JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

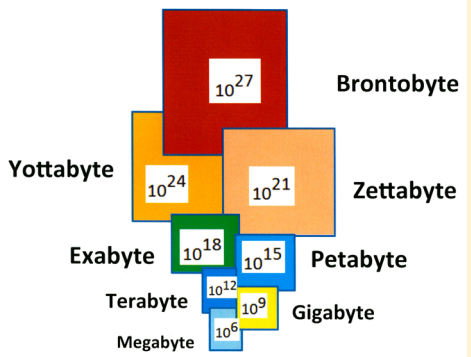
Big Data and Chemical Education

Harry E. Pence*[,†] and Antony J. Williams[‡]

[†]Department of Chemistry and Biochemistry, SUNY at Oneonta, West Oneonta, New York 13820, United States

[‡]ChemConnector, Inc., 904 Tamaras Circle, Wake Forest, North Carolina 27587, United States

ABSTRACT: The amount of computerized information that organizations collect and process is growing so large that the term Big Data is commonly being used to describe the situation. Accordingly, Big Data is defined by a combination of the Volume, Variety, Velocity, and Veracity of the data being processed. Big Data tools are already having an impact in chemical industry and research, and it is time to discuss whether at least an introduction to these developments might be incorporated into undergraduate chemical education.



KEYWORDS: *First-Year Undergraduate/General, Second-Year Undergraduate, Upper-Division Undergraduate, Chemoinformatics, Computer-Based Learning, Internet/Web-Based Learning, Student-Centered Learning*

INTRODUCTION

Big Data is one of the latest buzz-phrases that is being hyped in the world of computer information. Big Data has become a trend because many businesses seem to think that there is no limit to the benefits available for any company that learns how to use Big Data to better identify the needs and desires of their customers. There are equally strong reasons why Big Data is important in Science. The use of large data sets in science is not really new, but in the past, it has not always been called Big Data. For example, Hey et al. have edited a book on data intensive scientific research but titled the book *The Fourth Paradigm*.[1] Academic journals, like *Science*[2] and *Nature*,[3] have run special issues on Big Data in science, although neither of these provide many examples in Chemistry. In addition, scientific conferences such as the American Chemical Society meeting have hosted Big Data sessions in their CINF Division session for fall 2015 in Boston.

Many chemical educators are either unaware of the Big Data discussion or else feel that it has little relevance in Chemistry; however, this is clearly untrue. Their students are almost certainly already executing queries of at least one of the popular search engines looking for chemistry data, articles or related information. While the resulting set of URLs may be quite constrained, the reality is that the query itself is launched against an enormous amount of data aggregated and indexed by the search engine. Big Data tools are already having an impact in chemical industry and research, and it is time to discuss whether at least an introduction to these developments might be incorporated into chemical education.

DEFINING BIG DATA

The amount of computerized information that organizations collect and process is growing exponentially. This development has been made possible by a combination of several factors: the rapidly decreasing cost of computers with increasing computing power and computer memory; the globalization of many businesses that makes it essential to share information from multiple sites; the creation of software tools that can process huge amounts of information; and the proliferation of social networks. Twitter generates more than 7 Terabytes (TB) of data a day; Facebook more than 10 TB, and some enterprises already store data in the petabyte (PB) range. (Note: 1 TB is 1000 GB and 1 PB is 1000 TB.) The Library of Congress holds about 10 TB of information (see Figure 1).

David Weinberger points out that according to researchers at the UC-San Diego, Americans consumed about 3.6 ZB of information in 2008.[4] In an attempt to visualize what a zettabyte means, Weinberger uses the example of Tolstoy's *War and Peace*. This physical book is about 1296 pages in print and 6 in. thick, or 2 MB if it were in digital format. Thus, 1 ZB equals 5×10^{14} copies of *War and Peace*. It would take a photon of light traveling at 186,000 mi/s 2.9 days to go from the top to the bottom of a stack of *War and Peace* that would be equivalent to 1 ZB. This is, indeed, an unimaginable amount of data, and information scientists are already suggesting that in the near future even larger data measures, like yottabytes (YB,

Special Issue: Chemical Information

Received: July 1, 2015
Revised: January 4, 2016

DOI: 10.1021/acs.jchemed.5b00524
J. Chem. Educ. XXXX, XXX, XXX−XXX

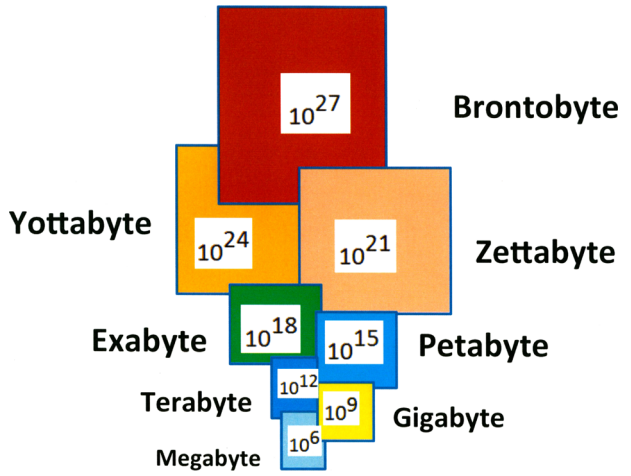


Figure 1. Data units in terms of bytes.

1000 ZB) and even brontobytes (1000 YB), will be part of our general discussions in terms of the challenges of data management. It is easy to see why the term Big Data is appropriate.

Big Data is not just about the amount of information involved. It has become common to describe Big Data in terms of four V's: Volume, Variety, Velocity, and Veracity. The need for Big Data tools depends upon the volume of stored data, but data can also be created in a bewildering variety of structured and unstructured formats. In addition, the velocity at which data are collected, analyzed or retrieved may require special software tools. Finally, the veracity or accuracy of the data is always important. Veracity problems might include duplicate records, missing or incorrect data, and insufficient information about the provenance of the data. Veracity problems may occur with any data, but they may be more difficult to detect with Big Data. In addition, the complexity of the data may require special software even though the total amount of information is not very large. Mike Gualtieri, a principal analyst with Forester Research, has pointed out that the data from one individual's sequenced DNA is only about 750 MB, but it would require 222 PB for storage of the entire population of the U.S.[5] Even analyzing the genome for a single person in order to find the best treatment for a disease would represent a massive computing problem that would require Big Data tools.

Thus far, most of the public attention has been on the use of Big Data tools for text and transaction analysis. Even if they are not aware of it, most chemists will have encountered Big Data tools without realizing it if they have used a search engine, like Google, since most search engines use a combination of Map Reduce and Hadoop to distribute a task among multiple servers to actually analyze the huge amounts of information and to determine search results.[6] More advanced tools of this type can determine how often words appear, how the sentiment changes in a document, or the people whose opinions have the most influence in a large text data set. When these tools are applied to data from social media, like a Twitter stream, they allow social scientists to gain new insights into how public opinion is shaped. These kinds of analyses have not been very interesting to chemists thus far, although they may be used to track public opinion and identify the individuals who are most influential toward chemical topics.

BIG DATA AND CHEMISTRY

Even though chemical applications of Data Analytics may not have received as much general attention, other kinds of Big Data tools are already being found to be useful in many areas of Chemistry, including molecular modeling, medicinal, environmental, and toxicological chemistry. According to Lusher and co-workers, Big Data methods have been used for over a decade in medicinal chemical research.[7] They argue that "Data-driven medicinal chemistry approaches have the potential to improve decision-making in drug discovery projects, providing that all researchers embrace the role of 'data scientist' and uncover the meaningful relationships and patterns in available data." Edwards et al. point out that environmental data may not only be large in size but may be complicated because it may consist of historical records in differing formats, different kinds of analyses, and even photographs.[8] The use of High Throughput Screening assays in Chemical Toxicology produces both structured and unstructured information that is so large and complex that it is difficult to analyze using traditional methods.[9] This is another area where Big Data provides better ways of data mining.

Big Data tools are also beginning to be adopted for industrial chemistry. The need is clear. Erickson argues that biomedical researchers are literally drowning in data.[10] He quotes Eric D. Green, director of NIH's National Human Genome Research Institute (NHGRI), as saying that, "The biomedical research enterprise is absolutely overwhelmed with data, and we cannot analyze it as fast as we can generate it." As a result, experimental results can be overlooked or repeated unnecessarily. Mullin reports according to one estimate, "... 40% of all R&D experiments are repeat runs necessitated by inefficient experimental design or inadequate IT."[11]

Thus far, most educational applications of Big Data in Chemistry seem to be at the graduate level.[12] Subdisciplines, like environmental chemistry, cheminformatics, or pharmacology, which deal with large data sets or complex data interactions, are leading the way at the graduate and research level, but it seems inevitable that as commercial software becomes more available these applications will also be introduced at the undergraduate level.

Synchronization of the research going on in several separate groups can clearly benefit from Big Data tools, and a number of vendors are acting to meet this need.[11] In 2013, Accelrys (acquired by Dassault Systems in 2014) announced the availability of a product called Experiment Knowledge Base, a laboratory informatics system that enables research and development organizations to analyze, manage and mine knowledge from their scientific experiments and data.[13] According to the press release, this software will reduce or eliminate repeat experiments, and provide greater extraction efficiency from collected data. IDBS, another scientific software company, produces E-Workbook, which consists of four components: an electronic laboratory notebook; Asset hub, which manages laboratory assets including samples, reagents, equipment and materials; ChemBook; and BioBook, both of which are designed for data capture and experiment design. The savings of time and money that result suggest that this type of product will become increasingly popular.

The reality for undergraduate chemistry students in general is that, other than queries via search engines across the expanse of online information and data, they are more commonly concerned with accessing big *databases* rather than big data.

DOI: 10.1021/acs.jchemed.5b00524
J. Chem. Educ. XXXX, XXX, XXX−XXX

As an example, at the time this is being written the CAS Scifinder database contains over 100 million unique organic and inorganic substances.[14] Meshing together the appropriate representations of a chemical compound with associated scientific papers, patents and experimental and predicted data certainly produces a large database of information, but in reality, this is likely only a few terabytes of data, i.e., not really *big*. For example, the Royal Society of Chemistry (RSC) ChemSpider database contains over 35 million unique compounds together with analytical data and, in many cases, measured experimental properties. The total amount of data in ChemSpider amounts to only 2 TB. Other major sources of chemical information available on the WWW include Reaxys, which is the largest collection of experimentally derived property data in chemistry,[15] and WorldWideScience, which allows anyone with Internet access to submit a single-query search that covers national scientific databases in more than 70 countries.[16]

There is a great deal of data available that also crosses the chemistry and biology domains, and much of this data is available as open data.[17] For example, the ChEMBL database provides access to almost 1.5 million compounds with 13.5 million bioassay data points associated with over 10,500 molecular targets. PubChem similarly provides access to a large compound library (over 68 million compounds) associated with almost a quarter of a million bioactivities. Both the ChEMBL[18] and PubChem[19] data sets can be downloaded from the relevant Web sites. Clearly, there is an increasing amount of data available for science students to access and utilize, and as a result, there are a number of possibilities for students to learn how to review and interpret data. This will be discussed in more detail later in this article.

Gibb predicts that in the near future, chemists will have a data management system that will automatically analyze the chemical literature to identify areas of recent interest and recommend which current research topics would be most likely to serve as a basis for a successful grant proposal.[20] After all, he notes, there are already impact factors for journals and individual researchers. Thus, it is not unreasonable to expect that there might also be impact factors for ideas. Whether or not science reaches this point any time soon, it is apparent that Big Data is already having an effect on chemical research and industry, so this topic should become part of the undergraduate training provided to the young people who will become the researchers of the future.

It should be noted that text-mining tools applied to chemistry articles continue to gain popularity and may ultimately make the largest contribution to expanding access to large quantities of chemistry data online. This will be especially true if such tools are utilized to mark-up scientific articles prior to release or are applied to archives of historical content for data mining. Murray-Rust and Rzepa outlined the basic principles of the value and approaches to chemical markup prior to the 21st Century.[21] In addition, software tools such as OSCAR (in its latest form known as OSCAR4[22]) and Chemical Tagger[23] have been discussed in mainstream scientific publications for many years.[24] While they are still not part of the conventional operations for the scientific publishers, interest remains high. This is indicated by the number of chemical entity extraction tools involved in the CHEMDNER challenge examining drugs and chemical names extraction.[25] These approaches are already being applied to the extraction of physicochemical properties from U.S. Patents,[26] as well as the extraction of NMR spectra from patents and publications.[27] Ultimately, it is expected these data will be available on the Internet via the appropriate chemistry databases and provide access to Big Data collections for the chemistry community.

BIG DATA AND CHEMICAL EDUCATION

Despite concerns about both the computer resources and the classroom time required for a full-scale treatment of the Big Data, chemistry departments need to find a way to introduce students to these techniques. Thus far, there have been relatively few published reports that deal specifically with the introduction of data handling techniques into the undergraduate Chemistry curriculum, although the proposed changes to the ACS Guidelines for Bachelor's Degree Programs do specifically mention the need to introduce a new skill on data and information management.[28] Reisner, Vaughn, and Shorish have described an exercise designed to improve how students named and organized data files,[29] and several educators, including Abrams,[30] Soulsby,[31] and Bennett and Pence[32] have used cloud computing to manage laboratory data. None of these has really applied Big Data tools to the undergraduate experience.

One important initial step might be to make sure that the InChI naming rules are included in the curriculum. Southan points out that, "The InChIKey indexing has therefore turned Google into a *de-facto* open global chemical information hub by merging links to most significant sources, including over 50 million PubChem and ChemSpider records."[33] This would allow students to become familiar with an almost ubiquitous way to do Big Data searches without the need to introduce specialized tools.

As discussed earlier, there are large amounts of chemistry data available online for students to use for the purposes of research and investigation. Should students wish to consider utilizing open source software tools to manipulate the available data, there are numerous cheminformatics software tools available for download. These include ChemAxon software tools[34] (freely available to students), the Indigo cheminformatics toolkit from GGA Software Services,[35] and various software components from companies including ACD/Labs.[36] With an abundance of software and data available online, there are numerous possibilities to investigate scientific data.

It is likely, however, that many faculty will not want to install and configure software, or learn the programmatic details for deploying cheminformatics software. As this article is being written, efforts are underway to provide training for students in cheminformatics standards, tools and platforms, by means of an online cheminformatics learning course (OLCC) led by Belford.[37] Despite these efforts to instill skills and understanding, the majority of student scientists will likely want to apply software and/or Web sites hosting large data collections to solving problems rather than spending time learning how to set them up.

For many educators, the first experience with Big Data may come not from applications in their specific disciplines but rather from efforts by colleges to improve student performance and retention.[38] Colleges have long had more information about students than could be used effectively. Learning Management Systems, online courses, and online homework assignments collect much more information about student behavior than has been previously put to use. More and more campuses are using Big Data applications to convert these

DOI: 10.1021/acs.jchemed.5b00524
J. Chem. Educ. XXXX, XXX, XXX−XXX

reams of data into actionable recommendations that are hoped to improve student performance.[39]

Seven years ago, Arizona State University was an early user of data-analytics with its eAdvisor system to improve the graduation rate of low income students. As a result, its four-year graduation rate for lower-income students is reported to have increased from 26% to 41%.[40] Other campuses are using Big Data techniques to monitor student attendance, keep track of how often students turn in required homework assignments, or make sure students are making the proper choice of courses for their major. Ball State University in Indiana has taken to heart the observation that students who are more engaged with college activities are more likely to graduate and so even monitors whether students are swiping in with their ID cards to campus-sponsored parties.[41]

CAN STUDENTS CONTRIBUTE TO THE BIG DATA OF CHEMISTRY?

The past decade has seen an explosion in participation in the social web and we owe much to this era of contribution and "crowdsourcing" to the development of sites such as Wikipedia[42] and to the availability of reviews for restaurants, books and movies contributed primarily by members of the public. There is a similar possibility to contribute to chemistry in terms of data, online articles, models and software code. While there are many databases online for chemists to use as a data source, only a small number allow direct contributions. Two of the most popular are the ChemSpider and PubChem databases. ChemSpider allows contributions to the database that may include anything from a single compound to large deposits of chemical property data, spectral data, and physicochemical properties. This requires one to establish a free RSC account.[43] Pubchem is primarily focused on hosting chemical compounds and associated assay data.[44] By the contribution of data into these systems, chemists are exposing their data to very large audiences. ChemSpider averages over 40,000 unique users per day,[26] and the data they contribute is integrated into the Google database.

Chemistry students can also contribute to the online data streams by means of a new form of scientific publication commonly called "micropublications".[45] These are many chemistry-based blogs online that include peer review occurring by means of comments on the posts. Examples include TotallySynthetic (now part of organicsynthesis.org),[46] where literature syntheses are analyzed in detail; the BRSM Blog;[47] and many others. Other more formal micropublishing platforms are appearing online, including Chips and Tips from the Royal Society of Chemistry[48] and ChemSpider SyntheticPages (CSSP)[49] also from the Royal Society of Chemistry.

CSSP is a micropublishing platform for chemical syntheses managed by a group of synthetic chemists who review the submissions. Generally, these are made available for public comment within a few days for the community to provide feedback and are highly accessed according to statistics collected by the CSSP leaderboards.[50] The chemicals discussed in the syntheses are deposited into ChemSpider, usually with the associated analytical data. These data become a part of the data index available through the public search engines. Efforts are presently underway to bring together the ChemSpider content, the content from the micropublishing platforms, and large-scale reaction and property content extracted using text-mining approaches in order to make it available through the RSC Data Repository.[51] It is likely that the drive of funding agencies to push for release of Open Data,[52,53] the availability of open repositories for sharing data (e.g., FigShare, Dryad and others), and the adoption of Open Notebook Science[54] will result in an increase in the amount of chemical Big Data available to users.

CONCLUSION

Big Data tools are already being used increasingly in industry and medicinal, environmental, and toxicological research. It seems likely that this usage will increase significantly in the future. Thus, it would be very helpful to incorporate at least an introduction to this topic into the undergraduate curriculum. At a minimum, students should learn best practices for giving descriptive names to their files, saving files in nonproprietary formats, and including appropriate metadata so that users will be able to understand what the data is and how it was collected. Several recent books go into greater depth discussing data management and can be suggested to readers who wish to go beyond the material in this article. *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success* is a recent text that provides an excellent overview of data management.[55] In *Big Data, Little Data, No Data: Scholarship in the Networked World*, Borgman describes the interactions of people, practices, technologies, institutions, material objects, and relationships that are necessary to give meaning to all data, including Big Data.[56]

It is always difficult to add a new subject into a program that is already crowded with essential material, but the importance of data science argues that this addition is necessary. The obvious place for this insertion would a chemical literature course, although portions of this material might be introduced into appropriate undergraduate courses or as part of the undergraduate research experience.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pencehe@oneonta.edu.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Hey, T.; Tansley, S.; Tolle, K. *The Fourth Paradigm*; Microsoft Research: Redmond, WA, 2009.

(2) Special Online Collection: Dealing with Data. http://www.sciencemag.org/site/special/data/ (accessed Oct. 31, 2015).

(3) Big Data. *Nature* **2008**, *455* (1), http://www.nature.com/news/specials/bigdata/index.html.

(4) Weinberger, D. *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room*; Basic Books: New York, NY, 2012; p 7.

(5) Gualtieri, M. Is 750MB Big Data? http://blogs.forrester.com/mike_gualtieri/12-12-05-is_750mb_big_data (accessed June 4, 2014).

(6) Langit, L. Hadoop MapReduce Fundamentals 1 of 5. https://www.youtube.com/watch?v=7FcMhTTG1Cs (accessed June 21, 2015,).

(7) Lusher, S. J.; McGuire, R.; vanSchaik, R. C.; Nicholson, C. D.; deVlieg, J. Data-Driven Medicinal Chemistry in the Era of Big Data. *Drug Discovery Today* **2014**, *19* (7), 859−868.

(8) Edwards, M.; Aldea, M.; Belisle, M. Big Data is Changing the Environmental Sciences; *Environmental Perspectives* **2015**, *1*, http://www.exponent.com/files/Uploads/Documents/Newsletters/EP_2015_Vol1.pdf.

DOI: 10.1021/acs.jchemed.5b00524
J. Chem. Educ. XXXX, XXX, XXX−XXX

(9) Zhu, H.; Zhang, J.; Kim, M. T.; Boison, A.; Sedykh, A.; Moran, K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants. *Chem. Res. Toxicol.* **2014**, *27*, 1643−1651.
(10) Erickson, B. E. Drowning in Data. *Chem. Eng. News* **2013**, *91* (7), 40−1.
(11) Mullin, R. Breaking Big. *Chem. Eng. News* **2013**, *91* (42), 19−21.
(12) Mullin, R. Enter the Data Scientist. *Chem. Eng. News* **2012**, *90* (46), 11−14.
(13) Accelrys Strikes Gold with Industry-First Experiment Knowledge Base for R&D. http://www.prnewswire.com/news-releases/accelrys-strikes-gold-with-industry-first-experiment-knowledge-base-for-rd-211309591.html (accessed May 30, 2015).
(14) CAS Registry - The Gold Standard for Chemical Substance Information. https://www.cas.org/content/chemical-substances (accessed June 25, 2015).
(15) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; David Evans, D. The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds,; American Chemical Society: Washington, DC, 2014; pp 127−148.
(16) World Wide Science Home Page. http://worldwidescience.org/ (accessed Oct. 31, 2015).
(17) ChemSpider: Search and Share Chemistry. http://www.chemspider.com/ (accessed June 25, 2015).
(18) ChEMBL. https://www.ebi.ac.uk/chembl/downloads (accessed June 2, 2015).
(19) PubChem Download Service. https://pubchem.ncbi.nlm.nih.gov/pc_fetch/pc_fetch.cgi (accessed June 25, 2015).
(20) Gibb, B. C. Big (Chemistry) Data. *Nat. Chem.* **2013**, *5*, 248−249.
(21) Murray-Rust; Rzepa, H. S. Chemical Markup, XML and the Worldwide Web. 1. Basic Principles, i. *J. Chem. Inf. Model.* **1999**, *39* (1999), 928−942.
(22) Jessop, D. M.; Adams, S. E.; Willighagen, E. L.; Lezan Hawizy, L.; Murray-Rust, P. OSCAR4: A Flexible Architecture for Chemical Text-mining. *J. Cheminform.* **2011**, *3*, 41; http://www.jcheminf.com/content/3/1/4110.1186/1758-2946-3-41.
(23) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A Tool for Semantic Text-mining in Chemistry. *J. Cheminf.* **2011**, *3* (17), 17 http://www.jcheminf.com/content/3/1/17,.
(24) Stutchbury, N. Chemical Information Mining: Possibilities and Pitfalls. http://www.rsc.org/images/ChemInfoMining_tcm18-153536.pdf (accessed June 21, 2009,).
(25) Krallinger, M.; Leitner, F.; Rabal, O.; Vazquez, M.; Oyarzabal, J.; Valencia, A. CHEMDNER: The Drugs and Chemical Names Extraction Challenge. *J. Cheminf.* **2015**, *7* (Suppl 1), S1 http://www.jcheminf.com/content/7/S1/S1,.
(26) Williams, A.; Lowe, D. Activities at the Royal Society of Chemistry to Gather, Extract and Analyze Big Datasets in Chemistry. http://www.slideshare.net/AntonyWilliams/activities-at-the-royal-society-of-chemistry-to-gather-extract-and-analyze-big-datasets-in-chemistry (accessed June 25, 2015).
(27) Williams, A.; Pshenichnov, A.; Corbett, P.; Lowe, D.; Coba,C. Providing Access to a Million NMR Spectra via the Web. http://www.chemconnector.com/2015/03/22/providing-access-to-a-million-nmr-spectra-via-the-web/ (accessed June 21, 2015).
(28) ACS Committee on Professional Training White Paper: Proposed Changes to the ACS Guidelines and Evaluation Procedures for Bachelor's Degree Programs. http://www.acs.org/content/dam/acsorg/about/governance/committees/training/guidelines-white-paper.pdf (accessed Dec. 17, 2014).
(29) Reisner, B. A.; Vaughan, K. T. L.; Shorish, Y. L. Making Data Management Accessible in the Undergraduate Chemistry Curriculum. *J. Chem. Educ.* **2014**, *91* (11), 1943−1946.
(30) Abrams, N. M. Combining Cloud Networks and Course Management Systems for Enhanced Analysis in Teaching *Laboratories*. *J. Chem. Educ.* **2012**, *89* (4), 482−486.
(31) Soulsby, D. Using Cloud Storage for NMR Data Distribution. *J. Chem. Educ.* **2012**, *89* (8), 1007−1011.
(32) Bennett, J.; Pence, H. E. Managing Laboratory Data Using Cloud Computing as an Organizational Tool. *J. Chem. Educ.* **2011**, *88* (6), 761−763.
(33) Southan, C. InChI in the Wild: an Assessment of InChIKey Searching in Google. *J. Cheminform*, 2013, *5* (10), http://www.jcheminf.com/content/5/1/10.
(34) ChemAxon: My Academic License. https://www.chemaxon.com/my-chemaxon/my-academic-license/ (accessed June 26, 2015,).
(35) epam Lifesciences. http://lifescience.opensource.epam.com/indigo (accessed Dec 29, 2015,).
(36) ACD Labs: Chemistry Software http://www.acdlabs.com/resources/freeware/ (accessed June 26, 2013).
(37) Belford, R. E.; Wild, D. J.; McEwen, L. R.; Williams, A. J. Cheminformatics OLCC: A CCCE Project in Intercollegiate Teaching and Learning. http://www.ccce.divched.org/P4Fall2013CCCENL (accessed June 26, 2015,).
(38) Bichsel, J. Analytics in Higher Education. http://net.educause.edu/ir/library/pdf/ers1207/ers1207.pdf (accessed June 2, 2015).
(39) Blumenstyk, G. Blowing Off Class? We Know. *New York Times*, http://www.nytimes.com/2014/12/03/opinion/blowing-off-class-we-know.html (accessed Dec 3, 2014).
(40) How Does eAdvisor Benefit Me? https://eadvisor.asu.edu/ (accessed June 2, 2015).
(41) Ball State Using Data to Boost Retention. *The JGJournal* [Online], **2015**. http://www.journalgazette.net/news/local/indiana/Ball-State-using-data-to-boost-retention-4783174 (accessed Jan 11, **2016**).
(42) Wikipedia. https://www.wikipedia.org/ (accessed June 26, 2015).
(43) Create a Royal Society of Chemistry account. http://www.chemspider.com/DepositionsMenu.aspx (accessed June 125, 2014).
(44) Wecome to PubChem Upload. https://pubchem.ncbi.nlm.nih.gov/upload/#welcome (accessed June 25, 2015).
(45) Nielsen, M. Micropublication and Open Source Research. http://michaelnielsen.org/blog/micropublication-and-open-source-research/ (accessed June 25, 2015).
(46) Taber, B. F.; Lambert, T. Organic Syntheses: Organic Chemistry Portal. http://organic-chemistry.org/totalsynthesis/ (accessed Dec 30, 2015).
(47) B.R.S.M. Blog: Total Syntheses. http://brsmblog.com/category/total-synthesis/ (accessed June 25, 2015).
(48) Walker, G. Chips and Tips. http://blogs.rsc.org/chipsandtips/ (accessed June 25, 2015).
(49) ChemSpider Synthetic Pages. http://cssp.chemspider.com/About.aspx (accessed June 25, 2015).
(50) ChemSpider Synthetic Pages: Leaderboard. http://cssp.chemspider.com/Leaderboard.aspx (accessed June 25, 2015).
(51) Williams, A. Activities at the Royal Society of Chemistry to Gather, Extract and Analyze Big Datasets in Chemistry. http://www.slideshare.net/AntonyWilliams/activities-at-the-royal-society-of-chemistry-to-gather-extract-and-analyze-big-datasets-in-chemistry (accessed June 25, 2015).
(52) Reynolds, E.; Strassel, G. Guidelines for OSTP Data Access Plan. http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/ostp.html (accessed June 25, 2015).
(53) Pampel, H.; Dallmeier-Tiessen. S. Open Research Data: From Vision to Practice. http://book.openingscience.org/vision/open_research_data.html (accessed June 25, 2015).
(54) Wikipedia. Open Notebook Science. http://en.wikipedia.org/wiki/Open_notebook_science (accessed June 25, 2015).
(55) Briney, K. *Data Management for Researchers: Organize, Maintain and Share your Data for Research Success*; Pelagic Publishing: Exeter, U.K., 2015.
(56) Borgman, C. L. *Big Data, Little Data, No Data: Scholarship in the Networked World*; The MIT Press: Cambridge, MA, 2015.

DOI: 10.1021/acs.jchemed.5b00524
J. Chem. Educ. XXXX, XXX, XXX−XXX