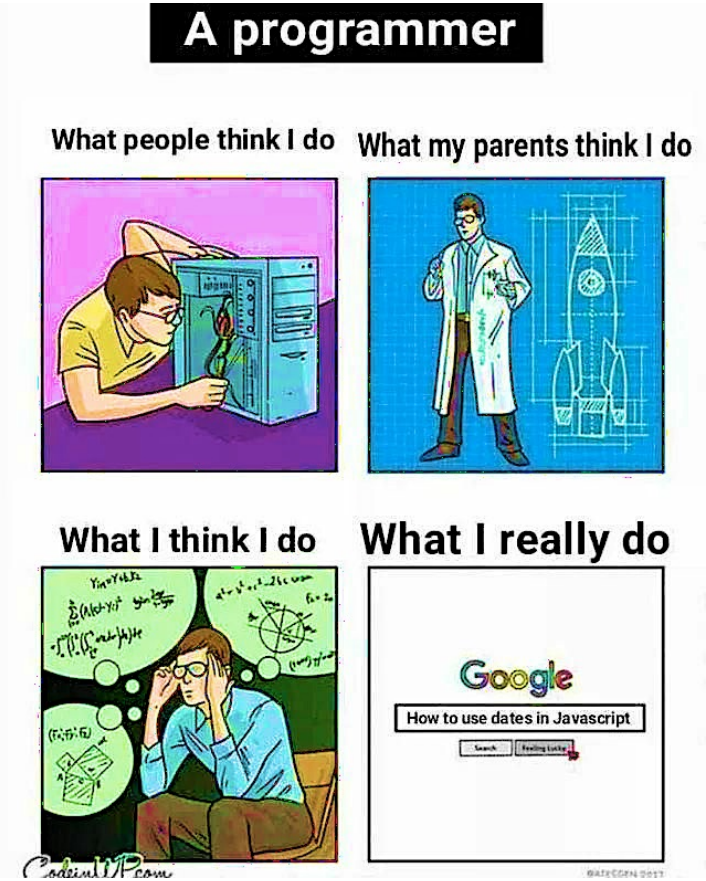
“I can’t code” and other reproducibility-blockers

Dr Lucy Whalley

Assistant Professor in Physics, Northumbria University
Associate Editor, Journal of Open Source Software

l.whalley@northumbria.ac.uk
Website: lucydot.github.io



Good research is reproducible research

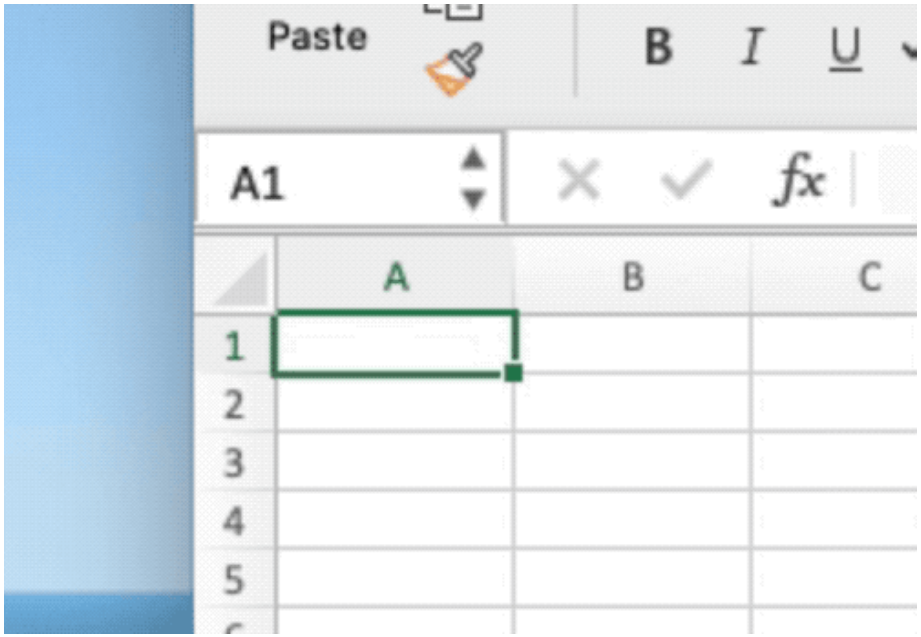
Vanilla sponge

Ingredients (data)

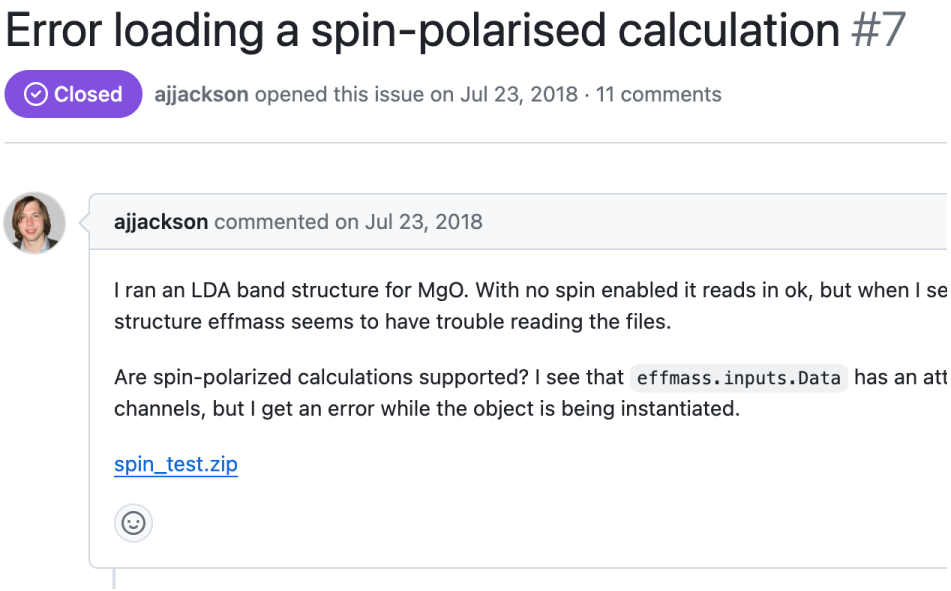
Recipe (methods)

Chocolate sponge

Researchers make errors



19.6% of genetic research crunched in excel contains errors[1]



My research code contains errors

Computational reproducibility

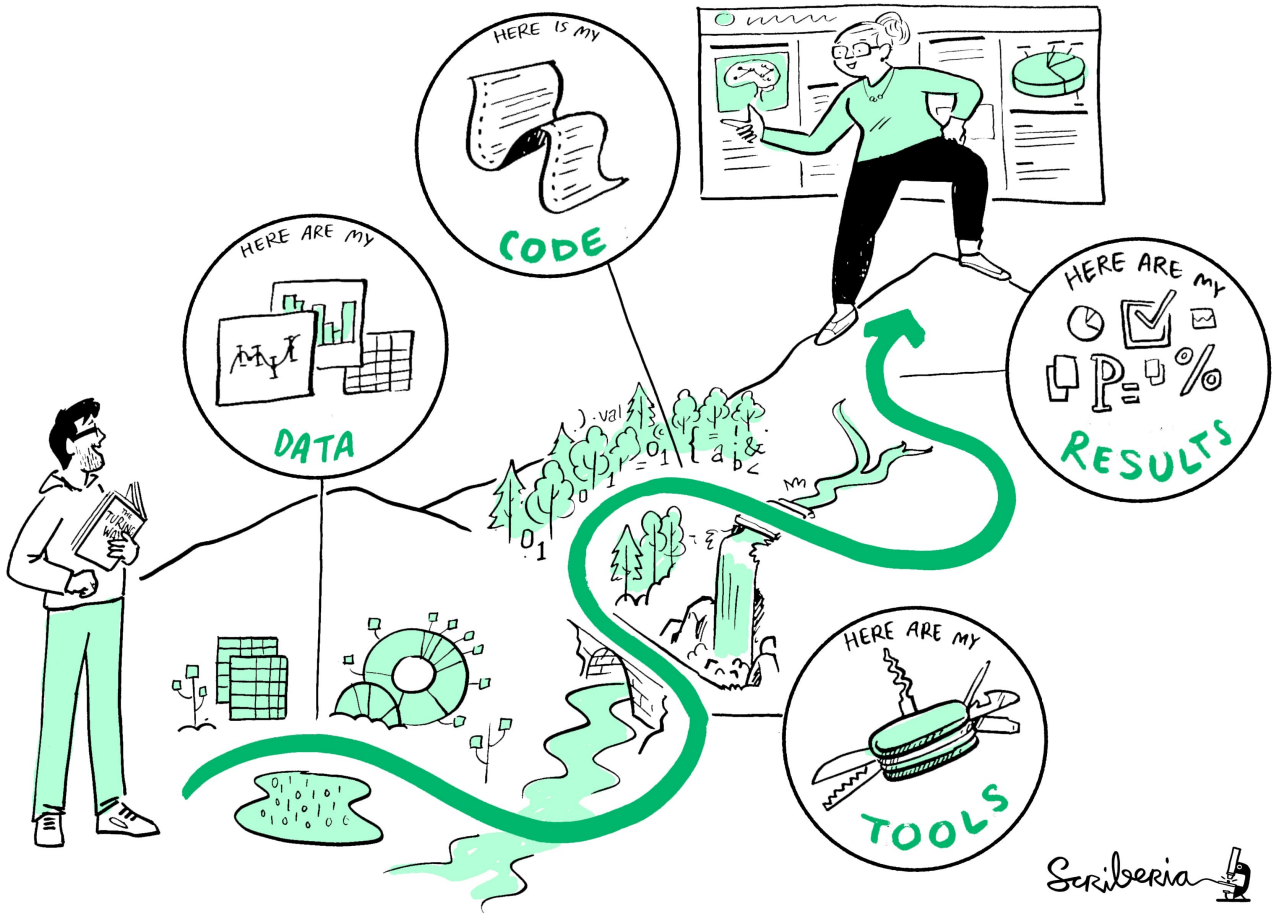
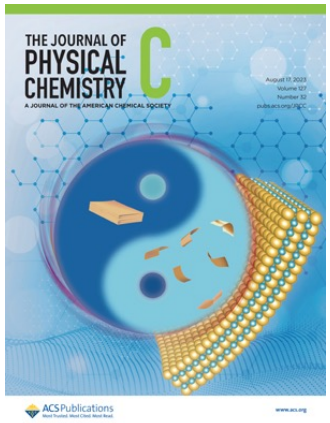


Image from
The Turing Way[2]

Our Approach

1. Journal Article[3]



2. Project-specific repository[4]

Data and analysis code for "Steric Engineering of Point Defects in Lead Halide Perovskites"

📖 This paper is published with open access in J. Phys. Chem. C here.

All of the code is distributed as Jupyter Notebooks. If you are looking for the code that implements the interpolation method used in the paper, please see this repository. If you are looking for raw DFT input and output files for the total energy calculations used to predict defect properties, please see this repository.

3. Domain-wide data repository[5]

NOMAD

Materials science data managed and shared

NOMAD lets you manage and share your materials science data in a way that makes it truly useful to you, your group, and the community. **Free and open source.**

Open NOMAD →

4. Pre-print[6]

arXiv

Our Approach

Project-specific repository

GitHub

+

Preview | Code | Blame | 191 lines (191 loc) · 75.4 KB | Raw

Symmetry Mode Analysis

In [8]:
```
import numpy as np
import csv
import re
import matplotlib
import matplotlib.pyplot as plt
from collections import OrderedDict
```

In [20]:
```
def get_data(filepath, cutoff):

    with open(filepath) as File:
        content = File.read()

    label = re.findall('Pm-3m\[\d*\/?\d*,\d*\/?\d*,\d*\/?\d*\]([A-Z]*\d*[+-?])',c
    label_content = re.findall('Pm-3m\[\d*\/?\d*,\d*\/?\d*,\d*\/?\d*\]([\s\S]*?)\

    totals = []
    for content in label_content:
        decimals = re.findall('(-?\d+\.\d+)',content)
        totals.append(sum([abs(float(entry)) for entry in decimals]))

    data = {label[i]: totals[i] for i in range(len(label))}
    data = OrderedDict(filter(lambda data: data[1] > cutoff ,data.items()))
    data = OrderedDict(sorted(data.items(), key=lambda data: data[1],reverse=True

    return data

def plot_data(data,amp):
    plt.style.use('seaborn-colorblind')
    plt.figure(figsize=(20,10))
    plt.bar(range(len(amp)), amp, align='center')
    plt.xticks(range(len(amp)), list(data.keys()),fontsize=20)
    matplotlib.rc('xtick', labelsize=20)
    matplotlib.rc('ytick', labelsize=20)
    plt.ylabel("Mode amplitude",fontsize=20)
    plt.axis(ymin=0,ymax=2.2)
    plt.show()
```

In [21]:
```
# all phonon modes with amplitude below this cutoff will not be plotted
```

+

open source initiative®

+

zenodo

Jupyter Notebook to map from Data to Code

Happy colleague

samhood 1:47 PM
Thank you!!

Your packages are so nice, you're amazing ❤️

lucydot 1:49 PM
I am totally screenshot-ing that comment for a future talk!!!

Big person in the field

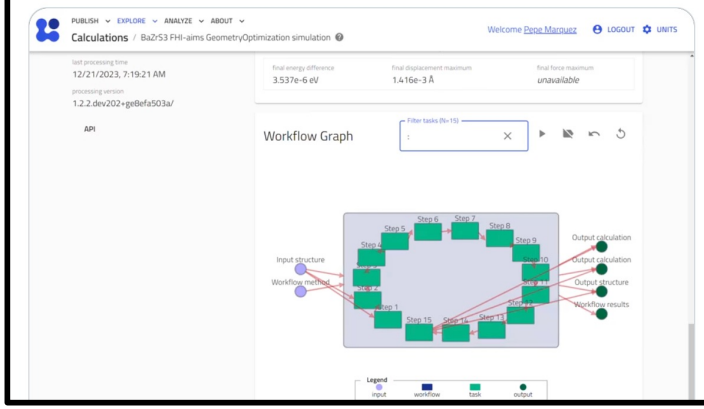
Volker Blum @AimsDuke · 16 Mar
Truly excellent news! This should be super-useful. Thank you!

> **Lucy Whalley** @lucydotwhalley · 16 Mar
> The effective mass code `effmass` has been updated! It now plays nicely with FHI-aims (thanks @matthias_golomb), Vasp, Castep and ASE. And it comes with a super trendy command line interface: github.com/lucydot/effmass.

Little person in the field

Pepe Marquez
@PepeMarquez9

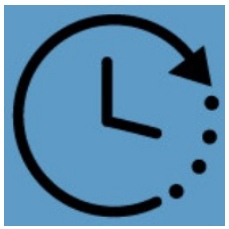
Nice nice!
Some cool DFT calculations on BaZrS3 🩵chalcogenide perovskites just made it to #NOMAD courtesy of @lucydotwhalley and @prakayastha314 !



Reproducibility-blockers

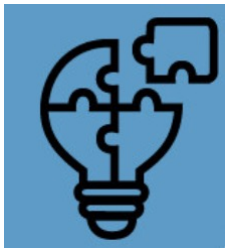
For most of the papers, there was little to provide any help to a researcher willing to reproduce the calculations: the crystal structures and input files were not provided.[7]

Time pressures



Lack of incentives

Fear of scooping



"I can't code"

Sensitive data

Coding has an image problem

Women invented programming

Ada Lovelace wrote the first computer programme

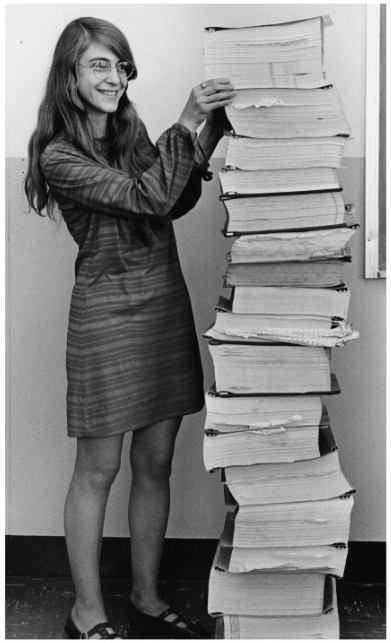
Grace Hopper invented the first compiler

Women were the first programmers

1969: 'Space age needleworker "weaves" core rope memory for [Apollo missions'] computers.' (Raytheon, 1969, p. 18)

1962: Mathematicians and programmers, Patsy Simmers, Gail Taylor, Milly Beck, Norma Stec, holding parts of the first computers.

c. 1972: African-American woman computer operator at the Office of Personnel Management.



1969: Margaret Hamilton with the code she and her staff wrote for the Apollo 11 mission.

Teaching coding inclusively: if this, then what?

Olivia Guest[1] and Samuel H. Forbes[2]

What happened in the 1980s?

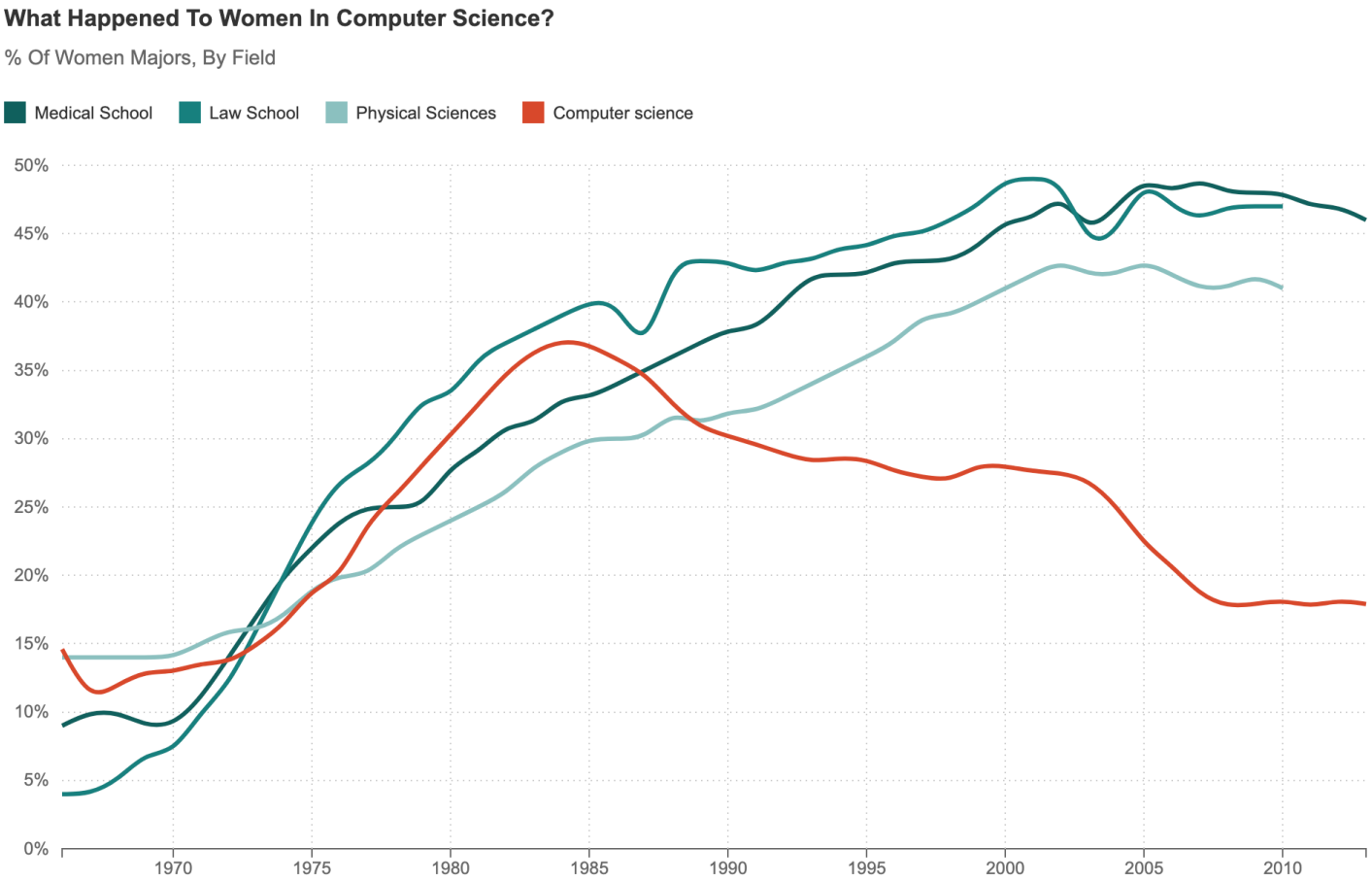
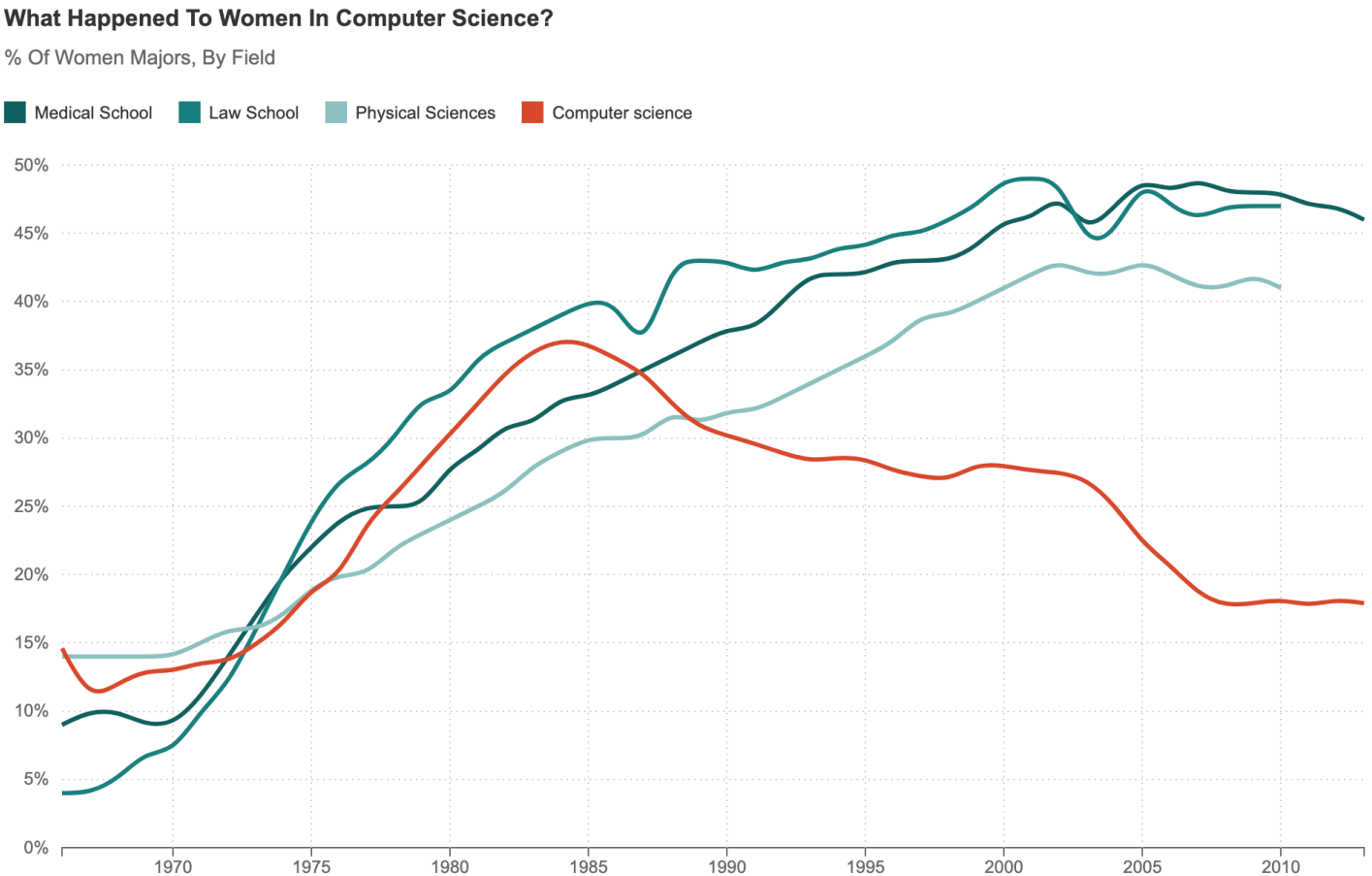


Image from NPR Planet Money "When women stopped programming"

What happened in the 1980s?



Changing (student) attitudes

“I can’t code”	→ like any other skill coding takes practice, and you *will* generate a lot of errors on the way
“I am too old to learn to code”	→ there is no critical developmental window for learning to code
“If we learn to code we will not have time to learn X”	→ Coding is an increasingly *necessary* part of research

Teaching coding inclusively: if this, then what?

Olivia Guest[1] and Samuel H. Forbes[2]

Communities of support

THE CARPENTRIES

We teach foundational coding and data science skills to researchers worldwide.

For those new to programming

For career advice

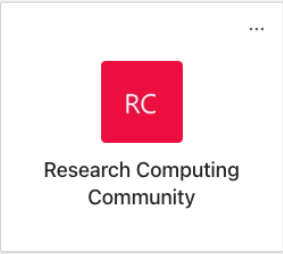
RSE

RESEARCH SOFTWARE ENGINEERS ASSOCIATION

Software Sustainability Institute

For reaching the wider community

Northumbria specific



Contribution takes many forms

Summary

1) Good research is reproducible

2) Join the “Research Computing Community” for further discussion

3) Jupyter Notebooks are a useful tool

4) Computing has an image problem: think about building confidence

5) Code contributions do not need to be technical

l.whalley@northumbria.ac.uk
lucydot.github.io

References

1) Gene name errors in Excel: https://doi.org/10.1186/s13059-016-1044-7
2) The Turing Way: https://the-turing-way.netlify.app/index.html
3) Steric engineering journal article: https://doi.org/10.1021/acs.jpcc.3c03516
4) Steric engineering project repository: https://github.com/NU-CEM/MACsPbI3_defects
5) Steric engineering NoMaD dataset: https://dx.doi.org/10.17172/NOMAD/2023.12.21-1
6) Steric engineering pre-print: https://arxiv.org/abs/2302.08412
7) Reproducibility in computational chem: https://doi.org/10.1021/acs.chemmater.7b00799

Further Reading

1) The Turing Way: https://the-turing-way.netlify.app/index.html
2) Teaching coding inclusively: https://osf.io/preprints/socarxiv/3r2ez