

Data Analysis and High Performance Computing

Presented by

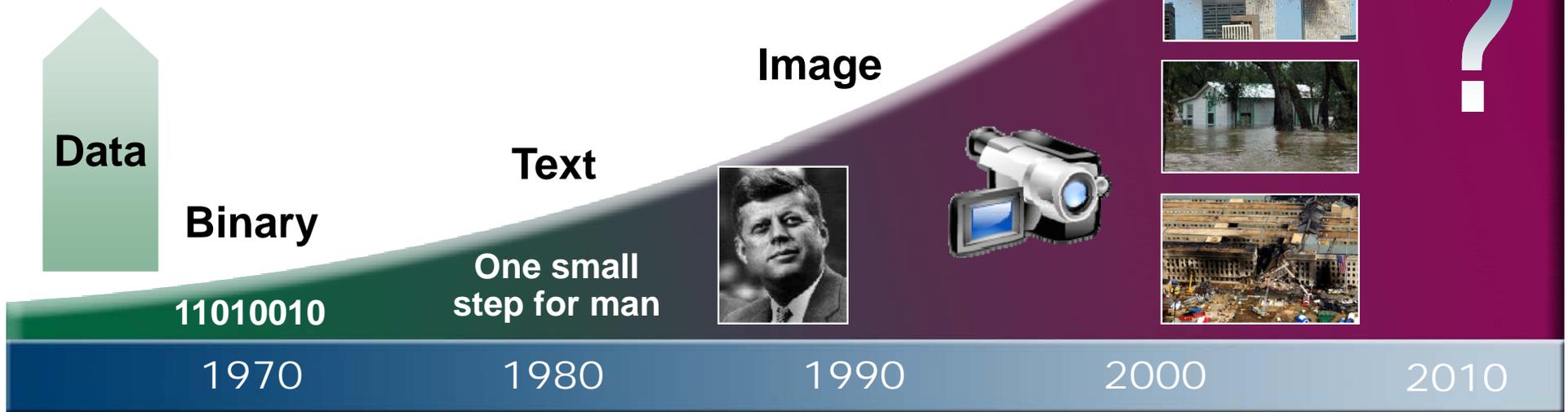
Yu (Cathy) Jiao, Ph.D.
Robert M. Patton, Ph.D.
Xiaohui Cui, Ph.D.

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



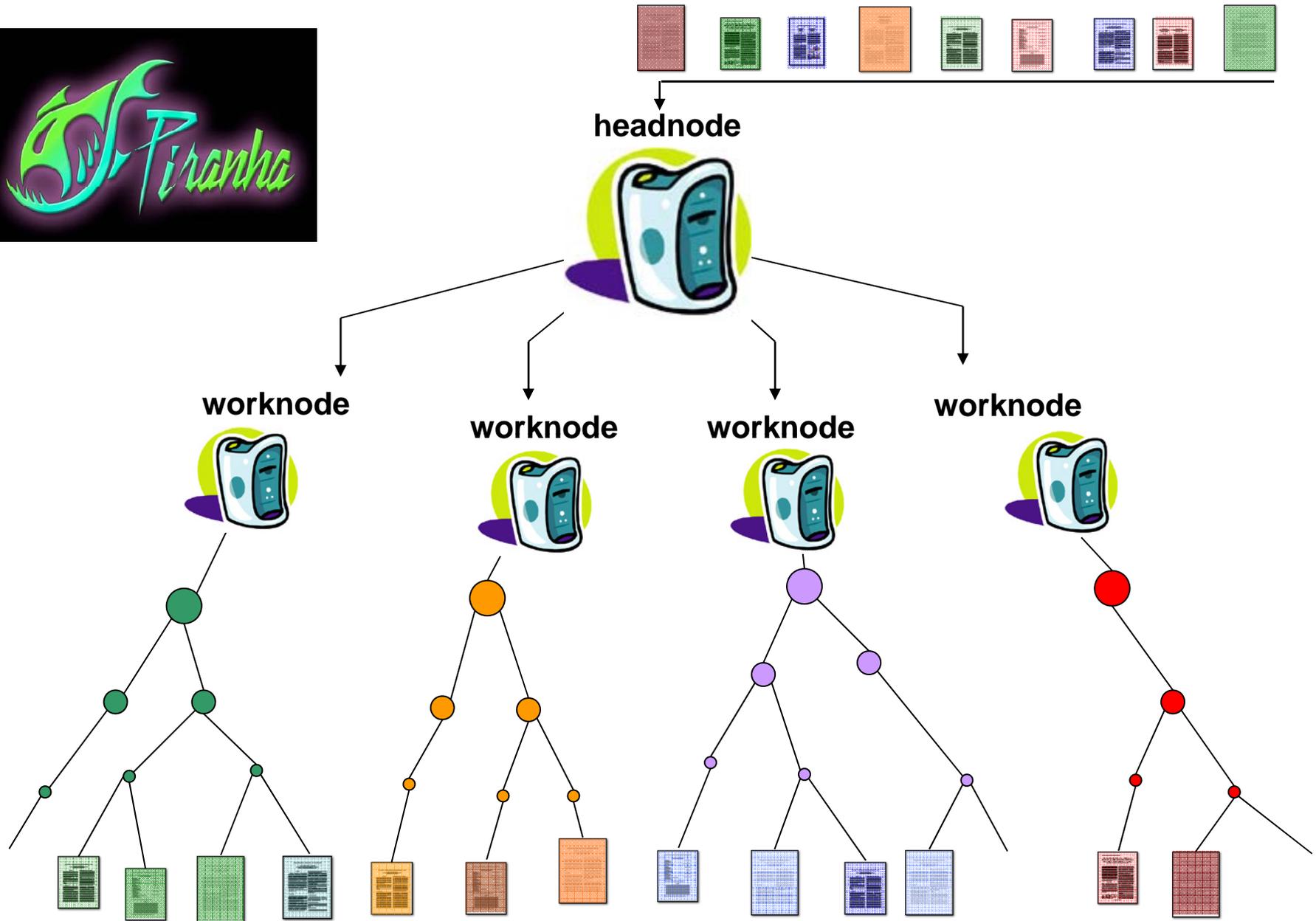
Streaming data analysis challenges

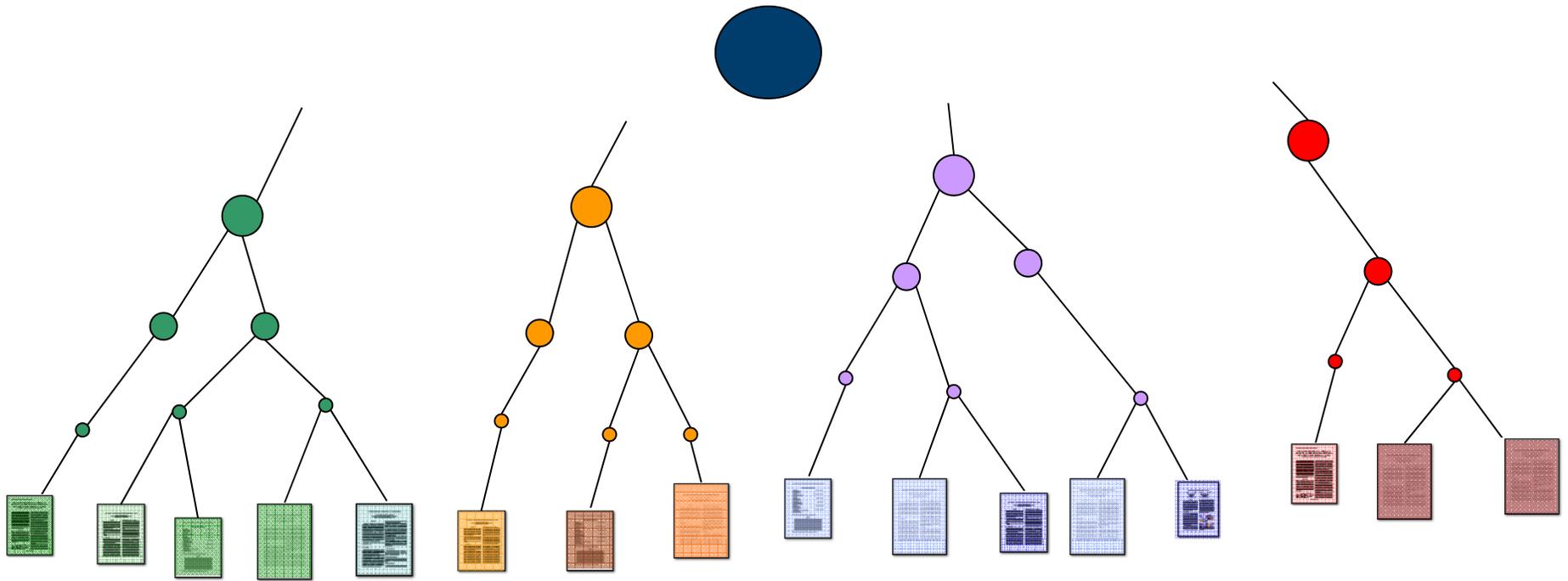
- High volume data streams are constantly generated.
- Traditional data encoding scheme is inefficient.
- Need a new solution to handle incremental clustering.



Distributed data stream mining with Piranha

- **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.**





Piranha Client File View Tools Settings Help

Home Refresh Print Save Open Close

Collections

- Collections (78)
 - Amphetamine (10)
 - Saddam Hussein and WMD (10)
 - Hoof and Mouth Disease (9)
 - Mortgage Rates (8)
 - Volcano (8)
 - Korea and Nuclear Capability (5)
 - Airline Safety (10)
 - Satanic Cult (4)

Top Collection Terms

- Amphetamine (50)
- Saddam Hussein and WMD (50)
- Hoof and Mouth Disease (45)
- Mortgage Rates (40)
- Volcano (40)
- Korea and Nuclear Capability (25)
- Airline Safety (50)
- Satanic Cult (20)

Top Words

- wmd (6)
- airlines, airline (5)
- amphetamine, amphetamines (5)
- saddam, saddams (5)
- volcanoes, volcano (4)
- amphetamine (3)
- cna (3)
- crew (3)
- mortgage (3)
- mortgages, mortgage (3)
- ocean (3)
- oncb (3)
- pollution (3)
- prodi (3)
- satanic (3)
- volcano (3)
- amphetamines (2)
- 10 - Shanghai Police Crack
- 39 - Taipei Police Seize 34

All Terms

- All Terms (15662)
- according (32)
- found (22)
- million (22)
- international (21)
- minister (21)
- government (20)
- major (19)
- military (19)
- president (19)
- united (19)
- authorities (18)
- case (18)
- chinese (18)
- clear (18)
- destruction (18)
- foreign (18)
- high (18)
- part (18)

User Term Management

Add Word Add Stemmed Term

Remove Term

User Terms

● User Terms (0)

Generate Document Cluster Generate Document Node Cluster

Main Document

10

Similar Documents

None

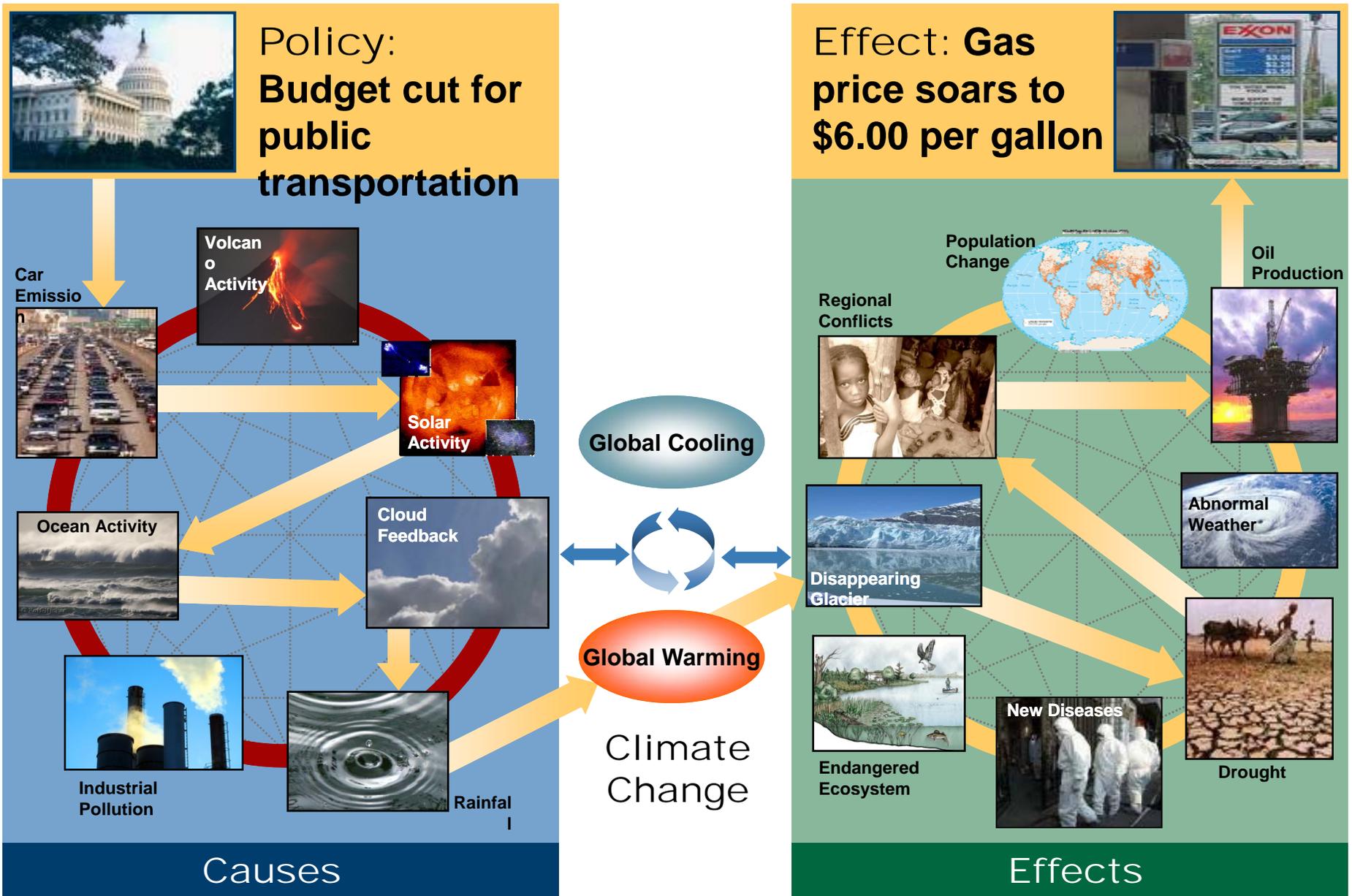
Category : Amphetamine

Title	Shanghai Police Crack Case Involving Amphetamines
AGENT_ASSIGNED_CATEGORY	Amphetamine
PIRANHA_ASSIGNED_CATEGORY	unknown
Content	<p>Shanghai Police Crack Case Involving Amphetamines</p> <p>Shanghai, March 10 (XINHUA) -- Police in China's largest metropolis have cracked the country's first case involving MDMA and MDA, drugs in the category of Amphetamines. In the early morning of February 15, a score of people were found taking the drugs in several public places of entertainment, and a man from Hong Kong was seized on the spot while selling the drugs, according to Zhang Shenghua, deputy director of the Shanghai Public Security Bureau. More than 120 pills were discovered on the man. Police sources said MDMA and MDA stimulate the nerve center, and that drug-users get excited easily and tend to act violently. The deputy director said that the city's police would carry out tougher crackdowns on the trafficking and use of "narcotic drugs."</p>

View Original Document

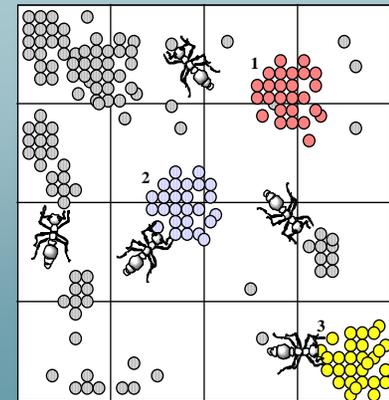
Ready 78 / 78

Modeling the impact of policies

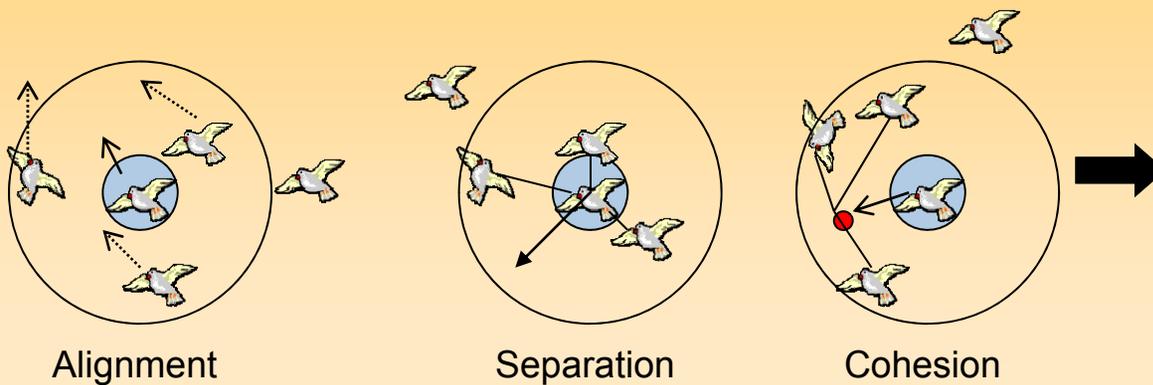


Breakthrough—bioinspired distributed solution

Ant colony optimization



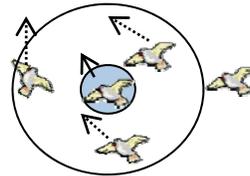
Bird flocking model



Photograph by José L. Gómez de Francisco
© 2009 National Geographic Society. All rights reserved.

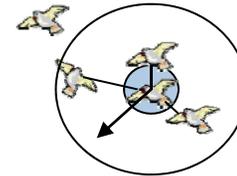
Visions of Earth
National Geographic magazine, March 2005

Multiple species flocking (MSF) document clustering



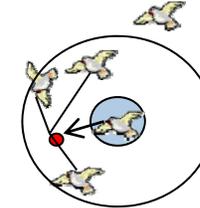
Alignment

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



Separation

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

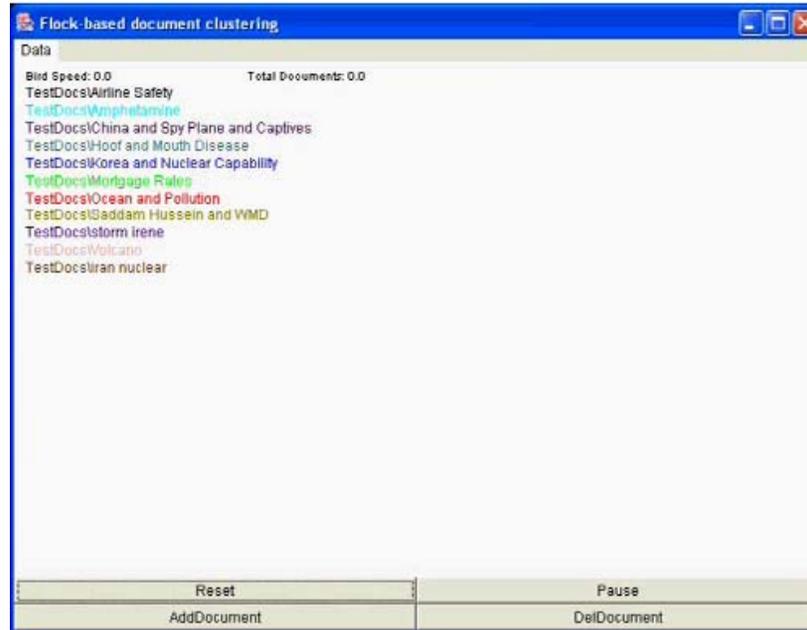


Cohesion

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

	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 document collection dataset



The clustering results of K-means, ant clustering and MSF clustering algorithm on synthetic and document datasets 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

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 significantly successfully deployed agent systems and research to our credit.**

Contacts

Yu (Cathy) Jiao, Ph.D.

Applied Software Engineering Research Group
Computational Sciences and Engineering Division
(865) 574-0647
jiaoy@ornl.gov

Robert M. Patton, Ph.D.

Applied Software Engineering Research Group
Computational Sciences and Engineering Division
(865) 576-3832
pattonrm@ornl.gov

Xiaohui Cui, Ph.D.

Applied Software Engineering Research Group
Computational Sciences and Engineering Division
(865) 576-9654
cuix@ornl.gov

