# sci.AI Methodology [1]

## Overview

Goal of sci.AI is to support researcher in synthesis of a new knowledge out of disjoint data points. It is achieved by presenting complementary facts from distributed sources together. Result can be produced on different levels of granularity, for example, epidemiological or molecular. And different levels can complement and validate each other. When there is a blank space in molecular detalization, high level observation might suggest direction of research.

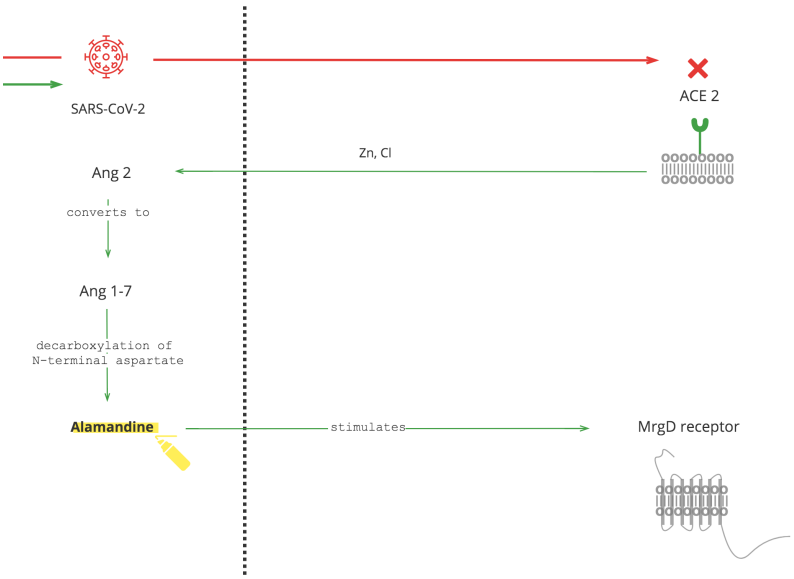


Fig.1. Example of an output produced by researcher with the help of sci.AI. Ultimate goal is to visualize results with such level of sophistication in fully automated mode.

1 Sketched by Roman Gurinovich. E-mail: roman@sci.ai

## High-level Architecture

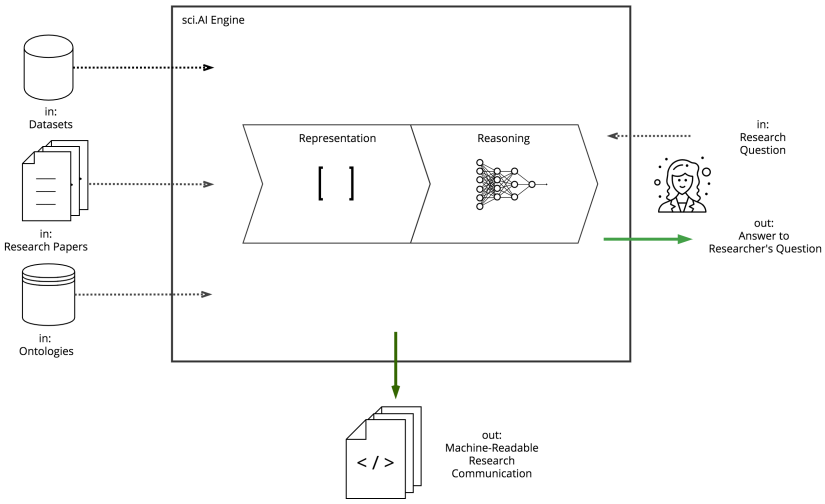


Fig.2. High-level view of analytical components. Data extraction, conversion, backend and frontend are out of focus of the current overview.

## How it works

Analytical part of the pipeline can be seen as two stages: Representation and Reasoning.

Representation algorithms translate unstructured individual papers, documents and files from heterogenous sources into machine-readable formats. This way documents format becomes suitable for computational operations at the reasoning stage. For instance, scientific papers are projected from natural language space into vector and semantic spaces. Representations in vector space can be seen as arrays of numbers reflecting probability distribution over textual sequences. Semantic representation includes recognition of named entities and relations between them with consequent semantic linking. For example, word *"Nfκβ"* is recognized as a member of `protein` class and linked to the concept `uniprot/P19838`. It goes beyond classic NER and finds composite structures like biomarkers and effects with the help of the reasoning mechanism described further, i.e. *"NT-proBNP levels were elevated* `<is_A>` `biomarker` `<correlated_with_condition>` *acute stroke"*. There is an additional positive side-effect of projecting into semantic space: by sending request to sci.AI API, one can semanticize and get any document in a standard machine-readable format like json or JATS xml, so that other systems can utilize extended structured representation for their own purposes.

As the result of Representation stage, all individual documents become embedded into multiple common spaces with recognized relations between them, waiting for user's request to find hidden connections and discover new knowledge as a whole, otherwise distributed across multiple disjoint sources.

New knowledge is being discovered at the Reasoning stage. Discovery process is triggered by user's textual query. This query is expected to contain particular objects of interest, for example, *"SARS-CoV-2"* and *"ARDS"*. Response consists of grouped lists of pieces of knowledge, like citations from scientific papers. These lists carry whole idea about interconnection between objects in the query. While it is possible to retrieve

all relevant components of knowledge at once, it might be overwhelming for the user to absorb all those results. That is why iterative traversal is preferred. It is done by the means of query >> response, query >> response etc. interactions with specification of the next query by user - domain expert. This way researcher has full control over knowledge formation and gets answer to the “why?” question in comparison to pure read mode.

Reasoner generates results by traversing through the interlinked representations and retrieving subgraph relevant to user’s request. Individual phrases or citations are arranged according to similarity of their context, for example, facts about “microglia” and “macrophage” will be positioned closer to each other.

Reasoner’s results are expected to provide detailed explanation of clinical observations or laboratory experiment. It can be used in hypothesis generation prior to experiment as well.

*{'text': 'Notably, experimental SARS-CoV infections of wild-type mice in vivo resulted in considerably reduced ACE2 expression in the lungs (Fig. 2b) suggesting that reduced ACE2 expression might have a role in SARS-CoV–mediated severe acute lung pathologies.', 'doi': '10.1038/nm1267',},*
*{'text': "Using mass spectrometry we observed that alamandine circulates in human blood and can be formed from angiotensin-(1-7) in the heart.', 'doi': '10.1161/circresaha.113.301077'},*
*{'text': 'These results indicate that microinjection of alamandine into the PVN increases blood pressure and sympathetic outflow via MrgD and the cAMP-PKA pathway.', 'doi': '10.1016/j.peptides.2018.03.014'},*

Example 1. Sample list of citations produced internally and to be returned via API or visualized in UI.

We are still experimenting with visualization and automation of human-machine interactions. As far as we can say, sky is the limit in perfecting representation of the results.

---

**Data Sources**

sci.AI works with textual data at the moment and we plan to extend it to any biomedical data like imaging and sequencing.

Standard data sources are NCBI abstracts and Open Access publications. There is an often question to our team whether OA subset is big enough to figure out substantial connections. Most likely it depends on particular branch of study. We’ve posted several results on COVID-19 recently, where one of them is the pathogenesis https://doi.org/10.6084/m9.figshare.12121575. For these studies we’ve used open subset and publicly available abstracts only. These data sources turned out to be enough to cover whole pathway on molecular level plus validating statement on tissue, organism and population levels https://doi.org/10.6084/m9.figshare.12121389.

While more data never hurts, OA subset is sufficient to train generalized models and test whole functional part of the system. So far, data availability never was a bottleneck. It is either computational capacity, development time or, paradoxically, producing results at a higher pace than their lab and clinical validation.

Also it is possible to analyse custom data source upon request as one-time processing of a given dataset or creating gateway to setup ongoing semanticization and reasoning.