

CAS Applied Data Science

Module 4

Ethics, best practices and tools

PD Dr. Sigve Haug, sigve.haug@unibe.ch

2023-10-20

Why ethics and best practices ?

Data and code is almost everywhere, and poor data and code
including documentation cost billions and maybe one day
humanity.

Therac-25

1. From 1985 to 1987 a computer controlled radiation therapy machine massively overdosed about six people. Some died.
2. Software controlled interlock failed due to a race condition (high dose was possible without appropriate shielding)



because multiple processes accessed
the same data at the same time

<https://de.wikipedia.org/wiki/Therac-25>

1991 Sinking of Norwegian Sleipner

1. 1991-08-23 The oil and gas platform Sleipner sinks
2. Economic loss about 700 MUSD

The post accident investigation **traced the error to inaccurate finite element approximation** of the linear elastic model of the tricell (using the popular finite element program NASTRAN). The shear stresses were underestimated by 47%, leading to insufficient design. In particular, certain concrete walls were not thick enough. More careful finite element analysis, made after the accident, predicted that failure would occur with this design at a depth of 62m, which matches well with the actual occurrence at 65m.



<http://www-users.math.umn.edu/~arnold/disasters/sleipner.html>

1996 Ariane 5 Explosion

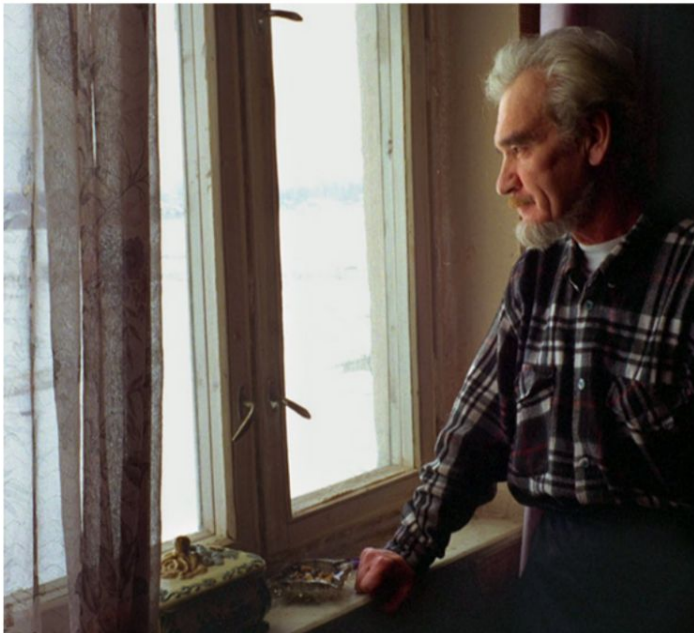
- June 4, 1996, unmanned rocket launched by ESA explodes
- Value about 500 MUSD
- Cause: Integer Overflow (conversion from 64 bit to 16 bit)



<https://www.youtube.com/watch?v=kYUrqdUyEpl>

1983-09-23 World War III (almost)

- Soviet early warning satellite reports 5 US missiles coming towards Soviet
- S. Petrov reports it as a false alarm (luckily)



- Could have caused massive attack from Soviet
- Was a misinterpretation of reflecting sun light from cloud tops

Stanislav Petrov : "I had a funny feeling in my gut"

AI kills woman March 2019



A woman crossing Mill Avenue at its intersection with Curry Road in Tempe, Ariz., on Monday. A pedestrian was struck and killed by a self-driving Uber vehicle at the intersection a night earlier. Caitlin O'Hara for The New York Times

Racial Bias in AI

← → ↺ nature.com/articles/d41586-019-03228-6

[nature](#) > [news](#) > article

NEWS | 24 October 2019 | Update [26 October 2019](#)

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

bias in training data

The Emergence of the Bug


Logbuch-Seite des Mark II Aiken Relay Calculator (Harvard) mit dem ersten *bug* (1947) - not so soft

9/9

0800 Antan started
1000 " stopped - antan ✓ { 1.2700 9.037 847 025
1300 032 MP-MC 2.130476415 9.037 846 995 correct
033 PRO 2 2.130476415 9.615 925059(-2)
correct 2.130476415

Relays 6-2 in 033 failed special speed test
in relay 11.000 test.

Relays changed
1100 Started Cosine Tape (Sine check)
1525 Started Multi Adder Test.

1545  Relay #70 Panel F
(moth) in relay.

First actual case of bug being found.
1600 Antan started.
1700 closed down.

Relay 3145
Relay 337



Mark II, general view of calculator frontpiece, 1948.

What can we do about it ?

- Conduct ethically
 - Do the right and not the wrong thing when doing our data science
 - Implies a choice

- What is the right thing ?
 - Often respecting the law and the regulations
 - Perform your work respectfully to the resources of the planet, including animals and other colleagues

International “Standard”

Data, code and other stuff should be FAIR

- Findable
- Accessable
- Interoperable Format that others can also use
- Reproducible

Fortunately there are commonly accepted Best Practices and Tools to help us being FAIR and ethical

Module 4

2023-10-20 Best practices and Documentation (half day - today)

2023-10-27 Cyber Security (full day)

2023-11-03 Git 1 (half day)

2023-11-10 Git 2 (half day)

2023-11-17 FOSS (half day)

2022-11-30 Deadline for module project (github repo)

Some Best Practices

Data Science Project - typical pipeline

- 1 Objective / research / business case formulation (CDR)
- 2 Data acquisition
- 3 Data cleaning / preprocessing
- 4 Feature engineering
- 5 Modelling
- 6 Results and evaluation
- 7 Presentation and publication
- 8 Production

Best Practices (BP) - Teamwork

You are assigned to 1 of 8 breakout rooms

- Room number defines the BP number in the article on Ilias to present (1 slide)
- Each room/team also considers and presents 1 slide on best practices on their pipeline number
- 30 minutes teamwork
- Each team has 3 minutes for their presentation (2 slides) (starts at 10:00)

Write your slides [here](#) !

Documentation (11:00-12:00)

Good Practice --> store documentation with the code

- Levels

- Good variable and method/function names
- Inline comments / docstrings (in, out, what it does)
- Readme files
- wiki pages
- Static webpages (automatically generated)
- Manual (pdf)
- Book
- Notebooks

dont overdo it either; dont say how you do things (can be seen in code), but why

- Use VCS (Git)

Tutorial - Inline and README

I will use my terminal and the nano editor on my MacOS. Windows users can use powershell, or better a terminal from Anaconda. You can also use an Integrated Development Environment (IDE) like Visual Studio, PyCharm or Spyder. A third option is subscribe to colab pro for 10 CHF per month. It provides a terminal.

Linux cheat sheet

<https://cheatography.com/davechild/cheat-sheets/linux-command-line/>

One simple editor is nano. If you don't have it, you can install it.

apt-get install nano

I will now write a python script and document it with inline comments and a README file.

nano perfect-numbers.py

python perfect-numbers.py

nano README.md

Tutorial - Webpages with Sphinx

Sphinx (and other tools)

- Generates webpage documentation of your project from Markdown files
- Files lives together with code in a doc/ subfolder
- When used with git, code and documentation are synchronised
- When pushing the project to GitHub, [ReadTheDocs](#) can automatically update its public webpages

This kind of documentation is rather for larger projects. If you want to practice it, you can do so in the module project ([instructions](#)), however, this is not mandatory.

More on Documentation

<https://coderefinery.github.io/documentation/>

Bias in Data Science

<https://www.youtube.com/watch?v=PWCtoVt1CJM>

12:00 - 12:20 Q&A