

UCLA

Presentations

Title

Where's all that good data I keep hearing about? Issues in the creation and reuse of research data

Permalink

<https://escholarship.org/uc/item/0jc6r4mp>

Author

Borgman, Christine L.

Publication Date

2013-02-06

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

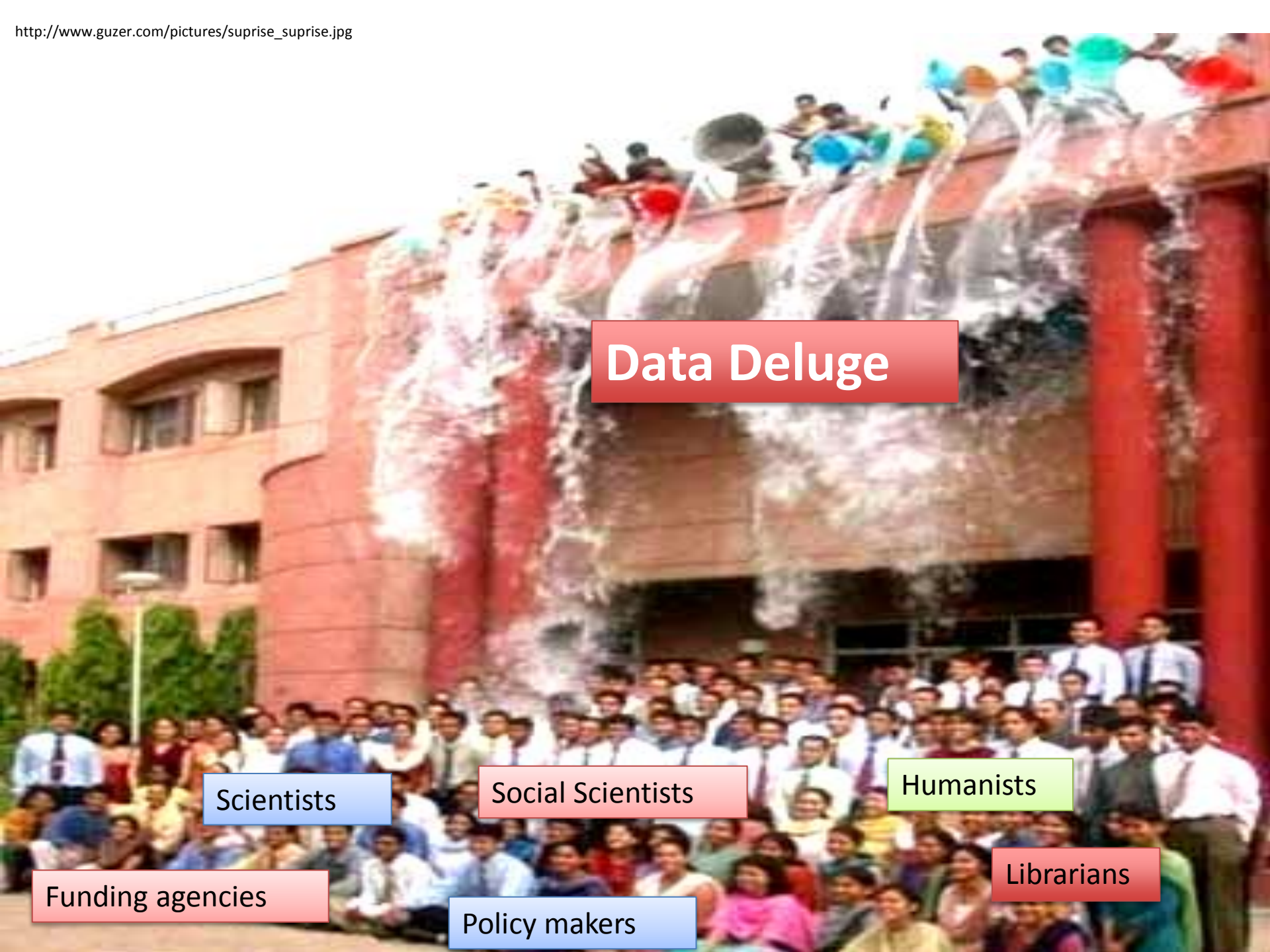
Where's all that good data I keep hearing about? Issues in the creation and reuse of research data

Christine L. Borgman

Oliver Smithies Visiting Fellow and Lecturer, Balliol College, Oxford

Professor and Presidential Chair in Information Studies
University of California, Los Angeles

Oxford Internet Institute, Brown Bag Talk, 6 February 2013



Data Deluge

Scientists

Social Scientists

Humanists

Funding agencies

Policy makers

Librarians

Data Practices Model

- Premise: multiple dimensions influence the use and reuse of data
 - List is not exhaustive
 - Dimensions are not mutually exclusive
- Data: sources of evidence
 - Observations, experiments, records, models
 - Identifying data is a scholarly act
 - Data are fuzzy objects

Case studies

- Astronomy
- Sensor networked science
- Social network behavior
- Digital humanities
- Qualitative interviews and ethnographies



Copyright Sydney

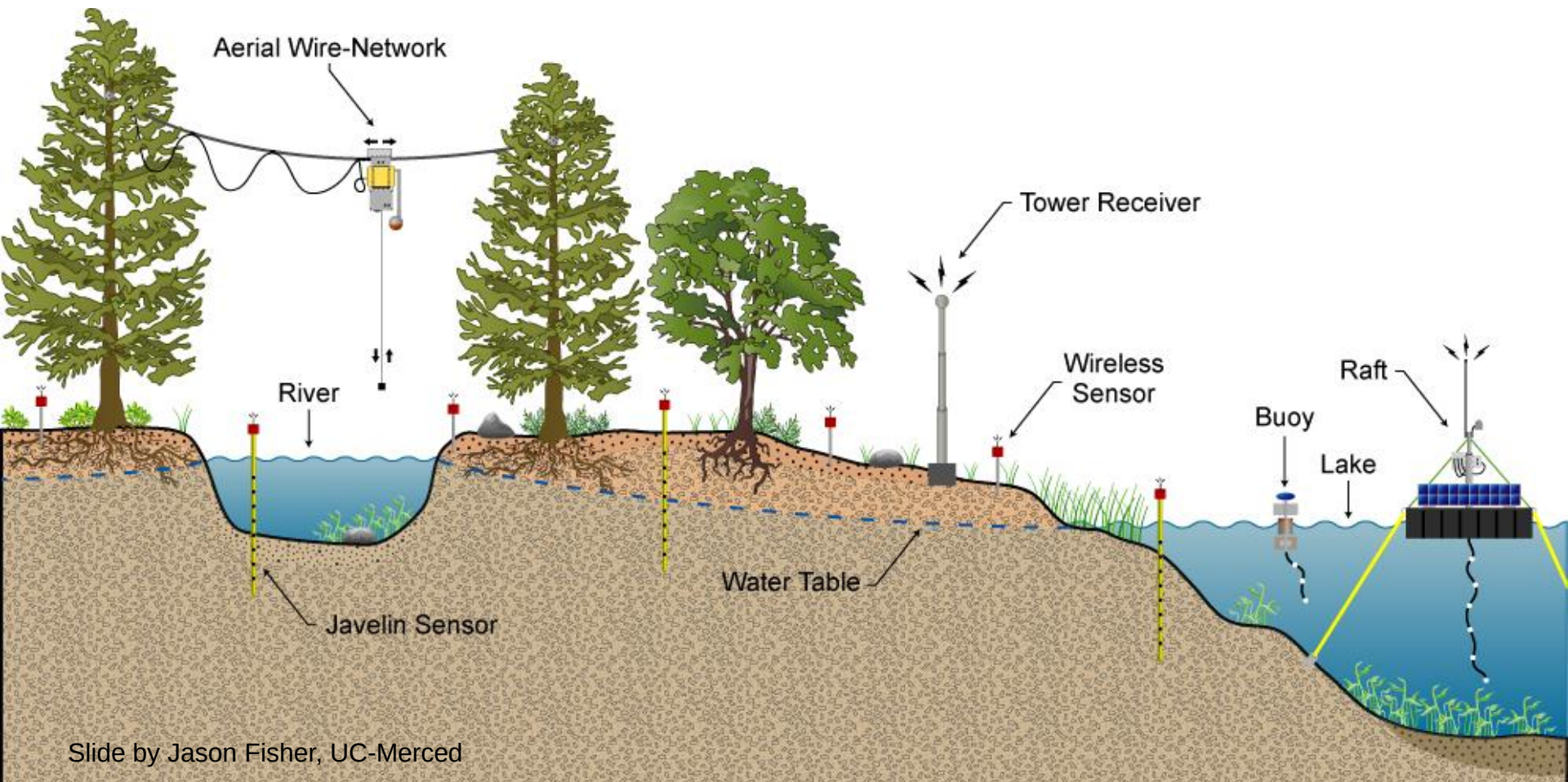
NASA photo of the day 22 Jan 2013

The Cygnus Wall of Star Formation

Credit & Copyright: [Nick Pavelchak](#)

Explanation: The [North America nebula](#) on the sky can do what the [North America continent](#) on Earth cannot -- form stars. Specifically, in analogy to the Earth-confined continent, the bright part that appears as [Central America](#) and [Mexico](#) is actually a hot bed of gas, dust, and newly formed stars known as the [Cygnus Wall](#). The [above image](#) shows the star forming wall [lit and eroded](#) by [bright young stars](#), and partly hidden by the dark [dust](#) they have created. The part of the [North America nebula](#) (NGC 7000) shown spans about 15 [light years](#) and lies about 1,500 light years away toward the [constellation](#) of the Swan ([Cygnus](#)).

Field Deployment of Embedded Sensor Networks

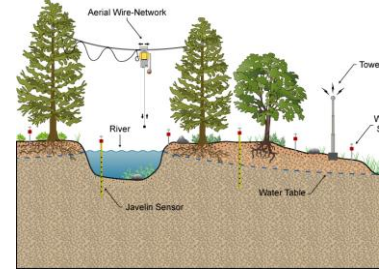


Dimensions of Data Practices

1. Big science, little science
2. Big data, long tail
3. Data Sources
4. Collaboration and Credit
5. Research Methods
6. Metadata
7. Provenance
8. Signal and Noise
9. Use and Reuse of Data
10. Innovation and Challenges



Big science, little science



Astronomy

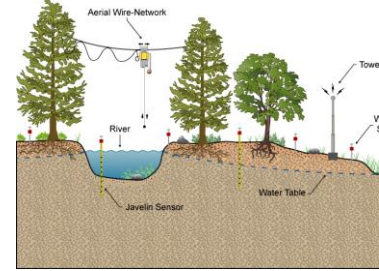
- Big science
- Massive community infrastructure
- Long term investments
- International

Sensor networked science

- Little science
- Minimal community infrastructure
- Short term adaptation to field conditions
- Local data collection



Big data, long tail



Astronomy

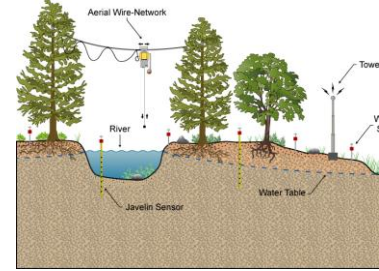
- Large volumes of data
- Continue to grow relative to phenomena
- Some small data astronomers
- More data than they can analyze

Sensor networked science

- Small volumes of data
- Unprecedented scale of data rates
- Parallel collection of physical samples
- Least publishable unit



Data Sources



Astronomy

- Observations from telescopes
 - Ground-based
 - Space-based
- Data from repositories
 - SDSS
 - PAN-STARRS
 - LSST
- Data papers

Sensor networked science

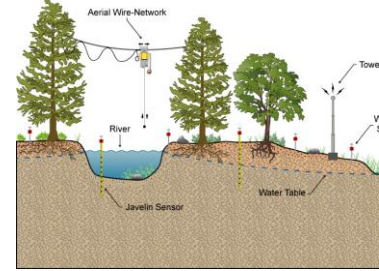
- Observations from sensor networks (S, T)
 - Scientific
 - Science to guide sensors
- Observations from sampling water, soil (S)
- Sensor performance data (T)
- Robotic guiding data (T)
- Background data from repositories (S)

S = science teams

T = technology teams



Collaboration and credit



Astronomy

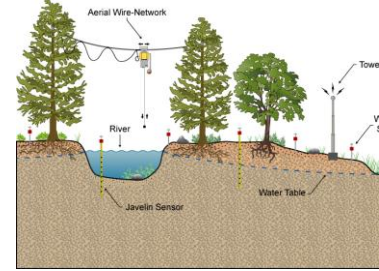
- Large, distributed, multiple
- Multi-disciplinary
 - Astronomers
 - Physicists
 - Computer science /software engineering
 - Instrument builders

Sensor networked science

- Joint science and technology deployments
- Tech teams learn the science
- Papers published separately in respective journals
- Data authorship diffuse



Research methods



Astronomy

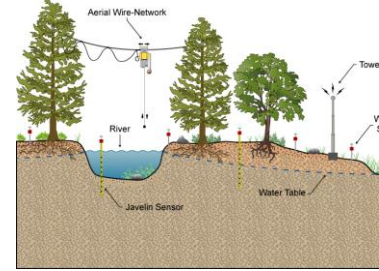
- Observational science
- Instruments determine available data
- Study celestial objects
- Study physics within spectra
- Theoreticians, model builders
- Computational methods
- Visualization

Sensor networked science

- Exploratory science and technology
- Test hypotheses in lab or field
- Develop new methods
- Return to same scientific sites
- Model
 - scientific phenomena
 - networks



Metadata



Astronomy

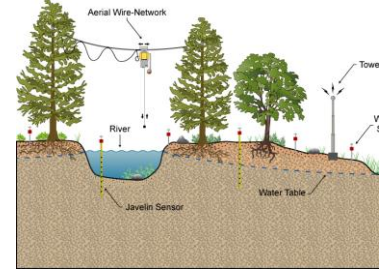
- FITS, ca 1980
- Astrophysics Data System
- CDS
- SIMBAD
- NED
- VizieR

Sensor networked science

- Few extant standards
- Standards not adopted, except seismology, genomics
- Local file naming conventions
- Spreadsheets to exchange data
- Documentation intended for local use



Provenance



Astronomy

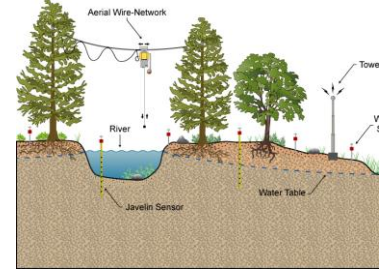
- Publications linked to objects
- Progress in
 - Publications linked to observations
 - Objects linked to observations
- Public sources used for calibration
- Dataverse: UNF, Handle, DOI for data

Sensor networked science

- Local data management
- Each team controls own data
- Tech teams calibrate sensor data
- Data dispersed at end of field deployment
- Few community repositories



Signal and noise



Astronomy

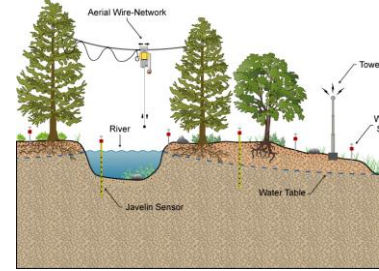
- Most astronomers want cleaned data
- Some want to do their own cleaning of observations
- SETI example
- Event-driven astronomy

Sensor networked science

- Ground truth sensors to local conditions
- Unstable sensor technology
- Indirect measurements require models of phenomena



Use and reuse of data



Astronomy

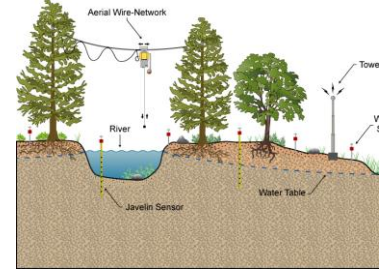
- Data abundance
- Trend toward data distribution
- Space-based missions better supported with repositories
- Journals encouraging data links
- Zooniverse, citizen science
- Astrometry.net
- Worldwide telescope
- Astrometry.net

Sensor networked science

- Data scarcity
- Minimal deposit or posting of data
- Will share privately on request
- Data exchange depends on trusted relationships
- Minimal infrastructure to curate data



Innovation /challenges



Astronomy

- ADS all sky survey integrating data from many sources
- Virtual observatory less successful than hoped

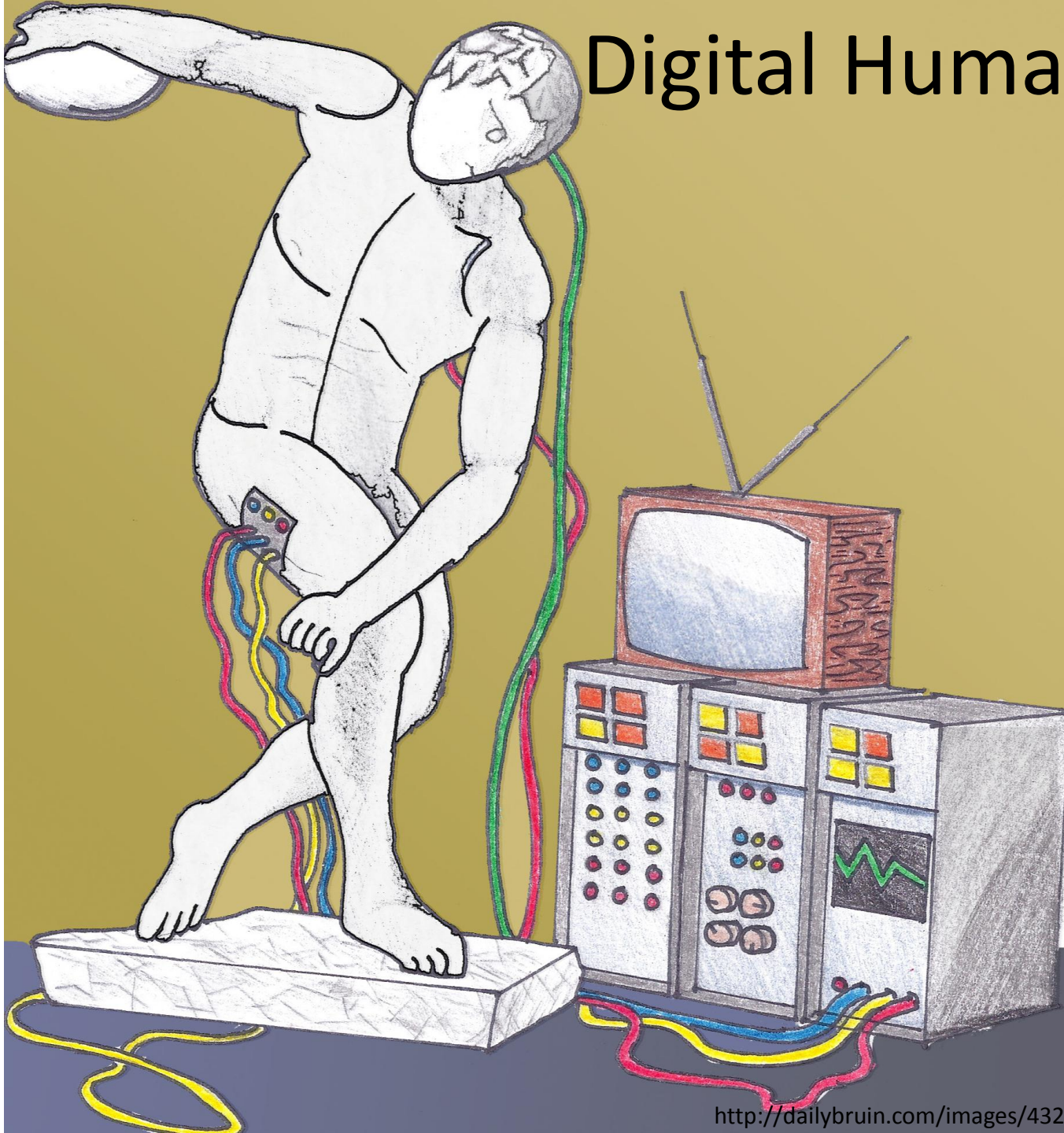
Sensor networked science

- Innovation in scientific application of sensor technologies
- Minimal success
 - Data registry
 - Deployment planning system
 - Data repository
- Value of CENS data lies in its heterogeneity

Social Media / Twitter

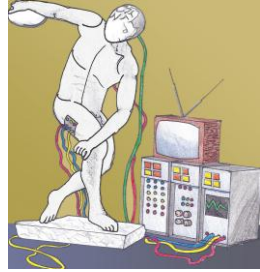


Digital Humanities





Big science, little science



Social Networks/Twitter

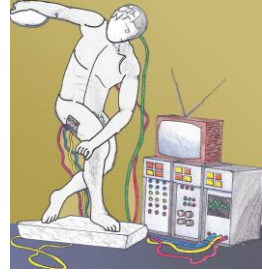
- Little science
- Corporate infrastructure
- Minimal research infrastructure

Digital humanities

- Little science
- Minimal community infrastructure



Big data, long tail



Social Networks/Twitter

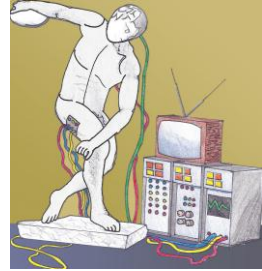
- Big data in volume
- Minimal data about social act
 - 140 characters of tweet
 - GPS coordinates, sometimes
 - URL shortened
- Computer generated data

Digital humanities

- Big data relative to phenomena
- Small volumes of data
- Long tail activity
- Digitized data
- Manual cataloging



Data Sources



Social Networks/Twitter

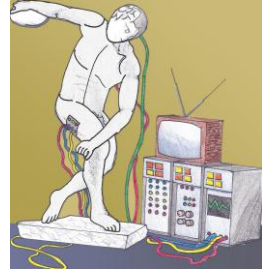
- Sole source: Twitter
- Multiple data providers
- Select data via public API
- Pay for data
- Proprietary, contractual access

Digital humanities

- Heterogeneous sources
- Observations of human activity
- Records, current and archival
- Text corpora
- Images
- Audio, video
- Maps...



Collaboration and credit



Social Networks/Twitter

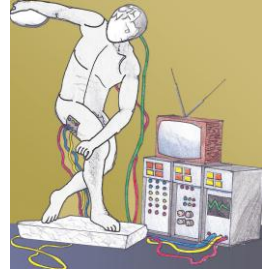
- Individuals and small teams
- Domain knowledge
 - Communication
 - Political science
 - Area studies
- Programming expertise
- Statistical expertise
- Methods: data selection

Digital humanities

- Small teams
- Humanities domain knowledge
- Computational knowledge
 - Data mining
 - Entity extraction
 - Network analysis



Research methods



Social Networks/Twitter

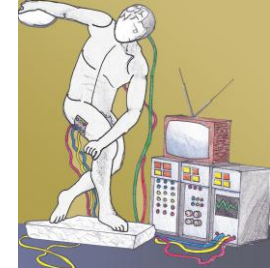
- Computational modeling
- Statistical analyses
- Combine with other data
 - Bibliometrics
 - URLs
 - Tags and annotations
 - Demographics
 - Geospatial
 - Temporal

Digital humanities

- Collection building
- Analysis of relationships, events
- Annotation, tagging
- Geospatial mapping
- Temporal mapping
- Visualization, modeling



Metadata



Social Networks/Twitter

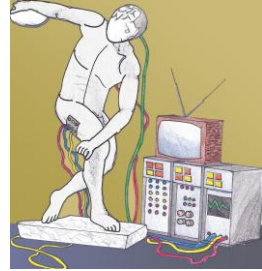
- Standard APIs, data structures
- GIS standards

Digital humanities

- Text encoding standards
- GIS standards
- Ontologies, thesauri
- Local, project-specific coding



Provenance



Social Networks/Twitter

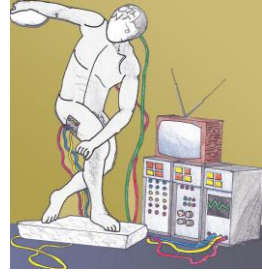
- Data origins clear
- Data selection not always clear
- Verify analyses
- Replication of behavior unlikely
- Context of tweets difficult to establish

Digital humanities

- Establishing provenance is a research question



Signal and noise



Social Networks/Twitter

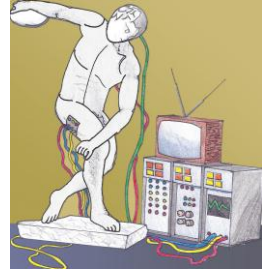
- Noisy data
- Anonymized tweets
- Randomization of data feed

Digital humanities

- Heterogeneous historical sources
- Sources often difficult to verify
- Authoritative texts
- Authoritative records
 - Museums
 - Archives



Use and reuse of data



Social Networks/Twitter

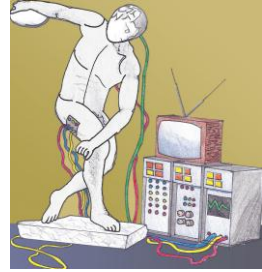
- Contractual release of data
- Rarely subject to deposit
- Proprietary control

Digital humanities

- Constructing digital libraries for reuse may be goal of project
- Increasing coordination
- Complex curation investments required
 - Sustainability of sites
 - Curation of back end data
 - Modeling standards



Innovation /challenges



Social Networks/Twitter

- Trust in proprietary sources
- Heterogeneous content
- Variation in scope and content
- Ethics and privacy issues

Digital humanities

- Building collections vs doing research
- Asking new questions vs asking same questions faster

THE I CAN'T DEAL WITH
ALL THIS DATA ERA IS OVER.



Your potential. Our passion.™
Microsoft

Microsoft Office has evolved. Have you?
microsoft.com/uk/evolve

 **Office**

Dimensions of Data Practices

1. Big science, little science
2. Big data, long tail
3. Data Sources
4. Collaboration and Credit
5. Research Methods
6. Metadata
7. Provenance
8. Signal and Noise
9. Use and Reuse of Data
10. Innovation and Challenges

Where's all that good data?

1. Big science invests in infrastructure for data curation; little science does not
2. Big, computer-generated data are easier to release than small amounts of hand-crafted data
3. Homogenous data are more sharable
4. Larger collaborations require agreements on data practices
5. Data from computational methods are easier to share than data from hand-crafted methods

Where's all that good data?

6. Metadata adoption makes data more sharable
7. Data with clear provenance are easier to reuse
8. Noisy data require more expertise to interpret
9. Reuse is more institutionalized in big science
10. Each field has distinct opportunities and challenges

Conclusions / themes

- Fields vary widely
 - Data types, sources, characteristics
 - Agreements on what are their data
- Data are available for reuse
 - Data abundance
 - Investments in infrastructure for curation and access
- Investments in infrastructure occur
 - Critical mass in size of collaborations
 - Homogeneous data with value for reuse
 - Data unencumbered by
 - Intellectual property
 - Privacy, confidentiality
 - Commercial value