Educational Data Sciences – Framing Emergent Practices for Analytics of Learning, Organizations, and Systems

Philip J. Piety
Learning Scientist for Education Systems
Ed Info Connections
3 Carole Court
Silver Spring, MD 20904
+001 301-332-2803
ppiety@edinfoconnections.com

Daniel T. Hickey
Associate Professor and
Learning Sciences Program Head
Indiana University School of Education
504 Eigenmann Hall, Bloomington, IN
+001 812-856-2344
dthickey@indiana.edu

MJ Bishop
Director, Center for Innovation and
Excellence in Learning and Teaching
University System of Maryland
3300 Metzerott Road, Adelphi, MD 20783
+001-301-445-1997
mjbishop@usmd.edu

ABSTRACT

In this paper, we develop a conceptual framework for organizing emerging analytic activities involving educational data that can fall under broad and often loosely defined categories, including Academic/Institutional Analytics, Learning Analytics/ Educational Data Mining, Learner Analytics/Personalization, and Systemic Instructional Improvement. While our approach is substantially informed by both higher education and K-12 settings, this framework is developed to apply across all educational contexts where digital data are used to inform learners and the management of learning. Although we can identify movements that are relatively independent of each other today, we believe they will in all cases expand from their current margins to encompass larger domains and increasingly overlap. The growth in these analytic activities leads to the need to find ways to synthesize understandings, find common language, and develop frames of reference to help these movements develop into a field.

General Terms

Theory, Measurement, Performance, Management, Design.

Keywords

Analytic approaches, methods, and tools for sense-making in learning analytics. Theories and theoretical concepts for understanding learning. Learning Analytics, Educational Data Mining, Educational Data Science, Learner Analytics, Big Data, Data-driven Decisions

.

LAK '14, March 24 - 28 2014, Indianapolis, IN, USA
Copyright 2014 ACM 978-1-4503-2664-3/14/03…$15.00.
http://dx.doi.org/10.1145/2567574.2567582

1. INTRODUCTION

In the past decade, several distinct movements have emerged around the use of data education. We refer to these areas as:

- Academic/Institutional Analytics,
- Learning Analytics/ Educational Data Mining,
- Learner Analytics/Personalization, and
- Systemic Instructional Improvement.

All of these movements are significantly related to digital technology and its ability to collect, share, and represent vast amounts of information with relative ease. However, this growth has also come with some fragmentation and terminological confusion. There are now distinct communities of discourse using similar concepts—often adapted from outside of education—and not always using them comparably. As these movements grow, conceptual overlap will also increase as concepts used to characterize one kind of contribution can conflate with other dissimilar work while at the same time areas of similarity might not be recognized. This will limit knowledge sharing and synergistic advance. A common and comprehensive language that can be used across these areas is needed. We believe that a broader notion of "Educational Data Sciences" will benefit both those producing and consuming information from these practices as well as those developing education programs aimed at building the human capital necessary to work with educational data.

In covering this topic, we are mindful that the elements that form a scientific and/or professional community are often *unclear*. We also recognize that conventional roles and distinctions (such as researcher, practitioner, developer) have become increasingly fluid and contingent. University-based researchers traditionally dominated the practices of producing high-quality evidence and using academic conferences and journals with peer review as a method for ensuring rigor and quality. These structures are increasingly challenged in a networked world where information can be collected and shared across communities at a low cost and where commentary and review can take many forms, including blog posts, community/topical web sites, and publication of primary sources. Rather than the kind of scientific community once described by Kuhn[1], this construct is now a much more fluid and dynamic endeavor. We contend that the notion of Educational Data Sciences embraces this reality.[2]

We begin this discussion with a high-level survey of these four emerging communities working with educational data. These communities each began from a slightly different vantage point. They all emerged relatively independently from each other roughly after the turn of the century. They are largely communities that involve different professionals and publication forums. At the same time, these communities are encountering similar issues involving rapid change and questions about their disciplinary boundaries. Their work challenges traditional evidentiary practices at the same time it creates new opportunities around data visualization and interpretation. In doing so, they raise common questions about things like culture and ethics/privacy. This suggests that they may really be segments of one larger and emerging field. While the kinds of data they have worked with are different, they all share some common features that we discuss below.

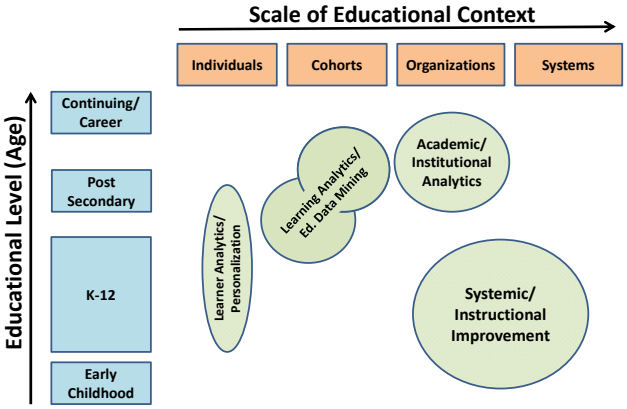

Figure 1- Early Educational Data Sciences Sub-fields

2. FOUR EMERGING COMMUNITIES

We begin with four broad categories of work involving educational data illustrated in Figure 1. With the realization that taxonomies are problematic, we base this classification on the actual communities and activities that have emerged to date. We believe that many will agree that (a) these four areas represent coherent communities that have identifiable research literatures and professional associations that will represent their work, and (b) these boundaries are neither fixed nor permanent.

The four communities illustrated in Figure 1 seemed to develop independently around different social levels and kinds of information, but all encountering similar issues. Rather than separate activities, we see them as the initial cells of the Educational Data Sciences. We believe they are now becoming sub-communities within this newly forming research/practice paradigm. This new paradigm has the potential to reshape both the ways educational organizations use information for their management and the way researchers collect evidence of effectiveness. In many cases, the same information collected from digital tools can be used for multiple purposes, and information collected for one purpose will likely be used for very different purposes in the end.

2.1 Academic/Institutional Analytics

The community we refer to as Academic/Institutional Analytics emerged from within what has traditionally been called Institutional Research. This began as a natural extension to the kind of analytic work and reporting that many institutions of higher education have been doing for years. In recent years, Institutional Research has begun to move forward with new kinds of data analyses enabled by many of the same digital technologies that have supported other educational uses of data outside of the IR context. While similar in some ways to these areas we will discuss below, the field of Academic/Institutional Analytics is *organizational*. It involves questions about who gets access to an institution of higher education, how they are admitted, how well they progress through the system and other "process" questions involving student services, finance, fundraising, administration, grants management, and the like.[3]

We use the term Academic/Institutional Analytics to make clear the focus of this area has been on the institution rather than on the processes of learning or the details about domains where learning occurs. While there is certainly consideration of learners, this consideration is generally from the perspective of the institution. For example, some use data in this community for early warning systems that can identify students at risk of future failure to complete.[4][5] Academic Analytics can focus on students, courses, programs, and even faculty characteristics. In some cases research productivity and publications that are useful in tenure decisions and in comparing programs against competitors are also part of Academic Analytics. Work in this area has also compared different institutions in terms of research funding, acceptance rates, faculty productivity, and the like.[6]

One notable development in this community is the Open Academic Analytics Initiative (OAAI) to develop a common set of tools that can be used by different institutions to develop their own analytics projects according to common definitions.[7] This initiative brings the potential to reuse various analytic tools and compare institutions of higher education according to common indicators. Relative to the arguments of this paper, this should make it easier for Educational Data Scientists in other communities to use that information. However, the use of common metrics may make it harder for some institutions to share their own strengths when those strengths are not represented in the common frameworks.

2.2 "Big Data" in Education: Learning Analytics & Educational Data Mining (EDM)

This "combined" community is perhaps the clearest example of the kind of convergence that we expect to occur more broadly within the Educational Data Sciences. The Learning Analytics community and the Educational Data Mining (EDM) community emerged around the same time and have similar roots in digital learning environments. The distinctions between them were blurry from the start and in recent years have converged. They now share so many aspects we are electing to treat them together within the broader EDS landscape.

The term "data mining" has been used in many fields as a general activity involving the search large datasets to discover patterns. The EDM movement cohered around 2005 and quickly developed a conference (the International Conference on Educational Data

Mining) and a *Journal of Educational Data Mining* (JEDM).[8] [9] This community is particularly rooted in the data from cognitive tutors and other immersive environments such as videogames[10] that produce detailed activity traces of student work. In 2013, Baker argued that EDM included prediction methods, structure discovery methods, relationship mining, model distillation, and distillation of data for human judgment.[11]

Learning Analytics emerged around the same time as EDM, but did so more around the analysis of data from learning management systems (LMSs) in higher education. LMSs generate a wealth of data that can be used to track student activity at various levels of detail, including assignment completion, participation in discussions, and assessments.[12] Learning Analytics also gained momentum with the explosion of massively open online courses (MOOCs) and their ability to generate data for thousands of students taking the same course. With these large course environments have come the frequent practice of using student-to-student interactions for assessment and evaluation. This allows researchers to study and understand in new ways the social dimensions of the learning process as is sometimes called Social Learning Analytics.[13] As with EDM, the Learning Analytics community began holding international meetings, including the Learning Analytics and Knowledge (LAK) and the Society for Learning Analytics Research (SoLAR) conferences.

Because of their similarities, these two communities have largely begun to fuse with many researchers publishing and presenting in both forums.[14] There may be more human interaction and interpretation with data in Learning Analytics, and more reliance on analysis and adaptations embedded in the technology in EDM. However, as Baker says: "Despite these differences, on the whole the two communities share a focus on discovering what can be learned from large-scale fine-grained educational data, and how it can be used to promote learning."[15]

As the world of Learning Analytics/EDM rapidly expands, the kinds of data that are being used also grows from test and assessment data to texts produced by online tools, interactions of students working in teams, peer assessments, surveys, and the like. This movement has begun to receive combined attention from both federal policymakers and foundation funders and is often seen as the community dealing with "Big Data" in education.[16] "Big data" is popular term for business management and its use in educational contexts is currently being explored. Of course, massive datasets do exist in the other EDS communities; however, it is within Learning Analytics/EDM that the datasets tend to be more fine-grained and rapidly updating in a way that makes them comparable to business applications of the "Big Data" concept. That said, the kind of analytics that are required for deep insights in many educational settings may require drawing broadly from data that more than one community currently works with.

In summary, by revealing what teaching methods and academic interventions are most likely to enhance learning of particular content with particular learners, Learning Analytics & EDM generally consider activity at the micro or "learning" level.

2.3 Learner Analytics & Personalization

There is also an emerging community collecting analytics at what might be considered the macro or "learner" level in order to explore how the differences among learners affect their persistence and overall college success.

When we think about differences among learners, we typically think of differences in their cognitive abilities. Cognitive abilities represent the learner's state of knowledge or "brain power." As recently as the 1960s, the prevailing belief was that the learner's cognitive abilities were the foremost determiner of what could be acquired from an educational experience, leading to a rather "selective" mode of education requiring learners to adapt to, and survive in, the learning environment as it has been designed. By the late 60s to early 70s, Glaser and others*[17] began calling for a more "adaptive" educational mode that considered other non-cognitive factors that might emerge as important in more interactive settings where there might be room for adjustment between abilities and modes of learning.

Learn*er* Analytics is therefore concerned both with collecting information around differences among learners with regard to cognitive traits like aptitudes, cognitive styles, prior learning, and the like, as well as the learners' non-cognitive characteristics such as differences in levels of academic motivation, attitudes toward content, attention and engagement styles, expectancy and incentive styles, personal experiences, extra-curricular interests, socio-economic status, and even family situations.

With these data, therefore, Learner Analytics attempts to predict things like which learners may have difficulty making the transition to college and identify the interventions best able support those at risk. Recent developments in the area of Learner Analytics have explored matching student characteristics to majors and career paths, increasing the likelihood they will remain engaged and persist through degree completion (see, for example, *Degree Compass* and *My Future*[18]). There has been recent interest in student persistence through adversity, what some have called "tenacity or grit" that allow some students to succeed even when other students with similar characteristics as shown through data like income and social background often fail.[19]

2.4 Systemic/Instructional Improvement

The field of research we call here Systemic/Instructional Improvement (SII) developed in the United States with the direct support of federal legislation in the Elementary and Secondary Education Act (ESEA), known for a time as No Child Left Behind (NCLB).[20] NCLB brought with it both annual testing requirements for many grades in math and reading as well as policies to direct educators to use data in their daily decision making processes. Within a few years, "Data-driven decision making" had become a popular topic across the nation.[21] Internationally, there was analogous work, but it is in the United States that the movement was given much momentum and encoded into policy.

Within this movement one can also see connections to many other parts of the Educational Data Sciences as well as some unique features. Like the other three communities, SII has been propelled by the collection of data. It has also, like other areas, seen a diversification of what counts as valid information for use. In its early days, test scores were considered almost synonymous with data. Data-driven decision making could also have been referred to at one time as "test-based decision making." As experience in the field accumulated and the design weaknesses of the NCLB legislation that encoded test-based accountability became clearer, additional forms of evidence beyond test scores became more widely used.

In the United States, with NCLB there also began a substantive research literature that looked at how district, school, and educators used data in different settings. Much of the research focused on determining the extent to which using data in school settings could be shown to improve achievement. While the research has yet to show clearly how data use is leading to broad achievement gains, the literature has documented the proliferation of digital data that are being used in the K-12 world. Much of this literature has also been marked by an inconsistency in terms of methodology. As Coburn and Turner noted when sketching out this field:

> "In many ways, the practice of data use is out ahead of research. Policy and interventions to promote data use far outstrip research studying the process, context, and consequences of these efforts. But the fact that there is so much energy promoting data use and so many districts and schools that are embarking on data use initiatives means that conditions are ripe for systematic, empirical study."[22]

The research community is still evolving and new studies are underway that use richer conceptual frameworks. Most of the research in this movement, early and current, has been focused on districts, schools, and classrooms. However, the movement has been given significant momentum from the development of a set of state level data infrastructures usually called state longitudinal data systems (SLDS) after the federal funding program that supported their development in many states. While each SLDS can differ from the others in important ways, they all link up student performance information across years with information about student demographics, teacher information, licensure, higher education, and in many cases workforce data. The SLDS systems then are becoming rich analytic resources that can be used by policymakers and researchers. In some studies these systems are now developing components to provide data about early childhood education and as the data quality in them increases, the kinds of analyses possible will also increase.

One important element of this movement that is connected to both test-based accountability and to SLDS systems is the use of data for teaching evaluations. This focus on teachers and potential for the evaluation of teaching is the most salient characteristic of this EDS community. Prompted by federal funding programs, many states and districts developed teaching evaluation systems that used test scores in what are called "value-added models" or VAMs. While proving controversial, VAMs are designed to estimate the "amount of learning" in each student's achievement can be attributed to the instruction provided by each teacher [23]. By factoring in student characteristics, a good VAM should be able to distinguish between the difference a good teacher can make in the progress or growth of a child. While many believe that proponents of VAM have yet to produce sufficiently reliable indicators, these efforts certainly highlighted the importance of many other kinds of information such as course classifications and learning goals and even student attendance/assignment information. Also, since only about a third of teachers would have test scores to be used for developing these models, and because the test scores had limitations in precisions, many districts started to collect many other kinds of information related to teaching, including having multiple observers rate teachers' work to produce evaluations that could be used later in developing a comprehensive teacher evaluation score. These observation data are often put into central databases where they can be mined for patterns and insights that are beyond their use in evaluating teaching.

Thus, in addition to achievement scores, SII is also encompassing student demographics, attendance, grades, course enrollment histories, graduation status, behavior data, special education information, teacher qualifications and professional development histories.[24] In some districts, the ways that educational data are used are growing such that they may include the use of community and social service data to understand larger patterns of family mobility and economic pressures. Geographic information tools, including complex, multileveled geospatial data analyses are beginning to allow the people responsible for the management of school systems to see different ways to address the problems of their constituencies and to see how they can manage their assets for better overall systemic performance with information being a vital and essential component.

While this EDS community has attracted significant funding and attention from researchers and policy makers, there are no dedicated journals or conferences about these topics. There have been several books from leading academic publishers and special issues in premier journals, including *Teachers College Record* and the *American Journal of Education*.[25] However, the scholars working in these areas have been distributed across several different kinds of departments and the leading educational research conferences have featured papers and symposia on these topics within existing research strands, including leadership, policy, and organizational studies.

3. COMMON FEATURES OF THE EDUCATIONAL DATA SCIENCES

Considering these four communities that appear to make up the Educational Data Sciences, we see a number of important features emerge across them. These following five features inform our description of this nascent field.

- Rapid evolution indicative of a broad sociotechnical movement
- Boundary issues indicating commonality
- Disruption in evidentiary practices
- Questions about visualization, interpretation, and culture
- Ethics, privacy, and information architecture

Below, we discuss these characteristics and how they are informing our conception of this field involved with sociotechnical topics where the technologies and social activity structures are interacting while both undergo pressures to change.

3.1 Rapid evolution indicative of a broad sociotechnical movement

All areas of the Educational Data Sciences have been rapidly evolving; in a few short years going from a small group of individuals to a wider circle of interest and involvement. In some cases, federal and foundation support have provided more momentum, but the movements themselves have all experienced forward motion and increasing interest. The fact that they all have followed this similar path at about the same time (starting in 2004-

2007) indicates they are likely part of a broad sociotechnical movement.

Sociotechnical movements—the printing press, the age of steam, the Internet—usually occur across many areas at about the same time and for a variety of reasons. One is that the enabling conditions and key technologies emerge across sectors giving rise to multiple sets of innovations that may at times seem disconnected, but are often related and interdependent.[26] Also, in many cases the societies' expectations are such that the innovations come at a time when there is other general interest in the kinds of changes that the innovations make possible. We see this certainly in the area of educational data where there has been both increasing capability to produce data and a greater public appetite for the kind of data-based accountability and the use of information that we see generally across all of these areas of education.

3.2 Boundary issues indicating commonality

All four EDS communities have been experiencing boundary issues." This is to be expected since they are growing and creating space for themselves within ecosystems of established academic communities. That they are all going through these kinds of boundary issues and their boundaries are coming into contact with each other is further evidence of their interrelationships in a broad sociotechnical movement. In many cases we can also see the boundaries of one intersect with the boundaries of other. Rather than being on distinct independent trajectories, their boundary issues involving each other speak to their becoming parts of a whole larger community more than separate fields. Just as Learning Analytics and Educational Data Mining are increasingly seen as a single community, similar blurring and synergy can be expected across the EDS landscape.

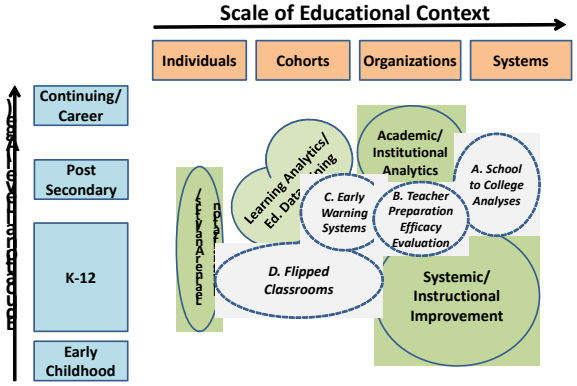

Figure 2- Boundary activities involving more than one EDS area.

Figure 2 illustrates four distinct areas of current activity that appear to cross boundaries of one or more of the original four cells of the Educational Data Sciences.

a. **School to College Analyses.** Student progress from high-school to college can now be reliably analyzed using datasets maintained by the National Student Clearinghouse, a not-for-profit organization that maintains records of college enrollment for verification purposes. While analyses to date are not sufficiently detailed to show which classes various students take and how they do, that kind of information is likely to emerge in coming years. Particularly promising is the eventual ability to look at transitions across currently different geographic regions and different career paths.[27]

b. **Evaluating Teacher Preparation Efficacy.** An emerging area of reform involves measuring the impact that teachers coming from different teacher education programs have on the students they teach. Such activity draws on information associated with both the Systemic/Instructional Improvement and the Academic/Institutional Analytics communities.[28]

c. **Early Warning Systems.** Another kind of boundary activity in higher education systems can come with early warning systems where institutions analyze student progress/risk factors and then can connect that information to the kind of student/class focused data coming out of Learning Analytics/EDM.[29]

d. **Flipped Classrooms.** The movement towards what Khan has called "flipped classrooms," which has roots in personalization and Learner Analytics, is beginning to make connections to both Learning Analytics/EDM and Systemic and Instructional Improvement with the development of learning maps and other tools focused on the learner.[30]

These examples are, we believe, just the beginning of a broadening of these areas from their current foci and the beginning of a new phase of Educational Data Sciences that can involve greater cross-community discussion. Whether or not all of these boundary activities grow and mature is less important, we believe, than their existence across these communities.

3.3 Disruption in evidentiary practices

Across all four of these communities, there have also been questions about how to use different kinds of information that were previously not available; how to make high quality inferences using the different kinds of evidence in ways appropriate to the context.[31]

The world of education has long been dominated by specific types of data and corresponding analytic methods. In a world ruled by assessment items, pre-post testing strategies and statistical processes built around assumptions of certain kinds of data, these new forms of evidence—log files, conversational records, peer assessments, online search and navigation behavior, and the like—are raising big questions about how to use them. They are, in effect, disrupting traditional ways of working, acting in a way similar to *disruptive innovations* that alters cultural, historical practices and activity systems.[32] Disruptive innovations are ones that reshape markets by delivering value in different ways. In the practices of educational evidence, these new kinds of data can act the same way providing new ways to understand student learning and educational processes.

Evidentiary economics is a part of the disruption. While these new data forms come with varying degrees of reliability, many of them cost very little to gather. This is what makes them so potentially disruptive. Collecting high-quality evidence of student learning has traditionally required secure tests that are expensive to develop and administer. Conversely, the kinds of evidence that can be harvested from digital tools may be less precisely engineered and subject to many kinds of potential bias and error for which one must account. In contrast, the cost of collecting these new kinds of information are usually negligible and so the value equations around collecting evidence can change with low cost data sources that in many cases can support different kinds of inferences.[33]

3.4 Visualization, interpretation, and culture

Across these four areas we also see the emergence of issues around visualization and interpretation of information. Visualization is usually the lead element in these discussions as different representational schemes. These include "dashboards" that rank and sort individuals and other similar tools used to make sense of the vast amounts of diverse information available in these four communities.

Many in the EDS movement are turning to the literature on visual representation of quantitative data, including the work of Tufte and others who explore the graphical dimensions of data displays.[34] While the graphics are often foregrounded, we believe that interpretation processes and the role of culture are also primary issues to be considered. As the data move through a communication process where visualizing the information can lead to certain kinds of judgments and actions[35], the visual representations become more than vehicles to communicate facts, but rather tools that reinforce certain kinds of value systems and cultural categories.[36]

3.5 Ethics and privacy

Across all four of these areas are issues of ethics and privacy; how the collection and use of the information about learners and teachers can be done responsibly while also safeguarding the privacy of those whose information is captured. These concerns are in many ways related to cultural matters since privacy and ethical choices related to the personal information can be influenced by what is culturally acceptable and normal. They are also in many cases shaped by the kind of legislative and regulatory frameworks that educational organizations operate under, especially in the United States where the Constitution delegates most educational matters to the states. This allows each state to implement its own approach to data governance. However, the U.S. Family Education Rights and Privacy Act (FERPA) details the minimum expectations for all custodians of educational data and describes what is allowable to share about students and under what circumstances. Generally, informed consent is required for the use of most identifiable information.

There is a paradox in the privacy landscape in that while public entities in the U.S. are required to adhere to FERPA and other regulations involving data sharing, in the private sector there are fewer restrictions and less regulation regarding the data collection and use. A person searching on the Internet is creating a log of activity that can be used by various entities engaged in marketing and use profiling. At this time, particularly in the U.S., students learning with online digital resources are likely to be providing data that the vendors of those environments might be able to use for a range of purposes without the same regulations as pertain to the data collected by schools and other institutions. The Children's Online Privacy Protection Act (COPPA) is potentially applicable in these cases, but it is not specific to education and it addresses different concerns.[37]

4. A UNIFIED PERSPECTIVE FOR THE EDUCATIONAL DATA SCIENCES

In advancing this view of a common field of Educational Data Sciences, we are arguing that the four areas will benefit from some common concepts and principles specific to their systemic and socio-technological nature. We see the four areas as important for guiding the development of this field to help researchers and developers use a common language, conceptual tools, and principles while also being adaptable to the range of contexts in which educational data are used productively.

The list of areas and our discussion is indicative rather than exhaustive. But we believe that these four areas present key language, concepts, and principles that are crucial to the success of the field of EDS and the work of educational data scientists. More specifically, in preparing future EDS professionals, we believe these four types of disciplinary knowledge are needed.

4.1 Appreciate the Distinctive Character of Education Data

Using data and information for systemic improvement is an important reason for there to be an Educational Data Sciences. In important ways this field is like the other areas where data are used in other domains such as health care, finance, and industry. In some other ways, educational data has unique properties. These unique properties include:

1. **Human/social creation.** Unlike most other fields that use data, much of educational data requires human manipulation, which increases the possibility of error and manipulation. Some have focused on cheating and gaming of the system in the area of tests, but this property is actually much more pervasive and affects areas like special education planning and school improvement plans as well as assessments.
2. **Measurement imprecision.** Educational data can be rife with issues of precision, especially when assessments of student learning or systemic capabilities are used. Compared to blood pressure readings or financial transactions, educational assessments are noisy. They can be sensitive to student background, instructional techniques, circumstances of testing, and the like
3. **Comparability challenges.** Comparisons across different areas of educational data can be sometimes impacted by contact variation. For example, different schools are often compared for many different kinds of analyses. However, programmatic variation often occurs from school to school and those programmatic differences may not always be apparent in the data streams.
4. **Fragmentation.** The world of educational data is fragmented. Many different organizations hold parts of educational information and there are still incomplete and partially adopted technical standards which impacts the ability to link some data without specific extra work. There are a number of efforts to create interoperable data standards. While progress has been made in these areas the road forward will be difficult as the governance of educational data is highly decentralized owing to the US Constitution's delegation of authority for education to states and across the states there are many different approaches to state-district interactions and almost 20,000 district and charter providers.

While these conditions do not make educational data impossible to use they do, in our view, impact the kind of work that educational data scientists can do and the kind of preparation they need.

4.2 Embrace Interdisciplinary Perspectives

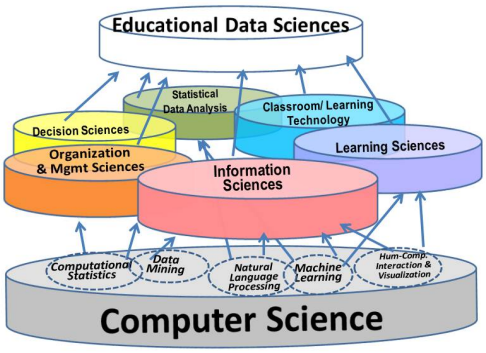

Figure 3 - Education Data Sciences interdisciplinary connections (Piety, Pea, and Behrens, 2013)

As we turn attention toward the nature of the Educational Data Sciences, it is important to begin with its interrelationships with other fields. In sketching out the disciplinary influences for the Educational Data Sciences, Piety, Behrens, and Pea identified seven different disciplinary connections. Six adjoin this new field. One, Computer Science, is generative in terms of innovations that enable data sciences and the fields that apply data sciences products.[38] They describe these fields, illustrated in Figure 3, as being joined by a set of dispositions and attitudes towards using data in the multileveled social context of education.

well as organizations, and some ways that data is used in other organizational settings, including business. As shown in Figure 3, the six fields of Classroom Learning, Learning Sciences, Information Science, Decision Sciences, Organizational Science, and Statistics are all adjoining fields to EDS and influence it. Computer Science, however, has a special relationship to EDS in that it not only generates innovations that can directly impact the tool complexes that are used, but also can provide these kinds of innovations for other fields that can amplify their impact on EDS. Natural Language Processes is a good example of this kind of innovation activity that can have broad applicability to EDS and its neighboring disciplines.

4.3 Recognize Social/Temporal Levels

One of the observed patterns of EDS is how it can relate to many different social and temporal scales or levels. It can look at the moment-to-moment interactions of students with learning material and with each other. It can also relate to that same student over the course of a semester, a year, or an educational career. Similarly the analysis can be at the level of students, classes, departments, programs, schools, and institutions and systems. The analysis can be of a single one of these kinds of entities or can cross multiple entities for comparison or group analysis. We see some of the work on ecosocial levels and timescales pioneered by the sociocultural semiotic theorist Lemke as an appropriate starting point for our theoretical model.[39]

Table 1 – Taxonomy of educational evidence and functions acording to timescale and ecosocial level (Hickey & Zuiker, 2013)

Timescale Context	Targeted Educational Content	Relative Time Frame	Format of Educational Evidence	Appropriate Formative Function for Students	Ideal Formative Functions for Others
Immediate	Specific Curricular Activity (lesson)	Minutes	Event-oriented *observations* (Informal observations of the enactment of the activity)	Discourse during the enactment of a particular activity.	Teacher: Refining discourse during the enactment of a particular activity.
Close	Specific Curricular Routines (chapter/unit)	Days	Activity-oriented *quizzes* (semi-formal classroom assessments)	Discourse following the enactment of chapter or quiz.	Teacher: Refining the specific curricular routines and providing informal remediation to students.
Proximal	Entire Curricula	Weeks	Curriculum-oriented *exams* (Formal classroom assessments)	Understanding of primary concepts targeted in curriculum.	Teacher/curriculum developer: providing formal remediation and formally refining curricula.
Distal	Regional or National Content Standards	Months	Criterion-referenced tests (external tests aligned to content standards)		Administrators: Selection of curricula that have the largest impact on achievement in broad content domains.
Remote	National Achievement	Years	Norm-referenced tests (external tests standardized across years, such as ITBS and NAEP)		Policy makers: Long-term impact of policies on broad achievement targets.

the next level up. For example, a distribution of student performance in a course assessment could lead to categories of students (proficiency levels) when looking at the whole course. When looking across courses, these proficiency categories could then be places into distributions again when being seen across courses. This characteristic of educational data can be seen broadly. Consider students entering a program being categorized demographically. When we look another level up, across programs, the key information is the distribution of students in the programs. At a higher level, it is likely that programs would be classified by the ways they attract and retain students of different backgrounds. The demographic categories become continuous variables (perhaps in combination with other variables) at the higher level. Borrowing from the field of energy transformation, Lemke called this principle of alternating symbol types the "Adiabatic Principle." While this principle is not a rule in that all educational data does not behave this way, it is a very common feature of this domain.

As we apply this timescales model to educational data we see opportunities to design appropriate uses of data. Take, for example, the accountability data collected under NCLB. These annual tests, imperfect measures under the best of circumstances, were proposed in the early educational data movement as being useful in informing classroom instruction. One of the signature difficulties of NCLB was that practitioners found these data had little relevance to what was being taught to students at a given point. This issue can be seen as a problem with testing as many have characterized it.[40] It can also be seen more positively when this timescales model is viewed as a *temporal mismatch* where semiotics that have some value for one ecosocial level (school accountability) are being used in an attempt to improve another level (classroom instruction).

Researchers who are considering the assessment and accountability from a sociocultural perspective are uncovering ways to re-conceptualize assessment practices in the context of their distance from the immediate enactment of classroom curricula.[41] Table 1 illustrates how assessment data can be seen as relating to different scales of activity. The more typical distinction between summative and formative purposes is set aside in favor of assessment *functions,* which leads to the argument that all assessments and, by extension all educational data, have formative *potential.* This potential varies from one level to the next. Recognizing that the timescales get longer and the evidence gets more formal across levels helps reveal and exploit the formative potential within and across levels.

4.4 Recognize Digital Fluidity

Digital fluidity refers to the capability for digital information of all kinds to be easily transported from one context to another. This means that the data collected for one purpose, for example the state longitudinal databases, can be used with relative ease by school teams or classroom teachers provided the information has value to them. Likewise, the data collected from classroom technologies could be easily transported for use in analysis at the school or district level. This characteristic of the digital age is one of the drivers behind a need for better conceptual frameworks. Figure 4 illustrates this concept from the domain of K-12 Systemic/Instructional Improvement community where there are three primary kinds of information resources that can be used by individuals at many different ecosocial levels from students to district to state analysis.

Digital fluidity comes from the very nature of digital technologies based on common standards of exchange. The photograph taken on a phone can be emailed and then posted in a social networking site quickly and with no additional cost.[42] This type of flexibility is only beginning to make an impact in education because educational data is often not standardized and because educational organizations are still developing a capacity for data similar to what other fields had many years ago. Both of these conditions are changing in education, but at a much slower pace the most other information sectors.

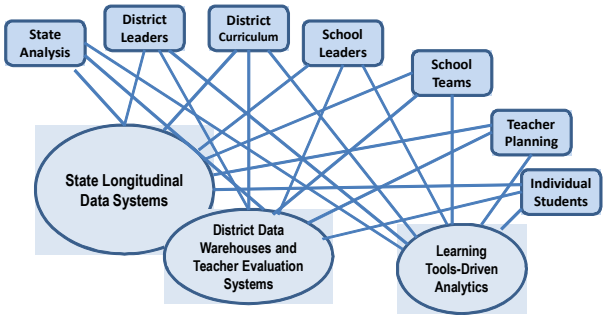

Figure 4 - Example of digital fluidity as same data can be used in many different settings

While educators have email and social networking and state-of-the-art digital media, they do not yet have the widespread availability of interoperable tools that allow information to flow freely. Many lack the requisite experience to use the range of data when it is available. As educational tools become more developed and interoperable, and as educational data exchange standards mature, the exchange of information across settings and for different uses will become easier so that the connections between data resources and analytic processes may resemble the combinations illustrated in Figure 4.

The fluid nature of digital data can be a catalyst for reshaping evidentiary practices for educators. If not today then in the near future, a classroom teacher will be able to draw upon information from different kinds of information resources in state, district, and classroom data to make decisions about groups of students or individuals. Although the contribution each kind of resource might make in each setting is different, all of these categories of information are potentially relevant to a range of data use practices. Likewise, the mix of information that one teacher. school team or department might choose to focus on could differ from another. Still, the broad availability of data that is developing in education means that recognizing and understanding these different possibilities calls for a kind of framework that can accommodate both a range of data and range of applications. The need for such knowledge naturally increases as new kinds of information resources proliferate and converge.

4.5 Understand Values Embedded in Information Architecture

One of the most important re-orientations for EDS is from an instrumental view of educational data to one that accounts for the complex and dynamic ways that these technologies can work in actual practice. Often, technological solutions are presented to educators as an instrument that can or should deliver a specific kind of change in practice (ex: data-driven decisions or blended

learning). We find that the relationships between technologies and educational practice are rarely that direct. In practice they are dynamic and contingent. Practitioners adopt and adapt various tools and how they do so can change and evolve over time. According to the sociotechnical theorist Bijker, "[one] should never take the meaning of a technical artifact or technical system as residing in the technology itself. Instead, one must study how the technologies are shaped and acquire their meaning in the heterogeneity of social interactions."[43] We see these realizations impacting the work that education data scientists do in order to understand the evolutionary nature of the technologies.

More specifically, there appear to be dialectical relationships between these information technologies and the broader activity systems they mediate.[44] These dialectical relationships can involve technologies that enable certain kinds of activities and constrain other types of activities. This is because these technologies make some evidential practices more possible or practical than others. As they become parts of infrastructures, these technologies—through their integration in daily praxis— make some kinds of activity easier than others. Their designs called at times *information architectures* then can be theoretical as well as practical. These architectures can encode various theories of learning that manifest themselves in the data the tools provide.[45]

We believe that we are witnessing a paradigm shift in the conception of what educational data means from a traditional focus on correct/incorrect student responses to a range of information about what is occurring in the educational processes.[46] One of the important implications of digital technology is that any activity that uses digital tools can produce digital artifacts that can be easily pulled into one or more spheres of analysis. Rather than needing external tests to produce some evidence of educational progress, the classrooms themselves can use local tools to produce meaningful data about what is occurring within them. This means, for example, that data collected externally and produced internally can be connected and used for analysis. Online activity, high-school graduation, college application and acceptance, and the different steps that students take within any PreK through College learning environment can generate a digital artifact. These digital artifacts can then be combined and put into an analytic frame. This creates opportunities to see educational practice in new, more public ways that are also limited and constrained by what the data allows and the evidentiary and interoperability standards the data were created to what?. Some data may be high quality, but not easy to exchange. Some may exchange well, but not be high quality.

5. CONCLUSION

In this paper we have covered four new and growing communities associated with educational data that we frame as nascent cells of an emerging field of Educational Data Sciences. This is a field that we believe is sociotechnical and trans-disciplinary. Working within it will call for a combination of technical and social skills; an aptitude for engineering and also the deep understanding of the complex world of educational practice and learning across settings.

The world of these new sciences is dynamic. It involves much change related to both technologies and the innovations that come from them as well as policies and national expectations around the use of data. Past efforts, including NCLB and VAMs in the K-12 space have led to great disappointment and much resistance from educators around data-oriented initiatives. Educational data scientists will need to navigate these expectations as they promote new solutions and approaches. They will need to be diplomats as well technocrats, to be students of the often particular ways that teaching and learning happen before they will see how their innovations and approaches can be used responsibly for the benefit of entire educational systems.

REFERENCES

[1] Kuhn, Thomas S. The structure of scientific revolutions . University of Chicago press, 1996.

[2] Fagerberg, Jan, and Bart Verspagen. "Innovation studies—The emerging structure of a new scientific field." Research policy 38.2 (2009): 218-233.

[3] Goldstein, Philip J., and Richard N. Katz. Academic analytics: The uses of management information and technology in higher education. Educause, 2005.

[4] Goldstein, Philip J., and Richard N. Katz. Academic analytics: The uses of management information and technology in higher education. Educause, 2005.

[5] Campbell, John P., Peter B. DeBlois, and Diana G. Oblinger. "Academic analytics: A new tool for a new era." *Educause Review* 42.4 (2007): 40.
Norris, Donald, et al. "Action Analytics: Measuring and Improving Performance that Matters in Higher Education." *Educause Review* 43.1 (2008): 42-44.

Baepler, Paul, and Cynthia James Murdoch. "Academic analytics and data mining in higher education." (2010).

[6] Goldstein, Philip J., and Richard N. Katz. Academic analytics: The uses of management information and technology in higher education. Educause, 2005.

[7] https://confluence.sakaiproject.org/pages/viewpage.action?pageId=75671025

[8] Romero, Cristobal, and Sebastian Ventura. "Data mining in education." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3.1 (2013): 12-27.

[9] Baker, Ryan. "Educational Data Mining: Potentials and Possibilities." Paper for the American Educational Research Association annual conference. April 29, 2013 San Francisco, California.

[10] Siemens, George, and Phil Long. "Penetrating the fog: Analytics in learning and education." Educause Review 46.5 (2011): 30-32.

[11] Baker, R. *Educational Data Mining: Potentials and Possibilities.* Paper for the American Educational Research Association annual meeting. San Francisco, CA. 2013.

[12] Siemens, George. "What Are Learning Analytics?" Elearnspace, August 25, 2010.http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/

[13] Buckingham Shum, S. and Ferguson, R., Social Learning Analytics. Educational Technology & Society (Special Issue on Learning & Knowledge Analytics, Eds. G. Siemens & D. Gašević), 15, 3, 3-26. http://www.ifets.info Open Access Eprint: http://oro.open.ac.uk/34092 (2012),

[14] Siemens, G., Baker, R. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. (2012).

[15] Baker, R. *Educational Data Mining: Potentials and Possibilities.* Paper for the American Educational Research Association annual meeting. San Francisco, CA. 2013.

[16] U.S. Department of Education Office of Educational Technology. Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief. U.S. Department of Education (2012) Office of Educational Technology Washington, D.C., 2012.

[17] Glaser, Robert. The new aptitudes and adaptive education. Council on Anthropology and education Newsletter volume 4, issue 2, pages 20-26, July 1973.

[18] http://www.apsu.edu/academic-affairs/degree-compass-and-my-future

[19] Nussbaum, A D. "Situational disengagement and persistence in the face of adversity." Journal of experimental social psychology 43.1 (2007):127.

[20] Act, No Child Left Behind. "Public Law 107-110." Washington, DC: US Congress. Available at: www2.ed. gov/policy/elsec/leg/esea02/107-110.pdf. 2002.

[21] Piety, Philip J. Assessing the Educational Data Movement. Teachers College Press, 2013.

[22] Coburn, Cynthia E., and Erica O. Turner. "Research on data use: A framework and analysis." Measurement: Interdisciplinary Research & Perspective 9.4 (2011): 173-206.

[23] McCaffrey, Daniel F., et al. Evaluating Value-Added Models for Teacher Accountability. Monograph. RAND Corporation. PO Box 2138, Santa Monica, CA 90407-2138, 2003.
Harris, Douglas. "Value-added models and the measurement of teacher productivity." (2011).

[24] Means, Barbara, Christine Padilla, and Larry Gallagher. "Use of Education Data at the Local Level: From Accountability to Instructional Improvement." US Department of Education (2010).

[25] American Journal of Education, 112(4); American Journal of Education 118.2 (2012); Teachers College Record 114.11 (2012).

[26] Eisenstein, Elizabeth L. The printing press as an agent of change. Vol. 1. Cambridge University Press, 1980.

[27] See the work of National Student Clearinghouse, including Hossler, Don, et al. "Transfer and Mobility: A National View of Pre-Degree Student Movement in Postsecondary Institutions. Signature Report 2." National Student Clearinghouse (2012).

[28]Goldhaber, Dan, Stephanie Liddle, and Roddy Theobald. "The gateway to the profession: Assessing teacher preparation programs based on student achievement." Economics of Education Review (2013).

[29] Arnold, Kimberly E. "Signals: Applying Academic Analytics." Educause Quarterly 33.1 (2010): n1

[30] Piety, Philip J. *Assessing the Educational Data Movement.* Teachers College Press, 2013.

[31] Behrens, J. Mislevy, R; Piety; and DiCerbo (2013). Evidence Centered Design for Learning Analytics. Commissioned paper for Educational Data Sciences Project. Roy Pea, principal investigator

[32] Christensen, Clayton. The innovator's dilemma: when new technologies cause great firms to fail. Harvard Business Press, 1997.

[33]Shapiro, Carl, and Hal Varian. Information Rules: A Strategic Guid. Harvard Business Press, 1998.

[34] Tufte, Edward R., and P. R. Graves-Morris. The visual display of quantitative information. Vol. 2. Cheshire, CT: Graphics press, 1983.
Shneiderman, Ben. "The eyes have it: A task by data type taxonomy for information visualizations." Visual Languages, 1996. Proceedings., IEEE Symposium on. IEEE, 1996.

[35] Goodwin, Charles. "Professional vision." American anthropologist 96.3 (1994): 606-633.

[36] Bowker, Geoffrey C., and Susan Leigh Star. Sorting things out: Classification and its consequences. The MIT Press, 2000.

[37].Children's Online Privacy Protection Act of 1998 (COPPA). United States federal law 15 U.S.C. §§ 6501–6506 (Pub.L. 105–277, 112 Stat. 2581-728, enacted October 21, 1998 http://www.gpo.gov/fdsys/pkg/PLAW-105publ277/html/PLAW-105publ277.htm

[38] Piety, P; Pea. R; Behrens, J (2013). *Big Data in Education: Arguing for an Educational Decision Sciences*. Paper for the American Educational Research Association annual meeting. San Fran., CA. 2013.

[39] Lemke, Jay L. "Across the scales of time: Artifacts, activities, and meanings in ecosocial systems." Mind, culture, and activity 7.4 (2000): 273-290.

[40] Nichols, Sharon Lynn, and David C. Berliner. Collateral damage: How high-stakes testing corrupts America's schools. Cambridge, MA: Harvard Education Press, 2007.

[41] Hickey, Daniel T., et al. *Balancing varied assessment functions to attain systemic validity: Three is the magic number.* Studies in Educational Evaluation 32.3 (2006):180-201.

[42] Shapiro, Carl, and Hal Varian. Information Rules: A Strategic Guid. Harvard Business Press, 1998.

[43] Bijker, Wiebe E. Of bicycles, bakelites and bulbs: Toward a theory of sociotechnical change. The MIT Press, 1997.

[44] Cole, Michael, and Yrjö Engeström. "A cultural-historical approach to distributed cognition." Distributed cognitions: Psychological and educational considerations (1993): 1-46.

[45] Bowker, Geoffrey C., Karen Baker, Florence Millerand, and David Ribes. "Toward information infrastructure studies: Ways of knowing in a networked environment." In International handbook of internet research, pp. 97-117. Springer Netherlands, 2010

[46] DiCerbo, K. E., and J. T. Behrens. "Implications of the digital ocean on current and future assessment." In R. Lizzets & H. Jiao (Eds.) Computers and their impact on state assessment: Recent history and predictions for the future (2012): 273-306.