Group #18: Community of Interest on the Future of Scientific Methodologies

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

|  |  |
| --- | --- |
| Participants | |
| Huolin Xin | Mike Hildreth |
| Kate Shattuck | Terry |
| James Ang | Terry Turton |
| Ilkay Altintas | Lee |
| bruce | \*fac- Nami |
| John Shalf | Terry Turton |
| Nagi Rao - ORNL | John Wu |
| shantenu jha | Manish Parashar |

Contents

[1 Day One - November 2, 2020 3](#_Toc55994298)

[1.1 Discussion: Breakout 1 - Define the Scope of the Problem. 3](#_Toc55994299)

[1.2 Discussion: Breakout 2 - Implications of this Problem. 5](#_Toc55994300)

[1.3 Brainstorm: Day 1 Reflections 7](#_Toc55994301)

[2 Day Two - November 5, 2020 7](#_Toc55994302)

[2.1 Brainstorm: Breakout 3 - Signposts 7](#_Toc55994303)

[2.2 Brainstorm: Breakout 4 - Signpost Plausibility 11](#_Toc55994304)

[3 Day Three - November 10, 2020 14](#_Toc55994305)

[3.1 Brainstorm: Breakout 5 - Pitfalls and Roadblocks 14](#_Toc55994306)

[3.2 Brainstorm: Breakout 6 - Keys to Success 14](#_Toc55994307)

[Appendix 14](#_Toc55994308)

[Live chat 14](#_Toc55994309)

1 Day One - November 2, 2020

1.1 Breakout 1 - Define the Scope of the Problem.

Participants: 0

**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?
  + (1) how to maximize the scientific productivity of nation user facilities to automated data processing (#1)
  + (3) Redefining the Scientific Method through the extension of Physical Science with Data Science (#4)
    - This where Data Science includes UQ, AI/ML, Streaming Analytics, Optimization, Graph Analytics, etc. (#6)
    - (1) Challenges with quality of scientific data exist for both computational Modeling and Simulation and Experimental Measurements. What Data Science capabilities, e,g., Uncertainty Quantification help here? (#12)
  + What is the tipping point that requires a revolution as opposed to a natural evolution of technology and trends? (#14)
  + (1) The move of computing between steady allocation to demand-driven allocation (#27)
  + (4) Meta-Topic: Automation (#30)
    - (2) A tool that knows all the past literature and makes expert recommendations (#28)
    - how to turn the data analysis output into actions in physical experiments (#5)
      * should add also feedback to simulation: steering of simulations (#25)
    - (1) Greater need for automation. Automation is the end product but which must be designed for, and which arises when the right underlying conditions are present. (#8)
      * Real-time adaptation of data collection based on accumulated data tested against a hypothesis (#7)
        + What would the computational ecosystem look like that enables refining simulation with collected data, and experimental control in real time? (#13)
      * (2) There are technological, cultural and scientific gaps to link experimentation and observation with simulation. (#11)
        + How to grapple with overwhelming volume and velocity of data (#9)
        + Yes: Example with the famous "Pauli effect" where experiments would fail if a good theorist were to enter the room. (see wiki page). The challenges are compounded by the cultural and methodology differences of the data science community. How does one make these different paradigms and methods convergent and work together. (#16)
  + Meta-Topic: Convergence (#31)
  + (3) Meta-Topic: Validation and Reproducibility (#35)
    - How to validate methods on a variable data stream? (#10)
      * Time-varying, heterogeneous data types (#26)
      * (2) Community driven validation sets for each significant problem. A tool that provides Kaggle competition type of validation and Kaggle artifacts finder. (#36)
      * (1) Reproducible pipelines, software stack, etc. (#32)
* Who would develop it (basic research to advanced deployment)?
  + Most likely the DOE community will have to develop the technologies to integrate physical science and data science (#17)
    - Don't forget the Universities! (#18)
  + (1) The development of new computing capabilities to support convergence of physical and data science will require inputs from DOE, and the greater science enterprise to "co-design" these computing capabilities. The commercial drivers for data science are focused on commercial workloads. (#21)
    - These co-design collaborations can span National Labs, Industry and Academia. (#22)
* Who would use it and what skills would they need to use it effectively?
  + The recommender systems to guide the experiments and simulation would be tremendously useful for science (#2)
    - sorry this seems to be on the wrong card (#19)
  + Initially this type of technology would be targeted for DOE science facilities. Hopefully, we will have a good technology transfer mechanism to bring it to wider communities (#3)
  + Expertise in the breadth of infrastructure and software ecosystem options, in a way that utilizes heterogeneity (#15)
  + (3) Echoing what Ilkay said in the first card, there are cultural differences between theorists, experimentalists, and data-scientists and there are significant differences in their workflows. Cross-training between the fields or perhaps pro-actively forming integrated teams for each new scientific target may be essential for the future of science. (#20)
    - Computational Scientists, Experimental Scientists, Data Scientists, Mathematicians, Computer Scientists, Diagnostics Developers, need multi-disciplinary collaboration skills. (#42)
    - Experimental teams embedded with data scientists will be critical (#41)
* When would it be expected to be in production use (N years in the future)?
  + It should be done now, but it might take years and concerted action to bring this in production. Right now this convergence seems very ad-hoc. (#23)
  + Initial implementations of Converged Physical and Data Science capabilities to begin the re-definition of the scientific method can begin immediately. The evolution of the method can take decades. (#40)
* Where, and how widely, would it be deployed?
  + For any new "big science" problem (big mission requirement defined by DOE), we should design this convergence in from the get-go. (#24)
    - This depends on the definition of convergence for DoE. It should be defined and the community should be trained on the convergence protocols and expectations before any deployment. (#29)
  + (3) Convergence will require heterogeneous computing which includes computing at all scales and in many places - we will not be focused on just the DOE LCF user facilities. (#33)
    - (1) There is likely a need for a continuum of computing resources, with some more widely distributed than others (#37)
    - This Convergence will require deployment of computing capabilities, both modeling & simulation and data science capabilities to be distributed throughout the DOE/scientific community enterprise. (#38)
* What is the setup time and/or process for using it?
  + some DOE experiments are already using real-time analysis capability at NERSC, presumably there are other examples as well (#39)
  + The integration of data science capabilities into experimental facilities, e.g. streaming analytics, inference engines, UQ, will require pre-work, e.g. training, to setup the in situ diagnostics. The "smart" data acquisition needs to occur in real time. (#43)
    - Training could occur at LCFs, but could also be distributed to cloud computing, edge servers, etc. (#44)
  + (1) Integration of data science capabilities into physical modeling and simulation is underway now with projects in domain-aware machine learning, etc. (#45)

1.2 Breakout 2 - Implications of this Problem.

Participants: 0

**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?
  + real-time machine learning and recommendation technology (#1)
    - Includes also natural language processing to create recommendations. (#6)
      * Need to be able to identify if there exists a theory that would explain the experimental data or if something \*novel\* and unobserved has happened. (#28)
    - Recommendations include what is the next experiment, but a complementary capability is to recognize if there are theories or hypotheses developed by others in the literature that match or contradict the data science or experimental result. (#7)
      * Need to trigger actuators to conduct the next experiment (no human in loop) (#23)
  + data integration across computing continuum (overlaps with data management group) (#2)
    - Universal data access (#10)
      * solving the universal data access is essential to connecting experimental science with data science and theory (#11)
  + Meta-topic: Advancing Scientific Method (#4)
    - This is integrating the theory, experiment, and data-science together for the discovery process (not just the advancement of each of these individually) (#41)
    - There are many implications of what is needed to succeed (as described in the other meta-bullets above and below) (#42)
  + Meta-topic: Continuum of compute resources (#5)
    - need for computing near the high-speed data sources (#8)
      * edge computing: for the real-time requirements, best to put the compute at the data source. (#9)
  + Need labeled data but for other data collected by experiments or theory need to have formal data description. (#12)
    - wide adoption of FAIR Principles (#13)
    - essential for machine learning (likely also connected to data management) (#14)
    - Making the description formal is essential for making it machine readable (connects with the FAIR principles as well) (#15)
    - formal description of data and label, to make the data machine readable (#16)
    - flexible analysis tools based on these formal data descriptors (#17)
    - unsupervised ways to generate synthetic training data (#53)
  + Meta-topic: Validation/Reproducibility (#21)
    - We need metrics and benchmarks to define and measure scientific data quality: where simulation-generated or experimentally measured. (#22)
    - Need to capture the experimental workflow in a manner that isn't overly heavy for the scientist. (#24)
    - data quality across all steps of the workflow (#25)
      * Needs to be fully automated in the future. (#26)
      * needs to record calibrations in all the steps (#27)
* Who is/will develop this companion technology/service?
  + Joint efforts spanning data science, experimentalists and simulation scientists (#29)
    - with possible collaboration with international scientific communities as well as industry (#33)
  + National Labs are primarily responsible for the cross-integration of these technologies. (#30)
    - But the component technologies may well come from industry and from universities. (#32)
    - some "off-the-shelf" technologies could be applied (#31)
    - "Computing Continuum" probably has contributions from a wide variety of sources, including commercial (#34)
  + The National Labs are adopting and developing scientific AI/ML methods using the frameworks that originated from the Commercial Sector (#37)
  + The DOE has invested heavily in Applied Math algorithms for Uncertainty Quantification, Optimization, Graph Analytics, Streaming Analytics. There have also been significant investments in Validation and Verification methods, Quantifying Margins of Uncertainty, which bind computational and experimental science (#39)
* What skills/knowledge does the end user require?
  + make sure the tools developed are usable (#35)
    - Usability of data-regularization tools is a huge impediment to enable FAIR sharing of data. (#36)
    - data regularization and formal description is necessary for sharing data \*across\* experimental, data-science, and theory (not just for ML/DL). (#38)
    - reduce the efforts required for hyperparameter tuning (#48)
  + The end user/scientist brings their domain expertise to the table. They may have a specialized role in capability development as well, but they also need to be able to define requirements for experimental, computational, and data science capability developers. (#49)
    - Bring domain knowledge into machine learning and other prediction and recommendation processes (#52)
  + Workforce concerns: Will enough able bodies enter the pipeline? Will we have enough diversity in the workforce? (#51)
* What are the training/support requirements?
  + Meta-topic: Cultural Diffs and cross training (#18)
    - Cultural differences between data-science, Experimental, and theory require cross-training across disciplines to overcome. (#19)
      * Google Translate for scientific conversation (#46)
    - Cross-training versus embedded teams -- teams with specific experts embedded, e.g., that have deeper domain knowledge but are trained data scientists? (#20)
  + Multi-disciplinary collaboration communication abstraction tools: e.g., Proxy Applications, Abstract Machine Models - from the HW/SW Co-design world. (#43)
    - Not each individual is multidisciplinary (jack of all trades), but the team needs to be multidisciplinary and ability to communicate with each other. (#44)
    - need a common language for communication within the team (#45)
      * Google translate for theorists and data scientists... (#47)
    - More advanced algorithms that require less hyperparameter tuning; so, the training requirement is reduced. (#50)

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.



* 1. We have a lot of actionable bullets, but in working through the details we might have lost some of the science-fiction implications. Need a way to tell the SciFi story better in this interface.
* 2. "Frictionless" use of technology will be important
* 3. Good interactive discussion with the participants. We identified some key challenges, but perhaps were not looking beyond 10 years out.
* 4. A common language is critical as is multi-disciplinary teams
* 5. I will be interested to read what the other "topic 8" working groups identified.
* 6. Common formal description across "information sources" (labeled data for data-science, experimental data for experimental science, and simulations from theory) and the automation for applying those formal descriptions is crucial to create cross-integration of physical/experimental/data.
* 7. On the SciFi front, a more human-like recommendation system will be really helpful.
* 8. Most far-reaching ideas sit on top of tedious infrastructure development

2 Day Two - November 5, 2020

2.1 Breakout 3 - Signposts

Participants: 0

**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.  30 year vision : Reimagine how we do science through an automated coupling of data science, experiment, simulation and theory. The research enterprise becomes a super-organism.
  + Comments
  + Flip the story on ML/DL. Instead of having labeled data to train an ML/DL to recognize patterns in the data. Rather, we have an AI expert system read/process the literature and identify theories that match the data. (#6)
  + Could redefine the way we do science. (#7)
  + "Theory machine" - could read all the literature and identify theory that matches data (#8)
  + develop a "theory machine" (digital expert agent) as helpers to scientists to improve productivity by automating the integration of experiment, theory, and simulation (#13)
  + Develop a digital expert agent that can (list of possibilities:) read and understand the scientific literature relevant to a given problem, synthesize experimental data, provide guidance on experiments (do the experiments?), provide theoretical predictions, refine theories (#19)
  + digital expert agent. (#32)
* 2. Compute increase -- @HPC centers 20-300x increase but 1M-1Gx increase if include non-HPC sources such as personal machines
  + Comments
  + Computing resources would be available everywhere to make digital expert agent accessible anywhere (#14)
  + Local (distributed) training of agent (#24)
  + I would clarify that the "time to complete" a scientific workflow will be reduced by 1M by 2040, or 1B by 2050 versus the time to complete the same scientific workflow in 2020. Through the use of more ubiquitous, distributed computing. Today these workloads might take weeks or months to complete because in 2020, sensor data is all pulled back to DOE LCFs or NSF HPC centers. (#31)
* 3. Data movement will change in 30 years. IOT, edge computing, etc will be much more distributed
  + Comments
  + Middleware that can link software-defined networks to data and computing in scientific applications are needed. (#25)
  + How to secure the distributed computing elements (#27)
  + Smart measurement systems that can integrate performance data across the network coupled with machine learning software that can turn it into information to steer/optimize scientific workflows. (#29)
  + Data movement constraints means that many scientific advances for distributed computing will drive the improvements in fog, edge server and IoT computing performance. (#33)
* 4. Fancy, automated lab -- what will ALS (light source) be like? An automated system will process samples, self driving facilities. Will include initial analysis, ship data back to sciencist in cloud or other technology
  + Comments
  + same for electron microscopy, Hubble, mars rovers (#5)
  + require end-to-end integration of data acquisition systems, analysis capability, decision support and control mechanisms (#15)
  + This applies to facilities driving and/or triggering simulations, using the science learned through experimentation for parameterization of simulations. (#22)
* 9. Even bigger theme of automating the connect between data-science, experimental science, and theory. What tools to facilitate that integration (automation of connecting theory to experiment or to data science).
  + Comments
  + Like the idea of automatically integrate the content of technical publications into digital expert agent (#18)
  + I would add that HPC modeling and simulation can be applied toward experiment design, apply corrections to experimental measurements. (#50)
  + The intersection of HPC M&S with experimental measurements plus UQ and AI/ML from Data Science supports increase confidence in simulation results through Validation and Verification processes (#53)
* 10. What are we lacking? Underlying Data descriptor/language to be machine readable -- SIGNPOST. Describe theory to be machine readable -- SIGNPOST. Middleware to support networks -- SIGNPOST
  + Comments
  + We don't even have the ability to describe data such that it is universally machine readable. (#11)
  + How we describe a theory so that it is machine processable? (#12)
  + Item 16 is likely the necessary 1st step in developing a digital expert agent (#20)
  + Universal data description, universal data compression, universal data reduction and universal modeling that are modular and easily deployable--either through edge computing, network computing, distributed computing or decentralized computing. (#21)
  + FAIR data is a good start, but doesn't take it all the way to automatability. (#35)
  + Data should come with metrics on responsibility (explainability, interpretability, bias, fairness, privacy, uncertainty, etc.). (#45)
* 16. Universal data description, universal data compression, universal data reduction and universal modeling that are modular and easily deployable--either through edge computing, network computing, distributed computing or decentralized computing.
  + Comments
  + Universal data description, universal data compression, universal data reduction and universal modeling that are modular and easily deployable--either through edge computing, network computing, distributed computing or decentralized computing. (#17)
  + Data reduction can also happen if summarized as a model rather than statistically summarized. (#23)
* 26. Distributed scientific facilities are common with an array of computing capabilities that span HPC centers for modeling and simulation with integrated deep learning methods, out to edge and IoT devices that integrate inferencing engines and streaming analytics at experiments.
  + Comments
  + Opportunity to reduce data movement by summarizing raw data with a learned model (at the edge) rather than forwarding raw data to the HPC centers. (#28)
  + There are inferencing engines and streaming analytics methods that can be used to analyze/filter raw data collected by sensors in the distributed computing facilities. (#57)
* 30. DOE proposals need more than a data management plan. This expands to ensure data is machine-readable.
  + Comments
  + this item is related to #16 (#37)
  + Data management plans should come with measurable metrics defined by the proposer and reported with annual reports. (#48)
* 34. SIGNPOSTS: How long? Precursor technology? Implications for the Labs -- what do we need?
  + Comments
  + Policy Changes (#71)
    - DOE proposals need more than a data management plan. This expands to ensure data is machine-readable. (#72)
      * this item is related to #16 (#73)
      * Data management plans should come with measurable metrics defined by the proposer and reported with annual reports. (#84)
    - Free access to all published articles would be required. (#59)
      * Articles would need to be freely accessible and machine-readable (#60)
    - Scientific Research and Grant proposals should be reviewed my multi-disciplinary committees to break down barriers among experimental, computational and data science cultures (#86)
    - Educational changes to cross-train the scientists involved (#87)
  + common (machine readable) knowledge representation: work need to start right away (#36)
    - Universal data description, universal data compression, universal data reduction and universal modeling that are modular and easily deployable--either through edge computing, network computing, distributed computing or decentralized computing. (#70)
    - Data should come with metrics on responsibility (explainability, interpretability, bias, fairness, privacy, uncertainty, etc.). (#69)
    - Ability to describe data such that it is universally machine readable. (FAIR gets us close, but not all the way to automatability). (#38)
      * common machine readable knowledge representation as the next level up from the data representation (#61)
        + needs to capture metadata/provenance/workflow (#65)
    - Common Uncertainty Description (#68)
    - Means of representing what analysis software does to process raw data into something intelligible and interpretable (#62)
      * (may be out of scope) (#63)
        + Shouldn't this feed into UQ or accuracy? (#66)

Good point (#67)

* + - * Provenance capture is another signpost. (#64)
    - Need a way to describe theories in a machine understandable manner (for the theory machine). We are way far away from that, but NLP literature mining is a growing field. (#39)
  + Advanced analysis capability (#88)
    - need inferencing engines that can take into account of physics underlying the scientific endeavor (#43)
      * Need more advanced ML methods for training, more advanced AI methods for synthesis (#51)
      * Inferencing engines may be deployed via IoT into smart sensors, that are trained via Modeling and simulation of the underlying physics integrated with deep learning methods. (#89)
    - Need "comparator engine" that can interpret data and compare (in some meaningful way) to models. Not sure what that means, but we need one. Maybe this is some sort of logical framework? (#42)
      * Comment on prev comment: yes, so after signpost of having a computational representation of a theory and of data, the next signpost is to be able to meaningfully infer that the data matches (or does not match) the theory. (#44)
      * We need standards for 'comparator engine" performance - quality, accuracy? (#47)
        + True: Theory can only be proven within some bounds of statistical uncertainty. (#49)
  + We need a system software stack can can be portable, and interoperable with this new distributed scientific computing enterprise. (#54)
    - middleware/software/system to connect the distributed computing resources together (#40)
  + If we can achieve this, then the world-wide research apparatus could act as a single organism (universal undestaning of all literature rather than limited to what one individual can read/understand). (#46)
* 85. 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?

2.2 Breakout 4 - Signpost Plausibility

Participants: 0

**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. To Committee Members: We have gathered our SIGNPOSTS under the main heading below (6 Signposts) as we had already gathered many of those those in the previous breakout session. Each of the major headings has been replicated below to address plausibility, precursors and timelines.
* 6. SIGNPOSTS: How long? Precursor technology? Implications for the Labs -- what do we need?
  + Comments
  + Policy Changes (#31)
    - Free access to all published articles would be required. (#19)
      * Articles would need to be freely accessible and machine-readable (#20)
    - DOE proposals need more than a data management plan. This expands to ensure data is machine-readable. (#32)
      * this item is related to #16 (#33)
      * Data management plans should come with measurable metrics defined by the proposer and reported with annual reports. (#34)
    - Scientific Research and Grant proposals should be reviewed my multi-disciplinary committees to break down barriers among experimental, computational and data science cultures (#35)
    - Educational changes to cross-train the scientists involved (#36)
  + common (machine readable) knowledge representation: work need to start right away (#7)
    - Data should come with metrics on responsibility (explainability, interpretability, bias, fairness, privacy, uncertainty, etc.). (#29)
    - Universal data description, universal data compression, universal data reduction and universal modeling that are modular and easily deployable--either through edge computing, network computing, distributed computing or decentralized computing. (#30)
    - Ability to describe data such that it is universally machine readable. (FAIR gets us close, but not all the way to automatability). (#8)
      * common machine readable knowledge representation as the next level up from the data representation (#21)
        + needs to capture metadata/provenance/workflow (#25)
    - Need a way to describe theories in a machine understandable manner (for the theory machine). We are way far away from that, but NLP literature mining is a growing field. (#9)
    - Means of representing what analysis software does to process raw data into something intelligible and interpretable (#22)
      * (may be out of scope) (#23)
        + Shouldn't this feed into UQ or accuracy? (#26)

Good point (#27)

* + - * Provenance capture is another signpost. (#24)
    - Common Uncertainty Description (#28)
  + Advanced analysis capability (#37)
    - Need "comparator engine" that can interpret data and compare (in some meaningful way) to models. Not sure what that means, but we need one. Maybe this is some sort of logical framework? (#11)
      * Comment on prev comment: yes, so after signpost of having a computational representation of a theory and of data, the next signpost is to be able to meaningfully infer that the data matches (or does not match) the theory. (#13)
      * We need standards for 'comparator engine" performance - quality, accuracy? (#15)
        + True: Theory can only be proven within some bounds of statistical uncertainty. (#16)
    - need inferencing engines that can take into account of physics underlying the scientific endeavor (#12)
      * Need more advanced ML methods for training, more advanced AI methods for synthesis (#17)
      * Inferencing engines may be deployed via IoT into smart sensors, that are trained via Modeling and simulation of the underlying physics integrated with deep learning methods. (#38)
  + We need a system software stack can can be portable, and interoperable with this new distributed scientific computing enterprise. (#18)
    - middleware/software/system to connect the distributed computing resources together (#10)
  + If we can achieve this, then the world-wide research apparatus could act as a single organism (universal undestaning of all literature rather than limited to what one individual can read/understand). (#14)
* 39. Policy Changes - Plausibility, precursors, etc.
  + Comments
  + Changes in the composition of proposal review committees is plausible and could be implemented in a relatively few number of years (#49)
    - Highly plausible, 5 year timeframe, precursor ability in existence. (#72)
  + DOE's continued support for FAIR Data principles can easily extend to ensuring that all DOE science R&D result in publication that are machine readable. But the converse is not assured. (#54)
    - Going forward, mandate is feasible but funding is needed. Going backward in time --- may require new technology (#73)
      * Technology development could help here by making it automatic to capture provenance, metadata, etc. (#74)
    - Plausible, 10+years at least. Some fields are underway (#75)
* 40. Common knowledge representation - Plausibility, precursors, etc
  + Comments
  + This whole initiative is dead in the water without common descriptions of all aspects of the scientific enterprise (#56)
    - Need to engage metadata/ontology specialists (#53)
  + common data representation likely to come first (#43)
    - Many groups are working on various aspects of this problem (#44)
      * Research Data Alliance (#46)
      * Materials Genome Initiative/Project (#48)
      * NIST (Bob Hanisch) (#50)
      * Likely to have a bottoms-up emergence where sub-disciplines or even facilities develop common descriptions first (#52)
      * Builds on FAIR Data Principles that also include machine readable requirements. (#60)
    - Software and provenance also need to be included (#80)
    - Plausible - many current examples but fragmented. Timeline - 5 years for examples with narrow scope. More global realization is probably at least 10 years (#76)
  + automated metadata/provenance capture (#57)
    - Precursor technology & use cases underway. Timeline - 5-10 years. (#77)
  + cross community education effort to propagate the common data standard(s) (#58)
    - Plausible, underway, will evolve as needed, 5-10 years; ties to policy and educational changes (#78)
  + Some work has been done on model abstraction for computability (#59)
    - not sure how advanced this is (#61)
    - This encompasses how to represent theoretical knowledge (#81)
    - No one (in our group) is expert on this but there is current work on this so plausible. Timeline: 30 years. (#79)
* 41. Advanced Analysis Capability - Plausibilty, precursors, etc.
  + Comments
  + recommender system/agent for experimental settings (#45)
    - Plausible, this is a first step. Timeline: 5-10 years (#82)
  + Need AI systems that can interpret data and compare with models (#68)
    - Plausible, Timeline is 10-15 years (#83)
  + Need AI systems that can adjust model parameters automatically and recompute predictions (#67)
    - Precursor technology, already underway -- 5 years (#84)
  + AI System that can compute which measurements most improve determination of theoretical parameters (#91)
    - Plausible, 5 years (#92)
  + The ML/AI systems to do this don't exist yet (#63)
    - Need NLP (or other) systems that can ingest theoretical arguments and extract semantics of theoretical models (#65)
      * Requires machine-readable articles, machine-readable representation of theoretical models (#69)
        + If everything was machine-readable, the needs here to process actual language are not as great. (#71)
      * This is one of our out-there 30 year goals (#85)
    - explainable AI: the AI systems need to be able explain its recommendations/decisions (#47)
      * Again, 30 year goal (#86)
    - AI system that can select the best models based on data (#87)
      * 30 year goal (#88)
  + R&D program to develop Scientific Analysis that integrates HPC with Domain-aware ML and other forms of Data Science: UQ, Streaming Analytics, etc. (#64)
    - Plausible, precursor technologies already underway 5-10 years (#89)
  + R&D Program in Validation and Verification methodology that integrates Data Science into traditional experimental and computational science (#66)
    - Some early precursor frameworks -- 10-15 years (#90)
* 42. Software Stack - Plausibility, precursors, etc.
  + Comments
  + R&D program for heterogeneous computing at all scales: HPC Center to Fog, to Edge to IoT. Also to cover distributed computing from Data Center to field. (#70)
    - coordinating distributed workflow on disparate computing hardware from handhold devices to HPCs (#62)
    - software technology to enable composition of large distributed workflows (#55)
    - data movement software tools that can respond to just-in-time workflow requirements (#51)
    - Plausible. Precursor technologies underway, but this is long-term 15+ years (#93)

3 Day Three - November 10, 2020

3.1 Breakout 5 - Pitfalls and Roadblocks

Participants: 1

**Brainstorm question or instruction:**  
Breakout 5 - Pitfalls and Roadblocks  
Detailed discussions on identifying pitfalls and potential roadblocks. If possible, list in rank ordering. o What could prevent the technology/service/device from being developed (funding, materials, policies, researchers, operations staff, etc.)? o How will progress be measured/evaluated? o How will lack of progress be measured/evaluated? o Who will decide if progress is being made? o What are the consequences of not engaging in this area?



**Sticky points:**

 Top Takeaways (5 points per participant)

* 1. Lack of understanding foundations of AI/ML so that they are appropriately applied. These methods are subject to computability, learnability, interpretability and explainability limits. It is a pitfall if ML/AI methods are applied to problems that are beyond their limits - if not properly applied the solutions can be unsound, incomplete, or misleading.

3.2 Breakout 6 - Keys to Success

Participants: 0

**Brainstorm question or instruction:**  
Breakout 6 - Keys to Success  
Identify who needs to be engaged, research communities, domain science communities, staff, management. Identify needed skills and knowledge (give examples) o What benefits would society obtain? o What benefits would the science/research community obtain? o What research communities need to be involved? o What domain science communities need to be involved? o What staff and management communities need to be involved? o What kind of management structure is required? o How broadly will this impact society and/or the science community?



**Sticky points:**

 Top Takeaways (5 points per participant)

Appendix

Live chat

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

* Nov 2, 2020, 17:40 UTC
  + Executing the scientific method with the convergence of physical science (experimental & computational) and data science: How does hypothesis generation change? How do we insure reproducibility of results? How are results validated? (John Shalf | Nov 2, 2020, 17:40 UTC)
  + Convergence of Data Science and Physical science (John Shalf | Nov 2, 2020, 17:41 UTC)
  + Are we allowed to think about how this topic connects with others? (John Wu | Nov 2, 2020, 17:50 UTC)
  + Why not? (Terry Turton | Nov 2, 2020, 17:50 UTC)
* Nov 5, 2020, 17:45 UTC
  + I'm back, but don't have the dots (Ilkay Altintas | Nov 5, 2020, 17:45 UTC)