Educational Big Data and the Digital Learner

*Privacy and Access Considerations*

Una-May O’Reilly
ALFA Group
Computer Science & AI Lab
MIT
Enabling technology

- MOOCdb project targets a Data Science Commons concept
- Includes multiple projects and open-source software that transform data and enable MOOC-related data science
Sharing MOOC data science software

The moocdb project

The hub becomes software, NOT DATA!
This is a mooc data science commons!
moocdb.csail.mit.edu
moocViz
featureFactory
labelMe Text

- Mooc data will be spatially distributed
- Mooc data access will be controlled locally
  - Chances of a big Data hub are low…
  - Maybe meta-data can be shared
Another Strategy

- MoocDB concept of software sharing shifts privacy into checking software outputs
- Return to the goal of openly sharing data
  - Must return to the issues of privacy protection
    » Risk of re-identification through linking
- In this context:
  - Technology is integral
    » But it’s only one element of the story…
  - Start by bringing clarity to the complexity of the system by identifying the stakeholders
The Stakeholders

- Analysts
- Subject/provider Controller?
- Controllers
- Instructor
- Institution of instructor
- Platform provider
- Coursera
- edX
- Doing Data Science

Who is likely to stop out?

Community detection
Weekly topic analysis
Shared data model
Shared analytics
Privacy as a service
Open access
Privacy Protecting Policy
Differential privacy
Crowd Sourcing
Machine Learning
Sharing Data and Protecting Privacy

The Present

Two extremes

1. Rich, raw data – highly vetted
2. Used, simple data - open
One Possible Future

- Sharing via gradualism: a consistent protection level provided while
  - Access increases
  - protection mechanism potentially changes:
    » This depends on nature of data being shared
      - We are interested in sharing transformed, advanced research-ready data
Data Privacy and Big Data

- Study from policy, access control and technology vantages in global, broad contexts
Closer to home: USA

• BigData@CSAIL initiative
  – Workshops on data privacy -&gt; Report
  – Working group report, sub-group on online learning

• Helped me to formulate a working context and its requirements for progress
A Working Context

Requirements

- Identify the data we want to share
- Assess: technology, control and policy practices, state of art
- Integrative Roadmap!!
Sharing transformed data that enables collaboration and cross-institute research will enable a MOOCDB data science commons concept.
Assessment

• Technology
  – Gap between practitioners’ needs and technology maturity

• Policy
  – Asilomar

• Controllers
  – MIT Registrar “challenge”
Roadmap

- Design in progress!
- Started a modest investigation thread
  - Theory/Social Science Sharing/Privacy-protected data
Sanitizing the Data Directly

### Post-Randomization

#### Perturb $(\gamma, \varepsilon)$

<table>
<thead>
<tr>
<th>Mid-Term Grade</th>
<th>Pass/ Fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>F</td>
</tr>
<tr>
<td>B</td>
<td>P</td>
</tr>
<tr>
<td>F</td>
<td>P</td>
</tr>
<tr>
<td>A</td>
<td>P</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>B</td>
<td>P</td>
</tr>
<tr>
<td>A</td>
<td>F</td>
</tr>
</tbody>
</table>

**Truthful data**

**Protected Data**

### $\varepsilon$-DP guarantee

Mid-Term Grade | Pass/ Fail |
----------------|------------|
| C              | F          |
| B              | P          |
| C              | P          |
| B              | F          |
| B              | P          |
| A              | F          |