

Semantic Abstraction and Integration Across Text Documents and Data Bases

Dan Roth

Department of Computer Science
University of Illinois at Urbana/Champaign



Homeland
Security



MIAS Mission

- Most of the data today is unstructured
 - books, newspaper articles, journal publications, reports, images, and audio and video streams.
- How to deal with the huge amount of unstructured data as if it was organized in a database with a known schema.
 - how to locate, organize, access and analyze unstructured data.

MIAS Mission:

- develop the theories, algorithms, and tools for analysts to
 - access a variety of data formats and models
 - integrate them with existing resources
 - transform raw data into useful and understandable information.



Task Perspective

- In the next decade, intelligence analysts will need to
 - monitor a huge number of interesting events and entities
 - formulate and evaluate hypotheses with respect to them.
- Analysts must interact, at the appropriate level of semantic abstraction, with a system that can
 - synthesize, summarize and interpret vast amounts of multimodal information,
 - integrate observed data with domain models and background information in multiple formats,
 - propose hypotheses, and help verify them.



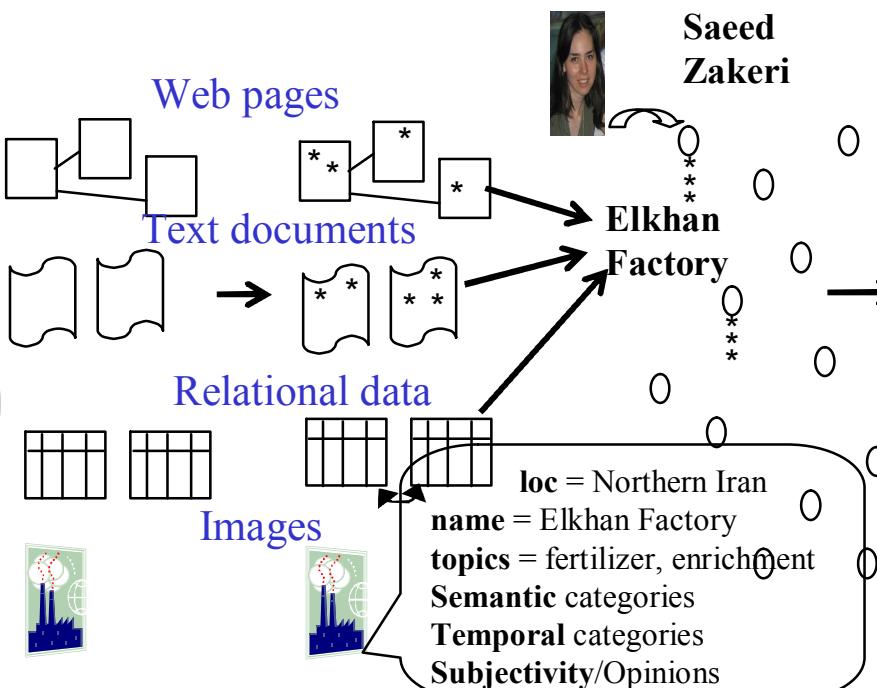
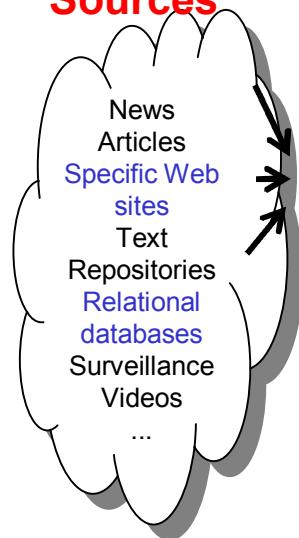
Scenario

- Consider an intelligence analyst researching a problem
 - Iranian nuclear program – generate a list of Iranian nuclear scientists, affiliations, specialties, biographies, photos, and notable recent activities.
 - Medical treatment – what is known about it; who are the experts; what do users say about it; what side effects have been reported
- Current technologies have solved the problem of collecting and storing huge amounts of information; it would be reasonable to assume that the information she is after does exist;
- However, multiple barriers exist on the way to a successful completion of the analysis, each posing a significant research challenge.



Multimodal Information Access & Synthesis

Online Data Sources



Discover unusual events, entities, and associations.

Continuous monitoring of events, entities & associations

Rapid retrieval of all info. about a particular entity

Efficient keyword search, querying, question answering,

browsing, mining,

Infer Metadata:
Semantic entities

Discover Relations
Between Semantic Entities

Focused Multimodal Data Retrieval

Semantic Disambiguation &
Integration across multiple
sources and modalities

Support Information Analysis, Knowledge Discovery, Monitoring



Semantic Categories

- Information Access and Extraction requires the identification of semantic categories in text.

Query: Aids Treatment

Federal health officials are recommending aggressive use of a newly approved drug that protects people infected with the AIDS virus against a form of pneumonia that is the No.1 killer of AIDS victims.

(AP890616-0048, TIPSTER VOL. 1)

Relevant documents may mention specific types of treatments for AIDS

Hemophiliacs lack a protein, called factor VIII, that is essential for making blood clots. As a result, they frequently suffer internal bleeding and must receive infusions of clotting protein derived from human blood. During the early 1980s, these treatments were often tainted with the AIDS virus.

(AP890118-0146, TIPSTER Vol. 1)

Many irrelevant documents mention AIDS and treatments for other diseases

- There is a need to identify that this phrase represent a name of an organization, a name of a person, a name of a disease, a medicine, etc.
- A narrow version of the problem is called: named entity recognition (NER)

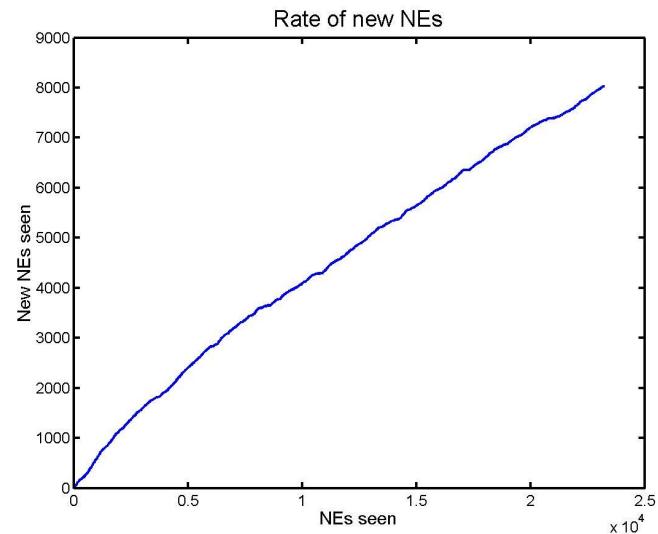


MIAS

Multi-Modal Information Access & Synthesis

Adaptation of Named Entity Recognition

- Entities are inherently ambiguous (e.g. JFK can be both location and a person depending on the context)
 - Can appear in various forms ; Can be nested.
 - Using lists is not sufficient
 - New entities are always being introduced
- A lot of Machine Learning work – significant over fitting
- Key difficulties – Adaptation to:
 - new domains/corpora
 - slightly new definition of an entity
 - new languages
 - New types of entities .
- How to reduce the requirements on the resources needed to produce a semantic categorization for a new domain/new language/new type of entities



NER Tools

Cognitive Computation Group

Named Entity Tagger Output

Screen shot from a CCG demo

<http://L2R.cs.uiuc.edu/~cogcomp>

The input sentences are:

- WASHINGTON (CNN) -- The FBI failed revealed Friday. The audit shows the F and a statement from the Justice Depa tolerated." In a statement released lat problems uncovered in the audit and w New York and Washington. The White House, Justice Department and the FBI have called it a vital tool in the battle against terrorism, but critics have said the act infringes on civil liberties.

Work in progress:

- Un-supervised discovery of entities in other languages
- Quick adaptation to new entity types and new domains.

The tagged output is:

- [LOC WASHINGTON] ([ORG CNN]) - The [ORG FBI] failed to report accurately how many people it snooped on using anti-terror measures , a government audit revealed Friday . The audit shows the [ORG FBI] " was sloppy when it should have been diligent " , one official with early access to the report said , and a statement from the [ORG Justice Department] said Attorney General [PER Alberto Gonzales] had told the [ORG FBI] " these past mistakes will not be tolerated . " In a statement released late Thursday , [ORG Justice Department] spokeswoman [PER Tasia Scolinos] said [PER Gonzales] was unhappy with the problems uncovered in the audit and was implementing changes . The [MISC Patriot Act] was passed after the September 11 , 2001 , attacks on [LOC New York] and [LOC Washington] . The [LOC White House] , [ORG Justice Department] and the [ORG FBI] have called it a vital tool in the battle against terrorism , but critics have said the act infringes on civil liberties .



Extracting Relations

- Information Access and Extraction requires the identification of relations between concepts in text.

Google search results for "what does hyundai produce":

- Automotive - **Hyundai** to build new plant and **produce** more model
Hyundai to build new plant and **produce** more models in China ... Despite the construction of a new factory, Beijing **Hyundai** does not intend to have new ...
auto.2456.com/eng/marketnews/details.asp?inewsiid=1&inid=57683 - 28k - Cached - Similar pages
- PDF Chevron Hydrogen Energy Station
File Format: PDF/Adobe Acrobat - View as HTML
How **does** the station **produce** hydrogen? ... technology to **produce** the hydrog...
Hyundai ... No, the station in Chino **does** not feature a stationary ...
technologyventures.chevron.com/pdf/QAChino.pdf - Similar pages
- Yahoo! Answers - **Hyundai** SUV'S-Why do Tucson and Santa Fe ha...
Why would one company (**Hyundai**) **produce** similar type vehicles that compete **does** my 1995 **Hyundai** Elantra feel like it needs rear shocks even after ...
answers.yahoo.com/question/index?qid=20060901141325AAZMB4O - 86k - Mar... Cached - Similar pages
- Yahoo! Answers - **Hyundai** Tiburon GT Manual 2003-please...
6 answers - I'm thinking of getting a 2003 **Hyundai** Tiburon GT manual with 1995 **Hyundai** Elantra feel like it needs rear shocks even after ...
answers.yahoo.com/question/index?qid=20061206011203AAJ7pZO - 103k - Cached - Similar pages
- ROMANIAN CARS SITE
In 1997 Dacia announced that it had signed a license agreement with **Hyundai** (...
The terms of the agreement called for Dacia to **produce** 500000 ...
www.geocities.com/MotorCity/Downs/4582/dacia.htm - 31k - Cached - Similar pages

- Relations expressed within a single sentence or paragraph
- Relations uncovered by processing large quantities of text (over time)

- There is a need to identify concepts (e.g., entities) and relations that hold between them in a given sentence.
- Closed set of relations:
 - [A causes B]
 - [A works for B]
 - [A prevents B]
 - [A lives in B]
- Open ended set of relations
 - Every predicate can be a relation



Extracting Relations via Semantic Analysis

Semantic Role Labeling Output

Input Text:

A car bomb that exploded outside the U.S. military base in Beniji killed 11 Iraqi citizens.

Result: Complete!

General Explanation of Argument Labels

A	bomb [A1]	killer [A0]
car		
bomb		
that		
	bomb (Reference) [R-A1]	
exploded	V: explode	
outside		
the	location [AM-LOC]	
U.S.		
military	temporal [AM-TMP]	
base		
in	location [AM-LOC]	
Beniji		
killed		
11		
Iraqi		
citizens		

Screen shot from a CCG demo

<http://L2R.cs.uiuc.edu/~cogcomp>

- Semantic parsing reveals several relations in the sentence along with their arguments.



- This level of analysis, however, cannot abstract over the inherent variability in expressing the relations. .
- Kill and Explode can be expressed in many different ways.



MIAS

Multi-Modal Information Access & Synthesis

Relations Extraction via Textual Entailment

- Given:

Q: Who acquired Overture?

- Determine:

A: Eyeing the huge market potential, currently led by Google, **Yahoo** took over search company Overture Services Inc last year.

~~Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc. last year.~~

Entails

Subsumed by

Yahoo acquired Overture

Overture is a search company

Google is a search company

Google owns Overture

.....



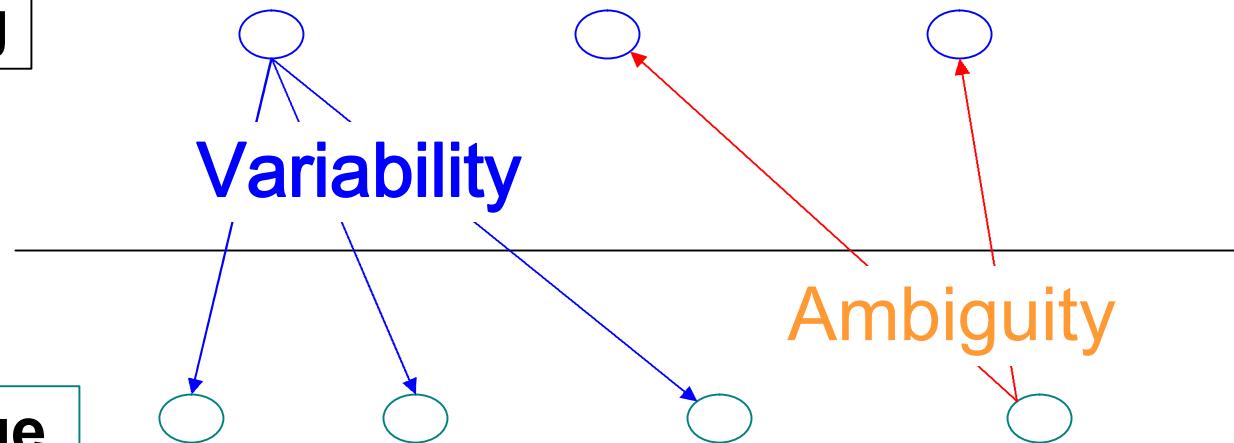
Why is it difficult?

Meaning

Variability

Language

Ambiguity



The same problem exists with other types of entities

The Reference Problem



Kennedy



Document 1: *The Justice Department has officially ended its inquiry into the assassinations of **John F. Kennedy** and Martin Luther King Jr., finding "no persuasive evidence" to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that **Kennedy** was "probably" assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the **Warren Commission**'s belief that Lee Harvey Oswald acted alone in **Dallas** on Nov. 22, 1963.*

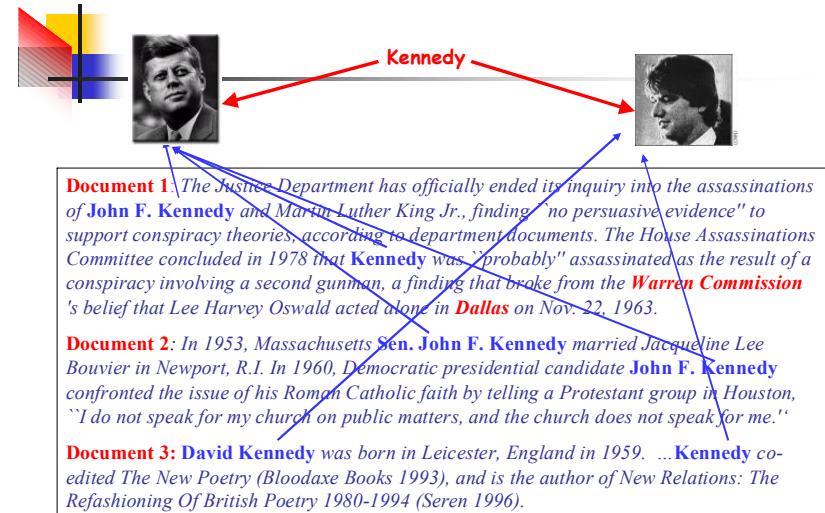
Document 2: *In 1953, Massachusetts **Sen. John F. Kennedy** married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate **John F. Kennedy** confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, "I do not speak for my church on public matters, and the church does not speak for me."*

Document 3: *David Kennedy was born in Leicester, England in 1959. ...**Kennedy** co-edited *The New Poetry* (Bloodaxe Books 1993), and is the author of *New Relations: The Refashioning Of British Poetry 1980-1994* (Seren 1996).*



Entity/Concept Identification in Text

- **Goal:** Given names in text documents and their semantic types, identify real-world entities they represent.
 - A similarity measure between names [entity type dependent]
 - A way to group different looking strings into one group
 - A context sensitive way to distinguish between identical/similar strings that represent different entities
- A generative Model
[Li, Morie, Roth, NAACL'04]
- A discriminative approach
[Li, Morie, Roth, AAAI'04]
- Summary: AI Magazine Special Issue on Semantic Integration'05



6

- **Goal:** Semantic Integration: Text, Databases and Institutional Recourses
 - Map concepts identified in text to entries in databases.
 - Construct/augment databases from textual information.
 - Aid discovery in text using existing knowledge bases.



Wilson - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop http://flake.cs.uiuc.edu/cgi-bin/connect.pl Search Print

Home Bookmarks

I-Track Look for Wilson as a Person

When searching for Wilson, we found 5 names you might be interested in:

Woodrow Wilson

Weinstein's study of President Woodrow Wilson argued that Wilson carried with him into the White House "psychological baggage" resulting from illnesses, which explained some of his policy positions and even his figures of speech.

Related Entities »

White House	» Bill Clinton	United States	Washington	New York City	New York Times News Service
San Francisco 49ers	» New York City	United States	Washington	California City	America
R-Roseville	» Woodrow Wilson	Ruben Barrales	Santa Ana	Seiji Ozawa	Koussevitzky
Pharmaceutical	» Petroleum Finance	William Jennings	Thank God	Mark Fabiani	Columbia University
Green Party	» George W. Bush	Bill Clinton	Ralph Nader	Al Gore Jr.	Texas

Relevant news articles (5):
New York Times September 21, 1998 - New York Times September 21, 1998 - New York Times September 22, 1998 - New York Times September 21, 1999 - New York Times September 23, 2000

Luke Wilson

Andy (Luke Wilson, who currently can be seen working the same sweet schlub quality in "Blue Streak"), is a recently dumped young bachelor who misses his ditzy girlfriend, Cheryl (Kathleen Robertson), perhaps a little less than he does their pooh Mogley, who is now a child of joint custody.

Related Entities »

Dog Park	» Dog Park	Bruce McCulloch
Dog Park	» Dog Park	Bruce McCulloch
Bruce McCulloch	» Dog Park	Dog Park
Janeane Garofalo	» Owen Wilson	Dog Park
Marshall	» New York Times News Service	New York City

Relevant news articles (2):
New York Times September 23, 1999 - New York Times September 23, 1999

Owen Wilson

In his low-key way, though, Wilson some

Related Entities »

Janeane Garofalo	» Detroit Tigers	Gold Rush
Detroit Tigers	» Alaska	Detroit Tigers
Patricia Arquette	» Gold Rush	New York City
Gold Rush	» Georgia Tech	

Relevant news articles (2):
New York Times September 23, 1999 - New York Times September 23, 1999

Pete Wilson

Sean Walsh, a spokesman for Gov. Pete Wilson, called the comparison of spending on corrections and higher education ludicrous.

Related Entities »

.....

Screen shot from a CCG demo
<http://L2R.cs.uiuc.edu/~cogcomp>

More work on this problem:
Scaling up
Integration with DBs
Temporal Integration/Inference

Related Entities – Context

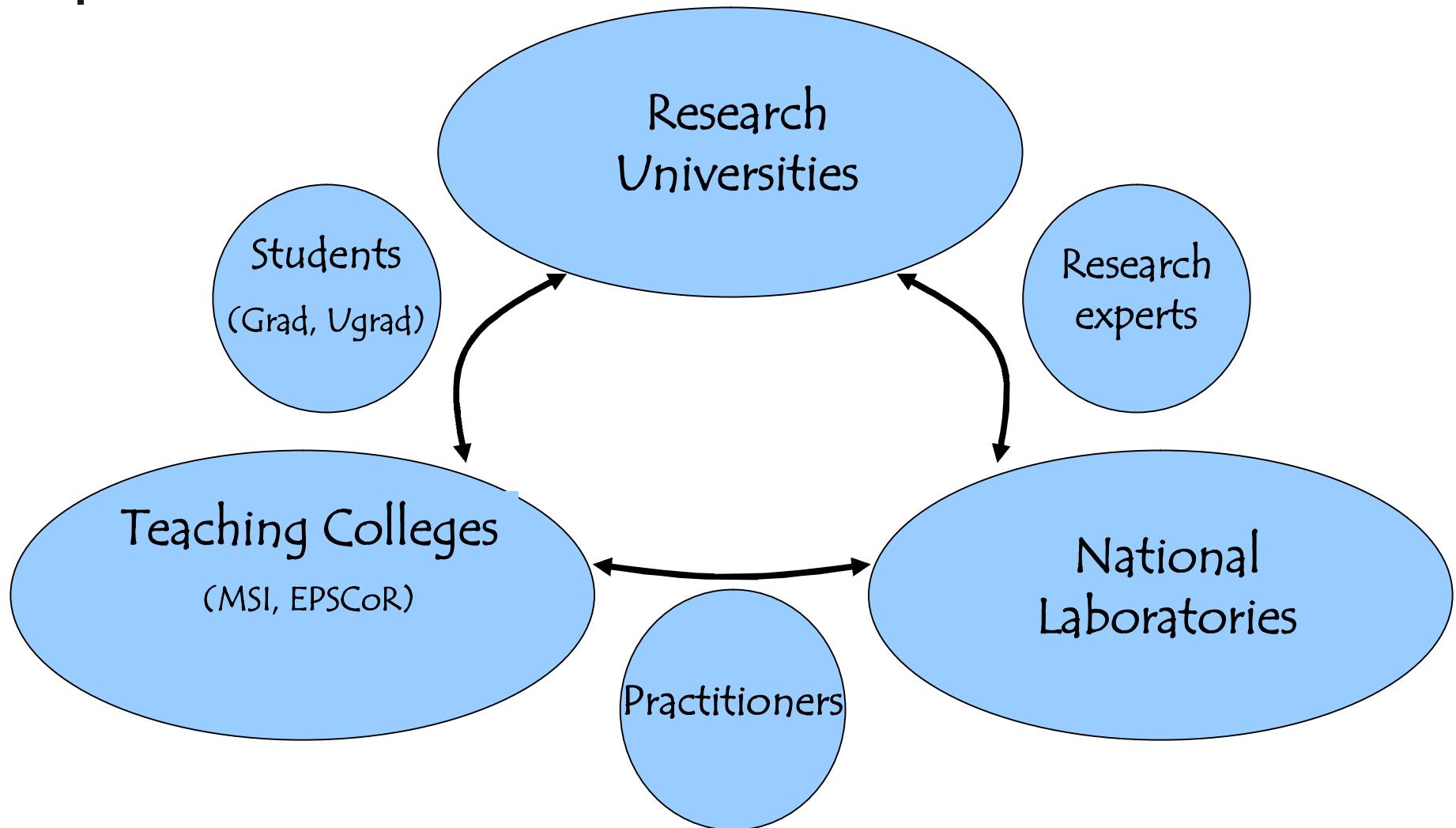
Integrated Mission: Research & Education

- Develop diverse human resources to enhance the scientific research, educational, and governmental workforce in MIAS
- Educational and Outreach Initiatives:
 - Encourage computer science students in universities with small research programs, particularly minority-serving, to pursue graduate studies
 - Expose them to the national labs
 - Open opportunities for bigger impact
- A comprehensive education program designed to increase participation in the study and practice of MIAS topics:
 - Provide substantive training for a new generation of experts in the field,
 - Serve as a tool for recruiting an experienced group of undergraduates into graduate study in one of the broad fields of information science
 - Be an intellectual community center, where participants at all levels of expertise come together in an enriched environment of collaboration.



Educational Initiatives

Diverse populations → Enriched collaborations

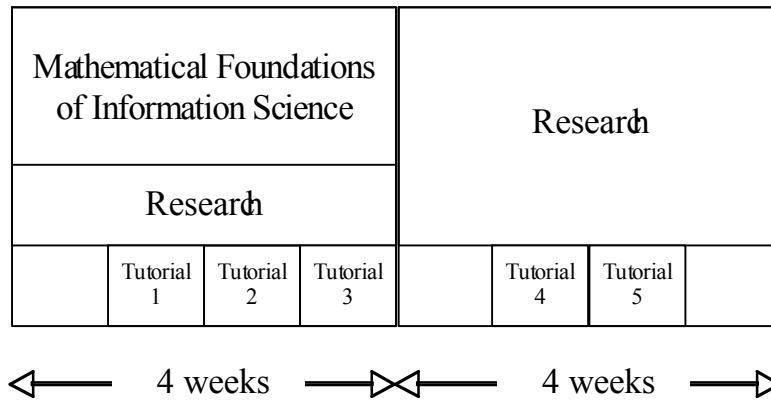


Data Science Summer Institute at UIUC

Intensive Course
in
The Math of Data
Sciences

- Probability and Statistics
- Linear Algebra
- Data Structures and Algorithms
- Optimization
- Learning & Clustering

Starting May 2007
Let us know if you want to send
**Students
Research projects ideas
Funding**



Research Projects

(Problems, possibly, from industry/national labs)

- Research institution resources
- Engages undergrads, grads, small colleges faculty, & national experts

Advanced MIAS Related Tutorials



MIAS

Multi-Modal Information Access & Synthesis

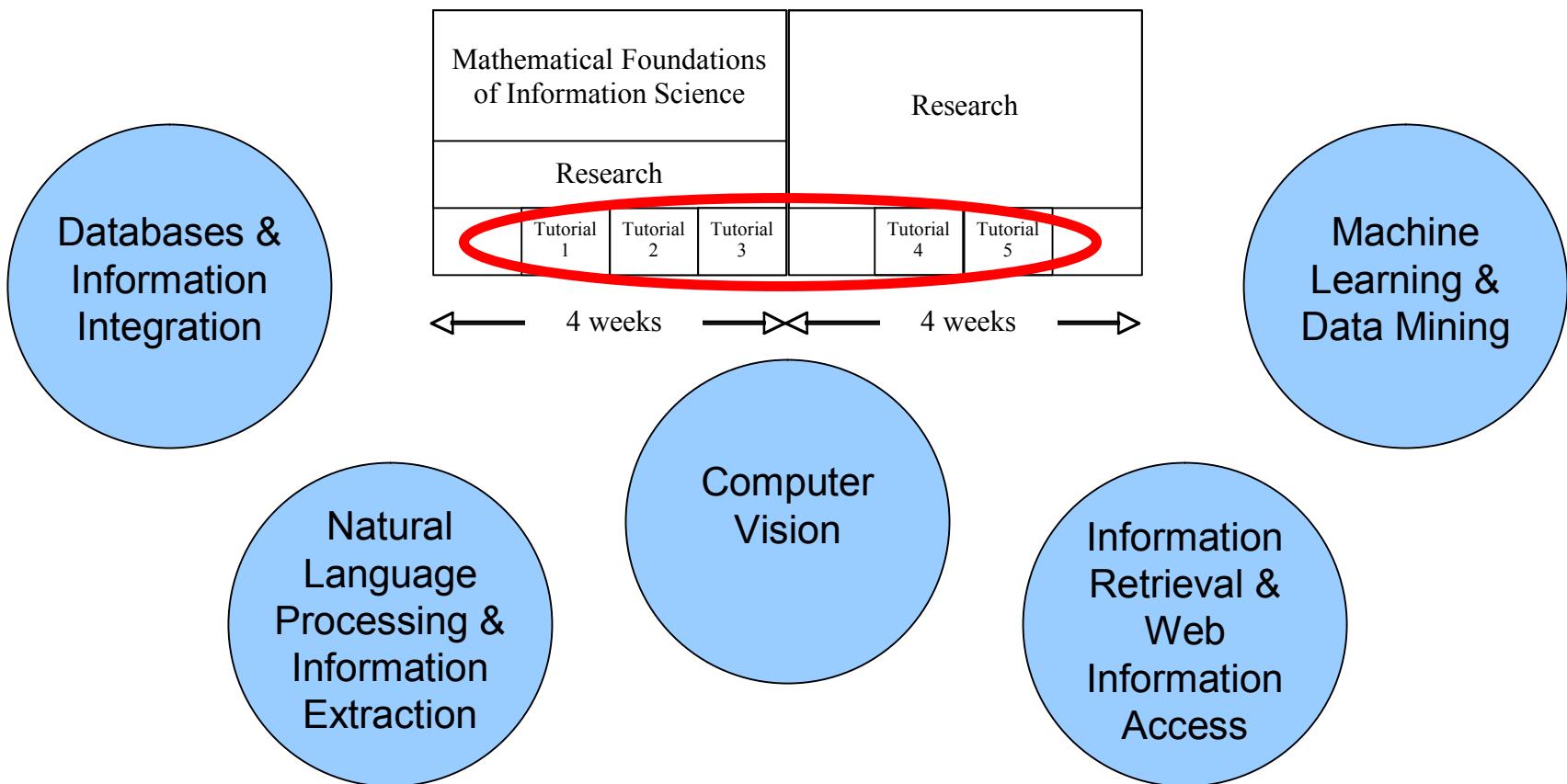
18



ILLINOIS

Data Science Summer Institute at UIUC

Advanced MIAS Related Tutorials



Information Access & Synthesis Processes

- Focused data retrieval and integration,
 - Identify and collect relevant data from multiple sources
- ★ Semantic data enrichment,
 - Infer semantics from unstructured data and images;
 - Allow navigation and search across disparate data modalities; augm
- ★ Entity identification and relations discovery,
 - Identify real-world entities and relations among them
 - Relate them to existing institutional resources for information integrat
- Knowledge discovery and hypotheses generation and verification
 - Construct the rich semantic structure and hidden networks of entit
- Foundations
 - Machine learning, database and data mining, natural language processing, inference and optimization and computer vision techniques
 - Called for and driven by the aforementioned problems.

Tools

Text

Processing&
Analysis

Semantic
Analysis &
Information
Extraction

Information
Integration

Machine
Learning &
Data Mining

Integrating
Text &
Images



MIAS

Multi-Modal Information Access & Synthesis