Problem Statement: *The NSF NRC (National Research Council) Committee is seeking comments on "The role that private industry and other federal agencies can play in providing advanced computing infrastructure—including the opportunities, costs, issues, and service models, as well as balancing the different costs and making trade-offs in accessibly (e.g., guaranteeing on-demand access is more costly than providing best-effort access)."*
Disclaimer: Note that I am not representing Google here and this document is a statement based solely on my personal experience and publicly available information.

NSF-NRC should focus significant R&D on important grand challenge problems of national interest such as:

1. WRF observes a weather pattern developing. NCAR, NCEP, …, should be able to ask the following questions and get immediate answers: a) have we seen similar weather patterns before?; b) how did those prior weather patterns develop and progress?; c) how can we use that information to fine tune the current weather pattern model to speedup the simulation and get better accuracy; d) what can we learn about what led to this pattern? e) what can we do to mitigate these challenges? f) what deep learning is possible to gain further insights into the root causes here to potentially address climate change and better control the trajectory of hurricanes and other damage causing weather patterns?
2. Jane Doe has a mammogram taken in Pennsylvania in 2014. She moves to California in 2015 and has another mammogram taken. The doctor in California should be able to ask and get answers to the following questions: a) retrieve the mammogram from 2014 and find out what has changed?; b) figure out which other patients have undergone similar change?; c) how were those patients treated and what were the results? d) Use those insights to design an effective treatment plan tailor made for Jane Doe; e) request expert opinion from other doctors for more complicated cases online; f) file interesting cases securely in a different database for possible use for educating medical students; Similar infrastructure can also serve to identify and eliminate rampant waste, fraud and abuse in the healthcare system.

There are a vast array of grand challenge problems like above that serve the national interest and NSF should be championing solutions to these problems. Below we summarize some of the key approaches that need to be taken to develop the infrastructure necessary and bring the necessary focus to solve the example grand challenge problems discussed above.

Summary Position: Plenty of evidence that there are several industry inflection points that suggests that NSF re-think its funding priorities. The key themes in this paper are:

1. **Adoption of public cloud infrastructure**: Leverage vast investments from industry in this space. Most open source services good enough. EoML implies rapid commoditization which translates to advantages with large volume.
2. **Invest in Velocity of Scientific discovery for grand challenge science problems**: Increase the velocity with which scientific discovery can occur; We discuss ideas below.

3. **Invest in Learning from Data**: Re-focus attention on rapid advances in machine learning and their application to science and learning through data (Bigdata);
4. **Invest in special purpose solutions**: Extreme scale serving, custom accelerators for national security specific services and challenges. Also a consequence of EoML.

This position paper is organized as a set of observations, and some recommendations. It also highlights some key limitations that NSF will still have to address in an ongoing fashion.

- **Observation 1**: Public Cloud infrastructure is growing at a very rapid rate (Amazon, Google, Microsoft, IBM, …) and there are a large number of startups developing value add services. This is one of the hottest areas industry is rapidly growing in today.
- **Observation 2:** Public SEC filing examples show companies investing several billion dollars each quarter in warehouse scale data centers development. Ability to leverage economies of scale clear from the vast investment in warehouse scale computing. It is going to be very hard for NSF/DOE/DARPA to outpace the industry in this space and there is no national interest served in this goal.
- **Observation 3**: Continued obsession with peak performance (aka Linpack) rather than sustained performance of important applications is distracting from the more fundamental goals of improved programming and data models that help with focus on the science and insights from data rather than focus on computer science. Good to see that this trend may be finally slowing.
- **Observation 4**: Lack of velocity in insightful science due to reliance on very complicated programming models. This has been a complaint from scientists for almost 25 years. Numerous efforts through X10, Fortress, and Chapel, PGAS models went through significant research over the years but still face significant adoption challenges. Need to decouple performance from usability. Its okay to try and improve both but usability should trump performance during the experimental phase and performance should be the focus in a serving system. A lot of scientific discovery is dominated by experimental work. **Anecdote**: MapReduce or BigQuery for some special cases may be inefficient in terms of runtime (in most cases they are just as efficient or in same order) compared to an MPI parallel program solving the same problem. But the time to write the program is dramatically lower with MapReduce. Need to focus on end user time rather than on machine time during the experimental phase. It takes me about 30 minutes to write a useful parallel program with MapReduce that I can successfully run on thousands of machines. To write a useful program in MPI to run on thousands of machines is at a minimum several weeks of serious work and the learning curve for writing good MPI programs is also very slow.
- **Observation 5**: The open source solutions for Operating systems, programming models, cluster file systems, workflows and resource management and tools are robust enough that it is very hard for proprietary solutions to have the ability differentiate and charge premiums to support significant investment into R&D to enhance their products.

- **Observation 6:** Many of the findings in the 2011 Magellan report need to be revisited. Findings 1, 5, and 9 are still valid. Findings 2, 3, 4, 6, 7, 8, probably need to be revisited.

- **Recommendation 1**: The vast majority of NSF/DOE supercomputing site workloads can and should migrate to Public Clouds (e.g. all batch workloads, workloads that scale to less than 25-50K processors).
- **Recommendation 2**: There will be need for some large NSF/DOE supported supercomputing centers that caters to a) extreme scale computing; b) real time serving systems; c) national security controlled data processing; c) applications support that are extreme (say require 100K processors or more). One or at most 2 sites should suffice. Note that Findings 1 and 5 of the Magellan report still apply.
- **Recommendation 3**: There is a need for investment in special purpose accelerators that serve critical national interest needs given the impending EoML. These should be well informed with the value to national interest and the advantages of leveraging special purpose acceleration. One area that needs more investment is - what can we do to accelerate the hardware development cycle.
- **Recommendation 4**: Machine learning has proven to be a remarkable technology that has come of age. It is a critical element of BigData, and infrastructure to improve services for a) classification; b) clustering; c) feature analysis; d) training; e) inference; f) prediction; g) recommendations; etc. are critical. Focus should be driven by these emerging use cases that can significantly increase velocity of scientific discovery. Matlab, R, Octave, BigQuery are far easier and should be the focus of speedup in turnaround time for experiments and improved user experience.
- **Recommendation 5**: The industry has realized the potential and has been investing heavily in deep learning models and infrastructure for deep learning. NSF should encourage the rapid adoption of this infrastructure and leverage the vast amount of data in its **"Learning from Data**".
- **Recommendation 6**: NSF should encourage rapid adoption of these new public infrastructure by allowing researchers to use get more cost effective value from Public clouds, allowing scientists to focus on the science and engineering insights rather than being bogged down by intricate computer science algorithms and art. NSF should invest in encouraging the rapid adoption of learning from data for the many scientists that have collected and generated vast amounts of data over the last several years. It is now time to realize deep scientific value from that data.