Softcite: Data-Driven Software Visibility in Science

James Howison, Patrice Lopez, Caifan Du, Norman Gilmore, Johanna Cohoon, Nick Adams, Karthik Ram

Howison Lab @ TEXAS

science-miner

Berkeley Institute for Data Science

TagWorks

Alfred P. Sloan Foundation

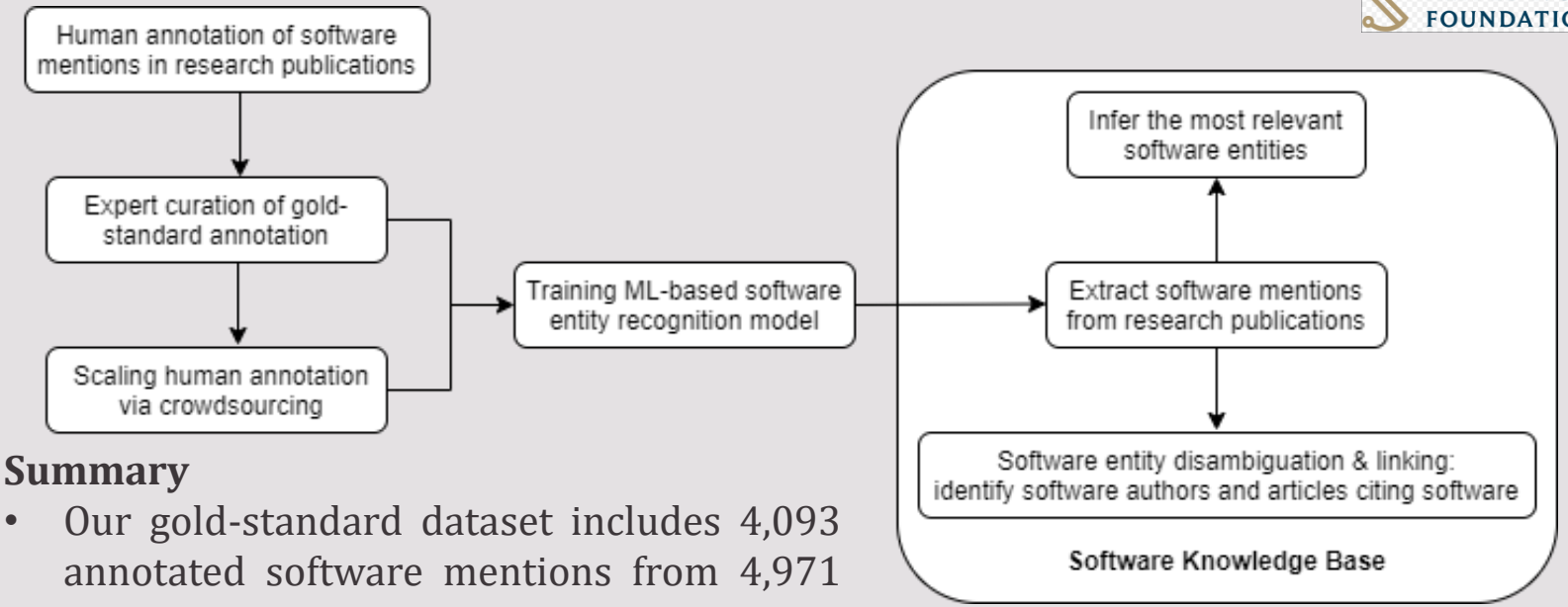
What is Softcite?

Softcite is a discovery engine for research software solutions. If you ever wonder the software toolkits used by other data science researchers, *Softcite* is conceived for meeting such needs. By mining software mentioned in research publications, *Softcite* informs technology decisions for data-intensive research.

Why software visibility?

Today, research software is still largely invisible to research databases & systems of information retrieval. The publishing, indexing, and citation of software are not yet well-standardized. As an outcome, good software misses its users while users miss good software. Besides, making others' software stack used in analysis workflow visible is critical to research reproducibility; while facilitating collaboration based on interoperability. Finally, making software visible makes credit for developers more likely.

Overview of Softcite



Summary

- Our gold-standard dataset includes 4,093 annotated software mentions from 4,971 publications
- In crowdsourced annotation, TagWorks collected 11,454 task responses on 2,743 article fragments from Mechanical Turk workers in one month

Outcome

- *Softcite* dataset release: http://doi.org/10.5281/zenodo.4445202
- *Softcite* software entity recognizer: https://github.com/ourresearch/software-mentions

Future work

Softcite enables scalable discovery of research software solutions, making research software visible while providing decision support for data-intensive research. We hope Softcite could support more utilities and systems that increase the visibility of research software, including CiteAs.org.