**Title of proposed program:** Coordinated Approaches to Big Data

**Submitting Source:** NIH

### What is the major obstacle/challenge/opportunity that the Common Fund should address?

The complexity of human biology in health and illness is increasingly being taken into account by research design, with individual studies collecting "omic," image, biosensor, and clinical data, along with information about sociocultural and environmental factors. These large, diverse datasets are almost always collected in digital form. Thus, modern biomedicine is confronted at once by great opportunity and great challenge. The opportunity presented by collecting multiple measures is to understand disease and gain insight to its prevention, treatment and cure, from a broad, encompassing perspective more likely to bear fruit than from studies limited to a small number of measures. The opportunity presented by collecting digital data is the ability to share, compare, reaggregate, reuse, and integrate data, as well as to use these data for models and simulations in ways that have been heretofore impossible. The challenge, however, is to be able to organize, present, analyze and manage these data to fully realize such opportunities. The challenge is one of "big data," where handling and working with complex data at large scale is both quantitatively and qualitatively different than at a smaller scale. This is a complex topic that was recommended through the 2011 Common Fund strategic planning process but was simultaneously articulated as a profound need in many other venues. Consequently, the NIH Director asked the Advisory Committee to the Director (ACD) to consider the challenges and to make recommendations for a trans-NIH approach to this topic. Those recommendations will be presented in June, 2012 and will inform the refinement of the concept described below.

### What would the goals of the program be?

- Improve storage, curation, integration, and analysis of diverse data sets, incorporating data on a continuous basis from the community at large
- Development and sharing of broad-based computational tools, making them available to the scientific community
- Train next generation of computational biologists
- Ensure core competency of all NIH trainees and fellows in quantitative biology
- Ultimately use the information contained in complex data sets to improve human health

### Why is a trans-NIH strategy needed to achieve these goals?

Currently, NIH-supported resources in this area are helpful but not sufficiently integrated, broad and/or powerful enough to address the growing need to integrate multiple data sets. Large datasets are being generated by many programs across the NIH, but the mechanism to combine data sets to make best use of resources and data are not in place. The integration of data systems to enable this type of synergy is inherently a trans-NIH activity.

### What initiatives might form the strategic plan for this topic?

The specific initiatives to be implemented will be informed by the recommendations of the ACD.

### If a Common Fund program on this topic achieved its objectives, what would be the impact?

This program will establish computational and informatics tools and technologies to improve analysis of large datasets and to combine data from disparate sources so that comparisons can be made. This has the potential to enable the establishment of virtual consortia in which individual studies are made more powerful through integration of their datasets. The result will be more rapid development of insight in basic and clinical research and the resulting increase in the rate at which basic data is translated into improved human health.