Aggregation and processing of data originating from heterogeneous sources for multidimensional analyses

Marcin Wichorowski

Katarzyna Błachowiak-Samołyk

Institute of Oceanology
Polish Academy of Sciences

IMDIS 2016

October 11th-13th, Gdańsk

4th paradigm

1. Empirical science
2. Theory and models
3. Numerical modelling
4. Data intensive science

Jim Gray, "The Fourth Paradigm", Microsoft Research, Redmond

Big Data V's

1. Volume
2. Velocity
3. Variety

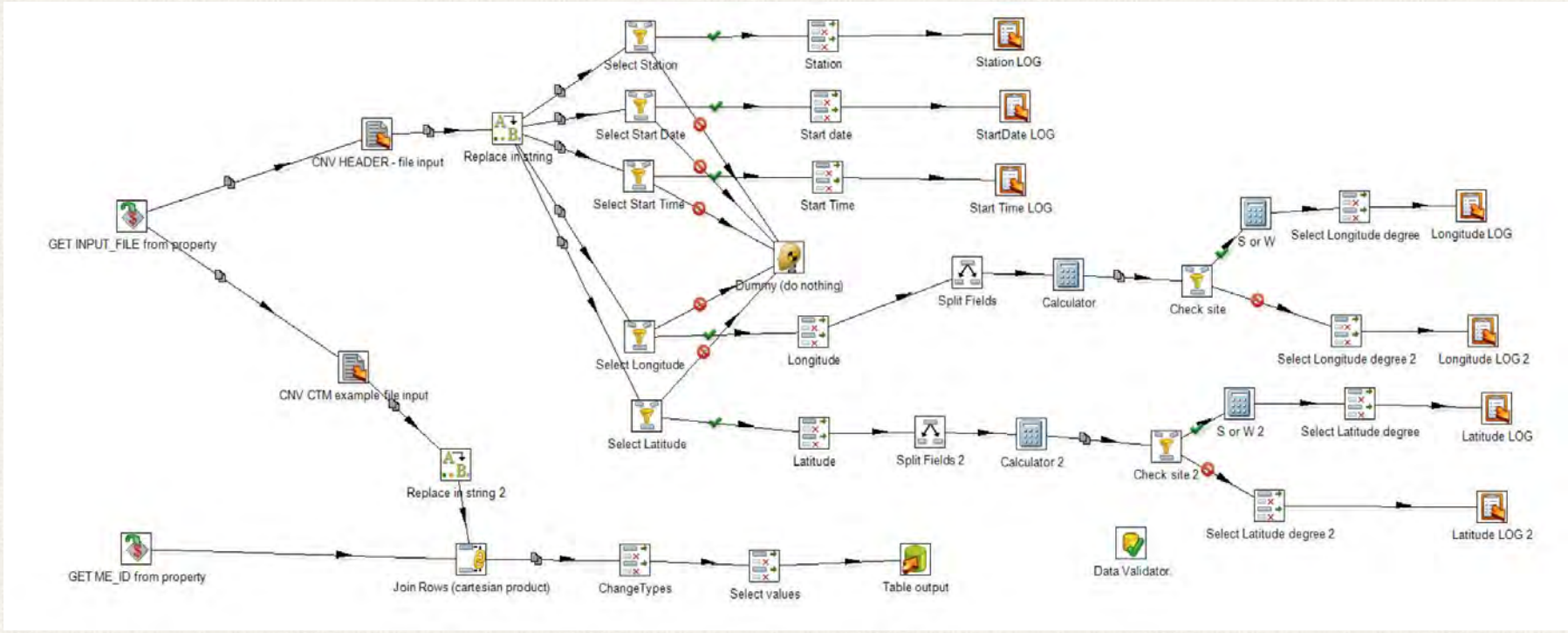
Doug Laney's 3 V's

4. Variability
5. Veracity
6. Visualisation
7. Value

Mark van Rijmenam

Graphical definition of ETL process

- Pentaho Business Analytics - GUI managed definition of ETL process
- Python – SQL Alchemy



The case: foraging grounds for little auk

Environment :

- currents
- temperature
- salinity
- water transparency

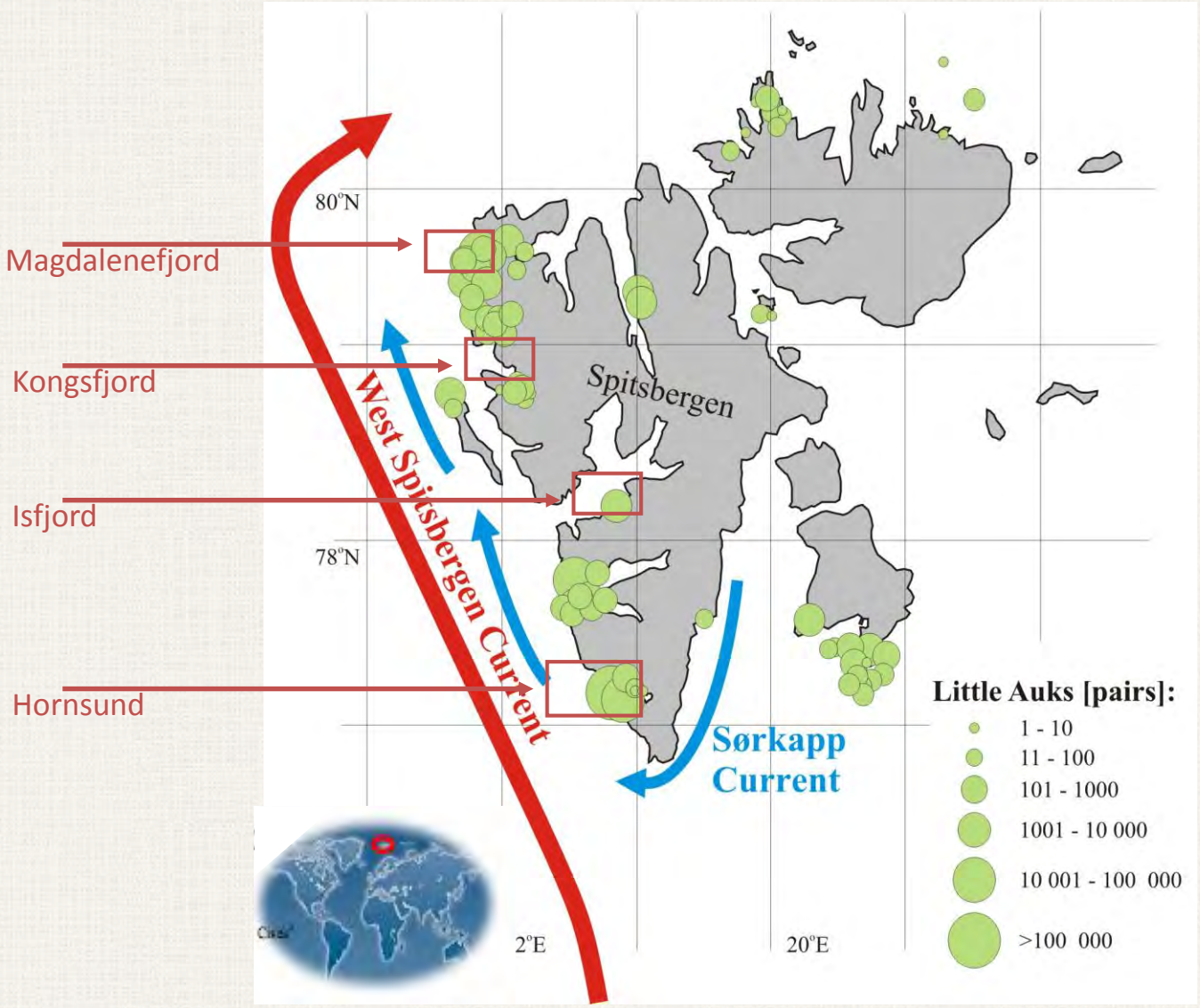
zooplankton

little auk

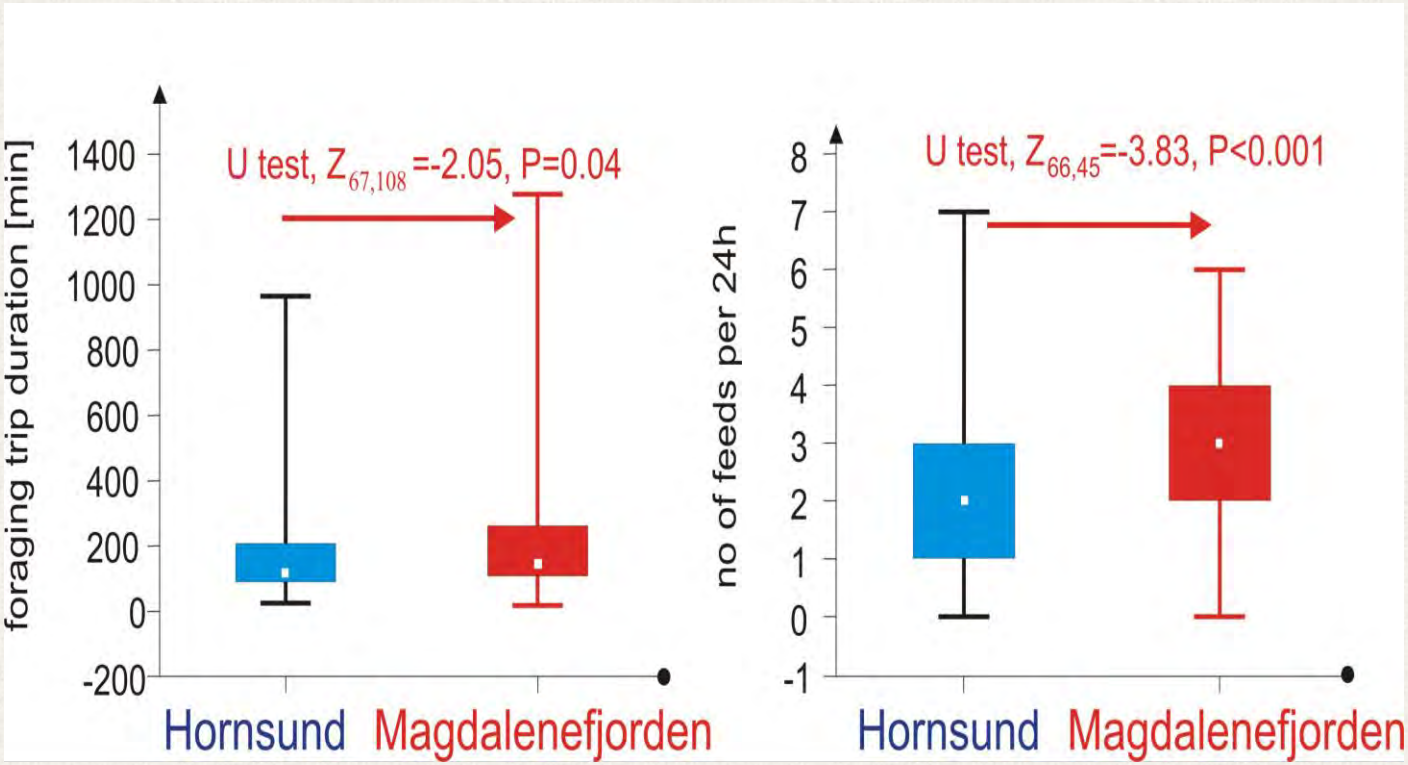
The goal of research is to establish different scenarios of Arctic ecosystem changes based on relationship between marine environment parameters, zooplankton, seabird and terrestrial communities

The challenge is defined as classification of certain area as potential feeding field for little auk

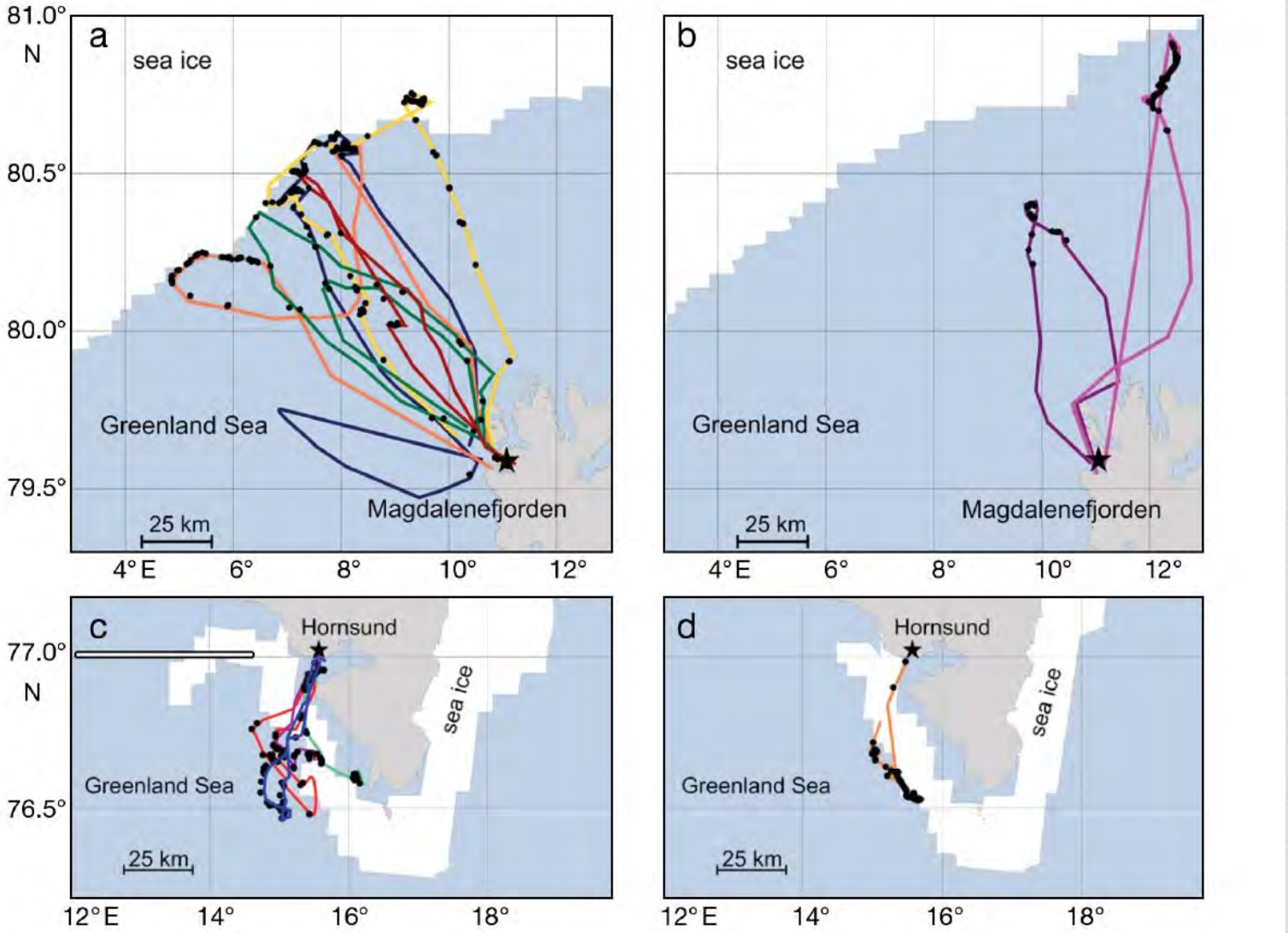
Location of little auk colonies at Svalbard



Little auk foraging trips

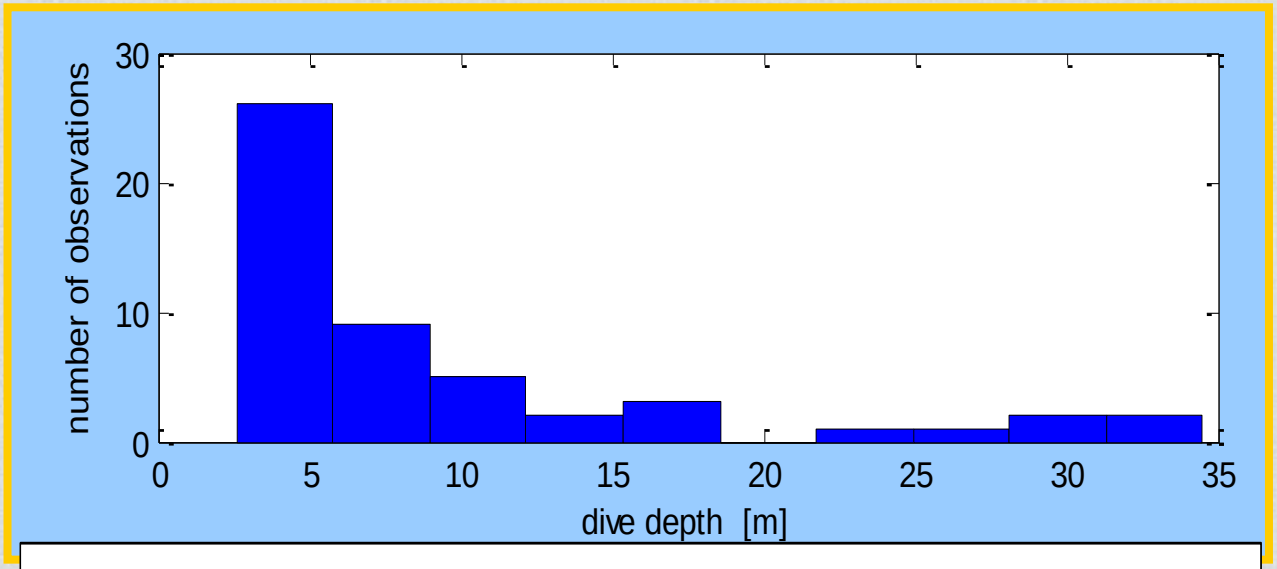


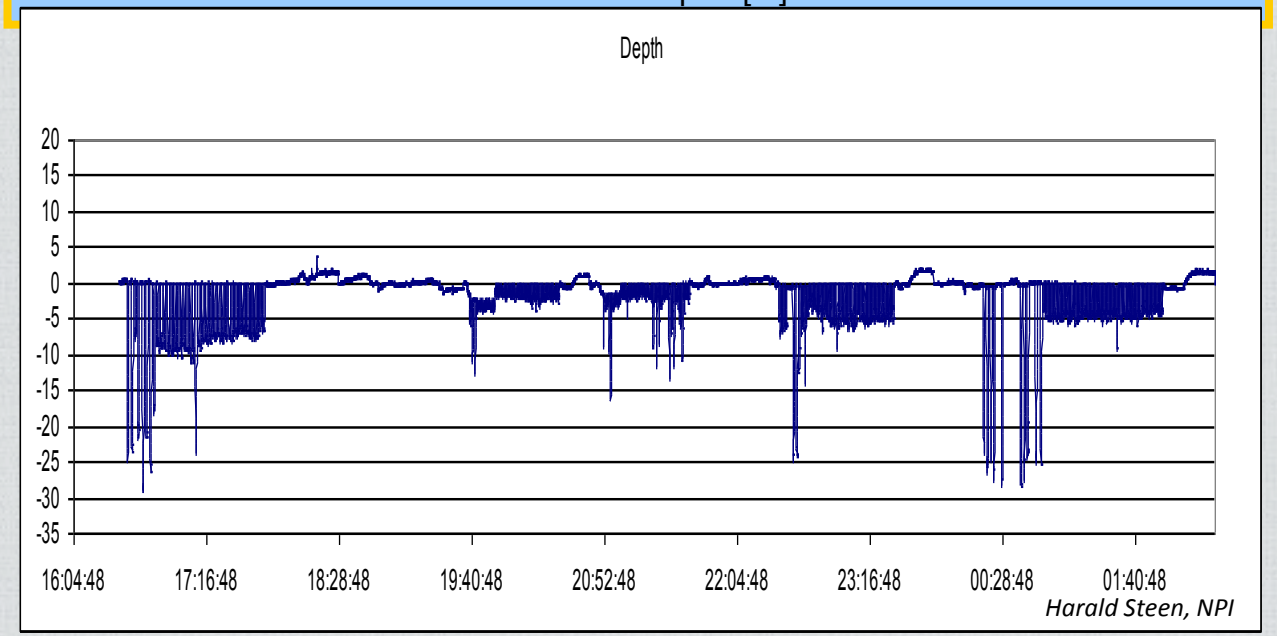
Little auk foraging trips



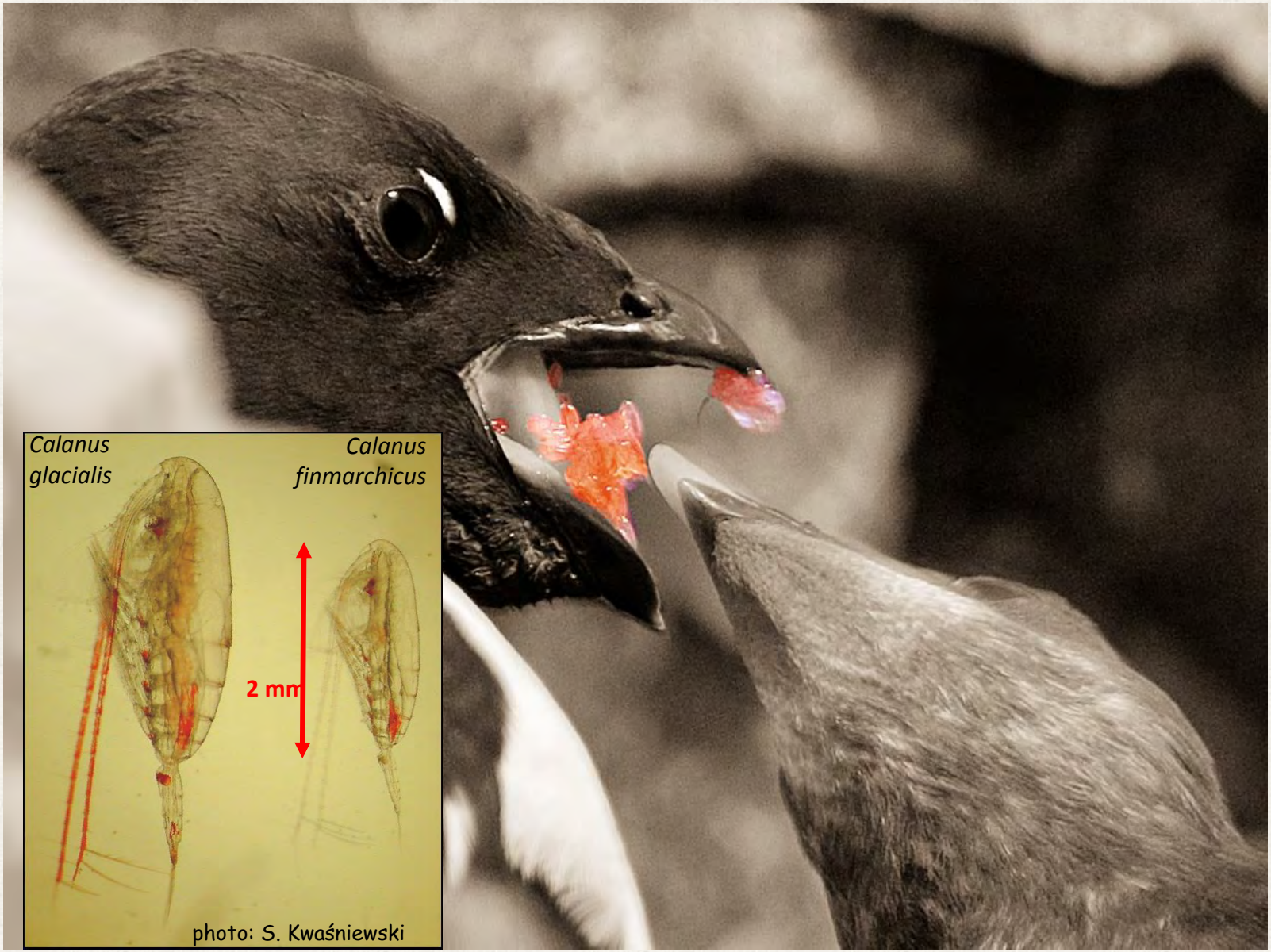
MARINE ECOLOGY PROGRESS SERIES, doi: 10.3354/meps10414, *Jakubas at all*

Little auk diving depths

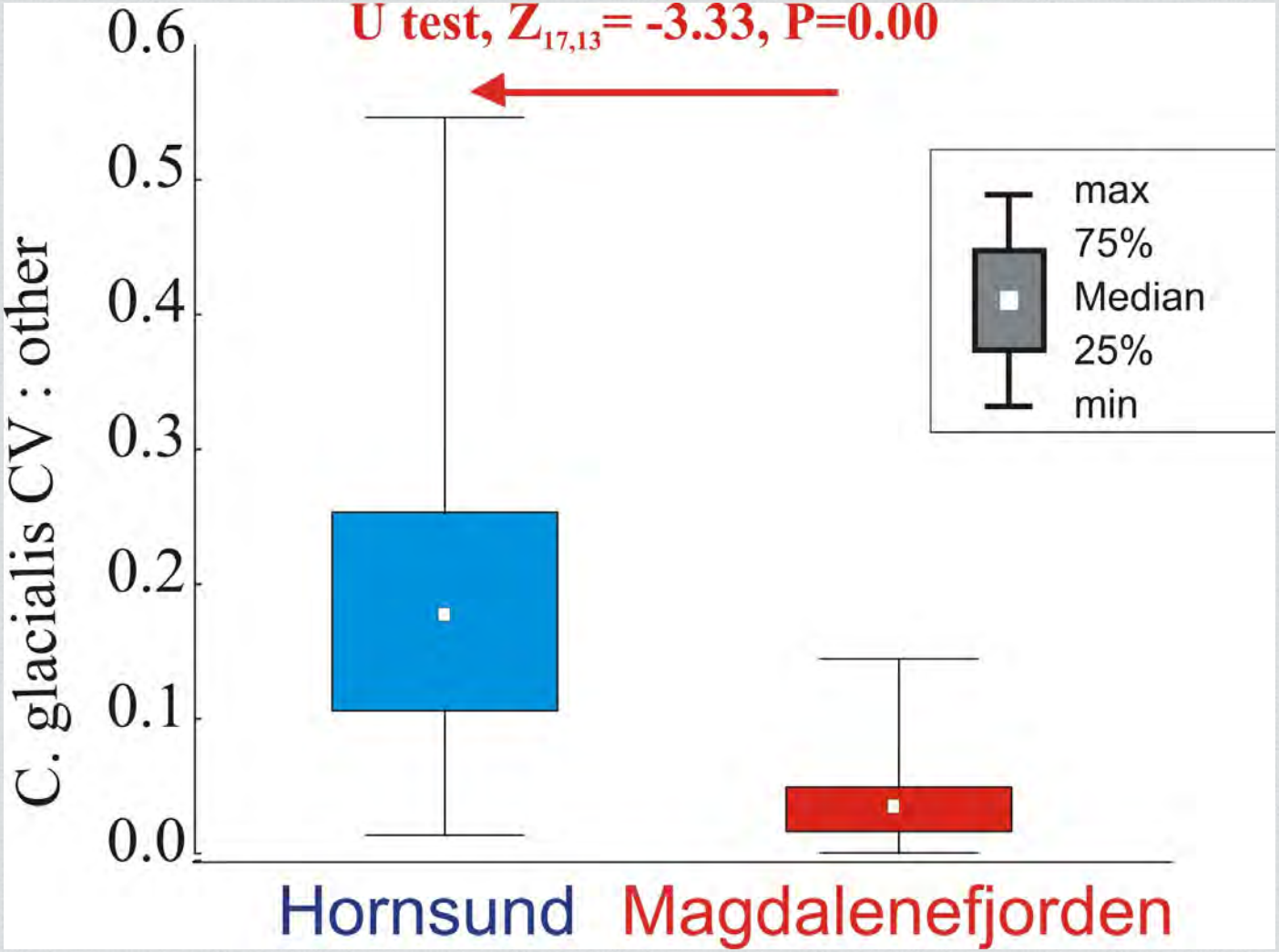




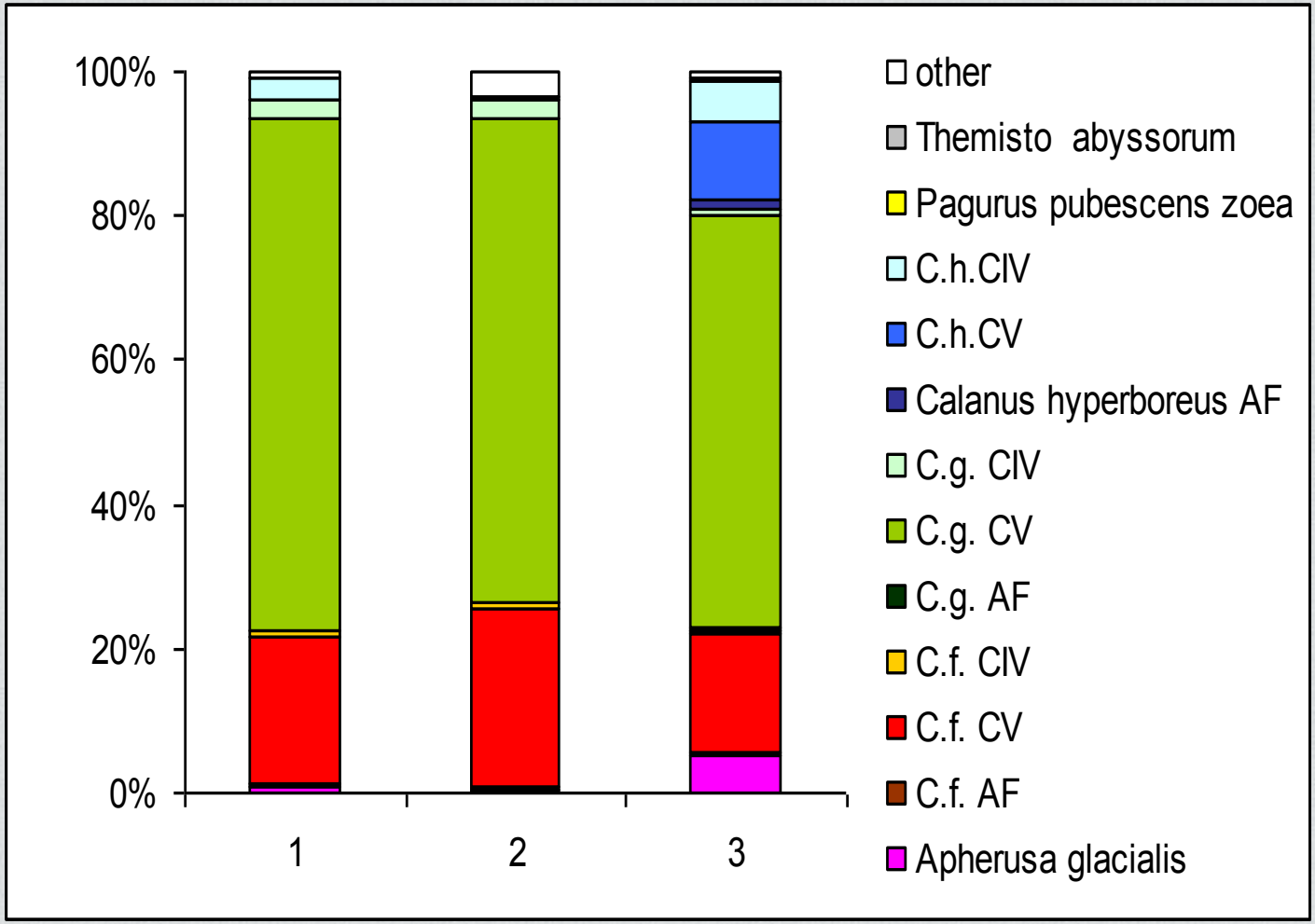
Little auk diet



Calanus glacialis in feeding grounds

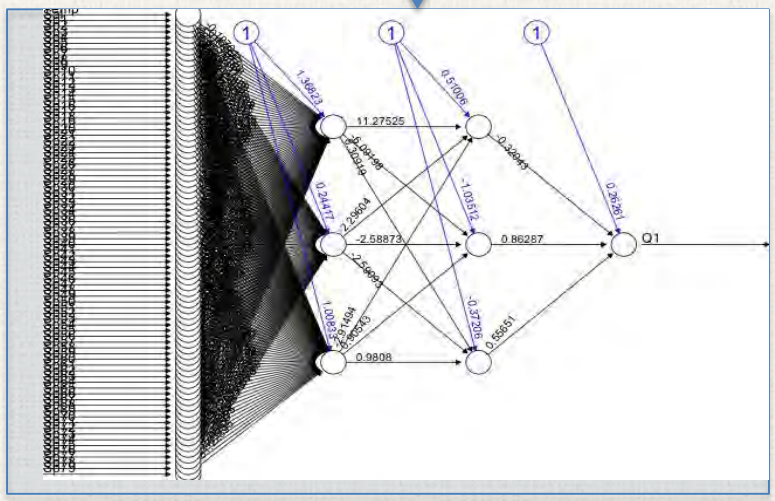


Little auk chick diet composition

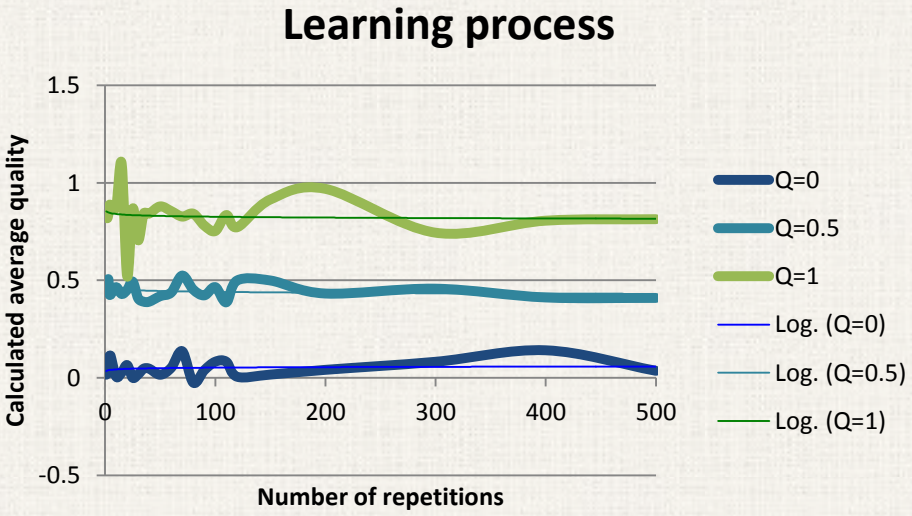


Methodology

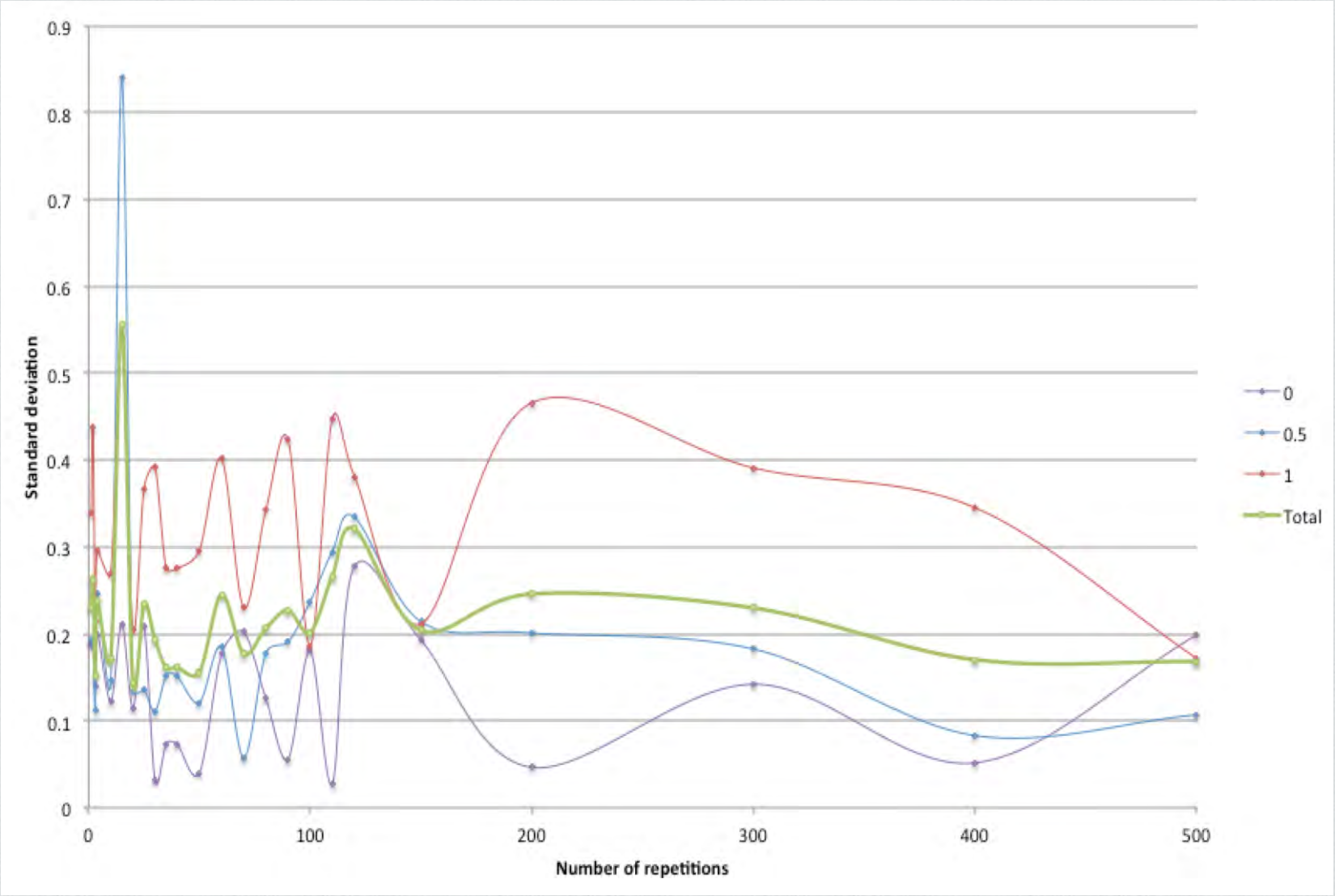
- about 50 species identified in diet samples
- about 250 observed taxa (incl. various development stages) in zooplankton nets samples
- optical properties of water (colour)
- physical properties of water



- Feeding fields were arbitrary assigned to the three classes of quality {0; 0.5; 1}, according to the structure of population.
- The structure of the neural network consists of the 252 neurons in the input layer.
- Data set is randomly divided into training set (75% of data) and control set (25% of data)



Estimation of the results against control set



Conclusions

Analysis of complex system is very often impossible to be performed in reasonable time. The limitations of traditional data analysis approach are fouling more when complexity of the environment and data heterogeneous nature express.

ANN can be used for rapid extraction of information from big data sets. This approach could provide system helpful to distinguish emerging areas for further scientific research and supporting decision systems in other areas of biological research based on Big Data.

Growing demand for data force development of autonomous measuring devices, speed up data processing and shorten time period between data acquisition and information available for use.

Growing volume of data, heterogeneity of data sources, force users to use more sophisticated tools for information retrieval. In the age of “Internet of Things”, things have to be smarter to “survive” in dynamically changing environment


THANK YOU !