Group #7: Community of Interest on the Future of Scientific Methodologies

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

|  |  |
| --- | --- |
| Participants | |
| Wes Bethel | Kathryn Mohror |
| Katrina Waters | Vas Vasiliadis (UChicago) |
| Ezra Kissel | Bogdan Nicolae |
| Darrell Long (again) | Amber |
| Javier Duarte | bruce |
| Kathryn Mohror | \*fac- Nami Ishihara |
| Anthony Degleris | Jordan Thomas |
| Line Pouchard | Robert Hanisch |
| Bei Wang Phillips | Liz Sexton-Kennedy |
| justin Hnilo | Maria Chan |
| Suren Byna |

Contents

[1 Day One - November 2, 2020 3](#_Toc55484623)

[1.1 Discussion: Breakout 1 - Define the Scope of the Problem. 3](#_Toc55484624)

[1.2 Discussion: Breakout 2 - Implications of this Problem. 7](#_Toc55484625)

[1.3 Brainstorm: Day 1 Reflections 11](#_Toc55484626)

[2 Day Two - November 5, 2020 11](#_Toc55484627)

[2.1 Brainstorm: Breakout 3 - Signposts 11](#_Toc55484628)

[2.2 Brainstorm: Breakout 4 - Signpost Plausibility 12](#_Toc55484629)

[Appendix 16](#_Toc55484630)

[Live chat 16](#_Toc55484631)

1 Day One - November 2, 2020

1.1 Breakout 1 - Define the Scope of the Problem.

**The following participants have not been active:**  
Suren Byna, \*fac- Nami Ishihara

**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?
  + (10) Findable data: Problem: search mechanisms for finding web pages are inadequate for scientific data, where we want to search over features present/not present in data, over assumptions made during problem setup or research question being asked, over various input/initial conditions, over author of the data, code/instrument that produced it, etc. (#5)
    - (1) I am going to be critical of the Google idea. Google returns (ranks) information based on popularity. That's not what we want. It says nothing about quality, or if it the data that we want. I think that we need something much more intelligent -- I am going to again mention a new profession of data archeologist -- probably AI assisted -- were you can describe precisely what you need and you get back matches that you can then evaluate. The number of choices has to be managable. (#41)
    - (1) Findability and Accessibiity depend a lot on the discipline. Astronomy, genomics, a few others are pretty good. Interoperability and Reusability are the tall poles. (#4)
    - Many kinds of scientific data are difficult to find, reuse, and reanalyze because of lacking metadata, provenance, search functionality, self-description, etc. (#1)
    - Search for scientific data; Google doesn't work! (#43)
    - "Dynamic" Google for scientific data - eearch capability beyond metadata and labels: features, etc. - may need dynamic processing and reinterpretation of data (#31)
      * This point and #5 appear to be pointing at the same problem/idea (a good thing!) (#35)
    - (1) Need to speak the language of domain scientists, both for basic discovery and for discovery based on higher level concepts that are not typically embedded in metadata. (#15)
    - Problem: Accessing, interfacing with, and manipulating data is challenging because datasets often include raw information (tables, HDF5 files, etc) but exclude appropriate tools for interacting with the data. (#7)
      * Google dataset search is trying to solve this problem, using schema.org for data description. Not sure it is very successful though. (#9)
    - General abstractions to support transformation of data is a difficult problem. Requires domain-specific tooling or "intelligence" to act appropriately on self-described data and the formats/layouts requested. (#11)
      * Impactful research builds on prior work spanning long periods of time. Problem: How do we ensure that we're not "re-discovering" things? A means to discover \_all\_ relevant data (likely machine-assisted) is key. (#12)
        + (Riffing off Jay's idea): Want the ability for changes made in to scientific data be immediately propagated to other replicas in distributed federations (#13)
  + (11) Accessible data. How do we go about preserving access to digital data (codes, experiment results) over time, as well as preserving the digital software environment needed to reproduce key analyses? (Preservation and Reproducibility over time: the data, the programs/codes, the surrounding software ecosystem) (#45)
    - Part of data curation is deciding what to keep, for how long. And there is the question of how to characterize data quality. (#27)
    - Problem: how do we go about deciding which data/programs/environments to keep and for long. (#48)
    - (1) Lack of data intelligence: self-descriptive (auto-metadata), self-reducing to important features (if space is not enough), lineage (capture relationships to other datasets and evolution in time), self-transforming (switch between representations, accessible to quantum accelerators), self-replicating (e.g.quantum entanglement) (#17)
    - One topic that seems important to me is the preservation of meaning over time. An easy example is a Lotus-123 file. Who can still read it? We talk about HDF5, for example, but when we design a data format we need to think about how it can be losslessly transformed to the next format. That doesn't just mean the "data" but also the metadata. (#26)
    - Documenting institutional knowledge. (#44)
      * (4) Accessible: All data globally available and accessible. DOE science is done for the public so it should be publicly available. (#6)
        + (3) Problems and issues: privacy, security, classification of data. Can we have proven/acceptable methods of sanitizing data at different levels for different security domains? (#10)

Problem: publication and dissemination of new data that is categorized appropriately with existing data. (#20)

Problems and issues: performance. How can people around the globe get data without waiting a week for it to be transferred? (#14)

* + - * + (1) Lack of policies and incentives to make data publicly available to the scientific community as an expectation of federally funded programs. (#8)
        + (2) Federal funding and long-term commitment for data archiving, preservation, indexing and query. (#18)
    - Lack of and/or inconsistent metadata, provenance, history and conditions under which the data was collected. (#2)
  + (8) Accessible Data : Storage and Supporting Infrastructure (#46)
    - Problem: migrating data from older storage technologies to newer ones will at some point take longer than the lifespan of the available technology (Darrell) (#50)
    - Data acquisition and data storage technologies have tended to advance kind of apace, so even with new storage capabilities we can also expect increasing data volumes from instrumentation and simulation. (#49)
    - Problem: How can we store all the data? When will we reach the physical limitations of storing data with respect to storage volume and data generation rates? Can we do this without reducing data such that (potentially) valuable details are lost forever? (#21)
    - Regarding storage costs. I hate to cite one of my own papers, but I do not know anyone else who is looking at this topic right now. (#30)
      * https://storageconference.us/2020/Papers/14.MeasuringCost.pdf (#32)
    - Lack of I/O performance portability: increasingly heterogeneous storage stack, need data models that focus on intent rather than data movements (which is what we mostly have today) (#25)
      * Does this argue for future infrastructure that "self-organizes/optimizes" based on metadata and access patterns? (#29)
  + (6) Data must be FAIR when it is born and we as a community need to recognize and reward the researchers that generate and curate their data appropriately. (#3)
    - (1) Problem: reduce cost, complexity of producing FAIR data (#19)
    - Problem: institutional priority of making data FAIR: presently, individual PIs (at least in DOE) are tasked with this, when it should be something that is of broad importance and visibility (#23)
      * Regulatory complexity as a barrier to data discovery and re-use. How can we "safely" combine open data with protected/restricted data to enrich the discovery process? (#24)
    - Retroactively making data FAIR is really hard! Expertise in the data may no longer be available. Perhaps focus more on looking forward. (#22)
    - Two FAIR mantras to bear in mind: data should be as open as possible, but as closed as necessary. (FAIR does not equal open.) And data should be as distributed as possible, as centralized as necessary. (Big centralized repos are not optimal.) (#28)
  + Interoperable data: what are the barriers to and incentives for scientific communities to share data, to create data models/formats that are shareable and interoperable across a broad set of science use cases, and that are usable with a diversity of software methods? (#51)
    - We need self-describing and machine-actionable data to improve interoperability. (#52)
    - And for data integration across disciplines, we need cross-walks between ontologies, metadata schema, etc. (#53)
    - Problem: How do we get data in a form/modality that is useful to the work we're doing today, given that it was probably originally created in a different format/context? (#54)
  + Findable data: Problem: search mechanisms for finding web pages are inadequate for scientific data, where we want to search over features present/not present in data, over assumptions made during problem setup or research question being asked, over various input/initial conditions, over author of the data, code/instrument that produced it, etc. (#55)
    - Many kinds of scientific data are difficult to find, reuse, and reanalyze because of lacking metadata, provenance, search functionality, self-description, etc. (#51)
    - Findability and Accessibiity depend a lot on the discipline. Astronomy, genomics, a few others are pretty good. Interoperability and Reusability are the tall poles. (#54)
    - I am going to be critical of the Google idea. Google returns (ranks) information based on popularity. That's not what we want. It says nothing about quality, or if it the data that we want. I think that we need something much more intelligent -- I am going to again mention a new profession of data archeologist -- probably AI assisted -- were you can describe precisely what you need and you get back matches that you can then evaluate. The number of choices has to be managable. (#91)
    - Problem: Accessing, interfacing with, and manipulating data is challenging because datasets often include raw information (tables, HDF5 files, etc) but exclude appropriate tools for interacting with the data. (#57)
      * Google dataset search is trying to solve this problem, using schema.org for data description. Not sure it is very successful though. (#59)
    - General abstractions to support transformation of data is a difficult problem. Requires domain-specific tooling or "intelligence" to act appropriately on self-described data and the formats/layouts requested. (#61)
      * Impactful research builds on prior work spanning long periods of time. Problem: How do we ensure that we're not "re-discovering" things? A means to discover \_all\_ relevant data (likely machine-assisted) is key. (#62)
        + (Riffing off Jay's idea): Want the ability for changes made in to scientific data be immediately propagated to other replicas in distributed federations (#63)
    - Need to speak the language of domain scientists, both for basic discovery and for discovery based on higher level concepts that are not typically embedded in metadata. (#65)
    - "Dynamic" Google for scientific data - eearch capability beyond metadata and labels: features, etc. - may need dynamic processing and reinterpretation of data (#81)
      * This point and #5 appear to be pointing at the same problem/idea (a good thing!) (#85)
    - Search for scientific data; Google doesn't work! (#93)
* Who would develop it (basic research to advanced deployment)?
  + A collaboration of computer scientists, domain scientists, and other stakeholders are needed to develop FAIR frameworks for scientific data (#16)
    - Agreed. We need both data users and data generators to determine how the data will actually be used and generated so that we support those scenarios. And computer scientists and regulators to make the system practical and to ensure that only appropriate data is exposed to the right audience (security/privacy) (#36)
  + Funding agencies must be partners in the development (#33)
  + Need to consider how such mechanisms will endure; perhaps by having professional "operators" so that it's not a burden on the PIs (#38)
* Who would use it and what skills would they need to use it effectively?
  + The scientific research community, regulatory agencies, schools/universities and perhaps even the public. (#34)
    - Agreed. analysis of data sets by the public can benefit DOE and vice versa (#39)
  + "The 99%"; not just projects/groups with lots of funding (#40)
* When would it be expected to be in production use (N years in the future)?
* Where, and how widely, would it be deployed?
  + Data sharing needs to be done in a distributed data lake over the wide area network, ie ESNET for labs and universities. Different centers will have different strengths, some will do high capability computing some will do high capacity computing. All we need to be able to cache the data but in addition certain centers will be need to be custodians for the data. As the expense and scale of these services increases, the scientific community will ned to be more organized and collaborative. (#37)
    - This is good. I like the custodians idea, since I guarantee that universities cannot and will not keep data for long periods. I have done some work on modeling this cost, and the paper is referenced in the other card comments. We need some kind of "data endowment" that will keep paying for maintaining the data after the project is gone. (#94)
* What is the setup time and/or process for using it?
  + Storage technologies will be a limiting factor here I believe. Not much research is being done in out of the box approaches to storing data (#42)
  + Ideally there would be no "setup" required; all investigators should have tools intuitively available to them on day one. (#47)

1.2 Breakout 2 - Implications of this Problem.

**The following participants have not been active:**  
Amber, Suren Byna, Maria Chan, Javier Duarte, \*fac- Nami Ishihara, Bei Wang Phillips, Line Pouchard

**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?
  + Message to organizers: Please ignore this card; all our notes from breakout #2 are in the next (blue) card. (#152)
  + Findable data: Problem: search mechanisms for finding web pages are inadequate for scientific data, where we want to search over features present/not present in data, over assumptions made during problem setup or research question being asked, over various input/initial conditions, over author of the data, code/instrument that produced it, etc. (#1)
    - Many kinds of scientific data are difficult to find, reuse, and reanalyze because of lacking metadata, provenance, search functionality, self-description, etc. (#-3)
    - Findability and Accessibiity depend a lot on the discipline. Astronomy, genomics, a few others are pretty good. Interoperability and Reusability are the tall poles. (#0)
    - Problem: Accessing, interfacing with, and manipulating data is challenging because datasets often include raw information (tables, HDF5 files, etc) but exclude appropriate tools for interacting with the data. (#3)
      * Google dataset search is trying to solve this problem, using schema.org for data description. Not sure it is very successful though. (#5)
    - General abstractions to support transformation of data is a difficult problem. Requires domain-specific tooling or "intelligence" to act appropriately on self-described data and the formats/layouts requested. (#7)
      * Impactful research builds on prior work spanning long periods of time. Problem: How do we ensure that we're not "re-discovering" things? A means to discover \_all\_ relevant data (likely machine-assisted) is key. (#8)
        + (Riffing off Jay's idea): Want the ability for changes made in to scientific data be immediately propagated to other replicas in distributed federations (#9)
    - Need to speak the language of domain scientists, both for basic discovery and for discovery based on higher level concepts that are not typically embedded in metadata. (#11)
    - "Dynamic" Google for scientific data - eearch capability beyond metadata and labels: features, etc. - may need dynamic processing and reinterpretation of data (#27)
      * This point and #5 appear to be pointing at the same problem/idea (a good thing!) (#31)
    - I am going to be critical of the Google idea. Google returns (ranks) information based on popularity. That's not what we want. It says nothing about quality, or if it the data that we want. I think that we need something much more intelligent -- I am going to again mention a new profession of data archeologist -- probably AI assisted -- were you can describe precisely what you need and you get back matches that you can then evaluate. The number of choices has to be managable. (#37)
    - Search for scientific data; Google doesn't work! (#39)
  + Findable data: Problem: search mechanisms for finding web pages are inadequate for scientific data, where we want to search over features present/not present in data, over assumptions made during problem setup or research question being asked, over various input/initial conditions, over author of the data, code/instrument that produced it, etc. (#40)
    - Many kinds of scientific data are difficult to find, reuse, and reanalyze because of lacking metadata, provenance, search functionality, self-description, etc. (#36)
    - Findability and Accessibiity depend a lot on the discipline. Astronomy, genomics, a few others are pretty good. Interoperability and Reusability are the tall poles. (#39)
    - Problem: Accessing, interfacing with, and manipulating data is challenging because datasets often include raw information (tables, HDF5 files, etc) but exclude appropriate tools for interacting with the data. (#42)
      * Google dataset search is trying to solve this problem, using schema.org for data description. Not sure it is very successful though. (#44)
    - General abstractions to support transformation of data is a difficult problem. Requires domain-specific tooling or "intelligence" to act appropriately on self-described data and the formats/layouts requested. (#46)
      * Impactful research builds on prior work spanning long periods of time. Problem: How do we ensure that we're not "re-discovering" things? A means to discover \_all\_ relevant data (likely machine-assisted) is key. (#47)
        + (Riffing off Jay's idea): Want the ability for changes made in to scientific data be immediately propagated to other replicas in distributed federations (#48)
    - Need to speak the language of domain scientists, both for basic discovery and for discovery based on higher level concepts that are not typically embedded in metadata. (#50)
    - "Dynamic" Google for scientific data - eearch capability beyond metadata and labels: features, etc. - may need dynamic processing and reinterpretation of data (#66)
      * This point and #5 appear to be pointing at the same problem/idea (a good thing!) (#70)
    - - (#76)
    - Search for scientific data; Google doesn't work! (#78)
* Who is/will develop this companion technology/service?
  + How much of the necessary technologies will come from the private sector, where we can ride the coattails? (#79)
    - The private sector is a for profit entity that does not have science as it's overriding and primary goal. Since they have more R&D dollars we must partner with them on technology development but we should not entirely depend on them. (#119)
  + Quantum technologies are likely to have a profound impact on capacity and capability. (#80)
  + Implication and consequences for... (#81)
  + Findable (#82)
    - If we can't do this: datasets may end up being recreated (#86)
      * Which would in turn require even more storage (#120)
    - Consequence of broadly findable data: the field of AI/ML will rapidly advance due to the prevalence of high quality training data for supervised methods. (#87)
    - Feature extraction: how can users search by data patterns instead of labels? can we do it in real-time at query time? (#88)
    - Discoverability: how do we build a lineage and relationships between data so that we can understand the big picture and explore for data even if we don't know what we are looking for? (#104)
    - If we can't do this new and integrative research will be unlikely to happen. (#89)
    - If we can do this: could this lead to compromised privacy in sensitive data? (#94)
    - Operating with "incomplete" data hinders progress; worst case: we miss new discoveries altogether (#97)
    - If done right, machine-augmented methods can help us find data that we would not have considered relevant (perhaps even in adjacent disciplines), hence opening the aperture to higher-quality research (#105)
    - Implication: achieving high levels of findability will require significant advances in feature-based search for scientific data, as well as advances in the feature detection/tracking/analysis methods themselves. (#111)
    - If we really do it right, in a distant future state we may have interesting data (and even new questions to research) "delivered" to us by intelligent machines, without explicitly requesting it/searching for iit (#117)
  + Accessible (#83)
    - if data is accessible to a wider community, it advances the state of the art more quickly, faster time to solution (#90)
    - Consequence of not doing it: science progress will be impeded by not having access to high quality training data for Ai/ML. (#91)
    - How do we balance openness with concerns for security and national competitiveness? (#92)
    - If we can: we can establish common benchmarks to test methods against (eg OpenAI Gym) (#106)
    - How do we incentivize data sharing? (#107)
    - With greater accessibility, new discoveries may come from unexpected sources/people/communities (perhaps even citizen scientists) (#112)
    - If data is not accessible, there may be problems that cannot be solved. For example if old datasets are deleted or inaccessible for experiments that cannot be reproduced, then the problems needing that data cannot be solved (#116)
    - How much infrastructure will we as a research community have to develop and how much can be adopted or adapted from commercial developments? (#118)
    - Performance, scalability, portability (#121)
      * Can we focus of properties of data instead of explicitly deciding actions (like data movements)? (#126)
    - Implication: need a place to store these exabytes of data for a long period of time. The program as a whole needs to be mindful of long-term data storage/data lifecycle needs and requirements. (#122)
      * Will require proactive, programmatic investment by the DOE (#123)
      * Can we store all the data? If not, can we replace it with its featues? Implication: we may need to develop novel feature extraction methods/tools. (#125)
  + Interoperable (#84)
    - If we can't: millions of PhD student hours will be lost to data wrangling and preprocessing that has already been done elsewhere (#109)
    - Quantum accelerators: how to we move between classic and quantum representations? Can we leverage entangelement to store many variations without collapse when we want to read data? (#93)
    - Consequence of not doing it: reinventing the wheel as individuals recreate key software methods, systems, infrastructure. (#95)
    - If data sets are interoperable, then results using them can be more robust because they use more data (#96)
    - If we do not solve the interoperability problem, we will miss out on much of the multi-disciplinary science opportunities enabled by big data (#98)
    - Interoperability is key for assuring reproducibility, robustness, and reliability of research. (#110)
    - If data sets are not interoperable, or there are no clear guidelines for mapping data labels from one set to another, scientists may incorrectly interpret datasets which could lead to erroneous results (#113)
    - Do we need to standardize APIs and/or representations to manipulate data? If not, can we rely on automated techniques to achieve this? (#114)
  + Reusable (#85)
    - Implication of improving data reusability: this is a quantifiable metric that could potentially incentivize creating/sharing FAIR data, as well as for measuring impact, not unlike how h-index and publications work now. (#115)
    - If we can: methods and algorithms could be more generalizable and robust by testing on a larger variety of datasets (#99)
    - If data is adequately notated so that it is reusable, scientists can more easily benefit from the data of other efforts (#100)
    - If we can: more credit will go to experimentalists who produce the data (#101)
      * how can we provide attribution of credit and data packages as a measured deliverables of programs (#127)
    - Consequence of not having broad reusability: scientific progress will be hindered as advances in one area are not able to be leveraged in other areas: making use of experimental data from imaging/microscope facilities to help with computational materials design. (#102)
    - If we do not make data reusable, then when new theories and understanding appear we will not be able to check old experimental data consistency against those new ideas. Some experiments are too expensive to redo. (#103)
      * and some samples are unreproducible. (#128)
    - If data is not reusable, then scientists may generate similar datasets to solve their problems. This is a waste of time and space (#108)
      * and resources (#124)
      * Intermediate data: capture, cache (or preserve it long term) and revisit to avoid recomputation (#139)
    - Need to make sure that researchers understand that re-use and re-purposing of data is first-class science. Might be a change of culture in some fields. (#129)
    - Versioning semantics: fork, merge, etc - Need git-like services for data (#149)
  + The Human Element: Skills/knowledge we need to develop (and continue to support). (#130)
    - Implication: the skills we currently teach will need to change if some of the above are realized. (#131)
      * New skills needed to next generation drive data management (e.g. skills needed to build AIs that auto-manage data) (#147)
    - Implication: New disciplines will emerge focused on training investigators how to manage data in this Brave New World (#138)
    - STEM education needs to evolve to make data science a central part of the curriculum (#136)
    - We may not be able to rely on the subject matter expert (scientist) alone for data maintenance. (#135)
      * Data now crosses disciplinary boundaries, so there is no single field where the expert lives. (#140)
      * We may need to create an entirely new profession of "data archeologist" (or "data curator") who understands the nature of the data and its associated metadata. (#145)
    - Implication: we'll need to change the mindset about competitiveness. Perhaps people hang onto their data and keep it private so that they can be the only ones that publish results from it. How can incentives to publish the data be more enticing than "winning"? (#132)
    - Science is becoming increasingly data dependent and data-driven, which means that scientists will need to evolve to become more aware of data-centric challenges, tools, opportunities. (#133)
    - A career path for domain data scientist must be supported in agency funding for universities and labs. (#134)
    - A common language for data - having people from different fields learn common terms used to communicate data (#137)
    - Implication: DOE will need to streamline procedures and processes for data release (#141)
      * sponsors will need to mandate data sharing guidelines and timelines (#144)
    - The FAIR world calls the new role "data stewardship" and there are already formal training programs being set up at universities across Europe. (#142)
    - Evolution to more data-driven scientific methods will require and encourage diversity of thought and openness to new ideas, such as reanalysis of existing data for new purposes. (#143)
      * How do we encourage people to challenge scientific orthodoxy? (#151)
    - Access to data for science is 'the great democratizer' when individuals can do science at home. (Stolen from Bob) (#148)
    - Implication: Improve the pipeline to STEM fields by making data available to US high schools and universities with less funding (increase the pool of trained candidates) (#150)
    - Intellectual property policies need to be evolved and standardized (#146)
* 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: 2

**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. The Universal FAIR Data Store - globally accessible by everyone; easy to contribute; automatically indexed; easy to search/discover; curated (to ensure accuracy/authenticity/etc.). May start out as a National Data Store and evolve. Alternatively a federation of National Data Stores, with mutually agreed-upon policies for sharing.
* 2. Examples of success in the direction of universal data store:
  + Comments
  + Examples of success: (#5)
  + HEP - data oceans, lakes, and "streams" between them (#6)
* 3. deprecated
* 4. deprecated
  + Comments
  + Climate: community-wide data stores like the Earth Systems Grid Federation (#7)
* 8. Collaboration documented here: https://docs.google.com/document/d/1k4oj11JHbJVAqignHg0Y2tuJ4qHr8NJOtsiehZnhEQI/edit#

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. Data Storage (hardware and software frameworks)
  + Comments
  + Signpost: Emergence of storage technologies with densities that far exceed today's devices (https://www.scientificamerican.com/article/dna-data-storage-is-closer-than-you-think/) (https://www.microsoft.com/en-us/research/publication/glass-a-new-media-for-a-new-era/) (#2)
  + Precursor technology: Development of new storage substrates/materials How to store all the data? Quantum? The Moon? This should be a cross cutting topic with the material scientist in DOE. We are running into physics limitations, but we have a lot of silicon. Writing to glass is essentially permanent, reliable, and we can read it fast -- much faster than we can write it. The moon has plenty of sand (being slightly sarcastic). DNA may be a possibility, but at this point it is easier to read (nanopore technology) than to write. Can we imagine a process for DNA synthesis that would be fast enough? (#3)
  + Precursor technology: Next generation "compression" mechanisms that can reduce storage requirements by order(s) of magnitude. I would say "synthesis" rather than compression. The reason is that we have two forms of compression: lossy and lossless, and what I would suggest is something that derives "meaning" from the data. Feature extraction If we decide to delete original data because we have synthesized it, we need to be aware that there is no recovering from that if there is a mistake. (#4)
  + Signpost: automatic curation and indexing (via bot) during and post storage; transparency and reproducibility for the curation pipeline (#5)
  + Technology 10yr: ability to store, share 100 EB/yr of curated scientific data; at the local scale, aggregate storage of 1 ZB/yr of fresh, off-the-instrument scientific data (raw). (#6)
  + Technology needed in N years: How fast is the network? Liz mentions HEP ESnet requirements review process. Mechanisms for staging computations close to data to avoid data movement. Radically new network technology to accommodate future anticipated data movement loads/requirements. (Vas showed a slide from ESnet giving perspective on how long it takes to transfer data of a given size within a given period of time. Discussion about using photon spin, etc. to encode information. https://fasterdata.es.net/home/requirements-and-expectations/ (#7)
* 8. Data Organization (Description/Ingest/Indexing/Tagging)
  + Comments
  + Signpost: Summarizing data, derived features, data organization (was Self-organizing data) Automatic capture of the evolution of datasets and relationships between them Automatic feature extraction and summarization (as additional metadata) and/or reduction to these features Precursor technology: Intelligent/machine-aided data transformation and reorganization, as new data are added. Precursor technology: versioning (github) for data Technology needed: automated "crawlers"/agents that process/analyze data in large data stores in the background that can help to do fundamental analyses, to generate hypotheses, to do hypothesis validation. (#9)
  + Precursor technology: need Network technologies that can transfer large amounts of data with reasonable latencies (e.g. quantum) far beyond today's technology See above for more comments about networking technology (#10)
  + Precursor technology: automated translation between data formats. May need to discover a path from format A to format B by potentially traversing many intermediate formats. What is the shortest path? (#11)
  + Signpost: Adding situational awareness to data and making data active in the sense the information in massive amounts of data actively reaches who seeks it or might seek it in future. Information is retrieved and organized or reorganized actively according to accesses of data AI techniques to understand data use and reach those who might need that information Data management systems to extract information proactively and manage relations among various datasets Technology needed : Tag data based on its importance - define value of data (based on the information content, etc.) - add beacons for information in data that would be visible to anyone and that info would "travel" to potential consumers of that information. Could be done by an agent: you could show the agent examples of what is important, and then the agent could do the tagging. Technology needed: Proactive Data placement by intelligent infrastructure (caching, but also explicit policy) based on consumer needs. "Logistical" notion of data, which can include online transformations. (#12)
  + Signpost: Automatic generation of metadata and management of ontology (Explainable data) Growing acceptance of AI/ML generated data and availability of mechanisms to "validate" accuracy The impact of AI on future data may lead to formats that are not easily understandable for humans (e.g., learning models or used to exchange information between AIs), which goes against the FAIR principles Precursor technology: keep track of provenance and enforce FAIR-ness (e.g. automated sensitivity analysis that labels models with a degree of confidence) Precursor technology: datasets must be "born FAIR", meaning that experiments are designed with metadata collection simultaneous with data generation, according to universal standards, ontologies, etc. Precursor technology: ability to track evolving ontology/history of data; flexible systems for interacting with and updating metadata in scientific data Precursor technology: Automated tools for capturing provenance information and storing with data sets (#13)
* 14. Data Protection (Security/Privacy/Policy)
  + Comments
  + Signpost: Accepted and widely deployed security measures to protect data store from attack or from unauthorized access of data There are no obvious barriers to using cryptographic protocols to prevent unwanted data manipulation (via digital signatures), the issues are sociological and political Precursor technology: Effective key management technology; this is the difficult (sociological and logistical) problem in cryptography, and there is no really good solution known. This capability will help to guarantee the integrity of data collections (prevent unauthorized modification, deletion), as well as to guarantee authorized access to data. The biggest problem -- and this cannot be overemphasized -- is quality of implementation. Virtually all security breaches are due to bad engineering. Essentially no one breaks modern cryptography, they find holes in the implementation Cryptographic algorithms preserve data content and meaning, so there is no issue in moving from one cryptographic protocol to the next -- provide you did not lose the key Precursor technology: Cryptographic protocols will need to be developed to enforce the policies. Specifically, engineering techniques to generate code that is provably bug-free: current exploits are based upon bugs in code, avoid turning data into gibberish. (#15)
  + Signpost: Availability of automated and secure mechanisms for anonymizing sensitive data so that it can be shared (e.g., health data remove identifying information from public store or don't return it on query to unauthorized person/requester) Precursor technology : robust anonymization tools to support doing useful things with data, but that does not permit de-anonymization. (#16)
* 17. Data Discovery (Search/Machine-aided methods)
  + Comments
  + Signpost: Data crawlers (similar to web crawling bots) to find relevant data and information (features) Precursor technology: Mechanisms for scoring scientific data sets in terms of matching/meeting given search criteria (Group relevant datasets - new scoring methods for grouping (science data, social data, etc.)) Google Dataset Search is a start, which allows dataset owners to populate metadata - A data crawler would be proactive without requiring data producers / owners to enter metadata Tech : features and metadata to be searched. search capability based not only on metadata (labels), but also features (may need dynamic query processing to match requested features to data features). An interesting side view of this space is a specific query for scientific data meeting certain conditions may not return a dataset, but may return "code" that can compute such data. Tech: recommender systems that suggest potential search matches for "fuzzy queries", for "queries with typographical errors", for data from unanticipated sources/collections. Includes publish/subscribe mechanism to alert you when new, updated data are available. Tech: searching tech for massive scientific data collections Precursor technology: Mechanisms for easily retrieving different granularities of data (raw data vs summarized/analyzed data) (#18)
* 19. Data access (interfaces, access methods)
  + Comments
  + Signpost: Human-computer interfaces to data, data tools, things beyond text and APIs (Visualizations of information / features in data) Access information with superior human-computer interaction than with programming interfaces (voice, hand gestures, reading thoughts (brain waves), etc. - "Minority Report" movie like interface) Holographic displays (in contrast to monitors), VR, AR? Access anywhere (in contrast to sitting at a desk) (#20)
  + Signpost: Declarative data to address portability (accessibility, performance) Precursors: Increasing heterogeneity of storage makes it impossible for applications to reason in terms of actions (read/write, data movements, etc.) Need to focus on data properties and constraints (age, sharing, scope, resilience, etc.) and features, let runtimes worry about actions expressive data models automated extraction of data patterns to optimize performance and scalability based on constraints Precursor technology: effective mechanisms for replicating/synchronizing lab/campus/regional/national data stores to increase accessibility at the edge (#21)
  + Signpost: Data from experimental facilities automatically transferred to data store (effectively implemented at scale) Precursor technology: storage and network infrastructure to support large scale data collection and transfer (#22)
* 23. Data Lifecycle
  + Comments
  + Signpost: mechanisms for data aging, purging. Over time, some data increases in value and is worth keeping, other data decreases in value to the point where it need no longer be kept around Precursor: model for assessing data value Signpost: bundling of data and software codes/environments for working with the data Precursor: data product equivalent of container technology Signpost: low barriers for ingest of curated data into federated storage (all data is "born FAIR") Precursor: vetting of scientific data (#24)
* 25. Sociological: Data Sharing
  + Comments
  + Signpost: social agreement on sharing of data (break the "habit" of everyone keeping their data to themselves) Signpost: policies and social agreement on use of public data, in terms of attribution and publishing Signpost: federal agreement on policies for data sharing driven by funding agencies based on the use of taxpayer dollars for research. Signpost: accepted policy on sharing data with commercial research partners/collaborators (no lawyers required every time we need to share one little thing) Signpost: wider community sees benefit of sharing data over keeping it private, what is the competitive advantage of sharing data? The competitive advantage of keeping it private is that you publish first/get the answer first Signpost: funding agencies mandate FAIR data as a requirement for funding Signpost: fully distributed storage a la blockchain model, smart contracts creating a data storage marketplace. Various projects implementing this now could be adopted for per-community needs. Signpost: training the scientific community (producers and consumers) in tools and infrastructure for such a data store. Experiments are designed from the ground up to produce FAIR data Signpost: data from store is available to limited users (e.g. DOE employees). Data stores are available more widely (e.g. US universities). Data store is open to all in the US. Data store is open globally (?) Precursor: community planning processes to roadmap long-term data/code sharing plans/strategies, eg 10yr plans (#26)
* 27. Sociological: Data Integrity
  + Comments
  + Signpost: policies to guide data verification emerge Signpost: communities within each discipline agree to adopt multiple ontologies/metadata schemas rather than block on single standards Signpost: need to come to community consensus on data format/metadata/provenance capture Signpost: developing community appetite for "opt-in" model of distributed and self-organizing curation and management. Distributed consensus to identify valuable data products vs the "bogus" Signpost: Crowd-curation to allow relevant groups of people curate datasets and improve quality of data. Curation, evaluation, review of data quality, correctness, etc. (like arXiv) Precursor: infrastructure for data-curation-at-home (like SETI at home). Example HEPData portal. Tech precursor: technology for automatic curation and indexing (via bot) during and post storage; transparency and reproducibility for the curation pipeline Signpost: provenance tracking and trust (#28)
* 29. Sociological: Data Access
  + Comments
  + Signpost: how can we enforce data equality (assess for all) as part of social equality (even for just scientific domains and users) Signpost: Timeliness of data access. Some communities will need to specify when data is of use, e.g. weather/climate/earth observation. Infrastructure support to deliver and notify when new data becomes available (mentioned this idea above). Signpost: Policies to decide who gets access to what and for how long. This is not a technological issue. Precursor: governance mechanism for deciding data aging/purging. Signpost: Methods for resolving the inherent conflict with data ownership -- especially in the classical scientific sense -- and public access to data (which they paid for) Some data, such as genomic or medical data, have legal and ethical privacy issues that must be respected. Some for the life of the patient, some beyond (who owns grandma's cell line?) (#30)
* 31. Economic: Funding Long-term Data Archives and Data Accessibility
  + Comments
  + Signpost: Funding to support long-term data archive and accessibility, including storage capacity, security, accessibility, curation and software. Precursor: models for sharing costs of infrastructure, operation across institutions, countries, diversity of science projects, consumers/producers of data Precursor: new funding models that go beyond pay-as-you go to foster stability and longevity. May include charitable donations, philanthropic, endowment-based, and tax-based models. (#32)
* 33. Economic: Incentives for sharing data
  + Comments
  + Signpost: In addition to awards, additional incentives for data sharing, including author citations for secondary use of data, similar to H-index, and sponsor-recognized deliverables such as data packages with high down-load or reuse as "high impact" datasets. Could be included in funding decisions for follow-on awards. (#34)

Appendix

Live chat

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

* Nov 2, 2020, 17:21 UTC
  + Testing chat (Ezra Kissel | Nov 2, 2020, 17:21 UTC)
  + Slack - "Just say NO" (Vas Vasiliadis - UChicago - Data Management | Nov 2, 2020, 17:22 UTC)
  + I have a conflict at 3pm EST, will have to drop off shortly for about 1h - apologies. (Ezra Kissel | Nov 2, 2020, 19:51 UTC)
  + OK, no worries. (Vas Vasiliadis - UChicago - Data Management | Nov 2, 2020, 19:59 UTC)