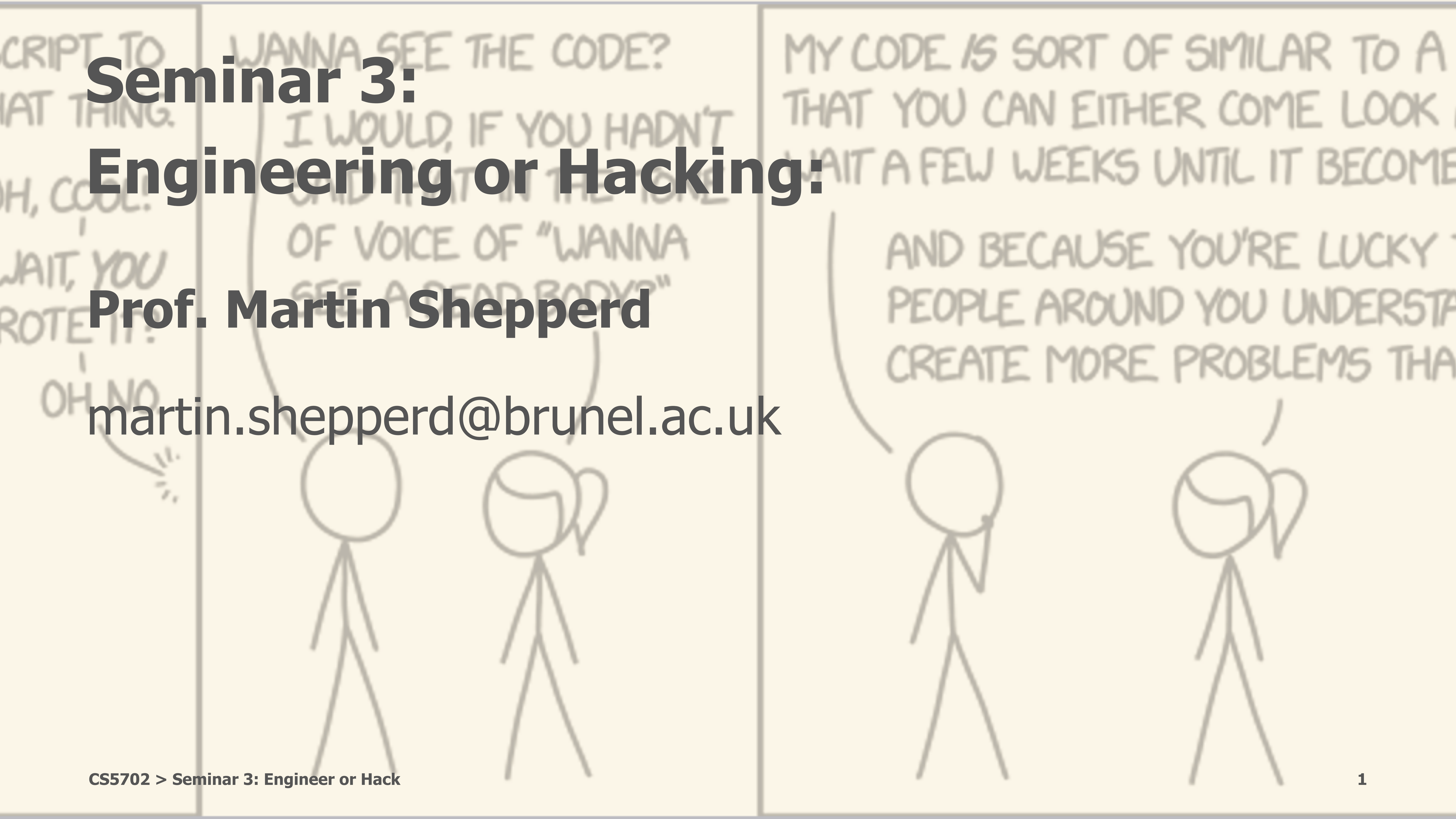
Seminar 3:
Engineering or Hacking:

Prof. Martin Shepperd

martin.shepperd@brunel.ac.uk


CRIPT TO
HAT THING.
OH, COOL!
WAIT, YOU
ROTE IT?
OH NO
WANNA SEE THE CODE?
I WOULD, IF YOU HADN'T SAID THAT IN THE TONE OF VOICE OF "WANNA SEE A DEAD BODY?"
MY CODE IS SORT OF SIMILAR TO A
THAT YOU CAN EITHER COME LOOK
WAIT A FEW WEEKS UNTIL IT BECOME
AND BECAUSE YOU'RE LUCKY
PEOPLE AROUND YOU UNDERSTA
CREATE MORE PROBLEMS THA

PollEv.com/mshepperd

Virtual seminar protocol

1. I will start at 1505 prompt; please be ready
2. Please mute your microphone
3. If you have a question can you use the chat facility or ...
4. ... or ask during in a question gap
5. ... or interrupt (it's fine)
6. Thumbs-up questions that are important to you
7. Nathan and Tasin are monitoring and answering chat - thanks guys!
8. Be aware, the seminar will be recorded

Seminars and Labs

- I will lead the seminar
- Maybe easier to watch (and answer the Pollev.com questions)
- **Then** do the exercises in the gaps or lab afterwards
- I will leave the seminar and then join the lab meeting
- There are answers in the Worksheets

Seminar + Lab Agenda

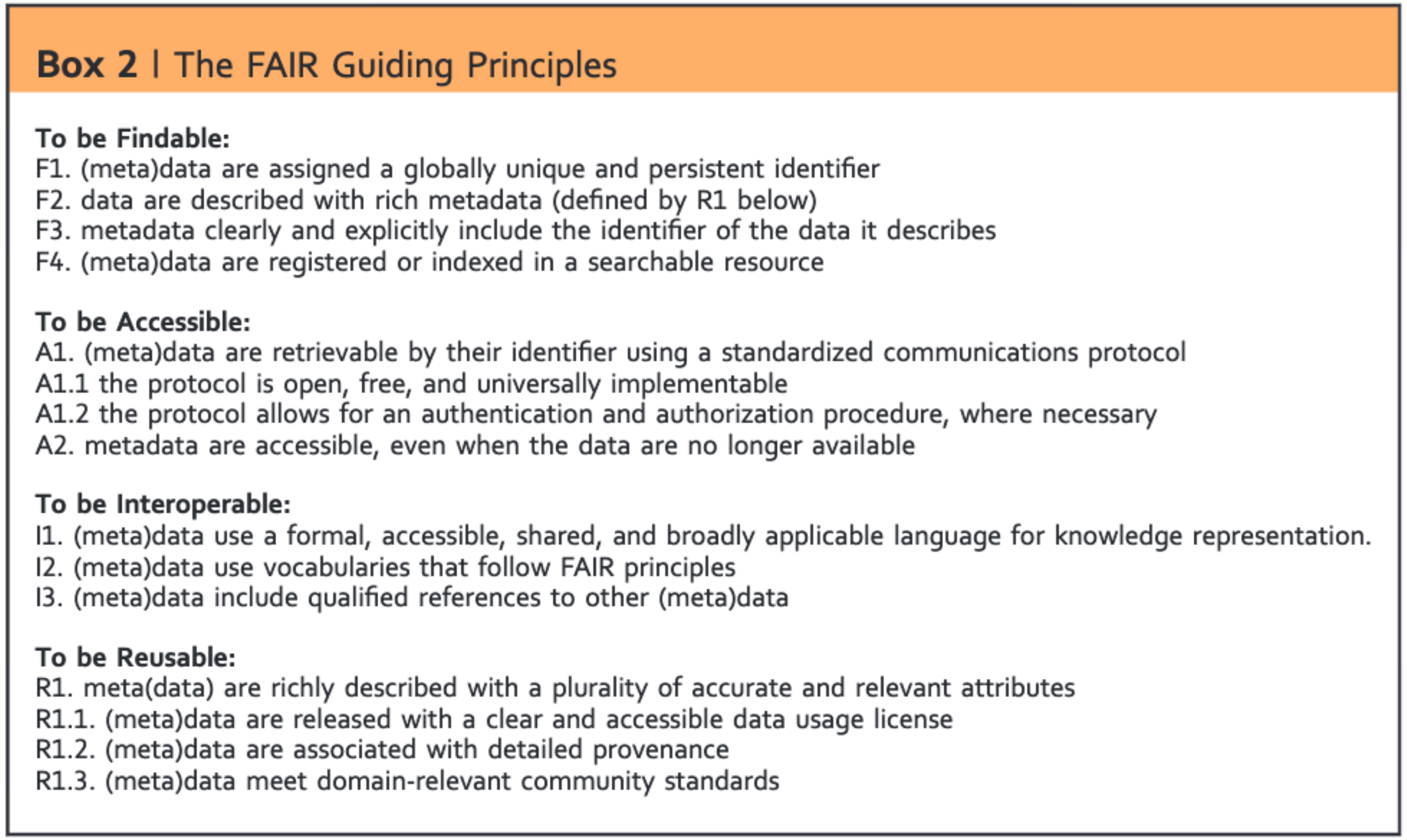
1. FAIR data
2. A (bad) data sharing example
3. Debugging and getting help
4. Lab Exercises
    — User defined functions

1. FAIR Data[1]

- **F**indability
- **A**ccessibility
- **I**nteroperability
- **R**eusability

[1] Wilkinson, Mark D, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Scientific Data 3 (1): 1–9.

FAIR principles



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

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

To be Reusable:
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

Source: Wilkinson et al., 2016

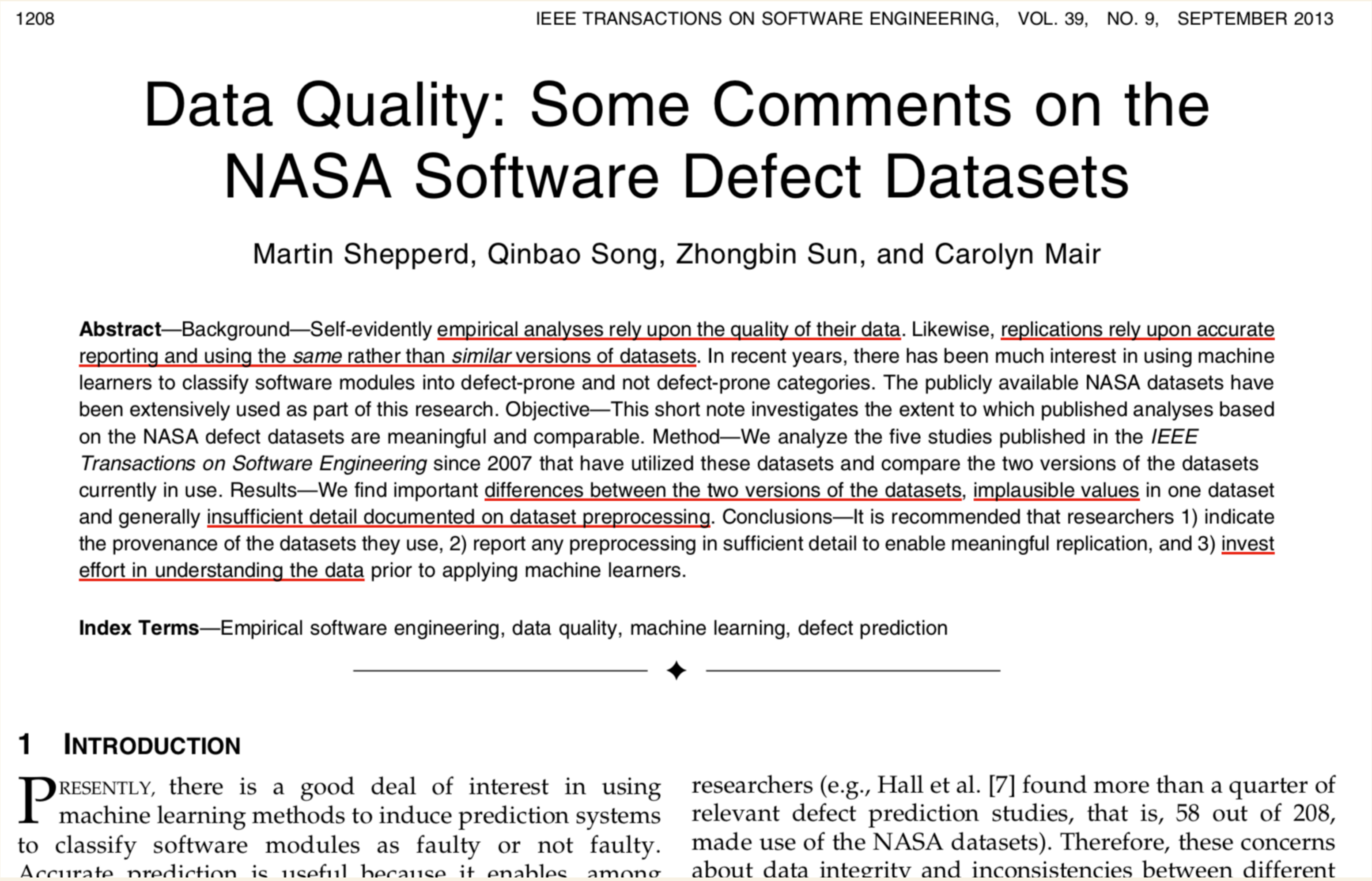
What is Meta-data?

- data about data
- meta-data describes how the data set is organised and the meanings of individual variables
- For example, we might say that the variable `Person ID`
  - comprises alpha-numeric characters (so it’s a string)
  - to be valid it must be (i) unique and (ii) exactly 8 chars
- it facilitates meaningful processing and analysis

Meta-data

- F2: data are described with rich metadata
- A1: meta-data are retrievable by their identifier [DOI] using a standardized communications protocol
- I1: meta-data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- R1.2: meta-data are associated with detailed provenance
- R1.3: meta-data meet domain-relevant community standard

2. A (bad) data sharing example[2]



1208 IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 39, NO. 9, SEPTEMBER 2013

Data Quality: Some Comments on the NASA Software Defect Datasets

Martin Shepperd, Qinbao Song, Zhongbin Sun, and Carolyn Mair

Abstract—Background—Self-evidently empirical analyses rely upon the quality of their data. Likewise, replications rely upon accurate reporting and using the *same* rather than *similar* versions of datasets. In recent years, there has been much interest in using machine learners to classify software modules into defect-prone and not defect-prone categories. The publicly available NASA datasets have been extensively used as part of this research. Objective—This short note investigates the extent to which published analyses based on the NASA defect datasets are meaningful and comparable. Method—We analyze the five studies published in the *IEEE Transactions on Software Engineering* since 2007 that have utilized these datasets and compare the two versions of the datasets currently in use. Results—We find important differences between the two versions of the datasets, implausible values in one dataset and generally insufficient detail documented on dataset preprocessing. Conclusions—It is recommended that researchers 1) indicate the provenance of the datasets they use, 2) report any preprocessing in sufficient detail to enable meaningful replication, and 3) invest effort in understanding the data prior to applying machine learners.

Index Terms—Empirical software engineering, data quality, machine learning, defect prediction

✦

1 Introduction

Presently, there is a good deal of interest in using machine learning methods to induce prediction systems to classify software modules as faulty or not faulty. Accurate prediction is useful because it enables among

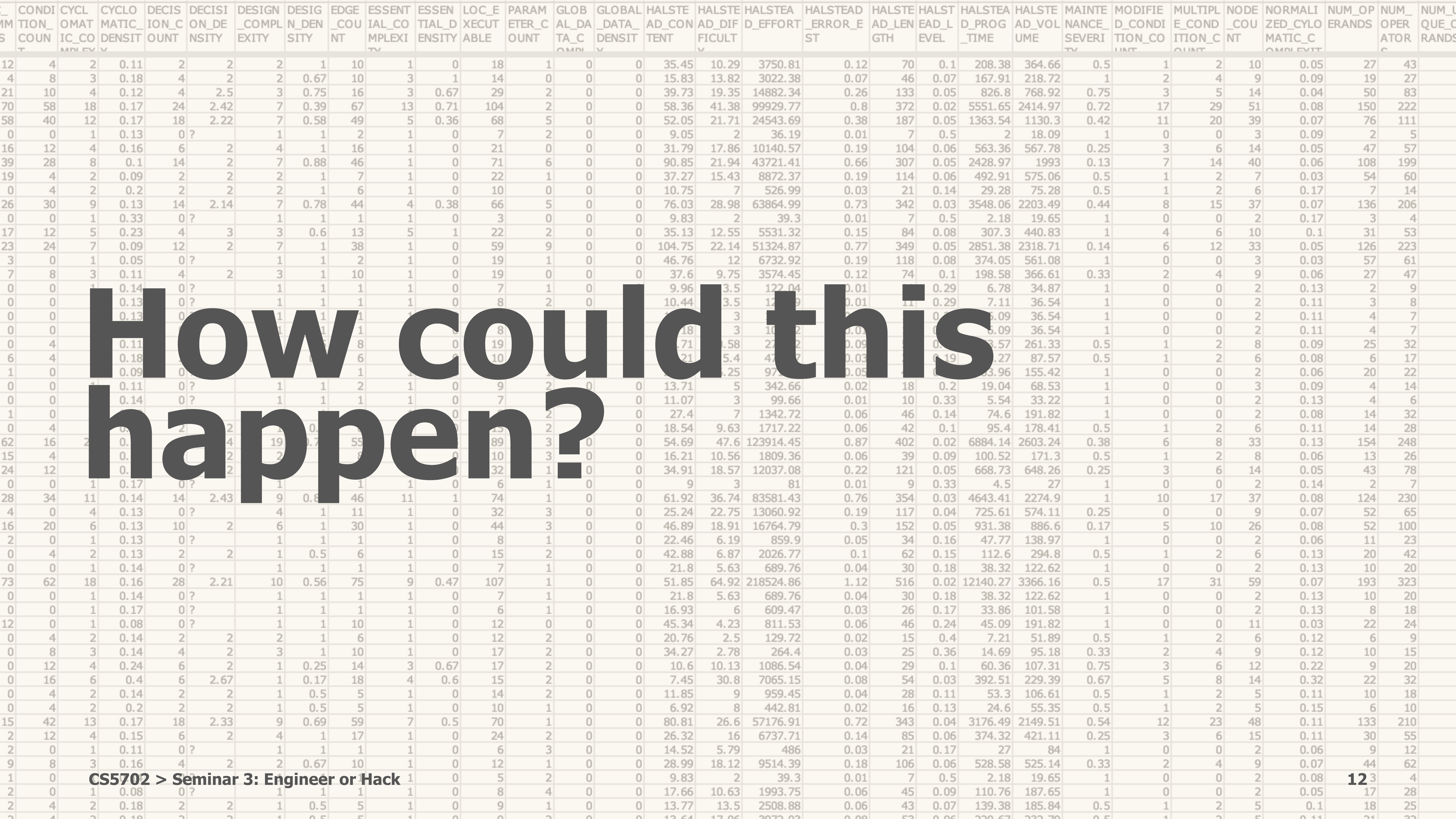
researchers (e.g., Hall et al. [7] found more than a quarter of relevant defect prediction studies, that is, 58 out of 208, made use of the NASA datasets). Therefore, these concerns about data integrity and inconsistencies between different

[2] Shepperd, M., et al., (2013). Data quality: Some comments on the NASA software defect datasets. IEEE TSE, 39(9), 1208-1215.

tl;dr

At least 100 research papers were based on public software defect data sets, until one day ...

we looked more carefully and realised that the data sets made no sense!



How could this happen?

Wisdom of crowds

Data checking in R

The following RMarkdown notebook is based on a simplified version of one of the software defect data sets described above.

(https://raw.githubusercontent.com/mjshepperd/CS5702-Data/master/CS5702_W3_NASAdata.Rmd) [https://raw.githubusercontent.com/mjshepperd/CS5702-Data/master/CS5702W3NASAdata.Rmd]

3. Debugging and Getting help …

1. find/read the relevant cheatsheet
2. perspiration e.g., see this five step approach
3. talk it over with a fellow student
4. module **FAQs** on Blackboard
5. Stack overflow
6. asking a member of the course team

For more suggestions visit the subsection 0.2 "vi) Learn how to get help" in the Modern Data book.

4. Lab Exercises

4.1 User defined functions

The RMarkdown Week 3 Lab Worksheet is available from:
https://raw.githubusercontent.com/mjshepperd/CS5702-Data/master/CS5702W3Lab.Rmd
and also on BBL (module content > Week 3)