October 26, 2015

Creating, Collaborating, and Celebrating the Diversity of Research Data

Christine L Borgman, University of California, Los Angeles

Available at: https://works.bepress.com/borgman/379/
Creating, Collaborating, and Celebrating the Diversity of Research Data

Christine L. Borgman
Distinguished Professor and Presidential Chair in Information Studies
University of California, Los Angeles
@SciTechProf

Seminar Presentation
Graduate School of Library, Information, and Media Studies
University of Tsukuba, Japan
October 26, 2015
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOUNT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.
Big Data, Little Data, Open Data, and Libraries

Christine L. Borgman
Professor and Presidential Chair in Information Studies
University of California, Los Angeles

University of Göttingen
Inaugural Göttingen Lecture on Library Futures
23 March 2015
Data <-> Publications

Publications are arguments made by authors, and data are the evidence used to support the arguments.

Theme issue ‘Celebrating 350 years of Philosophical Transactions: life sciences papers’ compiled and edited by Linda Partridge
19 April 2015; volume 370, issue 1666
Open access policies

• Australian Research Council
  – Code for the Responsible Conduct of Research
  – Data management plans
• National Science Foundation
  – Data sharing requirements
  – Data management plans
• U.S. Federal policy
  – Open access to publications
  – Open access to data
• European Union
  – European Open Data Challenge
  – OpenAIRE
• Research Councils of the UK
  – Open access publishing
  – Provisions for access to data
Big Data, Little Data, No Data: Scholarship in the Networked World

- Part I: Data and Scholarship
  - Ch 1: Provocations
  - Ch 2: What Are Data?
  - Ch 3: Data Scholarship
  - Ch 4: Data Diversity

- Part II: Case Studies in Data Scholarship
  - Ch 5: Data Scholarship in the Sciences
  - Ch 6: Data Scholarship in the Social Sciences
  - Ch 7: Data Scholarship in the Humanities

- Part III: Data Policy and Practice
  - Ch 8: Releasing, Sharing, and Reusing Data
  - Ch 9: Credit, Attribution, and Discovery
  - Ch 10: What to Keep and Why
Celebrating the diversity of data

• Defining data
• Creating data
• Collaborating with data
• Consolidating data value
Data
Big Data
Long tail of data
Open Data: Free

- A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike

State Library and Archives of Florida, 1922. Flickr commons photo

Open Data Commons. (2013).
Open Data: Useful

- Openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability, and sustainability.

What are data?

Marie Curie’s notebook

Pisa Griffin

http://www.census.gov/population/cen2000/map02.gif

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php

http://hudsonalpha.org

http://aip.org

http://ncl.ucar.edu
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

Creating research data

Sloan Digital Sky Survey Telescope, Apache Point, New Mexico

http://astro.uchicago.edu/~frieman/SDSS-telescope-photos/

Sensor networks

http://enl.usc.edu/~jpaek/data/cyclops/bird_nest_2008/figures/nestbox2.jpg
Tools for Astronomical Big Data
Tucson, Arizona, March 9-11, 2015

Program

Monday, March 9, 2015

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00-9:00</td>
<td>Registration/Continental Breakfast</td>
</tr>
<tr>
<td>9:00-9:15</td>
<td>Introductory Remarks</td>
</tr>
<tr>
<td>9:45-10:15</td>
<td>Carlos Scheidelger (University of Arizona) How do you look at a billion data points? Exploratory Visualization for Big Data</td>
</tr>
<tr>
<td>10:15-11:00</td>
<td>Break</td>
</tr>
<tr>
<td>11:00-11:20</td>
<td>Joshua Peek (STScI) Machine Vision Methods for the Diffuse Universe</td>
</tr>
<tr>
<td>11:50-12:00</td>
<td>Lunch</td>
</tr>
<tr>
<td>12:00-12:30</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:00-13:30</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30-14:00</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:00-14:30</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:30-15:00</td>
<td>Lunch</td>
</tr>
<tr>
<td>15:00-15:30</td>
<td>Lunch</td>
</tr>
</tbody>
</table>
Research process

• Models and theories
• Research questions
• Methods
  – Practices
  – Data sources
  – Software
  – Instruments
  – Infrastructure
  – Domain expertise
Random walk

http://www2.ess.ucla.edu/~jewitt/oort2-random.html
Collaborating with data
Big Science <-> Little Science

• Large instruments
• High cost
• Long duration
• Many collaborators
• Distributed work
• Domain expertise

• Small instruments
• Low cost
• Short duration
• Small teams
• Local work
• Domain expertise

Sloan Digital Sky Survey

Sensor networks for science
Telescope for the Sloan Digital Sky Survey, Apache Point, New Mexico
A role for self-gravity at multiple length scales in the process of star formation

Alyssa A. Goodman 1,2, Erik W. Rosolowsky 1,2, Michelle A. Borkin 1, Jonathan B. Foster 1, Michael Halle 1, Jens Kauffmann 1,2 & Jaime E. Pineda 1

Self-gravity plays a decisive role in the final stages of star formation, where dense cores (size ~0.1 parsecs) inside molecular clouds collapse to form star-plus-disk systems. But self-gravity's role at earlier times (and on larger length scales, such as ~1 parsec) is unclear; some molecular cloud simulations that do not include self-gravity suggest that 'turbulent fragmentation' alone is sufficient to create a mass distribution of dense cores that resembles, and sets, the stellar initial mass function. Here we report a 'dendrograms' (hierarchical tree-diagram) analysis that reveals that self-gravity plays a significant role over the full range of possible scales traced by 13CO observations in the L1448 molecular cloud, but not everywhere in the observed region. In particular, more than 90 per cent of the compact 'pre-stellar cores' traced by peaks of dust emission are projected on the sky within one of the dendrogram's self-gravitating 'leaves'. As these peaks mark the locations of already-forming stars, or of those probably about to form, a self-gravitating cocoon seems a critical condition for their existence. Turbulent fragmentation simulations without self-gravity—even of unmagnetized isothermal material—can yield mass and velocity power spectra very similar to what is observed in clouds like L1448. But a dendrogram of such a simulation shows that nearly all the gas in it (much more than in the observations) appears to be self-gravitating. A potentially significant role for gravity in non-self-gravitating simulations suggests inconsistency in simulation assumptions and output, and that it is necessary to include self-gravity in any realistic simulation of the star-formation process on subparsec scales.

Spectral line mapping shows whole molecular clouds (typically tens to hundreds of parsecs across, and surrounded by atomic gas) to be marginally self-gravitating. When attempts are made to further break down clouds into pieces using 'segmentation' routines, some self-gravitating structures are always found on whatever scale is sampled. But no observational study to date has successfully used one spectral line data cube to study how the role of self-gravity varies as a function of scale and conditions, within an individual region.

Most past structure identification in molecular clouds has been explicitly non-hierarchical, which makes difficult the quantification of physical conditions on multiple scales using a single data set. Consider, for example, the often-used algorithm CLUMPfind. In three-dimensional (3D) spectral-line data cubes, CLUMPfind operates as a watershed segmentation algorithm, identifying local maxima in the position-position-velocity (p-p-v) cube and assigning nearby emission to each local maximum. Figure 1 gives a two-dimensional (2D) view of L1448, our sample star-forming region, and Fig. 2 includes a CLUMPfind decomposition of it based on 13CO observations. As with any algorithm that does not offer hierarchically nested or overlapping features as an option, significant emission found at prominent clumps is typically either appended to the nearest clump or turned into a small, usually 'pathological', feature needed to encompass all the emission being modelled. When applied to molecular-line

Figure 1 | Near-infrared image of the L1448 star-forming region with contours of molecular emission overlaid. The channels of the colour image correspond to the near-infrared bands J (blue), H (green) and K (red), and the contours of integrated intensity are from 13CO (1-0) emission. Integrated intensity is approximately linear but not quite linear (see Supplementary Information), related to column density, and it gives a view of 'all' of the molecular gas along line of sight, regardless of distance or velocity. The region within the yellow box immediately surrounding the protostars has been imaged more deeply in the near-infrared (using Calar Alto) than the remainder of the box, to emphasize the central regions as well as the scattered starlight known as 'circumstellar' and dust (which appear orange in this colour scheme). The four yellow line labels indicate regions containing self-gravitating dense gas, as identified by the dendrogram analysis, and the labels identify areas that are best shown in Fig. 2a. Asterisks show the locations of the four most prominent embedded young stars or compact stellar systems in the region (see Supplementary Table 1), and yellow circles show the kilometre-dust emission peaks identified as star-forming or 'pre-stellar cores'.

Figure 3 | Schematic illustration of the dendrogram process. Shown is the dendrogram in 3D, where each leaf represents a self-gravitating core, with each node and its associated edges connecting to other nodes, thus forming a tree structure.
Center for Embedded Networked Sensing

- NSF Science & Tech Ctr, 2002-2012
- 5 universities, plus partners
- 300 members
- Computer science and engineering
- Science application areas

Slide by Jason Fisher, UC-Merced, Center for Embedded Networked Sensing (CENS)
Science <-> Data

Engineering researcher: "Temperature is temperature."

Biologist: "There are hundreds of ways to measure temperature. ‘The temperature is 98’ is low-value compared to, ‘the temperature of the surface, measured by the infrared thermopile, model number XYZ, is 98.’ That means it is measuring a proxy for a temperature, rather than being in contact with a probe, and it is measuring from a distance. The accuracy is plus or minus .05 of a degree. I [also] want to know that it was taken outside versus inside a controlled environment, how long it had been in place, and the last time it was calibrated, which might tell me whether it has drifted."
Center for Dark Energy Biosphere Investigations

International Ocean Discovery Program
lodp.tamu.org

- NSF Science & Tech Ctr, 2010-2020
- 20 universities, plus partners (35 institutions)
- 90 scientists
- Biological sciences
- Physical sciences

Repository for seafloor cores. Photo: Peter Darch
Self-descriptions C-DEBI scientists

Aquatic chemistry
Aquatic microbial ecology
Astrobiology
Biochemistry
Bioenergetics
Biogeochemistry
Biogeography
Bioinformatics
Biology
Chemical ecology
Chemical oceanography
Deep sea biogeochemistry
Deep sea microbiology
Ecology
Ecophysiology
Environmental chemistry
Environmental microbiology
Genomics
Geobiology
Geochemistry
Geobiology
Geology
Geochemistry
Geophysic
Microbial oceanography
Microbial physiology
Microbiology
Mineralogy
Molecular biogeochemistry
Molecular biology
Molecular microbiology
Molecular physiology
Paleoceanography
Paleoclimatology
Paleogeomicrobiology
Petrology
Physiology
Plant biochemistry
Sedimentary biogeochemistry
Sedimentary geochemistry
Sedimentology

Peter Darch, UCLA
Conservatives report, but liberals display, greater happiness: Wojcik et al - Behavioral Happiness - Study 1 data

Principal Investigator(s): Wojcik, Sean; Hovasapian, Arpine; Graham, Jesse; Motyl, Matt; Ditto, Peter;

**sesladder:country (Subjective SES)**

<table>
<thead>
<tr>
<th>Category</th>
<th>N=1433</th>
<th>Valid N=1431</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: Lowest</td>
<td></td>
<td>408</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>56</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>-122</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>-112</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>-180</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>219</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>-214</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>-70</td>
</tr>
<tr>
<td>10: Highest</td>
<td></td>
<td>23</td>
</tr>
</tbody>
</table>

Invalid Cases
Invalid N=2
System Missing

Citation: Wojcik, Sean; Hovasapian, Arpine; Graham, Jesse; Motyl, Matt; Ditto, Peter. Conservatives report, but liberals display, greater happiness: Wojcik et al - Behavioral Happiness - Study 1 data. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-02-21. http://doi.org/10.3886/E26998V1

Persistent URL: http://doi.org/10.3886/E26998V1
The Pisa Griffin Project

The aim of this project is to perform a comparative study of three artworks (bronze casts of Islamic provenance), to discover evidence of similarities and to get new insight on their origin.

Probably produced within the Islamic Mediterranean in the eleventh century, the Griffin has incised on its body a long inscription in Arabic expressing good wishes. Captured by the Pisans, it underwent an extraordinary transformation: for centuries it was a terrifying, sound-producing guardian figure on top of the roof of Pisa Cathedral. The present project is focused on the Griffin but also includes alongside it other bronze animal sculptures such as a Lion and a Falcon. It is hoped that the interdisciplinary study of the Griffin will shed light on the significance of such objects in a global Mediterranean culture.

Videos

The Pisa Griffin: an introduction

http://vcg.isti.cnr.it/griffin/
6 BEASTS THAT ROARED: THE PISA Griffin AND THE NEW YORK LION

Anna Contadini, Richard Camber and Peter Northover

The Pisa Griffin
Anna Contadini

My interest in the Pisa Griffin (pl. 6.1) goes back to my childhood, when my parents took me to visit Pisa for the first time. I am still as impressed by the beast as I was then, but I am now equally intrigued by the mystery that further substantiates, to be learnt about it. One might call a gynus through the opening a discovered that it had Griffin and attached with slightly overriden the animal (pl. 6.3).

The Composition of the Lion and the Griffin

Peter Northover

The discussion of the compositions of the two sculptures will be made in the order in which the analyses were carried out, that is with the Lion first, followed by the Griffin. All the analyses have been made by electron probe microanalysis (EPMA) using wavelength dispersive spectrometry; this method has been well standard against other current techniques so the results were broadly comparable with those from other laboratories.
Precondition:

Researchers share data
Lack of incentives to share data

- Labor to document data
- Benefits to unknown others
- Competition
- Control
- Confidentiality...
Lack of incentives to reuse data

• Identify useful data
  – Documentation
  – Interpretation
  – Software

• Cleaning

• Trust

• Credit

• Licensing...

Consolidating value in data

July 19, 1922. State Library and Archives of Florida. Flickr commons

Page 105 of "The Street railway journal" (1884); Flickr Commons
Better Data: Better Research

Why manage data?

- Preserve the integrity of the research
- Allow data to be made available for others to use
- Assist researchers to reduce the risk of data loss
- Secure continued access to the value in data

Why connect data?

- Interlink data to people to projects to publications
- Improve the discoverability of data
- Tie data to research achievements
- Provide richer context for data value

Why make data discoverable?

- Enable the demonstration of research excellence
- Allow researchers to build upon existing data, instead of recreating it
- Foster innovation
- Provide the ability to solve big problems across discipline boundaries

Why reuse data?

- Verification of research claims
- New discoveries from existing data
- Integration of sets of data for new analysis
- Re-analysis of expensive, rare or unrepeatable investigations
- Reduction of duplicated effort
Ways to pool data

• Centralized data production
  – Top down investments in data
  – Pooled data resources for the community

• Decentralized data production
  – Bottom up investments in data
  – Local data resources pooled later
The Sloan Digital Sky Survey has created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects. Learn and explore all phases and surveys—past, present, and future—of the SDSS.
ICT Diffusion and Distribution Dataset, 1990-2007 (ICPSR 23562)

Principal Investigator(s): Howard, Philip N., University of Washington; Busch, Laura, University of Washington; Cohen, Spencer, University of Washington

Summary:
This dataset covers the years 1990 through 2007 and contains two types of indicators for the global distribution of information, communication and technology (ICT) resources. The data includes gini coefficients for the distribution of Internet access within countries, and a technology diffusion index that weights the distribution of broadband subscribers, personal computers, mobile phones, Internet users, and international Internet bandwidth by economic output. The data are secondary source data... (more info)

Access Notes
- These data are available only to users at ICPSR member institutions. Because you are not logged in, we cannot verify that you will be able to download these data.

Dataset(s)

- Dataset - Download All Files (4.9 MB)
  - Documentation: Codebook.pdf
  - Data: SAS, SPSS, Stata, ASCII, Delimited
    - SAS Setup, SPSS Setup, Stata Setup

Study Description

Citation

Persistent URL: http://doi.org/10.3886/ICPSR23562.v1

Export Citation:
- RIS (generic format for RefWorks, EndNote, etc.)
- EndNote XML (EndNote X4.0.1 or higher)

Funding
This study was funded by:
- Peoples and Practices Group (IIS-0713074)
- National Science Foundation (IIS-0713074)

Scope of Study
Subject Terms: communications systems, computer use, information dissemination, information systems, Internet, technology
Discovery and Interpretation

- Identify the form and content
- Identify related objects
- Interpret
- Evaluate
- Open
- Read
- Compute upon
- Reuse
- Combine
- Describe
- Annotate...
Describing and attributing data

• Compound objects
  – Observations
  – Software
  – Protocols...

• Attribution
  – Investigators
  – Data collectors
  – Analysts...

• Ownership, responsibility

Mary Jane Rathbun (1860-1943), working with crab specimens
Metadata

• Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.
  – descriptive
  – structural
  – administrative
Provenance

- Libraries: Origin or source
- Museums: Chain of custody
- Internet: Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. (World Wide Web Consortium (W3C) Provenance working group)

British Library, provenance record: Bestiary - caption: 'Owl mobbed by smaller birds'
Reuse across place and time

- Reuse by investigator
- Reuse by collaborators
- Reuse by colleagues
- Reuse by unaffiliated others
- Reuse at later times
  - Months
  - Years
  - Decades
  - Centuries

Image from Soumitri Varadarajan blog. Iceberg image © Ralph A. Clevenger. Flickr photo
MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- Able to engage with senior management
- Story telling skills
- Translate data driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau
Data Curation and Stewardship

- Services and tools
- Data management planning
- Selection and appraisal
- Metadata, provenance
- Migration
- Economics
- Infrastructure

http://www.librarygirl.net/2013/08/putting-your-best-foot-forward-tl.html
Research workforce

https://jakevdp.github.io/blog/2014/08/22/hacking-academia/
Knowledge Infrastructures
Knowledge Infrastructures: Intellectual Frameworks and Research Challenges

Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation
University of Michigan School of Information, 25-28 May 2012
## Economics of the Knowledge Commons

<table>
<thead>
<tr>
<th></th>
<th>Subtractability / Rivalry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td><strong>Exclusion</strong></td>
<td></td>
</tr>
<tr>
<td>Difficult</td>
<td><strong>Public Goods</strong></td>
</tr>
<tr>
<td></td>
<td>General knowledge</td>
</tr>
<tr>
<td></td>
<td>Public domain data</td>
</tr>
<tr>
<td>Easy</td>
<td><strong>Toll or Club Goods</strong></td>
</tr>
<tr>
<td></td>
<td>Subscription journals</td>
</tr>
<tr>
<td></td>
<td>Subscription data</td>
</tr>
</tbody>
</table>

Adapted from C. Hess & E. Ostrom (Eds.), *Understanding knowledge as a commons: From theory to practice*. MIT Press.
Data Repositories

Proportion of Repositories by Continent - Worldwide

This chart is based on the number of repositories in each Continent. However, some organisations have two or more repositories - over 20 in some cases - and this arguably skews the results.

For a different viewpoint, please see the equivalent chart for Repository Organisations, in which each organisation only counts once, regardless of how many repositories it hosts.

For further data, please see the corresponding table of repositories sorted by country.

Show embedding code
Show legacy chart and embedding code
No Data

- Data not available
- Data not released
- Data not usable
Conclusions

• Defining data
  – Representations used as evidence
  – One person’s signal is another’s noise

• Creating data
  – Models, questions, methods
  – Domain expertise
  – Data science expertise

• Collaborating with data
  – Documentation, description, identity, linking
  – Incentives for release and reuse
  – Curation and stewardship expertise

• Consolidating data value
  – Infrastructure and workforce investments
  – Value propositions
  – Trust fabric
Acknowledgements

UCLA KI Team

*Peter Darch, Milena Golshan, Irene Pasquetto, , Bernie Randles, Ashley Sands

*Former members:
Rebekah Cummings, David Fearon, Ariel Hernandez, Elaine Levia, Jaklyn Nunga, Rachel Mandel, Matthew Mayernik, Alberto Pepe, Kalpana Shankar, Katie Shilton, Francee Williams, Jillian Wallis, Laura Wynholds, Kan Zhang

*Research funding: Alfred P. Sloan Foundation, National Science Foundation, Microsoft Research, DANS-Netherlands

*University of Oxford: Balliol College, Oliver Smithies