



Ready for BioData Management?



# Introduction to Data Management in Science

Daniel Faria



# Data, Information & Knowledge

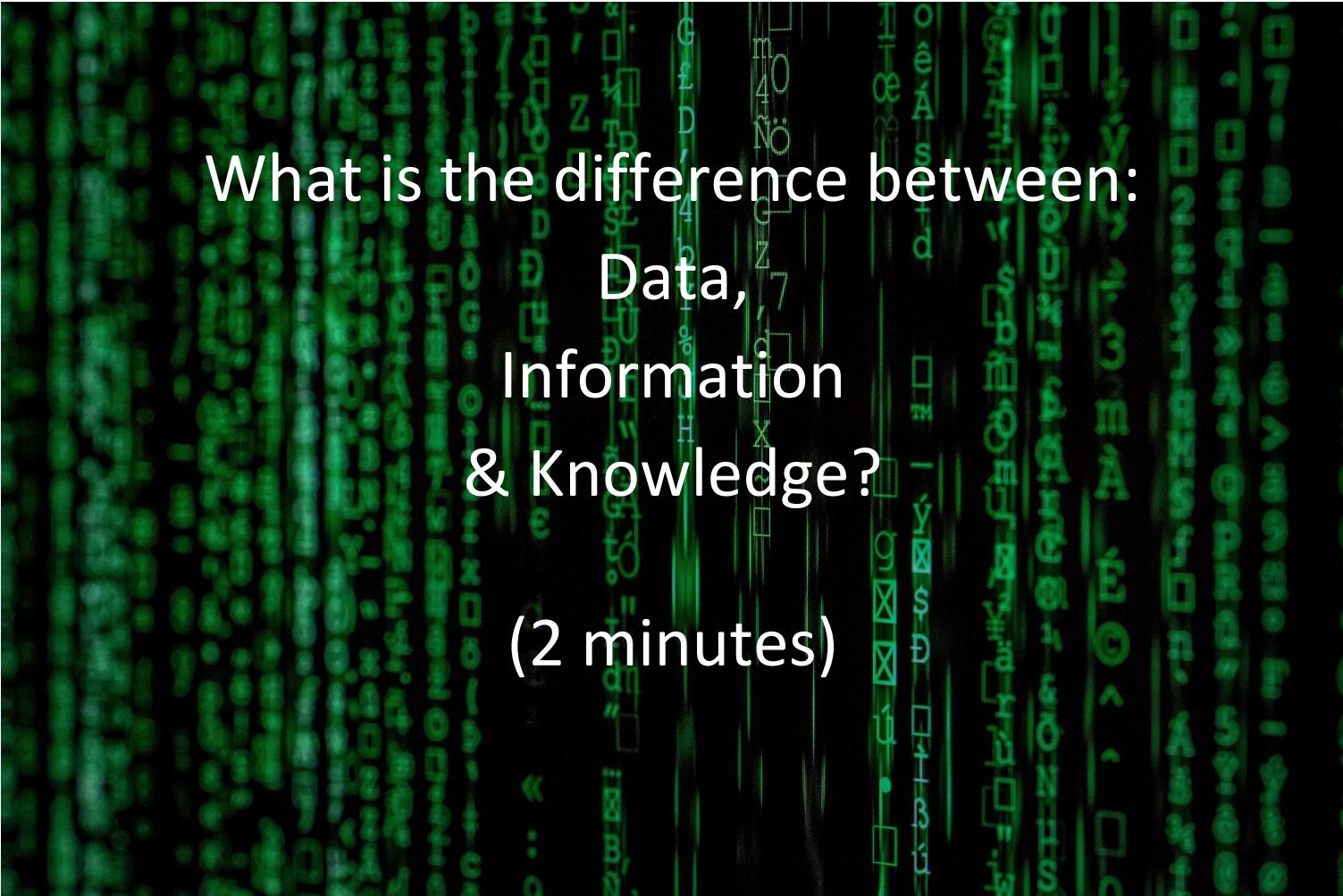
## Learning Outcome 1:

Distinguish between Data, Information and Knowledge

# Introduction

- The main goal of science is to discover knowledge
- This typically requires gathering / generating data
- Data, information, and knowledge are concepts that are often used interchangeably
- In this section, we will try to dissect their differences which heavily impact the dissemination of scientific outputs

# Group Discussion



What is the difference between:  
Data,  
Information  
& Knowledge?  
(2 minutes)

By Markus Spiske [temporausch.com](https://temporausch.com) from Pexels

# Data, Information & Knowledge

- Datum: an atomic fact or piece of “information”
  - Gender: male
  - Weight: 80 kg
  - Height: 1.75 m
- Information: data + context (metadata)
  - Object: Assessing obesity
  - Location: Portugal
  - Date: 12-02-2020
- Knowledge: information + (actionable) understanding
  - The average Portuguese is slightly overweight



By Marta Longas from Pexels

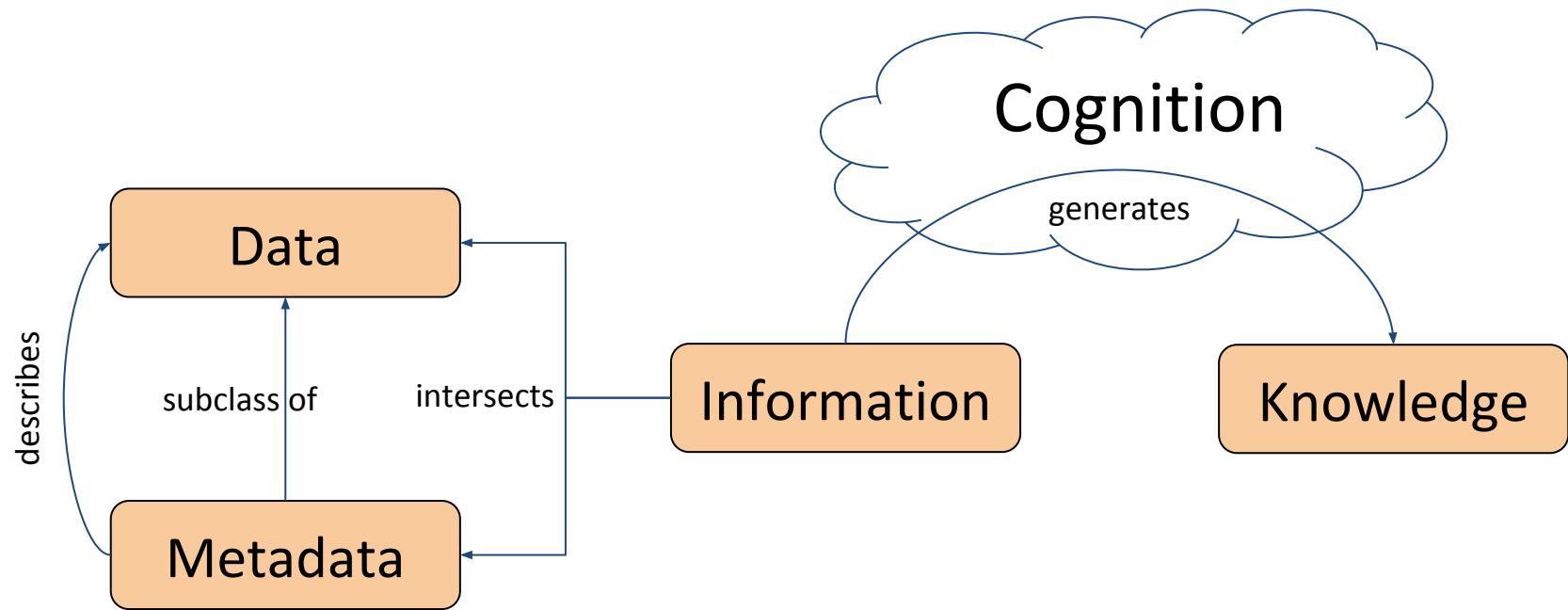
# What is Metadata?

- Metadata = data about data
- Metadata = context
  - Who produced the data?
  - When was the data produced?
  - What is the data about?
  - Why was the data produced?
  - How can the data be used?  
(license)

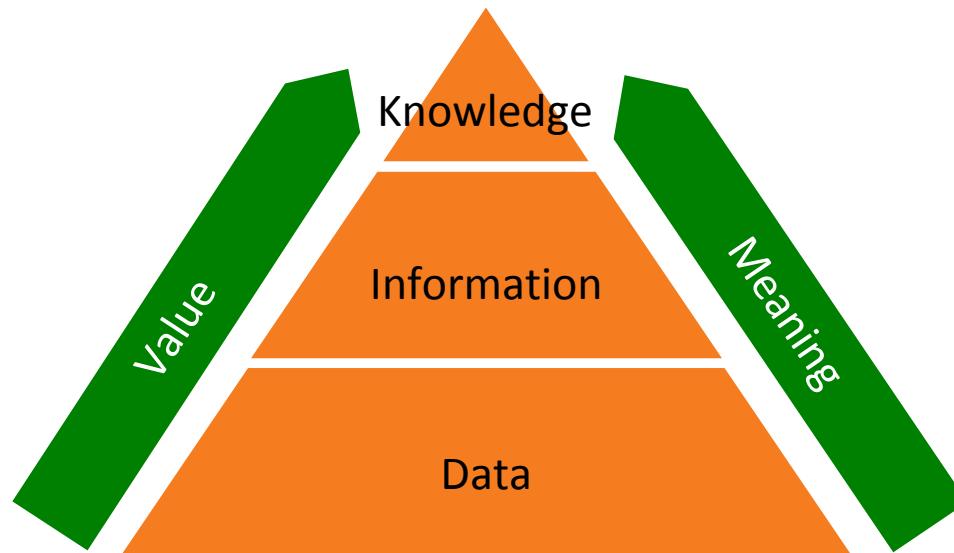


By Dr. Marcus Gossler - Own work, CC BY-SA 3.0

# Data, Information & Knowledge



# Data, Information & Knowledge





# The Knowledge Discovery Problem

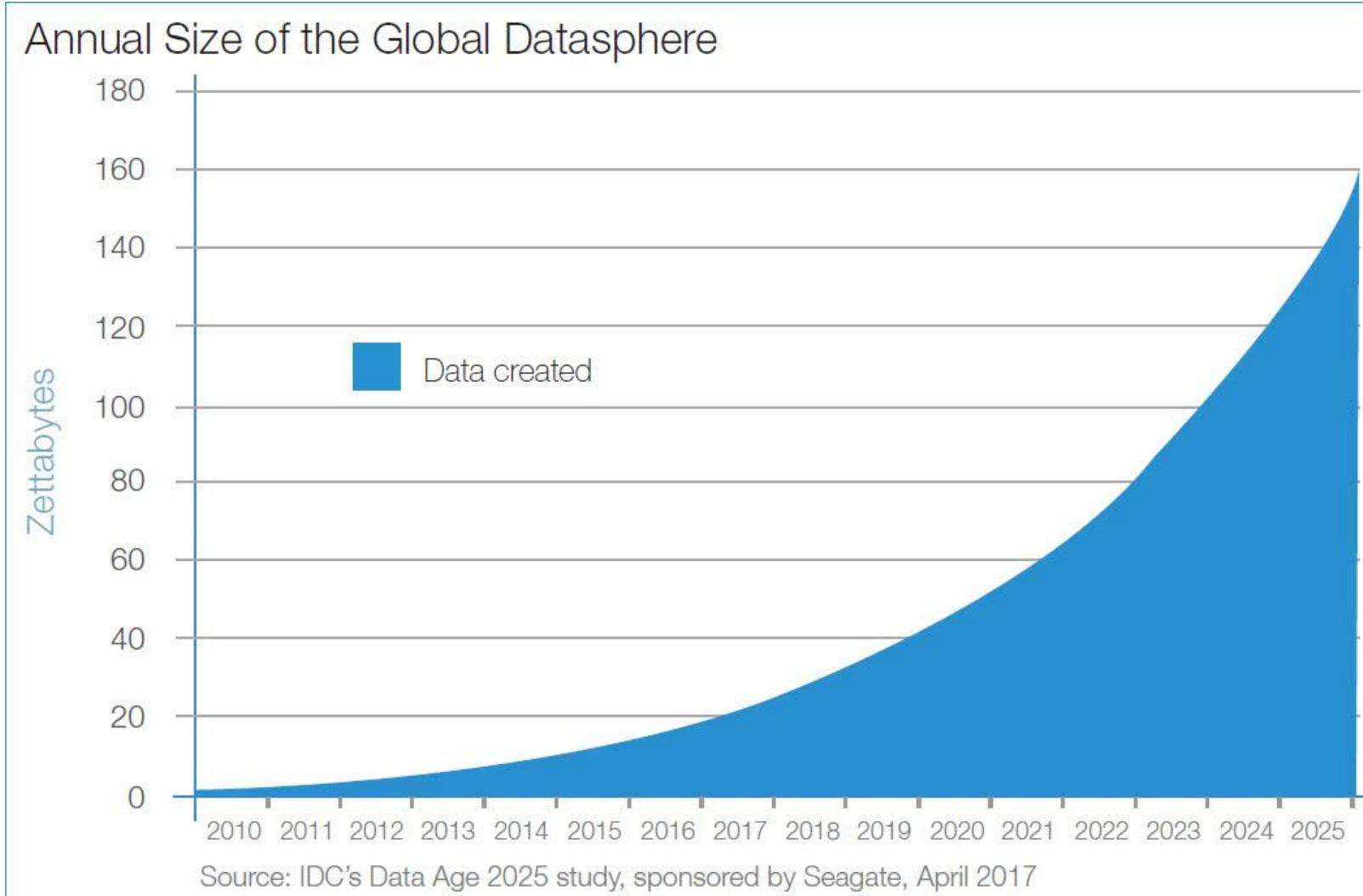
## Learning Outcome 2:

Identify the solutions to the three components  
of the knowledge discovery problem

# Introduction

- Scientists generate data to discover knowledge, but there is often additional value in the data (reuse)
- Scientists are assessed for sharing the knowledge, while data sharing takes on a supporting role
- To reuse research data, we often must:
  - Read paper wherein it was described
  - Extract the metadata needed for interpreting it
  - Figure out if the data is relevant, accessible and usable
- In this section, we will debate how this is not scalable and creates a bottleneck, and what can be done to address that

# The Knowledge Discovery Problem



# The Knowledge Discovery Problem

## Findability:

- More data ⇒ harder search
- Things can get lost amid a sea of things
- If it is not findable, it might as well not exist



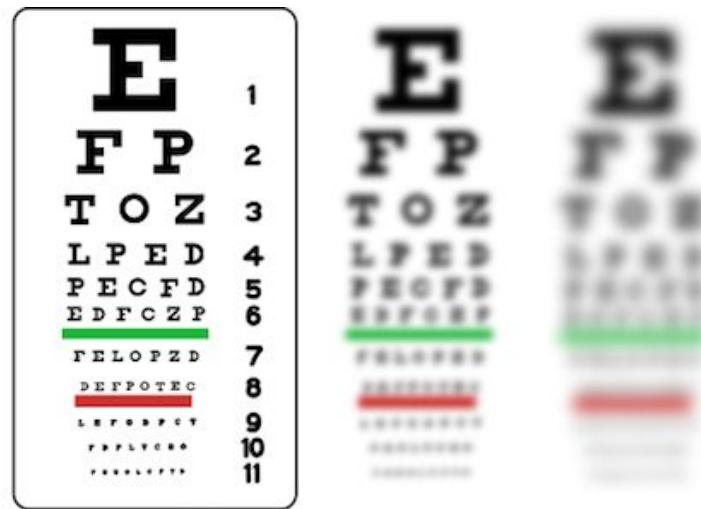
By Martin Handford, retrieved from:

<https://exploringyourmind.com/how-does-our-brain-find-waldo/>

# The Knowledge Discovery Problem

## Interpretability:

- More data  $\Rightarrow$  more costly to interpret
- We become myopic by necessity—can't afford the time to read the fine-print (i.e. research papers)
- If we cannot interpret it readily, then it is nearly useless



By Daniel P. B. Smith, CC BY-SA 3.0

# The Knowledge Discovery Problem

## Interoperability:

- More data & specialization  
⇒ vocabulary and viewpoint divergence
- Use of local dialects leads to sundered data and knowledge
- If we don't find common ground, we cannot integrate data from related domains



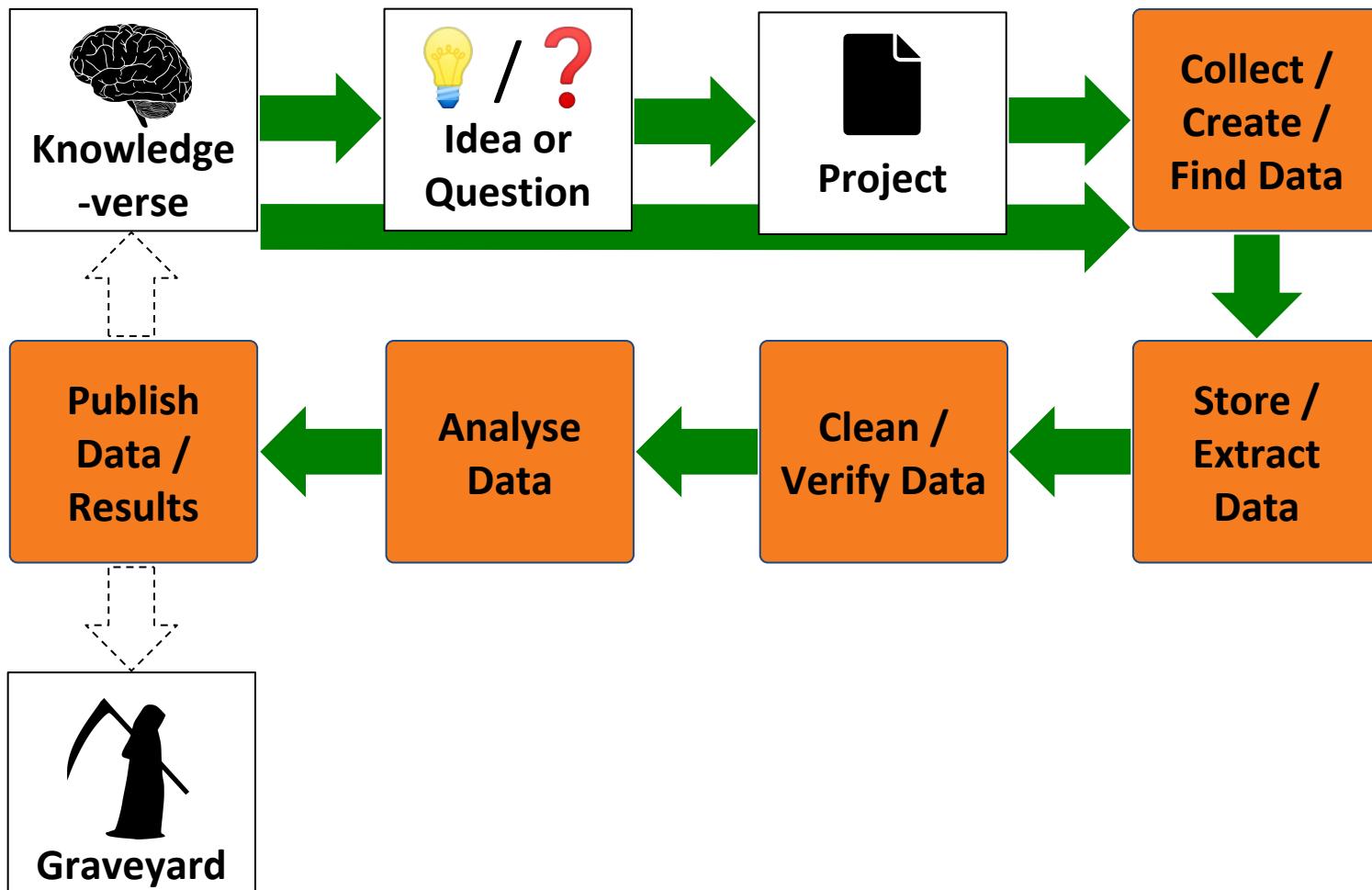
By Abel Grimmer, retrieved from:  
<http://cbcnews.net/cbcnews/the-tower-of-babel/>

# The Knowledge Discovery Problem

## Wrap-Up:

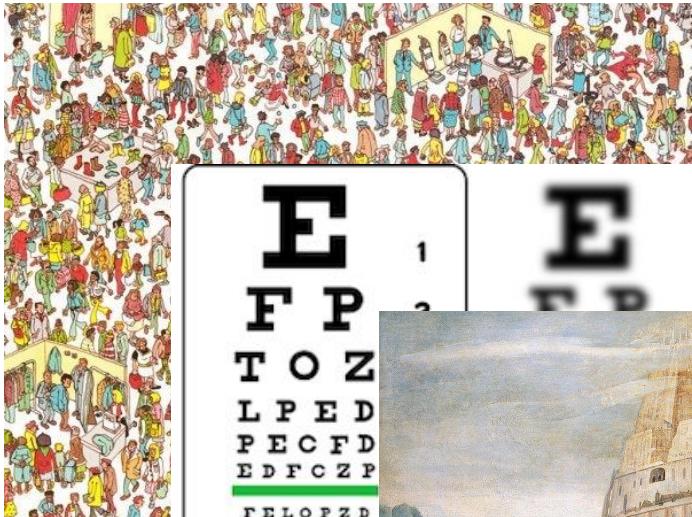
- Publishing data only in scientific papers is not enough
  - Papers are not efficient vehicles for knowledge transfer!!!
- If we want our data to effectively contribute to the knowledge-verse:
  - We must publish it in a form that is:
    - Findable
    - Interpretable
    - Interoperable

# The Data Lifecycle



# Group Discussion

# How to make data:



# Findable?



# Interpretable?



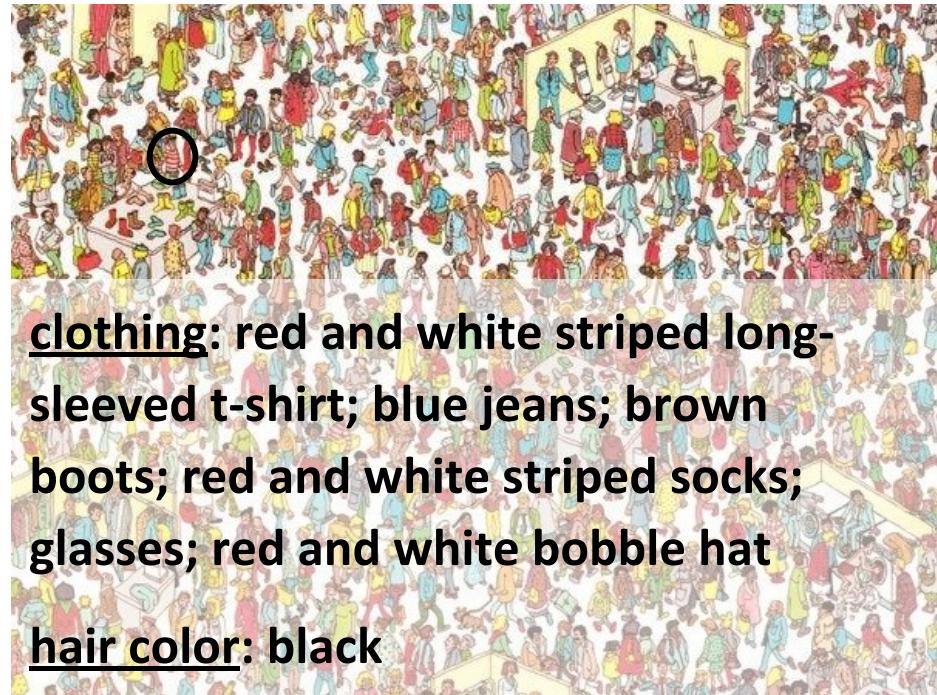
# Interoperable?

**(5 minutes)**

# The Knowledge Discovery Solution

## Findability:

- Describe data with precise metadata useful for searching
- Use a common (structured) controlled vocabulary for metadata fields and values
- Put data in a repository that:
  - Uses persistent unique identifiers
  - Indexes metadata and allows searches



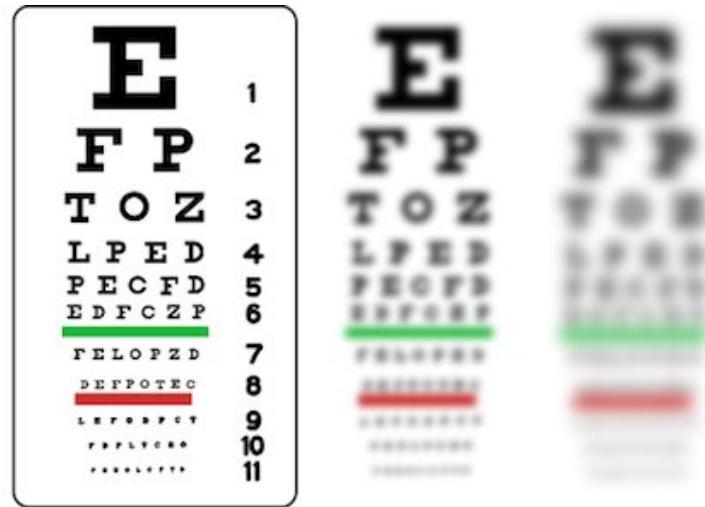
By Martin Handford, retrieved from:

<https://exploringyourmind.com/how-does-our-brain-find-waldo/>

# The Knowledge Discovery Solution

## Interpretability:

- Describe data with sufficient metadata for interpreting it and understanding the experimental context—each dataset should be fully self-contained
- Use a common (structured) controlled vocabulary for metadata fields and values



By Daniel P. B. Smith, CC BY-SA 3.0  
<https://en.wikipedia.org/wiki/File:Snellen-myopia.png>

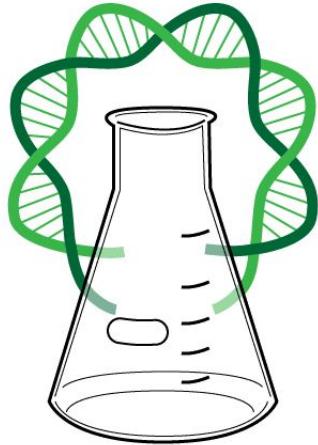
# The Knowledge Discovery Solution

## Interoperability:

- Use a common (structured) controlled vocabulary for metadata fields and values
- Include cross-references to external data objects whenever suitable (e.g. NCBI taxon id)

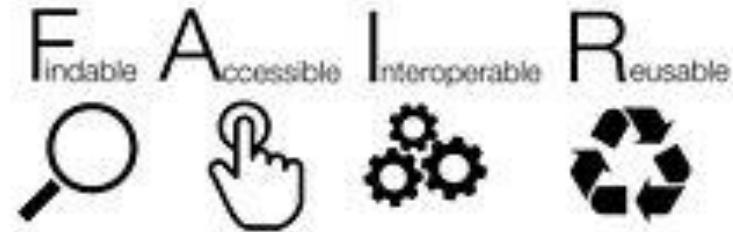


By Abel Grimmer, retrieved from:  
<http://cbcnews.net/cbcnews/the-tower-of-babel/>



**open science**

By Greg Emmerich, CC BY-SA 3.0



By SangyaPundir - Own work, CC BY-SA 4.0

# Open Science & FAIR Principles

## Learning Outcome 3:

Recognize the demands of science funders  
and debate their pros and cons

# Introduction

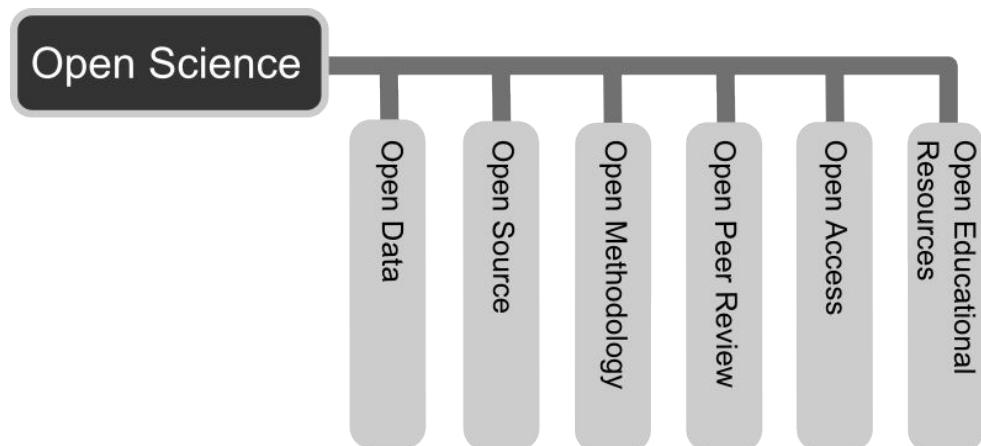
- The need to improve scientific dissemination has been recognized by research communities and publishers
- Leading to initiatives such as Open Science and the FAIR principles
- Funders recognized and are endorsing these initiatives (H2020 projects now require FAIR compliance)
- In this section, we will analyse these initiatives and the demands of funders and illustrate why complying with them is ultimately to the advantage of all



# What is Open Science?

## Goals:

- Scientific research and its dissemination accessible to all levels of society
  - publications
  - data
  - physical samples
  - software
  - ...
- Transparent and accessible knowledge shared and developed through collaborative networks



By Andreas E. Neuhold, CC BY 3.0

# What is Open Science?

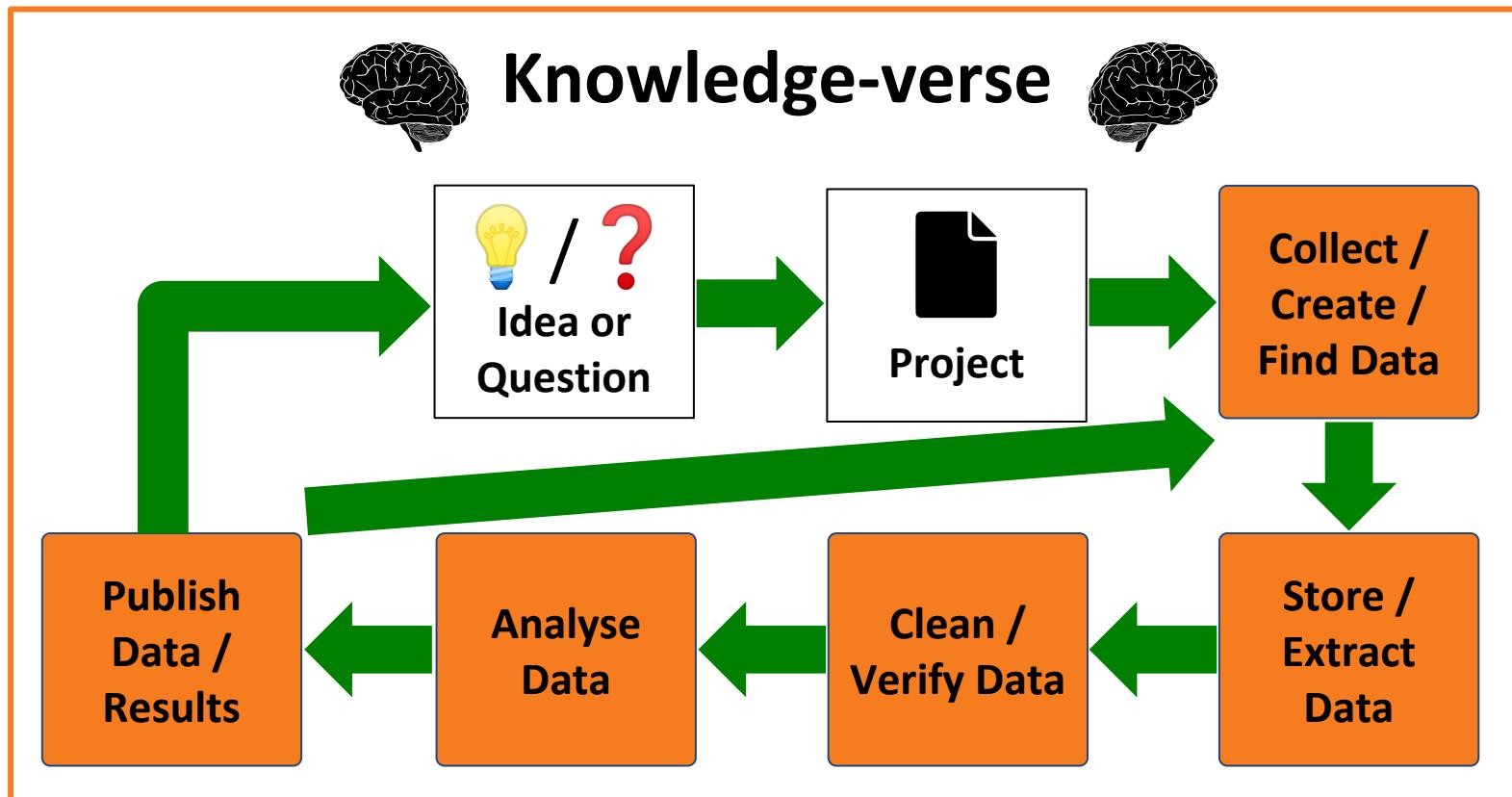
## Layers:

- **Open Access:** research outputs distributed online, free of cost or access barriers
- **Open Research:** data, result and methodology clearly documented and freely available online
- **Open-Notebook Science:** primary record of a research project publicly available online as it is recorded—no insider information



# What Is Open Science?

Everything documented & freely available!



# What are the FAIR Data Principles?

A set of four principles detailed in fifteen guidelines, that establish what Open Research should aim for.

**Findability** – (Meta)data should be easy to find for both humans and computers

**Accessibility** – (Meta)data should have a defined access protocol with authentication and authorization rules

**Interoperability** – (Meta)data should be integratable with other similar datasets and interpretable by applications or workflows for analysis, storage, and processing

**Reusability** – (Meta)data should be well described so that it can be interpreted and reused

# The FAIR Data Principles

II

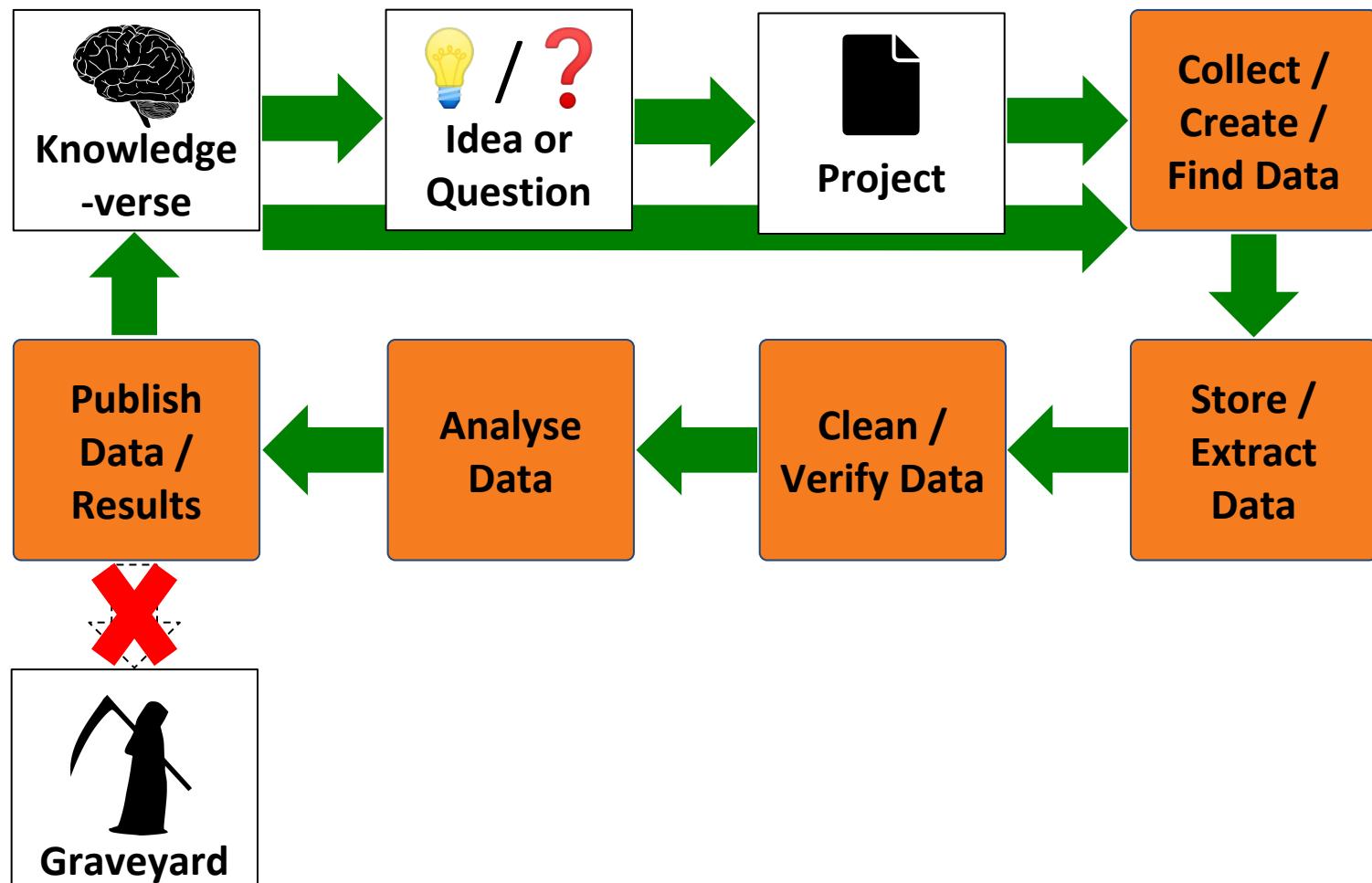
## The Knowledge Discovery Solution

**Wait, that was FII, Not FAIR...**

- Reusability is the end-goal, not the problem—it is contingent on Interpretability and Interoperability.
- Accessibility concerns data repositories, not really researchers, and it is already well addressed. As long as you publish your data in a well-established repository and define an authorization policy (when applicable, such as for sensitive data) you are well off.

# What Are The FAIR Data Principles?

Goal: Enable Data & Knowledge Discovery



# Group Discussion



To be or not to be  
Open & FAIR?  
(5 minutes)

# Pros & Cons

## Pros:

- Facilitates knowledge discovery
- Promotes reproducibility / impedes fake science
- Enables networking
- Helps demystify science for the general public

## Cons:

- Care with sensitive data and with knowledge that has dangerous misuse potential
- Harder to make money off of your research
- Harder to stay ahead of your competitors

# Pros & Cons

## FAIR & Open Data Payoff Matrix:

	Me	Others
Private data		
Open data		
FAIR data	➡	

# FAQ

- **Can I receive credit for publishing data?**
  - This is not yet well established, but we are amidst a shift towards crediting data publishers as much as paper publishers.
- **Can't someone publish a paper ahead of me if I release my data?**
  - If someone can write a paper using your data ahead of you that supersedes yours, shame on you. If it does happen, you at least get credit for the use of your data, and will likely still be allowed to publish your paper as the original author of the data.
- **What if someone uses my data without giving me credit?**
  - The same can happen with paper publication. Reviewers and editors are expected to police this. Authors that do so can be red flagged.

# To Be or Not to Be Open & FAIR???

**It Helps Science!**

- Enables others to apply your knowledge in contexts beyond your foresight
- Enables others to reuse your data to make new research

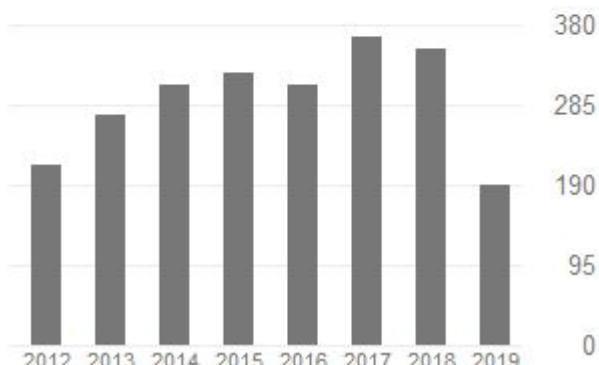


# To Be or Not to Be Open & FAIR???

## It Helps You!

- It is easier to find and reuse your own data
- It is easier to write and submit a research paper
- If others apply or reuse your research, you get more citations (citing or crediting datasets is becoming common practice)

Cited by	All	VIEW ALL
	All	Since 2014
Citations	2724	1856
h-index	21	20
i10-index	29	27



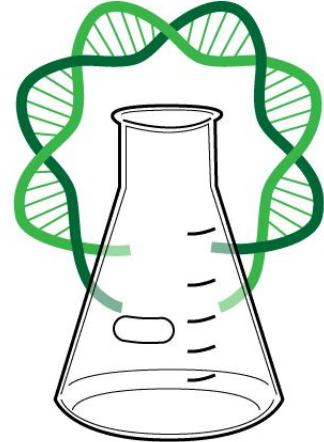
<http://scholar.google.com>

# To Be or Not to Be Open & FAIR???

You'll Need It To Get Funded!

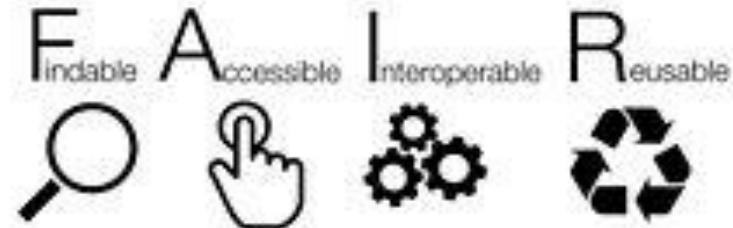
- Soon it will be impossible to get public funding in Europe without adherence to Open Science and FAIR
- FAIR compliance is starting to be verified
- A good track record will contribute to project approval





**open science**

By Greg Emmerich, CC BY-SA 3.0



By SangyaPundir - Own work, CC BY-SA 4.0

# Open Science & FAIR Principles (Again!)

## Learning Outcome 4:

Comply with the demands of science funders

# Introduction

- We've seen that adherence to Open Science and compliance with the FAIR principles are being increasingly demanded by funding agencies
- And debated the merits and demerits of compliance
- In this section, we will dissect what must be done in practice to comply with these demands



# How to be Open & FAIR?

## Step 1 – Do Your Homework

- Is there a default public database or data repository for your research domain?
  - Does it have a metadata schema?
- Are there community metadata standards for your research domain?
- Are there ontologies that cover your research domain?
- Are there default data (open) file formats in your research domain?



# How to be Open & FAIR?

## Step 1b – Do More Homework

- Is there are no standards in one or more of the previous categories for your domain, you have three alternatives:
  - Use a generic option (not valid for ontologies, and not very FAIR)
  - Use / adapt the closest option
  - Rally your community and develop / agree on a standard



# How to be Open & FAIR?

## Step 2 – Do Your Work-Work

- Organize, Document & Annotate:
  - Your code / scripts / workflows
  - Your protocols
  - Your data & metadata
  - According to the applicable guidelines / standards or the repository where you're depositing your data / materials
  - Using domain ontologies, recommended file formats
  - Cross-referencing all relevant information objects



Photo by Kateryna Babaieva from Pexels

# How to be Open & FAIR?

## Step 3 – Deposit

- Deposit your data and materials in an appropriate public repository:
  - Code / scripts / workflows: GitHub, BitBucket
  - Protocols: Zenodo, FAIRDOMHub, Dataverse
  - Data: Domain database, one of the above
  - Metadata: Together with the data (as an accessory file, in the form of the repository)
  - Under a declared usage license
  - With a clear versioning policy



Photo by cottonbro from Pexels

# How to be Open & FAIR?

## Example – Transcriptomic Data

- RNAseq data:
  - Data repository: GenBank / ENA
    - Metadata:
      - Bioproject – describe experiment
      - Biosamples – describe biological material
  - Data file format: FASTQ
- Differential expression data:
  - Metadata standard: MIAME
    - Document experiment, sequencing equipment, biological material, and differential expression software
  - Data file format: MAGE-Tab (metadata) + tabular text (data)

# How to be Open & Fair?

## The Main Hurdles

- The Biomedical Ontology landscape is complex and hard to navigate:
  - There are often overlapping ontologies for a given domain
  - And worse, the same concepts appear in several ontologies, sometimes with the same URI!!!
  - But there are also domains with no (suitable) ontology
- Metadata standards exist only for a few domains, and not all specify a data format for publication
- Generic data repositories (e.g. FAIRDOMHub, Zenodo, Dataverse) have rigid data models that are not compatible with all domains / standards

# How to be Open & FAIR?

**That sounds like a lot of work!**

- It is, especially if you only do it at the time of publication:
  - Have to trace all the data—risk of data loss
  - Have to trace all the details about the experiment—risk of metadata loss
  - It is a lot of boring work to do at once—inertia and rush lead to poor documentation

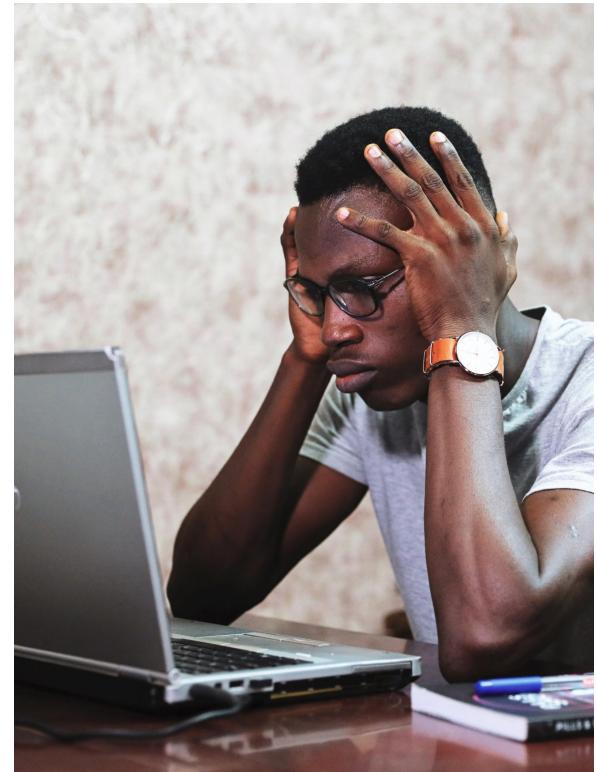
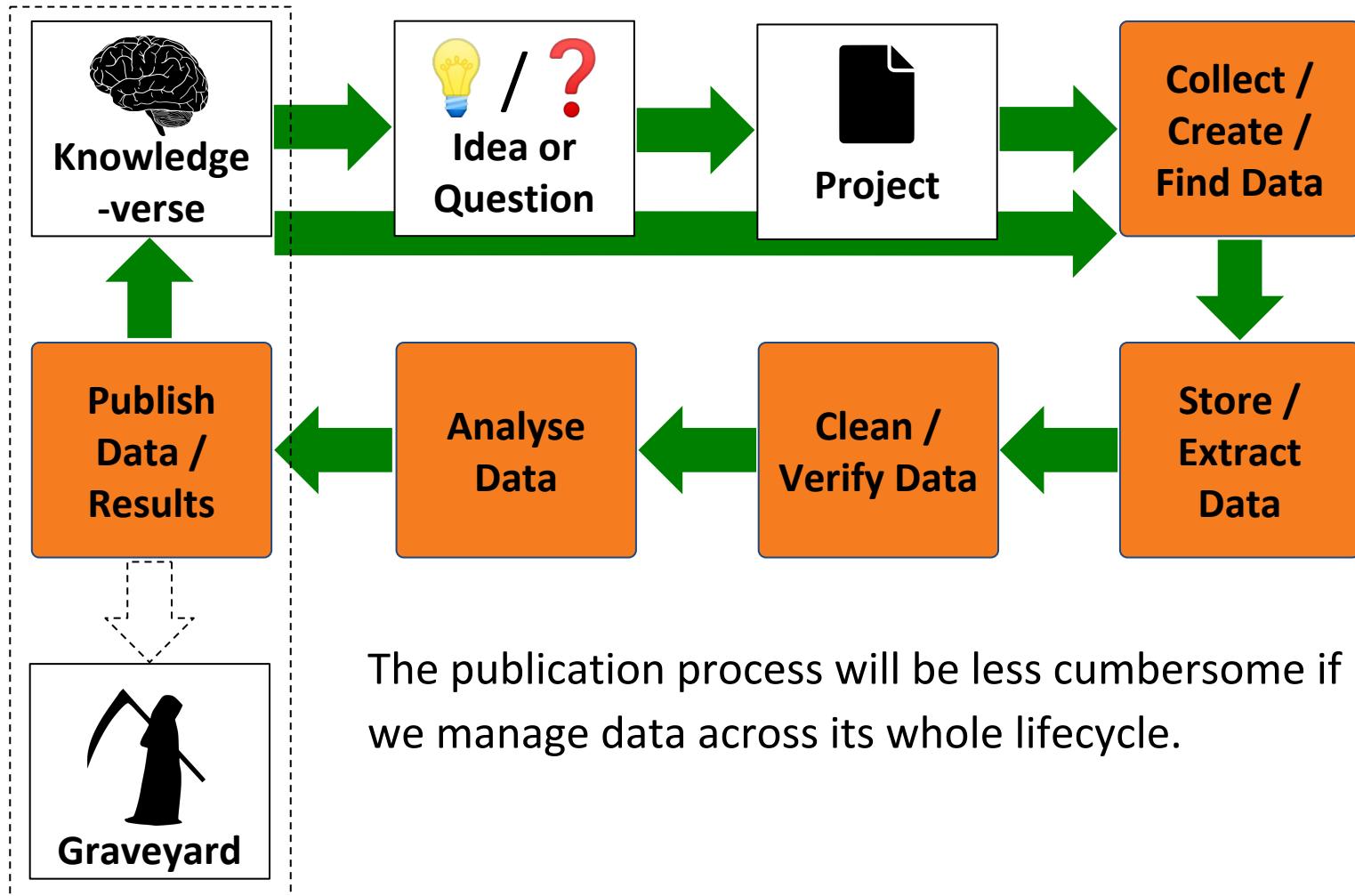


Photo by Oladimeji Ajegbile from Pexels

# How to be Open & FAIR?



The publication process will be less cumbersome if we manage data across its whole lifecycle.



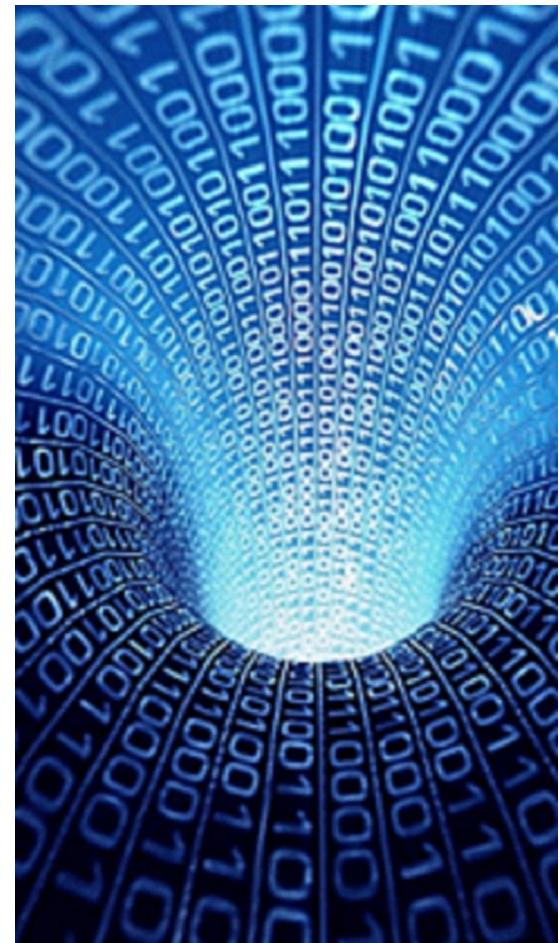
# Data Management

## Learning Outcome 5:

Recognize the supportive role of data management in science

# Introduction

- Data management is a research domain in each own right
- Devoted to topics such as: Data Architecture, Data Modeling, Data Storage & Maintenance, Data Security, Data Integration, Metadata, Data Quality
- Researchers needn't be data management experts
- But just like driving or using a computer, basic knowledge of data management is invaluable for a life in research



By OnePoint Services, CC BY 2.0

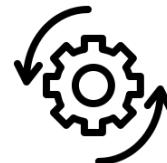
# Why Should I Care About Data Management?

Improve research:



By LAFS,  
CC BY 2.0

**Effectiveness** –  
obtain more/better  
results



By Youmena,  
CC BY 2.0

**Efficiency** – improve  
productivity and  
cost-efficiency



By ROZMOWA,  
CC BY 2.0

**Security** – reduce  
data loss / control  
access to data



By Nithinan Tatah,  
CC BY 2.0

**Impact** – facilitate  
dissemination and  
knowledge discovery

# Data Management Commandments

- Thou shalt make a Data Management Plan for thy research project, even if it isn't funded by a grant
- Thou shalt allocate some time after each day experimenting to document everything, preferably in a digital platform (e.g. electronic lab notebook, local shared repository)
  - Thou shalt document the documentation process
  - Thou shalt use version control (e.g. git)
  - Thou shalt use controlled vocabularies (public or your own, documented)
- For every data file (or collection thereof) thou shalt create a metadata file

# Take Home Messages

## Do the Best You Can!

- FAIRness is a spectrum, and halfway there is a step forward

## Reach Out For Help!

- Data stewards and data management experts can provide guidance

## Things Will Get Easier!

- There are people working towards more user-friendly data management solutions—they need feedback on what can be improved