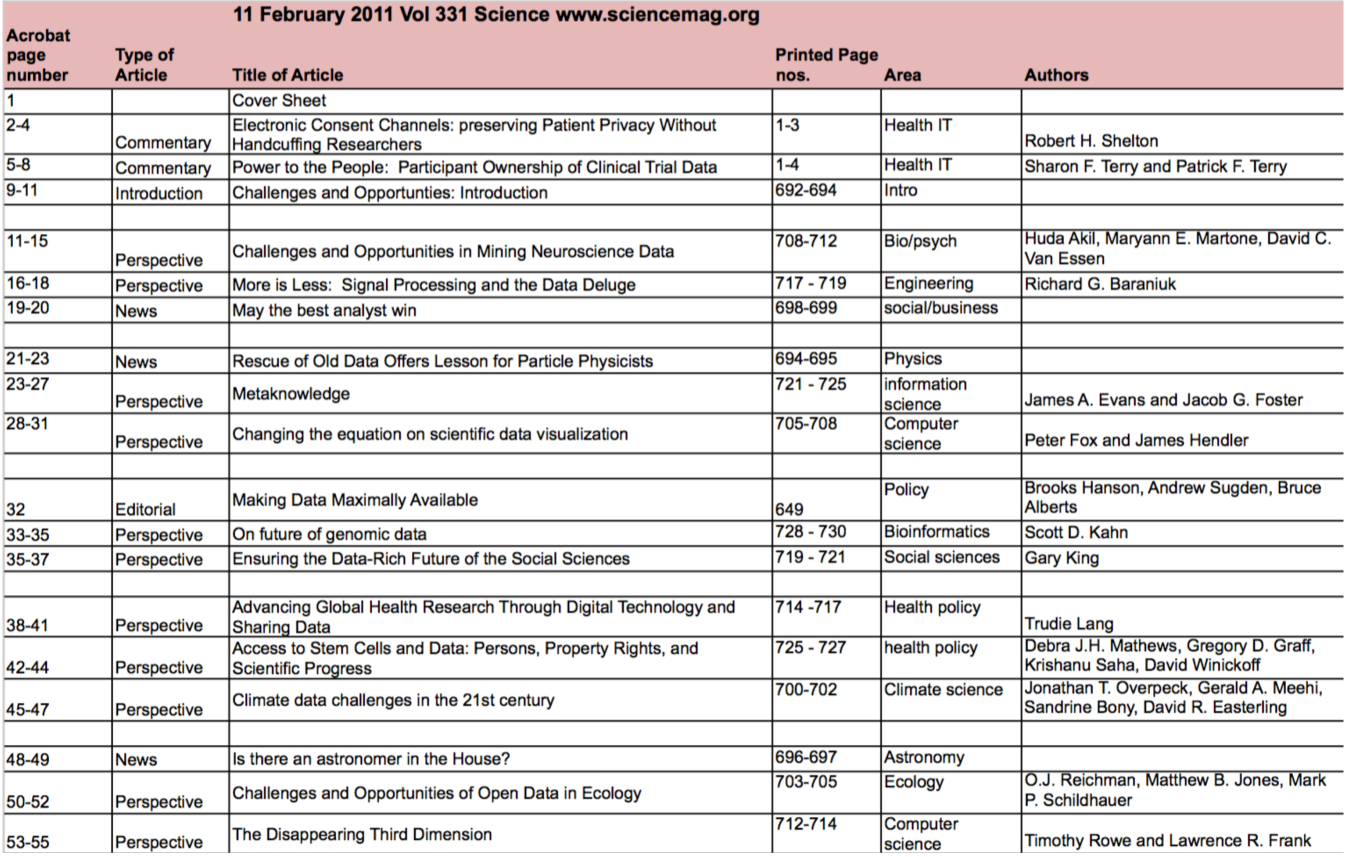
Carlos Sathler | cssathler@gmail.com

INFO-I 535 – MGMT ACCESS USE BIG DATA (Fall 2016)

Big data in Science

All answers are based on the following readings from articles listed below:



11 February 2011 Vol 331 Science www.sciencemag.org

Acrobat page number	Type of Article	Title of Article	Printed Page nos.	Area	Authors
1		Cover Sheet			
2-4	Commentary	Electronic Consent Channels: preserving Patient Privacy Without Handcuffing Researchers	1-3	Health IT	Robert H. Shelton
5-8	Commentary	Power to the People: Participant Ownership of Clinical Trial Data	1-4	Health IT	Sharon F. Terry and Patrick F. Terry
9-11	Introduction	Challenges and Opportunities: Introduction	692-694	Intro	
11-15	Perspective	Challenges and Opportunities in Mining Neuroscience Data	708-712	Bio/psych	Huda Akil, Maryann E. Martone, David C. Van Essen
16-18	Perspective	More is Less: Signal Processing and the Data Deluge	717 - 719	Engineering	Richard G. Baraniuk
19-20	News	May the best analyst win	698-699	social/business	
21-23	News	Rescue of Old Data Offers Lesson for Particle Physicists	694-695	Physics	
23-27	Perspective	Metaknowledge	721 - 725	information science	James A. Evans and Jacob G. Foster
28-31	Perspective	Changing the equation on scientific data visualization	705-708	Computer science	Peter Fox and James Hendler
32	Editorial	Making Data Maximally Available	649	Policy	Brooks Hanson, Andrew Sugden, Bruce Alberts
33-35	Perspective	On future of genomic data	728 - 730	Bioinformatics	Scott D. Kahn
35-37	Perspective	Ensuring the Data-Rich Future of the Social Sciences	719 - 721	Social sciences	Gary King
38-41	Perspective	Advancing Global Health Research Through Digital Technology and Sharing Data	714 -717	Health policy	Trudie Lang
42-44	Perspective	Access to Stem Cells and Data: Persons, Property Rights, and Scientific Progress	725 - 727	health policy	Debra J.H. Mathews, Gregory D. Graff, Krishanu Saha, David Winickoff
45-47	Perspective	Climate data challenges in the 21st century	700-702	Climate science	Jonathan T. Overpeck, Gerald A. Meehi, Sandrine Bony, David R. Easterling
48-49	News	Is there an astronomer in the House?	696-697	Astronomy	
50-52	Perspective	Challenges and Opportunities of Open Data in Ecology	703-705	Ecology	O.J. Reichman, Matthew B. Jones, Mark P. Schildhauer
53-55	Perspective	The Disappearing Third Dimension	712-714	Computer science	Timothy Rowe and Lawrence R. Frank

1. For each of the articles of type "Commentary" or "Perspective" identify the 1-2 primary challenges facing the area. Further, identify 1-2 primary opportunities that the area could realize through better organization and access to data

   1.1. Electronic Consent Channels: preserving Patient Privacy Without Handcuffing Researchers
   Primary challenge(s): obtain and keep patient information to support research, particularly clinical trials; reach out to individuals that could benefit from clinical trials as part of R&D efforts
   Primary opportunity(ies): create application systems that provide individuals with digital means to ensure ownership of their personal health, and that allow them to securely consent to release their health information, selectively per their discretion, to specific organizations and purposes.

   1.2. Power to the People: Participant Ownership of Clinical Trial Data
   Primary challenge(s): participants of clinical trials don't own their health information and don't have control over it; we want the persons who are contributing their data to research projects, in the academia or otherwise, to have ownership and control over their health data and its use in research (clinical trials in particular).
   Primary opportunity(ies): data sharing and crowdsourcing, if only we can incent people to make their electronic medical records and personal health records available selectively for research and clinical trials, through the assurance that they will receive that they fully own their medical

information, and that they have full control over it. Specifically, who is allow access to it, for what purpose, and for how long.

1.3. Challenges and Opportunities in Mining Newroscience Data
Primary challenge(s): The brain is extremely complex. A metaphor to describe how difficult it would be to represent the brain digitally through data is to imagine the task of "deciphering" a "neural choreography"; a dynamic system with many layers, micro and macro pathways and a large number of functions. Scientists study the brain in "silos" and don't share data; still, it is extremely difficult to aggregate the data they use.
Primary opportunity(ies): Launch a new field – Neuroinformatics to more adequately tackle the effort to aggregate existing brain data and incorporate new data bound to become increasingly available in the coming years; continue to collect data on micro and macro pathways in the brain; continue to enhance NIF (Neuroscience Information Framework) to facilitate discovery from the existing data (fourth paradigm).

1.4. More is Less: Signal Processing and the Data Deluge
Primary challenge(s): Sensors are part of "sensor systems" for which they have the function of acquire digital data; other functions of a sensor system are data processing (immediately after data is acquired), communication of the data (sensor are normally connected to mobile devices), and data storage (increasingly in the cloud). The capacity current sensors have to acquire digital data exceeds the capacity it has to process that data; transmission rates for communication of that data is also less than required to handle the volume of data sensors capture; and finally, there is not enough storage capacity to save all the data being captured.
Primary opportunity(ies): Redesign of sensor systems so that improvements can be made to all components of the systems, resulting in less data loss. For example, signal processing algorithms could be improved; and data compression and triage could be improved.

1.5. Metaknowledge
Primary challenge(s): The current existence of more scientists, more conferences, more exchanges, more communication channels among researchers has increased the complexity of how we can represent and understand the totality of our scientific knowledge, i.e., metaknowledge.
Primary opportunity(ies): Using data mining and inference on the output of scientific work, such as journal articles, to draw insights about opportunities for new research, to generate metaknowledge that will include understanding of the organization of current knowledge, and to improve our understanding of the processes that facilitate knowledge production.

1.6. Changing the equation on scientific visualization
Primary challenge(s): Data visualization tools do not perform well enough against big data. Hence scientist cannot use them to communicate intermediary results of their work, thus compromising their ability to exchange ideas with other scientists while their research is ongoing.
Primary opportunity(ies): Build new visualization tools that scale well for use with big data would allow for easy-to-generate visualizations during research, would allow information sharing, invite feedback, and would allow for insights into intermediary results of research work.

1.7. On future of genomic data
Primary challenge(s): The challenge here is similar to the challenges related to sensor systems. Genome sequencing is done faster than ever before and it's generating more data than can be

managed by researchers. The data is being generated too fast, and output of the sequencing process is creating larger data sets to represent the genomes (much like the sensors in the sensor article, which have the capacity now to digitalize more features). So, it's hard to store the data, and it's hard to process these large data sets as well. Just like in the case of sensors there is an effort to process the data before storing it, but that creates issues of provenance and reproducibility of research. Data compression is also being explored as way to save and disseminate the data among researchers. In parallel to these challenges, there is the issue of privacy and the concern about imperfect de-identification methods. The authors state that "research has shown that the genomic data" is inherently identifiable".
Primary opportunity(ies): Creating accepted ways to generate "derived information" along with proper compression techniques seem to be the most promising opportunities to deal with the vast amounts of genomic data being produced with the current sequencing technologies.

1.8. Ensuring the Data-Rich Future of the Social Sciences
Primary challenge(s): There is so much more information collected about individuals and their preferences, choices, opinions, political preferences, consumer profile, location, health information, to name a few areas; there are so many new ways to capture information that can be linked to individuals, such as smart phones, web applications, wearables, surveillance devices; even satellites are becoming capable to zoom in into much smaller areas. Most challenges in the big data for the social sciences relate to the need to archive and share this data for research and the public in ways that preserve individuals' privacy and access control over their data.
Primary opportunity(ies): Ensuring privacy, for example through "privacy-enhanced data sharing protocols" will allow researchers to more readily make data available for each other; another opportunity lies in the creation of improved archival standards (formatting, metadata structure, for example) that will facility the use of the data by scientists who are not necessarily savvy. In summary, there is so much data that the authors compare the current situation with the one immediately following the discovery of the microscope. Social scientists have so much to learn from the multitude of data available that it is crucial to address the issues of privacy and data sharing in the field.

1.9. Advancing Global Health Research Through Digital Technology and Sharing Data
Primary challenge(s): Lack of resources in general, including lack of electronic equipment for data collection, internet access and bandwidth make it extremely hard to reach remote communities in poor countries.
Primary opportunity(ies): Since these areas are so poor and vulnerable to diseases and their consequences, collection of health data (to support health data analytics) in these areas has the potential to have a huge impact towards disease prevention and effectiveness of medical care. Mobile devices with free applications for data collection, advances in satellite technology with the power to bring internet service to remote areas, and the expansion of mobile phone networks are some of the technological changes that could help bring the power of big data and analytics to less privileged regions of the world.

1.10. Access to Stem Cells and Data: Persons, Property Rights, and Scientific Progress
Primary challenge(s): In the context of stem cell research it is difficult do differentiate between what constitutes a "person vs an artifact", "information vs. material" and what's "private vs. public". For example is a stem cell line just a "thing" or does it represent personhood at any level? If it represents personhood, it is reasonable that privacy concerns should apply, otherwise, should

we care simply about ownership rights and who owns the stem cell line? The blurred lines between these distinctions pose many challenges around ownership rights as well, because intellectual property laws are quite different than laws governing ownership of physical entities.
Primary opportunity(ies): By addressing these challenges, it will be possible to make stem cells and stem cell research results more accessible for researchers and that alone will generate progress in the field.

1.11. Climate Data Challenges in the 21st Century
Primary challenge(s): Like in all other fields there is data deluge of new climate data available for researchers. Unlike many other fields however, systematic collection of climate data goes back to the 1800's with some important climate records even dating back to the 1600's. The new data is itself quite varied in its structure because it is collected and gathered in a variety of manners, such as through space-borne instruments, "reanalyzes" and simulation. Integrating the new data with old data, which in many cases have not been systematically digitalized yet is a big challenge in the field. Additionally, there is an interest in incorporating data captured as "paleoclimatic proxy records" which are records about natural entities such as corals and ice code that help understand climate change. Integrating this data clearly represents a challenge, as it is dissimilar to direct measures of climate.
Primary opportunity(ies): Improved access to better organized climate data will inform decisions and policies around extremely important aspects of our lives such as agriculture, water resource management, wildlife conservation among many others. The authors state that "climate data provide the backbone for billion-dollar decisions."

1.12. Challenges and Opportunities of Open Data in Ecology
Primary challenge(s): Ecology is a multidisciplinary discipline unlike many others. As the authors put it, it is a "synthetic" discipline. Its development draws from the very well established, mature science fields of "earth, life and social" sciences. The primary challenge for big data in ecology appears to be how to obtain and integrate data that is dispersed across these fields, and naturally heterogeneous, for it relates to different areas of knowledge. The article talks about technical challenges ("data dispersion, heterogeneity, and provenance") as well as socian/cultural challenges (data sharing is common intra-discipline, but not so much inter-discipline)
Primary opportunity(ies): Improve data sharing through cultural change that encourage collaboration; establish processes for data acquisition and management that allow for reproducible research.

1.13. The Disappearing Third Dimension
Primary challenge(s): There is no policy supporting archival and dissemination of 3D images; images are "lost" in the sense that they are not readily unavailable for research; there are ownership issues that need to be addressed.
Primary opportunity(ies): Voxel data (the 3D equivalent of pixels) currently capture high quality representations of a variety of types of entities (e.g. bones, rocks, and even soft tissue). If archival and dissemination improves, leading to better access to the data from research teams in universities, government and industry, it is likely this data will promote scientific discovery.

1. What in your mind are the significant takeaways from the perspective article "Challenges and Opportunities" (p. 692-694)

Carlos Sathler | cssathler@gmail.com
INFO-I 535 – MGMT ACCESS USE BIG DATA (Fall 2016)

The data deluge that occurred in the business world is also occurring in the science fields, bringing to the sciences the same promise of insights, discoveries and progress towards a better world. However, dealing with the huge amounts of data available for research poses several technical challenges, such as insufficient computational capacity to process huge volumes of relevant data, storage limitations, and in many cases the need to selectively discard data.

Still, some of the issues that scientist have to overcome are not purely technical; there needs to be a cultural change in the sciences. This cultural change calls for a new way of doing science that takes into account properly dealing with the huge volumes of data that are now available for research. It is in the best interest of the scientific community that this data is properly documented through metadata, properly archived and properly curated for possible future access.

Many of the questions that Science sent to 1,700 researchers for the introduction of its Feb 2011 issue centered on data sharing and data re-use. These questions seem to address two major concerns: first, the concern that all this data that has the potential to be so useful for the advancement of the sciences may not be at the disposal of the science community as a whole; second, and even worse, there is a concern that much of this data could be lost altogether, in the same way that early anthropologists burned wood from archeological sites to make coffee.