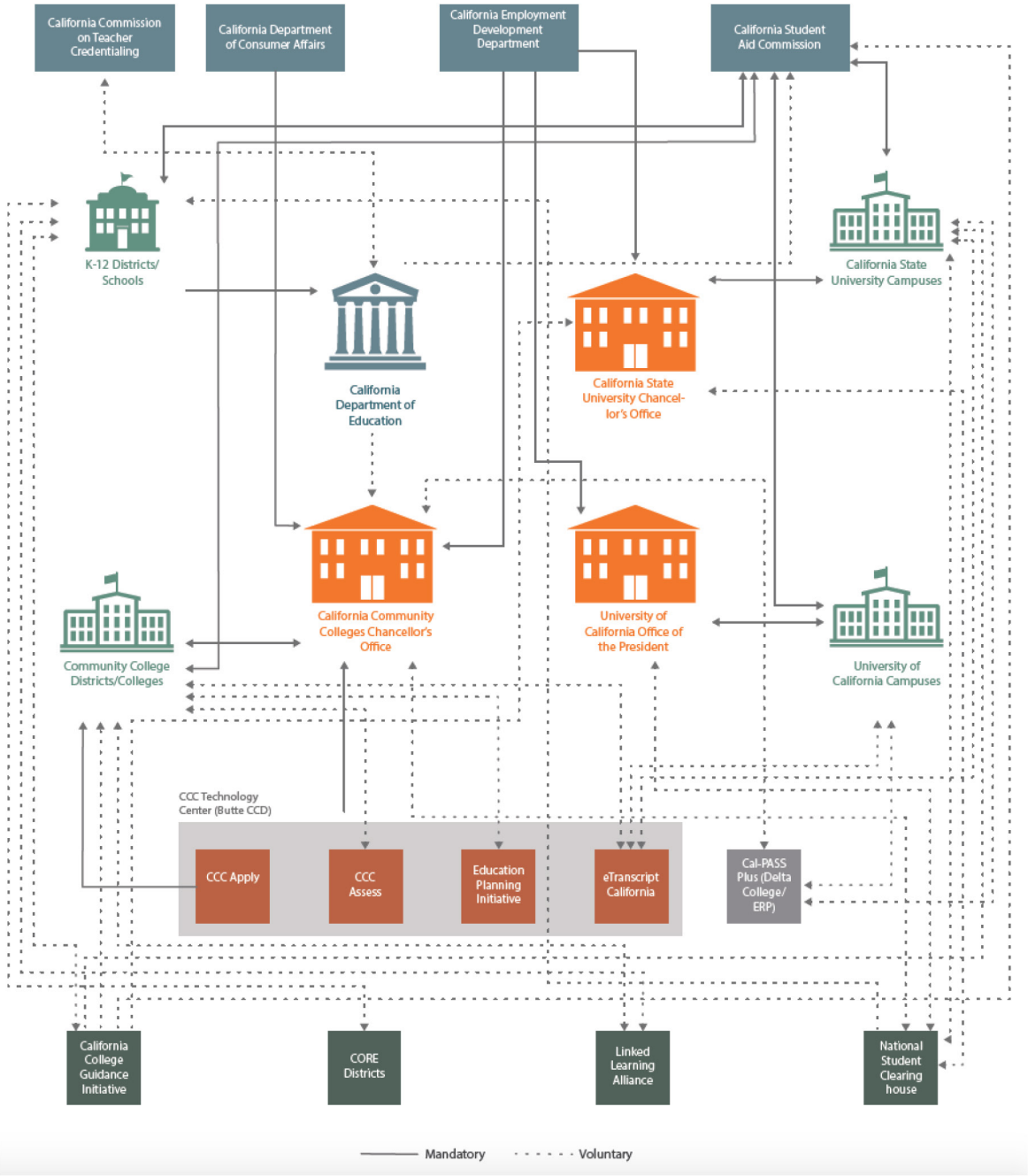
Building a Student-Centered Data System in California

With over 37 million residents, California is the most populous state in the country. California's primary and secondary schools enroll over 6.2 million students,[1] and there are 3.4 million undergraduates attending 683 postsecondary institutions in California. Yet, because of the lack of a strong data infrastructure, we are unable to answer basic questions about student progress and completion.

To improve educational outcomes and meet future workforce demand, California must establish a longitudinal data system that can answer the following critical questions about student pathways and outcomes:

- Where are California's high school graduates applying to and enrolling in college?
- What are the workforce outcomes of students who do not graduate from high school, or who do graduate from high school, but do not attend college? Are they employed? Do they later go to college?
- What is happening to students who are eligible to apply to the University of California (UC) or California State University (CSU) but are not accepted due to capacity constraints?
- How many students are enrolled simultaneously in high school and college? Or enrolled simultaneously in a community college and a four-year college or university?
- Who is applying for and receiving financial aid? How are they using it?
- Are students successfully transferring from community colleges to UCs, CSUs, or other four-year institutions? Are they successful?
- How long are students taking to complete their degrees? Are they successfully entering the workforce and earning living wages?
- Do the answers to these questions vary by students' race/ethnicity, income, region, gender, military status, parents' education, and age (i.e., what is the effect of entering college as an adult on graduation and post-college outcomes)?
- **Without a centralized postsecondary data system, California is seriously lagging behind the rest of the nation in this respect.** Even between the California Community Colleges, CSU system, and UC system, data sharing is cumbersome and time-consuming, so the simplest questions are difficult to answer. Add in the siloed early childhood, K-12, financial aid, and workforce systems, and California's data is, as the Education Insights Center calls it, a "maze of disconnected systems" (see Figure 1).

Figure 1. California's Data Framework – A "Maze of Disconnected Systems"[2]



BUILDING A LONGITUDINAL DATA SYSTEM

State longitudinal data systems (SLDS) combine data about individuals from different state agencies and programs so stakeholders can answer critical questions about K-12, postsecondary, and workforce outcomes. Complete SLDS have data on individuals from their early education through primary and secondary school, into college (if applicable), and the workforce. States have established their SLDS for several reasons, including to track student progress and success, to provide information for institutional resource allocation (for example, to track the necessary data for funding formulas), to help with federally mandated data reporting, and to track state progress against educational attainment or economic improvement goals.

Kentucky Center for Statistics

The Kentucky Center for Statistics (KYStats) was created by statute in 2012 when state leaders identified a need to consolidate disparate data sources and centralize information to improve educational attainment and economic outcomes in the Commonwealth. KYStats maintains the Kentucky Longitudinal Data System (KLDS). KLDS integrates data from the Kentucky Department of Education, the Council on Postsecondary Education, the Education Professional Standards Board, the Higher Education Assistance Authority, and the Kentucky Education and Workforce Development Cabinet. KYStats also develops reports and provides data to policymakers, agencies and the general public. Through this comprehensive statewide data system, Kentucky has been able to better understand how high school experience affect college going and success, improve teacher education, and plan to one day report on the economic outcomes of Kentucky's college graduates.

The data systems that have the most robust use, stability, and longevity have commonalities that are considered best practices. These include:

- Involvement or establishment of a cross-sector body or council (including representation from participating agencies) to act as a data governance body, coordinate agencies, facilitate data standardization, and manage collection, storage, and use;
- Formalized structure and ensured sustainability through codification in statute;
- Transparent policies to ensure access, data security, and individual privacy;
- Agreements to share data across state, through agreements with other states' longitudinal data systems, the National Student Clearinghouse, or the Wage Record Interchange System (WRIS).

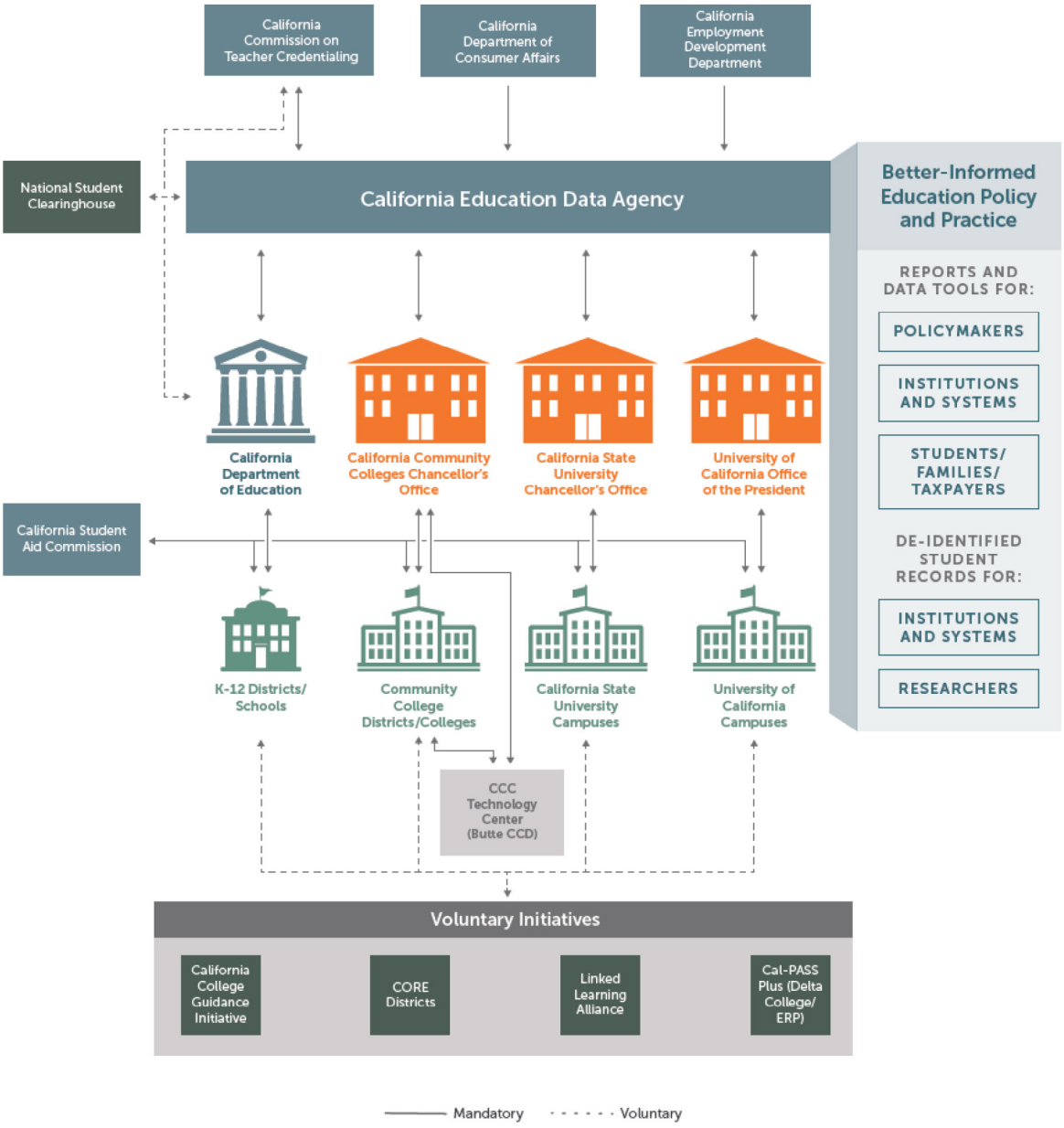
Based on the best practices as seen through the development of data systems across the country, **we recommend building a centralized data system in California.** This is preferable to a federated system with a maze of linkages, which is similar to what currently exists and has proven ineffective at getting cross-sector information into the hands of policymakers and the public to make better-informed decisions. Instead, this administration should take steps to build a centralized state longitudinal data system.

THE PATH FORWARD

The first step towards realizing a SLDS is to establish a data entity to manage and govern the data system that includes representation from the entities currently collecting separate data and formalize that body in statute. Kentucky is an exemplar in state longitudinal data systems. The legislative language that established what is today the KYSTATS formalizes the body, sets the purpose and mission of the body, describes overarching governance (in Kentucky, there is an executive director who oversees KYSTATS), identifies the participating agencies, and identifies the participating agencies and funding streams.

In California, at a minimum, legislation should establish a cross-agency body consisting of: the California Department of Education (which collects data from K-12 districts and schools), each of the public California postsecondary systems (or a newly established coordinating body), the California Student Aid Commission, the Bureau for Private and Postsecondary Education, the California Commission on Teacher Credentialing, the California Department of Consumer Affairs, and the California Employment Development Department. Figure 2 illustrates how the Education Insights Center envisions such an agency improving the efficiency of data exchanges and the availability of information about student progress and outcomes.

Figure 2. Proposed structure for California's longitudinal data system [3]



This body should be given the authority to establish data governance policies (including ensuring privacy and security of data), determine contents of the data system (and oversee collection of those items), report regularly on specific metrics, and provide access to researchers.

Data Governance Policies -The specific policies governing the data system should be established by the cross-sector body and would include guidance on how data elements are to be collected and stored, how the privacy of those elements would be ensured, and the security procedures to protect the data.

- Clear memoranda of understanding (MOUs) should be established between the agencies contributing data to the centralized warehouse that include guidelines for what data are to be contributed, by when, and in what format.
- The cross-sector entity should determine common data definitions and technical standards. As a starting point, the cross-sector entity should compile the existing data dictionaries to find commonalities and should rely upon national standards such as the Common Education Data Standards (CEDS) to establish data definitions.
- Privacy and security of data should be established by adhering to common privacy and security practices, such as stripping personally identifiable information after data have been matched, providing aggregate reports to avoid revelation of personally identifiable information, setting standards for data storage and encryption, providing access to data to approved researchers and through rigorous application processes, and setting policies and practices for addressing a data breach.

Data Contained in a Longitudinal Data System - A fully-realized centralized longitudinal data system would include data from early childhood education, K-12, postsecondary education, and the workforce. The cross-sector entity should be given the flexibility to determine the exact data elements to be included. National best practices, however, suggest that at a minimum, the following should be included:

- **Key demographics** - age/birthdate, gender, income, race/ethnicity, military status
- **K-12 indicators** - high school attended, graduation, grade point average, college entrance test scores
- **Postsecondary indicators** - institution(s) attended (including enrollment dates), enrollment status, attendance intensity, credential-seeking status, program of study, credit accumulation, credit completion, remedial placement, retention, applicable transfer credits, transfer date, graduation rate, net price, financial aid received, cumulative debt
- **Workforce indicators** - employment status, wages, loan repayment status

Inclusion of these data elements can help ensure that student-focused policies are in place, implemented faithfully, and effective. For example, Assembly Bill 705 (Irwin, 2017) mandates the use of high school grade point average in determining whether students should be placed in remedial education courses. A data system that includes both high school grade point average and college course information would easily be able to provide evidence that (1) high school grade point average is indeed being used; (2) whether students are able to progress to college level courses more quickly; and (3) if this legislation leads to better outcomes for students.

Not all centralized data systems across the states collect all the information above. In states with less robust data systems, crucial information such as income, attendance, and completion are left out of the data system. More commonly, information about students at private postsecondary institutions, licensure rates, cost and repayment information, data on participation in state and federal assistance programs, and workforce data not included in unemployment insurance wage records are limited or

omitted completely in many states. Challenges to filling these data gaps can be legal, limiting access to or usage of data; technical, lacking common identifiers or technological infrastructure; and logistical, including limited funding and siloed operations across state agencies. California should strive to build as complete a data system as possible.

Georgia's Academic and Workforce Analysis and Research Data System (GA-AWARDS)

GA-AWARDS is Georgia's centralized longitudinal data system that includes information from pre-K through workforce. It is managed by the Governor's Office of Student Achievement, and includes information from all institutions in the state. Georgia uses a matching algorithm and then provides deidentified data back to researchers from participating agencies. In addition, there are a number of data elements and tables made available for download on the website and data requests can be submitted by academic researchers.

Reporting and Access for Researchers - There are several examples of states that collect and maintain data systems, but do not make their data transparently available or accessible to researchers. The value in maintaining a centralized longitudinal data system is for the information such a system can provide. The cross-sector entity should report regularly (ideally through a regularly-updated dashboard) on a set of metrics that measure progress towards statewide goals including primary and secondary school progression and success, college access, college success, and workforce outcomes. These indicators should be disaggregated by race/ethnicity, income, and education level, at a minimum.

In Washington, the Education Research and Data Center (ERDC) within the Office of Financial Management (the Governor's budgeting and forecasting agency) manages the states' centralized longitudinal data system. The system has data from the three postsecondary data systems in the state, the National Student Clearinghouse, K-12, and the workforce. The budget funds six employees to manage the data system. ERDC regularly produces reports and, under the privacy policies that have been established, grants access to researchers.

COST OF A CENTRALIZED LONGITUDINAL DATA SYSTEM

In many states, the development and expansion of centralized longitudinal data system was made possible in part to the grant program administered through the Institute of Education Statistics since 2005, which encourages collaboration and linking of state data between K-12, postsecondary and the workforce. Forty-seven states, the District of Columbia, Puerto Rico, the Virgin Islands, and American Samoa have all received grants. California was awarded two of these grants, one in 2006 and another in 2009.

Because of differences in approach, structure, and administration a centralized longitudinal data system, the cost of building these systems can vary dramatically depending on the state context. According to a recent brief, some states constructed systems for $2.5 million and others have spent upwards of $7 million. Contributing factors to cost include: architecture (centralized versus federated systems), hardware, software, participating agencies, amount and quality of data, and data governance and policy procedures. Costs will also vary based on analytical capacity, pre-existing infrastructure and capacity, in-house or external development of the system, and ease of negotiating data sharing agreements. Once a state builds the system, it also must factor in maintenance and improvement costs, which will vary depending on analytic needs, continuous improvement, and the amount of data stored and shared.

Building a centralized longitudinal data system as described above would require two separate types of investment. The first is a start-up investment to establish the cross-sector entity, align the current

data systems, and develop (or adapt) the technology to administer a data system. The second would be ongoing annual costs.

- **Start-up costs** - A recent evaluation of the state longitudinal data system grants revealed that most grants awarded were under $10 million. However, it is important to take into consideration that California would not be starting from scratch, rather working to streamline existing data structures, definitions, storage, and policies. Thus, it is reasonable to assume that the "start-up" costs in California would be significantly lower. Documents from the planning phases for what is now KYSTATS suggest that an investment of less than $3 million could be sufficient to establish a cross-sector data entity and establish a data system.

- **Ongoing annual costs** - The operating budget of the California Postsecondary Education Commission (CPEC) when it was defunded was $2.3 million. It is reasonable to assume that the annual costs of a cross-sector data agency would be commensurate with a state agency such as CPEC. Moreover, with a centralized data system, the state could save in terms of shared services. For example, federal reporting could be streamlined, reducing time and burden for each of the individual agencies currently reporting. Furthermore, while there are currently individual agreements between vendors such as the National Student Clearinghouse, a centralized data system would offer savings by necessitating only one contract.

CHALLENGES ASSOCIATED WITH A CENTRALIZED LONGITUDINAL DATA SYSTEM

Although centralized longitudinal data systems offer many benefits to state policymakers, education leaders, students, taxpayers, and other stakeholders, there are critics of challenges associated with establishing and sustaining these systems include:

Privacy - One of the main arguments against longitudinal data systems is that they jeopardize student privacy. However, with clear data security, storage, access, and use policies, privacy is protected and data can be used within the requirements of the Family and Education Rights Privacy Act (FERPA). Many states have clear privacy policies and practices that allow data to be collected and used for the betterment of education, the economy, and the state as a whole. Organizations like the Data Quality Campaign can provide technical assistance on the best way to protect student privacy.

Lack of political and public will - Especially with the experience around the inBloom data effort, states have had to create general buy-in for creating a state longitudinal data system. This depends on strong leadership, the creation of a cross-sector entity, and a communications effort that can show the value of such a data system.

Lack of infrastructure - There are several data systems in California that currently exist, so the technology capacity is present in the state. New and ongoing funding infrastructure to combine the data systems would need to be built and there would need to be staff to manage the system.

Lack of funding - The overall cost estimates for a centralized longitudinal data system are small relative to the budget, and the benefits of developing a centralized longitudinal data system far outweigh the costs. In fact, **a centralized system can present significant savings in terms of streamlining current data collection and reporting efforts, ensuring efficiency in higher education delivery, and increasing student success, which will ultimately result in statewide economic payoff.**

Montana is one example of a state that matches its existing education and employment data to answer critical policy questions about outcomes. Colleges, students, businesses, and policymakers leverage data through the longitudinal data system for strategic planning, collaboration across agencies and sectors, and policymaking. This effort originated from the states' need to align education and workforce to combat the tight labor market and retiring labor force. With the support of the governor, the Montana Department of Labor and Industry (MTDLI) and the Office of the Commissioner of Higher Education (OCHE) matched data from public postsecondary institutions with age records from the state's unemployment insurance system. They also worked with private colleges in the state to include student-level data from those institutions in the data match. By connecting this information with market demands, policymakers could see whether the colleges produced graduates in in-demand fields, highlight where programs and occupations are oversupplied, and help identify geographical disparities between supply and demand. These analyses led to collaboration across colleges and universities with employers and provided students with better information on potential workforce outcomes.

BACKGROUND ON CALIFORNIA LONGITUDINAL DATA POLICY

The 1960 Master Plan for Higher Education called for the creation of a coordinating and planning entity for postsecondary education, which eventually became the California Postsecondary Education Commission (CPEC). As early as 1971, a joint legislative committee identified a lack of comprehensive information on key educational measures and inability to compare meaningful data to inform policy, budget, and planning decisions. **As technological capacity also evolved, it became clear that the state could do more to make data more transparent, accessible, and actionable.** In 1999, then Assembly Speaker Antonio Villaraigosa introduced AB 1570, which was enacted to task CPEC with additional and more specific responsibilities in the collection of longitudinal data (i.e. implementing use of a unique student identifier, making data accessible online), supported by an associated appropriation of $420,000.

However, shortly after CPEC's mandate was expanded, its capacity to achieve this charge was undermined. Between 2001 and 2009, repeated budgetary reductions to CPEC cut agency funding by over 60 percent.[4] After 2002, more than a dozen bills sought to eliminate or restructure CPEC in various fashions, ranging from modifying commission membership to amending agency responsibilities. Simultaneously, policymakers were considering the most effective means of integrating data systems. In 2008, Governor Schwarzenegger signed SB 1298 (Simitian) and charged the state Chief Information Officer and a working group to determine the best governance structure for a longitudinal P-20 data system and a strategic plan for its technical implementation. The recommendations that came from this working group were reflected in a 2011 bill, SB 885 (Simitian), which would have authorized CPEC, the California Department of Education (CDE), State Board of Education, Commission on Teacher Credentialing, California School Information Services, public higher education segments, and the Employment Development Department to develop a joint powers agreement (JPA) to implement a longitudinal P-20 data system. Unfortunately, 2011 was a difficult year in which to advance longitudinal data. In the first year of his term, Governor Brown had committed to identifying means to overcome the significant budgetary shortfalls faced by the state. Governor Brown not only vetoed SB 885, citing the agencies' ability to convene a JPA on their own authority, but also expressed doubt about whether they should do so due to "fiscal constraints." That same year, against the recommendation of the Legislative Analyst's Office, Governor Brown also vetoed funding for CPEC along with numerous other state agencies, boards and commissions.

Following the elimination of CPEC in 2011, legislators sought both short- and long-term solutions to preserving access to longitudinal data, but were unsuccessful in obtaining either. Senator Carol Liu introduced SB 1138 in 2012, which would have transferred the responsibilities of CPEC as a central data management system for higher education to the CDE on an interim basis, but was held in the

Senate Appropriations Committee. That same year, Assemblymember John Perez also introduced AB 2190, which would have established the California Higher Education Authority and transferred the prior responsibilities of CPEC to this proposed agency, but the bill was held on Suspense in Assembly Appropriations. Additional bills, AB 1348 (Perez, 2014), SB 42 (Liu, 2015), AB 1837 (Low, 2016) AB 217 (Low, 2017), and AB 1936 (Low, 2018) have all sought to establish a new higher education coordinating entity with responsibility for managing a longitudinal data system. SB 1224 (Glazer, 2018) took another approach by focusing exclusively on the integration of data platforms by requiring CDE and public higher education segments to develop a collection system to track student outcomes from K-12 through postsecondary education and into the workforce. Each of the above bills introduced since 2014 were either held on Suspense by an Appropriations Committee or vetoed by Governor Brown.

Stakeholders from K-12, higher education, early childhood education, and the workforce agree that California needs a strong data system to ensure that our students' needs are being met, to hold educational institutions at all levels accountable for serving our students well, and to ensure that our economy stays strong.

THE TIME IS NOW IN CALIFORNIA

California is the fifth-largest economy in the world, the most populous state, and home to the largest educational system in the nation, with technology industry clusters with global reach. However, **California lags behind the rest of the nation in terms of fully understanding the opportunities, roadblocks, and outcomes of its students and its citizens.** The growing equity gaps in educational success and income disparities are evidence that California is not able to adequately address the educational needs of its citizens. Without a robust, centralized longitudinal data system, the information that is necessary to do so is not available. **We need strong leadership to build the political will and establish a data system that can inform our policymakers, educational institutions, and provide critical information to students and the public.** Instituting truly data-informed decision-making in California will not only better guide state efforts advancing educational equity, but also provide a positive, national model for how 21st century technologies can be deployed effectively to further the impact of public policy in addressing generational challenges like the persistent opportunity gaps in education.

ENDNOTES

[1] Janice, A. & Voight, M. (2016, May). Toward convergence: A technical guide for the postsecondary metrics framework. Retrieved from http://www.ihep.org/sites/default/files/uploads/postsecdata/docs/resources/ihep_toward_convergence_low_2b.pdf

[2] Institution for Higher Education Policy. (2016, July 28). Mapping postsecondary and workforce information gaps in state data systems [Infographic]. Retrieved from http://www.ihep.org/sites/default/files/uploads/postsecdata/docs/resources/infographic_pdf_final.pdf

[3] Education Insights Center. (2018, December). A centralized data warehouse increases efficiency and availability of information [Infographic]. Retrieved from http://edinsightscenter.org/Publications/Research-Reports-and-Briefs/ctl/ArticleView/mid/421/articleId/2198/California-Education-Policy-Student-Data-and-the-Quest-to-Improve-Student-Progress

[4] Taylor, M. (2010, January). The Master Plan at 50: Greater than the sum of its parts—coordinating higher education in California. Retrieved from: https://lao.ca.gov/reports/2010/edu/ed_coordination/ed_coordination_012810.pdf

The Campaign for College Opportunity

ONE MISSION: Making College Dreams a Reality

15 Years

WWW.COLLEGECAMPAIGN.ORG

LOS ANGELES OFFICE
1149 S. Hill Street, Ste. 925
Los Angeles, CA 90015
Tel: (213) 744-9434
Fax: (877) 207-3560

SACRAMENTO OFFICE
1512 14th Street
Sacramento, CA 95814
Tel: (916) 443-1681
Fax: (916) 443-1682

WASHINGTON DC OFFICE
1825 K Street
Suite 720
Washington DC, 20006

@CollegeCampaign

@CollegeOpp

BOARD OF DIRECTORS