# Challenges in Preserving and Reconstructing Computer-Assisted Decision Processes

**Peter Bajcsy, Ph.D.**
**Research Scientist,**
**Image Spatial Data Analysis (ISDA) Group, NCSA**
**Adjunct Assistant Professor, CS and ECE Departments, UIUC**

**Contributions by Peter Bajcsy, Sang-Chul Lee and Rob Kooper,**
**National Center for Supercomputing Applications (NCSA)**

# Outline

- **Introduction**
- **Understanding Future Preservation and Reconstruction Requirements**
  - Focusing on Information Gathering and Reconstruction
  - Information Gathering Challenges
  - Proposed Approach
  - Simulation Framework for Understanding Preservation and Reconstruction Requirements for Computer Assisted Decision Processes
- **Example Tradeoff Studies Using the Simulation Framework**
  - Basic components: Image Viewer, Event Category Tracker, Event Viewer, Report Generator
  - Example Experimental Tradeoff Results
- **Discussion**

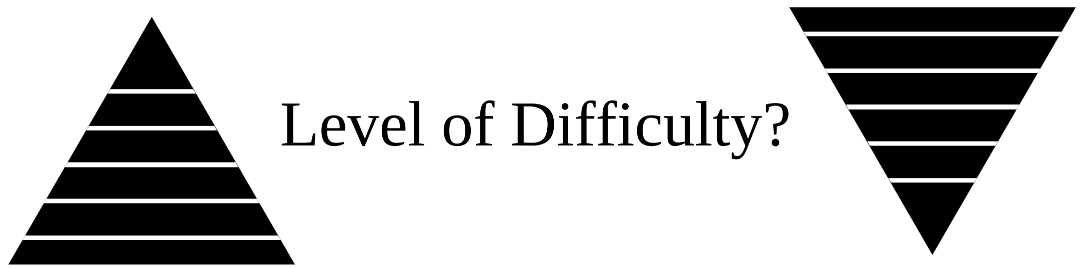
# Introduction: Trends

- **Two Trends in the Context of Decision Processes**
  - Decision processes are moving from paper based to electronic record based (~ **computer assisted decision processes**)
  - Usability of electronic records depends on rapidly changing information technology

- **Implications on Preservation and Reconstruction of Computer Assisted Decision Processes**
  - **What can we reconstruct from the past?**
  - **Can we preserve better today?**
  - **How much is preservation and/or reconstruction going to cost us tomorrow?**

# Introduction: Challenges

- **Top Level Challenges of Preservation & Reconstruction of Computer Assisted Decision Processes**
  - Dimensionality
  - Complexity
  - Volatility
  - Cost




- **Dimensionality and Complexity Example:**
  - **Temporal** axis (past, today and future processes)
  - **Human** axis (spectrum of agencies, socio-economics environments, human-computer interactions, …)
    - who had access/knowledge/resources/connections
    - what, where and when
  - **Electronic record** axis (spectrum of data formats, document formats, software languages, software platforms, …)
  - **In-silico** axis (computational hardware, computers/ peripherals/ networking/…)

# Introduction: Challenges

- **Volatility Example:**
  - Volatility of software/hardware/storage media
    - **Updates:** Windows operating systems since 2000: **Two major** new releases, **two minor** service pack updates, around **fifty security patches** since SP2
    - **Upgrades:** Microsoft Office Pro for Windows 95/98/ME/2000/XP/2003
    - **Media life expectancy:** Optical ~5 years, Disk ~ 15 years, Microfiche ~ 100, microfilm ~ 300, newspaper ~ 50, clay tablet ~ 10,000 (life expectancy vs. information density – [P. Conway, 1996] )
- **Cost Example:**
  - Cost of software/hardware/storage media
    - **Operating System:** Windows 3.1/95/98/NT/2000/XP/Vista: Windows 95 = $209; Windows NT = $280; Windows XP = $300; Windows Vista = $399
    - **128 MB of SDRAM: Year 1999** ~ $120-> $40 -> $200-250 due to Earthquake in Taiwan -> **March 2000** ~ $55->**March 2007** ~ $8.96 (flash card) - www.pricewatch.com (**1TB ~$500**)
    - **High performance computers: 2006: DARPA** awards approximately $500 million to Cray and IBM; **2007 NSF** $200 million to (NCSA)

# Introduction: General Approach

- **How can we answer basic questions?**
  - **What can we reconstruct from the past?**
  - **Can we preserve better today?**
  - **How much are preservation and reconstruction going to cost us tomorrow?**
- **Our Approach: Analysis and Synthesis of Focused Problems**
  - **Understanding Preservation and Reconstruction of Scientific Simulations Used In Past Computer Assisted Decision Processes**
  - **Simulation Framework for Understanding Preservation and Reconstruction Requirements of Future Computer Assisted Decision Processes**

# Simulation Framework for Understanding Preservation and Reconstruction Requirements of Future Computer Assisted Decision Processes



NCSA

# Problem Formulation

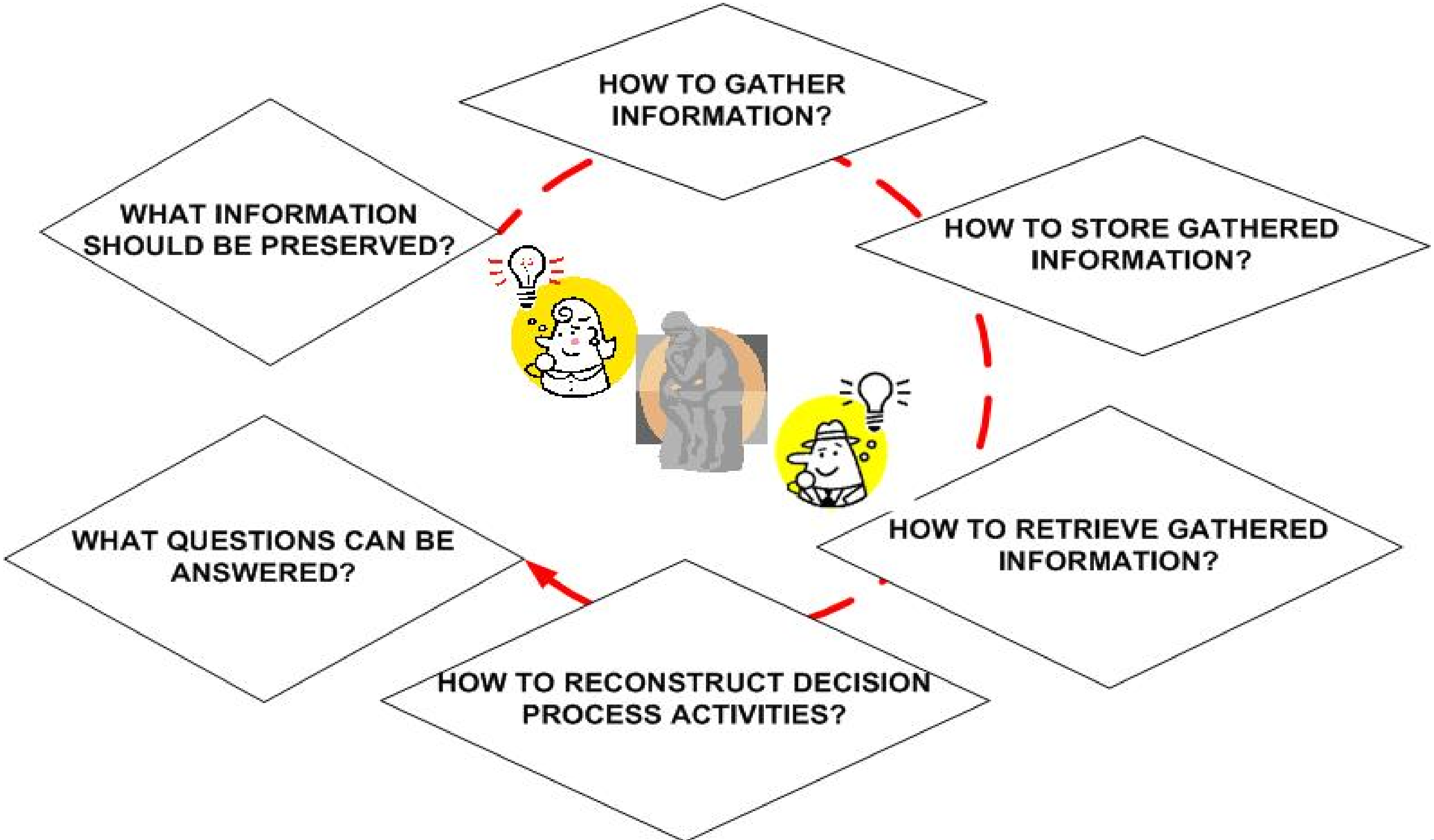
- **Problem**
  - Given a simulation of future computer assisted decision processes,
  - Predict computational requirements for preservation and reconstruction of decision processes

- **Motivation**
  - How decision making processes should be documented and archived in the future?
  - What computational performance and cost implications would be introduced if decision processes had been documented at the appropriate information granularity?

# Information Gathering: Challenges

- **Known challenges for building information gathering systems:**
  - System requirements:
    - Scalable, general and customizable, including trust, security and preservation [Groth et. al., 2004], as well as to providing automatic and systematic provenance gathering [Greenwood et. al., 2003].
  - Data sizes:
    - Large amounts and a wide variety of provenance information
  - Standards:
    - Paucity of standards, components and techniques for recording provenance [Groth et. al., 2004].
    - Lack of standards in semantic and syntactic descriptions about data leads to difficulties when the data sets are used, stored and retrieved.
    - Short history of standards since the first provenance systems were reported only about a decade ago (Geolineus in 1993 to track ArcInfo GIS operations, GOOSE in 1994, and Geo-Opera in 1997 [Bose & Frew, 2005]).
  - Scale:
    - Provenance of a single human (Gordon Bell and his MyLifeBits [Cherry, 2005]) or a community (collaborative environments and experiment management [Bose & Frew, 2005]).
- **There is no provenance system that would meet all above requirements, as well as specific requirements defined by multiple application domains.**

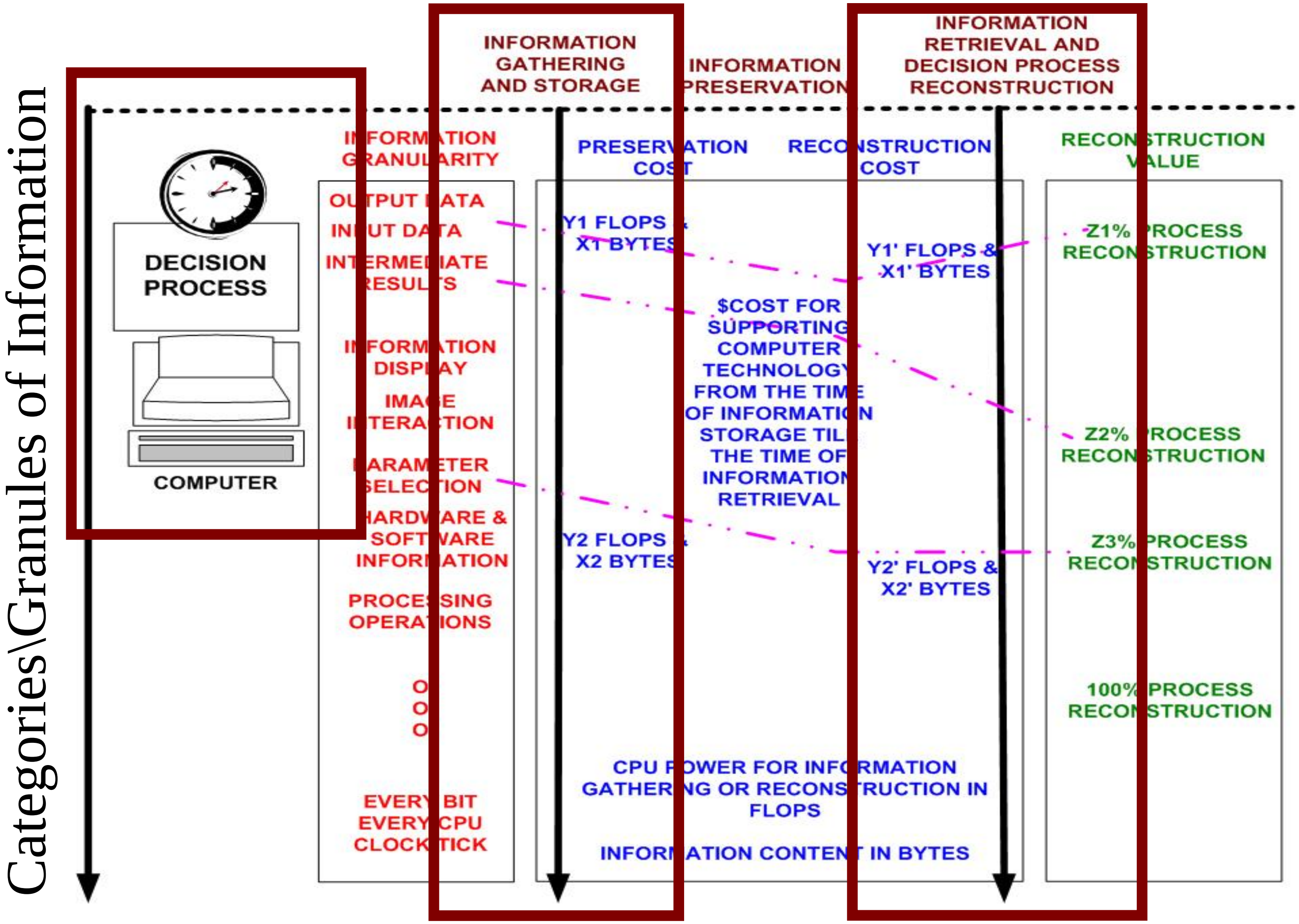
# Basic Questions



# Proposed Approach

- **Divide the problem into three sub-problems**
  - Identify a class of computer assisted decision making processes
  - Research information gathering about activities associated with decisions and semi-automated report generation about decision processes
  - Research and evaluate reconstructions of decision activities and authenticity of preserved records
- **Develop a simulation framework with three major components**
  - Solution to each sub-problem forms one major component
  - Basic archival questions could be studied by prototyping multiple solutions

# Simulation Framework Components

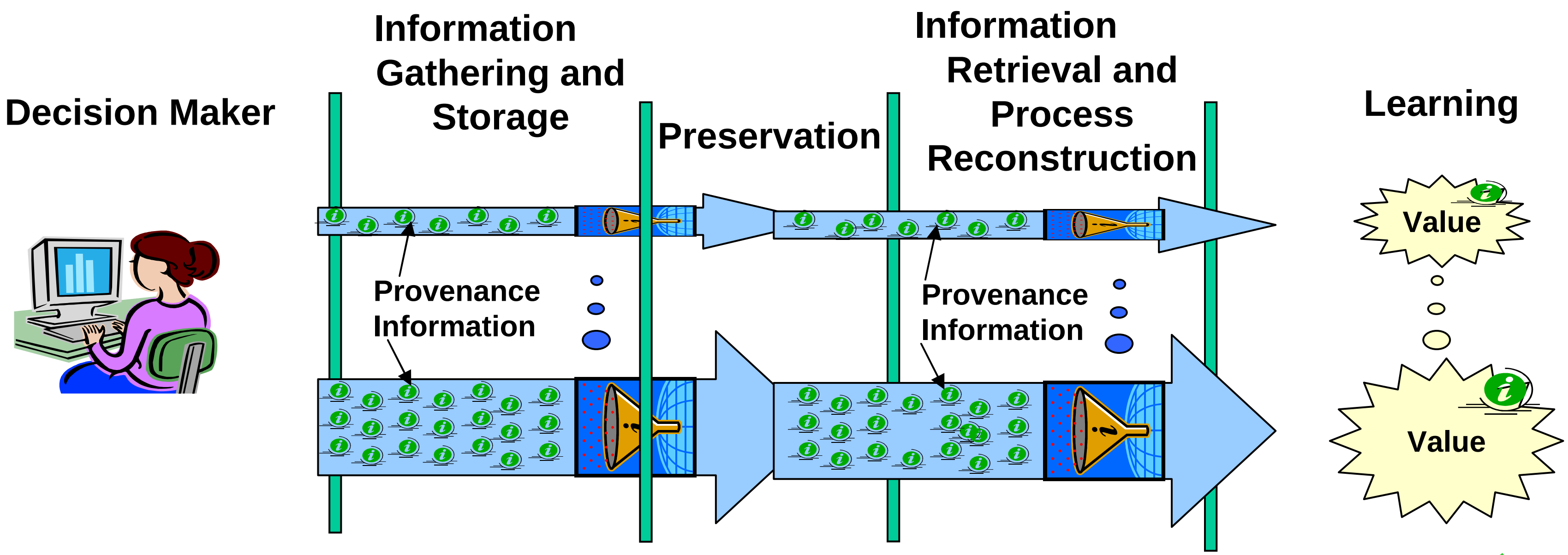



Three Major Components

# Selection of a Class of Computer Assisted Decision Processes

- **Objective:**
  - Find decision processes requiring common computer functionality
  - Select computer functionality that supports large number of decisions
- **Selection:**
  - A class of computer assisted decisions based on image inspection
  - Computer functionality to perform image inspections
- **Selection Motivation:**
  - Image properties
    - Language independent, omnipresent, multi-spectral/multi-dimensional
    - Basic form of important information for many decisions
  - What is of future interest?
    - Was the file viewed?
    - Did the file contain relevant information?
    - Was the image viewed at the appropriate magnification and spatial location?
    - Was image viewed using adequate intensity brightness and color bands?

# Simulation Framework



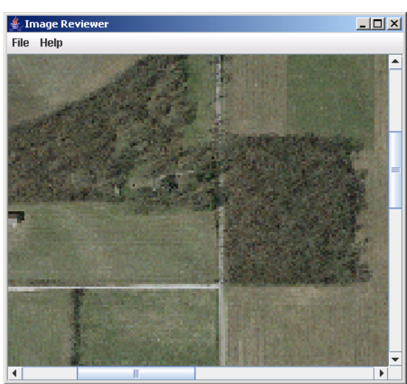
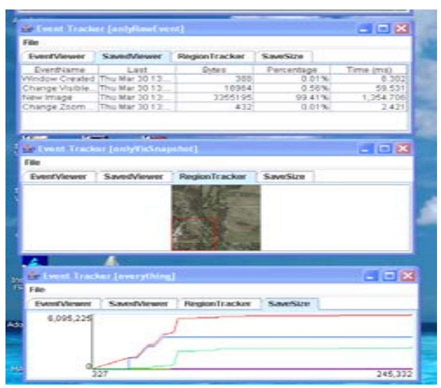
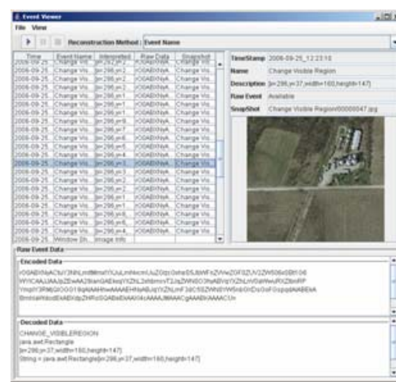



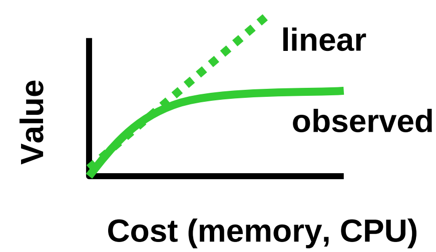

Image Viewer



Information Gathering System



Process Reconstruction System

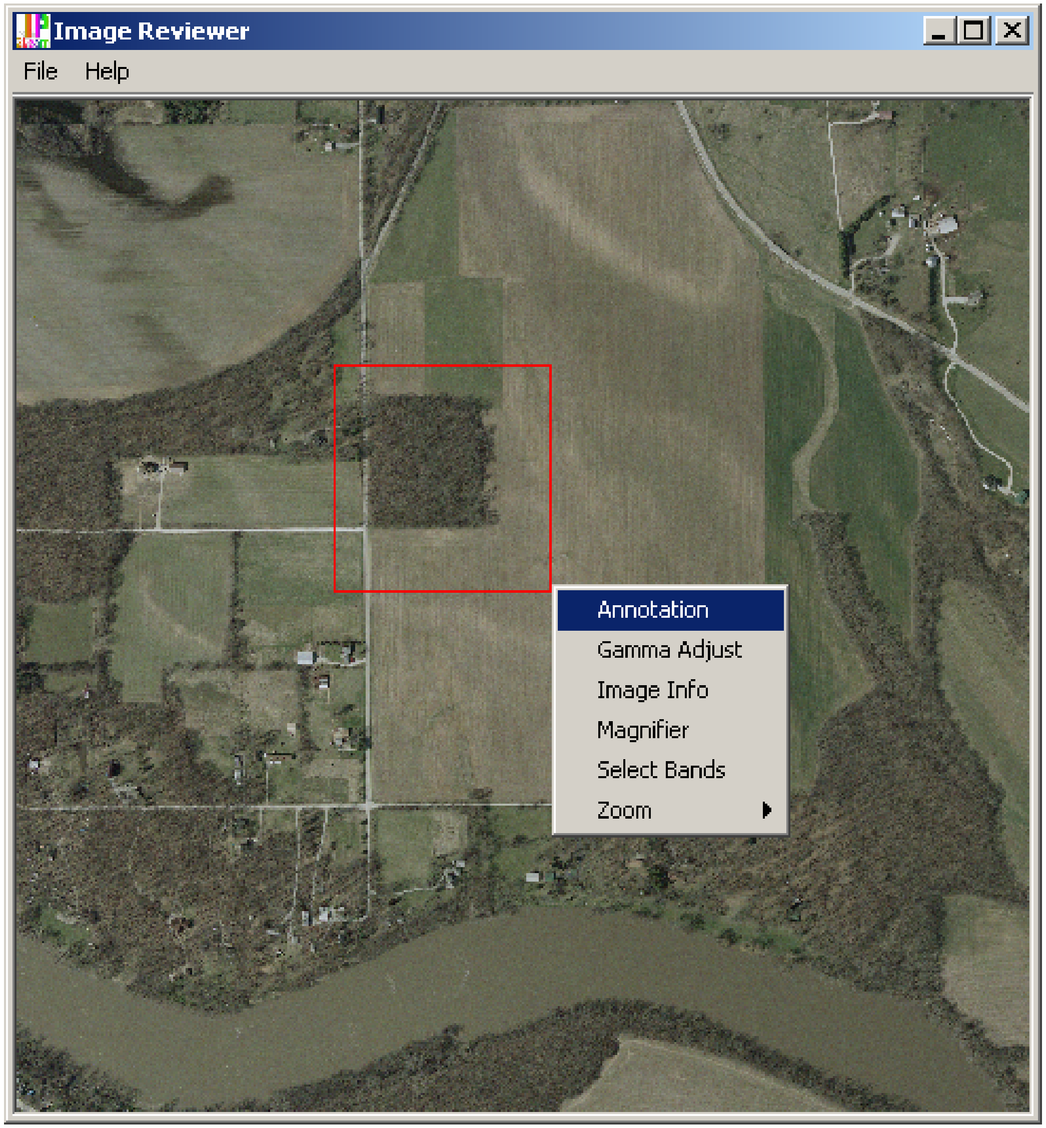



Cost / Information Granularity Analysis

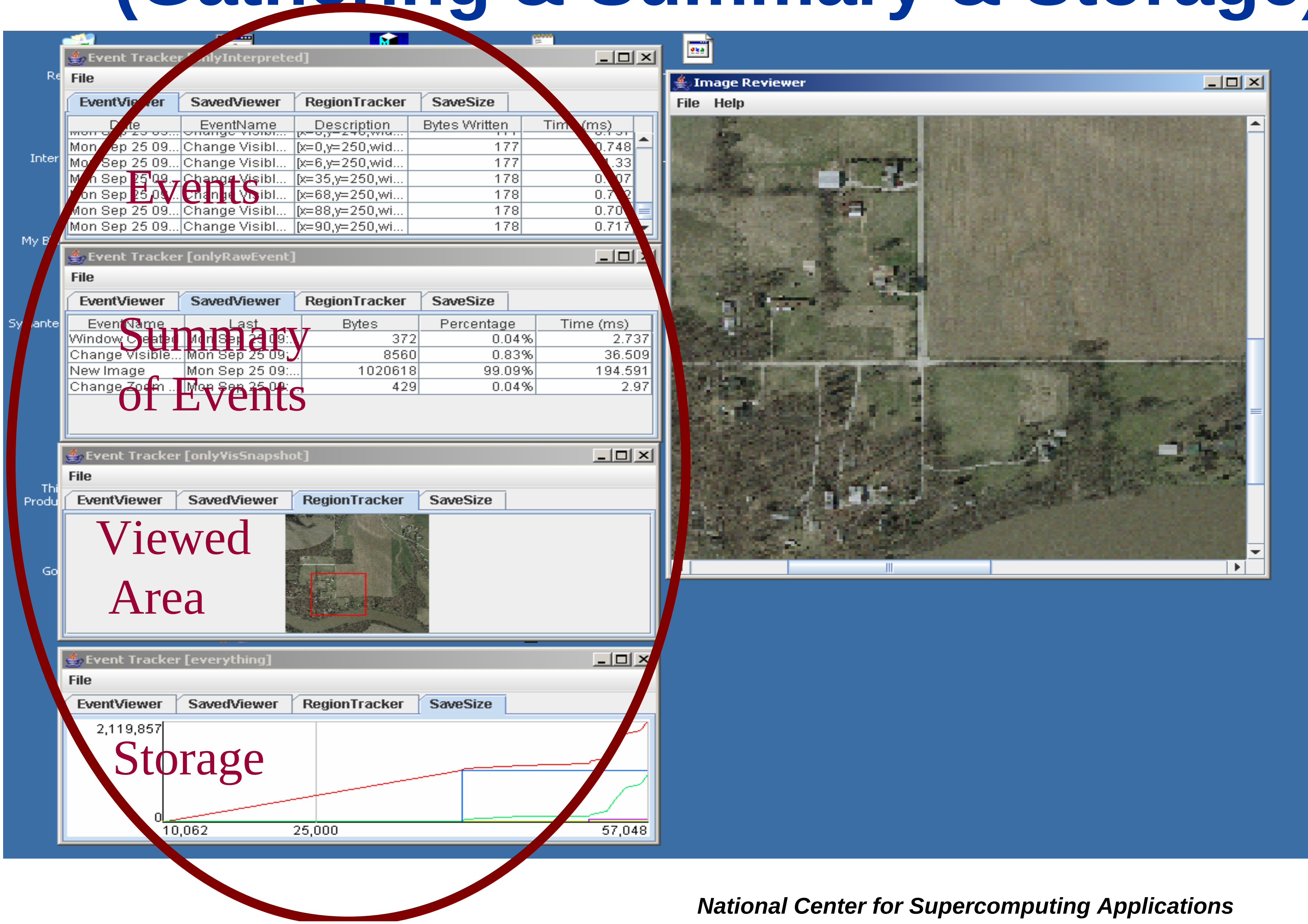
# Developed Simulation Framework

- **1. Image reviewer (Decision Process)**
  - Contains Image Frame and Image Panel, and includes all image manipulation functionality
- **2. Image event category tracker (Gathering & Summaries & Storage & Reporting)**
  - Tracks events in Image reviewer, allows setting preferences for information gathering and storage, lists event activities, summarizes event activities, displays inspected area and displays computational requirements
  - Enables to generate reports with protected authentic information about decision processes, speeds up reporting and provides a foundation for tracking versions of reports
- **3. Image event viewer (Retrieval & Reconstruction & Learning)**
  - Retrieves gathered and stored information by Image event category tracker, reconstructs processes with selected information granularity, and displays hierarchy of events and replays image inspection events according to the available information granularity

# 1. Image Reviewer (Decision Process)



# 2. Image Event Category Tracker (Gathering & Summary & Storage)




National Center for Supercomputing Applications

# 2. Image Event Category Tracker (Semi-Automatic Reporting)

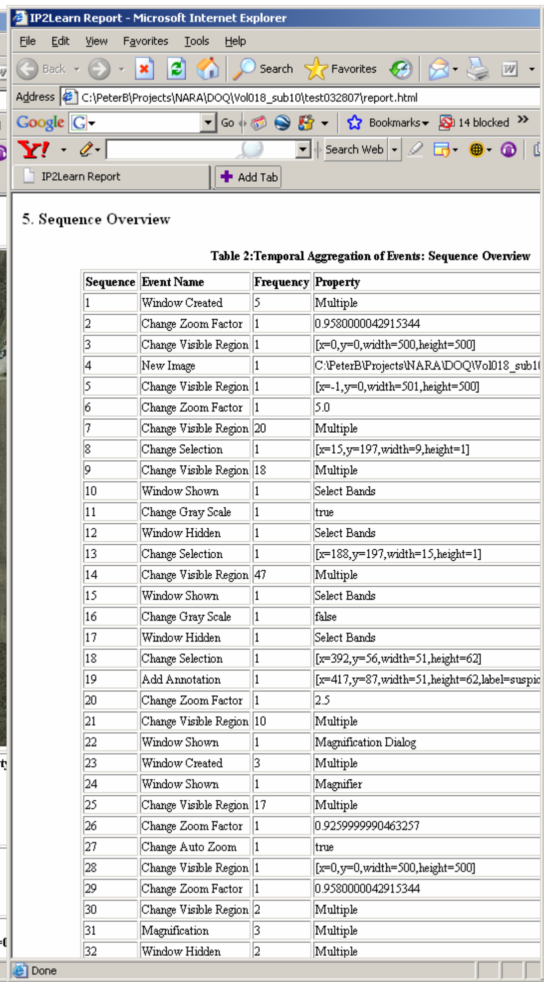
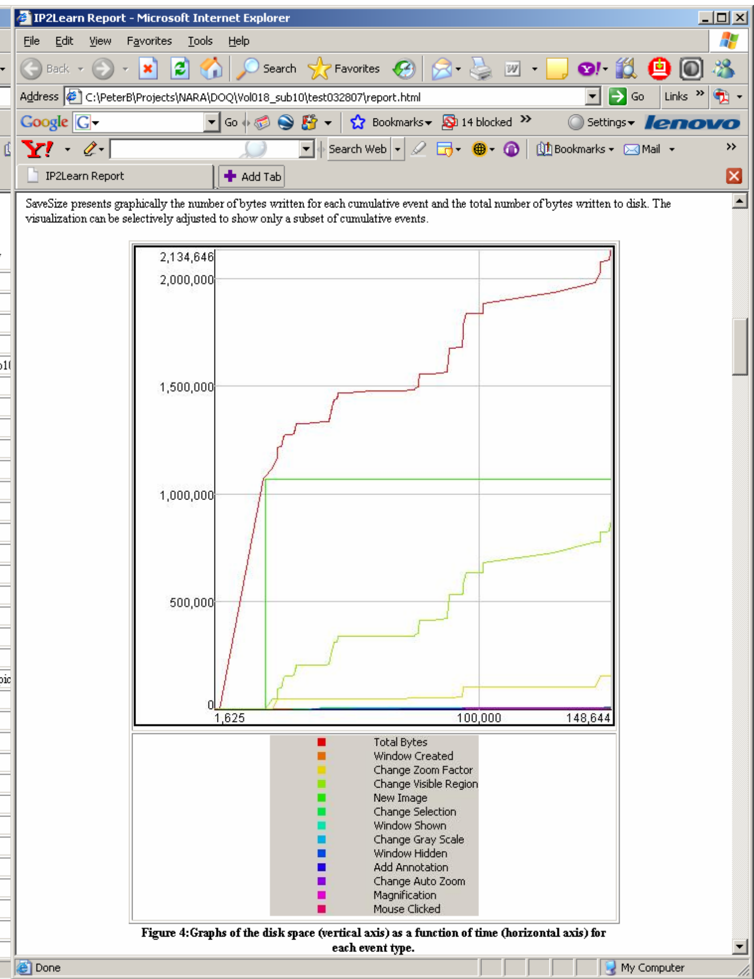
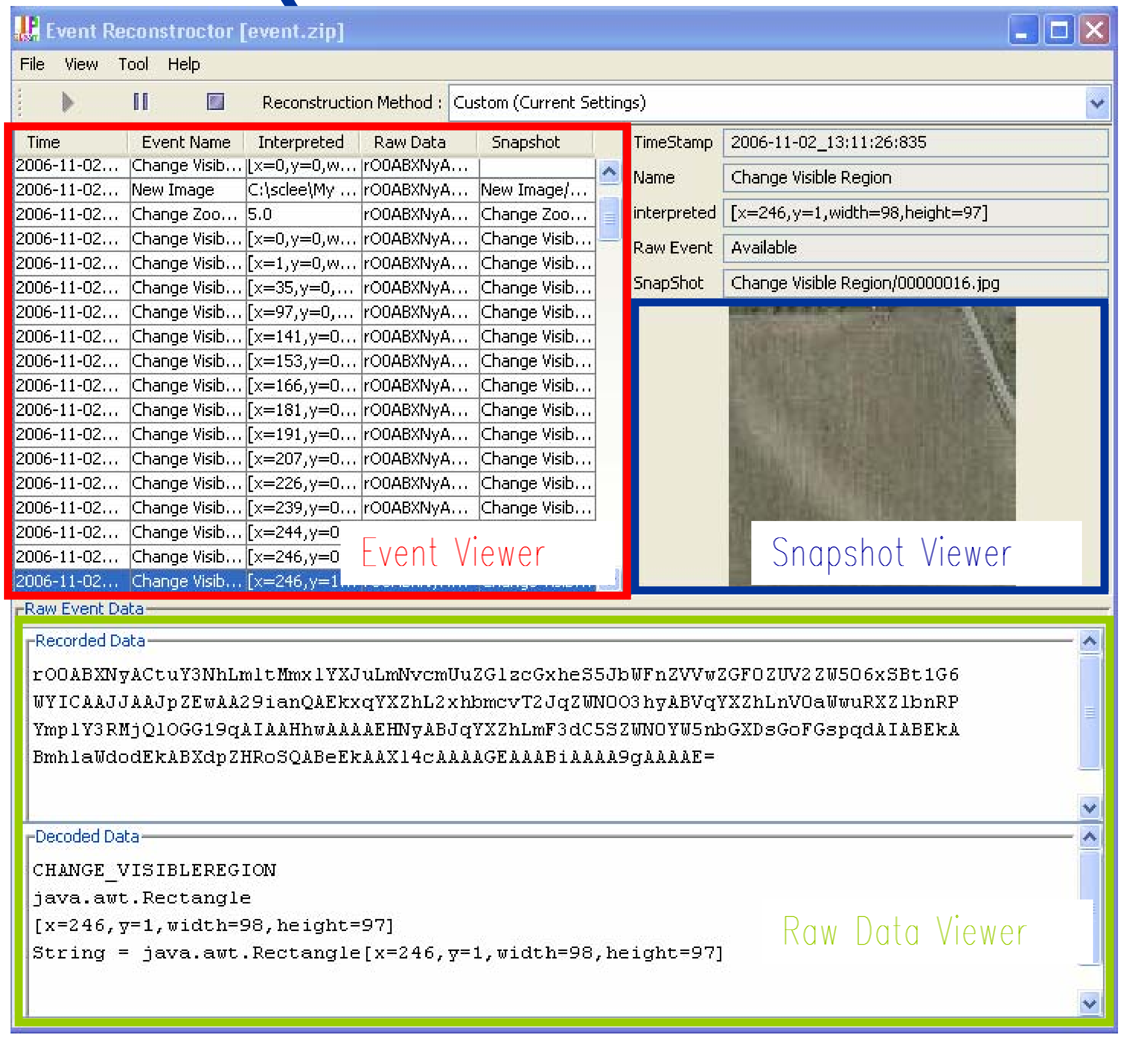


5. Sequence Overview

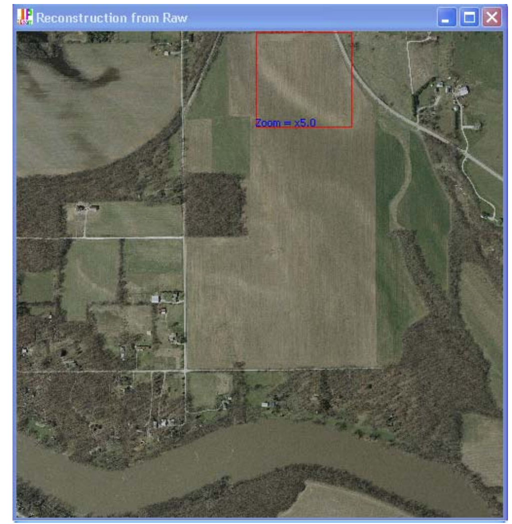
Table 2:Temporal Aggregation of Events: Sequence Overview

| Sequence | Event Name | Frequency | Property |
|---|---|---|---|
| 1 | Window Created | 5 | Multiple |
| 2 | Change Zoom Factor | 1 | 0.9580000042915344 |
| 3 | Change Visible Region | 1 | [x=0,y=0,width=500,height=500] |
| 4 | New Image | 1 | C:\PeterB\Projects\NARA\DOQ\Vol018_sub1( |
| 5 | Change Visible Region | 1 | [x=-1,y=0,width=501,height=500] |
| 6 | Change Zoom Factor | 1 | 5.0 |
| 7 | Change Visible Region | 20 | Multiple |
| 8 | Change Selection | 1 | [x=15,y=197,width=9,height=1] |
| 9 | Change Visible Region | 18 | Multiple |
| 10 | Window Shown | 1 | Select Bands |
| 11 | Change Gray Scale | 1 | true |
| 12 | Window Hidden | 1 | Select Bands |
| 13 | Change Selection | 1 | [x=188,y=197,width=15,height=1] |
| 14 | Change Visible Region | 47 | Multiple |
| 15 | Window Shown | 1 | Select Bands |
| 16 | Change Gray Scale | 1 | false |
| 17 | Window Hidden | 1 | Select Bands |
| 18 | Change Selection | 1 | [x=392,y=56,width=51,height=62] |
| 19 | Add Annotation | 1 | [x=417,y=87,width=51,height=62,label=suspic |
| 20 | Change Zoom Factor | 1 | 2.5 |
| 21 | Change Visible Region | 10 | Multiple |
| 22 | Window Shown | 1 | Magnification Dialog |
| 23 | Window Created | 3 | Multiple |
| 24 | Window Shown | 1 | Magnifier |
| 25 | Change Visible Region | 17 | Multiple |
| 26 | Change Zoom Factor | 1 | 0.9259999990463257 |
| 27 | Change Auto Zoom | 1 | true |
| 28 | Change Visible Region | 1 | [x=0,y=0,width=500,height=500] |
| 29 | Change Zoom Factor | 1 | 0.9580000042915344 |
| 30 | Change Visible Region | 2 | Multiple |
| 31 | Magnification | 3 | Multiple |
| 32 | Window Hidden | 2 | Multiple |

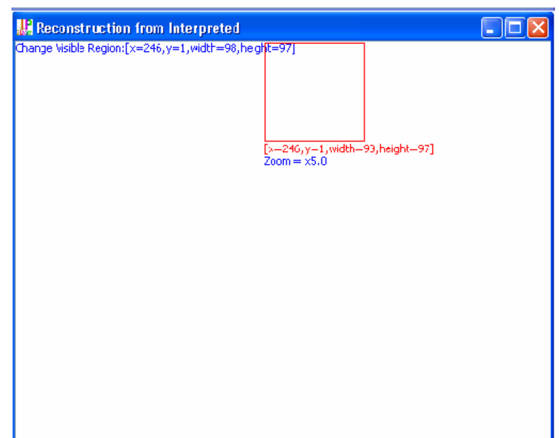


# 3. Image Event Viewer (Retrieval & Reconstruction)



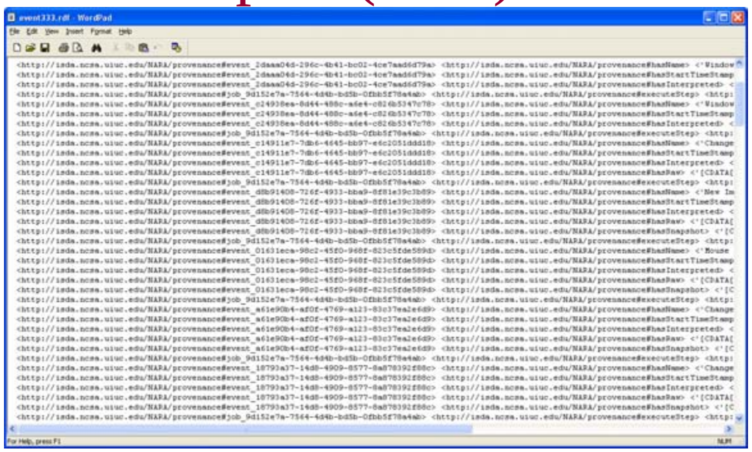

Reconstruction from Raw Data



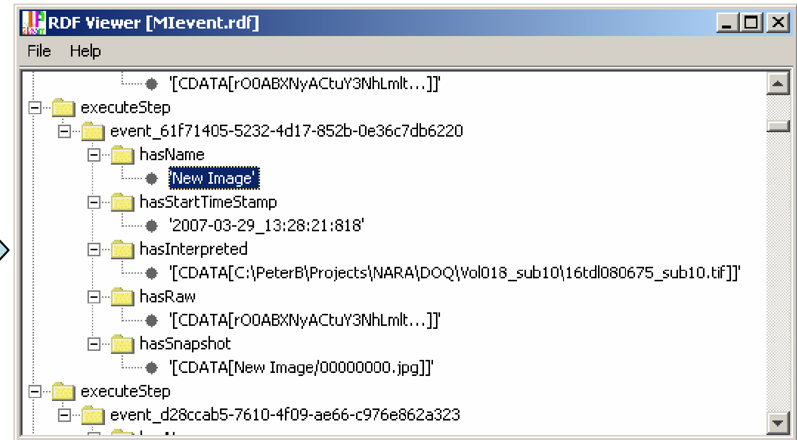

Reconstruction from Interpreted Data

# 3. Image Event Viewer (Learning)

RDF Triples (Raw)



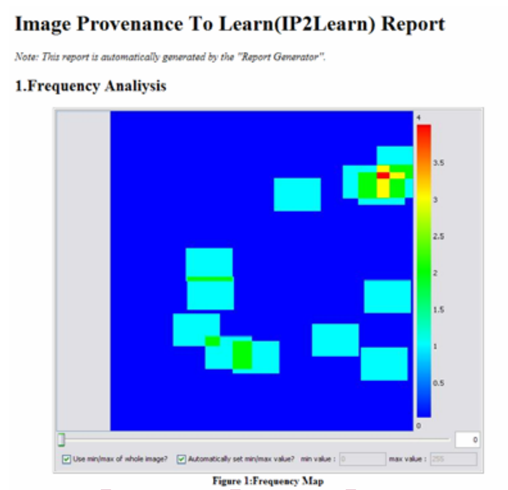
RDF Tree Viewer




Spatial aggregation of activities based on:

Time duration

Frequency of visits




Pseudo-colored version

# Example Tradeoff Studies
# or
# How to Use the Simulation Framework?

# Variables of Tradeoff Simulations (1)

- **Information granularity**
  – Categories of observable and measurable variables
- **Information gathering method**
  – Mechanisms how gather information about computer system activities ranging from hardware to user program levels
- **Information organization, representation and compression types**
  – Multiple 'standard ways' how to organize, represent and compress metadata
- **Types of information authentication**
  – Several purpose methods: encryption, watermarking or steganography

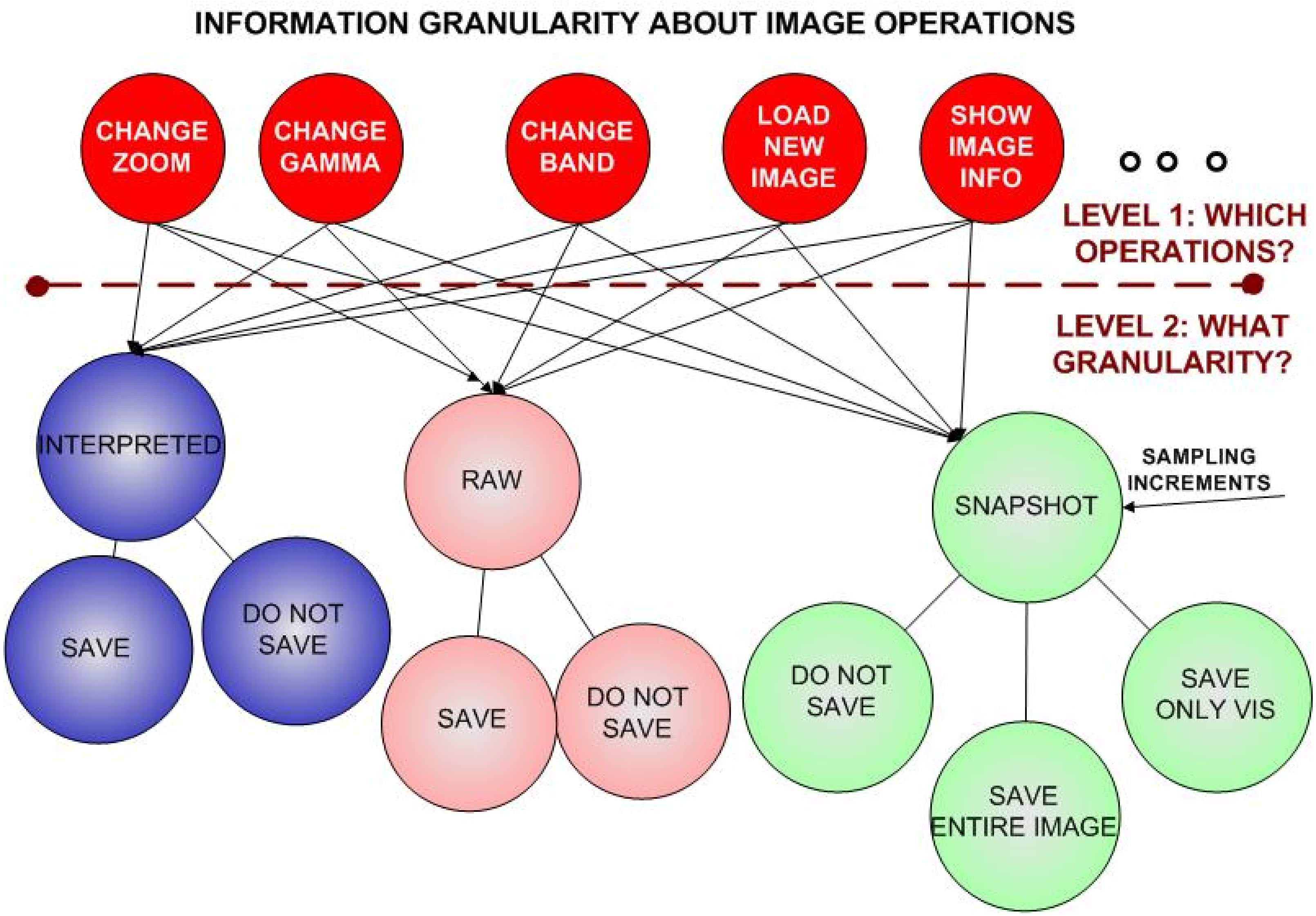
# Variables of Tradeoff Simulations (2)

- **Storage formats of gathered information and summary reports**
  - Multiple wide spread (standard?) formats of “final” reports
- **Information retrieval types**
  - Remote or local (distributed or centralized) retrieval, compressed or uncompressed files
- **Types of decision process reconstructions**
  - Activity levels ranging from human interaction to computer execution
- **Reconstruction value metrics**
  - Application dependent considerations about reconstruction value

# Storage vs. Granularity Tradeoffs

- **IP2Learn supports about 26 image inspection operations**
  - events during image inspection based decision processes
- **Three Major Information Granules**
  - **Interpreted** – what the programmer encoded as a textual description (interpretation) of an image operation
  - **Raw** – what the computer worked with when an image operation was recorded
  - **Snapshots** – what was rendered on the computer screen at the time of an image operation

# Information Granularity

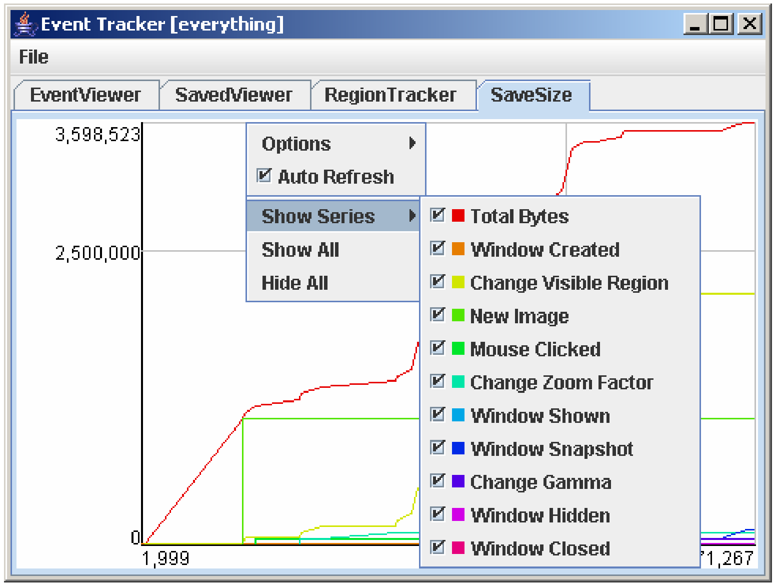


# Example of Information Granules

- **Example – Zoom image operation:**
  - The word “zoom” (interpreted),
  - Java event message reporting zoom (raw),
  - snapshot of zoomed image (snapshot)
- **Actual data:**
  - Interpreted zoom event:
    - Change Zoom Factor 2.5
  - Raw zoom event:
    - Recorded Java event about zoom operation:
      - rO0ABXNyACtuY3NhLmltMmxlYXJuLmNvcmUuZGlzcGxheS5JbWFnZVVwZGF0ZUV2ZW506xSBt1G6WYICAAJJAAJpZEwAA29ianQAEkxqYXZhL2xhbmcvT2JqZWN0O3hyABVqYXZhLnV0aWwuRXZlbnRPYmplY3RMjQlOGG19qAIAAHhwAAAABHNyABBqYXZhLmxhbmcuRG91YmxlgLPCSilr+wQCAAFEAAV2YWx1ZXhyABBqYXZhLmxhbmcuTnVtYmVyhqyVHQuU4IsCAAB4cEAEAAAAAAAA
    - Decoded Java event about zoom operation:
      - CHANGE_ZOOMFACTOR String = 2.5
  - Snapshot of zoom event:
    - Change Zoom Factor/00000021.jpg

# Storage vs. Granularity Test Case

- **Example Image Inspection Sequence**
  - Defined Task: find a house in a large image located at (machine + directory + file name)
  - Image Inspection Sequence:
    - load a file ->
    - zoom to 50% so that the entire image could fit into a viewable window ->
    - then zoom to 250% so that with the help of a scroll bar the house location can be found ->
    - adjust the gamma intensity value for better visualization.




Storage = function of time, image operation and its granularity

# Test Case Conclusions (1)

- **10 out of 26 image operations were used in this test case**
- **When <u>all information granularities</u> were recorded then Change Visible Region (59.55%) and New Image (29.88%) events dominated the storage requirements of about total 3,600,000 bytes.**
- **When <u>only 'interpreted' information</u> was recorded then Change Visible Region (70.26%) event required the most storage space and the total space was only about 10,600 bytes.**
- **When <u>only 'raw' information</u> was recorded then New Image (97.95%) event dominated the total of 1,000,000 bytes of storage.**
- **When <u>only 'snapshot' information</u> was recorded then Change Visible Region (54.74%), Window Snapshot (27.26%) and Change Zoom Factor (11.30%) were the main contributors to the storage requirements from the total of 547,000 bytes.**

# Test Case Conclusions (2)

- **Estimations of relative storage and execution time requirements:**
- **The ratios for ‘only interpreted’ to ‘only raw’ to ‘only snapshot’ to ‘everything’ are**
  - (a) approximately 1 : 100 : 55 : 360 for storage requirements and
  - (b) about 1 : 3.46 : 16.49 : 81.42 for execution time requirements.

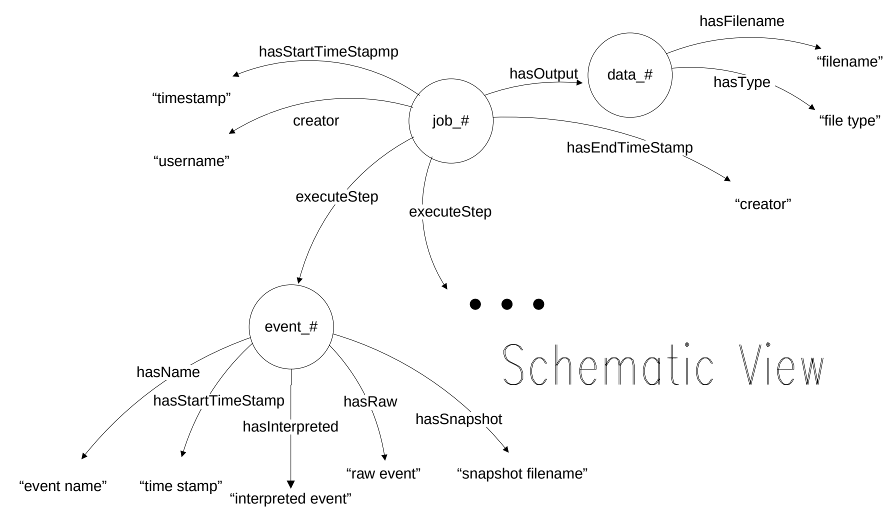
# Storage vs. Information Organization Tradeoffs


```
- <event name="Change Zoom Factor">
  <timestamp>2006-10-26_14:16:16:581</timestamp>
- <interpreted>
- <![CDATA[ 5.0 ]]>
  </interpreted>
- <rawevent>
- <![CDATA[
      rO0ABXNyACtuY3NhLmltMmxlYXJuLmNvcmUuZGlzcGxheS5Jb
      WFnZVVwZGF0ZUV2ZW506xSBt1G6WYICAAJJAAJpZEwAA29
      ianQAEkxqYXZhL2xhbmcvT2JqZWN0O3hyABVqYXZhLnV0aWw
      uRXZlbnRPYmplY3RMjQlOGG19qAIAAHhwAAAABHNyABBqYX
      ZhLmxhbmcuRG91YmxlgLPCSilr+wQCAAFEAAV2Wx1ZXhyAB
      BqYXZhLmxhbmcuTnVtYmVyhqyVHQuU4IsCAAB4cEAUAAAAA
      AAA  ]]>
  </rawevent>
  <snapshot file="Change Zoom Factor/00000003.jpg" />
  </event>
- <event name="Change Visible Region">
  <timestamp>2006-10-26_14:16:16:800</timestamp>
- <interpreted>
- <![CDATA[ [x=0,y=0,width=97,height=97] ]]>
  </interpreted>
- <rawevent>
- <![CDATA[
      O0ABXNyACtuY3NhLmltMmxlYXJuLmNvcmUuZGlzcGxheS5Jb
      WFnZVVwZGF0ZUV2ZW506xSBt1G6WYICAAJJAAJpZEwAA29
      ianQAEkxqYXZhL2xhbmcvT2JqZWN0O3hyABVqYXZhLnV0aWw
      uRXZlbnRPYmplY3RMjQlOGG19qAIAAHhwAAAAEHNyABJqYX
      ZhLmF3dC5SZWN0YW5nbGXDsGoFGspqdAIABEkABmhlaWdo
      dEkABXdpZHRoSQABeEkAAXl4cAAAAGEAAABhAAAAAAAAA
      AA=  ]]>
  </rawevent>
  <snapshot file="Change Visible Region/00000004.jpg" />
  </event>
```


**Key Pairs**

Information Organization

**Resource Description Framework (RDF) Triplets**



# XML Key Pairs: Example

- **Key Pairs (tag, value):**
  - Example:
    - <**event name**="New Image">
    - ~(**tag1 = event name**; value = New Image)

    - <**timestamp**>2007-03-29_13:28:21:818</timestamp>
    - <**interpreted**> <![CDATA[ C:\PeterB\Projects\NARA\DOQ\Vol018_sub10\16tdl080675_sub10.tif ]]> </interpreted>
    - <**rawevent**> <![CDATA[ rO0ABXNyACtuY3NhLmltMmxlYXJuLmNvcmUuZGlzcGxheS5JbWFnZVVwZGF0ZUV2ZW506xSBt1G6…………………d254 ]]> </rawevent>
    - <**snapshot file**="New Image/00000000.jpg" />

# RDF Triples: Example

- **Resource Description Framework (RDF)**
  - Representation: **Subject – Predicate - Object** format
- **Example:**
  - Subject = event ID: http://isda.ncsa.uiuc.edu/NARA/provenance#event_61f71405-5232-4d17-852b-0e36c7db6220
  - Predicate: <http://isda.ncsa.uiuc.edu/NARA/provenance#hasName>
  - Object: <'New Image'>;

# Storage Requirements vs. Storage Format Tradeoffs

## Storage

- **Format:**
  - Text: XML
  - Images: JPG or TIFF
- **Compression**:
  - Text and Images with or without ZIP

| **Tradeoff Study Environment** | **Specifications** |
|---|---|
| Java Virtual Machine | Version 1.6.0 (build 1.6.0_01-b06) |
| Operating System | Microsoft Windows XP Version 2002, Service Pack 2 |
| Hardware | CPU: Intel Pentium 4 (Northwood) 3.00GHz<br>Memory: 1GBytes |

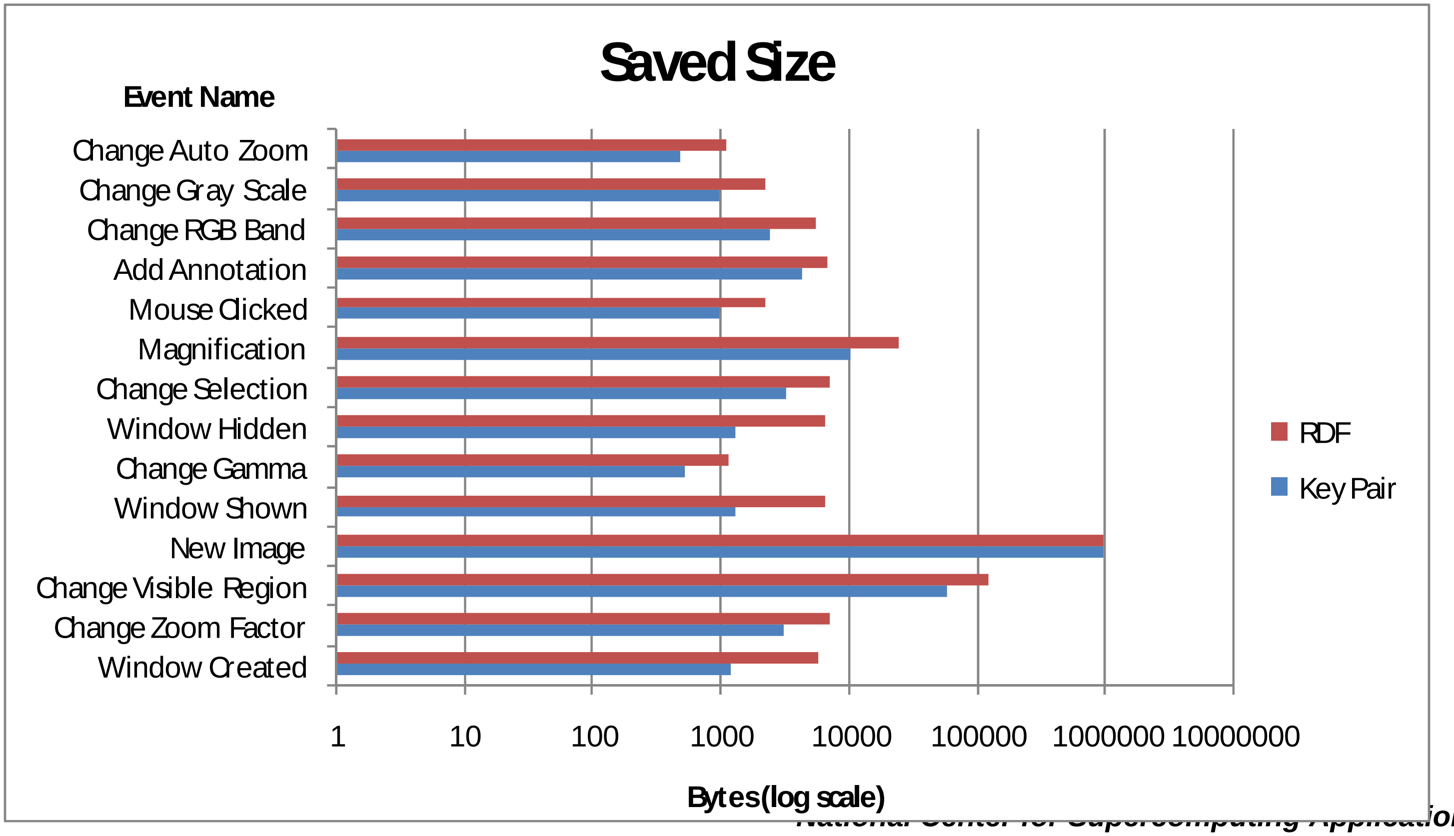
# Test Sequence

| Event Name | Frequency | Property |
|---|---|---|
| New Image | 1 | C:\sclee\My Documents\data\NARA\image\16tdl065660_sub10.tif |
| Change Gamma | 1 | 1.1 |
| Window Shown | 1 | Image Info |
| Change Selection | 1 | [x=50,y=164,width=50,height=52] |
| Magnification | 21 | Multiple |
| Change Selection | 1 | [x=403,y=42,width=77,height=78] |
| Mouse Clicked | 2 | Multiple |
| Add Annotation | 1 | [x=403,y=42,width=77,height=78,label=null,type=2] |
| Add Annotation | 1 | [x=50,y=166,width=40,height=38,label=null,type=2] |
| Add Annotation | 1 | [x=98,y=259,width=109,height=163,label=null,type=2] |
| Add Annotation | 1 | [x=131,y=4,width=51,height=39,label=null,type=2] |
| Change RGB Band | 1 | [R=0,G=-1,B=-1] |
| Change RGB Band | 1 | [R=-1,G=1,B=-1] |

| Event Name | Frequency | Property |
|---|---|---|
| Change RGB Band | 1 | [R=-1,G=-1,B=2] |
| Change RGB Band | 1 | [R=-1,G=-1,B=-1] |
| Change Gray Scale | 1 | true |
| Change RGB Band | 1 | [R=0,G=1,B=2] |
| Change Gray Scale | 1 | false |
| Change Zoom Factor | 1 | 1.0 |
| Change Visible Region | 1 | [x=0,y=0,width=500,height=500] |
| Change Zoom Factor | 1 | 5.0 |
| Change Visible Region | 67 | Multiple |
| Change Selection | 1 | [x=464,y=453,width=14,height=0] |
| Change Visible Region | 32 | Multiple |
| Change Zoom Factor | 1 | 1.5479999780654907 |
| Change Auto Zoom | 1 | true |
| Change Visible Region | 1 | [x=0,y=-1,width=500,height=501] |
| Change Zoom Factor | 1 | 1.5779999494552612 |
| Change Visible Region | 1 | [x=0,y=0,width=501,height=501] |

# Storage vs. Information Organization Tradeoffs: Test Case

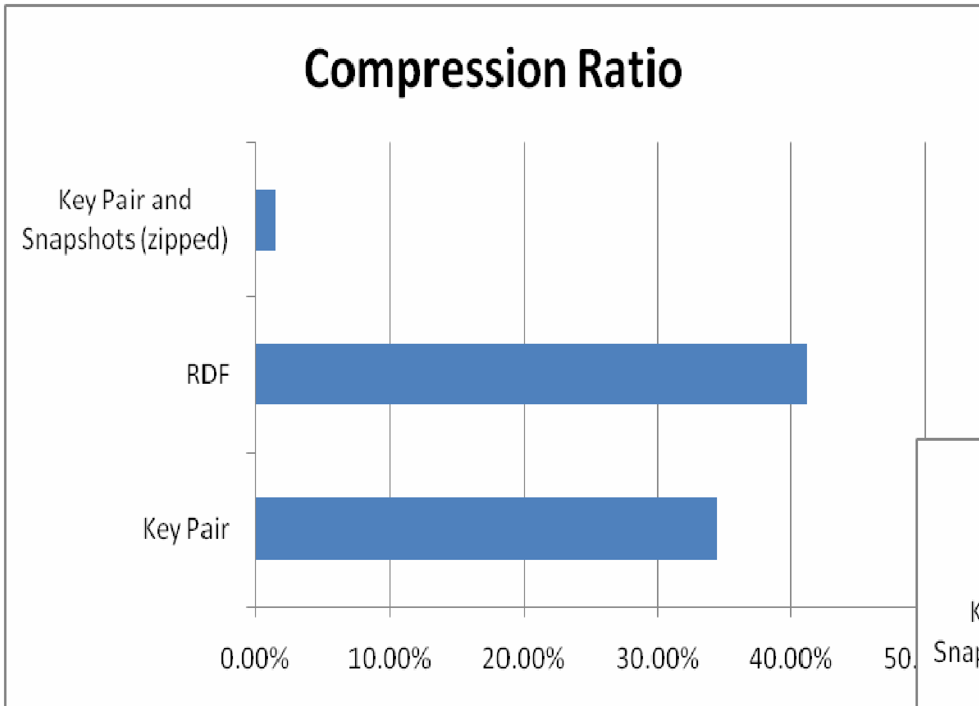
- **Information granules include interpreted, raw and snapshots**
- **Files were not compressed**

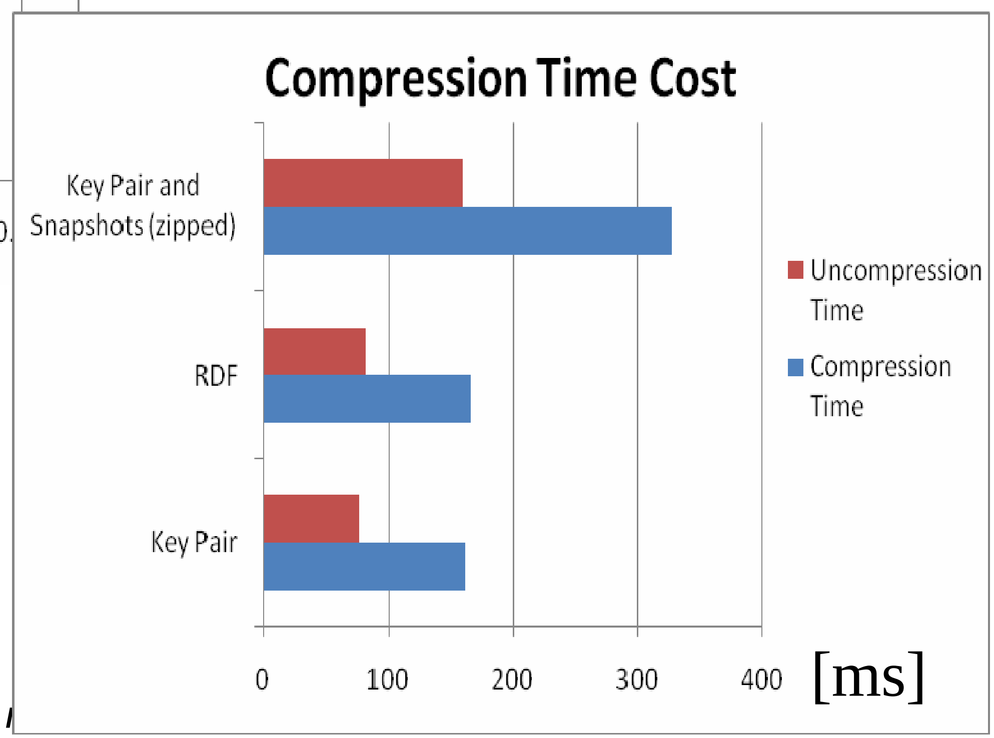




NCSA

# Storage vs. Compression Test Case



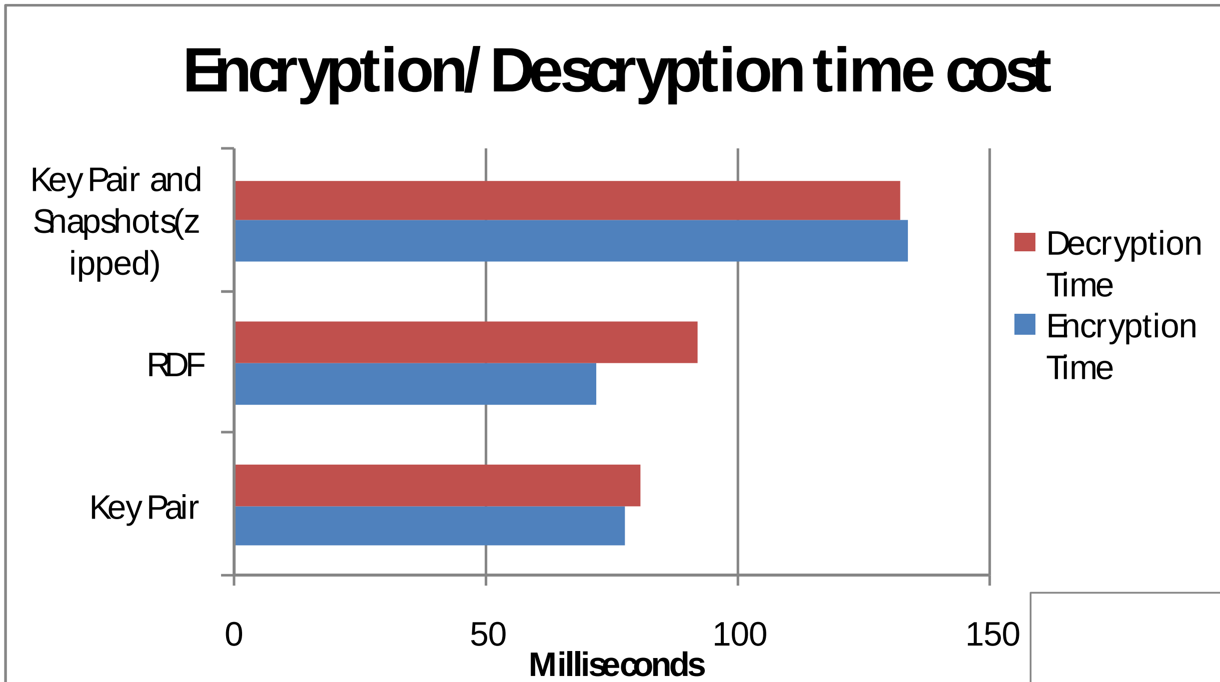
# Test Case Conclusions

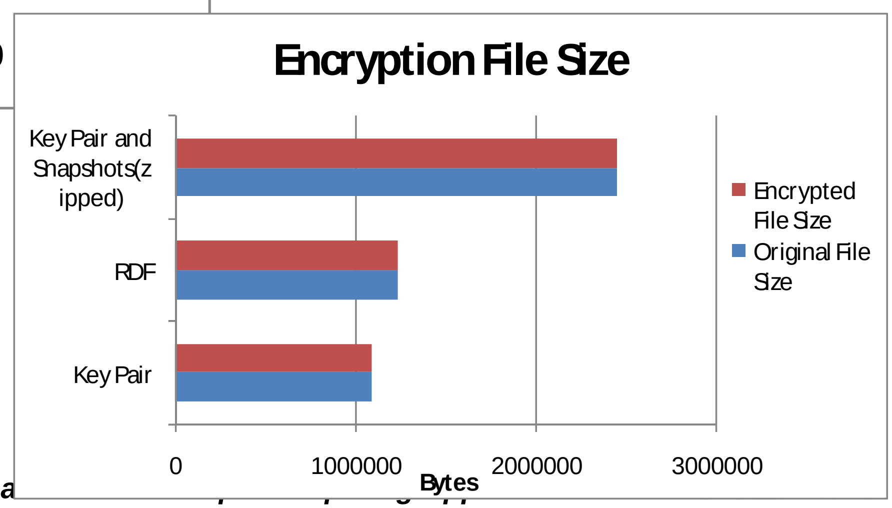
- Storage Requirements (RDF) > Storage Requirements (KeyPairs)
- Compression ratio (RDF) > Compression ratio (KeyPairs)
- Zipped (KeyPairs and Images) have a small compression ratio
- Uncompressing RDF or KeyPairs takes the same amount of CPU time and it is much less than zipped files
- Storage Requirements (Load New Image) > Storage Requirements (Other events)
- On average, all other events but “Load New Image” took about 2K-5KB while “Load New Image” took about 1MB (200-500 times more)

# Authentication of Records

- **Records to Authenticate:**
  - gathered provenance information
  - summary reports in HTML (Hypertext Markup Language) or in PDF (Portable Document Format) format
- **Methods:**
  - Encryption - following the Advanced Encryption Standard
  - Watermarking – labeling every image with **visible** text
  - Steganography – inserting **invisible** text or image into preserved images
- **Computational requirements of these methods**
  - The same test case (image inspection sequence) as before

# Computational Requirements vs. Authentication Method Tradeoffs



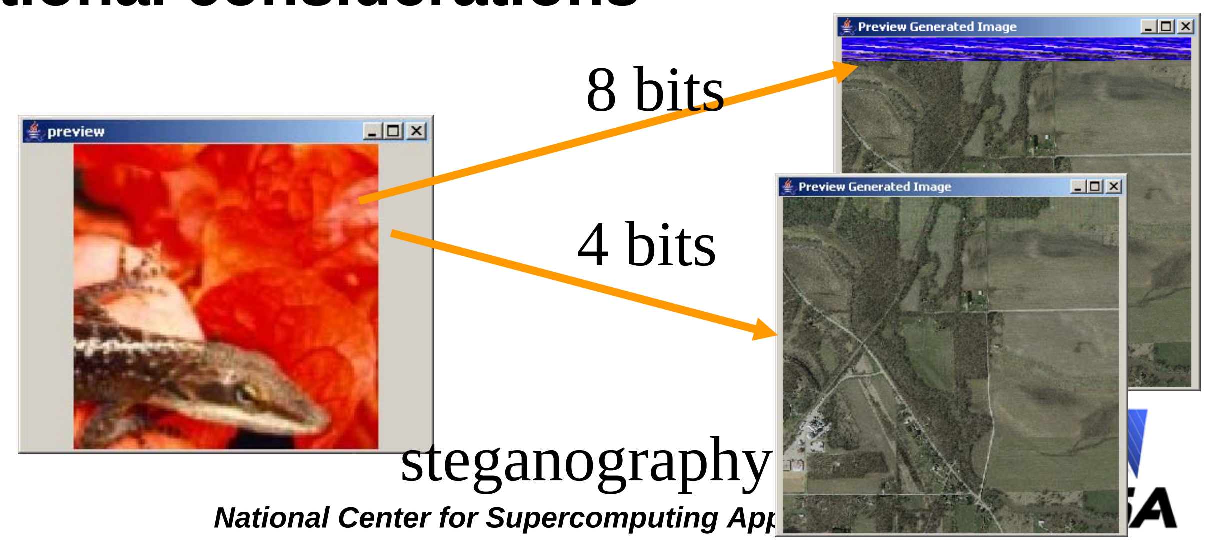




*Nationa*

# Test Case Conclusions

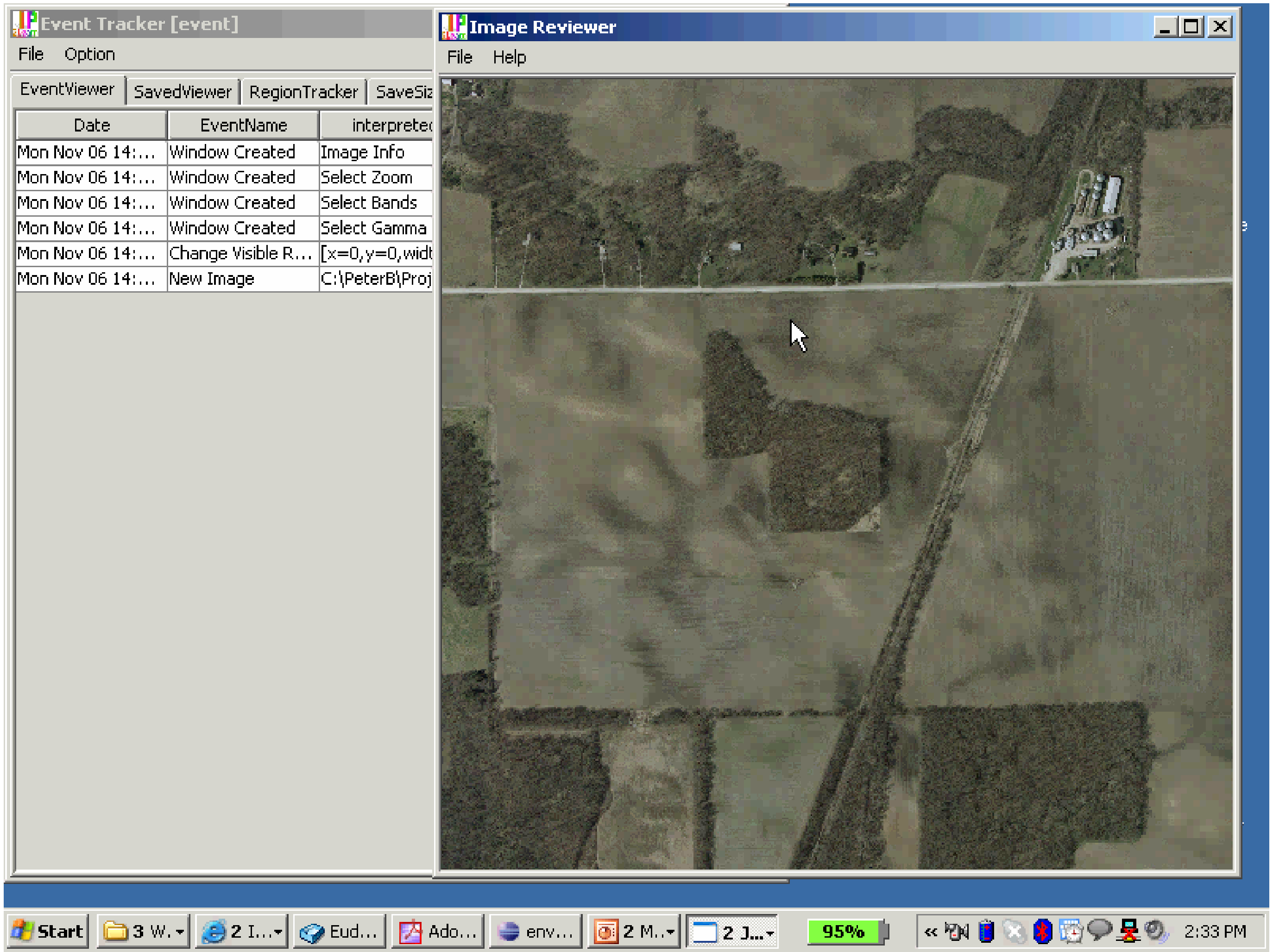
- File size does not change with encryption/decryption (as expected)
- Decryption CPU time requirements are slightly higher than the encryption CPU time requirements
- The largest difference between encryption and decryption CPU time is for files with RDF information organization
- Comparison of CPU requirements of authentication methods needs additional considerations


watermarking

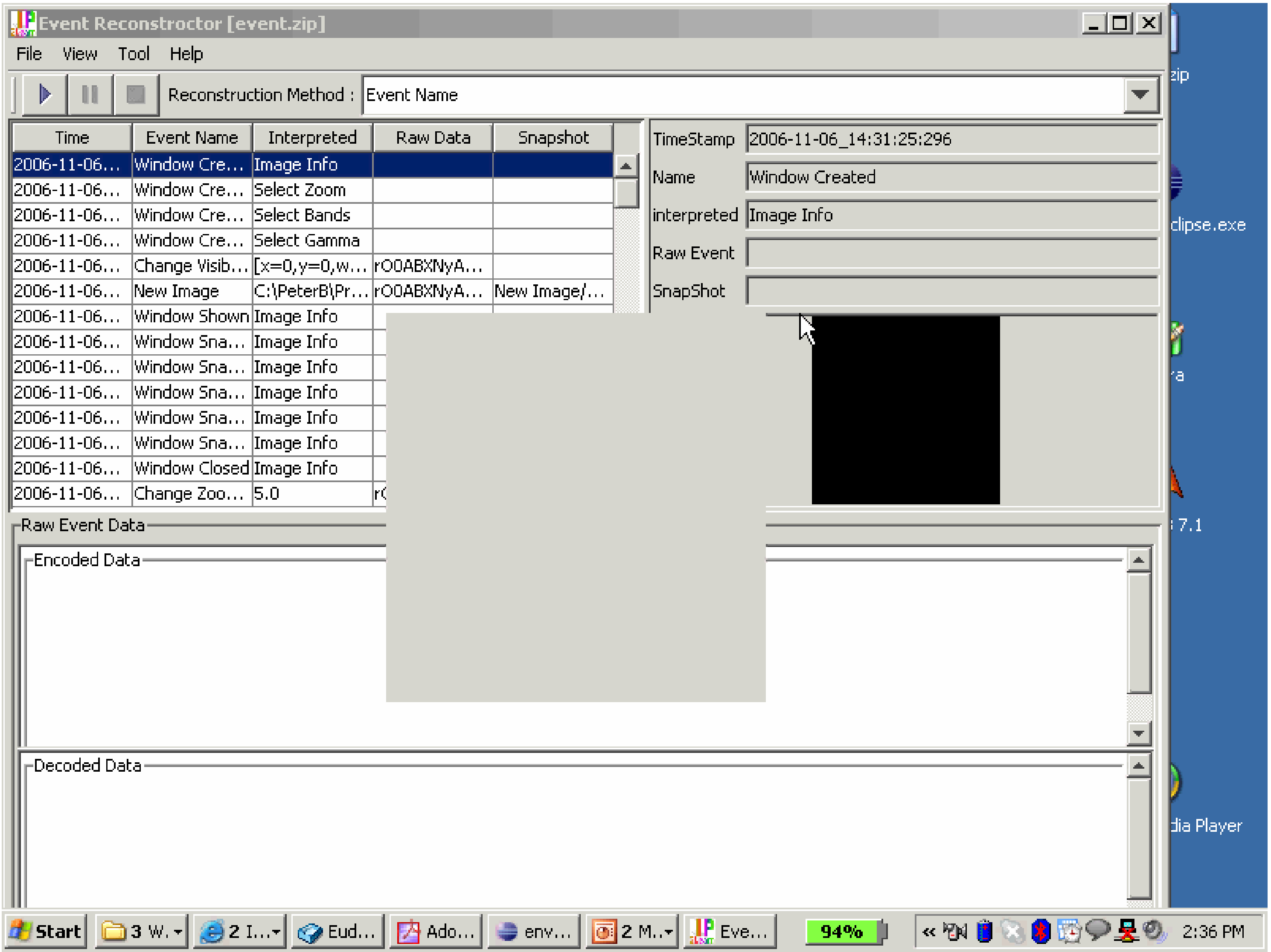



steganography

# Movie: Image Reviewer and Event Tracker



# Movie: Image Event Viewer



# Discussion

- **Simulation Framework for Understanding Preservation and Reconstruction Requirements of Future Computer Assisted Decision Processes**
  - Image Provenance To Learn (IP2Learn) software with sample data
  - Download: open source & documentation: http://isda.ncsa.uiuc.edu/download/
- **Comments? Feedback?**
- **POC: Peter Bajcsy, pbajcsy@ncsa.uiuc.edu**