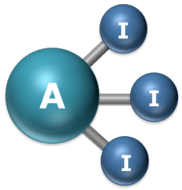

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

The Physical Sciences Data-Science Service (PSDS)

Patterns

Failed it to Nailed it: Data Tips & Tricks
22/10/2020
Online Event

Dr Samantha Kanza & Dr Nicola Knight
University of Southampton

18/11/2020

AI3SD-Event-Series:Report-19

Failed it to Nailed it: Data Tips & Tricks
AI3SD-Event-Series:Report-19
18/11/2020
DOI: 10.5258/SOTON/P0032
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

An EPSRC National Research Facility to facilitate Data Science in the Physical Sciences: The Physical Sciences Data science Service (PSDS)
This Facility is EPSRC Funded under Grant No: EP/S020357/1

Principal Investigator: *Professor Simon Coles*
Co-Investigators: *Dr Brian Matthews, Dr Juan Bicarregui & Professor Jeremy Frey*

Contents

1 Event Details

Title	Failed it to Nailed it: Data Tips & Tricks
Organisers	AI^3 Science Discovery Network+, Patterns Journal & Physical Sciences Data-Science Service
Dates	22/10/2020
Programme	AI3SD Event Programme
No. Participants	48
Location	Online Event
Organisation / Local Chairs	Dr Samantha Kanza & Dr Nicola Knight

2 Event Summary and Format

This event was the first of the 'Failed it to Nailed it' online data seminar series. The event was hosted online through a zoom conference. The event ran for approximately three hours in an afternoon session.

There were three talks given on the topics of data management and data sharing both from a domain agnostic point of view and from specific domain experts. These talks were followed by an interactive panel session with six early career researchers (ECRs) talking about their experiences in working with and sharing data.

3 Event Background

This event is part of the 'Failed it to Nailed it' data seminar series. This event series, currently comprised of four online events is a collaboration between the AI^3 Science Discovery Network+, Patterns & the Physical Sciences Data-Science Service (PSDS). This event series follows on from a data sharing survey that was undertaken earlier in 2020. Each event in the series handles a different aspect of dealing with data aiming to educate and inform researchers about how to work well with their data, as well as encouraging discussion along the way. Following on from these events the organisers hope to be able to organise more face-to-face events in 2021 which will expand this event series.

Good data management and sharing are central to the work that any researcher undertakes. In these events we want to encourage researchers to consider this as a fundamental aspect of their work rather than an afterthought, or something they wish they did but never got round to. This 'Tips and Tricks' event focuses on good practice that researchers can integrate into their work to ensure better data management, storage and sharing.

4 Talks

4.1 Love notes to the future: the importance of metadata - Ms Isobel Stark (University of Southampton)

http://orcid.org/0000-0002-8026-3315

Figure 1: Ms Isobel Stark

The full video of Isobel's talk can be viewed here: https://youtu.be/bvD7k5xKnXo [1]

Isobel Stark is the University of Southampton's Research Data Development Manager; she holds an MA in Librarianship from the University of Sheffield, and has had extensive experience of working in academic libraries throughout her career. Isobel works in partnership with other university services and researchers to ensure that research outputs are effectively captured, recorded, stored and shared, maximising the exposure of research to public and academic audiences. She has extensive expertise in Open Access, Data Management, Measuring Impact (Bibliometrics) Systematic Reviews, Theses and other research related matters.

Isobel's talk was focused on the importance of good research data management and how this can pay off in the future. This talk discussed four main aspects of data management: The data management plan, data storage, finding your data and sharing your data. Good data management is really important. Ultimately, managing your data well will save you time and effort in the future, making it easier to find, use, and distribute to others later on. A core part of data management is the data management plan, which should set out the full plan for what data is going to be gathered, how it is going to be catalogued and in what formats it is going to be stored in (paper/electronic/physical data).

The key points of data storage to consider are where the data is going to be stored, what is going to be stored (both long and short term), and what methodologies have been used to create or collect the data. Following naturally on from storing data, is finding it at a later date. Isobel emphasised the importance of having a sensible file/folder setup system to enable data to be found at a later date, including the use of metadata tags for files, incorporating file versioning in file names, using sortable date formats: (Year, Month, Day) and potentially including version control tables and document registers if required.

When it is time to share data, this will often involve depositing data in a repository and receiving a DOI (Digital Object Identifier) that will link to where the data is stored. DOI's don't have a large amount of required metadata (title, authors, abstract) but Isobel emphasised the importance of including a README file that provides a lot more information about the data. The talk then touches on the notion of interoperable metadata, and how important it is

to use structured machine readable metadata for your data.

Finally, Isobel concludes by warning that even though metadata is incredibly useful, the use of metadata alone is not enough to preserve your data going forward. You also need to consider the file formats you are using for long term storage, and ensure that these files are preserved digitally.

Tip 1: Make a Data Management Plan!
Often your university will have their own resources, but failing that, the DMPOnline Website [2] contains basic templates for data management, and has information on the different requirements from the different funders. It is always a good idea to create these as they help you think about how to manage your data throughout the whole project. Additionally, most of the UK Research Councils require a data management plan as part of your research applications, and even though EPSRC doesn't, they require it as an early deliverable.

Tip 2: Identify significant data for short and long term storage
You need to identify what data should be stored in the short term and what needs to be retained for the longer term. It isn't always necessary to store all data long term e.g. streaming data or training data (if your algorithm is the primary tool you are developing). Further, consider how expensive the data storage is vs how easy the data is to recreate. Data that is expensive to store but easy to recreate doesn't always need to be stored, but in that instance you would require very detailed records for replication.

Tip 3: Use the 3-2-1 Rule for your Data Storage
You should never assume that having a physical copy is enough. You should store 3 copies of data, within 2 types of media and 1 of these should be stored on a separate site. This ensures that you have covered all of your bases and means that if one of the media types fails, or you lose a copy, there are still alternative backups. NB: If you decide to store your data in the cloud, remember to check out what kind of backup systems are being implemented, as it may not be as robust as you think!

Tip 4: USB's are NOT a storage solution
They are good for transferring data, but they should not be used as a primary storage method. They have a tendency to fail and are also easy to lose!

Tip 5: Use sensible folder/file structures
Ultimately the first person you are sharing your data with is your future self, so you need to figure out the most sensible way of storing your data for you. There is no one true folder structure, but there are some things you can implement to make this process easier. Use a sensible folder structure with some hierarchy, and have a system for file names that is both human readable and meaningful. Also, remember to keep a document of any abbreviations that you use as part of this system, and be consistent with them! Further, you can keep a document register to lay out where your documents are stored[1].

Tip 6: Version Control your work!
In addition to backing up your work, version control it so you can go back to previous versions if you need to. There are a variety of ways to do file versioning, either through the use of specific software e.g. GitHub, or simply by including version numbers in file names and creating new files for each version. You can also create version control tables as part of your documents, these are mainly used in health research, ethics and business management, but may be beneficial if

[1] An example of a document register can be found at 20:07 in Isobel's video: https://youtu.be/bvD7k5xKnXo

you are working on a document with other people as they provide metadata information but also a running log of the changes made to the document[2].

Tip 7: Include README Files in your DOIs
When you deposit your data and receive a DOI, you should ALWAYS include a README file to go alongside this data to provide additional information such as information on the researchers, date, location, restrictions, licenses, methodology, links to publications, and any equipment used[3]. NB: Following all of the previous tips about planning your data management, keeping a list of your files and abbreviations, and using sensible structures will all be immensely helpful when you get to this stage.

Tip 8: Metadata alone will not future-proof your data
Metadata is incredibly useful, but it is not enough by itself. In order to ensure that your data is accessible in the future you need to think about what file formats the data is in; for example if your data is in a proprietary format can you export it out to a non proprietary format that everyone can access and that can be used going forward. To achieve this, it is worth capturing data in text files as well (e.g. CSV/XML/JSON free-form text) as this will ensure that the data will remain usable even if the formatting is lost. Alternatively, use file formats with openly published specifications.

[2] An example of a version control table can be found at 17:26 in Isobel's video: https://youtu.be/bvD7k5xKnXo
[3] A fuller list of information to include can be found at 25:46 in Isobel's video: https://youtu.be/bvD7k5xKnXo

4.2 Pitfalls and Gotcha's with bioactivity data - Professor John Overington (Medicines Discovery Catapult)

https://orcid.org/0000-0002-5859-1064

Figure 2: Professor John Overington

The full video of John's talk can be viewed here: https://youtu.be/EoZfoQlAKqo [3]

Professor John Overington is currently the Chief Informatics Officer at the Medicines Discovery Catapult, a not-for-profit national facility connecting the UK community to accelerate innovative drug discovery. He holds degrees in Chemistry and a PhD in comparative protein modelling, drug design, and sequence-structure relationships. John has extensive experience in Molecular Informatics and modelling having held positions in many institutions including InPharmatica, EMBL-EBI and BenevolentAI.

John's talk focused on his experiences working with bioactivity data and drug discovery research along with some of the problems and errors that people have encountered when working in this sphere. In the past researchers could reasonably know 'most' of the research within a field, but now we have much larger scale research, more participants and more data but without a lot of the groundwork being laid for good data sharing and reusability. Now there is a lot of messy data out there; inaccessible data, cryptic data and poorly described data. John talks about some of the bioactivity resources that are available for researchers and some of the successes that these data sources have had when handling large amounts of data. John gives plenty of tips about things to look out for when examining chemical and biological data, with plenty of examples.

Tip 1: Be wary when using Excel
Excel has a number of issues that can be encountered when running data analysis. Prime examples of these include data values (e.g. Genes [4], compounds etc.) being irreversibly converted into dates or currencies through its smart data handling.

Tip 2: Data Archiving is very important
Data doesn't live forever [5]. Keeping good data in a format that allows you to look back at it after some time will help you immensely in your studies / research. Electronic lab notebooks can be a really good investment in your research.

Tip 3: There should always be a degree of error associated with large scale chemical databases
The estimated errors in bioactivity databases do vary but the complexity of the data, in particular chemical structures, can cause relatively high errors and inconsistencies across all data-

bases [6]. Many of these structural errors are concerned with stereochemistry, which may not affect all cheminformatics approaches as many of these are stereochemistry insensitive. The error rate of structural depictions in publications is estimated around 3.5-4% and where databases extract data from publications these errors can be carried over. So an error estimate of around 3-4% is probably what should be associated with large scale chemical databases and analyses.

Tip 4: Trust but verify the data that you use
You can find lots of articles and commentaries on the reproducibility crisis with biomedical research and related fields, including examples such as Begley & Lee's commentary in Nature [7] and Prinz et al's correspondence in NRDD commenting on the reproducibility of pre-clinical data [8]. However, months later there was a Clarification to Begley's commentary [9], which is as close to a retraction as you can get, as the work presented in that commentary was in itself not reproducible. So while things might not be quite as bad as they are made out to be, always verify the data that you use.

Tip 5: Validate your data against other sources where you can
The quality of primary biological / pharmacological data can be very variable, even experiencing 20-50 fold variations. The inter-lab variance in assay measurements can also be quite significant, in some cases even producing a variance similar to inter species variance [10]. Cell based data tends to be even more variable than biochemical data. So if you have alternative sources for your data then try and check against them.

Tip 6: Chemical structures are not always correct
As well as incorrect stereochemistry in chemical structures there have been numerous examples of errors within vendor supply chains and information provision, in particular Bosutinib and Voxtalisib [11] . Possibly even as many as 1-1.5% of compounds sold by vendors are incorrect. Another case in which cross checking the data is very useful.

Tip 7: Be aware of unit errors
A frequent error that can occur is the thousand fold error, e.g. units are put in as micro instead of nano molar or a similar mismatch of units. This error can occur due in the publication of data or in the data capture. Potentially as much as 1-2% of quantitative bioactivity data in literature, atleast captured in ChEMBL, has this thousand fold offset.

Data Resource 1: ChEMBL
ChEMBL is a great reference source for bioactivity data, this database is open and this data is shared and utilised in many other resource databases. This database captures lots of medicinal chemistry, pharmacology and drug discovery data from the literature. ChEMBL is truly Open Data and is the basis for a vast amount of AI research in compound design / optimisation.

ChEMBL is extracted from the literature, the usual data includes; compound data, chemical structures, bioactivity endpoints and associated assays. The data is marked up to attempt to add rich semantic tagging to the literature data. But ChEMBL has contained many errors of different types and John reflected on some of the types of error that might be encountered:

- Chemical structures in papers are not always depicted in an easily extractable form.
- The process was manually done causing transcription errors to occur in data extraction and curation.
- Mistakes are present in publications and experimental data which get propagated into ChEMBL.

Data Resource 2: SureChEMBL
SureChEMBL is a public domain chemistry patent resource, donated into the public domain by Digital Science. This is not a manual database but is automatically extracted from full-text patents. The generation of this resource relies on text mining and automatic structure extraction. This is updated daily and has a data download option. You can build systems on top of this chemistry data download.

Data Resource 3: UniChem
UniChem is a chemical integration service, really useful for handling chemical information data and the cross referencing across different databases.The API / web service was developed to do live chemical structure lookup using InChI keys. It covers $>$ 144 million structures across $\sim$30 sources and can provide real time integration across services.

4.3 Digitising your Chemistry for Recordability, Shareability and Reproducibility - Dr Mark Warne (DeepMatter)

Figure 3: Dr Mark Warne

The full video of Mark's talk can be viewed here: https://youtu.be/BLf-M4xWITM [12]

Dr Mark Warne is the Chief Executive Officer of DeepMatter Group plc, a position he has held since July 2018. DeepMatter is a company focused on digitising chemistry across both hardware and software solutions. Mark has extensive experience in the life sciences sector having held several positions in intellectual property and drug discovery companies as well as holding degrees in Computational Chemistry, Colloid Science and Chemistry.

Mark's talk focused on three specific areas as to how you can digitise your data and workflows to improve your productivity, increase discovery of your data and make your research more reproducible. These tips were broken down into smaller areas in which you could implement them, with examples taken from the chemistry and life sciences domains. The three main areas which Mark included in his tips were: binning the old fashioned write up, collecting data throughout the whole experiment and sharing your data in accessible and transferable formats. The talk gave examples using the Digital Glassware^{TM} products offered by DeepMatter, in addition to ways to incorporate the tips in different systems. Mark concludes his talk by commenting on the large proportion of science that is currently irreproducible and the ways in which human interaction introduces opportunities for error. These tips aim to resolve these issues, increasing the reproducibility of science and reducing the errors.

Tip 1: Bin the old fashioned write-up. Write first!
Record your intent before the experiment. This doesn't mean that you can't change what you are doing, but it will aid in your notetaking. Providing your intent and context within your writeup provides you with a really strong base for doing your retrospective analysis. Make sure you store your information on all experiments, regardless of their outcome. It is just as important to know when a reaction failed as when it succeeds. When you come to do your analysis this means you have a repository of negative data as well as positive data. Codify your chemistry, move away from sketch and text to a more defined structure such as XML for your experiment plans. These experiment plans (digital protocols) are ideally stored in the cloud and can easily be accessed and shared.[4] Codifying your experiments is particularly useful when trying to do any form of automation or machine learning on your experimental data. It allows you to bridge the gap between human and machine-readable procedures.[5]

[4]Sample data can be found at https://public.deepmatter.tech/

[5]An example using codified experiments can be found at 10:35 in Mark's video: https://youtu.be/BLf-M4xWITM

Tip 2: Collect data throughout the whole experiment, not just the end point!
Collecting more data can provide you with a much richer data set to examine. But you should collect data from multiple feeds rather than a single source to provide you with an array of data across the time course of the reaction. In addition to traditional measurements, such as temperature and spectra, combine these with feeds such as video or sounds. Annotate your experimental records with observations and photos by the user. This richer data means you can better understand your reaction, what happened in it and potentially intervene or adapt your experimentation in real time. Rather than having a failed reaction, you can potentially save your reagents, or at least your time.[6]

Tip 3: Sharing your data in an easily accessible and transferable format
If we all had access to a wider repository of data with trusted data from other researchers then we would be able to learn much better as a community. Make it easy to share data with your colleagues. Having cloud storage will allow you to access your data from anywhere you need it, whenever you need it. Make sure it is in a secure and well structured stored format.[7]

Questions on presentation:

Q: Have you experienced cultural difficulties in trying to change people to writing first? And how have you overcome these?
A: *It is definitely a cultural shift and it does take time to adjust, but the changes are largely being driven in industry by health & safety and the incorporation of robotics. Making the system easy to interact with and reducing the number of times information has to be entered eases the barriers to adoption.*

Q: Arguably some of the skill of a chemist is to change the approach as an experiment develops; why not just let the technology track the process rather than requiring it to be defined in advance?
A: *A recipe is what you intend to do, but there is no requirement to exactly follow it. You can then compare your experimental run to your initial plan giving you additional insight and allowing you to adjust your recipe for the future if necessary. Additionally, it is always easier to record against a template, or plan, and it provides more context that raw data recording alone may not be able to provide.*

[6] An example using data collection can be found from 14:50 in Mark's video: https://youtu.be/BLf-M4xWITM
[7] An example using data sharing can be found from 18:45 in Mark's video: https://youtu.be/BLf-M4xWITM

5 Panel Session

Following on from the talks an ECR panel was convened, the panel members comprising of 6 ECRs who are either undertaking a PhD or have recently graduated and are carrying out Post-doctoral research.

The panel members were:

- Dr James Cumby - University of Edinburgh
- Dr Paul Dingwall - Queens University Belfast
- Dr Ella Gale - University of Bristol
- Dr Grant Hill - University of Sheffield
- Dr Jennifer Hiscock - University of Kent
- Ms Laura Powell - University of Southampton

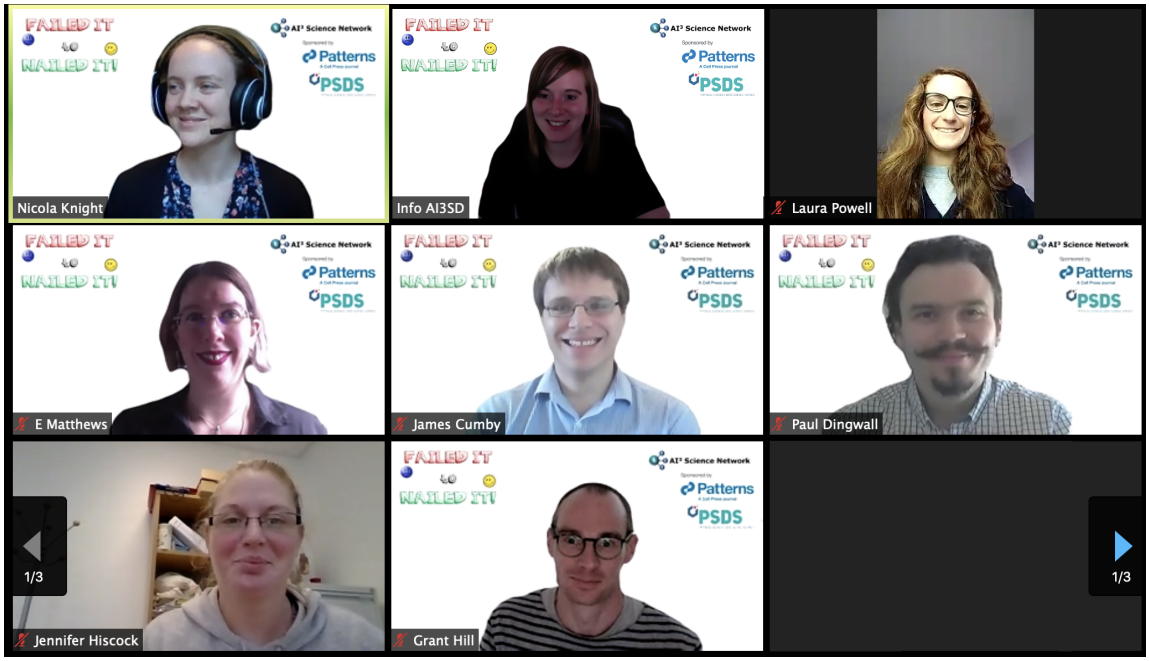

Figure 4: The ECR Panel members

The ECR panel was chaired by Dr Nicola Knight & Dr Samantha Kanza. The questions asked to the panel members were a mixture of pre-prepared questions and questions asked by members of the audience. The report content below outlines the questions asked and summarises the discussion and responses that followed these questions.

Q1: If you could go back and give yourself one piece of advice about managing your data, what would it be?
The main pieces of advice given here are predominantly centered around advance organisation, evolving strategies, and sensible data management.

- **Think about it at the start:** Consider the data, the files, the metadata, who do you want to share it with which might determine your file formats, how are you going to share it? Just for you? Or online with others? This will have different requirements.
- **Have an evolving strategy:** Projects can rapidly evolve and expand and this needs to be managed. Have an initial strategy but make sure to review it so it can expand to meet your needs.

- **Set out protocols and make your intentions clear:** Think of this as a Hansel and Gretel approach, leave your future selves breadcrumbs to follow. You will thank yourself later.
- **Create data dictionaries:** When you go back through an old lab book, it would be good to know what you have done and what abbreviations or codes you have made.
- **Don't be afraid to rename:** No naming / filing system is that important, if you feel like you could benefit from going back and renaming things more sensibly, don't be afraid to do that.

Q2: When adopting different data management methods, e.g. lab notebook or system. Have you had to convince any colleagues or supervisors to change their ways and do you have any advice on good ways to go about this?
There was a range of advice on how best to approach this:

- **Lead by example:** People won't necessarily change their opinions quickly, but you can lead by example. If you are working on a paper draft and you are sending drafts back and fourth, some collaborators will use the same name and that causes issues. If you send it back with a version number then they start to take on that approach.
- **Start with you:** Make it understandable for you first and then you can help others, as you can't help others until you have done it for yourself. Be clear and direct with them and remind them that this is about preparing for the future.
- **Do the leg work yourself:** It might be a lot of effort to set these procedures up, but if you can put in that time then it will be easier to get others on board.
- **Catch people early:** We are working with some new kit, and this has made it easier to implement certain tools and methods early on. Changing people's minds later on can be much trickier.
- **Make it attractive:** Create systems that people want to adopt and be part of that culture! Show people why they need to start adopting it!
- **It may take time:** There has been an ongoing battle to use a collaborative tool across our group, which given that it's a small group should have been easy but it wasn't. However, the global pandemic demonstrated this need and was the push needed to start the widespread usage of these tools.

Q3: The pandemic and restrictions have accelerated some of the transitions to digital technologies, which are likely to be focused around digitising notes, but have you seen any changes in the laboratory space?
Our panelists had noticed and indeed been part of implementing some key changes in the laboratory:

- **Reduction in paper forms:** It was typical prior to covid to use a paper COSHH form, but with the current pandemic I don't want to touch a piece of paper that other people have touched! So I have now put it in my OneNote Lab book that they can sign. I am encouraging the project students to use OneNote and lab based tools so they reduce their paper usage and their work can be accessible from home.
- **Improvement in data procedures:** What I've noticed is particularly in terms of facilities, the procedures within the universities things weren't very clear. However, now that we have had to implement strict covid procedures, and people aren't allowed to touch certain things, there are a lot of systems in place to make sure that the data is well labelled, well archived and accessible from anywhere. This will have a long term benefit to everyone.
- **ELNs are useful:** This is more of a comment than an answer, but to weigh in on the ELN side of things, these are brilliant for experimentalists. If you are away from the

laboratory then you can access your lab notebook remotely and see the exact details of what you did and when you did it. This can be very useful in situations such as being at a conference (when they were allowed!) and being able to locate exact details about your work for any questions.

Q4: There is a belief in (some of) the elder generation of chemists/researchers that 'all this data management stuff' is being done by the younger generation and they are fixing it. Either that or they are devolving responsibility. To what extent do you think this is actually the case (I believe most ECRs are 'brainwashed into the group way of doing things' at a very early stage - and don't therefore get on the data management soapbox?

The overall responses to this suggested that the panelists and indeed audience members didn't necessarily agree with the belief that data management is done purely by the younger generation, and also addressed some general issues with being the person to handle data management within a project.

- **Age isn't a criterion:** I do know an emeritus professor who is trying really hard to use teams, so maybe we need to be careful of assumptions. There's too much generalisation about age, it's about your experience and whether technology has been put in your science backpack. Some younger students know how to use snapchat and facebook but don't necessarily know how to use Excel or Word. Really good supervisors allow their students to explore different ways of managing their data. You need a level of curiosity about technology. People just want to get their hands on it and play with it, and that curiosity enables them to figure out how to use it.
- **It just depends:** I think this depends a bit on the supervisor. I had a student who couldn't handle watching me try and manage my data, they sorted it out for me and showed me how to do it and now I do use their methods. This does have a pay off, bigger projects can work better because these systems are in place.
- **This doesn't come with much reward:** People should be doing this, but there is little recognition for doing this either among the group or with the wider community.
- **It isn't always the top priority:** Unfortunately in a lot of cases, if data management hasn't been made a priority by the funders or project leaders than it remains a secondary concern.
- **Tools are often written for specialists:** We should take learnings from the tech industry to lower the barrier to entry to data input, and that will result in better data output.

Q5: When you go to publish or share your data, what are the issues you encounter?

Our panelists and audience members raised a range of different issues that they had encountered:

- **Lack of appropriate storage/access:** We have a university data repository that you can deposit data into when you publish, unfortunately it's not quite big enough if you are doing big data and machine learning. Even more unfortunately myself and the postdocs do not have access to it, the account belongs to our PI who doesn't know how to use it, which means our data doesn't get shared. For the last two papers I have published I have used other storage solutions which could be less permanent, which isn't the best solution. Moving forward you could consider using TenserFlow or Kaggle.
- **Time:** Data is important, but the time it takes to collect the data is also really important. If you come to write up an experiment and you're missing something vital then that can be a real issue. All of the tools that we are discussing here can help you save time.
- **How much to share for your discipline:** Coming from a crystallography point of view, the issue I face is how much to share? Should we share everything or just what is

relevant? We work with big data, as you can easily generate a few terabytes per synchrotron experiment, so it isn't that feasible just to upload it somewhere. Crystallography has standard file format for its results data, and in some senses has a standard format for most of the raw data but that hasn't been widely adopted like the results data. There is CIF (Crystallographic Information File), which is great but it doesn't cover everything. There are ways in which you can describe some of your raw data within a CIF but it is generally not done. All the instruments out there don't produce raw data in a particularly usable data format. Crystallographers are good at making things look good on the surface, but under the surface there are issues. In the past it was typical to just share the process data but then you can miss things. With modern crystallography experiments where you can see some of the nuances in the raw data in what we call in between the spots, that previously we'd have thrown away, but that are actually really important. We now know it's really important and need to go back and collect some of that missing data that we did throw away. The European Synchrotron Radiation Facility (ESRF) and Diamond have data policies - they are keeping all the raw data coming off all their beam lines. The Free Electron Lasers of the future are even bigger data generators, and they are confident they will be able to store everything.

- **Requirement for Further Digital Tools:** There has been a movement for ELNs in the supramolecular chemistry community to record lab data. In this community, we spend a lot of time working out the strength of molecular interactions between a host and guest species, or self associative strength of those non-covalent interactions there and this used to be published as somebody would take the maths equations on origin or excel and try and fit them and then you'd end up with numbers in a table (sometimes with a graph). Then bindfit was created, this is an online platform that allows you to input your raw data points and then save the data via a URL. You can send this to others, and it is possible to access other users' data via their URLs, but you cannot alter other users' data. This is a really good step forward as it allows you to get at some of the intricacies of the data. This would be a very good data source for an AI/ML Project!

Q6: Where have you been able to find resources to help you with your data management and sharing?
Here is a collection of the resources mentioned by our panelists:

- **University Resources:** There are often more resources at your university than you realise. Our university has great resources in the library, we have great access to people who know how to manage data and manage the tools for managing data. Some universities also have groups to help with software and offer courses or sessions on how to use and create software.
- **The Turing Way Website**[8]**:** This website is centered towards data science and informatics. It has some good tips on reproducible research, and is a very accessible website for people who might not be as technical.
- **Ask Google & People:** When I had to write a data management plan I didn't know what to do. I tried to google other data management plans and asked people. From this I learnt how to actually do these. This might seem like common sense but if you are just starting on something, this is a really good place to find the very basic levels to get you going. I also learnt a lot from reading and writing data management plans, especially working with industry collaborators on them.

[8] https://the-turing-way.netlify.app/welcome.html

Q7: What is your biggest data management disaster, and what did you learn from it? Or What is your biggest data triumph?
Our panelists recalled a number of triumphs and disasters:

Triumphs:

- **Averted data loss disaster:** At the end of my PhD I ran out of funding so I went home and wrote up from home. Everything was backed up from my work PC to my home PC every night. I handed in my PhD and went travelling. When I came back to write some papers on it a year later I found out that the University had given my computer to a new student who had wiped everything on the computer which backed up to mine. Thankfully I had it backed up to various other places so I was able to get that data back. Otherwise I wouldn't have been able to write those papers.

Disasters:

- **Small Disaster Story:** This wasn't a huge disaster but I did learn a lesson from it and it has stayed with me. When I was working as a placement student, I wanted to move some files, tried copy and paste, but ended up doing a cut and then another cut and lost all the files. This was right at the start of my studentship so it didn't really make too much of an effect. But the fact that I had lost that level of data did stay with me and now I am very careful!
- **Large Disaster Story:** I've got a large disaster story. Back in the first year of my PhD, I was working on a supercomputer with linux for the first time. I had gotten to the stage 6 months in where I thought I understood what I was doing. I had run lots of calculations, and tried to aggregate the data. I put in a find command and mistyped it and ended up deleting all of the data. . . .! Unfortunately the backup policy on the server was not fast enough to keep the data from the past. The tools are there to do quite exciting things but they need to be used properly and with caution. The moral of this story is, make sure you check your commands before executing them!

Q8: Does anyone have any final tips?

- **Never be afraid to ask:** Never be afraid to step out of your discipline or worry that what you say will be stupid. We all have to start somewhere and we will make mistakes. Acknowledge that and try and move past it. Have a coffee with someone and ask them for help (socially distanced).
- **Have plain text copies of your data:** As noted by Isobel's talk, having your data in a plain text form can be very useful. If you end up losing your accounts (e.g. Mendeley once you lose your student accounts) then that can be catastrophic.
- **Consider available cloud options:** People have been talking about data analysis; the idea is to make this easier for the users to experience rather than needing to be a tech or data science specialist. Encourage people to think about the cloud options that are available. Don't be afraid of this, there are clever approaches to live data and archiving and don't hesitate to move away from the old fashioned methods.
- **The narrative is important:** I want to reiterate that having the narrative that goes along with the data is really important. The negative results and what you think might be a negative result can be impacted by other factors, such as if you are having a bad day.
- **Negative results are still results:** There is a stigma attached to publishing negative results, and as such a lot of the data that doesn't get digitised are experiments that didn't work or earlier versions of protocols that had problems and have been improved later down the line. However this data is still really useful as it can inform others and stop them from repeating your mistakes. This can also be really helpful to you.

6 Participants

Participants attended from a wide variety of backgrounds due to the online nature of the event. While the majority of attendees were from the UK, there were a number of registrations from other countries.

7 Conclusions

This event captured a variety of viewpoints from across different disciplines to bring together a selection of tips and guidelines for how and why researchers should implement good data management practices in their work. In particular, researchers should think about how their data would be able to be interpreted or reproduced several months or years down the line when they have forgotten the details of an experiment or analysis.

While considerations about data management should ideally be implemented at the beginning of your project it is never too late to adjust these practices or implement new ones if they are not working optimally. We strongly encourage all researchers to revisit their data management practices and see if there are improvements that can be made. The tips included in this report will hopefully help researchers on their journey, but they should not be afraid to reach out and ask for advice. Together we can all work towards data that is better described, easier to understand and more accessible.

8 Related Events

Details of the other events in the Failed it to Nailed it data seminar series can be found here:
https://www.ai3sd.org/ai3sd-online-seminar-series/data-seminar-series-2020/
Each of these events will have video recordings and a report associated with it.

Details of other AI3SD events and events of interest can be found on the AI3SD website events page:
https://www.ai3sd.org/ai3sd-events/
https://www.ai3sd.org/events/events-of-interest/

References

[1] Stark I. AI3SD Video: Love notes to the future: the importance of metadata;. AI3SD, PSDS & Patterns Failed it to Nailed it: Getting Data Sharing Right Seminar Series 2020. 2020. Available from: http://dx.doi.org/10.5258/SOTON/P0067.

[2] Digital Curation Centre (DCC). DMPonline [Internet]; 2020. [cited 2020 Oct 29]. Available from: https://dmponline.dcc.ac.uk.

[3] Overington J. AI3SD Video: Pitfalls and Gotcha's with bioactivity data;. AI3SD, PSDS & Patterns Failed it to Nailed it: Getting Data Sharing Right Seminar Series 2020. 2020. Available from: http://dx.doi.org/10.5258/SOTON/P0068.

[4] Vincent J. Scientists rename human genes to stop Microsoft Excel from misreading them as dates [Internet]; 2020. [cited 2020 Oct 29]. Available from: https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates.

[5] Wren JD. URL decay in MEDLINE—a 4-year follow-up study. Bioinformatics. 2008;24(11):1381–1385. Available from: https://doi.org/10.1093/bioinformatics/btn127.

[6] Tiikkainen P, Bellis L, Light Y, Franke L. Estimating error rates in bioactivity databases. Journal of chemical information and modeling. 2013;53(10):2499–2505. Available from: https://doi.org/10.1021/ci400099q.

[7] Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012;483(7391):531–533. Available from: https://doi.org/10.1038/483531a.

[8] Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery. 2011;10(9):712–712. Available from: https://doi.org/10.1038/nrd3439-c1.

[9] Correction: Editorial note. Nature. 2012;485(7396):41–41. Available from: https://doi.org/10.1038/485041e.

[10] Kruger FA, Overington JP. Global analysis of small molecule binding to related protein targets. PLoS Comput Biol. 2012;8(1):e1002333. Available from: https://doi.org/10.1371/journal.pcbi.1002333.

[11] Halford B. Bosutinib Buyer Beware [Internet]; 2020. [cited 2020 Oct 29]. Available from: https://cen.acs.org/articles/90/i21/Bosutinib-Buyer-Beware.html.

[12] Warne M. AI3SD Video: Digitising your Chemistry for Recordability, Shareabilty and Reproducibility;. AI3SD, PSDS & Patterns Failed it to Nailed it: Getting Data Sharing Right Seminar Series 2020. 2020. Available from: http://dx.doi.org/10.5258/SOTON/P0069.