Title
Big data, little data, or no data? Scholarship and stewardship to build the UC digital library.

Permalink
https://escholarship.org/uc/item/5tc1z9n9

Author
Borgman, Christine L.

Publication Date
2018-02-27

License
CC BY-NC-ND 4.0
Big Data, Little Data, or No Data?
Scholarship and Stewardship to Build the UC Digital Library

Christine L. Borgman
Distinguished Professor &
Presidential Chair in Information Studies
Director, Center for Knowledge Infrastructures
https://knowledgeinfrastructures.gseis.ucla.edu
University of California, Los Angeles
http://christineborgman.info
@scitechprof

Digital Library Federation X Conference
University of California, Riverside
February 27, 2018
Theme issue ‘Celebrating 350 years of Philosophical Transactions: life sciences papers’ compiled and edited by Linda Partridge
19 April 2015; volume 370, issue 1666
Data sharing policies

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

Researchers share data
Big Data
What are data?

Marie Curie’s notebook

Pisa Griffin

Date: 1/2/07-75 Place: Sakaltutan Zafor
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. Es came from Germany and did the plastering. He arranged the carpentry in Kayseri Çok para gitti (much money went) Has a tractor.

Date: July 1980 Place: Sakaltutan Zafor:
Household now Zafor and wife; Nazif Unal and wife and youngest son, still a boy. They run two dolmus; one with a driver from Suleymanli. Goes in and out once a day. He gets 8,000 a month. Zafor then said, keskin deöili (not sharp - i.e. not profitable) I said he did very well on 8,000 TL with only two journeys a day. Nazif Unal has “bought” a Durak (dolmuş stop) from Beledie and works all day in Kayseri.

http://www.census.gov/population/cen2000/map02.gif
http://ncl.ucar.edu
http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

Research process

• Models and theories
• Research questions
• Methods
  – Domain expertise
  – Practices, protocols
  – Data sources
  – Instruments, software
  – Infrastructure
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. Goodman1,2, Erik W. Rosolowsky3,4, Michelle A. Borkin5, Jonathan B. Foster6, Michael Halle1,4, Jens Kauffmann1,5,6 & Jaime E. Pineda2

Self-gravity plays a decisive role in the final stages of star formation, where dense cores (size ∼0.1 parsec) 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 ‘denrogram’ (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% of the compact ‘pre-stellar cores’ traced by peaks of dust emission are projected on the sky within one of the denrogram’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 denrogram 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 sub-parsec 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 between 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 monotonically, but not quite linearly (see Supplementary Information), related to column density, and it gives a view of ‘all’ of the molecular gas along lines 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 Cerro Tololo) than the remainder of the box (2MASS data only), revealing protostars as well as the scattered starlight known as ‘Circumstellar’ and outflows (which appear orange in this colour scheme). The four billiard ball labels indicate regions containing self-gravitating dense gas, as identified by the dendrogram analysis, and the leaves they identify 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 millimetre-dust emission peaks identified as star-forming or ‘pre-stellar’ cores. | 83 | 10.1038/nature07609 | Vol 457 | January 2009 | 63
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.."
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/

Arte islamica, ippogrifo, XI sec 03, own work
Grey Literature

- Reports
- Working papers
- Conference papers
- Preprints
- Patents
- Datasets
- Audio
- Video
- Slides
- Posters
- Codebooks
- Course syllabi
- Proposals
- Memos

http://www.greynet.org/
Grey Data

- Student applications
- Registrar records
- Learning management systems
- University ID cards: library, health, recreation, dorms, food service, transportation...
- Academic personnel dossiers
- Regulation and compliance data
- Staff surveys
- Sensor networks
- Security cameras
- Network traffic
- Street traffic...


Networks of data


http://humannaturelab.net/wp-content/uploads/2015/01/Fig1-no-text-village-2-only-selection.png
Publications <-> Data: Role

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

Publications <-> Data: Mapping

- Article 1
- Article 2
- Article 3
- Article 4
- Article n
- Dataset time 1
- Dataset time 2
- Observation time 1
- Visualization time 3
- Community collection 1
- Repository 1
Publications <-> Data: Attribution

- Publications
  - Independent units
  - Authorship is negotiated
- Data
  - Compound objects
  - Ownership is rarely clear
  - Attribution
    - Long term responsibility: Investigators
    - Expertise for interpretation: Data collectors and analysts

Data citation and analytics

• Credit
• Attribution
• Discovery

Bibliographic styles

Zotero Style Repository

Here you can find Citation Style Language 1.0.1 citation styles for use with Zotero and other CSL 1.0.1–compatible software. For more information on using CSL styles with Zotero, see the Zotero wiki.

Style Search

<table>
<thead>
<tr>
<th>Format:</th>
<th>author</th>
<th>author-date</th>
<th>label</th>
<th>note</th>
<th>numeric</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Fields:</th>
<th>anthropology</th>
<th>astronomy</th>
<th>biology</th>
<th>botany</th>
<th>chemistry</th>
<th>communications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>engineering</td>
<td>generic-base</td>
<td>geography</td>
<td>geology</td>
<td>history</td>
<td>humanities</td>
</tr>
<tr>
<td></td>
<td>linguistics</td>
<td>literature</td>
<td>math</td>
<td>medicine</td>
<td>philosophy</td>
<td>physics</td>
</tr>
<tr>
<td></td>
<td>psychology</td>
<td>science</td>
<td>social_science</td>
<td>sociology</td>
<td>theology</td>
<td>zoology</td>
</tr>
</tbody>
</table>

8970 styles found:

- 3D Printing in Medicine  (2016-02-13 20:40:33)
- 3D Research  (2015-04-21 12:08:45)
- 3D-Printed Materials and Systems  (2015-04-21 12:08:45)
- 4OR  (2014-05-18 01:40:32)
- AAPS Open  (2016-02-13 20:40:33)
- ABI Technik (German)  (2015-12-16 02:32:01)

1797 unique styles (27 Feb 2018)
If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology

Jillian C. Wallis, Elizabeth Rolando, Christine L. Borgman

Published: July 23, 2013 • https://doi.org/10.1371/journal.pone.0067332

Abstract

Research on practices to share and reuse data will inform the design of infrastructure to support data collection, management, and discovery in the long tail of science and technology. These are research domains in which data tend to be local in character, minimally structured, and minimally documented. We report on a ten-year study of the Center for Embedded Network Sensing (CENS), a National Science Foundation Science and Technology Center. We found that CENS researchers are willing to share their data, but few are asked to do so, and in only a few domain areas do their funders or journals require them to deposit data. Few repositories exist to accept data in CENS research areas. Data sharing tends to occur only through individual research projects.

Published July 23, 2013; screenshot Feb 27, 2018
Searches for author: Christine Borgman, Christine L. Borgman, CL Borgman (excluding other C Borgman authors) on July 28, 2014 and February 25, 2016 for Google Scholar, Web of Science, Scopus

*UCLA cancelled Scopus subscription by 2016*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Scholar (Google)</td>
<td>380</td>
<td>443</td>
<td>7766</td>
<td>9701</td>
<td>39</td>
<td>43</td>
</tr>
<tr>
<td>Web of Science (Thomson-Reuters)</td>
<td>145</td>
<td>150</td>
<td>1629</td>
<td>1967</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td><em>Scopus – July 2014 (Elsevier)</em></td>
<td>77</td>
<td></td>
<td>1314</td>
<td></td>
<td>14 (after 1995)</td>
<td></td>
</tr>
</tbody>
</table>
Attributing responsibility

- Legal responsibility
  - Licensed data
  - Specific attribution required
- Scholarly credit: contributorship
  - “Author” of data
  - Contributor of data to this publication
  - Colleague who shared data
  - Software developer
  - Data collector
  - Instrument builder
  - Data curator
  - Data manager
  - Data scientist
  - Field site staff
  - Data calibration
  - Data analysis, visualization
  - Funding source
  - Data repository
  - Lab director
  - Principal investigator
  - University research office
  - Research subjects
  - Research workers, e.g., citizen science...

Discovery and Interpretation

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

Photo by @kissane; presentation by Jason Scott (@textfiles)
Identity and persistence

• Identity
  – Identifiers
    • DOI, Handles
    • URI, PURL...
  – Naming and namespaces
    • Authors/creators: ORCID, ISNI, VIAF...
    • Generic/specific: registry number...
  – Description
    • Self-describing
    • Metadata augmentation

• Persistence
  – Perishable
  – Long-lived
  – Permanent
Intellectual property

• What can I do with this object?
  – Reuse
  – Reproduce
  – Attribute

• What rights are associated?

• Who owns the rights?

• How open are data?
  – Open data
  – Open bibliography

http://pzwart.wdka.hro.nl/mdr/research/lliang/mdr/mdr_images/opencontent.jpg/
Information and Autonomy Privacy

Data Stewardship: The Ideal

Data Stewardship: the Reality

We just need to migrate the data from these systems to fit into that hole over there.

I’ll get the hammer.


Graduate students

http://gsa.nice.edu/

Mount Wilson Solar Observatory, 2017

Post-doctoral fellows

https://med.nyu.edu/our-community/life-nyu-school-medicine/life-postdoc
Data

If you can’t protect it, don’t collect it.

(privacy and security aphorism)

Therefore:

If you collect it, you must protect it.
Protect Data and Privacy

open by design

http://democracyos.eu/blog/open-by-design


https://privacybydesign.foundation/en/
Lifecycle Model

The DCC Curation Lifecycle Model provides a graphical high-level overview of the stages required for successful curation and preservation of digital material.

Full Lifecycle Actions
- Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres.
- Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.
- Conceive and plan the creation of data, including capture method and storage options.
- Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at this stage.
- Appraise and select. This may be done to ensure that the data meets the criteria for curation and preservation. Return data which fails validation procedures for further appraisal and reselection.
- Conserve and curate. This involves ensuring that the data is in a suitable format for long-term preservation. Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from obsolescence.
- Return data which fails validation procedures for further appraisal and reselection.
- Store the data in a secure manner adhering to relevant standards.
- Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of the digital material and the associated metadata.
- Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should be taken to prevent loss of data and to ensure its integrity. This may include actions such as data cleaning, data validation and data consolidation.
- Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian.
- Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand the data and its context. This may include information such as software, hardware or other contextual information.
- Complex Digital Objects are discrete digital objects, made by combining a number of other digital objects, such as websites.
- Simple Digital Objects are discrete digital items, such as textual files, images or sound files, along with their related identities, references and access information.
- By creating a subset, by selection or query, to create newly derived results, perhaps for publication.
- By migration into a different format.
- Create new data from the original, for example. By migration into a different format.
- Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should be taken to prevent loss of data and to ensure its integrity. This may include actions such as data cleaning, data validation and data consolidation.
- Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian.
- Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand the data and its context. This may include information such as software, hardware or other contextual information.
- Complex Digital Objects are discrete digital objects, made by combining a number of other digital objects, such as websites.
- Simple Digital Objects are discrete digital items, such as textual files, images or sound files, along with their related identities, references and access information.
- By creating a subset, by selection or query, to create newly derived results, perhaps for publication.
- By migration into a different format.
- Create new data from the original, for example. By migration into a different format.
- Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should be taken to prevent loss of data and to ensure its integrity. This may include actions such as data cleaning, data validation and data consolidation.
- Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian.
- Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand the data and its context. This may include information such as software, hardware or other contextual information.
- Complex Digital Objects are discrete digital objects, made by combining a number of other digital objects, such as websites.
- Simple Digital Objects are discrete digital items, such as textual files, images or sound files, along with their related identities, references and access information.
- By creating a subset, by selection or query, to create newly derived results, perhaps for publication.
- By migration into a different format.
- Create new data from the original, for example. By migration into a different format.
- Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should be taken to prevent loss of data and to ensure its integrity. This may include actions such as data cleaning, data validation and data consolidation.
- Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian.
- Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand the data and its context. This may include information such as software, hardware or other contextual information.
- Complex Digital Objects are discrete digital objects, made by combining a number of other digital objects, such as websites.
- Simple Digital Objects are discrete digital items, such as textual files, images or sound files, along with their related identities, references and access information.
- By creating a subset, by selection or query, to create newly derived results, perhaps for publication.
- By migration into a different format.
- Create new data from the original, for example. By migration into a different format.
- Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should be taken to prevent loss of data and to ensure its integrity. This may include actions such as data cleaning, data validation and data consolidation.
- Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian.
- Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand the data and its context. This may include information such as software, hardware or other contextual information.
- Complex Digital Objects are discrete digital objects, made by combining a number of other digital objects, such as websites.
- Simple Digital Objects are discrete digital items, such as textual files, images or sound files, along with their related identities, references and access information.
- By creating a subset, by selection or query, to create newly derived results, perhaps for publication.
- By migration into a different format.
- Create new data from the original, for example. By migration into a different format.
Promote Responsible Data Practices

• Respect information and autonomy privacy
  – Open data: release and reuse
  – Data collection and use
  – Data management
  – Collaborations
  – Publications

• Community
  – Faculty
  – Librarians
  – Staff
  – Students
  – External partners

• Joint governance process

https://www.commontrends.org/views/2014/09/20/corporations-your-diet
Scholarship and Stewardship to Build the UC Digital Library

• Mission-drive stewardship
  – Research
  – Teaching
  – Services
• Steward the scholarly record
  – Integrated workflows
  – Version of record
  – Record of versions (Van de Sompel)
• Support discovery at scale
  – Human readable
  – Machine readable
  – Lawyer readable
• Sustain trust of community
  – Privacy: information, autonomy
  – Academic freedom
  – Stewardship and governance
UC Leadership in Data Policy

• We must maximally enable the **mission** of the University by supporting the values of **academic and intellectual freedom**.

• We must be **good stewards** of the **information entrusted** to the University.

• We must ensure that the University has **access to information resources** for **legitimate business purposes**.

• We must have a University community with **clear expectations of privacy**—both **privileges and obligations** of individuals and of the institution.

• We must make decisions within an **institutional context**.

• We must acknowledge the **distributed nature** of information stewardship at UC, where **responsibility for privacy and information security** resides at every level.

---

Acknowledgements

UCLA Center for Knowledge Infrastructures