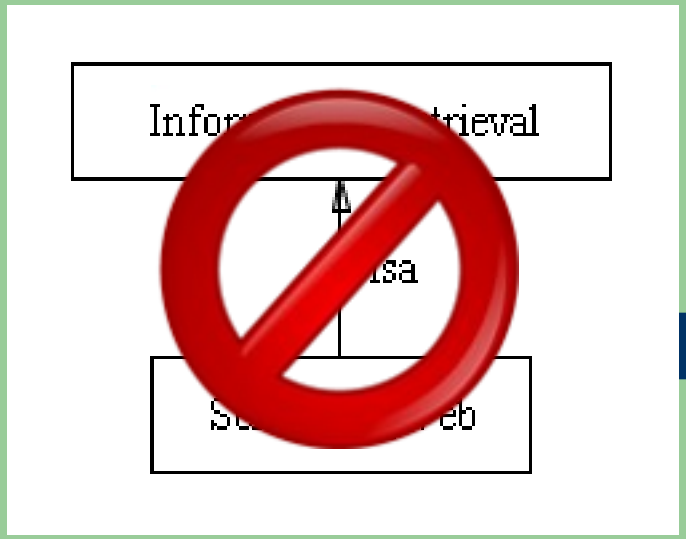
# The Semantic Web: It's not just for searching anymore!



Ken Baclawski
Northeastern University
Vistology

# The Semantic Web and the wide variety of emerging applications

- Introduction to the Semantic Web
- General classification and recognition of opportunities
  - Interoperability and integration
  - Web Services and composite applications
  - Records management
- Examples of projects and applications
  - Project Halo
  - Collaboration tools
  - Cognitive radio
  - Policy awareness
  - Behavioral health
  - Epidemiology and disease tracking
- Recent developments

# The Semantic Web

- The World Wide Web is a versatile infrastructure for basic data availability.
- The main emphasis was on human-mediated interactions via web browsers but new uses are rapidly increasing.
- These new uses can benefit from semantic technologies.

# The Resource Description Framework

- RDF is a language for representing information about resources in the web.
- While RDF is expressed in XML, it has different semantics.
- The document-centric semantics of XML is replaced by a semantics based on triples: (subject, predicate, object).
- RDF decouples information from the containing document.

# RDF Semantics

- All relationships are explicit and labeled with a property resource.
- The distinction in XML between attribute and containment is dropped, but the containment relationship must be labeled on a separate level. This is called *striping.*

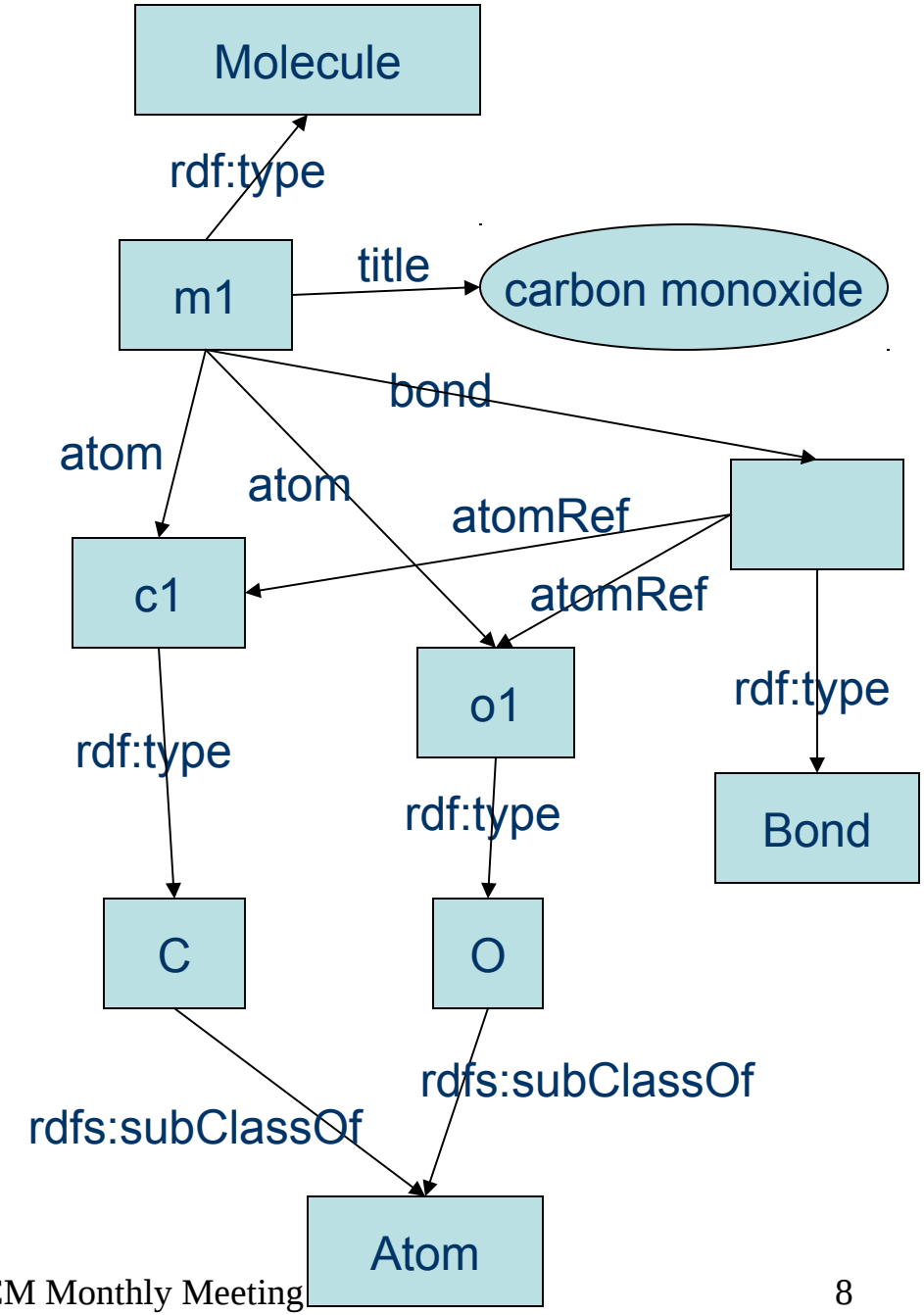
# XML Element Hierarchy

```
<bioml>
  <organism name="Homo sapiens (human)">
    <chromosome name="Chromosome 11" number="11">
      <locus name="HUMINS locus">
        <reference name="Sequence databases">
          <db_entry name="Genbank sequence" entry="v00565" format="GENBANK"/>
          <db_entry name="EMBL sequence" format="EMBL" entry="V00565"/>
        </reference>
        <gene name="Insulin gene">
          <dna name="Complete HUMINS sequence" start="1" end="4992">
          1 ctcgaggggc ctagacattg ccctccagag agagcaccca acaccctcca ggcttgaccg
          ...
          </dna>
          <ddomain name="flanking domain" start="1" end="2185"/>
          <ddomain name="polymorphic domain" start="1340" end="1823"/>
          <ddomain name="Signal peptide" start="2424" end="2495"/>
          <exon name="Exon 1" start="2186" end="2227"/>
          <intron name="Intron 1" start="2228" end="2406"/>
        </gene>
      </locus>
      <locus>
      ...
      </locus>
    </chromosome>
  </organism>
</bioml>
```

```
...
    <locus name="HUMINS locus">
      <contains>
        <gene name="Insulin gene">
          <isStoredIn>
            <db_entry name="Genbank sequence" entry="v00565"
                      format="GENBANK"/>
            <db_entry name="EMBL sequence" format="EMBL"
                      entry="V00565"/>
          </isStoredIn>
          <isCitedBy>
            <db_entry name="Insulin gene sequence" format="MEDLINE"
                      entry="80120725"/>
            <db_entry name="Insulin mRNA sequence" format="MEDLINE"
                      entry="80236313"/>
            <db_entry name="Localization to Chromosome 11" format="MEDLINE"
                      entry="93364428"/>
          </isCitedBy>
          <hasSequence>
            <dna name="Complete HUMINS sequence" start="1" end="4992">
              1 ctcgaggggc ctagacattg ccctccagag agagcaccca acaccctcca ggcttgaccg
              ...
            </dna>
          </hasSequence>
        </gene>
      </contains>
    </locus>
...
```

# RDF graph for carbon monoxide

```
<Molecule rdf:id="m1"
      title="carbon monoxide">
 <atom>
  <C rdf:id="c1"/>
  <O rdf:id="o1"/>
 </atom>
 <bond>
  <Bond>
   <atomRef rdf:resource="c1"/>
   <atomRef rdf:resource="o1"/>
  </Bond
 </bond>
</Molecule>
```



# The Web Ontology Language

- OWL is based on RDF and has three increasingly general levels: OWL Lite, OWL-DL, and OWL Full.
- OWL adds many new features to RDF:
  - Functional properties
  - Inverse functional properties (database keys)
  - Local domain and range constraints
  - General cardinality constraints
  - Inverse properties
  - Symmetric and transitive properties

# Class Constructors

- OWL classes can be constructed from other classes in a variety of ways:
  - Intersection (Boolean AND)
  - Union (Boolean OR)
  - Complement  (Boolean NOT)
  - Restriction
- Class construction is the basis for *description logic.*

# OWL Semantics

- An OWL ontology defines a theory of the world. States of the world that are consistent with the theory are called *interpretations* of the theory.
- A fact that is true in every interpretation is said to be *entailed* by the theory. Logical inference in OWL is defined by entailment.
- Entailment can be counter-intuitive, especially when it entails that two resources are the same.

# Identifying opportunities

- Domain knowledge
  - Technical background
  - Community organization
- Identify urgent needs
- Understand the trends
  - Short-term evolution
  - Possible paradigm shifts
- Semantic technology is only one part of any solution but it can be an important enabler.

# Search and retrieval

- Data is typically stored in either record/data structures or natural language.
- Need is to search and retrieve both kinds of data for a single query.
- There are several trends.
  - More semantics
  - Integration with other services
- Semantic technologies are more than just a fancy search and retrieval mechanism.

# Interoperability of legacy systems

- Legacy systems and databases are characterized by:
  - A large variety of formats
  - High degree of complexity
  - Many technologies of various ages
- Need to interoperate and integrate
- Trend is toward encoding more semantics in the data representation itself.
- Opportunity to develop products and services for interoperability and integration.

# Web services and composite applications

- The web is being used not only for retrieval of data but also for using tools and services.
- The need is to find the required services, and to get them to communicate with each other.
- The trend is to use semantic annotation to describe/advertise services, to express requests, and to represent the responses, but the level of semantic annotation is very uneven.
- The opportunity is to built agile workflow management tools that can deal with the differing levels of semantic annotation.

# Simple Semantic Web Architecture and Protocol (SSWAP)

- SSWAP is a protocol for semantic web services.  See http://sswap.info
- Unlike other protocols, SSWAP uses a single format and protocol for description, registration, discovery and invocation.
- SSWAP was developed using OWL as its basis, and OWL inference is fundamental to its operation.

# Records management

- Solving an electronic record problem will add little to the existing paper-based records if the systems are not interoperable.
- Simply automating paper-based processes has relatively little impact on productivity.
- Gains in efficiency and improved customer relationships require a change in the overall process of service delivery.

# Records Opportunity

- Develop event ontologies that:
  - Support interoperability
  - Are independent of workflows and processes
  - Are compatible with existing processes
- Develop products that:
  - Assist organizations to evolve toward electronic data management
  - Serve the interests of many stakeholders

# Halo Program at Vulcan

- Knowledge Representation in Practice: Proj ect Halo and the Semantic Web
  by Mark Greaves
- The vision: a scalable knowledge representation and reasoning system
  - Gets better with increasing scale
  - Embraces uncertain and incomplete information
- The system: scientific question-answering

# Halo Pilot

- Pilot project was on AP Chemistry.
- Typical question: “What are the reaction products if metallic copper is heated strongly with concentrated sulfuric acid?”
- Answer: $Cu^{2+}$, $SO_2(g)$, and $H_2O$
- Should also be able to explain the answer.

# Halo Pilot

- SRI, Ontoprise and Cycorp competed.
- The challenge achieved an AP level 3 on 70 pages of the Chemistry AP syllabus.
- Cost: $10K per page
- Most errors were due to lack of domain expertise by the ontology developers.

# Halo Phase II

- Knowledge acquisition performed by subject matter experts (not computer scientists)
- Expanded to cover Physics and Biology
- Cost: $100 per page
- Achieved the same AP level.
- http://www.projecthalo.com

# Halo Project today

- Goal is to achieve an AP level 4.
- Scale up the knowledge acquisition
  - Offshoring in India
  - Large scale collaborative ontology development
  - Semantic Wikis
- Ultimate goal is a *Digital Aristotle*
- Semantically enabled collaboration is an important new emphasis.

# Collaboration tools

- People need to collaborate to solve problems.
- The need is to support rapid team formation and problem solving even when the people are geographically dispersed.
- The trend is to use wikis and blogs rather than face-to-face meetings.
- The challenge is to develop tools that facilitate collaboration over the web without losing the advantages of face-to-face meetings.

# Wikis

- Wikis are a popular tool for collaboration.
- They have been used for rapid team formation and collaboration.
- They have a number of disadvantages:
  - Mix of natural language and untyped links.
  - Focus is on simplicity and presentation, not structure and semantics.

# Semantic Wikis

- A wiki with an underlying knowledge model (ontology) is a *semantic wiki.*
- Data in the wiki is annotated with meta-data in RDF or OWL.
- Links are typed and annotated, also in RDF or OWL.
- Machines can infer new facts from the explicitly asserted facts.
- Search and retrieval are facilitated by the semantics.
- Interoperability is greatly improved.

# Semantic Media Wiki

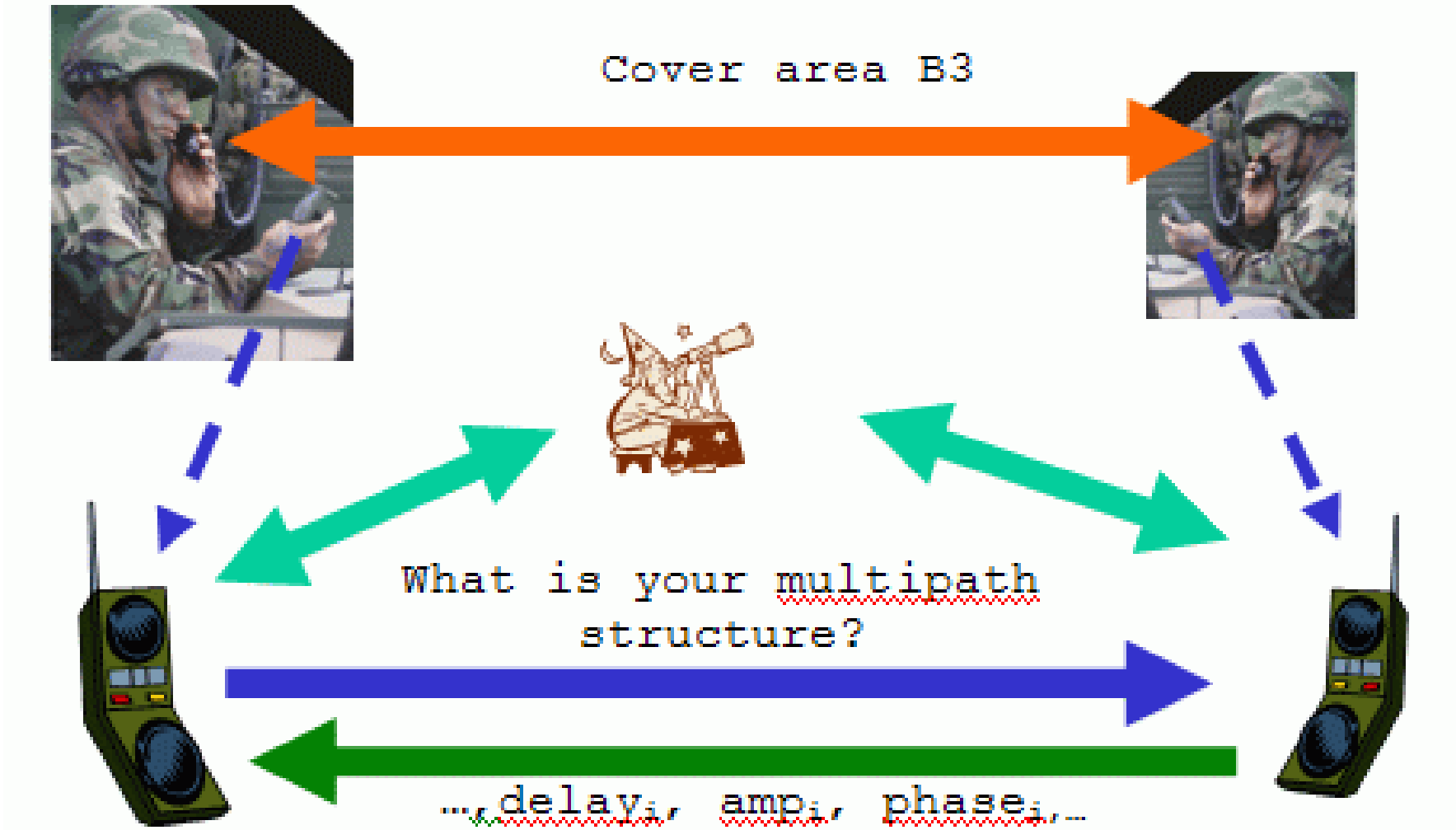
- Media Wiki is the technology of Wikipedia and related web sites.
- Semantic Media Wiki is a large ($100M+) EU project based in Karlsruhe.
- The Halo project provided the Halo extension .
- Fine grained access will soon be available via the PMWX project.

# Cognitive Radios

- Capabilities of a cognitive radio:
  - information collection and fusion;
  - self-awareness;
  - awareness of constraints and requirements;
  - query by user, self or other radio;
  - command execution;
  - dynamic interoperability at any stack layer;
  - situation awareness and advise;
  - negotiation for resources.

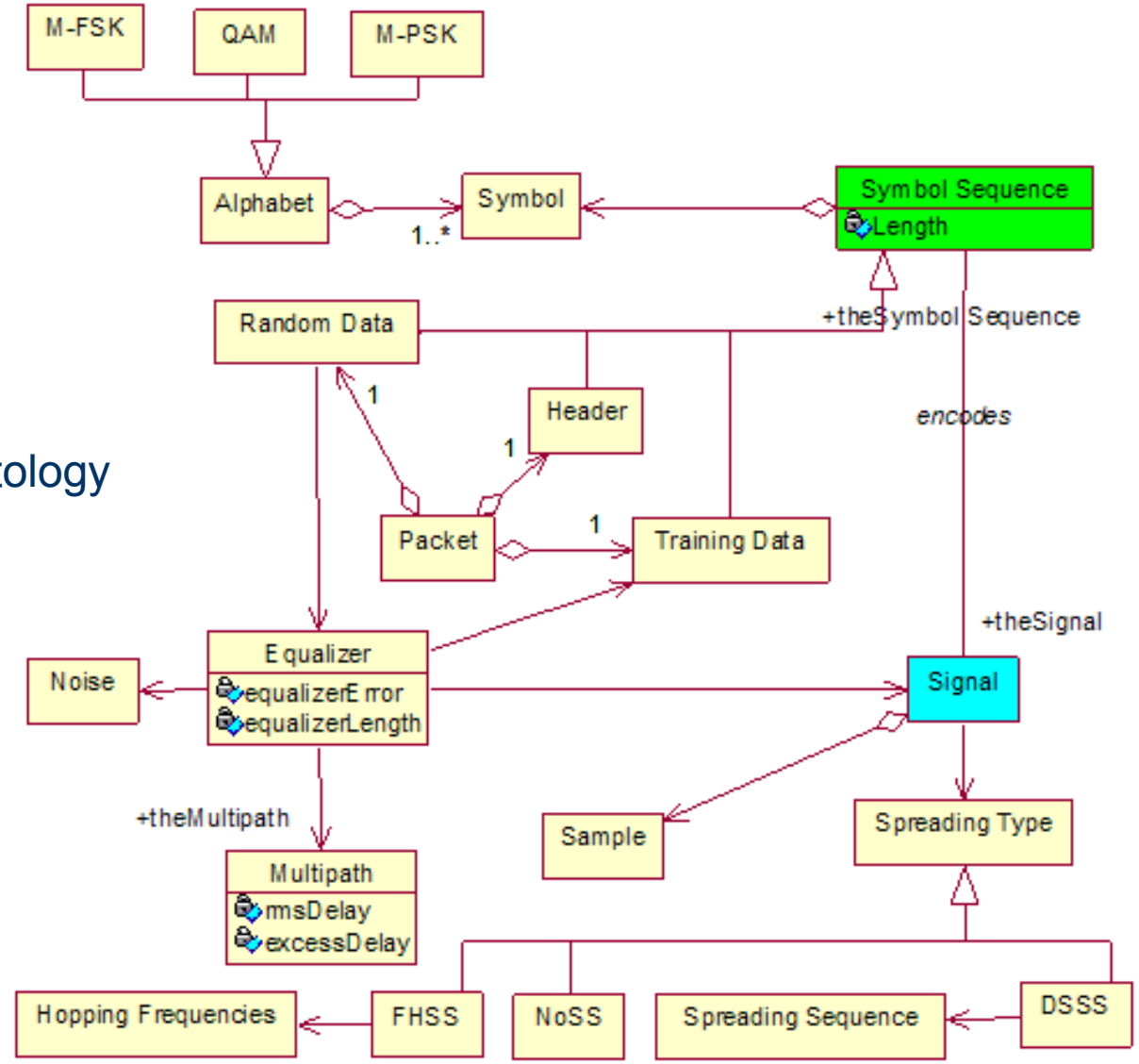
# Definition of a cognitive system

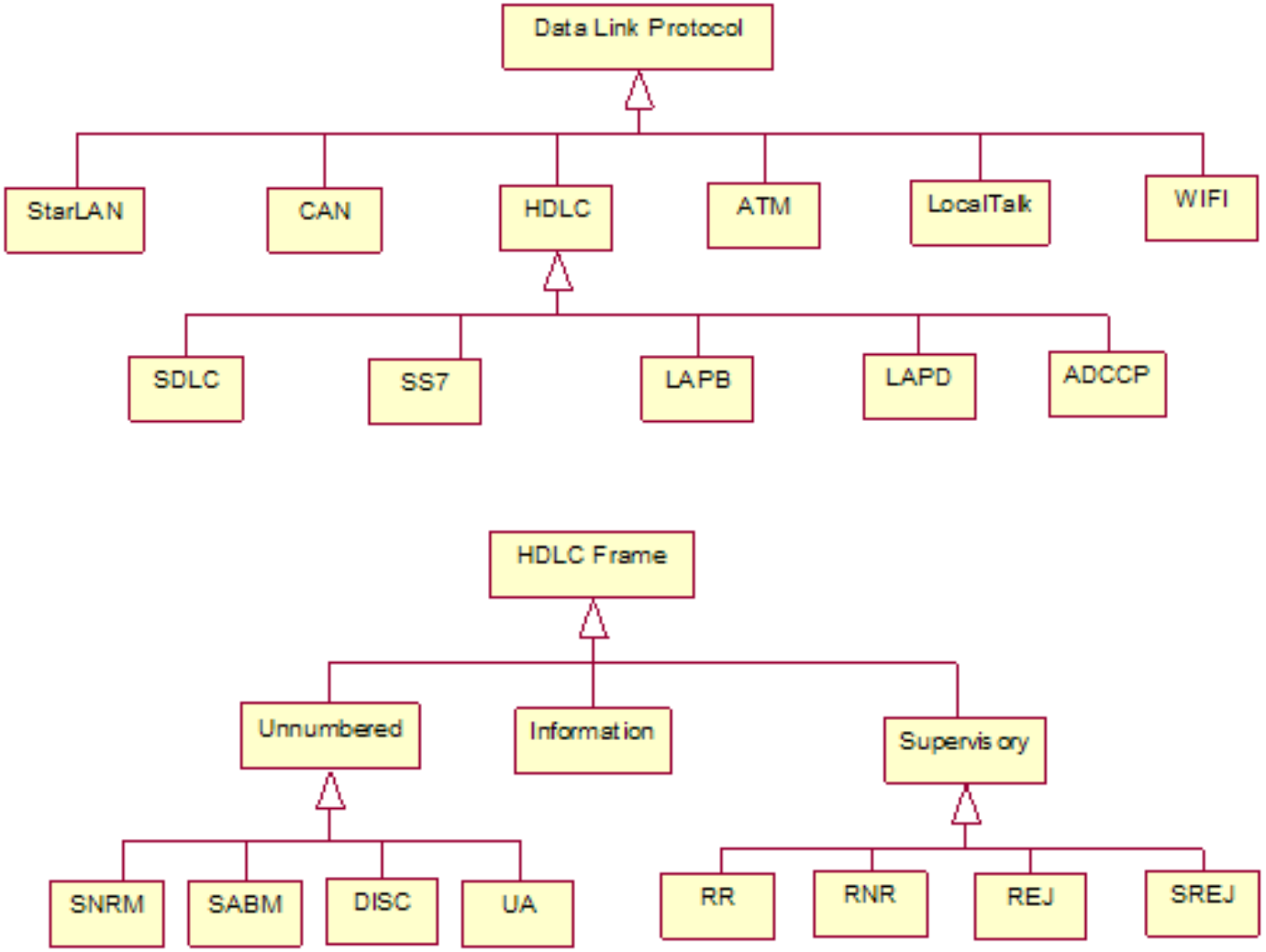
- can **reason**, using substantial amounts of appropriately represented knowledge
- can **learn** from its experience so that it performs better tomorrow than it did today
- can **explain itself** and be told what to do
- can be **aware** of its own capabilities and **reflect** on its own behavior
- can **respond robustly** to surprise



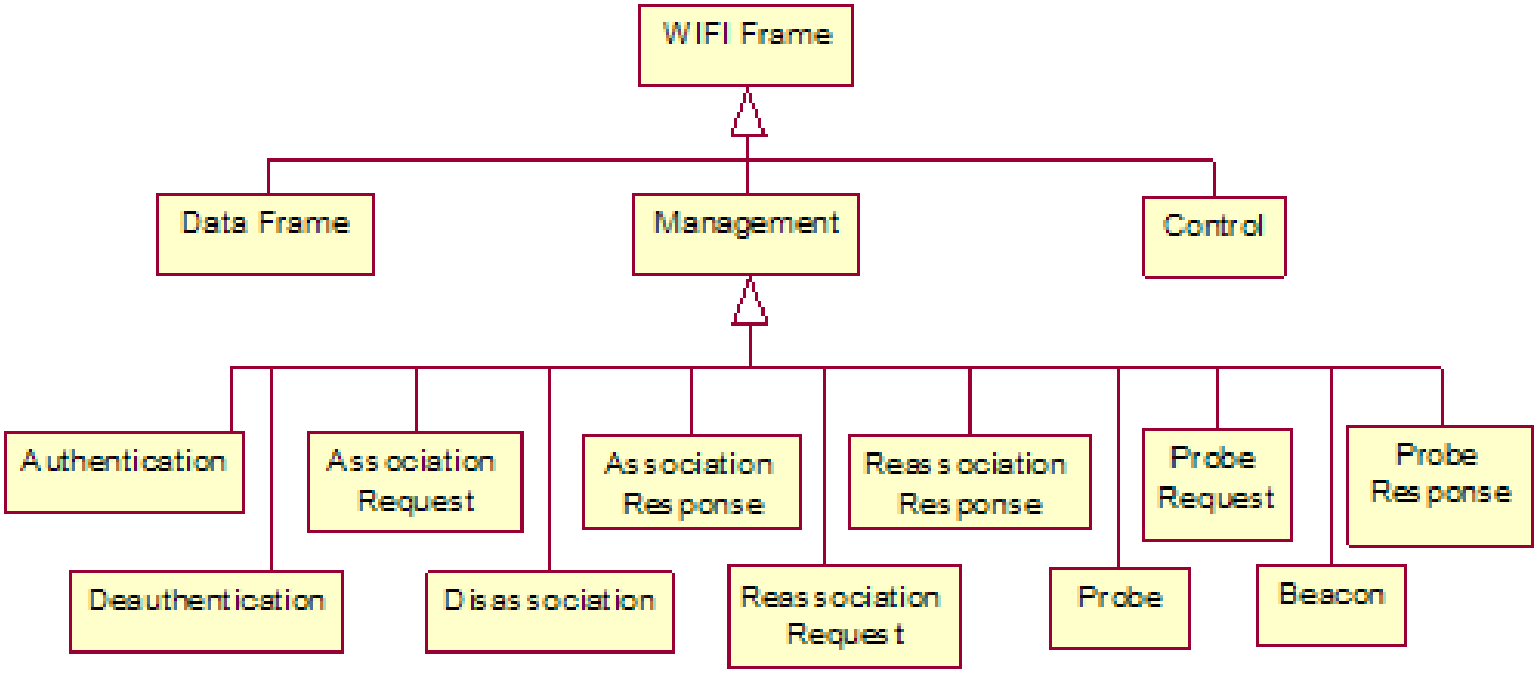

Multiple levels of communication

## Physical Layer Ontology






Some Data Link layer hierarchies




Data Link WiFi Frame Hierarchy

# Role of Semantic Technology in Cognitive Radio

- *Interoperability*
- *Flexible querying and Run-time modifiability*
  - Programming language reflection allows the algorithm to be queried at run time without having any explicit pre-programmed monitoring capability.
- *Validation*
  - Formalization allows one to check the consistency of protocols.
- *Self-awareness*
  - Communication nodes can understand their own structure and modify their functioning at run-time based on this understanding.
- *Policy management.*

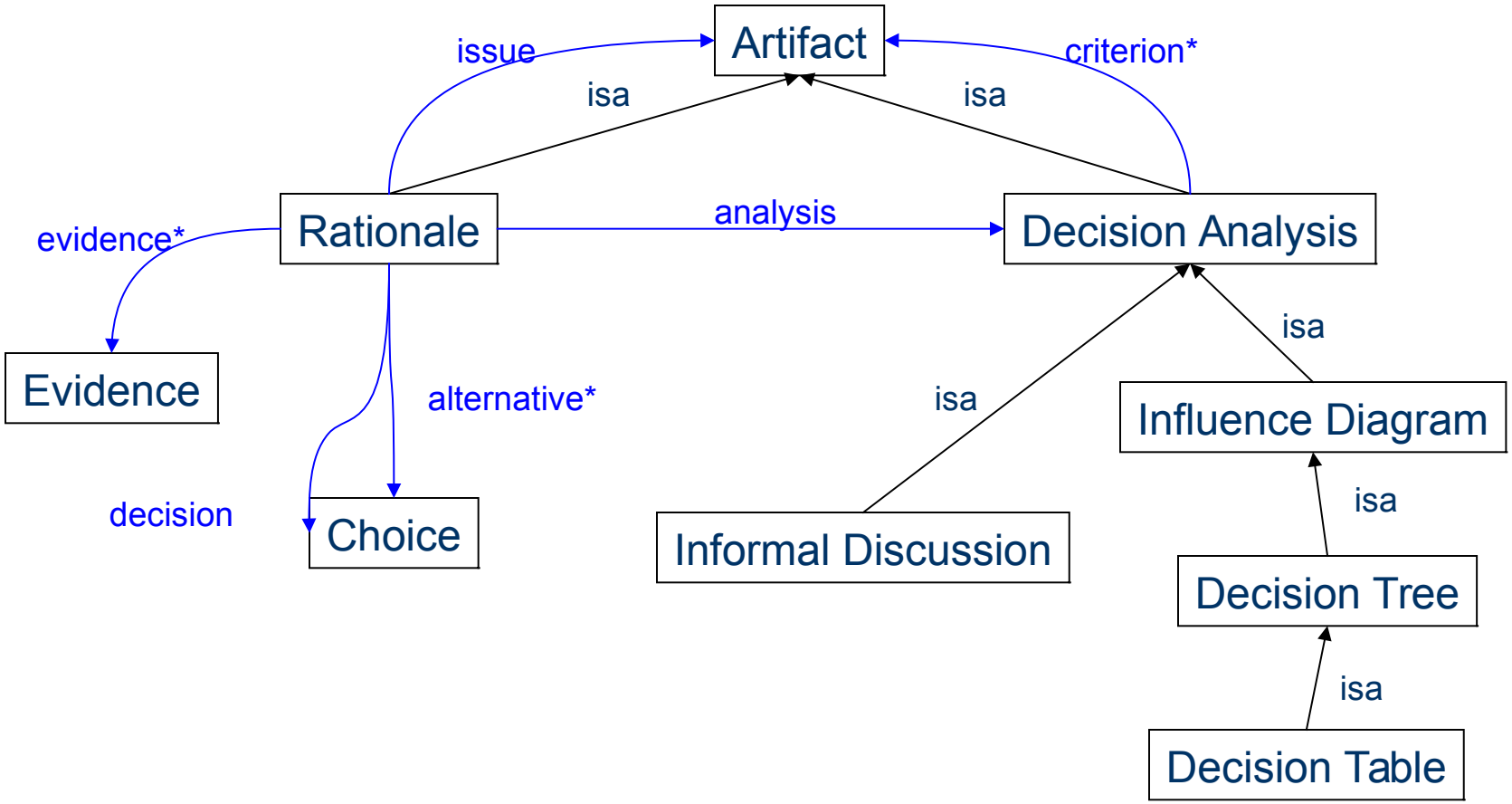
# Policy Awareness

- An important trend that is driving cognitive radio is the need for radios for flexible use of spectrum
- However, any use of the spectrum must conform to legal policies.
- Policies are expressed as rules.
- Ontologies make it possible to specify regulations for wireless communications, including complex, dynamic policies for spectrum management.

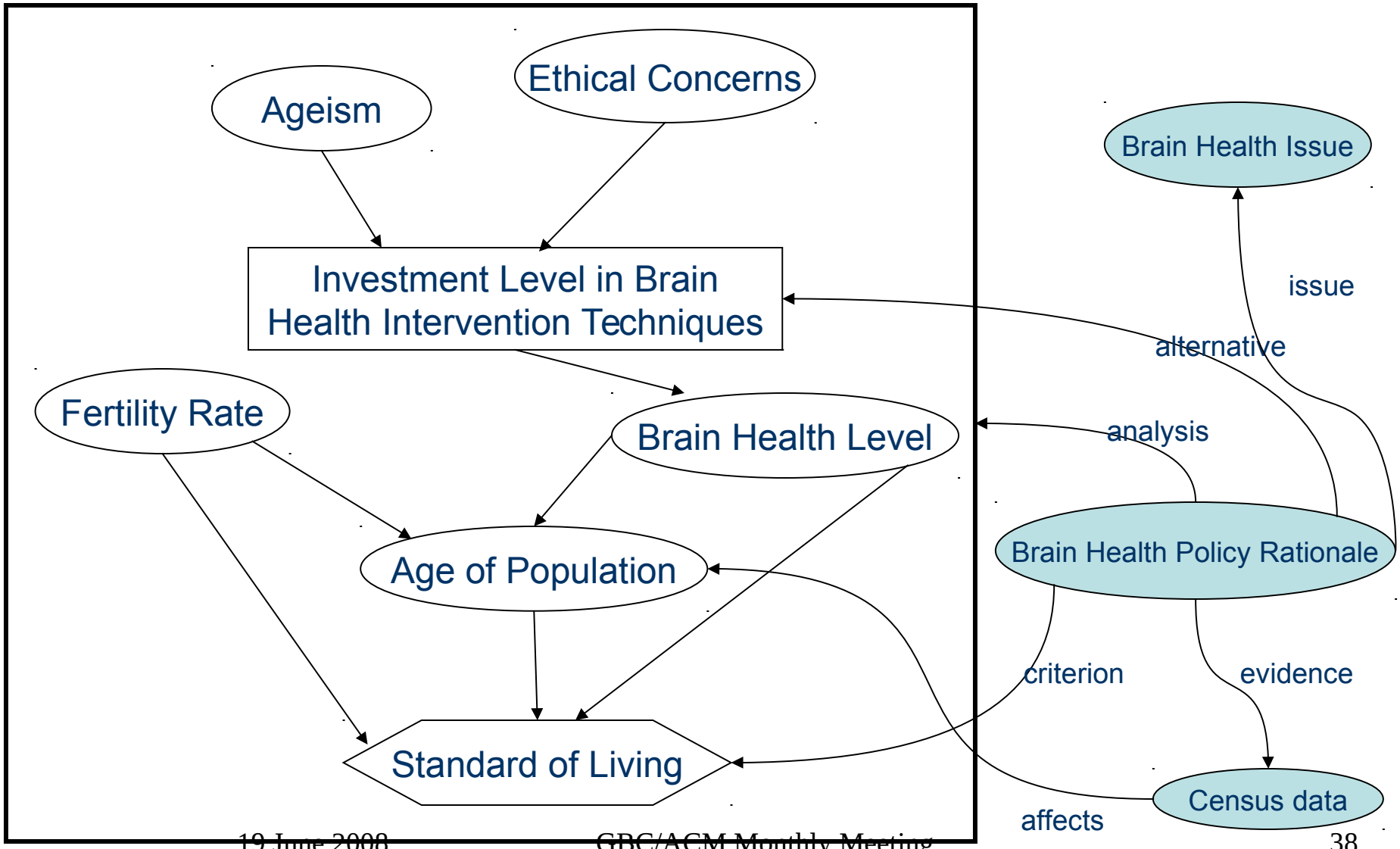
# Decision Analysis

- Important part of policy and development processes.
- Formal annotation of decisions and their analyses can have many benefits.
  - Integration with the process
  - Recognition of need to reconsider when circumstances evolve
  - Decisions can be delayed
  - Decisions can be reused for other situations
- An annotated decision is called a **rationale**

# Rationale Ontology



# Policy Decision Example

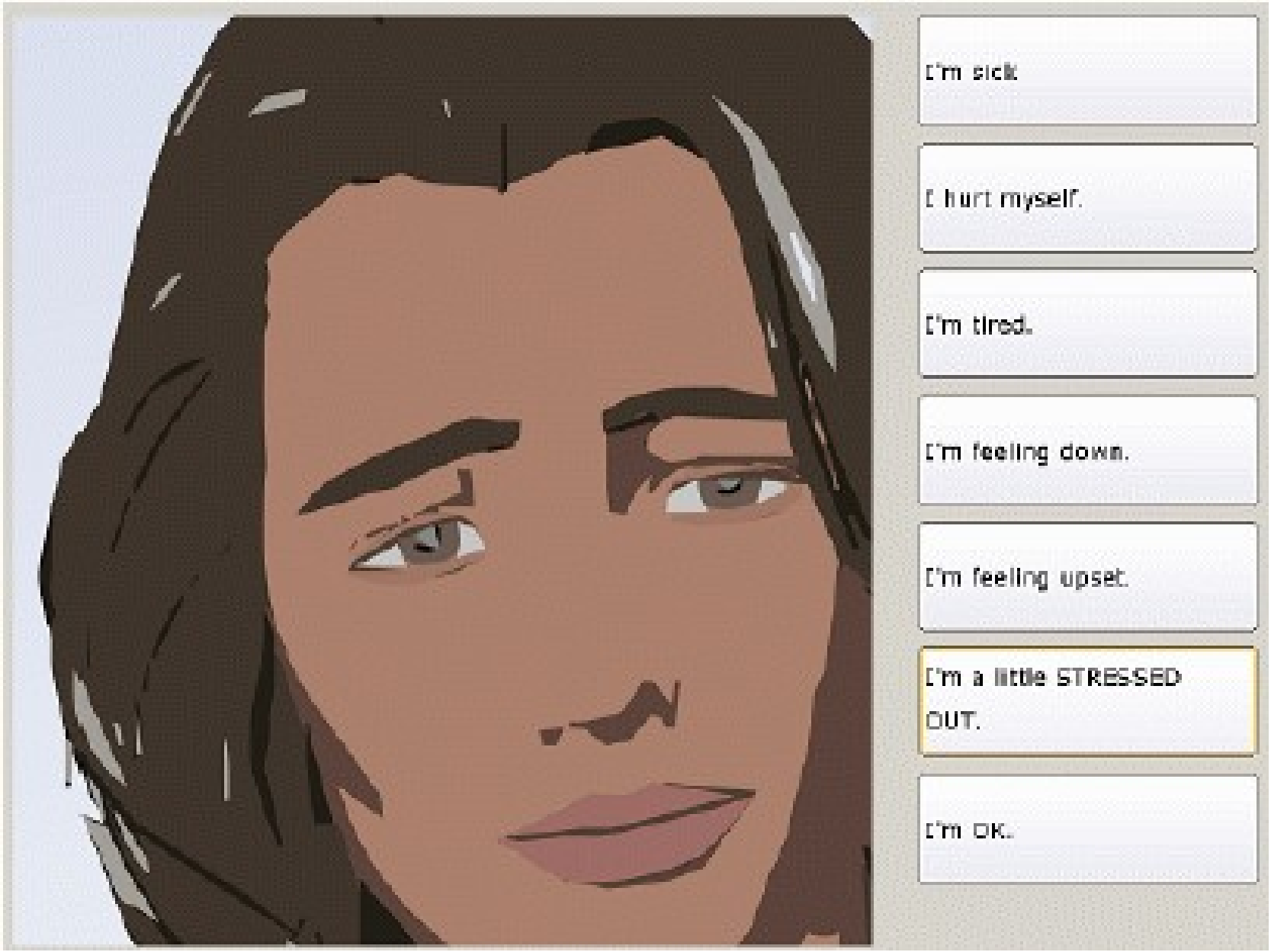


# References

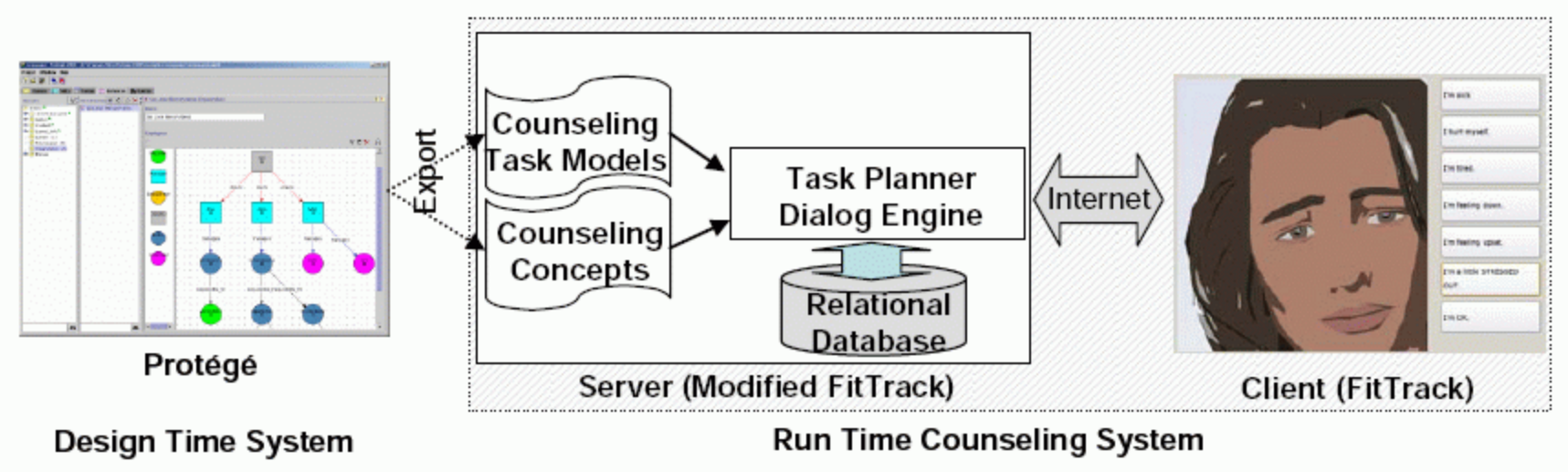

- M. Kokar, K. Baclawski and D. Brady. Uses of Ontologies for Cognitive Radios. In *Spectrum Efficiency and Cognitive Radio Technology*, Bruce, A., Fette (Ed). Newnes. (August, 2006)
- V. Duggar and K. Baclawski. Integration of Decision Analysis in Process Life-Cycle Models. In *International Workshop on Living with Uncertainties*. (November 5, 2007)

# Behavioral Health

- Medical ontologies have resulted in advances in standardization, information sharing and automation not previously possible in medicine
- In contrast, the development of ontologies for behavioral medicine is decades behind.
- Ontologies for behavioral health have the potential for important advances
  - Facilitating the growth of the discipline itself
  - More rapid development of automated systems for effecting health behavior change
  - Improving scalability, tailorability and adaptability


I'm sick.
I hurt myself.
I'm tired.
I'm feeling down.
I'm feeling upset.
I'm a little STRESSED OUT.
I'm OK.

# System Architecture



# Concepts in the ontologies

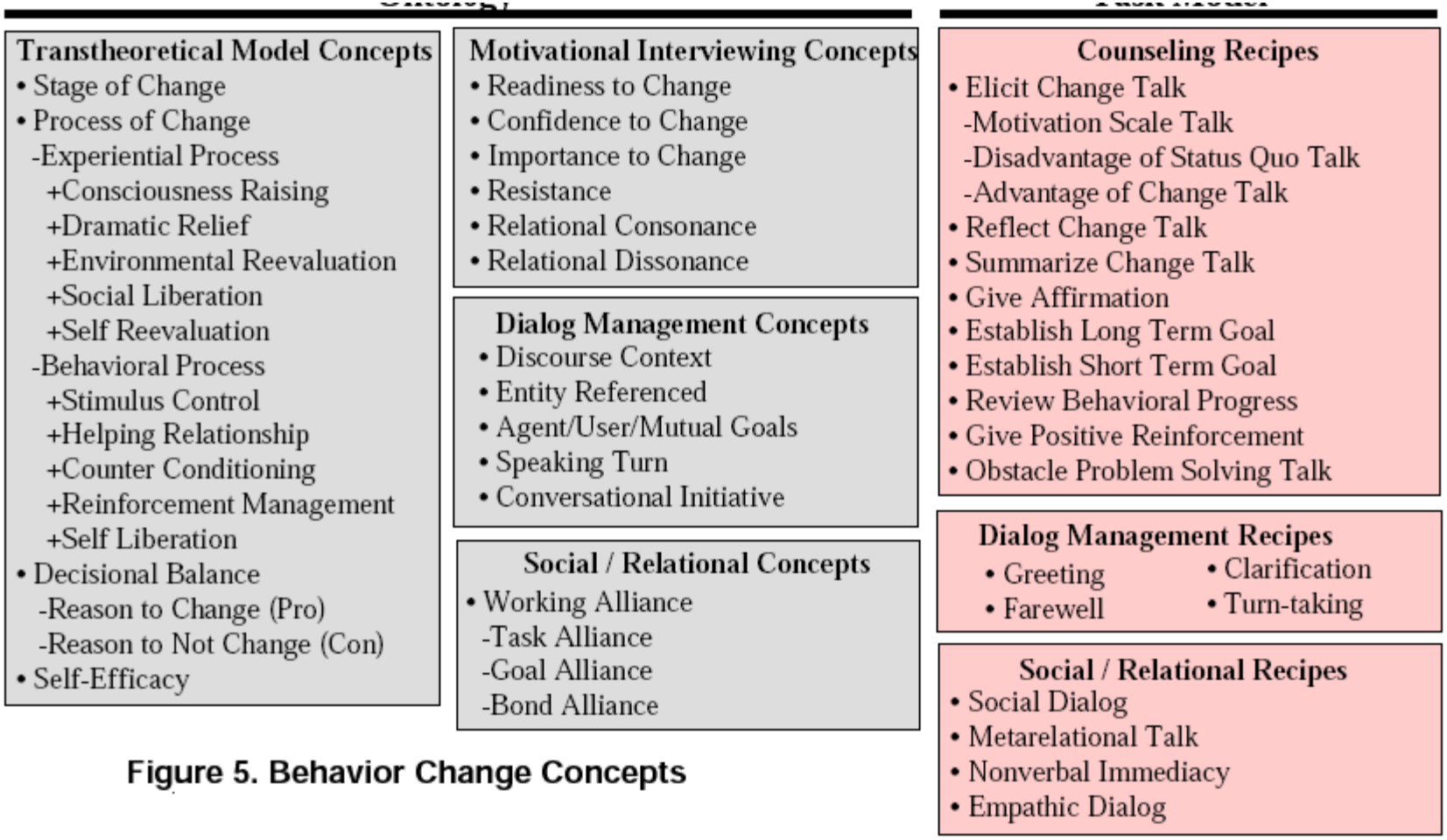
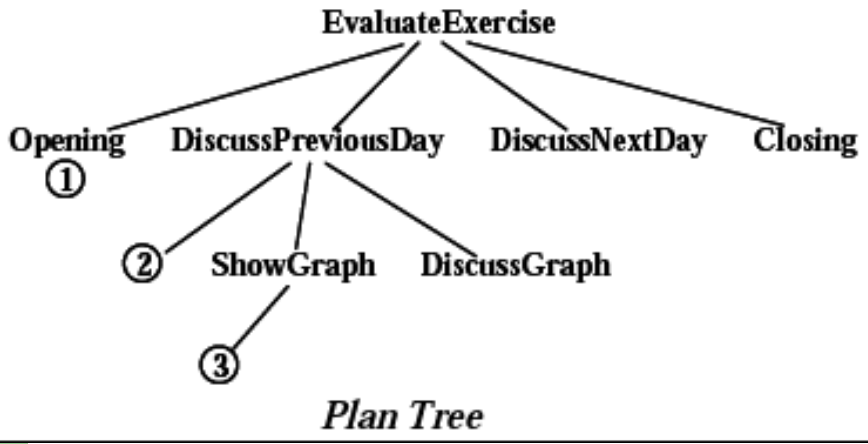



**Figure 5. Behavior Change Concepts**

# Conversational Planning




*Plan Tree*

*Interacting about exercise.*
① *Done opening interaction.*
*Discussing previous day.*
② *Done client identifying steps walked yesterday as 1000.*
Coach says "How many steps did you walk yesterday?"
Client says "1000."
*Coach showing graph for the week.*
③ Coach says "Here's how you've done this week."
*Next expecting coach to show graph for the week.*
*Expecting to discuss graph.*
*Expecting to discuss next day.*
*Expecting to close interaction.*

# Disease Knowledge Using Biological Taxonomy, and Environmental Ontologies

- Collaboration with Neil Sarkar of the Marine Biological Laboratory
- Biomedical knowledge relevant to the study of infectious diseases is currently in a variety of heterogeneous data sources
  - Citation databases
  - Health reports
  - Molecular databases
- Understanding infectious diseases requires
  - Environmental and geo-location
  - Biodiversity and biomedical resources

# Disease Knowledge Sources

- Research Literature Citation Indexes
  - Medline of the US National Library of Medicine
  - Agricola of the US National Agricultural Library
- Health Reports
  - Global Outbreak Alert and Response Network (GOARN) of the World Health Organization
  - Program for Monitoring Emerging Diseases (ProMED) of the International Society for Infectious Diseases

# Biodiversity Sources

- Biodiversity Heritage Library
- Global Biodiversity Information Facility (GBIF) hosted by the University of Copenhagen
- Encyclopedia of Life
- Many others…

# Some Background Ontologies

- NCBI Taxonomy of the US National Center for Biotechnology Information
  - Alpha taxonomy associated with molecular data (GenBank)
- Environmental ontology (EnvO)
  - Emerging Open Biomedical Ontology (OBO) of biological habitats
- Geo-location instance hierarchy (Gaz)
  - Emerging OBO instance hierarchy of geo-locations

# Example of integration of disease knowledge, genetic information, biodiversity information and geographical information

Geographic distribution of hantavirus disease outbreaks (boxes) and genetic samples (helices)

Geographic distribution of biodiversity information for the two most common US deer mouse species

# Recent Developments

- RDF storage provided by database vendors
  - Oracle has both a product and an active Database Semantic Technologies Group
  - Many RDF stores are layered on a general purpose RDBMS: Jena, Sesame, RDQL, …
- Non-relational RDF storage products
  - Siderean, Tucana, OWLIM, Allegro Graph, …

# Open Ontology Repository (OOR)

- Recent initiative of the Ontolog Forum
- The purpose of the initiative is to promote the global use and sharing of ontologies by:
  - 1. establishing a hosted registry-repository;
  - 2. enabling and facilitating open, federated, collaborative ontology repositories;
  - 3. establishing best practices for expressing interoperable ontology and taxonomy work in registry-repositories.

# Semantic Technology Conference

- Drew more than 1,000 attendees from 35 countries.
- Included many sessions on experiences and best practices.
- http://www.semantic-conference.com/

# Caveats

- The examples shown in this presentation were for educational purposes only. They are not complete, and there are technical details that were omitted.
- While RDF can be written using XML, there are other formats such as N3 and N-triples that are much simpler.