Computational Social Science in “Science” & “Nature”

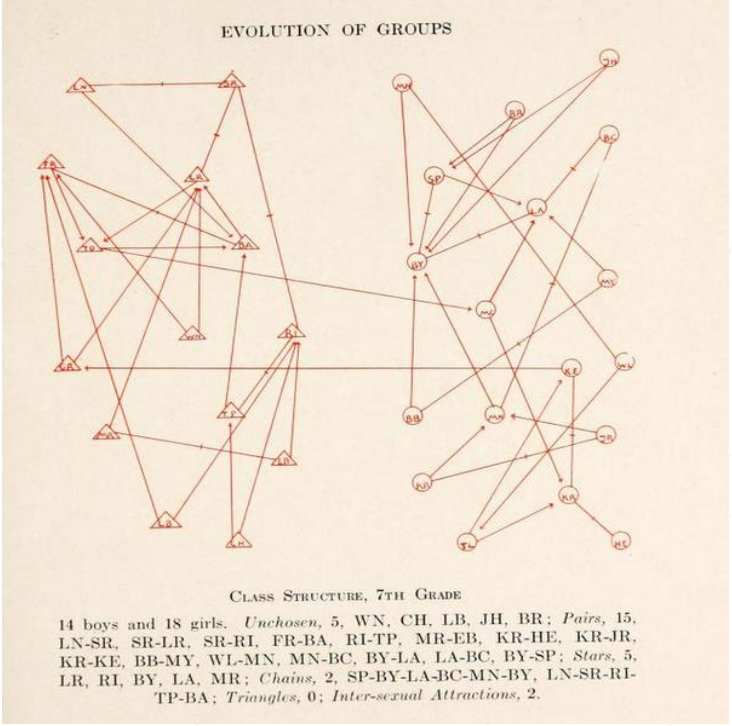
Joseph A. E. Shaheen
ORISE Intelligence Community Postdoc Fellow

Host Institution: Department of Computational & Data Science
George Mason University

This research was supported by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at George Mason University, administered by Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence.

History of Computational Social Science?

- Moreno (1934)
- Bott (1957)
- Erdos & Renyi (1959)
- Traverse & Milgram (1969)
- Granovetter (1973)
- Freeman (1970s)
- Holland & Reinhard (1981); Krackhardt (1987)
- Barabasi, Wattz and Strogatz (1990s)
- Newman, Snijders (2000s)



10 Year Anniversary

SOCIAL SCIENCE

Computational Social Science

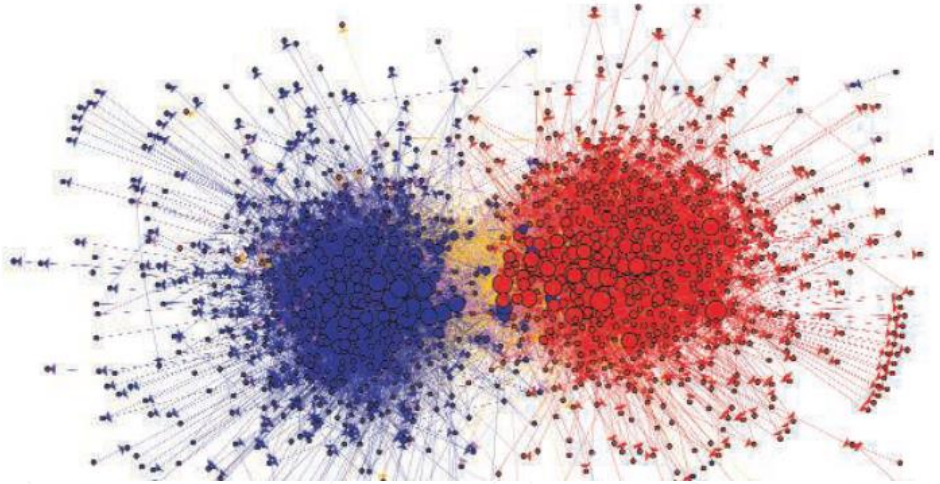
David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

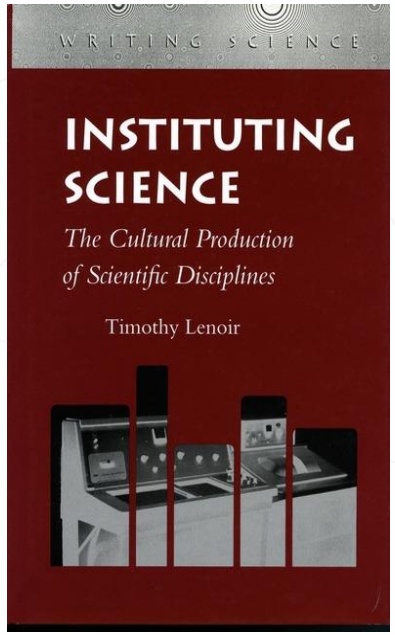
The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

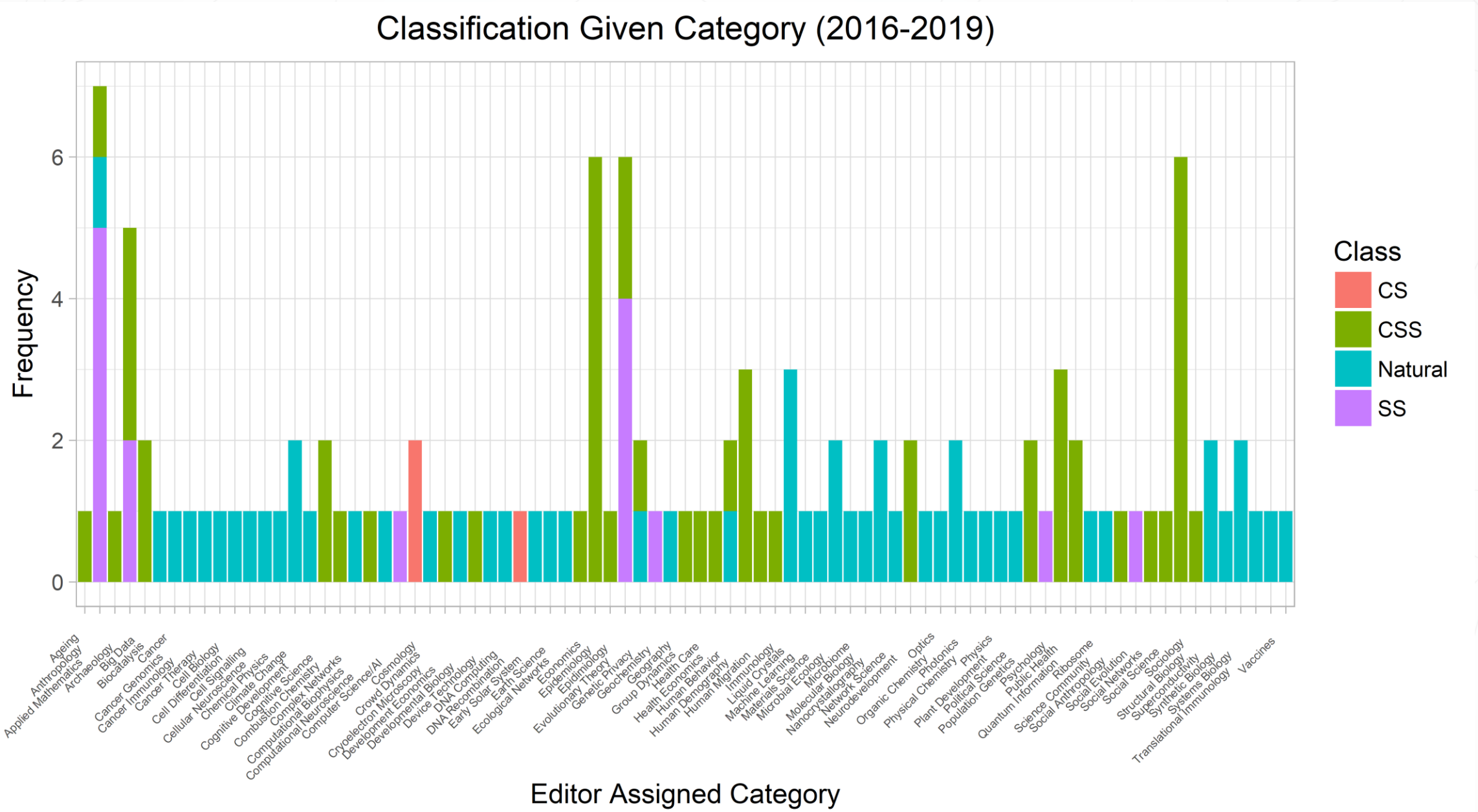


Central Question?

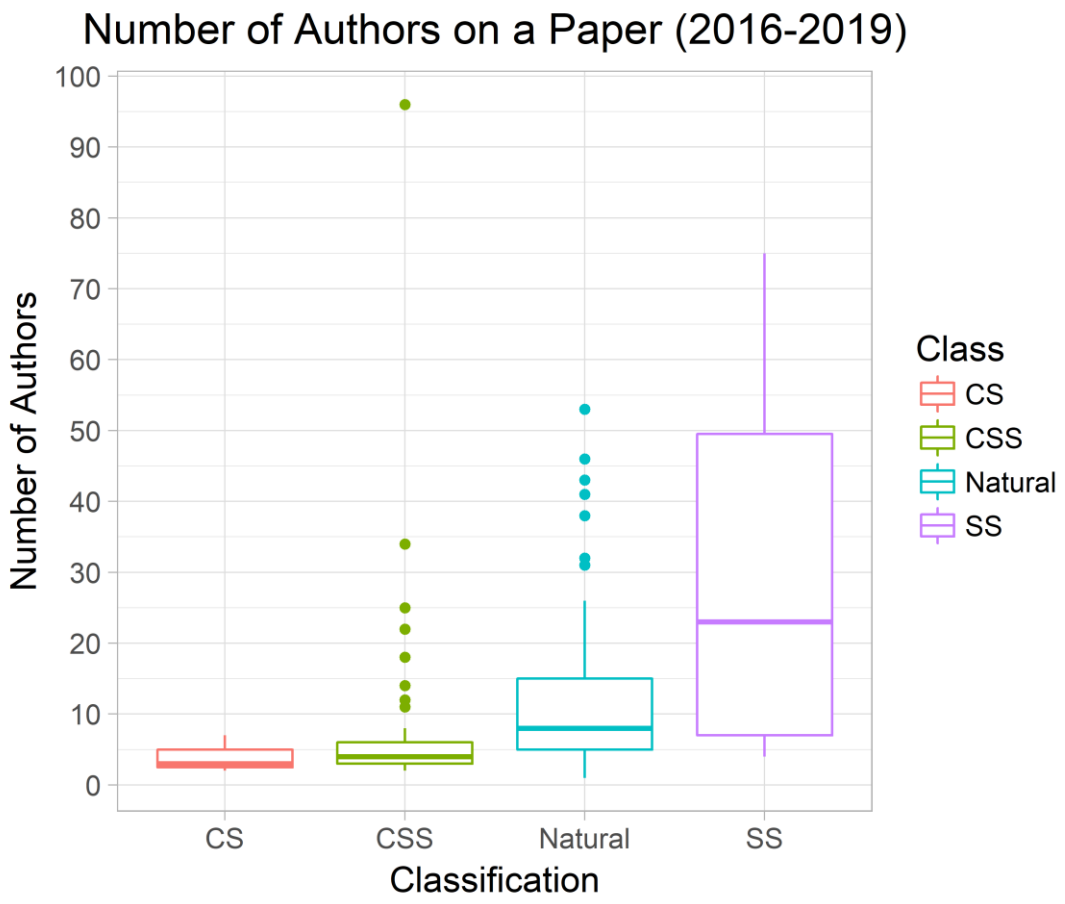
- Can we identify a field or a subfield, especially if interdisciplinary through the meta features of the papers published in the domain?
- The influence of top journals would be paramount in defining an interdisciplinary field.
- Nature and Science are natural targets.

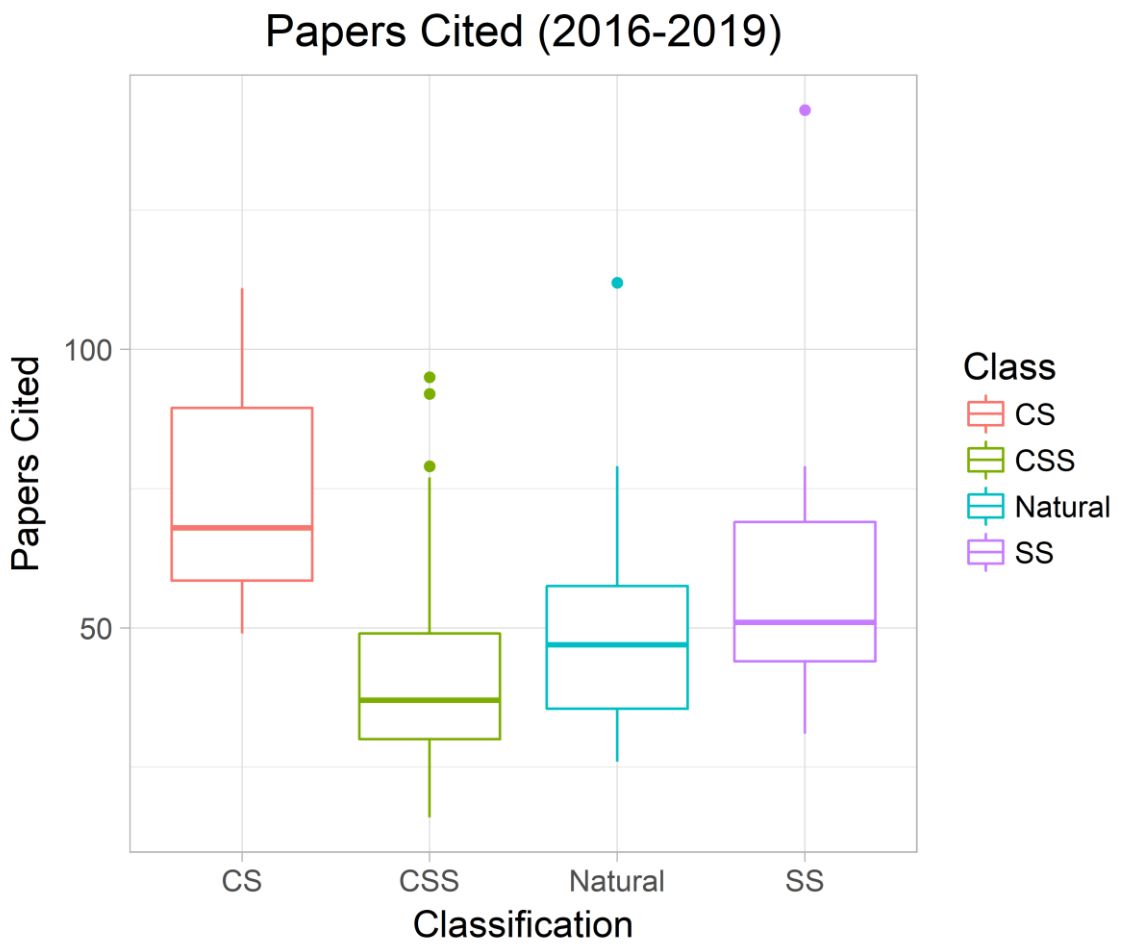


Lenoir, Timothy. *Instituting science: The cultural production of scientific disciplines*. Stanford University Press, 1997.

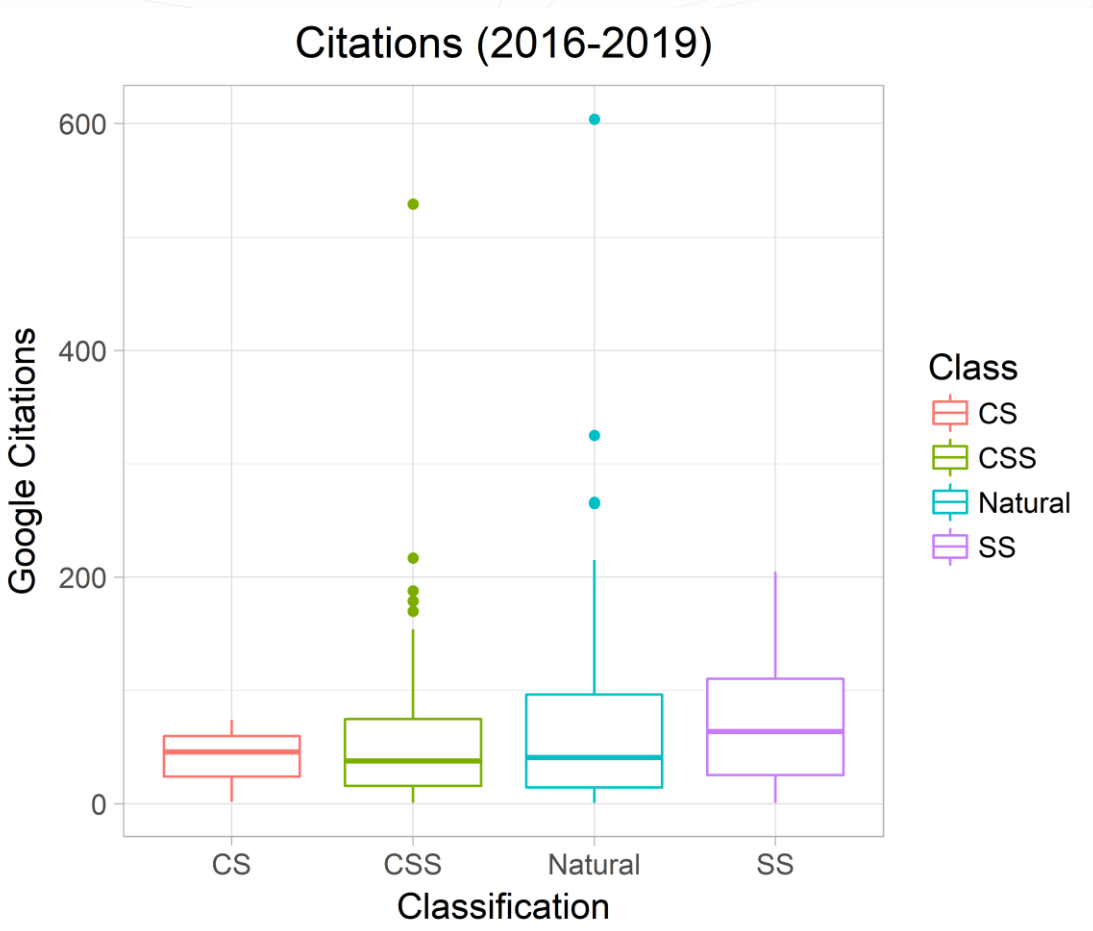

Classification Given Category (2016-2019)
Frequency
0
2
4
6
Class
CS
CSS
Natural
SS
Ageing
Anthropology
Applied Mathematics
Archaeology
Big Data
Biocatalysis
Cancer
Cancer Genomics
Cancer Immunology
Cancer Therapy
Cell Biology
Cell Differentiation
Cell Signalling
Cellular Neuroscience
Chemical Physics
Climate Change
Cognitive Development
Cognitive Science
Combustion Chemistry
Complex Networks
Computational Biophysics
Computational Neuroscience
Computer Science/AI
Cosmology
Crowd Dynamics
Cryoelectron Microscopy
Development Economics
Developmental Biology
Device Technology
DNA Computing
DNA Recombination
Early Solar System
Earth Science
Ecological Networks
Economics
Epidemiology
Epidimiology
Evolutionary Theory
Genetic Privacy
Geochemistry
Geography
Group Dynamics
Health Care
Health Economics
Human Behavior
Human Demography
Human Migration
Immunology
Liquid Crystals
Machine Learning
Materials Science
Microbial Ecology
Microbiome
Molecular Biology
Nanocrystallography
Network Science
Neurodevelopment
Optics
Organic Chemistry
Photonics
Physical Chemistry
Physics
Plant Development
Political Science
Population Genetics
Psychology
Public Health
Quantum Information
Ribosome
Science Community
Social Anthropology
Social Evolution
Social Networks
Social Science
Sociology
Structural Biology
Superconductivity
Synthetic Biology
Systems Biology
Translational Immunology
Vaccines
Editor Assigned Category

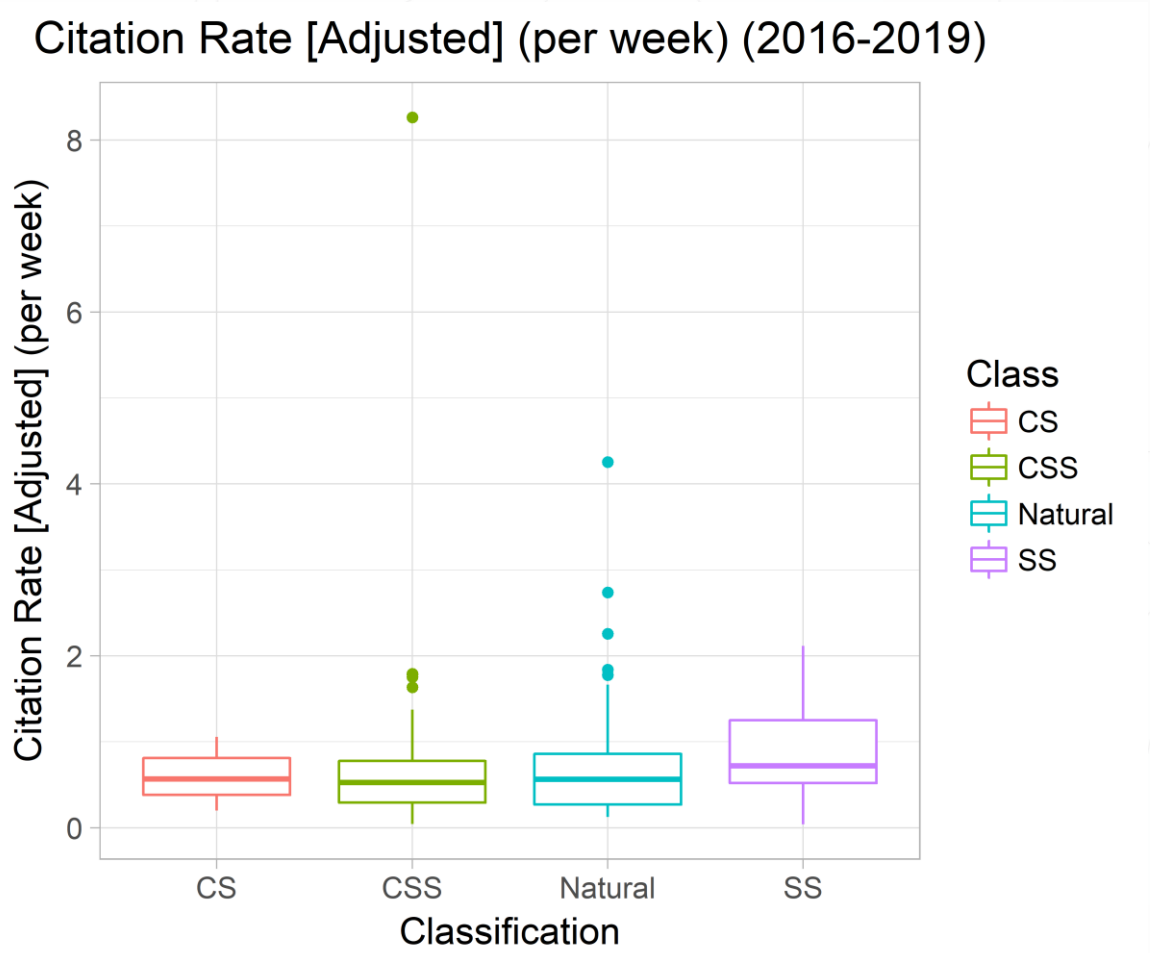
Features


Number of Authors on a Paper (2016-2019)
Number of Authors
100
90
80
70
60
50
40
30
20
10
0
CS
CSS
Natural
SS
Classification
Class
CS
CSS
Natural
SS

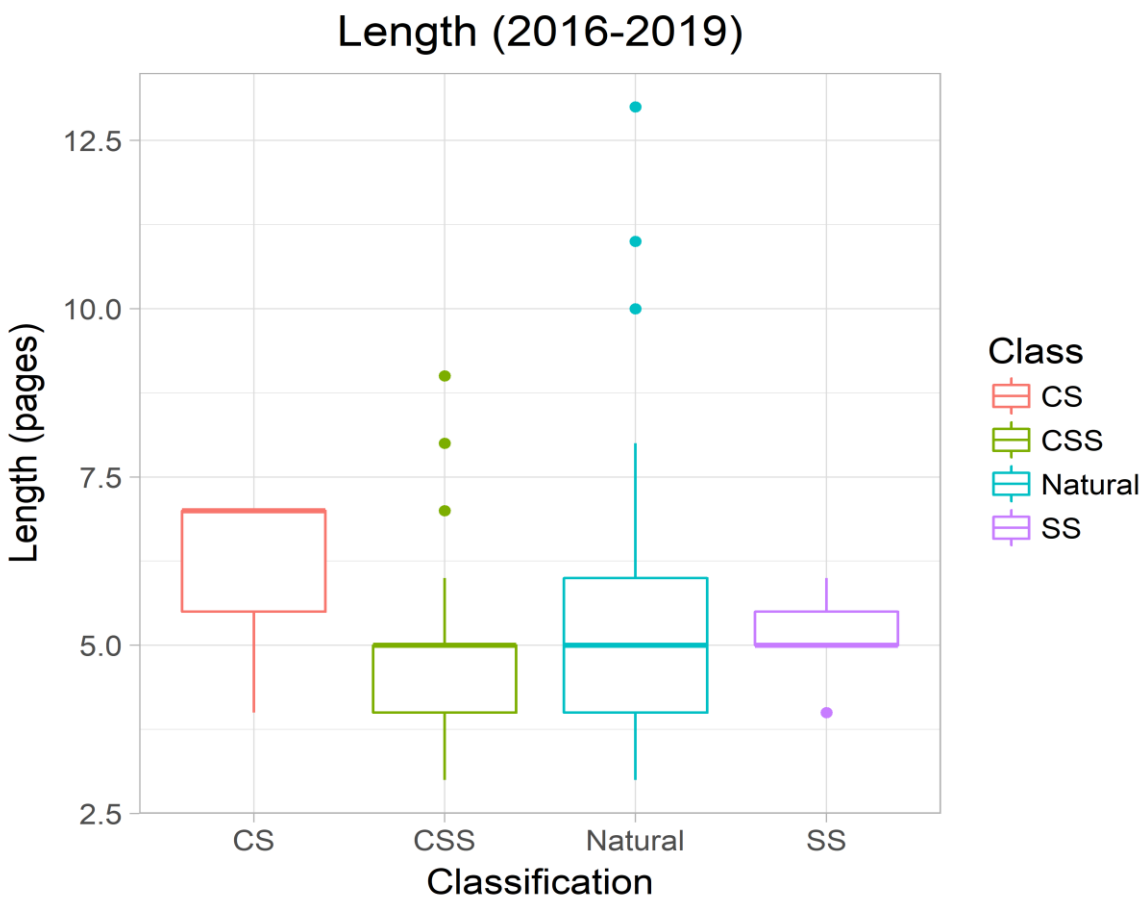

Papers Cited (2016-2019)
Papers Cited
100
50
CS
CSS
Natural
SS
Classification
Class
CS
CSS
Natural
SS

Features

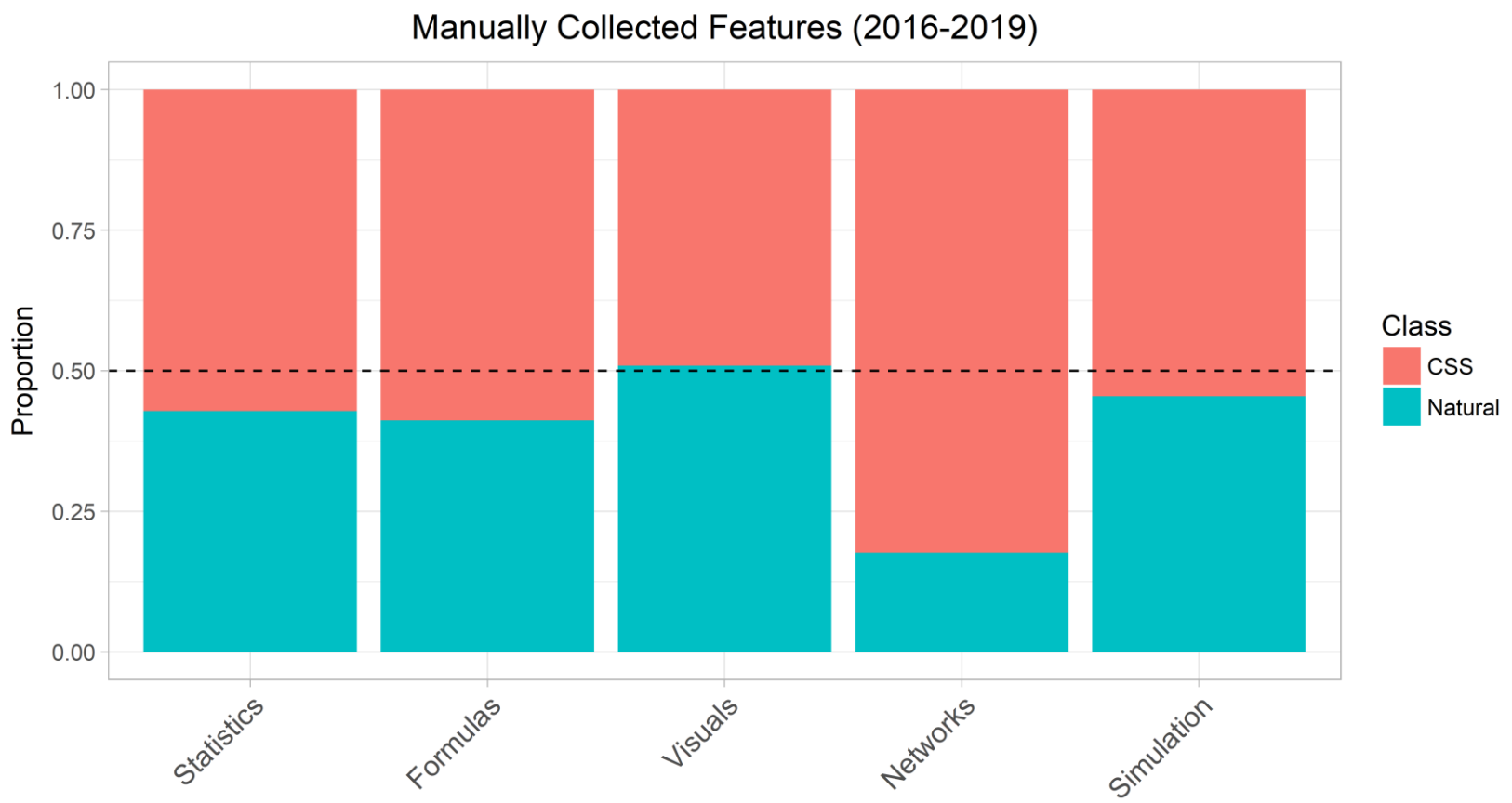

Citations (2016-2019)
Google Citations
600
400
200
0
CS
CSS
Natural
SS
Classification
Class
CS
CSS
Natural
SS


Citation Rate [Adjusted] (per week) (2016-2019)
Citation Rate [Adjusted] (per week)
8
6
4
2
0
CS
CSS
Natural
SS
Classification
Class
CS
CSS
Natural
SS

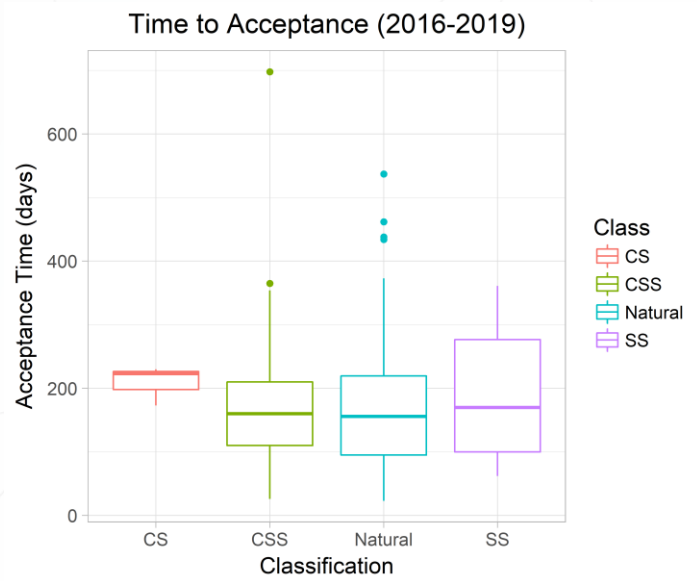
Features

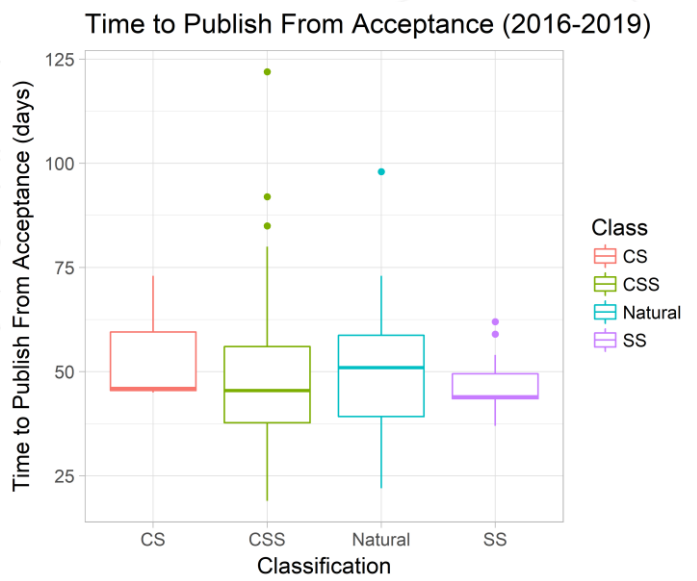

Length (2016-2019)
Length (pages)
12.5
10.0
7.5
5.0
2.5
CS
CSS
Natural
SS
Classification
Class
CS
CSS
Natural
SS

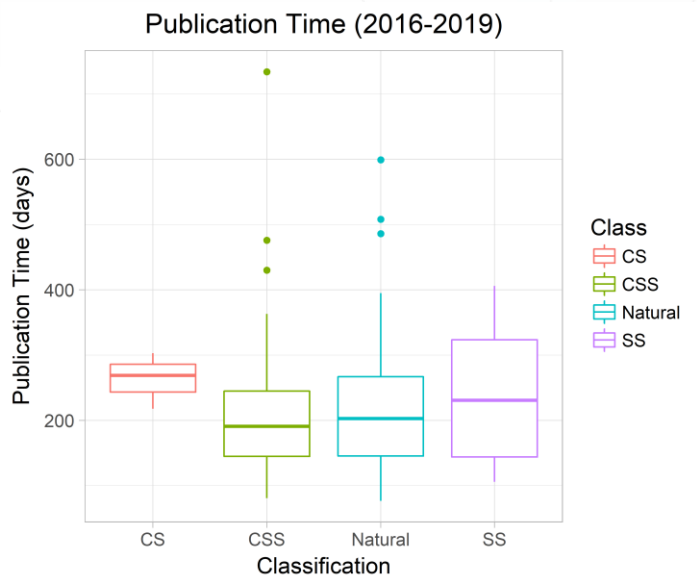
Features (Manual & Qualitative)



Features (editorial)





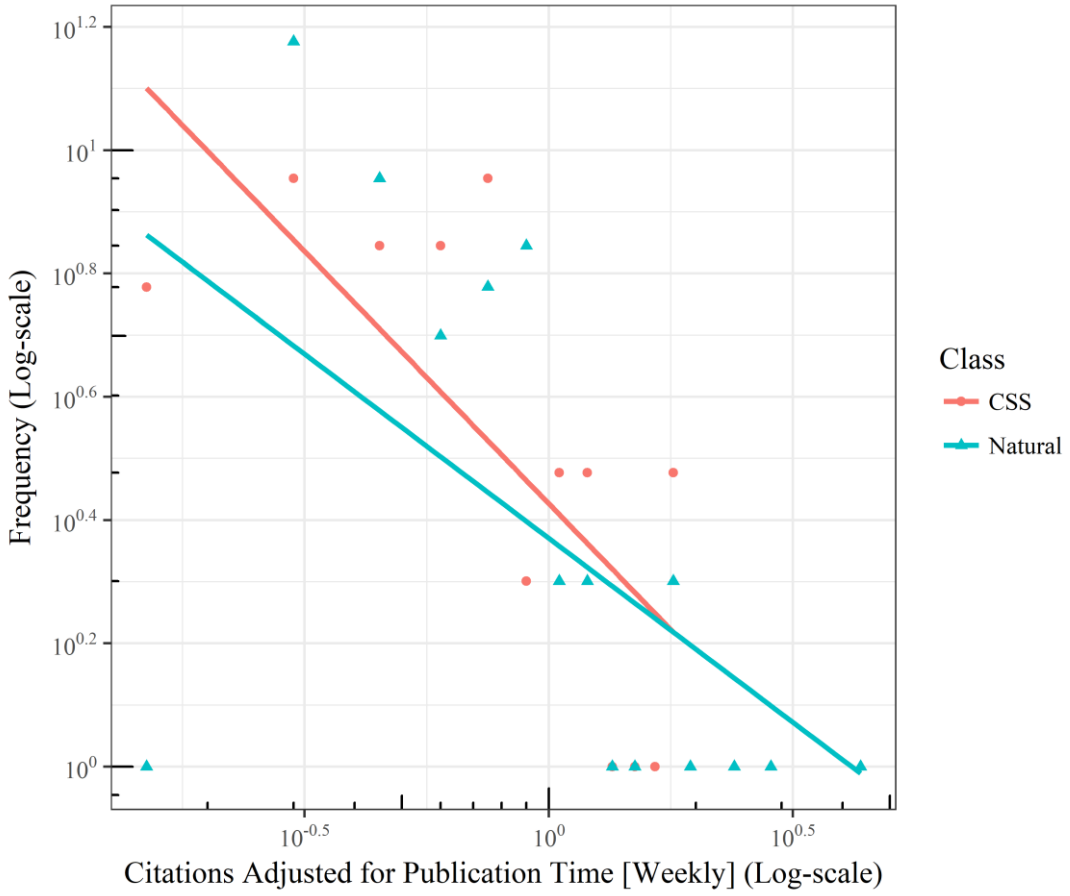


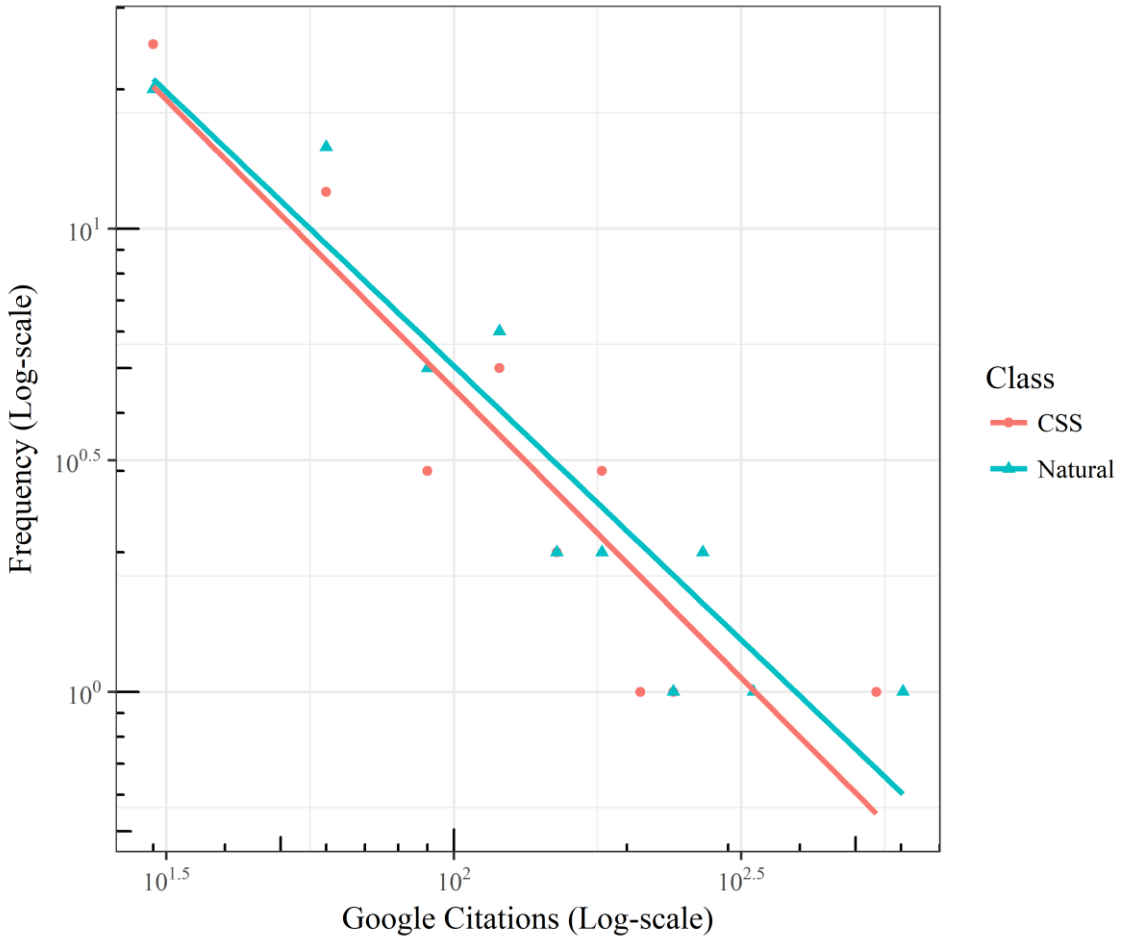
Non-Gaussian Features

- Citations (article/author) have been shown to follow highly skewed distributions (exponential, log-normal, ***power law***)
- Constraints and bounds from below and above. e.g. Length, Number of Authors, Time to Acceptance
- Possible violations of central tendency assumptions

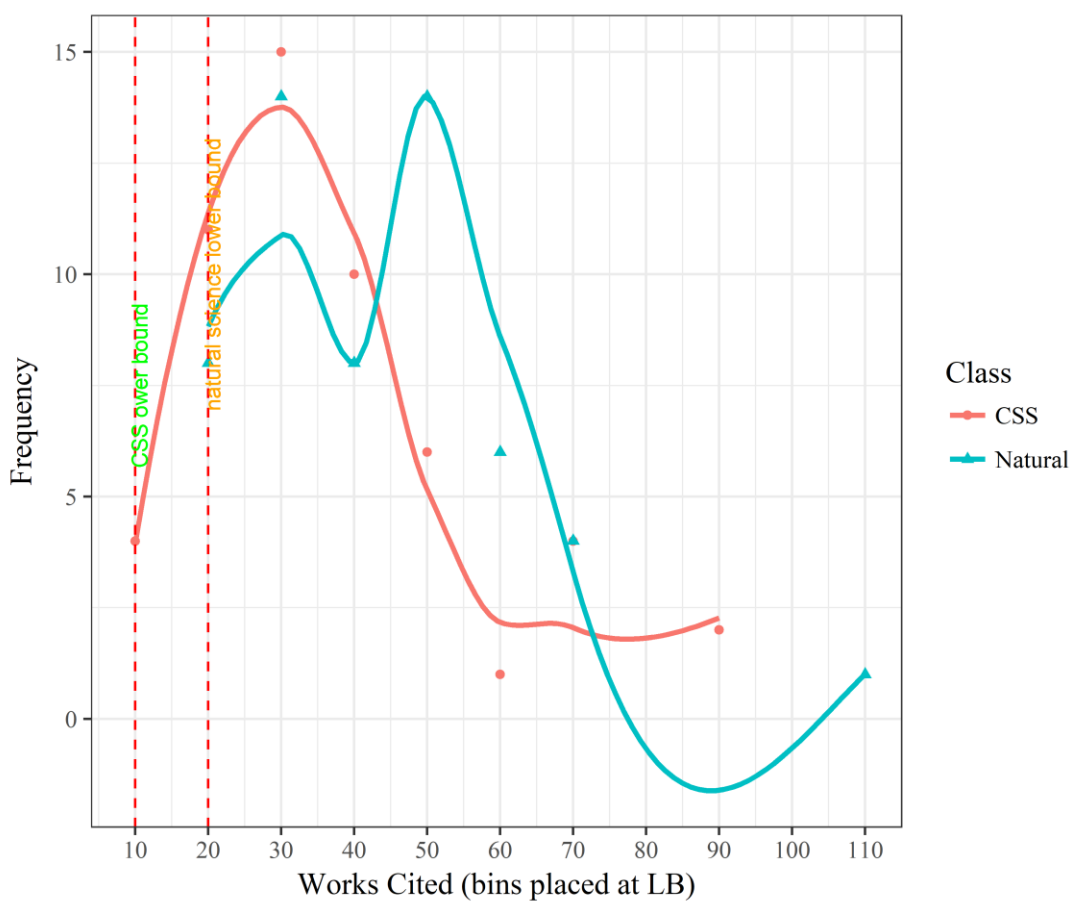
D. J. de S. Price, Networks of scientific papers. Science 149, 510–515 (1965).

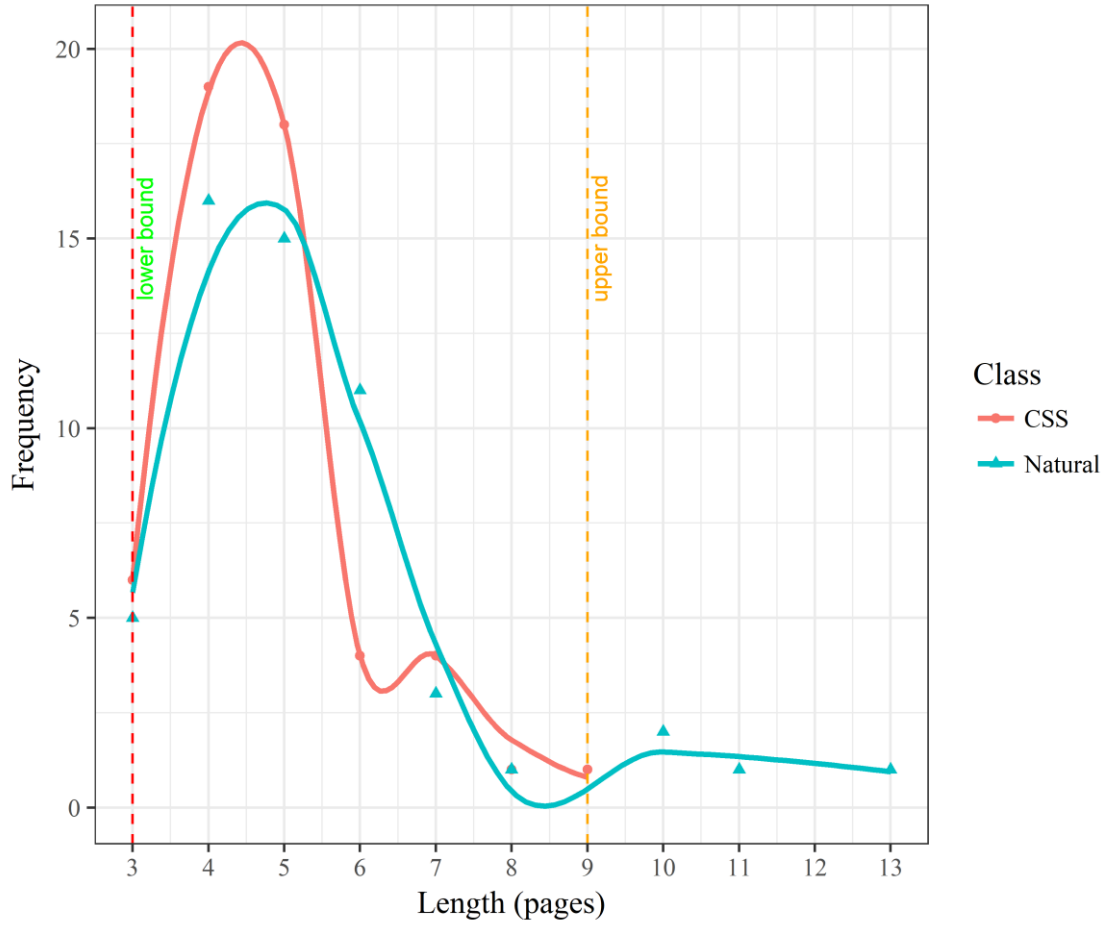
Highly Skewed Features



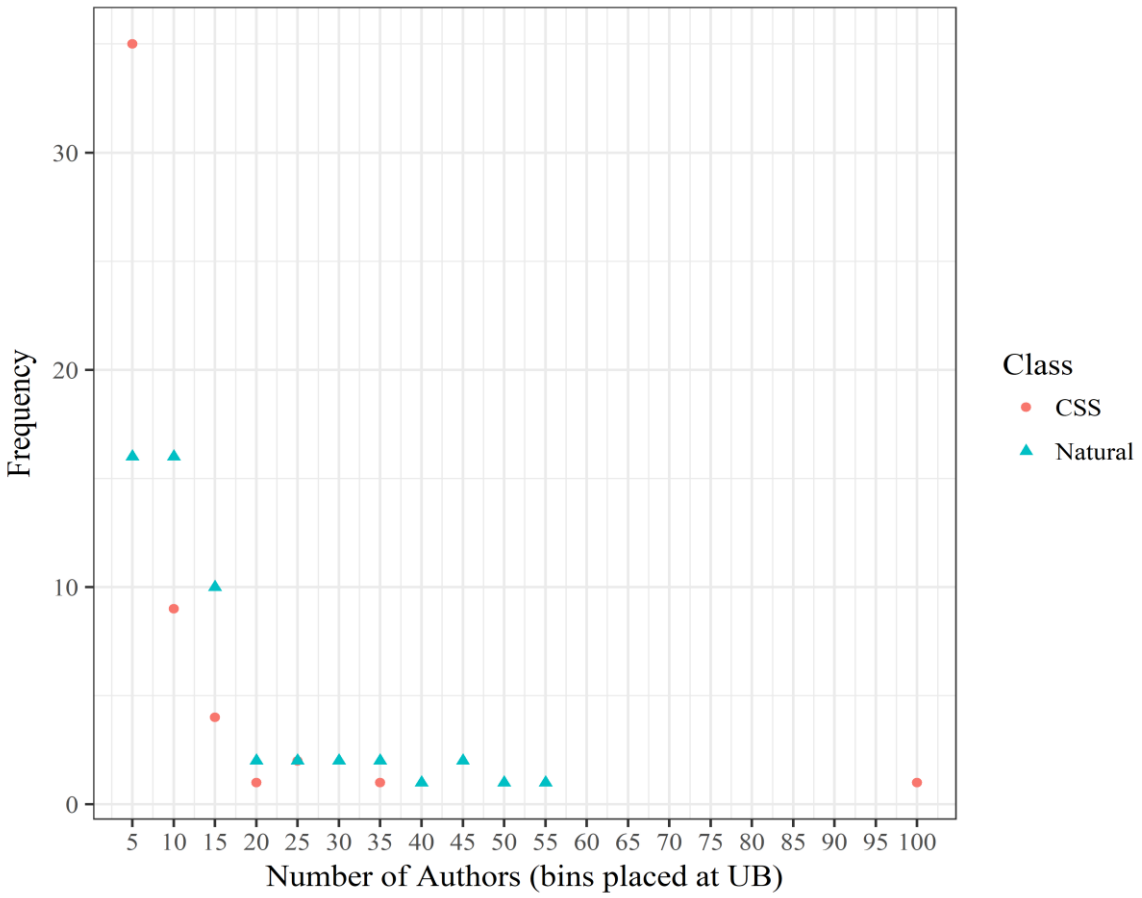


Highly Skewed Features



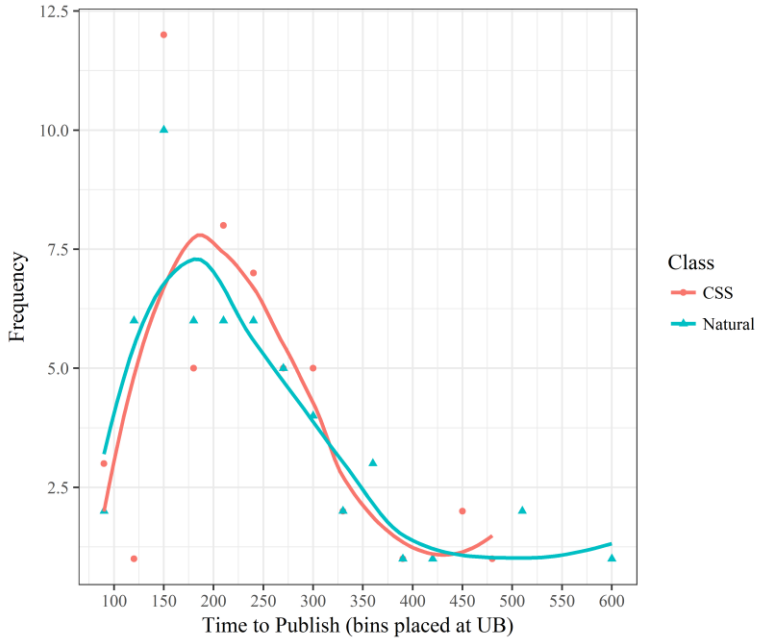


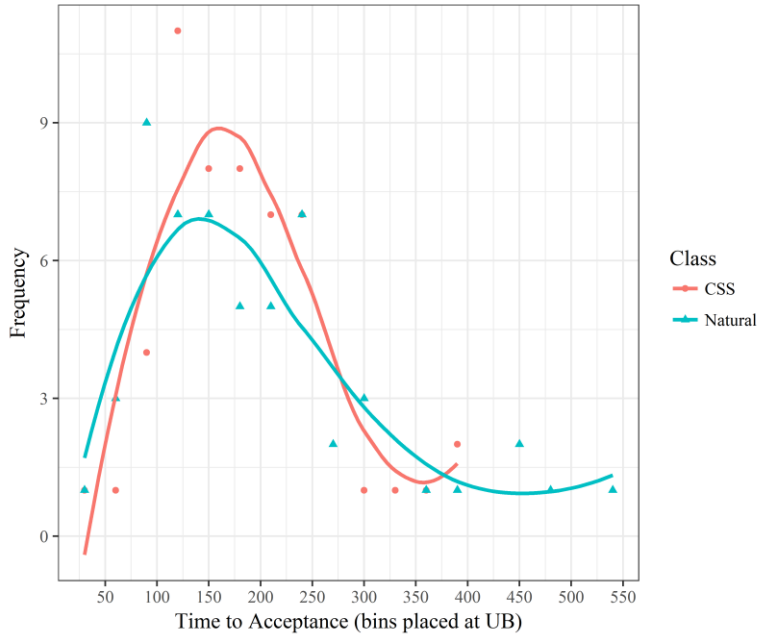
Constraints

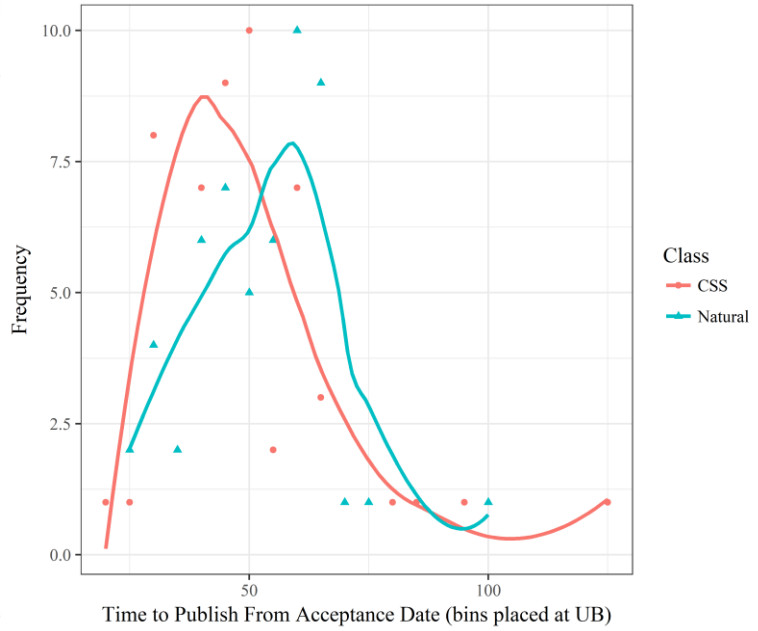


- ***Editorial constraints*** on length at top journals.
- Incentive to add material online
- Necessary not to be careless with claims.

Editorial Decisions



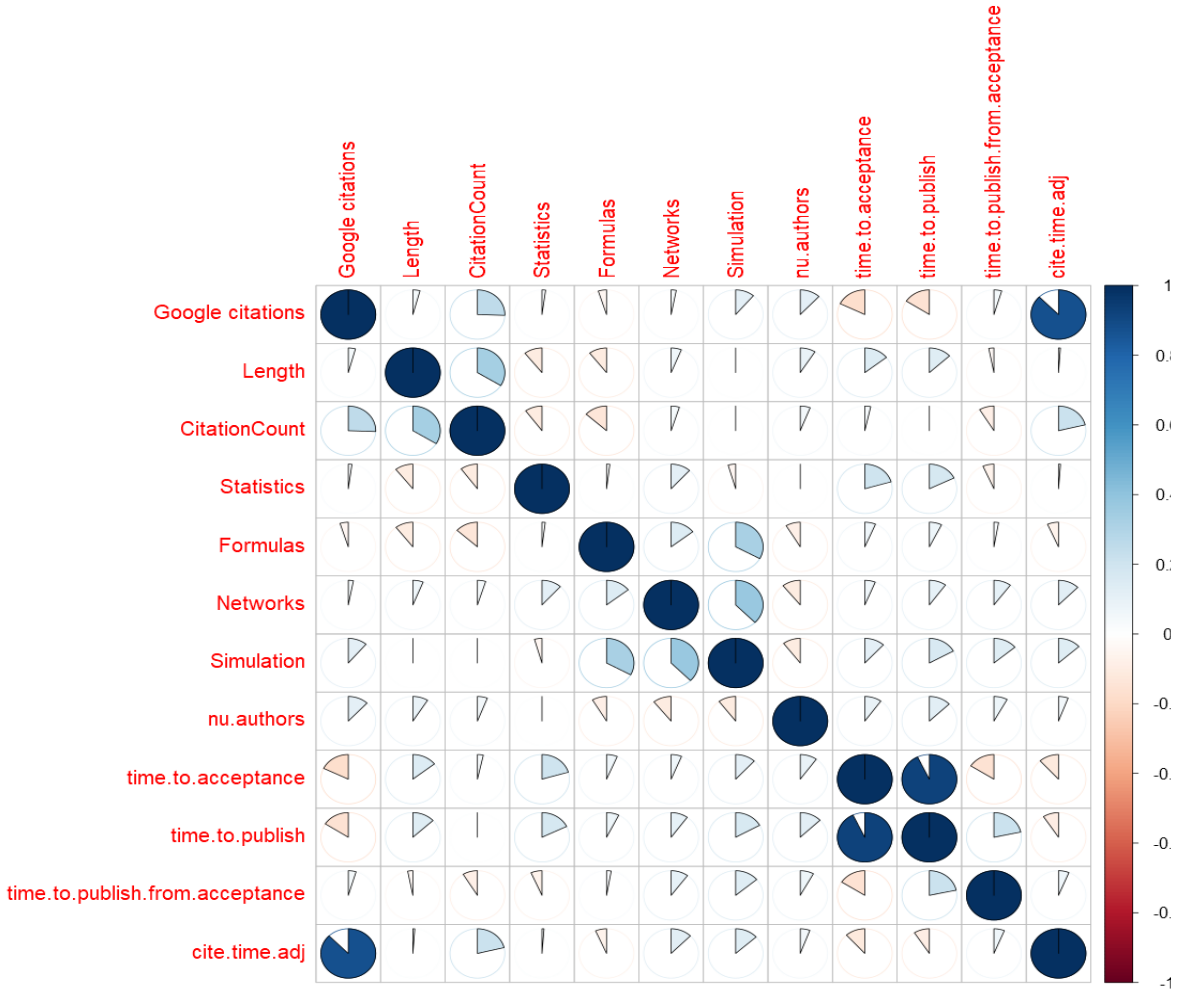


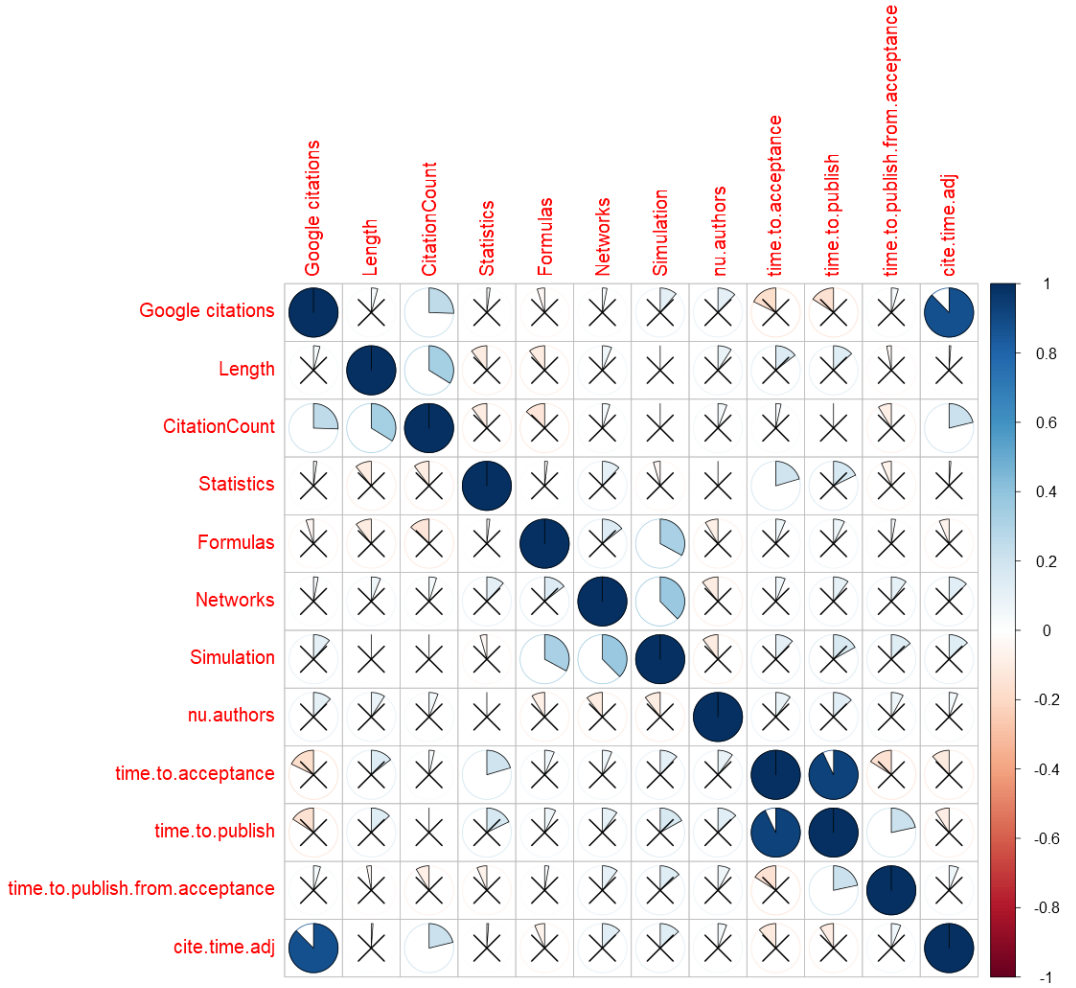


Do we need a model?

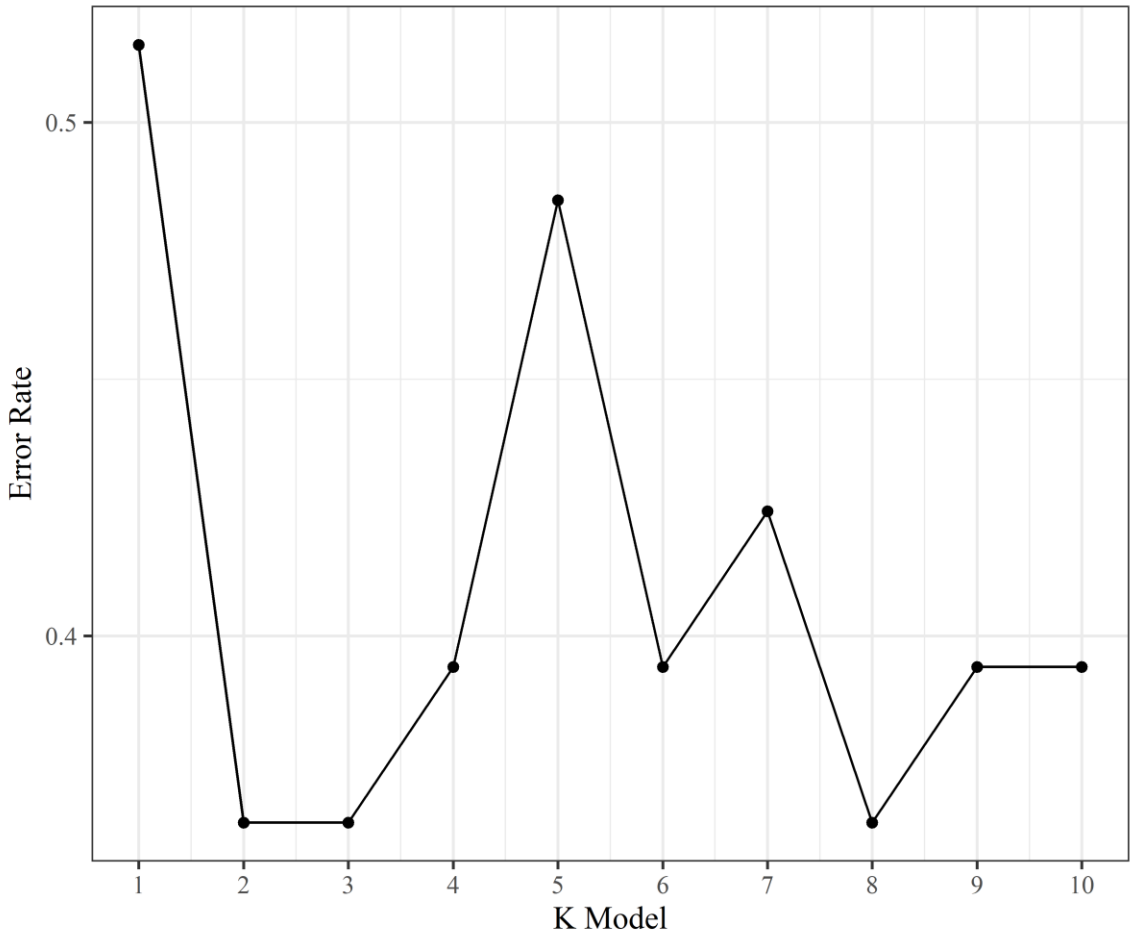
- All models are bad, but some are…worse..
- Violations of non-normality
- ***General Linear Model*** is still robust, but it's not clear if it would scale well when adjusting for a larger sample when considering the non-linearity of our features.
- Binary (dichotomized) variables restrict our choices
- Non-parametric, classification, or clustering.

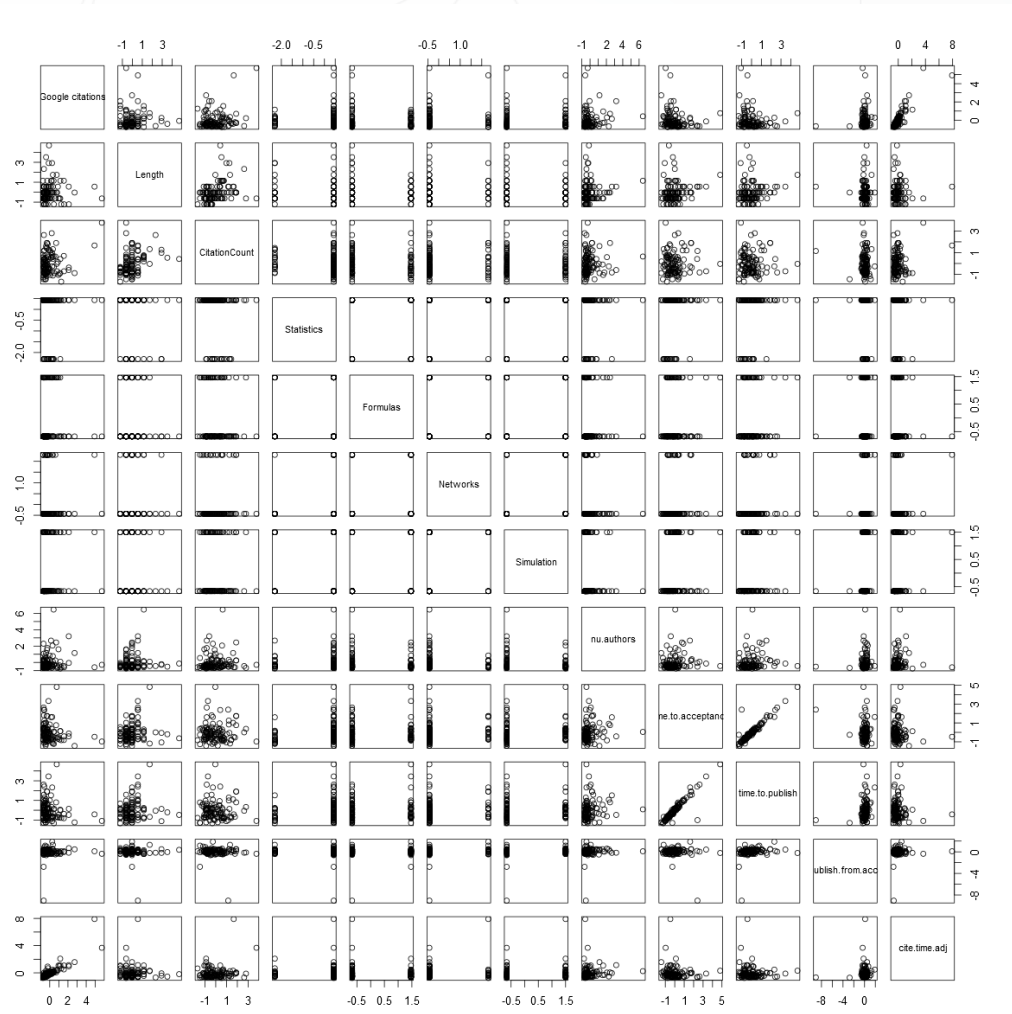
K-Nearest Neighbors (KNN)



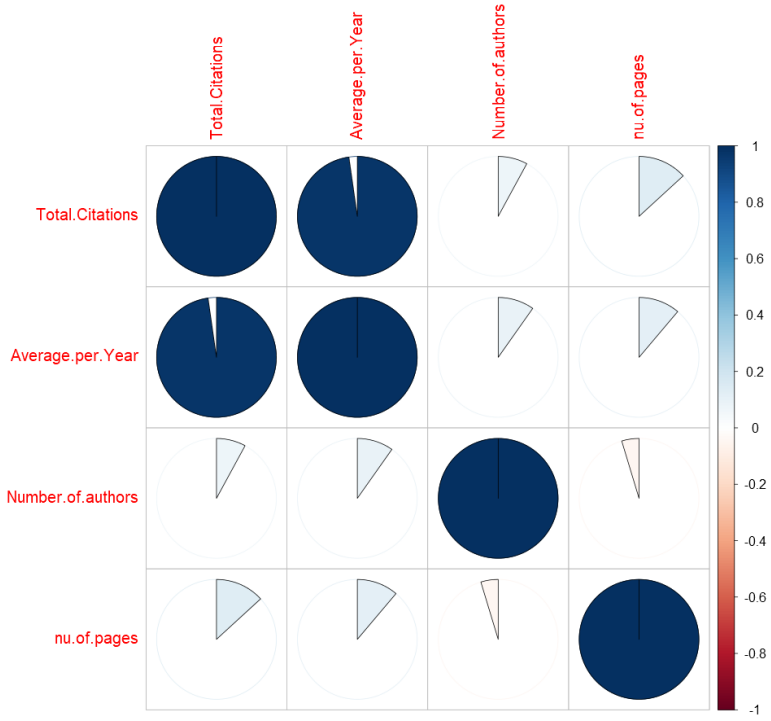


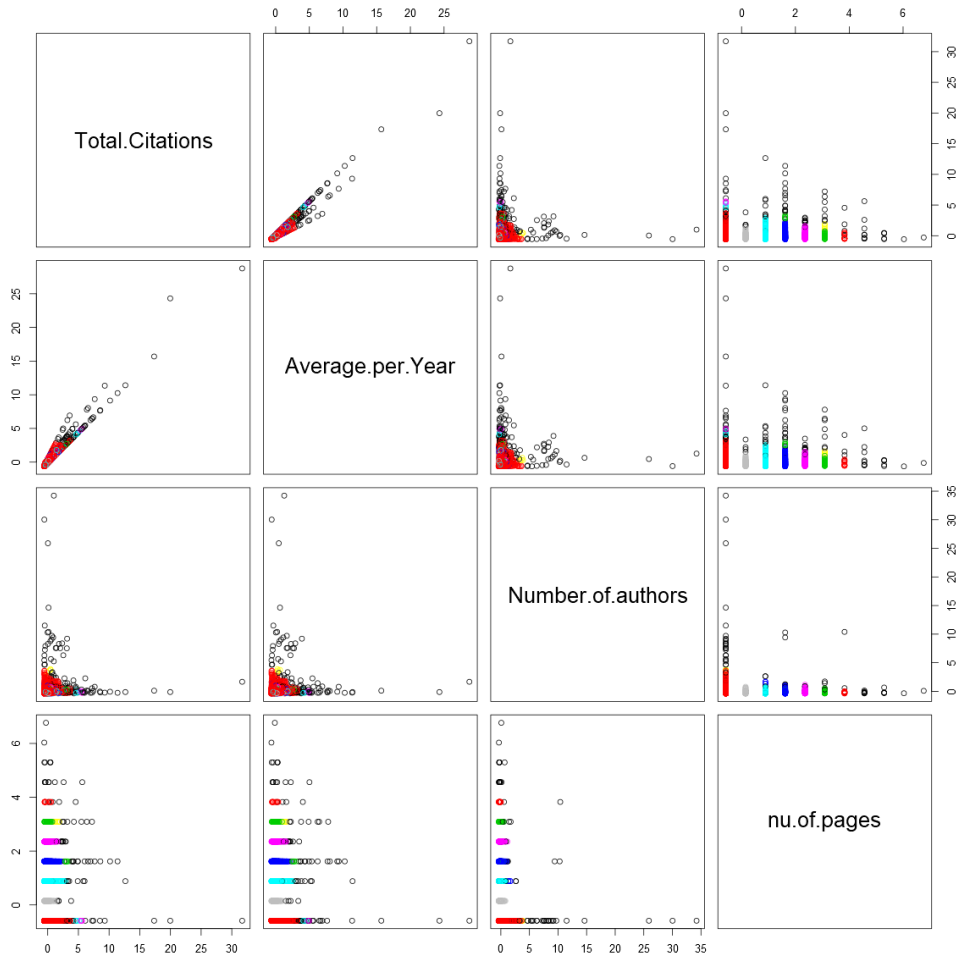
K-Nearest Neighbors (KNN) – 2 or 3 Groups



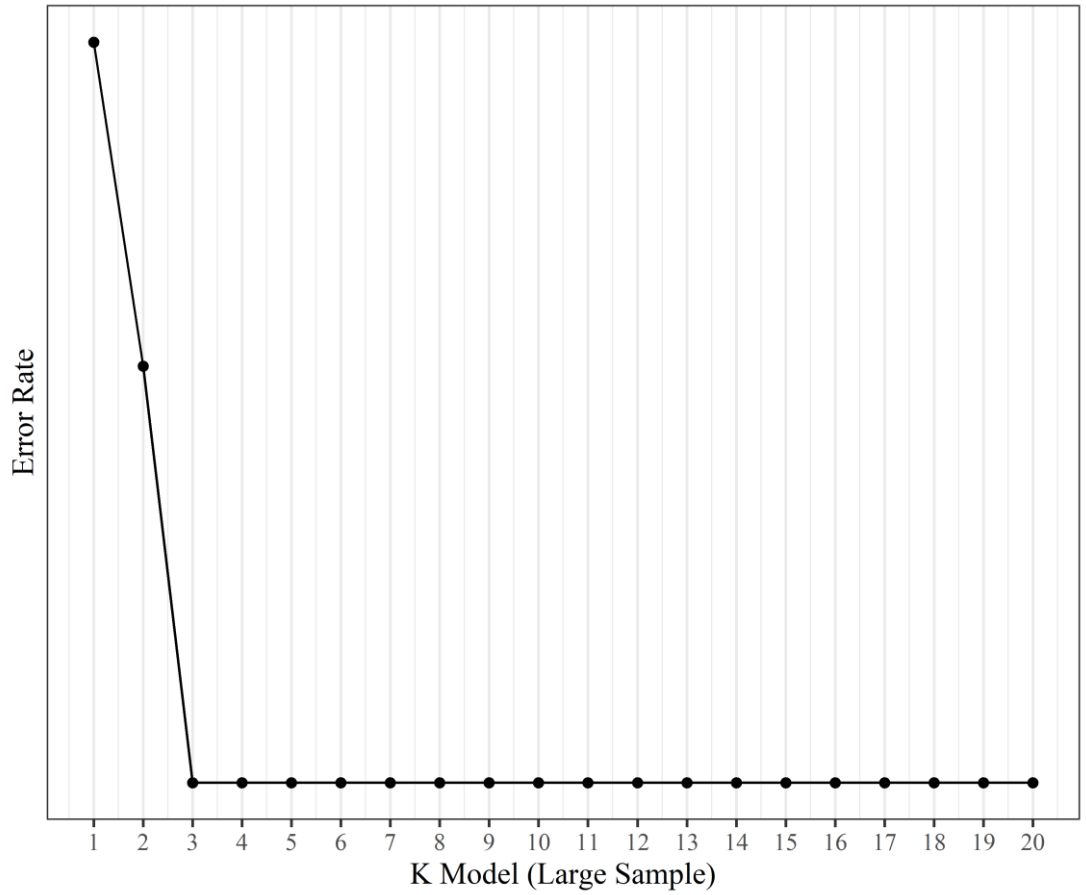


Dropping the Manual Features + Extracting the Population





Result: Distinct Groups of 3



Future Work

- Our model ***does discriminate*** between groups/clusters of Nature articles, but it's not clear what those groups are.

- The model has a long way to go:
  - Inconclusive
  - Data Difficult to collect
  - NLP will play a big role but full text articles are hard to come by.