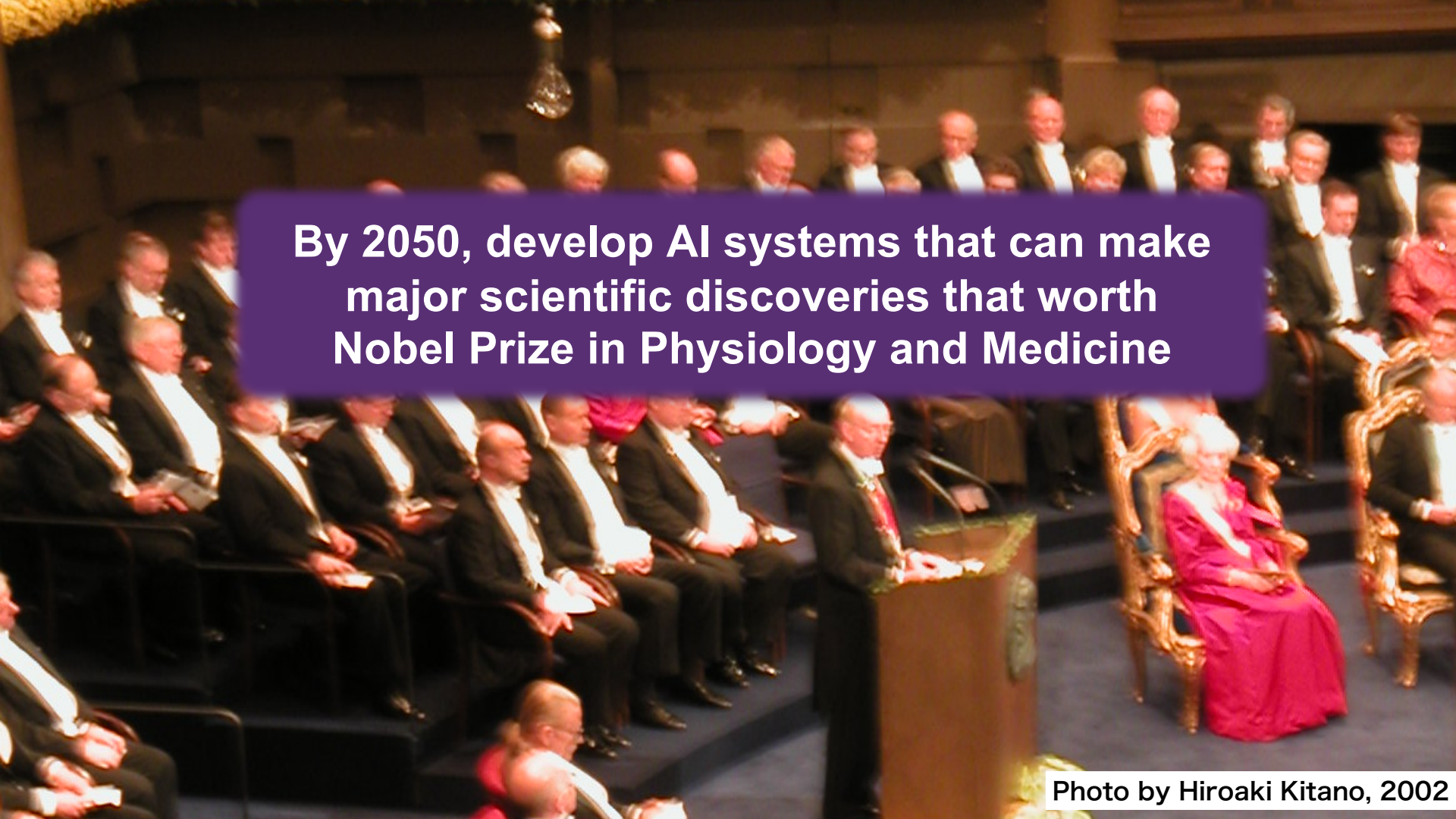
Creating the Engine for Scientific Discovery: Nobel Turing Challenge as a grand challenge project in AI and Systems Biology

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
Photo by Hiroaki Kitano, 2002

AI Magazine, Spring 2016　　人工知能学会誌　２０１６年３月号　　日経エレクトロニクス

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

Hiroaki Kitano

■ *This article proposes a new grand challenge for AI: to develop an AI system that can make major scientific discoveries in biomedical sciences and that is worthy of a Nobel Prize. There are a series of human cognitive limitations that prevent us from making accelerated scientific discoveries, particularity in biomedical sciences. As a result, scientific discoveries are left at the level of a cottage industry. AI systems can transform scientific discoveries into highly efficient practices, thereby enabling us to expand our knowledge in unprecedented ways. Such systems may outcompute all possible hypotheses and may redefine the nature of scientific intuition, hence the scientific discovery process.*

What 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 suffers particularly from information overflow, along with the limitations of human cognition, 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, while researchers are often overwhelmed by the sheer amount of data, as well as by the complexity 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 succession of discoveries that have immediate medical implications, saving millions of lives, and totally changing the fate of the human race.

Article アーティクル

人工知能がノーベル賞を獲る日，そして人類の未来
―究極のグランドチャレンジがもたらすもの―

The Day AI Win the Nobel Prize and the Future of Humanity
— An Ultimate Grand Challenge in AI and Scientific Discovery —

北野　宏明　Hiroaki Kitano
システム・バイオロジー研究機構，株式会社ソニーコンピュータサイエンス研究所，理化学研究所統合生命医科学研究センター，沖縄科学技術大学院大学
The Systems Biology Institute. / Sony Computer Science Laboratories, Inc. / RIKEN Center for Integrative Medical Sciences. / Okinawa Institute of Science and Technology.
kitano@sbi.jp

Keywords: systems biology, scientific discovery, grand challenge.

1. はじめに

人工知能が成し得る最も重要なことは何であろうか？　答えは一つではなく，いくつかのことがある．しかし，大きな科学的発見を行うことができる人工知能システムが構築されるなら，それは，人類の科学的知識を飛躍的に拡大させるという意味で，極めて重要な貢献となるであろう．この領域における研究を飛躍的に加速する手段として，新たなグランドチャレンジを提案したい．このグランドチャレンジは，「2050年までに，ノーベル賞級かそれ以上の科学的発見を行う人工知能を開発する」である[Kitano 16]．特に，生命科学分野は，大量の情報と複雑な対象の理解と制御という問題に直面しており，人工知能によるブレークスルーが期待される．この場合，ノーベル賞としては，医学・生理学賞または化学賞が対象となる．物理学賞の対象となるためには，物理学の新しい理論の発見が必要であり，現在の人工知能技術ではこれは難しい．ノーベル賞は，人類の文化的営みの最高峰を対象とするものであり，人工知能による発見であっても，人類の文化的営みの一部として認められるものでなければならない．

一方，人工知能による発見は，必ずしも人類にとって理解しやすいものである必要はない．人工知能による発見は，人類の理解を超えたものであるかもしれない．しかし，人工知能による発見であっても，人類の文化的営みの一部として認められるものでなければならない．

その後，データマイニングなど大規模データからの知識獲得の研究が，Knowledge Discovery from Database (KDD)として盛んになった．しかし，個別の相関や因果関係の推定に幅広い応用が見いだされたこともあり[Kulkarni 90, Zupan 07]，その背後の法則性を自律的に見いだすという研究には発展していない．

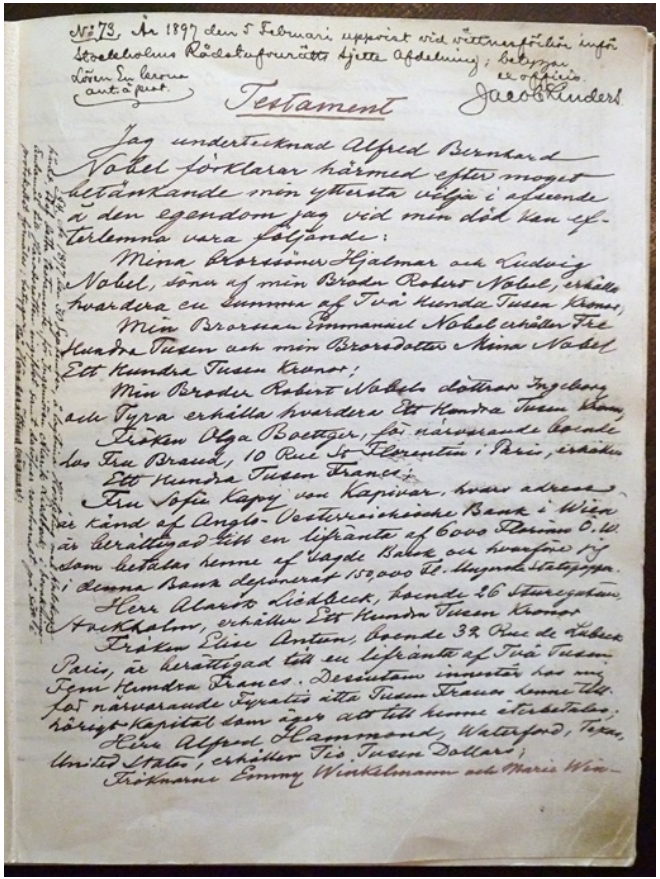
しかし，現在多くの科学分野で大規模データを扱うことが一般化し，膨大な計算を可能とする各種のインフラストラクチャが実現している．この状況の変化は，新たに人工知能による科学的発見という分野に，再度取り組むべき時期に来たと確信させるものである．

1·2　人工知能とグランドチャレンジ

人工知能の発展において，グランドチャレンジは，重要な役目を果たしてきた．コンピュータチェスは，多くにでもわかりやすい表現も時に有用であるので，あえて「ノーベル賞級の発見」という表現を使うことにしたが，「ノーベル賞を獲る」ということ自体が真の目的ではない．繰り返すが，このグランドチャレンジの目的は，科学的発見の理解とそれを自律的に実行するシステムの開発であり，その結果として人類が直面する多くの問題を解決することである．これは，RoboCupが，「2050年までに，FIFAワールドカップのチャンピオンに勝利する完全自律型ヒューマノイドのチームを開発する」[Kitano 97]としているが，真の目的は，その過程で開発された技術を世の中に広く応用していくというものと同じロジックである．同時にこれは，RoboCupとは対極的で，人間の知能の側面にフォーカスしており，RoboCupがカバーしていない領域へのチャレンジとなる．

1·1　科学的発見への再挑戦

Nobel's Will



Testament

Jag undertecknad Alfred Bernhard Nobel förklarar härmed efter moget betänkande min yttersta vilja i afseende å den egendom jag vid min död kan efterlemna vara följande:

Mina Brorssöner Hjalmar och Ludvig Nobel, söner af min Broder Robert Nobel, erhålla hvardera en Summa af Två Hundra Tusen Kronor;

Min Brorson Emmanuel Nobel erhåller Fyra Hundra Tusen och min Brorsdotter Mina Nobel Ett Hundra Tusen Kronor;

Min Broder Robert Nobels döttrar Ingeborg och Tyra erhålla hvardera Ett Hundra Tusen Kronor;

Fröken Olga Boettger, för närvarande boende hos Fru Braud, 10 Rue St. Florentin i Paris, erhåller Ett Hundra Tusen Francs;

Fru Sofie Kapy von Kapivar, hvars adress är känd af Anglo-Oesterreichische Bank i Wien är berättigad till en liffrente af 6000 Florins O. W. som betalas henne af Sagde Bank och hvarför ej i denna Bank deponerat 150,000 Fl. Utgnde Statsskuld;

Herr Alarik Liedbeck, boende 26 Sturegatan, Stockholm, erhåller Ett Hundra Tusen Kronor;

Fröken Elise Antoni, boende 32 Rue de Lübeck Paris, är berättigad till en liffrente af Två Tusen Fem Hundra Francs. Dessutom innestår hos mig för närvarande Fröken ytterligare Fem Tusen Francs hennes tillhörigt Kapital som ägas att tillkomma efterlefvande;

Herr Alfred Hammond, Waterford, Texas, United States, erhåller Tio Tusen Dollars;

Fröknarne Emmy Winkelmann och Marie Win-

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;

Nobel Turing Challenge

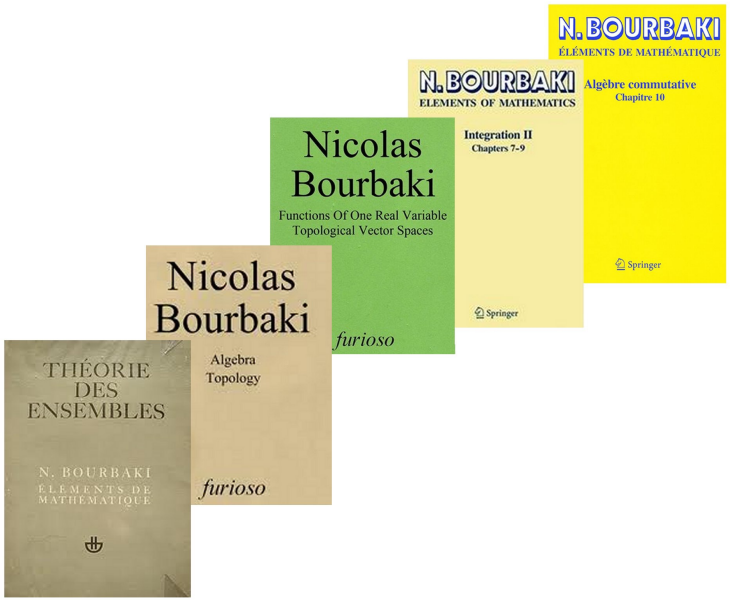
The Nobel Committee to give AI system the Nobel Prize without noticing it is an AI system, not a human scientist.

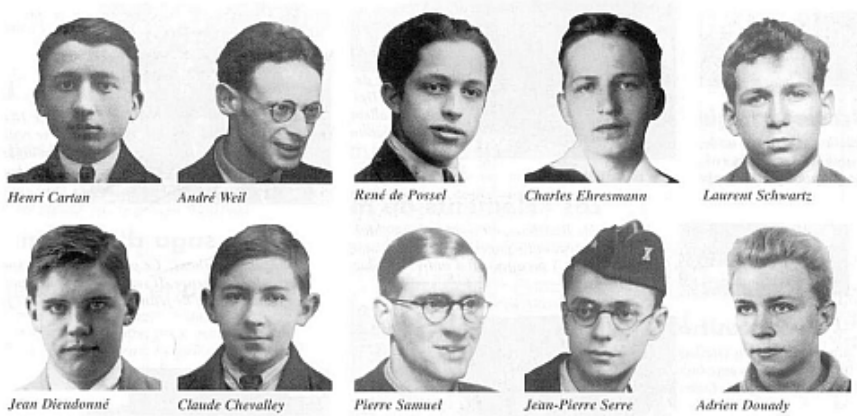
Alfred Nobel

The Turing Test at the Nobel-quality scientific activities

Alan Turing

Nicolas Bourbaki





Satoshi Nakamoto

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 the proof-of-work. The longest chain not only serves as proof of the sequence of events witnessed, but proof that it came from the largest 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.

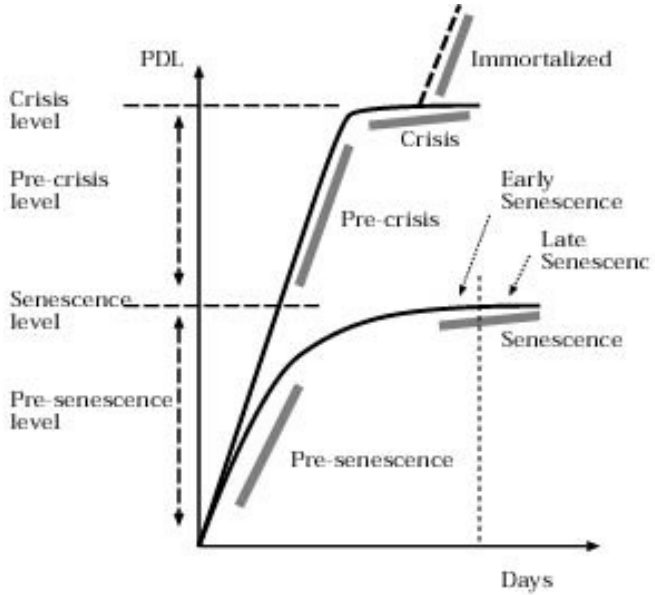
Two sub-goals

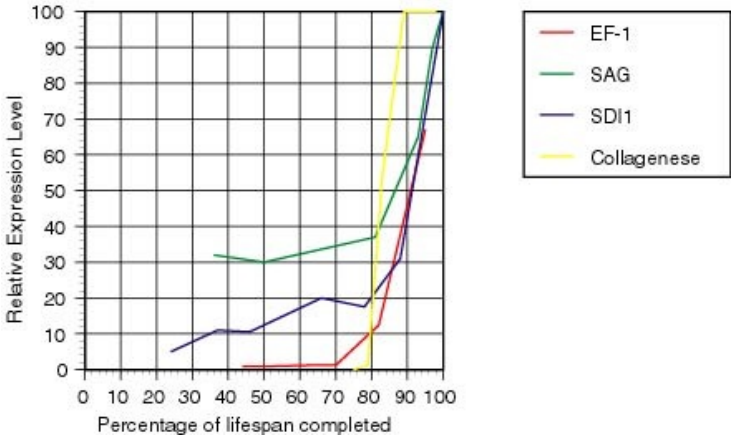
First, we will need to develop an AI and robotics system that can perform biomedical and biotechnology research fully autonomously that leads to major discoveries.

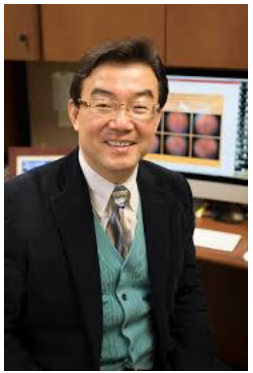
Second, the machine must be able to make strategic choice of the topic of research, communicate in form of publications and other means to explain the value, methods, reasonings behind the discovery, and their applications and social implications.

An Introspective Review of Scientific Discovery in Biomedical Sciences

Case Study: Prediction of Aging Mechanisms 1994 - 1998







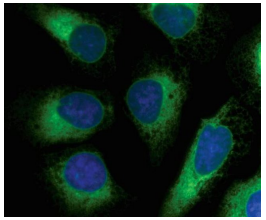
Prof. Shin-ichirou Imai
Washington Univ. St. Louis

Kitano and Imai, *J. Exp. Gerontology*, 1998

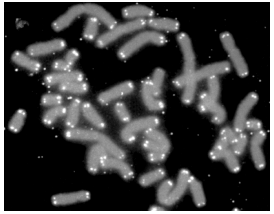
Generating Hypotheses on Mechanisms

Molecular Mechanisms

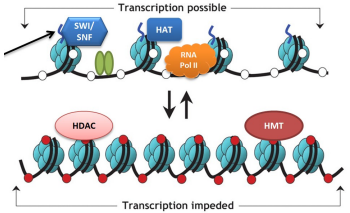
Accumulation of molecules



Telomere Shortening



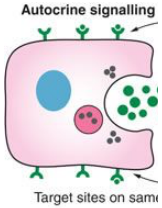
Heterochromatin Re-modelling



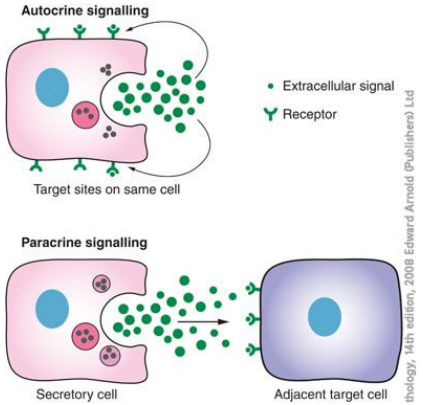
Random Mutations

Cellular Mechanisms

Cell autonomous



Autocrine

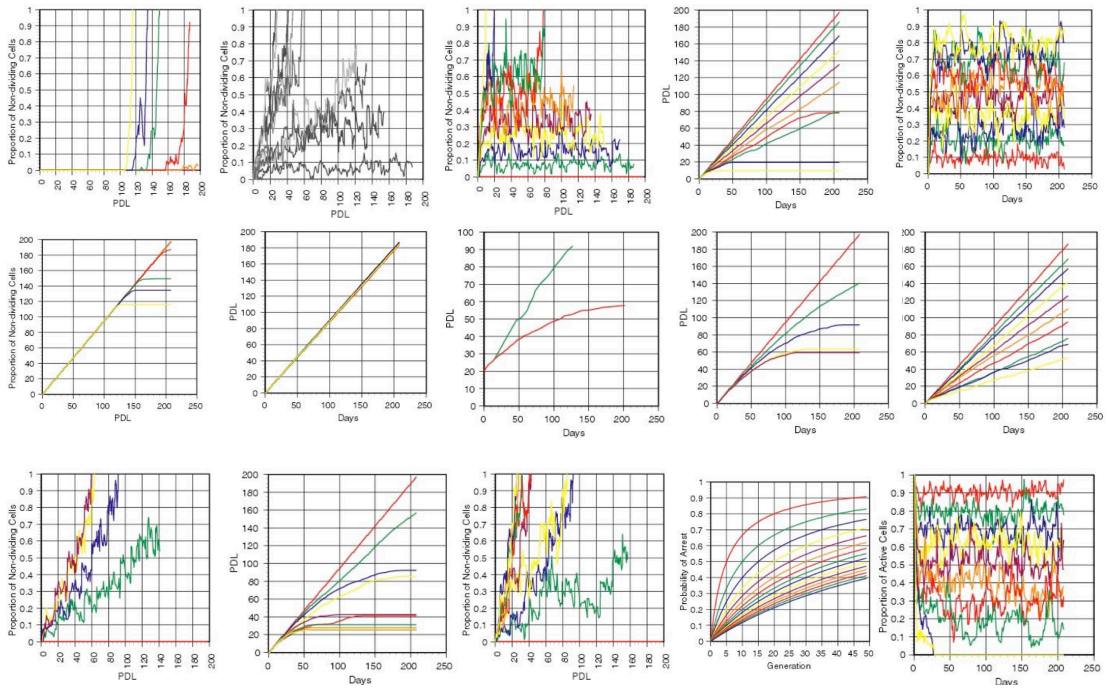


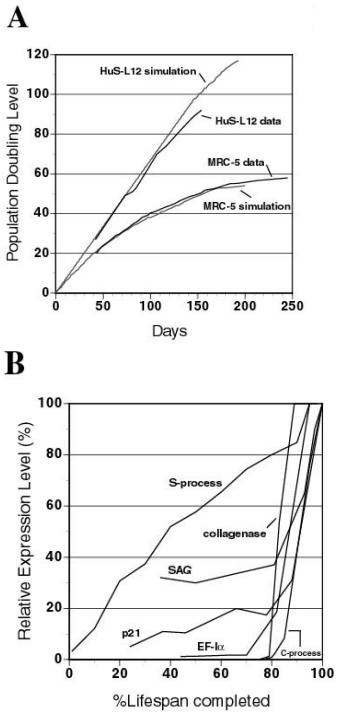
Paracrine

Identification of a model that explains the data set

Exhaust Search of cellular aging mechanisms using computer simulation

Over 500,000 simulation runs

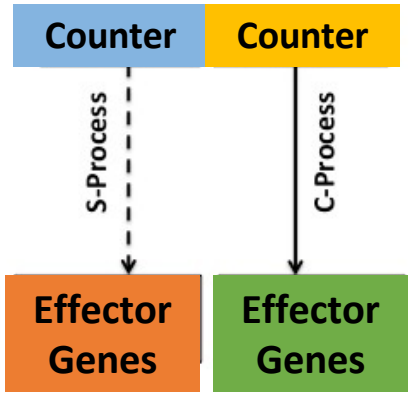


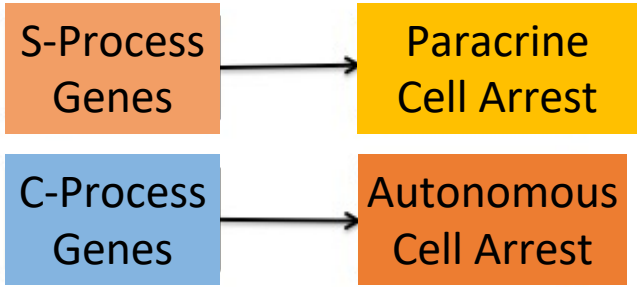


Kitano and Imai, *J. Exp. Gerontology*, 1998

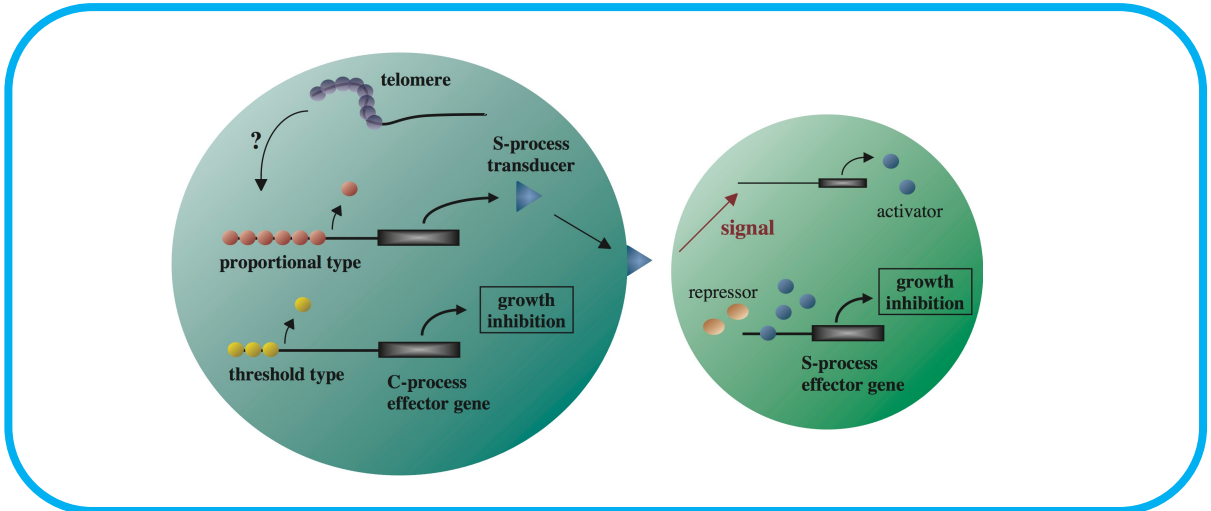
The only model that explained the data set

Math

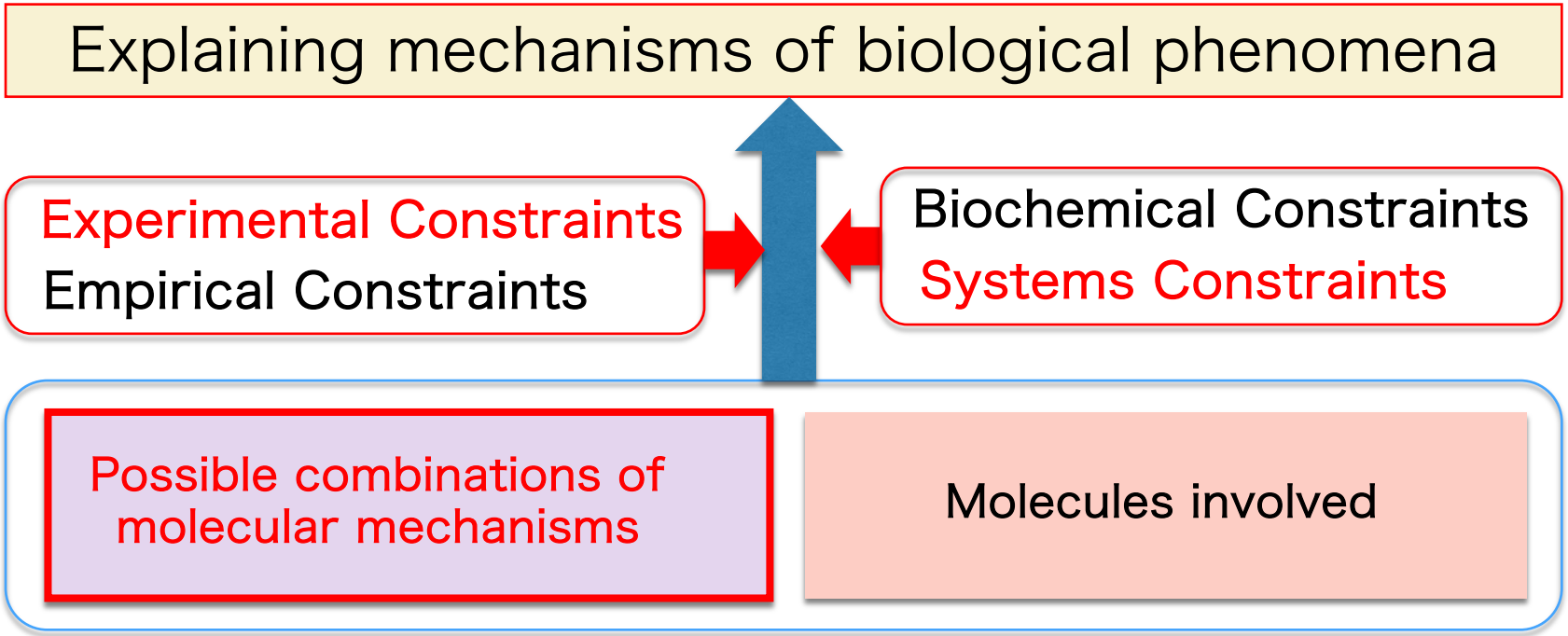




Biology



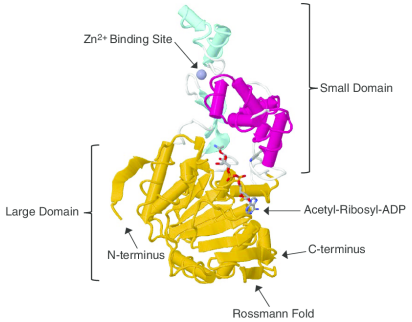
Structure of Biological Discovery



Inferring Sir2 to be the critical factor for aging and predicting the role of Sir2 in aging and verifying it

1. Implicated in Yeast aging
2. Histone Modifier
3. Conserved among species

Budding Yeast Sir2



Kitano, H. and Imai, S., "The two-process model of cellular aging," *Exp. Gerontol*, Aug; 33(5):393-419, 1998

Imai, S. and Kitano, H., "Heterochromatin islands and their dynamic reorganization: a hypothesis for three distinctive features of cellular aging," *Exp. Gerontol,* Sep; 33(6):555-70. 1998

Experimentally verified

Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase

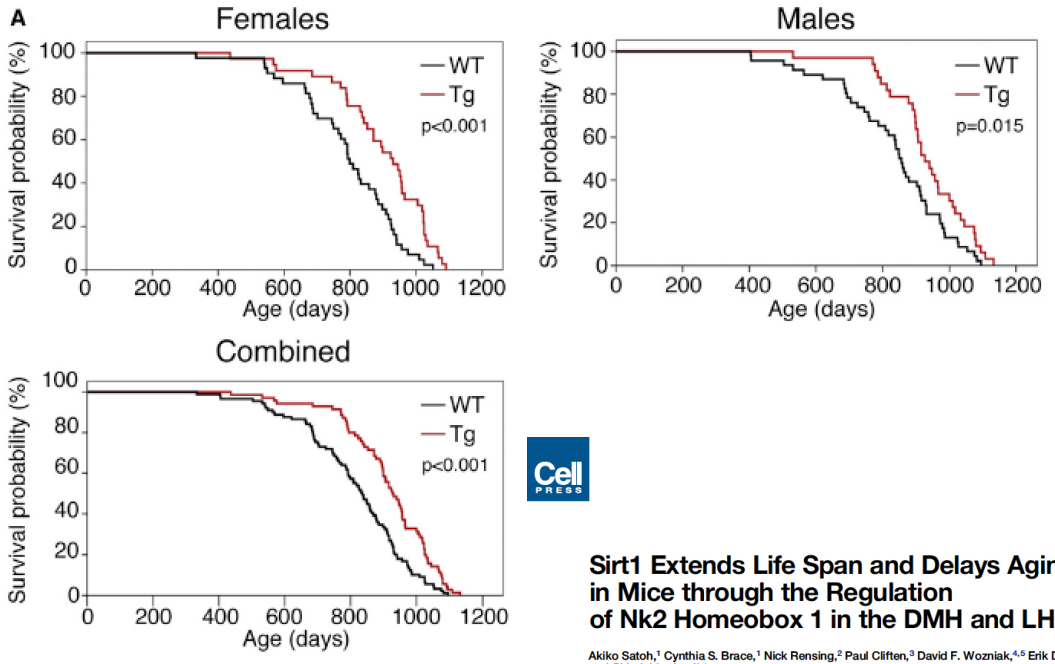
Shin-ichiro Imai, Christopher M. Armstrong, Matt Kaeberlein & Leonard Guarente

Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Imai, S., et al., *Nature* 403(6771): 795-800, 2000

Sirtuins Family

Life-span extension by SIRT1 over-expressing mouse

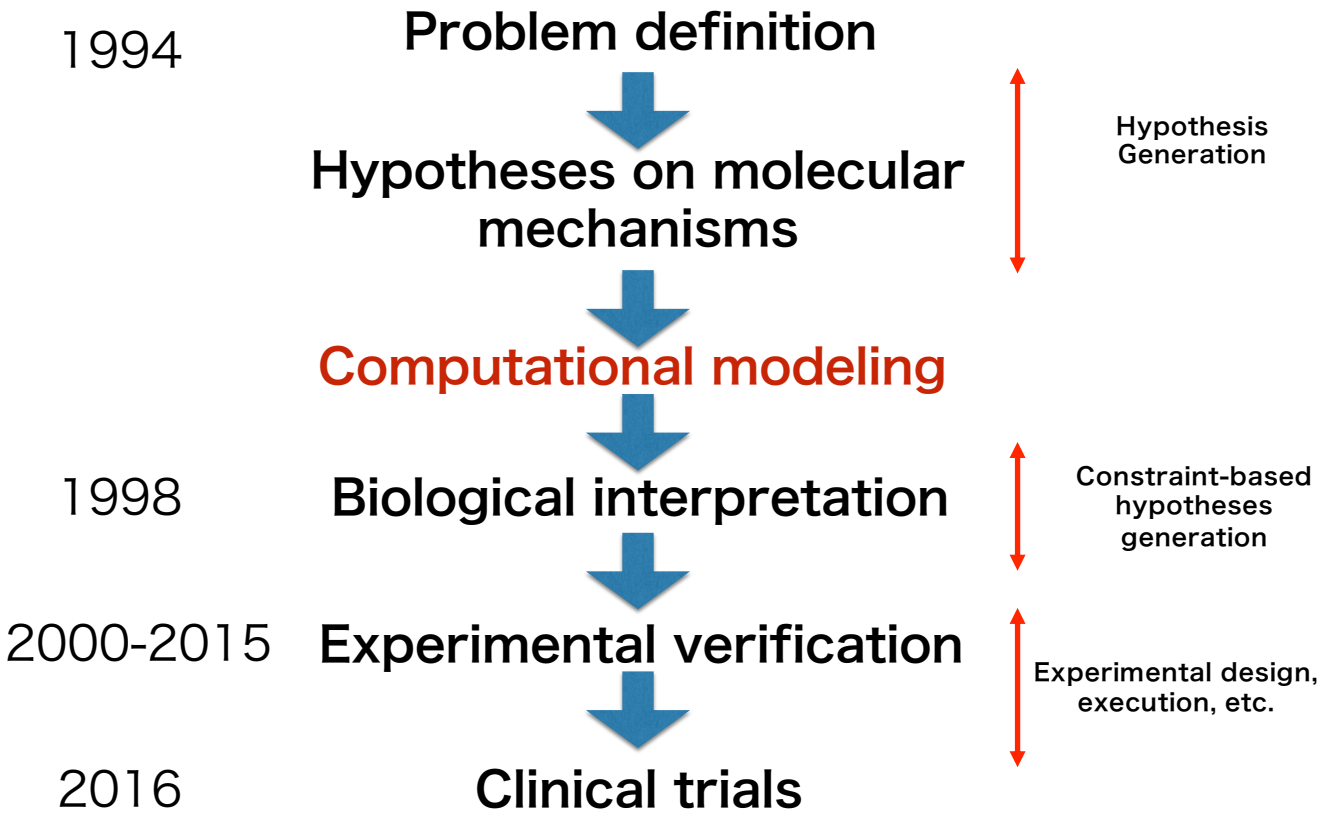


Cell Metabolism
Article

Sirt1 Extends Life Span and Delays Aging in Mice through the Regulation of Nk2 Homeobox 1 in the DMH and LH

Akiko Satoh,[1] Cynthia S. Brace,[1] Nick Rensing,[2] Paul Clifften,[3] David F. Wozniak,[4,5] Erik D. Herzog,[5,6] Kelvin A. Yamada,[2] and Shin-ichiro Imai[1,*]

The process of discovery




Scientific discovery is at pre-industry revolution level
Photo by Hiroaki Kitano, 2010

Process of Scientific Discovery

Serendipity

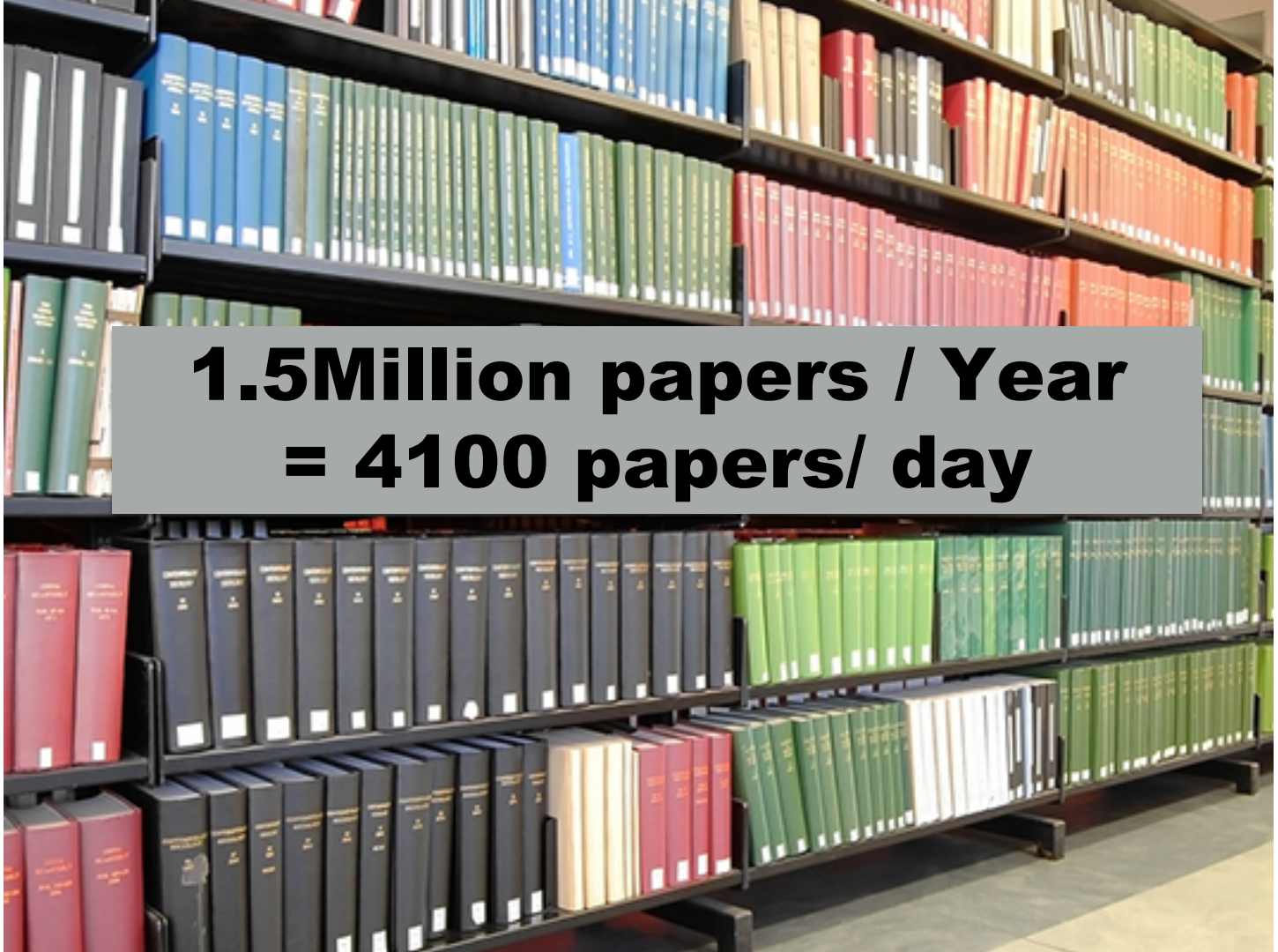
By Accident

Scientific Intuition

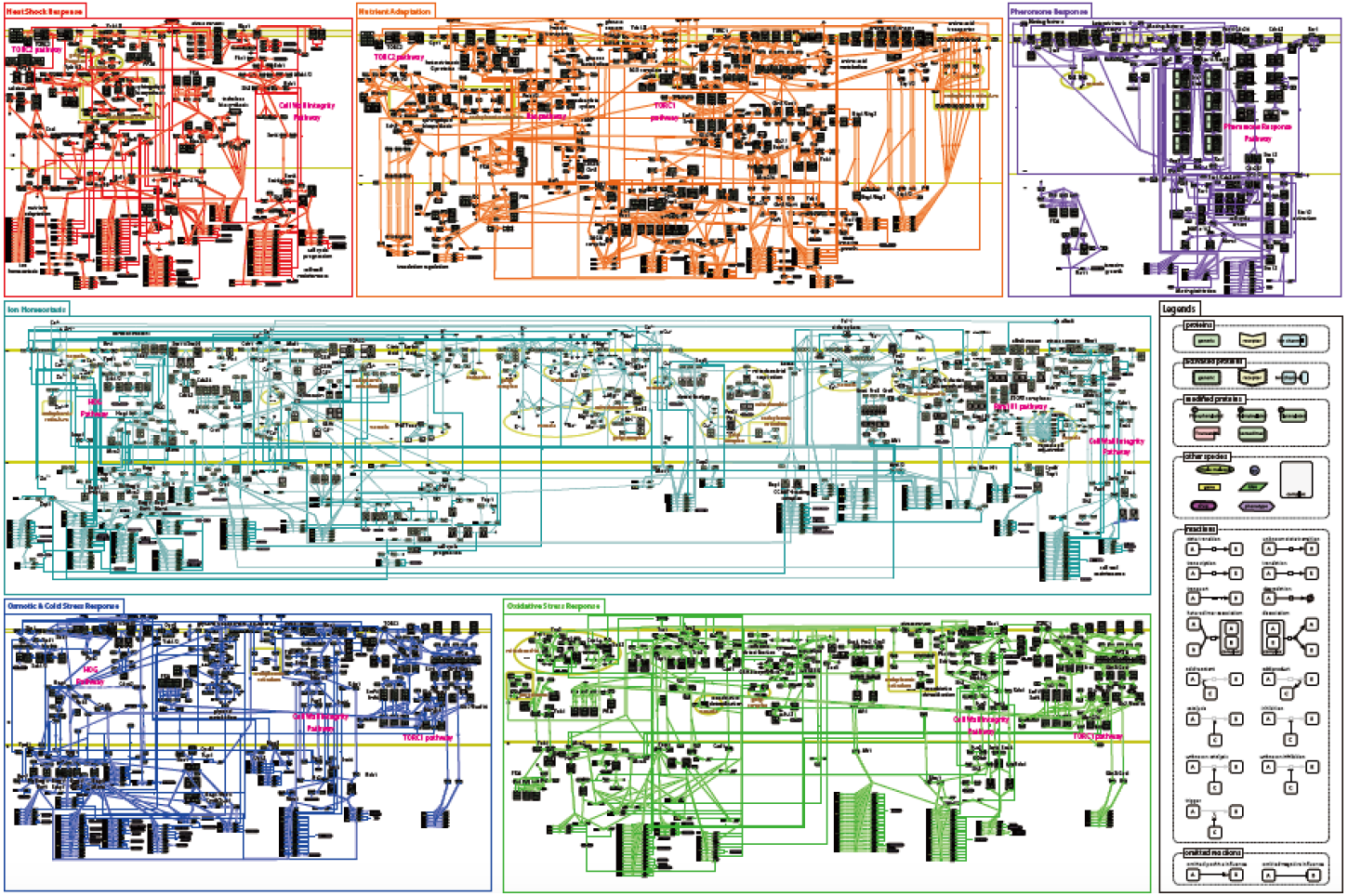
Systems Biology is Science for AI or AI-Human Hybrid System

Cognitive Problems in Scientific Discovery

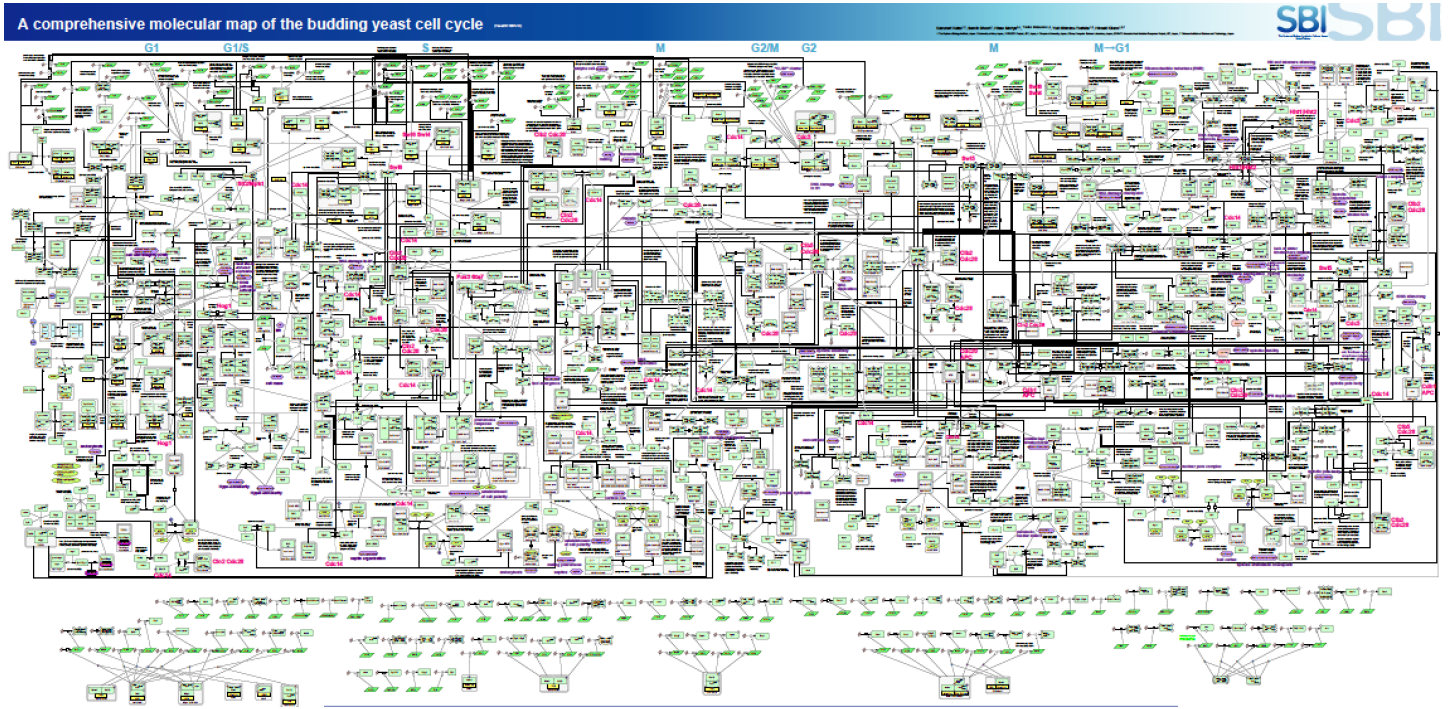
1. Information Horizon Problem
2. Information Gap Problem
3. Phenotyping Inaccuracy Problem
4. Cognitive Bias Problem
5. Minority Report Problem

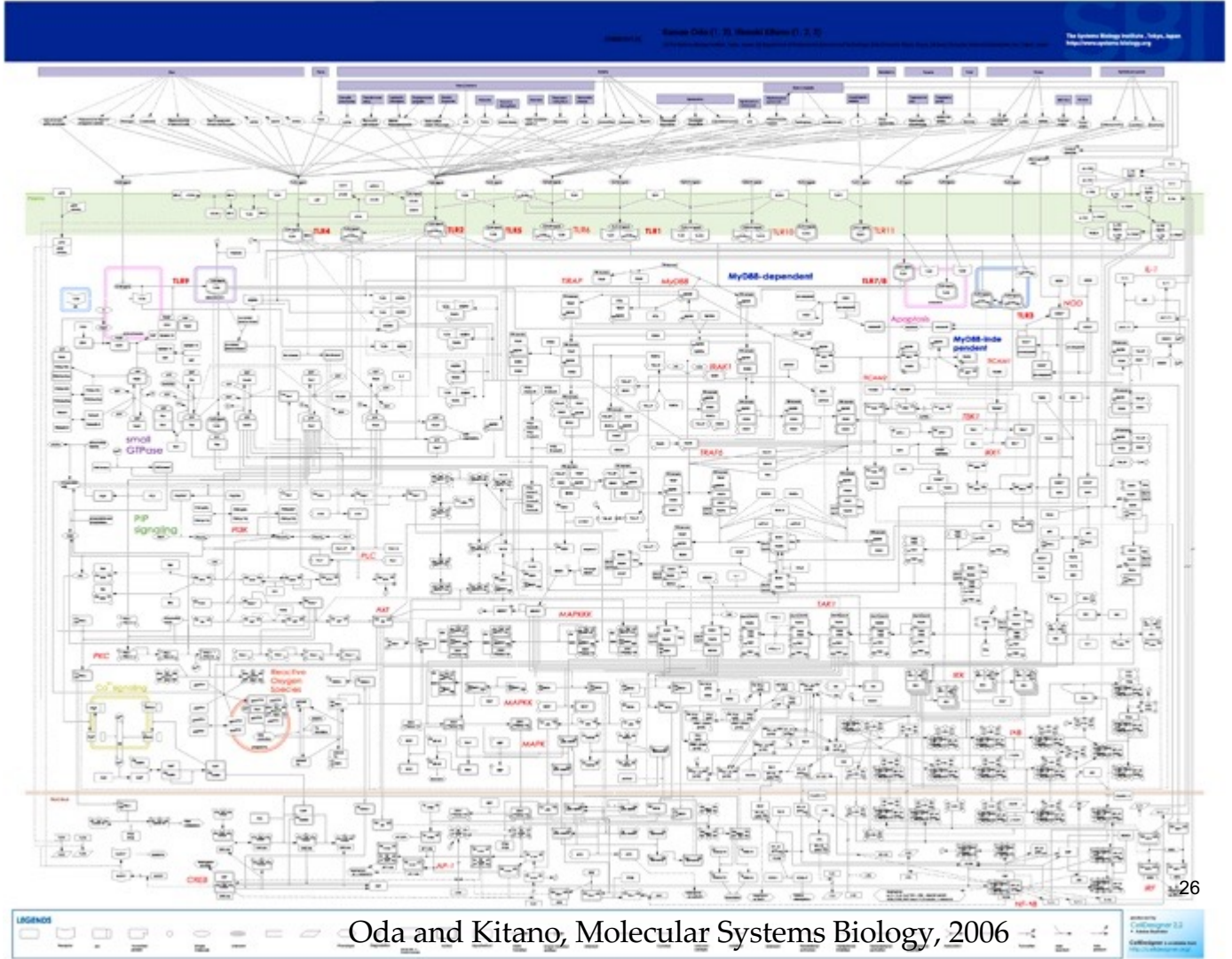

1.5Million papers / Year
= 4100 papers/ day

Yeast Signaling

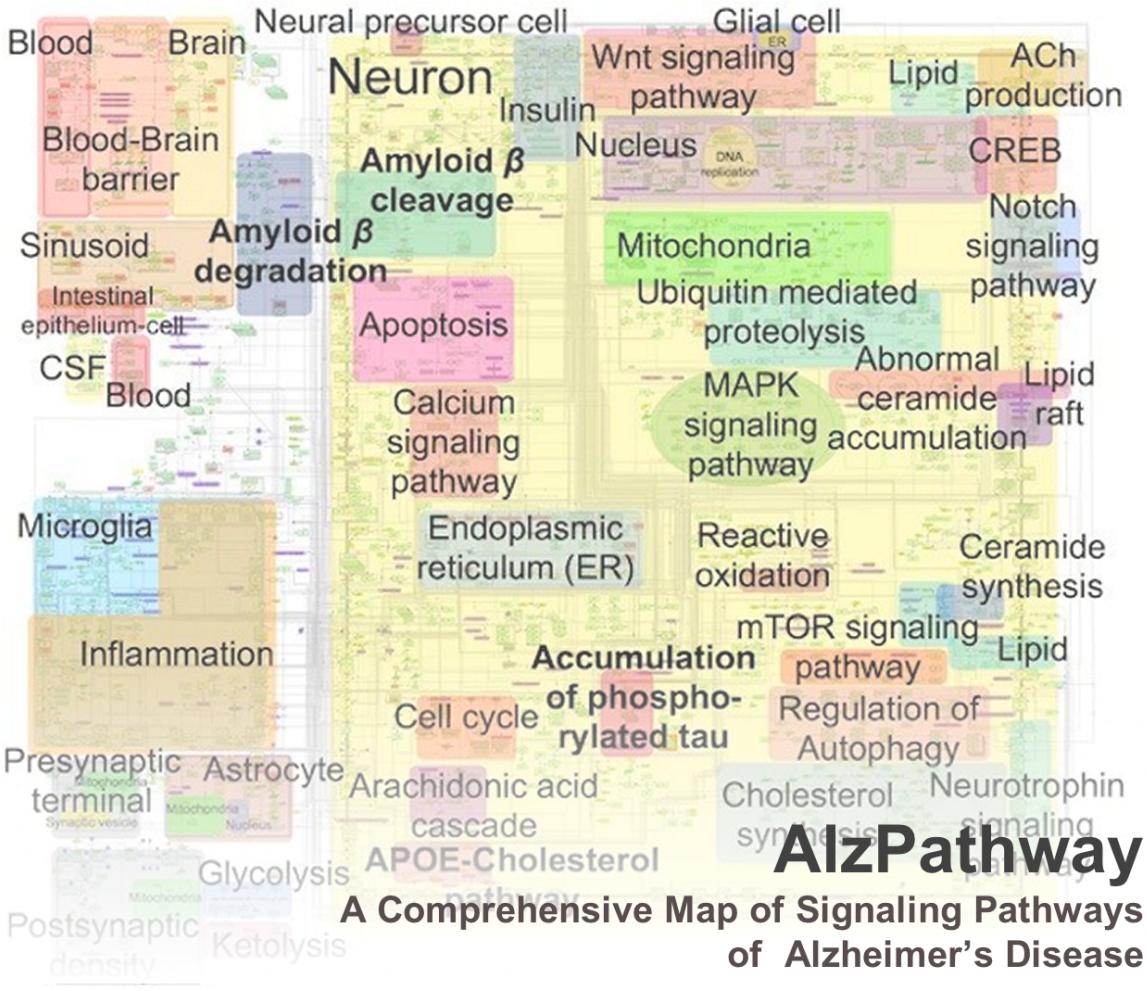


Yeast Cell Cycle

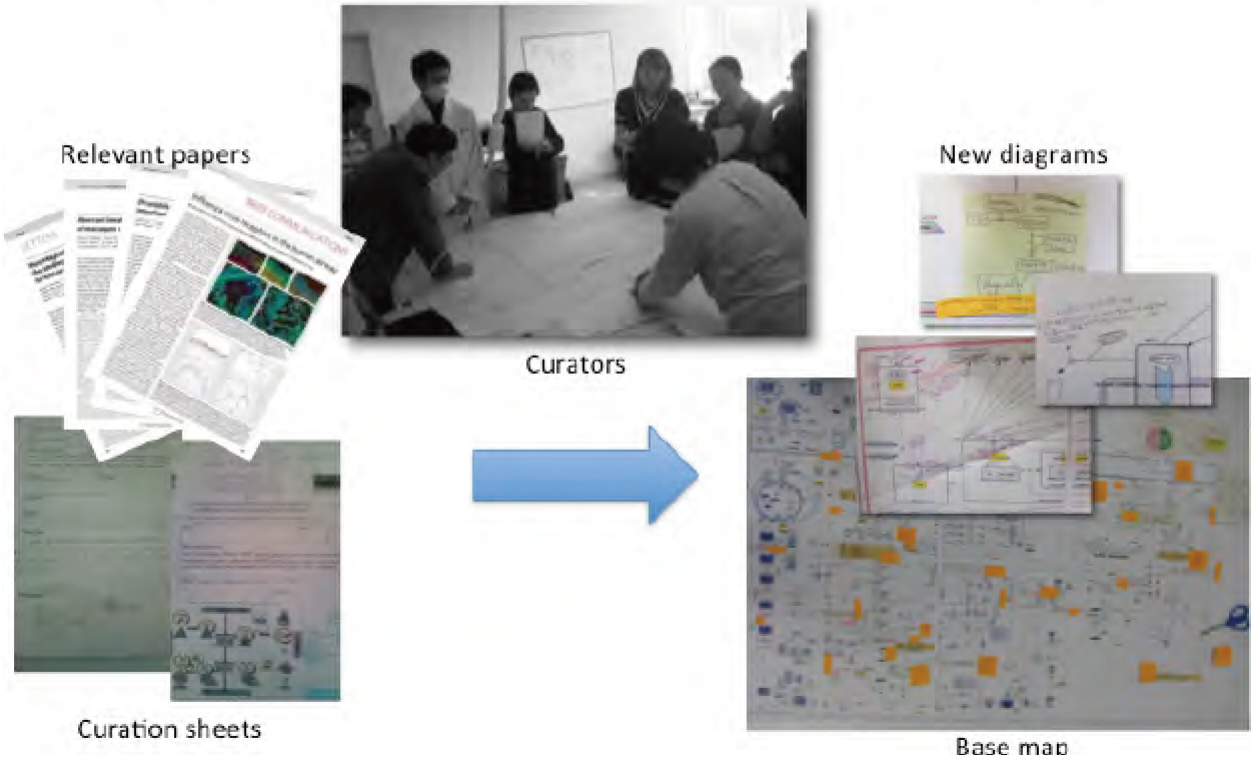




Oda and Kitano, Molecular Systems Biology, 2006


Blood
Brain
Neural precursor cell
Neuron
Glial cell
Wnt signaling pathway
Lipid
ACh production
Insulin
Nucleus
CREB
Blood-Brain barrier
Amyloid β cleavage
Amyloid β degradation
Sinusoid
Mitochondria
Notch signaling pathway
Intestinal epithelium-cell
Ubiquitin mediated proteolysis
Apoptosis
CSF
Blood
Abnormal ceramide accumulation
Lipid raft
MAPK signaling pathway
Calcium signaling pathway
Microglia
Endoplasmic reticulum (ER)
Reactive oxidation
Ceramide synthesis
mTOR signaling pathway
Lipid
Inflammation
Accumulation of phospho-rylated tau
Cell cycle
Regulation of Autophagy
Presynaptic terminal
Astrocyte
Arachidonic acid cascade
Cholesterol
Neurotrophin signaling
APOE-Cholesterol
Glycolysis
Postsynaptic
Ketolysis
AlzPathway
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

MINORITY REPORT

- 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.

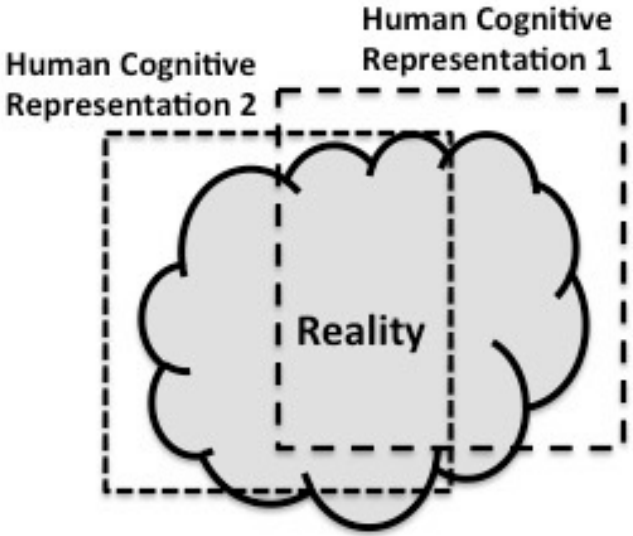
(B-2) Reports are from diverse labs

Kitano, H., AI Magazine, March 2016 (to appear)

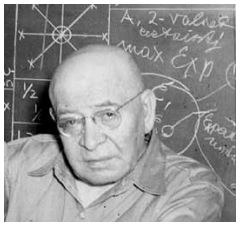
Human Cognitive Bias and Limitations of Semantic Mapping

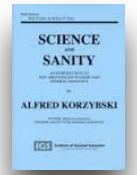
Cognitive Bias in Clinical Reasoning

- **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

Orignal Image

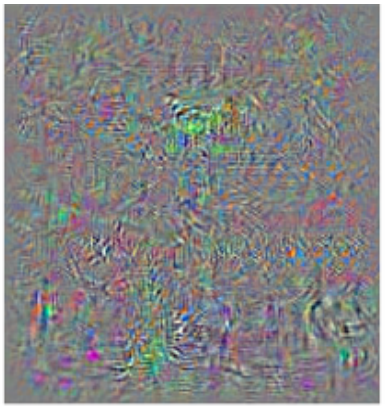
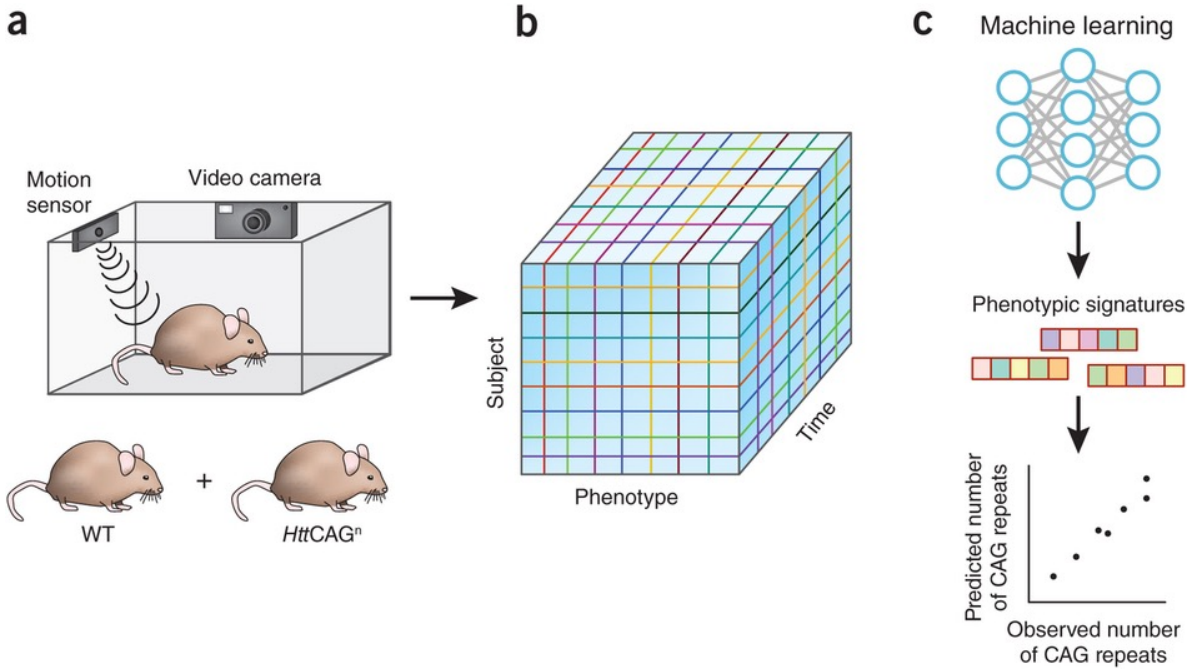


Image of Noise (amplified)

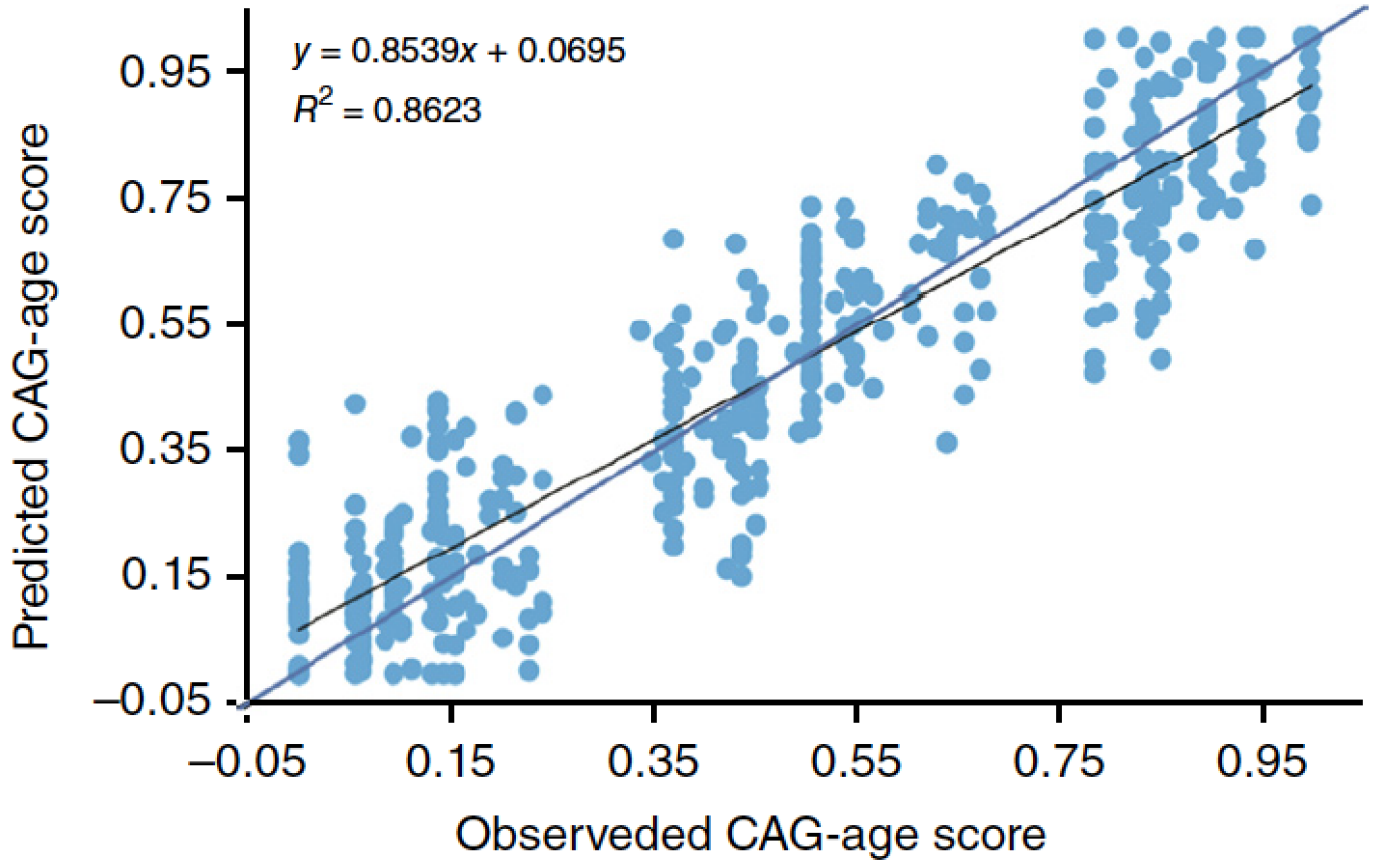
Deep Learning identified as Ostrich

Free Image （https://pixabay.com/）

Szegedy, et al., “Intriguing properties of neural networks,” International Conference on Learning Representation, 2014



Ruderfer, D. and Dudley, J., Nature Biotech, 34 (8) 2016



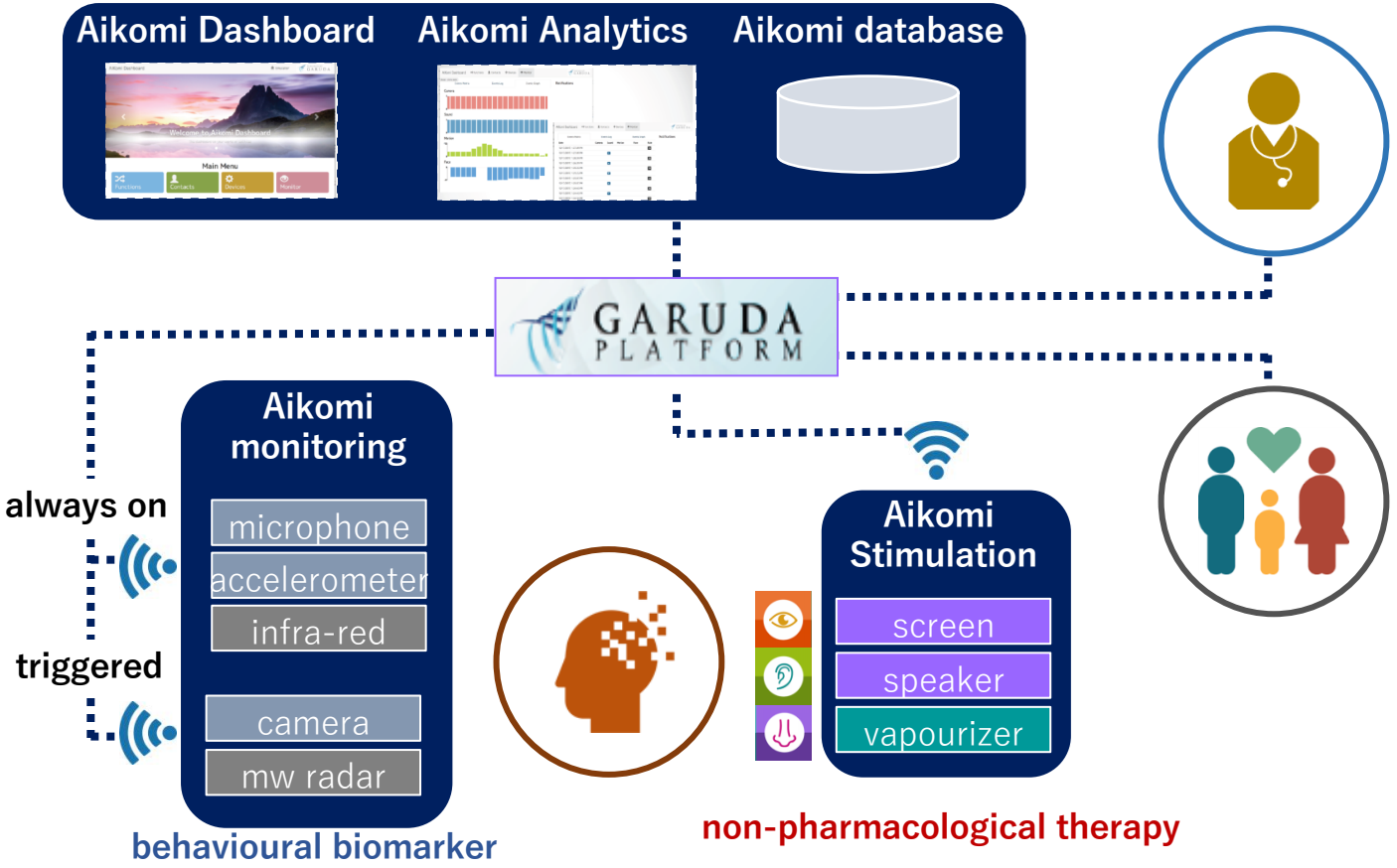
Alexandrov, V., et al.,Nature Biotech, 34 (8) 2016

寄り添う人たちを、支えたい

Aikomi

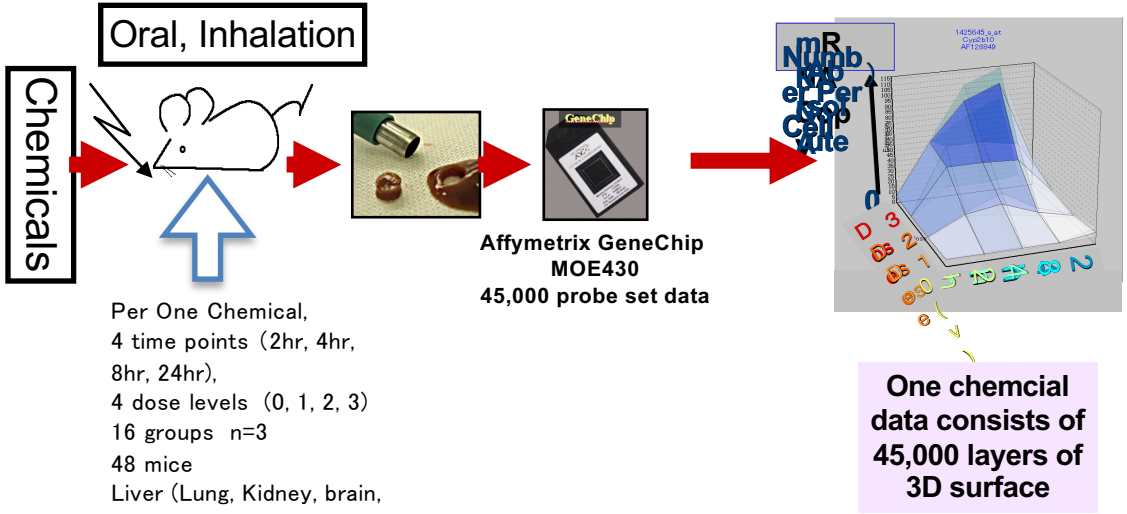
株式会社Aikomiは、認知症の方々が家族や介護者とともに
地域社会の中で、自分らしく暮らし続けるための
ソリューションを提供します。

Non-Pharmacological Intervention for Dementia

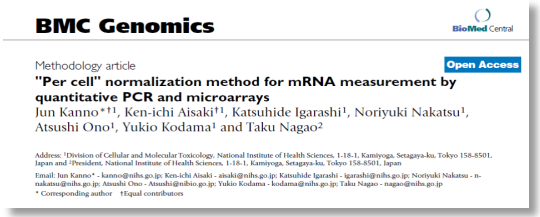


DTOX: Deep neural network based computational framework to analyze omics data in TOXicology

Percellome “Per Cell” -ome Database



Acephate	DMSO	Paclitaxel (Taxol)
Acetaldehyde	Domoic acid	Paraquat dichloride
Acetaminophen	Estragole	Pentachlorophenol
Agaritine	Ethanol	Permethrin
Alpha lipoic acid	4-Ethylnitrobenzene	Phenobarbital sodium
3-Amino-1H-1,2,4-triazole	Ethynyl estradiol	Phenytoin
4-amino-2,6-dichlorophenol	FK506	Phytol
2-Aminomethylpyridine	Formalin	5-Pregnen-3beta-ol-20-one-16alpha-carbonitrile
AraC	Forskolin	Pyriproxyfen
Aspirin	Fullerene	9-cis retinoic acid
Azacytidine	Genistein	All trans retinoic acid
Benzene	Genistin	Rifampicin
Bisphenol A	Glycyrrhizin2K	Sodium Arsenite
Bromobenzene	Hydroxycitric Acid	Sodium Dehydroacetate
Caffeine	Hydroxycitric Acid	Tamoxifen
Carbaryl	Ibuprofen (dl-p-isobutylhydratropic acid)	TCDD(2,3,7,8-Tetrachlorodibenzo-p-Dioxin)
Carbon tetrachloride	Indigo	TCDF(2,3,7,8-Tetrachlorodibenzofuran)
2-Chloro-4,6-dimethylaniline	Isoniazid	Tebufenozide
Cisplatin	Kanamycin monosulfate	Testosterone propionate
Citric acid-calcium salt	Levothyroxine	Thalidomide
Clofibrate	Maltol	Toluene
Coenzyme Q10	MEHP	Transplatin
Curcumin	Menthyl Valerate	1,2,3-Triazole
Daizein	Methanol	1,2,4-Triazole
Deet	Methoprene	Tributyltin chloride
DEHP	Methoprene acid	Troglitazone
Dexamethasone	Methyl dihydro jasmonate	Valproic acid sodium salt
Dibutyltin dichloride	3-methylcholanthrene	Verbenone
1,2-Dichloro-3-nitrobenzene	Monocrotaline	2-Vinylpyridine
Diethylnitrosamine (C57BL/6)	Nerolidol	Warfarin
Diethylstilbestrol	N-ethyl-N-nitrosourea	青色1号 Blue No.1
Digitoxin	N-Methylaniline	青色2号 Blue No.2
2,4-dinitrophenol	Omeprazole	赤色40号 Red No.40
		赤色226号 Red No.226


BMC Genomics

BioMed Central

Methodology article

Open Access

"Per cell" normalization method for mRNA measurement by quantitative PCR and microarrays

Jun Kanno*†1, Ken-ichi Aisaki†1, Katsuhide Igarashi1, Noriyuki Nakatsu1, Atsushi Ono1, Yukio Kodama1 and Taku Nagao2

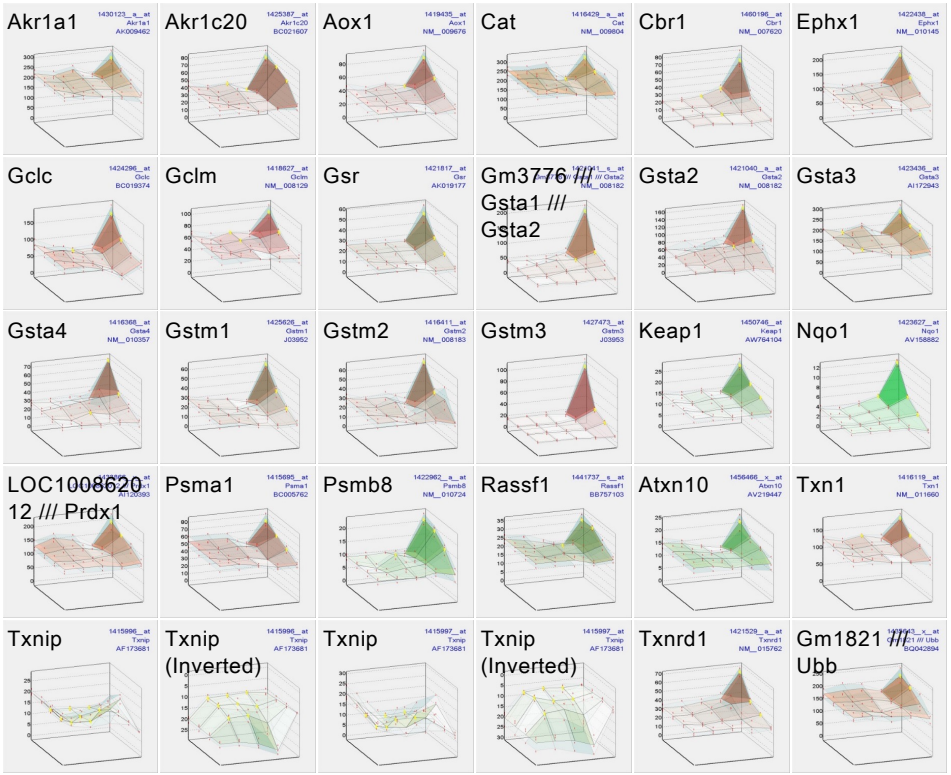
Address: 1Division of Cellular and Molecular Toxicology, National Institute of Health Sciences, 1-18-1, Kamiyoga, Setagaya-ku, Tokyo 158-8501, Japan and 2President, National Institute of Health Sciences, 1-18-1, Kamiyoga, Setagaya-ku, Tokyo 158-8501, Japan

Email: Jun Kanno* - kanno@nihs.go.jp; Ken-ichi Aisaki - aisaki@nihs.go.jp; Katsuhide Igarashi - igarashi@nihs.go.jp; Noriyuki Nakatsu - n-nakatsu@nihs.go.jp; Atsushi Ono - Atsushi@nihsis.go.jp; Yukio Kodama - kodama@nihs.go.jp; Taku Nagao - nagao@nihs.go.jp

* Corresponding author †Equal contributors

Open Access : ***BMC Genomics. 2006 Mar 29;7(1):64***

Examples of Percellome Data



J Tox Sci, Vol. 38-4, 2013

DTOX: Deep neural network based computational framework to analyze omics data in TOXicology

Training data

Class balanc Gray scaled/resized

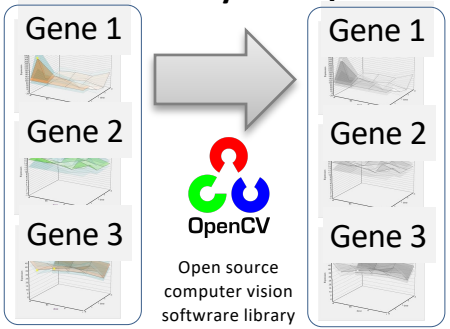
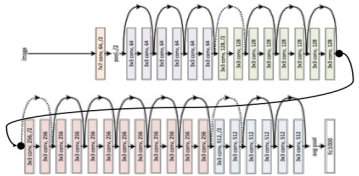


Image size: 321 x 321

Image size: 32 x 32

Deep convolutional neural network



Resnet (https://arxiv.org/abs/1512.03385)
We used network composed of 18 convolutional layers and that of 34 convolutional layers

Test data

Positive:	**1,148 images**
Negative:	**95 images**
Rough:	**124 images**
Intermediate:	**125 images**

Classifier model

The model classify surface images into 4 classes, positive, negative, rough, and intermediate classes

Performance of the model (Accuracy)

98%

Creating the Engine for Scientific Discovery

DENDRAL (1969)

Joshua Lederberg, Bruce Buchanan & Ed Feigenbaum

APPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR ORGANIC CHEMISTRY

The DENDRAL Project

DENDRAL Team

Robert K. Lindsay

Research Scientist
University of Michigan

Bruce G. Buchanan

Adjunct Professor of Computer Science
Stanford University

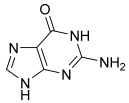
Edward A. Feigenbaum

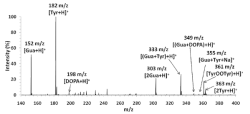
Professor of Computer Science
Stanford University

Joshua Lederberg

President, The Rockefeller University

Formerly, Professor of Genetics
Stanford University





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] Pınar Pir,[2] Larisa N. Soldatova,[1] Andrew Sparkes,[1] Kenneth E. Whelan,[1] Amanda Clare[1]

Sparkes *et al. Automated Experimentation* 2010, **2**:1
http://www.aejournal.net/content/2/1/1

AUTOMATED EXPERIMENTATION

REVIEW Open Access

Towards Robot Scientists for autonomous scientific discovery

Andrew Sparkes*[1], Wayne Aubrey[1], Emma Byrne[3], Amanda Clare[1], Muhammed N Khan[1], Maria Liakata[1], Magdalena Markham[2], Jem Rowland[1], Larisa N Soldatova[1], Kenneth E Whelan[1], Michael Young[2] and Ross D King[1]

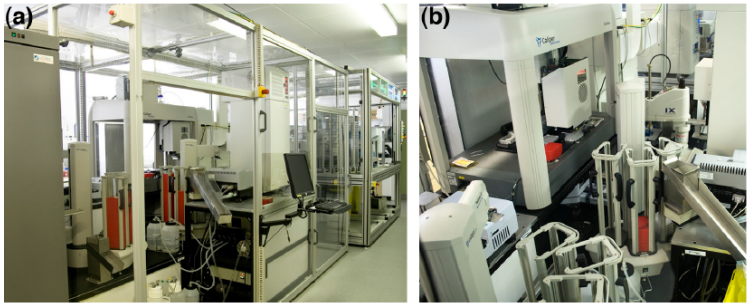


Figure 2 Adam's laboratory robotic system. (a) An external view of Adam's laboratory robotic system, also showing Eve's on the far right, and (b) a view looking down through the middle of Adam's robotic system, again with Eve's beyond.

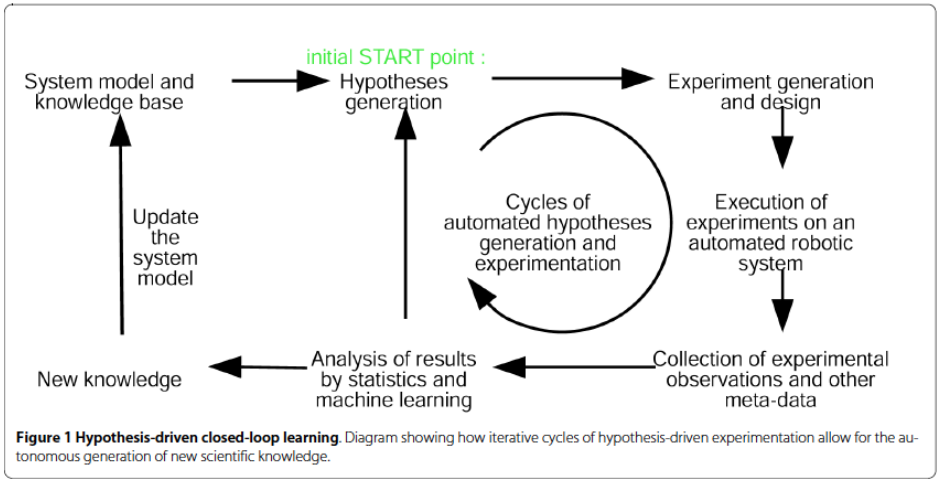


Figure 1 Hypothesis-driven closed-loop learning. Diagram showing how iterative cycles of hypothesis-driven experimentation allow for the autonomous generation of new scientific knowledge.

King, et al., Science 324, 85 (2009)

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

ACCIDENT, 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

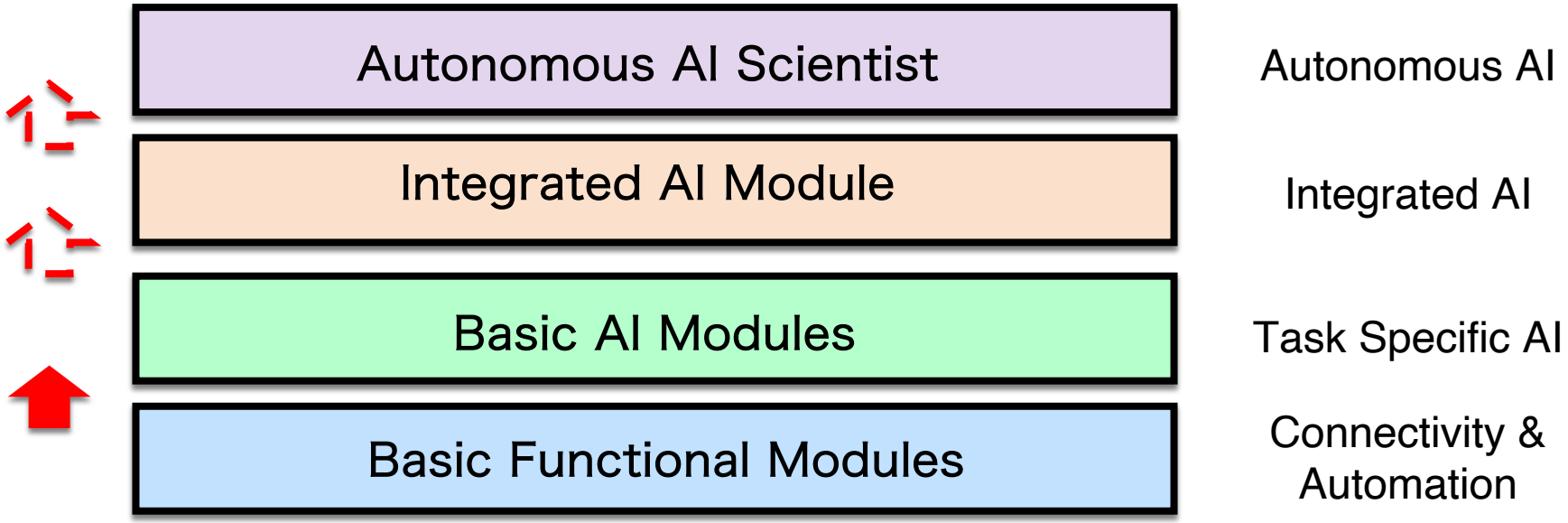
SEARCH & OPTIMIZATION

Prof. Alan Heeger

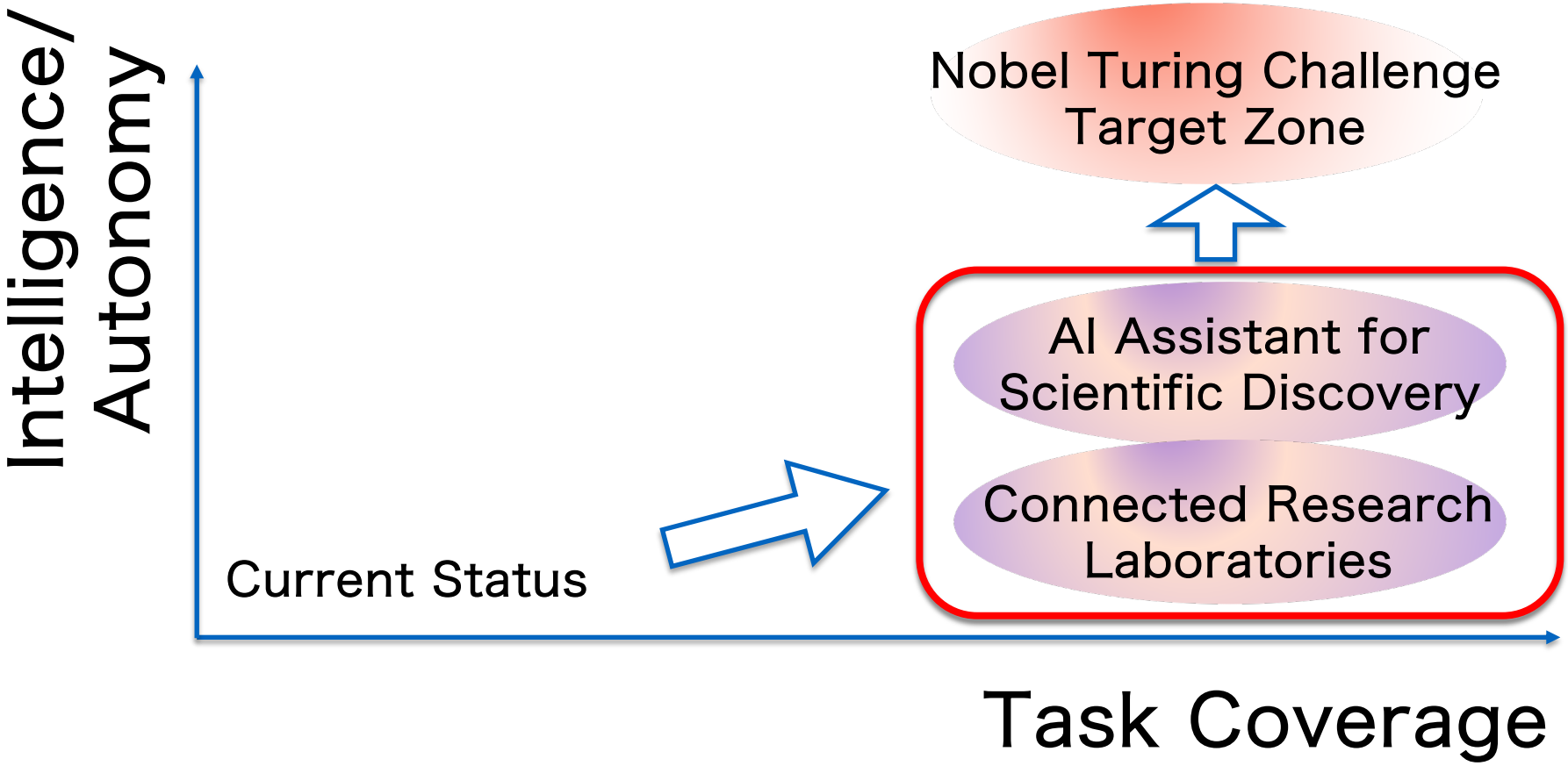
Conducting polymer thin film

Nobel Prize in Chemistry 2000

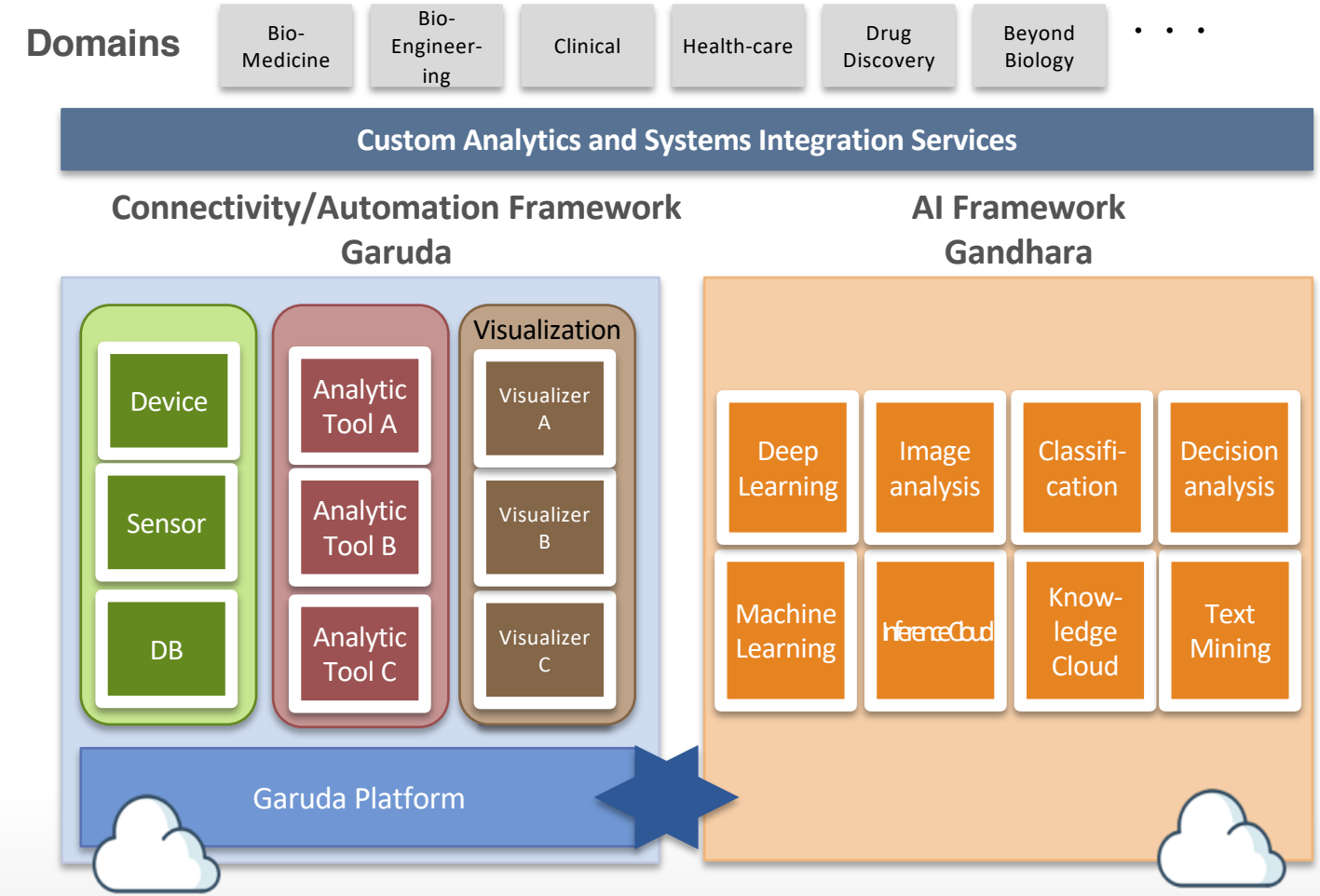
Technology Platform



Strategic Path



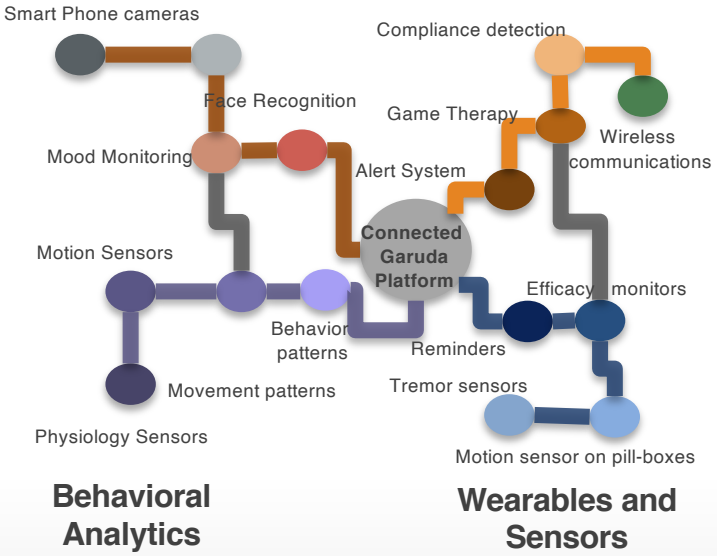
Technology Platform (SBI/SBX)

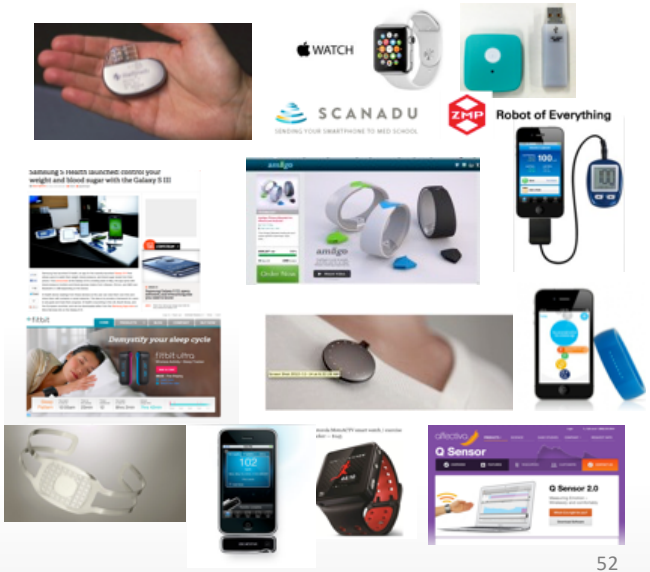


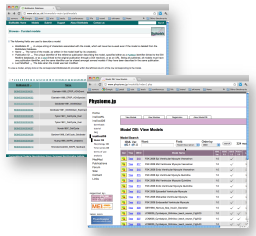
Garuda & Connected Devices

Connect to diverse spectrum of sensors and devices

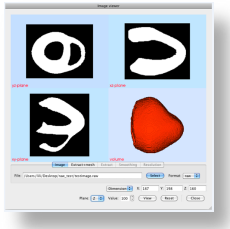
The Garuda Platform provides the capability to build connect digital solutions by integrating off-the-shelf wearables and monitoring devices



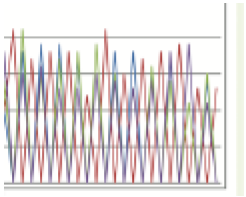




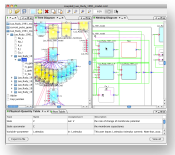
Model DB



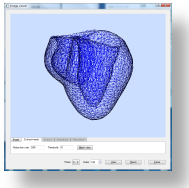
Medical Imaging



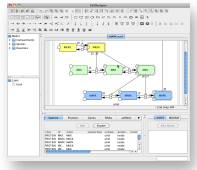
Genomic Data



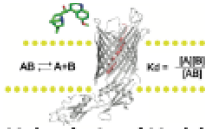
Physiological Modeling



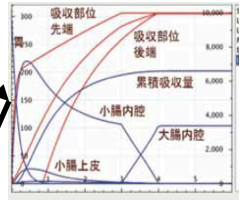
Organ-Scale Modeling

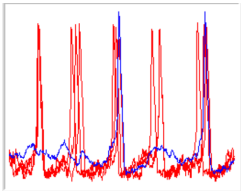


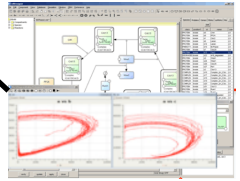
Cellular Modeling

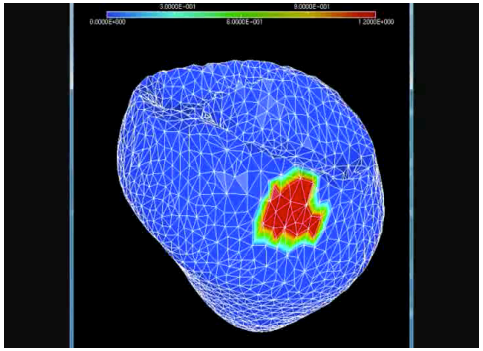


Ion-channel Modeling

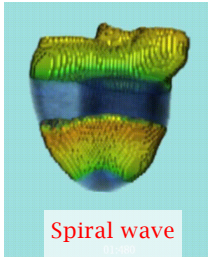




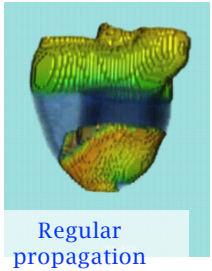


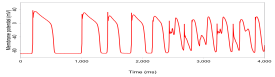


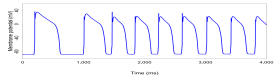
High Vesnarinone



No Vesnarinone







Shimadzu Collaboration Project

SBI The Systems Biology Institute

SBX Technologies

SHIMADZU Excellence in Science

大阪大学 OSAKA UNIVERSITY


分析装置にAI搭載 島津製作所

International Organization for Standardization ISO 9001

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

Shimadzu Data Explorer (Proto)

VizCloud*

Metabolites Information Query

Gene Annotator*

Explore KEGG pathway for KEGG metabolite IDs

iPath

Metabolic Pathway Enrichment in Reactome

Reactome Pathways

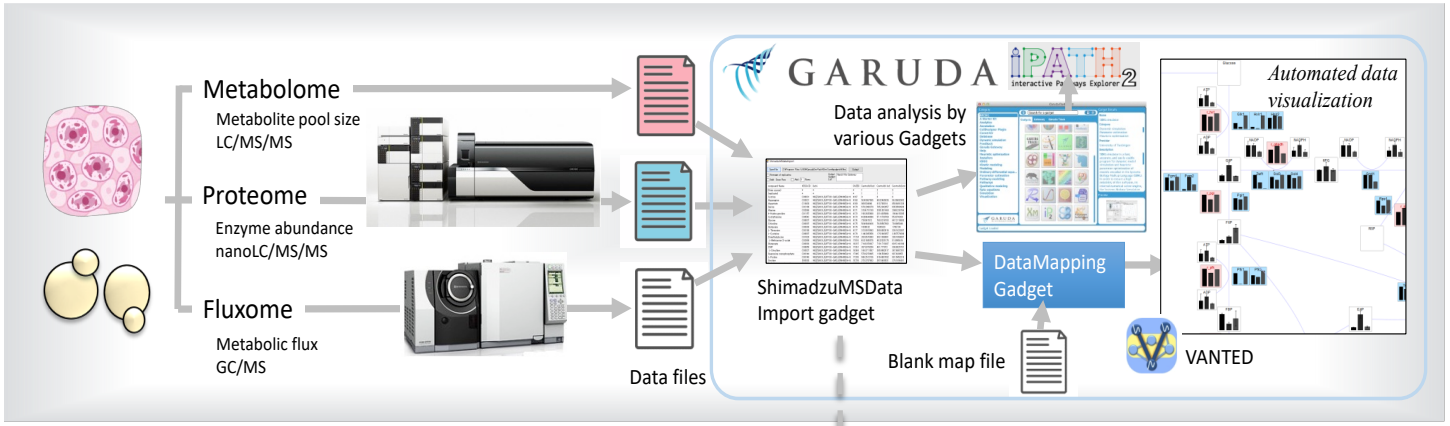
Load experimental data and visualize

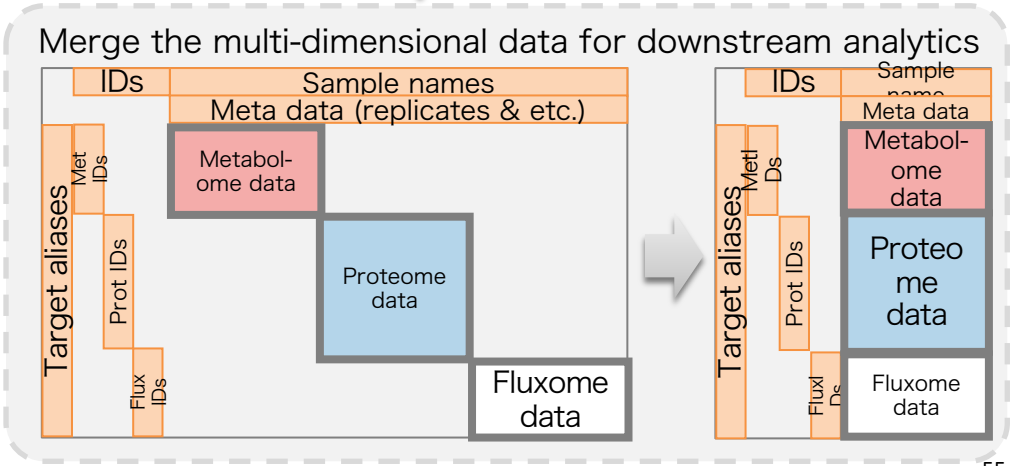
VANTED

GARUDA

Case Studies

Multi-Omics data analysis configuration





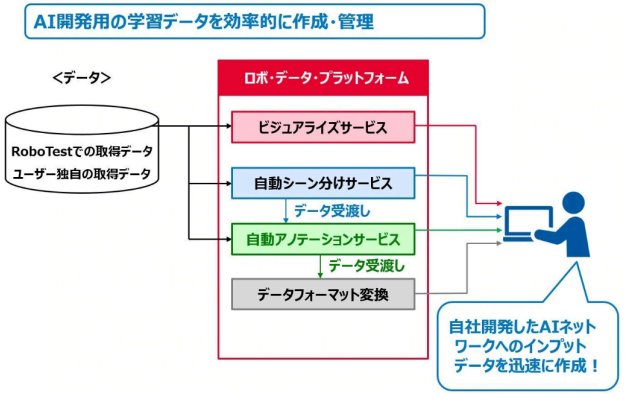
SBI The Systems Biology Institute

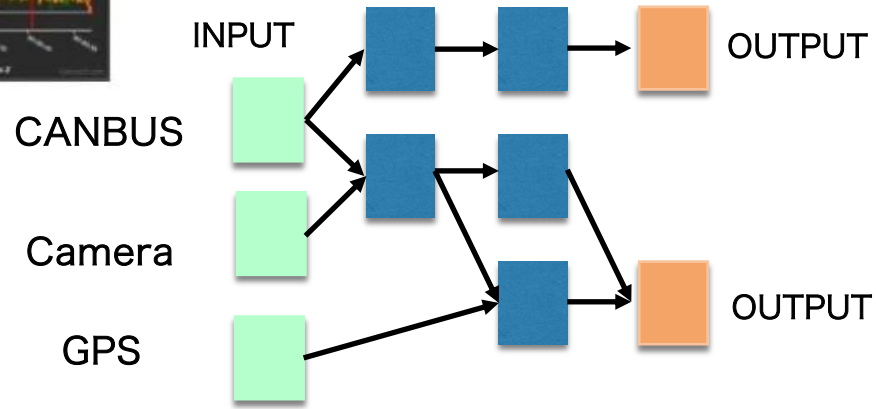
SBX Technologies

SHIMADZU Excellence in Science

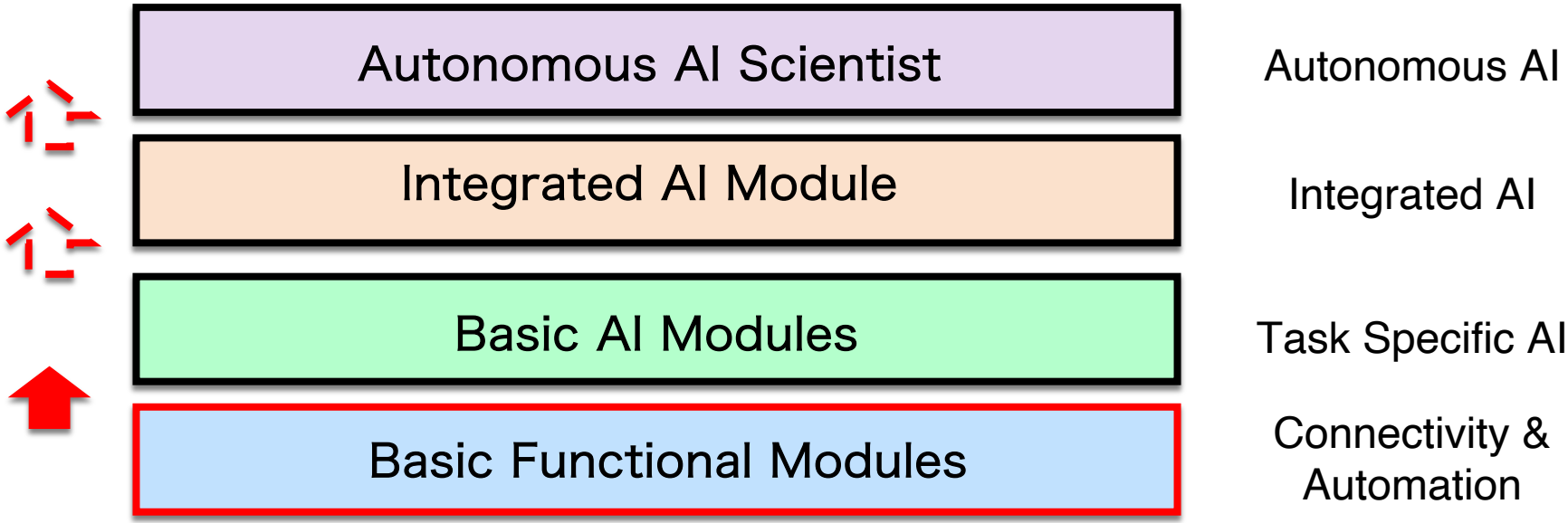
大阪大学 OSAKA UNIVERSITY

自動走行車の学習データ作成などへの応用例





Technology Platform



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

JST AMED Institute of Medical Science, University of Tokyo SBI The Systems Biology Institute OIST

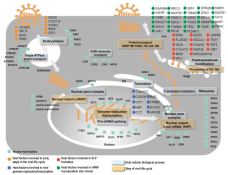
Network-based Compound Screening (NCS)

Case Study

369 Compounds

439 host proteins from Influenza Virus lifecycle

Protein	Uniprot acc.	Uniprot ID	PDB	Remarks
ABL1	P00519	ABL1_HUMAN	2HZI	Abl kinase domain in complex with PD180970 JIN
ABL2	P42684	ABL2_HUMAN	3GVU	The crystal structure of human ABL2 in complex with GLEEVEC STI



Watanabe et al., 2014

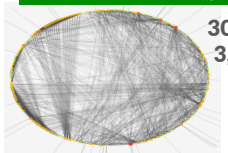
147 proteins

ML-based Docking simulation

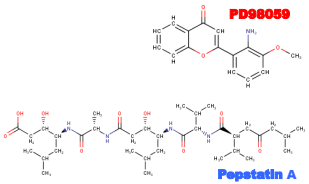
systemsDock[1]

Kun-Yi et al., 2016

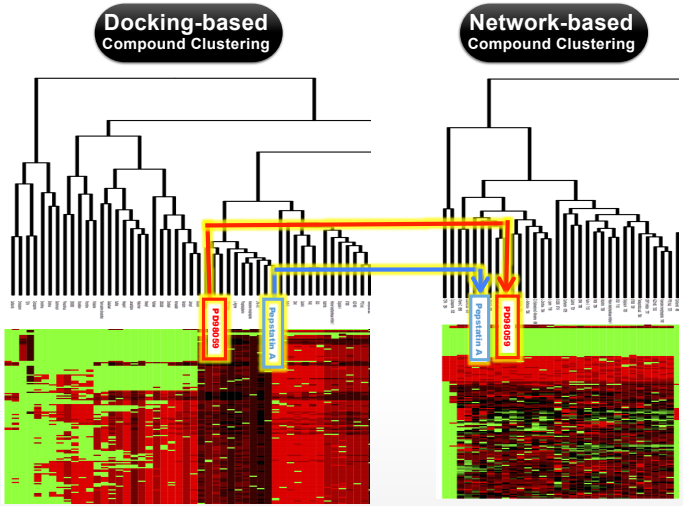
PPI Network Analysis



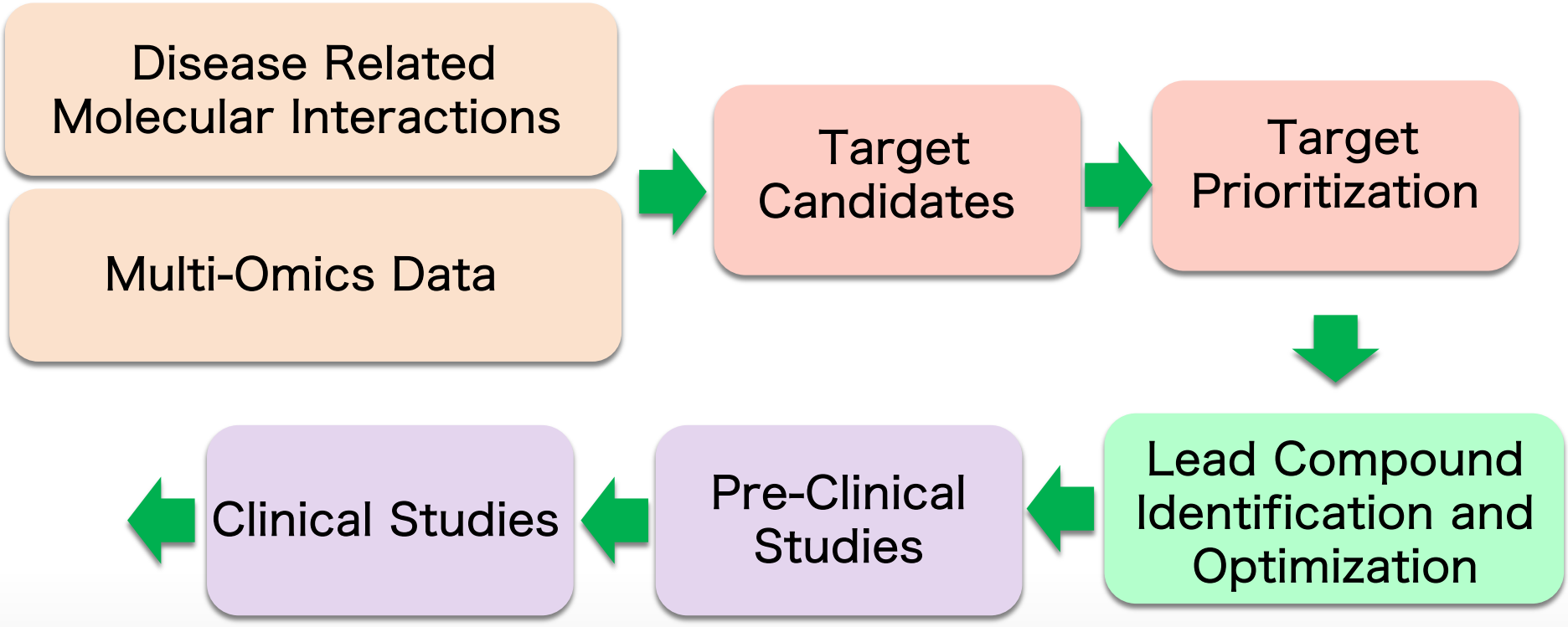
304 proteins
3,321 edges



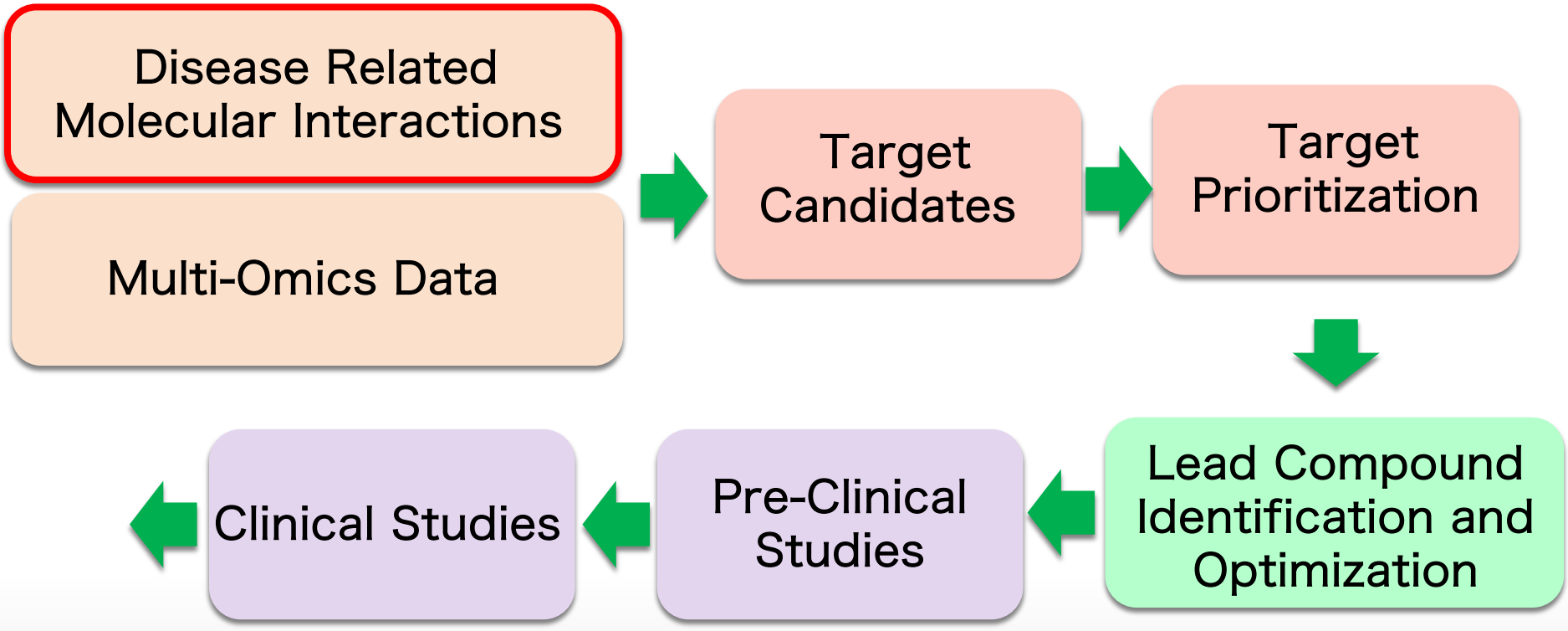
- ✓ For proof of concept, here we describe the example of compounds Pepstatin-A (protease inhibitor) and PD98059 (Map Kinase inhibitor) marked by blue and red boxes respectively
- ✓ Previous studies have shown that these compound possess activity against viral replication in human cell cultures (Barber 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



Very Simplified Process of System-Driven Drug Discovery

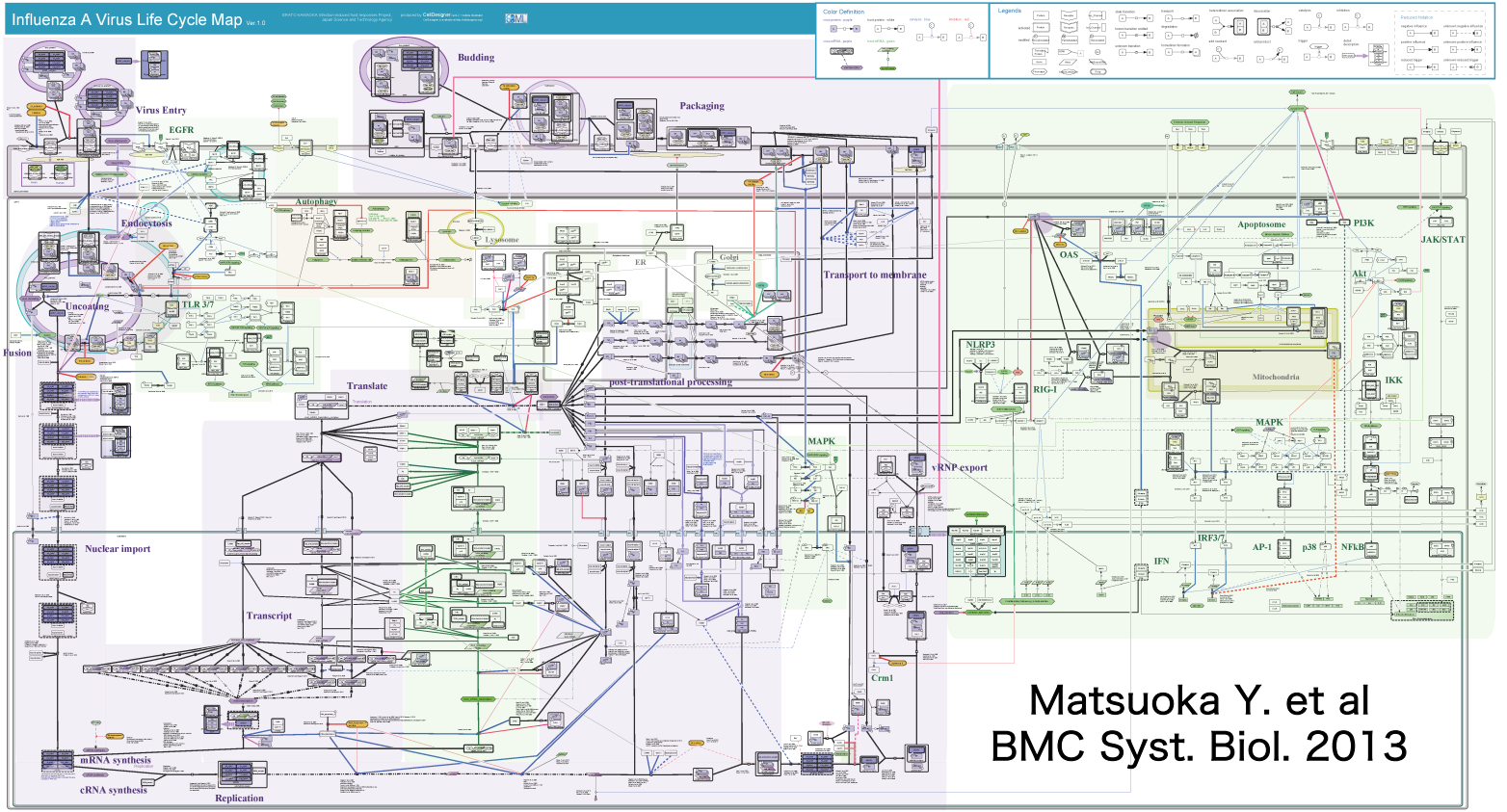


Very Simplified Process of System-Driven Drug Discovery



Literature-driven Approach

Influenza infection and replication network



Matsuoka Y. et al
BMC Syst. Biol. 2013

Large-Scale Knowledge Extraction from Text

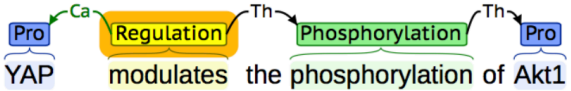


Figure 2: Example sentence with NLP event representations extracted.

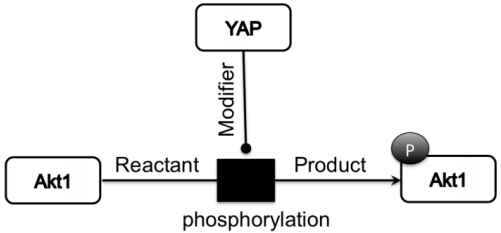
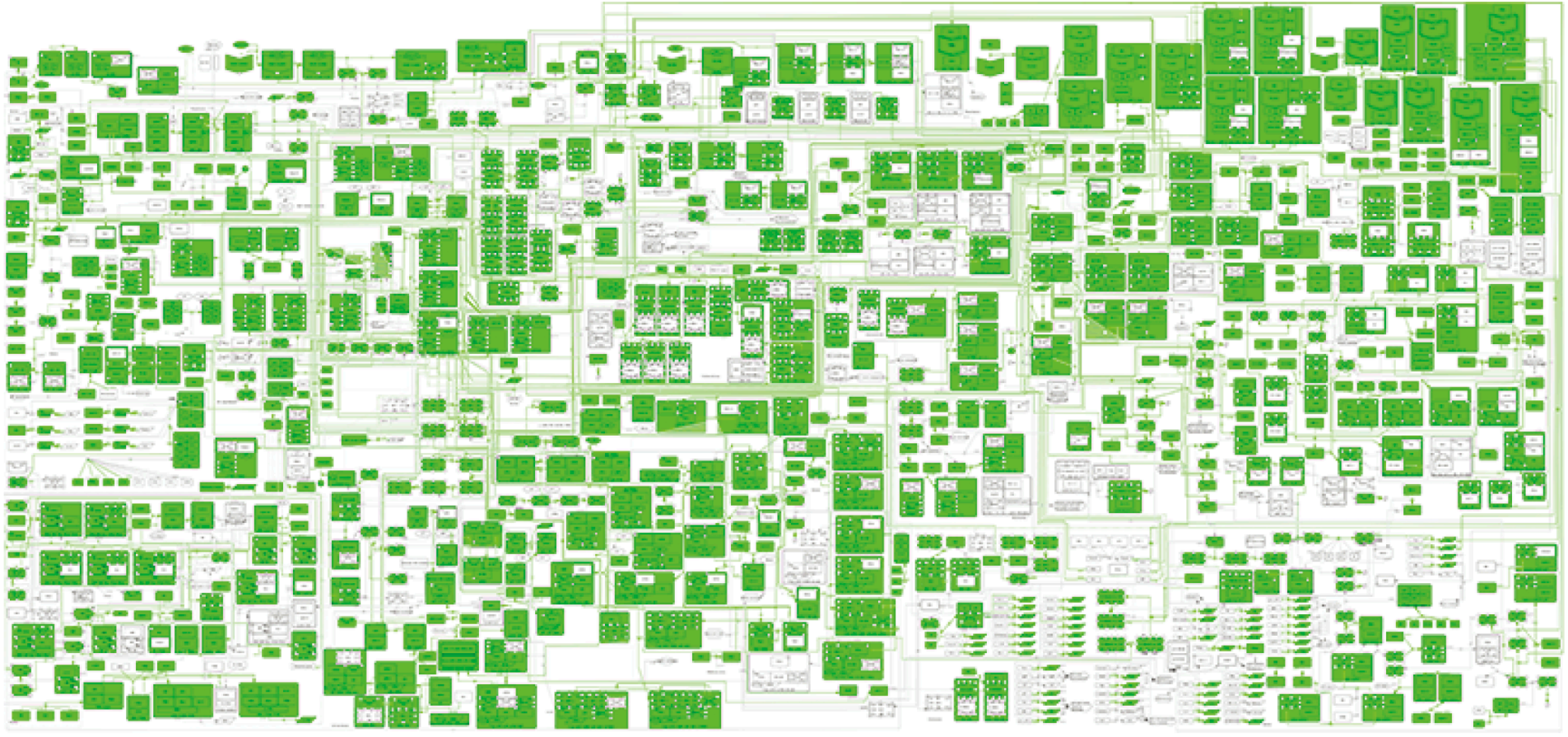


Figure 3: Phosphorylation reaction.

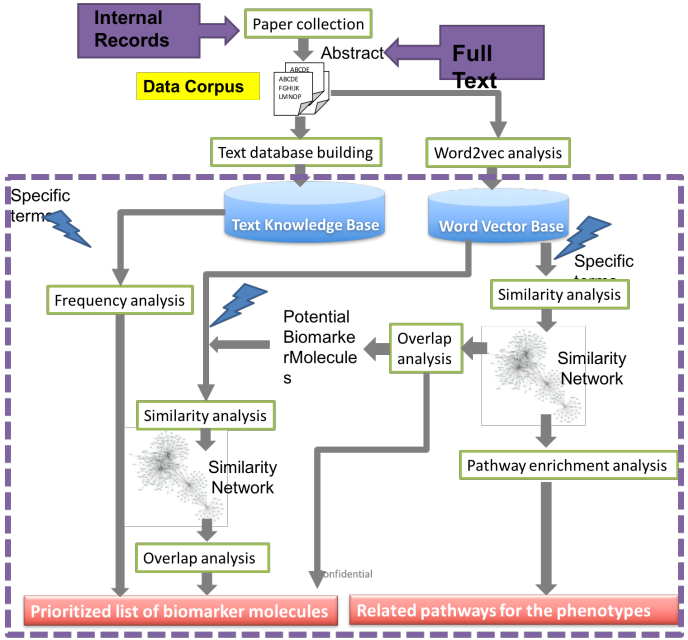
BioNLP 2016



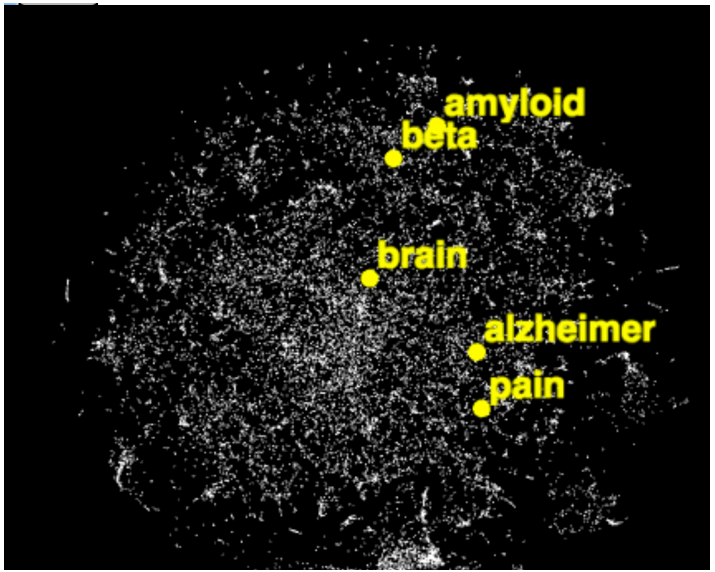
Key Case Studies

Text mining approach for BioMarker Discovery

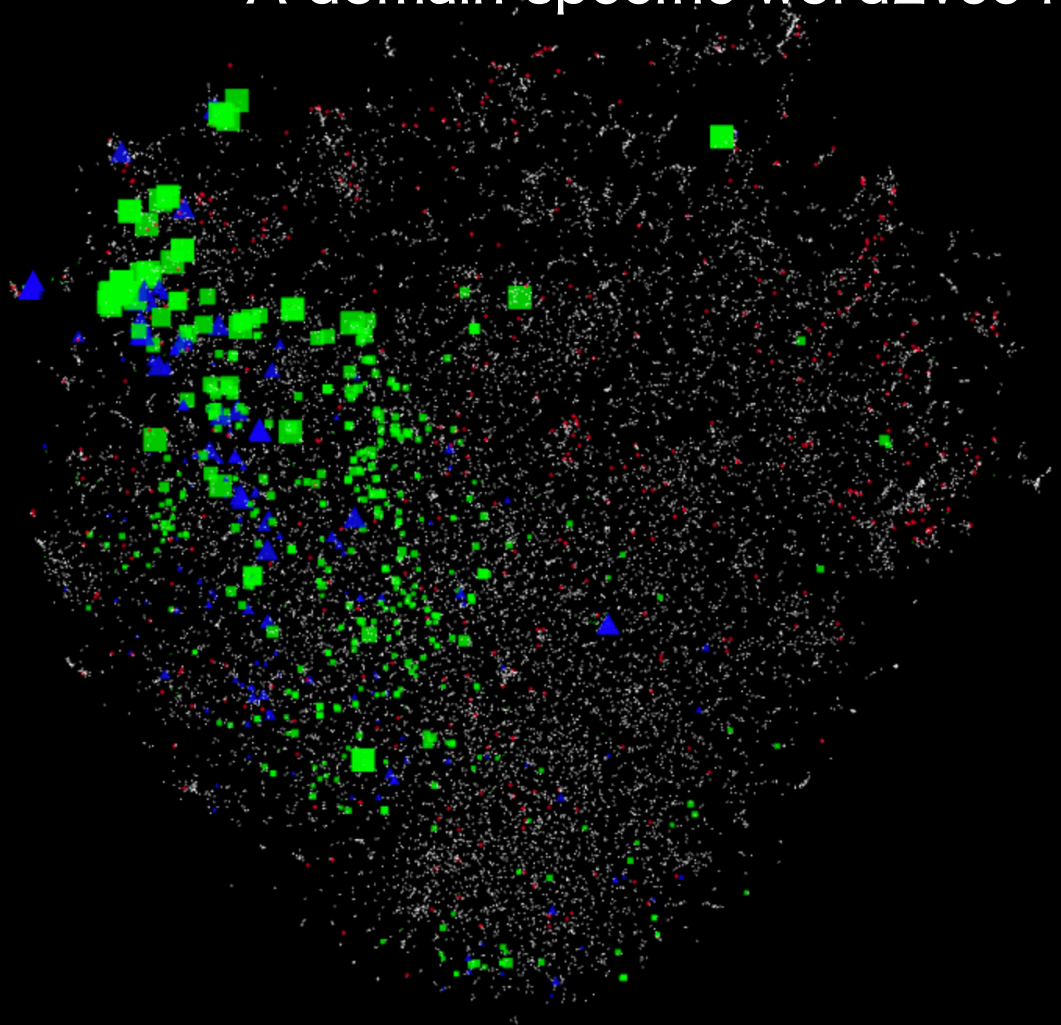
Biomarker Knowledge Mining Pipeline



Biomarker Galaxy
Associate biomarkers with diseases



A domain specific word2vec Model



150,000 Articles on Alzheimer's Disease

Over 15,000,000 words

Legend

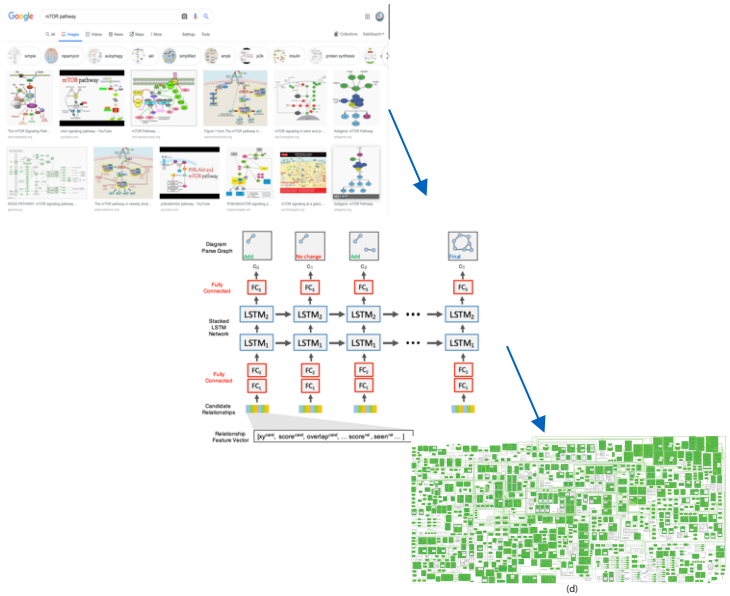
Gene

Chemical

Disease

Other

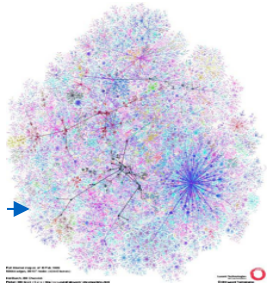
Image + Text based Pathway Reconstruction



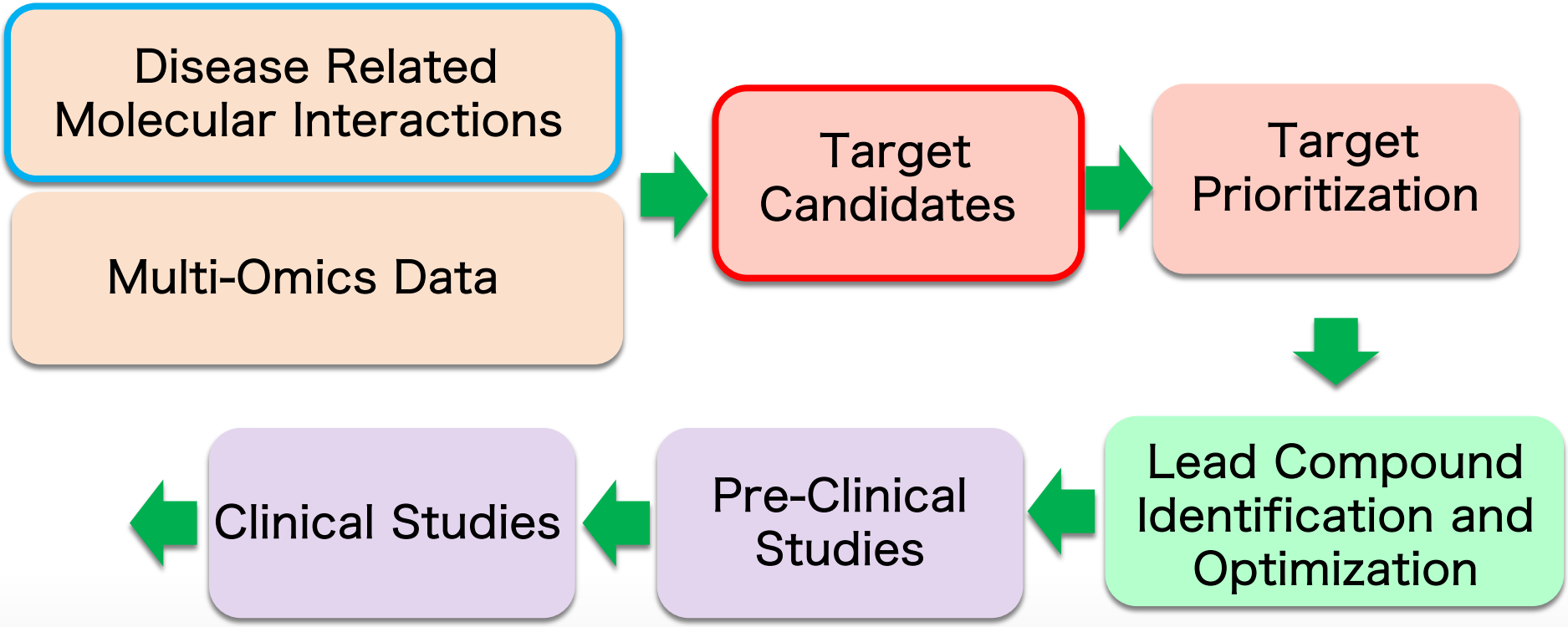
PUBMED based hypothesis generation system

PUBMED
> 20 million articles

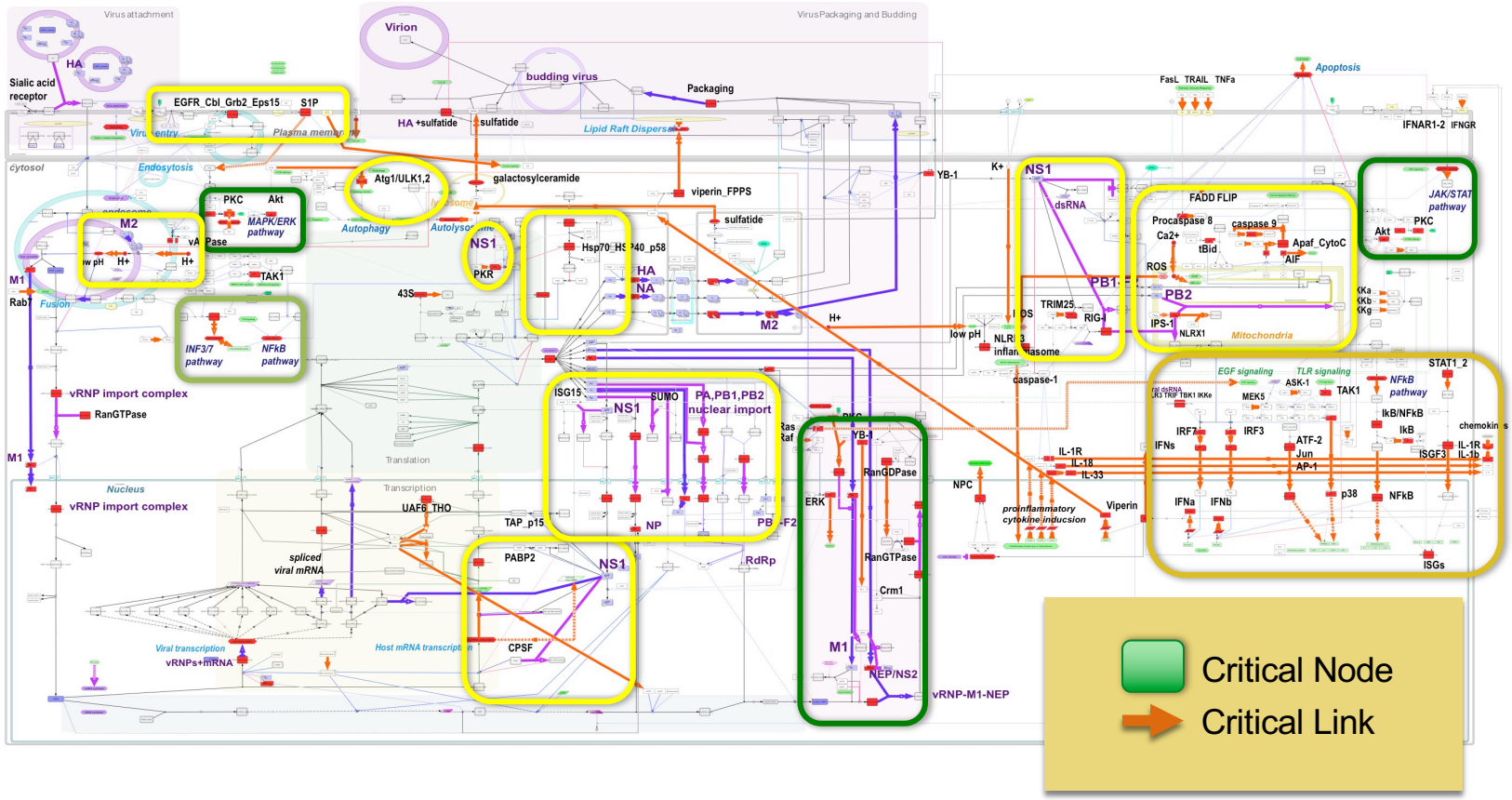
Generate hypotheses



Very Simplified Process of System-Driven Drug Discovery



Network controllability Analysis



Network Controllability Analysis

Class-I Controllability Analysis: Full Controllability

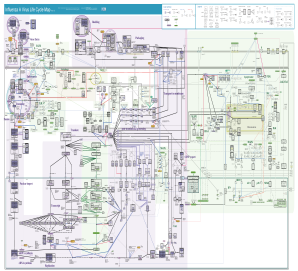
(Liu et al., Nature 2011)

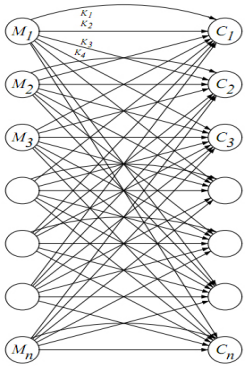
FluMap (State transition diagram)

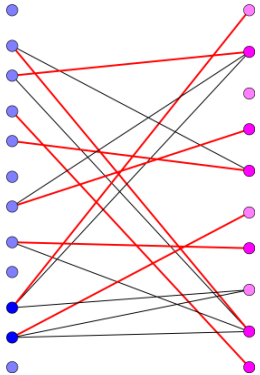
Convert to Bipartite graph

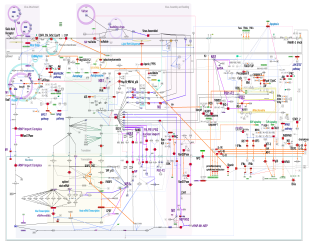
Calculate **maximum matching** with **Hopcroft-Karp** algorithm

Reflect results on FluMap









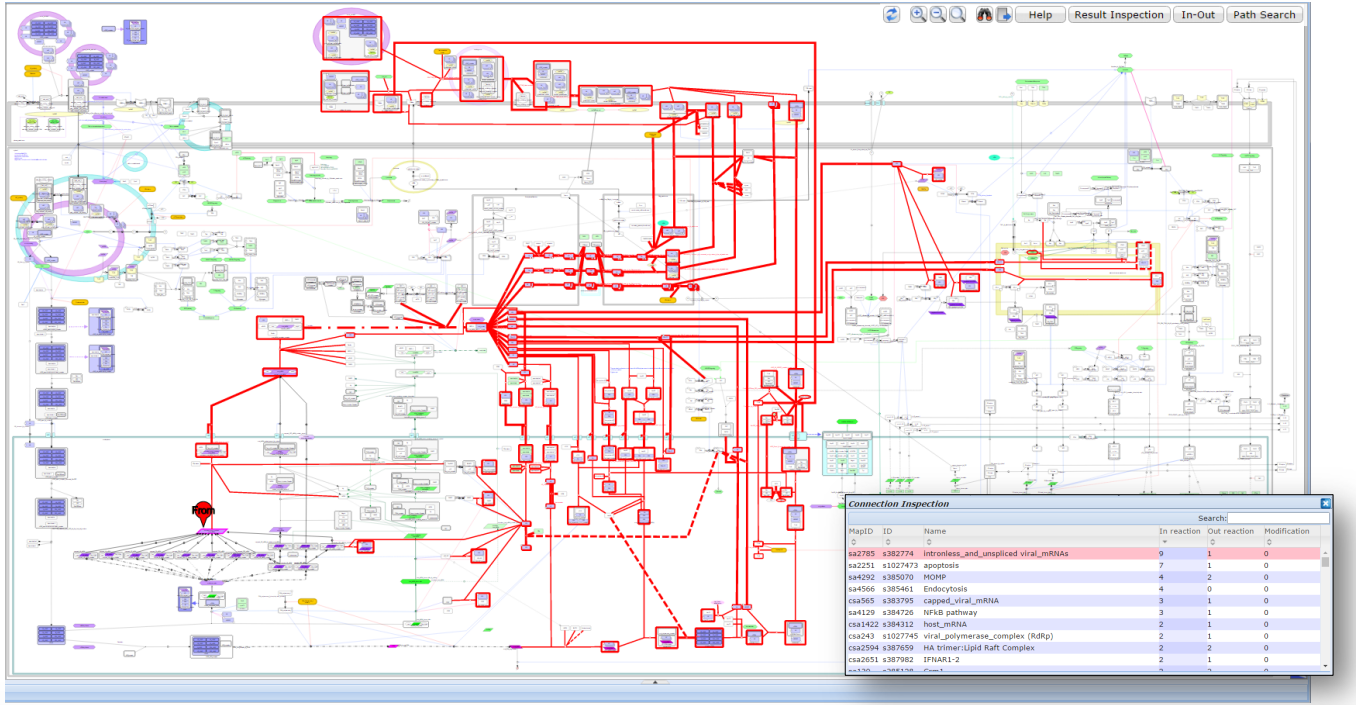
*Identify
Critical Nodes
Critical Links

Computational Network Analysis

(To identify and prioritize potential therapeutic targets)

Controllability:

- **Who is the major hub on the map?**
- Interactions/paths between particular molecules?

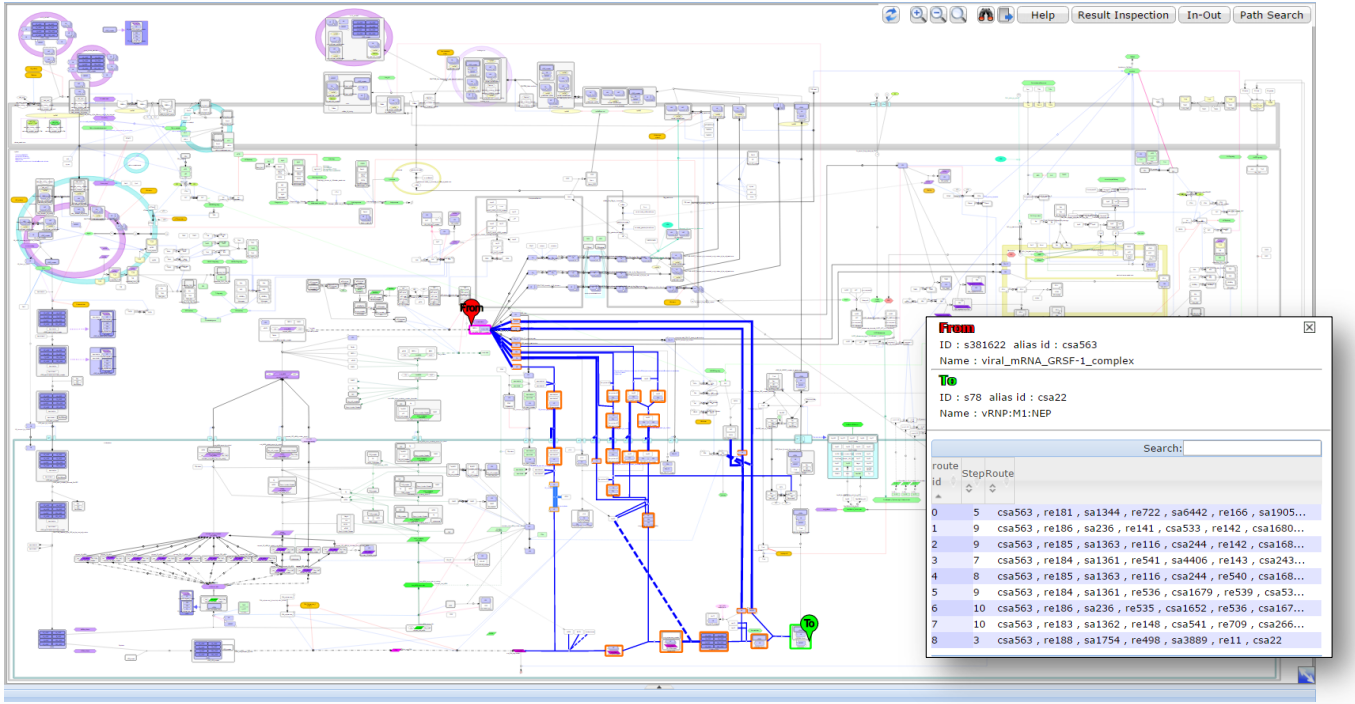


Computational Network Analyses

(To identify and prioritize potential therapeutic targets)

Controllability:

- Who is the major hub in the map?
- **Interactions/paths between particular molecules?**

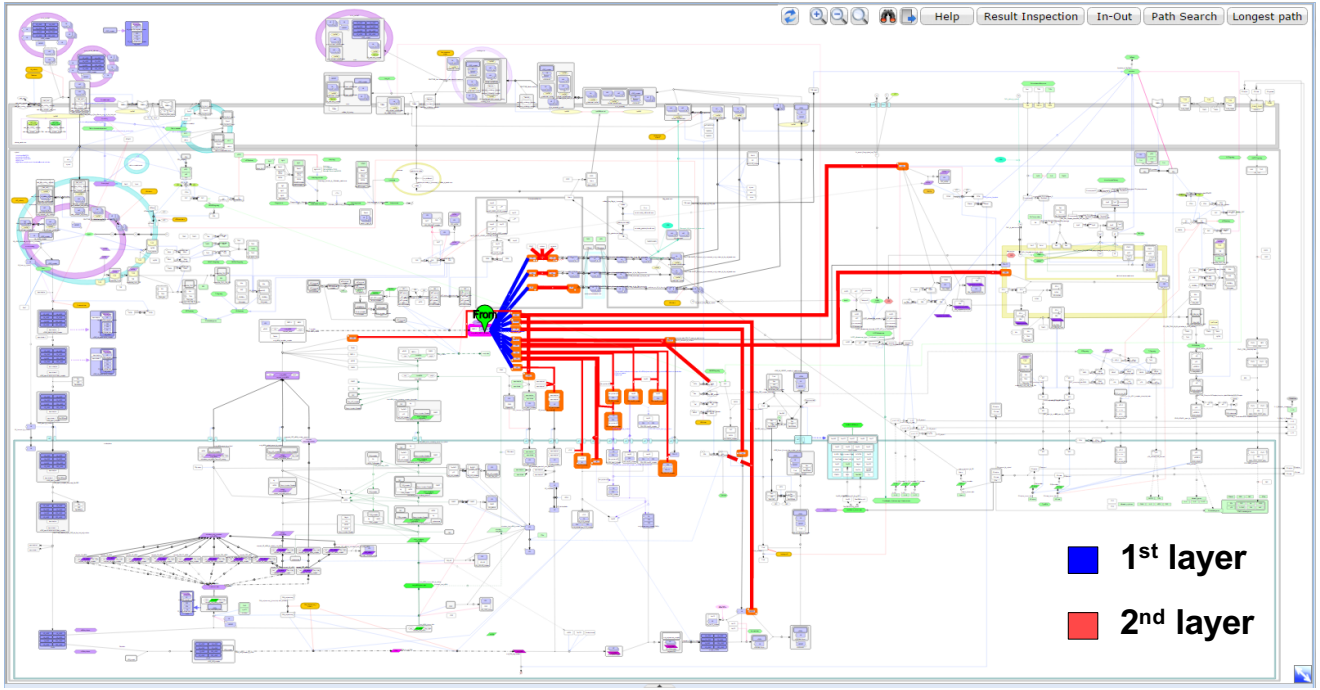


Computational Network Analyses

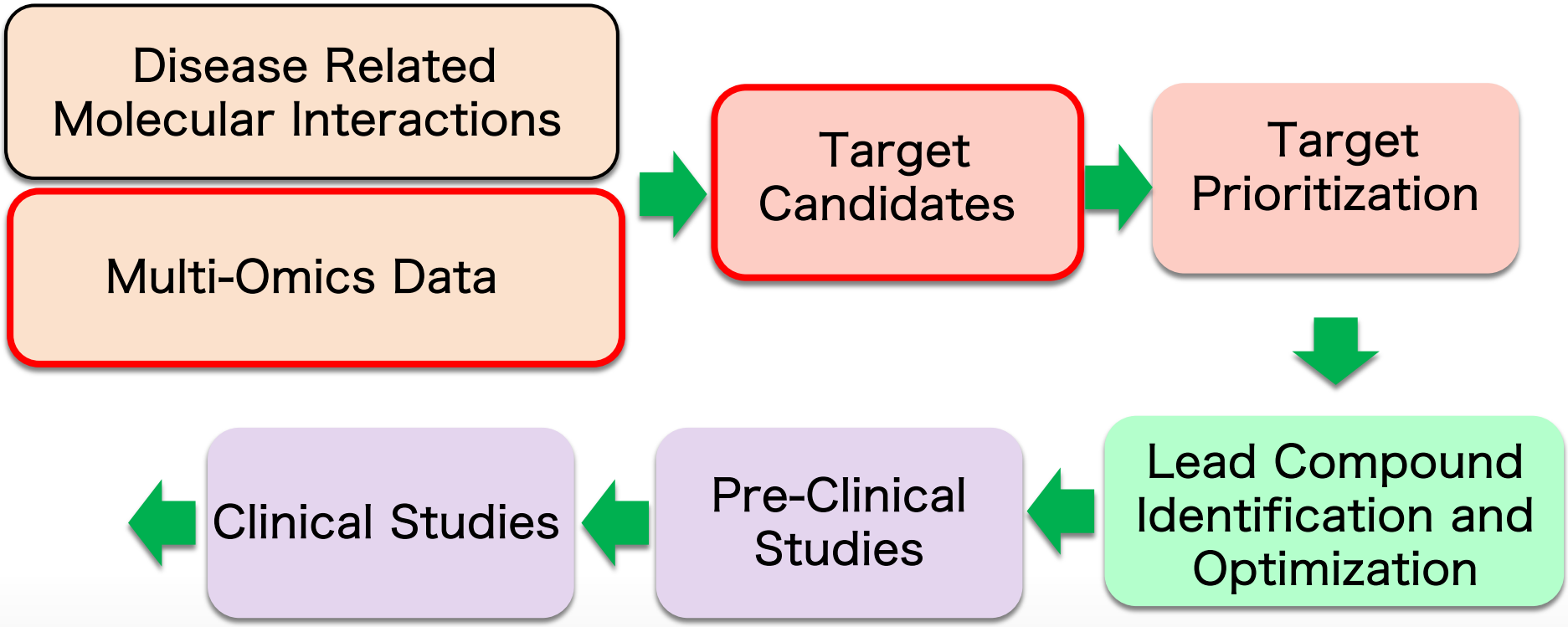
(To identify and prioritize potential therapeutic targets)

Controllability:

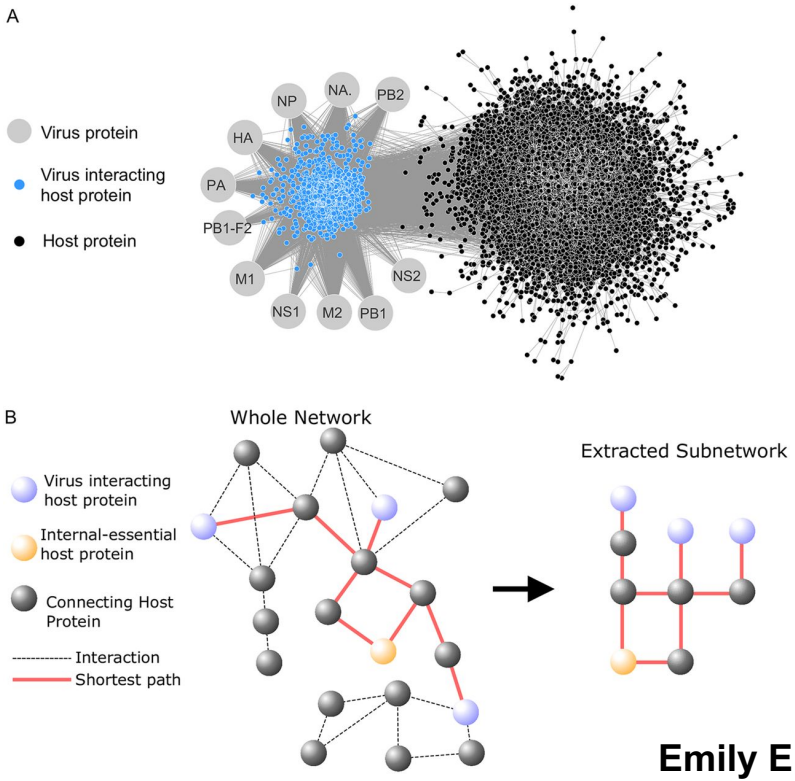
- Who is the major hub in the map?
- **Interactions/paths between particular molecules?**



Very Simplified Process of System-Driven Drug Discovery

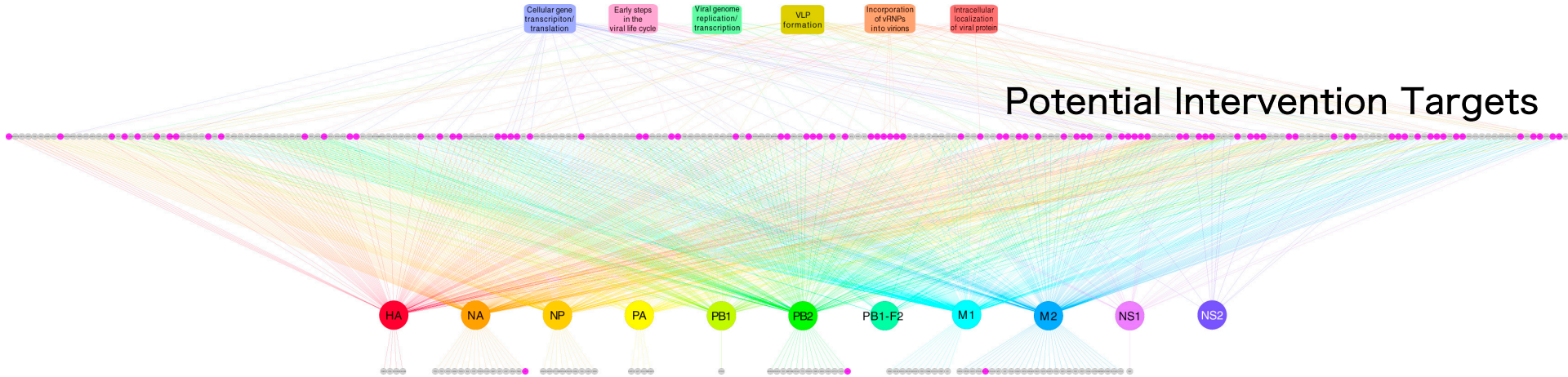


The virus-interacting network and the virus subnetwork



Emily E. Ackerman et al. mBio 2018; doi:10.1128/mBio.02002-18

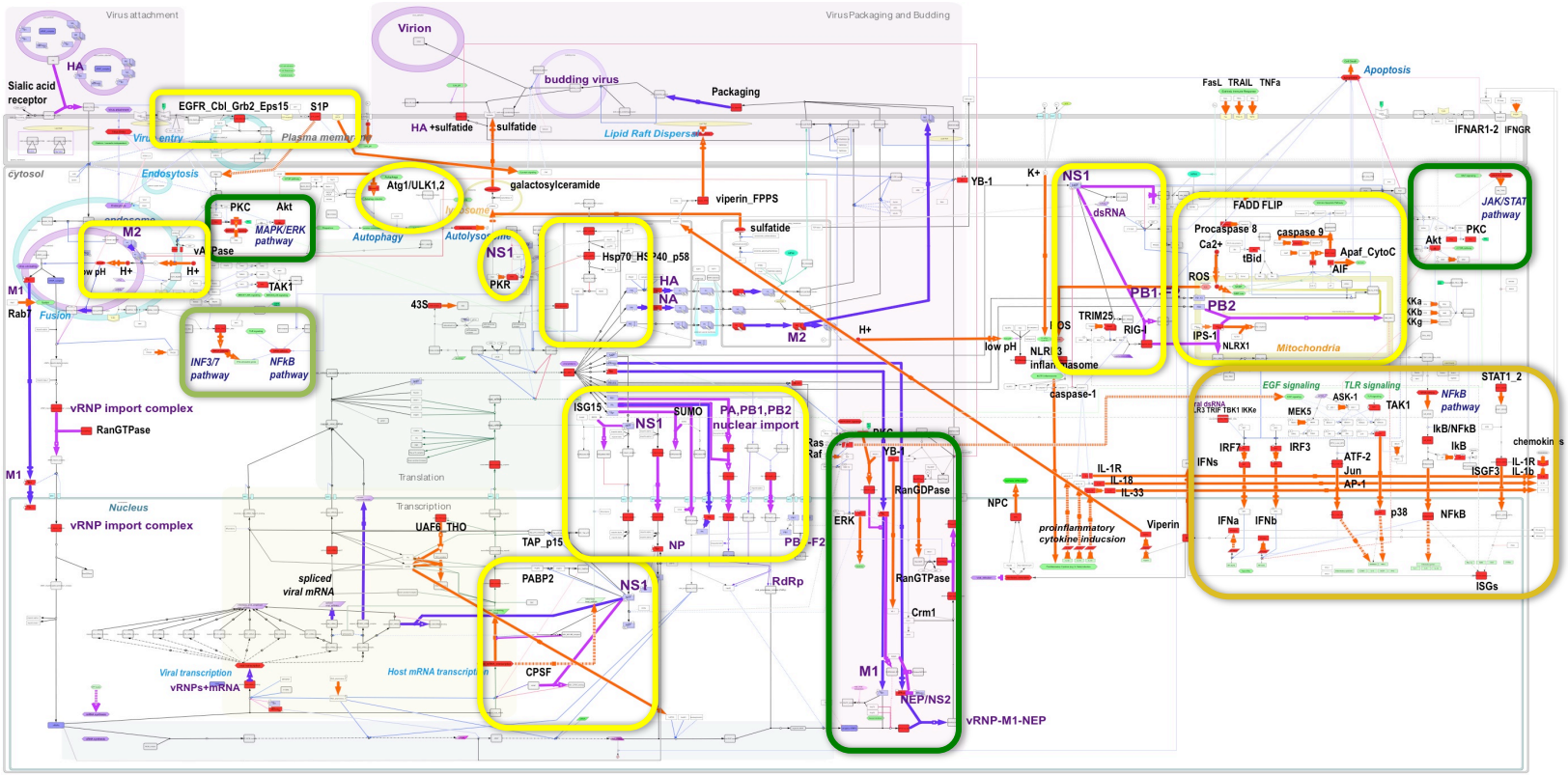
Target Biological Processes



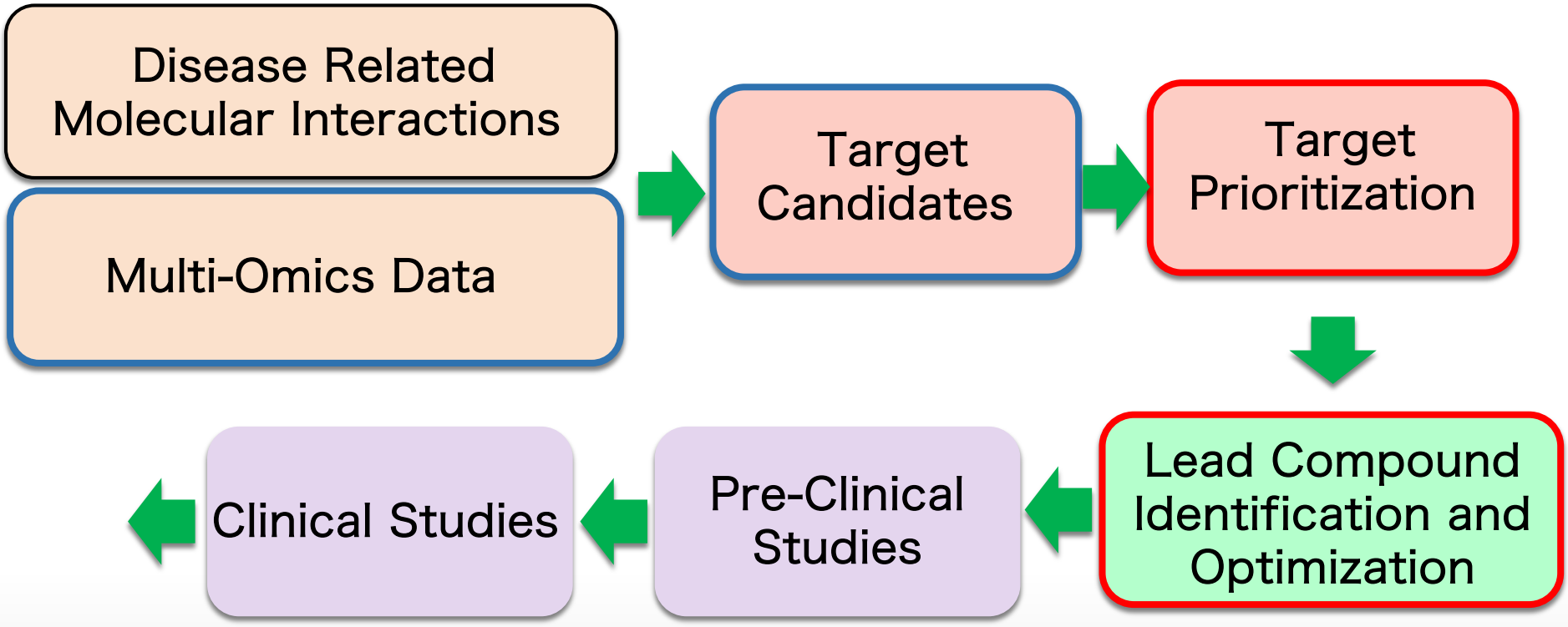
Virus Associated Molecules

Cell Host & Microbe 2014 16, 795-805DOI: (10.1016/j.chom.2014.11.002)

Evaluating compounds that targets host factors

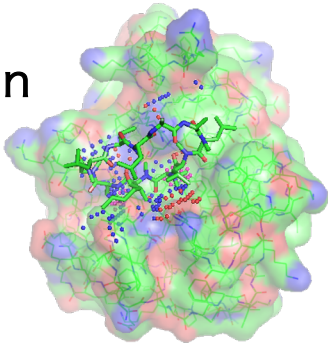


Very Simplified Process of System-Driven Drug Discovery

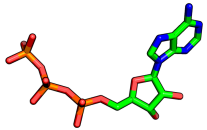


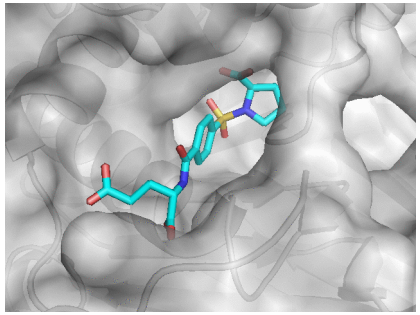
Docking Simulation

Target protein

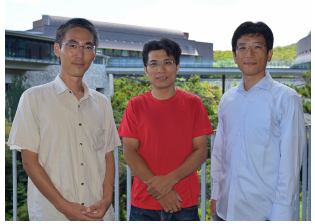


Test compound



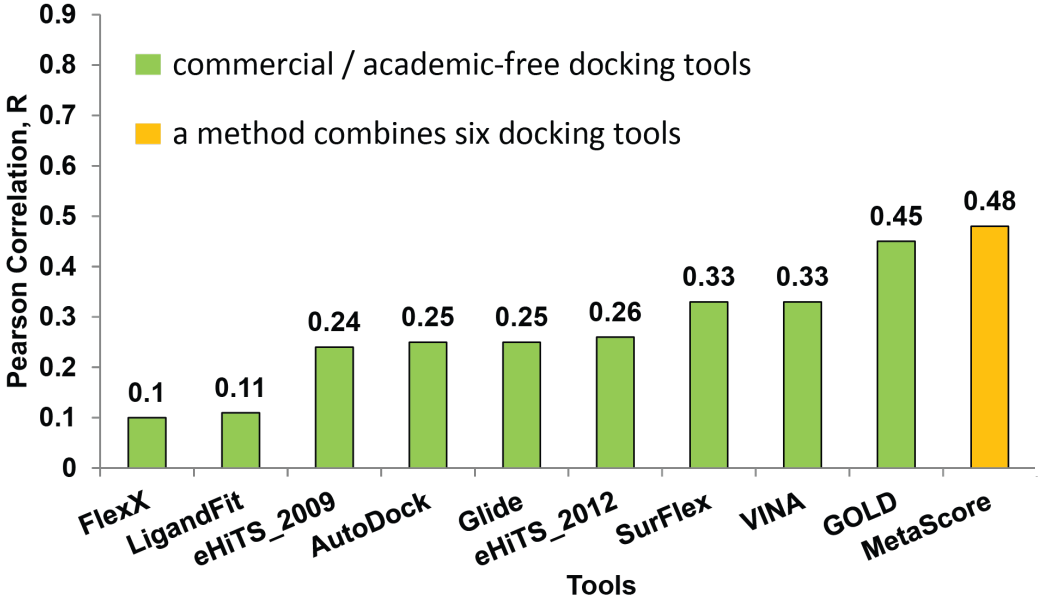


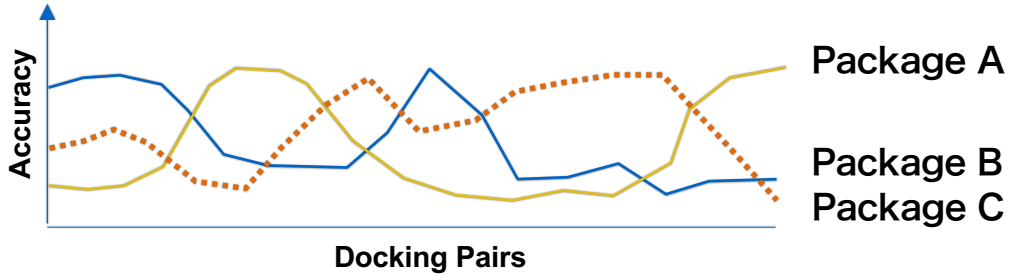
- Posing
- Scoring



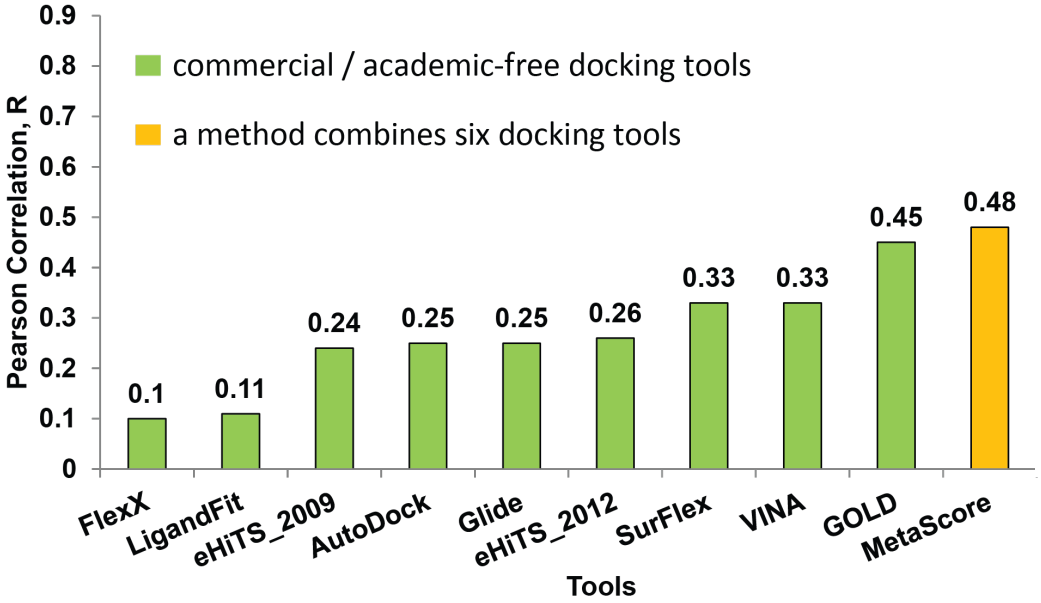
Hsin, Kamiyoshi, Asai
@ OIST

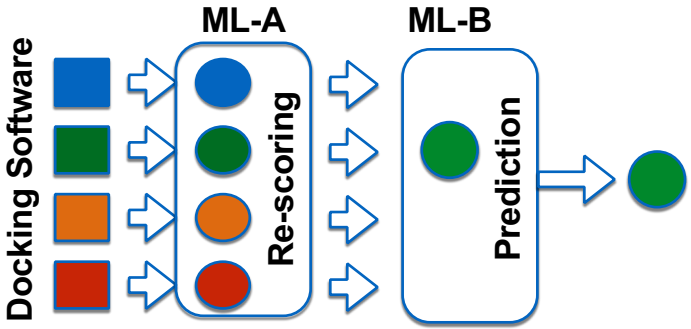
Performance Evaluation of Docking Simulation





Performance Evaluation of Docking Simulation

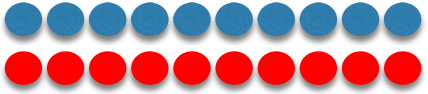




Performance Evaluation using Kinase Inhibitors

Selectivity = 1.0

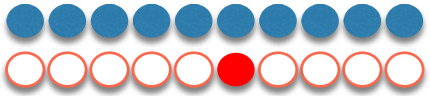
Target Molecules Tested



Kd < 3uM Interactions

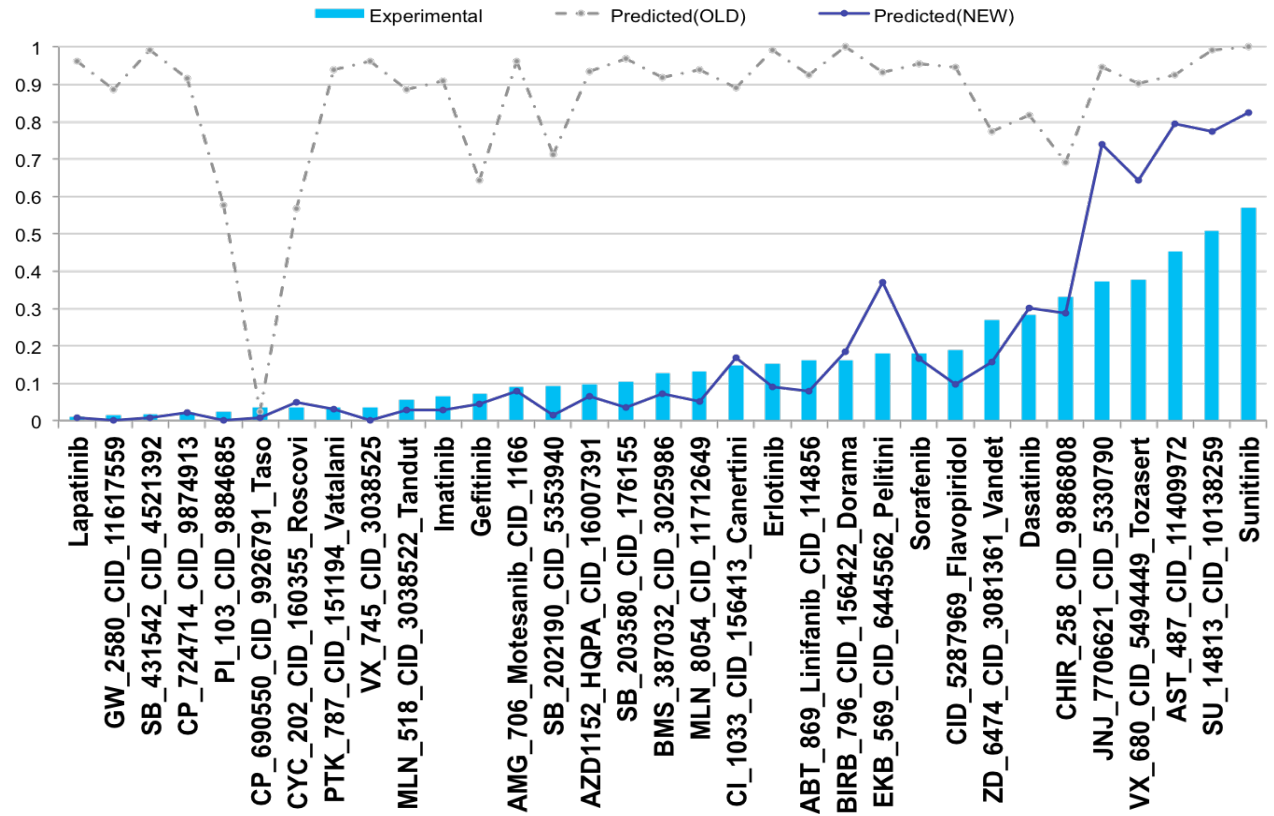
Selectivity = 0.1

Target Molecules Tested



Kd < 3uM Interactions

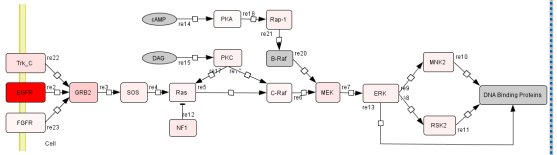
X-axis (selectivity score) = number of interactions with Kd < 3 uM / number of kinases tested



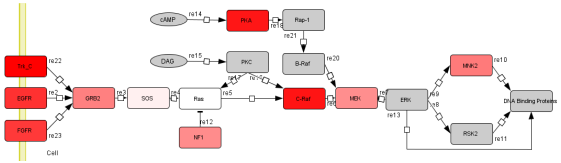
K.Y.Shin, Ghosh, Kitano, PLoS One, 2013

A case study on a simple MAPK pathway

Lapatinib

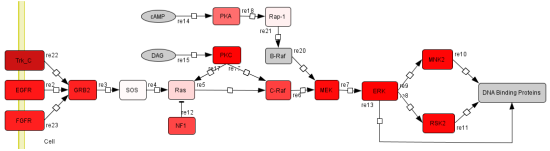


Bioassay contrast



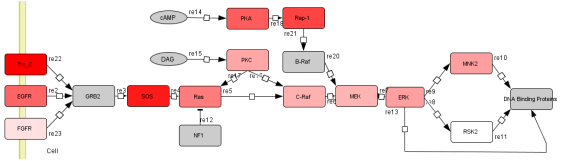
eHITS

Sunitinib

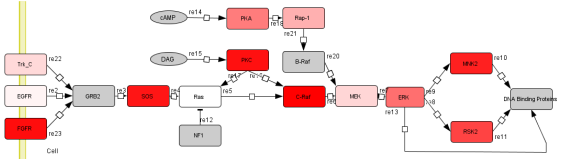


Bioassay contrast

1
0
0
pKd/pK
i



GOLD

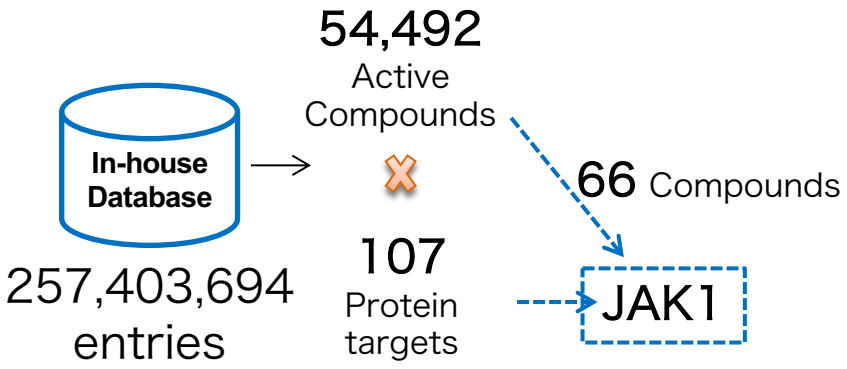


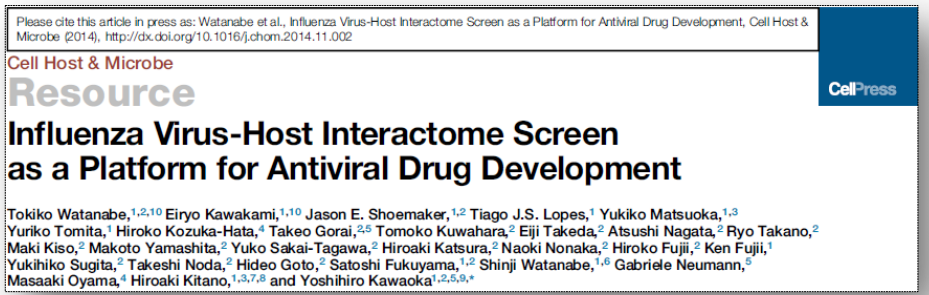
AutoDoc-VINA

Our Method

Bioassay data is from Karaman, et al., Nature Immunology, 10 356-360, 2009

Computational Screening → Experimental Validation




Please cite this article in press as: Watanabe et al., Influenza Virus-Host Interactome Screen as a Platform for Antiviral Drug Development, Cell Host & Microbe (2014), http://dx.doi.org/10.1016/j.chom.2014.11.002

Cell Host & Microbe

Resource

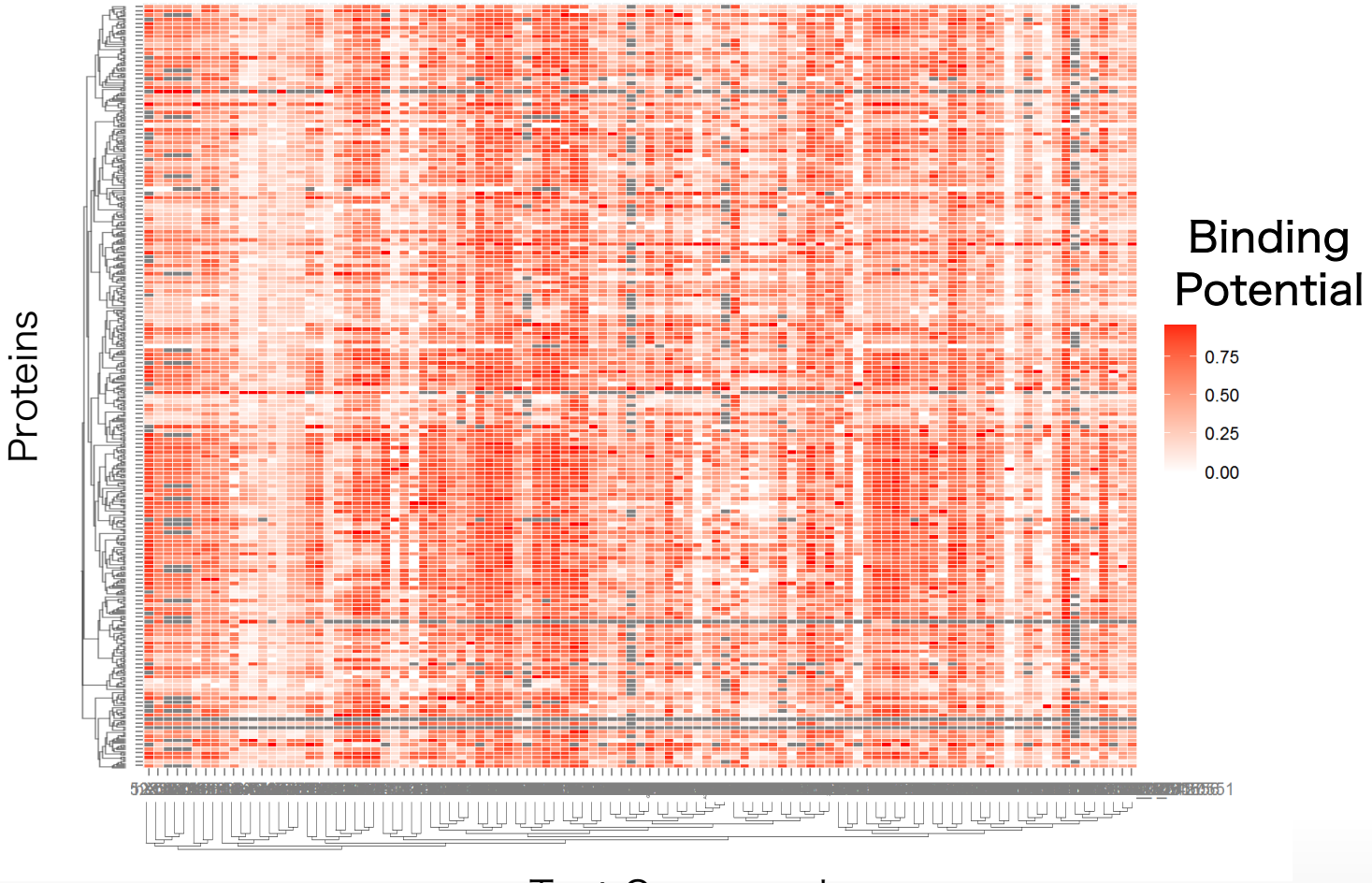
CellPress

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

Tokiko Watanabe,[1,2,10] Eiryo Kawakami,[1,10] Jason E. Shoemaker,[1,2] Tiago J.S. Lopes,[1] Yukiko Matsuoka,[1,3] Yuriko Tomita,[1] Hiroko Kozuka-Hata,[4] Takeo Gorai,[2,5] Tomoko Kuwahara,[2] Eiji Takeda,[2] Atsushi Nagata,[2] Ryo Takano,[2] Maki Kiso,[2] Makoto Yamashita,[2] Yuko Sakai-Tagawa,[2] Hiroaki Katsura,[2] Naoki Nonaka,[2] Hiroko Fujii,[2] Ken Fujii,[1] Yukihiko Sugita,[2] Takeshi Noda,[2] Hideo Goto,[2] Satoshi Fukuyama,[1,2] Shinji Watanabe,[1,6] Gabriele Neumann,[5] Masaaki Oyama,[4] Hiroaki Kitano,[1,3,7,8] and Yoshihiro Kawaoka[1,2,5,9,*]

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

Results in Color Matrix



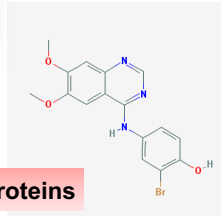
The Institute of Medical Science
The University of Tokyo

Prof. Yoshihiro Kawaoka

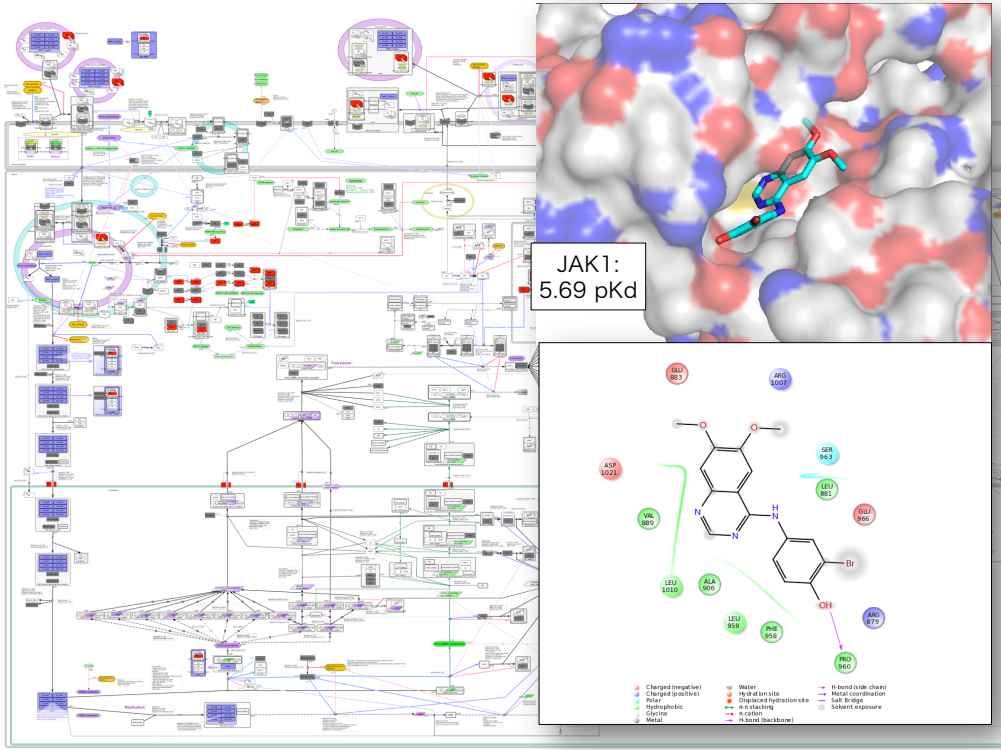
FluMap

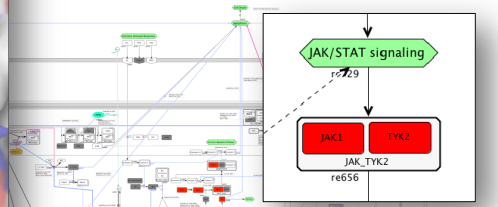
SBI The Systems Biology Institute

CID 3795



18% proteins





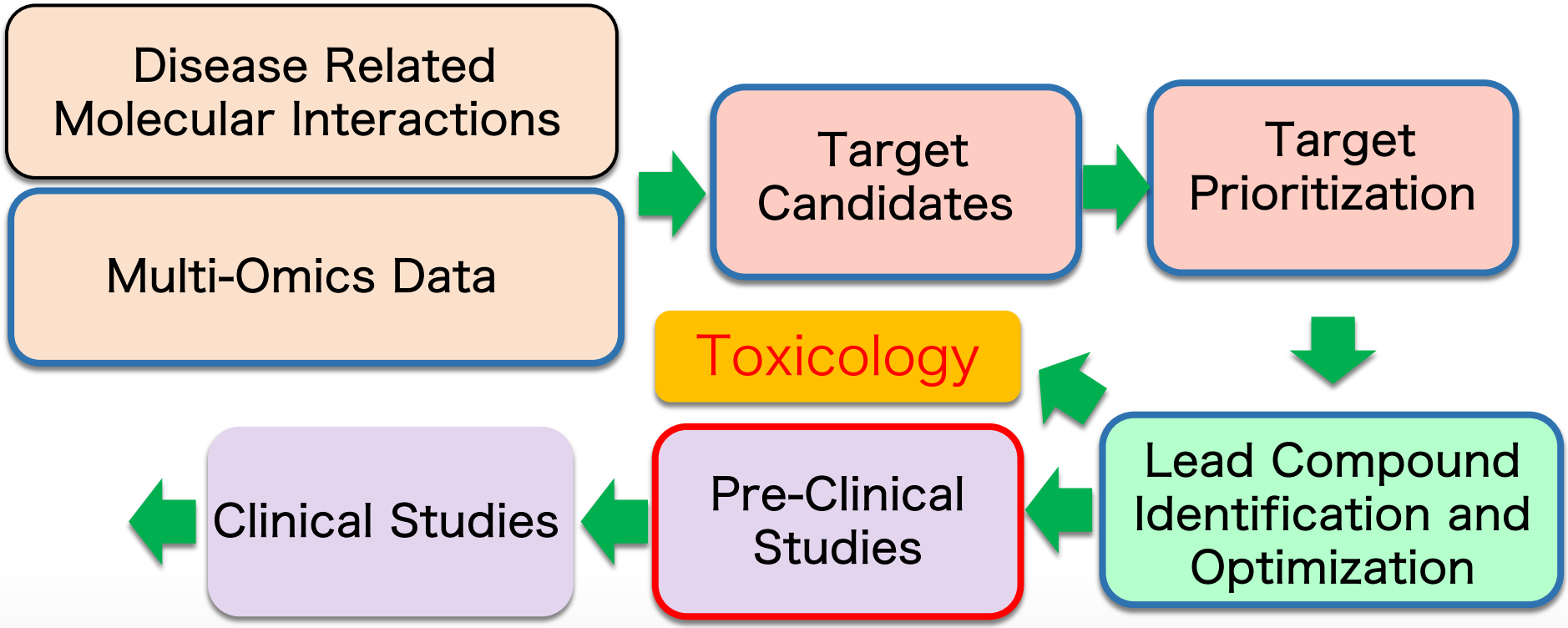
Nucleic Acids Research

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

Kun-Yi Hsin[1,*], Yukiko Matsuoka[2], Yoshiyuki Asai[1], Kyota Kamiyoshi[1], Tokiko Watanabe[3], Yoshihiro Kawaoka[3,4] and Hiroaki Kitano[1,2,5,*]

April 29, 2016

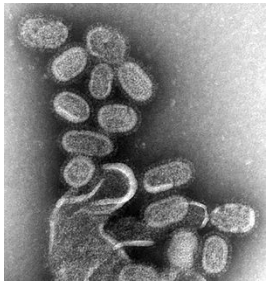
Very Simplified Process of System-Driven Drug Discovery



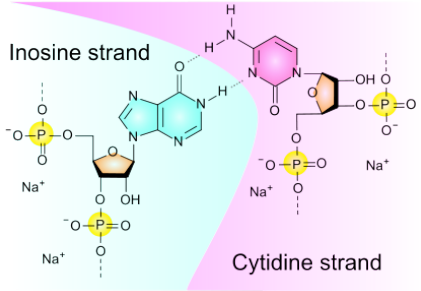
Uncovering Mechanism of Action of Maoto (麻黄湯) for Influenza

Complexity：100s X 1000s X 1000s X …

27



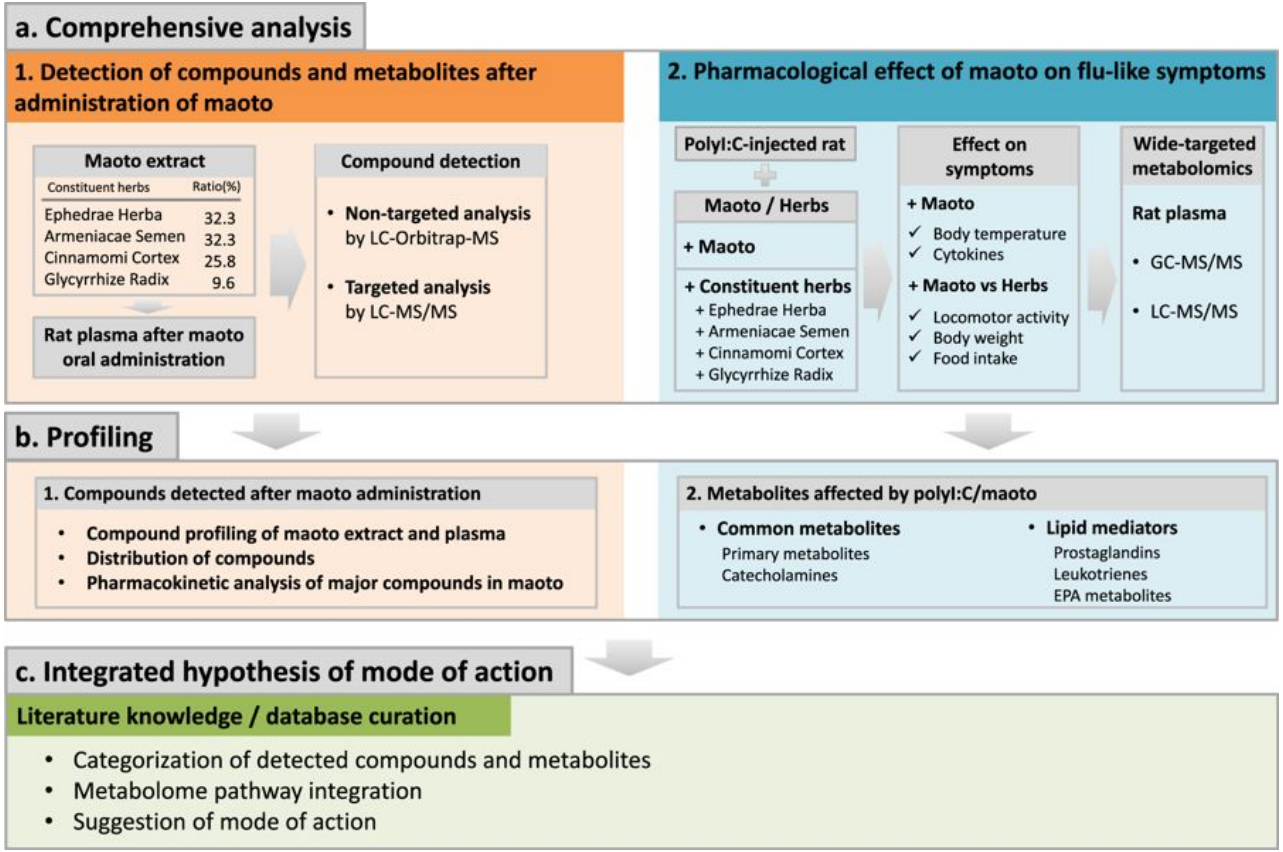
Influenza Virus



Poly(I:C)

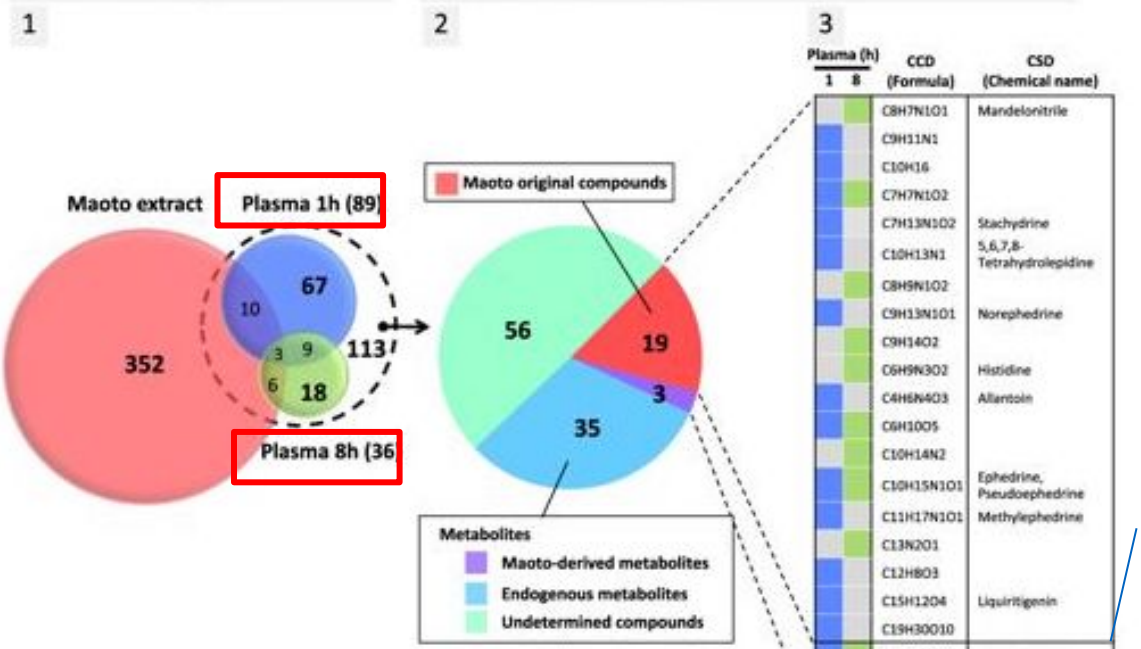
(Polyinosinic-polycytidylic acid sodium salt)

Rat



Nishi, A. et al., Deconstructing the traditional Japanese medicine “Kampo”: compounds, metabolites and pharmacological profile of maoto, a remedy for flu-like symptoms, *npj Systems Biology and Applications,* volume 3, Article number: 32 (2017)

Measuring Maoto Extract and Rat Plasma Metabolites



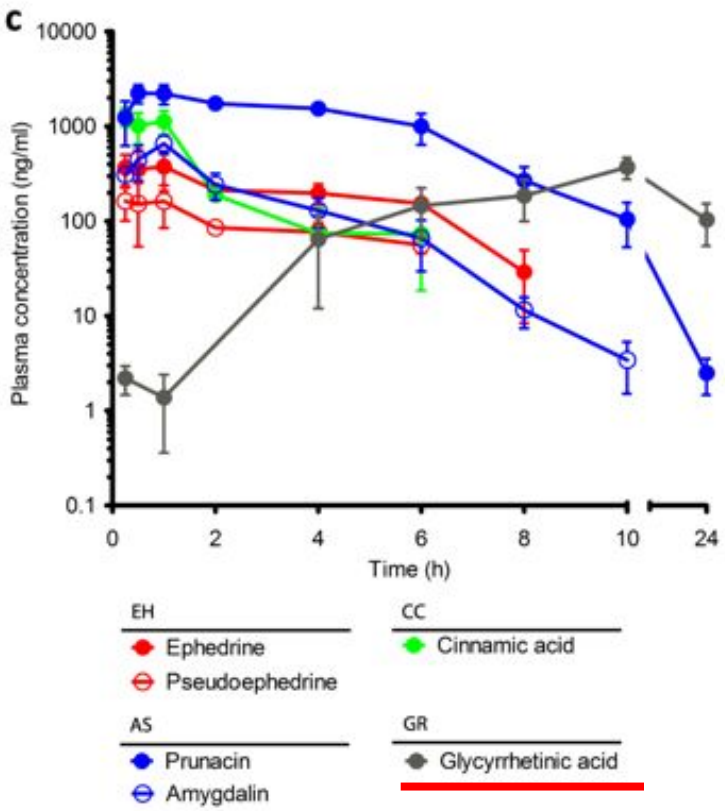
Number of CCDs in Maoto extract and rat plasma

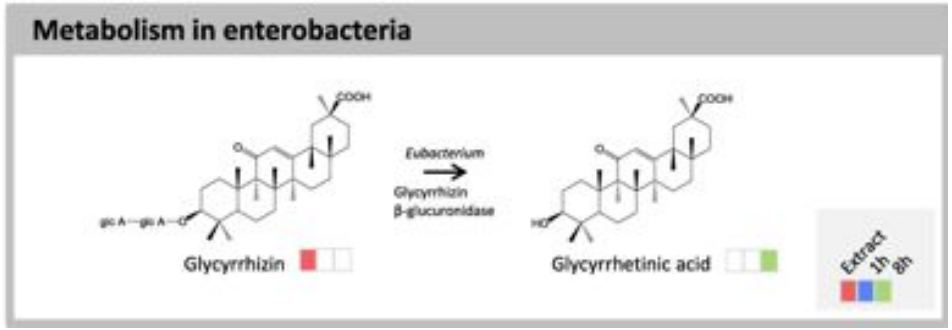
CCD = Composition determined compound

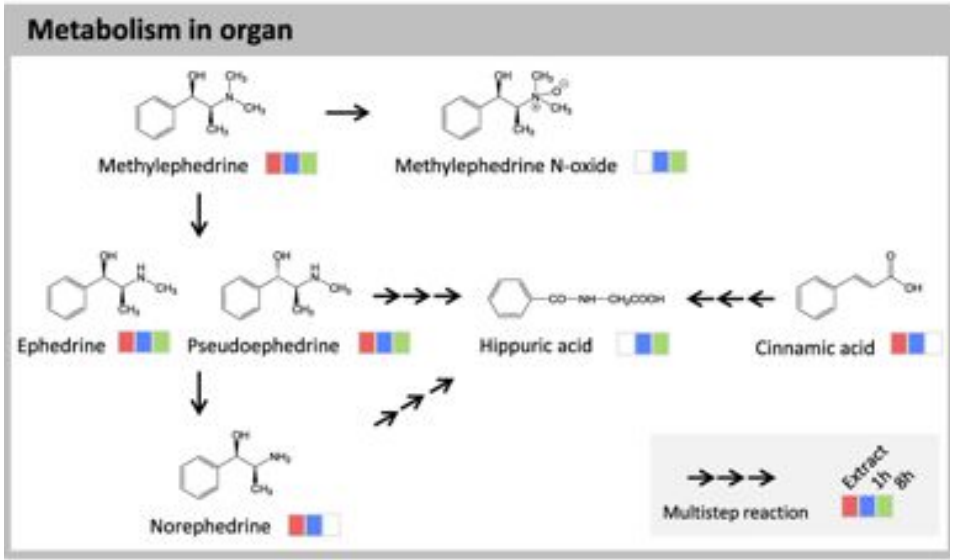
113 compounds detected in Rat plasma after Maoto treatment

Hippuric acid
(馬尿酸)
Methylephedrine N-Oxide
(メチルエフェドリン Nオキシド)
Glycyrrhetinic acid
(グリチルレチン酸)

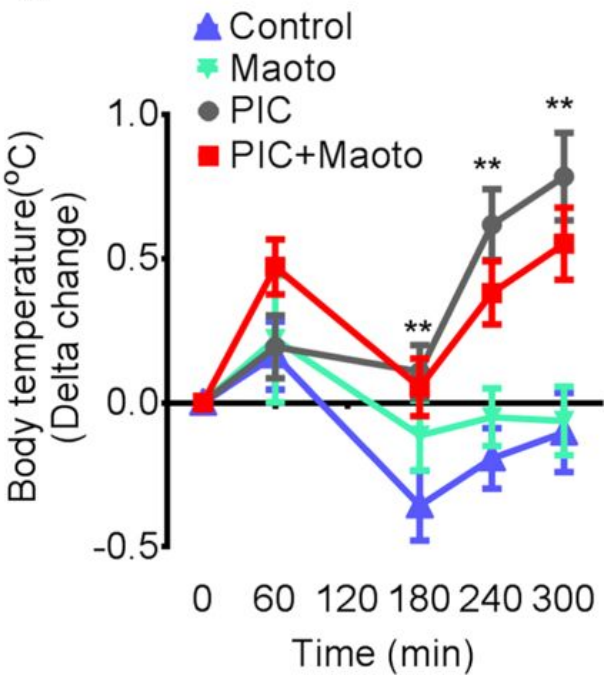
Nishi, A. et al., Deconstructing the traditional Japanese medicine “Kampo”: compounds, metabolites and pharmacological profile of maoto, a remedy for flu-like symptoms, *npj Systems Biology and Applications,* volume 3, Article number: 32 (2017)


c
10000
1000
100
10
1
0.1
Plasma concentration (ng/ml)
0
2
4
6
8
10
24
Time (h)
EH
Ephedrine
Pseudoephedrine
CC
Cinnamic acid
AS
Prunacin
Amygdalin
GR
Glycyrrhetinic acid

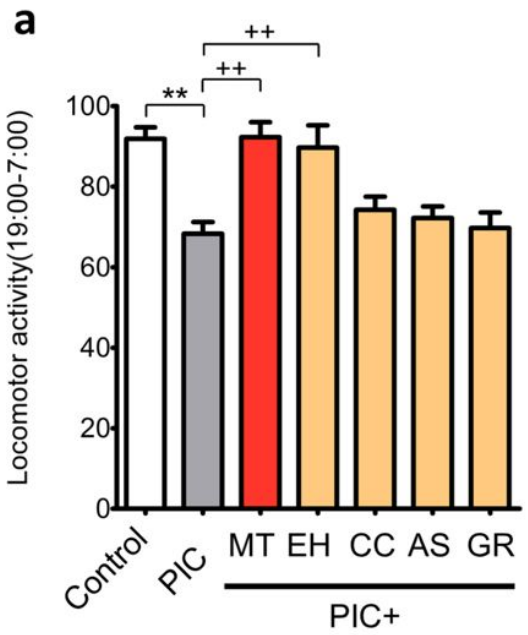

Metabolism in enterobacteria
Glycyrrhizin
Eubacterium
Glycyrrhizin
β-glucuronidase
Glycyrrhetinic acid
Extract
1h
8h


Metabolism in organ
Methylephedrine
Methylephedrine N-oxide
Ephedrine
Pseudoephedrine
Hippuric acid
Cinnamic acid
Norephedrine
Multistep reaction
Extract
1h
8h

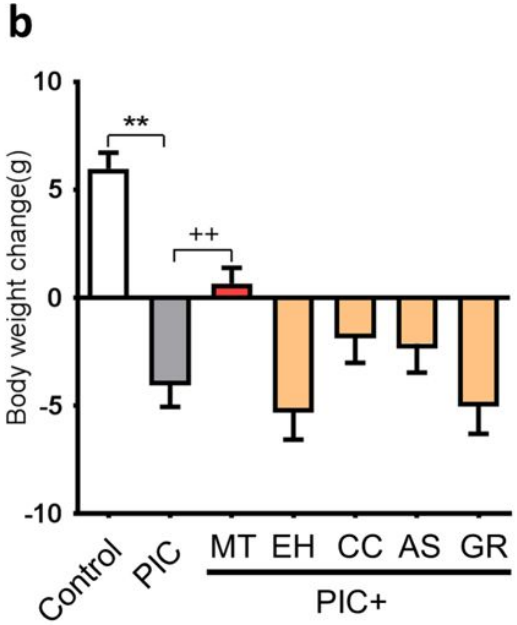
ラットにおける麻黄湯の生理学的影響



体温

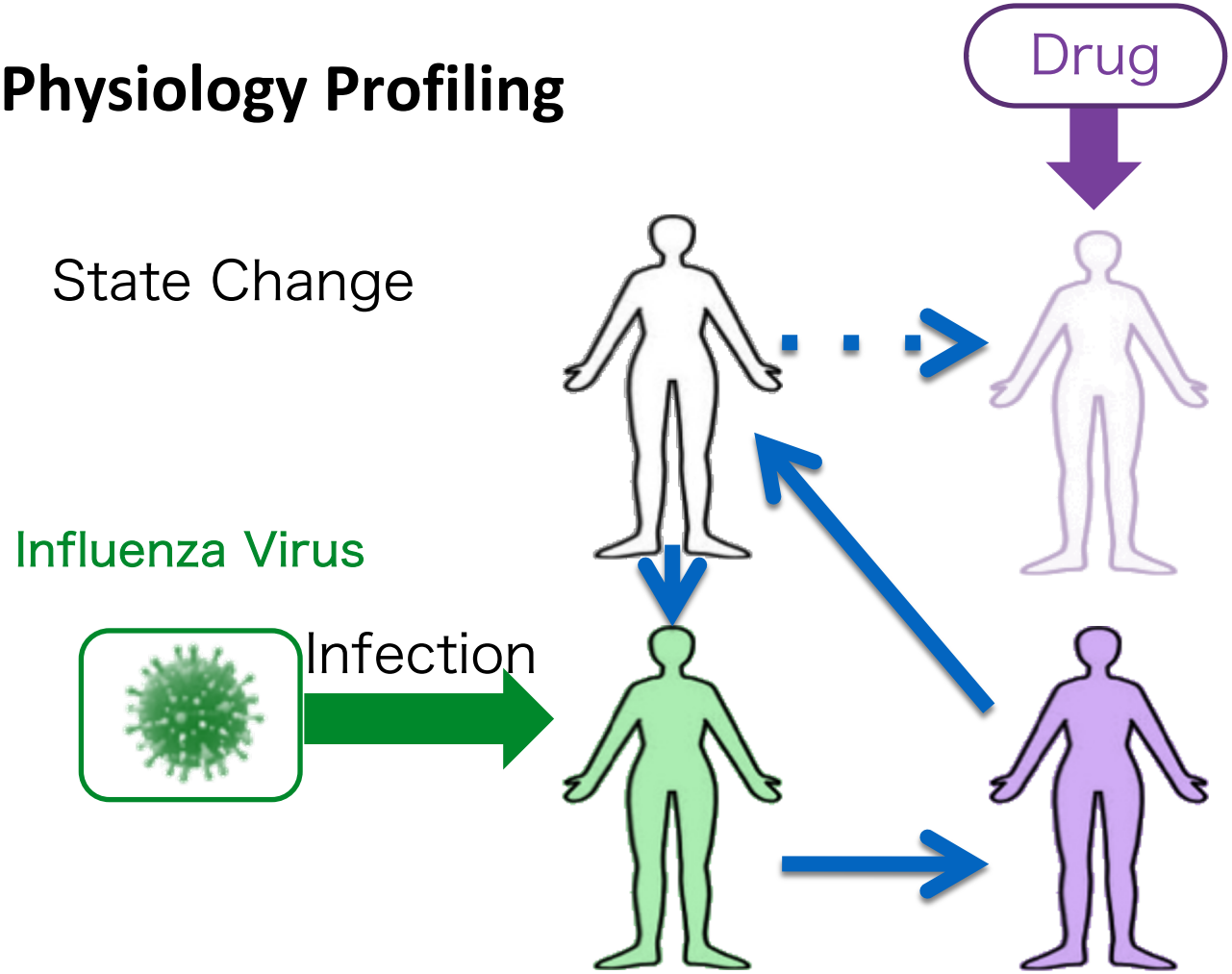


運動量



体重

Host Physiology Profiling

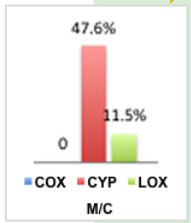


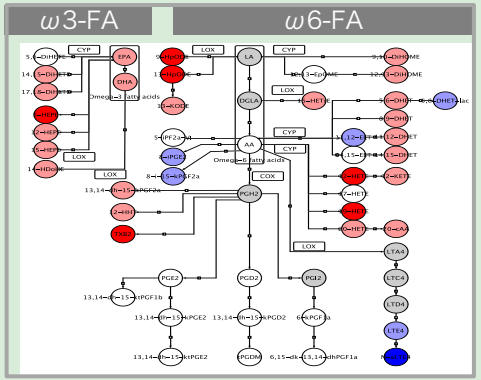
Host Physiology Profiling

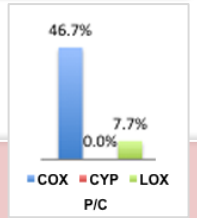
Drug 麻黄湯

Maoto/Control

Control

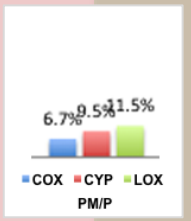


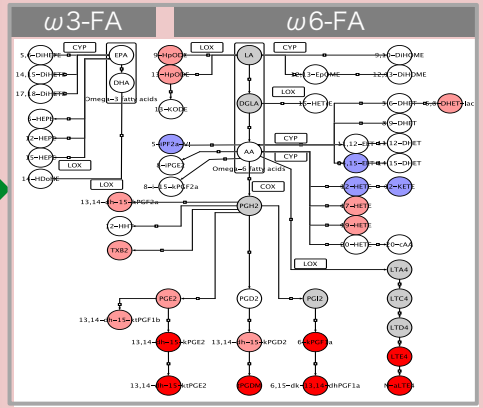


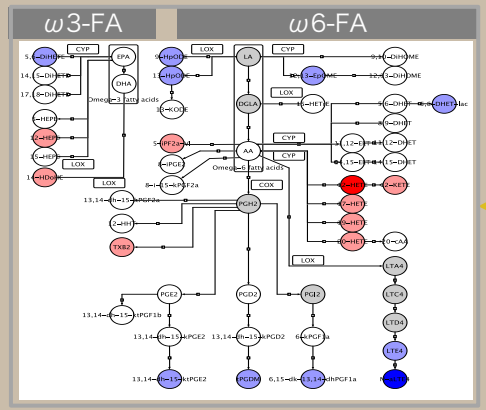


PIC/Control

PIC+Maoto/PIC





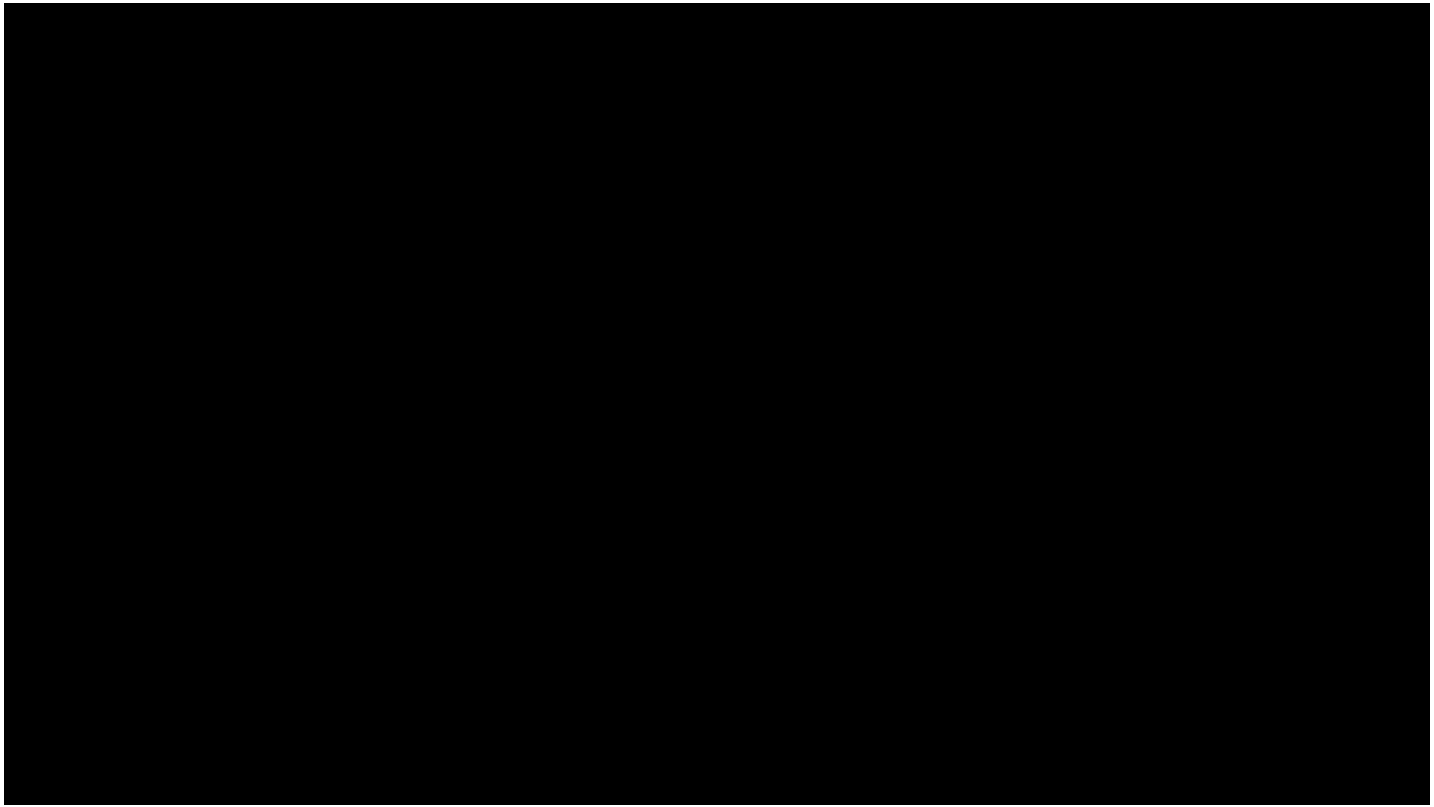


Infection

Poly I:C

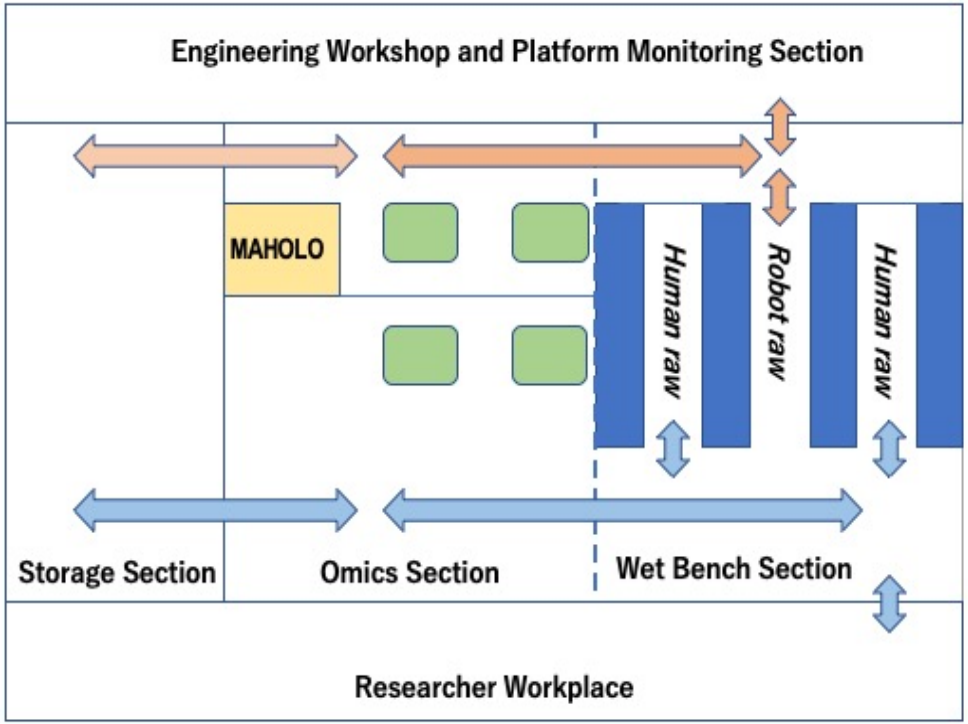
Joint research with Tsumura

Two Arms Robot for Lab



Crowd Robotics Consortium

Automation of Science Project

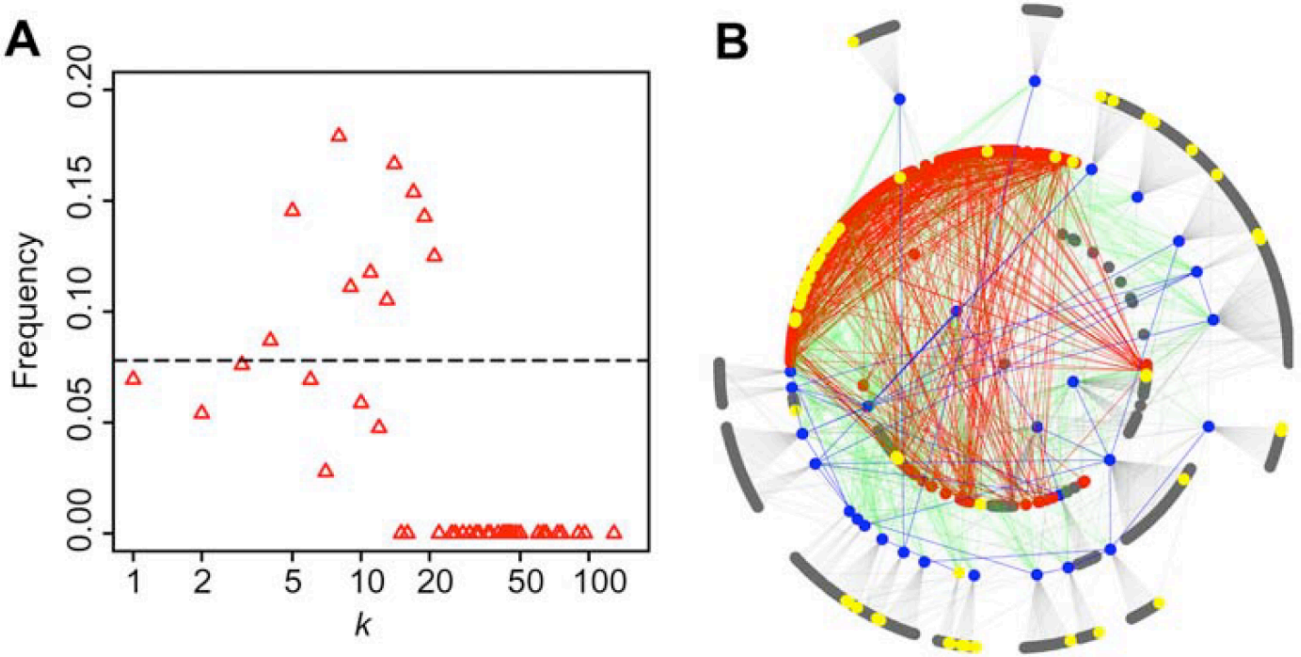


OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Structure of Protein Interaction Networks and Their Implications on Drug Design

Takeshi Hase[1☯], Hiroshi Tanaka[2☯*], Yasuhiro Suzuki[3], So Nakagawa[4], Hiroaki Kitano[5,6,7*]



Nobody can develop an entire pipeline alone

Open Platform

STUDY DESIGNS

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,

nature REVIEWS GENETICS

Nov 2011
Ghosh et al.

GARUDA

Open Collaboration

commentary

Social engineering for virtual 'big science' in systems biology

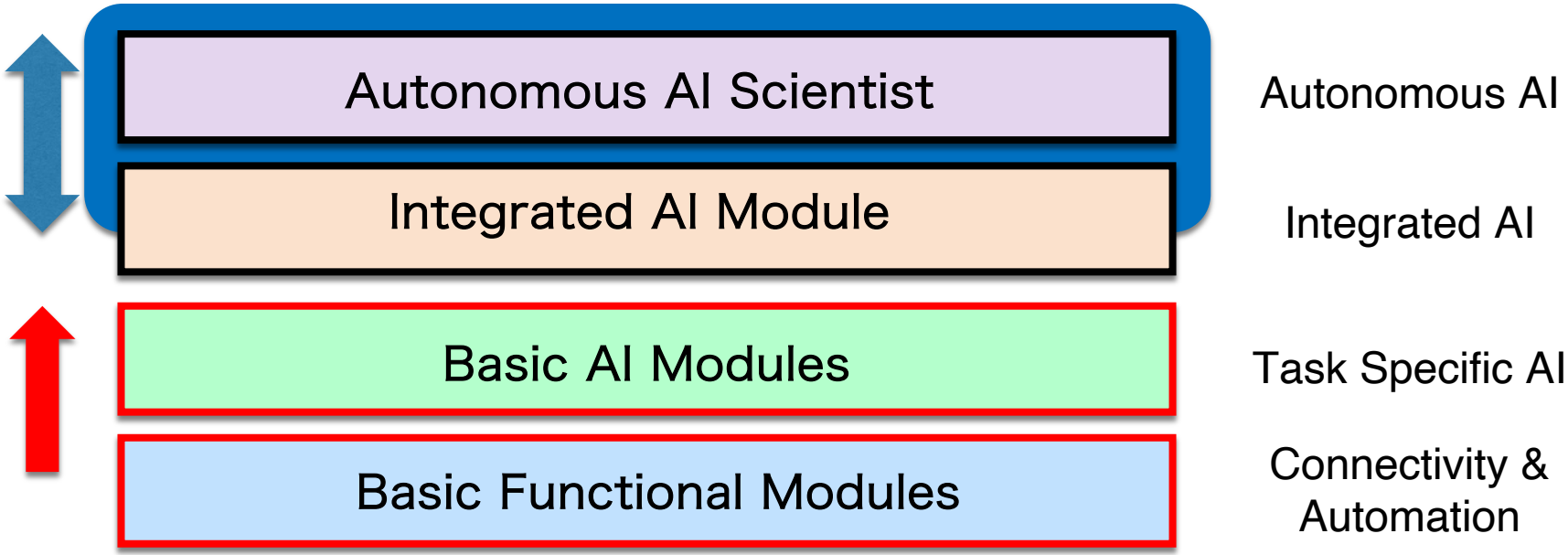
Hiroaki Kitano, Samik Ghosh & Yukiko Matsuoka

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.

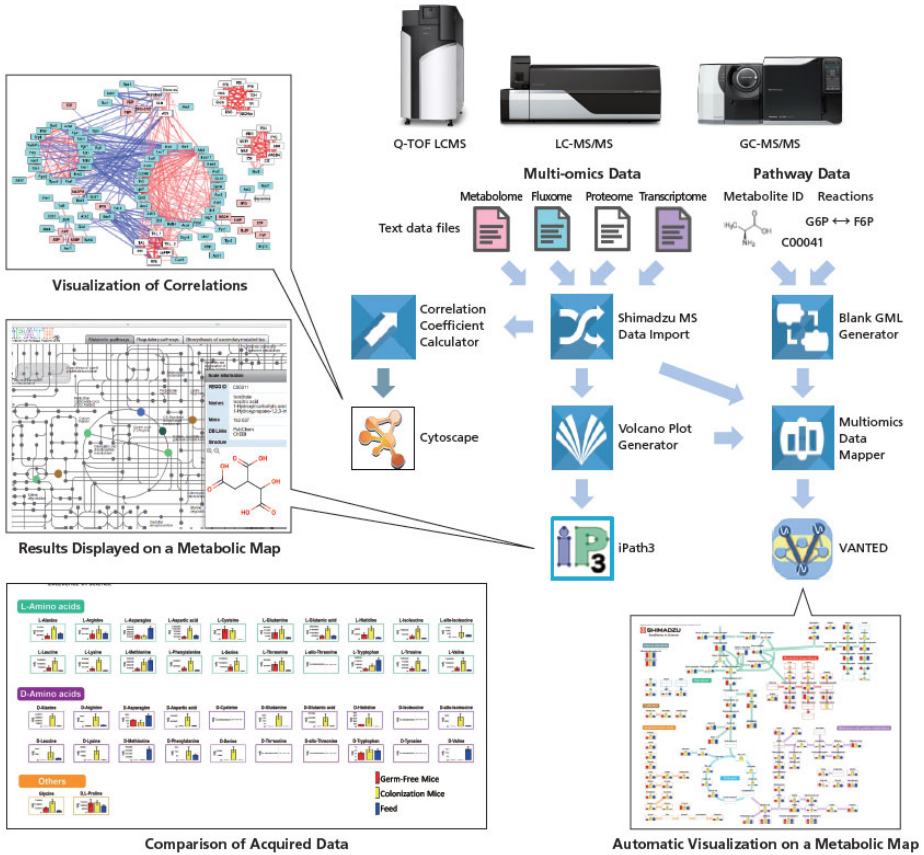
nature chemical biology

May 2011
Kitano, Ghosh, Matsuoka

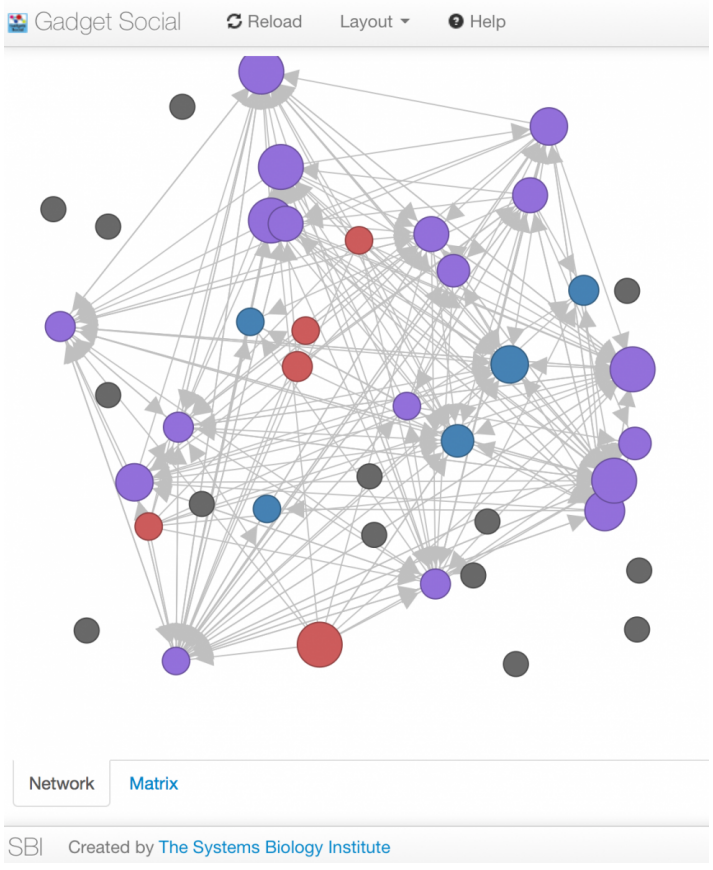
Technology Platform



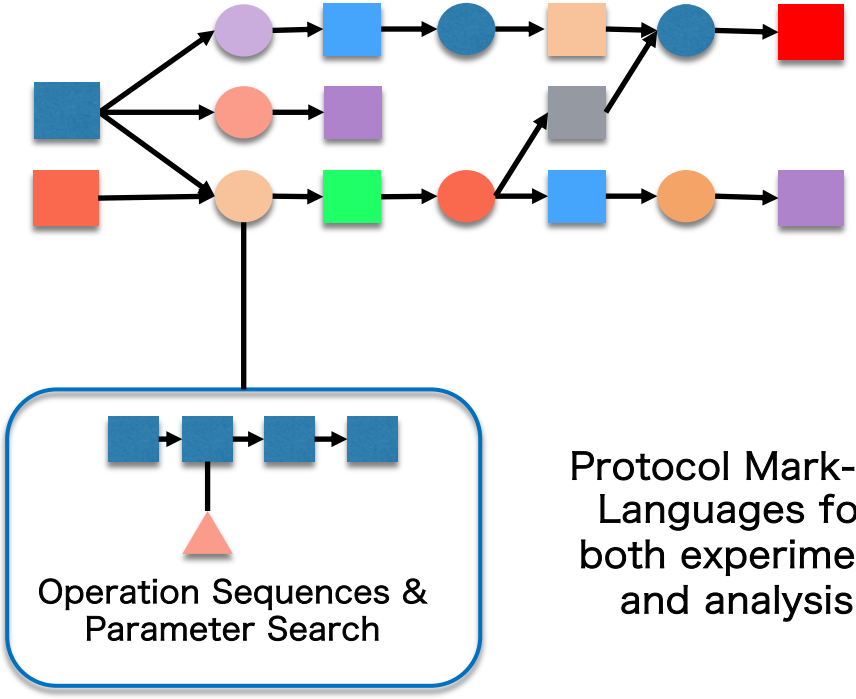
Pre-Defined Sequences



Experiment and Analysis Pipeline



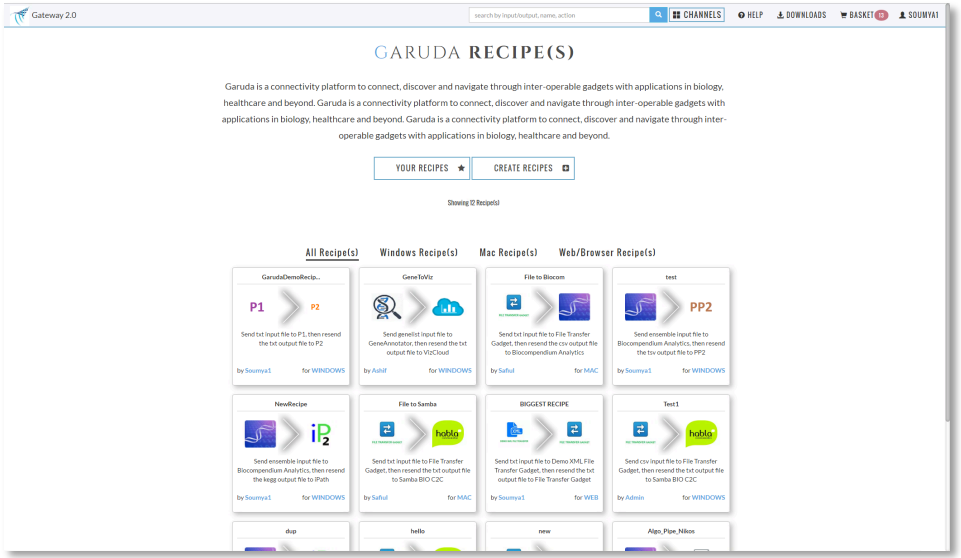
End-to-End Gadget Path



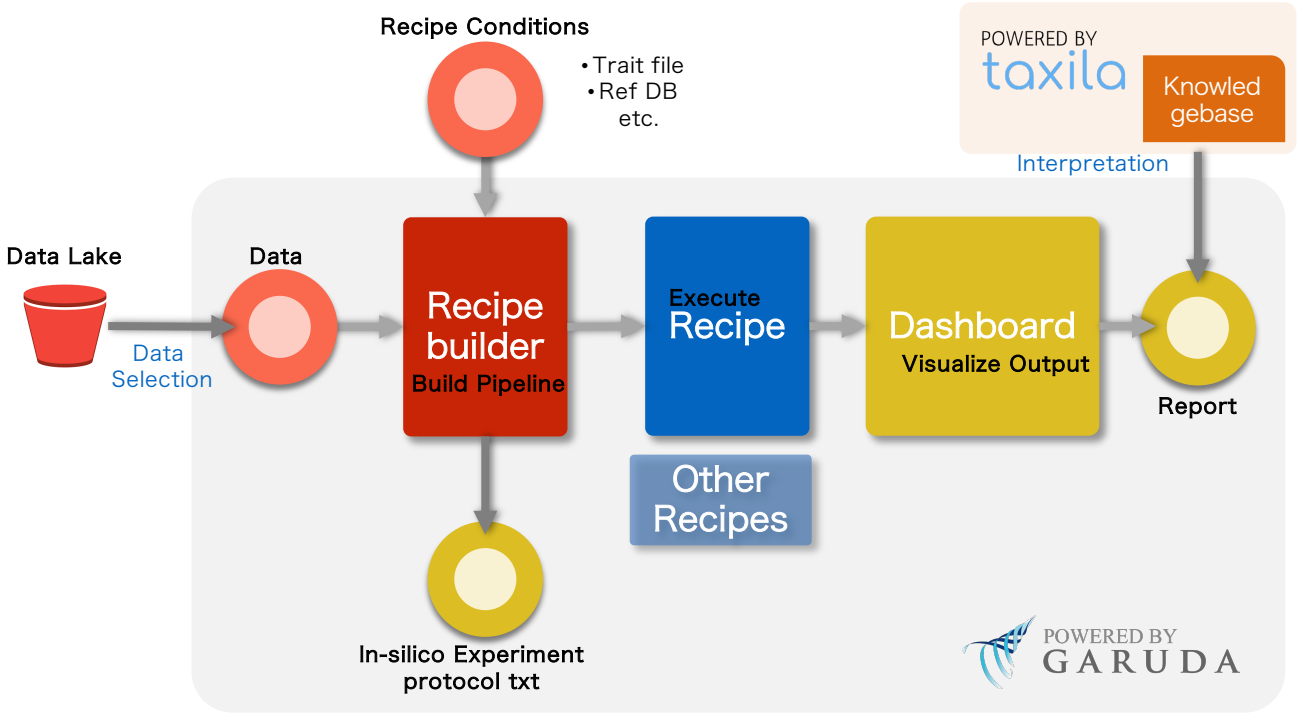
Protocol Mark-up Languages for both experiment and analysis

GARUDA 2x

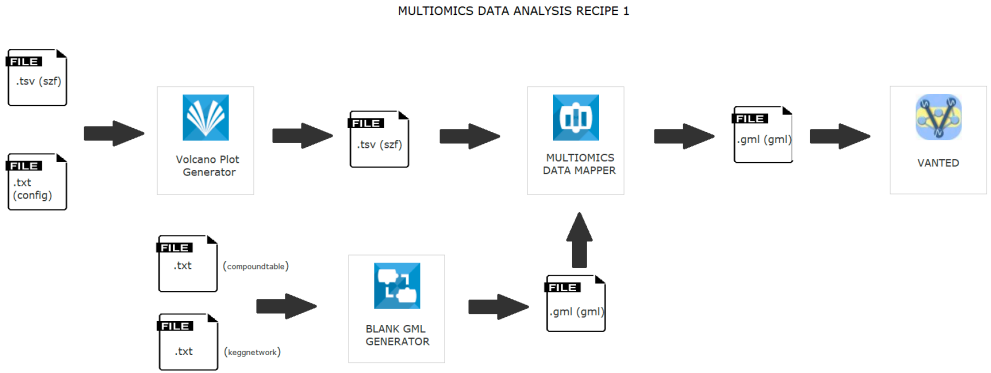
Automation and Recipes: Support for multi-device and cross-device automation of processes and workflows through “Garuda Recipes”



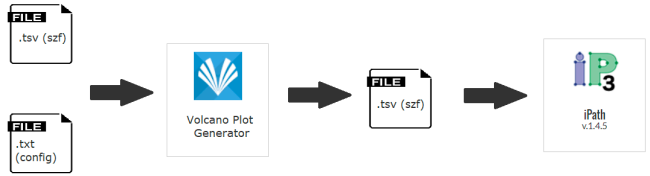
GarudaとTaxilaを組み込んだ In-silico Experimental Platform



Recipe 1 steps:



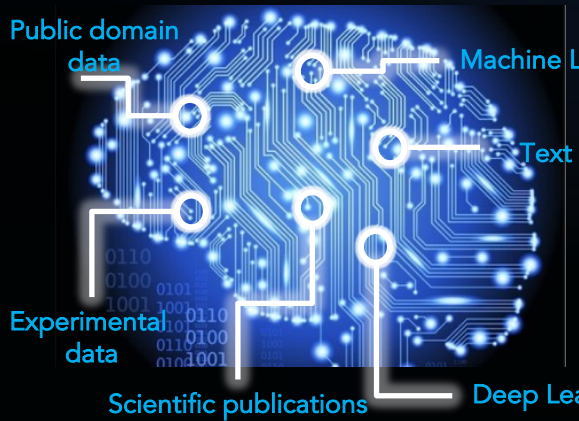
Recipe 2 steps:



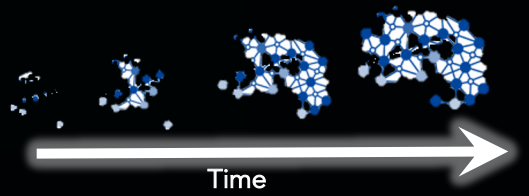
Taxila

- 次世代ナレッジ集積インテリジェンス解析プラットフォーム
- **Convert text to actionable insights:**
  - Combine powerful aggregation, search, text mining, machine learning and AI technologies
    - 多次元データと解析をつなぐクラウドベースのプラットフォーム
    - 新しいデータや解析手法をダイナミックに構築可能なアーキテクチャ
    - 対象分野情報と解析を提供し迅速な意思決定を支援

taxila



A "Living" system which *grows* and *learns* by forming new connections

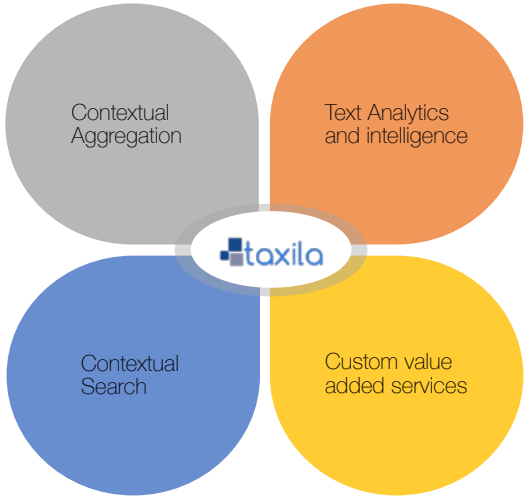


SBX

taxila

center of learning :
text to knowledge

Convert text to actionable insights: combine powerful aggregation, search, text mining, machine learning and AI technologies.



Taxila provides automatic context-aware aggregation and search of relevant information, driven by AI-powered mining and analytics engine for driving actionable insights with a intuitive user interface.

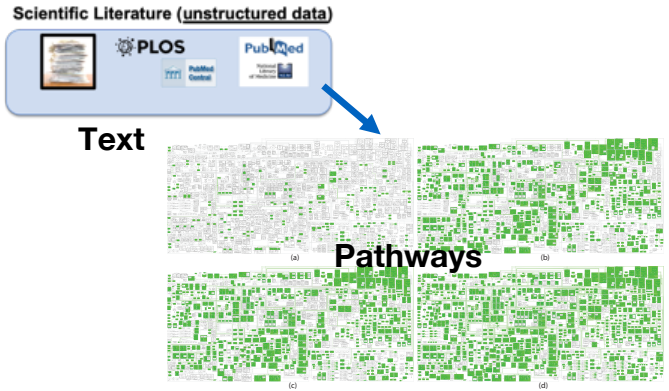
Industry and Research Case Studies

POWERED BY taxila

New Hypothesis Generation	Real World Evidence Analysis	Drug Repurposing	
Context specific knowledge base creation and Concept/ Relationship Identification	Biomarker Discovery, Target identification, selection	Risk Assessment in Genotoxicity prediction and assessment	Text driven network/ pathway reconstruction
Summarization, Visualization and interpretation of large knowledge bases	Trend Analysis/ Anomaly detection	Key Opinion Leader Identification	Sentiment Analysis, Patent and Publication Analytics, Social Media Analytics

Text driven network/ pathway reconstruction

Unstructured Data -> Structured representations -> Insights



RWE analysis for novel understanding of Neurological diseases (Unstructured Data -> Insights)

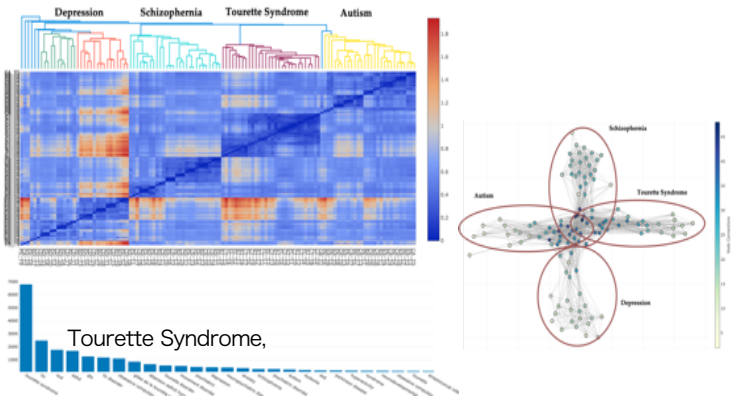
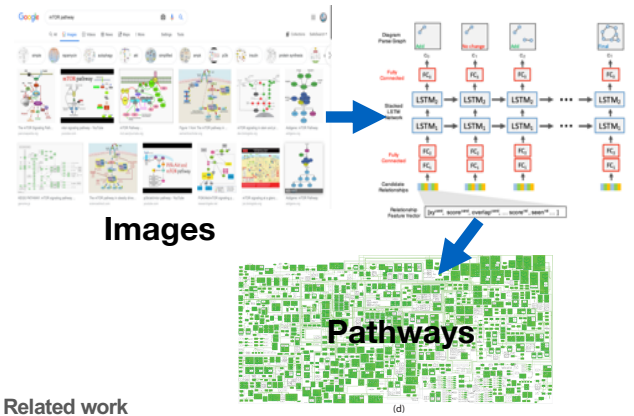


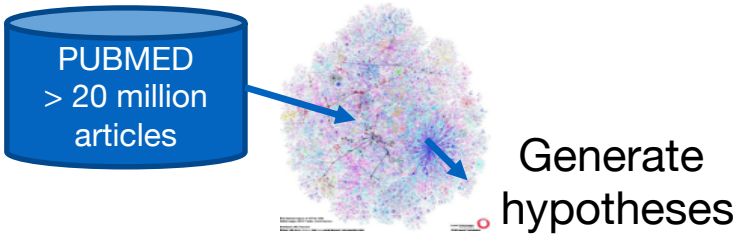
Image + Text based Pathway Reconstruction

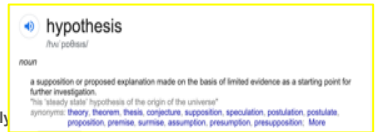


Related work

Kembhavi, Aniruddha, et al. "A diagram is worth a dozen images." *European Conference on Computer Vision*. Springer, Cham, 2016.

PUBMED scale hypothesis (link prediction) hypothesis generation





Related work

Sybrandt, Justin, Michael Shtutman, and Ilya Safro. "Moliere: Automatic biomedical hypothesis generation system." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017.

Real World Evidence Analysis

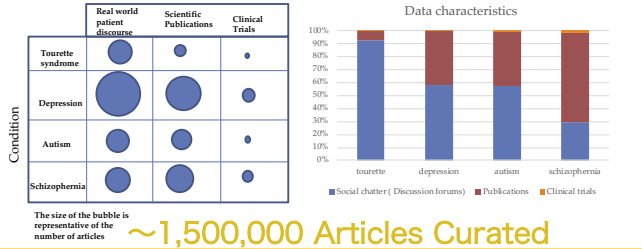
Joint work with major pharmaceutical company

Comprehensive real world evidence collection, curation and analysis to understand novel mechanism of specific neurological conditions to uncover new hypotheses.

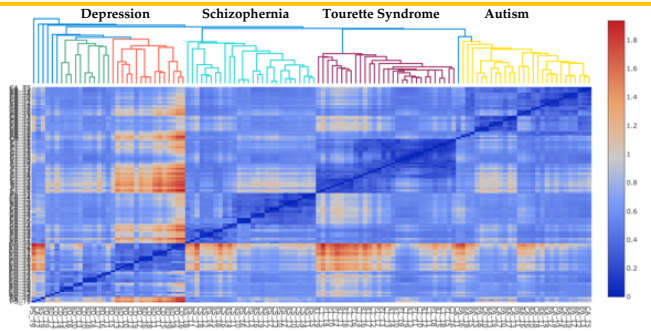
Data source: Publications, Blogs, Discussion Forums, Clinical Trials, Social Media

Focus: Tourette Syndrome, Depression, Autism and Schizophrenia

Large scale data collection

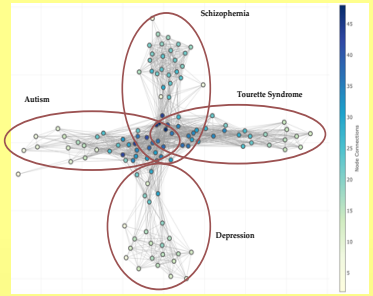


~1,500,000 Articles Curated



Text Analysis, NLP and Machine Learning

Novel insights and new hypothesis generation

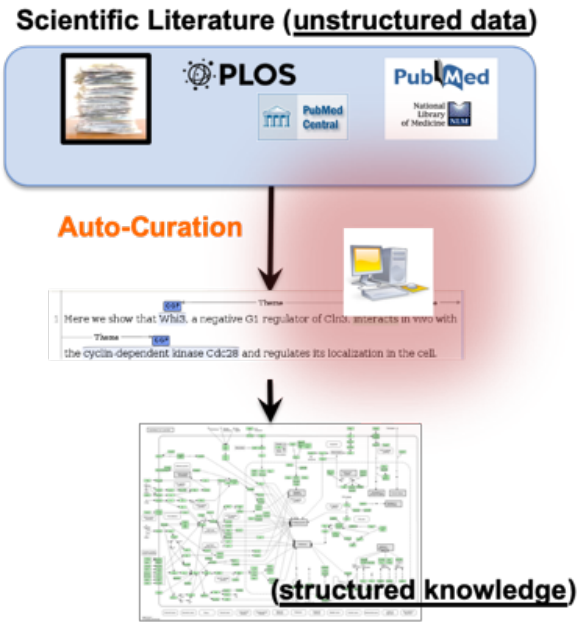


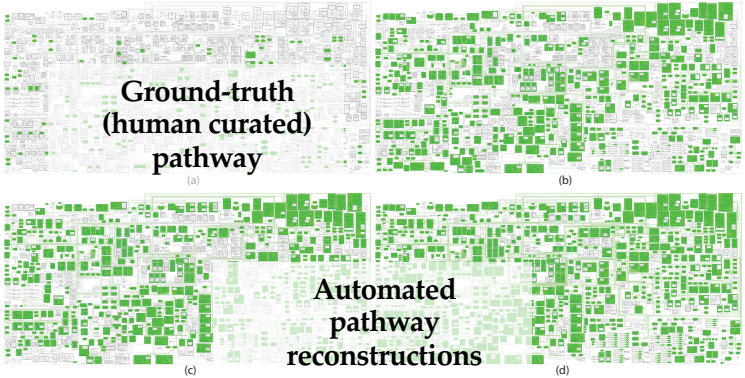
Outcome: Novel links between 'Tourette' and Schizophrenia

Text driven network/ pathway reconstruction

To Structured Representations from Unstructured text

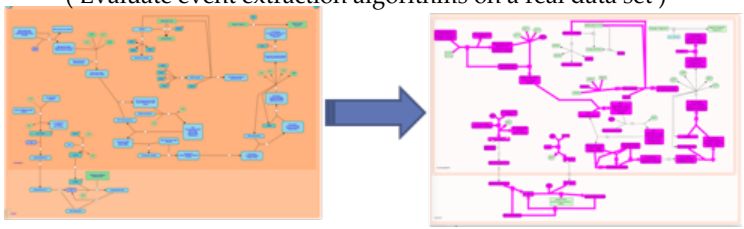
Reconstructing structured pathway representations from unstructured text





Future Reconstruction Challenge.

(Evaluate event extraction algorithms on a real data set)



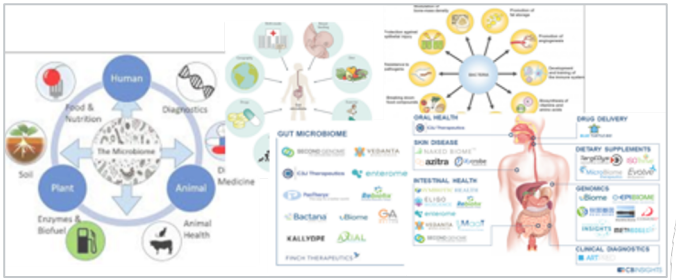
Spranger, M., Palaniappan, S., Ghosh, S.: Measuring the state of the art of automated pathway curation using graph algorithms - a case study of the mtor pathway. In: Proceedings of the 2016 Workshop on Biomedical Natural Language Processing (BioNLP 2016), pp. 119–128. ACL

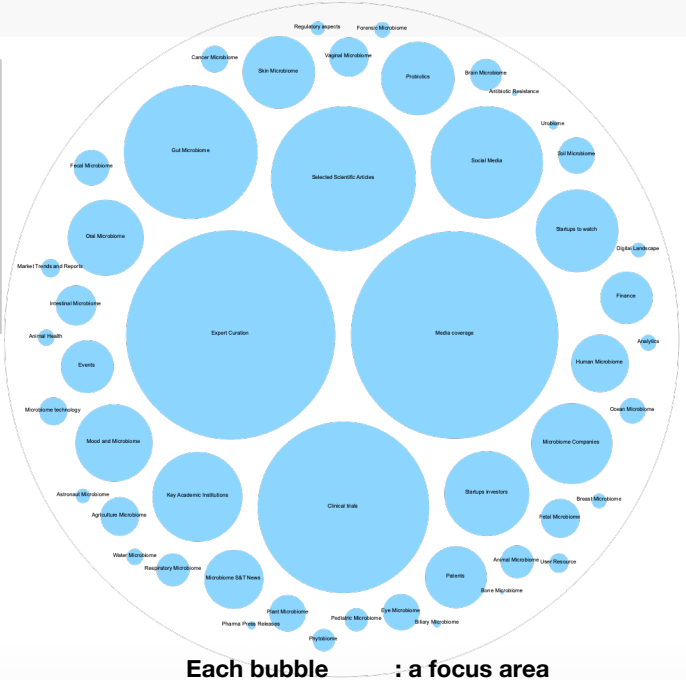
M. Spranger, S. Palaniappan, and S. Ghosh. Extracting biological pathway models from nlp event representations *BioNLP 2015*, pages 42—51. ACL, 2015

Microbiome Landscape Analysis

Work for due diligence analysis

- Collect, curate and comprehend opportunities in science, technology and business dimensions in the areas of personal omics with specific focus on microbiome
- Deep analysis of state of the horizon in these areas conducted based on the Taxila platform to identify strength, weaknesses, opportunities and threats (SWOT) for business in this domain



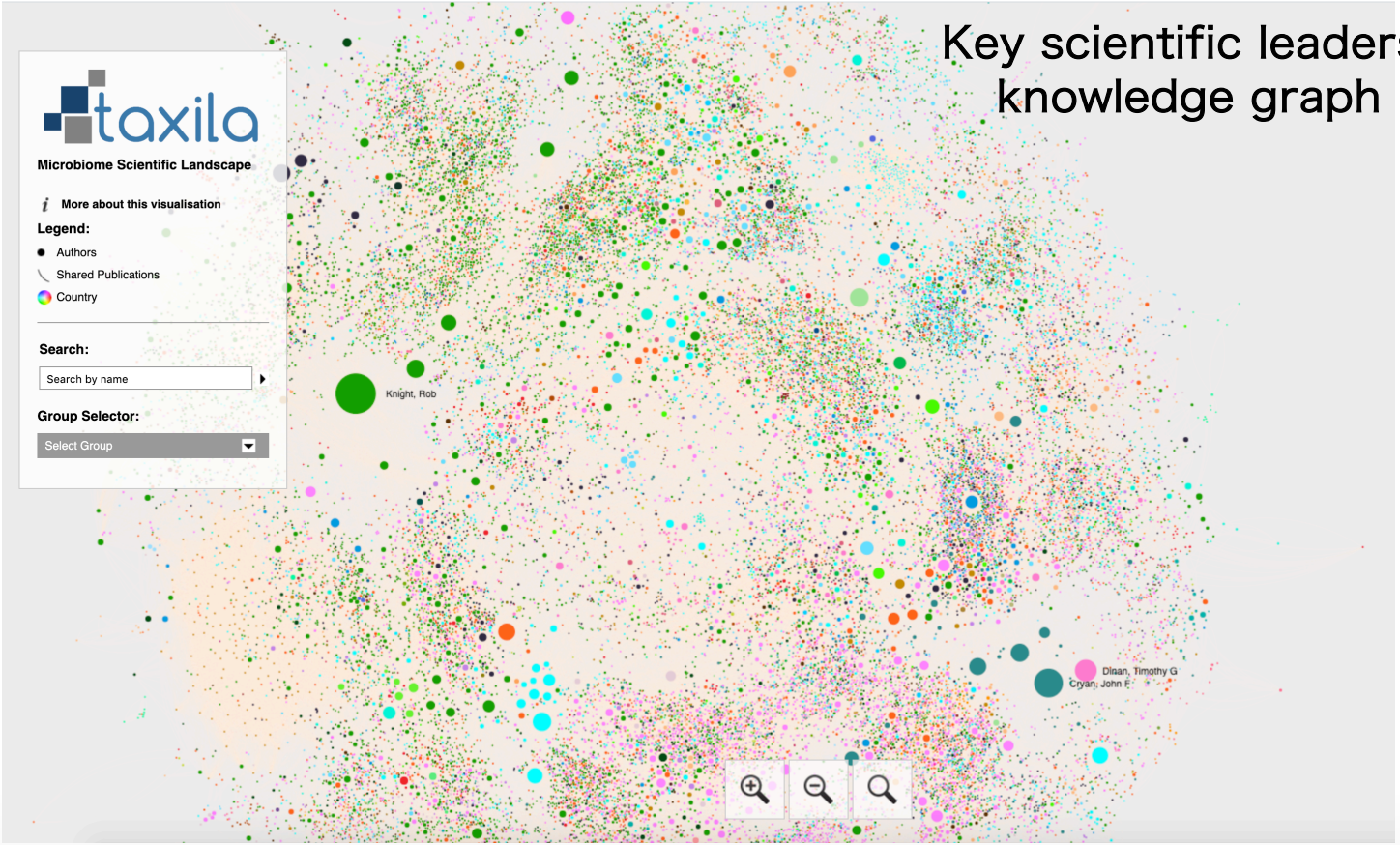


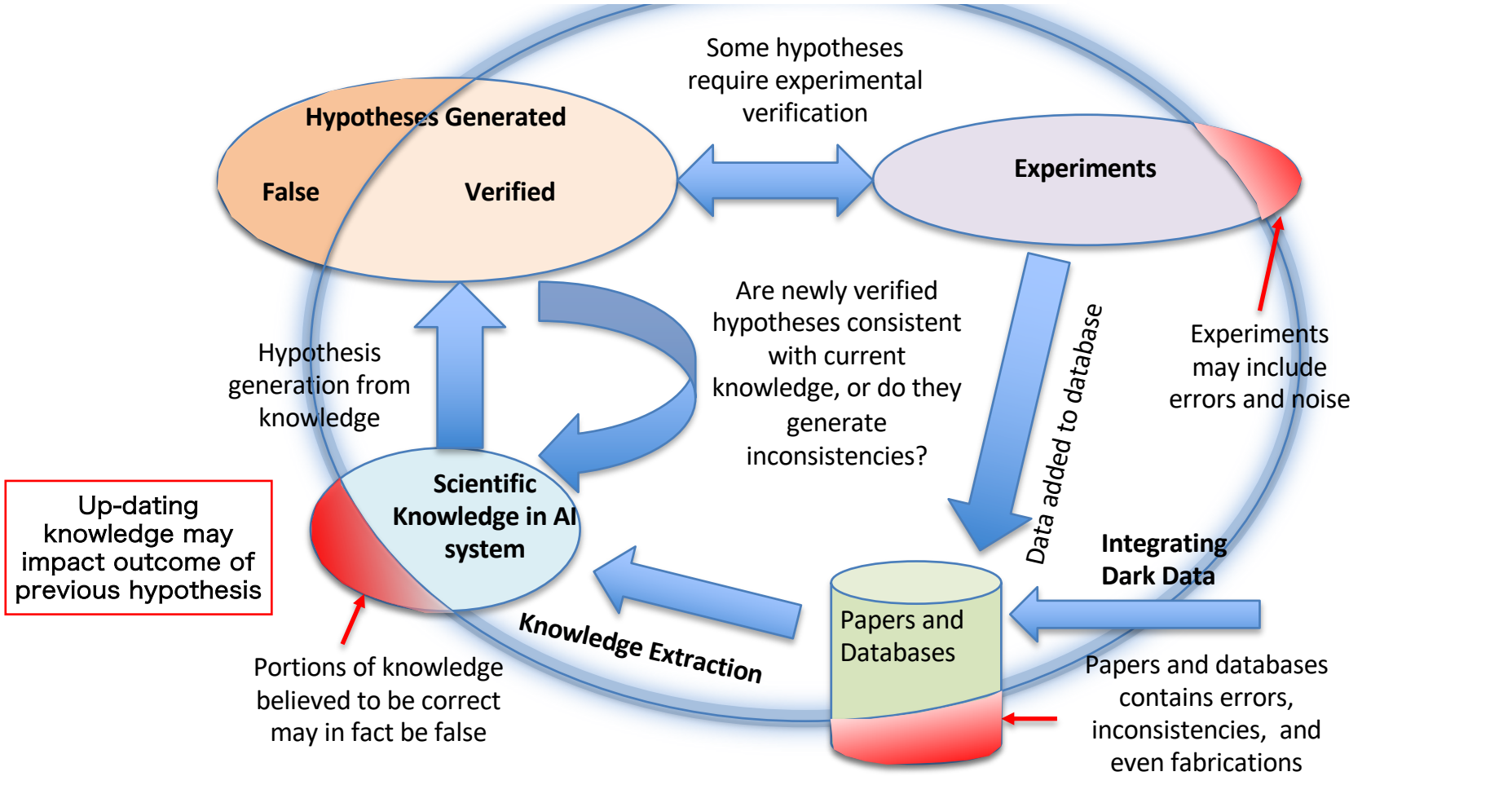
Each bubble : a focus area
Size of the bubble: the number of articles

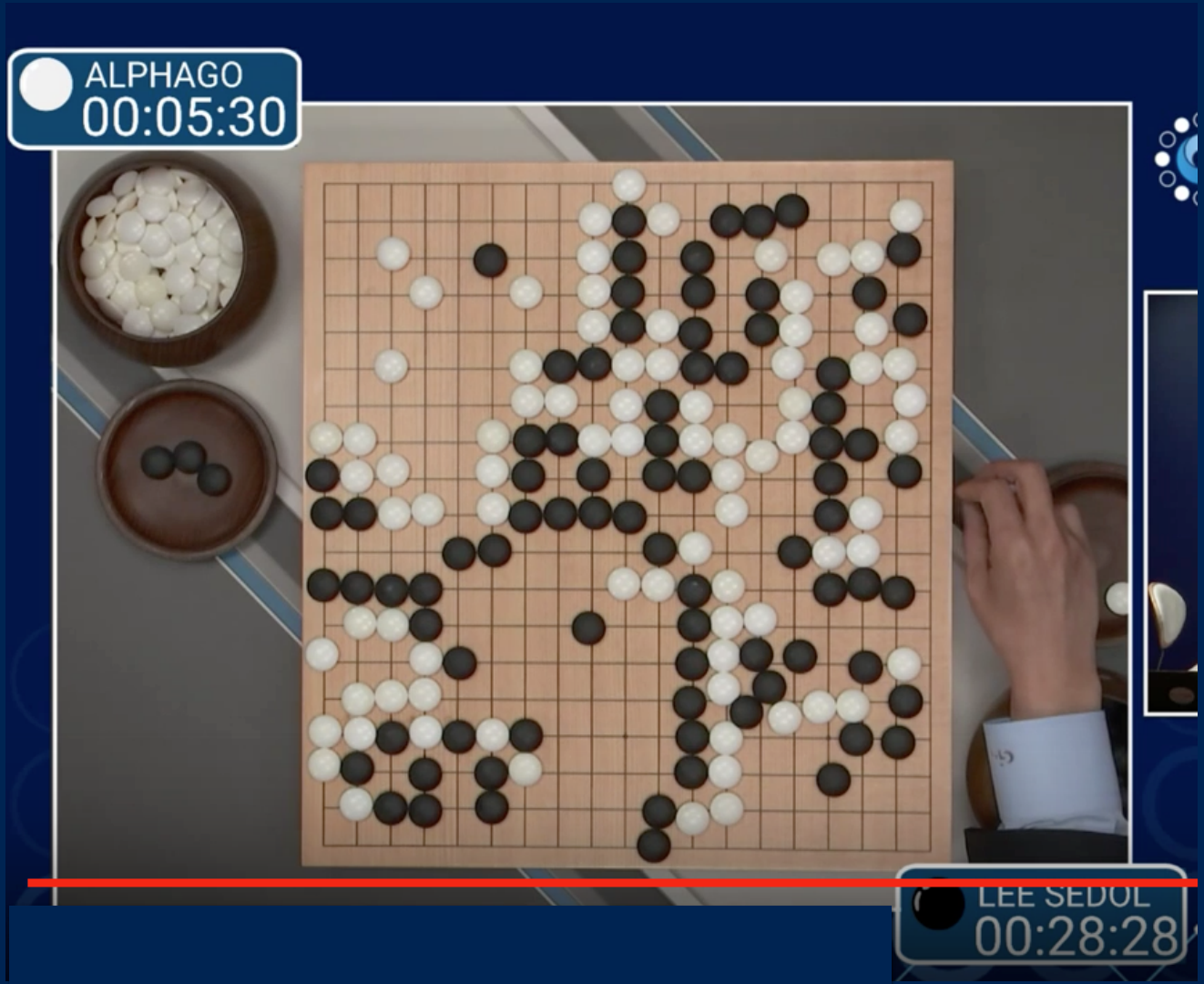
Article Landscape, 360 view, updated everyday

Total no. of focused microbiome articles	>50000
Total no. of channels	> 55
Total no. of KOLs identified	> 5000
Total no. of Key institutions identified	> 1000

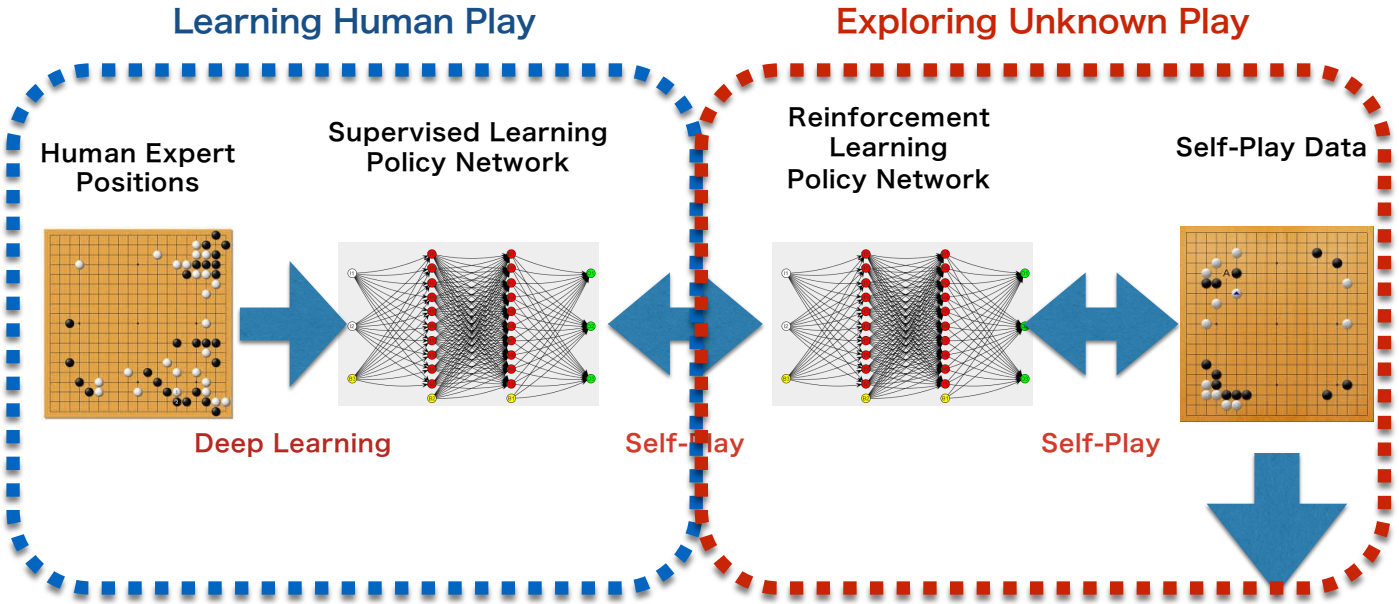
Galaxy view


taxila
Microbiome Scientific Landscape
More about this visualisation
Legend:
Authors
Shared Publications
Country
Search:
Search by name
Group Selector:
Select Group
Knight, Rob
Dinan, Timothy G
Cryan, John F
Key scientific leaders knowledge graph

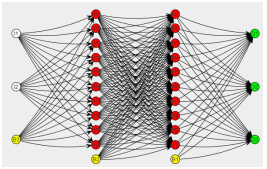

Entire Hypothetical Body of Scientific Knowledge
Some hypotheses require experimental verification
Hypotheses Generated
False
Verified
Experiments
Are newly verified hypotheses consistent with current knowledge, or do they generate inconsistencies?
Hypothesis generation from knowledge
Data added to database
Experiments may include errors and noise
Up-dating knowledge may impact outcome of previous hypothesis
Scientific Knowledge in AI system
Integrating Dark Data
Papers and Databases
Knowledge Extraction
Portions of knowledge believed to be correct may in fact be false
Papers and databases contains errors, inconsistencies, and even fabrications
Kitano, H., AI Magazine, March 2016


ALPHAGO
00:05:30
LEE SEDOL
00:28:28

Hypothesis Generation and Verification in AlphaGo



Game status evaluation is possible in the game of GO

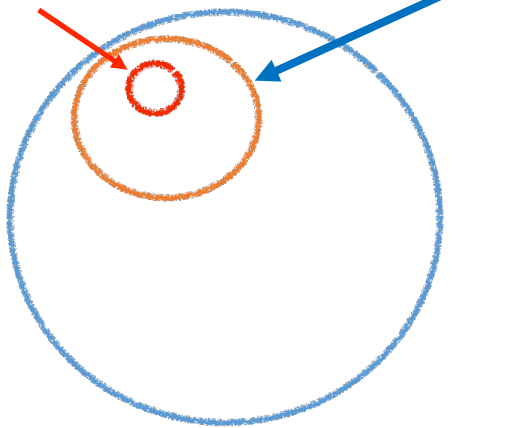


Value network

AlphaGo

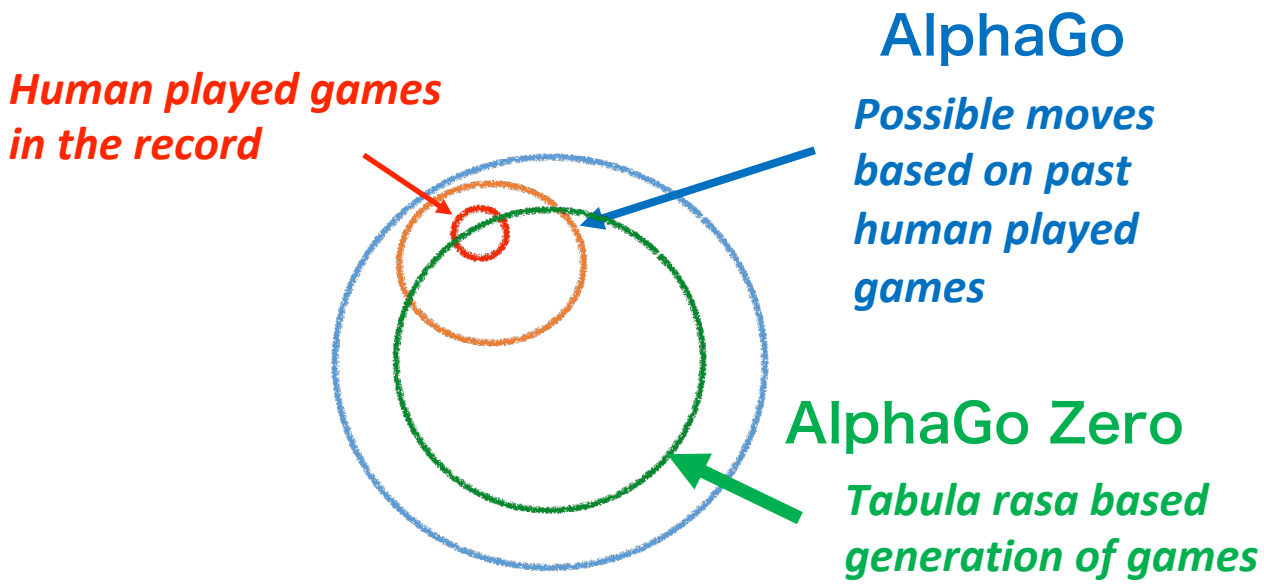
Human played games in the record

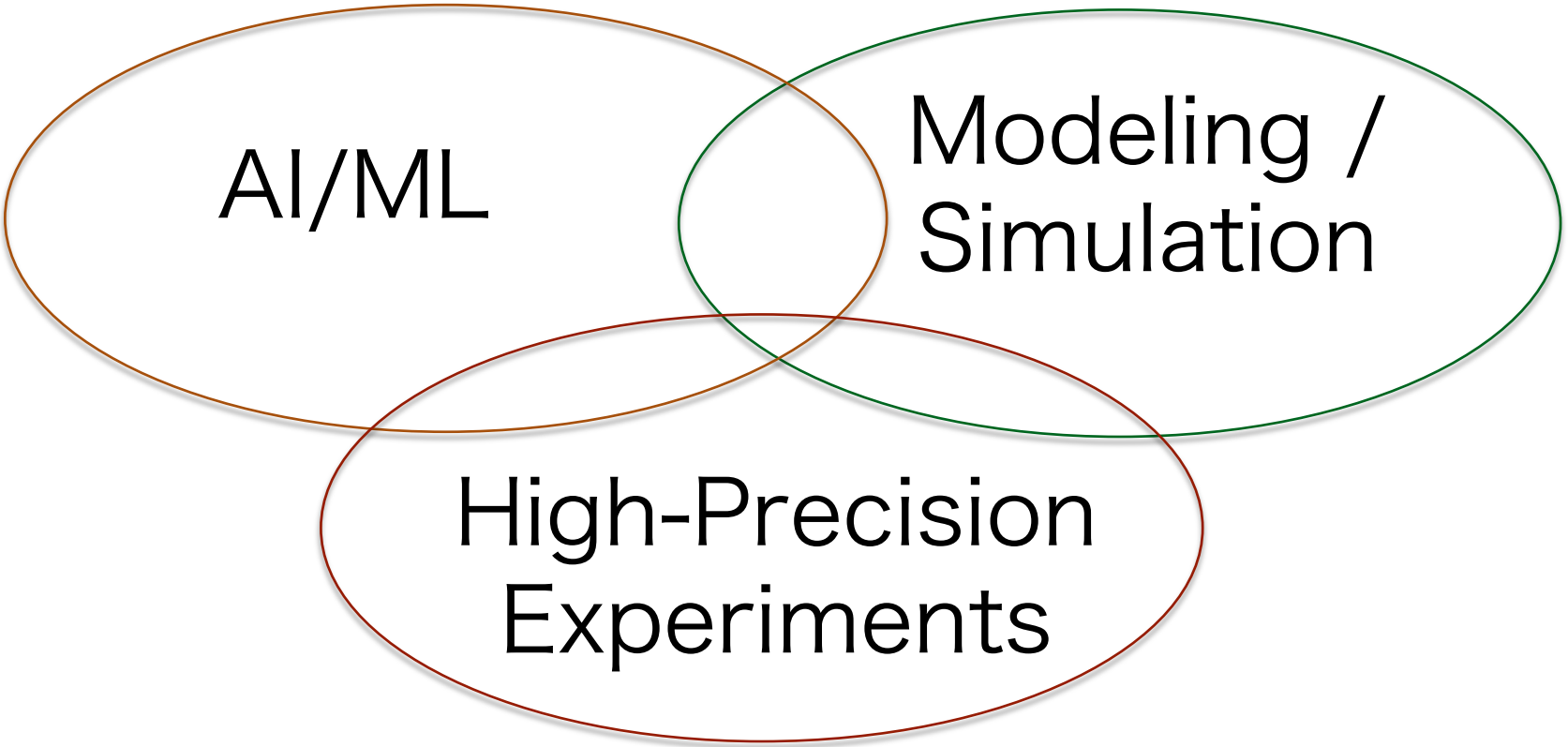
Possible moves based on past human played games

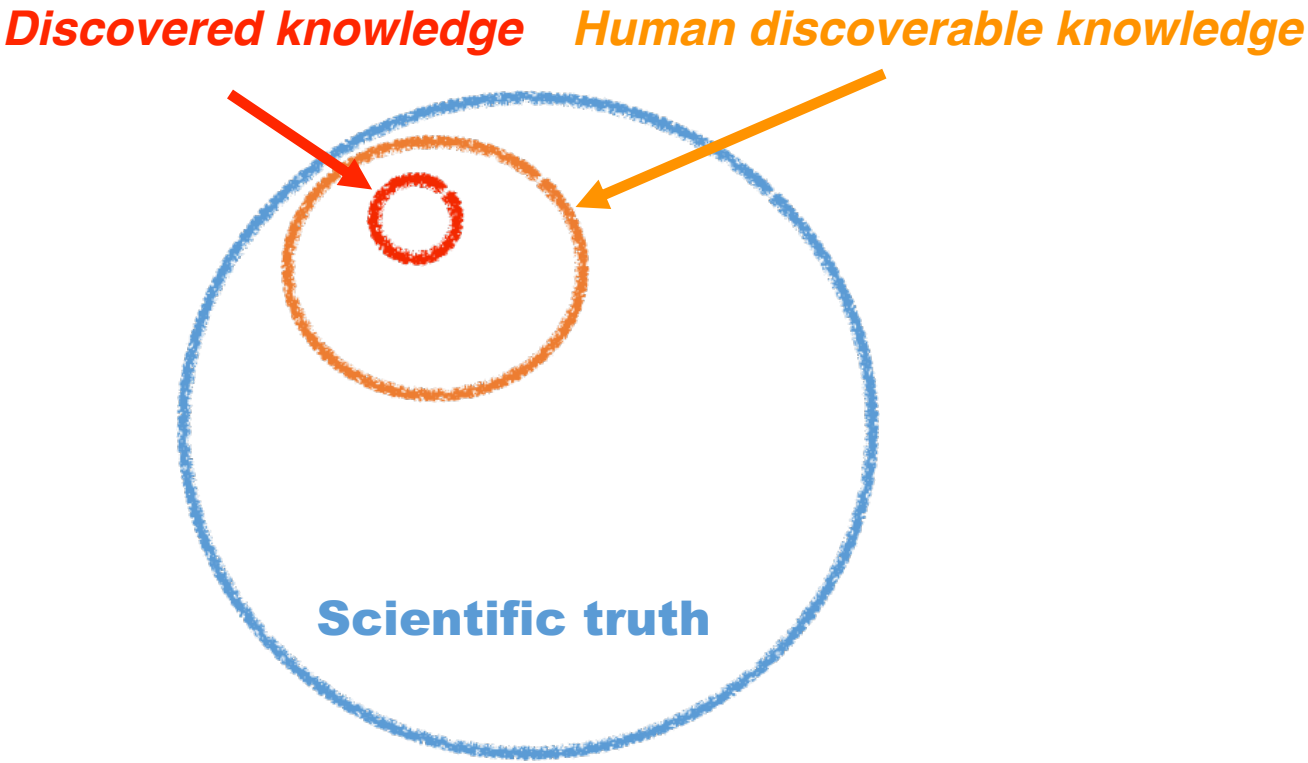


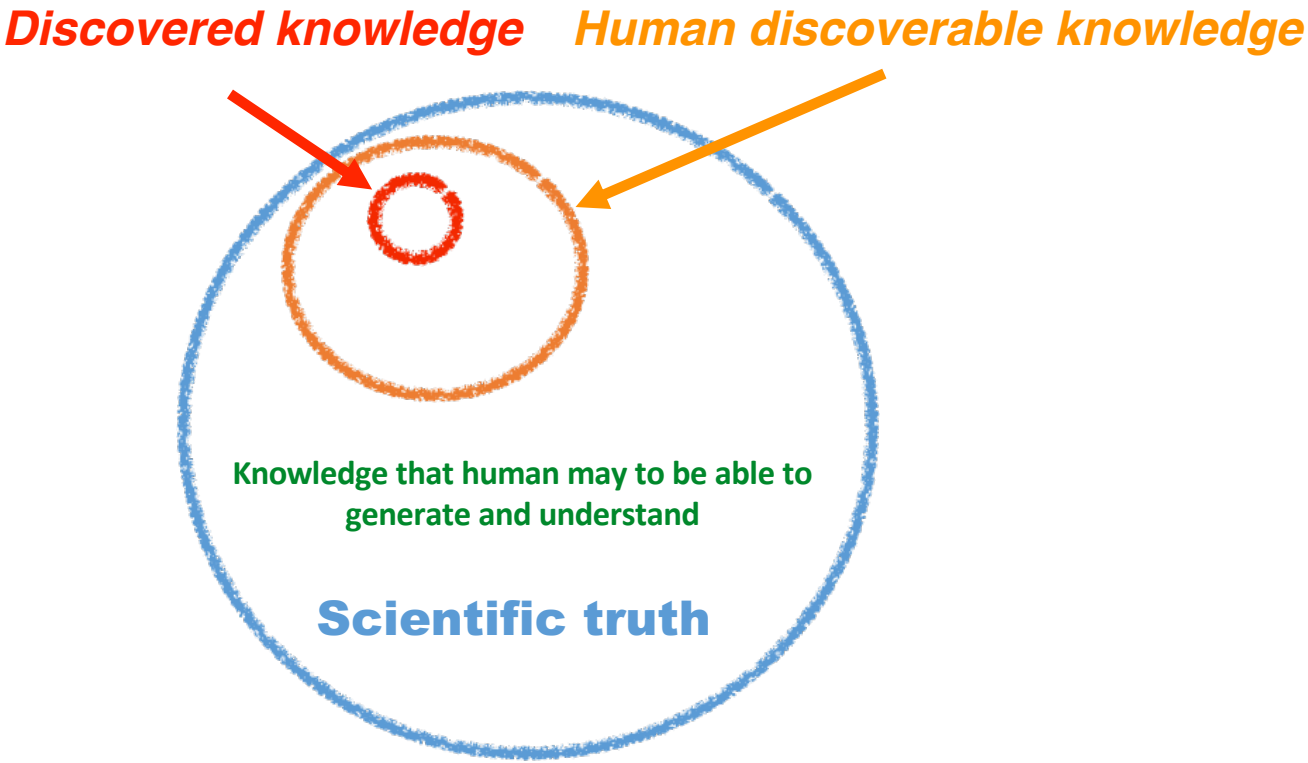
All possible moves on GO

AlphaGo ZERO




AI/ML
Modeling /
Simulation
High-Precision
Experiments


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)

Data Acquisition and Generation

~~Existing Data~~

Discard biased data
Turn over to more accurate data

＋

Data Acquisition

＋

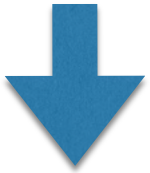
Virtual Data Generation

（Populating Hypothesis Space）

Can we develop a proper evaluation system?

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

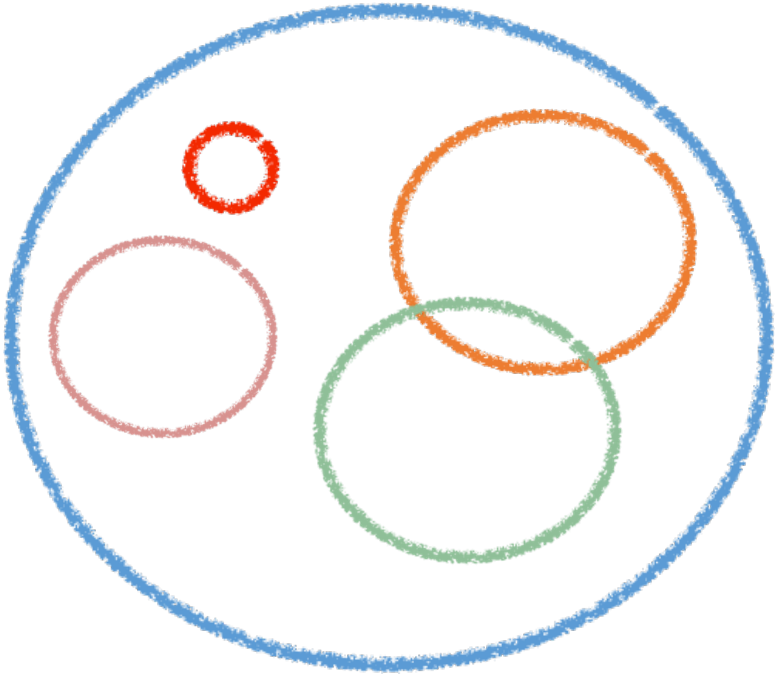
Build up ways to run super-fast hypothesis-testing cycle, so that every questions can be answered quickly
“Asking Right Questions” may not be that important anymore
Or
Human sense of “Right Questions” may be suboptimal

Redefining Scientific Discovery

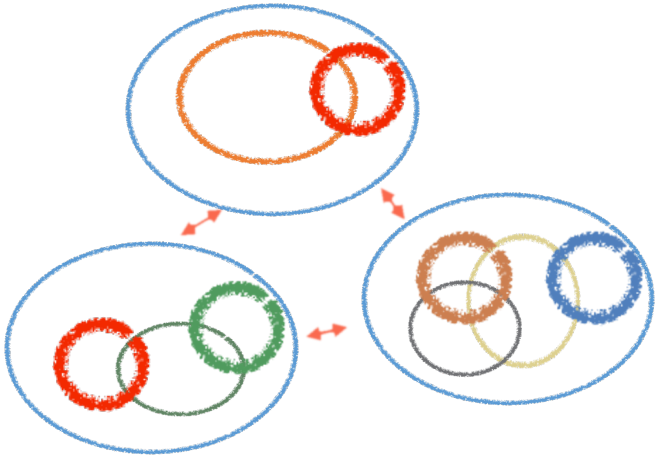
Massive search and verification of hypotheses space

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

Implications

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