Bigdata

- Internet: 5 million terabytes ($5 \times 10^{18}$ bytes) (Google)

- Data generated everyday: 2.5 EXB Bytes ($2.5 \times 10^{18}$ bytes) (IBM)
  - 90% are generated in the last two years

- Personal data increase dramatically
  - Web 2.0, Mobile
  - User transits from data user to data generator
  - Facebook 2.7 billion “likes” and comments per day
Impact of Bigdata on Privacy

- Data
  - Large and diverse collection of personal data

- Data analytics
  - Reveal (sometimes unexpectedly) personal data

- Application scope of law and regulations
Privacy under Bigdata

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.
Privacy under Bigdata

- Where laws can be applied
  - HIPPA
- Who is subject to HIPPA: clinics, hospitals, …, and Target?
Principles of Privacy Protection

- Collection limitation
- Data quality
- Purpose specification
- Use limitation
- Security safeguards
- Openness
- Individual participation
- Accountability
Bigdata Impacts on Privacy

- Data quality
  - Data collected relevant to purposes
  - Accurate, complete and keep up-to-date

- With Bigdata
  - Unpredictable use and analysis of data (for different purpose)
  - Hard to specify purposes before hand
Bigdata Impacts on Privacy

- Individual participation
  - Obtain his/her own data from data controller

- With Bigdata
  - Hard to ensure individual participation with high-volume high-velocity data
  - Hard to know all data controllers
  - Not sure whether applicable to derived data
Promising Privacy Technologies

- Advances in crypto theory and application
- System built with privacy in mind
  - CryptDB
  - Anonymization
- Non-disclosure agreement
- Obstacles
  - Privacy/utility tradeoff, privacy/performance tradeoff, privacy/?? Tradeoff?