Constructing a Knowledge Base on Aging

An Automated Approach

Mark Farrell

Bioinformatics Researcher

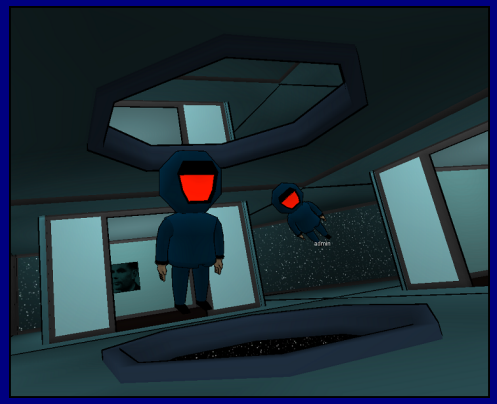
Center for Research and Education on Aging
Lawrence Berkeley National Laboratory
University of California, Berkeley

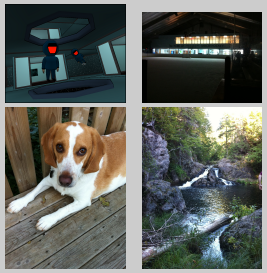
September 4th, 2014

Outline

Who I am

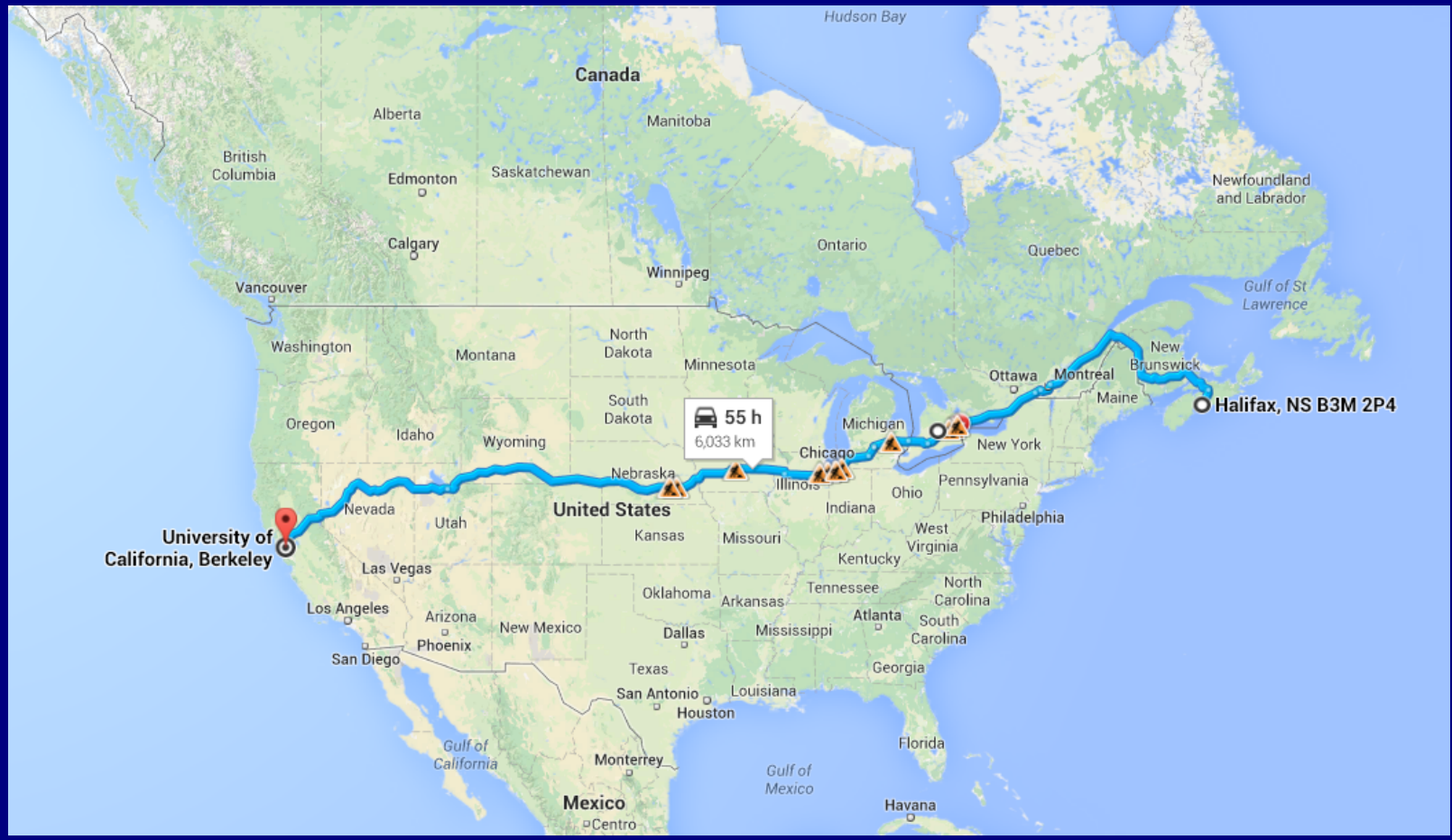
Mark Farrell

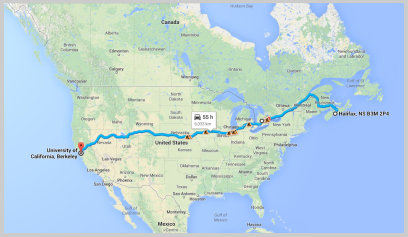




I'm going to start off by introducing myself. This is my first time being here in person, so I doubt many of you know who I am. I've lived most of my life in Canada's East Coast. I was back home again this summer, working as a co-op student for Defence R&D Canada.

Where I'm from
Atlantic Canada





So, I've had to travel quite a bit over the past few days in order to arrive here at Berkeley.

What I'm Studying

Year	2
Program	Bachelor of Computer Science
Faculty	Mathematics
Institution	University of Waterloo
Location	Waterloo, Ontario, Canada

What I'm Studying

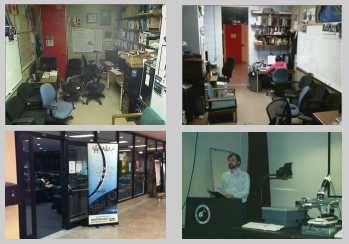
Year 2
Program Bachelor of Computer Science
Faculty Mathematics
Institution University of Waterloo
Location Waterloo, Ontario, Canada

And in case you're wondering, I'm going into my second year of undergraduate studies at the University of Waterloo, majoring in Computer Science.

How I Became a Researcher at CREA

Prof. Garan Emailed the University of Waterloo Computer Science Club

How I Became a Researcher at CREA



That's all very well and good: I haven't explained how I got involved with research at CREA. Steve Garan (you might know him) emailed the University of Waterloo Computer Science Club back in January, looking for current members interested in an AI internship; he used to be President back in the 80s. It looked interesting, so I began working with him, meeting weekly on Skype to discuss progress.

Why was I Interested in Research at CREA?

Hobby	CREA Research
Reverse Engineering	Natural Language Processing
Game Development	Artificial Intelligence
Health	Bioinformatics

Why was I Interested in Research at CREA?

Hobby	CREA Research
Reverse Engineering	Natural Language Processing
Game Development	Artificial Intelligence
Health	Bioinformatics

So, why was I interested in research at CREA? I've worked on some software projects involving reverse engineering network protocols and game development in spare time. And, I felt that CREA was offering me an opportunity to hone my interests, allowing me to deepen knowledge of computing beyond what I would likely accomplish on my own.

Outline

Ok, so let's talk about 'Understanding Aging' and where I come in.

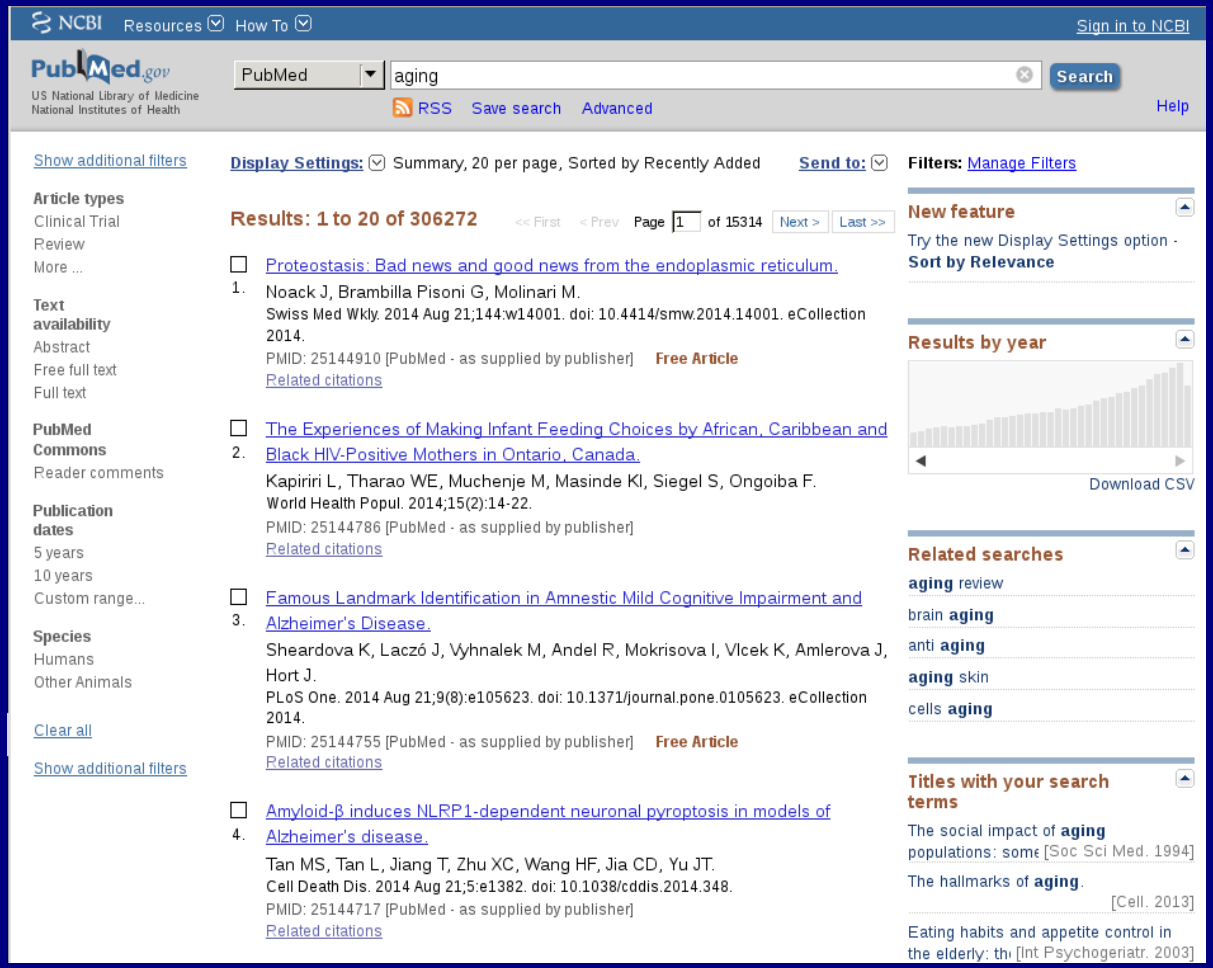
What does CREA want to do?

CREA Wants to Understand Aging and Increase Human Lifespan

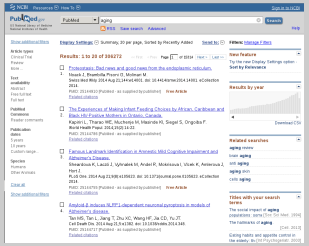
What does CREA want to do?
CREA Wants to Understand Aging and Increase Human Lifespan

As you might already know, CREA wants to understand aging and increase human lifespan.

Why is it Difficult to Understand Aging?



Why is it Difficult to Understand Aging?



But, Why is it Difficult to Understand Aging? If you look at this figure here, you can see that not only is there a large volume of publications made on the subject (on the order of 300 000), but also the rate at which new publications are being made has been doubling over the past few decades.

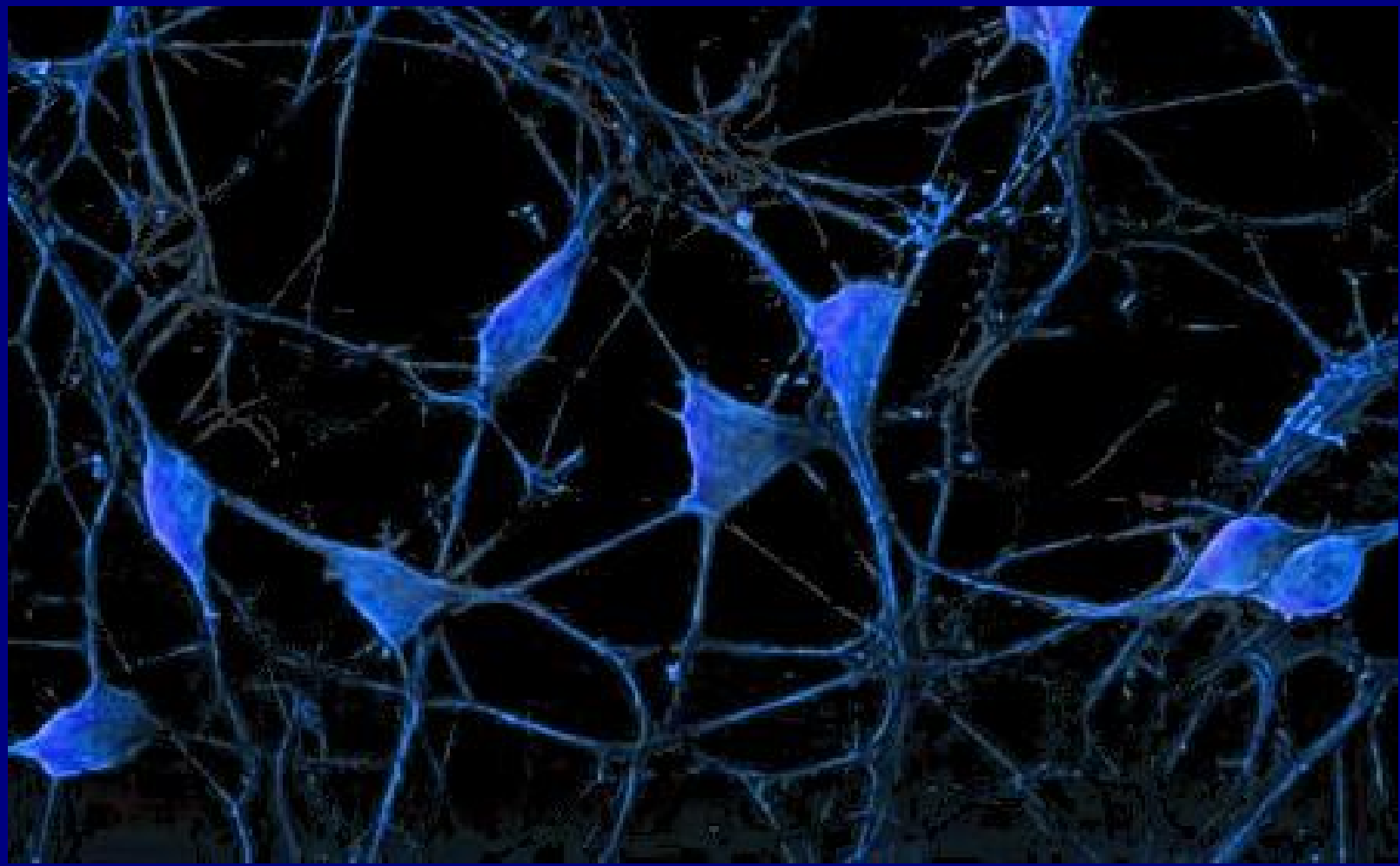
Why is it Difficult to Understand Aging?

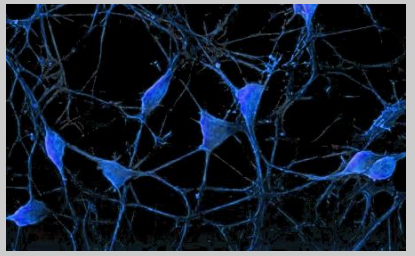
Why is it Difficult to Understand Aging?

So, we need a better way to manage past and future knowledge that will help us understand aging; this is a problem.

How can CREA Understand Aging?

Build a Knowledge Base that Describes and Reasons about Aging

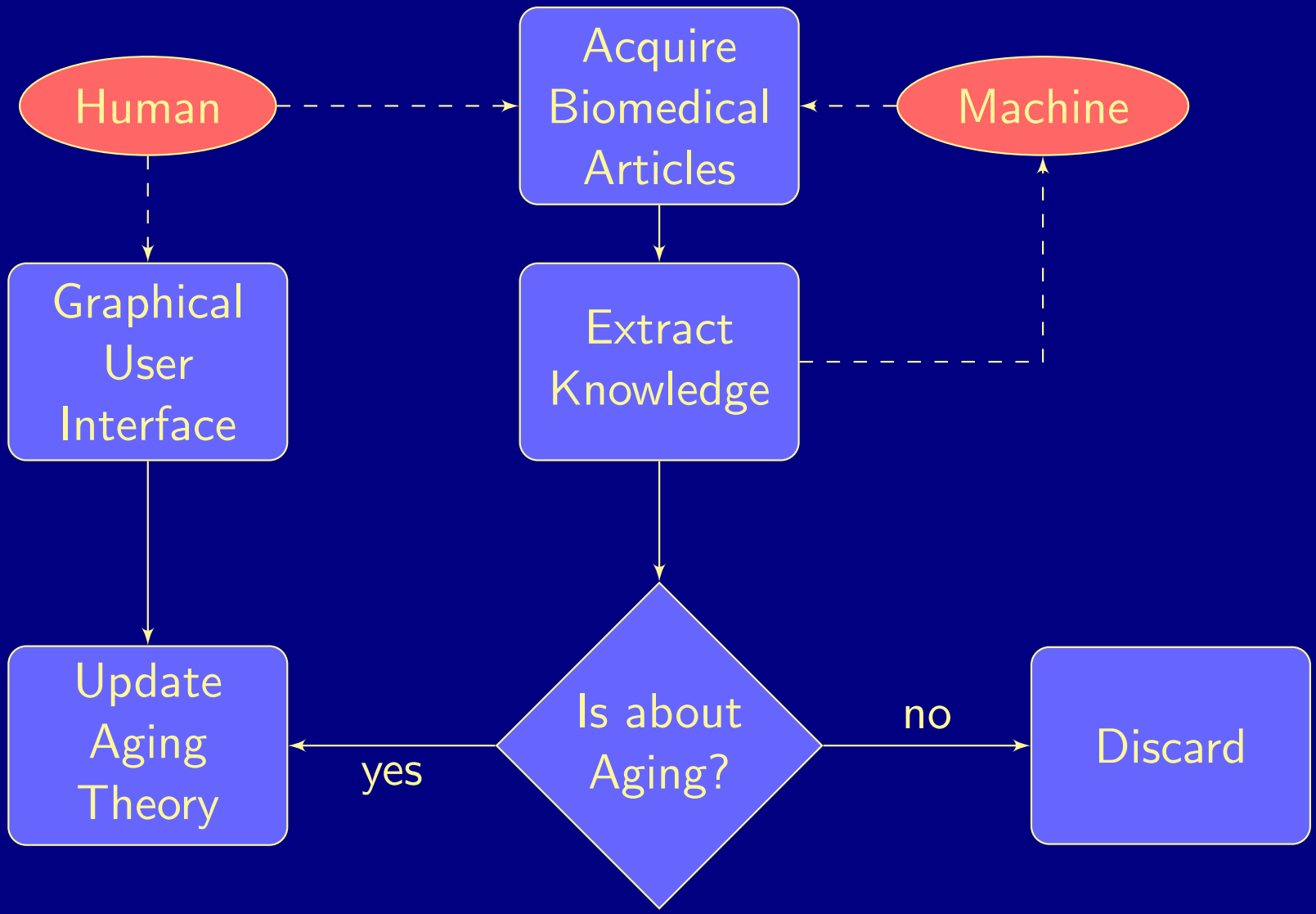


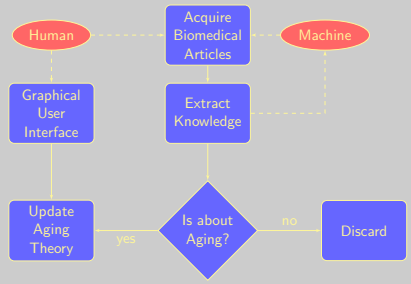


And that's where I come in. I've been trying to work on a solution to this problem, which involves building a knowledge base, software, that can assist humans in order to understand aging and develop theories about it.

How can CREA Understand Aging?

Build a Knowledge Base that Describes and Reasons about Aging





Have a look at this figure: the idea behind the software project I'm working on is to get a machine to read text and try to understand aging, just like how human's can try understand aging by reading texts and journal articles. And also, to present a machine's understanding of aging to humans and allow them correct and modify the machine's understanding of aging in a intuitive way.

Outline

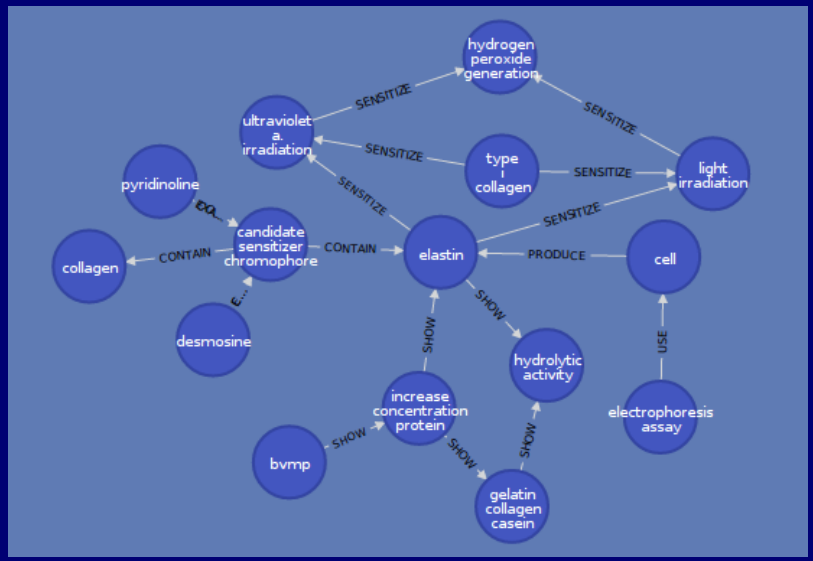
To date, I've made quite a bit of progress in allowing a machine to read and extract knowledge from text articles, and have also been able to display that knowledge in a manner that helps humans understand aging. However, I haven't made much progress on allowing the machine to form a theory of aging and reason about aging itself; i.e. it can't really decide what knowledge is relevant to aging and what it is not.

Knowledge Extraction
Overview of Progress

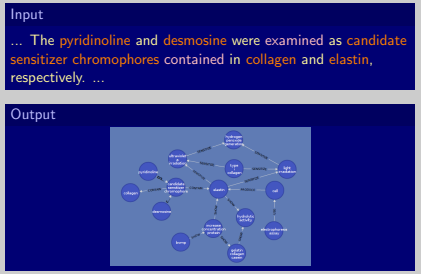
Input

... The pyridinoline and desmosine were examined as candidate sensitizer chromophores contained in collagen and elastin, respectively. ...

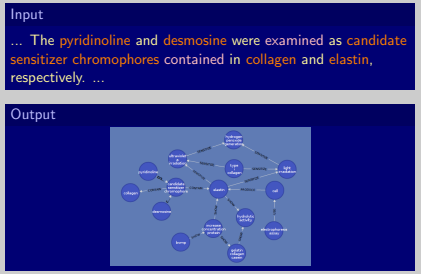
Output



Knowledge Extraction
Overview of Progress



As Steve Garan put it, aging can be looked at as the degradation over time and eventual failure of an organism's systems. So let's start off talking about knowledge extraction with an excerpt from a journal abstract that is related to elastin degradation. Have a look at the example input sentence: I've highlighted the literal nouns found in the sentence and the literal verbs that relate nouns to one another. Now look at the output: knowledge extraction is the process of identifying literal nouns and relating them to each other, i.e. making logical assertions about the actions that they can perform on one another.



Natural language data is rather unstructured, and so I have to decide on what components of natural language are essential for describing how systems and processes work when translating it into a structured format that a machine can understand and reason about. There are problems with the same literal nouns and verbs being used to represent multiple concepts in natural language. Take people's names for example: i.e. two different people can have the same name. However, I need to be able to blindly extract knowledge first before worrying about those problems, and extracting knowledge blindly is still quite useful for us when it comes to understanding and reasoning about aging; we just have to be aware that there could be problems.

Knowledge Extraction
Method

Tokenization Input a text document and read it, one sentence at a time.

Knowledge Extraction
Method

Tokenization Input a text document and read it, one sentence at a time.

Parsing For each sentence, generate a constituent tree that describes its phrase structure.

Knowledge Extraction
Method

Tokenization Input a text document and read it, one sentence at a time.

Parsing For each sentence, generate a constituent tree that describes its phrase structure.

Compilation Extract facts by pattern matching on each constituent tree.

Knowledge Extraction
Method

Tokenization Input a text document and read it, one sentence at a time.

Parsing For each sentence, generate a constituent tree that describes its phrase structure.

Completion Extract facts by pattern matching on each constituent tree.

And this is how my software extracts knowledge from text articles. Read the sentences left to right, one at a time, and extract knowledge from each sentence. Examine the grammatical structure of each sentence and translate it into a format that a machine can understand and reason about; that's knowledge extraction.

Constituent Tree Tags

Word Level

Tag	Meaning	Example
DT	Determiner	the
IN	Preposition	of
JJ	Adjective	blue
RB	Adverb	quickly
CC	Coordinating Conjunction	and
NN	Singular Noun	monkey
NNS	Plural Noun	monkeys
VB	Base Verb	fall
VBZ	Singular Present Verb	falls
VBD	Past Tense Verb	fell
VBN	Past Participle Verb	fallen
VBG	Gerund Verb	falling
...	...	...

Constituent Tree Tags

Tag	Meaning	Example
DT	Determiner	the
IN	Preposition	of
JJ	Adjective	blue
RB	Adverb	quickly
CC	Coordinating Conjunction	and
NN	Singular Noun	monkey
NNS	Plural Noun	monkeys
VB	Base Verb	fall
VBZ	Singular Present Verb	falls
VBD	Past Tense Verb	fell
VBN	Past Participle Verb	fallen
VBG	Gerund Verb	falling
...	...	...

But, there needs to be some way for a machine to examine sentence structure. Let's first talk about how linguistics might formally examine sentence structure, so that it becomes obvious how a machine might do so as well. First, they'd label what part-of-speech each word is in sentence: e.g. blue is an adjective, dog is a singular noun and so on and so forth.

Constituent Tree Tags
Phrase Level

Tag	Meaning	Example
NP	Noun Phrase	the woman
VP	Verb Phrase	calls the man
PP	Prepositional Phrase	from the store
ADVP	Adverb Phrase	quickly and quietly
ADJP	Adjective Phrase	blue and red
CONJP	Conjunctive Phrase	as well as
...	...	...

Constituent Tree Tags
Phrase Level

Tag	Meaning	Example
NP	Noun Phrase	the woman
VP	Verb Phrase	calls the man
PP	Prepositional Phrase	from the store
ADVP	Adverb Phrase	quickly and quietly
ADJP	Adjective Phrase	blue and red
CONJP	Conjunctive Phrase	as well as
...	...	...

Then, they'd label groups of words that contain certains parts-of-speech - these are called phrases. For example "the woman" is a phrase, containing a determiner and singular noun.

Constituent Tree Tags
Clause Level

Tag	Meaning	Example
S	Declarative Clause	the dog walks
SBAR	Conjunction + Clause	that the dog walks
...	...	...

Constituent Tree Tags
Clause Level

Tag	Meaning	Example
S	Declarative Clause	the dog walks
SBAR	Conjunction + Clause	that the dog walks
...	...	...

And lastly, groups of phrases that, together, form a logical proposition, are called clauses. For instance "the dog walks" is a clause because, when read, it makes a logical assertion about the dog.

Knowledge Extraction
Example

Input Token

The man walks the dog.

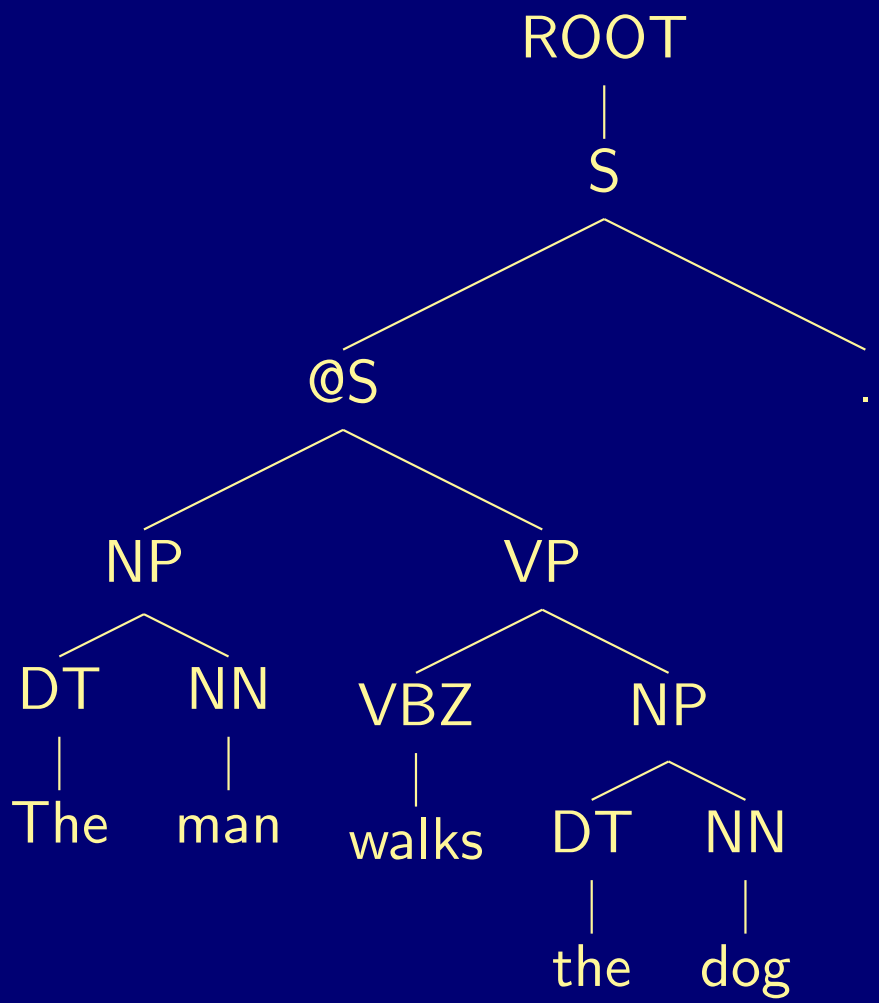
Knowledge Extraction

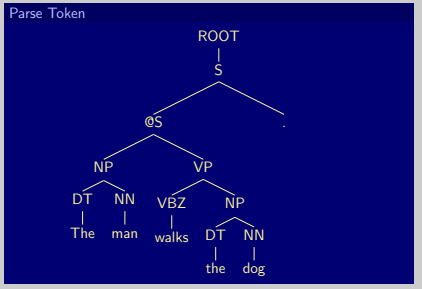
Input Token
The man walks the dog.

Now let's look at this example sentence and its grammatical structure.

Knowledge Extraction
Example

Parse Token

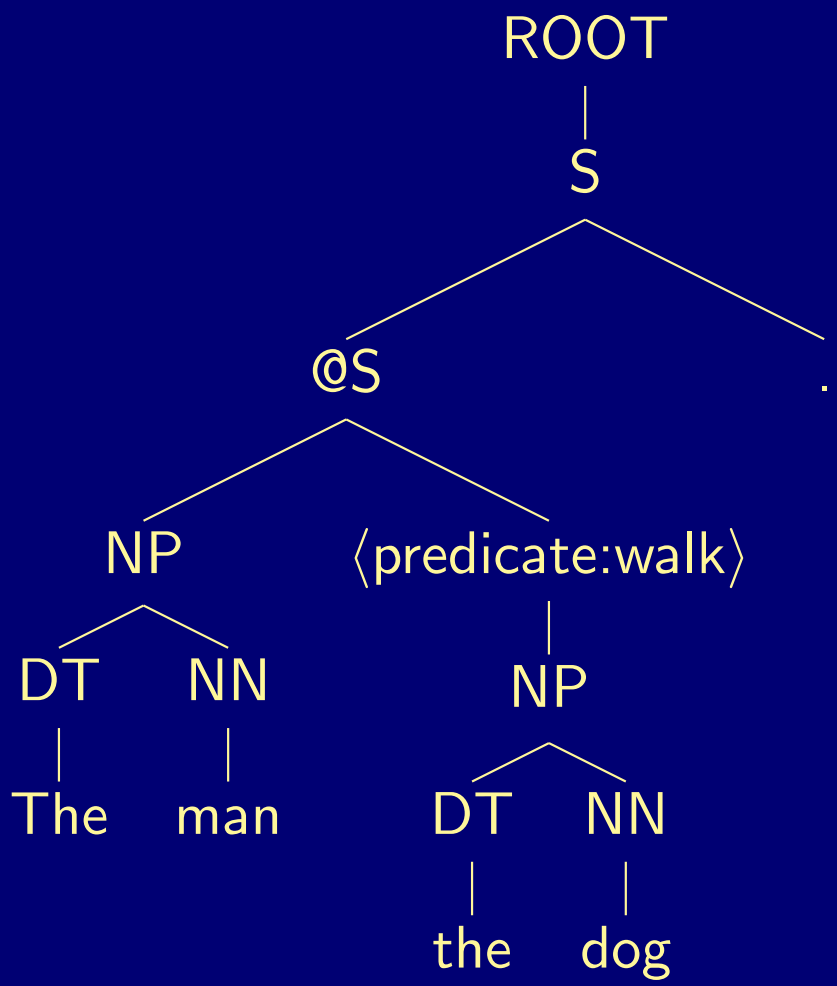




Based on the information I just gave you, we could have created this illustration of the sentence's structure ourselves. But, a machine needs to be able to do this as well, it needs to be able to parse sentences and generate constituent trees, in order to extract knowledge from them. So, there are several software tools available for training a machine to parse sentences, some using different methods than others; I have been using software called The Berkeley Parser, which worked fairly well, i.e. it has been fairly accurate when parsing sentences.

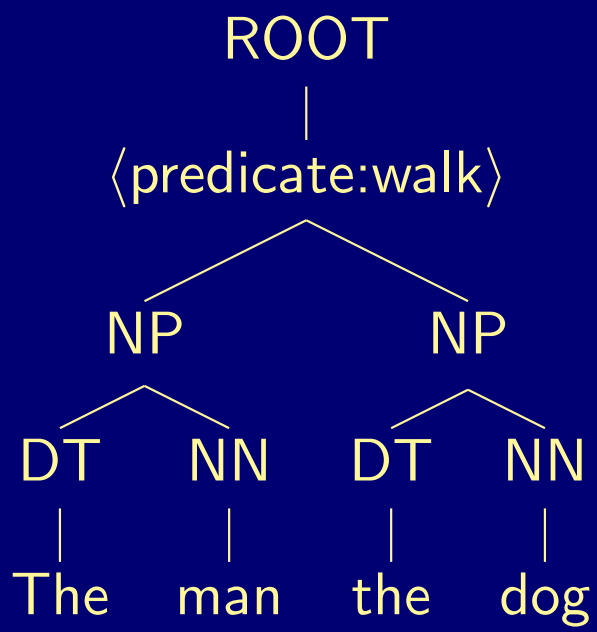
Knowledge Extraction
Example

Compile Token



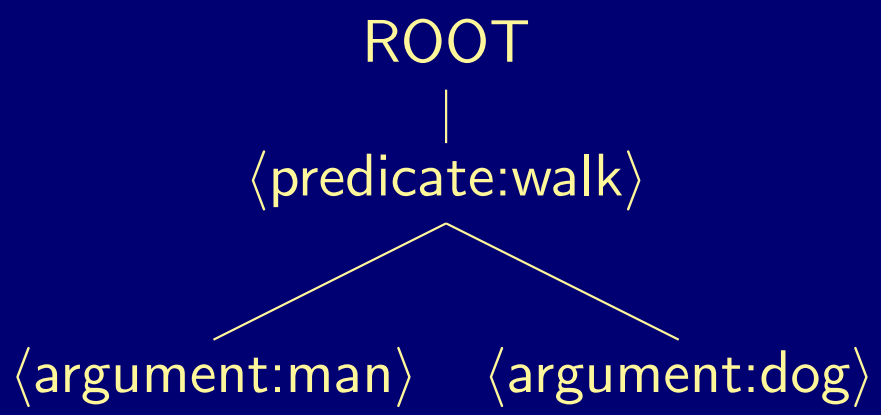
Knowledge Extraction
Example

Compile Token



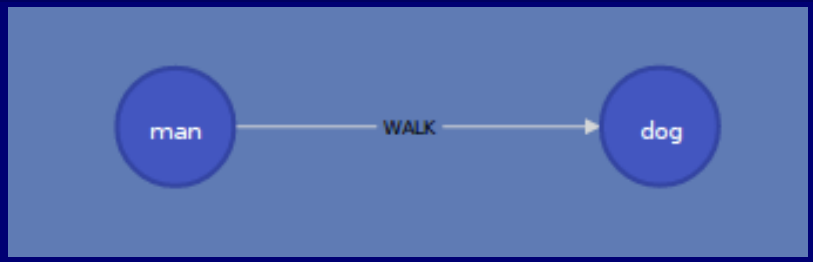
Knowledge Extraction
Example

Compile Token



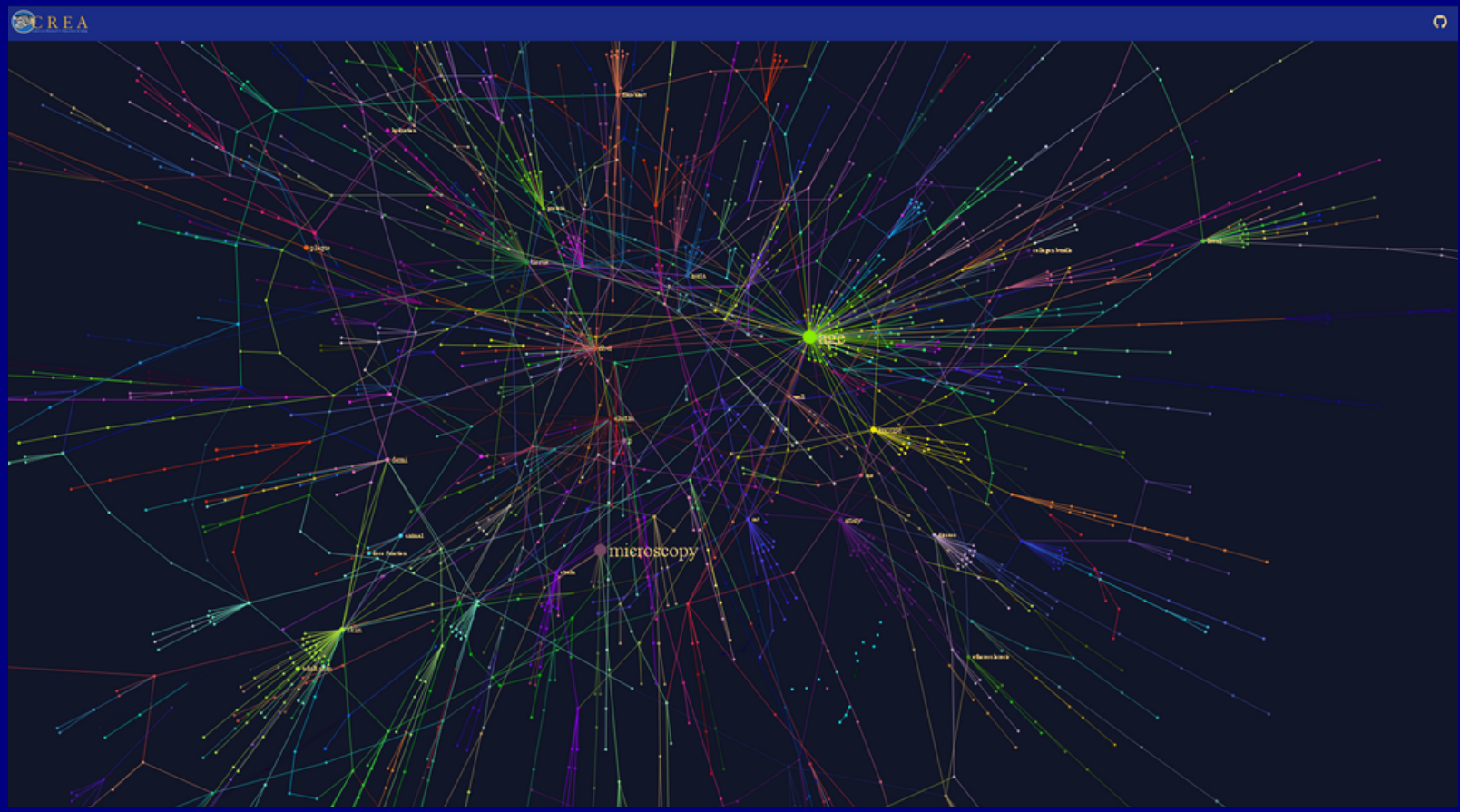
Knowledge Extraction
Example

Output Fact



Software Demonstration

A preview of CREA's knowledge base, compiled from PubMed abstracts.



High Performance

Knowledge can be extracted from many sentences at the same time.

Outline

What Next?

Task	Assignees
Research Direction	Steve Garan
Knowledge Extraction	Mark Farrell
Text Article Retrieval	Grace Park, Jeremy Wan
Knowledge Visualization	Mark Farrell
Graphical User Interface	Mark Farrell
Biomedical Spam Filtering	—
Automated Reasoning	—

Get Involved
Contact Me for More Information

Email m4farrel@csclub.uwaterloo.ca

Website markfarrell.ca