Research Data Management – An Institutional Perspective

Patricia Rankin
Associate Vice Chancellor for Research

University of Colorado
Boulder

Outline

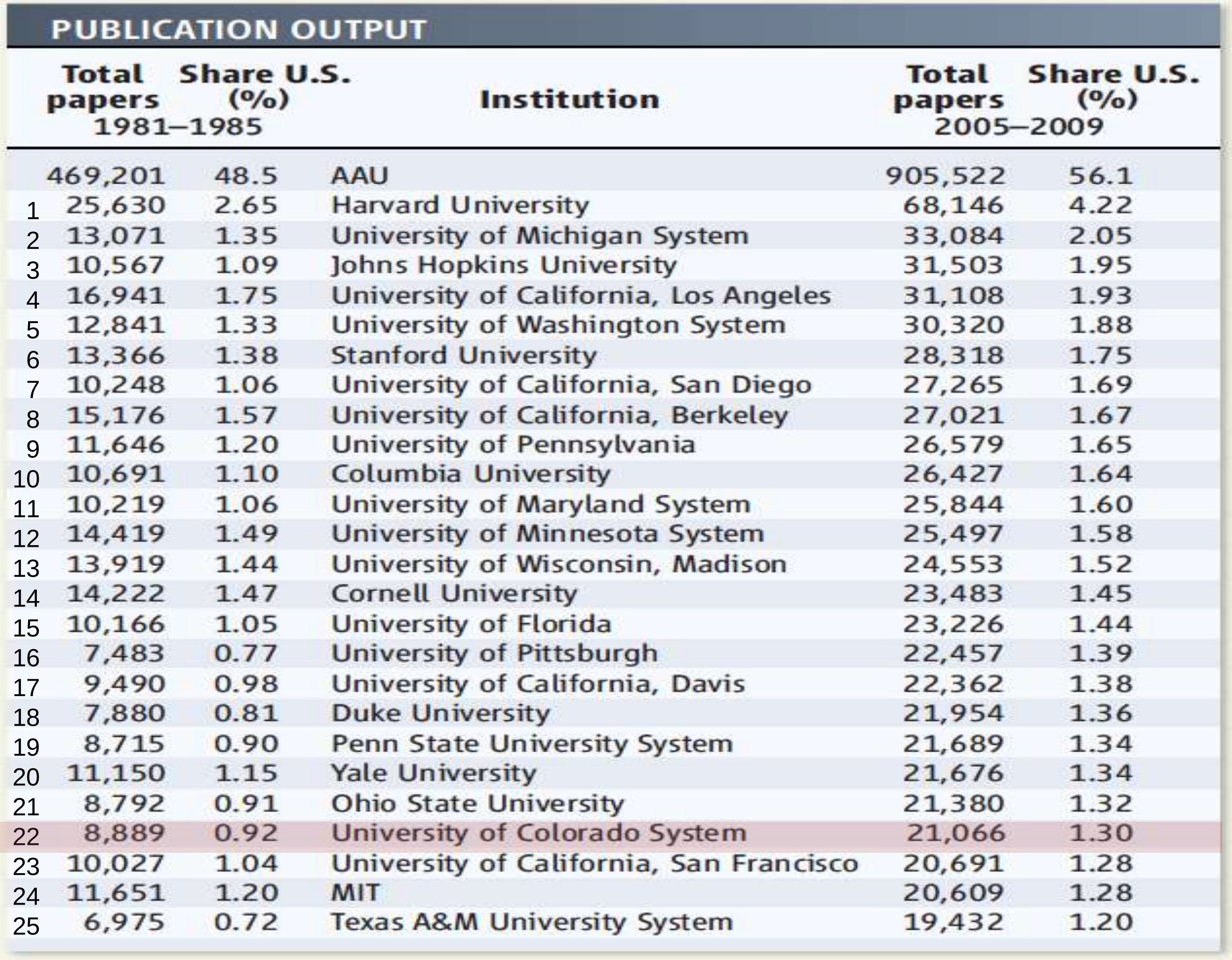
- **Overview of University of Colorado Boulder**
  - – AAU member, Research Intensive
  - – Limited State Support
- **Open Questions**
  - – Shifting Paradigms
- **Some Ideas**
  - – Carrots work better than sticks….

Sponsored Research at CU

- $351.9 million in federally sponsored research (FY 2013)
- Annual research awards have roughly doubled over the last ten years
- Publish about 4,800 articles a year
- Lead the publics in NASA funding
- More atmospheric scientists per square mile than anywhere else in US
- Undergraduate (800+) and graduate students (1,160) participate in research

41% of sponsored research revenue goes to local salaries.

Approximately half of U.S. research output is generated by 25 universities

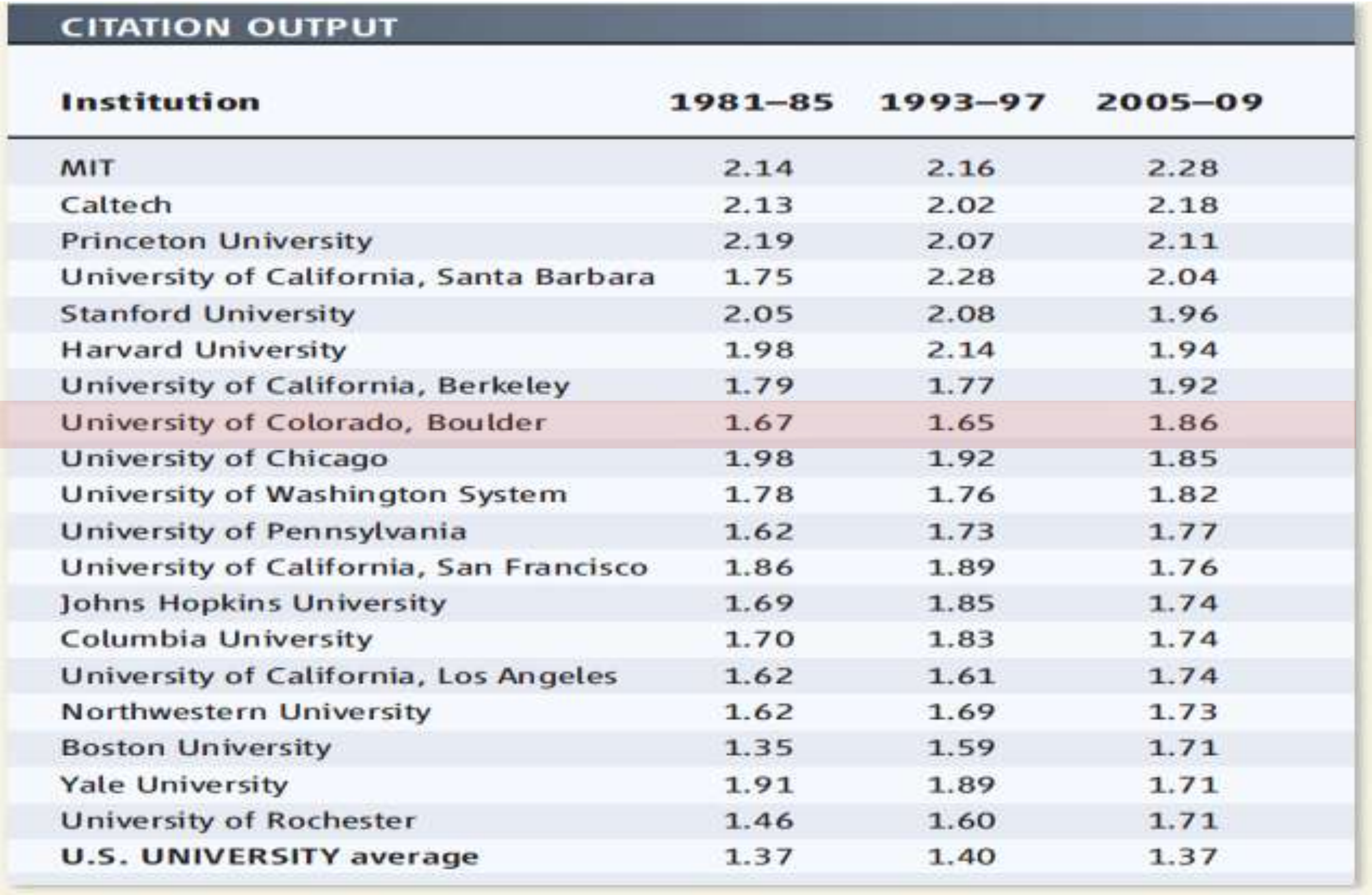


PUBLICATION OUTPUT

	Total papers	Share U.S. (%)	Institution	Total papers	Share U.S. (%)
	1981–1985			2005–2009	
	469,201	48.5	AAU	905,522	56.1
1	25,630	2.65	Harvard University	68,146	4.22
2	13,071	1.35	University of Michigan System	33,084	2.05
3	10,567	1.09	Johns Hopkins University	31,503	1.95
4	16,941	1.75	University of California, Los Angeles	31,108	1.93
5	12,841	1.33	University of Washington System	30,320	1.88
6	13,366	1.38	Stanford University	28,318	1.75
7	10,248	1.06	University of California, San Diego	27,265	1.69
8	15,176	1.57	University of California, Berkeley	27,021	1.67
9	11,646	1.20	University of Pennsylvania	26,579	1.65
10	10,691	1.10	Columbia University	26,427	1.64
11	10,219	1.06	University of Maryland System	25,844	1.60
12	14,419	1.49	University of Minnesota System	25,497	1.58
13	13,919	1.44	University of Wisconsin, Madison	24,553	1.52
14	14,222	1.47	Cornell University	23,483	1.45
15	10,166	1.05	University of Florida	23,226	1.44
16	7,483	0.77	University of Pittsburgh	22,457	1.39
17	9,490	0.98	University of California, Davis	22,362	1.38
18	7,880	0.81	Duke University	21,954	1.36
19	8,715	0.90	Penn State University System	21,689	1.34
20	11,150	1.15	Yale University	21,676	1.34
21	8,792	0.91	Ohio State University	21,380	1.32
22	8,889	0.92	University of Colorado System	21,066	1.30
23	10,027	1.04	University of California, San Francisco	20,691	1.28
24	11,651	1.20	MIT	20,609	1.28
25	6,975	0.72	Texas A&M University System	19,432	1.20

Source: Mervis, *Science* Vol 330: 1032, 2010

Approximately half of U.S. research citations generated by 19 universities



CITATION OUTPUT

Institution	1981–85	1993–97	2005–09
MIT	2.14	2.16	2.28
Caltech	2.13	2.02	2.18
Princeton University	2.19	2.07	2.11
University of California, Santa Barbara	1.75	2.28	2.04
Stanford University	2.05	2.08	1.96
Harvard University	1.98	2.14	1.94
University of California, Berkeley	1.79	1.77	1.92
University of Colorado, Boulder	1.67	1.65	1.86
University of Chicago	1.98	1.92	1.85
University of Washington System	1.78	1.76	1.82
University of Pennsylvania	1.62	1.73	1.77
University of California, San Francisco	1.86	1.89	1.76
Johns Hopkins University	1.69	1.85	1.74
Columbia University	1.70	1.83	1.74
University of California, Los Angeles	1.62	1.61	1.74
Northwestern University	1.62	1.69	1.73
Boston University	1.35	1.59	1.71
Yale University	1.91	1.89	1.71
University of Rochester	1.46	1.60	1.71
U.S. UNIVERSITY average	1.37	1.40	1.37

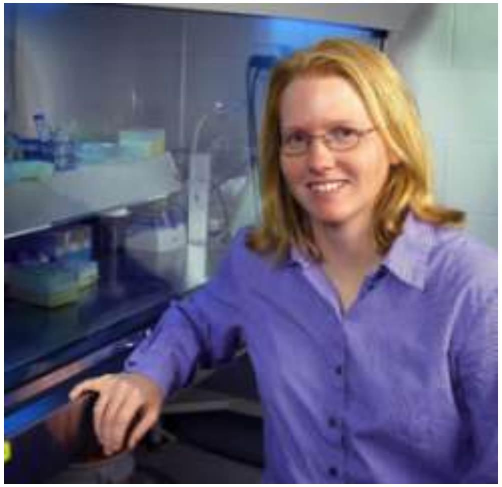
Source: Mervis, Science Vol 330: 1032, 2010

Research Initiatives

Key Areas

- Aerospace Sciences and Engineering
- Biotechnology and Biosciences
- Renewable and Sustainable Energy
- Geosciences/Environmental Sciences
- Computational Sciences
- STEM Education (Science, Technology, Engineering, Mathematics)

World-class interdisciplinary research at CU-Boulder advances society and the economy.



Computational Sciences

A Broad Spectrum of Faculty Partner with Universities, Government and Industry in:

- High-performance scientific computing
- Artificial intelligence
- Nanotechnology
- Next-generation internet
- Biotechnology
- Genomics
- Fluid dynamics
- Climate modeling
- Laser sciences

A great deal of research in the science and engineering disciplines is driven by simulations, requiring significant advances in computational technologies.

Data sets include

- Artifacts from Indian tribes in arctic regions
- Bee population studies
- Sounds from endangered languages
- NMR scans
- Ice Cores
- Collision data from LCH Higg’s search and reconstructed events
- Musical Performances
- Genomic studies
- Simulations of likely material behaviors

General guidelines

- Major collaborations and networks tend to have discipline specific archives
- Some agencies require data to be stored in specific repositories
- In many cases computing/data management is delegated to a postdoc or graduate student (aka “technically savvy native”)
- Many assumed technically savvy natives are not (and often information does not cross the barrier when a postdoc or graduate student moves on)

Big Data

- Universities are becoming major consumers of analytics
  - Research Productivity/Rankings
  - Student Retention “Smart” systems
- What questions can we answer because we have
  - Access to larger data sets?
  - Better ways to connect data sets?
  - More compute power?
- Who gets to use the data?
- Who sets the standards for allowed use?

Changing Times

- Federal funding of basic research is increasingly becoming a political issue
  - Economic Driver/Translational Research
  - Value of Social Science
  - Distrust of “expert” opinion
    - ***Data produced in research funded by the public should be available to public***
    - ***Results of research should be broadly disseminated/easily available***
      - Data Management Plans, Open Access

Details Matter

- NSF has indicated that people can budget for data management plans in their proposal requests… *but*
  - Budgets not growing to accommodate extra demands
  - Not clear that quality of data management plans matters to many reviewers yet
  - Communities still working to define data management plan standards

Questions I have

- How long do researchers get to keep data private?
  - IP issues, publication rights
- How do we determine a sensible amount of time to preserve data for?
  - Some simulations that took a few months some years back can be redone in a fraction of the time
  - The raw data may require analysis code that has evolved over time
- What happens if a researcher’s data management plan requires campus level resources that they don’t ask for in advance?
- Who pays once grant has ended?

More questions

- And how do we deal with publications based on data that are not high quality/do not meet discipline standards?
  - Statistics
  - Data selection
  - Equal time or proportional representation?
    - ***BBC in UK has changed policy on allowing all sides in a debate to speak…***

Open Access

- How does this impact tenure/promotion?
  - How do we figure out merit factors for open access journals?
    - ***Peer Review***
    - ***Quality of other papers published***
    - ***Long term reliability, reputation, accumulated social capital***
  - How does providing a data set weigh towards tenure/promotion?
    - ***Reward what we value***
- How do we sustain?
  - $2K publishing fee multiplied by 4,800 articles…not going to work

Peerage of Science

- www.Peerageofscience.org
- Interesting model
  - Authors submit manuscripts and deadlines for four stages
    - ***Reviews***
    - ***Peer Review of Peer Review (reviews get a quality index)***
    - ***Manuscript Revision***
    - ***Final Evaluation – breadth, impact, originality, data, methods, inference, literature coverage – leads to a quality index***

Next Steps

- Subscribing journals can offer to publish or authors can choose to submit to another journal (that journal can have access to existing reviews)
- Quality indexes include article quality, number of reviews, quality of reviews
- Issues
  - Seems mostly bio related right now
  - Early days – will be interesting to see adoption rate

What we are doing at CU Boulder?

- Research Computing – reports to Office of Information Technology and Office of Vice Chancellor for Research
  - Regular meetings between Head of Research Computing and Associate Vice Chancellor for Research
  - Regular meetings of both with Library leadership
- Research Data Management Task Force
- Data Management Audit

Research Data Advisory Committee

- Mix of disciplines and roles
  - – Co-chairs from English, Evolutionary Biology
  - – Research Staff, Library Staff
  - – Looking to add post-doc, graduate student
- Goal – to develop definitions (what is “data”), policies, best practices, campus outreach

Data Management Plans

- Now required for campus competitions (competitions run to select CU nominee if have a restriction on allowed number of proposals)
- Seed grant competition
  - About 80 proposals from across campus
  - RDAC Committee analyzed data managements plans
    - ***Not a selection criteria this year – will be next***
    - ***Lots of information on current state of data management practices – lots of room to improve***

So -

- Running a competition to search for the best data management plans
  - 5 broad areas – including arts and humanities, social sciences
  - Open to graduate students, post docs, and faculty
  - Encouraging use of tools available to develop data plans, review of best practices documents developed from seed grant study

Closing Words

- Data Management is an emerging field
- Interesting mix of technical, social issues
  - How do we store
  - What do we store
  - Why do we store
  - How do we use
  - When do we delete
- Important to form broad alliances