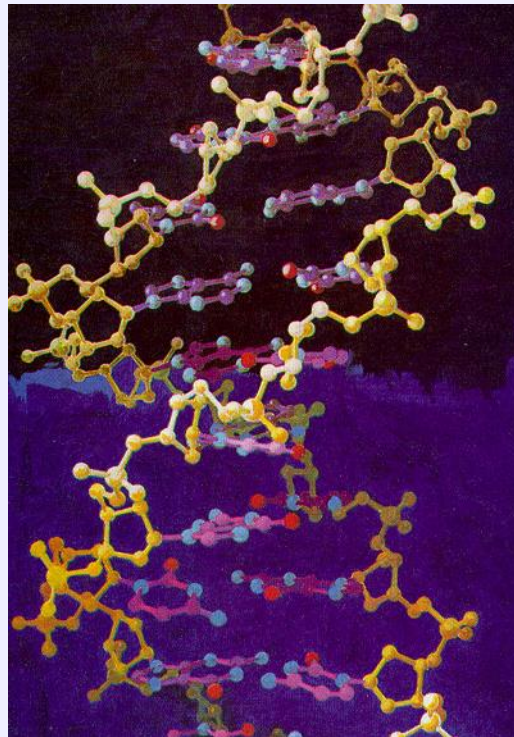# The Use of Active Structure in Biotechnology
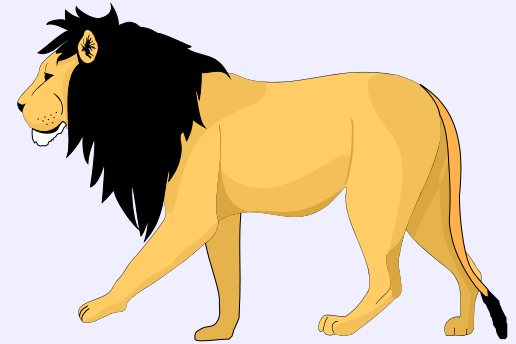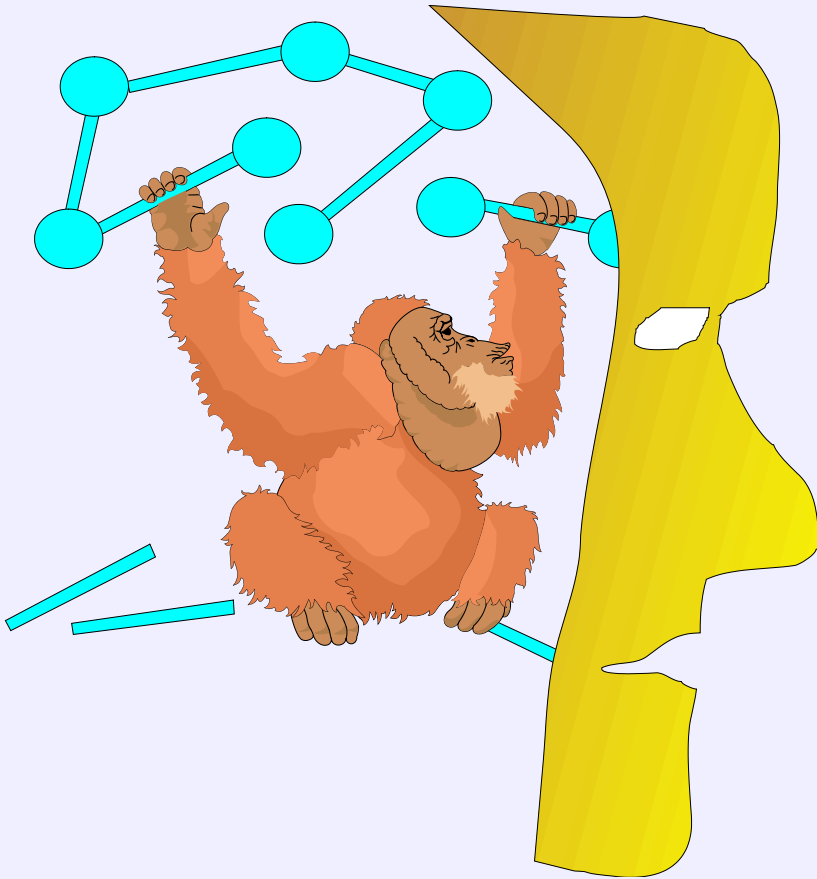

Orion Technology

We won't bother to tell you how hard, how complex, how dynamic, is knowledge in the field of biotechnology.

We will assume you are interested in a means of handling this dynamic complexity that doesn't rely on the heroic assumptions implicit in statistical or other static or pipelined methods

We use a cognitive approach - all of the knowledge is alive all the time

Active Structure

As humans, we build an internal model of the world, so we can predict future behavior - we make the model out of structure so we can combine it with other structure - we pass complex messages through the structure

Orion technology does all these things too...

Capturing Knowledge

It may seem obvious, but if we are to use knowledge in our machines, we have to capture it with all its possible inferences intact

That is why it needs to be captured in an undirected form, and why it has to be captured in a context

The IF...THEN... rules of programming or Expert Systems are about throwing away all the stuff we didn't think we would need before we knew exactly what the problem was
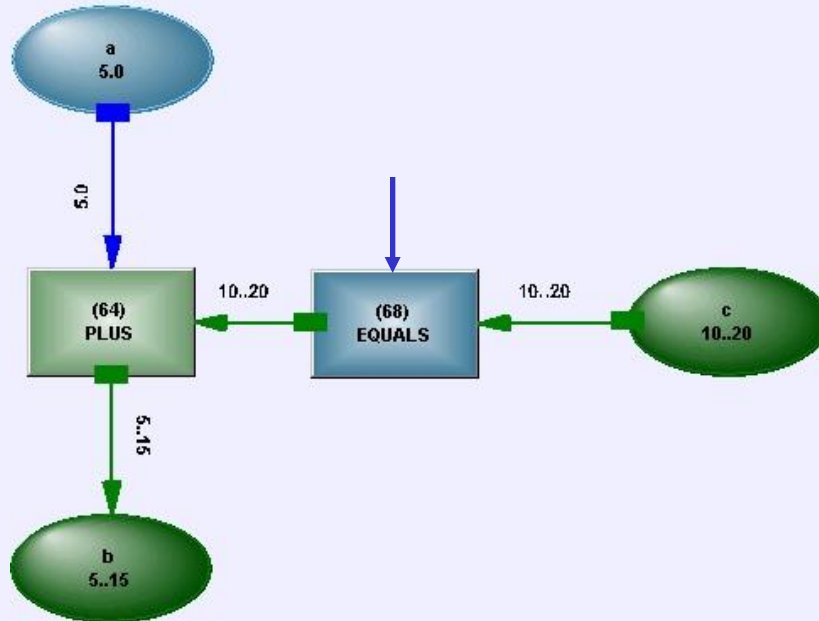
Active Structure
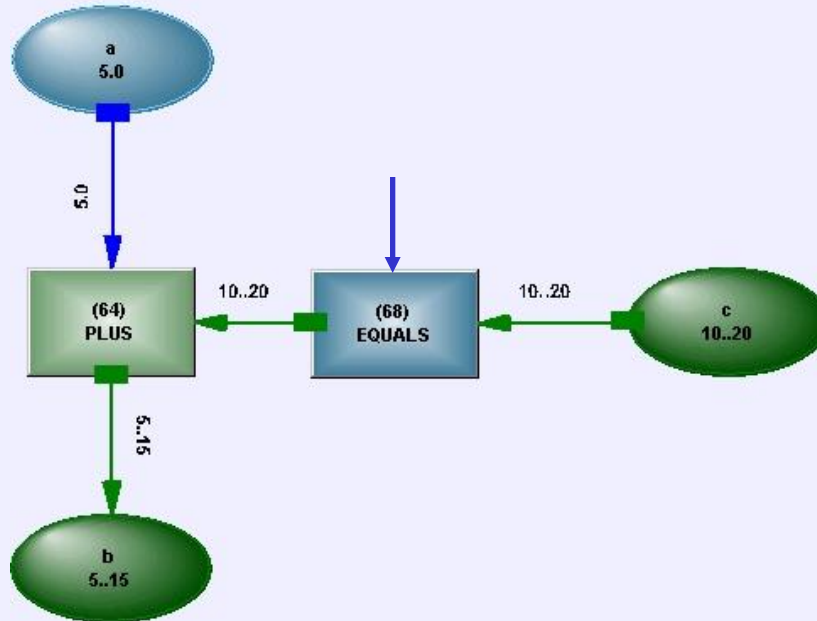
- Active
- Visible & Connectable
- Undirected
- Dynamically Extensible
- Self-Modifying
- Structural Backtrack
- Complex Messages

What has A + B = C got to do with biotechnology?
It can explain the paradigm simply
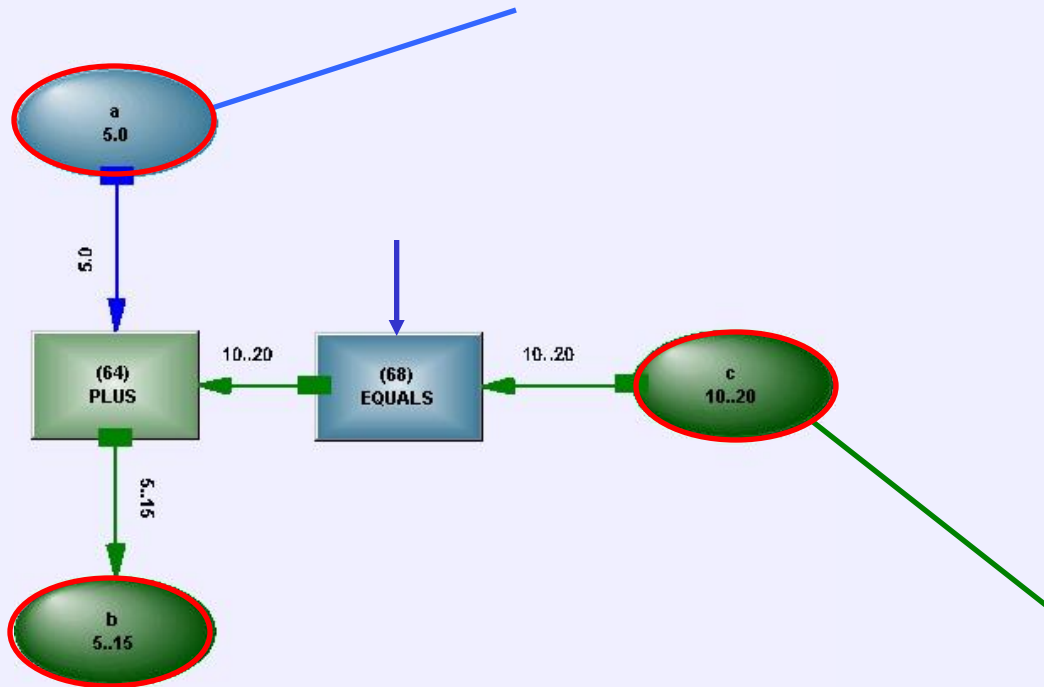
Active

The messages flow in the structure - there isn't something else looking at the structure and trying to figure out what happens - there can't be, because the information flow reverses and the structure keeps changing

Visible & Connectable

The variables in the structure can be seen and can be connected to, with information immediately flowing in or out of the connection

Undirected
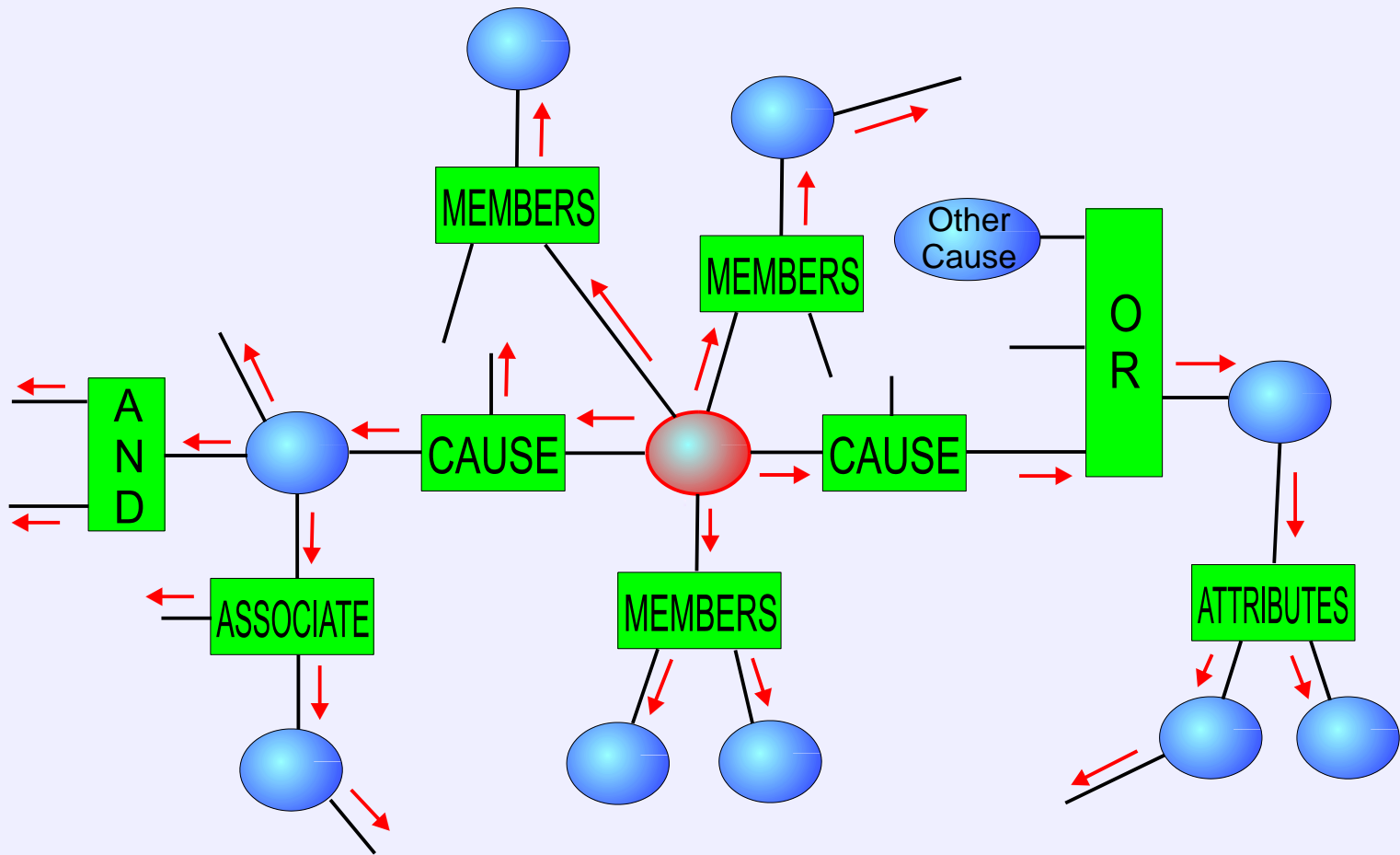
Information can flow in any direction in the structure - we can start with genes and get to diseases, or vice versa - it is not necessary to put in direction initially

Dynamically Extensible

In automatically reading text and extending the structure, the structure is extending itself dynamically

There is no compilation - the structure has been continually extended from a root and all the information it used to extend itself is still there

Self - Modifying

As states and values in the structure change, so the structure has to change

A noun phrase is detected, a new structure has to be built to represent it

An intransitive clause is detected, so the connections to the relation the verb represents need to be rearranged

Not only is the structure dynamically extensible, it is continually modifying its own structure to reflect its current state

Structural Backtrack

Sometimes it is necessary to build something and observe the result, then tear it down and do something else

This is common when trying to work out ambiguities in text, but also happens when trying to find a link between causes and effects
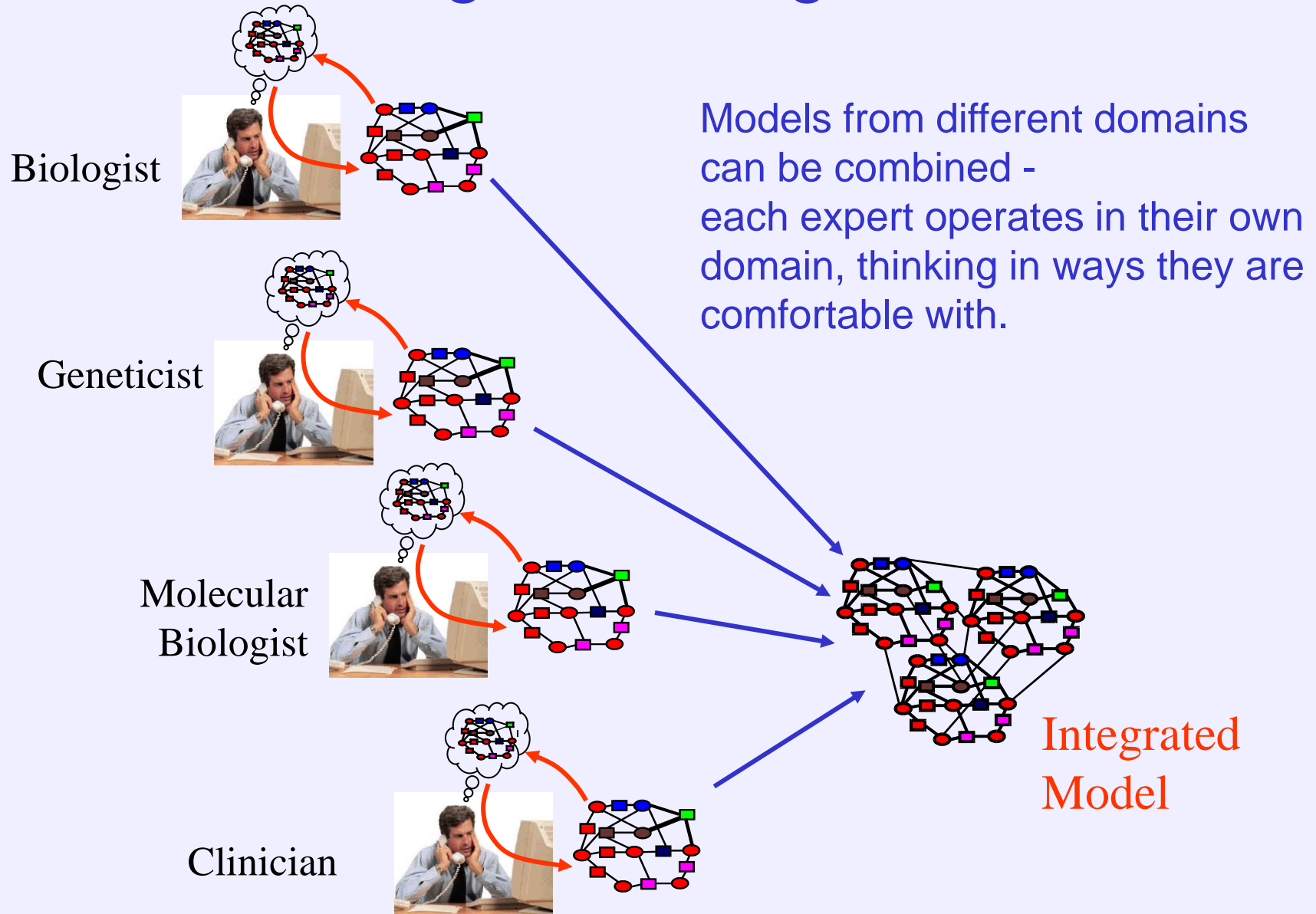
This building and destroying of structure is done automatically

Active Structure

- Active
- Visible & Connectable
- Undirected
- Dynamically Extensible
- Self-Modifying
- Structural Backtrack
- Complex Messages

Active Structure allows all these things to happen together - losing any one of them would greatly degrade its performance

Combining Knowledge Domains

Biologist

Geneticist

Molecular Biologist

Clinician

Models from different domains can be combined -
each expert operates in their own domain, thinking in ways they are comfortable with.

Integrated Model

Turning Scientific Text
Into
Active Knowledge

Scientific Text

Scientific text has many characteristics that distinguish it from other kinds of text:

- The use of words defined within the document
- The use of prepositional chains to establish context
- The manipulation of existence (the presence of...)
- The need to find complex objects whose description is distributed through the text
- The overwhelming complexity when many documents need to be read together

The Reading Process

A document is read, paragraph by paragraph, sentence by sentence, word by word.

As the words are read, they are turned into objects that can be manipulated - objects that have the properties both of words and of the objects they represent - a ligand, a gene

The word objects are assembled through grammar into larger objects - receptor or gene structure

And into larger structures, using the relations between the objects provided by nouns and verbs

Examples

ABSTRACT: The synergistic effect of serine and CheW binding to Tsr is attributed to distinct influences on receptor structure; changes in the conformation of the Tsr dimer induced by serine binding improve methylation efficiency, and CheW binding changes the arrangement among Tsr dimers, which increases access to methylation sites.

ABCA1: ATP-binding cassette, sub-family A (ABC1), member 1

LocusID: 19

Overview

The membrane-associated protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intracellular membranes. ABC genes are divided into seven distinct subfamilies (ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White). This protein is a member of the ABC1 subfamily. With cholesterol as its substrate, this protein functions as a cholesteral efflux pump in the cellular lipid removal pathway. Mutations in this gene have been associated with Tangier's disease and familial high-density lipoprotein deficiency.
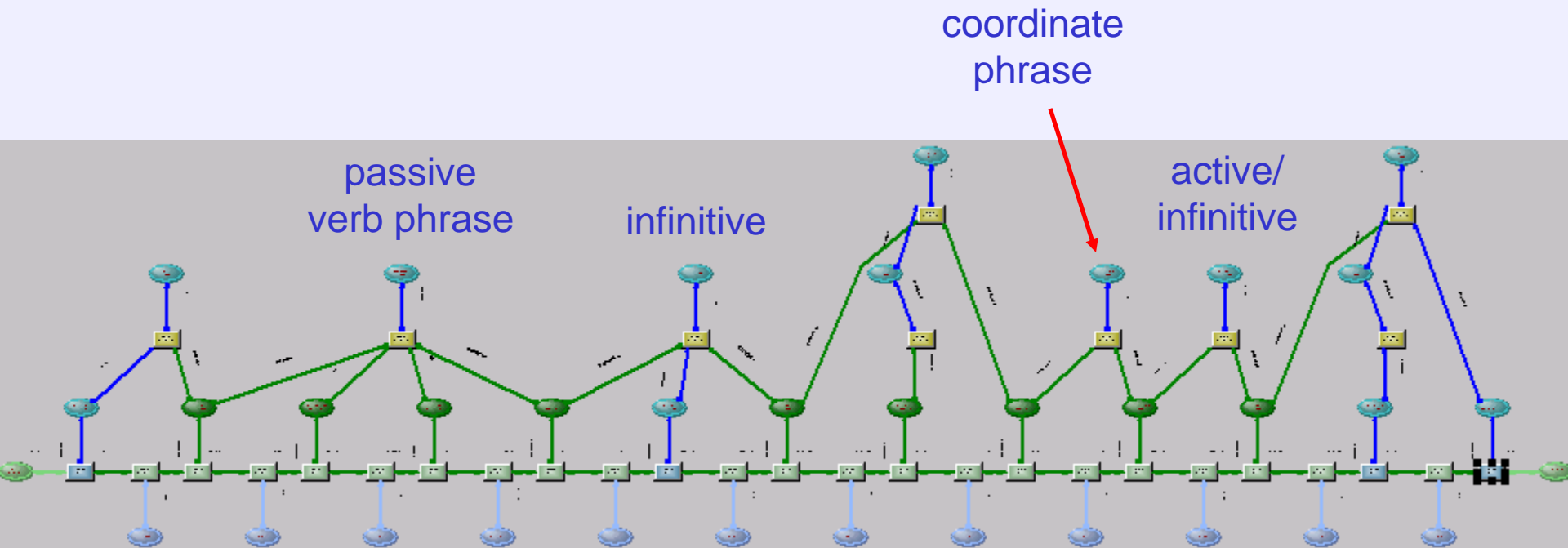
Textual Forms

There are two forms here

- One form describes a new structure in qualitative terms
- The other form introduces relatively little new
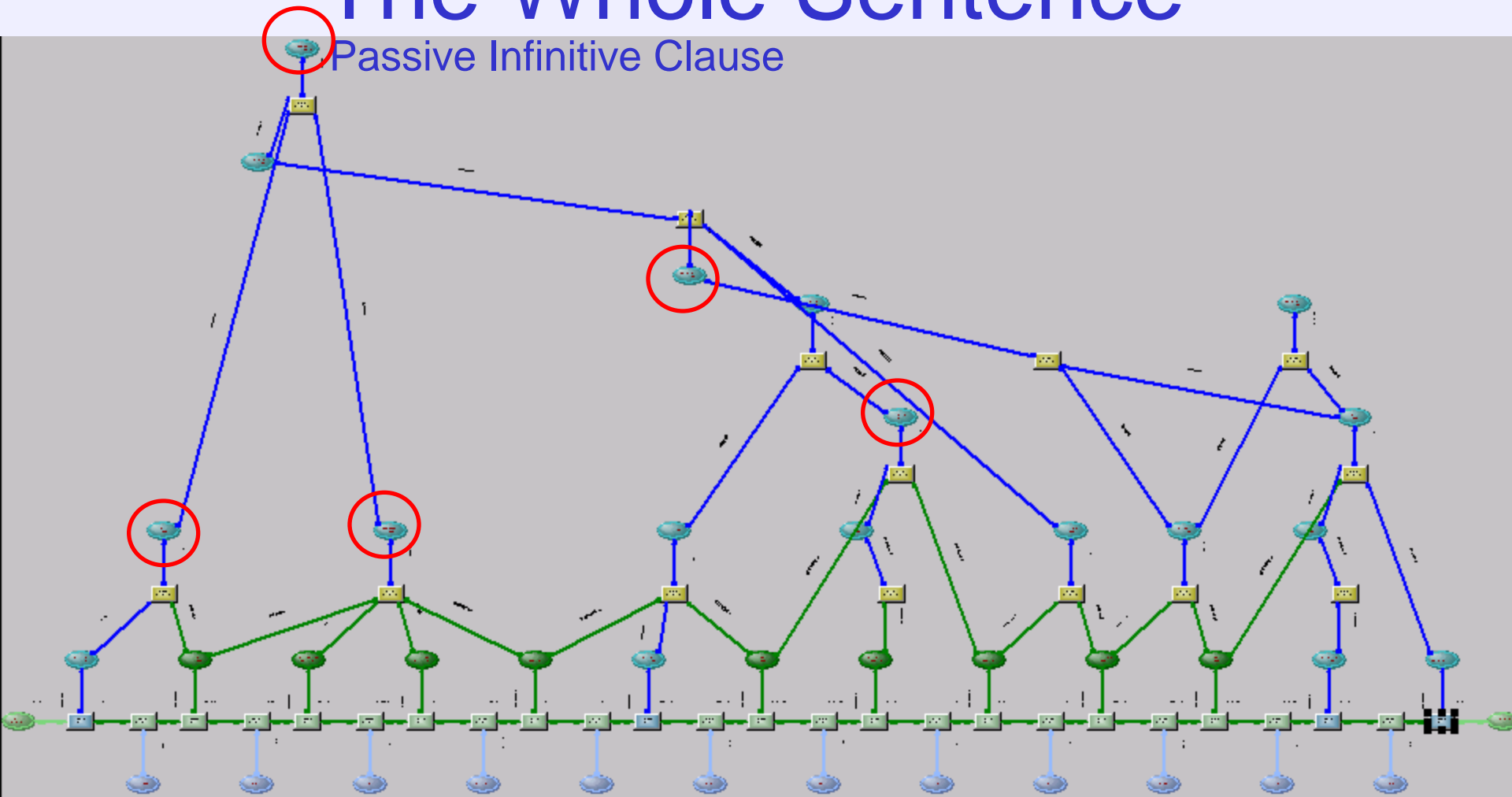   structure among a huge array of existing objects

Examples of each follow

Parse Chain

The words are already linked to the objects they represent - a ligand, a gene, as well as to noun or verb
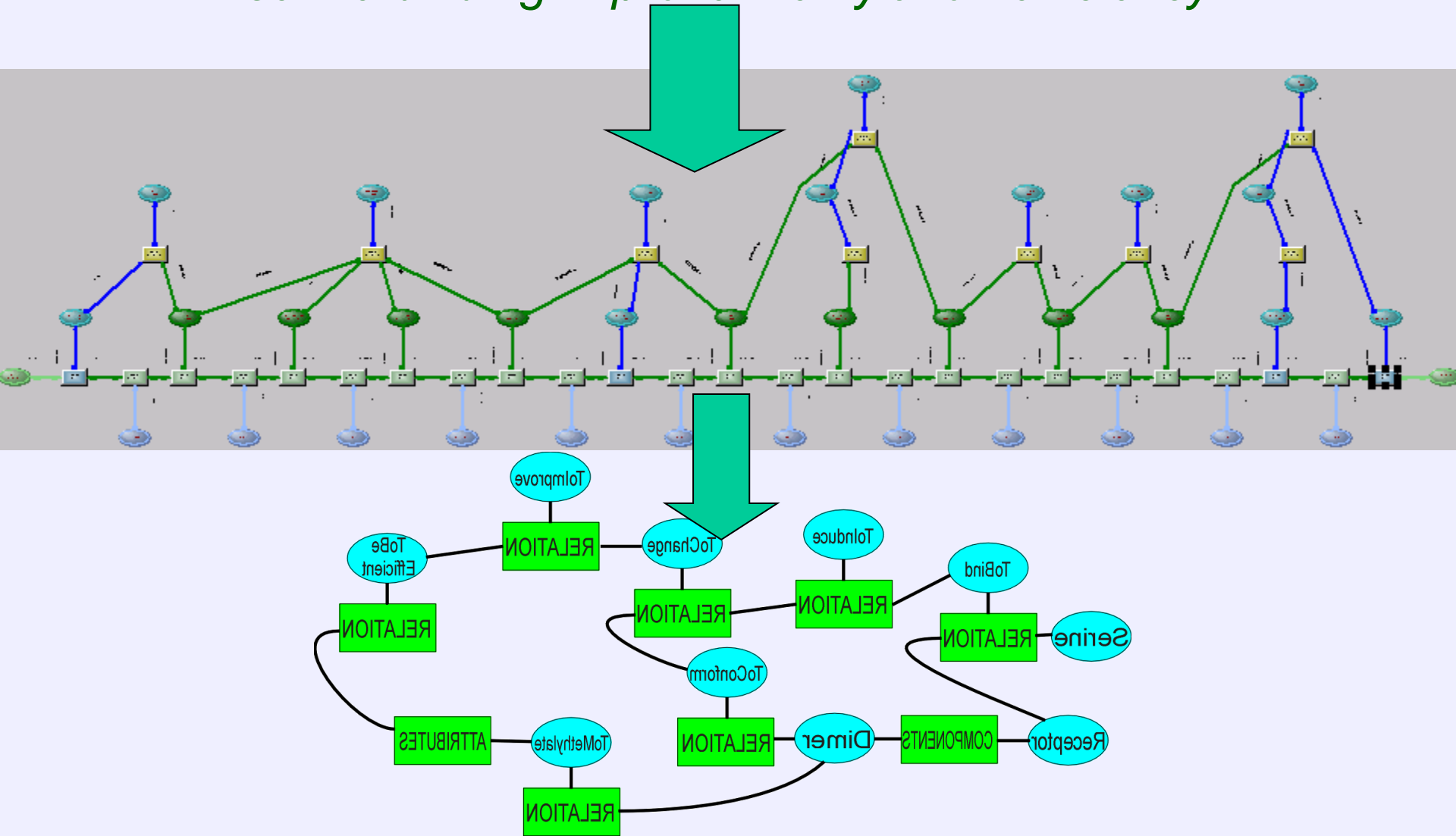
The Whole Sentence
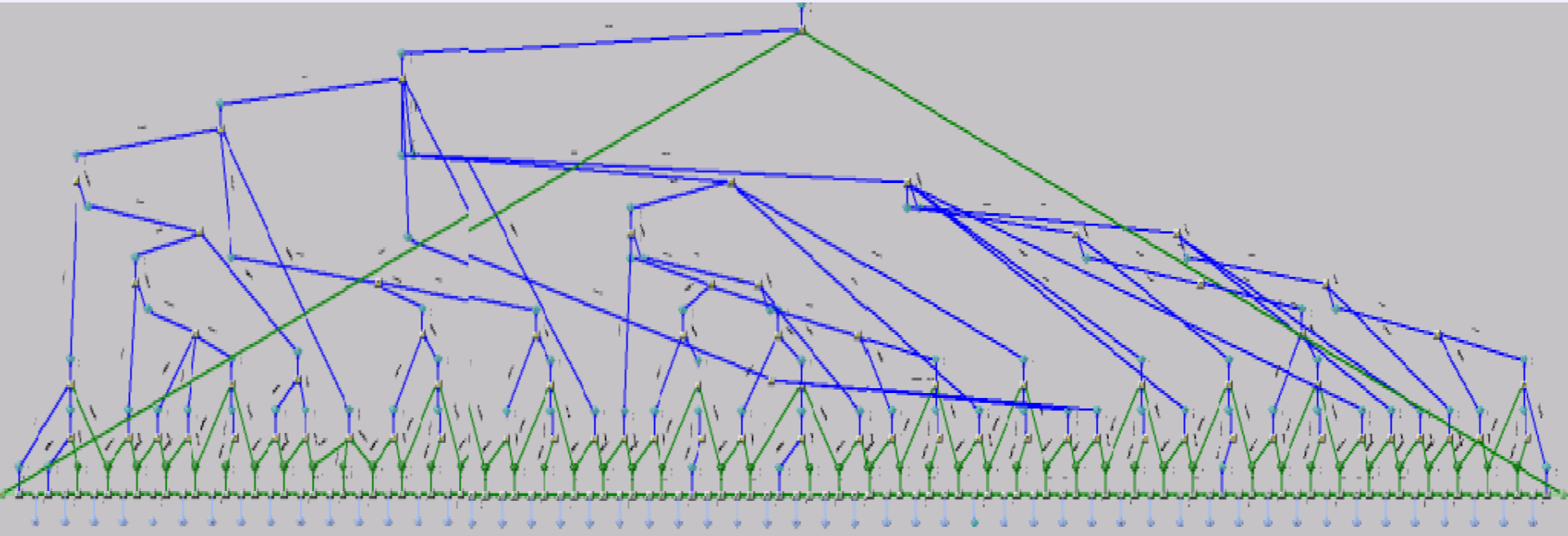
Passive Infinitive Clause

As the grammatical structure builds, it either builds the object/relation structure in parallel, or identifies the corresponding preexisting structure

Transformation

changes in the conformation of the Tsr dimer induced by serine binding improve methylation efficiency

Scientific Sentences Are Complex

The synergistic effect of serine and CheW binding to Tsr is attributed to distinct influences on receptor structure; changes in the conformation of the Tsr dimer induced by serine binding improve methylation efficiency, and CheW binding changes the arrangement among Tsr dimers, which increases access to methylation sites.

Grammar Is Not Enough

Grammar alone would turn meaningful scientific text into sludge - the participial phrase "induced by..." has to be anchored on the right object, the relative pronoun "which" has to be anchored on the relation

The reading process demands that domain knowledge be available at every turn - knowledge that is held in object hierarchies and relations, and which is seamlessly intermingled with grammatical knowledge during the parsing

Preexisting Structure
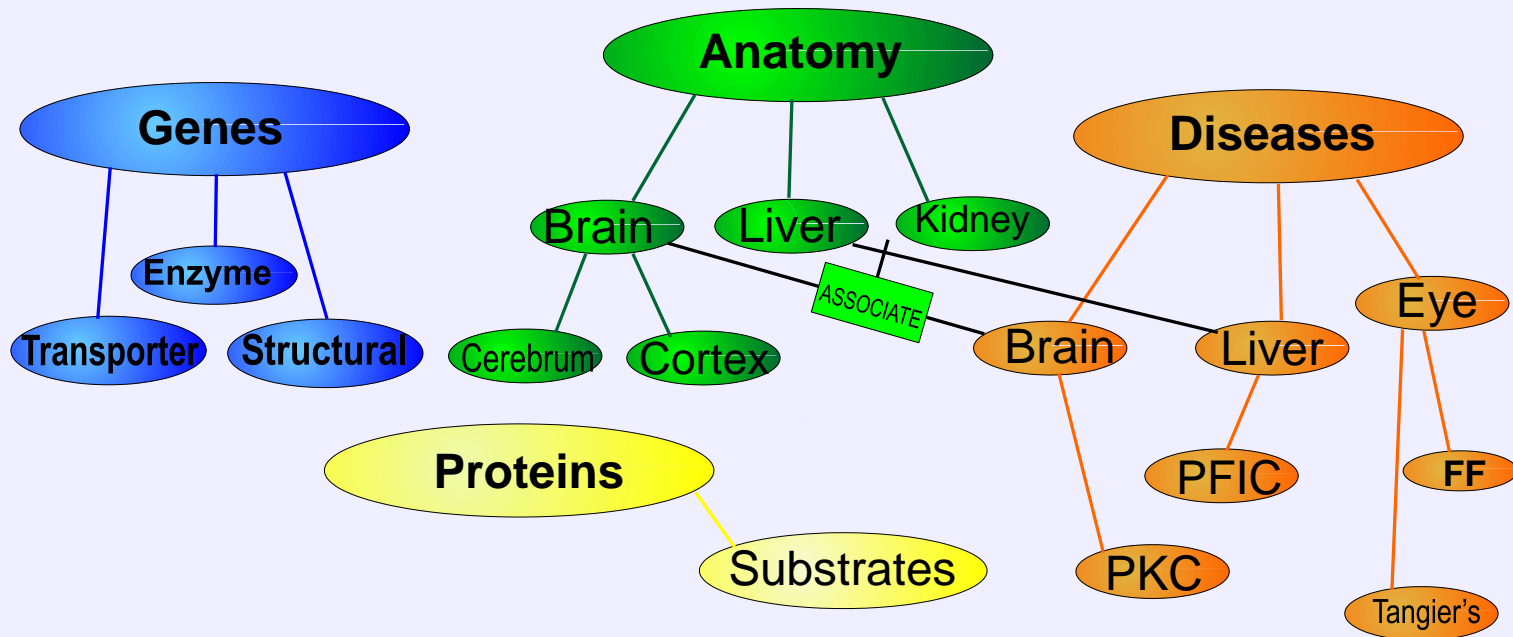
A beginning structure is built - gene families, anatomy, cell structure, diseases

Gene Summaries

ABCA1: **ATP-binding cassette, sub-family A (ABC1), member 1**

LocusID: 19

Overview

The membrane-associated protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intracellular membranes. ABC genes are divided into seven distinct subfamilies (ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White). This protein is a member of the ABC1 subfamily. With cholesterol as its substrate, this protein functions as a cholesteral efflux pump in the cellular lipid removal pathway. Mutations in this gene have been associated with Tangier's disease and familial high-density lipoprotein deficiency.

Family ABC (transporter across membranes)
Subfamily ABC1 (members ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White)
Gene ABCA1
Protein NP_005493
Substrate Cholesterol
Function cholesterol efflux pump associated lipid removal pathway
Mutation causes Tangier's disease, familial high-density lipoprotein deficiency.

Chromosome: 9 mv **Cytogenetic:** 9q31.1 RefSeq

All Together Now

ABCA1: ATP-binding cassette, sub-family A (ABC1), member 1

LocusID: 19

Overview

The membrane-associated protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intracellular membranes. ABC genes are divided into seven distinct subfamilies (ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White). This protein is a member of the ABC1 subfamily. With cholesterol as its substrate, this protein functions as a cholesteral efflux pump in the cellular lipid removal pathway. Mutations in this gene have been associated with Tangier's disease and familial high-density lipoprotein deficiency.

Family ABC (transporter across membranes)
Subfamily ABC1 (members ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White)
Gene ABCA1
Protein NP_005493
Substrate Cholesterol
Function cholesterol efflux pump associated lipid removal pathway
Mutation causes Tangier's disease, familial high-density lipoprotein deficiency.

Chromosome: 9 mv **Cytogenetic:** 9q31.1 RefSeq

We are not just picking the eyes out of the abstract - all of it is being represented against what we already know- gene structure, substrate behaviour, anatomy, drugs.
All the relations in the text are built, including known or potential causality as well as the family structure

The resulting structure seamlessly combines objects and relations, and is active

Linking

ABCA1 ATP-binding cassette, sub-family A (ABC1), member 1

LocusID: 19

Overview

The membrane-associated protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intracellular membranes. ABC genes are divided into seven distinct subfamilies (ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White). This protein is a member of the ABC1 subfamily. With cholesterol as its substrate, this protein functions as a cholesteral efflux pump in the cellular lipid removal pathway. Mutations in this gene have been associated with Tangier's disease and familial high-density lipoprotein deficiency.

Family ABC (transporter across membranes)
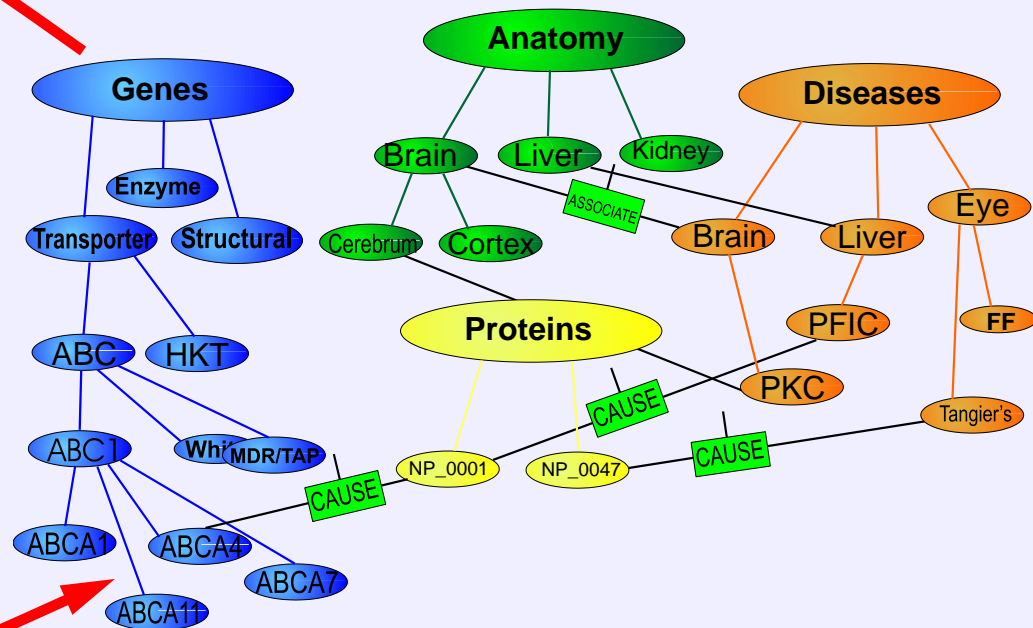**Subfamily ABC1 (members ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, White)**
Gene ABCA1
Protein NP_005493
Substrate Cholesterol
Function cholesterol efflux pump associated lipid removal pathway
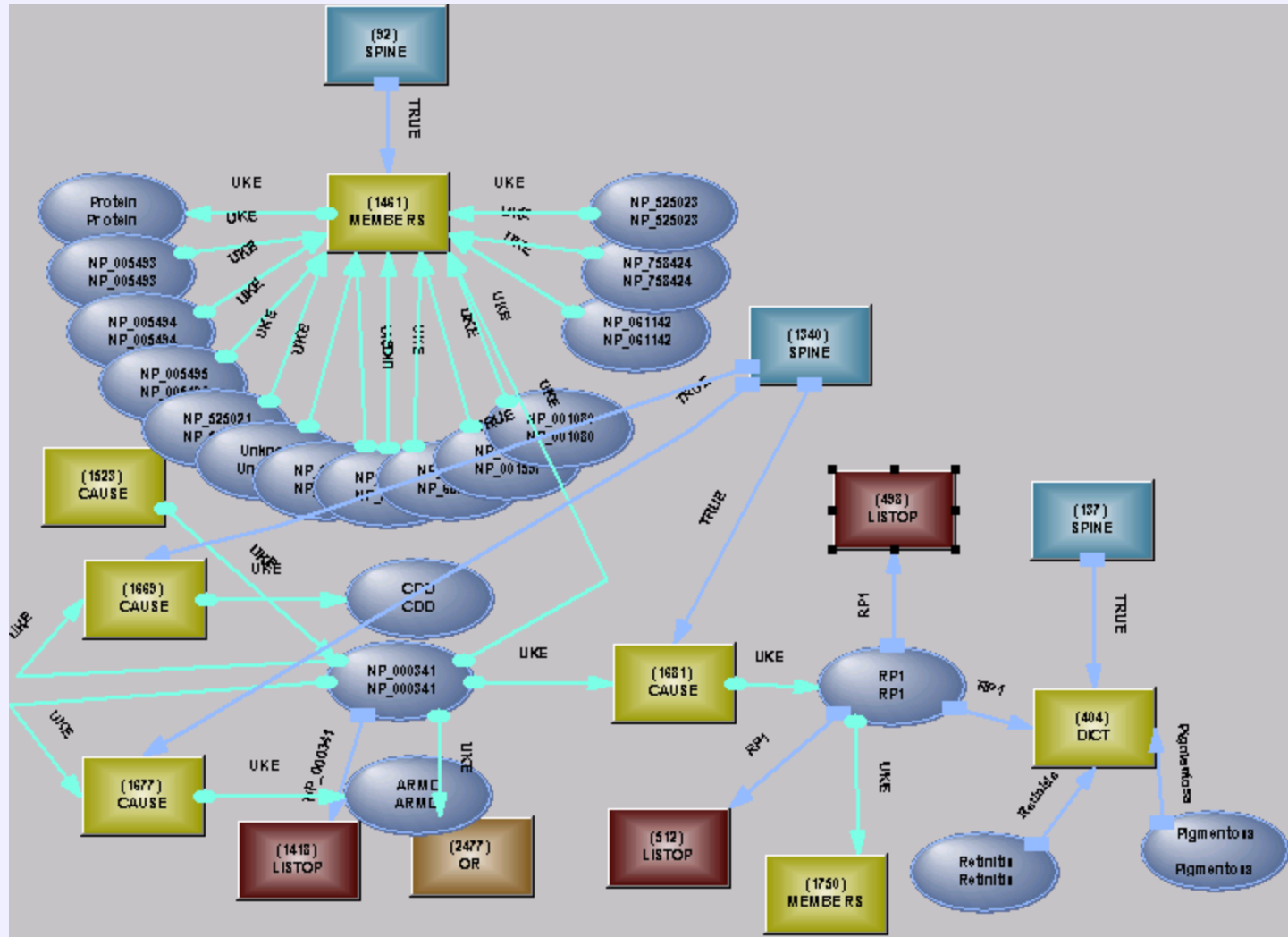Mutation causes Tangier's disease, familial high-density lipoprotein deficiency.

Chromosome: 9 mv **Cytogenetic:** 9q31.1 RefSeq

The structure is used to understand the text -
then the text is used to extend the structure

The Real Structure

Here is a tiny fragment of the real structure, illustrating the objects, relations, links and states which make it up.

Everything It Had

This seemingly simple operation - from structure to text and back - used everything the structure could do

- Active - the values flowing within the structure changed the structure
- Visible - the objects were found and connected to
- Undirected - it didn't matter what was found first
- Dynamically extensible - new objects, new relations, new structure
- Self-modifying - the structure extended itself
- Structural backtrack - ambiguities were resolved
- Complex messages - objects, causation, existence, gene code

Connected and Complete

The structure now represents all the knowledge held in the abstracts or summaries and all the foreknowledge needed to understand them, because every aspect of the text is represented by operations in the structure, operations that change the states in the structure

All of the documents are now connected together, in a way that text alone doesn't do

Invariant?

The structure that is built is not invariant - it uses the words described in the text except where it already has synonyms or existing structure

- you say induced, he says cause
- she says conformation, you say arrangement

But it can use synonyms, senses, inferential chaining (it knows how objects and relations connect together and can skip over gaps in the chain) and defined terms in the document to make the structure seem invariant for searching

Two Uses

The structure that results can be activated - things can be made to happen within it - to change its states or its connections.
Something can begin to exist, or disappear, or cause something else to happen - it can even add 2 + 2
This covers causation, simulation

The structure can also be passively searched - by creating a query in the same way the structure was read, using natural language, so a complex structure built from the query is used to probe for a match in a larger structure
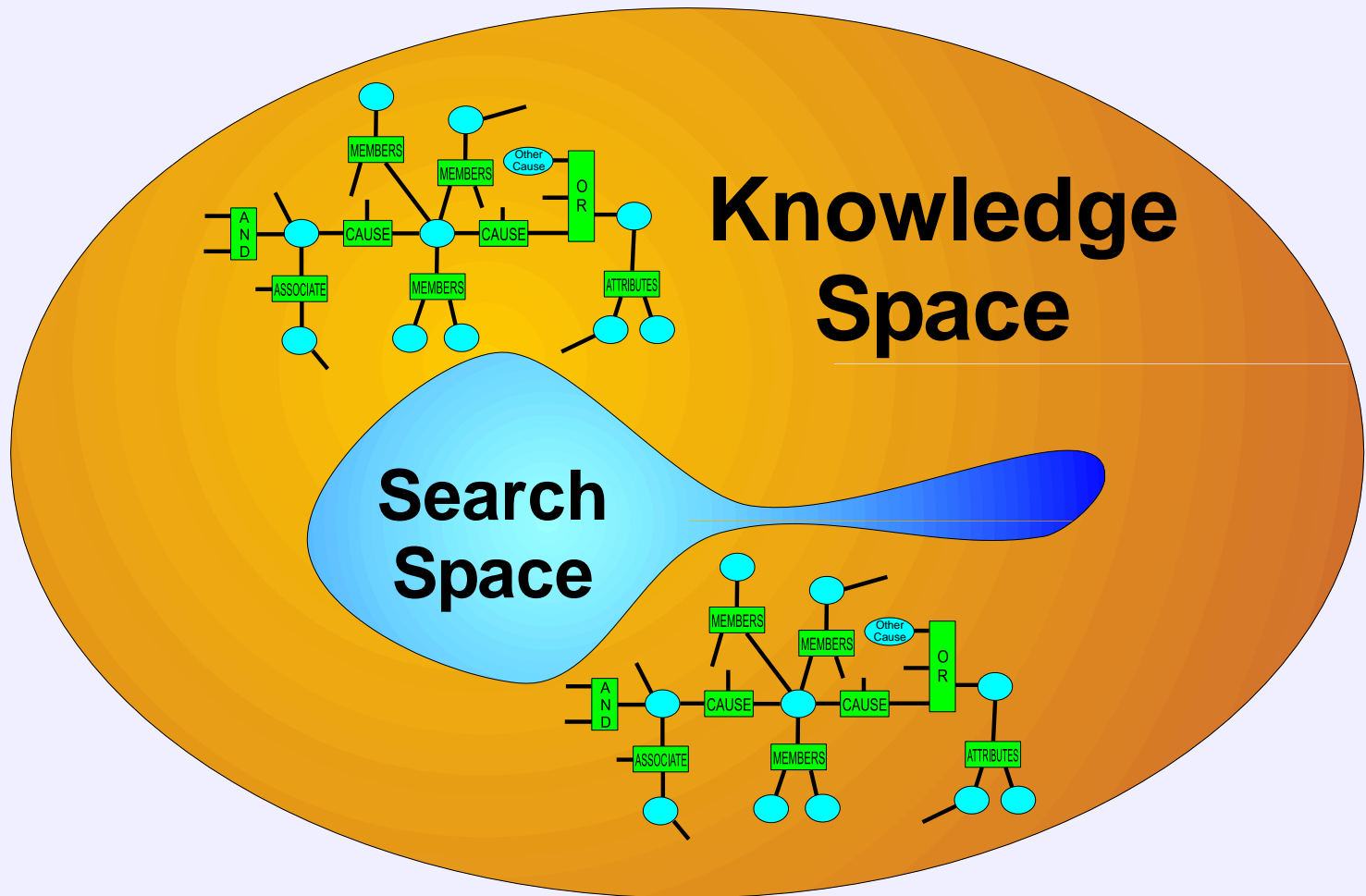
Active Searching

Sometimes passive probing isn't enough - the structure has to be activated to find something because otherwise the states are indeterminate, or effects are synergistic

Searching by Activation
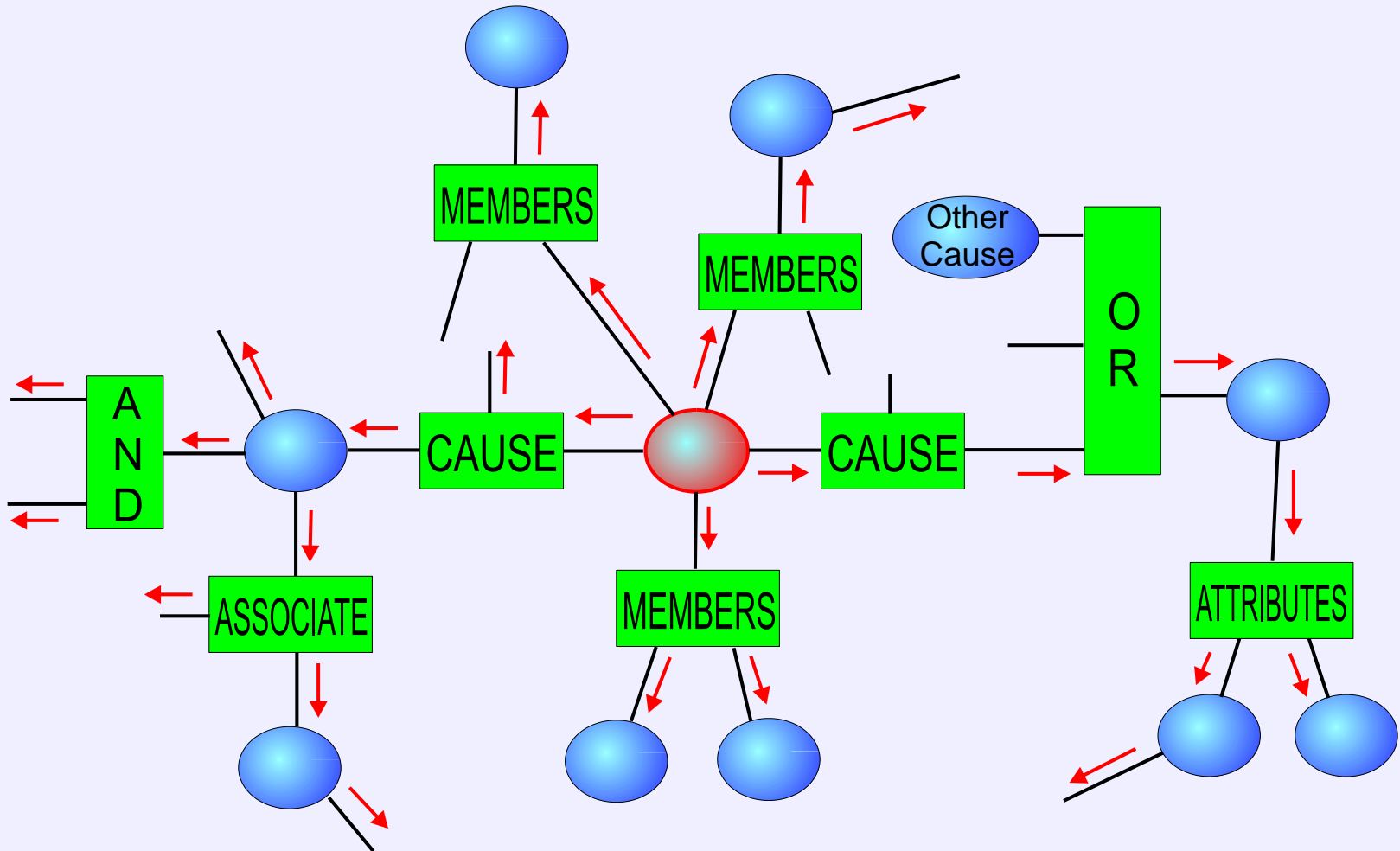
An existence pulse travels through the structure in all directions, bouncing off things it should ignore

Moldable Search Space

When there is an overwhelming amount of information, you need a powerful and flexible way of deciding what to ignore - held in the same structure

Other Methods

Key word search
  Misses too much, or needs lots of person time to search through too many hits
Magic search algorithms
  They know nothing about subject, throw up too much garbage
Templates or ontologies
  Knowledge in area keeps changing its structure as new linkages discovered - fixed structure too limiting - no seamless combination of objects and relations
Build structure by hand
  Knowledge in area too vast and too dynamic