Data Science Journals

NYC Datascience Academy

Web Scrap Project

In Suk Jang

8/2/2017

Data Science Research Papers

- Research hot topics?
- Active researchers?
- Frequency of publications?

DATA SCIENCE JOURNAL

Takahashi, M 2017 Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations. *Data Science Journal,* 16: 37, pp. 1–17, DOI: https://doi.org/10.5334/dsj-2017-037

RESEARCH PAPER

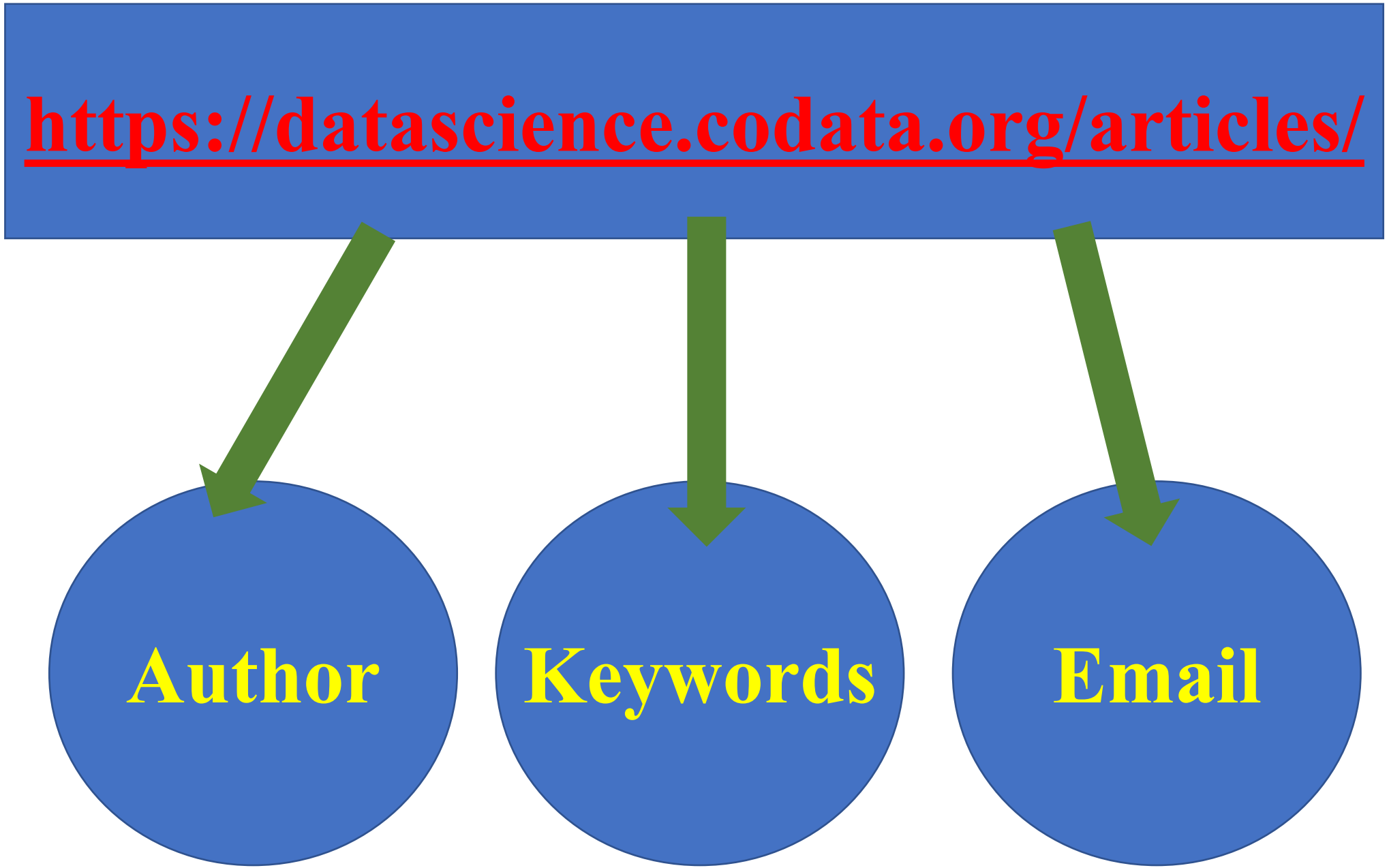
Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations

Masayoshi Takahashi

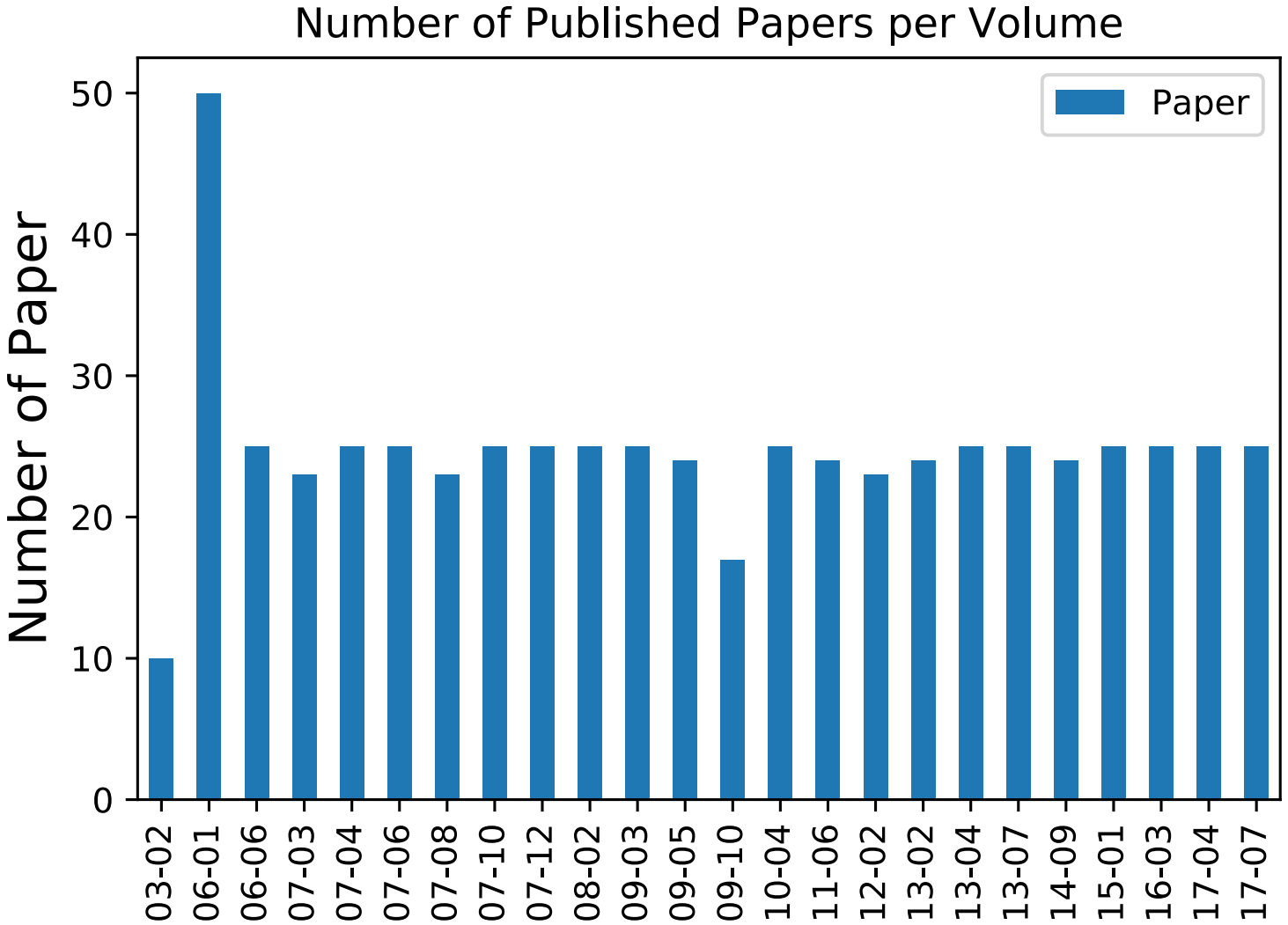
IR Office, Tokyo University of Foreign Studies, Tokyo, JP
mtakahashi@tufs.ac.jp

Incomplete data are ubiquitous in social sciences; as a consequence, available data are inefficient (ineffective) and often biased. In the literature, multiple imputation is known to be the standard method to handle missing data. While the theory of multiple imputation has been known for decades, the implementation is difficult due to the complicated nature of random draws from the posterior distribution. Thus, there are several computational algorithms in software: Data Augmentation (DA), Fully Conditional Specification (FCS), and Expectation-Maximization with Bootstrapping (EMB). Although the literature is full of comparisons between joint modeling (DA, EMB) and conditional modeling (FCS), little is known about the relative superiority between the MCMC algorithms (DA, FCS) and the non-MCMC algorithm (EMB), where MCMC stands for Markov chain Monte Carlo. Based on simulation experiments, the current study contends that EMB is a confidence proper (confidence-supporting) multiple imputation algorithm without between-imputation iterations; thus, EMB is more user-friendly than DA and FCS.

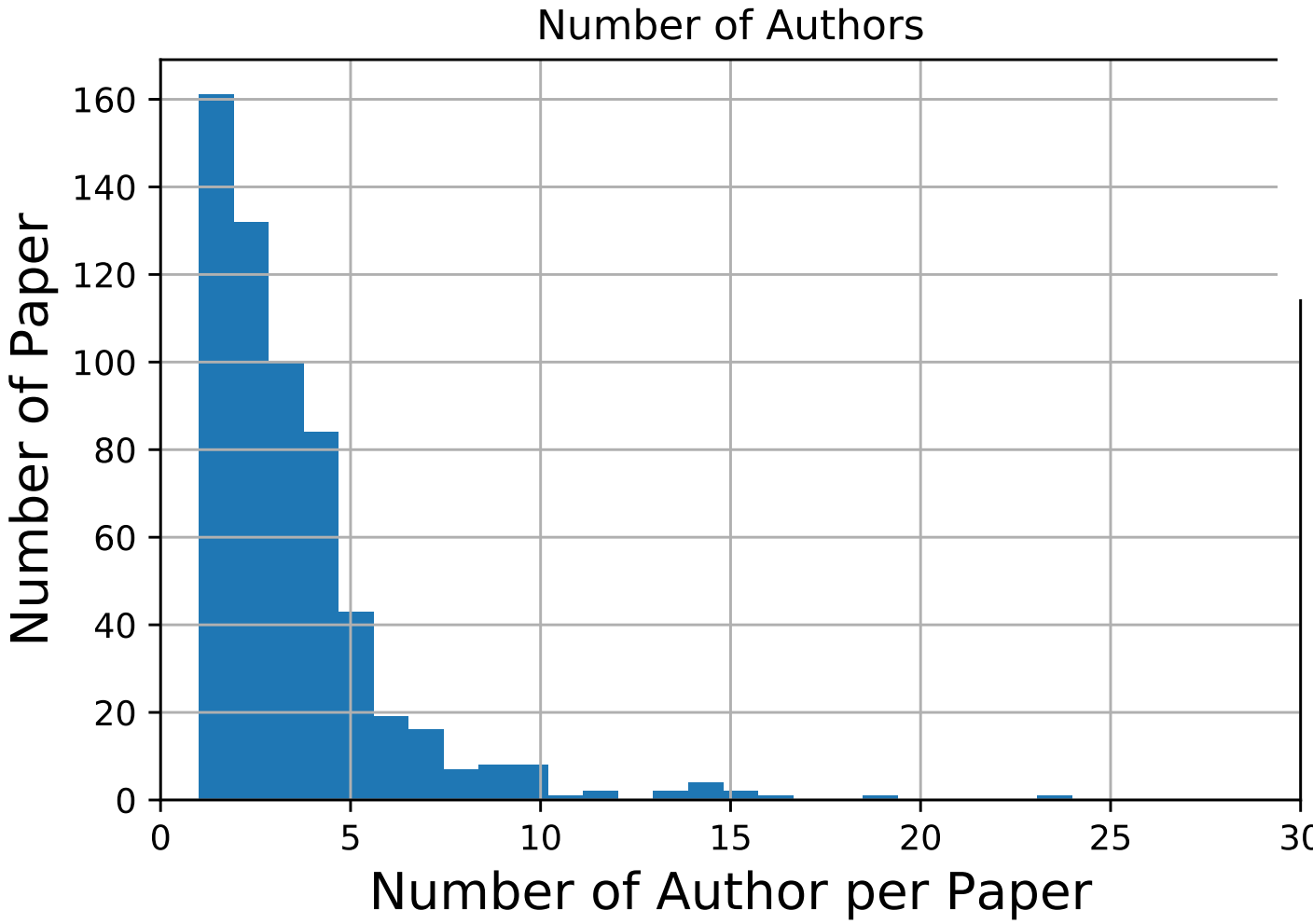
Keywords: MCMC; Markov chain Monte Carlo; Incomplete data; Nonresponse; Joint modeling; Conditional modeling


https://datascience.codata.org/articles/
Author
Keywords
Email

Frequency of Publications

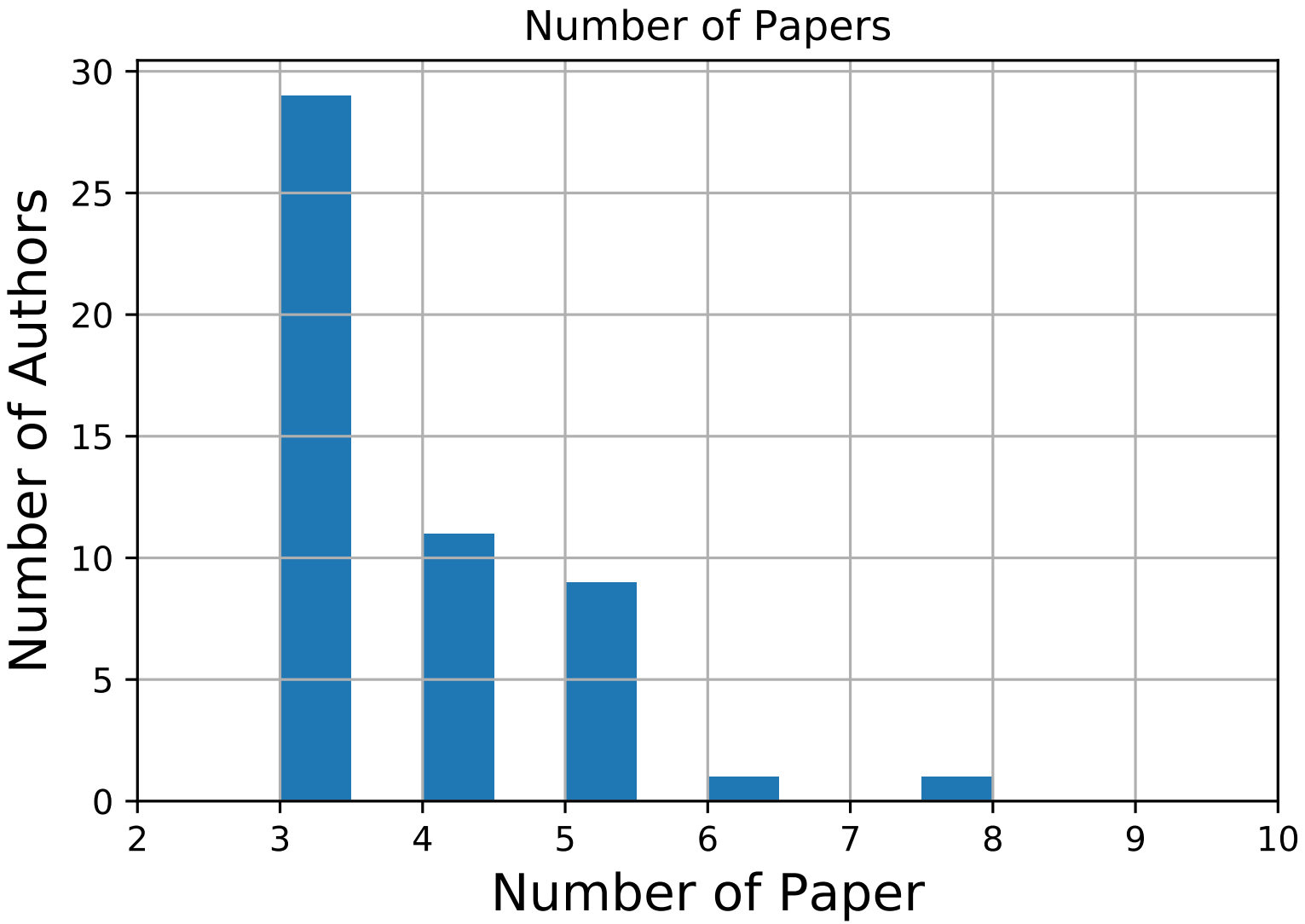


	Paper
count	24.000000
mean	24.666667
std	6.349712
min	10.000000
25%	24.000000
50%	25.000000
75%	25.000000
max	50.000000



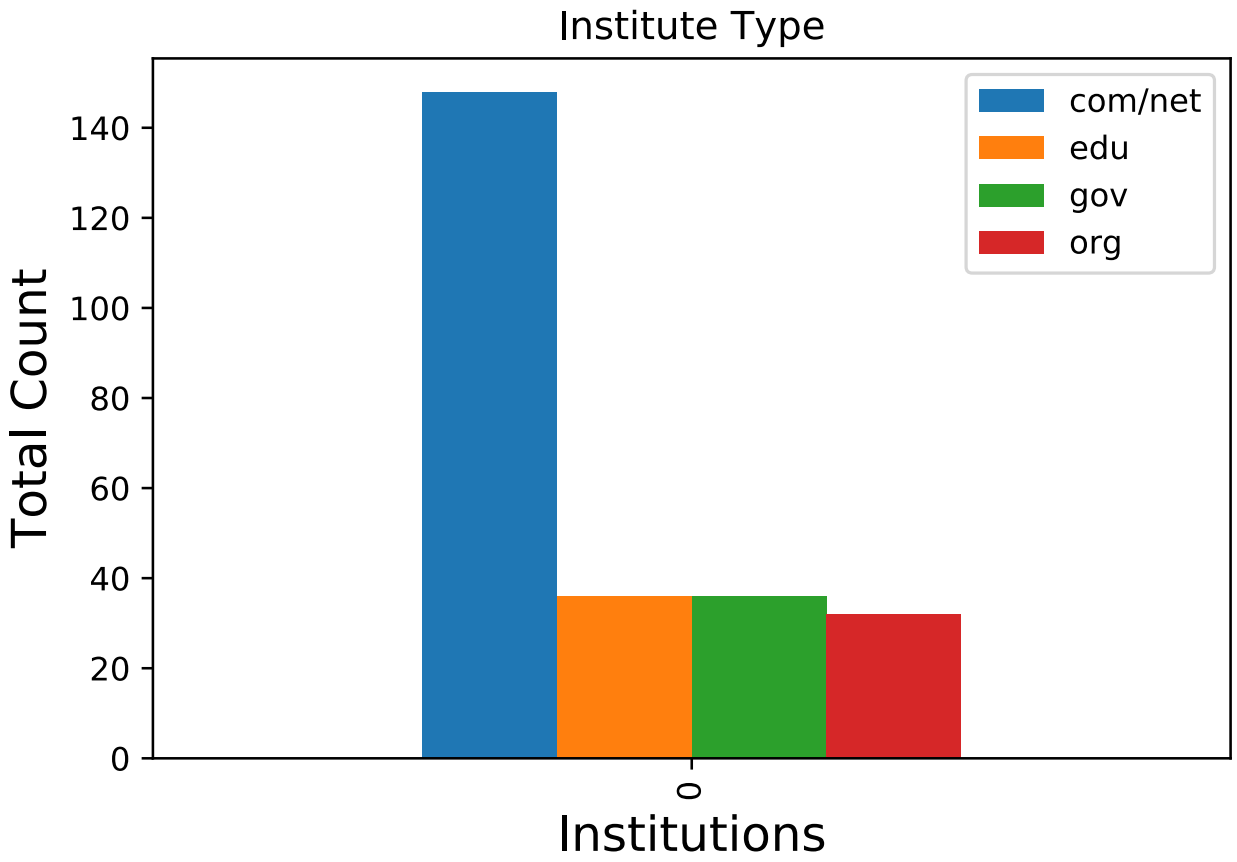

```
count     592.000000
mean        3.236486
std         2.711413
min         1.000000
25%         1.000000
50%         3.000000
75%         4.000000
max        24.000000
Name: Number, dtype: float64
```

```
count     115.000000
mean        7.434783
std         3.319613
min         5.000000
25%         5.000000
50%         6.000000
75%         9.000000
max        24.000000
```

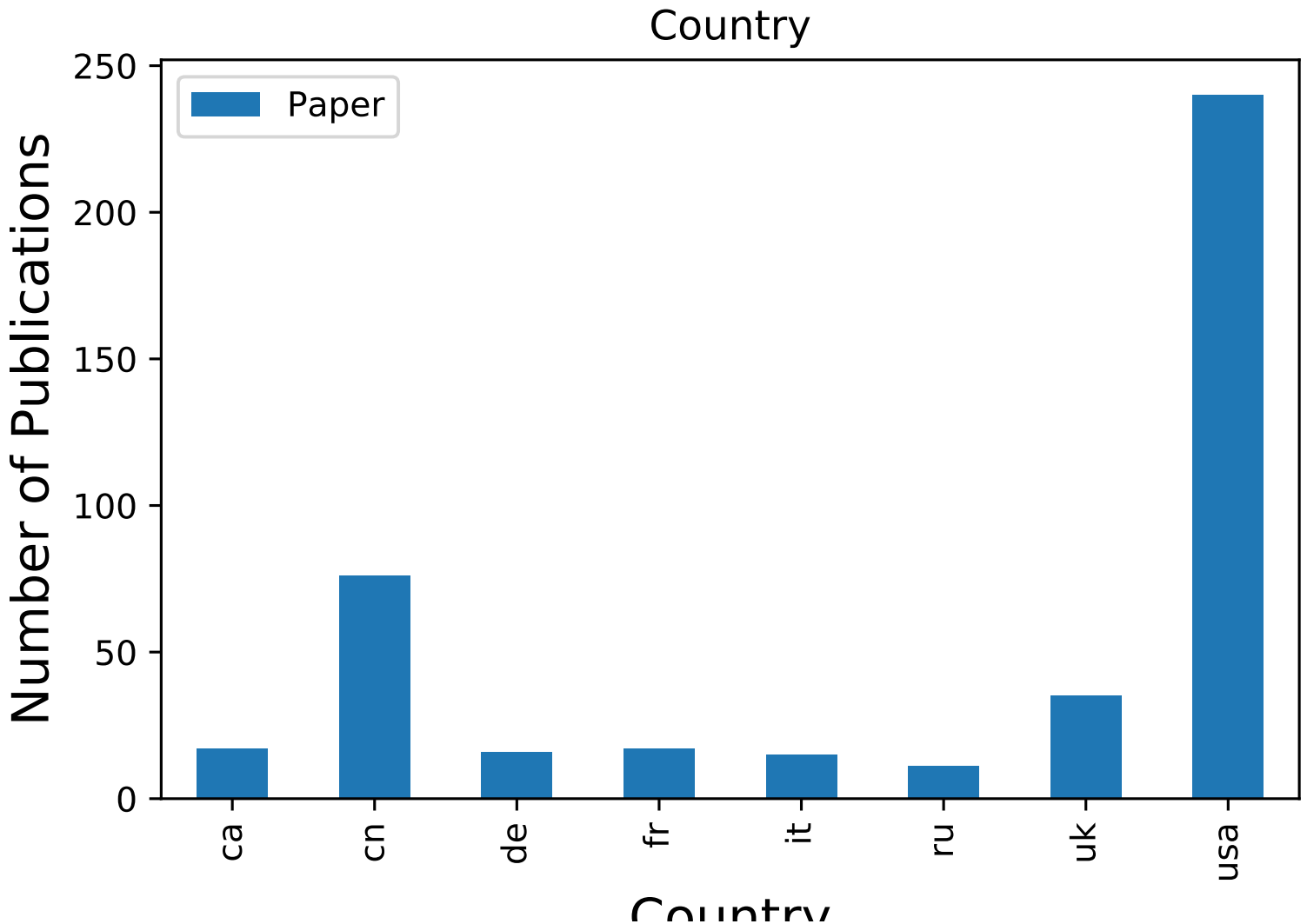


Paper	Author
1	1416
2	155
3	29
4	11
5	9
6	1
8	1

com/net	edu	gov	org
148	36	36	32

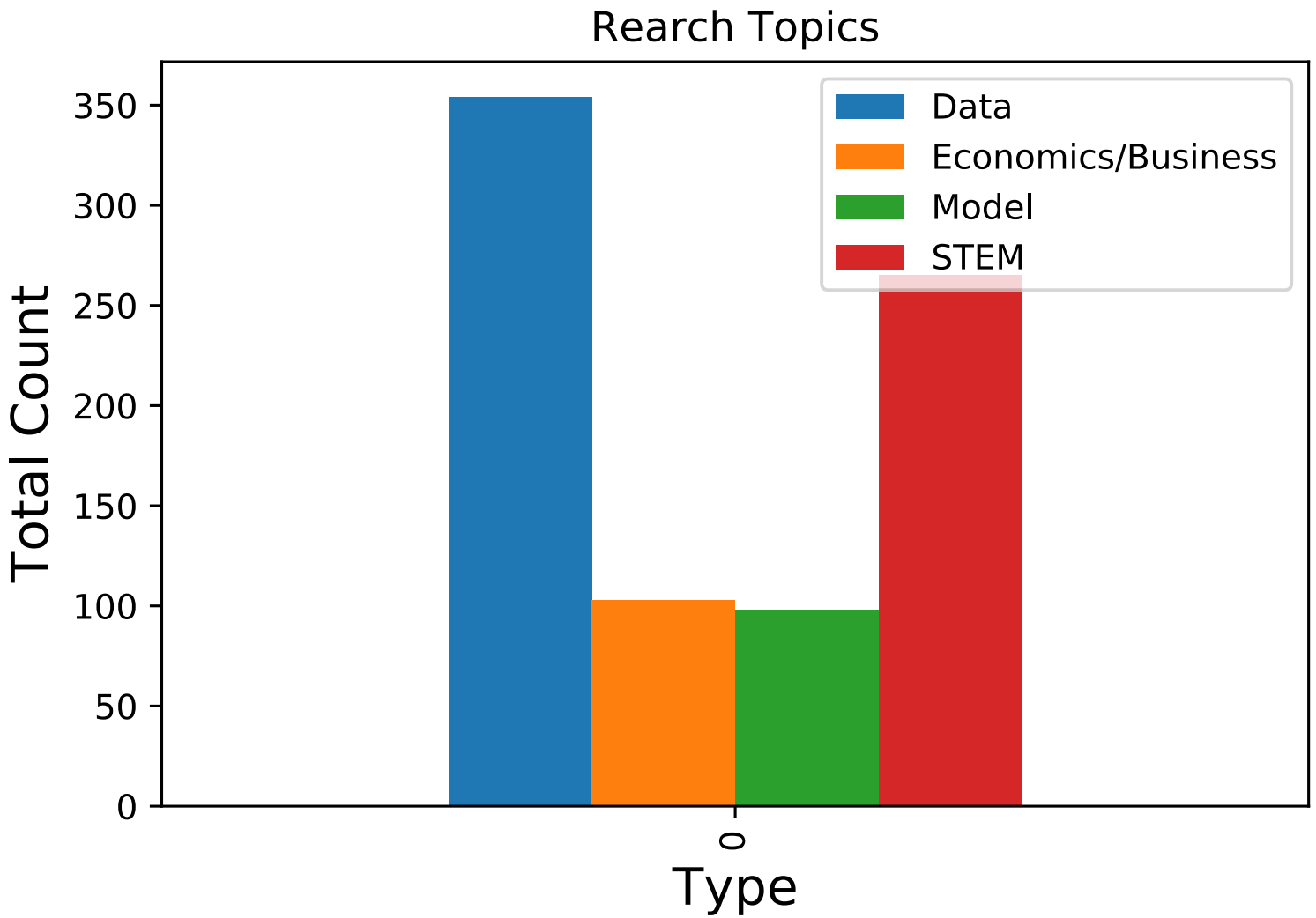


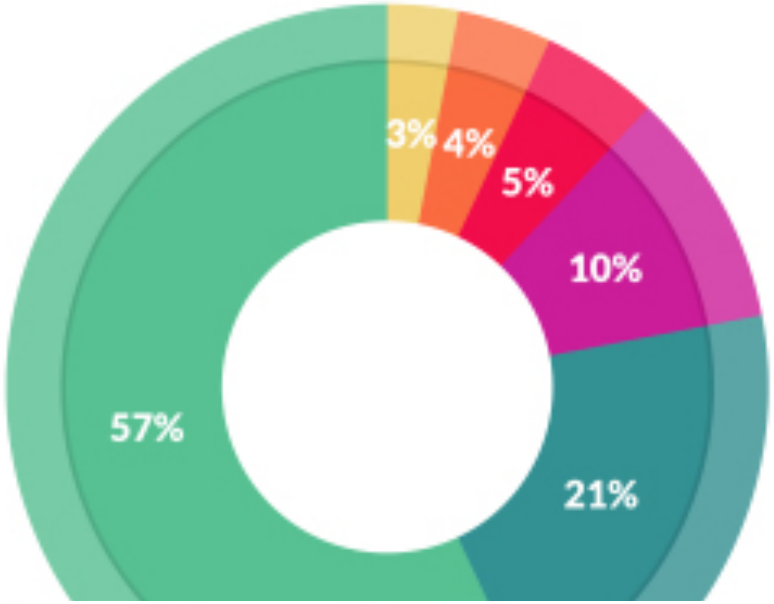
Name	Number
163.com	10
cass.org.cn	11
gmail.com	27
ubiquitypress.com	32
yahoo.com	12
yahoo.com.cn	10



	Country	Paper
0	ca	17
1	cn	76
2	de	16
3	fr	17
4	it	15
5	ru	11
6	uk	35
7	usa	240

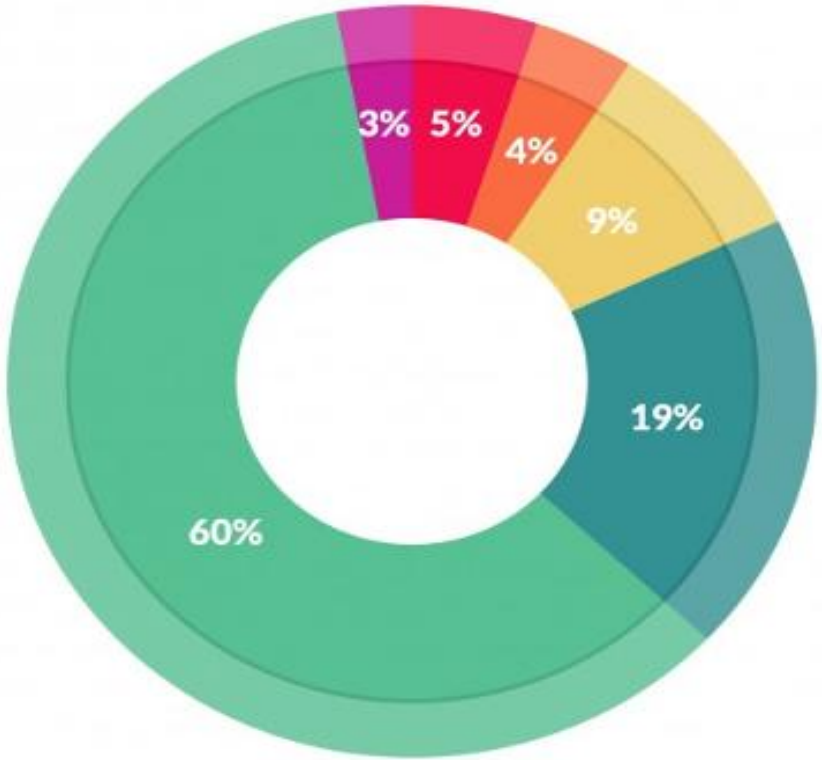
Data	Economics/Business	Model	STEM
354	103	98	265





What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*