Group #8: Community of Interest on the Future of Scientific Methodologies

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

|  |  |
| --- | --- |
| Participants | |
| Mike Hildreth | James Ang |
| Ilkay Altintas | Irene Qualters (LANL) |
| Sergei V. Kalinin | Ray Grout |
| Nagi Rao - ORNL | Lee Ann Kiser |
| Stuart Gluck | Verónica G. Melesse Vergara (ORNL) |
| Manish Parashar | Bethany Lusch |
| Ramana Madupu | Huolin Xin |
| James Fairbanks | Adam Arkin |
| Josh Levine | Thomas Britton (JLAB) |
| Lee | Bruce |
| \*fac- Nami | Shane Canon |
| Kate Evans | Torre Wenaus |
| joshua elliott | Maria Chan |
| Stefan Wild | Kari Heffner |

Contents

[1 Day One - November 2, 2020 3](#_Toc55484683)

[1.1 Discussion: Breakout 1 - Define the Scope of the Problem. 3](#_Toc55484684)

[1.2 Discussion: Breakout 2 - Implications of this Problem. 6](#_Toc55484685)

[1.3 Brainstorm: Day 1 Reflections 8](#_Toc55484686)

[2 Day Two - November 5, 2020 9](#_Toc55484687)

[2.1 Brainstorm: Breakout 3 - Signposts 9](#_Toc55484688)

[2.2 Brainstorm: Breakout 4 - Signpost Plausibility 12](#_Toc55484689)

1 Day One - November 2, 2020

1.1 Breakout 1 - Define the Scope of the Problem.

**The following participants have not been active:**  
Lee, Stuart Gluck, \*fac- Nami, Lee Ann Kiser, Manish Parashar, Torre Wenaus

**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?
  + Integrating the traditional scientific method with data science to accelerate discovery and understanding. The technologies exists in a limited form today but will presumably advance and mature. (#2)
  + Where can data science complement experimental/computational science? (#4)
  + Knowledge representation for key ecosystems science as an example that allows causal and mechanistic modeling to be connected to pure statistical approaches effectively. (#5)
  + To what extent is data science a separate discipline vs. a way of doing physics/chemistry/etc. (i.e. you can approach physics through theory, experiment, computation, or data science) (#7)
  + (1) What will an ecosystem that integrates data science with modeling and simulation look like? How can we integrate experimental data sources into computing facilities? (#8)
  + 1. The bulk of the data science is now based on the progressively more massive deep learning networks, correlative in nature, with the progress associated with the increase of depth, complexity of architecture, and nature of the objects (from tensors to graphs to transformers). This extensive development pathway is limited in how far it can scale (simply by e.g. amount of produced electricity). How do we transition or integrate this approach with deep Bayesian hypothesis driven character of physical sciences, where even small-data evidence can lead to fundamental change in paradigm? (#9)
  + Identification from large data where areas of ignorance are preventing the ability to answer key questions and model-driven designs of experiments to address these areas. (#10)
  + Useful models need to be a seamless combination of theory, computational, experiment without friction in the system - how to remove the 'iterative' nature of the process and shorten it to near real-time? (#11)
  + (1) Data science approaches often have less domain knowledge built-in, and tend to have difficulty learning "concepts" that generalize. How can we build in more existing domain knowledge? (#15)
  + Even in explainable AI, people are explaining models not explaining phenomena (#16)
  + (1) How do we make non-data scientists be aware of and able to effectively use the exceptional complex layers of primary and inferential data and models data science will produce with confidence and understanding. How can we support the 'einsteins' in forming hypotheses? Tools for focusing attention on critical points of interest without forming 'echo-chambers' will be key. (#17)
  + How can we achieve absolute transparency into research and maintain competitive advantage? (#18)
  + Can we eliminate 'data science' as a distinct activity from 'science'? I do not believe that distinction is helpful anywhere outside of career advancement. (#21)
  + Facilities need to evolve to effectively and simultaneously support data science workloads side-by-side with modeling & simulation workloads. How can that be achieved? (#32)
  + (2) The convergence of traditional science method plus deeply integrated data science fundamentally changes the scientific process. This is the next revolution in science. (#36)
  + Can we establish the relationship between the physical abstractions at different level and data sets and characteristics that can represent them (#47)
    - Closer connection of knowledge to actionable decisions - we often have the right models, just not for the questions we really need to answer. (#14)
  + Data Science Advancements (#48)
    - For the labs which do have some applied focus in addition to fundamental science, is prediction without understanding acceptable. Is this a danger? (#28)
    - (3) Data science lacks the foundations and rigor common to areas such as physics. There are multiple aspects, ranging from limits of data, and quantification of performance of answers/solutions derived using data science. (#1)
  + Application of Data Science and this Convergence (#49)
    - Biogeoengineering for agriculture, wetlands, dessert and ocean to mitigates issues of climate change, pollution and land use: requires integration of IOT sensors of MANY different types; population, industry and activity demographics; and integration of data science with planetary physics models to determine critical lever points, tipping points and design interventions. (#3)
  + (1) Data Science as a disciple as well as a foundational skill (#50)
* Who would develop it (basic research to advanced deployment)?
  + In addition to the platforms and infrastructure, we still need to advance the methods.. E.g. explainable AI. (#6)
  + (3) Foundations and basic laws of data science need to be developed so that look more like physical sciences so that we know what can be done with data. This area can be developed by combination of people from physical science, math, CS and others. Data science can help science discovery and facility provision. Also, domain science can help identify, fine-tune, optimize data science solutions. (#13)
  + We need to develop efficient methods to add deep heterogeneous Bayesian priors (past data, constitutive laws, etc.) in the deep learning models, from physical constraints to symbolic reasoning. Training networks on the data set obtained from physical simulations, if you think about it, is the worst of both worlds, since we use extensive correlation to substitute for physical knowledge. Also, when applied to experimental data, confounders and observation biases make comparisons difficult. (#20)
  + Mission-drive programs of scale to solve problems of great importance (e.g. greenhouse gas reduction from land and ocean) might focus staged deliverables but with aspirational long term goals. (#23)
  + (1) When figuring out how to integrate domain knowledge into data science, domain experts (such as chemists) are necessary (#24)
  + (1) Motivating our workforce needs anchor challenge problems - advanced deployment as a pathway to identify necessary basic research. Challenge is where seed corn for ideas come from. (#27)
  + Symbolic reasoning, analysis of programs and automated theorem proving are under-deployed in scientific enterprise, but are relevant to algorithms that understand scientific theories, models, and data. (#31)
* Who would use it and what skills would they need to use it effectively?
  + (3) Eventually all domains of science would use this. It could create a new class of researchers and personnel. We would have an increasing need for people that understand both the domain and emerging data science methods. (#12)
  + Foundations and laws of data can used by multiple parties: scientists to figure out when and how to use data, e.g. show ML solvability before using ML, incorporate science domain information in solving ML problems. In addition, facility providers monitor, diagnose and optimize using measurements - data science will guide it and can also what is beyond its capability. (#22)
  + The foundational technology is general purpose; but ontological and semantic layers; domain specific models and data will be used by specific communities with different working styles and needs. These will need to be customized. (#25)
  + At least in industry, data science is often described along the lines of intersection of software engineering, math/stats/computer science, and domain knowledge. (#26)
  + We desperately need effort at the intersection between physics and data science. It is somewhat ironic that ML is now applied in biology, medicine, economy, sociology (scientific areas without clear ground truth) much more extensively then hard core physics. (#30)
  + We need different technologies for different levels of users: Some will just want to use 'search' for relevant information/predictions effectively and be able to understand what to trust; others will want to add functionality and need a deep understanding of both theory and programming/compute paradigms (#33)
  + How do labs impact the world around us? (#35)
  + (1) Decision support, but not the way ASCR has so far defined decision support. Needs to include emphasis on context, UQ, but also science to provide answers to questions the country needs to deploy its resources to answer. (#38)
  + Science facilities are also very important part of science discovery - it uses significant funds. They can benefit from data science - they use measurements to monitor, provision, diagnose and optimize. (#51)
* When would it be expected to be in production use (N years in the future)?
  + Next 10 years... AI/ML methods become ubiquitous in the scientific process (at various levels). Maybe by 15-20 years data science is helping to "steer" hypothesis generation. (#19)
  + Physical laws and foundations of data will be used by scientists for discovery, and facility providers to monitor and diagnose systems. (#29)
  + I would argue that discoveries are already happening in the intersection of data science & physical science, so in some sense, it's in the beginning stages of "in production." However, there is much work to be done. (#37)
  + (1) This is something that builds in sophisticating over time. We are already seeing a vast convergence in the technology needed to access data and even label it if there is an interested community; but over the next 10 years increasing the FAIR nature of this data and getting far more sophisticated in crowd and automated labeling will transform the types of algorithms used. The modeling and knowledge layers will take longer to develop for each area of application. (#40)
  + (2) Reducing cycle time between defining a challenge and an actionable results is a continuous process - from improving productivity in next 5 years to real time connection between theory and experiment progress in 30 years (#41)
* Where, and how widely, would it be deployed?
  + Data science + physical science will be used by scientists and facilties (#34)
  + (1) I think that data science will be integrated as one approach to science, complementing theory, experiments, & computation. (#42)
  + (2) It will be ubiquitous. The danger will be in comparing the quality of sources, understanding uncertainty, and not falling into 'data-drive' feedback loops that limit attention to a few topics (e.g. how news echo chambers are operating today) (#43)
  + Breadth of deployment will start in labs with appropriate Capex. As cycle time in labs decreases diffusion of approach will increase impact. (#45)
* What is the setup time and/or process for using it?
  + (1) Foundations of data science will take much time to develop fully. They will be used as they are being developed but expect to span 30 years and beyond to be fully developed, (#39)
  + (2) Scientists focus on a problem of interest and the infrastructure provides an integrated view of what is known in this space and can potentially spot outliers or patterns that haven't been previously researched. This guides the scientists to formulating hypothesis that couldn't have easily reached without this. (#44)
  + (3) These types of things shift effort from setup time to execution. i.e., whereas today we do design of experiments, in future we will do 1 experiment as a seed and then have computation drive the n+1th experiment. (#46)

1.2 Breakout 2 - Implications of this Problem.

**The following participants have not been active:**  
Bruce, Thomas Britton (JLAB), Irene Qualters (LANL), Ilkay Altintas, Adam Arkin, Maria Chan, joshua elliott, Lee, Stuart Gluck, Ray Grout, Mike Hildreth, \*fac- Nami, Sergei V. Kalinin, Lee Ann Kiser, Kari Heffner, Torre Wenaus, Stefan Wild, Huolin Xin

**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?
  + Implications/consequences topic (#1)
    - (2) Intended Consequences: By developing foundational data science, the solutions to science and facilities can be made rigorous to characterize performance and confidence, provide explanations, recognize patterns and trends. There is significant potential for new discoveries, insights and new physical laws; and in facilities area new ways to monitor, diagnose, automate and optimize. (#2)
    - Intended consequence: Can accelerate discovery by replacing expensive parts of computations with data-driven model. (#46)
    - Intended consequence: can discover/suggest physical laws, conservation laws, and then test them (#48)
    - Coupling these methods to automated experimentation both for efficiency but also discovery. (#47)
  + Change in how scientists are trained (#3)
  + If foundations are required, then theoretical advances are needed, just as with physics. (#7)
  + Looking back at session 1, we need more formalized "rules" and understanding about how to apply these methods correctly. (#8)
  + Scientists will have to know more computer science in order to operate these systems. Bench work + Math no longer sufficient (#10)
  + (1) Adverse Impacts (#35)
    - (2) Adverse Impacts: How do we avoid this new approaches from introducing unintended biases or "looking under the lamp post" (#6)
      * Human biases in how we've done science so far can affect the data & then bias the machine learning. On the other hand, could use Bayesian methods to explicitly incorporate a trade-off between exploration & exploitation: https://www.nature.com/articles/s41586-019-1540-5 (#45)
    - Adverse impacts: over-trusting a non-robust model (#9)
    - Unintended Consequences: There significant risk in "black box" application of data science methods. If applied to non MIL-solvable problem, the solution will be incomplete or unsound. For example, it can lead to non-true physical laws begin discovered. One example, inferring science laws from data alone: :chaotic behavior of TCP based on ns-2 simulations - it turned out to be non-chaotic. (#13)
      * Adverse Impact: Greater emphasis on prediction (which is useful) versus understanding (which is science). (#29)
        + I'd say that understanding that leads to reliable predictions is the cornerstone of science. If the understanding doesn't let you make reliable predictions about novel scenarios, it isn't an accurate/useful explanation of the phenomenon. (#30)
  + Effects on human cognition: nobody remembers phone numbers anymore, what about when learning physical laws from data is as ubiquitous as smartphones? (#16)
  + Need to develop better uncertainty quantification or ways to test how trust-worthy data-driven models are (#15)
  + deleted (#17)
  + (1) Need to track the limitations of a data-driven model (such as: this was trained on data where there is no friction) (#19)
  + Emergent properties by definition cannot be determined by understanding the parts of a system, how can data driven science learn about emergence? (#21)
  + Computers and automation require a level of precision and rigor not currently employed by scientists. Machines do exactly what you tell them to do, not what you intend for them to do. (#27)
  + (2) Infrastructure/Platforms/federations/etc (#36)
    - (1) Orchestration services across instruments and facilities to bring data closer to large-scale compute systems or farm out compute to where data is stored/produced. (#11)
    - (1) There needs to be a more integrated ecosystem of platforms and services to support this new model. (#12)
    - Change in computing infrastructure, such as interest in GPUs (#5)
  + Data (#50)
    - (3) We need FAIR compliant data and data curation to enable many of these advances. (#44)
    - Scientific knowledge must be curated, maintained and accessed in machine understandable ways. The scientific publishing model will have to adapt to the primacy of machine usable scientific outputs (#4)
* Who is/will develop this companion technology/service?
  + (2) Continued improvement in data science platforms across the board (architecture, networks, storage, software). Where will industry create gaps that science domains must fill. (#14)
  + Statisticians have experience with quantifying uncertainty (#22)
  + Need to expand curriculum for scientists at the university level (#24)
  + (1) This requires applied math, computer science, statistics and their sub-domains to develop foundations. For tools, we need CS and software engineers (#25)
  + Improved tools for how we organize information and knowledge and connect these together. (#26)
    - This is what... this would have to come from the research community. (#39)
* What skills/knowledge does the end user require?
  + (2) Increased emphasis on data science as a skill and discipline. Scientists need a deeper understanding of this as part of the scientific method to avoid mistakes. (#18)
    - Critical thinking about the ways that data & statistics can be misleading (#28)
    - Example was given of using a curve fit in Excel to Corona virus that led to wildly incorrect extrapolation. (#42)
  + (1) Ideally, the end user should be provided "ready-to-use" tools and should not be required to have much specialized skills. End users come from diverse science and facility areas - less required of them to use, the higher will be the adoption. (#31)
    - This is probably about ease of use for the technologies but maybe not the application of the data science (for example). (#40)
  + (3) End users need to have a way (tools, domain knowledge, something else?) to verify/validate the results produced from the model. (#37)
    - is this something we should put on users? Is anyone happy with the V&V techniques currently used in practice? (#38)
* What are the training/support requirements?
  + (2) New courses and teaching in data science and its proper applications. This would a continuous process since the area will likely continue to evolve. (#20)
    - Concepts like UQ for data science. This would include over-fitting, P-Value hacking but needs to go well beyond that. (#49)
  + (1) Data science as a stand alone discipline to help both advance the field but also to further understand its limitations and risks. (#23)
  + HPC training for people who run simulations shouldn't necessarily be the same as people who are doing, say, deep learning (#32)
  + The requirements should be minimal. (#33)
  + There are many courses for data science applied to business use cases but not as many focus on data science methods applied to traditional scientific applications/workflows. (#34)
    - How might the training be different? Is it just the use cases that are covered or is it more. (#41)
  + (5) Two categories of training/support requirements: (a) developers of foundations and tools - it requires CS, applied math, statistics and other areas, and (b) end users from science and facilities areas - they should be required to have a detailed knowledge - they should be provided with tools with checks and confidences (#43)

1.3 Day 1 Reflections

Participants: 3

**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.



* 1. The integration of data science with traditional the scientific method will fundamentally change science.
* 2. A new discipline around data science is needed to better develop the rules, practices, etc.
* 3. Scientist will need a new set of fundamental training in the area of data science so they know both how to use the tools but how to use them properly and understand the potential risks.
* 4. A potential vision example might be the combination of data science with the traditional method will lead to a new class of discoveries that would have been impossible.   
    
  Another example... in 2050 a computer will have generated a testable hypothesis with little, to know human guidance that is tested and confirmed.
* 5. Expansion on Item 2. Foundational principles and laws of data and analytics with rigor and depth of physics, etc. We need to understand, design, optimize and test data solutions co-developed with domain knowledge for both (a) scientific discovery, and (ii) facility and federation services. We need analytical models and principles of data to (i) establish solvability, and (ii) provide optimized, trustworthy, provable, explainable solutions
* 6. Need to capture structure of scientific knowledge
* 7. Transform biology, where we don't already have robust rules/theory
* 8. Need to be able to verify & test AI discoveries
* 9. Publication model may change with open access. Should results be reported in a different way if they are data-driven and we don't have an understanding (and lacking causality)?
* 10. In 30 years, should be able to understand what the limits are of data-driven methods, and when we can trust results.
* 11. We need ways to organize & retain massive amounts of data so that it's feasible to reproduce/verify results.
* 12. In 30 years, need more transparency, reproducibility, etc. so public can trust science.

2 Day Two - November 5, 2020

2.1 Breakout 3 - Signposts

**The following participants have not been active:**  
Thomas Britton (JLAB), Ilkay Altintas, James Ang, Maria Chan, joshua elliott, Kate Evans, James Fairbanks, Stuart Gluck, Mike Hildreth, Sergei V. Kalinin, Lee Ann Kiser, Kari Heffner, Josh Levine, Ramana Madupu, Manish Parashar, Torre Wenaus, Stefan Wild, Huolin Xin

**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?



**Sticky points:**

 Top Takeaways (5 points per participant)

* 1. Recap
  + Comments
  + In some fields, where there are strong underpinnings we use those to validate the models. In the future, AI is allowing us to develop theories in areas that lack them today. (#2)
  + Given the early state of DS today, we lack the confidence in these methods to trust these today. (#3)
  + Science approach could be flipped where DS methods make predictions prior to formal hypothesis generation. (#4)
  + Dangers: as we rely on these we are at the mercies of the biases and limitations. (#5)
    - If we do this right then in the 30 years this will be largely addressed. (#9)
    - This could erode confidence in science. So we need to integrate these approaches while preserving fundamental aspects of the scientific method (e.g. reproduciblity, testableility ) (#10)
  + So the goal is 30 years from now we understand these dangers and we understand the limits. (#6)
  + We need Facilities geared to this... (#7)
  + We haven't discussed reproducibly.... How do we retain the right data in the right way to ensure this. (#8)
  + Note: I added more notes on this recap under the "Day 1 Reflections" tab (#62)
* (1) 11. Foundational unified theory of data science that encompasses scientific discovery (physics, chemistry, biology) and scientific facilities? - this includes analytical models for data science that encompasses disciplines including physics, chemistry: (i) Ways to best utilize data science to enable scientific discovery and optimally provide federations of facilities, and (ii) identify the limits of what is possible - speed-of-light limit of data based inferences and designs, Godel, Turing and Vapnik limits. Example1:, automatic discovery of physics laws based only on data has ML limits show that some laws may not be discovered this way - this limit does not apply to discovery by humans
  + Comments
  + These laws can be used to develop optimal, explainable and trustworthy solutions by co-developing them with domain laws such as ones from physics, etc. An example is power-level estimation of reactor systems using cooling tower measurements - ML solution is customized with physics laws. (#19)
  + Could we see computers pick up multiple strands of information and come up with strong hypotheses and proposed experiments for detailed scientific questions- e.g. could this new data science have noted the appearance of repetitive DNA in microbial genomes, clustering with proteins of various types, homology of he small repeat regions with prophage and phage sequence, and inferred the existence of CRISPR like phage defense systems faster than human scientists? What would it take in structuring machine learning to do that? (#22)
    - Computers by themselves are subject to Turing and Godel limits. But Computer+human can achieve that. (#31)
* 12. How will we recognize this is happening? Are there in key events or signals that we might see?
  + Comments
  + Multiple science domains will start adoping similar procedure/philosophies (#16)
  + Step change in breadth of engagement around papers/artifacts (#17)
  + More direct and obvious linkages between end use and early stage research (#18)
  + Need help from others on how to word this but... some new novel ways where computer generated hypothesis is coupled directly to the experiment to rapidly develop and test hypothesis. This should go beyond simply exploring a search space though. I could see some form of this happening in 5-10 years. Maybe it already is. So what is the evolution of that. (#20)
  + Experiments will be designed and operated in an automated way based on proven computational and data analytical programs. (#25)
  + Widespread engagement - more "moonshots" (#30)
    - Agreed! When we begin to see programs DESIGNED to create the data science-driven large scale science missions: Biogeoengineering of earth for climate, food and environmental health (while ensuring equity and inclusion in the creation and use of resources); creation of automated (bio)manufacturing for deep space missions to explore and utilize resources; (#40)
  + When exploratory data analysis and integration is a far more minimal time investment than it is now and our interactions with data when we are designing and interpreting experiments are less 'search' and visualization and more asking for predictions. (#32)
  + A major breakthrough (e.g. fusion, a new fundamental theory, curing cancer) occurs as a result of this. Timeframe: 10-20 years. (#37)
  + Negative signpost: The first major retraction of a result due to the misapplication or unknown bias in the data that led to it. (#39)
  + Previously intractable understanding of complex systems is embodied in new theories for complex systems (#41)
  + Acceleration of scientific discovery and the ease with which resources can be federated is an indication of progress. For, example, federation of accelerator, supercomputer and cloud services must be provided in few minutes to a scientist. (#42)
* 13. How do we present and publish these results and how is that different from today? What are the sign post around this? How do we convey these to the public to maintain trust?
  + Comments
  + How does publication change when semantically labeled data streams allow auto summarizers to 'create' custom papers from results? For example--summarizing a years worth of papers on micro metabolism to summarize the most impactful results in human readable 'text' for scientists, journalists and program managers? How then do we deal with biases in the corpus, ontologies, and keywords and prevent 'YouTube' algorithm problems? (#15)
  + Convergence of societal and academic discussion (#21)
  + FAIR principles really need to be amped up and there will need to be an increased emphasis (including funding changes) to make sure this happens. This is happening today but will have to be greatly increased in the next 5 years. (#23)
  + Shift in the news/discussion cycle to see the time and attention cycle increase (#24)
  + Public summaries and access to the methodologies, data, results, limitations are expected and required. public and multiple types of reviews (not just 2-3 peers but scientists and public from different areas?) that throw out the current peer review process (#26)
  + arxiv and the like have been created to solve a problem, but its not perfect. Can we build a better ark(iv) ? (#28)
  + Standardized ways to report limits of a model so that people don't accidentally extrapolate (#38)
    - Very important! (#43)
  + Standard, common ways to report history of dataset (like this Datasheets for Datasets idea: https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf) (#44)
  + Experimental results from the laboratory get logged into machine-readable databases (#45)
* (1) 14. How does the methodology and training evolve over this period? What needs to happen?
  + Comments
  + Science literacy widespread. (#27)
  + Science focus on matters of widespread importance. (#29)
  + Development of curriculum that compliments existing courses in areas like statistical analysis. Since the methods change and evolve, more theoretical training about how to rationalize new methods. This needs to happen immediately. Signpost: new courses part of most physical science courses (#33)
  + People might need specialized training in this multi-disciplinary area (#34)
  + is the PhD advisor model outdated? should we have multiple, targeted advisors to cover different areas? (#35)
  + Agreement on what the scientific method looks like in data-driven contexts (#36)
* (1) 46. Interesting Statements:
  + Comments
  + We will know when a human in the loop is important. And we will know when they aren't. (#47)
  + What happens when we can simulate a human brain? (#48)
  + Do we see a fundamental paradigm shift in the scientific model? (#49)
  + Current DOE proposal process makes it hard push the envelope with disruptive ideas. (#50)
* 51. DOE/lab policies
  + Comments
  + DOE can develop standards about reporting constraints on models (#52)
  + Standard code and data repositories with standards will be helpful. (#53)
  + Removing barriers for radical ideas (#54)
  + Support for radically new proposals, ideas (#55)
    - Could we reserve larger fraction of funding for these more radical approaches. (#57)
    - ARPA-E is nice, but restricted to certain areas (#56)
    - Potentially separate issues: funding for time for brainstorming/thinking vs. funding for bigger projects that require more resources (#59)
    - Who peer reviews this? Someone not competing for the same money? No one? (#60)
    - Example (potentially mythical) with no peer review: Google's 20% time (#61)
  + Labs have lots of structure that take time away from being creative (#58)

2.2 Breakout 4 - Signpost Plausibility

Participants: 3

**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. Signpost
  + Comments
  + Foundational data science co-developed with domain science (#2)
    - precursors: domain-incorporated ML/AI ML theory, theory of computation, generalization theory (#13)
    - Need: combination of current ML/AI theories with domain science theories (#19)
    - SciML DOE workshop is example of applied math and ML; ML/AI townhall meetings (#33)
    - Need: Notebook technologies that integrate both software and analytical aspects of these co-developed theories (#48)
  + Standardized ways to report limits of a model so that people don't accidentally extrapolate (#3)
    - Very important! (#8)
    - Future calls include language on how to bound the predictions and report those limits. (#32)
  + Increased emphasis from DOE on FAIR and related concepts to ensure access to data both for discovery but also reproducibility and publishing. (#9)
  + Foundational data science for facilities and their federations (#10)
    - precursors: Measurement-driven data analytics for profiling, optimization, diagnosis (#15)
    - Need: Combination with foundational ML/AI theories (#21)
    - Need: Software defined frameworks to provision and optimize these systems and federations (#25)
    - Need: scientific methods to capture the effectiveness of federations computing and experimental facilities in terms of scientific productivity (#40)
      * There are point methods that charactrize facility utilization but they do not adequately capture quantities such as scientific productivity (#49)
  + A major breakthrough (e.g. fusion, a new fundamental theory, curing cancer) occurs as a result of this new convergence. Timeframe: 10-20 years. (#14)
    - Self-driving experiments over federations (#16)
  + DOE program that are open to radically new proposals (#17)
    - precursors: not many (#18)
    - Need: Establish specific DOE programs (#27)
  + Formation of cross-office programs (#28)
  + DOE: Do we need new forms of SciDAC and co-design centers to include theory and foundational aspect of this convergence and discipline. (#29)
  + Precursor: SCIDAC, co-design centers - but are mainly HPC-driven (#30)
  + DOE: A future major project (future experimental facility) has participation from multiple program offices. (#36)
  + DOE: Future ECP-scale project focused on the convergence of data science and experimental and theoretical science. Time-frame 5-10 years. (#44)
  + Competitions: A series of CASP-like competitions to advance the data science field and focus on some domain. (#46)
* 11. A major breakthrough (e.g. fusion, a new fundamental theory, curing cancer) occurs as a result of this. Timeframe: 10-20 years.
* 12. A major breakthrough (e.g. fusion, a new fundamental theory, curing cancer) occurs as a result of this. Timeframe: 10-20 years.
* 20. Discussion
  + Comments
  + There current harbingers... NMDC, Jacobson work. What are the next set of results? (#22)
  + What are the Grand Challenge problems but maybe with something that is knowable? (#23)
  + Other pratical demonstrations? Hilbert's conjectures/problems. (#24)
  + DOE Facilities: How we link these. Do we need new metrics? (#26)
    - Federation of science instruments (#31)
    - Computational facilities need to be designed and optimized for these new DS use cases (#34)
    - If you need to integrate the capabilities more, when do we see a joint facility proposal that is not all in a single office but perhaps multiple offices (#35)
    - When you hook up facilities together how do you quantify that it is enabling science? (#38)
  + Superfacility concept... what are some signpost for that. (#37)
  + The main goal is scientific impact/advancement. Solving big problems. (#39)
    - Making health care more affordable, energy security, national security (#41)
  + We have the Exascale Computing Project. There is a current effort going on from the AI Townhalls. What is the evolution of that? If there is going to be an effort that needs to be pitched, what are the components of that project? (#42)
  + Run competitions for data science (#43)
    - Predicting which of the many vaccines for flu work and getting better and better at it (#45)
    - Which energy technology has had the biggest impact in a given year? (#47)