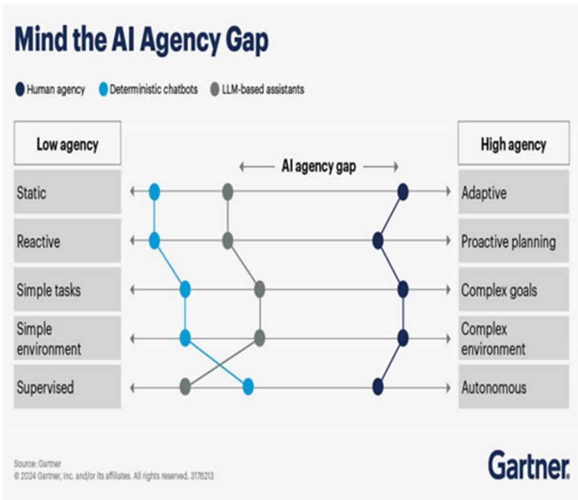
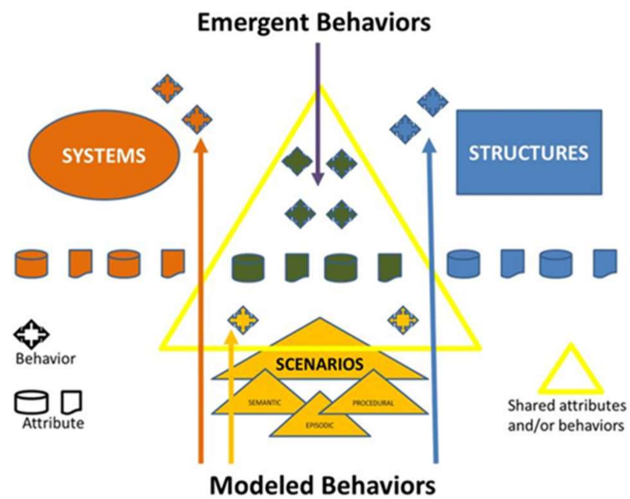
Agentic AI – Roadmap to High Agency

Joe Glick, Cofounder, Chief Innovation Officer – **RYLTI** Inc.

Enabling High Agency by Modeling Real World Ecosystems





Contents:

Gartner Perspective

Gartner believes that the future of AI is about agency, but today there are challenges. These include:
- Agentic AI proliferating without governance or tracking
- Agentic AI making decisions that are not trustworthy
- Agentic AI relying on low-quality data
- Employee resistance
- Agentic-AI-driven cyberattacks enabling “smart malware”

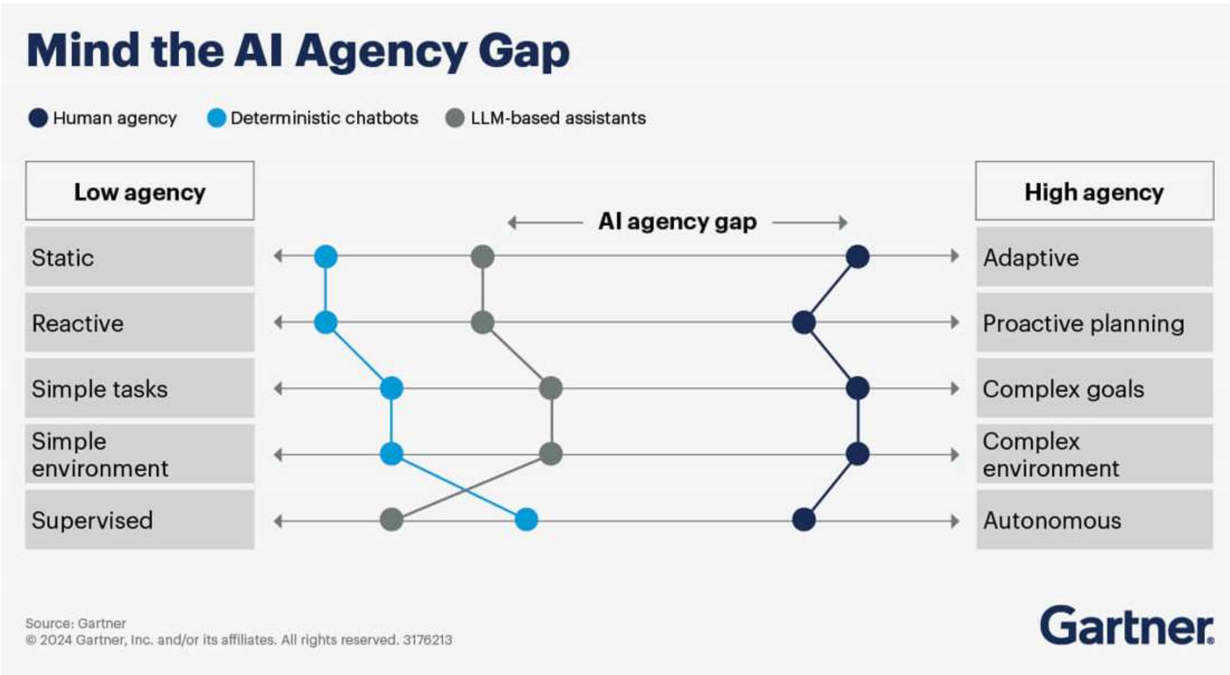
Effectively managing the risks of software entities acting autonomously requires advanced tools and strict guardrails.

Can the agency gaps be filled? Is high agency, comparable to human capabilities, achievable?

Strategy for Filling the Gap

The greatest value and challenges are associated with complex and adaptive agents. Gartner identifies LLM-based assistants as the current peak AI agents.

However, that technology is on the low end of the agency continuum, without a roadmap to high agency. Assistants are apps wrapped around LLMs which cannot identify multidimensional and multiscale complexity and cannot adapt to new and dynamic environments.



Required Strategy – BIOMIMICRY
The world is composed of innumerable ecosystems that behave as agents and are composed of agents, all dynamically adapting in highly complex environments.

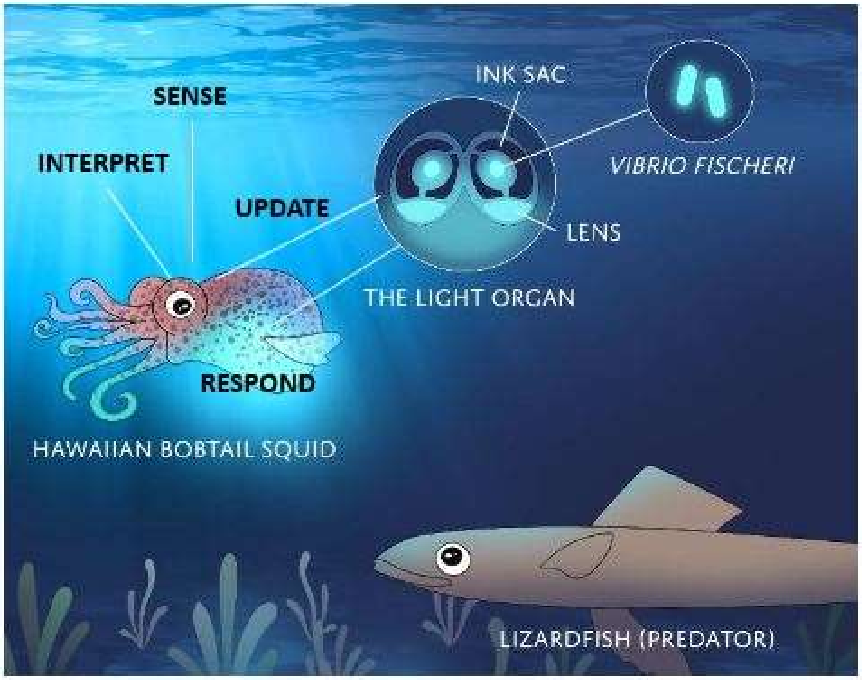
The NAS published a Physics of Life Report which concluded, "An important lesson from the long and complex history of neural networks and artificial intelligence is that revolutionary technology can be based on ideas and principles drawn from an understanding of life, rather than on direct harnessing of life's mechanisms or hardware." h�ps://doi.org/10.17226/26403

Enabling High Agency – MODELING MULTI-AGENT ECOSYSTEMS
Princeton researchers studying the stealth mode of the Hawaiian bobtail squid opened the door to discovering the microbiome. What began as marine zoology research was expanded to explain how this squid becomes invisible to its predators when it hunts at night.

It turned out that a foreign, bioluminescent bacteria enter the squid after it hatches and colonizes a light organ. This organ is immature until the bacteria enters and a gene expression stimulates the light organ to complete its development.

The squid uses a sensor on its head to measure the light above which varies based on changing atmospheric conditions and needs to adjust the light organ cover to the extent necessary to match the overhead light intensity, making it invisible to predators below.

DYNAMIC ADAPTABILITY TO CHANGING LIGHT



This research provided evidence that systems thinking can be used to predict macroscopic phenomena while bypassing the need to explicitly unmask all the quantitative dynamics operating at the microscopic level.

Value Delivered – DISCOVERY
Aside from learning to understand this amazing behavior, the discovery of the symbiotic relationship between the squid and the foreign bacteria was the first step to what is now our growing knowledge about the human microbiome. Most recently we are learning that our microbiome with non-human DNA may be as individual as our fingerprints (https://linkinghub.elsevier.com/retrieve/pii/S1931312824000568).

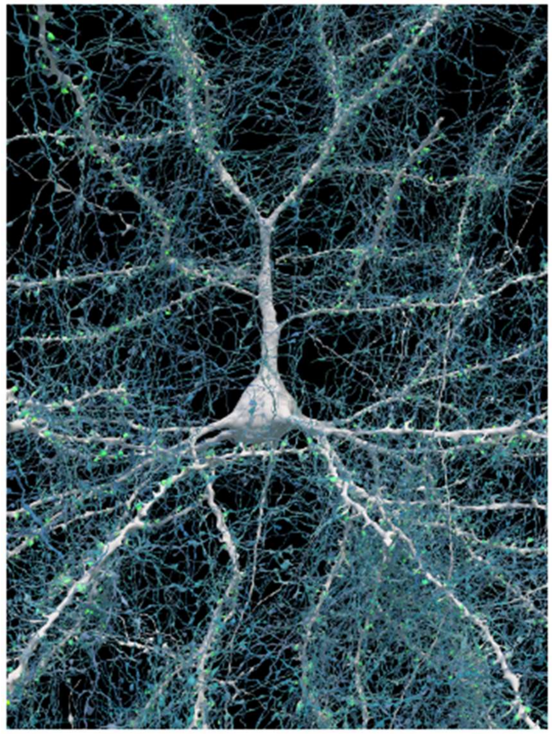
Whether we work in biopharma, product development or any other complex and dynamic field, what we do not know is our largest source of opportunity, and or risk.

Discovering what we do not know is critical and of high value.

BIOMIMETICS – TECHNOLOGY IMITATING LIFE

Researchers in the 1950s launched AI science with the early neural networks aiming to replicate the human brain so biomimicry was the intent from the start, but the earliest achievements took half a century. Why?

1. PRESUMPTION: human neurons are a collection of homogeneous nodes repeating defined processes.
   REALITY: Current brain research has identified over 3000 neuron types, and 1 cubic mm. (detailed image yielding 1.4 Petabytes of data.)



A single neuron is shown with 5,600 of the nerve fibers (blue) that connect to it. The synapses that make these connections are in green. *Google Research & Lichtman Lab, Harvard University. Renderings by D. Berger, Harvard*

2. PRESUMPTION: intelligence emerges from scaling data and repetitive computations.
   REALITY: infants learn as many languages as they are expected to by contextualizing small, diverse data to build a mental model of the world.

MODELING THE REAL WORLD IS POSSIBLE

Our knowledge of biological complexity is rapidly expanding, especially in the field of genomics. We used to refer to DNA as a blueprint, but now we understand it to be a molecular ecosystem. Artificial intelligence (AI) and other technologies, particularly Large Language Models (LLMs) hold promise for driving advances in biomedical research, but, along with their benefits, these technologies have limitations. Traditional AI normalizes data and removes outliers, preventing the identification of “dark” data – key relationships and critical interactions in the real world. AI and LLMs also require huge training datasets, and the associated scale of combinatorial math limits the ability of the algorithms to explore biological complexity.

To address these issues and to provide guidance to the biomedical community, the National Academies of Sciences, Engineering and Medicine (NAS), sponsored by the National Institutes of Health (NIH), the National Science Foundation (NSF) and the Department of Energy (DOE), began advocating research into the use of biomimetic digital twins technology to more effectively model multidimensional and multi-scale biological complexity (the complete report - http://nap.nationalacademies.org/26894).

While reporting on the multi-agency workshop on biomedical digital twins, Bissan Al-Lazikani MD, Anderson Cancer Center, explained that bottlenecks in drug discovery arise owing to the challenges of multidisciplinary and multiscale data integration and multiparameter optimization. To alleviate the issues associated with integrating data from disparate disciplines that span scales, instead of integrating the data points themselves, she suggested integrating how all of the data points interact with each other—essentially establishing edges that can be modeled graphically. This approach, which is especially useful when data are sparse, is advantageous in that different data are captured in the same logic

Key Challenges from NAS Digital Twin Ecosystem Report

1. A digital twin is more than just simulation and modeling. Digital twins have been the subject of widespread interest and enthusiasm; it is challenging to separate what is true from what is merely aspirational, due to a lack of agreement across domains and sectors as well as misinformation.
2. Many of the potential uses of digital twins are currently intractable to realize with existing computational resources.
3. Hybrid modeling approaches are a productive path forward for meeting the modeling needs of digital twins, but their effectiveness and practical use are limited ***by key gaps in theory and methods.***
4. Integration of component/subsystem digital twins is a pacing item for the digital twin representation of a complex system, especially if different fidelity models are used in the digital twin representation of its components/subsystems.
5. Digital twins will typically entail high-dimensional parameter spaces. This poses a significant challenge to state-of-the-art surrogate modeling methods.

Our Theory and Methods

We believe that the fundamental theory required to begin addressing these challenges was identified in a Physics of Life Report published by the NAS which concluded, "An important lesson from the long and complex history of neural networks and artificial intelligence is that revolutionary technology can be based on ideas and principles drawn from an understanding of life, rather than on direct harnessing of life's mechanisms or hardware." This indicates the need for a biomimetic engineering approach to building and maintaining the digital twin ecosystem. (http://nap.nationalacademies.org/26894)

We have evolved the methods below in the real-world lab of delivering solutions to explore highly complex, multidimensional, and multiscale problem domains for global organizations.

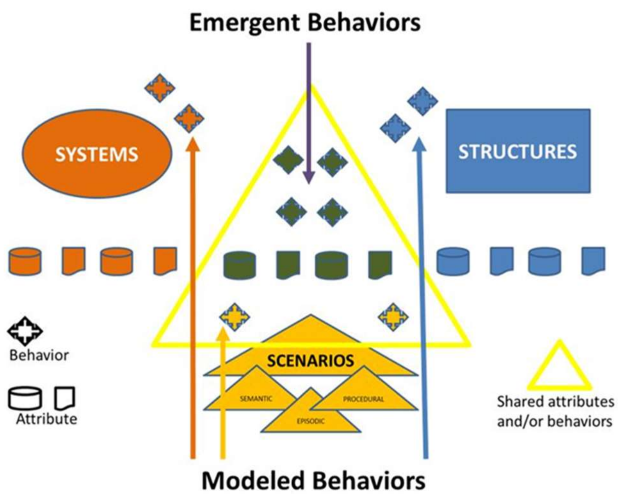
Method categories:

- **Real-world complexity modeling methods**
- **Real-world reasoning methods**
- **Real-world learning and adaptation methods**

To support genomic research, we built a biomimetic digital twin ecosystem composed of four twin classes – patient profile, phenotype, gene variant and protein variant. We incorporated the ecosystem into a leading geneticist's advanced genomics experimental protocol, which we believe is the first report of including this methodology in research to understand the complex pathophysiology of disease. The results were published in the Journal of Molecular Diagnostics. https://www.jmdjournal.org/article/S1525-1578(24)00062-X/fulltext

Biomimetic Digital Twin Model & Ecosystem Design

Models are scoped around known behaviors and designed by imitating (twinning) the understood structures, systems, and scenarios of the modeled behaviors. Emerging behaviors are not predictions, but evidence to be considered by experts.



Ecosystem Architecture

- Each twin models a discrete component of the analytical scope of the ecosystem.
- Internal properties and behaviors must be modeled to a level of sufficient comprehensiveness to enable the reactions that are required for the ecosystem to reflect the real world to the scope of its design.
- Each twin can initiate an interaction with others or respond as prompted.
- Mitigation of bias is achieved by:
    - Independent design of each twin
    - Abstract knowledge graphs populated without defining specific problems or events
    - Autonomous interactions between the twins

Roadmap to High Agency – Adaptive Contextualization Engine

All meaning is based on context. Zero can be a number, or the concept of absence. We understand the distinction intuitively, but the question of how the brain handles the challenging complexities is just beginning to be researched combining mathematics theory, biology and neuroscience. [How the Human Brain Contends With the Strangeness of Zero | Quanta Magazine]

We tend to see data as real-wold input, but the reality is that the interpretation of data is limited by:
1-A deterministically engineered system that can only see what it is designed to see
2-The cleansing and normalization of the inputs into those systems

To enable contextual interpretation of real-world relationships and interactions we need to model the ecosystems using a CONTEXT ENGINE.

What is a Context Engine?

An ecosystem of computing models designed using principles and methods discerned from the human brain. The primary applications of the engine are:

- Practical modeling of highly complex, multi-disciplinary domains to enable holistic discovery and analytics to find realities that are invisible to current ML/NLP, e.g., multi-factor commonality, dark data (Gartner), latent variables
- Chaotic and stochastic simulation that is traceable and repeatable
- Capture elements of human expertise as a computable asset that can be leveraged by standard architectures via an API

Approaches used to build and continually evolve the engine

In the brain, one neuron firing can trigger another neuron firing, which in turn can trigger a third, and so forth, and the message sent is an attenuated stimulation or suppression. The significance of a neuron is all about its connections to other neurons, and the processes are not algorithms but rather dynamic combinatorial interactions. In a context engine:

- Fundamental components and potential connections are defined in a metaontology as structures, systems and known scenarios
- Attributes of the components are defined as categories that are mapped across relational, hierarchical, and unstructured data sources
- Behaviors of the components are defined as expressions that include static and/or dynamic variables and operators

The human brain receives enormous amounts of sensory input. Rather than trying to process it all equally, it quickly filters out data likely to be irrelevant and processes the relevant in great depth. In a context engine the scenarios drive relevance computation, and:

- Are composed of semantic, procedural and/or episodic elements
- Can be situated across the behaviors of systems and structures (scenarios share functional properties with the systems and structures)
- May be emergent functional properties of the systems and structures (scenarios share attributes with the systems and structures)

When humans solve a problem, we rarely invent methods from scratch based solely on our own experience. We usually apply methods learned through education, or copied from others, and often take input from peers and specialists. In a context engine the model content is integrated and rationalized by a combination of:

- Automated and manual processes
- SME input

This real-world reasoning approach enables the construction of models that integrate highly diverse elements and information sources to enable exploration and discovery to a scope that traditional information architecture cannot accommodate.

Systems Thinking and Real-world Reasoning (RWR)

The NAS also recommends addressing complexity using systems thinking. Key observations are:

- Bottom-up, mechanistic, linear approaches to understanding macro-level behavior are limited when considering complex systems.
- Bottom-up, reductionist hypotheses and approaches can lead to a proliferation of parameters; this challenge can potentially be addressed by applying top-down, system-level principles.
- Systems thinking can be used to predict macroscopic phenomena while bypassing the need to explicitly unmask all the quantitative dynamics operating at the microscopic level.

While all knowledge engineering efforts seek to incorporate elements of cognitive science, a key aspect of our innovation strategy is the driving role of a cognitive methodology, which is enabled by biomimetic information architectures. Brain processes are systemic and leverage what neuroscientists label plasticity and sparsity.

- Plasticity is the ability to engage diverse combinations of neurons and synapses by relevance to the purpose of the analysis, and to dynamically adapt internal functional architectures.
- Sparsity is the ability to identify the minimum data required. The brain can respond to situations that are simultaneously new on multiple dimensions and can even categorize one data point.

The neuronal and synaptic architecture of the brain is an ecosystem, which according to the National Academies of Science contains 100 trillion neurons. Systemic architecture, plasticity and sparsity are core to biological learning but are NOT like ML algorithms. The biomimetic technologies that enable elements of RWR are:

- Expertise Graphs
- Neural System Dynamics Digital Twins

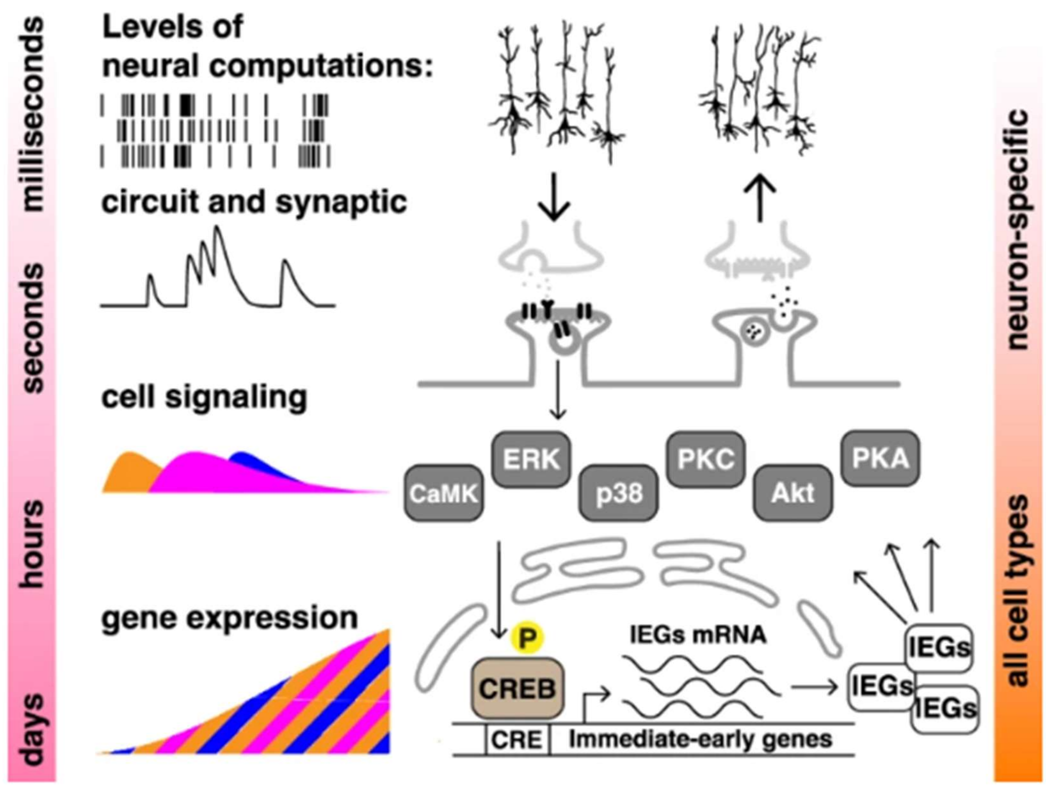
We can imitate principles of plasticity and sparsity by implementing qualitative expertise graphs and leveraging them for contextual selection of data and methods from the in-memory model library. Unlike the deterministic methods to which traditional application engineering is limited of necessity, systemic modeling requires the coexistence of chaotic and stochastic model elements, as well as their ability to dynamically interact with the deterministic elements.

Identify the Relevant Context Scope

Learning and memory are interdependent. Analysis of this relationship lead to the discovery that non-brain cells help with memory. [The massed-spaced learning effect in non-neural human cells | Nature Communications]()

The article explains: Learning and memory in animals exhibit a peculiar feature known as the massed-spaced effect: training distributed across multiple sessions (spaced training) produces stronger memory than the same amount of training applied in a single episode (massed training). This effect is highly conserved across the animal kingdom and is observed at both behavioral and synaptic levels.

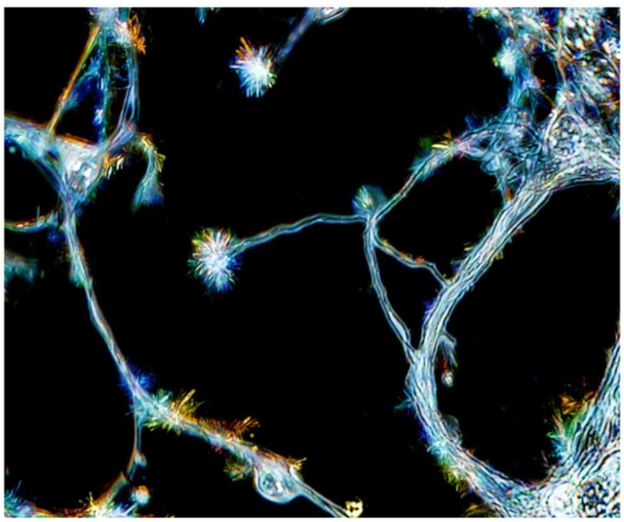
The spacing effect is typically associated with neural systems, but it enables the exploration of “cellular cognition” beyond neural systems.



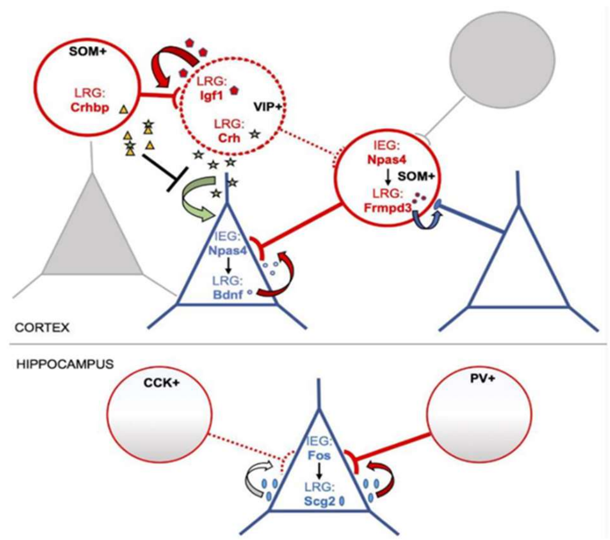
Computations are generally taken to occur at the circuit and synaptic levels of neuronal function on the timescales of milliseconds to seconds. These, however, are nested within slower, cellular computations that occur on the levels of cell signaling (seconds to hours) and gene transcription (hours to days and beyond)—left column. Although both the inputs and outputs of this system are neuron-specific, the conversion of transient signals into more sustained CRE-dependent transcription can be studied in non-neuronal cells—right column.

Key Context Modeling Method - Inhibition

Our brains contextualize using neuronal ecosystems. The image model depicts examples of activity-induced genes which modulate connectivity and/or excitability of distinct neuronal subtypes, overall enhancing the inhibitory tone onto activated pyramidal neurons. Excitatory and inhibitory neurons of the activated neuronal ensemble are depicted as blue triangles and red circles. (Molecular Neuroscience - Roles and Transcriptional Responses of Inhibitory Neurons in Learning and Memory - doi: 10.3389/fnmol.2021.689952)



Neurons in the hippocampus help to pick out patterns in the flood of information pouring through the brain. Credit: Arthur Chien/Science Photo Library



The model depicts examples of activity-induced LRGs which modulate connectivity and/or excitability of distinct neuronal subtypes, overall enhancing the inhibitory tone onto activated pyramidal neurons. Excitatory (blue triangles) and inhibitory (red circles) neurons of the activated neuronal ensemble are depicted. NIH PMC8239217

BIOMIMETIC CONTEXT ENGINE

To enable contextualization in real-world models we can imitate neurological findings:

1-Higher cognitive functions such as learning, memory and processing of sensory perceptions, are strictly dependent on the correct flow of information within neuronal circuits made of both inhibitory and excitatory neurons.

2-The proper balance between excitation and inhibition is particularly important for the correct execution of these brain functions.

3-Subtype-specific activity-dependent transcriptional programs shape circuit rearrangements in response to experience.

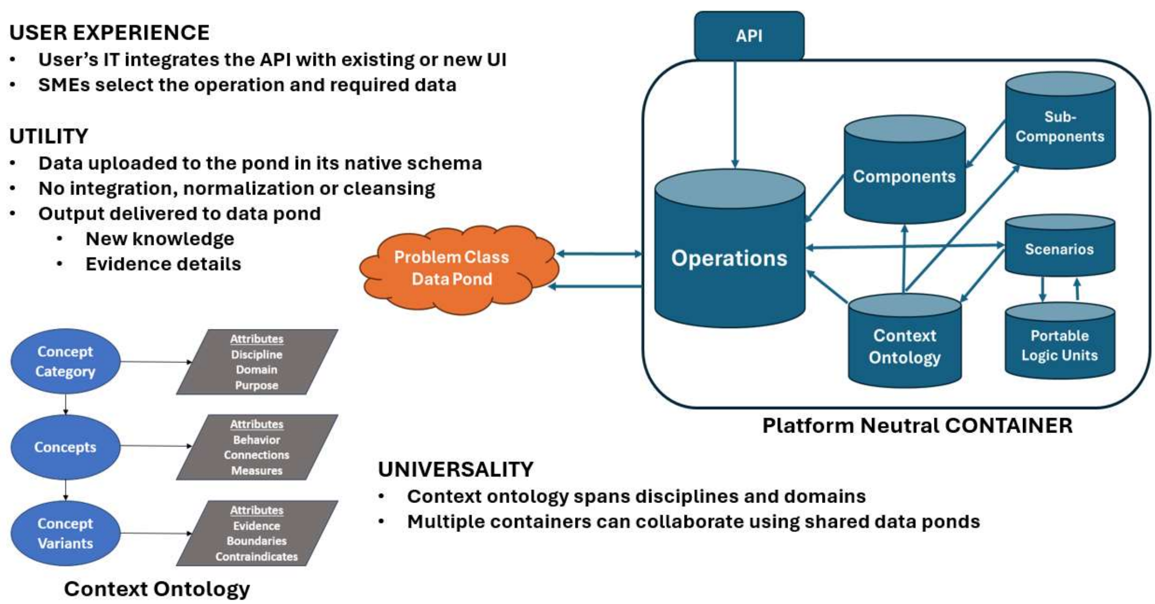
4-Memories are encoded by sparse ensembles of neurons called engrams, activated during memory encoding and reactivated upon recall. An engram consists of a network of cells that undergo long-lasting modifications of their transcriptional programs and connectivity.

5-Neurons forming an engram ensemble are connected by strengthened synapses, providing a conceptual link with the synaptic plasticity paradigm for learning and memory.

6-Memory-driven behavior is represented by the experimental phenomena of synaptic plasticity, characterized by enduring changes of synaptic strength such as long-term potentiation or depression.
7-Encoding of information relies on the synaptic strengthening of both existing and newly formed neuronal connections, along with increased excitability of ensemble neurons.

MODELING METHODS (corresponding to above findings)
1-Abstract ontology architecture to scope relevant knowledge and interactions
2-Attenuation attributes in the ontology
3-Component and subcomponents behaviors and interactions modeled in the ontology
4-Minimum information requirements for component and subcomponent behaviors
5-Abstract modeling of inter/intra-component relationships
6-Contextual relevance is established and strengthened by experimentation processes
7-Contextual relevance is computed leveraging existing and new information

RKE Biomimetic Digital Twin Ecosystems can fill the gap:



- The components and sub-components act as independent agents
- The Context Ontology captures multidomain expertise from diverse SMEs
- The relationships between agents are modeled based on real-world interactions

This methodology and architecture form a new paradigm that can fill the gap between the current Low Agency of LLMs and the High Agency of envisioned future applications – today!

Clearly this biomimetic roadmap is significantly different from traditional technology and methods. It requires an early adopter mindset but enables developing high agency solutions.

Author contact: jglick@rylti.com