CAS Applied Data Science

Module 4

Ethics - best practices and tools

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

2024-11-01

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)

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
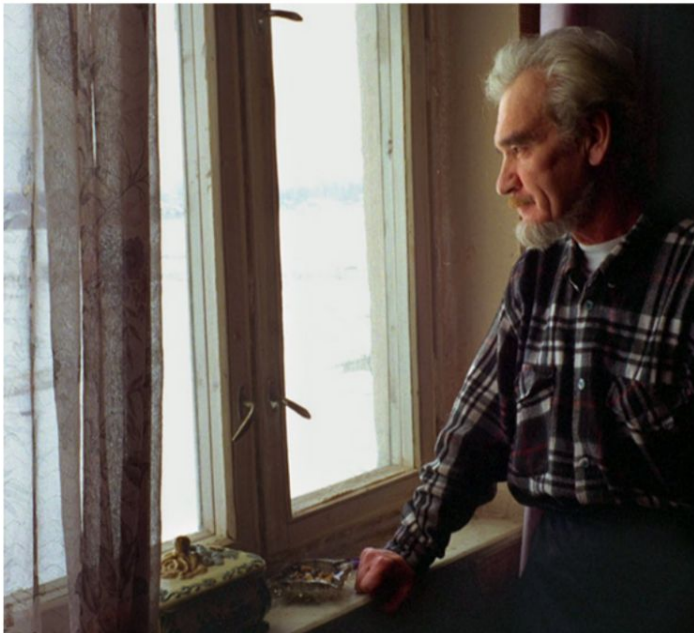
- June 4, 1996, unmanned rocket launched by ESA explodes

- Value about 500 MUSD

- Cause: Integer Overflow (conversion from 64 bit to 16 bit)

https://www.youtube.com/watch?v=kYUrqdUyEpI

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 > 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.

The Emergence of the Bug

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

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
- Reproducible

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

Module 4

2024-10-18 Git and GitHub 1

2024-10-25 Git and GitHub 2

2024-11-01 <mark>Best practices and Documentation (today)</mark>

2024-11-08 Cyber Security

2024-11-15 FOSS

2024-11-30 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 about 1 slide on their BP number
- ○ 30 minutes teamwork
- ○ Each team has 3 minutes for their presentation

Write your slides [here](#) !

Documentation

- Levels
  - ○ <mark>Good variable and method/function names</mark>
  - ○ <mark>Inline comments / docstrings (in, out, what it does)</mark>
  - ○ <mark>Readme files</mark>
  - ○ wiki pages
  - ○ Static webpages (automatically generated)
  - ○ Manual (pdf)
  - ○ Book
  - ○ Notebooks
- <mark>Use VCS (Git)</mark>

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 - Web pages 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

16:30 - 17:00 Q&A