SUMMARY: This white paper highlights current and promising research in educational data analysis of online student activities, most prominently with Vanderbilt's MOOCs, but increasingly with other vehicles for digital learning and education. The white paper begins with a summary of educational data mining on campus and elsewhere, moving to the possibility of controlled hypothesis testing in learning science research using online material, and then possibilities stemming from deliberative design of online materials with educational data analysis in mind. A list of priority activities in educational data analysis is given in Section 7.

White papers by the Vanderbilt Institute for Digital Learning (VIDL) and its university partners highlight current and future activities related to VIDL's many-fold mission

- to facilitate means by which on-campus education and learning are enhanced through online and other digital technology, and vice versa;
- to promote research and other scholarship at Vanderbilt on digitally-enhanced education and learning;
- to encourage and oversee world and local community service by Vanderbilt that involves digital learning, to include production and management of Vanderbilt's massive, open, online courses (MOOCs);
- to serve as a university center for expertise and resources in digital learning, to include knowledge of such activities across our campus; and
- to search out and act on synergies and dependencies across these thrusts.

VIDL anticipates a steady stream of white papers within and between the mission areas of education, research, outreach, and inreach. White papers highlight Vanderbilt research within a larger landscape and are intended largely for Vanderbilt's internal use. In contrast, VIDL Technical Reports, again with Vanderbilt partners, are scholarly documents, ideally published in the literature.

There is important research that will be advanced by online and digital learning that is outside of educational data analysis (e.g., human-computer interfaces, natural language processing, cultural accommodation and hybridization, organizational change theory), much of it to include various forms of data analysis, but we restrict this report to (educational) data analysis intended to advance the learning sciences primarily. Other white papers will address other subjects as time and campus interest dictates.

**1. Educational Data Mining**

Vanderbilt's massive open online courses (MOOCs), currently offered exclusively through Coursera, create many possibilities for investigating learning and teaching. Much touted are opportunities for *educational data mining* (EDM), largely the search for patterns in student behavioral data from both online and in-class sources.
EDM, while still a young field, nonetheless predates MOOCs by five or more years. Investigators at Vanderbilt have been on the leading edge of EDM activities since the field’s inception. For example, Gautam Biswas and his colleagues have been using data mining in the context of the “Betty’s Brain” project for many years [1]. In this project, middle school students ostensibly teach an intelligent software agent (or a “smart” avatar) named Betty about concepts of climate change, ecosystem dynamics, and other problems of contemporary interest. The software agent, Betty, builds a conceptual understanding based on what it is taught by the middle-school student, and the middle-school student’s own conceptual understanding is then assessed based on what he/she is teaching the agent.

Importantly, the interactions between the K-12 “student teacher” and the Betty software agent are recorded as data, and this data is mined. Discovered patterns in student behaviors that are associated with high performance are contrasted with those patterns associated with low performance by the K-12 student, and differences between these patterns can then be used to reinforce or remediate, as appropriate, each K-12 student’s learning, as well as to redesign course lessons so as to improve expected learning outcomes. This approach of using EDM to discover patterns in student behavior that then informs instructional design is generalizable across many educational settings.

MOOCs and online learning generally promise to expand by orders of magnitude the data from which learning patterns can be mined. Moreover, MOOCs are ideal venues for questions of motivation and the like. VIDL is pairing with data mining researchers on campus to mine Vanderbilt’s COURSERA data. A high-level analysis by Vanderbilt’s various MOOC instructors and TAs has already begun (e.g., Schmidt and McCormick [2], Pope [3]). With his Stanford colleagues, Vanderbilt’s Brent Evans has been identifying patterns of student persistence at various levels of granularity (i.e., lectures, courses, students, communities) in MOOCs offered by other universities [4].

Moving forward, VIDL and its partners are looking for correlations between various kinds of MOOC activities, such as student participation on discussion forums and scores on homework and exams. We are looking to see whether these correlations are informed by the course area (e.g., computer science, English), by the type of assessment (e.g., auto-graded or peer graded), and by other demographics. VIDL is also looking for early markers of success (and the inverse!) in courses using prognostic methods developed in earlier Vanderbilt research [11, 12].

VIDL has identified other collaboration opportunities with researchers in Engineering, Peabody, A&S, and the School of Medicine. In addition to “standard” data mining questions using common data representations, Vanderbilt experts in natural language processing enable mining of discussion board posts, essay assignments, and other text-rich sources. Moreover, there is interest and expertise among Vanderbilt researchers on other data mining topics, to include privacy-preserving data mining, mining data from mobile platforms, GIS data, and images. In consultation with data scientists on campus, VIDL is creating and will manage a repository of student behavioral data from Vanderbilt’s online courses that can be accessed and analyzed for IRB-approved and exempt analyses by researchers and educators, notably those at Vanderbilt.

Much of the data analysis to date with MOOCs is opportunistic, and indeed, opportunism is implicit in the mining metaphor: given an environment, we mine for patterns. Data mining will grow into other media beyond MOOCs, to include other online courses, and other digital media such as Wikimedia and other open textbook initiatives. VIDL and its partners will continue this mining activity with vigor, but we also aspire to move beyond this paradigm in several ways.
2. From Opportunistic Mining to A/B Testing

VIDL is identifying lightweight experimental interventions that can be integrated into Vanderbilt’s MOOC offerings. For example, suppose that in upcoming Vanderbilt MOOCs, we pose the following question to one half of the “active” students, chosen at random, midway through the MOOC:

“If you finish this course with distinction, would you like to be considered for a community TA position in a subsequent offering of this course?”

The hypothesis that VIDL and its partners will want to test is whether the group of students that is asked this question will exhibit different rates of completion and distinction from the half that were not asked. In fact, we are interested in different rates and types of activities of all kinds in the A/B populations, and how the shapes of student activity curves (e.g., linear, exponential, power) are affected. There are many possible lightweight interventions (“lightweight” because they do little or nothing to affect the structure of the course). For example, in conjunction with an earlier cited study, Brent Evans and colleagues asked randomly chosen MOOC students to commit to lecture viewing schedules to see whether such a commitment would influence persistence throughout the course [4].

Dependent dimensions of course persistence and performance are of current interest in our A/B testing, but related dimensions of social connectedness are not far off. There are many independent dimensions that we wish to investigate such as the course topic, the timing of the intervention in the course, whether the incentive is punitive or rewarding, the strength of “local” learning communities, and many others.

In addition to lightweight interventions, holistic A/B tests are desirable. For example, Vanderbilt’s Heather Johnson and colleagues have compared professional development (PD) courses for teachers with content “equivalent” online and face-to-face PD versions [5], intending to identify and 1st-order differences in teacher outcomes between the two modalities, and any 2nd-order differences in the outcomes of the students of the teachers from these two populations! Such studies can inform the design of online programs, which a number of academic units on campus are interested in pursuing. In turn, online materials from these programs, our MOOCs, and other sources will be used in our on-campus courses,1 which will enable holistic and specific A/B testing.

3. From Cloud to Campus

Each of Vanderbilt’s previous Coursera instructors (i.e., Schmidt, Owens, Pope, Clayton, Harris, Duda, Wehbe) is currently using, or plans to use, their online material for on-campus education (e.g., enabling instructors to flip their classes). Moreover, Bruff, Fisher, McEwen and Smith [6] report on Fisher’s use of MOOCs (by others) to enhance on-campus courses. More generally, a Gates-foundation funded study is examining how MOOCs like Jamie Pope’s Nutrition course can be used in on-campus courses at the University of Maryland [7]. In many cases the MOOC is repurposed as a “closed instance” or small, private, online course (SPOC), but unique data analysis possibilities occur when a campus course is tracking an ongoing and global MOOC.

---

1 See VIDL’s white paper entitled “Synergies between learning and teaching in the cloud and on the campus” (forthcoming).
Education and social science researchers will be very interested in looking at how local communities of students (e.g., at Vanderbilt) interact among themselves, as well as with the global MOOC community [e.g., 7, 8]. Even as we collect data on MOOC students, we will want complementary data on local interactions among Vanderbilt students who are taking on-campus courses that integrate MOOCs. Moreover some of the questions that we will want to ask are not about the MOOC’s use per se, but about (for example) the active learning that takes place in a “flipped” class because of the use of MOOCs and like technologies. As we move to flipped models that have rich social interactions in the classroom, the implications for persistent social relationships among Vanderbilt students and alums will be of great interest, to say nothing of interactions between students at Vanderbilt with students across the globe.

4. From Analysis within MOOCs to the World Beyond

Data analysis within a MOOC is the norm. Within-MOOC analyses look at student behaviors, ranging from keystroke patterns (“nano” analysis) to quiz performance, discussion board posts, and other lesson-specific “micro” analysis. Vanderbilt’s current analyses focus on within-course comparisons across students.

4.1 MOOCs as anchors in social networks

There is a world beyond that of single MOOCs, to include traditional social networks like Twitter and Facebook (intersecting with the local communities discussed above). Interactions between students on MOOC discussion boards and in peer assessments can lead to relationships (e.g., “following” and “friend”) on (other) social networks. These studies will be particularly important in understanding how student support groups emerge, and whether these MOOC-centered, ad hoc communities morph, but nonetheless persist, beyond the particular contexts in which they first emerge. In general, there are many mining and other data analytic studies that can be done in a MOOC’s extended “gravitational field.”

Resource repositories such as YouTube are now fully social networks as well, and the ways and amount that MOOC (and on-campus) students use these resources are of great interest to VIDL and it’s partners. Consider Figure 1’s example of activity on one YouTube channel in which MOOC students have been known to come for remediation. Are the peaks in activity synched with a MOOC offering? Are there traces in MOOC discussion boards of this and other auxiliary resources? Are there changes in the MOOC in response to students looking to other material, or are community resources embraced and endorsed by the MOOC provider? And as with other social networks, what are the short-lived and persistent relationships that emerge between people?

4.2 Distributed course customization and crowdsourcing

The interaction between MOOC students and (other) online social networks and resource repositories creates possibilities for course customization, notably by the addition of remedial and advanced material, which will often reside on a different platform than the MOOC. Moreover, as remedial and advanced materials (e.g., on Youtube) are brought into a MOOC’s gravitational field, some will be more widely endorsed, and consensus favorites will emerge as a result of crowdsourcing.

---

2 The merging of MOOC platforms like Coursera and local learning management systems (LMSs) like Oak, so that local and global student data can be analyzed seamlessly, is the subject of another white paper, “The merging of MOOCs, e-books, and LMSs” (forthcoming).

3 Ad hoc communities arise to achieve goals; the idea of ad hoc communities arises from Larry Baralou’s concept of ad hoc categories [9].
Figure 1: YouTube channel activity over time – are patterns due to MOOCs?

We can track some cross-platform use through surveys of our MOOC students. In general, however, tracking student activities across different platforms such as Coursera and YouTube is difficult, if not impossible at finer granularity. A recent partnering of edX, Stanford, and Google highlights the possibility that some groups will be able to do such cross-platform analysis soon, probably using sophisticated link identification algorithms that will identify the same individual by different names and on different platforms. Partnering with other platforms may therefore be of interest, not simply for purposes of delivering content, but for the research opportunities that these partnerships will enable.

Even within a platform, such as Coursera, there are student interactions that extend beyond any single MOOC. Notably, we can track students across MOOCs taken on that platform. We can certainly do such tracking with MOOCs that are offered by Vanderbilt. We can also request data from Coursera partners and look at student pathways in this larger collection of courses. Undoubtedly, we will be asked to reciprocate and VIDL wants policies in place to share MOOC data with other institutions.

5. From Courses to Curricula

The most obvious pathways among MOOCs are in related areas. As early as October 2012, for example, it was possible to take courses that would ostensibly satisfy the requirements for an undergraduate computer science major. These MOOCs spanned levels of expertise, from beginner to graduate level, and there were (and are) different options that students could take among comparable courses.

The same customization and crowdsourcing possibilities exist for curricula as they do for "single" courses. Figure 2 illustrates the idea that MOOCs in a related area (e.g., computer science) support customized paths, each path representing a possible curriculum that might be taken by different students. Again, some of these pathways may be variably endorsed, leading to crowdsourced consensus favorites for differing student populations. Studying customization and crowdsourcing of curricula may be relatively easy since student pathways through this curricular space may be dominated
by same-platform (e.g., Coursera) student “enrollment” and “completion” data. Figure 2, however, highlights the possibility of cross-platform paths (e.g., Coursera, edX).

Figure 2: One path (i.e., curriculum) through a small part of the space of courses.

Another important area for research is data analysis of "support groups" (working groups, “affinity groups”) that develop among students during a course, some of which persist and morph across courses and a curriculum. How do peer groups change across MOOCs when such community formation is not facilitated and not hindered?

6. From Analysis to Design (and more analysis)

Vanderbilt will design online content and infrastructure to enhance on-campus education, a theme that we elaborate in VIDL’s white paper entitled “Synergies between learning and teaching in the cloud and on the campus” (forthcoming). However, deliberative design of content and infrastructure can also provide a stable platform for deeper data exploration, which is the focus here.

6.1 Design of Open Curricula

Figure 2’s example of the core computer science curriculum is generalizable to other disciplines as well. Online programs stick closely to the structure of their on-campus antecedents, and we imagine that many MOOCs will develop that fill in the various slots of traditional disciplinary curricula. Within each of these largely standardized areas of study, opportunistic sequencing of MOOCs by students interested in curriculum-level learning will naturally follow traditional course structures. That is not to say that all students will want to pursue an area of study in depth, but a significant proportion will do so. We can imagine that formal recognition for in-depth study of a field will emerge, ranging from badges to diplomas, with courses that span institutions. But as we get to
advanced courses, the well-trodden paths are less obvious, and sequencing MOOCs by design becomes more important.

Vanderbilt has taken a lead role in moving beyond opportunistic path-finding of MOOCs. In particular, Doug Schmidt and Jules White of Vanderbilt and Adam Porter of the University of Maryland are organizing the first trans-institutional MOOC sequence, with Porter’s mobile apps-development MOOC serving as a soft prerequisite to Schmidt and White’s Pattern-Oriented Software Architecture MOOC.

**Figure 3:** MOOC sequences are macro-level curricular constructs that are subject to data mining and A/B testing.

VIDL and its partners will mine student behavioral data in and across MOOCs of the sequence. Is performance correlated across the two MOOCs? Are there some students that seem to suffer from MOOC fatigue in the latter Vanderbilt course, but may in fact return in a subsequent offering? Are there some students who appear to do better in the second Vanderbilt course because they took the first course from Maryland, as opposed to other students who are demographically similar, but who did not take the first course from Maryland? We are also interested in planning A/B tests in advance, investigating factors that improve retention, comprehension, and social connectedness.

The Vanderbilt/Maryland sequence is a nascent step towards richer curricular structures that are open and trans-institutional. We can imagine other MOOC sequences developing, first as total orderings (i.e., each MOOC has a unique predecessor), and later, collections of MOOCs with choices among courses at various points. We might even facilitate such analysis across courses by taking longer MOOCs of 10-16 weeks and breaking them into smaller MOOCs of 4-5 weeks each. Collections of related MOOCs, whether in computer science, medicine, education, or history would provide rich sources of student networking data. We know that some institutions that have already developed such collections will not share student data from those courses.

---

4 Many of the data analyses with MOOCs are equally applicable to traditional campus scenarios, and are therefore of possible interest to VIRG. An analog to our question about how student performance varies with and without the Maryland (soft) prerequisite, for example, is whether students who skip a course prerequisite, perhaps allowed by instructor discretion, and perform differently in that course than students with the prerequisite.
One reason to build collections of MOOCs within particular subject areas, in addition to our current strategy of sampling very different subject areas, is so that Vanderbilt researchers have an accessible pool of data on the migration of individual students and groups of students across MOOCs in the same area! Without such collections, Vanderbilt may be precluded from engaging in across-course data analysis that have important implications for curricular design and tracking student cohorts across time.

6.2 Design of Extended Courses
As with any course, online courses, to include MOOCs, are carefully designed. In most cases established courses have benefitted from generations of “crowdsourcing” to include synthesis that led to textbook creation. As we have noted earlier, however, students and instructors can pull remedial and advanced material into a MOOC’s extended field in ways that we would not typically characterize as deliberative design. Nonetheless, such extended material can be created very deliberatively.

One example of this deliberative design of auxiliary materials is found in the Blended and Online Learning Design (BOLD) initiative of Vanderbilt’s Center for Teaching [10], which coordinates graduate student production of online course materials to supplement on-campus courses. This same idea can be applied to the creation of auxiliary material for online courses, where the supplemental material can be text based (e.g., Wikipedia), video (e.g., on Youtube), or other multimedia.

Content creation by students and faculty alike can strategically target MOOCs. Appropriate places in a MOOC for supplemental content can be identified based on knowledge of the area and insights into what many students will have difficulty with or be bored with, as well as explicit indicators from MOOC students on discussion boards and other social media (e.g., Twitter, Facebook). Strategic creation of remedial and advanced supplements for Vanderbilt’s MOOCs or anyone else’s MOOCs, gives students and faculty a worldwide and “authentic” audience for their content (again, see the Synergies white paper), but also can be done so as to facilitate mining and other analysis by Vanderbilt researchers of networked student behavior.

6.3 Design of student support groups
Students on campus pursue study within a cohort, and online students can and will as well. As we have noted, we can track relationships that arise and evolve between individuals and groups, perhaps starting with discussion boards and peer assessments in a MOOC, but then crossing into other spheres, be it other courses or social networks.

VIDL is interested in characterizing emergent behavior, but also in cyber technology, informed by behavioral sciences, that will facilitate the formation of student peer support/working groups. What are the benefits to students working with physically non-local partners? What are the variable benefits of different methods of forming groups, for different participants in the groups, etc? Assuming that groups are formed by MOOC students with different skill sets, and include stronger students and weaker students, do the groups benefit the stronger? Do the groups benefit the weaker? We can peer into the behavior (even the key strokes) of all group members and judge whether the weaker students are really benefiting or being carried by the stronger students, and under what conditions do these differing outcomes arise?

We are also interested in recommender systems, informed by all the analyses of which we have been speaking, that will help individuals and cohorts to navigate the online education space, which will direct students to courses in which they are apparently ready, and redirect students who are not ready. In contrast, students who are advanced can be moved forward and/or recommended as community TAs. Facilitating and
managing peer support groups can enhance educational practice, and it will be a rich environment for student behavioral data analysis, as well as research in areas such as human-computer interfaces and cultural blending, to name a few.

7. Near-term VIDL Priorities on Educational Data Analysis

This white paper has laid out directions of near-term EDA research, much of which Vanderbilt can and will lead. To facilitate Vanderbilt activity in EDA research, VIDL is taking several specific steps.

(1) With partners across schools of Vanderbilt, VIDL is performing data mining of behavioral data from Vanderbilt’s previous Coursera courses.

(2) VIDL is preparing A/B tests for upcoming Coursera courses in conjunction with the instructors and learning science researchers.

(3) We are designing core pre, intermediate, and post MOOC survey data of students that will help condition data analysis research (as well as in support of administrative and educational services), informed by previous efforts coordinated by the Associate Provost for Undergraduate Education and Digital Learning, Cynthia Cyrus.

(4) With partners, VIDL is designing and implementing databases to store Vanderbilt’s MOOC student behavioral data in standardized forms that preserve security, privacy, correctness, consistency and provenance of the data. In addition, VIDL is implementing standard procedures for acquiring this data from Coursera to insure data integrity.

(5) Building on activities (1)-(4), VIDL is preparing general procedures for access to MOOC student behavioral data by Vanderbilt researchers going forward. VIDL is also preparing plans for institutional review board (IRB) submission, to include exempt and expedited proposals, for the benefit of Vanderbilt researchers.

(6) VIDL is evaluating proposals on digital learning for internal seed funding based on the potential of the proposed work to advance research. The potential of a digital learning project is but one factor used in evaluation.

(7) VIDL is participating and leading research proposals on digital learning to funding agencies and foundations. In addition, VIDL is writing letters of support for external proposals by faculty in areas other than digital learning. In most cases, support is for the broader impacts aspects of the proposal, but a factor in deciding to lend support is whether the proposed project can be leveraged to advance learning sciences research in the digital arena, including in EDA.

Beyond these ongoing activities, VIDL recommends a strategy of pursuing disciplinary depth, as well as breadth, in Vanderbilt’s Coursera offerings, thus giving Vanderbilt faculty and graduate students a rich and reliable source of data to pursue unique learning sciences research, to say nothing of the educational and outreach advantages that we will write of elsewhere. The curriculum-level innovations of Sections 5 and 6, which Vanderbilt is leading on, can give Vanderbilt a leader’s advantage in curricula-level data analysis research.

8. References
[1] The Teachable Agents Group at Vanderbilt University
http://www.teachableagents.org/research/bettysbrain.php

http://www.dre.vanderbilt.edu/~schmidt/PDF/POSA-MOOC.pdf


[4] Brent Evans, Persistence Patterns in Massive open Online Courses, Presentation at the Leadership, Policy, and Organizations Symposium at Peabody School of Education, Vanderbilt University, Nov 21, 2013
http://www.youtube.com/watch?v=MeFmpmLKO0M&feature=youtu.be

http://jte.sagepub.com/content/early/2013/05/24/0022487113494413.full.pdf+html

http://jolt.merlot.org/vol9no2/bruff_0613.htm


Acknowledgements

Douglas H. Fisher, VIDL’s Director, authored this white paper, which benefitted from discussions with colleagues at Vanderbilt, most notably

- Gayathri Narasimham, VIDL’s Assistant Director for Education and Research
- Gautam Biswas, Professor of Computer Science and Senior Research Scientist at the Institute for Software Integrated Systems
- Derek Bruff, Director of the Center for Teaching
- John Kinnebrew, Research Associate at the Institute for Software Integrated Systems
- Douglas C. Schmidt, Professor of Computer Science and Associate Chair of Electrical Engineering and Computer Science

Selected text from this (unpublished) white paper has been used and adapted in “The Past, Present, and Future of MOOCs” by Balakrishnan Dasarathy, Kevin Sullivan, Douglas C. Schmidt, Douglas H. Fisher, and Adam Porter (submitted for publication).