



THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE ■

University of
Hertfordshire **UH**

Computational Scientific Discovery in Social Sciences

AI-2023

Forty-third SGAI International Conference on Artificial
Intelligence

Cambridge, Tuesday December 12



Session 1

(11.00-12.30)

- Fernand Gobet (40 minutes)
 - Introduction to Computational Scientific Discovery in (Social) Sciences
 - GEMS
- Laura Bartlett (20 minutes)
 - Using GEMS to develop theories in Psychology (1)
 - Posner task
- Noman Javed and Dmitry Bennett (30 minutes)
 - Using GEMS to develop theories in Psychology (2)
 - Verbal learning
- Lunch break

Session 2

(13.15-14.45)

- Peter Lane
 - Tutorial of the GEMS system, covering:
 - *Setting up task definitions for scientific experiments*
 - *Defining a search space of candidate models*
 - *Searching techniques, such as Genetic Programming*
 - *Visualization and analysis of results*

Introduction to Computational Scientific Discovery in (Social) Sciences

Fernand Gobet



THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE ■



European Research Council

Funding for top researchers
from anywhere in the world



Computational Scientific Discovery (CSD)

- Scientific Discovery is the process by which scientists create or find new knowledge
 - A new class of objects
 - A new class of celestial objects in astronomy
 - A new species in biology
 - New taxonomy
 - Linnaeus' systematic categorization of plants and animals
 - An empirical law
 - Kepler's law of planetary motion
 - An explanatory theory
 - Newton's theory of gravity
- Computational scientific discovery aims to automate certain aspects of this process



Very Brief History of Computational Scientific Discovery

- Not a new field!
- Dates back to the 1970s
- Scientific discovery can be seen as a form of heuristic search through a problem space
 - Simon (1966)
- Originally the idea of CSD met with great resistance
- Relatively few researchers until a couple of years ago
- Now, explosion of interest with the new waves of AI and machine learning



Two Main Traditions

1. Discrete Structures

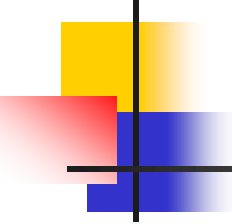
- Early research aimed to find laws or models expressed in terms of standard scientific formalisms
- Good example: Langley et al. (1987)
- Uses discrete symbolic expression
 - Graphs
 - Mathematical equations
 - Computer code
- Researchers address various discovery tasks
 - Induction of mathematical laws from data
 - Qualitative models that explaining phenomena using hidden structures or processes
 - Replicating historical discoveries



Two Main Traditions

2. Continuous Structures

- ~ last 10 years
- Interest in computational discovery in physics, applied mathematics, medicine, etc.
- Focuses on continuous mathematics
- Carries out search through a *parameter space*
- Relies on neural networks and continuous optimization
- A 2023, an AAAI symposium brought together these two traditions



Recent Examples from Natural Sciences

- DeepMind AlphaFold
- Understanding protein structures is crucial for insights into biological and disease mechanisms
- Traditionally, determining protein structures has been a lengthy and complex process
- AlphaFold predicts 3D protein structures with remarkable accuracy, using deep learning
- Solution to a 50-year-old grand challenge in biology
- Huge implications for drug discovery, disease understanding, and biotechnology



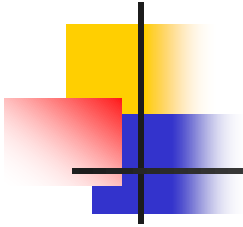
Recent Examples from Natural Sciences

- AI-Driven Personalized Cancer Treatment
- AI algorithms process complex genomic data to identify specific mutations in cancer cells
 - Helps in distinguishing cancer characteristics unique to each patient
 - Treatments that are more effective and have fewer side effects
- Potential for AI to transform cancer care



Example from Social Sciences

- To develop and refine theories of decision-making
- Peterson et al. (2021) collected a large data set on risky-choice decisions
 - Participants choose between different gambles
- Human-generated models of trained on the data
 - Subjective utility models
 - Prospect theory
 - Explained the data relatively poorly
- Machine learning discovered new, better theories
 - Strategies that are a mixture of previously proposed theories



The GEMS Approach

Genetically Evolving Models in Science



The Original Discovery Problem

- To develop process models in cognitive psychology
 - Explaining how people perform in standard *experiments*
 - How they memorise a list of items
 - How they select a previously shown item out of two items
 - etc.

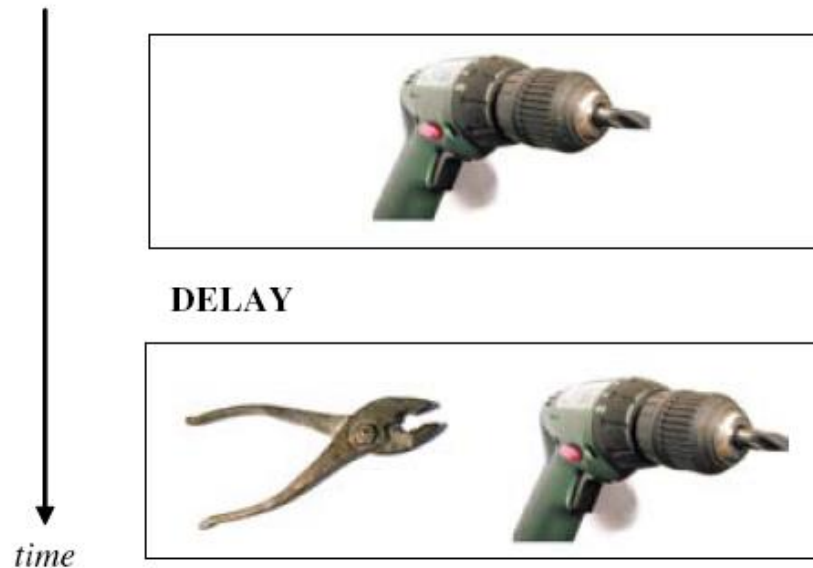


Example: Delayed Match to Sample (DMTS) Task





Example: Delayed Match to Sample (DMTS) Task



- Typical empirical data collected:
 - Percentage correct
 - Mean response time

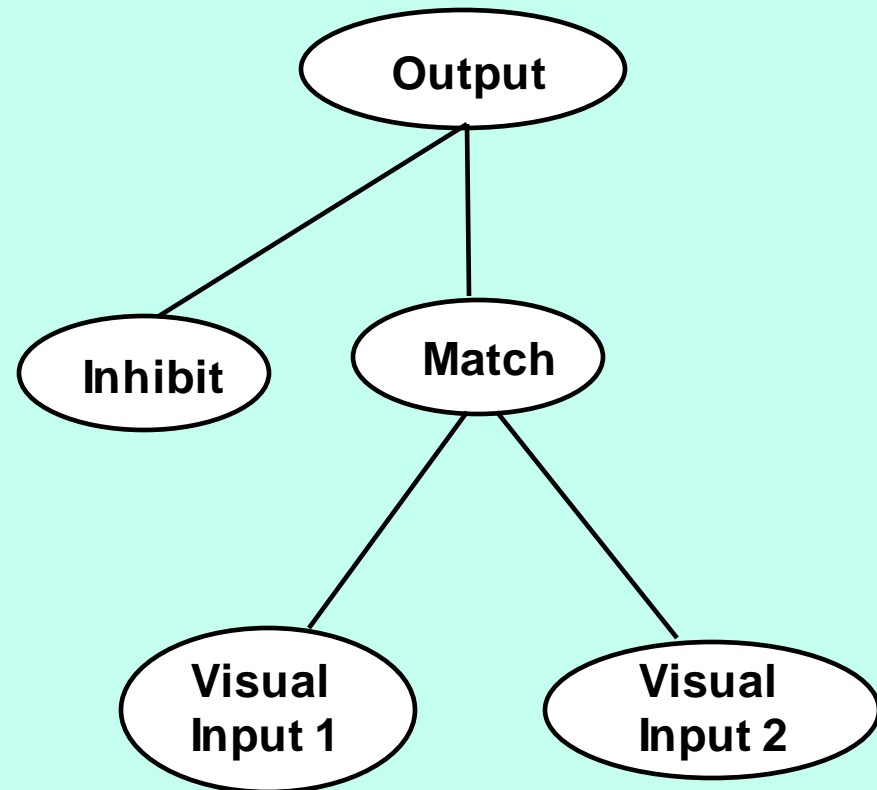


How Is the Problem Formulated Computationally?

- Heuristic search through a problem space of discrete structures (models/programs)
- The programs carry out the same experiment(s) as humans
- The programs
 - Embedded in a high-level, domain-specific cognitive architecture
 - Are interpreted by a virtual machine
- The search is carried out using genetic programming

Scientific Theories as Computer Programs

- Scientific theories can be represented as computer programs
- These theories can be represented as trees
- These programs (trees) can be evolved





Cognitive Architecture

- Specifies
 - The common, non-changeable structures of the models
 - Whether items are subject to activation and decay
- Very simple at the moment
 - Attention slot
 - Short-term memory
 - Long-term memory

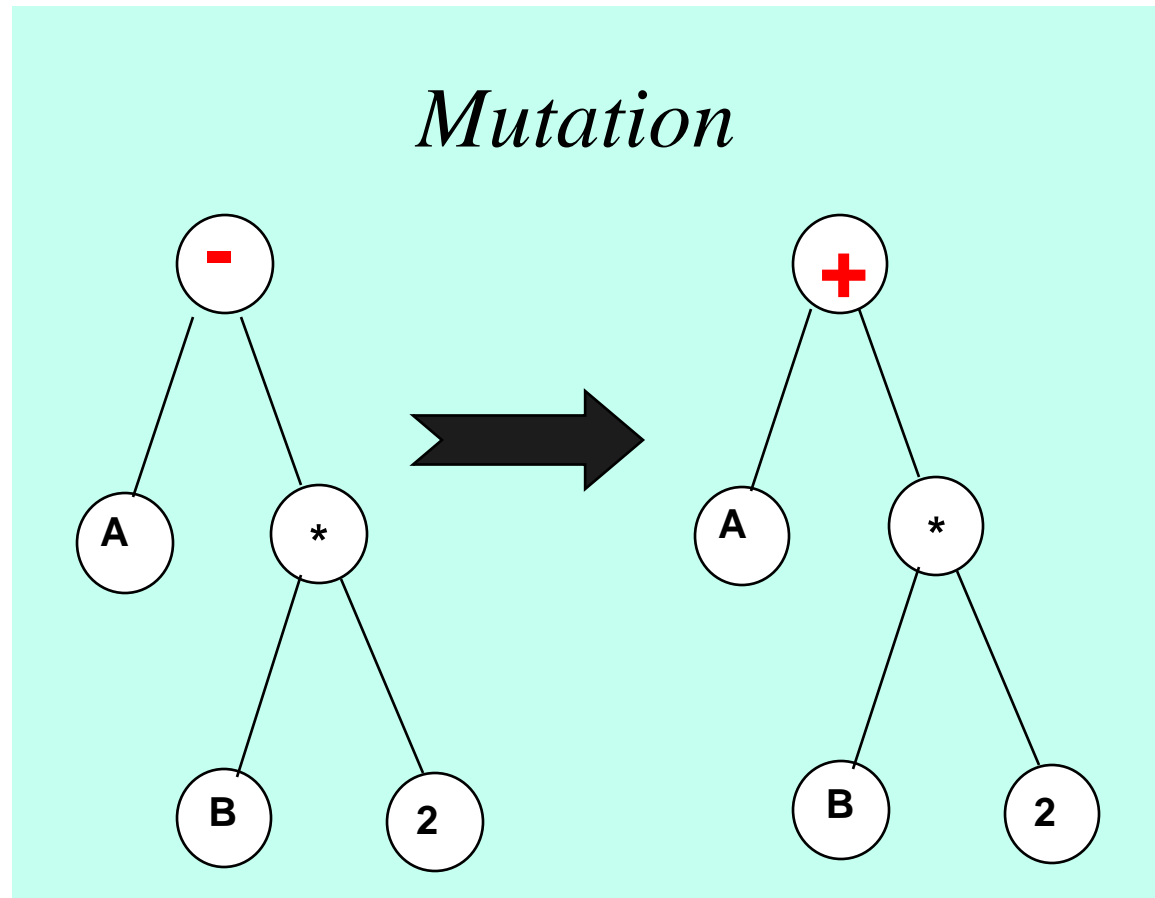


How Is the Problem Formulated Computationally?

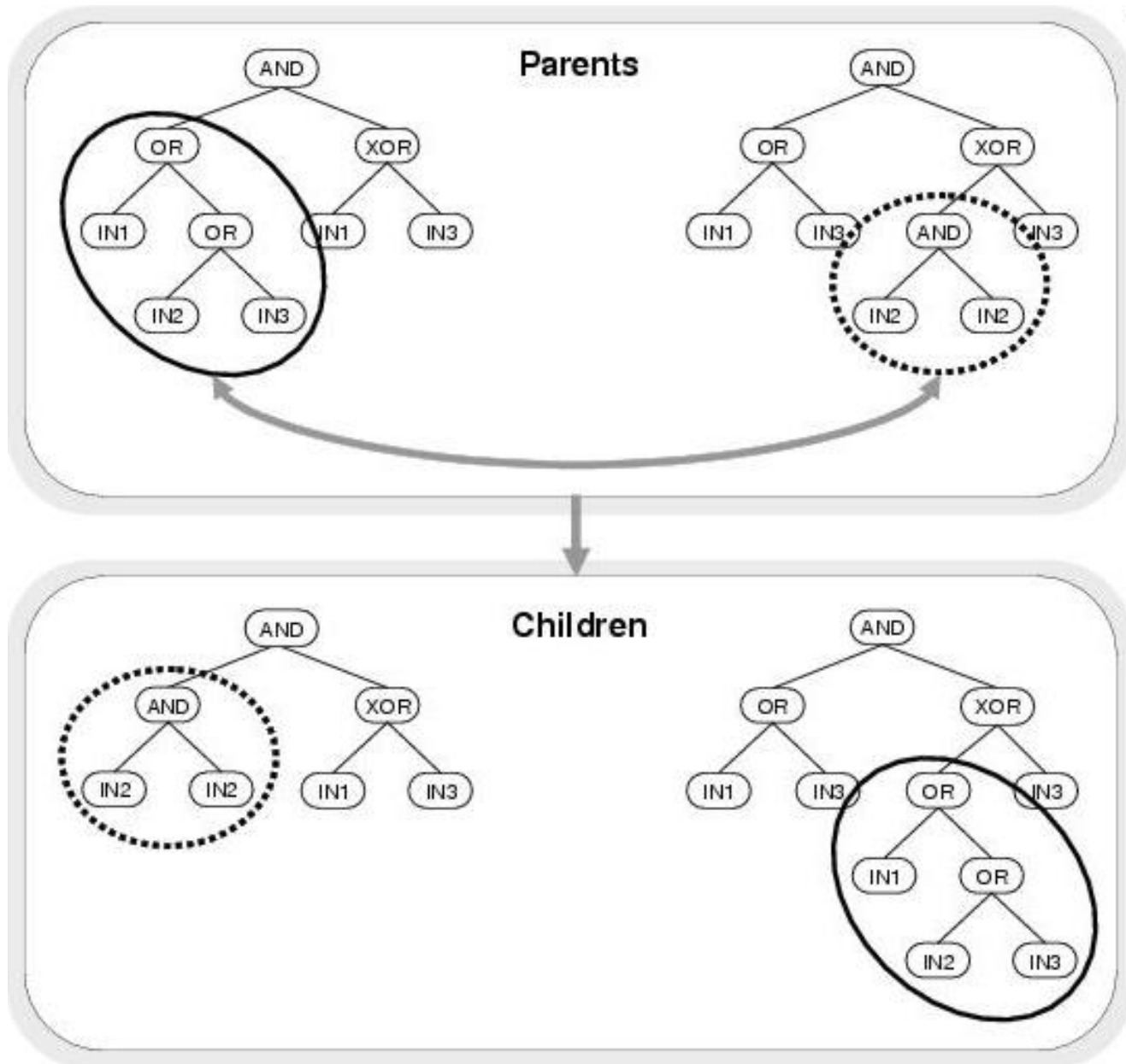
- Heuristic search through a problem space of discrete structures (programs)
- The structures/models are programs
 - for a high-level, domain-specific cognitive architecture
 - interpreted by a virtual machine
- The search is carried out using genetic programming

Genetic Programming (GP)

- Breeds and evolves entire computer programs
- Three main mechanisms
 - Selection
 - Mutation
 - Crossover



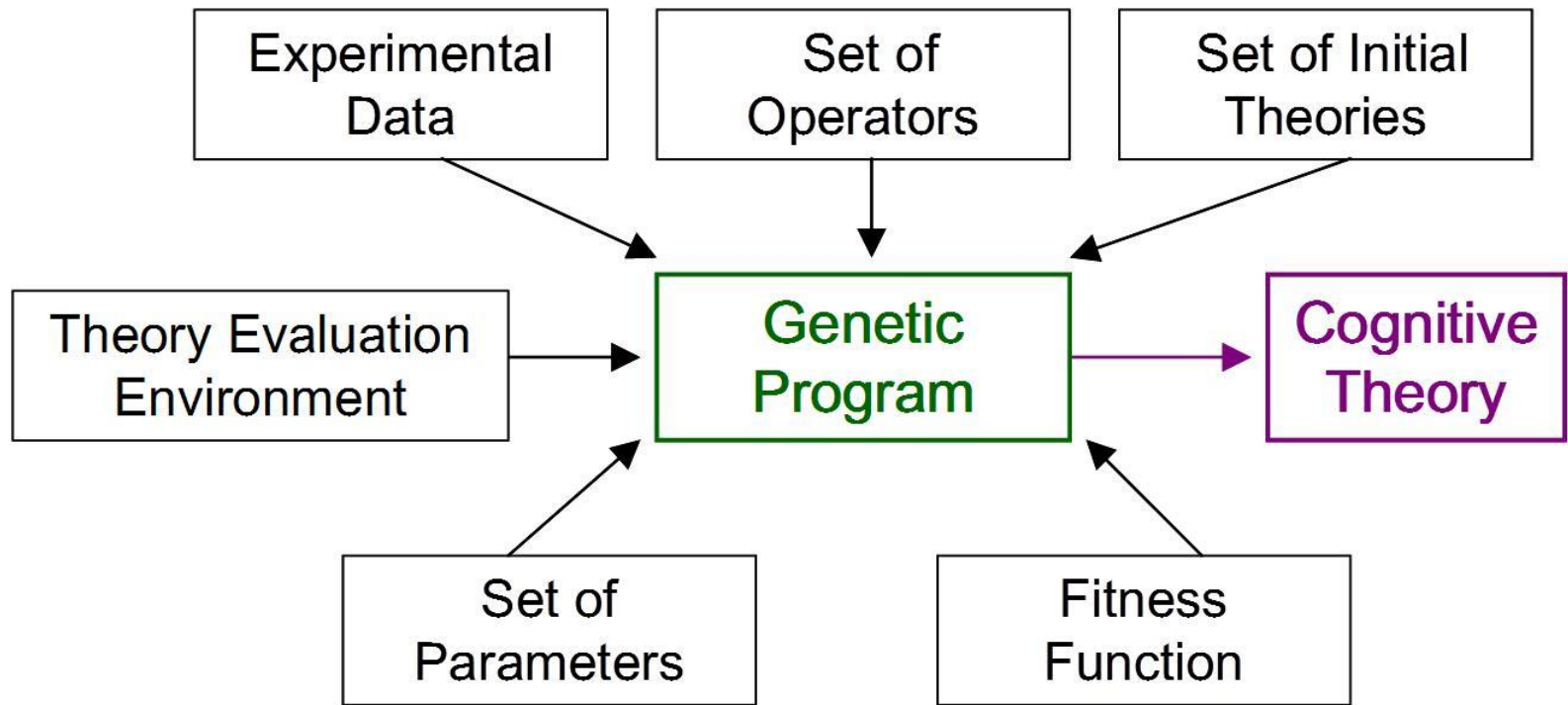
Crossover





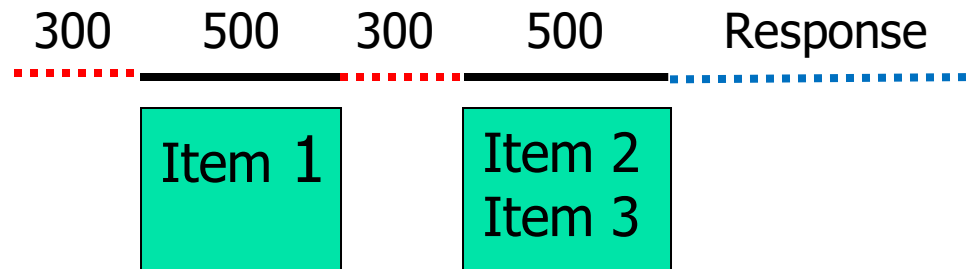
The Key Idea

1. A population of programs/models is (randomly) generated using basic operators
2. The predictions of the models in a specific task are compared with the actual empirical data
3. The fitness value of each model is computed using step 2
4. The best models are selected for producing the next generation, together with crossover
5. Steps 2 – 4 are repeated until stopping condition is satisfied



What Data and Knowledge are Provided to the System?

1. Description of the experimental methodology, including
 - The independent and dependent variables
 - The stimuli
 - The timeline of the experiment





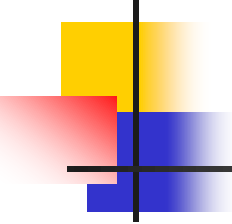
Data and Knowledge (II)

2. The results obtained in the corresponding human experiments, for each of the experimental conditions
 - percentage correct
 - response times
 - type of errors
 - etc.



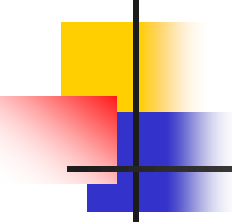
Data and Knowledge (III)

3. Description of the *architecture*
4. Description of the *operators* to be used
 - Specify basic cognitive operations
 - e.g. “put item into short-term memory”
 - Include a time cost
- Specification of the architecture and the operators is based on the literature
 - Many options are possible for both



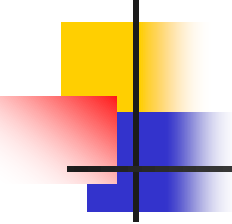
How Are the System's Inputs Represented?

- Inputs related to the experiment
 - The experimental methodology and the timeline consist of Lisp code
 - The stimuli are symbols (typically numbers or letters)
 - The experimental results are vectors of real numbers



How Are the System's Inputs Represented? (II)

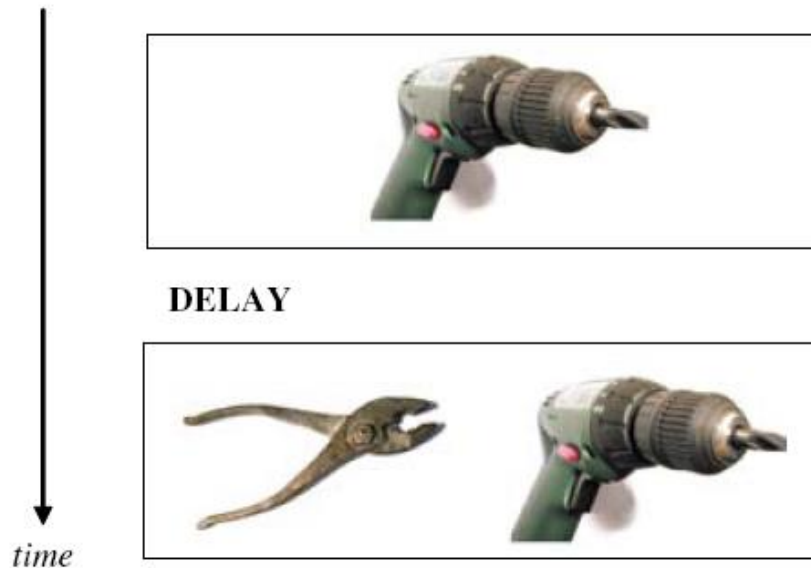
- Inputs related to the models
 - The *architecture* is specified by a virtual machine with *operators* supporting a simple interpreted language



How Are the System's Outputs Represented?

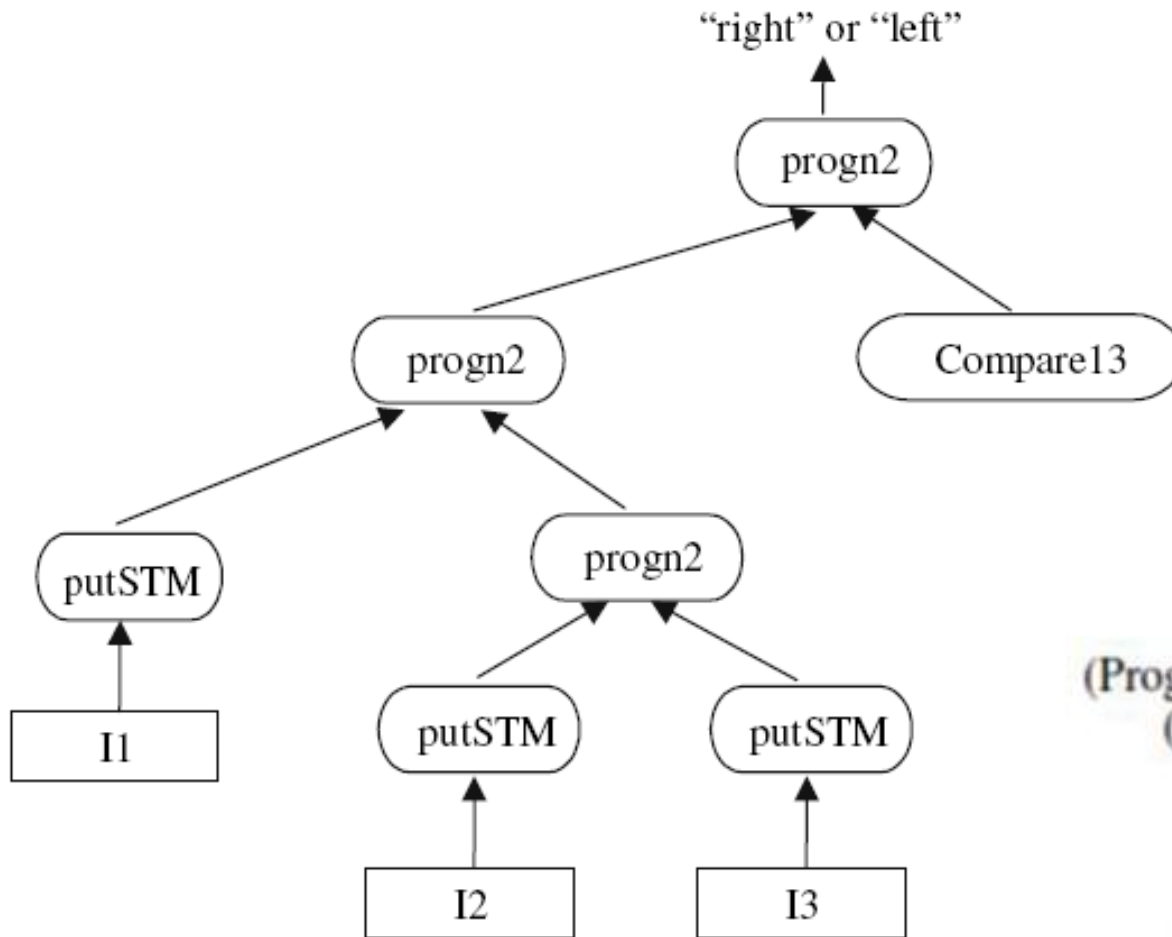
- The *outputs* are models with measures of goodness of fit

Example: Delayed Match to Sample (DMTS) Task



- Empirical data
 - Percentage correct in various conditions

Example of Generated Theory



```
(Progn2  
  (Progn2  
    (PutStm Input1)  
    (Progn2 (PutStm Input2)  
            (PutStm Input3)))  
  (Compare13))
```



The Space of Candidate Models that the System Searches

- All potential programs that can be generated from the operators
 - In principle, an infinite space
 - In practice, the space is much smaller due to constraints imposed by the operators' time cost



What Criteria are Used to Evaluate the Candidate Models?

- Fitness function
 - Computes the match between the predictions of a model and the human data
 - Several measures (e.g. percentage correct, response time) can be used
 - They can be given different weights
 - Criteria such as parsimony
 - e.g., size of the program/model

How are the Results Generated by the System Interpreted?



- The models consist of operators for a high-level, symbolic cognitive architecture
- Thus, they are easily interpretable
 - Abstract-syntax trees
 - Pseudo-code



Interpretation of Results (II)

- One important problem, typical of genetic programming
 - Bloating
 - The models can be long and complicated
- Actions taken
 - Methods for simplifying models
 - Methods for reducing the number of similar models



Interpretation of Results (III)

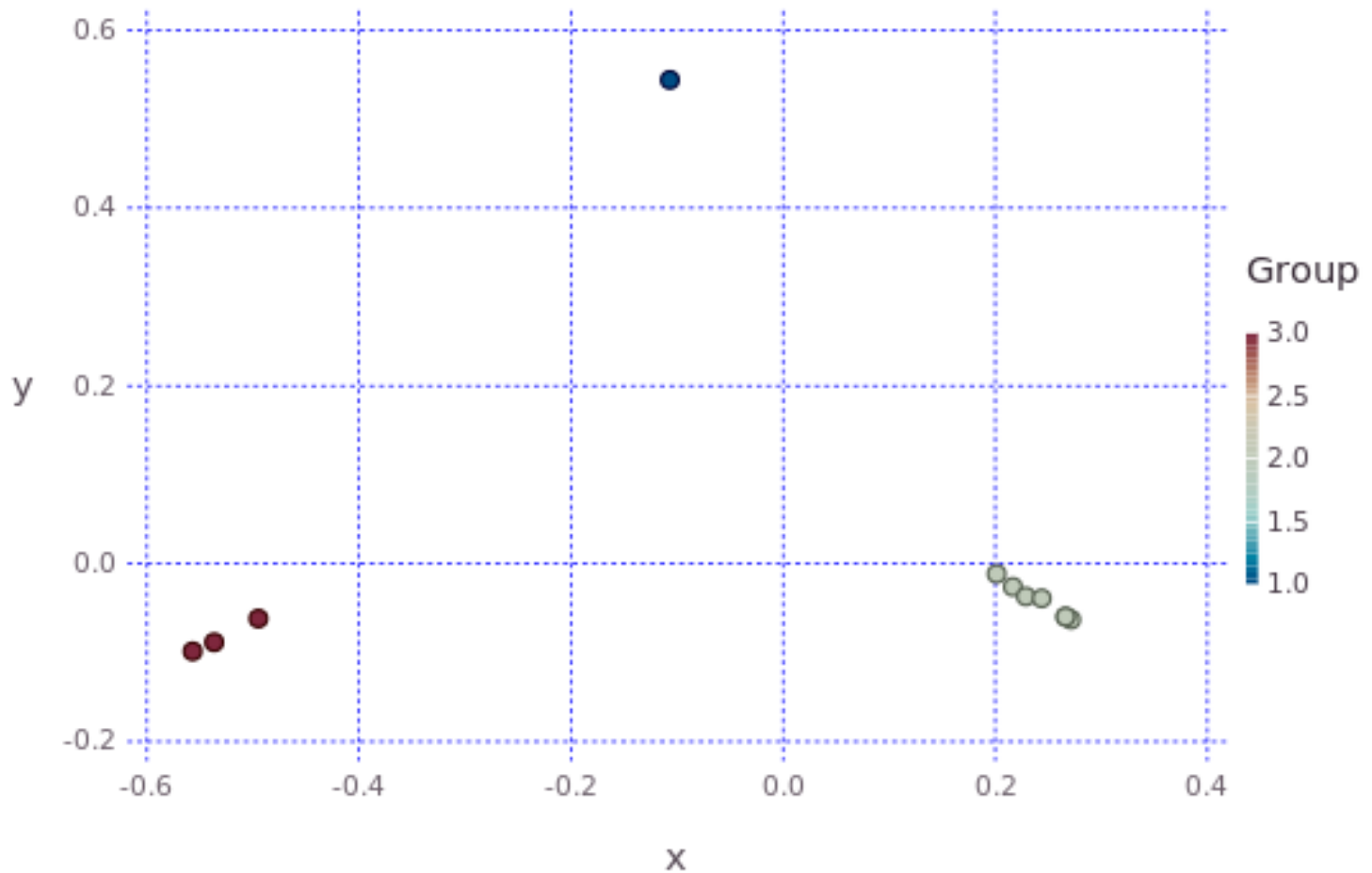
- Models can be visualised as clusters based on syntactic similarity
 - These clusters represent semantically different solutions to the task



Example

- DMTS task with 6 runs, 500 individuals and 2,000 generations
 - 1,164 “good” models
 - fitness < 0.1
- Post-processing techniques
 - Remove dead-code → 248 distinct models
 - Remove time-only operators → 11 distinct models

Example of Model Clustering





Advantages of the Methodology

- Increases likelihood of finding theories to account for the empirical data
- Produces theories that meet the criterion of sufficiency
 - The theories can indeed carry out the tasks under study



Potential Objections

- Size of the search spaces is very large
- Bloating with genetic programming
 - Trees can become large, and parts of them may be redundant
- Computational cost of measuring theories' goodness of fit is high
- The operators may be the wrong ones
- Empirical data might not be reliable



Using GEMS to Develop Theories in Psychology

- Laura Bartlett
 - Posner task
- Noman Javed and Dmitry Bennett
 - Verbal learning