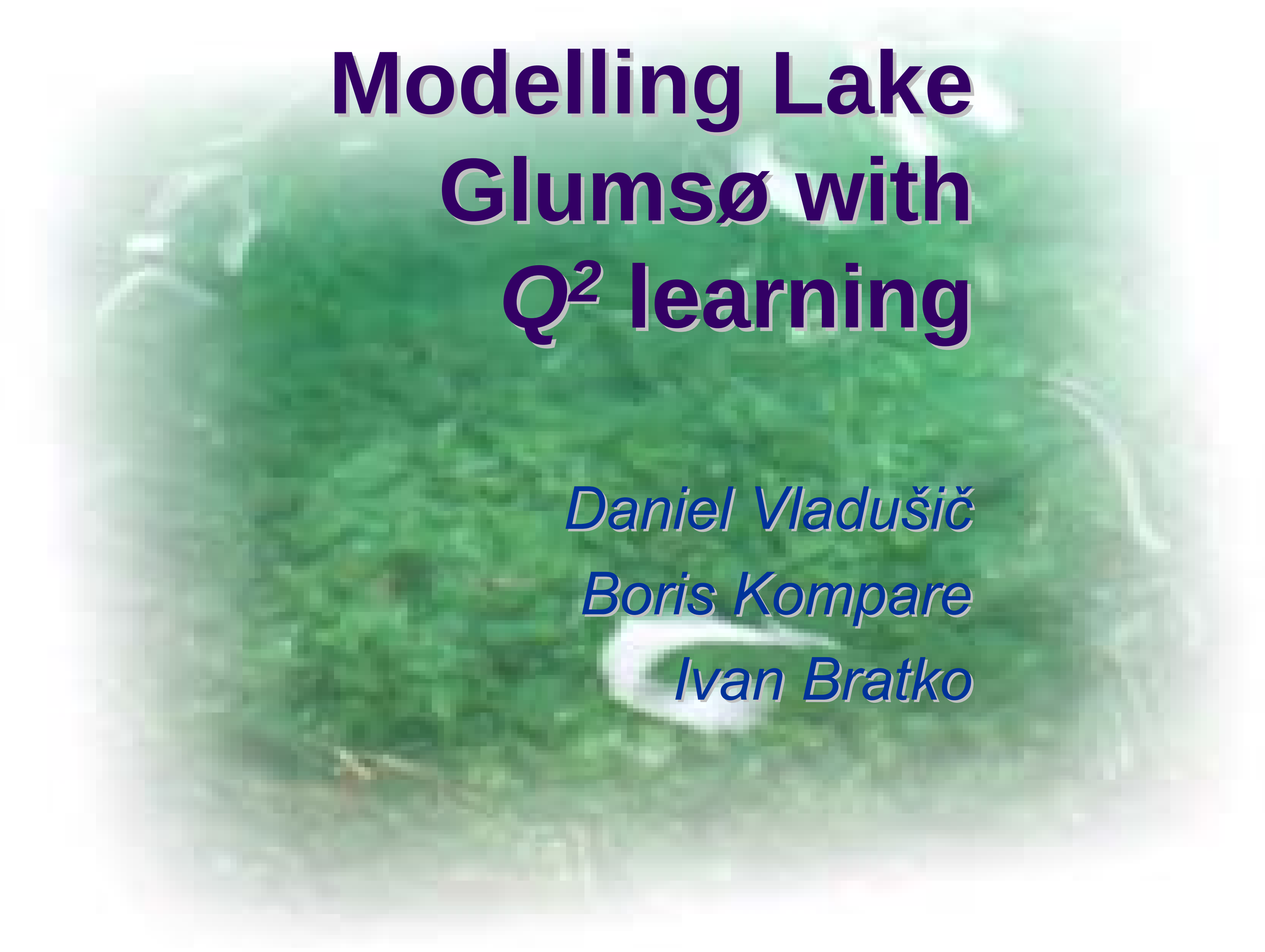


Modelling Lake Glumsø with Q^2 learning

Daniel Vladušič

Boris Kompare

Ivan Bratko

About

- Modelling of algae growth in Lake Glumsø
- Data – S.E. Jørgensen*
- Model extracted from data with ML
- ML used: Q^2 learning

*Data collected by S. E. Jørgensen –
Jørgensen, S., Kamp-Nielsen, L., Chirstensen, T., Windolf-Nielsen, J., and Westergaard, B.: Validation of a prognosis based upon a eutrophication model. *Ecological Modelling* **32**, 165, 1986.

Related Work

At least …

- S.E. Jørgensen
- S. Džeroski
- L. Todorovski
- B. Kompare
- I. Bratko
- V. Križman
- D. Demšar

Jørgensen, S., Kamp-Nielsen, L., Chirstensen, T., Windolf-Nielsen, J., and Westergaard, B., 1986. Validation of a prognosis based upon a eutrophication model. *Ecological Modelling* **32**, 165.

Križman, V., Džeroski, S. and Kompare, B., 1995. Discovering dynamics from measured data. *Electrotech. Rev.* 62, 191–198.

Demšar, D., 1996. Experiments in automated modeling of ecological processes in Lake Glumsoe. BSc Thesis, Faculty of Computer and Information Science, University of Ljubljana, Slovenia (in Slovenian).

B. Kompare, L. Todorovski and S. Džeroski. Modeling and prediction of phytoplankton growth with equation discovery: Case study - Lake Glumsø, Denmark. *Verhandlungen - Internationale Vereinigung fuer Theoretische und Angewandte Limnologie*, 27: 3626-3631, 2001.

Todorovski, L., Džeroski, S., and Kompare, B., 1998: Modelling and prediction of phytoplankton growth with equation discovery, *Ecological Modelling* 113: 71--81.

Džeroski, S, Todorovski, L., Bratko, I., Kompare, B. and Križman, V., 1999: Equation discovery with ecological applicatins. In: Fielding, A.H. (ed.). *Machine learning methods for ecological applications*. Boston; Dordrecht; London: Kluwer Academic Publishers, 185--207.

Lake Glumsø

- Location and properties:
  - Lake Glumsø is located in a sub-glacial valley in Denmark
  - Average depth 2 m
  - Surface area 266000 m^2

- Pollution
  - Receives waste water from community with 3000 inhabitants (mainly agricultural)
  - High nitrogen and phosphorus concentration in waste water caused hypereutrophication
  - No submerged vegetation
    - low transparency of water
    - oxygen deficit at the bottom of the lake

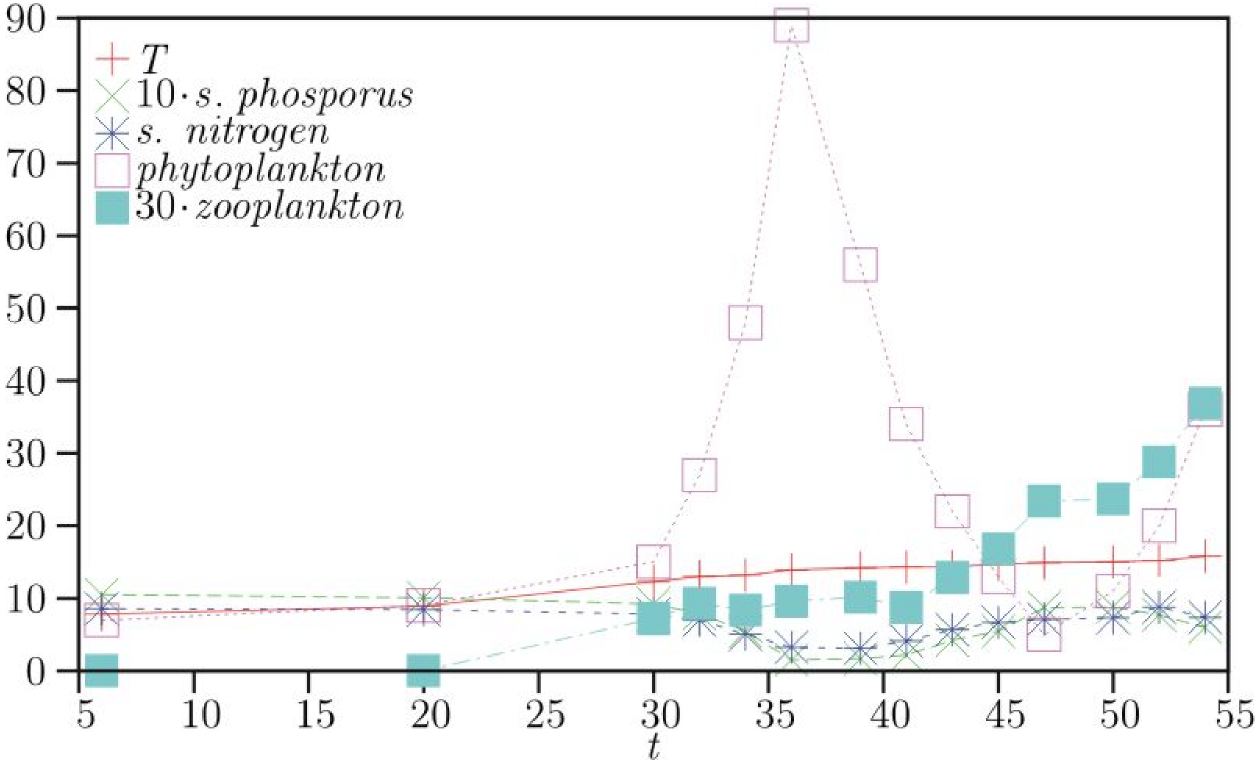
Lake Glumsø – data

- Relevant variables for modelling are:
  - phytoplankton *phyto*
  - zooplankton *zoo*
  - soluble nitrogen *ns*
  - soluble phosphorus *ps*
  - water temperature *temp*

EAML ’04
Bled

Lake Glumsø – data (*contd.*)

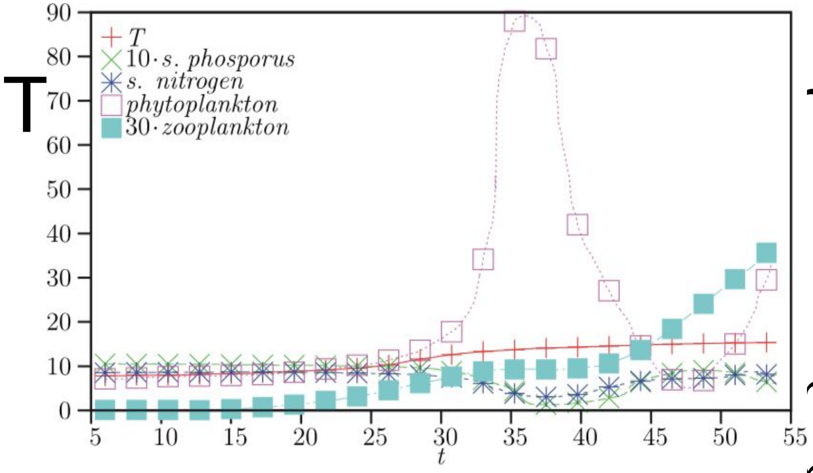
- Taken at 14 distinct time points over a period of two months



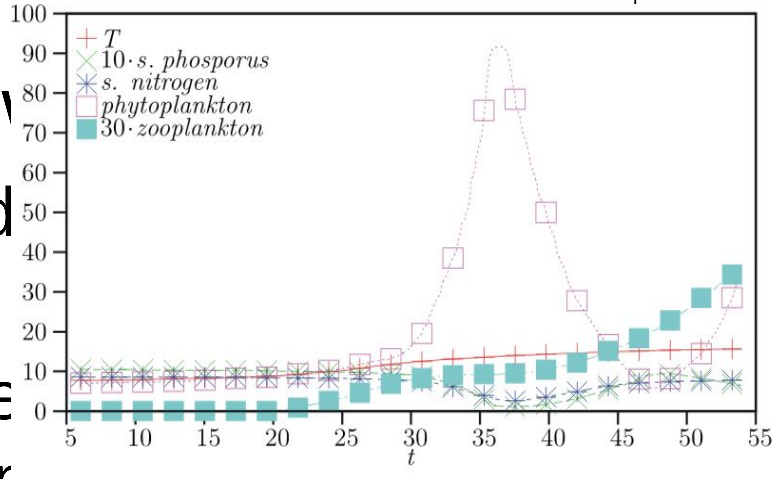
Lake Glumsø – introducing expert knowledge

EAML ’04
Bled

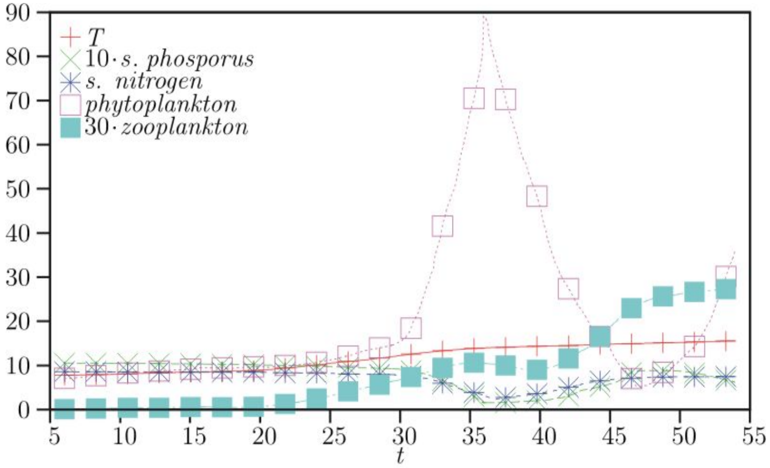
T ms \
and
n Bε
ne ir



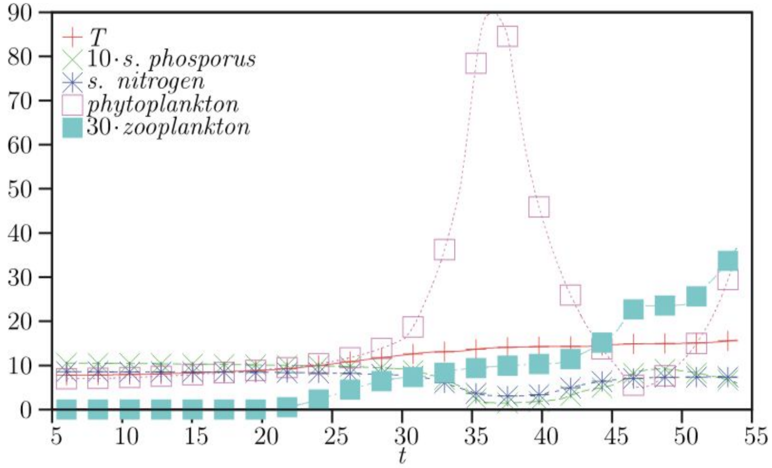
First expert



Second expert



Third expert



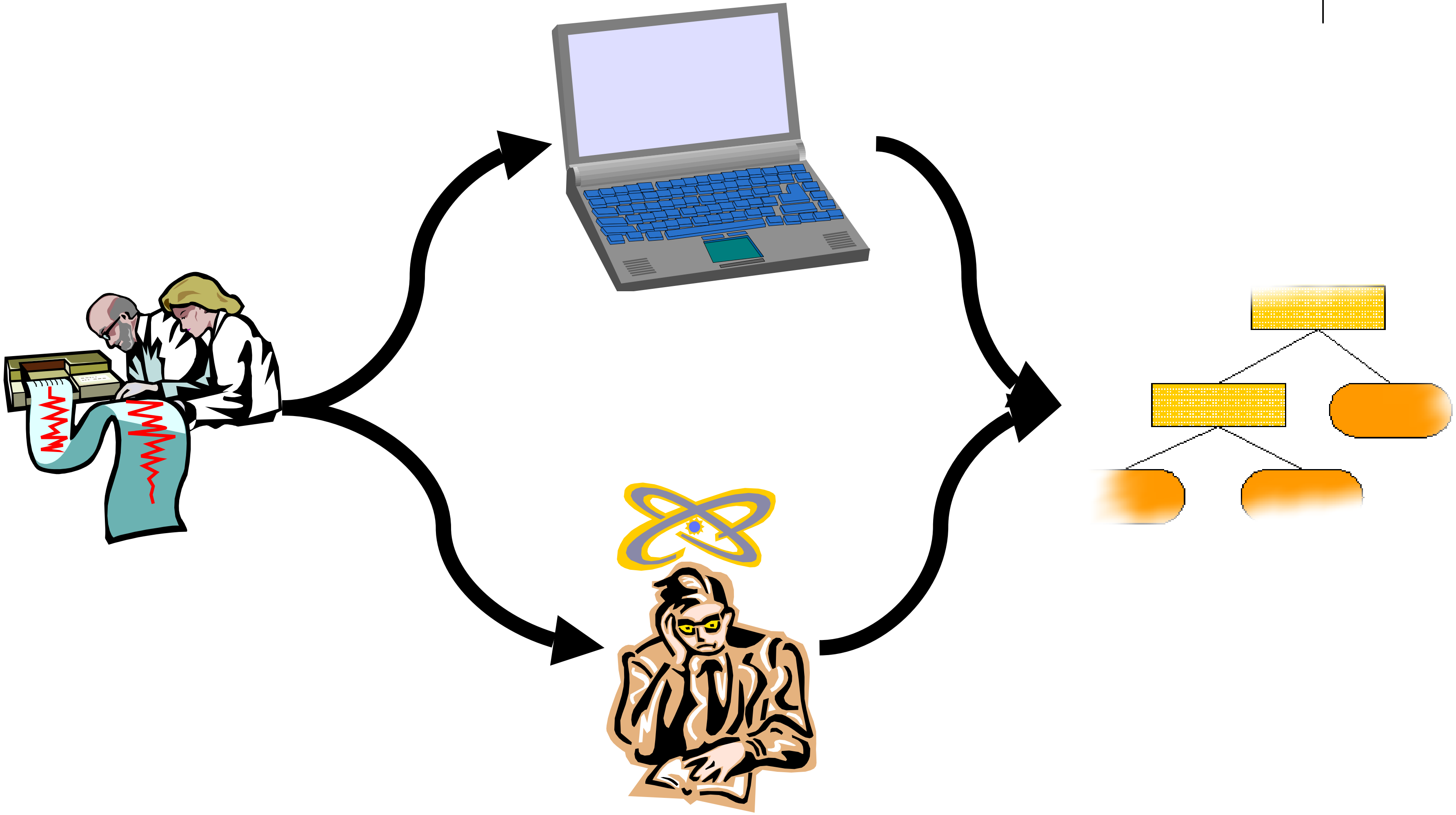
Fourth expert

Lake Glumsø – model outline

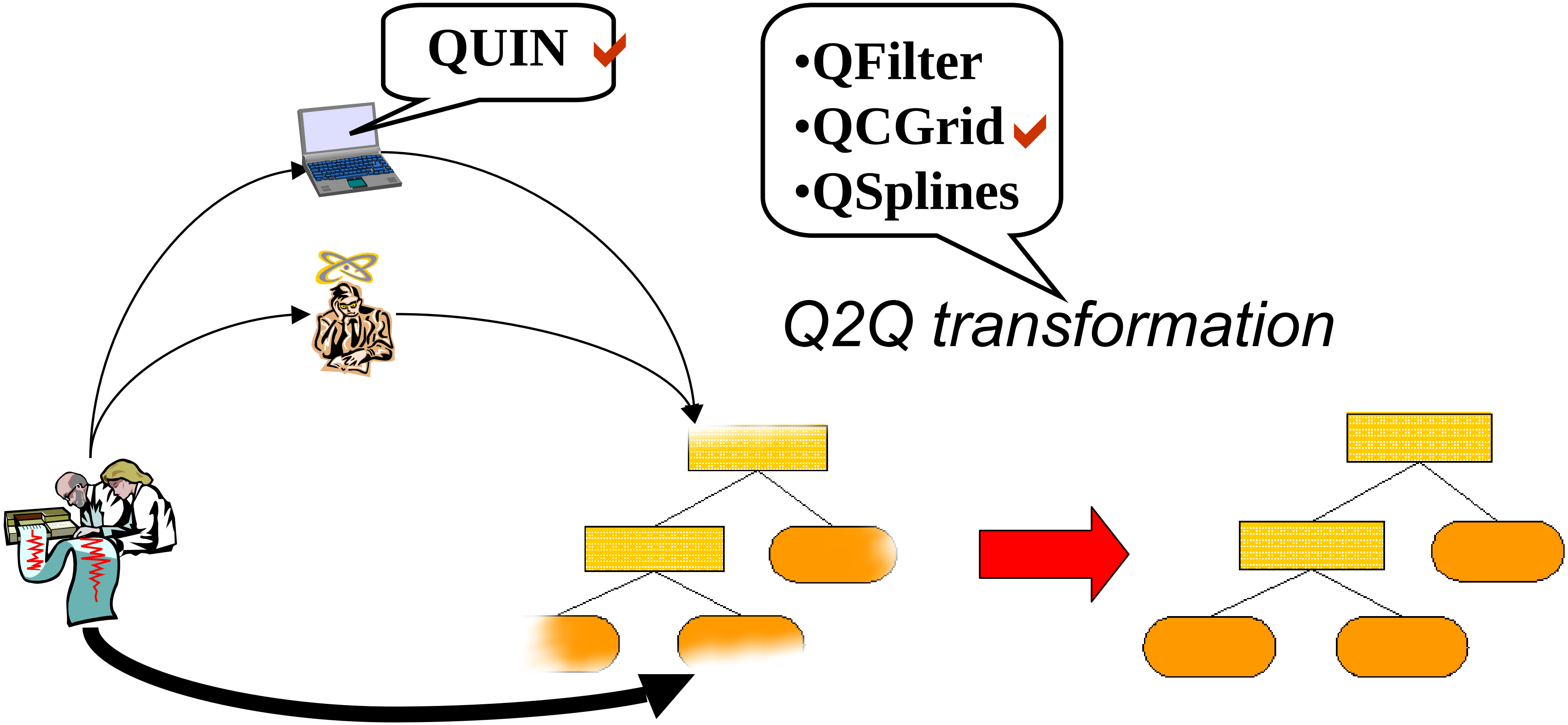
$$\begin{aligned}\boxed{phyto_{t+1}} &= f_1(phyto_t, zoo_t, temp_t, ps_t, ns_t)\\ zoo_{t+1} &= f_2(phyto_t, zoo_t)\end{aligned}$$

Concentrated on prediction of *phyto* - harder to predict

Q^2 approach to machine learning



Q^2 approach (*contd.*)



QUIN ✓

- QFilter
- QCGrid ✓
- QSplines

Q2Q transformation

Features of the Q^2 approach

Qualitative consistency of the prediction with the learning data can:

- help with the interpretation of the phenomena in the modeled domain
- considerably improve numerical accuracy

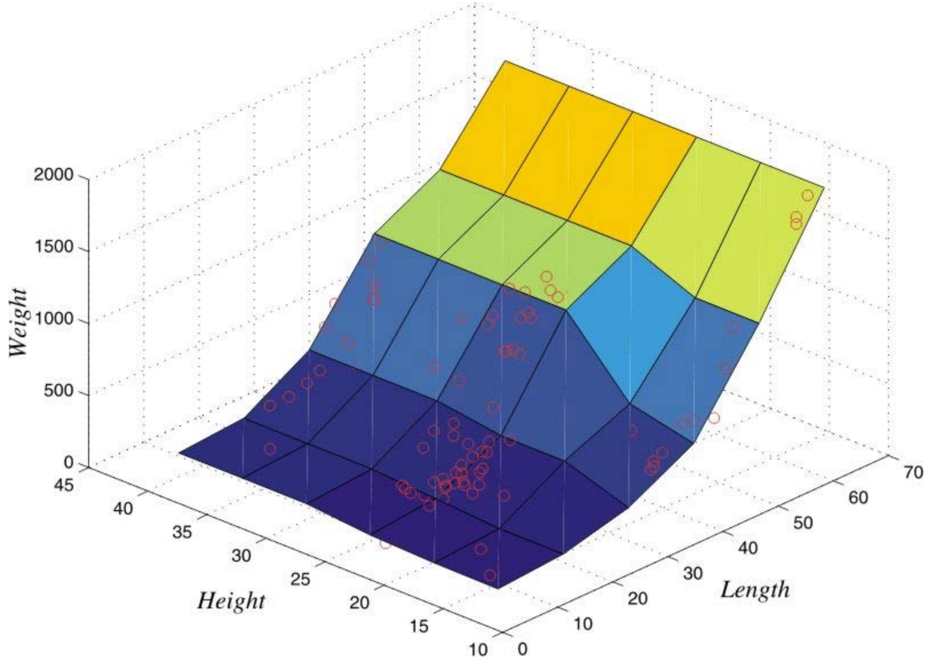
Q^2 – a simple example

Fish data*

Weight = *f*(Length,Height)

Weight = $M^{+,+}$(Length,Height)

Monotonic Qualitative Constraint

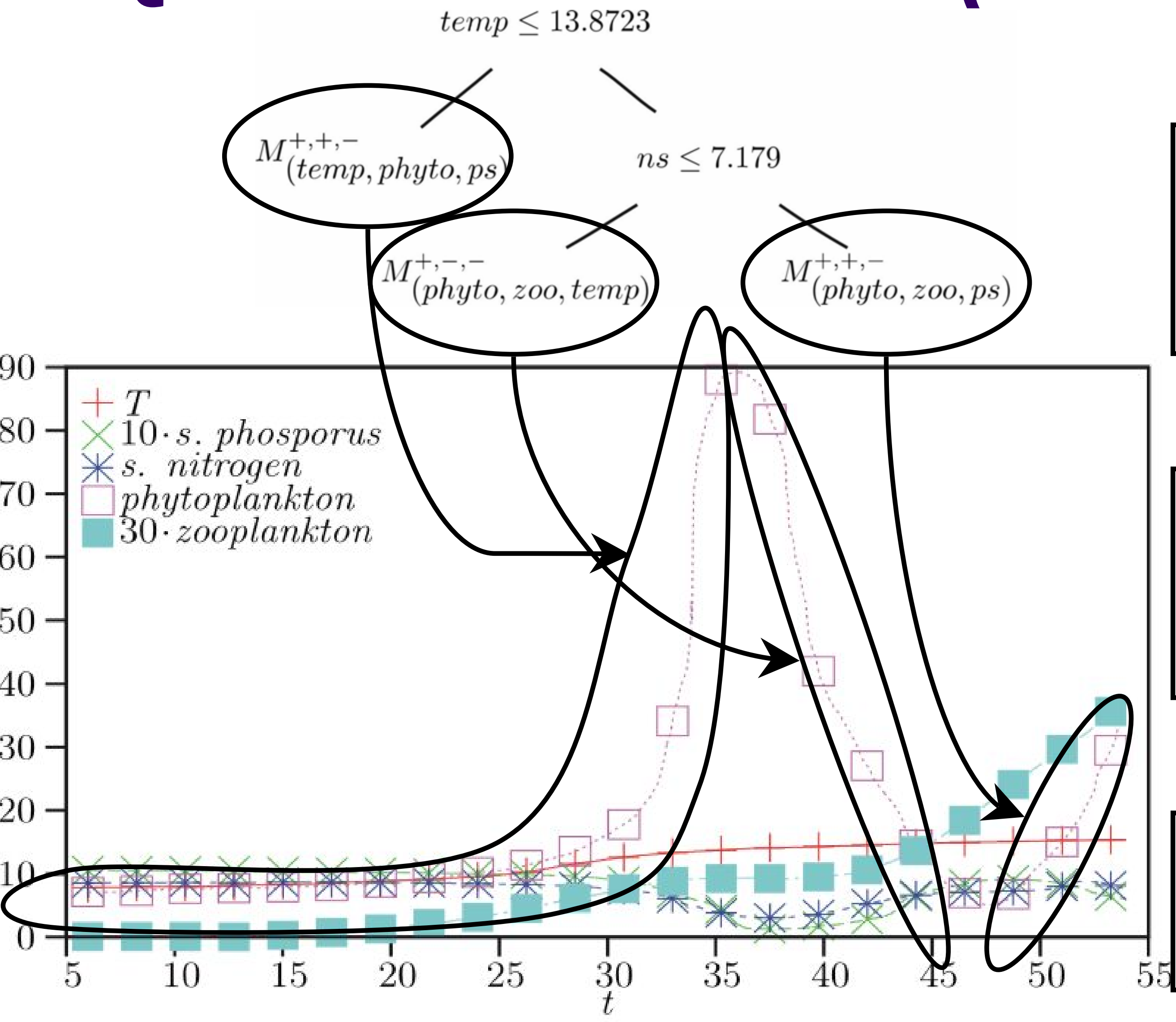


*Fishcatch data from UCI repository

Lake Glumsø – dataset difficulties

- Originally, the dataset too scarce
- Enhanced with different expert knowledge
- Dynamics of the system not well represented:
    - *temp* and *zoo* always increasing → time index → used as splits in the tree, but not necessarily biologically meaningful
    - only one peak of *phyto* captured

Qualitative models *(1. expert)*



Phyto increasing thus food *(ps)* decreasing.
Temp is an index (always increasing)

Zoo still increasing while *phyto* decreasing.
Temp is an index (always increasing)

Start of the second rise:
zoo increasing while *ps* decreasing due to *phyto*

Numerical evaluation – experimental setup

40 experiments

- Methodology
  - Each trace cut into 10 consecutive segments
  - 9 segments used for learning, 1 segment for testing (a kind of 10-fold cross-validation)
  - Prediction methods: Q^2, compared with LWR and M5
  - Parameter fitting:

    4-fold cross-validation on the learning set

Numerical results

	Q^2	LWR	M5
$\overline{\text{RMSE}}_{phyto}$	7.7	17.1	40.4
$\sigma\text{RMSE}_{phyto}$	11.5	22.3	91.4

- Over all (40) experiments.
- Q^2 better than LWR in 75% (M5, 83%) of the test cases
- The differences were found significant (t-test) at 0.02 significance level

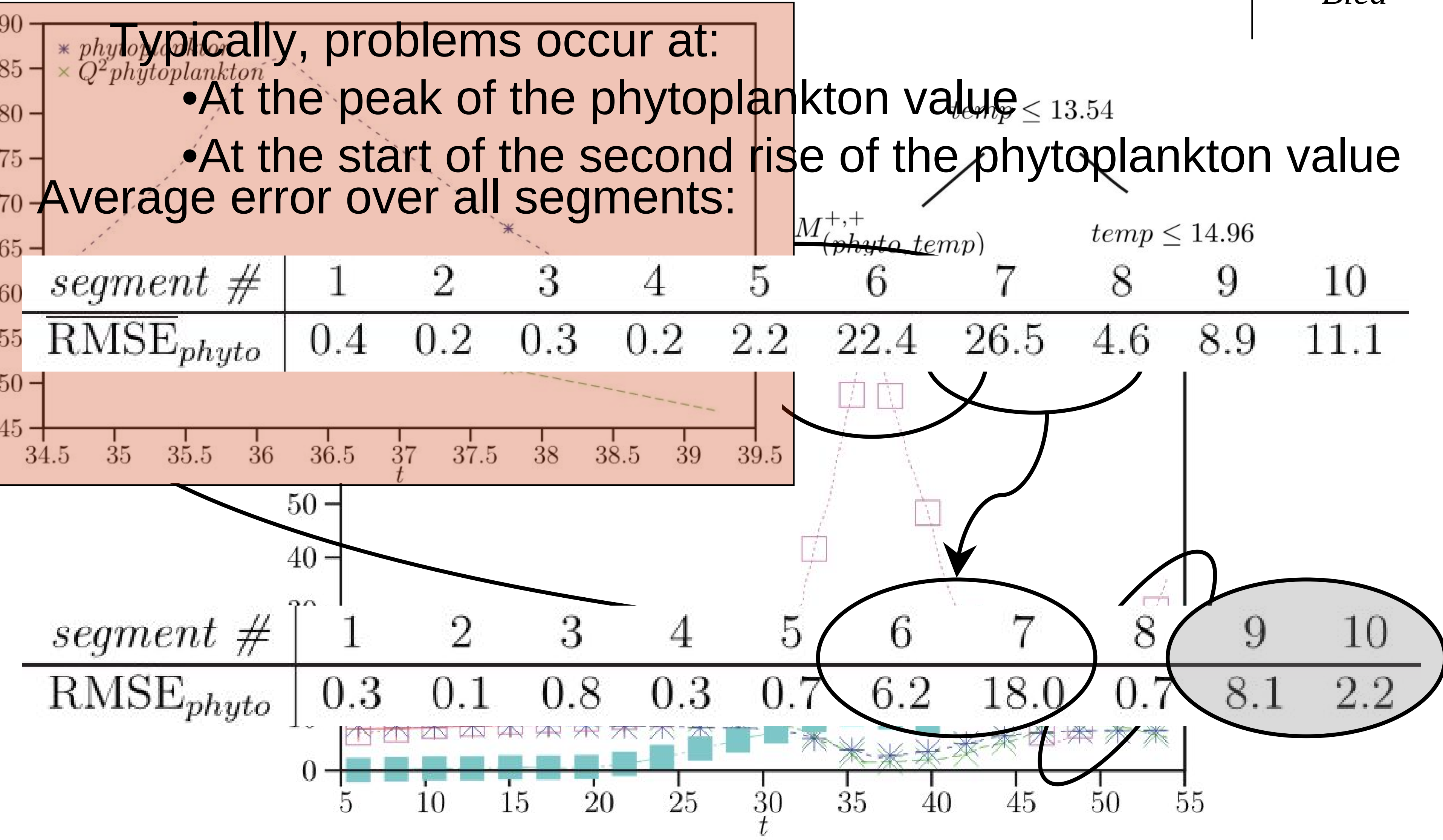
Typical problems

Typically, problems occur at:

- At the peak of the phytoplankton value
- At the start of the second rise of the phytoplankton value

Average error over all segments:

segment #	1	2	3	4	5	6	7	8	9	10
RMSE_{phyto}	0.4	0.2	0.3	0.2	2.2	22.4	26.5	4.6	8.9	11.1



segment #	1	2	3	4	5	6	7	8	9	10
RMSE_{phyto}	0.3	0.1	0.8	0.3	0.7	6.2	18.0	0.7	8.1	2.2

Conclusions

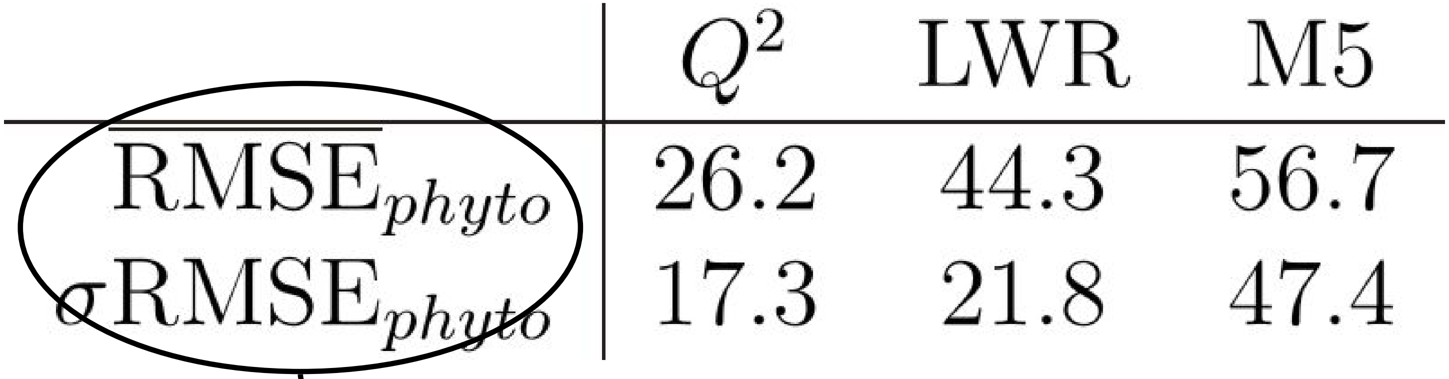
- Q^2 is numerically significantly better than other representative ML methods

- The dataset insufficiently represents the dynamics of the system

Thank You!

Questions, please

Numerical results (additional)

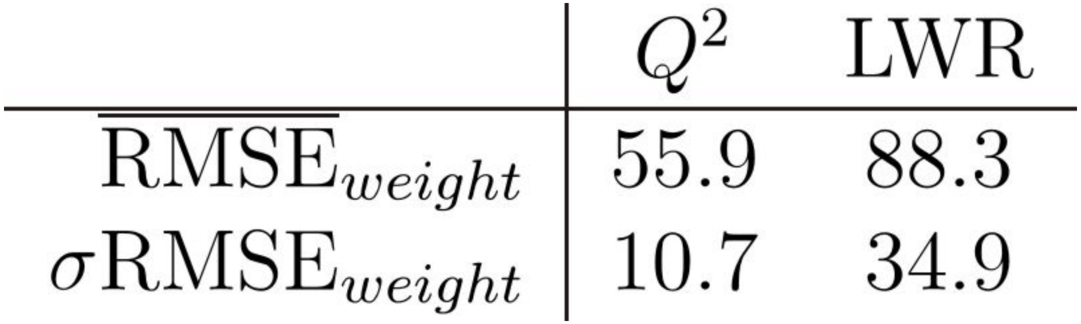


	Q^2	LWR	M5
$\overline{\text{RMSE}}_{phyto}$	26.2	44.3	56.7
$\sigma\text{RMSE}_{phyto}$	17.3	21.8	47.4

- Each trace was divided into four (4) segments
- Results shown over all (16) experiments.

Q^2 – a simple example (*contd.*)

- 5-fold cross-validation
- Q^2 better than LWR in 80% of the test cases



	Q^2	LWR
$\overline{\mathrm{RMSE}}_{weight}$	55.9	88.3
$\sigma\mathrm{RMSE}_{weight}$	10.7	34.9