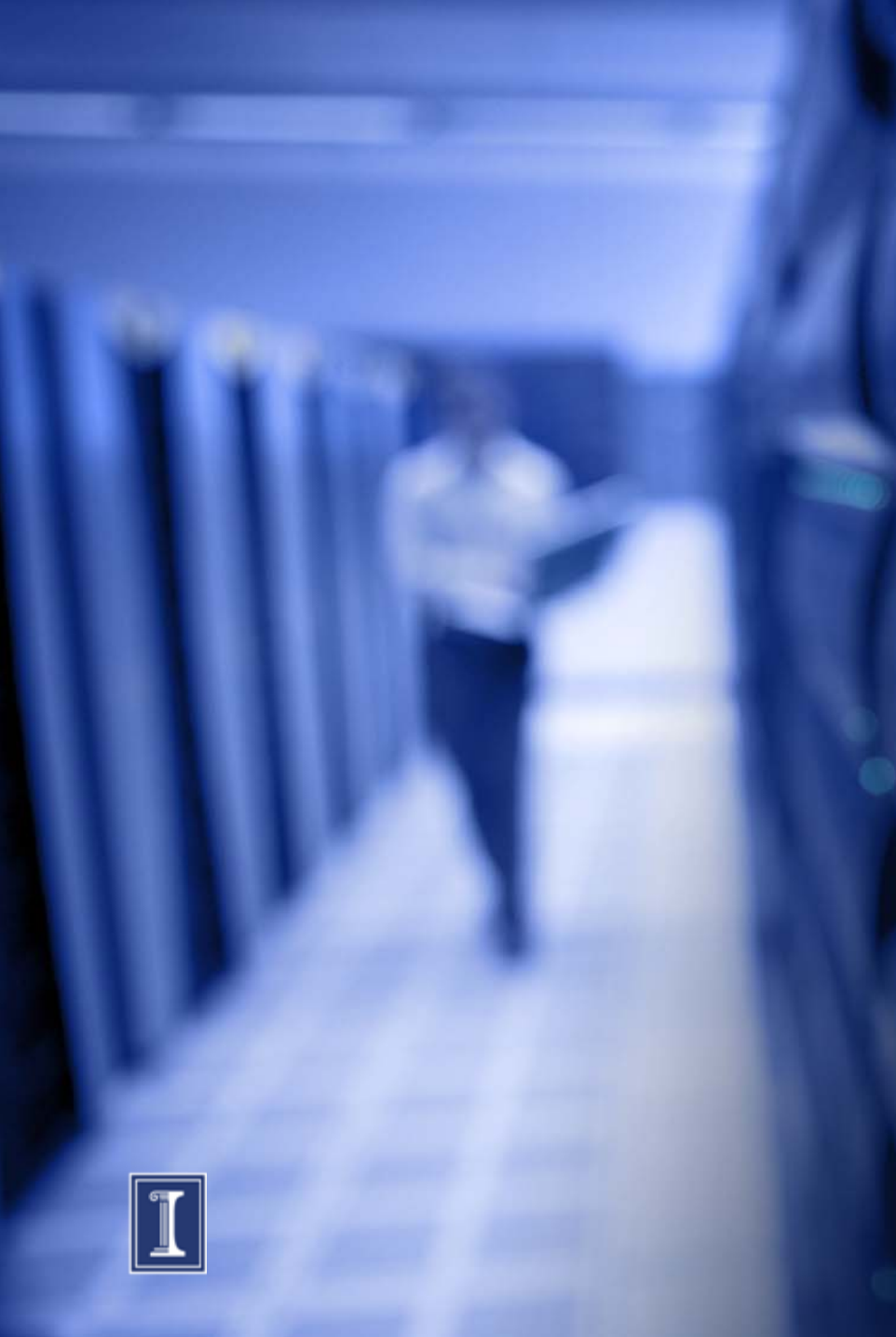


Preserve or Not To Preserve? How Can Computers Help with Appraisals.

Peter Bajcsy, PhD

- Research Scientist, NCSA

- Adjunct Assistant Professor ECE & CS at UIUC

- Associate Director Center for Humanities, Social Sciences and Arts (CHASS), Illinois Informatics Institute (I3), UIUC

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Date: October 16th, 2008

Acknowledgement

- This research was partially supported by a National Archive and Records Administration (NARA) supplement to NSF PACI cooperative agreement CA #SCI-9619019 and by NCSA Industrial Partners.
- The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the National Archive and Records Administration, or the U.S. government.
- Contributions by: Peter Bajcsy, Sang-Chul Lee, William McFadden, Kenton McHenry and Alex Yahja

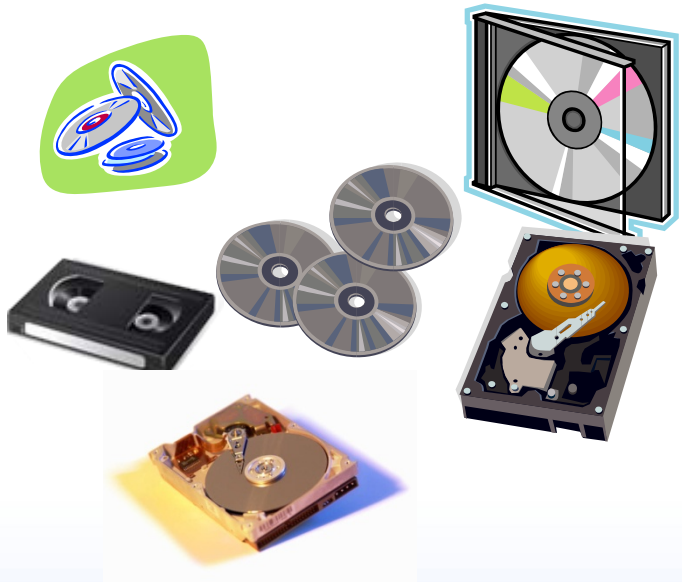
Outline

- Introduction
  - The Strategic Plan of The National Archives and Records Administration 2006–2016
- Motivation
  - Past & Current Research
- Computer-Assisted Appraisal of Documents
  - Approach
  - PDF Documents
  - Methodology
- Experimental Results
  - Grouping, Ranking and Integrity Verification
  - Computational Scalability
- Conclusions

Introduction: To Be Preserved!

Information transfer ?

Digital representation of information & knowledge

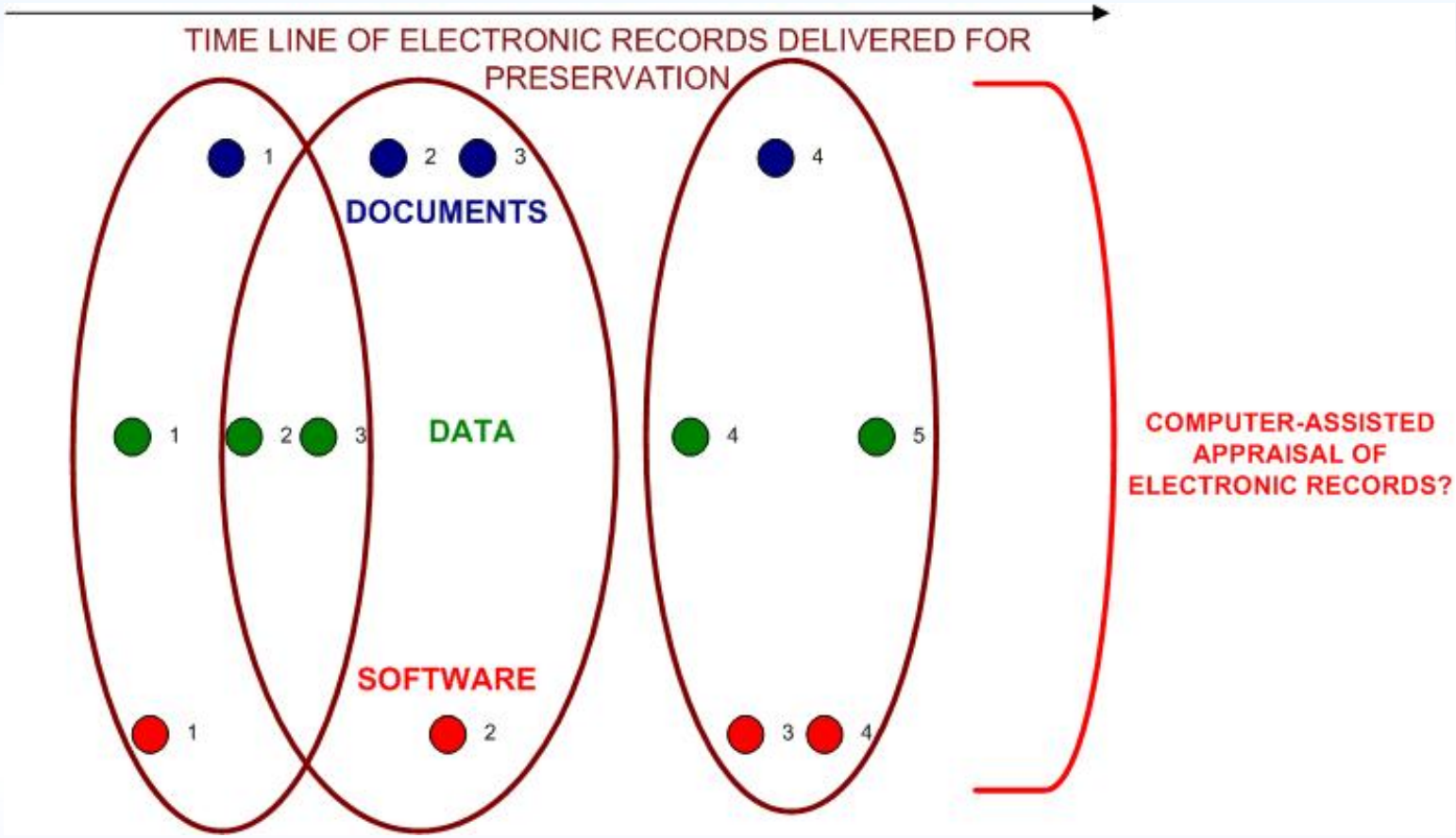


Preservation

AGENCY

ARCHIVES

Introduction: What Should Be Done?



- Can People Do It Manually?
- Human versus Computer or Human with Computer?

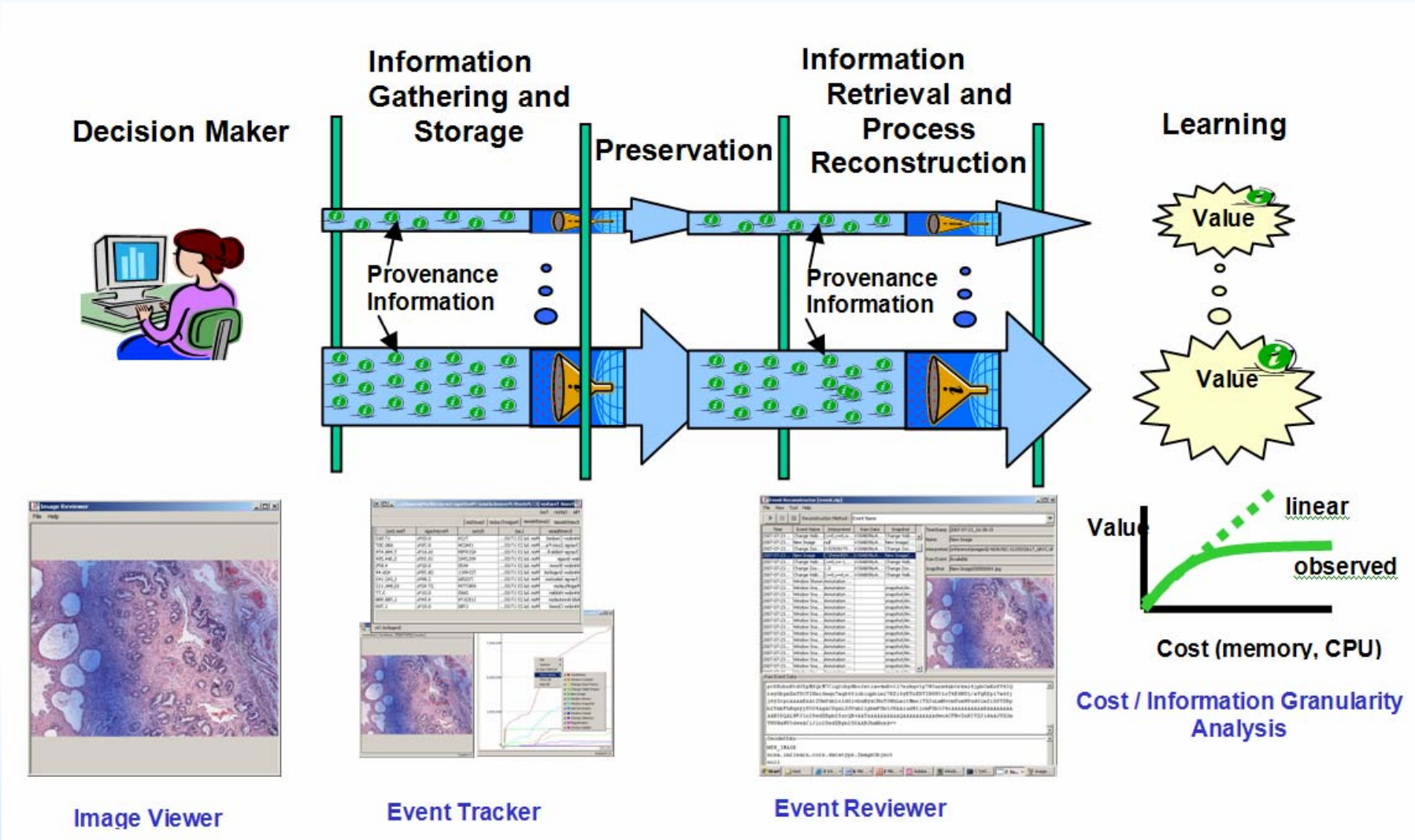
Introduction: Strategic Plan

- According to *The Strategic Plan of The National Archives and Records Administration 2006–2016.* "Preserving the Past to Protect the Future"
  - **"Strategic Goal 2:** We will preserve and process records to ensure access by the public as soon as legally possible"
    - *"D. We will improve the efficiency with which we manage our holdings from the time they are scheduled through accessioning, processing, storage, preservation, and public use."*
- The management and appraisal of electronic documents have been identified among the top ten challenges in the 34th Semi-annual Report to Congress by National Archives and Records Administration (NARA) Office of Inspector General (OIG) in 2005.
- Official appraisal policy of NARA adopted in May 17, 2006, and issued as NARA Directive 1441

Motivation (past research)

- To address the *Strategic Plan of The National Archives and Records Administration – specifically*
  - (1) Understand the tradeoffs between information value and computational/ storage costs by providing simulation frameworks
    - Information granularity, organization, compression, encryption, document format, ...
    - Versus
    - Cost of CPU for gathering information, for processing and for input/output operations; cost of storage media, upgrades, storage room, …
- **Prototype simulation framework**: Image Provenance To Learn available for downloading from isda.ncsa.uiuc.edu

Simulation Framework: Architecture



Motivation (current research)

- To address the *Strategic Plan of The National Archives and Records Administration – specifically*
  - (2) Assist in *improving the efficiency with which archivists manage all holdings from the time they are scheduled through accessioning, processing, storage, preservation, and public use."*
    - Are the records related to other permanent records?
    - What is the timeframe covered by the information?
    - What is the volume of records?
    - Is sampling an appropriate appraisal tool?
- **Prototype computer assisted appraisal framework**: Doc To Learn – work in progress

Objectives

Design a methodology, algorithms and a framework for document appraisal by

- (a) enabling exploratory document analyses
- (b) developing comprehensive comparisons and integrity/authenticity verification of documents
- (c) supporting automation of some analyses and
- (d) providing evaluations of computational and storage requirements and computational scalability of computer-assisted appraisal processes

Electronic Records of Interest

Electronic Records of Interest

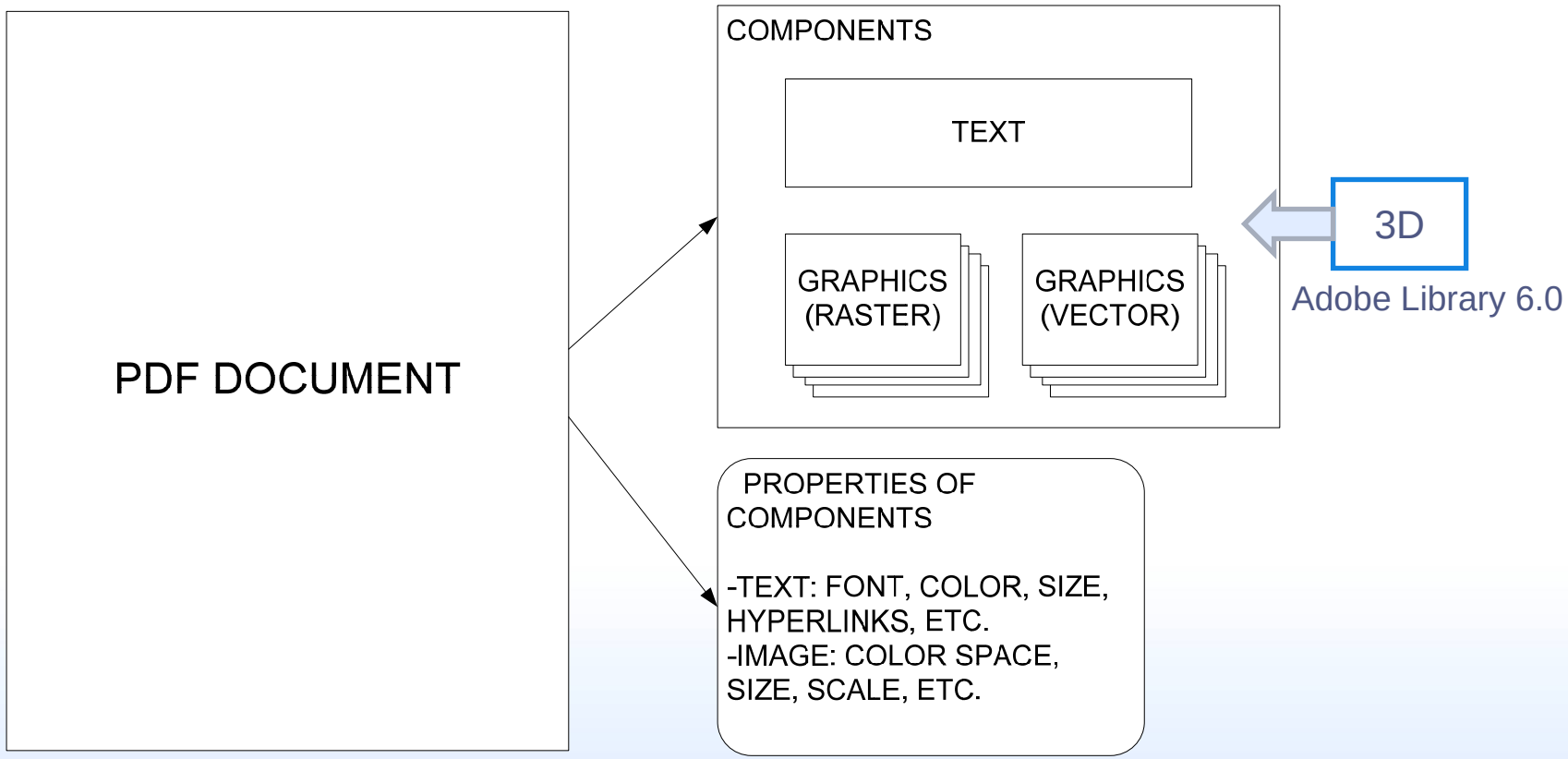
- Characteristics of a class of electronic records of interest
  - (a) Records contain information content found in software manuals, scientific publications or government agency reports
  - (b) Records have an incremental nature of their content in time, and
  - (c) Records are represented by office documents used for reporting and information sharing.
- File formats of electronic records of interest
  - Adobe PDF, PS,
  - MS Word, RTF,
  - TXT, HTML, XML, …

Focusing on Adobe Portable Document Format (PDF)

- Motivation:
  - Libraries for loading and writing PDF files are available for free to the academic community
  - PDF is one of the most widely used file formats for sharing contemporary office and publication information
  - PDF has the PDF/A type designed for archival purposes
    - For example, New York Times rented computational resources from Yahoo to convert 11 million scanned articles to PDF
  - PDF has been adding support for 3D and other data types

Adobe Portable Document Format (PDF)

- **Contemporary PDF documents**



Approach to Exploratory Document Analyses

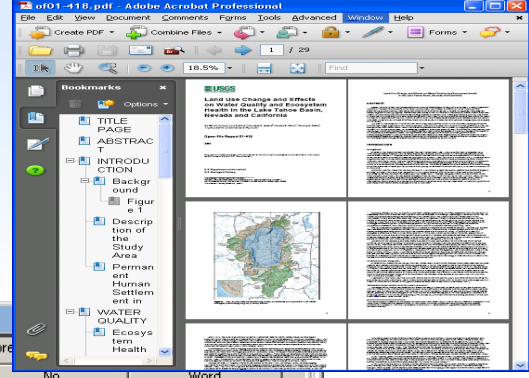
Exploration of PDF Components

- PDF Viewer presents information as a set of pages with their layouts
- PDF Viewer renders layers of internal objects (components) and hence only the top layer is visible

---

- Viewer of PDF docs for appraisal analyses presents information as a set of components and their characteristics
  - Text – word frequency
  - Images (rasters) – color frequency (histogram)
  - Vector graphics – line frequency
- Exploration of PDF docs for appraisal analyses includes visible and invisible objects

Prototype: Text Components



Document Analyzer

Document List

No	Filename	File Size	File Date	Num Pages	Num Images
1	pf01-418.pdf	1644KB	Wed Aug 29 15:30:04 CDT 2007	29	4
2	pubAboutLakeTaheo_fs-100-9...	1744KB	Mon Apr 16 15:15:58 CDT 2007	6	6

LOADED FILES

Occurrence of words

Occurrence of numbers

Word Frequency

All Frequency

No	Word	Frequency
0	analyzing	2
1	Service's	1
2	Changes	4
3	materials	2
4	slope	1
5	levent	1
6	Surface	1
7	Nev	1
8	generation	1
9	discharges	1
10	hydrology	1
11	timber	2
12	James	1
13	influen	1
14	$	1
15	http//www....	1
16	answer	1
17	+	3
18	regard	2
19	meet	1
20	panchromatic	3
21	arameters	1
22	.	969
23	Team	1

Minimum frequency: 0 Set

Text Frequency

No	Word	Frequency
0	analyzing	2
1	Service's	1
2	Changes	4
3	materials	2
4	slope	1
5	levent	1
6	Surface	1
7	Nev	1
8	generation	1
9	discharges	1
10	hydrology	1
11	timber	2
12	James	1
13	http//www....	1
14	$	1
15	influen	1
16	answer	1
17	+	3
18	regard	2
19	Team	1
20	.	969
21	arameters	1
22	panchromatic	3
23	meet	1

Minimum frequency: 0 Set

Integer Frequency

No	Word	Frequency
0	36	1
1	39	1
2	155	1
3	1970	4
4	1971	2
5	103365	1
6	43	1
7	103366	3
8	42	1
9	1975	2
10	1976	5
11	40	6
12	1978	1
13	1979	2
14	12921	1
15	200	1
16	1900	2
17	22	2
18	23	1
19	24	2
20	25	4
21	1990	6
22	26	2
23	27	2

Minimum frequency: 0 Set

Float Frequency

No	Word	Frequency
0	21.5	1
1	9.6	1
2	0.46	1
3	0.011	1
4	0.006	1
5	0.043	1
6	56.5	1
7	0.009	1
8	0.027	1
9	41.2	1

Minimum frequency: 0 Set

Ignore

No	Word
1	
2	&
3	=
4	A
5	Data
6	For
7	In
8	Only
9	The
10	These
11	This
12	We

New Remove Load Save

"Ignore" words

Compare List

No	Word	Info

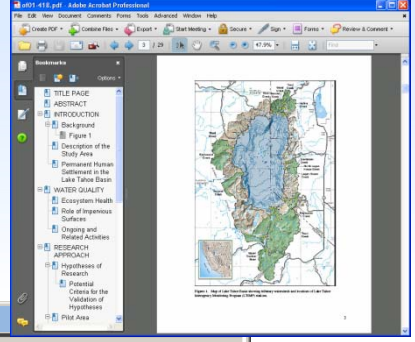
Save

Load Launch Document Extract Images Remove Use Filter Compare Group Move to Ignore List Reset Done

Performing Word Comparison...

NCSA

Prototype: Image Components



Document Analyzer

Document List

No	Filename	File Size	File Date	Num Pages	Num Images
1	of01-418.pdf	1644KB	Wed Aug 29 15:30:04 CDT 2007	29	4
2	pubAboutLakeTaheo_fs-100-9...	1744KB	Mon Apr 16 15:15:58 CDT 2007	6	6

LOADED FILES

Ignore List

No	Word
1	Col_0_0_0
2	Col_255_255_255

"Ignore" colors

List of images

Occurrence of colors

Preview

Image List

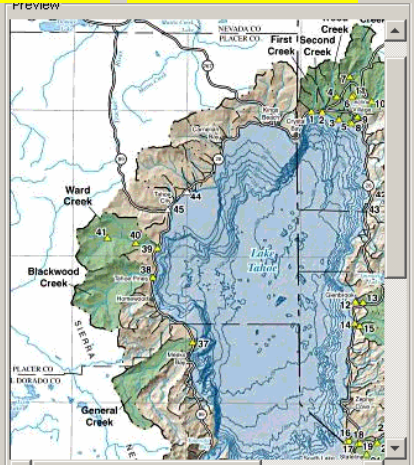
No	Filename
1	(numrows=630, numcols= 43...
2	(numrows=397, numcols= 29...
3	(numrows=436, numcols= 43...
4	(numrows=574, numcols= 61...

Color Frequency

No	Color	Frequency
0	RGB_85_198_170	1
1	RGB_142_198_255	487
2	RGB_113_85_28	3
3	RGB_28_85_85	41
4	RGB_198_227_227	4176
5	RGB_170_170_0	2
6	RGB_227_198_227	82
7	RGB_142_198_113	51
8	RGB_255_255_28	1
9	RGB_142_142_170	336
10	RGB_57_113_28	4
11	RGB_113_113_113	2197
12	RGB_255_198_198	7
13	RGB_57_113_170	75
14	RGB_142_170_198	2556
15	RGB_198_255_255	594
16	RGB_170_113_85	1
17	RGB_227_255_227	1110
18	RGB_142_170_227	76
19	RGB_227_227_85	28
20	RGB_255_227_85	1
21	RGB_85_170_170	46
22	RGB_198_198_255	14
23	RGB_0_85_170	1
24	RGB_170_170_170	4507
25	RGB_198_198_113	8

Minimum frequency: 0 Set

Preview



Down Color Bins: 10 Set

Show Color Reduced Image

New Remove Load Save

Compare List

No	Word	Info

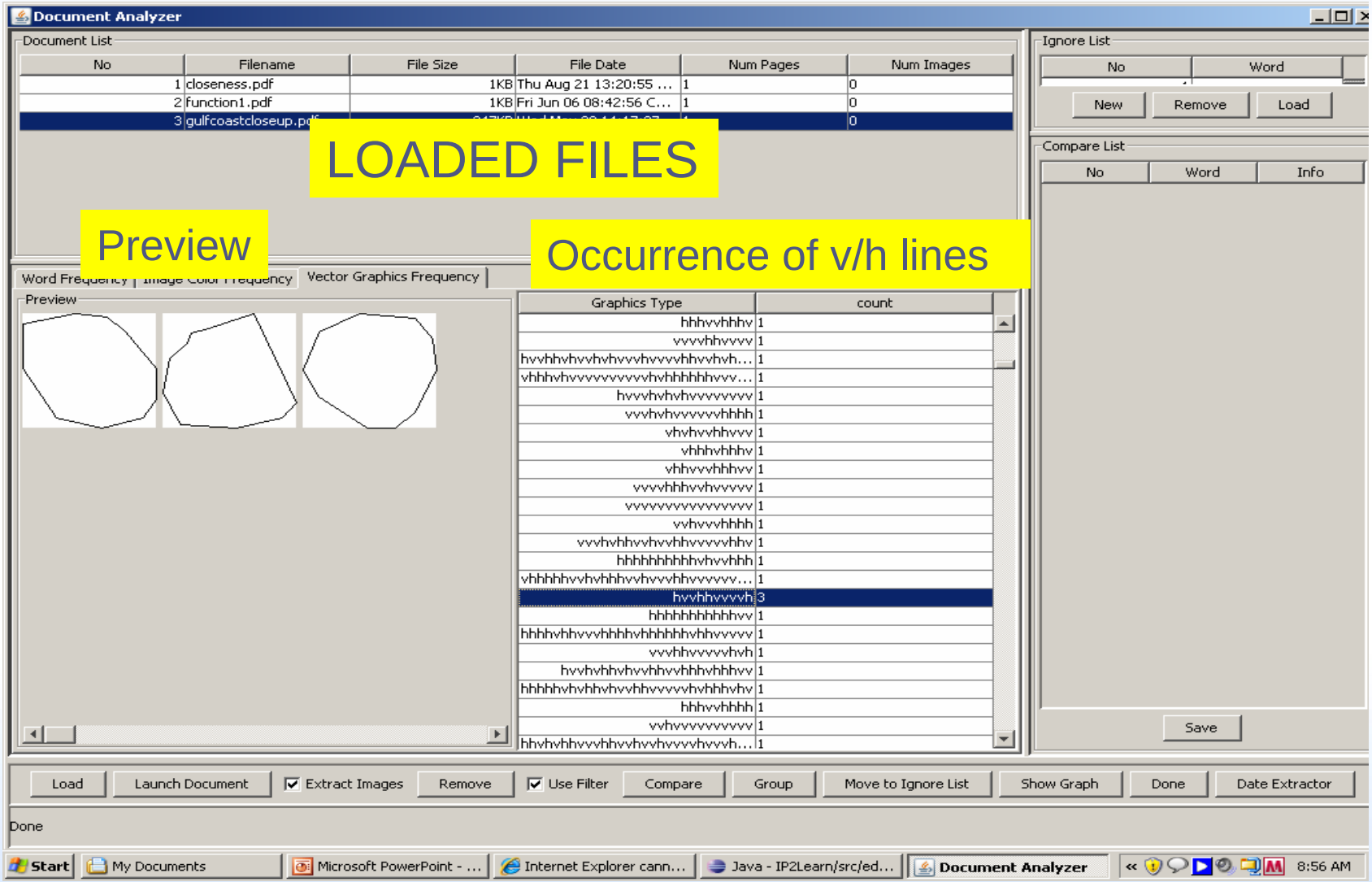
Save

Load | Launch Document | Extract Images | Remove | Use Filter | Compare | Group | Move to Ignore List | Reset | Done

Performing Word Comparison...

NCSA

Prototype: Vector Graphics Components



Be Aware of Visible And Invisible Objects in PDF Documents

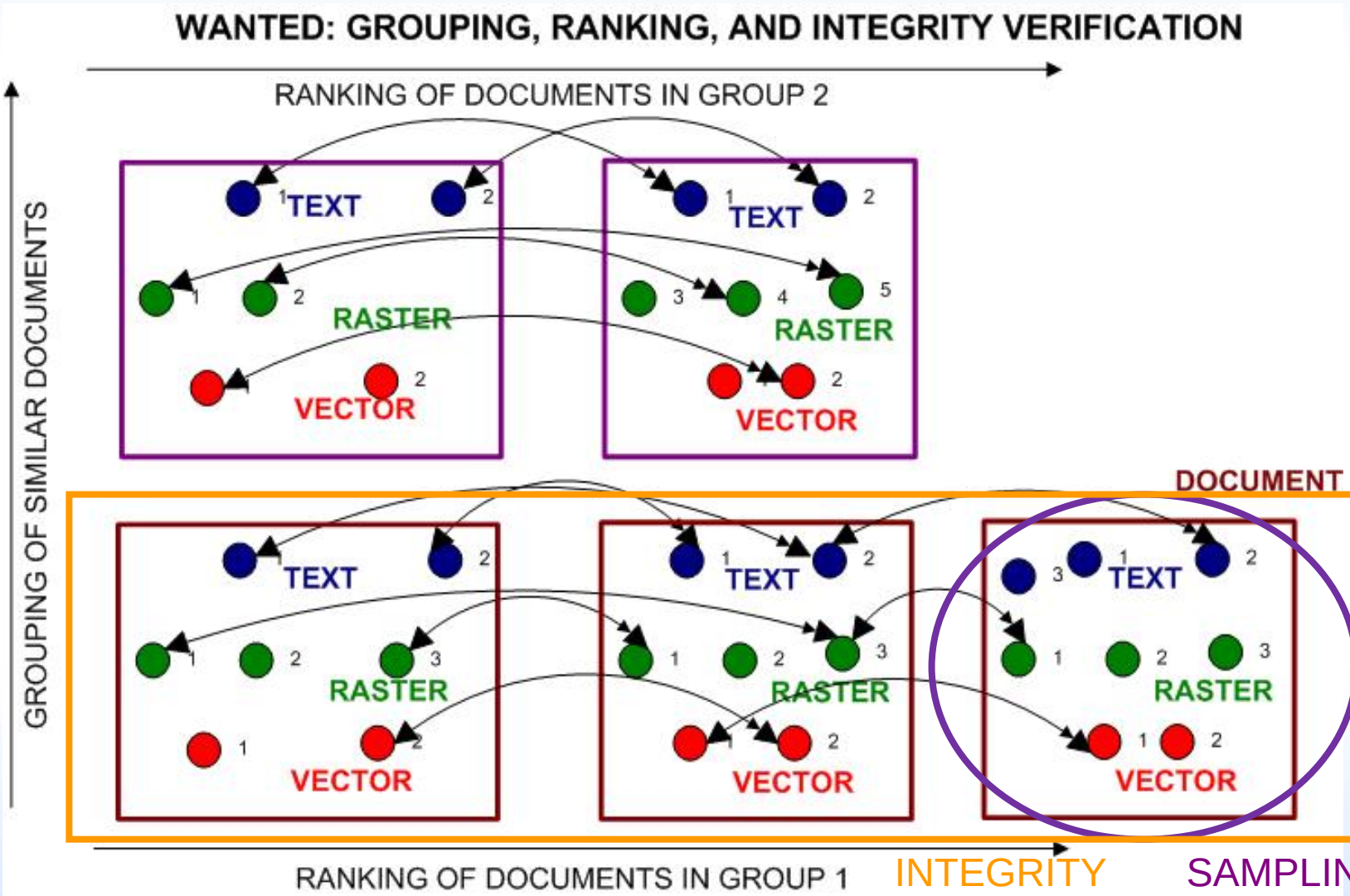
Approach to Developing Comprehensive Comparisons and Integrity/Authenticity Verification of Documents

Approach

Decompose the series of appraisal criteria into a set of focused analyses:

- (a) find groups of records with similar content,
- (b) rank records according to their creation/last modification time and digital volume,
- (c) define inconsistency rules and detect inconsistencies between ranking and content within a group of records,
- (d) design preservation sampling strategies and compare them.

Overview of the Approach



Related Work

- Past work in the areas of
  - (a) content-based image retrieval,
  - (b) digital libraries, and
  - (c) appraisal studies.
- We adopted some of the image comparison metrics used in (a), text comparison metrics used in (b), and lessons learnt from (c) to achieve a comprehensive comparison based on text, image/raster and vector graphics PDF components.

Mathematical Framework Needed for Document Comparisons

- Similarity of two documents

$$sim(D_i, D_j) = w_{TEXT} \cdot sim(T_i, T_j) + w_{RASTER} \cdot sim(\{I_{ik}\}_{k=1}^{K}, \{I_{jl}\}_{l=1}^{L}) + w_{VECTOR} \cdot sim(V_i, V_j)$$

- Weighting coefficients

$$W_{IMAGE}(D_i, D_j) = \frac{R_{IMAGE}(D_i) + R_{IMAGE}(D_j)}{2} \qquad R_{IMAGE}(D) = \frac{Area_{IMAGE}(D)}{Area_{IMAGE}(D) + Area_{VECTOR}(D) + Area_{TEXT}(D)}$$

$$W_{IMAGE}(D_i, D_j) + W_{VECTOR}(D_i, D_j) + W_{TEXT}(D_i, D_j) = 1 \qquad R_{IMAGE}(D) + R_{VECTOR}(D) + R_{TEXT}(D) = 1$$

- Intra- and inter-doc image-based similarity

$$sim(I_{ik} \in D_i, I_{il} \in D_j) = \sum_{k1,k2} \omega_{i,k1}\omega_{i,k2} \qquad sim(\{I_{ik}\} \in D_i, \{I_{jl}\} \in D_j) = \sum_{k1,k2} \omega_{i,k1}\omega_{j,k2}$$

Intra-document Inter-document

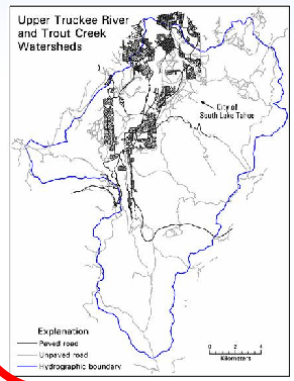
$$\omega_{ik} = \frac{f_{ik} \log(N / n_k)}{\sqrt{\sum_{l=1}^{L} (f_{il})^2 (\log(N / n_l))^2}}$$

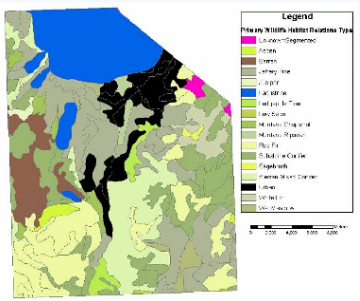
- Text-based and v/h line count similarity

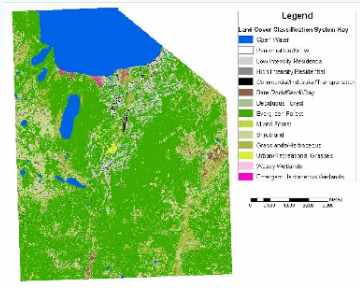
$$sim(T_i, T_j) = \sum_{k1,k2} \omega_{i,k1}\omega_{j,k2}$$

f –frequency of occurrence of a feature (word/color/line)
L - number of all unique feature primitives
n - number of documents that contain the feature (n=1 or 2)
N – number of documents evaluated

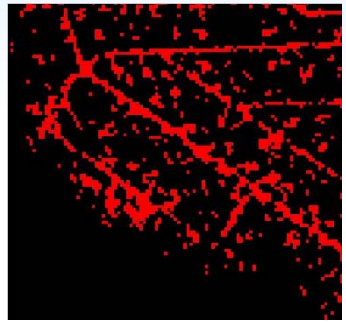
Example: Image Grouping



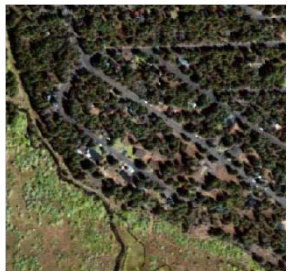




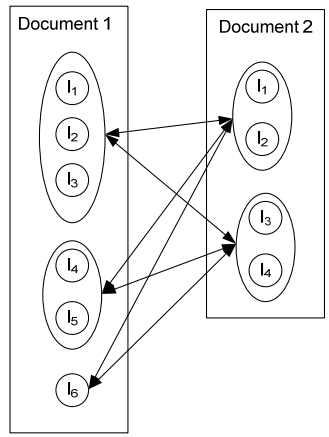
a

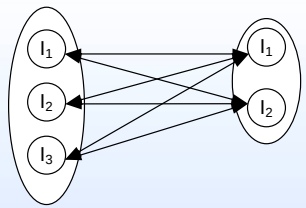


c



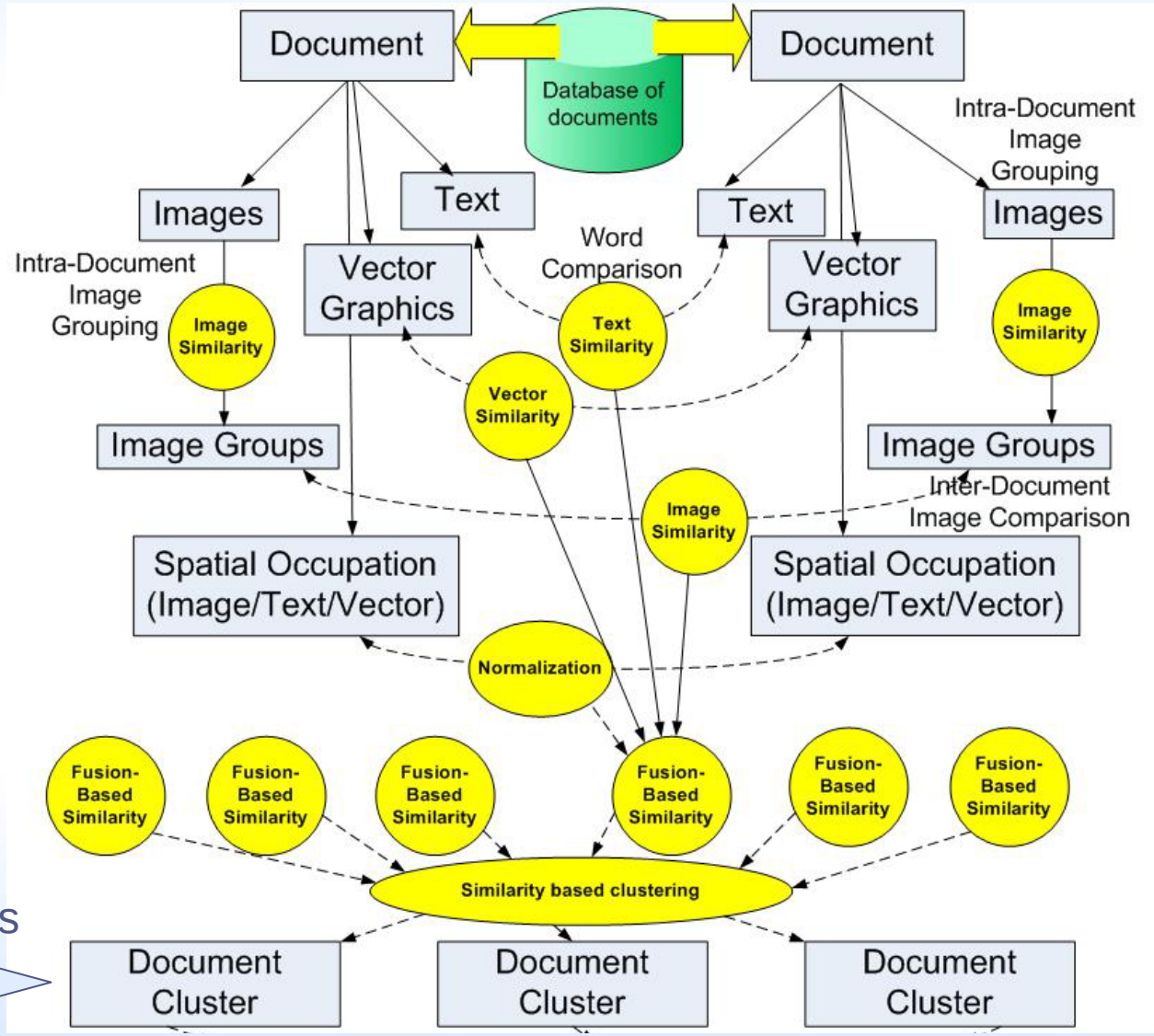
b





	Average similarity between image pairs	Standard deviation of the similarity
Group (a)	0.9565310641762074	0.045131416130196965
Group (b)	0.873736726083776	0.1746431238539268
Group (c)	1.0	0.0

Methodology

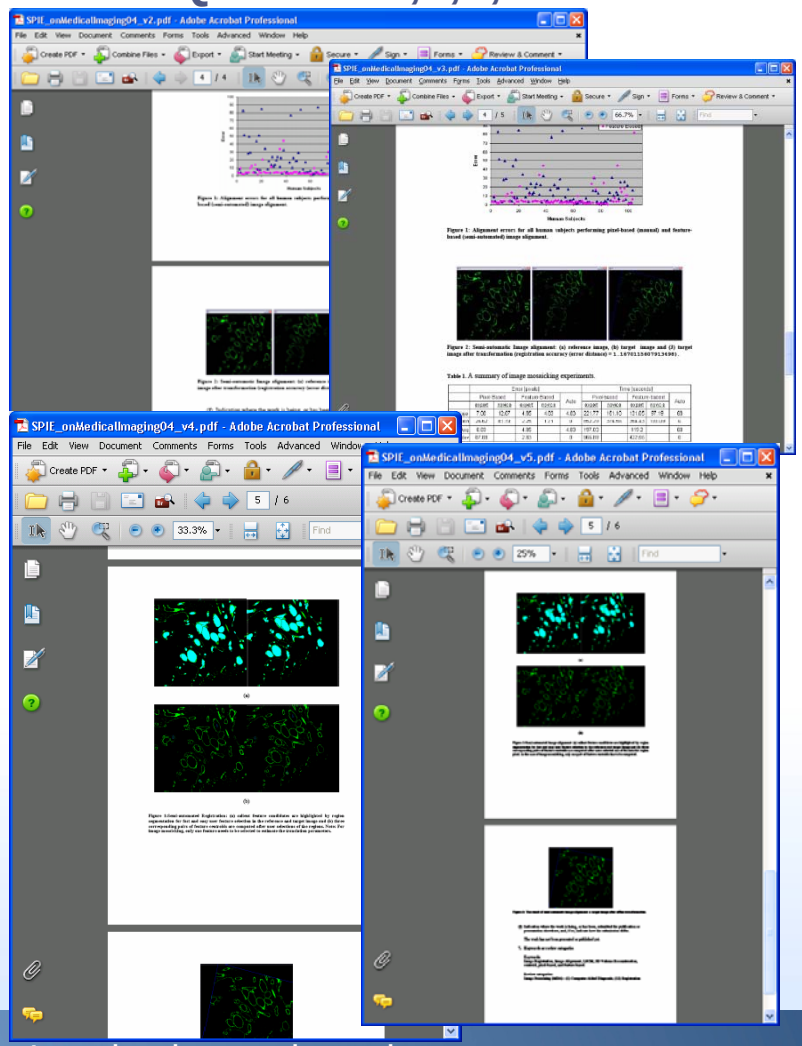


Relationship to Permanent Records

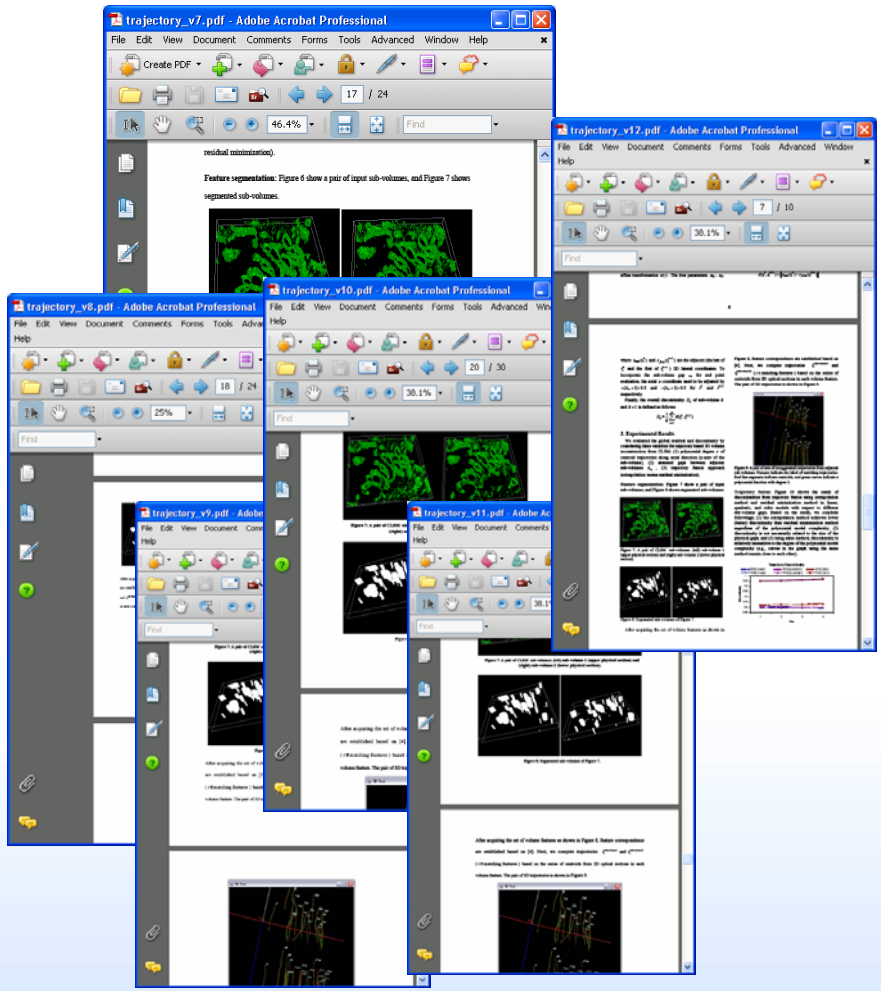
Illustrative Experimental Study

INPUT = 10 PDF docs (4 & 6 Groups)

UNIQUE ID= 1,2,3,4



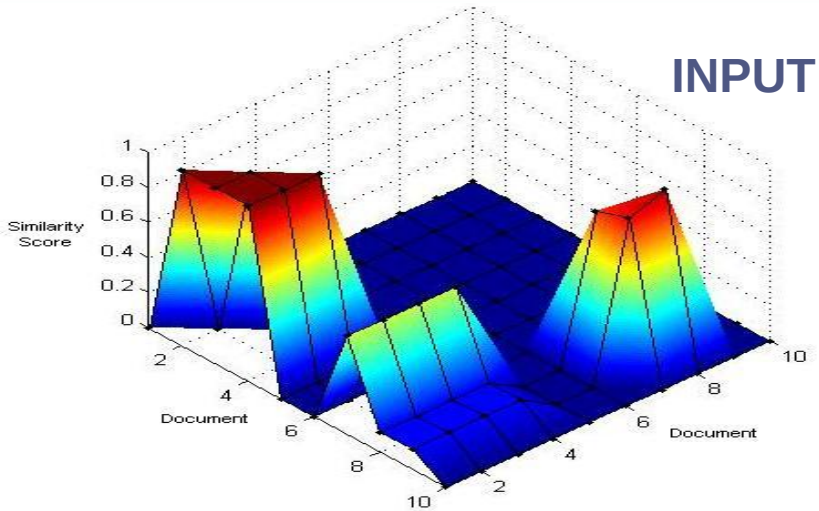
UNIQUE ID= 5,6,7,8,9,10



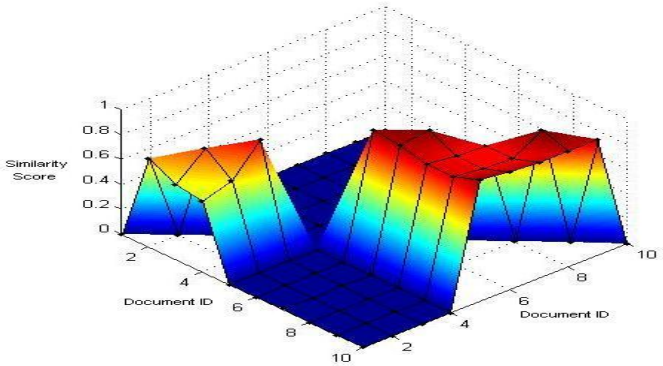
NCSA

Comparative Experimental Results

INPUT = 10 PDF docs (6 & 4 Groups)



Vector-based similarity



Text-based similarity

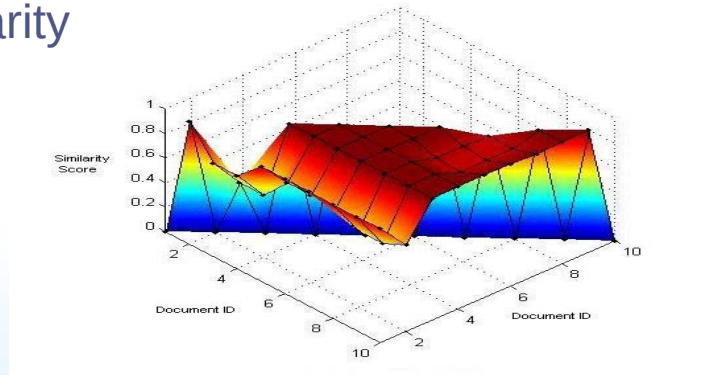
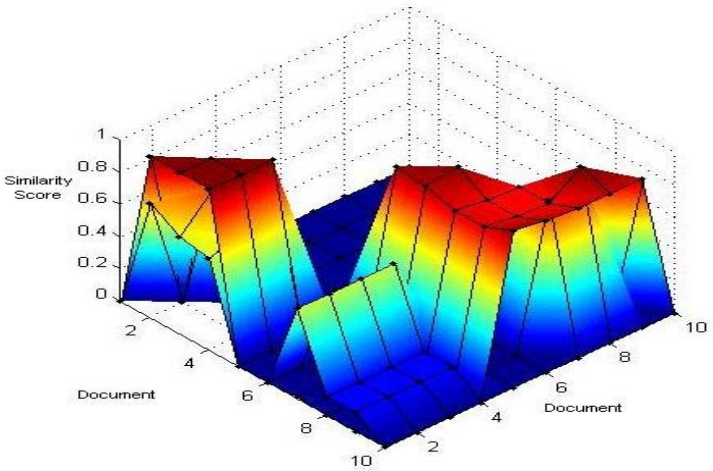
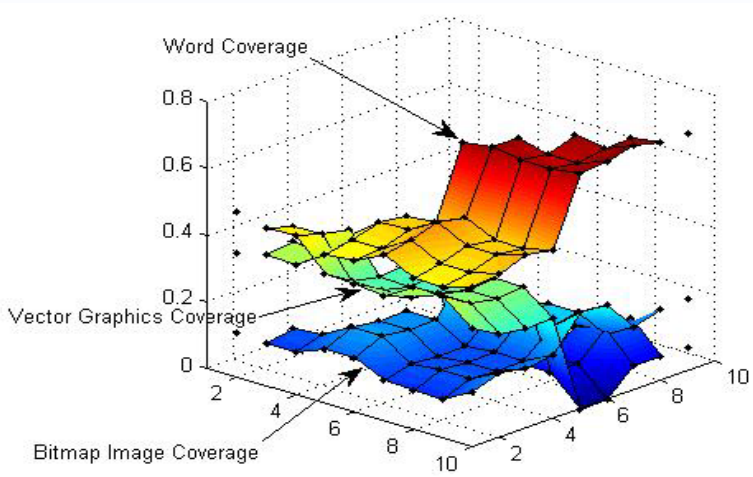


Image-based similarity

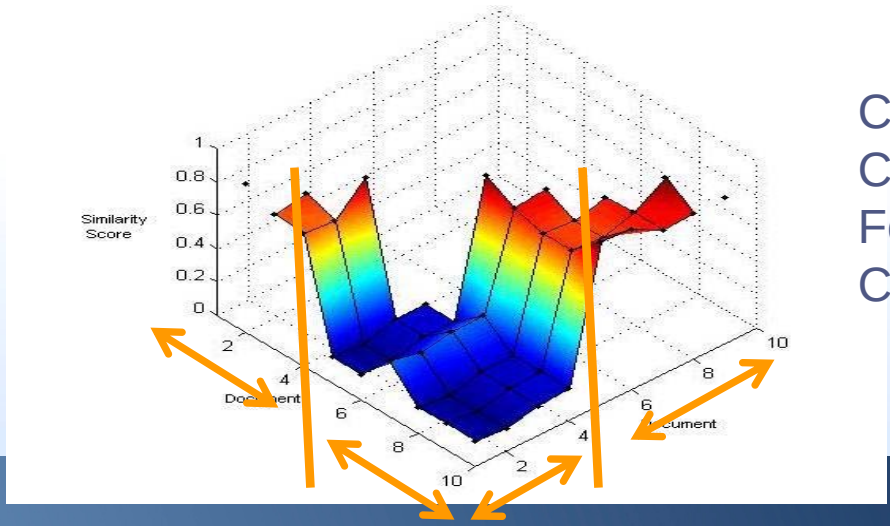
Comparative Experimental Results



Vector Graphics Similarity and Word Similarity Combined



Portion of Document Surface Allotted to Each Document Feature



Comparison Using Combination of Document Features in Proportion to Coverage

Accuracy Comparisons

Method	Average Similarity of Group 1	Average Similarity of Group 2	Average Similarity Across Group 1 & 2
TEXT ONLY	1	0.489	0
TEXT & IMAGE & GRAPHICS	0.906	0.520	0.075

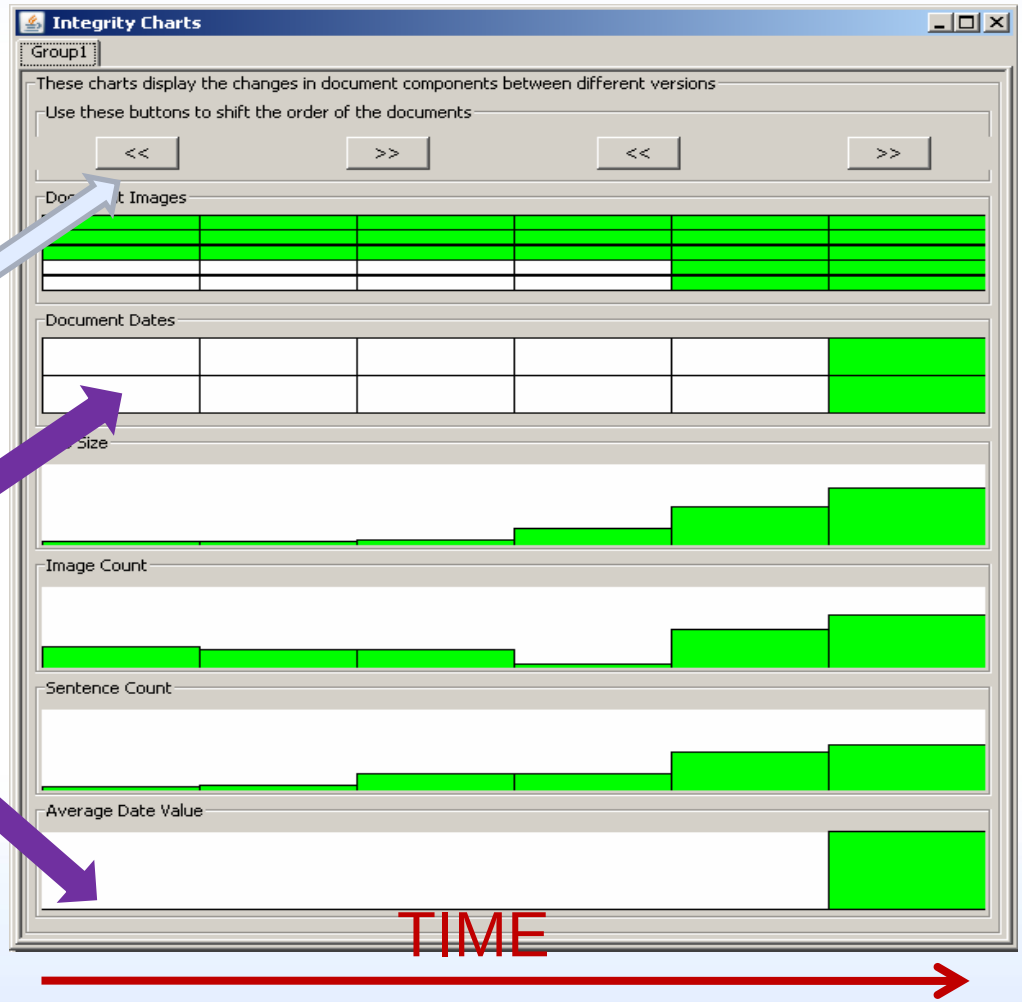
One refers to high similarity & zero refers to low similarity

Conclusions:

- Differences in similarity are up to 10% of the score
- Documents in Group 2 would likely be misclassified as 0.5 similarity would be the threshold between similar and dissimilar documents

Document Ranking According to Time

- Chronological ranking based on time stamps of files
  - Last modification (current implementation)
- Ranking can be changed by a human
- Content referring to dates can be used for integrity verification



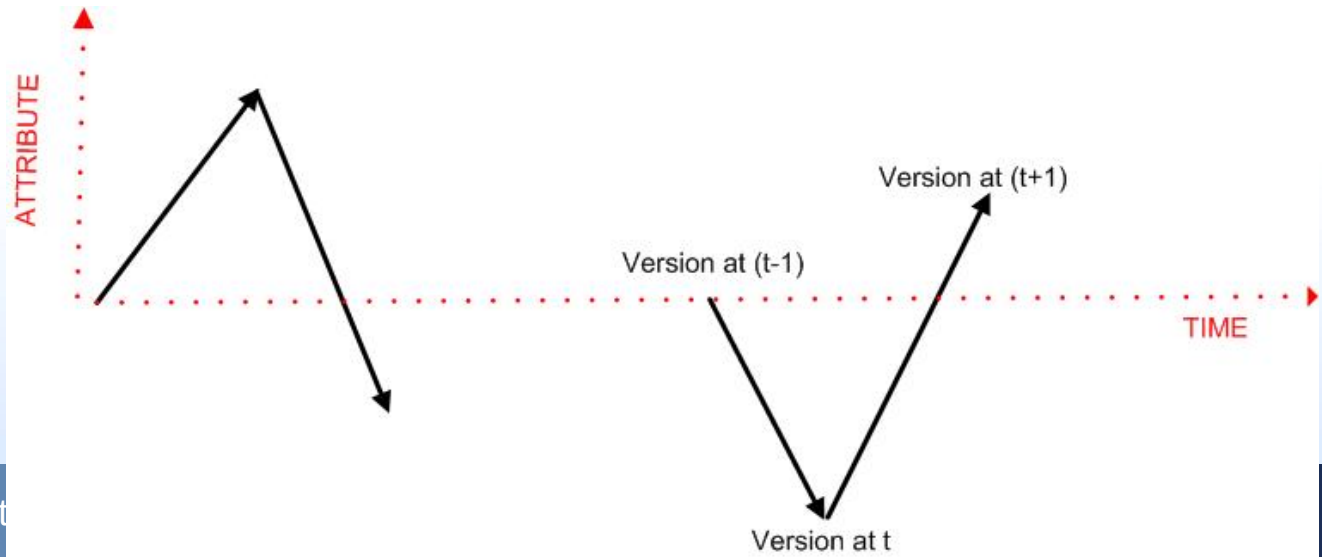
Integrity Verification

- **Document integrity attributes**:
  - appearance or disappearance of document images
  - appearance and disappearance of dates embedded in documents
  - file size
  - count of image groups
  - number of sentences
  - average value of dates found in a document

- **Approach:** rule based verification

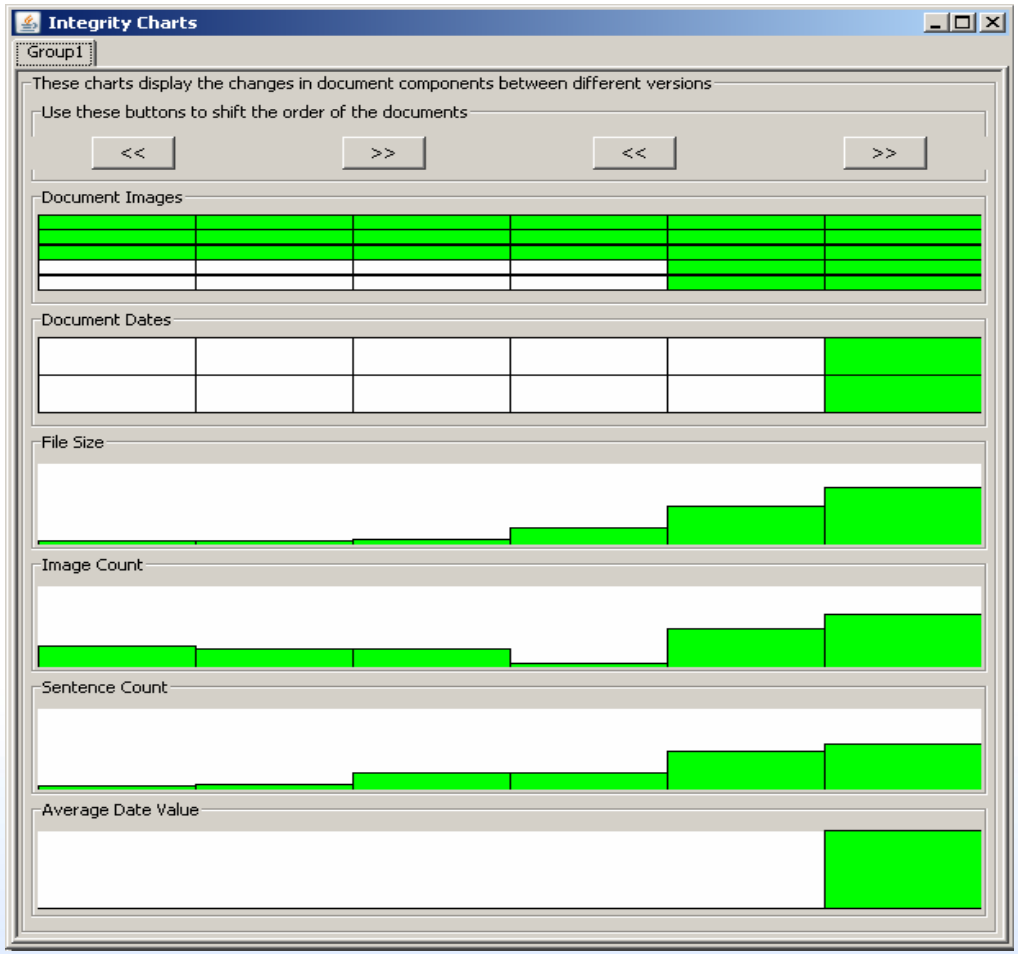
Integrity Verification Rules

- **Rule #1:** if (attribute (t-1) - attribute(t)) > thresh && (attribute (t+1) - attribute(t)) > thresh && attribute(t+1) >attribute(t-1) then fail
- **Rule #2:** if (attribute (t-1) - attribute(t)) < -thresh && (attribute (t+1) - attribute(t)) < -thresh && attribute(t+1) <attribute(t-1) then fail
- If rules fail for more than three attributes then alert for a document sequence



Integrity Verification - Passed

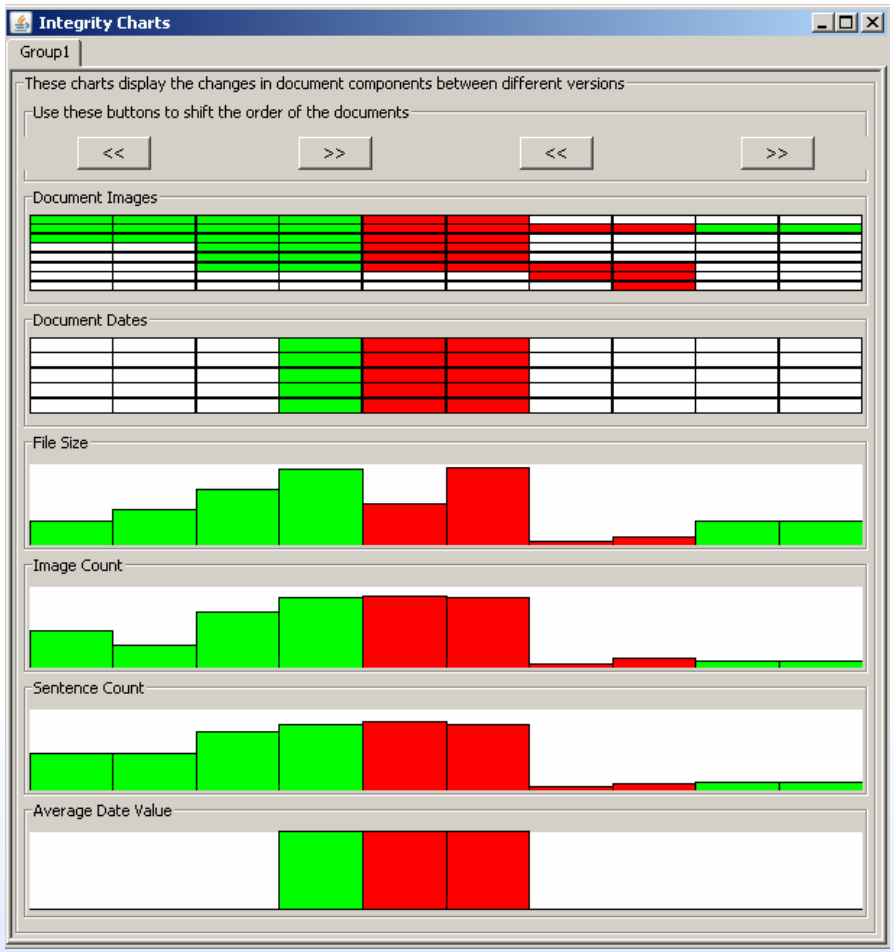
TIME



(1) appearance or disappearance of document images,
(2) appearance and disappearance of dates appearing in documents,
(3) file size,
(4) image count,
(5) number of sentence, and
(6) average value of dates found in document.

Integrity Verification - Failed

TIME



(1) appearance or disappearance of document images,
(2) appearance and disappearance of dates appearing in documents,
(3) file size,
(4) image count,
(5) number of sentence, and
(6) average value of dates found in document.

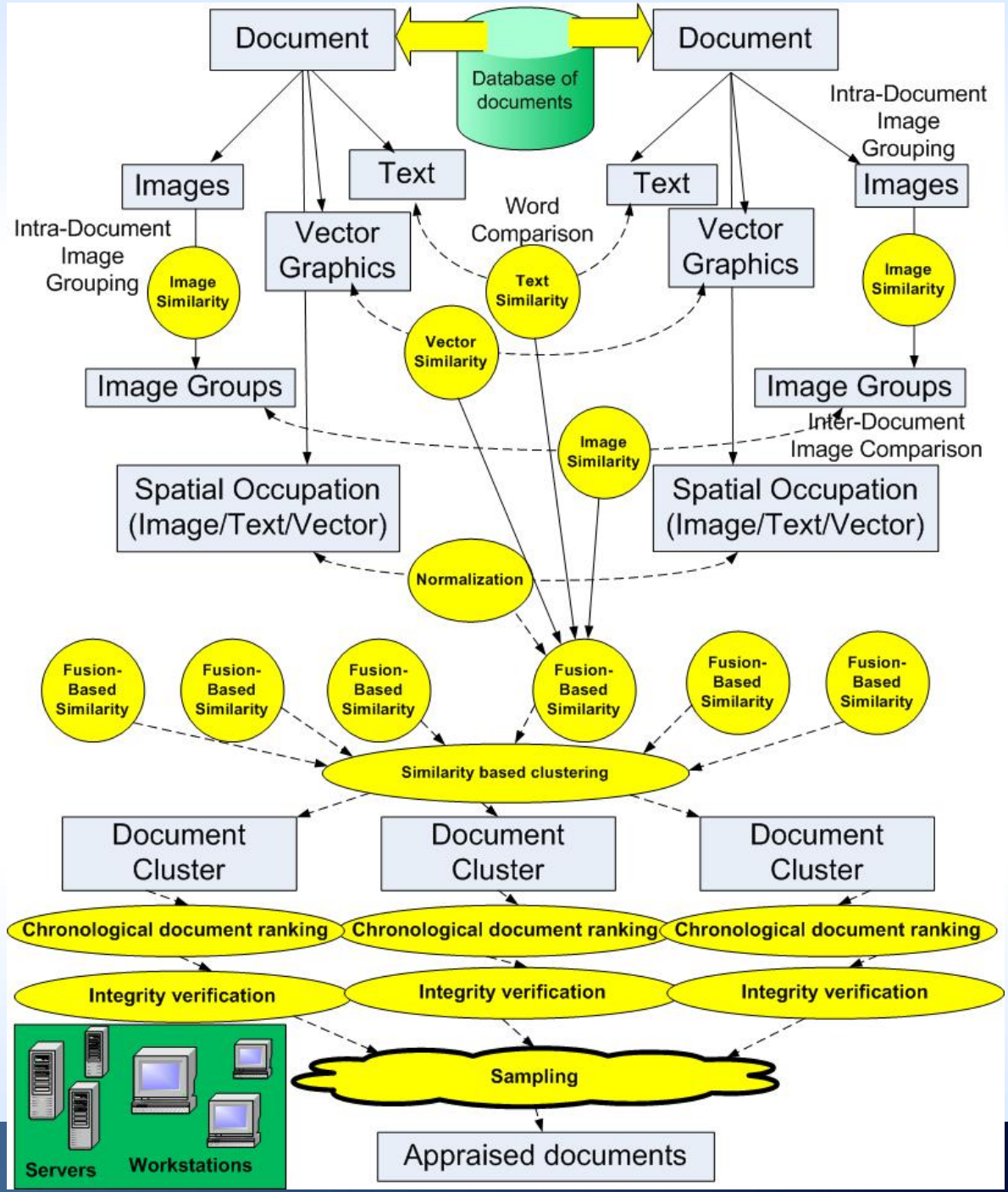
Approach to Providing Computational Scalability

Computational Requirements for Executing the Methodology

Yellow indicates computations

Relationship to Permanent Records

Appraisal & Sampling



Scalability of Document Appraisals

- **Options for parallel processing**
  - message-passing interface (MPI)
    - MPI is designed for the coordination of a program running as multiple processes in a distributed memory environment by using passing control messages.
  - open multi-processing (OpenMP)
    - OpenMP is intended for shared memory machines. It uses a multithreading approach where the master threads forks any number of slave threads.
  - Google's MapReduce for commodity clusters
    - It lets programmers write simple Map function and Reduce function, which are then automatically parallelized without requiring the programmers to code the details of parallel processes and communications

Simple Experiment with Google's MapReduce

- **Test data:** We downloaded 15 PDF files from the Columbia investigation web site at http://caib.nasa.gov/. We extracted text from the PDF documents using the Linux's pdftotext software to create a set of test files.
- **Software configuration:** We installed Linux OS (Ubuntu flavor) on three machines and then the Hadoop implementation of Map and Reduce functionalities. One machine was configured as a master and two as slaves.
- **Hardware configuration:** three machines – two laptops and one desktop; heterogeneous hardware specifications

Scalability of Document Appraisals

Machine\parameters	Processor	RAM	Hard Disk
1 - desktop	a quad-core Core 2 Duo processor 2.7 GHz	8 GBytes	750 GBytes
2 – laptop IBM Thinkpad T60	a dual-core Intel Core Duo processor 2 GHz	2 GBytes	80 GBytes
3 – laptop IBM Thinkpad T30	a single-core Intel Mobile Pentium 4-M processor 1.6 GHz	512 Kbytes	40 GBytes

Master & slave configuration	Performance time [sec]
Machine 1	49
Machines 1 and 2	35
Machines 1, 2 and 3	95

Conclusion: MapReduce (Hadoop implementation) does not perform very well in heterogeneous environments
Confirmed also by the most recent tech. report by Zaharia et al, UC Berkeley, August 2008

Conclusions

- **Accomplishments:** We have designed a framework for computer assisted document appraisal
  - A methodology
  - A prototype for grouping, ranking and integrity verification of PDF documents – support for document explorations
  - Identified computational challenges
- **Key contributions**:
  - Comprehensive comparison of PDF documents (text, images & graphics objects)
  - Initial integrity verification metrics
  - Automation and initial scalability studies
- **Future work**
  - Sampling is still an open question
  - Scalability of document analyses
    - Each file is large and the number of files is large
    - Exploring the TeraGrid resources
  - Inclusion of 3D data into the framework

Questions

- More information
  - Peter Bajcsy; email: pbajcsy@ncsa.uiuc.edu
  - Project URL: http://isda.ncsa.uiuc.edu/CompTradeoffs/
  - Publications – see our URL at http://isda.ncsa.uiuc.edu/publications