Draft Report: Strengthening and Sustaining UC Berkeley's Position in the Disciplines of Data Science: A Proposal for Administrative Structures and Organizational Sustainability

Submitted by: Proposal Committee for the Formation of a Department of Data Science

Submission Date: 17 April 2023

Introduction

The new *College of Computing, Data Science and Society* (CDSS) [regental approval pending] represents a major step for UC Berkeley, providing institutional support for faculty involved in the intellectual transformations emerging from data science tools and the creation of the space for these new disciplines to grow and thrive. Aspects of the institutional form of CDSS can be traced to the establishment of the Center for Computational Biology (CCB), the Berkeley Institute for Data Science (BIDS), the reimagining of the School of Library and Information Studies as the School of Information and the deep collaborations between the Departments of Statistics and the Department of Electrical Engineering and Computer Science. A series of reports and actions dating back nearly a decade, including the 2015 Faculty Advisory report, converged on a recommendation to form a unit spanning what was then thought of as the entire scope of the new field of data science. Concurrent with this early development was the creation of the Division of Data Science and the Data Science Undergraduate Major through a joint venture between the College of Engineering and the College of Letters and Science. The major has grown rapidly, to the point that it is now one of the most in-demand majors on our campus. The Division is on a path to becoming a College and is the primary manager of the DS major and minor.

Our committee was charged with envisioning the next phase of the development for this intellectual transformation. We were asked to consider how to provide institutional support for emerging disciplines. We find that new data science disciplines are lacking support because they are either (a) not recognized as core to the departments within CDSS or (b) not viewed as core in Berkeley departments outside of CDSS, or both. Consequently, there is no established process or body with the responsibility for taking the lead in creating and nurturing these new disciplines. Many of these possible fields can be described as Data + X, where the “Data” reflects the part of the field closest to EECS or Statistics and the “X” reflects the aspect closest to any other discipline across humanities, social sciences, and physical sciences.

These challenges are felt most acutely in the context of faculty hiring. Researchers who are most accurately described by the combined descriptor “Data + X” are, to our minds, falling through the cracks too often as individuals; moreover, we are not investing in the critical mass of scholars in any X that would be needed to seed the development of a new field. Although we recognize that there are examples (such as through joint hires) where UCB has overcome institutional barriers to hiring extraordinary individual scholars in the mold we describe, institutional support for such hires is neither consistent nor reliable. From the perspective of faculty broadly involved in data science, each of these hires appears to rely on an individual act of heroism by a visionary in a disciplinary department who recognizes a scholar of exceptional potential and launches tireless campaigns with colleagues in their own unit and a CDSS partner unit to create a joint hire. Each hire, then, bears the scars of the idiosyncratic nature of the process, and each hire does not present as a model that can be replicated. Consequently, as a

faculty community, we do not view this as a strategy capable of providing scholars the support they need to create, sustain and grow new fields on our campus.

Rapid advances in the data-driven analysis of natural, physical, social, and cultural phenomena–and the development and optimization of new methods for supporting those analyses–represents an enormous new opportunity for advancing human knowledge. Berkeley, with CDSS, has the exceedingly rare opportunity to further distinguish itself from the many tier-one global universities lining up to develop sophisticated, supportive environments for research and teaching in these intellectual spaces. This potential has been echoed in all of the deliberations of the committee over the past several months, and there has been a great deal of effort to ensure that our next steps are confident and set us on a trajectory toward success.

We describe our deliberative and consultative processes and our recommendations in more detail below. We begin by recognizing the two questions that have guided our thinking:

What is the intellectual landscape that is not being adequately filled with new faculty in the current ecosystem of departments and AGGs [augmented graduate groups] on campus?

The number of interrelated and cross-cutting disciplines and their attendant faculty that engage in some form of data science at Berkeley is staggering. Indeed enumerating them is a somewhat absurd exercise, not least because nearly all the traditional STEM fields now have a considerable intersection with data science, but also because the non-STEM fields engaging in data science now range from public health, social welfare, business, and law among the professional schools, to anthropology, sociology, political science, languages and literature, art practice, film and media, history, and linguistics. Yet, what is missing is an institutional framework that would facilitate creating a coherent intellectual home for, say, a historian working on machine learning approaches for accessing and analyzing heterogeneous archival data from 17th-century France, or a scholar in public health working with natural language processing (NLP) and network methods for ethnographic and epidemiological understanding of a viral outbreak in South Asia. The possibilities spiral outward from the list above. For many scholars (faculty and graduate students) working across disciplinary boundaries from an existing departmental base is adequate. For others, who see themselves creating new fields, this is less so. To address this opportunity. CDSS and the data science institutional structures we outline below will provide important platforms for (a) creating and supporting the emergence of these transdisciplinary research communities engaged in groundbreaking integrative data science, (b) attracting and fostering the best talent in newly emergent fields that have no single “core” in any one department, (c) creating vibrant communities of inquiry that are attractive to the very best graduate students and future researchers, and (d) provide a model of learning and inquiry that will be an inspiration for undergraduates across all the diverse parts of campus.

At the start of our deliberations committee members spoke wistfully of numerous potential hires that, in recent years, due to the barriers within and between schools, departments, and disciplines, were either not offered positions or who decided against joining our faculty. Many were eventually hired at our Tier 1 competitor schools. Across the discussion, a critical gap was

identified, namely the lack of a central framework under which interdisciplinary hires could occur and from which intellectual homes could grow. Institutional structures that support communities of research that celebrate the disciplinary border-crossing that is a hallmark of “data science” across fields is necessary if this effort is to succeed.

While much of our report focuses on new units that would support the emergence of new disciplines, we also are attentive to the need to strengthen existing structures and opportunities for interdisciplinary research across pairs of disciplines represented by existing departments and AGGs. Some of the failed searches we discussed were of potential colleagues who sit on the bubble of existing fields, not necessarily creating new ones, but also not viewed as at the center of either field. Methods to ensure the robustness and fairness of joint appointments across two departments are likewise needed, in particular for junior faculty. At more senior levels, such methods could alleviate an observed lack of understanding in the “core” discipline which today translates into diminished opportunities to develop research communities. With such support, it should become easy to attract the best graduate students to faculty groups spanning departmental or college boundaries.

What structure(s) would allow interdisciplinary scholars to thrive and to create new disciplines?

Fortunately, some of these barriers can be addressed through institutional structures that recognize the differing – and shifting – needs of researchers at all stages of their careers and that make investments in new disciplines emerging from interdisciplinary research. Newly hired data science faculty need to receive institutional support at the level of an established department, require access to graduate students and should understand their pathway to tenure in order to thrive with the campus.

The productive discussions of representatives from departments, centers, programs, schools, and divisions across campus have led to an agreement that there is no single solution to the complex inter- and transdisciplinary landscape that characterizes the broad and integrative aspects of data sciences described above in our first question. What is abundantly clear is that the model of “one Data Science department” as the locus of all data science activity would be counterproductive and undermine the remarkable progress made across the campus to date. Rather, the committee suggests pursuing a multiple structures approach to ensure flexibility, to recognize existing structures that have allowed new fields to emerge (e.g. CCB), and to recognize that flexibility, varying sizes, and varying institutional designations might best provide an environment to attract, support, and nurture the most exciting researchers working across what are often traditionally conceptualized as separate disciplines. The opportunity to create synergies where such opportunities are currently hard to discern should be seen as a considerable advantage. We believe that a series of small(er) units, including departments, graduate groups (and augmented graduate groups), centers and the like, however seemingly “messy” on an organizational chart, will turn out to be a remarkable feature of this constellation, and not a short-lived solar flare.

Committee's Charge

As described in the committee's charge letter the Committee was formed to determine what potential academic structures could be put in place to "...advance the frontiers of data science [and] play a role in leading collaborative interactions and joint appointments across the campus…" The campus has been slowly building momentum in formalizing the intellectual home for data science since at least 2015. Starting with the growing interest and research in the field which led to the launch of the Data Science major (in 2018) through to the current state where a proposal to convert the nascent Division of Computing Data Science and Society into a fully mounted college is pending the UC Regents' consideration. The Proposal Committee's inception is just one piece of this multi-layered journey.

The questions outlined in the charge letter about the prospects to create a world-class program, expanding data science opportunities for the campus, and logistical details (funding, faculty allocations, etc) are addressed by our recommendations, illustrated below. Ultimately, we came to understand that there are more opportunities for exciting new DS units than the campus can support with new faculty FTE. Recognizing this, **we recommend creating a process for colleagues across the campus to propose institutional forms for intellectual clusters that will seed the development of new fields.**

Intellectual Landscape

The Committee reviewed a number of examples of exceptional scholars who don't fit neatly into the intellectual and research priorities of existing UCB departments. The list was idiosyncratic and not subject to a single set of criteria, but it was long enough to give the impression that many departments could be filled with exciting scholars a few times over.

The Committee also reviewed the history of the Center for Computational Biology (CCB). CCB is illustrative of the interdisciplinary transformation that motivates growth in Data Science. CCB is a thriving unit that includes scholars who are leading a transformation of biology. CCB began as a new initiative center and later became an AGG. Over its 20+ year history, the field of computational biology has emerged, with new journals dedicated to the subject and exciting new insights into biological systems. However, recently it has been challenging to sustain and support the intellectual transformation that CCB has led on our campus (see, for example, comments from BIR reflected in the Senate's letter of May 5, 2022 on the CDSS college formation proposal). On the one hand, many scholars who are focussed on some form of "big data" are being hired directly into our biology departments. On the other hand (per Dean Michael Botchan of L&S Biological Sciences), the sort of scholars who are the core of CCB still would not be a priority for our biology departments. CCB has become to our committee a prime example of the type of faculty holding unit we believe could provide founding leadership in other intellectual areas that are currently viewed as interdisciplinary but might, with institutional support, become new disciplines.

It is our committee's view that the areas of focus for new CDSS units be ones where there are unique opportunities for scholars who both create new disciplinary tools and then evaluate and/or apply them from a disciplinary perspective. Those scholars will be creative in fundamentally different ways from either scholars who focus most of their energy on building new tools and from scholars who focus most of their energy applying tools developed by others in their disciplines. They will drive a more rapid and distinct cycle of discovery and will have a deeper understanding of how artificial intelligence, machine learning, and similar budding specialties are advancing adjacent fields. Investing in these emerging new fields would give a first mover advantage to our campus, allowing UCB to be a leader in new data science enabled disciplines.

It is the view of our committee, validated by discussions with senior faculty, department chairs, and deans in fields that could intersect with data science, that scholars of this type are not prioritized in the current UC Berkeley ecosystem of faculty searches. The interdisciplinary nature of their work leaves them in danger of not fitting into department faculty searches, even the open area searches. **We have concluded that institutional leadership with a focus on growing this "in between" space is essential to data science's success at Berkeley.**

It is also worth noting that we are persuaded that current structures do not support transformational graduate education across all interdisciplinary fields, which would be framed as

Data Science + X (X being a “traditional” academic field, as described earlier). There are important successes supporting data science within existing X fields. The D-lab, for example, has provided foundational support for students in some current disciplines. However, it is not designed to lead an interdisciplinary transformation in those fields. Similarly, the DATA 100 and 200 courses have been invaluable for PhD students seeking a foundation on which to ground exploration of new opportunities to advance their research with data science tools. Still, there is no critical mass for artificial intelligence, machine learning or other data science education among graduate students in our disciplinary departments. Further, it is unlikely that the most talented interdisciplinary faculty or graduate student scholars will achieve their leadership potential in the disjointed manner that we are currently organized.

Our peer institutions (e.g., MIT, Stanford and Columbia) have approached the explosion of opportunity in data science through a combination of substantial enhancements (25 or more new faculty positions) in FTE allocated to departments of computing and/or statistics and a nearly equal amount of joint FTE for those departments and other departments. UC Berkeley has done nothing at this scale. Mostly, we have transformed existing FTE within EECS, Statistics and the I-School to focus on development, application and evaluation of some of the societal impacts of artificial intelligence and machine learning. The Center for Computational Biology is a prominent exception that has led the way in creating a new field with allocation of seed FTE funding and exceptional hires who have contributed to establishing Computational Biology as its own field – distinct from other biology and computational disciplines.

Structures and Process

Deciding the intellectual scope, breadth and balance of new fields is challenging in the absence of concrete proposals and faculty who are self-identified advocates. Our committee is convinced the success of new units will depend on groups of faculty committed to creating new fields, to mentoring junior faculty, and to establishing graduate programs as the new field takes shape. To ensure that proposed programs meet all of these (and other criteria), **we recommend that the campus establish an opportunity to describe small, intellectually focused, faculty-led programs in emerging fields of study that intersect with data science**. Inspired by the success of the new initiative centers and the specific example of CCB, we propose that the campus create a process that would lead to the creation of three to five new intellectual programs with approximately 25 new FTE (15 philanthropic). The number and size of each new unit will depend on the scope of the intellectual opportunity, and the existing faculty expertise around campus in addition to new hires. We expect that new units will consist largely of faculty holding > 0% FTE within CDSS and that the majority of the new FTE will be >50% in CDSS, but not exclusively so; we encourage a system of joint appointments and faculty affiliations across campus. We emphasize that the process we propose is intended to seed new disciplines. They are not intended to substitute for the form of collaborative hiring between CDSS and other Schools and Colleges that is already occurring and strengthening our existing disciplines. We recommend that the campus process be short, initiated in Fall ‘23 and completed by Spring ‘24, as we are nervous about losing a first mover advantage in some fields and about losing momentum more broadly.

The appendices to this report provide a set of sample intellectual areas that faculty might coalesce around. They are not fully fleshed-out ideas, but rather they are examples that emerged in our conversations that illustrate the wide potential for new fields. They are not intended as recommendations or to define winners in the campus process outlined below. We do note that four philanthropic FTE are associated with the BIDMaP gift. We expect those FTE to count toward the 25 we envision and toward the 15 philanthropic. The form of a new unit that incorporates those FTE should emerge as a proposal to the process outlined below. Our committee discussed whether we should specify the size of new units. We believe a critical mass of 10 or so faculty is essential and that 5 new faculty with commitments from current faculty should reach that number of committed individuals. We also discussed the possibility that smaller groups of faculty might want to band together to build something bigger, for example a unit that might flourish by combining CCB with other natural science targets such as BIDMaP might be of interest to some. We choose to leave that choice to faculty proposing new units.

Forms of new units

The canonical example of new units we have in mind is an AGG and this would entail an expectation that any new FTE be 50% or more in that AGG. The process should be open to any other institutional forms that make a commitment to creating new fields of research and

scholarship. For example, two or more departments could collaborate on creating a new field in the form of a graduate group. In this case, proposals could indicate a plan for the new FTE to be located in those departments, possibly along with development of a recognized disciplinary division or cohort within the department and a commitment to formation and sustenance of a graduate group that would be the seed of the new field.

Efficacy review and possible sunset

While it is likely that most of the units the campus chooses to invest in will be successful, we recognize that some might not be. Accordingly, we recommend that new programs be reviewed on the typical departmental time scale. This should be approximately eight years after completing the hiring of the initially proposed faculty cohort and no more than 12 years after the start of the new unit. Criteria for review should emphasize the extent to which a ***new field of scholarship*** has emerged/is emerging and the extent to which UCB is recognized as a leader in that field. Excellence alone in research and graduate education should be seen as insufficient for continuation. Units that are dissolved should terminate their graduate programs and the associated faculty should either find homes in existing departments or return to their home department.

Proposals should include the following elements:

- What is the intellectual need for the proposed area, and how is it not already served on campus? Proposals should focus on the opportunity to seed and nurture creation of a new field. Description of the research domain should represent the bulk of the proposal.

- Which current faculty are committed to support, mentor, lead hiring, and shepherd growth of the new field? This should include commitments to leadership within the campus and the larger intellectual community in creating the new field. It should also include specific commitments to leadership within the new unit to search for and mentor new hires, and to create a graduate curriculum. Shifts in FTE (%) from current units to the new unit should be described and sufficient to allow current faculty to devote time and energy to leadership of a new unit.

- What institutional structure (e.g. departmental collaboration and a graduate group, augmented graduate group, department) is planned?

- What is the relationship of this field to other units on campus? Are there existing partners who would support joint appointments and searches.? How will creation of the new field envisioned by the proposal impact fields at its boundaries?

- What would likely be essential elements of a graduate curriculum? To expedite a selection process and reduce the burden of writing proposals, we ask for sufficient detail to know that a curriculum might emerge, but not a full proposal for a new degree.

- What is the rationale for the new FTE requested? We are imagining the campus supporting five focussed units with up to five new FTE each. However, proposals could be larger or smaller. Include proposed timing for hiring (e.g. all at once, over five years)

- Describe the potential for philanthropic funding of the requested FTE.

Proposals need not say much about undergraduate teaching. They should briefly summarize the unit's commitment of effort to undergraduate teaching (e.g. percentage of faculty teaching). Our intention is that all of the units formed by this process will contribute their undergraduate teaching in the DS Major with assignments made by the Data Science Governance Committee. Developments of new undergraduate majors, if any, at some later time, would likely need to draw students away from the DS major as we cannot afford to hire DS related FTE that do not contribute to reducing the present overload on faculty who are teaching in the DS major. We don’t see that situation changing any time soon.

Similarly, proposers should not say much about administrative support. We recommend CDSS will provide clustered administrative support to any units formed within the College. To the extent possible, these new units, or at least the new hires, should be situated in the Gateway Building. We recommend proposals omit discussion of space and that space plans be developed after selection of the new units.

Timeline for Program Development & Launch

Our committee feels a sense of urgency—in some areas, we have already lost the opportunity to be first movers to our competitor institutions—while also appreciating the time required to consolidate faculty interest and perform due diligence in the review process. To that end, we suggest the following timeline:

- Summer/Fall 2023: Review of this recommendation
- Summer/Fall 2023: Announcement; faculty self-organize into groups.
- January 2024: Letters of intent due; Chancellor's Advisory Committee on Data Science (described below) appointed in Fall 2023 helps organize related proposals.
- March 2024: Proposals due to the Advisory Committee for evaluation and also to be vetted by CDSS development to evaluate potential for philanthropic support.
- May 2024: Initial recommendations of the Advisory Committee conveyed back to the proposers with a copy to the Vice Provost for the Faculty and the Budget Committee.

Selection Process
We recommend the creation of a Chancellor's or Provost's Advisory Committee on Data Science to solicit and review proposals. After the selection process, this Advisory Committee would have responsibilities as part of their core objective to serve as the campus think tank for advancing data science in a coordinated manner across multiple colleges and schools.

The Advisory Committee would be tasked with endorsing a proposal before it moves forward for resource allocation approval by the Vice Provost for the Faculty, the Vice Provost for Academic Planning and review by the Academic Senate following procedures outlined in the compendium. The Advisory Committee would effectively serve as the control point through which a comprehensive view of the campus landscape for data science is presented and be positioned to advocate for resources based on that well-informed perspective.

The initial key responsibilities of the Advisory Committee should include:

Managing the process of proposal submission, evaluation and recommendations for new small units. This should include encouraging merging of similar efforts, encouraging faculty from a wide range of fields to participate and ensuring an outcome that seeds ideas where Berkeley takes appropriate risks to ensure exciting outcomes.

Advise the Chancellor and Provost on the range of opportunities represented by the proposals and how they can best be supported from initial concept through to implementation.

Ongoing management after establishment

After the establishment of the initial cohort of three to five small units described in this proposal the Advisory Committee could continue in a modified capacity.

It could pivot to focus on being the campus advocate for interdisciplinary hires. The advisory committee should also continue to work with established academic units to ensure that requests for cluster and joint hires result in adequate coverage, while avoiding unnecessary overlap. In addition, it could advocate for occasional faculty hires who are exciting for their interdisciplinarity but falling through the cracks in departmental searches. While there is a strong consensus on this committee that the majority of the ~25 FTE we propose should be “pre-allocated” into the three to five small units, there is also support for holding some out for interdisciplinary opportunities. While this idea had wide appeal in concept, we struggled with a recommendation for implementation other than allowing an advisory committee to add its weight to a recommendation for an exceptional hire.

Proposed Relationship to Existing Campus Programs and Units

Our committee had extensive discussions about how to establish relationships between emerging new disciplines and existing units. The challenges we recognized include the movement of disciplinary divisions within the EECS and Statistics departments as well as the prevalence of scholars within existing departments who are using new data science tools to advance their scholarship.

During the course of our conversations as a committee, both EECS and Statistics have begun internal discussions about creating new disciplinary divisions within their units. The new divisions would recognize that artificial intelligence/machine learning (AI/ML) and other data science ideas are distinct fields that are developing independently from other sub-disciplines of statistics or EECS. As these internal conversations are ongoing, it is important that the creation of new units not interfere with the development and identity of scholars centered in these two departments. Similarly, the new units created should not be seen as efforts that are core to any other existing department.

At the same time, new fields are emerging at the boundaries of old ones and it will no doubt be appropriate for joint hires between the new units and existing departments. We encourage such appointments as a way of strengthening the pool of available faculty mentors.

CDSS is already in the process of joint hiring with other units in areas that might be nucleating new fields. For example, a joint hire between CDSS and the College of Chemistry will support the Bakar Institute of Digital Materials for the Planet (BIDMaP), and a joint hire in progress between CDSS and MCB will grow the intellectual presence of research at the intersection of cancer and AI. We envision that the usual rules of joint appointments and participation in faculty governance of CDSS would apply to faculty positions created through the process outlined in this report. Any new faculty hired with joint or full appointments in CDSS would be members of the governing faculty of CDSS, and any existing faculty who move appointment lines fully or partially into CDSS as part of the creation of new units would also be governing members of CDSS.

It is important that the new units created fill an unmet intellectual need and do not overlap with existing units. This is especially challenging in units that have been supporting the DS revolution with interdisciplinary appointments and that have, as a result, been undergoing substantial internal transformations and resetting of the boundaries of their fields.

We solicited comments on the gist of ideas being developed in this proposal from the three CDSS departments. The Department of Statistics had a very active conversation in parallel with our committee. Data Science has become an important element of the identity of scholars in the

field of Statistics. Our proposal to seed new disciplines in the space between data science and other fields is not intended to usurp territory within the field of Statistics. Similarly, it is not intended to usurp territory that is clearly in the bounds of any other field. The idea that there is intellectual space outside of existing departments, schools and colleges and their divisions is the subject of a healthy debate on our campus. As we noted above, in those areas where scholars do not see new fields emerging, we support the usual process for allocating FTE in our existing departments as that is where leaders are responsible for building consensus about the future of their fields.

Some of the comments from the CDSS departments follow below. These comments should not be taken as endorsements by the authors of the full text of this proposal.

Statistics
The Statistics Department has a long history of cross-disciplinary engagement and faculty hires. In response to the opportunity and challenge implied by the formation of our [Data Science Department] committee, the Statistics Department has had internal discussions about a proposal to change its name from the Department of Statistics to [the] “Department of Statistics and Data Science,” and to establish a division within the department to house and foster applied research. Statistics is supportive of the idea of small focused units with a core that is separate from their department and looks forward to participation/collaboration in these new intellectual spaces.

EECS
The EECS Department’s Executive Committee discussed this committee’s charge letter. While not in general supportive of a single DS Department, the EECS Executive Committee expressed support for a structure for hiring interdisciplinary researchers in data science. A group of faculty from EECS presented a proposal to our committee which aligns with the structure discussed in this proposal.

I-School
The School of Information emphasizes that it serves as a logical home for many of the proposed interdisciplinary scholars named in the sample intellectual areas [of this proposal]. These include cultural analytics, social science applications of data science, computational social sciences, and topics relating to technology and social justice, and the study of information and misinformation. To the degree to which interdisciplinary data science is already core to its mission, the I-School encourages the creation of interdisciplinary searches in this space, and would welcome faculty in particular who explore data science in its intersection with the social sciences and humanities and other topics described above.

Implications for Existing Augmented Graduate Groups (AGGs) Within CDSS (CCB/CPH)

We see no direct implications for the two AGGs that already reside in CDSS, Computational Precision Health and Computational Biology, as they preexist, have FTE allocations and a

mechanism to express a need to grow and would not necessarily be among the small units to be formed as part of the process discussed in this report. The question of whether and when Computational Biology should become a department also need not bear directly on the process we propose. Both groups were an inspiration for the model we are proposing. In other words, should the Computational Biology or CPH faculty decide to propose department formation, that department need not count as one of the three to five units envisioned in this report, but they could, as the other new units, request new FTEs allocated for hiring in interdisciplinary data science.

Recommendations of Staffing and Resource Requirements

These new DS units will each require administrative support on a scale comparable to small departments existing elsewhere on campus. Rather than piecemeal support, we envision clustered support (reporting up through CDSS) of dedicated staff that will serve all small DS units. CDSS is already in the planning stages for clustered support for CCB, CPH and the DS undergraduate program. We imagine that, for the scale of new units proposed, an initial administrative unit of approximately 5.0 staff FTE will be needed including: a DS units manager (akin to a department manager), AP analyst, graduate student affairs officer, IT liaison (including purchasing), and a DS units financial and events coordinator (responsible for fiscal management, travel coordination, visitors, and events management). Additionally, the HR needs and pre/post awards needs of new DS unit members will either be served by CDSS's regional shared service partner (ERSO) or their home departments (in the case of split appointments). As new undergraduate programs are envisioned, the undergraduate student advisor services would continue to be provided through the data science major and the college. We recommend that CDSS also allocate a resource to facilitate fundraising in coordination with the central development office. Ideally, CDSS would put staff in place to help facilitate these new units' incubation and support the onboarding process for the new faculty FTE soon after (or simultaneously with) recommendations that proposals go forward. If new units are primarily centered in existing departments, those departments would be responsible for staff support.

We feel that the in-person connection between DS unit members and their students (as well as amongst different DS units) is key to facilitating the sum-greater-than-the-parts success of this program. Ideally, dedicated office and meeting space for all the DS unit members–all sharing the same "water cooler" in the same building and hallways--would be allocated. Given the pressing space constraints however, we recognize that new dedicated space may be too challenging.

At a minimum, we recommend that a flexible meeting and office space in the Gateway Building be allocated to DS units. DS unit heads/chairs and the DS unit staff would have fixed spaces and all other members would have hoteling space. Meeting space for small and big groups would be arranged following Gateway allocation rules. We envision each DS unit conducting its own weekly seminar/colloquium series in the Gateway Building, open to all members of the community. It is our vision that the intellectual project being proposed here will best be served by being in close physical proximity to other data science faculty, e.g. in the Gateway. The

committee chose not to explore the space constraints or their solutions in any detail at this time, as we think those are best explored by a group more knowledgeable about space. We imagine this could occur with support from campus space planners in parallel with additional vetting and discussions that build on this report.

The other key resource for new units is graduate student support. Primary support for graduate students will be through GSRs from individual faculty research grants. It will be important for the campus and CDSS to develop some expectations for institutional support of graduate programs as GSIs in the DS major or other programs, initial support for new graduate programs, and allocation of admissions to new programs.

Philanthropic Prospects

We recognize that in this time of austerity supporting three to five new units is a tall order. Further, given the frozen size of the UCB faculty over the last two decades, 25 new FTE is a similarly tall order. Philanthropy will be essential to the success of our recommendations. Collaborations with existing units who agree to invest some of their FTE in new fields at the boundaries of their discipline will also be essential.

We recommend that the model for new units be three philanthropically funded FTE for every two state-funded FTE. Current models for philanthropic FTE are a mix of state (⅓) and donor-funded (⅔) support. Consequently, with this recommendation, we are suggesting the allocation of 15 state-funded FTE to new fields from a combination of FTE that are not assigned to the floor of any existing department and from departments that choose to invest some of their floor in a new discipline.

In her three years at UCB, Associate Provost Chayes has raised funds for 14 philanthropically funded FTE. Of those, we suggest the four BIDMaP FTE be considered part of the cohort of 15 FTE which is the target outlined here. Once fundraising for the Gateway is complete, we believe that an additional 11 FTE is eminently possible over the next three to five years. AP Chayes has expressed commitment to fundraising for FTE in all disciplines that have a home in CDSS.

Philanthropic support for the graduate and undergraduate educational mission of CDSS is also robust and will likely be strengthened by the excitement of the new ventures represented by the creation of new fields and institutions to support them.

Appendices

Examples of themes for new small units

- Appendix A: Proposal for a “small unit” in Culture Analytics
- Appendix B: Proposal for a “small unit” in Data Science + Material Science
- Appendix C: Proposal for a “small unit” that interfaces Data Science + Applied Disciplines
- Appendix D: Proposal for a “small unit” in Data Science + Societal Scale Infrastructure Systems
- Appendix E: Proposal for a “small unit” in Computational Biology
- Appendix F: Proposal for a “small unit” in Social Science Applications
- Appendix G: Proposal for a “small unit” in the Physical Sciences
- Appendix H: Proposal for a “small unit” bridging applied work in computer science and economics

- A “small unit” in Human Technology Futures
  At the time of this writing, a group of faculty are in the process of proposing a new department called Human Technology Futures (HTF) within CDSS (see https://htf.berkeley.edu/ for more information on the intellectual focus of this group of faculty). The graduate research area of this potential unit exists as designated emphasis (DEs) in Science and Technology Studies (STS) and other DEs, such as New Media and Critical Theory.

Committee Consultation

- **Benjamin Hermalin**, Executive Vice Chancellor and Provost - Oct. 5, 2022
- CDSS College Formation Advisory Committee - Oct. 7, 2022
- **Jennifer Tour Chayes**, Associate Provost, Computing, Data Science, and Society and Dean, School of Information - Oct. 12, 2022
- College of Engineering Chair’s Meeting - Oct. 13, 2022
- Center for Computational Biology Executive Committee - Oct. 4, 2022
- **Elizabeth Purdom**, Director of the Center for Computational Biology - Nov. 30, 2022
- **Giles Hooker**, Professor, Statistics - Dec. 7, 2022 & Jan. 7, 2023
- **Cathryn Carson**, Professor, History - Jan. 25, 2023
- **Bin Yu**, Professor, Statistics - Feb. 1, 2023
- **Scott Shenker**, Professor, Computer Science - Feb. 1, 2023
- **Jitendra Malik**, Professor, Computer Science - Feb. 1, 2023

Committee Membership & Staff Support

- **Ron Cohen**, Chemistry (Chair)

- **Adrian Aguilera**, Social Welfare
- **David Bamman**, School of Information
- **Josh Bloom**, Astronomy
- **David Harding**, Sociology and D-Lab
- **Haiyan Huang**, Statistics
- **Jennifer Listgarten**, EECS and CCB
- **Ziad Obermeyer**, Public Health
- **Claire Tomlin**, EECS
- **Aaron Streets**, Bioengineering
- **Tim Tangherlini**, Scandinavian and Folklore

- **Sumali Tuchrello**, Office of the Executive Vice Chancellor and Provost

Appendix A: Proposal for a "small unit" in Cultural Analytics
By: Tim Tangherlini, Professor, Scandinavian and Foklore

Possible small departments in CDSS with a strong data science orientation

1. Culture Analytics (alternatively: Cultural Analytics)

What is it?

Culture Analytics simultaneously studies the dynamics of culturally informed interactions between people, and the cultural expressive forms that result from these interactions, and it does so at scales hitherto unimaginable. Researchers in the discipline, often working collaboratively, aim to identify, document, and integrate concepts, methods and tools that provide an intellectually and ethically sound approach to the study of cultures across time and across space, leveraging the enormous gains made in the past decade in computation and machine-actionable cultural archives, from libraries and museum collections to the born-digital cultural expressions of billions of people on the internet. A macroscopic approach to cultural problems that allows a researcher to move from the microscale of close reading, up through various mesoscales of analysis, and onto the macroscale of distant reading, is a hallmark of the discipline. The fundamental need for intellectual engagement across disciplines that traditionally have little interaction is a hallmark of Culture Analytics and ensures that data-driven studies of culture are both responsible and meaningful.

Culture Analytics is, by definition, a collaborative, translational data science that explores culture and cultural interaction as a multi-scale / multi-resolution phenomenon, an approach that is not common in the Humanities and qualitative Social Sciences. Consequently, the concept of a small department with an intellectual focus on this area makes a great deal of sense. As we have discussed in numerous meetings, potential future hires may not be appealing as a "stand-alone" hire for either a Humanities or Social Sciences department (e.g. Film and Television) or a traditional CS or Statistics department, yet their work at the intersection of data science and film/television might provide important new insights into our understanding of production and reception. As noted below, there is a considerable "critical mass" of tenured faculty who, through joint appointments, could provide mentorship, guidance, and the appropriate institutional home for early career faculty, while also creating rich networks of affiliation to "south campus" departments. Equally importantly, a small department in CDSS focused on culture analytics would provide an attractive home for potential graduate students eager to work on cultural phenomena without the constraints of a traditional department in the Humanities (e.g. in Scandinavian, all students must have near-native competence in a Nordic language and take graduate level Old Norse), which often precludes student applications where the goal is to work on a data-driven analysis of cultural phenomena writ large. Similarly, a small department in CDSS for culture analytics would be a more appropriate home for various grants focused on the data-driven analysis of culture.

Challenges

A main concern of Culture Analytics researchers is how best to formalize the validity of cultural hypotheses and models obtained via the analysis of organic data either collected passively or collected for other purposes. This challenge is particularly crucial given the enormous amount of effort directed at the analysis of cultural phenomena from sources such as the internet, where controls are not put in place by design. While researchers in the Social and Behavioral Sciences are used to generating data from experiments that are carefully designed, and while Humanities researchers rely on rich historical archives of textual and material representations of past cultural activity, these scholarly traditions may need to be updated in the context of the explosion of digital resources that reflect cultural processes. Although some people are, and perhaps rightly so, skeptical about any claims made from large-scale organic cultural data, it is precisely this type of data that fuels a great deal of investigation in the data-driven study of culture as currently practiced. Consequently, there is room for considerable formal work to establish metrics for evaluating when a data set can be trusted. We consider this one of the central open challenges for Culture Analytics. The ethics of cultural data analysis is also a central concern of Culture Analytics. The data-driven analysis of culture can lead to the discovery of both socially constructive and socially destructive patterns. Similarly, there is a strong possibility that individuals' privacy as well as safety can be at risk. Recognition of these risks, while also recognizing the massive social and cultural benefits that can derive from cultural analysis at scale is crucial to the success of the discipline.

2. Networks and Complexity Sciences

I throw this out there in the context of a series of extremely attractive and successful programs at institutions such as Indiana University (CNetS with scholars such as Santo Fortunato, Johan Bollen, Fil Menczer, YY Ahn, Katy Börner), Univ Vermont (Complex Systems and Data Science with scholars such as Peter Dodds, Chris Danforth, Laurent Hebert-Dufresne), Santa Fe Institute, Stanford (Center for Computational Social Science), UC Davis (Complexity Sciences Center), Univ Michigan (Center for the Study of Complex Systems--Mark Newman), and Northeastern (Center for Interdisciplinary Research on Complex Systems; Network Science Institute with David Lazer, Alessandro Vespignani, Tina Eliassi-Rad, Alberto Barabaszi).

This could be a very dynamic locus for the study of complex phenomena, including epidemiology, the rise of political movements, urban development. It also intersects a huge number of areas, including somewhat underrepresented areas such as data visualization.

Appendix B: Proposal for a "small unit" in Data Science + Material Science
By: Haiyan Huang, Professor, Statistics

Possible small unit (e.g., a graduate group or an augmented graduate group): Data Science + Material Science
(Haiyan Huang draft 1/10/23)

This unit would focus on the development and application of modern computational methods alone or in conjunction with experimental techniques to discover new materials and investigate existing (inorganic, organic and hybrid) materials. As computational and statistical techniques advance, modern computational technologies (such as machine learning, artificial intelligence and high-dimensional statistical analysis) hold new promises in understanding various properties and phenomena of materials and achieving designing and making better materials for society. In some cases, a computational approach may become the only way to handle materials under extreme and hostile conditions that can never be reached in a laboratory such as in the presence of toxic substances. Computational approaches can also help develop more accurate guidance or decisions in terms of material design beyond what traditional experimental approaches or human intelligence could offer. [Here is an example. Synthetic random heteropolymers (RHPs), consisting of a predefined set of monomers, offer an approach toward the design of protein-like materials. These RHPs, if designed appropriately, can mimic protein behavior and function. However, despite the fact that RHPs can serve as great biofunctional materials, designing RHPs with desired function is challenging. Traditional RHP designs are largely empirical and depend on time-intensive lab screenings over various monomer compositions and chain lengths, leading to a limited design space for RHPs. As such, effective computational tools are in demand to efficiently guide RHP design.] That is to say, in today's data-driven world, deep expertise in computing and statistics, combined with expertise in materials and other relevant domain disciplines, will bring unique perspectives on and contribute to addressing the most pressing questions facing materials scientists.

This proposed unit, by recruiting scholars (from both on and outside the campus) working on methodological approaches to big data and applications in material science and engineering, would help rapidly advance the cycle of knowledge about existing and new materials. The co-location in a common transdisciplinary academic unit of experts working at the interface of computing and statistics with applications in materials sciences, will facilitate the essential cross-fertilization between approaches to data-intensive research in the field, so that novel computational and statistical methods are developed with timely and high-impact questions directly in mind.

This proposed unit is closely aligned with the Bakar Institute of Digital Materials for the Planet

(BIDMaP), which was just established in September 2022 in the Division of Computing, Data Science, and Society (CDSS), with the aim of developing cost-efficient, easily deployable new materials that help limit and address the impact of climate change. [Narrative of BIDMap: Two classes of materials with immense potential are ultra porous materials known as metal-organic frameworks (MOFs) and covalent organic frameworks (COFs). However, realizing the promise of such materials requires the development of novel machine learning methods and scalable computing platforms. The advances in statistical and computational infrastructure will have broad applicability beyond BIDMaP to problems that include protein structure prediction and drug development. The gift from the Barbara and Gerson Bakar Foundation to launch BIDMaP includes four philanthropically-funded FTEs which will boost Berkeley's excellence in machine learning, scalable open-source computing platforms, chemistry, and materials science and make BIDMaP's vision a reality.]

The aligned visions of BIDMaP and the proposed unit and the gift to launch BIDMaP provide a great opportunity to both seed the new unit and connect with other departments on campus in a hub-and-spoke fashion. The philanthropically-funded FTEs (open-rank) could be split into half FTEs to be appointed in this new unit under consideration and completed with state-funded half FTEs to be appointed in partner departments such as Materials Science and Engineering, Chemistry, Physics, Statistics and EECS.

The proposed unit, with a DE (for Data Science undergraduates or PhD students in other programs) in computational material sciences, will also enhance the education at both undergraduate and graduate level in this field.

Appendix C: Proposal for a "small unit" in Data Science + Applied Disciplines
By: Department of Statistics Taskforce, 2022

The Statistics Department Taskforce has been asked to comment on proposals to create a number of small departments at the interface between data science and applied disciplines, with examples in physical sciences, materials science, cultural analytics, computational biology, etc.

We want to start by reiterating our support for developing new research initiatives in this space. The statistics department faculty have not weighed in on specific proposals for new small departments, but they generally support plans that involve creating graduate groups at the interface between data and applications. The faculty of statistics expressed interest in joining such groups, and willingness to help, for example, by serving on search committees.

Graduate groups are often formed with the expectation that they may ultimately become new departments; it is reasonable to expect this to be the destiny of these groups. The Statistics faculty perceived considerable risk of duplicated effort or competition if this leads to a new generalist department of data science. In contrast, more focused departments are more clearly distinct from Statistics and are more likely to complement rather than compete with our pedagogical offerings. As was the case for graduate groups, the Statistics Department is prepared to consider appropriate joint hires, bearing in mind our need to maintain strength in the core concerns of our discipline. The Statistics Department enthusiastically supports Computational Biology, for example.

While the Statistics faculty support initiatives that foster deeper involvement between data and applications, the taskforce is concerned that separating the faculty in these departments from each other may miss opportunities for synergies between fields; the process of using Artificial Intelligence to screen potential products for chemical properties is very similar to that used to screen genotypes of crops for productivity, and the tools used in cultural analytics may also be applied to medical notes. Reproducibility standards, hypothesis generation and uncertainty quantification for physical sciences reasonably apply to all application areas. The Statistics Department has for decades contributed foundational theory and methods to those areas. Current examples include the PCS frameworks being developed in Bin Yu's group; uncertainty quantification and hypothesis generation methods developed by Giles Hooker; methods for uncertainty quantification, inverse problems, and selective inference, and work on reproducibility in the philosophy of science and foundations of statistics by Philip Stark; and selective inference methods developed by Will Fithian, as individual examples.

Facilitating interactions among all faculty working in these areas will improve both the speed of research advancements and the attractiveness of Berkeley to the faculty we would like to hire. The Statistics Department faculty's preference for developing an internal unit is partly driven by the potential for these synergies. This could, naturally, also be accomplished using joint appointments between Statistics and other departments, although presumably, not all faculty would hold joint appointments. We believe it is important to develop structures that encourage cross-talk between all the faculty involved in data-intensive science, without adding substantial

friction, for instance, without requiring substantial additional effort by those faculty. Departmental membership has the advantage of naturally providing cross-pollination of perspectives, research and pedagogy through business and meetings that are generally thought of as part of membership expectations, as opposed to individual's research commitments. That said there is a substantial cost associated with service to multiple departments and added complications for merit and promotion reviews.

Ultimately, we support appointments that augment, rather than replace faculty expertise in the core of statistics. Adding a large number of cross-appointments with state-funded positions without accounting for these in Statistics' FTE floor risks compromising our ability to maintain our preeminent position in the field, in both research and teaching.

Appendix D: Proposal for a "small unit" in Data Science + Societal Scale Infrastructure Systems
By: Claire Tomlin, Professor, Electrical Engineering and Computer Science

Data Science for Societal Scale Infrastructure Systems

DRAFT – Jan 12 2023

Claire Tomlin

Proposed Vision: Starting from technological advances in wireless sensor networks to cyber-physical systems, big data analytics and machine learning have now made it possible to analyze the large amounts of data that networks of embedded sensors generate. The translation of such sensor information with decision-oriented data analytics into Societal Scale Infrastructure Systems (SSIS) is an exciting new frontier that researchers across many engineering domains are tackling. Four technology domains in this are:

1. Big Data analytics for cyber-physical systems. Large numbers of distributed sensors are generating real-time data, which need to be analyzed using provably correct algorithms to provide actionable information in real-time.

2. Integrating human decision-making into SSIS. Even with substantial machine intelligence "in the cloud", it is critical to keep the human decision-maker in the loop to enable preference input and safety intervention. This requires having cognitive models of human decision-making and the interaction of human decision-making with machine intelligence.

3. Mechanism design for monetizing the operations of SSIS. The most critical use of the analyzed data is to develop new services, sharing economies and mechanisms of resilience to faults, physical and cyber-attacks on the infrastructure.

4. Resilience: privacy and security. It is critical to provide users with utility-based privacy contracts, as well as resilient operations through cyber-attacks.

Berkeley is uniquely poised to work on the development of novel design techniques critical to the evolution of four key infrastructures for future smart cities: (a) urban mobility, (b) air transportation, (c) the electric power grid, and (d) the urban water system. All four feature big data, and moreover require real-time analytics of this data for critical decision-making; in all four, human decision-making is central to the system behavior, and thus designs that feature incentive and mechanism design are paramount; and all four are safety critical and use a scarce resource, so designed resilience is critical. Such research will enable urban mobility in which the energy footprint of individual travelers is minimized through personalized incentive schemes, an energy-efficient air transportation system that meets customer demand and is flexible enough to incorporate unmanned aircraft, an electric power grid that seamlessly incorporates heterogeneous distributed energy resources, and a secure and efficient water supply network. It represents an

exciting synthesis and melding of techniques from cyber-physical systems, machine learning, big data analytics, game theory and mechanism design, security, privacy, human-machine interface design, verification, and validation to develop design tools for critical urban infrastructures, to develop a new field of Societal Scale Infrastructure Systems (SSIS) Engineering. It involves the resolution of complex economic, cognitive, privacy, business and public policy issues, spanning several academic departments in Engineering, CDSS, Social Sciences, Business, Law and Public Policy.

Appendix E: Proposal for a "small unit" in Computational Biology
By: Aaron Streets, Professor, Bioengineering

The Center for Computational Biology (CCB)
A model interdisciplinary unit for CDSS and Data Science

Overview
Computational biology is a field that develops and implements computational techniques to analyze and interpret biological and biomedical data. It combines principles from computer science, statistics and biology to develop algorithms, models, and software tools for gaining insight from large biological datasets such as DNA sequences, protein structures, gene expression patterns, human and public health records, and biological data from non-human animals, microorganisms and viruses. Computational biology has become a field that is critically important for, but distinct from biology broadly and has led to many important discoveries in fields such as genomics, evolution, and human development and disease.

Background
Biology has been fundamentally transformed to a highly quantitative science, relying on advanced methods in computer science and statistics. Computational biology now takes center stage in the interpretation of the new wealth of biological data for improving our understanding of complex biological systems. The field consists of the development of computational and statistical methodologies specific to the problems found in the field of biology, particularly to handle data from emerging biotechnical advances and large increases in the scale and scope of genomic-related data. The importance of computational biology is recognized nationally with NSF, NIH and other funding agencies providing substantial research support in areas of computational biology; large private philanthropic efforts also specifically support research in computational biology, such as the Chan-Zuckerberg Initiative. It is the close and sustained collaboration of biologists, computer scientists, and statisticians that enables the success of large-scale and high-impact projects such as the international Human Genome Project and the Obama-launched BRAIN Initiative.

As a result, in the last 20 years, computational biology has been recognized as a distinct field of research, with separate journals and conferences dedicated to research in computational biology. Importantly, prominent research in this area is not dominated by faculty with degrees or affiliations in biology departments. Instead, such researchers represent a range of departments, both in the biological and computational sciences. This reflects the fact that computational biology is not a subfield of biology, but a truly interdisciplinary area of research requiring integration of both computational and biological expertise. This is not unlike other areas, such as Bioengineering or Neuroscience, where vibrant new academic fields of research developed from the need to integrate traditionally different domain knowledge areas. The co-location in a common academic unit of experts in the biological and computational sciences is essential to ensure cross-fertilization between these fields, so that the latest computational and statistical methods are brought to bear soundly and rigorously to address timely biological questions. The

silo model, where computationally-oriented scientists are hired in biology (or other domain application) departments, has clearly shown its limits. Indeed, in the more general context of data science, institutions worldwide are reorganizing to create transdisciplinary academic units that bring together the computational and statistical foundations of data science with domain applications and implications. As such, the hiring of faculty through CCB fits squarely within the mission of Berkeley's proposed College of Computing, Data Science, and Society (CDSS).

Not surprisingly, many universities have invested heavily in gaining a foothold in computational biology, hiring new faculty, expanding their research programs, and developing undergraduate and graduate programs. Peer universities have aggressively expanded their computational biology faculties and research programs; Harvard and Stanford, for instance, experienced explosive growth around 2015, hiring 10 to 15 researchers to build up their programs. Most universities have formal organizations that bring together the faculty who concentrate in this interdisciplinary research. These organizations range in format from loosely formed groups to complete departments, with Harvard, Stanford, and CMU being examples of universities with full-fledged computational biology departments. Separate degree programs in computational biology are also common in most peer institutes, at either the graduate or undergraduate level. Included among these are programs at other top institutions, for example Stanford, Harvard, MIT, U of Chicago, Duke, and Penn.

Research Mission of CCB faculty

A critical role of CCB has been to initiate the hiring of prominent computational biologists to ensure UC Berkeley's preeminence in this important field. CCB both initiates FTE hires as well as embraces FTEs hired from departments across campus. These two types of hires are not mutually exclusive, but are essential to the success of the research and educational missions of both CCB and computational biology on campus more broadly. Hires initiated by other departments with expertise in computational biology continue to remain critical for the computational biology community, with two main sources being biology departments and computational departments, such as EECS and Statistics. Hires initiated by CCB enables UC Berkeley to hire the methodologically-focused computational biologists needed to maintain excellence in the field. Together these two types of CCB faculty hires span two broad research categories in CCB, methodological development and biological applications. A non-comprehensive sample of research areas in and between these categories is listed below:

Methodological areas of high-dimensional data analysis, multiple testing, and machine learning: Statistics, Biostatistics, Statistics, EECS.

Computational methods and mathematical and stochastic modeling of dynamic systems for systems biology: CCB, NTS, Statistics, EECS

High-throughput technologies and their use in understanding cellular regulation; BioEngineering, MCB, CBE

Evolutionary genomics: Integrative Biology, Statistics, CCB, Molecular Cell & Biology, Physics

Health-Related Genomics and Precision Medicine: CCB, EECS, Molecular Cell &

Biology, Public Health - Epidemiology,
Systems biology: Statistics, EPSM, CCB, NST

Educational Mission of CCB faculty

CCB hosts two graduate programs in computational biology, a PhD and a designated emphasis (DE). In addition to their educational value, our graduate programs serve an important role in supporting the research of computational biology at UC Berkeley. CCB PhD students have a mix of computational and biological training, and also have an interest in research specifically at the interface of computational and biological problems. Many CCB faculty, both FTEs and non-FTEs, look to CCB PhD students for research that requires these traits, not always easily found in the pool of PhD students in their other affiliated department. The CCB DE program connects with the campus even more widely, providing training in computational biology for PhD students across campus, including from research labs unconnected with CCB.

CCB FTEs support educational efforts in computational biology more widely across campus. The courses taught by CCB faculty also support the DE, which consists of students from many related disciplines on campus. The teaching efforts of CCB FTE faculty often consist of teaching courses cross-listed with other departments. Thus the teaching efforts for CCB helps other departments to maintain frequent offerings of computational biology courses appropriately targeted for their student bodies – offerings which might otherwise be limited due to the demand for teaching courses core to the department's degree programs.

CCB FTEs will also allow for the development of new modern programs for undergraduate instruction. In particular, UC Berkeley does not offer an undergraduate degree in computational biology, nor is there a large range of undergraduate courses offerings in computational biology. This is unlike several other peer institutes, some of which offer complete undergraduate degrees in computational biology such as Harvard, CMU, UCLA, MIT, and UCSC. Currently the Data Science major offers a domain emphasis in Computational Methods in Molecular and Genomic Biology, but the lack of available upper-divisional courses in computational biology limits the viability of this emphasis for many students. Similarly, biology departments on campus offer several related courses, but these are mostly introductory courses; more advanced coursework is limited. CCB runs a Berkeley Connect program in computational biology, and the success of this program shows that there is clearly interest in more undergraduate offerings in computational biology. CCB currently supports computational biology education across campus, through the teaching of CCB FTE faculty and TAS support of courses in computational biology taught by CCB faculty, both FTE and non-FTE. More advanced coursework appropriate for graduate students and advanced undergraduates, could also be addressed by CCB through expanded course offerings. This year CCB has tasked a committee to propose strategic improvements for undergraduate curriculum on campus, both in the short and long-term. In addition to filling an important need on campus, increased course offerings at both the undergraduate and graduate level will also provide opportunities for revenue generation, including a professional masters in computational biology, certificate programs, summer course offerings, and courses of interest to concurrent enrollment students.

Appendix F: Proposal for a "small unit" in Social Science Applications
By: David Harding, Professor, Sociology

SMALL DS UNIT: COMPUTATIONAL SOCIAL SCIENCE (Harding draft 1/5/23)

This unit would focus on the development and application of computational methods and tools for social science research and would interface with scholarship in both traditional social science disciplines and professional schools. The primary motivation for this unit is the increasing need for social science researchers to deploy computational methods, broadly defined, in the service of data collection and analysis, including the development of scholarly expertise in the translation and adaptation of approaches from computer science, statistics, and adjacent fields to social science problems and research questions. In some instances, the need for such tools reflects new topics of inquiry that have emerged from the digitization of social life and social interaction (e.g. social and political polarization on social media, online communities), while in other instances the sheer scale of social data from administrative databases and "digital exhaust" necessitates new methods for data collection, processing, and analysis. In still other instances, new tools can provide social scientists with increased capacity for analysis (e.g. the combination of close reading of text with computational tools for text analysis), new approaches to integrating artificial intelligence with policy analysis and applied decision-making (e.g. "prediction policy problems"), and new digital platforms for conducting social experiments.

Yet, despite the potential for computational social science advances in the study of social life and policy analysis, challenging intellectual problems remain, including but not limited to:

- Integrating computational methods with longstanding social science research methods and paradigms
- Adapting methods and approaches from other disciplines and fields to social science problems
- Developing new computational approaches to answering social science research questions
- Responsible and ethical use of computational methods and tools in social science research on human subjects

Addressing these problems requires the collective expertise of social scientists with domain-specific knowledge and long-standing social science research methodologies with statisticians, computer scientists, and others with an applied interest in social science and social policy problems.

The following are examples of the types of problems and approaches that would be included in research themes for a unit on computational social science:

- "Data Engineering:" Leveraging large language models and other AI tools to develop "human in the loop" methods for the cleaning and preparation of large-scale but unstructured data from administrative and other sources (e.g. Berkeley's EPIC lab).

- The “augmented social scientist:” Combining traditional methods of content analysis with natural language processing tools to extend the amount and complexity of data that can be analyzed.
- Text/images/video as data: Computational text analysis methods have greatly expanded the potential for extracting insights from spoken and written language. The same goes for image processing methods for pictures and videos.
- Causal inference: Harnessing machine learning algorithms to improve causal inference from both experimental and observational research designs, from adaptive experiments to effect heterogeneity.
- Online Experiments: Digital platforms (and online surveys) now provide social scientists with virtual labs in which to design experiments involving novel primes and complex social interactions.
- AI and decision-making: Deploying machine learning algorithms in the service of understanding and improving decision-making by practitioners and policymakers (e.g. work by Obermeyer, Mullainathan, and colleagues)
- Online interaction and communities: As social life has gradually moved onto online communities, platforms, and social media, social scientists who wish to study online social phenomena must be equipped to collect, analyze, and in some cases intervene in digital life.
- Mass collaboration using digital platforms: Distributed computing resources now mean we have the capability to engage large numbers of “citizen scientists” in social research in the service of data collection, coding, and analysis.
- Modeling complexity in social life: The advent of ever-larger data sets and the data science tools to process and analyze them means that social scientists can move beyond general patterns to understand complex interactions between variables and small subpopulations.
- Transparency and Reproducibility: Ever-larger datasets also raise questions about transparency and reproducibility, which are especially challenging when working with data on human subjects, where concerns about privacy and confidentiality are also paramount.
- App-based data collection: Mobile apps provide social scientists with new opportunities for both data collection and targeted interventions.

Appendix G: Proposal for a "small unit" in the Physical Sciences
By: Josh Bloom, Professor, Astronomy[1]

SMALL DS UNIT: **AI & The PHYSICAL SCIENCES (MPS+Chemistry)**

This unit is focused on the approaches, theory, and practices to imbue artificial intelligence/machine learning (AI/ML) into every aspect of the scientific method in the physical sciences: where AI/ML systems become more than just off-the-shelf tools in a large toolbox of analysis approaches operating on data already acquired but a first-class actor empowered to make data-taking decisions, a central figure in the generation of hypotheses and inferences, and the source of the creation and testing of new fundamental insights. Just as the best chess player in the world is now a combination of algorithms, computation, and people, the physical scientist of the future will be inextricably AI-enhanced.

Background: The physical sciences have already begun to embrace algorithms and toolkits developed in the computational sciences and throughout industry. However, there is a divergence of objectives: whereas industry-centric AI focuses on business outcomes and is perfectly content with retraining models for better performance as the user landscape changes, the scientific endeavor *seeks to probe the limits of generalization and extrapolation about a physical universe with fundamentally unchanging laws*. Physics aims to discover and understand what works when and where—and to explain why. In the best light, AI/ML tools adopted today in the physical sciences can be seen as essentially very flexible interpolators: trained with sufficient data they have shown remarkable success, at scale, in predicting properties of new data of similar structure. More pessimistically, AI/ML applied blindly to physical science data acts as a petri dish for automated p-hacking, yielding irreproducible outcomes that lack any scientific interpretation. Those in this unit seek to understand how we can build AI/ML systems that can generalize accurately beyond the regimes in which they are trained, exploiting known symmetries, physical laws, and other constraints and structure inherent in physical systems.

While national-level attention and funding appears to be flowing towards AI/ML in the physical sciences, it is largely doing so piecemeal, through the traditional domain-specific silos. This Program envisions an orthogonal agenda, bringing together scientists across domains with statistical and computational methodologists, to answer cross-cutting questions about the ultimate (and perhaps fundamental) utility of AI/ML in the service of science.

Research Themes of the Unit: This unit envisions a purpose-built, symbiotic relationship between the physical scientist of tomorrow, AI/ML systems, instrumentation, experimentation, and data analysis. The following research themes will serve to develop this relationship:

[1] This small DS unit proposal is adapted from a New Program Areas pre-proposal submission to the Sloan Foundation (2020). The Co-Is were: Joshua S. Bloom (Astronomy), Fernando Perez (Statistics), Laura Waller (EECS), Uros Seljak (Physics), and Philip B. Stark (Statistics).

1) Theoretical foundations and practical tools for Uncertainty Quantification (UQ)

We need to understand the reliability and replicability of predictions made by AI/ML systems when they are trained with data from the physical world. This effort should include the development of novel AI/ML architectures whose approximation and error behavior is consistent with that of our physical theories. Traditional methods for UQ for inverse problems in physics are based on either Bayesian posterior analysis, such as Monte Carlo Markov Chain methods, or frequentist UQ methods such as minimax optimization and resampling methods. Many AI/ML methods are entirely data-driven, and the data can be sparse and exhibit complicated patterns of missingness. This is even more essential when AI/ML methods are combined with physics-based models, and the physics parameters are the ultimate goal of the analysis. Current architectures, such as those used for vision and image classification, can suffer from catastrophic failures under small perturbations of the inputs, that make them unsuitable for modeling the physical universe.

2) Automated and Assistive Hypothesis Generation and Experimental Design

AI/ML systems should provide not only prediction and classification tools or accelerators for complex physical simulators, but assist the scientists in the generation of novel hypotheses and the design of experiments and data acquisition. Recent successes in computational imaging, for example, come from embracing the advantages of co-designing hardware, acquisition, and inference . Using end-to-end learning, one can employ AI/ML not just for reconstructing an image, but also for designing the imaging system hardware and data capture strategy. Can this approach be extended to other fields, such as high-energy physics, space-based astrophysics experiments, climate modeling, next-generation gravitational-wave experiments, or laboratory searches for dark matter, to name a few? This unit looks ahead to a time when AI/ML systems can assist the physical scientist in the creation of novel ideas for Discovery.

3) Physically Interpretable Modeling

Many predictive methods from AI/ML, including boosted decision trees and deep neural networks, frequently lead to models with high predictive power but with little to no interpretability. In order to leverage such methods effectively in scientific contexts, it is essential to have new tools that can extract theoretical insights, interpretation and explanatory power after having learned patterns and structure from simulations and real-world data. This will require novel developments beyond current work in model distillation and explanation in the ML literature, extending these ideas to the capture of physically-relevant knowledge from black-box systems.

4) Reproducibility

This unit will prize and elevate the development of both computational tools and community practices around the transparent, reproducible and robust use of AI/ML tools in scientific

research. This is an extension of current research in the statistical community regarding significance and p-hacking: in developing and training AI/ML systems, myriad opportunities exist for obscuring “researcher degrees-of-freedom.” Current work on selective inference should be extended to the workflows that practitioners follow when applying AI/ML techniques to scientific data analysis and prediction.

Appendix H
Notes on a "small unit" bridging applied work in computer science and economics
Ziad Obermeyer, Associate Professor, Public Health & Computational Precision Health

The tools and goals of computer science and economics are highly complementary. A small unit bridging these fields could accelerate the research of a number of Berkeley faculty members who are already at the forefront of applied research in this dynamic new area.

Of note, given the applied and multidisciplinary nature of this work—in mechanism design, health care, environmental policy, law, entrepreneurship, etc.—we believe this small unit would have a broad appeal well beyond just computer science and economics. The of gravity of the unit would likely be shared across faculty members in EECS, Economics, Statistics, and Berkeley's professional schools and applied departments—ARE, Bioengineering, Business, Information, Public Health, Public Policy, and more.

A non-exhaustive list of the areas in which Berkeley faculty are already working in this area, which could be part of the focus of this small unit, include:

1. Algorithmic approaches to the design and optimization of **markets**
2. Targeting interventions and measuring impact in **global development and public policy**
3. How algorithms interact with human **behavior**, in particular how to avoid training algorithms on human biases, and how to design algorithms to improve biased human decision-making
4. Measuring and countering **racial biases** (and other biases), both algorithmic and human
5. **Health**, where algorithmic advances can generate new insights from high-dimensional image and waveform data, and novel experimental designs; and must also work within the complex landscape of policies, incentives, and behaviors studied by health economics
6. **Environmental policy** and the management of planetary resources, including the measurement of resources and program evaluation using climate and satellite data
7. Using data science to bring empirical foundations to **macroeconomics**
8. Finding synergies between empirical work in **econometrics and computer science**, with particular emphasis on novel data collection and causal inference

Caveat: This is (obviously) not a full proposal, but rather the distillation of several conversations I've had with faculty members in this area over the past year.