**IBM** has gifted the Center for Digital Economy Research at the NYU Stern School of Business $200,000 over two years.

CeDER will use the funds to support faculty research, primarily in the area of "Data Risk," where the objective is to find ways to help organizations better manage the risks associated with the use of data, especially as it relates to its customers.

The two projects listed below describe two projects in progress by CeDER faculty.

**IBM/NYU Collaborative Research Project**

**Is Data an Asset or a Liability? A Risk-Based Cost/Benefit Analysis of Data**

## Concept

Data is usually viewed as an asset within an organization. With this view, archiving more data is always a benefit, and the worst case scenario is that it goes unused; in which case, the only cost involved is that of storage which is rapidly decreasing.

Over the last few years, however, it has become apparent that data can be a liability as well. There have been hundreds of incidents of data breaches and other types of occurrences that have resulted in monetary damages, and there is a fear that there could be some serious incidents around the corner.

Professors Dhar and Sundararajan of the Stern School of Business have been considering research questions related to data governance. The professors intend to collaborate with IBM in this endeavor starting in the Fall of 2008. In this research, our objective is to build a model that can be used by decision makers to assess the benefits as well as the costs involved in maintaining data, typically customer data. Our objective is to build a tool for Chief Executive and Information officers that can be used to make decisions about the corporate governance of data instead of such decisions happening "by default" without an explicit considerations of the costs and benefits. We envision this to be some type of "scorecard" that is based on taxonomy or risk factors and their associated importance in specific situations.

## Background

Is there anything unique to data that makes its governance different from that of any other type of asset? Is the problem of data governance a subset of, or similar to, the area of "operational risk?"

Let us start with the second question. Operational risk is concerned with the design of business process to eliminate the risk of adverse outcomes. The processes typically specify the "correct" process, deviations from which are considered risky. While this is essential part of any organization, it ignores or fails to quantify explicitly the risks associated with the possession, ownership and internal sharing of data. Once an organization chooses to *store* any data about its customers, it faces the risk of data loss from breach of its systems, and the liability associated with such breach. The risk increases when it takes *ownership* of customer data, in contrast with simply "licensing" data for specific business uses. The potential for profitable data use increase with ownership, but so does the liability. As the breadth of access to customer data increases, an organization might increase the value it can realize from its customer data, but by increasing the number of points of employee contact, it multiplies the potential misuse and breach opportunities, whether malicious or accidental. The liability associated with this data risk goes beyond customer restitution and legal costs; the reputational damage and revenue loss to an organization whose customer data is breached or lost can be substantial. After all, trust is fundamental to sustained online interaction.

Our proposed research is grounded in a recognition of the fact that data are fundamentally different from other assets, and pose unique risks by virtue to the fact that they tend to be digital. The properties of digital goods have been well documented by many researchers. It is virtually costless to replicate and transport digital data. Once data are "lost", they cannot be recovered, since they are non-rival, existing with identical fidelity in multiple locations. Further, much like intellectual property, the use of data in one physical location does not preclude its ownership or use in others. These properties make them immensely useful to organizations for customer support. For example, the more data a support representative can access and the better they can be analyzed, the higher the potential quality of customer support. At the same time, however, these properties of data enable abuse as well. The broader question or research will contribute towards eventually addressing is whether it possible to make an optimal decision on data access that balances these costs and benefits. Clearly, a judicious assessment of the risks and related sources of liability associated with data retention, ownership and access provision is a necessary first step.

**IBM/NYU Collaborative Research Project**

**Information acquisition for sentiment analysis**

This statement of work covers the joint project described below, between Prem Melville of IBM Research, and Shengli Sheng, Foster Provost and Panos Ipeirotis of NYU, and others (such as students) who are added explicitly later by written communication between the parties.

The IASA project looks at data analysis techniques to help assess and potentially reduce a company's risk due to market perceptions of its products, services, and brands. The modern world of user-generated content on the web provides a new lens through which a company can view the perceptions of customers and other market participants. For example, blog postings can give an early view of negative perceptions of a product. The challenge is that there are millions upon millions of blog postings daily, far more than can be examined manually, even with the help of sophisticated search engines.

Building on the prior work of the researchers from IBM and NYU, this project will study the triage of blog postings to improve (hopefully substantially) the efficiency of sentiment analysis. Specifically, the project will study the combination of automatic statistical text classification models and the manual acquisition of sentiment classifications by (low-cost and potentially noisy) human "labelers." The acquired labels will both provide direct classification of some blog postings, and also will provide labeled training data to improve the statistical models.

A main operational goal is to minimize the cost to achieve a given level of quality of analysis or efficiency of an analytical team, or alternatively, to maximize quality/efficiency for a given cost. A main research goal is to understand how best to manage the outsourcing of low-cost and potentially noisy labelers, to nevertheless achieve operational goals such as high data quality.