IBM Research

Project ES2:
Mining and Analytics for IBM Intranet Search

Shivakumar Vaithyanathan
Sr. Manager Search, Analytics and Information Integration
IBM Almaden Research Center

June 12, 2008

Information Management at Almaden

- **Advanced IM Architectures (Hamid Pirahesh)**
  - Guy Lohman (Blink)
  - Eugene Shekita (Cloud9 → Cloud Infrastructure)
  - John McPherson (Cloud analytics)
- **Information Systems (Shiv Vaithyanathan)**
  - Search and Analytics (Avatar Group)
  - Information Integration (Clio Group)

Acknowledgements

- **Sriram Raghavan**
- **Sandeep Tata**
- **Huaiyu Zhu**
- **Vuk Ercegovac**
- **Rajasekar Krishnamurthy**
- **Yunyao Li**
- **Frederick Reiss**
- **Fei Chen (summer student from Wisconsin)**

What makes IBM Intranet search hard?

Axioms from Fagin *et al* www 2003

- **Web**
  Economic & Social incentives to be in the top 10 results of Web search queries
  Drive traffic to your web site

  Intranet
  No economic incentives
  Isolated pages that may not be linked heavily if at all

- **Significant fraction of queries are "navigational"**
  There is often exactly "one" right answer to a query and that needs to be ferreted out

  (Most of top 6500 queries are navigational)

- **Geographically disperse organization**
  350K users across over 80+ countries. Why do we care?
  Because **the same query may have a different "right" answer depending on who is issuing it**

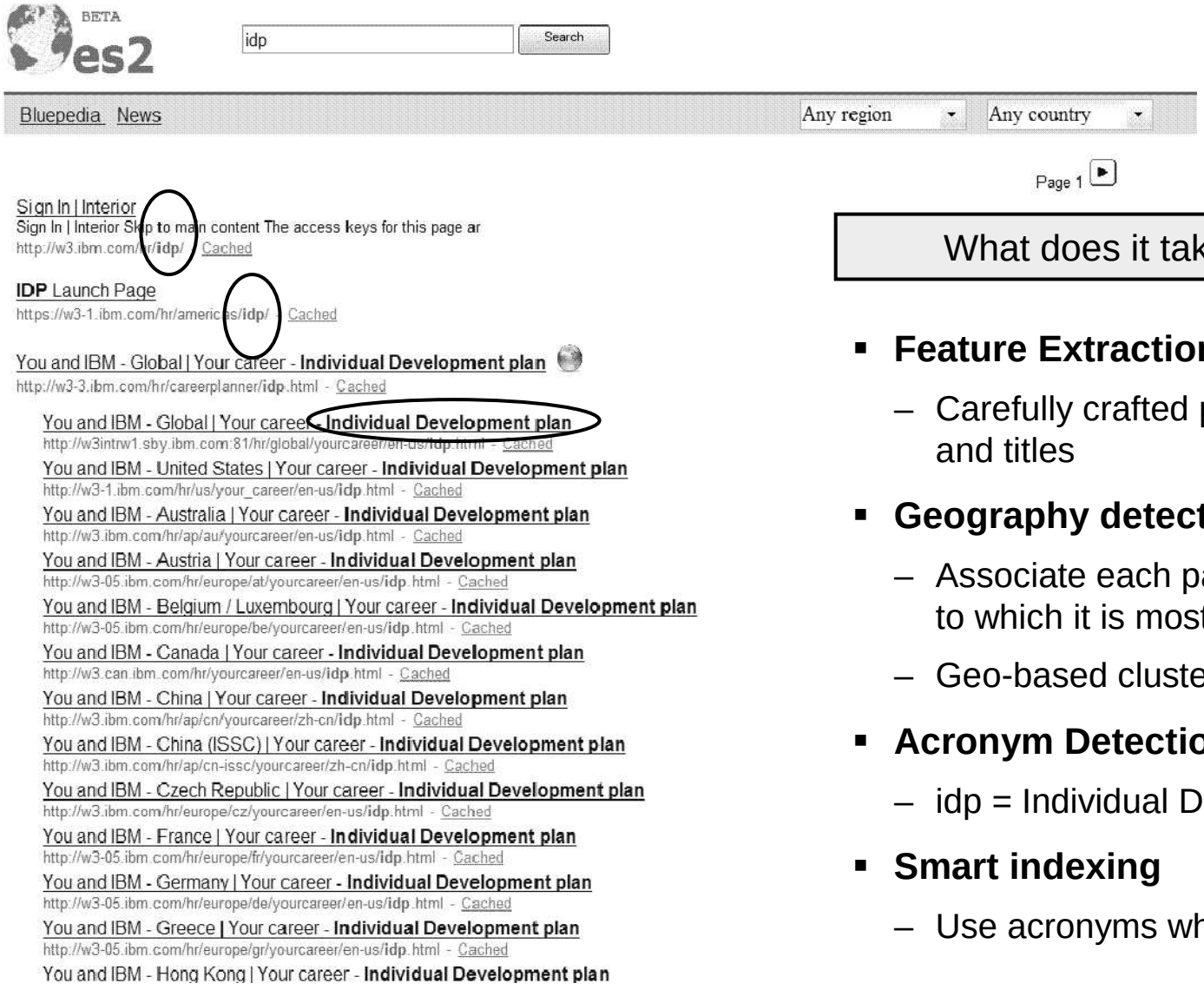
Project ES2 Goal

While improving search on the IBM Intranet, build a testbed to invent and deploy analytics on very large scale enterprise data driven by real users and queries

Outline of the Talk

- **Motivating search queries**
- **Analytics and Mining in ES2**
  - Overall workflow
  - Local Analysis (LA)
  - Global Analysis (GA)
- **Migration to Hadoop**
  - LA & GA on Hadoop
  - Mapping analysis/mining algorithms onto MapReduce
    - Three specific examples

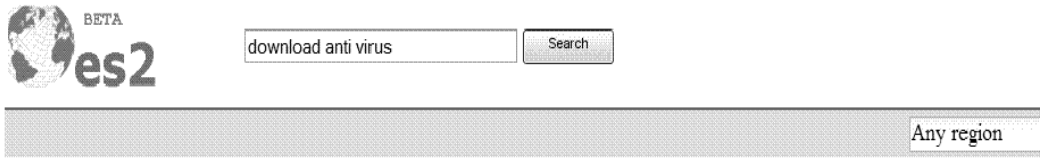
Search Example 1: idp



What does it take to get this result ?

- **Feature Extraction**
  - Carefully crafted patterns applied to URLs and titles
- **Geography detection**
  - Associate each page with the country/region to which it is most relevant
  - Geo-based clustering for grouping results
- **Acronym Detection**
  - idp = Individual Development Plan
- **Smart indexing**
  - Use acronyms when generating index terms

Search Example 2: download anti virus



IBM Standard Software Installer | Symantec AntiVirus 9.0.5.1000 (EMEA)
Standard Software Installer | Symantec AntiVirus 9.0.5.1000 (EMEA... Standard Software Installer > Symantec AntiVirus 9.0.5.1000... by Symantec AntiVirus Parent Servers in EMEA. If you already have Symantec Antivirus 9 installed, you
http://w3-1.ibm.com/download/standardsoftware/nav/nav951emen/nav951emen.html?GeoName=Europe, Middle... - Cached

download IBM **AntiVir**us in French
http://ibmav.watson.ibm.com/**download**/french.html - Cached

IBM **Virus** CERT | Downloads
http://w3-03.ibm.com/**virus**/**download**/certsoftware.html - Cached

LTCDesktop/Developer - LTC Wiki
https://ltc3.linux.ibm.com/wiki/LTCDesktop/Developer - Cached

IBM - Desktop Security
Featured literature IBM Proventia Desktop Endpoint Security Access Control Updated... your desktops ahead of the threat with multi-layered protection from IBM Internet Security Systems... - Desktop Security
http://www-935.ibm.com/services/us/index.wss/offerfamily/iss/a1026607 - Cached

IBM Export Regulation Office | Chapter 1
https://w3-01.ibm.com/chq/ero/ero.nsf/Pages/USERP-CHAPTER 1 - Cached

K0QA Policies: **Virus** Protection
K0QA Policies: **Virus** Protection... . **Virus** Protection... This document describes the process for maintaining current **virus** protection on hosts
http://pandora.lenexa.ibm.com/is/policies/security/**virus**_protection.html - Cached

IBM GTSS | pSeries mail tips by Tesch
IBM GTSS | pSeries mail tips by Tesch... pSeries by Tesch... GTSS
http://tesch.dfw.ibm.com/pseries/teschmail.html - Cached

[TITLE UNAVAILABLE]
http://www-05.ibm.com/services/ecalib/doc/2158-007.txt - Cached

Virus Alert Page
http://w3.pok.ibm.com/organization/solution/question/**virus**ale.htm - Cached

What does it take to get this result ?

- **Page categorization**
  - Through custom patterns applied to URLs, titles, and other features of a Web page
  - Categorize these pages as "ibm standard software installer" pages
- **Intelligent query interpretation**
  - download anti virus →
    anti virus category="ibm standard software installer"

Search Example 3: gj chaitin

G J **Chaitin** Home Page
G J **Chaitin** Home Page " Dieu a choisi celuy qui est... le plus simple.... 32, 33 for the above texts.] G J **Chaitin** Home Page This website contains Greek... of **Chaitin**'s published papers, many book chapters, and the LISP, Java, C, and Mathematica software
http://w3.watson.ibm.com/~**chaitin**/index.html - Cached

Chaitin, The Unknowable
Chaitin, The Unknowable THE UNKNOWABLE G J **Chaitin**, IBM Research... Gödel's Proof . To future understanding! G.J. **Chaitin** 11 February 1999 **chaitin**@watson.ibm.com http://www.umcs.maine.edu /~**chaitin** http://www.cs.auckland.ac.nz/CDMTCS/**chaitin**
http://w3.watson.ibm.com/~**chaitin**/unknowable/index.html - Cached

Le Hasard des Nombres
Le Hasard des Nombres Le Hasard des Nombres La Recherche 22, N o... théorie des nombres ont des réponses tout aussi aléatoires que le résultat...;noncé dans le langage de la théorie des nombres, laquelle constitue le soubassement des
http://w3.watson.ibm.com/~**chaitin**/paris.html - Cached

Chaitin, The Unknowable
http://w3.watson.ibm.com/~**chaitin**/unknowable/bib.html - Cached

Review of "Exploring RANDOMNESS" by Newton C. A. da Costa
Review of "Exploring RANDOMNESS" by Newton C. A. da Costa Folha de S. Paulo... a aleatoriedade Newton da Costa especial para a Folha Exploring Randomness de G.J.... extraordinariamente original e idiossincrático de **Chaitin**. Newton C. A. da Costa é
http://w3.watson.ibm.com/~**chaitin**/ait/costa.html - Cached

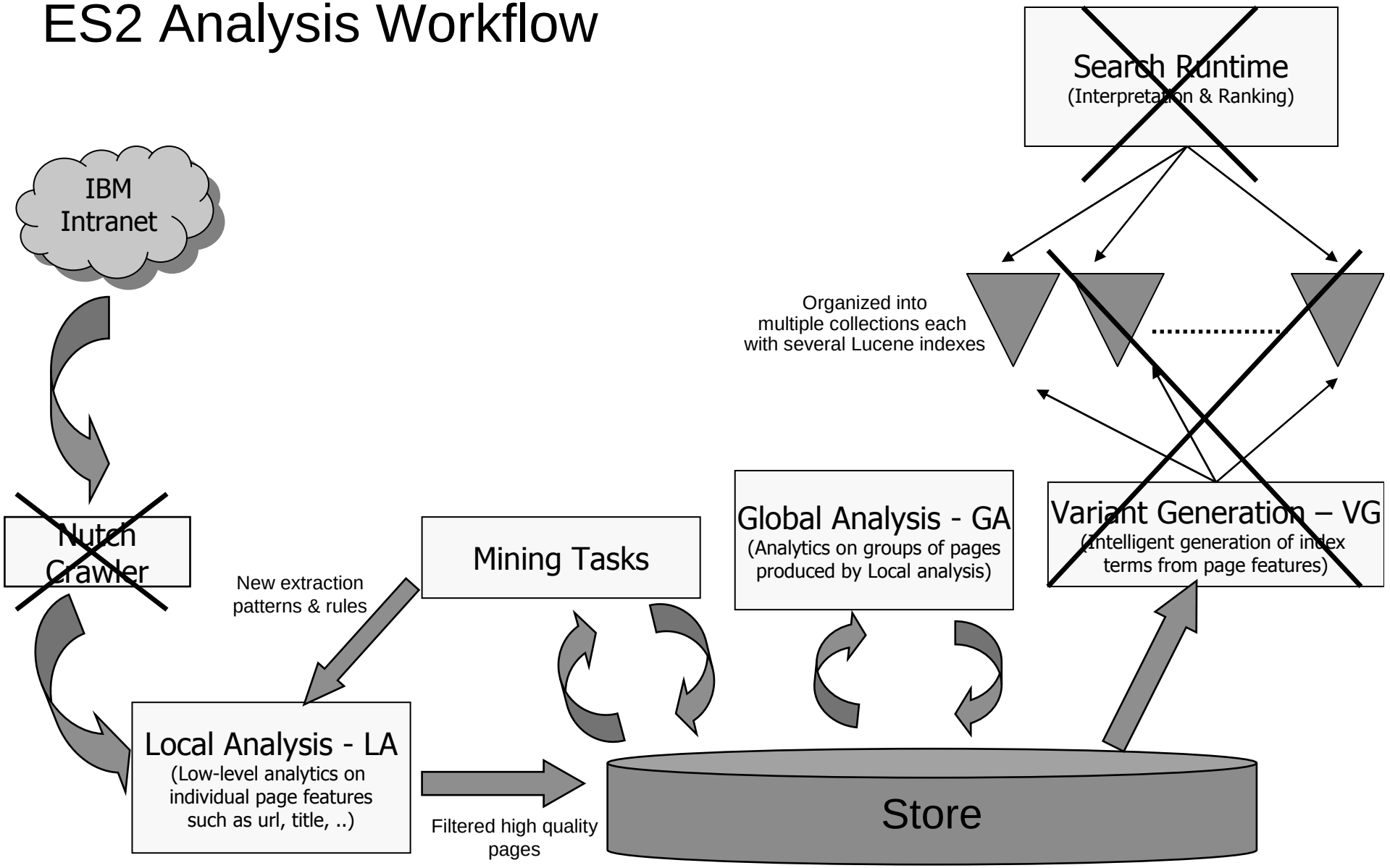
Randomness everywhere
and views / NATURE / vol 400 Mathematics / Randomness everywhere... Randomness everywhere
http://w3.watson.ibm.com/~**chaitin**/nature.html - Cached

Probability and Program-Size for Functions
Probability and Program-Size for Functions Probability and Program-Size... between algorithmic probability and program-size for enumerating sets, in the case of the graphs... theory [1] deals with the program-size complexity and algorithmic probabilities for computing
http://w3.watson.ibm.com/~**chaitin**/fnc.html - Cached

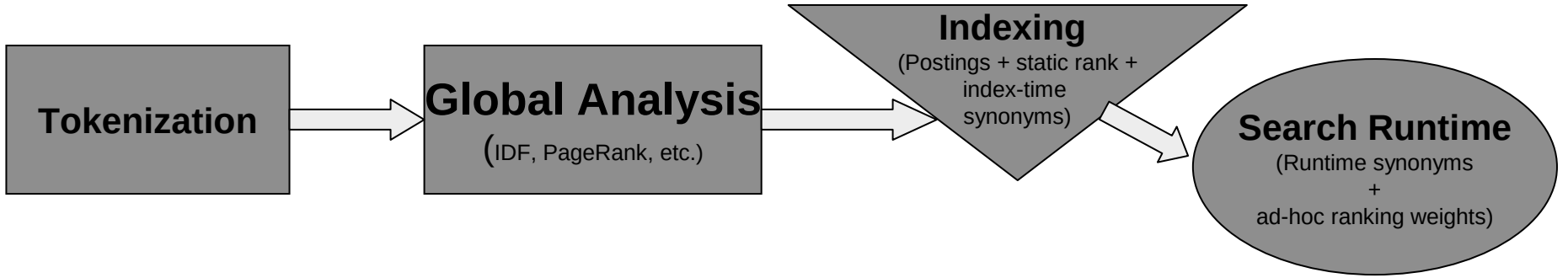
What does it take to get this result ?

- **Semantics-driven index term generation**
  - Having recognized a personal home page, we can generate "variants" that exploit the different ways in which person names are written out
    - G J Chaitin → GJ Chaitin
    - G J Chaitin → Chaitin, G J
    - ……

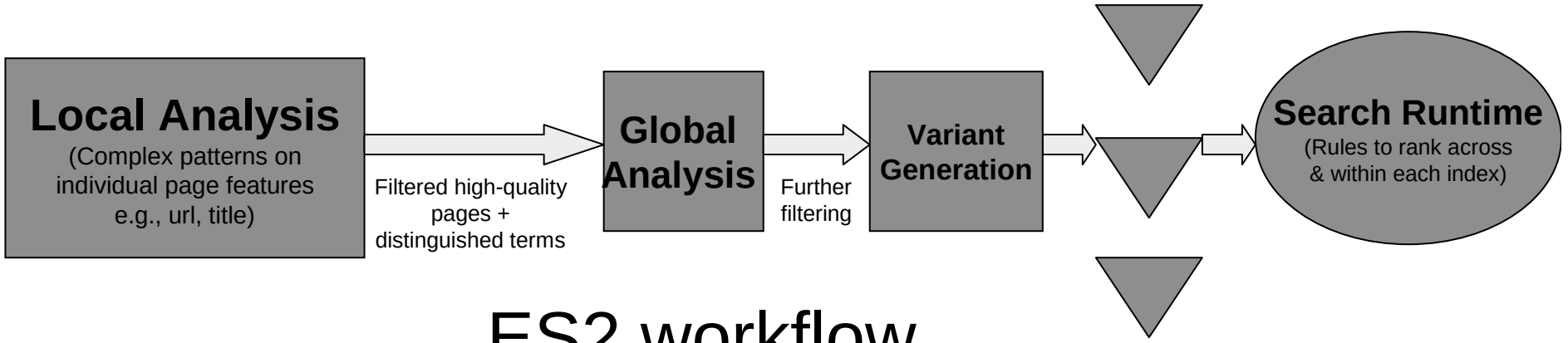
ES2 Analysis Workflow



Standard IR Workflow



LA is more sophisticated than Tokenization — *More complex site-root analysis* — *Multiple indices with custom term generation* — *Rule-driven ranking to exploit LA/GA/VG semantics*



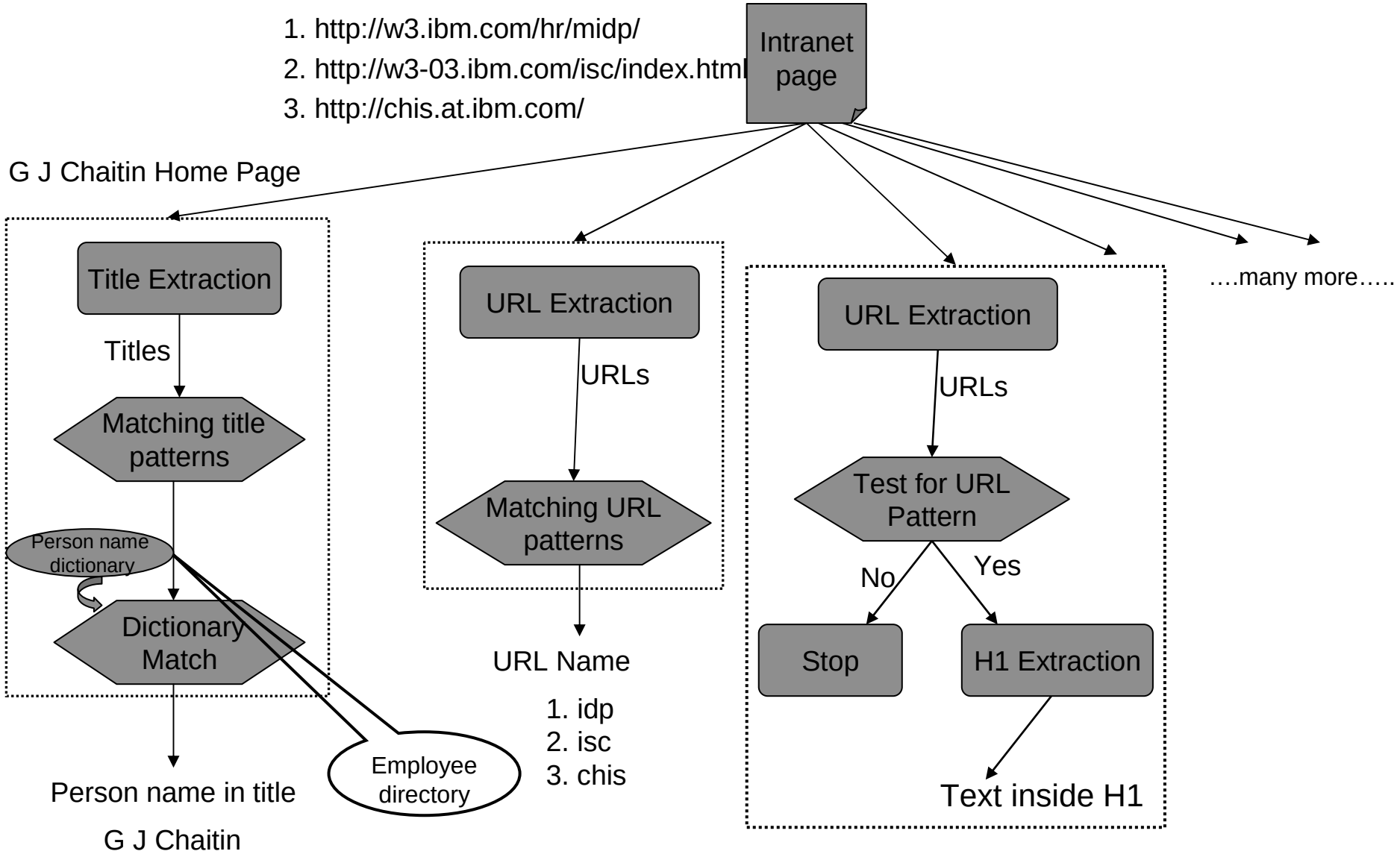
ES2 workflow

Local Analysis (LA)

- **Broadly, three types of analyses**
  - **Type 1**: Navigational page detection & navigational feature extraction
  - **Type 2**: Extraction of page-level attributes
    - E.g., identifying that a page is an IBM Standard Software Installer page
  - **Type 3:** Extraction to drive mining algorithms
    - E.g., acronym detection, feature extraction for input to geo classification, …
- **Mechanism**
  - LA involves complex rules consisting of
    - Carefully crafted regular expressions and dictionary matches against specific features of a page (e.g., META headers, title, URL, H1 and H2 tags, etc.)

Examples of Type 1 Local Analyses

1. http://w3.ibm.com/hr/midp/
2. http://w3-03.ibm.com/isc/index.html
3. http://chis.at.ibm.com/



Why Global Analysis ?

- **We have individual navigational pages and associated feature values e.g.,**

G J Chaitin → **http://w3.watson.ibm.com/~chaitin/index.html**

Can we put the feature values directly into the index !!

Unfortunately not so simple !!

All these pages have the title “G J Chaitin Home Page”

G J Chaitin Home Page - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://w3.watson.ibm.com/~chaitin/index.html

IBM Business Transfo... IBM Standard Softwa... IT Help Central Join World Communit... Windows Marketplace

"**Dieu a choisi celuy qui est... le plus simple en hypotheses et le plus riche en phenomenes**"
[God has chosen that which is the most simple in hypotheses and the most rich in phenomena]

"**Mais quand une regle est fort composée, ce qui luy est conforme, passe pour irrégulier**"
[But when a rule is extremely complex, that which conforms to it passes for random]

--- Leibniz, *Discours de métaphysique,* VI, 1686

[The *Discours* is also available online from Gallica; see pp. 32, 33 for the above texts.]

G J Chaitin Home Page

This website contains Greek letters and other mathematical symbols. If "Ω" isn't a capital Greek letter Omega, you should switch to another browser, for example, MS IE or Mozilla Firefox.

This website contains most of Chaitin's published papers, many book chapters, and the LISP, Java, C, and Mathematica software for Chaitin's Springer-Verlag trilogy. It also contains interviews and reviews of Chaitin's books.

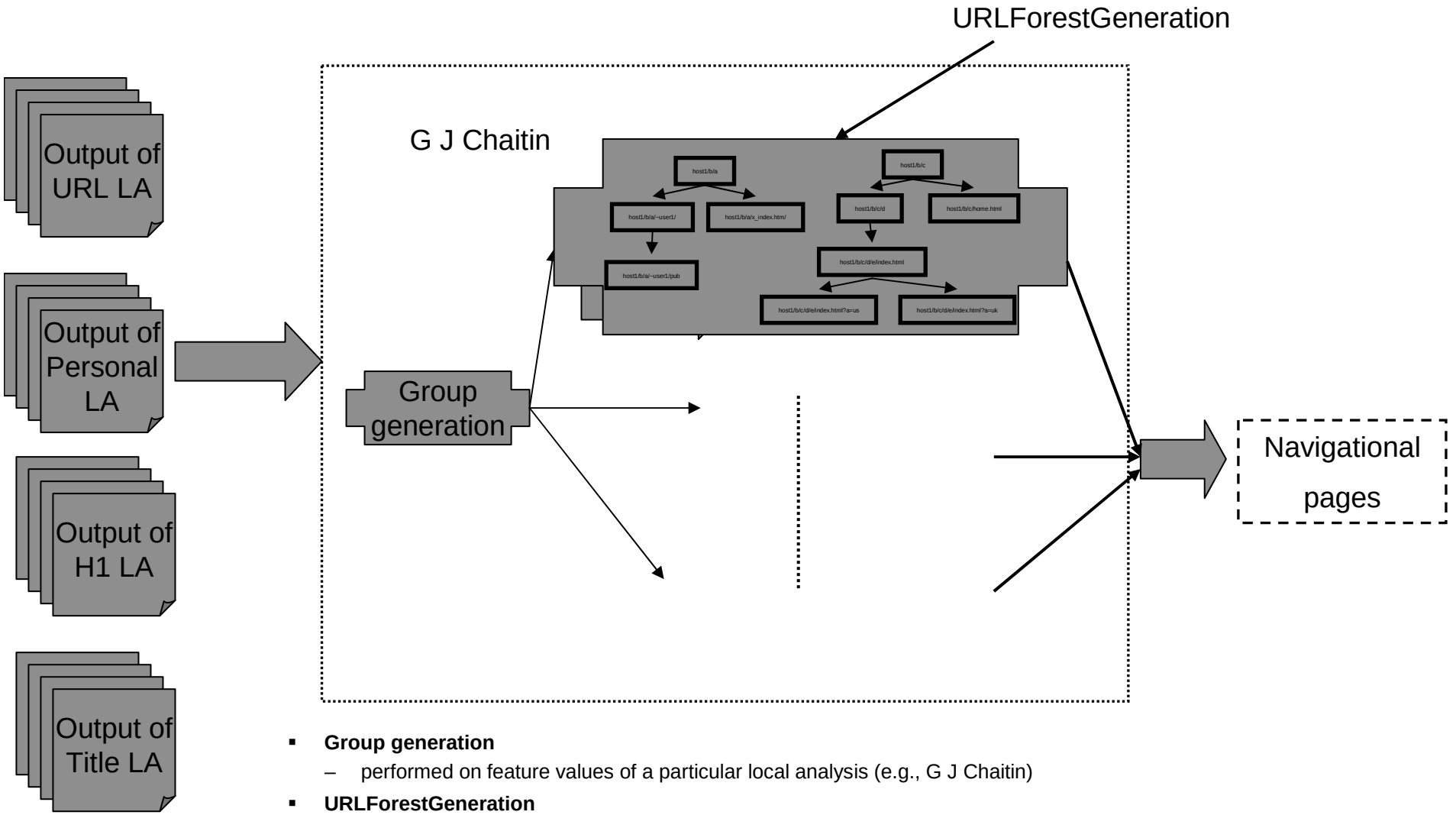
- **E-mail:** chaitin@us.ibm.com
- **Website:** http://www.umcs.maine.edu/~chaitin
- **Mirror:** http://www.cs.auckland.ac.nz/~chaitin
- **Phone:** 914/945-2785
- **Fax:** 914/945-4506
- **Mailing address:**
  IBM Research, P O Box 218
  Yorktown Heights, NY 10598, USA

Contents

- Latest News
- Recent Books
- LISP Applet
- Books with LISP Software
- Collections of Interviews
- Essays on Leibniz & Other Popular Articles
- Photo Album & Time-Line
- Brief Biography
- Latest Publications
- Complete List of Publications

Transferring data from w3.watson.ibm.com...

GA: Site Root Analysis (Zhu et al, WWW 2006)



- **Group generation**
  - performed on feature values of a particular local analysis (e.g., G J Chaitin)
- **URLForestGeneration**
  - URL hierarchy provides clues about navigational pages

Summary

- **Local Analysis**
  - Large number of complex rules involving a combination of dictionaries and regular expressions. We require
    - Mechanism to express and execute these rules
      - **SystemT & AQL** (http://www.alphaworks.ibm.com/tech/systemt) Declarative information extraction system
    - Ability to curate dictionaries
      - E.g., Person names (tap the corporate directory)
    - Complex regular expression learning (given below)
- **Mining**
  - Automatically extract acronyms and their expansions
  - Geo classification
  - Learning regular expressions for use in LA

Outline of the Talk

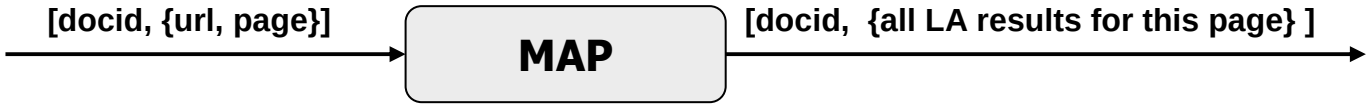
- **Motivating search queries**
- **Analytics and Mining in ES2**
  - Overall workflow
  - Local Analysis (LA)
  - Global Analysis (GA)
- **Migration to Hadoop** 
  - LA & GA on Hadoop
  - Mapping analysis/mining algorithms onto MapReduce

Challenges in migrating to Hadoop

- **Algorithmic challenges**
  - Designing efficient MapReduce versions of the analysis and mining algorithms
    - In some cases straightforward mappings work!!
    - In other cases, smart mappings to MapReduce have to be designed
- **Data management challenges**
  - Representing and manipulating the data associated with the analysis steps

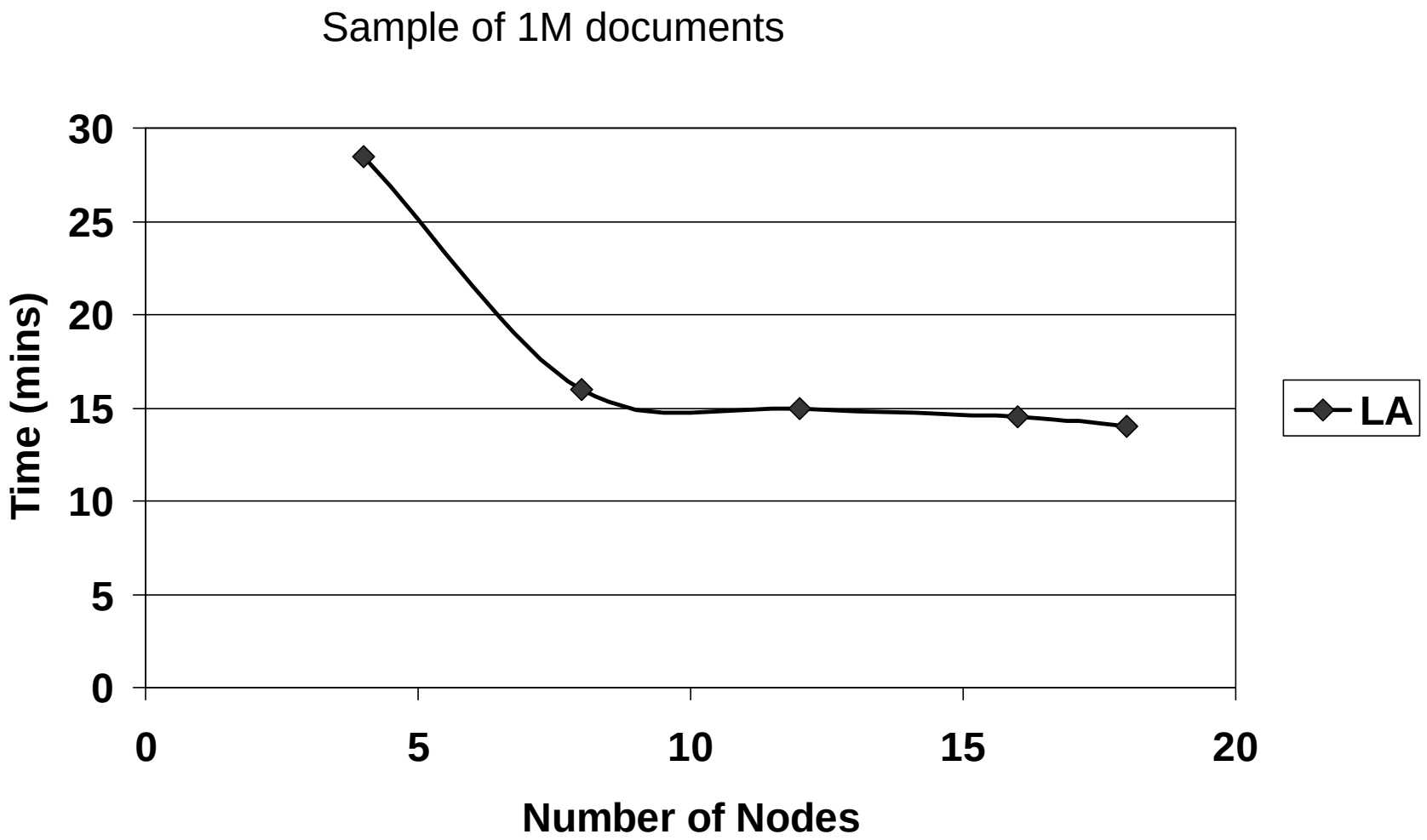
LA on Hadoop

- **By definition,**
  - Page-at-a-time analysis ➔ easily parallelizable
- **MapReduce instantiation**
  - A MAP-only job where LA analysis is performed inside the mapper one document at at time
  - Conceptually,

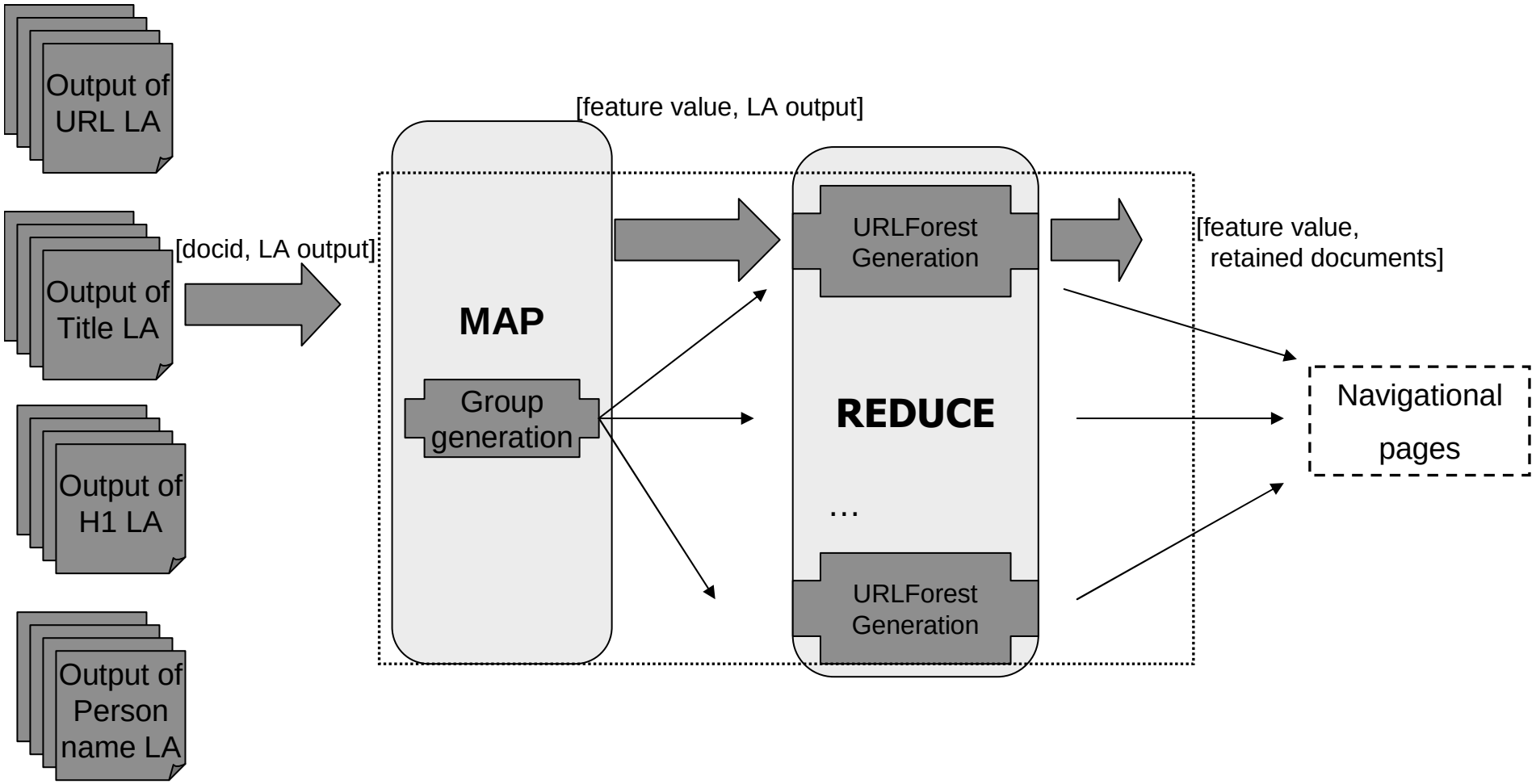


Expect to scale linearly with the number of nodes in the cluster

LA on Hadoop



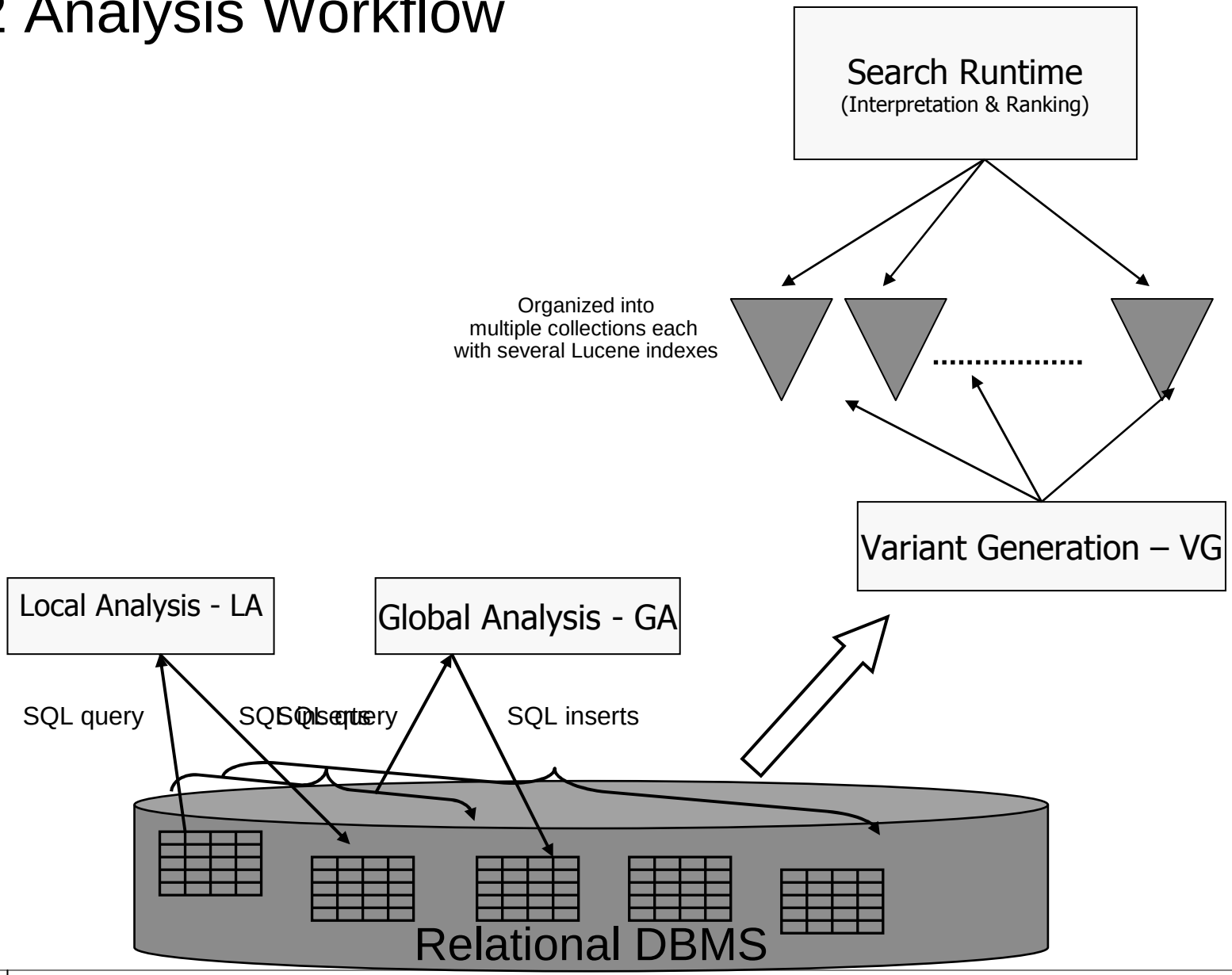
GA on Hadoop



LA & GA on Hadoop

- **Algorithms appear to map directly to MapReduce**
- **Remaining challenge**
  - Deal with the data management problems

ES2 Analysis Workflow



To migrate LA & GA on Hadoop

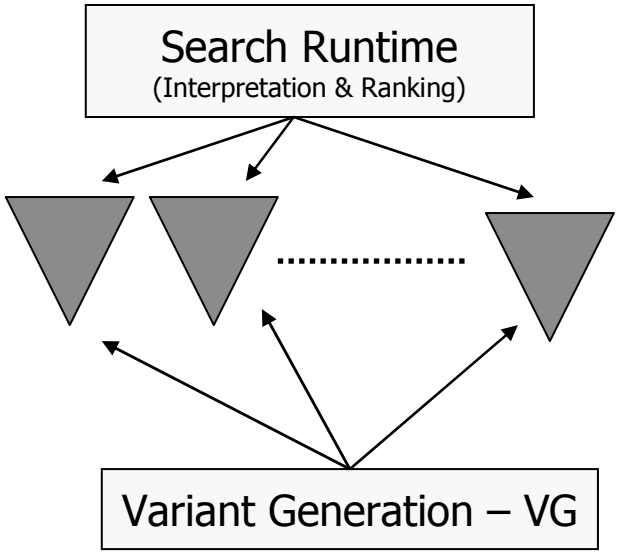
- Either, we invent
  - Data formats to represent LA output and other intermediate results in Hadoop
  - Write custom data manipulation operations in Java
- Instead, we can use a query language
  - JAQL (http://code.google.com/p/jaql)
    - Uses JSON as the data model
    - Designed to process massive quantities of semi-structured data
    - Exploit map-reduce for parallelism
    - Easily extend by plugging in user-defined functions

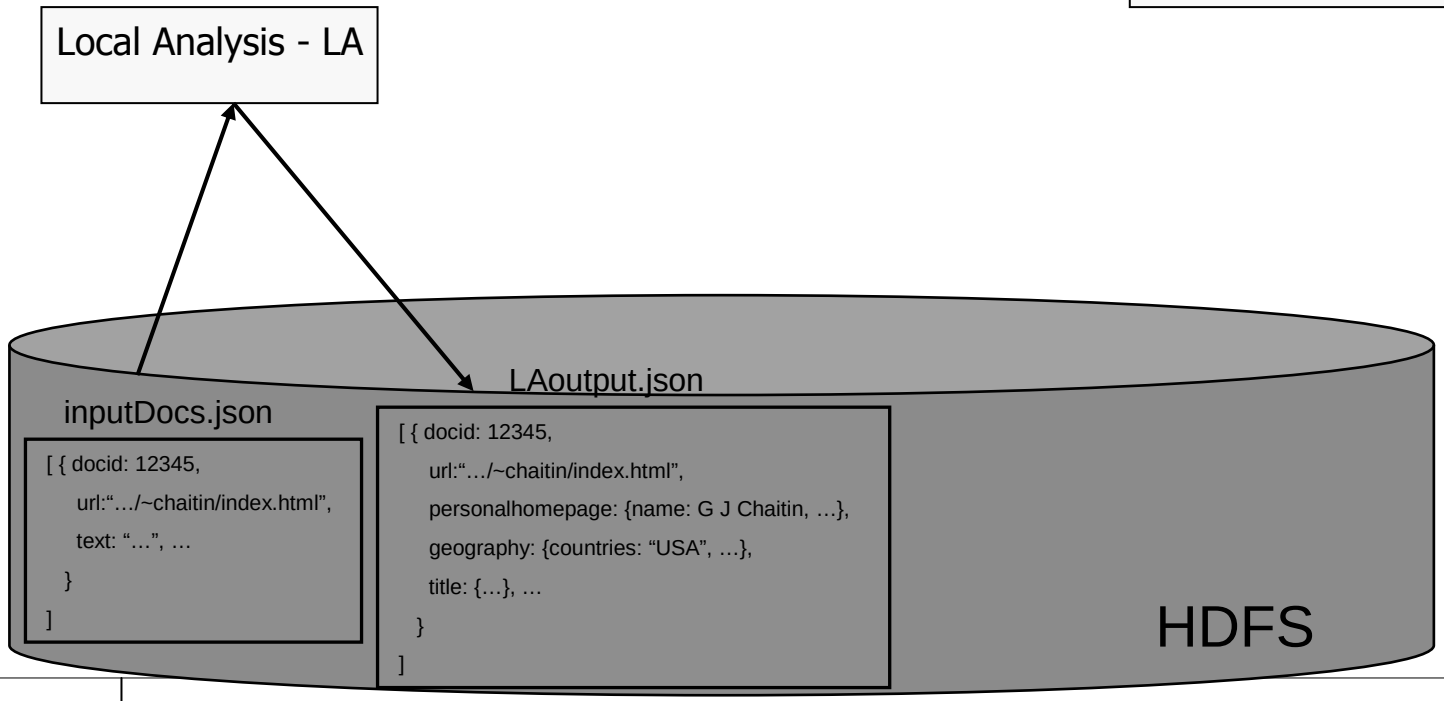
ES2 Analysis Workflow

JAQL LA query

```
$alldocs    = file 'inputDocs.json;
$results    = file 'LAoutput.json';

$alldocs
→ map LocalAnalysis($.docid, $.text, $.url)
→ write $results;
```

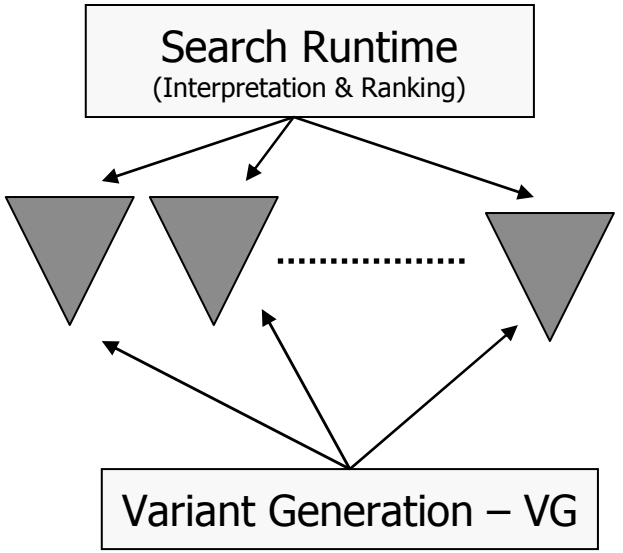


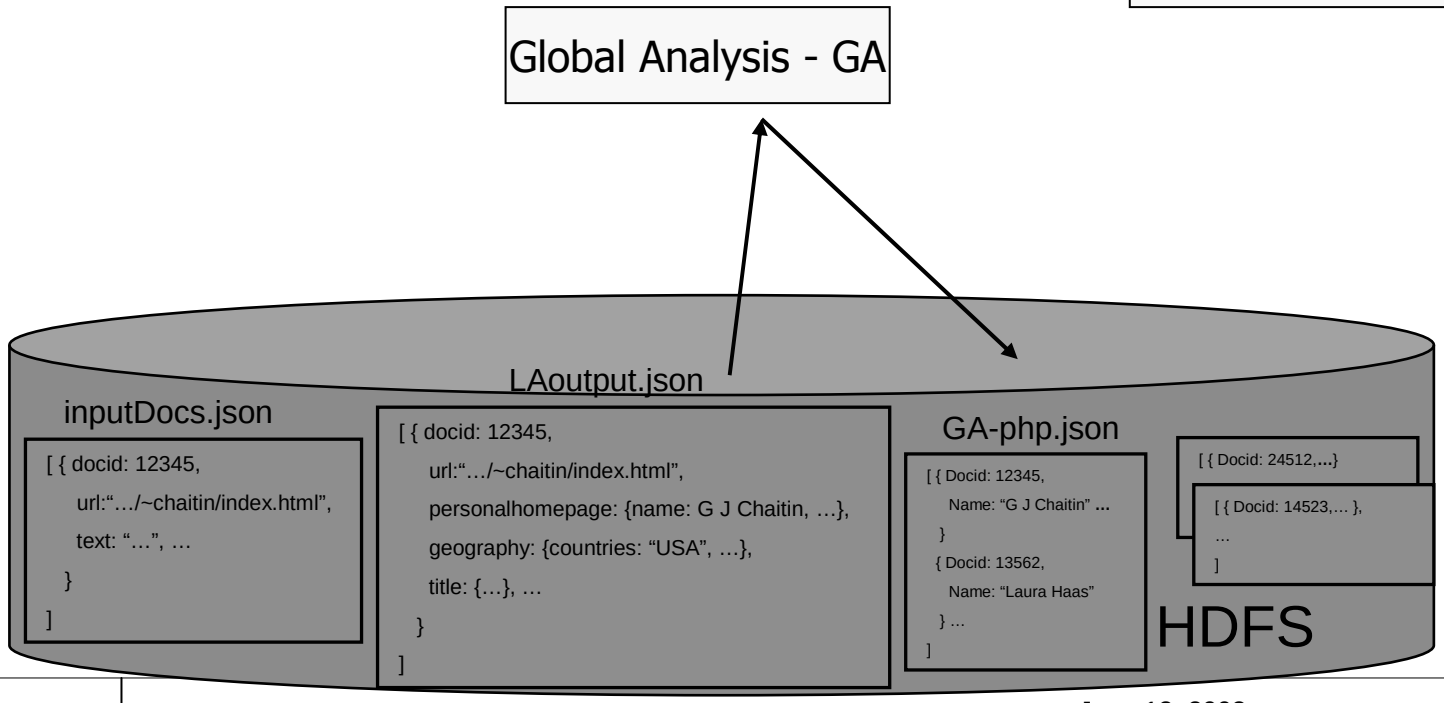


ES2 Analysis Workflow

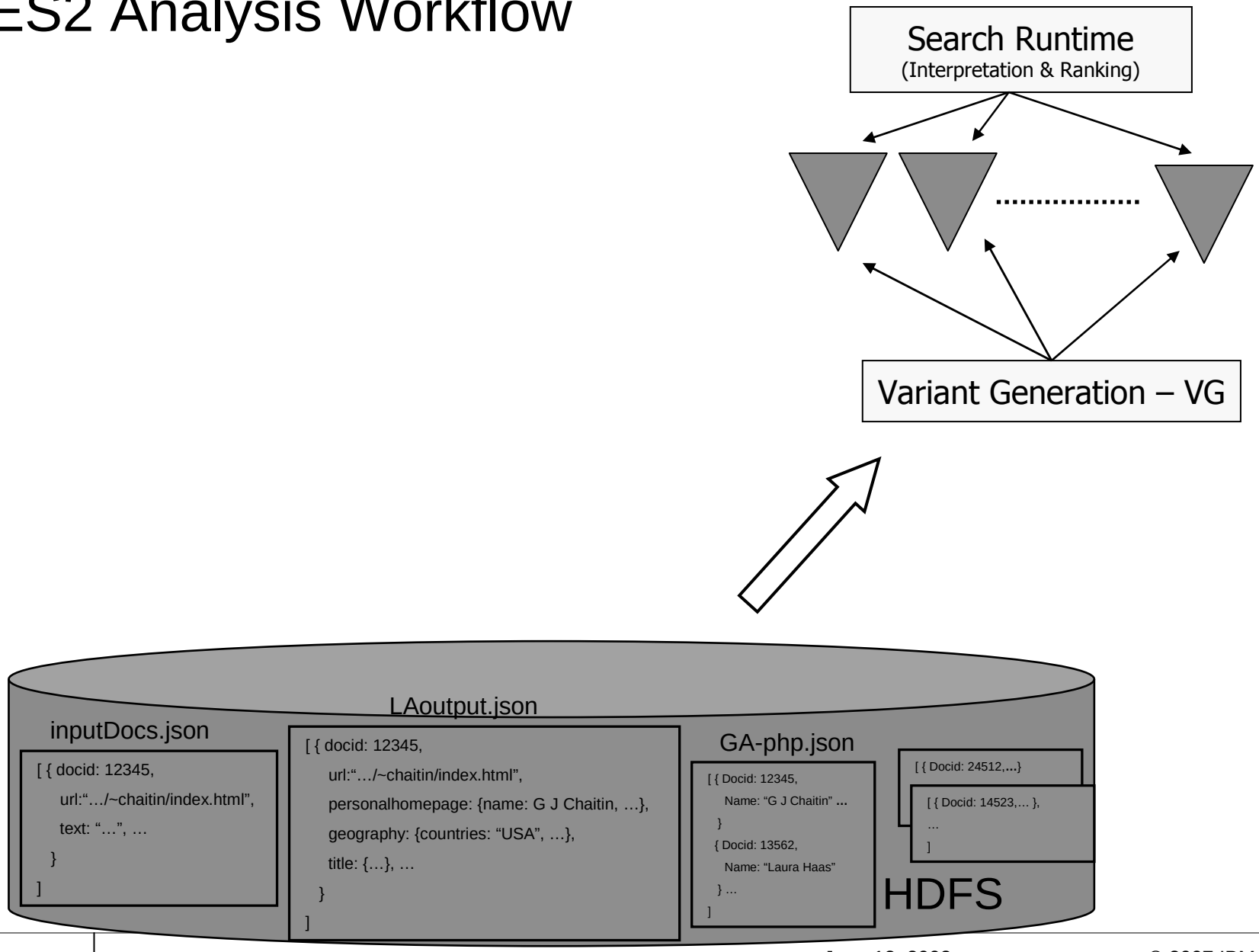
Example JAQL GA query

```
$ alldocs        = file 'inputDocs.json';
$results         = file 'GA-php.json';
$alldocs
→ filter  not(isnull($i.personalhomepage.NAME))
→ partition by $t = $.personalhomepage.NAME
     |-  UrlForestGeneration($t, $);
→ write $results;
```





ES2 Analysis Workflow



Outline of the Talk

- **Motivating search queries**
- **Analytics and Mining in ES2**
  - Overall workflow
  - Local Analysis (LA)
  - Global Analysis (GA)
  - Variant Generation (VG)
- **Migration to Hadoop**
  - LA & GA on Hadoop
  - Mapping analysis/mining algorithms onto MapReduce 
  - Performance results

Three Tasks

- **Acronym extraction** 
- **Geo classification**
  - Frequent item set mining
- **Learning regular expressions**

Acronym variants examples

- **ACM**
  - Total 32 long form variants.
  - After case-folding:
    - Accessibility Configuration Manager
    - Accommodation System
    - Accumulated Call Meter
    - AFE Contract Management
    - Application Component Manager
    - Area Calculation Method
    - Asbestos Containing Material
    - Association for Computing Machinery
    - Asynchronous Communication Monitor
    - Atlas Call Management

- **PBC**
  - Total 60 long form variants.
  - After case-folding:
    - People's Bank of China
    - Performance Business Commitments
    - Personal Business Comitments
    - Personal Business Comittments
    - Personal Business Commitement
    - Personal Business Commitements
    - Personal Business Commitment
    - Personal Business Commitments
    - Personal Business Committment
    - Personal Business Committments
    - Personal Business Contribution
    - Personal Business Controls
    - Personal Business Objectives
    - Personnal Business Commitments
    - Personnel Business Commitment
    - Personnel Business Commitments
    - Pesonal Business Commitments
    - Plant Biotechnology Centre

Acronym Extraction: Algorithm

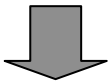
1. **Scan document for patterns that match:**
   - *longForm* ‘(‘ *shortForm* ‘)’ OR *shortForm* ‘(‘ *longForm* ‘)’
   - Example: *… in recent develpments,* *Bank of America (BofA)…*
2. **For each match apply following heuristics to check if this is a good candidate:**
   - Match characters in *shortForm* to characters in *longForm* starting from right to left. *Bank of America (BofA)*
   - If no match is found for some character in the *shortForm*, return null.
   - Find beginning of first word in *longForm* to match first letter in *shortForm* *Bank of America (BofA)*
3. **Resolve variations in different *longForms* for a given *shortForm***
4. **Aggregate counts for each [*shortForm*, *longForm*]**

Ref: Ariel Schwartz and Marti Hearst, “*A Simple Algorithm for Identifying Abbreviation Definitions In Biomedical Text*”, PSB 2003

Acronym Extraction MapReduce Implementation

MAP: EXTRACT CANDIDATES

1. **Scan document for patterns that match:**
   - *longForm* '(' *shortForm* ')' OR *shortForm* '(' *longForm* ')'
   - Example: *… in recent develpments, Bank of America (BofA)…*
2. **For each match apply following heuristics to check if this is a good candidate:**
   - Match characters in *shortForm* to characters in *longForm* starting from right to left.
   - If no match is found for some character in the *shortForm*, return null.
   - Find beginning of first word in *longForm* to match first letter in *shortForm*

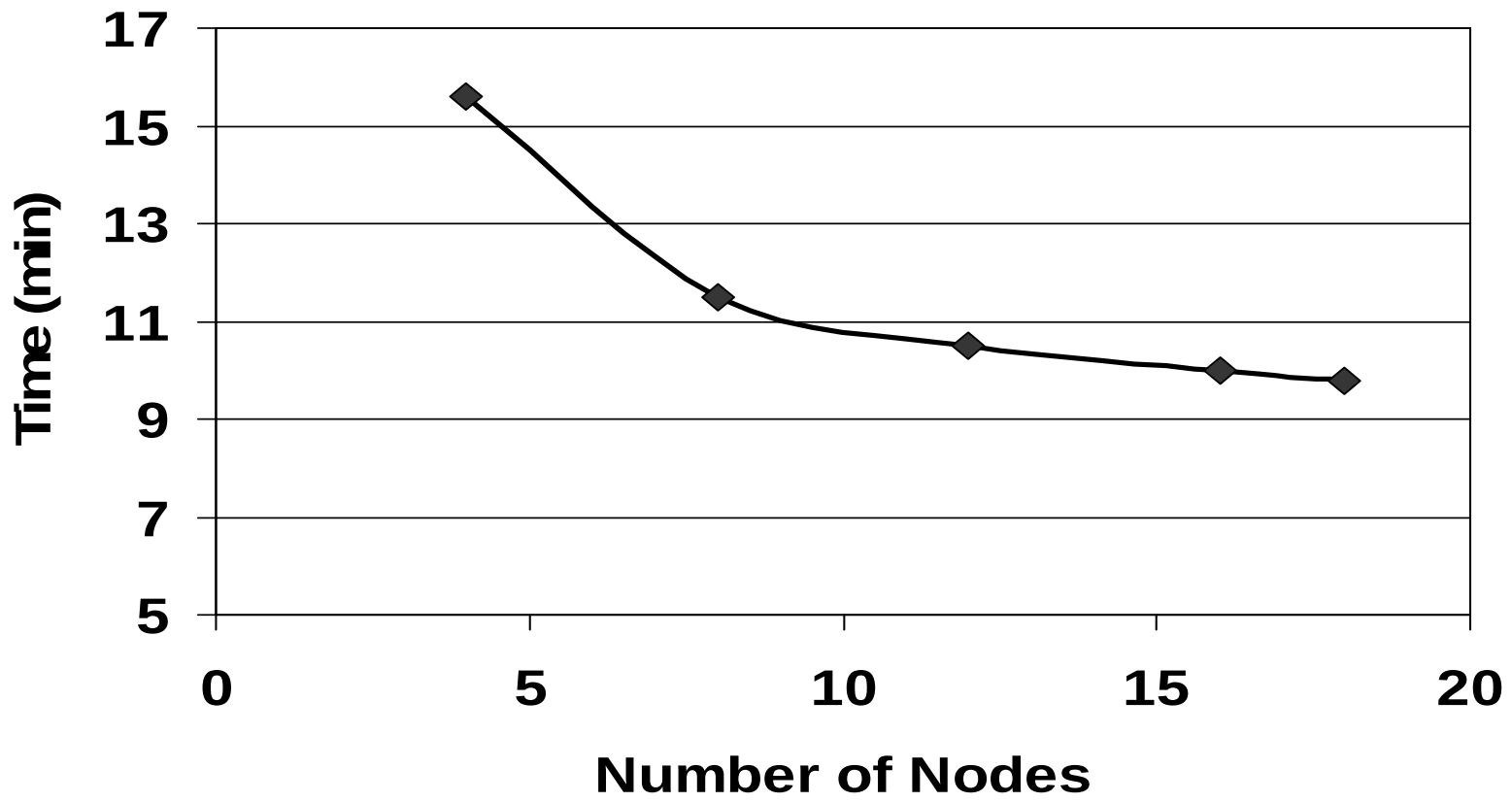


[shortForm, longForm]

REDUCE: RESOLVE & COUNT

3. **Resolve variations in different *longForms* for a given *shortForm***
4. **Aggregate counts for each [*shortForm*, *longForm*]**

Acronym Extraction on Hadoop



Lessons being learned and more to come …

- **Mining (such as acronym extraction) operate as a single task thereby using Hadoop more efficiently**
- **LA runs each crawler segment as a task (same data multiple tasks) → does not scale as well**
  - Speculatively execution may use resources for existing job instead of devoting resources to a new job. (*As of v16*)
- **Dependency on number of data-nodes versus processing nodes**
- **Problems with asymmetry:**
  - Slower nodes may be serious bottlenecks especially if a larger partition gets assigned to one

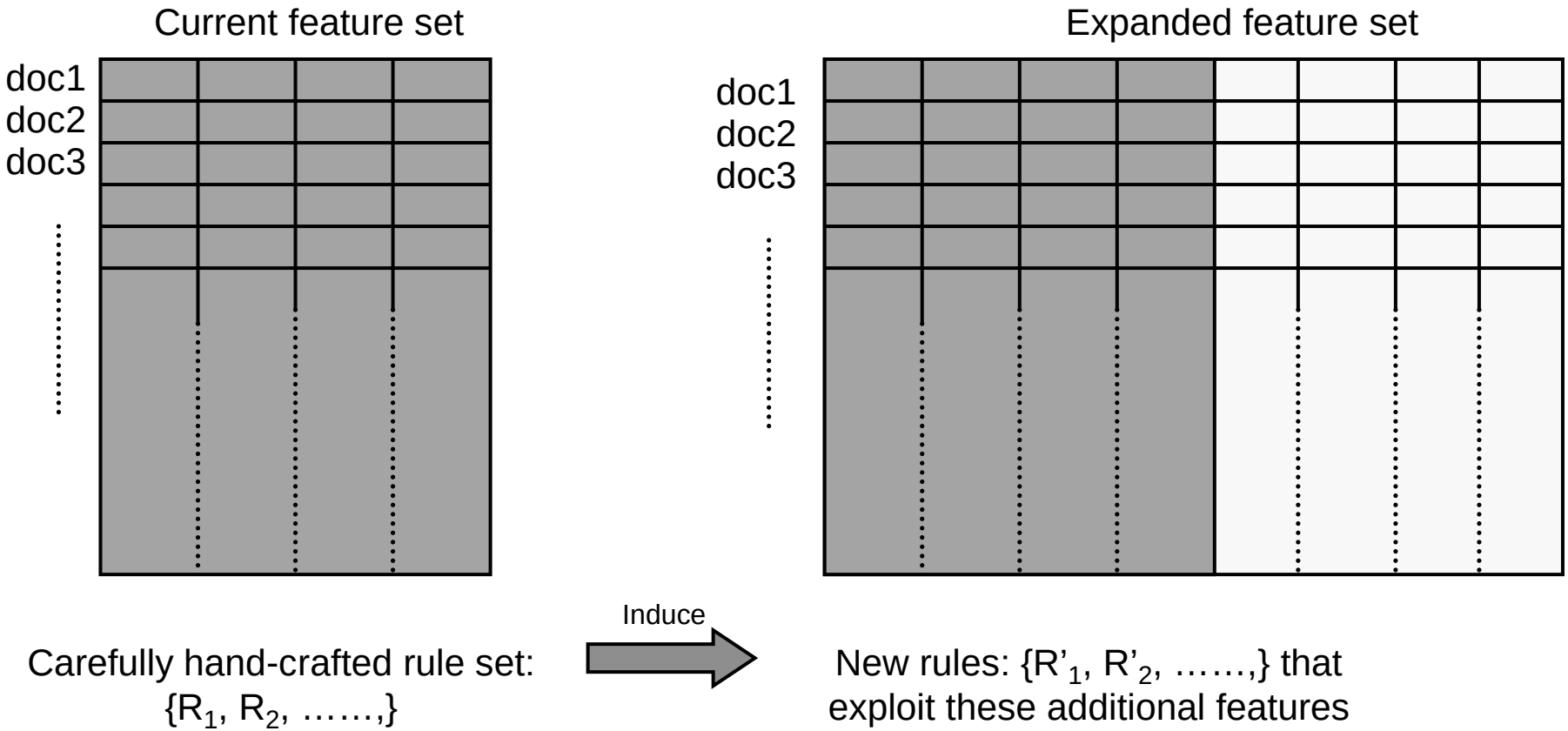
Three Tasks

- **Acronym extraction**
- **Geo classification** 
  - Frequent item set mining
- **Learning regular expressions**

Geo classification

- **Goal**
  - With each page, associate the IBM location/country/region for which it is most relevant
    - Simple approaches don't work (e.g., cannot assume all pages hosted at "foo.bar.de" are for "German employees")
- **Current approach**
  - Uses a limited number of features extracted during LA
    - By matching dictionaries of country, region, and location names with the title, URL, meta headers, etc.
  - Manually specified rules (that use a subset of features) to classify
- **Problem**
  - Tuned for **precision** but suffers from poor recall (less than 1/4th of the pages are labeled)

Improving recall for Geo classification



Leads to the following sub-problem: **Finding association rules over the geo features**
In turn, this requires: **Computing frequent item-sets involving geo features**

Scalable Frequent Item-set Computation

- **Challenges**
  - Very sparse date set in a feature space of several hundred
  - Minimum support for item-sets must be set very low
- **Work done this summer with Fei Chen @ Univ. of Wisconsin, Madison**
  - Developed and implemented a scalable frequent item-set algorithm on Hadoop
  - Series of algorithms from naïve to progressively more sophisticated

Performance problems with a Priori

- **Performance**
  - On our earlier small dataset
    - 36368 rows
    - 20 columns
    - Minimum support = 0.1%

    this implementation takes over 1 hour
- **Problems**
  - # scans of entire data set = # iterations = Length of longest freq. itemset
  - Because of the low-support requirements,
    - Large number of potential candidates
    - Negatively impacts the performance of the “generate and test” approach used in a priori

FPGrowth Algorithm (Han et. al., SIGMOD 2000)

- **Based on the following key ideas**
  - A pattern growth approach as opposed to candidate generation
  - A recursive divide-and-conquer method to decompose the mining task into a set of smaller mining tasks on sub-databases called projected databases
- **Mapping to Hadoop (FPGrowth-H)**
  - Design MapReduce jobs to repeatedly project the input data set until each projection can be completely mined in memory using the FP-tree structure

Experiments

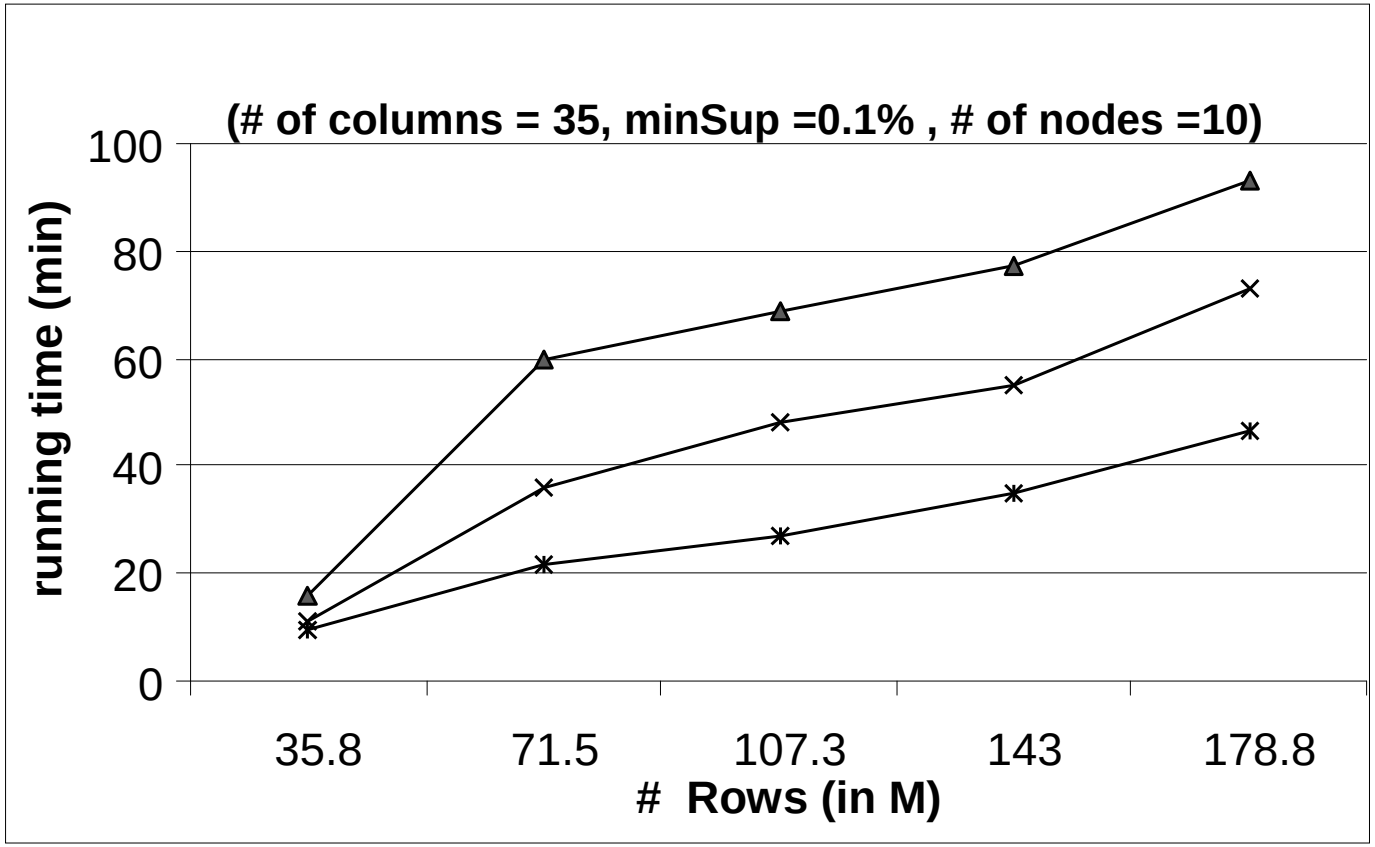
- **Small data set**
  - 36368 rows, 20 columns, s_min = 0.1%,

Naive	a priori	FPGrowth-H
> 2.5 hrs	**> 1 hr**	**25s**

On a 10-node Hadoop cluster

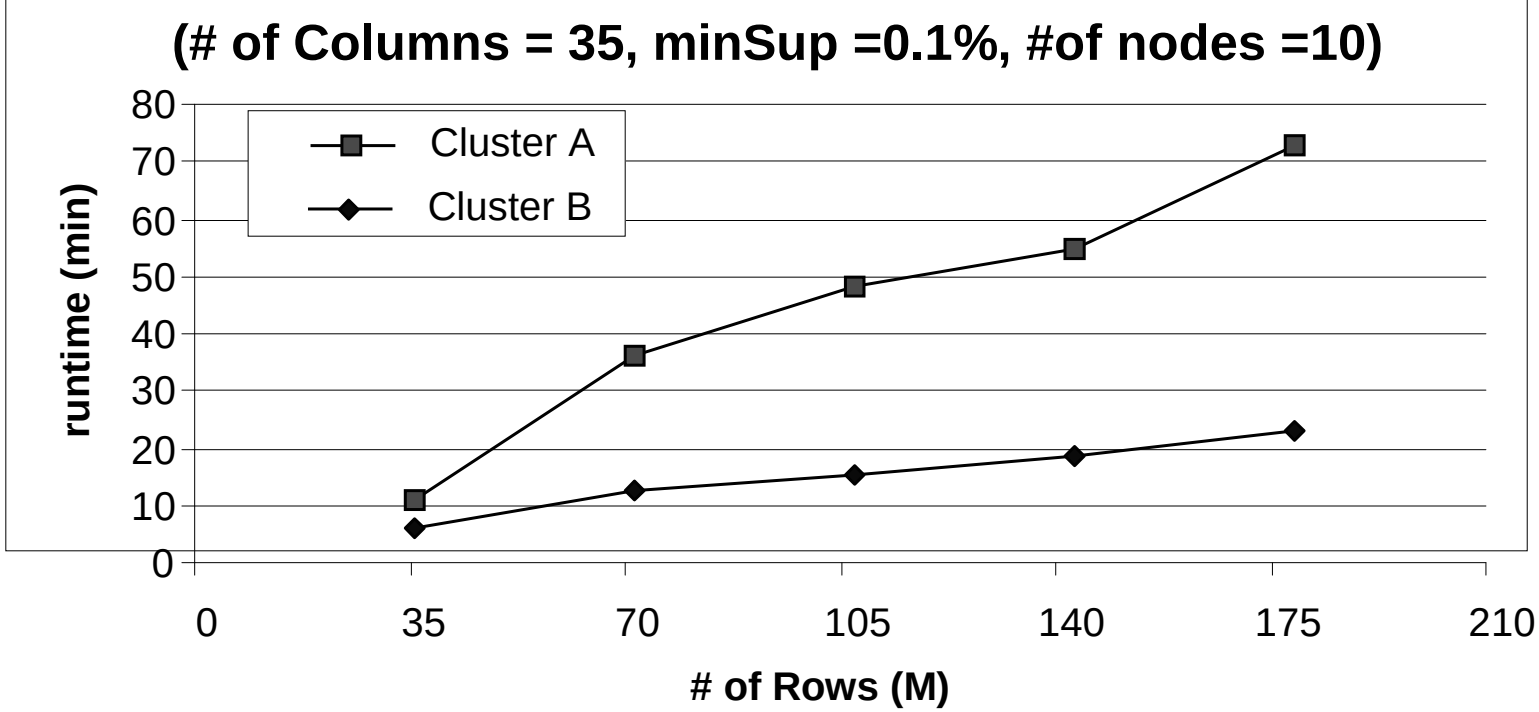
  - We see the impact of small min-support even on small data sets

FPGrowth-H Performance on larger data sets (up to 10G)



FPGrowth-H Performance on larger data sets (up to 50G)

	Cluster A	Cluster B
Vendor	**Intel XEON**	**AMD Opteron**
Cores	**2 X 1**	**2 X 4**
Clock Rate	**1000 MHz**	**1050 MHz**
Memory	**4 GB**	**16 GB**



Comparisons with prior work

- **In general, hard to compare numbers**
  - Variations depends on the hardware, architectures, min_sup, data distribution, etc.
  - Not all of these parameters are reported in literature
  - Also, not sure whether prior work pushed min_sup to as low as 0.1%

- **However, one comparison point**
  - Osmar R. Zaane et. al.: Fast Parallel Association Rule Mining Without Candidacy Generation
    - Shared memory *SGI Origin 2400* with 64 processors, 50M rows, average 12 columns, and min_sup is not reported
    - Runtime ~ 63 mins (3831 seconds)
  - FPGrowth-H
    - 10 cluster, 70M rows, 35 columns, min_sup = 0.1%
    - Runtime ~ 39 min

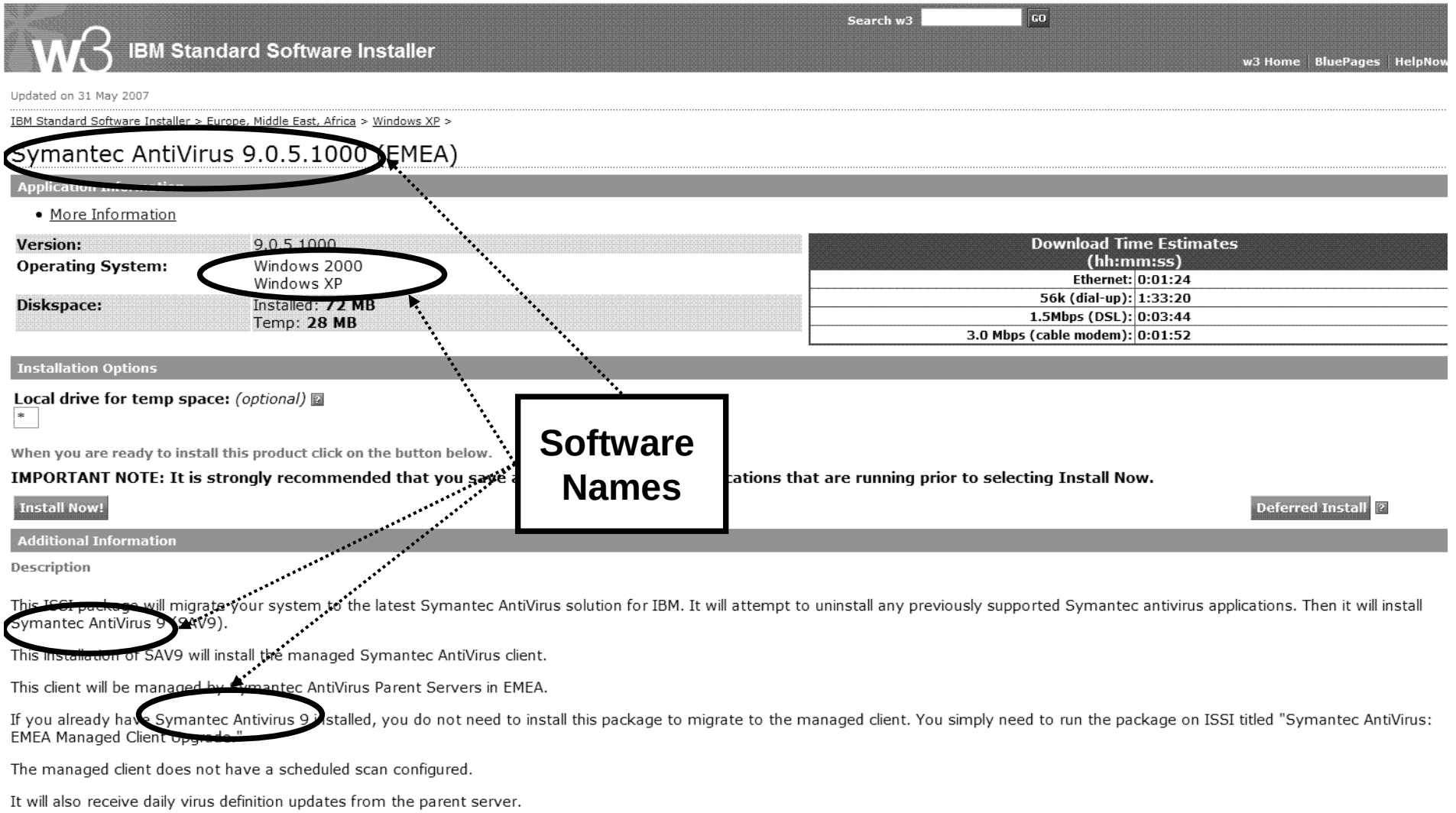
Lessons Learned

- **Combiners versus custom aggregation**
  - Lack of flexibility in controlling when and how the Combiner will be called (Disclaimer: as of Hadoop v17.1)
    - Hadoop performs local sort-merge before calling the combiner
    - The combiner is triggered whenever Hadoop decides the output buffer pool is full
  - We obtained better performance by using custom aggregation code within the mapper
- **Lack of flexibility in scheduling MapReduce jobs**
  - Potentially better performance by streaming data directly from one MapReduce job to another
    - E.g., Begin “Single Frequent Item” even as output is being generated from Reducer of “Project-and-Mine”
  - Likely to be true for most iterative mining algorithms

Three Tasks

- **Acronym extraction**
- **Geo classification**
  - Frequent item set mining
- **Learning regular expressions** 

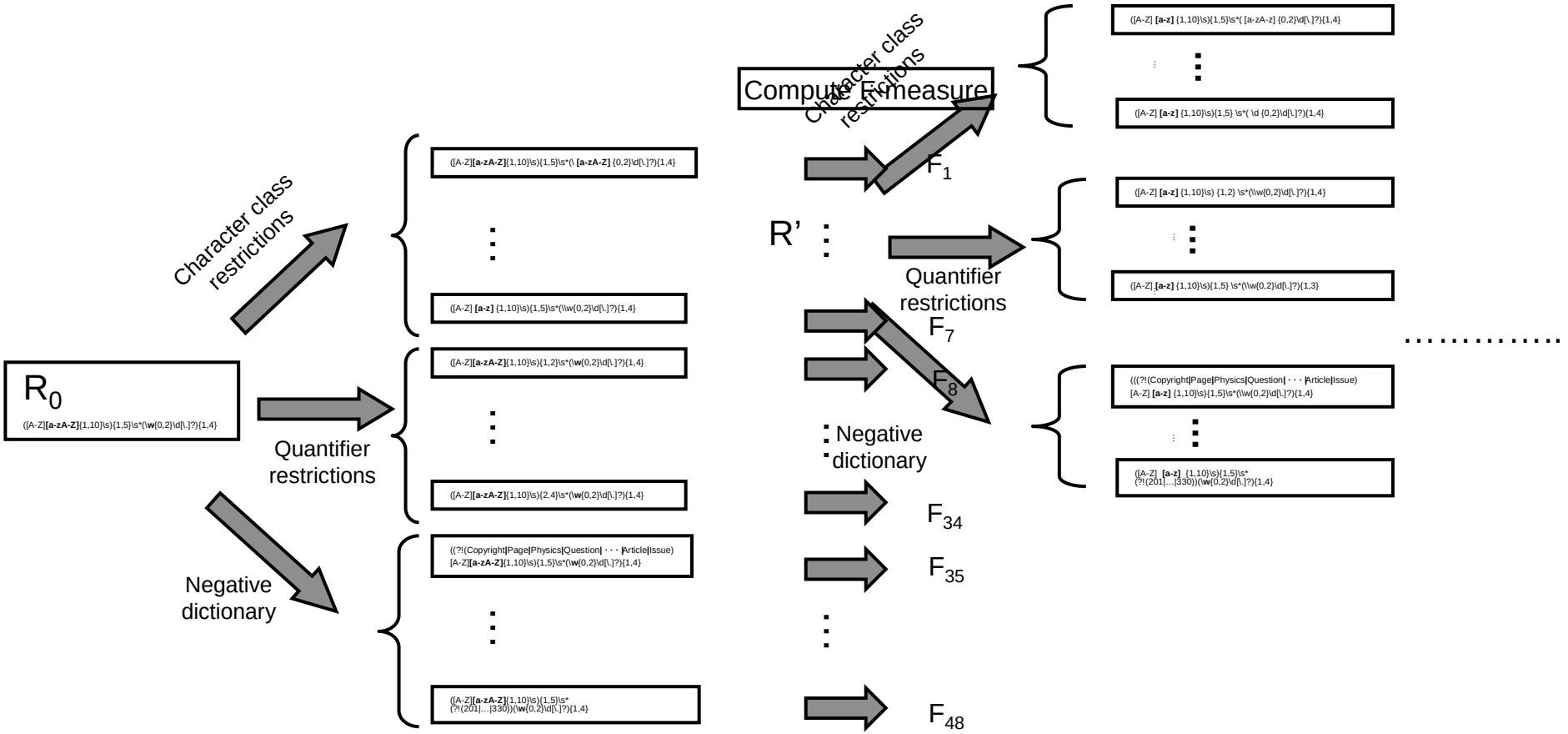
Extracting Software Names, Yunyao Li et al, SIGIR 2006



ReLIE Intuition …

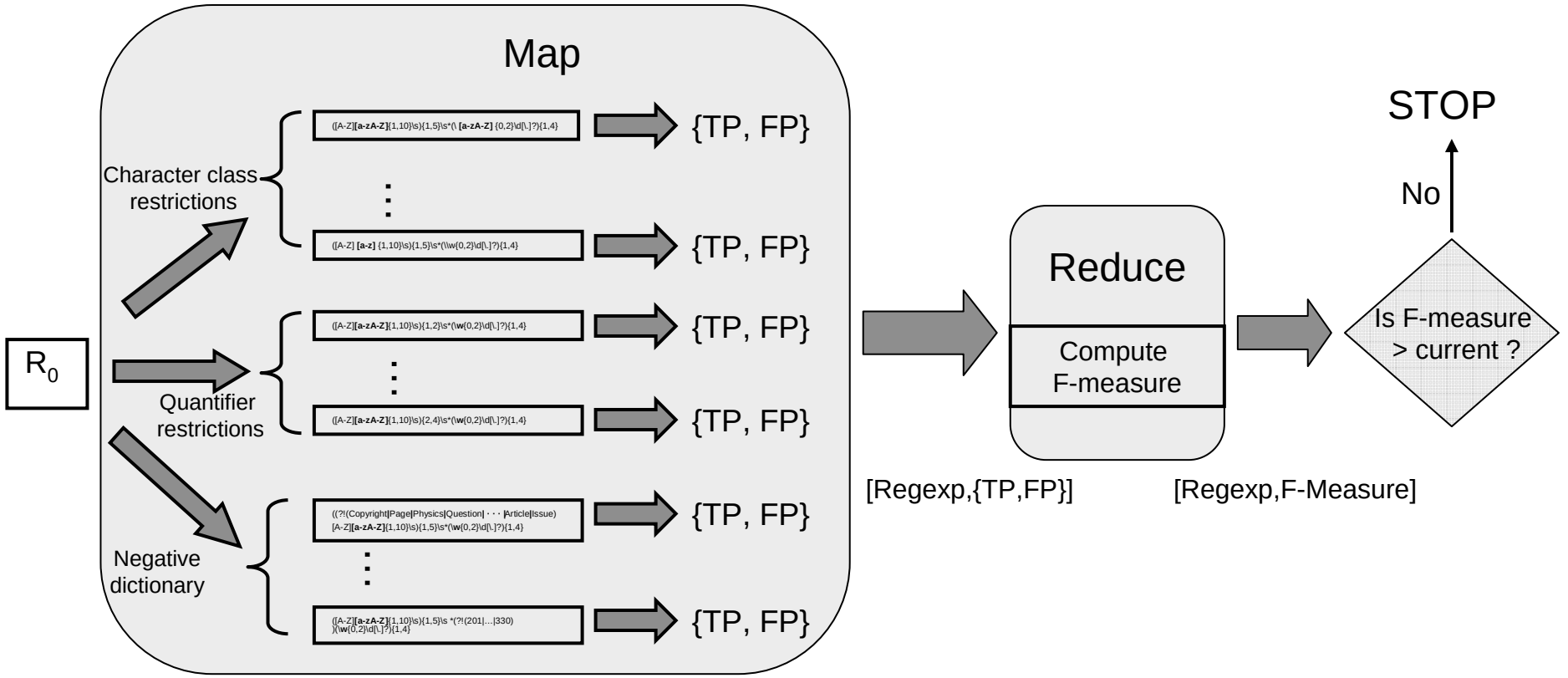
- **Start with an initial regular expression R_0**
  - \b([A-Z]**[a-zA-Z]**{1,10}\s)**{1,5}**\s*(**w**{0,2}\d[\.]?){1,4}\b
  - Identifies correct instances
    - Norton Antivirus 5.03.61, Windows 2003, Eclipse 3.2
  - Matches False positives
    - Physics 201, Room 330, Chapter 2.2
- **Modified to obtain R_{final} with higher precision**
  - \b(**(?!(Copyright|Page|Physics|Question|** · · · **|Article|Issue))**[A-Z] **[a-z]** {1,10})\s)**{1,4}** \s*(**[a-zA-Z]**{0,2}\d[\.]?){1,4}\b
- **$R_{improved}$ obtained by making several local transformations**
  - Character class restrictions [a-zA-Z] → [a-z]
  - Quantifier restrictions {1,5} → {1,4}
  - Negative Dictionary of terms {Copyright, Page, Physics, …}

Regular Expression Learning Algorithm Yunyao Li et al, EMNLP 2008



- Generate candidate regular expressions by applying a single transformation
- Select the “best candidate” R’ based on F-measure on training corpus
- If R’ has better F-measure than current regular expression, repeat the process

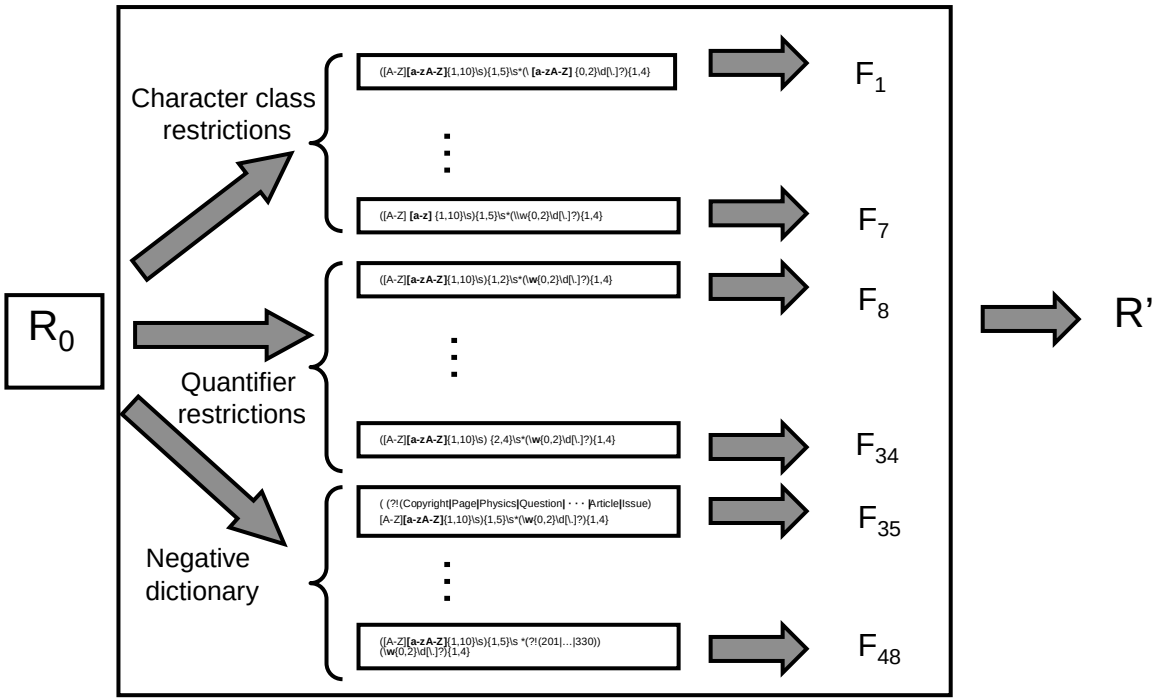
Mapping if we have a large training corpus (“gedanken experiment”)



Most expensive part of an iteration is computing the F-measure for all candidates

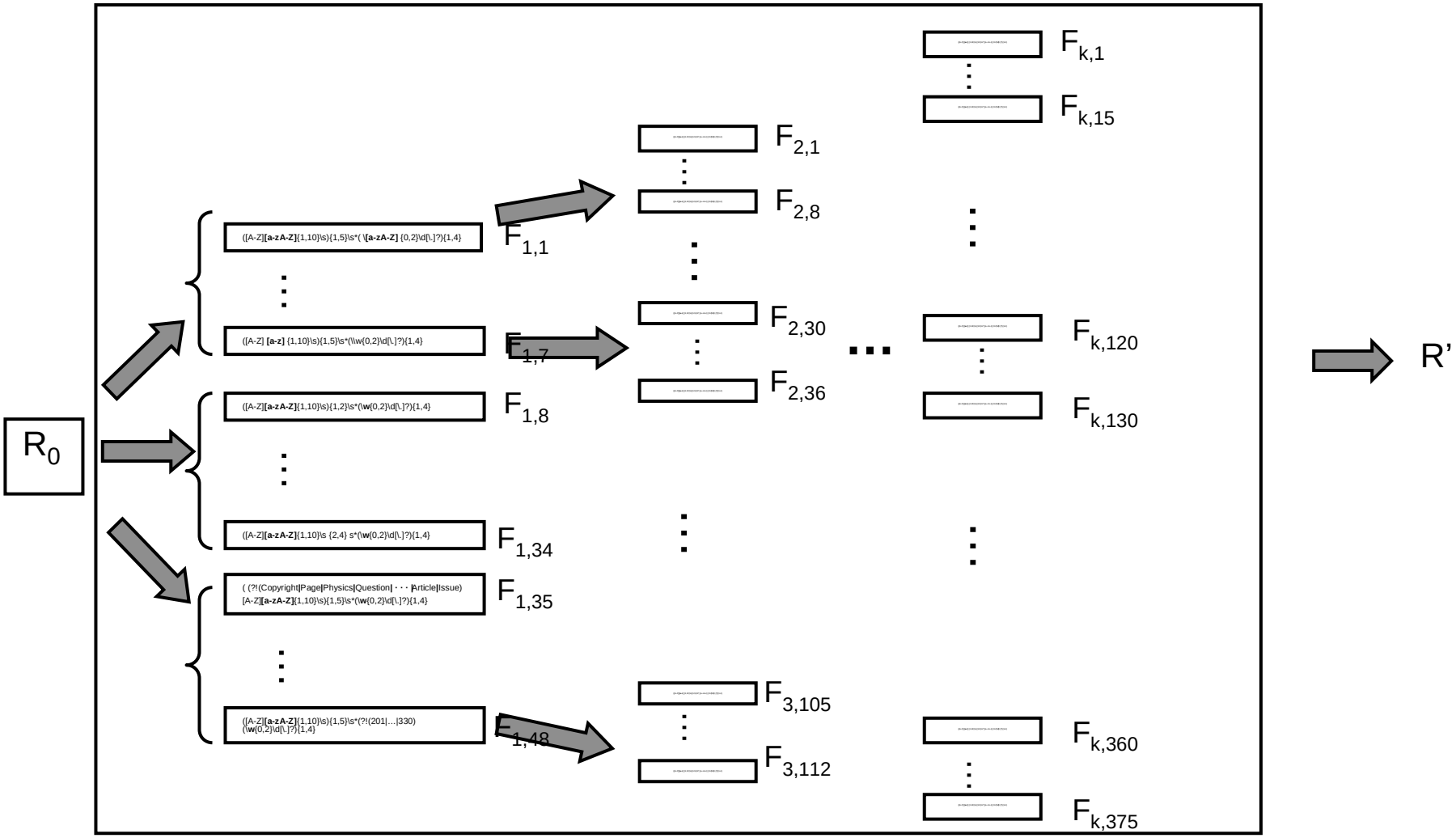
We can translate this computation into a map-reduce job by partitioning the document corpus

Training corpus is usually small : Exploring larger portion of the search space

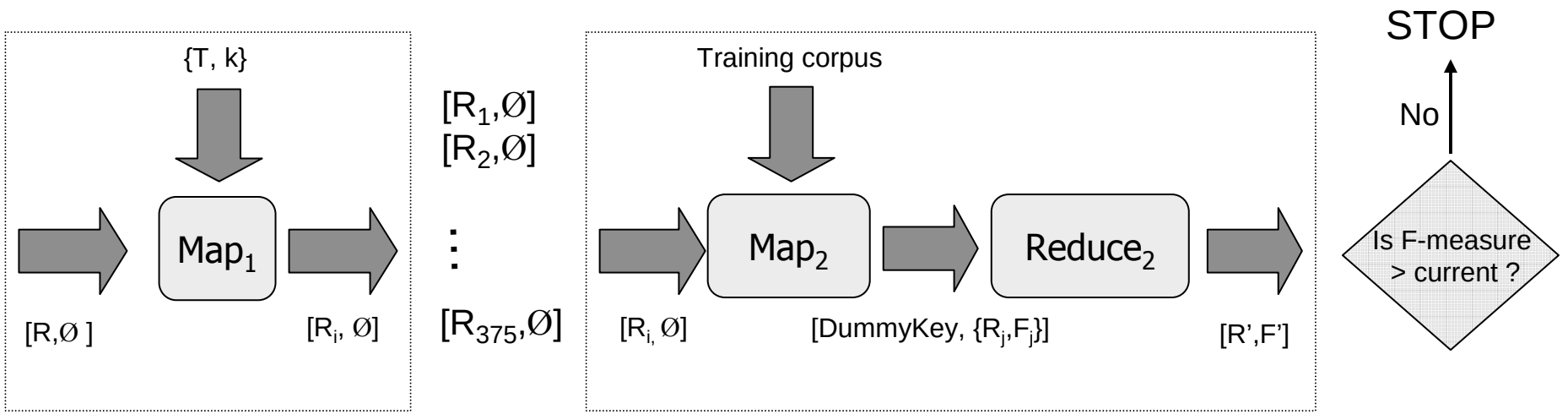


ReLIE uses a greedy heuristic where candidates are “1-transformation away” from R0 !

Instead, generate k-transformation neighborhood …



Expanding the search space : Parallelizing the F-measure computation of candidates



Map :	generate all candidates by applying k-transformation to R
Reduce :	none

Map :	for each candidate, compute F-measure over training corpus
Reduce :	choose "best" candidate

IBM/Google Cloud Academic Initiative

- One year anniversary – launched 10/2007
- Fostering next-generation learning environment with flexible yet powerful virtual IT infrastructure
- Allow users to tap into massive computing resources not previously available
- One of the largest production clouds in existence (1100+ servers across three locations)
- 12 universities participating in the program, 25 by end of 2008
- Over 700 students & researchers served to date
- Promoting advanced research & learning activities
- Supported by the National Science Foundation
- Promoting open-source software on Linux using Eclipse plug-in

Thank you !

