The relevance of measurement data in environmental ontology learning

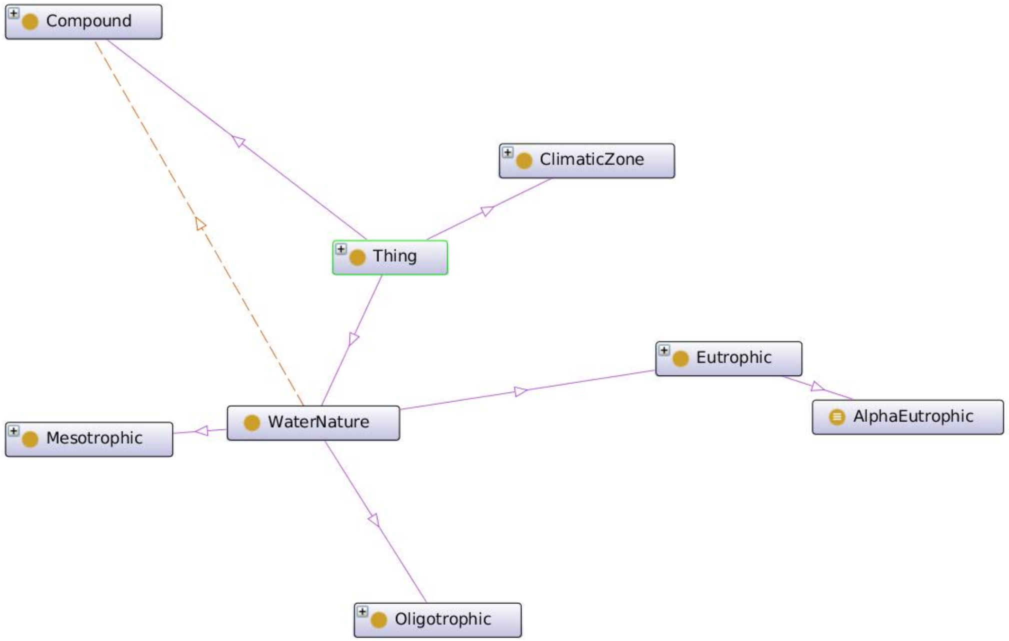
Markus Stocker, Mauno Rönkkö, Ferdinando Villa, Mikko Kolehmainen

ISESS 2011 · June 27-29, 2011 · Brno, Czech Republic

UNIVERSITY OF EASTERN FINLAND

bc³
BASQUE CENTRE FOR CLIMATE CHANGE
Klima Aldaketa Ikergai

Introduction
Ontology



Introduction
Ontology

- Schema
  - $\exists \mathit{poorIn}.\mathrm{Compound} \sqsubseteq \mathrm{Oligotrophic}$
  - $\exists \mathit{richIn}.\mathrm{Compound} \sqsubseteq \mathrm{Eutrophic}$
  - $\mathrm{Nitrogen} \sqsubseteq \mathrm{Compound}$
  - DataProperty(*totalNitrogen*)
- Individual
  - *totalNitrogen*(lakeSuperior, 2.1)

Introduction

- What does rich and poor mean, numerically?
  - From literature, classification using Trophic State Index (TI)
    - TI <30-40: Oligotrophic, thus poor in nutrients
    - TI 50-70: Eutrophic, thus rich in nutrients
  - From measurement data, using data-driven methods
    - Learn a central tendency for ”being rich” (or poor) from measurement data
- Measurement
  - Process of assigning numbers to the properties of objects
  - Fundamental to environmental science
- Hypothesis
  - Measurement data is relevant to environmental ontology learning

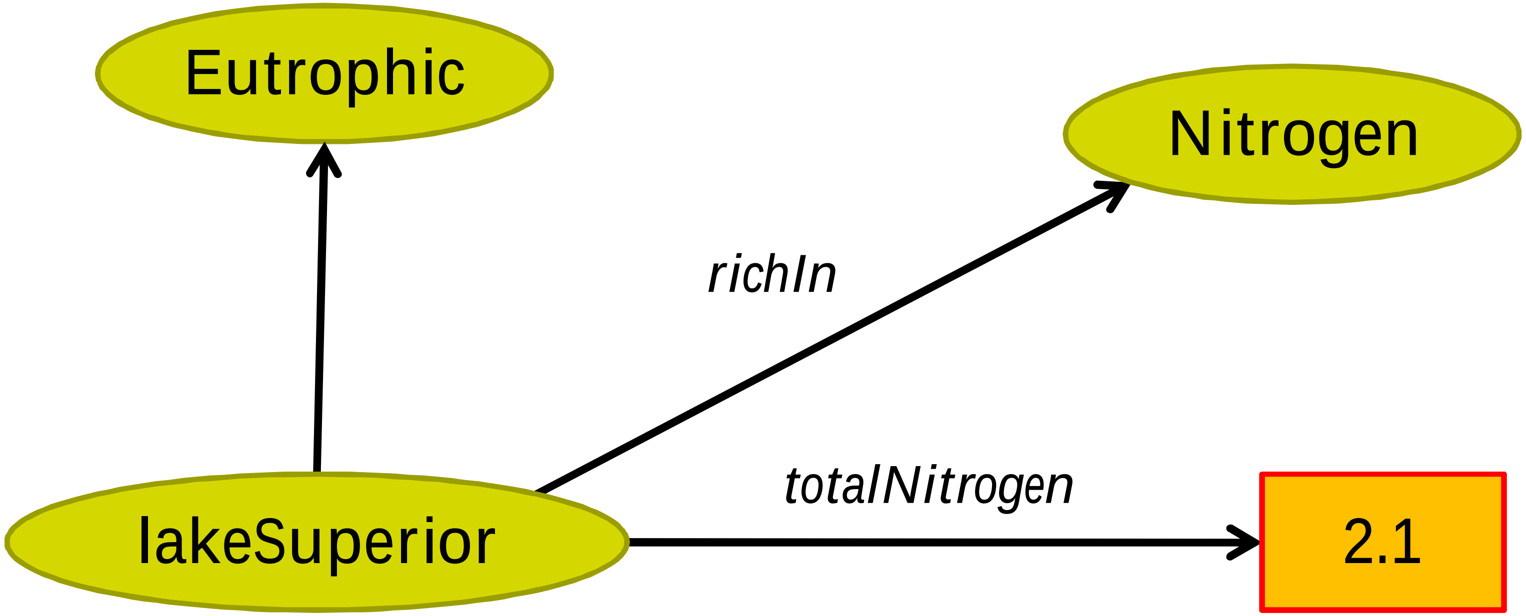
Materials

- Taxonomy of lakes formalized in OWL
  - In particular two relations *richIn* and *poorIn*
- Data on the nutrient concentration of European lakes (EEA)
  - Including mean annual total nitrogen concentration
- Jena for ontology management (RDF, OWL) and query (SPARQL)
- WEKA for data mining

Methods

- Rules
  - (Lake *totalNitrogen* X) $\wedge$ (X $\leq$ Y) $\rightarrow$ (Lake *poorIn* Nitrogen)
  - (Lake *totalNitrogen* X) $\wedge$ (X > Y) $\rightarrow$ (Lake *richIn* Nitrogen)
- Learn threshold Y
  - K-means clustering
  - Two clusters and two centroids
  - Interpreted as central tendencies for lakes being *poorIn* and *richIn*
  - Threshold Y is calculated as the mean for the centroids
- Example for centroids C_1 = 0.8 and C_2 = 2.4; Y = 1.6
  - (Lake *totalNitrogen* X) $\wedge$ (X $\leq$ 1.6) $\rightarrow$ (Lake *poorIn* Nitrogen)
  - (Lake *totalNitrogen* X) $\wedge$ (X > 1.6) $\rightarrow$ (Lake *richIn* Nitrogen)

Results
Rule-based reasoning



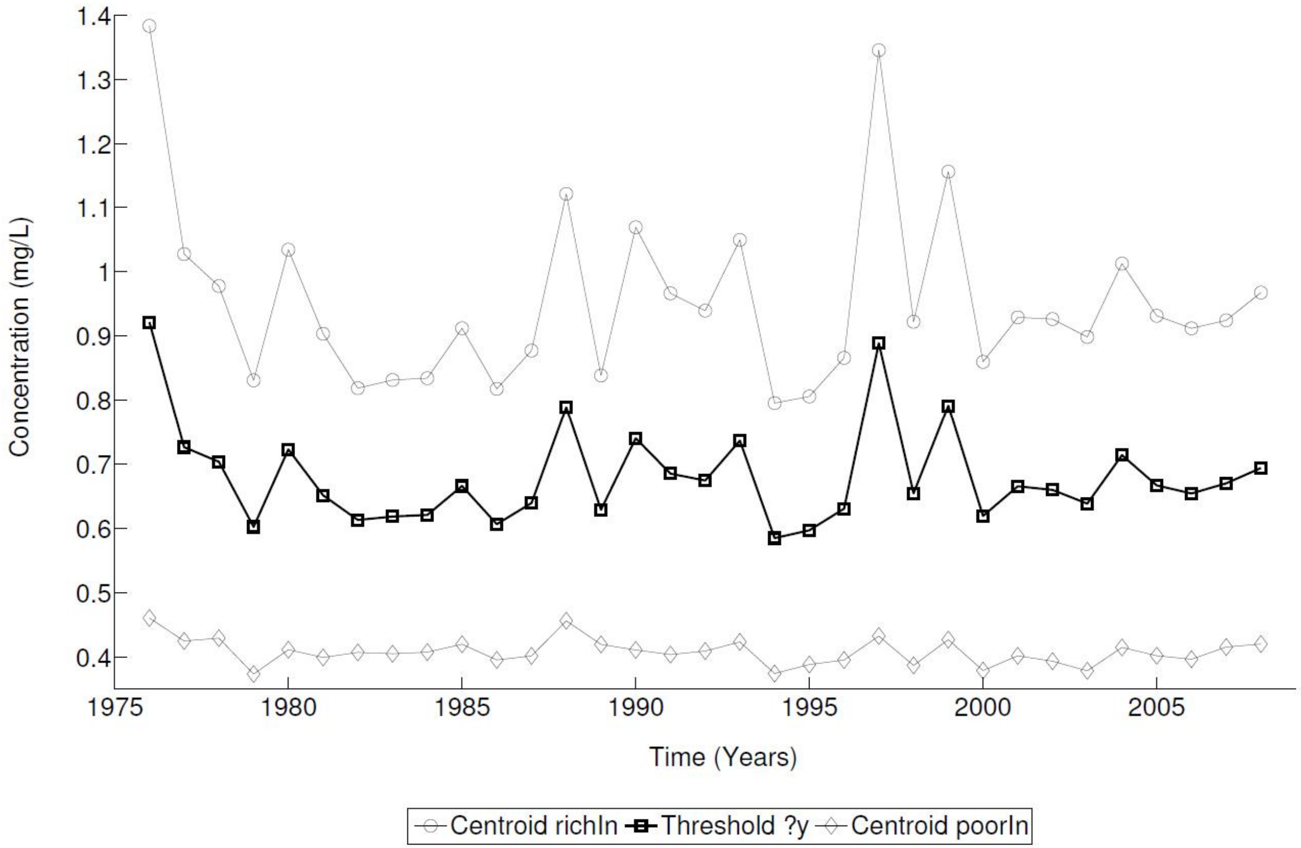
(Lake *totalNitrogen* X) $\wedge$ (X > 1.6) $\rightarrow$ (Lake *richIn* Nitrogen)

$\exists$*richIn*.Nitrogen $\sqsubseteq$ Eutrophic

Results
Spatial variation

- Using Finnish lakes in 2008
  - $C_1 = 0.39$ and $C_2 = 0.88$; $Y = 0.63$
  - 150 lakes *poorIn* and 53 *richIn* total nitrogen
- Using Spanish lakes in 2008
  - $C_1 = 0.78$ and $C_2 = 8.36$; $Y = 4.57$
  - 137 lakes *poorIn* and 12 *richIn* total nitrogen
- Tests for Denmark, Germany, Great Britain, Italy and Switzerland

Results
Temporal variation



Discussion

- Experiment suggests
  - Measurement data relevant to environmental ontology learning
  - Thus, learning methods beyond text collections needed
- Interaction
  - Feedback acquired knowledge to ontology
  - Use ontological knowledge in data mining
  - Cyclical interaction between data mining and ontologies

Discussion

- Spatial and temporal variation
  - To what extent should environmental ontologies reflect this?
  - Methods for time-space localization of ontologies
- Query
  - Lakes with mean annual total nitrogen concentration ≥1.6
  - Or, simply, Lakes *richIn* nitrogen
- Learning beyond simple ontological rules

Related work

- Data mining with ontologies cycle (Nigro *et al.*)
- Rule-based reasoning for environmental ontologies (Henson *et al*.)
  - Wind $\geq$ 35 miles/h $\rightarrow$ HighWinds
  - Similar use case
  - Authors give no indication on threshold value
    - May be expert opinion
  - We learn threshold value from the data

Conclusions

- Aims
  - Demonstrate the learning of ontological rules
  - Using numerical measurement data and clustering methods
- Hypothesis
  - Measurement data is relevant to environmental ontology learning
- Is the interpretation given to centroids valid?
  - Bridge between data mining and ontology
  - Open for discussion

References

- Zafar, A.: Taxonomy of lakes. Hydrobiologia 13(3), 287-299 (1959)
- Sydenham, P.H.: Handbook of Measurement Science: Volume 1 Theoretical Fundamentals. John Wiley & Sons (1982)
- Nigro, H.O., Císaro, S.E.G., Xodo, D.H.: Data mining with ontologies: Implementations, findings, and frameworks. Information Science Reference (an imprint of IGI Global) (2008)
- Henson, C.A., Pschorr, J.K., Sheth, A.P., Thirunarayan, K.: SemSOS: Semantic Sensor Observation Service. In: Proc. of the 2009 International Symposium on Collaborative Technologies and Systems (CTS 2009). Baltimore, MD (May 2009)
- http://en.wikipedia.org/wiki/Trophic_state_index
- http://www.openjena.org/
- http://www.cs.waikato.ac.nz/ml/weka/
- http://en.wikipedia.org/wiki/Lake