Towards AI Scientists: Critical Partnerships for Future Discoveries

Yolanda Gil
Information Sciences Institute
and Department of Computer Science
Viterbi School of Engineering
University of Southern California

December 1, 2022
Towards AI Scientists

SCIENTIFIC KNOWLEDGE & SKILLS

Crowdsourced vocabularies
Common entities
Software and models
Scientific data repositories
Data analysis processes
Automating discoveries
Provenance
Collaborative methods
Model integration

AI reproducing articles  AI as research assistant  AI as co-author
Will AI Write the Scientific Papers of the Future?

What would need to be represented in order for an AI system to automatically generate a paper that provides an accurate report of its analysis and findings?

Benefits:
- Accurate reporting
- Customizable
- Updates & reuse
- What-ifs
- Comparisons
- Creativity

SCIENTIFIC KNOWLEDGE & SKILLS
- Crowdsourced vocabularies
- Common entities
- Software and models
- Data analysis processes
- Collaborative methods
- Provenance
- Automating discoveries
- Model integration
- Open reproducible publications
- W3C PROV
- OntoSoft
- Scientific data repositories
- Data Science
- SCIENTIFIC KNOWLEDGE & SKILLS

Reproducibility and FAIR Principles with Stanford, OSU, OHSU (DARPA, NIH, NSF)

We characterized the variations among top solutions to the challenge, and designed an abstract method that our AI system could elaborate into any solutions and find the best one while explaining its merits over others.

**AI approach:** Document all resources following best practices of reproducible research, open science and FAIR principles, and digital scholarship.

- **Use of Semantic Workflows to Enhance Transparency and Reproducibility in Clinical Omics.** Zheng, C. L; Ratnakar, V.; Gil, Y.; and McWeeney, S. K *Genome Medicine,* 7(73). 2015.
We reproduced a seminal cancer study and explored systematic alternative tools/sources, finding that 35% of protein identifications are not robust to changing even just one analysis step. Our AI system can run all methods and do comparisons and ensembles.

**AI approach: intelligent workflow system captures general method that can be automatically elaborated into alternative implementations and customized to the data**

- Cancer multi-omics: systematic, continuous analysis
Cognitive Architecture for Hypothesis-Driven Discoveries (ONR)

We characterize types of questions about dynamic systems, and develop methods to analyze time series data

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association</td>
<td>Is X associated with Y?</td>
</tr>
<tr>
<td>Diagnostic</td>
<td>What was observed about X given the last few events?</td>
</tr>
<tr>
<td>Intervention</td>
<td>What would I observe if I change X?</td>
</tr>
<tr>
<td>Counterfactual</td>
<td>What outcome if event E occurred in the past?</td>
</tr>
<tr>
<td>Prefactual</td>
<td>What outcome if event E occurs now?</td>
</tr>
<tr>
<td>Semifactual</td>
<td>Same outcome if event E had occurred instead of F?</td>
</tr>
<tr>
<td>Prediction</td>
<td>What future outcome if event E occurs now?</td>
</tr>
<tr>
<td>Backcasting</td>
<td>What event G will have future outcome, what E,F before G?</td>
</tr>
<tr>
<td>Postdiction</td>
<td>What event E caused F that caused current outcome?</td>
</tr>
<tr>
<td>Prefactual</td>
<td>What outcome if event E occurs now?</td>
</tr>
</tbody>
</table>

AI approach: Cognitive framework designed to capture how scientists think about questions to set up computational experiments


Accelerating Discoveries through AI Automation with USC LONI/ENIGMA (NIH)

We automate computational experiments, updating findings when new data becomes available in repository

AI approach: For a type of question pattern, develop lines of inquiry that express what data to retrieve, select method based on data available, and analyze findings

- Neuroscience: automating analysis of large data repositories
  - **Towards Automated Hypothesis Testing in Neuroscience.** Garijo, D.; Fakhraei, S.; Ratnakar, V.; Yang, Q.; Endrias, H.; Ma, Y.; Wang, R.; Bornstein, M.; Bright, J.; Gil, Y.; and Jahanshad, N. In Proceedings of the Fifth Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH), held in conjunction with the 45th International Conference on Very Large Data Bases (VLDB), 2019.
Towards AI Scientists

**SCIENTIFIC KNOWLEDGE & SKILLS**

- Crowdsourced vocabularies
- Common entities
- Software and models
- Provenance
- Collaborative methods
- Model integration
- Data analysis processes
- Automating discoveries
- Scientific data repositories
- Open reproducible publications

- W3C PROV
- OntoSoft
- WINGS
- diskit
- Mint

**AI as co-author**

- AI reproducing articles
- AI as research assistant
- AI as co-author
AI Scientists: The Next Two Decades

2030: AI can generate and test sophisticated hypotheses about complex physical phenomena.

2025: AI can generate automatically new complex scientific analyses using open data.

2025: AI detects when it is missing knowledge and can seek and read new scientific papers on target topics.

2030: AI can reproduce the results in 80% of the articles in a scientific journal.

2035: AI can design a scientific experiment and discuss sophisticated aspects of it.

2035: AI can compare scientific experiments and papers and contrast their merits.

2040: AI can teach advanced theories in some scientific domain effectively to students.

2040: AI can formulate research questions and generate novel contributions in some scientific domain.
AI Grand Challenge: AI Scientists that are Partners in Scientific Discovery

AAAI Presidential Address, February 2020
https://vimeo.com/400177695

Article
https://doi.org/10.1609/aimag.v42i4.18149

2025

AI reproduces articles
2030

AI as research assistant
2035

AI as co-author
2040

AI as investigator
2045

Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery
Hiroaki Kitano

https://doi.org/10.1609/aimag.v42i4.18149