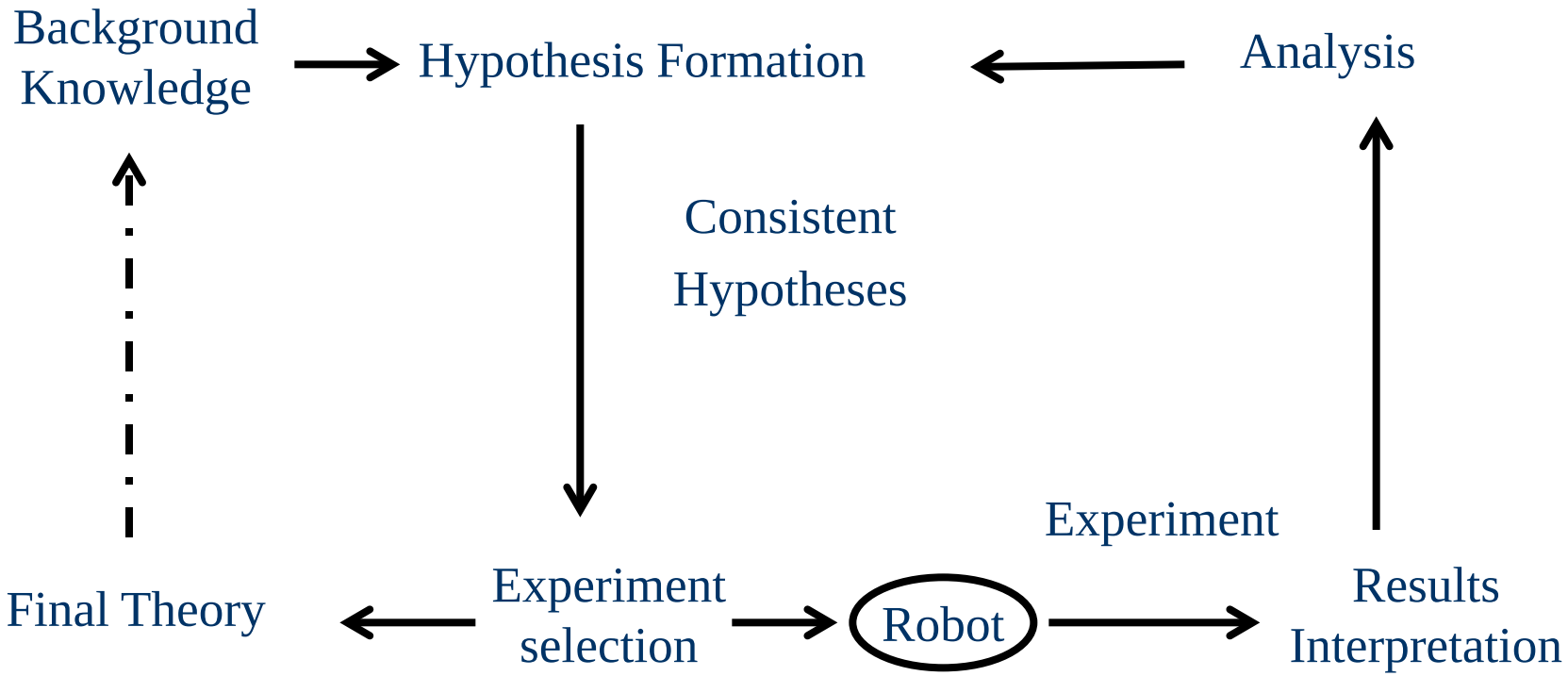
Automating Science

Ross D. King
University of Wales, Aberystwyth

Background

The Concept of a Robot Scientist

We have developed the first computer system that is capable of originating its own experiments, physically doing them, interpreting the results, and then repeating the cycle*.



Motivation: Philosophical

- What is Science?
- The question whether it is possible to automate the scientific discovery process 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.

Motivation: Technological

- In many areas of science our ability to generate data is outstripping our ability to analyse the data.
- One scientific area where this is true is in Systems Biology, where data is now being generated on an industrial scale.
- The analysis of scientific data needs to become as industrialised as its generation.

Technological Advantages

- *Robot Scientists have the potential to increase the productivity of science* - by enabling the high-throughput testing of hypotheses.

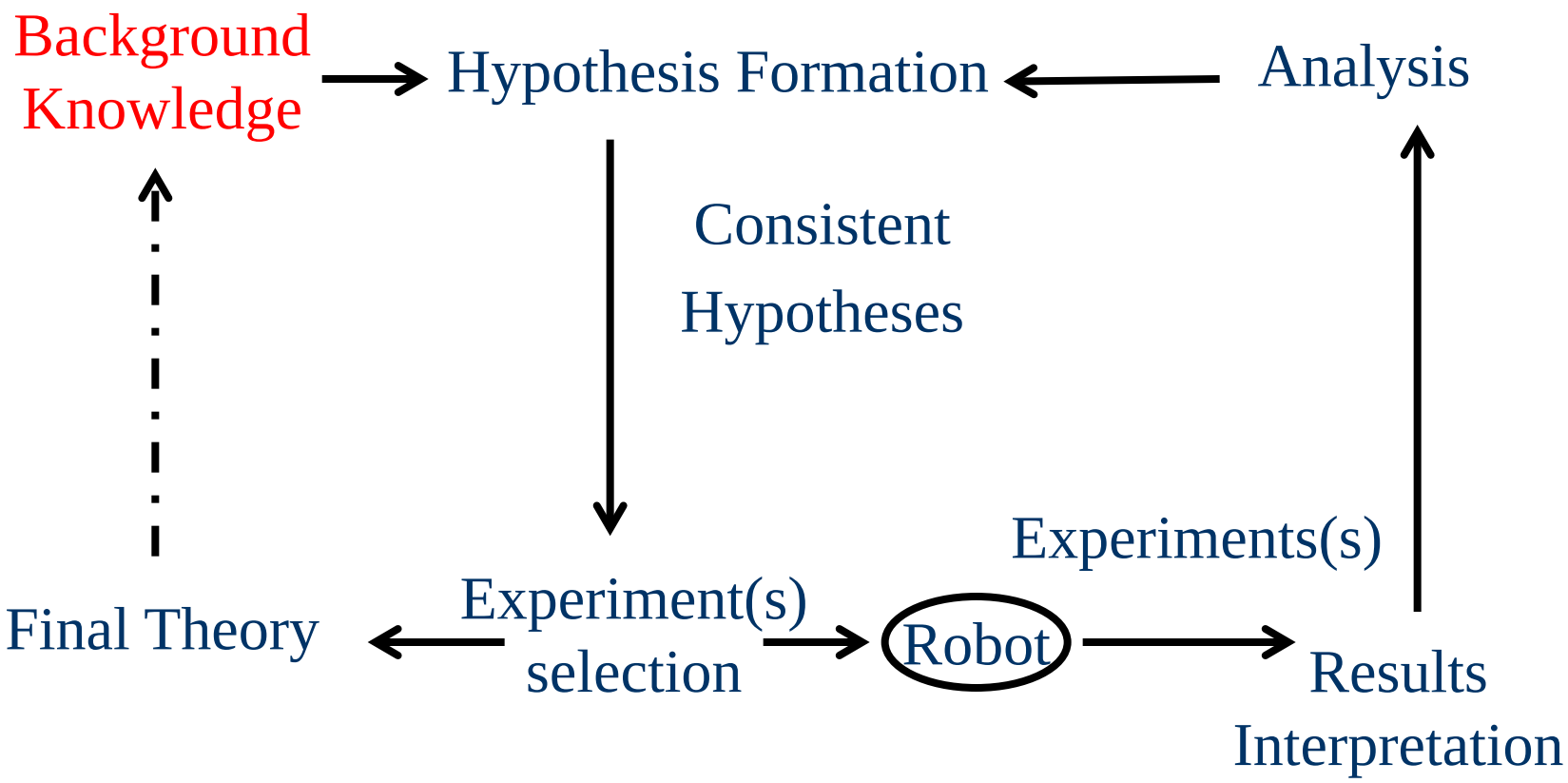
- *Robot Scientists have the potential to improve the repeatability and reuse of scientific knowledge* - by enabling the description of experiments in greater detail and semantic clarity

Scientific Discovery

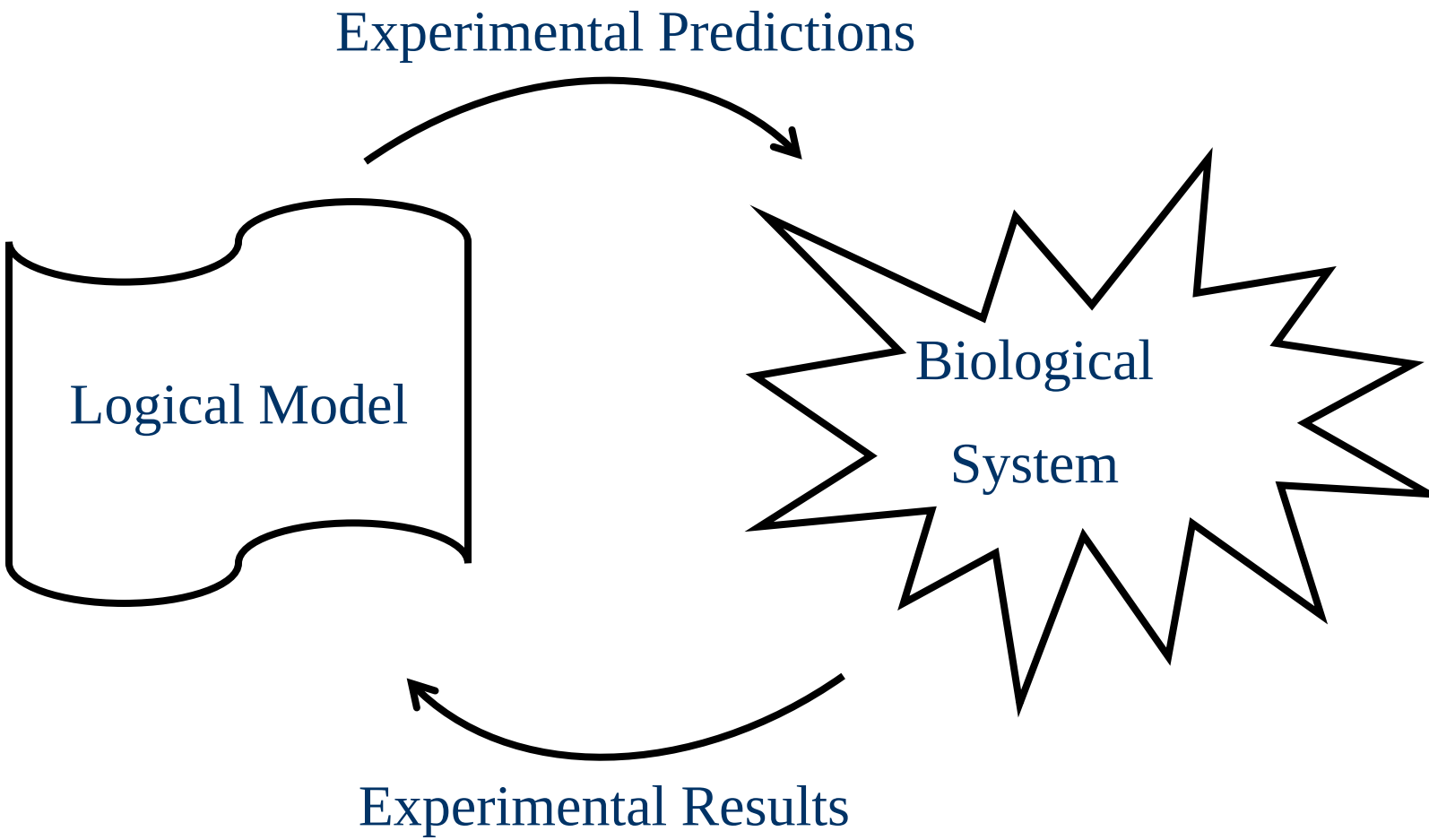
- Meta-Dendral: Analyis of mass-spectrometry data. Buchanan, Feigenbaum, Djerassi, Lederburg (1969).
- Bacon: Rediscovering physics and chemistry. Langley, Bradshaw, Simon (1979).
- Automated discovery in a chemistry laboratory. Zytkow, Zhu, Hussman (1990).

Adam

The Experimental Cycle



Model v Real-World



The Application Domain

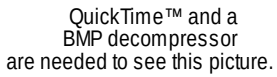
- Functional genomics
- In yeast (*S. cerevisiae*) ~15% of the 6,000 genes still have no known function.
- EUROFAN 2 has knocked out each of the 6,000 genes in mutant strains.
- Task to determine the “function” of the gene by growth experiments comparing mutants and wild type.

Logical Cell Model

- We have developed a logical formalism for modelling metabolic pathways (encoded in Prolog). 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.

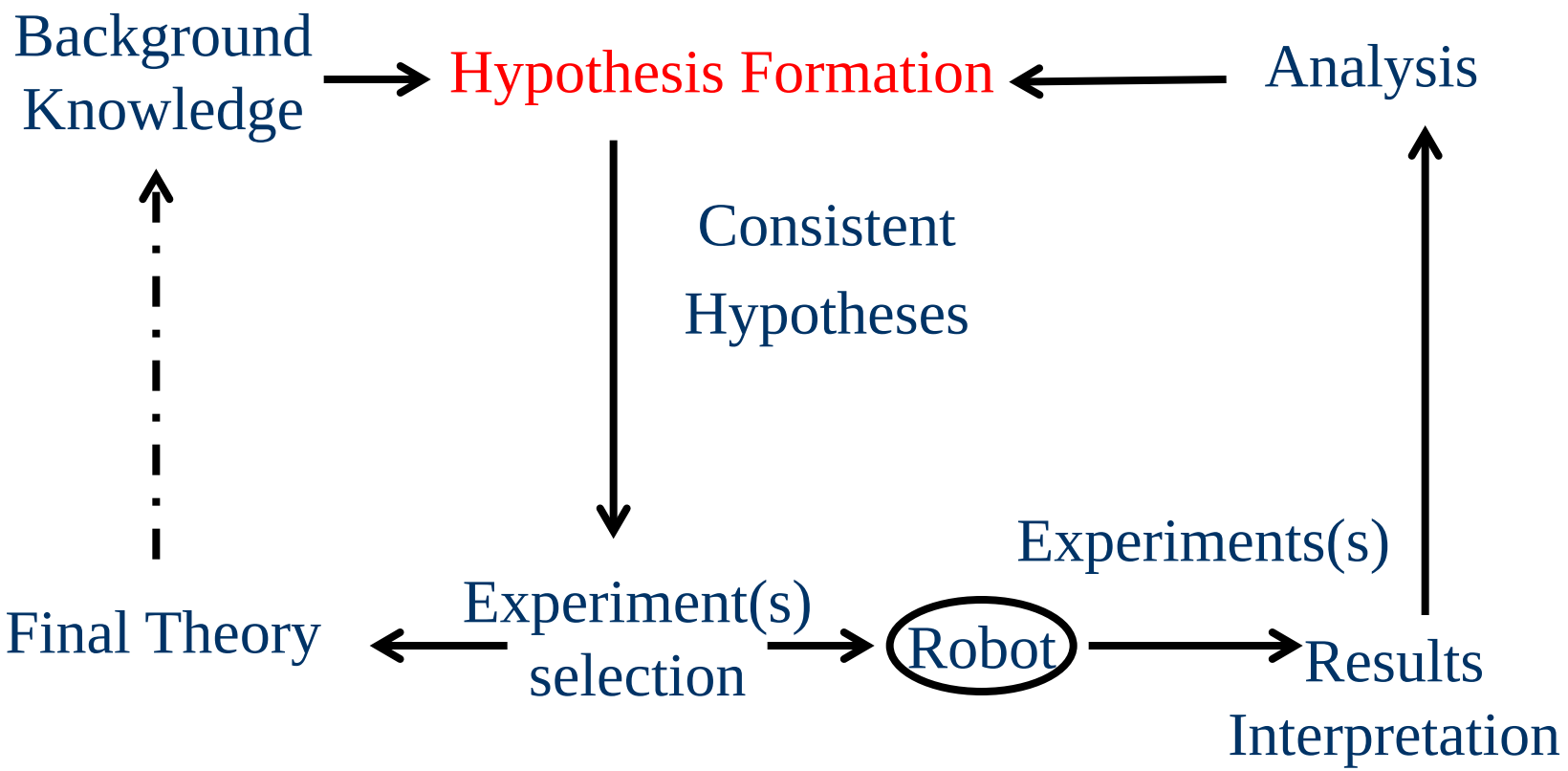
ß


QuickTime™ and a
BMP decompressor
are needed to see this picture.

Genome Scale Model of Yeast Metabolism

- It covers most of what is known about yeast metabolism.
- Includes 1,166 ORFs (940 known, 226 inferred)
- Growth if path from growth medium to defined end-points.
- State-of-the-art accuracy in predicting cell viability

The Experimental Cycle



Inferring Hypotheses

- In the philosophy of science. It has often been argued that only humans can make the “leaps of imagination” necessary to form hypotheses.

- We used Abduction to infer missing arcs/labels in our metabolic graph. With these missing nodes we can explain (deductively) all the experimental results.

Reiser et al., (2001) ETAI 5, 233-244;

Types of Logic

Deduction

Rule: If a cell grows then it can synthesise tryptophan.

Fact: cell cannot synthesise tryptophan

$\therefore$ Cell cannot grow.

Given the rule $P \rightarrow Q$, and the fact $\neg Q$, infer the fact $\neg P$

(*modus tollens*)

Abduction

Rule: If a cell grows then it can synthesise tryptophan.

Fact: Cell cannot grow.

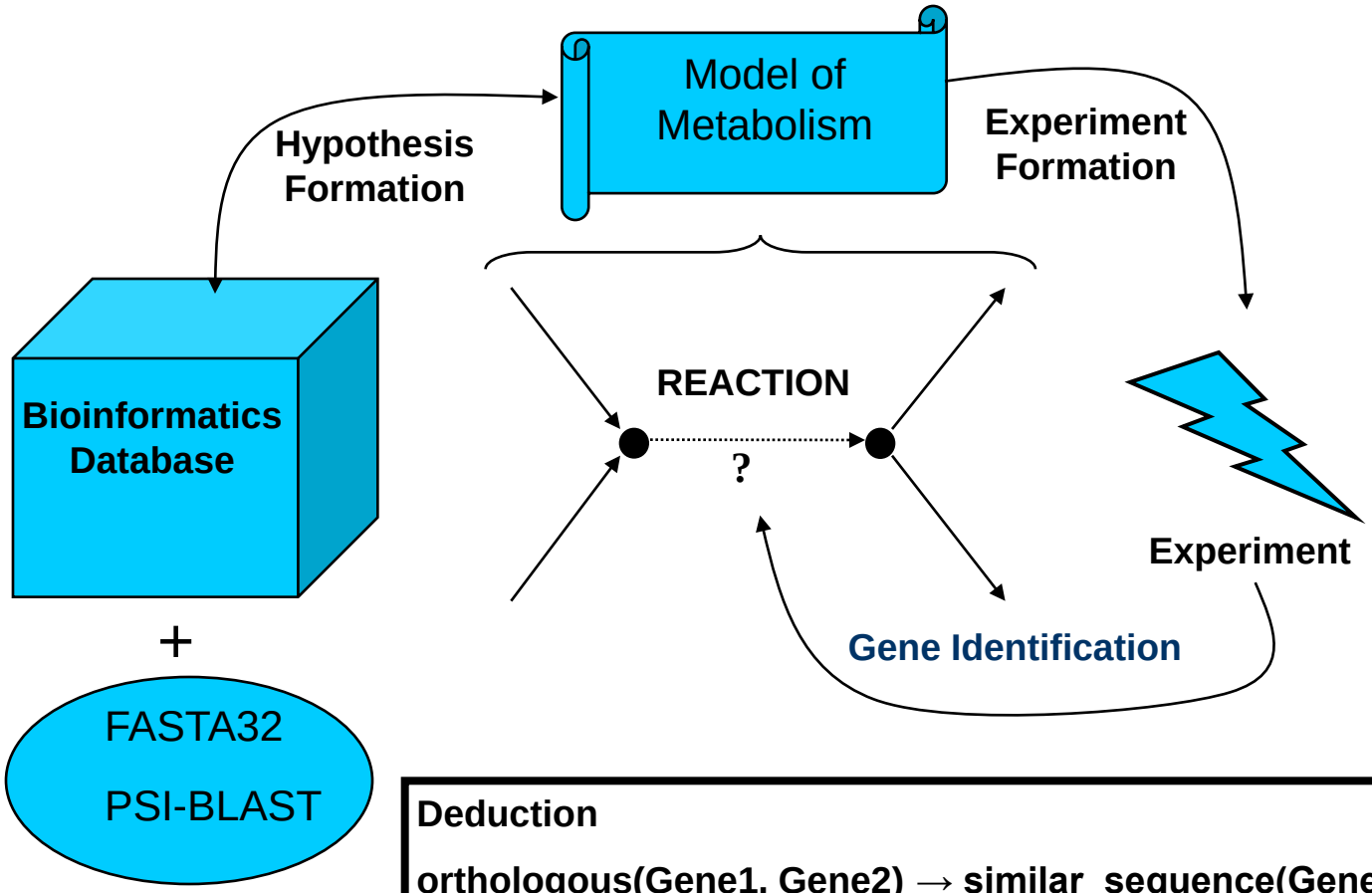
$\therefore$ Cell cannot synthesise tryptophan.

Given the rule $P \rightarrow Q$, and the fact $\neg P$, infer the fact $\neg Q$

Orphan Enzymes

- Our model of yeast metabolism has “locally orphan enzymes” enzymes which catalyse biochemical reactions known to be in yeast, but which do not have identified parent genes
- We use bioinformatics to abduce genes which encode for these orphan enzymes.

Automated Model Completion



Deduction

orthologous(Gene1, Gene2) → similar_sequence(Gene1, Gene2).

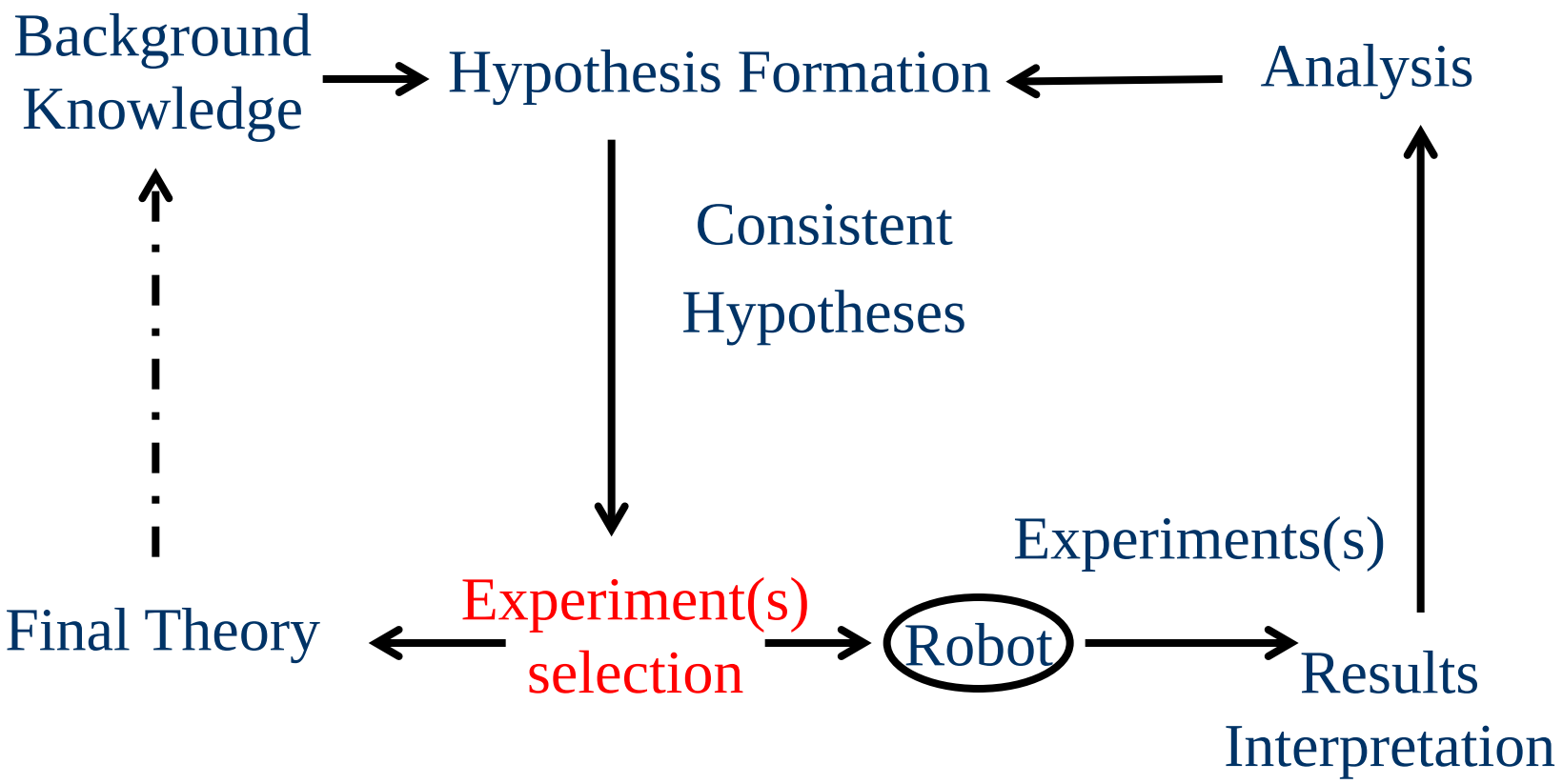
Abduction

similar_sequence(Gene1, Gene2) → orthologous(Gene1, Gene2).

ß


QuickTime™ and a
BMP decompressor
are needed to see this picture.
?

The Experimental Cycle



Form of the Experiments

- Hypothesis 1: Gene X codes for the enzyme the reaction: chorismate $\rightarrow$ prephenate.
- Hypothesis 2: Gene Y codes for the enzyme the reaction: chorismate $\rightarrow$ prephenate.

- These can be tested by comparing the wild-type with strains
  - without Gene X / with and without prephenate.
  - without Gene Y / with and without prephenate.

ß


QuickTime™ and a
BMP decompressor
are needed to see this picture.
?

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:

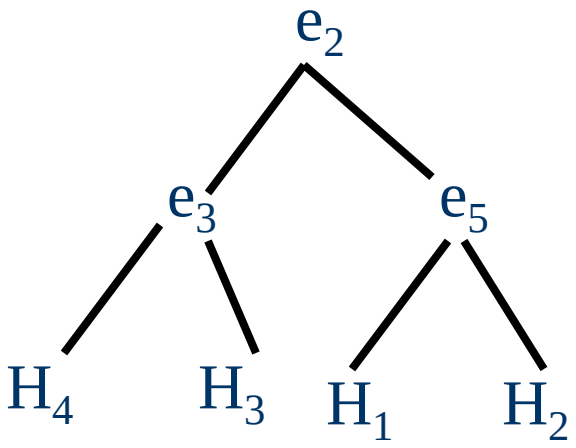
- To choose a series of experiments which minimise the expected cost of eliminating all but one hypothesis.

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.

How to choose the best experiment

	e_1	e_2	...	e_m
H_1	T	F	...	T
...	F	T	...	F
H_n	F	T	...	T



Choosing the best experiment is equivalent to choosing the best node in a decision tree.

Bryant et al. (2001) ETAI 5, 1-36.

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*:

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

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

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

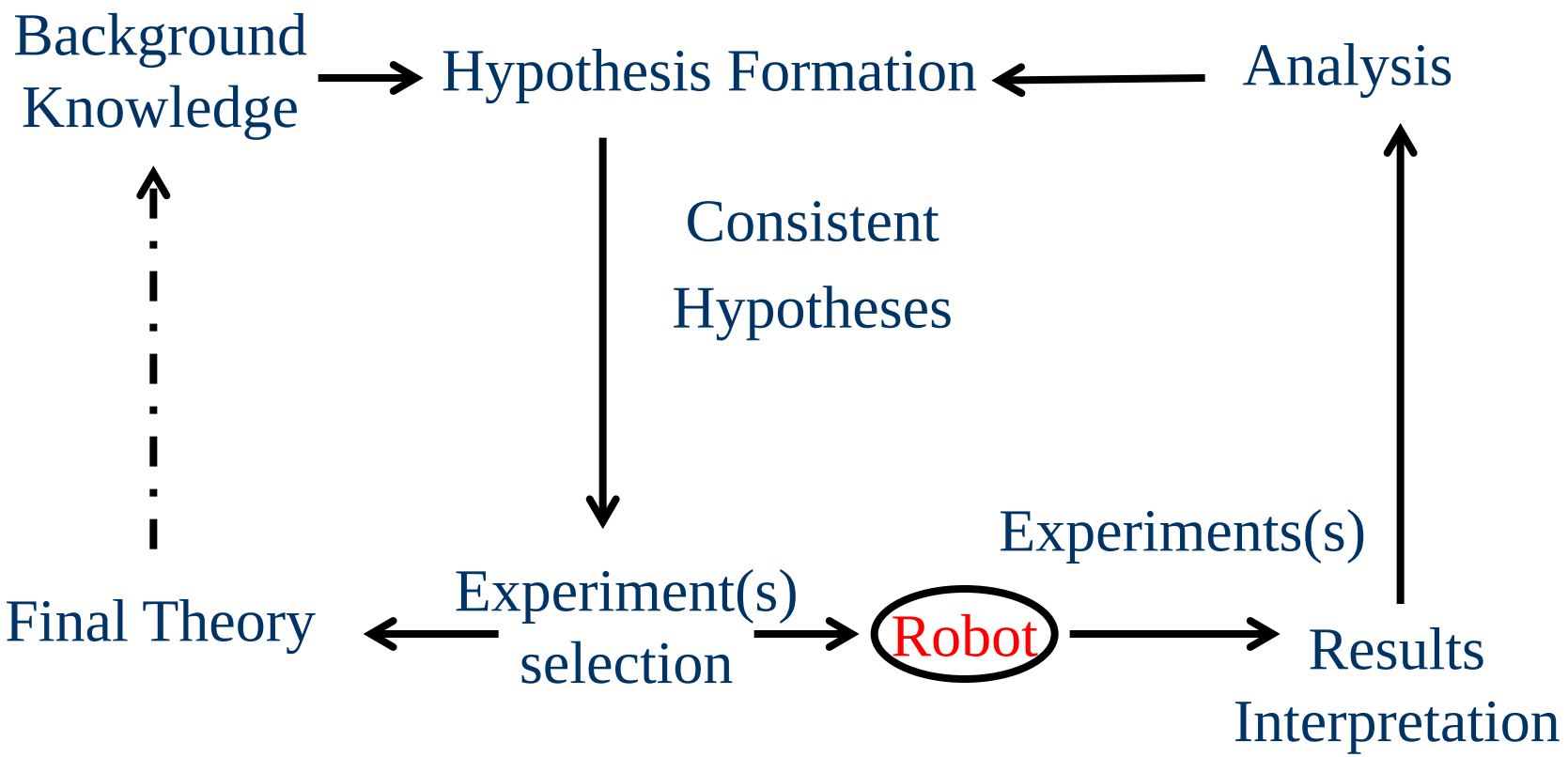
$$J_H = -\sum_{h \in H} p(h)\lfloor \log_2(p(h))\rfloor$$

C_t is the monetary price of the trial *t*

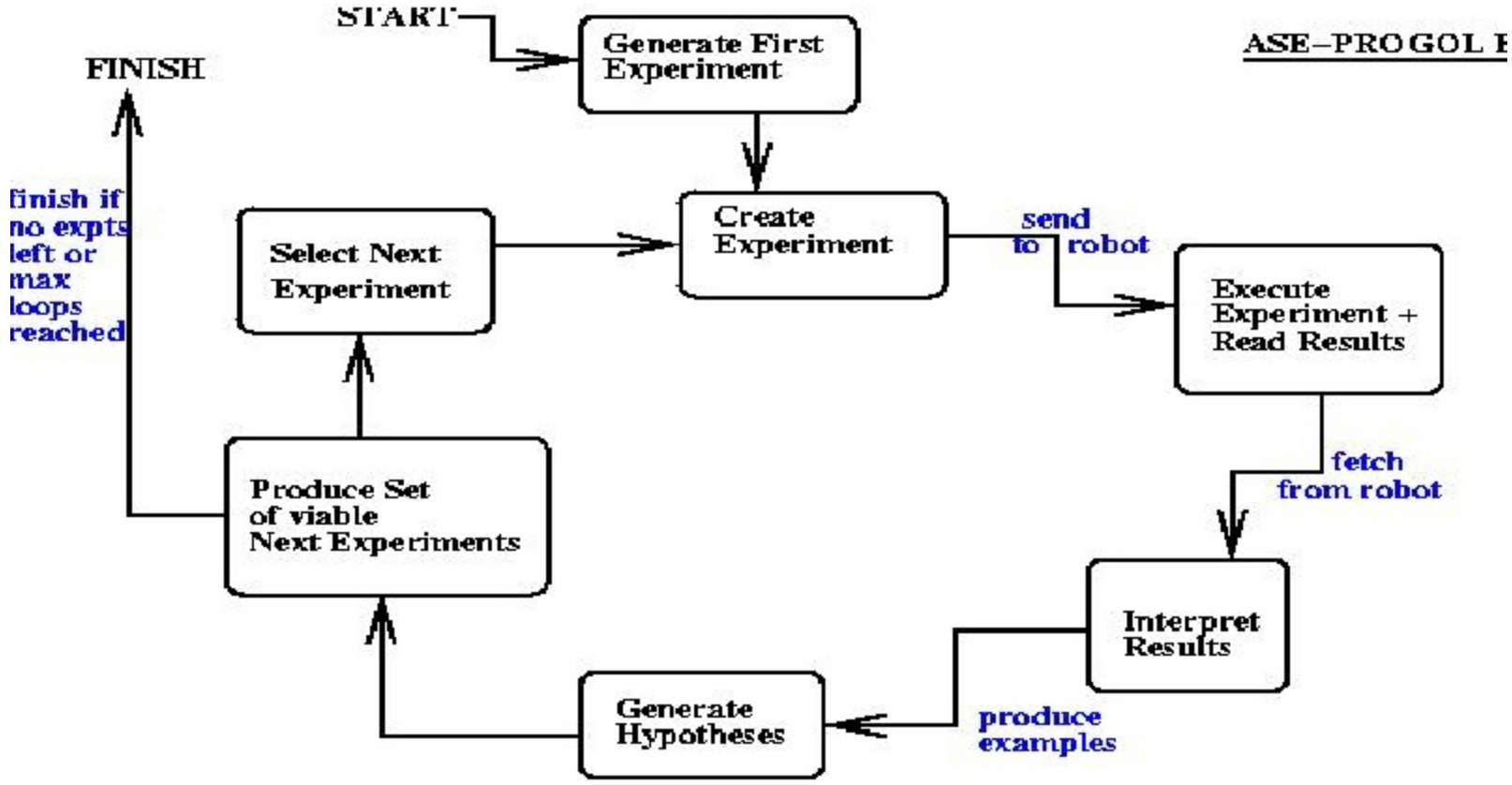
p(t) is the probability that the outcome of the trial *t* is positive

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

The Experimental Cycle



LIMS Setup


START
Generate First Experiment
ASE–PROGOL
FINISH
finish if no expts left or max loops reached
Select Next Experiment
Create Experiment
send to robot
Execute Experiment + Read Results
fetch from robot
Produce Set of viable Next Experiments
Interpret Results
Generate Hypotheses
produce examples

Adam

- Designed to fully automate yeast growth experiments.
- Has a -20C freezer, 3 incubators, 2 readers, 3 liquid handlers, 3 robotic arms, 2 robot tracks, a centrifuge, a washer, an environmental control system, etc.
- Is capable of initiating ~1,000 new experiments and >200,000 observations per day in a continuous cycle.

Plan of Adam

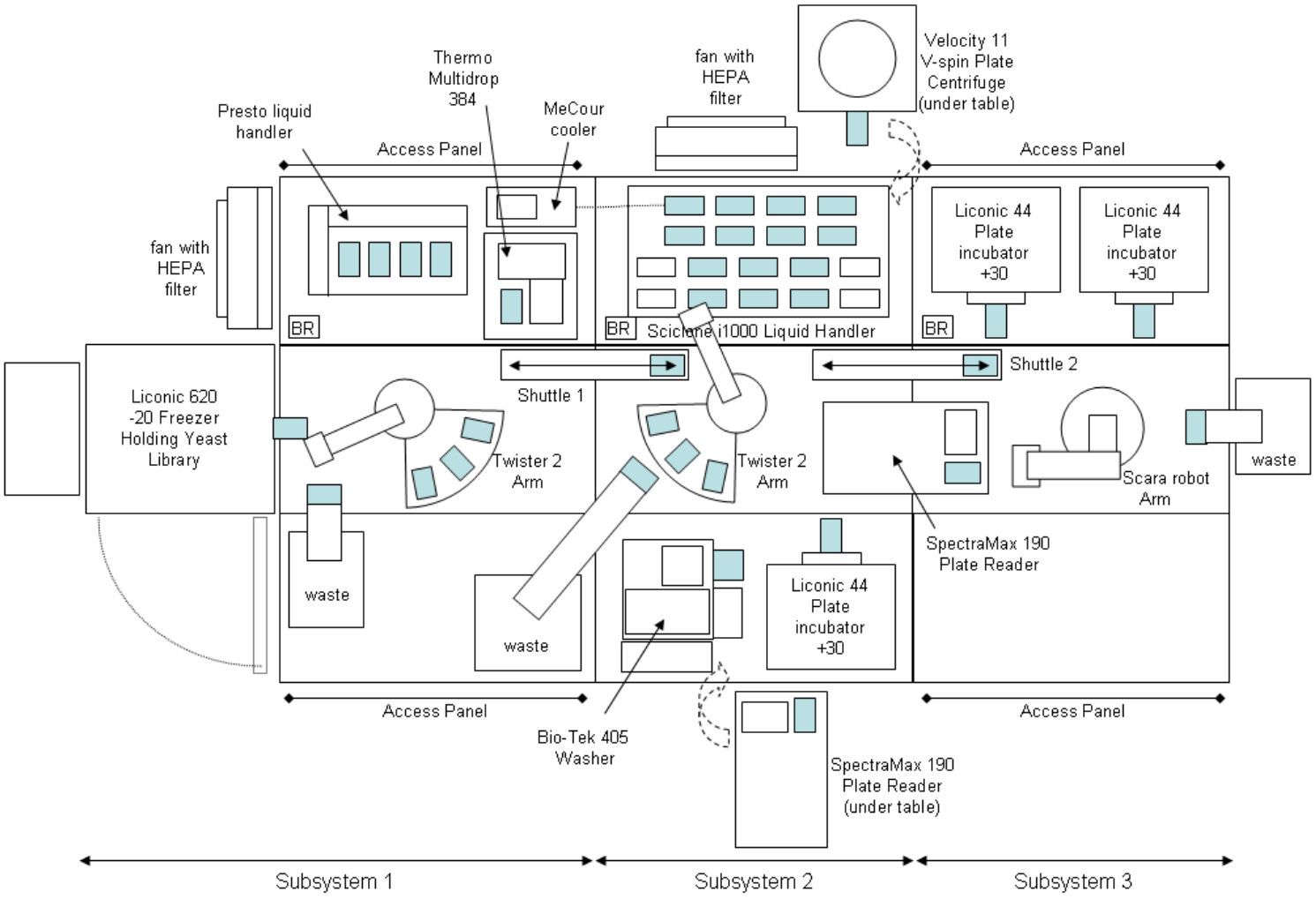
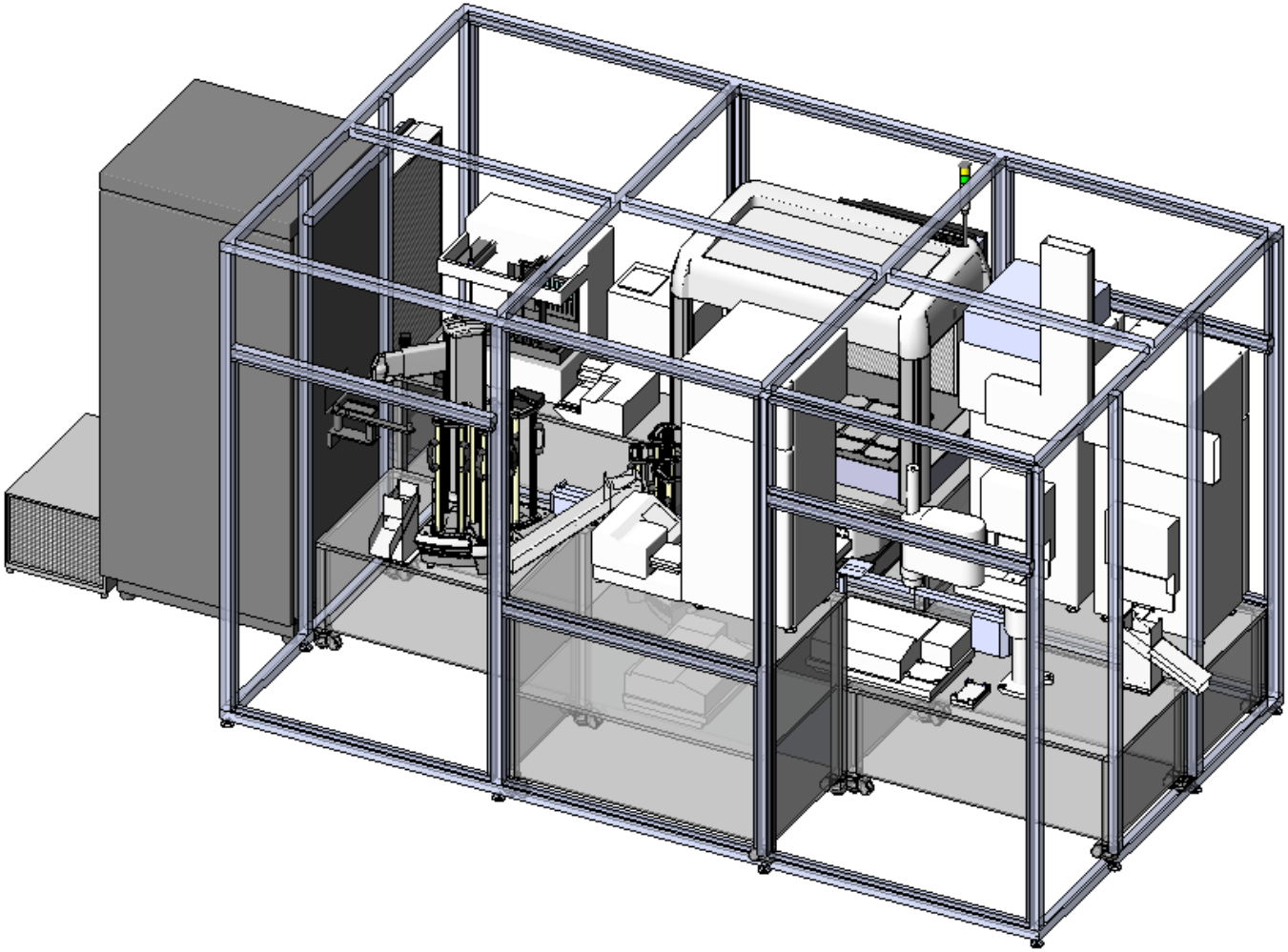
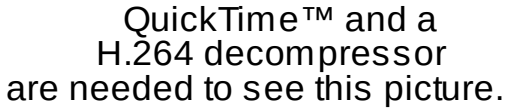

Velocity 11
V-spin Plate
Centrifuge
(under table)
Thermo
Multidrop
384
fan with
HEPA
filter
Presto liquid
handler
MeCour
cooler
Access Panel
Access Panel
fan with
HEPA
filter
Liconic 44
Plate
incubator
+30
Liconic 44
Plate
incubator
+30
BR
BR
Sciclone i1000 Liquid Handler
BR
Shuttle 2
Shuttle 1
Liconic 620
-20 Freezer
Holding Yeast
Library
Twister 2
Arm
Twister 2
Arm
Scara robot
Arm
waste
SpectraMax 190
Plate Reader
waste
waste
Liconic 44
Plate
incubator
+30
Access Panel
Access Panel
Bio-Tek 405
Washer
SpectraMax 190
Plate Reader
(under table)
Subsystem 1
Subsystem 2
Subsystem 3

Diagram of Adam

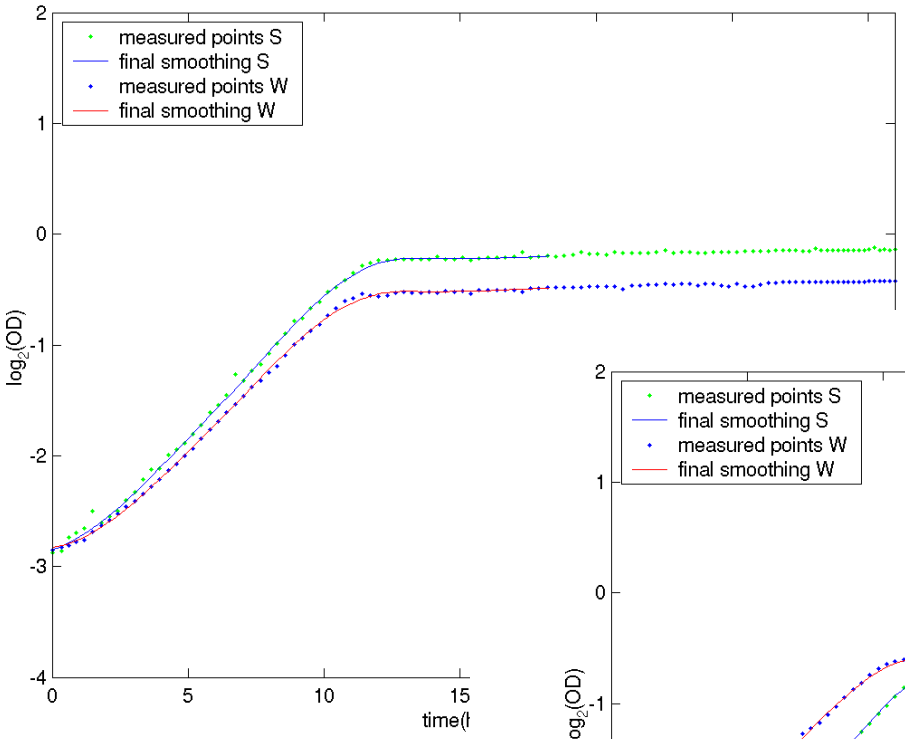


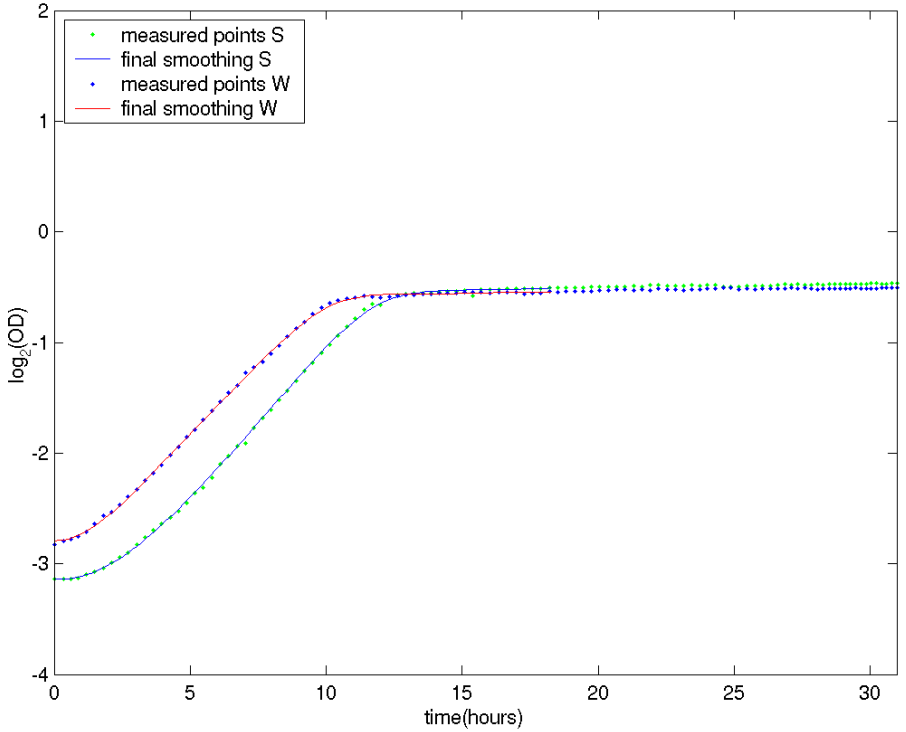
Adam During Commissioning

Adam in Action

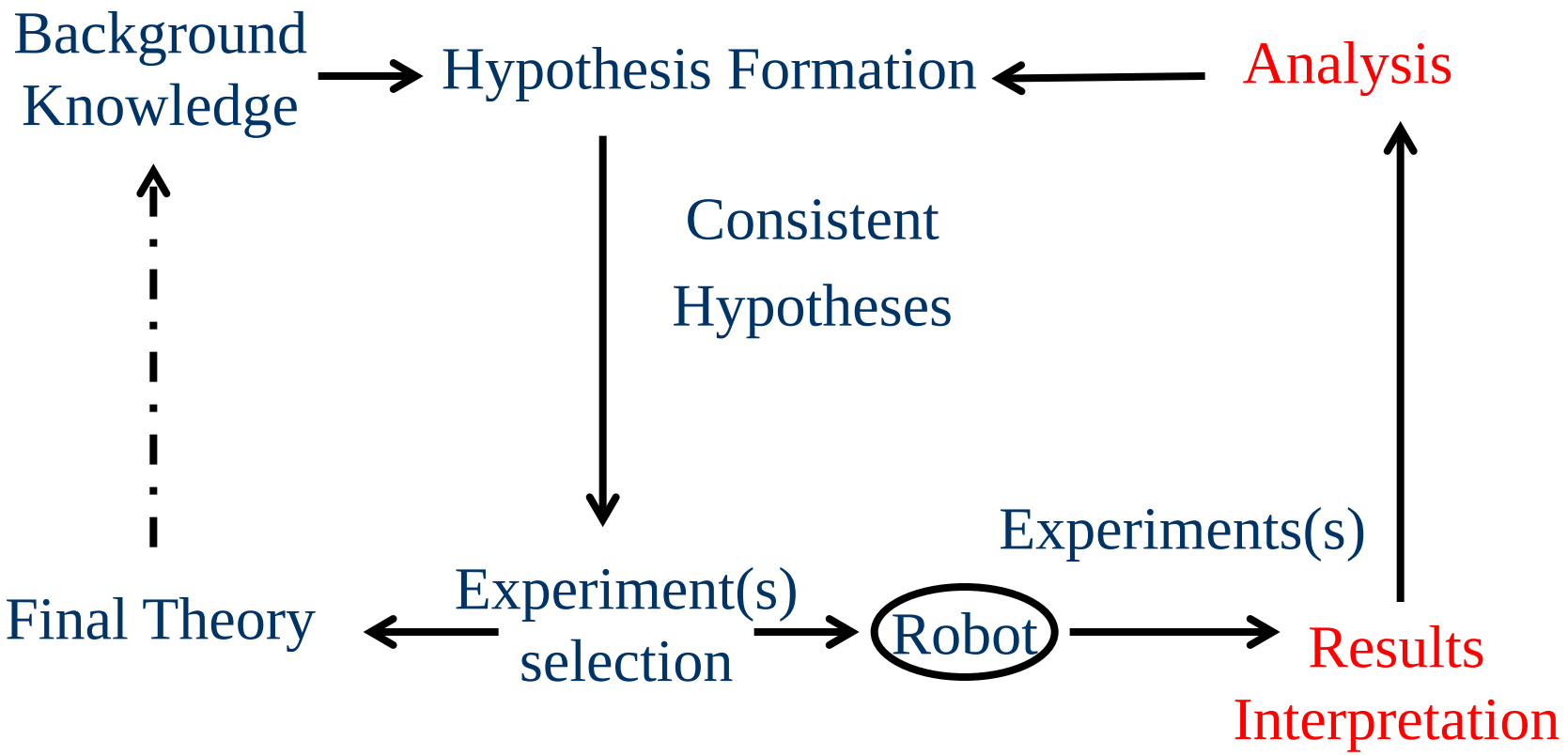

QuickTime™ and a
H.264 decompressor
are needed to see this picture.

Example Growth Curves





The Experimental Cycle



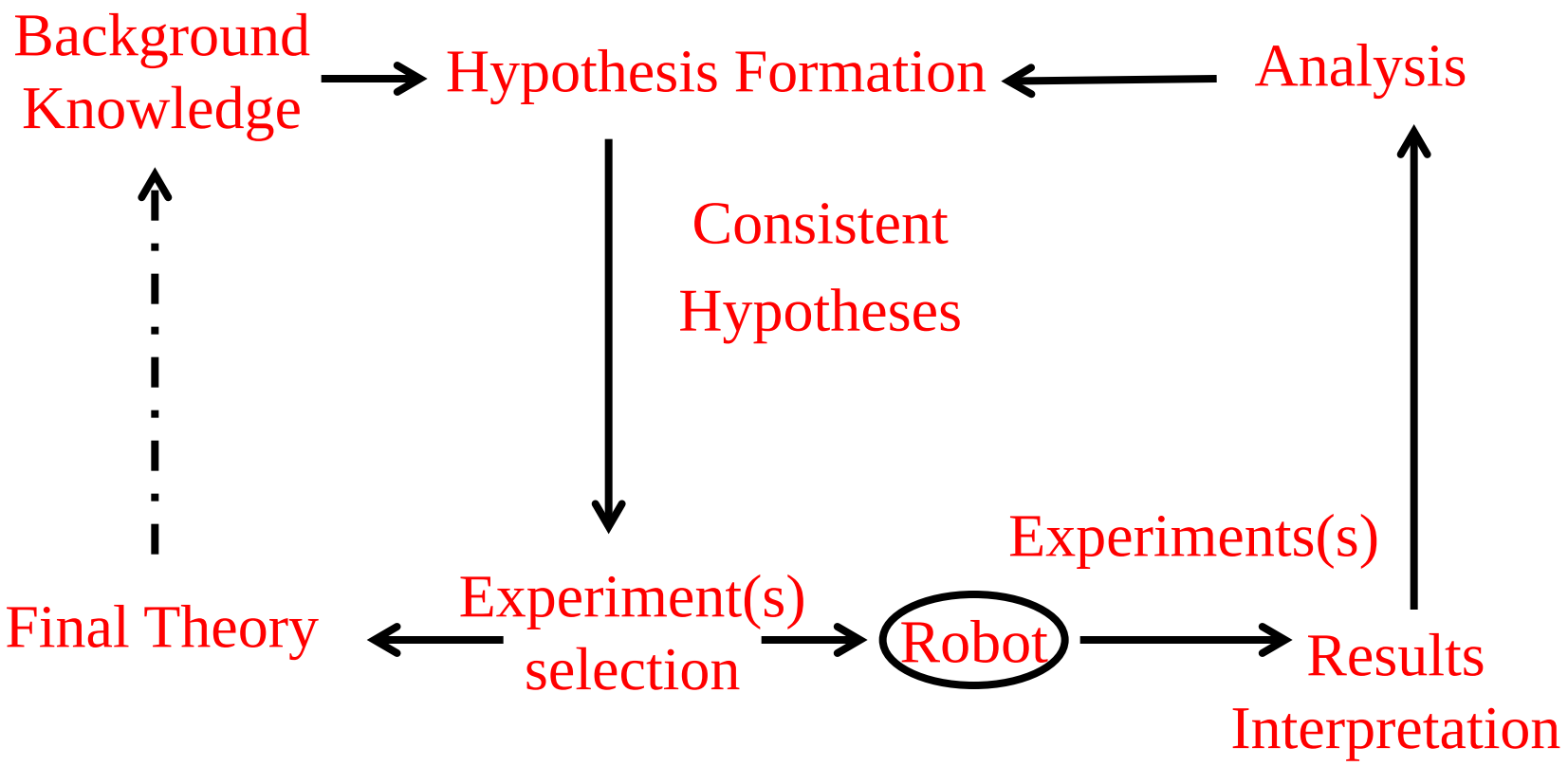
Qualitative to Quantitative

- The functions of most genes that when they are knocked out result in auxotrophy (no growth) have already been discovered.
- Most genes of unknown function only affect growth quantitatively.
- They may have slower growth (bradytrophs), faster growth, higher/lower biomass yield, etc..

Experimental Design

- Adam used a 2 factor design on each 96 well plate
  - Wild-type, Wild-type + metabolite
  - Knockout, Knockout + metabolite
  - 24 repeats using Latin square designs
- Look for a statistically significant difference in the response to the knockout to the metabolite.
- Use decision trees to discriminate between differences in growth curves.

The Experimental Cycle



Closing the Loop

- We have physically implemented all aspects of Adam.
- To the best of our knowledge Adam is the most advanced AI system that can both explicitly form hypotheses and experiments, and physically do the experiments.

Discovery of Novel Science

Novel Science

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

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

King *et al*. (2009) *Science*.

Novel Results

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	✓	-

A 50 Year Old Puzzle

- The enzyme 2-aminoadipate: 2-oxoglutarate aminotransferase is missing from our model.
- It is in the lysine biosynthesis pathway which has been studied for 50 years in fungi: target for antibiotics, and on path to penicillin.

- Adam formed three hypotheses for the gene to encode this enzyme: YER152C, YJL060W, and YGL202W (in that order of probability).

- Currently KEGG states that YGL202W is the gene.
- Evidence from 1960's that 2 iso-enzymes involved.

Confirmed New Knowledge

- Adam's differential growth experiments were consistent with all three genes encoding 2-oxoglutarate aminotransferase.
- Manual experiments: purified protein + enzyme assays, are consistent.
  - YGL202W literature confirmed.
  - YJL060W (was annotated as an arylformamidase, new (08) annotation kynurenine aminotransferase)
  - YER152C (currently not annotated)
- YGL202W & YJL060W double knockout is lethal

Systems Biology Prospects

- We are using Adam to develop a quantitative model of metabolism that maps genotype (list of deletion mutants) and defined growth medium (environment) to predicted quantitative growth.
- Combines ideas from logical and FBA modelling.
- Experiments with Adam are ongoing.

Eve

Eve

- First Drug Screening / Drug Design equipment in a Computer Science Department.
- Design Features:
  - During the screening process Eve will be able to decide to switch to QSAR mode.
  - Eve will use cycles of active learning to learn QSARs.
  - Use yeast assays to target 3rd World diseases.

Eve


QuickTime™ and a
BMP decompressor
are needed to see this picture.

Formalisation

Formalization of Science

- The goal of science is to increase our knowledge of the natural world through the performance of experiments.
- This knowledge should, ideally, be expressed in a *formal logical language*.
- Formal languages promote semantic clarity, which in turn supports the free exchange of scientific knowledge and simplifies scientific reasoning.

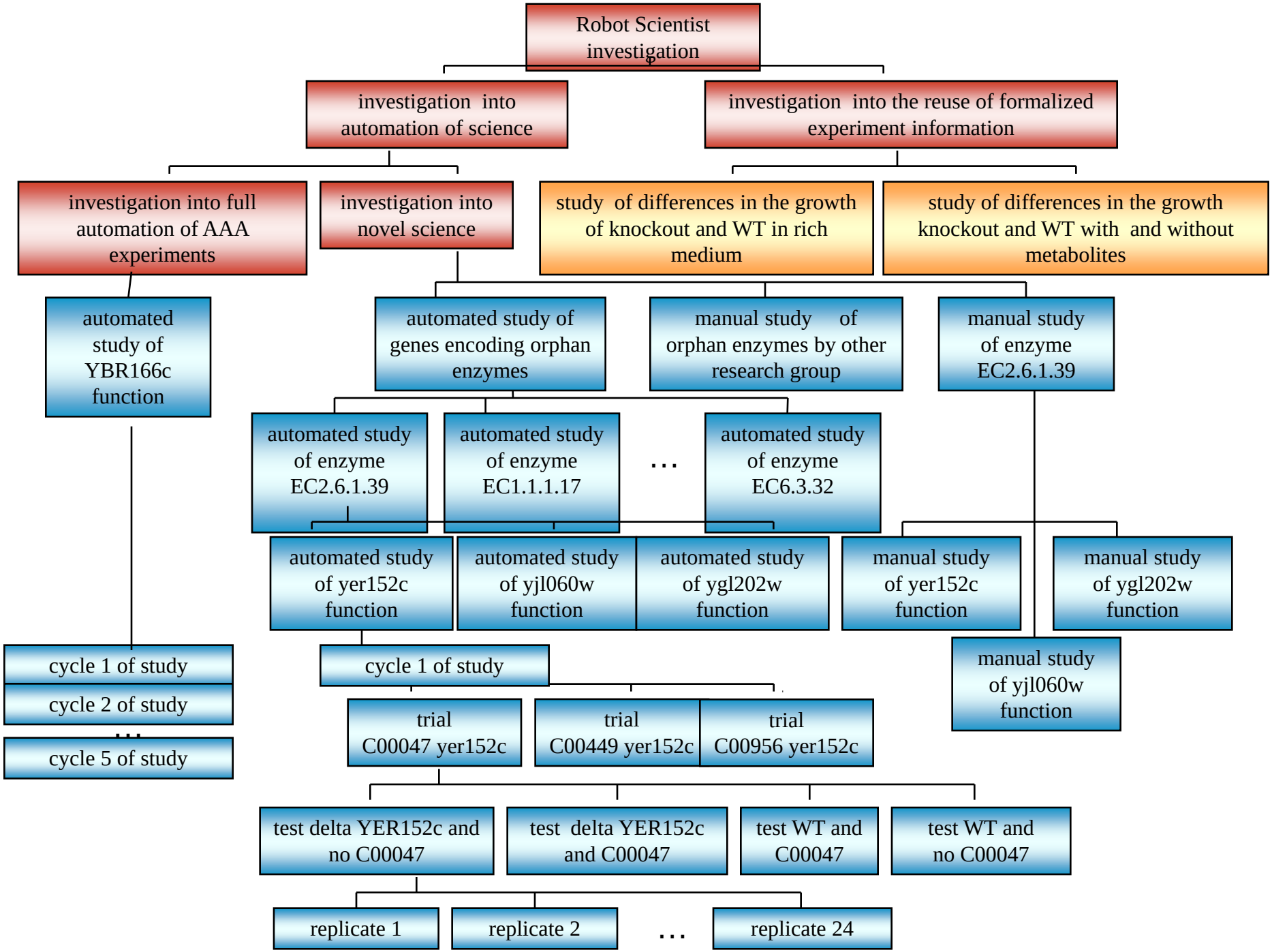
Robot Scientist & Formalisation

- Robot Scientists provide unsurpassed test-beds for the development of methodologies for the curation and annotation of scientific experiments.
- As the experiments are conceived and executed by computer it is possible to completely capture and digitally curate all aspects of the scientific process: hypotheses, experimental goals, results, conclusions, etc.
- The ontology LABORS is designed to enable the open access of the Robot Scientist experimental data and metadata to the scientific community.

Soldatova, Sparkes, Clare, & King (2006) *Bioinformatics*

The Formalisation of 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 automation of science
investigation into the reuse of formalized experiment information
investigation into full automation of AAA experiments
investigation into novel science
study of differences in the growth of knockout and WT in rich medium
study of differences in the growth knockout and WT with and without metabolites
automated study of YBR166c function
automated study of genes encoding orphan enzymes
manual study of orphan enzymes by other research group
manual study of enzyme EC2.6.1.39
automated study of enzyme EC2.6.1.39
automated study of enzyme EC1.1.1.17
...
automated study of enzyme EC6.3.32
automated study of yer152c function
automated study of yjl060w function
automated study of ygl202w function
manual study of yer152c function
manual study of ygl202w function
manual study of yjl060w function
cycle 1 of study
cycle 2 of study
...
cycle 5 of study
cycle 1 of study
trial C00047 yer152c
trial C00449 yer152c
trial C00956 yer152c
test delta YER152c and no C00047
test delta YER152c and C00047
test WT and C00047
test WT and no 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

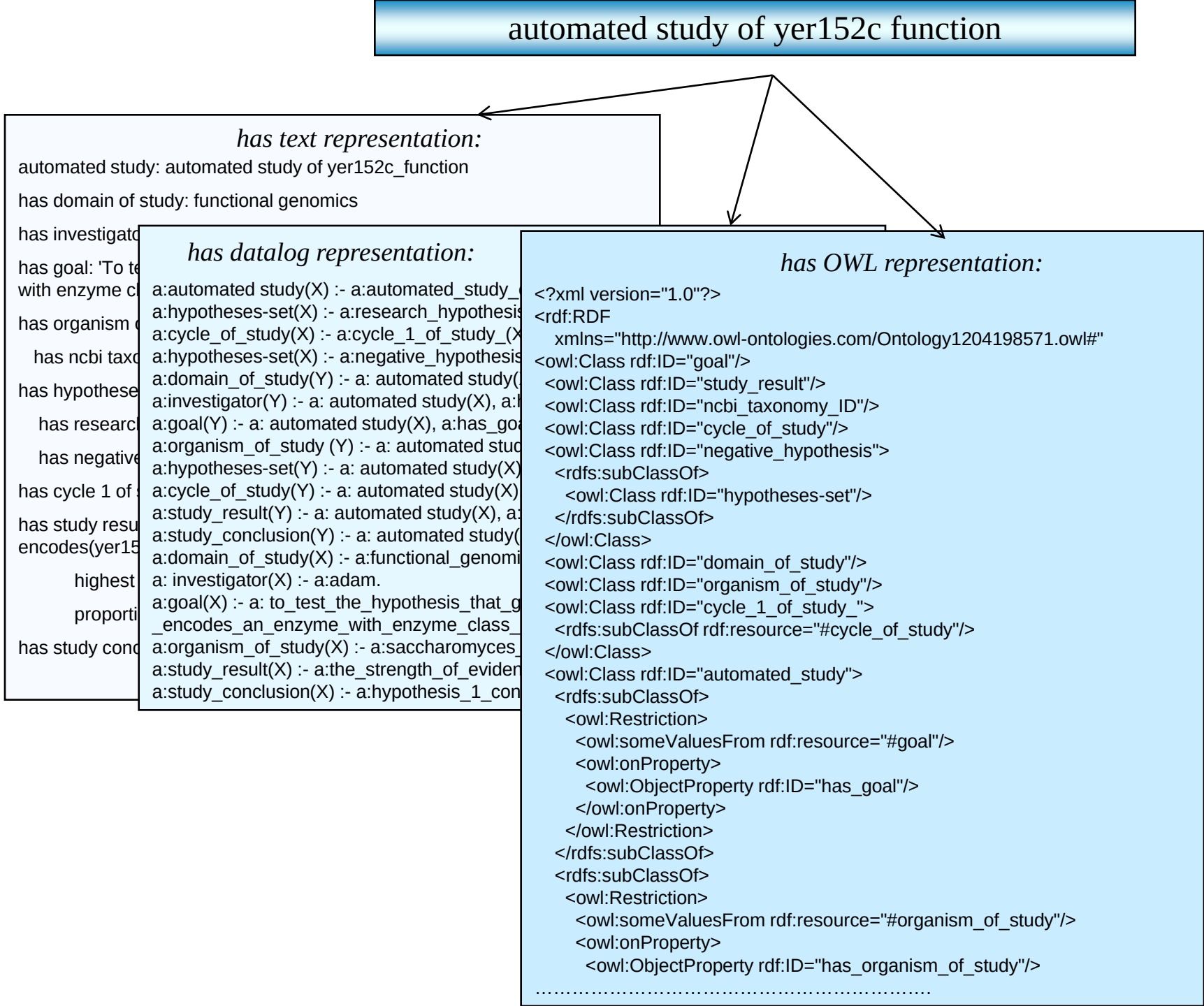
YER152C and Lysine automated trial

Experiment 1 (wild-type no metabolite)

Replicate 1 (well)

Observation 1

b)


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
has ncbi taxo
has hypothese
has researc
has negative
has cycle 1 of
has study resu
encodes(yer15
highest
proporti
has study cond
has datalog representation:
a:automated study(X) :- a:automated_study_
a:hypotheses-set(X) :- a:research_hypothesi
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(
a:investigator(Y) :- a: automated study(X), a:
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"/>
……

Conclusions

- Automation was the driving force of much of 19th and 20th century change, and this is likely to continue.
- Automation is becoming increasingly important in scientific research e.g. DNA sequencing, drug design
- The Robot Scientist concept represents the logical next step in scientific automation.
- We have physically built a proof-of-principle Robot Scientist, Adam, for application to functional genomics.

- Adam has used automated techniques to generate novel scientific knowledge.

Acknowledgments

Amanda Clare

Jem Rowland

Mike Young

Ken Whelan

Larisa Soldatova

Maria Liakata

Andrew Sparkes

Wayne Aubrey

Magda Markham

Steve Oliver

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
- 2004-2008 Adam Project
  - Sophisticated Laboratory Automation
  - Collaboration with Steve Oliver (Cambridge).
  - King *et al*. (2009) *Science* (in press)
- 2008-2011 Eve Project