Data Science: History repeated ?
The heritage of the Free and Open Source Science Community

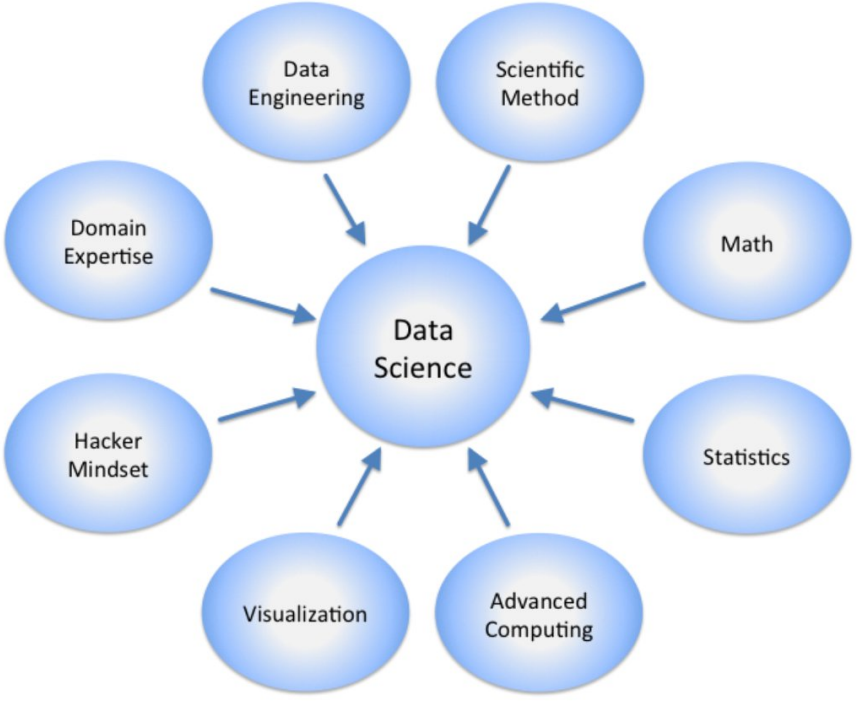
Peter Löwe, Technische Informationsbibliothek TIB, Hannover, Germany
peter.loewe@tib.uni-hannover.de

Markus Neteler, Fondazione Edmund Mach, S. Michele all'Adige, Italy

EGU General Assembly 2014
2.5.2014

What is Data Science ?

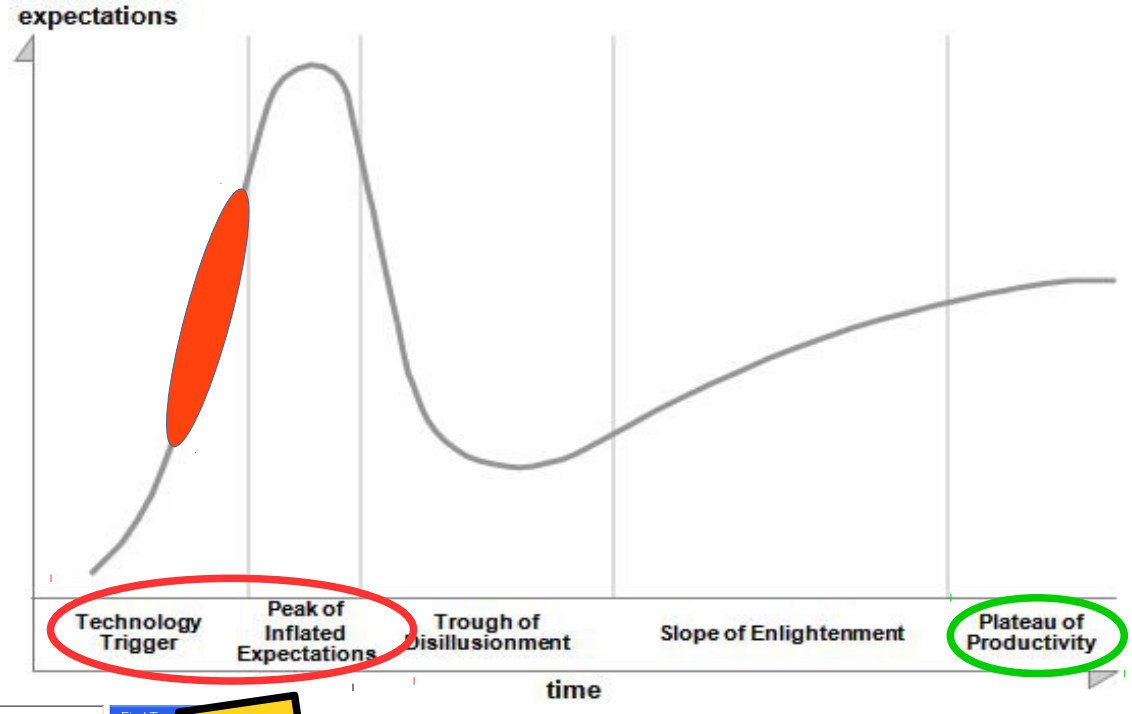
- Knowledge extraction from large data sets by means of scientific methods.
- Uses techniques and theories from many fields,
- which are jointly used to furthermore develop information retrieval on structured or unstructured very large datasets.

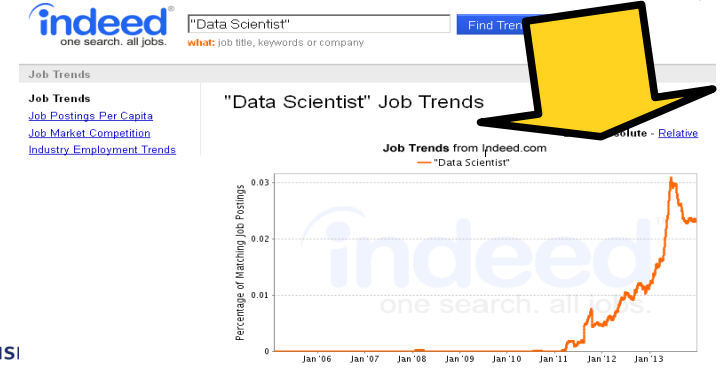


http://en.wikipedia.org/wiki/File:DataScienceDisciplines.png

Data Science – on the Gartner Hype Cycle

The current perception of this field is still in the first section of the Gartner hype cycle.





Forbes on Data Scientists: „*the sexiest career of the 21st century*“

*„The **Harvard Business Review** has declared 'Data Scientist' to be **the sexiest career of the 21st century**.*

*Because if there's one thing that gives a job an indefinable allure, **it is everybody else being kind of unsure what it is you really do***

— a quality that data scientists [...] embody. „

http://www.forbes.com/sites/gilpress/2012/09/27/data-scientists-the-definition-of-sexy/

Data Science – the GIS of the next decade ?

Geographic Information Systems (GIS)

- processing and analysing of spatially referenced content

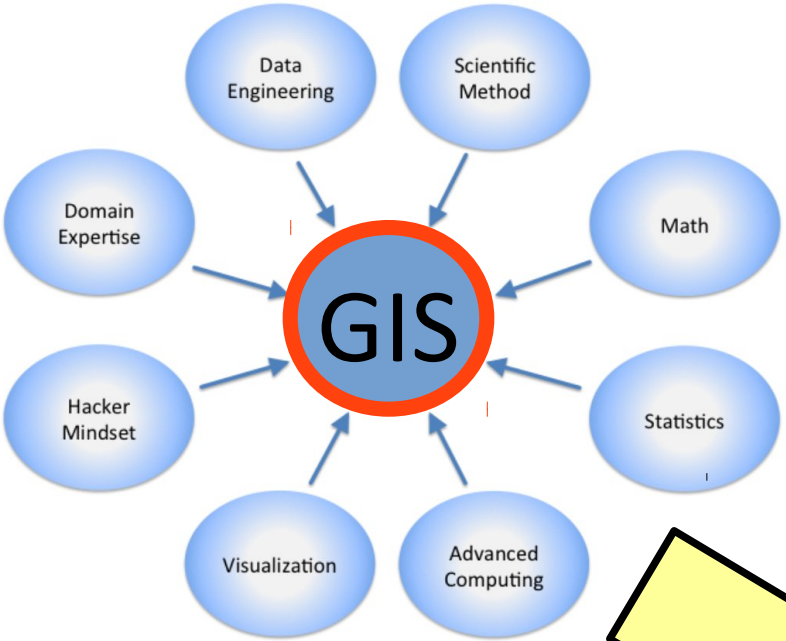
- integration and storage of spatial information
- from heterogenuous sources,
- **data analysis**,
- sharing of reconstructed or aggregated results in visual form

Geo-... and. ...**-Informatics**: „**Hyphenated Computer Scientists**“

- Applications in:

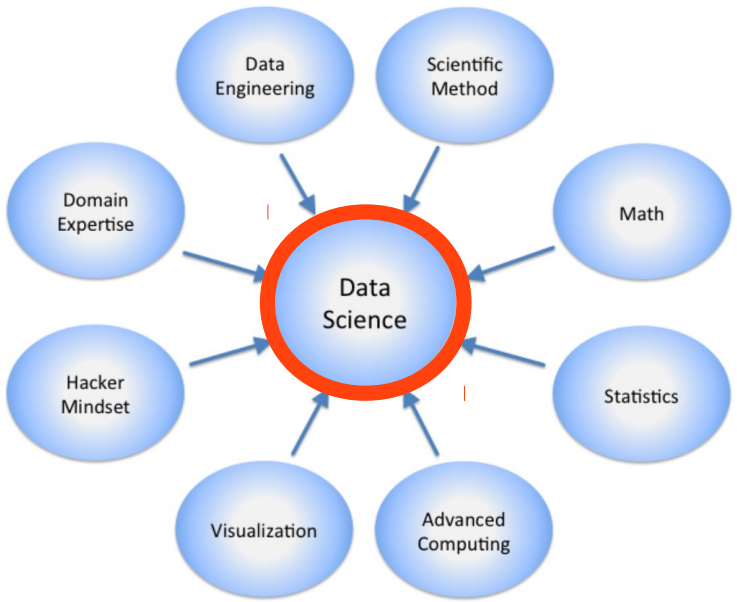
  Biology, Ecology, Medicine, Physics, Chemistry, WWW Studies, Archeology, Agriculture, Politics, etc.

History repeated ? What lessons can be learned from GIS ?



„Have I Been a Data Scientist from the Start? *Parallels from the Geographic Information Science Community in the Early 1990s“, AGU Poster IN43A-1639, 2013*

Dawn J. Wright, Environmental Systems Research Institute (Esri), Redlands, CA, USA

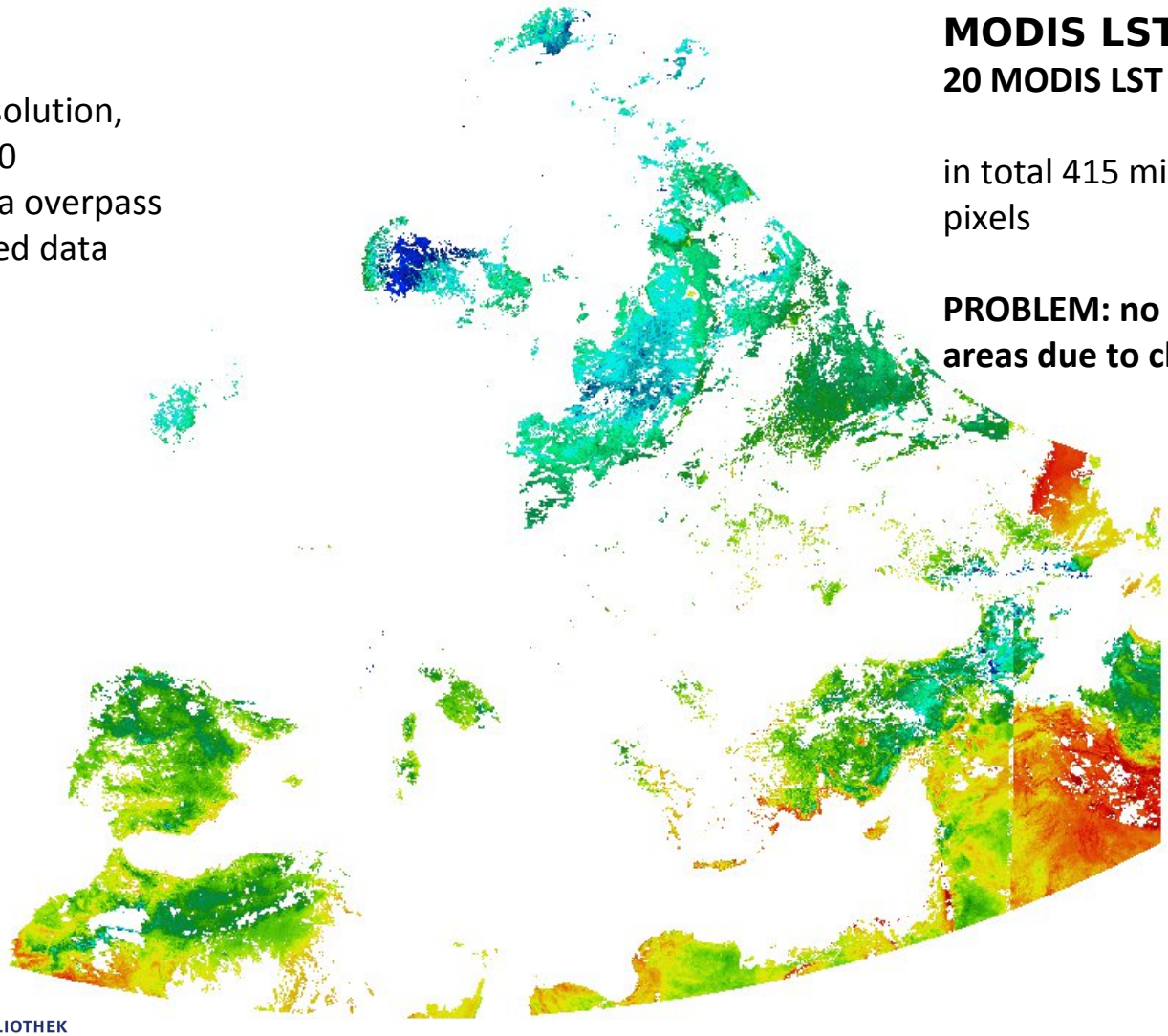


http://en.wikipedia.org/wiki/File:DataScienceDisciplines.png

GIS = Ecoinformatics = Data Science ?

Remote Sensing for Ecology

1000m resolution,
2010-05-30
01:30 Aqua overpass
Raw-filtered data



MODIS LST mosaic of
20 MODIS LST tiles

in total 415 million pixels

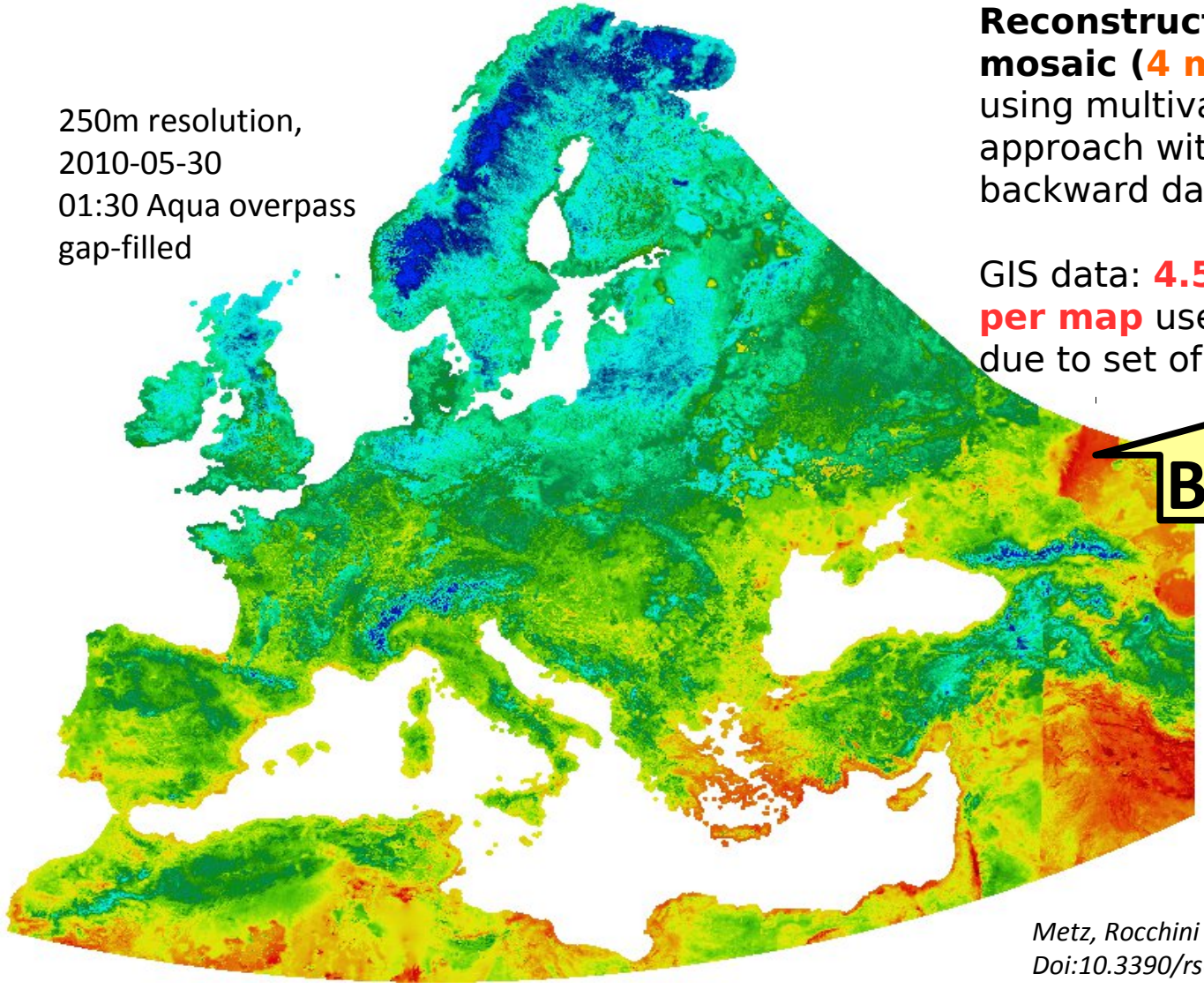
PROBLEM: no data areas due to clouds

Metz, Rocchini & Neteler, 2014, Doi:10.3390/rs6053822

GIS = Ecoinformatics = Data Science ?

Remote Sensing for Ecology

FONDAZIONE EDMUND MACH
CENTRO RICERCA e INNOVAZIONE



250m resolution,
2010-05-30
01:30 Aqua overpass
gap-filled

Reconstructed MODIS LST mosaic (4 maps per day) using multivariate approach with forward/ backward data lookup

GIS data: **4.5 billion pixels per map** used in calculation due to set of input variables

Big Data

Metz, Rocchini & Neteler, 2014
Doi:10.3390/rs6053822

The five tiers of GIS

- **Standardisation, Standardisation, Standardisation**
- **Free and Open Source Software (FOSS)**
  - freely accessible program code enabling
  - analysis, education and improvement.
- Applications based on **closed-source software,**
  - without the option of access to
  - and analysis of the implemented algorithms

FOSS GIS in Ecology...
FOSS Ecoinformatics

Review

Cell PRESS

Special Issue: Ecological and evolutionary informatics

Ecoinformatics: supporting ecology as a data-intensive science

Data be free!

William K. Michener[1] and Matthew B. Jones[2]

[1] University Libraries, University of New Mexico, Albuquerque, NM
[2] National Center for Ecological Analysis and Synthesis, University

Ecology is evolving rapidly and increasingly changing into a more open, accountable, interdisciplinary, collaborative and data-intensive science. Discovering, integrating and analyzing massive amounts of heterogeneous data are central to ecology as researchers address complex questions at scales from the gene to the biosphere. Ecoinformatics offers tools and approaches for managing ecological data and transforming the data into information and knowledge. Here, we review the state-of-the-art and recent advances in ecoinformatics that can benefit ecologists and environmental scientists as they tackle increasingly challenging questions that require voluminous amounts of data across disciplines and scales of space and time. We also highlight the challenges and opportunities that remain.

run on powerful distributed computing systems. For example, Kepler includes facilities for easily executing models on pre-existing computing grids, in cloud-computing environments and in ad hoc networks of workflow systems [65,66], while capturing a full provenance trace of the process; and VisTrails is built to generate effectively scientific visualizations while also capturing the provenance of the analysis [61].

Supporting the full data life cycle

New ground, aerial and satellite-based environmental observing systems coupled with the rapid growth in the use of in situ environmental sensor networks for field research and monitoring, as well as an ever-growing number of citizen-science programs, will soon push ecology and the environmental sciences into a new era where petabytes of data are being collected annually. Powerful informatics platforms will be required to support scientists as they move into this age of data-intensive science. Several such platforms are being designed and built at various scales, including the LTER NIS, the DataONE Federation, LifeWatch, NEON, GLEON and OOI.

The US LTER Network is presently building a network information system that will support synthetic science by: (i) using standardized metadata management and access approaches; (ii) providing middleware programs and workflow solutions that facilitate the creation and maintenance of integrated LTER data sets; and (iii) supporting standardized applications that facilitate discovery, access and use of LTER data [25,67].

DataONE represents a new type of research platform

Box 3. Open science for society

Global problems require open access to global data from many disciplines. Such data arise from scientific disciplines that often have very different cultures with respect to data sharing, development and adoption of standards, and practice of good data stewardship. Incentives from research sponsors, societies and institutions (e.g. requiring data management plans) combined with the availability of new informatics tools and platforms, such as DataONE, will be necessary to facilitate data intensive science. Three avenues of research and development offer particular promise: (i) automated provenance-tracking mechanisms that allow scientists to understand and replicate scientific findings fully [76]; (ii) advanced visual analytics that enable scientists to interpret complex, large data volumes more rapidly [68]; and (iii) usability analysis and software engineering support that enable scientists to use advanced ecoinformatics tools more easily.

Tracking the provenance of scientific results is particularly important as advances in environmental science are applied to issues important to society. Open data provide the feedstock on which good science is based, replicable analysis and modeling practices lead to robust findings, and open-access publication disseminates these critical results to the broadest audiences, ensuring the greatest impact of open science for society.

research must be openly available and the approaches used in deriving scientific findings must be transparent to ensure that science and society maximally benefit (Box 3).

Remaining challenges

Despite the emergence of ecoinformatics solutions that enable science, several technical and sociocultural challenges and research opportunities remain. First, from the technical side, it is difficult to transport terabyte- and petabyte-sized data sets. Possible solutions include adding

Open science wants Open Source!

Trends in Ecology and Evolution June 2012, Vol. 27, No. 6

Let the four freedoms paradigm apply to ecology

Duccio Rocchini and Markus Neteler

Fondazione Edmund Mach, Research and Innovation Centre, Department of Biodiversity and Molecular Ecology, Via E. Mach 1, 38010 S. Michele all'Adige (TN), Italy

In 1985, Richard Stallman, one of the most brilliant minds in computer science, founded the Free Software Foundation and launched the concept of 'copyleft', the opposite of copyright. The aim, outlined in the GNU Manifesto (http://www.gnu.org/gnu/manifesto.html, [1]), was to make software programs 'free' as in 'freedom'.

The famous 'four freedoms' expounded by Stallman [1] are: (i) the freedom to run the program for any purpose; (ii) the freedom to study how the program works and adapt it to one's own needs; (iii) the freedom to redistribute copies; and (iv) the freedom to make improvements to the program and release them to the public. Thus, the whole (scientific) community benefits from software development. These freedoms are also inherent in several free software licenses, the GNU General Public License (GPL) being one of the most popular.

Approximately a quarter of a century after Stallman put forward his ideas, William K. Michener and Matthew B. Jones, in an article in *TREE* [2] focusing on the analysis of ecological data, stated that: 'analytical processes are fundamental to most published results in ecology'. Explicit reference to the analytical procedures adopted in generating scientific results is crucial for reproducibility, yet these processes are rarely documented in published ecological papers [2]. Scientific workflow applications, such as Kepler (https://kepler-project.org), attempt to address the problem [2], but are only partially successful because the underlying algorithms may still be opaque.

In our view, the explicit use of Free and Open Source Software (FOSS) with availability of the code is essential for completely open science: 'scientific communication relies on evidence that cannot be entirely included in publications', but 'anything less than the release of source programs is intolerable for results that depend on computation' [3].

The idea of FOSS and the public availability of the code has been around for almost as long as software [4]. Nonetheless, as far as ecologists are concerned, the open source philosophy is only just taking off, as Stokstad has also pointed out [5].

The increasing availability of open ecological data through networks such as the Global Biodiversity Information Facility (GBIF, http://www.gbif.org, [6]) or the Data Observation Network for Earth (DataONE) federated data archive (http://www.dataone.org, [7]) makes it increasingly possible to test cutting-edge ecological theories, such as dark diversity [8], evolutionary paths [9] and climate change scenarios [10]. In using a shared open-source code for testing these ecological theories, researchers can be sure that their results are reliable and also that the code they have used is robust [11]. This is particularly true when complex algorithms (or statistical approaches) are involved.

To avoid black box calculations and built-in user interfaces, criticized in [2], researchers have recourse to several examples of FOSS in areas of ecological research, such as ecological statistics (e.g. R Language and Environment for Statistical Computing, http://www.R-project.org, [12]) and spatial ecology [e.g. Geographical Resources Analysis Support System (GRASS) GIS, http://grass.osgeo.org, [4]). The modular design of such software means decentralized contributions can be made to the source code and allows different institutions and individuals around the world to improve the code base.

If FOSS were more widely employed in ecology and the code used in data analysis provided in scientific papers, more researchers [11] would be able to rely on and replicate peer-reviewed functions. Efforts still need to be made in this area to improve the processes for sharing what is in effect the backbone of ecological software: its code. Therefore, there is an urgent need to embrace Stallman's four freedoms paradigm in ecology.

Acknowledgments

We would like to thank Anne Ghisla, Luca Delucchi and Tessa Say for valuable suggestions. DR is partially funded by the Autonomous Province of Trento (Italy) within the ACE-SAP project (University and Scientific Research Service regulation number 23, June 12, 2008).

Corresponding author: Rocchini, D. (duccioroccchini@gmail.com), (duccio.rocchini@fmach.it).

Why FOSS is used

Letters

Richard Stallmann's four Freedoms
How can non-free software be scientific ??

0. Freedom to run the program as you wish.

1. Freedom to study the source code of the program and then change it so the program does what you wish.

2. The freedom to redistribute the exact copies of the software when you wish.

3. Freedom to contribute to your community. That's the freedom to distribute copies or modified versions when you wish.

If a software allows for all four essential freedoms, then it is free software.

The heritage of „Free and Open Source Software GIS“ **Approaches that will work**

- **Acknowledgement of a meritocratic attitude is crucial.**
- Best-practices, which evolved from long duration FOSS projects (up to 30 years).
- Community-driven global umbrella organisations
- Evolutionary processes of establishing and maintaining a web-based democratic culture spawning new kinds of communication and projects.

Meritocracy

Actions speak louder than words.

- Political philosophy: **Power should be vested in individuals according to merit.**
- The **Apache Software Foundation** (among other projects) is an example for open source software projects that officially claim to be a meritocracy.
- Meritocratic community culture in science:
  - **transcends the established compartementation and stratification of science**
  - by creating **mutual benefits for the participants**,
  - ***irrespective of their research interest and standing***.

OSGeo: Interlinked communities
Trust, share, reuse, improve

- The **Open Source Geospatial Foundation** (**OSGeo**)
  - a **non-profit NGO**
  - **an „umbrella“ for multiple community software projects**
  - Mission: to support and promote the collaborative development of open geospatial technologies and data.

- OSGeo draws governance inspiration from several aspects of the **Apache Foundation**,
  - including **a membership composed of individuals drawn from foundation projects**
  - who are selected for membership status based on their active contribution to foundation projects and governance.

OSGeo Mission Statement

Trust, share, reuse, improve

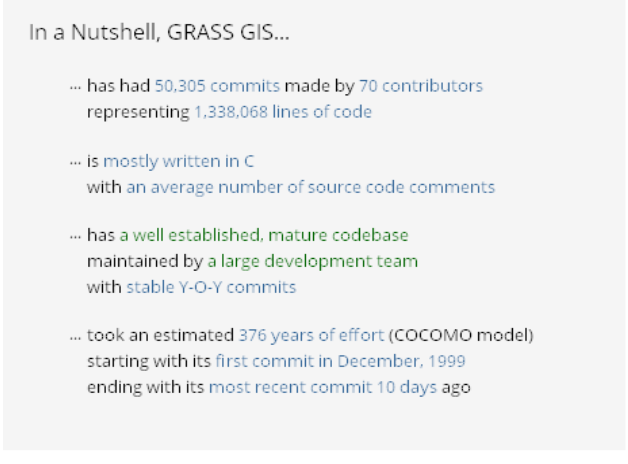
Support and promote the highest quality Open Source Geospatial Software

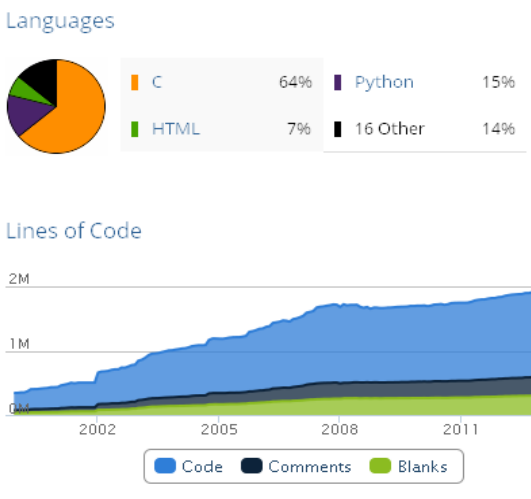
OSGeo Principles („the way“)
Trust, share, reuse, improve

- Projects should **manage themselves**, striving for **consensus** and encouraging **participation** from all contributors - **from beginning users to advanced developers.**
- **Contributors are the scarce resource and successful projects court and encourage them.**
- Projects are encouraged to **adopt open standards** and **collaborate with other OSGeo projects**.
- **Projects** are **responsible** for reviewing and controlling their code bases to **insure the integrity of the open source baselines**.

Project Incubation Process
Trust, share, reuse, improve

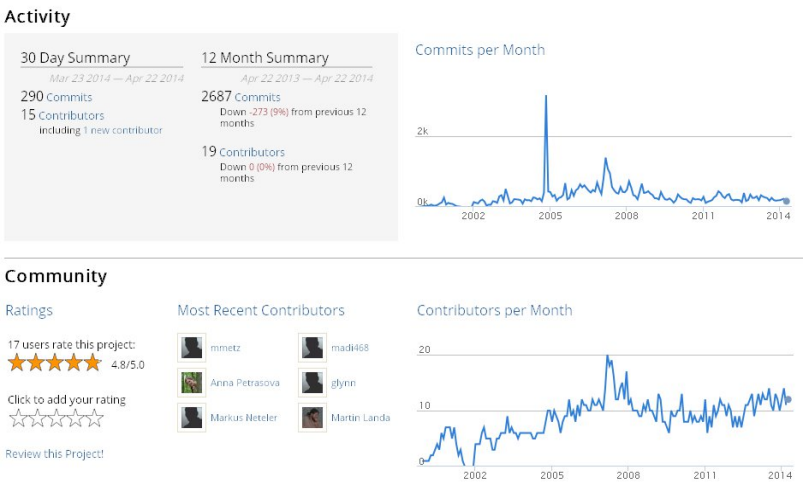
- Projects being part of OSGeo...
  - have a successfully operating open and **collaborative** development **community**
  - have **clear IP oversight of the code base** of the project
  - adopt the OSGeo principles and operating principles
  - are mentored through **the incubation process**





Project Operating Principles

- document how they **manage themselves**.
- maintain developer and user **documentation**.
- maintain a **source code management system**.
- maintain an **issue tracking system**.
- maintain **project mailing lists**.
- actively promote their participation in OSGeo.
- have automated **build and test systems**.



Software Quality: Sharing trustworthy building blocks

OSGeo works because **people** participate.

Participation includes

- **using,**
- **learning and then**
- **contributing back to the community**
- **user review of the contributions through repository commit mailing lists**

Quality-checked
Free and Open Source
makes
science a safe investment.

Conclusion: Proven patterns for Data Science

- **Free and Open Source Software**
- **Meritocratic community culture** transcends the established compartementation and stratification of science.
- **Umbrella Organisation**
- **Establish the „trust, share, reuse, improve“-paradigm**

These best practice patterns will enable the emerging Data Science communities to avoid known pitfalls and quickly establish a productive environment.

TIB | Technische Informationsbibliothek

Fondazione Edmund Mach – Centro Ricerca e Innovazione

Thanks for listening !

peter.loewe@tib.uni-hannover.de