science of intelligence

The Measure of Intelligence

François Chollet

Psychology + AI

currently:
skill at tasks

Created by Saeful Muslim
from Noun Project

Intelligence Benchmark

- definition?
- no general set of tests
  - e.g. Turing: subjective

- driver of progress: measurable, quantifiable, objective

“Intelligence”: skill-acquisition efficiency

scope, generalization difficulty, priors, experience

Intelligence

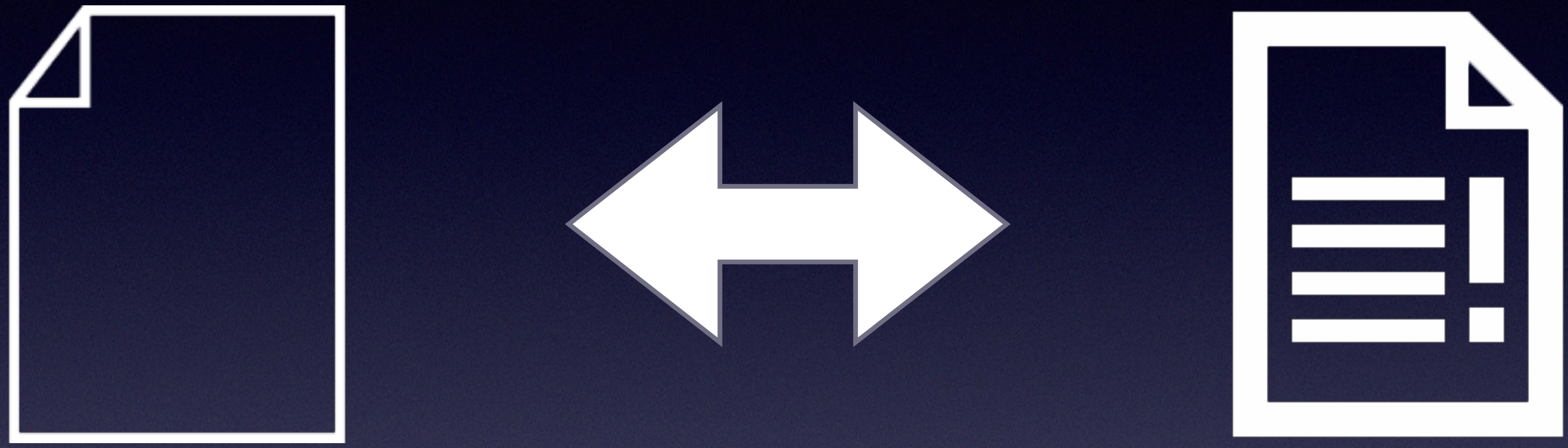
- *“Intelligence measures an agent’s ability to achieve goals in a wide range of environments”* - Legg and Hutter

- Crystallized and fluid intelligence - Catell

knowledge,
acquired skill

ability to
acquire new skills

Artificial Intelligence



blank slate
tell *how* to acquire skill itself

connectionism

cognitive psychology

hard-coded
set of rules

Symbolic AI

evolutionary psychology:
cognition result of adaptation

Narrow tasks as Benchmarks

AI Effect

Why? Non-human/short-cut skill acquisition

Goal: Generality, robustness, flexibility
<u>not task-specific performance</u>

Generalization

System-centric

situations
unknown to system

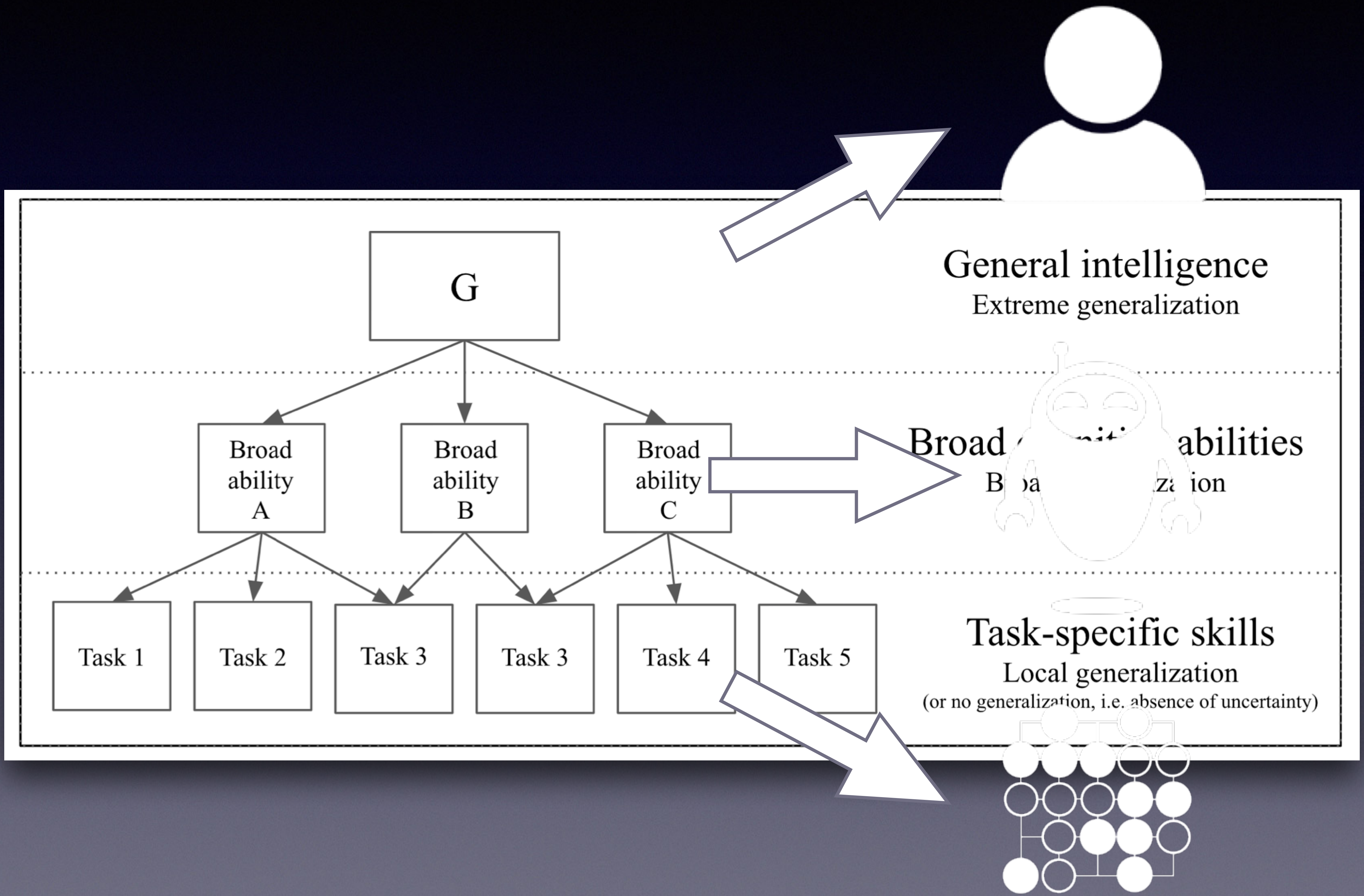
implicit prior knowledge
of developer

Developer-aware

situations
unknown to system +
developer

accounts for prior knowledge
of developer

Generalization



General intelligence
Extreme generalization

Broad c[illegible] abilities
B[illegible]a[illegible]za[illegible]ion

Task-specific skills
Local generalization
(or no generalization, i.e. absence of uncertainty)

Measuring Broad Generalization

- Tools from Psychometrics
- broad battery of tasks + unknown tasks
- not Multi-Task benchmarks: tasks known to developer (*prior + external knowledge*)
- Example Benchmarks (Reinforcement Learning): Animal-AI Olympics, GVG-AI

Psychometrics Tests for AI

- implicit skill assumptions (crystallized): reading, writing
- Principles apply to AI Benchmark
  - Measuring abilities
  - Batteries of tasks
  - Reliability, Validity, Standardization, Freedom from bias

Broad Generalization

- generalization orthogonal to priors / experience
- Deep Learning: currently local generalization / “robustness”, maybe able to achieve broad generalization
    - current Benchmarks fail at testing this level of generalization

Human-like intelligence as goal for AI

- Human possess g-factor: “general intelligence”
  - different cognitive abilities to varying degrees but correlated across tasks
- Human intelligence either best implementation of intelligence or best for our set of tasks
  - Problem: often other systems only considered as intelligent if they display human-like behaviors (language, tool use) not match broadly accepted definitions of intelligence

Human-like intelligence as goal for AI

- Human intelligence is not universal and biased for human-relevant tasks
    - Fail Traveling Salesman Problem for longest path
    - Fail tasks in >3 Dimensions
- “General intelligence”: Spectrum of Scope ✕ Efficiency ✕ Generalization Difficulty

Human-like intelligence as goal for AI

BUT

human relevant tasks are more assessable, approachable and easier to understand

Core Knowledge

A. Objectness and elementary physics

- principles of cohesion, persistence, contact

B. Agentness and goal-directedness

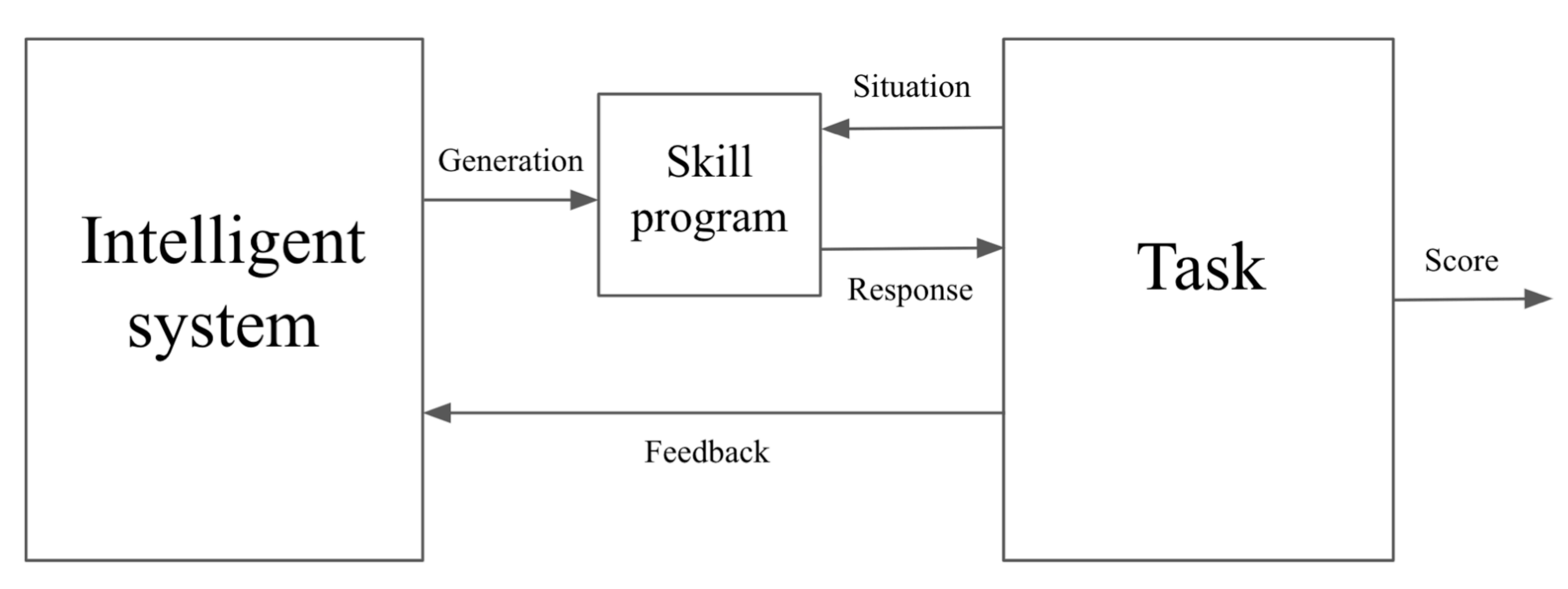
C. Natural numbers and elementary arithmetic

D. Elementary geometry and topology

- distance, orientation, in/put relationships in environment

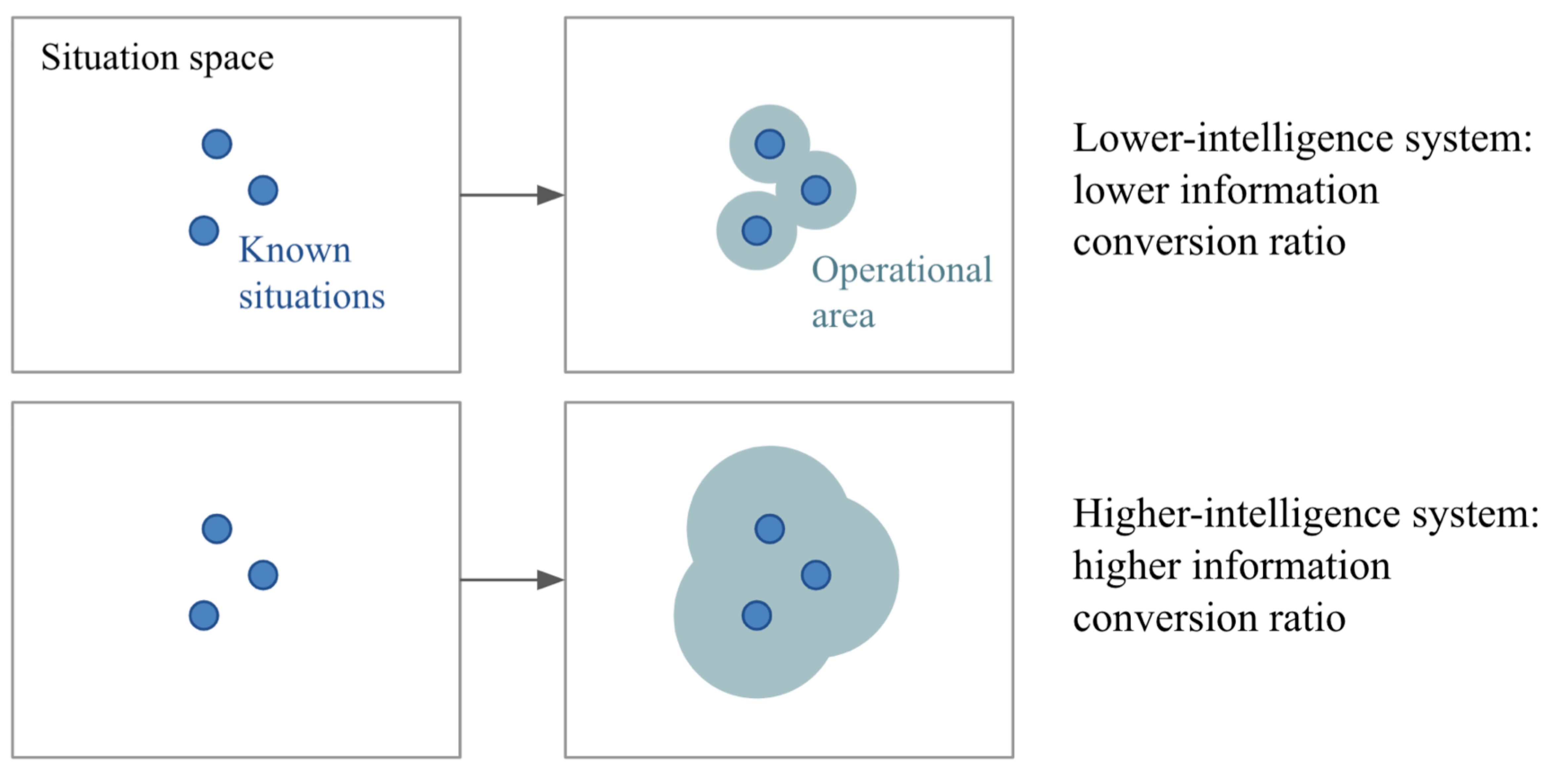
?

Intelligence formalized



intelligence is the rate at which a learner turns its experience and priors into new skills at valuable tasks that involve uncertainty and adaptation

Intelligence formalized



Lower-intelligence system: lower information conversion ratio

Higher-intelligence system: higher information conversion ratio

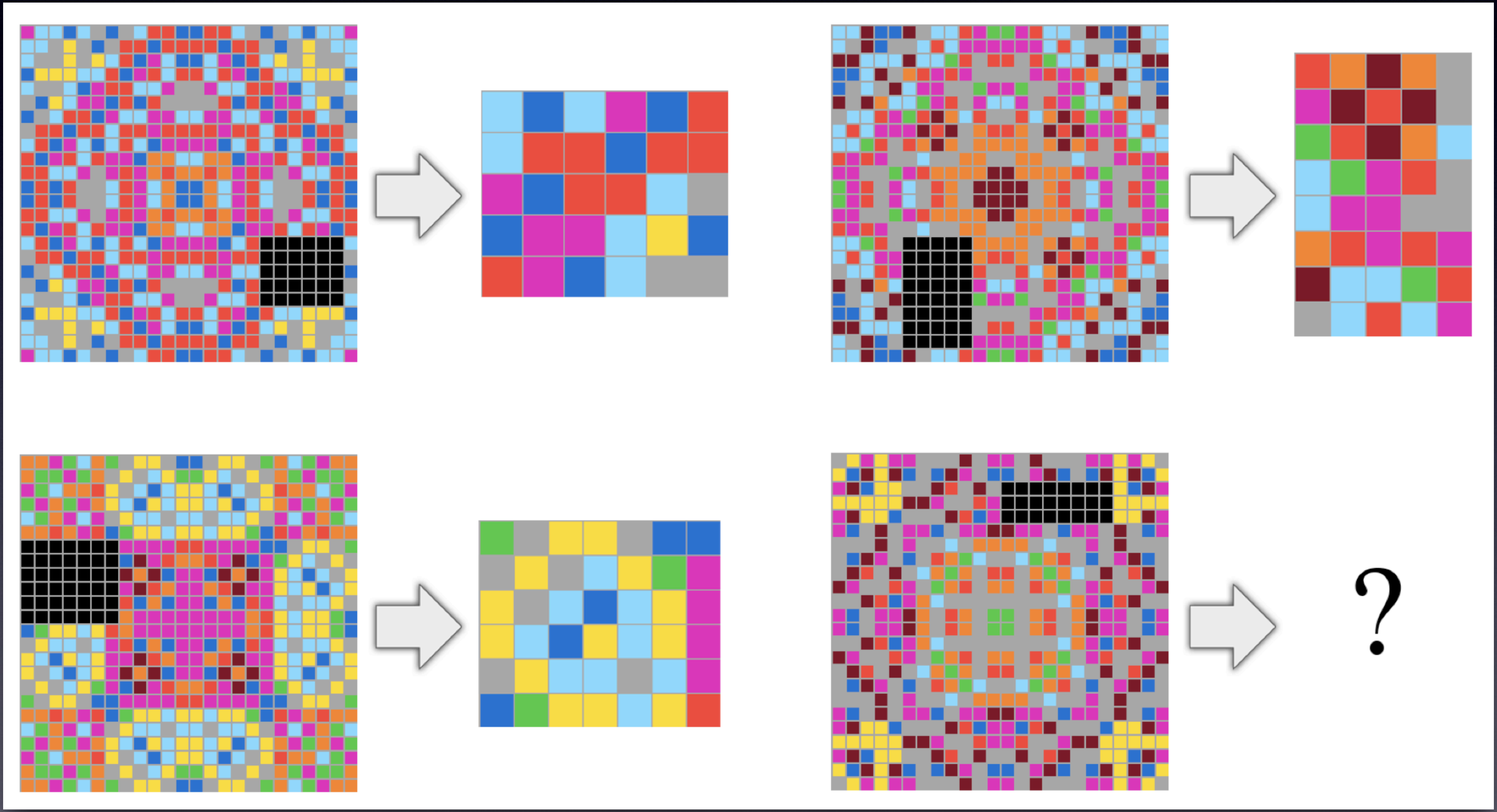
Intelligence formalized

- Possible additions:
  - Computation efficiency (skill program + intelligent system)
  - Time, energy, risk efficiency
- Practical implications: program synthesis, curriculum development

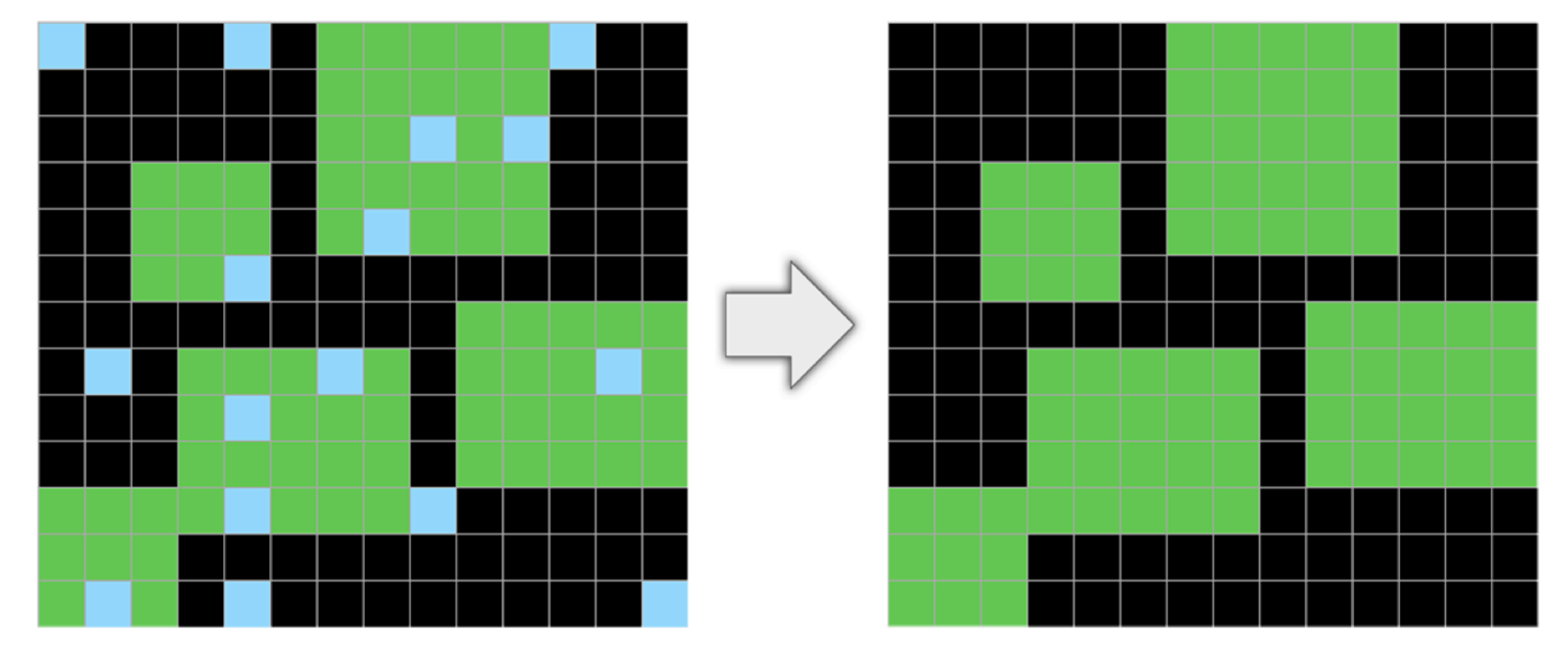
Benchmark:
Abstraction and Reasoning Corpus (ARC)

- similar to Raven's Progressive Matrices (classical IQ test)
- close to psychometric IQ test
  - human- + machine-approachable
  - no specific training required
- developer-aware generalization: evaluation features only novel tasks
- assumes only Core Knowledge priors

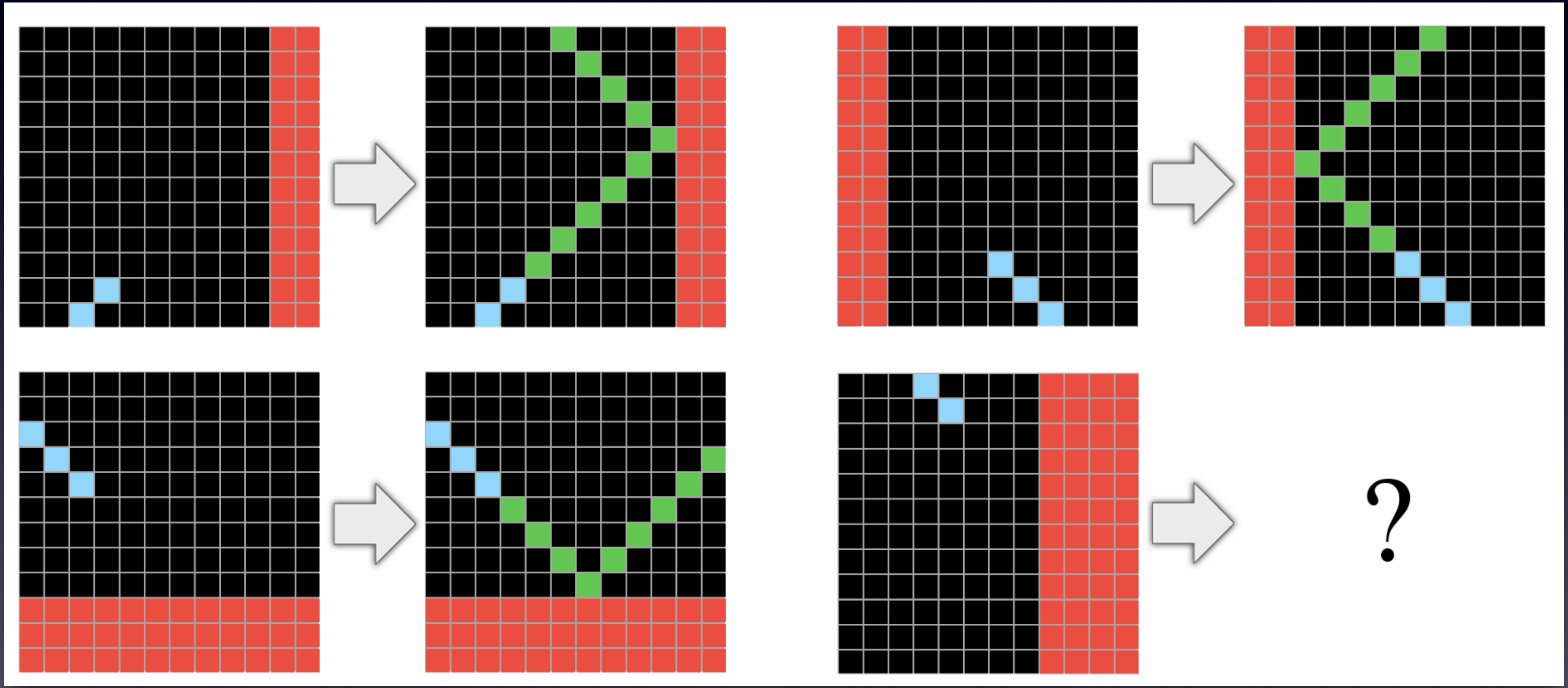
Benchmark:
Abstraction and Reasoning Corpus (ARC)



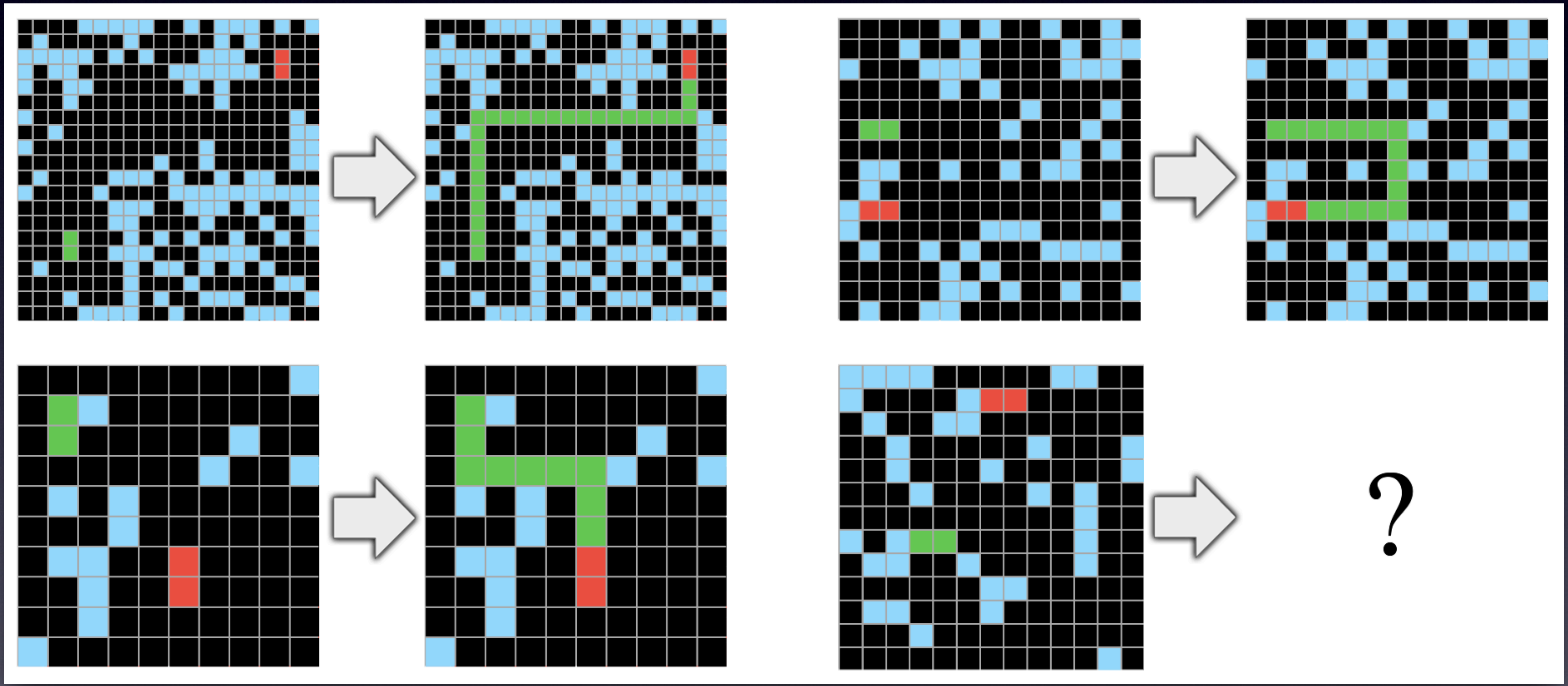
Benchmark: Abstraction and Reasoning Corpus (ARC)



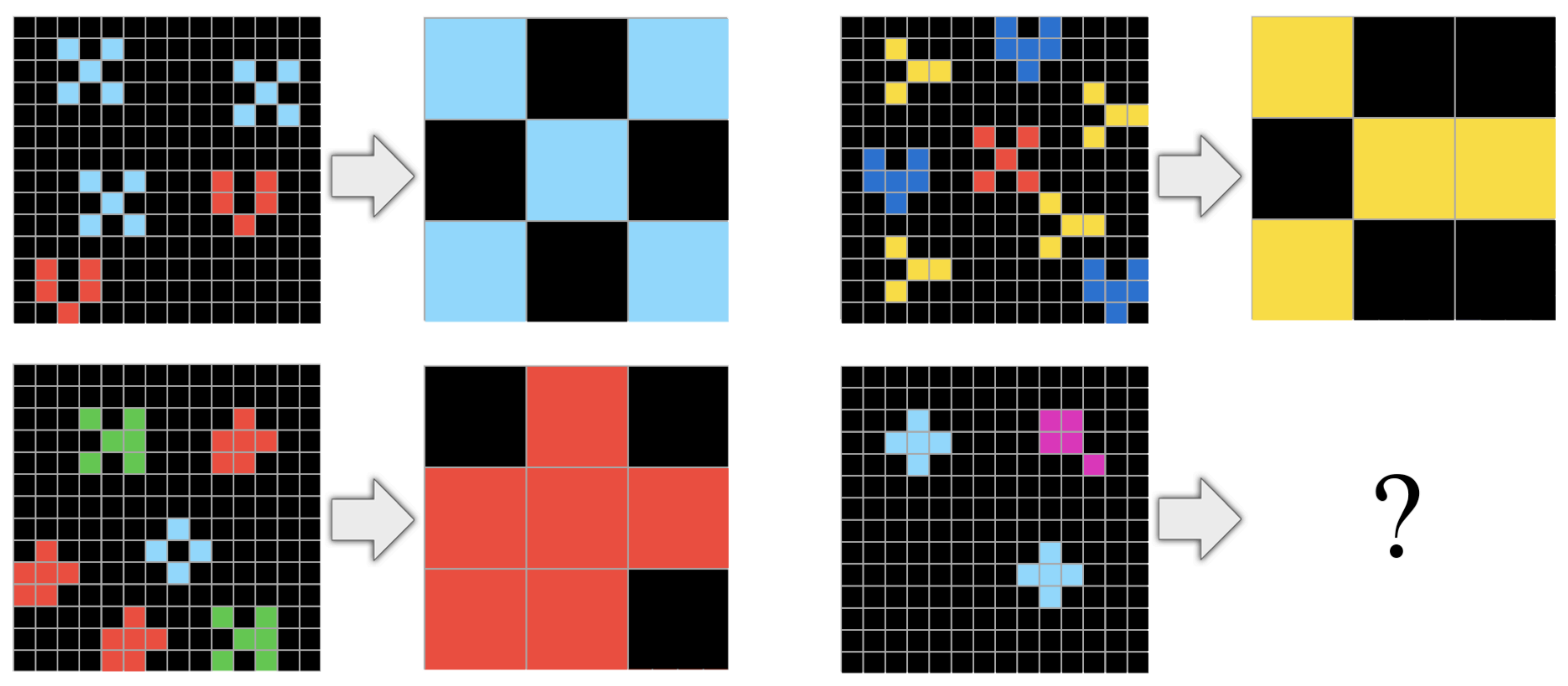
Benchmark:
Abstraction and Reasoning Corpus (ARC)



Benchmark:
Abstraction and Reasoning Corpus (ARC)



Benchmark:
Abstraction and Reasoning Corpus (ARC)



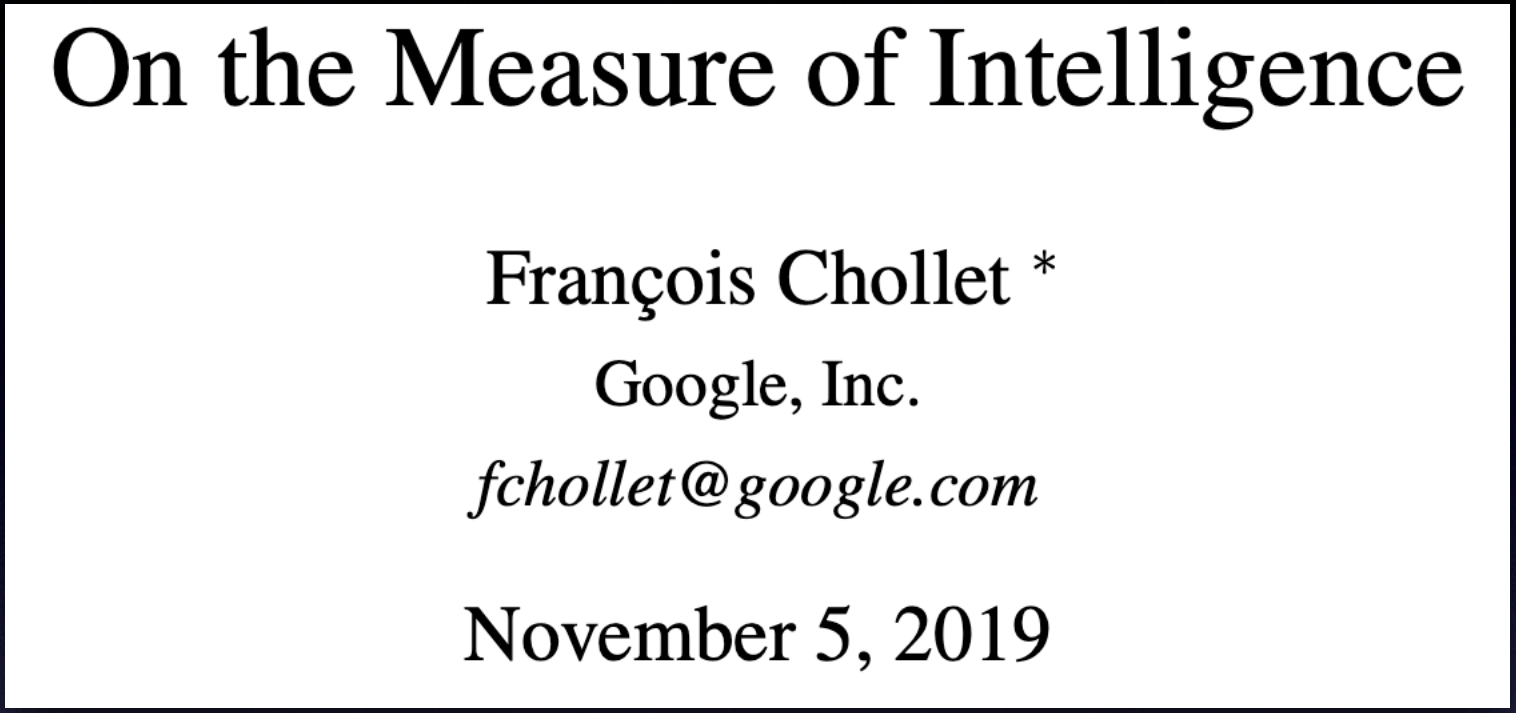
Benchmark:
Abstraction and Reasoning Corpus (ARC)

- Pros:
  - no crystallized intelligence required
    - language, real-world images, common sense
  - diverse tasks
  - unique tasks
  - not programmatically generated

Benchmark:
Abstraction and Reasoning Corpus (ARC)

- Cons:
  - no generalization difficulty of tasks
    - assess via human performance
  - validity: transfer to real-world problems
  - limited dataset size (1000 tasks)
  - Evaluation feedback only binary

Thank you :)



On the Measure of Intelligence

François Chollet *
Google, Inc.
fchollet@google.com

November 5, 2019