Report from Dagstuhl Seminar 17091

Computer Science Meets Ecology

Edited by

Gustau Camps-Valls[1], Thomas Hickler[2], and Birgitta König-Ries[3]

1 **Universitat de València, ES,** gustau.camps@uv.es
2 **Senckenberg Research Centre, DE,** thomas.hickler@senckenberg.de
3 **Universität Jena & German Centre for Integrative Biodiversity Research (iDiv), DE,** birgitta.koenig-ries@uni-jena.de

Abstract

This report summarizes the program and main outcomes of the Dagstuhl Seminar 17091 entitled "Computer Science Meets Ecology". Ecology is a discipline that poses many challenging problems involving big data collection, provenance and integration, as well as difficulties in data analysis, prediction and understanding. All these issues are precisely the arena where computer science is concerned. The seminar motivation was rooted in the belief that ecology could largely benefit from modern computer science. The seminar attracted scientists from both fields who discussed important topics in ecology (e.g. botany, animal science, biogeochemistry) and how to approach them with machine learning, computer vision, pattern recognition and data mining. A set of specific problems and techniques were treated, and the main building blocks were set up. The important topics of education, outreach, data and models accessibility were also touched upon. The seminar proposed a distinctive perspective by promoting cross-fertilization in a unique environment and a unique set of individuals.

Seminar February 26 to March 3, 2017 – http://www.dagstuhl.de/17091
1998 ACM Subject Classification D.2.12 Interoperability, J.3 Life and Medical Sciences – Biology and genetics, I.6.5 Model Validation and Analysis
Keywords and phrases ecology, biodiversity, earth observation, earth system, remote sensing, computer science, citizen science, big data, data integration, modeling, semantics, society
Digital Object Identifier 10.4230/DagRep.7.2.109
Edited in cooperation with Ivaylo Kostadinov

1 Executive Summary

Gustau Camps-Valls
Joachim Denzler
Thomas Hickler
Birgitta König-Ries
Markus Reichstein

Ecology is a discipline that shows clearly the potential but also the challenges of computer supported research described as the 4th scientific paradigm by Jim Gray. It is increasingly data driven, yet suffers from hurdles in data collection, quality assurance, provenance, integration, and analysis.

We believe that ecology could profit from modern computer science methods to overcome these hurdles. However, usually, scientists in ecology are not completely aware of 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.

Motivated by our belief that a stronger bond between the disciplines that goes beyond viewing computer science as a "service provider" is of vital importance, we proposed this Dagstuhl seminar. The aim of the Dagstuhl seminar was to establish such links between (geo-)ecologists, ecoinformaticians and computer scientists.

The seminar: perspective and self-evaluation

Before the seminar. It turned out that it was not an easy task to motivate non-computer scientists to attend the seminar. For many, travel costs were a hurdle ultimately preventing attendance. This resulted in an unusually large number of declined invitations (often accompanied by "I would love to attend, but..." emails.

Despite these initial problems, we believe that the aim to start building links among the communities was reached at the seminar: We had fruitful discussions in numerous working groups resulting in some very concrete plans for future work.

Organization of the seminar. A total of 27 attendees gathered at the seminar. The wide variety of expertise and backgrounds constituted an initial challenge for the organization. The agenda considered a first round of presentations of the individuals and their research groups with a clear outline and items to treat (personal background, Research Areas/Interests, prospective links to „Computer Science meets Ecology" seminar). After this, the main topics of interest for a wide audience were designed: essentially, three breakout groups were set up in the very first day of the meeting. Over the course of the seminar, these groups were adjusted, split up, or merged, several times. This resulted in quite a number of topics being touched upon with concrete results ranging from a working example for the application of a new method to a modeling problem to concrete plans for publications, a proposal and follow-up activities. Reports on these groups were given in the plenary session, and can be found in this report.

Broad results of the seminar. Results from the seminar can be categorized in three types: (i) collaborative and networking, as new joint works on specific topics came out of the meeting; (ii) knowledge transfer between fields, as computer scientists learned about the main problems in ecology involving data, while ecologists became aware of what kind of problems data scientists can solve nowadays; and (iii) educational, as several young PhD students and postdocs attended and participated in high level discussions.

Conclusions. The seminar brought together top scientists in the fields of ecology and computer science. The group of individuals was largely interdisciplinary, with a wide range of interests and expertises in each community too: from botany and animal science, to machine learning and computer vision. The seminar was organized in two main types of

sessions: plenary and working group sessions to better focus on particular topics. Interesting developments and discussions took place in both, and a high level of cross-fertilization and future collaborations was initiated. On top of this, there was a broad consensus among the participants that the seminar should be the start of a series of yearly or bi-yearly meetings. We hope that the success of this first seminar will encourage broader participation in follow-up activities.

2 Table of Contents

3 Introduction

3.1 Why Computer Science needs to meet Ecology

Gustau Camps-Valls (University of Valencia, ES), Benjamin Adams (University of Auckland, NZ), Joachim Denzler (Universität Jena, DE), Thomas Hickler (Senckenberg Research Centre, DE), Birgitta König-Ries (Universität Jena, DE), Markus Reichstein (MPI für Biogeochemistry – Jena, DE), and Johann Wolfgang Wägele (ZFMK – Bonn, DE)

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 terabytes 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.

In the following, we will look at the 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.

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" [1] – several thousands of samples can be collected within reasonable time frames. Satellite remote sensing

data offer temporally continuous and spatially contiguous estimates of the states of land and aquatic ecosystems [2]. Monitoring biologically mediated fluxes of CO2 between land and atmosphere exchanges allow monitoring of ecological processes [3] (http://fluxnet.ornl.gov/). Soundscapes of birds [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 [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 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. [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 [7], whereby the data often suffers from extremely biased (in space and time) sampling [8] and few data 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), 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.

In order to provide a more detailed understanding of some of the problems involved, let us have a look at three concrete examples that highlight different problem areas and different possible links between computer science and ecology.

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 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", 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 and biochemical markers, images and sounds, the storage and sorting of a large amounts of data and finally, when several stations are combined, modeling 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 [11] 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 [12]. 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 [13]. 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 modeling, 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 [14, 15]. Yet, there is a large uncertainty in future model predictions for these dynamic systems [16]. 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 need to move away from simple model-data comparisons, to pattern-oriented model evaluation, calibration and interpretation in a system-oriented way [17]. Examples of this include approximate Bayesian computation [18] and the concept of emerging constraints [19]. 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. At lot of analogies between video data and dynamic Earth System data have been identified and ideas generated of how applying methods of one domain in the other.

In summary, we strongly believe that a closer interaction between ecologists and computer scientists is needed to tackle the challenges in Ecology and that both disciplines will profit from such interaction: Ecologists will be able to solve problems currently beyond their reach. Computer Scientists will be exposed to a challenging set of real-world problems requiring the development of new methods and approaches.

References

1. P. Taberlet, E. Coissac, M. Hajibabaei, L. H. Rieseberg, "Environmental DNA", Molecular Ecology, 21(8), 1789–1793, 2012.
2. M.-N. Tuanmu, W. Jetz, "A global 1-km consensus land-cover product for biodiversity and ecosystem modelling", Global Ecology and Biogeography 23:1031–1045, 2014.
3. D. Baldocchi, "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, 2014.
4. E. P. Kasten, S. H. Gage, J. Fox, W. Joo, "The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology", Ecological Informatics, 12, 50–67, 2012.
5. F. Hartig, J. Dyke, T. Hickler, S. I. Higgins, R. B. O'Hara, S. Scheiter, A. Huth, "Connecting dynamic vegetation models to data – an inverse perspective", Journal of Biogeography, 39(12), 2240–2252, 2012.
6. M. C. Rillig, W. Kiessling, T. Borsch, A. Gessler, A. D. Greenwood, H. Hofer, ..., F. Jeltsch, "Biodiversity research: data without theory – theory without data", Frontiers in Ecology and Evolution, 3, 20, 2015.
7. M. Balint, M. S. Domisch, C. H. M. Engelhardt, P. Haase, S. Lehrian, J. Sauer, K. Theissinger, S. U. Pauls, and C. Nowak, "Cryptic biodiversity loss linked to global climate change", Nature Clim. Change, 1:313–318, 2011.
8. C. Meyer, H. Kreft, R. Guralnick, W. Jetz, "Global priorities for an effective information basis of biodiversity distributions", Nature Communicatins, 6:8221, 2015.
9. EU COST Action: "Mapping and the citizen sensor", http://www.citizensensor-cost.eu/
10. H. Ter Steege, N. C. Pitman, D. Sabatier, C. Baraloto, R. P. Salomão, J. E. Guevara, ..., P. V. Fine, "Hyperdominance in the Amazonian tree flora", Science, 342(6156), 1243092, 2013.
11. F. T. Maestre, J. L. Quero, N. J. Gotelli et al., "Plant Species Richness and Ecosystem Multifunctionality in Global Drylands", Science, 335:214–218, 2012.
12. H. M. Pereira, S. Ferrier, M. Walters, G. Geller, R. Jongman, R. Scholes, M. W. Bruford, N. Brummitt, S. Butchart, A. Cardoso, "Essential biodiversity variables", Science, 339:277–278, 2013.
13. M. C. Hansen, 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, J. R. G. Townshend, "High-Resolution Global Maps of 21st-Century Forest Cover Change", Science, 342:850–853, 2013.
14. G. B. Bonan, S. Levis, S. Sitch, M. Vertenstein, K. W. Oleson, "A dynamic global vegetation model for use with climate models: concepts and description of simulated vegetation dynamics", Global Change Biology, 9(11):1543–1566, 2003.

15 P. M. Cox, R. A. Betts, C. D. Jones, S. A. Spall, I. J. Totterdell, "Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model", Nature, 408: 184–187, 2000.
16 M. Heimann, M. Reichstein, "Terrestrial ecosystem carbon dynamics and climate feedbacks", Nature, 451: 289–292, 2008.
17 M. Reichstein, C. Beer, "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, 2008.
18 J. A. Vrugt, M. Sadegh, "Toward diagnostic model calibratn and evaluation: Approximate Bayesian computation", Water Resources Research 49.7 (2013): 4335–4345.
19 P. M. Cox, D. Pearson, B. B. Booth, P. Friedlingstein, C. Huntingford, C. D. Jones, C. M. Luke, "Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability", Nature 494, no. 7437 (2013): 341–344.

4 Overview of Talks

4.1 Life-long Learning with Applications in Monitoring Biodiversity

Joachim Denzler (Universität Jena, DE)

Joint work of Joachim Denzler, Erik Rodner, Alexander Freytag, Christoph Käding, Marcel Simon, Clemens-Alexander Brust

Most of today's impressive results in computer vision and machine learning arise from two major changes during the past 20 years: Firstly, the increased performance of hardware together with the advent of powerful graphical processing units (GPU) applied in scientific computing beyond pure displaying. Secondly, the huge amount of, in part, annotated image data provided by today's generation of Facebook and Twitter users and available easily over databases (e.g., Flickr) and/or search engines. Consequently, tasks like face recognition and identification can be solved using powerful methods, like Convolutional Neural Networks [13], and millions of face images for training.

For visual monitoring of biodiversity, for example, to keep track of species distribution of certain mammals, no such databases or collection of annotated or even weakly annotated images exists at a size such that systems can be directly trained. This first challenge, the collection of training data bases for computer vision algorithms, links directly to the citizen science activities. We need to motivate people to also share their annotated images of animals, insects, and other species, or at least to help collecting such databases.

Although training data is now one limiting factor for visual monitoring at a certain level of quality, there are several other and equally important challenges from the computer vision and machine learning perspective:

1. Number of species to be distinguished: although current computer vision systems can differentiate between up to 10.000 different categories (see ImageNet [2]), this number is far from being sufficient for the number of species to be expected in Germany. In addition, the classification of such many different objects has been demonstrated at the category level only, i.e. to differentiate dogs from cats, but not certain races of dogs and cats.
2. Generic classifiers: although certain systems already exist for analyzing images of moths, chimpanzees, or other specific class of objects [12, 4], those systems have been carefully developed using handcrafted and optimized features and individual domain knowledge. At

present, it seems not possible that such specialized system can be individually developed for all the different classes of animals and insects to be monitored. Thus, there is the need for generic classifiers that learn their feature representation from data, at best in an unsupervised manner [1, 11].

3. Fine-grained recognition: As mentioned earlier, most computer vision system for classification of objects in an image, are already powerful if it comes to distinction of categories, like cups, cars, dogs, etc. Within category classification, i.e. the distinction between a Great Spotted Woodpecker and a Middle Spotted Woodpecker, it is a much more challenging problem, and currently the focus in fine-grained recognition. For certain categories, like birds, cats, and dogs, solutions are already available [3]. However, there is still a generic method missing that identifies the relevant, visual parts of objects that allow reliable classification within a category of visually similar species.
4. Detection of the unexpected: Today's machine learning system work under the closed-world assumption, i.e. they will map any input image to one of the known classes. Species not known to the system will not be correctly classified, but even worse might be wrongly assigned to a known class. Since the unexpected is often the driver of progress in science, such wrong assignments might prevent some insight in the monitored ecosystem. Thus, methods for novelty and anomaly detection is another big challenge to not miss the probably important insight from unexpected observations [5, 10].
5. Keeping the human in the loop: Today, it cannot be expected that automatic monitoring systems will work error-free from scratch. The challenge arising from difficult and changing recoding conditions in the wild, hiding and only partially visible animals will result in erroneous assignment or even misses of objects visible in the image for the human. Thus, acceptance of such systems in the monitoring community will heavily depend on reliability of the automatically generated statistics and properties of the observed species. Consequently, one additional challenge is to provide a feedback mechanism from the machine to the human, to report about uncertain or undetermined results. However, the feedback from the human to the machine is equally important by correcting results or adding additional information for refinement and optimization of the automatic system [7].

In summary, we believe that automatic visual monitoring should be framed in a life-long learning cycle that has been recently applied to monitor mammals in Portugal [6]. The key ingredients are initial, generic classifier, for example, powerful CNN architectures [13], active learning to reduce costly annotation effort by experts [8, 5], fine-grained recognition to differentiate between visually very similar species [3], and efficient incremental update of the classifier's model over time [9]. For most of these challenges, initial solutions exist. Building first visual monitoring systems, possibly for a restricted area or set of species, will definitely help to improve all parts over time, if biodiversity and computer vision researchers are working closely together.

References

1 A. Freytag, E. Rodner, M. Simon, A. Loos, H. Kühl, J. Denzler, "Chimpanzee Faces in the Wild: Log-Euclidean CNNs for Predicting Identities and Attributes of Primates", German Conference on Pattern Recognition (GCPR), 51–63, 2016.

2 J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database", International Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

3 M. Simon, E. Rodner, "Neural Activation Constellations: Unsupervised Part Model Discovery with Convolutional Networks", International Conference on Computer Vision (ICCV), 2016.
4 J. Balfer, F. Schöler, V. Steinhage, "Semantic Skeletonization for Structural Plant Analysis", 7th Intern. Conf. on Functional-Structural Plant Models (FSPM 2013), Saariselkä, Finland, June 9-14, 2013.
5 Ch. Käding, A. Freytag, E. Rodner, P. Bodesheim, J. Denzler, "Active Learning and Discovery of Object Categories in the Presence of Unnameable Instances", IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 4343–4352, 2015.
6 Ch. Käding, E. Rodner, A. Freytag, J. Denzler, "Watch, Ask, Learn, and Improve: A Lifelong Learning Cycle for Visual Recognition", European Symposium on Artificial Neural Networks (ESANN), 2016.
7 Ch. Käding, A. Freytag, E. Rodner, A. Perino, J. Denzler, "Large-scale Active Learning with Approximated Expected Model Output Changes", German Conference on Pattern Recognition (GCPR), 179–191, 2016.
8 A. Freytag, E. Rodner, J. Denzler, "Selecting Influential Examples: Active Learning with Expected Model Output Changes", European Conference on Computer Vision (ECCV), 562–577, 2014.
9 E. Rodner, A. Freytag, P. Bodesheim, B. Fröhlich, J. Denzler, "Large-Scale Gaussian Process Inference with Generalized Histogram Intersection Kernels for Visual Recognition Tasks", International Journal of Computer Vision (IJCV), 253–280, 2017.
10 P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, J. Denzler, "Kernel Null Space Methods for Novelty Detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3374–3381, 2013.
11 E. Rodner, M. Simon, G. Brehm, S. Pietsch, J. W. Wägele, J. Denzler, "Fine-grained Recognition Datasets for Biodiversity Analysis", CVPR Workshop on Fine-grained Visual Classification (CVPR-WS), 2015.
12 A. Loos, A. Ernst, "An automated chimpanzee identification system using face detection and recognition", EURASIP journal on image and video processing, 2013.
13 Y. LeCun, J. Bengio, G. Hinton, "Deep Learning", Nature 521, 436–444. doi:10.1038/nature14, 2015.

4.2 Systematic Evaluation of Land Surface Models Using the International Land Model Benchmarking (ILAMB) Package

Forrest Hoffman (Oak Ridge National Laboratory, US), Nathaniel O. Collier, Gretchen Keppel-Aleks, Charles D. Koven, David M. Lawrence, Mingquan Mu, James T. Randerson, William W. Riley

Main reference F. M. Hoffman et al., "2016 International Land Model Benchmarking (ILAMB) Workshop Report", Technical Report DOE/SC-0186, 2017.
URL http://dx.doi.org/10.2172/1330803
URL http://dx.doi.org/10.18139/ILAMB.v002.00/1251621

As Earth system models (ESMs) become increasingly complex, there is a growing need for comprehensive and multi-faceted evaluation of model predictions. To advance understanding of biogeochemical processes and their interactions with hydrology and climate under conditions of increasing atmospheric carbon dioxide, new methods are needed that use observations to

constrain model predictions, inform model development, and identify needed measurements and field experiments. Improved process parameterizations are needed to constrain energy and water predictions in land surface models and better representations of biogeochemistry – climate feedbacks and ecosystem processes in ESMs are essential for reducing uncertainties associated with projections of climate change during the remainder of the 21st century. The International Land Model Benchmarking (ILAMB) project seeks to 1) develop internationally accepted benchmarks for land model performance, 2) promote use of benchmarks for model intercomparison projects, 3) strengthen linkages between experimental, remote sensing, and modeling communities, and 4) support the design and development of an open source benchmarking software system. Leveraging work on past model evaluation studies, we have developed two generations of such benchmarking software packages that assess model fidelity on 24 variables in four categories from about 45 data sets; produce graphical global-, regional-, and site-level diagnostics; and provide a hierarchical scoring system. The ILAMBv2 package, publicly released in May 2016, has become an integral part of model verification workflow for rapid model development and calibration cycles for the U.S. Department of Energy's Accelerated Climate Modeling for Energy (ACME) model and the Community Earth System Model (CESM). We will present results from model analysis using the ILAMB packages, discuss techniques for routine model evaluation, propose coordinated evaluation of the Sixth Phase of the Coupled Model Intercomparison Project (CMIP6) output, and describe new metrics that integrate across carbon, surface energy, hydrology, and land use disciplines.

4.3 BExIS 2 – An open source data management platform for collaborative projects in Biodiversity Research and beyond

Birgitta König-Ries (Universität Jena, DE)

URL http://bexis2.uni-jena.de

In many collaborative projects, there is a strong need for data preservation and sharing. BExIS 2 is an open source data management platform that meets these needs and supports data management throughout the entire data lifecycle. It is a modular platform that can easily be adapted to the specific needs of particular projects with respect to, e.g., access rights, data structure, or metadata schema used. Further information including an online demo and a download link can be found on the BExIS 2 website: http://bexis2.uni-jena.de.

4.4 Real time monitoring of vegetation phenology with the PhenoCam network

Andrew Richardson (Harvard University – Cambridge, US)

Phenology – the seasonal rhythms of plants and animals – has been shown to be a robust integrator of the effects of year-to-year climate variability and longer-term climate change on natural systems. At the level of ecosystems, phenology is important because it influences

productivity, carbon sequestration, nutrient cycling, and feedbacks to the atmosphere and climate system.

There is a demonstrated need to better document biological responses to a changing world, and improved phenological monitoring will contribute to achieving this goal. In this talk, I will describe a collaborative research network called "PhenoCam" (http://phenocam.sr.unh.edu/). PhenoCam uses networked digital cameras – webcams – for phenological monitoring in a range of ecosystems (almost 400 sites, and 750+ site-years of archived data) across the North American continent. Images are captured every 30 minutes, uploaded to the PhenoCam server for display in real-time, and processed to yield quantitative measures of vegetation "greenness." I will conclude by talking about some of the challenges we face with managing this ever-expanding image archive.

4.5 New technologies for biodiversity monitoring

Johann Wolfgang Wägele (ZFMK – Bonn, DE)

Biodiversity is one of the most valuable resources of our planet. With possibly more than 10 million living species and most of these still unknown to science, the biosphere of our planet guarantees future generations a wealth of hitherto untapped genetic resources, which are relevant for food production, medicine, bioenergy production, and life-supporting ecosystem functions. In contrast to global warming, a steady loss of biodiversity is irreversible and leads to an impoverished world that will not recover its original richness within the next 5 million years. Already more than 20 years ago the large-scale destruction of habitats and losses of biodiversity alarmed researchers and policy makers. Until today, the biodiversity crisis is accelerating and a trend reversal is not achievable with political treaties and resolutions solely. One important reason is the lack of reliable high resolution, large scale data. Such data are needed as a basis for informed decisions, to analyze causes of local extinctions, to prove that trends are really happening, to model scenarios that explain ongoing changes and that can predict future processes, and to define actions based on scientific information. In analogy to climate scientists, who were able to raise awareness for ongoing climate changes at a global scale, biologists need data to advice policy makers, to convince stakeholders and the general public. The most significant impediment for large-scale and fine-grained biodiversity monitoring is the taxonomic one. Even when sampling campaigns are well planned and executed, the samples have little value if the majority of species cannot be identified. This difficulty is mainly due to the lack of time to sort and identify all species found, combined with the fact that taxonomist are scarce and largely specialized for selected taxa, which again makes the majority of identifications very time-consuming and not doable by untrained ecologists. Another problem is that monitoring schemes usually are not comparable, and programs do not run long enough to document trends. Climate monitoring using satellite images and automatized weather stations has been organized at a large scale everywhere on earth. In contrast, large-scale and long-term monitoring of biodiversity does not exist, among others, because the required technology has not been developed. It is therefore crucial to adapt existing technologies for the development of automatized biodiversity motoring. We need "weather stations for species diversity". It is possible to construct an automatized multisensor station for monitoring of species diversity (an AMMOD) using already available

technology: bioacoustics sensors, automated image analyses, DNA-barcoding, analyses of volatile organic compounds (VOCs). Thus it is possible to detect mammals and birds (mainly via images and sounds), insects (mainly with DNA barcoding), plants (via barcoding of pollen and VOCs), and soil microorganisms (via emitted VOCs). AMMODs allow for a continuous detection of a large number of species, Main challenges are in the field of computer science: pattern recognition and comparison of environmental signals with reference databases has to be improved to increase resolution.

5 Working groups

5.1 Biodiversity Weather Stations

Tilo Burghardt (University of Bristol, GB), Yun-Heh Jessica Chen-Burger (Heriot-Watt University – Edinburgh, GB), Joachim Denzler (Universität Jena, DE), Birgitta König-Ries (Universität Jena, DE), Miguel Mahecha (MPI für Biogeochemistry – Jena, DE), Shawn Newsam (University of California – Merced, US), Natalia Petrovskaya (University of Birmingham, GB), and Johann Wolfgang Wägele (ZFMK – Bonn, DE)

The working group focused on the question how an increasing need for data availability on global biodiversity information can be met by introduction of automated field stations for in-habitat sampling. The discussion was underpinned by previous conceptual work on a concept laid down in proposals for BioM-D (Deutsches Zentrum fuer Biodiversitätsmonitoring)/AMMOD (Automatic Multi-sensory station for Monitoring Of species Diversity).

Motivation

The group emphasised that biodiversity is one of the most valuable resources of the planet; and that changes due to species extinction are irreversible. Monitoring biodiversity must therefore be a key component and precursor for taking informed decisions about ecosystem management and conservation. Currently, major obstacles prevent large-scale monitoring of biodiversity at species level: 1) the difficulty of taxonomic identification, 2) the difficulty of spatial-temporal coverage, 3) the difficulty of meaningful spatio-temporal reference, and 4) the workload problem: automatic workflows are in their infancy.

Concept

Faced with these impediments, the group supported the concept that, to be able to observe global change of our biosphere, we need an infrastructure comparable to that used by climate researchers; that is 'weather stations for species diversity'", which operate in a similar fashion to traditional weather stations sampling the breadth of species presence at a particular sampling location over time. In fact, biologists have started to adopt various technologies to enable such measurements – bringing these technologies together for an automatic multi-sensory station for monitoring of species diversity establishes a clear, interdisciplinary development goal. The group reiterated previously identified candidate modalities for automated monitoring; these include DNA barcoding, bio-acoustic monitoring, computer vision-based surveillance, and the analyses of 'smell-scapes'.

Discussions

Driven by the difficulties experienced in raising appropriate funding to progress this agenda at scale, the group discussed proof-of-concept options for the introduction of a few prototypical stations that could demonstrate the value and practical operation of the concept first hand. Key conclusions here included the identification of existing tower infrastructures for the commissioning of systems, the focus on well established sites for best cross-referencing of data, and the limitation of developments for a few species most relevant for showcasing the capabilities of a prototype. We also discussed the technologies underpinning these stations and practical ways of utilising the expertise of scientists and research groups to best conduct development work towards the establishment of prototypes.

Conclusions

The group concluded to work on a detailed positioning paper that may include authorship of the wider community over the following months, and continued efforts towards funding of the concept as next steps. The working group made clear that the technological foundations for an AMMOD concept are widely available today, and that a strong effort is needed to turn this foundation into a practical, working infrastructure to support the gathering of biodiversity information at scale.

5.2 Blending machine learning methods and process-based approaches in dynamic ecological models

Florian Hartig (Universität Regensburg, DE), Martin Bücker (Universität Jena, DE), Gustau Camps-Valls (University of Valencia, ES), Forrest Hoffman (Oak Ridge National Laboratory, US), Kazuhito Ichii (JAMSTEC – Yokohama, JP), Martin Jung (MPI für Biogeochemistry – Jena, DE), Bertram Ludäscher (University of Illinois at Urbana-Champaign, US), Markus Reichstein (MPI für Biogeochemistry – Jena, DE), and Jakob Zscheischler (ETH Zürich, CH)

In many areas of ecology and earth-system sciences, process-based computer simulations are central for understanding the dynamic response of ecological systems to external forcings. An example are dynamic vegetation models, which describe the response of vegetation ecosystems, typically represented by soil water, nutrient and plant state variables, to disturbances and climatic forcings (Zaehle & Friend, 2010; Forkel et al. 2016). The processes governing these dynamics are often complex, and can only partially be observed. It has therefore become common to statistically calibrate model parameters to field observations, for example to vegetation inventories, measured gas exchange, or remote-sensing data (e.g. Hartig et al., 2012).

The issue with this approach is that statistical methods, while accounting for the fact that parameter values are uncertain and some stochastic error is present, are contingent on the assumption that the underlying data-generating model is correct. In other words, statistical conclusions are generally only correct if the fitted model is approximately correct. While this assumption is of lesser concern in simple regression problems, it becomes a major concern in complex, dynamic models, where errors may propagate through nonlinear processes into

other model compartments or times, which can create a range of problems for the correct estimation of model parameters, their uncertainty, and associated forecasts.

An obvious solution is making the structure of process-based models more flexible. This could mean, for example, that process-descriptions, which are typically specified by relatively rigid formulae, are replaced by a flexible statistical approach. An early attempt at implementing such an approach is Wood et al. (2001), who replaced fixed formulae by flexible generalized additive models (see also Nisbet et al., 2004). Another possibility is to make the model structure itself flexible, by adding or removing state variables, or their connections (e.g. Babtie et al., 2014).

In this working group, we discussed those and other technical approaches to tackle the problem of creating flexible models that blend machine learning and process-based models. In particular, we considered the problem of a complex dynamic system, where very little prior information about a particular subprocess is available. The challenge is thus to train a flexible statistical algorithm to learn the dynamical response of the subprocess from observing the system as a whole, while at the same time keeping the problem computationally tractable. A possible solution identified by the group was the use of automatic differentiation methods (Griewank and Walther, 2008), which seemed promising for creating a computationally efficient blend of process-models with machine-learning methods.

Acknowledgements: the working group would like to acknowledge useful suggestions from Shawn Newsam and Andrew Richardson.

References

1 A. C. Babtie, P. Kirk, M. P. H. Stumpf (2014). *Topological sensitivity analysis for systems biology.* Proceedings of the National Academy of Sciences 2014 111 (52), 18507–18512

2 M. Forkel, N. Carvalhais, C. Rödenbeck, R. Keeling, M. Heimann, K. Thonicke, S. Zaehle, M. Reichstein (2016). *Enhanced seasonal CO2 exchange caused by amplified plant productivity in northern ecosystems.* Science 351, 696–699.

3 A. Griewank, A. Walther (2008). *Evaluating Derivatives.* Society for Industrial and Applied Mathematics.

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

5 R. M. Nisbet, E. B. Muller, K. Lika, S. A. L. M. Kooijman (2000). *From molecules to ecosystems through dynamic energy budget models.* Journal of Animal Ecology 69 (6), 913–926.

6 S. N. Wood (2001). *Partially specified ecological models.* Ecological Monographs 71(1), 1–25.

7 S. Zaehle, A. D. Friend (2010). *Carbon and nitrogen cycle dynamics in the o-cn land surface model: 1. model description, site-scale evaluation, and sensitivity to parameter estimates.* Global Biogeochemical Cycles 24, 1–13.

5.3 Improving data discovery and integration for global ecological analyses

Ivaylo Kostadinov (Jacobs University Bremen, DE), Martin Bücker (Universität Jena, DE), Matthew Evans (University of Hong Kong, HK), Thomas Hickler (Senckenberg Research Centre, DE), Donald Hobern (GBIF – Copenhagen, DK), Birgitta König-Ries (Universität Jena & iDiv, DE), Bertram Ludäscher (University of Illinois at Urbana-Champaign, US), Frank Pennekamp (Universität Zürich, CH), Brody Sandel (Santa Clara University, US), and Bernhard Seeger (Universität Marburg, DE)

This working group collected key challenges in contemporary ecological research, which are based on different aspects of data, including discoverability, standardization, integration, size, and metadata description. Current and future approaches in addressing these challenges were discussed in a dialogue between scientists covering the whole range between ecology and informatics. In the later part of the seminar the group focused on the (semi-)automated distribution and harvesting of data and metadata for ecological analyses.

Motivation

Understanding ecological systems, for example detecting biodiversity change and pinpointing its main drivers, is a major challenge which nowadays requires close cooperation between ecology and computer science to tackle. Studying global patterns requires the acquisition, management and integration of large, heterogeneous datasets, whose complexity increases rapidly. Heterogeneity is among the key challenges in working with biodiversity data. Differences in acquisition protocols, study areas and strategies in dealing with gaps in the observations are only some of the problems in integrating data from different sources, disciplines and scale.

Discussion

Ecological communities can be described by four main data types (abundance, distribution, traits, genetic) and their environment is described by direct measurements and remote-sensing approaches. One of the key challenges is trying to bring together different layers of biodiversity change and some of the drivers of that change. For example, traits can constrain species distribution and are therefore one key factor to explore. However, not many dedicated databases for trait data exist and linking to other data (sources) is often difficult. Optimally, there would be a single point for searching and doing at least basic analysis. One possibility would be including traits (e.g. habitat preference) in distribution plots or mapping traits instead of species distributions. This could be attempted in the near future with the GFBio Visualization, Analysis and Transformation (VAT) system (https://vat.gfbio.org) which already offers different plots of publicly available data resources like the Global Biodiversity Information Facility (GBIF, www.gbif.org). Including habitat, temporal, climate continuity in analyses was also identified as a desirable future outcome. This could also be achieved with VAT, by developing a spatio-temporal database in cooperation with scientists, for example to determine apparent correlations between species co-occurrence maps, based on GBIF data. GBIF also strives to improve the data it offers for re-use, for example by including the

taxonomic target and completeness as part of the metadata of the sampling event, so that presence/absence can be inferred more correctly.

Data discovery and integration. A common example of the difficulties researchers face when searching and integrating ecological datasets is that even basic metadata like units of measurement is often not available or impossible to compare. Therefore, dedicated tools to systematically extract candidate metadata (e.g. locations, times) from journal texts would be very helpful. A solution, currently employed by GBIF and others, is a wizard-style interface for authors to input basic metadata. However, minimal requirements almost always differ in scope and detail, as shown by some of the most widely used frameworks: Ecological Metadata Language (EML,https://knb.ecoinformatics.org/#tools/eml), Dublin Core (http://dublincore.org/), Darwin Core (DwC, http://rs.tdwg.org/dwc/). In general, one could distinguish between rich and complex data (e.g. most XML-Schema based data) and simple, flat records. One idea to obtain a rich information network is to increase the semantic load of lightweight DwC properties. Some of the main emerging themes to work on the next years were identified to be (1) the development of an integrated data ecosystem, where global players like GBIF provide highly linked data, (2) the automatic extraction of descriptions for curation, integration and (3) increasing the incentives for data producers to deliver high-quality, standardized metadata, for example by introducing or supporting alternative or additional measures of scientific contribution (e.g. micro-crediting system for data publication).

Metadata enrichment. Metadata is essential for the correct integration of data from different resources. However, it is often not readily available or suffers from lack of standardization. The group outlined a multifaceted approach for addressing this challenge. The main roles would be the original data producer, the later data consumer and any software components for automated metadata extraction. The usual workflow was identified to be: the data producer publishes the data together with the associated journal article (some of the metadata might reside in the article or its appendices). Therefore, the possible sources to extract additional metadata from are (1) author-provided metadata, (2) the data itself, (3) the paper. Optimally, part of the metadata will provide (machine readable) links and to further literature, related projects, funding sources, and involved institutions. A combination of different methodologies was deemed as the best way to cover the different aspects of metadata enrichment. One possibility is to use machine learning techniques for text mining and deducing the research domain context, improving user interfaces for user annotation, curation, and validation, all the while applying established terminologies to standardize the outcome. Finally, a feedback-loop to feed back manually annotated and approved content back to the automation steps would improve their performance.

Conclusion and Outlook

The major challenges for improving descriptive metadata of datasets, and consequently their discoverability and interoperability, are (1) providing the right tools and the right incentives for the data producers to provide the metadata in a standardized way, (2) determining the minimal set of parameters, required for interoperability and (3) providing the tools to harvest the required metadata from available resources like the data itself, the corresponding journal article or even program code automatically.

Several informatics techniques like machine learning offer promising solutions for increasing the automation of metadata extraction. In order for them to be applied meaningfully, priorities for improved information capture must be identified. This includes determining

what essential biodiversity variables ought to be captured. Recording of re-usable workflows from user operations, as already employed by VAT, can deliver some insight to the way scientists (re-)use biodiversity data.
Further improving and standardizing the interfaces for data exchange between data repositories, taking into account emerging serialization formats and access means (e.g. protocols), is important. Large, integrated data resources like GBIF will continue to play a key role in paving the road to Linked Open Data for ecology and biodiversity research.
Raising community awareness to the problems at hand is important. This includes clarifying the added-value of high quality metadata and supporting an appropriate credit system for data publication.
As a continuation of the efforts initiated during the seminar, the working group will organize a Hackathon with the following preliminary topics in mind:

- Rapid prototyping tools for automatically extracting metadata
- Integration of different metadata schema
- Exploring different techniques, e.g. deep learning and traditional statistics, and data sources, e.g. Catalogue of Life (www.catalogueoflife.org) for improving data discovery and linkage.

6 Panel discussions

6.1 Reproducibility and teaching needs – a dialogue between ecology and computer science

Frank Pennekamp (Universität Zürich, CH), Martin Bücker (Universität Jena, DE), Tilo Burghardt (University of Bristol, GB), Gustau Camps-Valls (University of Valencia, ES), Yun-Heh Jessica Chen-Burger (Heriot-Watt University – Edinburgh, GB), Joachim Denzler (Universität Jena, DE), Matthew Evans (University of Hong Kong, HK), Florian Hartig (Universität Regensburg, DE), Thomas Hickler (Senckenberg Research Centre, DE), Donald Hobern (GBIF – Copenhagen, DK), Forrest Hoffman (Oak Ridge National Laboratory, US), Kazuhito Ichii (JAMSTEC – Yokohama, JP), Martin Jung (MPI für Biogeochemistry – Jena, DE), Birgitta König-Ries (Universität Jena, DE), Ivaylo Kostadinov (Jacobs University Bremen, DE), Bertram Ludäscher (University of Illinois at Urbana-Champaign, US), Miguel Mahecha (MPI für Biogeochemistry – Jena, DE), Laetitia Navarro (iDiv – Leipzig, DE), Shawn Newsam (University of California – Merced, US), Natalia Petrovskaya (University of Birmingham, GB), Markus Reichstein (MPI für Biogeochemistry – Jena, DE), Andrew Richardson (Harvard University – Cambridge, US), Ribana Roscher (Universität Bonn, DE), Brody Sandel (Santa Clara University, US), Bernhard Seeger (Universität Marburg, DE), Johann Wolfgang Wägele (ZFMK – Bonn, DE), and Jakob Zscheischler (ETH Zürich, CH)

Abstract

This document summarizes the plenary discussion about reproducibility and teaching needs during the seminar. The discussion started with perspectives from ecologists and computer

scientists about the state of the art and challenges of reproducibility in their fields. A key question discussed was the time-scale of reproducibility, which may pose large challenges for computational researchers. On the practical side, data and code archiving practices were discussed, and efforts to provide incentives for reproducible research highlighted. The second part of the discussion covered whether there is a need for joint training efforts by ecologists and computer scientists to generate the next generation of eco-informaticians trained for the challenges of a largely data-driven science. We concluded the plenary discussion with agreement that both disciplines would benefit from a better dialogue.

Reproducibility in ecology and computer science

Ecology is increasingly becoming a data-driven science and hence ecologists need to work with large, complex datasets for which they often lack the appropriate training [2]. This can lead to issues with reproducibility, which is not unique to the field of ecology, but for science in general [4].

The discussion was started with perspectives from ecologists and computer scientists. The available ecologists believed that a large fraction of papers currently published is not fully reproducible, the attitude toward reproducibility is changing within the field. This is partly due to tools like the statistical computing environment R [5] which is increasingly used for analyses in ecology, biology and the life sciences. It is also due to the rise of literate programming tools in R such as knitr and Rmarkdown [3], which allow to intersperse code and text and are increasingly used.

The change is also fostered by changing journal policies that increasingly require data archiving [9] and also the provisioning of computer code for modeling/simulation studies and methods development [1]. Nowadays most of the major journals in ecology and evolution require data archiving for publication, and guidelines and best practices how to make data available can be found [10]. Nevertheless, current practice is still lacking behind [6]. Whereas many of the modern tools for reproducible research were developed by computer scientists (e.g. version control, unit tests, code review), the available computer scientists in the audience expressed doubts whether the practice of reproducibility is actually better developed within their field [4].

A major discussion topic was what reproducibility actually implies. Whereas data archiving and providing scripts may guarantee the correctness and validity of the results at the time of publication, it is not guaranteed that results will be reproducible in the future (e.g. in ten years time). A major challenge is, for instance, the use of high performance computing in many fields of computer science, including computer vision. Ideally, information about software and hardware architecture should be preserved to reproduce results in the future, but this is often prohibitive. Whereas these issues require careful consideration, they probably only concern a small fraction of ecologists, whose research questions still can be addressed on common desktop or laptop computers most of the time.

Towards more reproducible research

Several guidelines for better reproducible results are available [11, 7]. A first step is better data archiving practices. Several recent publications highlight the value of data archiving and give practical advice how to prepare data for long-term archiving. In addition, journals could require that the scripts used for data analysis are submitted for peer review and check that the output of the scripts corresponds to the results reported in a paper [3]. The Association for Computing Machinery, for example, uses a system in which badges are assigned for

research results which can independently be reproduced. In the Life Sciences the ReScience Journal aims to publish independent reproductions of published research. They implemented a fully transparent review process in which the reproduction is first peer-reviewed and then published online, providing researchers with incentives to reproduce other's work.

Another possible avenue for improved reproducibility is the use of work flows that document data input and provenance [8]. These work flows often produce visual representations of data sources and processing steps, which can be understood without knowing the data processing language itself.

Students and researchers should also be exposed to reproducible research early on in their careers. One possible way to do so is a reproducible research journal club. Instead of just reading and discussing research papers, the goal is to reproduce the analysis or model of a given paper. To do so, the students have to access the data from a publicly available source such as Dryad or contact the authors directly. Consequently, the analysis of the paper is performed independently from the authors of the study, just based on the information provided in the paper. Ideally, one does not only reproduce the results of the study, but also learns and understands a method, and learns about the steps a researcher has taken during the analysis. An example of such a reproducible research journal club is run by Owen Petchey at the University of Zurich[1], with successful reproductions publicly available: http://opetchey.github.io/RREEBES/.

Training needs for better computing practices in ecology?

The challenges in managing increasingly large and complex datasets require appropriate training of ecologists. Two non-profit organizations are dedicated to provide training for scientific computing and data management: Software Carpentry[2] is primarily dedicated to train scientists and engineers basic principles to make their scientific computing applications reliable. Data Carpentry[3] on the other hand focuses on providing training in teaching the basic skills to conduct data-driven research such as cleaning, integrating, managing and visualizing large datasets covering the full data life cycle for a variety of research fields (e.g. biology, life sciences, ecology, social sciences). Both organizations organize short, domain-specific workshops of two to three days in which basic principles are taught by trained instructors. Both organizations adhere to a particular teaching style that values hands-on programming by participants, and live coding of instructors. Whereas Data Carpentry is primarily directed at learners without prior programming experience, Software Carpentry workshops often require a basic knowledge of programming languages such as R or Python.

Besides focused workshops, the challenges of data-driven science may ask for more formal training in terms of dedicated Master's programs. Such Master's programs could provide specialized training at the cross-section of ecology and computer science, covering advanced topics such as database creation and management, computer vision and machine learning algorithms, geographical information systems (GIS) and modeling of complex ecological systems.

We concluded the plenary discussion with agreement that both disciplines would benefit from a better dialogue.

---

[1] https://github.com/opetchey/RREEBES
[2] https://software-carpentry.org/
[3] http://www.datacarpentry.org/

References

1 Freckleton, R. P. and Iossa, G. (2010). *Methods in ecology and evolution.* Methods in Ecology and Evolution, 1(1):1–2.

2 Michener, W. K. and Jones, M. B. (2012). *Ecoinformatics: supporting ecology as a data-intensive science.* Trends in Ecology & Evolution, 27(2):85–93.

3 Mislan, K. a. S., Heer, J. M., and White, E. P. (2016). *Elevating The Status of Code in Ecology.* Trends in Ecology & Evolution, 31(1):4–7.

4 Peng, R. D. (2011). *Reproducible Research in Computational Science.* Science (New York, N.y.), 334(6060):1226–1227.

5 *Development Core Team (2016).* R: A language and environment for statistical computing.

6 Roche, D. G., Kruuk, L. E. B., Lanfear, R., and Binning, S. A. (2015). *Public Data Archiving in Ecology and Evolution: How Well Are We Doing?* PLOS Biology, 13(11):e1002295.

7 Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). *Ten Simple Rules for Reproducible Computational Research.* PLOS Computational Biology, 9(10):e1003285.

8 Shade, A. and Teal, T. K. (2015). *Computing Workflows for Biologists: A Roadmap.* PLOS Biology, 13(11):e1002303.

9 Whitlock, M., McPeek, M., Rausher, M., Rieseberg, L., and Moore, A. (2010). *Data Archiving.* The American Naturalist, 175(2):145–146.

10 Whitlock, M. C. (2011). *Data archiving in ecology and evolution: best practices.* Trends in Ecology & Evolution, 26(2):61–65.

11 Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., and Wilson, P. (2014). *Best Practices for Scientific Computing.* PLOS Biology, 12(1):e1001745.

Participants

- Martin Bücker
Universität Jena, DE
- Tilo Burghardt
University of Bristol, GB
- Gustau Camps-Valls
Universitat de València, ES
- Yun-Heh Jessica Chen-Burger
Heriot-Watt University – Edinburgh, GB
- Joachim Denzler
Universität Jena, DE
- Matthew Evans
University of Hong Kong, HK
- Florian Hartig
Universität Regensburg, DE
- Thomas Hickler
Senckenberg Research Centre, DE
- Donald Hobern
GBIF – Copenhagen, DK
- Forrest Hoffman
Oak Ridge National Laboratory, US
- Kazuhito Ichii
JAMSTEC – Yokohama, JP
- Martin Jung
MPI für Biogeochemistry – Jena, DE
- Birgitta König-Ries
Universität Jena, DE
- Ivaylo Kostadinov
Jacobs University Bremen, DE
- Bertram Ludäscher
University of Illinois at Urbana-Champaign, US
- Miguel Mahecha
MPI für Biogeochemistry – Jena, DE
- Laetitia Navarro
iDiv – Leipzig, DE
- Shawn Newsam
University of California – Merced, US
- Frank Pennekamp
Universität Zürich, CH
- Natalia Petrovskaya
University of Birmingham, GB
- Markus Reichstein
MPI für Biogeochemistry – Jena, DE
- Andrew Richardson
Harvard University – Cambridge, US
- Ribana Roscher
Universität Bonn, DE
- Brody Sandel
Santa Clara University, US
- Bernhard Seeger
Universität Marburg, DE
- Johann Wolfgang Wägele
ZFMK – Bonn, DE
- Jakob Zscheischler
ETH Zürich, CH