Report out:
Consensus: distinction between data management/standards/meta-data vs. drawing inference from data that includes variability, measurement error, bias. To get to the "to knowledge" part we need focus on the latter. Be aware inference: large sample size does not get rid of bias.

Analysis:

Methods developers need to understand the science they're operating on (the subtleties and the specific details in solving these problems).

- **Knowledge about the methods that are available, need someone knowledgeable in the methods to properly apply them; methods are not interchangeable. Need experts in data analysis involved in the analysis.**
- **New methods: need as many methodology developers as possible to have access to the data, need methods to extract knowledge from many kinds of data.**
  Through collaboration between biomedical people and methods developers
  Through access: put the data and the challenges/difficulties/needs out there and let data scientists operate on it. "free the data" (immediately)
  Don't handicap the methods developers by insisting they collaborate within BD2K or some circumscribed group.
  Education is important to avoid re-inventing the wheel.
  Need awareness of disciplines outside biology (e.g. astronomy, economics and their methods).

Demonstration of Success:

- **Communicating how the methods are integral to discovery and how they were used.**
- **Encourage people to write useable software: either developers themselves or through a third party (who makes the software useable).**
- **Computer scientists, statisticians aren't necessarily good software developers: facilitate the integration of software engineering into this whole endeavor.**
- **Need benchmarking data sets to test algorithms, approaches, comparisons.**

Data Science Methodology Breakout Session
BD2K AHM Friday, November 13th 2015;

The purpose of this breakout is to provide a forum for the discussion of the main technical data science analytical hurdles that stand in the way of progress in biomedical science. Focus on methods, approaches, software for doing data science: what are the big technical hurdles, what are the big opportunities for advancement, what's standing in the way of progress

High-level Challenges:

- Can we identify key technical challenges or opportunities in biomedical data science?
- Avoid re-inventing the wheel. Through education efforts assure that the new generation of data scientists have deep understanding of existing methods
- How to facilitate collaboration between data scientists who have the technical skills and biomedical researchers who have questions needing answering? How do we maximize

the number of method developers working with new challenges. How to incentivize sharing knowledge? Can we open new datasets to everybody? Collaboration of computer scientists, statisticians and software engineers. Data scientists need to express impact of their findings (public relations to make everyone aware of the role of a method in a discovery). Journal for RMarkdown or Shiny Apps, e.g. analogous to JOVE). DREAM challenge. Engaging analytics experts at the *beginning* of an experiment.

- Learning from other Applied Big Data fields (e.g. physics, astronomy, econometrics, etc.) Nuclear physics use of simulations, testability - learning from their methods development. Challenges in acquisition
- Facebook/wiki/stack exchange/Rbloggers/Hubzero (KnowEng/BBDT)/CrossValidated for data scientists to find expertise (person). Specialization of biomedical analytic challenges distinct from business analytics
- In the big data era, how to make methodological research financially sustainable and attractive for talented researchers? How can we maximize the number of independent investigators that focus on method development.
- How can we (as BD2K/data scientists) reach out to the broader biomedical community to find out what their needs are?
- What are the emerging data types needing data science analytics? (e.g. CLARITY in neuroscience, high res. EEG, cell mechanics, 3-D spatial data, time series data, big behavioral data); examples of behavioral big data include social media (FB, Twitter, YouTube), data (e.g., physical activity; sleep), the environment (e.g., food availability; weather/pollution), and internet of things (e.g., use of refrigerator to measure food intake)
- How do we assure that public data (for example the data on The Commons) is available in the optimal format for methods development? How can the data providers get feedback from analysts/method developers.
- Promoting reproducibility (e.g., source control, jupyter notebooks, etc.)

Discussion of pressing technical problems, for instance:

- Sampling bias, replicability, large sample size improves precision but does not remove bias
- Inference from analysis
- Data management
- Dissemination: Software Engineering, developing disseminable software. Modern methods for software deployment (software containers, VMs, Github). Balance with education of end-users to use appropriately. TopCoder crowdsourcing of SW eng.
- Scaling: Build, explore, scale *loop*. Use of workflow/pipelining systems. What are the "FAIR" principles for software? Provide developers access to data and informed on latest challenges. User feedback to identify useful tools. Starting with use cases to develop story of analysis
- Characterization of suitability of various software *programs* when facing specific analytic challenges.
- Bring more citizen scientists in, and help them contribute usefully
- Large matrix factorization and estimation
- Deep learning
- Dimensionality reduction
- Human-computer interaction (including user interfaces, visualization, visual analytics, usability, human-centered design, etc.) — gateways/portals for non-technical scientists to leverage developed methods/tools/visualizations; tools/methods for visualization-based exploratory analysis which integrate with underlying statistical models

- Missing data
- How to compare and contrast methods in a productive way? Evaluation approaches, standardized evaluation datasets, standaized simulation tools
- High performance computing
- Social aspect: how to connect well trained data scientist with lab in need of specific expertise but also how to connect data scientists with each other for specific subproblems
- Office with data scientists for helpdesk style service
- Open source tools in popular languages (matlab, python, R)
- Multiple-comparison challenges and pre-registration of shared data experiments
- Interface with infrastructure (the long tail of data sharing)
- Sharing FAIR (findable, accessible, interoperable, reusable) data
- Guidelines to provide rich metadata
- Electronic Health Record linkage
- Data integration
- Meta-analysis
- Uncertainty assessment
- Biomedical big data are often collected from a convenient, opportunistic sample/cohort, rife with issues including selection bias, missing data, mis-classification, and measurement error that may complicate data analysis and lead to biased estimation and prediction.
    - The existing methods may not be adequate for big data due to their scale and complexity.
    - “Smart” study designs for big data research that balance resource use and sound science. For example, combining a naturalistic design for data collection from a large cohort and a more traditional design for data collection from a small randomly sampled cohort, which may help alleviate and address aforementioned analytical issues.

- Problems of identifying mechanisms and causal relations. The problems there are diverse, for example:
    - Where can parallelization be effective for high-dimensional problems?
    - Are search algorithms most effective when the search problem is resolved into stages to which different algorithmic strategies apply?
    - Can scoring algorithms be developed for classes of problems for which only constraint based searches are presently available, e.g., search with unknown, unmeasured confounders?
    - Are there causal motifs that are transferrable across biomedical systems?
    - Systematic strategies for validating or invalidating search results by comparisons with experimental literature.
    - Causal inference under more realistic assumptions.
    - develop of computational primitives for distributed and parallel computing
    -

Broad Topics:

- What is data science in the context of biomedical big data?
- Essential elements of data science methodologies (e.g., probabilistic interpretation, algorithmic efficiency, domain specificity, software availability)

Report back Summary:

- In 10 years be able to demonstrate success related to methods
    - Software engineering/dissemination - metrics to measure impact (downloads flawed, phonehome, forum registration, community of users)
    - Methods benchmarks. Evolving/growing datasets
- Maximize # of methods developer that have access to latest data. Metadata about methods
- Software use cases to determine which method when. Methods are not interchangeable. Training needed to understand this.