



Universität  
Zürich<sup>UZH</sup>

Hauptbibliothek

# Data Information Literacy

GEO 802 Fall 2020

Gary Seitz, MA


Anna C. Véron, Dr. sc. nat.

# Who we are...

UZH Mammut

Human geographer

Table Tennis Trainer




Gary Seitz, dipl. geogr.  
Subject Specialist Geography

Physical chemist

Knows how painful science can be

Nature Enthusiast

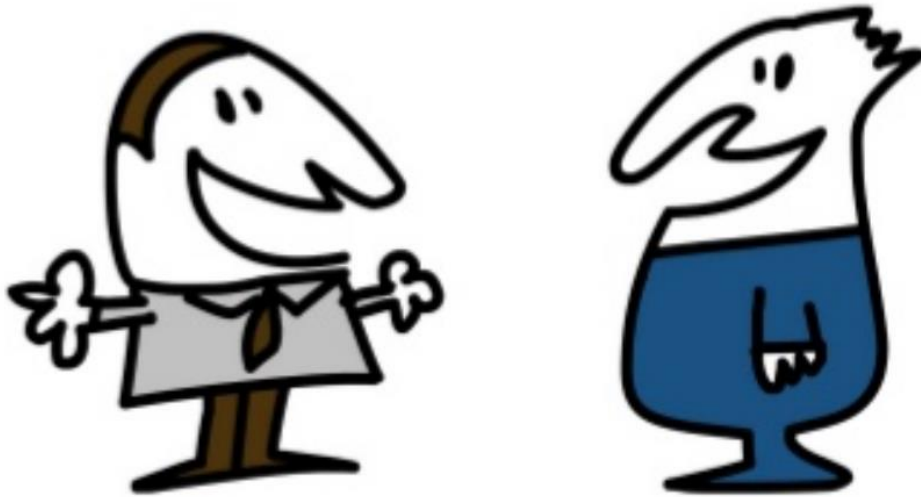


Dr. Anna C. Véron  
Subject Specialist Chemistry, Physics,  
Mathematics & Computer Science

[hbz.uzh.ch](http://hbz.uzh.ch)

# Tell us something about you!

- Scientific **background**
- **Motivation** to attend this course





# Course Schedule

## Day 1

1. Introduction	Anna	09:00-10:00
-----------------	------	-------------

---

*Break*

2. Discovery & Acquisition	Gary	10:15-11:30
----------------------------	------	-------------

---

*Lunch Break*

3. Data Entry / Creating Data	Anna	13:00-13:45
-------------------------------	------	-------------

---

4. Organizing Data	Gary	13:45-14:45
--------------------	------	-------------

---

*Break*

5. Data Types & Formats	Gary	15:00-16:00
-------------------------	------	-------------

---

# Course Schedule

## Day 2

6. Data Documentation & Metadata	Anna	09:00-09:45
----------------------------------	------	-------------

---

7. Storage, Backup, Security & Preservation	Anna	09:45-10:30
---	------	-------------

---

*Break*

8. Data Sharing, Reusing & Citation	Gary	10:45-11:45
-------------------------------------	------	-------------

---

*Lunch Break*

9. Ethics & Copyright	Gary	13:00-13:45
-----------------------	------	-------------

---

10. Data Management Planning	Anna	13:45:14:45
------------------------------	------	-------------

---

*Break*

---

<b>Exercise</b>		15:00-16:00
-----------------	--	-------------

---

# Your course goals

- You'll be able to apply efficient research data management techniques during your Master / PhD research project (and during your further career).
- We'll give you a «buffet» of knowledge and tools for data management.
- Only you can decide and pick what you need for your research project!



## To be handed in

- Your very first DMP (for your research project)
- Exercises solved during the class

Create a folder structure and file naming convention and upload your files to SWITCHdrive.

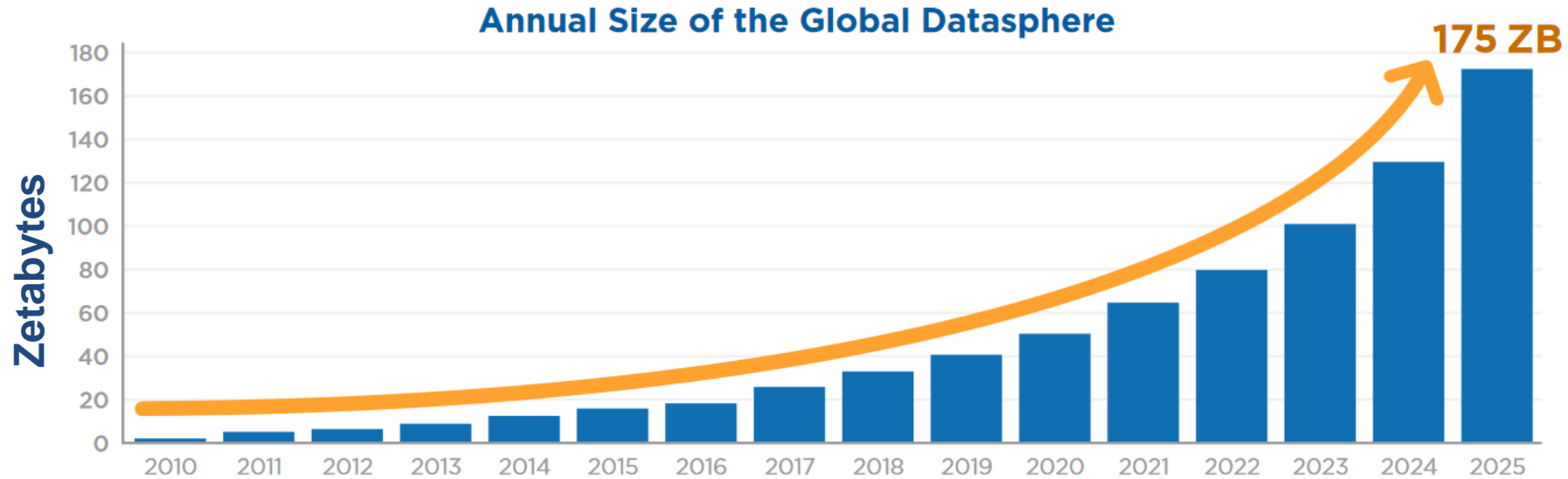


# Lesson 1: Introduction

→ **What is Research Data?**

- **The Importance of Data Management**
- **The Data Lifecycle**

# The Global DataSphere



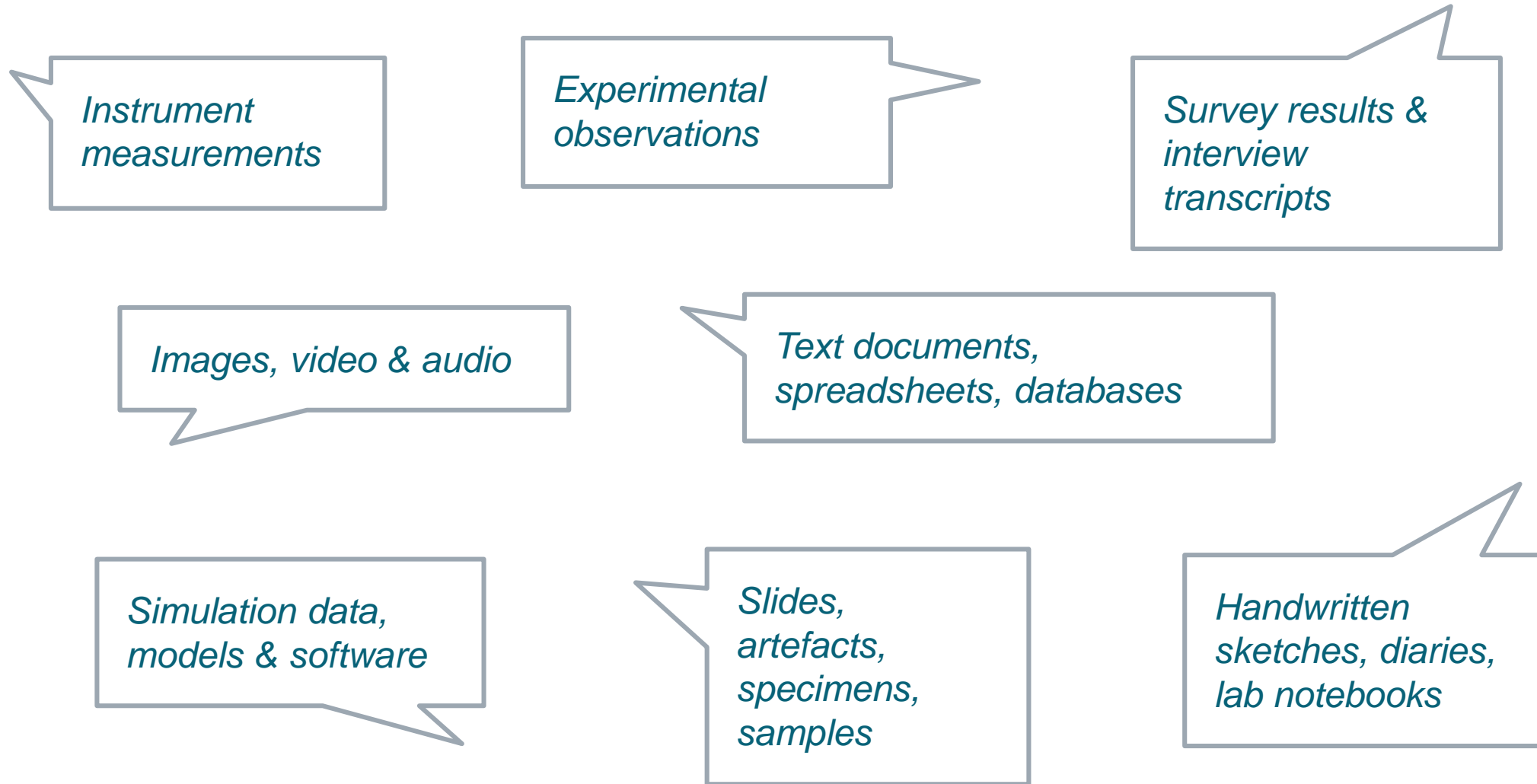
## How big is 175ZB?

Sometimes it can be difficult to get our minds around such a large number. Here are some illustrations of just how large 175ZB is.

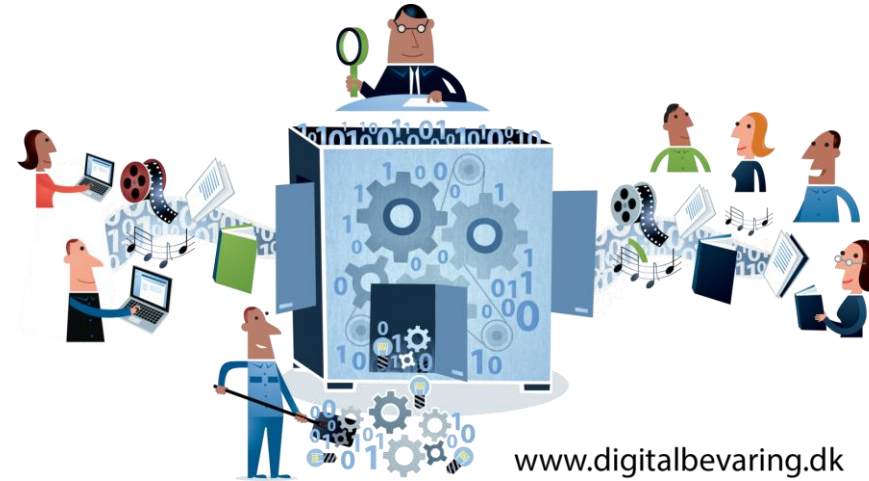
- One zettabyte is equivalent to a trillion gigabytes
- If you were able to store the entire Global Datasphere on DVDs, then you would have a stack of DVDs that could get you to the moon 23 times or circle Earth 222 times.
- If you could download the entire 2025 Global Datasphere at an average of 25 Mb/s, today's average connection speed across the United States, then it would take one person 1.8 billion years to do it, or if every person in the world could help and never rest, then you could get it done in 81 days.

Source: IDC White Paper, Doc# US44413318, November 2018. The Digitization of the World – From Edge to Core <https://www.seagate.com/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>

# What is Research Data?



# What is Research Data?



«Research data is collected, observed or generated factual material that is commonly accepted in the scientific community as **necessary to document and validate research findings.**»

[Open Research Data – Swiss National Science Foundation](#)

«Unlike other types of information, research data are collected, observed, or created, **for the purposes of analysis to produce and validate original research results**»

University of Edinburgh, MANTRA Research Data Management Training, 'Research Data Explained'

# Data Deluge

Increasing quality,  
resolution,  
precision,  
coverage...



Photo courtesy of <http://modis.gsfc.nasa.gov/>



CC image by CIMMYT on Flickr



Image collected by Viv Hutchinson

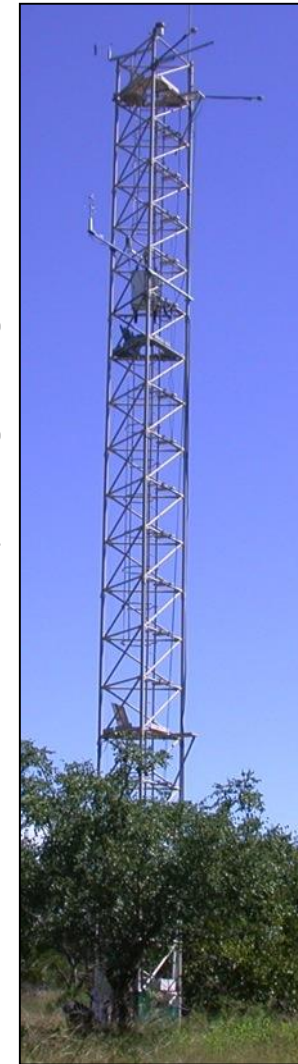
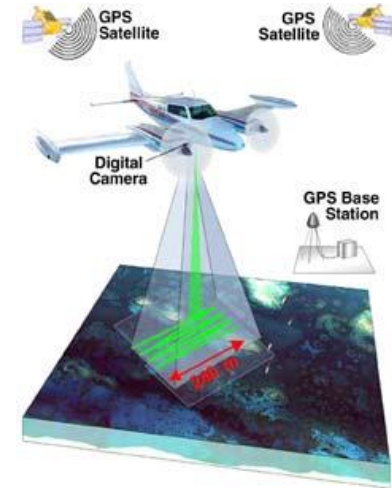


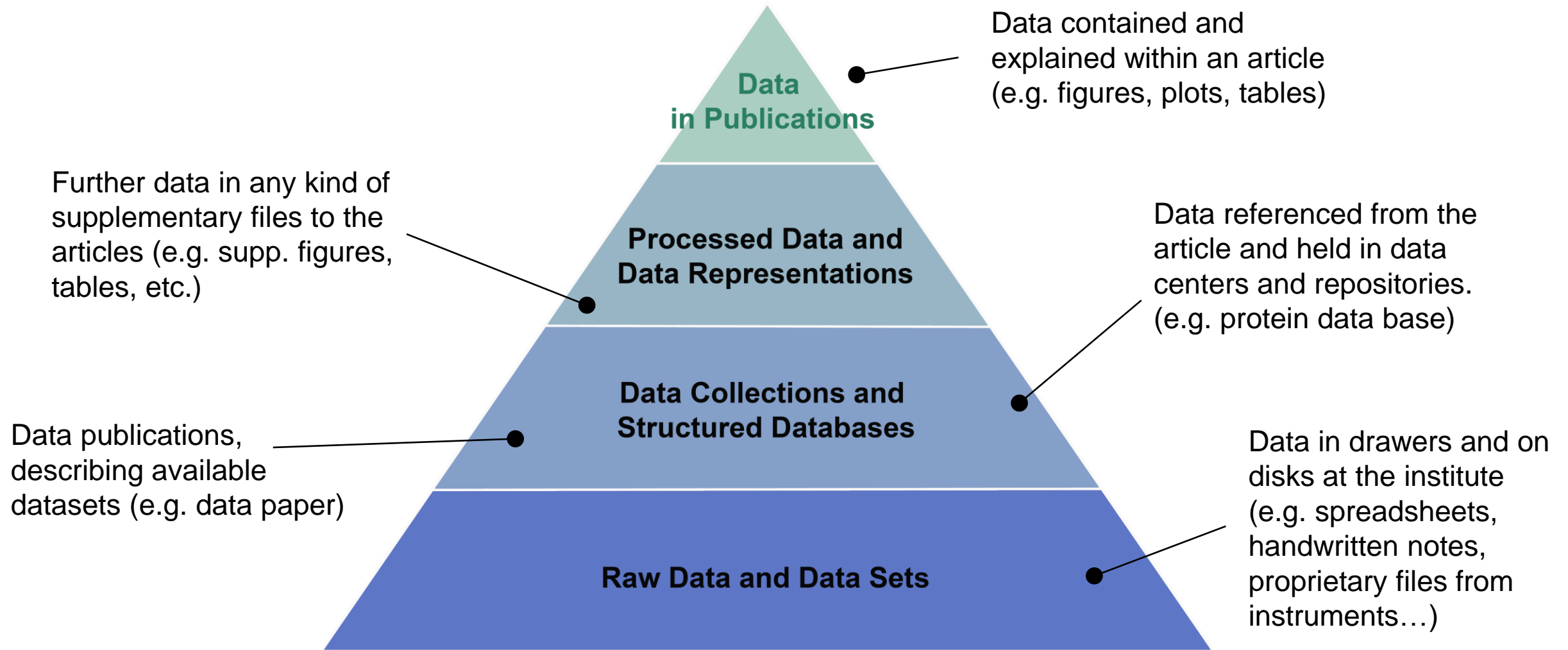
Photo courtesy of [www.carboatfrica.net](http://www.carboatfrica.net)



CC image by tajai on Flickr



# Different classes of research data



Adapted from: Reilly S et al. (2011) Report on integration of data and publications. Opportunities for Data Exchange (ODE).

# Research data are precious!

**Your  
research  
data**

=

**Time  
Resources  
Hard work  
Nerves  
Grey hair...**

**\$\$\$**

# Lesson 1: Introduction

✓ **What is Research Data?**

→ **The Importance of Data Management**

**Reproducibility Crisis**

• **The Data Lifecycle**



## Exercise 1.1: Why research data management?

As you watch the cartoon jot down the data management mistakes which interest or appal you.

[https://youtu.be/66oNv\\_DJuPc](https://youtu.be/66oNv_DJuPc)



# Why Data Management?

## Data Loss



CC image by Sharyn Morrow on Flickr



CC image by momboleum on Flickr

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- External dependencies (e.g. PKI failure)
- Format obsolescence
- Legal encumbrance
- Human error**
- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations and requirements

# Why Data Management?

## Data Loss



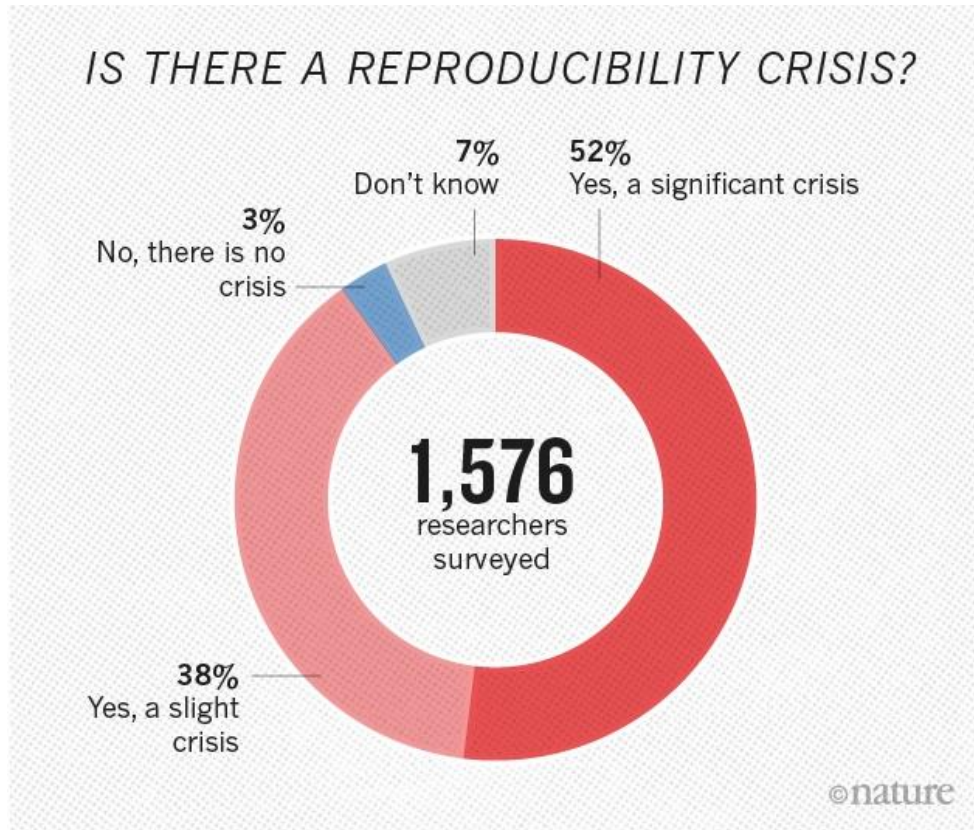
<https://www.flickr.com/photos/quinnanya/3239528185/in/gallery-wlef70-72157633022909105/>



Blick am Abend, 25.10.2018



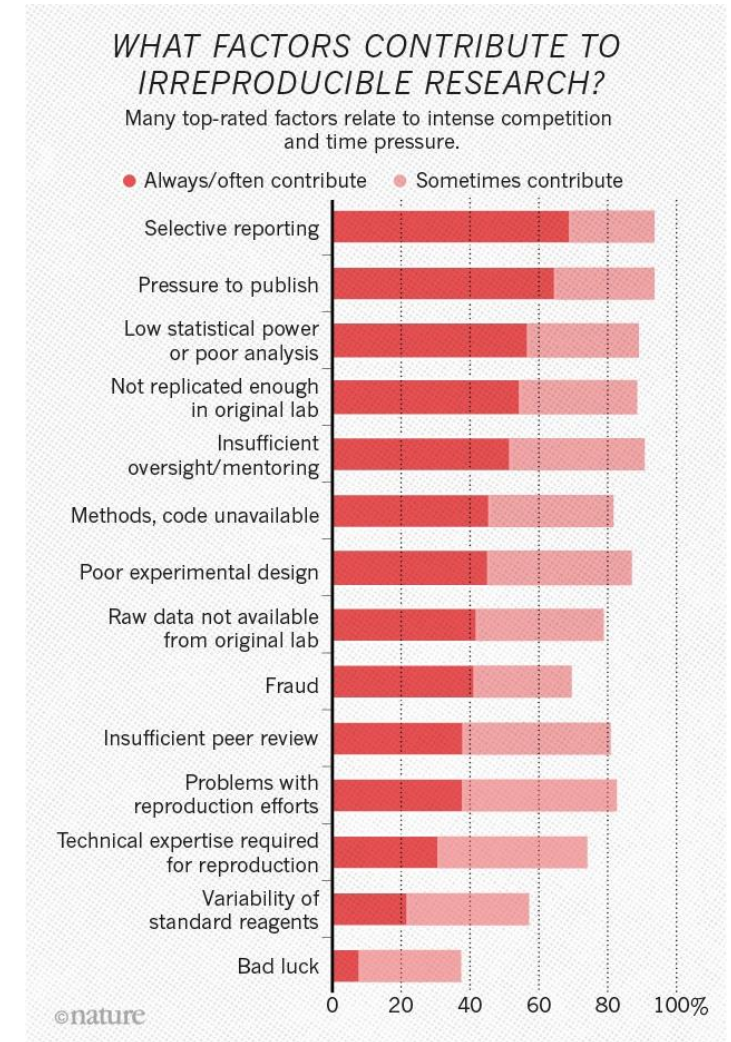
# Reproducibility crisis



“More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments.”

<http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970>

<http://www.nature.com/news/reality-check-on-reproducibility-1.19961>



# Exercise 1.2

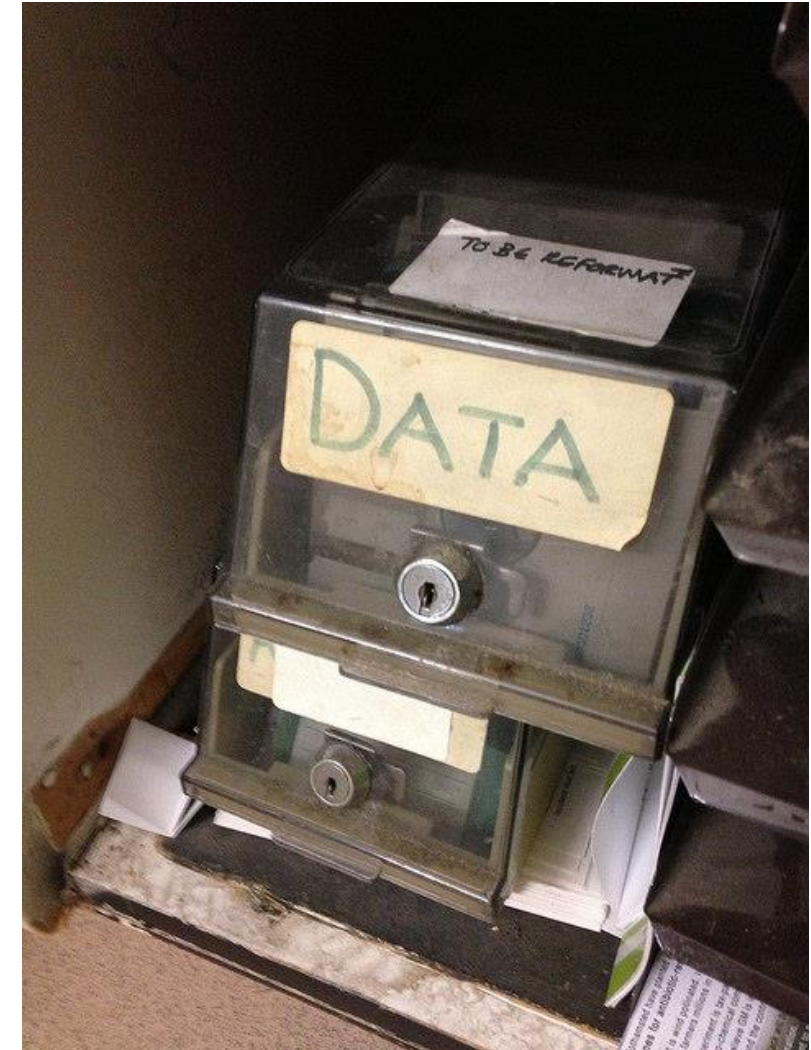
## **Please follow the instructions**

1. Put the paper with the blank side up in front of you
2. Fold in half
3. Fold in dotted line
4. Fold the outer edges up
5. Turn over

# Exercise 1.0

# What is research data management?

- **Research data management** is about how you **organize, describe, store and archive** the information used or generated during a research project
  - It includes: **How you deal** with data on a day-to-day basis **over the lifetime of a project**
    - folder structure, file name, format and versioning
    - metadata for retrieval
    - data storage and security
    - documentation for the publication
  - **What happens to data in the longer term** (after the project)



## Data management is not...

Data science

Computational science

Database administration

A research method:

- what data to collect
- how to collect them
- how to design an experiment



# Benefits of Data Management



## research integrity

- avoid fraudulent research
- difficult for people to produce «fake data»
- increased trust in your work

## Impact

- trust and reproduce experiments you find in the literature
- save costs of repeating similar experiments over and over

## save time & money

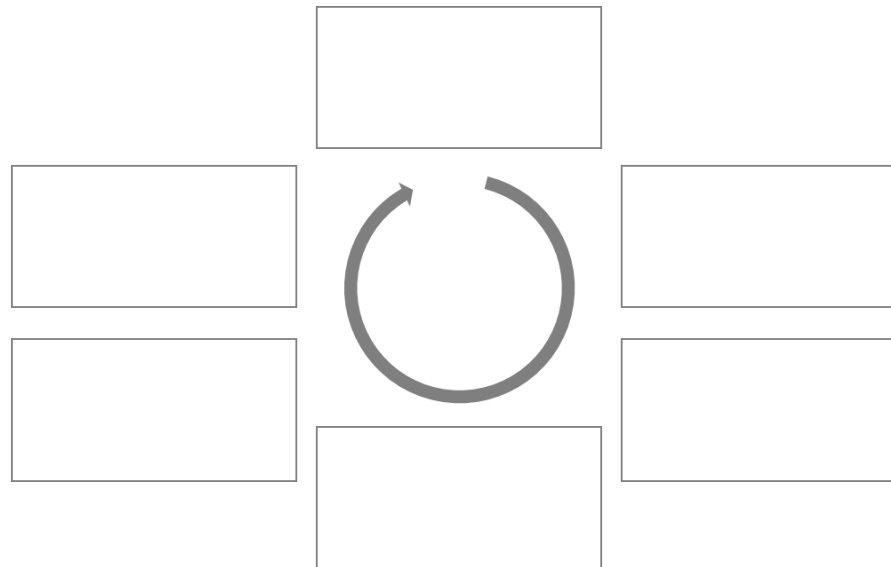
- increase the lifetime of your work
- reuse of your data will lead to an increased visibility, citations and impact
- when you write a paper or thesis, you know where to find which data
- after a student left the group, you are still able to find and understand their data
- new students don't have to repeat old work over and over again

# Lesson 1: Introduction

- ✓ **What is Research Data?**
- ✓ **The Importance of Data Management**
  - **The Data Life Cycle**
    - Sharing Data, Open Data**

# Exercise 1.1: Research Data Life Cycle

1. Arrange the Post-Its as a lifecycle in way that makes sense to you.
2. Read the snippets and arrange them within the lifecycle.
3. Create a PDF of your finished product and save it for the Upload to SWITCHdrive.



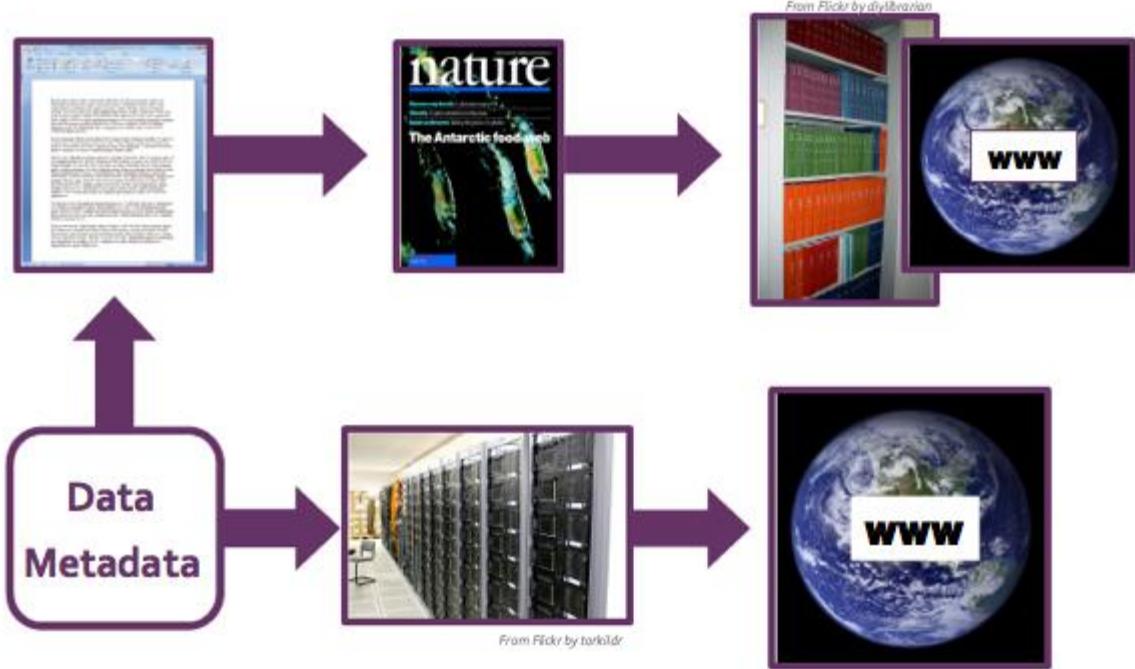
# Exercise 1.1: Research Data Life Cycle

# Where a majority of data end up now



Recreated from Klump et al. 2006

# If data were more accessible



Recreated from Klump et al. 2006

# Thoughts about Sharing Data

## Barriers

Lack of experience with open science in your institute	PI demands conventional science
Lack of appropriate journals to publish open access	In some disciplines, careers are built on impact factors
Worries of being scooped	Lack of resources/money to publish openly

vs.

## Incentives

Institutional support	Requirement of funding organizations
Save time and money in the role term	Be a role model
Receive more citations and visibility	Increased credibility

# Summary of Lesson 1

Data deluge: **The amount of data created every year is increasing exponentially**

Improper data management can be **costly**

Data management allows you to **find, access, understand, integrate and re-use data.**

## If data are:

- ✓ Well-organized
- ✓ Documented
- ✓ Preserved
- ✓ Accessible
- ✓ Verified to accuracy and validity



## The benefits are:

- ✓ High quality data
- ✓ Easy to share and re-use
- ✓ Citation & credibility for the researcher
- ✓ Saving costs