Group #19: Community of Interest on the Future of Scientific Methodologies

|  |  |
| --- | --- |
| Date | November 2, 2020 |

|  |  |
| --- | --- |
| Participants | |
| Kari Heffner (Jefferson Lab) | Thomas Britton (JLab) |
| Lee Gimpel | Bruce |
| fac- Nami Ishihara | Stuart Gluck |
| Torre Wenaus | Peter Nugent |
| Bei Wang Phillips | Stefan Wild |
| Rob Roser | Dan |
| Stuart Gluck |  |

Contents

[1 Day One - November 2, 2020 3](#_Toc55484898)

[1.1 Discussion: Breakout 1 - Define the Scope of the Problem. 3](#_Toc55484899)

[1.2 Discussion: Breakout 2 - Implications of this Problem. 5](#_Toc55484900)

[1.3 Brainstorm: Day 1 Reflections 8](#_Toc55484901)

[2 Day Two - November 5, 2020 8](#_Toc55484902)

[2.1 Brainstorm: Breakout 3 - Signposts 8](#_Toc55484903)

[2.2 Brainstorm: Breakout 4 - Signpost Plausibility 11](#_Toc55484904)

[Appendix 14](#_Toc55484905)

[Live chat 14](#_Toc55484906)

1 Day One - November 2, 2020

1.1 Breakout 1 - Define the Scope of the Problem.

**The following participants have not been active:**  
Lee Gimpel, Stuart Gluck, Kate Shattuck

**Question or instruction for the discussion:**  
Breakout 1 - Define the Scope of the Problem.  
The purpose of this session is to lay the foundation for the next 5 sessions. That is, each breakout group will define a key piece of technology, a new device, or methodology that would have an impact on how the labs/scientists operate. The details should include answers to the questions below.



**Sticky points:**

 Top Takeaways (5 points per participant)

* What is the problem, issue, technology, device, methodology?
  + critical mass to "do" science can limit scientific output (#2)
  + pedagogy of data sets/working with specific data sets (#1)
  + (2) Accelerate (provable) hypothesis generation with the aid of data scientists (#3)
  + How should we curate data sets and package them so that other scientists can use them (#4)
  + Add in folks doing the large-scale simulation work. i.e. data + models + simulation --> hypothesis generation Followed by feedback into the simulations and direction for new data taking and revised models. (#5)
  + (2) How to best match scientific goals and data science techniques/toolkits (#6)
  + (2) Is AI/ML the "end of theory?" What role will AI play--replace scientists or augment scientists? What parts of scientific methodology will AI contribute to, and what parts will scientists still need to do? (#8)
  + (2) Planning in the 'democratization' of data from the beginning in an experiment. Its product is data usable by the scientific community, not just experiment members. Challenges in metadata as well as data itself. Impact on experiment design, which has a ~20 year timeline eg in HEP (#9)
    - Need more "green field". There is a lot of historical baggage surrounding how science has been done.... (#22)
  + (2) Predicting extensive uses for data beyond original "experiment" (#10)
  + (1) How can scientists deal with the effects of AI/ML models being black boxes? Will scientists understand the models/theories they create? (#11)
  + How should data-centric scientific methodology work? (#14)
  + How should we integrate many disparate data resources with X-AI and then deploy the resulting data products as community resources? (#18)
  + Data and analysis preservation offers a development path from today, where it is sorely needed, to a 30 year vision for community data, analysis and hypothesis testing outside the originating experiment, true reproducibility (#21)
  + (1) For science, do we want to explain black-box machine learning models or do we want to use human interpretable models in the first place (as argued by Cynthia Rudin)? (#24)
  + (1) Which aspects of hypothesis generation and validation are replaceable by AI and which parts are not? (#28)
  + Are AI/ML models simulation experiments or some other sort of tool? (#32)
  + What should be "the scientific method" now that we have AI? Is AI just a new tool, or will it transform the nature of science itself? (#33)
  + (1) What role does data management play in data-centric science? (#34)
  + How do we convey robustness and uncertainty of the data? (#35)
* Who would develop it (basic research to advanced deployment)?
  + Accomplishing the adoption of data science tools/techniques will require developments from a large body of people that cuts across many scientific domains (#13)
  + (1) Collaboration between scientists, computer scientists/IT people, data scientists, and philosophers of science. (#15)
* Who would use it and what skills would they need to use it effectively?
  + Basically this could be used by all branches of science. (#12)
  + Presumably it needs to be used by all or most working scientists, but a question is how specialized each instance should be. For example, should tools/conventions be for all branches of science, for physics, for particle physics, for neutrino physics? (#19)
  + Everyone would use it! But open science is natural place to adopt first (#20)
  + (1) The more curated a too/collection of data sets is, the more easily it can be used. the tradeoff is that then it is less democratic and more "steered"--more open to influence from a small number of stakeholders. (#23)
* When would it be expected to be in production use (N years in the future)?
  + I am not sure this has a specific end date. Most of these things will be adopted over time and/or changed over time in response to the development of the technologies... (#17)
  + In some sense it is already in deployment, but in a haphazard, organic way. It would be probably a 10-year horizon for developing intentional tools/standards and then an ongoing effort to continually evolve and maintain it. (#26)
* Where, and how widely, would it be deployed?
  + A tighter data science / domain bond should allow for a very widely deployed solution (eg common lexicon, robust metadata) (#7)
  + (Meta)Data/lab notebook standards (#16)
  + Meta data including assumption, approximations, uncertainties (especially helpful in data-limited environments) (#25)
  + It seems like a universal need for all of science. (#27)
* What is the setup time and/or process for using it?
  + This should be seamless. We should be able to pull the desired data sets, simulations, and underlying model and run immediately. (#29)
  + We should start on this yesterday. It should be ongoing with constant scrutiny and development (#30)
  + Scientists using AI/ML seems like it will need to be on demand and always available, for smaller scale applications. For applications that require supercomputers, obviously there will need to be scheduling, but these needs will only become more common in the future, so expanding capabilities is a key. (#31)

1.2 Breakout 2 - Implications of this Problem.

Participants: 3

**Question or instruction for the discussion:**  
Breakout 2 - Implications of this Problem.  
Each group will now develop a list of issues and implications for the issue/technology/community they settled on. There are lots of implications for how a technology can be used, or further developed.



**Sticky points:**

 Top Takeaways (5 points per participant)

* What other/companion technologies, services, software/hardware must also be developed and deployed?
  + Prior session (#5)
    - \*Matching\* up : people, technologies, data, etc. (#6)
    - Training and support requirements (#7)
    - Democratization at many levels/stages (#8)
    - Use of data beyond original experiment (#9)
    - Collaboration at many levels (#10)
    - Balance between democratization and steering (#11)
  + Trying to come to consensus, adding notes here (#12)
    - quicken the hypothesis/testing \* coming up with hypothesis + testing it \* metadata stuff is a means to that (#13)
    - connection between data science and hypothesis generation gap between doing data mining and explaining the underlying physicscurrently more correlation, want to see more causation (#16)
    - standards + management of metadata - real thread (#15)
    - Need to be additional incentives, resources for investments in long-term code (#14)
      * Some times the code should be engineered and sometimes not (#17)
      * Sometimes there should be diversity in the way we code things (#18)
      * Same thing for data curation and data management (#21)
    - Having metadata standards would be useful so people could learn what data might be useful to them (#20)
    - Compared to 30 years back, the toolset is much better for software engineering etc. So much better that it makes it possible to curate in a way that wasn't possible before. Easy to integrate good practices - the result is something much more capable of being curated than before. (#22)
    - The idea of having fair sharing of data so that we can use each other's data is changing scientific practice. The science 30 years from now will be different because of data and AI. Science has to focus on prioritizing the production of data for the sake of just having enough data for everyone to use - and finding standards and tools for sharing that data with each other. (#25)
      * Does that necessitate the changing of funding and publication requirements? (#26)
        + Even 20 years ago, if all you did was generate data, that wasn't considered science. Now, it is. We have DOIs etc to get credit for data (#27)

Evolving, but need to overhaul incentives to encourage this. Needs to be included in the journal editorship/prioritization and funding agencies. Some funding agencies are coming out with calls for data mining, but few and far between. (#28)

* + - * Agree with prior comments about incentives. In computer science, don't have good journals for this - to publish a software paper. (#33)
    - Science model is different than a data model. Data science is more like a collection of preparation and analysis. (#29)
    - Thinking forward, use backward look to help think of vision. 30 years ago the internet was created. What does this look like going forward? Data sharing and data silos? (#30)
      * Look at private sector and what they are doing with data too. Fundamental societal problem because it's easy to talk about sharing data for better science but there need to be real incentives to promote this. (#31)
    - Want to be able to generalize the process for hypothesis generation. Compare two data sets that have similar characteristics. Can our model for comparing data sets be adapted for hypothesis generation? Can we transfer the learning to the hypothesis generation process? (#32)
    - After analysis, have the semantic layer for interpretation. Data curation -> Analysis. In climate science, the data analysis requires a lot of computation. Then interpretation requires some steering. (when in the process does this happen? need some curation process) (#34)
    - Even the centralized aspects of genomic science / data curation is becoming less centralized/curated. (#35)
      * Is it a good thing to democratize of all of these stages? Maybe need to pick places where good versus where steering is good? Where do we need expert guidance? (#36)
        + Democratization means that related scientists can get access to the information. Still experts, but the data etc is more available in the community. (#37)

When we talk about democratization are we just talking about better/more open sharing of data? (#38)

Could look to astronomers as a model for sharing data. Also climate. (#39)

Very different versions of physics resources - colliders versus light sciences. Light sciences tend toward individual PIs. Consider the natures of the facilities. (#40)

* + - With COVID-19 work, it's a very different attitude/spirit - can we learn from this? Much more cooperation/sharing. We all have a common interest. How do we drive this feeling of common interest into other areas of science? (#41)
      * We have a few reasons why our area is a little different. 1) Data complexity, 2) culture (people have invested in their own labs), and 3) funding. All the COVID resources was instantaneous funding that was dependent on sharing of information. Without changing of incentives, can't see things changing. This is how it was done in astronomy - using the Hubble telescope as an example. One year to use data, then it will become public. (#42)
        + Use carrot and stick model, worked well in other domains too. (#43)
        + 1/3 of astronomy budget goes toward data management etc. Even for simple projects. (#44)

Totally different in other domains (e.g., HEP) (#45)

* + - The rise of data science and AI is changing scientific practice and should in principle allow the analysis of shared data or the separation of processes of data gathering and analysis. To truly enable FAIR data sharing will require social engineering in the scientific community to create awareness of issues, incentives for sharing, and conventions for standardization and interoperability of data management. Organizations or templates will be needed for maintenance and evolution of standards. Labs of the future should serve as nodes supporting collaboration and sharing research data throughout the scientific ecosystem. (#64)
  + Getting closer to consensus (#24)
    - Be able to formulate and test a hypothesis based on data that isn't your data (#23)
    - Need to change culture of science (#46)
  + What other/companion technologies, services, software/hardware must also be developed and deployed? (#1)
    - Cultural changes - need to start right now to see wide adoption many years down the road. (took 30 years in astronomy) (#48)
      * In HEP for example the culture change is huge, and technically very difficult because of data complexity (for one). If for example the experiments at the 30 year lab of the future were mandated to open their data after N years of exclusivity, that would generate ~30 years of work to make it possible. And that stick would require the complement of the carrot that pays for it, cf. Peter's comment that ⅓ of the LSST budget is data management. (#63)
    - Can we have a stronger form of AI that can perform both inductive and deductive reasoning? (#50)
    - Consider data governance - databases of metadata (#59)
      * Set the standards that everyone else uses when they collect/categorize data (#60)
      * Groups to define data standards (#61)
      * Want the metadata to be automated - want the AI to be doing a lot of this (#62)
  + Who is/will develop this companion technology/service? (#2)
  + What skills/knowledge does the end user require? (#3)
    - Could have specialists - some who gather data and others who analyze data. (#51)
      * Could have people who engineer the experiments separate too (#52)
      * Lots of different types of people (#53)
      * Simulation people - where the science should go next (#54)
      * What i'd love is that everyone is sitting at the table together working on this. (#55)
        + Engineers, CS, everyone in between (#56)
        + This does exist in particle physics - but it's still in the same fence because it's so complex (#57)
      * Even with specialization, need to focus on how they collaborate. AI will not fully replace a human to break the interface/specialization boundaries. (#58)
    - Something that I'm doing - when someone develops another tool, have them teach/document it. They have to hand it over to someone. (#65)
    - Balance between using "black box" libraries versus reinventing the wheel (#66)
      * See this as a general trend - using open source library to plug into the code - this is dangerous because we don't know the details/quality of the library. (#67)
        + Then the reverse - Someone working on matrix reductions - wrote his own library instead. (#68)
      * Shouldn't reinvent the wheel every time either - there's a balance somewhere that we struggle to find. (#69)
  + What are the training/support requirements? (#4)
    - incentivize good practices. Keeping in mind the give and take of diversity of code (#19)
    - Money - dramatic ⅓ of budget that we see in astronomy (#49)
  + The rise of data science and AI is changing scientific practice and should in principle allow the analysis of shared data or the separation of processes of data gathering and analysis. To truly enable FAIR data sharing will require social engineering in the scientific community to create awareness of issues, incentives for sharing, and conventions for standardization and interoperability of data management. Organizations or templates will be needed for maintenance and evolution of standards. Labs of the future should serve as nodes supporting collaboration and sharing research data throughout the scientific ecosystem. (#47)
* Who is/will develop this companion technology/service?
* What skills/knowledge does the end user require?
* What are the training/support requirements?

1.3 Day 1 Reflections

Participants: 0

**Brainstorm question or instruction:**  
Day 1 Reflections  
This area is for the Moderator to note key discussion points to summarize what was accomplished in Day one. Remember that day one is focused on Identifying a new technology or methodology and identifying the implications and possible consequences of it. The moderator can populate this individually at the end of the day or request input from the group here.



2 Day Two - November 5, 2020

2.1 Breakout 3 - Signposts

Participants: 1

**Brainstorm question or instruction:**  
Breakout 3 - Signposts  
What we are looking for is technology or social trends that would give us clues that we are on the right track. o How would precursor technologies/services be identified? o What are the precursor technologies/services? o Is there a rank order for when specific technologies/services need to be available? o What DOE or Lab policies need to be in place now, in 5 years? o What facilities need to be in place now, in 5 years?



* 1. Revisiting "what is our idea" - consensus is that we were not thinking out far enough into the future
  + Comments
  + What is our 30-year goal?: AI to participate in deductive reasoning, design of the experiments and simulations runs to go from theory to hypothesis to observation to confirmation. (#8)
  + Need to consider the "Captain Kirk" approach to doing science. Issue commands from your chair - find data, etc. (#2)
  + inductive versus deductive reasoning, let's say that we want the AI to be doing the deductive reasoning. (#3)
  + Design next generation experiments too (#4)
  + Then would it also come up with the hypotheses?Here's a data set and what are explanations about how the data might be explained (#5)
  + Explainable AI/machine learning is doing exactly this - explainable components. Response - But you set up the explainable components - The data would request new data sets and design new experiments for us Talking about AI as an autonomous scientist Take it to that level - and maybe bring it back to something that's reasonable for 30 years Whatever AI means, it needs to be fully explainable (#6)
  + AI could be developed to the point where it could be the devil's advocate Model dependence reasoning - At least a recommender of known hypotheses Then this takes us back to the data format standards - needs to be accessible forms Maybe a signpost Also consider social engineering challenges, incentives (#7)
  + In the future AI/data science techniques will have developed to a point where they can aid in the deductive reasoning; sifting through the masses of data and delivering competing models to describe a data set. These suggestions could form the basis to further study or experimentation . The role of the scientist would then be gatherers/curators of new data sets for analysis, not stripped of the theorizing but collaboratively generating these new theories with AI. (#9)
  + What our goal look like for the future for various domains? (#10)
    - Nuclear Physics: Particle tracking with AI, lots of things with AI. (#11)
    - Astrophysics: Driven by theory simulations that are then merged with observations. (#12)
    - Biology Could imagine a massive AI model that crunches data and designs experiments to test the hypothesis. Robot-driven labs (ginko facility?) huge lab spaces are empty because they are automated - everyone is working on the data. Human in the loop - how to focus those efforts. There's a cost matrix about what you can afford for your AI model too. Already automating phenotyping, a range of different types of laboratory procedures. How can you tell the AI what's interesting and affordable. (#13)
    - HEP: Not really looking for AI to write the code for us, but want it to alleviate the need for us to write the code. Movement in the last few years has been toward the acceptance of the black box in the machine learning process. Has had some success - understanding of what's going on in the process of the neural network. The most forward looking adventurous approaches are looking back at the raw data - even the uncalibrated raw data. Throw massive data into massive resources. The automation and anomaly detection are going to be very important. In our area, detector design and cost optimization will be important in the future. (#14)
      * NP - data collection - treat the accelerator and experiment as one whole system. Automate data taking, diagnosing problems, etc. Zero person shifts - or remote access to information. (#15)
  + General thoughts on what else might be included (#16)
    - Example - anomaly detection is a good example - cuts across all the fields but implementation would be different in different domains. (#17)
    - if we are going to think about things like autonomous data gathering by data systems - there's an area of logic called paraconsistent logic. What if I am getting conflicting readings from different sensors - how do i reason through that to find out what's going on? (#18)
    - The big debate has been about the role of theory making. There is no doubt that AI can take on the role of looking at correlations, etc. Can also do hypothesis generation in terms of making predictions for future experiments. It has not been able to do - theorize, understand what is going on and develop a theory. May also depend on the scientific domain how far we can take AI. (#19)
    - Biology - already finding equations from data. At some reasonable scale and coming up with approaches. If you consider a mathematical model to be theory, then AI is getting us toward theory. A lot of people use a conceptual model as a theory, which is a pretty picture of a set of hypotheses. Doing 21st century science and converting to 19th century publication model. Converge the story into something small enough that an editor can grok. (#20)
    - Mundane thing about black boxing. For example, climate - modeling hurricanes means making pictures/videos of hurricanes that you share. But the new model is that you don't even know what the AI is doing, what is the definition of a hurricane, etc. Ideas of model includes mathematical model, hypothetical model, ability of the model to explain. (#21)
    - There is something to be gained by using the entire system that builds the experiment - may be too complex for any single person. But maybe the AI can understand all of it - and we could explain parts of the system. Maybe AI isn't a black box - maybe it's a non-newtonian fluid. Maybe we can pull out small pieces of it that can be understood. (#22)
    - Given we are talking 30 years out... We want to push AI to be a full partner in the ability to make new predictions. AI models are already looking at data and doing things like understanding the ideal gas law. It doesn't know what the ideal gas law is, but the more that we push down these paths, the closer we get to where the AI can start proposing new hypotheses, linking it to a theory and making new predictions. (#23)
    - We want AI to be a colleague in hypothesis testing. (#24)
* 25. Breakout 3 questions
  + Comments
  + o How would precursor technologies/services be identified? (#26)
  + o What are the precursor technologies/services? (#27)
    - Data needs to be accessible and captured with the metadata. When publish a paper, need to include the data necessary to make the plots (x, y data points that were generated. Need to make all of the data accessible. (#40)
    - Agree with explainable AI as a research area. (#43)
    - Agree with data governance, FAIR data standards, metadata governance. (#44)
    - Deep learning, neuromorphic computing. Paraconsistent logic, some of the math underlying these tools. (#45)
  + o Is there a rank order for when specific technologies/services need to be available? (#28)
  + o What DOE or Lab policies need to be in place now, in 5 years? (#29)
    - Not exactly DOE but most aligned here... Financial market engineering - huge investment in scientific publications, have zero interest in this changing. Not trivial - from corporate point of view but also from point of view of society. Is definitely harming science. Forced to dumb things down. On top of that, we've created a scenario in which the certain types of scientists are recognized over other types of scientists and that restrains science from moving forward, all because of this 19th century construct that we are working in. (#32)
    - Should also address the need to look at different models for acknowledgement when creating or curating datasets or creating algorithms that mine the data sets. Many of the most key people are being pushed out of the academic world into industry. (#41)
    - Need to fund with sharing caveats. For example, you have a year to work with your data and then it needs to be open. (#42)
  + o What facilities need to be in place now, in 5 years? (#30)
    - Then more supercomputers and more capable supercomputers (Exascale, etc). (#46)
      * From NP - have tracking code... Most of the exascale computers have power in the graphics card, not usable by the guy who wrote the tracking code. (#47)
      * Exascale massively parallel - need to invest in software that can exploit this resource. (#48)
      * Nowadays, less resource intensive. Bottleneck isn't as much with the processing itself as with the management of the data/information (network, etc). (#49)
      * If you have more cycles, you can solve more complex problems or you can try more things. Can think of having two models instead of one really complex model. (#50)
      * Having spare cycles would give us the opportunity to try more things. (#51)
      * Would like to see more powerful machines that are more powerful than the added complexity of the new challenges - want overhead room. (#52)
    - Other ideas (#53)
      * When you develop a brand new experimental facility, must have networking, tie to DOE HPC facilities, storage and data curation (#54)
      * Need someone to take on software engineering and maintenance challenges. (#55)
      * Need an ecosystem that includes the intertwining pieces of the science. In calls from ASCR for example, should see a wholistic picture of the science that includes things like software engineering, data curation, etc. (#56)
      * Long term maintenance of code used in experiments (#57)
      * Can there be an office of software production and maintenance (#58)
        + The danger there is that there needs to be a tight coupling between the application space and the domain space. Otherwise, you will lose the plot. (#59)
        + When doing computer science for computer science's sake, often go off in useless directions (#60)
      * Can DOE take the risks on new cutting edge ideas? (#61)
      * At least force the programs to think through the data storage and management and software engineering, etc. (#62)
      * If we are thinking about the labs, we should think about them not just as physical labs but also in terms of their role supporting the data infrastructure. Physical design should include data storage, edge computing, etc. Need to also serve as nodes in the data ecosystem. (#63)
      * Consider better sharing of tools within domains. (#64)
  + Other comments (#31)
    - What role does pedagogy play in science. If someone comes up with a great theory and it is inexplicable to all of mankind, of what use is it. Maybe this is the role of the scientists. How important in the future is pedagogy to the field of science? (#33)
    - History - Planck came long and found equations for black body radiation - phenomenological. Then Einstein came along with more fundamental insight/theory. AI is starting to do the phenomenological theorizing pretty well. Can it do the fundamental theorizing. (#34)
    - Explainable AI does give us the why. We are an increasingly data-driven society and that is the wave of the current/future. Big data/machine learning/etc. Does that drive us away from the Einstinian model of science? (#35)
    - Will not push people out of theorizing - will push them into a deeper more fundamental theorizing. I wouldn't be surprised if the phenomenological theory would be done more by AI. (#36)
    - What kind of science will be funded in the future? Will we only be doing phenomenological theoretical work? Or will we continue to fund the Einsteinian model of science. (#37)
    - With modern science and the nobel prize - how big a team do you give it to? For some big projects, thousands of people produce the science. As opposed to "the olden days" where it was easier to find the lone scientist who would produce a paper. Humans tend to do very badly with very large numbers of people. So modern collaborations in science might include people who don't know anything about each other. (#38)
    - An aside - nobel prizes have always favored experimentalists not theoretical work (#39)

2.2 Breakout 4 - Signpost Plausibility

Participants: 1

**Brainstorm question or instruction:**  
Breakout 4 - Signpost Plausibility  
Now that we have the list of signposts, the groups need to consider how plausible they are and what DOE needs to do to either ensure they happen or the implications of them not happening. o Who is actively working on these precursors? o When would these precursor technologies/services be needed? o What active or pending research programs need to be in place now? In 5 years? 10? o What existing or planned facilities need to be in place now? In 5 years? 10? o What software services or capabilities need to be in place now? In 5 years? 10? o How successful has the community been in meeting previous goals?



**Sticky points:**

 Top Takeaways (5 points per participant)

* 1. What are the precursor technologies/services/etc?
  + Comments
  + Less that we need new things than that we need to pull things out of the shadows (#2)
  + We aren't doing most of the building - we are drawing upon the wider community (well, we are doing some of the building) (#3)
  + Keep in mind that there is a community of data scientists - apply AI / ML models to whatever applications that they come across (business world, election polling, etc) (#4)
    - Not being intentional in designing it for use by scientists (#5)
  + In trying to hire AI/ML employees (#6)
    - Pay structure is not amenable to competing with private sector in these areas (#7)
    - Suggest investment to keep these individuals in domain of science (partnered with scientists) (#8)
    - Get excited about compelling domain use cases (#9)
    - In the future may be competing with startups rather than the big players (#10)
    - Is there an appropriate reward/incentive structure for non-domain scientists (ML/AI/CS) (#11)
      * Yes, it's part of the supervisor's responsibility Then there are people who aren't a fit for that career track (#12)
    - Incentives are important - having a prestige associated with it is important (#13)
    - Training is important, in terms of a pipeline (#14)
      * In bio, they have a specialty of bioinformatics Otherwise, might be trained in overlap of multiple disciplines (bio, IT, etc) Why isn't there a specialty in grad school for coding/etc? Should be a viable grad school path (#15)
  + Interagency cooperation on data management standards/plans (domain by domain) (#16)
    - In CS, called ontology - categorization of data model (#17)
    - Debates on gene ontology in biology - include philosophers in this discussion (#18)
    - Community curation is much harder as the ontological space becomes more complex (#19)
    - Also no incentive structure - no incentive for annotating 10K genes (#20)
    - Facilitate early career training to take people who are in a specific domain and pull them into AI: naturally progress from "i got my phd in physics, i have an interest in AI/data science, help me hone those skills" (#21)
    - Consider developing new markers to identify career growth/progress (#22)
      * E.g., use of software library by others is an equivalent of a citation (#23)
      * Analogous markers that give credit for data curation, sharing, metadata development (#24)
* 25. Facilities
  + Comments
  + Finding an adequate gpu to accelerate my training was not easy (#26)
    - Current facilities need to make sure the resources are adequate (#27)
  + Care about development time in iterative development - need something that's a development and training platform (#28)
  + DOE should fund the platforms, encourage and provide recognition (#29)
  + Possibly a separately funded facility just for training and development (#30)
    - Local facilities have advantages too (#31)
  + Motivation is the building of a new user base (#32)
  + Summit has 27K (#33)
  + Interesting queuing policy implications for this use (#34)
    - Many short jobs might be disincentivized by the queuing policies (#35)
  + In 10-15, have specialized facilities for exactly this work (#36)
    - Quick turnaround time, quick startup time (#37)
    - Was a program called Panda that used Titan to do exactly this (#38)
      * Trying to do it on Summit (#39)
        + Software stack is very fluid, rapidly changing (#40)
        + Containerization is the answer but it's a complicated answer (#41)
  + Need to consider how users are managed (#42)
    - Want to make the resource available for "any researcher" - very large community (#43)
  + Signpost: Some examples where some ML is coupled directly to an automated instrument to rapidly generate and test hypotheses. (#80)
* 44. Research Programs
  + Comments
  + NSF has done AI institutes - DOE should do something similar (#45)
    - Form a table that everyone can then come to (#46)
  + "Coopetition": Cooperation versus competition between the laboratories - should encourage more cooperation (#47)
    - Cross appointments to get the labs to work together is a good idea (#48)
    - Need freedom, incentive structure to work together (#49)
    - Take the COVID work as an example (#50)
    - Virtual labs (e.g., NVBL) as a model (#51)
    - Extramural cooperation - maybe pull people in from outside the laboratory complex (or vice versa) for sabaticals (#52)
    - Safety nets for people who want to take a risk to learn/grow (#53)
  + Formation of Cross-Office programs - to incorporate data science with the domain science. Maybe beyond SciDAC and efforts like ECP. (#77)
  + Signpost: Data science as a standalone discipline and integrated component of scientific education and training. Establish foundational principles about the application of data science. (#78)
* 54. How successful has the community been in meeting previous goals?
  + Comments
  + The facts that we haven't done it suggests that maybe it can't be done (#55)
  + Will require a lot of careful steering (#56)
  + If you don't do anything, the 30-year goal will never happen- have to meet some signposts (#57)
  + Indications of success - (#58)
    - New faces, new people working to implement AI into the workflow of NP - would be able to see if things are changing (#59)
    - How long do teams need to wait in order to do their work - how quickly are they able to run the models? (#60)
      * Can depend on the size and style of the compute power needed. Creating and training the model can take a long time (#61)
  + Are the current metrics for using the facilities good metrics? (#62)
    - Data driven models now starting to replace some of the simulation (#63)
* 64. Signpost - major breakthrough
  + Comments
  + Would it be good for DOE to have a program to fund breakthrough/grand challenge programs? (#65)
    - Start with a bunch of seed projects in first 5-10 years (#66)
    - Set stage for bigger top-down projects in the future (#67)
    - Universities would like to see smaller ideas bubble from bottom up right now (#68)
    - Any long-term applications project should be defined in terms of applying the tools to the requirement (#69)
* 70. Have to fund tools and techniques but even now, these tools and techniques are rapidly changing (consider rapid development of TensorFlow)
  + Comments
  + With QIS, there is the question of how we do Quantum AI. (#71)
    - How much effort do we spend making sure a tool is right versus actually using a tool to build something (#72)
  + QIS will bring a completely different code stack for everything - everyone will have to start from scratch (#73)
  + Consider this as a potentially hugely disruptive force in our own 30-year roadmap (if it generalizes) (#74)
  + Standardized ways to report the limitations of models and the incorporation of that in publications, funding calls, etc. This is both important for experts as well as communicating to the broader public. (#79)
* 75. Other notes
  + Comments
  + Signpost: Future ECP-scale efforts around convergence (#76)

Appendix

Live chat

**The following messages were exchanged via Live chat.**

* Nov 2, 2020, 19:42 UTC
  + based on my listening in the first session and comments from the other moderators, i felt that having everything in one card would be useful (Kari Heffner (Jefferson Lab) | Nov 2, 2020, 19:42 UTC)
  + so if you agree, i've put all the card topics on the green card (Kari Heffner (Jefferson Lab) | Nov 2, 2020, 19:42 UTC)
  + i've also done my best to capture the top-voted comments from the previous session on the green card too (Kari Heffner (Jefferson Lab) | Nov 2, 2020, 19:43 UTC)
  + sounds good (Thomas Britton (JLab) | Nov 2, 2020, 19:45 UTC)