Data Science and Online Education

Geoffrey Fox, Sidd Maini, Howard Rosenbaum, David Wild
School of Informatics and Computing
Indiana University
Bloomington, IN, 47408, USA
gcf@indiana.edu, smaini@indiana.edu, hrosenba@indiana.edu, djwild@indiana.edu

***Abstract*—we discuss the Data Science program at Indiana University, which is offered in both traditional residential and online formats. We describe Data Science, our chosen curriculum and its motivation. We describe experience in online delivery for both traditional lectures and online programming laboratories, and discuss implications for the technology used.**

Keywords—Data Science; big data; MOOC; online education; curriculum; clouds; programming laboratory

I. Introduction

Here we discuss our experiences in offering a Data Science curriculum and in delivering the courses online in a MOOC (Massive open online course) like fashion. Sec. II introduces Data Science and big data in general while in Sec. III we describe the Data Science curriculum. Sec. IV discusses our experience in online education while Sec. V describes delivery technology and lessons from its use. We note that our main experience is online classes with 20-100 students and not the large audiences (tens of thousands) of the most popular MOOCs. Conclusions are in Sec. VI

As context, we describe the School of Informatics and Computing at Indiana University (IU) as its broad structure supports a Data Science program. The School of Informatics was established in 2000 as first of its kind in the United States while Computer Science, established much earlier in 1971, became part of the school in 2005. Information and Library Science was established in 1951 and became part of the school in 2013. Now named the School of Informatics and Computing or SOIC, the School has an unusually broad perspective from curation, ethics through complex systems, machine learning and parallel programming. SOIC has approximately 100 faculty, 1500 undergraduates, 650 Masters students and 250 PhD students.

Data Science was added January 2014 and is set up as a cross disciplinary program made up of faculty from the three departments in SOIC. The school will add a fourth department in intelligent systems engineering over the next year. Data Science already has members from outside the School – in particular from the Statistics department which is in the College of Arts and Sciences. In the near future we expect an expansion to include faculty in domains from social science to physics where Data Science is of great importance.

The well-known McKinsey report [1] was an important motivator in establishing a Data Science program. It forecast that by 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions. Our Data Science program is designed to satisfy both job opportunities with respectively technical and decision-maker paths that both lead to Data Science degrees.

At the time Data Science was being discussed, there was growing attention to MOOCs and more generally online education with Georgia Tech's online Masters in Computer Science, both of which motivated us [2]. So we started the program in a purely online form with a certificate requiring 4 courses and we followed Georgia Tech in discounting the tuition. The certificate had advantage of requiring a shorter approval cycle than the full Master's degree.

II. Data Science and Big Data

A. Data Science

The question is often raised as to "What is Data Science?" Given the breadth of the field illustrated by the McKinsey report, a precision definition is not likely to be possible but we were guided by the work of the NIST Public Big Data working group [3]. Data Science lies at intersection of fields like statistics and data mining, analysis, algorithms, modelling, visualization, programming, systems (engineering), curation and data management with an understanding of how to apply these techniques across domains. Data Science has many flavors differing often according to university unit offering it. We see information science, computer science, statistics and business schools/departments hosting Data Science. This seems to us healthy and there is no need to standardize now. Further there seems no difficulty in designing curricula built on interesting fundamental/core knowledge (not just fads) that will prepare students for many promising jobs.

B. Big Data

There are related questions: "What is big data?", "What is data-intensive science?"; "What is the fourth paradigm of science?" Of course there has always been data and much of science has been observational and heavily based on data. However data has increased both in volume and breadth. We are familiar to using physics models to interpret data from particle accelerators; the LHC at CERN is larger than previous accelerators and the data sizes much larger and more complex, but the analytic approach is not so different from that of 50 years ago. However now we have Amazon and Netflix gathering data on user preferences and using this to build a

model (collaborative filtering) to predict what others might like. Another interesting feature of today's big data is the diversity of spaces. The LHC dealt in space defining multi-particle states; Amazon in the space of users or items; search in the space of documents defined over so-called bags of words (axes in space). In the past Fox worked in a field called particle physics phenomenology comparing data with simple theoretically motivated models; in retrospect, this was just a form of Data Science and so perhaps that's why he likes this new field. The current big data field has spurred incredible developments in infrastructure (clouds, Hadoop, Spark etc.) and algorithms (machine learning). These are remarkable but have roots in earlier work in statistics, algorithms and parallel computing. So Big Data is both the same and very different!

C. Computational Science

Computational science [4-6] (sometimes called scientific computing) has important similarities to Data Science but with a simulation rather than data analysis flavor. Both are positioned at the interface between computer science and application domains with similar interdisciplinary characteristics with perhaps applied mathematics swapping with statistics in the methods area. Although a great deal of effort went into computational science with meetings and several academic curricula/programs, it didn't take off [7-9].

Will Data Science suffer a similar promising start and then not make it? From experience, the number of students interested in computational science was not large and the academic job opportunities were not great. It is clear that Data Science has more jobs and so one can expect that it will do much better. We could in fact integrate these concepts into a single field, as they have many ideas such as parallel computing in common.

III. Data Science Curriculum

The current Data Science offering [10] from IU consists of 3 approved degrees: a) a purely online 4-course Certificate in Data Science has been running since January 2014; b) a campus wide Ph.D. Minor in Data Science (also 4 courses) and c) a Masters in Data Science (10-course) approved October 2014 and starting with a few students in Spring 2015. The Masters is offered as a residential program (you come to IU) or online as in Certificate. There is also a hybrid model where you mix residential and online. We are exploring a PhD in Data Science and hope to propose this in next few months.

This fall is the first semester that we can judge student interest in our Masters curriculum and we see encouraging results; we have had approximately 300 applicants of which one third are online or hybrid. Given the excitement over MOOCs and online education perhaps most interesting is that two thirds of our applicants are traditional residential students. This high student interest makes this curriculum quite significant in the school; only the basic computer science Masters has larger numbers while Data Science pages on our web site have higher numbers of page views than all other curricula SOIC pages showing topical value of field.

A. Types of Students and Their Needs

We have already noted the breadth of Data Science and in Section I mentioned the two distinct paths students can take, which is expanded in Part B. We can further identify the three types of students:

- Professionals wanting skills to improve job or required by employee to keep up with technology advances
- Traditional students interested in IT Masters
- Students in non IT fields wanting to do domain specific Data Science

What do students want? They are interested in the degree with a good curriculum but for many a key consideration is ability to follow the degree with Optional Practical Training (OPT) which allows graduated students with an appropriate visa to work for US companies. OPT requires a residential component in education and this can be lowering interest in the online option. It has been noted that online degrees have a larger fraction of domestic students than residential degrees and we see that in data science. A typical online participant is a USA professional aiming to improve credentials and/or switch jobs.

B. Requirements for a Masters Degree in Data Science

The Data Science Masters degree currently has no required courses in order to allow a broad range of faculty and student involvement. There is a general Masters degree track and specialized tracks are allowed but currently only a computer science oriented track "Computational and Analytic Data Science" has been designed. We expect other tracks including one focused in biomedical and life sciences area. Note tracks are visible as they appear on final degree transcript. The two pathways (technical, decision-maker) are informal corresponding to different course choices leading to the same degree. Students must pass 10 courses to be awarded a Masters in Data Science with the following constraints. We have a list of sanctioned courses divided into four areas described in detail in Sec. IIIC. Students must choose one course from two of the three technological areas namely Data analysis and statistics; Data lifecycle (includes handling of research data); and Data management and infrastructure. Further a student must take one course from an (big data) application course area. The other 7 courses must be chosen from list maintained by Data Science Program or from outside this list with permission. A capstone project is optional. Note that all students are assigned an advisor who approves course choice.

C. The Four General Data Science Masters Course Areas

As noted above we classify courses into four areas that are described broadly below with available courses given online [11]. Note that only a few of these are currently available online; we are increasing this number as program grows.

1. Data analysis and statistics area: gives students skills to develop and extend algorithms, statistical approaches, and visualization techniques for their explorations of large scale data. Topics include data mining, information retrieval, statistics, machine learning, and data visualization and are

examined from the perspective of big data, using examples from the application focus areas described in category 4.

2. Data lifecycle area: gives students an understanding of the data lifecycle, from digital birth to long-term preservation. Topics include data curation, data stewardship, issues related to retention and reproducibility, the role of the library and data archives in digital data preservation and scholarly communication and publication, and the organizational, policy, ethics and social impacts of big data.

3. Data management and infrastructure area: gives students skills to manage and support big data projects. Data have to be described, discovered, and actionable. In Data Science, issues of scale come to the fore, raising challenges of storage and large-scale computation. Topics in data management include semantics, metadata, cyberinfrastructure, cloud computing, databases, document stores, and security and privacy.

4. Big data application domains area: gives students experience with data analysis and decision making and is designed to equip them with the ability to derive insights from vast quantities and varieties of data. The teaching of Data Science, particularly its analytic aspects, is most effective when an application area is used as a focus of study. The degree will allow students to specialize in one or more application areas which include, but are not limited to Business analytics, Science informatics, Web science, Social data informatics, Health and Biomedical informatics.

D. Computational and Analytic Data Science Track

This computer science oriented track has a slightly different structure [12] with a core set of courses. CSCI B503 Analysis of Algorithms is required while the student must take two courses out of CSCI B561 Advanced Database Concepts, STAT S520 Introduction to Statistics and CSCI B555 Machine Learning or INFO I590 Applied Machine Learning. The remaining courses must be selected from categories: data systems, data analytics and applications which are similar to those in Sec. IIIC and defined precisely online [12].

E. Typical Courses Available for Masters Degree

Here we give three possible sets of courses that could be used to be awarded a Masters. These are labelled by home department CSCI (Computer Science), INFO (Informatics), ILS (Information & Library Science), STAT (Statistics).

1) General Masters: Technical Pathway

INFO I523: Big Data Applications and Analytics
STAT S520: Intro to Statistics
CSCI B649: Cloud Computing
INFO I524: Big Data Software and Projects
ILS Z637: Information Visualization
STAT S670: Exploratory Data Analysis
CSCI B679: High Performance Computing
CSCI B555: Machine Learning
STAT S670: Exploratory Data Analysis
INFO I590 Complex Networks and their Applications

2) General Masters: Decision-maker Pathway

INFO 523: Big Data Applications and Analytics
ILS Z604 Big Data Analytics for Web and Text
STAT S520 Intro to Statistics
INFO I590: Management, Access, and Use of Big Data
ILS Z637: Information Visualization
ILS Z653: Semantic Web
ILS 605: Internship in Data Science
ILS Z604 Data Curation
ILS Z604 Scholarly Communication
INFO I590: Data Provenance

3) Computational and Analytic Data Science track

CSCI B503 Analysis of Algorithms
CSCI B561 Advanced Database Concepts
STAT S520 Introduction to Statistics
CSCI B649 Cloud Computing
ILS Z534: Information Retrieval: Theory and Practice
CSCI B555 Machine Learning
CSCI Y790: Independent Study in Data Science
CSCI B565 Data Mining
INFO I520 Security for Networked Systems
CSCI B656: Web mining

IV. Online Education and MOOCs

A. Online Education and MOOC Background

There are several reasons why online education is growing in popularity with cost a primary driver. This is clearly important for basic courses with large student interest such as Introduction to Computer Science. However there is another course type where online delivery is relevant; that is courses that are not offered at many universities and so the only way many students can take them is through online access. Data Science falls into this course type. It has broad interest and still many universities do not offer programs in Data Science. Further we can distinguish online education from MOOCs where former has classes similar in size to residential courses (10's to 100's) and the latter much larger. Size and cost imply that MOOCs need different approaches to assessment but online education does not need new approaches here as even for residential classes, homework is typically submitted electronically. MOOCs suffer from a high attrition rate and one approach in making MOOCs or online learning much more effective is creating new and innovative pedagogical models.

So MOOCs are a disruptive force in the educational environment because they change business model as can be seen with Coursera, Udacity, Khan Academy and many other examples. However MOOCs have courses and technologies and the latter are partly shared with online education and here the basic ideas are quite old but work much better these days due to tremendous advances in Web2.0/Cloud technologies (network, software, compute infrastructure) and supporting areas such as streaming video and discussion groups. In spite of these changes the core technology for online education is just a learning management system and one can either use traditional approaches like Blackboard or newer approaches typically aimed at MOOCs but applicable to all online education. Here we use open source where Google Course Builder [13] and OpenEdX [14] are best known systems. Our current work uses Course Builder but we switch to OpenEdX

this fall [15] for one class with an expectation that this will be the future direction; a plan that Google has in fact endorsed.

B. Some Online Data Science Classes Taught

Fox has taught two classes in Data Science; in different modes but always with a major online aspect. The first class is Informatics I523: BDAA: Big Data Applications & Analytics which was originally called X-Informatics [16, 17]. This is in the decision maker track and requires a small amount of Python work but no sophisticated programming. It contains about 38 hours of video mainly discussing applications (The X in X-Informatics or X-Analytics) in the context of big data and clouds. The second class is Informatics I524: BDOSSP: Big Data Open Source Software and Projects [18]. This contains about 27 Hours of video discussing Big Data software and the use of our local cloud, FutureSystems [19]. It stresses learning to architect and specify big data systems rather than solving particular problems.

Both classes are divided into sections (coherent topics), units (equivalent to lectures) and lessons. The latter are chosen to be about 5-20 minutes in length; short enough that viewers can maintain attention. Classes have been offered in four modes a) MOOC with no grades or credits, b) Purely online class with credit for online students c) Purely online class with credit for students in residence and d) Largely online class for residential students but additional face-to-face weekly sessions. The latter is a type of flipped classroom course.

The BDAA class videos covered Motivation/Introduction (What are Big Data, Clouds, Data Analytics and X-Informatics), Health Informatics, Sports Analytics, 51 NIST use cases [20] (and consequent features of Big Data), Physics Informatics (Discovery of Higgs and statistics of events), Web search and Information Retrieval, e-commerce, Internet of Things and Sensors, Remote sensing (Radar Informatics), Parallel Computing, Cloud computing for Big Data, with technologies like Python, Use of OpenStack (optional), recommender engines, clustering, visualization, MapReduce and PageRank.

The BDOSSP videos included a Big Data overview; a summary of the ~350 so-called HPC-ABDS software components [21] followed by details with about one slide per component; a discussion of 11 different Big Data software integration patterns; features of Big Data applications with some cases done in more detail (Internet of Things and Image-based problems) and the NIST study [12]; key concepts like NoSQL and machine learning; Clouds compared to HPC. This was accompanied by an extensive tutorial on the use of OpenStack and virtual clusters to host Big Data problems [22].

C. Big Data MOOC

The BDAA class was offered as a MOOC starting November 2013 and then updated and reopened December 2014. This was open to everybody and interestingly used no University resources with a basic Course Builder site hosted on a Google cloud and videos hosted on YouTube. It is one of two SoIC MOOCs named in the "7 great MOOCs for techies" by ComputerWorld [23]. It has only attracted about 4000 enrolled students and we attribute this partly to the lack of technology pizzazz; EdX and Coursera have much greater visibility and MOOC enrollments. Students come from 108 countries with leading percentages 29% USA; 26% India; 5.1% Brazil; and 3.7% each for France, Spain and United Kingdom. Possible Chinese interest would be limited by our use of Google technologies that cannot be accessed there. There is a broad age distribution with an average age of 34 years.

V. Technologies for Online Education

We now discuss in detail some technology lessons that we have learnt. What special features do learning management systems need to support the type of classes we have discussed? We divided this by functionality: video preparation and web hosting, synchronous and asynchronous interaction with students, homeworks and quizzes and programming laboratories. We avoid the possible face to face interactions for residential classes using online technologies

A. Video Preparation and Web Hosting

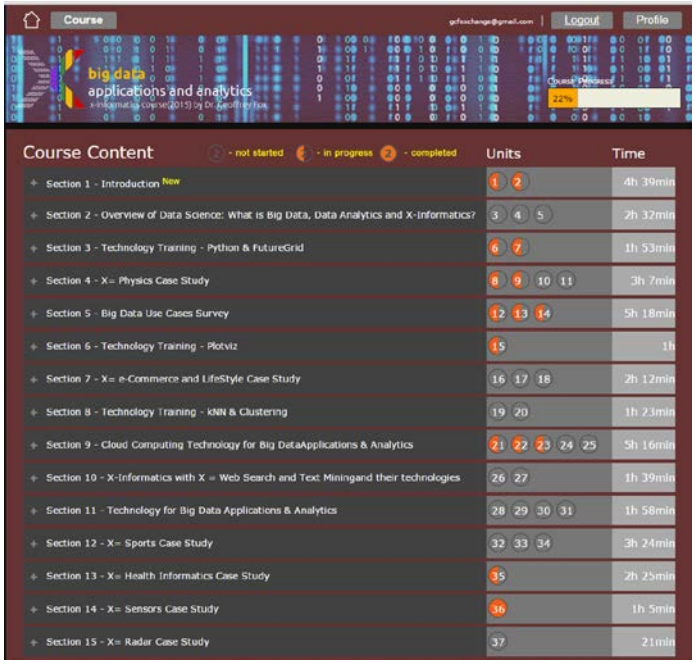

Fig. 1. Part of Spring 2015 BDAA class home page.

Clearly any online class must have a web site with a key feature that video streaming is supported. However in our work, the streaming is always occurring off site at a Google (YouTube) or Microsoft (Office Mix) page. Course Builder can prepare a very well organized page shown in Fig. 1, supporting (with our enhancements [24]) the 4-way Course-Section-Unit-Lesson Hierarchy with Spring 2015 BDAA class having 15 sections, 37 units and around 250 lessons. Fig. 2 shows a drill down with embedded video playing in page; students can switch to richer YouTube interface if they prefer. Note the rather mundane slide plus talking head format.

Our original videos were prepared by Adobe Presenter and edited by Camtasia. Later we adopted Office Mix [25] for PowerPoint presentations which integrates the video with the slides and is much easier to manage for editing than the

original approach although videos used in programming tutorials (without PowerPoint) do not benefit from this.

As we discuss the different capabilities we notice a plethora of approaches – YouTube, Google Sites, Microsoft, Social media, several choices of forums and mailing lists, Skype, Google Hangouts, email, computer system support/help systems, University homework submission. This creates a chaotic environment not enjoyed by the students or instructor.

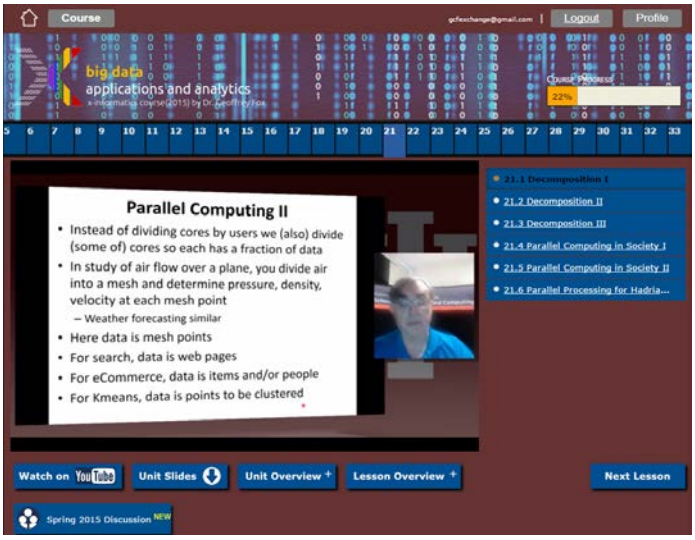

Fig. 2. Spring 2015 BDAA class showing the first lesson of unit 21 (Parallel Computing).

B. Synchronous and Asynchronous Interaction with Students

Each offering had a separate set of electronic discussion forums which were used for class announcements and for assigned discussions. The students were given participation credit for posting here and based on initial feedback, we encouraged even greater participation through students both posting and commenting in forums. Subjects for BDAA forums discussion included topics such as "Will physical shopping malls survive e-commerce?", "What fields cannot exploit big data?"

So far we have used familiar social media forums such as Google Community Groups and we have not used the forums available in learning management systems as those in the Indiana University learning management system IU-LMS Oncourse (superseded by Canvas) seemed clumsy. In fact in our last BDOSSP course, the online students used sporadically both their assigned IU email and the IU-LMS forums. We improvised to instead use Google group email lists and the students' original email. As well as discussion topics, all class announcements were made in the Instructor forum category although no sensitive material such as returned homework was posted on discussion site.

For the purely online offering, we supplemented the asynchronous approach described above with real-time interactive sessions offered once a week. Given varied time zones and weekday demands on students, these were typically held around 1pm Eastern on Sundays; alternatively in our last class, we repeated interactive session – either around 9am Eastern in the morning or 7pm in the evening. We used Google Hangouts or Adobe Connect for these sessions. Google Hangouts are conveniently scheduled from community page and offer interactive video and chat capabilities that were well received. Other technologies such as Skype or Google Hangout are obviously also possible with real time video for all participants limited to about a dozen people. However chat and streaming video from instructor is satisfactory for larger sessions. These synchronous sessions focused on general Data Science issues, the mechanics of the class and for BDOSSP common issues in using our OpenStack cloud.

C. Homeworks and Quizzes

Quizzes are online within Google Course Builder for the MOOC with peer assessment. In the credit offerings, all graded material (homework and projects) is handled identically to residential classes through IU-LMS. IU-LMS was additionally used to assign which videos should be watched each week and the discussion forum topics described in Sec. VB.

We found that the students in the non-residential classes were on a variable schedule as they were typically working full time and had many distractions; one for example had faculty position interviews. So considerable latitude had to be given for video and homework completion dates.

This is an area where large MOOCs see very different issues from the graded online classes of up to about 100 students where we have most experience.

D. Programming Laboratories

The programming for BDAA could be performed on a laptop although we also offered as a cloud service. However BDOSSP definitely required a serious programming laboratory. Now this should be straightforward as computers are typically accessed remotely and have well developed (largely asynchronous) help or ticket systems. The last offering of BDOSSP class involved 40 online students from around the world. One unexpected difficulty was that their employer policies made it impossible for some of them to access Indiana University computers from their business location. Many resorted to Internet cafes which increased our security risk. The many time zones and variable job demands, made it impossible to have communal help sessions while students chose interactive help at times outside our normal business hours.

This class highlighted value of very structured organization of material seen in BDAA course, as students tended to jump around our current hyperlinked tutorials. Regular videos and homeworks need less support and so do not suffer as much from the irregular patterns of the online cohort. The programming lab material [22] will be reorganized for our next offering to mitigate the consequences of the inevitable erratic online profile. We will review security issues consequent from heavy use of rather dangerous client locations.

VI. Conclusions

Data Science is a very healthy area and we expect it to grow in interest at Indiana University although its setup as a program has some strange side effects. We cannot easily

translate success into new faculty as latter are appointed by regular departments. Fox teaches all his classes – residential or online -- with online lectures and this helps cope with research related travel. In spite of large interest in MOOCs, it is not clear how important online sections are for University (graded) education. If we look at technology, there is no clear winner although OpenEdX is gaining some traction and combines quality with name recognition. It is still harder to use than Course Builder (in terms of support) but that should improve, a recent development being the release of the OpenEdx stack by Bitnami [26], which we see as a step in the right direction. We suggest that there is no reason to distinguish the LMS used for online and traditional education; partly as key difference – video streaming – is handled outside the LMS. So the winner for MOOCs may win everything! We list some short comments about available technologies

Canvas: (Indiana University Default) Best for interface with IU grading and records; **Google Course Builder:** Best for management and integration of components; **Ad hoc web pages:** alternative easy to build integration; **Microsoft Mix:** Good faculty video preparation interface; **Adobe Presenter/Camtasia:** More powerful video preparation that support subtitles but not clearly needed; **Google Community Groups:** Good social interaction support; **YouTube:** Best user interface for videos (without Mix PowerPoint support); **Google Hangouts:** Good for instructor-students online interactions (one instructor to 9 students with live feed). Hangout on air mixes live and streaming (30 second delay from archived YouTube) and more participants; **OpenEdX:** possible future of Google Course Builder and getting easier to use but still requiring significant developer (support) effort

ACKNOWLEDGMENT

We thank the Indiana University Data Science faculty, especially Professor Judy Qiu for their support.

REFERENCES

[1] McKinsey Global Institute. *Big data: The next frontier for innovation, competition, and productivity* 2011 May [accessed 2011 November 3]; Available from: http://www.mckinsey.com/mgi/publications/big_data/index.asp.

[2] Georgia Tech. *Online Master of Science in Computer Science (OMS CS).* [accessed 2015 July 21]; Available from: http://www.cc.gatech.edu/academics/degree-programs/masters/online-ms-cs.

[3] NIST. *Big Data Working Group Reports from V1*. 2013 [accessed 2014 March 26]; Report at http://bigdatawg.nist.gov/V1_output_docs.php Available from: http://bigdatawg.nist.gov/home.php.

[4] J. Tinsley Oden, Omar Ghattas, and John Leslie King, *Final Report of NSF-OCI Task Force on Cyber Science and Engineering Chapter 7: Education, Training and Workforce Development in Computational Science and Engineering*. October, 2010. https://www.nsf.gov/cise/aci/taskforces/TaskForceReport_GrandChallenges.pdf.

[5] Fox, G., *Parallel computing and education.* Daedalus Journal of the American Academy of Arts and Sciences, 1992. 121(1): p. 111-118.

[6] Sharon Glotzer, David Keyes, J. Tohline, and J. Leszczynski., *Computational Science*, in *NSF MPS Advisory Committee White Paper*. May 14, 2010. http://www.nsf.gov/mps/advisory/mpsac_white_papers/mpsac_computational_science_white_paper.pdf.

[7] Swanson, C.D. *COMPUTATIONAL SCIENCE EDUCATION*. 2003 November [accessed 2009 December]; Available from: http://www.krellinst.org/doecsgf/cse_survey/all.php.

[8] Rice, J.R., *The CSE Community Converges: The First IEEE Computer Society Workshop on Computational Science and Engineering.* IEEE Computational Science & Engineering, 1997. 4(2): p. 10-11. DOI:10.1109/MCSE.1997.609826. http://portal.acm.org/citation.cfm?id=615490

[9] Fox, G., *From Computational Science to Internetics: Integration of Science with Computer Science*, Chapter in *Computational Science, Mathematics and Software*, R.F. Boisvert and E. Houstis, Editors. 2002, Purdue University Press: West Lafayette, Indiana. p. 217-236. http://www.old-npac.org/users/gcf/internetics2/.

[10] *Data Science Program in School of Infomatics and Computing at Indiana University*. 2014 [accessed 2014 March 25]; Available from: http://www.soic.indiana.edu/graduate/programs/data-science/index.shtml.

[11] *Masters in Data Science Degree Checklist.* [accessed 2015 July 23]; Available from: http://www.soic.indiana.edu/graduate/degrees/data-science/ms-data-science/degree-checklist.html.

[12] *Computational and Analytic Data Science Track of Indiana University Bloomington Data Science Degree*. [accessed 2015 July 23]; Available from: http://www.soic.indiana.edu/graduate/degrees/data-science/ms-data-science/computational-and-analytic-data-science-track.html.

[13] Google. *Online Open Education.* [accessed 2015 July 26]; Available from: https://www.google.com/edu/openonline/.

[14] *OPENedX Online education platform*. [accessed 2015 July 26]; Available from: https://open.edx.org/.

[15] Indiana University Data Science Program. *Big Data Applications and Analytics Fall 2015 course hosted on Open edX*. 2015 [accessed 2015 Sept. 15]; Available from: http://openedx.scholargrid.org/.

[16] Geoffrey Fox. *Big Data Applications and Analytics (X-Informatics) MOOC*. 2013 August 14 [accessed 2015 July 25]; Available from: https://bigdatacourse.appspot.com/preview.

[17] Geoffrey Fox. *Big Data Applications and Analytics Spring 2015 I400 (Undergraduates) and I590(Graduates).* 2015 Available from: https://bigdatacoursespring2015.appspot.com/preview.

[18] Geoffrey Fox. *Data Science Curriculum: Indiana University Online Class: Big Data Open Source Software and Projects*. 2015 [accessed 2015 March 31]; Available from: http://bigdataopensourceprojects.soic.indiana.edu/.

[19] Digital Science Center. *FutureSystems Homepage*. [accessed 2015 July 25]; Available from: https://portal.futuresystems.org/.

[20] *NIST Big Data Use Case & Requirements.* 2013 [accessed 2015 March 1]; Available from: http://bigdatawg.nist.gov/V1_output_docs.php.

[21] *HPC-ABDS Kaleidoscope of over 300 Apache Big Data Stack and HPC Tecnologies.* [accessed 2014 April 8]; Available from: http://hpc-abds.org/kaleidoscope/.

[22] Digital Science Center. *Tutorial on OpenStack for Virtual Clusters and Big Data.* 2015 [accessed 2015 July 26]; Available from: http://cloudmesh.github.io/introduction_to_cloud_computing/class/bdossp_sp15/week_plan.html.

[23] Jake Widman. *7 great MOOCs for techies -- all free, starting soon!* 2014 November 21 [accessed 2015 July 24]; Available from: http://www.computerworld.com/article/2849569/7-great-moocs-for-techies-all-free-starting-soon.html.

[24] Sidd Maini. *CGL MOOC Builder*. 2014 [accessed 2015 July 25]; Available from: http://moocbuilder.soic.indiana.edu/.

[25] Microsoft. *Office Mix.* [accessed 2015 July 25]; Available from: https://mix.office.com/.

[26] *Bitnami OpenEdX release*. [accessed 2015 July 29]; Available from: https://bitnami.com/stack/edx.