Repurposing Administrative Data for Statistical Purposes

Aaron D. Schroeder, Ph.D.
Senior Data Research Scientist
Social & Decision Analytics Lab
Biocomplexity Institute of Virginia Tech
Every Repurposing Is a New “Investigation”

• Locating the Data Repurposing Discussion
• Overview of the SDAL “Investigative Process” for Repurposing Data
• Recommendations for Research-Enabling Standards for Integrated Administrative Data Systems to Aid Future Investigations
Data Repurposing
Locating the Discussion

Secure/Privacy-Protecting Linkage

FISMA

Dataset

DFARS

Dataset

HIPPA

Dataset

DFARS

Dataset

HIPPA

Dataset

FERPA

Dataset

State Law

Specific Question

DATA GOVERNANCE
SDAL
Data Science Processes & Platforms for Evidence-Based Policy

- Data Analytics Process
  - **Data Fitness Analysis**
  - Data Analysis & Hypothesis Testing
  - Creation of Community Data Tools

- Data Fitness Analysis
  - **Profiling**
  - Preparation
  - Linkage
  - Exploration & Assessment
Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**

Structure, Quality, Metadata & Provenance

---

**Missing Variables**
values in column headers instead of variable names
- e.g. Value-ranges being used as column headers (0-9|10-19|20-29|...)

**Combined Variables**
more than one variable represented in a attribute (column) value
- e.g. An attribute combining gender and age (m25, f32,...)

**Multiple Observation Directions**
variables in both columns and rows
- e.g. A dataset with an element(column) for each day of the month (horizontal) and an element(column) for 'month' (vertical)
  - note. the messiest and can be dealt with multiple ways according to the needs of the specific analysis

**Combined Observation Unit Types**
more than one observation unit type per table
- e.g. A table containing both individual demographic data and a periodic measurement like weekly attendance where demographic data and weekly attendance are separate observational units and need to be in separate datasets.

**Divided Observation Unit Type**
observation unit type is split among multiple tables
- e.g. Individual demographic information split among several datasets; for example, separate tables for gender, ethnicity, and surname.
This is a single record with 128 fields all keyed to the variable “List Number”.

Structured this way, it is not possible to analyze property changes over time.

Pulling out a definitive list of unique properties using “Parcel ID” seems like a possibility.

However, “Parcel ID” is left blank in over 7% of entries – extra work required – perhaps including address, but address is not standardized.
Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**

*Structure, Quality, Metadata & Provenance*

**Combined Observation Unit Types**

Ideal Restructuring of MLS Data

- Property Characteristics
- Property Sales Information
- Property Tax Information
Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**

**Structure, Quality, Metadata & Provenance**

**Divided Observation Unit Types**

<table>
<thead>
<tr>
<th>gender1</th>
<th>id</th>
<th>gender2</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>43XXX13</td>
<td>M</td>
</tr>
<tr>
<td>F</td>
<td>43XXX14</td>
<td>M</td>
</tr>
<tr>
<td>M</td>
<td>76XXX46</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>74XXX98</td>
<td>M</td>
</tr>
<tr>
<td>F</td>
<td>76XXX23</td>
<td>M</td>
</tr>
<tr>
<td>F</td>
<td>77XXX40</td>
<td>M</td>
</tr>
<tr>
<td>M</td>
<td>74XXX98</td>
<td>F</td>
</tr>
<tr>
<td>M</td>
<td>78XXX73</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>78XXX74</td>
<td>M</td>
</tr>
<tr>
<td>M</td>
<td>77XXX84</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>79XXX87</td>
<td>M</td>
</tr>
<tr>
<td>M</td>
<td>71XXX95</td>
<td>F</td>
</tr>
<tr>
<td>M</td>
<td>21XXX96</td>
<td>F</td>
</tr>
<tr>
<td>M</td>
<td>71XXX54</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>71XXX55</td>
<td>M</td>
</tr>
<tr>
<td>F</td>
<td>77XXX86</td>
<td>M</td>
</tr>
<tr>
<td>F</td>
<td>80XXX24</td>
<td>M</td>
</tr>
<tr>
<td>M</td>
<td>76XXX79</td>
<td>F</td>
</tr>
</tbody>
</table>

**NC Student Data**

Demographics Recorded in Multiple Tables

- Actual 2011 data from different tables linked via unique ID
- Many more tables with apparently separately collected demographics
- Derivation of Demographic Truth is now Probabilistic
Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**
Structure, **Quality**, Metadata & Provenance

### Completeness
percentage of elements properly populated
- e.g. Testing for NULLS and empty strings where not appropriate

### Value Validity
percentage of elements whose attributes possess meaningful values
- e.g. A comparison constraint like \{male; female\} or an interval constraint like \[0,110\]

### Consistency
a measure of the degree to which two or more data attributes satisfy a well-defined dependency constraint — relationship validation
- e.g. Zip-code — state consistency or gender — pregnancy consistency

### Uniqueness
the number of unique values taken by an attribute, or a combination of attributes in a dataset
- e.g. Frequency distribution of an element note. The more homogeneous the data values of an element, the less useful the element is for analysis

### Duplication
a measure of the degree of replication of distinct observations per observation unit type
- e.g. Greater than 1 registration per student per official reporting period note. Duplication occurs as a result of choice of level of aggregation
Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**
Structure, **Quality**, Metadata & Provenance
**Completeness**

- Seems straight-forward -- Nope
- A set of data is complete with respect to a *given purpose* if the set contains all the relevant data for that purpose
- A common measure is the proportion of data that has values to the proportion that “should” have values.
  - Completeness is *application-specific*
  - Incorrect to simply measure number of missing field values in a record without considering which fields are necessary
    - MLS Data had MANY highly incomplete fields that were not necessary for the study at hand
- Data that are missing can be categorized as:
  - record fields not containing data
  - records not containing necessary fields
  - datasets not containing the requisite records
Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling
Structure, Quality, Metadata & Provenance
Value Validity

- Data elements with proper values have **value validity**
- The percentage of data elements whose attributes possess values within the range expected for a legitimate entry is a measure of value validity
- Checking for value validity generally comes in the form of straight-forward domain constraint rules
  - How many entries contain non-valid values for a non-empty text field representing gender?
    - `<count gender where gender is not (male, female)>`
  - How many entries contain non-valid values for a non-empty integer field representing age?
    - `<count age where age is not between [0, 110]>`
Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**
Structure, **Quality**, Metadata & Provenance
**Value Validity**

Pulled from current James City County MLS Data

<table>
<thead>
<tr>
<th>zip_code</th>
<th>area</th>
<th>subdivision</th>
<th>neighborhood</th>
<th>zoning</th>
<th>parcel_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>23185</td>
<td>JCC</td>
<td>Governors Land</td>
<td>River Reach</td>
<td>R-4</td>
<td>4511000022</td>
</tr>
<tr>
<td>23188</td>
<td>JCC</td>
<td>Wellington</td>
<td></td>
<td>RESIDENT</td>
<td>1330800178</td>
</tr>
<tr>
<td>23188</td>
<td>JCC</td>
<td>Powhatan Secondary</td>
<td></td>
<td>RES</td>
<td>3741600013</td>
</tr>
<tr>
<td>23185</td>
<td>JCC</td>
<td>Kingsmill</td>
<td>Padgetts Ordinary</td>
<td>R 4</td>
<td>5041100213</td>
</tr>
<tr>
<td>23185</td>
<td>JCC</td>
<td>Pointe @ Jamestown</td>
<td></td>
<td>RES</td>
<td>4640600108</td>
</tr>
<tr>
<td>23185</td>
<td>JCC</td>
<td>Paddock Green</td>
<td>Paddock Green</td>
<td>R1</td>
<td></td>
</tr>
</tbody>
</table>

Comparison constraint: **zoning 2015, James City County** = \{A-1, R-1, R-2, R-3, R-4, R-5, R-6, R-7, R-8, LB, B-1, M-1, M-2, RT, PUD, MU, PL, EO\}

- During Data Profiling issues are described, not “fixed”
- The appropriate fix depends upon the needs of the research
- It may be appropriate to simply normalize all zoning entries to the five major categories of zoning: Residential, Mixed Residential-Commercial, Commercial, Industrial, and Special
Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling
Structure, Quality, Metadata & Provenance
Consistency

- The Degree to Which Two or More Attributes Satisfy a Dependency Constraint
- Simple example
  - Location disagreements like zip and state (Record-Level)
- More complex example (Longitudinal)
  - Consistency with locally derived “truth”
  - VDOE Student Record, no definitive list of student demographics
  - Truth must be derived from multiple observations
    - Student Record has multiple observations per school year
    - Query here shows disagreement on gender for some of the observations when Student Record is matched to itself
      - select count(distinct a.internal_id) from vdoe.student_record a
      join vdoe.student_record b on a.internal_id = b.internal_id
      and a.gender <> b.gender
      - 16,310 / 2,346,058 individuals have more than one value for gender
Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**
Structure, Quality, **Metadata & Provenance**

---

**Observation Unit Definition**
Datasets (tables) without definition and/or non-meaningful/confusing naming

**Observation Unit Attributes Definition**
Attributes (columns) without definition and/or non-meaningful/confusing naming

**Semantic Confusion**
Attributes with the same name but different definitions
- e.g. An attribute named “Grade” can refer to both a ‘score’ for a test or the ‘level/year’

**Multiple Attribute Names**
Attributes with different names but the same definition
- e.g. Attributes name “Grade” and “Year” both referring to ‘level/year’ of schooling

**Inconsistent Attribute Formats**
Attributes of the same type that are formatted differently
- e.g. Most commonly an issue when dealing with dates and times

**Data Process History**
Attributes collected at different locations, with different tools

**System of Origin**
Where was this data originally collected?

**Intermediate Storage Systems**
Chain of Custody

**Contact Information**
Who can I contact with my questions?

**Transformation**
What happened to the data since collection and why?

---

Getting this stuff in order is a BIG part of Data Repurposing!
Research-Enabling Standards for Integrated Administrative Data Systems

• Metadata
  – Minimum: Table and Field Definitions, Field Valid Values and Definitions
  – Extra: Valid Value Timing and Relationship Data
• Provenance
  – Minimum: System of Origin/Collection and Contact
  – Extra: Intermediate Storage Systems and Contacts
  – Extra +: Transformations Used and Reasons Why
• Just starting with the Minimums will accomplish quite a bit