Project Halo:
Making Sense of Questions in a Knowledge-Rich Environment

Peter Clark

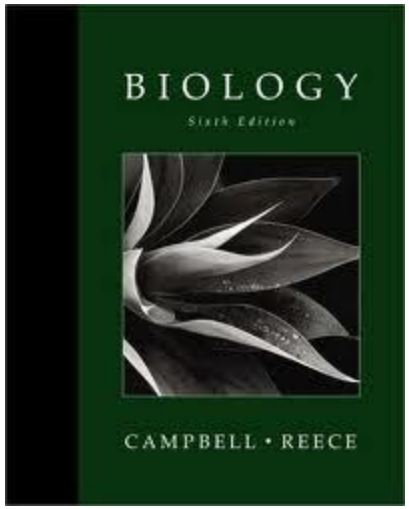
Vulcan Inc.

PROJECT HALO

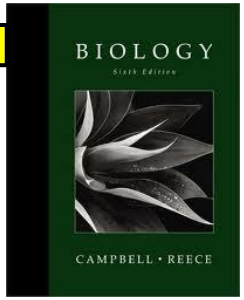
Project Halo

an approximation of
part of

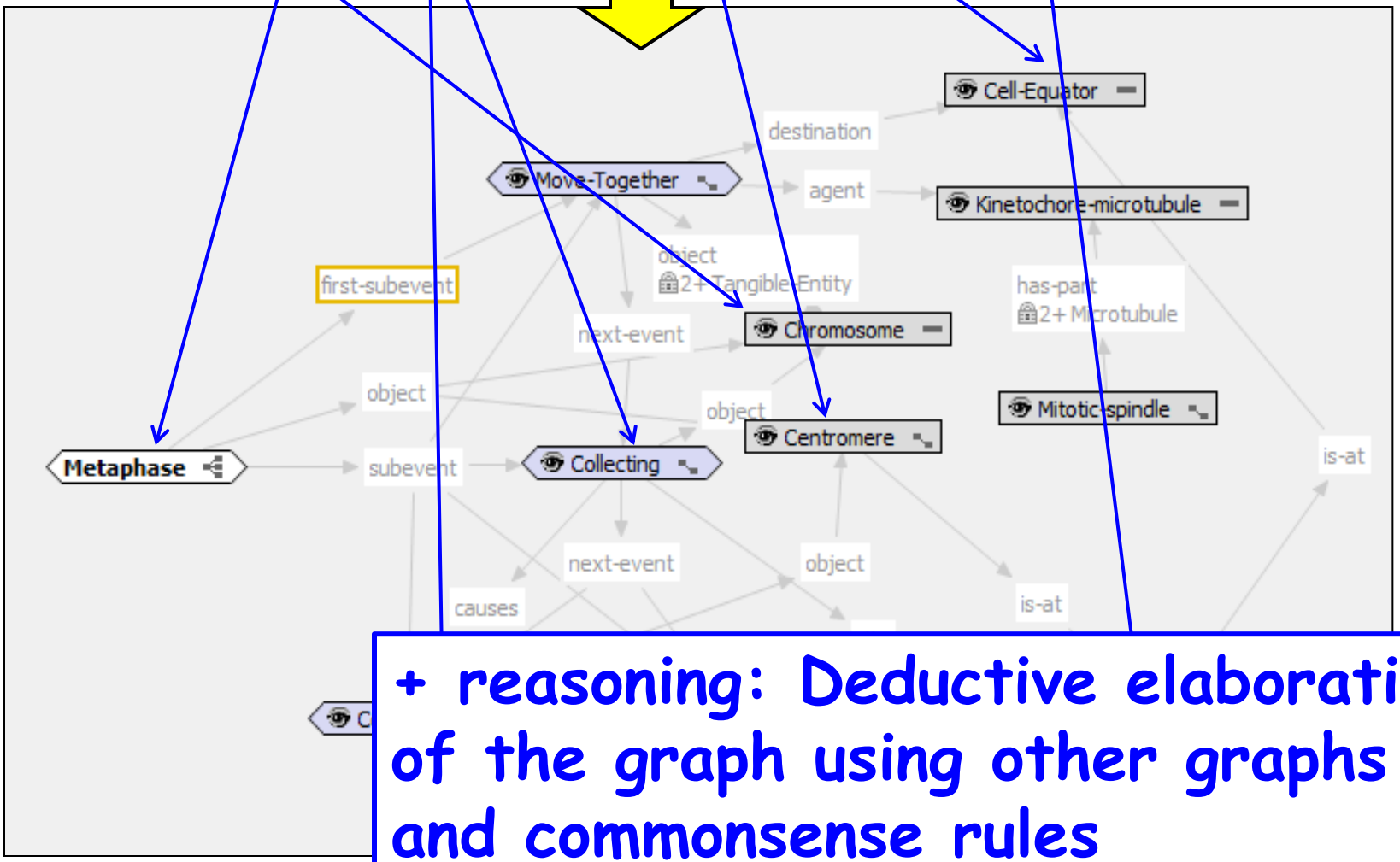
- Formally encoding a biology textbook as a KB
  - The “knowledgeable book”, for educational purposes
- Manually encoded using graphical KA tools



PROJECT HALO



....During metaphase, the centromeres of all the duplicated chromosomes collect along the cell equator, forming a plane midway between the two poles. This plane is called the metaphase plate....

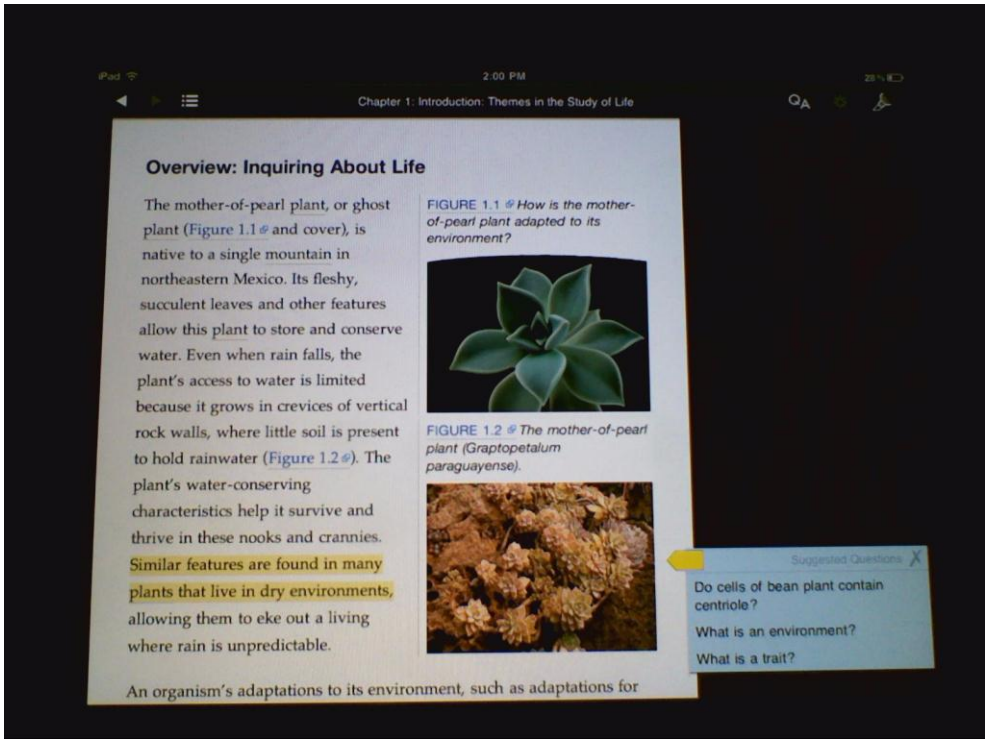


+ reasoning: Deductive elaboration of the graph using other graphs and commonsense rules

Project Halo

an approximation of
part of

- Formally encoding a biology textbook as a KB
  - The “knowledgeable book”, for educational purposes
- Manually encoded using graphical KA tools
- Developing an iPad platform for it



Project Halo

an approximation of
part of

- Formally encoding a biology textbook as a KB
  - The “knowledgeable book”, for educational purposes
- Manually encoded using graphical KA tools
- Developing an iPad platform for it

- This talk:
  - **Not** about the KB, **not** about scalable KA
  - About **understanding questions, given the KB**
    - i.e., bridging the gap between what is asked and what is known

PROJECT HALO

Typical examples of questions the system can answer:

During mitosis of plant cells, when does the cell plate begin to form?

What happens during DNA replication?

What do ribosomes do?

During synapsis, when are chromatids exchanged?

What are the differences between eukaryotic cells and prokaryotic cells?

How many chromosomes are in a human cell?

In which phase of mitosis does the cell divide?

But: There are a lot of questions it can't answer….

The Problem

For many questions:

- Biology knowledge is (somewhat) adequate
- Users avoid high linguistic complexity
- But still: many questions fail (unless worded “just right”)
  - Linguistic fluidity/variability
  - General lexical and world knowledge missing

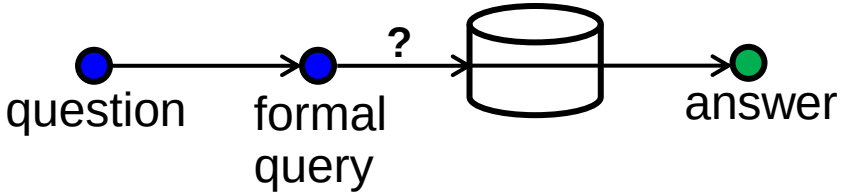
QN: Does a mitochondrion ***require*** oxygen for its function?
KB: Mitochondria -agent-of→ Synthesis -***raw-material***→ Oxygen

QN: Does the nuclear membrane ***break into fragments*** during metaphase?
KB: Metaphase -subevent→ ***Destroy*** -object→ Nuclear-Membrane

PROJECT HALO

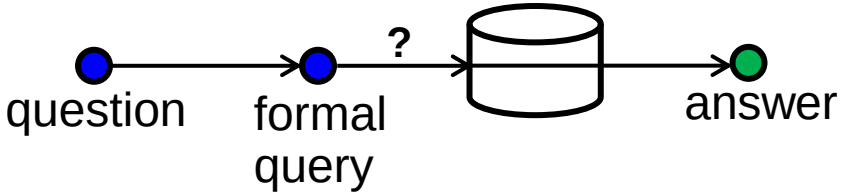
Attempts

1. Pipeline

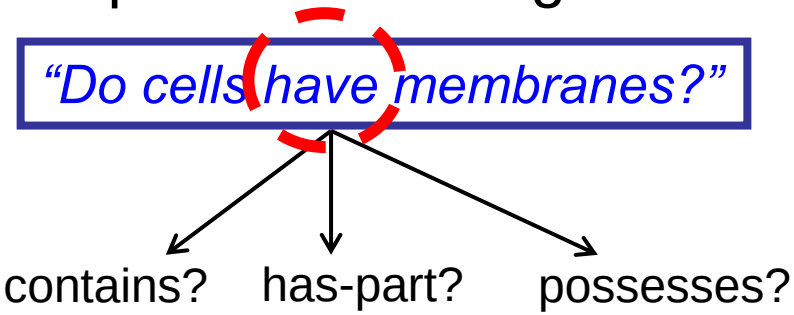


Attempts

1. Pipeline



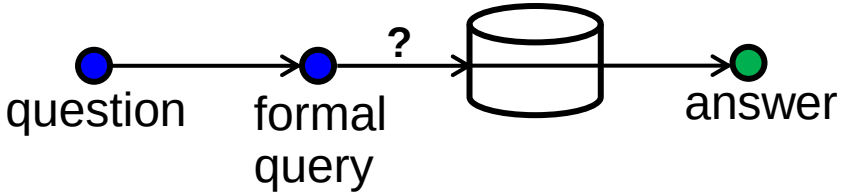
- Hard to do up-front disambiguation reliably



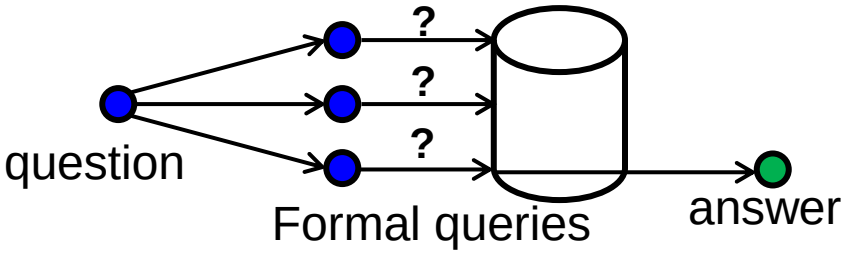
PROJECT HALO

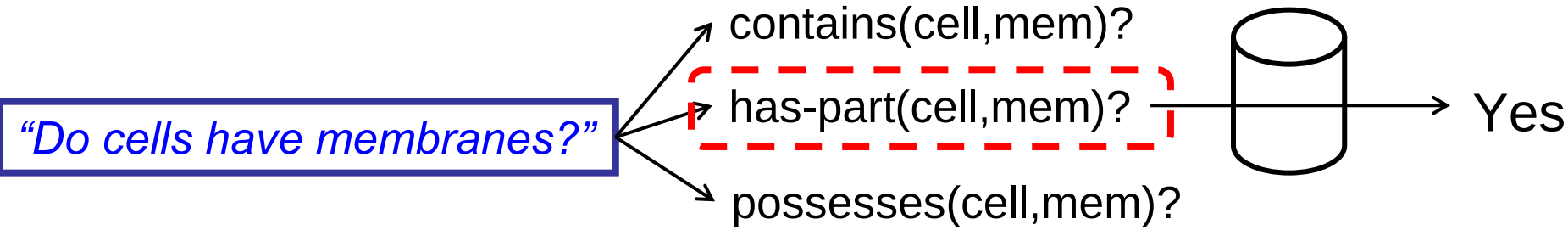
Attempts

1. Pipeline



2. Deferred commitment
   - Try valid disambiguations, prefer those that answer

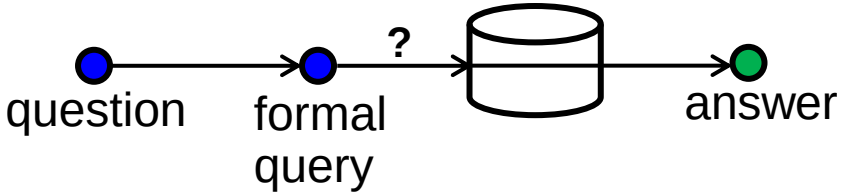




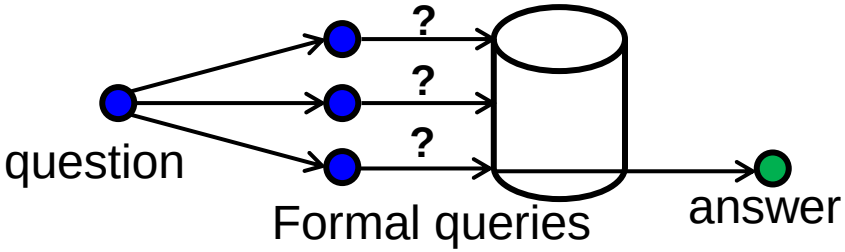
PROJECT HALO

Attempts

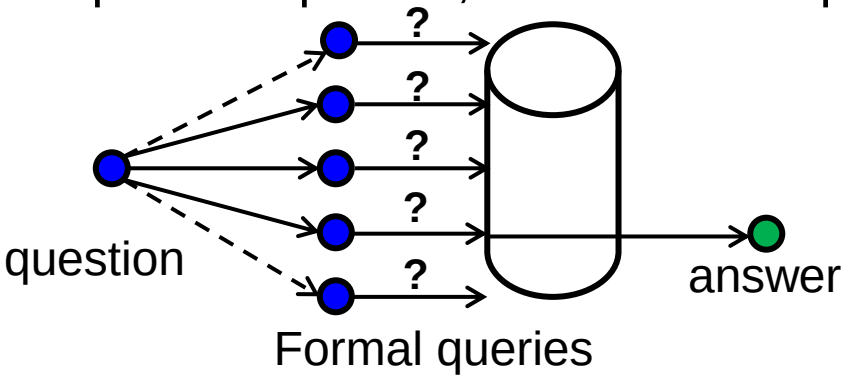
1. Pipeline



2. Deferred commitment
   - Try valid disambiguations, prefer those that answer



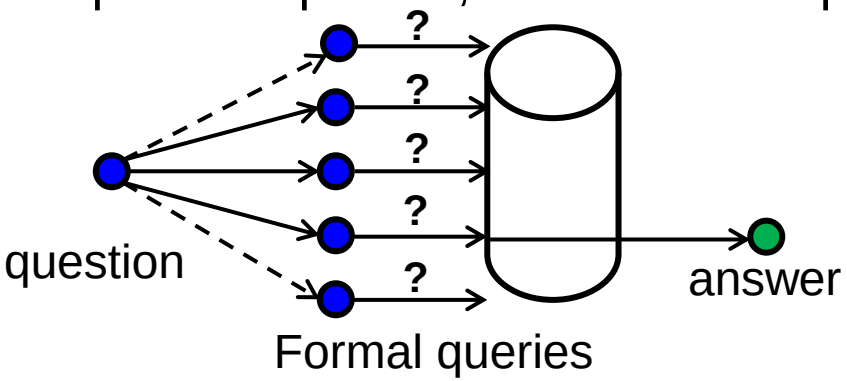
3. Paraphrase-expanded set of formal queries
   - Larger, noisier space of queries, user in the loop



PROJECT HALO

Attempts

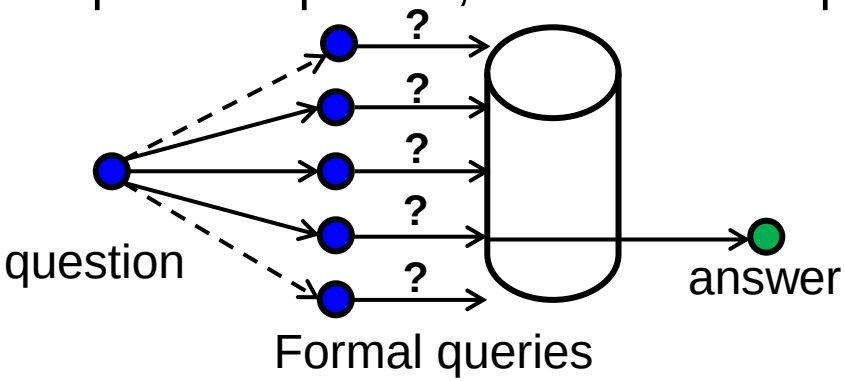
3. Paraphrase-expanded set of formal queries
   - Larger, noisier space of queries, user in the loop



PROJECT HALO

Attempts

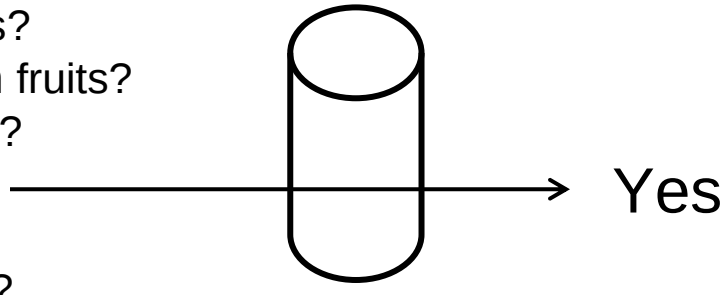
3. Paraphrase-expanded set of formal queries
   - Larger, noisier space of queries, user in the loop



"Are seeds found in fruits?"

DIRT

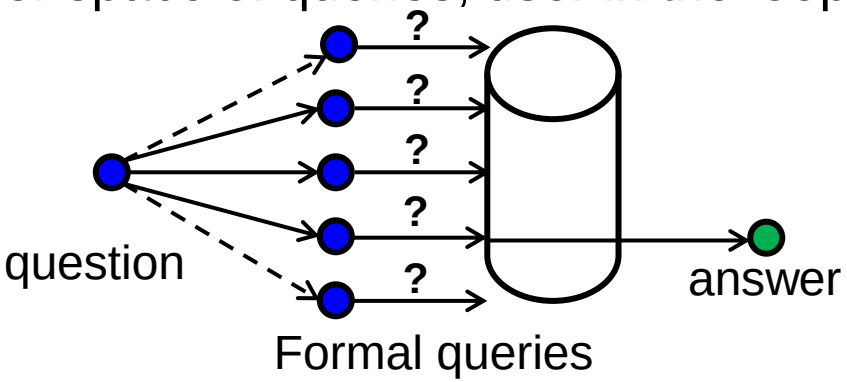
...
Do seeds come from fruits?
Are seeds excavated from fruits?
Do seeds explode in fruits?
Do fruits contain seeds?
Are seeds hid in fruits?
Are seed stashed in fruits?
Do fruits get into seeds?
Are fruits smuggled into seeds?
Are seeds stolen from fruits?
Are seeds used in fruits?
...



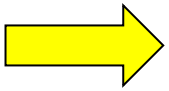
PROJECT HALO

Attempts

3. Paraphrase-expanded set of formal queries
   - Larger, noisier space of queries, user in the loop



   - Helps a bit, ***if*** the question is very close to something answerable
   - But: the problem is more fundamental
     1. We **cannot reliably account** for all possible transformations
        - Language is too fluid
     2. Sometimes the question is **slightly but not significantly different** to what is in the KB
        - Strictly the KB can't answer the question
        - But it could answer something essentially the same



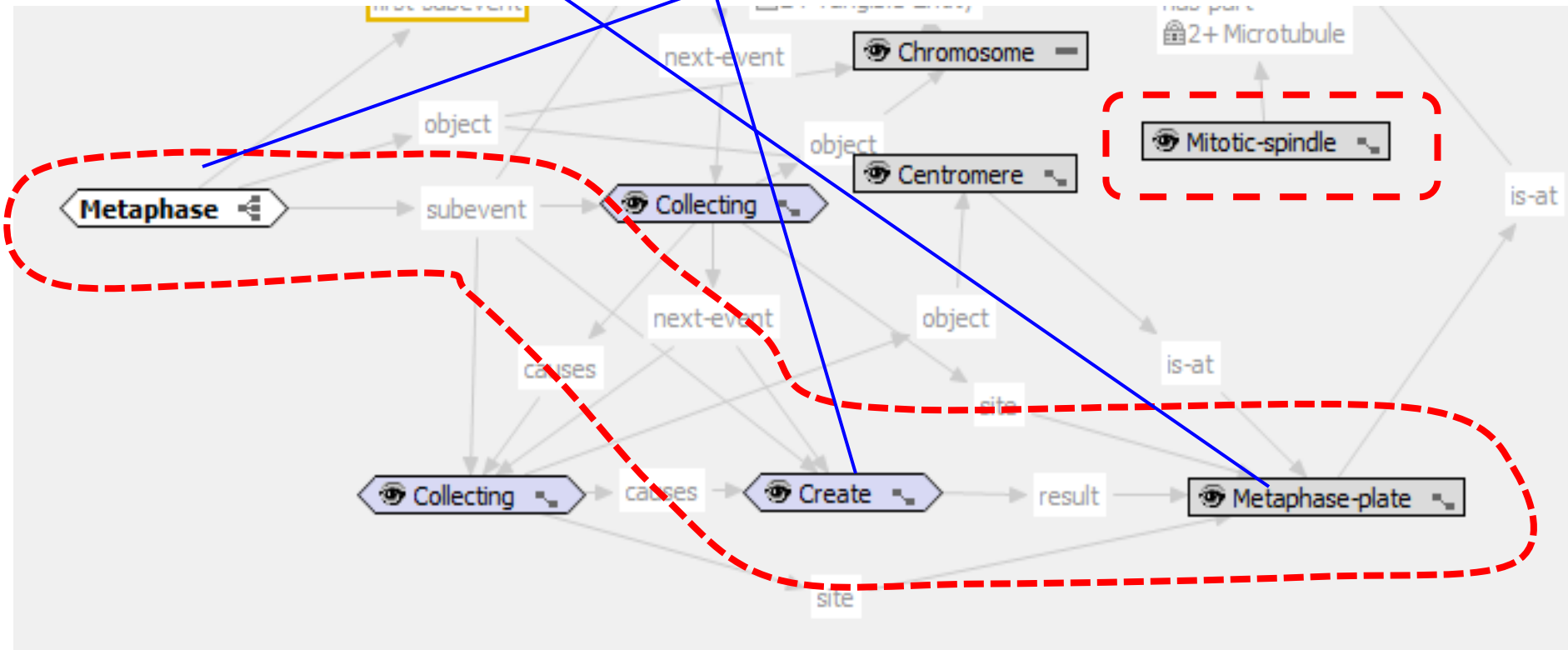
Need to "jump the gap" from language to knowledge

Match what is asked with what is answerable

Example #1

Qn: When is the equatorial plate of the mitotic spindle formed? [Metaphase] ??

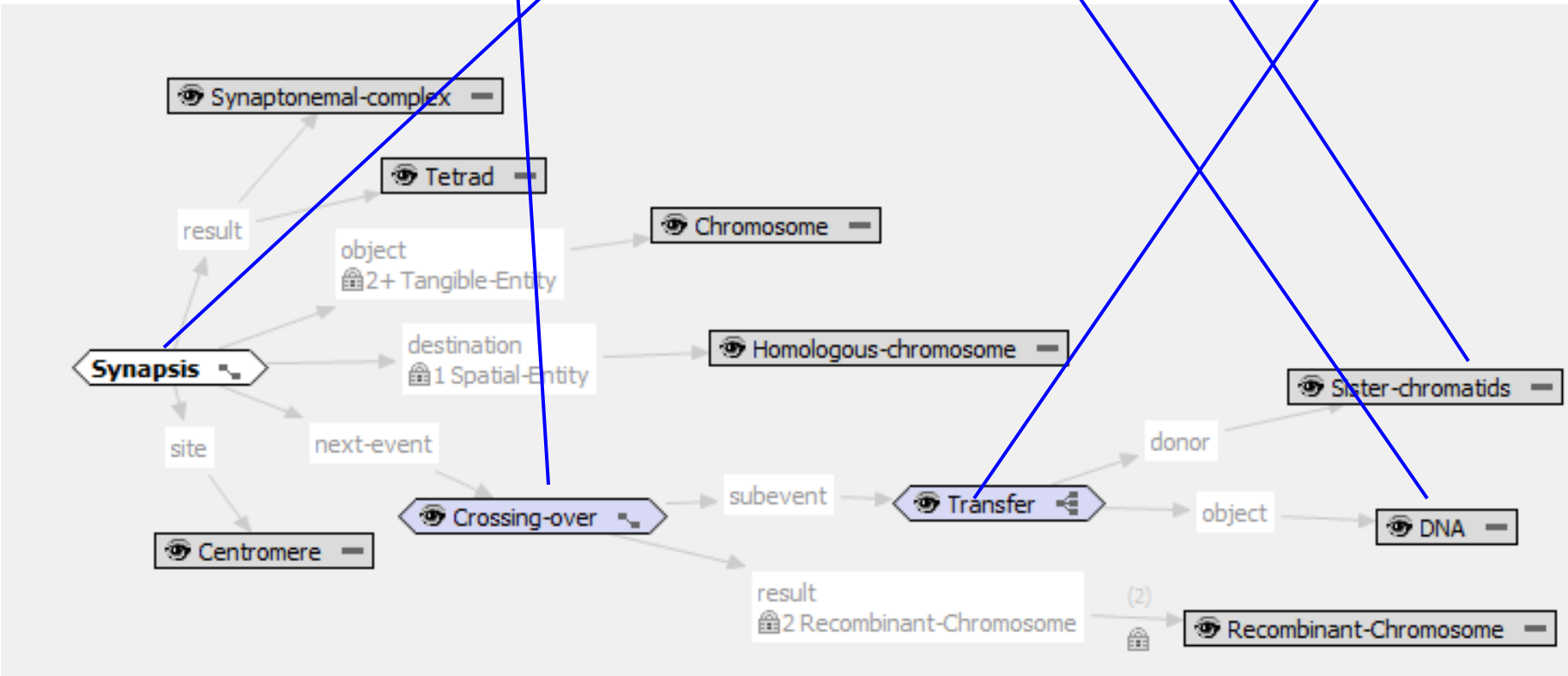
KB: The equatorial plate is formed during metaphase.



Example #2

Qn: When during synapsis are segments of chromatids exchanged? ??

KB: During crossing over during synapsis, DNA of chromatids is transferred.



Extended Q-A Strategy:

- User asks a question
  - Try and interpret and answer it
  - But also find the ***closest, answerable*** question(s)
    - Not provably exactly the same as original question
    - But ideally is essentially the same
    - i.e., provides the information the user is seeking
  - Present these as "suggested questions" to the user

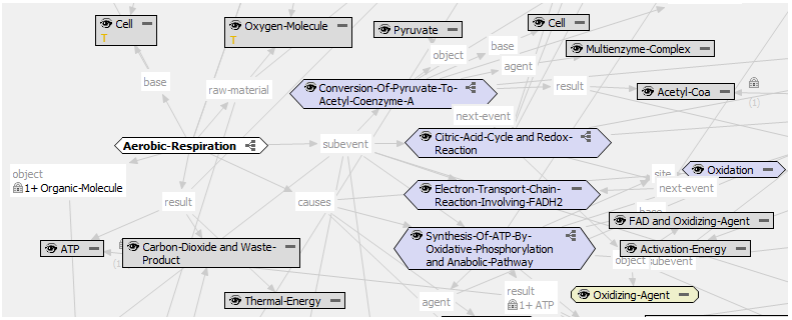
User: When is the equatorial plate of the mitotic spindle formed?
System: Do you mean:
- When is the equatorial plate formed?

An "answerable question"

What is the space of answerable questions?

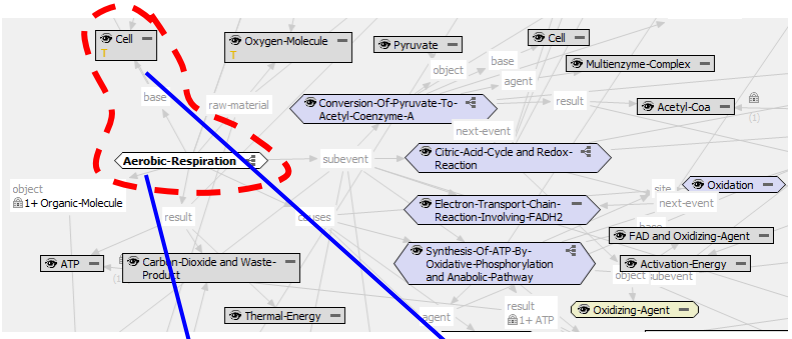
PROJECT HALO

The KB as a formal/compressed textbook



PROJECT HALO

The KB as a formal/compressed textbook

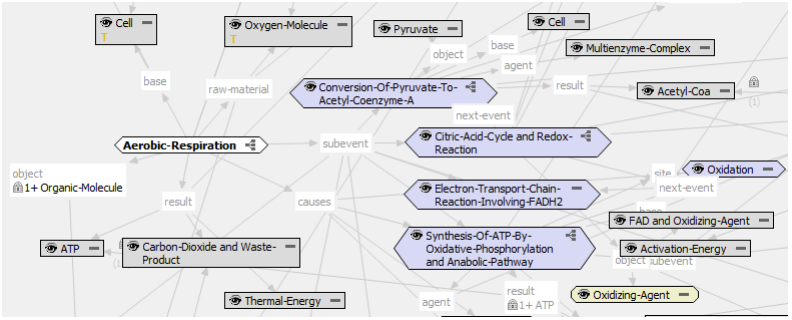


Assertions explicitly in the KB:

Aerobic respiration is performed by cells.

Aerobic respiration is performed by cells.

The KB as a formal/compressed textbook



Assertions explicitly in the KB:

Aerobic respiration is performed by cells.

Aerobic respiration uses oxygen.

Aerobic respiration produces ATP.

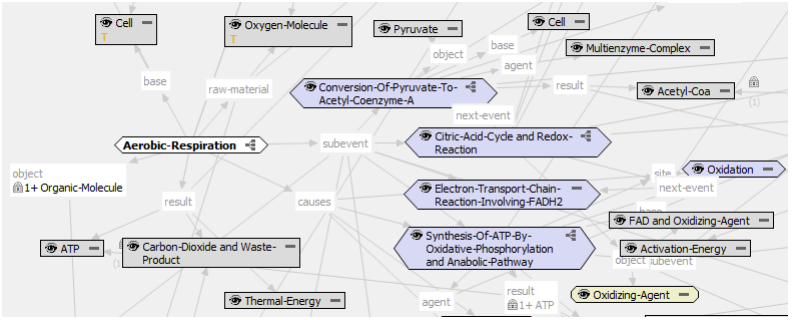
Aerobic respiration involves glycolysis.

….

Aerobic respiration is performed by cells.
Aerobic respiration uses oxygen.
Aerobic respiration produces carbon dioxide and ATP.
Aerobic respiration involves glycolysis.
…

PROJECT HALO

The KB as a formal/compressed textbook



Synonym phrases:

Aerobic respiration is done by cells.

Cells do aerobic respiration.

Aerobic respiration consumes oxygen.

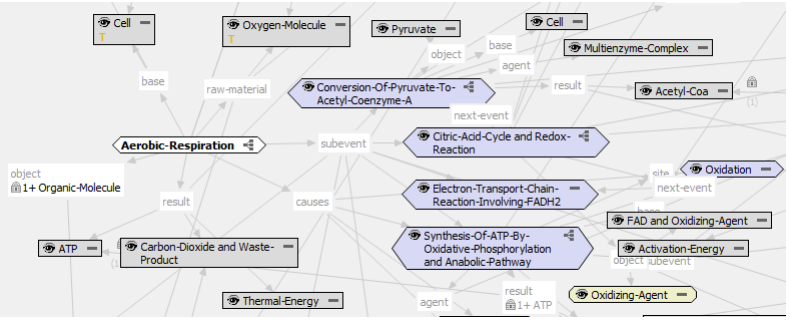
Carbon dioxide is a result of aerobic respiration.

...

Aerobic respiration is performed by cells.
Aerobic respiration uses oxygen.
Aerobic respiration produces carbon dioxide and ATP.
Aerobic respiration involves glycolysis.
…
Aerobic respiration is done by cells.
Cells do aerobic respiration.
Aerobic respiration consumes oxygen.
Carbon dioxide is a result of aerobic respiration.
...

PROJECT HALO

The KB as a formal/compressed textbook



Generalizations and Specializations:

Aerobic respiration is performed by cells.

Aerobic respiration is performed by eukaryotic cells.

Aerobic respiration is performed by plant cells.

Aerobic respiration is performed by bean plant cells.

Respiration is performed by cells.

Respiration is performed by eukaryotic cells.

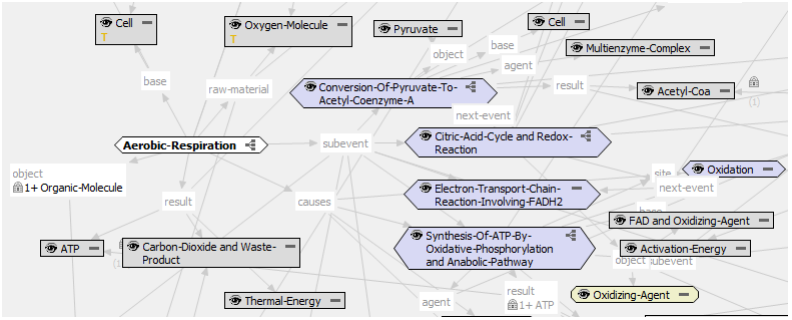
...

Aerobic respiration is performed by cells.
Aerobic respiration uses oxygen.
Aerobic respiration produces carbon dioxide and ATP.
Aerobic respiration involves glycolysis.
…
Aerobic respiration is done by cells.
Cells do aerobic respiration.
Aerobic respiration consumes oxygen.
Carbon dioxide is a result of aerobic respiration.
...

Aerobic respiration is performed by cells.
Aerobic respiration is performed by eukaryotic cells.
Aerobic respiration is performed by plant cells.
Aerobic respiration is performed by bean plant cells.
Respiration is performed by cells.
Respiration is performed by eukaryotic cells.
...
...

PROJECT HALO

The KB as a formal/compressed textbook



Inference:

ATP synthase is used in aerobic respiration.

Pyruvate is an intermediate product in aerobic respiration.

Aerobic respiration produces chemicals.

Aerobic respiration produces energy for use in the cell.

Aerobic respiration is performed by plants.

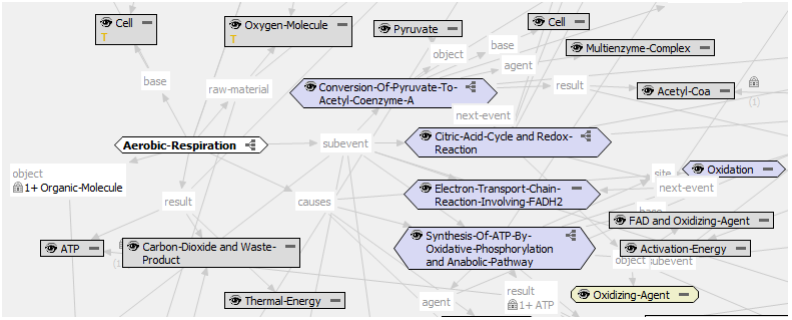
Aerobic respiration is performed by bean plants.

...

Aerobic respiration is performed by cells.
Aerobic respiration uses oxygen.
Aerobic respiration produces carbon dioxide and ATP.
Aerobic respiration involves glycolysis.
…
Aerobic respiration is done by cells.
Cells do aerobic respiration.
Aerobic respiration consumes oxygen.
Carbon dioxide is a result of aerobic respiration.
...
Aerobic respiration is performed by cells.
Aerobic respiration is performed by eukaryotic cells.
Aerobic respiration is performed by plant cells.
Aerobic respiration is performed by bean plant cells.
Respiration is performed by cells.
Respiration is performed by eukaryotic cells.
...
ATP synthase is used in aerobic respiration.
Pyruvate is an intermediate product in aerobic respiration.
Aerobic respiration produces chemicals.
Aerobic respiration produces energy for use in the cell.
Aerobic respiration is performed by plants.
Aerobic respiration is performed by bean plants.
Aerobic respiration requires oxygen.
Respiration requires oxygen.
Breathing requires oxygen.
Oxygen is required to generate ATP in respiration.
Glycolysis requires pyruvate in aerobic respiration.
Glycolysis is a subevent of aerobic respiration.
ATP synthase produces ATP during aerobic respiration.
Glycolysis is a metabolic pathway in aerobic respiration.
Glycolysis is a pathway in aerobic respiration.
Glycolysis is a pathway in respiration.
Glycolysis is a pathway used in respiration.
A pathway used in respiration is glycolysis.
Glycolysis occurs in the cytosol of cells.
Glycolysis occurs in cells.
Cytosol is the location of glycolysis reactions in cells.
During glycolysis, glucose is converted to pyruvate.
Pyruvate is produced via glycolysis.
...

PROJECT HALO

The KB as a formal/compressed textbook



Large space of things that the knowledge base knows

→ large space of *answerable questions*

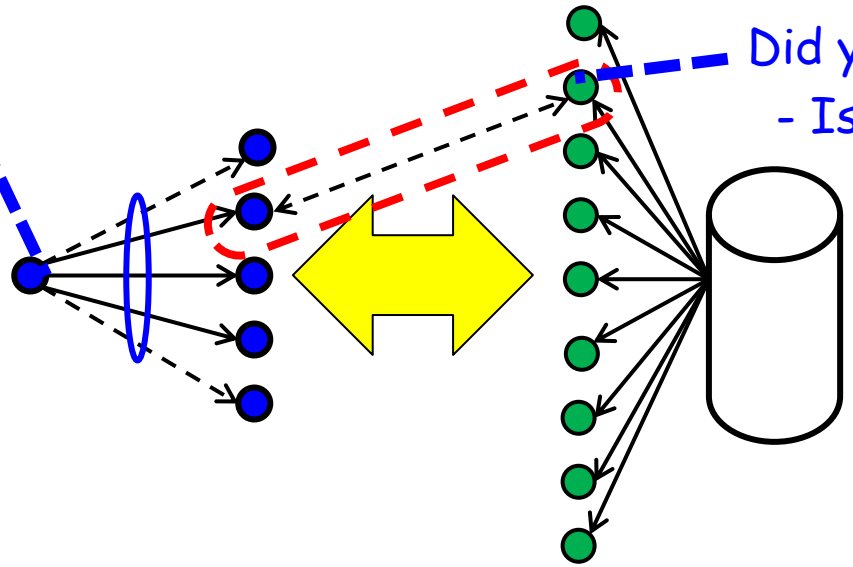
When is ATP synthase used?
What is the role of pyruvate in aerobic respiration?
What are the products of aerobic respiration?
Where is the energy produced by aerobic respiration used?
Do plants perform aerobic respiration?
Do bean plants perform aerobic respiration?
...

Aerobic respiration is performed by cells.
Aerobic respiration uses oxygen.
Aerobic respiration produces carbon dioxide and ATP.
Aerobic respiration involves glycolysis.
…
Aerobic respiration is done by cells.
Cells do aerobic respiration.
Aerobic respiration consumes oxygen.
Carbon dioxide is a result of aerobic respiration.
...
Aerobic respiration is performed by cells.
Aerobic respiration is performed by eukaryotic cells.
Aerobic respiration is performed by plant cells.
Aerobic respiration is performed by bean plant cells.
Respiration is performed by cells.
Respiration is performed by eukaryotic cells.
...
ATP synthase is used in aerobic respiration.
Pyruvate is an intermediate product in aerobic respiration.
Aerobic respiration produces chemicals.
Aerobic respiration produces energy for use in the cell.
Aerobic respiration is performed by plants.
Aerobic respiration is performed by bean plants.
Aerobic respiration requires oxygen.
Respiration requires oxygen.
Breathing requires oxygen.
Oxygen is required to generate ATP in respiration.
Glycolysis requires pyruvate in aerobic respiration.
Glycolysis is a subevent of aerobic respiration.
ATP synthase produces ATP during aerobic respiration.
Glycolysis is a metabolic pathway in aerobic respiration.
Glycolysis is a pathway in aerobic respiration.
Glycolysis is a pathway in respiration.
Glycolysis is a pathway used in respiration.
A pathway used in respiration is glycolysis.
Glycolysis occurs in the cytosol of cells.
Glycolysis occurs in cells.
Cytosol is the location of glycolysis reactions in cells.
During glycolysis, glucose is converted to pyruvate.
Pyruvate is produced via glycolysis.
...

PROJECT HALO

The KB as a formal/compressed textbook

Does photosynthesis need CO_2?



Did you mean:

- Is CO_2 used in photosynthesis?

→ An alignment of question syntax with knowledge

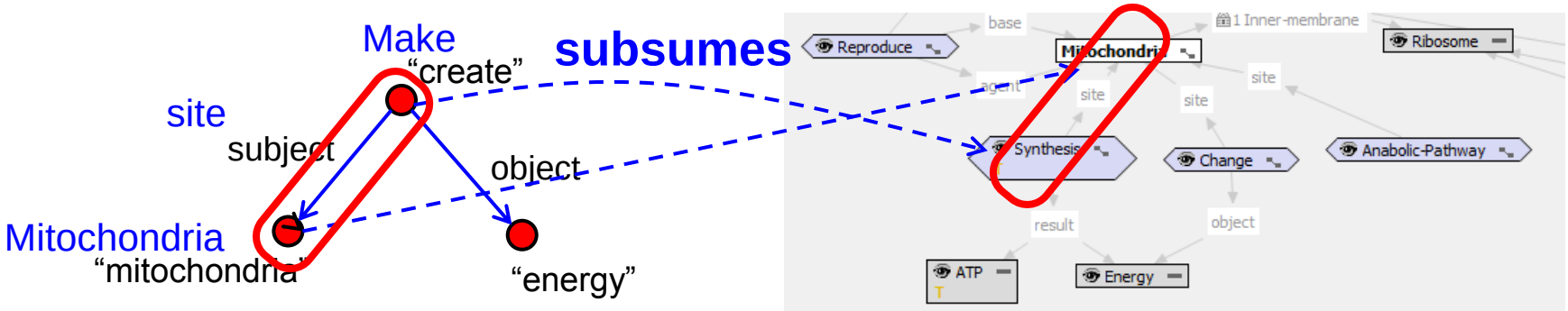
Aerobic respiration is performed by cells.
Aerobic respiration uses oxygen.
Aerobic respiration produces carbon dioxide and ATP.
Aerobic respiration involves glycolysis.
…
Aerobic respiration is done by cells.
Cells do aerobic respiration.
Aerobic respiration consumes oxygen.
Carbon dioxide is a result of aerobic respiration.
...
Aerobic respiration is performed by cells.
Respiration is performed by eukaryotic cells.
...
ATP synthase is used in aerobic respiration.
Pyruvate is an intermediate product in aerobic respiration.
Aerobic respiration produces chemicals.
Aerobic respiration produces energy for use in the cell.
Aerobic respiration is performed by plants.
Aerobic respiration is performed by bean plants.
Aerobic respiration requires oxygen.
Respiration requires oxygen.
Breathing requires oxygen.
Oxygen is required to generate ATP in respiration.
Glycolysis requires pyruvate in aerobic respiration.
Glycolysis is a subevent of aerobic respiration.
ATP synthase produces ATP during aerobic respiration.
Glycolysis is a metabolic pathway in aerobic respiration.
Glycolysis is a pathway in aerobic respiration.
Glycolysis is a pathway in respiration.
Glycolysis is a pathway used in respiration.
A pathway used in respiration is glycolysis.
Glycolysis occurs in the cytosol of cells.
Glycolysis occurs in cells.
Cytosol is the location of glycolysis reactions in cells.
During glycolysis, glucose is converted to pyruvate.
Pyruvate is produced via glycolysis.
...

PROJECT HALO

Finding Answerable Questions

1. Full Interpretation:
   - ***Transform*** the parse tree into something ***fully provable*** from the KB
   - The transformed tree = the interpreted question

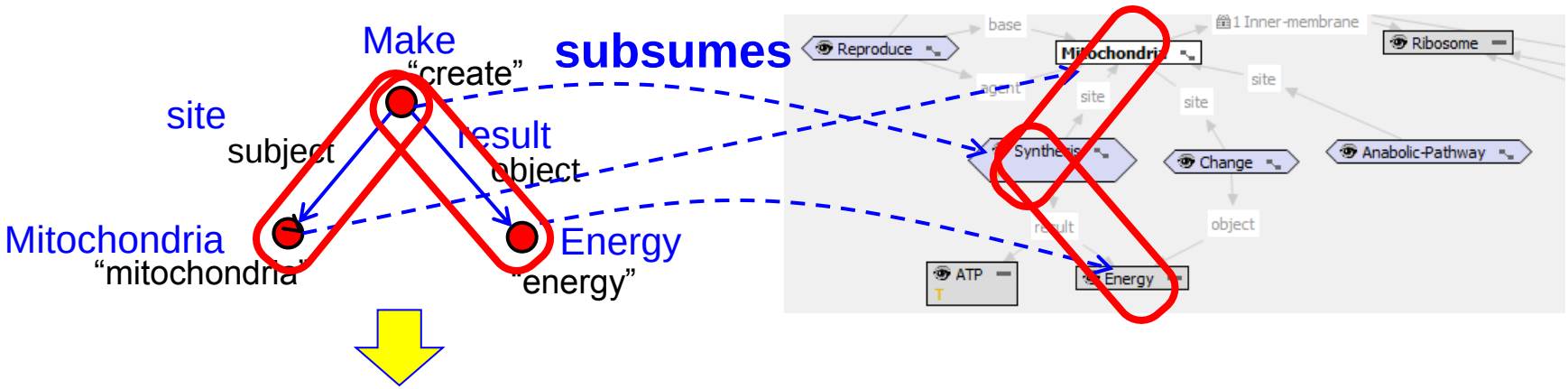
Qn: “Do mitochondria create energy?”



Finding Answerable Questions

1. Full Interpretation:
   - ***Transform*** the parse tree into something ***fully provable*** from the KB
   - The transformed tree = the interpreted question

Qn: "Do mitochondria create energy?"



Interpreted Question:

isa(mito01,Mitochondria).

?- site-of(mito01,?m), result(?m,?e), isa(?e,Energy).

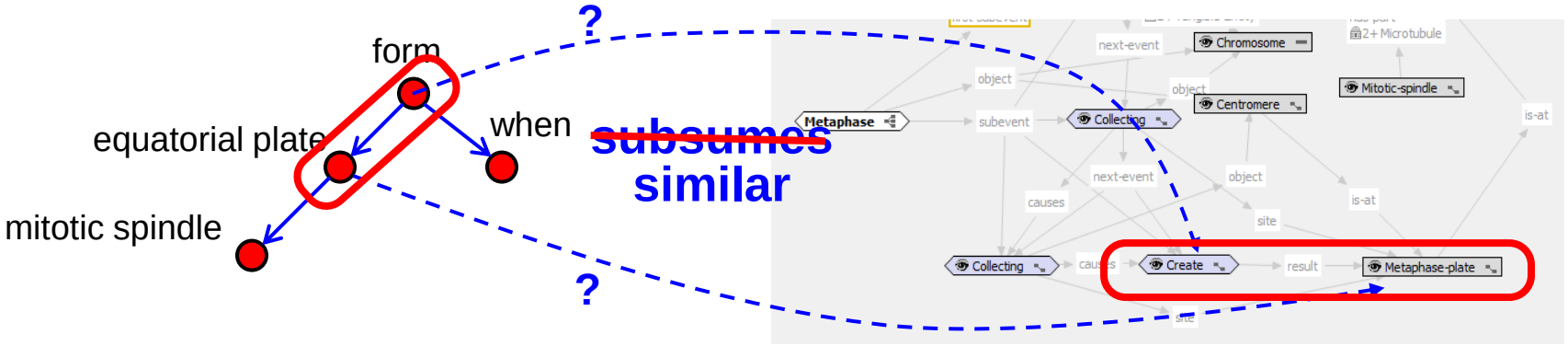
"Do mitochondria make energy?" "Yes!"

PROJECT HALO

Finding Answerable Questions

1. Full Interpretation:
   - ***Transform*** the parse tree into something ***fully provable*** from the KB
   - The transformed tree = the interpreted question
2. Finding nearest answerable question:
   - Transform the tree into something ***best matching*** part of the KB
   - That part of the KB = the ***best answerable question***

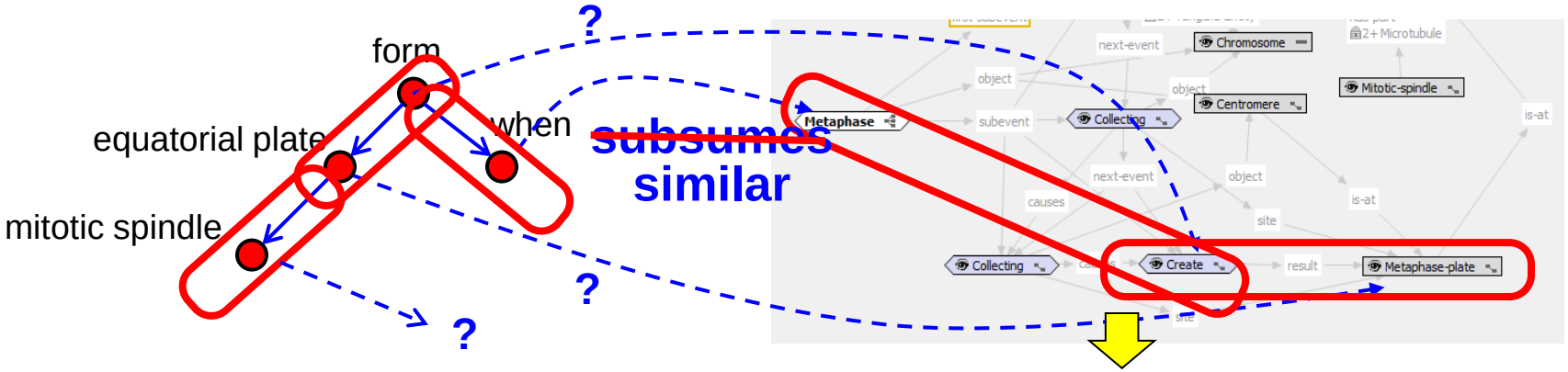
Qn “When is the equatorial plate of the mitotic spindle formed?”



Finding Answerable Questions

1. Full Interpretation:
   - ***Transform*** the parse tree into something ***fully provable*** from the KB
   - The transformed tree = the interpreted question
2. Finding nearest answerable question:
   - Transform the tree into something ***best matching*** part of the KB
   - That part of the KB = the ***best answerable question***

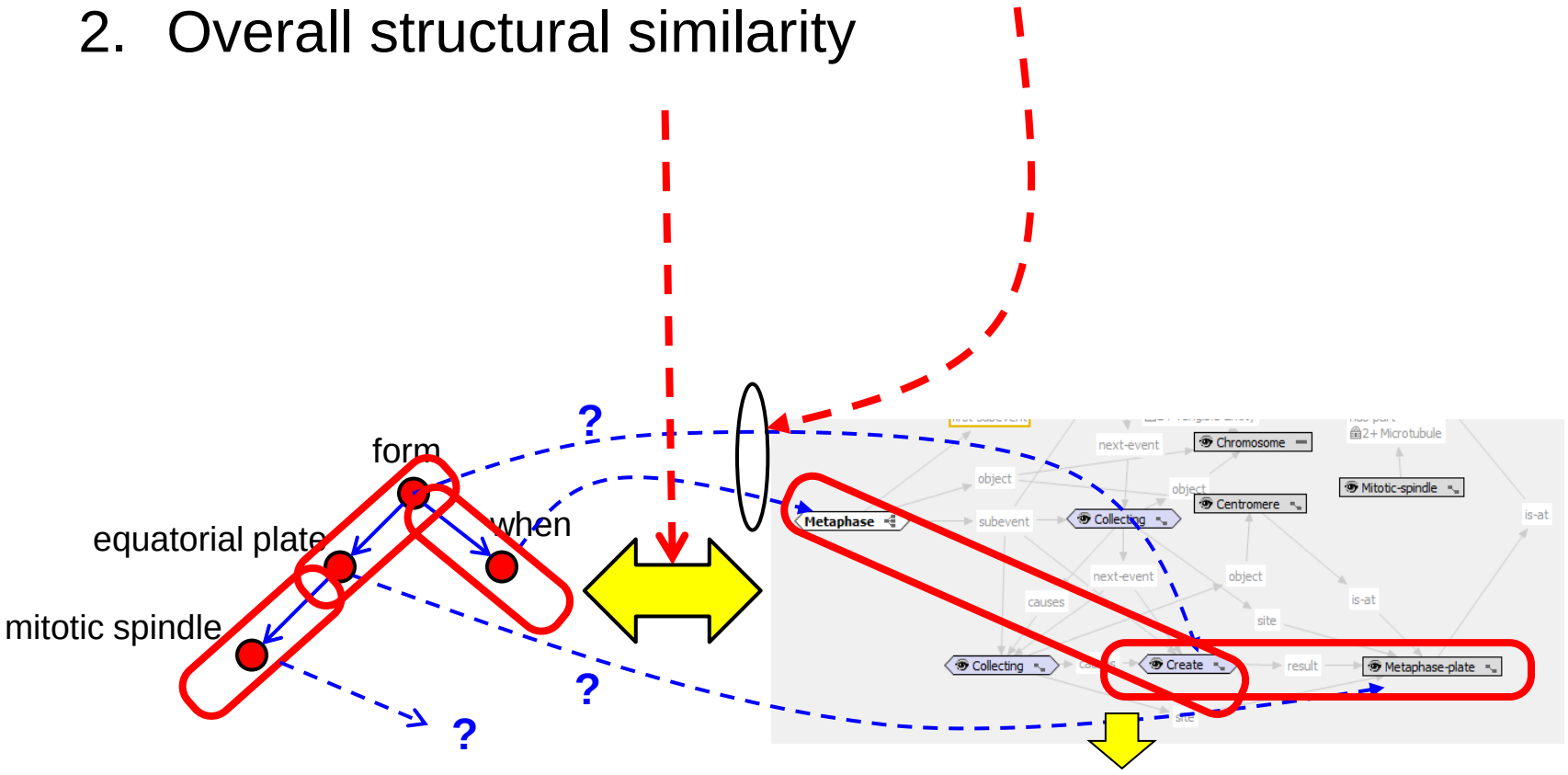
Qn "When is the equatorial plate of the mitotic spindle formed?"



Answerable Question:
"When is the equatorial plate created?"

Finding Answerable Questions

- Two trainable similarity metrics
  1. Subgraph similarities (e.g., node-node)
  2. Overall structural similarity



Answerable Question:

"When is the equatorial plate created?"

Finding Answerable Questions

- Two trainable similarity metrics:

1. Subgraph similarities (e.g., node-node)

   KB may not know a concept can be realized as a word/phrase

   But other knowledge sources offer evidence, e.g.,

   "consume" ↔ raw-material:

   - Wordnet: "raw material": material used; "use": consume fully
   - DIRT: "material for" → "used for" → "used in" → "consumed in"
   - Pivot-based paraphrasing: "raw material for the" → consumption of"

PROJECT HALO

Finding Answerable Questions

- Two trainable similarity metrics:

1. Subgraph similarities (eg node-node)

	WordNet distance	Google distance	Lexical overlap	Proposition store evidence	Etc.
Create ↔ "form"	0.9	0.8	0.0	0.7	…
raw-material ↔ "consume	0.3	0.5	0.0	0.3	…
has-part ↔ "present in"	etc. etc.				

2. Structural similarity

	concept overlap	relation overlap	# orphans	∑ Similarities	Etc.
	0.92	0.83	1	0.74	…
	etc. etc.				

PROJECT HALO

Sources of Training Data

- Existing question+answer suites
  - Best suggest question produces same answer → good

User: What organelles are generated at a nucleolus? **Ribosome**
System: Do you mean: What organelles are synthesized at a nucleolus? **Ribosome**

User: Which vesicles break down cellular debris? **Lysosome**
System: Do you mean: What is outside a cell? **Extra-Cellular-Matrix**

+

-

- Click feedback from user:

User: When is the equatorial plate of the mitotic spindle formed?
System: Do you mean:
- When is the mitotic spindle formed?
- When is the equatorial plate formed? ⇦
- When does the equatorial plate break up?
- …

PROJECT HALO

Two Final Thoughts

1. Is querying a DB/KB and text-based QA so different?…
   - Matching text against formal sentences
     - Or rather, against linguistic expressions of those sentences

2. What about knowledge acquisition?
   - Making a best guess at *what is meant* in terms of *what you know* is broader than just understanding questions

Summary

- **Project Halo:** A biology book as a knowledge base
- **Interpreting Questions:**
  - **Can't** always fully prove a question
  - Doesn't matter how well authored the knowledge is, there's still a QA problem
- **Current work:**
  - Extending to find the "nearest answerable question"
    - Add weaker evidence to suggest nearness
    - Use machine learning to control weights
    - User click responses → potentially new training data
  - "Jumping the gap" from language to knowledge

Thank you!

PROJECT HALO