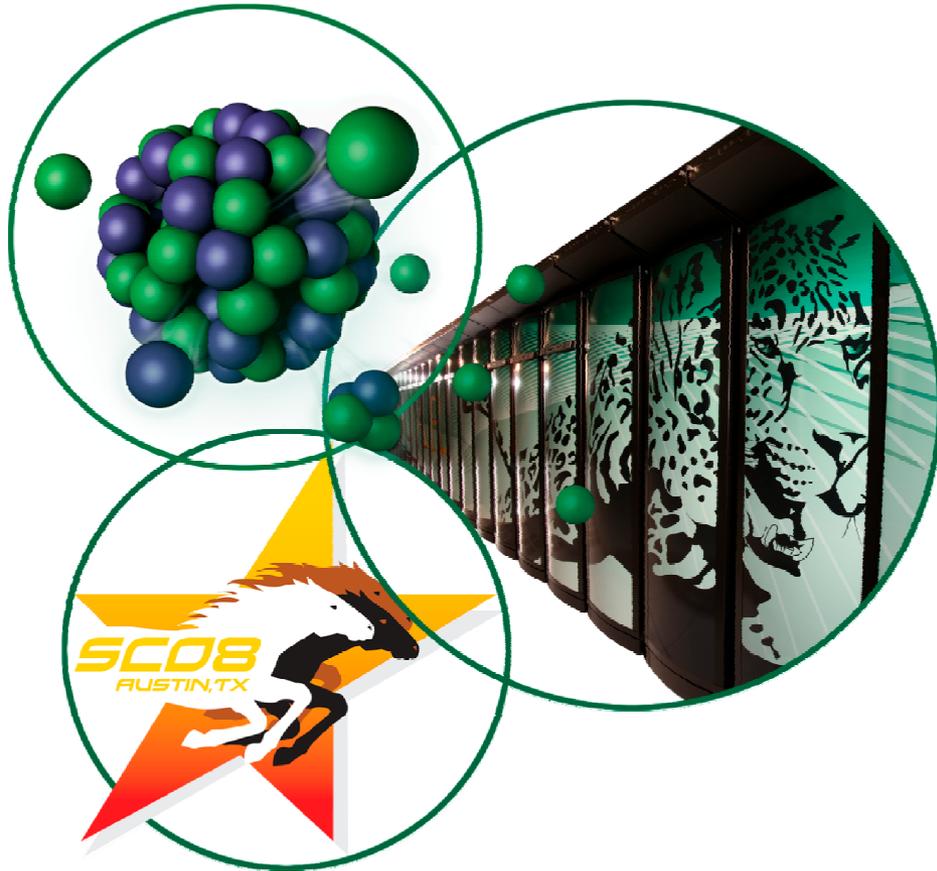


# Data Analysis and High-Performance Computing

Presented by

**Justin M. Beaver**  
**Yu (Cathy) Jiao**

Applied Software Engineering  
Research Group  
Computational Sciences and  
Engineering Division



# Streaming data analysis challenges

- High-volume data streams are constantly generated
- Traditional data-encoding scheme is inefficient
- Need a new solution to handle incremental clustering

## Sensors



## Multimedia

### Image



### Text

One small step for man

### Binary

11010010

### Data

1970

1980

1990

2000

2010

# The

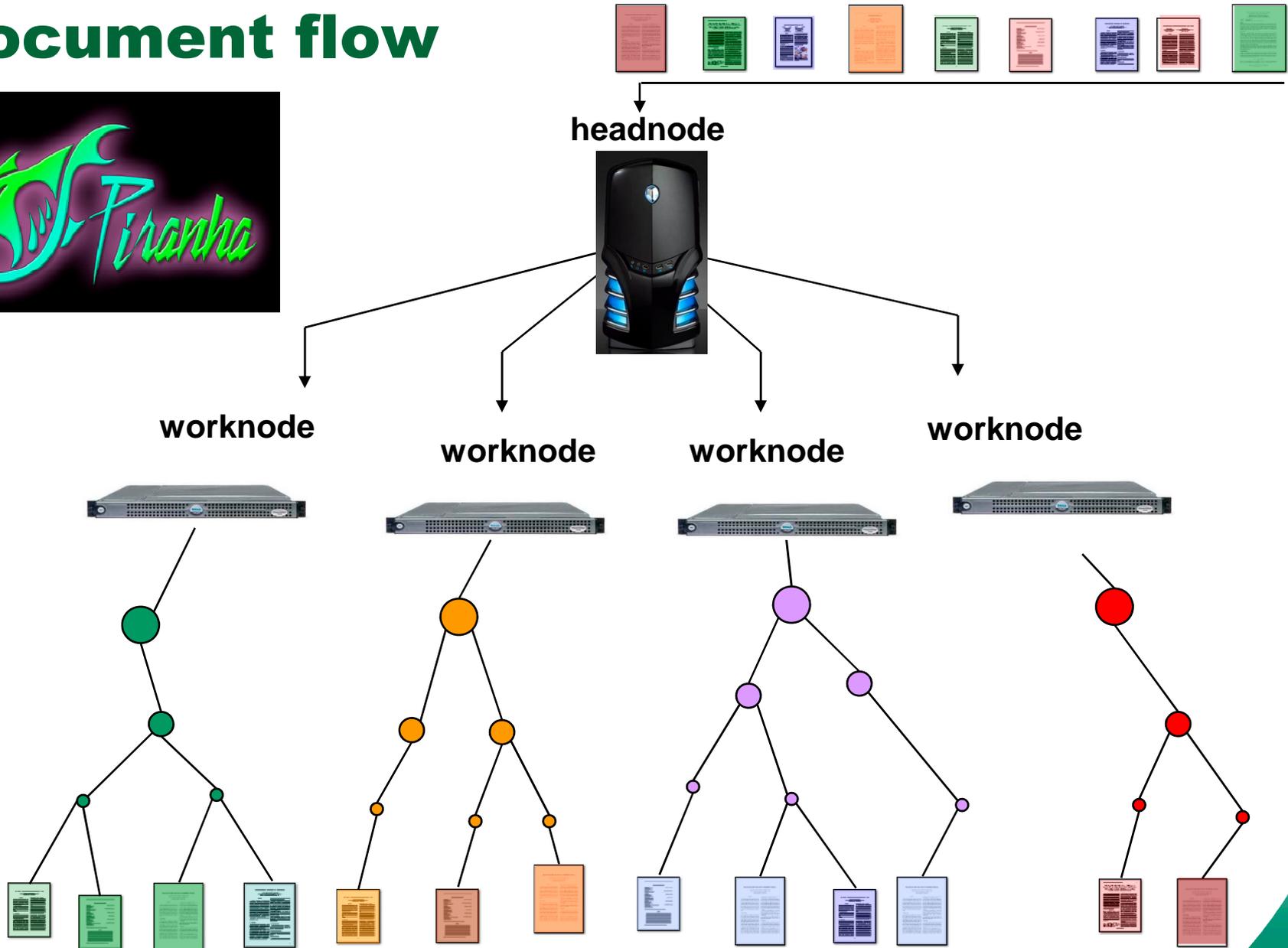


# approach

- Piranha utilizes distributed and parallel data clustering to process data streams
- Piranha applies a novel data-encoding scheme, Term Frequency-Inverse Corpus Frequency (TF-ICF)
- Piranha handles incremental clustering using a threshold-based solution
- Ability to search AND analyze
  - Organize documents based on content
  - Identify similar and dissimilar documents
  - Identify duplicate and near-duplicate data
- 2007 R&D 100 Award winner
  - Awards are based on each achievement's technical significance, uniqueness, and usefulness compared to competing projects and technologies



# Document flow



# Document flow



headnode

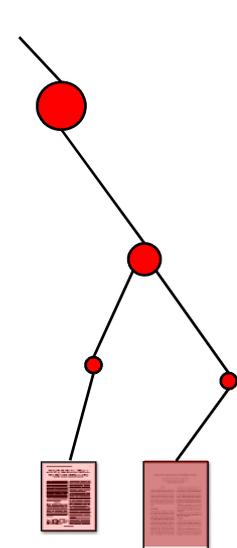
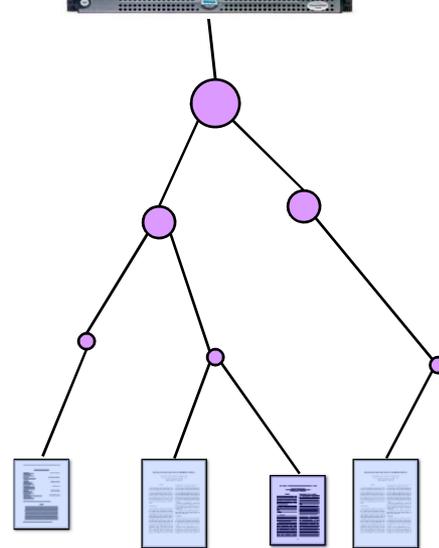
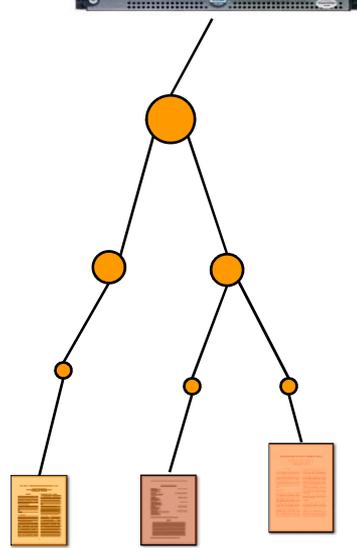
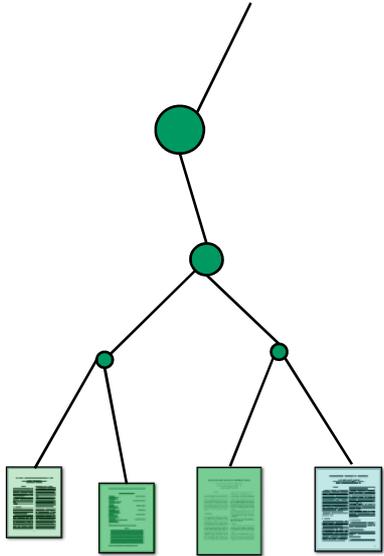


worknode

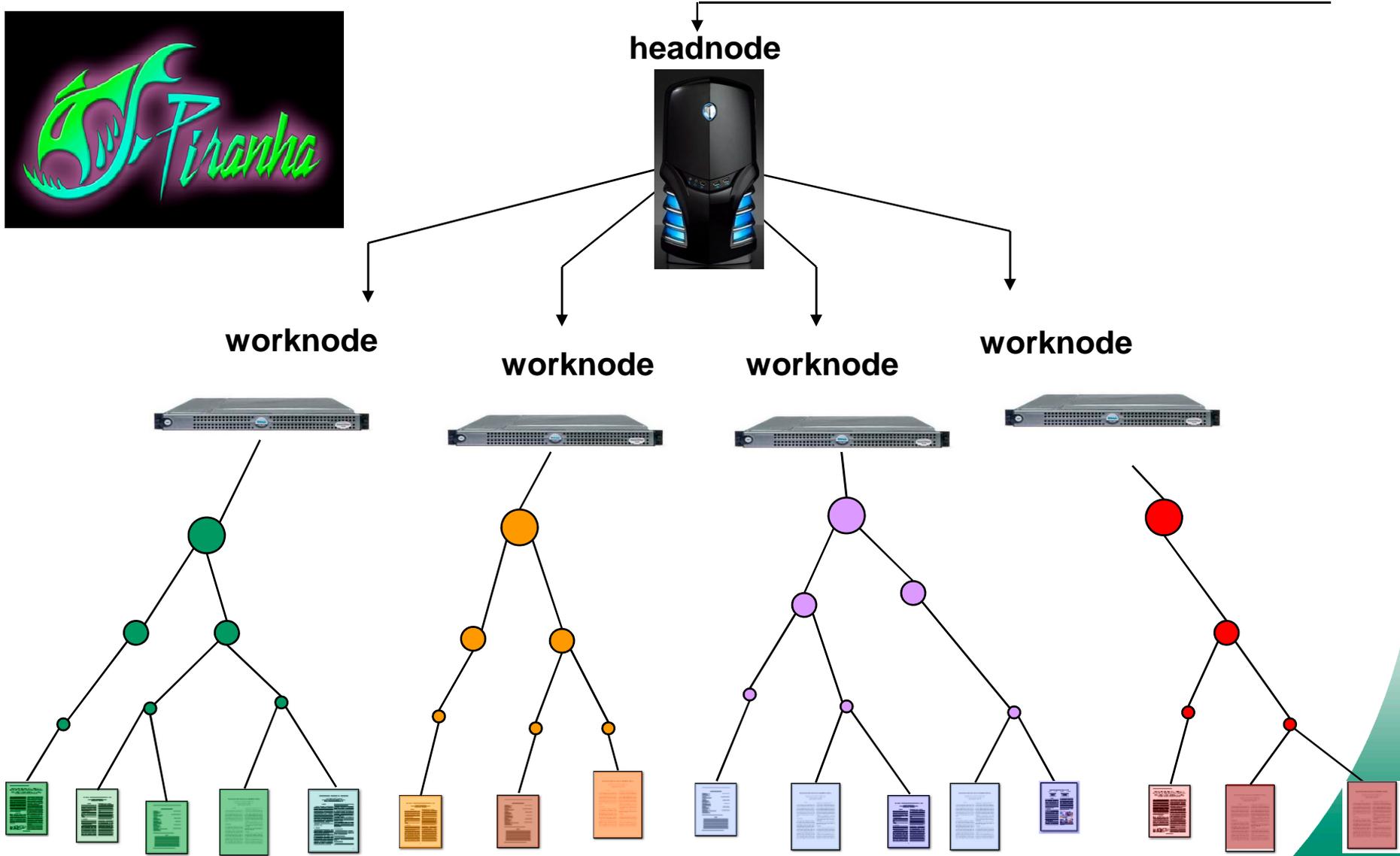
worknode

worknode

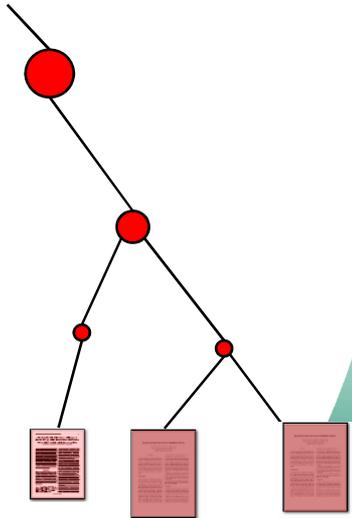
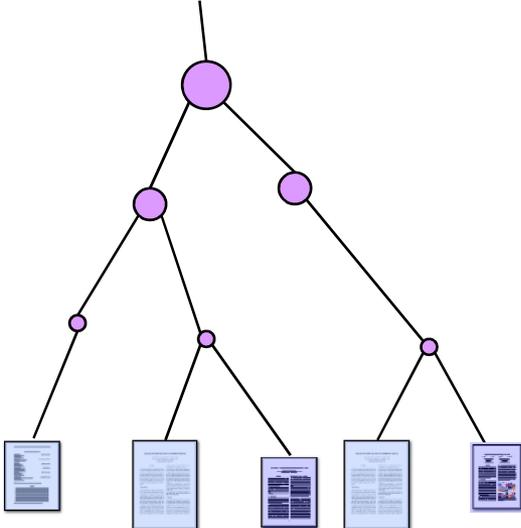
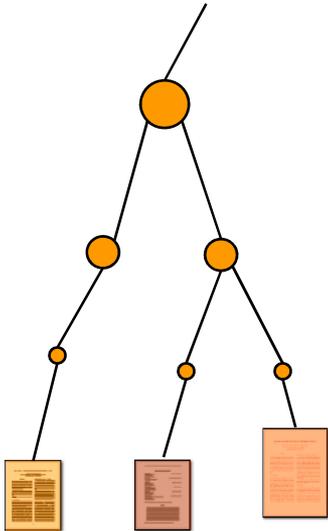
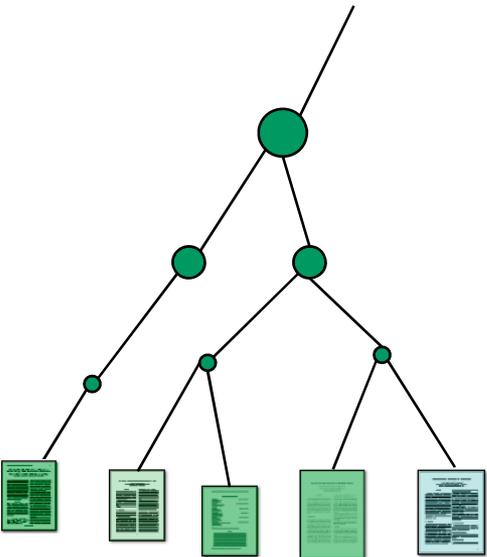
worknode



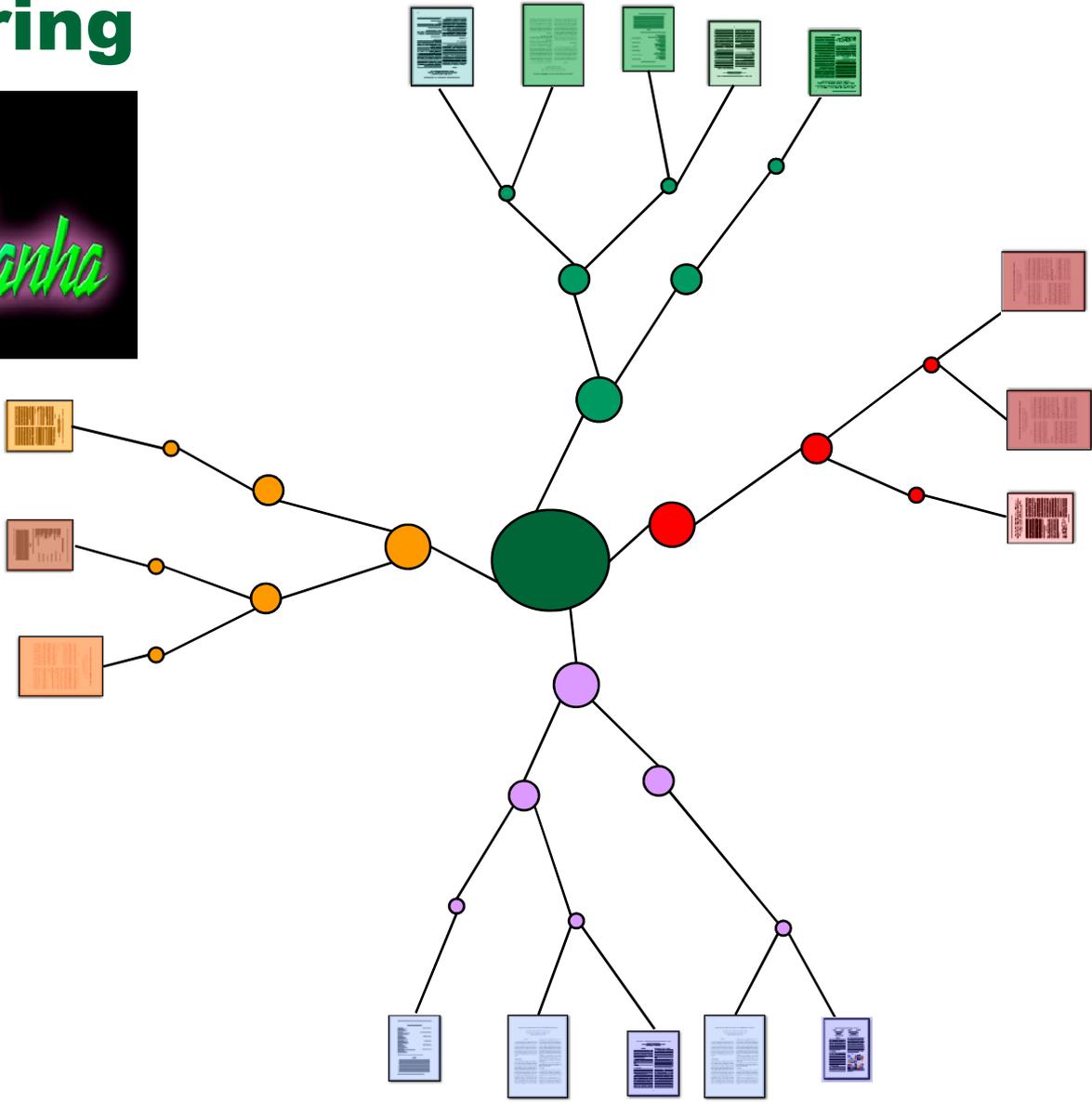
# Document flow



# Clustering



# Clustering





# Piranha single-term word analysis

Piranha Client : 2008 Presidential

File View Tools Settings

Help



## Collections

- Collections (28)
- ▶ obama (4)
- ▶ mccain (5)
- ▶ huckabee (5)
- ▶ edwards (5)
- ▶ clinton (5)
- ▶ giuliani (4)

## Top Collection Terms

- Top Collection Terms
- ▶ obama (11)
- ▶ mccain (10)
- ▶ huckabee (16)
- ▶ edwards (11)
- ▶ clinton (8)
- ▶ giuliani (9)

## Top Words

- Top Words (65)
- Top Words (65)
- ▶  hampshire (4)
- ▶  barack (3)
- ▶  edwards (3)
- ▶  huckabee (3)
- ▶  mccain (3)
- ▶  obama (3)
- ▶  arkansas (2)
- ▶  giuliani (2)
- ▶  giulianis, giuliani (2)
- ▶  huckabees, huckabee (2)
- ▶  isbn (2)
- ▶  mccain, mccains (2)
- ▶  obama, obamas (2)
- ▶  arkansans, arkansan (1)
- ▶  arkansass (1)
- ▶  authorized (1)
- ▶  azcentral (1)
- ▶  bracewell (1)
- ▶  buildmenurow (1)
- ▶  caretpos (1)
- ▶  cfr, cfrs (1)
- ▶  clinton (1)
- ▶  committee (1)
- ▶  cond (1)
- ▶  defaultconsole (1)
- ▶  dumonds, dumond (1)
- ▶  eustace (1)

## Top Phrases

## All Words

- All Words (13556)
- All Words (13556)
- ▶  john (13)
- ▶  health (12)
- ▶  act (11)
- ▶  care (11)
- ▶  foreign (11)
- ▶  hampshire (11)
- ▶  high (11)
- ▶  washington (11)
- ▶  budget (10)
- ▶  deductible (10)
- ▶  federal (10)
- ▶  george (10)
- ▶  january (10)
- ▶  leadership (10)
- ▶  military (10)
- ▶  rights (10)
- ▶  won (10)
- ▶  americans, american (9)
- ▶  bill (9)
- ▶  children (9)
- ▶  committee (9)
- ▶  david (9)
- ▶  defense (9)
- ▶  edwards (9)
- ▶  february (9)
- ▶  giuliani (9)

## All Phrases

## User Term Management

Add Stemmed Term

Remove Term

## User Terms

- User Terms (0)

Generate Document Node Cluster

Document Category:

Piranha Category:

View Original Do...

# Piranha multi-term phrase analysis

Piranha Client : 2008 Presidential

File View Tools Settings

Help

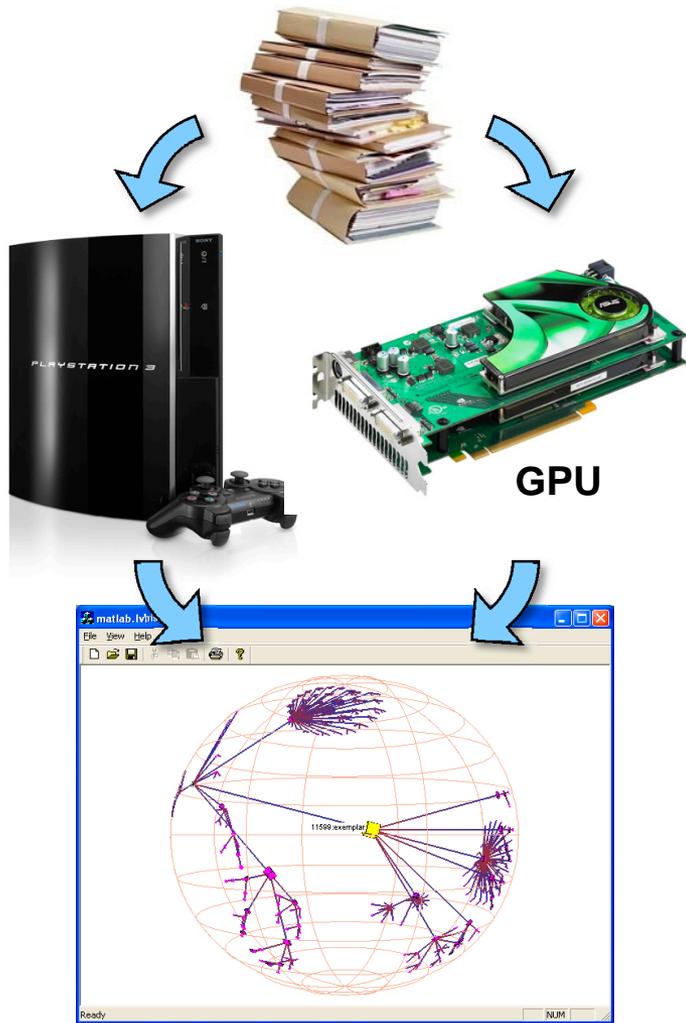
The interface is divided into several main sections:

- Collections (28):** A list of folders representing different groups of documents, including obama (4), mccain (5), huckabee (5), edwards (5), clinton (5), and giuliani (4).
- Top Collection Terms:** A list of terms extracted from the collections, such as obama (11), mccain (10), huckabee (16), edwards (11), clinton (8), and giuliani (9).
- Top Phrases (75):** A list of phrases with their frequency counts, including "john edwards (3)", "barack obama (2)", "john mccain (2)", "mike huckabee (2)", "mike huckabee, mikehuckabee (2)", "rudy giuliani (2)", "action script (1)", "ann coulter (1)", "approval ratings, approval rating (1)", "arizona republic (1)", "barackobama, barack obama (1)", "benazir bhutto (1)", "bill clinton, bill clintons (1)", "bill muller (1)", "boston globe (1)", "budget resolution raise (1)", "chicago tribune (1)", "city hall (1)", "clinton announce funding (1)", "clinton announces measure (1)", "clinton calls (1)", "clinton city hall (1)", "clinton questions general (1)", "clintoncity, clinton city (1)", "dan nowicki (1)", "draft rudy giuliani (1)", and "edwards signals, edwards signaling (1)".
- All Phrases (6935):** A list of all phrases found in the collections, including "documentation license (6)", "gore john kerry (6)", "tax deductible nonprofit (6)", "wikimedia foundation (6)", "barack obama (5)", "barack obama bill (5)", "boston globe (5)", "congressional support fundraising (5)", "convention hillary clinton (5)", "convention hugh cort (5)", "dan gilbert rudy (5)", "dodd dal lamagna (5)", "duncan hunter alan (5)", "gilmore ray mckinney (5)", "giuliani mike huckabee (5)", "gravel dennis kucinich (5)", "john cox jim (5)", "john edwards mike (5)", "keyes john mccain (5)", "national convention (5)", "ralph nader (5)", "richardson evan bayh (5)", "romney fred thompson (5)", "ron paul mitt (5)", "timeline straw (5)", and "tom tancredo tommy (5)".
- User Term Management:** A section for managing user-defined terms, featuring an "Add Stemmed Term" button, a "Remove Term" button, and a "Generate Document Node Cluster" button.

Document Category:  
Piranha Category:

View Original Do...

# Fast, distributed image and text analysis on heterogeneous platforms



- **Problem**

- Medical image and text information are buried in enormous, dynamically evolving databases
- Clustering-based approaches for searching these large data sets are typically too slow

- **Technical Approach**

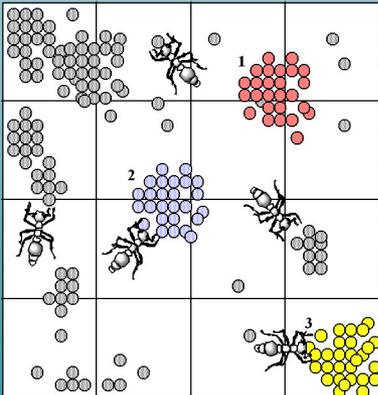
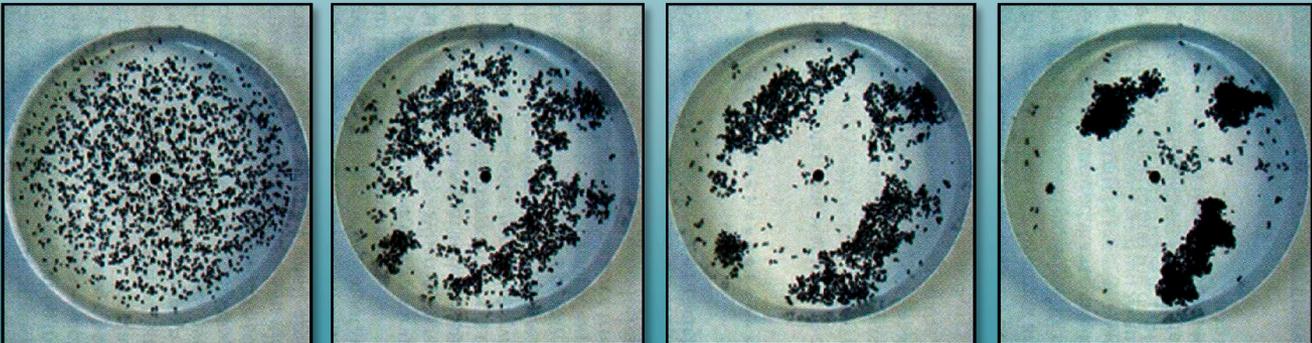
- Developed fast data clustering and analysis algorithms
- Well-suited for inexpensive, next-generation computing platforms such as Graphics Processing Units (GPUs) and Playstation 3 (PS3)

- **Benefit**

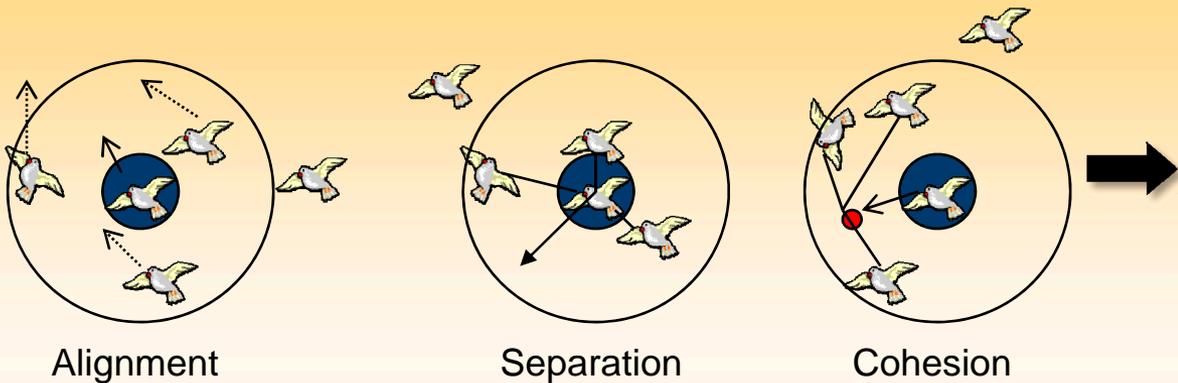
- Dramatic improvement in speed of clustering and analysis of dynamic data
- Order of magnitude (10X) performance improvement using a single PS3 versus a high-end desktop machine

# Breakthrough—bioinspired distributed solution

## Ant colony optimization

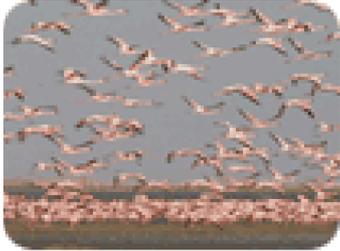


## Bird flocking model

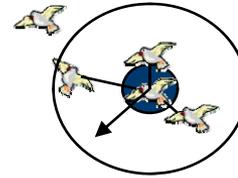


Photograph by José L. Gómez de Francisco  
© 2009 National Geographic Society. All rights reserved.  
Visions of Earth  
National Geographic magazine, March 2005

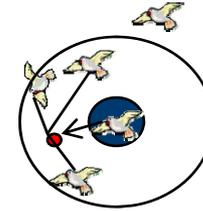
# Multiple species flocking (MSF) document clustering



Alignment



Separation



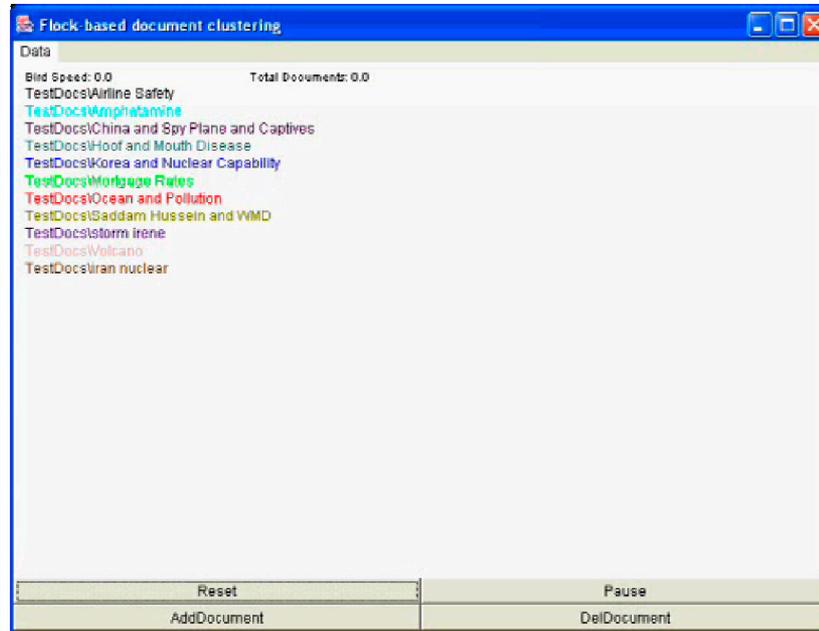
Cohesion

$$d(P_x, P_b) \leq d_2 \Rightarrow \bar{v}_{sr} = \sum_x^n \frac{\bar{v}_x + \bar{v}_b}{d(P_x, P_b)}$$

$$d(P_x, P_b) \leq d_1 \cap d(P_x, P_b) \geq d_2 \Rightarrow \bar{v}_{ar} = \frac{1}{n} \sum_x \bar{v}_x$$

$$d(P_x, P_b) \leq d_1 \cap (P_x, P_b) \geq d_2 \Rightarrow \bar{v}_{cr} = \sum_x^n (\overline{P_x - P_b})$$

	Category/topic	Number of articles
1	Airline safety	10
2	China and spy plane and captives	4
3	Hoof and mouth disease	9
4	Amphetamine	10
5	Iran nuclear	16
6	North Korea and nuclear capability	5
7	Mortgage rates	8
8	Ocean and pollution	10
9	Saddam Hussein and WMD	10
10	Storm Irene	22
11	Volcano	8



The clustering results of K-means, ant clustering, and MSF clustering algorithm on synthetic and document data sets after 300 iterations

	Algorithms	Average cluster number	Average F-measure value
Synthetic dataset	MSF	4	0.9997
	K-means	4	0.9879
	Ant	4	0.9823
Real document collection	MSF	9.105	0.7913
	K-means	11	0.5632
	Ant	1	0.1623

The document collection data set

# Summary

- Current technology **cannot solve** emerging national challenges
- Intelligent software agents are a **significant breakthrough** technology
- Results indicate **high potential** to help solve these national challenges
- We have a progression of very successfully deployed agent systems and research to our credit

# Contacts

## **Justin M. Beaver, Ph.D.**

**Applied Software Engineering Research Group  
Computational Sciences and Engineering Division**

**(865) 576-0327**

**beaverjm@ornl.gov**

## **Yu (Cathy) Jiao, Ph.D.**

**Applied Software Engineering Research Group  
Computational Sciences and Engineering Division**

**(865) 574-0647**

**jiaoy@ornl.gov**