The Automation of Science

Ross D. King,
University of Manchester & AIST, ross.king@manchester.ac.uk

Scientific Discovery

Technology Drivers

- Improved computer hardware:
  - faster processors, more processors, GPUs, …
- Improved data availability:
  - computers recording almost everything, deep data, …
- Improved computer software:
  - new machine learning methods, deep mining, …

Artificial Intelligence (AI)

- There have been multiple AI hype cycles, but this time it seems different.

- AI, especially machine learning, is now the hottest technology on the planet. Speed of Advance has surprised me.

- Machine Learning is the core technology of Google, Facebook, Amazon, …Tencent, Alibaba, Baidu, …

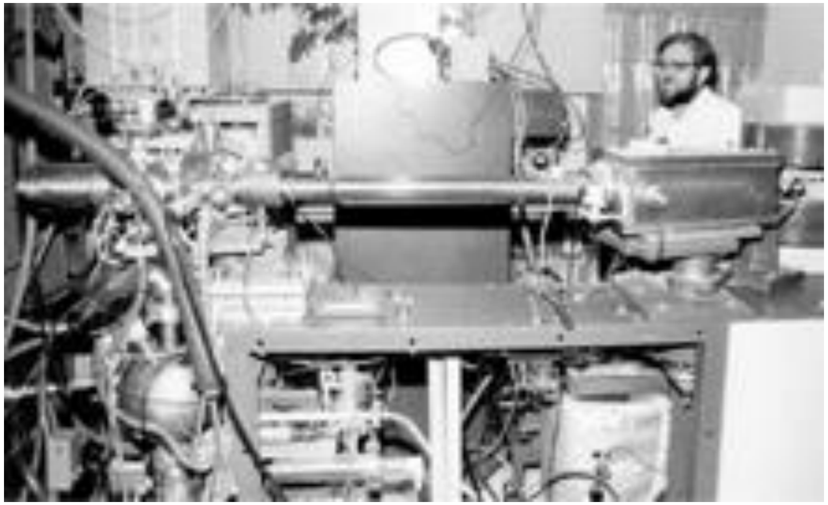
AI Systems have Superhuman Scientific Reasoning Powers

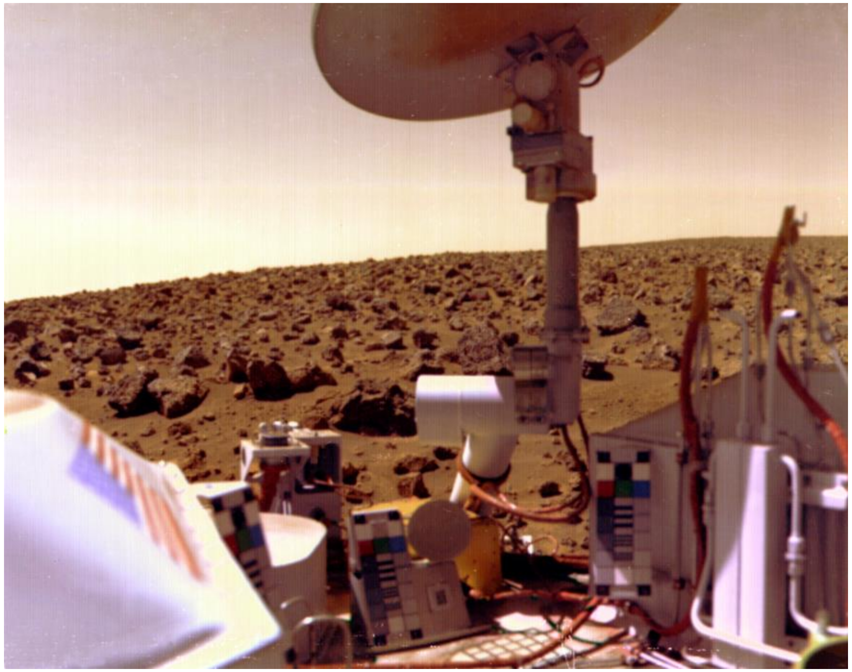
- Flawlessly remember vast numbers of facts
- Execute flawless logical reasoning
- Execute optimal probabilistic reasoning,
- Learn more rationally than humans
- Learn from vast amounts of data
- Extract information from millions of scientific papers.
- Etc.

Scientific Discovery

- Scientific problems are abstract, but involve the real-world.
- Scientific problems are restricted in scope – no need to know about “Cabbages and Kings”.
- Nature is honest – no malicious agents.
- Nature is a worthy object of our study.
- The generation of scientific knowledge is a public good.

Meta-Dendral

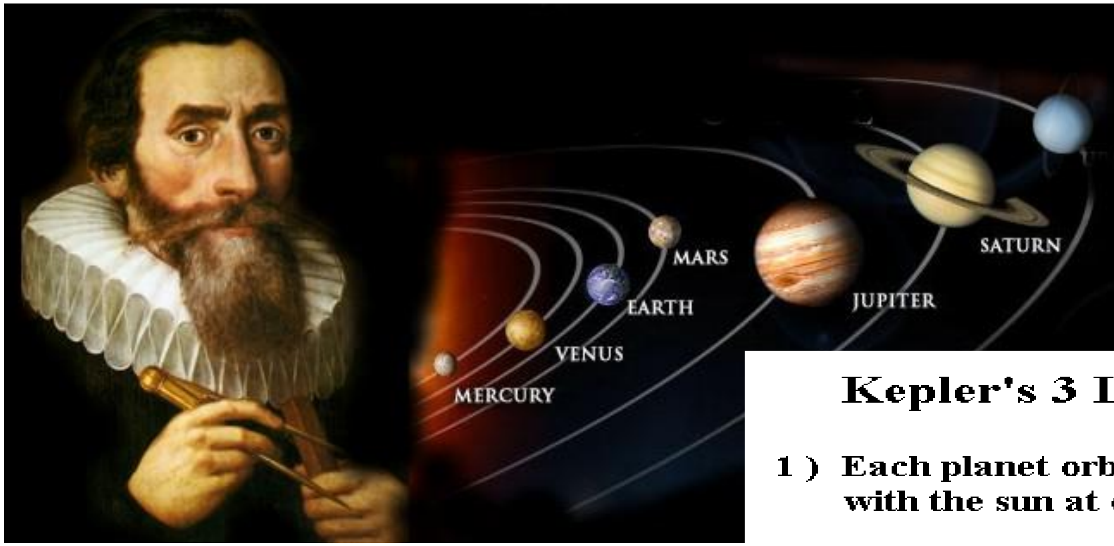




Analysis of mass-spectrometry data.

Joshua Lederburg, Ed. Feigenbaum, Bruce Buchanan, Karl Djerassi, *et al*. 1960-70s.

Bacon



Kepler's 3 Laws of Planetary Motion

1) **Each planet orbits the sun in an elliptical path with the sun at one focus**

2) **The radius vector (from sun to planet) sweeps out equal areas in equal time intervals**

3) **The square of the period is proportional to the cube of the semi-major axis of the orbit**

i.e. $T^2 = k\ a^3$ for some constant k

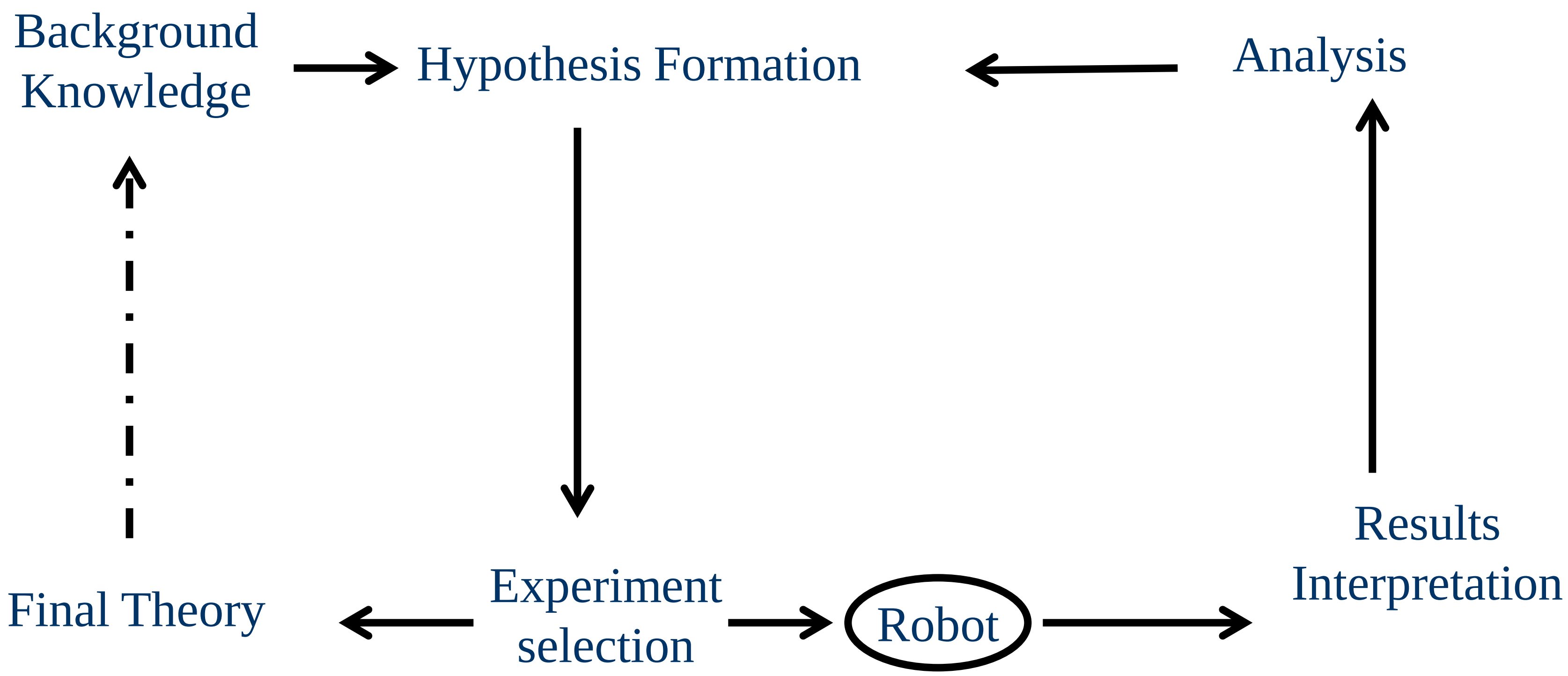
Figure 11.1

Rediscovering physics and chemistry: Langley, Bradshaw, Simon (1979).

Robot Scientists

The Concept of a Robot Scientist

Computer systems capable of originating their own experiments, physically executing them, interpreting the results, and then repeating the cycle.



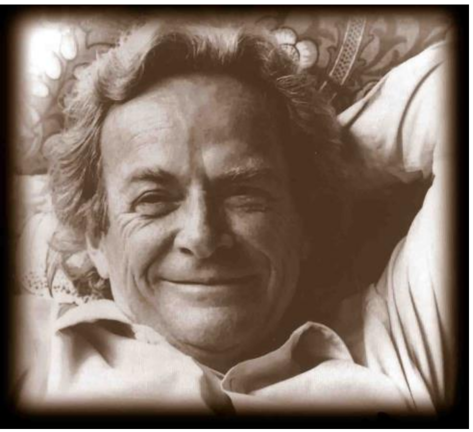
Motivation: Philosophical

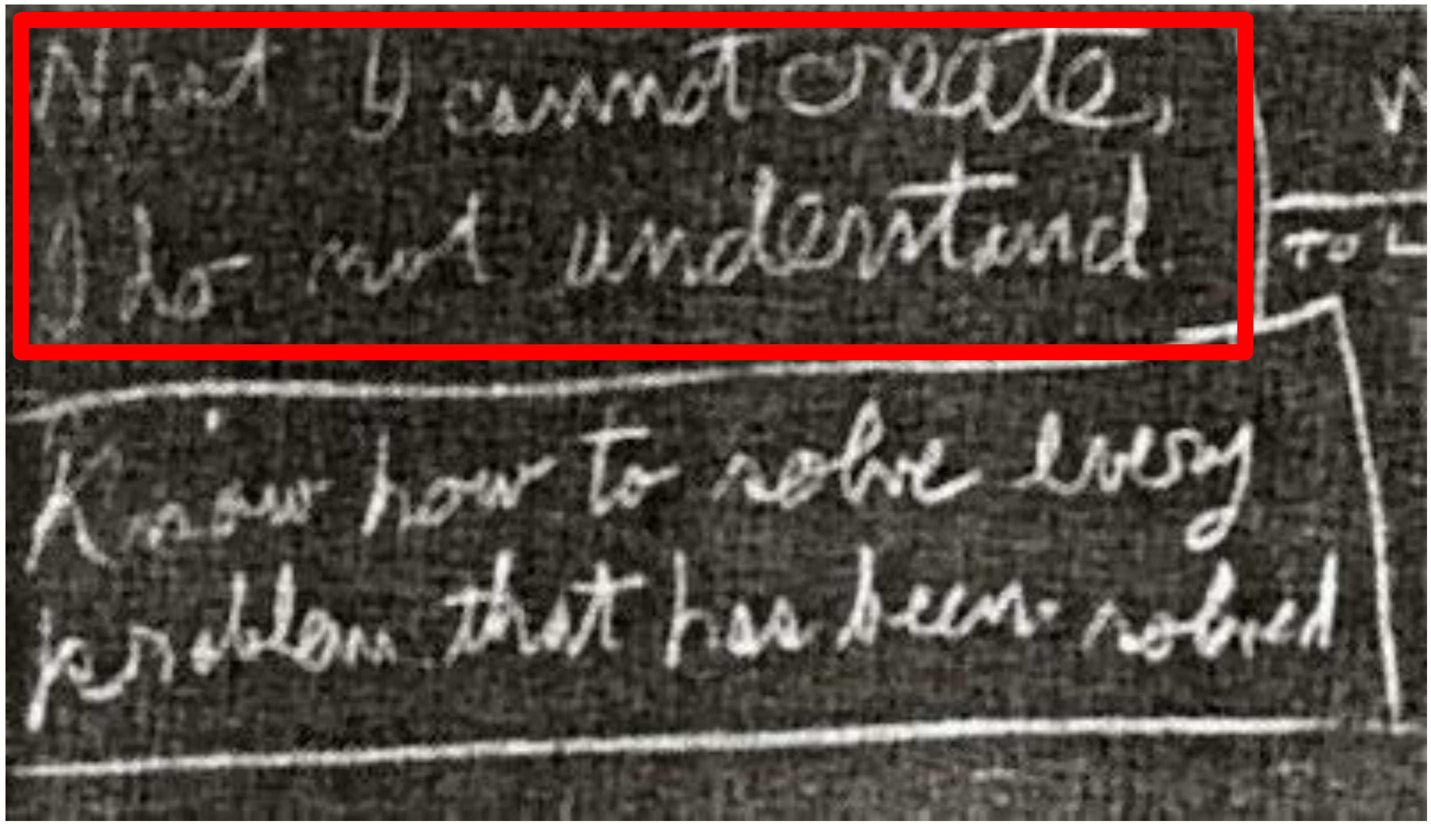
- What is Science?

- The question whether it is possible to automate scientific discovery seems to me central to understanding science.

- There is a strong philosophical position which holds that we do not fully understand a phenomenon unless we can make a machine which reproduces it.

Richard Feynman's Blackboard





"What I cannot create, I do not understand"

Motivation: Technological

- Robot Scientists have the potential to increase the productivity of science. They can work cheaper, faster, more accurately, and longer than humans. They can also be easily multiplied.
  - *Enabling the high-throughput testing of hypotheses*.

- Robot Scientists have the potential to improve the quality of science.
  - *by enabling the description of experiments in greater detail and semantic clarity.*

Robot Scientist Timeline

- 1999-2004 Initial Robot Scientist Project
  - Limited Hardware: Collaboration with Douglas Kell (Aber Biology), Steve Oliver (Manchester), Stephen Muggleton (Imperial)

  King *et al*. (2004) *Nature*, 427, 247-252

- 2004-2011 Adam – Yeast Functional Genomics
  - Sophisticated Laboratory Automation: Collaboration with Steve Oliver (Cambridge).

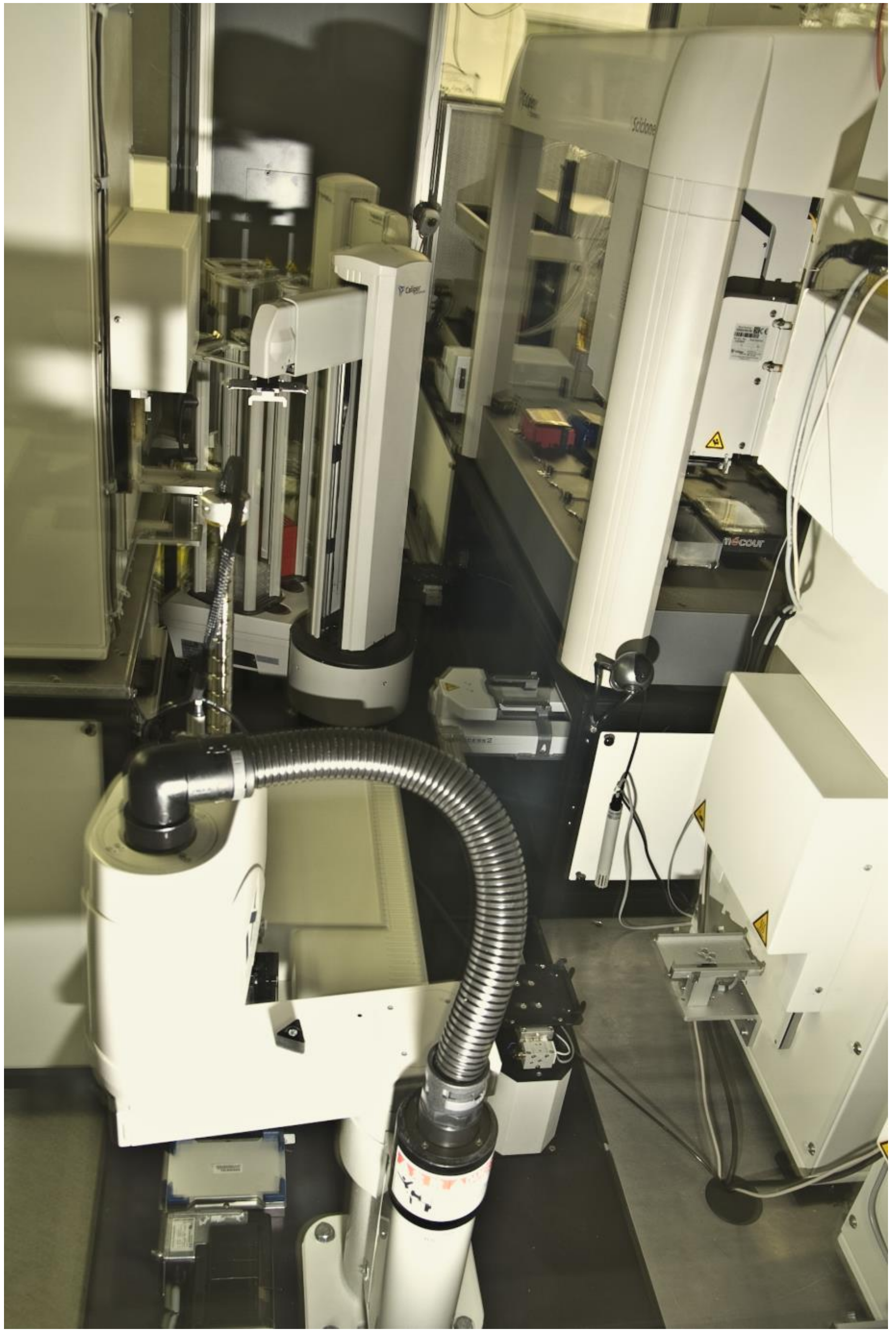
  King *et al*. (2009) *Science*, 324, 85-89

- 2008-2015 Eve – Drug Design for Tropical Diseases
  - Sophisticated Laboratory Automation: Collaboration with Steve Oliver (Cambridge)

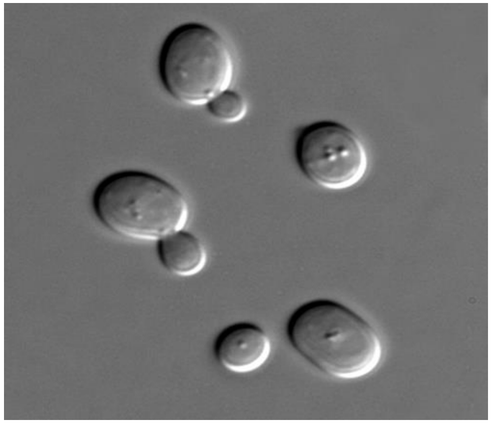
  Williams *et al*. (2015) *Royal Society Interface*, DOI 10.1098/rsif.2014.1289

- 2015-2018 Eve – Human cells - Cancer, Yeast - Aging
  - DARPA, CHIST-ERA

Adam



The Application Domain



- Functional genomics
- In yeast (*S. cerevisiae*) ~15% of the 6,000 genes still have no known function.
- EUROFAN 2 made all viable single deletant strains.
- Task to determine the “function” of a gene by growth experiments.

Formalising the Problem

- Use logic programming to represent background knowledge: metabolism modelled as a directed labeled hyper-graph.
- Use abduction to infer new hypotheses:
  - Abductive logic programming.
  - Techniques from Bioinformatics.
- Use active learning to decide efficient experiments: cost of compounds and time.
- Use machine learning to decide meaning of experimental results.

Types of Logical Inference

Deduction

Rule: All swans are white.

Fact: Daffy is a swan.

∴ Daffy is white.

Abduction

Rule: All swans are white.

Fact: Daffy is white.

∴ Daffy is a swan.

Induction

Fact: Daffy is a swan and white.

Fact: Tweety is a swan and white

∴ All swans are white.

Types of Logical Inference

Deduction

Rule: All swans are white.

Fact: Daffy is a swan.

∴ Daffy is white.

Abduction

Rule: All swans are white.

Fact: Daffy is white.

∴ Daffy is a swan. Daffy is a duck.

Induction

Fact: Daffy is a swan and white.

Fact: Tweety is a swan and white

∴ All swans are white.

Types of Logical Inference

Deduction

Rule: All swans are white.

Fact: Daffy is a swan.

∴ Daffy is white.

Abduction

Rule: All swans are white.

Fact: Daffy is white.

∴ Daffy is a swan.

Induction

Fact: Daffy is a swan and white.

Fact: Tweety is a swan and white

∴ All swans are white.

Types of Logical Inference

Deduction

Rule: All swans are white.

Fact: Daffy is a swan.

∴ Daffy is white.

Abduction

Rule: All swans are white.

Fact: Daffy is white.

∴ Daffy is a swan.

Induction

Fact: Daffy is a swan and white.

Fact: Tweety is a swan and white

∴ All swans are white.

Bruce is a black swan.

Novel Science

- Adam generated and confirmed novel functional-genomics hypotheses concerning the identify of genes encoding enzymes catalysing orphan reactions in the metabolic network of the yeast *S. cerevisiae.*

- Adam's conclusions have been manually verified using bioinformatic and biochemical evidence.

- Adam was the first machine to autonomously discover novel scientific knowledge: hypothesise, and experimentally confirm.

Novel Scientific Knowledge

Orphan Enzyme		Hypothesised Gene	Prob.	Acc.	No.	Existing Annotation	Dry	Wet
1	glucosamine-6-phosphate deaminase (3.5.99.6)	YHR163W (SOL3)	$<10^{-4}$	97	8	'6-phosphogluconolactonase' ida	-	-
2	glutaminase (3.5.1.2)	YIL033C (BCY1)	$<10^{-4}$	92	11	'cAMP-dependent protein kinase inhibitor ' ida	✗ ?	-
3	L-threonine 3-dehydrogenase (1.1.1.103)	YDL168W (SFA1)	$<10^{-4}$	83	6	'alcohol dehydrogenase' ida	-	-
4	purine-nucleoside phosphorylase (2.4.2.1)	YLR209C (PNP1)	$<10^{-4}$	82	11	'purine-nucleoside phosphorylase' ida	✓	-
5	2-aminoadipate transaminase (2.6.1.39)	YGL202W (ARO8)	$<10^{-4}$	80	3	'aromatic-amino-acid transaminase' ida	✓	✓
6	5,10-methenyltetrahydrofolate synthetase (6.3.3.2)	YER183C (FAU1)	$<10^{-4}$	80	4	'5,10 formyltetrahydrofolate cyclo-ligase' ida	✓	-
7	glucosamine-6-phosphate deaminase (3.5.99.6)	YNR034W (SOL1)	$<10^{-4}$	79	2	'possible role in tRNA export'	-	-
8	pyridoxal kinase (2.7.1.35)	YPR121W (THI22)	$<10^{-4}$	78	1	'phosphomethylpyrimidine kinase' iss	-	-
9	mannitol-1-phosphate 5-dehydrogenase (1.1.1.17)	YNR073C	$<10^{-4}$	78	6	'putative mannitol dehydrogenase ' iss	-	-
10	1-acylglycerol-3-phosphate O-acyltransferase (2.3.1.51)	YDL052C (SLC1)	0.0001	80	6	'1-acylglycerol-3-phosphate O-acyltransferase' ida	✓	-
11	glucosamine-6-phosphate deaminase (3.5.99.6)	YGR248W (SOL4)	0.0002	78	2	'6-phosphogluconolactonase' ida	-	-
12	maleylacetoacetate isomerase (5.2.1.2)	YLL060C (GTT2)	0.0003	76	3	'glutathione S-transferase' ida	-	-
13	serine O-acetyltransferase (2.3.1.30)	YJL218W	0.0005	78	2	'unknown function'	-	-
14	L-threonine 3-dehydrogenase (1.1.1.103)	YLR070C (XYL2)	0.0052	75	6	'xylitol dehydrogenase' ida	-	-
15	2-aminoadipate transaminase (2.6.1.39)	YJL060W (BNA3)	0.0084	73	3	'kynurenine aminotransferase' ida	-	✓
16	pyridoxal kinase (2.7.1.35)	YNR027W	0.0259	76	2	'involved in bud-site selection' iss	-	-
17	polyamine oxidase (1.5.3.11)	YMR020W (FMS1)	0.0289	78	4	'polyamine oxidase' ida	✓	-
18	2-aminoadipate transaminase (2.6.1.39)	YER152C	0.0332	74	3	'uncharacterized'	-	✓
19	L-aspartate oxidase (1.4.3.16)	YJL045W	0.1300	72	1	'succinate dehydrogenase isozyme' iss	-	-
20	purine-nucleoside phosphorylase (2.4.2.1)	YLR017W (MEU1)	0.1421	72	6	'methylthioadenosine phosphorylase' ida	✓	-

Formalising Science

Formalisation of Science

- The goal of science is to increase our knowledge of the natural world through the performance of experiments.

- This knowledge should be expressed in *formal logical languages*.

- Formal languages promote semantic clarity, which in turn supports the free exchange of scientific knowledge and simplifies scientific reasoning.

Robot Scientist & Formalisation

- Robot Scientists provide excellent test-beds for the development of methodologies for formalising science.
- Using them it is possible to completely capture and digitally curate all aspects of the scientific process.
- The ontology LABORS is designed to enable the open access of the Robot Scientist experimental data and metadata to the scientific community.

Ontologies

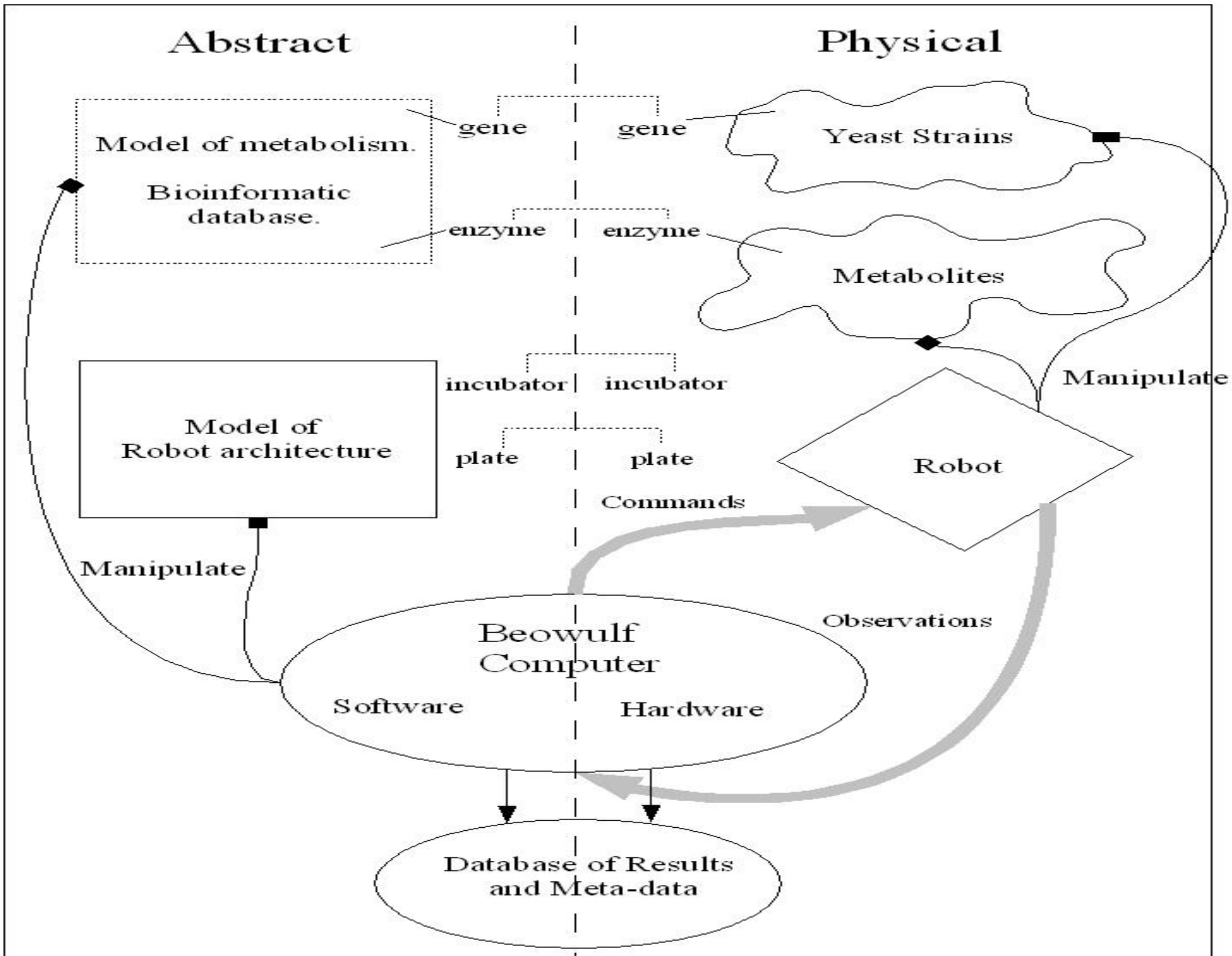
An ontology is "a concise and unambiguous description of what principal entities are relevant to an application domain and the relationship between them"*.

*Schulze-Kremer, S., 2001, Computer and Information Sci. 6(21)

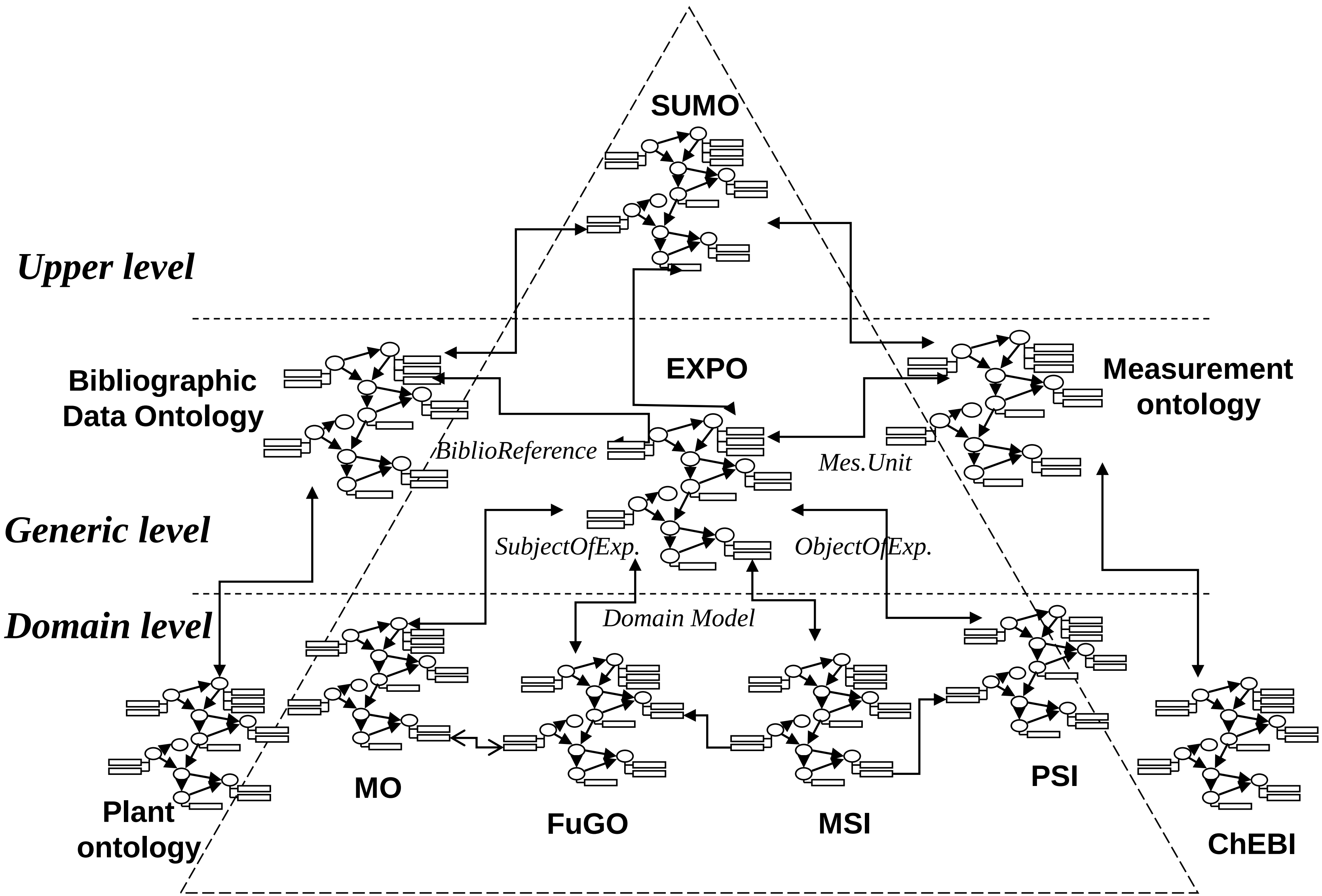
Dualism

- The most fundamental ontological division in our design of Adam is between <abstract> and <physical> objects
- We argue for this ontological division because it makes explicit the separation between models and reality.
- All the objects which Adam deals with computationally are <abstract>, and all the objects it deals with physically are <physical>.

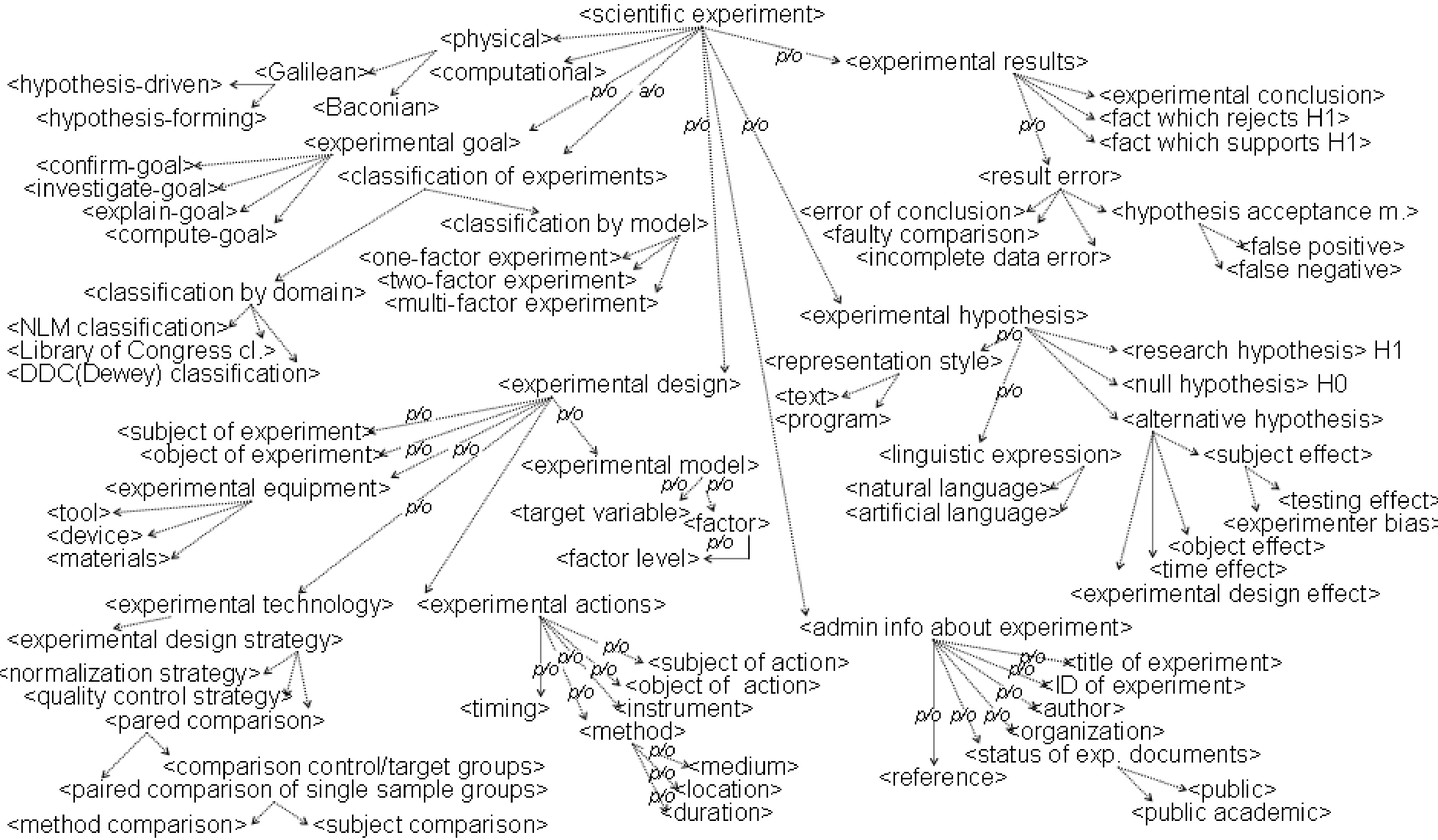
Overall View of the Universe



The Position of EXPO

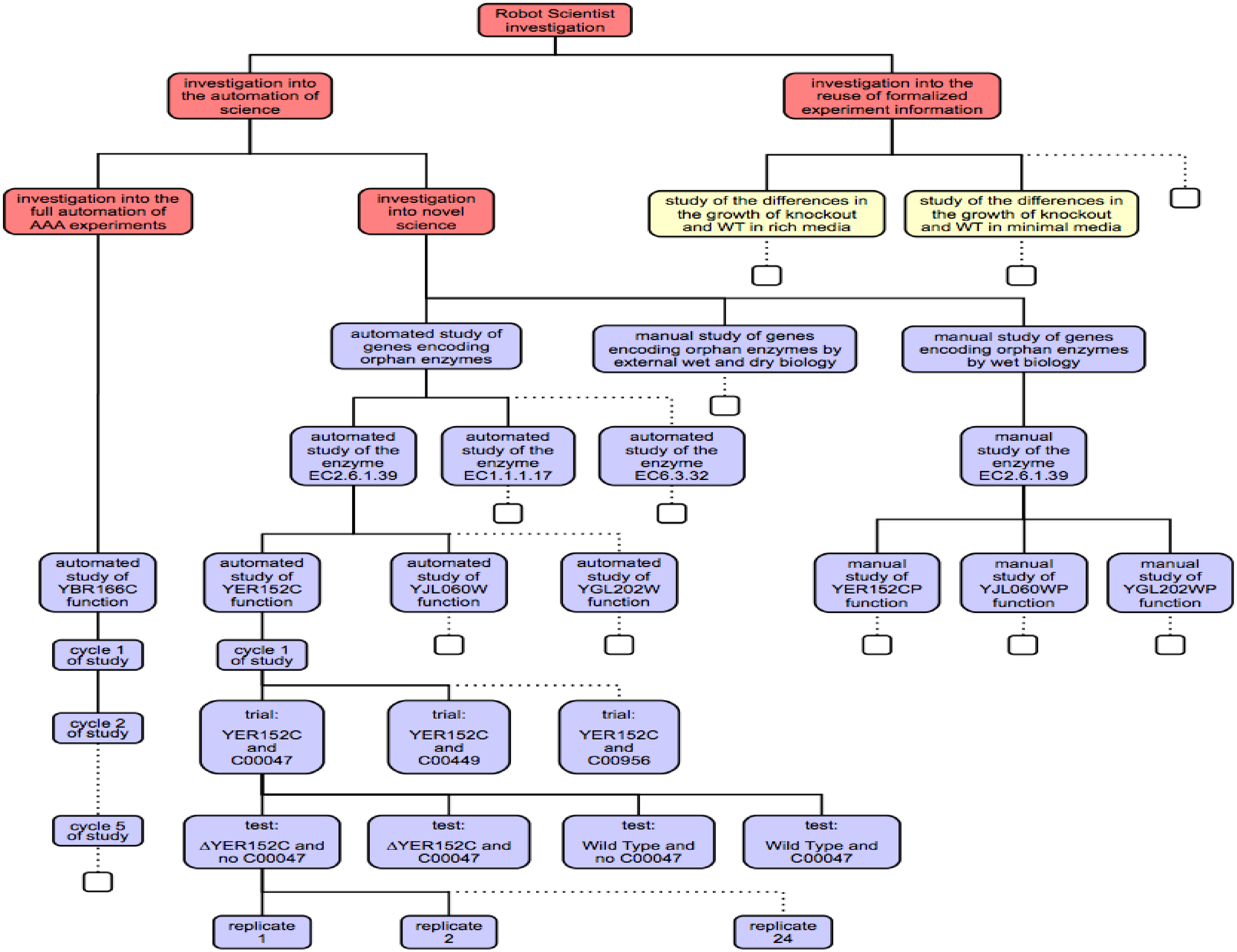


Small Section of EXPO



Adam's Investigations

- This formalisation involves >10,000 different research units in a nested tree-like structure 11 levels deep.
- It logically connects >6.6 million $OD600_{nm}$ measurements to hypotheses, experimental goals, results, etc.
- No previous large-scale experimental work has been so comprehensively described and recorded.



Robot Scientist investigation
investigation into the automation of science
investigation into the reuse of formalized experiment information
investigation into the full automation of AAA experiments
investigation into novel science
study of the differences in the growth of knockout and WT in rich media
study of the differences in the growth of knockout and WT in minimal media
automated study of genes encoding orphan enzymes
manual study of genes encoding orphan enzymes by external wet and dry biology
manual study of genes encoding orphan enzymes by wet biology
automated study of the enzyme EC2.6.1.39
automated study of the enzyme EC1.1.1.17
automated study of the enzyme EC6.3.32
manual study of the enzyme EC2.6.1.39
automated study of YBR166C function
automated study of YER152C function
automated study of YJL060W function
automated study of YGL202W function
manual study of YER152CP function
manual study of YJL060WP function
manual study of YGL202WP function
cycle 1 of study
cycle 2 of study
cycle 5 of study
cycle 1 of study
trial: YER152C and C00047
trial: YER152C and C00449
trial: YER152C and C00956
test: ΔYER152C and no C00047
test: ΔYER152C and C00047
test: Wild Type and no C00047
test: Wild Type and C00047
replicate 1
replicate 2
replicate 24

Levels in the Formalisation

Investigation into the automation of Science

- **Investigation into the automation of novel science**
  - **Investigation into the automated discovery of genes encoding orphan enzymes**
    - **Automated study of E.C.2.6.1.39 encoding**
      - **Cycle 1 of automated study of YER152C function**
        - **YER152C and Lysine automated trial**
          - **Experiment 1 (wild-type no metabolite)**
            - **Replicate 1 (well)**
              - **Observation 1**

automated study of yer152c function

has text representation:

automated study: automated study of yer152c_function

has domain of study: functional genomics

has investigato

has goal: 'To te
with enzyme cl

has organism o

has ncbi taxo

has hypothese

has research

has negative

has cycle 1 of s

has study resu
encodes(yer15

highest

proporti

has study conc

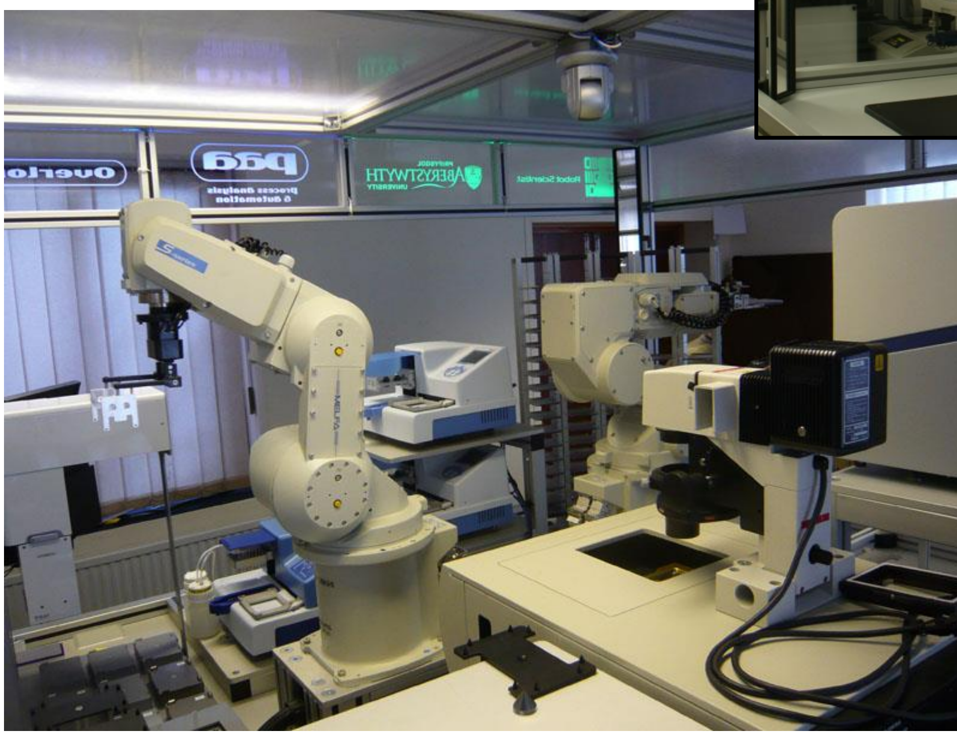
has datalog representation:

a:automated study(X) :- a:automated_study_o
a:hypotheses-set(X) :- a:research_hypothesis
a:cycle_of_study(X) :- a:cycle_1_of_study_(X
a:hypotheses-set(X) :- a:negative_hypothesis
a:domain_of_study(Y) :- a: automated study(X
a:investigator(Y) :- a: automated study(X), a:h
a:goal(Y) :- a: automated study(X), a:has_goa
a:organism_of_study (Y) :- a: automated stud
a:hypotheses-set(Y) :- a: automated study(X)
a:cycle_of_study(Y) :- a: automated study(X),
a:study_result(Y) :- a: automated study(X), a:
a:study_conclusion(Y) :- a: automated study(
a:domain_of_study(X) :- a:functional_genomi
a: investigator(X) :- a:adam.
a:goal(X) :- a: to_test_the_hypothesis_that_g
_encodes_an_enzyme_with_enzyme_class_
a:organism_of_study(X) :- a:saccharomyces_
a:study_result(X) :- a:the_strength_of_eviden
a:study_conclusion(X) :- a:hypothesis_1_con

has OWL representation:

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns="http://www.owl-ontologies.com/Ontology1204198571.owl#"
<owl:Class rdf:ID="goal"/>
 <owl:Class rdf:ID="study_result"/>
 <owl:Class rdf:ID="ncbi_taxonomy_ID"/>
 <owl:Class rdf:ID="cycle_of_study"/>
 <owl:Class rdf:ID="negative_hypothesis">
  <rdfs:subClassOf>
   <owl:Class rdf:ID="hypotheses-set"/>
  </rdfs:subClassOf>
 </owl:Class>
 <owl:Class rdf:ID="domain_of_study"/>
 <owl:Class rdf:ID="organism_of_study"/>
 <owl:Class rdf:ID="cycle_1_of_study_">
  <rdfs:subClassOf rdf:resource="#cycle_of_study"/>
 </owl:Class>
 <owl:Class rdf:ID="automated_study">
  <rdfs:subClassOf>
   <owl:Restriction>
    <owl:someValuesFrom rdf:resource="#goal"/>
    <owl:onProperty>
     <owl:ObjectProperty rdf:ID="has_goal"/>
    </owl:onProperty>
   </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
   <owl:Restriction>
    <owl:someValuesFrom rdf:resource="#organism_of_study"/>
    <owl:onProperty>
     <owl:ObjectProperty rdf:ID="has_organism_of_study"/>
………………………………………………………
```

<u>Eve</u>



<u>Drug Design</u>

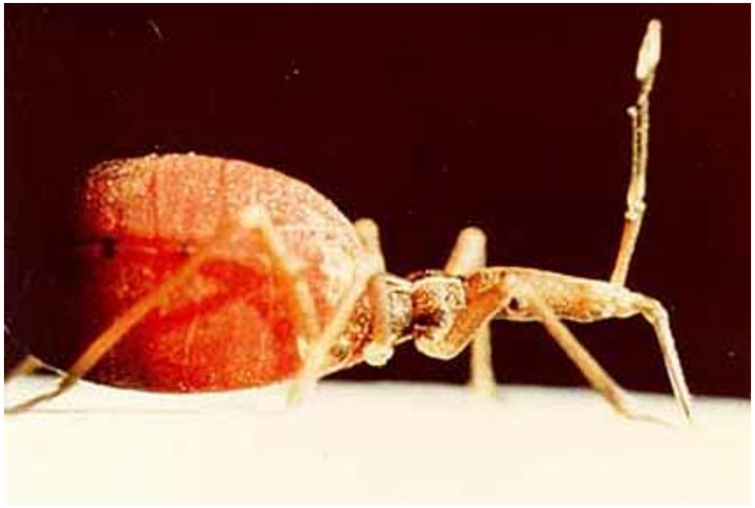
The Application Domain

Malaria

Shistosomaisis

Leishmania

Chagas



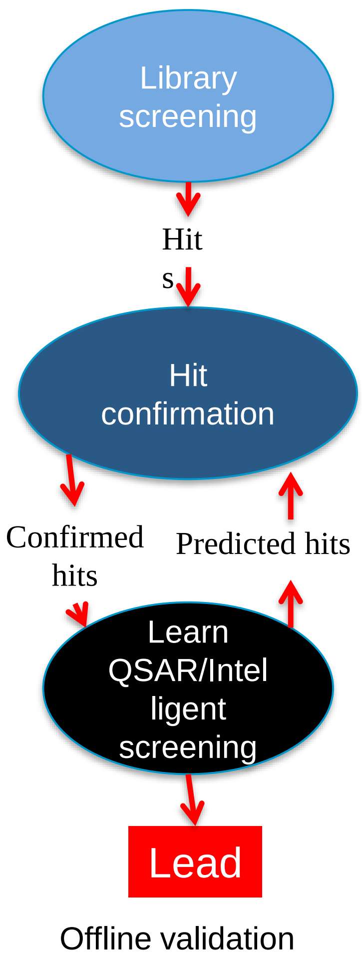
Why Tropical Diseases?

- Millions of people die of these diseases, and hundreds of millions of people suffer infection.
- It is clear how to cure these diseases – kill the parasites.
- They are “neglected”, so avoid competition from the Pharmaceutical industry.

Formalising the Problem

- Use graphs and standard chemoinformatic methods to represent background knowledge - the use of relations is planned.
- Uses induction (quantitative structure activity relationship – QSAR learning) to infer new hypotheses.
- Use active learning to decide efficient experiments, and econometric model to decide what compounds to test.

Eve's Automation of Pipeline



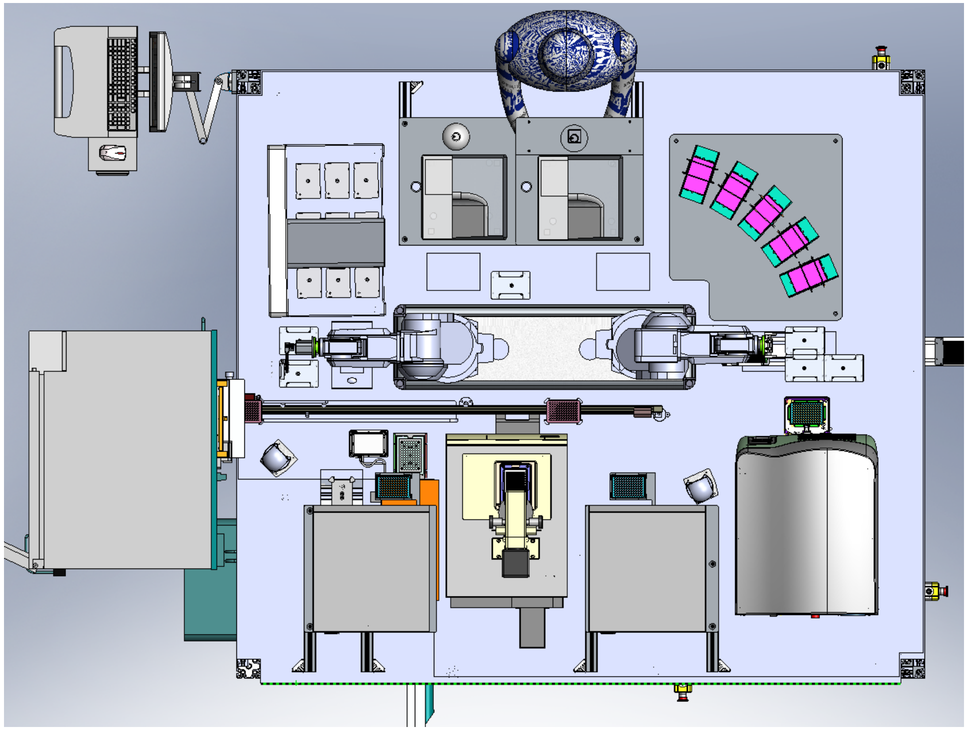
- Standard library screening is brute force:
- Eve uses intelligent screening

- In the standard "pipeline" the 3 processes are not integrated.
- In Eve automated and integrated.

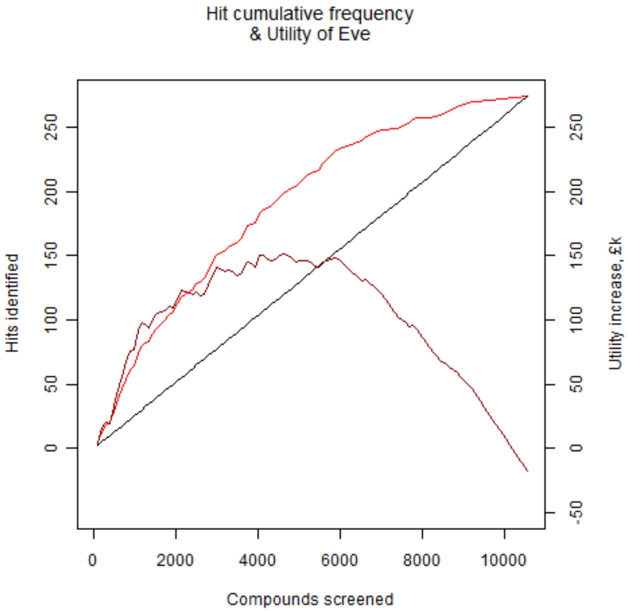
Eve’s Hardware

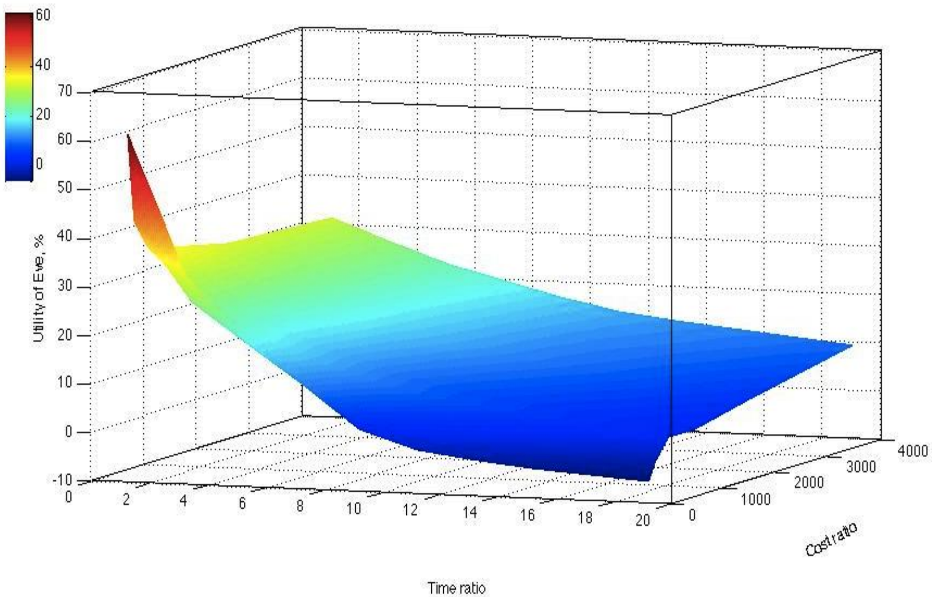
Highlights of Eve's hardware:

- Acoustic liquid handling
- High throughput 384 well plates
- Two industrial robot arms
- Automated 60x microscope
- Liquid handlers, fluorescence readers, barcode scanners, dry store, incubator, tube decapper ...



The Economics of Intelligent Screening

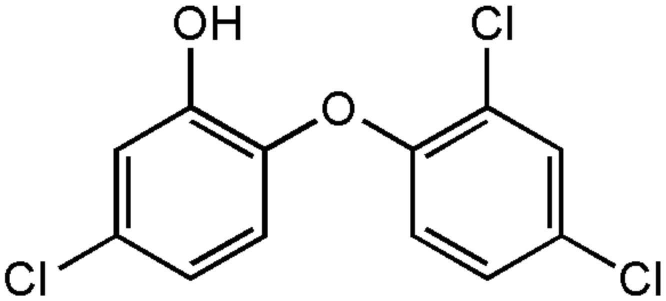




$$\Delta\, Utility\ of\ Eve = \sum_{1}^{Nm}(Tm + Cm) + \sum_{1}^{Nx}(Tc + Cc - Uh) + \sum_{1}^{Ne}(Tm - Tc + Cm - Cc)$$

Nm	-	Number of compounds not assayed by Eve
Tm	-	Cost of the time to screen a compound using the mass screening assay
Cm	-	Cost of the loss of a compound in the mass screening assay
Nx	-	Number of hits missed by Eve
Tc	-	Cost of the time to screen a compound using a cherry-picking (confirmation or intelligent) assay
Cc	-	Cost of the loss of a compound in a cherry-picking assay
Uh	-	Utility of a hit
Ne	-	Number of compounds assayed by Eve

Triclosan Repositioned for Malaria



- Simple compound
- Known to be safe – used in toothpaste.
- Targets both DHFR and FAS-II – well established targets.
- Demonstrated activity using multiple wet experimental techniques.
- Works against wild-type and drug-resistant *Plasmodium falciparum*, and *Plasmodium vivax*.

Future Prospects

The Future?

- In Chess/Go there is a continuum of ability from novices up to Grandmasters.
- I argue that this is also true in science, from the simple research of Eve, through what most human scientists can achieve, up to the ability of a Newton or Einstein.
- If you accept this, then just as in Chess/Go, it is likely that advances in computer hardware and software will drive the development of ever smarter Robot Scientists.
- In favour of this argument are the ongoing development of AI and laboratory robotics.

Vision

- The collaboration between Human and Robot Scientists will produce better science than either can alone – human/computer teams still play better chess that either alone.

- Scientific knowledge will be primarily expressed in logic with associated probabilities and published using the Semantic Web.

- The improved productivity of science leads to societal benefits: better food security, better medicines, etc.

- The Physics Nobel Frank Wilczek is on record as saying that in 100 years' time the best physicist will be a machine

Conclusions

- Science is a wonderful application area for AI.
- Automation is becoming increasingly important in scientific research e.g. DNA sequencing, drug design.
- The Robot Scientist concept is the logical next step in scientific automation.
- The Robot Scientist Adam was the first machine to have discovered novel scientific knowledge.
- The Robot Scientist Eve has found new lead compounds for neglected tropical diseases.
- The Robot Scientist Eve can accelerate systems biology modelling.

Acknowledgments

- The Robot Scientist team: Manchester, Aberystwyth, Cambridge, Brunel, Leuven, Thailand. (BBSRC)
- Chicago Big Mechanism consortium. (DARPA)
- AdaLab consortium. (CHIST-ERA)
- AIST (Tokyo)