# The University Libraries as Data Science Research Partner

Dr. Bruce Herbert, Director

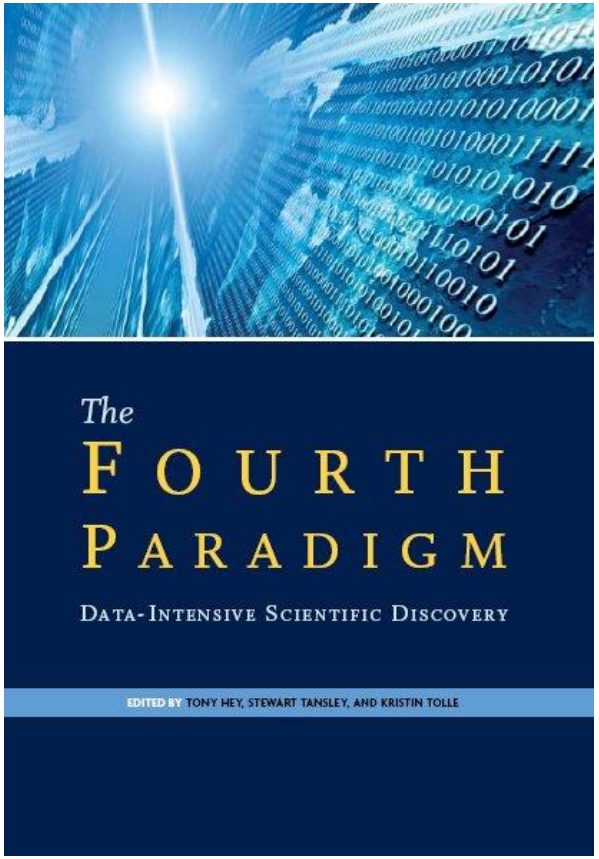
Office of Scholarly Communications

# The Fourth Paradigm: Data-Intensive Scientific Discovery

"In Jim Gray's last talk to the Computer Science and Telecommunications Board on January 11, 2007, he described his vision of the fourth paradigm of scientific research.

He outlined a two-part plea for the funding of tools for data capture, curation, and analysis, and for a communication and publication infrastructure.

He argued for the establishment of modern stores for data and documents that are on par with traditional libraries."




*http://research.microsoft.com/en-us/collaboration/fourthparadigm/*

Role of Libraries:

# A Transformed Scientific Method

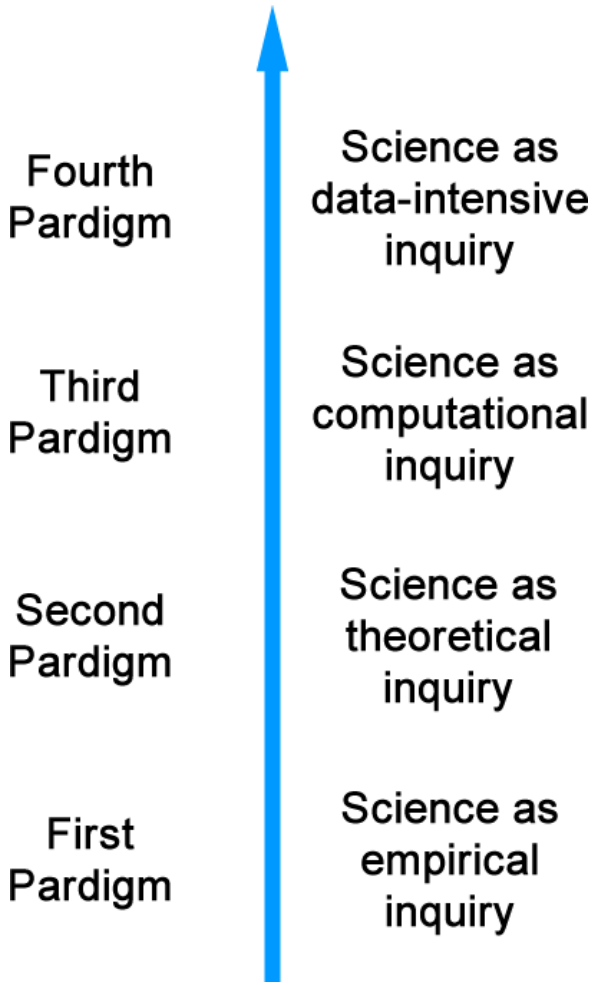
***Research***

*Ontological or data model frameworks*

*Library Collections as Big Data*

*Institutional Research*

*Scholarly Practices*




***Services***

*Data Management*

*Data Curation*

*Publishing/Dissemination*

University Libraries as:

**Service Provider**

Source of Big Data

Research Partner

# Federal Mandates For Public Access to Research

**The Library Supports:**

Publication repositories

Training

Funding for OA publication fees

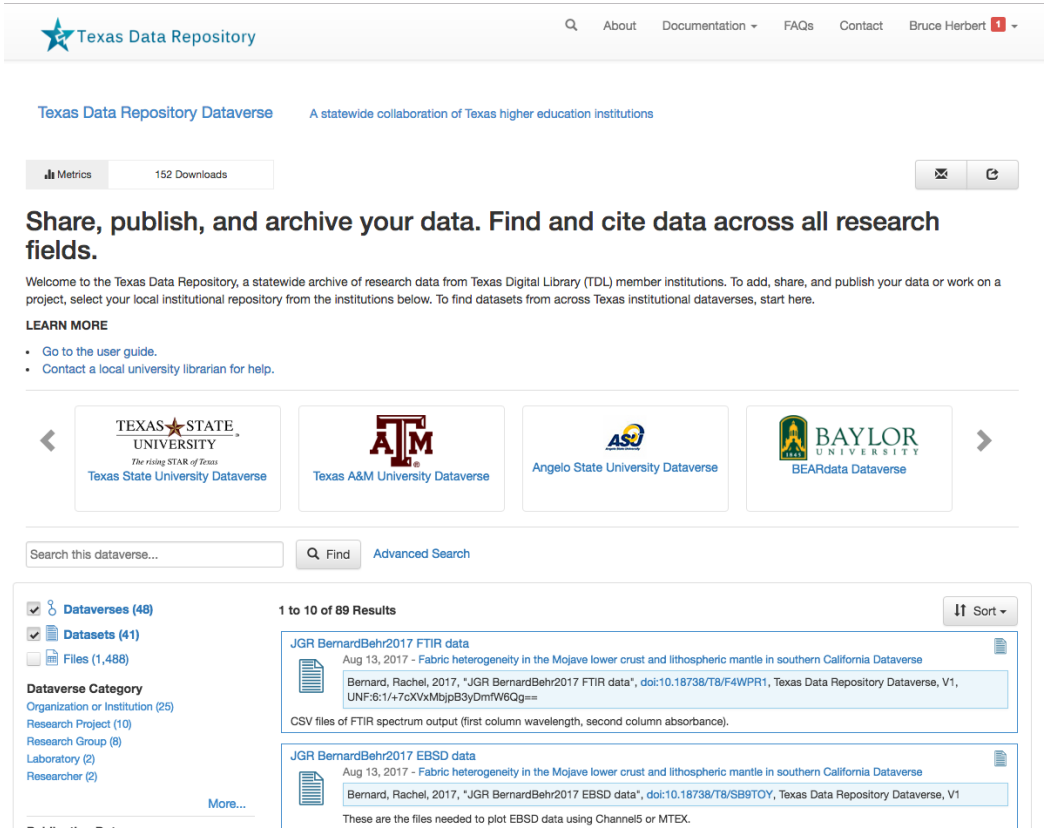
Data repositories

http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

LIBRARIES
TEXAS A&M UNIVERSITY

Research Innovation:

# Data Management




Data is citable

Meets funder mandates

Supports move to open science

Link graduate student data to ETDs in TAMU repository

Library training programs

https://data.tdl.org/

LIBRARIES
TEXAS A&M UNIVERSITY

University Libraries as:

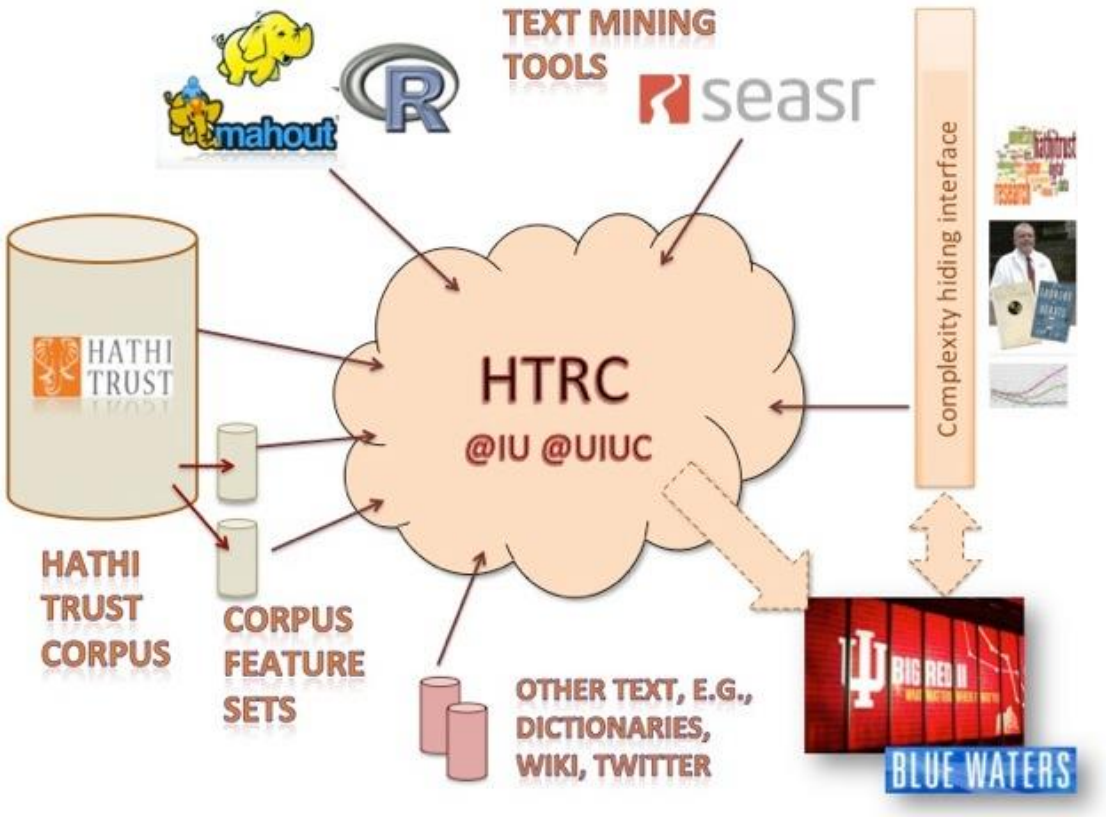
Service Provider

**Source of Big Data**

Research Partner

Library Collections as Data:

# Text Mining Library Collections



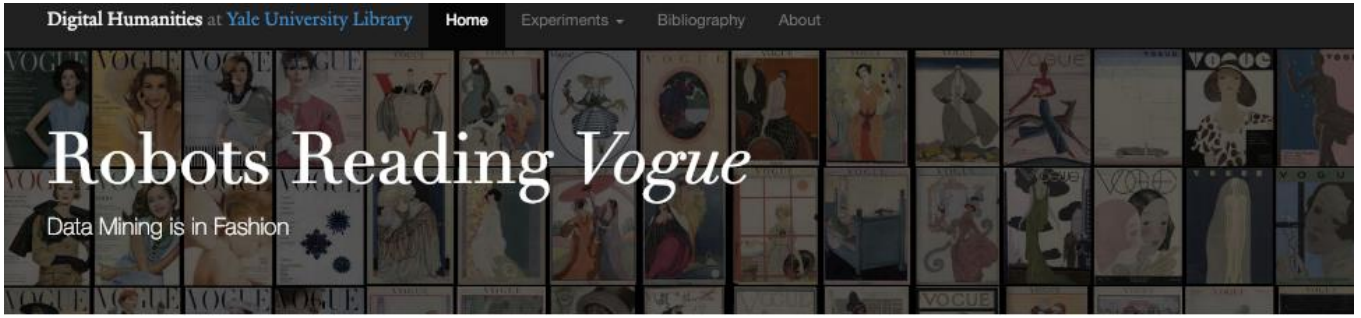

HATHI TRUST Research Center

## HathiTrust Research Center

https://analytics.hathitrust.org/

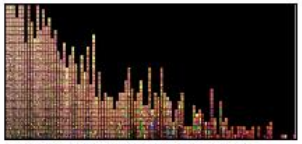
Library Collections as Data:

# Text Mining Library Collections




Few magazines can boast being continuously published for over a century, familiar and interesting to almost everyone, full of iconic pictures — and also completely digitized and marked up as both text and images. What can you do with over 2,700 covers, 400,000 pages, 6 TB of data? Students, librarians and faculty are excited about the possibilities of working with *Vogue* to explore questions in fields from gender studies to computer science. We highlight some early experiments below:

## Slice Histograms



Direct visualization of color patterns.

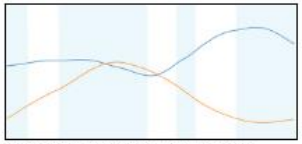
Explore »

## Cover Averages

Visual continuity and change across the decades.

Explore »

## *n*-gram Search



Search and compare word usage in 400,000+ pages within ads, articles, or all texts.

Explore »

## Topic Modeling

Self-organizing themes determined by word coöccurrence.

Explore »

## Advertisements

Sort ads by frequency, date, and industry.

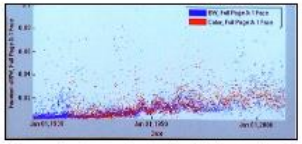
Explore »

## Statistics

View figures on circulation, ratio of articles to advertisements, price per issue, and number of pages per year.

Explore »

## Student Work



See projects conducted by Yale students using this data.

Explore »

## Colormetric Space

Hue, Saturation and Lightness.

Explore »

http://dh.library.yale.edu/projects/vogue/

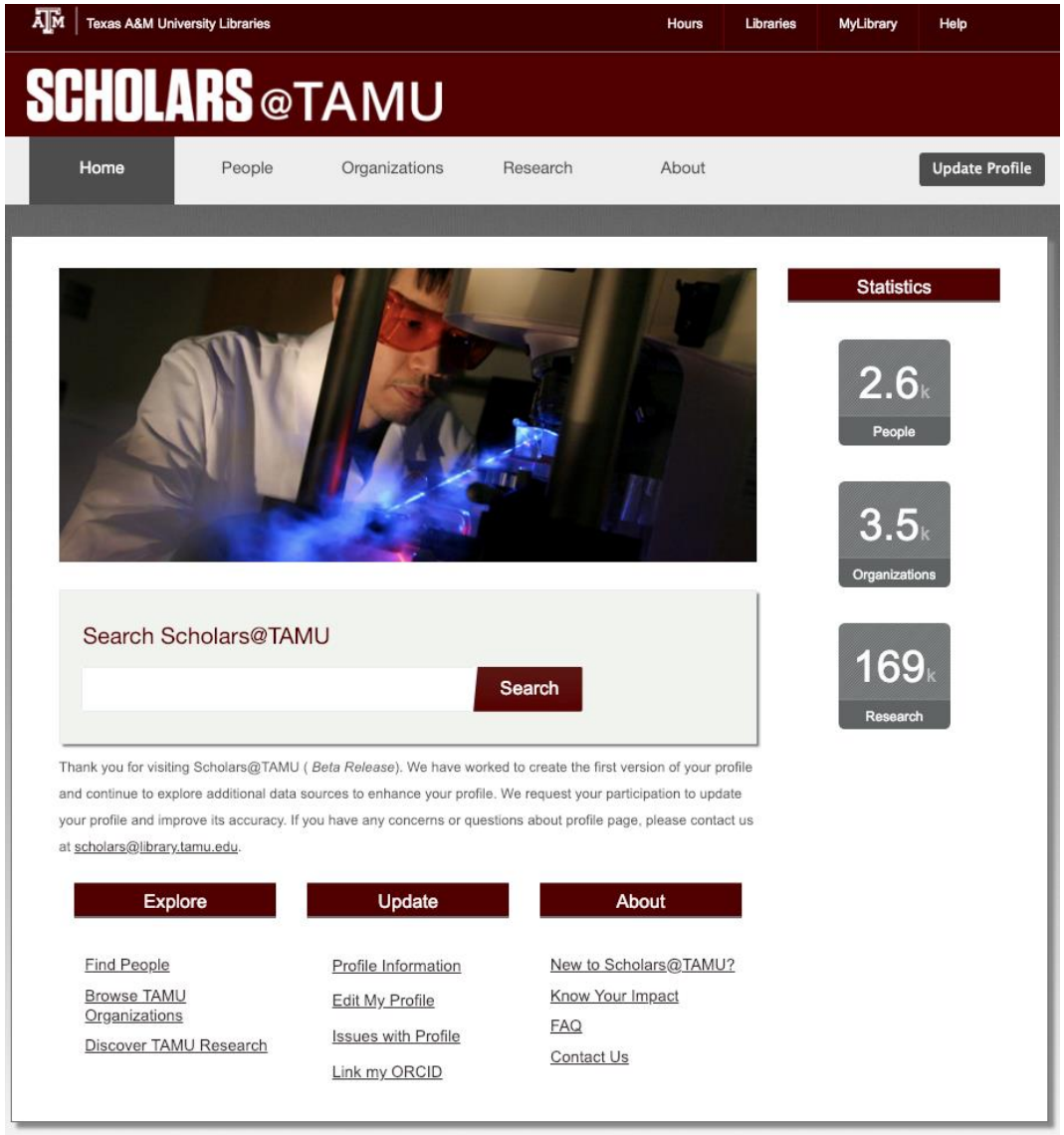
University Libraries as:

Service Provider

Source of Big Data

**Research Partner**

# Institutional Research



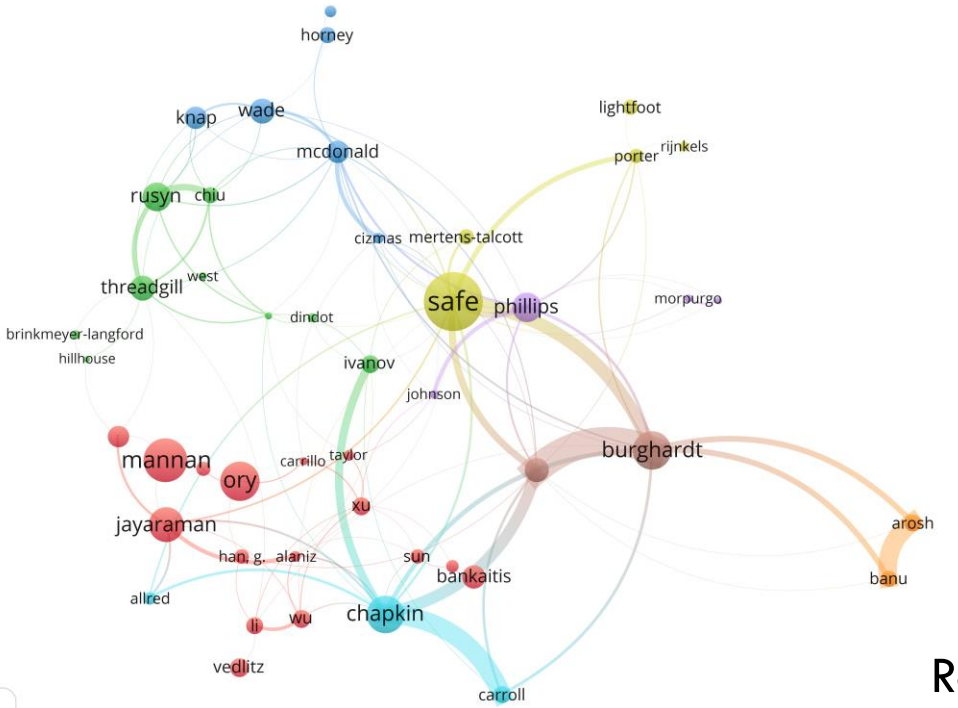

Scholars@TAMU

- Integrated Impact Metrics
- Publications
- Grants and Awards
- Teaching
- Thesis & Dissertations

http://scholars.tamu.edu/

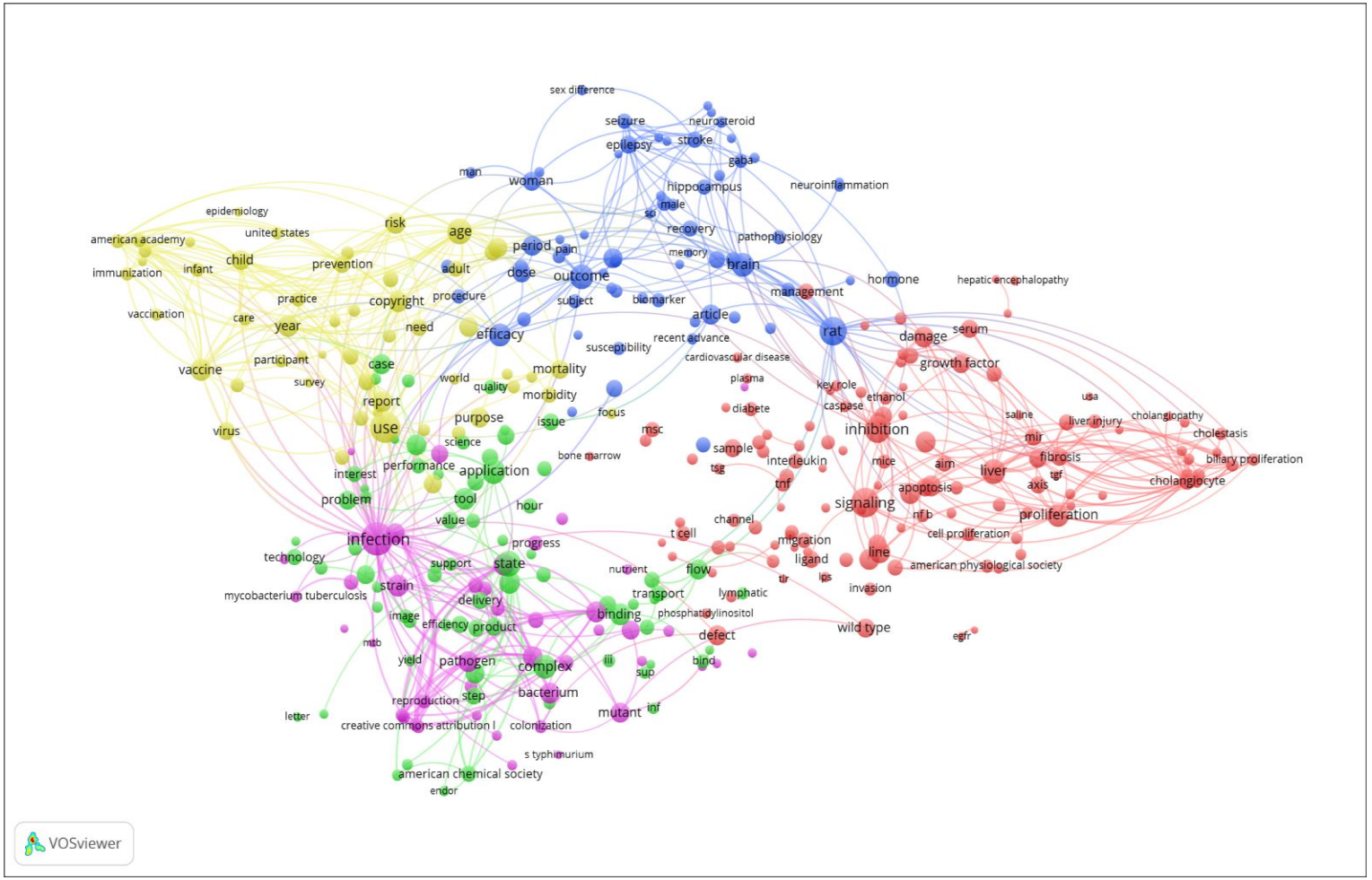
LIBRARIES
TEXAS A&M UNIVERSITY

# Institutional Research




Research Strengths
Scholarly Impact & Reputation
Collaboration Networks
Societal Impact

# Institutional Research




Network visualization of research areas in the College of Medicine

# Impact of Team Collaboration

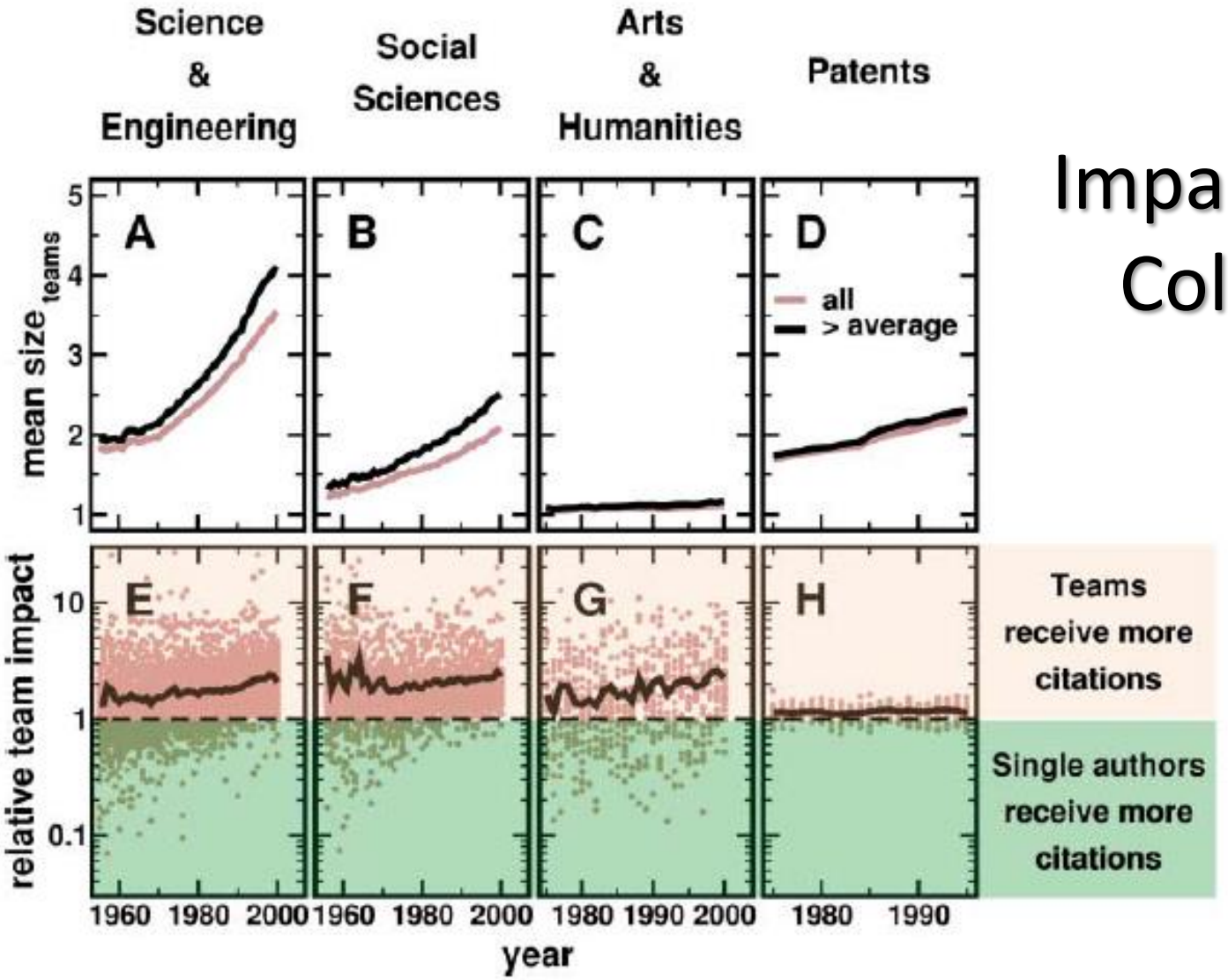



**Fig. 2.** The relative impact of teams. (**A** to **D**) Mean team size comparing all papers and patents with those that received more citations than average in the relevant subfield. (**E** to **H**) The RTI, which is the mean number of citations received by team-authored work divided by the mean number of citations received by solo-authored work. A ratio of 1 indicates that team- and solo-authored work have equivalent impact on average. Each point represents the RTI for a given subfield and year, whereas the black lines present the arithmetic average in a given year.

Wuchty et al. (2007). Science 316(5827): 1036-1039.

LIBRARIES
TEXAS A&M UNIVERSITY

# Collaboration Across Institutions

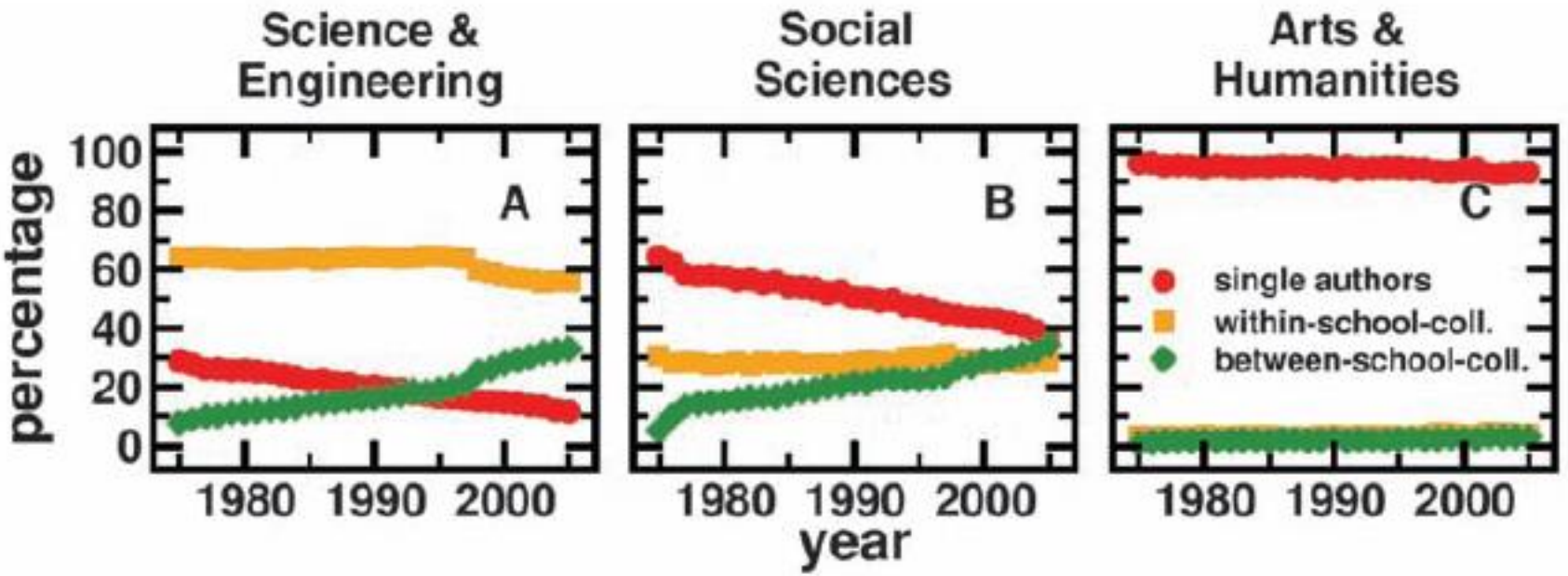



**Fig. 1.** The rise in multi-university collaboration. By comparing the incidence of papers produced by different authorship structures, we see that the share of multi-university collaborations strongly increases from 1975 to 2005. This rise is especially strong in SE (**A**) and SS (**B**), whereas it appears weakly in AH (**C**), in which collaboration of any kind is rare. The share of single-university collaborations remains roughly constant with time, whereas the share of solo-authored papers strongly declines in SE and SS.

Jones et al. (2008). Science 322: 1259-1262.

University Libraries as:

**Service Provider**
**Source of Big Data**
**Research Partner**

Dr. Bruce Herbert
beherbert@tamu.edu

LIBRARIES
TEXAS A&M UNIVERSITY