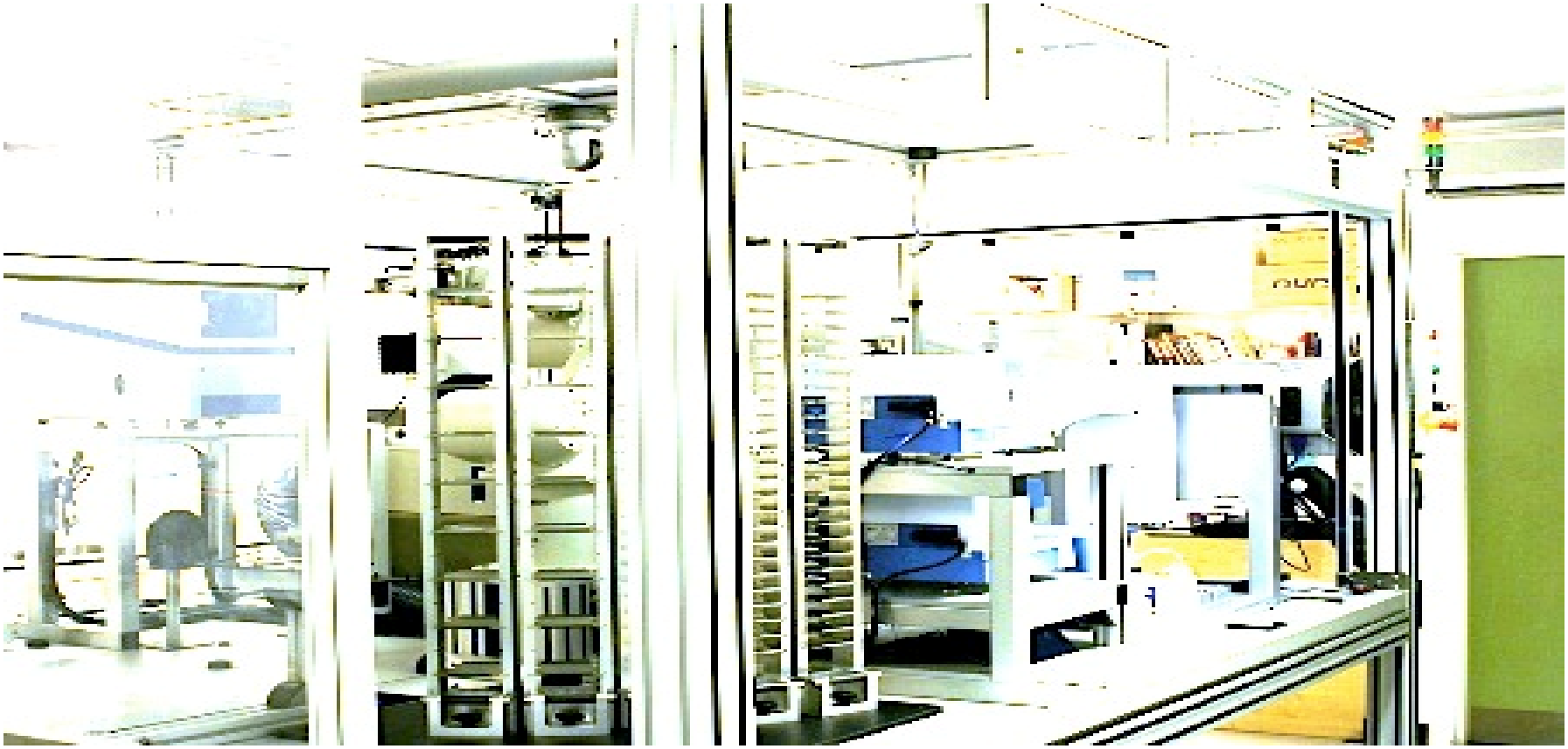
Automating Science using Robot Scientists

Ross D. King,
Alan Turing Institute, Chalmers University, rdking@turing.ac.uk



Scientific Discovery

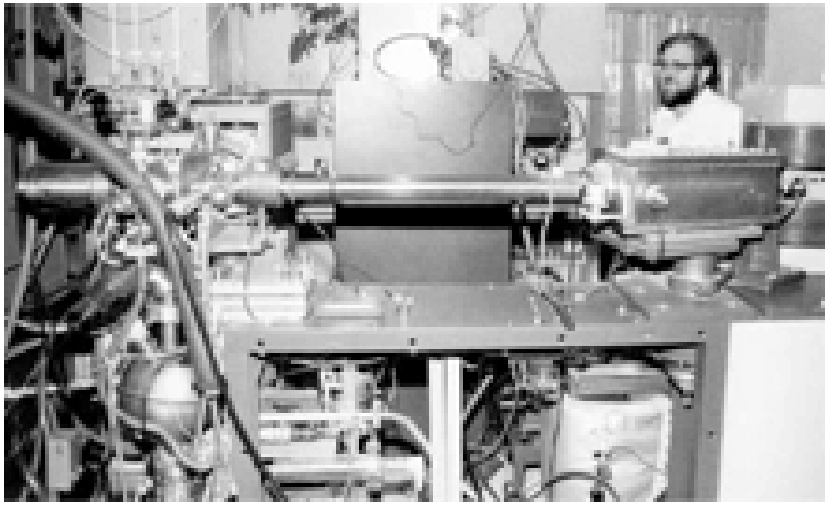
AI Systems have Superhuman Scientific Powers

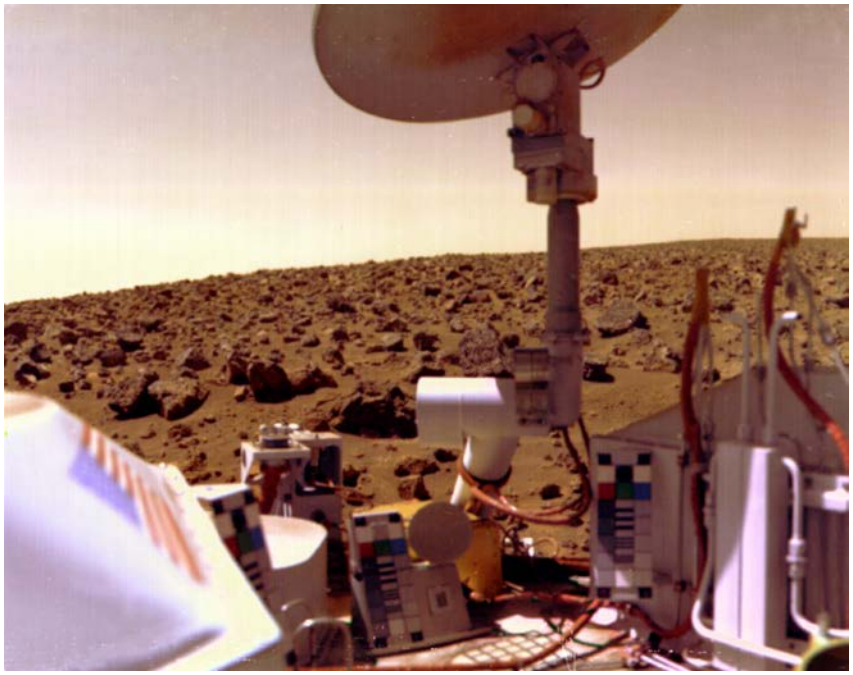
- Flawlessly remember vast numbers of facts
- Flawless logical reasoning
- Near optimal probabilistic reasoning
- Learn from vast amounts of data
- Extract information from millions of scientific papers.
- Etc.

Science as an AI Task

- 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.
- 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**

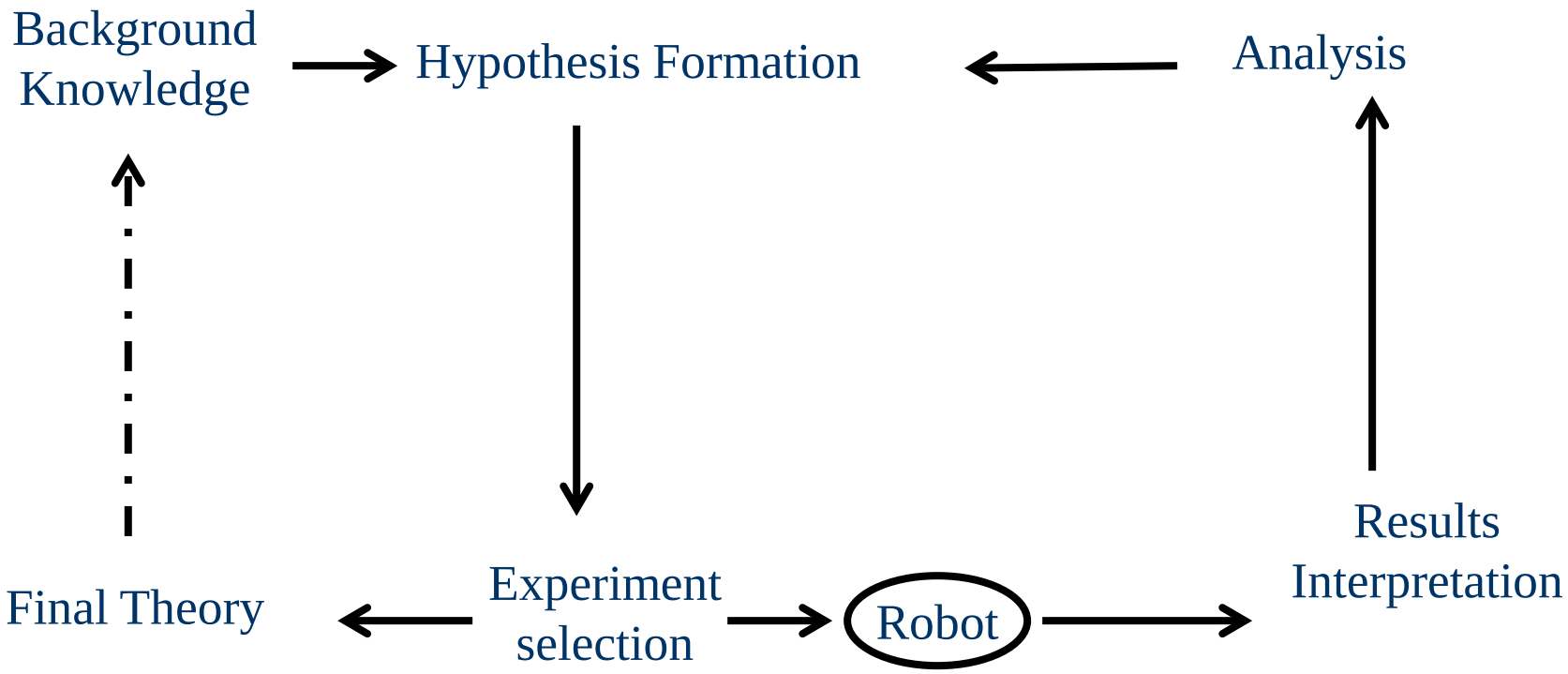
Figure 11.1

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

Robot Scientists

The Concept of a Robot Scientist

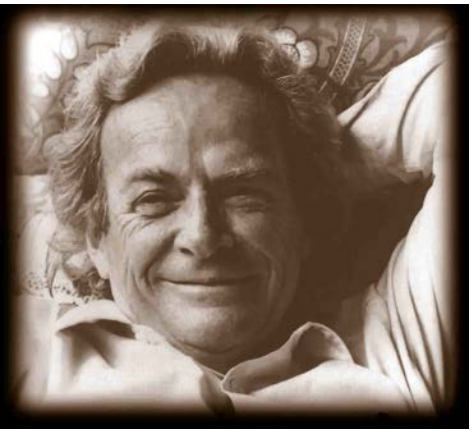
Computer system capable of originating its own experiments, physically executing them, interpreting the results, and then repeating the cycle.

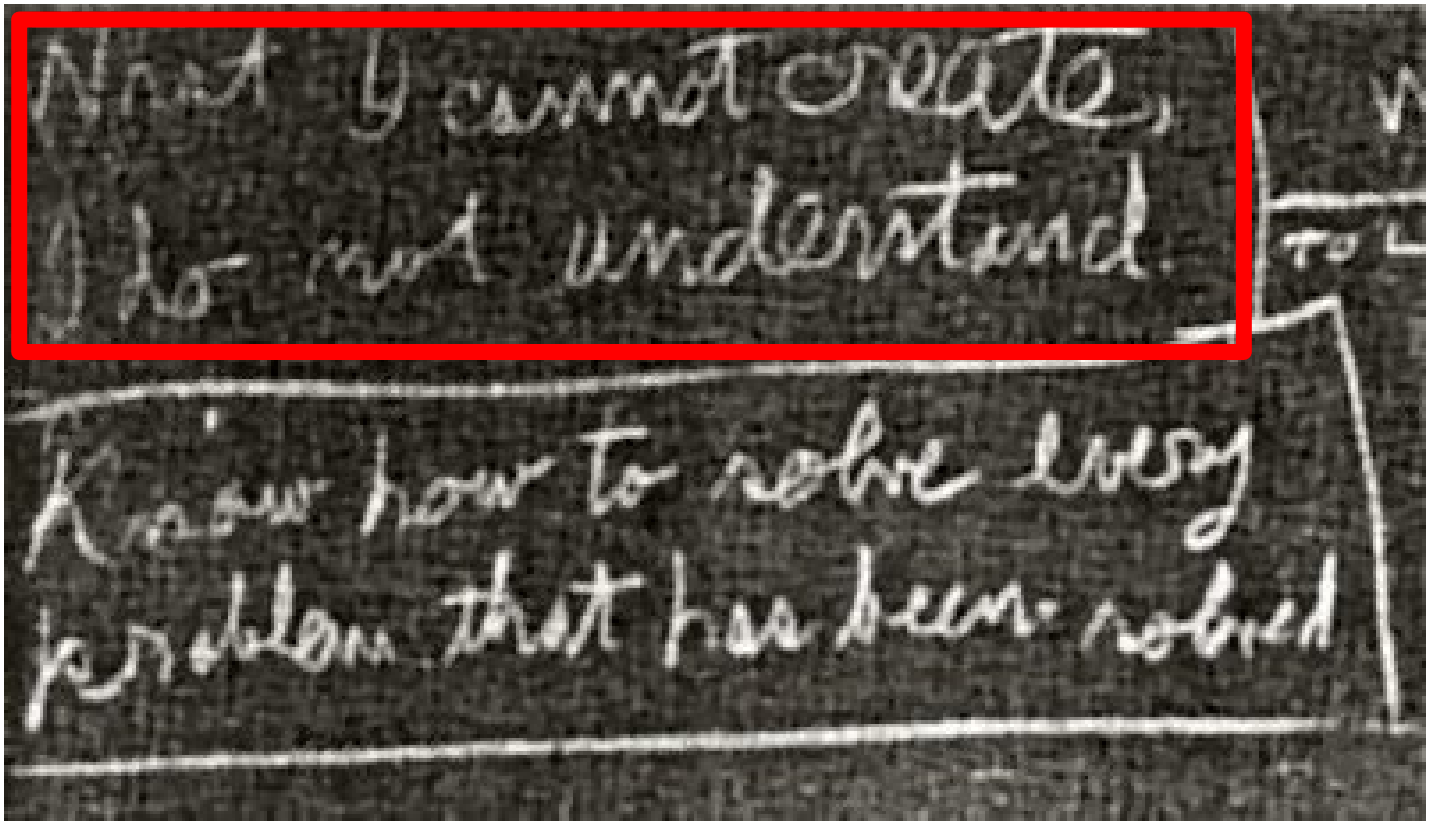


Motivation: Philosophical

- 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

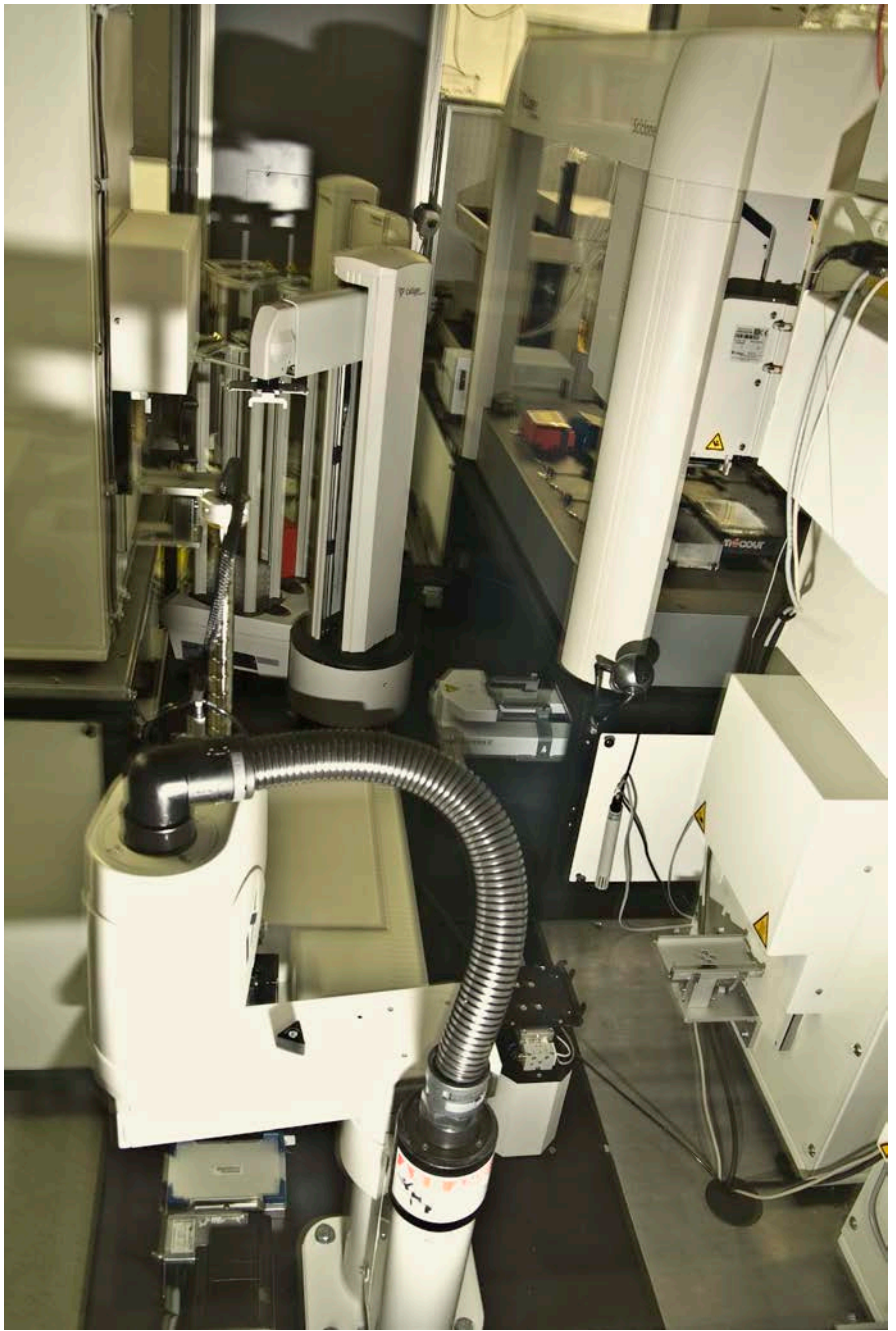
- 2008-2015 Eve – Drug Design for Tropical Diseases
  - Sophisticated Laboratory Automation: Collaboration with Steve Oliver (Cambridge)

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

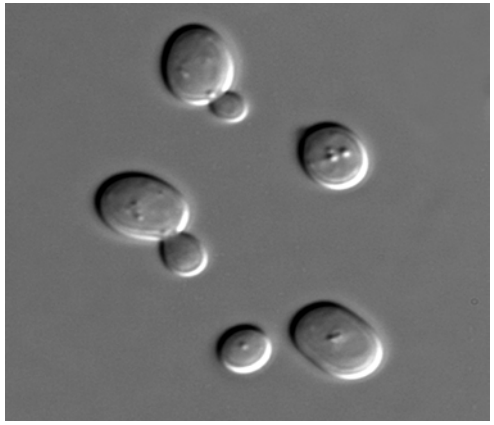
- 2015-2019 Eve – Systems Biology

  Coutant et al. (2019) *Proc. Nat. Acad. Sci. U.S.A*. 201900548

Adam



The Application Domain

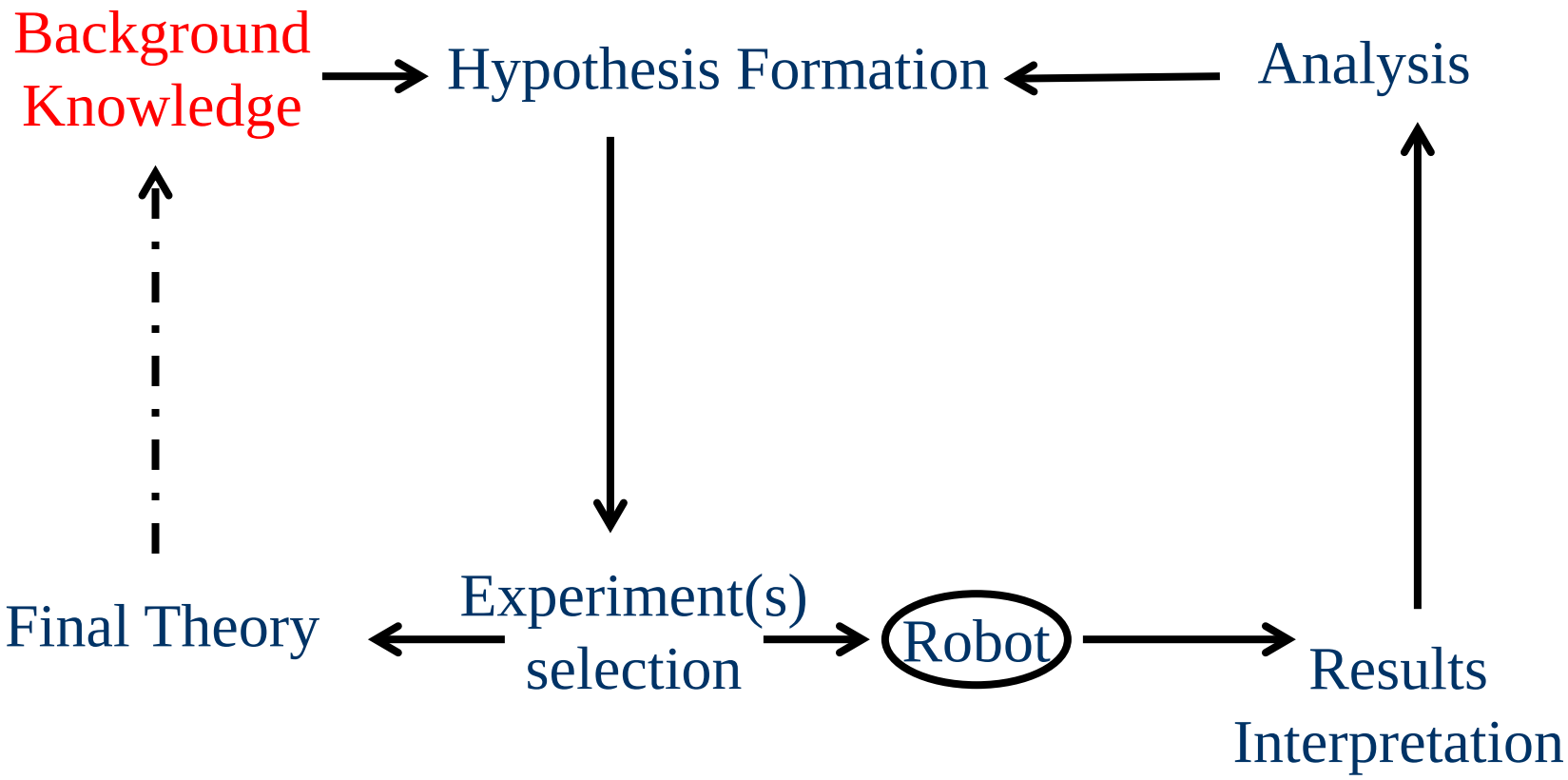


- Functional genomics
- In yeast (*S. cerevisiae*) ~20% of the 6,000 genes still have no known function.
- EUROFAN 2 made all viable single deletant strains.
- Task to determine the “function” of a gene by growth experiments.

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.

The Experimental Cycle

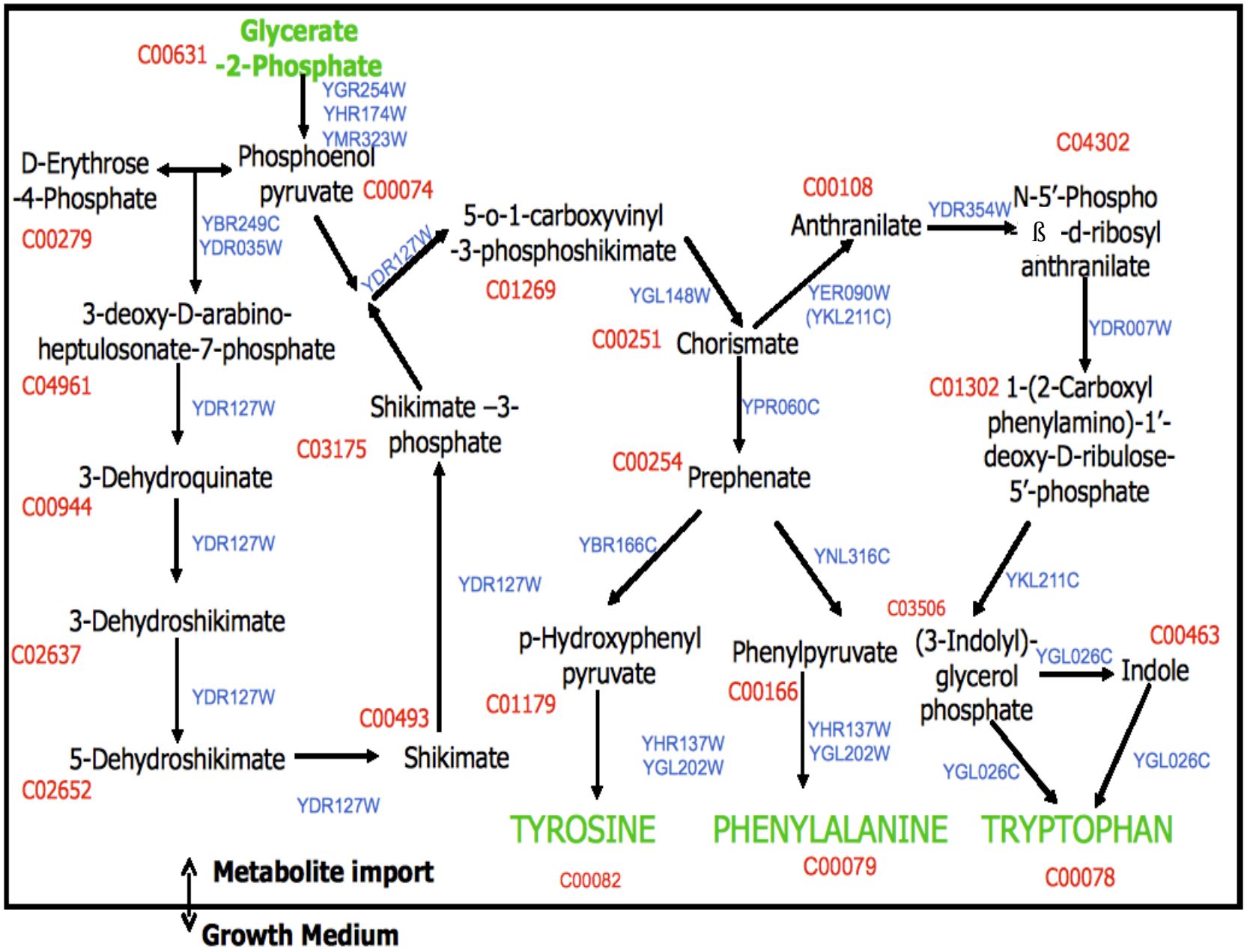


Logical Cell Model

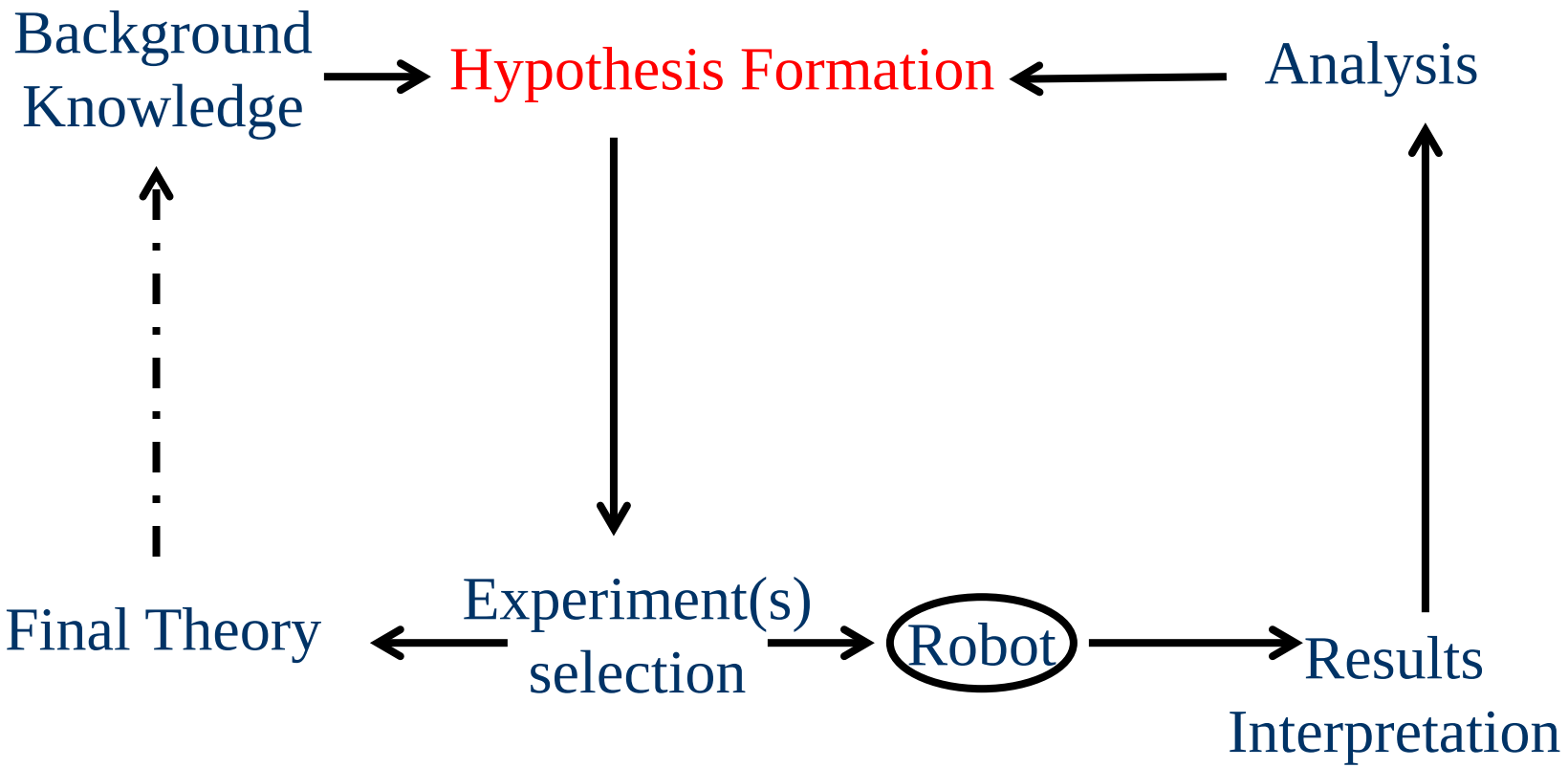
- We developed a logical formalism for modelling metabolic pathways (encoded in logic programs). This is essentially a directed labeled hyper-graph: with metabolites as nodes and enzymes as arcs.

- If a path can be found from cell inputs (metabolites in the growth medium) to all the cell outputs (essential compounds) then the cell can grow.

Phenyalanine, Tyrosine, and Tryptophan Pathways for *S. cerivisae*



The Experimental Cycle



Inferring Hypotheses

- Science is based on the hypothetico-deductive method.
- In the philosophy of science. It has often been argued that only humans can make the "leaps of imagination" necessary to form hypotheses.
- In biology most hypothesis generation is abductive, not inductive.
- Adam used abductive inference to infer missing arcs/labels in its metabolic graph - hypotheses. With these missing nodes Adam could then deductively infer (explain) the observed 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.

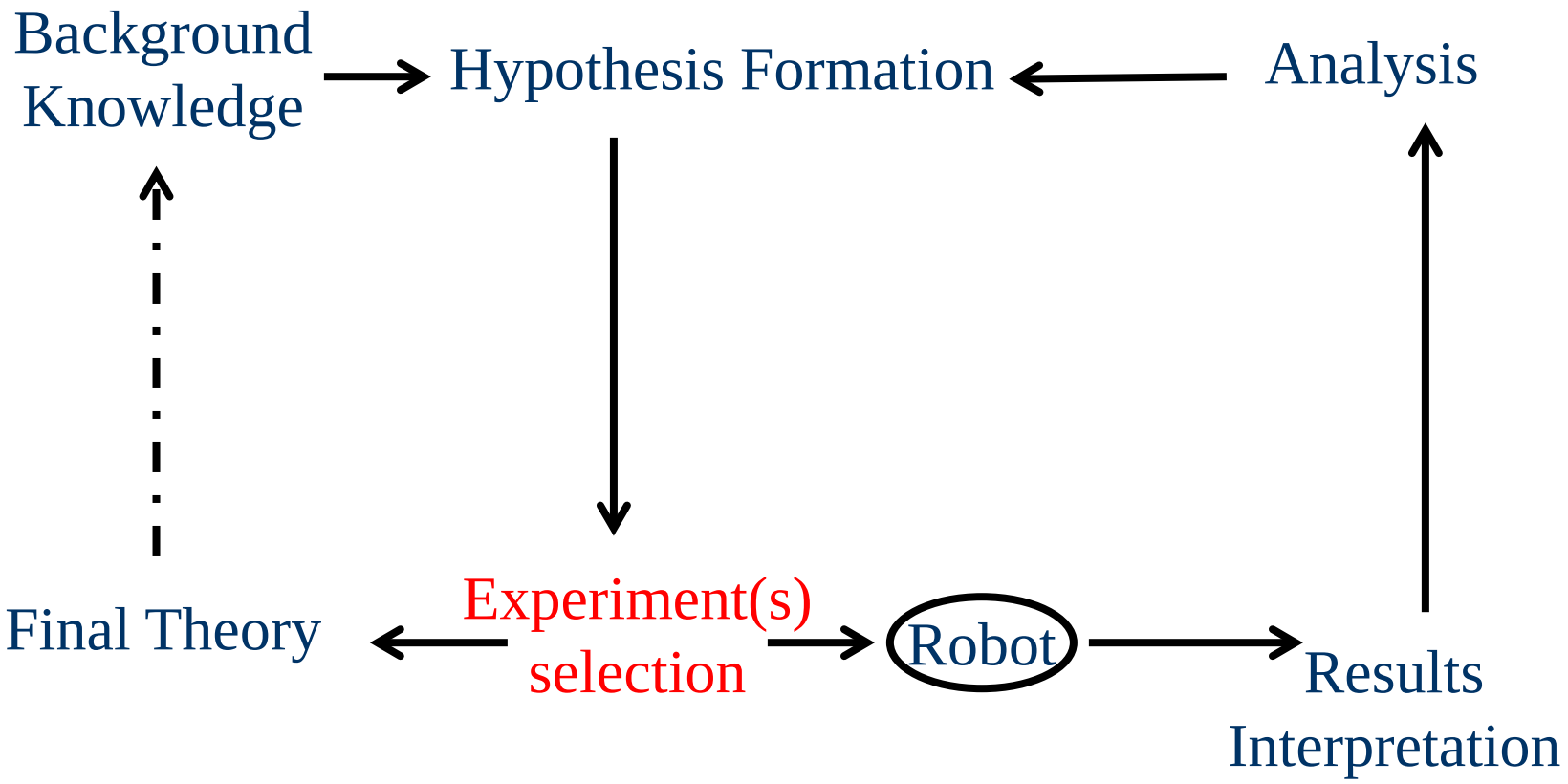
Abduction

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.

The Experimental Cycle



Inferring Experiments

Given a set of hypotheses we wish to infer an experiment that will efficiently discriminate between them

Assume:

- Every experiment has an associated cost.
- Each hypothesis has a probability of being correct.

The task:

- To choose a series of experiments which minimise the expected cost of eliminating all but one hypothesis.

Active Learning

- In the 1972 Fedorov (Theory of optimal experiments) showed that this problem is in general intractable (NP complete).
- However, it can be shown that the problem is the same as finding an optimal decision tree; and it is known that this problem can be solved “nearly” optimally in polynomial time.
- We have shown that this strategy can outperform (get answer *faster* and *cheaper*) than simply choosing the cheapest experiment. Also better than humans on test problem.

Recurrence Formula

EC(H,T) denote the minimum expected cost of experimentation given the set of candidate hypotheses *H* and the set of candidate trials *T*:

$$EC(\varnothing, T) = 0$$

$$EC(\{h\}, T) = 0$$

$$EC(H, T) \approx \min_{t \in T}\left[C_t + p(t)(mean_{t' \in (T - t)} C_{t'})J_{H[t]} + (1 - p(t))mean_{t' \in (T - t)} C_{t'})J_{H[\bar{t}]}\right]$$

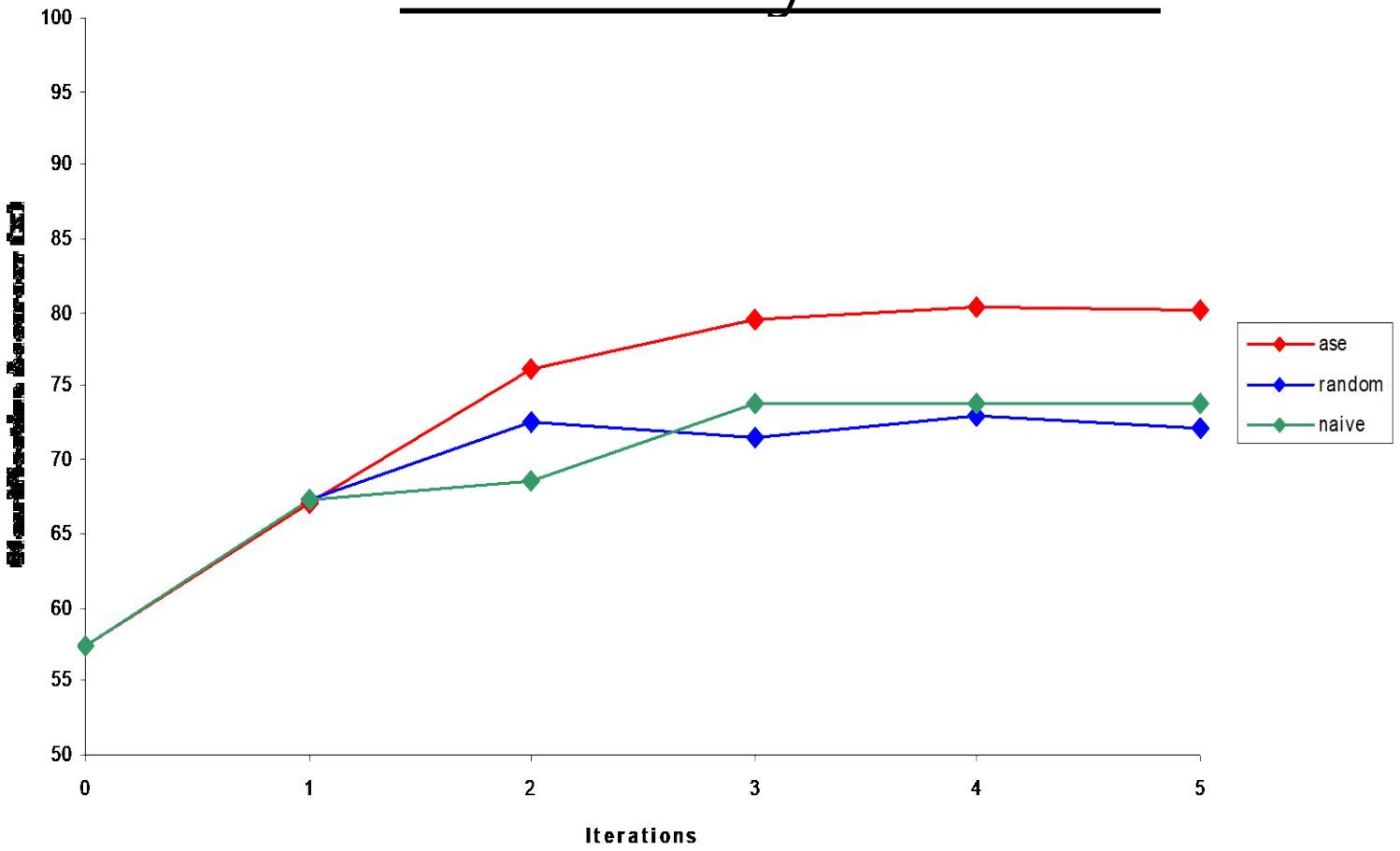
$$J_H = -\sum_{h \in H} p(h)\lfloor \log_2(p(h))\rfloor$$

C_t is the monetary price of the trial *t*

p(t) is the probability that the outcome of the trial *t* is positive

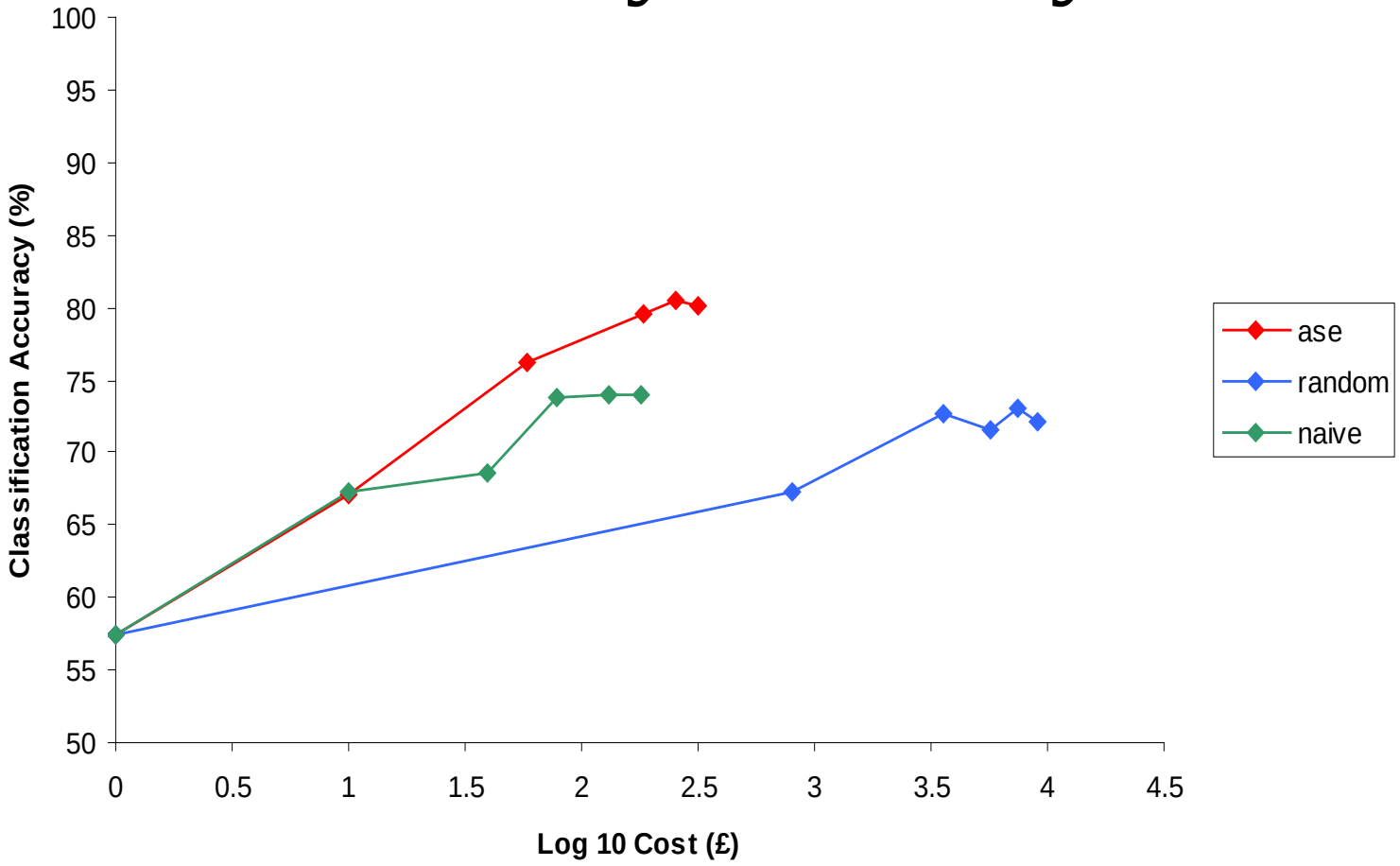
p(t) can be computed as the sum of the probabilities of the hypotheses (*h*) which are consistent with a positive outcome of *t*

Accuracy v Time



At the end of the 5th iteration: ASE 80.1%, Naïve 74.0%, Random 72.2%. ASE was significantly more accurate than either Naïve ($p < 0.05$) or Random ($p < 0.07$) using a paired t-test.

Accuracy v Money



Given a spend of ≤£$10^{2.26}$, ASE 79.5%, Naïve 73.9%, Random 57.4%. ASE was significantly more accurate than either Naïve ($p < 0.05$) or Random ($p < 0.001$).

The Experimental Cycle

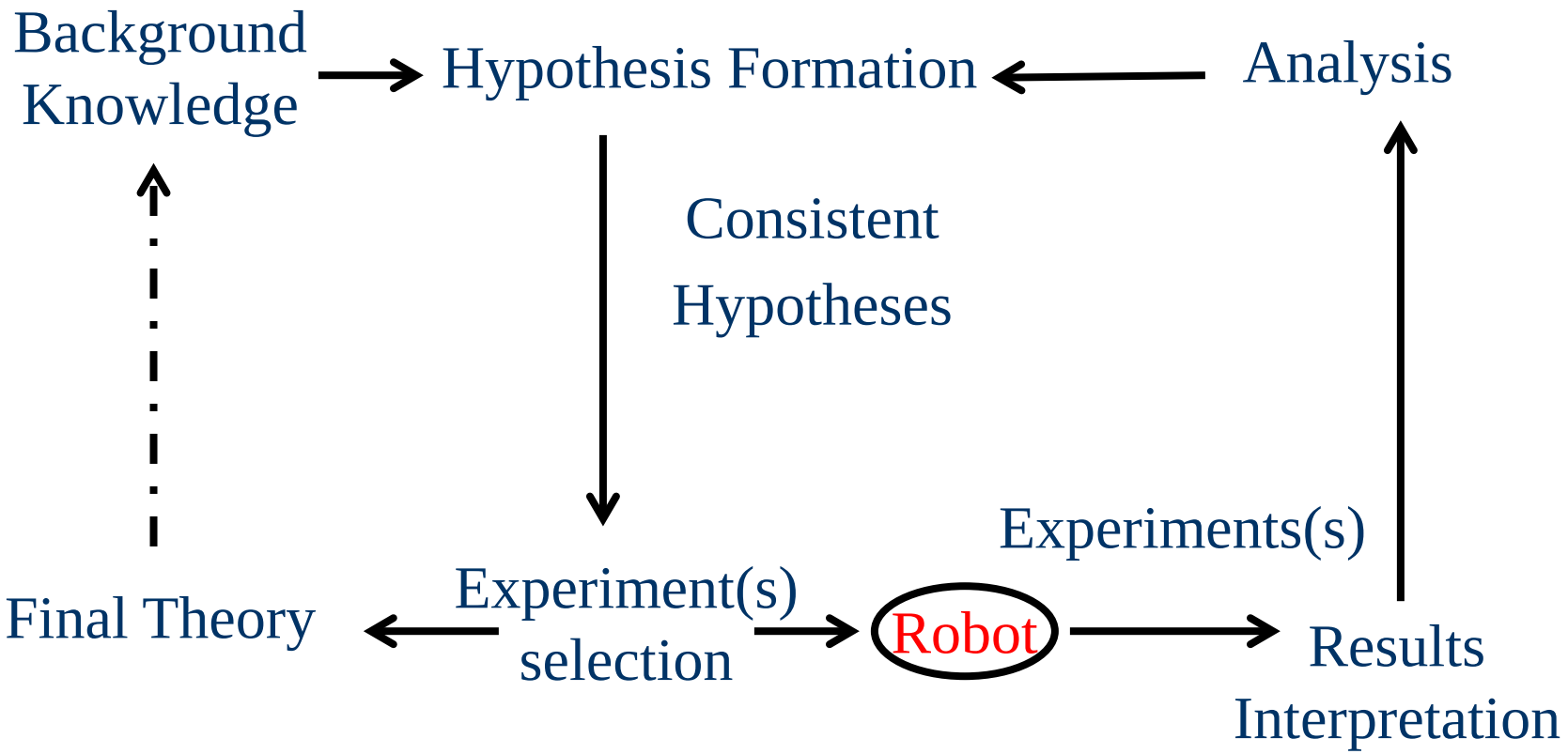
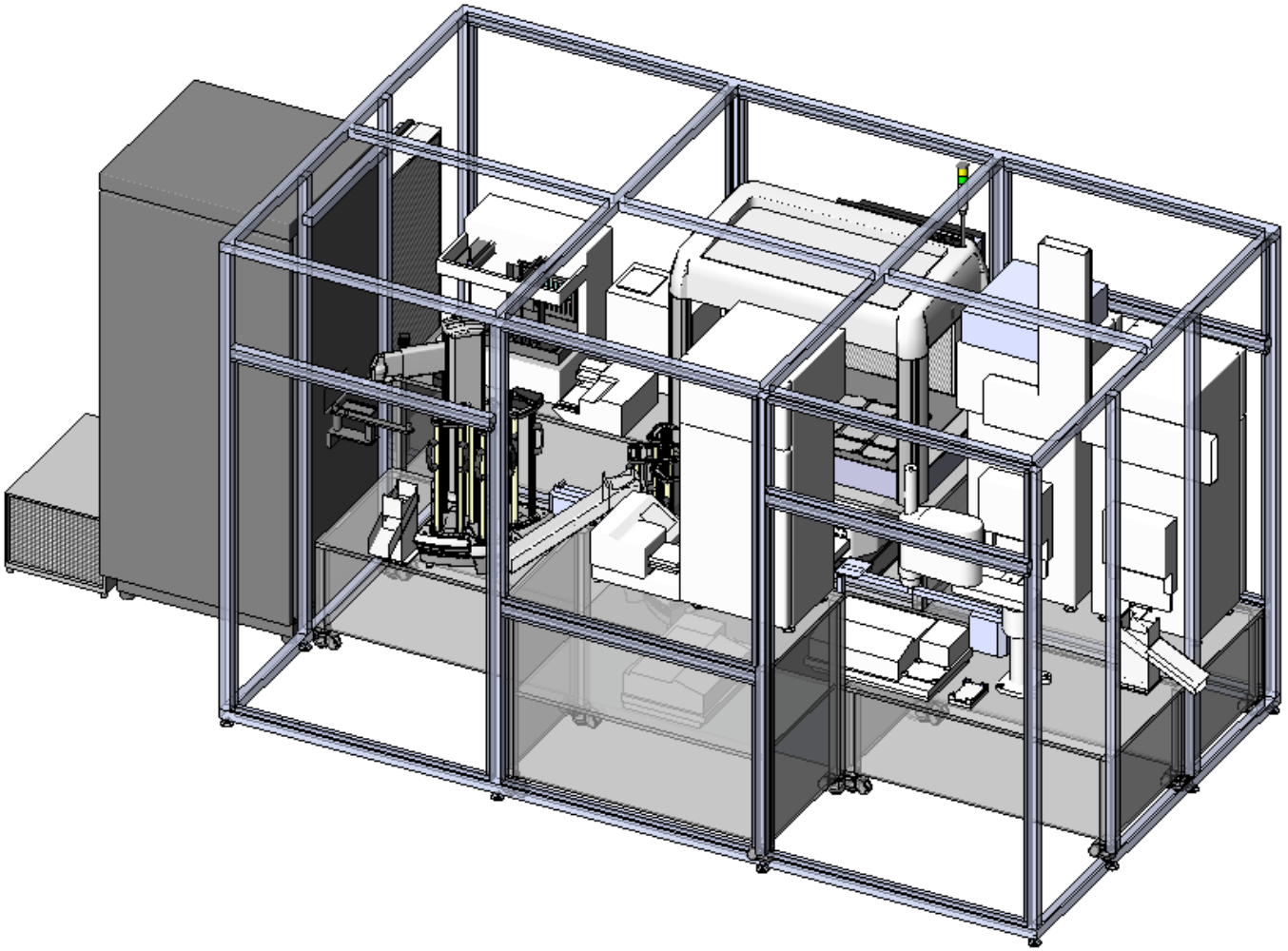


Diagram of Adam



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

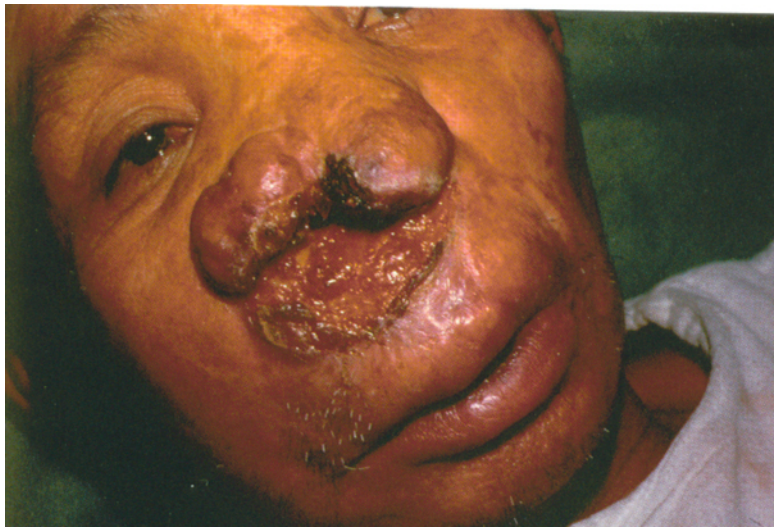
Orphan Enzyme		Hypothesised Gene	Prob.	Acc.	No.	Existing Annotation	Dry	Wet
1	glucosamine-6-phosphate deaminase (3.5.99.6)	YHR163W (SOL3)	$<10^{-4}$	97	8	'6-phosphogluconolactonase' ida	-	-
2	glutaminase (3.5.1.2)	YIL033C (BCY1)	$<10^{-4}$	92	11	'cAMP-dependent protein kinase inhibitor ' ida	✗ ?	-
3	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	✓	-

Eve

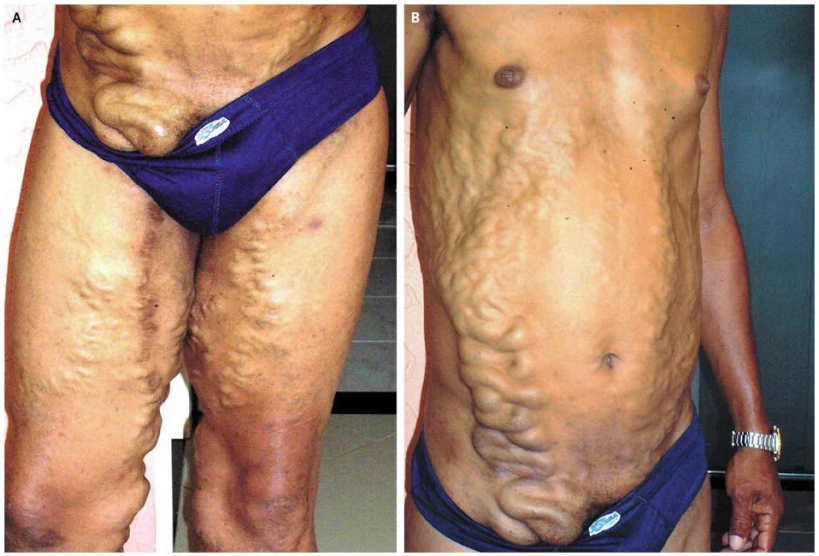
Drug Design

The Application Domain

Malaria



Shistosomaisis



Leishmania

Chagas

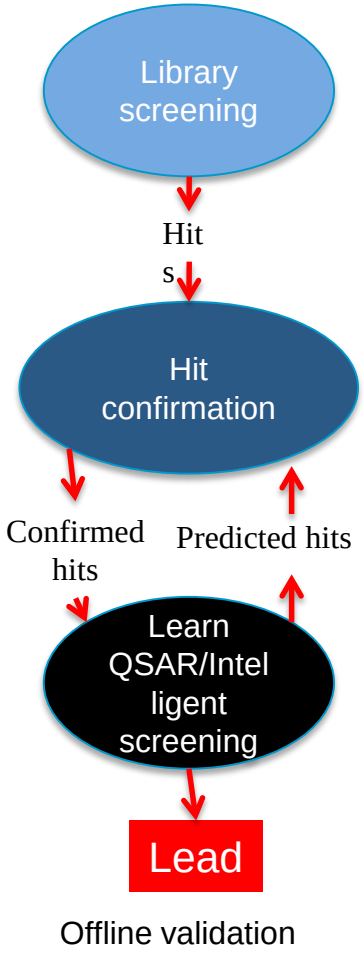
Why Tropical Diseases?

- 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 standard chemoinformatic methods (Boolean fingerprints) to represent background knowledge.
- Uses induction (quantitative structure activity relationship – QSAR learning) to infer new hypotheses.
- Use a combination of active learning and active learning to decide efficient experiments.

Eve's Automation of Pipeline



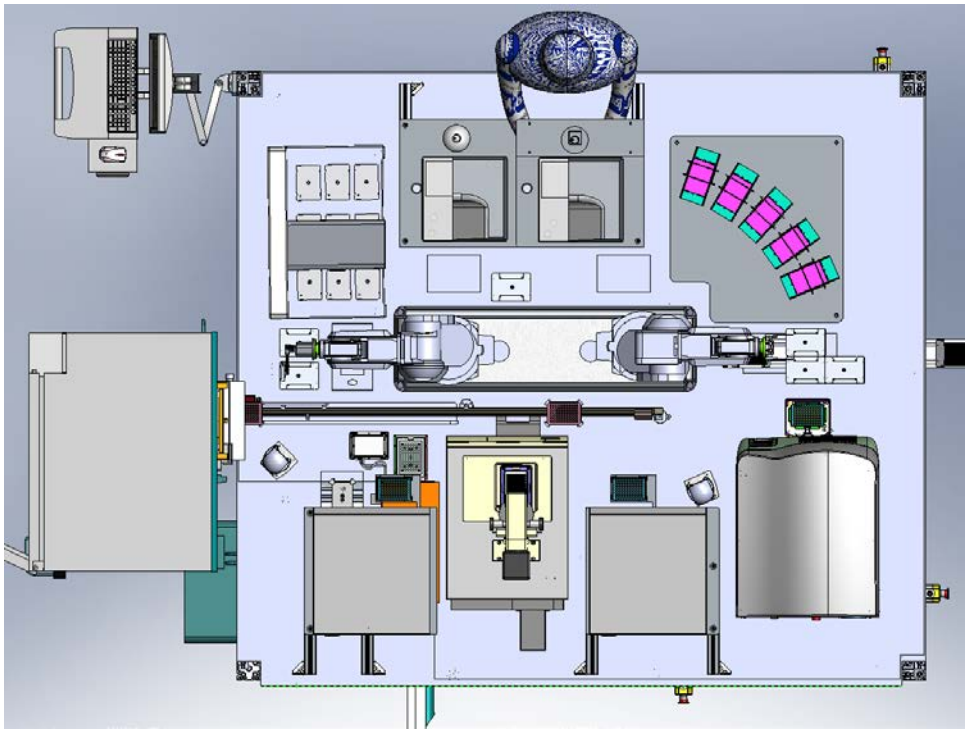
- Standard library screening is brute force:
- Eve uses intelligent screening

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

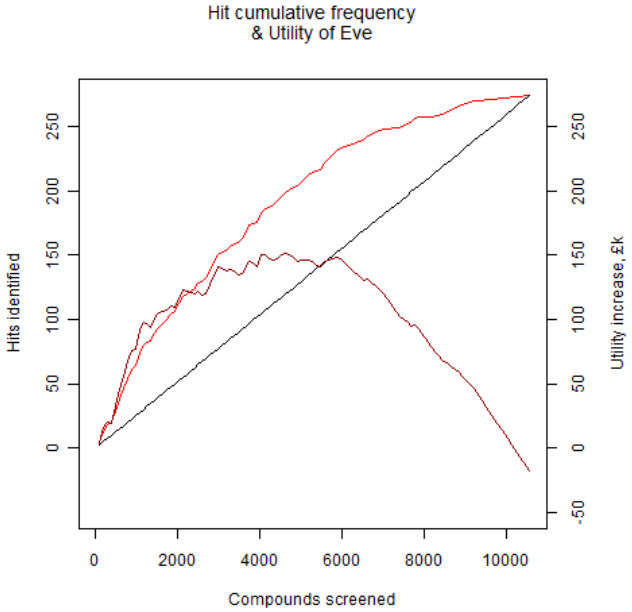
Eve’s Hardware

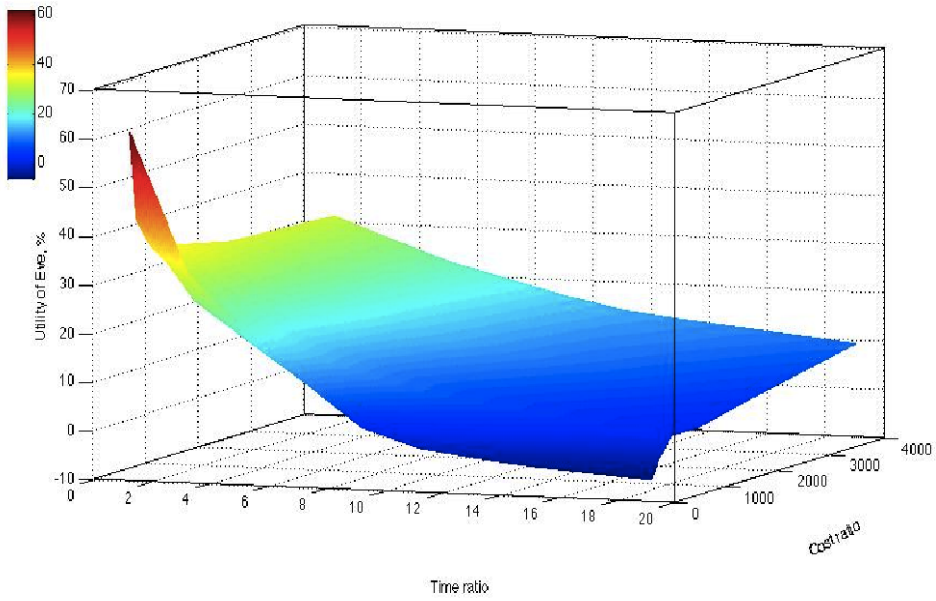
Highlights of Eve's hardware:

- 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

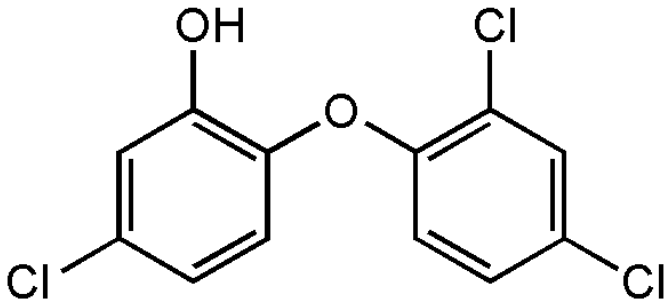




$$\Delta\, Utility\ of\ Eve = \sum_{1}^{Nm}(Tm + Cm) + \sum_{1}^{Nx}(Tc + Cc - Uh) + \sum_{1}^{Ne}(Tm - Tc + Cm - Cc)$$

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



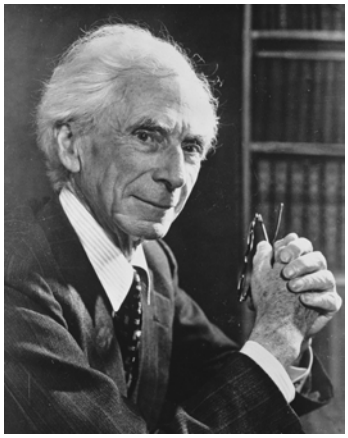
- Simple compound
- Known to be safe – used in toothpaste.
- Targets both DHFR and FAS-II – well established targets.
- Demonstrated activity using multiple wet experimental techniques.
- Works against wild-type and drug-resistant *Plasmodium falciparum*, and *Plasmodium vivax*.

Relational Learning

Relational Learning

- The use in machine learning of representations based of tuples of attributes is in essence the use of propositional logic.
- Use of the richer, more expressive, language of 1st-order predicate logic is termed Relational Learning.
- Not fashionable now, but in my view, sooner or later relational learning will become essential.
- Used to be considered too inefficient to be practical, but Deep Neural Networks have moved the bar here!

1st-Order Predicate Logic



- Developed a 100 years ago to describe mathematical structure.
- Propositional logic assumes the world contains facts.
- 1st-order Predicate logic assumes the world contains:
  - Objects: numbers, molecules, atoms, stars…
  - Relations: greater-than, bonds, distances …
  - Functions: plus1,charge…

Machine Learning: Attributes

- Propositional Logics: Statistical, and neural network, methods use tuples (vectors) of attributes (descriptors) to represent examples.
- An attribute is something true about a whole example.
- Examples can be put into a single row in a table, the columns are the attributes.

Attributes and Chemical Structure

- Difficult to represent arbitrary chemical structure using propositions, i.e. a single table. Either the table grows exponentially, or a compromise must be made in fidelity.

- What should the attributes be?
  - Boolean fingerprints
  - topological indices
  - Voxols
  - Quantum mechanic descriptors

Relational Learning

- Uses richer representation language

  - Enables relations: abstract structure

  - Explicit background knowledge (constraints): expressed as computer programs. Arbitrary complex pieces of prior knowledge.

  - No need for examples to be a single row in a table

King, R.D., et al. (1996) *Proc. Nat. Acad. Sci. U.S.A.* **93**, 438-442.

Relational Approach: Atoms

- The predicate atom is used, e.g. in compound 127 (3,4,3’-trinitrobiphenyl).
- atom(127, 127_1, c, 22, 0.191).
- States that in compound 127, atom no. 1 is a carbon of type 22 (aromatic carbon in a six membered ring) with a partial charge of 0.191.

Relational Approach: Bonds

- The predicate bond is used, e.g.

  bond(127, 127_1, 127_6, 7).

- States that in compound 127, atom no. 1 and atom no. 6 are connected by a bond of type 7 (aromatic bond).

High Level Background Knowledge 1

A definition of a Methyl group is:

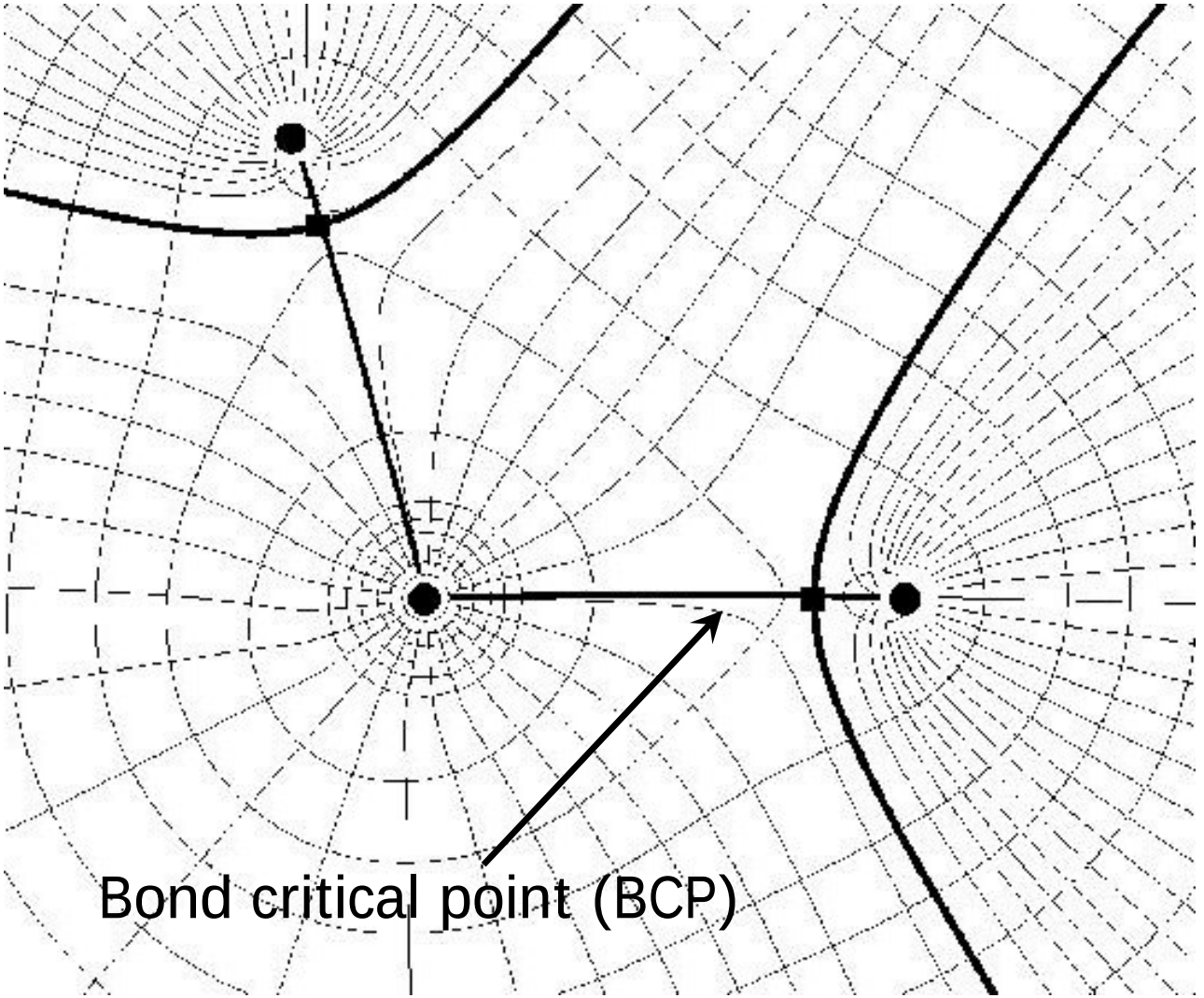
```
methyl(Drug,[Atm0,Atm1,Atm2,Atm3,Atm4]) :
    atom(Drug, Atm0, Type,…,…),          % Link atom
    Type not h,
    atom(Drug, Atm1, c. 10, …),          % Aryl carbon
    atom(Drug, Atm2, h, 3, …),           % Hydrogen
    atom(Drug, Atm3, h, 3, …),           % Hydrogen
    atm(Drug, Atm4, h, 3, …),            % Hydrogen
    bond(Drug, Atm0, Atm1,1),            % Single bond
    bond(Drug, Atm1, Atm2, 1),           % Single bond
    bond(Drug, Atm1, Atm3, 1),           % Single bond
    bond(Drug, Atm1, Atm4, 1),           % Single bond
    Atm2 @> Atm3,
    Atm3@> Atm4.                         % Atm2,3,4 different
```

High Level Background Knowledge 2

- Extend to 3D: simply add Pythagoras' theorem.
- Extend to multiple confirmations: simply add and label confirmations.
- Extend to quantum mechanics.

Richard F.W. Bader

3D Electron Density



Example StruQT predicates

- cp(Id,Field,[x,y,z],Hessian) % critical point
- signature(Id,Signature). % signature of Hessian
- rank(Id,Rank). % rank of Hessian
- con_path(Id_1,Id_2) % connectivity of points
- hessian(Id,L_1,L_2,L_3) % eigenvalues of Hessian
- nuc_attr(Id). % nuclear attractor
- bcp(Id) % bond critical point
- rcp(Id). % ring critical point
- ccp(Id). % cage critical point

Alsberg, et al. (2001) *Analytica Chimica Acta* **446**, 3-13

Transformational Learning

The New Idea

- The new idea is to represent examples using predictions made by ML models learnt on other tasks: extrinsic features.
- This transformation has the dual advantages of significantly improving predictive performance, and providing scientific insight.

Standard Machine Learning

+

Size	Ears	Cute	Donkey
Big	Big	No	1.0
Small	Big	No	0.3

→ Donkey

+

Size	Ears	Cute	Kitten
Small	Small	Yes	1.0
Small	Small	No	0.1

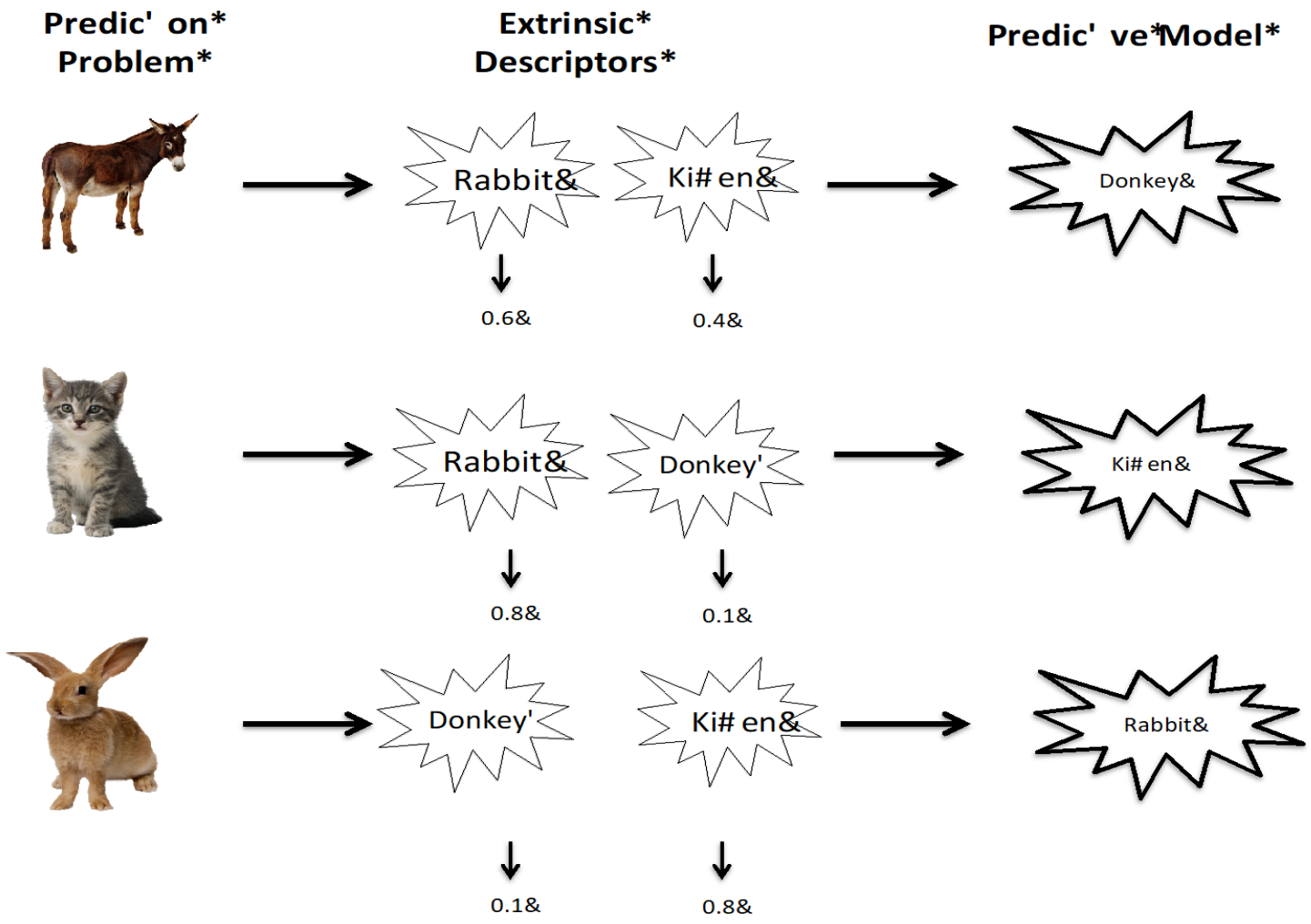
→ Kitten

+

Size	Ears	Cute	Rabbit
Small	Big	Yes	1.0
Big	Small	Yes	0.2

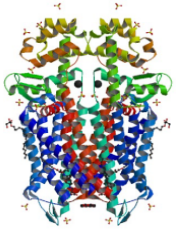
→ Rabbit

Transformative Machine Learning

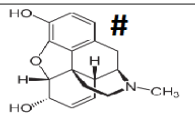
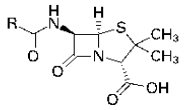


Standard Machine Learning

Predic' on*Problem* **Intrinsic*Descriptors*** **Predic' ve*Model***



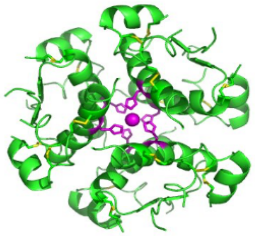
+"

	G1#	G2#	...#	G1024#	Activity#
 #	#	#	#	#	#
	1#	0#	1#	1#	0.9#
...#	...#	...#	...#	...#	...#
	#	#	#	#	#
	0#	0#	1#	1#	0.3#

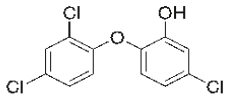
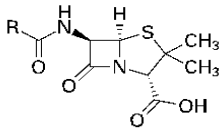
→ $F_1(X)$

"

...



+"

	G1#	G2#	...#	G1024#	Activity#
	#	#	#	#	#
	0#	0#	1#	0#	0.5#
...#	...#	...#	...#	...#	...#
	#	#	#	#	#
	1#	0#	1#	1#	0.4#

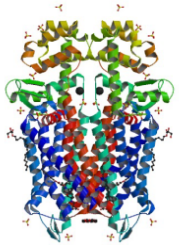
→ $F_n(X)$

"

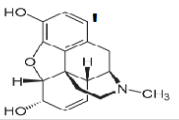
...

Transformative Machine Learning

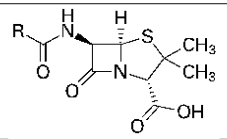
Predic' on*Problem* **Extrinsic*Descriptors*** **Transforma' ve* Learning* Predic' ve*Model***



+"



	$\mathcal{F}_1$	$\mathcal{F}_2$	...	$\mathcal{F}_n$	Activity'
	0.3'	0.4'	0.7'	0.9'	0.9'
...'	...'	...'	...'	...'	...'
	0.1'	0.8'	0.7'	0.5'	0.3'



→ $T_1(X)$

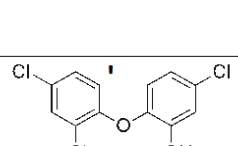
"

...

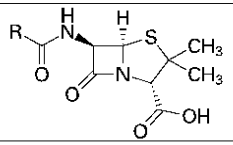
"

...

+"



	$\mathcal{F}_1$	$\mathcal{F}_2$	...	$\mathcal{F}_n$	Activity'
	0.2'	0.5'	0.9'	0.1'	0.5'
...'	...'	...'	...'	...'	...'
	0.5'	0.8'	0.9'	0.1'	0.4'



→ $T_n(X)$

Intuition

- Transformational learning combines ideas from multi-task learning and transfer learning.
- Instead of using a predefined similarity measure to pre-select a set of similar tasks, we project the different tasks into one joint numeric representation.
- Then use a meta-learning algorithm to learn from this new representation how to make accurate predictions for the task at hand.

Experimental Setup – Machine Learning Methods

- To in investigate the utility of Transformative Learning we selected four machine learning methods:
  - Random forests (RF).
  - Support-vector machines (SVMs).
  - k-nearest neighbour (KNN).
  - Neural-networks (NN).

- These represent the main families of non-linear Machine Learning methods.

Experimental Setup – Problems

- We applied the four machine learning methods to three typical real-world scientific problems:
  - Drug-design (quantitative structure activity relationship learning - QSAR). 2,219 target proteins and associated chemical compounds and activities from the ChEMBL database
  - Predicting 1,000 human gene expression (across different tissue types and drug treatments) – LINCS.
  - Meta-machine learning (predicting how well machine learning method will work on problems).

Results

Problem Area	ML Method	Standard NRMSE (RMSE)	Transformed NRMSE (RMSE)	Stacked NRMSE (RMSE)	Difference % NRMSE (RMSE)
QSAR					
	RF	0.164 (0.651)	0.148 (0.632)	0.148 (0.632)	+10.05 (3.07)
	SVM	0.169 (0.672)	0.152 (0.649)	0.152 (0.649)	+10.10 (3.42)
	KNN	0.167 (0.711)	0.171 (0.734)	0.167 (0.711)	0 (0)
	NN	?	?		?
LINCS					
	RF	0.245 (0.069)	0.236 (0.066)	0.236 (0.066)	+3.68 (4.32)
	SVM	0.255 (0.069)	0.227 (0.068)	0.227 (0.068)	+10.9 (2.18)
	KNN	0.257 (0.072)	0.249 (0.072)	0.249 (0.072)	+2.92 (-0.28)
	NN	0.260 (0.074)	0.263 (0.071)	0.260 (0.074)	0 (0)
Meta-ML					
	RF	0.257 (0.118)	0.124 (0.053)	0.124 (0.053)	51.8 (55.57)
	SVM	0.296 (0.134)	0.214 (0.097)	0.214 (0.097)	27.8 (27.2)
	KNN	0.290 (0.133)	0.274 (0.126)	0.274 (0.126)	5.39 (5.40)
	NN	0.323 (0.148)	0.248 (0.113)	0.248 (0.113)	23.5 (23.7)

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

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.

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.
- Robot Scientists are needed for 21st century science.

Acknowledgments

- The Robot Scientist team: Manchester, Aberystwyth, Cambridge, Brunel, Leuven, Thailand. (BBSRC)