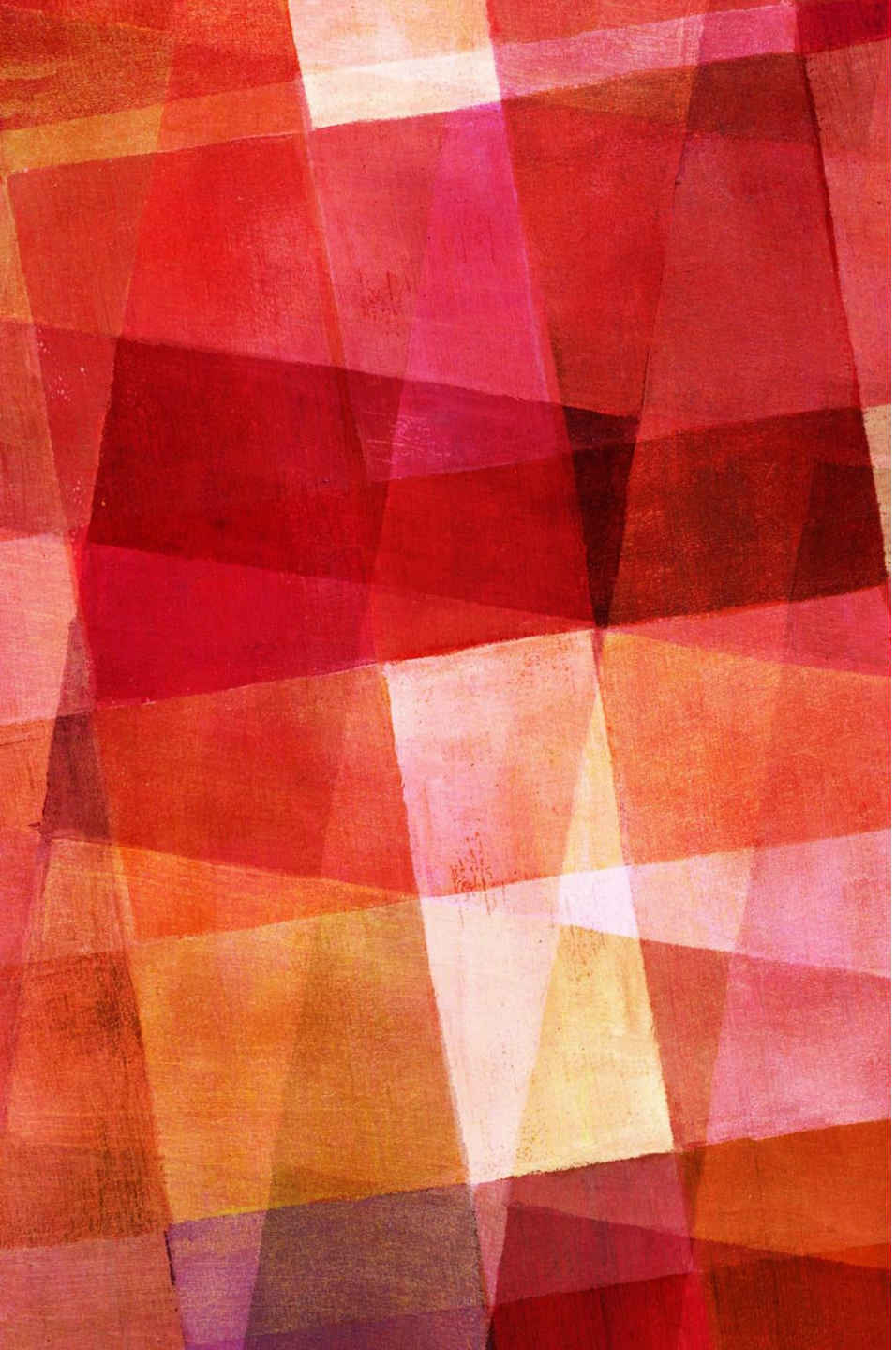
When the subjects of metadata embrace the statistical learning

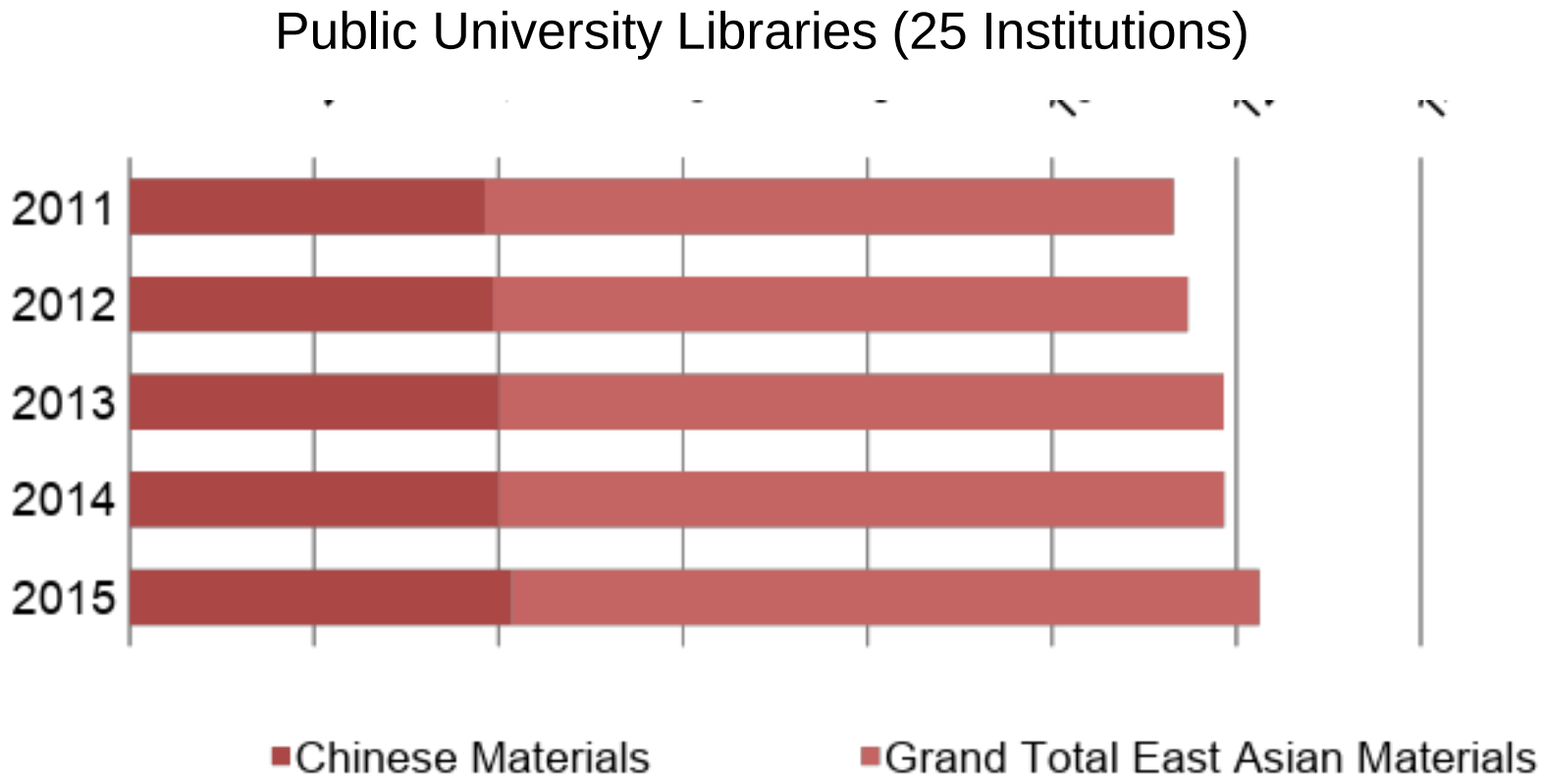
Anlin Yang, East Asian Cataloging Librarian

University of Iowa Libraries



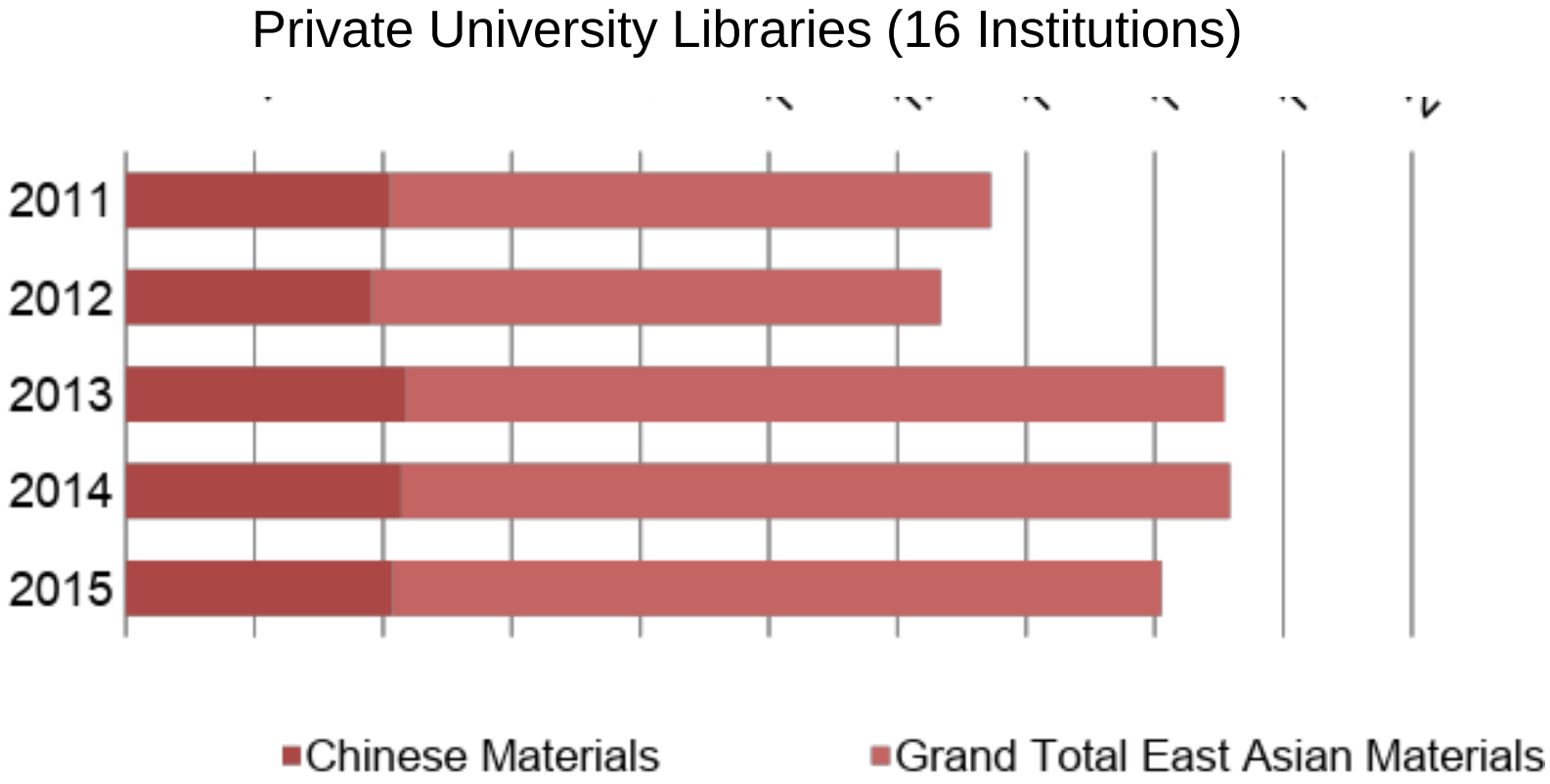
- INVESTIGATION
  - The growing number of Chinese materials
  - The new change of subjects
- IMPLEMENTATION
  - Current statistical learning methods/frameworks
  - The assumptions of statistical learning application on subjects of metadata

The Number of Chinese Volumes in U.S. University Libraries



Source: CEAL Statistics Data

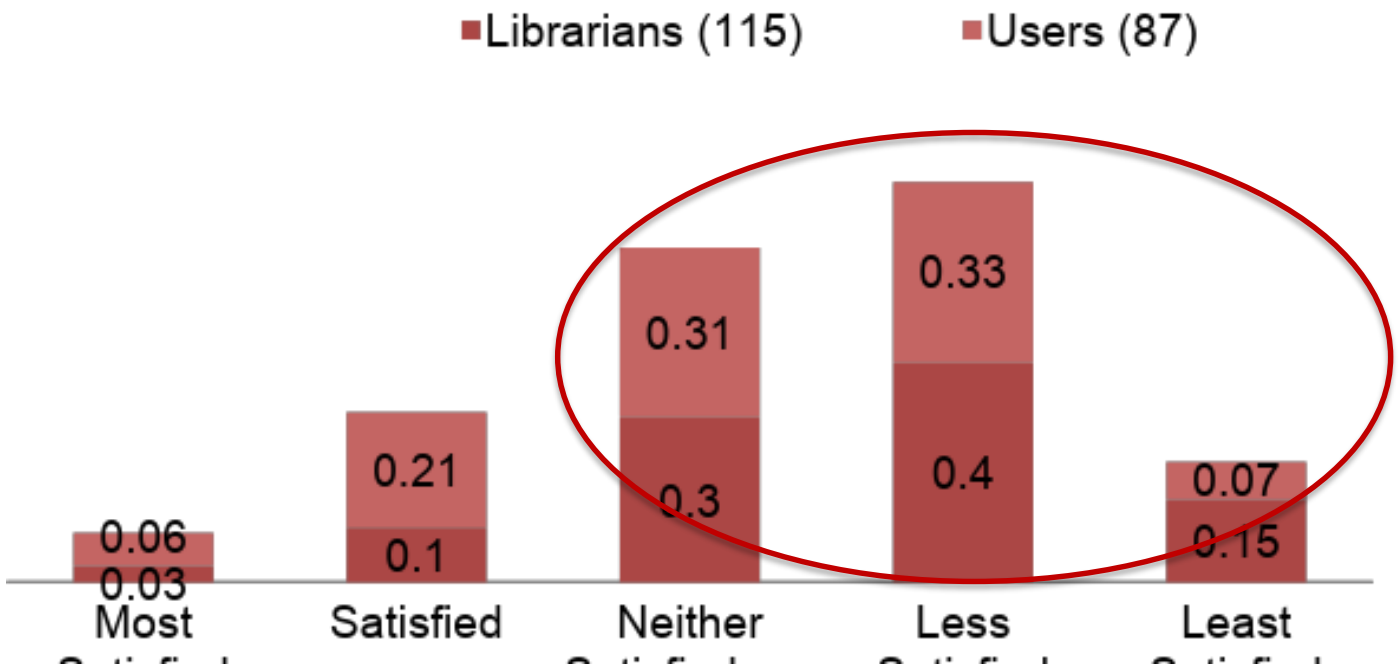
The Number of Chinese Volumes in U.S. University Libraries



Source: CEAL Statistics Data

The Satisfaction to Search Non-Roman Scripts

Satisfaction with using controlled English subjects to find Non-Roman scripts



Source: El-Sherbini, M., & Chen, S. (2011). An assessment of the need to provide non-Roman subject access to the library online catalog. *Cataloging & Classification Quarterly, 49*(6), 457-483.

http://dx.doi.org/10.1080/01639374.2011.603108

Some Controlled Vocabularies and Thesaurus Release Timeline

1898
LCSH

1940
MeSH
For Medicine

1996
ERIC Thesaurus
For Education

1997
Transportation Research Thesaurus
Based on NCHRP 20-32(2)

1998
Getty Vocabularies

- The Art & Architecture Thesaurus (AAT)
- The Getty Thesaurus of Geographic Names (TGN)
- The Cultural Objects Name Authority (CONA)
- The Union List of Artist Names (ULAN)

2016
COAR Vocabularies

- Resource type to identify the genre of a research resource: October 2016
- Access mode to declare the degree of 'openness 'of a resource (draft): May 2017

- *Increasing update frequency*
- *More detailed on subject classification*

What challenges we meet on the subjects of metadata?

- The continuous growing number of Chinese materials

  How fast we could manage those metadata?

- The difficulties of language barriers for searching Chinese resources

  How easy we could swing between different languages?

- The rapid changes on academic studies

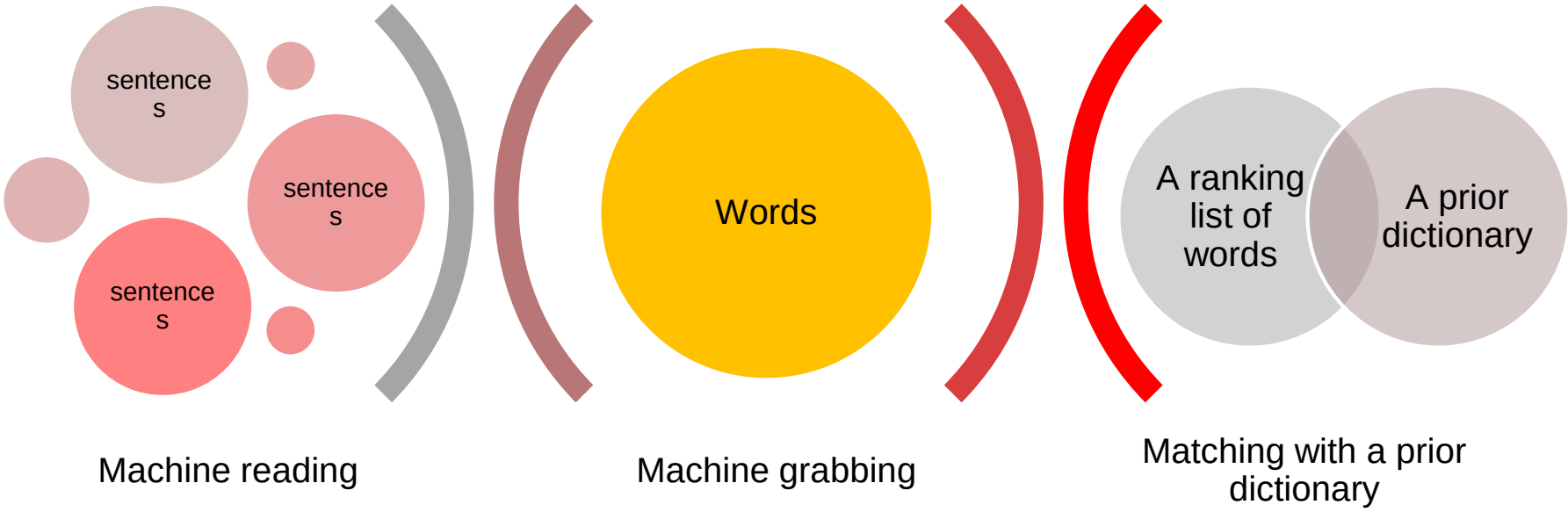
  How possible we could learn brand new academic knowledge continually?

Core Ideas For Us

- Segmentation
- Words / phrases discovery
- Build lexicon

Segmentation

- Word Dictionary Model (WDM)



Source: Olivier, D. C. (1968). *Stochastic grammars and language acquisition mechanisms: a thesis.* Harvard University.

Words / Phrases Discovery

- The ambiguity of Chinese words segmentation

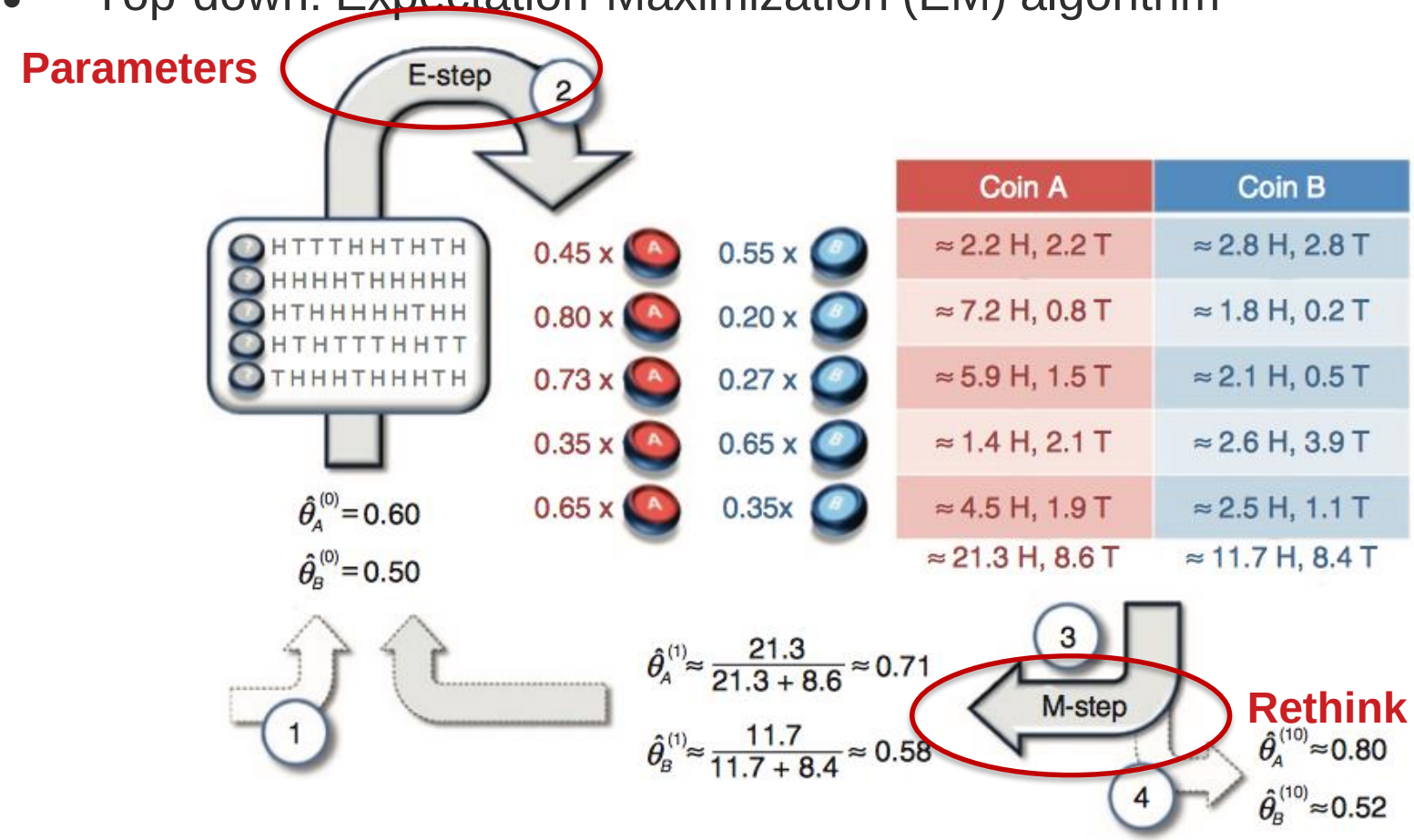
“土地公有政策”(Policy of Public Ownership of Land)

- Correct: 土地 公有 政策 (land, public ownership, policy)
- The possibility of words segmentation: 土地公 有 政策 (the earth god, has, policy)

Source: Chang, J. S., & Su, K. Y. (1997). An unsupervised iterative method for Chinese new lexicon extraction. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 2, Number 2, August 1997*, *2*(2), 97-148.

Segmentation

- Top-down: Expectation-Maximization (EM) algorithm

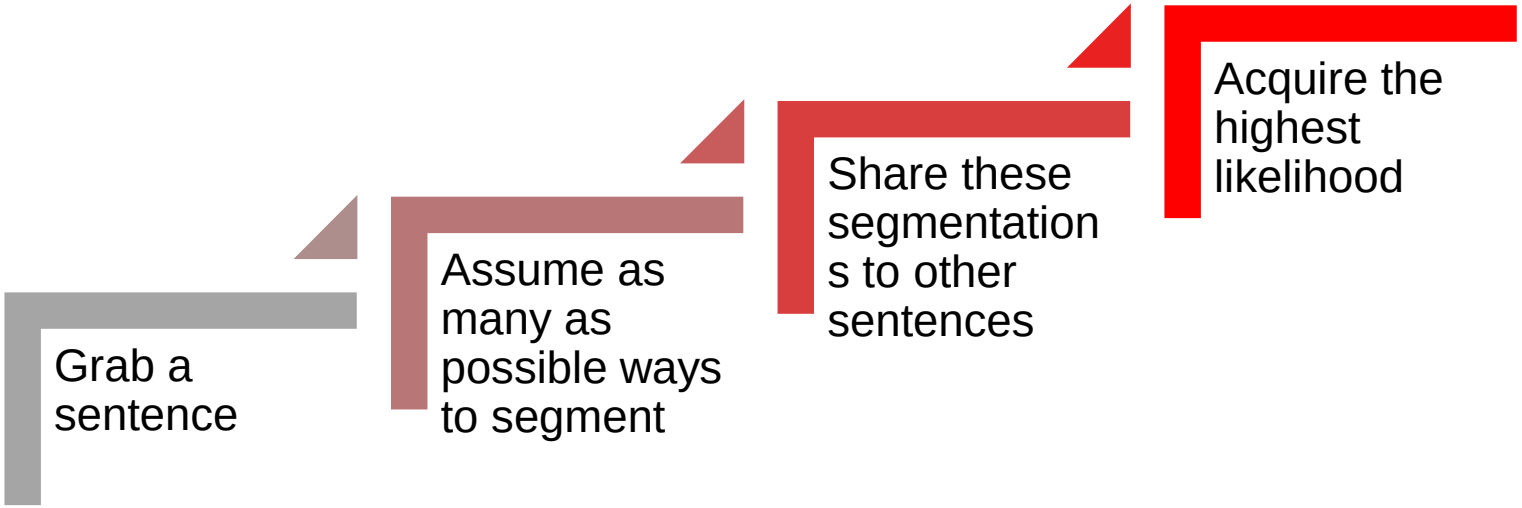


Source: Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm?. *Nature biotechnology*, *26*(8), 897.

doi:10.1038/nbt1406

Segmentation

- EM algorithm and its Chinese words catching

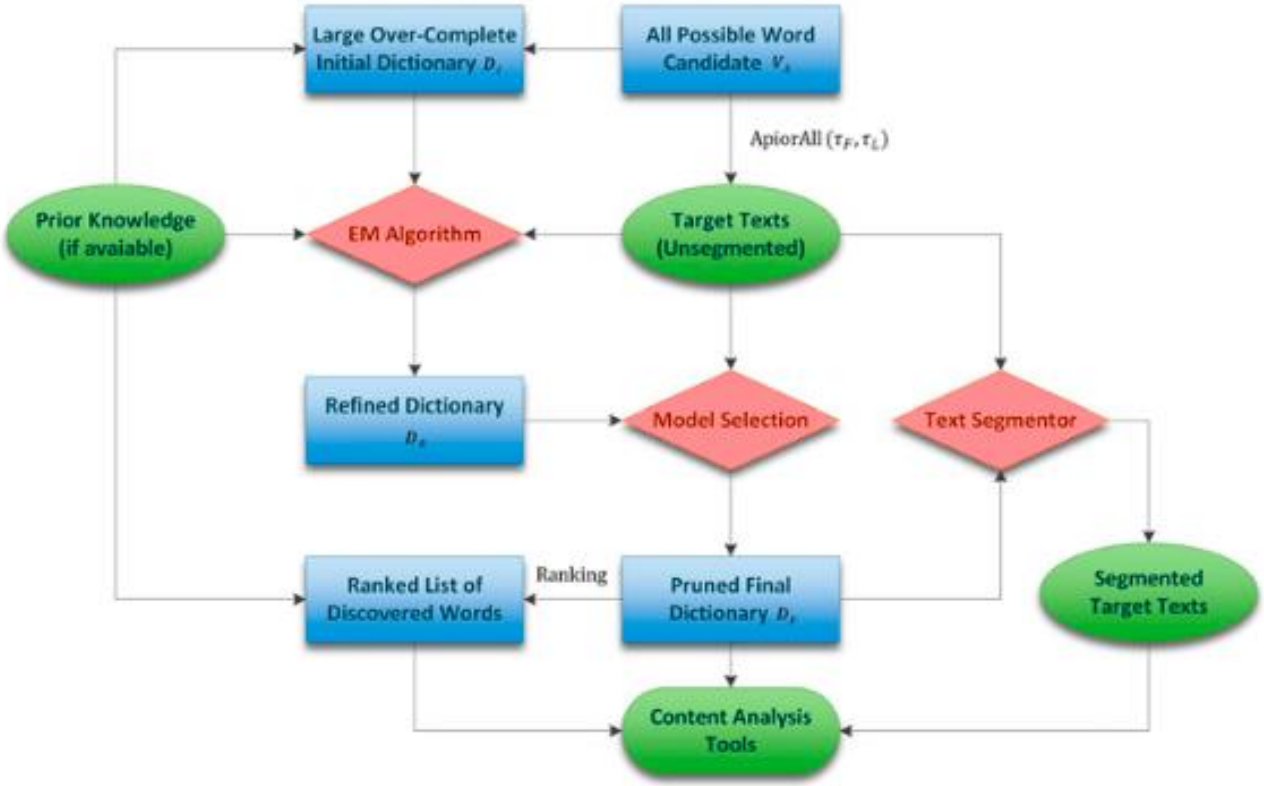


Segmenter	Recall(%)	Precision(%)
Soft-counting	65.65	71.91
(after postprocessing)	97.72	91.05

Source: Ge, X., Pratt, W., & Smyth, P. (1999, August). Discovering Chinese words from unsegmented text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 271-272). ACM.

Build Lexicon

- TopWORDS: The unsupervised analysis of Chinese texts



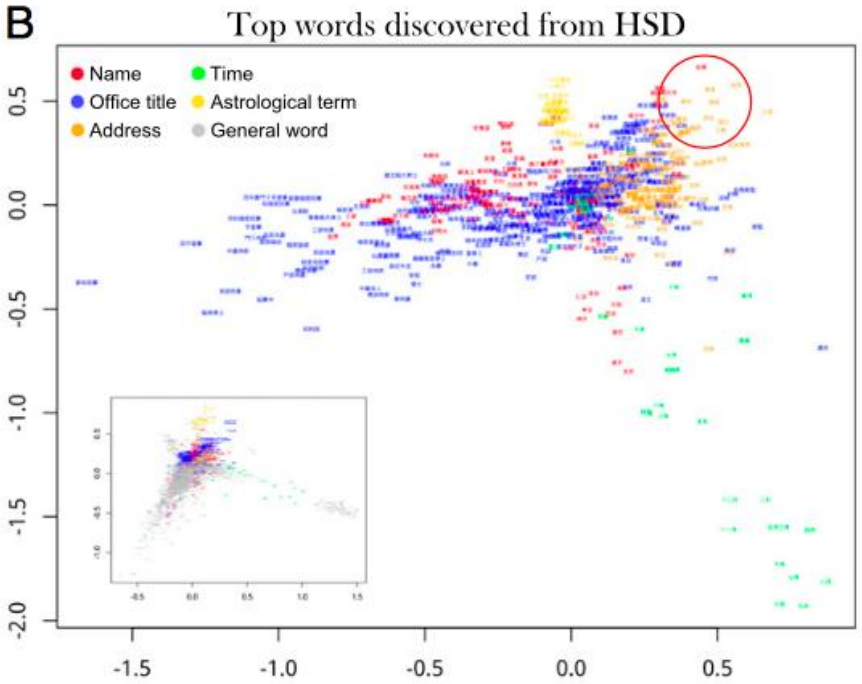
Source: Deng, K., Bol, P. K., Li, K. J., & Liu, J. S. (2016). On the unsupervised analysis of domain-specific Chinese texts. *Proceedings of the National Academy of Sciences*, *113*(22), 6154-6159.

https://doi.org/10.1073/pnas.1516510113.

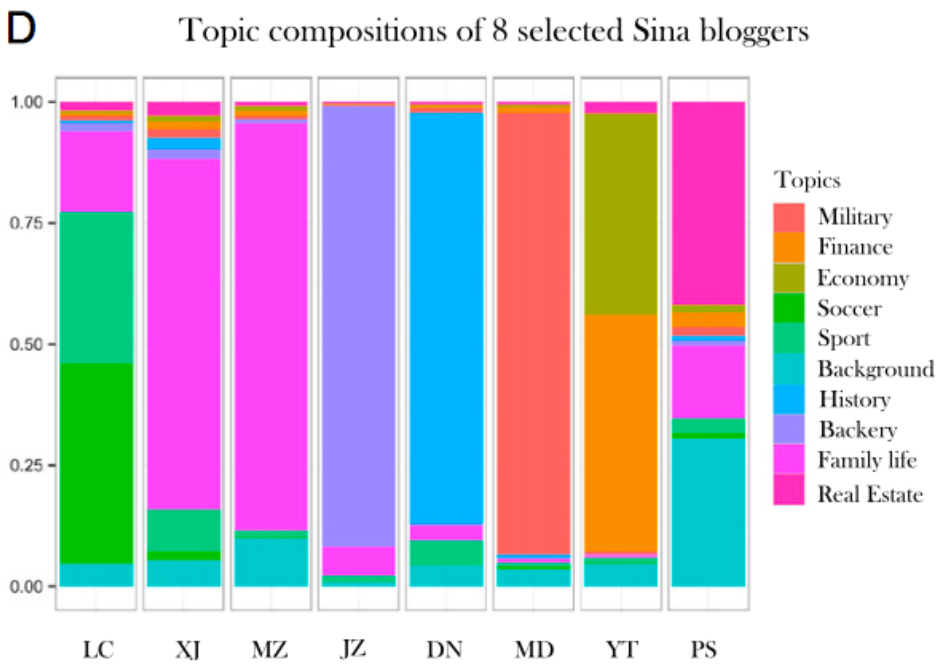
Build Lexicon

- TopWORDS and its analysis results (search: Deng Lab, Tsinghua)

- The word frequency of *History of the Song Dynasty*



- The top topics from Sina bloggers



Source: Deng, K., Bol, P. K., Li, K. J., & Liu, J. S. (2016). On the unsupervised analysis of domain-specific Chinese texts. *Proceedings of the National Academy of Sciences*, *113*(22), 6154-6159.

https://doi.org/10.1073/pnas.1516510113.

What could we get from statistical learning?

- A lexicon or ranked list of words

  To help us extract the subjects of metadata

- Word frequency

  To help us catch keywords and discard non-critical information

- Text probable preferences and topics

  To help us figure out the academic areas

Assumption of Frameworks

- Integration
- Workflow
- Cooperation

Integration

Statistical learning

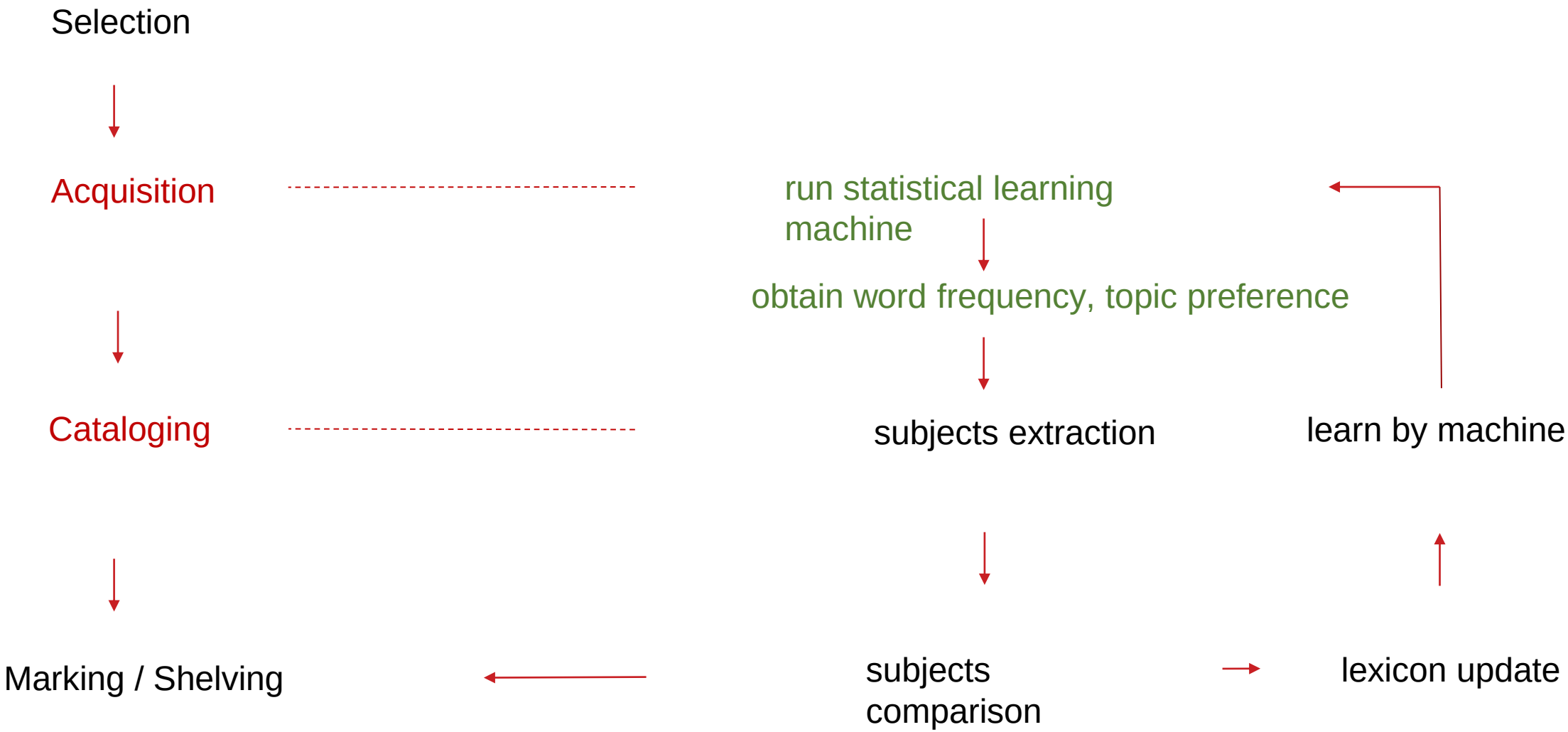
Library job

Word frequency, text preferences & topics

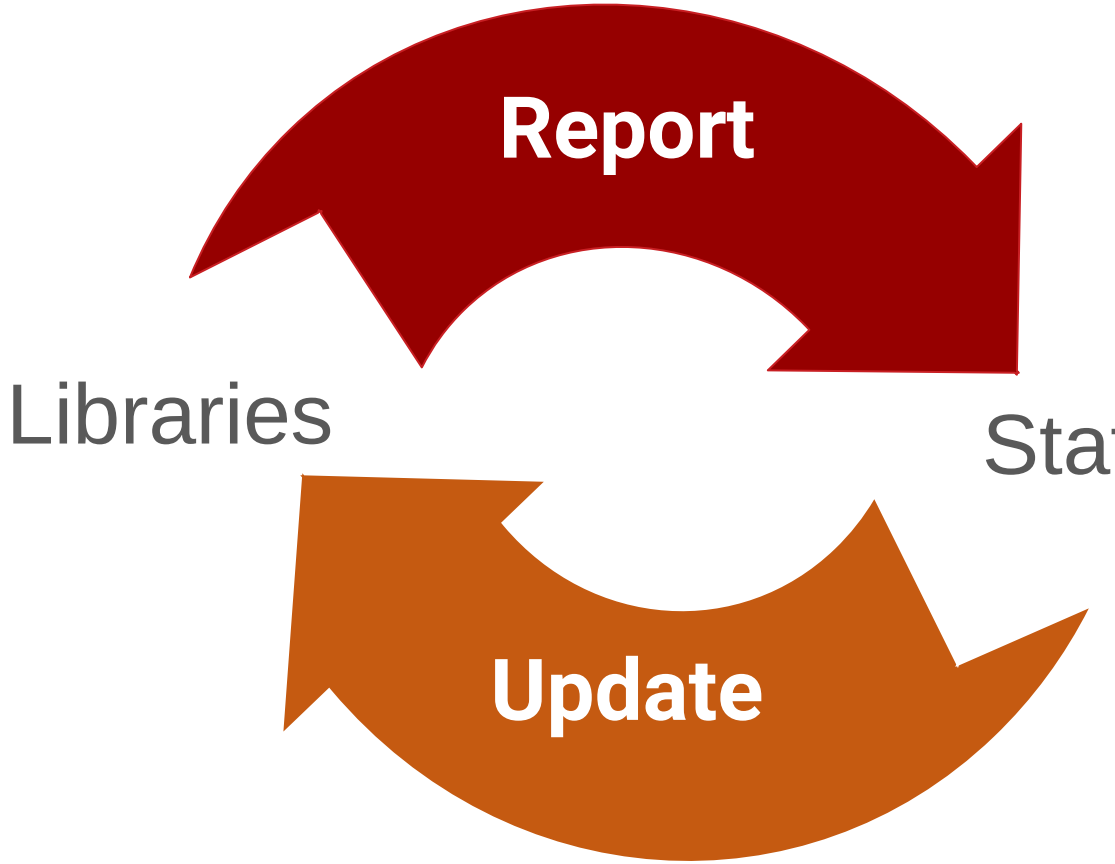
Source: Deng, K., Bol, P. K., Li, K. J., & Liu, J. S. (2016). On the unsupervised analysis of domain-specific Chinese texts. *Proceedings of the National Academy of Sciences*, *113*(22), 6154-6159.

https://doi.org/10.1073/pnas.1516510113.

Workflow


Selection
Acquisition
Cataloging
Marking / Shelving
run statistical learning machine
obtain word frequency, topic preference
subjects extraction
learn by machine
subjects comparison
lexicon update

Cooperation



Thank you!

anlin-yang@uiowa.edu