Piano Nazionale di Ripresa e Resilienza

Data Infrastructure

Lecture 1: Data & Research Data Management

Federica Bazzocchi
24/3/2025

Brief introduction to Data Infrastructure section:

- Welcome
- Goals
- Calendar
- Topics and Organization

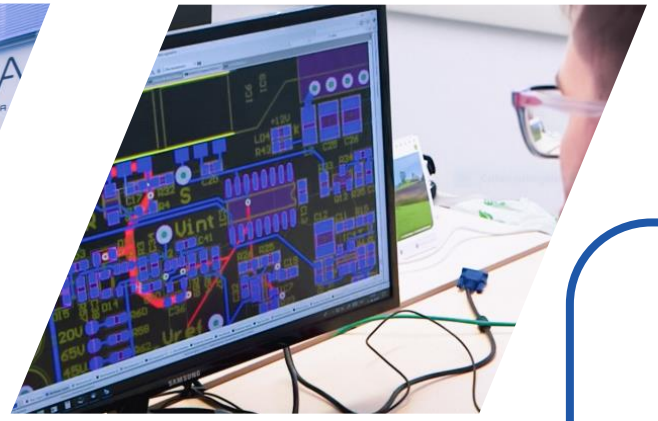
Welcome

- Few words about you
- My contact: Federica.Bazzocchi@areasciencepark.it
- My institute: RIT@AreaScience Park

Three laboratories active in creating an **integrated system** of research infrastructures and platforms.

GENOMICS AND EPIGENOMICS LABORATORY
LAGE

ELECTRON MICROSCOPY LABORATORY
LAME



DATA ENGINEERING LABORATORY
LADE

DATA INFRASTRUCTURE

DATA MANGEMENT AND DATA CURATION

DATA SCIENCE

Welcome

- @LADE we host students'internship and have different possibilities for undergraduated and graduated fellowhips! (if interested ask me!)
- We organize together with SISSA a Master in Data Management and Curation ! The pilot edition is finishing but the next one will start on September 2025

- **Open Science principles and methodologies**, within the context of Horizon Europe Framework programme and EOSC;
- **FAIR principles**: data FAIR-by-design approach and FAIRification of data processes;
- **Tools and software for data acquisition and metadata enrichment;**
- **Tools and methods for preliminary data and metadata analysis.**

Master in Data Management and Curation (MDMC)

Goals

- Introduce/review the concepts of data management
- Discuss the concepts of data infrastructure and storage ecosystem
- Data infrastructure sustainability
- Present some examples/tools and some parallelism between research and companies approach
- Discuss some specific examples: DECOS and OFED

Calendar :

Timetable	24-mar	27-mar	07-apr	10-apr	14-apr	17-apr	28-apr	05-may
09:00-11:00	online		online		online		online	online
11:00-13:00		presence		presence		presence		

Topics:

- Research Data Management ad tools
- Large scale data infrastructure and hardware/software stack for large data management
- Parallel and distribute storage
- Cloud storage and associated services

Organization:

- Frontal lessons
- Active participation from you
- Interactive session on an example of data management in material science: OFED and its services
- Seminar to discuss of an example of data management: DECOS in the PRP project

LECTURE 1 OUTLINE:

- Some Reflexion on Data
- Research Data Management
- Introduction to Data Infrastructure

REFLEXION ON DATA

WHAT IS

DATA

?

My old (ingenuous) BIAS as young physics' student (that I am not anymore) and science enthusiast (that I still am)

- Data is a results of a measurement;
- It is objective;
- It is quantitative;
- It is related to a physical law/phenomena;
- It is analyzed by mean of statistics and is used to validate a/make prediction by a (analytical) mathematical model.

DATA is a wider concept

- "Data is any set of characters that has been gathered and translated to some purpose, usually analysis. It can be any character, including text and numbers, pictures, sound, or video"
  (https://www.computerhope.com/jargon/d/data.html)
- "Data is information that has been translated into a form that is efficient for movement or processing“
  (https://searchdatamanagement.techtarget.com/definition/data)
- · "Data is a collection of facts, such as numbers, words, measurements, observations or even just description of things"
  (https://www.mathsisfun.com/data/data.html)

DATA journey

- The word “data” derives from the Latin word “datum” (singular), which means the “thing given”
- A **data** can be defined as a fundamental unit of **raw information**, represented in different form that can be transferred and then **recorded, processed, analyzed** and then **interpreted**.
- 19000 B.C-**calculation**- Ishango bone (baboon tool) : the first **mathematical data** (intended as information registered)
- I640s –**medical data-** John Graunt started collecting information regarding deaths in London (number of death, mortality rate per age, causes)
- 1880s-**data processing-** the German-American statistician Herman Hollerith had the idea of using punch cards in writing and processing data. With this invention Hollerith helped the American government complete the US census within the same year.
- 1928 – **magnetic tape**- German engineer Fritz Pfleumer patented a magnetic tape that he used to replace wire recording for storing data.
- 1960s – **relational database** – idea introduced by the computer scientist Codd
- 1990s **–** Internet and then Google **– the rise of big data**

DATA is raw and processed information

Key Characteristics

- Primality – A data point on its own is neutral and meaningless without context.
- Representation – expressed in different formats (numerical, textual, binary, images, sound signals)
- Storage – It can be recorded on physical or digital media.
- Processing and Analysing –extraction of **useful information**
- Transferability – It can be transmitted and exchanged between systems, individuals, or devices.

Kind of Data

- **Observational**: real-time captures (e.g. brain images, survey data)
- **Experimental**: from experimental results (e.g. from lab equipment)
- **Simulation**: generated from test models (e.g. economic or climate models)
- **Derived or compiled**: resulting from processing or combining 'raw' data (e.g. compiled databases, text mining, aggregate census data) ·
- Reference or canonical: collection of datasets, usually published and curated (e.g. gene databanks, crystallographic databases)

Types of Data

- Structured Data: Organized in tables or databases (e.g., name, age, address).
- Unstructured Data: Texts, images, videos, audio, without a predefined structure.
- Semi-structured Data: JSON, XML, which have some organization but are not as rigid as relational databases.
- Qualitative vs. Quantitative Data: Words vs. numbers.

Data Sources

- Data are not only created anymore to write scientific papers, but they are created with the notion of being reused in different contexts which is revolutionary in many disciplines;
- Data are produced by almost everything
- Advanced statistical methods (machine learning/deep learning) are required and have allowed to detect the patterns and correlations hidden in the data (without LLM would have been impossible NLP)

DATA-CENTRIC WORLD

In 2006, mathematician Clive Humby coined the phrase "**data** is **the new oil**."

Forbes

INNOVATION

Data Is The New Oil -- And That's A Good Thing

By Kiran Bhageshpur , Forbes Councils Member.
for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

Nov 15, 2019, 08:15am EST

INNOVATION

Data Is The New Business Fuel, But It Requires Sound Risk Management

By Morgan Palmer , Former Forbes Councils Member.
for Forbes Technology Council, COUNCIL POST | Membership (fee-based)

Apr 28, 2022, 08:00am EDT

Why Are We Moving Towards A Data-Centric World?

Ian Gerald King · Follow
4 min read · Feb 23, 2017

CYBERSECURITY

We need a new era of data responsibility

Jan 21, 2018

A parentheses

- Data intended as digital recorded infomation is a technical concept and is processed /analyzed by highly specialized techniques and professionals (data scientists, data and AI engineers)

- Nevertheless our society is so data-centered (and data-obsessed) that also humanities/social studies are interested in the influence of data for our society. They enter into the dialogue about DATA and their own studies are influenced by DATA.

LECTURE/PRESENTATION/TALK

Mellon Sawyer Seminar Series: Catastrophe, Data, and Transformation (Dagomar Degroot and Jessica Otis)

Sponsored by Center for Spatial and Textual Analysis (CESTA)

May 21st, 2024 | 6 min read Arts & Humanities

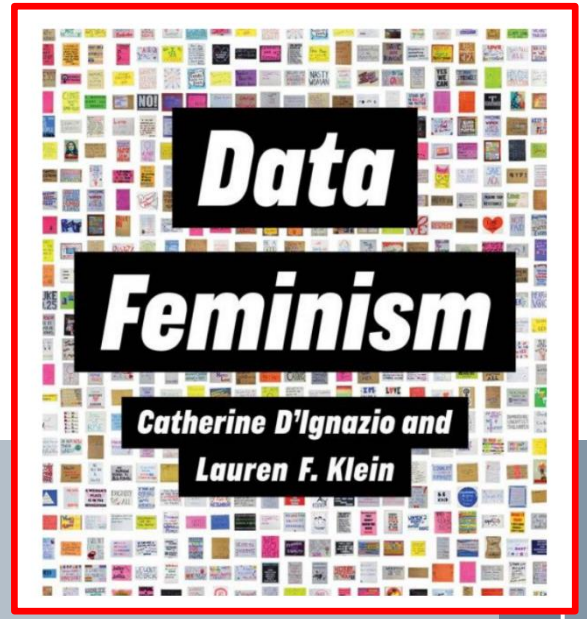
Lectures explore data's place in the humanities

A seminar series takes a humanistic approach to extracting meaning from the numbers that saturate our world.

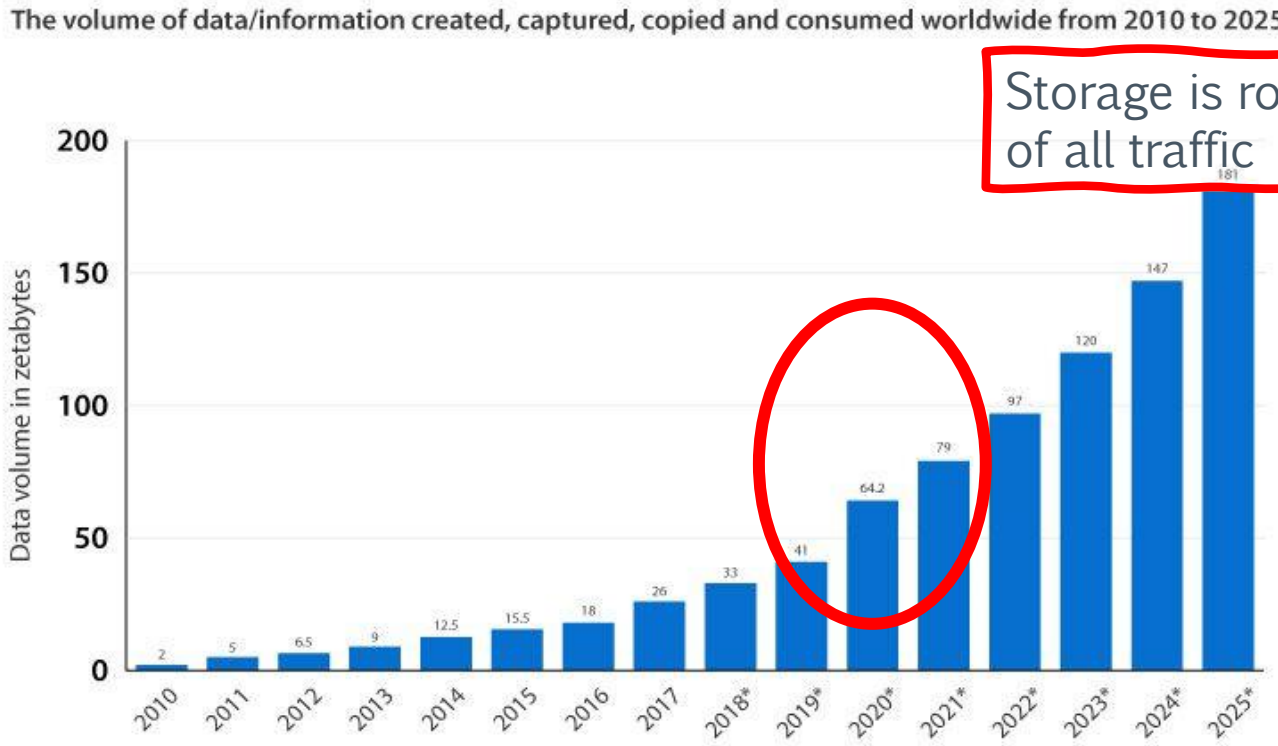
WHAT IS DATA ETHICS?

Data ethics encompasses the moral obligations of gathering, protecting, and using personally identifiable information and how it affects individuals.

"Data ethics asks, 'Is this the right thing to do?' and 'Can we do better?'" Harvard Professor Dustin Tingley explains in the Harvard Online course Data Science Principles.

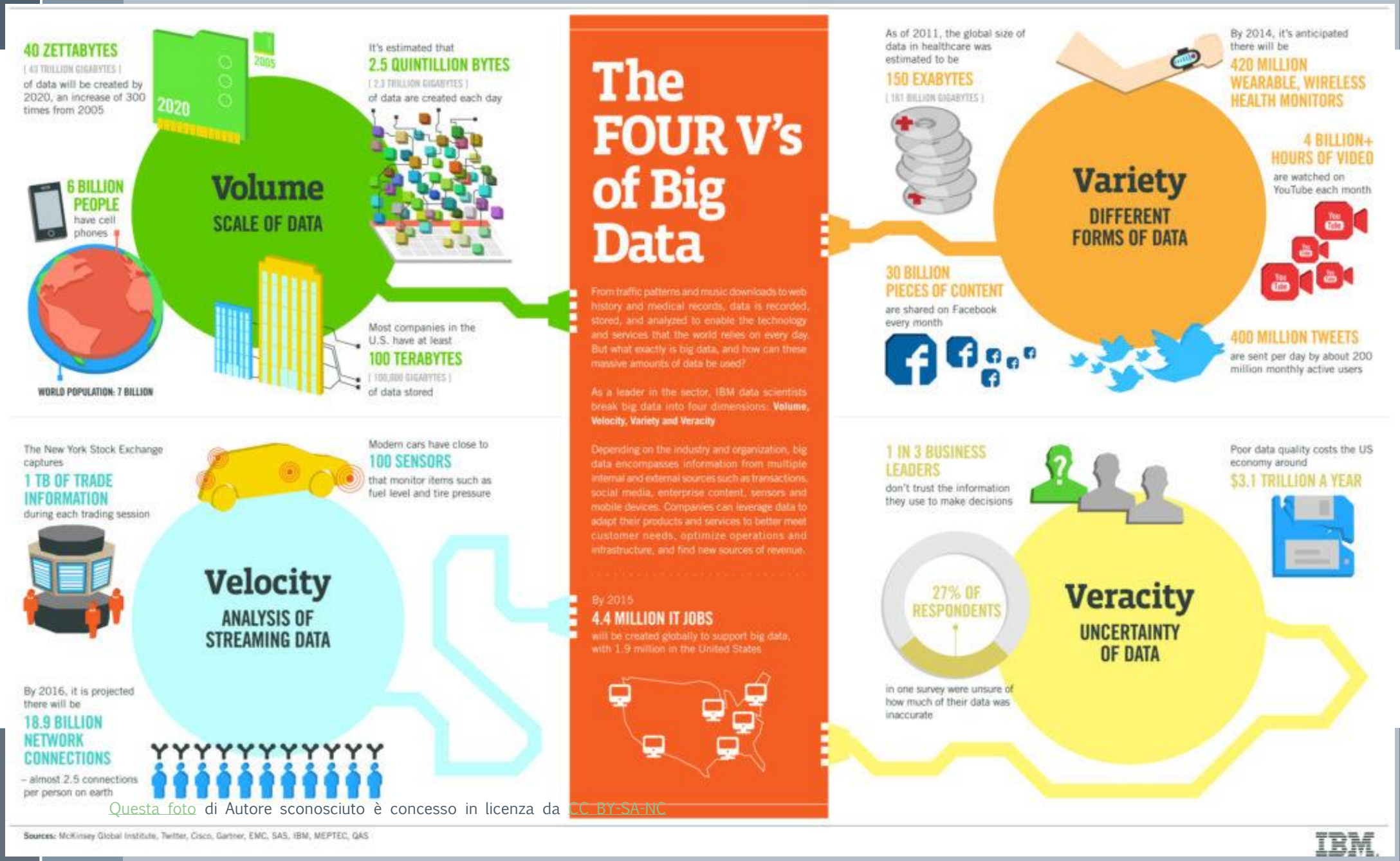


DATA explosion in the last decades



Storage is rouhghly 10% of all traffic

Year	World Storage Size (Exabytes)
1986	2.6 EB
1993	15.8 EB
2000	54.5 EB
2007	295 EB
2014	5000 EB
2020	6800 EB



HOW WE MANAGE THIS

HUGE AMOUNT OF DATA

?

WHERE WE STORE THIS

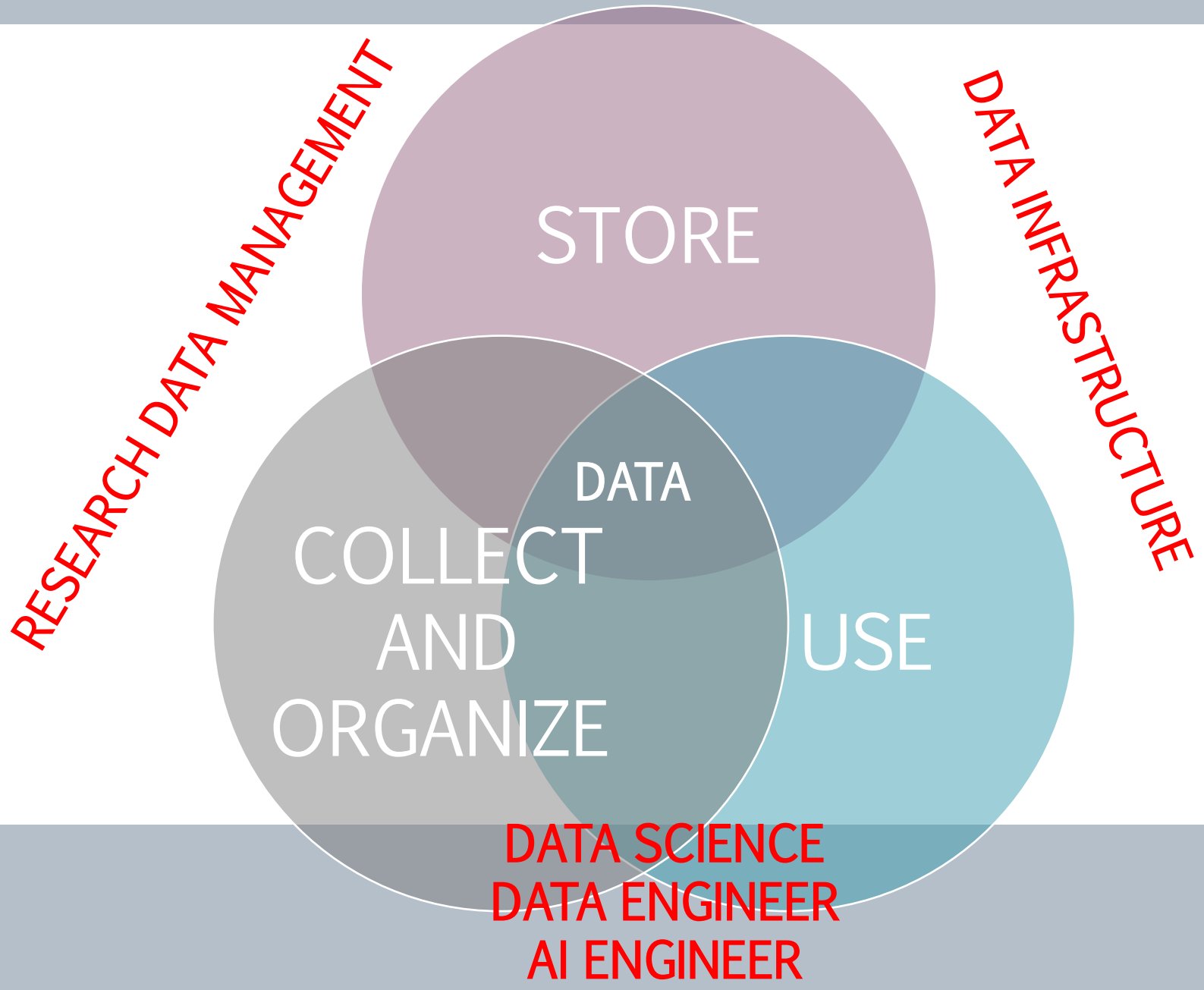
HUGE AMOUNT OF DATA

?

HOW WE USE (AND WHAT WE MAY DO WITH) THIS

HUGE AMOUNT OF DATA

?


RESEARCH DATA MANAGEMENT
DATA INFRASTRUCTURE
STORE
DATA
COLLECT AND ORGANIZE
USE
DATA SCIENCE
DATA ENGINEER
AI ENGINEER

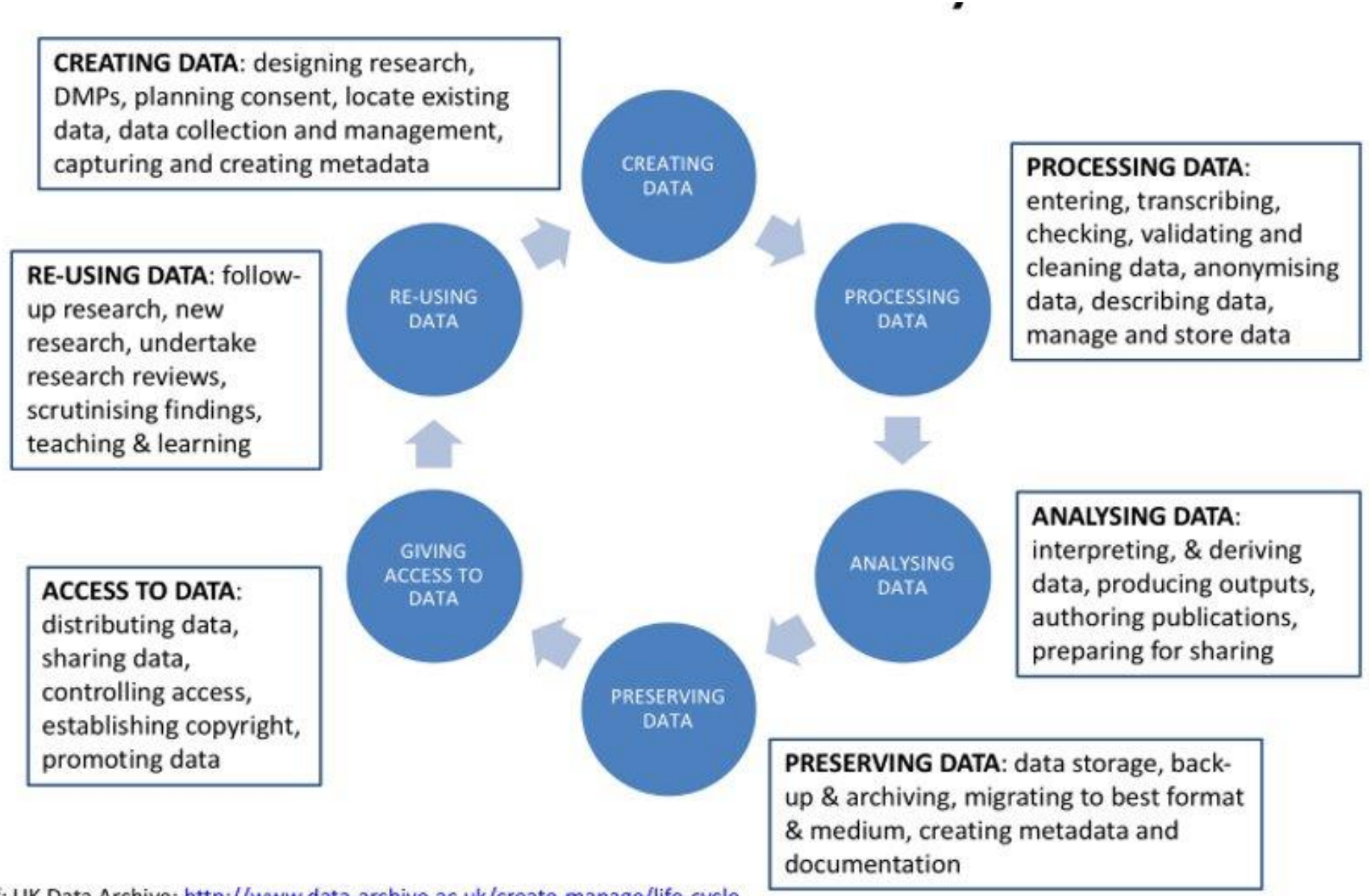
RESEARCH DATA MANAGEMENT

BIG DATA IN SCIENCE

- Data growing exponentially, in all sectors and therefore also in all science ·
- All science is becoming data-driven and this is happening very rapidly
- Data becoming increasingly open/public
- A scientific revolution in how discovery takes place => a rare and unique opportunity
- Data have **to be managed** adequately

Cross domain/context problem!

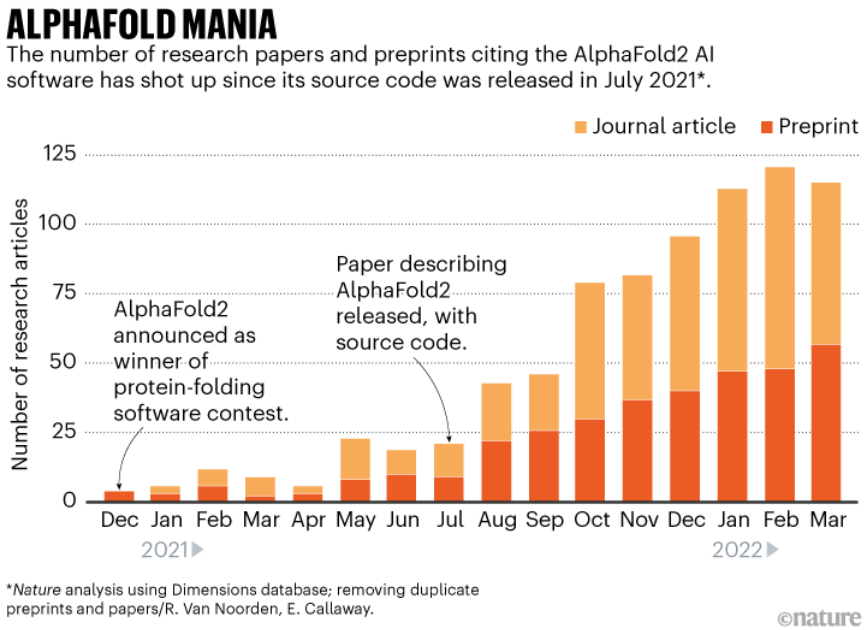
DATA LIFE CYCLE



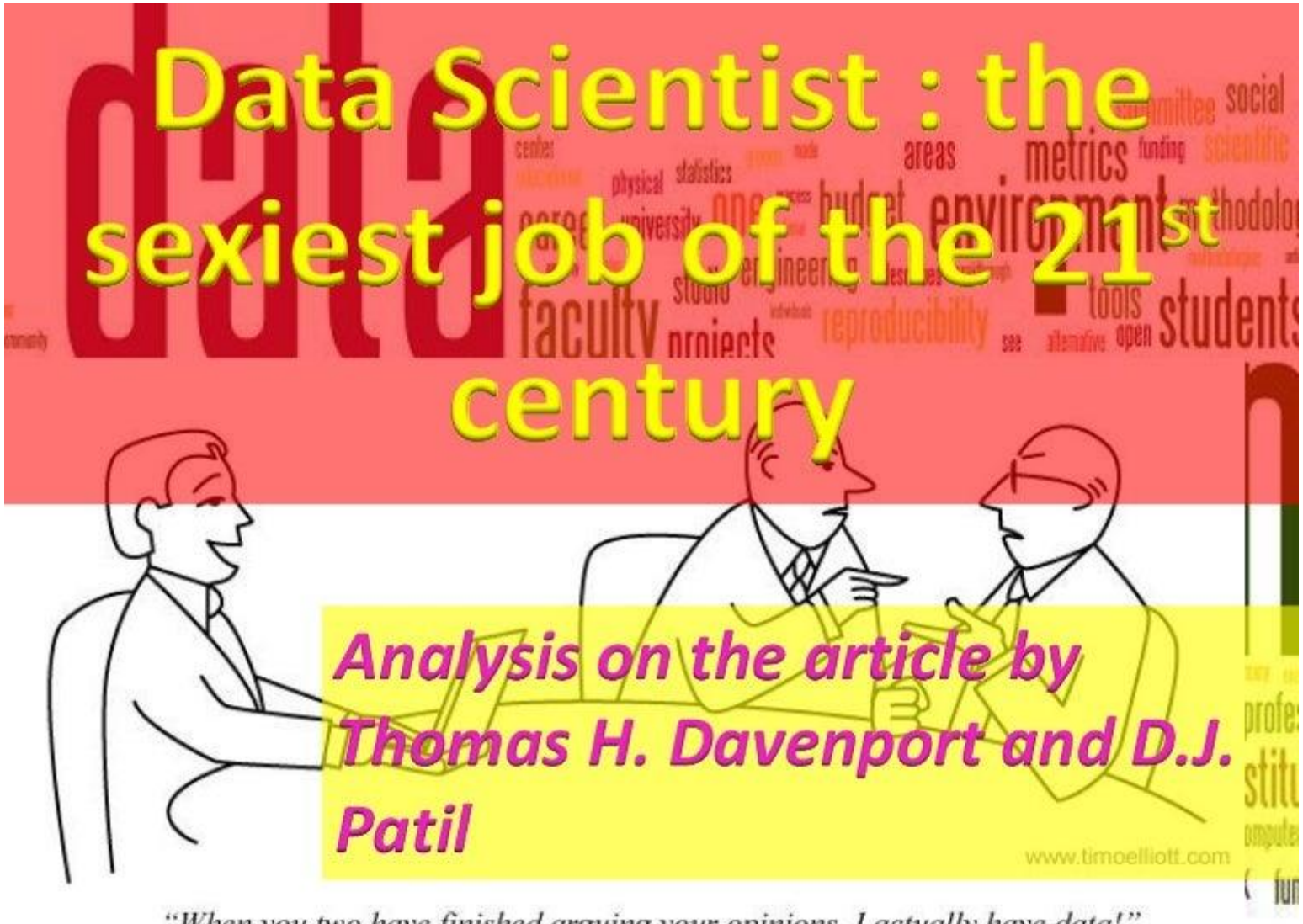
Ref: UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

NOT – INCREMENTAL CHALLENGES!

- Multi-faceted challenges in the analysis as well
- New computational tools and strategies
- … not just statistics, not just computer science, not just astronomy, not just genomics…
- Science is moving increasingly from hypothesis driven to data-driven discoveries and now to LLM-driven discoveries (think to alphafold/matgen)



Data Scientist : the sexiest job of the 21st century



Analysis on the article by Thomas H. Davenport and D.J. Patil

"When you two have finished arguing your opinions, I actually have data!"

SQUIRRELS

- They eat too much at once

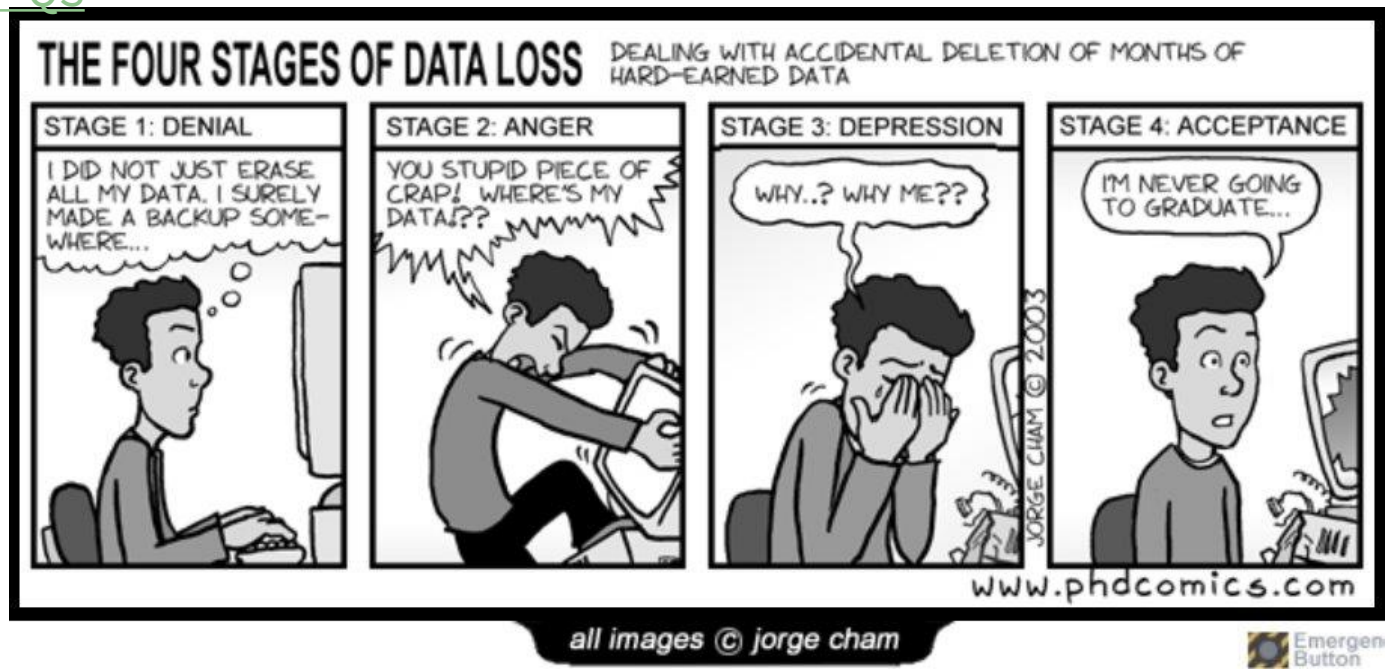
They forget where they store their food

SCIENTISTS ARE LIKE SQUIRELLS

They collect too much data at once

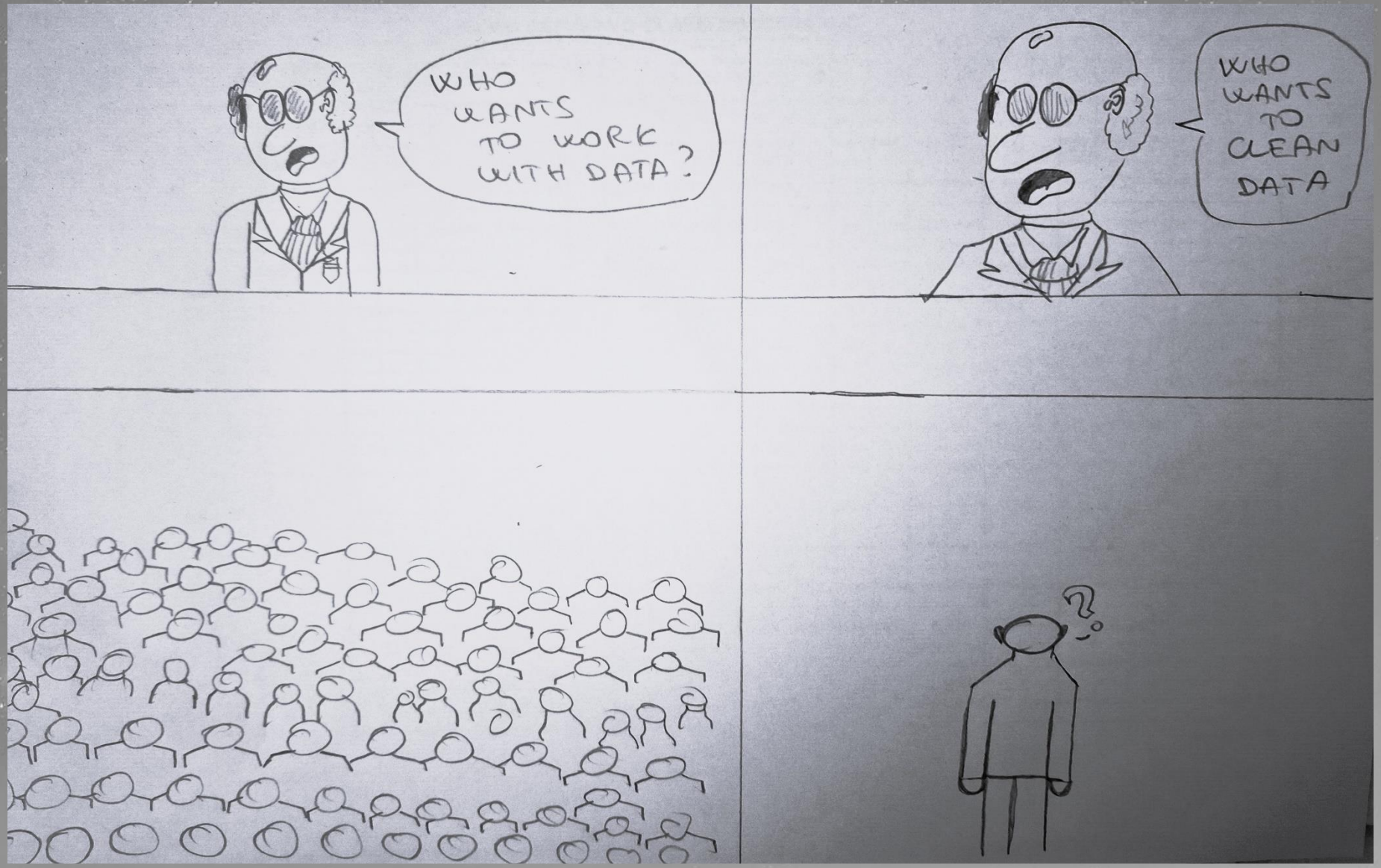
[QS World University Rankings by Subject 2015 - challenges and developments - QS]

They forget where they stored them!



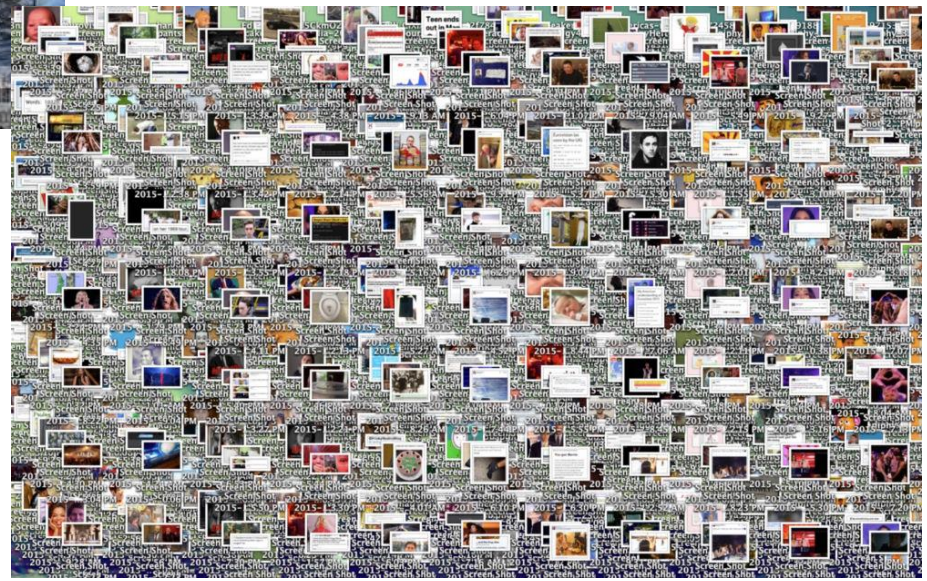
PHD Comics: Stages of Data Loss

DATA SCIENTIST IS THE SEXIEST JOB IN 21st CENTURY?



DATA MINING FOR SURE NOT!!!

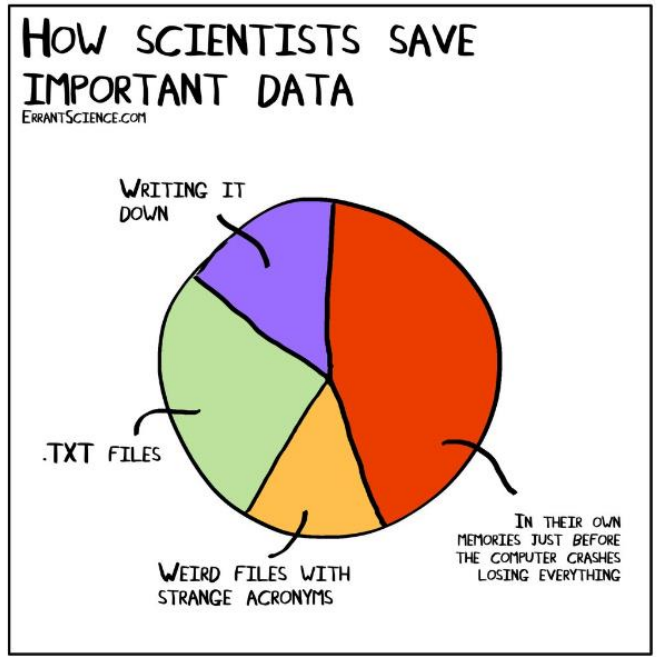
DATA SCIENTIST IS THE SEXIEST JOB IN 21st CENTURY?



DO SCIENTISTS NEED DATA MANAGEMENT?

subject: External hard disk lost
Organization: S.I.S.S.A.
Date: Wed, 4 Jul 2018 13:54:48 +0200
From: Students' Secretariat <XXXX@sissa.it>
To: SISSA Users:;

An external hard disk has been lost, most probably on the 4th floor, black, in a white box.
It contains a lot of work data of a SISSA PhD student.
If you happen to find it, please leave it at the reception desk or at the students' secretariat. Alternatively you can leave it in the Students' Secretariat mailbox in the lower level.



Marconi: scratch is almost full – quota imposed

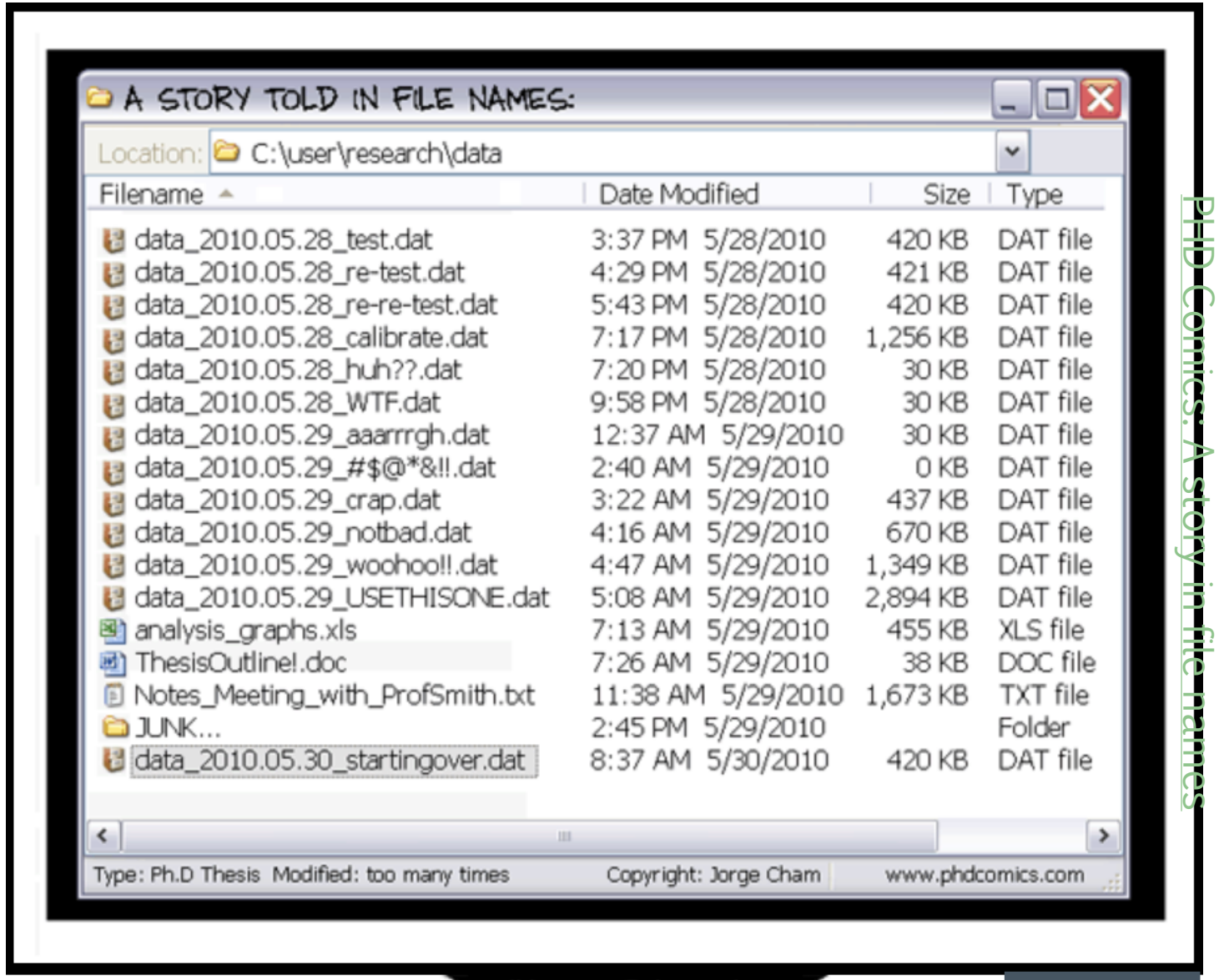
16 May 2024

Dear Marconi Users,

we inform you that the scratch space has reached the occupation of more than 87% today. This may cause malfunctions to the filesystems. To avoid reaching a 100% occupancy, we temporarily set a quota of 20 TB on the scratch folder of each user. We encourage you to clean your scratch folders by removing useless data or by moving data to work and dres spaces. We will inform you as soon as normal occupancy will be restored and the quota removed.

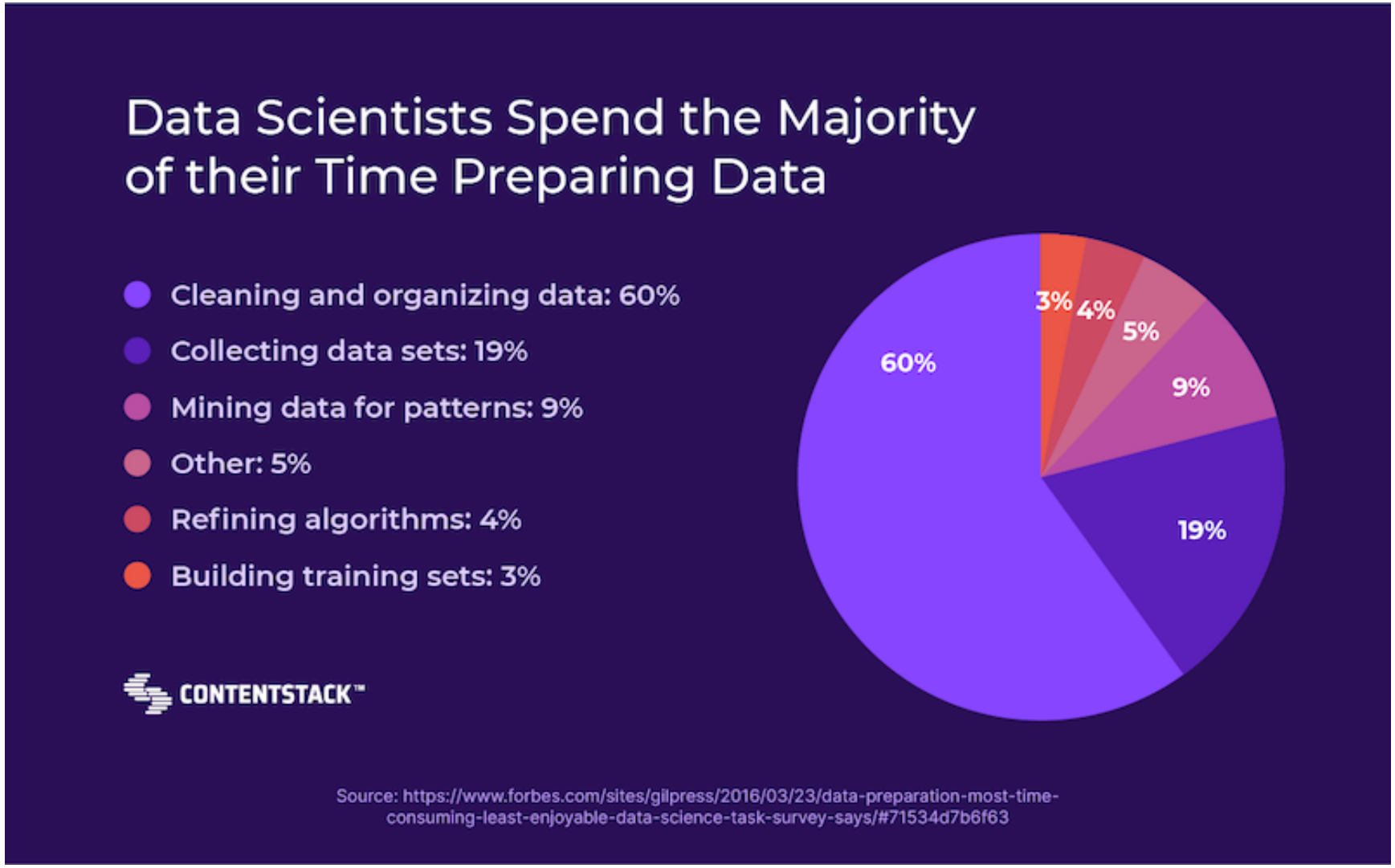
Best regards,
HPC User Support @ CINECA

DO SCIENTISTS NEED DATA MANAGEMENT?



PHD Comics: A story in file names

DO SCIENTISTS NEED GOOD DATA MANAGEMENT?



DATA MANAGEMENT ROLE IN SCIENCE

- "*Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly*".

- "*Data management should be woven into every course in science, as one of the foundations of knowledge*"

'Editorial: Data's Shameful Neglect'

(**10 September 2009**) in Nature 461, p. 145, doi:10.1038/461145a.

What is Research Data Management (RDM)?

Data management refers to all aspects of creating, housing, delivering, maintaining, and archiving and preserving data. It is one of the essential areas of responsible conduct of research

Ensures data integrity, accessibility, and compliance with regulations

Supports reproducibility, transparency, and long-term usability of research outputs

Increasingly, universities and research center now encourage all researchers (including postgraduate students) to undertake data management plans (DMPs) at the start of their research project

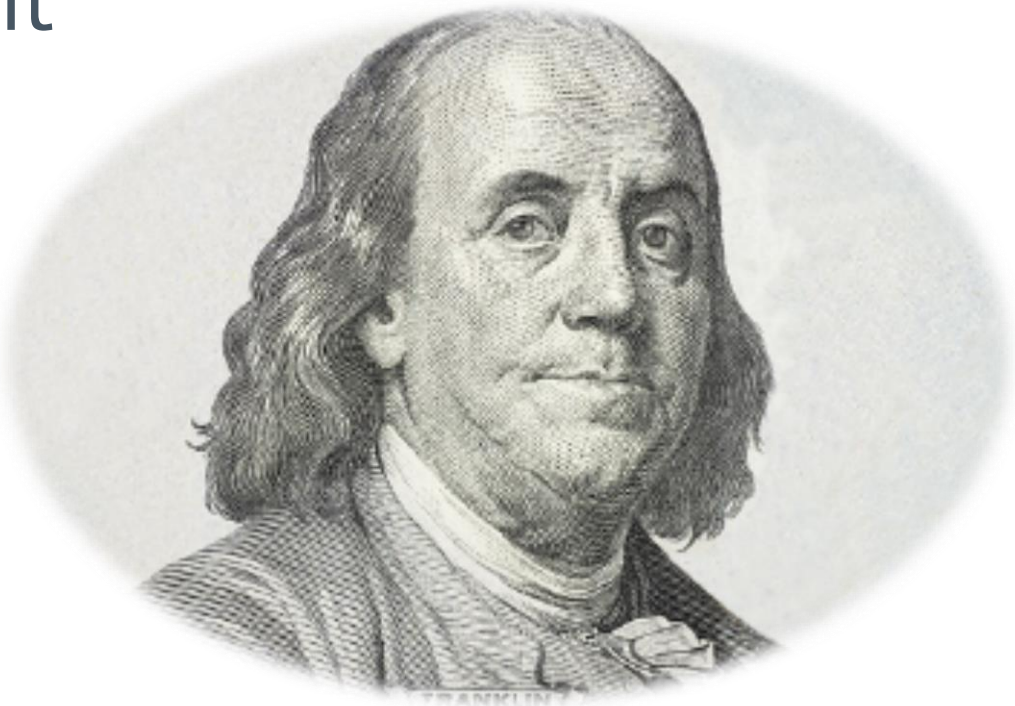
Before starting a new research project, Principal Investigators (PIs), research teams, and postgraduate students must address issues related to data management

EU-funded projects require a data management plan

Why is RDM Important?

- Enhances data quality and integrity
- Facilitates data reuse and collaboration
- Complies with funding agency and institutional policies
- Prevents data loss and ensures long-term preservation
- A good data management allows progress in research in a more direct way, without reinventing the wheel each time, both locally and within the community

“If you think ~~education~~
data management is expensive,
Try ~~ignorance~~ without it”



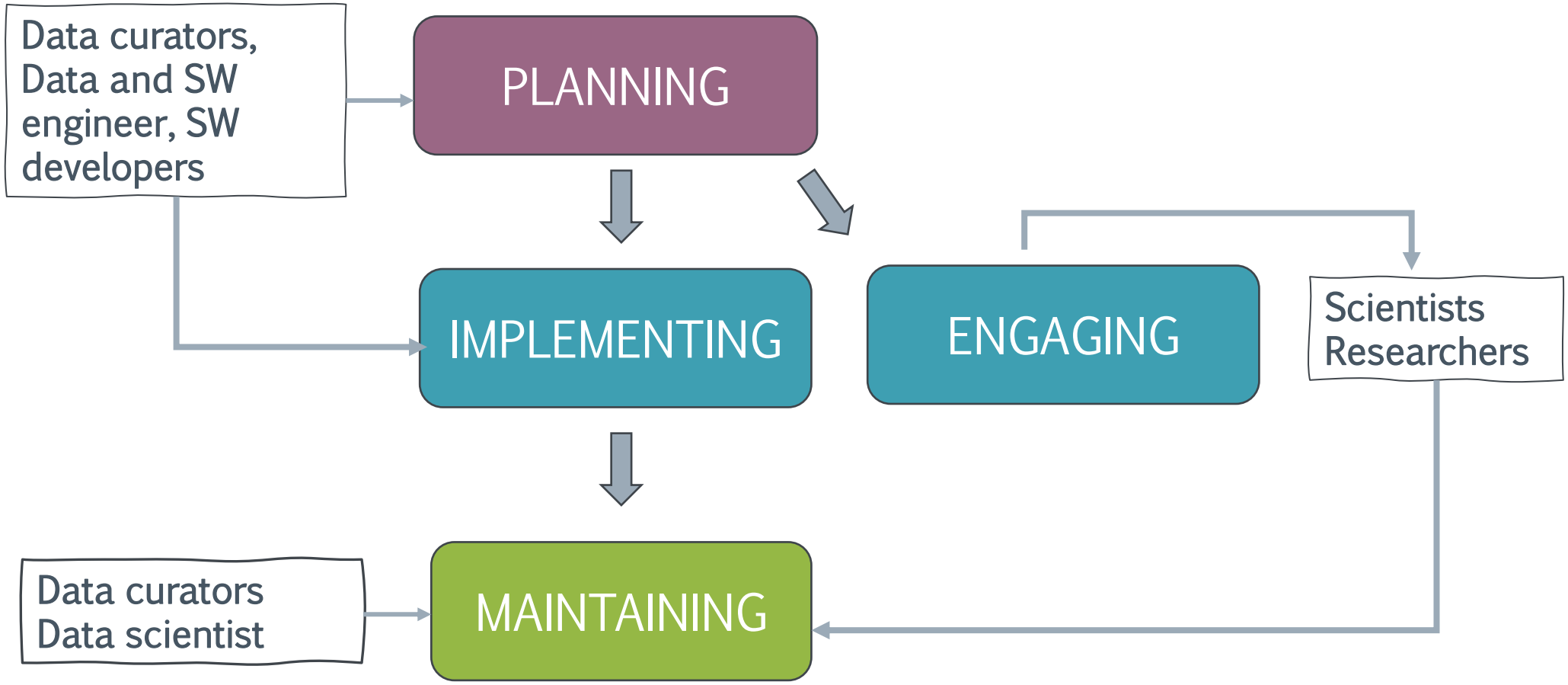
Benjamin Franklin

Data Management Priorities

- A higher degree of **interoperability** is required to overcome the huge fragmentation;
- Data scientists have to face too much detail in an increasingly **complex** data and tool landscape;
- Data scientists need wide scale data tracing and **reproducibility** mechanisms to facilitate trust and verification;
- Improved ways are needed to automatically create scientific **annotations** to capture and exploit knowledge

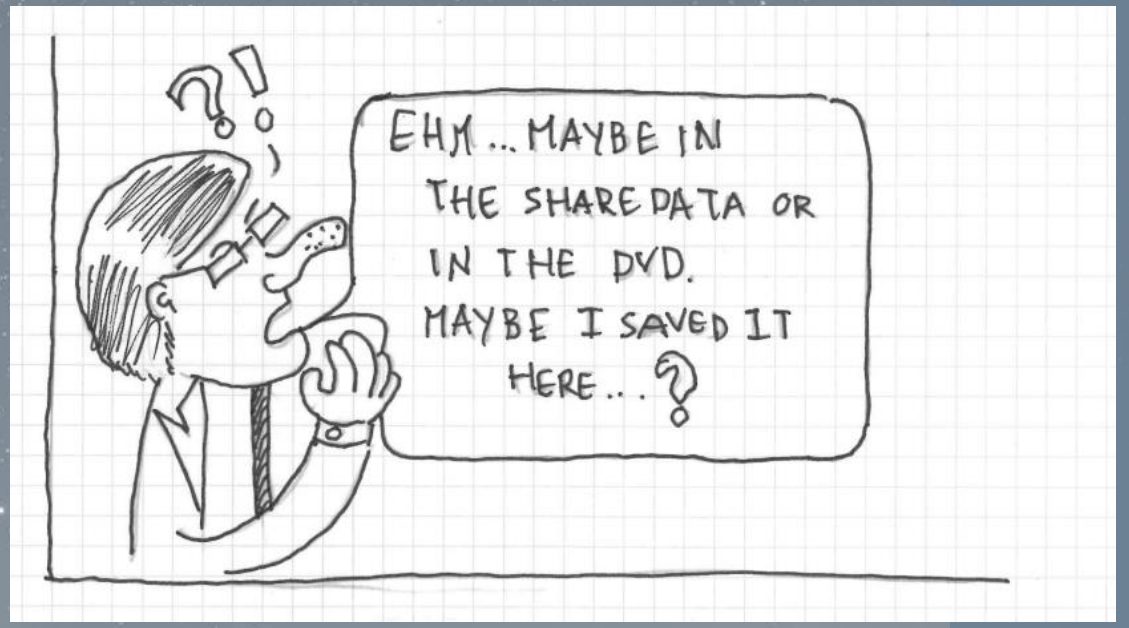
Roles in data intensive science:

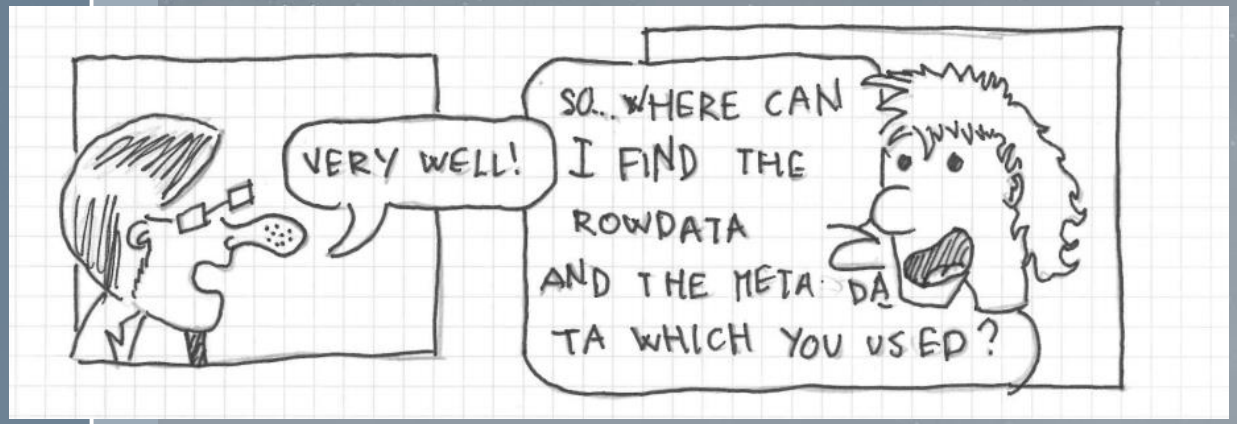
- **Scientists/researchers**: acquire, generate, analyze, check, organize, format, document, share, publish research data
- **Data scientists/users**: access, understand, integrate, visualize, analyze, subset, and combine data
- **Data engineers**: develop infrastructure, standards, conventions, frameworks, data models, Web-based technologies
- **Software developers**: develop tools, formats, interfaces, libraries, services
- **Data curators**: preserve data content and integrity of science data and metadata in archives
- **Research funding agencies, professional societies, governments**: encourage free and open access to research data, advocate elimination of most access restrictions

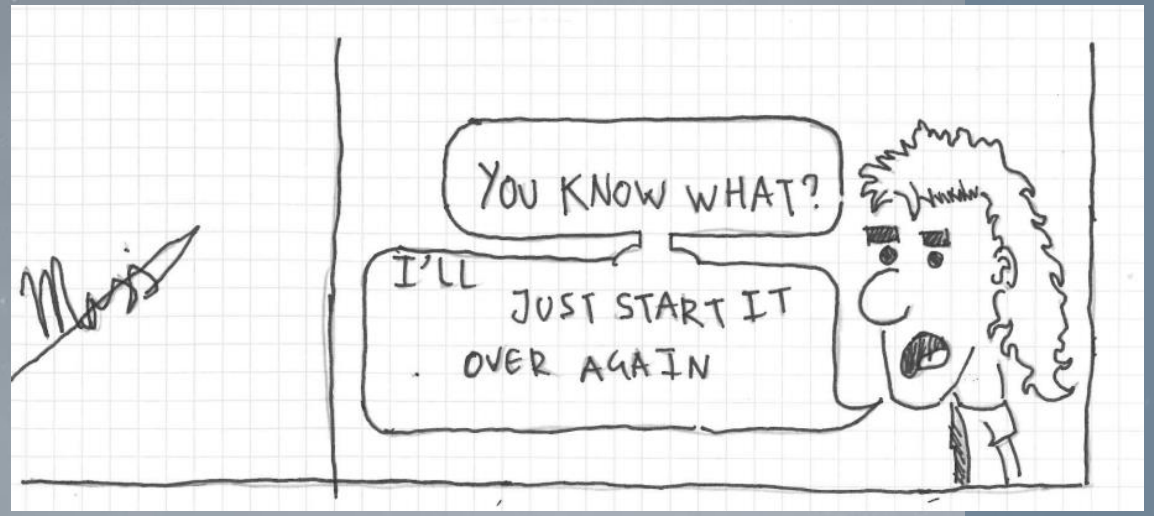


GOOD DATA MANAGEMENT REQUIRES COORDINATION AND COLLABORATION AMONG ALL THE PLAYERS !!!

Scientists need a Data Management Plan







DATA MANAGEMENT PLAN (DMP)

- How will the data be created?
- How will the data be documented?
- Who will access the data?
- Where will the data be stored?
- How will the data be shared?
- How long will the data be preserved?
- Who will back up the data?

A living document updated any time is needed

TOOLS (and guides) TO WRITE A DMP

easy.DMP

DAMAP
A tool for machine actionable DMPs

DMP Tool

DSW

argos

TU WIEN research data management

RDMkit

DMPONLINE

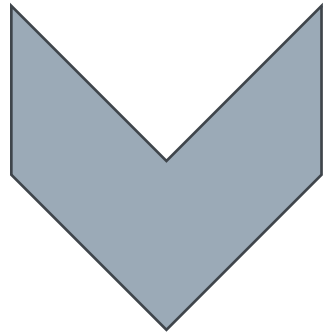
- Each funding agency could require or **recommend a specific** DMP template.
- Your institution could require and **recommend** a DMP template.
- Template could be presented as list of questions in text format or in a **machine-actionable format**.

Brief recap on DMP

DMP is at the heart of fair (FAIR) DATA MANAGEMENT

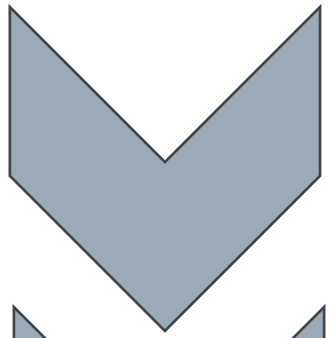
1.1st Generation DMPs (Structured Data):

- In the 1960s, the concept of DMPs began with organizations emphasizing professional training and quality assurance metrics.
- DMPs primarily focused on managing structured data, such as relational databases. (datadiversity.net)



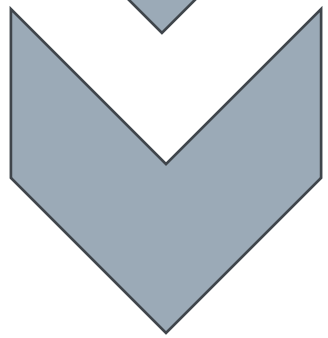
2. 2nd Generation DMPs (Big Data Analytics):

- With the rise of big data, DMPs adapted to handle diverse data types (structured, semi-structured, unstructured). (sparkfish.com)



3. Current Trends in DMPs:

- Metadata and FAIR data principles (Findable, Accessible, Interoperable, Reusable) play a crucial role.
- Ensuring data security, confidentiality and ethical compliance remains essential



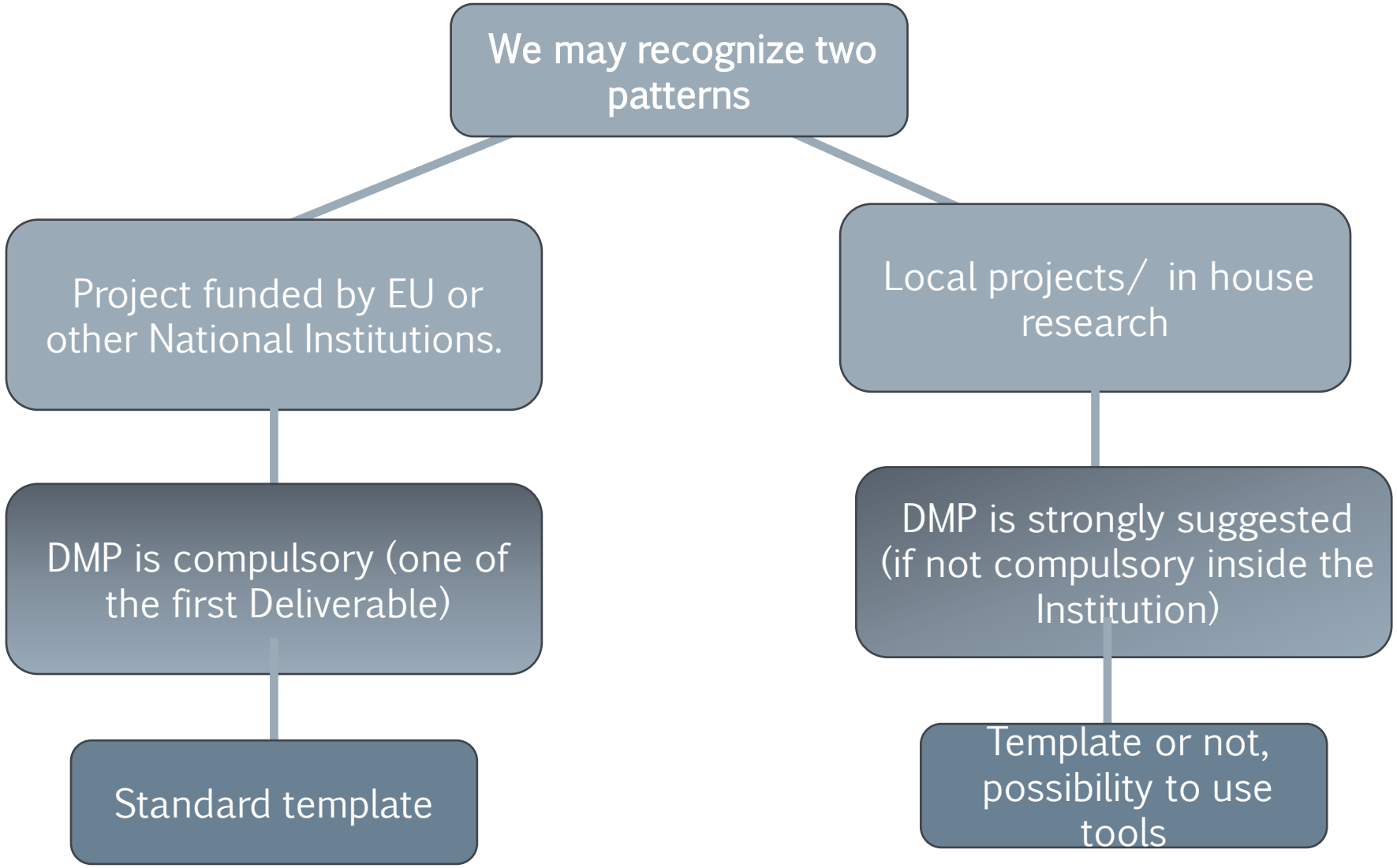
Brief recap on DMP

The **Data Management Plan (DMP)** became **mandatory** for all EU projects, including **Horizon Europe** grants (rdm.mpdl.mpg.de)

Unlike in Horizon 2020, where an opt-out option existed, Horizon Europe no longer allows skipping the DMP requirement. This trend aligns with the goal of making research data **FAIR** (findable, accessible, interoperable, and reusable) across all projects

The huge amount of data produced nowadays in all Sciences requires a deep planning of data management in all projects. Scientific team as well as individual investigator should always start their project with a DMP for sake of (their own) science

Approaches


We may recognize two patterns
Project funded by EU or other National Institutions.
DMP is compulsory (one of the first Deliverable)
Standard template
Local projects/ in house research
DMP is strongly suggested (if not compulsory inside the Institution)
Template or not, possibility to use tools

We said that …

DATA is raw and processed information

Key Characteristics

- Primality – A data point on its own is neutral and meaningless without context.

What do you think of when I say "metadata"?

WHAT IS METADATA?

- Data describing other data
- It provides information about the content, i.e., an image may include metadata describing the picture size, colour depth, image resolution, creation date...
- It describes individual files, single objects, or complete collections

WWW.FLICKR.COM/PHOTOS/CENTRALASIAN/8071729256/

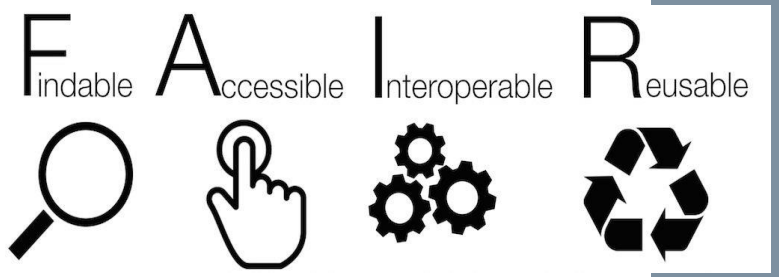
THE IMPORTANCE OF BEING METADATA

- Gives the context, gives meaning to the data
- Ensures that resources will survive and continue to be accessible in the future
- Is searchable, aiding the identification and retrieval of resources
- Helps users in managing, mantaining, and preserving digital collections
- Supports archiving, security, and authentication of data

WHAT MAKES METADATA GOOD?

- Be complete and consistent ! (collect all metadata)
- If exist, use standards, if not exist define ad hoc schema and gives it a URI
- Controlled vocabularies for unambiguous keywords
- Persistent identifiers (DOIs)
- Clearly stated data limitations
- Explanation for appropriate reuse (indicate licences etc)
- Machine readable (interoperability)

(GOOD) METADATA HAVE A CORE ROLE IN FAIR PRINCIPLES



F

F1: **(Meta) data** are assigned globally unique and **persistent identifiers**
F2: Data are described with rich metadata
F3: Metadata clearly and explicitly include the **identifier of the data** they describe
F4: **(Meta)data are registered** or indexed in a searchable resource

A

A1: **(Meta)data** are **retrievable** by their identifier using a standardised communication protocol
A1.1: The protocol is open, free and universally implementable
A1.2: The protocol allows for an authentication and authorisation procedure where necessary
A2: **Metadata should be accessible** even when the data is no longer available

I

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

R

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**

http://www.nature.com/articles/sdata201618

CHALLENGES IN RDM

Ensuring compliance with evolving data policies

Managing large volumes of diverse data types

Encouraging researchers to adopt best practices

Balancing data security with open access principles

DATA POLICY

A documented set of guidelines for ensuring the proper management of the data in an organization

Establishes who is responsible for data under various circumstances, and specifies what procedures should be used to manage it

Regulated data usage, data sharing, and data citations

Requires synergy of executive committee, finance, IT, management, and other data stewards within the organization

Is a flexible document, which can be changed in response to changing needs of the community

INTRODUCTION TO DATA INFRASTRUCTURE

WHAT IS DATA INFRASTRUCTURE?

- The set of technologies and processes for collecting, storing, processing, and managing data
- Includes hardware, software, and policies to handle structured and unstructured data
- Aims to provide reliable, scalable, and secure data management solutions
- A well-structured data infrastructure enables efficient data management, security, and innovation

THE RISE OF DATA INFRASTRUCTURE

1960s-1970s: Emergence of databases (e.g., IBM's IMS, relational databases by E.F. Codd)

1980s-1990s: Data Warehousing and ETL processes

2000s: Big Data explosion and cloud computing

2010s-present: Modern data infrastructure with scalable architectures and AI integration

KEY COMPONENTS OF DATA INFRASTRUCTURE

- Data Storage: Databases, Data Lakes, Cloud Storage
- Data Processing: ETL Pipelines, Batch & Streaming Processing
- Data Governance: Security, Compliance, Data Quality
- Data Access & Analytics: BI Tools, Dashboards, AI/ML Models

DATA INFRASTRUCTURE

- Store large datasets and large data rates from experiments
- Allow reliability by replicating data sources
- Allow accessibility by copying source to several places
- Monitor and check resource usage
- Provide a set of integrated services which are compatible between domains
- Increase interoperability through common standard schemes
- Guarantee secure, broadband, remote access to data

DATA INFRASTRUCTURE

› Data preservation to allow long-term availability of data

› High quality of data and metadata to enable advanced and cross-disciplinary access and enrichment operations

› Economic justification: as the scientific community is operating on increasingly larger datasets and want to preserve the information concerned, the infrastructure provided should have a clear roadmap of technology exchange and backwards compatibility.

› Provide the infrastructure to allow fine-grained access control

WHY DATA INFRASTRUCTURE IS RELEVANT

- Supports data-driven decision-making
- Enables real-time analytics and business intelligence
- Ensures data quality, security, and compliance
- Facilitates scalability and efficiency in modern enterprises

TAKE AT HOME MESSAGE N.1

Data wrangling:

Different formats (often proprietary)

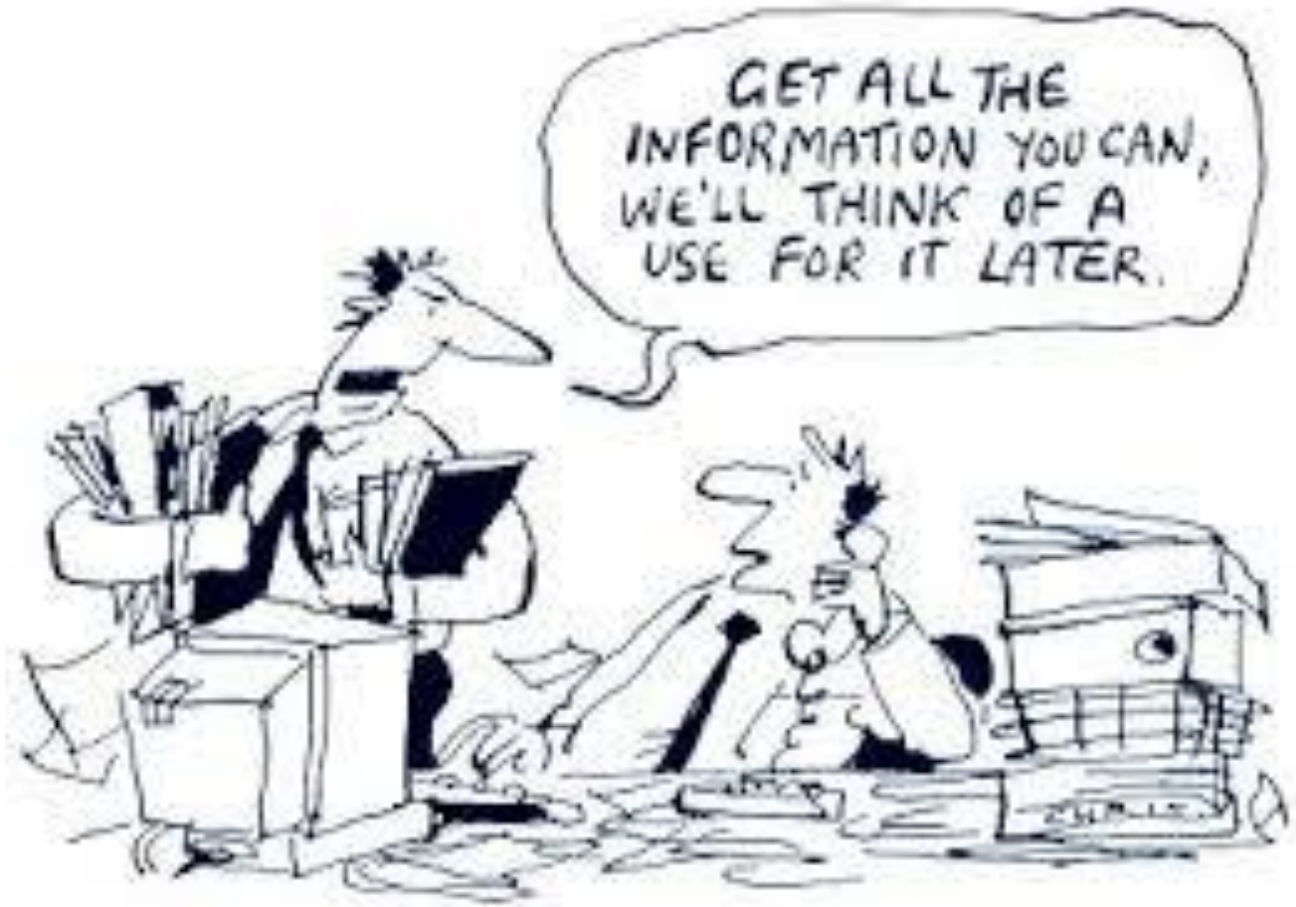
No standardization

Different information

Different units

Incomplete data

Incompatible data

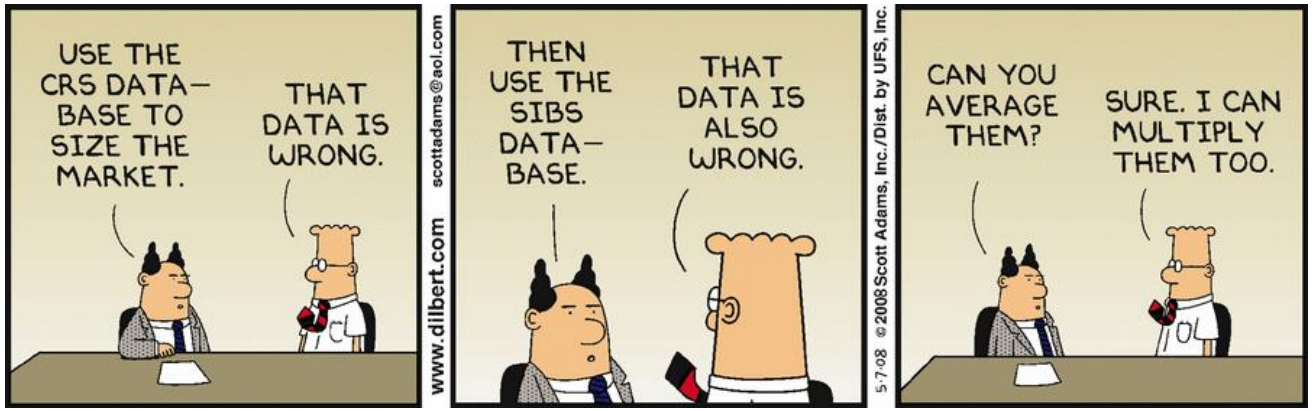


TAKE AT HOME MESSAGE N.2

Lack of agreement among scientists and IT on how to treat data

Manual metadata registration only at publication time

No clear and common (meta)data models



TAKE AT HOME MESSAGE N.3

Inactivity of scientists:

› Old "handmade" programs

› Pen and copybook

› Data intellectual propriety

› Not familiar with technology

› Sharing data by physical drives (external hard disk, usb pen, …)

› *"Metadata registration is a waste of time"* (cit.)