Title
THE UC CLIOMETRIC HISTORY PROJECT AND FORMATTED OPTICAL CHARACTER RECOGNITION

Author
Bleemer, Zachary

Publication Date
2018-02-10

Research & Occasional Paper Series: CSHE.3.18

CSHE | Center for Studies in Higher Education

UNIVERSITY OF CALIFORNIA, BERKELEY

http://cshe.berkeley.edu/

The University of California@150*

THE UC CLIOMETRIC HISTORY PROJECT AND FORMATTED OPTICAL CHARACTER RECOGNITION

February 2018

Zachary Bleemer**
UC Berkeley

ABSTRACT

In what ways—and to what degree—have universities contributed to the long-run growth, health, economic mobility, and gender/ethnic equity of their students' communities and home states? The University of California ClioMetric History Project (UC-CHP), based at the Center for Studies in Higher Education, extends prior research on this question in two ways. First, we have developed a novel digitization protocol—formatted optical character recognition (fOCR)—which transforms scanned structured and semi-structured texts like university directories and catalogs into high-quality computer-readable databases. We use fOCR to produce annual databases of students (1890s to 1940s), faculty (1900 to present), course descriptions (1900 to present), and detailed budgets (1911-2012) for many California universities. Digitized student records, for example, illuminate the high proportion of 1900s university students who were female and from rural areas, as well as large family income differences between male and female students and between students at public and private universities. Second, UC-CHP is working to photograph, process with fOCR, and analyze restricted student administrative records to construct a comprehensive database of California university students and their enrollment behavior. This paper describes UC-CHP's methodology and provides technical documentation for the project, while also presenting examples of the range of data the project is exploring and prospects for future research.

Keywords: History of Higher Education, Big Data, Natural Language Processing, University of California

Longitudinal studies of higher education in the United States have long been limited by uneven archival and statistical record availability and by highly decentralized record collection (with each university campus maintaining its own archives). Indeed, these limitations are shared by all varieties of institutional history, from long-run analyses of primary and secondary schools to those of prisons and hospitals. As a result, until recent years scholars largely focused (by necessity) on macro-level narratives of administrative change reconstructed from contemporaneous statistical publications or on institution-specific narratives shaped by a relatively small number of participants' or authorities' written accounts.

Such records rarely include detailed personal or financial information (due to self-censorship, external censorship, and selective archiving), and statistical data rarely extended beyond the walls of any specific institution, challenging measurement of the longer-run impacts of schools and other organizations on participants like students, teachers, or patients.

---

* The University of California celebrates its 150 years since establishment in 1868 by an act of the California legislature. This article is the second in a series to be published by the Center for Studies in Higher Education related to the history of the University of California, and more broadly America's unique investment in public universities.

** Zachary Bleemer is UC-CHP Director and Lead Researcher at the Center for Studies in Higher Education and a PhD student in the Department of Economics, UC Berkeley: bleemer@berkeley.edu. Thanks to John Douglass and the UC-CHP research team as well as Renata Ewing, Mary Elings, the California Digital Library, and the HathiTrust Digital Library. Thanks as well to seminar participants at UC Berkeley and UC San Francisco and conference participants at the HathiTrust Research Center UnCamp 2018. Financial support from the UC Office of the President, the All-UC Group in Economic History, and DH@Berkeley is gratefully acknowledged. All errors that remain are my own.

A number of recent library-led digitization efforts, from the massive HathiTrust Digital Library to university-specific projects like the Digital Collections of Stanford University Libraries, have increased the availability of a third variety of historical institutional information by making millions of detailed administrative records publicly-available from any internet-enabled computer. With regard to universities, these records include annual student registers, course catalogs, budgets, and other documents that are too detailed for any individual to conceptualize and summarize independently, but which could feasibly be processed and analyzed using modern computational software.[1] However, the digital records' typical format—downloadable PDF images, sometimes accompanied by unformatted computer-generated transcriptions—does little to improve the records' accessibility for large-scale empirical analysis.

The University of California ClioMetric History Project (UC-CHP), occasioned by the 2018 sesquicentennial (150 year) anniversary of the UC system and based at UC Berkeley's Center for Studies in Higher Education, arises from a broad question about higher education: in what ways—and to what degree—have universities contributed to the long-run growth, health, economic mobility, and gender/ethnic equity of their students' communities and home states?

This paper describes UC-CHP's methodology and acts as a technical paper for the project, while also providing examples of the range of data the project is exploring, and prospects for future research.

The Project extends prior research in two ways. First, we have processed thousands of volumes of historical university records using a newly-developed digitization protocol—formatted optical character recognition (fOCR)—which transforms scanned structured texts like directories and catalogs into high-quality computer-readable databases convenient for statistical analysis.

Many of these databases are made publicly-available, including annual student enrollment records at most large California universities in the early 20th century (1890s-1940s), four universities' annual faculty directories for the entire 20th century, the same four universities' annual course offerings from 1900 to the present, and detailed budget records (including departmental allocations and, in some years, faculty and administration salaries) for the entire University of California system from 1911 to 2012.[2] These databases are then linked to each other—e.g. connecting courses to their faculty professors—and to supplemental datasets, like the 1940 US Census and various years' teacher and doctor licensing databases for the state of California.

Second, UC-CHP is working with university registrars across the state of California (and particularly across the University of California system) to photograph, process, and analyze historical student transcript records, which were maintained on paper 'hard cards' until the late 20th century. These records are processed with fOCR and then integrated with modern digital student records, producing a complete record of student identifying information, demographic characteristics, and course completion/evaluation back to the 1950s or earlier.[3]

The student records of one university campus (UC San Francisco) have been fully-processed, and the other nine University of California campuses are expected to be successively completed by the end of 2018. The resulting database is not available to outside researchers (due to privacy restrictions), but its derivatives can be leveraged to analyze both long-run student outcomes and changes in students' behavior across long-run transitions in university policy.[4]

While statistical analysis estimating the magnitude of universities' contributions to individual outcomes across the state of California is outside the scope of this introductory paper, a number of suggestive figures are presented illuminating the role of higher education, and of the public University of California system in particular, in 20th century California. Universities' tremendous enrollment and program expansion in the early 1900s—fueled in part by low tuition and the spread of college-educated role models across the state—and their mid-century transition towards engineering-oriented instruction were likely both central to California's century of extraordinary economic growth.

---

[1] Similarly-detailed administrative records have been directly recorded in computer-readable form since the 1980s, and recent research using those administrative records has proven extremely fruitful for understanding the individual and social ramifications of university-related decisions and policies. These studies' primary limitation is their necessary restriction to studies of contemporary university practices without the possibility of long-run evaluation. See, for example, Chetty *et al* (2017), Kirkeboen, Leuven, and Mogstad (2016), and Arcidiacono, Aucejo, and Hotz (2016).

[2] As of January 2018, the budgetary records are still being processed.

[3] Restricted and FERPA-protected student data are maintained and analyzed in accordance with university policy (including Institutional Review Board guidelines on human subjects research), contractual obligations, and relevant law.

[4] Examples include the University of California's shift away from tuition-free education and its trend towards increasing numbers of female faculty across academic disciplines.

About 15 percent of University of California medical students were female as early as 1900, and UC's gender- and ethnicity-blind admissions policies around the turn of the century were an early cornerstone of the state's equity identity. Complementing the interactive graphics available on UC-CHP's website, these figures depict nuanced aspects of California higher education that could not be observed without the detailed data collected by UC-CHP.[5]

Section 1 of this introductory paper presents an overview of the fOCR protocol, which can be flexibly adapted to a number of similar settings in higher education and other historically-oriented areas of study. Sections 2-5 provide additional details about the four initial databases constructed by UC-CHP, which cover students, faculty, courses, and budgetary allocations for a number of California universities, respectively. Section 6 discusses the student transcript records collected by UC-CHP, with a particular focus on fOCR data quality. Section 7 concludes.

1. Formatted Optical Character Recognition

Current best practices for the computational processing of unstructured text—like novels, newspaper articles, and printed speeches, in which words' conceptual content depends on syntax but not their spatial location on the page—are relatively successful.[6] A number of organizations and firms—such as the HathiTrust Digital library for books, LexisNexis for articles, and HeinOnline for speeches—have photographed tens of millions of text documents and used proprietary optical character recognition (OCR) software, like ABBYY FineReader or open source Tesseract, to produce computer-readable text versions of each document.

Though OCR software is error-prone and can only read typewritten text, processed corpuses are often massive enough to estimate statistical relationships despite typo noise. The growing discipline of natural language processing, which is expanding the frontiers of several academic disciplines in the humanities and social sciences, is distinguished by the development of statistical tools to study unstructured text.[7]

Best practices for generating computer-readable data from structured or semi-structured paper records are limited or costly to scale. Structured records use strict tabular organization to format the presented information, while semi-structured records like student directories and course catalogs rely on indentation, non-standard punctuation, vertical gaps, and other local spatial organization to classify text.

There are two available methods for digitizing structured text. The more popular is human transcription, which has produced digital versions of the 1790-1940 complete-count US Censuses, historical Surveys of Consumer Finances, and the 1915 Iowa State Census, and many other surveys and statistical summaries of varying magnitudes. While the results of human transcription are high-quality, the method is very costly and difficult to scale, rendering it infeasible for large or commercially-unviable databases. Alternatively, some OCR software has native tabular functionality that allows it to directly produce structured output using the scanned page's layout, but the results tend to be highly error-prone, requiring substantial human correction of errors. This human-augmented OCR output is only available when documents are typewritten and share a uniform tabular layout; while more time-efficient than human transcription, it remains costly for large databases. fOCR provides a cost-efficient and scale-able alternative to these transcription methods which produces similar-quality output without any human transcription.

We are only aware of one study, Brunet (2017), which produces computer-readable data from semi-structured text. Brunet processes four volumes of WWII-era financial records using human-corrected OCR output (purchased from a California data processing firm) and pattern-recognition software that identifies firm names, dollar amounts, and dates in each line of text, generating a large high-quality database used for economic analysis. fOCR builds on this method by replacing human OCR

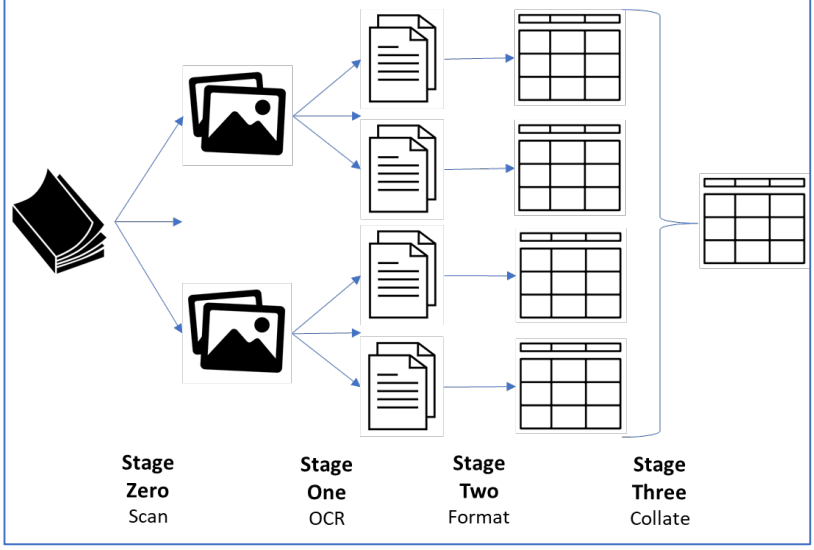
[5] Our website, uccliometric.org, hosts the three types of UC-CHP products. In addition to downloadable publicly-available data, the website contains a series of interactive graphics (produced in Tableau) that visualize the publicly-available and restricted databases collected by UC-CHP. The website also presents scholarly papers and policy/topic briefs produced by UC-CHP using its expansive data holdings. Our first working paper, which measures the effect of increasing number of female physics/chemistry teachers and doctors in early 20th century rural California communities on young women's college-going and major selection, was released last year (Bleemer 2016).

[6] I use the term 'unstructured' differently than it is used in either the economics or machine learning literatures when referring to text data. In economics, 'structure' refers to underlying statistical relationships governing textual meaning, while in machine learning 'structure' refers to contextual relationships between words (like syntax and topic ordering). In this paper, 'structure' refers to the purposeful spatial organization of words on a page, as in table formatting and the location of page numbers. All three 'structure' definitions are mutually independent.

[7] See, for example, the work of Matt Gentzkow (2016) in economics, Gerard Hoberg and Gordon Phillips (2016; forthcoming) in finance, and Ted Underwood (2015; 2016) in English.

correction with repeated computer 'readings' of the same text (and algorithmic selection of the highest-quality transcription of each entry), substantially increasing efficiency while maintaining quality.

Figure 1: Stages of Formatted Optical Character Recognition (fOCR)



Like standard OCR processing, formatted optical character recognition software begins with digital image files of each document to be processed, usually stored as a single PDF file for each document (which could be one page or a many-paged volume). Unlike OCR, fOCR benefits from multiple image files representing each document, either produced from multiple scans from the same document or scans from identical documents (e.g. two identical volumes owned by different libraries). Multiple scanning is a common (if unintended) result of contemporary digitization practices led by organizations like the HathiTrust Digital Library, which frequently scan complete libraries without sufficient regard for the existence of previous digital copies of each volume.[8] Each additional image file of a single document improves the quality of fOCR output, though a single uniformly-high-quality scan is often sufficient. In the diagram on the next page, (multiple) image scanning is referred to as 'Stage Zero' of fOCR, as it must be completed prior to beginning the fOCR protocol.

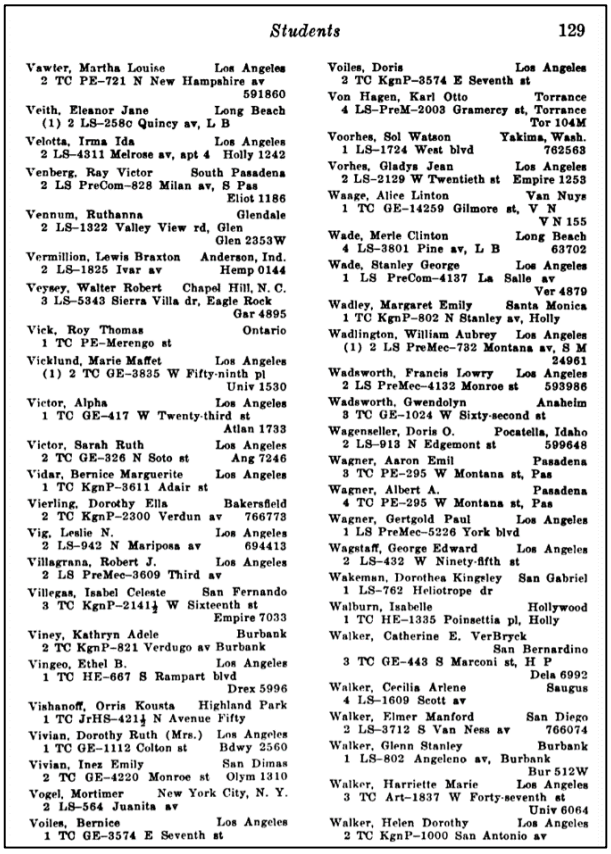
Formatted optical character recognition (fOCR) software proceeds in three stages, depicted graphically in Figure 1. In the first stage, each image of each page is processed by multiple OCR programs, resulting in a large number of similar transcriptions (each with its own typographical errors). For semi-structured documents, standard OCR output ('flat' text files, which preserve lineation but not spacing or other spatial organization) is sufficient, but structured documents require that the OCR program produce XML or HOCR files, which contain x- and y-coordinates for the top, bottom, right side, and left side of each character or word identified on the page. Most proprietary OCR programs accommodate both capabilities, as well as 'batch' interfaces enabling large-scale document processing; UC-CHP uses four programs: ABBYY FineReader 12, Adobe Acrobat Pro DC 2018, OmniPage Ultimate, and Tesseract 4.0 Alpha.

Each OCR program offers standard settings to identify a large number of fonts and languages in multiple columns or other formats; while several programs claim to recognize and format tables, they have not proven effective, and we disallow that functionality. As we discuss below, we have found OmniPage to be the highest-quality software for structured fOCR.

[8] See, for example, the five overlapping digital collections of University of California annual registers available from the HathiTrust Digital Library (records 007130126, 011249103, 007910193, 100024883, and 003915007). Two of the collections were scanned from volumes owned by UC libraries, while the other three were scanned from volumes at three other university libraries.

The second stage of fOCR implements document-specific pattern recognition algorithms to reorganize the text of each transcription into a tabular database. For semi-structured documents, these patterns are formalized syntactically as 'regular expressions', allowing the computer to 'read' a given page, identify all matches, identify the types of information inside each match, and organize that information tabularly.

For example, consider the 1925 UCLA student register, one page of which is displayed to the right. The following pattern could be used to identify each student record: a capitalized word on a new line, a comma, several capitalized words, another new line, a number, one or two words with multiple capital letters, and then a dash.



Students 129

Vawter, Martha Louise Los Angeles
2 TC PE–721 N New Hampshire av 591860

Veith, Eleanor Jane Long Beach
(1) 2 LS–258c Quincy av, L B

Velotta, Irma Ida Los Angeles
2 LS–4311 Melrose av, apt 4 Holly 1242

Venberg, Ray Victor South Pasadena
2 LS PreCom–828 Milan av, S Pas Eliot 1186

Vennum, Ruthanna Glendale
2 LS–1322 Valley View rd, Glen Glen 2353W

Vermillion, Lewis Braxton Anderson, Ind.
2 LS–1825 Ivar av Hemp 0144

Veysey, Walter Robert Chapel Hill, N. C.
3 LS–5343 Sierra Villa dr, Eagle Rock Gar 4895

Vick, Roy Thomas Ontario
1 TC PE–Merengo st

Vicklund, Marie Maffet Los Angeles
(1) 2 TC GE–3835 W Fifty-ninth pl Univ 1530

Victor, Alpha Los Angeles
1 TC GE–417 W Twenty-third st Atlan 1733

Victor, Sarah Ruth Los Angeles
2 TC GE–326 N Soto st Ang 7246

Vidar, Bernice Marguerite Los Angeles
1 TC KgnP–3611 Adair st

Vierling, Dorothy Ella Bakersfield
2 TC KgnP–2300 Verdun av 766773

Vig, Leslie N. Los Angeles
2 LS–942 N Mariposa av 694413

Villagrana, Robert J. Los Angeles
2 LS PreMec–3609 Third av

Villegas, Isabel Celeste San Fernando
3 TC KgnP–2141½ W Sixteenth st Empire 7033

Viney, Kathryn Adele Burbank
2 TC KgnP–821 Verdugo av Burbank

Vingeo, Ethel B. Los Angeles
1 TC HE–667 S Rampart blvd Drex 5996

Vishanoff, Orris Kousta Highland Park
1 TC JrHS–421½ N Avenue Fifty

Vivian, Dorothy Ruth (Mrs.) Los Angeles
1 TC GE–1112 Colton st Bdwy 2560

Vivian, Inez Emily San Dimas
2 TC GE–4220 Monroe st Olym 1310

Vogel, Mortimer New York City, N. Y.
2 LS–564 Juanita av

Voiles, Bernice Los Angeles
1 TC GE–3574 E Seventh st

Voiles, Doris Los Angeles
2 TC KgnP–3574 E Seventh st

Von Hagen, Karl Otto Torrance
4 LS–PreM–2003 Gramercy st, Torrance Tor 104M

Voorhes, Sol Watson Yakima, Wash.
1 LS–1724 West blvd 762563

Vorhes, Gladys Jean Los Angeles
2 LS–2129 W Twentieth st Empire 1253

Waage, Alice Linton Van Nuys
1 TC GE–14259 Gilmore st, V N V N 155

Wade, Merle Clinton Long Beach
4 LS–3801 Pine av, L B 63702

Wade, Stanley George Los Angeles
1 LS PreCom–4137 La Salle av Ver 4879

Wadley, Margaret Emily Santa Monica
1 TC KgnP–802 N Stanley av, Holly

Wadlington, William Aubrey Los Angeles
(1) 2 LS PreMec–732 Montana av, S M 24961

Wadsworth, Francis Lowry Los Angeles
2 LS PreMec–4132 Monroe st 593986

Wadsworth, Gwendolyn Anaheim
3 TC GE–1024 W Sixty-second st

Wagenseller, Doris O. Pocatella, Idaho
2 LS–913 N Edgemont st 599648

Wagner, Aaron Emil Pasadena
3 TC PE–295 W Montana st, Pas

Wagner, Albert A. Pasadena
4 TC PE–295 W Montana st, Pas

Wagner, Gertgold Paul Los Angeles
1 LS PreMec–5226 York blvd

Wagstaff, George Edward Los Angeles
2 LS–432 W Ninety-fifth st

Wakeman, Dorothea Kingsley San Gabriel
1 LS–762 Heliotrope dr

Walburn, Isabelle Hollywood
1 TC HE–1335 Poinsettia pl, Holly

Walker, Catherine E. VerBryck San Bernardino
3 TC GE–443 S Marconi st, H P Dela 6992

Walker, Cecilia Arlene Saugus
4 LS–1609 Scott av

Walker, Elmer Manford San Diego
2 LS–3712 S Van Ness av 766074

Walker, Glenn Stanley Burbank
1 LS–802 Angeleno av, Burbank Bur 512W

Walker, Harriette Marie Los Angeles
3 TC Art–1837 W Forty-seventh st Univ 6064

Walker, Helen Dorothy Los Angeles
2 TC KgnP–1000 San Antonio av

After matching this pattern, each determinant would be assigned to a different category of information: the word in front of the comma must be the last name, the following word is the first name, the number is the student's year in school, et cetera. Some words are indeterminant—like the second capitalized word after the comma, which could be either the student's middle name or the first word in their home town's name (spatial information like spacing is rarely preserved in flat OCR output)—and are assigned to multiple information types (appearing in both columns of the resulting table) for the time being.

Moreover, notice that the pattern listed above would fail in some cases—it would miss Catherine Walker because of the line break between her name and home town, as well as Karl Von Hagen due to his two-word last name—and would need to be generalized for each of these (and many other) exceptions.

The second stage of fOCR for structured documents organizes text using explicit spatial patterns rather than textual patterns. Text is assigned to different categories of information based on its location on the page. For example, a certain template of student transcript may always record students' names between 0.5 and 2 inches from the left edge of the page and between 1 and 1.3 inches from the top edge; the coordinates assigned to each character or word on the page can be rescaled into inches or another standard unit, matched to the spatial patterns, and (if the match is successful) pulled into 'Name' column of a table.[9] When the set of pages to be processed includes multiple tabular templates—e.g. a set of university transcripts that changed format in different years—a pre-processing step to identify each page's template is necessary (usually by matching the locations of standard words like "Name" or "University", which will differ on different templates).

For either semi-structured or structured documents, the second stage of fOCR results in a large table, one row per individual per transcription per document and one column for each category of information derived for that individual/transcription.[10] The third and final stage of fOCR discards repeat and lower-quality duplicate rows, maintaining only the best row of information for each individual by eliminating typos and other transcription errors.

[9] Spatial information derived from scanned documents can run into other problems as well. If the scanned documents were not identically-aligned on the scanner, then the locations of various tabular information may differ slightly from page to page; a pre-processing step can be added in which documents are aligned using 'fingerprint' features (say, fixing certain words like "Name" or "Major" to specific locations and measuring other coordinates in relation to those words). Scanned images could also be skewed or unintentionally-resized; similar pre-processing could also be used to return the page images to their original orientation and magnitude.

[10] Additional columns contain information that identifies the page from which the information was derived, the scanned copy of that page used for the transcription (if multiple images of each page were available), and the OCR software used to generate the transcription.

For semi-structured text, the easiest algorithm to implement is to rank the number of individuals identified on each transcription of each page, keeping only information from the transcription that included the most individuals (and discarding all other information from the final table). An improvement on this algorithm would be to algorithmically-identify high-quality information (e.g. students' year in school should be between 1 and 5, and students' home town should be an identifiable town, state, or country) and keep pages with the most so-estimated high-quality information.

For structured text, one could determine quality field by field, keeping only algorithmically-determined high-quality information; if the four transcriptions of a date are "/24/", "224461", "2/24/46", and "12/24/Ab", the third transcription can be kept and the rest discarded, with similar arbitrations for each field. Ultimately, the product is a database with one entry for each individual, with the quality of the resulting information greater than any OCR transcription could be alone.

The first and third stages of fOCR are very similar for all different kinds of documents, and a small set of OCR and quality-identification algorithms can be generally applied to many circumstances. Stage two, however, requires specific tailoring for each document type, with new patterns developed to identify the various types of available information. Nevertheless, given that the primary alternative to algorithmic pattern recognition is human transcription, fOCR is highly efficient for large documents or corpuses. Moreover, as will be discussed below, the output quality of fOCR is high, and enables a large variety of historical statistical analysis that would otherwise be impossible or extremely time-consuming.

Below, we discuss five applications of fOCR to university records collected and organized by the UC ClioMetric History Project. The first four record types—student registers, faculty registers, course catalogs, and university budget allocations—are semi-structured, and both the original images and the resulting databases are publicly-available on UC-CHP's website. The last, student transcript records, are structured documents; while the resulting database is not externally-available, the data's structure allows for an extended discussion of OCR and data quality.

2. University Student Database

Until the 1940s, most California universities published annual Registers containing, in addition to other institutional information, complete directories of enrolled undergraduate and graduate students (an example page is displayed above). In the late 19th century, these lists often only included students' names and home towns, but over the years more information was added, including fields of study and local addresses (in order for students at the time to find and contact each other).

UC-CHP has obtained public access and fully processed these semi-structured student records for eight universities, in order of the number of available student-year records: UC Berkeley (1893-1946), UCLA (1921-1946), Stanford University (1893-1946), the University of Southern California (1904-1924), the health-oriented professional schools that became UCSF (1893-1946), the California Institute of Technology (1912-1946), Mills College (1910-1940), and Hastings College of the Law (1893-1946), excluding a small number of years in each case. Detailed descriptions of the data available for each of these universities is available in Appendix 1.

2.1 Data Construction

UC-CHP obtained between one and five scanned copies of each processed student register. In many cases, the available PDF documents had already been transcribed using OCR software (e.g. by the HathiTrust Digital Library), and we used these transcriptions for Stage One of fOCR rather than generating our own; where no transcription was available, we produced one using Tesseract 4.0. Common typos in all transcriptions were 'corrected' using a standard cleaning algorithm.[11] In Stage Two of

[11] It is well-known that all available proprietary optical character recognition (OCR) software is typo-prone. UCCHP has developed its own multi-stage cleaning algorithm, with several general components as well as components that are specific to certain kinds of documents (like student registers or course catalogs). The following general corrections, which are constructed as a series of regular expressions, are implemented on all OCR transcriptions prior to stage two of fOCR:

1. Text encoding is switched to UTF-8 to avoid character encoding errors.
2. Line breaks are replaced with *þ* for tractability.
3. Lines that end with hyphens followed by lines that begin with either a lower-case letter, a number, or two capital letters are combined, omitting the hyphen.
4. The euro sign is replaced by *E*; the dollar sign by an *S*; brackets by *l*. All other unusual punctuation marks are replaced with the letter *q*, our chosen marker for unusual characters. Five spaces in a row are also replaced by *q*; any remaining sets of spaces are reduced to a single space. Tabs are replaced by commas.
5. Cyrillic characters are replaced by their closest Roman counterpart. Accents are removed from all letters.
6. Empty lines, or lines with no more than two characters, are omitted.

fOCR, regular expressions identified individual student records on each transcribed page, and in Stage Three the transcription of each page with the largest number of identified students is selected. If more than one transcription has the maximum number of identified students, then whichever of those transcriptions is missing the fewest home towns (the most-frequently-missing component) is selected.

Following selection, we match the non-selected records to the selected transcription by name and fill in any missing or incomplete information (home town, field of study, year in school) from the non-selected records before discarding them.[12] Finally, an additional field-specific cleaning algorithm removes numeric characters from names, corrects errors in common fields of study, and corrects other typos.[13]

Next, we use algorithmic tools to identify each student's gender and home town location. We infer gender by matching students' first names with Social Security Administration records, which include all names assigned to at least five newborn American children of one gender in each year since 1880.[14] Spelling errors and name changes challenge town identification; we match towns to a comprehensive list of populated areas in California compiled from Wikipedia (along with the names of other states and nations to identify out-of-state university students), allowing for small spelling changes and frequently-occurring errors.[15]

Each town is matched to geographic coordinates using Wikipedia's GeoHack database, and the coordinates are then matched to California counties (the borders of which do not change through the 20th century) using data from the National Historical Geographic Information System. Finally, we algorithmically link student records across years into a panel using combinations of parts of their first and last names, home towns, fields of study, and year of study, generating a unique ID number for each student (though, due to remaining typos and over-matching, the quality of cross-year IDs is lower than that of the raw fOCR records).[16]

---

7. There are a large number of common letter-specific corrections. For example: .\ is replaced by *A*, *'I'* is replaced by *T*, *VV* by *W*, *I-I* by *H*, et cetera.
8. Words that only appear as the first in a multi-word last name (like *De La*, *Mac*, *St.*, *Van*, *Von* etc.) have the following space deleted, rendering the last name to a single word (with multiple capital letters).

Student registers tend to have similar formats, and additional cleaning algorithms were included to preserve those formats:

1. Parentheses almost never appear; *(* was replaced with *C*.
2. Many lines begin with a single-digit number; those which begin instead with single-digit character (with punctuation) are replaced by the most-similar-looking number. For example, *g* is replaced by 2, *S* by *3*, and *U* and < by *4*.
3. Certain major or college abbreviations are very common, while similar character-combinations are not; the latter are transformed into the former, e.g. *L8* and *LC* to *LS* (Letters and Sciences) and *Uec* to *Mec* (Mechanical Engineering).

[12] The name-to-name match across these records is imperfect due to typographical errors. Individuals who appear in the non-selected transcription but not the selected transcription are nevertheless discarded; we assume that those individuals *do* appear in the selected transcription, but the match fails due to typographical imperfection.

[13] Rather than listing students' year in school, Stanford reports their total number of completed credits; these are transformed into years in school based on the number of credits earned per year (30 until 1917; 45 thereafter). USC does the same prior to 1909, with 30 credits per year. Names are converted to title case and split into first, middle, and last.

[14] These records include more than 2,000 names for each gender in each year. We begin by matching students to SSA records from 20 years earlier (with a floor at 1880, the first year in which the records are available), and then continue matching using subsequent and previous years. Individuals with names that are less than 10 times more likely for one gender, or those whose names do not match SSA records, are not assigned to a gender (about 3 percent). Data available at https://www.ssa.gov/oact/babynames/limits.html.

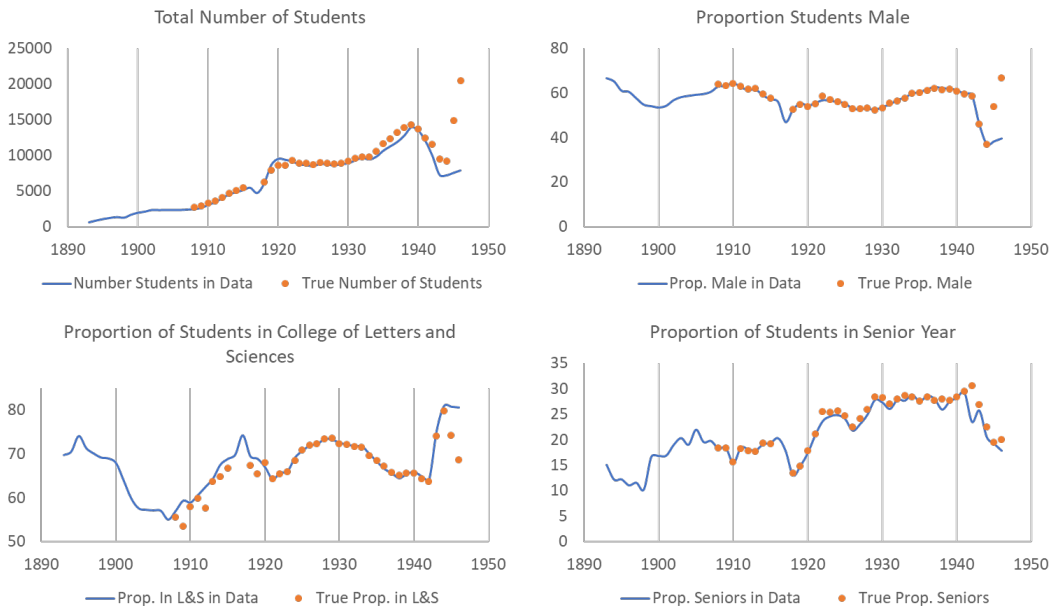
[15] In particular, a match is successful if the recorded town name is no more than one generalized Levenshein distance away from the true town name, omitting spaces (see Levenshtein 1966). Populated areas in California include those mentioned in one of the following Wikipedia categories: unincorporated communities in California (overall and by county); incorporated cities and towns in California; Census-designated places in California (overall and by county); former Census-designated places in California; former populated places in California; neighborhoods in Los Angeles; neighborhoods in Newport Beach, California; neighborhoods in San Diego; or populated coastal places in California. Some towns have previous or alternative names not mentioned on these lists, which are added manually. When multiple towns with the same name occur (which is very rare), students are assumed to come from the town with the larger 2010 population.

[16] Students are linked in three stages. First, students who attend the same campus in consecutive years and years-in-school with similar first and last names are linked, where two names are similar in either of the following cases:

- Their first names are the same and the first four letters of their last names are the same
- Their last names are the same and the first four letters of their first names are the same

Commonly-misconstrued letters are unified for these similarity measures: for example, the names *Lily* and *Liiv* are considered identical, since *l*'s and *i*'s, as well as *v*'s and *y*'s, are often transcribed for each other by OCR programs. Second, individuals who satisfy either of those name conditions and list the same home town, but may have a one-year or one-year-in-school gap, are matched. Finally, students who follow the name rule and appear within 2 years or years-in-school are matched. These ID links are used to correct the degree and location fields; if a linked student's degree changes in one year and then changes back the next, the change is assumed to be a typo and the data corrected.

Figure 2: UC Berkeley Student Statistics, 1908-1946



2.2 Data Quality

In order to measure the quality of the fOCR-generated student records, we have collected a number of annual statistics describing UC Berkeley from the 1908-1946 annual Statistical Summaries published by the university (omitting 1916 and 1917, when the Summaries were not published). Figure 2 shows that the statistics estimated from the UC-CHP database closely match the number of students and the gender, college, and seniority distributions of students throughout those years, with the true number of students only slightly exceeding the number of students captured in the database.

The largest exception is around World War Two, when UC Berkeley's student registers appear to substantially undercount male students outside of the College of Letters and Sciences, likely student veterans whose late or unusual registration prohibited them from inclusion in the 1946 Register.[17] There is also some noise in the number of students in Letters and Sciences in the 1910s, when Berkeley was forming the College of Letters and Sciences from the Colleges of Letters, Social Sciences, and Natural Sciences (completed in 1915). In general, Figure 2 provides evidence of the high-quality nature of the student database produced by fOCR.

2.3 Data Visualization

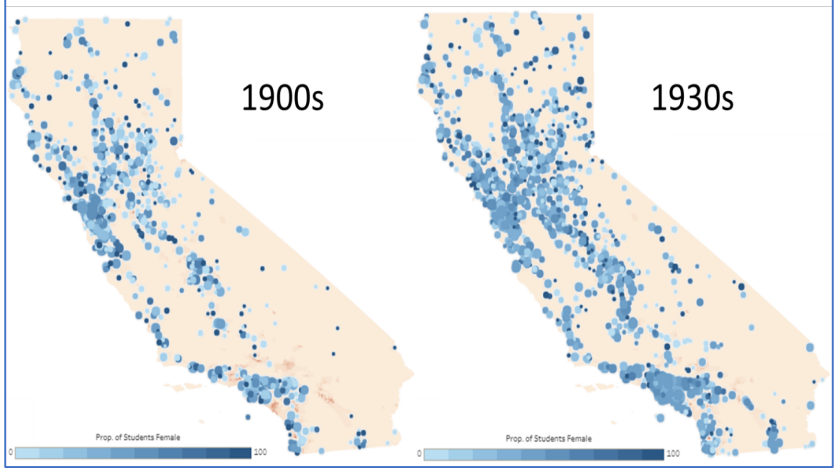
The UC-CHP student database constitutes one of only two available large databases of early-20th-century university students in the United States (the other being the 1915 Iowa Census; Goldin and Katz 2010). While the UC-CHP student database includes only limited information about each student—their name, home town, major or 'college' of attendance, and years of matriculation and exit/graduation—even just these fields, combined with empirical learning algorithms and links to other databases, provide substantial insight into college-going at the time.

Figure 3 maps the geographic and gender distribution of California university students. Even in the first decade of the 20th century, the small proportion of young adults who chose to enroll at a large California university (fewer than 5 percent) arrived

[17] Indeed, universities likely ceased publishing annual Registers in 1946 due to the influx of students attending college under the GI Bill, increasing student enrollment so much as to make the directories' publication infeasible.

from every corner of the state. Even using a restrictive definition of rural areas, almost 20 percent of public university students—and almost 30 percent of Stanford students—came from rural California communities.[18]

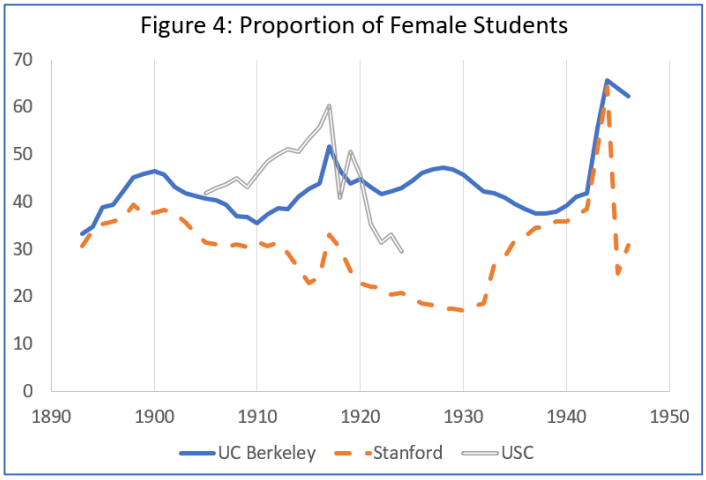
Figure 3: Home Towns of California University Students, 1900s and 1930s



This was true for both male and female matriculants; Figure 3 shows that there were many rural communities that sent either mostly-male or mostly-female students to university, along with those that sent similar proportions of each (darker circles represent towns that sent a larger proportion of female students to university). By the 1930s, when nearly 10 percent of American youths were attending college, the number of rural students remained large despite shrinking proportionally (to 15 percent at the University of California), especially due to the rise of the Los Angeles region (both in terms of population growth and university presence, with UCLA expanding to a four-year university in 1921).

The proportion of female students at the University of California stayed largely unchanged through the first decades of the 20th century, fluctuating between 35 and 50 percent of students. Figure 4 shows that USC had a somewhat-higher proportion of female students through most of the 1900s and 1910s, but that the proportion of women at Stanford slowly declined between 1905—when Jane Stanford, who founded Stanford University with her husband Leland, died and left a will stipulating that no more than 500 Stanford could enroll no more than 500 female students at a time, despite the universities' ensuing growth—and 1933, when the will's stipulation was overturned in court.

Figure 4: Proportion of Female Students



The proportion of female students also jumped during the first and second World Wars as young men left campus to join the military, falling thereafter as male veterans returned to school.

Finally, in order to measure long-run outcomes for early 20th-century college-goers, UC-CHP has linked its student database to the 1940 US Census; a working paper describing our findings, along with linking code and match IDs, will be released in summer 2018. In the meantime, we employ Olivetti and Paserman's (2015) "name score" technique to estimate the class background of California college-goers.

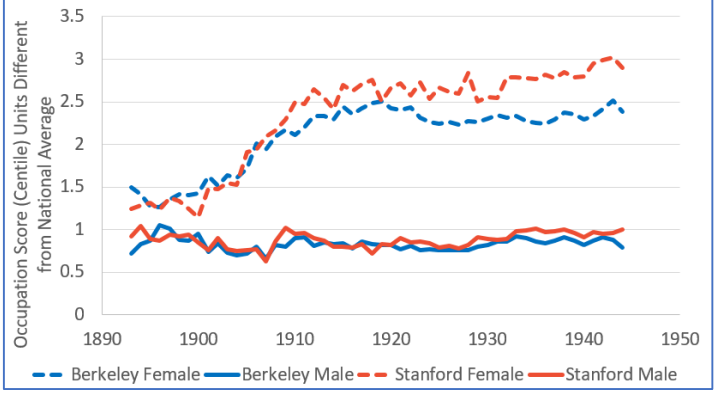
Using one-percent extracts of the 1890-1940 Censuses, we estimate each student's parental income by their first name assignment (which Olivetti and Paserman show to explain about six percent of variation in parental income in this period), estimating so-called "name scores" for each student.[19]

[18] A town is defined as rural if it is unincorporated or in the bottom half of populations of CA incorporated towns (ranging from 1,600 in 1900 to 3,100 in 1930). Town populations are interpolated from high-order polynomial fits to decennial Census counts and biannual population estimates by municipal clerks made for tax purposes, weighing the two sources equally. See the Annual Reports of Financial Transactions of Municipalities and Counties of California.

[19] Name scores are estimated using the one percent Census extracts available from IPUMS. Students are matched to children under the age of 15 in any US Census such that they were born between 15 and 25 years before the year the student appeared in the register. Because pre-1940 Censuses did not elicit

Figure 5 displays annual average name scores for male and female students at UC Berkeley and Stanford University from 1893 to 1946. The y-axis is measured in name score units above the national average for similar-age Americans; name scores have a standard deviation of 3 across the university student population.[20] All estimated values are above 0, indicating that university students tended to come from wealthier-than-average families, and female students tended to come from substantially-wealthier families (almost one standard deviation above average) than their male counterparts, especially after 1910.

Figure 5: Average 'Name Scores' of UC Berkeley and Stanford Students



The substantial increase in name scores of female students between 1900 and 1910 remains unexplained, and is an interesting subject for future research. UC Berkeley and Stanford students appear to have had similar income backgrounds until the 1910s, after which Stanford students tended to come from wealthier families; not coincidentally, Stanford only began charging tuition in 1916 (and UC Berkeley charged no tuition throughout this period).

Botany (A Graduate Group)

Cell and Developmental Biology (A Graduate Group)

Chemistry

Cantonese

See Asian American Studies

The Major Programs

A.B. Major Requirements:

B.S. Major Requirements:

157

3. University Faculty Database

Throughout the 19th and 20th centuries, most university course catalogs began each department's dedicated section with a list of the department's faculty members. These lists typically included the faculty member's rank, and many included information about each member's degrees and research interests. Cross-listed faculty were independently listed in multiple departments, sometimes with reference to their primary field.

As universities have phased out paper catalogs in favor of online course guides, these faculty lists have disappeared, an unintended casualty of the transition from print to computerized records; while department websites list current faculty members, public records of past years' faculty members are typically no longer maintained.

3.1 Data Construction

UC-CHP has fully-digitized faculty records for four universities—UC Berkeley, Stanford University, UCLA, and UC Davis—from 1900 until each school ceased publication or digitization of print course catalogs.[21] In each case, UC-

income information directly, parental income is itself estimated between 1 and 100 using 1950 occupation score centiles; higher-earning occupations correspond to a higher centile of earnings, and therefore a higher occupation score.

[20] Name score averages in the state of California are very similar to those across the US.

[21] A number of course catalogs are omitted, largely due to the volume's exclusion from digitized collections (likely because it has been lost in the respective institution's archives). These include the 1905, 1907, 1959, 1960, 1972, and 1982 UC Berkeley volumes, the 1902-1904, 1921, 1961, and 1967 Stanford volumes, and the 1943 UCLA volume. Moreover, UC Berkeley and UCLA began publishing their course catalogs every *two* school years in 1996, though UCLA returned to an annual publication in 2007; faculty records are only available biannually in those years. Our procedure for 'smoothing' faculty members' presence at each campus, filling in missing years with faculty present in prior and subsequent years, is discussed below.

CHP obtained a single scanned version of each universities' catalog directly from the respective universities' library's web site, where they are publicly available.

Each image file was processed into flat text files using three OCR programs: OmniPage Ultimate, ABBYY FineReader 12, and Tesseract 4.0. Moreover, in most cases the libraries had already used unnamed OCR software to render the PDFs computer-readable, so we extracted their processed text into a text file using the *pdfminer* Python package and included it as a fourth transcription of the document.[22] As above, common typos in all transcriptions were 'corrected' using both standard and course-catalog-specific algorithms.[23]

For course catalogs, Stage 2 of fOCR begins by splitting each transcription of the course catalog into academic departments by recognizing the patterns common to the header text announcing each new department (which often includes the department's name, chair, and office location), conducted in part by matching the text to a complete list of department names used at any campus in any year.[24]

Within each department, the text listing that department's professors is identified using name and degree patterns, and the remaining text (like major requirements and course descriptions) is discarded.

Professors are distinguished using line breaks and punctuation, depending on the catalog format, and are similarly matched to preceding and subsequent personal information like earned degrees and academic rank (assistant professor, lecturer, etc.). This information is pulled into a table for each department, and these tables are concatenated with an additional column identifying the department.

Finally, as above, additional field-specific cleaning algorithms remove numerical characters from names and fix other obvious typos. Additional available information like research interests and more specific professional information (e.g. endowed chairs held by faculty) is stored as a character string in an overflow column.[25]

In Stage 3 of fOCR, faculty names from each of the four transcriptions of each school-year's course catalog are merged into a single table (with a new column specifying a transcription code).[26] Similar names in the same academic department are merged together, and the most-common identified name of each individual faculty member (across years and transcriptions) is preserved, with typos and repetitions dropped.[27] Names which appear in only a single transcription are omitted; most are typos. ID numbers are assigned to each faculty member.

---

[22] We use the *pdfminer.six* program *pdf2txt.py* implemented in Python 2.7.13: https://github.com/pdfminer/pdfminer.six

[23] See footnote 11 for a description of the standard cleaning algorithm. Additional corrections for the course catalogs include:

1. Common headers and footers (like *COURSES OF INSTRUCTION*) are deleted.
2. Common degree names are corrected, e.g. *Pll,D,* is replaced by *Ph.D.*
3. Spaces are removed from parenthetical professorial titles, to ease professor identification.
4. Commas and line breaks are added to separate professors' names when they are erroneously missing, based on common name patterns like middle initials as well as surrounding information like titles and academic degrees.

[24] A total of six course catalog templates were used in the 334 school-years processed. The list of department names was produced manually, and includes more than 700 department names used by the four schools (and some other universities) since 1900.

[25] Until the 1950s, UC Davis was an agricultural campus of UC Berkeley. In that period, Davis professors were listed in the UC Berkeley course catalog with a note indicating their presence at UC Davis. Such professors are included as *Davis*, not Berkeley, professors in the UC-CHP database, for which reason UC Davis faculty records predate their course records (which did not easily distinguish between the two campuses).

[26] Clearly-erroneous faculty names are omitted (e.g. those with names identical to that of their academic department, or those with first or last names more than 20 characters long).

[27] The matching algorithm for names was developed by extensive trial and error. Names are stratified by Area, which agglomerates similar departments. Each pair of names is assigned points for achieving a variety of measures of similarity, based on Levensheim distances between names (abbreviated below; e.g. if two names are 2L, they are at most two Levensheim distance units away from each other, excluding spaces and punctuation; see Levenshtein 1966):

- Last names are 2L (or 1L if fewer than 6 characters), or one of the last names is 2L from the other middle name (or 0L if the last name is fewer than 6 characters) with non-missing middle name: *3 points*
- Last names are 3L: *1 points*
- Middle names are 1L and non-missing: *2 points*
- First names are 2L (or 0L if one-letter initial) and non-missing, or one first name is 2L from the other middle name (or 0L if first name is one-letter initial): *2 points*

Pairs of names with at least 5 points are matched. Matching is transitive, even if the ultimately-matched pairs are not all within 5 points. Resulting matches are all combined into a single ID; the most-frequently-appearing name among them is chosen as the unified name across all instances.

If a faculty member's name is missing for a 'gap' of fewer than 7 years, we assume the disappearance to be erroneous (resulting from interstitial typos or other errors), and 'fill in' the missing years using the nearest available year's record; we add a column denoting such interpolated records.[28] Interpolated faculty records make up about 9.5 percent of records from years with observed course catalogs.[29]

Finally, as with the student records above, we infer faculty members' gender using Social Security Administration records. Departments are grouped into Areas and General Areas (Humanities, Social Sciences, Natural Sciences, Engineering, and Professional) by our discretion, and degrees and academic rank are standardized and encoded.

Due to the more complex format of course catalogs relative to student registers—professors' names are not always at the beginning of a line of text, and catalogs' multiple columns frequently confuse OCR transcription and (in particular) lineation—the faculty database is somewhat lower-quality than the student database. Summary statistics and a large number of ocular comparisons between the database and the original volumes suggest that the data are relatively reliable for aggregate descriptive figures and statistical analysis, but future iterations of the UC-CHP data will continue to improve these records' quality, in addition to expanding the sample of available universities.

3.2 Data Visualization

Between 1900 and 2010, the number of ladder-ranked faculty positions at UC Berkeley increased by more than 30 times, while the number of departments offering courses increased nearly threefold. Stanford grew at a rate only slightly less than Berkeley's, while UCLA and Davis became independent campuses rivaling the growth of their older peers. Many of the 20th century's most disruptive social and technological movements manifest themselves in universities' faculty composition; the two charts below provide a birds-eye view of these trends.

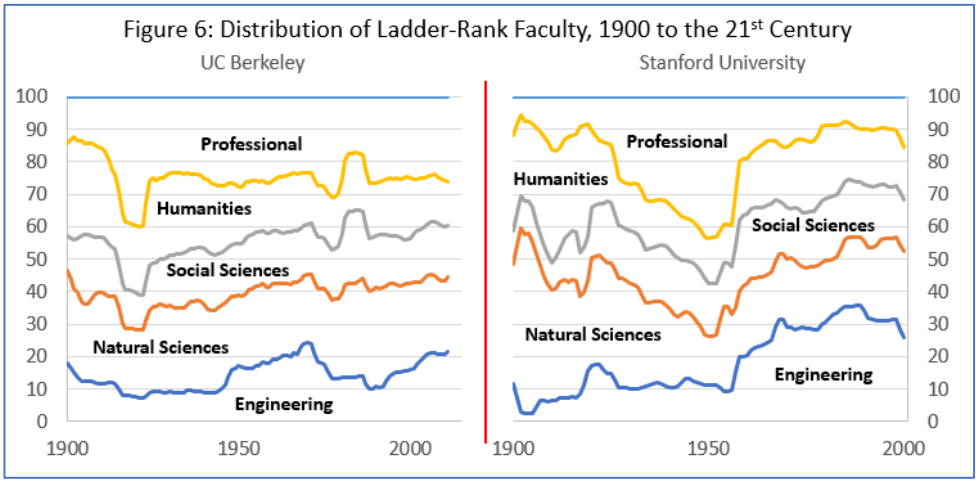


Figure 6 shows the distribution of faculty across five broad areas of the university, double-counting faculty cross-listed in multiple departments. Both universities sharply expanded their engineering faculties in the mid-century, but with different timing: Berkeley grew the 1940s and '50s with the rise of the military-industrial complex and the space race, while Stanford's engineering programs' sharpest growth began in the late 1950s, simultaneous with William Shockley's move to Mountain View and the rise of Silicon Valley. Stanford's faculty following the Second World War was dominated by its medical school, which Berkeley did not

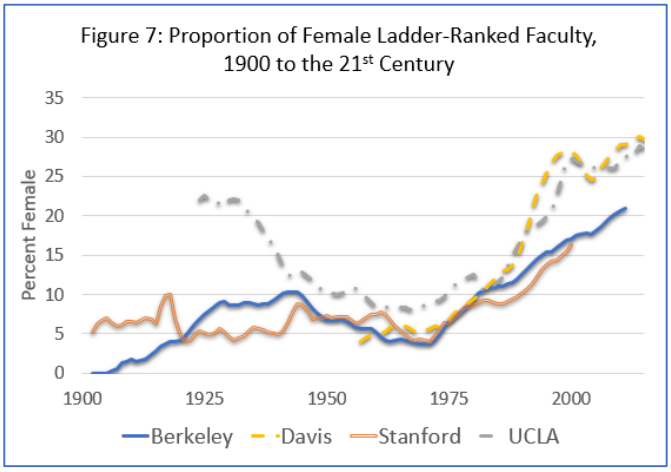
[28] For example, assume that John K. Smith appears as an assistant professor in the Economics Department in 1924 and 1925, is not present in the 1926 or 1927 records, and then appears as an assistant professor in the Economics Department in 1928 and many years thereafter. It's possible that Prof. Smith left the school for two years, but it is more likely that John Smith was an economics professor throughout the period but failed to be observed in 1926-1927 due to typos or other processing errors. As a result, we generate records for John Smith in 1926 and 1927 as an economics professor, listing him as an assistant professor in 1926 and an associate professor in 1927.

[29] When a year's course catalog is missing, either due to non-digitization or non-publication (as in the case of universities that only published course catalogs every two years), faculty are filled in using prior and subsequent volumes. As a result, including these interpolated faculty member, a total of 21 percent of the faculty records are interpolated. Interpolation is lowest-quality at the start and end of each schools' available course catalogs, since there are few earlier or later records from which missing records can be interpolated.

have (the University of California's medical school was located in San Francisco) but which required a large number of professionally-oriented professors.

Both schools have presided over similar declines in humanities faculty relative to the rest of the university; humanities professors made up between 25 and 30 percent of each university's faculty in 1900, but were fewer than 20 percent of each faculty by 1950 and in recent years have been closer to 15 percent of the faculty. Scientists and engineers have made up about forty percent of UC Berkeley's faculty since 1900, while at Stanford they have made up half the faculty since the 1960s.

Figure 7 displays the gender distribution of ladder-rank faculty of four California universities since 1900. UCLA opened as a four-year university in 1922 but maintained its adjacent originally-two-year teacher's college, dramatically increasing its proportion of female faculty (many of whom were primary and secondary school teachers who taught UCLA evening courses) until the college was phased out over the following decades. UC Berkeley and Stanford have historically hired approximately similar proportions of female faculty, though Berkeley has been leading Stanford in its acquisition of female faculty in recent decades.



The University of California campuses did not consistently have even ten percent female faculty until the early 1980s, nor Stanford until the 1990s, but each university has steadily increased its number of female faculty since those years; today, the faculties of UCLA and UC Davis are nearly 30 percent female.

4. University Course Database

How has humanistic and technical instruction—and, indeed, the very material available for collegiate instruction—evolved over the past century? While databases like Columbia University's Open Syllabus Project display the tremendous variety of courses available to American students, the large majority of their available course descriptions are from the 21st century, limited (as elsewhere) by the availability of computerized records.[30]

Though course catalogs contain more limited information than complete syllabi, they typically provide a one-paragraph course description along with a wealth of additional meta-data: whether the course was taught in a given year, when it was taught, and (in many cases) who taught it. These data are combined into the UC-CHP course database.

4.1 Data Construction

As with the faculty database described above, the course database spans four universities—UC Berkeley, Stanford University, UCLA, and UC Davis—from 1900 until each school ceased publication or digitization of its print course catalogs. The first stage of fOCR is shared by the two databases, as is the beginning of the second stage: each scanned file is transcribed by three OCR programs (OmniPage, ABBYY, and Tesseract) and a fourth transcription is scraped from the downloaded file; the transcriptions are corrected using both standard and course-catalog-specific cleaning algorithms; and each transcription is broken into academic departments using text patterns that identify department name headers.[31]

Faculty and summary information is discarded, leaving only course descriptions. Each course is identified using start-of-line pattern-recognition, as each begins with a course number; all subsequent information, until the following course number, is

[30] See https://opensyllabusproject.org/.

[31] In a small number of cases (e.g. UC Berkeley engineering in the 1970s and 1980s) the order of information in course catalogs was shuffled: a number of departments' faculty-members were listed sequentially, followed by the courses taught in each of the departments. A special algorithmic module attempts to rejoin each departments' faculty with their respective courses, but errors likely remain, leaving some courses omitted altogether and others wrongly-assigned to different departments.

associated with the most-recently-identified number.[32] Each course is assigned a row in a new table, with fields for the full course description as well as its number, name, department, and school.

Next, courses are matched to faculty. Each course description is searched for the last names of each faculty member known to teach in that school-department-year. Up to two faculty members can be assigned to each course; their ID numbers are added as new fields in the database. Some courses are listed as being taught by "Staff", and other courses list no faculty-member at all; for those courses, the instructor fields are left empty. Courses taught be faculty members in other departments are left unmatched due to the likelihood of common last names across departments.

Course descriptions are searched for textual indicators of which semester they're taught; those with any semester information are listed as being 'taught' in a separate field, so long as they don't explicitly state that they are not taught in that year. Some courses state that they are only taught in even or odd years; the 'taught' field accurately reflects these designations. As with faculty, departments are grouped into Areas and General Areas by our discretion.

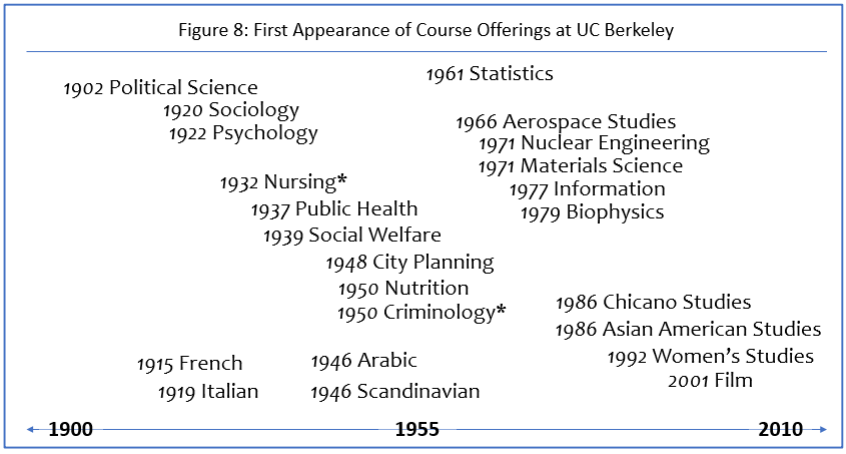
While a future iteration of the UC-CHP database may attempt to further 'clean' the transcribed course descriptions, which could improve the quality of natural language processing, currently the description is presented to include the description proper along with other subsequent text (like professors' names).

Finally, in Stage Three of fOCR, the transcription of each school-department-year with the largest number of identified courses is preserved, with the remaining three transcriptions discarded.

4.2 Data Visualization

As with trends in university faculty hiring and retention, the arrival of new departments and courses often mirrors large-scale social trends outside the university. Figure 8 displays the year in which various academic departments first offered courses at UC Berkeley, as reflected in the course database.

Figure 8: First Appearance of Course Offerings at UC Berkeley



These departments appear in waves: social sciences in the first decades of the 20th century, urban-oriented fields like public health and city planning in the 1930s and 1940s, new science and engineering disciplines like nuclear engineering and biophysics in the 1970s, and a variety of new ethnic studies programs in the last decades of the 20th century.

With two exceptions—nursing and criminology—these departments continue to offer courses up to the present day; strong path dependence makes the disappearance of new academic disciplines rare.

5. University of California Budget Database

The University of California has been publishing detailed annual budgets since at least 1911, the first year in which the publication was bound and preserved in UC Berkeley's archival Bancroft Library. In the early years, these budgets provided detail on the university's income (by investment, fee, and grant) and expenses (including individual salaries and expenditures by department) for each UC campus. As the university system grew, this level of detail became impractical; starting in 1953, expense information was aggregated to the department level by category (academic salary, staff salary, equipment, etc.), with tables displaying the number of positions or full-time-equivalent employees in each category. The document's title changed in the

[32] In some cases, an extraordinary amount of text can be associated with a single course number. This most frequently occurs when the department-finding algorithm fails to identify an entire department, leaving its entire faculty list and summary text to be assigned to the previous department's final course. In such cases, the final course and all subsequent courses are omitted completely (since the subsequent courses likely belong to a different department); they will be replaced by courses from another transcription that accurately identified the second department.

1960s to "Departmental Allocations" from "Budget for Current Operations", though the included information remained largely unchanged. The complete publication had reached 2 volumes and 2,500 pages by the time the university decided to replace print publication with a PDF document (stored on a CD-ROM); in 2012, when the document was more than 3,500 pages long, publication was ceased altogether.

UC-CHP has worked with the Bancroft Library to produce digital scans of all available detailed UC budget volumes, from 1911 to 2012.[33] With the exception of 1983-1987, which are missing from the Bancroft's collection, those volumes are now publicly available on UC-CHP's website.[34] The volumes which had been previously-digitized (2003-2012) are computer-readable, but the older volumes are still being processed using the fOCR protocol.

Despite their limited computer-readability, this release provides scholars with a massive and exceptional source of detailed information about a public university system's financial status throughout the 20th and 21st centuries. In a future release, a complete century-long database of UC budget records will be processed and made publicly available, further expanding the research accessibility of these records.

6. University Student Transcript Database

While student registers and other public records provide general information about students and their university attendance, they fail to present more detailed information about students' educational abilities, decisions, or outcomes. University administrative student records, especially the course transcripts maintained by Registrars offices, provide far more information. In addition to listing the courses students' enrolled in each term (and the grades they received), student transcript records often include additional individual identifying information (like birth dates and social security numbers) that can be used to link students to long-run outcomes; additional geographic information like location of birth, home address, and residency status; and secondary education records like high school, graduation year, and collegiate-level coursework completed.

This supplemental information enables additional insight into longitudinal changes in California college-going over the 20th century—documenting changes in field preference and the effect of faculty role models, showing trends in grade inflation and college preparedness, and more—but also requires rigorous safeguards to protect the records' privacy.

Universities typically maintain student records in multiple formats. Records were usually kept on hand-written cards until the 1940s, when most universities replaced handwriting with typewriting. These typewritten cards were used until the late 1970s, when many schools switched to computerized record-keeping. Paper cards were subsequently converted to microfiche in the 1990s and early 2000s, and more recently have been converted to PDF or other digital image files (sometimes from the original paper records, and sometimes indirectly from the microfiche records).

UC-CHP is working with a number of universities to collate these diverse records into a single comprehensive student database—at least going back to the 1940s, due to the absence of high-quality OCR software for hand-written records—and by the end of the year our team hopes to have completed such a database for the 10 University of California campuses. At present, we have completed student course transcript digitization for UC San Francisco and are nearing completion for UC Berkeley and UC Santa Cruz.

6.1 Data Construction

Prior to fOCR, in some cases we have organized teams of research assistants who physically scan student transcripts that had not previously been digitized. Scanned documents stored as .tif, .png, .bit, or .jpg files are converted to PDFs and rotated to their upright orientation. Some universities maintain multiple digital scans of each student record, all of which are used for fOCR.

Because student transcript records are fully-structured documents, the first stage of fOCR requires that each transcription denote the x- and y-coordinates of every letter or word on each image.[35] Each scanned image of each record is processed by four OCR programs—OmniPage Ultimate, ABBYY 12, Adobe DC Pro 2018, and Tesseract 4.0—but instead of producing flat text files, the

[33] The Bancroft Library was also missing the 2008 CD-ROM-stored Departmental Allocations; it was obtained from UC Davis's Shields Library. Digitization was coordinated in a partnership with the HathiTrust Digital Library, which has also made the scanned volumes available.
[34] See http://uccliometric.org/budget/.
[35] For a definition and discussion of 'fully-structured' documents, see Section 1 above.

settings of each program are altered to generate either XML files (OmniPage), hOCR files (Tesseract), or text-searchable PDF files (ABBYY and Tesseract).[36] Cleaning algorithms transform each transcription into a table, with rows for each word and columns containing the word's coordinates.

Stage Two of fOCR begins by identifying the template used by each student record. Universities have changed their student records' formatting many times over the past century—for example, UC San Francisco used 16 transcript templates between the 1890s and 1975, including some variation across its colleges—but each template contains unique textual formatting that enables algorithmic identification. Some templates print the word 'Absence' in the bottom-right-hand corner; others include 'MEMORANDA' on the back; still others, like a nursing transcript from the 1920s, print "TUBERCULOSIS" and "Personality" on the right-hand side of their only page. The presence of these 'fingerprint' words on at least one of the record's transcriptions identifies each record's template.[37]

Next, the spatial offset of each record is measured. Because record scans may not always be similarly-centered on a given page, the fixed locations of fingerprint words can be compared to their estimated locations on the scan to slightly adjust the coordinate locations of each word in the transcription. These fingerprint words can also be used to rescale documents that appear larger or smaller in a given scan, or to straighten skewed scans of records.

Following coordinate adjustment, the observed templates of each transcription are in near-perfect alignment, allowing them to be transformed into a new tabular format (one row per transcription of each record, front and back) using template-specific spatial algorithms. For example, a certain undergraduate record template may record student last names between 0.7 and 1.6 inches from the left and between 0.6 and 0.9 inches from the top of the front page of a document; any words from the transcription within that region are combined and added to the 'Last Name' column of the new table.[38] The space on each record designated for a list of completed courses and grades is sliced into rows based on word location, and each row is added to a second new table, this one with one row per course (along with an identification number linking the courses to the student named at the top of the record).[39]

Finally, in Stage Three of fOCR, we implement an evaluation algorithm that estimates the quality of each field of information from each transcription of each record, selecting a transcription with the most high-quality fields and correcting low-quality fields with higher-quality information from other transcriptions.[40] For example, if the four transcriptions of the "Home" field (which records students' home town) are "Bakrsfield1", "Bakersfield", " ", and "Bakersfield", the final table would record the student's home town as "Bakersfield", a town in California. All other information is discarded. As above, gender is imputed from Social Security Administration records and home towns are matched to geographic coordinates using the GeoHack database.

---

[36] Text-searchable PDF files are transformed into XML files using the Python *pdfminer* tool. The XML text formats differ from program to program, but the information they contain is roughly the same. Each transcription begins by stating the size of the piece of paper in its preferred unit of measurement. For every word on the page (where 'word' is defined as spatially-consecutive characters the program deems inseparable), the document lists four coordinates: the distance from the left side of the page to the left side of the word, the distance from the page's left side to the word's right side, and the distance from the top of the page to the top and bottom of the word. We discard the second and fourth coordinates, defining each word's location by the coordinate of its upper-left-hand corner, and normalize the measurements by the size of the page.

[37] In the rare case that a template cannot be identified, the record is omitted, with a random selection of omitted records inspected by hand; most are low-quality scans and must be discarded from analysis. Sometimes different transcriptions of the same record identify different templates due to erroneous transcription; the majority-identified template, or that chosen by the more reliable OCR transcription, is selected.

[38] Cleaning algorithms delete unnecessary fingerprint words (like the word 'Name', which often appears in the range of a student's last name), isolate and standardize names and dates, convert symbols (like an unchecked box in front of the word "Resident", which is often transcribed as a capital O or 0) into standard language ('Non-Resident'), and make a number of additional adjustments. A number of field-specific cleaning algorithms are also implemented: typos in majors and cities are corrected, degrees awarded by the school itself and by schools previously attended are standardized, et cetera. Some schools include students' subsequent names (as after marriage) recorded *above* students' original names; such names are identified and stored in a separate column.

[39] Each semester's courses begin with a header naming the semester; headers are identified by the presence of standard vocabulary ("Spring") or years, removed from the course list, and assigned to all subsequent courses until the next semester header. Some schools' course records contain aggregate course credit sums and other extraneous information, which is deleted. Some schools' records contain hand-written grades despite their typewritten course records; grade information is lower-quality, though still often available, in these cases. The course table is also cleaned using field-specific algorithms, correcting frequent typos in grades and course numbers among many other adjustments.

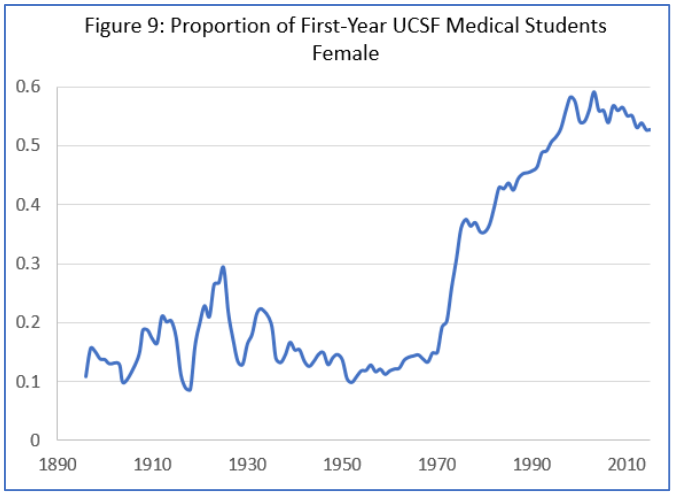
[40] For example, high-quality dates have valid months, days, and years; high-quality locations include either cities in California, states, or countries; and high-quality grades are single capital letters like A, D, or P (but not M or Z). Some transcript records include certain information (like names and majors) on both the front and the back; this information is all collated, with the highest-quality text selected. If multiple 'high-quality' transcriptions exist (as often happens for names, which only require a single word beginning with a capital letter), the information that appears most frequently in the various transcriptions is selected; if there is a tie, then the highest-quality and most-frequent information from the highest-quality transcription (averaged across the other fields) is selected.

Finally, we integrate the information gleaned from historical student records with schools' contemporary computerized student records, producing a single database covering all enrolled students from the transition to typewritten text (usually in the late 1940s) to the present day. The integrated records typically include:

- Student name, location of birth, birth date, student ID, home address, gender, California residency status
- Previous institutions, degrees, year degrees earned
- Degree enrolled, major, start date, end date, whether graduated
- *For each student-course*: Term, subject/department, course name, course number, grade

6.2 Data Visualization

The first school to provide administrative student record access to UC-CHP was the UC San Francisco, in summer 2017. Though UC San Francisco has only been an independent university campus since the 1960s, several of the colleges that comprise UCSF have their roots in the late 19th century.



By combining the relevant portion of the student directory database—which includes medical students since 1893, dental students since 1917 (when dentistry became a four-year degree), and pharmacy and nursing students since 1940, all up until 1946—UCSF's paper administrative student records—which were typewritten from about 1948 to 1975—and UCSF's current computerized database—which contains all student records 1975 to present—we produce a comprehensive 125-year record of UCSF's student population.

An interactive graphic displaying these records is available on our website, uccliometric.org; Figure 9 shows the proportion of UCSF medical students who were female throughout the period; while non-negligible numbers of women were earning medical degrees every year at the turn of the 20th century, the crucial increase in the female student population occurring during the 1970s Women's Movement.

Table 1: Baseline accuracy of fOCR Student Transcript Records

Condition	Proportion 1970s Records	Proportion All Typewritten Records
Has Identifiable Field of Study	97.1%	91.9%
Has Identifiable Start Year	95.9%	92.6%
Has Identifiable Home Town	95.0%	85.9%
Same Name across Transcriptions	91.8%	84.6%
First Name Matches Gender Records*	94.3%	90.0%

6.3 Data Quality

Because UCSF maintained no computerized records of students whose administrative records were preserved on paper 'hard card' transcripts, there is no way to directly compare records produced using the fOCR protocol to a baseline human-generated database to test the accuracy of the former. Instead, we measure the records' internal consistency by presenting the proportion of records which satisfy a variety of estimable conditions. These are shown in Table 1.

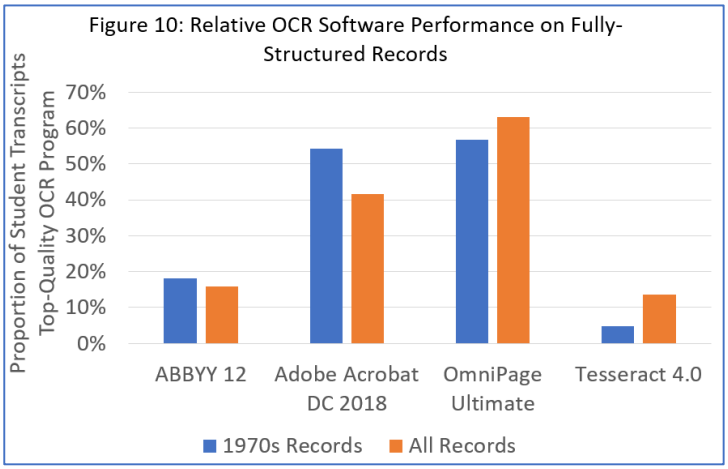
The corpus of UCSF student transcripts contains about 22,600 type-written cards, spanning from the 1940s until 1975.[41] Of those, about 11,700 are recorded on the two most popular student transcript templates, which were used across all of UCSF's colleges from the late 1960s until paper records were discontinued. These later records, being less aged and better-formatted, generate higher-quality fOCR transcription than the older records. Table 1 shows estimates of transcription quality for both these more recent records and for the inclusive set of all typewritten transcripts.

The first three rows of Table 1 show whether fOCR is able to isolate plausible information of various kinds for UCSF students, though we cannot directly test whether the isolated information is accurate. UCSF offers a limited number of fields of study to its students; fOCR was able to produce text identifying one of those fields for 97.1 percent of 1970s students and 91.9 percent overall. It isolated students' start year—four-digit numbers between the 1940s and 1975—for 95.9 (92.6) percent of 1970s (all) students, and home towns that could be matched to actual California cities, states, or countries for 95.0 (85.9) percent of students. Similar statistics were obtained for birth years and birth towns.

The fourth row reports the proportion of individuals whose first and last names were identically-transcribed in at least 60 percent of transcriptions. Consider, for instance, a student transcript on which the student's name appears on both the front and back of the card. In the first stage of fOCR, that name will be transcribed eight times, twice each by four OCR programs. Such a record would satisfy the stated condition if at least five of those eight transcriptions provided identical output for the student's first name, and five provided identical output for the last name.

This is a relatively high bar--many names are likely to be reported correctly despite failing this condition—but it's also feasible that some small number of names is identically *but incorrectly* read by different OCR programs. About 92 (85) percent of 1970s (all) student records satisfy this condition, and it's likely that most of the remaining records nevertheless have correct name transcriptions despite several OCR programs wrongly transcribing them.

Finally, the fifth row shows the proportion of student records with first names that could be matched to a gender using contemporary Social Security Administration records. As described above, first names are matched to an SSA database the contains every name assigned to (at least five) children born in the United States around the year of the students' births; students with names at least 10 times more likely to be given to babies of one gender relative to the other are assigned that gender. Since some names are androgynous or highly-uncommon (the most-common among these students being 'Marion', which appears on 34 records), it is impossible to match 100 percent of students to genders; in the student register database, about 2.5 percent of student names cannot be assigned genders. The UCSF database includes 5.7 (10.0) percent of student records that are either androgynous or cannot be matched to the SSA database.



Turning to the quality of individual OCR transcriptions, Figure 10 displays the relative performance of each of the four OCR programs used to fOCR UCSF's fully-structured student transcript records. In particular, the figure shows the percent of student records (1970s or overall, as defined above) for which each OCR program produced the highest-quality transcription, where quality is defined by the number of fields of information in each transcription which satisfy field-specific quality guidelines (e.g. years must be four digits starting with *19*).

[41] Typewritten cards are identified by their template, as identified in Stage Two of fOCR. Document templates that are used for both handwritten and typewritten records—namely, those used during the transition from handwritten to typewritten cards in the late 1940s—are considered handwritten for the purposes of this exercise. UCSF used a total of 16 record templates between the late 19th century and 1975.

The proportions sum to more than 100 percent because many records have two equally-high-quality transcriptions. We find that OmniPage produces the highest-quality transcriptions of the student records, especially for the full set of early and late records (many of which are more difficult to read); among the more recent records, the performance of OmniPage is similar to that of Adobe Acrobat DC, the next-best software. ABBYY comes in a distant third, and Tesseract—the only freeware OCR software used—a near-negligible fourth. OmniPage is the least-expensive of the three proprietary OCR programs used in our analysis, but also the most successful.

7. Conclusion

Institutional histories of higher education have long been limited by the inaccessibility of detailed historical university records. Using the new formatted optical character recognition (fOCR) protocol, the University of California ClioMetric History Project has generated a number of novel publicly-available databases reflecting the growth, diversification, and increasing egalitarianism of California universities since the late 19th century, along with an expanding collection of reports and analysis summarizing these data—with a particular focus on economic mobility and gender/ethnic equality—for academics and policy-makers.

As we enter the 2018 sesquicentennial year of the University of California system, we hope that these new databases provide fodder for future research on the history of both Californian and American higher education as a whole.

The databases released in January 2018 are UC-CHP version 1.0. Over the next year, a number of additions are expected, including:

- fOCR detailed UC budget records (See section 5)
- ID-number links between the University Student Database and the full-count IPUMS 1940 Census
- Machine-learned ethnicity identification for the University Student and Faculty Databases
- Links between the University Student Database and contemporaneous high school teacher and doctor licensing records, as used in Bleemer (2016)
- Additional universities added to the Faculty and Course Databases

Each of these additions will be documented on the UC-CHP website.

∞∞∞∞∞∞

Bibliography

Arcidiacono, Peter, Esteban M. Aucejo, and V. Joseph Hotz. "University Differences in the Graduation of Minorities in STEM Fields: Evidence from California". 2016. *American Economic Review* 106(3): 525-562.

Bleemer, Zachary. 2016. "Role Model Effects of Female STEM Teachers and Doctors on Early 20th Century University Enrollment in California". Center for Studies in Higher Education Research and Occasional Paper Series 10.16.

Brunet, Gillian. 2017. "Stimulus on the Home Front: The State-Level Effects of WWII Spending". Manuscript.

Chetty, Raj, John N. Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. 2017. "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility". NBER Working Paper 23618.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2016. Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech". NBER Working Paper 22423.

Goldin, Claudia and Lawrence Katz. 2010. The 1915 Iowa State Census Project. ICPSR28501-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-12-14. https://doi.org/10.3886/ICPSR28501.v1

Hoberg, Gerard and Gordon Phillips. 2016. "Text-Based Network Industries and Endogenous Product Differentiation". *Journal of Political Economy* 124(5): 1423-1465.

Hoberg, Gerard and Gordon Phillips. "Text-Based Industry Momentum". Forthcoming. *Journal of Financial and Quantitative Analysis*.

Kirkeboen, Lars, Edwin Leuven, and Magne Mogstad. 2016. "Field of Study, Earnings, and Self-Selection". *The Quarterly Journal of Economics* 131(3): 1057-1111.

Levenshtein, Vladimir I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." *Soviet Physics–Doklady* 10(8): 707–710.

Olivetti, Claudia and M. Daniele Paserman. 2015. In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940. *American Economic Review* 105(8): 2695-2724.

Underwood, Ted. 2015. "The Literary Uses of High-Dimensional Space". *Big Data and Society* 2(2).

Underwood, Ted. 2016. "The Life Cycles of Genre". *Journal of Cultural Analytics* 05.23.16.

Data Appendix

The following is a list of primary sources and digital providers from which the data described in this paper were constructed. HathiTrust records can be accessed online at https://catalog.hathitrust.org/Record/###, where ### is the record number.

University Student Database

- University of California system: University of California Register, 1893-1946, published by the University in Berkeley, California. Available in HathiTrust records 007130126, 011249103, 007910193, 100024883, and 003915007, which were scanned by partnerships between Google and the University of California, the University of Illinois at Urbana-Champaign, Cornell University, and the University of Michigan. The 1903 Register is available in print from the UC Berkeley Bancroft Library, which digitized it for UC-CHP (currently not publicly-available); the 1945 Register is unavailable.
- Stanford University: Leland Stanford Junior University Annual Register, 1893-1918, and the Stanford University Annual Register, 1919-1946. Published by the University in Stanford, California. Scanned by and available from the Stanford Publications division of the Stanford University Library: https://exhibits.stanford.edu/stanford-pubs/browse/annual-register-1891-1947.
- University of Southern California: University of Southern California Year-Book, 1905-1921, and University of Southern California Circular of Information, 1905-1924. Published by the University in Los Angeles, California. Available in HathiTrust records 100630461 and 000056358, which were scanned by partnerships between Google and both the University of Illinois at Urbana-Champaign and the University of Michigan.
- California Institute of Technology: Throop College of Technology Annual Catalogue, 1912-1919, and Bulletin of the California Institute of Technology, 1920-1946. Published by the Institute in Pasadena, California. Available in HathiTrust record 100607120 (1912-1921), which was scanned by a partnership between Google and the University of Illinois; scanned by and available online from CaltechCampusPubs at the Caltech Library (1920-1946): http://caltechcampuspubs.library.caltech.edu/view/publication/Bulletin_of_the_California_Institute_of_Technology.html. No Bulletins were published in 1942 and 1943.
- Mills College: Annual Catalogue of Mills College, 1903-1940. Published by the College in Oakland, California. Available in HathiTrust record 005808070, which was digitized by partnerships between Google and both of the University of Illinois at Urbana-Champaign and the University of Michigan. The 1920-1923 and 1925 Catalogues are unavailable.

University Faculty and Course Databases

- Stanford University: Stanford University Announcement of Courses (1900-1952), Stanford University Courses and Degrees (1952-1994), and Stanford Bulletin (1995-2000). Published by the University in Stanford, California. Scanned by and available from the Stanford Publications division of the Stanford University Library: http://exploredegrees.stanford.edu/archive/. The 1902-1904, 1921, 1961, and 1967 volumes are unavailable.
- University of California, Berkeley: Register of the University of California, 1900-1958, and UC Berkeley General Catalogue,1961-2013. Published by the University in Berkeley, California. Available in HathiTrust record 007130126 (1900-1955), which were scanned by partnerships between Google and the University of California and the University of Illinois; scanned by and available online from the University of California, Berkeley Library (1900-2013): http://digitalassets.lib.berkeley.edu/generalcatalog/. Catalogues were published biannually starting in 1995.
- University of California, Davis: UC Davis General Catalog, 1955-2016. Published by the University in Davis, California. Scanned by and available from the UC Davis Office of the Registrar: http://catalog.ucdavis.edu/pdf.html. Catalogs were published biannually starting in 2000.
- University of California, Los Angeles: UCLA General Catalog, 1900-2015. Published by the University in Los Angeles, California. Scanned by and available from the UCLA Registrar's Office: http://www.registrar.ucla.edu/Archives/General-Catalog-Archive/UCLA-General-Catalog. Catalogs were published biannually between 1995 and 2007.

University of California Budget Database

- University of California Budget (1911-1961) and Departmental Allocations of the University of California: Budget for Current Operations (1962-2012). No publication information; not for public use. Scanned by the UC Berkeley Bancroft Library on behalf of the UC ClioMetric History Project; available in HathiTrust record 102153438. Departmental Allocations are unavailable between 1983 and 1987.