

MeTA: A Unifying Toolkit for the Management and Analysis of Text Data

Chase Geigle and ChengXiang (“Cheng”) Zhai

Department of Computer Science

University of Illinois at Urbana-Champaign

geigle1@illinois.edu; czhai@illinois.edu

<https://meta-toolkit.org/>

Text data cover all kinds of topics

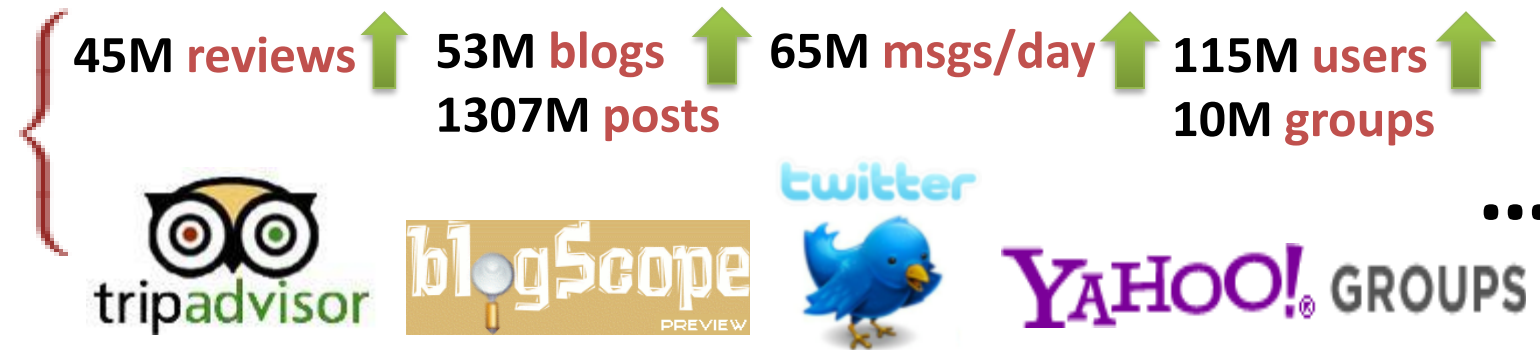
Topics:

People
Events
Products
Services, ...

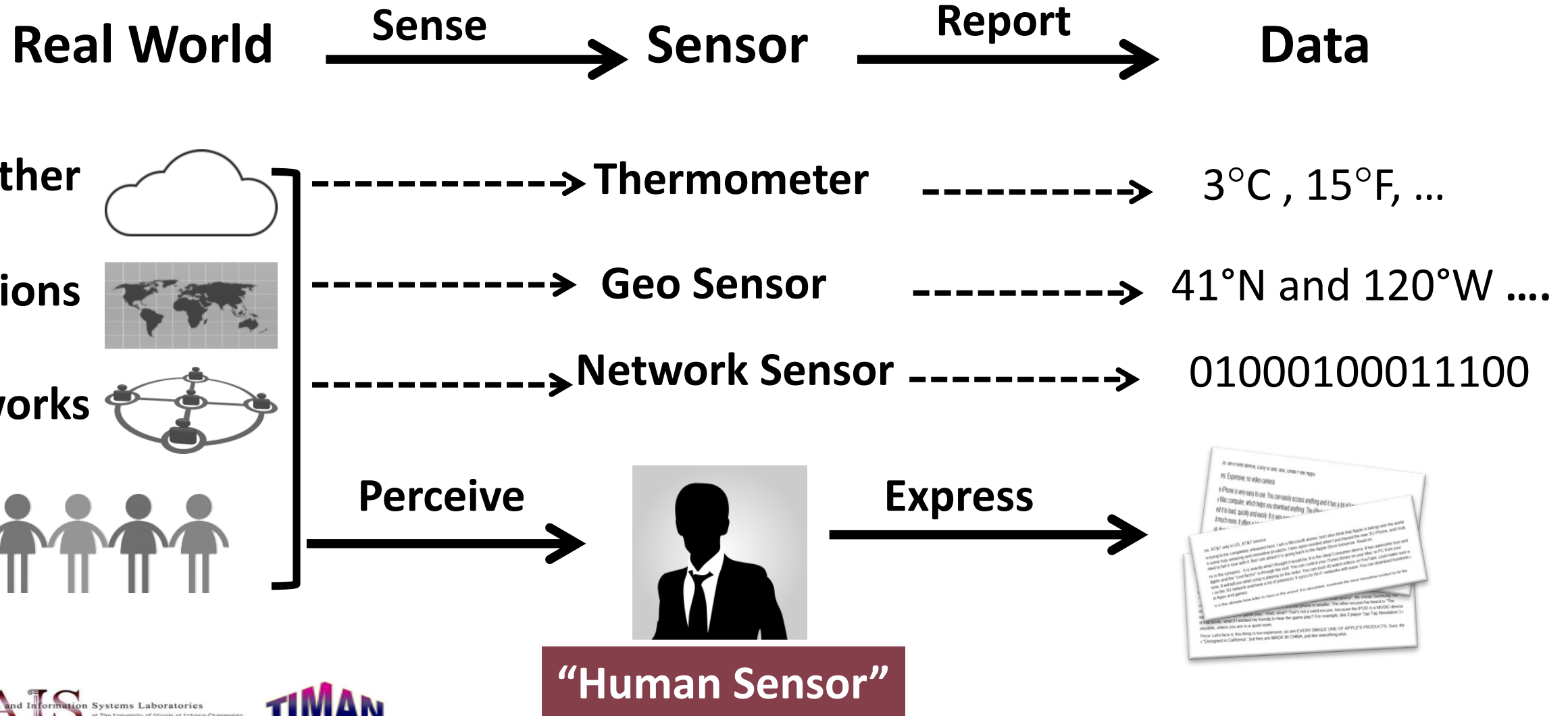


Sources:

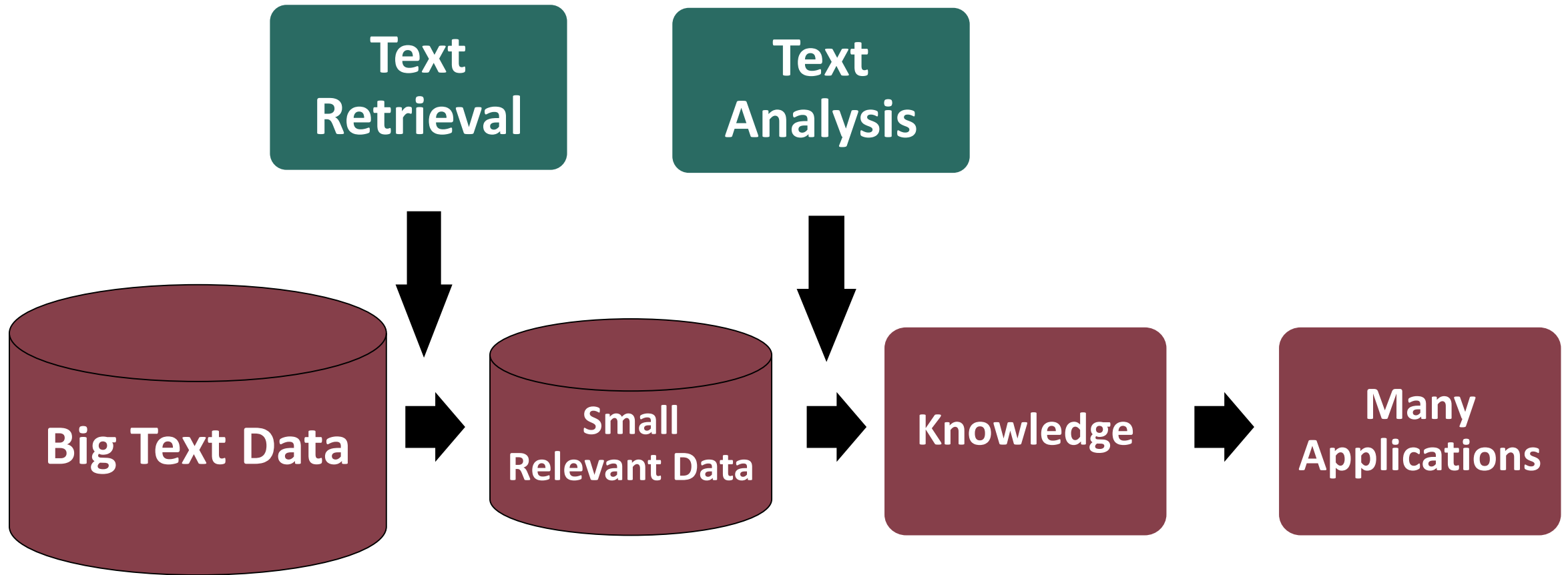
Blogs
Microblogs
Forums
Reviews ,...



Humans as Subjective & Intelligent “Sensors”



Main Techniques for Harnessing Big Text Data: Text Retrieval + Text Analysis

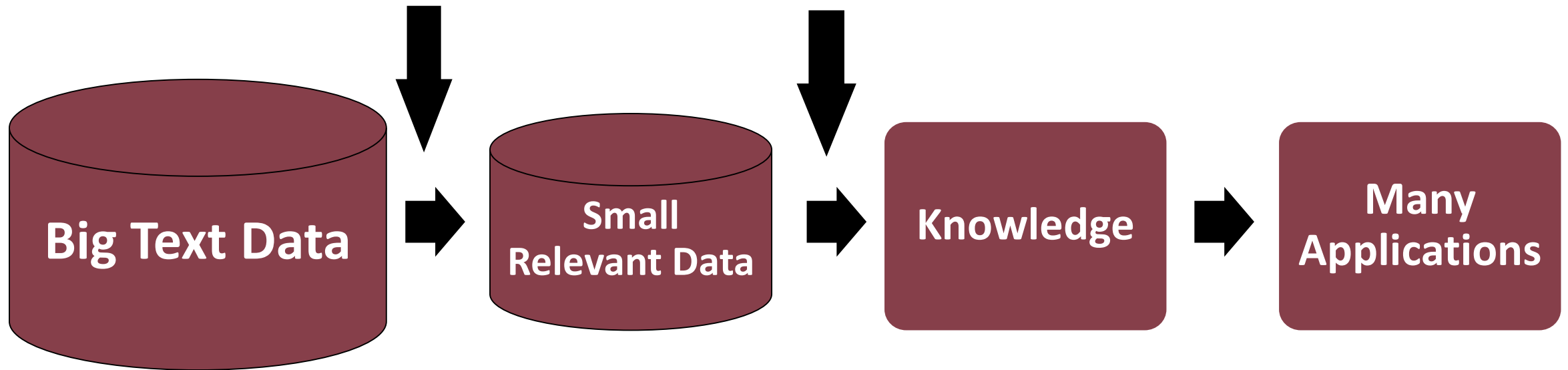


This tutorial: MeTA Toolkit

<https://meta-toolkit.org/>

Text
Retrieval

Text
Analysis



Overview of MeTA

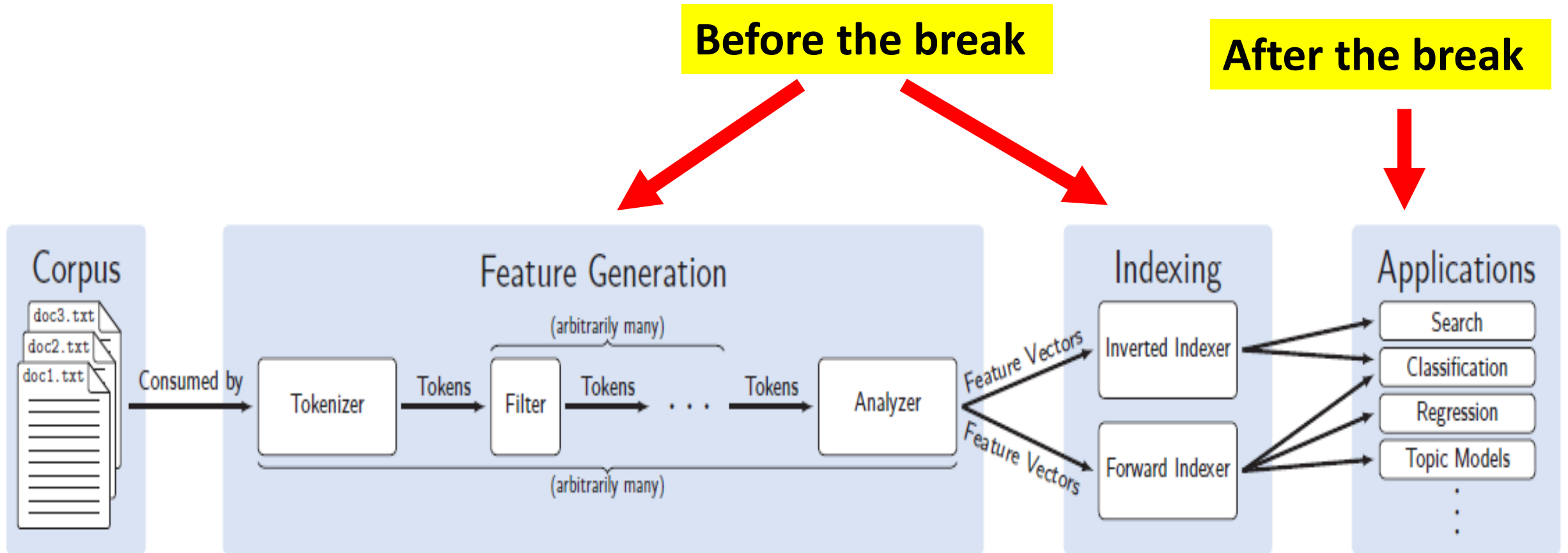
- **Founders & Key contributors:** Chase Geigle, Sean Massung
- **Advisor:** ChengXiang Zhai
- **Design philosophy**
 - Unified toolkit for full support of text data retrieval and analysis
 - Facilitate education, research, and application development (open source)
 - Highly efficient and extensible (C++ with Python binding)
 - Continuously updated to reflect research progress (hope you can help!)
- **Current uses**
 - Mostly education (MOOCs on Coursera, courses at UIUC)

MeTA vs. Related Toolkits

| | Indri <i>IR</i> | Lucene <i>IR</i> | MALLET <i>ML/NLP</i> | LIBLINEAR <i>ML</i> | SVM ^{MULT} <i>ML</i> | scikit <i>ML/NLP</i> | CoreNLP <i>ML/NLP</i> | MeTA <i>all</i> |
|--------------------|--------------------|---------------------|-------------------------|------------------------|----------------------------------|-------------------------|--------------------------|--------------------|
| Feature generation | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ |
| Search | ✓ | ✓ | | | | | | ✓ |
| Classification | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Regression | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| POS tagging | | | ✓ | | | | ✓ | ✓ |
| Parsing | | | | | | | ✓ | ✓ |
| Topic models | | | ✓ | | | ✓ | | ✓ |
| <i>n</i> -gram LM | | | | | | | | ✓ |
| Word embeddings | | | ✓ | | | | ✓ | ✓ |
| Graph algorithms | | | | | | | | ✓ |
| Multithreading | | ✓ | ✓ | | | ✓ | ✓ | ✓ |

Massung, Sean, Chase Geigle, and ChengXiang Zhai. "Meta: A unified toolkit for text retrieval and analysis." *ACL 2016* (2016): 91.

MeTA Pipeline



Anticipated Schedule

- 8:30am-8:40am: Introduction and overview
- 8:40am-8:50am: MeTA setup
- 8:50am-9:20am: Raw text processing
- 9:20am-9:30am: Indexing
- **9:30am-10:00am: BREAK**
- 10:00am-10:45am: Information retrieval (Search engine competition)
- 10:45am-11:30am: Text classification (competition if time permitting)
- 11:30am-11:55am: Clustering
- 11:55am-12:00noon: Wrapup and feedback