DATA IS IT GREY, MALIGNED OR MALIGNANT?

Julia Gelfand & Daniel C. Tsang

University of California, Irvine

16th international conference on grey literature,

Washington DC: Library of Congress

8-9 December 2014

DEFINITIONS OF GREY LITERATURE/GREY DATA

Grey Literature:

- 1997- Luxembourg definition - “that which is produced on all levels of government, academics, business & industry in print & electronic formats, but which is not controlled by commercial publishers”
- 2010-Prague definition - “Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers, ie, where publishing is not the primary activity of the producing body.” Reinforces more digitally converted & born digital content with 4 caveats:
  - Documents character of greyness – more multi & interdisciplinary in content & form
  - Includes legal nature of created outputs – intellectual property protected
  - Reinforces quality level for peer review, validation
  - Links to intermediation – bridging collection status to readers/users

Data: created from research enterprise – from scientific and non-scientific communities

- OMB definition for data – “the recorded factual material commonly accepted in the scientific community as necessary to validate research findings”

Data Science: “May encompass the full spectrum of theories and methods that use data to understand, &

MALIGNED OR MALIGNANT?

Speaking ill of someone/thing; being evil, harmful, injurious, slanderous, defaming

wordinfo.info

Tending to produce a bad outcome, being highly invasive as in pathology, dangerous, cause harmful influence; out of control but must be explained

www.thefreedictionary.com

MIGRATION OF GREY LITERATURE TO GREY DATA

- Step-child image - Research data considered inappropriate for mass consumption nor ready for lay public without interpretation
- Serious interest in findings – migration from file cabinet/computer drives to publication
- Role of funding agencies changes future of data – less forgotten
- DataPaper – introduces searchable metadata for the data rather than just text
- Process of finding, manipulating, repurposing, curating, archiving, sharing, managing

RISKS AND CONDITIONS OF DATA

http://www.cresol.co.in/cloudservices.aspx

www.truecount.com

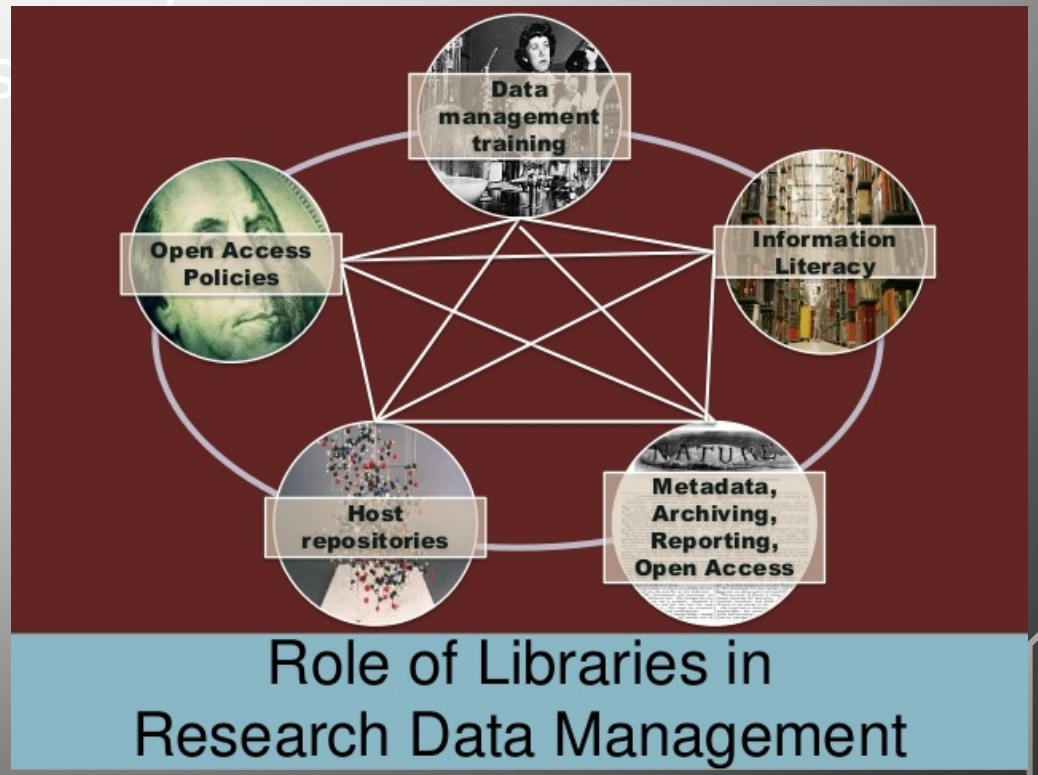
RESEARCH FRONTIER

- Not the wild west anymore
- Data Sharing
- Maturing cyberinfrastructure
- Networked environments
- Statistical analysis
- Evidence-based orientations

IPTROLLTRACKER2.WORDPRESS.COM

ROLE OF LIBRARIES

- Finding data – traditional information seeking behavior
- Utilizing technologies – QR Codes, RFID, discovery systems, social networks
- Conducting & Supporting Research
- Internet Manifesto 2014
- Staffing Needs
  - Curation
  - Data Management
  - Preservation
  - Repositories



http://image.slidesharecdn.com

DATA PROFESSIONALIZATION

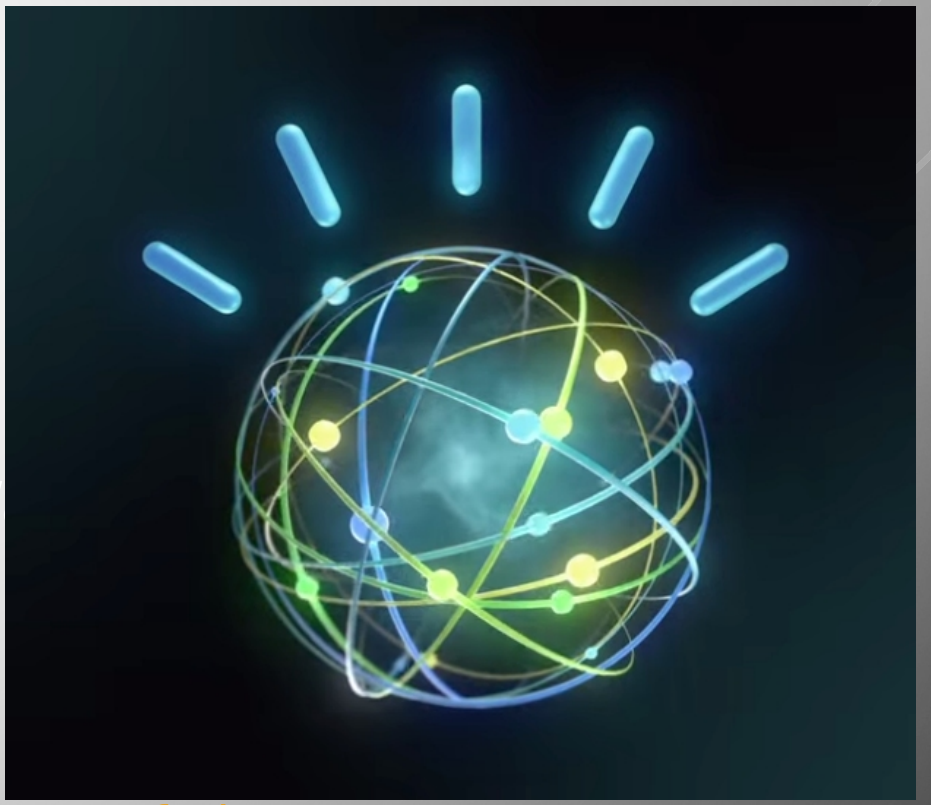
- Industry, Business/Commerce
- Interdisciplinary
- Innovation – data mastery + data vision
  - Amplify
  - Reduces building & investment costs
  - Allows for better cash flow during projects
  - Encourages new opportunity costs
  - Promotes a more green & sustainable environmental landscape

VALUE OF REPOSITORIES

- Reinforces Open Access
- Increasingly soliciting data– becoming less text-centric
- Subject or thematic : ArXIV, ePrint servers
- Institutional : eScholarship ; serves UG & graduate students
  - Intersections between Scholarly communication & information literacy
- 2015 = 10th anniversary of Conference on Open

DATA STRUCTURE

- New Initiatives & Examples
  - UCI's Data Science Initiative
  - Statistics vs Computing
  - Robots or Robotic Programming – welcome W



www.forbes.com

www.cnn.com

UCI Data Science Initiative

IMPACTS OF OPEN ACCESS

library.clemson.edu

- Data Intensive Databases
- Government Information – ERIC, Agricola, PubMed
- New Journals – PLoS; hybrids
- GenomeBanks
- Technical Reports
- eBooks – OP & digitized works, born digital, hybrid

EXAMPLES OF DATA

- One flavor does not apply to all
- Data has many properties & structures – demographic, geographic, etc to describe it

http://legendairyicecream.weebly.com/store.html

SURVEY DATA

www.mtabsurveyanalysis.com

Forms include:

- Demographic
- Public opinion
- Real estate valu
- Consumer behav
- Geographic
- Political participation
- Trends

All require structures that applies to form analysis & communicate results

www.universalsurvey.com

BIG DATA

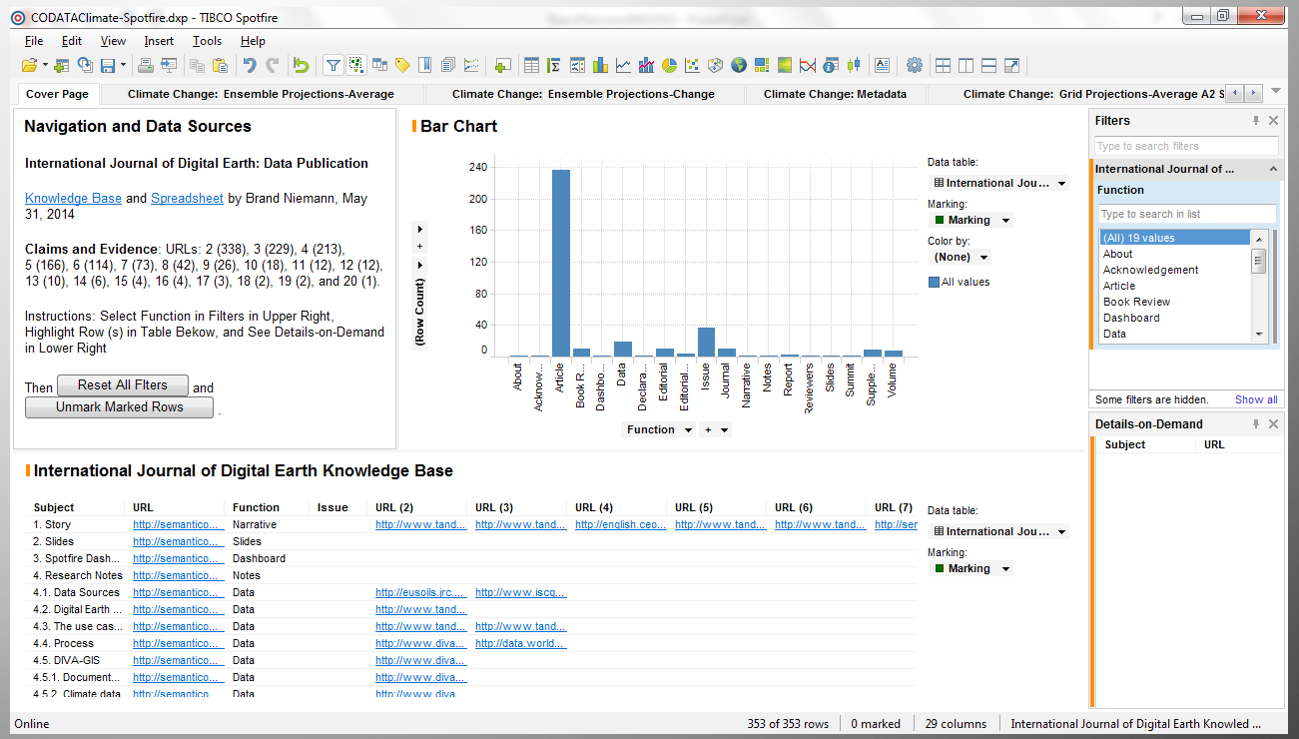
"Capability to manage a huge volume of disparate data, at the right speed and within the right time frame to allow real-time analysis and reaction." Typically broken down into 1) volume, 2) velocity, 3) variety

www.dreamstime.com

www.businessbrio.com

DATA TOOLS

- Spreadsheets – Excel, SPSS
- ClearStory Data
- Statista
- Trifacta
- Paxata
- MapReduce
- Data Conversion Laboratory



semanticommunity.info

CREDIBILITY OF DATA

- More information but perhaps more false information
- All data not equal
- Trust & source credibility

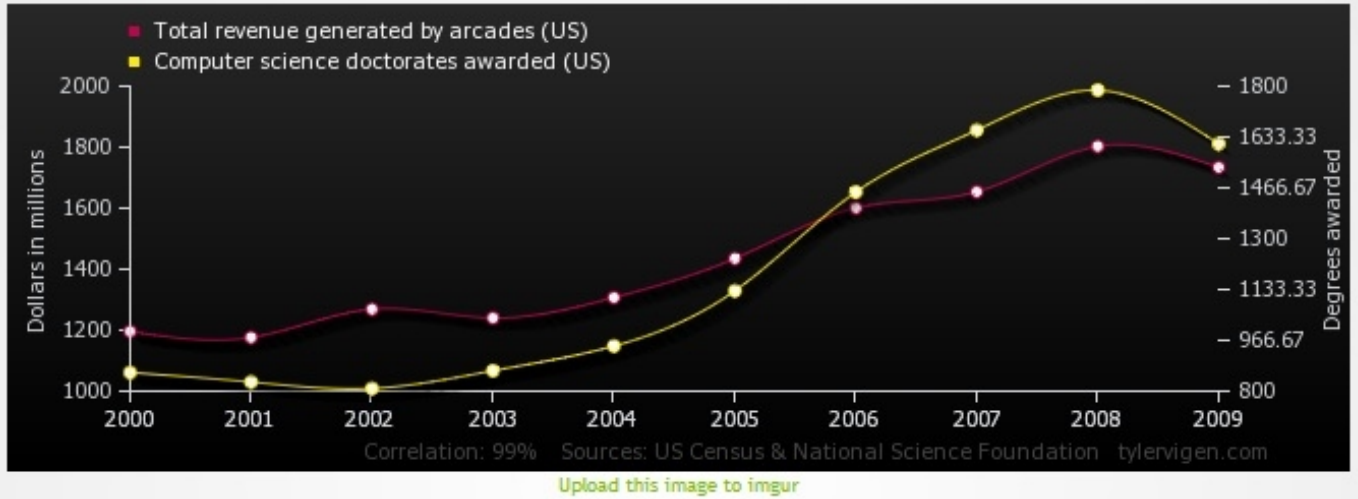
datasupport.researchdata.nl

SPURIOUS DATA: CORRELATIONS

Total revenue generated by arcades (US)

correlates with

Computer science doctorates awarded (US)



Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Total revenue generated by arcades (US) Dollars in millions (US Census)	1,196	1,176	1,269	1,240	1,307	1,435	1,601	1,654	1,803	1,734
Computer science doctorates awarded (US) Degrees awarded (National Science Foundation)	861	830	809	867	948	1,129	1,453	1,656	1,787	1,611

Correlation: 0.985065

Permalink - Mark as interesting (3,970) - Not interesting (2,013)

View all correlations - Discover a new correlation

Re-Chart

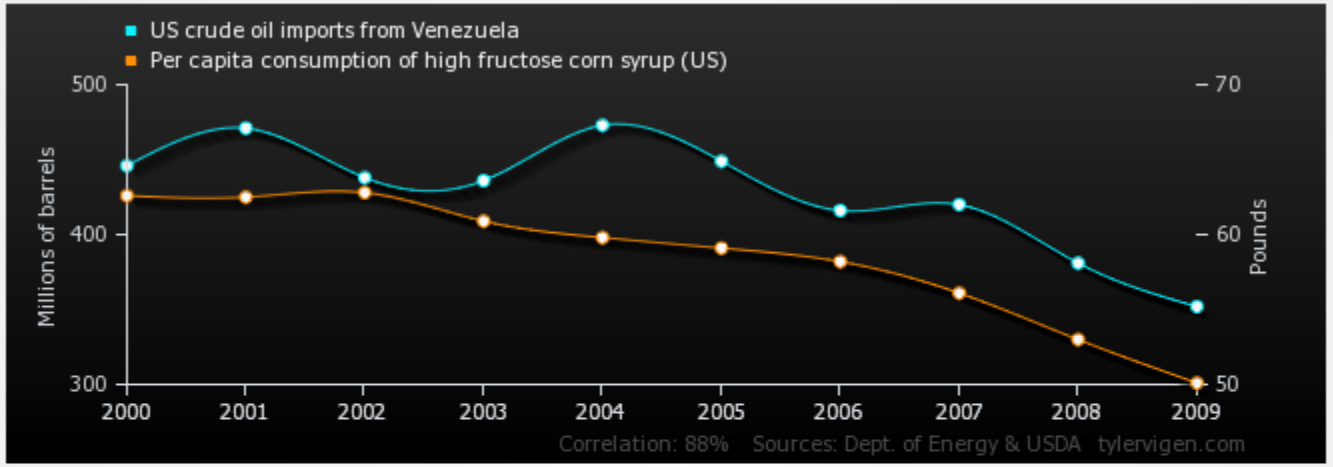
http://www.tylervigen.com/view_correlation?

SPURIOUS DATA: CORRELATIONS

US crude oil imports from Venezuela

correlates with

Per capita consumption of high fructose corn syrup (US)



Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US crude oil imports from Venezuela *Millions of barrels (Dept. of Energy)*	446	471	438	436	473	449	416	420	381	352
Per capita consumption of high fructose corn syrup (US) *Pounds (USDA)*	62.6	62.5	62.8	60.9	59.8	59.1	58.2	56.1	53	50.1

Correlation: 0.884883

Permalink - Mark as interesting (2,354) - Not interesting (2,271)

View all correlations - Discover a new correlation

Re-Chart

http://www.tylervigen.com/

DATA PUBLISHING & PUBLICATIONS

- Data Management Plans – CDL
- Retention vs de-selection

http://www.cdlib.org/

http://www.cdlib.org/uc3/

https://escholarship.org/

LOOKING FORWARD: FUTURE FOR GREY DATA

- Becomes central value in academic libraries – an evolving process
- Sufficient training available for data management
  - Strategists, statisticians, technologists, designers, curators, media specialists
- Grey data to be introduced like Grey

MALIGNED OR MALIGNANT?

wordinfo.info

www.thefreedictionary.com

And how grey?

Thank you!

Julia Gelfand (jgelfand@uci.edu)
Daniel C. Tsang (dtsang@uci.edu)