

08

**Fall**

**From Genes to Stars:**

**Leading the Science of Big Data**

***When Leeuwenhoek peered through the first microscope, he discovered a new world that revolutionized medicine. An equivalent moment in   
history is upon us.***

*We are crafting a radically new lens through which to peer not just at microbes but at nature’s many laws as embedded in “Big Data.” When this “datascope” finally brings today’s massive datasets into focus, it, too, will reveal a whole new world. Johns Hopkins has a choice. We can lead this revolution or we can follow.*

“If I had asked people what they wanted, they would have said faster horses.” Henry Ford

## What is the Problem?

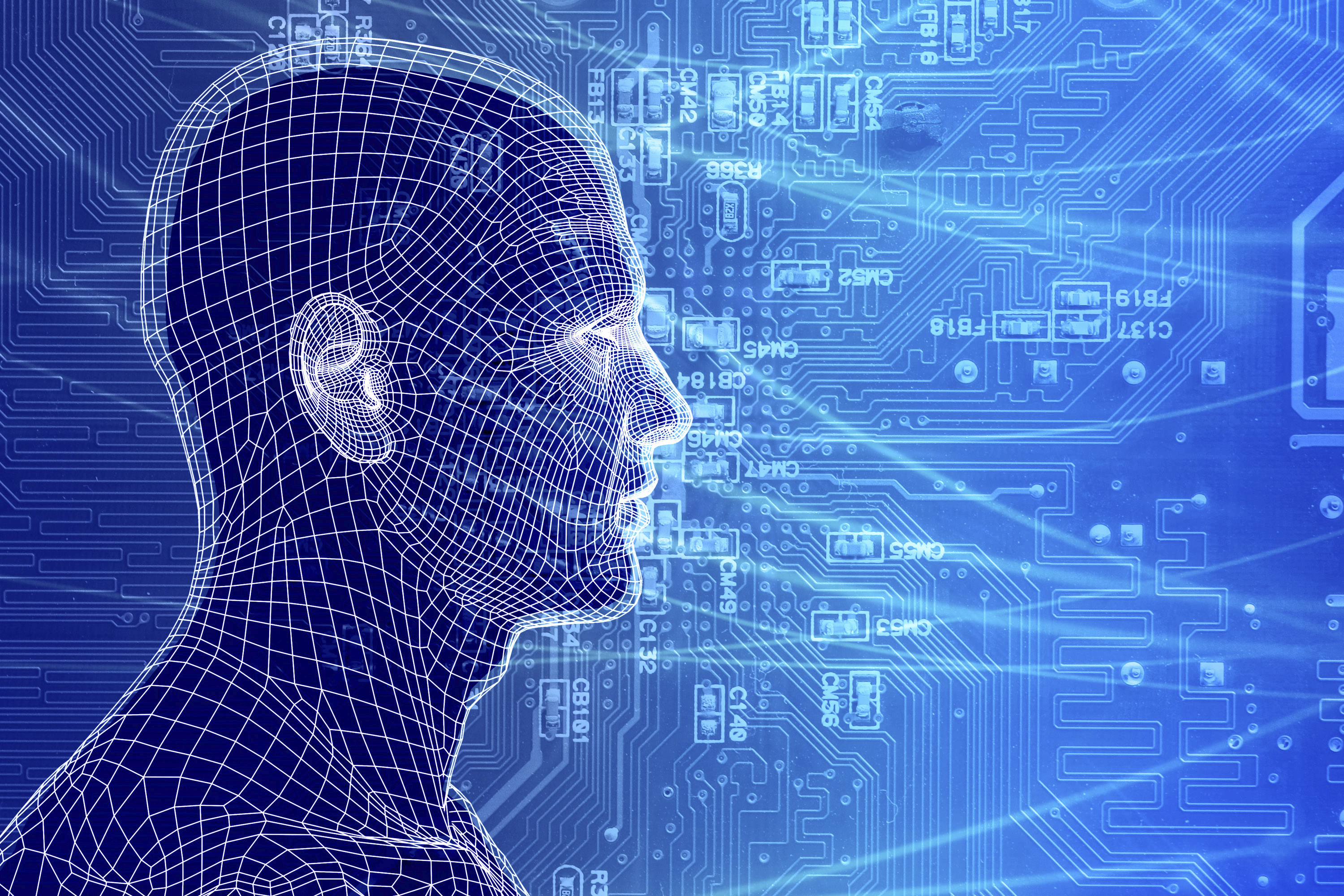
Today’s scientific instruments and computer simulations are churning out data about natural phenomena from genes to stars at exponential rates. This data deluge is hurtling modern science toward an impasse. Although datasets represent an unprecedented resource for advancing knowledge, their sheer size and complexity defy our ability to extract meaning from them. Now is the time to discover novel theories and methods that will guide the exploration of vast and diverse collections of data and drive new scientific insights. Otherwise, all the computer power in the world won’t lead us to new knowledge.

Data-intensive scientific analysis has been called the *fourth paradigm* because it is an entirely new way of *doing* science. Still in its infancy, it is the next great advance in the evolution of the scientific method. The earliest science was empirical, relying on the careful observation of the natural world. Later investigations became more abstract, generating theories based on mathematical and other models to explain reality. The third paradigm—computational science—employs computers to produce simulations of complex natural phenomena for study. Data-intensive science synthesizes all three earlier methods and overlays them with statistical tools to probe digital information for answers to scientific questions. *Developing this revolutionary approach and pressing it into the service of humanity is one of the greatest challenges of our time.*

## Why is this important?

All sciences, both physical and social, as well as some humanities, are increasingly data-intensive. Whoever develops and applies the “science of Big Data” first will leapfrog the competition in every field. No spectacular advance in any single discipline can catapult Johns Hopkins into the top echelon of universities. But creating the science of Big Data to drive discovery in fields from individualized health to the science of learning to energy and the environment will do just that. This opportunity promises a kind of leadership that Johns Hopkins has not known before.

Success in manipulating Big Data will be transformational, enabling discoveries and advances that are currently out of reach. This new method of inquiry is a key to solving our most intractable problems and to answering our most fundamental questions. With it we will discover new cures for disease, new solutions to public health challenges, new knowledge for preserving the environment, and unimagined insights into the universe. By deciphering data, we will understand our world as never before—and across the entire spectrum of mysteries from genes to stars.

This new approach is essential for achieving the promise of the proposed signature initiatives of the university’s fundraising campaign. The science of Big Data will make **individualized health** possible and will propel needed transformations in health care, including controlling costs. It will enable researchers to analyze vast amounts of complex genomic and clinical data (from gene sequencers, electronic patient records and imaging equipment) in search of optimum prevention and treatment strategies. The science of Big Data will fine-tune our ability to protect and preserve the **environment**, anticipate natural disasters and reduce the threat from **emerging diseases**. It will enable researchers to tease the truth out of new types of environmental data generated by sensors in the ground and satellites in the sky. The science of Big Data will allow researchers to use massive datasets generated by brain scans to radically advance knowledge about **how we learn** and how we can ensure that every individual reaches his or her potential. And the science of Big Data will capitalize on modern terrestrial and space-based telescopes—prolific data producers—to reveal **the origins and fate of our universe**, and to tell us whether or not we are alone.

This new science will extend its reach far beyond these important initiatives. It will, for example, open new vistas for understanding how the world’s many languages—past and present—have evolved and how they work. It will provide both economists and political scientists with unprecedented tools for understanding human behavior. Indeed, the promise of this new scientific method cannot be understated, and it is already under way. Recently, a collaboration of Johns Hopkins scientists in medicine, biology, statistics and computation applied innovative methodologies to analyze large amounts of data they collected to study the epigenetic basis of cancer. When they studied the whole genome in ways that could not be done before, their “datascope” brought into focus something entirely unanticipated and *hitherto unobservable* – in cancer tissue, areas of the epigenome that enable normal cells to differentiate had lost their boundaries and become chaotic. The discovery is potentially a major advance in cancer biology, suggesting that effective cancer therapy might include helping cancer cells return to normal.

Johns Hopkins can catapult itself into the top tier of research universities by aggressively exploiting its current lead in the 'Science of Big Data,’ an emerging field cutting across substantive sciences, mathematics, statistics, and computer science. It is about discovering knowledge amidst complex, oft-confusing information in large, complex data sets. Big Data science is an essential foundation for our university-wide initiatives in Individualized Health, Space Science, Learning, and Water. It will also provide us the opportunity to re-invent PhD education, reaffirming our leadership role established at the inception of JHU.

Leading the development of the science of Big Data will distinguish Johns Hopkins among all universities, and will give our faculty an unmatchable advantage in advancing their disciplinary and interdisciplinary work. And because this new scientific methodology requires a new kind of researcher, Johns Hopkins will again lead the reinvention of doctoral education, as it did more than 100 years ago.

## Why Johns Hopkins?

Faculty members at Johns Hopkins were among the first to recognize both the promise of Big Data and the impending investigative brick wall. Two decades ago, a small group of astronomers began pioneering solutions. With more ingenuity than resources, they have improvised their way into a worldwide lead. As part of the monumental Sloan Digital Sky Survey, they built an accessible astronomical data archive equivalent to the Human Genome Project. Later they parlayed a private gift of $350,000 into an unprecedented academic computational system that rivaled the performance of most supercomputers in its ability to handle large data. That feat quickly led to several initiatives involving other Hopkins researchers —including computer scientists, applied mathematicians, and environmental modelers—to develop new tools for dealing with the large amounts of data in their respective fields. Now they are constructing the biggest academic-based data supercomputer in the world, aptly named the Data-Scope, to be applied to fields as diverse as genomics, turbulent fluid flow, cosmology, and ocean circulation. Johns Hopkins’ competitive advantage also extends into data publishing and curation, two critical areas for the future of this new science. (The Libraries are leading the national-scale Data Conservancy project funded by the NSF.)

At Johns Hopkins, developing the science of Big Data also offers a powerful force for achieving the goal of One University. All sciences are becoming data-intensive, so all investigators have a compelling interest in the rapid development of solutions. And these solutions—as well as the pursuit of new knowledge itself—are inherently multidisciplinary. The convergence of the physical and the life sciences is happening now. Data-intensive science at Johns Hopkins is an engine for collaboration and coordination among investigators across many different fields. And our premier standing in biomedical and physical sciences offers an unparalleled environment for securing our position in the vanguard of this revolution.

## The Importance of the Science of Big Data Across the University

### Medicine

* New methodologies for unlocking the secrets of Big Data are integral to a wide array of medical breakthroughs, from cancer research to treatment protocol evaluation to genomics to neuroimaging. New tools to understand data will allow us to transform the way we monitor, study, and treat patients and how we unlock the secrets of disease.

### Nursing

* Harnessing Big Data at the School of Nursing will allow us to gain a deeper understanding of and insight into very large datasets that affect the health of the population.  It will allow our researchers to generate scientific hypotheses from large collections of patient cases or post-approval drug safety surveillance reports to detect patterns that were not previously visible.

### School of Arts and Sciences

* Ensuring our continuing excellences in a range of fields such as physics and astronomy, biology, neuroscience, and cognition depends on the development of new intellectual paradigms of Big Data. KSAS faculty are conducting simulations of Earth's atmosphere, oceans, and climate system, as well as of complex biophysical and chemical processes, generating huge amounts of data. An investment in the Science of Big Data will enable data exploration and visualization that would otherwise not be possible.

### Public Health

* Increasingly extensive and electronically archived patient records offer an opportunity for public health researchers to improve understanding of the connections among disease processes, care processes, and environmental influences as they affect human populations. We need new Big Data methodologies to delve into this deluge of information and find the clues that will lead us to better detection, treatment, and prognoses strategies. In addition, the Department of Biostatistics is a global leader in innovative methods for genomics and brain science that are only limited by our capacity to use ever-growing amounts of relevant information.

### Engineering

* Developing novel concepts and tools to deal with Big Data is crucial for fields as diverse as environmental modeling, speech processing, computational medicine, materials science, and robotics.  The strengths of the Departments of Computer Science and Applied Mathematics in databases and statistics, and an emerging focus on systems engineering, will play a central role in enabling such developments.

### School of Business

* Research, especially in marketing and finance, increasingly involves building and verifying sophisticated explanatory and predictive models that are based on data-intensive computing for analysis and the discovery of often subtle patterns and trends.

### Libraries

* The Hopkins Library system is building a data curation infrastructure for grand challenge science problems, such as modeling and mitigating vulnerability to climate change, natural disasters, energy shortages, and food insecurity. The Science of Big Data’s new methodologies will allow us to unlock this data and use it to help solve some of the world’s most pressing and complex problems.

What will we do? How will we do it?

Our most urgent need is to spread this revolution from a small core group out across the university and beyond. This transformation requires a critical mass of specialists, the means to encourage high risk and innovation, and the participation of more sciences, so as to discover general principles and methods of Big Data science from the similarities of problems across many fields. Establishing an institute dedicated to the science of Big Data will provide an overarching presence, the impetus to create and nurture a community of scholars, and the resources to harness and apply their collective intellectual power.

Foremost is the need to bring more data-intensive science experts onto the faculty. We seek people who have expertise in a scientific discipline—such as astrophysics—plus statistics, plus applied computer science. In the future, such multidisciplinary people must be commonplace. The science of Big Data requires investigators who can be conceived as “π-shaped” (deeply knowledgeable in at least two fields and broadly knowledgeable in others), rather than today’s “T-shaped” researchers who are deep in a single discipline and broadly knowledgeable in others. Today’s geneticists, for example, must also be adept in statistics and computer science to capitalize on the promise of massive data-generating tools such as gene sequencers. We must attract to Hopkins some of the rare experts who exist and enable them to straddle more than one departmental home without compromising their careers.

Developing the “Science of Big Data” is an integrative task unlike any in science before. It requires the collaboration of statisticians, computer scientists, and disciplinary experts who can merge their skills to create new tools (computer hardware and software), new approaches (algorithms and statistical methods), and new methodology that leads to discovery through data-driven rather than hypothesis-drive inquiry. In this “brave new world” of Big Data, here are some of the challenges ahead:

* As the size of data impedes the ability to move it from one place to another, we must be able to “push the computation to the data.”
* We need to design large parallel computers with extreme data-handling capabilities – massive, high- speed storage, and comp­rehensive analytics and visualization capabilities – all in one system.
* We need easy-to-use tools to manipulate and transform the data to recognize new patterns.
* With petabytes of data, spreadsheet-like views are unfeasible. We need visual, immersive explorations of data.
* We need to be able to get approximate results from our queries to the data quickly, using advanced statistical sampling ideas.
* We need to be able to “learn” typical phenomena in very large data sets automatically and recognize anomalies.
* We need to find appropriate transformations or projections of the data to reduce its very high dimensionality.

It is also a priority for Johns Hopkins to lead the training of this new breed of investigator so that they are no longer rare. Postdoctoral fellows and graduate students trained in this science at Johns Hopkins will be highly sought. Fellows and students ready to plunge into this new world will be difficult to fund through traditional sources because these new scientists do not fit the single-discipline mold.

We should build and endow unique data collections of great significance to science and serve them up from Johns Hopkins. This will help us secure an exceptional strategic and competitive advantage. By establishing control over valuable datasets in such areas as genomics, protein folding, turbulence or ocean simulations, for example, we can solidify our leadership while maximizing our freedom to experiment in designing them.

To capitalize on these human, data and computing resources, we need a dedicated Hopkins-wide institute. This entity will establish broad visibility for the effort, provide support for innovative collaborative research and foster a new level of intellectual excitement, communication, and camaraderie. The institute could facilitate “internal sabbaticals” to allow faculty to spend time in other departments; offer seed funds for highly imaginative research equipment and ideas; host visiting scholars, workshops and conferences; assist with grant seeking; and provide both a virtual and a physical home on both campuses for Big Data scientists and emerging enthusiasts.

## What will it take?

New Faculty: These faculty members will be jointly appointed to the institute and any other departments. We propose 10 endowed professorships in the science of Big Data—five based at Homewood and five in East Baltimore, for a total of $25 million. We further propose five named assistant professorships in this science (to be funded for five years each), for a total of $6 million. (After year five, these positions would be covered by the departments.) The combined faculty cost is $31 million.

Endowed Postdoctoral Fellows and Graduate Students: 10 PhD students (at $350,000 per year) and five postdoctoral fellows (at $400,000 per year) dedicated to the science of Big Data will require an endowment of $15 million.

Endowed Data Collections: Establishment of highly valuable and visible data collections requires constant support from a group of four data curation experts (at $400K per year), or an endowment of $8 million. (Physical computational facilities to house the data collections should be made available as part of the high-performance computing infrastructure initiative.)

### Support for the Institute

**Physical space:** Capital costs involve renovating at least 10,000 square feet of space on each campus, for a total of approximately $10 million.

**Venture Fund:** To foster innovation and support start-up projects, the institute requires $2 million per year for the first five years, or a total investment of   
$10 million.

|  |  |  |
| --- | --- | --- |
|  | Quantity/Specifics | New Resources Required |
| **Faculty Positions** | 10 full professorships for Homewood and East Baltimore  5 assistant professorships funded for  5 years | $31 million (after five years, the cost of the assistant professorships would be covered by individual departments). |
| **Endowed Postdoctoral Fellows and Graduate Students** | 10 graduate students at  $35,000 per year each  5 postdoctoral fellows at $80,000 each | $750,000 per year or  $15 million endowment |
| **Endowed Data Collections** | 4 data curation experts at  $100,000 per year each | $400,000 per year or  $8 million endowment |
| **Capital/Facility Needs** | Renovation of at least 10,000 square feet on each campus    Venture fund | $10 million  $2 million per year for the first five years or $10 million total |
| Totals | 10 full professorships  5 assistant professorships  10 graduate students  5 postdoctoral fellows  4 data curation experts  Space Renovation  Venture Fund | $74 million in total  $31 million for faculty  $15 million for graduate students and postdoctoral fellows  $8 million for data curation experts  $10 million for campus renovation  $10 million for venture fund |

Copyright Johns Hopkins University 2011, All Rights Reserved