BCC: Developing a Comprehensive Regional Approach to Data Set Integration to Support Data-intensive Research in Education in Silicon Valley

Proposal Submitted to the National Science Foundation

By

University of California, Santa Cruz

Principal Investigator: Rodney T. Ogawa, Ph.D.

Co-Principal Investigators: Douglas Bonett, Ph.D.

Scott Brandt, Ph.D.

Ronald Glass, Ph.D.

Carlos Malzahn, Ph.D.

Project Summary

Developing A Comprehensive Regional Approach to Data Set Integration to Support Data-Intensive Research in Education in Silicon Valley

Overview

This project will lay the social, intellectual, and technological groundwork to develop a comprehensive, regional data set to support data-intensive research in education in Silicon Valley, California. Despite unprecedented access to data and the use of sophisticated methods of data analysis, research has yet to offer detailed and comprehensive answers to the most fundamental educational questions which reflect the most pressing educational challenges confronting the nation: a) What are the multiple causal vectors that explain the academic achievement of youth in the U.S.? b) What are the multiple causal vectors that contribute to the achievement gap between low-income, racially, culturally, and linguistically diverse students and their more academically successful White and Asian peers?

To answer these and other fundamental educational questions will require existing and newly emerging data sets with a breadth of data—ranging from fine-grain measures of instruction to larger-grain measures of classroom and school conditions—and a depth of data to include data on non-school factors that are collected by agencies in institutional sectors other than education, including health, social services, employment, and housing.

To plan and design an approach to constructing such an education data set, a cross-disciplinary group of researchers in education, the social sciences, and engineering at the University of California, Santa Cruz will partner with consultants from the National Lab for Educational Transformation, which is a private non-profit organization with expertise in developing and analyzing large-scale education data sets and data security and with ties to public agencies and the private sector in the region. This cross-disciplinary and cross-sector team will examine and develop solutions to 4 sets of challenges to designing, constructing, and analyzing such a data set: a) cultural, legal and regulatory issues that prevent sharing data across institutional sectors; b) analytic approaches to analyze such a data set; c) technical infrastructures to integrate and manage the data set; and d) ethical issues, including concerns for data security and privacy. The team will engage regional stakeholders from public agencies and the private sector in 5 convenings to discuss and receive feedback on the parameters for designing and developing the comprehensive, regional data set in education.

Intellectual Merit

The **intellectual merit** of this project lies in its potential to advance knowledge through empirical research in education and across other public domains. It holds the promise for opening opportunities for research in education that can yield new understandings of student achievement, the achievement gap, and other critical educational issues. Such research will apply advanced statistical methods from the social sciences and from other disciplines, where advanced analytics such as agent based modeling are being applied. Additionally, just as the data set will provide new prospects for education researchers, it will similarly create opportunities for researchers in other fields to examine the impact of factors in other domains on important outcomes in their focal fields.

Broader Impacts

This project has the potential to have several **broader impacts**. Just as the development of a comprehensive, regional data base can open opportunities for research, it can also facilitate regional planning, advance cross-institutional coordination and cooperation in the region and inform efforts to improve regional services in education and other public service domains, and contribute to advancements in the technology for data mining in education.

BCC: A REGIONAL "DATADROME": DEVELOPING A COMPREHENSIVE REGIONAL APPROACH TO DATA SET INTEGRATION TO SUPPORT DATA-INTENSIVE RESEARCH IN EDUCATION IN SILICON VALLEY

A. Need

Despite unprecedented access to data and the use of sophisticated methods of data analysis, research has not offered detailed and comprehensive answers to the most fundamental educational questions which reflect the most pressing educational challenges confronting the nation:

- What are the multiple causal vectors that explain the academic achievement of youth in the U.S.?
- What are the multiple causal vectors that contribute to the achievement gap between low-income, racially, culturally, and linguistically diverse students and their more academically successful White and Asian peers?

This project, which focuses on planning and design, addresses two key barriers that stand between educational researchers and the ability to answer these central and crucial questions by: a) designing a comprehensive, regional data set to support data-intensive research in education and b) organizing a cross-disciplinary community of researchers and regional stakeholders,

Presently, researchers attempt to answer these questions by combining the results of multiple studies that employ a variety of analytic approaches—ranging from ethnographies to randomized trials and large-scale surveys. This approach is used to examine a broad array of topics ranging from teachers' instructional practices to student characteristics and teacher quality (e.g., Farkas, 2009).

Researchers must have access to data sets with greater **depth** and **breadth** in order to mount more comprehensive studies to analyze the causal vectors of student achievement and the achievement gap. In addition, they must be able to employ the full panoply of analytic methods to combine and analyze disparate data sets.

Large-scale data sets in education tend to lack **depth.** They generally include only relatively large-grain data from surveys and existing records about classroom and school conditions but do not include the fine-grain measures of instructional activity needed to understand teaching and learning. This lack of depth is in part the result of the tendency for educational studies to employ either unstructured data or structured data, but rarely both. While unstructured data can capture details of instructional interaction and their meaning to teachers and students, structured data are needed to provide measures of specific dimensions of instruction and conditions of classrooms and schools.

Education data sets also lack **breadth** by not including data on potential causal vectors outside of schools. This is a serious limitation because research reveals that much of the variance in student achievement is attributable to non-school factors (e.g., Coleman et al., 1966). Specifically, research documents the negative impact of poverty on academic achievement (Jones & Schneider). Consequently, in order to illuminate the mechanisms by which non-school factors, generally, and poverty, specifically, affect academic achievement and the achievement gap, research will require data on a variety of family and community conditions as potential causal vectors. Such data are collected by agencies operating in sectors other than education, including health, social services, employment, housing, and the like. In addition, private firms collect data on non-school educational services that students receive as well as data in non-educational social and economic domains.

These limitations of education data sets are reinforced by the tendency for education researchers not to collaborate with scholars in other academic fields. While it is widely recognized that researchers must work across disciplines to answer complex research questions—such as the causal vectors of academic achievement and the achievement gap—educational researchers generally confine themselves to their disciplinary roots and tend not to work with colleagues outside the field of education (Gamoran, 2009). This isolation contributes to limitations that characterize education data sets in at least 3 ways: a) data from any given study tend to reflect the unit of analysis and variables dictated by its disciplinary perspective,

limiting the depth and breadth of measures included in the data set; b) data tend to focus on in-school factors, ignoring factors in other domains (e.g., health, nutrition, and social and cultural capital); and c) approaches to data management and analysis rarely engage advances in non-social science disciplines such as engineering, where rapid developments are occurring in the management and mining of "big data." This project will lay the social, intellectual, and technological groundwork to develop a comprehensive, regional data set to support data-intensive research in education. A cross-disciplinary group of researchers at UC Santa Cruz will partner with a private non-profit organization involved in education transformation to work with public and private sector stakeholders in Silicon Valley to identify potential data sources and to develop strategies to overcome challenges to the development and analysis of data across many institutional domains that constitute the social, economic, and cultural ecology of education.

Many barriers must be overcome in order to design, build, and utilize a comprehensive education data set, with the depth and breadth required to support data-intensive research.

- Cultural, legal and regulatory issues prevent sharing data across different institutional sectors.
- Ethical issues, including the prominent concern for data privacy, can arise from constructing and utilizing a comprehensive data set.
- Cutting-edge technical infrastructures will be required to integrate and manage the large quantities of data from multiple institutional sectors.
- Advanced analytic approaches, including mathematical modeling techniques developed in fields outside education research, must be applied.

To confront these barriers, collaboration across academic disciplines and multiple stakeholders must occur. By bringing together cross-disciplinary researchers, school and community leaders, and corporate technology entrepreneurs, this project can create a unified approach to regional data and youth outcomes that combine social, technical and education research components, and that is scalable.

The feasibility of designing, constructing, and analyzing such a comprehensive education data set must be tested on a regional scale. Existing large-scale data sets employ representative, national samples (e.g., NELS). Ideally, a comprehensive data set with adequate breadth and depth of measures would employ such a national sample, but first it must be designed and tested at a more manageable scale.

B. Purpose and Objectives

The purpose of this year-long project is to design a model, regional data base in Silicon Valley, California that will support data-intensive research in education by integrating structured and unstructured data that capture both the depth of analytic levels and breadth of measures to integrate in-school and out-of-school factors affecting academic achievement.

The project's objectives are the following:

1. To establish a cross-disciplinary community of scholars, representatives of institutional agencies, and leaders of private companies to collaborate in developing a comprehensive regional data set that can be analyzed to understand the causal vectors of student achievement and the achievement gap.
2. To design a regional-level pilot project to inform the development of a comprehensive national data set for data-intensive research in education.
3. To identify existing structured databases.
4. To identify existing unstructured databases.
5. To determine the computational and analytic issues that must be resolved to make it possible to answer the core questions
6. To assess the technical, organizational and regulatory barriers to the aggregation and integration of those data bases into a single architecture. To establish the search, user and system analytics required to make such an architecture functional.
7. To examine the ethical, security and policy issues that inhibit the development of an aggregated database needed to answer the core questions.

C. Relation to Present State of Knowledge

The proposed project is informed by the present state of knowledge that reveals the limitations of research aimed at analyzing the causal vectors of academic achievement and the achievement gap in the U.S. In addition, we describe work undertaken by members of the project team that shapes the approach we will take to design a comprehensive, regional database to support data-intensive research in education. This work highlights the importance and challenges of integrating data from education with data from other institutional sectors.

Our understanding of the causal vectors of academic achievement and the achievement gap is compromised in part because we lack studies that provide detailed analyses of existing as well as newly emerging data sets that provide both a depth of data—which includes measures of instruction and individual learning and measures of classroom and school factors—and breadth of data, which includes measures of non-school factors—including data on factors from institutional sectors outside of education.

Large-scale data sets typically lack depth in that they do not include fine-grain measures of instruction and students' responses to instruction. This is reflected in the research literature which highlights factors that can be measured through surveys or are readily available in public records, such as teacher quality (Hanushek, Kain & Rivkin, 2004; Lankford, Loeb & Wyckoff, 2002), demographic composition of student bodies (Hanushek et al., 2004; Zimmer & Toma, 2000), student academic placement (Oakes, Gamoran & Page, 1992), and teacher expectations (Farkas, 2003). While these school factors have been documented to affect the academic achievement of students, they do not illuminate factors that arguably are most proximal to student learning in schools: teaching and students' responses to teaching.

Another limitation of education data sets is that they lack **breadth** by not including data on potential causal vectors outside of schools. For the past half-century research has documented that most of the variance in student achievement is attributable to factors outside of schools (Coleman et al., 1966; Rumberger & Palarty, 2004). Specifically, research reveals the negative impact of neighborhood poverty on academic achievement (Massey & Denton, 1993; Wilson, 1987). However, care must be given to separating the impact of families from the impact neighborhoods, or communities (Duncan & Raudenbush, 1999; Jencks & Mayer, 1990). Duncan and Murnane (2011) argue that income inequality is undermining the central goal of public education—to provide children with an equal chance at academic and economic success. They posit that rising social and educational inequality operates not through race, but through poverty-related factors such as disadvantaged neighborhoods, insecure labor markets, and economic segregation. Similarly, Burdick-Will et al. (2010) encourage researchers to broaden the scope of their studies to examine the influence of neighborhoods, including such factors as housing and violence.

This project builds on a prototype with roots in Silicon Valley, which highlights two critical issues: a) the foundational requirements for scaling a similar effort across the region and b) extending it to include newly defined variables such as unstructured data and out-of-school factors affecting academic achievement of students. San Jose Unified School District (SJUSD), the largest school district in the region with 33,000 students, developed, manages, and utilizes one of the most comprehensive data warehouses in the country. Two individuals responsible for this warehouse are core members of the proposed project team. The SJUSD warehouse architecture integrates a variety of data from many non-academic sources over time in addition to both standard and expanded measures of academic performance. The data warehouse houses 16 years of data, with more than 50 million records, 113 objects, and over 3,500 attributes.

At the student level, the warehouse comprises student attributes including–but not limited to: state-level test scores in Language Arts and Math (grades 2-11), Science (grades 5, 8, 9-11) and History Social Studies (grades 8-11) including End-of-Course scores for eleven specific high school courses, District-level interim assessments in Language Arts and Math (grades 2-11), district performance-based assessments in writing and math (grades K-11), course rosters (K-12), course-taking patterns K-12, course grades (grades 6-12), GPA (grades 6-12), high school credits, SAT/ACT test scores, Advanced Placement/International Baccalaureate scores, Physical Education test scores, college enrollment and completion, behavior records, student attendance, student perception data, English Learner data, special education status and

chronic health issues. Student demographic data include gender, ethnicity, socioeconomic status, language proficiency, home language, program participation for gifted and talented education, immigrant education and migrant education, zip codes, dates in which students entered the District and/or the United States, birth cities, and reclassification dates for English Learners.

Teacher variables include gender, ethnicity, courses taught and course rosters by education, credentials, endorsements and years of experience.

SJUSD established mutually beneficial partnerships and data-sharing agreements with county agencies and non-profit student support networks, including the following:

- Santa Clara County Mental Health Services spearheaded a School Engagement Initiative Project (SEIP) that helped the District focus on non-cognitive factors such as behavior and attendance to develop indicators for off-track students.
- Juvenile Justice probation officers work closely with the District's Student Services team to share data regarding suspended, truant or expelled students.
- With the largest health database in the County, the District partnered with the Lucille Packard Children's hospital to place 4 full-time nurses in high poverty schools. The results of this highly successful effort were published in the Journal of American Medicine (JAMA, 2012).
- Internally, key data reports and data protocols were developed through extensive teacher and administrator input and implemented district-wide. The impact of these teacher cycles of inquiry were seen in dramatic student achievement increases, closing the achievement gap between Hispanic and White subgroups by 36%; and 8 inner city schools exited Program Improvement.
- Ten years of perception data captured from asset-based Climate Surveys that were administered to over 300,000 Grade 3-12 students, parents and teachers in Spanish and English framed broader data-driven Community Conversations involving over 5000 participants.

This rich variety of variables allowed for in-depth analysis of critical factors potentially impacting the whole student. Efforts to expand the research and information sharing into the community revealed numerous barriers, including the difficulties of melding external databases. The cross-agency dialogues required to permit data sharing under research exemptions specified in the California Welfare and Institutions Code, the California Education Code and HIPAA regulations become critical gatekeepers in efforts to integrate data from different institutional sectors.

This project will address those barriers to data sharing and define the foundational requirements for scaling a similar effort across the region. The effort would extend research to include unstructured data and out-of-school factors affecting academic achievement of students.

In addition, the National Laboratory for Education Transformation (NLET), a partner in one of the research centers collaborating on this project (see Plan of Work) organized a project to apply advanced modeling techniques to the San Jose Unified data warehouse to develop a model of 8th grade achievement in mathematics. Through a National Science Foundation DRK12 grant, this multi-institutional and organizational team has been involved with a collaborative team of computer scientists, statisticians, educational researchers, and data visualization experts from the Los Alamos National Laboratory; the University of Texas, Austin Learning Technology Center; and the University of California Santa Cruz in pioneering the application of agent-based models to the analysis of "big" educational datasets extracted from district-level and state-level data warehouses that will make these models visually accessible to educators and policy makers. This group has succeeded in operating a diverse stakeholder data project.

D. Expected Significance

Data-intensive research in education can contribute to answering fundamental research questions needed to solve or attempt to solve the central enduring problems confronting education in the U.S. However, such research must be supported by data sets that have both the depth, reaching into instruction and learning, and breadth, combining data from multiple institutional domains, and thus provide opportunities to conduct comprehensive examinations of causal vectors of individual student achievement and the

achievement gap in the U.S. This project will lay the social, intellectual, and technological groundwork to develop a comprehensive, regional data set that is characterized by both data depth and breadth. The expected significance of this project is four-fold.

First, this project will identify, examine, and posit solutions to challenges that compiling and analyzing such a data set will pose. The cross-disciplinary project team will work with public agencies and private companies to identify potential data sources and to negotiate issues such as proprietary interests, privacy and security, that can present barriers to sharing and integrating across multiple domains.

Second, the groundwork laid for this project can inform similar work by other researchers in other parts of the country. Indeed, this project's broad and ambitions intent is to provide a model for the compilation of a comprehensive, national data set that will support data-intensive research in education and across institutional sectors.

Third, the results of this project will result in work to develop a comprehensive, regional data set down to the level of individual students. This data set can be employed by researchers at the many universities and research and development laboratories in and around Silicon Valley and beyond. This body of research will contribute to understanding the causal vectors of academic achievement and the achievement gap, and related educational questions and issues.

Fourth, in this project, academic researchers will collaborate with representatives of public agencies that collect and store data as well as with representatives of private firms in the region that possess human and technological capital that are relevant to this project. A large divide has opened between education research and the fast-paced big data and smart systems approach of commerce, finance, and security systems. Education researchers potentially could benefit from advances in the private sector's management of "smart-data," which refers to the ability for diverse data to be compiled, exchanged, integrated, analyzed, and employed to produce integrated responses at any level of granularity. Smart-data also enables data exchange and management within ethical and secure regulatory and technical parameters.

E. Rationale for Regional Scope

Silicon Valley, encompassing San Mateo, Santa Clara and Santa Cruz counties, reflects the discontinuity between new systems used by consumers, commerce, and government and those used in the education and social services sectors. In fact, the Valley is the most significant contributor to the growth of the global information economy in the last decade because of its social-technical innovations. Yet, within the shadow of this economic marvel is a high level of social, economic and educational disenfranchisement and contrast. The high-tech and bio-tech firms in Silicon Valley explain that they cannot rely on growing their workforce locally, regionally or even in the U.S.

Silicon Valley's dense social networks and open labor market have long encouraged entrepreneurship and experimentation. In a network-based system such as the Valley, the porous organizational boundaries within and between companies, as well as between companies and local institutions such as universities and government agencies is a cultural strength (Saxenian, 1994). A cross-disciplinary research collaborative focused on the development of this comprehensive regional database could inform regional efforts to increase school achievement and workforce readiness and would be a natural extension of the Valley's culture.

There is regional incentive for this collaboration. The 2013 Silicon Valley Index reported greater disengagement of the more than 400,000 K-12 students in the region with increased dropout rates, lower graduation rates, increased juvenile felony drug offenses, and increased numbers of substance abuse rehabilitation clients. For the first time in 4 years, the percentage of Silicon Valley eighth graders scoring advanced on the Algebra I test fell and overall math proficiency levels also declined. Combined with a call from the Valley leadership to better coordinate efforts to support economic development across the region, this proposed collaborative can leverage the cross-disciplinary and research capacity of the Valley by engaging its scholars, scientists and practitioners from diverse disciplines to apply their skills to the development of new, large-scale, next-generation data resources for education.

F. General Plan of Work

The project's plan of work includes 4 components:

1. Description of the 3 University of California, Santa Cruz (UCSC) research centers and the 2 University of California multi-campus research initiatives that will contribute to the project
2. Descriptions of the 4 major areas of work:
    a. Regional Data on Causal Vectors of Academic Performance
    b. Advanced Analytics
    c. Technical Infrastructure, and
    d. Security and Ethical issues
3. Integration of the 4 major areas of work to design the comprehensive educational data base
4. Description of 5 convenings to engage researchers, education practitioners, non-education institutional stakeholders, and business and technology innovators to provide input on the project team's work on the 4 major work foci and their integration in a data base design.

1. **Collaborating Research Centers**

Three research centers of the University of California, Santa Cruz (UCSC) will collaborate on this project: the Silicon Valley Educational Research and Development Center (SV-ERDC), the Center for Statistical Analysis in the Social Sciences (CSASS), and the Institute for Scalable Scientific Data Management (ISSDM). In addition, two centers that are multi-campus research initiatives will also contribute to the project's leadership: the UC Center for Collaborative Research for an Equitable California (CCREC) and the Center for Information Technology Research in the Interests of Society–Data and Democracy Initiative (CITRIS-DDI. See the section on "Expertise" for the expertise that individual team members bring to the project.

SV-ERDC is a partnership of UCSC's Center for Educational Research in the Interest of Underserved Students (CERIUS) and the National Laboratory for Education Transformation (NLET), a private non-profit organization. SV-ERDC engages faculty across disciplines (education, social sciences, and engineering) to collaborate with educators and individuals and organizations in the private sector to conduct research and development to transform education. The Co-Directors of SV-ERDC are Rodney Ogawa, Professor of Education and Principal Investigator on this project and Gordon Freedman, President of NLET and a consultant on this project. NLET also includes Marcy Lauck and Bill Erlendson, who were the principal architects of the Data Warehouse in the San Jose Unified School District.

CSASS involves quantitative researchers from across the Division of Social Sciences to advance the application of statistical analyses in the study of human problems and conditions. The mission of CSASS is to stimulate and support collaborative research across social science disciplines and to inform public policy. Douglas Bonett, Co-PI on this project, is Professor of Psychology and Director of CSASS with expertise in psychometrics and the application of advanced statistical analyses in the social sciences.

ISSDM is a collaboration between the University of California Santa Cruz (UCSC) and Los Alamos National Lab (LANL). Bringing together researchers in systems, machine learning, visualization, processing, algorithms and other areas, the ISSDM promotes UCSC/LANL research collaborations in Storage Systems, Knowledge Management, Machine Learning, Human Computation, Data Visualization, Visualization and Analysis of Cosmology. Scott Brandt, Co-PI on this project, is Professor of Engineering and Director of ISSDM, and Carlos Malzahn, also Co-PI on this project, is Associate Adjunct Professor of Engineering, contribute their expertise in storage systems and knowledge management.

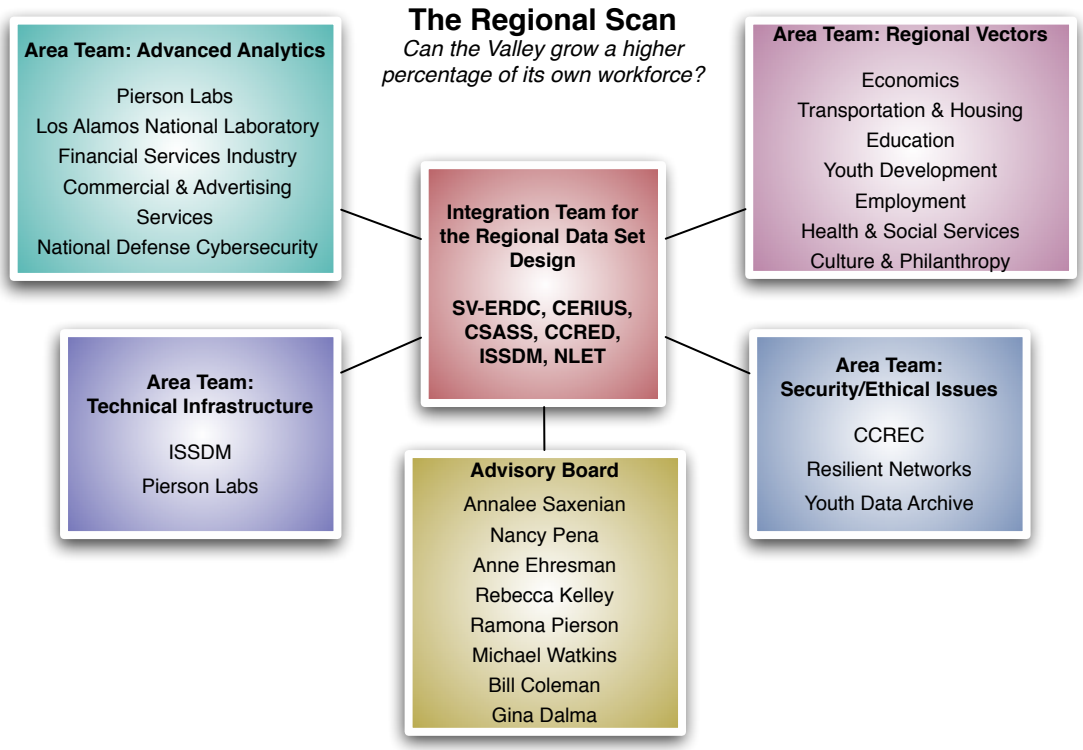
CCREC is a University of California system-wide multi-campus research program initiative that is supported by the U.C. Office of the President. CCREC supports multi/trans-disciplinary and collaborative regional research projects designed to address the state's intersecting crises in the economy, education, employment, environment, health, housing, and nutrition. CCREC's Faculty Fellows include scholars in the social sciences, public health, computer science, and the humanities. Ronald Glass, Co-PI on this project and Professor of Educational Philosophy is PI/Director of CCREC, oversees the Center's multi-disciplinary and cross-sector projects and is also Principal Investigator for a project supported by a grant

from the Spencer Foundation to examine the complex ethical issues embedded within collaborative community-based research, such as those pertaining to informed consent, confidentiality and anonymity.

CITRIS is a University of California multi-campus (Berkeley, Davis, Merced, and Santa Cruz) research enterprise that has garnered ~ $1B in grants in the past ten years to create information technology-related tools and systems to solve some of the most pressing social, environmental and health problems. CITRIS is at the cutting edge of incorporating computational capacities in the service of the public good. **DDI** develops tools to support dynamic relationships between digital media and democratic practices, such as the use of innovative mobile, Internet and social media applications to facilitate online deliberation, participatory decision-making, and rapid mobilization. Ron Glass is Acting Associate Director, Center for Information Technology Research in the Interests of Society–Data and Democracy Initiative

2. Four Major Areas of Work

The Project Team will engage in work on 4 major areas that present challenges to developing a comprehensive regional data set to support data-intensive research in education: Determining Sources of Regional Data on Causal Vectors, Advanced Analytics, Technological Infrastructure, and Security/Ethical Issues. Leaders of the 4 area teams will form the Integration Team to Design the Regional Data Set. The Integration Team will confer with the Advisory Board, which includes among its members representatives of key regional stakeholders: public agencies, public schools, private businesses, and philanthropic organizations. The results of the work produced by the 4 area teams and by the Integration Team will be shared with a broad base of regional stakeholders at 5 convenings over the course of the project year.



a. Regional Data on Causal Vectors

Regional Scan of Education and Human Resource Development

In Stage 1 of this initiative we will conduct a scan of the region to identify prospective sources of data that could be employed to analyze causal vectors of individual academic achievement and the achievement gap. The project team, which will be led by the Silicon Valley Educational R & D Center, will identify multiple sources within the region as well as relevant sources beyond the

region to construct a broad profile of data sources on the educational conditions, challenges and opportunities in the Silicon Valley. We will seek sources of both structured and unstructured data for in school factors, ranging from fine-grain data on instructional practices and student responses to larger-grain measures of classroom and school conditions. The scan will also identify prospective sources of structured and unstructured data on non-school factors, focusing on regional stakeholders in housing, workforce, transportation, health, mental health, and human services agencies.

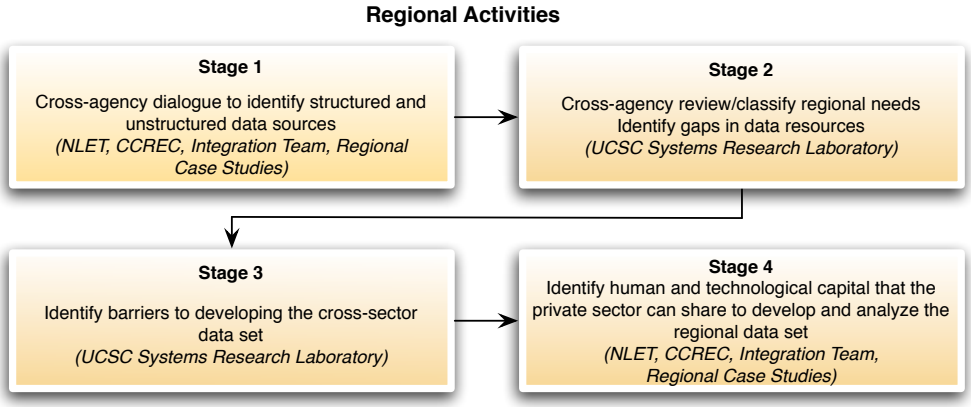
Regional Vectors Detail - Potential Community Stakeholders

In Stage 2 we will identify gaps in the data from institutional sources that are identified in Stage 1. This will be accomplished in two steps. First, the Integration Team will review research on causal vectors of academic achievement and the achievement gap to identify those vectors illuminated by previous research and other potential vectors identified in the literature. Second, we will compare the vectors identified by prior research with the data and sources identified in Stage 1. This will inform the team's discussions with stakeholders at a regional convening to consider how potentially important gaps in data can be filled.

In Stage 3 of the regional scan, the Integration Team will consult with agencies identified as potential data sources in Stage 1 to identify barriers to combining data from multiple sources in the comprehensive, regional data set. This work will be informed by the efforts of the Area Team that examines issues of security, privacy, and ethics. The potential barriers to data integration that are identified in this stage of the regional scan will be discussed with regional stakeholders at a project convening.

In Stage 4 we will scan the region's private sector to identify human and technological capital possessed by companies that can contribute to the development and analysis of the regional database. Silicon Valley is home to many companies engaged in activities relevant to this project, including the aggregation and linking of data from multiple and diverse sources, the application of advanced analytics to large-scale data, and data security. Project team members will confer with companies identified in stage 4 to catalogue their project-relevant human and technological capital. The information gathered in this stage of the regional scan will inform the discussion of the potential role of the private sector in developing and analyzing the comprehensive, regional education data set during a convening of regional stakeholders.

Regional Activities



b. Advanced Analytics

Work in this area will be led by the Center for Statistical Analysis in the Social Sciences. This project will explore a very large data set concept in terms of both sample size and number of variables. We anticipate using a wide variety of statistical methods to analyze these data. In general, many of the statistical analyses will require the use of random coefficient (multilevel) statistical models to accommodate the hierarchical structure of the data in which students are nested within classrooms, classrooms are nested within schools and schools are nested within communities.

Some student performance measures will be categorical. The analysis of these student performance measures will require generalized mixed modeling methods to accommodate the hierarchical structure and the categorical responses.

The proposed database will contain repeated measurements for each student. The availability of repeated measurements with both time varying and time invariant covariates will provide information for answering research questions that cannot be properly addressed using only cross sectional data. The use of modern mixed model statistical packages will provide the tools for analyzing multi-level data with autocorrelated responses over time with the flexibility of modeling an appropriate autocorrelation structure.

We anticipate the use of latent variable structural equation models to assess causal chain models in which one set of variables is assumed to predict a second set of variables and the second set is assumed to predict a third set. With repeated measurements of predictor and outcome variables, it will be possible to define latent variables that are free of certain types of measurement error, and this will provide less biased estimates of relations between predictor variables and outcome variables.

The multivariate data that will be collected on multiple family members (student, father, mother, siblings) can be analyzed using behavior genetic covariance structure models that will provide information regarding both paternal and material influences on student performance. Given the large sample size, it may be possible to obtain subsets of families in which students have half-siblings, non-identical twins, or identical twins so that multi-group covariance structure models can be used to tease out the important environmental effects of family characteristics on student performance.

Although nonlinearity and interaction effects are the rule rather than the exception, most social and behavioral research is forced to make simplifying assumptions because of difficulties of estimating and testing parameters in complex models with small samples. The proposed design will likely produce a very large sample from which appropriately complex models can be applied.

Agent-Based Modeling (ABM)

In 2011, a NSF DRK12 grant was organized by NLET and awarded to UCSC and UT Austin with a contract to the Computer, Computational, and Statistical Sciences Division of the Los Alamos Na-

tional Laboratory (LANL). This grant, *Collaborative Research: An Agent-Based Simulation Environment for Predictive Longitudinal Modeling of High School Math Performance*, takes the multi-agency data in the San Jose Unified School District data warehouse and uses that data as a guide to the construction of an agent-based model. In that investigation, a synthetic data structure is constructed where all the data types are assigned to each student, or agent. From the point of view of the investigators, by making the student a point of integration, the factors affecting their eighth grade math performance could be examined within a holistic representation of each student over time. This early pioneering work with agent-based modeling will also inform this grant.

c. Technical Infrastructure

The Institute for Scalable Scientific Data Management will lead work in this area. There are five technical issues to address in creating a large-scale educational database: ingestion, entity resolution, normalization, storage, and processing. There exists a significant body of technical expertise in each of these areas to draw upon.

The first three are standard database questions that arise when receiving data from multiple datasets with different information, data formats, etc. Ingest is the process of taking in data from a dataset and primarily involves remapping the data into the internal format of the common database. Entity resolution and normalization involve determining when multiple entries from the same or different datasets represent the same person, and revising the database to reflect that fact. The Technical Infrastructure Team will work with database practitioners to determine the best ways to accomplish these tasks, which we expect to involve relatively standard solutions.

The last two are big data questions involving the selection of appropriate infrastructure to support the storage, processing, and querying goals of the system---how best to store and process the data, with sufficient capacity to handle the expected data sizes and sufficient robustness to provide relatively high availability, and how to provide sufficient processing efficiency and capacity to support the expected analysis load. Co-PIs Brandt & Maltzahn have significant expertise in this area, having developed the Ceph (Weil, Brandt, Miller, Long & Maltzahn, 2006) distributed object-based storage system that is part of the Linux kernel distribution, and also SciHadoop (Watkins, Lefevre, Ioannidou, Maltzahn, Polyzotis & Brandt, 2011), a distributed processing architecture based on MapReduce (White, 2012) / Hadooop (Dean & Ghemawat, 2004) for processing structured data.

This project will examine a range of alternative storage and processing architectures to determine the best approaches, our goals will be to use standard techniques, but with an eye toward the scalability that would be required to support a true national-scale educational database. We will examine alternatives including standard databases (which have scaling limitations) and new data processing architectures such as MapReduce/Hadoop, Spark (Chowdhury, Das, Dave, Ma, McCauley, Franklin, shenker & Stoica, 2012) / Shark (Engle, Lupher, Xin, Zaharia, Franklin, Shenker & Stoica, 2012), and interactive database-like processing architectures like Dremel (Melnik, Gubarev, Long, Romer Shivakumar, Tolton & Vasilakis, 2010). We will also look at new research in immutable database design, intended for databases (like this one) where the data does not change after it has been ingested.

d. Security and Ethical Issues

The Center for Collaborative Research for an Equitable California and the SV-ERDC will lead the work in this area. A data set of the breadth and depth that we are proposing presents a myriad of security and ethical challenges. The data set will contain profiles of individuals that reach into sensitive domains, detailing personal physical and mental health conditions, educational achievement and behavioral information, criminal justice system judgments, and data on health conditions and services. Virtually all of these data are highly regulated, with significant barriers that prevent integration and coordination of the data (e.g., California Welfare and Institutions Code, the California Education Code, and HIPAA regulations). This project is committed to the protection of personal privacy, yet recognizes that the barriers must be made permeable so that data can be aggregated and analyzed in ways that make it possible to examine causal vectors impacting student achievement and workforce

preparedness. Therefore, the project will work closely with the regulating agencies to develop guidelines and standards to insure security of the data and privacy of individuals.

In the regional scan of data and data sources that this project contemplates, the investigators will map the diversity of security issues present in the region on two levels. The first is a security scan, of vulnerability factors. What threats and opportunities are available to distract, disable or destroy data and systems in the interaction with school-based systems? The second is data integration, interoperability and performance levels, or success factors. At this level, we will scan and compile the systems used by schools and regional agencies and map their ability to produce coherent and integrated data pictures of students' learning progressions and support networks.

In addition to the two levels of security scan, the project team, joining the expertise of NLET staff and faculty from the UCSC Baskin School of Engineering, will create a security and data management forecast for the region. This forecast will communicate the need for regional data to be secure, interoperable, and pointed toward broader student performance gains. By creating a way to forecast for whole student success in a region, a template can be created that can be expanded upon, tested and put into practice in other regions of the country.

NLET has partnerships that will provide unique consultation expertise on modern data integration, data mining, identity management, and security. These include Resilient Networks, Inc, which has a primary NIST grant (National Institute of Standards and Technology) to pilot cyber-security trust network technology with school districts and allied agencies and corporations.

3. Integration of 4 Major Areas of Work for Data Set Design

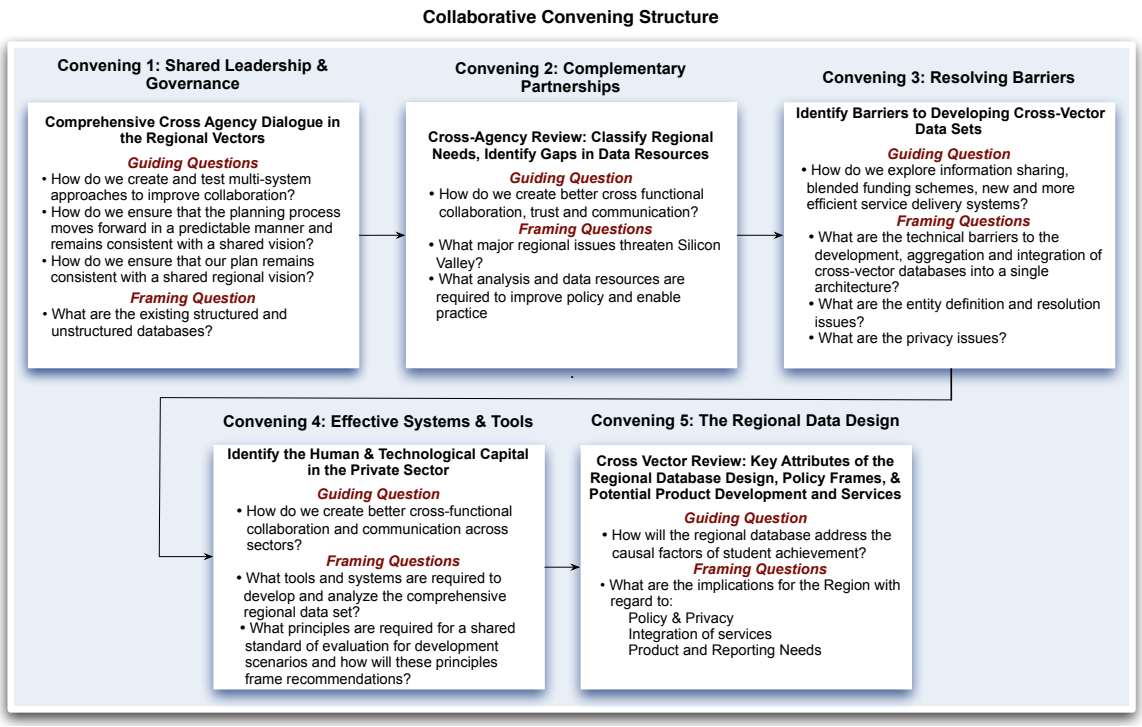
a. Integration and Design Team

During the course of the project year, the PI, the Co-PIs who lead each of the teams focusing on a major area of work (Data on Causal Vectors, Advanced Analytics, Technical Infrastructure, and Security/Privacy/Ethics),and NLET will form the project's Integration and Design Team. This team will meet monthly to share emerging results. These discussions will identify points of intersection across the areas of work and their implications for the overall design of the comprehensive, regional data set. In month 10 of the project, the Integration and Design Team will develop an overall set of recommendations for integrating the results of the work on each major area in a design for a comprehensive, regional data base, including approaches to analyzing the data set, the technical infrastructure to support the data set, and the security, privacy, and other ethical issues that must be addressed. The Integration and Design Team will draft a report which will be distributed and discussed at the final convening of regional stakeholders.

b. Collaborative Convenings

Convenings of the collaborative will explore and delineate the requirements and barriers to the integration of data variables from disparate multi-agency data sources into a data-sharing model with defined data standards and software tools that can address the causal factors of student achievement.

Through 5 regional convenings in 2013-14 between strategic Valley partnerships and education practitioners and researchers, community stakeholders and business and technology innovators, the Design Team will define the roadmap for a scalable set of K-12 solutions in the Valley utilizing the best data methods and systems available. Participants will be selected from among researchers and educators with expertise in research areas relevant to the development of this robust, regional database. Built around key guiding and framing questions, convenings would also include plenary speakers overviewing both current and planned activities and strategies, breakout sessions to ideate on research directions, data requirements, barriers to implementation and panels to begin the process of synthesis.



G. Intellectual Merit

This project will lay the social, intellectual, and technological groundwork to develop a comprehensive, regional data set to support data-intensive research in education. This work has the potential to advance knowledge through empirical research in education and across other domains.

The project team crosses institutional sectors with members from the faculty of the University of California, Santa Cruz (UCSC) and from a private non-profit organization with strong ties to both public agencies and private, for-profit firms in Silicon Valley. The team is also cross-disciplinary; it involves 3 research centers at UCSC and 2 research centers that are multi-campus research initiatives sponsored by the University of California system. Participating UCSC faculty represent the Psychology, Economics, and Education Departments in the Division of Social Sciences and the Computer Science Department in the Baskin School of Engineering. The project team will work with public agencies and private companies in to identify potential sources of data and to explore and negotiate issues, such as proprietary interests, privacy and security, that can present barriers to sharing and integrating across multiple domains.

The project will apply the social and intellectual capital of the cross-sector and cross-disciplinary team to design a data set that possesses the depth and breadth of data lacking in existing data sets and that includes the addition of potentially emerging data sets enabled by new software applications. The data set will have depth by including data ranging from fine-grain measures of instruction to larger-grain measures of classroom, school, and district conditions and by including both structured and unstructured data. It will have breadth by including data from education and from other public domains, such as health, social services, and housing, that can affect academic achievement as well as data from public and private sources that include measures of out-of-school learning opportunities and outcomes.

Importantly, it has always been a challenge to get social agencies, education entities (schools, districts and intermediate education agencies) and outside providers of services to work together across institutional and organizational settings. As a result, individual students in large numbers have lost out on better support for their academic journey. Today, as each institutional or organizational structure is producing its own data within its own data structures, the discontinuity problem is multiplied. The intent of this project is to address this discontinuity in order to service and support the whole student.

This comprehensive, regional data set will open opportunities for research in education that can yield new understandings of student achievement, the achievement gap, and other critical educational issues. Such research will apply advanced statistical methods from the social sciences and from other disciplines, where advanced analytics such as agent based modeling are being applied. In addition, the analysis of a comprehensive, multi-sector data set holds the promise for yielding results that reveal causal vectors, the impact of which can be directly assessed in experimental studies, an approach that has seen increasing application in economics.

Because the comprehensive, regional data set will include data from multiple domains, it will provide opportunities for research that enhances our understanding of cross-domain relations, such as education and health. Moreover, just as the data set will provide new prospects for education researchers by including data from sources beyond the education system, it will similarly create opportunities for researchers in other fields to examine the impact of factors in other domains on important outcomes in their focal fields, e.g., enabling researchers to examine how education and work force factors affect health.

H. Broader Impacts

Underlying the challenges and promises of education are information and computing technologies that are needed to enable large scale aggregations of disparate data sets as well as facilitate geographically and temporally distributed activities associated with student outcomes. This project has the potential to have several broader impacts: facilitate regional planning, advance cross-institutional coordination and cooperation in the region, improve regional services in education and other public service domains, and contribute to advancements in the technology for data mining in education.

In addition to supporting data intensive research in education, a comprehensive, regional data set that includes data from education and other public service domains could inform planning and decision making of public agencies, including schools districts, public health, social services, employment services, housing, and transportation. In the past decade, educators have been encouraged to engage in data-driven decision making. However, in many schools and districts, educators have access to limited data, with an emphasis on standardized test scores. The availability of a comprehensive data set would greatly advance the ability of educators, as well as officials in other public sectors, to make better informed decisions and to engage in data-based planning.

The process of working with multiple agencies to develop the regional data set could contribute to advancing regional coordination and cooperation as described by the research of Annalee Saxenian, a member of the project's Advisory Board. Silicon Valley may be poised for such regional coordination and cooperation, as reflected in discussions at the State of the Valley conference (February 8, 2013), where regional leaders from government and industry discussed the advantages of offering region-wide services ranging from refuse disposal, transportation, business development, and education.

Currently, it is estimated that hundreds of individuals and start-up companies in Silicon Valley are developing educational programs and applications without any basis in research. The findings of research employing a comprehensive, regional data set to identify causal vectors of academic achievement and the achievement gap would provide direction to the development of efficacious educational interventions. In a sense, this work could provide a research-based incubator for educational innovations that could be developed by educators, university-based scholars, educational R & D firms, and private entrepreneurs.

I. Expertise

1. Project Team

a. Principal Investigator: Rodney Ogawa, Professor of Education, Director of the Center for Educational Research in the Interest of Underserved Students and Co-Director of the Silicon Valley – Educational R & D Center, UC Santa Cruz

Professor Ogawa's research has focused on educational policy, school organization, and educational reform. His current study examines how educational policy and the consequent organization of schools shapes educational opportunities for low-income Latina/o youth in 3 high schools in Silicon Valley.

b. Co-Principal Investigator: Douglas Bonett, Professor of Psychology and Director of the Center for Statistical Analysis in the Social Sciences, UC Santa Cruz

Professor Bonett's research has focused on developing statistical methods that perform properly under realistic conditions. He has developed new confidence interval methods, hypothesis testing procedures, and sample size determination techniques. Professor Bonett will take a lead role in coordinating the development of the proposed database structure and data collection processes to issure that the necessary data will be available and organized in a format that will accommodate all of the proposed statistical and analytical analyses.

c. Co-Principal Investigator: Scott Brandt, Professor of Computer Science, Associate Dean for Graduate Studies, Baskin School of Engineering, and Director of the Institute for Scalable Scientific Data Management, UC Santa Cruz

Professor Brandt's research is broadly in the area of Computer Systems. He specializes in Storage Systems, Real-Time Systems, and System and Storage Performance Management. His Storage System research includes high-performance peta-scale object-based storage and the use of new storage technologies to improve storage system performance and reliability. More recently, this work has evolved into big distributed computing, information, and I/O systems. His real-time research focuses on integrating real-time and non-real-time processing into a uniform processing environment and multiprocessor realtime scheduling. His performance management research integrates the two to provide processing and I/O performance guarantees in local and distributed systems.

d. Co-Principal Investigator: Ronald Glass, Associate Professor of Philosophy of Education, PI/Director of the Center for Collaborative Research for an Equitable California, UC Santa Cruz, Acting Associate Director, Center for Information Technology Research in the Interests of Society–Data and Democracy Initiative

Professor Glass' research focuses on integrating research practices with public learning processes to strengthen deliberative democracy and foster equity; public school reform in low-income, racially, culturally, and linguistically diverse communities; issues in moral and political philosophy and their relationship to public education.

e. Co-Principal Investigator: Carlos Malzahn, Associate Adjunct Professor of Computer Science, Direcor of the Systems Research Labo, and Associate Director of the Institute for Scalable Scientific Data managament,UC Santa Cruz

Professor Malzahn joined UCSC in January 2005 after five years at Network Appliance. His current research interests include scalable file system data and metadata management, storage QoS, data management games, network intermediaries, information retrieval, and cooperation dynamics.

f. Co-Researcher: Carlos Dobkin, Associate Professor of Economics, UC Santa Cruz

Professor Dobkin research is primarily focused on the areas of health insurance, substance abuse and education with a particular interest in evaluating the impact of policies intended to mitigate harms or improve outcomes. In recent work he has estimated the causal effect of class attendance on exam performance by conducting an experiment in which half of the students in several large classes were encouraged to increase their class attendance.

g. Co-Researcher: Eduardo Mosqueda, Assistant Professor of Education, UC Santa Cruz

Professor Mosqueda's research analyzes large-scale data sets to examine the influence of families and school structure on the academic performance of Latina/o students in mathematics.

2. Project Consultants

Consultation will be provided by the National Lab for Education Transformation (NLET), a private non-profit organization that is partnering with UC Santa Cruz in the Silicon Valley Educational R & D Center, which is the research center leading this project.

a. Gordon Freedman, President, NLET and Co-Director of the SV-ERDC

Gordon Freedman was vice president of global education strategy with Blackboard, Inc. from 2005 to 2011 and executive director of the Blackboard Institute from 2009-2011. In those positions, Freedman visited governments, universities and schools in eighteen countries and became conversant

with most of the technology efforts in education. Freedman founded NLET in 2011 to bring together the best cross-sector expertise, methods and tools to help analyze the current system of education in order and assist in redesigning learning and education infrastructures for the future.

a. Marcy Lauck, NLET, Director, National Data Strategy.

Ms. Lauck brings expertise in guiding strategic institutional change and data-based quality management processes. Under her leadership, San Jose USD built a comprehensive K-12 educational warehouse, encompassing 16 years of data, more than 50 million records, 113 objects and over 3,500 attributes. These data provided the foundation for the development of repeatable, data-informed processes, predictive analytics and the means to create and evaluate individualized student support systems.

b. Bill Erlendson, Ph.D., NLET, Director, Public Policy Development.

Building on twenty-five years of teaching and administrative experience in public school and post secondary education, Dr. Erlendson served as Assistant Superintendent for Educational Accountability and Community Development for the San Jose Unified School District and Adjunct Professor for National University in the field of Public Policy Development. As Assistant Superintendent, Dr. Erlendson developed national models in public engagement that served to guide and inform system wide cultural change and policy development.

c. Nancy Netherland, President, Netherland and Associates

Nancy Netherland develops strategic content through data analysis and research, and guides strategies for securing public and private grants, cultivating private sector donors, and informing policy and social change. Her areas of expertise are safety net health systems and educational entities ranging from cradle to career.

Advisory Board

The project's Advisory Board (see Regional Scan/Advisory Board diagram on page 8) includes leaders of public agencies that operate in domains that bear on education, private non-profit organizations that provide programs in and related to education, other academic institutions, private sector firms that are involved in technology related to the project, and educational philanthropy in Silicon Valley. Their letters of support for this project are attached.

1. William T. Coleman III, Chair and Chief Executive Officer, Resilient Network Systems
2. Gina Dalma, Program Officer, Silicon Valley Community Foundation
3. Rebecca Kelley, Chief Innovation and Knowledge Officer, STRIVE Network
4. Anne Ehresman, Exective Director, Project Cornerstone, YMCA of Silicon Valley
5. Nancy Pena, Director Santa Clara County Mental Health
6. Ramona Pierson, Founder and Chief Executive Officer, Pierson Labs
7. AnnLee Saxenian, Professor and Dean, School of Information, UC Berkeley
8. Michael Watkins, Superintendent, Santa Cruz County Office of Education

References

Buck, J., Watkins, N., Lefevre, J., Ioannidou, K., Maltzahn, C., Polyzotis, N., & Brandt, S. (2011). SciHadoop: Array-based query processing in Hadoop, *International Conference on High Performance Computing, Networking, Storage and Analysis (SC)*, pages 1-11.

Burdick-Will, J., Ludwig, J., Raudenbush, W., Sampson, R. J., Sanbonmatsu, L., & Sharkey, P. (2010). *Converging evidence for neighborhood effects on children's test scores: An experimental, quasi-experimental, and observational comparison*. Brookings Institution.

Dean, J. & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters," in OSDI 2004. San Francisco.

Duncan, G. J., & Murnane, R. J. (2011). *Whither opportunity?: Rising inequality, schools, and children's life chances*. New York : Chicago: Russell Sage Foundation.

Duncan, G. & Raudenbush, S. (1999). Assessing the effects of context in studies of child and youth development. *Educational Psychologist*, 34, 28-41.

Engle, C., Lupher, A., R. Xin, R. Zaharia, M., Franklin, M. J., Shenker, S., & Stoica, I. (2012). Shark: Fast data analysis using coarse-grained distributed memory, in *SIGMOD '12*. Scottsdale, AZ.

Farkas, G. (2003). Racial disparities and discrimination in education: What do we know, how do we know it, and what do we need to know? *Teachers College Record, 10*5, 1119-1146.

Gamoran, A. (2009). The disciplinary foundations of education policy research. In G. Sykes, B. Schneider & D. Plank (Eds.), *Handbook of Education Policy Research*, pp. 106-110. New York: Routledge.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *The Journal of Human Resources, 39*, 326-354.

Jencks, C. & Mayer, S. (1990). The social consequences of growing up in a poor neighborhood. In L. Lynn & M. McGeary (Eds.), *Inner-city poverty in the United States* (pp. 111-186). Washington, D.C.: National Academy Press.

Jones, N. D. & Schneider, B. (2009). Social stratification and educational opportunity. In G. Sykes, B. Schneider & D. N. Plank (Eds.), Handbook of education policy research (pp. 889-900). New York: Routledge

Lankford, H., Loeb, S, & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*, 37-62.

Massey, D. S. & Denton, N. A. (1993). American apartheid: Segregation and the making of the underclass. Cambridge, MA: Harvard University Press.

Melnik, S., Gubarev, A., Long, J. J., Romer, G., Shivakumar, S., Tolton, M., and Vassilakis, T. (2010). Dremel: Interactive analysis of web-scale datasets, in *Proc. of the 36th Int'l Conf on Very Large Data Bases*, pp. 330–339.

Oakes, J., Gamoran, A., & Page, R. (1992). Curriculum differentiation: Opportunities, outcomes, and meanings. In P. Jackson (Ed.), Handbook of research on curriculum (pp. 570-608). New York: MacMillan.

Rumberger, R. W. & Palardy, G. J. (2004). Multilevel models for school effectiveness research. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 235-258). Thousand Oaks, CA: Sage.

Sage A. Weil and Scott A. Brandt and Ethan L. Miller and Darrell D. E. Long and Carlos Maltzahn, "Ceph: A scalable, high-performance distributed file system," The Symposium on Operating Systems Design and Implementation (OSDI}, 2006, pages 307-320.

Saxenian, A. (1994). *Regional advantage: Culture and competition in Silicon Valley and Route 128.* Cambridge, MA: Harvard University Press.

White, T. (2012). Hadoop: The Definitive Guide. O'Reilly Media Inc.

Wilson, W. J. (1987). *The truly disadvantaged: The inner city, the underclass, and public policy*. Chicago: University of Chicago Press.

Zaharia, M. Chowdhury, M., Das, T., Dave, A., Ma, McCauley, J. M., Franklin, M. J., Shenker, S., & Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing in *NSDI 2012*.

Zimmer, R. W. & Toma, E. F. (2000). *Peer effects in private and public schools across countries. Journal of Policy Analysis and Management, 19*, 75-92.