All right so another aspect of storage that many people use are as we were talking a little bit as some of these cloud storage options. And while many people use cloud storage you should always be cautious when storing research data in the cloud. The first thing you want to do when using alternative cloud storage options are to check out the ownership and claim of ownership with different cloud storage data options. And second is it's often recommended to pick more than one ownership provider with cloud storage solutions in case as many people have often come across is there are cases where cloud storage options the ownership and of your data the owner of the cloud storage there have been bankruptcy issues that people have run across occasionally with cloud storage. Especially if it's a smaller provider. We have had people not necessarily here but different places come across. So in case your cloud storage provider goes down you often have you have more than one provider especially if you're looking outside of the university, it often as a safety solution is recommended.

Another thing to discuss is where your data is stored at different parts of your workflow. And to document where you're storing your data during those different parts of the workflow. Where your datawill be stored during data collection, processing, and analysis is crucial to have figured out at the beginning of your project. It's not something you want to decide during the process but rather to figure it out at the beginning and document that. It's especially important when working as part of a large team. All your folders where data will be stored should be indicated and where that data is stored at the end of the process should be figured out and documented as well. Here is an example of research data that was stored on one device and that device was stolen and all the raw data was lost. This is something that is not as common now but it has still happened. So it's something you just want to keep in mind that you don't want to be storing your data in one place only.

Another thing to consider is to save if you're still using a physical device is to save multiple copies of your data and disperse them geographically. Especially if there's something like environmental disasters. And again if we're not necessarily talking so much about cloud storage but rather physical devices if you're storing. When storing your data there should also be several reasons why you should be there are several reasons why you should be concerned about security. So let's go over those. If you're collecting personal health information you want to make sure you're keeping your data is secure in a way that's HIPAA compliant. So you got to keep in mind compliance with HIPAA. If you're working on patients or patents or commercial data you want to ensure that that too is protected and secure from theft and/or damage. And if you're concerned just about the intellectual property you've worked hard to collect your information and data and you don't want to lose all your hard work. So security is just something very very important to keep in mind.

All right so when thinking about security make sure you're working with IT to understand what they can do for you about your about security. And what extra steps they can help you with to make sure you're keeping your data safe. On your own there are several things that you can do. Add passwords to files or folders to keep them safe. Locking your machine when you walk away to make sure that nobody can just walk up and access your information. Just things that you can just do on a daily basis to make sure that nobody can access your information. Have other researchers who work with your data sign use agreements. These agreements will make sure your collaborators are held accountable for what they can do with your data. Use agreements can also restrict what someone can or cannot do with your data. And when storage is needed for personally identifiable data it's important to consider a number of factors due to HIPPA. From a data management perspective HIPPA indicates that you should have documentation for who is responsible for managing stored data at the original site during transfer, and at the new storage site. And indicated who the authorized users are at the beginning of the project is best to go over during your standard procedures. So just make sure you have all of this documented. So as you're getting here documentation is key with everything we're doing here. Just from storage, security, and with HIPAA requirements. And HIPPA states that PHI data requires end-to-end encryption meaning that data must be encrypted at its original location, the connection that data must travel through to get to its destination must also be encrypted, and the data must remain encrypted at in its other storage location. The encryption key must be stored separately. And keeping HIPAA at on a portable device is not recommended. So keep that all in mind when we're dealing with HIPAA data. And if you're using cloud storage you need to make sure you're following all HIPAA regulations and cloud storage is allowed with HIPAA. Cloud storage is allowed under HIPAA, provided that your covered entity or business associate enters into a HIPAA compliant business agreement with the CSP that will be creating receiving maintaining or transmitting electronic protected health information on its behalf and otherwise complies with HIPAA rules. So while you couldn't use a personal Google Drive for storing that HIPAA information you could however use a work Google Drive if you entered into a BAA with Google Drive for work. So if that was a professional business agreement with that provider. So something like that it just have to depends on what your provider allowed. So learn more about cloud storage and HIPAA we're gonna send out a URL that will go over all of these details.

Alright let's get away from storage and go into a little bit more about preservation which will be a little less technical. And I will jumble a little less. Difference is a little bit are basically it's important to understand the differences first off, storage is mostly about your storage preservation is going to be going over how to make sure your information is not just stored but also accessible long-term. So just because we're storing our data doesn't necessarily mean that you're going to be able to access it long-term. Preservation focuses on making sure your data will be available and available in a way that you can use it in the same way that you were able to collect it. So looking back on your previous research you want it to others to be able to still access it. And as we move forward preservation will be increasingly important with more sharing agreements and sharing requirements emerging from funders and publishers.

So the first thing we're going to talk about is our hardware obsolescence. Preservation helps you helps protect you from hardware obsolescence. And you will want to avoid avoid scenarios where all your data is saved on in an old format like a Jaz drive, only not to be able to open it later because you know Jaz drives are no longer supported. So you always want to migrate to a new hardware format so that your data will be able to be available long term. If you're using a proprietary or specific hardware format have a plan for how you can migrate it to a more universal format. And like hardware obsolescence there's also software obsolescence. So you want to be able to think about how you can save your data to open software formats. If you're using a proprietary software in the lab or homegrown software you created to collect data there may not be a way for others to access it. So moving data to open formats will also protect you from going back to old data opening it only to find that it's unreadable. Because of this it's important to just to make a distinction between how you collect data versus how you will disseminate it. You may still need to collect data in one way using specific hardware or software but you should have a plan to how you will transform it into a format that will be suitable for collection, collaboration, and dissemination.

So the difference between collecting it one way and being able to disseminate it another way. So our best practices for preservation is to save your data on preservation formats. So there's a few gold there are four gold standard preservation formats that we're gonna talk about here. So the first and they can the good thing about these gold standard formats are that they can be used and viewed on basically any operating system with any kind of software. So we've got the XML extensible markup language and this is used to ensure simplicity, generability, and usability across the internet and can be used to save documents or web service content. The CSV (Comma separated values) is an ideal way to save spreadsheets in preservation formats. We've got the PDF and the nice thing about PDFs is they can be used to save documents in perpetuity. The one thing about PDFs to keep in mind is generally you can't really edit them. So you're not really going to be able to go back and edit them but it is way to get a way to capture them in a lasting format. So you won't be able to change them but you can freeze a document in a PDF format. And then you have TIFF - tagged image file format - and it's the gold standard for saving images. And they are already preservation ready format and they can ensure the quality of an image over time. So those are typically four of our gold the gold standard for preservation. format.

So we didn't really talk about this question but looking at this list what would you guys say would be the file format for an open data for documents? You would probably for when looking at a basic document you would probably want to use a .txt file. Because it's just an open it's an open software. That said it's not always possible to use a preservation format so when at all possible use one but if you can't keep in mind that some files are more ready available than others. Microsoft Excel or Microsoft Word is probably going to be around a lot longer than some more obscure lab-created software. So use your best judgment when deciding what to use for preservation format. So if you've created your own homegrown software and you don't know if you're going to be around or you know that you were going off to the middle of nowhere and nobody's going to be able to reach you, but you know your data is gonna have to exist forever, see if you can you know transfer it to a format that other people will be able to access. Like Microsoft Excel if that's the only thing that you could use. That said it's not always possible but do your best to find something more stable. Chances are Excel, Word, Microsoft Office suite is going to be around. We're all gonna have problems if they disappear in the near future. Remember if your if your data is irreplaceable and you can't collect it again or it take took years to collect, it's best to make sure you're saving it in preservation format so it's available and accessible for a long-term. So just kind of think about what you're collecting, figure out you know figure out what it means to you when you are trying to preserve it and make those decisions for yourself.

Other ways to protect your data from degradation and ensure its preservation is to try to avoid encrypting or compressing your data when possible. Now this is not always possible. Encryption is useful for keeping your data safe but encrypting data can also make it difficult to access your data later on if you haven't kept good documentation. Key word: documentation another word that's coming up a lot here on how to access your encrypted data. When compressing your data you risk damaging your data each time as it can remove the quality and integrity of your data over time. Always keep an original data set, make a copy, and then compress or encrypt that data to send to others.

Finally with preservation it's important to know who owns the data you've collected and you can't assume that you own your data. So before you share it, collaborate, or delete data it's important to check funder or institutional policies to understand your user rights. Now we're gonna pass back on to Tom to talk a little bit about providing access.

And it's the final stretch. I promise to make this quick. I'm just curious does anybody know what the actual university policy is for data ownership? They own it. If you go to the research homepage they do have a policy there. And so that's important to know if you're going to be leaving to go somewhere else and you cannot automatically assume you will be able to take your data with you. You probably will. I don't think it's hard and fast. But the basic policy is that the university does own data produced using their facilities. So it's a good thing to know.

Ok access and standards. Again access is the issue here. Can be it's one thing to store it, it's another thing that people can actually get to it. So where are you going to put it? Most people choose repositories. There's two basic types of repositories. There's a cross-disciplinary adoptable repository you can put anything in there it is not subject specific. We'll talk about a couple of these that are very popular. Probably more in interest to this group would be discipline specific repositories. Dealing with specific areas of research. The nice thing about that is other people who are doing the research probably know to go there to get the data. So it makes it a little bit easier to find for other people doing the same kind of research. One of the nice things is the NIH has a nice list on their webpage of data repositories dealing with medical research issues. If you just go to their webpage do a keyword search repositories you'll get this list. Also the slideshows will be available so you can see the link. So you can see they do a really good job there's over 60 listed here. And they get very specific you've got Cancer Nanotechnology Library, the Cancer Imaging Archive. All sorts of things. i-Genes, peptides...And it will give you the repository name, a description of what's in there. They will give you the submission policies for how to put it in what they require when you do deposit data. And then how to access the data. So this would probably be the most interesting place for you guys to go. We will talk about some other options but just keep in mind the NIH has kind of done this for you. See if there's a place that's appropriate for your research data on their list.

R3data.org is a very popular research data repository. I was actually gonna bring up their page but since we're having trouble showing other web pages now I'm not going to do that. There is a link to our re3data on my research guide. Hopefully when we're done here I can get that page up but I don't want to interrupt the slideshow right now. Re3data is not subject specific. They have their own metadata standards. Very easy to use they are fee based. The feed the varies depending upon how much data you're inputting. They've got policies on their home page you can look at. Also OpenDoar the directory of open access repositories. They've got a open-access, non discipline-specific. So these are two just popular places you can google them up, or get there the links off the slide here.

Figshare. Figshare's another open repository. The good thing about Figshare is it's a great place for open source code. They do a lot of other deposit other datasets there but it tends to be a very popular site for open source code. So as Sarah mentioned earlier, if you can use open source software for your data manipulation, data analysis. a final product whatever, that always helps other people who want to use the data later on because they will probably be able to get to the software easier Figshare is a great place for that. They're very good at keeping things updated when a new (whoops sorry) when a new version of the software gets uploaded, it will be immediately there will be a notice about it. So a good place to keep in mind.

And again the issue here is access versus meaningful access. You put it up somewhere that's great. It's somewhere how do people get to it? Meaningful access means that you've got some metadata that will be encoded with the file when you upload it. So that it is documented and people can get to it. Most repository sites do have metadata standards on the site that will usually be a very obvious link. You can just go there and see what metadata standards they offer. Usually it will be discipline-specific. There are also different disciplines specific metadata standard sites. There is a link to that on my research guide which hopefully I can show you later. And it's the kind of thing where you can just google it up very easily. Metadata standards biochemistry whatever. And you'll find those sites. So the point is try to get some well-documented metadata associated with your file so that other people can find it. A good example of people who are really pretty stringent about their metadata standards is the NIH National Heart Lung and Blood Institute. The data repository actually started back in 1975 so they've been doing this for a long time. Their guidelines state and this is a quote: "The documentation must be sufficiently complete such that a person responsible for determining use and selection of biospecimens could tell how the study was conducted, how the specimens were prepared and stored, and how associated data was collected." So they'll have a list here of the study names. You can just bring those up. They're pretty strict about these standards and they're so strict that they've actually got a biorepository guide to building biospecimen collection so they will tell you exactly what they want to have. Basically their standard is that somebody who is not associated with the experiment can go find the data as easy as someone who's associated with the experiment. This is the only one I know where they actually have a guide for this. I'm assuming again that this will become more common and it will be research area dependable as well.

Tools. Tools you can use to help with your data management. The biggest most popular one is the DMP tool. Anybody actually use this yet, I'm curious? Okay, wonderful. So this has been up for about four or five years. It is basically just an online tool which will walk you through the process of creating a data management plan. One of the nice things about it is it will have links to the funding agencies so you can find out what they require. It will have templates for each funding agency, so if you're doing NIH or NSF or whatever and sometimes there's sub-disciplines in there NSF has a lot of sub disciplines for different areas of engineering. Because they might require different data elements and different information in their data management plan. So you'll be able to go find out exactly what the funding agency that you're dealing with requires. They will have a template there when you bring the template up they will walk you through the process of filling out each field. You can cut and paste you can share your account with other people. If you are creating a data management plan with someone else on the team you can add them to there, you can both review it at the same time. They will have previously published data management plans that people have agreed to make public. So if you want to see what someone else looks like for a certain field you can go see what they did. Again the important thing to remember about a DMP is it does not have to be long. In fact they usually want it short. Two pages or less. It's just important that you provide the elements of information that they are interested in having. You wanna make sure that when you don't want to have I actually had a call from somebody over in the Materials Research Center that had (or the Energy Conversion Center I'm sorry) who had submitted it had submitted a grant application did not supply the DMP. They rejected it. Sent it back - I got a panicked call "What do I do?" Took two days to fix the problem, no big deal but you do not want to have be in that situation. This is available off of my research guide, once again I hope I can show you. If you just go to that link as long as you're on a University of Delaware campus network - so you're on the UD network - when you click on that get started link it will then bring up a drop-down menu of all the participating institutions. University of Delaware will be one of those on there. Simply select it say "ok". It will then throw you into your CAS-authentification. That's how you get in. Once you're in you've got an account created, you can start creating plans, you can save them as you go along. They're very good at updating when new templates came out. You'll see there's a new template for what DOT. And NASA funded research. So they'll have News, they'll have old DMPs you can look at, they'll have the requirements by agency and some private institutions as well. Private grant funders. Again if you need to find out what the requirements are for a DNP for an agency or public funder that you are dealing with and you cannot find it contact us we'll see if we can help track it down. This is just an example of a new template notification that they put up for the genomic data sharing policy. Again they're very good on the DMP tool about keeping things very up to date. We've had no complaints about that.

Github. Github is another good place to go for actually, Github's another place for software. As long as well as Figshare. In fact, Github is probably better. So the nice thing about Github is they will give you notifications once things gets updated. It will tell you how to upload things, how to share things. This would be a notice of something that one changed file, two additions, two deletions. They're very good at letting you know about that. There's also links to Figshare and Github off of my research guide. A good place to go for open source software.

Lab notebooks. If you do not want to use traditional lab notebooks. This is another interesting thing I found out from the from the civil engineering people, a lot of them are still using that but a lot of people don't like having paper in the lab anymore. You might spill on it might do things. So there are some online lab notebooks "lab archives", which is pretty much a general-purpose site. "Labguru" which does really inventory management for equipment and processes. And "Benchling" which is more specifically for DNA tool integration. There's other ones available these just tend to be the most popular. Again if you don't feel like keeping traditional lab notebooks… Another interesting thing I noticed with the civil engineering people is, and I'm not picking on them they're just people I just did this project with so I know about it, a lot of them just have a little file cabinets stacked up with this stuff, you know. That's saving it in a way, but if you want to go online there are options so just be aware of that.

So what have we concluded? Plan your data management plan before starting the research. When we do the general workshops this is something we talked about at the very beginning and we actually have a slide of the data management cycle. It's very easy to think that you don't have to think about these issues till the end. Oh no I've done the research and I've got this data, and where am I going to put it? It's a really good idea to start planning this before your grant. Start thinking how what am I going to do? What am I going to need? What are the funding requirements going to be? Where am I thinking of publishing? What what might their requirements be? If you have a team, again assign somebody to be in charge of all this so you're not wondering who's taking care of things. So take care of it throughout the research process. It's not just something you need to be thinking about at the end. And do not ignore the march toward more data management and sharing both from grant providers and publishers. That is it. I really want to thank you for letting us come down here.