Answering geo-analytical questions with QuAnGIS

Simon Scheider

Department Human Geography and Spatial Planning, Utrecht University

https://questionbasedanalysis.com/

Universiteit Utrecht

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

QuAnGIS
Question-based analysis of Geographic Information with Semantic Queries

06-09-2023

Overview

1. Motivation
2. Principles
3. Conceptual theory
4. Methodology – From question to transformation request
5. Methodology – Generating abstract workflows for answer matching
6. Prototype
7. Conclusion

Motivation

Motivation

Question-answering (QA)

- People prefer to use natural language questions to interact with computers

- Question-answering has been around for a long time in AI
- Recent boost because of transformer based models and LLM

Motivation

GeoQA

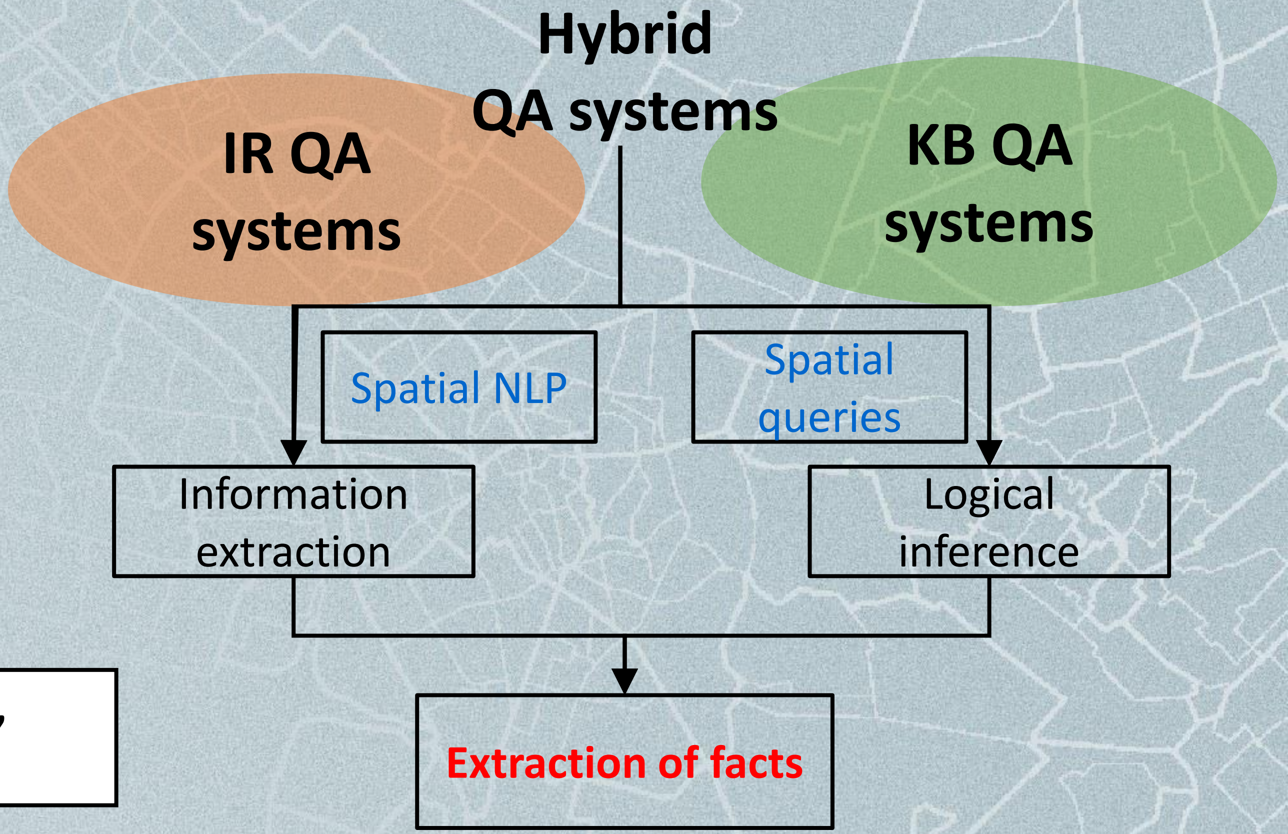
- Punjani, et al. 2018 "Template-based question answering over linked geospatial data.“
- Mai et al. 2019 "Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model.“
- Hamzei 2021. “Place-related question answering: From questions to relevant answers”
- …

“Which bridges cross the river Thames?”

List of crossings of the River Thames

From Wikipedia, the free encyclopedia

The River Thames is the second-longest river in the United Kingdom. It is crossed by over 200 bridges,

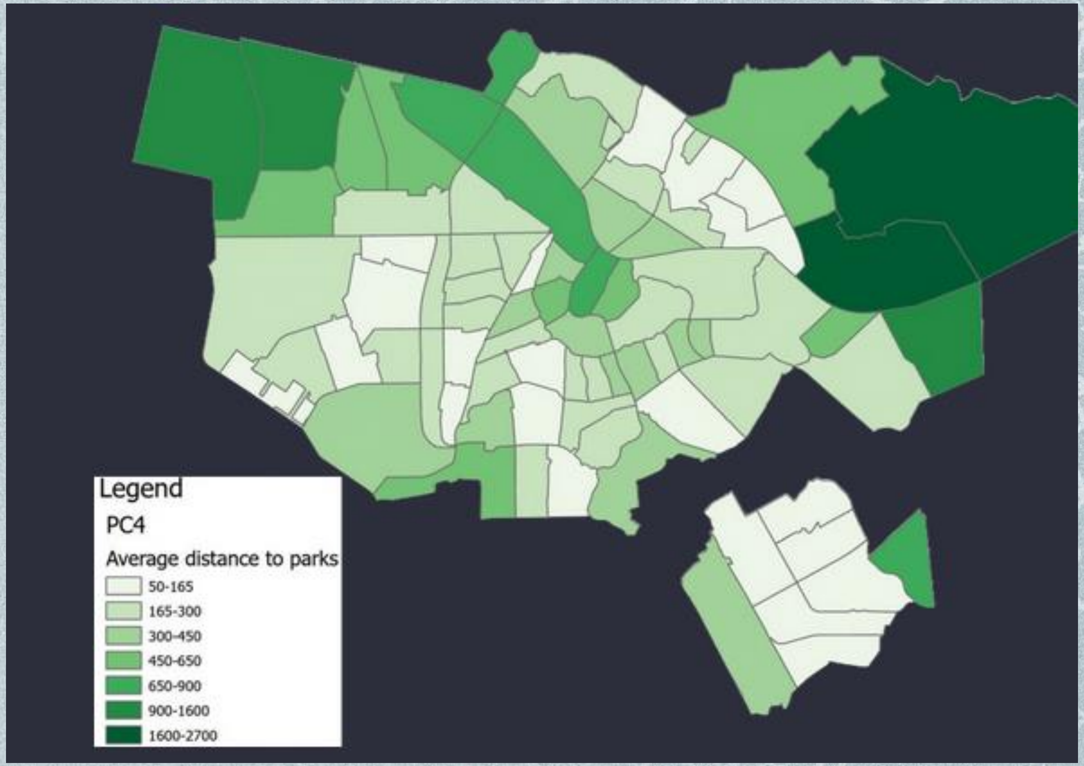


QA with Geographic Information Systems (GIS)

However, Human Geographers may ask:

"What is the average Euclidean distance to parks for each PC area in Amsterdam?"

- Answer is a *map generated for a specific purpose*:
  green: low accessibility
  light: high accessibity



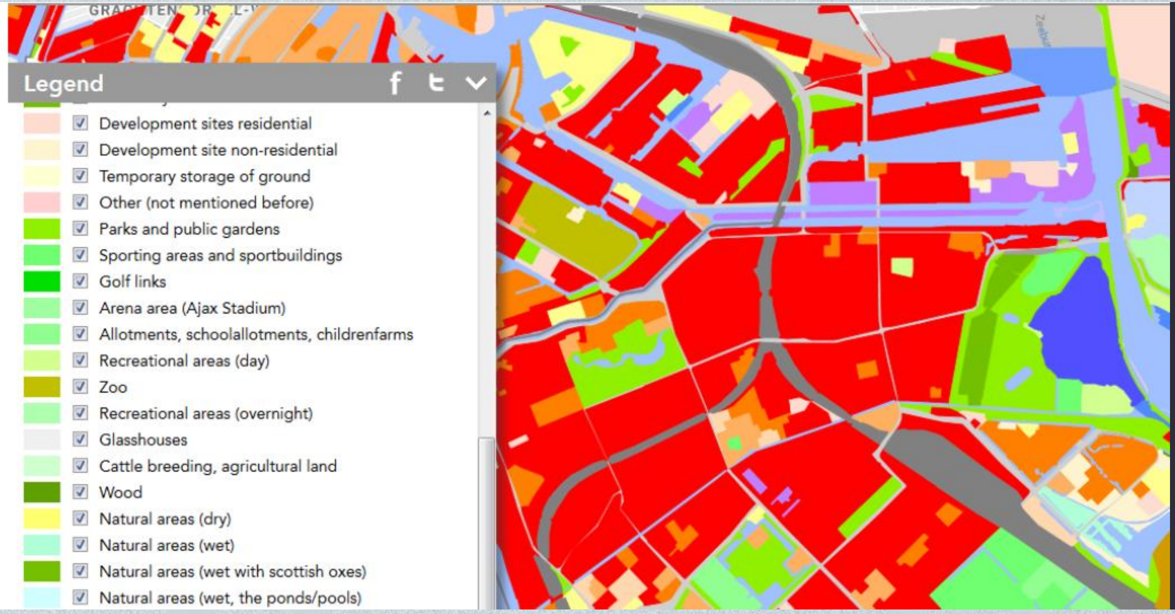
Motivation

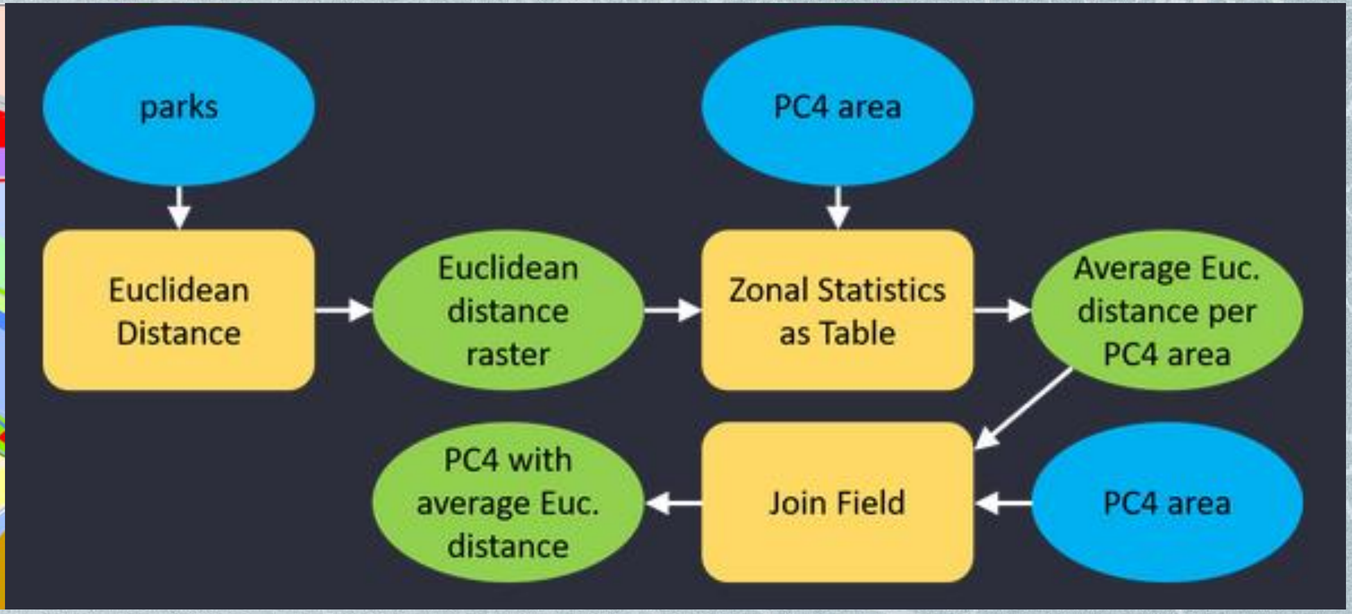
QA with Geographic Information Systems (GIS)

However, Human Geographers may ask:

> "What is the average Euclidean distance to parks for each PC area in Amsterdam?"

Answer requires a *workflow* generating maps from other maps:

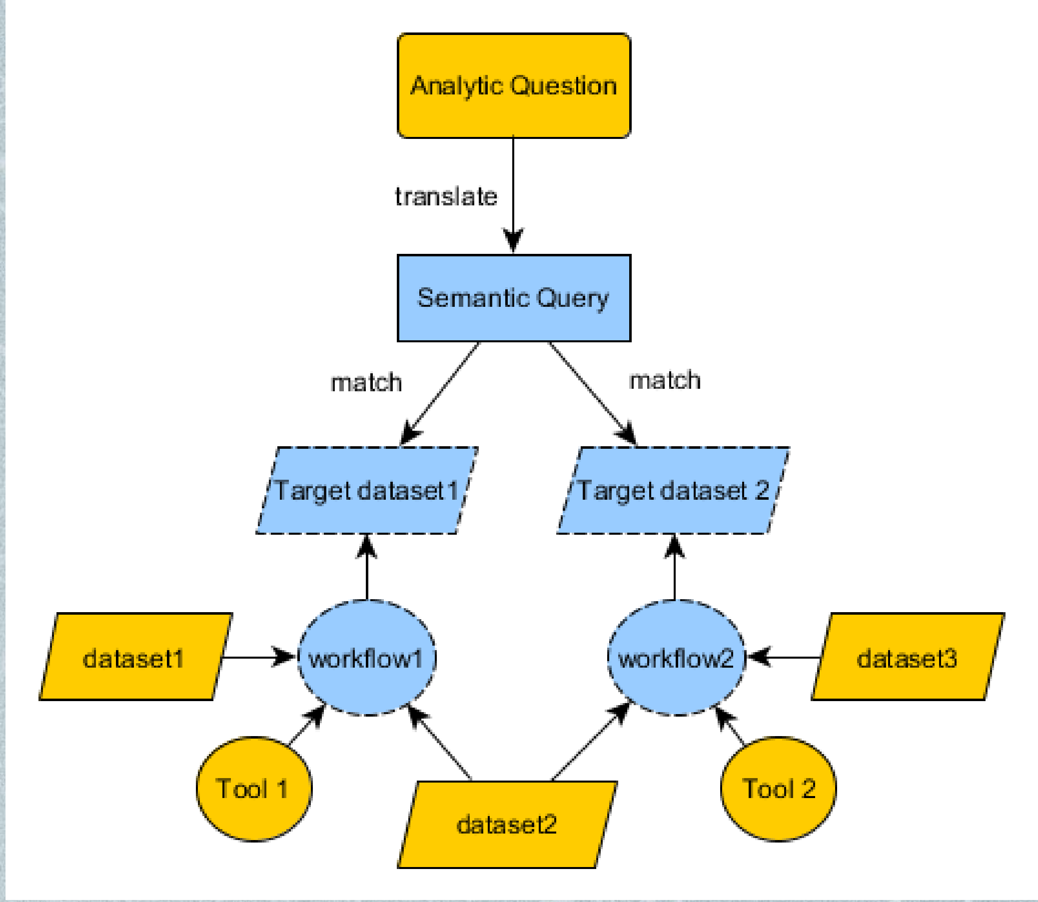




Geo-analytical QA (= indirect QA with GIS)

… ≠ knowledge extraction or question-answer matching:

- Answer is unknown, so *knowledge needs to be generated* (vs. retrieved)
- … requires *creativity (vs.* optimal solution):
  - There is more than one answer workflow
  - There is more than one possible data source
- … requires *procedural knowledge* (vs. declarative knowledge)



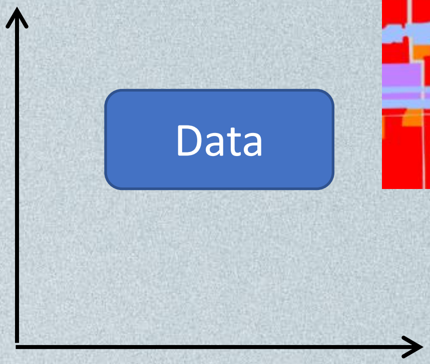
Scheider et al. 2021
"Geo-analytical question-answering with GIS."

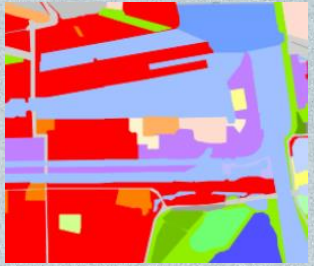
Principles – Procedural knowledge (QuAnGIS in a nutshell)

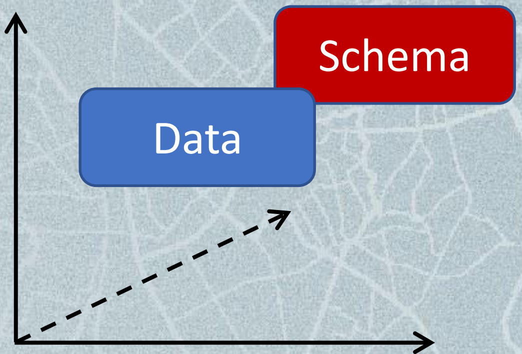
Principles

Understanding data with data schemas

1. *interpret* data into schema
   (orthogonal to data dimensions, so schema is *not* contained in data!)







Dimensions explicit in the data:

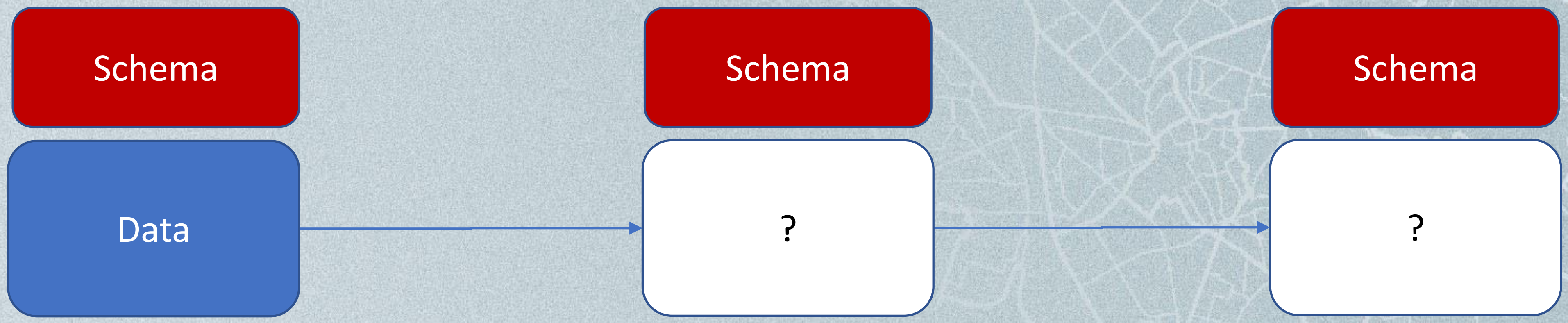
- Polygons
- Landuse classnames

Dimensions *implicit* in the data

- “Object (Park)”
- “Nominal”

Understanding question as schematic transformation request

2. *interpret* question into transformation schema with unknowns

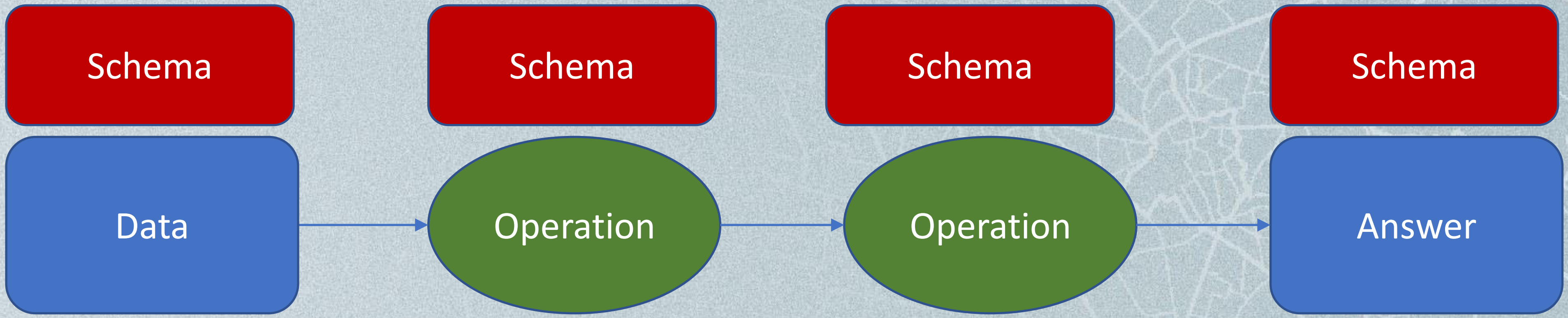


"Average Euclidean distance to objects"

Principles

Transform data into answer

3. *transform*: data into answer (map)



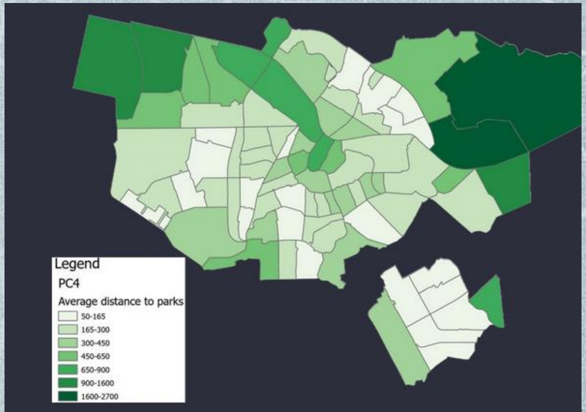
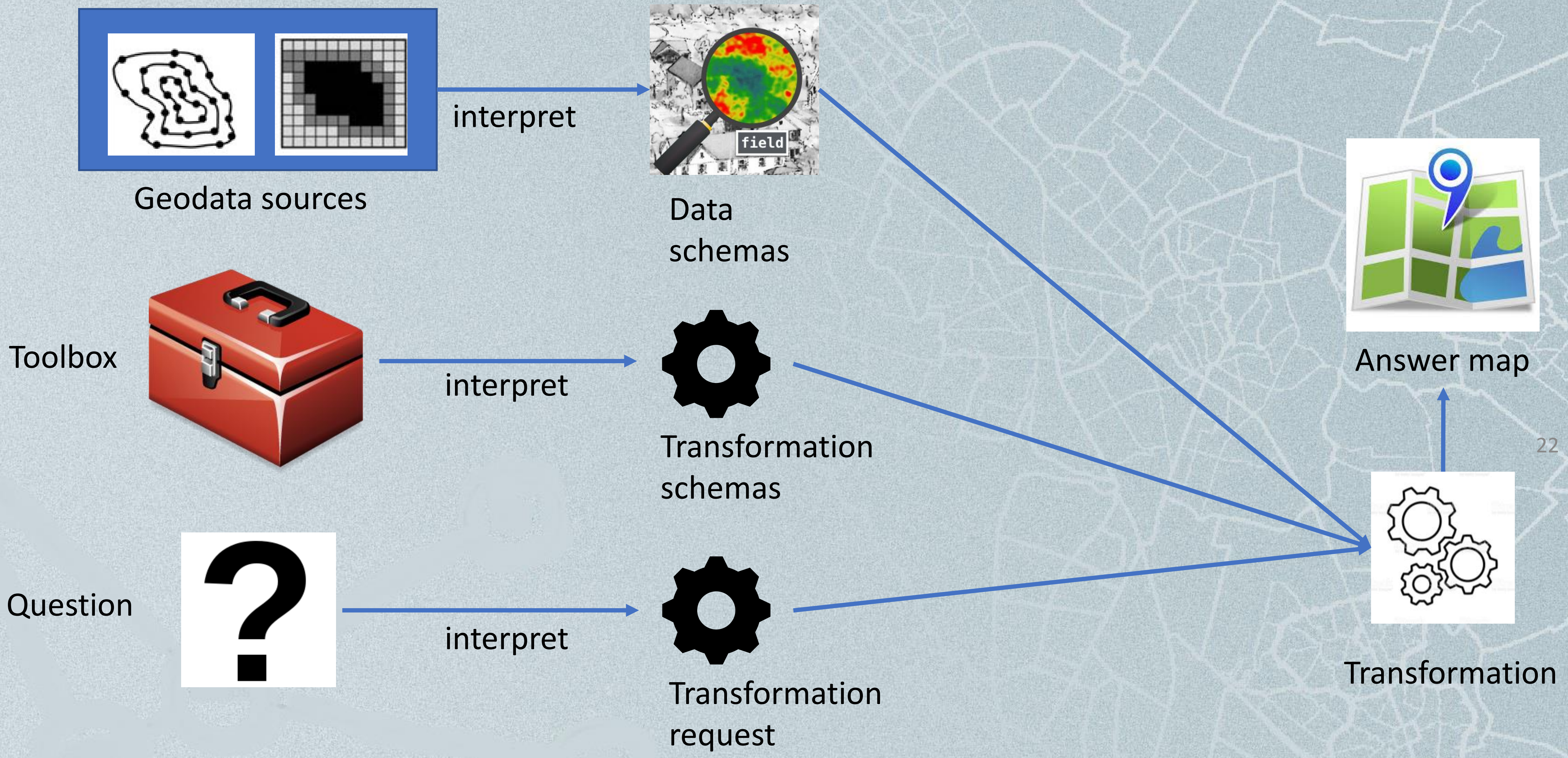


Figure 2. A map produced from the workflow's output.

Principles

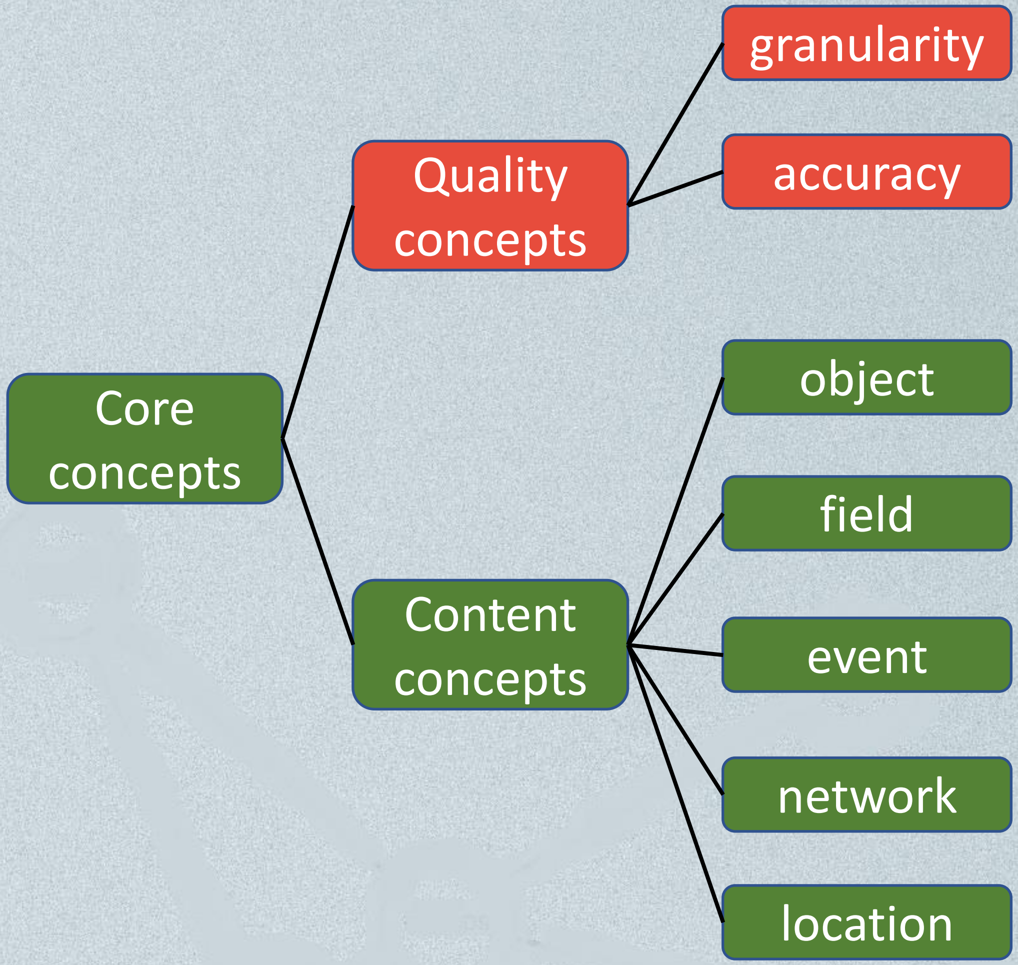
GIS = procedural question-answering system



SCHEMAS – Core Concepts and measurements

Conceptual Theory – Core Concepts

Core Concepts of Spatial Information

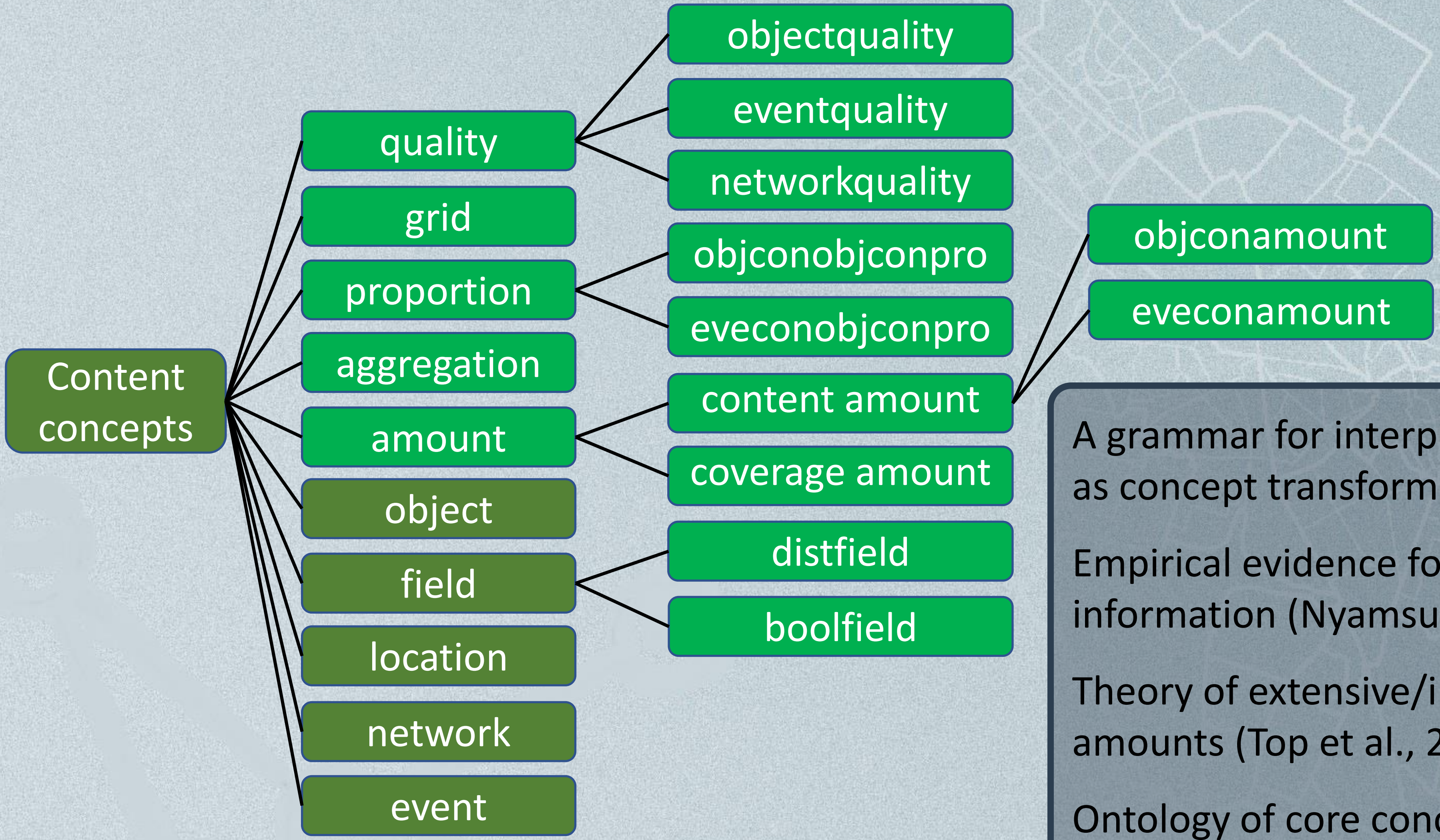


- Core concepts are a **representational model** of spatial information
- Each content concept has a **unique set of spatial properties**
- A spatial entity (data, operation) can **be interpreted** as a particular content concept.

Designing a Language for Spatial Computing (Kuhn & Ballatore, 2015)

Conceptual Theory – Core Concepts

Extended core concepts for GeoQA (+ measurement levels/amounts)



A grammar for interpreting geo-analytical questions as concept transformations (Xu et al., 2023)

Empirical evidence for concepts of spatial information (Nyamsuren et al, 2022)

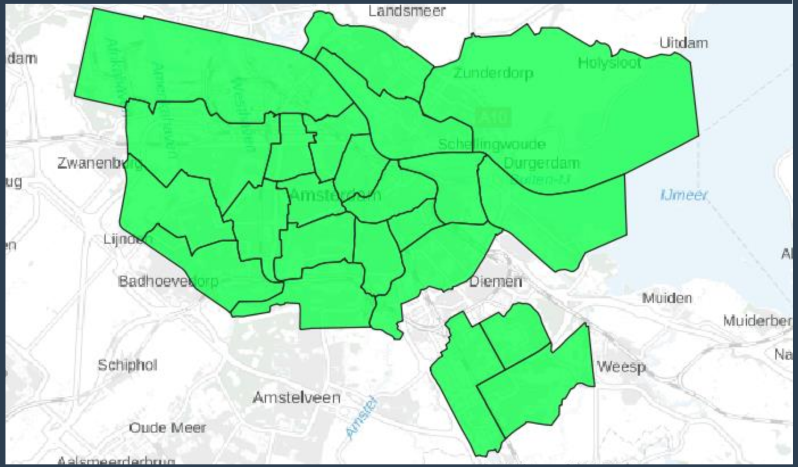
Theory of extensive/intensive measurement of amounts (Top et al., 2022)

Ontology of core concept data types for answering geo-analytical questions (Scheider et al., 2020)

Conceptual Theory – Core Concepts

Core concept examples

Object (Nominal)

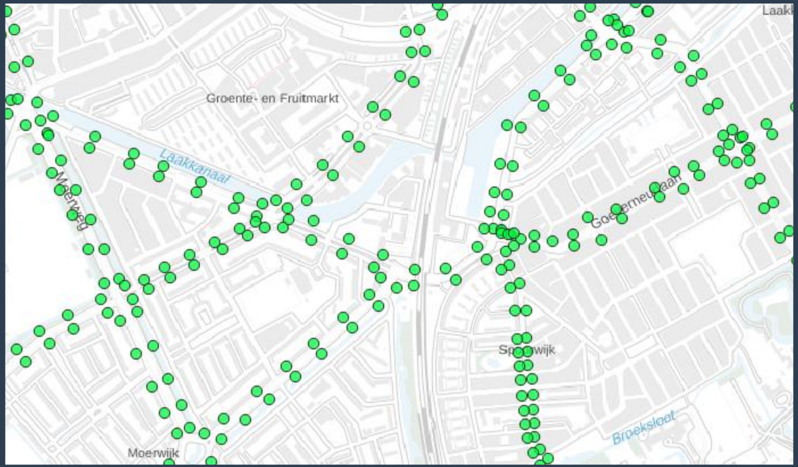


22 management areas of Amsterdam (e.g., Bijlmer-Oost, Westerpark, Slotervaart)

Field (Ratio)

Buffers showing distance ranges from a road

Field (Ratio)



Micrograms of NO2 per cubic meter (e.g., 31.853, 38.29, 44.339)

Conceptual Theory – Core Concepts

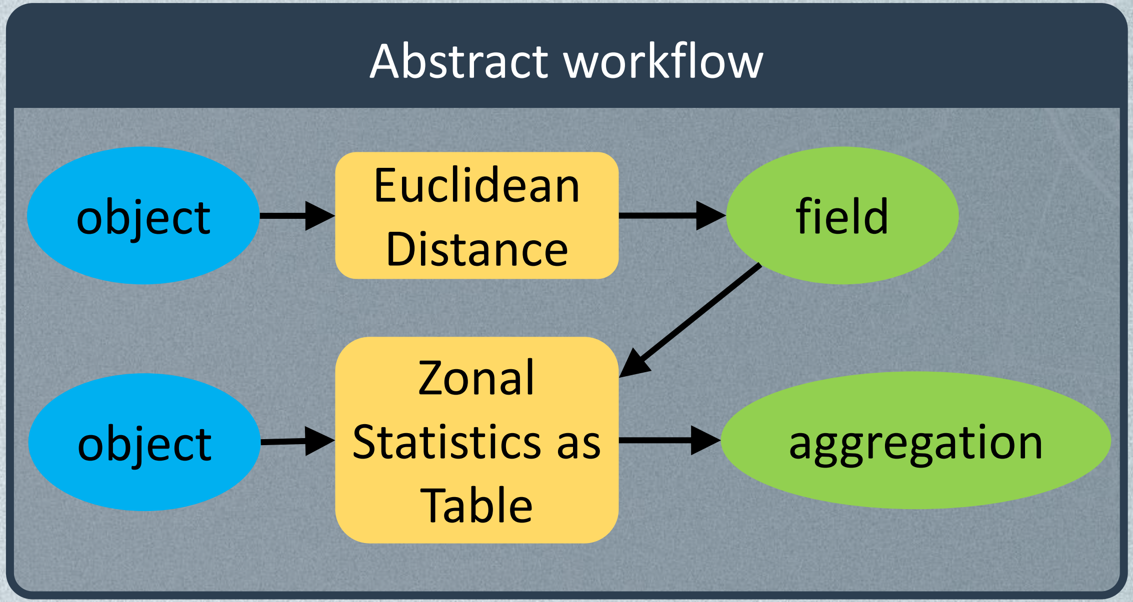
Using the core concepts to annotate questions

What is the average*(aggregation, ratio)* Euclidean distance(field,ratio) to parks(object) for each PC4 area(object) in Amsterdam(object)

What is the density(proportion,ratio) of people(object) older than 65(object quality) for each PC4 area(object) in Amsterdam(object)

Conceptual Theory– Core Concept Transformation

Tool signatures and workflows specify a concept transformation



- Core concepts define in abstract terms the **input and output signatures** of GIS operations.
- A GIS workflow **transforms** one core concept into another core concept.

METHODOLOGY – From questions to transformation requests

Functional roles

Functional roles are

- based on Sinton's spatial measurement framework (Sinton, 1978)
- define roles that spatial information can take within the analytic process specified by the questions

Functional role	Description	Cardinality
Measure	provides the question **goal** and **starting inputs** for GIS workflows	1
Condition	defines the **criteria** (of relevance) for estimating the measure	0 or more
Support	provides the **spatial control** on the measure	0 or 1
Extent	defines the **spatial boundary** for the other functional roles	1
Etc.		

A grammar for interpreting geo-analytical questions as concept transformations (Xu et al., 2023)

METHODOLOGY– Functional roles

Functional roles: examples

defines the **spatial boundary** of the entire process

What is the
average Euclidean distance to parks *(measure)*
for each PC4 area (support)
in Amsterdam (extent)

defines the spatial span of the measure

defines features to identify the relevant measure

What is the density of people *(measure)* **older than 65** *(condition)* **for each PC4 area** (support) **in Amsterdam** (extent)

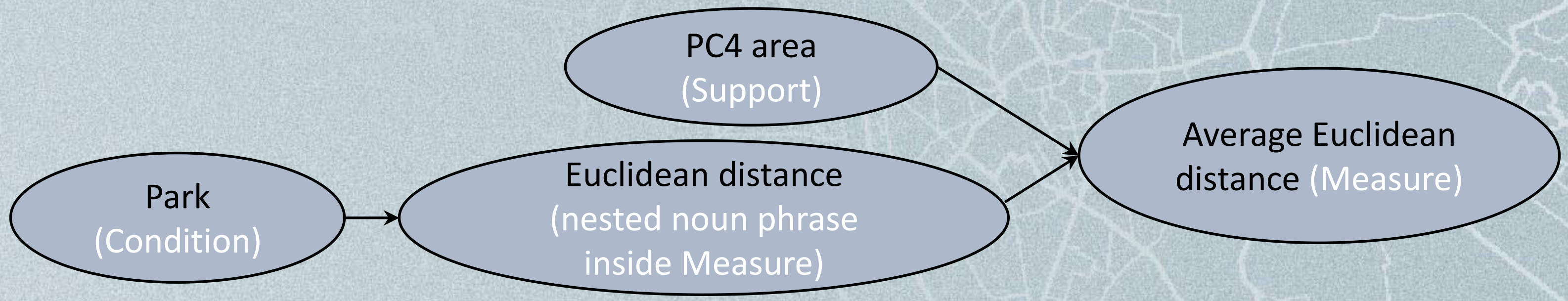
METHODOLOGY– Extracting concept transformations

Functional grammar and NLP

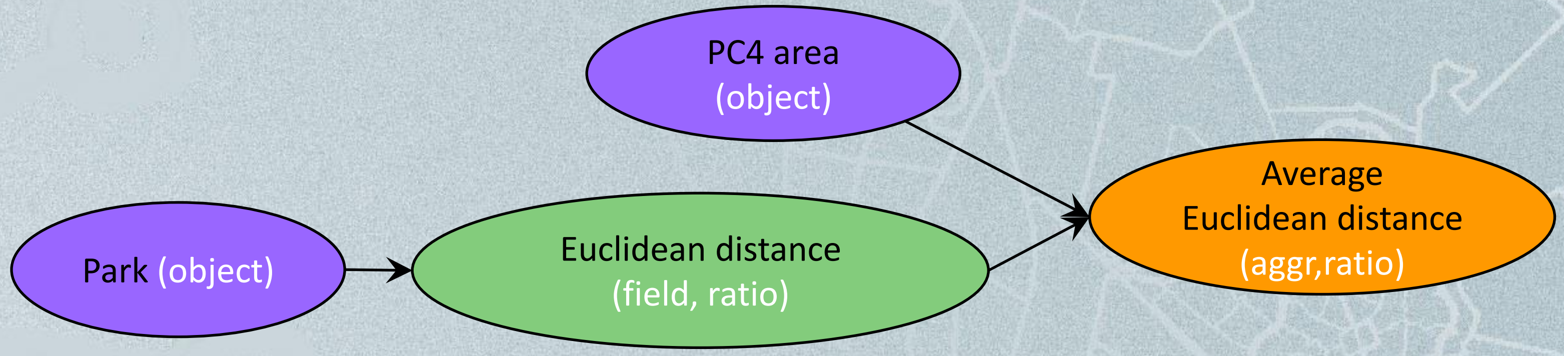
1. Identify concepts with NLP and functional roles with grammar

What is the [average *(aggr,ratio)* Euclidean distance *(field,ratio)*] *(Measure)*
[to parks *(object)*] *(Condition)* [for each PC4 area *(object)*] *(Support)*
[in Amsterdam *(object)*] *(Extent)*?

2. Establish generation order based on roles



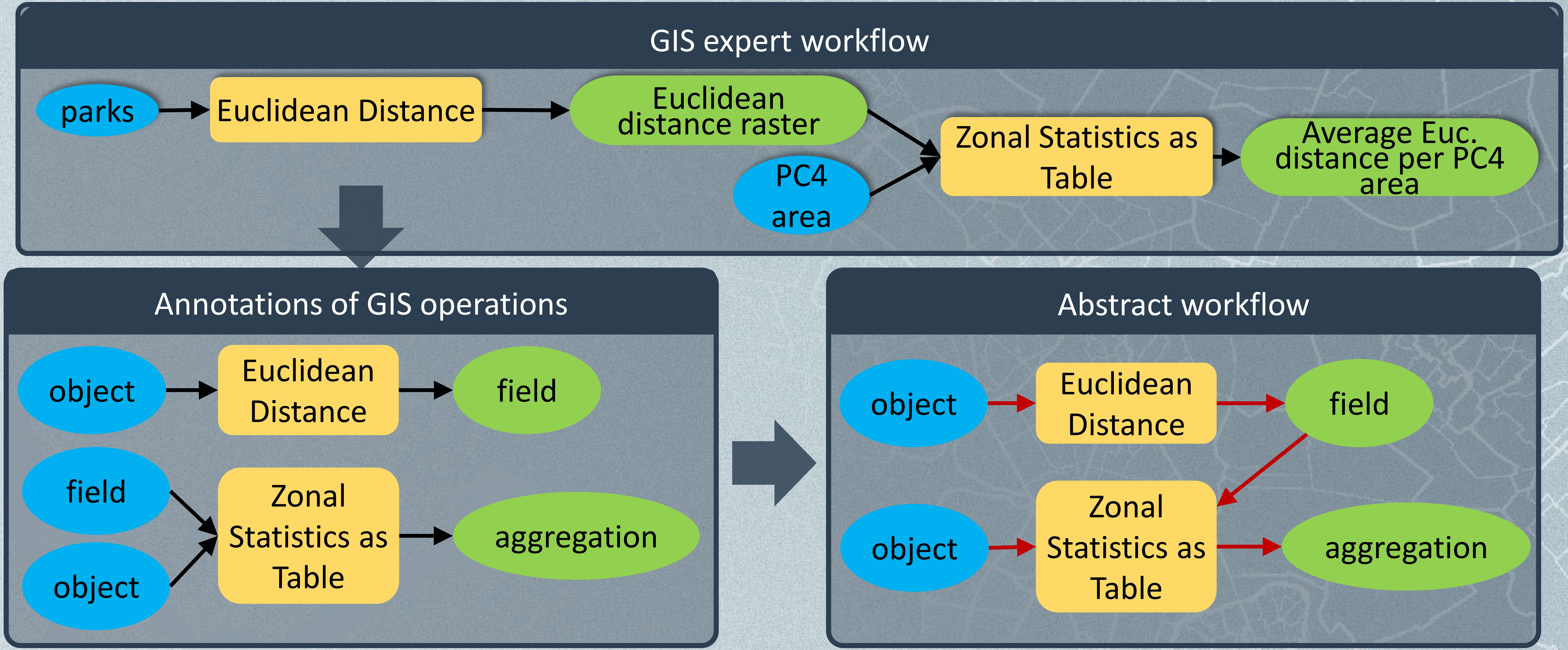
3. Order core concepts



METHODOLOGY – Generating abstract workflows for answer matching

Methodology – Generating abstract workflows

GIS expert workflow => operation annotations => abstract workflows



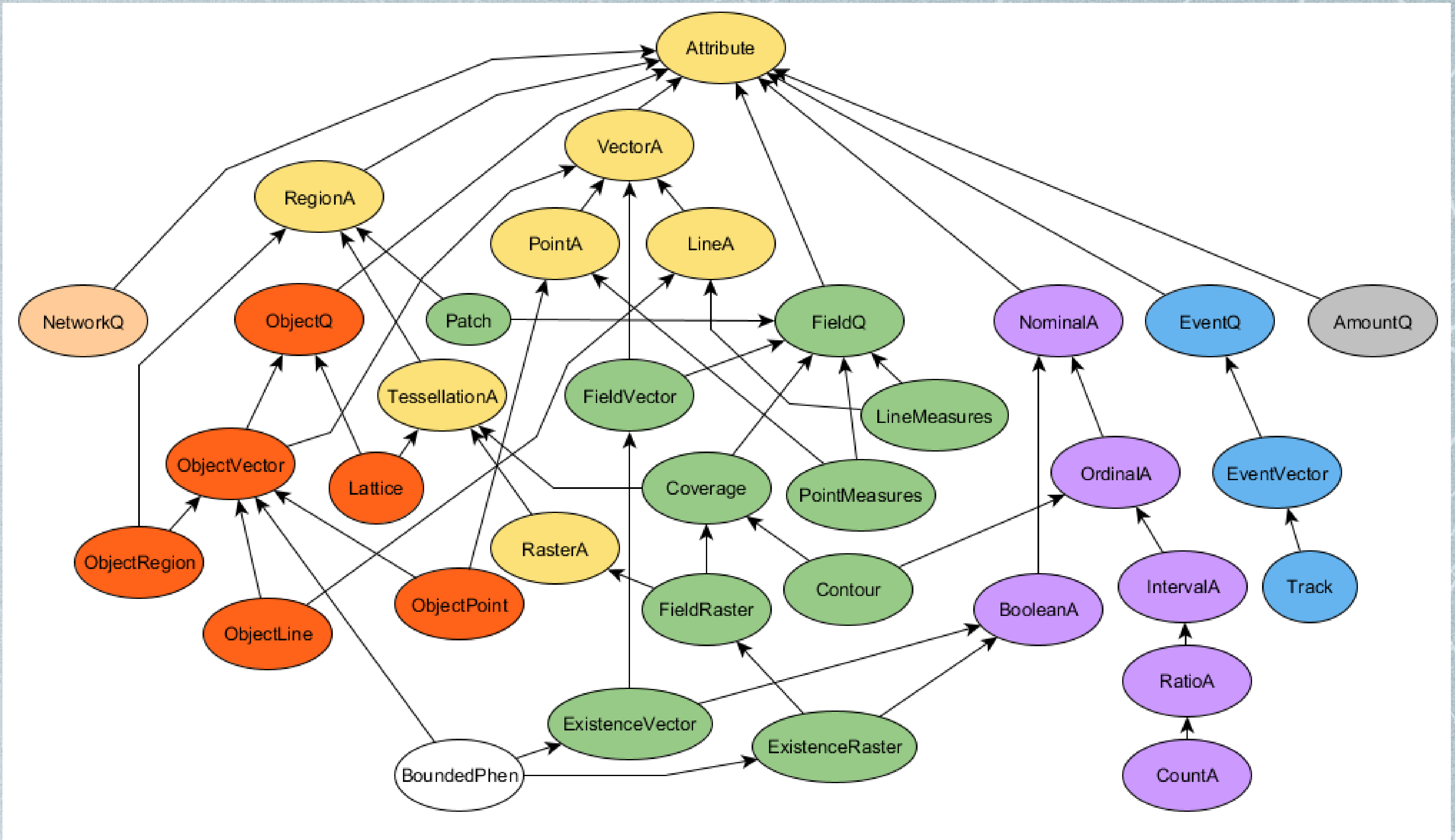
By matching abstract input and output signatures, we can **automatically generate technically valid abstract workflows (CSP problem).**

Methodology – Generating abstract workflows

Core Concept Datatype Ontology

The ontology combines three dimensions:

1. Core concepts field, object, network, event
2. Layer types
3. Levels of measurement



(Scheider et al. 2020a: Ontology of core concept data types for answering geo-analytical questions)

Methodology – Generating abstract workflows

Example tool annotation

ZonalStatisticsMeanRatio

Tool URI

https://quangis.github.io/tool/abstract#ZonalStatisticsMeanRatio

Signature

input2	ObjectQ	An attribute that represents the quality of an object (core concept)
	PlainVectorRegionA	An attribute of a vector region layer that is not a tessellation
input1	FieldRaster	Field rasters are raster representations of continuous fields. For example, raster layers representing a terrain height (DTM).
	RatioA	Attribute on ratio scale level
output	IRA	Intensive region attribute. An attribute that is independent from the size of its support region.
	ObjectQ	An attribute that represents the quality of an object (core concept)
	PlainVectorRegionA	An attribute of a vector region layer that is not a tessellation

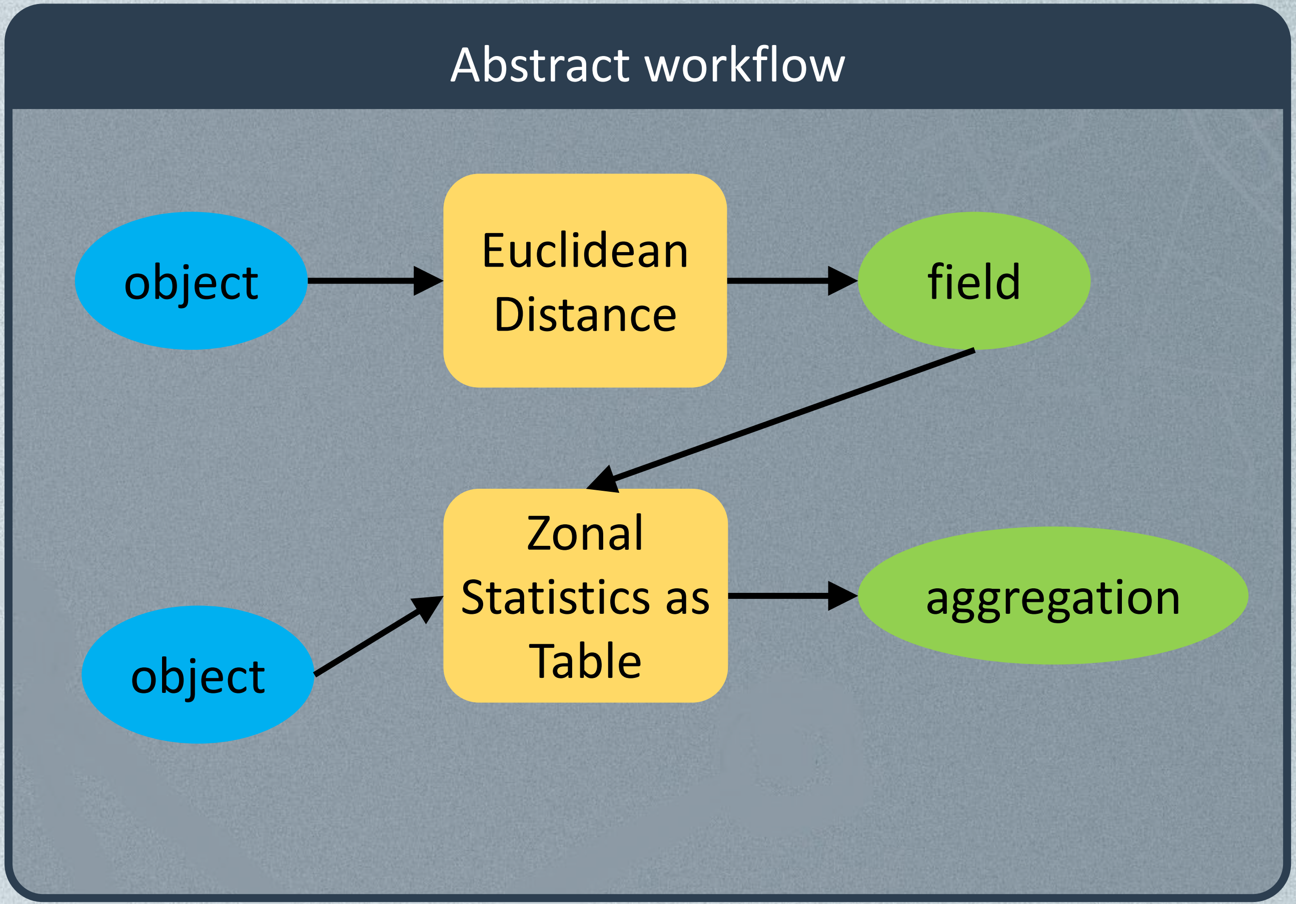
Summary

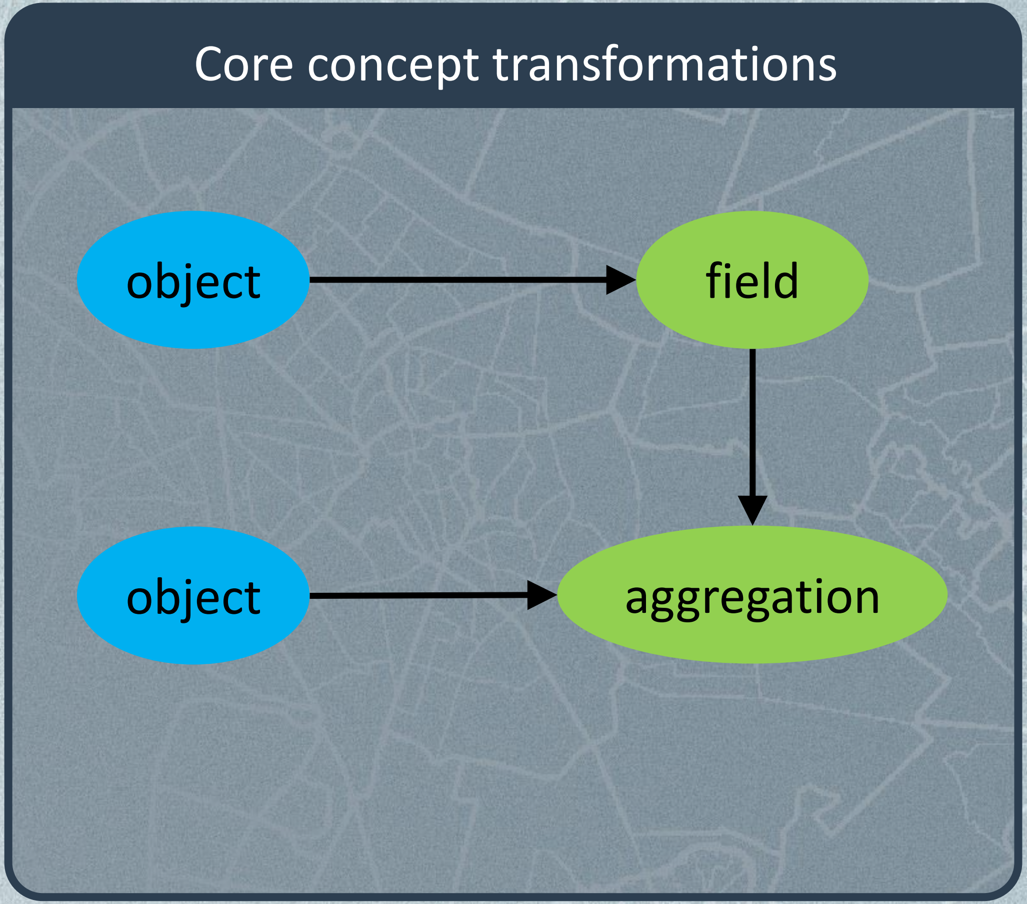
Calculates the average value of a field raster within the zones given by object regions. Outcome is intensive.

https://131.211.60.19/docs-tool for the full list of tool annotations

Methodology – Core Concept Transformation (CCT)

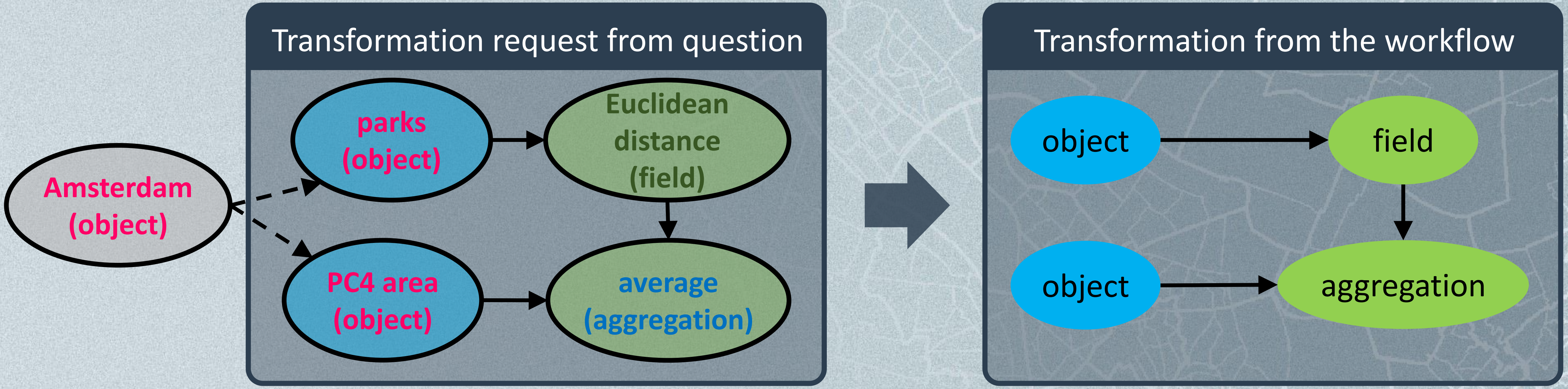
Abstract workflow to concept transformation graph





Methodology – Core Concept Transformation (CCT)

Matching transformations



Transforge library (Steenbergen et al. 2022)
https://github.com/quangis/transforge

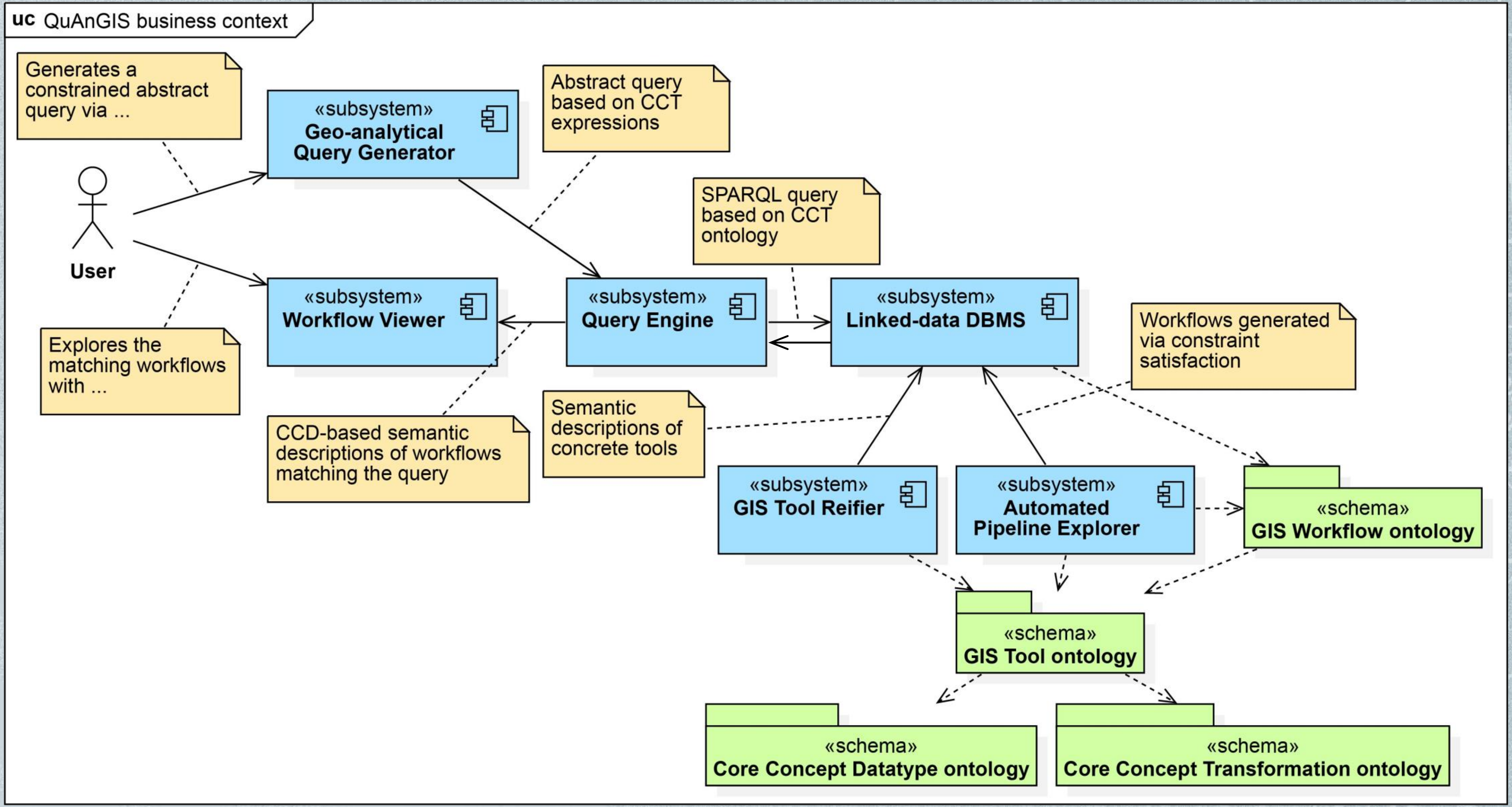
- Infers types over abstract workflow and represents it as linked data
- Expresses transformation request in SPARQL
- Matches request with types in workflow

PROTOTYPE

30

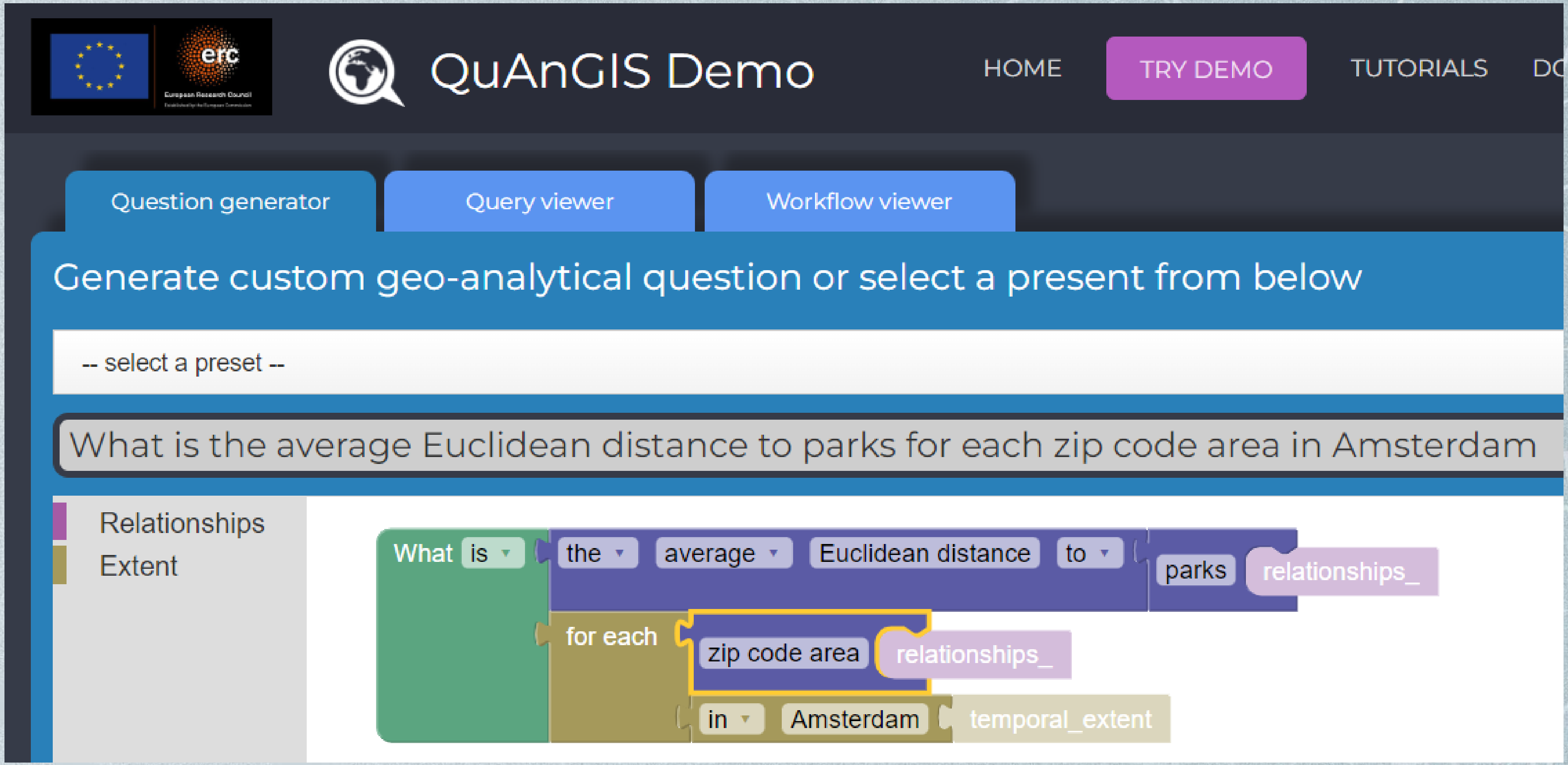
Prototype Architecture

Business context



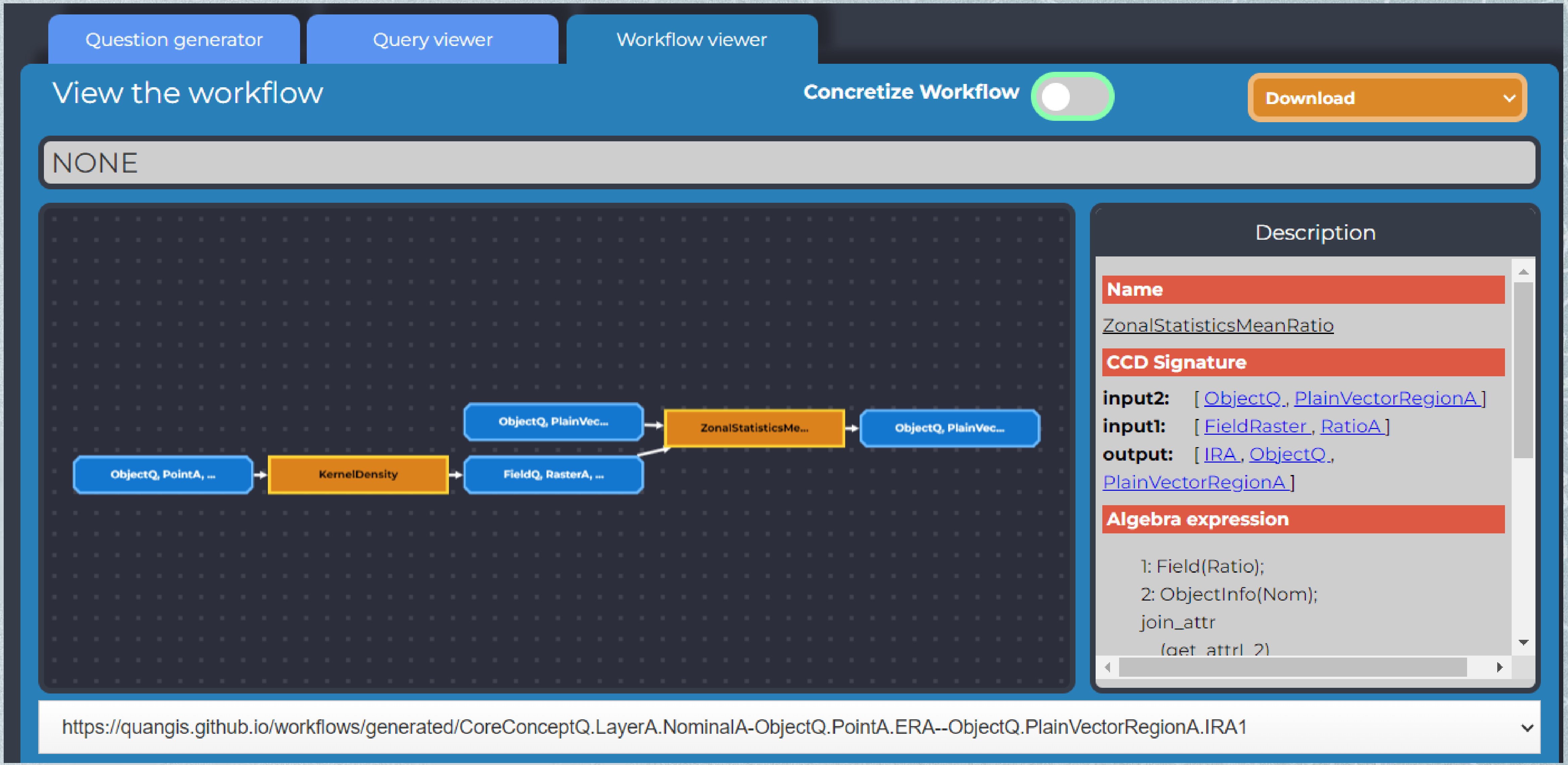
Prototype

Demo (https://131.211.60.19)



Prototype

Demo (https://131.211.60.19)



Conclusion

Pros:

- Generative
- White-box
- Explainable (AI)
- Incorporates explicit expert knowledge
  - Declarative and procedural
- Simple to use
  - Practical and education purposes

Cons:

- Domain modelling needed for generalization
- Limited scalability
- Moderate accuracy (missing functions and data sources in matching)

Future direction:

- Using a LLM for
  - Question interpretation
  - Domain model generation
- Improving accuracy and scalability with
  - IR
  - Data Matcher
  - Matching transformation functions (distance vs. density)

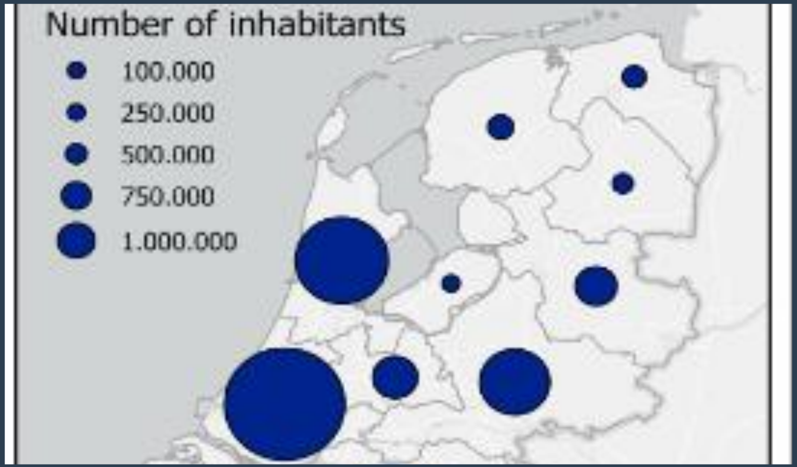
References

- 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 25(1), 10.1111/tgis.12692
- 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
- Scheider, S., Meerlo, R., Kasalica, V., & Lamprecht, A. L. (2020a). Ontology of core concept data types for answering geo-analytical questions. Journal of Spatial Information Science, 2020(20), 167-201
- Scheider, S., Nyamsuren, E., Kruiger, H., & Xu, H. (2020b). Geo-analytical question-answering with GIS. International Journal of Digital Earth, 1-14, doi: 10.1080/17538947.2020.1738568
- Scheider, S. and de Jong, T. (2022). A conceptual model for automating spatial network analysis. Transactions in GIS 26(1), 10.1111/tgis.12855
- Steenbergen, N et al. (2022): Algebra of core concept transformations. Procedural meta-data for geographic information. Preprint: https://osf.io/j6krv
- Top, E., Scheider, S., Nyamsuren, E., Xu, H. , Steenbergen, N. (2022). The Semantics of Extensive Quantities in Geographical Information. Applied ontology 17(3), 10.3233/AO-220268
- Xu, H. et al., (2022). A Grammar for Interpreting Geo-analytical Questions as core concept transformations. International Journal of Geographical Information Science 37(2), 10.1080/13658816.2022.2077947

Conceptual Theory – Measurements

Measurement examples

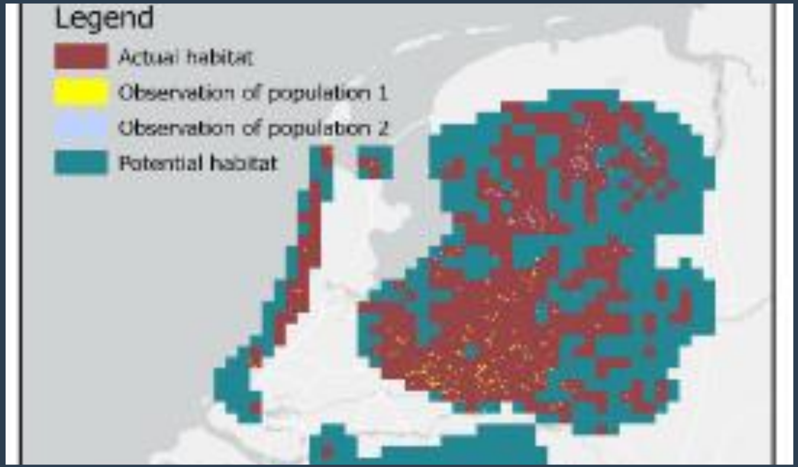
Content amount (Ratio)



Number of inhabitants in each province of the Netherlands

“Region, Obj -> Number of objects”

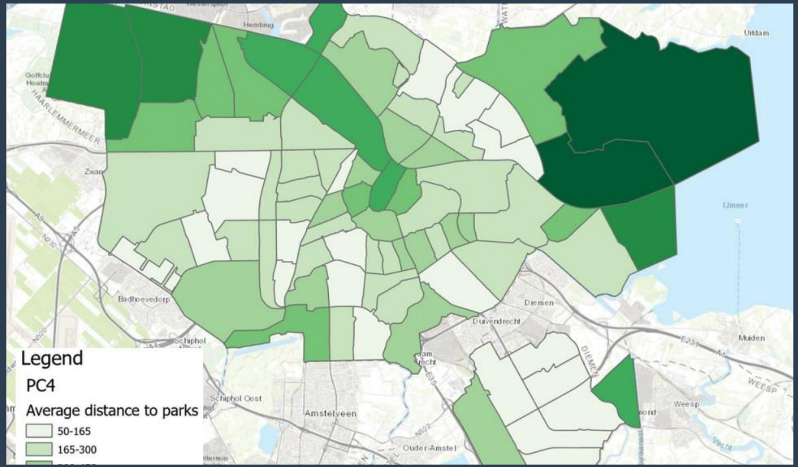
Coverage Amount (Ratio)



Living area of the European Pine Marten in the Netherlands

“Obj -> Region”

Aggregation (Ratio)



Average distance to parks in each neighborhood in Amsterdam

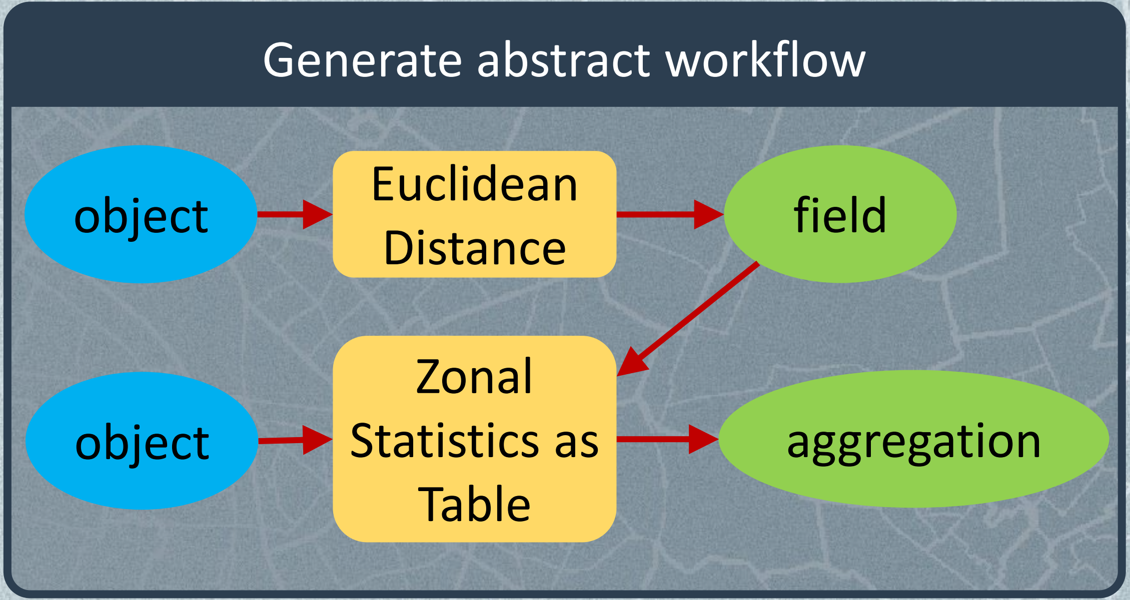
“Region, DistField -> Average distance“

(Top et al. 2022)

Methodology – Generating abstract workflows

CSP

- Generating an abstract workflow is a constraint satisfaction problem
  - Semantic constraints (matching input/output signatures)
  - Parametric constraints:
    - Length of a workflow
    - Branching factor
    - Etc…
  - Logical constraints (goal, source concepts)



Loose programming of GIS workflows with geo-analytical concepts (Kruiger et al. 2021).

Workflow Discovery with Semantic Constraints: The SAT-Based Implementation of APE (Kasalica & Lamprecht, 2020)

APE: A Command-Line Tool and API for Automated Workflow Composition (Kasalica & Lamprecht, 2020)

https://github.com/sanctuuary/APE

Prototype Architecture

Deployment