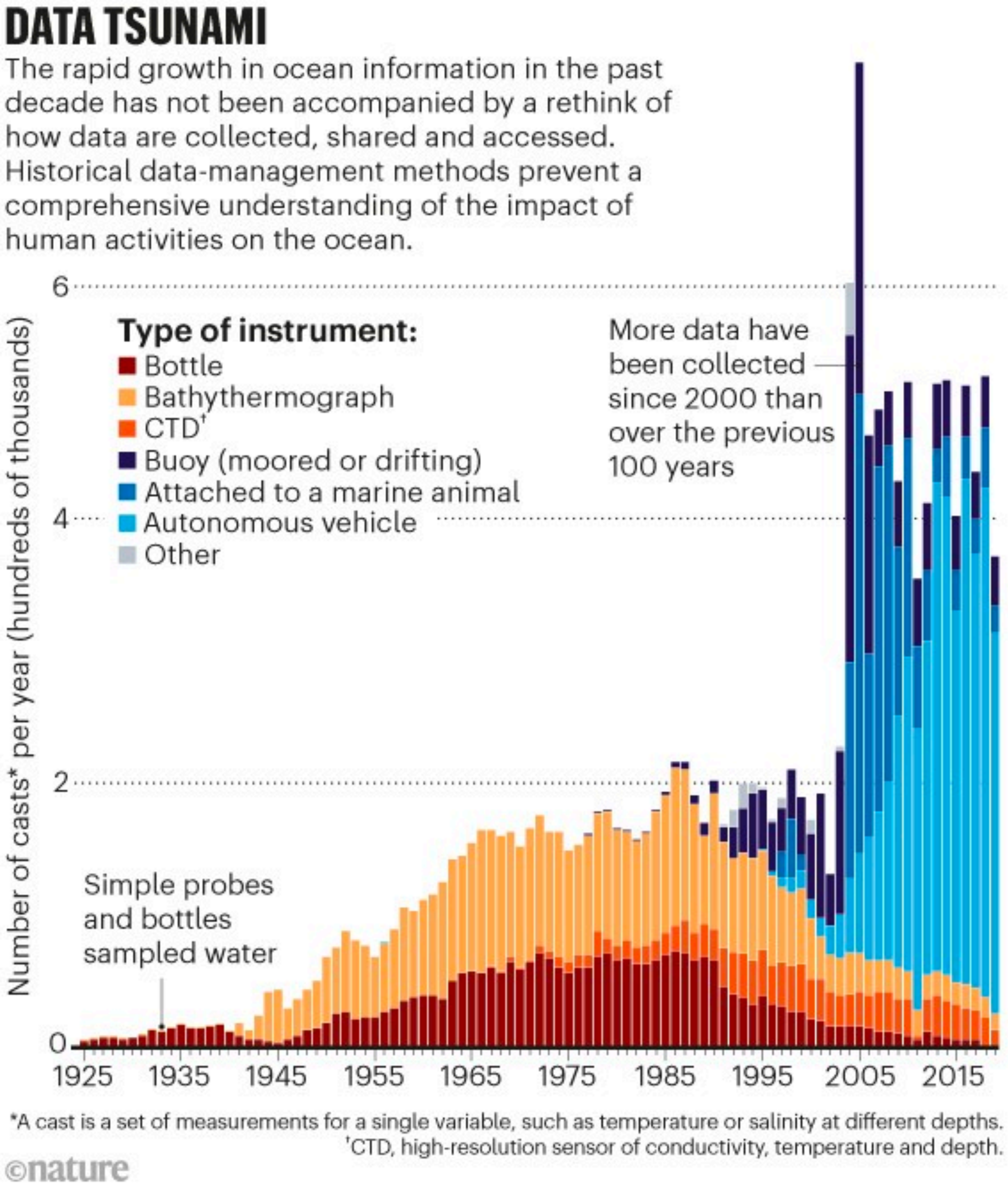
OEAS 895 Env. Data Sci.

Spring 2023
Week 1

Dr Sophie Clayton, sclayton@odu.edu
Class times: OCNPS 403 T/Th 9:30 - 10:45
Office hours: OCNPS 423 M 13:00 - 15:00



DATA TSUNAMI

The rapid growth in ocean information in the past decade has not been accompanied by a rethink of how data are collected, shared and accessed. Historical data-management methods prevent a comprehensive understanding of the impact of human activities on the ocean.

*A cast is a set of measurements for a single variable, such as temperature or salinity at different depths.
†CTD, high-resolution sensor of conductivity, temperature and depth.

©nature

Brett et al., 2020, Nature,
doi: https://doi.org/10.1038/d41586-020-01668-z

Learning objectives

1. Understand FAIR data principles and how to apply them when generating, sharing and accessing data.
2. Develop a working knowledge of existing ocean and earth science databases and how to efficiently access data from them, including via APIs.
3. Students will develop their own data analysis toolbox using, but not limited to, Python and shell scripts.
4. Understand and use version control (e.g. git), environments (e.g. conda) and code repositories (e.g. GitHub) to manage and share code.
5. Understand the underlying principles of machine learning techniques for regression and classification, including supervised and unsupervised learning and apply them to a targeted research question.
6. Understand the process of model evaluation and optimization and commonly used metrics for reporting model performance.

Course schedule:

Modifications may need to be made to this schedule as the semester progresses.

Week	Topic	Assignment
1 (1/10)	Open Science framework FAIR data Version control (git, github)	PS 1 – git and github
2 (1/17)	Initial data access and exploration Basic plotting in python Oceanographic databases and repositories	PS 2 – exploratory data analysis
3 (1/24)	Oceanographic toolboxes Mapping toolboxes	PS 3 – building a function for accessing data using an API
4 (1/31)	NO CLASSES	
5 (2/7)	Building packages and sharing code Collaborative workspaces	PS 4 – documenting and sharing your code (builds on PS3)
6 (2/14)	Machine Learning overview Introduction to scikit-learn	
7 (2/21)	Supervised Learning Overview of algorithms Training and testing algorithms	PS 5 - building a simple classification model with scikit-learn
8 (2/28)	Unsupervised learning Clustering Classification	PS 6 – building a regression model with scikit-learn
9 (3/7)	NO CLASSES – SPRING BREAK	Deadline for capstone project approval
10 (3/14)	Model evaluation Cross-validation (dealing with small training sets)	
11 (3/21)	Capstone project development	PS 7 – project outline
12 (3/28)	Machine Learning applications in oceanography Paper discussion (student-led)	
13 (4/4)	Capstone project hacking	
14 (4/11)	Capstone project hacking	PS 8 – project and code review (peer evaluation)
15 (4/18)	Data analysis project hacking In-class capstone presentations	Capstone Project Due (published github repository)

Capstone Project

- combine and apply the skills learned in class in the context of a real-world research problem.
- data analysis and visualization
- developing and evaluating machine learning models
- dataset(s) and general scope of their capstone project approved by the instructor prior to spring break.

Grading Summary

Problem sets (8)	50%
Capstone project report	20%
Capstone project repository	15%
Capstone project presentation	15%

Course Housekeeping

Tools: laptop, git, GitHub, Slack, conda, python

To Do:

- set up GitHub account before Thursday
- check course info on GitHub:

  https://github.com/sophieclayton/OEAS805_envdatasci
- join Slack group (invites sent):

  shorturl.at/dnOR5

Questions?

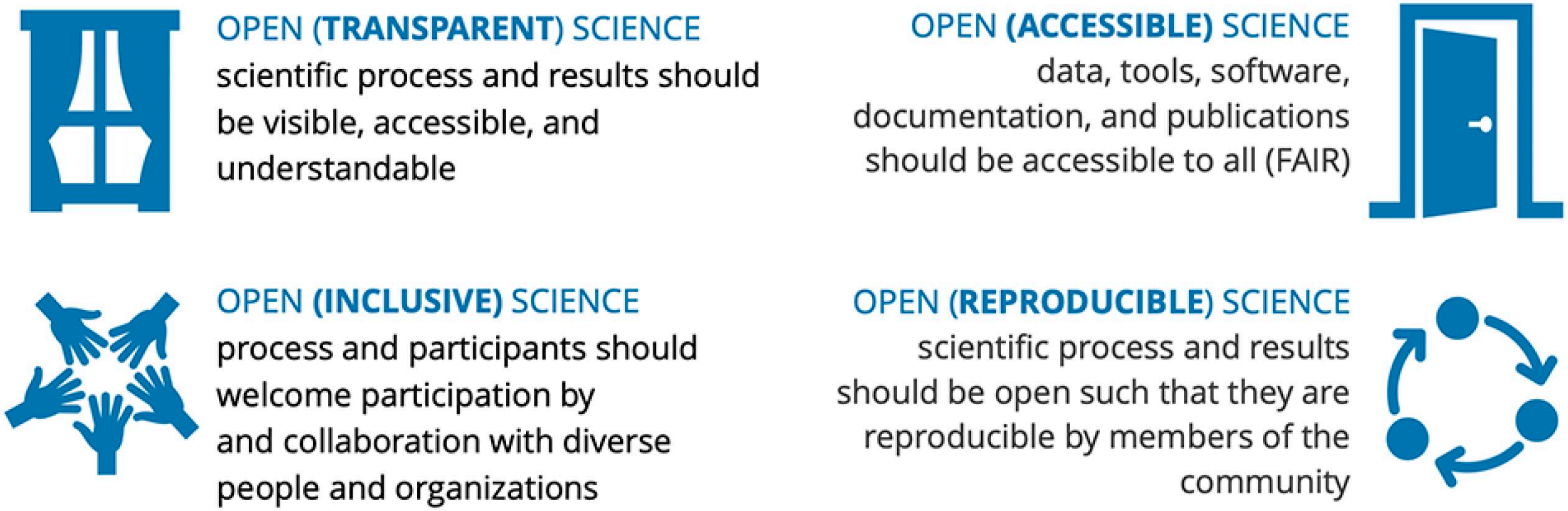
Intros

- research interests
- coding background
- types of data you’ve worked with (generated?)
- goals for this course

(Data) Science needs (good) data

- Open science framework
- FAIR data principles
-

"Open-source science requires a culture shift to a more inclusive, transparent, and collaborative scientific process, which will increase the pace and quality of scientific progress."



OPEN (**TRANSPARENT**) SCIENCE
scientific process and results should be visible, accessible, and understandable

OPEN (**ACCESSIBLE**) SCIENCE
data, tools, software, documentation, and publications should be accessible to all (FAIR)

OPEN (**INCLUSIVE**) SCIENCE
process and participants should welcome participation by and collaboration with diverse people and organizations

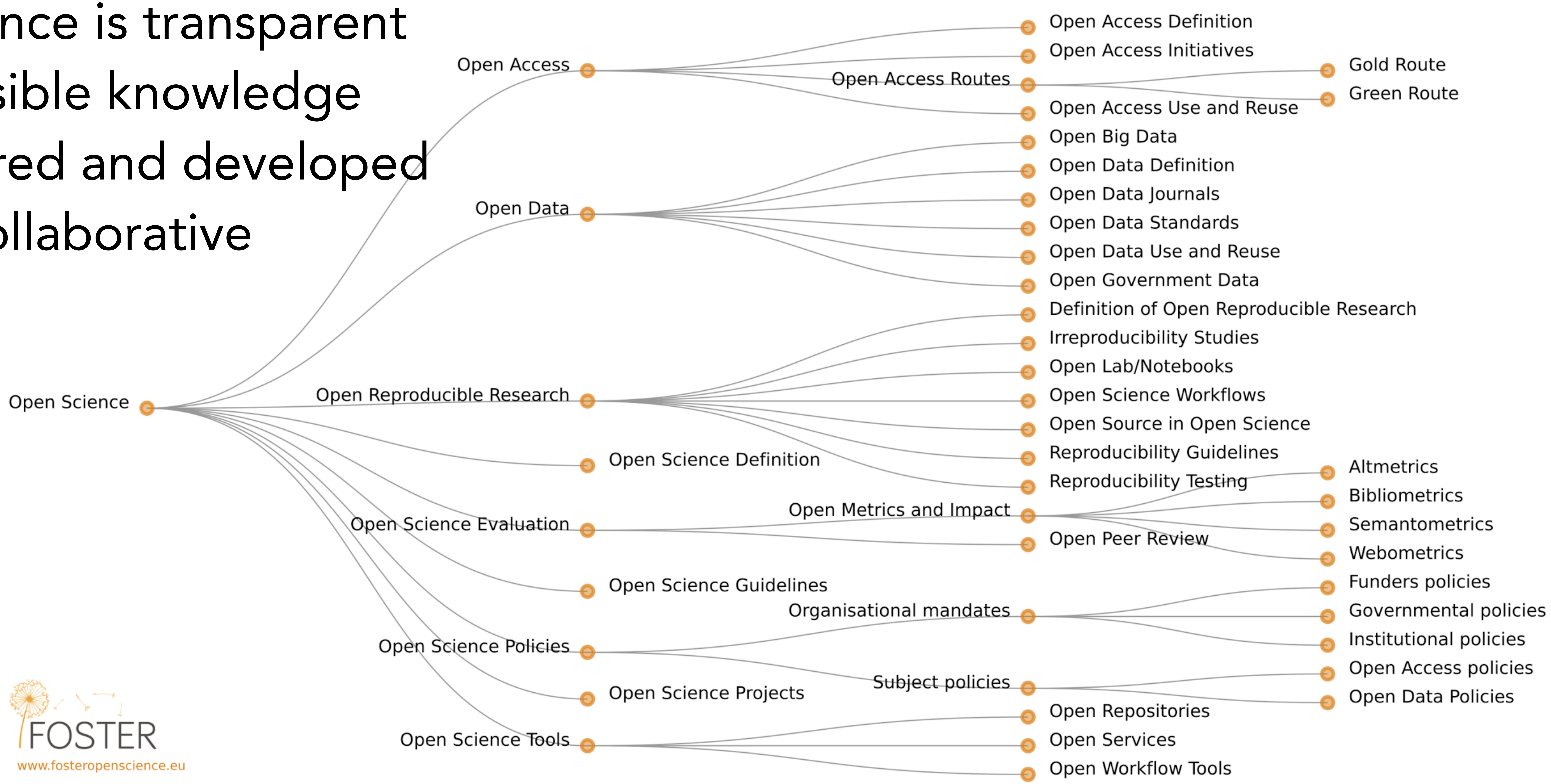
OPEN (**REPRODUCIBLE**) SCIENCE
scientific process and results should be open such that they are reproducible by members of the community

NASA Open Science Initiative
https://science.nasa.gov/open-science-overview

Open Science Framework

Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks.

Open Science Taxonomy



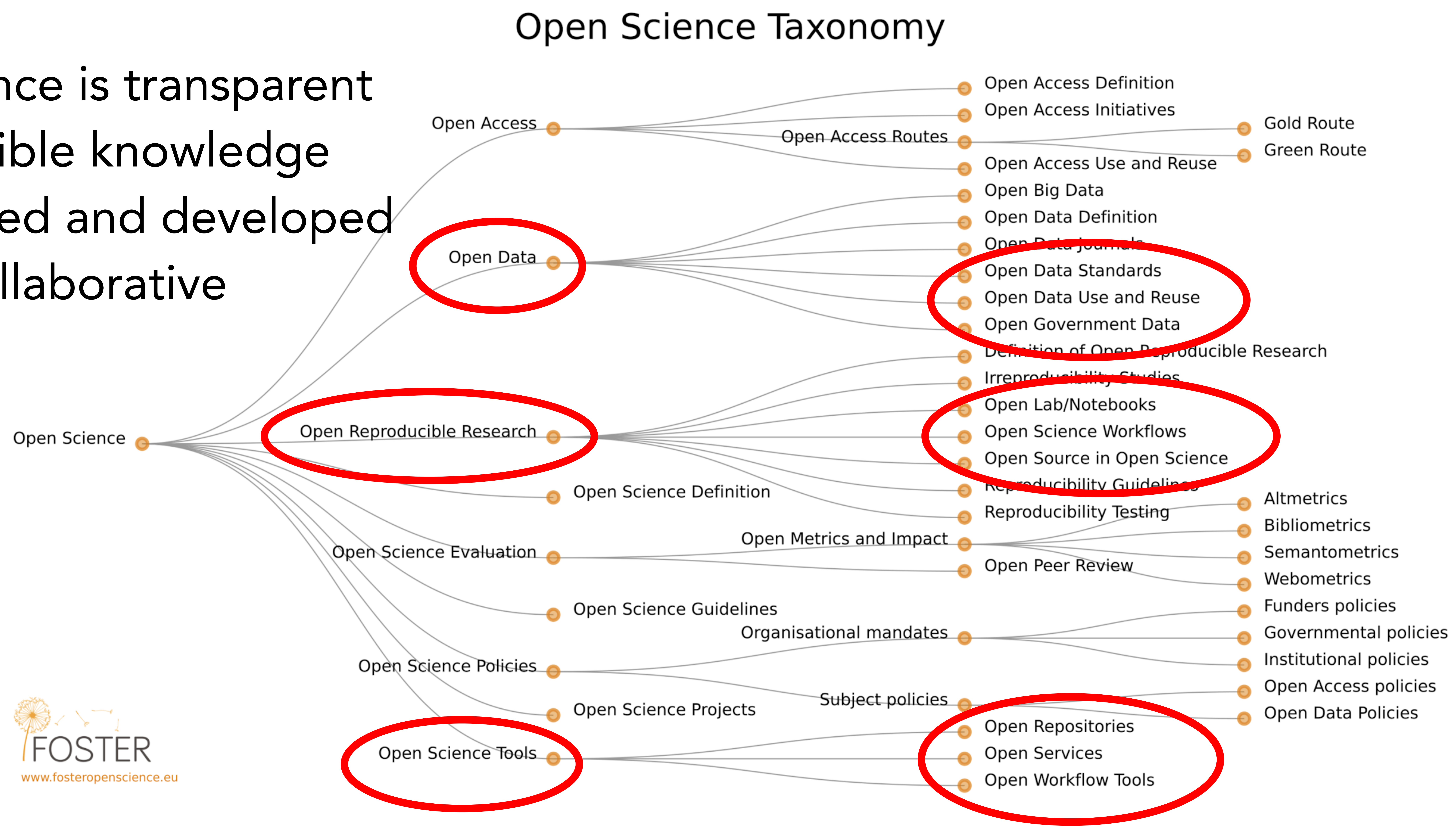
- Open Science
  - Open Access
    - Open Access Definition
    - Open Access Initiatives
    - Open Access Routes
      - Gold Route
      - Green Route
    - Open Access Use and Reuse
  - Open Data
    - Open Big Data
    - Open Data Definition
    - Open Data Journals
    - Open Data Standards
    - Open Data Use and Reuse
    - Open Government Data
  - Open Reproducible Research
    - Definition of Open Reproducible Research
    - Irreproducibility Studies
    - Open Lab/Notebooks
    - Open Science Workflows
    - Open Source in Open Science
    - Reproducibility Guidelines
    - Reproducibility Testing
  - Open Science Definition
  - Open Science Evaluation
    - Open Metrics and Impact
      - Altmetrics
      - Bibliometrics
      - Semantometrics
      - Webometrics
    - Open Peer Review
  - Open Science Guidelines
  - Open Science Policies
    - Organisational mandates
      - Funders policies
      - Governmental policies
      - Institutional policies
    - Subject policies
      - Open Access policies
      - Open Data Policies
  - Open Science Projects
  - Open Science Tools
    - Open Repositories
    - Open Services
    - Open Workflow Tools

FOSTER

www.fosteropenscience.eu

Open Science Framework

Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks.



Open Science Taxonomy

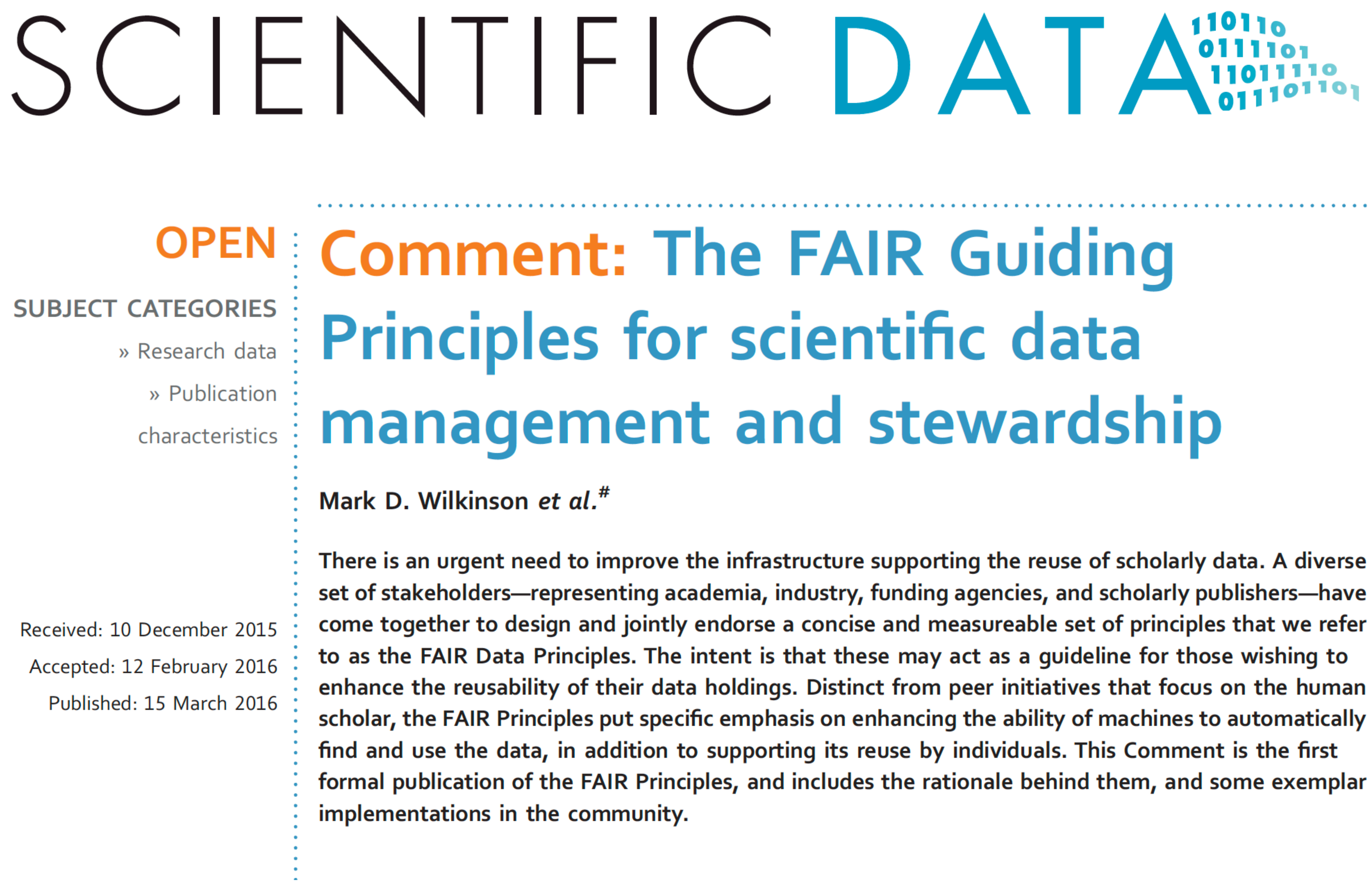
- Open Science
  - Open Access
    - Open Access Definition
    - Open Access Initiatives
    - Open Access Routes
      - Gold Route
      - Green Route
    - Open Access Use and Reuse
  - Open Data
    - Open Big Data
    - Open Data Definition
    - Open Data Journals
    - Open Data Standards
    - Open Data Use and Reuse
    - Open Government Data
  - Open Reproducible Research
    - Definition of Open Reproducible Research
    - Irreproducibility Studies
    - Open Lab/Notebooks
    - Open Science Workflows
    - Open Source in Open Science
    - Reproducibility Guidelines
    - Reproducibility Testing
  - Open Science Definition
  - Open Science Evaluation
    - Open Metrics and Impact
      - Altmetrics
      - Bibliometrics
      - Semantometrics
      - Webometrics
    - Open Peer Review
  - Open Science Guidelines
  - Open Science Policies
    - Organisational mandates
      - Funders policies
      - Governmental policies
      - Institutional policies
    - Subject policies
      - Open Access policies
      - Open Data Policies
  - Open Science Projects
  - Open Science Tools
    - Open Repositories
    - Open Services
    - Open Workflow Tools

FOSTER
www.fosteropenscience.eu

FAIR data principles

A set of principles aimed at making it easier to find and re-use scholarly data.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3, doi:10.1038/sdata.2016.18



SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Research data

» Publication characteristics

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

FAIR data principles: definitions

The principles refer to three types of entities:

- data (or any digital object)
- metadata: information about that digital object
- infrastructure: a searchable resource that houses data and metadata

FAIR data principles

F = FINDABLE

A = ACCESSIBLE

I = INTEROPERABLE

R = REUSABLE

FAIR data principles

F = FINDABLE

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers.

FAIR data principles

F = FINDABLE

A = ACCESSIBLE

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

FAIR data principles

F = FINDABLE

A = ACCESSIBLE

I = INTEROPERABLE

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

FAIR data principles

F = FINDABLE

A = ACCESSIBLE

I = INTEROPERABLE

R = REUSABLE

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Open Science and FAIR data principles: discussion (15 mins)

Group 1: you have generated data

- What do you need to do to make it useful for other researchers?

Group 2: you are using someone else's data

- What do you need to know for it to be useable for your research?

Open Science/FAIR principles need to be baked into scientific workflows, can rarely be reverse engineered.

F = FINDABLE

A = ACCESSIBLE

I = INTEROPERABLE

R = REUSABLE

Before Thursday

- set up GitHub account
- check course info on GitHub:

  https://github.com/sophieclayton/OEAS805_envdatasci
- join Slack group (invites sent):

  shorturl.at/dnOR5
- read Wilkinson FAIR data paper (in readings folder on GitHub)

Thursday (every) class:

- bring laptop