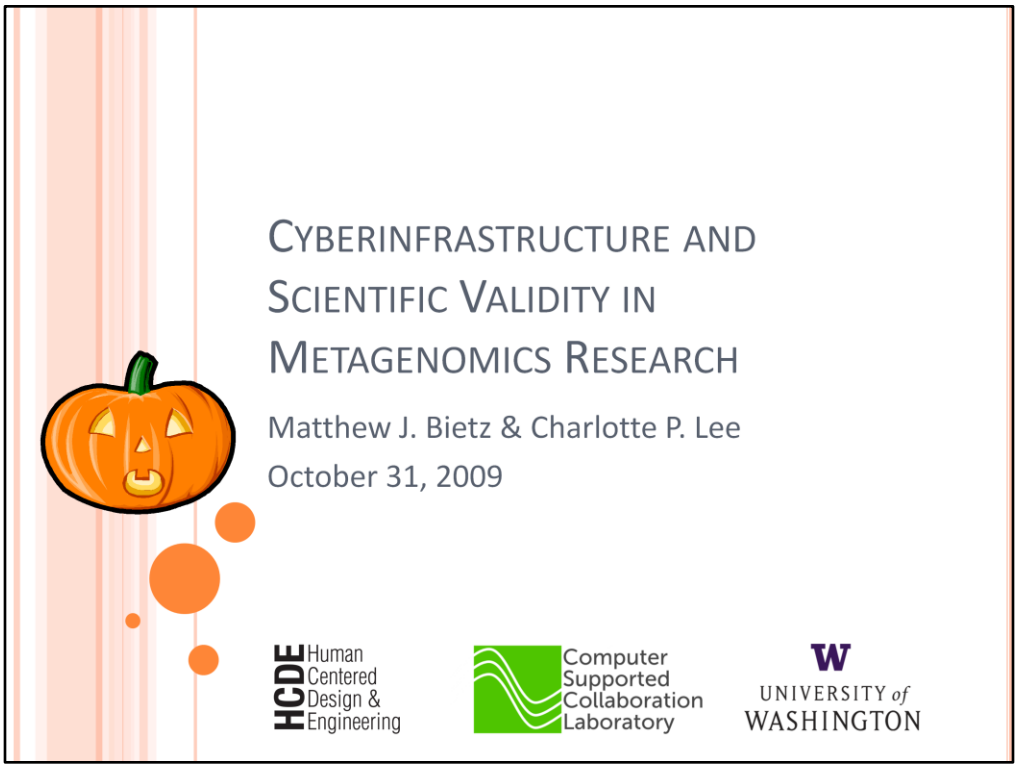


[This talk was presented at the 4S 2009 conference, Washington, D.C., October 31, 2009.]

Today I'm going to be talking about scientific validity in the emerging field of metagenomics, or, put another way, I'll be looking at how scientists create metagenomic facts, and how large genetic databases play a role in creating those facts.

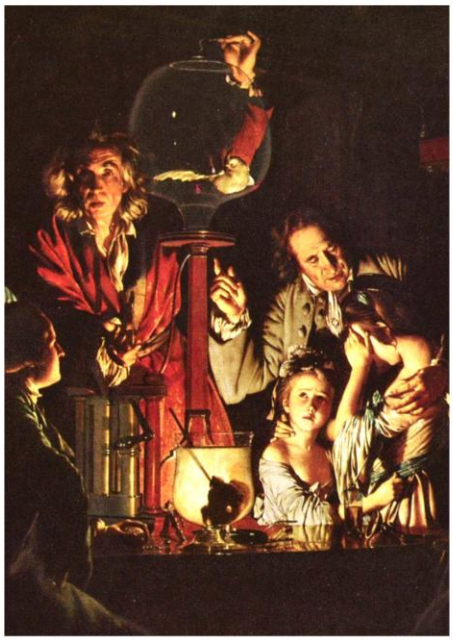
As a matter of fact...

- *Witnessing* is necessary to make matters of fact
  - Woolgar and Coopmans , 2006
  - Shapin & Schaffer, 1985
  - Shapin, 1984

2

I'm going to draw on Woolgar and Coopmans discussions about STS approaches to e-Science, and use the notion of scientific witnessing that was described by Shapin and Schaffer. Shapin and Schaffer discuss Boyle's use of the air pump as a demonstrative device, so that the "fact" of vacuums could be witnessed by others. The performance of an experiment in and of itself was not enough to create a matter of fact—the experiment had to be witnessed. The fact becomes stronger when more people can testify to the experiment—requiring what Shapin, in *Pump & Circumstance,* calls the "multiplication of witnesses."

Shapin describes three mechanisms for creating and multiplying witnesses. The first was eye-witnessing—seeing the experiment first-hand, usually in a social space, and being able to testify to its results. For example, Boyle would demonstrate his air-pump experiments in front of the Royal Society. The second means of witnessing was through replication. You may not be able to see me do my experiment, but if you have the protocol, you (or anyone else) can conduct the experiment yourself. Both of these first two ways of witnessing have significant problems. Eyewitnessing is limited by how many people can see the experiment first-hand. Replication is notoriously difficult. Often others may not have the resources or expertise to do the experiment themselves. And even if they did, trying to redo someone else's experiment is fraught with problems and frequently fails.



Through virtual witnessing the multiplication of witnesses could be in principle unlimited…. What was required was a technology of trust and assurance that the things had been done and done in the way claimed.

- Shapin: *Pump & Circumstance*

3

The third and "far more important" way of multiplying witnesses is through what Shapin calls the "literary technology of virtual witnessing."

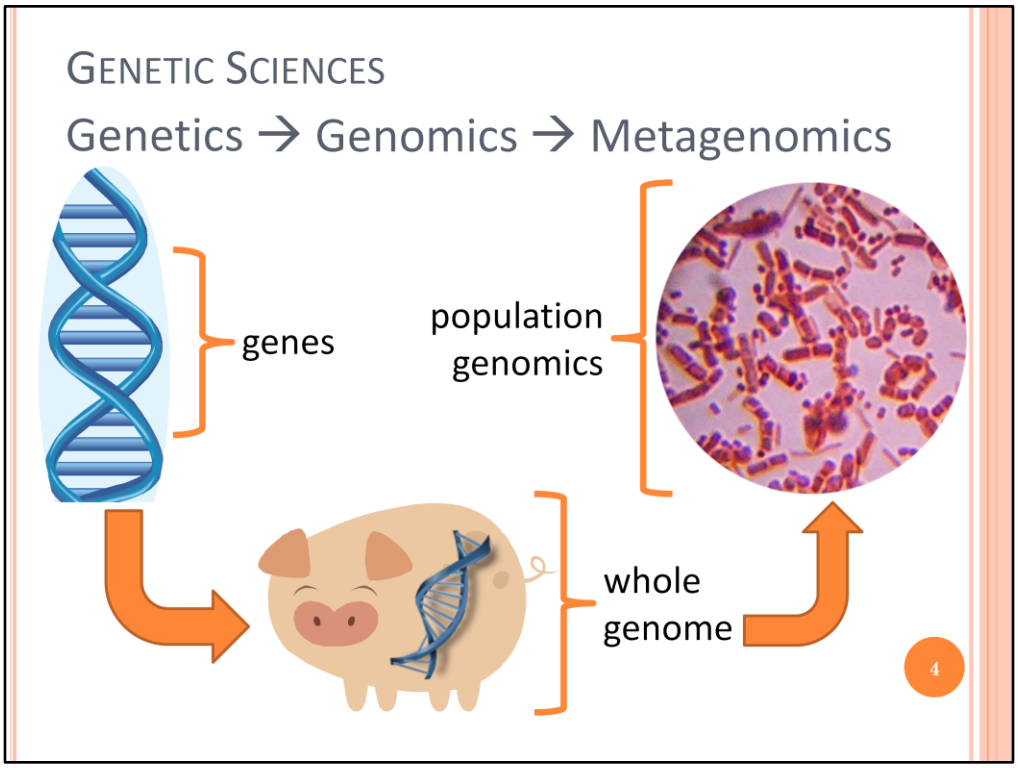
"The technology of virtual witnessing involves the production in a reader's mind of such an image of an experimental scene as obviates the necessity for either its direct witness or its replication. Through virtual witnessing the multiplication of witnesses could be in principle unlimited. It was therefore the most powerful technology for constituting matters of fact. …What was required was a technology of trust and assurance that the things had been done and done in the way claimed." (Shapin, Pump & Circumstance)

Shapin here is referring to the written accounts of Boyle's experiments that were used to convince even larger numbers of witnesses that the experiments had been done and produced certain results. In other words, through the use of words and pictures it would be possible to create a matter of fact, even among those who had not directly seen or replicated the experiment.

I'm going to using Shapin's idea of witnessing in science to frame a discussion of what's happening with databases in metagenomics. In particular, I want to focus your attention on that last sentence—the technology of trust and assurance. Keep that in mind as we go forward.

The work I'm going to be talking about today comes from a study of the development of cyberinfrastructure for metagenomics research. Cyberinfrastructure is a specific class of infrastructure that that brings together people, information, and technologies to support research. Work in this area is also sometimes called collaboratories, virtual science, or e-science. In particular, we're studying the development of infrastructures for metagenomic data storage, sharing and analysis.

Before I go any farther, I want to tell you a bit about metagenomics, or as it is sometimes called, population genomics. I'm guessing at least someone in the room hasn't heard about it before today. I'll share the origin story I often hear from our informants when they describe metagenomics.

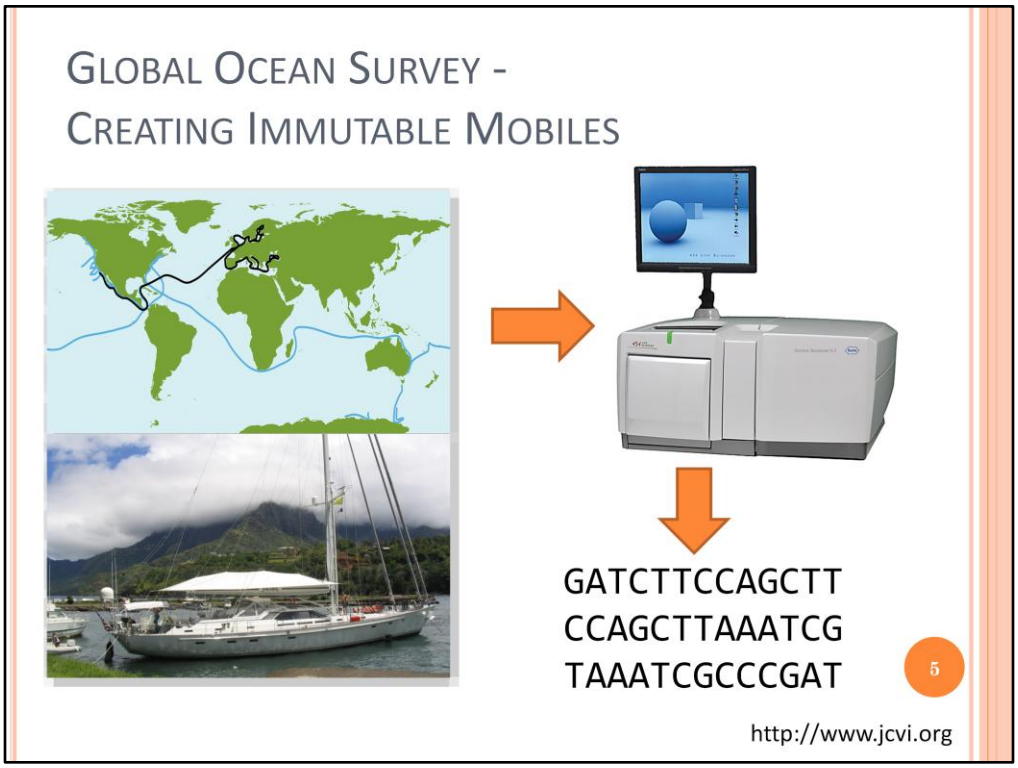


This slide shows a story that we hear from our informants about the origins of metagenomics. The story basically says that genetics came first, and genetics focuses on genes, or functional segments of DNA. As technologies improved, genomics allowed scientists to look at the entire DNA of an organism. Organisms could be compared, scientists could look at how genes interacted, etc. Then, more recently, metagenomics came along. Metagenomics goes to the next "logical" step by looking at the genetic material of entire populations of microorganisms. So that we can ask about the diversity of organisms, or the prevalence of certain kinds of genetic functions within a population.

Obviously there are interesting things to talk about just in this version of scientific history, but the basic idea is that this is an extension of techniques from genetics and genomics. And as a result of this, a lot of the things that I say today may apply as well to genetics and genomics. But my own data and research are in the field of metagenomics, so I'll be limiting my comments to what's happening in this new field.

The term metagenomics is quite new – it was coined in 1998, and a 2007 National Research Council report called it a "new science." Three key factors are often cited as crucial for the development of metagenomics. First, new laboratory methods make it possible to do DNA sequencing without first having to culture organisms in the laboratory. Second, high-throughput DNA sequencing has dramatically lowered the cost of sequencing. Third, high-performance computing allows scientists to store, manipulate and analyze the huge amount of data that is being produced.

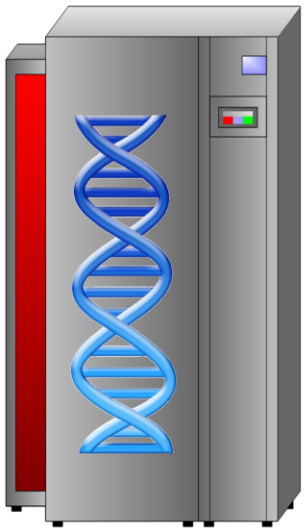
But let me tell you about a specific metagenomic study, the Global Ocean Survey.



This is an ongoing project to attempt to map the metagenome of the ocean. Craig Venter's yacht is sailing around the world, now on its second trip, and every 200 miles or so, it stops and sucks up 200 liters of seawater. The scientists filter the water to collect the viruses and bacteria, and they send those filters back to the lab. The DNA is extracted from all the microorganisms on the filter, sequenced, and they end up with nice inscriptions representing the genetic code of the population of organisms that are present at that spot in the ocean. There are all kinds of interesting STS things to talk about just on this slide, but I'm going to put those aside for the moment and move on to how the data is shared.

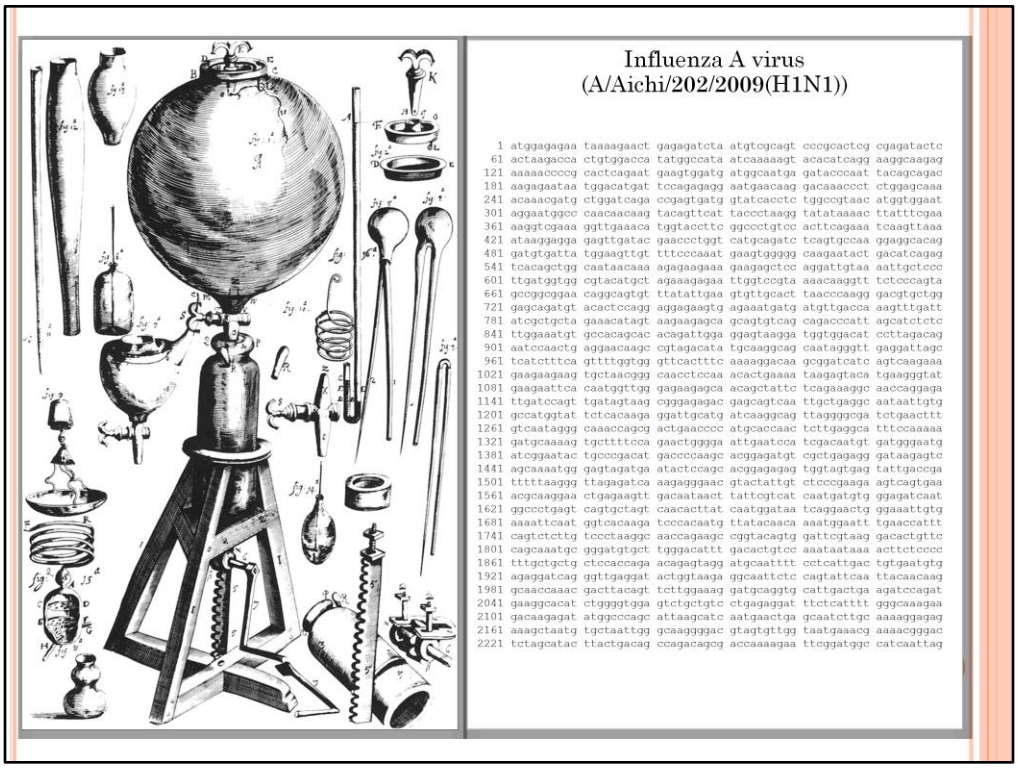
Motivating Data Sharing in Metagenomics

- Required for comparison techniques
- Efficient use of collected data
- Combine datasets for new discoveries
- Ability to replicate analyses and findings



6

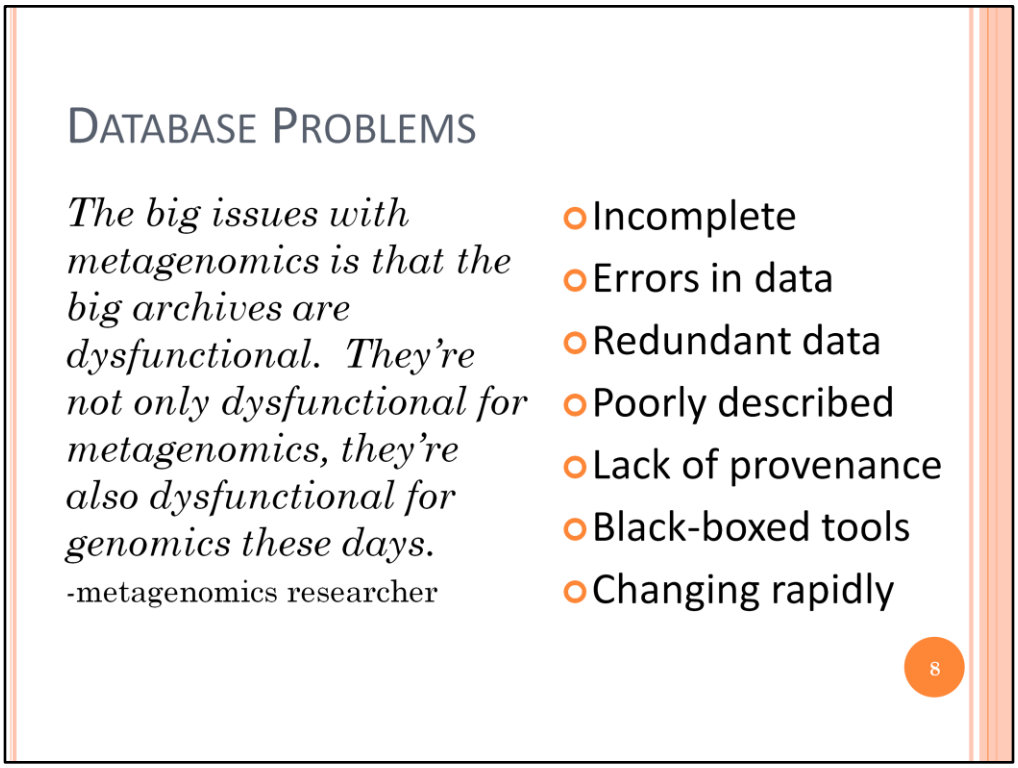
Once these inscriptions are created, there is a lot of pressure to share the data, and a lot of reasons are cited to justify data sharing. Many of the analyses that scientists need to do in the genetic sciences require comparing the new DNA sequence to other known DNA sequences. Also, metagenomic datasets are often of a size and complexity that single investigators simply do not have the time to do thorough analyses, and there is sort of an "if we share it, they will come" sense that someone else will take up the analytical slack. There is a belief that researchers will be able to discover new things by combining datasets that they couldn't have found by looking at the data in isolation. But one of the most common motivations is that sharing data will make the science more robust by providing the ability to check someone else's work—in essence, sharing data creates an opportunity for replication.



That word is important: opportunity. In the same way that Boyle's realistic and technical drawing of the air pump suggests that you could build it and test it if you wanted to, so too does the existence of shared data in the database suggest that you could replicate the results. But, at least among the scientists we talked to, direct replication was nonexistent. This doesn't mean that mistakes weren't found, but it was rare to go looking for them, and there are simply too many sequences to check them all. Essentially, the genre of the database, its conventions, and scientists' beliefs about what it means to have data in the database means that they are willing to testify that this is the genetic code from a specific organism, that it is a particular gene, and that that gene has some specific function. The database and its associated tools allow a scientist to believe that two organisms share a branch on the phylogenetic tree, even though the scientist has not done, seen, or replicated that experiment. The genetic database becomes a technology for virtual witnessing.

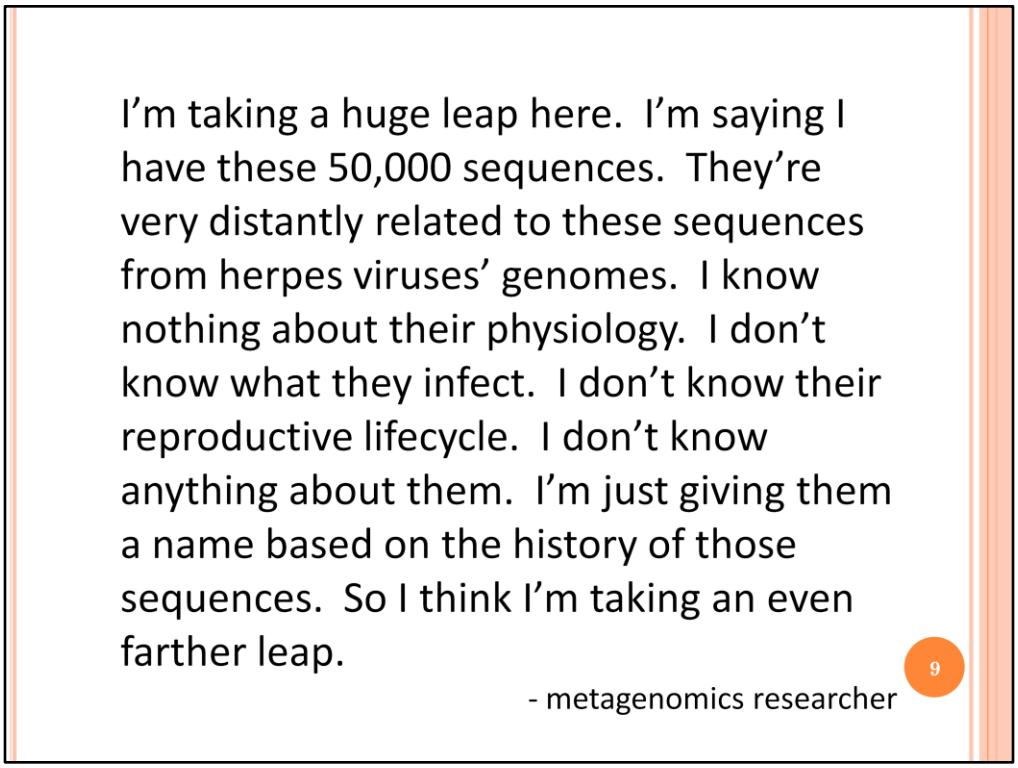
And this brings us back to the quote from Shapin: "What was required was a technology of trust and assurance that the things had been done and done in the way claimed." For Boyle, that technology was a literary one. In metagenomics, the technology is digital, and there's a big question about whether it is a "technology of trust and assurance."

So now I want to bring in some of the words of our informants, and give you sense of how they think about these databases.



What does this researcher mean when he says, "the big archives are dysfunctional?" There is an awareness among our informants that there are some serious problems with the databases they are using. The scientists are telling us that the databases are incomplete, they've got errors and redundant data, the data are poorly described, there's no metadata, and it's not clear where the data came from or how they were processed before going into the system. It's often not clear exactly how the various query and analysis tools provided by the databases are operating behind the scenes. And on top of that, with the pace of sequencing going on today, the databases are changing extremely rapidly. The upshot of all of this is that our informants are concerned that if they run an analysis on System A, they can never be sure if they will get the same result from System B (even if System B claims to have the same data). In fact, they don't even trust that they'll get the same answer from System A tomorrow that they got today. This is a big concern for them - the senior scientists in our study all reported spending more than 90% of their research time in silico.

And metagenomics is a new, and some would say unproven, way of doing science. The scientists using the techniques are already fighting battles for legitimacy on other fronts.

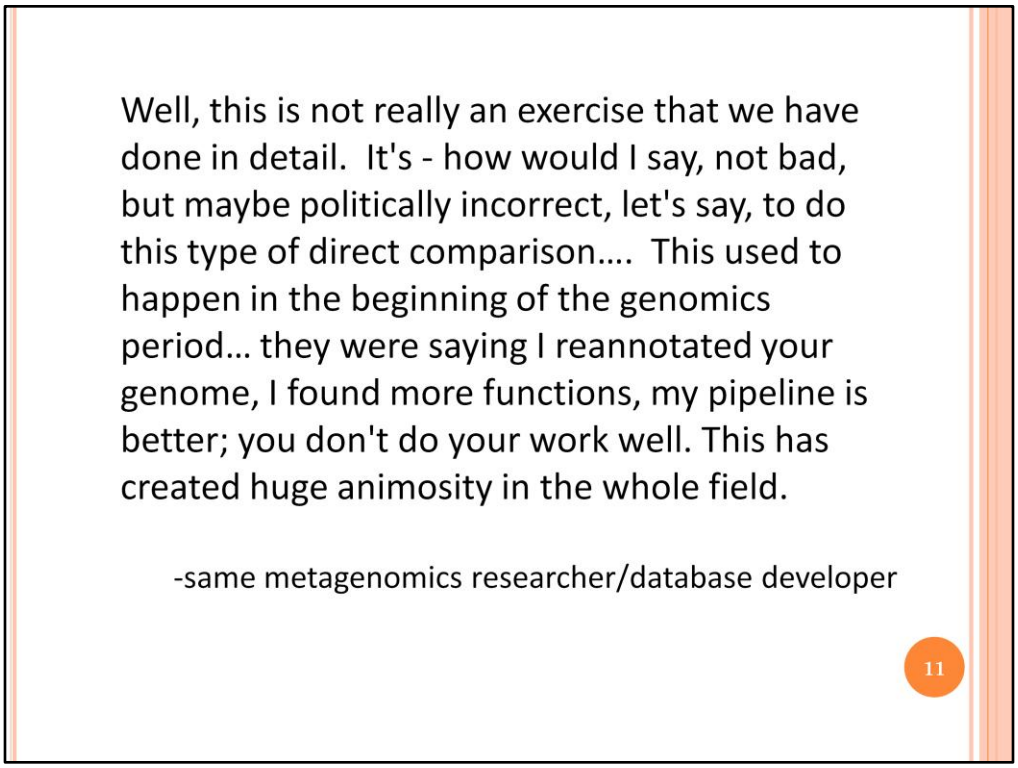


This quotation is from a scientist who is using metagenomics to identify pathogenic viruses in an environment. [read quote] This is a scientist who found it very difficult to get papers published. There was pushback from traditional virologists about the legitimacy of metagenomics as a method. A contributing factor is that the evidentiary status of the databases, one of the primary tools for making these kinds of claims—that one genome looks like another—is in question.

So when we did just an informal comparison, we found that one organism has 500 genes, but the other doesn't. But when we went to compare a much more detailed study, we found that the genes were actually present on the second organism as well. They have just not been identified, those genes…. The reason was that the two tools, the two different methods were working differently. So where one method was predicting that there was a gene, the other method was predicting that this is not a gene.

-metagenomics researcher/database developer 10

Here another metagenomics research who is also working on developing one of these databases talks about this problem in particular—that if you compare on different database systems that have used different methods for annotation of the sequences, you might get different answers. [read quote] What he's saying here, is that different database systems were giving differing results because of minor details in the behind-the-scenes analysis going on in the database tools. But when I asked him whether he was discussing this or publishing about these issues, he responded like this: [next slide]



We heard stories like this from several of the researchers we interviewed, where problems were discovered that could have far-reaching impacts on the fact-ness not only of the database systems but also of the results generated from them, yet the issues were either ignored or dealt with quietly, with maybe the only recognition being in the list of "bug fixes" when the next version was released. There is a sense that everyone is aware that the database has problems, but at the same time, there are social pressures not to report them widely.

There are a lot of people working on trying to make the databases better, to fix the problems, develop quality control mechanisms, etc. And for the mean time, many scientists are simply going on with their work – what else can they do?

I don't view it currently as an issue… [metagenomics] is another tool in the box of tools which you can use to answer more things. It's just a slightly larger hammer…. If there is bias in it, we haven't been able to detect any. If there is any bias, then so be it. There's bias in every technique that we have in our toolbox at the moment, and you can't avoid that. I don't view it as being a particular issue.

-metagenomics researcher

12

Here another researcher discusses his approach to the potential problems with the tools they are using: [read quote] For this scientist and many others, the databases work. Maybe not perfectly, but well enough.

But there was another and even more intriguing response some scientists gave when I asked how they proved that their results were true. [next slide]

We don't really have any hypothesis and that's kind of an interesting thing. I would say that metagenomics is not necessarily a hypothesis-driven discipline.

…

So, it's all about data interpretation, right? You make interpretations as best as the data can actually provide you. It's just the nature of incomplete datasets and it sounds like a really bad thing, but it's just the nature of metagenomics. There is no finish line and there never will be a finish line.

-metagenomics researcher 13

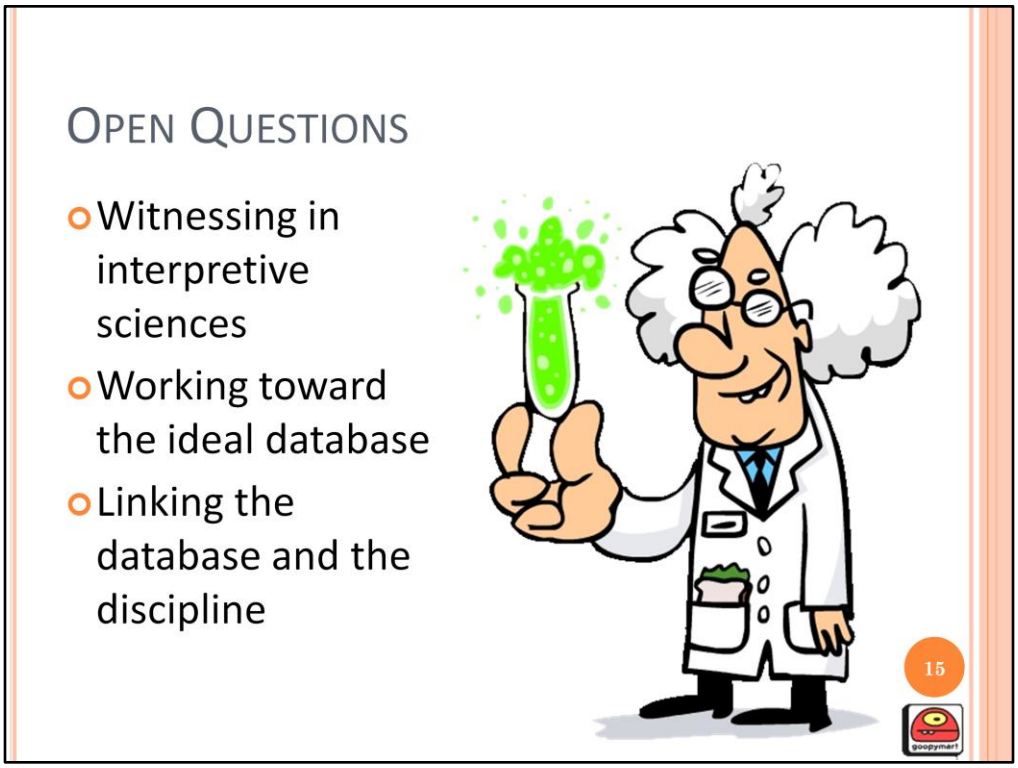
Similarly, another scientist said: [next slide]

Kind of where we're headed in science is away from - the data sets are so large anymore that we're really going to be doing things, you know, initially with, like, regression analysis. But we're just going to go the point of where you don't start with a hypothesis, you look at your data and you extract a hypothesis back out of it. That's kind of what you do with metagenomes already. I mean, it gets so complicated that, yes. It's kind of this petabyte version of the world.

-metagenomics researcher

14

Both of these scientists were trained in and have published hypothesis-driven science, but both of them are suggesting that their expectations are changing about how they are using these large shared data sets to make facts. I think trying to make a causal claim here would get us into a chicken and egg situation—which came first, the problems of the databases or the move toward interpretive science. But in these quotations we can see that the two issues are linked. This is the "petabyte version of the world."



I want to wrap up with a few thoughts about where this line of inquiry could lead. When we look at the database as a technology for the multiplication of witnesses, we end up raising as many questions as we answer. What does it mean to "witness" in an interpretive science as opposed to a hypothesis-drive one?

And I haven't talked about it much, but there is a lot of work going on now to get closer to that ideal of replicability, ranging from improvements in quality inside of the database to a whole set of activities outside of it, including publishing algorithms and releasing software under open-source licenses, publishing standard operating procedures, developing data standards and data sharing protocols, etc. Can we also understand these efforts and artifacts, which are both descriptive and prescriptive, as technologies for virtual witnessing?

Finally, I want to ask how the evidentiary status of the database is implicated in the legitimacy of the discipline, and vice versa? Shapin talks about "linguistic practices in the making," and how these linguistic practices are bound up not only in the production of facts, but also in the maintenance of scientific communities. These linguistic practices still operate in many forms—the database has not supplanted publication. But in this instance, are we seeing database practices in the making, and how are they implicated not only in matters of fact, but also matters of community.


THANKS!
Matthew J. Bietz – mbietz@uw.edu
Charlotte P. Lee – cplee@uw.edu
http://depts.washington.edu/csclab/
Images: goopymart, kaibara87, J. Craig Venter Institute, 454 Life Sciences, Robert Boyle, Joseph Wright of Derby
HCDE Human Centered Design & Engineering
Computer Supported Collaboration Laboratory
UNIVERSITY of WASHINGTON

Thanks!