

# Bridging Disciplines: Assessing the Interdisciplinary Impact of Open Data

*Robert R. Downs<sup>1</sup> and Robert S. Chen<sup>2</sup>*

<sup>1</sup> [rdowns@ciesin.columbia.edu](mailto:rdowns@ciesin.columbia.edu)

<sup>2</sup> [bchen@ciesin.columbia.edu](mailto:bchen@ciesin.columbia.edu)

NASA Socioeconomic Data and Applications Center (SEDAC)  
Center for International Earth Science Information Network  
(CIESIN)  
The Earth Institute, Columbia University

IASSIST 2015

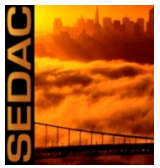
Minneapolis, Minnesota, 2-5 June 2015

Session: Impact of Data Citation  
Wednesday 3 June 2015, 1:15 p.m.





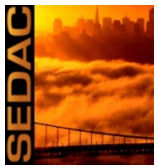
# Abstract



Freely disseminating scientific data can contribute to multiple disciplines across the physical, social, health, and engineering sciences. If the impact of data centers is not measured, stakeholders will not know whether data centers, archives, and libraries, and the data that they disseminate, are having a positive impact on the conduct of science. **Data citations provide evidence** on the use of data in various stages of the research process, including problem definition, statistical analysis, modeling, and validation. Measuring the interdisciplinary citation of scientific data disseminated by a data center can reveal the degree to which the data center is supporting cross-disciplinary research. Analysis of a **decade of data citations demonstrates the interdisciplinary use** of scientific data and the impact that one data center has had across disciplinary boundaries.



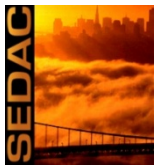
# Potential Impact of Scientific Data



- Science
  - Methods and results
  - Disciplinary and interdisciplinary advancements
- Policy
  - Government decision-making
  - International relations
- Economics
  - Private decision-making
  - Innovation and efficiency
- Education
  - Formal: K-12, undergraduate, graduate
  - Informal: home schooling, lifelong learning
- Journalism and Mass Media
  - Informing those who inform the general public and public discourse



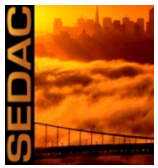
# Impact of Scientific Data on Science



- Contributions to Science
  - Creating knowledge in one or more disciplines
- Metric: data citations in scientific literature
  - Journal articles, book chapters, working papers, and other scientific publications (e.g., assessment reports, NRC studies, reference works)
  - Secondary citation of literature
- Possible Alternative Metrics: evidence of data usage
  - Data cited in science presentations
  - Data cited in science blogs and unpublished reports
  - Downloads of data and related information (documentation, maps)
  - Inclusion in science models, tools, services, apps, and catalogs
  - Views of data landing pages, descriptions, and metadata
  - Bookmarks and tweets by scientists and research organizations



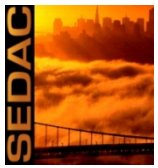
# Metrics for Other Potential Impacts of Scientific Data: Policy and Economics



- Contributions to Policy
  - Use of data to support government decision making
  - Metric: data citations in government reports (e.g., environmental impact assessments) and decision records; data citations in other reports prepared by stakeholders
  - Possible Alternative Metrics: downloads and views of data by government agencies; designations of scientific data as electronic records; use of scientific data in legal proceedings
- Contributions to Economics
  - Use of data to support private decision making, including improved efficiency, innovation in processes, reduced externalities
  - Metric: data citations in industry publications, reports
  - Possible Alternative Metrics: downloads, views, and permission requests by private sector entities; citations of data in patent filings.



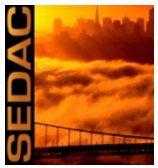
# Metrics for Other Potential Impacts of Scientific Data: Education and Media



- Contributions to Education
  - Improving knowledge of students and teachers
  - Metric: data citations in textbooks and published instructional resources including course syllabi and lab exercises
  - Possible Alternative Metrics: data cited in learning resources and tools; page visits and downloads by educators and students; data mentions in educational apps, bookmarks, tweets; formal and informal assessments of “data literacy”
- Journalism and Mass Media
  - Expanded data use in mass media and via online channels
  - Metric: data cited in newspapers, magazines, news programs, science programs (including radio, tv), science web sites, blogs, and other
  - Alternative Metrics: downloads and page visits; use of graphics based on data appearing in conventional and social media; data referenced in apps, bookmarks, and tweets



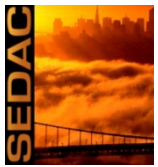
# Analyzing Disciplinary Impact of Data from a Scientific Data Center



- NASA Socioeconomic Data and Applications Center (SEDAC)
- Focus: “human interactions in the environment”
- Mission: “develop and operate applications that support the integration of socioeconomic and Earth science data and to serve as an "Information Gateway" between the Earth and social sciences”
- Currently disseminates 193 datasets in 36 online data collections
  - Each dataset receives extensive review and preparation prior to dissemination
- Users are encouraged to cite SEDAC data used for a publication
  - Each dataset landing page contains a recommended citation for the dataset



# Diverse SEDAC Data Collections

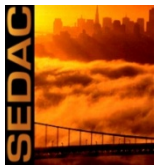


Anthropogenic Biomes	Global Reservoir and Dam (GRanD)	Low Elevation Coastal Zone (LECZ)
Archive of Census Related Products (ACRP)	Global Roads	Millennium Ecosystem Assessment (MA)
China Dimensions	Global Rural-Urban Mapping Project (GRUMP)	National Aggregates of Geospatial Data Collection (NAGDC)
Climate Effects on Food Supply	Gridded Population of the World (GPW), v3	Natural Disaster Hotspots
Compendium of Environmental Sustainability Indicators	Gridded Species Distribution	Natural Resource Management Index (NRMI)
Energy Infrastructure	Historical Anthropogenic Sulfur Dioxide Emissions	Population Exposure to Natural Disasters
Environmental Performance Index (EPI)	Human Appropriation of Net Primary Productivity (HANPP)	Poverty Mapping
Environmental Sustainability Index (ESI)	Indicators of Coastal Water Quality	Satellite-Derived Environmental Indicators
Environmental Treaties and Resource Indicators (ENTRI)	Intergovernmental Panel on Climate Change (IPCC)	Socioeconomic Downscaled Projections
Georeferenced Population Data sets of Mexico	Land Use and Land Cover (LULC)	Superfund Site Footprints
Global Agricultural Lands	Last of the Wild, v2	U.S. Census Grids
Global Fertilizer and Manure	Last of the Wild, v1	Urban Landsat





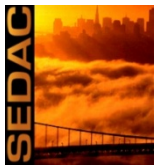
# Methodology



- **Collected citations of SEDAC data**
  - Obtained list of citing journal articles published during 2004 - 2013
- **Obtained classifications to identify disciplines of each citing journal**
  - Obtained Web of Science® (WoS) Category assignments
  - Obtained Web of Knowledge® Subject Classification of WoS Categories (WoK5.3) and Equivalent General Categories and Subject Areas
  - Obtained ScienceWatch® Field Definitions of Major Fields
- **Identified multidisciplinary use of SEDAC data**
  - Identified WoS Categories assigned to journals citing SEDAC data, and used Scopus®, journal titles, and publisher sites when Categories were not assigned
  - Paired assigned WoS Categories to Equivalent General Categories and Subjects
  - Identified Major Fields corresponding to assigned WoS Categories and Subjects
  - Normalized journals with WoS Categories, (WoK5.3) and Equivalent General Categories and Subject areas, and Field Definitions of Major Fields
  - Identified Categories, Subjects, and Major Fields of journals citing SEDAC data



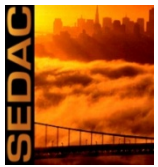
# Collection of SEDAC Data Citations



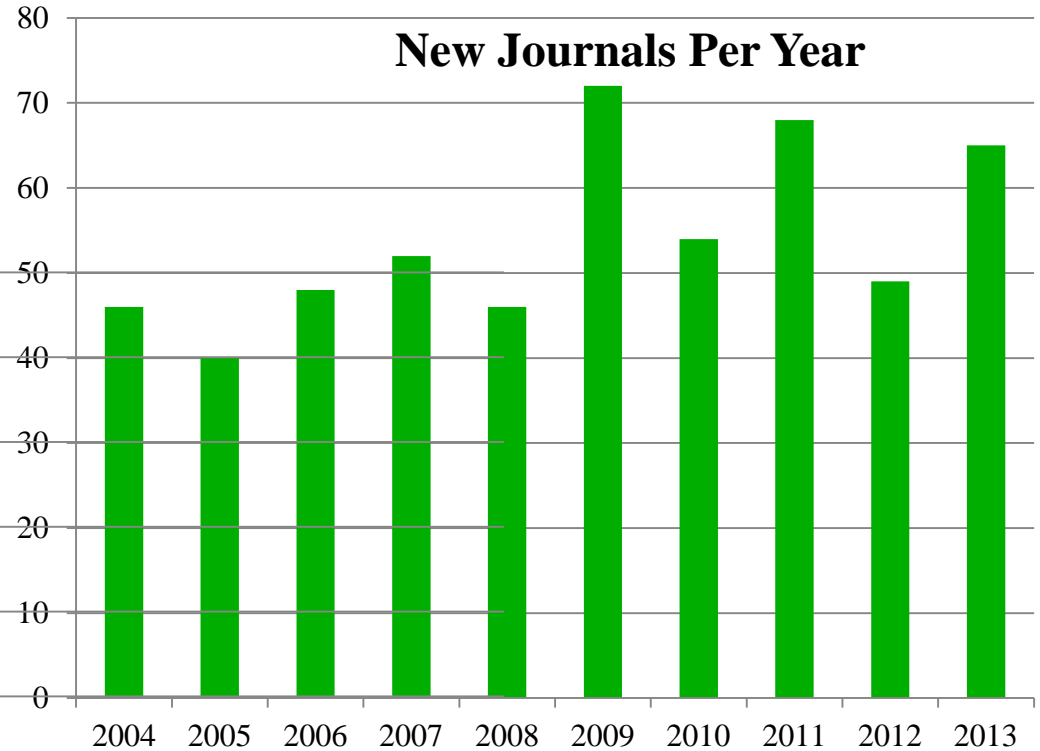
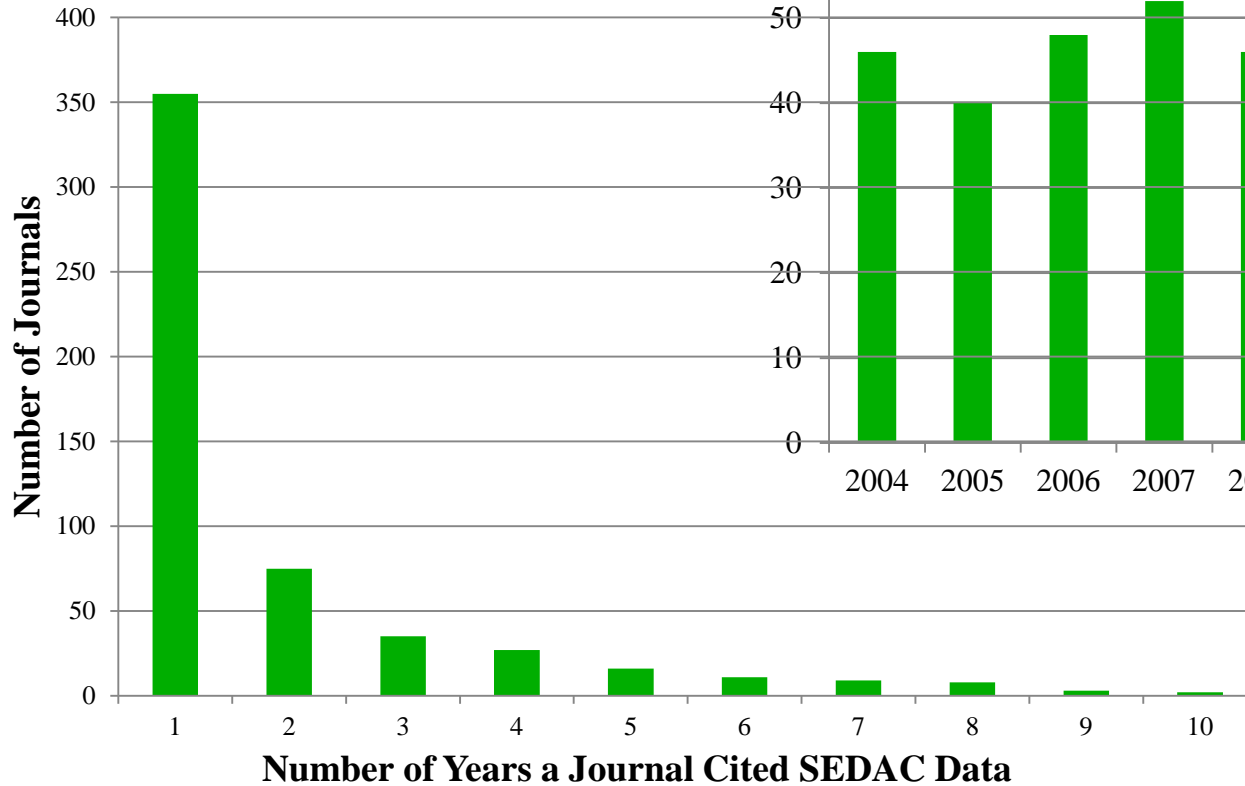
- Routine tracking and reporting of SEDAC data citations \*
  - Content alerts received from bibliographic databases
  - Searches of bibliographic databases and Google Scholar®
  - Notifications received from authors who used SEDAC data
  - SEDAC Routinely reports data citations to NASA
- 1480 journal articles cited SEDAC data in 2003-2014
  - Citing articles were published in 540 journals
- Other data citations also are tracked
  - Books, book chapters, reports, proceedings, atlases and textbooks
  - Broadcast and online media, newspapers, magazines, newsletters, etc.



# 540 Journals Published 1480 Articles Citing SEDAC Data During 2004-2013

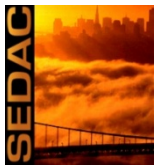


**26% (125 of 480) of Citing Journals  
Cited SEDAC Data in Multiple Years**





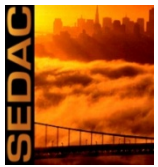
# Journals that Published 10 or More Articles Citing SEDAC Data from 2004 through 2013



PLoS One	66
Ecological Economics	41
Global Environmental Change	30
Atmospheric Environment	24
Climatic Change	23
Proceedings of the National Academy of Sciences	22
Ecological Indicators	21
Biological Conservation	21
Journal of Geophysical Research	21
Energy Policy	19
Atmospheric Chemistry and Physics	18
Environmental Science & Technology	17
Environmental Research Letters	16
Remote Sensing of Environment	16
Malaria Journal	15
Geophysical Research Letters	15
Journal of Environmental Management	15
PLoS Neglected Tropical Diseases	14
Conservation Biology	14
Global Ecology and Biogeography	12
Nature	12
Science	12
Bioscience	11
International Journal of Health Geographics	11
Agriculture, Ecosystems & Environment	11
Global Change Biology	10
Proceedings of the Royal Society B: Biological Sciences	10
Water Resources Research	10
Environmental Science & Policy	10



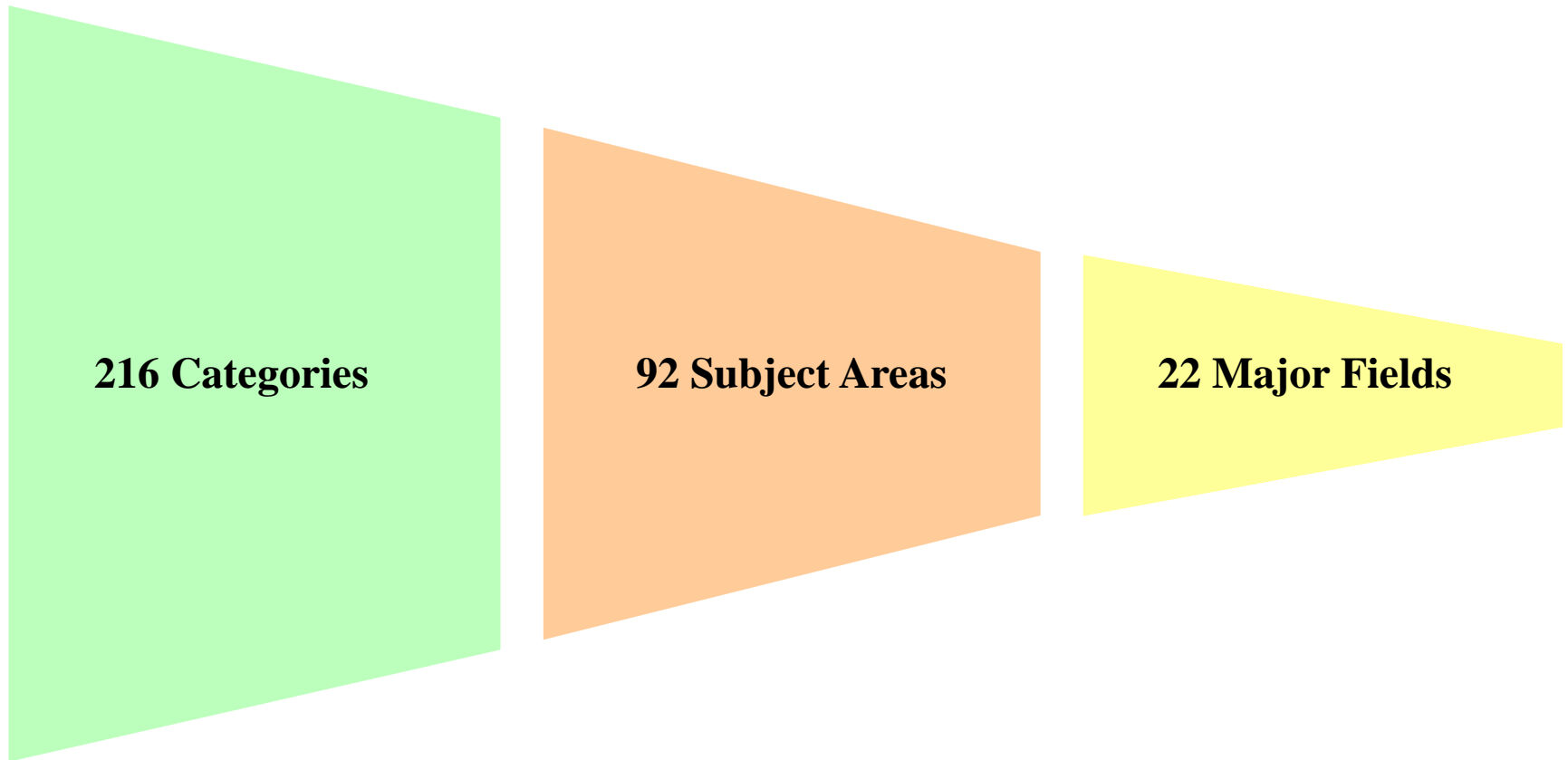
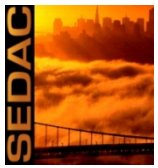
# Example Classifications of Journal Titles



Journal Title	Category	Subject	Major Field	General
Agricultural Economics	Agricultural Economics & Policy	Agriculture	Agricultural Sciences	Science & Technology
Agricultural Economics	Economics	Business & Economics	Economics & Business	Social Sciences
Agricultural Systems	Agriculture, Multidisciplinary	Agriculture	Agricultural Sciences	Science & Technology
Journal of Infectious Diseases	Immunology	Immunology	Immunology	Science & Technology
Journal of Infectious Diseases	Infectious Diseases	Infectious Diseases	Immunology	Science & Technology
Journal of Infectious Diseases	Microbiology	Microbiology	Microbiology	Science & Technology

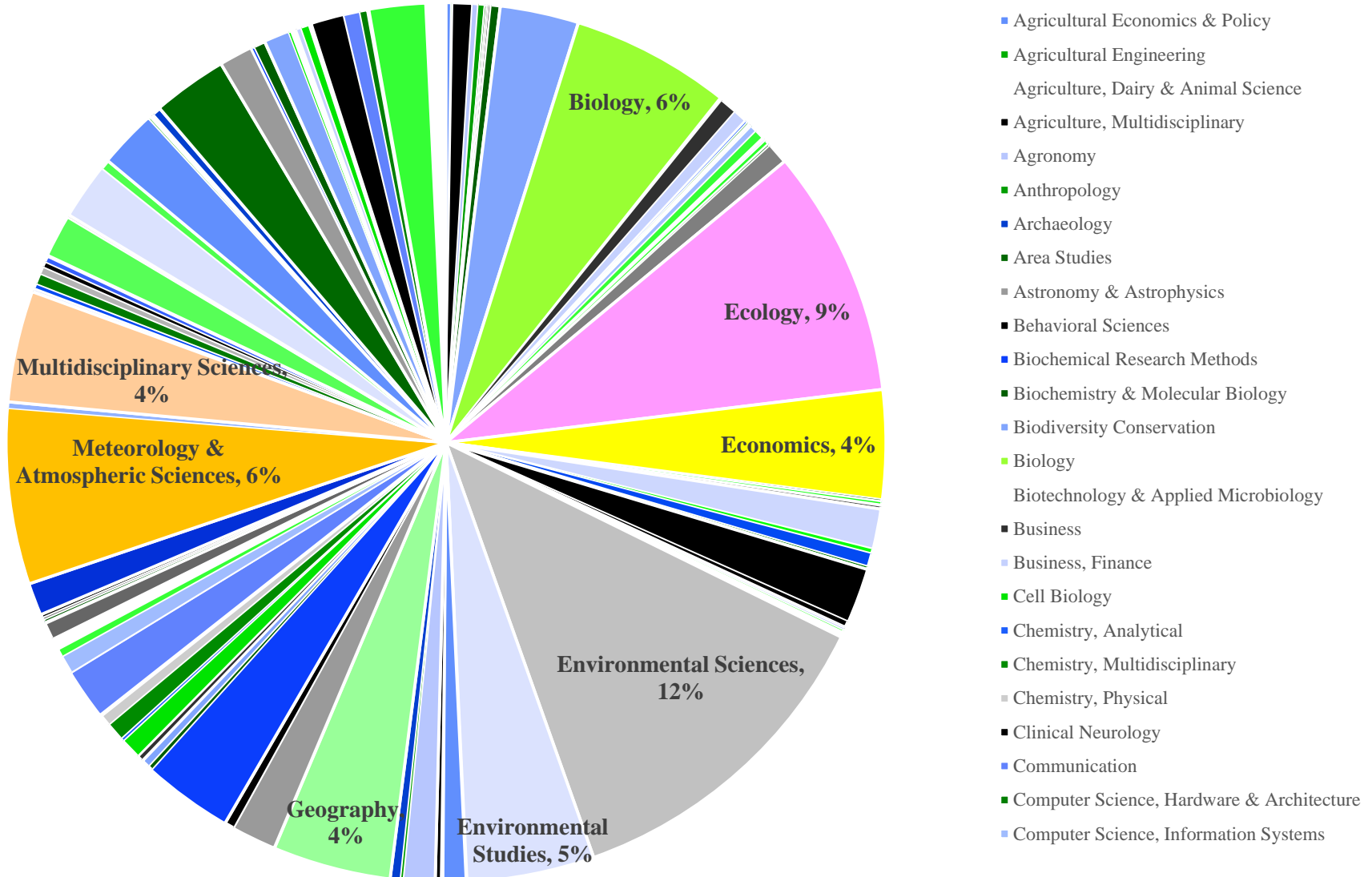
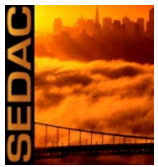


# Multidisciplinary Impact of SEDAC Data: 2004-2013



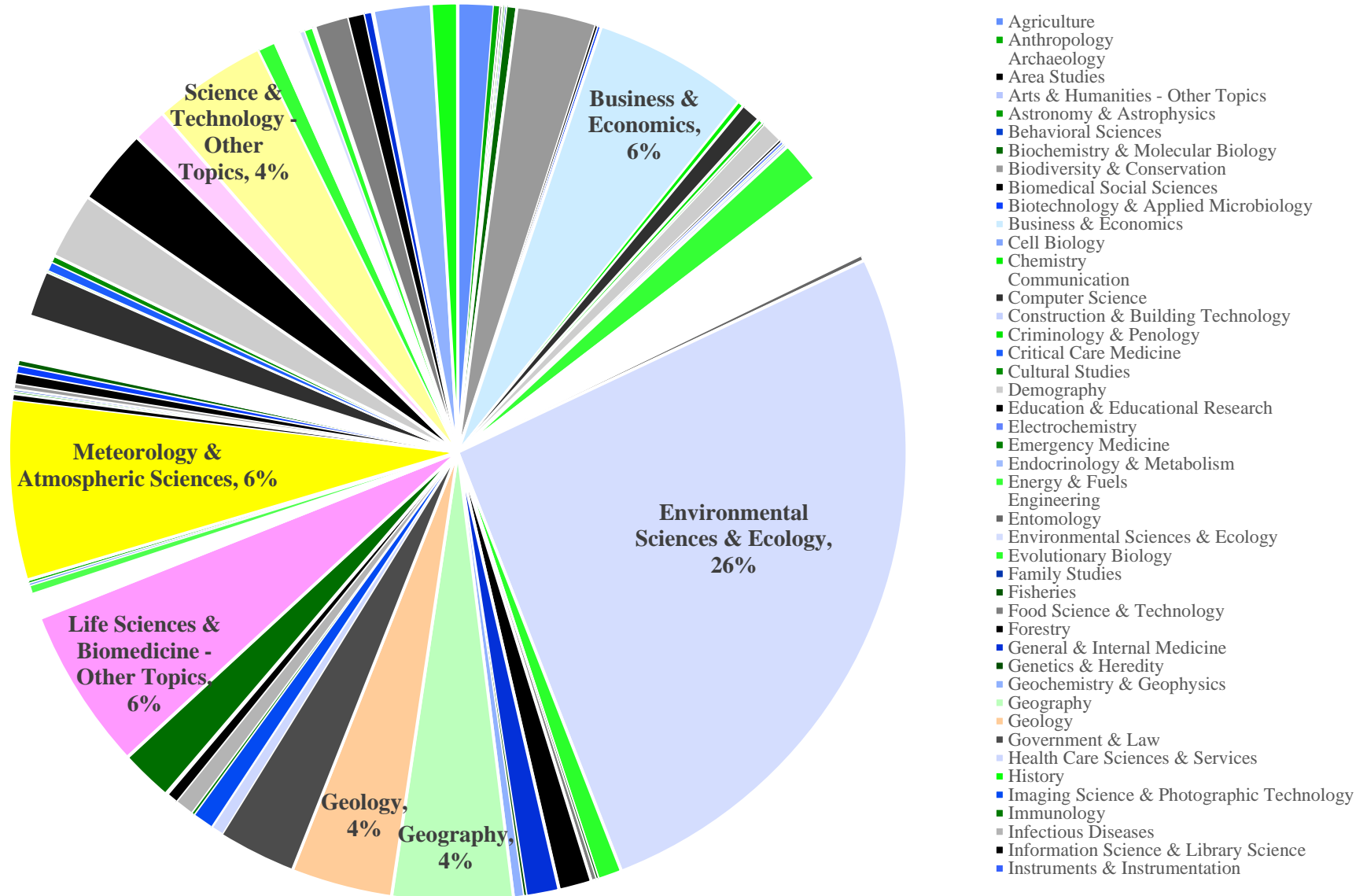
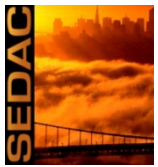


# 216 WoS Categories of Journals Containing Citations of SEDAC Data from 2004 through 2013





# Proportional Representation of 92 WoK Subject Areas of Journals that Cited SEDAC Data 2004-2013







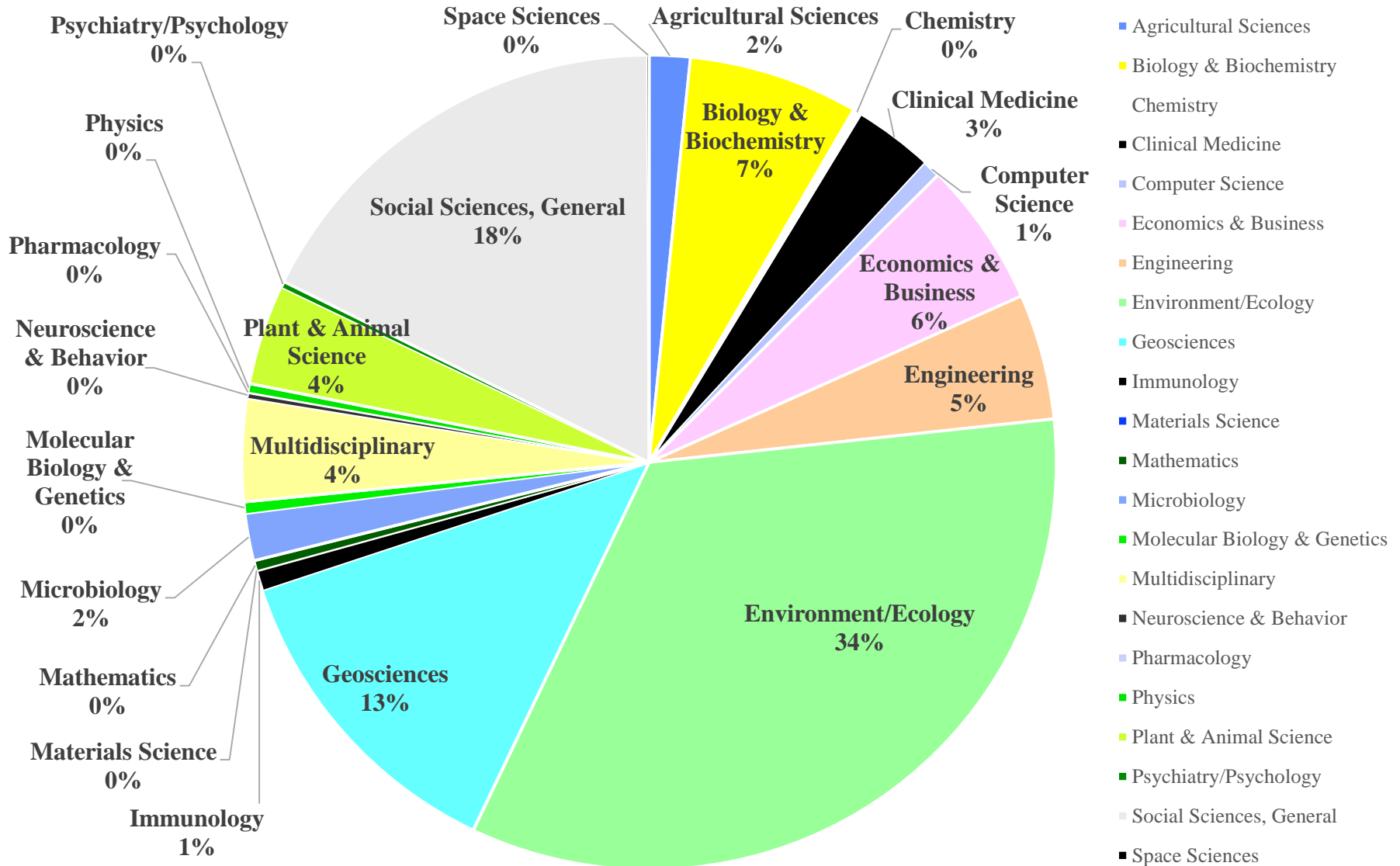
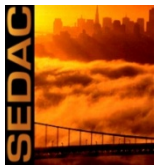
# 92 WoK Subject Classifications Represented by Journals that Cited SEDAC Data 2004-2013



Agriculture	Fisheries	Nutrition & Dietetics
Anthropology	Food Science & Technology	Oceanography
Archaeology	Forestry	Oncology
Area Studies	General & Internal Medicine	Operations Research & Management Science
Arts & Humanities - Other Topics	Genetics & Heredity	Orthopedics
Astronomy & Astrophysics	Geochemistry & Geophysics	Paleontology
Behavioral Sciences	Geography	Parasitology
Biochemistry & Molecular Biology	Geology	Physical Geography
Biodiversity & Conservation	Government & Law	Physics
Biomedical Social Sciences	Health Care Sciences & Services	Physiology
Biotechnology & Applied Microbiology	History	Plant Sciences
Business & Economics	Imaging Science & Photographic Technology	Psychology
Cell Biology	Immunology	Public Administration
Chemistry	Infectious Diseases	Public, Environmental & Occupational Health
Communication	Information Science & Library Science	Remote Sensing
Computer Science	Instruments & Instrumentation	Science & Technology - Other Topics
Construction & Building Technology	International Relations	Social Issues
Criminology & Penology	Life Sciences & Biomedicine - Other Topics	Social Sciences - Other Topics
Critical Care Medicine	Linguistics	Sociology
Cultural Studies	Marine & Freshwater Biology	Surgery
Demography	Materials Science	Telecommunications
Education & Educational Research	Mathematical & Computational Biology	Thermodynamics
Electrochemistry	Mathematics	Toxicology
Emergency Medicine	Mathematics, Applied	Transportation
Endocrinology & Metabolism	Medical Informatics	Tropical Medicine
Energy & Fuels	Meteorology & Atmospheric Sciences	Urban Studies
Engineering	Microbiology	Veterinary Sciences
Entomology	Mycology	Virology
Environmental Sciences & Ecology	Neurosciences & Neurology	Water Resources
Evolutionary Biology	Nuclear Science & Technology	Zoology
Family Studies	Nursing	

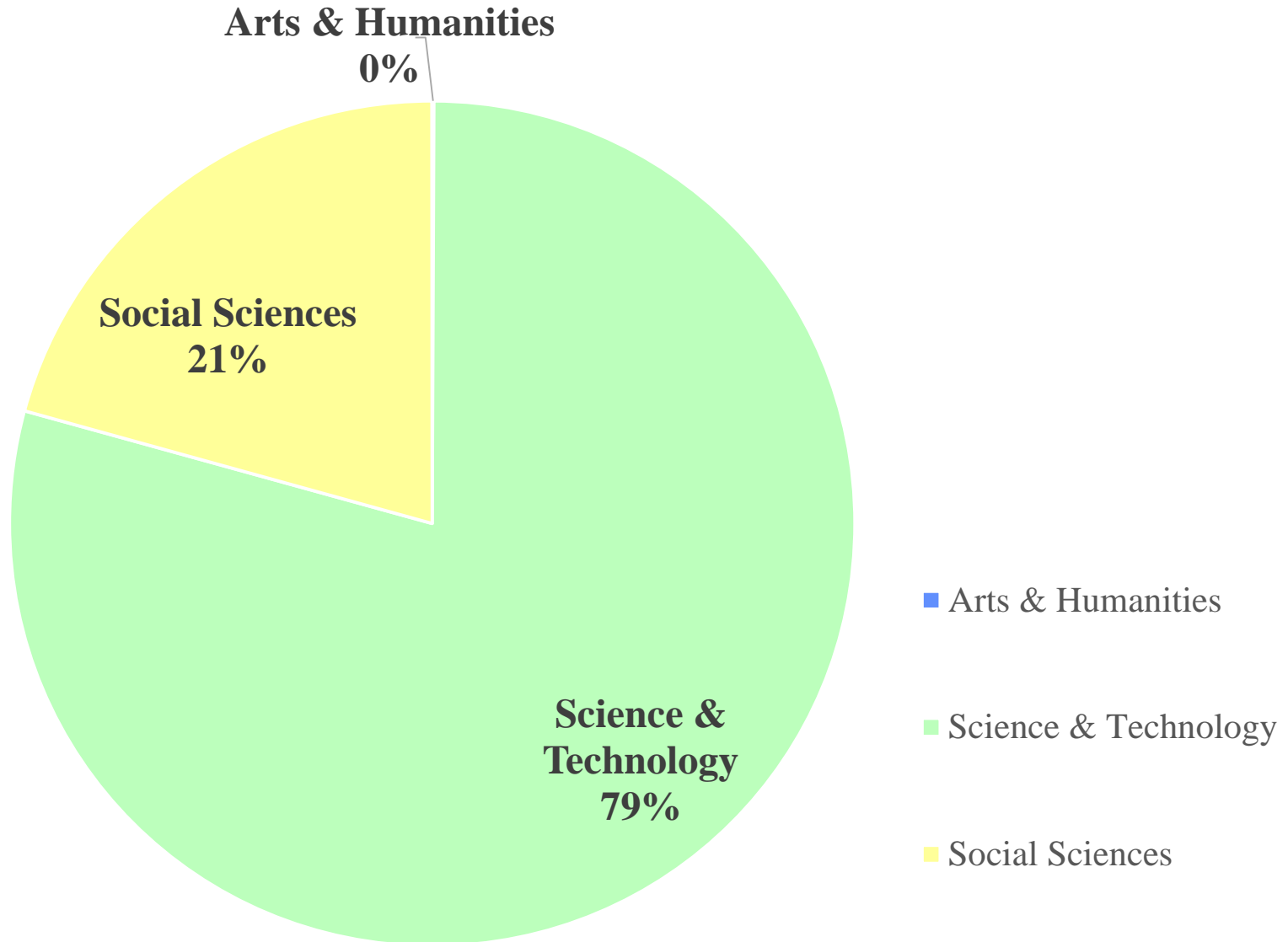
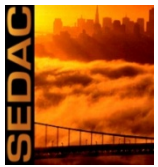


# 22 Major Fields of Science Represented by Journals that Cited SEDAC Data 2004-2013



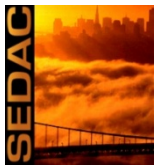


# WoK General Categories Represented by Journals that Cited SEDAC Data 2004-2013

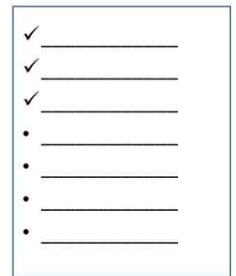




# Implications of Multidisciplinary Data Use

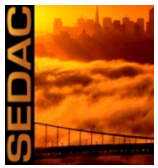


- Data use extends across traditional disciplinary boundaries
  - Users do not represent a limited subset of science disciplines
  - Use may exceed uses planned during data creation and development
  - Discipline-specific presumptions about use could restrict potential users
  - Opportunities for new uses should be fostered
- Design data products, services, tools, and support for broad use
  - Develop affordance to support use by multiple disciplines
  - Seek design guidance from non-traditional disciplinary users
  - Offer capabilities to enable unanticipated uses





# Limitations

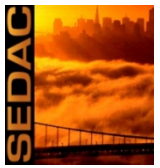


- Only analyzed identified journal articles that cited SEDAC data.
  - SEDAC data could have been used for a publication, but not cited.
  - All SEDAC data citations might not have been identified.
- Individual judgements could influence assignments and pairings.
  - WoS Category assignments are conducted by WoK staff.
  - Pairing of WoS Category to Major Field was completed by the first author.
  - Data collection of WoS Category assignments was conducted over 4 years.
- Disciplinary impact of scientific data could include other factors.
  - Citations in journal articles do not reflect entire disciplinary impact.
  - SEDAC data also are cited in books, chapters, and reports.
  - The extent of data use for each article was not considered in the analysis.
  - Alternative metrics may inform disciplinary impact of SEDAC data.





# Questions for Further Study



- How do other journal discipline classifications compare with those that were used?
- How would assessing type or extent of use, as reported in each journal article, improve understanding of data impact?
- How do data citations reflect the use of SEDAC data in conjunction with data from other sources?