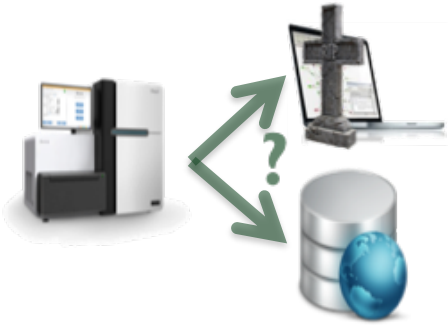
Research Data Management

Niclas Jareborg, NBIS
niclas.jareborg@nbis.se

Introduction to NGS course

Why manage research data?

- To make your research easier!
- To stop yourself drowning in irrelevant stuff
- In case you need the data later
- To avoid accusations of fraud or bad science
- To share your data for others to use and learn from
- To get credit for producing it
- Because funders or your organisation require it



Well-managed data opens up opportunities for re-use, integration and new science

Accusation of fraud

Science

LETTERS

Cite as: J. Berg., *Science* 10.1126/science.aan5763 (2017).

Editorial Retraction

Jeremy Berg

Editor-in-Chief

After an investigation, the Central Ethical Review Board in Sweden has recommended the retraction of the Report "Environmentally relevant concentrations of microplastic particles influence larval fish ecology," by Oona M. Lönnstedt and Peter Eklöv, published in *Science* on 3 June 2016 (*1*). *Science* ran an Editorial Expression of Concern regarding the Report on 1 December 2016 (*2*). The Review Board's report, dated 21 April 2017, cited the following reasons for their recommendation: (i) lack of ethical approval for the experiments; (ii) absence of original data for the experiments reported in the paper; (iii) widespread lack of clarity concerning how the experiments were conducted. Although the authors have told *Science* that they disagree with elements of the Board's report, and although Uppsala University has not yet concluded its own investigation, the weight of evidence is that the paper should now be retracted. In light of the Board's recommendation and a 28 April 2017 request from the authors to retract the paper, *Science* is retracting the paper in full.

REFERENCES

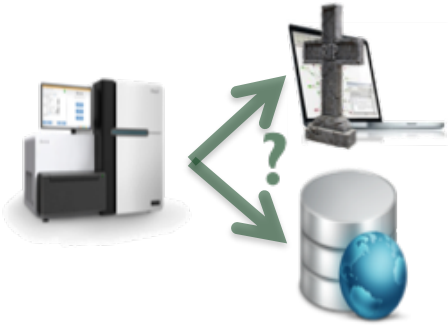
1. O. M. Lönnstedt, P. Eklöv, *Science* **352**, 1213 (2016).
2. J. Berg, *Science* **354**, 1242 (2016); published online 1 December 2016.

Published online 3 May 2017
10.1126/science.aan5763

- Be able to show that you have done what you say you have done
- Universities want to avoid bad press!

Why manage research data?

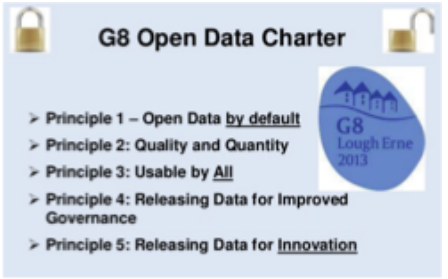
- To make your research easier!
- To stop yourself drowning in irrelevant stuff
- In case you need the data later
- To avoid accusations of fraud or bad science
- To share your data for others to use and learn from
- To get credit for producing it
- Because funders or your organisation require it



Well-managed data opens up opportunities for re-use, integration and new science

Open Access to research data

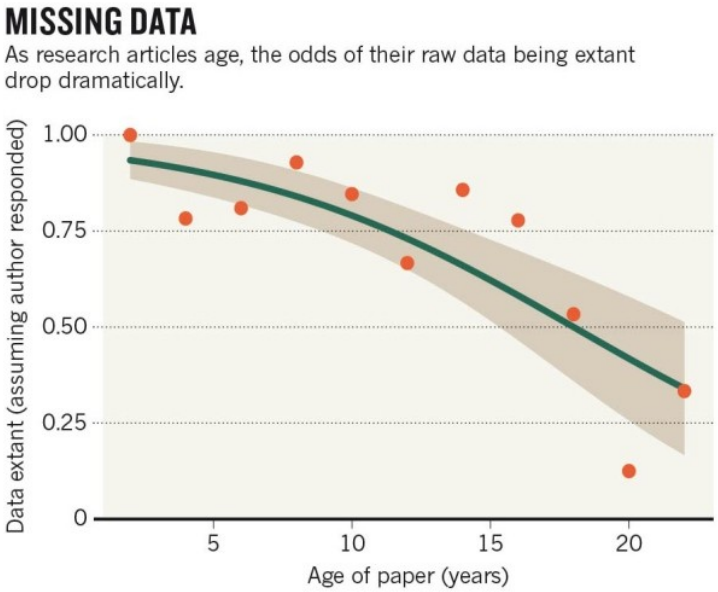
- *The practice of providing on-line access to scientific information that is free of charge to the end-user and that is re-usable.*
  - Does not necessarily mean unrestricted access, e.g. for sensitive personal data
- Strong international movement towards Open Access (OA)
- European Commission recommended the member states to establish national guidelines for OA
  - Swedish Research Council (VR) submitted proposal to the government Jan 2015
- Research bill 2017–2020 – *28 Nov 2016*
  - "*The aim of the government is that all scientific publications that are the result of publicly funded research should be openly accessible as soon as they are published. Likewise,* ***research data*** *underlying scientific publications should be* ***openly accessible*** *at the time of publication.*"
    [my translation]



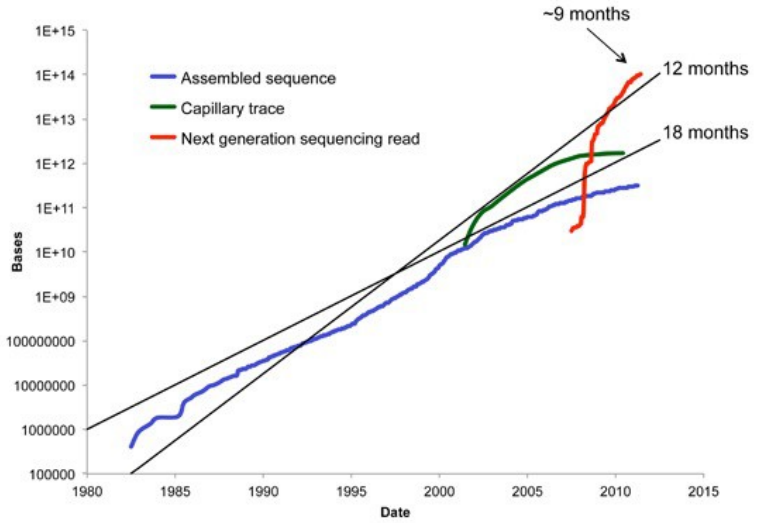
Why Open Access?

- Democracy and transparency
  - Publicly funded research data should be accessible to all
  - Published results and conclusions should be possible to check by others
- Research
  - Enables others to combine data, address new questions, and develop new analytical methods
  - Reduce duplication and waste
- Innovation and utilization outside research
  - Public authorities, companies, and private persons outside research can make use of the data
- Citation
  - Citation of data will be a merit for the researcher that produced it

Data loss is real and significant, while data growth is staggering



Nature news, 19 December 2013



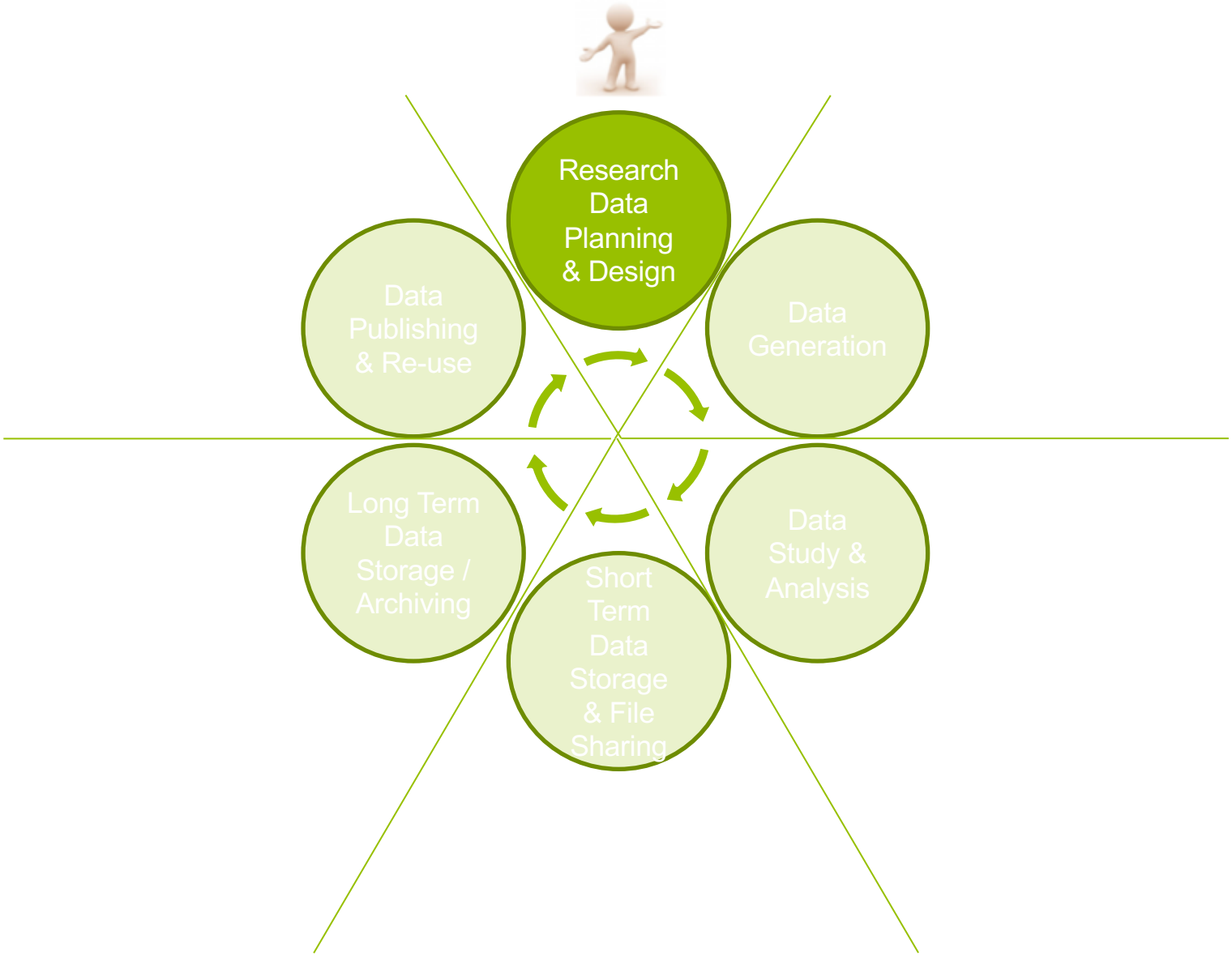
- DNA sequence data is doubling every 6-8 months and looks to continue for this decade
- Projected to surpass astronomy data in the coming decade

'Oops, that link was the laptop of my PhD student'

Slide stolen from Barend Mons

The Research Data Life Cycle

Planning & Design



Planning & Design

- Data Management planning
  - Data types
    - Sizes, were to store, etc
  - **Metadata**
    - Study, Samples, Experiments, etc
    - Use standards!

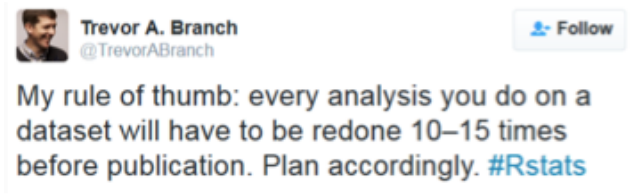
- *Data Management Plans*
  - Will become a standard part of the research funding application process
  - What will be collected?, Size?, Organized?, Documented?, Stored and preserved?, Disseminated?, Policies?, Budget?

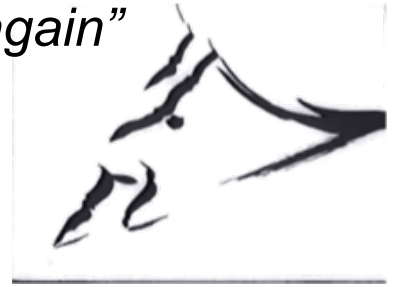
Study & Analysis

Human derived data

Study & Analysis

- Guiding principle
  - *“Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.”*
- Research reality
  - *"Everything you do, you will have to do over and over again”*
    – Murphy’s law

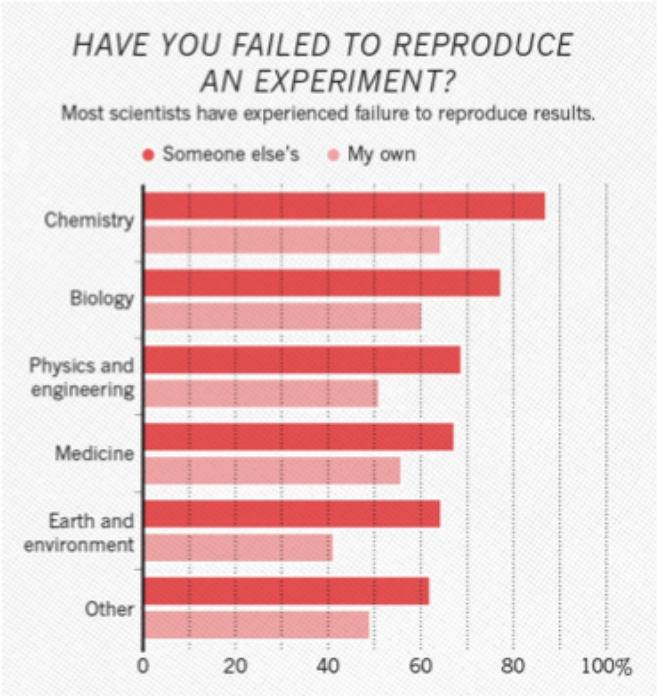


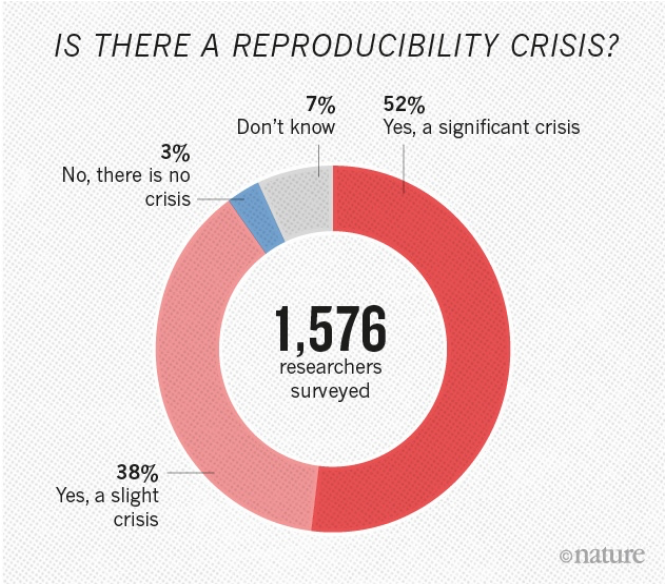


- Structuring data for analysis
  - Poor organizational choices lead to significantly slower research progress.
  - It is critical to make results reproducible.

"I think you should be more explicit here in step two."

A reproducibility crisis





A recent survey in Nature revealed that irreproducible experiments are a problem across all domains of science[1].

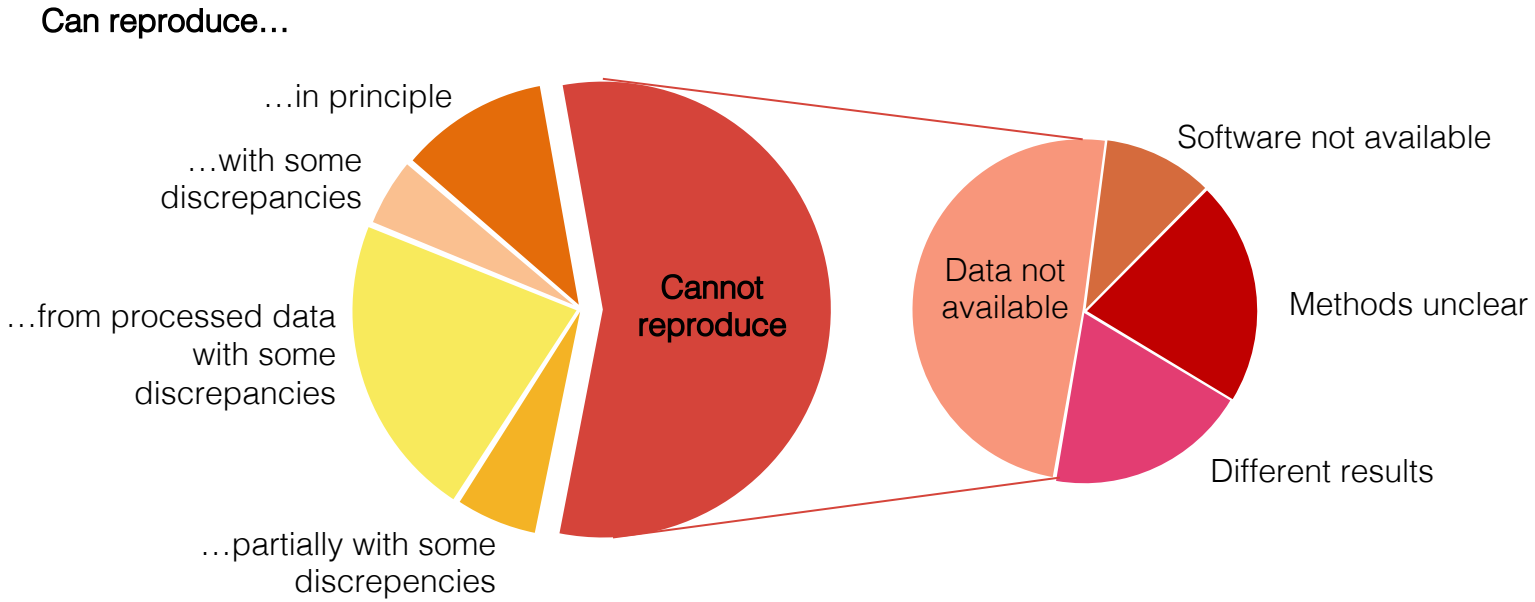
Medicine is among the most affected research fields. A study in Nature found that 47 out of 53 medical research papers focused on cancer research were irreproducible[2].

Common features were failure to show all the data and inappropriate use of statistical tests.

[1] "1,500 scientists lift the lid on reproducibility". Nature. 533: 452–454
[2] Begley, C. G.; Ellis, L. M. (2012). "Drug development: Raise standards for preclinical cancer research". Nature. 483 (7391): 531–533.

A reproducibility crisis

Reproduction of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:



Summary of the efforts to replicate the published analyses.
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses. *Nature Genetics* **41** (2009) doi:10.1038/ng.295

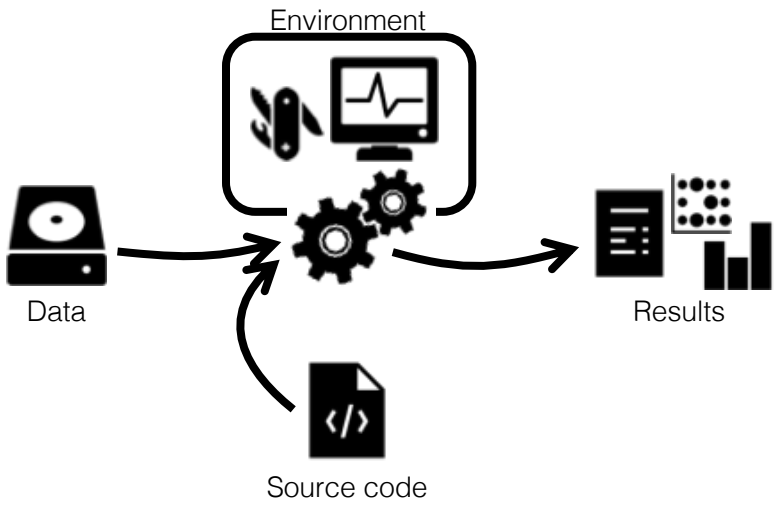
What do we mean by reproducible research?

Code	Data: Same	Data: Different
Same	Reproducible	Replicable
Different	Robust	Generalizable

Is it really any point doing this?

- Primarily for ones own benefit! Organized, efficient, in control. Dynamic team members.
- Transparent what has been done
- Some will be interested in parts of the analysis. Make it easy to redo, then adapt to own data.

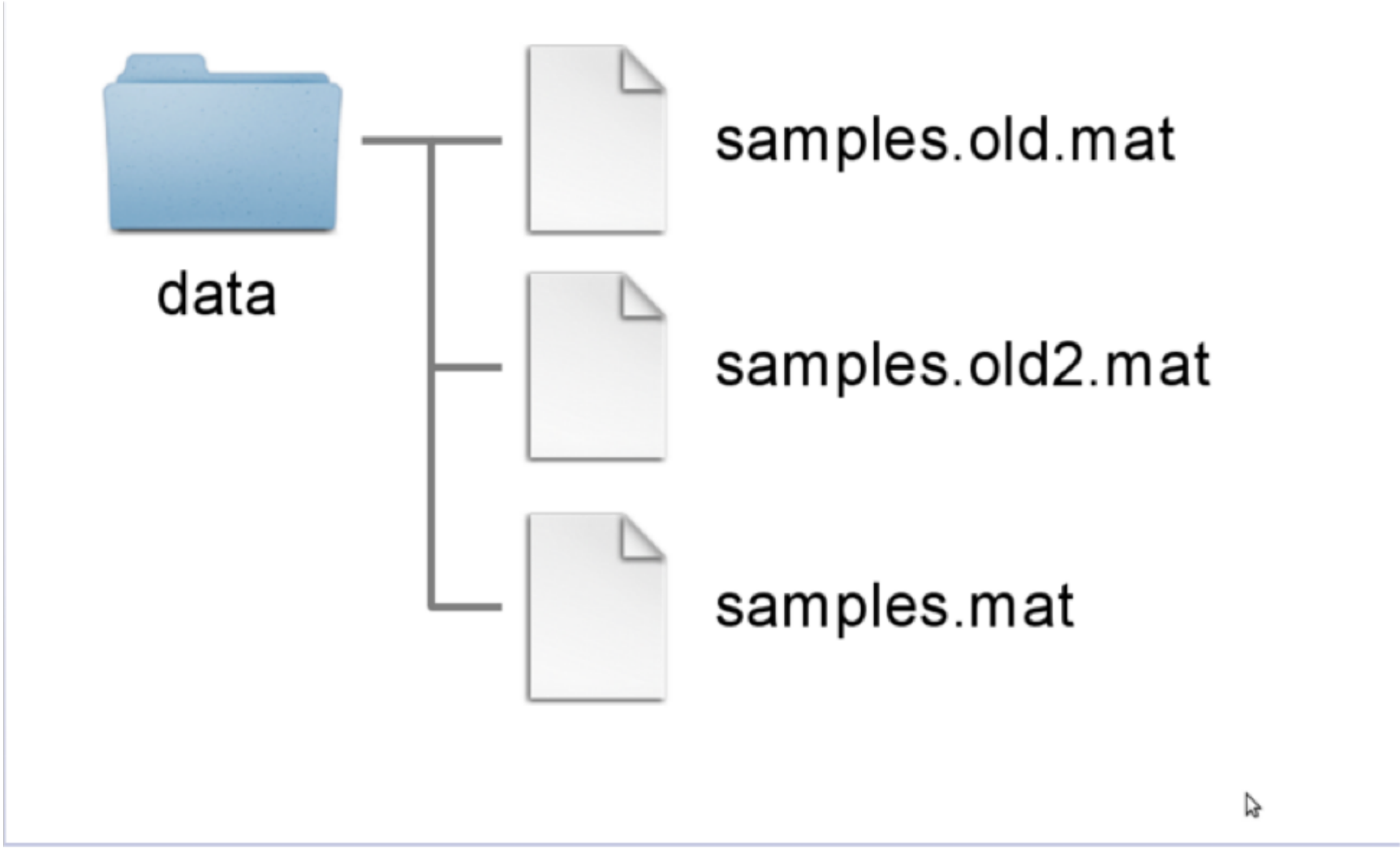
All parts of a bioinformatics analysis have to be reproducible:



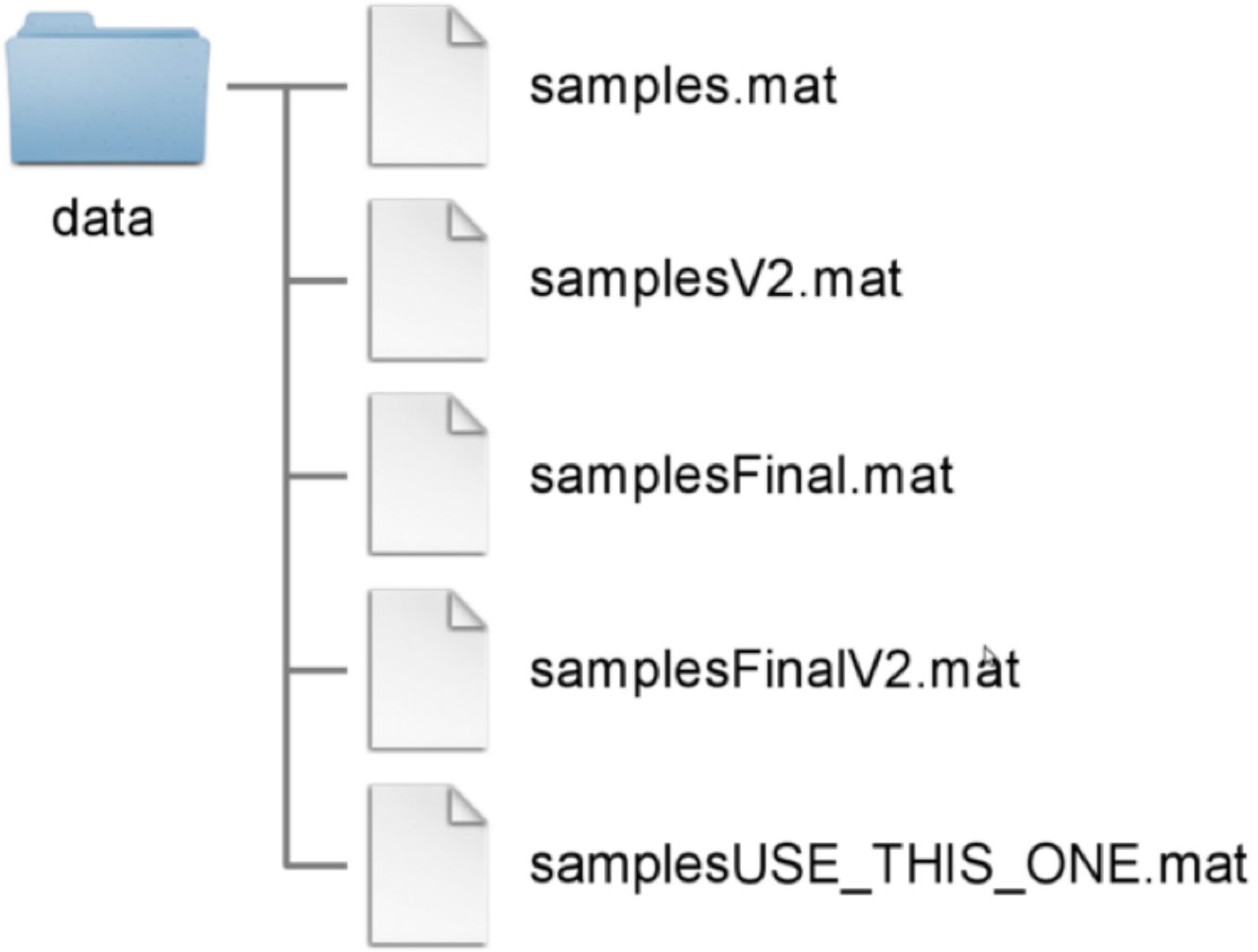
First step - Organization

Now what?

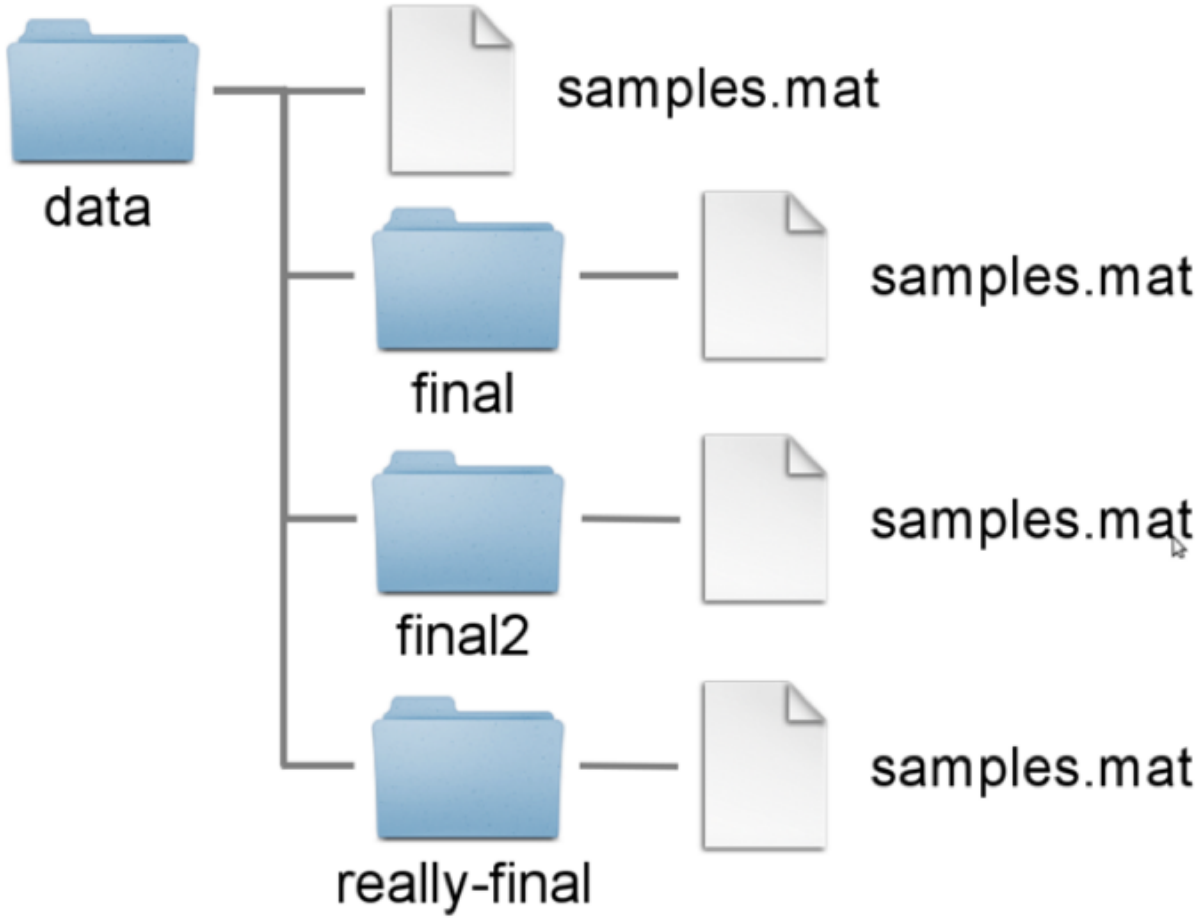
I guess this is alright



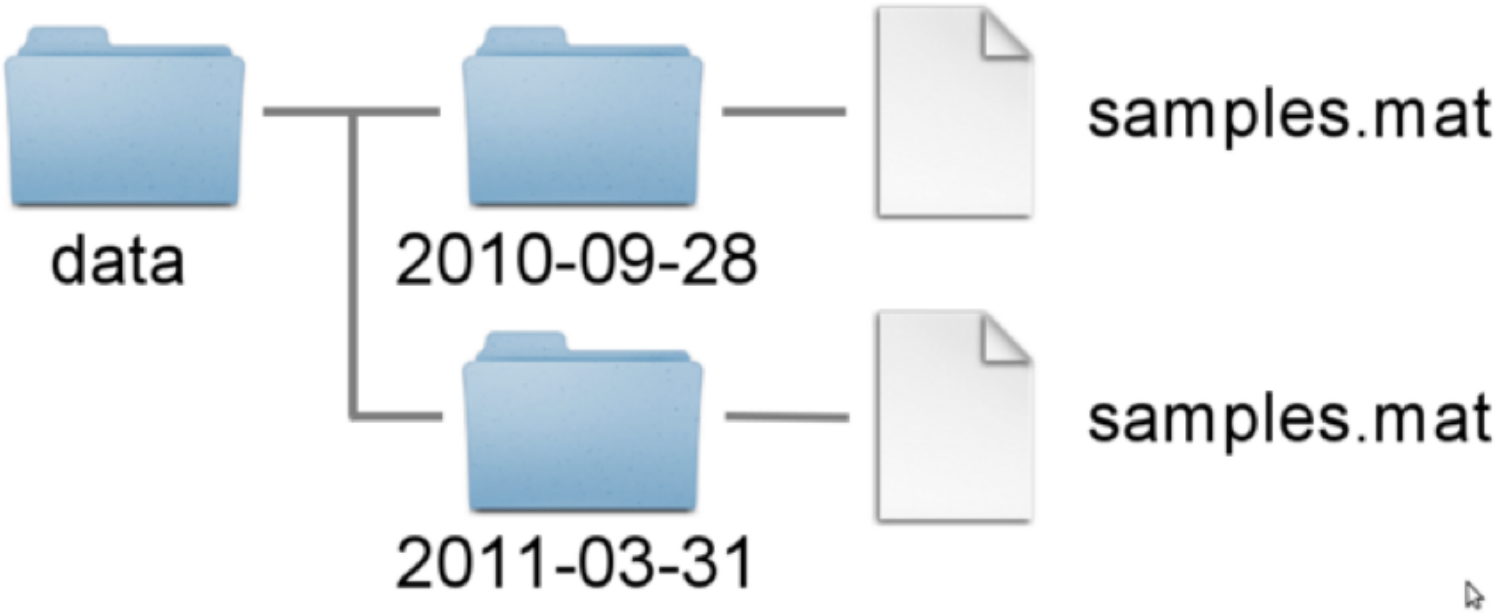
Which one is the most recent?



Another (bad) common approach



A possible solution

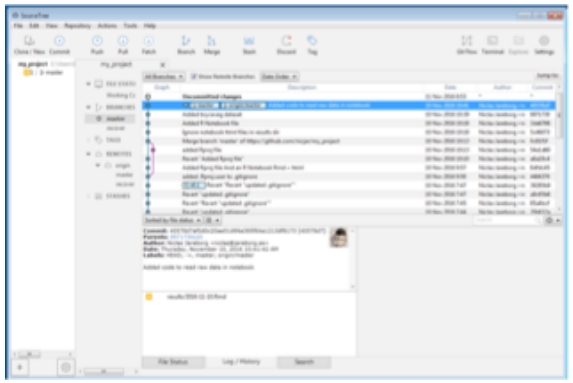


Suggested best practices

- There is a **folder for the raw data**, which do not get altered, or intermixed with data that is the result of manual or programmatic manipulation. I.e., derived data is kept separate from raw data, and **raw data are not duplicated**.
- **Code is kept separate from data**.
- Use a **version control system** (at least for code) – e.g. **git**
- There is a **scratch directory for experimentation**. Everything in the scratch directory can be deleted at any time without negative impact.
- There should be a **README in every directory**, describing the purpose of the directory and its contents.
- Use **non-proprietary formats** – *.csv* rather than *.xlsx*
- Etc…

Version control

- What is it?
  - A system that keeps records of your changes
  - Allows for collaborative development
  - Allows you to know who made what changes and when
  - Allows you to revert any changes and go back to a previous state
- Several systems available
  - Git, RCS, CVS, SVN, Perforce, Mercurial, Bazaar
  - Git
    - Command line & GUIs
    - Remote repository hosting
      - GitHub, Bitbucket, etc



Suggested best practices

- There is a **folder for the raw data**, which do not get altered, or intermixed with data that is the result of manual or programmatic manipulation. I.e., derived data is kept separate from raw data, and **raw data are not duplicated**.
- **Code is kept separate from data**.
- Use a **version control system** (at least for code) – e.g. **git**
- There is a **scratch directory for experimentation**. Everything in the scratch directory can be deleted at any time without negative impact.
- There should be a **README in every directory**, describing the purpose of the directory and its contents.
- Use **non-proprietary formats** – *.csv* rather than *.xlsx*
- Etc…

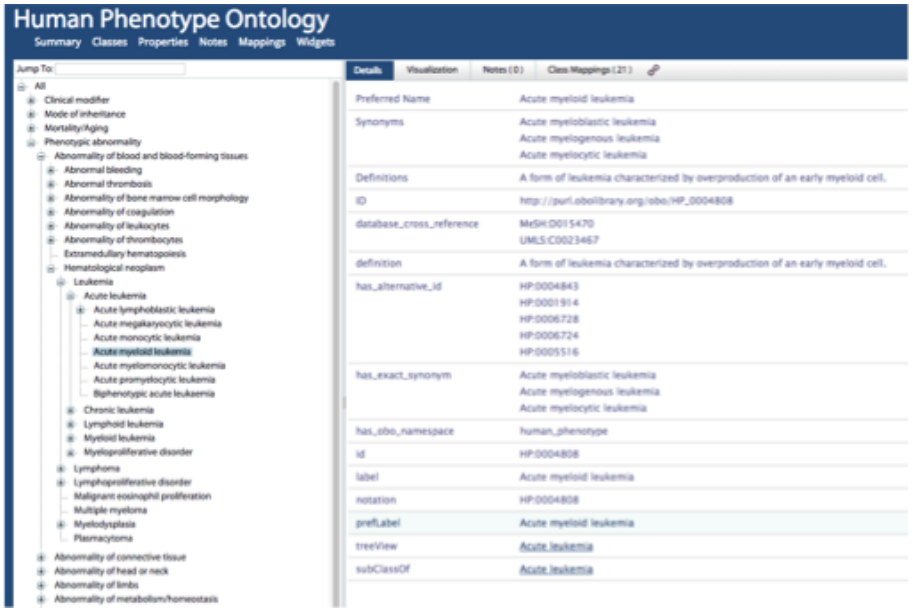
Non-proprietary formats

- A text-based format is more future-safe, than a proprietary binary format by a commercial vendor
- ***Markdown*** is a nice way of getting nice output from text.
  - Simple & readable formating
  - Can be converted to lots of different outputs
    - HTML, pdf, MS Word, slides etc

- *Never, never, never use **Excel** for scientific **analysis**!*
  - Script your analysis – bash, python, R, …

Still missing something

- Need context → document **metadata**
  - How was the data generated?
  - From what was the data generated?
  - What where the experimental conditions?
  - Etc
- Use standards
  - Controlled vocabularies / Ontologies
  - *Not straight-forward…*

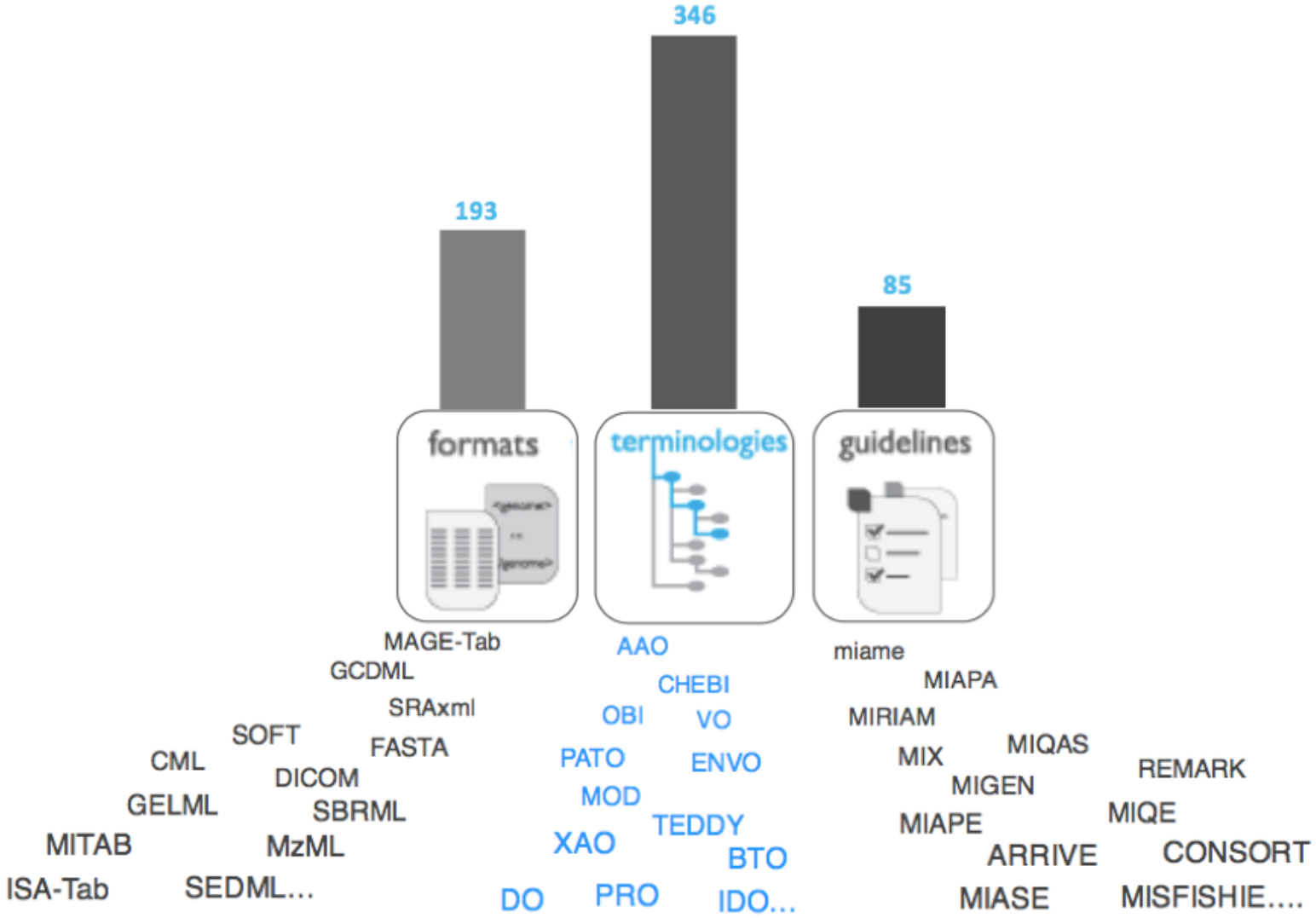


NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

FAIRsharing.org

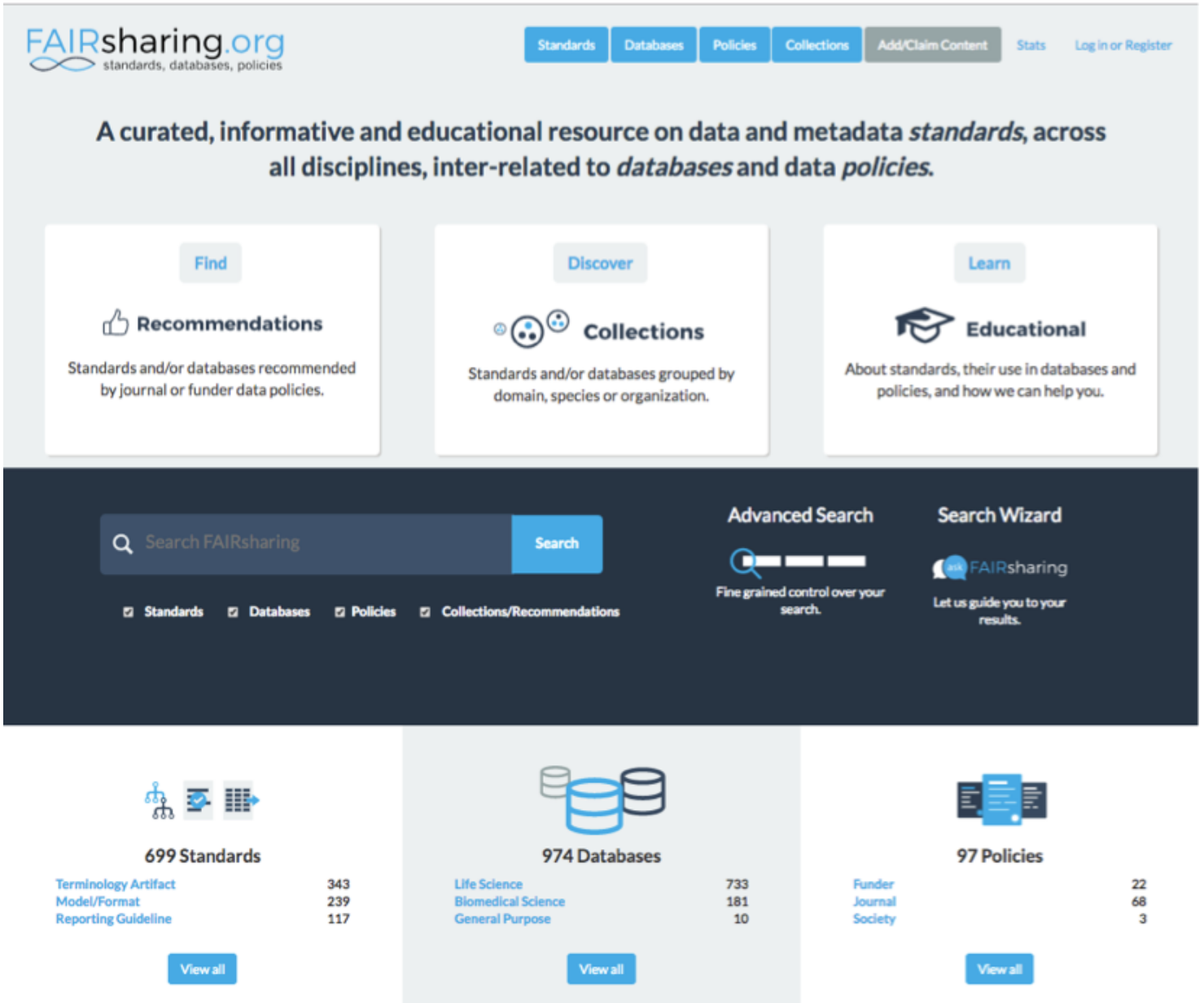
SciLifeLab

In the life sciences there are >600 *content standards*



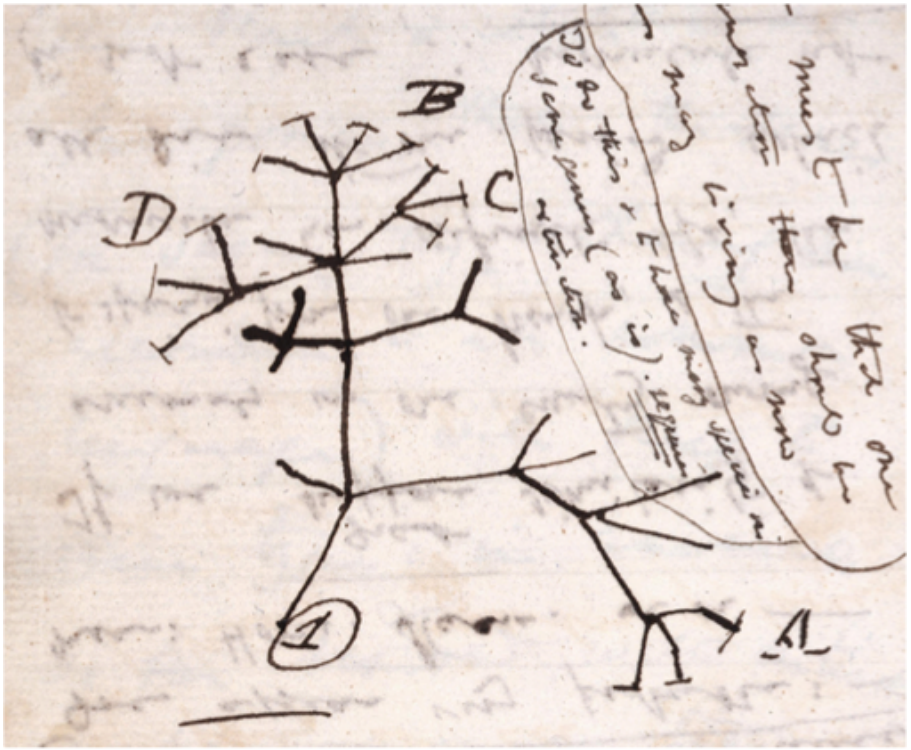
FAIRsharing.org

(was biosharing.org)



Lab notebooks

- Why?
  - You have to understand what you have done
  - **Others should be able to reproduce what you have done**

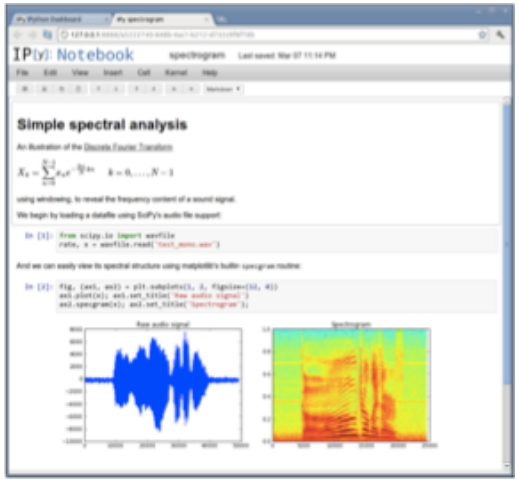


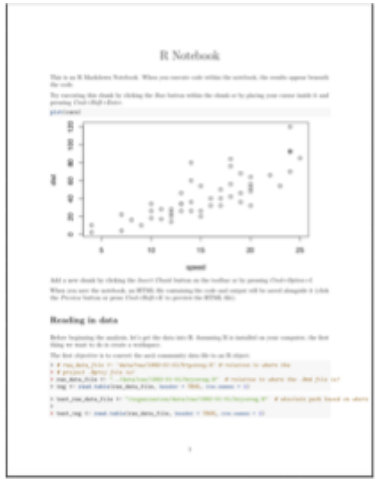
Lab notes – useful practices

- Put in *results* directory
- *Dated* entries
- Entries relatively verbose
- Link to *data* and *code* (including versions)
    - Point to commands run and results generated
- Embedded images or tables showing results of analysis done
- Observations, Conclusions, and *ideas* for future work
- Also document analysis that *doesn't* work, so that it can be understood why you choose a particular way of doing the analysis in the end

Where to keep notes

- Paper Notebook
- Word processor program / Text files
- Electronic Lab Notebooks
- 'Interactive' Electronic Notebooks
  - e.g. jupyther, R Notebooks in RStudio
  - Plain text - work well with version control (Markdown)
  - Embed and execute code
  - Convert to other output formats
    - html, pdf, word





R Markdown

- R Markdown makes your analysis more reproducible by connecting your code, figures and descriptive text.
- You can use it to make reproducible reports, rather than e.g. copy-pasting figures into a Word document.
- You can also use it as a notebook, in the same way as lab notebooks are used in a wet lab setting

```
---
title: "Differential expression analysis using linear models"
output:
  pdf_document
---

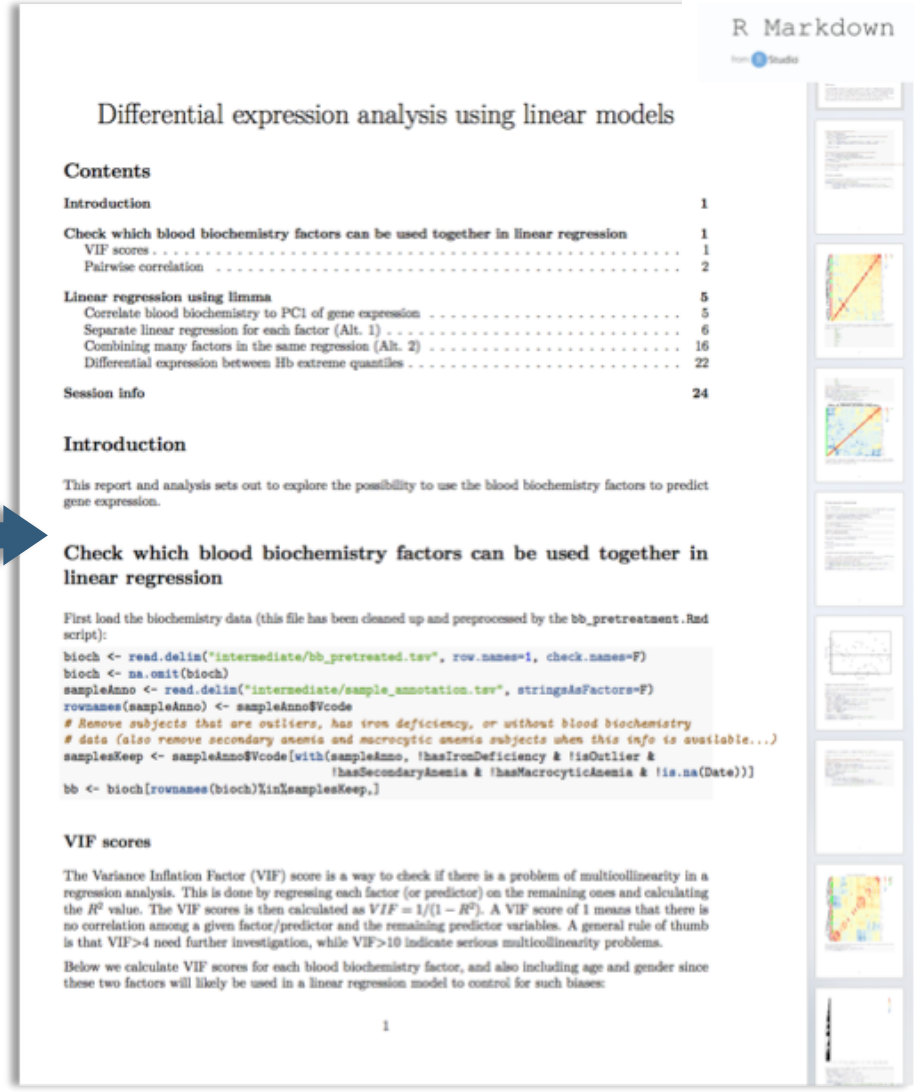

```{r setup, echo=F, message=FALSE, warning=FALSE, eval=T}
library(knitr)
opts_knit$set(root.dir=normalizePath('../'))
```
```{r, echo=F, message=FALSE, warning=FALSE, eval=T}
source("source/rmarkdown_functions.R")
```

Introduction
This report and analysis sets out to explore the possibility to use the blood biochemistry
factors to predict gene expression.

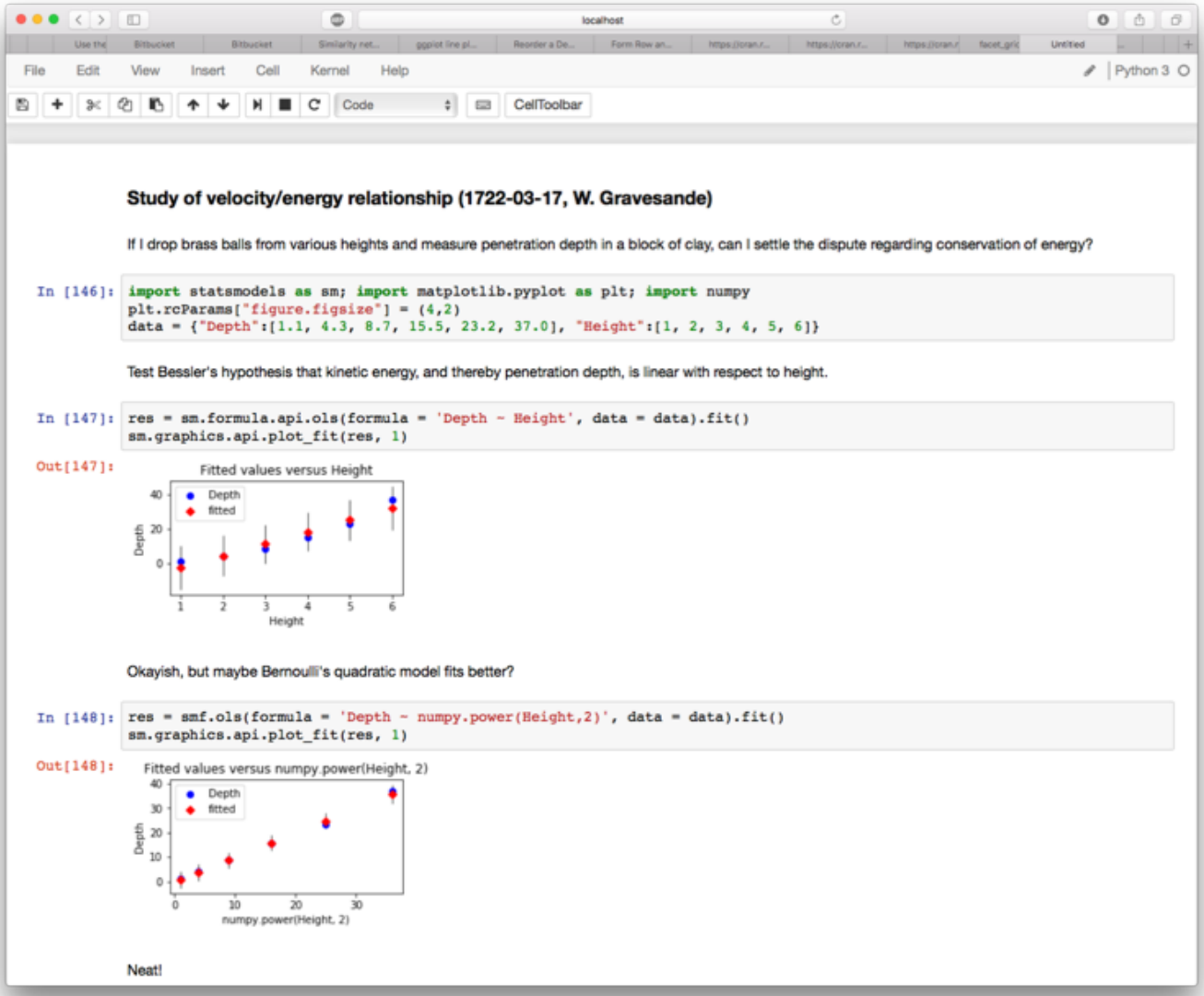
Check which blood biochemistry factors can be used together in linear regression

First load the biochemistry data (this file has been cleaned up and preprocessed by the `bb_pretreatment.Rmd`
script):
```{r, echo=TRUE, message=FALSE, warning=FALSE}
bioch <- read.delim("intermediate/bb_pretreated.tsv", row.names=1, check.names=F)
bioch <- na.omit(bioch)
sampleAnno <- read.delim("intermediate/sample_annotation.tsv", stringsAsFactors=F)
rownames(sampleAnno) <- sampleAnno$Vcode
# Remove subjects that are outliers, has iron deficiency, or without blood biochemistry
# data (also remove secondary anemia and macrocytic anemia subjects when this info is available...)
samplesKeep <- sampleAnno$Vcode[with(sampleAnno, !hasIronDeficiency & !isOutlier &
                                     !hasSecondaryAnemia & !hasMacrocyticAnemia & !is.na(Date))]
bb <- bioch[rownames(bioch)%in%samplesKeep,]

```
```

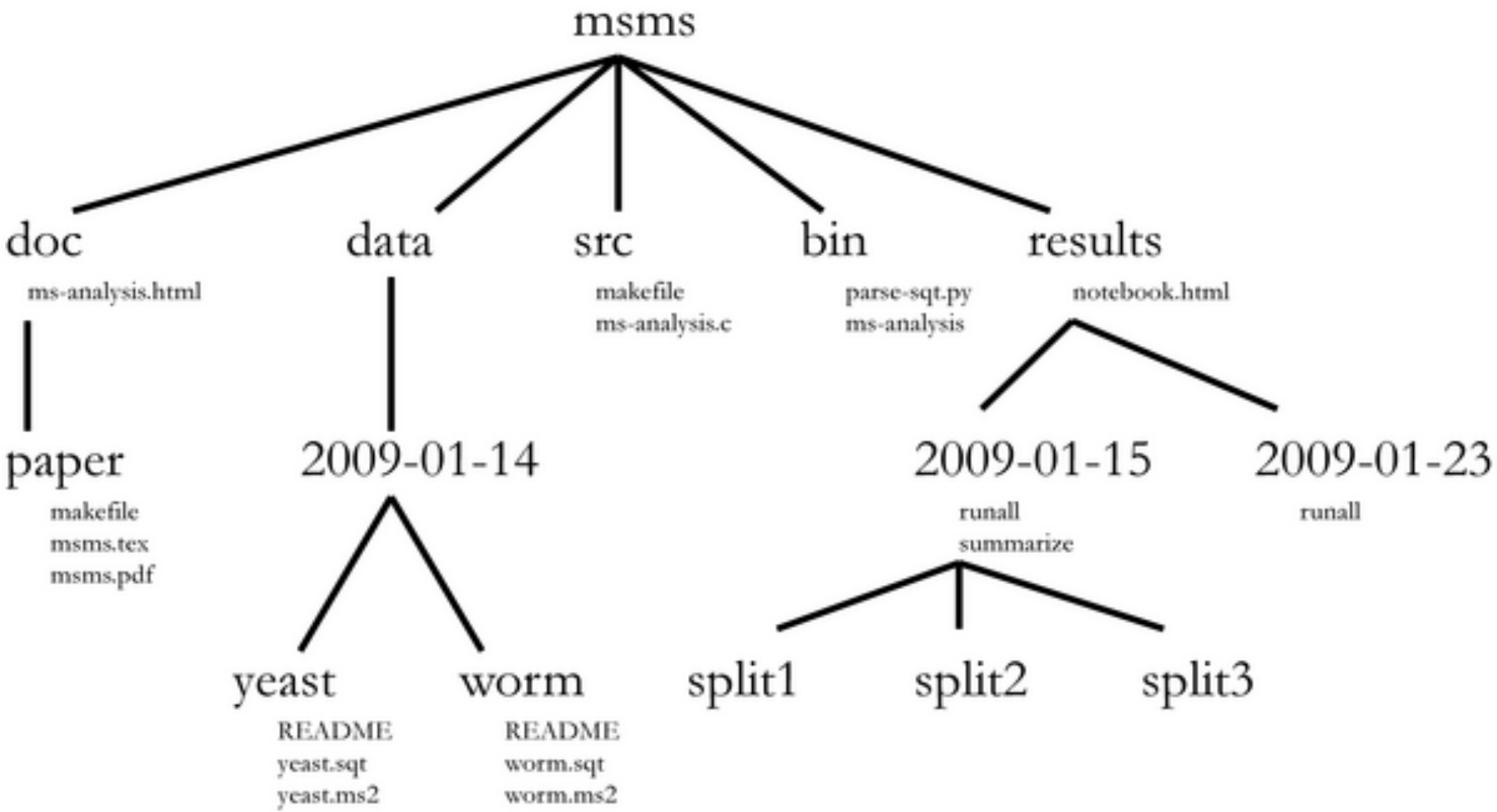


jupyter



- In-browser editing for code, with automatic syntax highlighting, indentation, and tab completion/introspection.
- The ability to execute code from the browser, with the results of computations attached to the code which generated them.

Directory structure for a sample project



Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. doi:10.1371/journal.pcbi.1000424
http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1000424

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

The project directory

```
project
|- doc/          documentation for the study
|
|- data/         raw and primary data, essentially all input files, never edit!
|   |- raw_external/
|   |- raw_internal/
|   |- meta/
|
|- code/         all code needed to go from input files to final results
|- notebooks/
|
|- intermediate/   output files from different analysis steps, can be deleted
|- scratch/      temporary files that can be safely deleted or lost
|- logs/         logs from the different analysis steps
|
|- results/      output from workflows and analyses
|   |- figures/
|   |- tables/
|   |- reports/
```

Adapted from https://github.com/Reproducible-Science-Curriculum/rr-init

Project organization

- There's no perfect set-up
  - Pick one! e.g.
    - https://github.com/chendaniely/computational-project-cookie-cutter
    - https://github.com/Reproducible-Science-Curriculum/rr-init
    - https://github.com/nylander/ptemplate
    - …
- Communicate structure to collaborators
- Document as you go
- Done well it might reduce post-project explaining

Reproducible research for bioinformatics projects

Leif Väremo (leif.varemo@scilifelab.se)
Rasmus Ågren (rasmus.agren@scilifelab.se)
Bioinformatics long-term support (WABI)

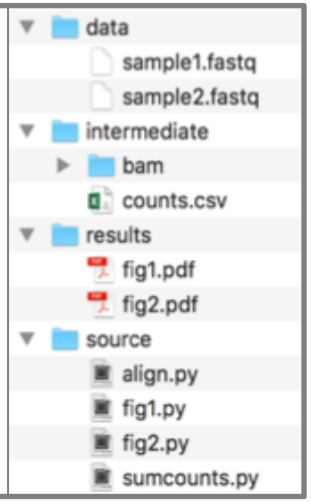
Everything can be a project

Divide your work into distinct projects and keep all files needed to go from raw data to final results in a dedicated directory with relevant subdirectories (see example).

Many software support the "project way of working", e.g. Rstudio and the text editors Sublime Text and Atom.

Tip! Learn how to use git, a widely used system (both in academia and industry) for version controlling and collaborating on code.

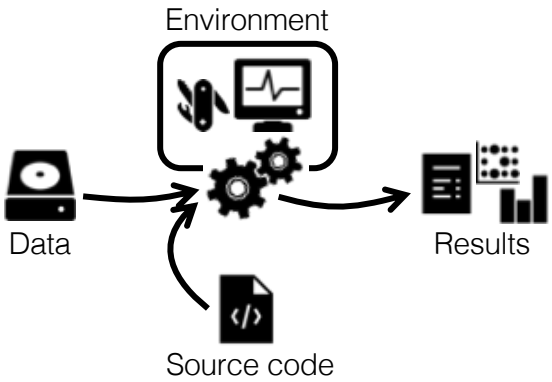
https://git-scm.com/



Take control of your research by making it reproducible!

By moving towards a reproducible way of working you will quickly realize that you at the same time make your own life a lot easier! The added effort pays off by gain in control, organization and efficiency.

Below are all the components of a bioinformatics project that have to reproducible.



Connect your results with the code

Rmarkdown and Jupyter notebooks blur the boundaries between code and its output. They allow you to add non-code text (markdown) to your code. This generates a report containing custom formatted text, as well as figures and tables together with the code that generated them.

R Markdown

http://rmarkdown.rstudio.com/

jupyter

http://jupyter.org/

Master your dependencies

- Full reproducibility requires the possibility to recreate the system that was originally used to generate the results.
- Conda is package, dependency, and env-ironment manager that makes it easy to install (most) software that you need for your project.
- Your environment can be exported in a simple text format and reinstalled by Conda on another system.

CONDA https://conda.io

For the advanced

- Conda cannot always *completely* recreate the system, which is required for proper repro-ducibility.
- A solution is to package your project in an isolated Docker container, together with all its dependencies and libraries.
- A vision is that every new bioinformatics publication is accompanied by a publically available Docker container!
- Singularity is an alternative to Docker which runs better on HPC clusters.

docker https://www.docker.com/

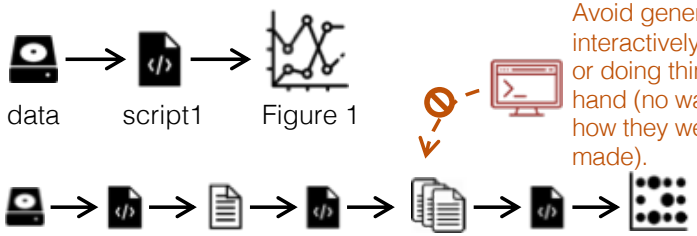
Singularity http://singularity.lbl.gov/

Treasure your data

- Consider your input data static. Keep it readonly!
- Don't make *different* versions. If you need to preprocess it in any way, script it so you can recreate the steps (see box below).
- Backup! Keep redundant copies in different physical locations.
- Strive towards uploading it to its final destination already at the beginning of a project (e.g. specific repositories such as ENA, or GeneExpress, or general repositories such as Dryad or Figshare).

Organize your coding

- Write scripts/functions/notebooks for specific tasks (connect raw data to final results)
- Keep parameters separate (e.g. top of file, or input arguments)



Avoid generating files interactively on the fly or doing things by hand (no way to track how they were made).

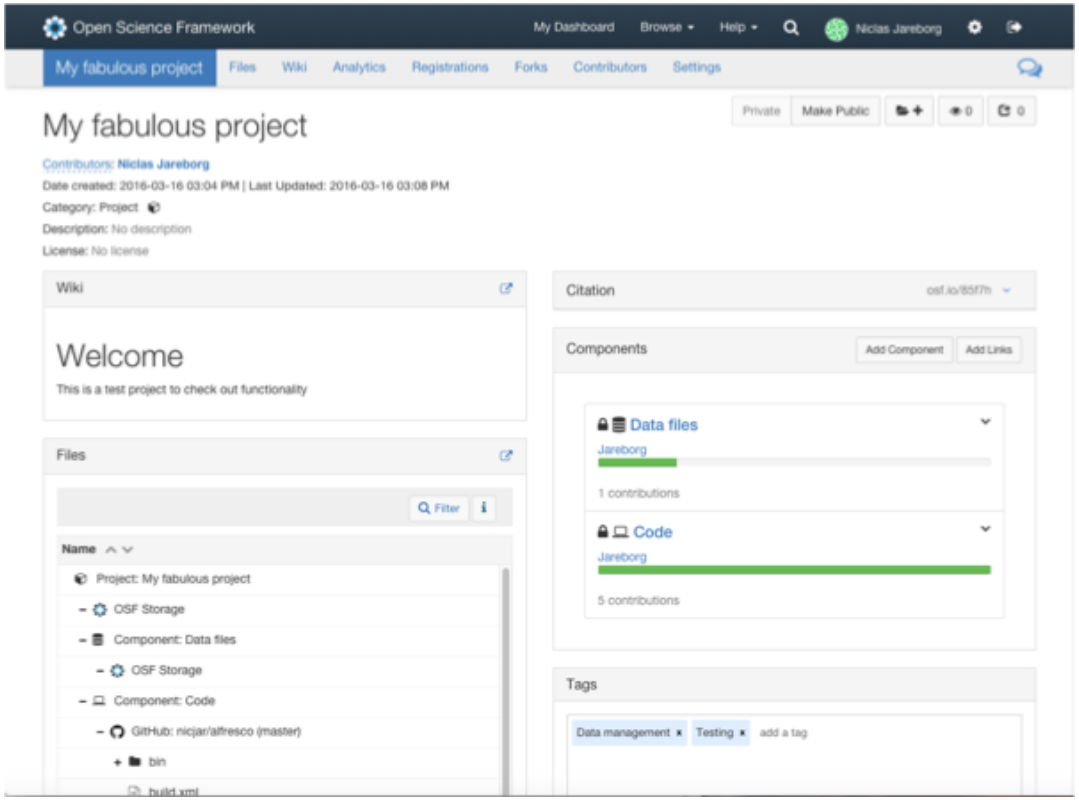
For the advanced

As projects grow, it becomes increasingly difficult to keep track of all the parts and how they fit together. Snakemake is a workflow management system that keeps track of how your files tie together, from raw data and scripts to final figures. If anything changes (script code, parameters, software version, etc) it will know what parts to rerun in order to have up to date and reproducible results.

Snakemake
https://snakemake.readthedocs.io/

Project collaboration tools

- Open Science Framework – http://osf.io
  - Organize research project documentation and outputs
  - Control access for collaboration
  - 3rd party integrations
    - Google Drive
    - Dropbox
    - GitHub
    - External links
    - Etc
  - Persistent identifiers
  - Publish article preprints



NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Personal data

Personal data - Legislation

- Personal Data Act (*Personuppgiftslagen (PUL)*)
- Act concerning the Ethical Review of Research Involving Humans (*Lag om etikprövning av forskning som avser människor*)

Personal Data Act

- All kinds of information that is directly or indirectly referable to a natural person who is alive constitute personal data

- Sensitive data
  - It is **prohibited** to process personal data that discloses *ethnic origin*, *political opinions*, *religious or philosophical convictions, membership of trade unions*, as well as personal data relating to ***health*** or *sexual life*.
  - Sensitive personal data can be handled for **research purposes** if person has given **explicit consent**

- The Data Inspection Board (*Datainspektionen*) is the supervisory authority under the Personal Data Act

- May 2018: **General Data Protection Regulation** (GDPR)
  - New Swedish laws
    - **Data Protection Act** (*Dataskyddslag*)
    - **Research Data Act** (*Forskningsdatalag*)
    - *Several other*

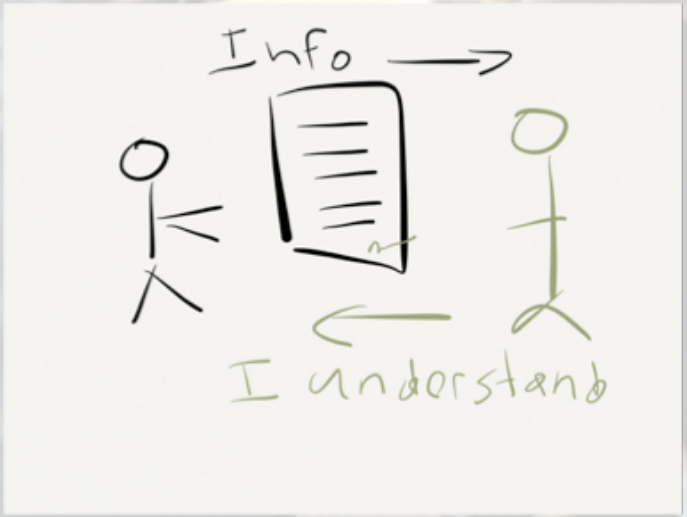
Personal Data Act – cont.

- The (legal) person that decides why and how personal data should be processed is called the **controller of personal data** (*personuppgiftsansvarig*)
  - e.g. the employing university
- The controller of personal data can delegate processing of personal data to a **personal data assistant** (*personuppgiftsbiträde*)
  - e.g. UPPMAX/Uppsala university
- A **personal data representative** (*personuppgiftsombud*) is a natural person who, on the assignment of the controller, shall ensure that personal data is processed in a lawful and proper manner
- Obligation to report handling of personal data to the Data Inspection Board
  - Or, notify the Board of the named representative

Act concerning the Ethical Review

- Research that concerns studies of biological material that has been taken from a living person and that can be traced back to that person may only be conducted if it has been approved subsequent to an ethical vetting

- Informed consent
  - The subject must be informed about the purpose or the research and the consequences and risks that the research might entail
  - The subject must consent



Genetic information

- The genetic information of an individual is personal data
  - **Sensitive** personal data (as it relates to health)
    - Explicitly defining in GDPR
  - Even if *anonymized* / *pseudonymized*
  - In principle, **no** difference between WGS, Exome, Transcriptome or GWAS data

- Theoretically possible to identify the individual person from which the sequence was derived from the sequence itself
  - The more associated metadata there is, the easier this gets
  - Gymrek et al. “Identifying Personal Genomes by Surname Inference”. Science 339, 321 (2013); DOI:10.1126/science.1229566

- *“The controller is liable to implement technical and organizational measures to protect the personal data. The measures shall attain a suitable level of security.”*

Bianca & Mosler

- **Bianca**
  - Swedish Research Council funded - *SNIC Sens* project
  - Implemented by SNIC/UPPMAX
  - 3200 cores / 1 PB
  - Opened april 2017

https://uppmax.uu.se/resources/systems/the-bianca-cluster/

- **Mosler**
  - e-Infrastructure for working with sensitive data for academic research
    - Developed & operated by NBIS
  - Inspired by Norwegian solution (TSD)
  - Designed to look like UPPMAX clusters
    - UPPMAX modules
    - UPPMAX can assist with installing custom tools
  - Implementation project completed Nov 2015
  - “Pilot-size system”
  - 24 nodes, 270 TB

- Provide users with a compute environment for sensitive data, with an *appropriate level of security*

https://nbis.se/infrastructure/mosler.html

Mosler

- High-performance computing in a virtualized environment (OpenStack)
  - Each project environment is **isolated** from all other projects
    - Separated private networks and file systems
    - No internet access
    - No root access
- Only accessible over remote Linux desktop (ThinLinc) via a web dashboard
- **2-factor authentication for login**
- **Restricted data transfer in/out**
  - Via a file gateway
  - Project members can transfer IN / only PI allowed to transfer out
  - Not possible to copy/paste out

TRYGGVE2: NORDIC COLLABORATION ON E-INFRASTRUCTURE FOR SENSITIVE DATA

https://wiki.neic.no/tryggve

TRYGGVE2 VISION
AND KEY TARGETS

Tryggve2 develops and facilitates access to secure e-infrastructure for conducting large-scale cross-border biomedical research studies utilizing sensitive data

Key targets

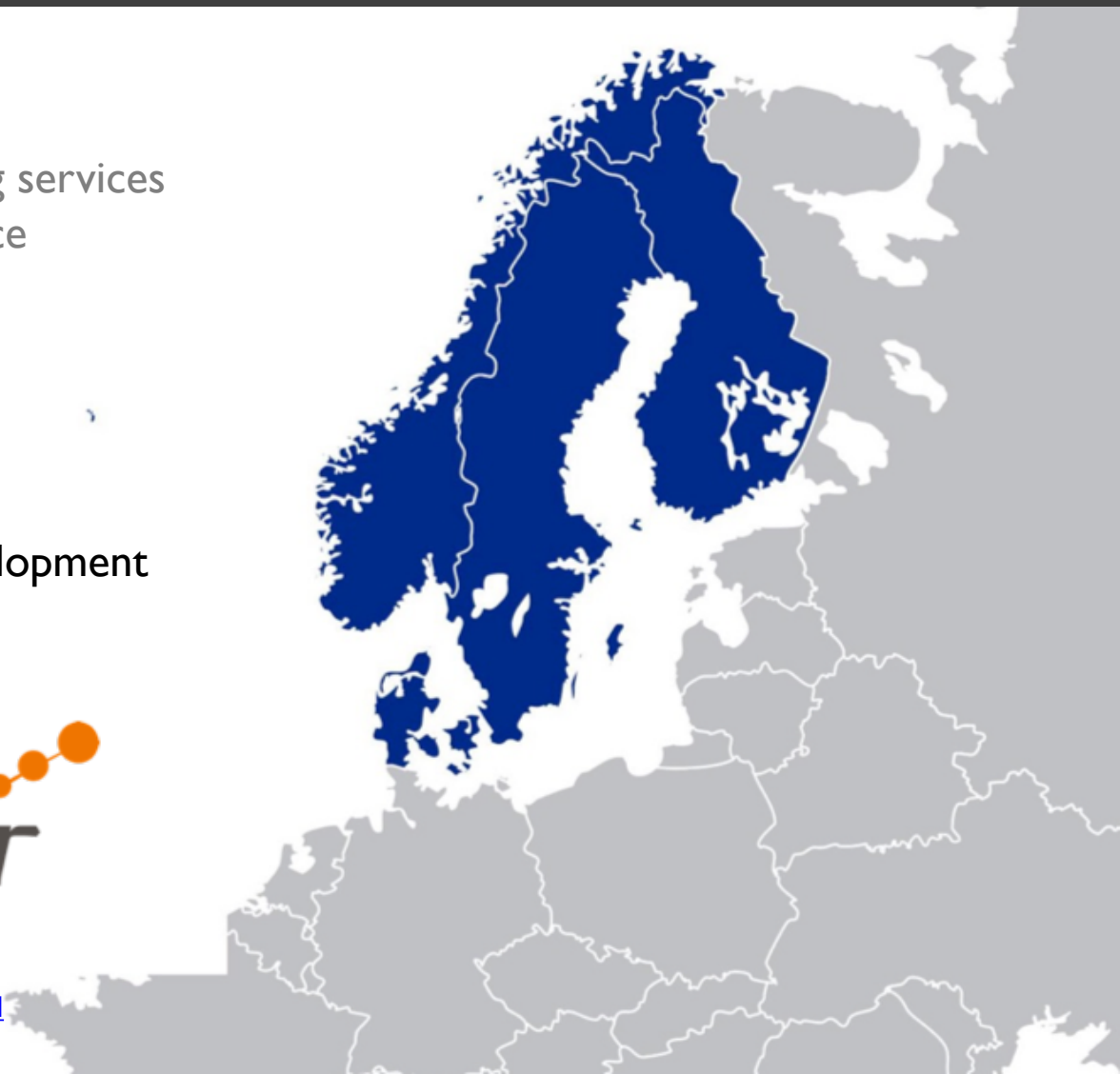
- Develop state-of-the-art infrastructure
- Support wide range of use cases
- Improve cross-border access and usability of services
- Leverage Nordic secure resources and expertise

TRYGGVE2 MAJOR DELIVERABLES

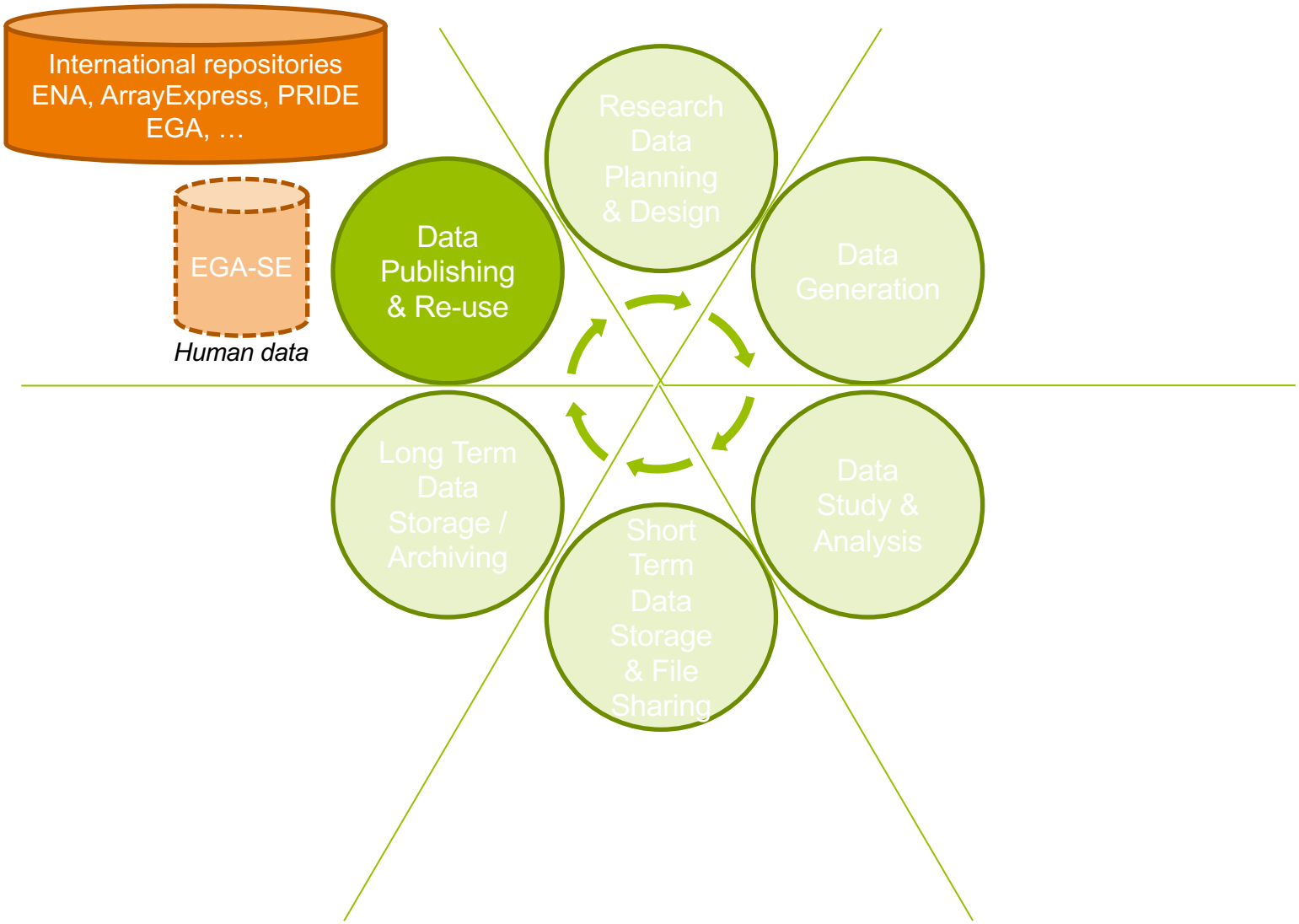
COLLABORATION FOR SENSITIVE BIOMEDICAL DATA

1. Sensitive data archiving
2. Production quality processing services
3. Homogenized user experience
    1. User mobility
    2. Workflow mobility
    3. Data mobility
4. Nordic use cases
        1. Research
        2. Infrastructure development
5. ELIXIR AAI
6. IT Security
7. ELSI Topics

https://wiki.neic.no/wiki/Tryggve_Getting_Started



Data Publishing & Re-use



Data Publishing & Re-use

- *Research Data Publishing is a cornerstone of Open Access*

- Long-term storage
  - Data should not disappear
- Persistent identifiers
  - Possibility to refer to a dataset over long periods of time
  - Unique
  - e.g. DOIs (Digital Object Identifiers)

- Discoverability
  - Expose dataset metadata through search functionalities

- *Strive towards uploading data to its final destination already at the beginning of a project*

Long tradition of data publication

- DNA sequence databases: *Genbank* and *EMBL db* 1982
- Protein structures: *PDB* 1969

Proc. Natl. Acad. Sci. USA
Vol. 86, p. 408, January 1989
Data Submission

1989

Submission of data to GenBank

CHRISTIAN BURKS AND LAURIE J. TOMLINSON

Theoretical Biology and Biophysics Group T-10, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545

In response to both the ever-increasing rate of determining nucleotide sequences (1) and the growing trend among journals to allow articles to appear that describe the results of determining a sequence without explicitly presenting the sequence (1), GenBank* (2–5) and a number of the journals that publish nucleotide sequence data are working together to promote the direct, timely submission of nucleotide sequence data to GenBank. The policy being established by the PROCEEDINGS is described in the editorial on p. 407; here, we will provide a brief summary, in the context of this policy, of

Electronic file transfer. Files can be
network to the network GenBank submi
above. This address—in most cases with
can be reached from various networks,
ARPANET, USENET, JANET, JUNET, etc.
work or system expert how to send electro
us for help. *Floppy disks.* We can read M
or 5¼-in diskettes written on MS-DOS s
that the submitted data be written as flat t
in a format specific to a given word

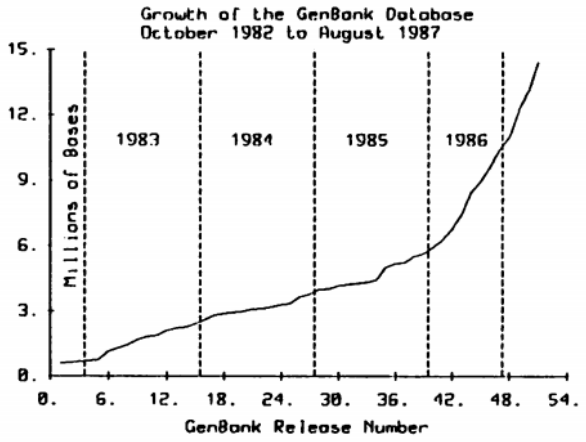


Figure 1.

"The author will provide the accession number to the PROCEEDINGS [PNAS] office to be **included in a footnote to the published paper**."

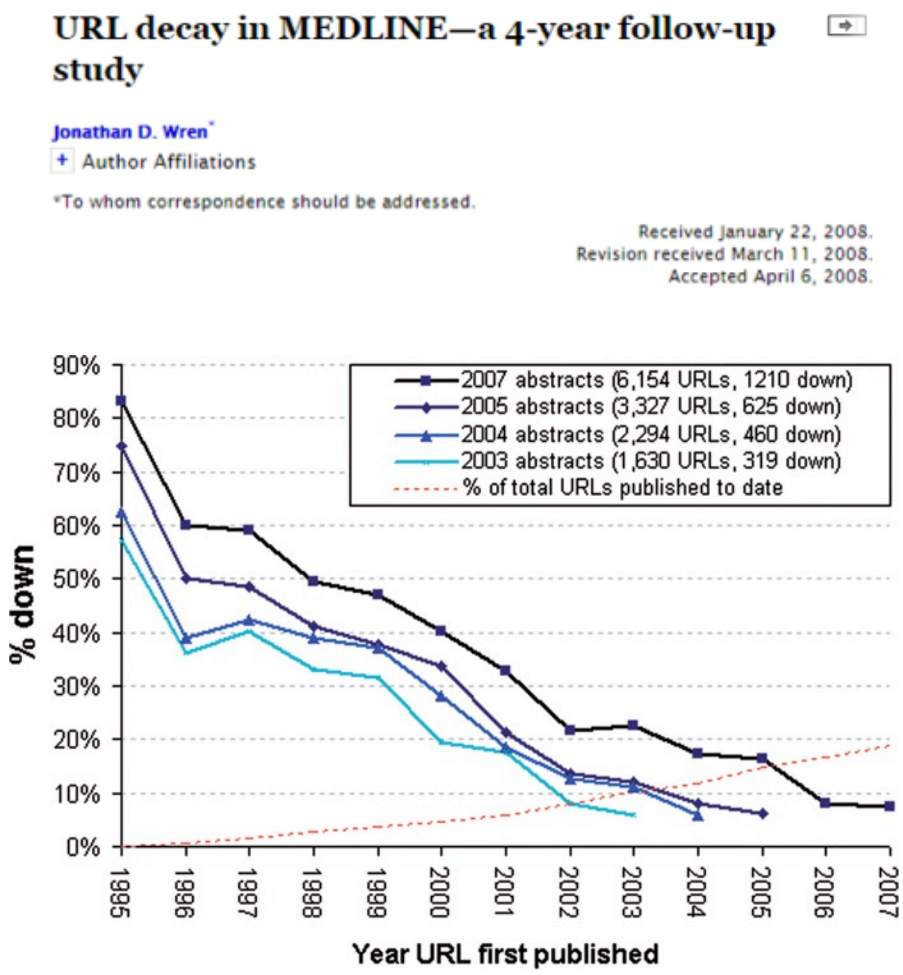
Bilofsky & Burks (1988)
Nucleic Acids Research v16 n5

Human genome project

Bermuda Principles for sharing DNA sequence data

- Automatic release of sequence assemblies larger than 1 kb (preferably within 24 hours).
- Immediate publication of finished annotated sequences.
- Aim to make the entire sequence freely available in the public domain

Data persistency issues



Jonathan D. Wren Bioinformatics 2008;24:1381-1385

- Link rot – more 404 errors generated over time
- Reference rot* – link rot plus content drift i.e. webpages evolving and no longer reflecting original content cited

* Term coined by Hiberlink http://hiberlink.org

FAIR

- To be useful for others data should be
  - ***FAIR*** - Findable, Accessible, Interoperable, and Reusable
    … for both Machines and Humans

Wilkinson, Mark et al. *"The FAIR Guiding Principles for scientific data management and stewardship"*. Scientific Data 3, Article number: 160018 (2016)
http://dx.doi.org/10.1038/sdata.2016.18

www.nature.com/scientificdata

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.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and

DOI: 10.1038/sdata.2016.18

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

'We support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR)'

European Commission

- European Open Science Cloud – EOSC
  - *Enable trusted access to services, systems and the re-use of shared scientific data across disciplinary, social and geographical borders.*
  - FAIR principles are a cornerstone of EOSC

Ref. Ares(2017)3468418 - 11/07/201

EUROPEAN COMMISSION
DIRECTORATE-GENERAL FOR RESEARCH & INNOVATION

The Director-General

Brussels, 10 July 2017

EOSC Declaration

RECOGNISING the challenges of data driven research in pursuing excellent science;

GRANTING that the vision of European Open Science is that of a research data commons, widely inclusive of all disciplines and Member States, sustainable in the long-term,

CONFIRMING that the implementation of the EOSC is a process, not a project, by its nature iterative and based on constant learning and mutual alignment;

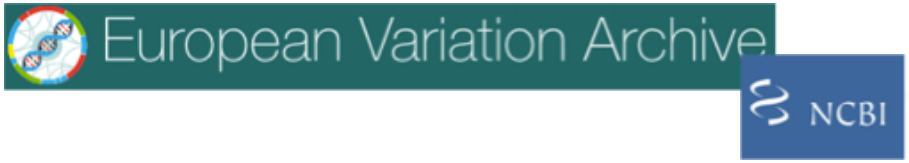
UPHOLDING that the EOSC Summit marked the beginning and not the end of this process, one based on continuous engagement with scientific stakeholders, the European Commission,

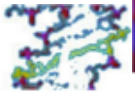
PROPOSES that all EOSC stakeholders consider sharing the following intents and will actively support their implementation in the respective capacities:

Data culture and FAIR data

- [**Data culture**] European science must be grounded in a common culture of data stewardship, so that research data is recognised as a significant output of research and is appropriately curated throughout and after the period conducting the research. Only a considerable cultural change will enable long-term reuse for science and for innovation of data created by research activities: no disciplines, institutions or countries must be left behind.
- [**Open access by-default**] All researchers in Europe must enjoy access to an open-by-default, efficient and cross-disciplinary research data environment supported by FAIR data principles. Open access must be the default setting for all results of publicly funded research in Europe, allowing for proportionate limitations only in duly justified cases of personal data protection, confidentiality, IPR concerns, national security or similar (e.g. 'as open as possible and as closed as necessary').
- [**Skills**] The necessary skills and education in research data management, data stewardship and data science should be provided throughout the EU as part of higher education, the training system and on-the-job best practice in the industry. University associations, research organisations, research libraries and other educational brokers play an important role but they need substantial support from the European Commission and the Member States.
- [**Data stewardship**] Researchers need the support of adequately trained data stewards. The European Commission and Member States should invest in the education of data stewards via career programmes delivered by universities, research institutions and other trans-European agents.
- [**Rewards and incentives**] Rewarding research data sharing is essential. Researchers who make research data open and FAIR for reuse and/or reuse and reproduce data should be rewarded, both

International public repositories





- Best way to make data FAIR
- Domain-specific metadata standards

ELIXIR Deposition Database list

Deposition Database	Data type	International collaboration framework [1]
ArrayExpress	Functional genomics data. Stores data from high-throughput functional genomics experiments.	
BioModels	Computational models of biological processes.	
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.	European Bioinformatics Institute and the Centre for Genomic Regulation
ENA	Nucleotide sequence information, covering raw sequencing data, contextual data, sequence assembly information and functional and taxonomic annotation.	International Nucleotide Sequence Database Collaboration
IntAct	IntAct provides a freely available, open source database system and analysis tools for molecular interaction data.	The International Molecular Exchange Consortium
MetaboLights	Metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.	

Deposition Database	Data type	International collaboration framework [1]
PDBe	Biological macromolecular structures.	wwPDB
PRIDE	Mass spectrometry-based proteomics data, including peptide and protein expression information (identifications and quantification values) and the supporting mass spectra evidence.	The ProteomeXchange Consortium
Pending incorporation into a Node Service Delivery Plan (see How countries join):		
BioSamples	BioSamples stores and supplies descriptions and metadata about biological samples used in research and development by academia and industry.	NCBI BioSamples database
BioStudies	Descriptions of biological studies, links to data from these studies in other databases, as well as data that do not fit in the structured archives.	
EVA	The European Variation Archive covers genetic variation data from all species.	dbSNP and dbVAR
EMDB	The Electron Microscopy Data Bank is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures.	

https://www.elixir-europe.org/platforms/data/elixir-deposition-databases

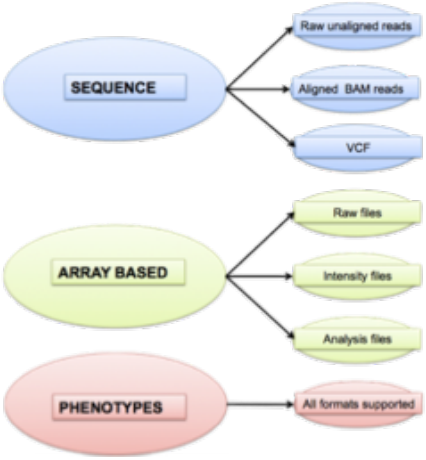
Surprisingly few submit to international repositories

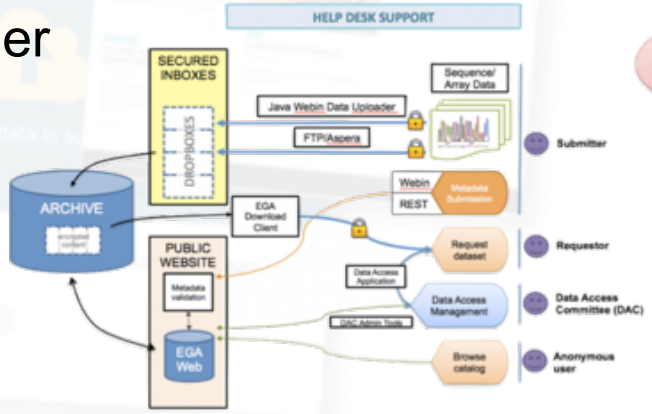
- NIH funded research
  - Only 12% of articles from NIH-funded research mention data deposited in international repositories
  - Estimated 200000+ “invisible” data sets / year

Read et al. “Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study” (2015)
PLoS ONE 10(7): e0132735. doi: 10.1371/journal.pone.0132735

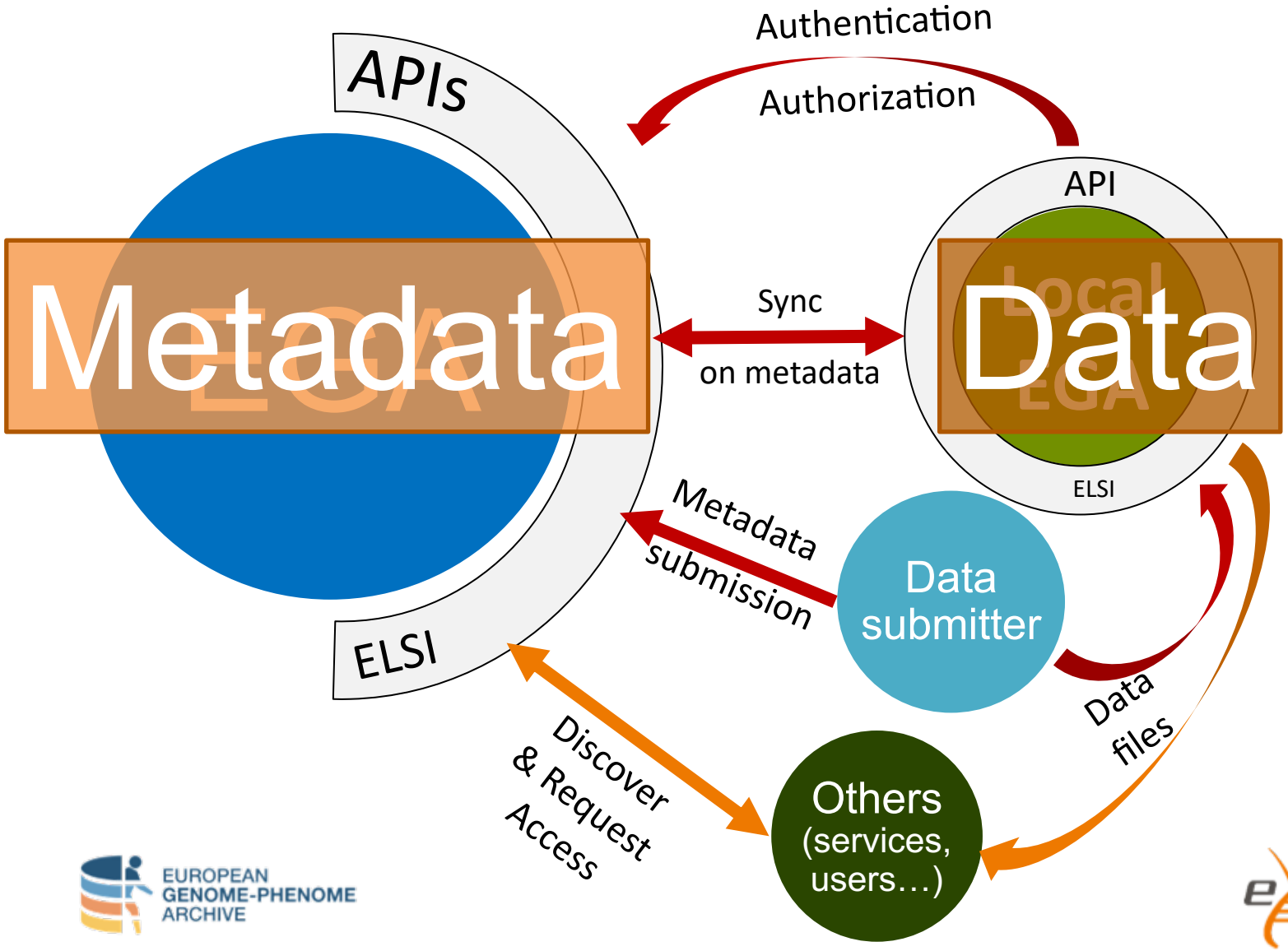
What about sensitive data?

- **EGA** – European Genome-phenome Archive
  - Repository that promotes the distribution and sharing of **genetic and phenotypic data** consented for specific approved uses but **not fully open, public distribution**.
  - All types of sequence and genotype experiments, including case-control, population, and family studies.
- Data Access Agreement
  - Defined by the data owner
- Data Access Committee – DAC
  - Decided by the data owner



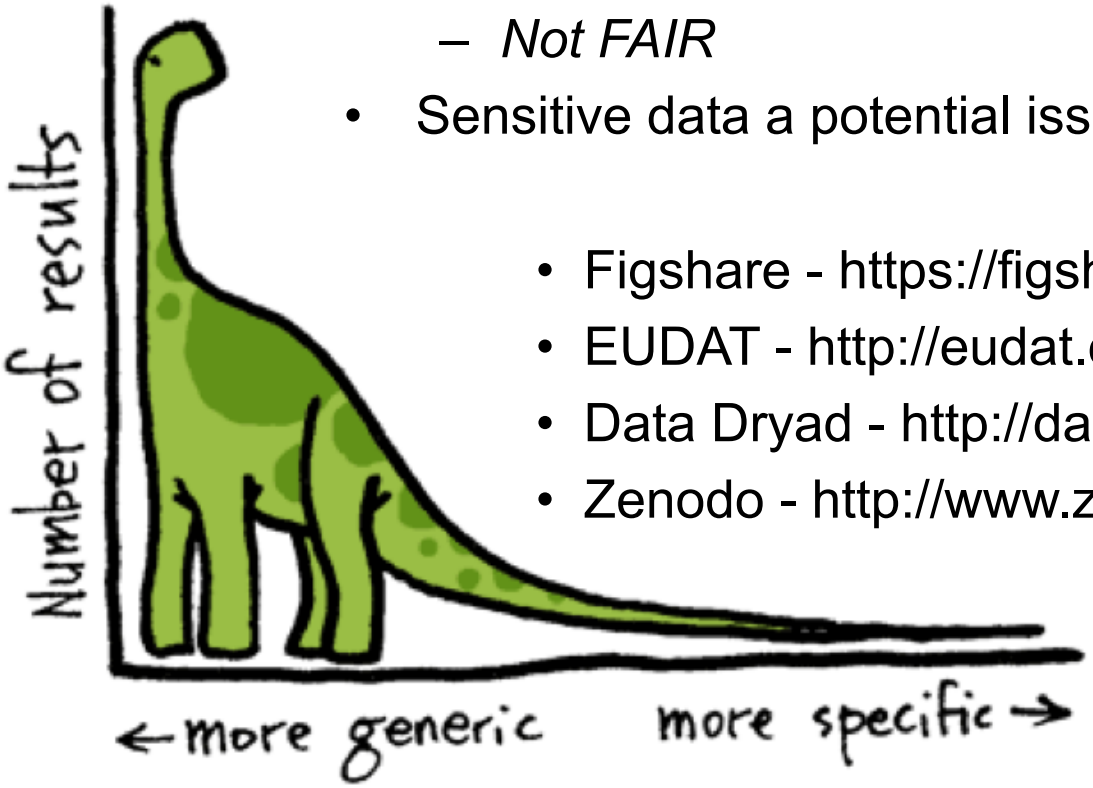


Local EGA


Authentication
Authorization
APIs
API
Metadata
Sync
on metadata
Data
Local
EGA
ELSI
Metadata
submission
Data
submitter
ELSI
Data
files
Discover
& Request
Access
Others
(services,
users…)

“Long-tail data” repositories

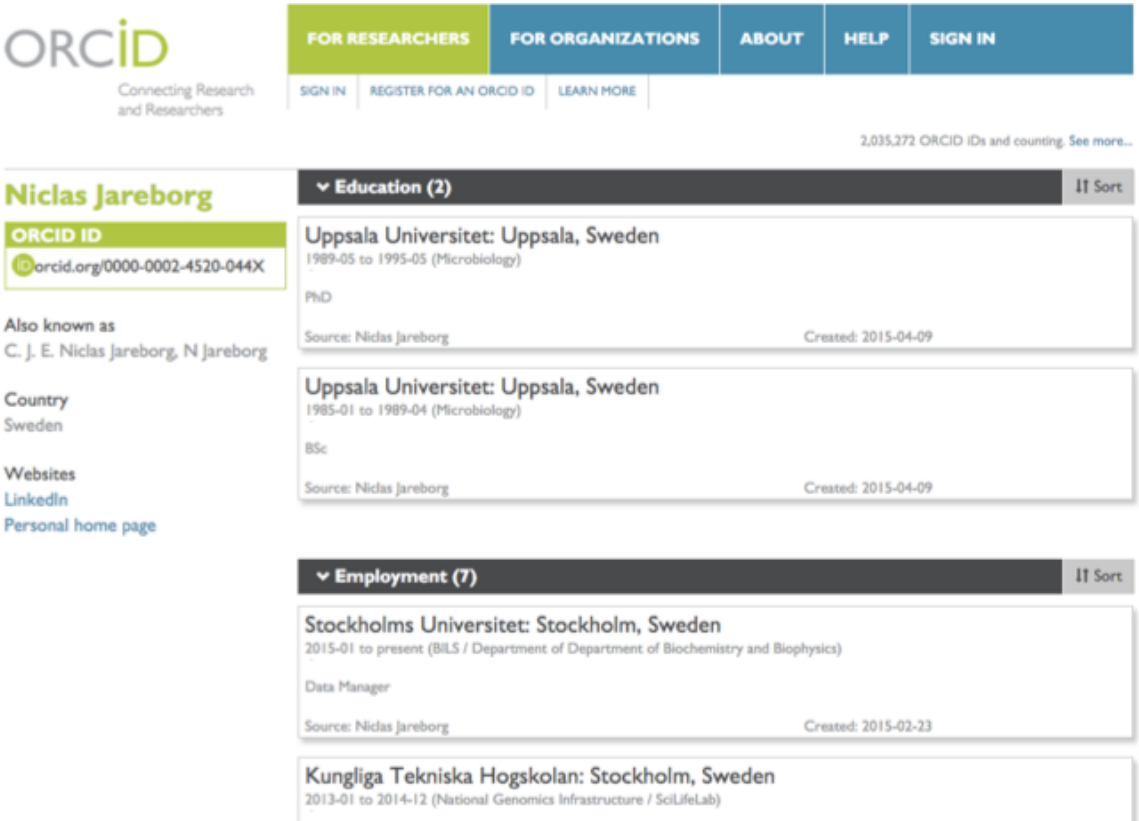
- Research data that doesn’t fit in structured data repositories
- Data publication – persistent identifiers
- Metadata submission – not tailored to Life Science
  - *Affects discoverability*
  - *Not FAIR*
- Sensitive data a potential issue



- Figshare - https://figshare.com/
- EUDAT - http://eudat.eu/
- Data Dryad - http://datadryad.org/
- Zenodo - http://www.zenodo.org/

Persistent identifier for yourself

- ORCID is an open, non-profit, community-driven effort to create and maintain a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers.
- http://orcid.org
- Persistent identifier for you as a researcher



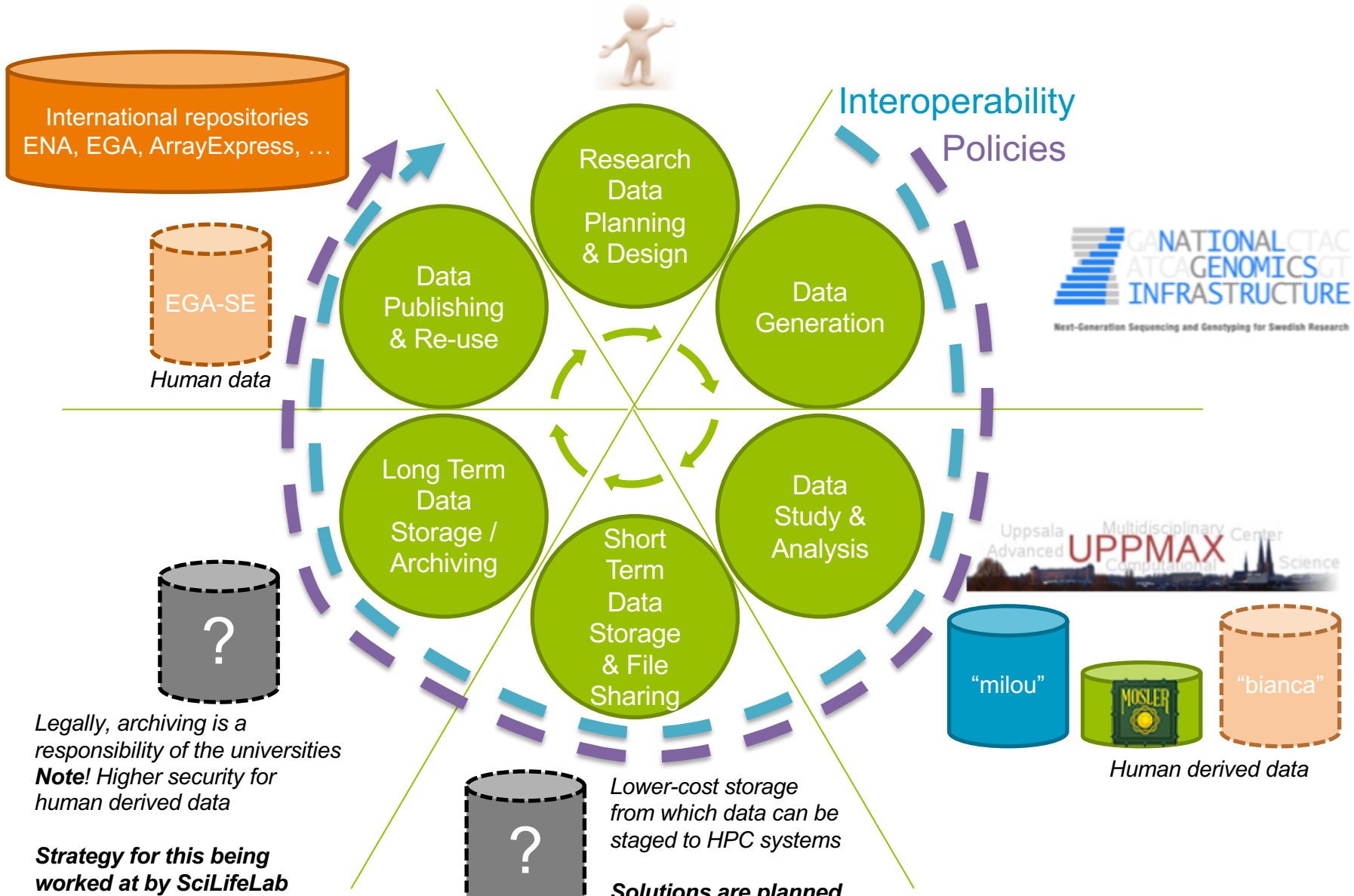
NBIS Data Management support

- Project planning
  - Metadata
  - File formats
  - Licensing
  - *Data Management Plans*

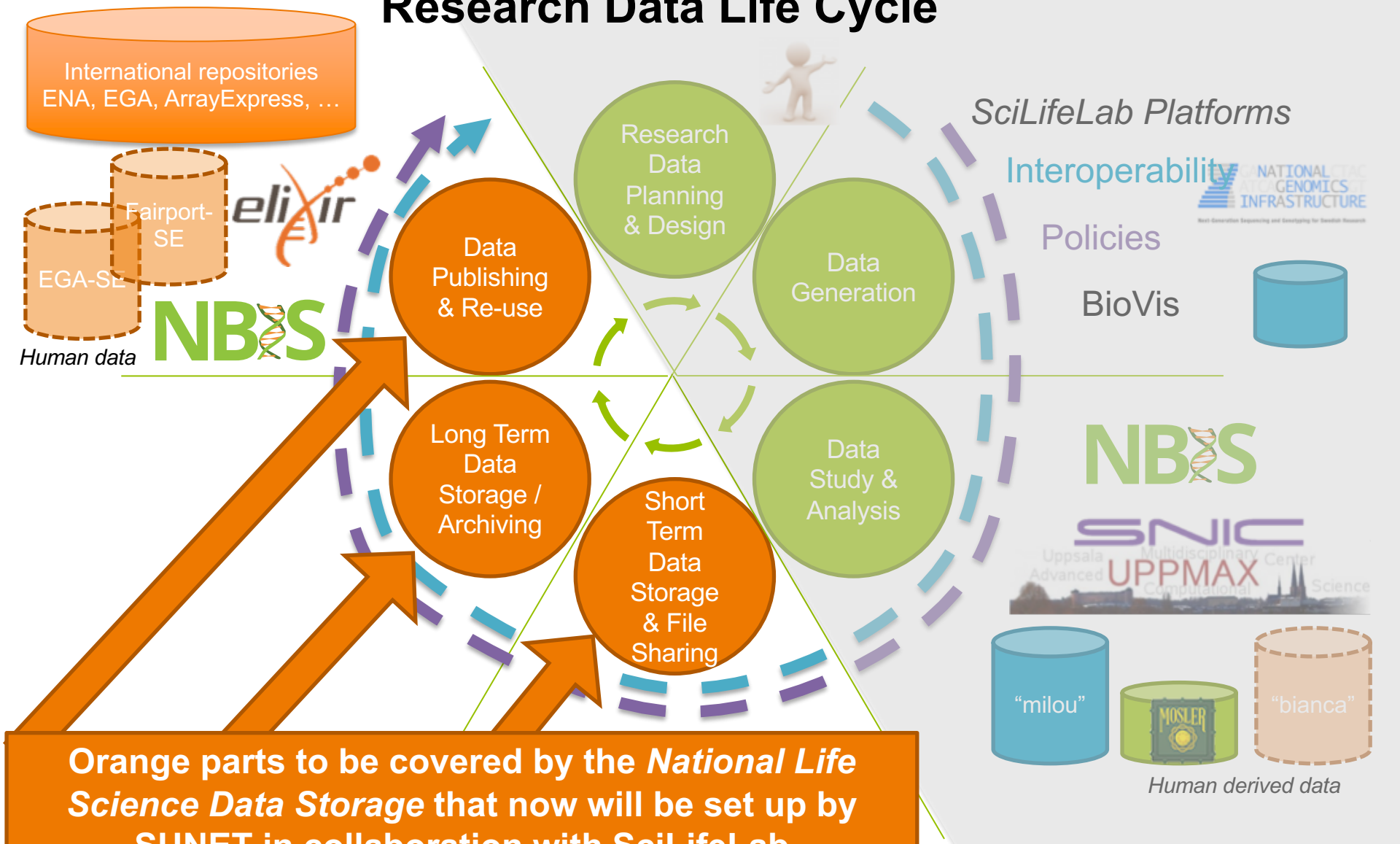
- Data analysis

- Data publication and submission
  - Automate submissions to public repositories
  - Metadata
  - Licensing

Gaps in the NGS Data Life Cycle



Research Data Life Cycle



Orange parts to be covered by the *National Life Science Data Storage* that now will be set up by SUNET in collaboration with SciLifeLab

Source Acknowledgements

- Research Data Management, EUDAT - http://hdl.handle.net/11304/79db27e2-c12a-11e5-9bb4-2b0aad496318
- Barend Mons – FAIR Data
- Antti Pursula – Tryggve https://wiki.neic.no/wiki/Tryggve
- Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. doi:10.1371/journal.pcbi.1000424
- Samuel Lampa - http://bionics.it/posts/organizing-compbio-projects
- Reproducible Science Curriculum – https://github.com/Reproducible-Science-Curriculum/rr-init
- Leif Wigge - https://bitbucket.org/scilifelab-lts/reproducible_research_example/src