Examining Citations of Natural Language Processing Literature

Saif M. Mohammad

Senior Research Scientist, National Research Council Canada

Saif.Mohammad@nrc-cnrc.gc.ca @SaifMMohammad

With the dawn of a new decade and NLP research becoming more diverse and more popular than it ever has been, this work looks back at the papers already published to identify broad trends in their impact on subsequent scholarly work.

Metrics of Research Impact (on subsequent scholarly work)

- Often derived from citations
  - number of citations, average citations, h-index, relative citation ratio, and impact factor (Bornmann and Daniel, 2009)
- However, citations do not always reflect quality or importance
  Impacted by:
  - systematic biases
  - atypical contributions
  - popularity of area
  - unethical practices (e.g. egregious self citations)

Nonetheless, given the lack of other easily applicable and effective metrics, citation metrics used as an imperfect but useful window into research impact

- often a factor in funding research and hiring scientists

Our Work

We extracted and aligned information from

- the ACL Anthology (AA)
  - full text and metadata for ~45K articles (as of June 2019)
- Google Scholar
  - Scholar profiles of authors who published ≥ 3 papers in AA
  - explicitly allowed by GS's robots exclusion standard

to create a dataset of tens of thousands of NLP papers and their citations

We examine nine questions to identify trends:

- overall, across paper types, publication venues
- over time (90s, 2000s,…)
- across research areas

Data

GScholar-NLP:

- citation information for 1.1 million papers
- includes citation counts for NLP papers and non-NLP papers

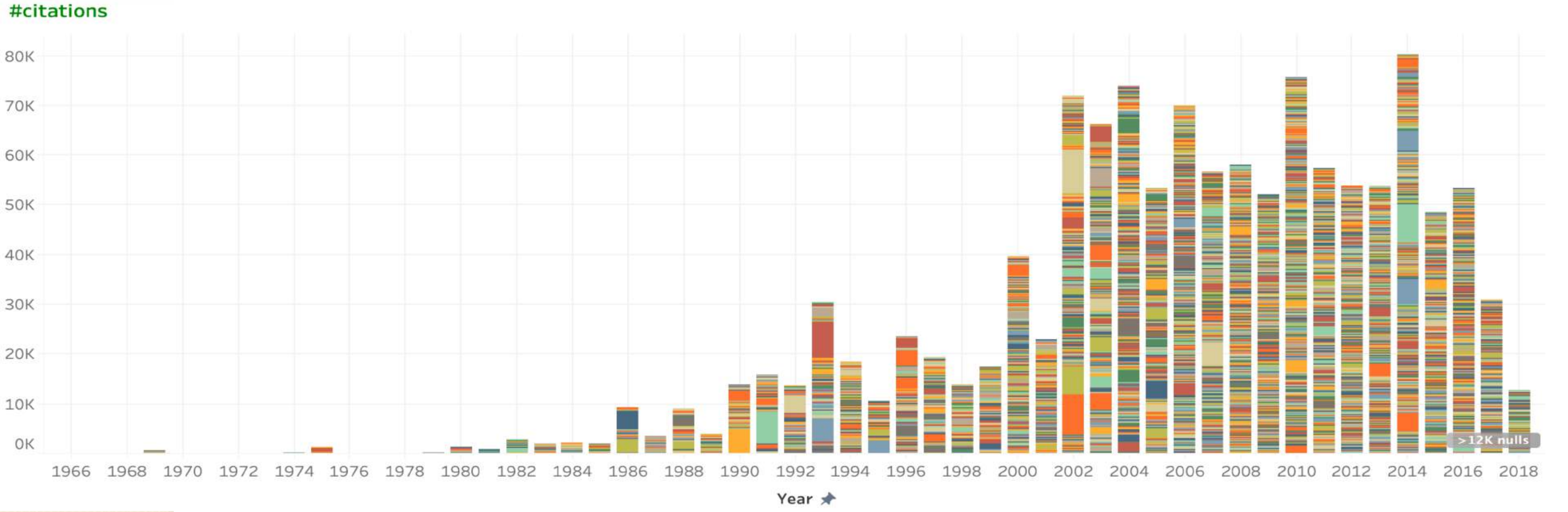
NLP Scholar:

- aligned the information across AA and GS
  - paper title, year of publication, and first author last name
- citation info for ~33K of the 45K papers in AA (about 74%): AA′

NLP Scholar: A Dataset for Examining the State of NLP Research. Saif M. Mohammad. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC-2020)*, May 2020, Marseille, France.

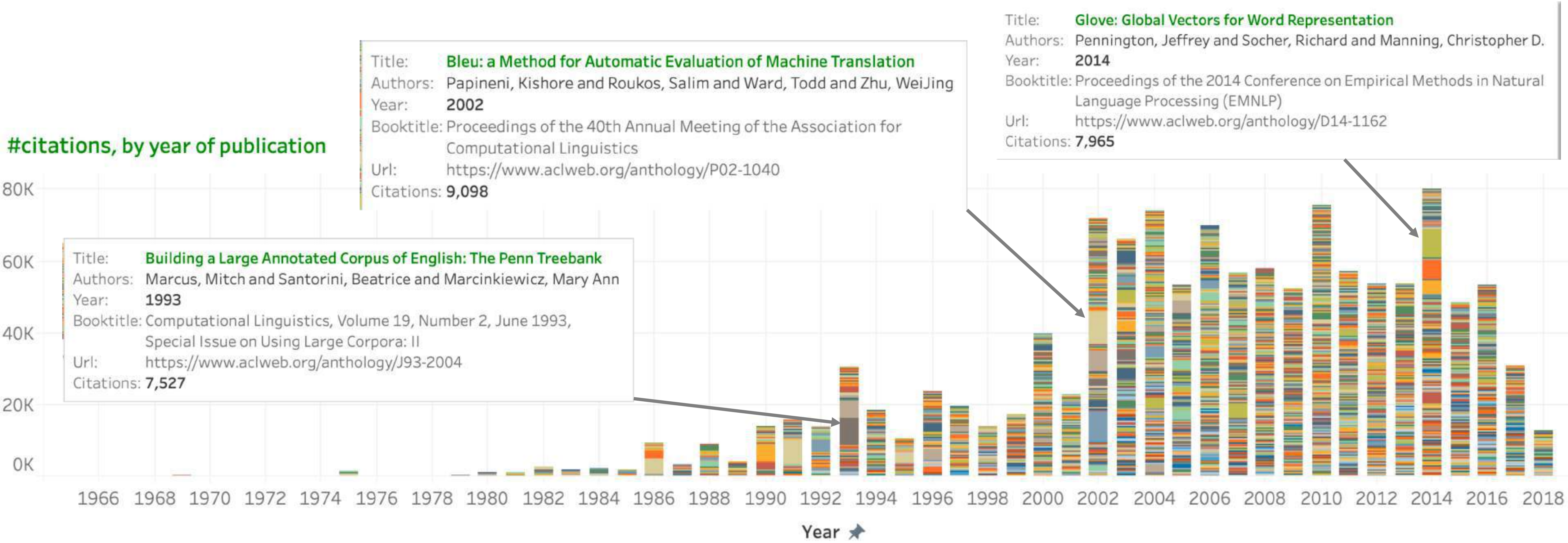
Q1. How many citations have the AA′ papers received? How is that distributed among the papers published in various years?

- ~1.2 million citations (as of June 2019)
- Figure below shows papers published over the years and their citations
  - colored segments corresponding to each of the papers
  - the height of a segment is proportional to the number of citations the paper has received



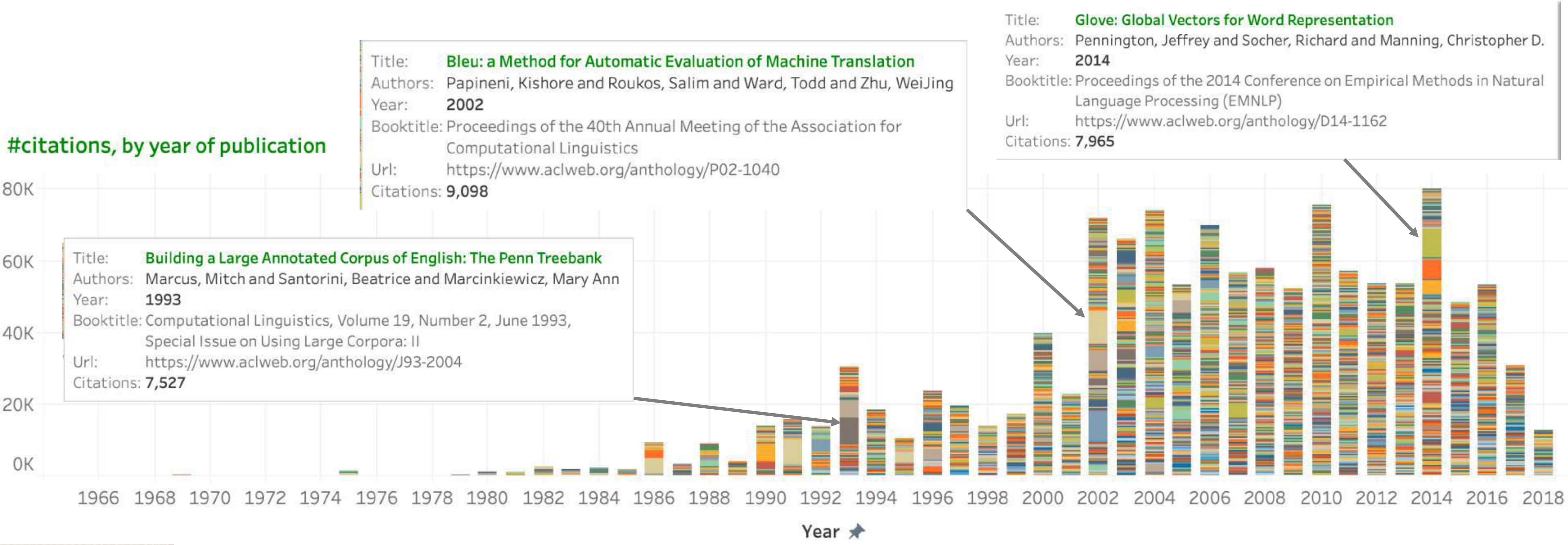
Q1. How many citations have the AA′ papers received? How is that distributed among the papers published in various years?

- ~1.2 million citations (as of June 2019)
- Figure below shows papers published over the years and their citations
  - colored segments corresponding to each of the papers



Q1. How many citations have the AA′ papers received? How is that distributed among the papers published in various years?

- marked jumps in the 1990s and then in the 2000s
- the 2010s papers will likely surpass the 2000s papers in the years to come



Q2. How well cited are individual AA′ papers?

What is the average, what is the median,
What is the distribution of citations?

How well cited are the different types of papers:
journal papers, main conference papers,
workshop papers, etc.?

For all further analyses:

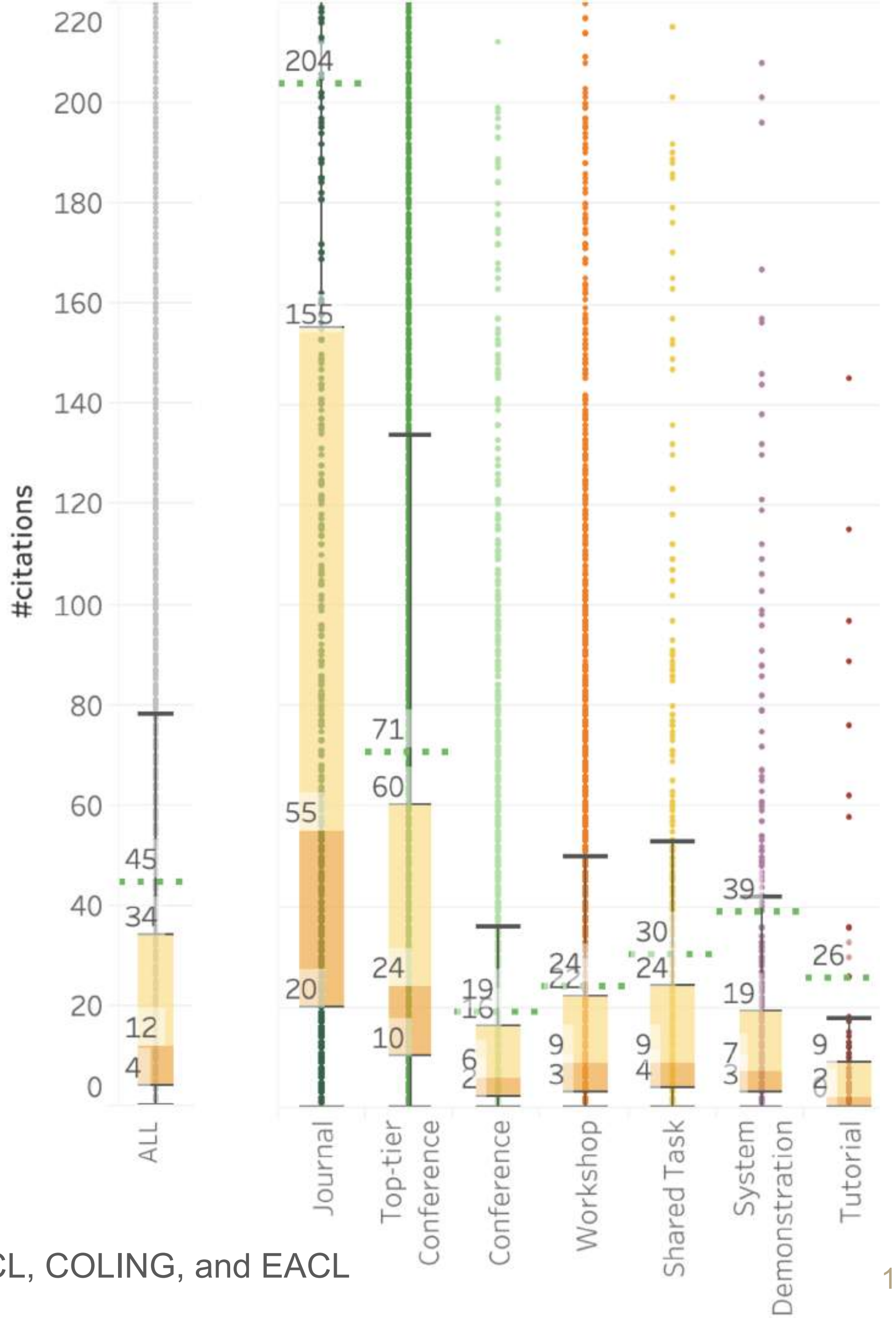
- we do not include AA′ papers published in 2017 or later
  (to allow for at least 2.5 years for the papers to collect citations)
- there are ~27K AA′ papers that were published from 1965 to 2016

Q2. How well cited are individual AA' papers?

What is the average, what is the median,
What is the distribution of citations?

How well cited are the different types of papers: journal papers, main conference papers, workshop papers, etc.?

- box and whisker plots
- shaded segments represent a quartile on either side of the median



* Top-tier: ACL, EMNLP, NAACL, COLING, and EACL

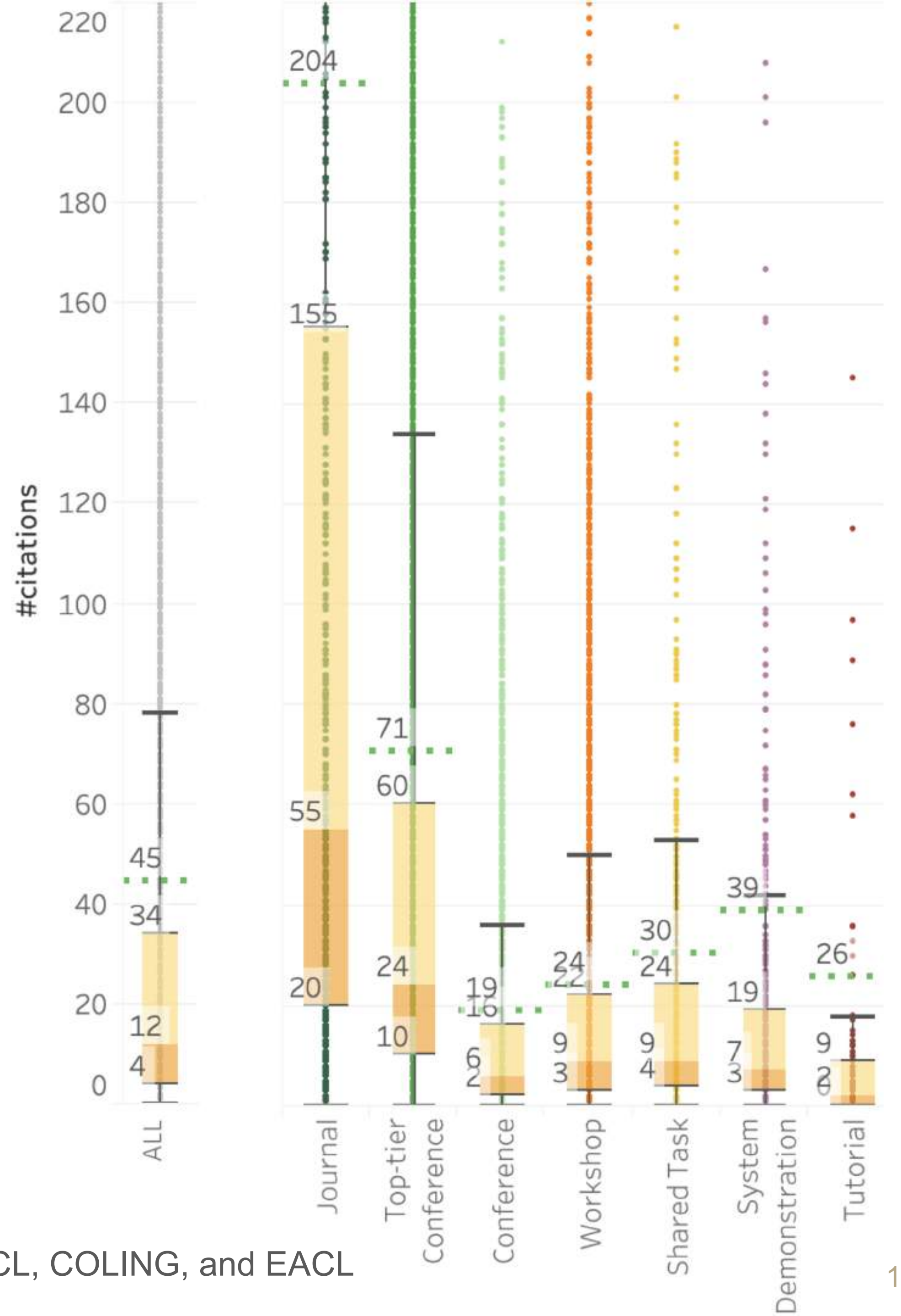
Q2. How well cited are individual AA' papers?

what is the average, what is the median,
what is the distribution of citations?

How well cited are the different types of papers:
journal papers, main conference papers,
workshop papers, etc.?

Overall:

- median: 12
  - 75% of the papers ≤ 34 citations
  - 25% of the papers ≤ 4 citations
- average: 45
  - markedly higher than the median
    (because of a small number of highly cited papers)



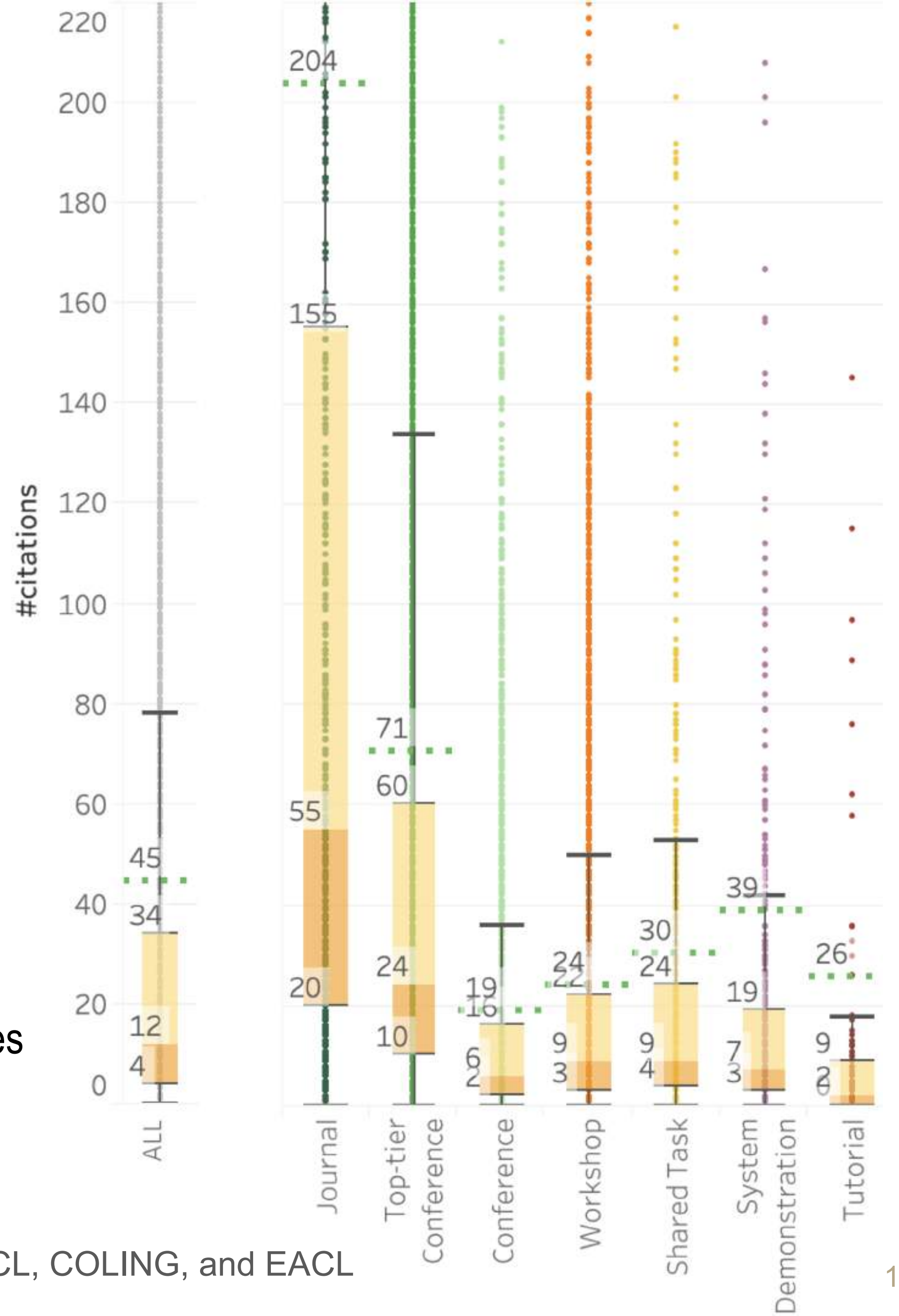
* Top-tier: ACL, EMNLP, NAACL, COLING, and EACL

Q2. How well cited are individual AA' papers?

what is the average, what is the median,
what is the distribution of citations?

How well cited are the different types of papers:
journal papers, main conference papers,
workshop papers, etc.?

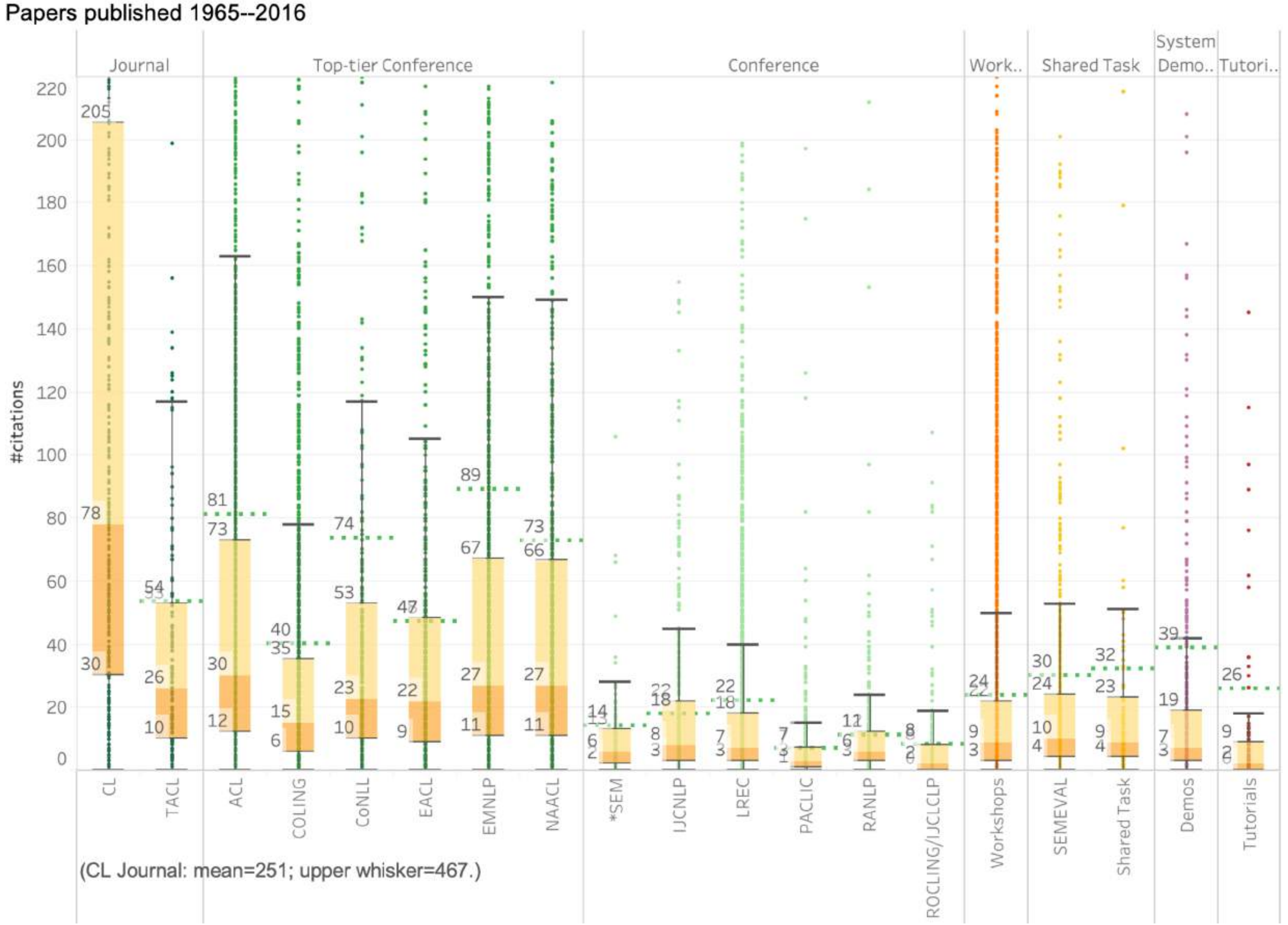
- highest median and average: journal papers
  - Journal papers are only 2.5% by volume
- journals > top-tier conferences > other conferences
  - statistically significant (Kolmogorov–Smirnov, p <.01)
- workshops, shared task papers > non-top-tier conferences
  - statistically significant (KS, p< .01)



Q5. How well cited are papers from individual NLP venues?

- CL Journal > ACL > EMNLP > NAACL
- SemEval, non-SemEval shared tasks, workshops > IJCNLP, LREC
- TACL< CL Jrnl., ACL, EMNLP, NAACL

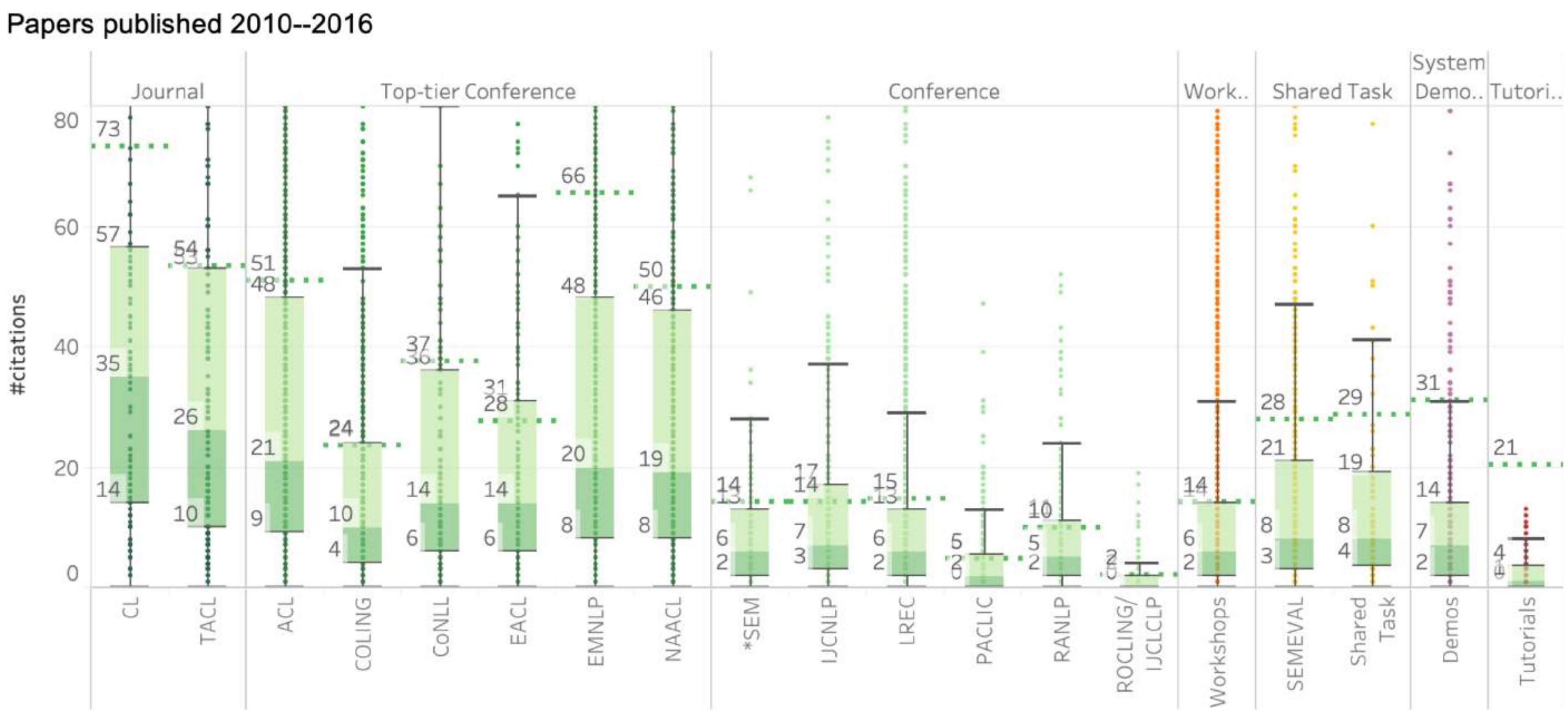
However, TACL only began publishing since 2013…



(CL Journal: mean=251; upper whisker=467.)

National Research Council Canada Conseil national de recherches Canada

Q5. How well cited are papers from individual NLP venues?



When considering only the 2010--2016 papers:

- TACL's citation numbers are second only to CL Journal
- The gap between CL Journal and ACL and EMNLP is considerably reduced

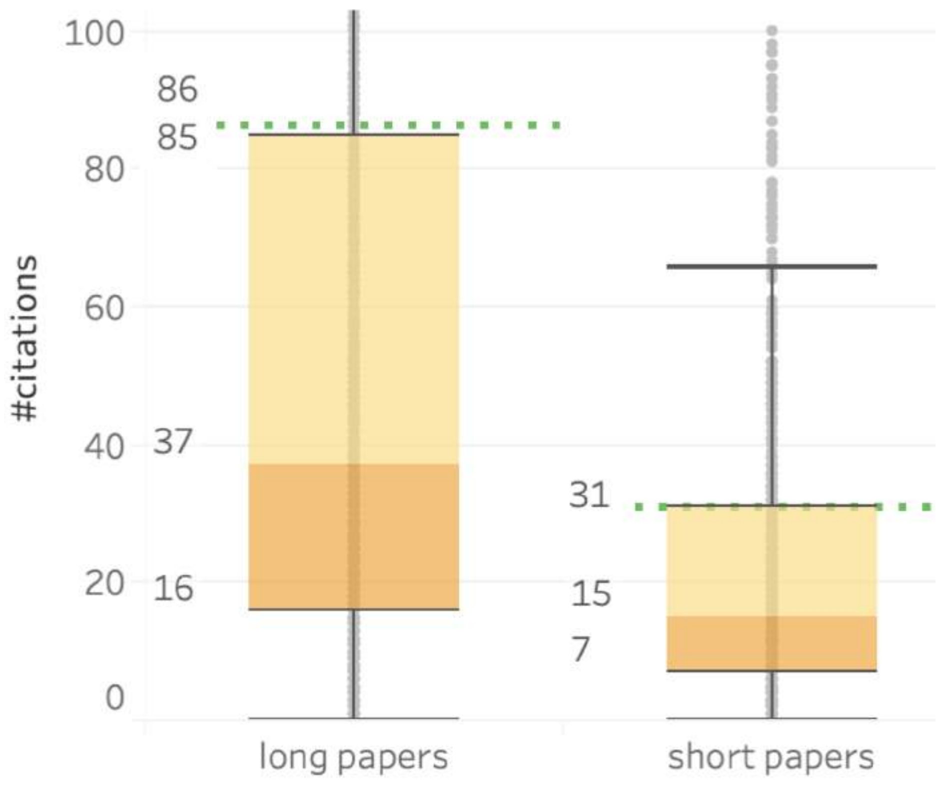
Q6. How well cited are long and short ACL main conference papers?

Short papers -- a place for focused contributions

- introduced: ACL 2003
- venue with most short papers: ACL
- acceptance rate: similar to long papers

So we compare long and short papers published at ACL since 2003

- the two distributions are statistically different (KS, $p < .01$)
- on average, long papers get almost three times as many citations as short papers
- the median for long papers is two-and-half times that of short papers



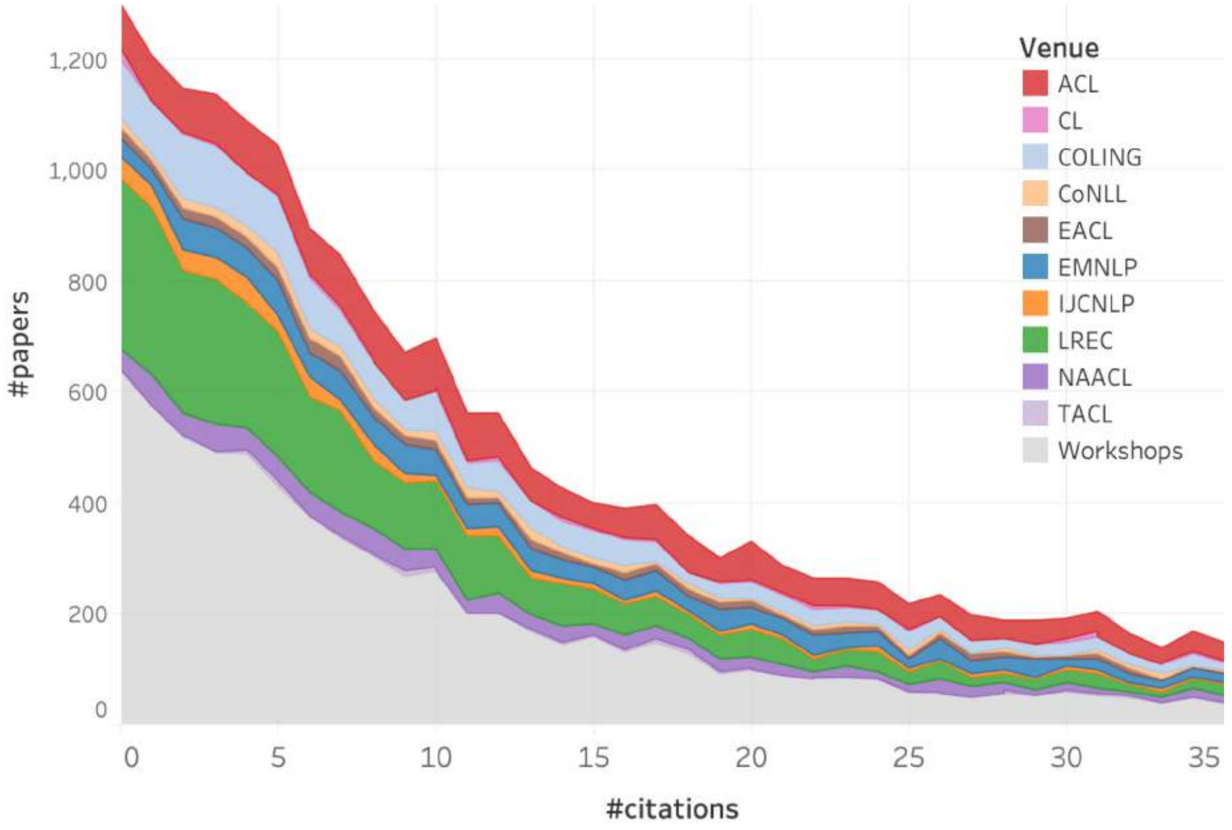
Q7. How do different venues and paper types compare in terms of the volume of papers pertaining to various amounts of citation?

A stream graph of #papers by #citations:

- contributions of each of the venues, paper types are stacked one on top of another
- for a given point on the citations axis (say k), the width of the stream corresponds to the number of papers with k citations

Observations:

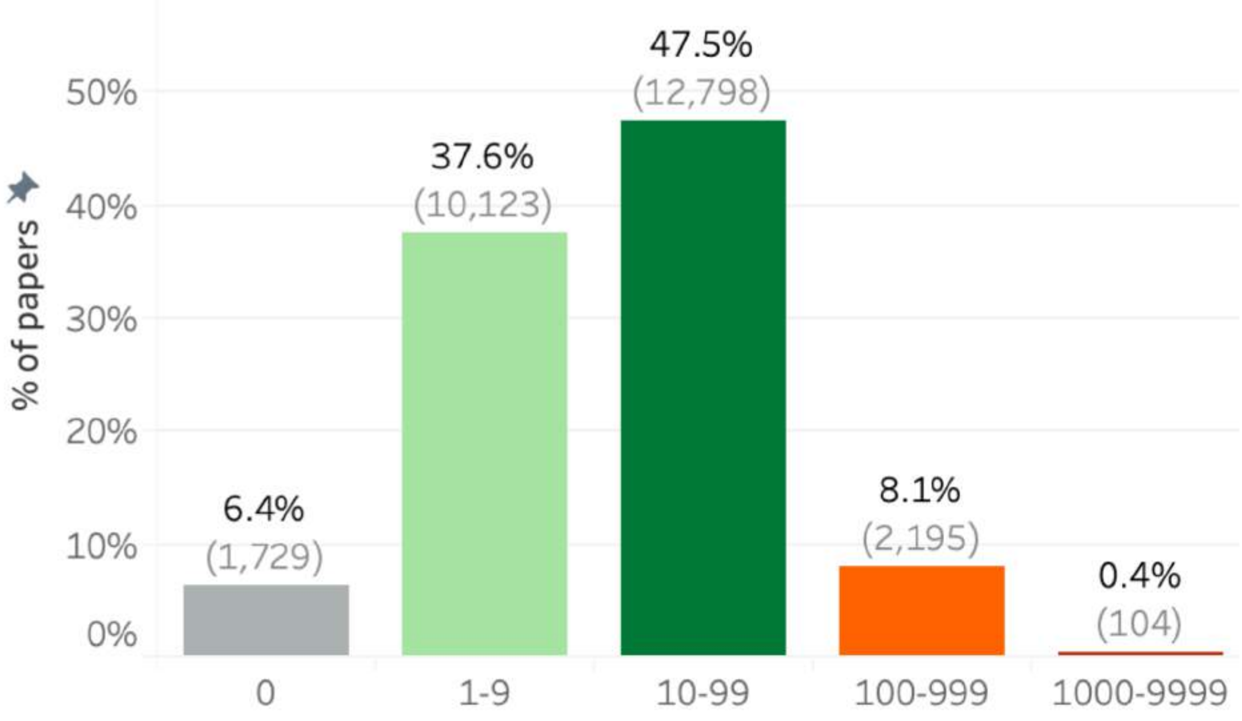
- a power law distribution
- workshop papers (light grey) are the most numerous, followed by LREC (green)
- the workshops and LREC produce lots of papers that are cited ten or more times
- as one considers higher citations, top-tier conferences become more dominant.



Q8. *What percentage of papers are cited more than 10 times? How many papers are cited 0 times?*

Figure shows the percentage of AA′ papers in citation bins: 0, 1–9, 10–99, 1000–9999.

- 6.4% of the papers have 0 citations
  - some papers in the 1-9 bin only received self-citations
- ~56% of the papers are cited ten or more times
- If a paper has more than 100 citations
  - It is in the 91.5th percentile



How do these numbers look in medical sciences, physics, linguistics, machine learning, psychology?

Summary

- Aligned information from the ACL Anthology and Google Scholar
- Analyzed ~27K NLP papers to examine patterns of citation
  - recorded how well NLP papers are cited (average, median, percentiles)
    - overall and across paper types, venues, etc.
  - only about 56% of the papers are cited ten or more times
  - CL Journal has the most cited papers
    - top-tier conferences are markedly closer to CL for recent years
  - on average, long papers get almost three times as many citations as short papers
  - workshops and LREC contribute a marked volume <35 citation papers

In Separate Work:

Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations. ACL 2020.

NLP Scholar: An Interactive Visual Explorer for Natural Language Processing Literature. ACL 2020 (Demo).

- A tool to find related work

Future Work:

- measure involvement of the industry in NLP publications over time
- analyze NLP papers that are published outside of the ACL Anthology
- to compare patterns of citations in NLP with those in other fields
- develop richer ways of capturing scholarly impact

Created by Symbolon from Noun Project

Project page for NLP Scholar: http://saifmohammad.com/WebPages/nlpscholar.html

- data
- Interactive visualizations
- limitations and ethical considerations

Contact: Saif.Mohammad@nrc-cnrc.gc.ca @SaifMMohammad