

Homeland Security: Research in Discrete Sciences

Eduard Hovy

USC/ISI

hovy@isi.edu

Representing the 4 IDS-UAC Center Directors

What is Discrete Science?

- Built on a foundation of the mathematical sciences:
 - Math, Computer Science, Statistics
- Deals with **data and information**
 - Massive, heterogeneous, dynamic
- Seeks patterns, and *departure* from patterns
- Develops powerful computer algorithms

Methods of discrete science have become important tools for homeland security, especially when combined with powerful, modern computer methods for analysis and simulation

IDS is a “Center of Centers”

Institute for Discrete Sciences

RUTGERS

DyDAn: The Center for Dynamic Data Analysis



CKID: The Center for Knowledge Integration and Discovery

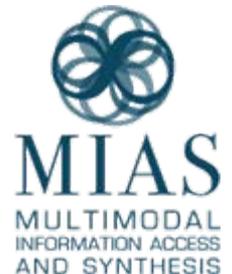


CERATOPS: Center for Extraction and Summarization of Events and Opinions in Text



ILLINOIS

MIAS: Multimodal Information Access and Synthesis



IDS partner network



RUTGERS

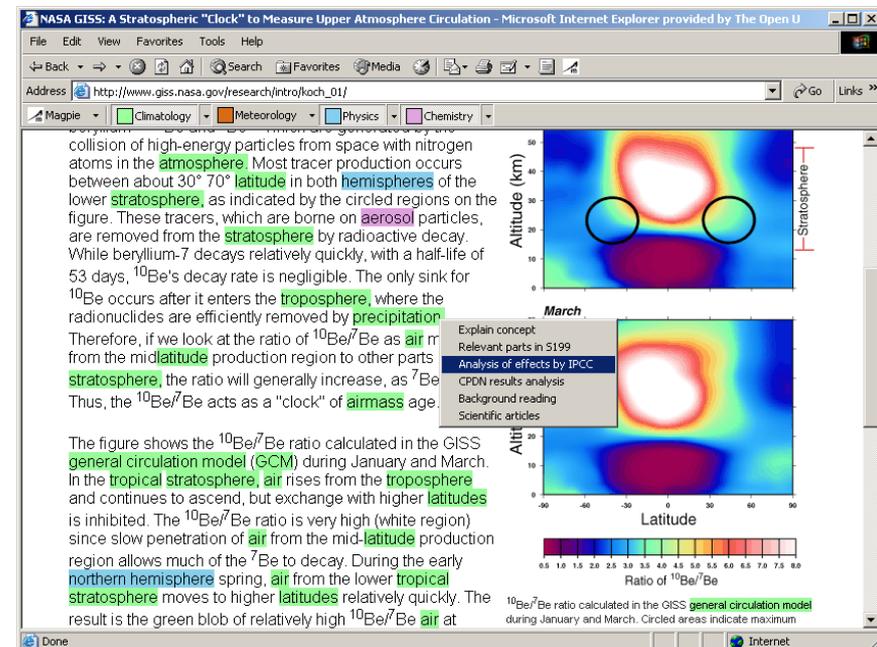


IDS research goals 1

IDS seeks to develop fundamental theories, algorithms, and tools that enable us to gain knowledge from data

- Problem 1: **Extracting pertinent information** from various media

- New capabilities in information identification, extraction, storage, and access across media
- E.g., robust techniques for extracting, summarizing and tracking information about events from unstructured text



Information Extraction

- Sources:

- Text

- Entities and events
- Opinions and goals

- Geospatial Data:

- Maps of various types
- Satellite images

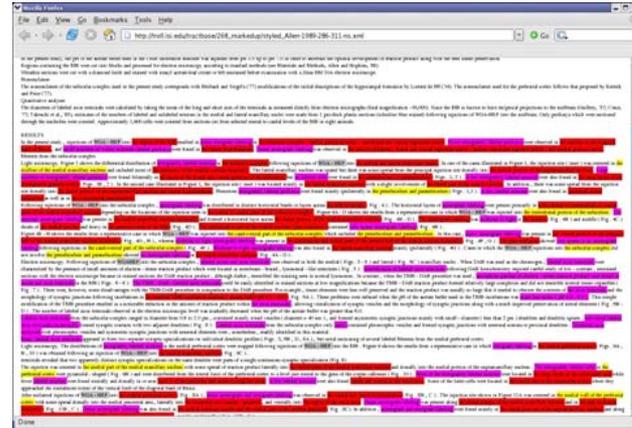
- Images

- Photos
- Diagrams

- Speech

- Content
- Emotions

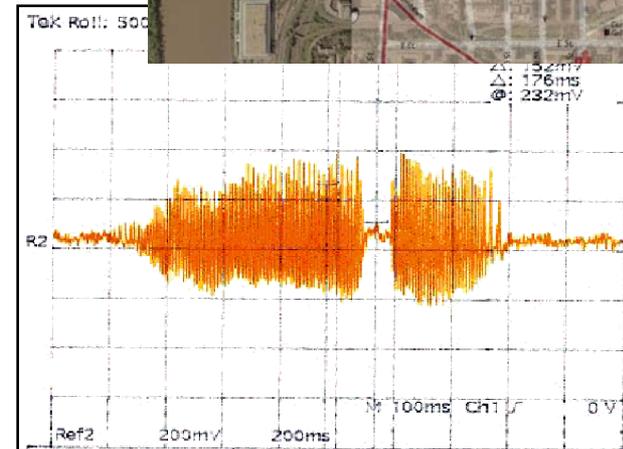
CERATOPS



MIAS



CKID



Text extraction for animal health surveillance

- Collaborative project between CERATOPS, PURVAC, and the Veterinary Information Network (VIN) using ProMEDmail, with funding from LLNL
- Goal: proof-of-concept of an end-to-end NLP-based visual analytics system for unstructured text
 - 73% of emerging infectious diseases are zoonotic in origin
 - Pets can provide early warning signs of disease outbreaks and exposures to toxic substances
 - Adverse pet reactions can be early indicators of food chain contamination

CERATOPS



Pet Owners Watch Animals Closely After Food

Recall

Donie Turner
Associated Press
Wednesday, March 21, 2007; 8:22 PM

ATLANTA (AP) -- A recall of potentially deadly pet
and cat owners studying their animals for even the slightest hint of
illness and swamping veterinarians nationwide with calls about

20 Million Chickens Given Tainted Feed

Birds Held From Market for Study

By *Rick Weiss*
Washington Post Staff Writer
Saturday, May 5, 2007; Page A08

Syntactic Analysis

3 chickens died from avian flu.

Extraction

3 chickens died from avian flu.
SUBJ VP PP

Coreference Resolution

Fact: DEATH
Victim: 3 chickens **Disease:** avian flu

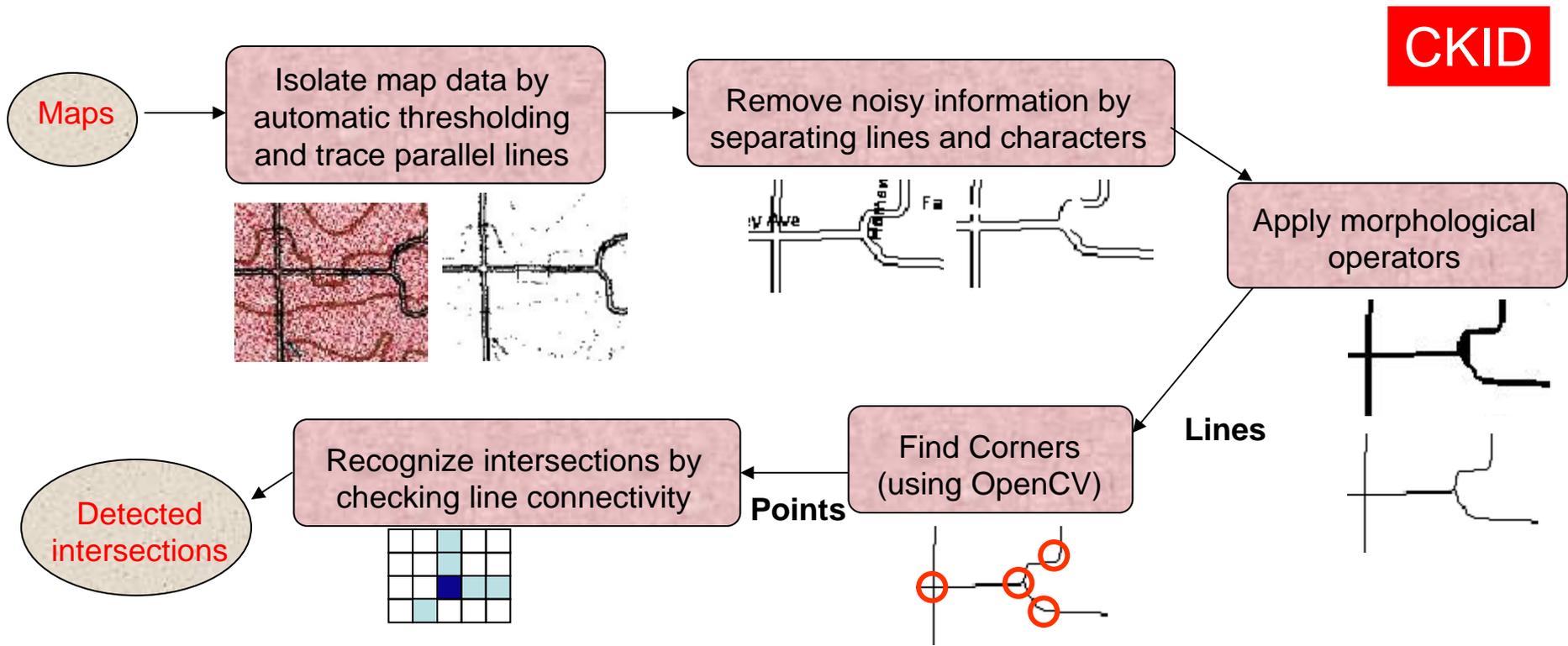
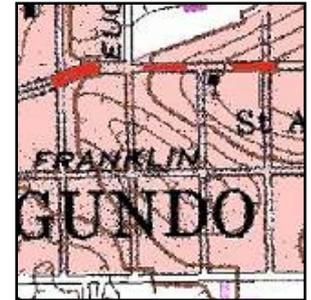
Template Generation

3 chickens died from avian flu.
The birds were found in Canada.

Event: Outbreak
Victim: 3 chickens / the birds
Disease: bird flu
Country: Canada

Finding intersections on maps

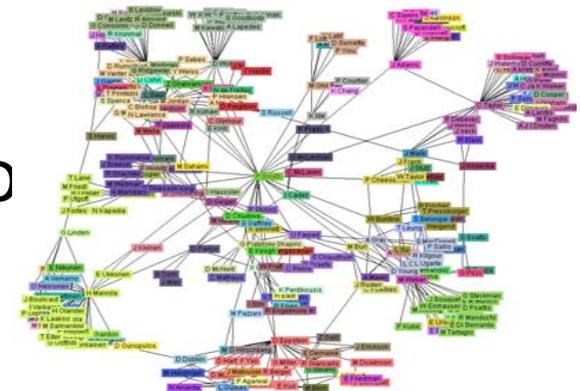
- Difficult to identify intersection points automatically and accurately
 - Varying thickness of lines
 - Single-line map vs double-line map
 - Noisy information: symbols and alphanumeric characters



IDS research goals 2

IDS seeks to develop fundamental theories, algorithms, and tools that enable us to gain knowledge from data

- Problem 2: **Fusing and storing** this information
 - Computationally efficient methods for representing and fusing information, while preserving privacy
 - E.g., applying inference rules to enforce data consistency and discover anomalies



Continuous, distributed monitoring of dynamic, heterogeneous data

- Need to understand massive amounts of data
- Problems:
 - Data takes numerous forms; requires data mining methods that span the modalities
 - Data inherently distributed from multiple sources
 - Data arrives rapidly and continuously
- Seek anomalies, patterns, emerging events
 - Run continuous queries to monitor incoming data stream



Data and Statistics

DyDAn

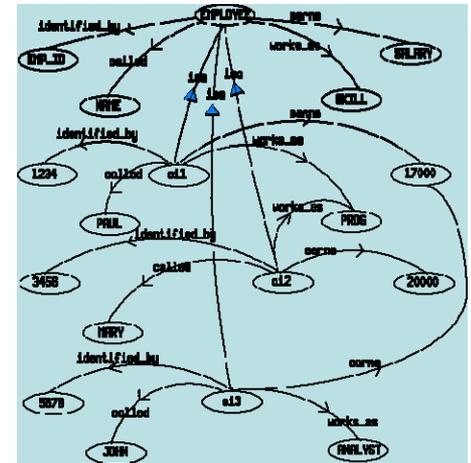
Representing and fusing info

- Multigraphs: Represent source information that is cross-linked multiple times by various relations
- Data management: Large databases
- Adding semantics:
 - Deploy large general ontologies over databases
 - Perform entity (co)reference

DyDAn

CKID

MIAS



Retrieving and accessing info

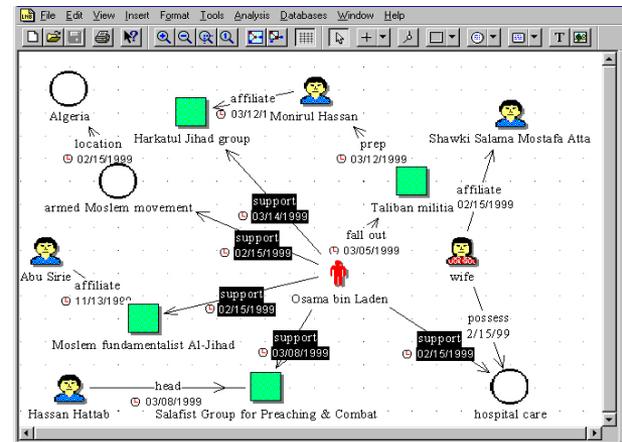
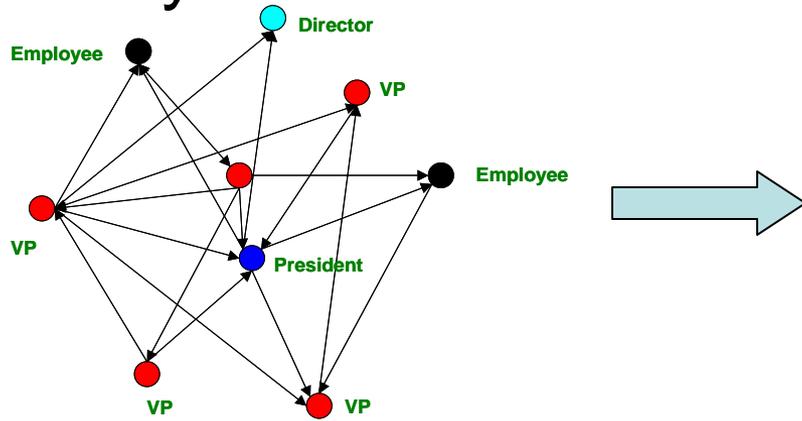
- Textual information retrieval:
 - Methods to enhance information search over the web and the ‘deep’ web

MIAS

CKID

- Database retrieval and optimization:
 - Methods for handling very large amounts of changing data in ‘optimally efficient’ ways

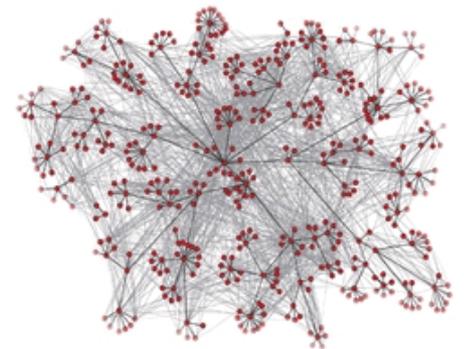
DyDAn



IDS research goals 3

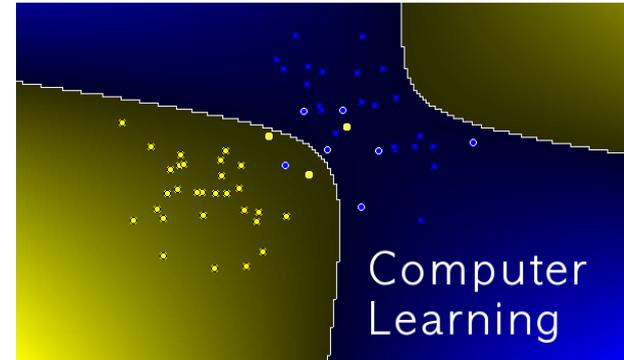
IDS seeks to develop fundamental theories, algorithms, and tools that enable us to gain knowledge from data

- **Problem 3: Finding trends and patterns of interest**
 - Novel technologies for identifying patterns and relationships in massive graphs and datasets that change rapidly
 - E.g., trainable learning algorithms to discover and extract events such as infectious disease outbreaks



Pattern detection

- Detect correlations and patterns over large (multi)graphs:
 - Count features (words, etc.) to classify and/or match
- Statistical & information-theoretic techniques for pattern and trend discovery over extracted info:
 - People, meetings, and events
 - Social network analysis
 - Analysis of blogs
 - Credit card and bank transactions
 - Fraud detection

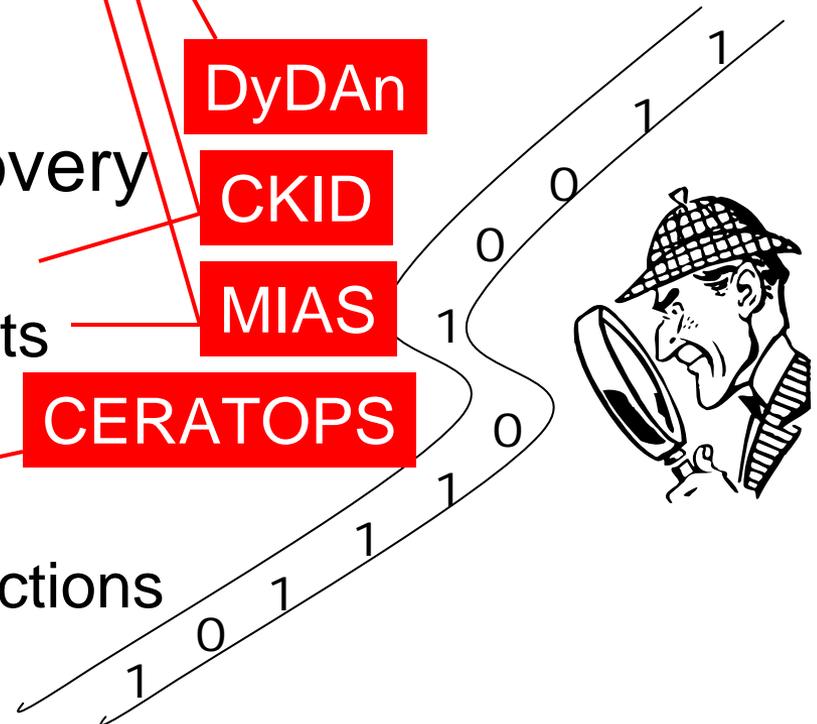


DyDAn

CKID

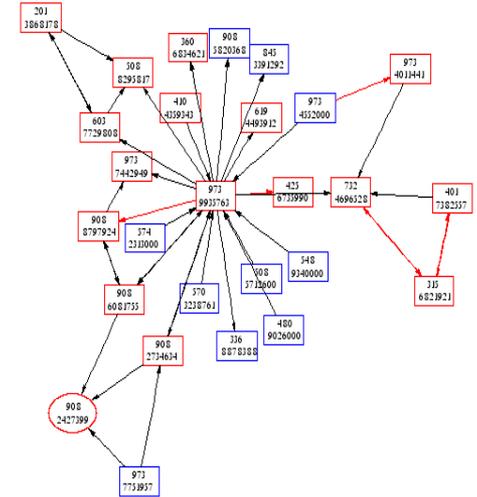
MIAS

CERATOPS



Statistical and graph-theoretical approaches to time-varying multigraphs

- A COI (Community of Interest) is
 - an effective **summary of significant connections** in a graph
- Use COI for very large scale analysis of a dynamic graph:
 - Stark change in COI indicates an anomaly
 - Has an entity changed its id?
 - New cliques?
- Goal: Analyze and apply automated anomaly detection to COIs of dynamic multigraphs in telecomm, blogs, and intelligence data
- Example: telecomm data at AT&T
 - Proven technique to summarize and analyze 300 million telephone nos. and 350 million calls daily, using Kalman Filters on contingency tables



DyDAn

Pattern detection in data

- Research focus: Analyzing data streams, frequent patterns, sequential patterns, graph patterns, and their applications
- Privacy preserving data mining
- Machine learning methods:
 - Developed many popular pattern detection algorithms, e.g., FPgrowth, PrefixSpan, gSpan, StarCubing, CrossMine, RankingCube, and CrossClus
 - Semantic analysis and data enrichment
 - Entity and relation identification and integration
 - Textual entailment

MIAS

Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc. last year	Does it say that?	
		Yahoo acquired Overture
		Overture is a search company
		Google is a search company
		Google owns Overture

IDS educational programs

- Sampling of events:
 - Summer 2007, 2008: 8-week school (UIUC + teachers from other centers)
 - Summer 2008: 1-week short course on Biosurveillance (DyDAn)
 - November 2007: Tutorial: Balancing Data Confidentiality and Data Quality (DyDAn)
 - 2008: Scholar program and summer interns: (DyDAn, USC, UIUC)
 - February 2008: Workshop: Privacy Preserving Data Analysis (DyDAn)



Open to All

Concluding remark

Discrete science is a “cross-cutting” capability that can provide support to **each** of the DHS divisions

Please contact us!

DyDAn: The Center for Dynamic Data Analysis

- Fred Roberts (froberts@dimacs.rutgers.edu)



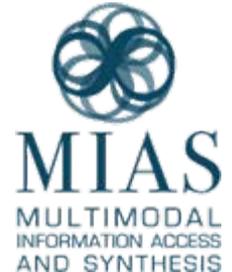
CKID: The Center for Knowledge Integration and Discovery

- Eduard Hovy (hovy@isi.edu)



MIAS: Multimodal Information Access and Synthesis

- Dan Roth (danr@cs.uiuc.edu)



CERATOPS: Center for Extraction and Summarization of Events and Opinions in Text

- Jan Wiebe (wiebe@cs.pitt.edu)

