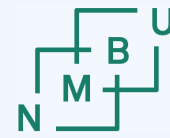# DAT390
Data science seminar

3   Methodology

3.1   The state of the art, and beyond it

3.2   Research data management

Highlight talk schedule for today

Monday, 9th October 2023

Peer feedback from …

15.15 – 15.19 #1 Sougata Bhattacharya   15.19 – 15.21 #1 Mathilde Haglund

Knowledge graphs for software security assessments and cyber threat intelligence

15.23 – 15.27 #2 Sujan Devkota   15.27 – 15.29 #2 Disha Preetha Kannan

Image super-resolution for sperm detections and prediction of motility and morphology

15.31 – 15.35 #3 Artush Mktrchyan   15.35 – 15.37 #3 Jony Karmakar

Explainable AI readiness of cancer research data

15.39 – 15.43 #4 Isak Vartdal-Gjerde   15.43 – 15.45 #4 Gurubaran Rajeshwaran

Evaluating the performance of video segmentation models for car accident detection

Highlight talk schedule for next week

Monday, 16th October 2023

15.15 – 15.19   #1   Awo Arab   15.19 – 15.21   #1   Sougata Bhattacharya

Genomic prediction of complex traits in wheat using multispectral time-series data

15.23 – 15.27   #2   Simen Holter   15.27 – 15.29   #2   Sujan Devkota

(undeclared topic)

15.31 – 15.35   #3   Julie Overrein   15.35 – 15.37   #3   Artush Mktrchyan

Building yield prediction models with remote sensing and deep learning

15.39 – 15.43   #4   Asim Rasheed   15.43 – 15.45   #4   Halvor Steffenssen

EEM spectroscopy and PARAFAC modelling of water quality in nanofiltration

15.47 – 15.51   #5   Mahrin Tasfe   15.51 – 15.53   #5   Isak Vartdal-Gjerde

Deep learning identification and classification of paddy disease in precision agriculture

15.54 – 15.58   #6   Ulrik Egge Husby   15.58 – 16.00   #6   Petter Bøe Hørtvedt

Exploring the landscape of explainable AI models: An empirical study
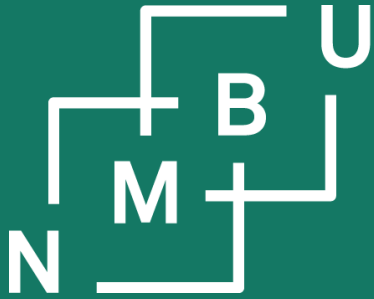
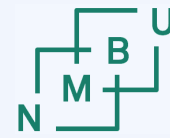**Noregs miljø- og biovitskaplege universitet**

Institutt for datavitskap
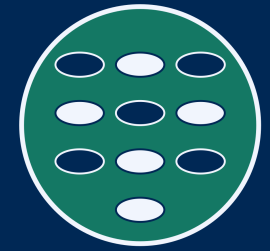
Digitalisering på Ås

3   Methodology

3.1   <u>State of the art & beyond</u>

First impression from the submissions

What looks most problematic?

  – From now **50 students** (initially 55), only **23 submitted** a document.
  – Only about **40% of these** even claim to look at 12 literature references.
  – This means that about 80% of the students will not be able to receive a targeted feedback, since they did not really submit what was asked for.

However, all who submitted something will of course also receive a feedback.

Regarging those who submitted: **What looks best?**

  – General **competency with LaTeX and BibTeX** is good.
  – About 70% of the submissions provide their citations at least in an acceptable format. The remaining ones do not look totally bad either.

For the single most important challenge met by the submissions, let us look at a few examples. (Blanking out any information that could identify the authors.)

First impression from the submissions

Context: In what setting are we looking at this – why even a literature review?

  1) The **DAT390** module is there to help you write a strong **master thesis**.
  2) The "**draft report**" (focus: **literature review**) and "nearly finished report" (focus: methodology) are there to guide you toward the **DAT390** report.

This means: The **literature review** was the first step to writing the **master thesis**.

The evaluation and feedback (and your revision!) must focus on the question:

Does the literature review explain the state of the art, concerning your topic and from your perspective, in a form and at a level suitable for a master thesis?

For the single most important challenge met by the submissions, let us look at a few examples. (Blanking out any information that could identify the authors.)

Some examples – what is the question they answer?

(███████ et al., 2018) is relevant, because it addresses the crucial issue of defending ██████ against ████████████████, which is vital for ensuring the reliability of AI models used in ██████████████. The approach leverages model robustness without the need for retraining or modifications, making it practical for immediate application. This aligns with the ████████ requirement for AI models in the context of ██████████████████.

(███████ et al., 2018) is important, because it introduces a methodology to evaluate the limitations of ████████ methods. It emphasizes the pitfalls of solely relying on ████████ ██████████ for explanations, a critical consideration in █████████████ where precise insights are vital. Additionally, it highlights the inadequacy of certain methods for tasks sensitive to data or model interactions, aligning with the challenges in this domain.

(███████ et al., 2021) is relevant, because it assesses AI system ██████████████████████ ██ detection, highlighting concerns about reliance on ██████████████████. This issue extends to broader ██████████████ contexts, emphasizing the need for ██████████████ The article underscores the necessity of ██████████████ for trustworthy ████████ models, aligning with the importance of ██████████ in ████████ research data analysis.

(████████, 2023) is crucial for my research because it improves AI models, which are essential for reliable ██████████████ analysis. Its impressive performance in ██████████ ████████ aligns perfectly with my focus on ██████████████ for ██████████████. Furthermore,

1

7

Some examples – what is the question they answer?

(███████ et al., 2018) is relevant, because it addresses the crucial issue of defending █████ against ██████████████, which is vital for ensuring the reliability of AI models used in ████████████. The approach leverages model robustness without the need for retraining or modifications, making it practical for immediate application. This aligns with the ██████ requirement for AI models in the context of ████████████████.

(██████ et al., 2018) is important, because it introduces a methodology to evaluate the limitations of ███████ methods. It emphasizes the pitfalls of solely relying on ████████ ████████ for explanations, a critical consideration in █████████ where precise insights are vital. Additionally, it highlights the inadequacy of certain methods for tasks sensitive to data or model interactions, aligning with the challenges in this domain.

(██████ et al., 2021) is relevant, because it assesses AI system ████████████████ ██ detection, highlighting concerns about reliance on ███████████. This issue extends to broader █████████████ contexts, emphasizing the need for █████. The article underscores the necessity of ████████████ for trustworthy ██████ models, aligning with the importance of ████████ in ██████ research data analysis.

(████████, 2023) is crucial for my research because it improves AI models, which are essential for reliable ██████████ analysis. Its impressive performance in ████████ ██████ aligns perfectly with my focus on ██████████ for ████████████. Furthermore,

1

Some examples – <u>what is the question they answer</u>?

(████ et al., 2018) is relevant, because it addresses the crucial issue of defending ████ against ████████, which is vital for ensuring the reliability of AI models used in ████████. The approach leverages model robustness without the need for retraining or modifications, making it practical for immediate application. This aligns with the ████...

(███...

the limita...

sights are...

sensitive t...

<u>What is the question</u> that these statements are answering?

"Xxx is relevant, because …"

"Xxx is important, because …"

"Xxx is crucial for my research, because …"

██ detecti... ...issue extends to broader ████████ contexts, emphasizing the need for ████████ The article underscores the necessity of ████████ for trustworthy ████ models, aligning with the importance of ████ in ████ research data analysis.

(████, 2023) is crucial for my research because it improves AI models, which are essential for reliable ████ analysis. Its impressive performance in ████ ████ aligns perfectly with my focus on ████ for ████. Furthermore,

1

Some examples – what is the question they answer?

2 Referansevurdering

[1]Jeg valgte denne kilden på grunn av dens relevans for ███████████ ███████████████ Resultatene indikerer betydningen av ████████████ ██████████

[2]Denne kilden presenterer forskning som benytter klassifikasjon og regresjonstrær for ████████████████████ ved hjelp av et datasett fra en studie om ██████ ███████████████████████████ Studien antyder at ████████████████ ███████████████████████████████

[3]Denne kilden utforsker hvordan ulike modeller kan ███████████████ ████████████████████, samtidig som de skiller mellom ulike ████████ ██████████████████████ anses som en vesentlig faktor for å forstå ████████████████

[4]Denne kilden gjennomgår studier fra perioden 2016 til 2020 som benytter seg av maskinlæring for å identifisere nøkkelvariabler i forbindelse ████████████ ████████████████ Videre utforsker studien ulike typer ████████████████████ og antyder at ████████████████ kan være mer tilpasset for ██████████████, mens andre er bedre egnet for ██████████████

Some examples – what is the question they answer?

1 Introduction

(██████████, 2022) is relevant because it provides clear-cut info about ██████████████, a topic I'm likely to focus on during my research.

(██████████, 2022) is relevant to my thesis as the proposed methodologies and findings could provide a valuable foundation and comparison for my research.

(███████ et al., 2020) is relevant because it provides valuable perspectives on ██████ conditions and challenges for ███████ systems. It also provides these matters in a context of data and AI.

(██████████ and ██████████, 2019) is relevant to my master thesis as it explores the use of ██████████ for predicting ███████████████, which aligns with my intention of utilizing ██████████████ to analyze ████████████████████████.

(██████████, 2021) is relevant due the mentioned approach of integrating ████████████ data to create a ████████████████ dataset, an operation I likely will perform for my research.

(████████ et al., 2016) is a relevant reference since it explores the development of an artificial neural network (ANN) model to predict ███████████████, while also considering factors like ████████████████ and █████████████, which aligns with my topic.

Some examples – what is the question they answer?

1   Introduction

(███████, 2022) is relevant because it provides clear-cut info about ████████████, a topic I'm likely to focus on during my research.

(███████, 2022) is relevant to my thesis as the proposed methodologies and findings could provide a valuable foundation and comparison for my research.

(██████ et al., 2020) is relevant because it provides valuable perspectives on ██████ conditions and challenges for ██████ systems. It also provides these matters in a context of data and AI.

(██████ and ████████, 2019) is relevant to my master thesis as it explores the use of ████████ for predicting ██████████████, which aligns with my intention of utilizing ████████████████ to analyze ████████████████████.

(███████, 2021) is relevant due the mentioned approach of integrating ████████████ data to create a ████████████ dataset, an operation I likely will perform for my research.

(██████ et al., 2016) is a relevant reference since it explores the development of an artificial neural network (ANN) model to predict ████████████, while also considering factors like ████████████████ and ██████████, which aligns with my topic.

Some examples – <u>what is the question they answer</u>?

1   Introduction

(░░░░░░░, 2022) is relevant because it provides clear-cut info about ░░░░░░░░, a topic I'm likely to focus on during my research.

(░░░░░░░░░░ ... findings could provid ...

(░░░░ et a ... con-ditions and ... ontext of data and AI

(░░░░░░ a ... s the use of ░░░░░░░ ... f utilizing ░░░░░░░░

(░░░░░░, ... data to create a ░░░░░░░░░░ dataset, an operation I likely will perform for my research.

(░░░░ et al., 2016) is a relevant reference since it explores the development of an artificial neural network (ANN) model to predict ░░░░░░░, while also considering factors like ░░░░░░░░░░ and ░░░░░░░, which aligns with my topic.

> **What is the question** that these statements are answering?
>
> "Xxx is relevant because …"
>
> "Xxx is relevant to my master thesis as …"
>
> "Xxx is a relevant reference since …"

Analysis and discussion

Step 1: What is the question that, implicitly, they and others were answering?

Step 2: How does this differ from review papers from NMBU that we looked at?

- Co-authored by Fadi al Machot (doi:10.1177/1550147716665520)
- Co-authored by Kristian Berland (arXiv:1412.6827 [cond-mat.mtrl-sci])

A review on applications of activity recognition systems with regard to performance and evaluation

Suneth Ranasinghe, Fadi Al Machot and Heinrich C Mayr

Abstract
Activity recognition systems are a large field of research and development, currently with a focus on advanced machine learning algorithms, innovations in the field of hardware architecture, and on decreasing the costs of monitoring while increasing safety. This article concentrates on the applications of activity recognition systems and surveys their state of the art. We categorize such applications into active and assisted living systems for smart homes, healthcare monitoring applications, monitoring and surveillance systems for indoor and outdoor activities, and tele-immersion applications.

van der Waals forces in density functional theory: The vdW-DF method

Kristian Berland,[1] Valentino R. Cooper,[2] Kyuho Lee,[3,4] Elsebeth Schröder,[5] T. Thonhauser,[6] Per Hyldgaard,[5] and Bengt I. Lundqvist[7]

[1] *Centre for Materials Science and Nanotechnology, SMN, University of Oslo, NO-0318 Oslo, Norway*
[2] *Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6114, USA.*
[3] *Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.*
[4] *Department of Chemical and Biomolecular Engineering, University of California, Berkeley, California 94720, USA.*
[5] *Microtechnology and Nanoscience, MC2, Chalmers University of Technology, SE-412 96 Göteborg, Sweden.*
[6] *Department of Physics, Wake Forest University, Winston-Salem, North Carolina 27109, USA.*
[7] *Department of Applied Physics, Chalmers University of Technology, SE-412 96 Göteborg, Sweden.*
(Dated: December 23, 2014)

A density functional theory (DFT) that accounts for van der Waals (vdW) interactions in condensed matter, materials physics, chemistry, and biology is reviewed. The insights that led to the construction of the Rutgers-Chalmers van der Waals Density Functional (vdW-DF) are presented

Analysis and discussion

Step 1: What is the question that, implicitly, they and others were answering?

Step 2: How does this differ from review papers from NMBU that we looked at?

- Co-authored by Fadi al Machot (doi:10.1177/1550147716665520)
- Co-authored by Kristian Berland (arXiv:1412.6827 [cond-mat.mtrl-sci])

Step 3: The aim of **analysing and reviewing the literature** is:

(a) to **understand the state of the art**, as a background to your own work,
(b) to **explain the state of the art** to the reader, also as a background to your own work.

In view of this, what question(s) should we be answering instead?

From the state-of-the-art review to a complete report

From the "nearly finished report" on (deadline 10th November), the work is expected to <u>cover all aspects</u>, not just reviewing the state of the art.

It must in this sense be <u>complete</u>.

What distinguishes a "nearly finished" DAT390 report from the master thesis?
- The **master thesis** needs to report on the finalized research <u>and results</u>.

- The **master thesis** is a comparably long document. (Usually.)

From the state-of-the-art review to a complete report

From the "nearly finished report" on (deadline 10th November), the work is expected to <u>cover all aspects</u>, not just reviewing the state of the art.

It must in this sense be <u>complete</u>.

What distinguishes a "nearly finished" DAT390 report from the master thesis?
- The **master thesis** needs to report on the finalized research <u>and results</u>. The **DAT390 report** is not expected to contain any final results. However, a **feasibility study** is needed – *can we see that it will work?*
- The **master thesis** is a comparably long document. The **DAT390 report** is like a conference paper, up to 12 pages (+ literature), 11pt, A4 paper.
- There is only one difference between the final **DAT390 report** and the **"nearly finished" report:** The DAT390 report determines your character grade in DAT390. The "nearly finished" report is not graded.

From the state-of-the-art review to a complete report

UHR Universitets-
og høgskolerådet

See the UHR's recommended **standardized assessment form** for master theses.

What distinguishes a "nearly finished" DAT390 report from the master thesis?

– The **master thesis** needs to report on the finalized research <u>and results</u>. The **DAT390 report** is not expected to contain any final results. However, a **feasibility study** is needed – *can we see that it will work?*

– The **master thesis** is a comparably long document. The **DAT390 report** is like a conference paper, up to 12 pages (+ literature), 11pt, A4 paper.

– There is only one difference between the final **DAT390 report** and the **"nearly finished" report:** The DAT390 report determines your character grade in DAT390. The "nearly finished" report is not graded.

What are the **criteria for evaluating a master thesis**? (See UHR document.)
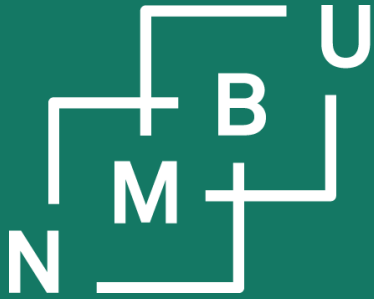
Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap
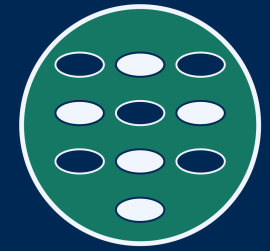
Digitalisering på Ås

3   Methodology

3.1   State of the art & beyond
3.2   **Research data management**

NMBU's guidelines for research data management

Research data management guidelines document

(Approved by NMBU's rector in 2018)

- Principle "**open as standard**"

- Comply with international standards for research data management

- Comply with the **FAIR principles** (acronym for "findable, accessible, interoperable and reusable").

- **Research data/metadata** should be available, searchable, and reusable.

- Furthermore, "interoperable" means that both data and metadata must be **manageable for machines** and that a **consistent vocabulary** is used.

NMBU's guidelines for research data management

Research data management guidelines document

(Approved by NMBU's rector in 2018)

1) Research data must be **stored safely**.

2) Research data should not (only) be stored on a local storage.

3) Follow agreed **good practices**, regulations, guidelines, and the law.

4) Persistent long-term **data preservation** for a minimum of ten years.

5) The research data must be **annotated with metadata**.

6) A **data management plan** (DMP), describing how data will be managed.

7) Open and available (**open-access**) research data.

8) Protect **personal data**.

Platforms for data storage and preservation

Dataverse.NO

Zenodo

Why do we need good practices?

Published data:
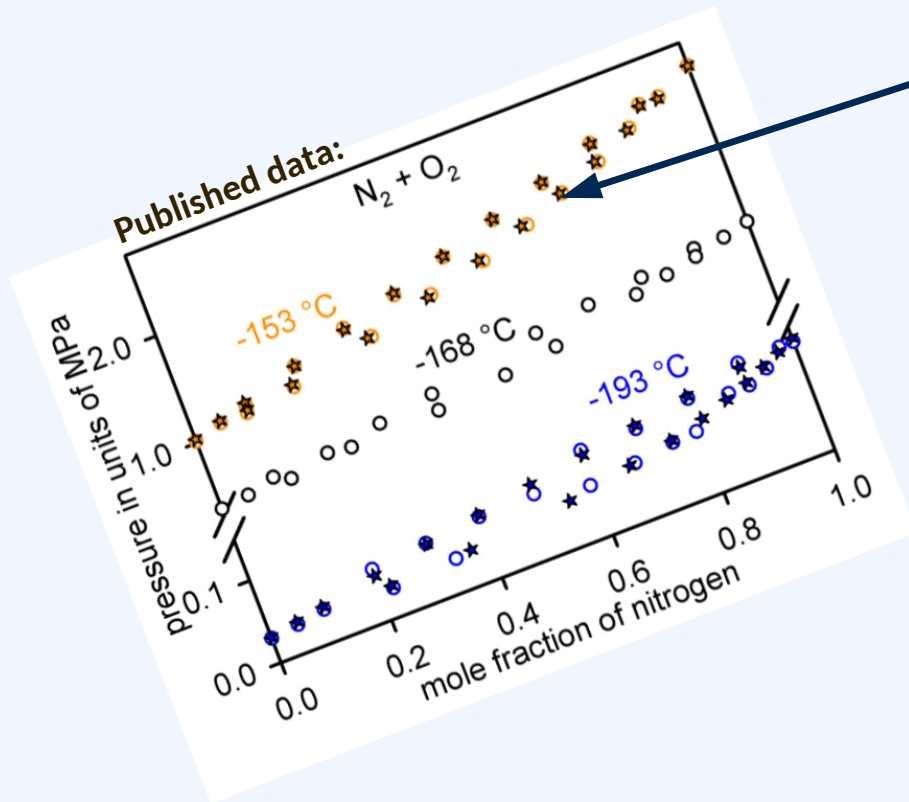
What values did x and p have?

How was the data point obtained?

What is the margin of error, how was the error defined, and what software (or experimental setup) was used?

Good practice in managing research data:

Make all data **findable**, **accessible**, **interoperable**, and **reusable** (FAIR).

FAIR principles[1] in detail

Findability

F1. Globally unique **persistent identifiers (PID)**
F2. **Enriched with metadata**
F3. Data identifier included in metadata
F4. **Registered** in **searchable platform**

Accessibility

A1. **Retrievable from PID** via a standard protocol
A1.1. Open and freely implementable protocol
A1.2. … **authentication/authorization** if necessary
A2. Metadata remain accessible (beyond data)

Interoperability

I1. **Formal language** used for **knowledge representation**
I2. Metadata use **vocabularies** that are themselves FAIR
I3. Semantic web principles, **data can refer to other data**

Reusability

R1. Metadata include a plurality of accurate and relevant attributes
R1.1. Release data and metadata with an accessible **data usage license**
R1.2. Data are annotated with a detailed **provenance description**
R1.3. Relevant **disciplinary and community standards** are fulfilled

[1]M. D. Wilkinson *et al.*, "The FAIR Guiding Principles …," doi:10.1038/sdata.2016.18, **2016**.

FAIR principles[1] in detail

persistent identifier

Findability

F1. Globally unique **persistent identifiers (PID)**
F2. **Enriched with metadata**
F3. Data identifier included in metadata
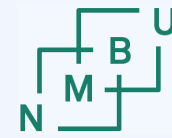F4. **Registered** in **searchable platform**

Accessibility

A1. **Retrievable from PID** via a standard protocol
A1.1. Open and freely implementable protocol
A1.2. … **authentication/authorization** if necessary
A2. Metadata remain accessible (beyond data)

Interoperability

I1. **Formal language** used for **knowledge representation**
I2. Metadata use **vocabularies** that are themselves FAIR
I3. Semantic web principles, **data can refer to other data**

Reusability

R1. Metadata include a plurality of accurate and relevant attributes
R1.1. Release data and metadata with an accessible **data usage license**
R1.2. Data are annotated with a detailed **provenance description**
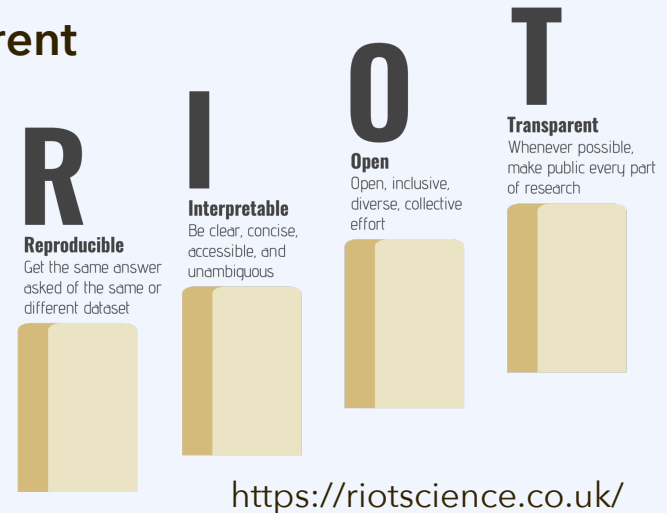R1.3. Relevant **disciplinary and community standards** are fulfilled

[1]M. D. Wilkinson *et al.*, "The FAIR Guiding Principles …," doi:10.1038/sdata.2016.18, **2016**.

RIOT principles

RIOT:[1] Reproducible, interpretable, open, transparent

- Origin: UK Reproducibility Network (UKRN)

- UKRN encouraged foundation of the other reproducibility networks, such as NORRN, the Norwegian Reproducibility Network

- Local "RIOT science clubs" were founded

R Reproducible
Get the same answer asked of the same or different dataset

I Interpretable
Be clear, concise, accessible, and unambiguous

O Open
Open, inclusive, diverse, collective effort

T Transparent
Whenever possible, make public every part of research

https://riotscience.co.uk/

RIOT, FAIR, and CARE principles

RIOT:[1] Reproducible, interpretable, open, transparent

- Origin: UK Reproducibility Network (UKRN)

- UKRN encouraged foundation of the other reproducibility networks, such as NORRN, the Norwegian Reproducibility Network

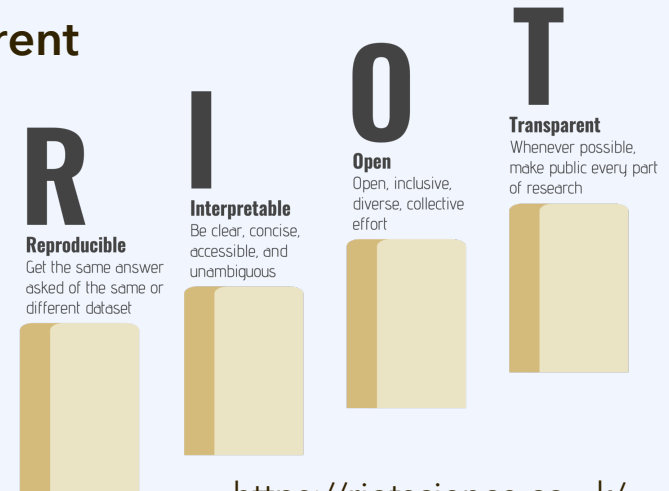- Local "RIOT science clubs" were founded

R Reproducible — Get the same answer asked of the same or different dataset

I Interpretable — Be clear, concise, accessible, and unambiguous

O Open — Open, inclusive, diverse, collective effort

T Transparent — Whenever possible, make public every part of research

https://riotscience.co.uk/

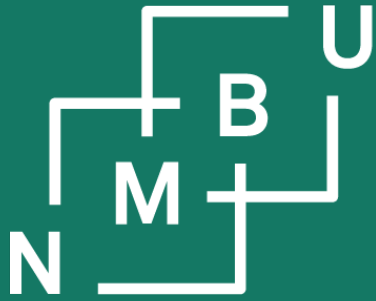**CARE:[2] Collective benefit, authority to control, responsibility, ethics**

- Origin: Global Indigenous Data Alliance

- Uptake supported by the Research Data Alliance

- Orientation: Sovereignty and epistemic justice

Be FAIR — Findable Accessible Interoperable Reusable

and CARE — Collective Benefit, Authority to Control, Responsibility, Ethics
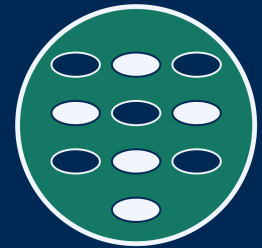
https://www.gida-global.org/care/

[1] E. Ganley *et al.*, *BMC Res. Notes* **15**: 51, doi:10.1186/s13104-022-05932-5, **2022**.
[2] S. Russo Carroll *et al.*, *Sci. Data* **8**: 108, doi:10.1038/s41597-021-00892-0, **2021**.

DAT390

Data science seminar

3   Methodology

3.1   The state of the art, and beyond it

3.2   Research data management