How and why do scientists reuse others’ data to produce new knowledge?
Background, Foreground, and Beyond

Christine L. Borgman, PhD
Distinguished Research Professor
Center for Knowledge Infrastructures
University of California, Los Angeles
http://christineborgman.info
https://knowledgeinfrastructures.gseis.ucla.edu
@scitechprof

Irene V. Pasquetto, PhD
Research Affiliate
UCLA Center for Knowledge Infrastructures
Fringe Event
Cochrane Colloquium
Edinburgh, 15 September 2018
Data sharing policies

- Research Councils of the UK
- European Union
- U.S. Federal research policy
- Australian Research Council
- Individual countries, funding agencies, journals, universities
Data Stewardship: The Ideal

Findable Accessible Interoperable Reuseable

What is “data reuse”?

Background and Foreground Reuses of data at CENS

Images: CKI and NSF archives
The DataFace Consortium for Data Sharing

GOAL:
Collect and release high-throughput “hypothesis free” biomedical data related to the genetics of facial formation and development in humans and animals.

DATA TYPES:
Whole genome sequences, gene expression data from ChiP-seq, RNA-seq, and microarrays, genotypes and phenotypes from GWAS studies, etc.

DOMAINS:
Developmental biologists, evolutionary experts, human geneticists, computational biologists, surgeons, etc.
Background Reuse at DataFace: Comparison, control, verification. (I)

UCSC Genome Browser – Search example (CAPZB gene)

UCSC Genome Browser – Zoom IN

Irene Pasquetto, DataFace Study, 2018
Background Reuse at DataFace: Comparison, control, verification. (II)
Foreground Reuse at DataFace: Data Analysis

Aligner software pairs "reads" using reference assemble genome

Data processing tool summarizes BAM information to compute likelihood of each possible genome

In-house script takes the ratio of mutant and allele frequencies to find the highest peak

R studio calculates relative frequency and generate plotting graphs

Annotation tool predicts consequences of gene function

Variants are annotated by gene names, variant impact, and type of variant

"RAW" DATA

Pipeline

RESULTS
Having access to the contact information of those who collected the data increases rates of foreground reuse.

Image source: https://www.123rf.com
The “Data Creator Advantage”

• Creator has most current annotations about the dataset
• Creator has most specialized knowledge of relevant literature
• Creator may have software pipelines locally customized for the dataset

Image source: https://www.siteminder.com
<table>
<thead>
<tr>
<th></th>
<th>BACKGROUND Reuse of Data</th>
<th>FOREGROUND Reuse of data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal of reuse</strong></td>
<td>“Ground truthing:” calibrate, compare, confirm</td>
<td>Analysis: identify patterns, correlations, causal relationships</td>
</tr>
<tr>
<td><strong>Example of reuse</strong></td>
<td>Instrument calibration, sequence annotation, review summary-level data</td>
<td>Meta-analyses, novel statistical analyses</td>
</tr>
<tr>
<td><strong>Frequency of reuse</strong></td>
<td>Frequent, routine practice</td>
<td>Rare, emergent practice</td>
</tr>
</tbody>
</table>

| Goal of reuse | “Ground truthing:” calibrate, compare, confirm | Analyses: identify patterns, correlations, causal relationships. |
| Example of reuse | Instrument calibration, sequence annotation, review of summary-level data | Meta-analyses, novel statistical analyses. |
| Frequency of reuse | Frequent - routine practice | Rare - emergent practice |

Questions: Trusted Evidence?

• When to reuse open data independently?
• When to collaborate with data creators?
• What information is needed, when, to trust evidence?
Questions: Informed decisions?

• What do you need to know about the data to inform decisions?
• When are data sufficient for decision making?
• When is further information about about data needed?
• How should data sharing and reuse be governed?
Questions: Better health?

- Where should community invest in data sharing and reuse?
- How should data resources be governed?
- Who should be responsible for sustaining access to health data?
- What are reasonable licensing agreements?
- What are appropriate funding models for data resources?
Acknowledgements

Christine Borgman
Bernie Boscoe
Peter Darch
Milena Golshan
Irene Pasquetto
Michael Scroggins
Cheryl Thompson
Morgan Wofford