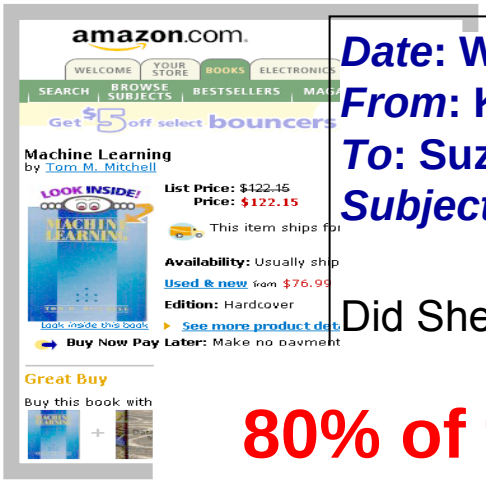
Optimizing Information Extraction over Evolving Text

Fei Chen
Database Group
Computer Science Department
University of Wisconsin – Madison

Lots of Text



Date: Wed Dec 20 08:57:00
From: Kay Mann <kay.mann
To: Suzanne Adams <suzan
Subject: Re: GE Conferenc

Did Sheila want Scott to parti

80% of the world's data, and growing!

⟨Operator⟩ Thank you
is John am
⟨Caller⟩ Yeah you ar
⟨Operator⟩ Morning ho
⟨Caller⟩ Oh I'm I a
um ah of co
ways. . . pro
one in and
I put in thi
what is the
you know. . .
doesn't mat
in and I hav
ing can you

MEDICAL RECORDS INFORMATION

PATIENT NAME: **ADKINS, PAUL J**		PATIENT NO: **000001**
ADDRESS: 1121 E MADISON EDWARDSVILLE, IL 62025		
HM TEL: 618-692-5545	WK TEL: 618-251-4784	SSN: 654-31-5818
OCCUPATION:	DOB: 01/16/1965(43 yrs)	CHARTNO: **PA1234**
GUARANTOR: ADKINS, PAUL J		SEX: MALE

HISTORY & PHYSICAL ON 03/03/2007

PROVIDER: MELMAN, IRVING G ENTERED BY: IGM

SUBJECTIVE

Mr. Adkins is a 41 yo male who noticed a recent bulge at his tailbone, prior to being seen he noted purulent drainage from this area. There has been no previous drainage from this area and no history of a known pilonidal abscess or cyst. He is not a diabetic. He was seen by his primary care MD and started on antibiotics. He states that the area has improved significantly since first developing.

Exploiting Text by Information Extraction (IE)

Disease News

Jan 31: Turkey confirmed an incident of foot&mouth disease.

Jan 30: H1N1 identified in California Turkey Flock.

IE program

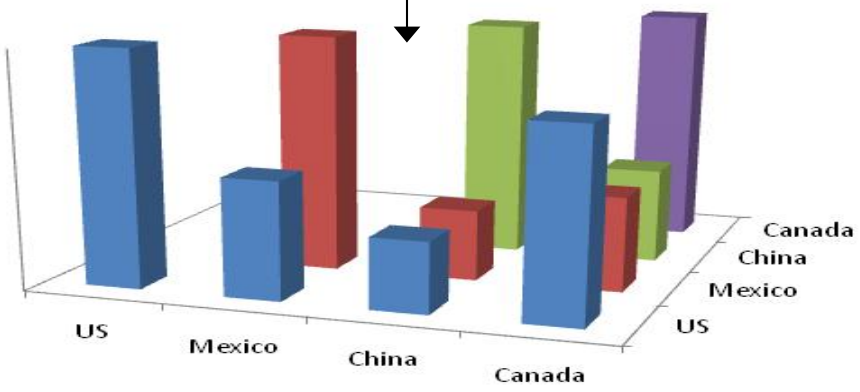
Incidents

disease	location
foot&mouth	Turkey
H1N1	California
…	…

```
SELECT  location
FROM    Incidents
WHERE   disease = 'H1N1'
```

California

Data Mining



Web Search

Advertising

Monitoring

Visualizing

Current State of Art

- **Many players**
  - AI/DB/DM/IR/NLP/Web communities: CMU, Columbia, Cornell, IIT, JHU, Max-Panck, PSU, Stanford, UCLA, UIUC, UMass, UMichigan, USC, UT Austin, UToronto, UWashington, UWisconsin…
  - industry: AT&T, Google, HP Labs, IBM, Microsoft, Yahoo!...
- **Mainly centered on improving extraction accuracy**
- **Recent work starts to consider improving runtime…**

Current Work on Improving Runtime

Pruning documents

[Agichtein *et al.* ICDE-03]
[Etzioni *et al.* WWW-04]
[Ipeirotis *et al.* SIGMOD-06]

Efficient pattern match

[Gravano *et al.* VLDB-01]
[Cho *et al.* ICDE-02]
[Chandel *et al.* ICDE-06]

Parallel processing

[Gruhl *et al.* IBMSJ-04]
[Lin SIGIR-09]

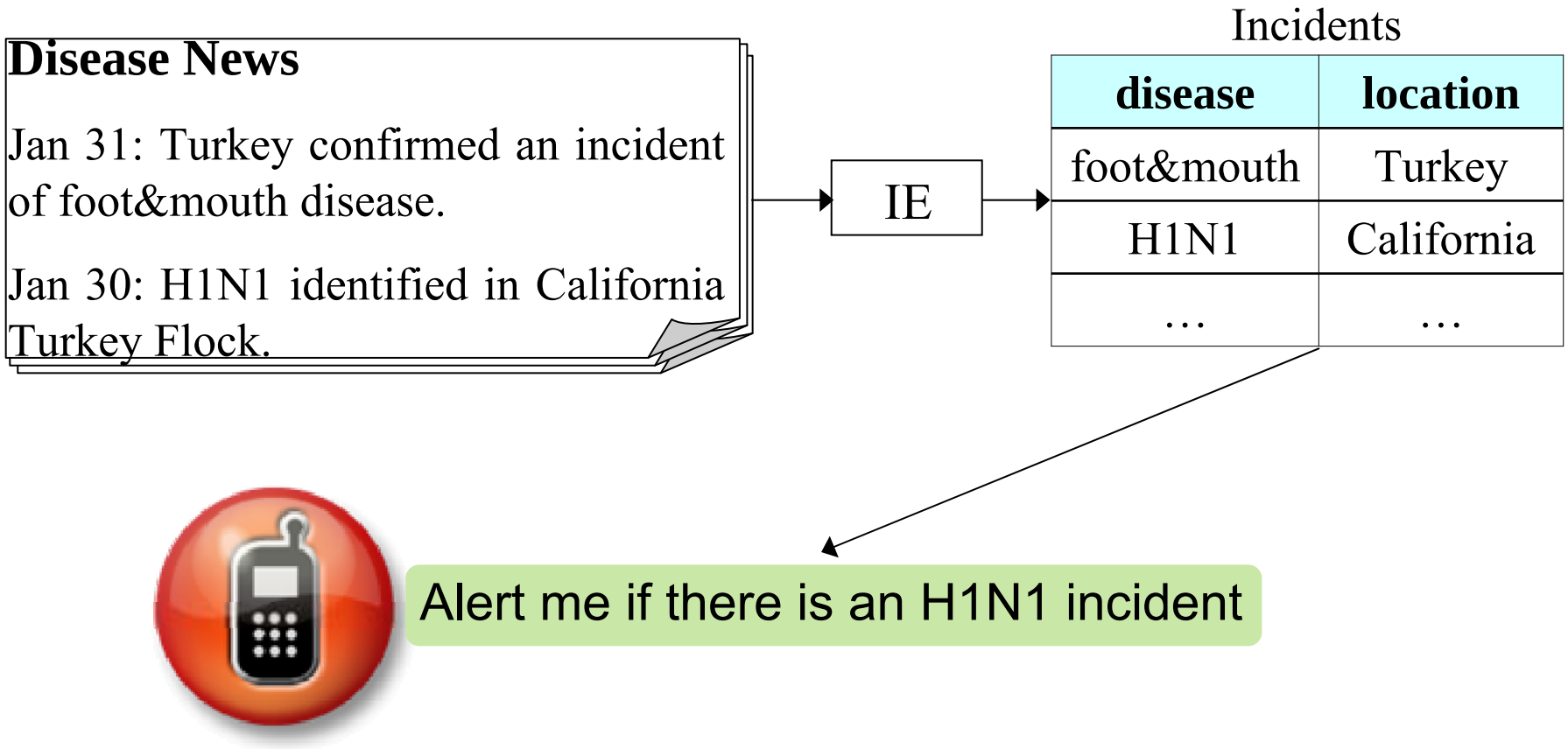
Relational-style optimization

[Shen *et al.* VLDB-07]
[Reiss *et al.* ICDE-08]

IE over evolving text

[Chen *et al.* ICDE-08]
[Chen *et al.* SIGMOD-09]
[Chen *et al.* TechReport-10]

IE over Evolving Text : An Example



day 1

Recent Incidents

Flu Situation Update

Influenza activity remained at the same levels.

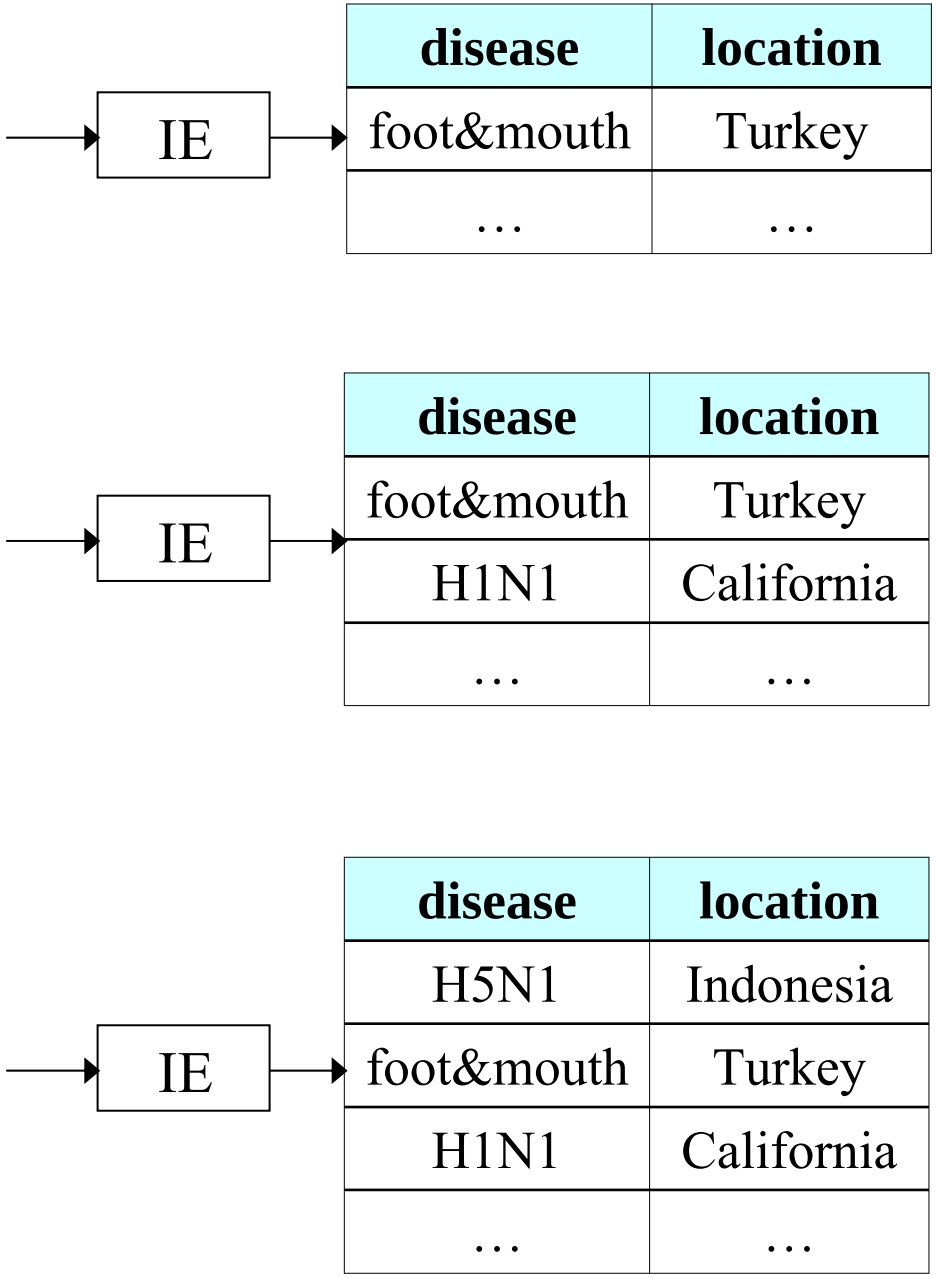
Reduction of Inventory at the

Disease News

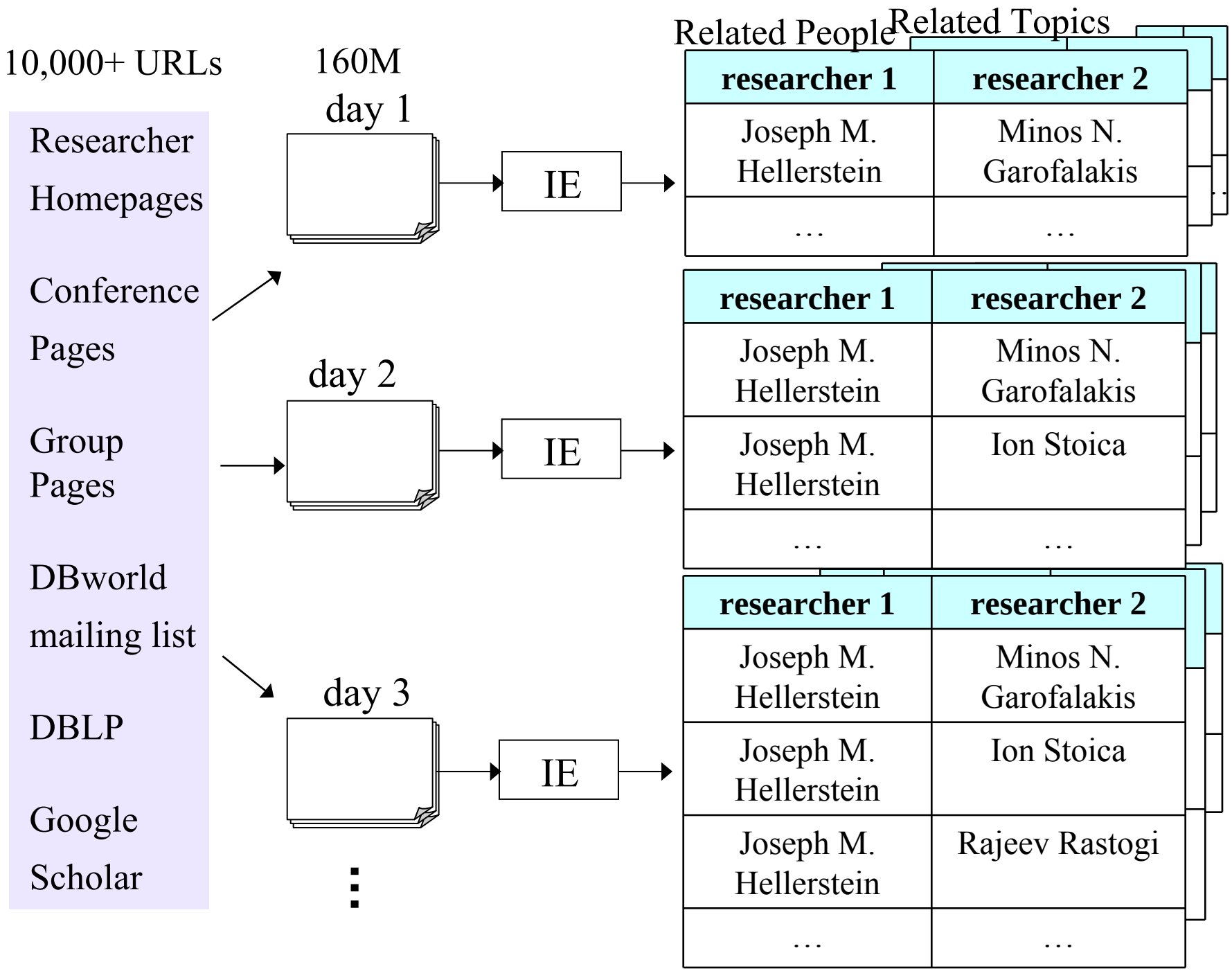
Feb 1: A new H5N1 case confirmed in Indonesia.

Jan 31: Turkey confirmed an incident of foot&mouth disease.

Jan 30: H1N1 identified in California Turkey Flock.



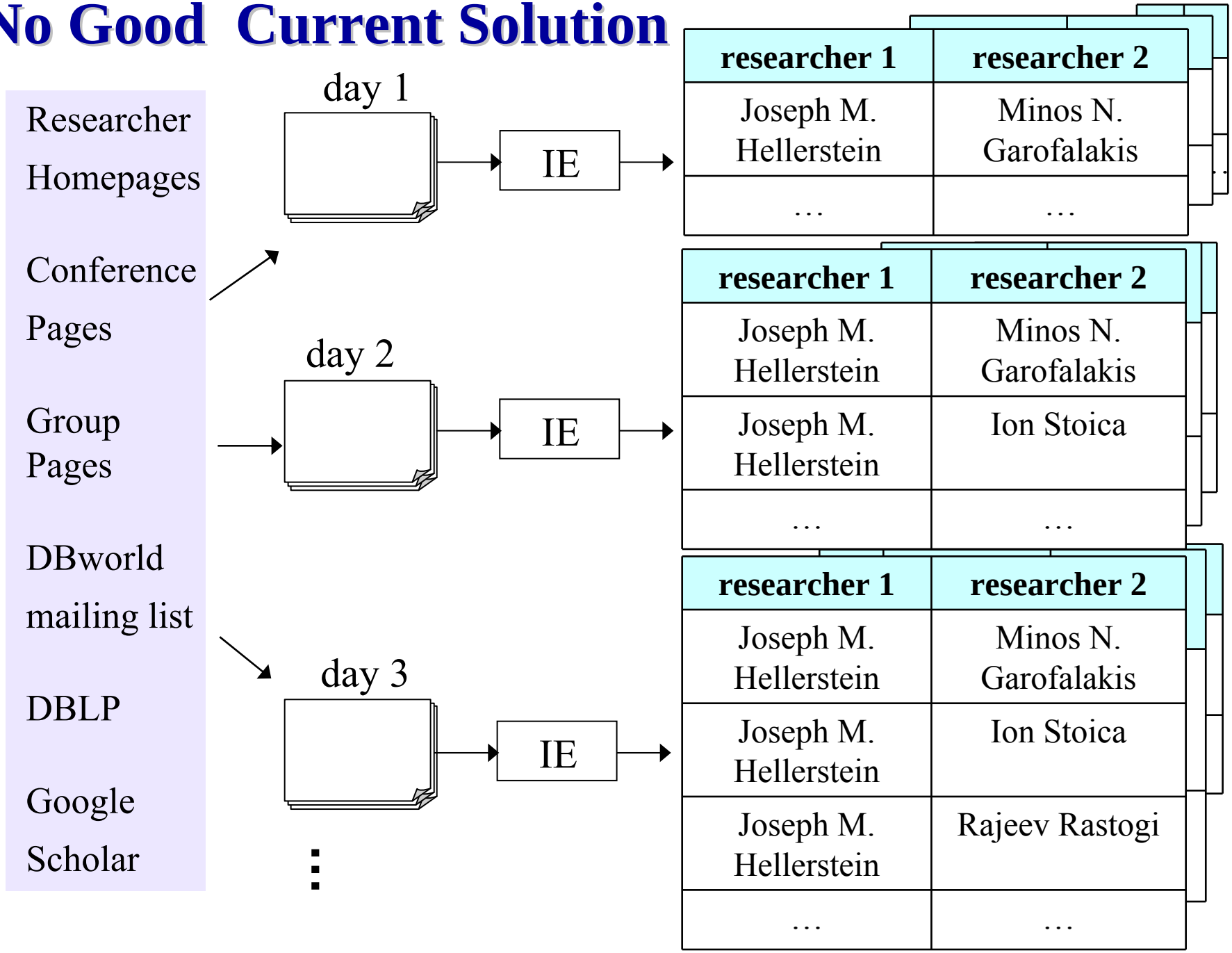
IE over Evolving Text: Another Example


10,000+ URLs
Researcher Homepages
Conference Pages
Group Pages
DBworld mailing list
DBLP
Google Scholar
160M
day 1
day 2
day 3
IE
IE
IE
Related People
Related Topics
researcher 1
researcher 2
Joseph M. Hellerstein
Minos N. Garofalakis
…
Joseph M. Hellerstein
Ion Stoica
Joseph M. Hellerstein
Rajeev Rastogi

Many Other Applications of IE over Evolving Text

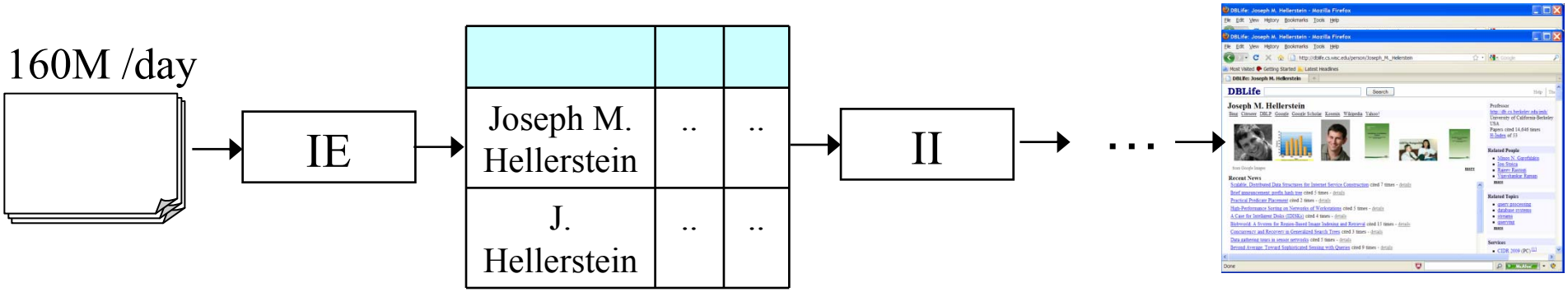
- **Impliance @ IBM**
  - manages information on enterprise intranets
  - finds the latest information
- **IWP @ Univ. of Washington and YAGO @ MPI**
  - extract structures from Wikipedia
  - keep extracted structures up-to-date
- **Blogscope @ Univ. of Toronto**
  - monitors the blogosphere
- **…**

No Good Current Solution



No Good Current Solution

- **Inefficient**
  - e.g., takes 8+ hours in DBLife everyday



- **Unsuitable for time-sensitive applications**
  - e.g., stock and auction
  - cannot finish extracting in time
- **Unsuitable for interactive debugging**
  - long debug loop

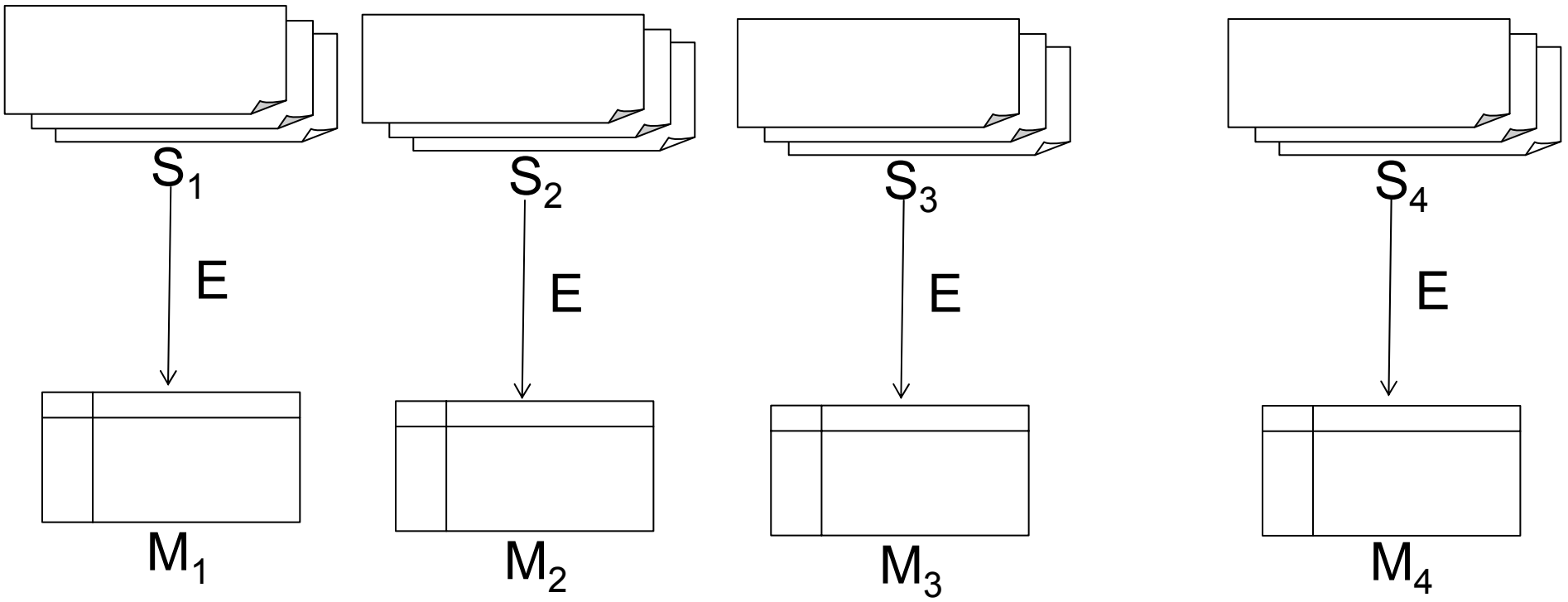
My Dissertation Contributions

Developed efficient IE solutions over evolving text that

- recycle previous IE efforts
- guarantee correctness
- deal with large text corpora
- deal with large IE programs
- deal with complex learning-based IE programs

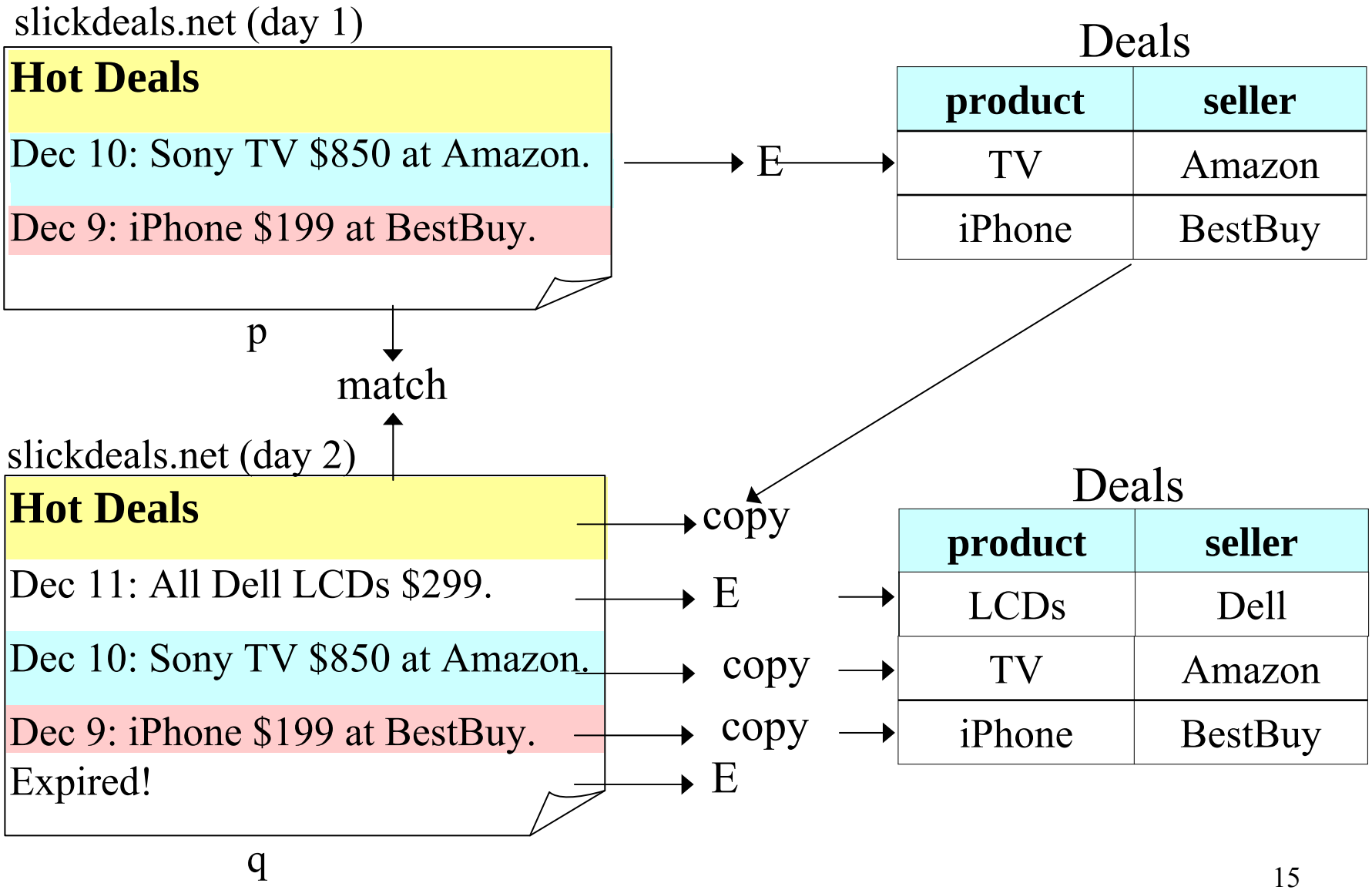
Results in [ICDE-08], [SIGMOD-09] and [TechReportA-10]

Problem Definition



Can we get M_4 faster than applying E to S_4 from scratch?

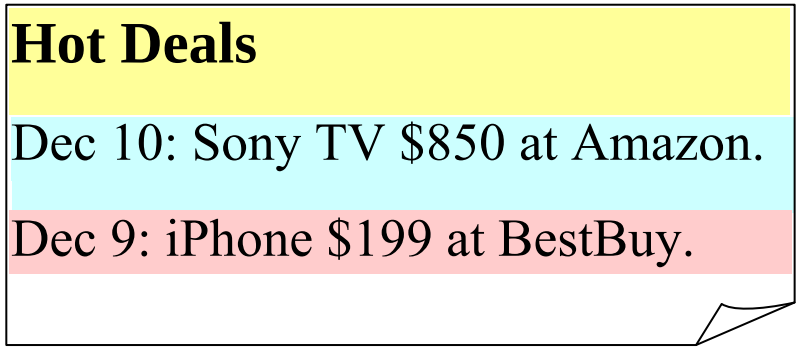
Cyclex [ICDE08]: Recycling Extraction



Not Always Correct to Copy Everything!

E extracts a deal if (a) product & seller are within a window of at most 20 chars
(b) no "Expired!" is within 10 chars around the window.

slickdeals.net (day 1)

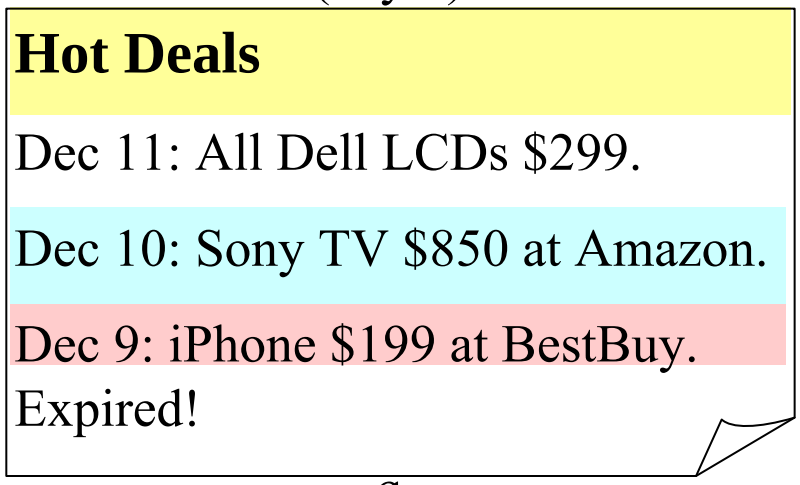


p

Deals

product	seller
TV	Amazon
iPhone	BestBuy

slickdeals.net (day 2)



q

Deals

product	seller
LCDs	Dell
TV	Amazon
iPhone	BestBuy

wrong!

Extractor Properties: Context

- **Many extractors only examine small "context windows" surrounding a mention to extract the mention.**

E extracts a deal if (a) product & seller are within a window of at most 20 chars
(b) no "Expired!" is within 10 chars around the window.

Hot Deals

Dec 11: All Dell LCDs $299.

Dec 10: Sony TV $850 at Amazon.

Dec 9: iPhone $199 at BestBuy.
Expired!

E's context = 10 chars

- **The text outside the context of a mention m is irrelevant for E to extract m.**

Revisit The Example

E extracts a deal if (a) product & seller are within a window of at most 20 chars
(b) no "Expired!" is within 10 chars around the window.

slickdeals.net (day 1)

Hot Deals

Dec 10: Sony TV $850 at Amazon.

Dec 9: iPhone $199 at BestBuy.

p

E's context = 10 chars

Deals

product	seller
TV	Amazon
iPhone	BestBuy

slickdeals.net (day 2)

Hot Deals

Dec 11: All Dell LCDs $299.

Dec 10: Sony TV $850 at Amazon.

Dec 9: iPhone $199 at BestBuy.
Expired!

q

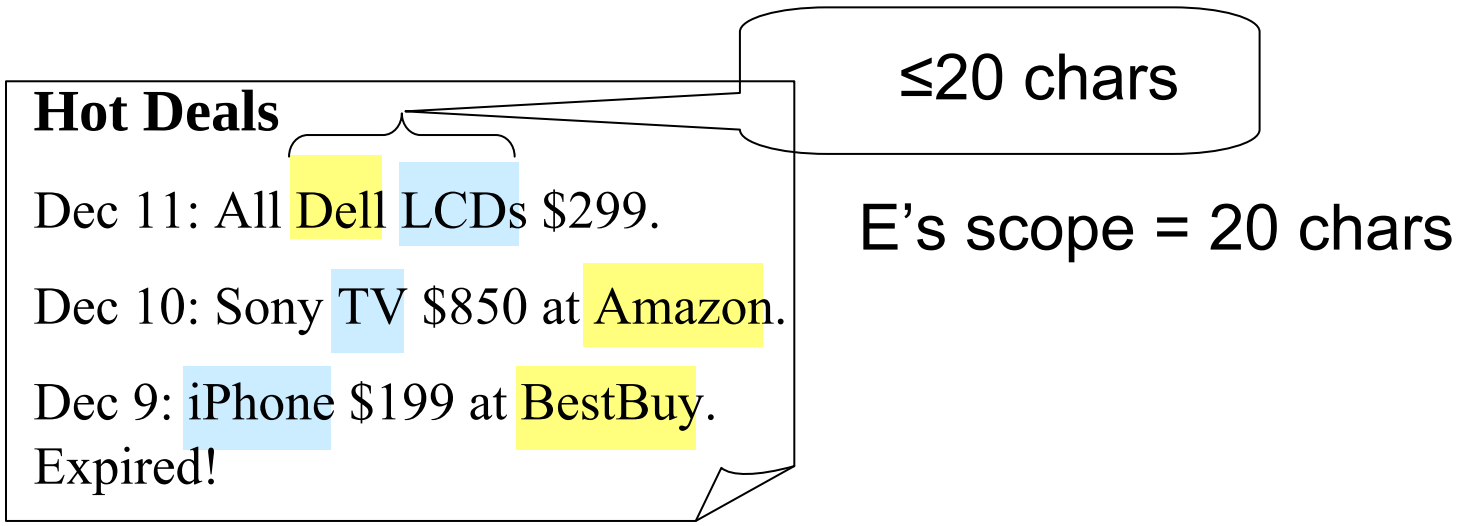
Deals

product	seller
LCDs	Dell
TV	Amazon

Extractor Properties: Scope

- **Mention attributes often appear in close proximity.**
  - an extractor E has scope α iff any mention produced by E at most spans α.

E extracts a deal if (a) product & seller are within a window of at most 20 chars
(b) no "Expired!" is within 10 chars around the window.



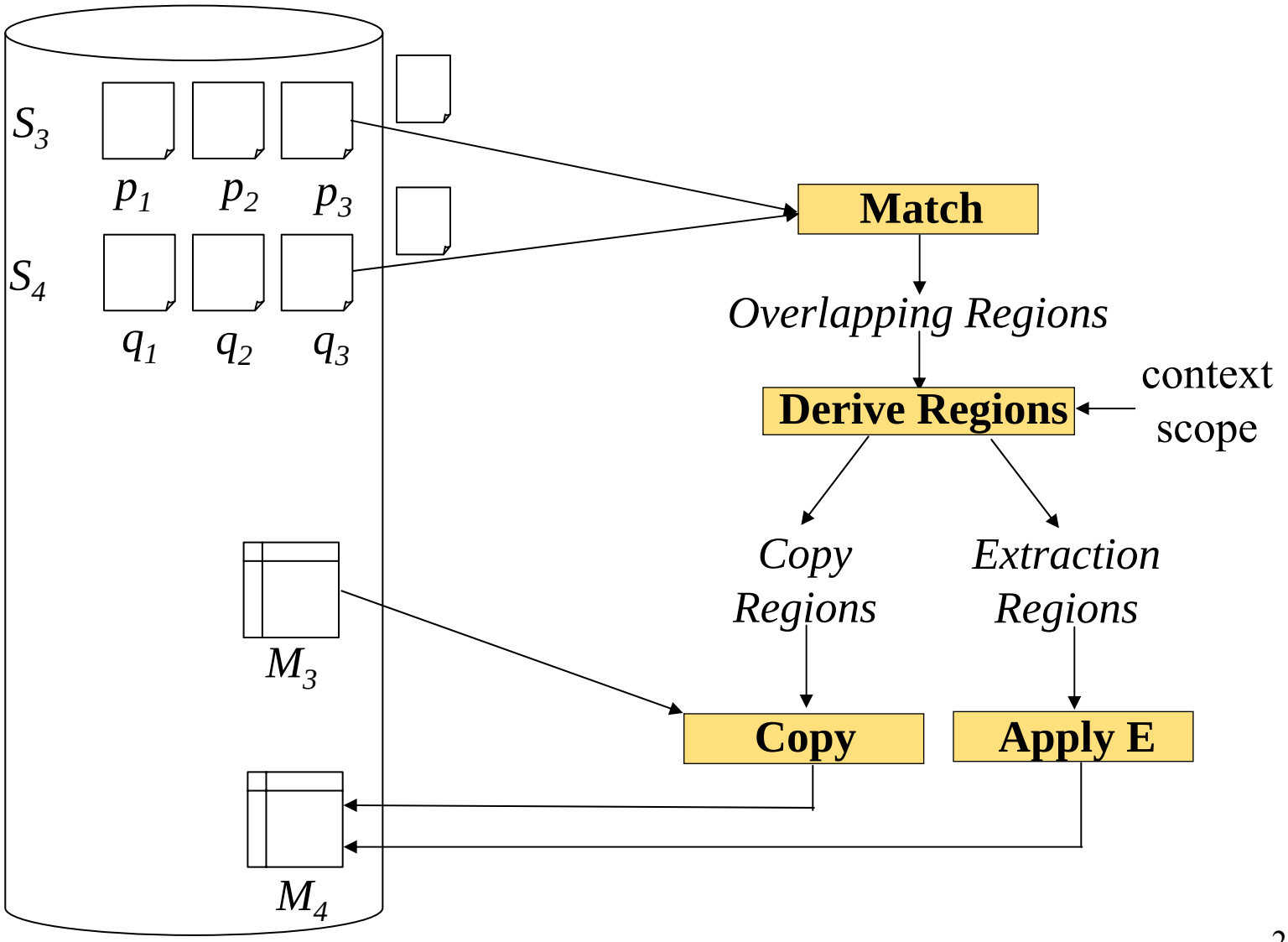
Obtain Scope and Context

- **Cyclex takes E, its scope & context**

- **Scope & context are provided by**
  - who writes E
  - who knows how E works

- **Even with loose scope and context**
  - can still guarantee correctness
  - can still reduce runtime
  - tighter scope/context ➔ more recycling

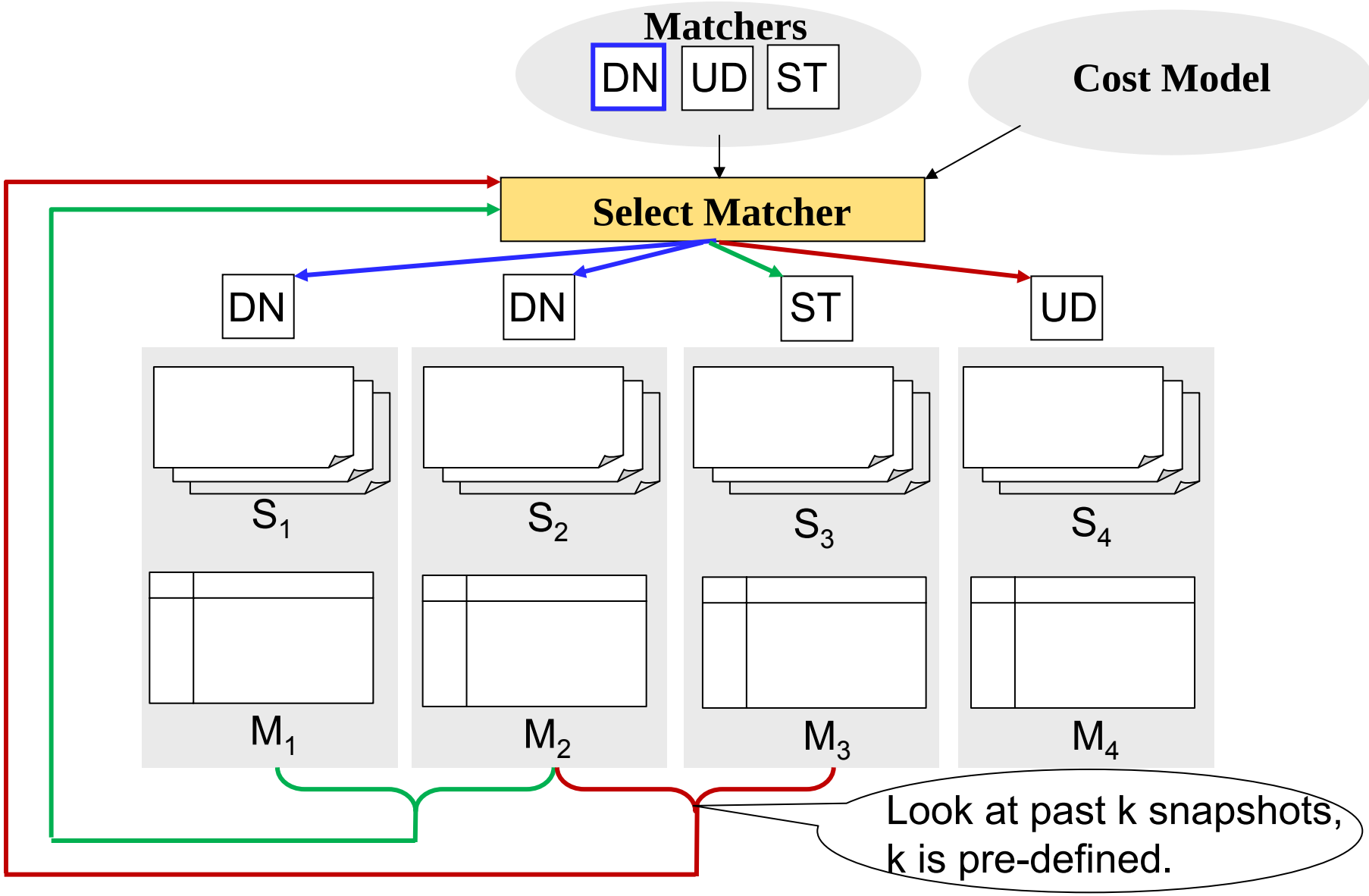
Baseline Approach



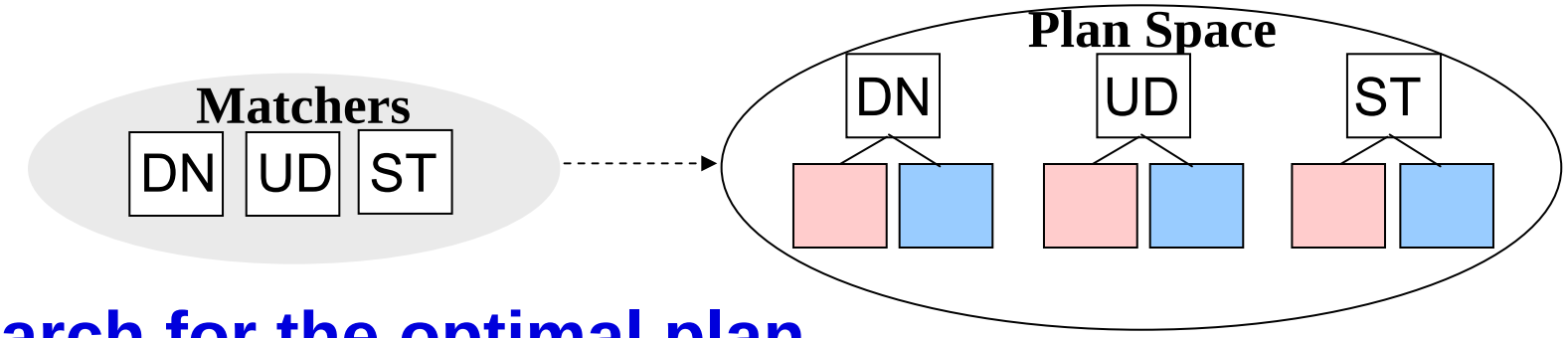
Matchers

- **Consider 3 matchers (more can be added)**
  - DN (Doing Nothing): 0 runtime, no overlapping regions
  - UD (Unix Diff): fast, find some overlapping regions
  - ST (Suffix Tree): relatively slow, all overlapping regions
- **Trade off runtime vs. size of overlapping regions**
- **No matcher is always optimal**

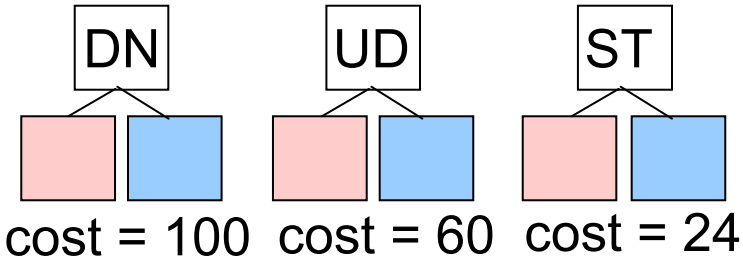
Select An Optimal Matcher: Adaptive and Cost-based Optimization

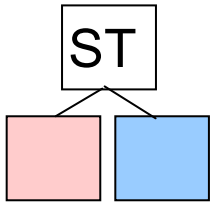


Select An Optimal Matcher: Cost-based Selection



- **Search for the optimal plan**
  1. enumerate all possible plans
  2. use cost model to estimate cost of each plan
  3. select plan with minimal cost

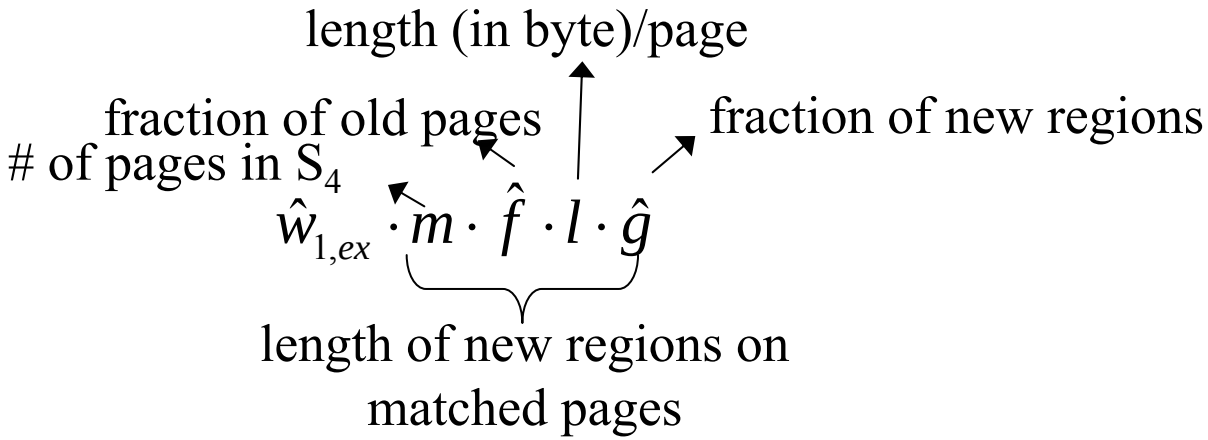




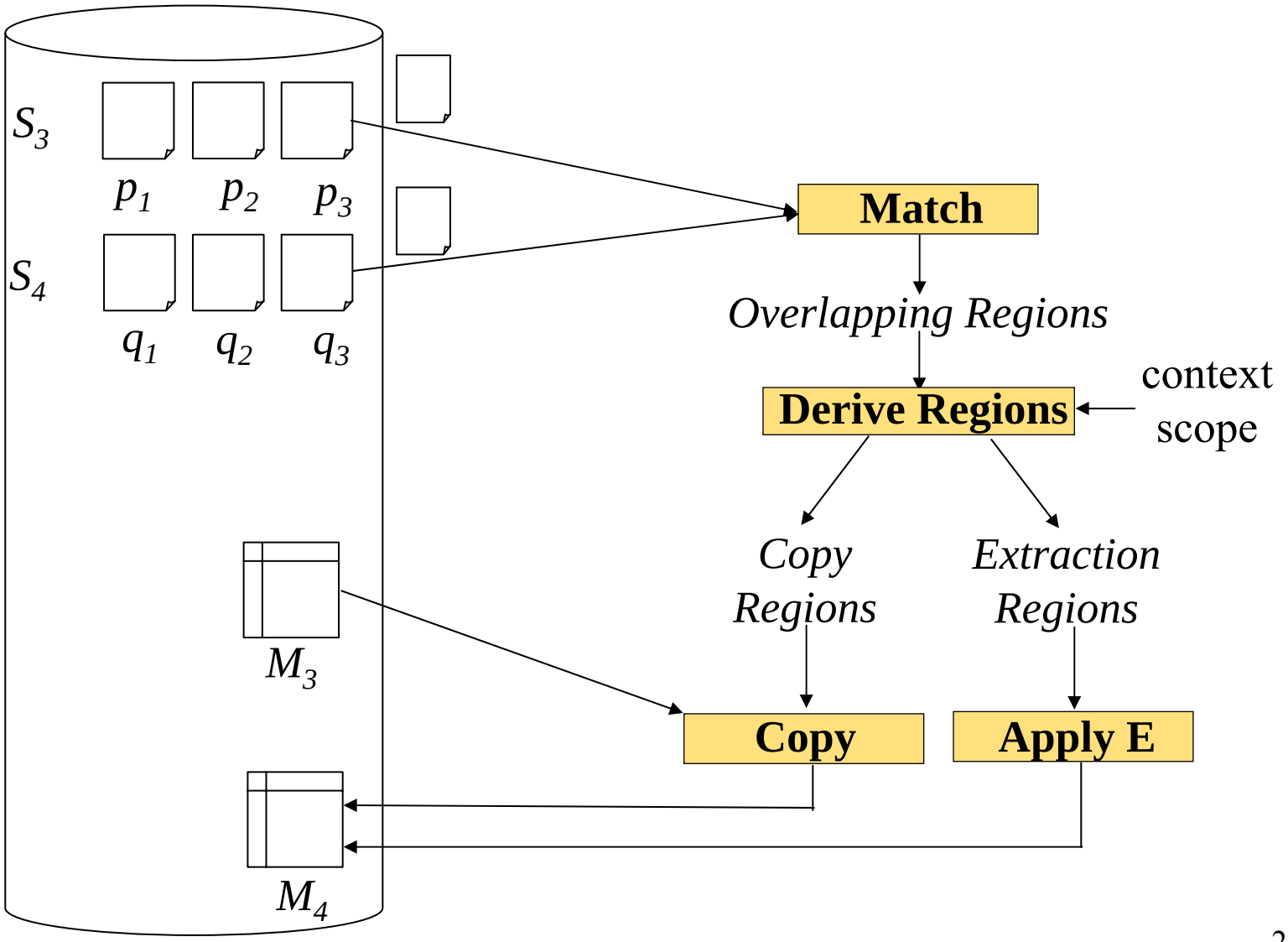
Cost Model

- **Captures matching/extraction/copy times**

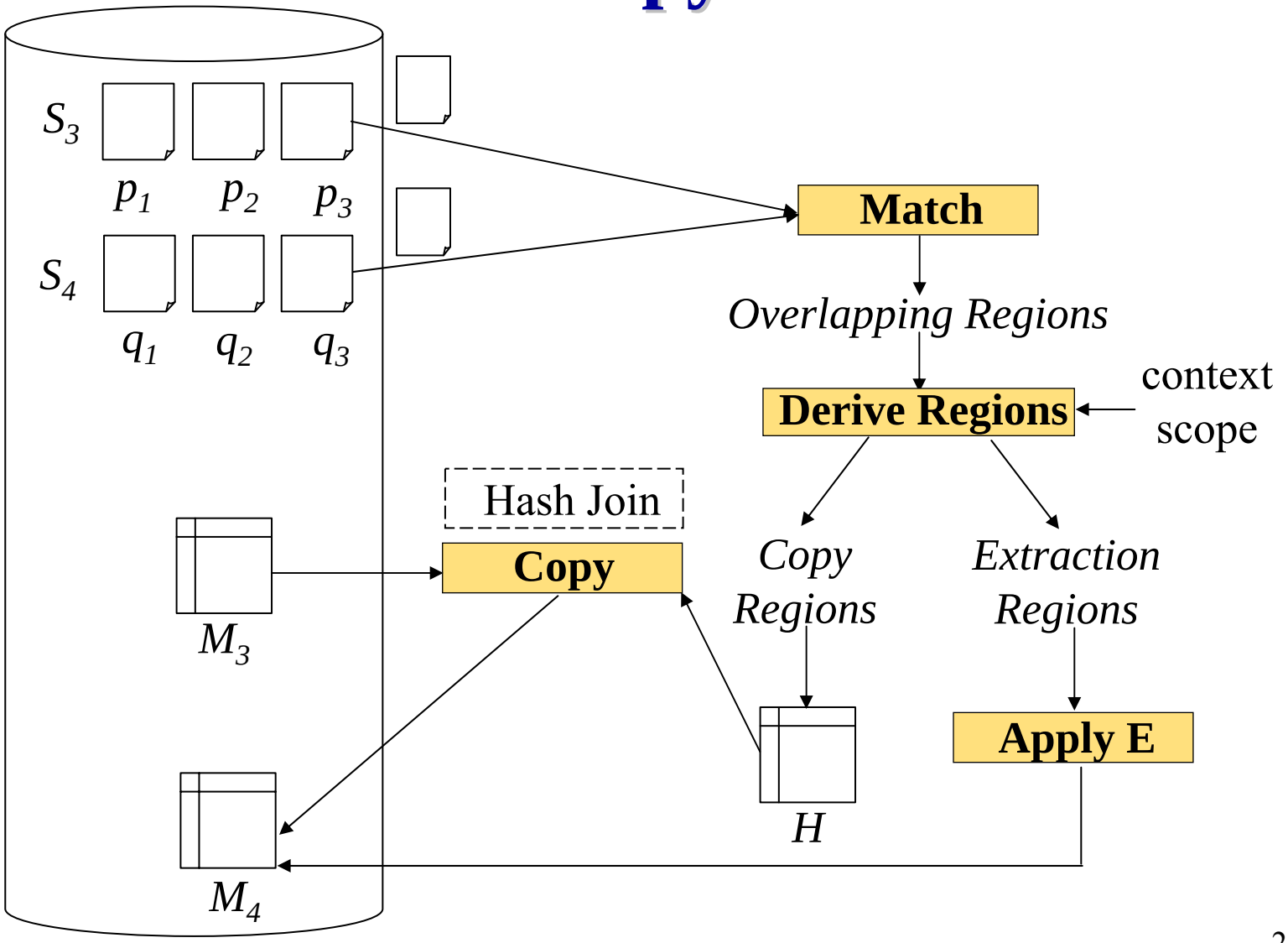
- **Captures text properties**



Baseline Approach



Interleave Matching, Extraction and Copy



Experimental Setup

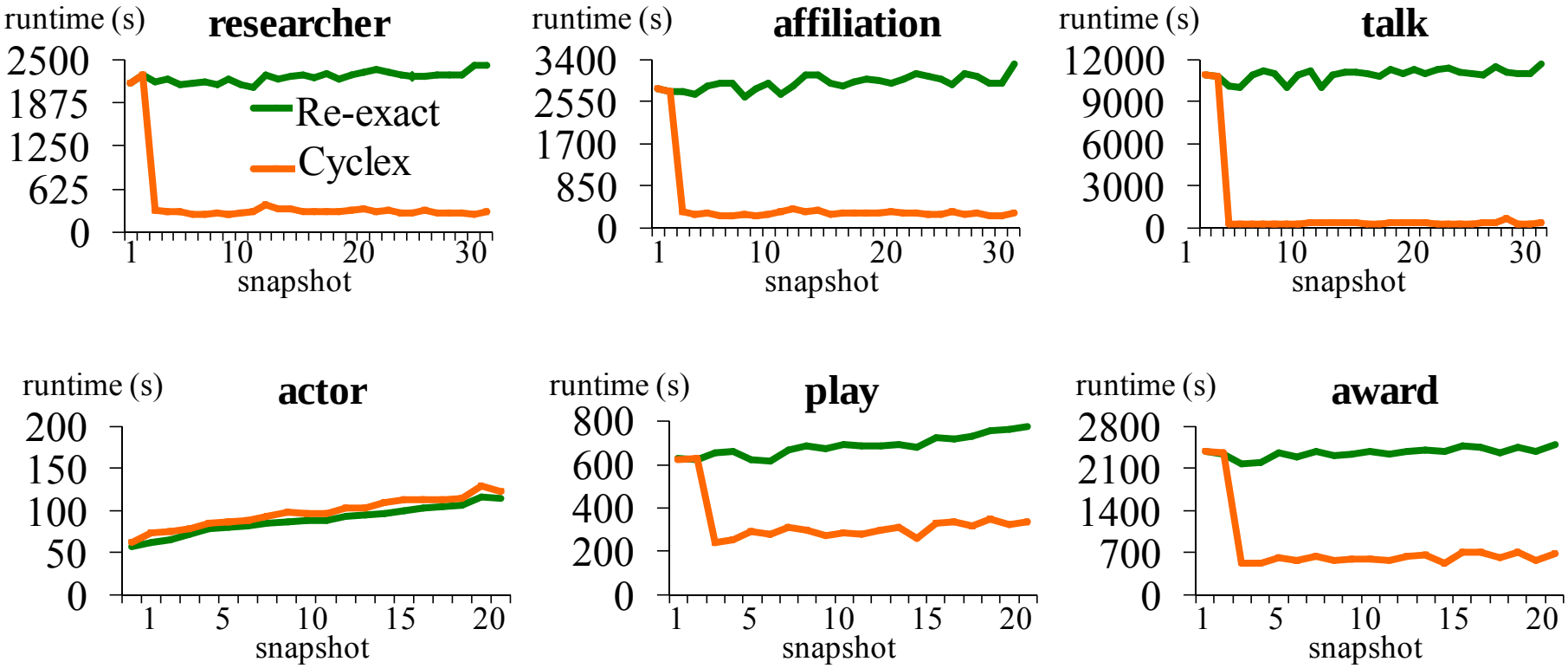
- **Datasets**

Data Sets	**DBLife**	**Wikipedia**
# Data Sources	980	925
# Snapshots	30	20
Time between snapshots	1 day	21 days
Avg # Page per Snapshot	10155	3038
Avg Size per Snapshot	180M	35M

- **Extractors**

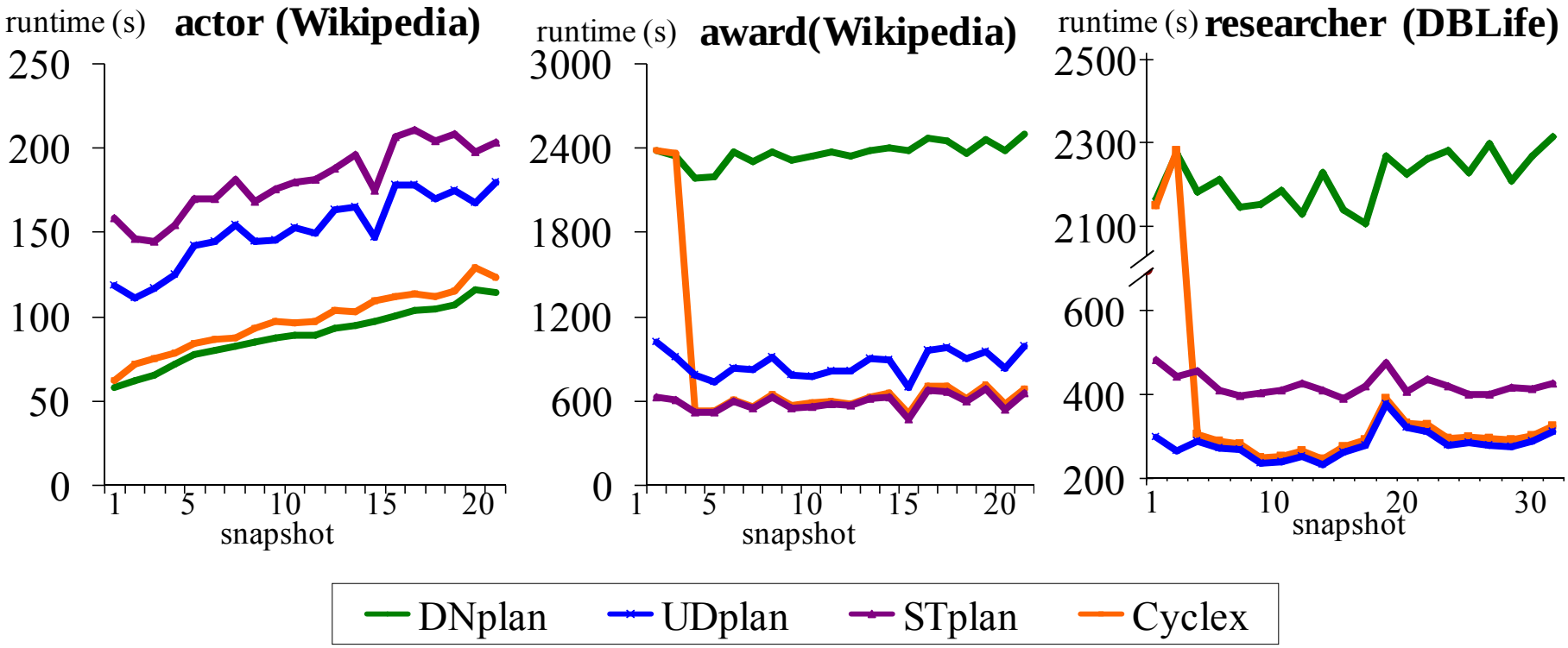
	DBLife			**Wikipedia**		
	researcher	affiliation	talk	actor	play	award
Scope α	32	93	400	35	96	250
Context β	3	7	10	3	4	10

Benefit of Recycling IE Results



- **In 5 out of 6, outperformed extract-from-scratch by 50-90%**

Importance of Optimization



- **No matcher is uniformly optimal**

Summary of Cyclex

- **Guarantee correctness**
  - model extractors using scope and context
- **Choose a good way to match pages**
  - adaptive optimization
  - cost-based decisions using text specific cost model
- **Efficiently interleave matching, copy and extraction**
  - a way to scan data once

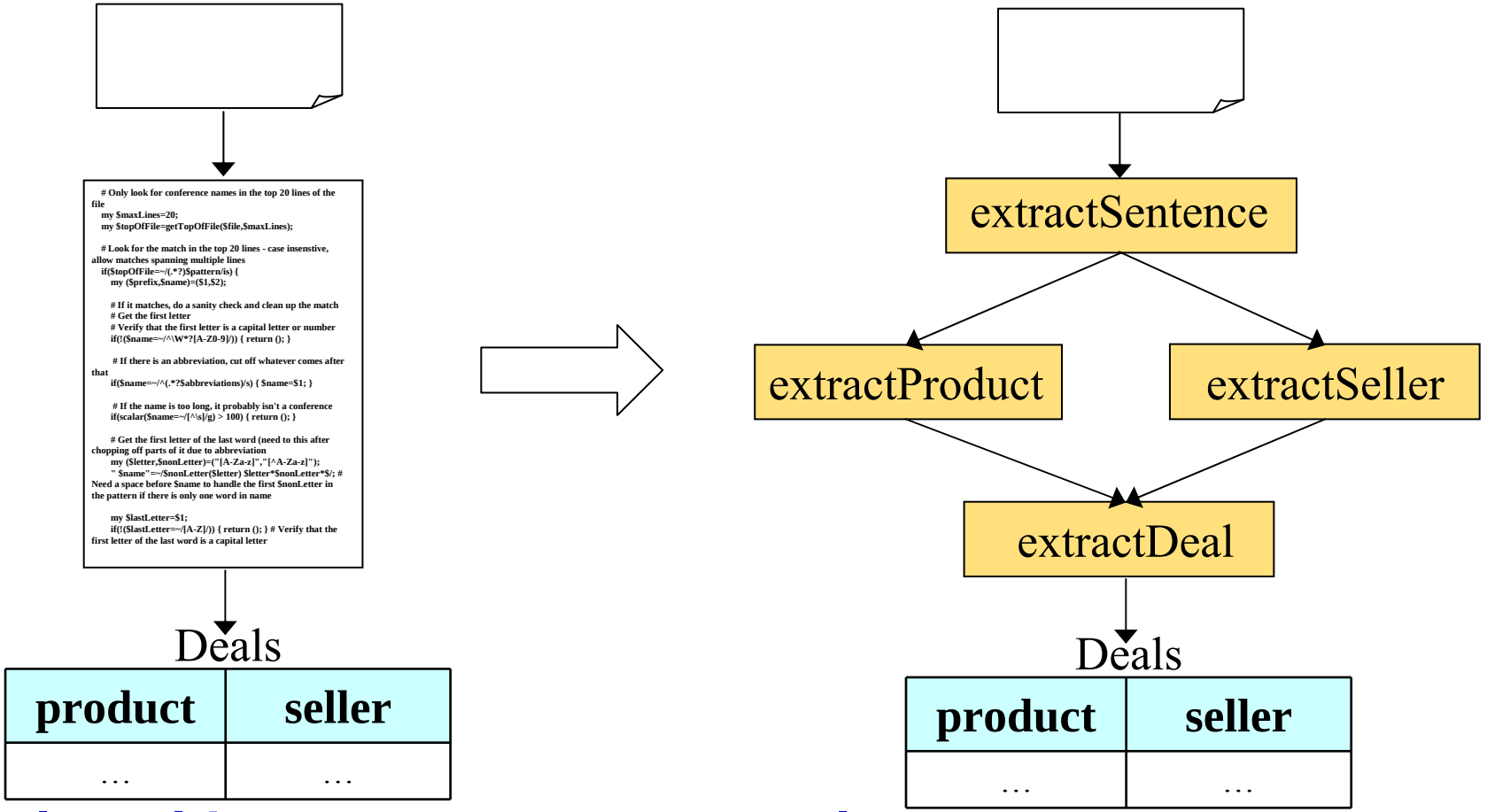
My Dissertation Contributions

Developed efficient IE solutions over evolving text that

- recycle previous IE efforts
- guarantee correctness
- deal with large text corpora
- deal with large IE programs
- deal with complex learning-based IE programs

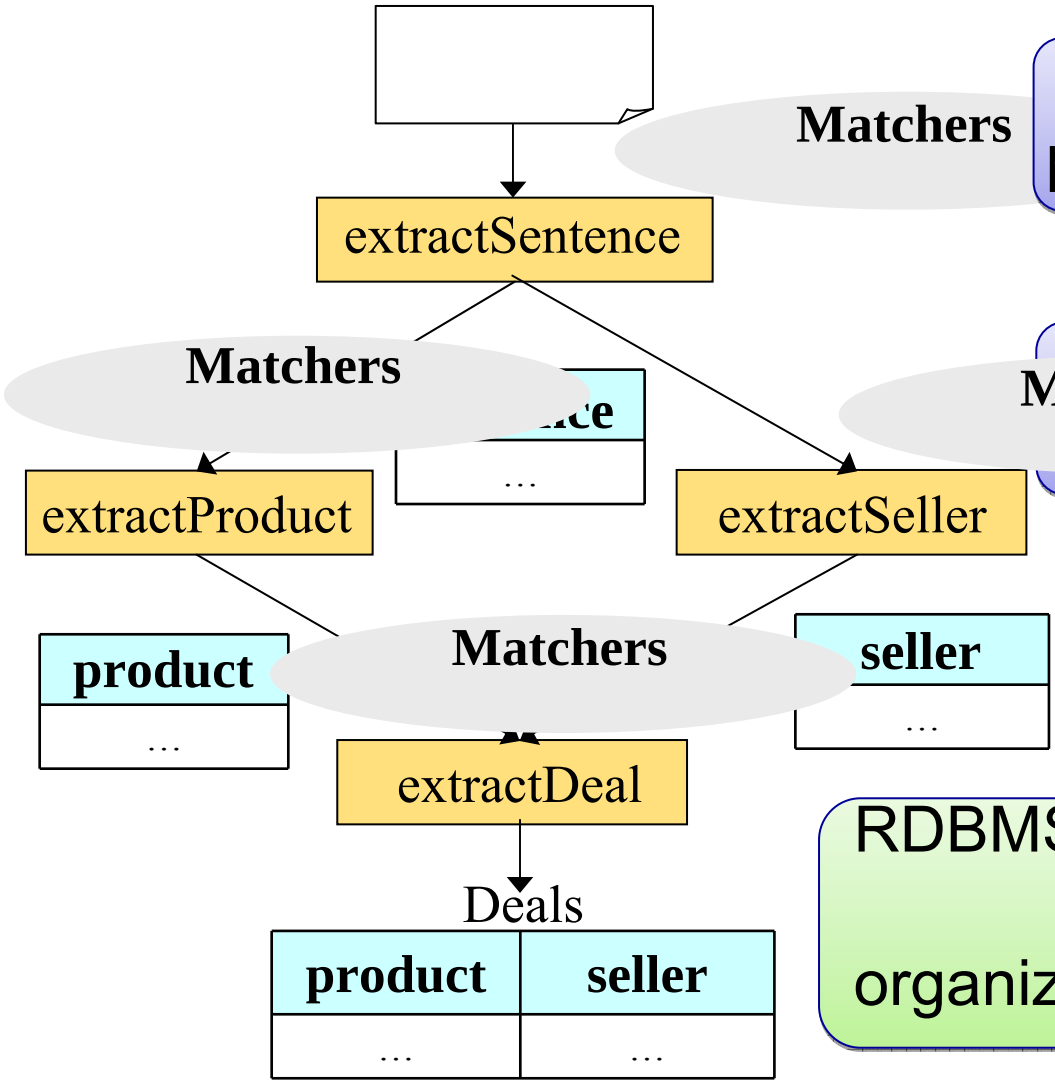
Results in [ICDE-08], [SIGMOD-09] and [TechReportA-10]

Large and Complex IE Programs



- **Real-world IE programs are complex**
  - Avatar@IBM: 25+ blackboxes
  - DBLife@WISC: 45+ blackboxes

Delex [SIGMOD09]: Decompose and Recycle



How to find a fast recycling plan?

Matchers

…ordinate recycling …blackboxes"?

exploit databases techniques!

RDBMS query optimization

organize files for sequential access

Experiment Setup

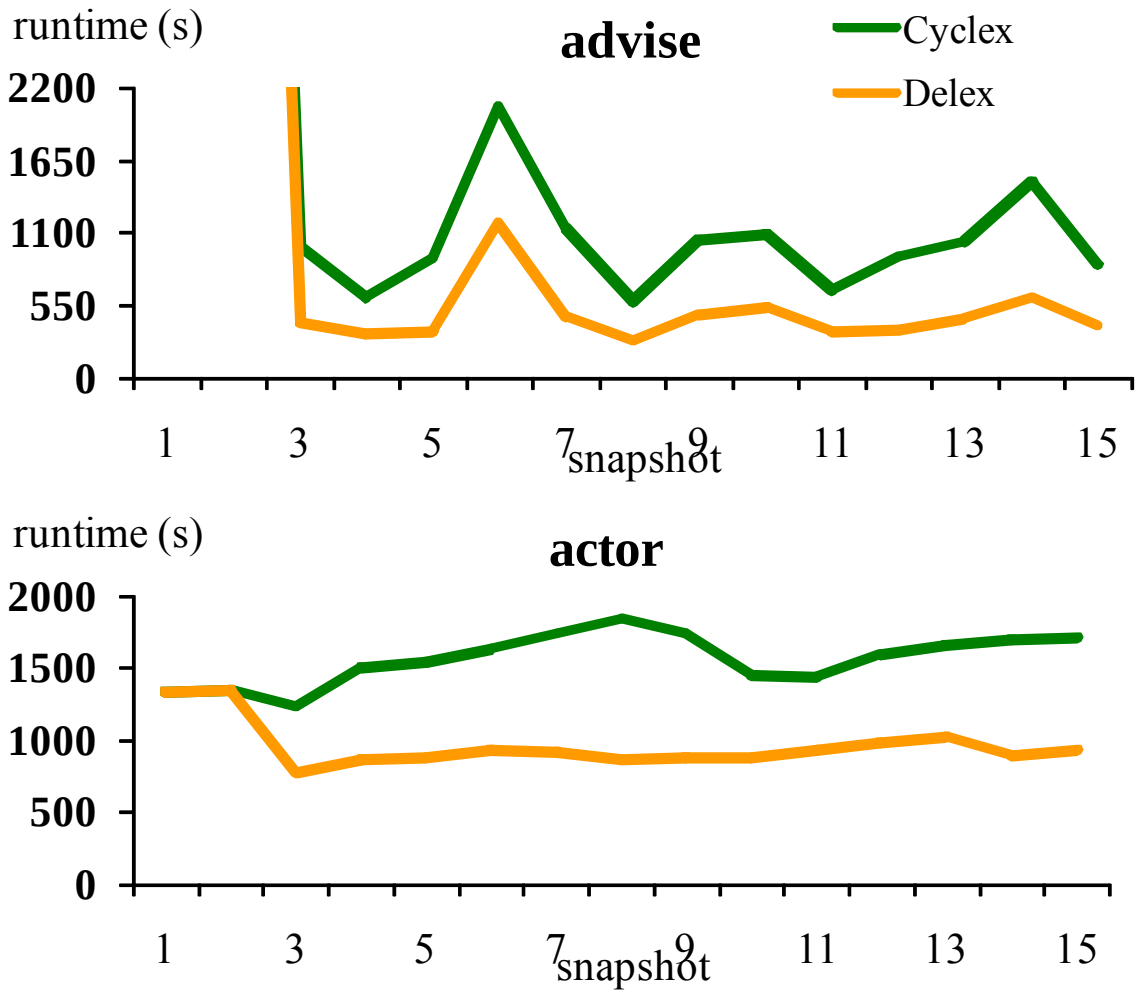
- **Datasets**

Data Sets	**DBLife**	**Wikipedia**
# Snapshots	15	15
Time between snapshots	2 days	21 days
Avg # Page per Snapshot	10155	3038
Avg Size per Snapshot	180M	35M

- **IE Programs : Rule-based and Learning-based IE Programs**

	DBLife (Rule-based)			**Wikipedia (Rule-based)**			**Wikipedia (Learning -based)**
	talk	chair	advise	blockbuster	play	award	actor
# of IE "Blackboxes"	1	3	5	2	4	6	5

Runtime Comparison



- **Delex drastically cuts runtime of Cyclex by 45-71%**

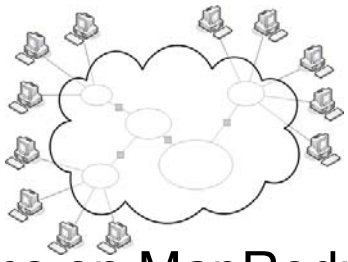
(See paper for more experiments)

Related Work

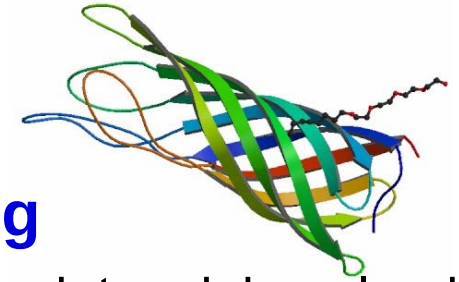
- **Early works on IE**
  - [Bikel UAI-97] [McCallum KDD-00]…
  - focus on improving accuracy
- **Recent works on IE**
  - [tutorial in KDD-06, SIGMOD-06] [Gravano et al, SIGMOD-06]…
  - consider developing efficient IE
  - do not consider evolving text corpora
- **Evolving text data**
  - [McCann VLDB05] [Lim WWW03]
  - consider problems other than IE
- **Incremental view maintenance**
  - [Gupta&Mumick][Widom et al, SIGMOD95]…
  - only consider relational operators
  - assume changes to the inputs are available

My Other Contributions

- **Web community systems**
  - DBLife [DE Bulletin-06, VLDB-07a, CIDR-09]



- **Text mining in cloud**
  - scalable mining algorithms on MapReduce clusters[TechReportB-10]



- **Biology sequences mining**
  - incorporate domain knowledge into mining algorithms [KDD-03, Bioinformatics-04]
  - interpret complex mining models with human-understandable rules [SFUThesis]

Future Work

- **Maintain IE and information integration (II)**
- **Develop/Maintain IE/II over large clusters**
  - build efficient, scalable and robust IE/II
  - declaratively develop and automatically optimize IE/II
- **Develop user-friendly and efficient data analysis tools**
  - analyze non-structured data, e.g., text and biology sequences
  - let non-database users such as biologists ask query easily and obtain answers efficiently

Conclusions

- **IE over evolving text is increasingly important!**

- **Developed solutions that**
  - recycle previous IE to reduce runtime
  - guarantee correctness
  - deal with large text corpora / large IE programs
  - deal with complex learning-based IE programs

- **Database techniques are increasingly critical for developing efficient IE solutions**
  - in collaboration with other communities: NLP, AI, Web, IR, KDD, HCI