Mining the Biomedical Research Literature

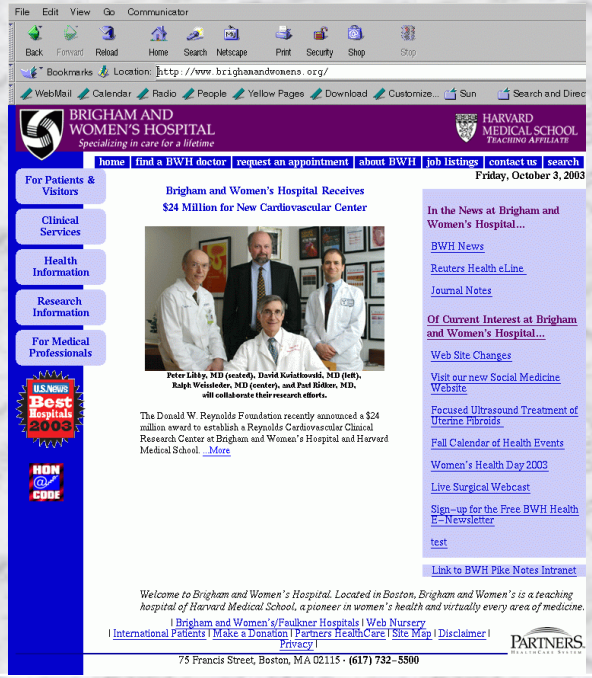
Ken Baclawski

Data Formats

- Flat files
- Spreadsheets
- Relational databases
- Web sites

```
011500   18.66   0   0  62   46.27102
011500   26.93   0   1  63   68.95152
020100   33.95   1   0  65   92.53204
020100   17.38   0   0  67   50.35111
```

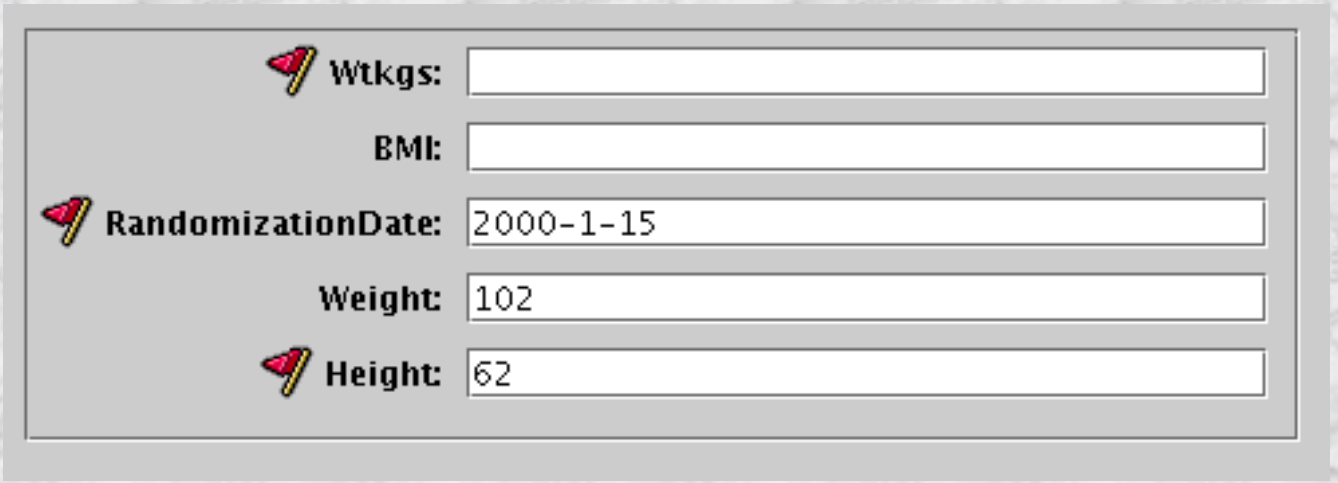
component	variable	initial_value	physical_unit	interface
membrane	u	-85.0	millivolt	out
membrane	Vr	-75.0	millivolt	out
membrane	Cm	0.01	microF_per_mm2	
membrane	time		millisecond	in
ionic_current	I_ion		microA_per_mm2	out
ionic_current	v			in
ionic_current	Vth		millivolt	in



XML Documents

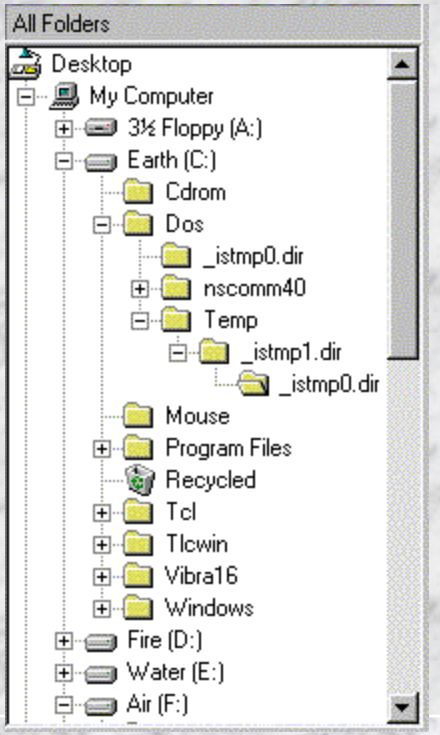
- Flexible very popular text format
- Self-describing records

```
<Interview RandomizationDate="2000-01-15" BMI="18.66" Height="62" Weight="102" ... />
<Interview RandomizationDate="2000-01-15" BMI="26.93" Height="63" Weight="152" ... />
<Interview RandomizationDate="2000-02-01" BMI="33.95" Height="65" Weight="204" ... />
<Interview RandomizationDate="2000-02-01" BMI="17.38" Height="67" Weight="111" ... />
```



XML Documents (continued)

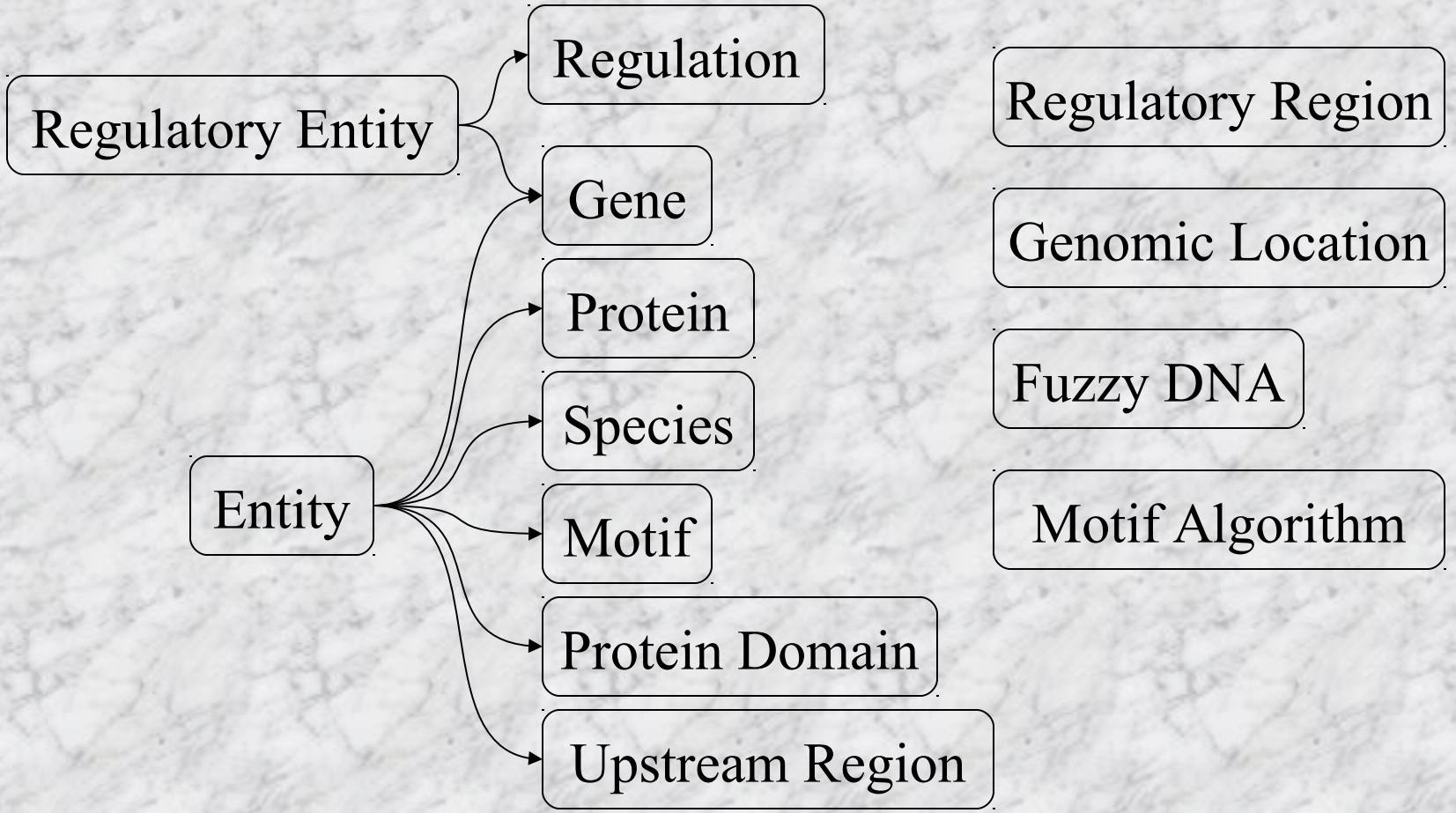
- Hierarchical structure

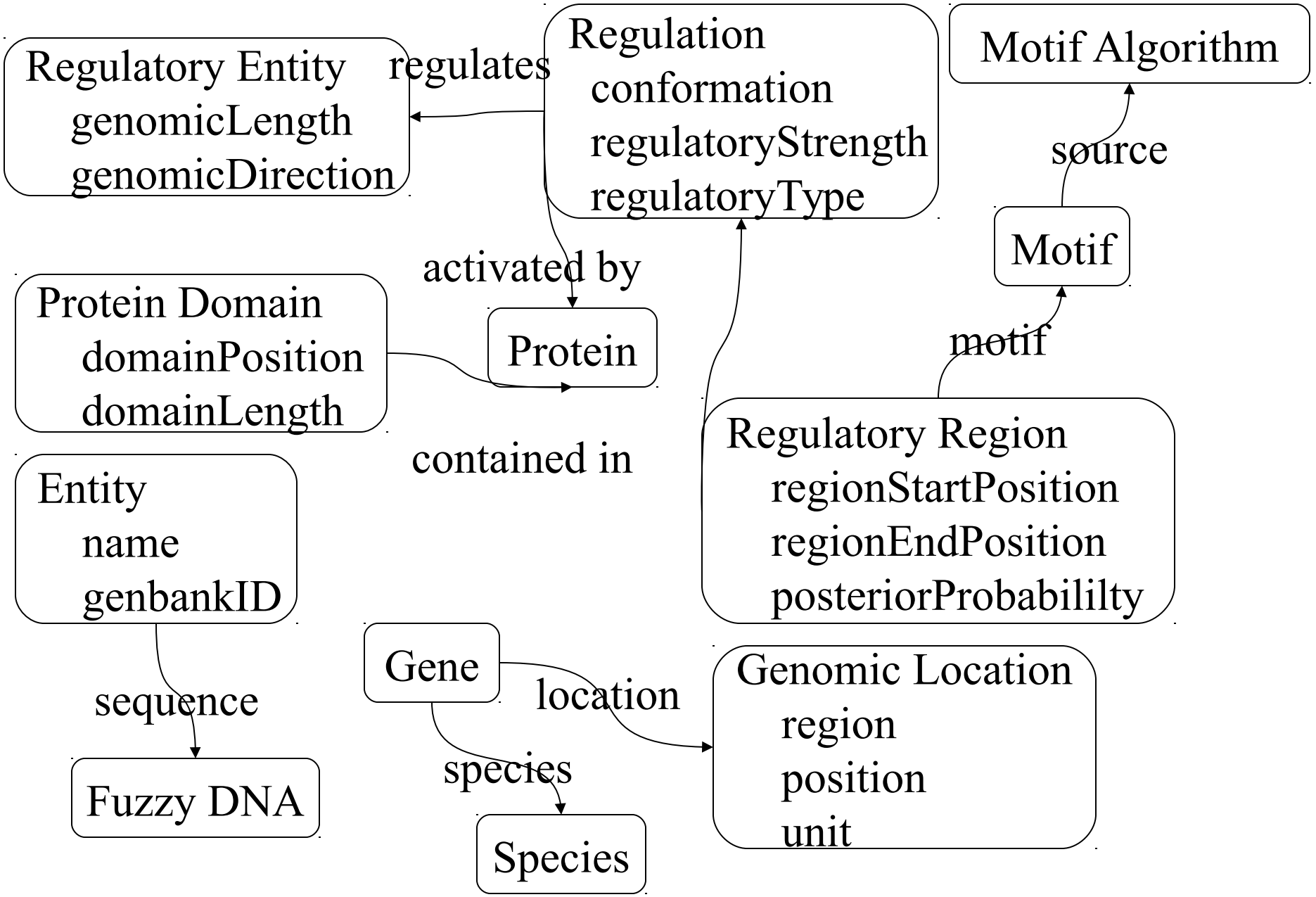


Ontologies

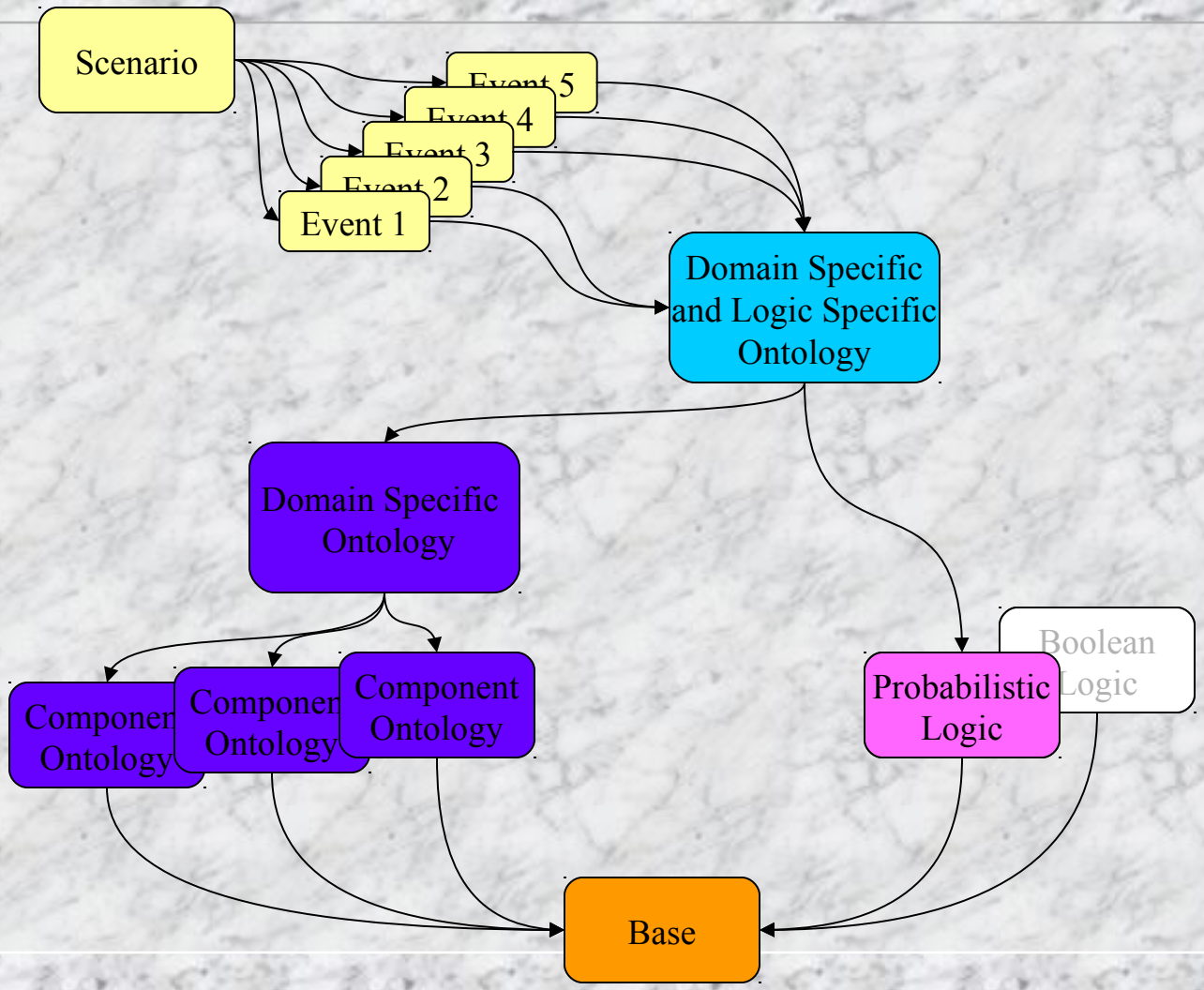
- An ontology defines the concepts and relationships between them in a domain.
- Philosophers speak of "the" ontology and define it informally. In Computer Science there are many ontologies and they are formally defined.
- The structure of data is its ontology.
  - Database schema
  - XML Document Type Definition (DTD)

Gene Regulation Ontology




Regulatory Entity
genomicLength
genomicDirection
regulates
Regulation
conformation
regulatoryStrength
regulatoryType
Motif Algorithm
source
Motif
activated by
Protein Domain
domainPosition
domainLength
Protein
motif
contained in
Regulatory Region
regionStartPosition
regionEndPosition
posteriorProbabililty
Entity
name
genbankID
sequence
Fuzzy DNA
Gene
location
Genomic Location
region
position
unit
species
Species

Constructing Large Ontologies



Some Ontology Languages

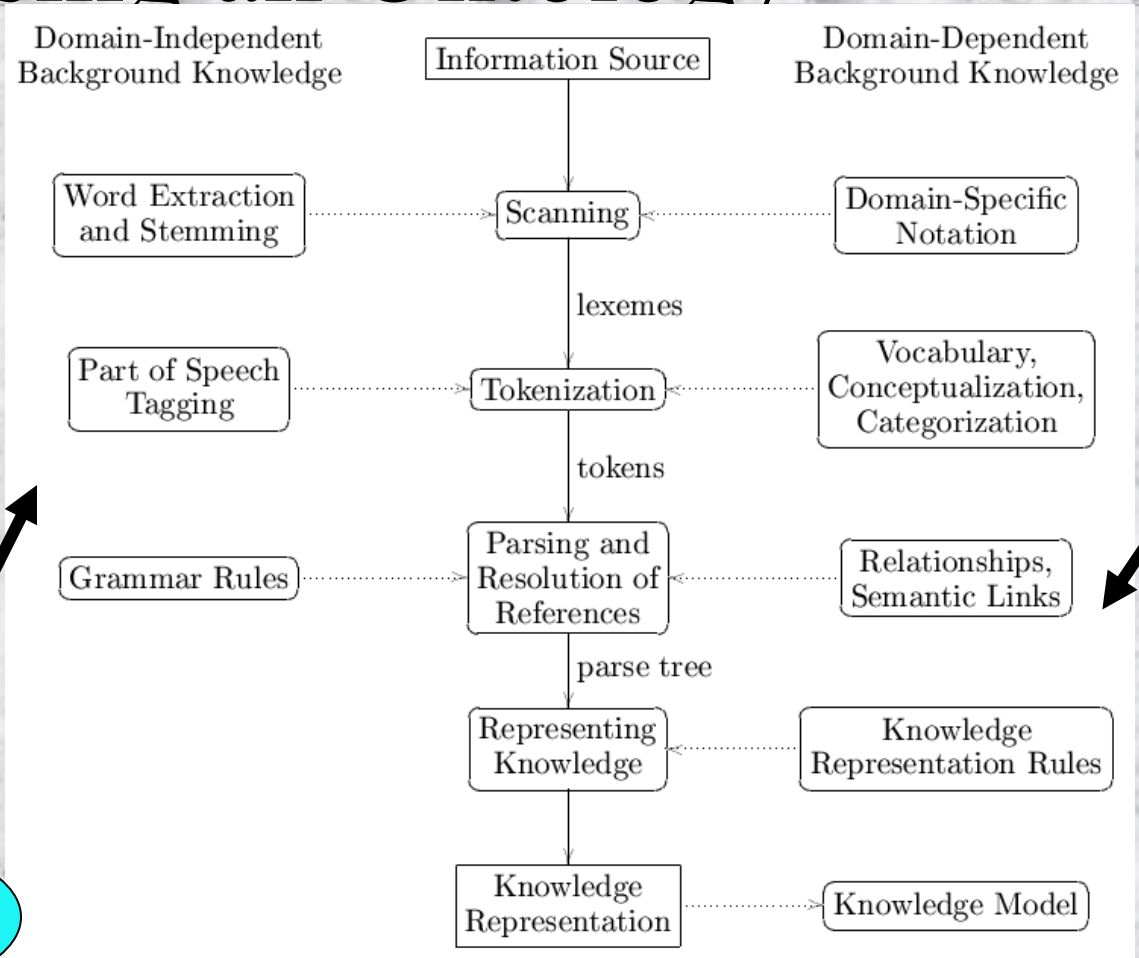
- Established languages
  - Knowledge Interchange Format (KIF)
  - XML Schema (XSD)
  - Resource Description Framework (RDF)
  - XML Topic Maps (XTM)
- Emerging languages
  - Common Logic
  - Web Ontology Languages (OWL)
  - Ontology Definition Metamodel (ODM)

Biomedical Ontologies

- Gene Ontology (GO)
- Unified Medical Language System (UMLS)
- BioPolymer Markup Language (BioML)
- Systems Biology ML (SBML)
- MicroArray Gene Expression ML (MAGE-ML)
- Protein XML (PROXIML)
- CellML
- RNAML
- Chemical ML (CML)
- Medical ML (MML)
- CytometryML
- Taxonomic ML (TML)

- Semantic Categories
  - > 130 semantic categories
- Semantic Relationships
  - " is a ", " part of", "disrupts"
- Semantic Concepts (Vocabulary)
  - > 1,000,000 concepts map to categories

Natural Language Processing using an Ontology

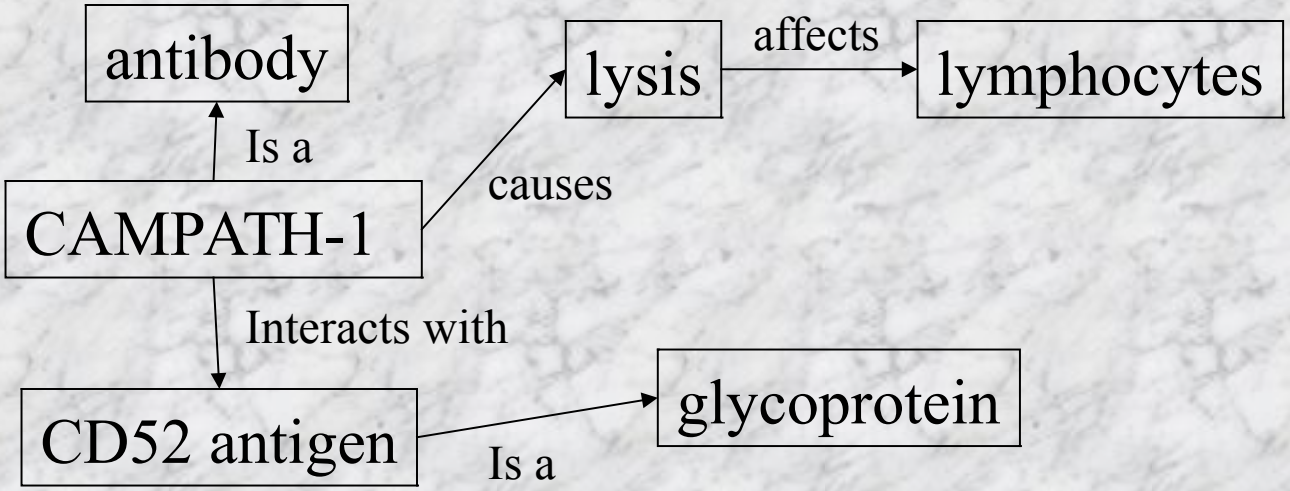


semantic

syntactic

Example of knowledge extraction

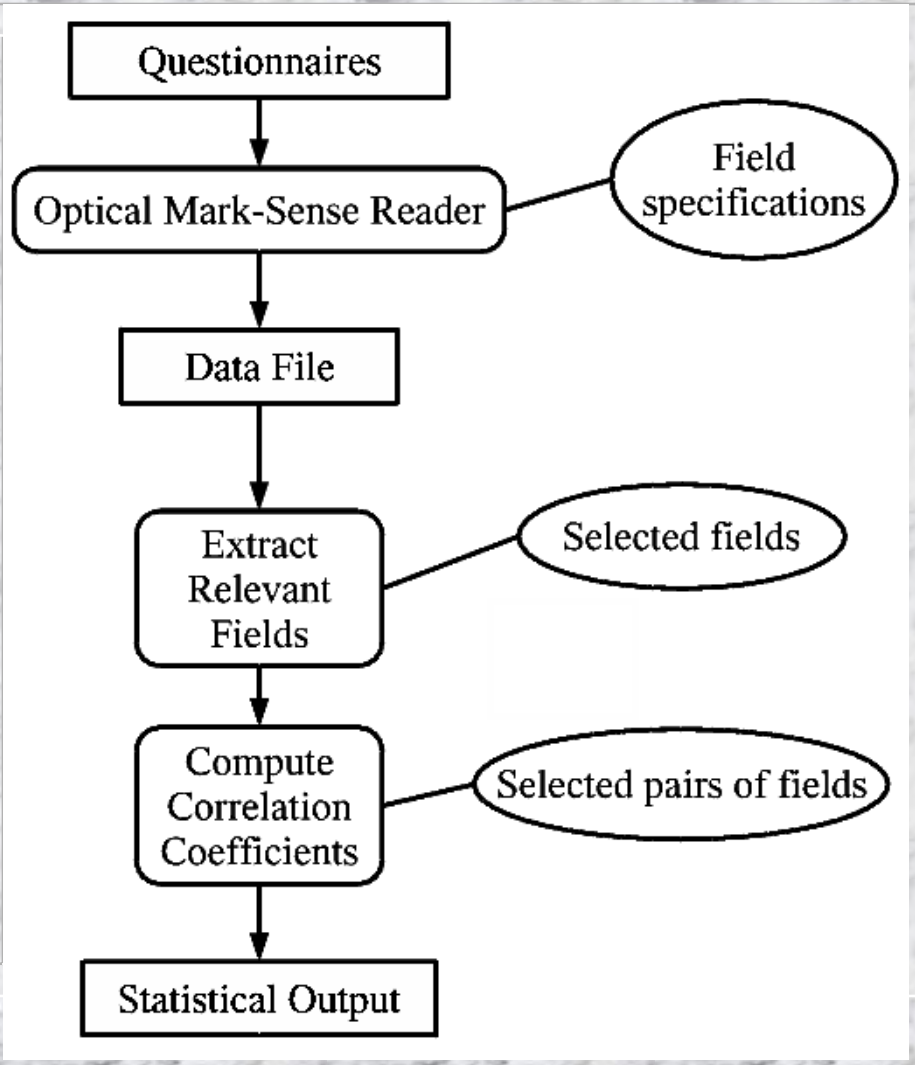
CAMPATH-1 antibodies recognize the CD52 antigen which is a small lipid-anchored glycoprotein abundantly expressed on T cells, B cells, monocytes and macrophages. They lyse lymphocytes …



Purpose of Data

- Data is collected and stored for a purpose.
- The format serves that purpose.
- Using data for another purpose is common.
- It is important to anticipate that data will be used for many purposes.
- Data is reused by *transforming* it.

Statistical Analysis



- Transformation consists of a series of steps.
- Specialized equipment and software is used for each step.
- Separation into steps reduces the overall effort.

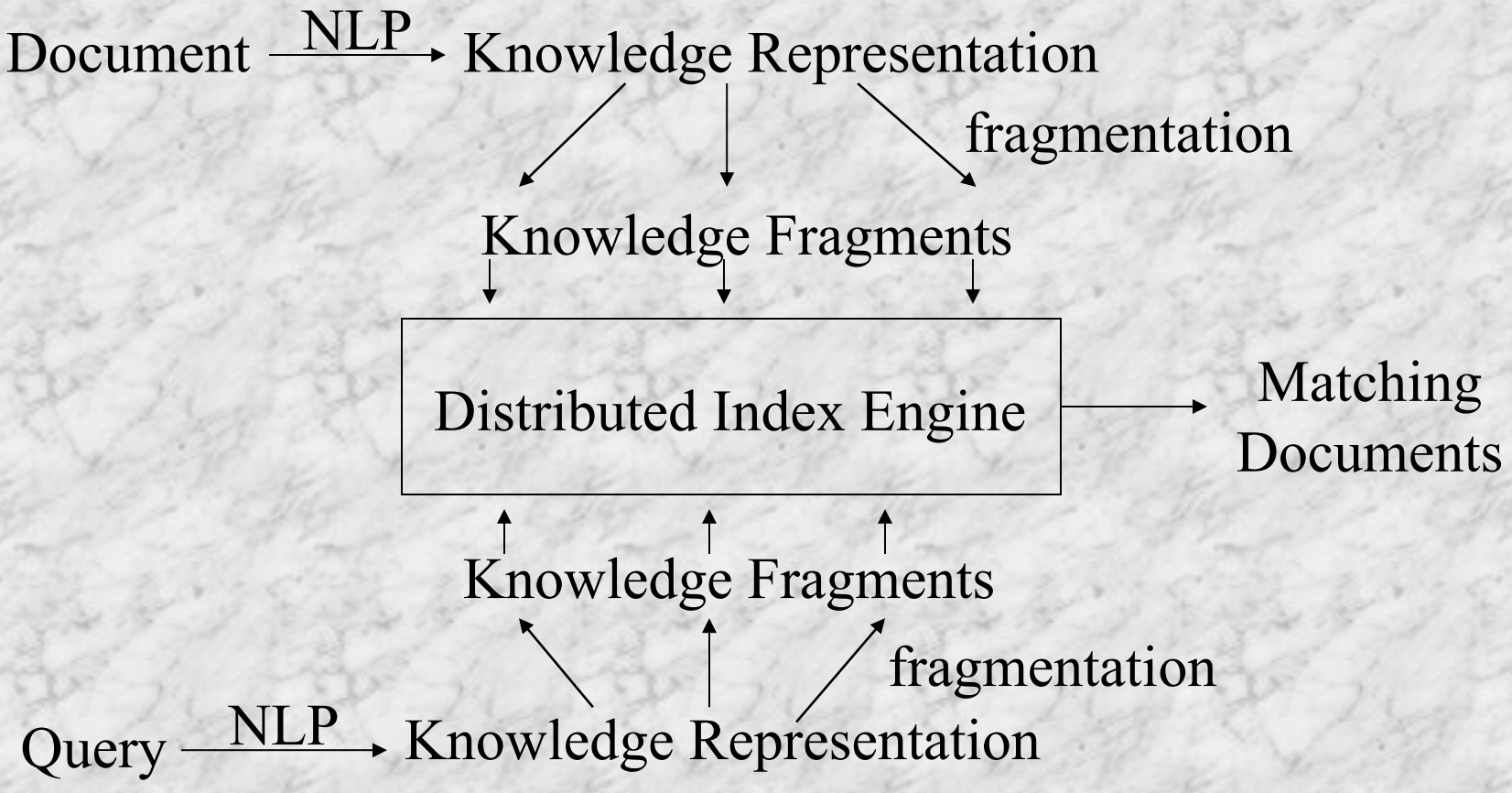
Statistical Models

- The “selection” step can involve much more than just choosing fields of a record:
  - Data can be rescaled, discretized, ...
  - Data in several fields can be combined.
  - The statistical model can be much more complex (such as a Bayesian network).
- In general, data is transformed to a different ontology: A statistical model *is* an ontology.

Transformation Languages

- Traditional programming languages such as Perl, Java, etc.
- Rule-based (declarative) languages such as the XML Transformation language (XSLT).
  - Rule-based rather than procedural
  - Transform each kind of element with a template
  - Matching and processing of elements is analogous to the digestion of polymers with enzymes.

High Performance Indexing



Consistency Checking

- Logical consistency means that a formal theory has at least one interpretation.
- Inconsistency is to be avoided.
- Probabilistic consistency means that a probabilistic model is likely to have an interpretation.
- Probabilistic inconsistency is significant.

Research Challenges

- Inference and deduction
  - Logical inference
  - Probabilistic inference
  - Scientific inference
  - Other forms of inference
- Integrating inference with
  - Data mining
  - Experimental processes

Phase Transitions and Undecidability