

# Introduction to Research Integrity and Good Scientific Practices

## Module 2: Data management

**Context:** This presentation was created for the “Science in action” course of the Pompeu Fabra University (UPF) PhD program in Biomedicine, which has been running since 1998 in the context of the PRBB (Barcelona Biomedical Research Park). The authors of this latest edition are Maruxa Martinez-Campos and Eva Casamitjana-Martinez. Ero Jimenez Tejero was involved in previous editions, and Jordi Camí and Elinor Thompson were the authors of the original edition.

**Last edition:** October 2024

**License:** [CC BY](#)



**BEFORE STARTING THIS MODULE, please have a look at this short video and reflect on this tale of bad data managing and sharing**



Data Sharing and Management Snafu in 3 Short Acts



# MODULE 2

## Introduction to data management



# At the end of this module you should:

- Understand the importance of planning and what is a Data Management Plan (DMP)
- Beware of the importance of a good record keeping for Reproducibility
- Beware of biases and be skeptical of your own results
- Know the basics of Data Sharing and be familiar with the FAIR principles



# Outline

- Planning
- Record keeping
- Best practices and data sharing
- Data modifications



# THE IMPORTANCE OF PLANNING



Data management includes:

- collecting, storing, protecting, sharing, analysing, interpreting and presenting the data

All of this needs to be properly **planned** in advance. Many funding bodies require a **Data Management Plan** (DMP) explaining how they are going to deal with the data resulting from their project, both during and after the project ends.



Even if not required by the funding agency, doing a Data Management Plan is always a very good practice!!! And it is **mandatory for Clinical Trials**.



A **Data Management Plan** is a living document which should be made at the onset of each project, and be updated during the project's life cycle.

Some of the issues included in a DMP are:

- Data generation (how much data will be created, what types or formats, where will it be stored and backed up, how much will it cost...)
- Roles and Responsibilities (who will create the data, etc)
- Software and Services required
- Naming and describing your data
- Data Sharing with Collaborators
- Storage - short & long term
- Dissemination, Restrictions and Permissions



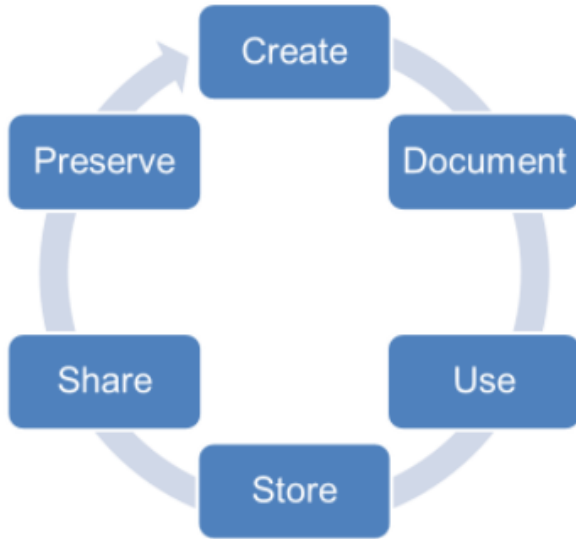


Diagram from the Digital Curation Centre, [www.dcc.ac.uk](http://www.dcc.ac.uk)

There are several online tools that can help you create your own Data Management Plan:

- [Research Data Management Plan / Pla de Gestió de Dades de Recerca \(csuc.cat\)](http://www.csuc.cat) (Spain)
- <https://dmponline.dcc.ac.uk> (UK) helps you write your DMP to follow rules of specific funders or institutions
- <https://dmptool.org> (US)

Check these and other resources on the [UPF website](#) and your own center policies and templates.



# What does the PRBB code say about it?

## Code of Good Scientific Practice



### **2.1. Written projects subject to scrutiny by outside parties**

All research projects that directly involve humans, experimental animals, or human embryonic material, must be formulated in a written research plan prior to their initiation. The text of the written plan must have been independently assessed by an ethics committee on clinical research and/or animal experimentation. This text generally coincides with the written proposal necessary to obtain approval and funding.

### **2.3. Extension or modification of the research plan**

In research involving humans, or experimental animals, or in some cases where the primary objectives of the research are extended or altered, or an unexpected or additional research question arises, a complementary written plan may be prepared prior to initiating research in that direction. If the implications of the new research question so require, the revised research plan must follow established procedures for external authorisation and supervision by the relevant committees.



# What does the PRBB code say about it?

## Code of Good Scientific Practice



Research institutions, organisations and researchers must ensure appropriate stewardship and curation of all data and research materials including the following points:

### **3.1. Data collection and storage**

All research plans must include a system for collection of data, registries, and biological or chemical material arising from the research, along with a data management plan (DMP) relating to their custody and storage.



# RECORD KEEPING





# Record keeping

Why it matters?

What should you record?

How should you record it?

Whose responsibility is it?

Who owns it?

# Why it matters?

Proper **record keeping** is crucial for **reproducibility**

As a reminder to yourself – difficult to remember what you do over several years...!

As a guide for others to replicate or verify your results - your project doesn't end when you leave the lab!

For retrospective audits by granting agencies in accountability and evaluation procedures

To enable review of primary data during drug and medical device approval processes

The original primary data with recorded dates may be required to resolve:

- Questions about the truth of published research results
- Legal challenges concerning ownership and/or originality in commercial, patenting or intellectual property cases



# What should you record?

- All processes from collection to analysis (including standard procedures and any changes made to them), parameters, manipulation of data... as detailed as possible
- Also your ideas, reasoning,... maybe meeting notes or email exchanges?
- Data filenames, formats and locations
- Evidence of approvals by ethics committees etc.

**Write what  
you do  
and  
do what you  
write**

## **RECORD EVERYTHING!**

Keep all your data - what, why, who, how, when, where, what happened, your interpretations, what's next... Except: **confidential information**



## What is considered “data”?

All measurements or observations used to make descriptions or inferences. Scientific data include:

An electrophoresis gel or a DNA sequence....

Completed questionnaires, videotapes, and photographs.

Microscope slides, cell lines, climate patterns, soil samples, astronomical measurements, spectrographic analyses, custom software or hardware, specialized methods...



[Image by Jacopo Werther](#), from Wikipedia

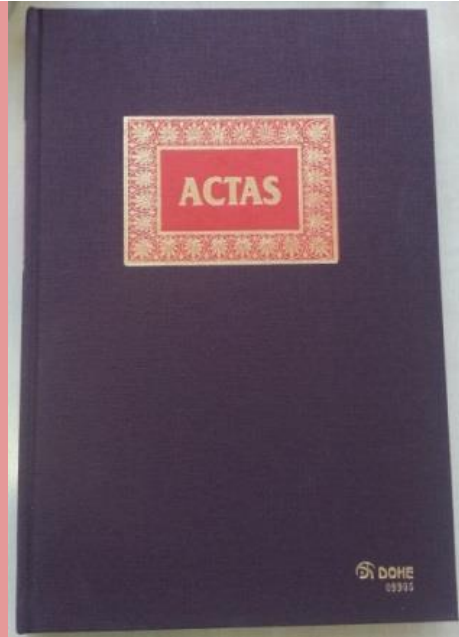


# How should you record it?

- You should use either a **paper** or an **electronic notebook** (ELN) in addition to diverse forms of material storage.
- Whichever form, it should be dated, and permanent.
- Witnessing (signing of the record by someone else) is:
  - Required in industrial research laboratories.
  - Less common in academic research, but may be necessary for some clinical trials and industry funded studies.
  - Indispensable if the work may lead to a patentable discovery or invention.




The PRBB provides registered notebooks free of charge to all personnel belonging to its constituent centres



Some centres (like the CRG) have in place an Electronic Lab Notebook (ELN) system for all its researchers



A photograph of a desk setup. On the left is a white ceramic mug. In the background is a silver laptop. In the foreground is a spiral-bound notebook with a black pencil lying across it. A green paperclip and a green pushpin are also on the notebook. A white ruler is visible on the right side of the desk.

However you do it (notebook, ELNs, ...)  
**Write what you do  
and  
do what you write**

## What does the PRBB code say about it?

### Code of Good Scientific Practice



#### **3.2. Recording of data and alterations**

Without exception, all data arising from experiments or research observations must be recorded in an accurate way to ensure traceability of the work. That information must remain permanently recorded in databases, registered notebooks, or other appropriate format, in a condition that facilitates external review. The records must also include changes, errors and negative, unexpected, or conflicting results, as well as an indication of the person who performed the experiment or made the observation.



# Whose responsibility is it?

- Your own! Individual scientists are responsible for maintaining their own notebooks and research records.
- Heads of labs and research departments are responsible for making sure that their team members' notebooks and research records are in order.



Photo by [Olu Eletu](#) on [Unsplash](#)



# Who owns it?

- When public bodies award a grant to a university or research centre, **all data collected are owned by the institution winning the award.**
- Subsequently the university or centre allows the principal investigator on the grant to be the steward of data collection, recording, storage, retention, and disposal.
- Depending on who funds the research, the funding agency itself may own the data...

Neither the Principal Investigator (PI), nor the Head of Department, as individuals, legally own either the data or data books collected by him or herself, or by any students or other scientists on a research project.



# Data transfer or what happens if someone leaves the lab

Removal of copies of original data and data books may be permitted on a variety of grounds:

- Copying and removal of data must be approved by the PI.
- If the PI is the mover, copying and removal of data must be approved by the head of department or the centre/ university director.
- When transfer of data ownership does occur, it is between one institution and another. In this case, a **Data Transfer Agreement** should be signed by the institutions.
- In industry-funded or privately funded research, data may belong to the sponsor (to retain the right to the commercial use of data). Depending on the contract, the right to publish the data may or may not be extended to the investigator.



# What does the PRBB code say about it?

## Code of Good Scientific Practice



### 3.5. Ownership of data and samples

All primary documentation (registered data-collection notebooks, databases, etc.) and biological or chemical material obtained in the course of a research project is the property of the centre to which the person responsible for the research is affiliated. Institutions and organisations have therefore a role in facilitating the recording, storage, and safekeeping of that material, although the primary responsibility lies with the individual responsible for the project. Should a researcher change institutions, the individual responsible for the project may make available a copy of part or all of the records, and/or aliquots of available biological or chemical materials, provided such sharing is necessary. A Material Transfer Agreement must be signed for all human biological samples (blood, serum, DNA, tissues etc). When the change involves the person responsible for the research, the director of the centre will take responsibility for supervising this process.



# The perfect notebook should be...

legible, well organized, accurate and complete, enable repetition, compliant with requirements, accessible to authorized persons, stored properly, and appropriately backed up...

... **and secure**, ensuring:

- **Confidentiality**: the information is accessible only to authorised people
- **Integrity**: accuracy and completeness of the data
- **Availability**: authorised users have access to information systems when required
- **Security**: Prevent theft, accidental loss or damage



# What does the PRBB code say about... availability?

## Code of Good Scientific Practice



### **3.4. Custody and access to collected data**

All individuals who belong to the research group must be able to access information on the data obtained and their interpretation. The individual responsible for the research will have a single record accessible to third parties, of the locations of all samples and data-collection instruments (registered notebooks, databases, etc).



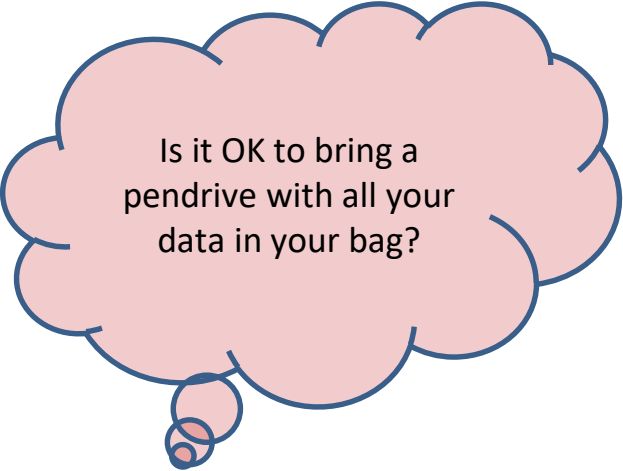


## How can you prevent damage to your data?

- Store lab books in a safe place.
- Back up computer files, keep backups secure and physically removed from the original data
- Samples should be appropriately saved so that they will not degrade over time.
- Take care to reduce risk of fire, flood and other catastrophes.



## Pause for thought



Is it OK to bring a  
pendrive with all your  
data in your bag?



What does the  
PRBB code say  
about...  
safety?

## Code of Good Scientific Practice



### 3.3. Storage of data

The necessary means and infrastructure must be provided by the institutions for correct storage and safekeeping of all documentation and biological or chemical material resulting from a research project. In the case of data recorded on electronic media, a specific plan will be included for the preparation and storage of backup copies.



## A special note on Personal Data

### Personal Data

Any information relating to an identified or identifiable natural person ('data subject')



Health Personal Data is the  
**MOST SENSITIVE** personal  
information

Personal Data should comply with specific Privacy and Data Protection regulations according to the [General Data Protection Regulation](#) (GDPR) in the EU. Check the policies in place in your center.





# How long should the data be kept for?

Depends on the institution and the kind of data.

- Generally a minimum of 5 years after publication of the results is recommended
- In the EU, data from Clinical trials should be kept for 25 years.
- Long term storage is space dependent, and should be secure, appropriate and geographically removed from the centre.



# What does the PRBB code say about it?

## Code of Good Scientific Practice



### 3.7. Length of storage of data and samples

All original primary information and biological and chemical material arising from a research project must be stored for a minimum of 5 years from the date of the first publication of the results, except in those cases in which the law allows shorter storage periods or requires longer periods to be applied. If the centre allows, the primary information and material may remain stored for longer periods, provided their final destination meets the approval of the person responsible for the research.

# Tips for a good record keeping

- Maintain all original records and have one (or several) backups – ideally at least one electronic
- Write down the information as soon as possible
- Include dates, names of people who did the experiment, etc.
- Make corrections with clear annotations and explanations (don't just erase!)
- Reference everything in one central record
- Other things to consider: organisation (chronological order or project-based?), language, etc.



# BEST PRACTICES AND DATA SHARING

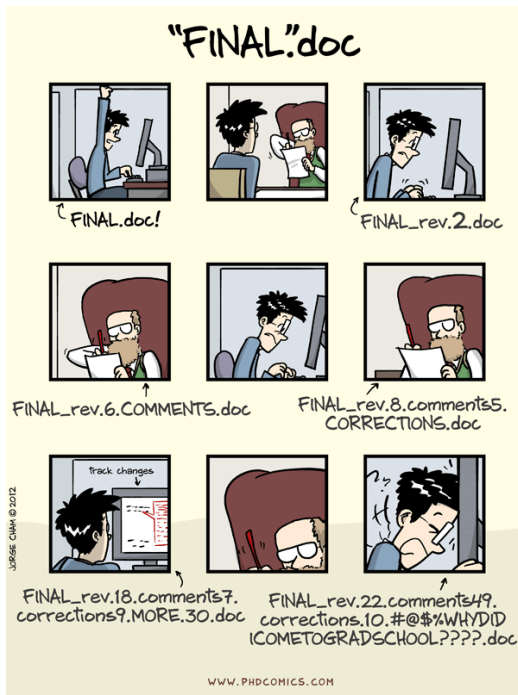


# Data management best practices

Data management will be very different for each research project/lab and it may be discipline-specific (eg [bioinformatics](#)). But some challenges are common to most researchers (**File naming, Version control, use of proprietary software...**) and it's worth keeping them in mind!

- Use logical and consistent file names, according to agreed conventions (in the lab, with collaborators, etc)
- Keep file names short and relevant (descriptive and without cryptic label!)
- Do not use special characters or spaces (except \_ and -)
- Never change the file extension

The [Open Science Framework](#), the [University of Cambridge](#) and the [University of South Hampton](#) give tips about [file naming](#) and organisation, creating a data management plan, etc. Check out also the Software Carpentry website for tutorials on [how to use version control](#) and others.



# Data sharing

In principle all data should be considered for sharing in the interests of:

- Free exchange of information.
- Spreading learning and avoiding waste through unnecessary repetition.

In addition, new regulations by EU grants (Horizon Europe) and other funders (NIH) make open access to research data

**compulsory!**

Open Data and Open materials badges from the Open Science Collaboration - *CC BY 4.0*

Ideally, the sharing process should have been described in the data management plan, or a specific **data sharing plan**.

Click [here](#) to see an example!





Whenever possible, make sure your data follows the [FAIR data principles](#) to ensure data are:

- **F**indable
- **A**ccessible
- **I**nteroperable (allows data exchange and reuse between researchers - i.e. use of ontologies, standards for formats, open software, etc.)
- **R**eusable



# Barriers for sharing data

## **“I don’t benefit...”**

- No recognition of sharing for CV/grants/assessments (Lack of credit)
- No obligation by funders

## **“It’s difficult...”**

- Lack of time and money (How important is it relative to doing the research?)
- Lack of infrastructure/repositories for data sharing or of data management knowledge and skills

## **“I shouldn’t anyway...”**

- Data Protection issues: data privacy/anonymity
- Commercial motives (patents,...)
- Others will benefit from my data and ‘scoop’ me; how will it affect my chances to publish?
- Others might misuse my data (to harm contributors, to support arguments,...)

## **“In reality I fear...”**

- My data is not of good enough quality / my results will not be validated

# Solutions for sharing data

- No recognition → **this is changing; new research assessment policies now consider Open Science**
- Lack of time and money/infrastructures/knowledge and skills → **proper planning from the beginning, so you get data in the right format will help. Also, ask your institution for training and infrastructures!**
- Data Protection issues and commercial motives (patents,...) → **share what you can; if you can't share all data, share only the metadata. If you can't share individual data, share pooled data. Share anonymised data, or shared with restrictions.**
- Others will benefit from my data and 'scoop' me / might misuse my data → **have a restricted access policy e.g. with an embargo, or in exchange of collaboration, or get people to explain what they will do with the data... Publish the data ASAP as a preprint.**
- My data is not of good enough quality / my results will not be validated → **make sure your data is good enough! And if results are not validated, better to find out ASAP thanks to other people's use of the data**

## What does the PRBB code say about it?

### Code of Good Scientific Practice



**3.6. Sharing of data and samples with outside parties** Researchers, research institutions and organisations must ensure access to data is as open as possible and as closed as necessary. Where appropriate, data and materials arising from a research project must be publicly available and in a condition to be shared with outside parties in line with the FAIR Principles (Findable, Accessible, Interoperable and Re-usable) for data management. Exceptions include cases where restrictions have been established on the basis of possible future commercial use.

(...)



# Sharing exceptions

dual

Data sharing restrictions do apply in 3 circumstances:

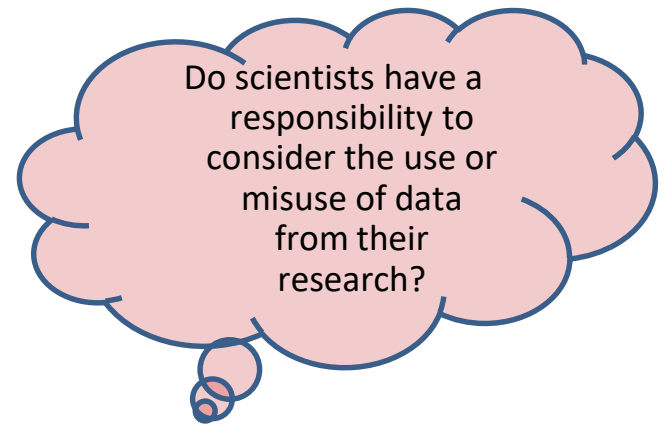
- Data prone to **dual use**: data that may be misused to pose a biologic threat to public health and/or national security\*
- Data that can lead to **patents, commercial interest**
- Data and information that can affect the **privacy of human subjects**

Balancing the obligation to protect and need to share:  
**Be open when you can and close when you must**



## Pause for thought

### Consider these cases...



- **De novo synthesis of poliovirus (2002)**

Cello J, Paul AV and Wimmer E. Chemical synthesis of poliovirus cDNA.... Science 2002; 297:1016-8.

- **De novo synthesis of 1918 influenza strain**

Tumpey TM et al. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. Science 2005; 310:77-80.

- **A mathematical model for a bioterrorist attack**

Wein LM and Yiu Y. Analyzing a bioterror attack on the food supply: the case of botulinum toxin in milk. PNAS 2005; 102:9984-89.



# What does the PRBB code say about it?

## Code of Good Scientific Practice



**3.6 Sharing of data and samples with outside parties** Provision of data or materials will require 1) that information be provided on the intended use by the person who has requested them; 2) that the research group is aware of the request; 3) that there is a material or data transfer agreement with the approval of the individual responsible for the research; 4) and that the person making the request is willing to pay all possible costs of production and shipping. Sharing may be restricted for reasons of availability, competition, or confidentiality. Material or data obtained from human subjects must be shared in such a way that the subjects cannot be identified; if identification of individual subjects is possible, those individuals must first consent.

### **4. Research projects funded by the healthcare industry or other commercial enterprises**

**4.1 Transparency** When knowledge and technology is exchanged or provided to private enterprises, public interests must always take priority, and any agreements must be transparent.



# What does the PRBB code say about it?

## Code of Good Scientific Practice



### 2.6. Collaborative research

When a planned research project involves the participation of several groups from the same or different centres, a formal agreement should be made where the terms of the collaboration are formalised in writing before initiating the definitive project. Also, all partners must take responsibility for the integrity of the research and its results.





# Gender perspective

When doing research, specially in biomedicine, it is essential to include a sex/gender dimension in the design and analysis, whenever relevant.

**WHAT?** This might mean to include both men and women in a study (and analyse the results by sex) or to take into account how different genders might be affected by the research, if applicable. It applies to humans, but also animal models or even cultured cells.

**WHY?** Fairness and quality of the research. There are differences in the way male and female bodies react to drugs, for example. That's why since 1994, the US has required all clinical trials funded by the National Institutes of Health (NIH) to include women.



# Gender perspective

Some funders have made the inclusion of this dimension compulsory, e.g. European Research Council (ERC) Horizon Europe programme, Swedish Research Council or Canadian Institutes of Health Research.

There are several guides to incorporate the gender perspective in research, for example this [Toolkit](#) from Community Hipàtia.

**INCORPORATING THE  
SEX AND GENDER  
PERSPECTIVE IN  
RESEARCH CONTENT:  
A TOOLKIT**

Community  
Hipàtia



*Sex and gender perspective in research content is different than gender equality in research teams - but the latter is just as important!*

# What does the PRBB code say about it?



## Code of Good Scientific Practice



**2.7. Gender and diversity perspective** Research projects must take into account and be sensitive to relevant differences among research participants, such as age, gender, sex, culture, religion, worldview, ethnicity, geographic location and social class, amongst others.



# DATA MODIFICATIONS



Once you have your data collected, you set to analyse it.

- Selecting which data you will use to draw conclusions, and which will be discarded if any
- Establishing significance and identifying potential weaknesses and limitations
- Choosing how to present it

Can the way you modify / transform your data affect its integrity?

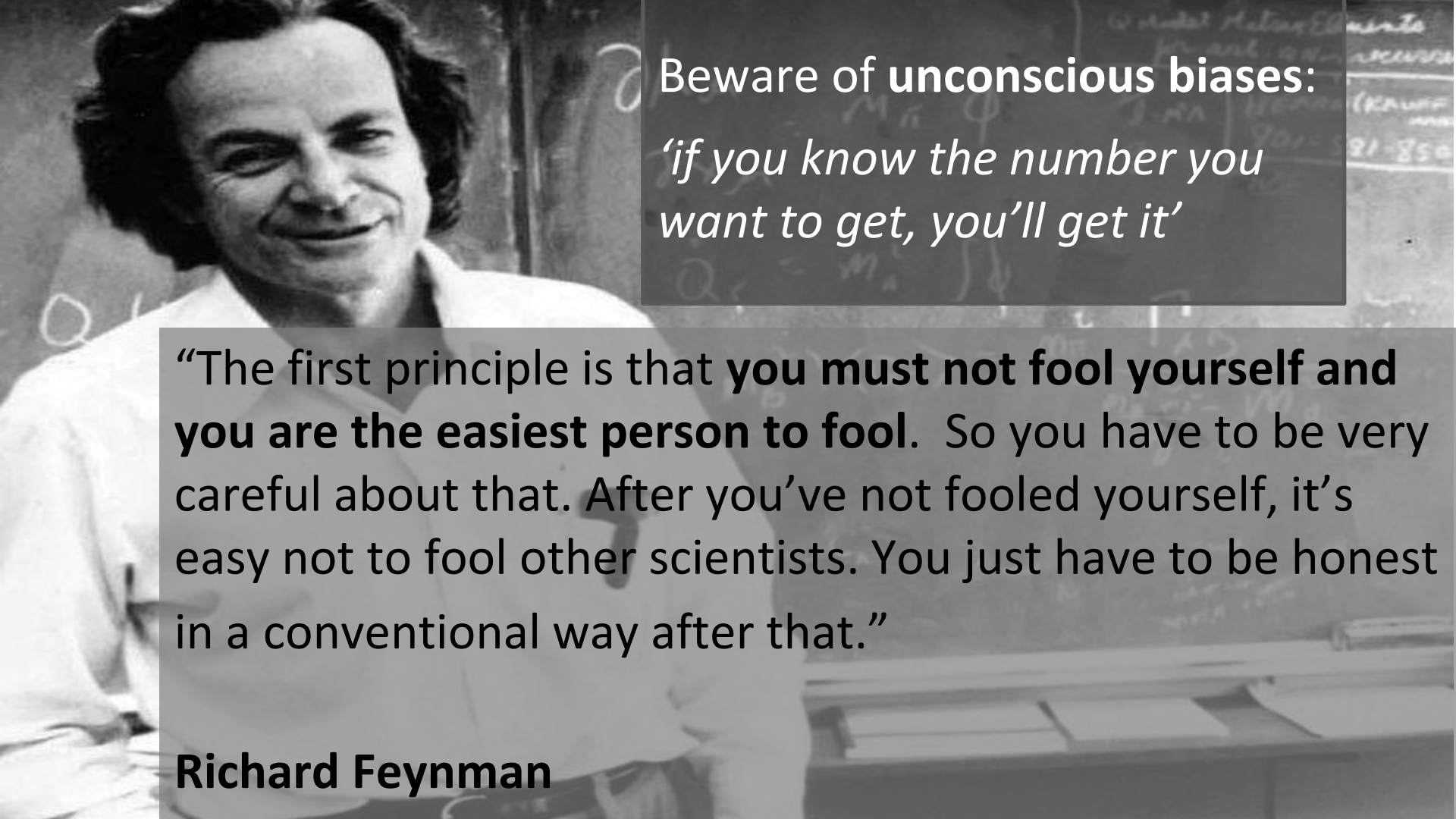
Can the way you look at your data (and the result you are expecting to get) affect its integrity?



# Cooking data...

“Scientific results can be distorted in several ways, which can often be very subtle and/or elude researchers' conscious control. Data, for example, can be “**cooked**” (a process which mathematician Charles Babbage in 1830 defined as “*an art of various forms, the object of which is to give to ordinary observations the appearance and character of those of the highest degree of accuracy*”); it can be “**mined**” to find a statistically significant relationship that is then presented as the original target of the study; it can be **selectively published** only when it supports one's expectations; it can **conceal conflicts of interest**, etc...”



A black and white photograph of Richard Feynman, a physicist, smiling and looking towards the camera. He is wearing a light-colored shirt. In the background, there is a chalkboard with some faint mathematical equations and diagrams. The image is overlaid with text boxes.

Beware of **unconscious biases**:  
*'if you know the number you  
want to get, you'll get it'*

“The first principle is that **you must not fool yourself and you are the easiest person to fool**. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists. You just have to be honest in a conventional way after that.”

**Richard Feynman**

Examples of types of biases\*:

- **hypothesis myopia/confirmation bias**: failing to look for evidence or consider alternative explanations to a data that fits your preconceived hypothesis
- **asymmetric attention to detail**: rigorous check of non-intuitive results, free pass to expected ones
- **p-hacking**: is the misuse of data analysis to find patterns in data that can be presented as statistically significant when in fact there is no real underlying effect

\*How scientists fool themselves – and how they can stop. Regina Nuzzo, Nature News, October 2015

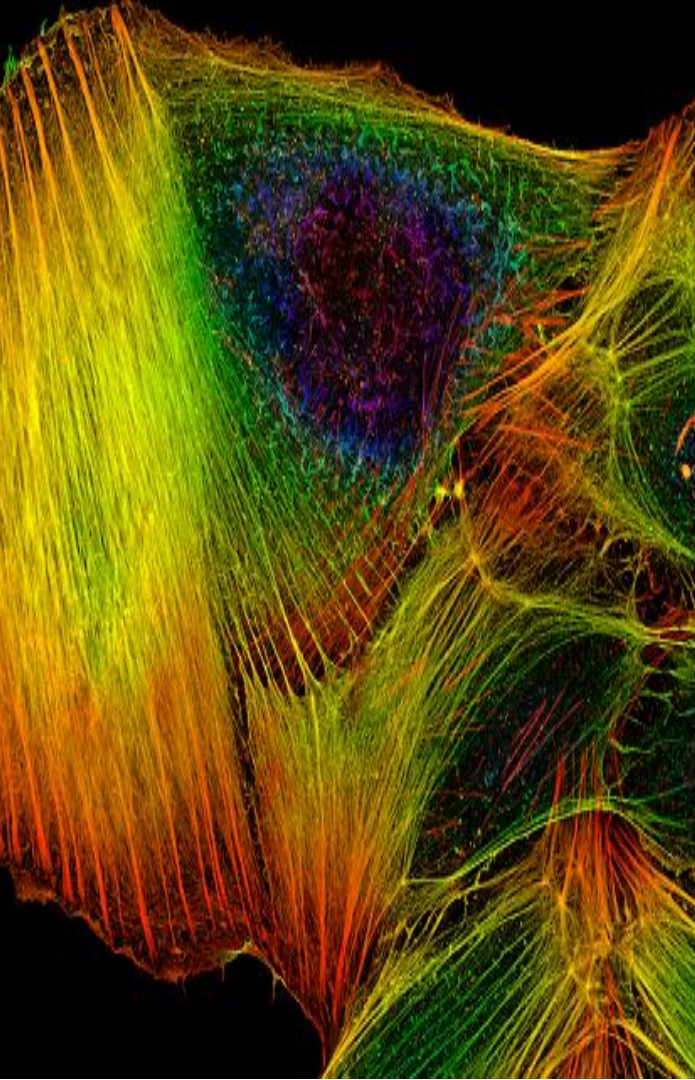
Many errors can occur  
→ **be skeptic of your  
own results!**





## Examples of ways to avoid bias:

- **Strong inference:** force oneself to explicitly consider competing hypothesis / alternative explanations and do the experiments to tell them apart
- **Data blind analysis:** computers shift the data so you are doing analysis with 'fake' data, but computer applies all your analysis to the real one - at the end you unblind and see the real results; adding errors or noise on purpose without researchers knowledge; introducing fake signals...
- **Planning!**
- **Transparency:** open science, open data, *registered reports*



# The special case of images

Some tips\* on how to treat digital images to preserve their integrity:

- Manipulation of digital images should always be done on a copy of the unprocessed image data file
- Simple adjustments to the entire image are usually acceptable.
- Digital images that will be compared to one another should be acquired under identical conditions, and any post-acquisition image processing should also be identical
- Manipulations that are specific to one area of an image and are not performed on other areas are questionable
- Use of software filters to improve image quality is usually not recommended for biological images (they may create artifacts)
- Copying objects into a digital image, from other parts of the same image or from a different image, is very questionable
- Magnification and resolution are important
- Be careful when changing the size (in pixels) of a digital image
- Avoid the use of lossy compression (when compressing to jpeg, some features might be changed...)

\* From <http://microscopy.arizona.edu/learn/digital-image-ethics>



The way you present your data can also be misleading...

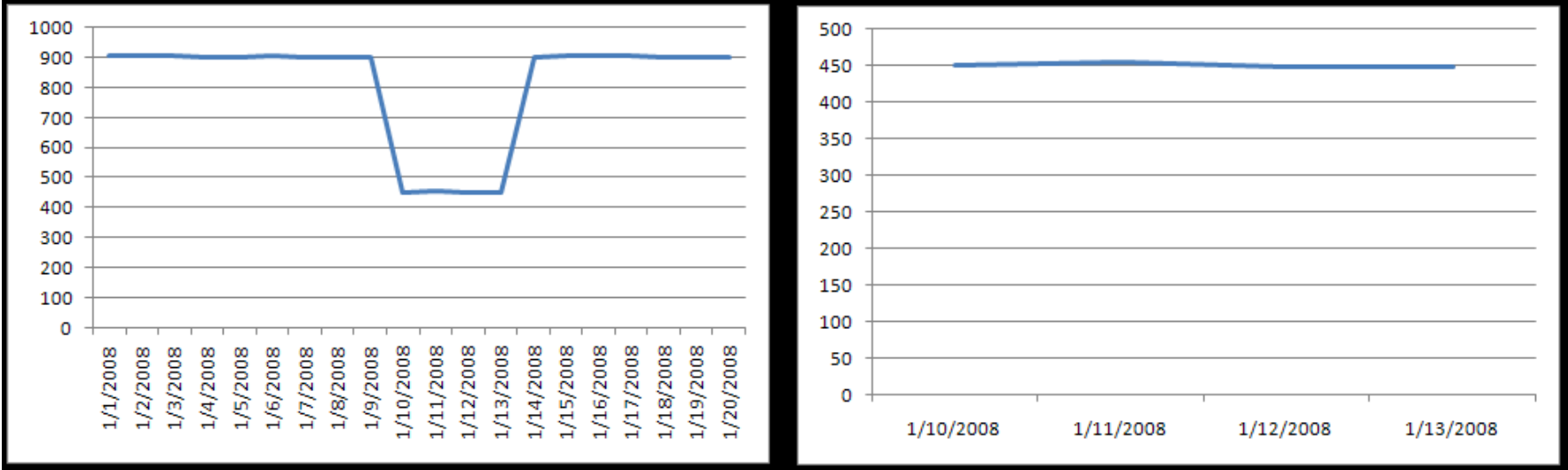


Image by [Tempshill](#) CC BY-SA 3.0



# Pause for thought

## On Raw data



ORI Case Study 5: To Proceed or Not to Proceed Without Raw Data?

## On Cherry Picking



ORI Case Study 3: Data Cherry Picking

## On Reproducibility



ORI Case Study 2: Reproducibility or Luck? Struggle to Get Results



# In summary....

- Plan what you will do
- Record what you do
- Beware of biases and be skeptical of your own results!



END OF THE MODULE

