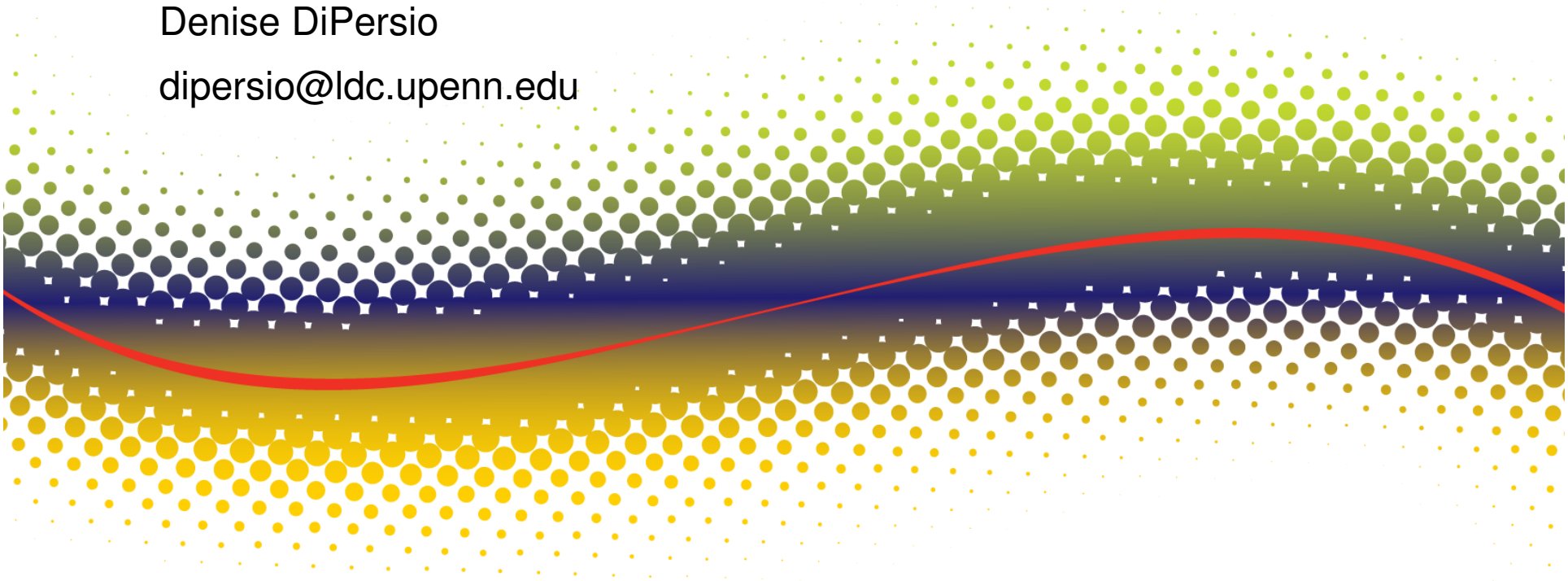
A US Perspective on Selected Legal and Ethical Issues Affecting the Development of Language Resources and Related Technology

Denise DiPersio

dipersio@ldc.upenn.edu



Overview

- Data Protection/Privacy
  - Relevant US laws and regulations
    - Facebook-Cambridge Analytica and its fallout: changes to the US system?
  - Human subjects research – The Common Rule
  - Ethical considerations
    - Threat of re-identification
    - Bias in algorithms
    - The AI Factor- intersection of robots with humans
- Public Sector Information
  - US federal, state and municipal data
- Web Data Collection
  - Copyright, terms of use

- Disclaimer – this is not legal advice

Data Protection/Privacy 1/5

- No single US data protection law
  - Importance of scientific and technical progress
  - Technology provides the solution
    - No human in the loop
- Sector-specific US laws, regulations, agencies
  - Fair Credit Reporting Act of 1970
  - Privacy Act of 1974
  - Cable Communications Policy Act of 1984
  - Electronic Communications Privacy Act of 1986
  - Video Privacy Protection Act of 1988
  - Computer Matching and Privacy Act of 1988
  - Driver's Policy Protection Act of 1994
  - Health Insurance Portability and Accountability Act of 1996
  - Children's Online Privacy Protection Act of 1998
  - Genetic Information Nondiscrimination Act of 2008

Data Protection/Privacy 2/5

- No single definition of personal information, PII
- Laws apply to different actors: government, private sector, both
- The advantage of flexibility?
  - Can react to new developments (with a new law)
  - Remedies have the potential for damages, orders to restrain bad behavior
- No system can police the magnitude of personal data collected and processed
- Many data breaches; many never reported
  - Privacy Rights Clearinghouse: since 2005, over 10 billion records breached in 8000+ data breaches

Data Protection/Privacy 3/5

- Pew Research Center survey (Rainie, 2016; Rainie & Duggan, 2015)
  - Americans are uncertain how their personal data is collected and used
  - But they are willing to share it if they think they will get something in return
    - Social media platform services, good e-commerce deals
  - Over 90% say that consumers have no control over how their personal information is collected and used by companies
  - Most support US government surveillance activities (not targeting them)

Data Protection/Privacy 4/5

- Facebook-Cambridge Analytica
  - App for personality test harvested data from respondents and their friends
  - 87 million affected profiles
  - Data used for political purposes (allegedly)
- Facebook settled claims of deceptive privacy practices in 2011
  - Federal Trade Commission consent decree
- Weaknesses of US framework exposed
  - Terms of FB Open Graph platform did not violate US privacy laws
  - A privacy law with opt-in process could have been helpful
- Users stay with the platform
- No privacy on the web; some information may be “less public”
- FB developers react – following FB terms?
- New privacy terms across platforms, services – not just a FB problem

Data Protection/Privacy 5/5

- Human Subjects Research
  - The Common Rule – designed to prevent abuses in human scientific experiments
  - The Belmont Report
    - Respect for persons: autonomy, consent
    - Beneficence: no harm to subjects
    - Justice: fair procedures
- Institutional Review Boards (IRBs) administer the Common Rule across US universities (ethics boards counterpart)
  - Some bias toward medical research
  - Most language-related studies are minimal risk and subject to expedited review
    - Provide for protection of personal information
    - Right to withdraw from study and to withdraw data
    - Consent to share data (in a corpus)

Segue to Ethics

- Problems with anonymization/de-identification
  - A few data points are enough to re-identify
  - MIT researchers: “need to reform our data protection mechanisms beyond PII and anonymity toward a more quantitative assessment of the likelihood of re-identification” (de Montjoye, et al., 2015)
  - Does not account for voice, facial recognition systems
- Harmful applications of algorithms – bias, discrimination
- The AI Factor – robots and ethics
  - Surveillance, access, social presence (Calo, 2012)
    - Greater penetration into zone of privacy
    - Smart home technology – your data is being collected and used; risk of unauthorized access
    - Humans interact with robot companions; personal solitude diminished
  - Personal data generated subtly, not through user’s deliberate action
  - Rethinking laws and regulations, individual expectations of privacy
- What are the implications of our work?

Public Sector Information

- Open data initiatives across US federal, state, local agencies
  - Data generated in the course of routine activities
  - Goals: transparency, accountability, civic engagement, promoting data use and innovation
  - Resources usually available at no cost and without restriction
- DATA.GOV
  - Federal government site – 200k+ data sets
  - Relevance for HLT/language-related applications
- 50% of US states have data portals and at least 10 have open data policies (Drees & Castro, 2014)
- Municipalities: Philadelphia, New York City
  - New York City Open Data Law – one of the most "robust" worldwide policies
- Consider using this data!

Web Data Collection 1/2

- Web data is an important source for language resources
  - Use may be constrained by copyright, terms of use
- Fair Use Doctrine
  - Alternative to permission from copyright holder
  - An exception to US copyright law
  - Four-factor, fact-intensive analysis
    - the purpose and character of the use, including whether use is commercial or for nonprofit educational purposes
      - Does the use add something new, a different purpose or character?
    - the nature of the copyrighted work
      - Fiction v. nonfiction – fair use more likely applies to the latter
    - the amount of the work used in relation to the whole work
      - Quantity + Quality, Importance
    - the effect of the use on the work's potential market
      - Market impairment v. substitute market

Web Data Collection 2/2

- Important Fair Use Rulings for Language Resources and Infrastructures
  - Fair use covers: full-text searchable databases, student papers used in plagiarism detection database, thumbnail images
  - Transformation is still the touchstone for uses like HLT
    - New use is different from the original, not a substitute; work can be transformative even if it is unchanged
    - Using whole work is generally allowed
    - No market threat
- Browser wrap terms of use
  - Can negate fair use
  - Regulate how material can be used, shared, modified; third party data problem
  - Be aware

Questions, Comments ?

- Questions, comments?
- Thank you!