Data Ethics

CUSP aims to utilize Big Data to help study and understand urban environments. As a part of this effort, we are planning to build an inclusive data warehouse at CUSP. Our vision for this data warehouse is to hold large quantities of data from multiple sources, including personal (most likely anonymized) data about individuals. But, obtaining, housing, and protecting these data come with many challenges and questions. We hope to answer some of these questions in this working session and converge on a set of principles that will guide our data practices moving forward.

Personal data is a new asset class touching all aspects of society. It is potentially as valuable a resource in the 21st century as heavily traded physical goods like oil have been in the past hundred years. However, throughout history, economic value creation has been linked to the ability to move and trade physical goods. Similarly, "data needs to move to create value. Data sitting alone on a server is like money hidden under a mattress. It is safe and secure, but largely stagnant and underutilized."

But, personal data lacks the trading rules and policy frameworks that exist for widely traded physical assets. As a result, there is little trust among the key stakeholders, - individuals, governments and the private sector, - which could undermine its long-term potential.

In response to surveys, individuals generally say that they want enhanced control over their personal data, increased transparency on how it is used, and some kind of fair value in return. However, their actions are often quite different. While many say they care deeply about privacy, they share information quite widely online. They often sign up for services not knowing how their data will be protected or whether it will be shared. They rarely read the privacy policies of the organizations providing these services, which are usually written in hard-to-comprehend legal language.

Companies, on the other hand, view the data they have captured or created about individuals as theirs. Data is an asset on which they have invested significant resources. They want to leverage the data to create business value, better understand the behavior of their customers and help themselves become more productive. They struggle with how to best protect all the data they now have access to, as well as trying to figure out the different regulations pertaining to its use.

Governments are trying to leverage all this data to stimulate innovation and drive growth, while simultaneously protecting individuals. This is indeed a tall order given the rapid pace of change and the lack of clear rules and overall transparency.

Individuals, companies and governments do not much trust each other regarding the use of personal data. This is not surprising given their different and sometimes conflicting interests. There is continuing debate among these different stakeholders, as well as among different regional jurisdictions on what the best approach might be for allowing data to flow in a trusted manner.

CUSP will be leading a new class of personal data users: non-profit research and academic institutions that will primarily use these data for educational and research purposes. In addition, CUSP will be tapping into novel data streams that have not traditionally been subject to privacy or security concerns.

We need to establish internal policies and principles to ensure proper protection of individual privacy without keeping our researchers and students from "doing science." These policies will lead the research industry in establishing industry-wide ethical standards and principles. Like all good science, we need to always keep in mind the societal benefits of our research.

Below are some of the questions we would like you to think about before the workshop. We expect a lively discussion and debate.

Some general questions to lead to a set of basic principles:

- What are the ethical limits to using personal data for scientific research in a number of disciplines, including social sciences, urban planning, and complex sociotechnical systems?
- What are some security measures that we have to take to ensure proper protection of private dataon our servers?
- One can identify an individual in an anonymized/deidentified dataset by cross correlating that dataset with others. What are the limits to cross correlating datasets?
- How can we best engage the public in shaping our policies and principles regarding these issues?
- What is the best effective enforcement mechanism for such principles? How can we ensure its robustness?

Some questions about specific data sources, how do they fit within the more general framework as addressed by the questions above:

- What level of resolution in synoptic imagery (visual, IR, radar, etc) starts to bring privacy concerns? How is it different than simply snapping a photo of a building or a street, which is done all of the time?
- What are the legalities of tracking and making public the pollution plume with enough resolution that it can be tracked to a given building?
- What are the ethics of environmental genomics on sewage to measure public health? How fine does the level of resolution have to be before privacy becomes an issue?
- What ethical issues do the other potential projects raise?

Further reading:
http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf
http://www3.weforum.org/docs/WEF_IT_RethinkingPersonalData_Report_2012.pdf