



HATHITRUST

A Shared Digital Repository

# HathiTrust, Collective Action, and Local Services

---

University of California Davis  
February 25, 2015  
Mike Furlough  
Executive Director, HathiTrust

# Mission

---

To contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge.

Efforts include, but are not limited to

- ...building comprehensive collections co-owned and managed by partners.

- ...enabling access by users with print disabilities.

- ...supporting computational research with the collections.

- ...stimulating shared collection storage strategies among libraries.



# Timeline: Highlights

---

- Google Library Project announced (2004)
- Launch (2008)
- TRAC certification (2011)
- Constitutional convention (2011)
- 10 million volumes (2012)
- New governance established (2012)
- Current bylaws and fee structure (2013)
- 13 million volumes (2014)



# Today's Conversation

---

- HathiTrust Today
  - Collections
  - Organization
  - Current Initiatives
  - Short term plans
- HathiTrust Tomorrow
  - How has the world changed?
  - How should we change it?



# Collections



# The Name

---

- The meaning behind the name
  - Hathi (hah-tee)--Hindi for elephant
  - Big
  - Never forgets
  - Secure
  - Trustworthy



THE Law of the Jungle  
—which is by far  
the oldest law in the  
world—has arranged  
for almost every kind  
of accident that may  
befall the Jungle-  
People, till now its  
code is as perfect as  
time and custom can make it. You will remember.



# Preservation with Access

---

- Preservation
  - TRAC-certified
  - Long-term commitments on digital content facilitate planning, decision-making
- Discovery
  - Bibliographic and full-text search of all materials
  - Mechanisms for local loading of records
- Access and Use
  - Full text search (all users)
  - Public domain and open access works (all users)
  - Print on demand (all users, selected works)
  - Collections and APIs (all users)
  - Lawful uses of in-copyright works (members)



# HathiTrust in 2015

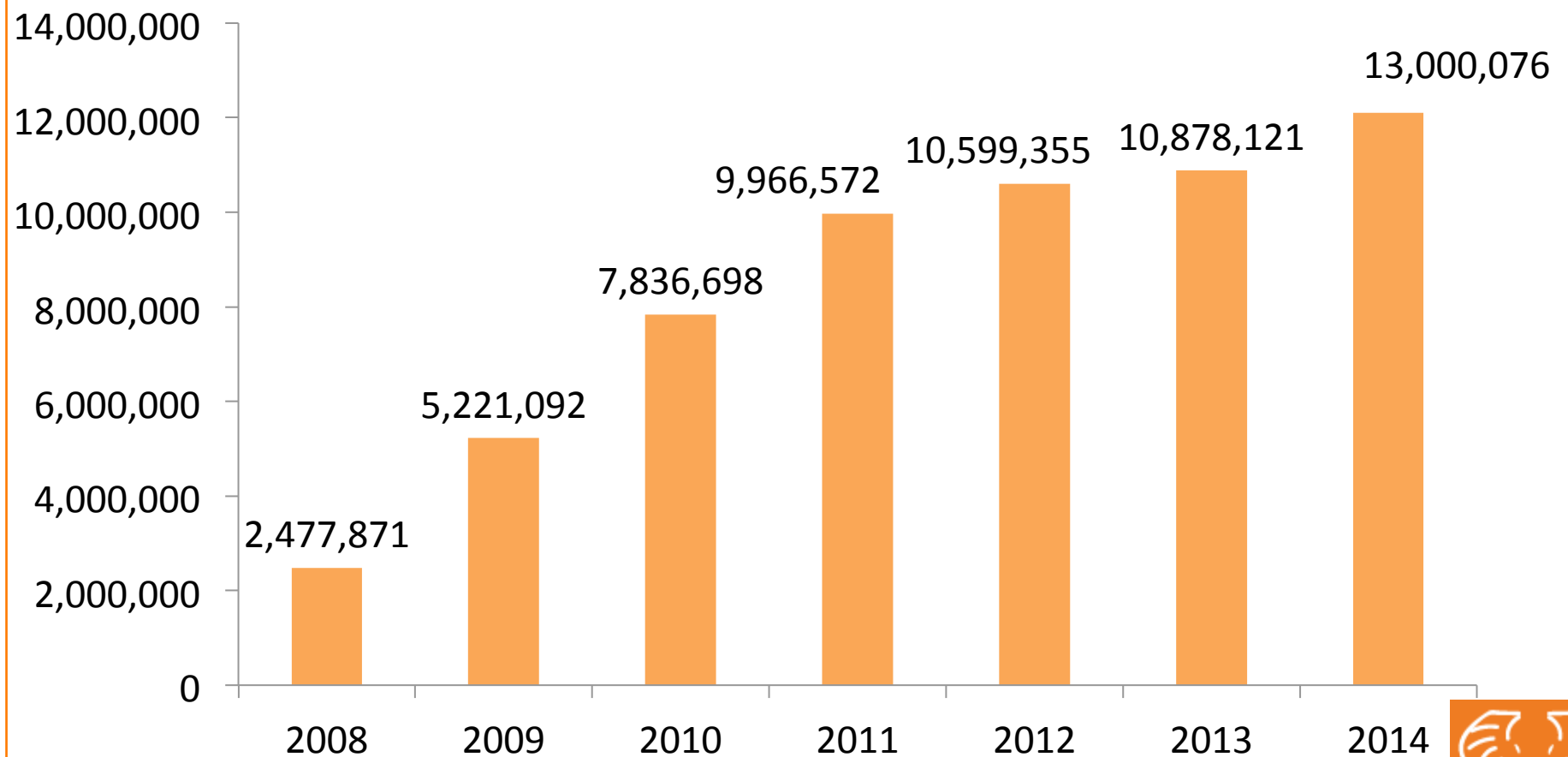
---

- 13 million total items
  - 6.7 million book titles
  - 345,000 serial titles
  - 604,000 US federal government documents
  - 4.9 million items open (public domain & CC-licenses)
  - A handful of images and thimbleful of audio files





# Growth of Collection



# Largest Contributors (end 2014)

Volumes Added	Total Volumes
University of Michigan	4,712,752
University of California*	3,612,596
Harvard University	838,110
University of Wisconsin	560,775
Indiana University	528,811
Cornell University	510,065
Penn State University	387,717
University of Illinois	318,131
New York Public Library	294,835
Princeton University	252,808

## **Davis:**

Total: 39,823

Google: 36,194 (all public domain)

IA: 3629

## **NRLF:**

Total: 2,600,753

Google: 2,530,463 - including  
323,738 public domain

IA: 70,290



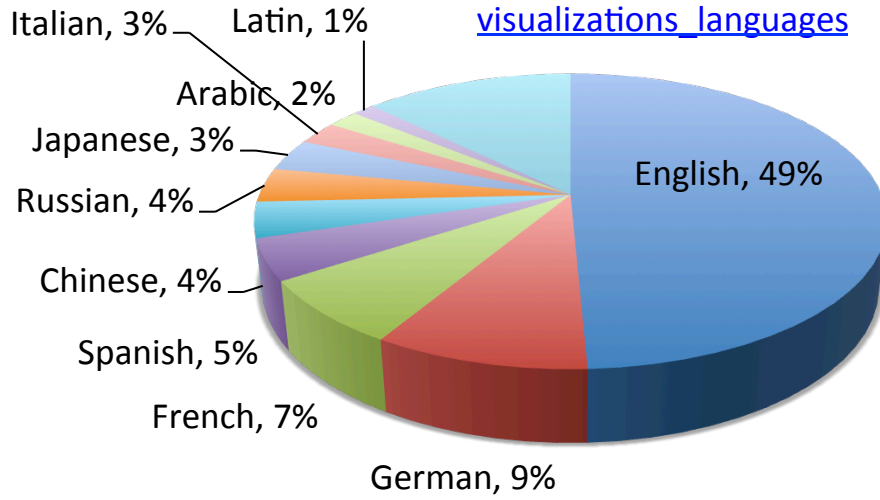
# 2014: Largest Contributions (by volumes)

Volumes Added	Jan-Dec 2014	Total Volumes	Pct Growth
Harvard University	600,675	838,110	71.67%
Indiana University	333,231	528,811	63.02%
Penn State University	319,513	387,717	82.41%
University of Illinois	205,156	318,131	64.49%
University of California	164,426	3,612,596	4.55%
Keio University	90,094	90,094	100.00%
University of Alberta	76,106	76,106	100.00%
Cornell University	72,574	510,065	14.23%
Ohio State University	61,129	61,129	100.00%
University of Michigan	46,720	4,712,752	0.99%

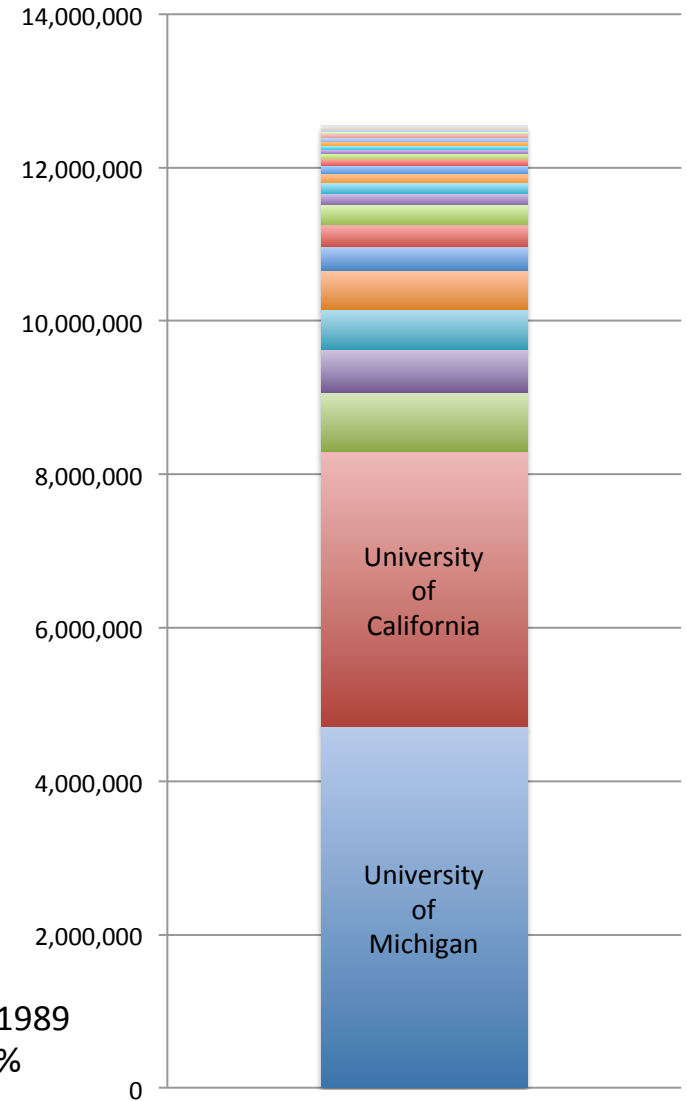
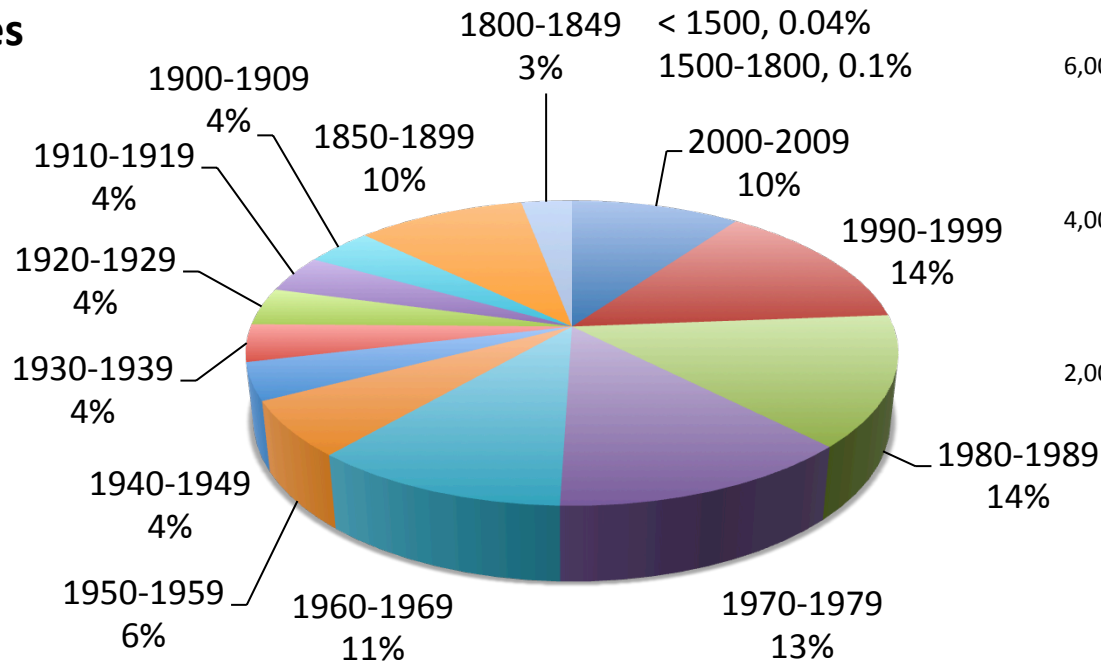


### Top 10 Languages

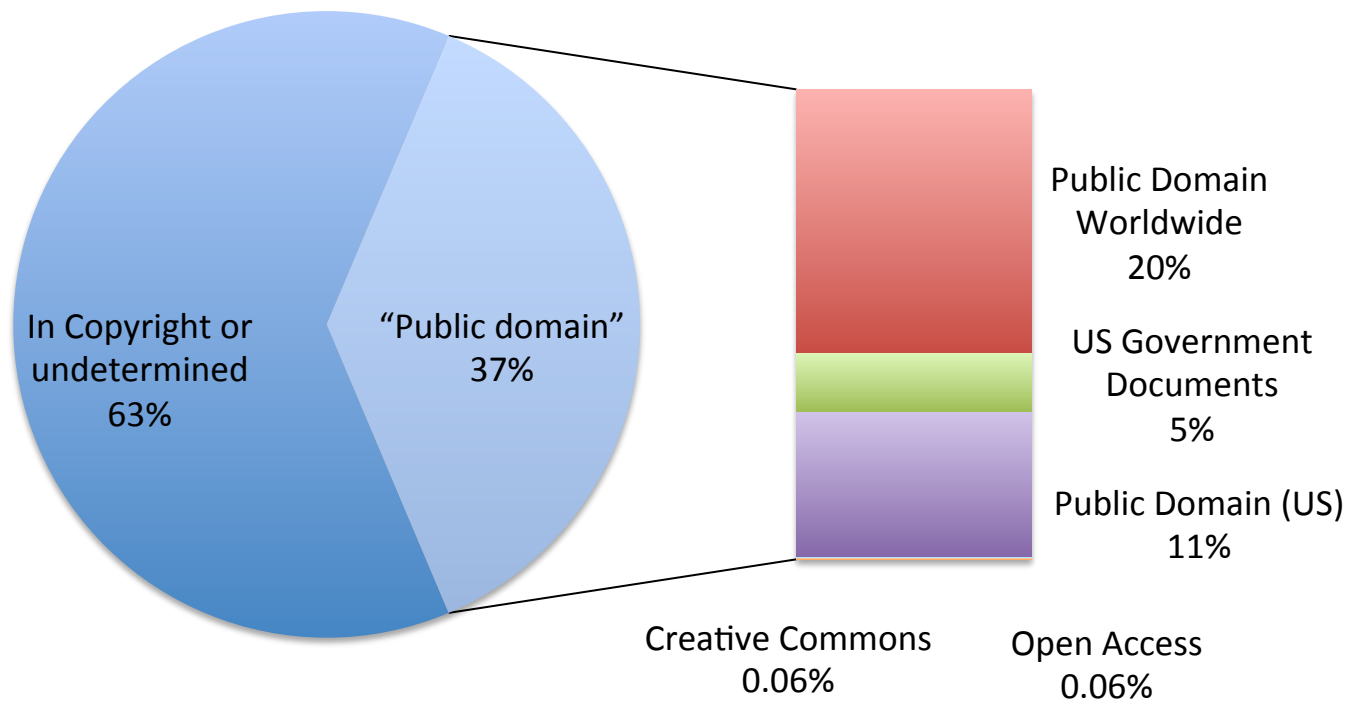
[http://www.hathitrust.org/visualizations\\_languages](http://www.hathitrust.org/visualizations_languages)



### Dates



# Copyright Distribution



Type of work	Searchable (bibliographic and full-text)	Viewable*	Full-PDF download	Print on Demand	Print disabilities*	Preservation uses (Section 108)*
Public domain worldwide	Worldwide	Worldwide	Partners only if 3 <sup>rd</sup> -party restrictions, if not, worldwide.	Worldwide	Worldwide	N/A
Public domain (US) – Non-US works published between 1873 and 1923.	Worldwide	When accessed from within the United States	Partners in the US if 3 <sup>rd</sup> party restrictions, if not, anyone in the US	Available within the United States	Partners in the US; partners worldwide where laws permit	N/A
Works that rights holders have opened access to in HathiTrust	Worldwide	Worldwide	Worldwide (if digitized by Google, full-PDF only available if opened with CC license)	Worldwide with permission	Worldwide	N/A
Works that are in-copyright or of undetermined status	Worldwide	Not available	Not available	Not available	Partners in the US; partners worldwide where laws permit	Partners in the US; partner worldwide where laws permit

\* Note: Access to in-copyright works is subject to conditions listed in HathiTrust's policies on [Access and Use](#).



# Access: Lawful uses of in-copyright works

---

- Sensitive to multiple legal regimes
  - Full-text search (everyone everywhere)
  - Access to users who have print disabilities (through member proxy in US, and where law permits)\*\*
  - Access works that are damaged or missing and also out of print and unavailable (members in US only)

\*\*Terms and conditions at

[http://www.hathitrust.org/access\\_use#ic-access](http://www.hathitrust.org/access_use#ic-access)



# Collective Action: Copyright Review

---

- Copyright Review Management System
  - Systematic manual review of copyright registrations to determine status of portions of the HathiTrust Collection
  - CRMS US: Published in US, 1923-1963
    - 318,887 reviewed / 168,248 PD (~53%)
  - CRMS-World: Published in UK (1874-1944), Canada, Australia (1894-1964)
    - 175,681 reviewed / 92,919 PD-world 9 (~53%)

Supported generously by IMLS





# Top Ten Titles January 2015

---

1. The Human Figure, by John H. Vanderpoel
2. Quicksand, by Nella Larsen.
3. Godey's Magazine, v.40-41, 1850.
4. Pennsylvania German pioneers: A Publication of the Original Lists of Arrivals in the Port of Philadelphia from 1727 to 1808, by Ralph Beaver Strassburger.
5. The Book of a Hundred Hands, by George Brant Bridgman.
6. Indian boyhood, by Charles A. Eastman.
7. Roster of the Confederate soldiers of Georgia, 1861-1865, v.2.
8. Solid mensuration, by Willis F. Kern and James R. Bland.
9. The Five Laws of Library Science, by S. R. Ranganathan.
10. Roster of the Confederate soldiers of Georgia, 1861-1865, v.1.



# Shared Stewardship



# HathiTrust Members

Allegheny College  
American University of Beirut  
Arizona State University  
Baylor University  
Boston College  
Boston University  
Brandeis University  
Brown University  
California Digital Library  
Carnegie Mellon University  
Case Western Reserve  
Colby College  
Columbia University  
Cornell University  
Dartmouth College  
Duke University  
Emory University  
Florida State University  
Getty Research Institute  
Georgetown University  
Georgia Tech  
Harvard University Library  
Indiana University  
Iowa State University  
Johns Hopkins University  
Kansas State University  
Lafayette College  
Library of Congress  
Massachusetts Institute of Technology  
McGill University  
Michigan State University  
Montana State University  
Mount Holyoke College  
New York Public Library  
New York University  
North Carolina Central University

North Carolina State University  
Northeastern University  
Northwestern University  
Oklahoma State University  
The Ohio State University  
The Pennsylvania State University  
Princeton University  
Purdue University  
Rutgers University  
Stanford University  
State University System of Florida  
Syracuse University  
Temple University  
Texas A&M University  
Texas Tech University  
Tufts University  
Universidad Complutense de Madrid  
University of Alabama  
University of Alberta  
University of Arizona  
University of British Columbia  
University of Calgary  
**University of California**  
**Berkeley**  
**Davis**  
**Irvine**  
**Los Angeles**  
**Merced**  
**Riverside**  
**San Diego**  
**San Francisco**  
**Santa Barbara**  
**Santa Cruz**  
The University of Chicago  
University of Connecticut

University of Delaware  
University of Houston  
University of Illinois  
University of Illinois at Chicago  
The University of Iowa  
University of Kansas  
University of Maine  
University of Maryland  
University of Massachusetts, Amherst  
University of Miami  
University of Michigan  
University of Minnesota  
University of Missouri  
University of Nebraska-Lincoln  
University of New Mexico  
The University of North Carolina at Chapel Hill  
University of Notre Dame  
University of Oklahoma  
University of Pennsylvania  
University of Pittsburgh  
University of Queensland  
University of Tennessee, Knoxville  
University of Texas  
University of Utah  
University of Vermont  
University of Virginia  
University of Washington  
University of Wisconsin-Madison  
Utah State University  
Vanderbilt University  
Virginia Tech  
Wake Forest University  
Washington University  
Yale University Library



# Cooperative Work

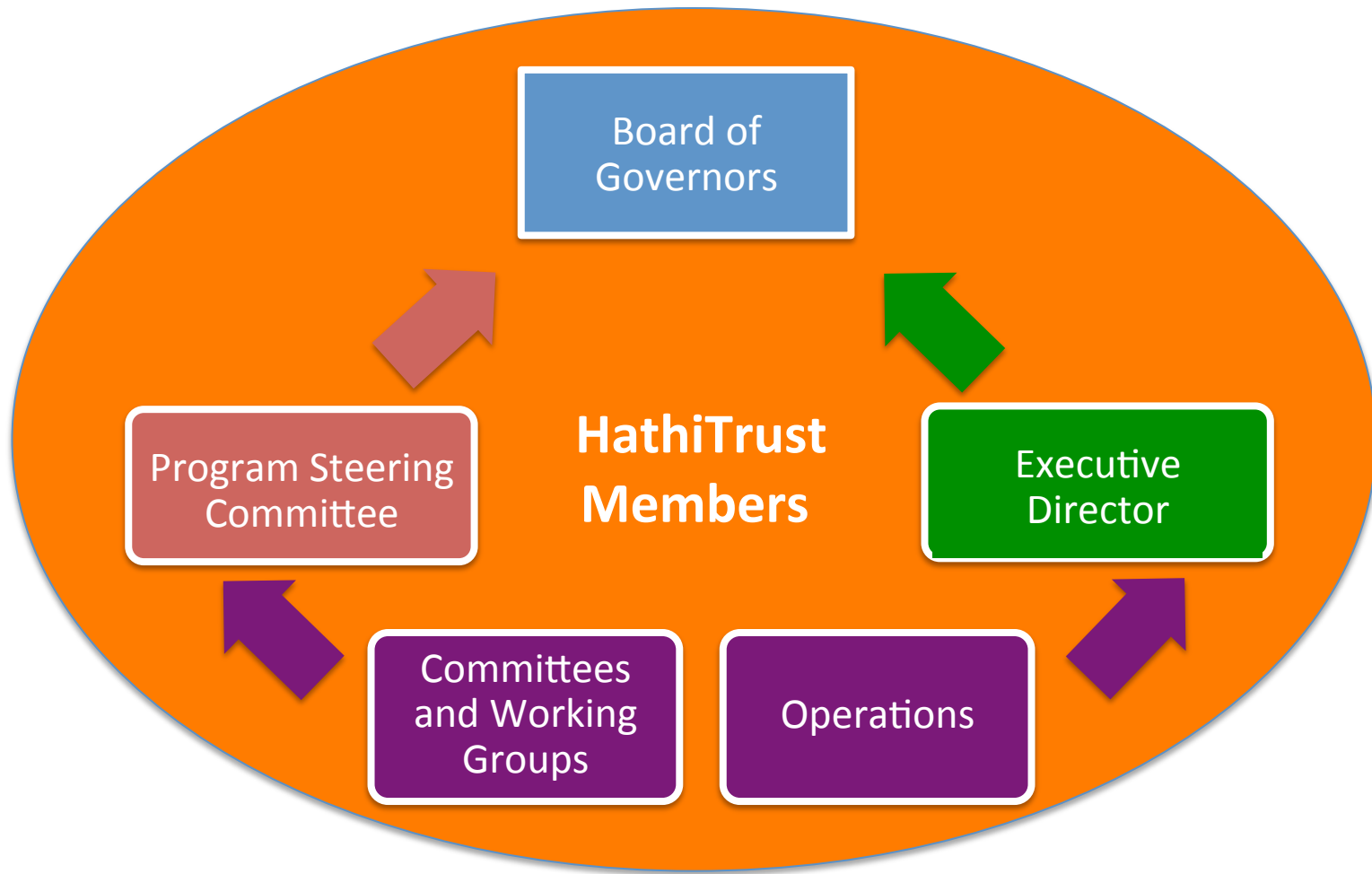
---

- Draw upon knowledge across institutions
- Distributed Functions and Services
  - Preservation repository and access services
    - University of Michigan
    - Mirror site: Indiana University
  - Metadata management services
    - California Digital Library
  - HathiTrust Research Center
    - Indiana University and University of Illinois



# Governance

---



# Committees and Working Groups

---

- Program Steering Committee
- Collections Committee
- Zephyr Advisory Group
- User Support Working Group
  
- Rights and Access Working Group
- Government Documents Initiative Planning and Advisory Group
- Print Monographs Archive Planning Task Force
  
- On Hiatus
  - Communications
  - User Experience



# Current Initiatives



# Current Initiatives

---

1. Developing a shared print monographs archive
2. Expanding coverage and access to US government publications
3. Expanding support for computational (non-consumptive) research
4. Development Priorities





# Shared Print Monographs Archive

---

- Ballot Initiative passed at the 2011 HT Constitutional Convention (Con-Con)
  - “To develop a print monographs archive corresponding to volumes represented within the HathiTrust”
- Focus
  - Ensure preservation of print and digital collections
  - Catalyze national/continental collective management of collections



# Why A Shared Print Archive Program

---

- Many regional efforts, but limited national/international coordination
- Strengthens preservation commitments
  - Connects both print and digital preservation
- Significant need and desire to reduce costs of collection management and associated footprint



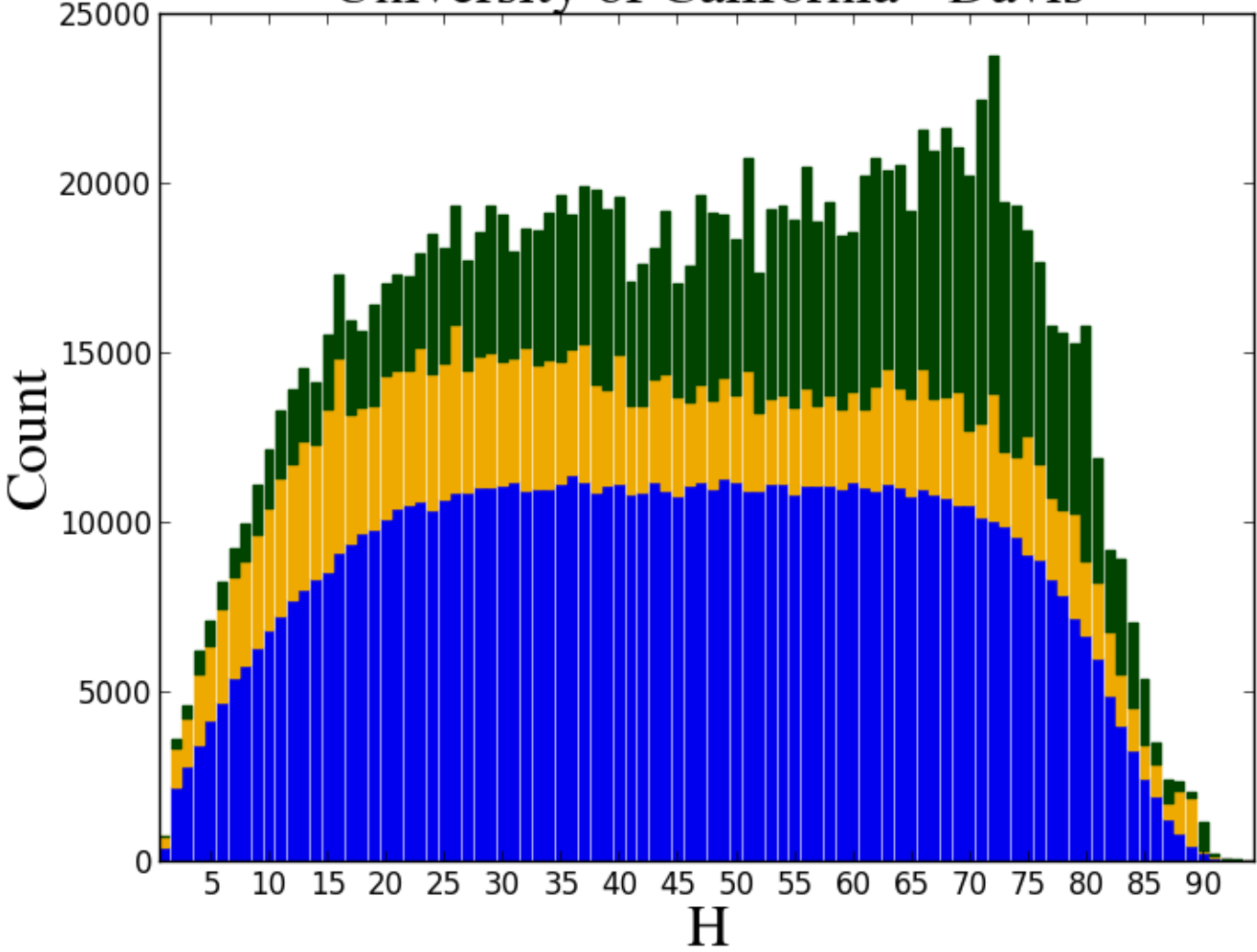
# UC Collection Overlap (by titles)

---

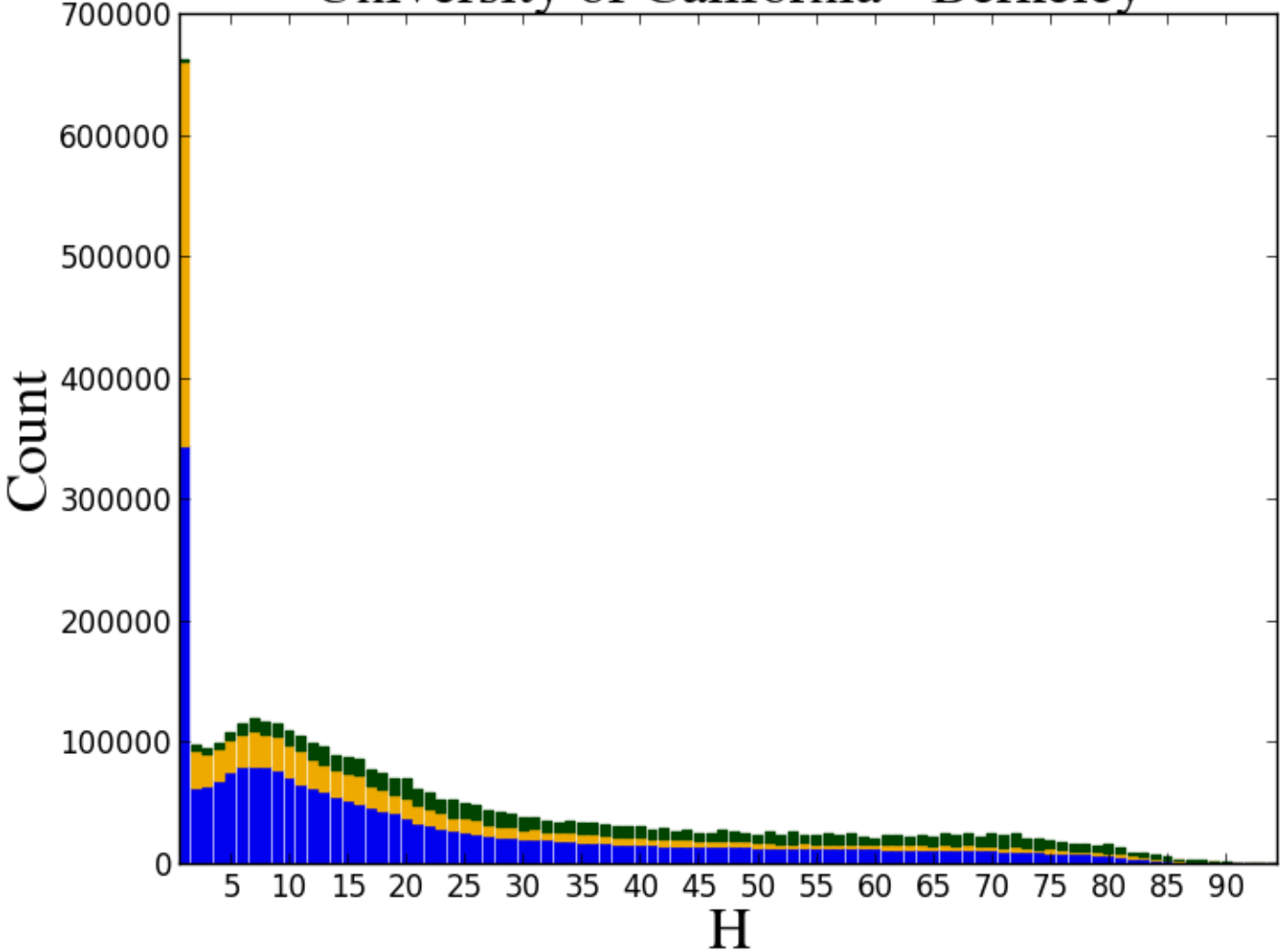
	Submitted	In Hathi	Percent
single-part monographs	9,766,951	3,332,926	34.1%
multi-part monographs	985,087	467,318	47.4%
serials	349,422	89,545	25.6%
<b>TOTAL</b>	<b>11,101,460</b>	<b>3,889,789</b>	<b>35.0%</b>



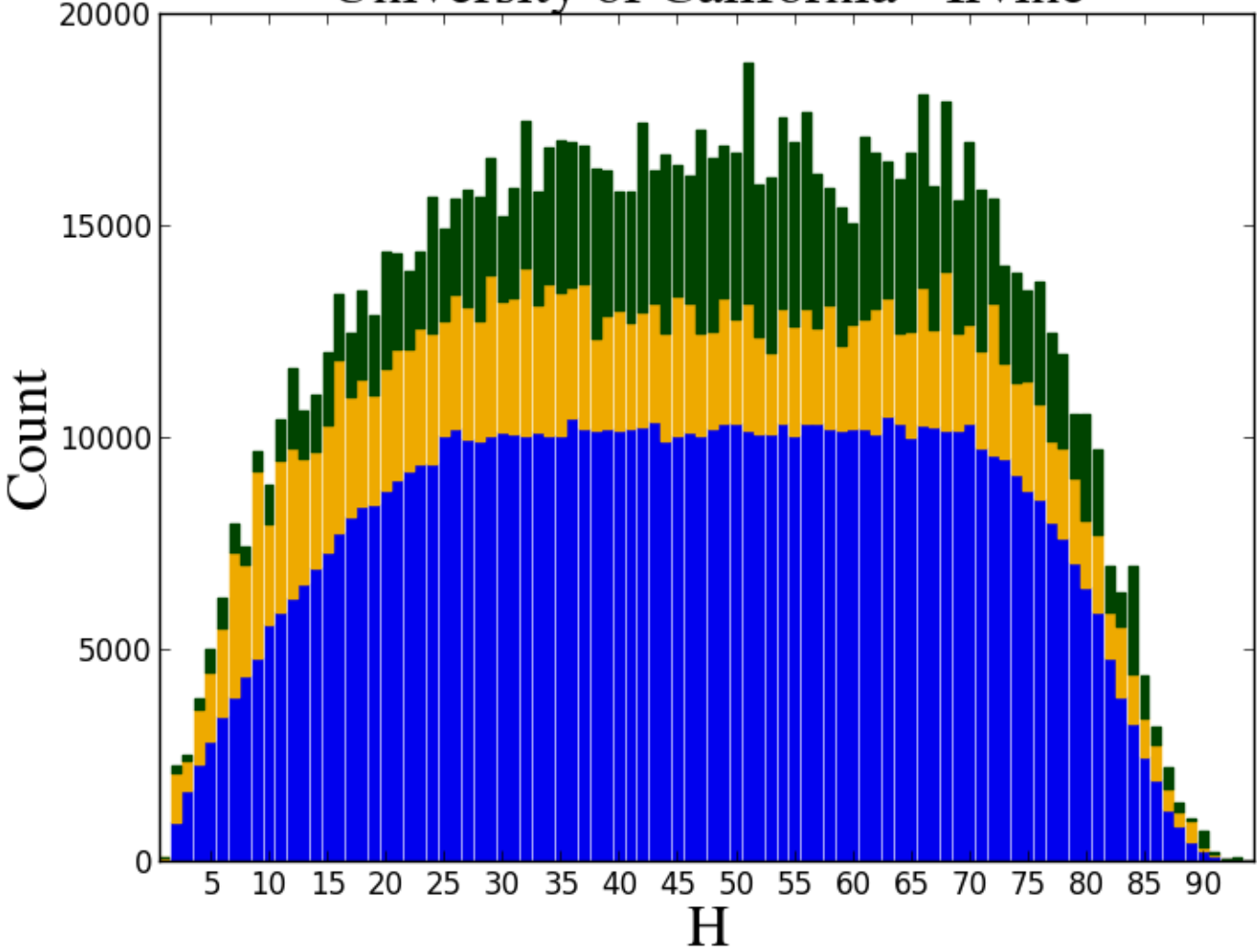
# University of California - Davis



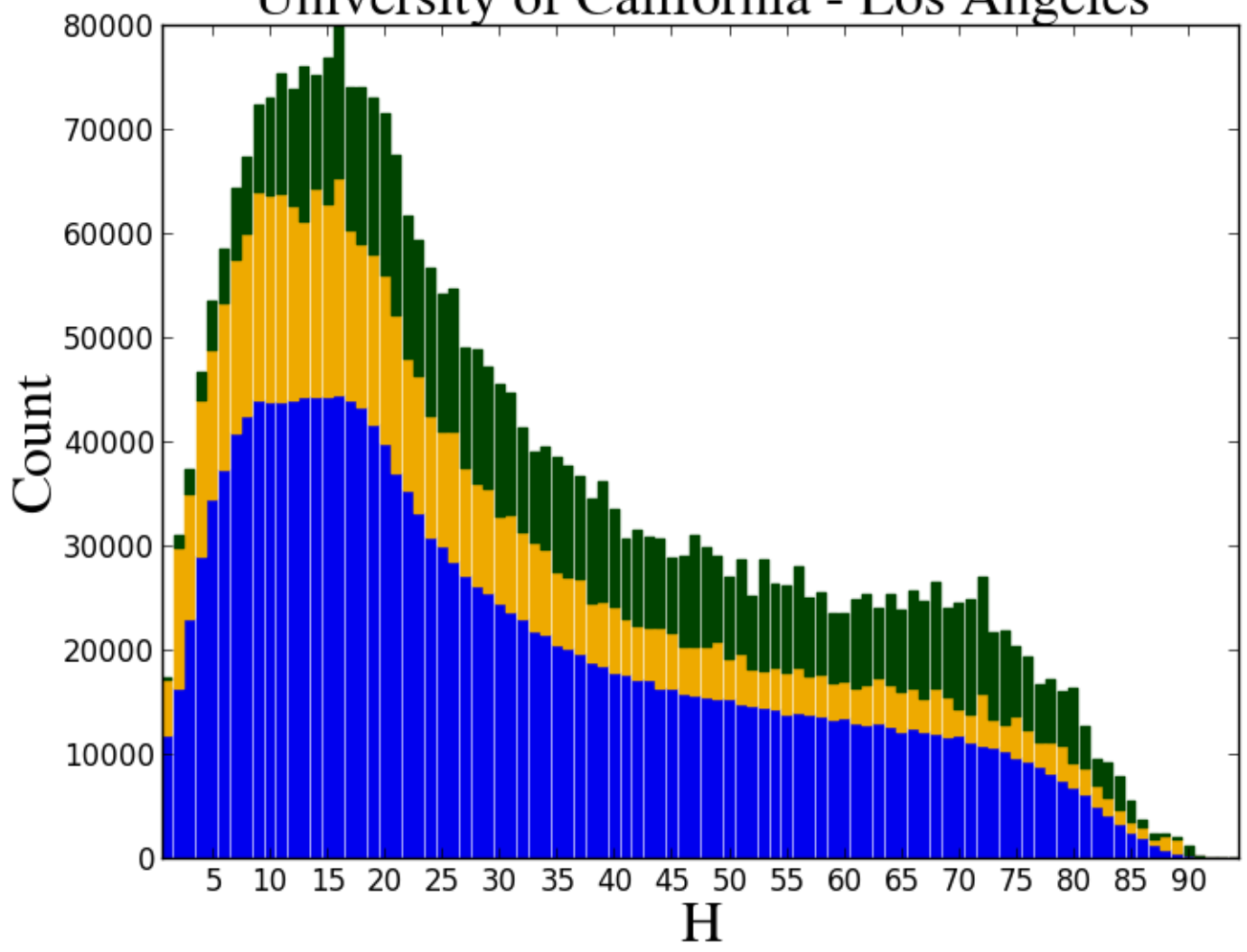
# University of California - Berkeley



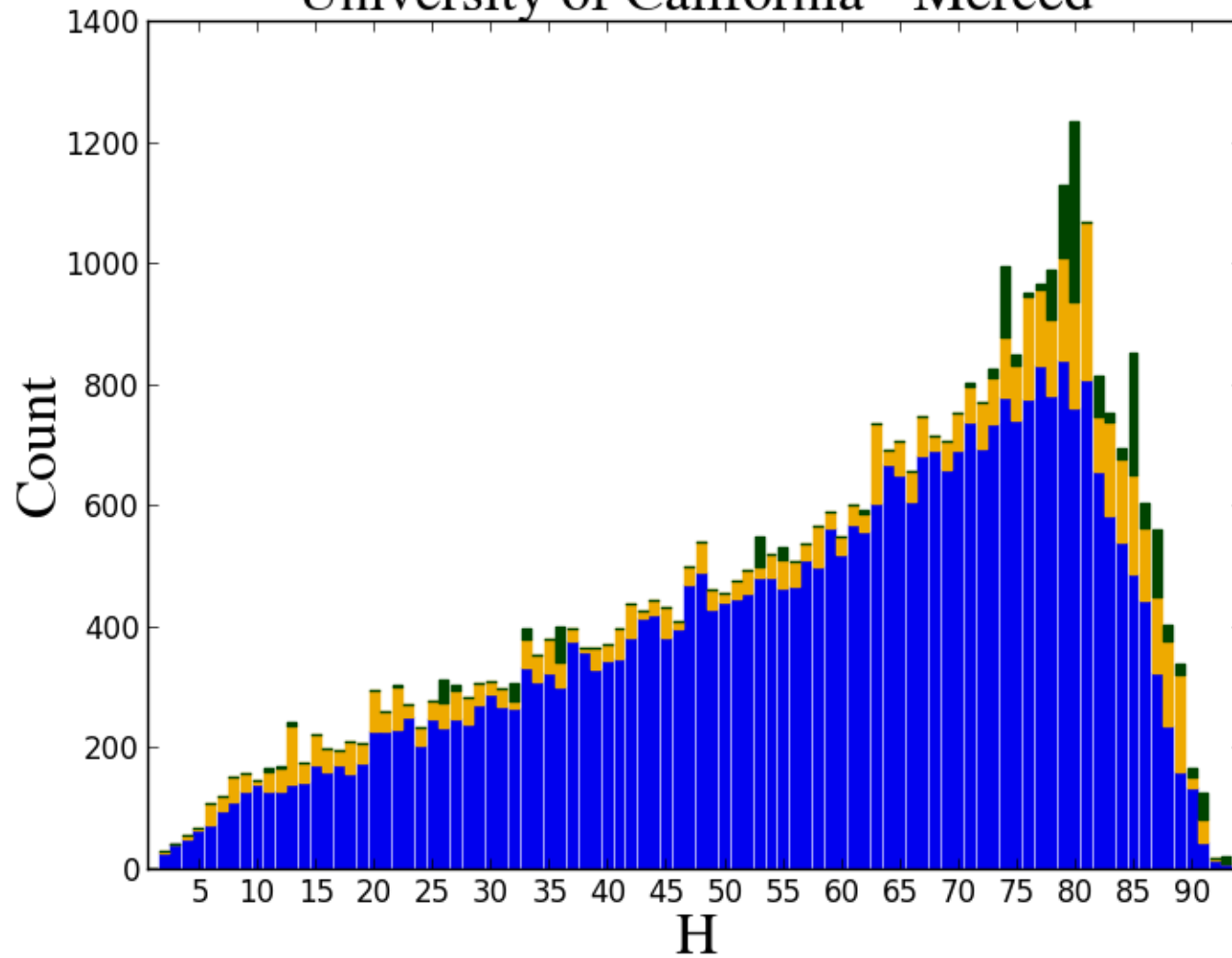
# University of California - Irvine



# University of California - Los Angeles

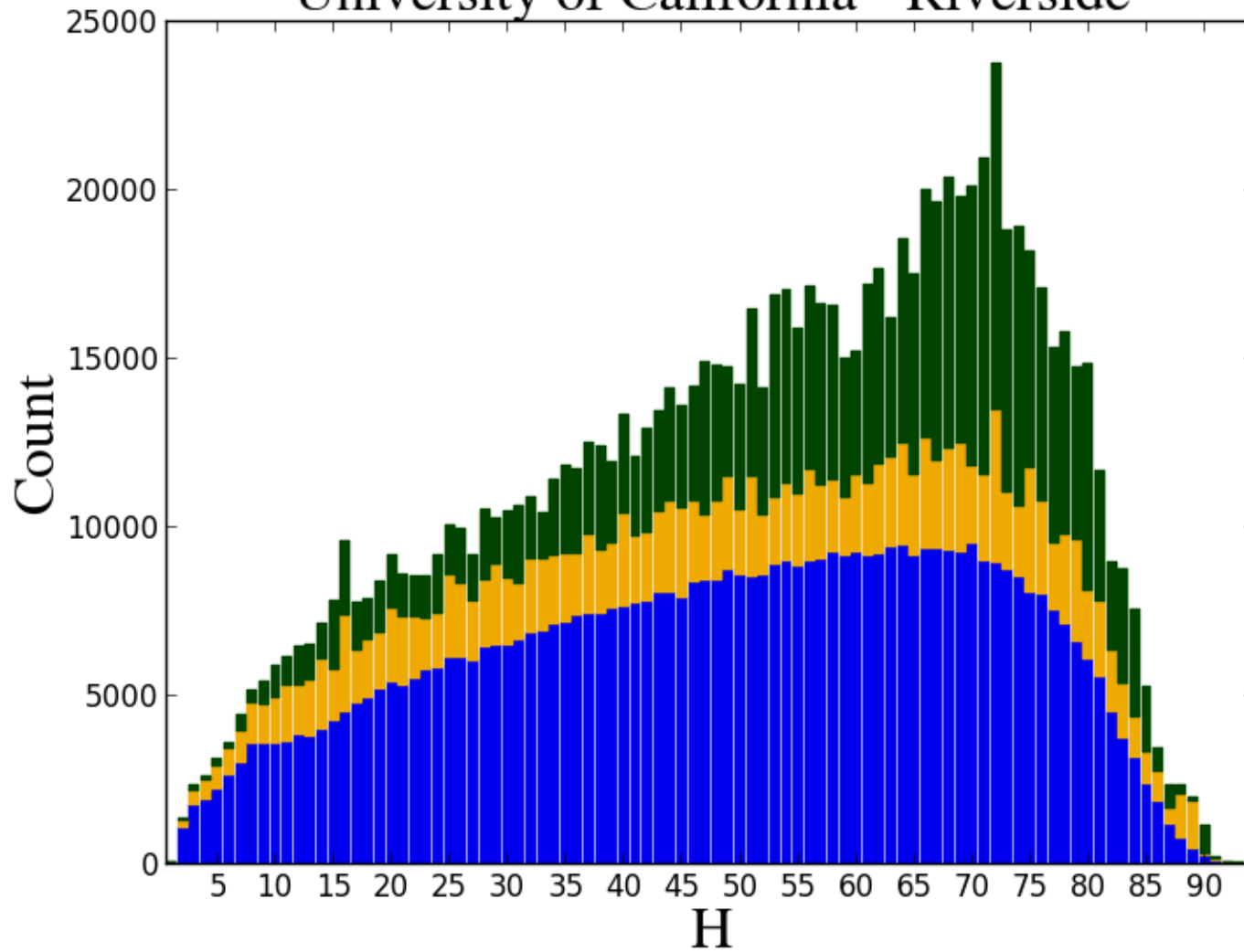


# University of California - Merced

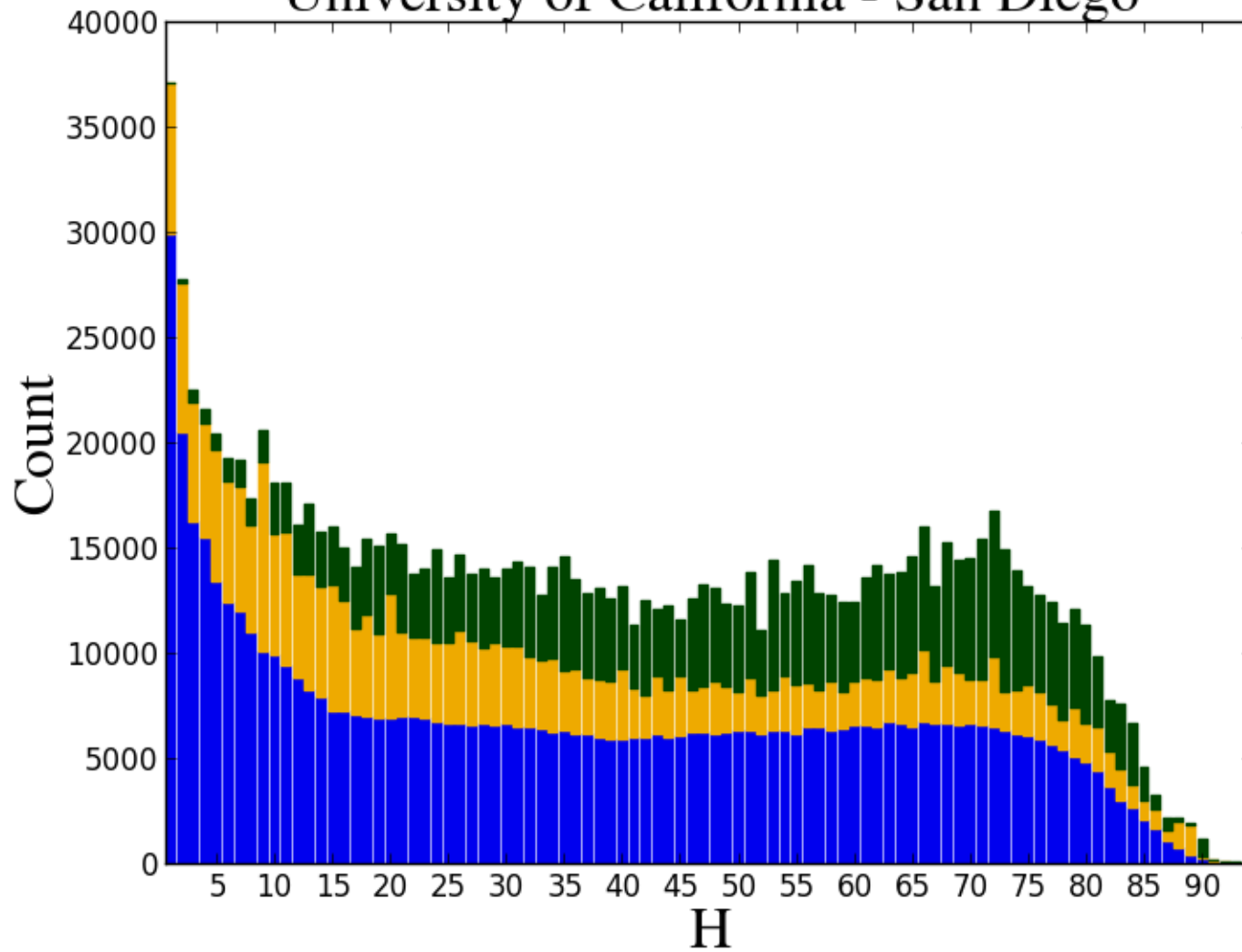




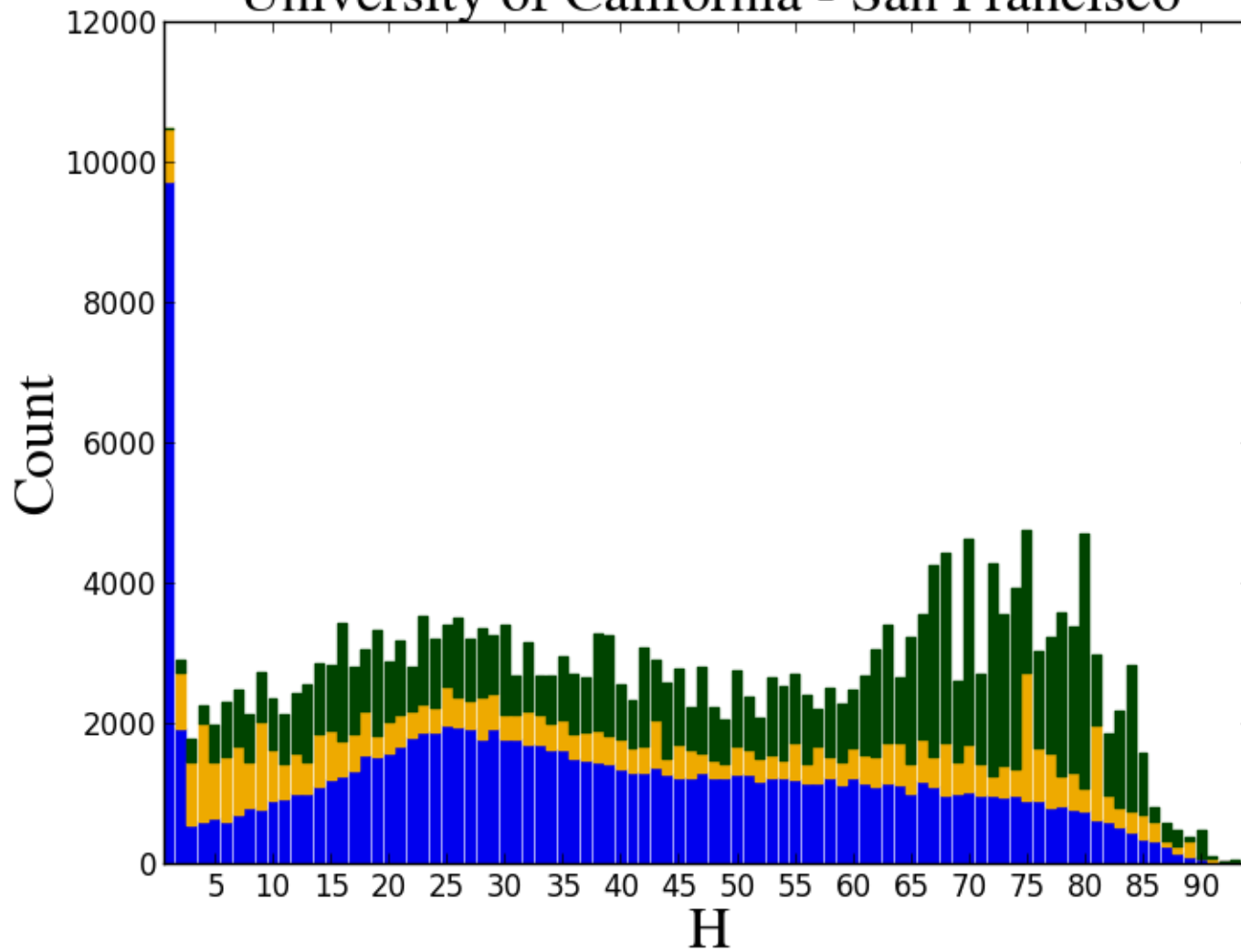
# University of California - Riverside



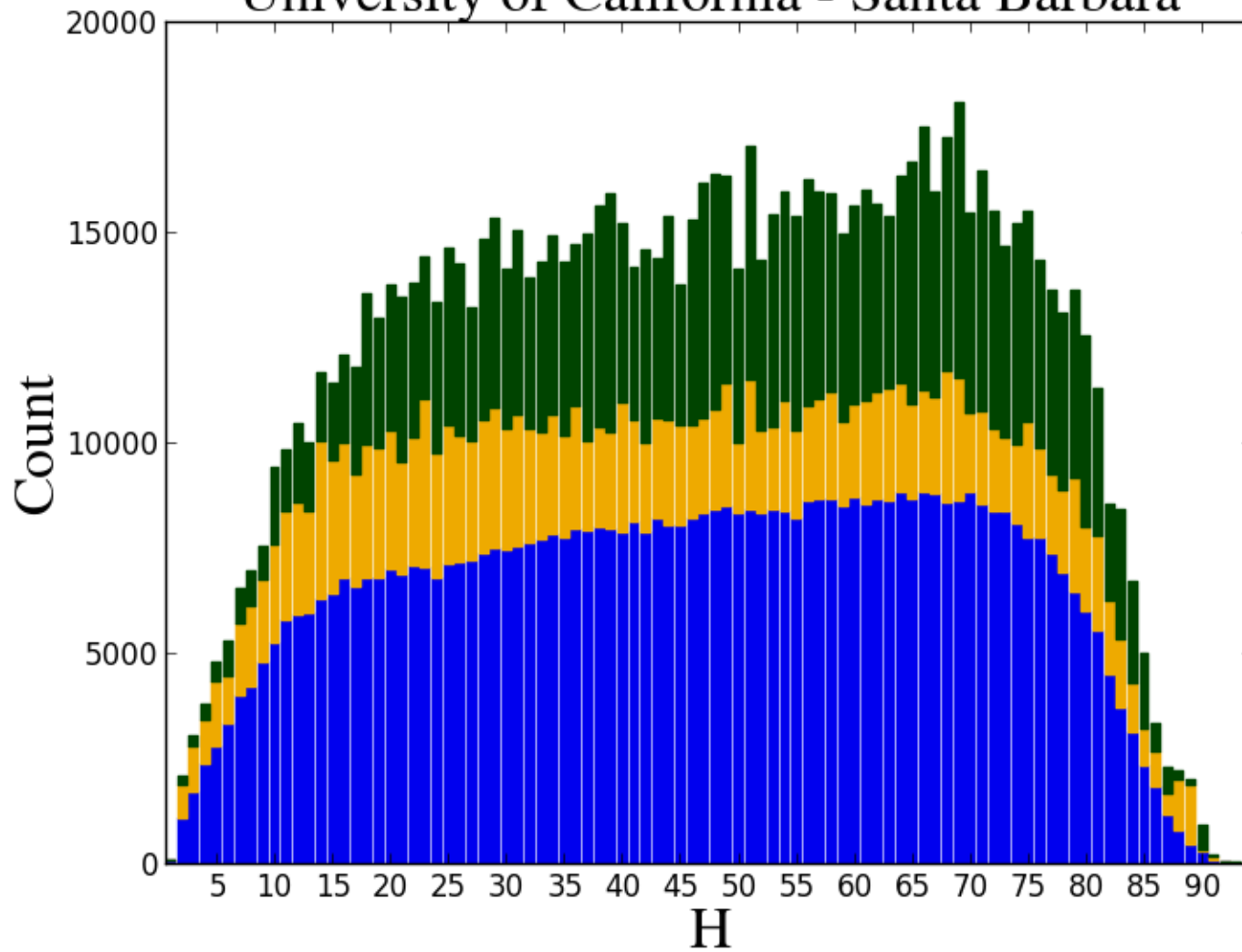
# University of California - San Diego



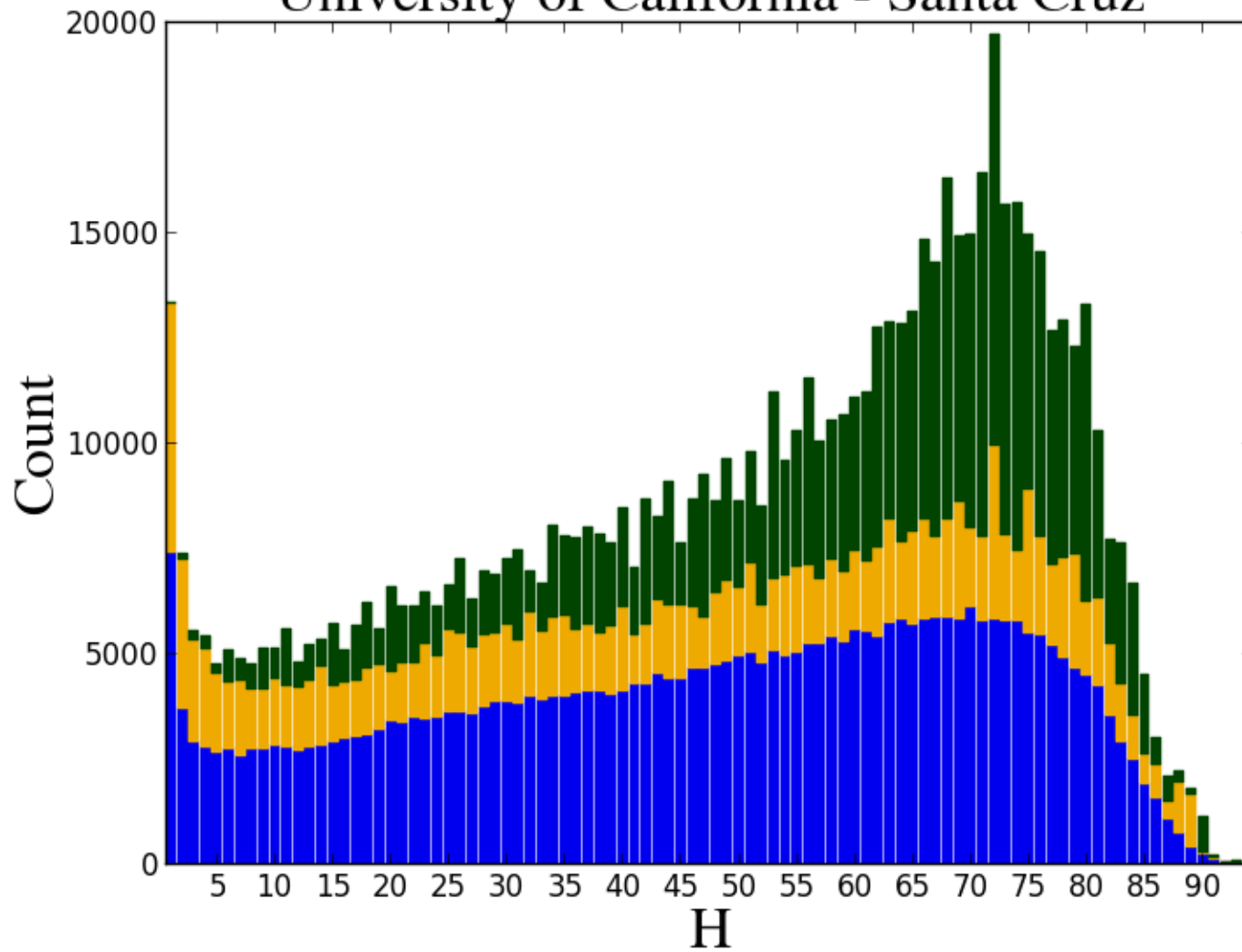
# University of California - San Francisco



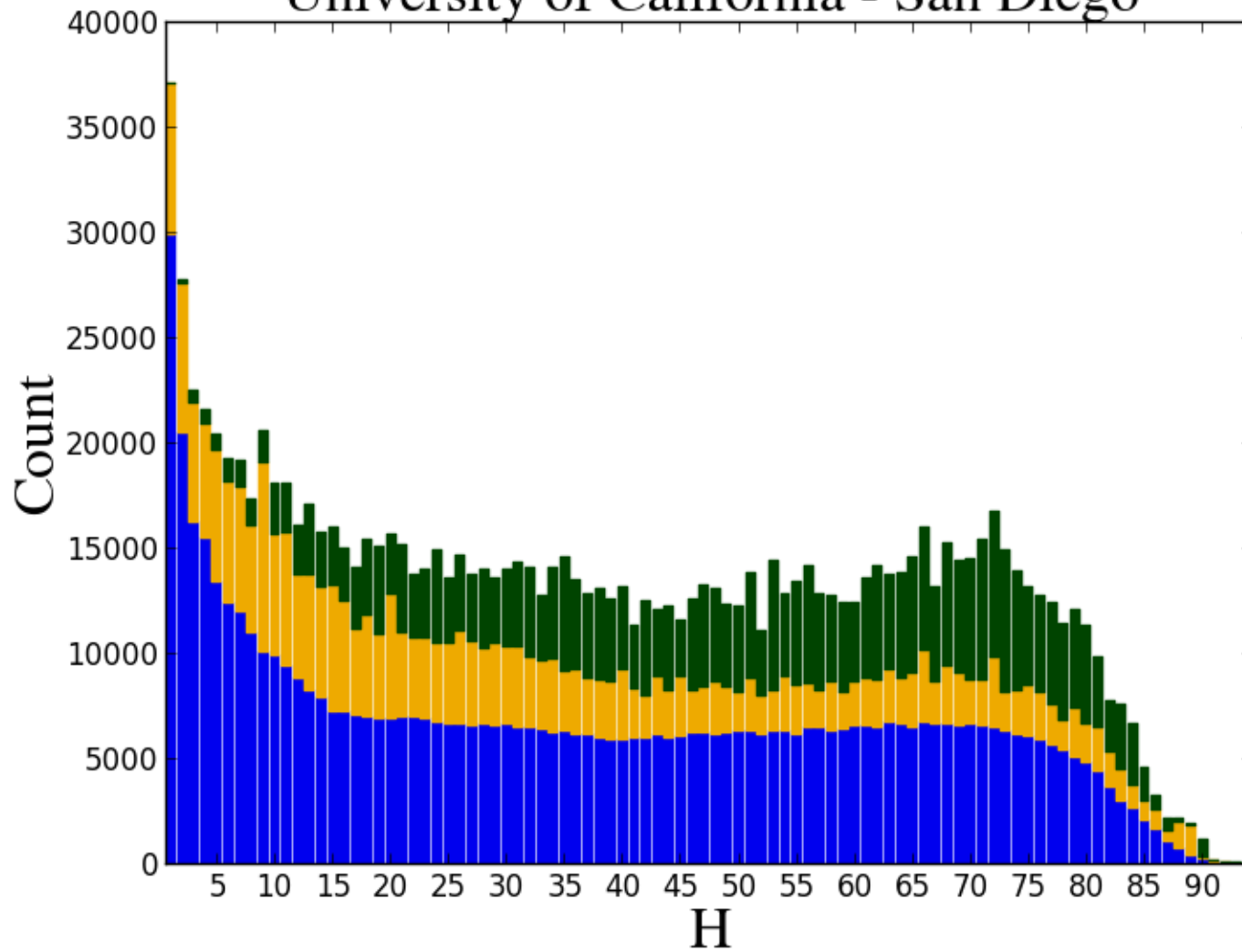
# University of California - Santa Barbara



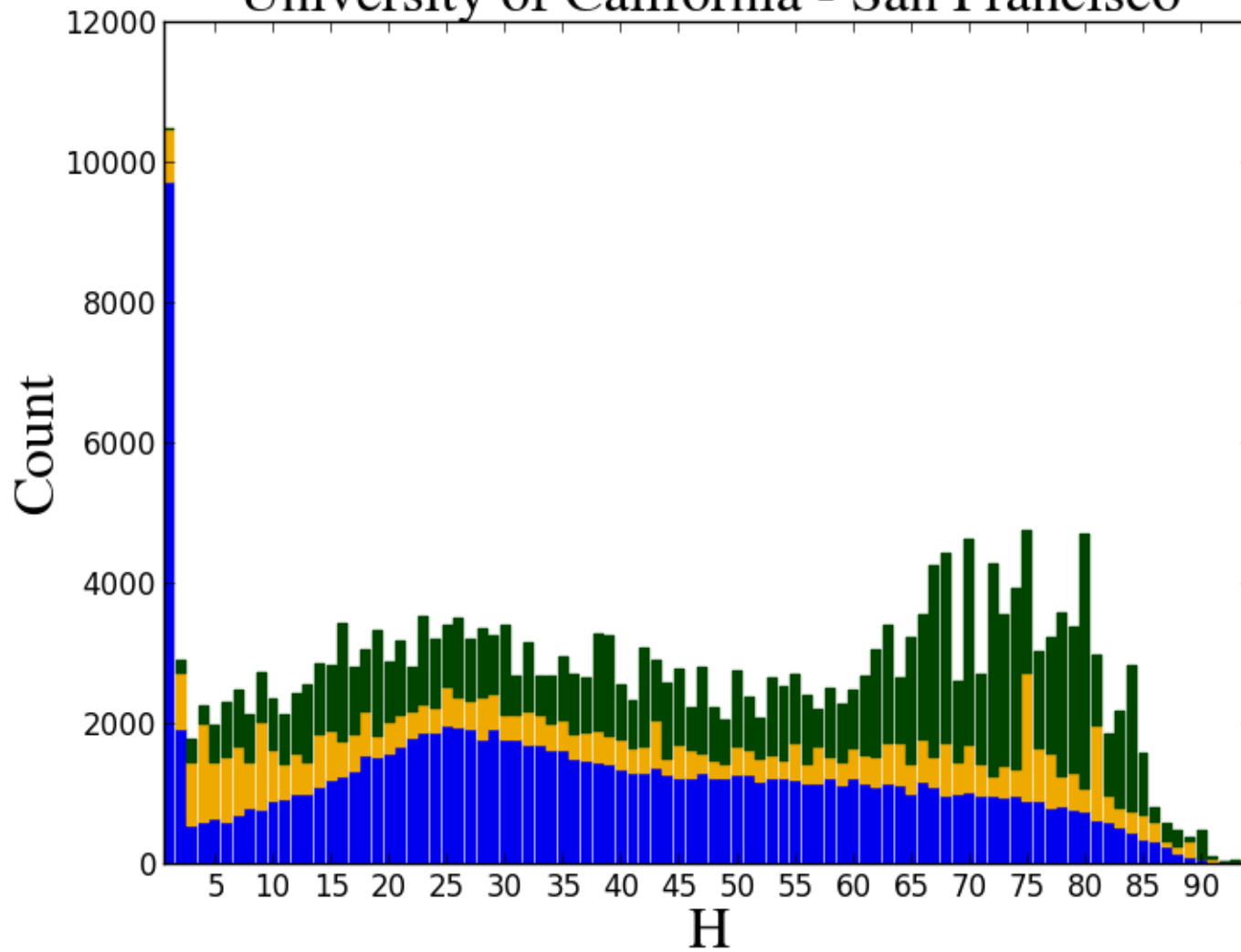
# University of California - Santa Cruz



# University of California - San Diego



# University of California - San Francisco



# Government Documents Initiative

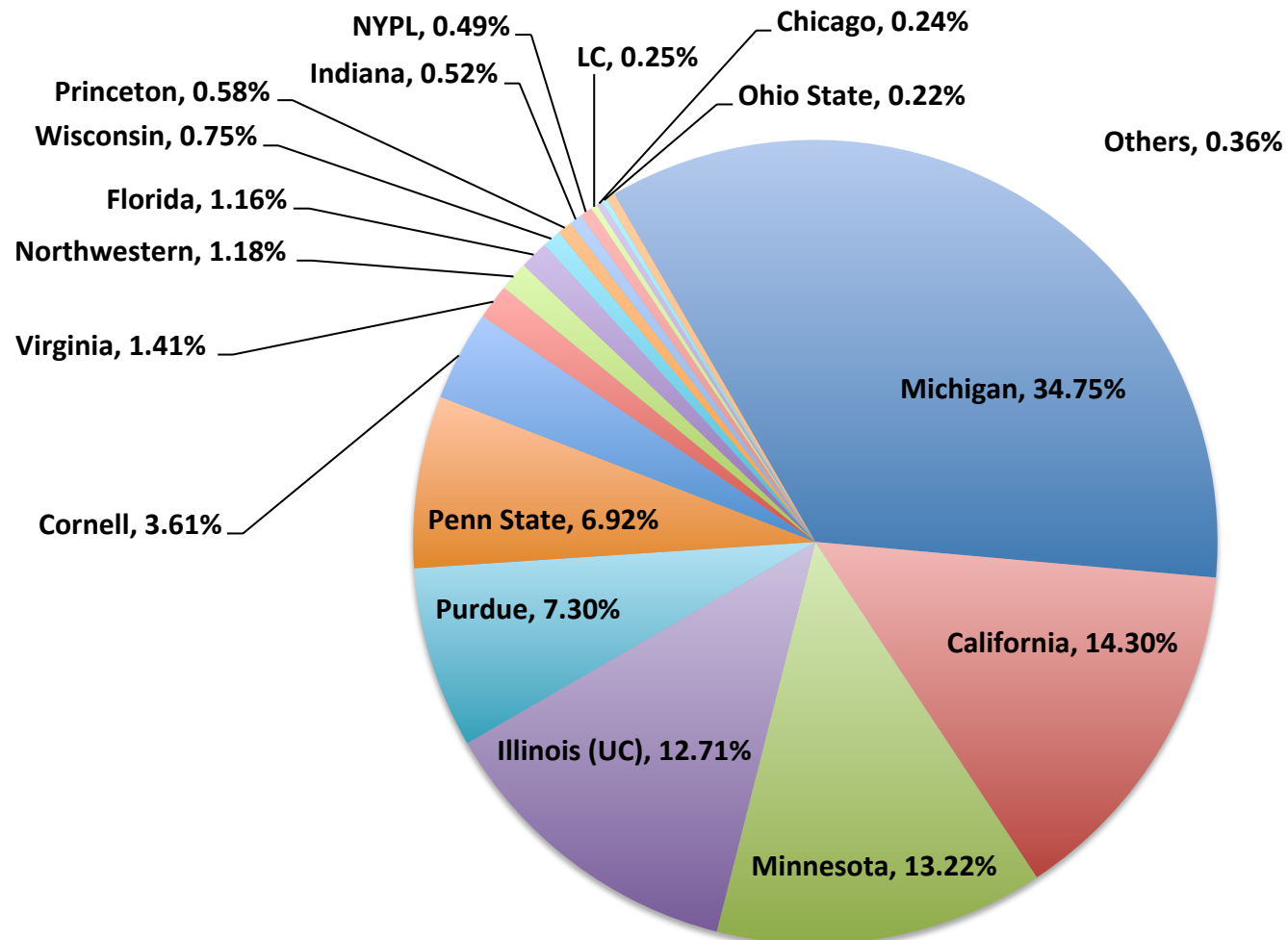
---

- Ballot Initiative: provide “expanded coverage & enhanced access to U.S. Government Documents.”
- Activities:
  - Developing a registry of US Federal Government Documents
  - Locate materials for inclusion in the collections
  - Improve search and discovery

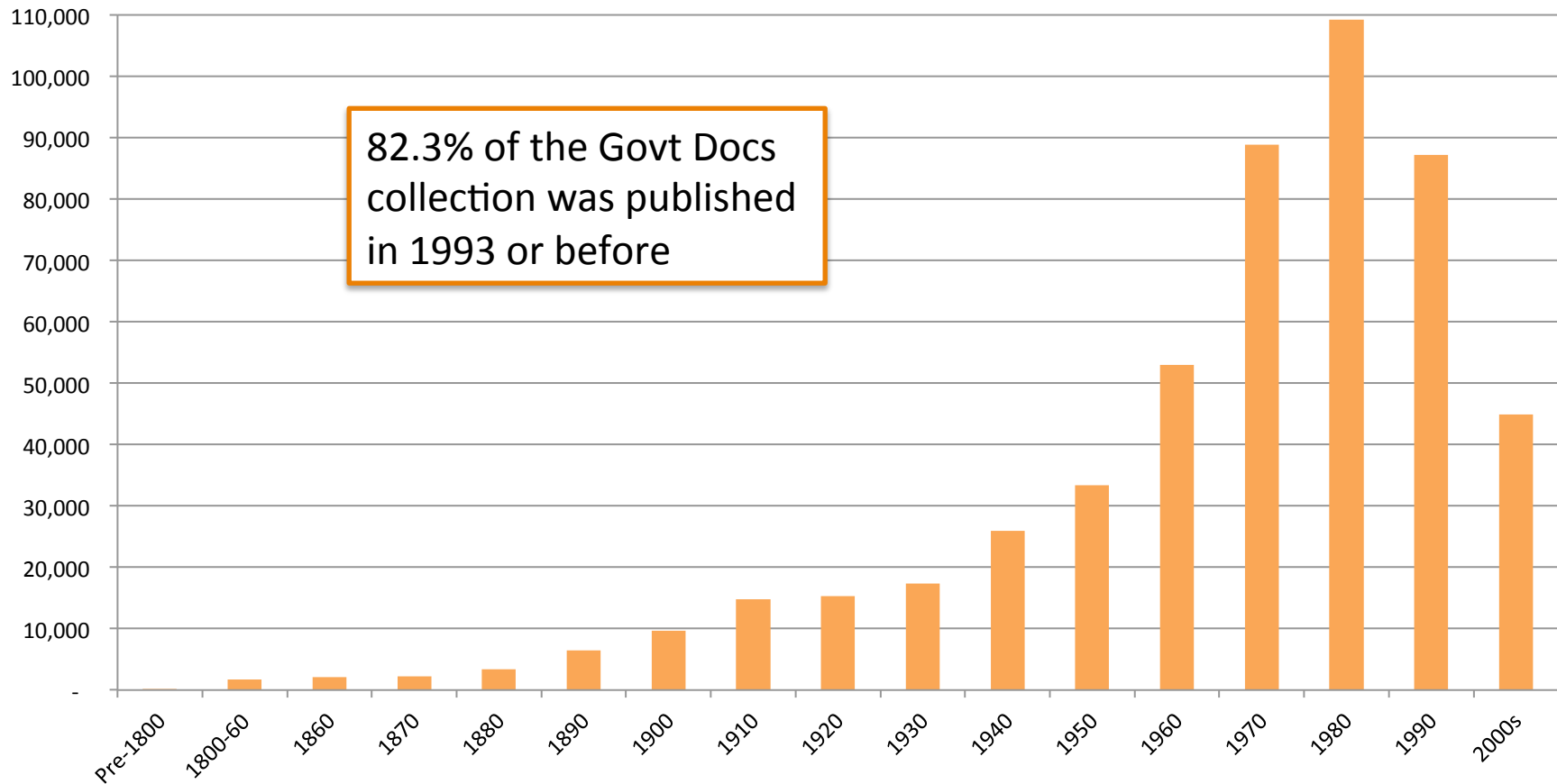




# US Gov't Publications by Source Library



# US Gov't Publications by Date



# The Registry

---

- Goal: “....include metadata for the comprehensive corpus of U.S. federal documents. This will include materials produced at U.S. government expense, in all formats, at the item level, from 1789 to the present.”
- Why?
  - Limited knowledge of this corpus.
  - Collection gap analysis
  - Digitization sourcing



# Near/Intermediate Term Activity

---

- Bibliographic and collections analysis
  - Registry and holdings work
- Focus first on known and cataloged materials
  - Prioritize print, post-1976 materials
  - Identify collections for inclusion (and get them)
  - Digitize where needed
- Publicize the efforts
  - Within the library community
  - To the general public



# Computational Access

---

- HathiTrust distributes public domain datasets
- HathiTrust Research Center
  - Developed collaboratively by Indiana University and University of Illinois; launched July 2011
  - Funding from the Sloan Foundation, Andrew W. Mellon Foundation, and NEH Office of Digital Humanities.
  - Partially Funded by HathiTrust (2014-2018)





RESEARCH CENTER

# Goals for the Research Center

- Research arm of HathiTrust
- Provide a persistent and sustainable structure to enable original and cutting edge research.
  - Leverage data storage and computational infrastructure at Indiana & Illinois
  - Stimulate community development of new functionality and tools
  - Use tools to enable discoveries that would not be possible without the HTRC
- Enable scholars to fully utilize content of HathiTrust Library while preventing intellectual property misuse within U.S. copyright law.
  - Provision secure computational and data environment for scholars to perform research using HathiTrust Digital LibraryIndiana University and University of Illinois



# Example Projects Supported by HTRC

---

- Muñoz, Trevor, University of Maryland. “Distributed Metadata Correction and Annotation.”
  - Correction, annotation and enhancement of HT records and export as linked data
- Page, Kevin, Oxford University. “EIEPHãT: Early English Print in HathiTrust, a Linked Semantic Workset Prototype”
  - Development of secondary worksets based on both HT and the Early English Books Online Text Creation Partnership (EEBO-TCP).
- Burton, Vernon. “The South as ‘Other,’ the Southerner as ‘Stranger.’”
  - Explore how attitudes expressed in print about slavery, southerners, and non-southerners have changed over both time and space.
- Ted Underwood, Associate Professor of English at the University of Illinois, Urbana-Champaign.
  - Using public domain texts received from HathiTrust to explore changing relationships in literary genres from 1700-1899.



# HTRC DataCapsule: Secure Access

Run all the demo codes there by clicking on "Cell" -> "Run All"

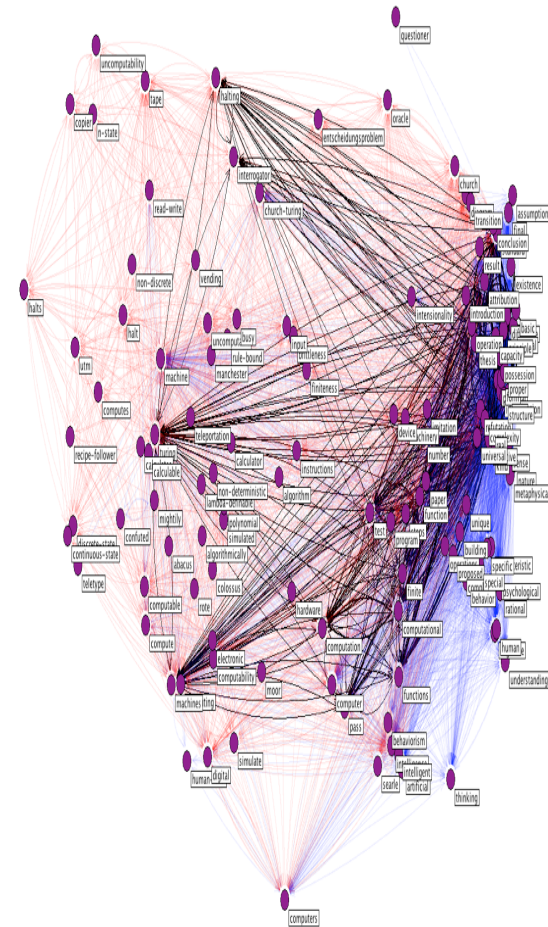
```
In [1]: # importing necessary libraries #  
from vsm.corpus import Corpus  
from vsm.model.ldacgsmulti import LdaCgsMulti as LDA  
from vsm.viewer.ldagibbsviewer import LDAGibbsViewer  
  
In [3]: # Uploading a saved Corpus object.  
plain_dir = '/home/demouser/demo/vsm/'  
c = Corpus.load(plain_dir + 'uc2.ark+=13960=t5w66bs1h-nltk-freq3.npz')  
Loading corpus from /home/demouser/demo/vsm/uc2.ark+=13960=t5w66bs1h-nltk-freq3.npz  
  
In [7]: # Building an LDA model #  
# LDA model takes a Corpus object,  
# context type (what we want to consider as documents),  
# and number of topics, K.  
lda = LDA(c, 'page', K=20)  
  
In [8]: # Training an LDA model #  
# number of iterations and number of processors (with  
# the multi-processing model) could be specified.  
lda.train(itr=20, n_proc=5)  
Iteration 0: log prob=-1147.156238  
Iteration 1: log prob=-243161.382093  
Iteration 0: log prob=-1147.156238  
Iteration 1: log prob=-243161.382093
```





# Scholarly Commons User Support Services

- Develop training materials
- Educational workshops
- Tool and workset support
- Collaborate with librarians and DH centers at HT institutions
- Assist researchers in HTRC text data mining research projects
- Collaboration: University Libraries, Illinois and Indiana



# Advanced Collaborative Support Awards

---

- **Detecting Literary Plagiarisms: The Case of Oliver Goldsmith.** Douglas Duhaime. University of Notre Dame: *....developing tools for detecting plagiarisms...to detect the literary thefts of Goldsmith.*
- **Taxonomizing the Texts: Towards Cultural-Scale Models of Full Text.** Colin Allen, Jaimie Murdock. Indiana University Bloomington. *...a cultural-scale investigation and topic modeling....random sampling to select collections according to the Library of Congress Subject Headings (LCSH).*
- **The Trace of Theory.** Geoffrey Rockwell, Laura Mandell, Stefan Sinclair, Matthew Wilkens, Susan Brown. University of Alberta, Texas A&M University, University of Notre Dame. *...aim to subset theoretical subsets from the HT public corpus and apply large-scale topic modeling... develop tools and computational methods for tracking the concept of "theory".*
- **Dr. Michelle Alexopolous**, University of Toronto...tracking technology diffusion through time using the HT corpus.



# HTRC UnCamp

---

- MAR 30-31, 2015, Ann Arbor, MI
  - Workshops, speakers, demonstrations
- Keynotes
  - Michelle Alexopoulos, Professor, University of Toronto  
March 30, 2015, 8:45 to 9:45 am
  
  - Erez Lieberman Aiden, Assistant Professor, Baylor College of Medicine  
March 31, 2015, 11:00 am to 12:00 pm

[http://www.hathitrust.org/htrc\\_uncamp2015](http://www.hathitrust.org/htrc_uncamp2015)



# Some Thoughts on the Present and Future



# How are we positioned?

---

- Our mission, collection, and the repository operations are all strong.
- Our brand reputation is outstanding.
- Our work is solidly supported by the law.
- We have expanded access in unprecedented ways.
- The partnership provides a solid base for action.
- We have very important programs underway.



# What needs thought?

---

- Strategy, mission, and role in the future
  - (Inter)National digital infrastructure
  - Public policy
  - Membership growth
  - Collections program
  - Services portfolio
- Organizational
  - Engagement with researchers and libraries
  - Enabling more participation in plans and action
  - Standing on our own



# Assumptions

---

- Our actions must align with the mission, goals, and purpose across our partnership.
- A few additional assumptions
  - We should pursue complementarity and cooperation, not competition and duplication.
  - Scale will continue to drive our strategies
  - Potential partners are not just other libraries and library organizations, but also readers, authors, publishers.



# How to find out more

---

- About: <http://www.hathitrust.org/about>
- Resources: <http://www.hathitrust.org/resources>
- Twitter: <http://twitter.com/hathitrust>
- Facebook: <http://www.facebook.com/hathitrust>
- Monthly newsletter:
  - <http://www.hathitrust.org/updates>
  - RSS [http://www.hathitrust.org/updates\\_rss](http://www.hathitrust.org/updates_rss)
- Contact us: [feedback@issues.hathitrust.org](mailto:feedback@issues.hathitrust.org)
- Blogs: <http://www.hathitrust.org/blogs>
  - Large-scale Search
  - Perspectives from HathiTrust





Thank you!

furlough@hathitrust.org  
@MikeFurlough

