Applying The Hacker Within lessons to a research project

Flaviu Gostin

18 March 2019

Introduce myself

- Field: Materials Science and Engineering
- Research area: corrosion of Ti alloys for medical implants
- No formal programming training

My path to The Hacker Within

- Excel catastrophes early on. Origin much better, but still not enough
- Keyword: unnatural
- I had enough of Windows and GUIs, all the good tools seem to be developed for GNU/Linux and as Python packages
- 19 March 2018, Matt's welcome talk - the right talk at the right time
- Switch to GNU/Linux, command line, Emacs, Python, Git

What this presentation is about

- My first attempt to make a new analysis workflow for my recent research project, which is (computationally) reproducibile
- Describe the research
- Show the results I want to reproduce
- Show how I did it
- Show the new workflow

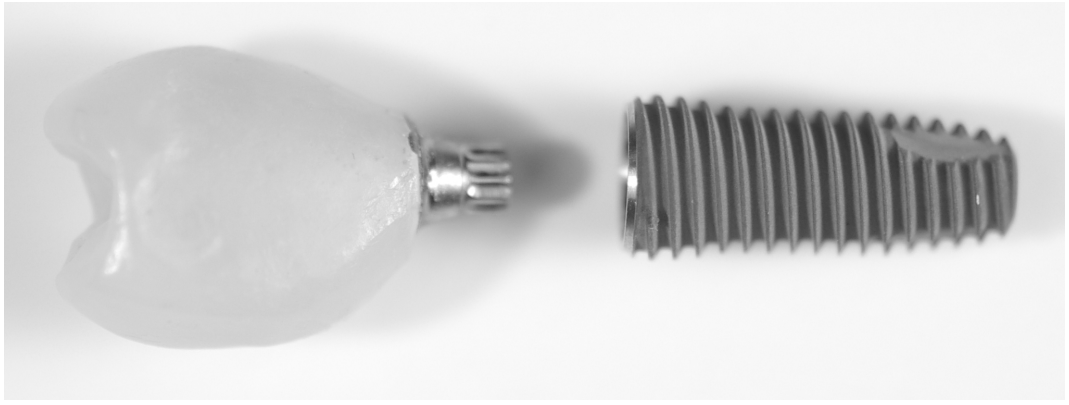
Practical stuff

- Let's see if my analysis can be reproduced on your machine:
- Navigate to:
  https://github.com/craicrai/xrd_analysis_workflow
- Fork
- Open terminal
- cd Desktop
- git clone
  https://github.com/*your-user-name*/xrd_analysis_workflow
- cd xrd_analysis_workflow
- make all

DESCRIBE THE RESEARCH

New titanium alloy for implants

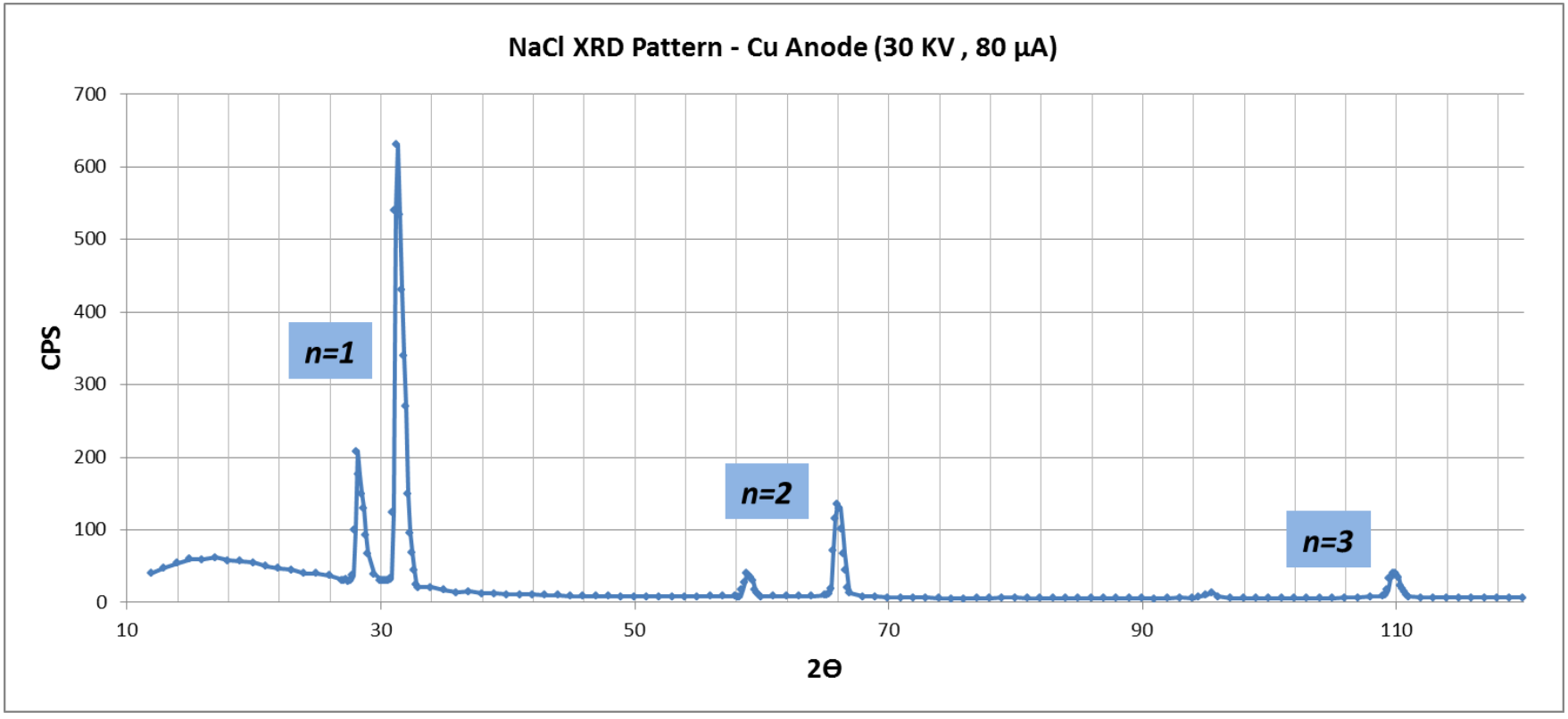
- More than 3000 tonnes of titanium alloys implanted in people every year
- Titanium is very corrosion resistant, but not perfect
- New alloy: Ti40Zr10Cu34Pd14Sn2 (at. %)
- It is important to know its corrosion products



Dental implant made of titanium

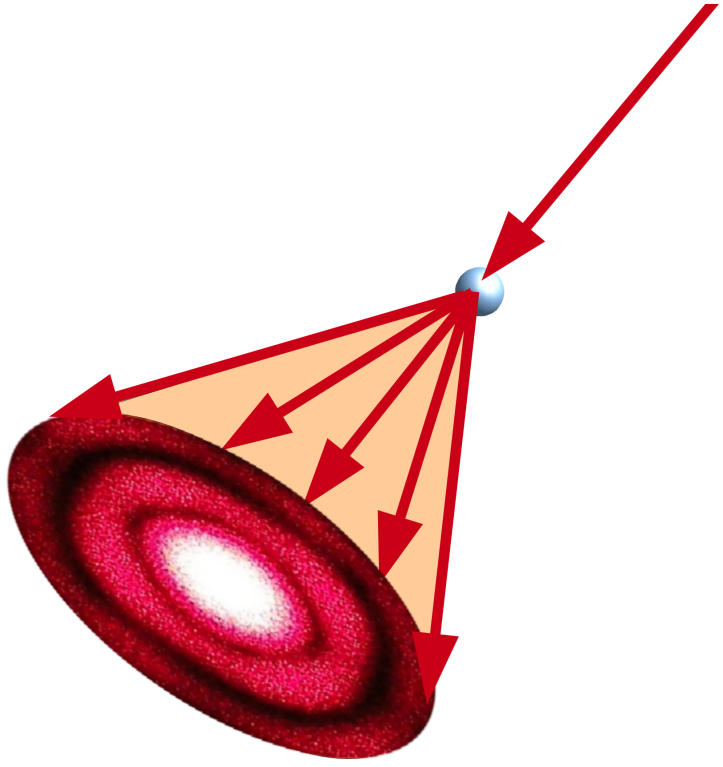
How can we see what the corrosion products are?

- Shine X-rays on corrosion products which diffract them
- Resulting diffraction pattern is like a fingerprint



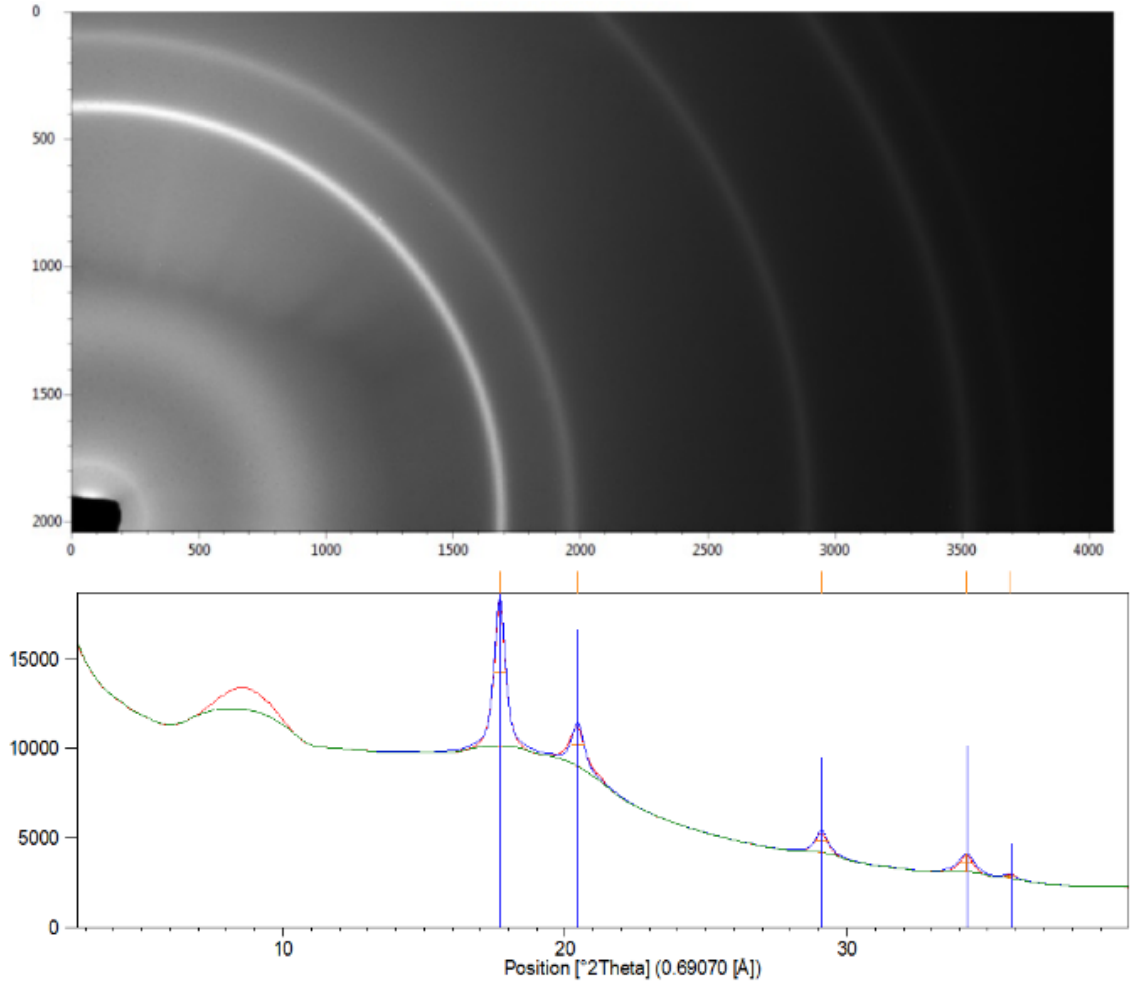
Diffraction pattern of NaCl (kitchen salt) from physicsopenlab.org

Diffraction experiment



Diffraction rings (from Wikipedia)

Azimuthal integration



Calibration required to find the centre and the detector tilt

SHOW THE RESULTS I WANT TO REPRODUCE

Selected raw 2D diffraction image

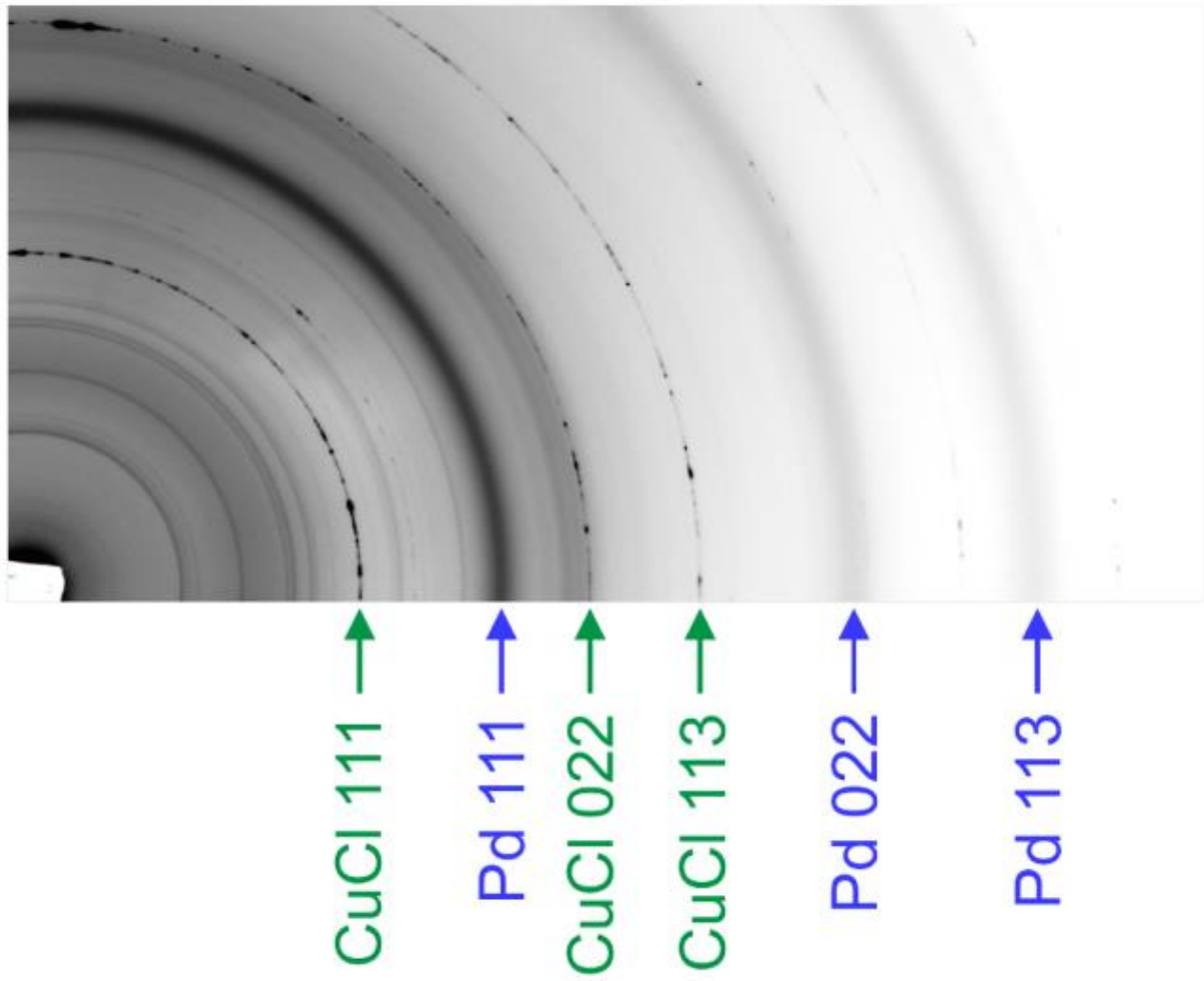


Figure S1 in paper

Stack of 1D diffraction patterns

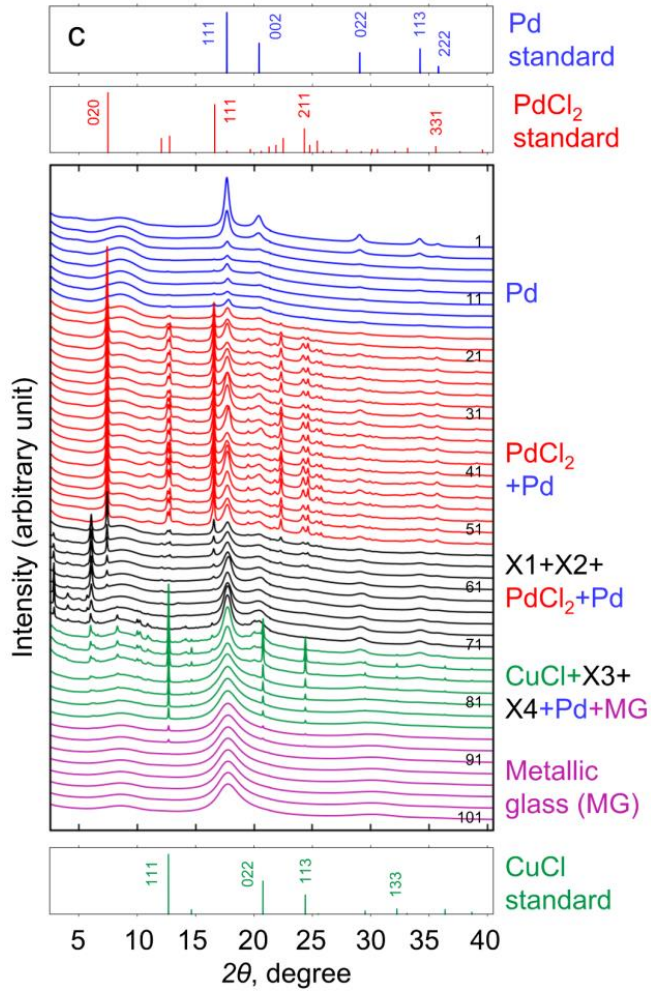


Figure 1c in paper

Extract peak info and calculate lattice spacing

Electrolyte/ Phase	Applied potential [V vs. Ag/AgCl]	Lattice parameter [Å]	Crystallite size [nm]	$n^{d)}$
PS[a)]		3.891(5)	6.8 ± 0.9	82
PS + A[b)]	1.3	3.887(3)	6.9 ± 0.5	43
PS + P[c)]		3.890(4)	5.0 ± 0.6	63
PS + A + P		3.892(3)	6 ± 1	21

Table 2 in paper

HOW I DID THE ANALYSIS ORIGINALLY

- complete chaos!
- conspicuously irreproducible

Documentation

- no repository, no appendix with details, just this:

  stepwise automatically with a constant predefined step size in the interval 3-200 m. Detector calibration and azimuthal integration of raw two-dimensional diffraction patterns were performed with the software DAWN.[45] Crystallite sizes were determined with Scherrer formula.[13] Instrumental broadening was assumed to have a value of 0.1°

  Excerpt paper

Project organization

- very poor organization
- afraid of losing track of which data is where: just leave it as it comes
- inconsistent structure
- mixed raw data with processed data with metadata with Python scripts etc.

Software

- DAWN Science for calibration and azimuthal integration
- Brucker X pert for peak detection, fitting and indexing
- Also used Match for the same thing as it had access to a different database
- CrystalDiffract
- CrystalMaker
- gnuplot for plotting

Version control

- ls | less

```
Manuscript Flaviu Synchrotron Advanced Functional Materials ajd.docx
Manuscript Flaviu Synchrotron Advanced Functional Materials.docx
Manuscript Flaviu Synchrotron Advanced Functional Materials.docx.pdf
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa.docx
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg-ajd.docx
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg-am.docx
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg-am-ki-al-srs-ylc-pfg-ajd2.do
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg-am-ki-al-srs-ylc-pfg-ajd.doc
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg-am-ki-al-srs-ylc-pfg.docx
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg-am-ki-al-srs-ylc-pfg-MS.docx
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg-am_ki.docx
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg.docx
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg(TrackedChanges).docx
Manuscript Flaviu Synchrotron Advanced Functional Materials-oa-pfg-ylc.docx
Manuscript Flaviu Synchrotron AFM final compressed images SUBMITTED.docx
Manuscript Flaviu Synchrotron AFM final.docx
Manuscript Flaviu Synchrotron AFM modif all authors ajd oa.docx
Manuscript Flaviu Synchrotron AFM modif all authors ajd oa pfg ajd2 compressed images.docx
Manuscript Flaviu Synchrotron AFM modif all authors ajd oa pfg ajd2 pfg2.docx
Manuscript Flaviu Synchrotron AFM modif all authors ajd oa pfg ajd2 pfg2.pdf
Manuscript Flaviu Synchrotron AFM modif all authors ajd oa pfg.docx
Manuscript Flaviu Synchrotron AFM modif all authors ajd oa pfg.pdf
Manuscript Flaviu Synchrotron AFM modif all authors.docx
Manuscript Flaviu Synchrotron AFM modif all authors.pdf
:
```

HOW I WOULD DO THE ANALYSIS NOW

Resources

- Previous THW presentations
- Wilson et al. (2017). Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510
- Millman et al. (2018). Teaching Computational Reproducibility for Neuroimaging. Front. Neurosci. 12:727. doi: 10.3389/fnins.2018.00727
- https://github.com/berkeley-stat159/project-alpha
- Matthew Brett. (2017) Curious git (0.2). https://matthew-brett.github.io/curious-git/index.html
- The Internet using DuckDuckGo, Stack Ovferflow

Tools

- Keep it simple!
- This presentation: done in Markdown, converted to pdf with Pandoc
- Version control: git. All git actions done in Bash, used GitHub only as remote repository
- Bash, Emacs, Python
- Instructions in README
- Left Jupyter and Binder for later

Ensuring a reproducible environment

- virtualenv
- made directory venv/ in project root
- pip freeze > requirements.txt, NO! better manually

“Always search for well-maintained software libraries that do what you need”

- fit2d, oldest (?) and most known
- pyFAI, in Python, faster than fit2d
- DAWN Science, it’s a GUI, in Java?
- GSAS-II, in Python and does a lot more!
  https://subversion.xray.aps.anl.gov/trac/pyGSAS

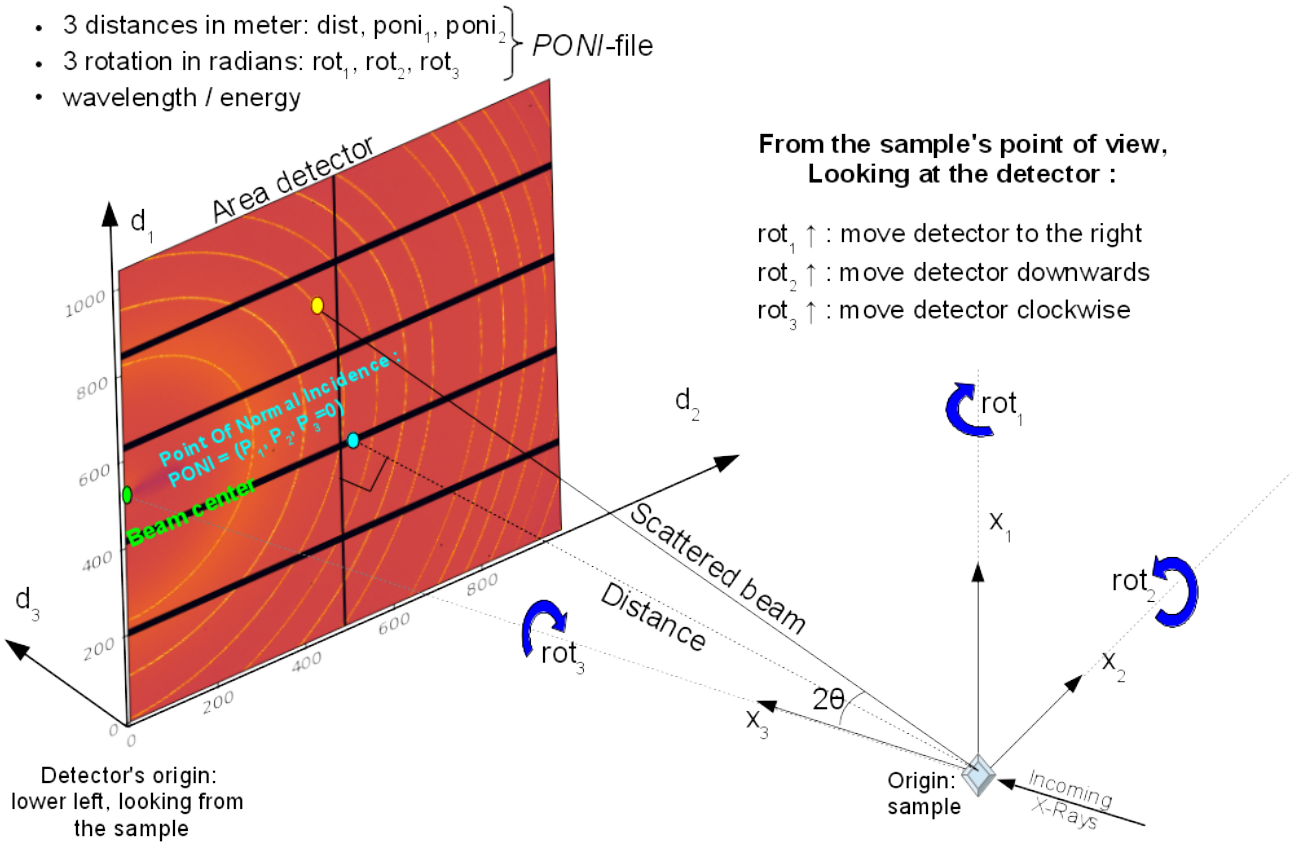
Data processing steps in pyFAI

- Calibration, calibrate experiment geometry and save it to a .poni file
- Azimuthal integration, use the .poni file

pyFAI import does not work

- import function does not work on my .hdf files! -> contact dev team, report bug, contribute?
- use h5py module. Spent a few good hours to understand how it works
- write small script to visualize the groups tree inside hdf files
- dataset is a 3D array (stack of 2D images)
- write function to extract individual diffraction images as 2D numpy arrays

Experiment geometry in pyFAI



PONI - point of normal incidence

Calibration

- normally, done using a GUI
- 'tell' the GUI which ring is which by clicking (!) five points on each ring
- how does one reproduce a click?
- ...
- the calibration determines the geometry of the setup, which is saved in a .poni file

Azimuthal integration

- create an AzimuthalIntegrator (ai) object with the .poni file
- ai.integrate1d(img to integrate as ndarray, etc) all diffration images

Test driven development?

- at the beginning, not really
- because struggling to figure out how everything should work together

Make raw data available

- uploads to Zenodo receive DOI as soon as the data is uploaded so there is no chance to modify it
- uploading an archive 1.3G to Google Drive did not work for me; wget cannot download directories from Google Drive
- finally uploaded 1.3G archive to Figshare
- when entire raw data set is ready, upload to Zenodo. Include code?

To do

- Remove variables from scripts and merge them in a txt file in data/ to avoid errors due to duplication, e.g. wavelength used in several scripts
- Create metadata and store in repository,
- Peak fitting, extract peak
- More plots

Check reproducibility by different people on different machines

- OceanNuclear, mkdir data and clarify README
- Greg, make README more concise and ImportError (tk . . .)
- Observations from auditorium?

Not the last slide

- Data processing workflow is reproducible
- . . . but it does not necessarily imply it is correct
- . . . but at least interested people have the chance to check it

My impressions

- This is so much fun!
- This is better than I imagined because
- I can go back to it anytime and see *exactly* how the analysis was done
- and I or someone else can re-use it for other projects
- this process actually helped better understand the processing of my data and build confidence