Concept for the Dagstuhl Seminar

„Computer Science meets Ecology"

Benjamin Adams, University of Auckland, New Zealand
Gustau Camps-Valls, Universität de València, Spain
Joachim Denzler, University of Jena, German
Thomas Hickler, Senckenberg Biodiversity and Climate Research Centre (BiK-F) and Goethe-University Frankfurt, German
Birgitta König-Ries, University of Jena, Germany
Markus Reichstein, MPI for Biogeochemistry, Germany
Wolfgang Wägele, Zoological Research Museum Alexander König, Germany

General Introduction

In his pioneering work, Jim Gray identified the 4th scientific paradigm, arguing that modern science needs computer supported research. Recent developments in many scientific disciplines prove him right: Huge amounts of heterogeneous, unstructured and multisource data can now be collected routinely, sometimes in a fully automatic manner. Due to the development of computer hardware and sensors even new data modalities are readily available. The main difference to the general "big data" hype is that in science collecting data always has the intention to gain insights into processes and mechanisms, or in general to gain knowledge from data, typically motivated by some hypothesis. So far, the main challenge is to manage the explosive growth in size, complexity, and rates of data accumulation. On the one hand, it is easy to collect Tera-Bytes of data per minute. On the other hand, analysing even a fraction out of it still remains a big problem for scientists, companies and international organizations

A discipline that shows the potential but also the challenges of this 4th scientific paradigm is Ecology.

Ecology is the study of the interactions amongst organisms and with their physical environment. For a long time, ecological analyses have been realized locally both with respect to both the geographical and phenomenological area of investigation. Today, scientists are interested in quantifying ecological relations globally and can consider multiple dimensions of interactions between atmospheric, oceanic, and terrestrial processes. Due to the possibilities to record data all over the world, the increase of resolution and quality in recordings from, e.g., satellite platforms, and international efforts to document the global distribution of biodiversity, increasing availability of heterogeneous data sets via the World Wide Web and computing in the cloud, new opportunities arise. These data may enable us to answer questions that are of fundamental importance for the future of our planet. In short: ecology is one of those sciences, affected in a significant way by the tremendous increase in possibilities to collect and analyse data, and there is significant societal interest in taking advantage of these possibilities.

However, usually, scientists in ecology are not completely aware about current trends and new techniques in computer science that can support their daily work. Such support could consist in the management, integration, and (semi-)automatic analysis of resources, like experimental data, images, measurements, in the generation of useful metadata, cloud computing, distributed processing, etc. Ecoinformatics is regarded as an important supporting discipline by many ecologists. However, up to now, very few computer scientists are involved in this discipline; mostly ecoinformatics (or biodiversity informatics) is done by people with a strong background in e.g. ecology and a long (mostly self-taught) experience in data management. It lacks a strong connection to cutting-edge computer science research in order to profit from the results of this area. On the other hand, computer scientists know too little about the domain to be able to offer solutions to relevant problems and to identify potential research avenues.

Over the last few years, all of the authors have been involved in interdisciplinary settings and projects bringing together scientists from these different disciplines. The idea for the Dagstuhl seminar is the

result of numerous discussions in these projects identifiying the need for such a meeting. We all believe that a stronger bond between the disciplines that goes beyond viewing computer science as a “service provider” is of vital importance. The aim of the Dagstuhl seminar is to establish such links between (geo-)ecologists, ecoinformaticians and computer scientists.

In-depth description

In the following, we will look at the seminar topic from two perspectives. First, from the perspective of ecological research: Where would it profit from computer science? And second, from the perspective of computer science: where could it support ecological research and gain challenging research questions from such a collaboration? We will start with a rather general discussion, but then narrow each topic down to one rather specific problem. These specific problems will serve as crystallization points for discussions and working groups at the seminar.

One example discipline, where the 4th scientific paradigm may revolutionize the epistemic foundations could be ecology: Ecologists have been collecting data all over the world and organizational scales ranging from microscopic processes to global phenomena. For instance, latest developments in metagenomics have opened the possibility to prove the occurrence of species across a wide range of taxonomic hierarchies via “Environmental DNA” (Taberlet et al. 2012)[1] - several thousands of samples can be collected inwithin reasonable time frames. Satellite remote sensing data offer temporally continuous and spatially contiguous estimates of the states of land and aquatic ecosystems (e.g. Tuanmu and Jetz 2014)[2]. Monitoring biologically mediated fluxes of CO2 between land and atmosphere exchanges allow monitoring of ecological processes (Baldocchi 2014, [http://fluxnet.ornl.gov/])[3]. Soundscapes of birds (Kasten et al. 2012)[4] offer new ways to determine species diversity. All these examples show that novel observational methodologies are currently revolutionizing this branch of science. In all cases, the resulting data streams are heterogeneous and often unstructured, even when the same processes are observed by different groups, or over different regions of the world. Nevertheless, model building is heavily supported by the collected data. Furthermore, increasingly sophisticated models are developed, which are parameterized or calibrated with different sources of data (e.g. Hartig et al. 2012)[5] and demand very substantial computing power. Most information cannot be extracted from the data without computer support during the analysis, storage, access, distribution, visualization.

Besides typical “big-data” problems caused by volume, velocity, variety and veracity of data, there are more important challenges: providing access to the right data (and in an appropriate structure), to extract the relevant information considering redundancies and knowledge, and to develop computationally efficient ways for data model linkages.

Therefore, at least three general topic areas can be identified:

- **Obtaining and Preserving Data**: This includes automatic monitoring schemes, automatic interpretation of e.g. remote sensing or image data, sampling bias analysis and gap-filling, data quality management, synthesis and curation. A particular challenge is the huge heterogeneity of data ranging from sequence data to remote sensing images, and from digitized natural history museum collections to manually collected observation data to audio files capturing acoustic

---

1 Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. Molecular Ecology, 21(8), 1789-1793.

2 Tuanmu, M.-N. and W. Jetz. 2014. A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. Global Ecology and Biogeography 23:1031-1045.

3 Baldocchi, D. (2014). Measuring fluxes of trace gases and energy between ecosystems and the atmosphere–the state and future of the eddy covariance method. *Global change biology*, *20*(12), 3600-3609.

4 Kasten, E. P., Gage, S. H., Fox, J., & Joo, W. (2012). The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology. *Ecological Informatics*, *12*, 50-67.

5 Hartig, F., Dyke, J., Hickler, T., Higgins, S. I., O’Hara, R. B., Scheiter, S., & Huth, A. (2012). Connecting dynamic vegetation models to data–an inverse perspective. *Journal of Biogeography*, *39*(12), 2240-2252.

diversity. A second important challenge is the increasing volume of such data evident already for remote sensing data and for sequence and related data, where new techniques and rapidly sinking prices lead to an explosion in data volume.

- **Pattern-recognition in highly dimensional and geo-tagged data sets**: The field involves developing sound and efficient algorithms able to capture structure and feature relations in empirical data, and mostly involve finding groups (*clustering*), anomalies (*detection*), automatic categorization and prediction (*classification/regression)*, and learning proper representation spaces (*visualization*) of generally unstructured, heterogeneous, multimodal data streams where quantifying uncertainty is mandatory.
- **Model development and Model-Data-Confrontation** (see e.g. Rillig et al 2015)[6]: This includes dealing with sampling bias and scale issues, methods for fitting model to data, scaling and parallelization for cluster or cloud computing.

Some areas of computer science that can contribute to these topic areas and derive research questions from them are:

- **Data and Model Management:** Data Management is certainly the part of computer science that has been used in ecology the longest and is one of the major focus areas of Ecoinformatics. Numerous data management platforms and workflow environments suitable for ecological data have been developed focussing on different stages of data management from data collection in the field (supported, e.g., by smartphone applications) to long term preservation of data. As major challenge remains the seamless integration of data management tasks in the usual workflows of the researchers. A key part of this challenge is identifying what data are useful for particular types of analysis and purposes. Capturing the pragmatic relationships between data and their use, including the tasks and methods for which data have been successfully used, remains a relatively unexplored area of research. Additionally, platforms are needed that can deal with the vast heterogeneity of the data and the expected future huge volumes of data. Increasingly, ecological data of high spatial and temporal resolution can be crowdsourced and streamed from sensors of variable quality, and despite the great potential for this data to be used for ecological analysis the heterogeneity of sources creates open research challenges for data management. New challenges arise also from the vast amount and poor quality of sequencing data; requiring new bioinformatics techniques to handle and preserve the data.

- **Data Integration:** The ability to integrate data is vital for ecological research. However, such integration is hampered by a number of factors where the application of modern approaches from computer science will be helpful. Over the last few years, considerable effort went into the development of formal, machine-readable taxonomies and metadata standards; the use of ontologies is relatively widespread. This requires ontology matching and modularisation. Often, integration problems are present at the *instance* rather than the *schema* level. Approaches for duplicate detection and data quality assurance are needed here. Provenance and uncertainty management are needed for gaining meaningful results from the integrated data. This area poses a real challenge for computer science since the information that needs to be encoded goes well beyond the rather simplistic e.g. simple probability distributions commonly used today.

- **Modern techniques from Computer Vision, Pattern Recognition, Data Mining and Machine Learning:** Over the last years, computer vision research already tackled problems that are of high relevance for ecological research as well. One example is the analysis of remote sensing data, which forms one of the basis for global analysis of terrestrial processes, for which several modern methods for automatic processing exist, for example, semantic segmentation. Other examples include large scale analysis of the distribution of animals, plants, and (increasingly genetically derived) populations (e.g. Balint et al. 2012)[7], whereby the data often suffers from extremely biased (in space and time) sampling (Meyer et al. 2015)[8] and few data

---

6 Rillig, M. C., Kiessling, W., Borsch, T., Gessler, A., Greenwood, A. D., Hofer, H., ... & Jeltsch, F. (2015). Biodiversity research: data without theory—theory without data. *Frontiers in Ecology and Evolution*, *3*, 20.

7 Balint, M., S. Domisch, C. H. M. Engelhardt, P. Haase, S. Lehrian, J. Sauer, K. Theissinger, S. U. Pauls, and C. Nowak. 2011. Cryptic biodiversity loss linked to global climate change. Nature Clim. Change 1:313-318.

8 Meyer, C., H. Kreft, R. Guralnick, and W. Jetz. 2015. Global priorities for an effective information basis of biodiversity distributions. Nature Communicatins 6:8221.

are available for organism groups where it is difficult to identify the species. Several computer-based methods have recently been developed to support ecological research. These include object recognition software for e.g. plants. However, since those objects offer not just very challenging problems but also call for new methods, that lead to the area of fine-grained recognition. Although today's state of the art systems achieve only recognition rates of 70-80%, in some scenarios machine vision systems are already better than the inexperienced user. Together with techniques from machine learning, like active learning (i.e. keeping the human in the loop as in recent activities[9]), and novelty detection, i.e. detecting if a new object or event is observed, preliminary life-long learning systems are currently under development. In such an iterative manner of building recognition systems and improving performance by specific feedback of users, it is expected that performance of automatic analysis of animals or plants from images and videos will reach the threshold that almost fully automatic observation of our environment will be possible. Having such methods will bring researchers from ecology closer to measurement stations equipped with cameras that could record the environment at a level that has not been possible before. Finally, computer vision techniques might support digitalization of existing ecological data sets.
Besides computer vision, modern machine learning techniques will play an important role in the future of ecology data analysis as well. For example, analysing huge amount of data by the human can be supported by automatic clustering into relevant groups. Dimensionality reduction methods, like non-linear or kernel PCA offer new potentials in data pre-processing. Detecting the unexpected, i.e. interesting in data streams can be supported by automatic analysis using novelty and anomaly detection methods, and thus can serve as clustering in the sense of reduction of human efforts to the most important parts of data streams.
Finally, machine learning techniques in general might help to *make the invisible visible* by solving regression problems using training data. Such mappings from input data to output might be the basis for future decision based on measurement. Estimation of bio-geo-chemical parameters using advanced retrieval methods currently provide accurate time-resolved estimations, but advances on uncertainty estimation (going beyond point-wise predictions to meaningful confidence intervals) and knowledge discovery capabilities (i.e. ranking input features to understand the underlying bio-physical processes) are still needed.

- **High-Performance and Cloud Computing** (bring computing power to the data): The growing amount of data and increasingly complex models require new ways of processing. It is no longer feasible – as is done today – to select data from some online source and download it for local processing. Rather than launching the data to the algorithms, the trend is to launch the algorithms to the data. Here, approaches for function shipping and/or parallelisation can be helpful and are successfully applied, e.g., by GBIF for (re-)ingest of data or in the Map of Life project. Ecological information analysis and modeling largely remains restricted in the size and complexity of problems that can be addressed due to lack of research into up-scaling ecological algorithms (e.g., analysis of ecosystem connectivity) from desktop applications to high performance computing. This requires a systematic approach of mapping ecological data structures and algorithms to well-understood techniques of parallel computation and communication that have been identified by the high-performance computing research community. Identification of how environmental simulations and analyses map to compositions of these well-established scientific computing patterns will be a necessary outcome of this research. Another challenge is model design to best meet recent advances in computer science. This includes, e.g., re-designing models to run on energy-efficient graphics processing units (GPUs). Running models on GPUs instead of conventional CPUs can decrease electricity costs very substantially[10].

Quite obviously, it will not be possible to address all of these issues within a week. Therefore, during the seminar, we will bring these two dimensions, i.e., the ecological and the computer science perspective together based on three concrete sample problems.

Example 1: Biodiversity Weather Stations/Automated Long-Term Monitoring: Traditionally, data in ecological research have been collected manually on a rather small scale. For instance, the traditional approach to analysing species richness in a tropical rainforest is to select a plot of manageable size and send scientists (typically PhD students) there, to map the species that occur on this plot. This approach

9 EU COST Action: „Mapping and the citizen sensor", http://www.citizensensor-cost.eu/

10 see, e.g. https://csc.uni-frankfurt.de/index.php?id=loewe-hw

has several drawbacks: First, it is extremely expensive. Second, since neither money nor personnel are unlimited resources, it scales poorly. Third, the quality of the result depends a lot on the expertise of the scientists in the field. The acknowledgements of a recent paper on tree flora in the Amazonian that aims at developing a large scale model and uses data from around 2000 plots, e.g., states "This paper is the result of the work of hundreds of different scientists and research institutions in the Amazon over the past 80 years."[11] Basically the same drawbacks exist for other types of data collection in ecological research. For instance, in the Biodiversity Exploratories, insect populations on research plots are determined by installing window traps in the field which collect insects. The species are then determined by manual analysis by large numbers of student helpers analysing every caught individual.

In the future, such monitoring schemes could be automated. Technologies like DNA-barcoding of environmental samples, visual and acoustic identification of animals, identification of plants via emitted chemicals are currently being combined to build an Automated Multisensor Station for Monitoring of Species Diversity (AMMOD). The AMMOD requires a combination of image and sound recognition, machine-readable reference libraries for genetic und biochemical markers, images and sounds, the storage and sorting of a large amounts of data and finally, when several stations are combined, modelling of species distribution in landscapes.

Example 2: Global Change Ecology: Key challenges for Ecology in our Global Change era are i.) to understand and predict the geographical distributions and abundances of species and populations and ii.) to improve our understanding of the role of biodiversity for the functioning of ecosystems (e.g. Maestre et al. 2012)[12] and their supply of services to the human society under Global Change. Addressing these challenges implies dealing with spatially biased data, e.g. for the occurrence of species, and integrating various data types on where species or populations occur, which functional traits they have, the environment in which they live (e.g. climate, soil types, land cover) and ecosystem processes, such as biomass productivity and carbon cycling (Pereira et al. 2013) [13]. Thus, it is necessary to integrate multiple types of data from the biological and geosciences, ranging from genetic data characterising populations or species to satellite-derived estimates of land cover change (e.g. Hansen et al. 2013) [14]. Thereby, the genetic and satellite data, in particular, have reached levels of complexity and sizes, which are sometimes beyond the capacities of normal desktop computers. Instead, massive RAM or parallel cluster computing are increasingly necessary to handle the data, even for relatively simple analyses. For more complex model-data fusion techniques, such as hierarchical Bayesian modelling, computational capacities are still highly limiting ecological research.

Example 3: Modelling ecosystem and Earth system processes: Modelling now also plays a crucial role for ecosystem science from the local to global scale. More and more ecological processes are currently integrated into so-called Earth System models, which integrate climate models with biosphere models (Bonan et al 2003)[15] (Cox et al 2000)[16]. Yet, there is a large uncertainty in future model predictions for these dynamic systems (Heimann and Reichstein 2008)[17]. One challenge now is to provide observation-based constraints which can confine future model behaviour. We need to understand better which patterns of the observations provide robust constraints for models. Hence, we

---

11 Ter Steege, H., Pitman, N. C., Sabatier, D., Baraloto, C., Salomão, R. P., Guevara, J. E., ... & Fine, P. V. (2013). Hyperdominance in the Amazonian tree flora. *Science*, *342*(6156), 1243092.

12 Maestre, F. T., J. L. Quero, N. J. Gotelli et al. 2012. Plant Species Richness and Ecosystem Multifunctionality in Global Drylands. Science 335:214-218.

13 Pereira, H. M., S. Ferrier, M. Walters, G. Geller, R. Jongman, R. Scholes, M. W. Bruford, N. Brummitt, S. Butchart, and A. Cardoso. 2013. Essential biodiversity variables. Science 339:277-278.

14 Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science 342:850-853.

15 Bonan, G. B., S. Levis, S. Sitch, M. Vertenstein and K. W. Oleson (2003). "A dynamic global vegetation model for use with climate models: concepts and description of simulated vegetation dynamics." Global Change Biology **9**(11): 1543-1566.

16 Cox, P. M., R. A. Betts, C. D. Jones, S. A. Spall and I. J. Totterdell (2000). "Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model." Nature **408**: 184-187.

17 Heimann, M. and M. Reichstein (2008). "Terrestrial ecosystem carbon dynamics and climate feedbacks." Nature **451**: 289-292.

need to move away from simple model-data comparisons, to pattern-oriented model evaluation, calibration and interpretation in a system-oriented way (Reichstein and Beer 2008)[18]. Examples of this include approximate Bayesian computation (Vrugt et al 2013)[19] and the concept of emerging constraints (Cox et al 2013)[20]. As a variety of data types, ranging from leaf-level measurement of photosynthesis to satellite-derived estimates of forest biomass, can be used to parameterize and constrain ecosystem models, such models might in the future rather serve as process-based linkages between multiple data types, instead of just being parameterized and tested with individual data sets at a time.

Organization of the seminar, objectives and results

To foster collaboration, most of the seminar will be organized in interdisciplinary wo*rking groups.*
We envision the following preliminary schedule:

Monday:
09:00 – 09:30 Welcome
09:30 – 11:00 Intro by Participants (2 min each)
11:00 – lunch Intro talk to Example 1
14:00 – 15:00 Intro talk to Example 2
15:30 – 16:30 Intro talk to Example 3
16:30 – 17:15 Intro to structure of working groups, organisation of working groups
17:15 – 18:00 Setup meeting of working groups
20:00 – 21:00 Poster session, then wine cellar

The introductory talks to the three examples will be prepared prior to the seminar by small, interdisciplinary groups of participants and organizers.

Tuesday:
Working group day
Evening: Tool demo session (see below for details)

Wednesday:
Morning: reports from working groups
Afternoon: excursion

Thursday:
Morning: plenary: Identification of further working groups
Depending on participants' interests, these working groups could focus on additional example scenarios provided by some of the participants or delve deeper into individual aspects identified in the first round of working groups.
Remainder of day: work in working groups

Friday:
Report from working groups
Assignment of homework (see below "Objectives")

Organisation of working groups
The first round of working groups shall be organized along the following lines:
Each working group will be based on one of the example problems identified in this proposal and further elaborated in one of the introductory talks.
The working group shall then:

---

[18] Reichstein, M. and C. Beer (2008). "Soil respiration across scales: The importance of a model-data integration framework for data interpretation." Journal of Plant Nutrition and Soil Science **171**(3): 344-354.

[19] Vrugt, Jasper A., and Mojtaba Sadegh. "Toward diagnostic model calibratn and evaluation: Approximate Bayesian computation." Water Resources Research 49.7 (2013): 4335-4345.

[20] Cox, Peter M., David Pearson, Ben B. Booth, Pierre Friedlingstein, Chris Huntingford, Chris D. Jones, and Catherine M. Luke. "Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability." Nature 494, no. 7437 (2013): 341-344.

- identify data; data types, characteristics of data relevant to this problem
- identify what scientists are interested in (classification, ...)
- identify suitable tools to solve these problems; where they exist: demo in evening session (e.g. Map of Life, GFBio,);
- identify gaps and needs for further research.
- identify possible funding sources for such research.

The seminar has **two main objectives**:

1. Joint authoring of a book on the state of the art and challenges in the intersection of computer science and ecology. This book shall be based on the results of the working groups. Based on the example scenarios it will introduce three important classes of approaches in Ecology. For these, it will provide an introduction to available tools, and will outline challenges for future research. Such a book can serve as a handbook for ecologists needing computer science but also as a roadmap for future research activities.
2. Define project ideas for cooperation between Computer Scientists and Ecologists and where possible identify suitable funding schemes.