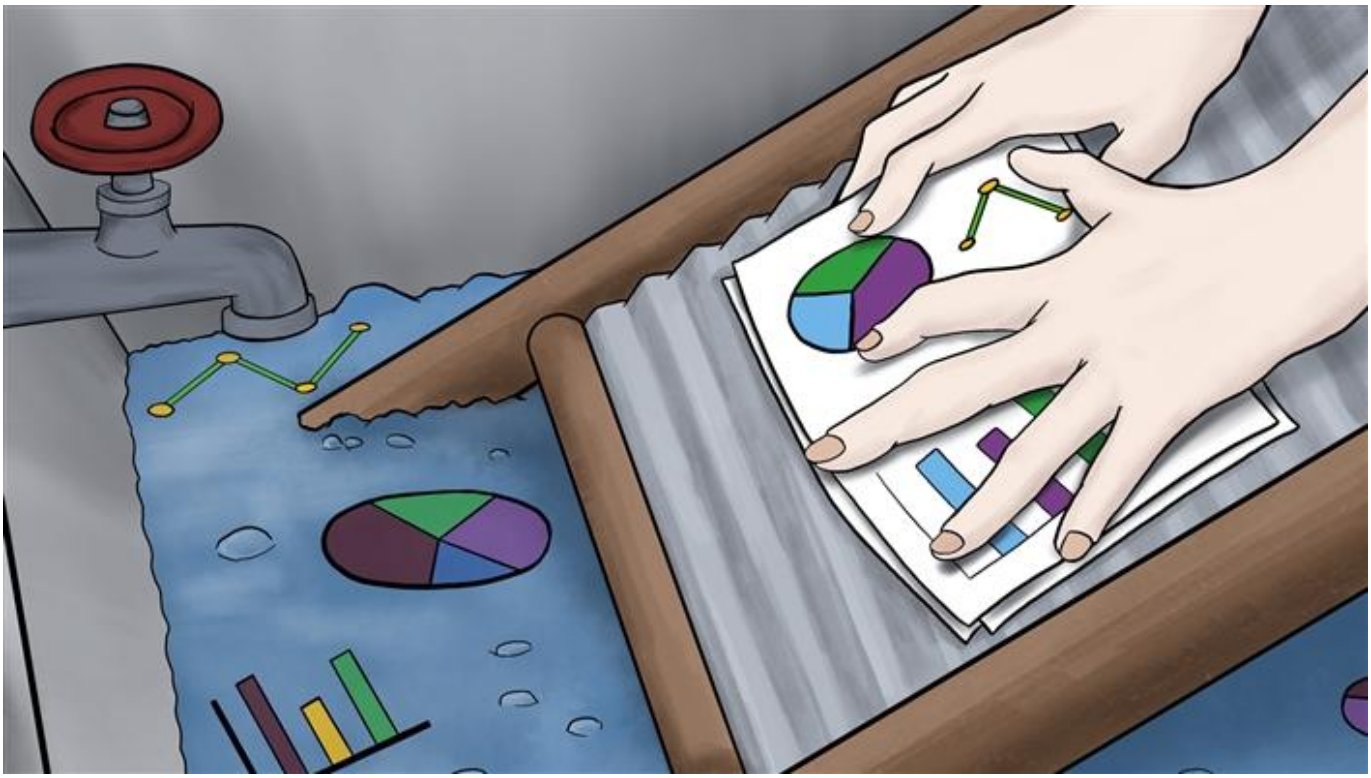
Why IT Matters to Higher Education
EDUCAUSEreview

Who Is Doing Our Data Laundry?

by **Brad Wheeler** Monday, March 13, 2017 Editors' Pick

We are seeing a surge in firms with offers to take institutions' data so that they can reformat it and make it available as dashboards, with trends and models. It is time to ask: Who is doing our data laundry, and why?



We all like to wear clean clothes, but few of us enjoy the chore of doing the laundry. Sorting, washing, drying, folding, ironing, buying soap, and repairing the washer are often viewed as a necessary distraction from how we would prefer to spend our time.

Sometimes we let the clothes pile up, and other times the laundry becomes urgent. Some of us dispense with the chore altogether and outsource it to a dry cleaner or hire help.

Similarly, the data at our institutions is piling up and taking on a new urgency for the challenges ahead. Who is or should be doing an institution's data laundry? By **data laundry**, I am referring to *the legitimate process of transforming and repurposing abundant data into timely, insightful, and relevant information for another context*. It is a mostly unseen, antecedent process that unlocks data's value and insights for the needs of decision makers. I am not referring to *data laundering,* which seeks to obscure, remove, or fabricate the provenance of illegally obtained data for nefarious purposes.[1]

At least two forces are motivating an accelerating demand for this process. First, market forces are changing the economics of higher education as financing a college education becomes more of a private good and less of a subsidized public good. Greater transparency of information also fuels increased shopping among colleges and universities by students, parents, donors, and state legislatures as they make choices for educational investments.[2] A second force is the growing belief that computational analyses of big data from colleges and universities will yield formerly unknown insights regarding new efficiencies and effectiveness in the competitive market.[3]

These motivating forces are pervasive across higher education for institutions of all sizes and missions, both public and private. The thirst for insightful information comes from students, faculty, department chairs, deans, provosts, presidents, boards, and others who seek a rapid means to more effectively manage their investments in the academy. They sometimes seek reports or real-time dashboards with newly tailored information, but increasingly they expect analysis, prediction, and benchmarking based on historical data to inform the future. Some of the thirst may also emerge when sales pitches assert cloud-based services as a quick "magic bullet" to solve local information needs.[4]

Our institutions are often quite data-rich and insight-poor. We have an abundance of primary data from our many systems of record (e.g., Student Information Systems, Finance, Learning Management) spanning a decade or more, yet few of us have made the substantial investments internally to repurpose our data for new insights. Add to that many new high-volume sources of data from social media or geolocation tracking,

and that gap between real or perceived internal inabilities and an institution's urgent needs has driven many to seek external help—to send the data laundry out to others—in hopes of a quick win by accelerating information insights.

Dirty Clothes and Institutional Data

What do our unanalyzed terabytes of 1's and 0's have in common with dirty socks and soiled blouses? They all involve a cost of transforming from a prior, undesired state to a future, desired state in a process that requires a mix of capital, labor, and expertise and that must be performed under time and quality constraints. Finally, the perceived quality of an outcome is deeply influenced by the context in which it is used.

Clothes Laundry

At home, most of us rarely consider the routine, tacit assumptions that shape our process choices for laundry. We just want clothes that are clean, folded, pressed, and hung up or put away. Some of us may even be meticulous with beautifully arranged closets. We make decisions for inputs of capital and labor, investments of time and supervision, and allowances for scheduling and workflow. Then we (literally) rinse and repeat.

Many of us choose to manage this process in-house. We buy a spiffy washer and dryer (capital investment); we add consumable expenses in soap, bleach, spot remover, and fabric softener; and then we supply our own labor and expertise from hamper to hanger. For some special cases, we may outsource the process to take advantage of greater expertise—to get ink out of the white shirt or to care for the cashmere sweater that needs extra special handling at a dry cleaner, for instance.

Timing is important also. Saturday may be the weekly planned laundry day at home, yet other times we do a quick load at 10 p.m. for an urgent need. Sometimes, however, we realize on Sunday that we need something dry-cleaned for a Monday trip.

In the end, our choices for clothes laundry illustrate our willingness to invest our capital, time, and expertise and our preferences for scheduling to achieve our goal of a regular supply of clean clothes.

Data Laundry

Our choices for data laundry similarly allocate capital, labor, and expertise in forms that yield near- and longer-term costs and benefits. Campus leaders spend much

money to implement and operate critical systems of record that pass audits, successfully schedule students, and enable courses, yet often they are then told that the data from these systems is not ready to be used for other purposes. Why not?

The seeming disconnect is triggered when data that works fine for its original purpose and timeframe needs to be accurately repurposed for secondary uses. The data laundry process is required to transform the data that had integrity in the original context into meaning for purposes of changed context, integration, unanticipated uses, and normalization.

- *Context Changes:* A financial system may have accurate data for the revenues and costs of each academic school and department over five years. Then, in year six, there was a small reorganization of schools that was followed by a major reorganization of departments in year seven. Any scenario-planning reports, dashboards, or decision support systems that repurpose that historical data will need to transform it so that it will not lose its historical or predictive value.
- *Integrating:* Data that may make perfect sense in the Learning Management System may not be coded or in a format to integrate deeply with information from the Student Information System enrollment and degree map data. New data elements have to be calculated or cross-referenced to tie disparate sources of data together for information.
- *Unanticipated Uses:* A new financial policy may compel sophisticated analysis of building and space use to account for charges to federal grants. The systems that manage building and square feet of architectural renderings may have no relationship at all to the academic department structure. New needs—especially for new legal compliance—accelerate repurposing of data.
- *Normalizing:* If a decision maker wants to understand student progress and hours completed at the beginning of a senior year, the data could skew meaning. For example, students who major in engineering might have about the same number of hours completed as political science majors, yet the engineering students may require 12 more hours to graduate. Thus, any useful measure of "progress toward degree completion" would need to represent that progress in a normalized way across majors. Likewise, comparative data between schools on a campus, between campuses in a university system, or between independent institutions often requires extensive normalizing and contextualizing before it can be repurposed in any credible way.

The data laundry work includes two distinct and sequenced phases. First, *data cleaning* prepares the data for reuse in a different context. It is not unusual that fields in various systems have taken on creative and poorly documented uses over time. This data thus may need to be extracted and transformed into proper fields that may involve human interpretation. This phase includes extensive testing and quality control to ensure that the data retains its accuracy even as it is repurposed for a new context. It is important to note here that *accuracy* has both an objective and a perceptual connotation, since the data is judged by the users of the information and also through the context in which they interpret it.

By analogy, the product at the end of data cleaning is like a pile of clean clothes. The clothes still need finishing work such as folding, ironing, and placing on hangers to be fully ready to use. Clean data likewise needs additional work, and that is done in the second phase: *data presentation* contextualizes the information into reports, graphs, dashboards, and decision support systems. When done well, this presentation becomes smoothly integrated into campus workflows alongside other essential systems. Both cleaning and presenting are essential steps to support decision makers, and each phase requires distinct expertise and capabilities.

Campus Data, from Hamper to Hanger

Institutions of all sizes have long had some local staff engaged in campus data laundry, mostly involved in data reuse within context. For example, staff transform transactional data into useful financial statements, enrollment reports, and dashboards for executives. Although this within-context work has satiated campus needs and is likely to continue, new needs for cross-context work are accelerating.

Sourcing the Work

Over the last three to five years, with the need for integrated information across contexts increasing, many large-scale campus investments have been made in initiatives with labels like Big Data, Business Analytics/Intelligence, Decision Support, and Reporting. Some institutions have chosen to create considerable internal capabilities in technical staff, data scientists, software, databases, and computational systems that may reside with the CIO or the Institutional Research Office and/or are distributed among administrative departments.

Others have turned to the burgeoning number of firms that offer a range of additional data laundry services. These services provide formats for receiving data and operate

under extensive contractual agreements regarding compliance with the law and an institution's policies. Institutions may enable real-time access to extract the designated data from local or cloud-based systems, or in some cases, institutions may take data "snapshots" to be uploaded daily or at other intervals. In almost all cases, outsourcing the data laundry still requires considerable internal work, so many institutions are doing a blend of both.

The complete work of data laundry is largely the same processes no matter if done in-house or outsourced to others, but the results can vary due to assumptions or the depth of understanding of the business logic that created the original data. These nuanced assumptions are critical in repurposing data from its original context to a new one. Inadequate attention to detail can quickly undermine confidence in the accuracy of new uses. As one campus leader recently noted: "Get the data right from the beginning . . . once people have lost trust in the data, you've lost."[5] Figure 1 illustrates the challenges of that very detailed technical work. It is an excerpt from an email sent by an institution to an outsourced servicer while working on a data import.

"We have been researching the Registered Credit issue. When the student you note withdrew from <COURSE> in Sum, <YEAR>, the business logic that is used in PeopleSoft is different in noting this fact based upon what table is being updated – for example:

- Registered Credits (UNT_PRGRSS) is on two different tables that we send to you
  - o Term table - the basis of this table in PS is not course specific and for the student below, PS reflects this UNT_PRGRSS field = 0
  - o Course table – the basis of this table in PS is course specific and for the student below, PS reflects this UNT_PRGRSS field = 3

While you may think this is goofy, it is all tied into what counts and what doesn't count in different places like transcripts. In any event, by sending you UNT_PRGRSS field both your term and course tables will result in a mismatch.

You are using this data for an entirely different purpose and it seems as though we need to use an algorithm for the course table instead of UNT_PRGRSS to get the two to match on your end for Registered Credits.

We think what we should be sending for UNT_PRGRSS in the course table is 0 credit hour (to agree with the value in the term table) because the student is no longer registered in the class once he/she drops or withdraws.

For Attempted Credits **new** field for a term, we would derive this count by including both the dropped courses (once term starts) and the registered courses.

For the Lifetime Attempted Credits issue, it would seem that that logic needs to be the same algorithm as above except for all terms combined. Please let us know your thoughts on this and if we need to actually have a phone conversation."

Figure 1. The challenges of maintaining data

The Data Laundry Process

The data laundry process of data cleaning and presenting can be further illustrated by the nine steps in figure 2. Red boxes indicate a *presumed* advantage and the likely sourcing of the work for those

steps. Steps with thick black borders can be especially challenging. Steps with dashed borders are sometimes optional depending on the project.

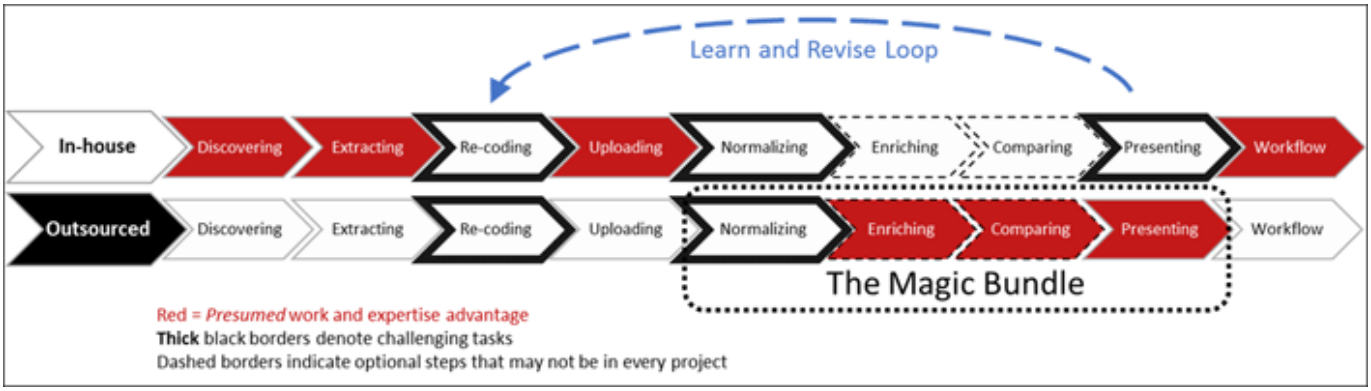


Figure 2. Data laundry steps: Who does what?

Even if you send all your home laundry to a dry cleaner, you still need to gather clothes, sort them, provide any special instructions, and possibly handle drop-off and pick-up. Similarly, campus data laundry requires multiple steps whether done fully in-house or in a combination of in-house and outsourced. Table 1 provides more detail for each task in figure 2 and illustrates the roles for various sourcing choices.

Table 1. Data laundry tasks

Tasks	In-House (or via Consultants)	Outsourced
Discovering data sources that support the end objective	Knowing where data resides in campus systems	Possibly knowing of external data sources
Extracting data from systems of record	Conducting considerable functional and technical work to extract data from core systems	Technical consulting
Re-coding data to preserve context and conform to new formats	Using local expertise or work to convey business logic and to format data	Knowledge of intended uses and models required

		to adapt context to new purpose
Uploading	Holding extensive internal reviews of uses for legal and policy compliance before data is uploaded to campus or vendor systems, real-time/batch	Technical consulting
Normalizing	Normalizing data to new context and as inputs for models	Normalized to vendor's models
Enriching	Enriching with local predictive models	Enriched with vendor's predictive models or peer benchmarks
Comparing	N/A or internal within/across an institution	Enriched with normalized peer group comparisons
Presenting	Providing dashboards, screens, reports	Dashboards, screens, reports
Workflow (often never achieved)	Integrating the new presentations into workflow and other campus systems	Technical consulting

As in any workflow, each step in this process also invokes costs to coordinate with the previous step and the next step. Thus, there is an implied efficiency when adjacent steps are under common management and within the same organization. Conversely,

there may be experience, expertise, or considerable cost advantages of sourcing some steps in-house or through an experienced vendor.

Many vended data laundry services are pitched as an input-output product bundle with the offer of "send us your data, and we'll send you back insights." This narrow focus is the "Magic Bundle" in figure 2. The term *magic* is intended to be neither descriptive nor pejorative, but many of these services do operate as a black box with proprietary operations for what happens between data-in and information-out. Finally, experience shows that each new insight from repurposed data often triggers deeper questions and a need for additional recoding or another analysis. This begins the "Learn and Revise Loop," which may again trigger additional policy and legal reviews regarding data use.

Sorting the Data Laundry

These days, many data laundry efforts, both small and large, seem to be popping up all over campuses, and each merits some deeper scrutiny before hiring another staff member or signing another contract. The following questions can help assess particular projects and possible outcomes from sourcing choices.

What's the Imperative?

First, what problems or motivations are driving current and future initiatives? Examples might include a need to

- rapidly remediate information reporting and dashboards to get everyone on a common basis of facts;
- accelerate student success goals in terms of graduation rates, retention, course sequencing, etc.;
- empower advisors and students with predictive information to support wise decisions in major and course selection;
- enable better financial decisions by deans and department chairs as they make choices in degrees to offer, frequency of courses, and teaching assignments; and
- benchmark an institution's performance on key performance indicator ratios for efficiency and effectiveness.

For example, EAB has identified "Ten 'No-Regrets' Analyses" that institutions should provide to every department to help with specific cost or quality imperatives.[6]

Will the Data Be Ready for Use?

Untrusted information, no matter the effort and expense, will not be used. We have all grabbed a shirt that we thought was clean, then discovered that it wasn't yet ironed and ready to wear. The same is true of data. When first uses of repurposed information have obvious errors or make predictions that defy experience, it is often quite difficult to regain users' confidence in the information. Likewise, even accurate data that is beautifully presented undermines confidence if it merely affirms common wisdom—for example, high school GPA is a strong predictor of college GPA.

What Pace Is Required?

The pace to achieve useful insights for data laundry efforts often proves disappointing. The hard work of achieving accuracy and integration for an institution's context and workflow can be underestimated by both vendors and internal leaders. The early work-products of these efforts often cause excitement as new insights first become available, but then the initial enthusiasm fades as the project struggles with accuracy and loses speed.

How Good Is the Enrichment?

One of the great hopes of data laundry is that it can be enriched with insights that emerge from integration, predictive modeling, and benchmarking. For example, can analysis of numbers of sections taught by faculty, credit hour revenues by course section, and payroll costs per section yield valuable insights for department chairs who assign faculty to courses? The data for such analysis must be accurately integrated from student, HR, and payroll systems and then presented in useful ways that enable insight through graphical displays and scenario modeling. If the data is not insightful, then it won't be used.

Repurposed data also provides a remarkable input for mathematical models that may make personalized predictions based on similar historical data. These insights may help students, advisors, faculty, department chairs, and others make more-informed choices of where to invest their time and efforts. In most cases, the core models that underlie these predictions originated from academic research whether based on regression, collaborative filtering, or machine learning. When models are proprietary

and hidden, however, we often have few means to assess the quality of what we are buying.

For example, one large research university had engaged a service to help with some predictive analytics in the student area. The vendor's system projected some performance indicators based on many years of historical data that the institution had provided. Concurrently, the institution also engaged a local faculty member to assess the data. When it compared the two results, the professor's linear regression model—transparent and informed by local expertise—had higher predictive validity than the hidden model of the vended service.

Finally, enrichment for benchmarking via cross-institutional data can be especially challenging. For example, aggregating and normalizing student data with other institutions through highly standardized and normalized models often washes out a lot of important context from the original data. Excessive normalization risks making the benchmark comparisons of little insight or real value to any institution.

Who Owns What?

Dropping off a suit at the dry cleaner yields a claim check that can be exchanged back for the suit, and ownership remains quite clear. The processes of data laundry and data use, however, create yet more data. Who owns the source data, the repurposed data, and the byproduct data—and under what terms? Unlike a suit, digital data can be replicated so that multiple parties can each "own" their own copy of the data. Ownership is important since ownership affords decision rights regarding what can and cannot happen with your possessions. For example, can data that is repurposed for a financial dashboard be easily ported and used again for a separate student, HR, or facilities dashboard? Who makes the ownership decision for which data can be used where and under what terms?

Contracts attempt to codify much of this, but by their nature, they often lag behind the emergent uses of data. Some contracts also grant a formal data license that conveys certain clear rights to owners of replicated or emergent data. Contracts and licenses may also grant rights for anonymized or de-identified uses of data for further aggregation and benchmarking. The latter services usually involve monetizing the data in some way, and licenses may also describe royalty and revenue-sharing arrangements. One thing is certain: all contracts should clearly describe how to completely retract an institution's data if it wishes to exit a service.

How Will the Data Evolve?

Data laundry efforts rarely have a definitive end, and this is particularly true when enriching involves predictive models and benchmarking. Each new insight spurs additional questions or desires for another dashboard feature. How will those features evolve and get prioritized and at what pace? Locally controlled systems sometimes struggle to keep pace with the innovation of commercial efforts, and vendors often prioritize features that help them acquire new customers over the wish lists of current customers. Who has the rights to make decisions for both pace and features? The answer to this question is important for the ongoing success of a system to serve evolving needs.

Beyond providing answers to these questions for any particular or urgent data imperative, institutions should also think about how numerous initiatives may fit together as part of an overall, multiyear data strategy.

Crafting a Data Laundry Strategy

Data laundry initiatives tend to grow in scope and complexity over time, and that is true whether they start as a limited-scope, in-house project or as a broader vended service. Future projects often build on and integrate with prior efforts, and projects with a seemingly narrow domain (e.g., admissions) may soon need to be integrated with other projects from the Finance or Curriculum Office as secondary uses of data grow in value when meaningfully integrated.

Thus, campus leaders are well advised to craft and execute an intentional, multiyear strategy for data laundry services. The thirst for insight will naturally propel many ad hoc initiatives, and sourcing choices for those projects will be made at different points of time by many separate departments that often have little knowledge of other choices across the institution. Some institutions have chosen to appoint a formal Chief Data Officer to set strategy to steer campus data efforts.[7] Some have worked with other institutions as a consortium to source data and benchmarking projects to the expertise of one of their members.[8]

With no explicit institutional strategy, "hope" becomes the tacit, de facto strategy to achieve institutional goals over time. It is possible that disparate choices will collectively turn out well for institutional goals, will be able to integrate where needed, and will be efficient in contracting and hiring, but that outcome is highly unlikely.

Implementing an intentional strategy requires communication, strong influence, and occasional authority (when necessary).

A well-crafted data laundry strategy informs trade-offs in achieving near-term goals, pace, longer-term economics, and institutional outcomes. It must not fall prey to the age-old siren song of *technological determinism,* which implies, "buy this technology, get that great organizational outcome!" Pitches from both internal staff and salespeople may lead executives to hear what they want to hear, in hopes of a magic bullet to rapidly cure their information thirst or a shortcut to solve a problem. Data laundry is hard to do, expensive over its lifecycle, and even harder to get right in ways that will materially matter to an institution's goals. Illusionary shortcuts often waste resources and, ultimately, disappoint.

The In-Sourcing Strategy

Institutions with sufficient size, scale of IT operations, and functional staff may choose to engage the work of data laundry through investments in additional staff, expertise, and technologies to create an internal capacity. There are some favorable and unfavorable considerations in this strategy.

Favorable:

- Greater control over features and pace across the entire set of data laundry steps (see figure 2), with lower coordination costs throughout the process
- Easier to maintain fidelity to an institution's context to accurately repurpose data
- Greater ability to tailor the fit of data presentation to an institution's real needs without adding things that are not valued
- Greater ability to integrate insight presentation into an institution's workflow with other existing systems
- Retaining institutional knowledge in-house for application across successive projects
- May draw on faculty and staff expertise for highly targeted and valued data enrichment with proven efficacy for an institution's context
- Potential life-cycle cost savings for a project and across a range of projects if resulting systems are owned and further repurposed rather than annually

rented

Unfavorable:

- Full responsibility to manage all the steps of the data laundry service from start to effective use, though unfavorable only if not executed well
- May be laboriously re-creating internal systems capabilities that already exist in a fairly priced, vended bundle of services
- May lack expertise or may struggle to hire and retain sufficient expertise for parity outcomes—particularly in the data-enriching and data-presentation steps
- May be slower to achieve near-term goals

The Outsourcing Strategy

Likewise, there are some favorable and unfavorable considerations in an outsourcing strategy.

Favorable:

- Greater speed, since vended solutions already exist and remaining work is pre- and post-bundle (see figure 2)
- Available to all and does not require hiring new staff for expertise, investing in technology, or managing the entire process
- Subscription service that may be turned off if no longer valued
- Access to data enrichment that may include proprietary knowledge and cross-institutional benchmarking

Unfavorable:

- May involve long internal legal, policy, and security reviews and approvals
- May incur costs of local staff to perform the pre-upload steps and the integration work (see figure 2), mitigating some of the presumed speed and value of the service

- Generic models and insights, which may lose so much institutional context that they offer little real insight
- May not integrate well with other efforts or other vended data laundering services from rivals
- Perpetual new annual cost or service ends

The Smart-Sourcing Strategy

For many institutions, a blended, smart-sourcing strategy is advised. I caution against overbuilding internal capabilities that lack agility for rapidly evolving institutional needs, and I also caution against stacking up a growing list of one-off contracts that may solve one problem but that, over time, create others.[9] Smart-sourcing seeks to avoid the dogma of either approach. It affirms a view that internal capabilities may be economically efficient over the lifecycle of a set of projects while targeted outsourcing can also be the best match for other needs. Outsourcing can be quite effective for well-understood, proven solutions. There is wisdom in the CIO adage of "outsource things you understand well."

Is We Data Illiterate?

Finally, I strongly encourage institutions to look among their own faculty—in schools or departments of business, computer science, informatics, mathematics, statistics, and other areas—to draw on their deep expertise. The underlying algorithms that power any predictive model for data enrichment are likely well understood within the academy, and magic bundles may seem less magical when we dig deeper. Many institutions, particularly research-intensive institutions, have within their ranks deep expertise in big data and modeling. Formally engaging the skills of faculty, post-docs, graduate students, and staff experts may be a very wise use of resources.

Will Clean Data Make Us Wiser?

Neither the best washing machine nor the most deluxe concierge dry-cleaning service can make us appear fashionable if our clothes don't fit. Likewise, the real efficacy for any data laundry effort cannot be measured unless the data is repurposed to fit our new needs. Such efforts must be assessed by their ability to inform—and then spur—wiser decisions as we navigate the shifting economics of higher education.

Almost fifty years ago—in the punched card and green-striped-line printer era—Russell Ackoff, professor and director of the Management Science Center at the University of Pennsylvania, brilliantly labeled some of the rising technologies of his day as "Management Misinformation Systems."[10] He observed that the information provided by these systems often compounded or reinforced the ineffective biases of many decision makers. The information sometimes even made their decision behavior worse. Today's rush to quickly repurpose our data can risk compounding the problems that we seek to remedy. I believe that by giving some careful thought to who's doing our data laundry and why, colleges and universities can skillfully match real institutional needs with the data, models, and predictions that will truly help us make better decisions.

Notes

1. Rob O'Neill, **"Cybercriminals Boost Sales through 'Data Laundering,'"** *ZDNet,* March 16, 2015.
2. Brad Wheeler, **"Speeding Up on Curves,"** *EDUCAUSE Review* 49, no. 1 (January/February 2014).
3. Gordon Wishon and John Rome, **"Institutional Analytics and the Data Tsunami,"** *EDUCAUSE Review*, December 12, 2016.
4. M. Lynne Markus and Robert I. Benjamin, **"The Magic Bullet Theory in IT-enabled Transformation,"** *MIT Sloan Management Review* 38 no. 2 (Winter 1997).
5. Personal correspondence with the author.
6. EAB, **"Ten 'No-Regrets' Analyses"** (infographic), IT Forum, July 28, 2015.
7. See Purdue University, **"Purdue Places Priority on Data Analytics with New Chief Data Officer,"** news release, November 12, 2013.
8. See **"UMETRICS,"** Big Ten Academic Alliance website, accessed January 22, 2017.
9. Brad Wheeler, **"The Flyswatter Strategy of IT?"** EDUCAUSE CIO Constituent Group Listserv, February 29, 2016.
10. Russell L. Ackoff, "Management Misinformation Systems," *Management Science* 14, no. 4 (December 1967). Ackoff listed five faulty assumptions that remain quite relevant today.

Brad Wheeler is Vice President for Information Technology and Chief Information Officer for Indiana University and a professor of information systems in IU's Kelley School of Business.

Academic Information Systems, Administrative Systems, Analytics, Business Intelligence (BI), Data Administration and Management, Data Curation, Data Management Planning, Data Mining, Data Privacy, Data Warehouse, Information Systems and Services