Evaluating GIBS Technology for Research

DRAFT 1

Prepared for

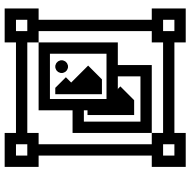
Lewis Mcgibbney
Charles Thompson
Chris Mattmann
NASA Jet Propulsion Laboratory

By

Andy Chase
Daniel Kirkpatrick
Kazuriah Buckley
Oregon State University

November 2015 - May 2016

Abstract: We investigated the use of GIBS to assist researchers in their work. We interviewed 8 researchers at the OSU College of Earth, Ocean, and Atmospheric Sciences. In conclusion, we found that they wouldn't find it useful it for research or data collection.



ARIA

Oregon State
UNIVERSITY

Contents

Part I

Meetings with the Researchers

Chapter 1

Meeting with Dr. Dudley Chelton

These are just notes; this is not a transcript.

Dr. Chelton have a seminar before the meeting which he suggested I go to.

Notes from the Seminar

Sea Surface Height [SSH]/Altimeter data data used mostly. SWOT used along with QuikSCAT/SeaWinds was the data source.

Because of noise, data has to be filtered. This lowers the spatial resolution.

Notes from the Meeting

Q.1

Question: Are Eddies hard to find? Can they be found using Sea Surface Temperature [SST] imagery for example?

Eddies can't be found using imagery alone. Maybe some of the really big ones but SST is influenced by so many factors including wind, sun, etc. Eddies cause SST differences, not the other way around. We use [...] altimeter data in combination with other things to find and track Eddies.

It took about a year to make the model we use to find Eddies and then we improved it over the course of seven years.

Q.2

Question: What are some other phenomena that don't have a specific location and time that researchers are often looking for?

Oceanic Fronts -- Gradient fields about >10km

Who uses? Fisherman use these because they are where the fishes are at.

Q.3

Questions: Do researchers spend a lot of time looking for things? Would helping [Ocean] researchers find things be beneficial?

Yes, researchers spend years looking for things.

You'd have to pre-determine the shapes ahead of time, but since things in the ocean move around, it would be helpful if you didn't just look for ellipses, because often in front for example ellipses that curve around and keep going.

Chapter 2

Meeting with Dr. Jennifer Hutchings

November 23 Assistant Professor
Sea ice dynamics

Difficulties with Remote Sensing: Uncertainty and Limited Resolutions

Q.1i

First off I want to tell you a bit about my project so you know where I'm coming from. I'm working with NASA JPL to help build a project to help earth researchers and specifically we're trying to use remote sensed images to help researchers do their work

[Who are you working with at NASA?] **Well, specifically it's the GIBS team, which is like the Global Imagery Browser Service Team, and they make WorldView Maps basically, and people's names are Lewis McGibbney and Charles Thompson.**

Hm, I think I'm familiar with the data.. but not the people :)

Q.2

Okay, so my first question is what challenges do you think researchers face– ice researchers face using remote sensed images in their research?

Oh, wow, that's a big question. So the biggest challenges is a lot of the remote sensed products– not necessarily the images but the products– are *not* necessarily provided with.. uncertainty...?

Uncertainty? You mean like bounds of error?

Bounds of error and bias. and this is actually a *huge* area of research because on some of the products we don't necessarily know what uncertainties and bias are. Thinking of particular some of the ice things/products that are out there...

Q.3

Like SAR? I saw you used SAR in your research.. the scatterometry?

Yes.. I do , I used– I've used the radarSAT geophysical– processing system? Which [? Ron Clark] worked on– which is looking at Ice Drift. Yeah! one of the biggest problems with that product is that it isn't fully documented. So that can be– a huge challenges when you realize you're working with a dataset and you don't understand what's causing differences between the data you're working on and another dataset.

Q.4

So I'm interested– you said –remote sensing– you don't the images at all in your work?

I do use the images, so typically I use visual or thermal images **[like modis?]**, modis yeah. Normally it's level 1b basically– the calibrated stuff I'm not interested in calibrating my own data [laughing].

Q.5

I noticed you have a Ph.D. in remote sensing is that right?

My Ph.D. advisor was doing remote sensing, I used remote sensing in my thesis, sometimes these titles get mixed up– so the majority of my Ph.D. was in modeling.

Q.6

Ok– I was interested in that because I noticed– some of your latest work is deploying buoys and bots– so you're actually collecting data on the ground [Yeah] **not necessarily using remote sensing.**

Yeah, that's right.

Q.7

I saw that–earlier work you looked at tracking ice using remote sensing– but it wasn't reliable enough to not need buoys– seems like you mostly used that to figure out where to put the buoys.

Yeah, so what happens is with satellite overpasses you only get a snapshot of the ice path– somewhere between every six hours in particular locations, or three days.. depending on the orbits. So I'm interested in sub synoptic scale– **[What's that?]**

which is meaning "below the weather scale", so you know weather synoptic time scales is the time scale a weather system exists over which can be 3 days to a week, or 2 days to a week. A lot of the work I'm doing is at inertial motion of the ice– or tidal motion of the ice pack, which is all on subdaily timescales, and so you can't actually look at that with remote sensing.

So it's nice to– when you're trying to understand how ice deforms and moves you have to look at it on high time resolution as well as high spatial resolution– so that's where we need to blend remote sensing together

Q.8

So how do you blend that?

So– I don't *blend* it right now [laughing] I'll say– That's a high level product that would be quite difficult to make.

Q.9

So you mostly use [remote sensing] for visualization purposes?

Yeah I'll use the imagery for visualization– what the field of drift looks like– and then overplot buoys on that and I use the buoys information for time series analysis. So yeah it's– you end up synthesize different data sets together that are not necessarily compatible with each other.

Q.10

How so? The resolutions are so different?

Yeah. And the errors on the position for remote sensing products is very different. GPS.

Because of clouds?

Actually for tracking ice drift what's happening is you're looking at– feature tracking– on the path so you do that in many different ways it's often cross-correlation analysis between two images to find those features and there's an error associated with that.

Q.11

So you might be tracking the noise, not the ice

So you're hoping you are tracking the ice– and they're normally quite good at doing that, but the position error associated with that can normally be quite large. Yeah,

sorry, I'm sending you in all different directions here. [No no, it's great.] The fundamental thing that it all comes down to is we need to know what the uncertainties are in the products to be able to use them.

Cool, makes sense. [..] How small are tracking?

So with drifting buoys– with GPS I'm limited to not being able to track below about a 1 km

Q.12

So you'd like to track more but you can't?

No, well I *could* if I used differential GPS methods, so you can go down scale if you improve your position error– so there's a fundamental limit on what you can actually track with remote sensing– which is not a bad thing, we just need to know what the limit is.

I should point out, you know a lot of people when they talk about their frustration with using remote sensing imagery will bring up this issue of them not providing error– is not provided with error in some cases– but I actually think for many of the products [error] is a high level research question to identify what those are.

If you're looking for recommendations for what a database should provide, this might be a tricky one.

Sharing Data in Ice Research & Binary Formats

Q.13

I'm just trying to see– in general problems researchers face, but I do think this is really helpful [..]

So another question I had is about the status of sharing science in research. I know I've heard from before that in some disciplines the sharing [of] data is pretty poor– a lot of people are repeating the same data collection– tell me about the status of sharing in ice– I saw that you run an Ice Watch service– was that in response to seeing a lot of fragmented data?

It was! Yeah- actually it wasn't my idea it came out of a workshop in 2009– but the climb to the cryosphere sponsored, and the work *theme* was to coordinate and standardize our measurement methods– and it was for industry field work– specifically.

There's always a need for sharing data and– if you're sharing data that's difficult to collect– it helps to standardize your method for collecting that. That's in relation to this particular field data, it's not necessarily how you would treat remote sensed data.

Q.14

Yeah, I noticed that that's just all on the ground data– you don't combine it all.

Yes, it is and I– at the moment we're all just working on collecting the data and then– the next step is providing it in formats the users need.

Q.15

So right now you're not providing the data for people?

It's free available in .csv format! Yeah, so I'm beta testing a different format that ice charting group use, called seger3– **[*writes it down*]**– yeah don't worry it's all jargon [laughs]

Q.16

Yeah I know– maybe this is more atmospheric and ocean– but I know they use netcdf and hdf formats–but I guess there's a wide range. I know a researcher was talking about a couple days ago that some research groups provide data in binary formats– and they don't say how to use them.

So the problem with binary format is unless you're given a piece of code of how to read it with the algorithm of how to read– and you know whether it's big endian or small endian you can't use it– I've seen providing binary data with no additional metadata as a way of preventing people from using the data.

What do you mean?

It's like– people will– the zero order is they say providing the data– but– if the data is not provided with a way of reading it– no one else can use it so– as you said it's useless.

So if you are a researcher that's protective of their the data– maybe that's what they want. There's two things that could be at work, they could be protective of their data or they might not just have enough time.

Q.17

So have you run into that– have any ships said they didn't want to provide their data?

Yeah, I've run into oil industry ships that didn't provide their data– but said would when they reach the embargo period which is a couple years?

Is that just because that's a trade secret?

Yeah, they just don't want to share where they've been, and that you can respect because they have an embargo period.

What's more frustrating is when you're trying to use data that other people have created and they *do* provide it to you, but you can't actually use it– I've run into *that*.

Q.18

Oh really? So they provide it to you but you can't use it?

It's impossible to use– at the moment. It might just be that more communication is required in order to extract the information that needed, but it's really hard to do that–

I've a big advocate for people providing data in netcdf format. The problem is is that it's expensive to get in netcdf format

Q.19

What is that, just writing the converters?

Well to get anything into a data format takes time, and mad hours. Because you have to write a piece of code to do it.

Code: Students using Python, and Trusting and Sharing Code

Q.20

So what do you work in when you work with data? MATLAB, IDL?

I do use IDL mostly, I have written some C code for processing imagery. I will use MATLAB, my students are using Python.

Q.21

[Excited] Oh, really? That's the first time I've actually heard any reference to a student or teacher using it

Really? It's becoming the standard.

Q.22

I've heard that, but when I talk to researchers they say: Well, I learned MATLAB in grad school and when my students come to me for help, I can only help them in MATLAB..

I'm stuck with IDL because– IDL is like breathing to me it's so easy for me to write in IDL, but.. for my students I would– *never* suggest they learn IDL

Q.23

Really? Because it's old, or not as easy to use?

and it's not free, and I think we're now in a world where free sharing information and algorithms is a good thing to do. It's seems that the Python community is actually growing– so I think Python is the way forward.

Q.24

So it's a good, strong reference on that one

Yeah, the MATLAB community is also good I see shared MATLAB code on the internet.

Q.25

Yeah, I was talking to a researcher that says: [exaggerating] if he's looking for something and it has five stars on the MATLAB central he just uses it, trust it

That's pretty amazing, I wouldn't trust everything.

It all comes down to where you find your code, so I've used R– because there was code available.

Q.26

So going off of that, how do you know if it's giving you the right answer, previous researchers have said if it looks right then they trust it– do you dive into reading it?

When I'm using someone else's code– if it's a researcher I work with I trust it. If I find bugs, we report them– really it's– mostly quality control is logical checks– what you're actually seeing in the data– I don't know I can't really describe how I do my debugging– it's really involved.

You always have to spend time convincing yourself that you're getting the right answer [laughing]. It's a process. It helps to have access to people's code when they have solved problems– yeah I don't know how you trust someone's code.

Q.27

That's just something I'm curious about because I know in studies you don't publish the code you use to get results so no one can really audit that without repeating your whole experiment basically– I know I'm from CS so we spend a LOT of time looking at other people's code and finding bugs– it's just a constant process of finding mistakes

So you know how endemic they are then– they are mistakes everywhere! **[yeah]**

There is a motion in researcher, in geophysiscs, towards having models and code archived– **[as part of the studies?]** yeah, if you publish something you should be able to go back and look at the code. It's just a discussion that's happening now, I haven't seen it really implemented

Q.28

From talking to researcher it sounds like there's a huge pride thing there– were researchers don't necessarily want to publish what they've been working because it's not– when you put something out there it's like you're putting your name attached to it [mhm] **– and so a lot that they put in, there's pressure, deadlines– they just want to get the right answer, they're not concerned with quality–**

You're talking about quality of the code: so how well commented it is, and how usable it is for someone else. I would say that's totally true.

Q.29

and if it works for someone else's parameter space. Even in the open source community I would say people don't want to publish their code unless they've but a lot of time into it.

I think that makes sense. Yeah.

Sharing Data Part 2

Q.30

Going back to the Ice Watch service, would you consider it a success? Or do you think there's still people that aren't really sharing data or are collecting data that already exists?

I think Ice Watch is a success, because people that *are* collecting data are now collecting data in the same format. We've haven't entrained the *whole* Sea-Ice community

yet– but I think it's achievable. I mean that's just one small– case though, one particular type of data **[Because you are mostly looking at- ships?]** It's just ships, it's just visual observations of Sea-Ice upon the bridge of the ship, so...

Q.31

So what happens to the Buoy data that you put out with Alison Kohout from New Zealand, does that get shared at all?

Ah, so the stuff that Allie did– She shared her data with me I'm not actually sure what she does with her data. So there's two different database for Sea-Ice drifting buoy data, one is the International Arctic Buoy program, and all of my drifters send data in realtime to the International Arctic Buoy program, and then it's distributed on the global telecommunication servers, the gts, which goes to weather services.

Q.32

Okay. So the weather services are using the data?

So the weather services are using the data. The international program for **an**tarctic buoys, which is kind of the sister to the I.A.B.P., they don't have quite the same logistics in place, so they archive data, and it might go out to the gts, but I'm not sure if all of it does, and I have *no* idea where you access the data. So Allie's buoys.. that'd be a question for her I'm not sure what she actually did. But that data doesn't go into a public forum where it's immediate available.

Q.33

It's that just not a priority?

It's not a priority I think that program doesn't have the infrastructure that the Arctic Buoy program has, because she needs someone in charge of the program who has time to make sure the data is going from one place to another.. yeah.

Yeah, and skills required for doing that.

Yeah, but the important thing is that the data gets archived.

Personal: Do you read your own research, and why become an Ice Scientist?

Q.34

This is more unrelated, but I was curious, I wanted to ask at least one researcher this: Do you ever read your own research after it's published?

[Big grin, laughing] Yes I do haha, I read my papers all the time because I forget what I've done.

Are you serious? So you are just like: "Oh I wonder what I did then?"

Yup, I am serious hahaha- they're kind of like bookmarks of haha where you were at- at that particular time.

You also have to re-read them when people read my papers and *think* they see something in there that I was like: "Really, did I write that?" **[hmm, interesting]**

So you have to kind of become an expert in your own work then.

haha yeah, or *re*-expert [laughing]

Q.35

What inspired you to go into this field and become a Sea-Ice scientist?

Oh wow, I wanted to go to Antarctica since I was eight. **[Really?]** Does that answer the question?

Did you ever go?

I've been once. So when I finished my physics degree I decided I'd like to try and do some– I'd like to go into Earth Science– I asked in my department if they knew anyone who was good to take on a Ph.D student who was doing anything Earth related and-- the *only* person I talked to was Seymour Laxon who was my Phd advisor and he was studying ice, and as soon I heard as I heard he was studying ice-- I'm doing that...

"Sign me up!"

Yeah, it was like hooked at first bait. Yeah, so I don't know if I really thought too deeply about why, it's just something I really wanted to do.

Have you ever regretted it?

No. Never.

General Public: Untapped resource?

Q.36

For our project- originally the idea was to help people identify things where they didn't know the space and time of where those things were located– the idea was to visually find things– and I don't really know of any good examples for Ice Research,

but we were looking at some things like Ocean Eddies [mhm] **, Coastal Pollution hazards– but one thing we ended up discovering is that this won't really be useful to most of these researchers because they really prefer to work with the raw data– you know, like level 1, level 2, not necessarily like.. whatever level we're working at..**

Derived data products are more useful to the general public.

Q.37

Yeah, so that's what we were thinking, is one direction we could go is, creating more of a tool for the general public to just explore the data and have fun basically vs helping researchers

Yeah, the best tools I've ever come across are ones where you– sattelite track simulators– software where you can plot up the images with the extract footprint shape of the sensors– so you know you can actually build swatches of images. The other thing that's useful for researchers is– understandable gridding tools? So modellers like to have their data gridded, and often remote sensing data is provided in gridded format? It's really important that we understand how that gridding is done– I'm at the moment working with a dataset where I– there's no documentation on how the gridding is done.

Q.38

Just so I understand– are you talking about– I know satellites float over and they're not going in Latitude/Longitude– are you talking about where they map the datasets to..?

So just mapping. So if you have a– model that you want to– compare directly with satellite data you need to map the satellite data into real space or if you have a field campaign

Like for images for example, you have to map those points onto the pixels? [Yeah] **So understanding how that's done, basically?**

Yeah, I mean I've actually some really cool IDL code that does that really slick– so you could basically have a satellite track analyzer that plots up the data for you.

Q.39

From the raw data?

Yeah, so that's kind of cool. So when it comes to gridded products which the public use all the time I find it frustrating that the documentation is not provided on how those gridded products were created. I'm working at the moment the GRPS product,

and it has a gridded product that's publically available and it does not look anything like my drifting buoy data. It's.. missing signal somewhere.. that's all I can say it's frightening.

Q.40

Could be pushing wrong data out there..? or maybe your model just doesn't match up somehow?

I don't know what's going on, but I can't figure out what's going on without documentation on how the product was gridded.

Q.41

[..] So that's the end of my questions, but one thing I wanted to show you 'cause you mentioned Python was an idea that I came up with as a replacement that we kind of scratched, but I thought was kind of interesting, [hands over paper prototype] the idea was basically to try as hard as we can to stream acquiring datasets from NASA.

Okay, so you can select regions and time in the classic way.

Q.42

Showing that to other researchers, what I heard was, first of all: acquiring data currently is not difficult, in fact a lot of the ways that are streamlined already they just don't use 'cause they can just go ftp in and download it

Yeah that's pretty much what I do. Actually I find, anytime there's a new web interface to get ahold of data I have to spend an extra half–to a few hours learning how to go through the interface in order to get the data– but on the other hand– so for a researcher's perspective I'll figure out a way to get the data– that's not the worry– from the public perspective– *getting/*being able to get ahold of data in a transparent way is probably important, I would imagine.

Q.43

I don't know how much the public deals with scientific data.

Yeah, who knows how much the public is using MODIS imagery.

Actually, I should show you something, have you seen this? Ok, so talking about the public using NASA data, there is a website, a blog that has a wide following of people that like Sea-Ice, for whatever reason I don't know why.

Maybe they all just want to go to Antarctica

Maybe they just enjoy looking at pretty pictures of ice, but what they *do* do is– this blog has access to data on here– the guy who is making it is pulling data from different websites where it's made available and presenting it here in a nice way and then people come to this blog for information about what the ice packs doing– and this guy is not a scientist he– I think he might be an engineer and he does this in his free time.

Just interest in ice?

Just interest in ice. And they [laughs]– they're doing crowdsourced science– they participated in something called the Sea Ice outlook were members of the blog voted on when we would see the summer minimum of sea ice and what the extend of the ice would be, at summer minimum in the arctic– kind of wacky stuff– but they're actually using data on this site! and they're parsing it– and it would be hard to find– so look they've got satellite imagery– they're pulling it from somewhere– this guy has plotted up trends in ice extent in area by month– so they're taking data from the national snow and ice data center–

so imagine these guys if they could search the NASA database and do anything they wanted with the data.

Q.44

Yeah, I think most of the NASA data is available for public consumption– maybe a bit hidden?

Yeah. So they're definitely using the MODIS imagery that NASA uses, and probably AVHR as well, but I've seen mostly MODIS– and they also tend to parse a lot of– data from researchers to the public. You know, they'll see something interesting like a researcher has done and pretty it up.

Q.45

That's interesting. That could be like a trend for the general public to more interested in Earth research, and playing around with it themselves

And *participating* in Earth research. I think there's a lot of people out there that if you give them data they'll actually look at it. Not just be passive, you know.. NASA always been providing pretty imagery to the public that everyone loves– the whole space telescope stuff comes to mind right now because that's in the news but– I think we underestimate how much the public would use the data too.

I think in the Computer Science community, we rely vastly on just general interest and– I mean the actual computer science academics– they're doing things

that the general public not going to use for the next 10-20 years but– I think a lot of stuff that's pushing us forward is coming from people that don't have CS degrees or maybe just doing it for fun or are in countries where they can't easily get CS degrees so– that's an interesting though that just making it easier for people to use could lead to more eyes looking at the data– and more discoveries.

Yeah, and that would be the polar opposite from what I need. I basically just don't want you messing with the data [laughing] I need to have the raw levels of the raw data there, and documented.

Discovering NASA Web Services Exist

Q.46

[..] so helping Earth researchers is a tall order, so trying to figure that out is.. hard.

Yeah.. so one of the things that's kind of cool is being able to look through imagery in real time. I use that all the time while I'm tracking field campaigns– at the moment– do you know the NASA Lance?

Yeah, that's the one that provides data every three hours

Yeah! So the near real time site has these little thumbnails of modis imagery and you're like– it's sitting there searching through it trying to find– which– and you start memorizing the orbit paths so you can remember which part of the screen I'm going to find that particular location [laughing] on the planet– that site can do with a bit of a "spiff up" make the data more–

Well they actually– the team that I work with GIBS will take that imagery and put it on a "Google Maps" type service called WorldView and then you can see it show up as it comes out

O.k.! Now we're talking haha

So that's already available. So they're dream is that researchers will use those images in their work.

So the connection there is to make the WorldView imagery searchable

[Reserved] Yeah... so that's basically something my client thinks is useful

It would be.

[Excited] Really?! So searchable in terms of visual- what are you looking for?

Looking for a position on the planet at a particular time.

[Disheartened] Okay.. so spatially and temporarily.

[Yeah, just] tell me what's there

Q.47

Yeah, so that's available.. you can actually see what datasets are available by putting in your space and time that you are looking at– and it will tell you what datasets are available and it will give you tons of metadata about the different datasets– it's called– it's not WorldView– [thinking search.earthdata.nasa.gov]– if you look up "NASA Search"– there's a word for it but there's a service they provide that uses– the team I'm working with's data and it will give you not only Satellite images but also– the drones– the ones that fly over and take pictures they'll show you.

Oh, really?

So, if you are looking for a space and time and you want to see all the data NASA has– you can put in the space and time and it will show you everything

That's pretty cool

I guess the big problem is just telling people about it [..] Researchers often will just talk to people and just bypass the easy-to-use tool.

Well we don't even realize that it's that easy to get the data haha, that's funny.

Chapter 3

Meeting with Dr. Robert Kennedy

November 20 Assistant Professor
Geospatial analysis, remote sensing, modeling, landscape ecology, disturbance dynamics, computational methods.

Q.1

For my senior project I'm working with NASA to basically help researchers– Earth Researchers– use remote sensed images to do their research– how we are doing that we're not 100% sure yet.

You're working with NASA, NASA is a big place– there must be some person or something right? [laughs]

Q.2

Yeah so very specifically it's NASA JPL, and it's the Global Imagery Browser Service Team– so that's the team that takes Earth Data and makes images out of it and provides it for people to look at it a tile format so they can like, zoom in and stuff.

Okay, cool.

Q.3

So, my question for you is– I saw that you did research looking at Landsat Imagery and how it's used and so I'm really interested in how researchers specifically use remote sensed images– what kind of challenges do they face when using it.

Okay.. that's a big question, haha. I mean my whole research career is built on image processing– so I can answer that but I'm not sure I'm your audience 'cause I already do that right? Like my whole world is using the imagery to do stuff.

So it's all pretty easy for you is that what you are saying, or?

Oh no! It's never easy [laughs] but I mean it sounded like part of your goal was to encourage people to use imagery in their work, and that's all my work is–

... and make it easier, yeah

Well you know– gosh, where to start. So I have traditionally done all of the processing and you know– all of the atmospheric cloud screening, radiometric normalization, geometric correction, all that stuff, but increasingly that's *done*, and so that's– those are key things to have available to have good atmospheric correction and cloud masking, but you can get those from the USGS for Landsat now so that's not an issue and the other part is is bringing the algorithms to the data– really the thing that has been the biggest issue for all of us in the community– I'm on the Landsat Science Team and we talk about how access to Landsat data [should] work.

and so– I'm actually also on the LP DAAC Advisory Committee, so we talk about– that's another place that serves up imagery for the community– they do MODIS stuff primarily but a lots of other things. [Checks for Student waiting to talk to him as this is his office hour]

So anyway, you know the *real* change is that a lot of the questions people are interested in asking– and the community is pushed forward to make people's expectations about answering– require LOTS and lots of imagery. Like, the stuff I do, I have 100 Terabytes of imagery sitting in a server here that I had to pull down, right? So– that's a lot to manage and it requires a certain level of expertise and interest in doing the computer management and all that stuff which not everyone has, so a real barrier to broader use is– having to managing all that data yourself if you're not already an expert in doing that.

So a lot of the prior barriers were the processing of the imagery, you know– the atmospheric correction, the cloud masking, the other things I mentioned– MODIS really changed a lot of the expectations in that, Landsat has pick that up, Sentinel is going to be having that– that European Satellite soon. And so that's great, now we have all these images and the next barrier is basically how to help people manage that.

And so the answer that a lot of people see is: instead of having somebody take terabytes of data from the USGS over there and bring it over here and do something over here is to set up a system where people can actually process it on the archive and so you're not moving terabytes of data back and forth, you're moving the *results*, but not– like there are hundreds of copies of the *exact* same image sitting around the

world that people have pulled down for whatever place, right?– that's redundant.

So that's that's the big change– so there's several changes, so NASA JPL– so I don't know actually about what you're describing so that's good for me to learn about, but NASA Ames has NEX, the "NASA Earth Exchange" which is– they also have the entire Landsat archive there and that is a place where you can plug into their super computers and access all that– so I don't know how that interacts with the JPL folks.

Q.5

So you process the images remotely, what are you trying to get back, just the answers to your research questions, or are you looking for a specific space and time or what kind of answers are you looking for for those questions?

Okay, so that's a good question– so I am interested in terrestrial systems, Land Cover change– what people and natural processes are doing to the observable surface of the earth, and Landsat– primarily in the Landsat era and the satellite era.

So that's looking at disturbance, people, or you know, or fires, or landslides– whatever changing what happens, and then what happens after that, what precedes it sometimes if there are biotic things that precede anthropogenic things, et cetera– so there are bugs that kill the forests that precede a salvage logger that precedes a fire, is there a fire that precedes a sal... like understanding how those things happen, what the processes are that drive them so that eventually we can model them better and you know– the whole goal is to better understand how what we're doing affects the planet so we better– have a brighter, greener future.

Q.6

So you are saying that it's more like– possibly interactive, like you're not sure when you're looking at the data– you might not have a specific– you might not know why this is happening but you'd kind of know that–

Oh, I see, well I'm just trying to paint a broad picture– we do have specific questions for project of course, right, but that's for the overarching thing– so one project I have is looking at Carbon, and so there is specific questions about forest Carbon– the first one is: just documentation.

Documenting every pixel of forest carbon over time: how much it is, if it go away, and if it goes away, who caused it to go away and where. That hasn't been done, it's a very fundamental– you know, observation thing but that's the starting point. You have to understand what's happening first, right, so, that project is looking at the– essentially the fade of carbon in Oregon, Washington, California– and trying to understand and

again just document and through the documentation hopefully come up with *some* understanding of what's happening with policy drivers or whatever.

Land cover change is another one so, biomass land cover descriptors are like forests, non-forest, you know agg, urban, that kind of stuff and watching how those– again, documentation. Watching how those change over time is sort of a critical first step for a looot of questions about what's actually happening on the landscape. So, you know, the first thing there is again, characterizing who is doing that and what kind of change is happening, and then that leads into questions about habitat and urbanization and what people do.

[...]

There are– Other things that– Just to sort of I guess final answer your question so– I also work with national parks, and national marine fishery service, and again a lot of those projects are trying to test specific hypotheses about how much land cover change matters. So, for example for salmon, what they want eventually is to have information about what is happening near streams– salmon bearing streams– on a yearly time step that's just very descriptive about what is change is happening because that affects the inputs to the steam– that affect the salmon.

Q.7

So a lot of like temporal type stuff, not necessary spatial? Looking at one place over time or looking at–?

Well, it's spatial and temporal. So the spatial part is because we are looking at large areas and we're looking at– you can look at anything adjacency or whatever–- like with the steams– the temporal is yeah– I mean I use– the reason I have so much imagery is because I'm looking at hundreds of images at the same place through time and then applying time series algorithms to them to extract interesting information.

Q.8

Alright cool, well that was the most pressing questions so I'll leave you to, you know help your other students and stuff [Okay!] **but thanks a lot for meeting with me** [Yup!] **I really appreciate it, it's been really helpful.**

Ok. Sure, yeah, yeah–

[These notes were taken as I was leaving while the recorder was off, not verbatim]

[Showing prototypes]

Q.9a

My advice to you is don't do something that's already been done.

- For example there's this LP DAAC service that nobody uses because people can already get the data elsewhere
- I use Google Earth for example because I can get the imagery and it provides a service on top of it
- If you provide a service on top of the data that's useful, people will use your service.
- Don't do something that's already been done

Chapter 4

Meeting with Kuuipo Walsh

November 19 GIScience Program Director
GIS, metadata, marine resource management.

Q.1i

I don't know if you know anything about our project.

Noo, tell me about it.

Uhmm, Yeah, so it's pretty exciting, we are working with NASA JPL to build software to help researchers, earth researchers specifically, do their research,[ah] **that's like the general overview.** [Right] **The people we are working with are from the global imagery browser service team, so what they do is take earth data and they, uh, make pictures of it, and make it so you can like zoom in on parts of the earth and see cool Satellite pictures and stuff.**

Great.

So, from what I understand from their project description... which if you want you can actually read if you want [hands over project description]...

Yeah that would actually be great.

From what I understand, their goal is to allow researchers to use that images to do their research

Right, and this is with NASA?

Yeah NASA JPL

How did you get involved with them?

It was for our senior project, so every CS student has to do a three term capstone [right] **and so Daniel is part of my team and this other kid too, and we were assigned this group.**

Oh so you were assigned this project.

Well we selected some choices and from there we were assigned

Hmm 'k. [Reading project description]

So they have imagery data, and they want a web service? [**Yes, I think so.]** And uhmm.. do they have services already set up and you're just creating another one...? Or...?

Q.2

Uhmm, yes the thing is we are very vague-- we are trying to meet with researchers to figure out what will help researchers because I don't think they know exactly what they want either, [Oh... Okay.] **Yeah, so GIBS is a service that provides images-- that's already set up** [Yeah, right.] **and they have like tools for seeing those images and looking through them already.** [Ok.] **Called WorldView I think, yeah WorldView.**

But do they want more researchers to come to that site and actually use the data?

Okay well from talking to them, what we understand this... [hands over prototype] here's my prototype for what we're thinking might be useful. Well specifically we're to try to help researchers find things ['kay.] **like visually finding things, using like shapes and things. And Lewis [our client] is really keen on "images as a proxy", which we have no idea what that means, but I think what that means looking through the images instead of the data.** [Okay, yeah, I think so too.] **Yeah it's a good hypothesis. Yeah.** [Yeah, Yeah.]

For example, we were looking at some remote sensing images around Syria yesterday, and we're actually looking at areas that have been affected by drought, and the information we were learning about was agricultural data, but you could *see* that, you could see what vegetation was affected by the drought so much easier just looking at the remotely sensed image.

Like the pictures?

Yup, just looking at the pictures you could see where the drought was and it was interesting because it followed political boundaries. So like in Turkey there were a lot of green and then at the border of Turkey and Syria you go down to Syria and it

was brown, and that was telling us a lot, that somehow Turkey was getting a lot of water resources, where as soon as you got to Syria they weren't, so that's a remote sensed image that was giving us a lot of information. Just by looking at the colors of the vegetation.

Q.3

So one question I have then is was that hard to find? Or did you already kind of know where you were looking for that information or were you just kind of searching all over?

I think another researcher that did this talk that would be really useful to talk to because he worked at NASA himself, and he has a lot of background information that would probably be helpful to you [windows error sound], is Jamon, **[Oh I'm meeting with him on [next Tuesday]]** oh, okay, perfect! Well it sounds like you guys know who to talk to.

Q.4

Oh well I mean, just information from Ted [Yeah.] **So he worked with you on…?**

So he was giving… Yesterday was "GIS day" [laughs], kind of our Woodstock [laughs]. and so we were met on campus and we heard different presentations about what's going on, and Jamon gave the keynote presentation at the end. So he was a NASA postdoc so he will know *exactly* where you guys are coming from, he'll probably be familiar with the databases, because he was telling us how the imagery he had isn't publicly available, so it was given to him because he worked at NASA.

Yeah, so he well, he will be perfect, he will tell you what data, so *HE* was given the data that he was doing his research on. **[The images?]** The remotely sensed images. So what other researchers do, especially students, they have to go and search for it themselves, but hopefully Jamon can enlighten you about where to find-

Q.5

GIBS provides lots of different Images types, and one gives images of resolution about 250 meters, I don't know like what resolution he's working with, but that's available almost every day they release those images.

Right, his resolution was probably better than that, I wanna to say one meter, but I'm not positive.

Yeah that sounds like something they would probably not provide for everyone. [Yeah, right.] [laughs]

And then there's the issue that you're talking about terabytes upon terabytes.

Q.6

Yeah a lot of data, that makes sense, so going back the question, so was that hard to find those areas of vegetation? Was that something you kind of knew about coming into it?

Oh so the way most research works, I would say in Geography, I couldn't speak for other disciplines, in Geography we're very interested in a particular place, so normally a geographer will have identified the place they are interested in. **[Okay]** Geographers are very interested in space and time. So they are going to know, they are going to have a research question that starts with: "In this particular place", be it a small locale or even globally on the entire earth, they will have a research question and it will definitely be attached to a place. At different scales from local, regional and global scales.

Q.7

So you have never really worked on any projects where you aren't sure when or where something is located?

The projects I have worked on have been all based in Oregon, so for me no, it's Oregon and the time frame has been recent, you know modern. That's a good question though, Geographers pretty much know the time and the place that they want to do their research, because what they're doing is they're comparing-- they also do comparisons to different places, why is this place different from this place? What's causing drought and more in Syria and not here? They're looking at comparisons and they are also looking at comparisons as things change through time, so it's usually defined,--

-- sometimes it's defined by the data itself, for example we have a researcher that's doing research with *National Geographic*, and she's looking at issues of *National Geographic* but she's looking at issues-- the magazines-- from the 80s until now, because she doesn't have access to issues before then. So your data can define your... **[the time you're looking for.]**[Yeah]. And that happens a lot with students because they tend to-- they aren't working in tons of research money, like maybe their professors are, so those students may be stuck with free data. and so they might only use landsat data-- and its landsat has been a satellite that's been up during a certain time period. So they might do their research during that time period. So the data may drive the research question.

Q.8

Okay yeah, that makes sense, yeah I was wondering about that.

Another thing I was curious about is the intersection– I mean you in kind of a unique place, since you are in GIS Systems and you are not like doing the [field] research necessarily [right.] **but I was interested in the intersection between software and researchers** [*Good* question.] **Coming into this project I assumed researchers didn't write software they just used tools available or something like that** [Shakes head no vigorously] **, coming into this. But after talking to Ted it, he was like "Oh man I work with software all the time I used to write Fortran but now I have some people do it for me", and I was really interested in the fact that he said that they write their software to answer their questions, and then just throw out the software after they are done.**

He also said something really interesting, which was that he doesn't really trust outside software packages, but he said that was probably unique to oceanographers though.

[Laughs] And I have to admit I have been around Oceanographers and Geographers and not so much other scientists.. so in our college we have Oceanographers, Atmospheric Science, and Earth Science, so that's pretty much my exposure, but I would definitely have to agree with what Ted said; My husband worked for a Atmospheric Scientist, and he did not trust-- he would not trust-- GIS Software, my husband his first job as a student was to write programs to analyze data, that is definitely true in our college: that researchers either like to write their own code, or hire students to write the code, or work with collaborators at other institutions or other colleges to write their own code.

Q.9

So is that a problem for you working with GIS systems?

So... a lot of researchers see GIS, Geographic Information Systems, as a way to *visualize* their data, once they've done all their analyses writing their own programs, they might use GIS to just make a map to visually display it. And they don't really see it as more than that. It's like I do all of the analysis of all my data myself, and at the very end I might throw it into a GIS for display purposes. If you talk to somebody who's a Geographic Information Scientist they wouldn't agree with that, they would say a great thing about a GIS is you can do all this analyses. That's what the important thing is.. but... there is definitely a difference of opinions depending on your background.

Q.10

I wonder what causes that, like distrust? Or maybe they just think they are smarter than the system they are using?

That's a really good question.

Q.11

Like for us we want to get past that barrier, to produce a software that will help researchers, how are they going to find it? Or whether if they are going to find it useful. [Right]

I think researchers aren't comfortable, with, let's say your GIS truncates the data at a certain-- so its numbers-- decimal points, at a certain point, the GIS is truncating your data and its a different precision than you thought it was. I mean, they would not be happy with that. So maybe GIS is just a "black box" so they don't know exactly what is happening to their data at every point of the analysis.

That makes sense.

You don't want rounding error, and the GIS may not be as accurate as you want your output to be.

Q.12

Yeah we were looking at a research project that was waste water being pumped into the ocean, and the researcher was able to get a resolution that we weren't able to get, and I think that part of that was the rounding processes was getting images out of the raw data itself.

Yep, yeah.

And I think some researchers just like to stay with the raw data, they just don't want that data manipulated *at all*. They are going to be the ones that manipulate the data, not a piece of software.

Q.13

So what do you think about researchers using software libraries or pre-prepared packages in their work? Do you think that's something that's common?

Yes, yeah. Especially Open Source tools and libraries, R is very popular right now.

So R is really is in all the different libraries you can get, so R is statistics, but then there are all these libraries you can be used for Geographics statistics and so many

of our students will start off by learning R then they will *quickly* start using all the libraries that can manipulate the statistics geographically.

Q.14

Do you know much about researchers acquiring data? Do you think that's something that is challenging or do you think-- I know Ted said something that might be helpful that I think might already exists might be integrating some of the data we have with GIS Systems because that would help researchers acquire it easier, although I saw NASA has a few packages already for that.

Is that something you have worked with before? Researchers trying to get information easier?

That's a very good question...

This might be a question that is better for the researchers that are working with it.

Right, yeah from a GIS perspective, over the years we have created these things called, uhmm, I'll show you, there's a whole chapter on it in in a textbook, and I can just refer it, the chapter on...

So. From the GIS perspective we have worked very hard to share our data, because that's actually very expensive, to create data and to create GIS data, I mean, there are researchers here that have driven on every single road in Oregon, and every single forestry road in Oregon, gathering data. And have canoed and kayaked down every river in Oregon or in the Willamette Valley gathering this data, it's really really expensive, really really time consuming. So in the GIS field we definitely want to share our data with each other. So we have these things called "Clearinghouses" where we share all these different datasets with each other, and we've worked long and hard to make it make it easy for other people to find that data. So that is a whole field of study in Geographic Information Science.

Q.15

Yeah I saw you worked with Oregon spatial data.

Right, right, and... but how successful that really has been- the jury is still out I think on how successful it's been, they had a spatial data infrastructure that was created, like the... national... [typing]...There it is, and it is a federal effort, and then on a Federal level for the United States we have the "National Spatial Infrastructure", and we have all this data and services and metadata exactly to help people find the data they need.

How successful this has been, we have had the best of intentions, but I don't know how successful it's been. We also have a global data infrastructure and that might, in my textbook they're saying that might be a little bit more successful for people to share data globally with each other.

Q.16

So why are you saying you're not sure whether it was successful? Why do you think researchers aren't being able to find or use-

Because researchers are still creating their own data, and there are still silos and groups that are creating their data from scratch and don't even know that other researchers exist or that other data exists.

Q.17

That's really interesting, this has been really helpful and really insightful.

Yeah, Yep, Sure, let me show you the chapter real quick, so this book is written by the people in the field that are kind the most knowledgeable about... GIS in general.

What's the book called?

Its: *Geographic Information Science and Systems*, by Paul Longley. And there's a little chapter, or little section in here, "Spatial Data Infrastructure: the US experience." And it tells about how we really intended to be able share our data with other researchers and hasn't really worked. And so... that might answer your question.

But yeah, people are still having trouble I think, finding data.

Q.18

Yeah, Ted Strub said he likes going out and putting buoys out in the ocean

Well, yeah! (excited)

and he said remote sensing data is... why did he say he didn't like it? I'm trying to remember, because it moves around a lot.

Daniel: Yeah the oceans always changing., So it's hard to gather information because you can't rely on prior's day data, or even minutes or hours data.

Right, interesting

So he spends millions of dollars on paddling boats out to float buoys and-

Yeah there's live data, and he is probably getting from those buoys.

Well thanks a lot for your time, we really appreciate it.

You are welcome! Good luck with your project.

Alright, thanks a lot, have a good thanksgiving.

Yeah you too. Bye.

Chapter 5

Meeting with Dr. Larry O'Neill

November 20 Assistant Professor
Air-sea interactions, satellite meteorology and oceanography, atmospheric boundary layer and ocean mixed layer dynamics.

A Bit about Research Data

Q.1

So a bit about our project [...] the initial project was essentially trying to help earth researchers like do earth research using remote sensed images easier [ok, that's good] **how we're going about that we're still in the process of trying to figure out.**

We're really looking for information about how earth researchers use remote sensed images and just information about how that works– especially any challenges or any problems that people run into.

Yeah, ok, well this is good. Have you talked to anyone else around here?

Um, yeah, I've talked to so far Ted Strub– Ku'uipo from GIS and then this morning Robert Kennedy

Ok, yeah, that's good one to talk to. Yeah so you probably know now that we all use remote really sensing differently

Yeah, that's right, especially oceanographers vs...

...land surface people. Yeah so I'll just tell you a little bit about what I do and then– I'm probably going to be a little bit like Ted Strub– he uses a lot of ocean data especially

near the coast. I end up using a lot of ocean data as well. Like ocean winds and moisture and temperature fields. I also use quite a bit of clouds like remote sensing like clouds and things like that as well. And so I do a lot of like– my visualization is really bad [laughing]

yeah but some of our datasets are really large like they measure into the tens or hundreds of gigabytes and so– is your project more in the visualization of data or just working with large datasets?

Q.2

Well specifically the team we're working with is from the NASA/JPL GIBS Team so they produce images from data and then they like serve those to researchers hopefully, ideally, to help them do research–

so maybe visualization– I know Ku'uipo said that earth research like using GIS Systems or prepared images for visualization purposes but not necessarily for actual research 'cause they prefer to use the raw datasets themselves

Yeah, and that's where I am at so I don't know any GIS systems actually. You know, at times wanted to, but most of my stuff very– kind of *raw* so I do a lot of stuff like– processing the data from various things and you know I like to deal with the raw data. And GIS isn't as good for that but, they on the other hand, produce much better visualization and I think there's a lot more tools in there for doing that– but I don't do as much of that so yeah as far as GIS goes no.

I end up doing a lot of my stuff in either MATLAB or Fortran or just like, shell scripting. I'm not very good at it [laughing] you guys are probably much better at shell scripting and stuff I end up having to you know, if it's anything else other than simple stuff I have to look it up online and see if anyone else [laughing] so yeah that's one thing is that earth, you know like people like me n stuff sometimes we end up having to do deal with very large data sets and shell scripting especially is very useful–

yeah as far as visualization stuff it's really I think that's a real challenges because you know a lot of times you know I just want to make a simple map or something– you know JPL's having really nice things, I'm actually on their user working group for PO DAAC **[yeah, I saw that]** yeah, so what's really nice is that they you know, you can– sometimes you want to just like a day's worth of data and you just go and say "before–", say: "show something interesting or something that I want" and it's really nice there because you can go online and do it without having to– 'cause otherwise you have to download the data and you have to know how to read it or write the reader

and sometimes if you just want to look at like a day's worth of like temperature or something– say if someone asks you for or you want to look at it, could take you a couple hours to make that map if you don't have the data– so the stuff is really nice.

So it's really nice to have it segmented temporarily is what you are saying?

Yeah

Q.3

What about spatial as well, do you ever identify an area of interest but you have to download more then [oh yeah]

Yeah actually one thing I think that's really interesting I don't know if this might be completely off topic but it's just maybe something that is interesting– so you end up downloading and you end up getting each data file for one time so you know it's a 2D field of one time– it might cover the globe so you have like you know, two thousand grid points in the longitude, and maybe 500 latitude or something and you know maybe–

so the data is stored in these 2D files and if you want, say do a time series of one point in there you have to actually *load* all the file for every time in one of these systems and– or you have to look through a bunch of 'em and so the data you know is really a 3D data but the file system like really good for 2Ds and I've always through about– you know so if you want if you just want to do a time series instead of a map at a point or something it's really a problem to, 'cause you know getting the data–

For someone like me, I think there might be solutions to this but it's not– as far as earth researchers– it hasn't really tricked down so I don't know.

Q.4

So the knowledge is there, but it must be available?

I don't know if it is or not, I'm sure it is 'cause it seems like some that's really– like other people– this happen in other types of data structures or something, it must be some sort of data structure problem and I just– you know I have so like I have like data that I can show you you know I.. have–

[Looks up computer stuff]

A lot of monitors, we should have a record for which researcher has the most amount monitors available haha [Yeah]

–so I have like this data set for Sea Surface Temperature so each one of these names is

a file name– and it's just a 2D field and so if I want to know– like recently was a good example I want to know the temperature– someone asked me about the temperature at a morning that they had rid off of hawaii– so, his one point– so there's one number in every single one of these files that has million numbers in each one– so you have to look through all of them and it's kind of a pain– and anyway that's one thing that I think of for– the difference between a map and a time series.

Working on Data Remotely

Q.5

Okay, well staying on this topic I know this morning I was talking to Robert and he said one development that could be done is kind of keeping the data where it comes from like USGS and then kind of working on it– he said "bring the algorithms to the data" vs vis-versa

Do think that's something that if it was available that you would take advantage or do you think that's something that you like– you personally like having the data available right on your own computer or in your own data center?

A lot of times I like having it available, probably because I've– I think the datasets that I'm involved with are kind of like "medium sized", in the order of hundreds of gigabytes or tens of hundreds of gigabytes– which now is.. it used to be that was a untrackable amount and now it's.. [small]

So I have a computer down in the– environmental computing center I just have a– RAID and its 70 terabytes– and it's mostly full now [laughing] and so it's– I have a bit of storage, and actually that was only like $20,000 a couple years ago so storage is really cheap–

A fair amount of money for like an intern student

Well, yeah, yeah it is, but as far as like some of these computer go, you know it's actually not bad.

Q.6

So you have a lot of experience then managing that kind of data– because you set up that RAID yourself, or–

Yeah, yeah I setup the raid– it's a RAID-2 it's have about 48GB of unique storage and it raids to ~72tb. So yeah so I have a lot of these datasets– like I have a bunch.

So back to your original question about– yeah it is nice to be able to access some of these bigger datasets so I think Robert Kennedy deals with a lot of Landsat data and he– so Landsat has– so it's like– images of the earth at like 20 meter resolution– so you have like– every single day you have like, you know, hundreds of gigabytes of all sorts of different things– for that that's a huge problem– he tries to store it locally so for him it makes more sense.

For me, it's not quite as much– although like I don't know– if you guys are thinking about like OpenDAP servers and things like

Is that the like the same thing as PO DAAC?

Yeah, yeah. So you can so I guess it's a protocol– I know JPL and PO DAAC they have– this thing it's called (pnep?) basically you can query data from– you know scripts you can say– you can point to– I think it's a web address or something. I don't use it very much.

Q.7

That's what's I think is really interesting, Robert said LP-DAAC does exist [Yeah] **, but not very many people use it** [No] **because they can already just get the data some other way(?)**

Yeah– right now when– storage is still kind of cheap so you know– for me I end up using the data for a lot of different projects and stuff so sometimes it's nice to have it so that I can–

So you don't have to find it again, or..?

Yeah. So you don't have to find it again and sometimes I just need a lot, like I'll want 5 years of global wind data– which you know each file gunna be a couple megabytes and then and then every day 365*maybe say 5mb and then so to– if I have this script that does store some sort of you know– it does something to the data– analyzes somehow– and then– do it. So if I you know first time you read the script and its starts loading and I've made some sort of mistakes and it screws up you know if the data– if I had to access the data– if it wasn't.. cached.. or something– I'd have to download it over the network every time I, you know–

Well the idea would be you would be running the script over there, so

Oh! Right, okay yeah–

So it wouldn't be pulling it as you ran it, it would goes through

Oh yeah yeah, okay. That's something else [Face becomes scrunched up with intense thinking and focus]

Q.8

But I can already see that it's like complicated enough where you are like: "Well I kind of like how I'm already doing it"

Yeah, it's like– I do kind of like having it locally– so I have done some of this "where you run it remotely" you do some scripting– one of the things, I'm not very good at– you know a lot of the scripting was, I think perl or javascript or something– and we don't learn that–

You really like your Fortran, MATLAB, you're not open to maybe–

I'm open to changing, but at this point sometimes I'm kind of like crunched for time or a deadline or something

because I know this you know– this might not help you and I know people don't like changing but they say once you learn one of these higher level languages you can achieve productivity of like 10x faster vs Fortran

Yeah, I wouldn't doubt that. I actually use a lot of– I actually mostly use MATLAB because..

That's more like javascript, that's more higher level so that makes sense

Yeah, I do it there and so I've– I learned it in Grad school and I have like a core– probably couple 100 analysis routines that I've written over–

Q.9

You have your like– util dump?

Kind of yeah, yeah, and so it's stuff that you know– over the years, you know– found all the mistakes in them and now I'm kind of comfortable with what they do and whatever and– so sometimes going to a remote thing you don't always have access to that so when you– so having it locally is good, so that's one drawback sometimes the–

We've talked about this at the user working meeting is– 'cause I guess PO-DAAC has something– they're playing with– they have like a PO-DAAC labs type thing– they're working on something and– I guess one of the things they had is kind of LDAP for whatever–where you can kind of do like simple subsetting of the data locally and doing some stuff and– I've always– so one of the things I've wondered about if you're doing that– if it's popular and lots of people do that– are they going to have the amount

of resources that you know– it's going to be so if you were all trying to do some things with the data–

So you're worried about basically sharing

Oh ya ya ya, sharing. but I– I haven't used it very much, so I imagine it's still pretty good right now because not many people use it hahaha, unfortunately

Acquiring Datasets

Q.10

Ok, so another questions that I have is more about your research in general. How do you personally acquire the datasets that are you're working with?

So– for the longest time it's just simple ftp so I'll run– I just have like a bash script or something or yeah a bash script that just kind of loops through and uses anonymous ftp– or like wget and you know put that in a bash script

Q.11

What datasets are you working with? Scatterometry, right?

Yeah, so wind scatterometry– any sea surface temperature datasets that there is I've probably– either have or used it– some like "goes" there's a satellite called "goes" (geo-stationary satellite) so it's pretty good to–pretty pictures of the earth of the clouds– so I use some of that data...

Q.12

So you FTP it down and keep it around– in case you want to ever go back?

Yeah in case I want to go back– so this is like an example [shows script on screen] it's just anonymous ftp it just builds the filename and I have to go through the anonymous or the hosting a server and figure out the directory structure and then I have to build the filename that I want.

The reason that I do that is so– 'cause a lot of these datasets are continually updating because the satellite collect the data and every couple months, you know when I say: "hey I need"– I want to look at the last couple months of data and I haven't run it in a while I'll just run the script and we'll take 20 minutes or something to download and– here actually– well I guess everywhere now but here they got– some really good

connections– like I can get some really good speeds– so this stuff downloads pretty quickly.

Q.13

So PO DAAC is supposed to pretty similar to that right? Isn't that supposed to be what PO DAAC is supposed to replace?

Yeah.

So you're on the user working group, but [it's] not really something that's used too much?

Yeah that sort of stuff I haven't used too much– and the reason is is that I think some of that stuff is– I think their design is set to be used kind of towards you know, kind of like novice to intermediate users who are– kind of still getting used to or just want small parts of the data– or something.

So you're worried that it's not all there, or?

No, I just – I worry about that– we'll not worry– but sometimes I need access to more than it's designed to have– like global– a lot of times they end up using global datasets for many years and– the tools they have aren't designed for things that are that big.

Yeah, so that, yeah– so that was my thing we've– so there's actually someone else in the working group who, uh– he tries as hard as he can to "break" some of their tools– because he knows a lot more about the backend stuff– I don't know anything about the backend and he breaks it pretty easily– with fairly small types of commands– things like a tenth– he tries to get like a tenth or a hundredth of the amount of data that I would try to get and that breaks their stuff because– one of the things is it just takes way too long, like if he tries to subset you know, maybe an area within 500km of hawaii or something of temperature you want it for two years and you will enter that command into LDAP and it just– it essentially freezes [laughs]– or it takes a long long time and one of the reasons is because they– I guess how their back end is that they– and i think they are working on fixing this but they– it's how they store their data and compress it– and so when you read it in.

So they store it compressed, they don't uncompress it until there's a user call– so when there's a user call it takes a file, uncompresses it, reads it in- and then it keeps it in cache, but it expires really quickly– and I guess there was– they were thinking about changing the file system to– I think there's another one– I forget what it's called– you guys might know– it's a different file system structure that it is already compressed– and so when you read it from disk the disk does the uncompressing.

yeah, it's ZFS

Yeah, ZFS! That's it [laughs] so they were thinking about doing that but it wasn't trivial

Yeah, they have a lot of people there, but.. they also have a lot of stuff to do

Yeah, Well no, it's a great idea if they can do it but I think–

Okay so I get it, you have something that works, and and it's like– why would you want to spend lots of time re-inventing the wheel when you already have the data

Yeah! That would be it, so if it could be made so that– I mean it's a great idea– I've thought about maybe having– you know because other people from that college sometimes have worked on similar data sets or something and so– you know, not everyone and not very many people but you having local copies or something or having a local thing where you can–

They should just mail hard drives like every month with all the data and then just everyone shares it! [Laughing]

Yeah, that would be awesome! [Tone of voice: seeing a long awaited package has arrived] "Ah! that's my new.. temperature dataset. Oh man!" [laughing]

Using Outside Code / Sharing Code

Q.14

Okay, going back to what you were saying about you have like your util that you are working.. what are you .. how do you feel about sharing code? I know some researchers are really hesitant on using outside source code.

Yeah. I share it fairly freely. I share my stuff– at least the stuff I know– I'm pretty sure it's not buggy [laughing]

Um, yeah, it's just– I mean there's some code I have– like MATLAB is really easy to not to– document your code and stuff it's really easy to right bad, bad code– and I have got some of that– and also some nicer stuff.

Sometimes like your own code, yeah you guys know this– it works for *your* project– and you're not 100% sure if someone else uses it– if they enter a parameter space that it wasn't designed to and it fails or something and– sometimes it's kind of like pride because here you worked really hard to just get it to work for yourself and it works really well and your stuff but if it somebody uses it for something that's slightly

different than what it was designed to then they're gonna say "AH! This guy writes shitty code" [laughs]

Q.15

But what about the other way around, do you ever use other people's code in your work?

Yeah, occasionally, sometimes I get code from other people that [laughs] doesn't work? It's because it works on their stuff and not mine.

Q.16

But it's just like random, individual researchers, not like necessary packages?

Yes, yes. Packages I do. Well, both. Yeah, so some days I'll take somebodies stuff and sometimes times it takes a little bit of time– to see how it's supposed to be used? or something or maybe there's– you know, how you're trying to use it is slightly different than what it was meant for and sometimes you can just look at it and see– you know maybe it's a simple thing– your array is, you know, transposed where it's supposed to be when you're putting it in or something. But yeah, usually– yeah it's just learning how to do it basically.

Q.17

How do you learn about these things, is it word of mouth or do you actually spend time looking up different packages?

Both, yeah so sometimes yah like, data readers are a big one– so sometimes you get like a data set– you know, now it's becoming better with using like netCDF or .hdf formats which are, you know, easy in MATLAB to get– but sometimes people have like their own binary format or something– or it's just put into an unformatted binary file– and you're like "Wow, the hell is this?" [laughing]

Yeah, and so you just have to know kind of– if it's like gridded data or something you just have to know how it was formatted or something– so yeah so definitely that.

Sometimes looking for utilities like– one example that comes to mind is a couple years ago I needed to fit an ellipse to points– you know so I had like points and I wanted a least squares fit to fit an ellipse to those points and– you know I didn't want to spend a whole day writing the code to do that so you just google it [and there it is, laugher]– and somebody like a mathematician who is very good– write this bulletproof ellipse routines. And yeah I download it and try it once and if it works like it's supposed to like that then "that's cool"

Q.18

One thing I'm always like curious about– me personally– is how can you be sure that that code doesn't have bugs necessary– looks like you have the right answer but you don't [yeah] **especially because I know a lot of researchers don't– the code isn't – you don't put your code at the end of your study – you might hide it, you might even delete it. How do you know?**

Yeah so it's a little bit faith.

I mean, you know there are two things: 1) yeah you use it and see if it gives you something reasonable, so if it gives you something reasonable then– like for the ellipse example if the ellipse looks like it fits okay– then it's pretty good– the other thing is too if I try to do it– you know sometimes you do it right the first time and sometimes you don't. It's the same thing about my own code– how good are you, and are you sure about your own stuff.

Yeah like over time I kind of have gotten better– like I don't find many instances where I downloaded– a piece of code that was just wrong– most of the stuff I find actually I feel like have for instances from like the matlab central– people can "rate" it or something

Haha, five stars = it's right

or you can read the comments for that stuff– the stuff you get from other researchers yeah it's you kind of trust the reputation. Like, I mean I have had people give me code that I didn't trust them so I didn't use it– but then there are– most of the time it's– you're working with someone that you know is good, so yeah

Q.19

Their reputation extends onto the code– even though you said for your own code maybe not the best all the time?

Yeah. Although over time the stuff I've written– obviously over time you get better but also some of the stuff you use a lot is you've kind of used it so many times that you've fixed it– that all of the.. **[well polished?]** yeah well polished– I have some routes that are really good – bulletproof I've– any time I get a funny answer from it and I go to look to see if there's a code error or something– so I've looked through it so many times that yeah it's– yeah. I'm pretty confident with it.

Q.20

Going into this project I didn't even know that researchers even wrote their own code, I just figured they used tools available, and so it's been a pretty eye opening

experience to see that intersection between researchers and software

When I was an undergrad I took a fortran class– and then I took a regular shell scripting kind of class– and then when I was in grad school I sat in on a MATLAB class but I basically taught MATLAB and once you learned the basic thing about what a code is– like a loop, conditional statements, the other statements you can kind of– they seem logical. It's like speaking a language or something– once you can at least where the bathroom is, it's not very hard to say "Oh, I'm thirsty, I want a beer" [laughing]

I end up doing a lot of coding, it's not necessarily very difficult coding– just analyzing the data and just stuff like that– it's not super challenging coding I don't think.

Front (Oceanography)

Q.21

[So where do you think your project is going?] **We'll you know, we're a little bit lost at the moment– [originally] the idea was to help researchers find things– for example an ocean example– your research is more into winds is that right?**

Yeah, actually I do both what's called air-sea interaction so I look at both, so how the ocean atmospheric influences each other– so one of the things that it does is surface winds it also has to do a lot of ocean data– ocean eddies–

So that's what I was going to mention, I talked to Dr. Chelton, Eddies would have been an example where you might not know exactly where and when it happens but from what I understand, finding eddies isn't too easy– it took him a long time to find algorithms to do that

Yeah, so actually Dudley was my Grad school advisor– and actually I've worked with him a bit on the project. One thing that's kind of interesting is I'm looking a lot on how storms affect the ocean, one thing that– one example is front– [pulls out images]

Yeah Dudley mentioned fronts

Yeah, so there's fronts in the atmosphere so this is a– looking like at hemisphere fronts or something. This might– I'll just give you an example of something I'm working on to see if it helps anything so here's like the east coast of the united states here [points to the map] and this is a cold front here–

so it's like a low pressure system here and this huge front– and this is kind of indication that temperature is top of clouds so the top of the clouds are really cold and the surface

is warm– and so its like this huge cloud system and its really windy here and a lot of rain and snow and stuff and so there's this big ocean front that comes under here and so I'm studying how storms like this are affected by the ocean front and one thing is the you get– so this is an example a surface winder field how it converges air–

so air converges along the front and it causes upward motion that causes rain underneath– and it turns out– these longs tails funnels stuff and with time they kind of propagate along– I'm trying to figure out ways of tracking fronts– how they evolve and stuff– and it's been difficult [laughing] it's been difficult to do that– so I mean I've ended up doing other things it– might be an example of something..

Q.22

Something you are tracking or monitoring?

Yeah, monitoring or something, so if that helps too– that might be an example for you to looking on. I was looking at weather models– here's just that same front in the weather model– so it actually looks– this was the satellite data here and this the model– so there's some differences so yeah– trying to compare them– yeah so things with like sharp funnel– eddies are one things– they have like little fronts around the eddie–

Resistance to Change in Research Tools

Q.23

Thanks. [...] one big thing that I've noticed– I think there's some resistance in terms of how information gets out there like– the GIS people would say: "GIS Systems can be used for analysis", but in order to convince a researcher that a tool can be reliable or can be actually used for analysis would be a huge barrier and people wouldn't be able to cross that

Yeah, I would be open to something except there's a learning curve to use these systems. Sometimes you don't want to climb that hill, but yeah I mean I would be open if there's an easier way to do something.

Yeah, learning– you don't want to take risks in terms of time if there's no payoff

Possibly, or if you're sometimes you get a system set up– like Dudley– when I was in Grad School– they do everything in like Fortran– they do everything in this old OLD 1970s... **[IDL?]**

No, no, it's a Fortran based program except it was for– they adapted it for so back in like the early 80s there was this old printer systems they used to make these maps it was really system, it was really slow but it was this plotting system and it was basically like a dot plotter– they called it– and so it has this code that did it and so basically to make a map you had to write a thing that said "GOTO POSITION X Y" put a DOT– and then "goto next one" and put a DOT, so anyway he adapted this system to make his own plotting thing– and that's what he likes.

and I tried using it once and it's not– it's ughhhh. It doesn't do very much and so– it's like well I started just using MATLAB so he used to give me all sorts of crap because, "Oh MATLAB :(", "Oh, you have to pay for a license :(", and "You can't take it everywhere :(", and "Oh, they might go out of business, and then you're up shit creek!" [laughing] and I got all this grief

Q.24

He might go out of business and then how are you going to use his plotter?

Yeah! So now, you know, he's retiring– and he's going to take that with him, and now I'm with MATLAB and now people are coming up with better things and I'm going to say: [Creaky, old voice] "Oh, MATLAB, it's great :}" [laughing] and it's, you know, not because MATLAB is great, it's just because all I know– well kind of what I know– yeah I do want to learn some of these other things because it's really cool to there's a lot of– I like to make better visualizations and that's hard– making research movies– seems simple.

Making Research Movies

Q.25

Are you talking about when it shows changes over time?

Yeah, so if I were to make– I've done this I've made animations like this front or something and I end up– the easiest thing– MATLAB has something– like an MPEG encoder– and it's not very good– it comes out really grainy with a really big file size and so I end up making PNGs of all my imagal thing and then I have a folder with a thousand PNGs for each frame and then I use this UNIX utility called FFMPEG. [..]

I played around a lot with it but I felt in the process of doing that I was completely over my head. I mean I'm not– you end up getting into the details of like how these things get encoded and as a researcher it's not– it's like you're axis are the same in every

frame but somehow in the file it has to like– every frame has to have like information about where the little axis are and so it makes this huge– to me it seems like it made this huge file size– and so necessary parts of it didn't need to be repeated necessary I don't know if– I'm sure there's a solution out there– better one then I thought

Q.26

there are people in that specialize in encoding video specifically– that's like its own field

Definitely– yeah so I feel completely– and you know I tried to like– [re-assuring yourself voice] "oh, I'm a smart guy I'll to figure this out"– and you get into and you're like "No! This is very complicated stuff", so I use FFMPEG– I'm one of the only people I know who even got that far– most people just use MATLAB– most people just use the simple tool which isn't very...

Customizable?

Yeah, not very customizable, and it doesn't do a very good job I don't think. FFMPEG does seem to do a little bit better if you can give it just– but then FFMPEG it's own limitations, you have to name the files sequentially

Have to have enough zeroes [padding]

YES! And you can only have ten thousand [images]– at least from what I can see– [..]

So yeah that's maybe my two cents, I would use animations a lot more if [it were easier]. You know using FFMPEG is not that hard I guess you can just write your loop in MATLAB and make a map. You know all your map and different and then just point FFMPEG to that directly.

Version Control

Q.27

So you mostly work by yourself on your research projects? I know there's a lot of other co-authors, but specifically– like do you view research projects as more like an indie novel or more like writing [meant making] an indie movie?

That's a good question. I guess starting off in my career– I mean you interact a lot with your advisor and maybe some other people on your committee it was mostly by myself and as I gotten more in my career have noticed– yeah now I work more with people. I still do a lot of– somebody says, "Oh can you..." you know, "Can we look at

this fruit– this thing and this variable", and I just kind of sit here and do it and then print it off and then show them. But yeah we don't have a very good system for you know– if you work on a little bit of code or something somebody else can go and look at your code and say...

Q.28

You guys don't use version control? (!)

No [laughs]

Actually I used to work I used to work with the naval research lab down in monterey– so they run like an operational weather model and they have– so people are continuously working various elements of the model and they have version control for that so– but they really need it.

But with this stuff– we don't work enough together were that would be as useful– I would really like– for a nascent individual sometimes the really good code I use all the time– actually I did this recently where I script I use to read in a lot of my data sets– and it's kind of really nice– it subsets and all this stuff– and I made a small change and it broke it. [laughing] and it was like [regret voice] "I just want to go back to the old one" but I can't– and I know that at the very least I should make copies or something I think I ended up having to ask the computer guy to get the backup because they do backups every night– yeah version control would be very good.

and it definitely sucked because I don't know how much trouble it would be.. [laughing] yeah we need that.

and it depends on what level researcher you get but from me– now that I'm kind of on the faculty doing the sort– making maps and analysis– I end up doing that a lot less then I used to– I just feel like sometimes I just don't have enough time to learn some of this newer stuff

It Takes a Generation

Q.29

I know at some level you get some sort of students to do all your code for you

[Laughing] Yeah, I'm not there yet. Yeah well I have an undergrad student– and I think she'll continue on to be a grad student under me but she's– she's learning MATLAB and stuff but yeah it would be kind of nice to actually eventually to be able to share some

of my code with her– so that she could learn how to use it or something– 'cause I mean there's some of my code that's okay.

Q.30

You got that one folder that's like "Perfect Code"

Haha, yeah there's some stuff that would be– a lot of my stuff would be good for her to learn or something– but for like the student perspective– usually it's good to just kind of– when you're starting out to kind of let them develop their own stuff and 'cause you know in the long term it will be good for them to learn how to do that and– It kind of gives them some examples but then you start in Grad school it's really good to be able to say "ok, can you look at this dataset" and let them go at it and figure it out how to download it and how to read it in and how to extract what they need to.

Letting them developing their own stuff while doing this, it might be soon that LDAP becomes more prevalent

So it takes a generation

Yeah I'd be really happy to figure that out especially now that these datasets– there's these repositories now you don't have to worry about: is it going to be there in two years? or something..

Q.31

That definitely used to be a problem with the web, things would be shut down we'd never seen again

Yeah research data sometimes this is a problem– you have a dataset you really like and the grant ends on that and it doesn't get renewed and so that dataset doesn't get– sometimes it just gets removed, and that's it, it's gone. And so if you didn't download it haha– but with PO DAAC & JPL that's not a problem..

Until NASA gets defunded

Well.. that's where my funding comes from soooo..

[Laughing]

Part II

Analysis of Interviews

Chapter 6

Problems with Current Ideas

- Idea 1 -- Streamline Data Acquisition
  - Currently, getting data is not a problem for scientists – Existing new solutions for improving this are not used
    * (Kennedy Q.9a) – I don't use streamlined solutions – current streamlined data solutions are a failure
    * (Hutchings Q.42) – Researchers like me can get the data – new tools require learning
    * (O'Neill Q.12-13) – I don't use it (Q.13) – "sometimes I need access to more than [streamlined solutions are] designed to have"
    * (Walsh Q.15) – Existing tools are supposed to help researchers find data but don't – "Jury is still out I think on how successful it's been"
    * (Van de Hook ~2:00) – "[The new stuff...] I don't know much people are using these"
    * How data acquisition is actually done today
      · (Hutchings Q.42) – FTP
      · (O'Neill Q.10) – FTP, Bash Scripts
      · (Kennedy Q.9a) – Google Earth and unknown
  - Scientists like collecting ALL the data
    * (Kennedy Q.3) – Has 100tb – would actually prefer to work on *all* the data remotely but technology isn't there yet – might prefer locally if

that he had less data or could manage it easier
    * (O'Neill Q.7) – "it's nice to have it" – Doesn't have to find it again
    * (Shell) – Don't want to lose access to data right before deadline
  - Problem with (1) prototype specifically
    * Scientists don't use Python
      · (Chelton via O'Neill Q.21) Fortran
      · (Strub) – Fortran, IDL
      · (Kennedy Q.9a) – Google Earth, unknown
      · (O'Neill Q.10) – Fortran, Matlab, Shell Scripts
      · (Scientists via Kuuipo Q.13) – R – "R is very popular right now."
      · (Hutchings, Q.22) – IDL – "IDL is like breathing to me"
      · (Shell) – NetCDF Command Language
      · **Exceptions**: Van de Hook & Hutchings's students
        1. (Van de Hook) – Uses it – had Python pulled up during meeting
        2. (Hutchings Q.21) – My students use it – "It's becoming the standard"
    * Scientists don't want to use novice / simplified tools
      · (O'Neill Q.13) (Talking about data acquisition tools): "I think their design is set to be used kind of towards you know, kind of like novice to intermediate" (and therefore I don't think it's useful to me)
- Idea 2 – Finding Phenomena visually where you don't know the Space/Time
  - Researchers like working with the RAW data / Wouldn't accept or trust a tool
    * (Hutchings, Q.4) – Works with Raw Data – "Normally it's level 1b basically" (Q.36) – "Derived data products are more useful to the general public."
    * (Scientists via Walsh Q.8) "[researchers] would not trust– GIS Software,"

    * (O'Neill Q.8) "[my work is] kind of *raw* so I do a lot of stuff like– processing the data from various things and you know I like to deal with the raw data."
    * (Strub) Data values are more important to research than imagery
    * **Exceptions to Using Raw Data**:
        · (Kennedy Q.6) – Robert works with Landsat images
        · (Shell) – Has to work with derived products otherwise work wouldn't be possible. Still programs, doesn't use a tool
        · (Van de Hook) – Works with Landsat images but still programs, doesn't use a tool
    * **Exception to Not Accepting a Tool**
        · (Kennedy Q.9a) Uses Google Earth, but ignores other scientific tools
        · (NASA via Van de Hook) Does say that people he worked with at NASA used tools like envi & QGIS and didn't know how to program – Van de Hook was a bit surprised by this and said they should probably learn to program
        · (O'Neill Q.23) – "I mean I would be open [to using GIS software or tool] if there's an easier way to do something." (Though this was a strange reversal to his attitudes up to this point in the meeting– possibly trying to stay open minded?)
- Finding Phenomena where Space/Time is unknown is very difficult – can't be done visually alone
    * (Chelton Q.1) Eddies can't be found using imagery alone [including SST, etc] – It took about a year to make the model we use to find Eddies and then we improved it over the course of seven years. (Q.3) researchers spend years looking for things
    * (O'Neill Q.21) "I'm trying to figure out ways of tracking fronts– how they evolve and stuff– and it's been difficult"
- Finding Phenomena where Space/Time is unknown is uncommon
    * (Hutchings Q.46) [Tool that would be useful is] Looking for a position on the planet at a particular time
    * (Geographers via Walsh Q.6) "normally a geographer will have identi-

fied the place they are interested in." (Q.7) "Geographers pretty much know the time and the place that they want to do their research"

* (Kennedy Q.7) [My work is] spatial and temporal – "I'm looking at hundreds of images at the same place through time"
* (Jamon) Looks at one place over time
* (Strub) Program/Project usually focused on one specific area
* **Exception:**
    - (Chelton) Ocean research – things in the ocean move around

Chapter 7

Possible Problems to Solve

- (P.1) Software packaging is inadequate
  - Documentation isn't always there
    * (Hutchings Q.40) [Talking about derived data] "can't figure out what's going on without documentation on how the product was gridded"
    * (O'Neill Q.16) "I take somebodies stuff and sometimes times it takes a little bit of time-- to see how it's supposed to be used?"
  - Outside code doesn't work / Isn't fully tested
    * (O'Neill Q.15) "Yeah, occasionally, sometimes I get code from other people that [laughs] doesn't work? It's because it works on their stuff and not mine."
    * (Hutchings Q.27) [About software bugs:] "So you know how endemic they are then-- they are mistakes everywhere!"
  - Code not available in all languages
    * (Hutchings Q.25) "It all comes down to where you find your code, so I've used R-- because there was code available."
  - Outside code isn't trusted
    * (O'Neill Q.18) "Like, I mean I have had people give me code that I didn't trust them so I didn't use it"
  - Researchers are open to using software packages

    * (Scientists via Kuuipo Q.13) [Are scientists open to using packages?] "Yes, yeah. Especially Open Source tools and libraries," -- "[For example] students will start off by learning R then they will *quickly* start using all the libraries that can manipulate the statistics geographically."
    * (O'Neill Q.13) "Sometimes [I look for] utilities like [ellipse routines] And yeah I download it and try it once and if it works like it's supposed to like that then "that's cool"
    * (Hutchings Q.26) "It helps to have access to people's code when they have solved problems" (Q.23) "I think we're now in a world where free sharing information and algorithms is a good thing to do."

- (P.2) Work is often re-done / Wasted work
  - Researchers write software that isn't saved or reused
    * (Chelton, not noted) -- Feel free to re-implement the algorithms I listed in my paper for finding Eddies
    * (O'Neill, Q.14) Shares only some code and only with certain people -- "Yeah. I share it fairly freely. I share my stuff-- at least the stuff I know-- I'm pretty sure it's not buggy [laughing]"
  - Derived Products aren't trusted
    * (Shell) A lot of work goes into derived products but many researchers don't use them
- (P.3) Version control is inadequate
  - (O'Neill Q.28) "I think I ended up having to ask the computer guy to get the backup because they do backups every night-- yeah version control would be very good."
- (P.4) Knowledge is not shared / Researchers have to learn about things outside their domain
  - (O'Neill Q.25) Had to learn how things were encoding for visualization "-you end up getting into the details of like how these things get encoded and as a researcher it's not-"
  - (Hutchings Q.47) "Well we don't even realize that it's that easy to get the data haha, that's funny."
  - (Kennedy, Q.3) "that's a lot to manage and it requires a certain level of expertise and interest in doing the computer management and all that stuff

which not everyone has"
  - (Scientists via Kuuipo Q.16) "researchers are still creating their own data" -- "and don't even know that other researchers exist or that other data exists"
  - (O'Neill Q.17) Researchers don't always use existing formats -- "but sometimes people have like their own binary format or something"
- (P.5) Data can be hard to work with
  - Not indexed in the right way (temporally)
    * (O'Neill, Q.4) Going through to find time series is a pain -- "so you have to look through [millions of files to find one point in each one] and it's kind of a pain"
  - Data not being in the right format / poorly documented formats
    * (O'Neill Q.17) "but sometimes people have like their own binary format or something-- or it's just put into an unformatted binary file"
    * (Hutchings Q.16) -- "if the data is not provided with a way of reading it-- no one else can use it so-- as you said it's useless."
  - Researchers have to deal with a lot of data
    * (Kennedy) Landsat 100tb
    * (Jamon) Landsat 100 gb (currently)-100tb (ideally)
    * (O'Neill) Various 70 terabytes
    * (Shell) Various 10tb
  - Data goes away (or at least used to)
    * (O'Neill, Q.30) "now you don't have to worry about: is it going to be there in two years? or something.."
    * (Shell) You don't want to lose access right before the deadline