Creating the Engine for Scientific Discovery: Nobel Turing Challenge as a grand challenge project in AI and Systems Biology

Hiroaki Kitano

Photo by Hiroaki Kitano, 2002
By 2050, develop AI systems that can make major scientific discoveries that worth Nobel Prize in Physiology and Medicine.
Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery

Hirotaka Kihara

W hat is the single most significant capability that artificial intelligence can deliver? What pushes the human race forward? Our civilization has advanced largely by scientific discoveries and the application of such knowledge. Therefore, I propose the launch of a grand challenge to develop AI systems that can make significant scientific discoveries. As a field with great potential social impacts, and one that relies on human intelligence, I believe that the initial focus of this challenge should be on biomedical sciences, but it can be applied to other areas later. The challenge is “to develop an AI system that can make major scientific discoveries in biomedical sciences and that is worthy of a Nobel Prize and far beyond.” While recent progress in high-throughput "omics" measurement technologies has enabled us to generate vast quantities of data, scientific discoveries themselves still depend heavily upon individual intuition, which are often overwhelmed by the sheer amount of data. AI, as the complement of the biological phenomena they are seeking to understand. Even now, scientific discovery remains something akin to a cottage industry, but a great transformation seems to have begun. This is an ideal domain and the ideal timing for AI to make a difference. I anticipate that, in the near future, AI systems will make a success of discoveries that have immediate medical implications, saving millions of lives, and totally changing the face of the human race.

Keywords: systems biology, scientific discovery, grand challenge.
The said interest shall be divided into five equal parts, which shall be apportioned as follows: one part to the person who shall have made the most important discovery or invention within the field of physics; one part to the person who shall have made the most important chemical discovery or improvement; one part to the person who shall have made the most important discovery within the domain of physiology or medicine;
The Nobel Committee to give AI system the Nobel Prize without noticing it is an AI system, not a human scientist.

Nobel Turing Challenge

The Turing Test at the Nobel-quality scientific activities

Alfred Nobel

Alan Turing
Bitcoin: A Peer-to-Peer Electronic Cash System

Satoshi Nakamoto
satoshin@gmx.com
www.bitcoin.org

Abstract. A purely peer-to-peer version of electronic cash would allow online payments to be sent directly from one party to another without going through a financial institution. Digital signatures provide part of the solution, but the main benefits are lost if a trusted third party is still required to prevent double-spending. We propose a solution to the double-spending problem using a peer-to-peer network. The network timestamps transactions by hashing them into an ongoing chain of hash-based proof-of-work, forming a record that cannot be changed without redoing all that came before it. The network as a whole uses simple economics to encourage nodes to help maintain this record by indicating whether a transaction is valid. These calculations verify the history of Bitcoin and protect the network from double-spending.

The longest chain not only serves as proof of the sequence of events witnessed, but proof that it came from the longest pool of CPU power. As long as a majority of CPU power is controlled by nodes that are not cooperating to attack the network, they'll generate the longest chain and outpace attackers. The network itself requires minimal structure. Messages are broadcast on a best effort basis, and nodes can leave and rejoin the network at will, accepting the longest proof-of-work chain as proof of what happened while they were gone.
Two sub-goals

First, we will need to develop an AI and robotics system that can perform biomedical and biotechnology research fully autonomously that leads to major discoveries.

Second, the machine must be able to make strategic choice of the topic of research, communicate in form of publications and other means to explain the value, methods, reasonings behind the discovery, and their applications and social implications.
Scientific discovery is at pre-industry revolution level
Process of Scientific Discovery

- Serendipity
- By Accident
- Scientific Intuition
Systems Biology is Science for AI or Al-Human Hybrid System
Cognitive Problems in Scientific Discovery

1. Information Horizon Problem
2. Information Gap Problem
3. Phenotyping Inaccuracy Problem
4. Cognitive Bias Problem
5. Minority Report Problem
1.5 Million papers / Year
= 4100 papers / day
Yeast Signaling
Yeast Cell Cycle
A Comprehensive Map of Signaling Pathways of Alzheimer’s Disease
“Mapathon” = Mapping Marathon

Matsuoka, et al., Weaving Knowledge of Biochemical Pathway in collaboration, to appear in Computational Systems Toxicology
• 99% of reports indicate “A activates B”
• 1% of reports indicate “A inhibits B”

(A) Ignore Minority Report?
(B) Examine quality of Minority Report
   (B-1) All reports are from one lab.
   (B-2) Reports are from diverse labs

Kitano, H., AI Magazine, March 2016
Human Cognitive Bias and Limitations of Semantic Mapping

Cognitive Bias in Clinical Reasoning

• Anchoring Bias
• Availability Bias
• Confirmation Bias
• Premature Closure
• Representativeness

Map is not the territory

Alfred Korzybski
Limits of Human Cognition
Power of Computation

Power of AI
Hypothesis Generation and Verification in AlphaGo

Game status evaluation is possible in the game of GO
Human played games in the record

Possible moves based on past human played games

All possible moves on GO
Human played games in the record

Possible moves based on past human played games

Tabula rasa based generation of games

All possible moves on GO
Creating the Engine for Scientific Discovery
DENDRAL (1969)

Joshua Lederberg, Bruce Buchanan & Ed Feigenbaum

APPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR ORGANIC CHEMISTRY
The DENDRAL Project

Robert K. Lindsay
Research Scientist
University of Michigan

Bruce G. Buchanan
Adjunct Professor of Computer Science
Stanford University

Edward A. Feigenbaum
Professor of Computer Science
Stanford University

Joshua Lederberg
President, The Rockefeller University
Formerly, Professor of Genetics
Stanford University

DENDRAL Team

META-DENDRAL

Hypothesis

HEURISTIC-DENDRAL

Evaluation
The Automation of Science

Ross D. King,1* Jem Rowland,1 Stephen G. Oliver,2 Michael Young,3 Wayne Aubrey,1 Emma Byrne,1 Maria Liakata,1 Magdalena Markham,1 Pinar Pir,2 Larisa N. Soldatova,1 Andrew Sparkes,1 Kenneth E. Whelan,2 Amanda Clare1

Towards Robot Scientists for autonomous scientific discovery

Andrew Sparkes1, Wayne Aubrey1, Emma Byrne1, Amanda Clare2, Muhammed N. Khan1, Maria Liakata1, Magdalena Markham1, Jem Rowland1, Larisa N. Soldatova2, Kenneth E. Whelan2, Michael Young1 and Ross D. King1

Figure 2 Adam’s laboratory robotic system. (a) An external view of Adam’s laboratory robotic system, also showing Eve’s on the far right; and (b) a view looking down through the middle of Adam’s robotic system, again with Eve’s beyond.

Figure 1 Hypothesis-driven closed-loop learning

System model and knowledge base

Hypotheses generation

Experiment generation and design

Cycles of automated hypotheses generation and experimentation

Update the system model

Analysis of results by statistics and machine learning

Collection of experimental observations and other meta-data

New knowledge

Figure 1 Hypothesis-driven closed-loop learning: Diagram showing how iterative cycles of hypothesis-driven experimentation allow for the autonomous generation of new scientific knowledge.

Redefining Scientific Discovery

Massive search and verification of hypotheses space

How efficiently can we execute?  
What are science specific constraints?  
What is computational definition of “Serendipity”?  

SEARCH and OPTIMIZATION

Goal: Reprogram Cell to gain Stemness

SEARCH

Search 24 genes from FANTOM DB

24 genes will enable reprogramming

OPTIMIZATION

Leave-one-out experiments

Yamanaka Factors identified

Nobel Prize in Physiology and Medicine 2012

ACIDENT, SEARCH and OPTIMIZATION

Accidental discovery of thin film formation in polyacetylene polymerization process

ACCIDENT

Goal: Polyacetylene thin film formation condition

OPTIMIZATION

Search optimal thin film formation condition

Prof. Alan MacDiarmid
Goal: Conducting polymer

SEARCH & OPTIMIZATION

Prof. Alan Heeger

Conducting polymer thin film

Nobel Prize in Chemistry 2000
Strategic Path

Intelligence/ Autonomy

Current Status

Task Coverage

Nobel Turing Challenge Target Zone

AI Assistant for Scientific Discovery

Connected Research Laboratories
Technology Platform (SBI/SBX)

Domains
- Bio-Medicine
- Bio-Engineering
- Clinical
- Health-care
- Drug Discovery
- Beyond Biology

Custom Analytics and Systems Integration Services

Connectivity/Automation Framework
Garuda

- Data
  - Device
  - Sensor
  - DB
- Analytic Tool A
- Analytic Tool B
- Analytic Tool C
- Visualization
  - Visualizer A
  - Visualizer B
  - Visualizer C

Garuda Platform

AI Framework
Gandhara

- Deep Learning
- Image analysis
- Classification
- Decision analysis
- Machine Learning
- Inference Cloud
- Knowledge Cloud
- Text Mining

Visualizer A
Visualizer B
Visualizer C
Connect to diverse spectrum of sensors and devices

The Garuda Platform provides the capability to build connect digital solutions by integrating off-the-shelf wearables and monitoring devices.
Case Studies

Shimadzu Data Explorer
Gadget receives and processes the metabolomics data (szf) file from the Machine

Shimadzu Data Explorer (Proto)

VizCloud*
Metabolites Information Query
Explore KEGG pathway for KEGG metabolite IDs
Gene Annotator*
Metabolic Pathway Enrichment in Reactome
Load experimental data and visualize
Reactome Pathways

VANTED
Multi-Omics data analysis configuration

- **Metabolome**: Metabolite pool size (LC/MS/MS)
- **Proteome**: Enzyme abundance (nanoLC/MS/MS)
- **Fluxome**: Metabolic flux (GC/MS)

**Data analysis by various Gadgets**

- **ShimadzuMSData Import gadget**
- **DataMapping Gadget**

**Merge the multi-dimensional data for downstream analytics**

- **IDs**
- **Sample names**
- **Meta data (replicates & etc.)**

**Gadgets**

- **SBI Systems Biology Institute**
- **SBX Technologies**
- **Shimadzu Excellence in Science**
- **Osaka University**
Target discovery using Large-Scale Molecular Interaction Map and Machine Learning based Docking Simulation

ERATO & AMED Project: Discovery of host response modulating factors for a novel influenza drug development with application to avian flu

Collaboration with Prof. Kawaoka at the Institute of Medical Science, the University of Tokyo
Case Study

369 Compounds

For proof of concept, here we describe the example of compounds Pepstatin-A (protease inhibitor) and PD98059 (Map Kinase inhibitor) marked by blue and red boxes respectively.

Previous studies have shown that these compounds possess activity against viral replication in human cell cultures (Barber et al., 2002 and Matarrese et al., 2011).

Based on docking score they were clustered in different groups, however, after including network property in the clustering algorithm both Pepstatin-A and PD98059 are clustered together.

Network-based Compound Screening (NCS)

439 host proteins from Influenza Virus lifecycle

369 Compounds

304 proteins, 3,321 edges

Docking-based Compound Clustering

Network-based Compound Clustering

Watanabe et al., 2014

Kun-Yi et al., 2016

ML-based Docking simulation

PPI Network Analysis

147 proteins

systemsDock

Network-based Compound Screening (NCS)
Very Simplified Process of System-Driven Drug Discovery

1. Disease Related Molecular Interactions
2. Multi-Omics Data
3. Target Candidates
4. Target Prioritization
5. Lead Compound Identification and Optimization
6. Pre-Clinical Studies
7. Clinical Studies
Very Simplified Process of System-Driven Drug Discovery

1. Disease Related Molecular Interactions
2. Multi-Omics Data
3. Target Candidates
4. Target Prioritization
5. Lead Compound Identification and Optimization
6. Pre-Clinical Studies
7. Clinical Studies
Literature-driven Approach

Influenza infection and replication network

Matsuoka Y. et al
BMC Syst. Biol. 2013
Large-Scale Knowledge Extraction from Text

Figure 2: Example sentence with NLP event representations extracted.

Figure 3: Phosphorylation reaction.
Pathways as succinct representations of molecular and process interactions
How good is automated curation vs expert curation?

**Text driven network/pathway reconstruction**

- **unstructured text**
  - Regulation modulates the phosphorylation of Akt1

**NLP techniques**

- **Stand-off format**
  - T1 Protein 0 3 YAP
  - T2 Protein 37 41 Akt1
  - T3 Regulation 4 33 modulates
  - T4 Phosphorylation 17 32 phosphorylation
  - E1 Phosphorylation:T4 Theme:T2
  - E2 Regulation:T3 Theme:E3 Cause:T1

**SBML + annotations**

- **Network and graph theory**

**Expert curated map (Gold Standard)**

**mTOR pathway as a case study**

Key Case Studies

Text mining approach for BioMarker Discovery

Biomarker Knowledge Mining Pipeline

- Internal Records
- Paper collection
- Full Text
- Data Corpus
- Text database building
- Word2vec analysis
- Word Vector Base
- Text Knowledge Base
- Potential Biomarkers
- Similarity Network
- Overlap analysis
- Frequency analysis
- Specific terms
- Prioritized list of biomarker molecules
- Related pathways for the phenotypes

Biomarker Galaxy

- Associate biomarkers with diseases

- amyloid
- beta
- brain
- alzheimer
- pain
A domain specific word2vec Model

150,000 Articles on Alzheimer’s Disease

Over 15,000,000 words
Topic Models
Topic Heatmaps (Disease topics naturally group into clusters)
Very Simplified Process of System-Driven Drug Discovery

Disease Related Molecular Interactions

Multi-Omics Data

Target Candidates

Target Prioritization

Clinical Studies

Pre-Clinical Studies

Lead Compound Identification and Optimization
Network controllability Analysis
Network Controllability Analysis

Class-I Controllability Analysis: Full Controllability
(Liu et al., Nature 2011)

FluMap (State transition diagram) → Convert to Bipartite graph → Calculate maximum matching with Hopcroft-Karp algorithm → Reflect results on FluMap

*Identify Critical Nodes Critical Links
Computational Network Analysis
(To identify and prioritize potential therapeutic targets)

Controllability:
• Who is the major hub on the map?
• Interactions/paths between particular molecules?
Computational Network Analyses
(To identify and prioritize potential therapeutic targets)

Controllability:

• Who is the major hub in the map?

• Interactions/paths between particular molecules?
Computational Network Analyses
(To identify and prioritize potential therapeutic targets)

Controllability:
• Who is the major hub in the map?
• Interactions/paths between particular molecules?
Very Simplified Process of System-Driven Drug Discovery

Disease Related Molecular Interactions

Multi-omics Data

Target Candidates

Target Prioritization

Clinical Studies

Pre-Clinical Studies

Lead Compound Identification and Optimization
Evaluating compounds that targets host factors
Very Simplified Process of System-Driven Drug Discovery

Disease Related Molecular Interactions

Multi-Omics Data

Target Candidates

Target Prioritization

Clinical Studies

Pre-Clinical Studies

Lead Compound Identification and Optimization
Docking Simulation

Target protein

Test compound

In-house Database

54,492 Active Compounds

257,403,694 entries

107 Protein targets

66 Compounds

Hsin, Kamiyoshi, Asai @ OIST

Resource

Influenza Virus-Host Interactome Screen as a Platform for Antiviral Drug Development

Watanabe et al., Cell Host & Microbe (2014), http://dx.doi.org/10.1016/j.chom.2014.11.002
Results in Color Matrix

Proteins

Test Compounds

Binding Potential

0.00

0.25

0.50

0.75
FluMap

18% proteins

Prof. Yoshihiro Kawaoka

JAK1: 5.69 pKd

Nucleic Acids Research

systemsDock: a web server for network pharmacology-based prediction and analysis

Yoshihiro Kawaoka and Hiroshi Kitano

April 29, 2018
Structure of Protein Interaction Networks and Their Implications on Drug Design

Takeshi Hase¹, Hiroshi Tanaka²*, Yasuhiro Suzuki³, So Nakagawa⁴, Hiroaki Kitano⁵,⁶,⁷*
Nobody can develop an entire pipeline alone

Open Platform  Open Collaboration

Software for systems biology: from tools to integrated platforms
Samik Ghosh*, Yukiko Matsuoka†, Yoshiyuki Asai*, Kun-Yi Hsin† and Hiroaki Kitano†‡
Abstract | Understanding complex biological systems requires extensive support from software tools. Such tools are needed at each step of a systems biology computational workflow, which typically consists of data handling, network inference, deep curation, etc.

Social engineering for virtual ‘big science’ in systems biology
Hiroaki Kitano, Samik Ghosh & Yukiko Matsuoka
A new type of big science is emerging that involves knowledge integration and collaboration among small sciences. Because open collaboration involves participants with diverse motivations and interests, social dynamics have a critical role in making the project successful. Thus, proper ‘social engineering’ will have greater role in scientific project planning and management in the future.

GARUDA
Technology Platform

- Autonomous Al Scientist
- Integrated Al Module
- Basic Al Modules
- Basic Functional Modules

- Autonomous Al
- Integrated Al
- Task Specific Al
- Connectivity & Automation
Taxila provides automatic context-aware aggregation and search of relevant information, driven by AI-powered mining and analytics engine for driving actionable insights with an intuitive user interface.

To Structured Representations from Unstructured Text

Reconstructing structured pathway representations from unstructured text

Future Reconstruction Challenge.
( Evaluate event extraction algorithms on a real data set )


M. Spranger, S. Palaniappan, and S. Ghosh. Extracting biological pathway models from nlp event representations
BioNLP 2015, pages 42—51. ACL, 2015
Galaxy of concepts from All Alzheimer’s Disease Papers
Galaxy view

Key scientific leaders knowledge graph
Concept Identification
Powered by Taxila

- Identify key concepts in a group of articles
- Understand the hidden relationships in the text w.r.t to the concepts
- Create Context Specific Knowledge bases
Predicting associations from publications

Akujuobi, T., et al., IEEE Transaction of Knowledge & Data Engineering, 2020
doi: 10.1109/TKDE.2020.3017687
Prediction associations - Retrospective study

<table>
<thead>
<tr>
<th>Association</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD-FO</td>
<td>Raynaud’s Disease (RD) – Fish Oil (FO)</td>
</tr>
<tr>
<td>MIG-MG</td>
<td>Migraine Disorder (MIG) – Magnesium (MG)</td>
</tr>
<tr>
<td>ARG-IGF1</td>
<td>Arginine (ARG) – Somatomedin C (IGF1)</td>
</tr>
<tr>
<td>AD-INN</td>
<td>Alzheimer Disease (AD) – Indomethacin (INN)</td>
</tr>
<tr>
<td>SZ-PA2</td>
<td>Schizophrenia (SZ) – Independent Phospholipase 2 (PA2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evaluation test year</th>
<th>Discovery year</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD - FO</td>
<td>0.04</td>
</tr>
<tr>
<td>MIG - MG</td>
<td>0.58</td>
</tr>
<tr>
<td>ARG - IGF1</td>
<td>0.00</td>
</tr>
<tr>
<td>AD - INN</td>
<td>0.74</td>
</tr>
<tr>
<td>SZ - PA2</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Akujuobi, T., et al., IEEE Transaction of Knowledge & Data Engineering, 2020
doi: 10.1109/TKDE.2020.3017687
<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Connected in future graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apnea</td>
<td>BRAT1</td>
<td>Yes</td>
</tr>
<tr>
<td>Diabetes</td>
<td>SelIL</td>
<td>Yes</td>
</tr>
<tr>
<td>Autoimmune diseases</td>
<td>DIPDHAQ§</td>
<td>Yes</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>miR-572</td>
<td>Yes</td>
</tr>
<tr>
<td>Trauma</td>
<td>PACAP</td>
<td>Yes</td>
</tr>
<tr>
<td>Mental retardation</td>
<td>DOCK8</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## 3 hops away

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Shortest Path</th>
<th>Connected in future graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOTCH1</td>
<td>Gcn2</td>
<td>- NOTCH1&lt;br&gt;- Encephalopathy&lt;br&gt;- Sleep deprivation&lt;br&gt;- Gcn2</td>
<td>Yes</td>
</tr>
<tr>
<td>Loss of consciousness</td>
<td>Ceftaroline</td>
<td>- Loss of consciousness&lt;br&gt;- Infections&lt;br&gt;- BAL9141&lt;br&gt;- Ceftaroline</td>
<td>Yes</td>
</tr>
<tr>
<td>Tenuifoliside</td>
<td>WAGR syndrome</td>
<td>- WAGR syndrome&lt;br&gt;- Cognitive difficulties&lt;br&gt;- Acetylcholine&lt;br&gt;- Tenuifoliside</td>
<td>No, but clues found in [48], [49]</td>
</tr>
<tr>
<td>interleukin-15</td>
<td>Antero**</td>
<td>- interleukin-15&lt;br&gt;- Cancer&lt;br&gt;- Dyskinesias&lt;br&gt;- Antero</td>
<td>No, but clues found in [50], [51]</td>
</tr>
<tr>
<td>Macroglobulinemia</td>
<td>HNF-1</td>
<td>- Macroglobulinemia&lt;br&gt;- Injury to cranial nerves&lt;br&gt;- C-Met&lt;br&gt;- HNF-1</td>
<td>No, but found in [52]</td>
</tr>
<tr>
<td>Sezary syndrome</td>
<td>Antibiotic</td>
<td>- Sezary syndrome&lt;br&gt;- Aermatoxes&lt;br&gt;- Amyloid beta peptide(1-42)&lt;br&gt;- Antibiotic</td>
<td>No, but found in [53]</td>
</tr>
</tbody>
</table>

Akujuobi, T., et al., IEEE Transaction of Knowledge & Data Engineering, 2020
doi: 10.1109/TKDE.2020.3017687
## 2 Hops away

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Shortest Path</th>
<th>Connected in future graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinocerebellar ataxia type 1††</td>
<td>c-KIT</td>
<td>- Spinocerebellar ataxia type 1</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- corticobasal degeneration</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- c-KIT</td>
<td></td>
</tr>
<tr>
<td>Triethyltin</td>
<td>Amphiregulin</td>
<td>- Triethyltin</td>
<td>No, but with internet evidence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Calcium</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Amphiregulin</td>
<td></td>
</tr>
<tr>
<td>Transient ischemic attack</td>
<td>Phycoerythrin</td>
<td>- Transient ischemic attack</td>
<td>No, but found in [54]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- P-selectin</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- phycoerythrin</td>
<td></td>
</tr>
<tr>
<td>Kawasaki disease</td>
<td>Aortitis</td>
<td>- Kawasaki disease</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Livedoid vasculitis</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Aortitis</td>
<td></td>
</tr>
<tr>
<td>BBS4</td>
<td>Rtl1/Mart1</td>
<td>- BB54</td>
<td>No, but clues found in [55], [56]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Amino acid</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Rtl1</td>
<td></td>
</tr>
<tr>
<td>Simpson-golabi-behmel syndrome ‡‡</td>
<td>Tetrasomy</td>
<td>- Simpson-golabi-behmel syndrome</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Cancer</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Tetrasomy</td>
<td></td>
</tr>
</tbody>
</table>

Akjuobi, T., et al., IEEE Transaction of Knowledge & Data Engineering, 2020  
doi: 10.1109/TKDE.2020.3017687
Scientific Knowledge in AI system

Hypotheses Generated
- False
- Verified

Hypothesis generation from knowledge

Experiments
- Some hypotheses require experimental verification
- Are newly verified hypotheses consistent with current knowledge, or do they generate inconsistencies?

Entire Hypothetical Body of Scientific Knowledge

Knowledge Extraction
- Data added to database
- Papers and databases contain errors, inconsistencies, and even fabrications

Integrating Dark Data

Papers and Databases

Portions of knowledge believed to be correct may in fact be false

Up-dating knowledge may impact outcome of previous hypothesis

Kitano, H., AI Magazine, March 2016
Structure of Biological Discovery

Explaining mechanisms of biological phenomena

Experimental Constraints
Empirical Constraints

Biochemical Constraints
Systems Constraints

Possible combinations of molecular mechanisms

Molecules involved
Discovered knowledge

Human discoverable knowledge

Scientific truth
Discovered knowledge

Human discoverable knowledge

Knowledge that human may to be able to generate and understand

Scientific truth
Data Acquisition and Generation

Existing Data
+
Data Acquisition
+
Virtual Data Generation
(Populating Hypothesis Space)

Can we develop a proper evaluation system?
Data Acquisition and Generation

- Existing Data
- Data Acquisition
- Virtual Data Generation (Populating Hypothesis Space)

Can we develop a proper evaluation system?

Discard biased data
Turn over to more accurate data
Asking Right Questions

This may be critically important under resource constraint situation. One cannot work on so many issues, so focus on an important issue.

Eliminating a bottleneck

Build up ways to run super-fast hypothesis-testing cycle, so that every question can be answered quickly

“Asking Right Questions” may not be that important anymore

Or

Human sense of “Right Questions” may be suboptimal
Redefining Scientific Discovery

Massive search and verification of hypotheses space

How efficiently can we execute?
What are science specific constraints?
What is computational definition of “Serendipity”?
Is discovery path dependent?

One Universal AI Scientist

A population of characteristic AI Scientists
Process of Scientific Discovery

Serendipity
By Accident
Scientific Intuition
Redefining Scientific Discovery

Cognitive Problems in Scientific Discovery
- Information Horizon Problem
- Information Gap Problem
- Phenotyping Inaccuracy Problem
- Cognitive Bias Problem
- Minority Report Problem

Massive Search of Large Hypotheses Space

AI as an engine for scientific discovery
Implications

- Alternative forms of scientific discovery
- Alternative forms of intelligence
- Accelerating sciences at unprecedented speed
- Machines to evolve by itself?