Next Generation Digital Libraries:

Mining Our Own Scholarly Documents

KAN Min-Yen

Web IR / NLP Group (WING)

NUS National University of Singapore | School of Computing

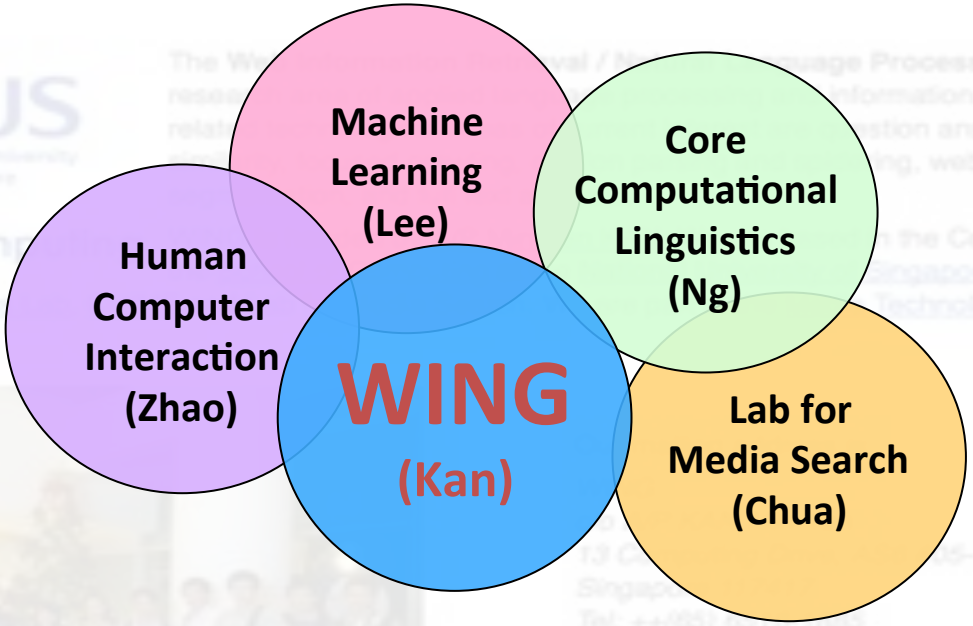
Research Topics in WING

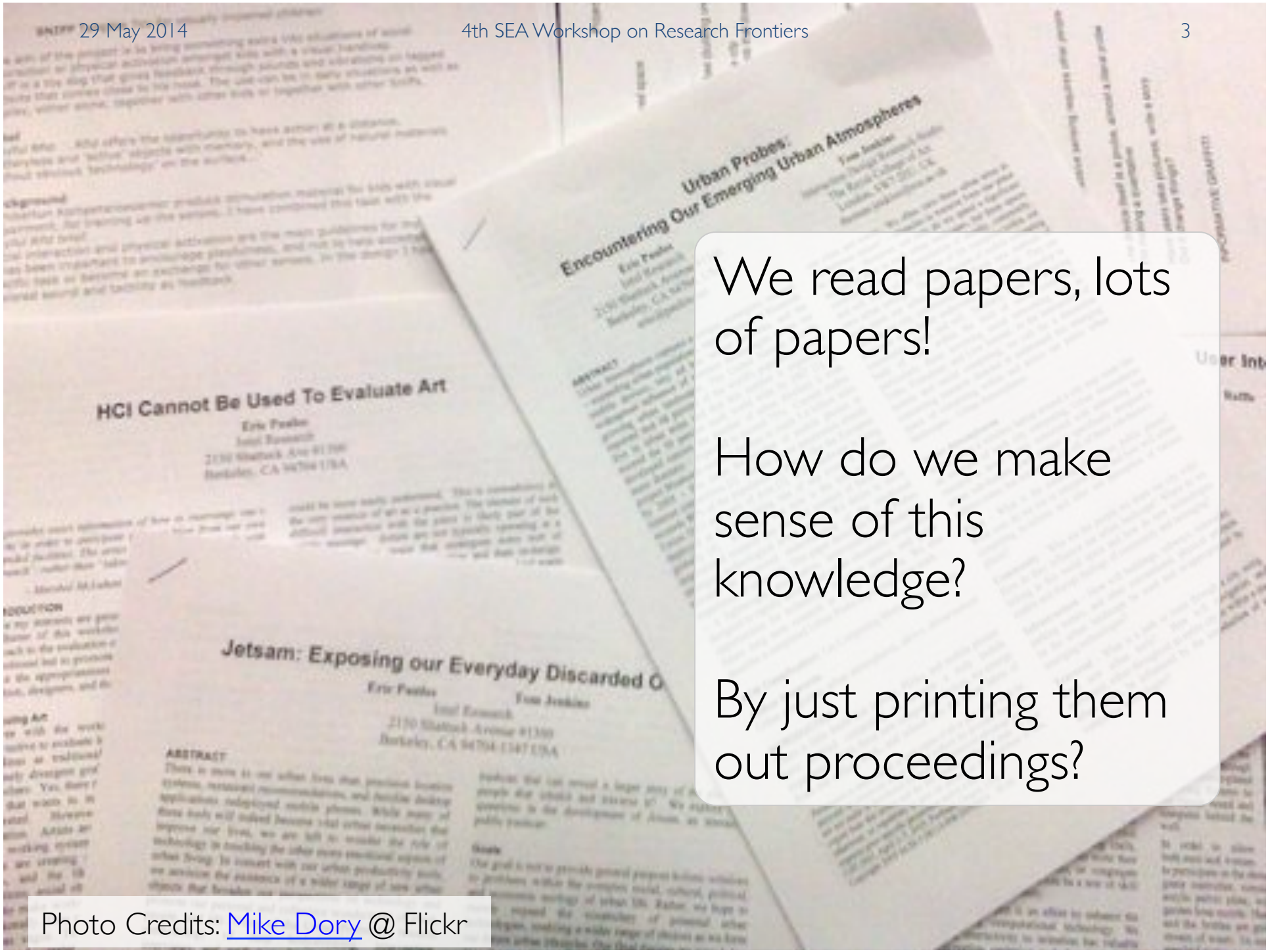
Digital Libraries << Focus Today

Natural Language Processing
- Machine Translation
- Information Extraction
- Automatic Summarization

Information Retrieval
- Understanding Users
- Multimedia Alignment

Human Computer Interaction
- Design of Browsing / Searching User Interfaces





We read papers, lots of papers!

How do we make sense of this knowledge?

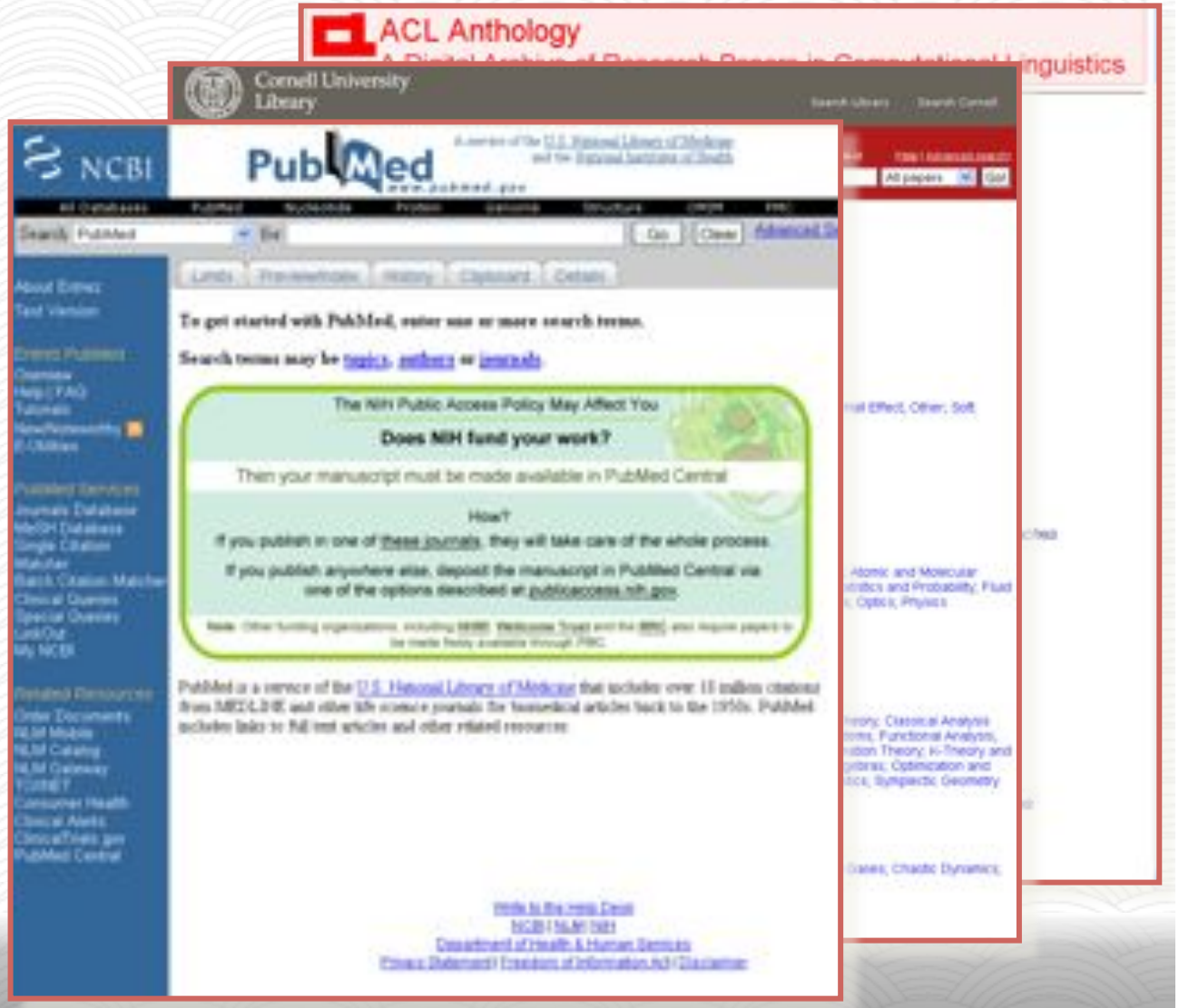
By just printing them out proceedings?

Photo Credits: Mike Dory @ Flickr

Scholarly digital libraries, circa 2014

Characteristics:

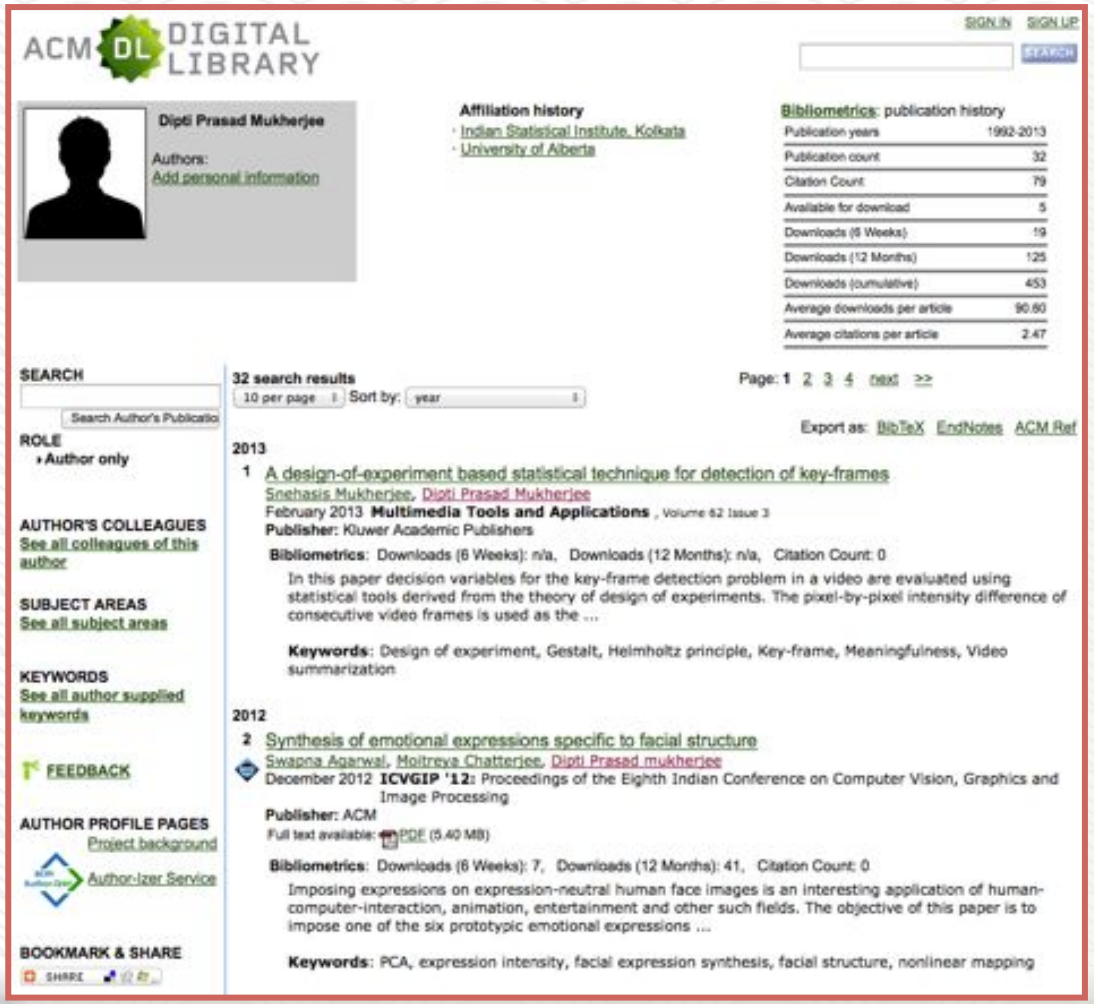
- PDFs and metadata of papers
- Searchable



Case in point: ACM Portal

Characteristics:

- Also has author statistics, co-authorship list and citation records
- Emphasize impact of author, vanity search

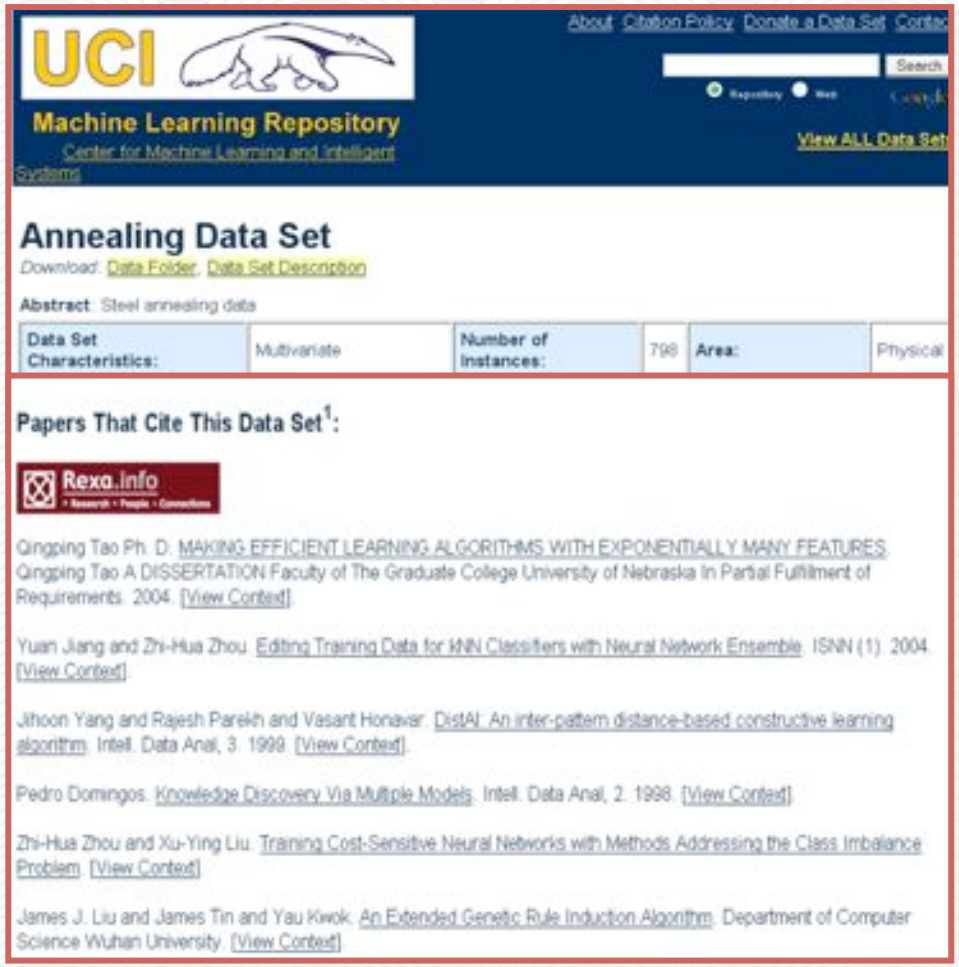


Rexa

- From the machine learning community
- Papers, authors and grants as first-order objects
- Requires login

Rexa's Service

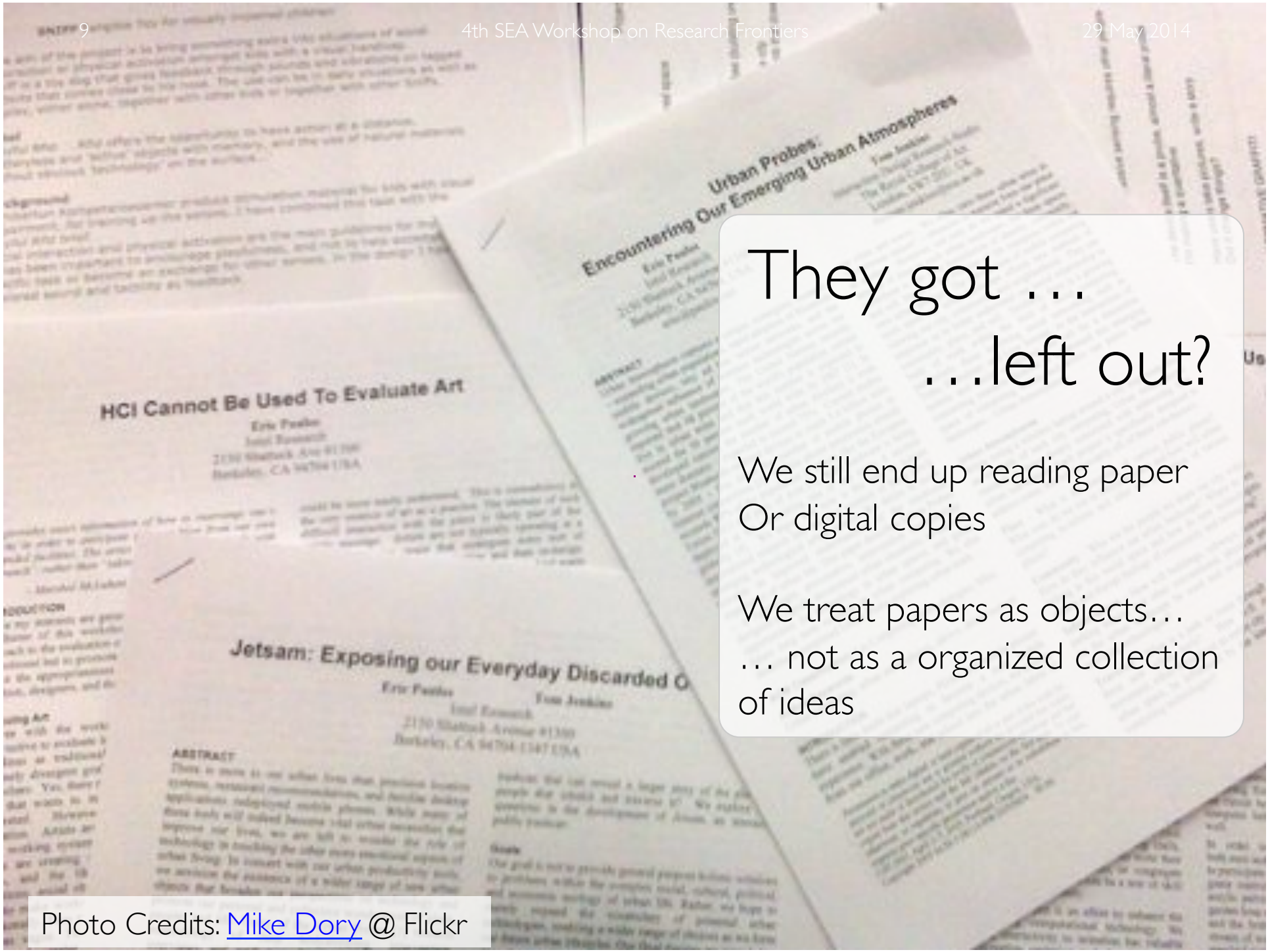
- Embed a layer serviced by Rexa
- Like Trackback for blogs



Revisiting the characteristics

- Led by computer science
- Digital library as a monolithic server
- Automation and scalability
- Document access, not document use
  - Focus on rights management → OpenURL
  - Focus on identifying records → DOI

Centered on the authors!
What happened to the readers?



They got …
…left out?

We still end up reading paper
Or digital copies

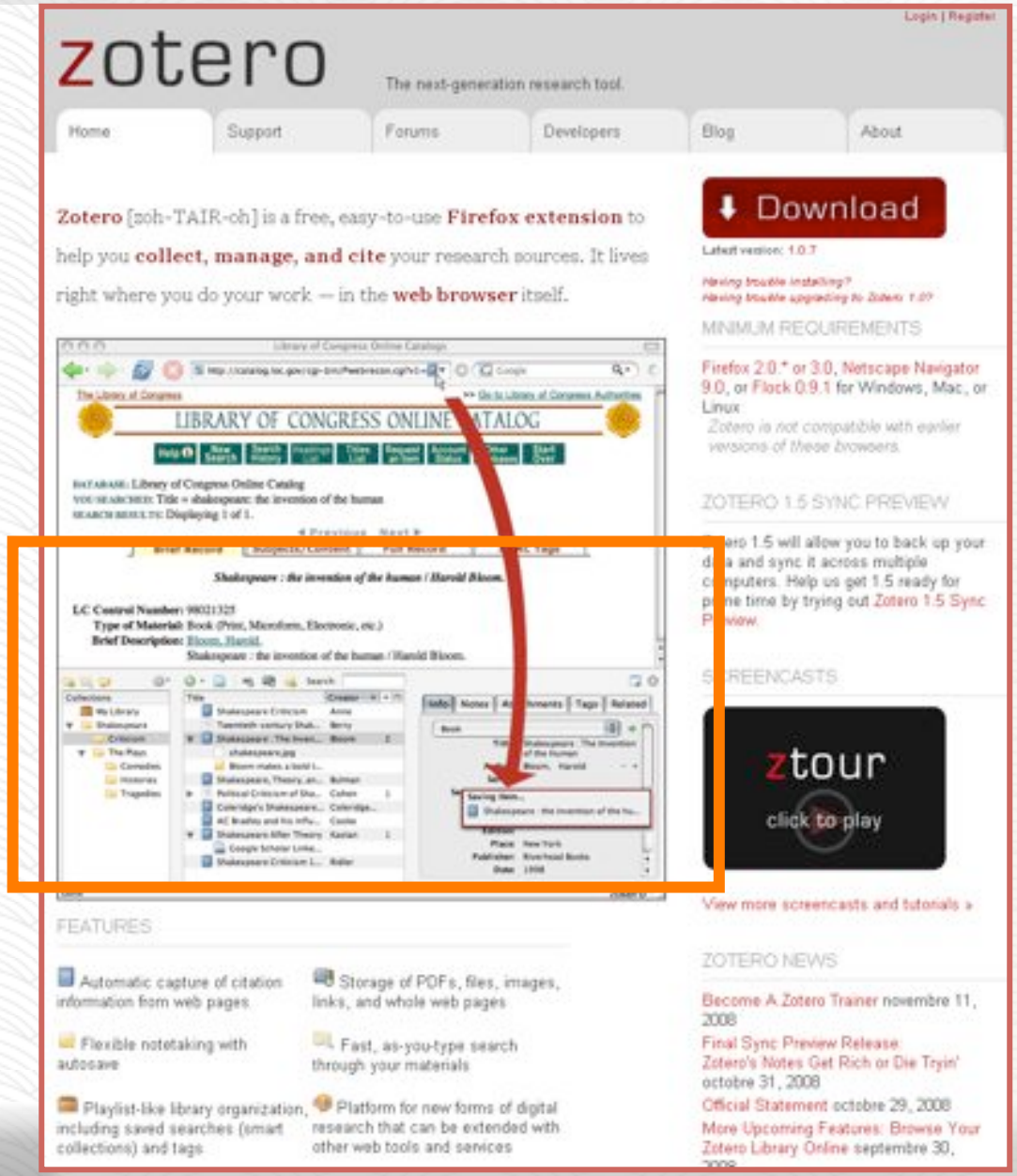
We treat papers as objects…
… not as a organized collection of ideas

Photo Credits: Mike Dory @ Flickr

CiteULike

- Web 2.0 style – powered by users
- Upload your own metadata
- Got so big, publishers allow ingestion directly

- Firefox management of reference data
- Scrapes info from specific webpages
- Loads manager in place
- EndNote, other systems also offering such functionality



Merging the two

- Librarians and end users have thought about the readers
  - Supporting re-use (Citation management)
  - Sharing metadata through group management
- Computer scientists, about the author
  - Prestige, finding authors and papers

But still not about using the document!

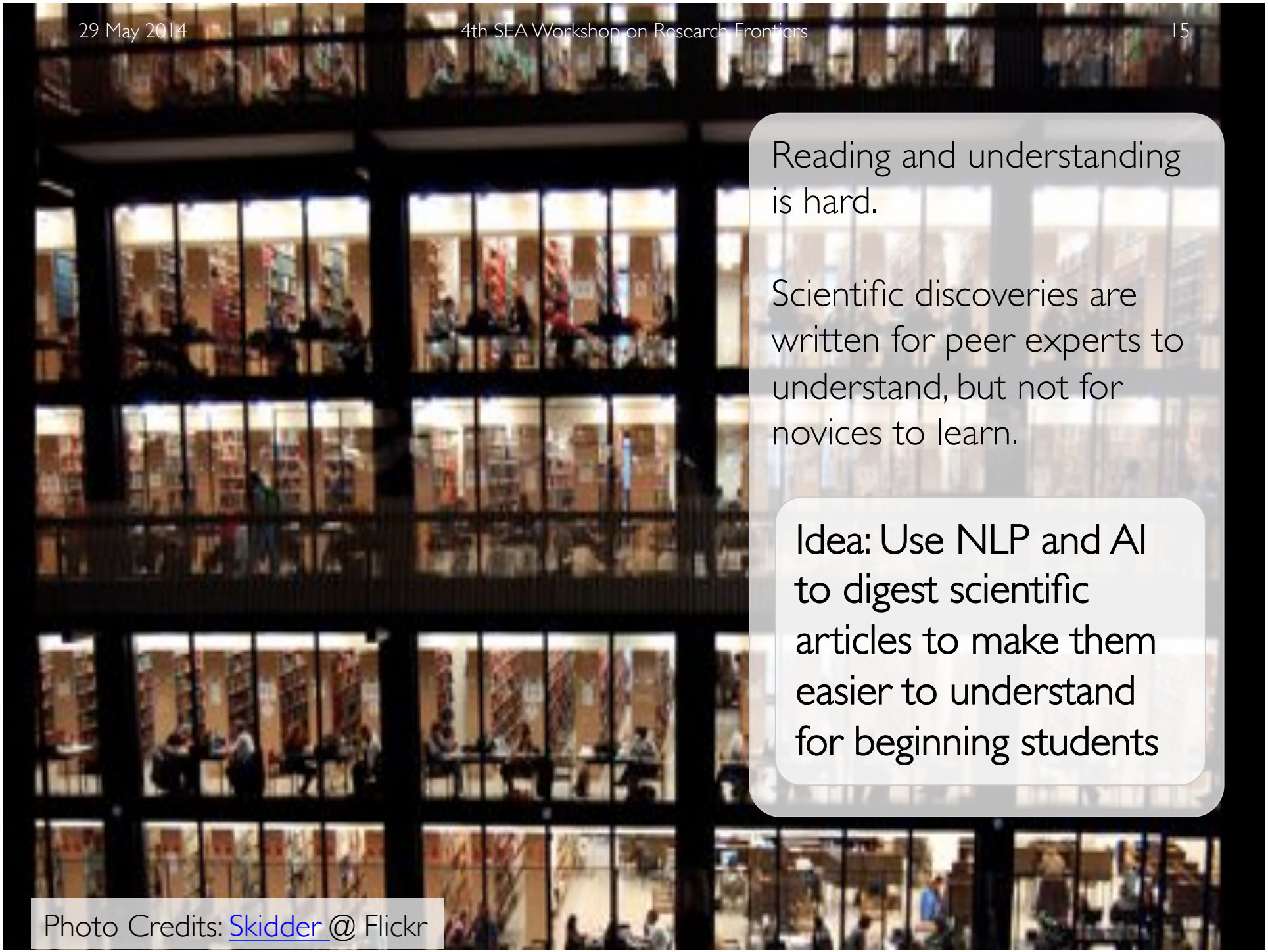
- Beginning researchers: undergraduate students
- Professors: managing students

Next Gen DL Elements

1. Summarizing Documents
2. Aligning Papers to their Presentations
3. Rediscovering Document Structure

Summarizing Scientific Documents

Photo Credits: Skidder @ Flickr



Reading and understanding is hard.

Scientific discoveries are written for peer experts to understand, but not for novices to learn.

Idea: Use NLP and AI to digest scientific articles to make them easier to understand for beginning students

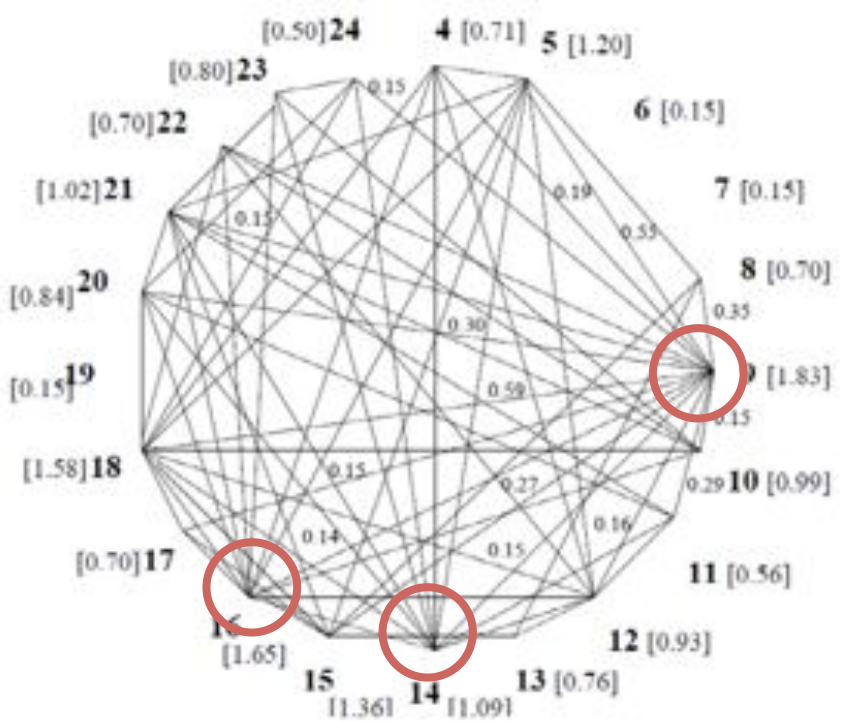
Photo Credits: Skidder @ Flickr

Document Summarization

- First try: To build a summary of a document
- To extract *k* sentences from a document of size *n*
- To rank the *n* sentences from a document with respect to some criterion

- Machine learned document summarization: Find features that help to select good sentences for a summary. Train from summaries.

Finding central sentences: TextRank



Build a graph of how much word "overlap" is there between sentences in a text.

Realize that the sentences that are more "central" are probably more important.

What about *scientific* documents?

Recognizing Human Action at a Distance in Video by Key Poses

Sreemoyee Mukherjee, Sujoy Kumar Biswas, *Member, IEEE*, and Dipti Prasad Mukherjee, *Senior Member, IEEE*

I. Introduction

A. Related works

Let's look at an example from Prof. Mukherjee

1. They have specific presentation conventions

I. Introduction

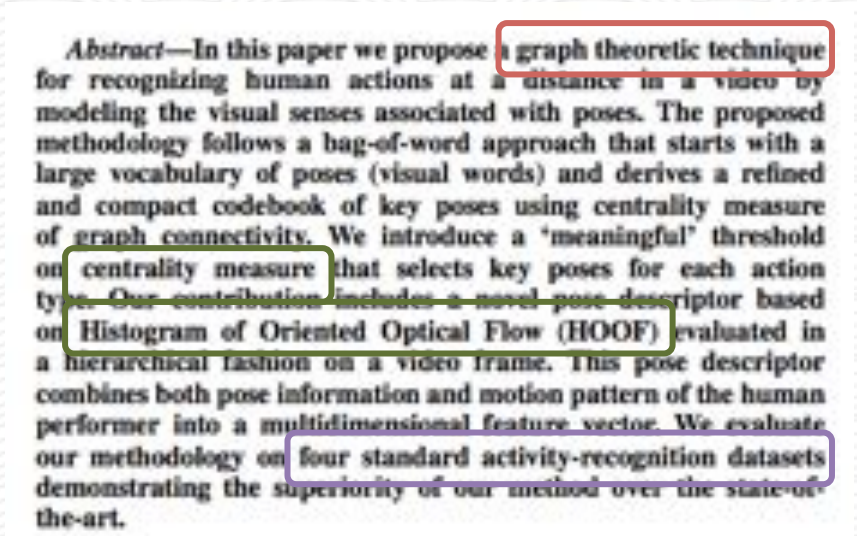
A Related Works

B Motivation and Overview

…

Abstracts!

Aim



Abstract—In this paper we propose a graph theoretic technique for recognizing human actions at a distance in a video by modeling the visual senses associated with poses. The proposed methodology follows a bag-of-word approach that starts with a large vocabulary of poses (visual words) and derives a refined and compact codebook of key poses using centrality measure of graph connectivity. We introduce a 'meaningful' threshold on centrality measure that selects key poses for each action type. Our contribution includes a novel pose descriptor based on Histogram of Oriented Optical Flow (HOOF) evaluated in a hierarchical fashion on a video frame. This pose descriptor combines both pose information and motion pattern of the human performer into a multidimensional feature vector. We evaluate our methodology on four standard activity-recognition datasets demonstrating the superiority of our method over the state-of-the-art.

Methods

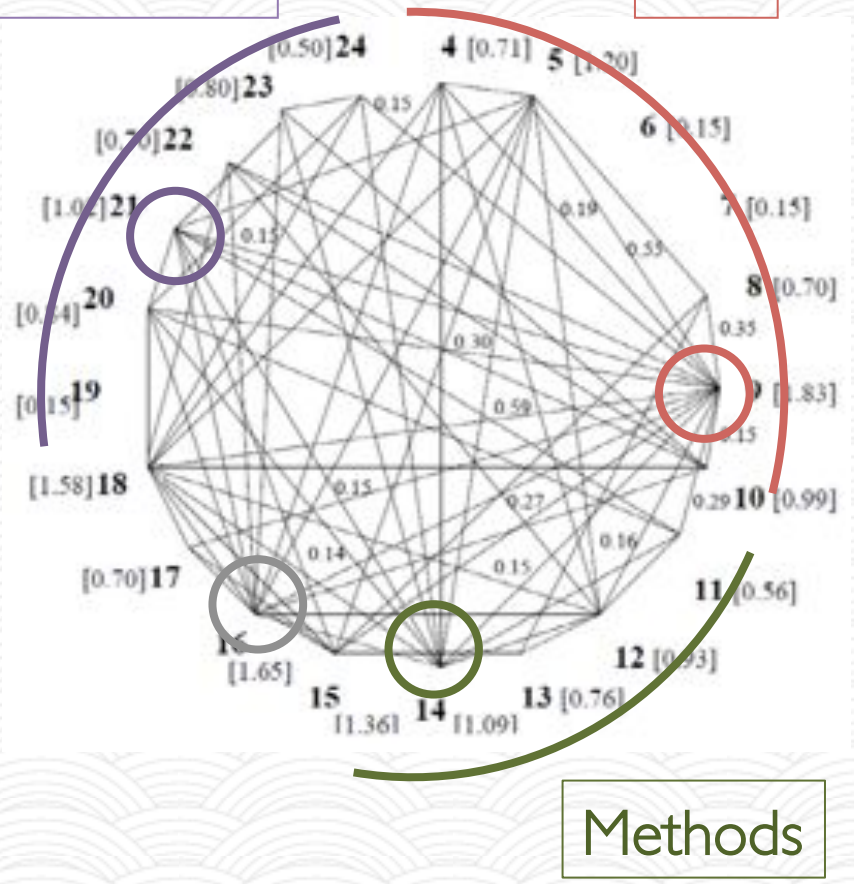
Evaluation

These are the "ground truth" for a summary of a single paper

Want to select sentences that overlap significantly with the abstract

But even here there is structure within an abstract

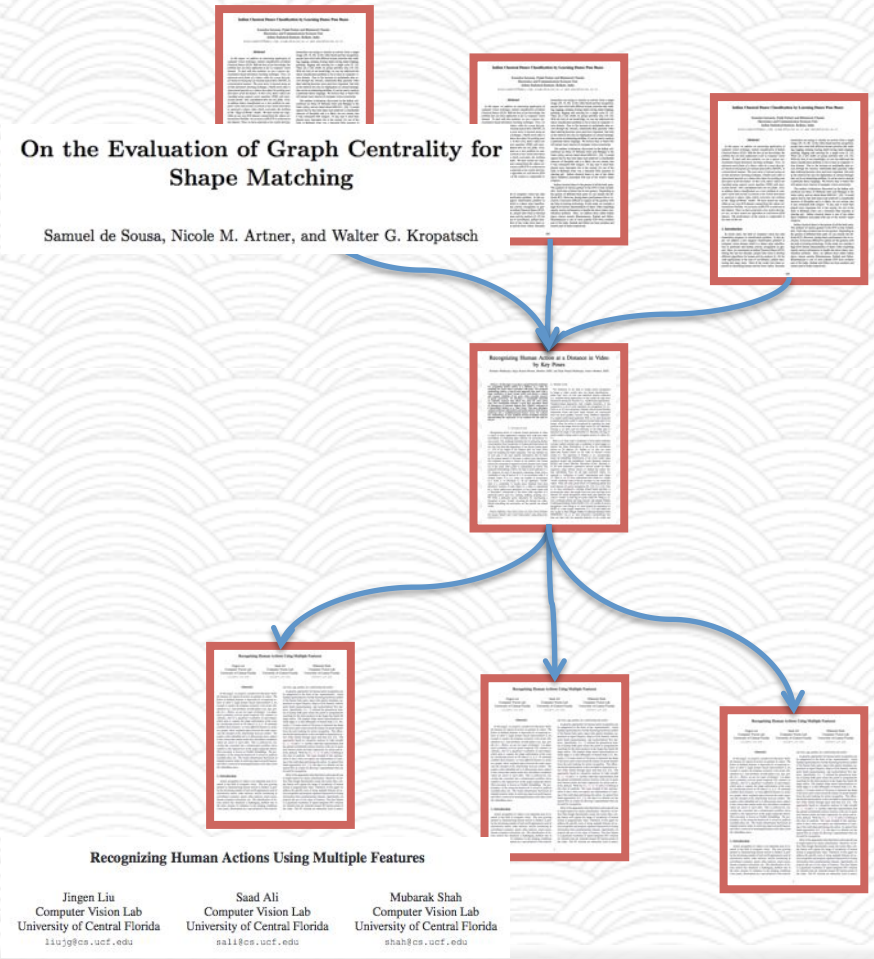
Revisiting TextRank



Capitalize on the conventional structure in documents.

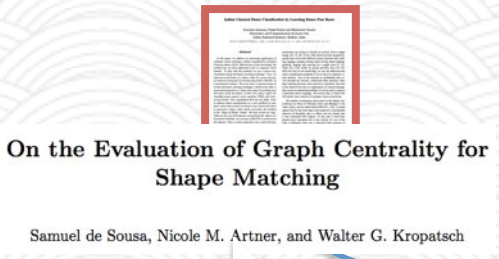
- Identify the logical structure of a scientific document
- Find the best sentences within the sections of the document

What about *scientific* documents?



- They have specific presentation conventions
- They have references and citations

Citation Sentences



On the Evaluation of Graph Centrality for Shape Matching

Samuel de Sousa, Nicole M. Artner, and Walter G. Kropatsch

- Often describe a paper from the community's point of view
- A representation of a key point of a work

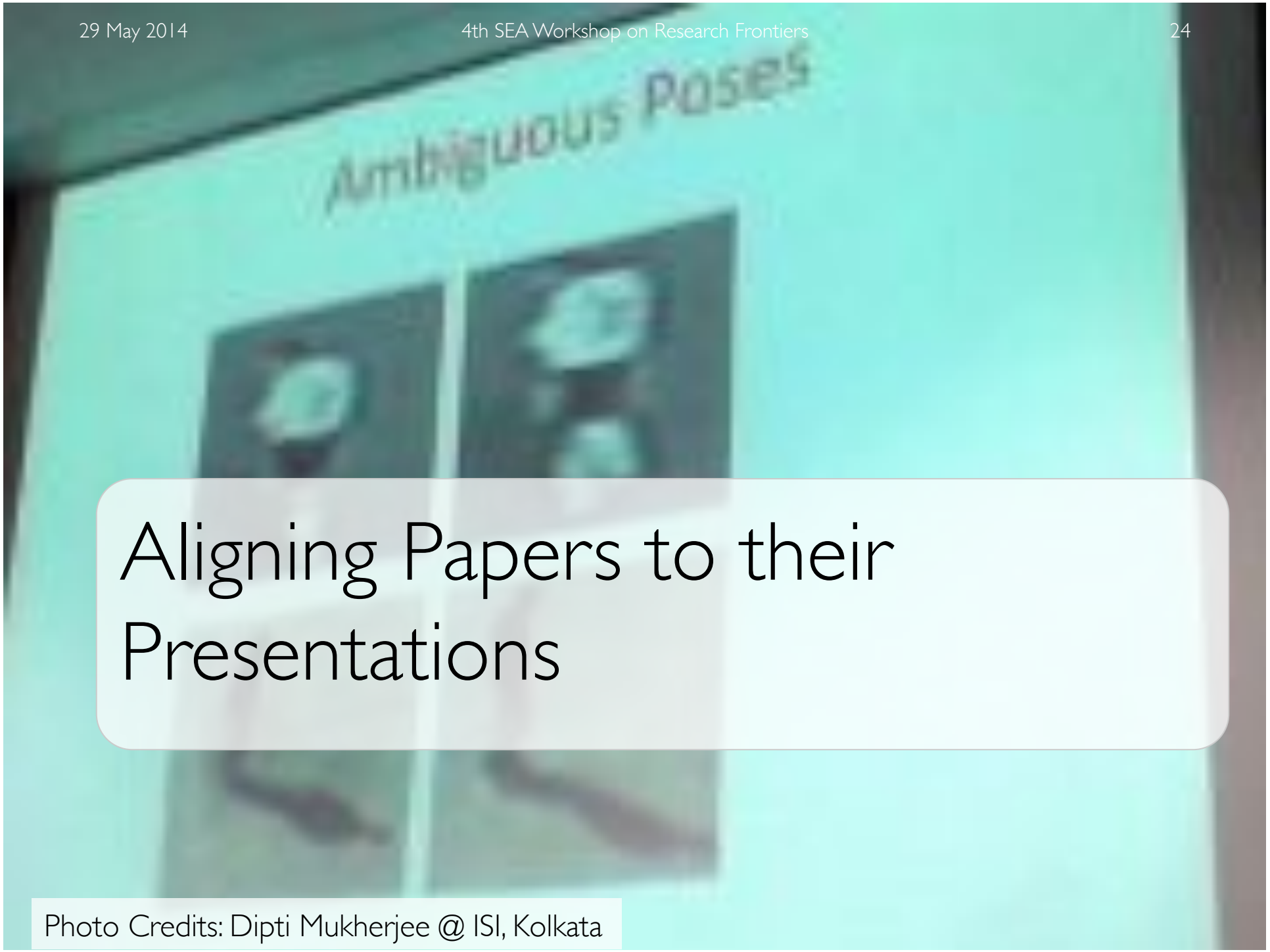
Mukherjee et al. [13] present an application of centrality in human action recognition. They employ centrality to create a compact codebook out of a large vocabulary of poses (bag-of-word approach). Cukierski et al. [7] use centrality

A star and his fans

Photo Credits: Bollywood Glamour

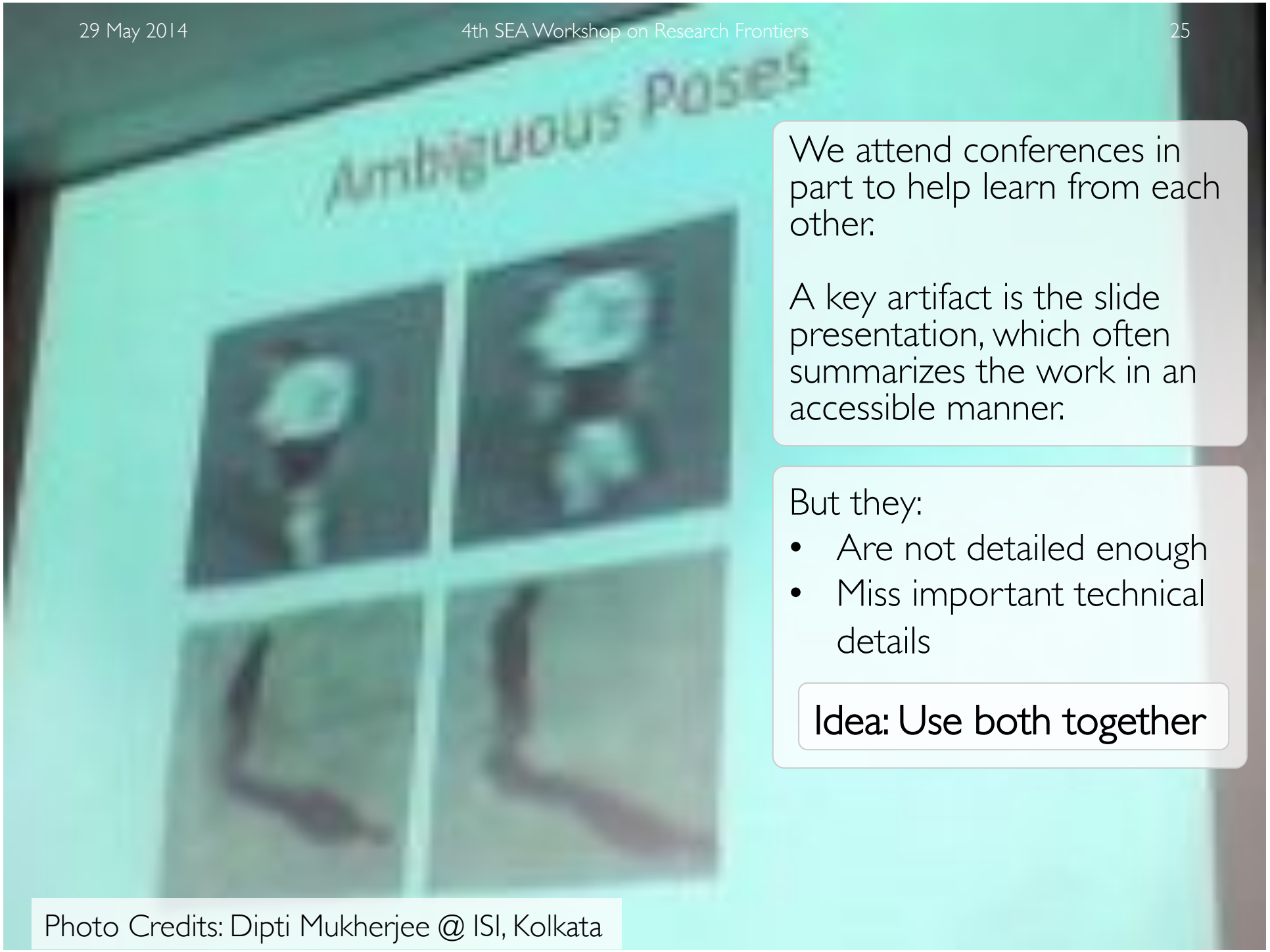
Citation sentences and in-article sentences have complementary purposes

- Results and evaluations usually not mentioned in citation sentences.
- Sentences in a paper that describe its method are usually too detailed for a summary.



Aligning Papers to their Presentations

Photo Credits: Dipti Mukherjee @ ISI, Kolkata



We attend conferences in part to help learn from each other.

A key artifact is the slide presentation, which often summarizes the work in an accessible manner.

But they:

- Are not detailed enough
- Miss important technical details

Idea: Use both together

Photo Credits: Dipti Mukherjee @ ISI, Kolkata

Better to juxtapose both media together in a fine-grained manner.

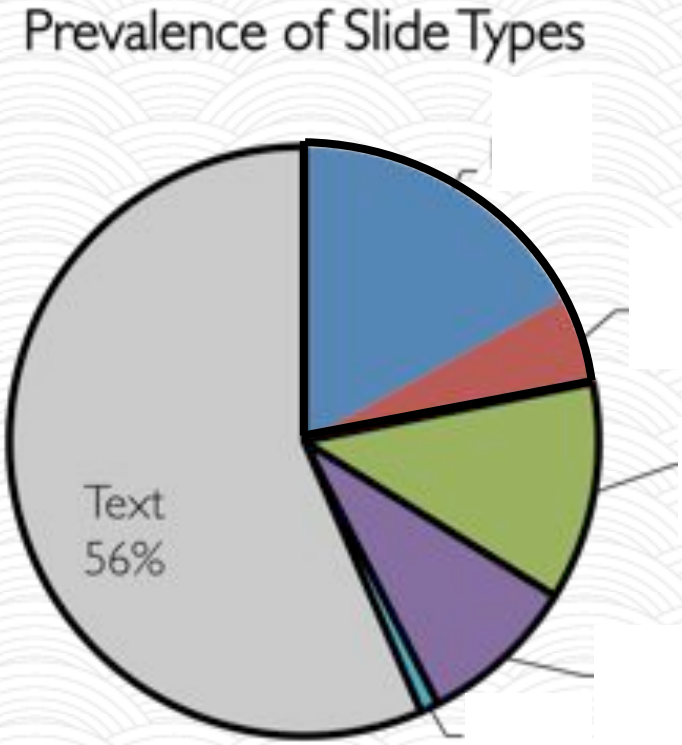
Output: an alignment map

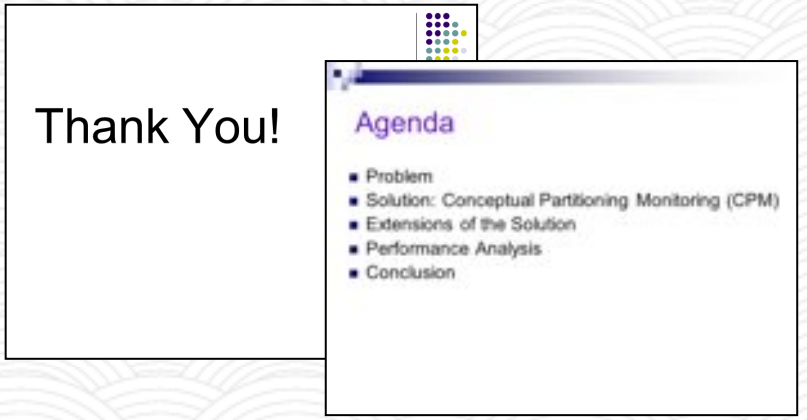
A prototype

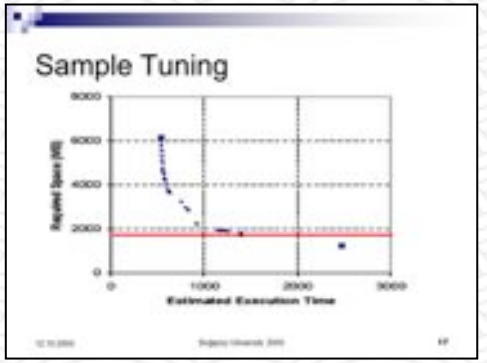
Document in focus

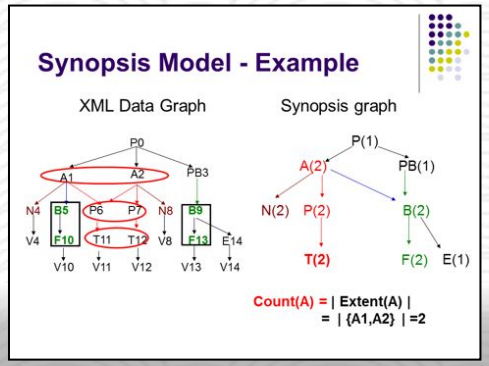
Slides in Focus

Demographics of an Existing Dataset









Combining Evidence

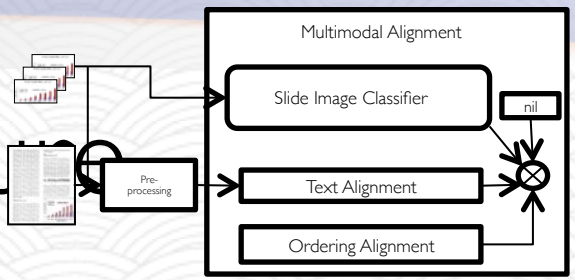
Represent each of the three sources as a probability distribution or preference

1. Text Similarity
2. Linear Ordering
3. Visual Content

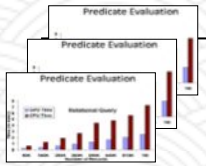
Handle obvious exceptions.

Weight distributions together to find most likely point as alignment.

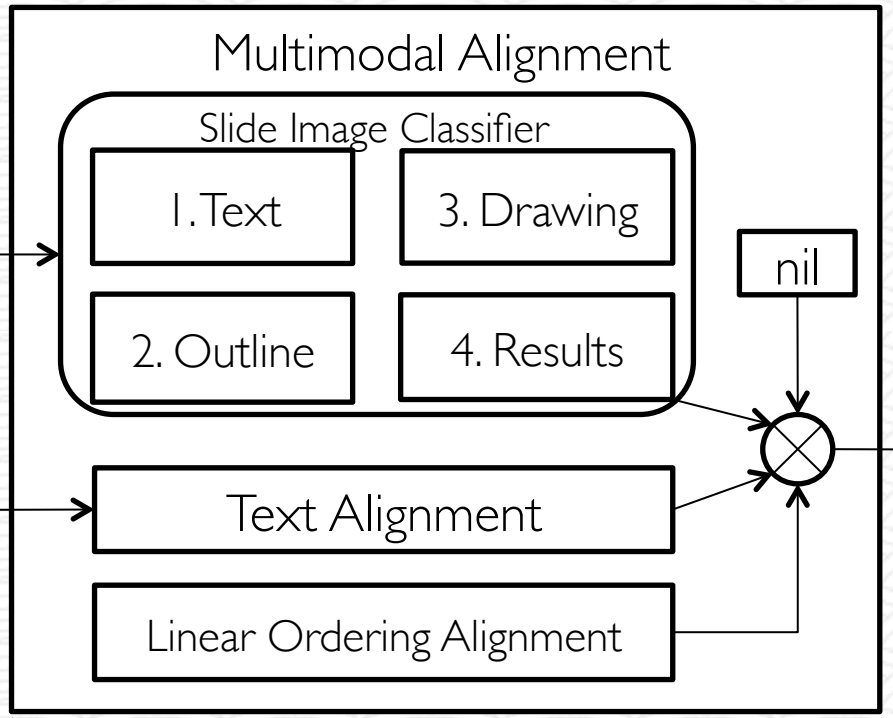
System Architecture

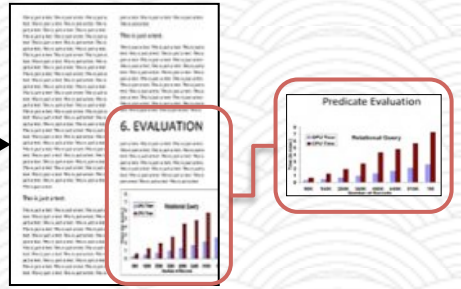


Input: Presentation



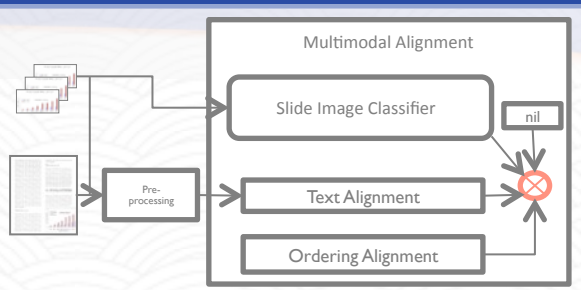
Input: Document





Output: Alignment map

Current architecture. Slightly different from published paper.



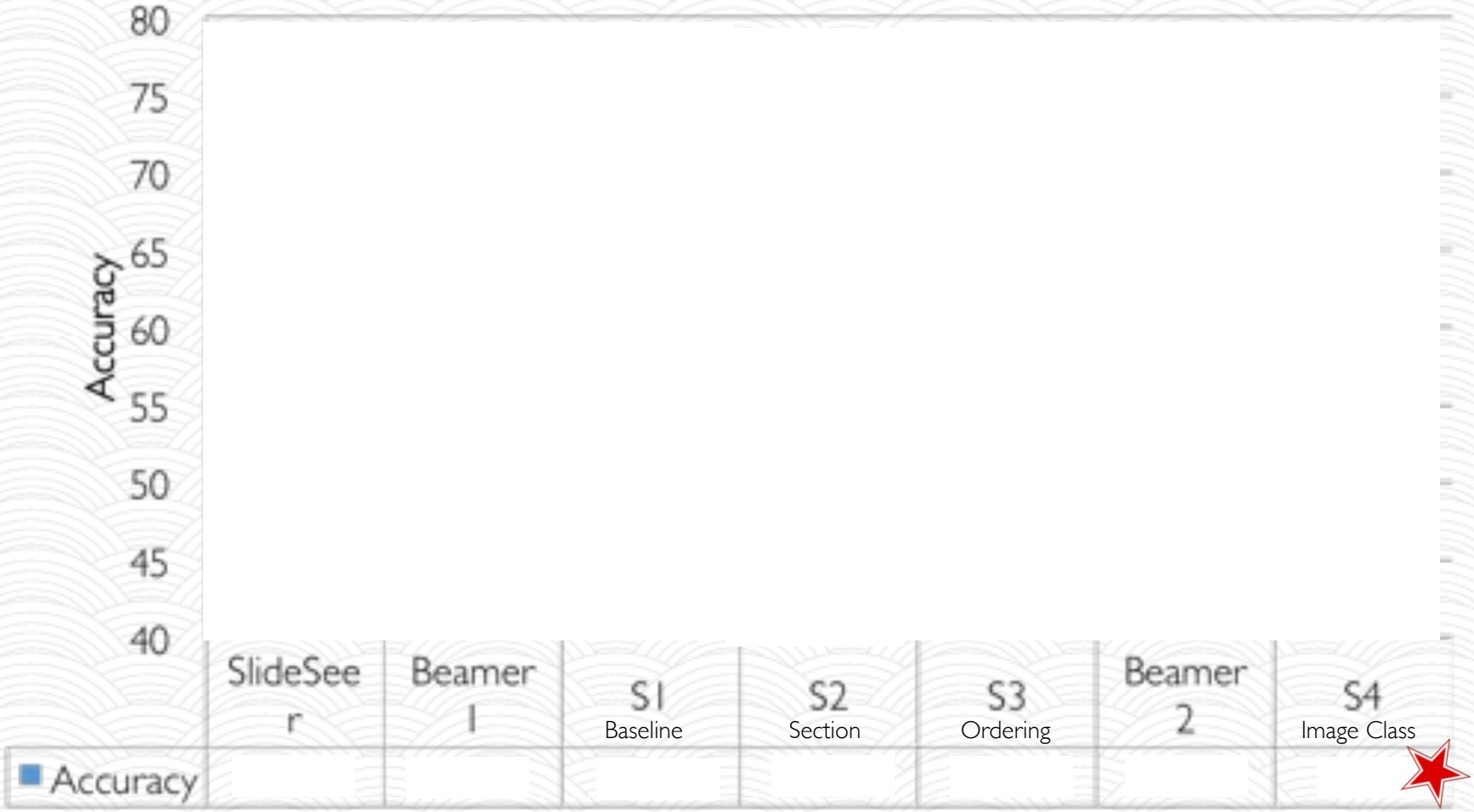
Multimodal Fusion

- Input for each slide:
  1. Text Alignment Vector → V_{Ts}
  2. Ordering Alignment Vector → V_{Os}
  3. Class assigned from image classifier

Uses Histogram of Oriented Gradients (HOG)

- Define 3 weights as: $W_{Ts} + W_{Os} + W_{nil} = 1.00$
- Tune weights according to image classes
- Apply *Nil* classifier
- Output for each slide: Final Alignment Vector → FAV_s

Results


80
75
70
65
60
55
50
45
40
Accuracy
SlideSee r
Beamer 1
S1 Baseline
S2 Section
S3 Ordering
Beamer 2
S4 Image Class
Accuracy

Rediscovering Document Structure

Photo Credits: Sound Doctrine Ministry

Publishers want authors to annotate the structure of the document in markup.

But history has shown time and time again that people don't always follow instructions.

We need tools that infer the logical document structure from finished documents

"Creating the cow from the hamburger" – Steve Pfeiffer

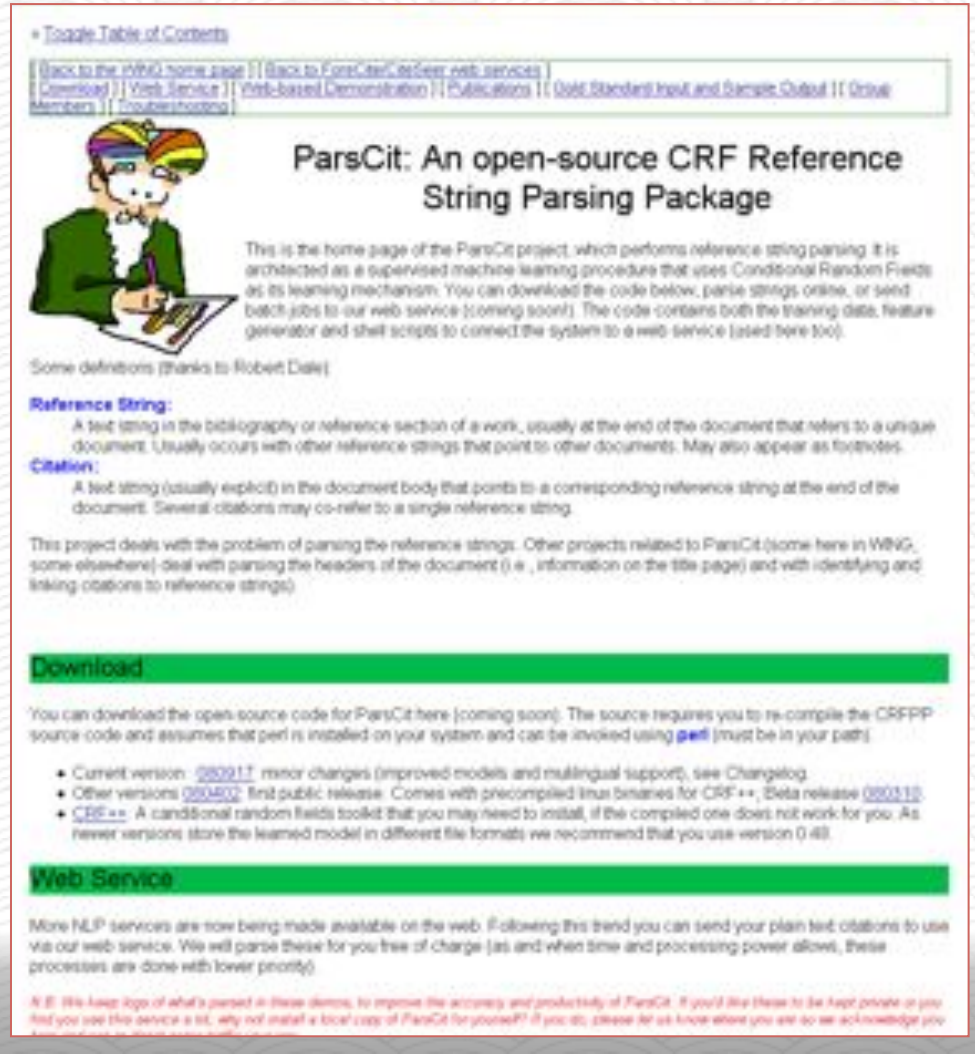
Photo Credits: Sound Doctrine Ministry

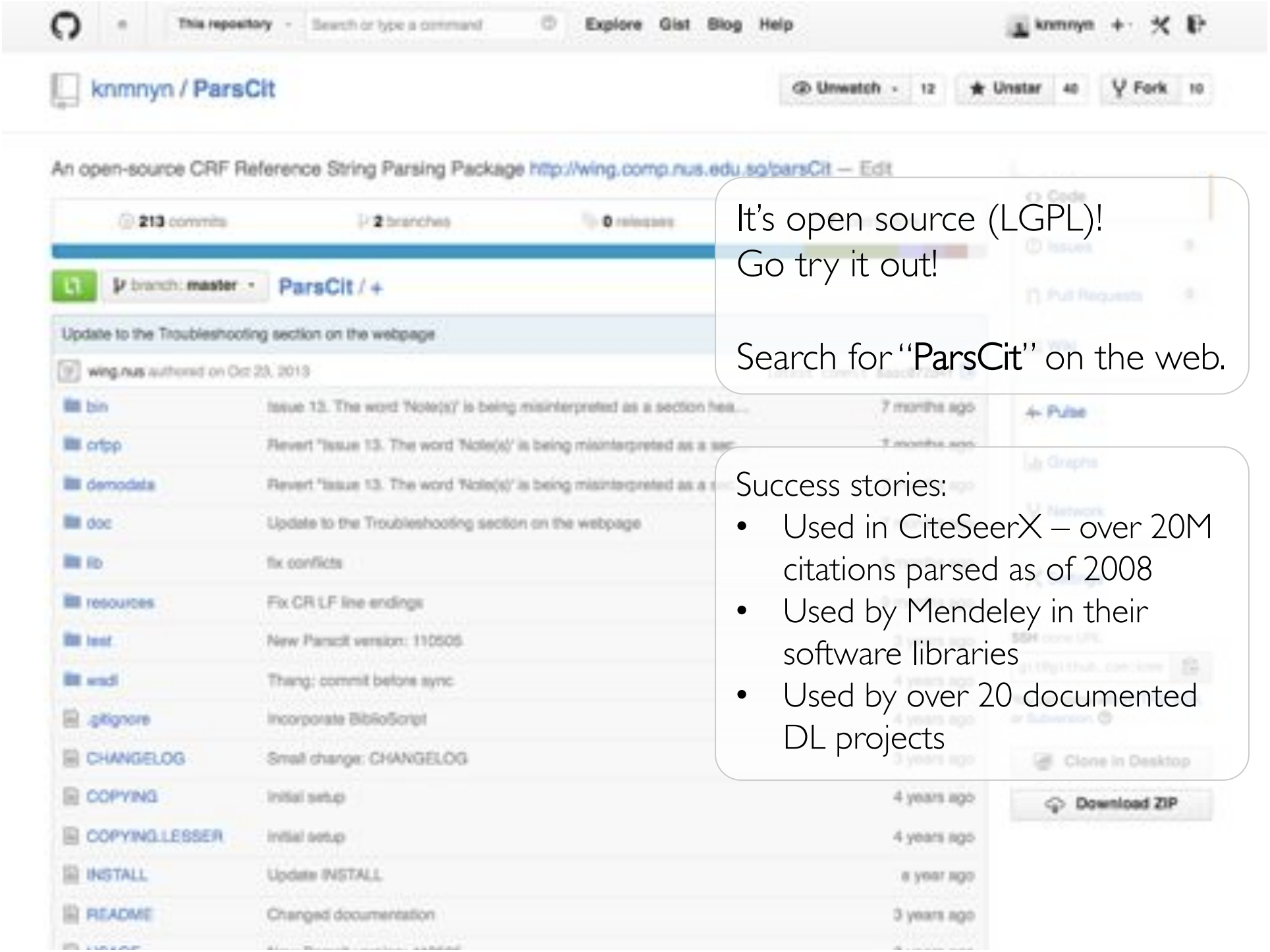
ParsCit

putting papers back together since 2008

- Reference String Parsing and Logical Document Structure Discovery Tool

- Other subsystems built on top:
  - Technical terminology finder,
  - Author and affiliation matcher, and
  - Document summarization.





It's open source (LGPL)!
Go try it out!

Search for "**ParsCit**" on the web.

Success stories:

- Used in CiteSeerX – over 20M citations parsed as of 2008
- Used by Mendeley in their software libraries
- Used by over 20 documented DL projects

Conclusions

- Building open-source tools for authors and readers

- Studying users as they interact with the DL
  - **Literature Review** – Single Document Summarization, Survey paper generation
  - **Downstream Data Mining** – Author affiliations, Technical terms, Zoning of the paper, Key metadata, **Scholarly document structure**

WING and visitors, Dec 2013

My question to all of you: What will the "publishing" model look like in 5-10 years?

THANK YOU