Tiered Access: A Framework and Practical Approach

Julia Lane, New York University
## Agency Actions

<table>
<thead>
<tr>
<th>Action</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Identify Data Needs to Answer Priority Agency Questions</td>
<td>✔️</td>
</tr>
<tr>
<td>2. Constitute a Diverse Data Governance Body</td>
<td>✔️</td>
</tr>
<tr>
<td>3. Assess Data and Related Infrastructure Maturity</td>
<td>✔️</td>
</tr>
<tr>
<td>4. Identify Opportunities to Increase Staff Data Skills</td>
<td>✔️</td>
</tr>
<tr>
<td>5. Identify Priority Data Assets for Agency Open Data Plans</td>
<td>✔️</td>
</tr>
<tr>
<td>6. Publish and Update Data Inventories</td>
<td>✔️</td>
</tr>
<tr>
<td>Executive Order on Maintaining American Leadership in Artificial Intelligence</td>
<td>✔️</td>
</tr>
<tr>
<td>Improving Implementation of the Information Quality Act (M-19-15)</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Utility and Risk must be simultaneously measured and compared

Or else we end up with a corner solution
Both Risk and Utility are moving targets

Risk: Increasing because of technology; decreasing because of technology
Utility: Can design access to increase utility
Fixing ideas: The New Zealand context

Budget 2018

Presentation: Sir Bill English, former NZ PM
What do we mean by integrated data?
Understanding risk
The risk profile of this cohort reveals a long tail of increased risk.

48,077 low-risk students have an average risk of 27%
6,990 high-risk students have an average risk of 76%

The hypothetical cost effectiveness of interventions increases when they are delivered to the students who need them most.
Aggregate liability valuation

**Total liability**
- 72.2 billion

**Liability reduction attributable to system reform**
- $13.7 billion

- Liability decrease related to experience, $billions
  - 2012: -5
  - 2013: -4
  - 2014: -3
  - 2015: -2
  - 2016: -1

- Increase due to policy package to reduce child poverty – because of the method we employ, this can be isolated
- Updated model assumptions
- Client numbers
Five Safes: Four of them mitigate risk

1. Safe People  
   Training; Clearance; indemnification
2. Safe Settings  
   Physical protection (new options)
3. Safe Data  
   Protect core elements (hashing, suppression)
4. Safe Outputs  
   SDL techniques (reidentification probability)
Five Safes: Three of them refer to utility tradeoff

1. Safe Projects
   Utility measure
2. Safe People
   Trust measure
3. Safe Outputs
   Risk vs. Harm
Practical Approach

Training

Data user

Data producer

Metadata

Usage

Feedback

Rich Context
Links metadata, code, tools, publications

Data Stewardship
Approval workflow, monitoring, reporting

Collaboration
Interactive chat and code sharing

Workspace and tools

Python, Jupyter, R

Data in cloud
Alternative: local servers

Security
FedRAMP security certified

Access
Workflows
Monitoring

Reporting
Risk Mitigation
Utility/quality incentives

(1) influential (number of policy documents citing the work)
(2) useful (number of uses)
(3) reproducible (number of methods reported)
(4) accessible (number of distinct users) and
(5) objective (quality of scientific publications).
Practical

Automate dataset search and discovery

Publish results in agency data inventory
Tie access to contribution
Practical next steps for implementation

• Identify and post utility measures
• Incentivize contributions to data quality
• Focus on protecting outputs not inputs
Comments welcome

• Julia Lane Julia.lane@nyu.edu
• ML Competition github.com/Coleridge-Initiative/rclc
• Github https://github.com/Coleridge-Initiative

Readings

• “Human-in-the-loop AI for scholarly infrastructure”
• “New initiative to help with discovery of dataset use in scholarly work”, Christian Zimmerman