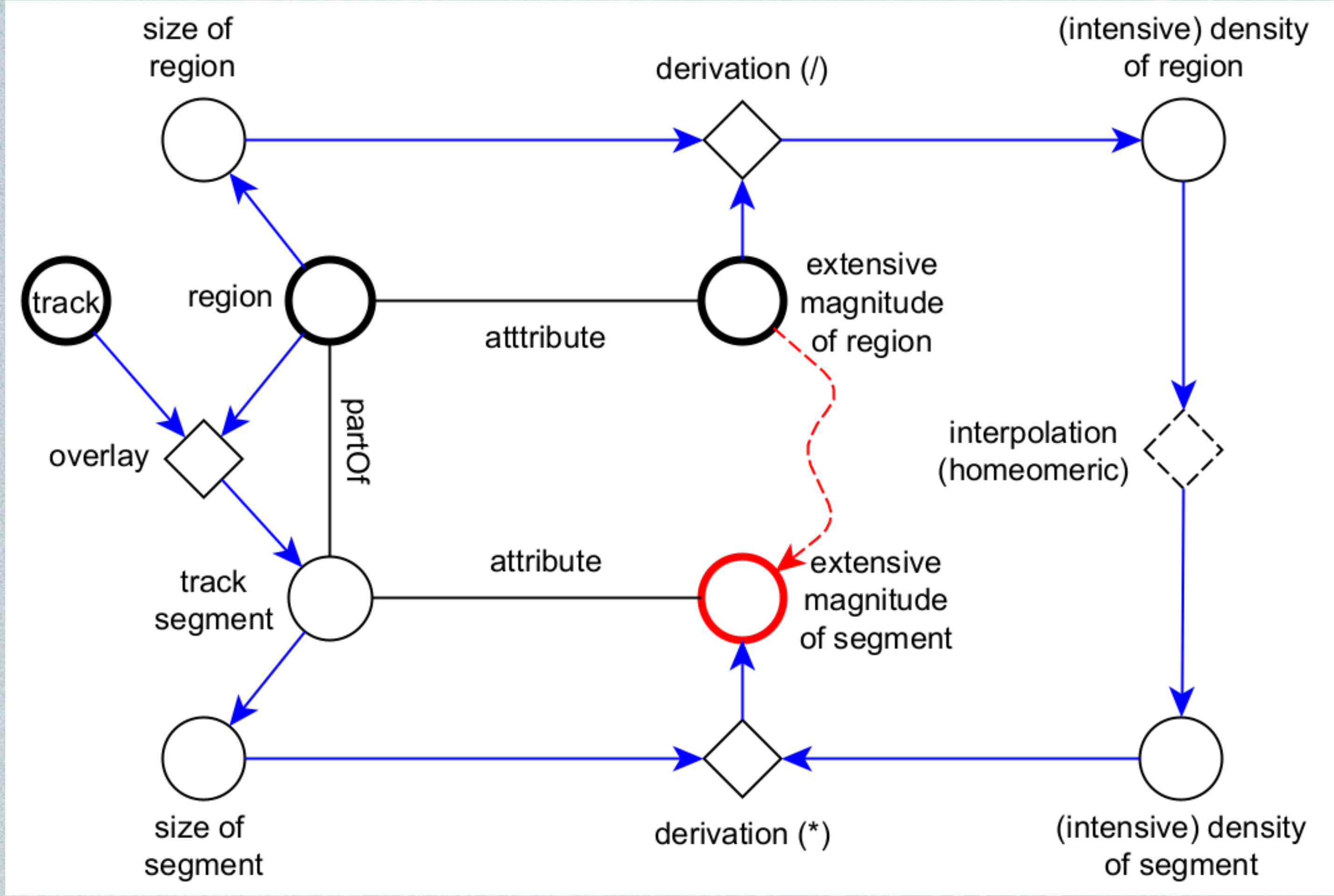
Define: Crime statistics





Crime schema inference invalid because…

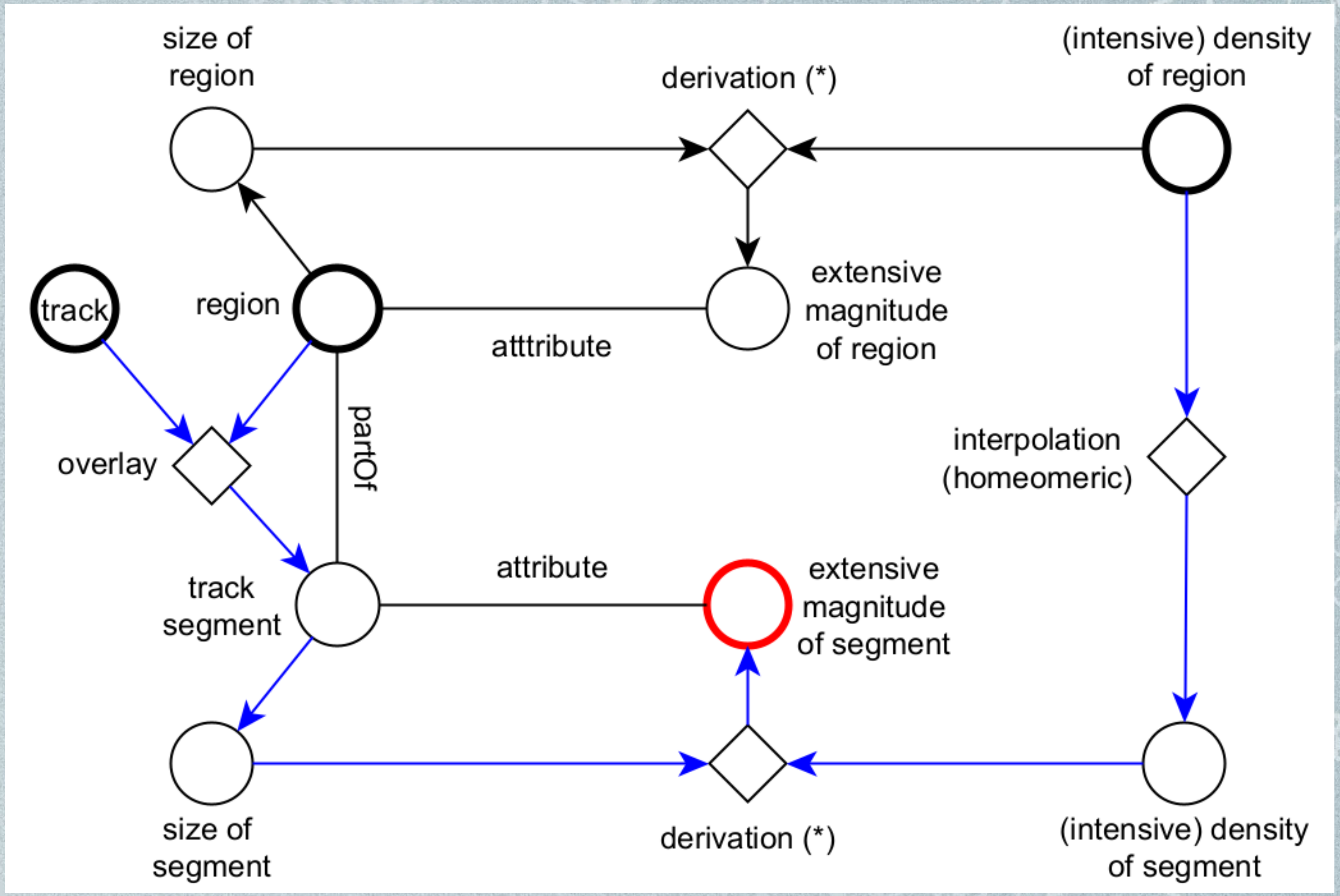
1. Number of crime is **extensive** (capacity measurement)
2. Extensive measurements are **never** homeomeric
3. Thus result is **not** the extensive magnitude **controlled by segment**



Land cover inference valid because…

1. Land cover cost is interpreted as **intensive** (density measurement)
2. Density is assumed homogeneous (and **thus homeomeric)**
3. **Thus** extensive magnitude of segment can be derived (using *)

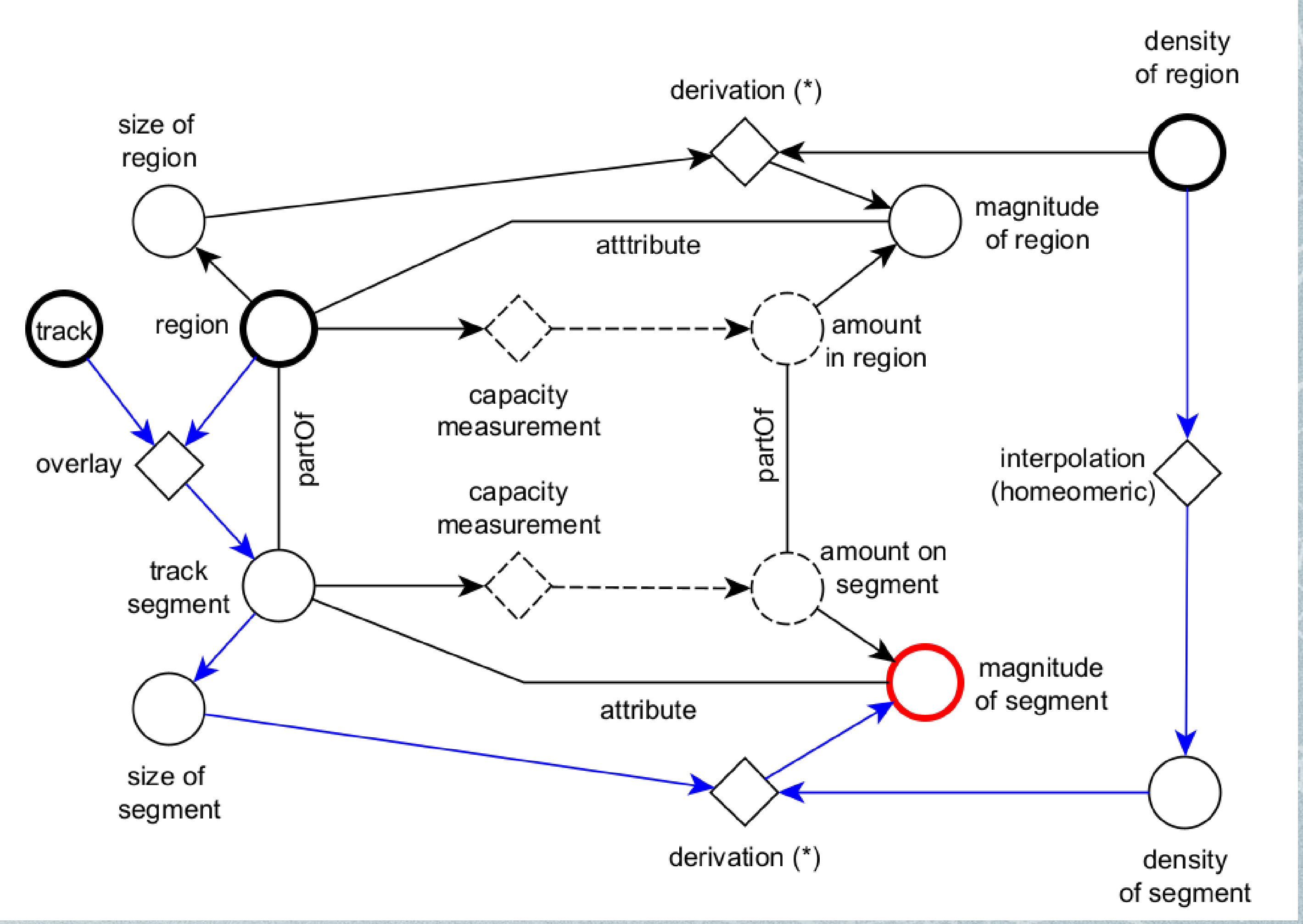
$$\mathrm{cost(t)} = \sum_{n=1}^{|t|} length(s_n) * cost(s_n) | \bigcup_{n=1}^{|t|} s_n = t$$



alternative solution I: measure amount (capacity)

1. Land cover cost is interpreted as **intensive** (density measurement)
2. Thus, there must be an implicit amount
3. This could also be measured directly, with a **capacity measurement**

(count the trees you are passing, sum NDVI index,…)

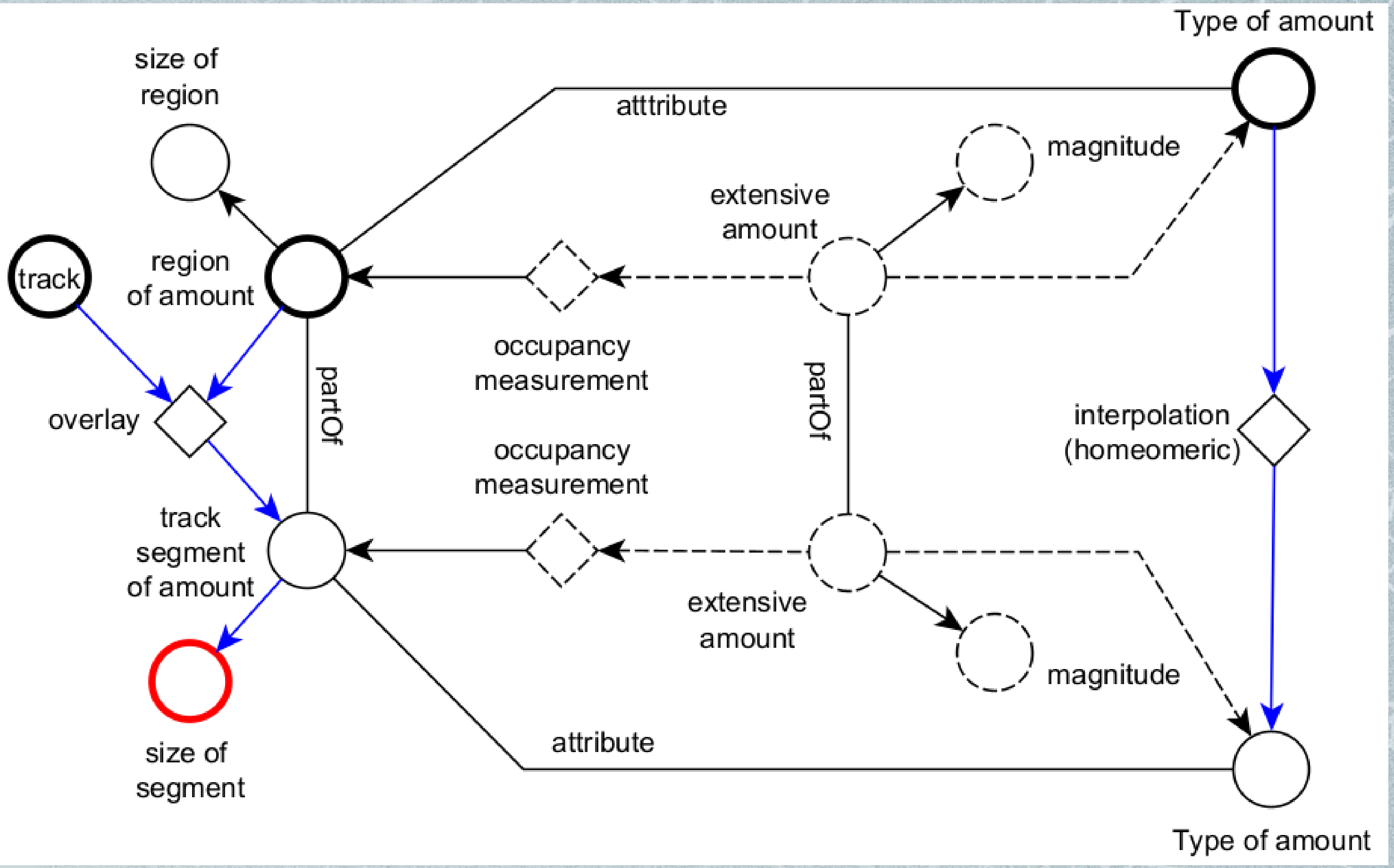


alternative solution II: measure occupancy of amount

Interpret attribute not as a cost, but as an **amount type** (amount *of green*)

Measure the *length of the segment occupied by this type of amount*

(amounts left implicit)



Outlook and conclusion

Core concept data (CCD) ontology

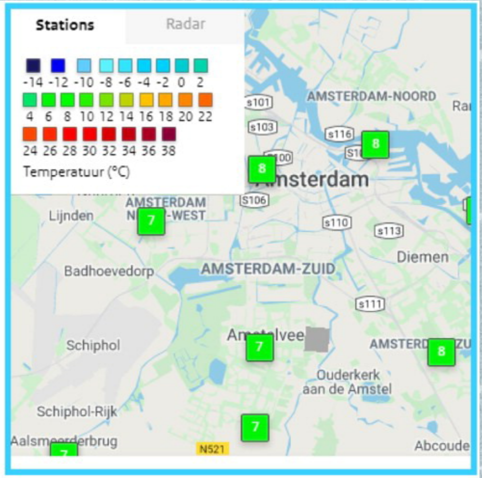
can be used to annotate geodata sources with concepts and data types.

Examples
from the Amsterdam data portal

https://maps.amsterdam.nl/open_geodata/

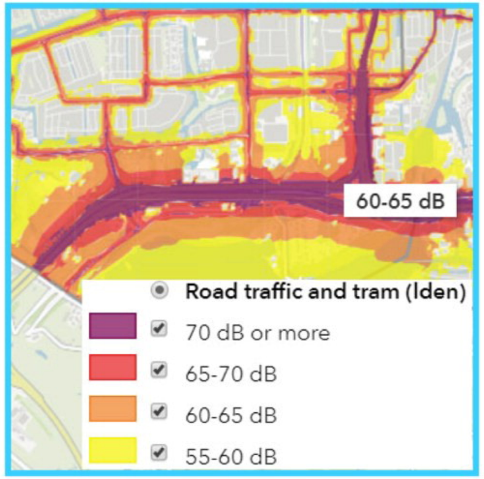
ObjectPoint

EventPoint

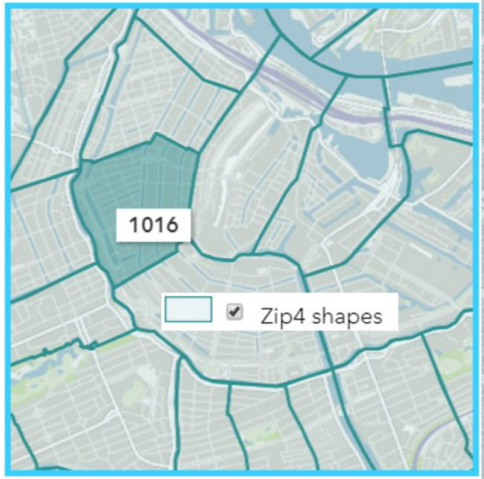


PointMeasures

Coverage



Contour

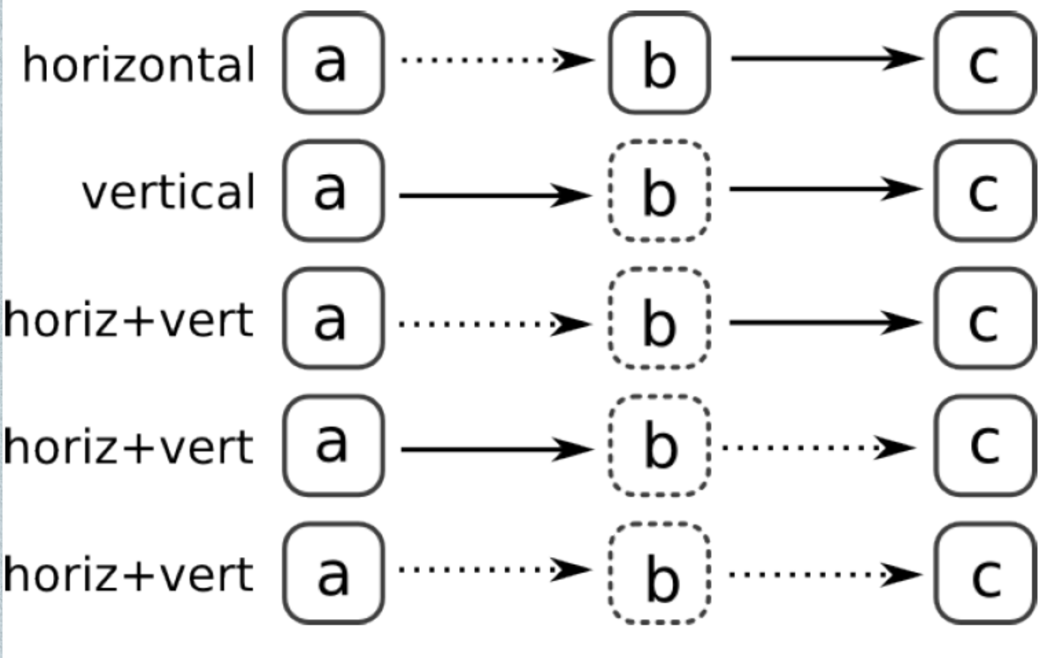


Lattice

Workflow synthesis by loose programming

- A way to 'program loosely' (without specifying each step in a workflow)
- Vertical: by leaving away semantic detail (taxonomy)
- Horizontal: by leaving away process detail (workflow nodes)

(Lamprecht et al. 2010)
(Kruiger et al. 2020)

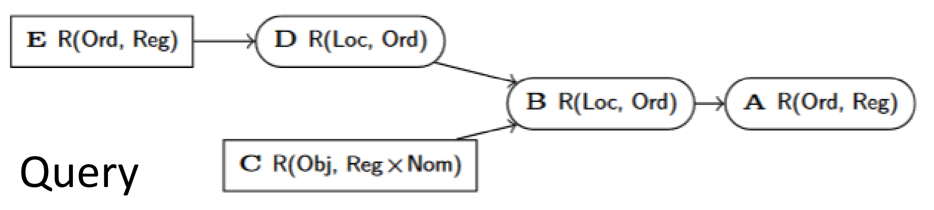


Automated Pipeline Explorer (APE)
https://github.com/sanctuuary/APE

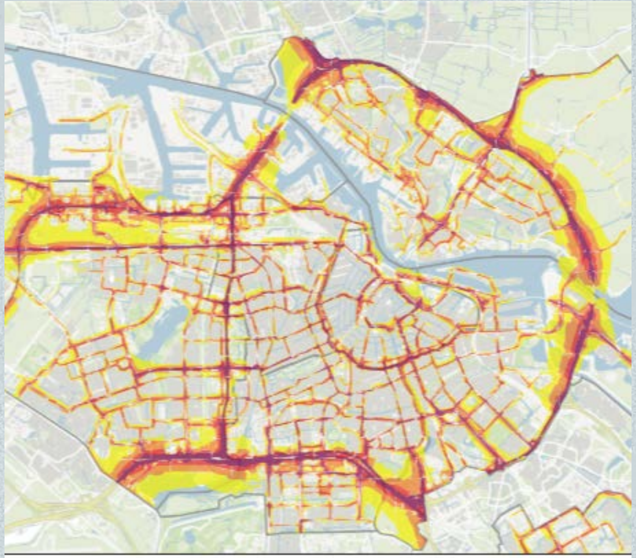
Query GIS workflows with CCT algebra

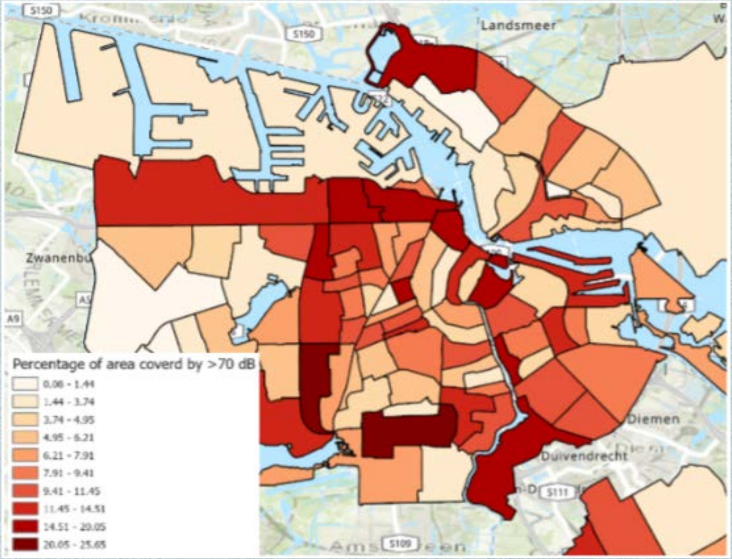
Describes and queries GIS workflows with conceptual transformations

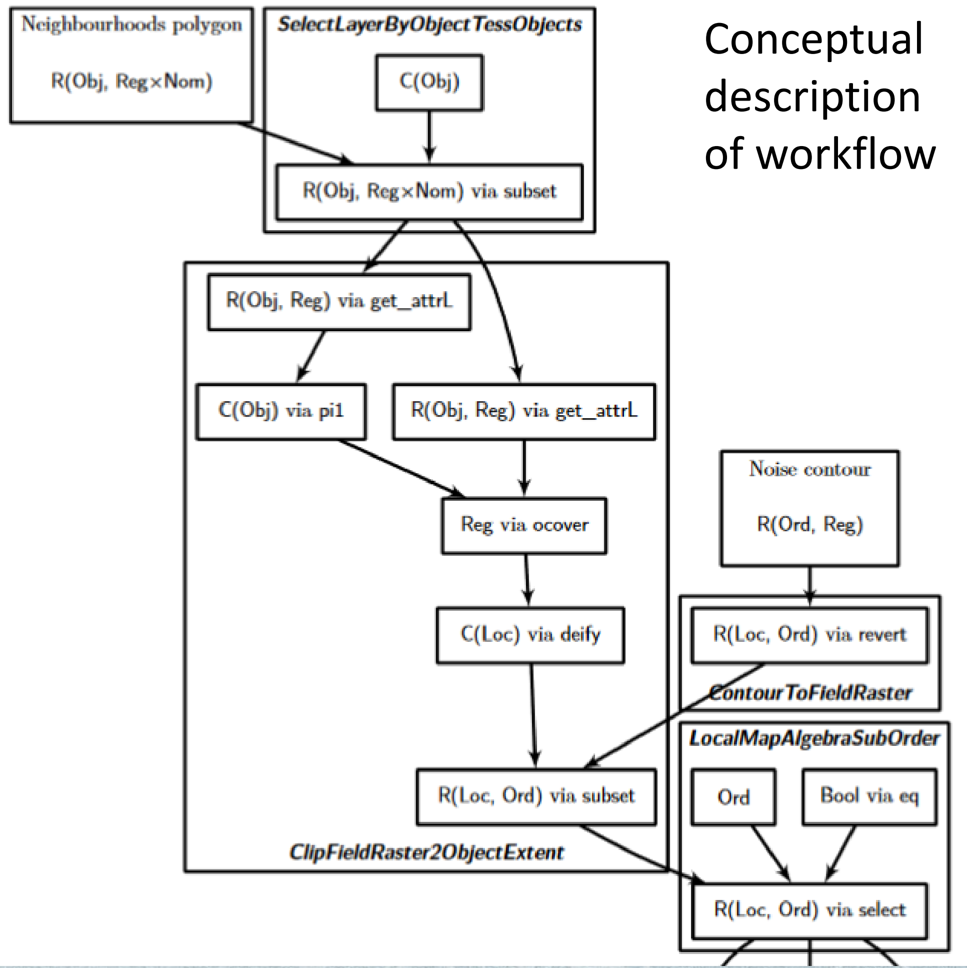
Question: What is the proportion of the area covered by 70 db noise?



Query







Conceptual description of workflow

CCT algebra
https://github.com/quangis/cct

Conclusion

- GIS know-how = valid transformation of geographic information
- Modeling this know-how is required for automating GIS methods
- Goes beyond retrieval (procedural) as well as statistical test/experiment
- Pragmatic aspects of validity:
  purpose, **concepts**, **inference**, **explanation** (valid only w.r.t. purpose)
- The concepts **amount**, **extensivity** and **homeomerocity** can be used to explain why exposure measurement is invalid
- Our work on amount theory is under review. As well as work on conceptual transformation models (CCT) for composing and querying workflows.
- Pragmatic knowledge models can be used for geo-analytical QA

QuAnGIS

Question-based analysis of Geographic Information with Semantic Queries

https://questionbasedanalysis.com/

@2019 -2024: Funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 803498).

Universiteit Utrecht

- Guizzardi, G. (2010). On the Representation of Quantities and their Parts in Conceptual Modeling. In FOIS (pp. 103–116).
- Janich, P. (2001): Logisch-pragmatische Propaedeutik. Velbrueck.
- Janich, P. (2006): Was ist Information? Suhrkamp.
- Kruiger, J.F., Kasalica, V., Meerlo, Rogier, Lamprecht, A.L., Nyamsuren, E. & Scheider, S. (2021). Loose programming of GIS workflows with geo-analytical concepts. Transactions in GIS.
- Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. International Journal of Geographical Information Science, 26(12), 2267-2276.
- Lamprecht, A. L., Naujokat, S., Margaria, T., & Steffen, B. (2010). Synthesis-based loose programming. In 2010 Seventh International Conference on the Quality of Information and Communications Technology
- Steenbergen, N et al. (forthcoming). Algebra of core concept transformations. Procedural meta-data for geographic information
- Top, E., Scheider, S., Nyamsuren, E., Xu, H. , Steenbergen, N. (forthcoming). The Semantics of Extensive Quantities with in Geographical Information
- Xu, H. et al., (forthcoming). A Grammar for Interpreting Geo-analytical Questions as core concept transformations