What I learned about Research in the Digital Age

Sanda Harabagiu

University of Texas at Dallas

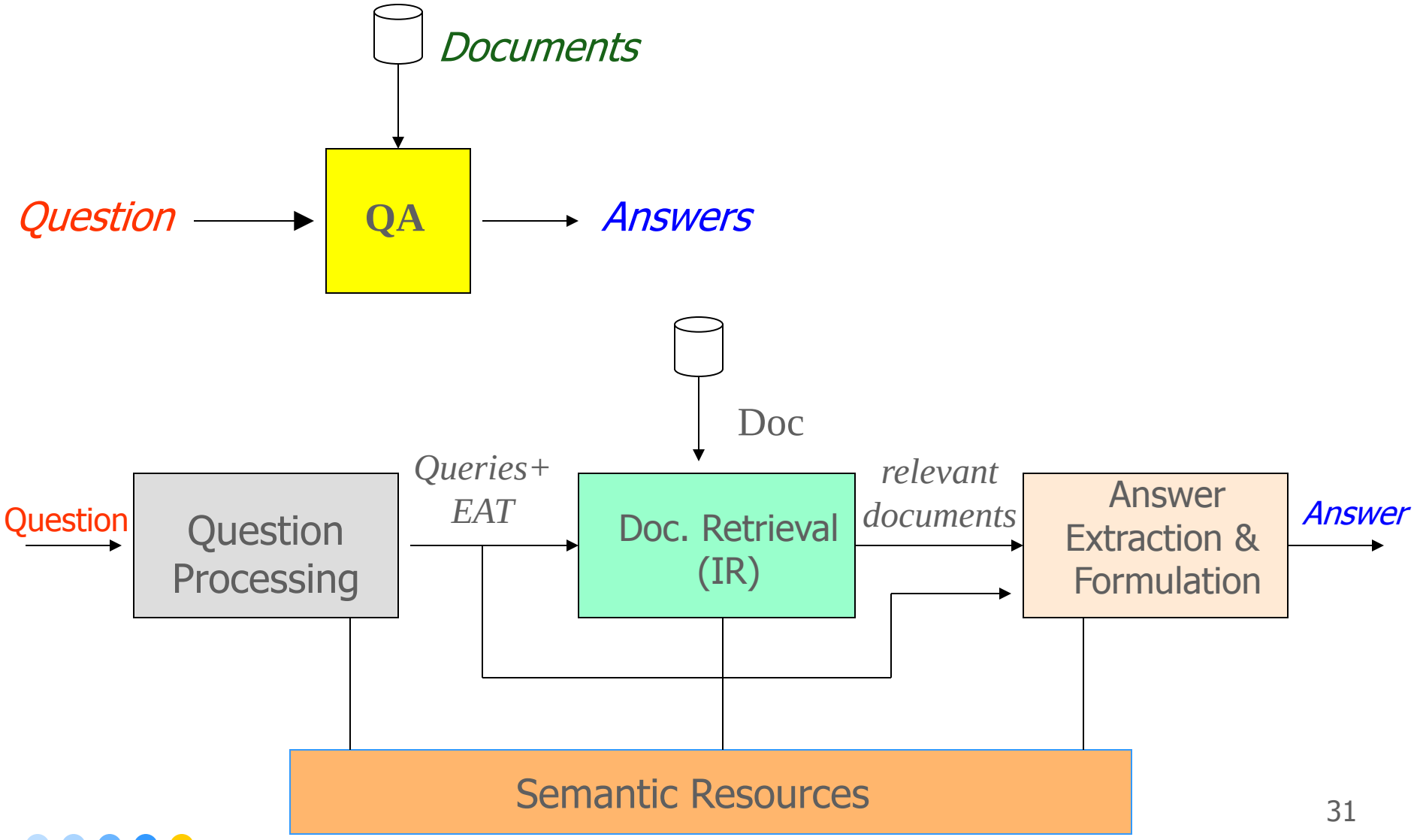
Why **Research** ????

- Because I always was *curious* and *enjoyed learning*
- Because being *creative* and *innovative* is akin to being an **ARTIST** of knowledge
- Research allows you to practice your art of knowledge !!!

Curiosity = Question Answering

- How could a system automatically find the answer to any question?
- How could you build a system?
- How could you represent knowledge?
- What is intelligence?
- How do humans communicate?
- Why we understand each other?
- Why is natural language difficult?
- What did I learn about semantics after building successful Q/A systems?

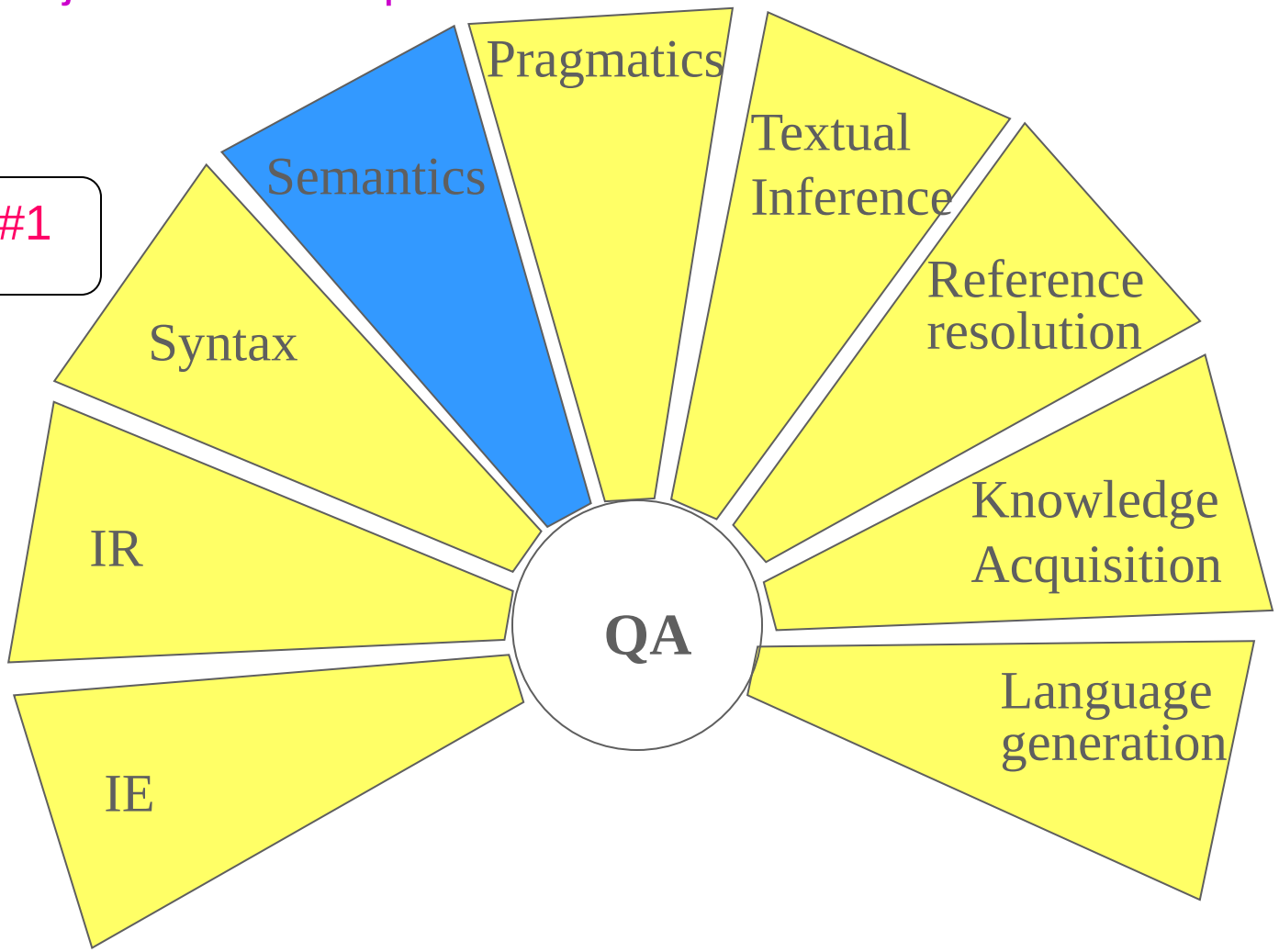
Textual Question Answering



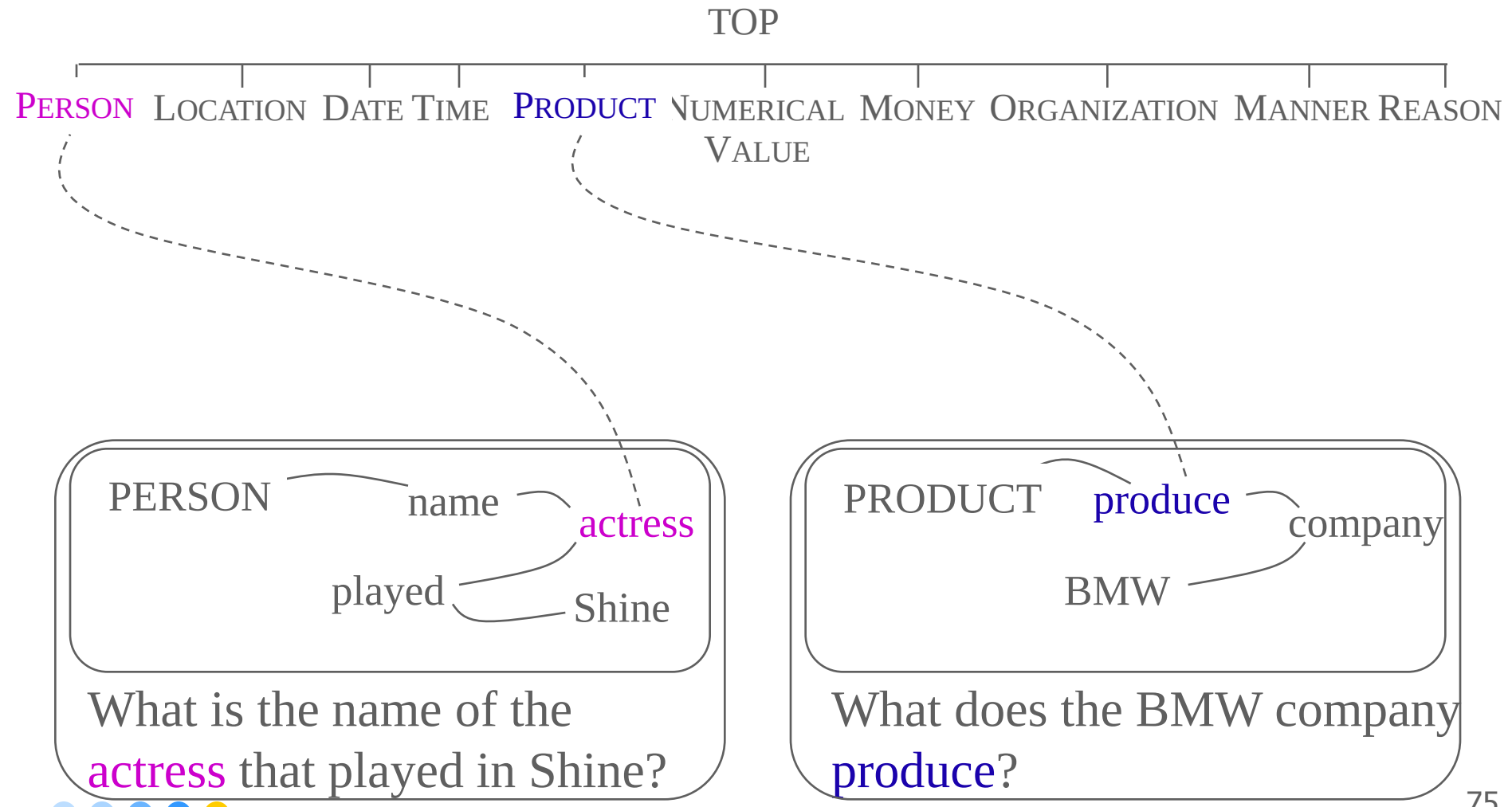
Problem complexity

Semantics is just one of the problems! But what kind of semantics do we need?

Lesson #1



Examples


TOP
PERSON
LOCATION
DATE TIME
PRODUCT
NUMERICAL VALUE
MONEY
ORGANIZATION
MANNER
REASON
PERSON
name
actress
played
Shine
What is the name of the actress that played in Shine?
PRODUCT
produce
company
BMW
What does the BMW company produce?

Semantic Problems with EATs

- The problem of assigning EATs
  - E.g. "manner questions":
  - Example *"How did Hitler die?"*

Lesson #2

- The problem of recognizing answer types/structures
  - Should *"manner of death"* be considered an answer type?
  - What other manner of event/action should be considered as answer types?

- The problem of recognizing EATs in texts
  - Should we learn to extract "manner" relations?
  - What other types of relations should we consider?
  - Is relation recognition sufficient for answering all types of questions? Is it necessary?

EAT = Manner-of-death

In TREC evaluations several questions asked about manner of death:

- *"How did Adolf Hitler die?"*

- Solution:
  - *We considered "Manner-of-Death" as an answer type, pointing to a variety of verbs and nominalizations encoded in WordNet*
  - We developed text mining techniques for identifying such information based on lexico-semantic patterns from WordNet
  - Example:
    - [*kill #sense1 (verb) – CAUSE → die #sense1 (verb)*]
  - Source of the troponyms of the [*kill #sense1 (verb)*] concept are candidates for the MANNER-OF-DEATH hierarchy
    - e.g., drown, poison, strangle, assassinate, shoot

Practical Hurdle

Lesson #3

Semantic information for EATs needs to be recognized by text mining techniques

- Not all MANNER-OF-DEATH concepts are lexicalized as verbs
  → we set out to determine additional patterns that capture such cases
- *Goal:* (1) set of patterns
  (2) dictionaries corresponding to such patterns
  → well known IE technique: (IJCAI'99, Riloff&Jones)

X { DIE / be killed } in ACCIDENT	seed: train, accident, car wreck (ACCIDENT)
X { DIE / be killed } {from|of} DISEASE	seed: cancer (DISEASE) AIDS
X { DIE } { after suffering / suffering of } MEDICAL CONDITION	seed: stroke, complications caused by diabetes (ACCIDENT)

- *Results: more than100 patterns were discovered*

How much information does the text wear on its sleeve?

By doing only:

- named entity recognition
- semantic classification of the expected answer type (off-line taxonomy + semantic info from WordNet)
- Text mining

→ 55% accuracy on factual trivia-like questions (*TREC-8, Moldovan et al.)*

What else is needed?

→ *Most questions cannot be processed successfully in this way, as they do not have a simple, conceptual EAT*

We need to consider additional forms of semantic knowledge and semantic processing.

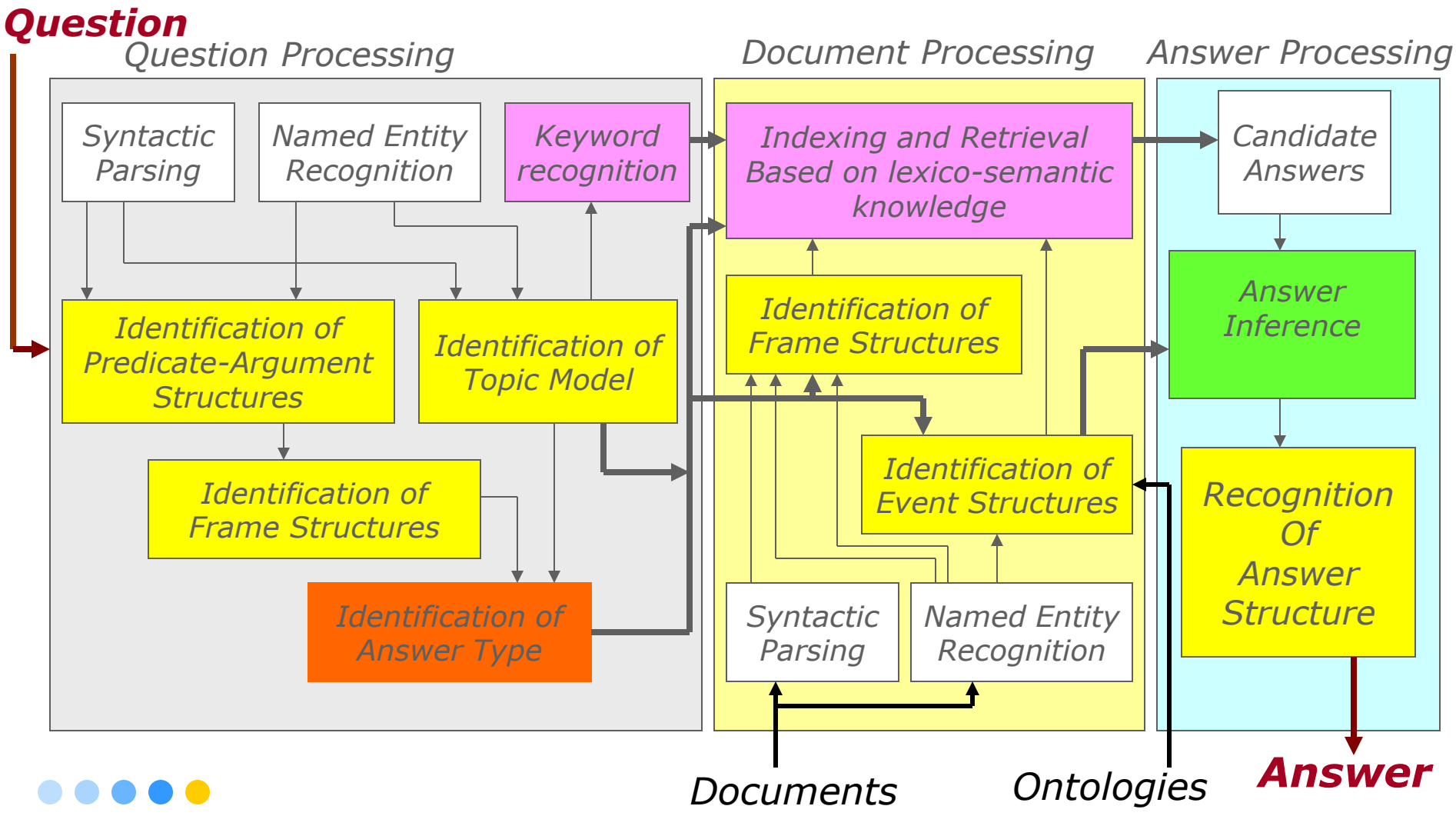
Lesson #4

AQUINAS – *Answering QUestions using INference and Advanced Semantics*

Collaborative Research Project
between UTD,
ICSI Berkeley and Stanford University

The driving rationale for our approach is that humans appear to have limited need for factoid question answering, but rather much more need to have systems that can deal with complex reasoning about causes, effects and chains of hypotheses.

QA architecture based on semantic structures



Applying Predicate-Argument Structures to QA

Predicate-arguments structures improve answer extraction!!!

Human Language Technology Research Institute

Lesson #5

- Parsing Questions

***Q**: What kind of materials were stolen from the Russian navy?*

PAS(Q): What [Arg1: kind of nuclear materials] were [Predicate:stolen] [Arg2: from the Russian Navy]?

- Parsing Answers

***A(Q)**: Russia's Pacific Fleet has also fallen prey to nuclear theft; in 1/96, approximately 7 kg of HEU was reportedly stolen from a naval base in Sovetskaya Gavan.*

PAS(A(Q)): [Arg1(P1redicate 1): Russia's Pacific Fleet] has [ArgM-Dis(Predicate 1) also] [Predicate 1: fallen] [Arg1(Predicate 1): prey to nuclear theft]; [ArgM-TMP(Predicate 2): in 1/96], [Arg1(Predicate 2): approximately 7 kg of HEU] was [ArgM-ADV(Predicate 2) reportedly] [Predicate 2: stolen] [Arg2(Predicate 2): from a naval base] [Arg3(Predicate 2): in Sovetskawa Gavan]

- Result: exact answer= *"approximately 7 kg of HEU"*

- Using FrameNet for QA
- *Example: What stimulated India's missile programs?*

FRAME: Stimulate

Frame Element CIRCUMSTANCES: ANSWER (part 1)
Frame Element EXPERIENCER: India's missile program
Frame Element STIMULUS: ANSWER (part 2)

FRAME: Subject_Stimulus

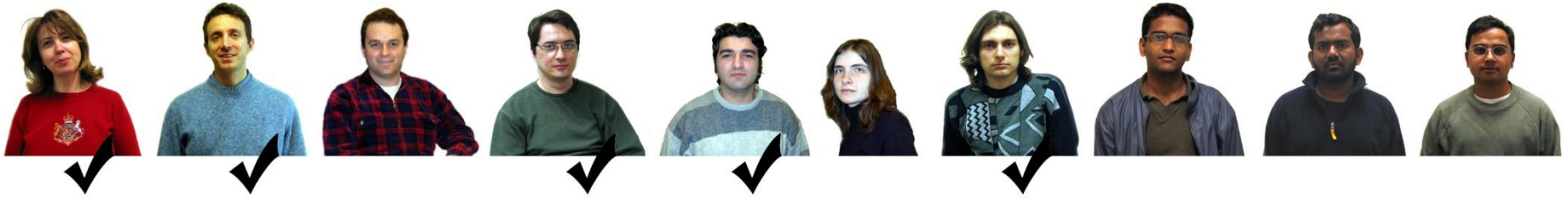
Frame Element CIRCUMSTANCES: ANSWER (part 3)
Frame Element COMPARISON SET: ANSWER (part 4)
Frame Element EXPERIENCER: India's missile program
Frame Element PARAMETER: nuclear proliferation

FrameNet Overview

- Produced by ICSI Berkeley [Baker et al.,1998] as a lexico-semantic resource encoding a set of frames (schematic representations of situations)

- Frames are characterized by:
  - target words or lexical predicates whose meaning includes aspects of the frame;
  - frame elements (FEs) which represent the semantic roles of the frame;
  - examples of annotations performed on the British National Corpus for instances of each target word.

- The project methodology was done on a *frame-by-frame basis*:
  - choosing a semantic frame (e.g. Commerce)
  - define the frame and its frame elements (e.g. BUYER, GOODS, SELLER, MONEY)
  - list the various lexical predicates which invoke the frame (buy, sell)
  - finding example sentences of each predicate in a corpus



Shallow Semantic Parsing Based on FrameNet

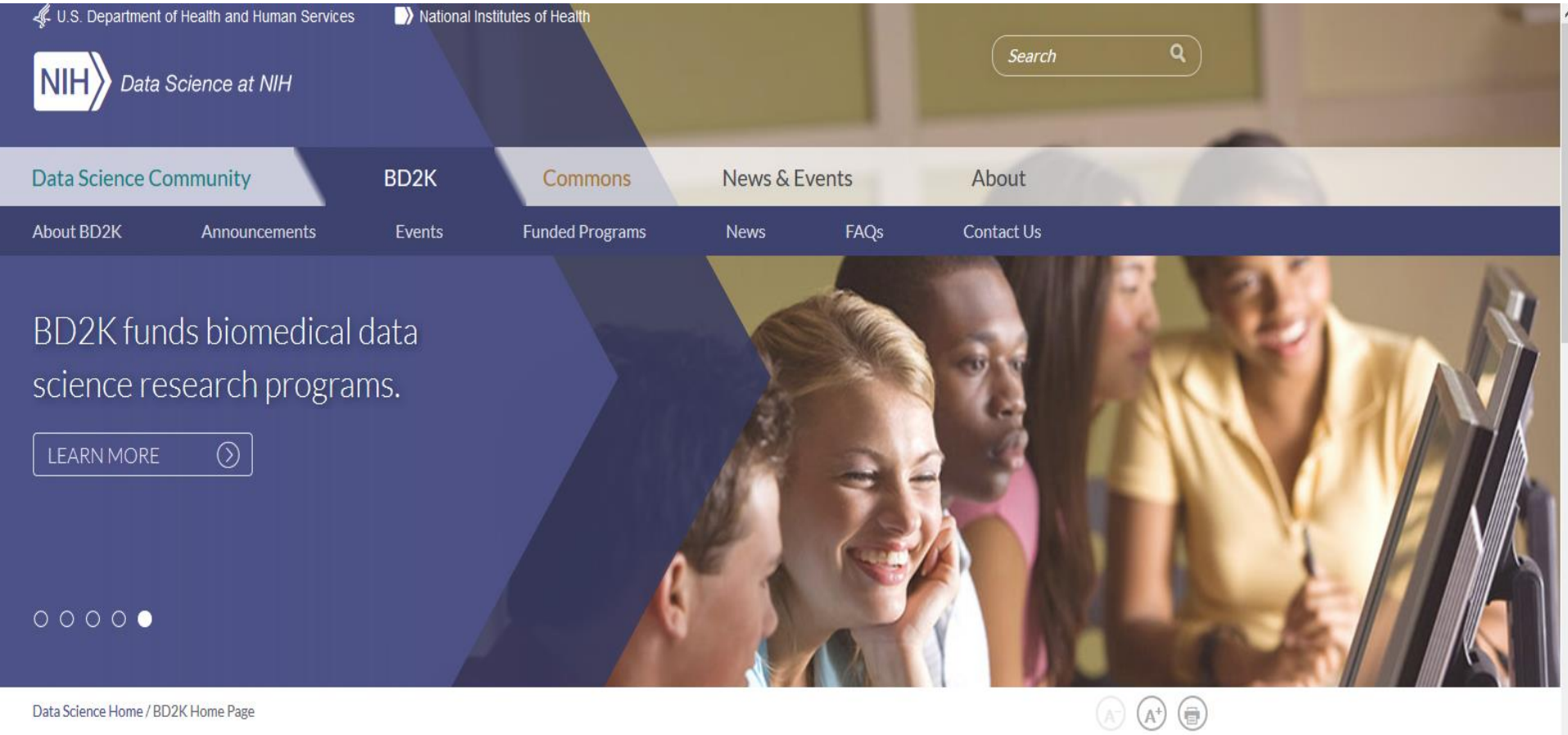
Cosmin Adrian Bejan, Alessandro Moschitti, Paul Morărescu, Gabriel Nicolae and Sanda Harabagiu

University of Texas at Dallas

Human Language Technology Research Institute

Capitalizing on Clinical Data

BIG DATA



Data Science Home / BD2K Home Page

Big Data to Knowledge (BD2K)

The ability to harvest the wealth of information contained in biomedical Big Data will advance our understanding of human health and disease; however, lack of appropriate tools, poor data accessibility, and insufficient training, are major impediments to rapid translational impact. To meet this challenge, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative in 2012.

BD2K Recent News

BD2KCenters Coordination Center Solicits Proposals for BD2K-Related Hackathons

The Big Data to Knowledge Centers Coordination Center is announcing a call for hackathon... *read more*

AUTOMATIC DISCOVERY AND PROCESSING OF EEG COHORTS FROM CLINICAL RECORDS

This award supports software development for automated explanatory modeling of complex healthcare data. The researchers develop a patient cohort retrieval system to provide big mechanism modeling capability for analysis of electroencephalogram (EEG) data.

Big mechanisms *have been defined as large explanatory models of complex systems with many causal interactions. The project is centered on the aggregation of clinical knowledge automatically discovered from EEG signals and EEG reports into a medical knowledge graph.*

The software framework established by this project could be transformative for mining the wealth of biomedical knowledge available from hospital medical records.

Research supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468.

Patient Cohort Retrieval:

Given a query modeling of the desired attributes of a patient sub-population, return a ranked list of patients that match this criteria.

In large EMRs:

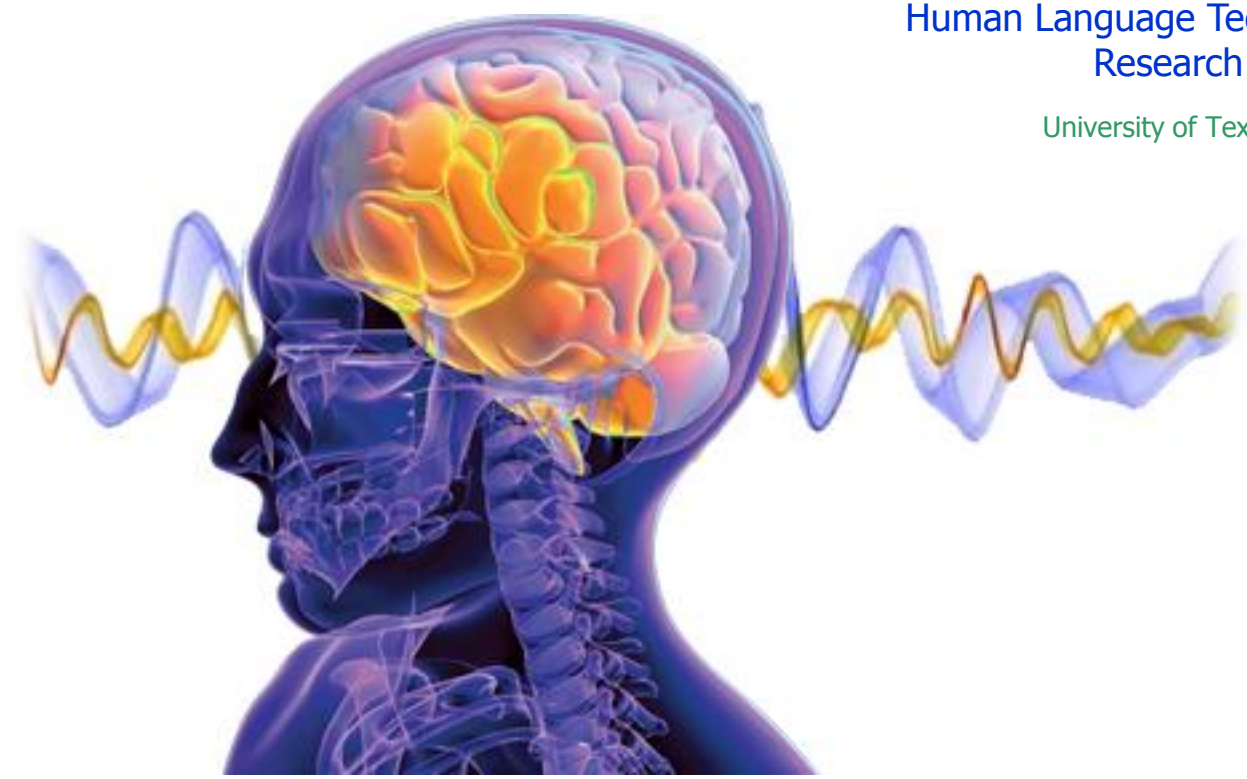
- 95k de-identified clinical records resulting from 18k hospital visits
- **Example Query:** *Patients diagnosed with localized prostate cancer and treated with robotic surgery.*

In EEG Reports:

- 20k de-identified EEG reports
- **Example Query:** *patients with occasional sharp waves suggesting a potential for seizures*

Human Language Technology
Research Institute

University of Texas at Dallas



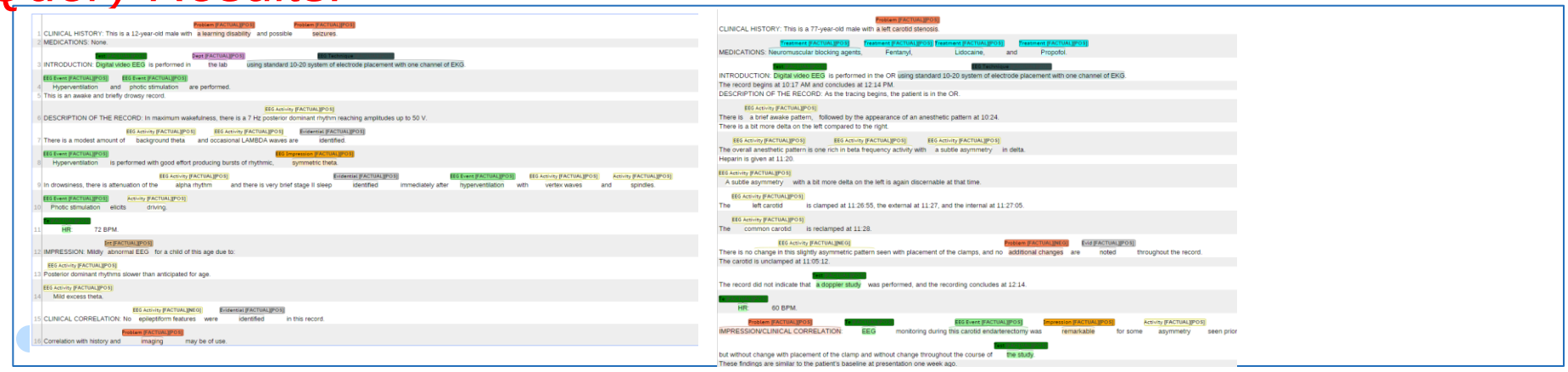
Electro-Enchephalography EEG

Patient Cohort Retrieval System

Query:

Patient Cohort

Query Results:



Temporal Disease Modelling:

The medical signs, symptoms, diagnoses, treatments, tests, and observations associated with a patient change over time. Thus, automatically reasoning about patients (e.g. for patient cohorts) can be improved by modelling the temporal aspect of the clinical picture and therapy.

Knowledge Discovery & Representation

Digital Age

BIG DATA →BIG KNOWLEDGE → BIG MECHANISMS

Thank you!!