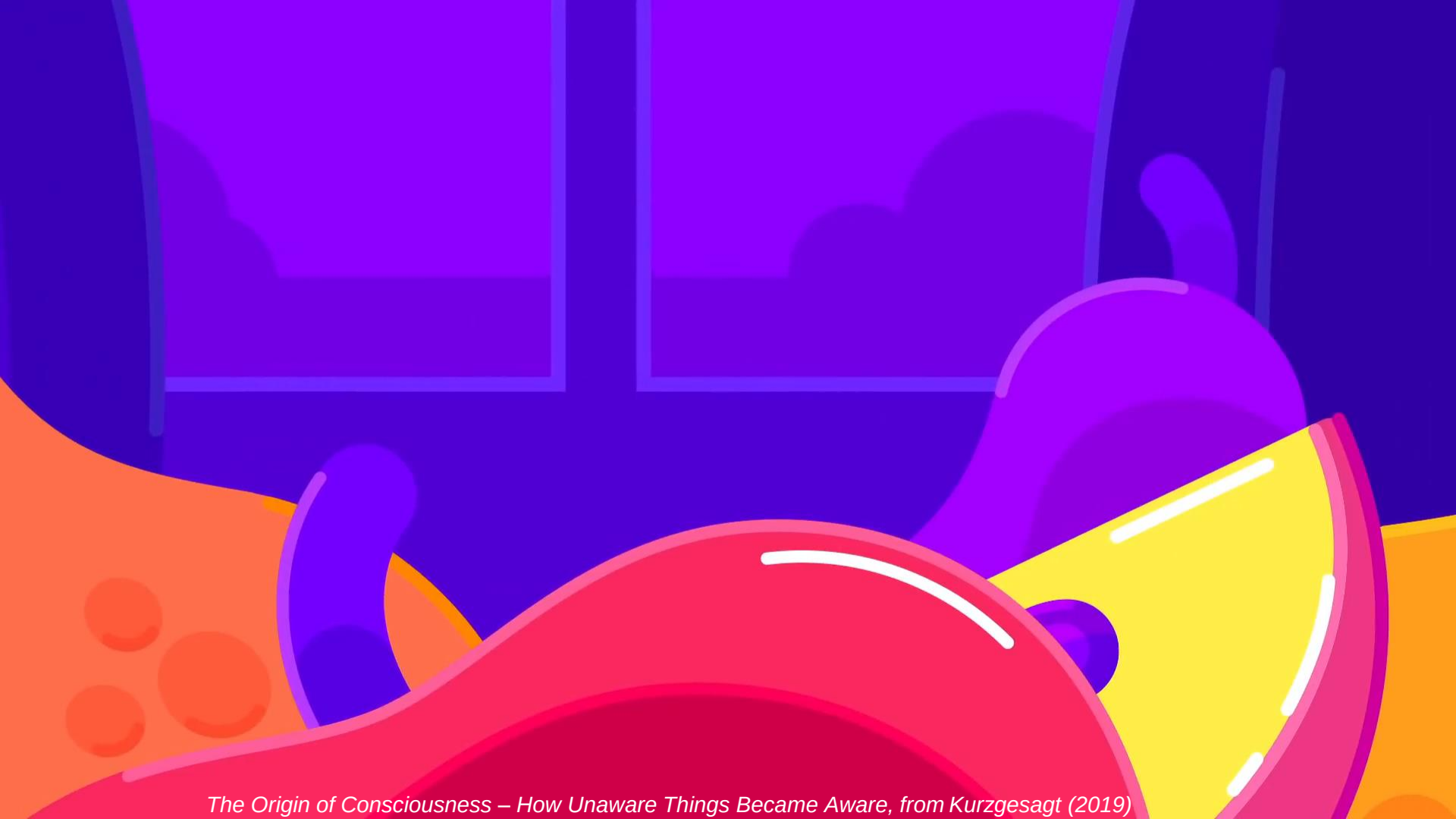
Explanatory learning

Può una macchina imparare a formulare teorie?

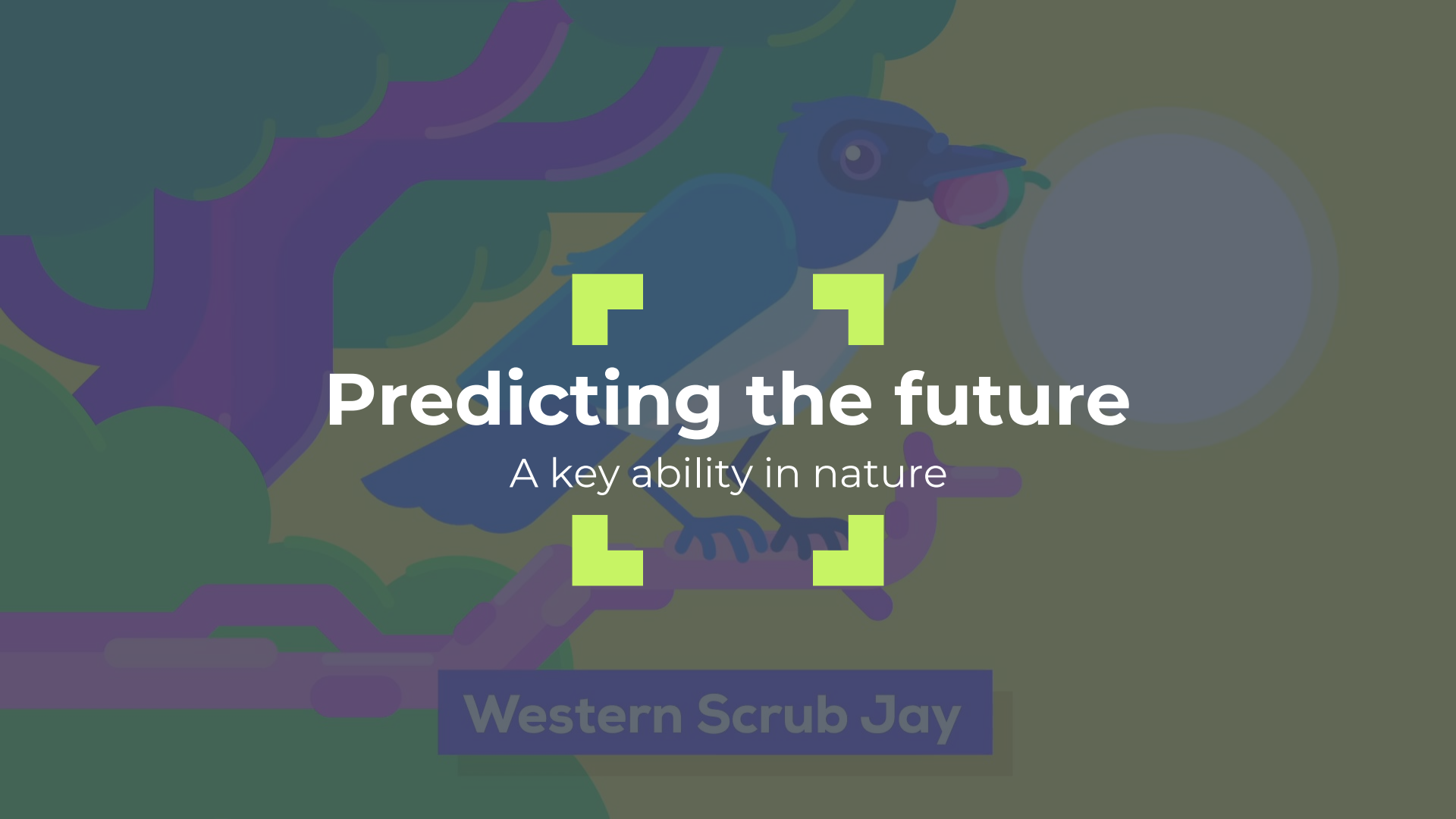
Antonio Norelli

III year PhD student in Computer Science at Sapienza
GLADIA research group - norelli@di.uniroma1.it

Dipartimento di Informatica
Sapienza Università di Roma
GLADIA



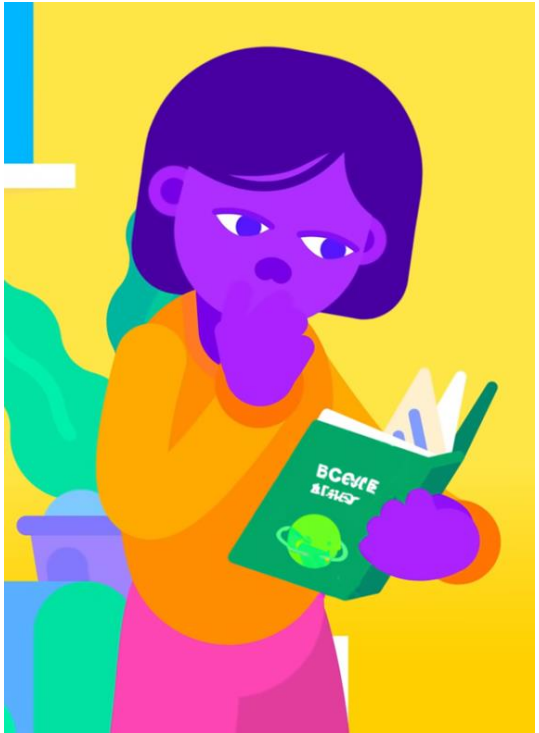
The Origin of Consciousness – How Unaware Things Became Aware, from Kurzgesagt (2019)


Predicting the future
A key ability in nature
Western Scrub Jay

Predicting the future

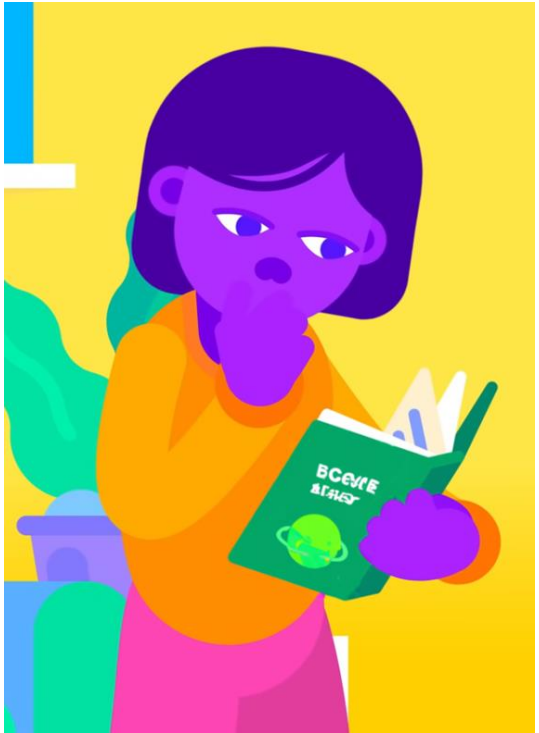
No animal come even close to humans

Human unique stystem



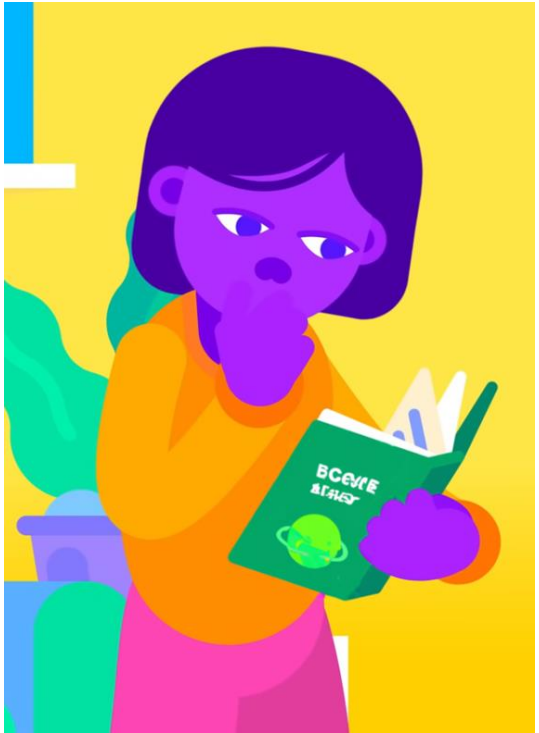
Human unique stystem

Four wandering stars having their period around a principal star



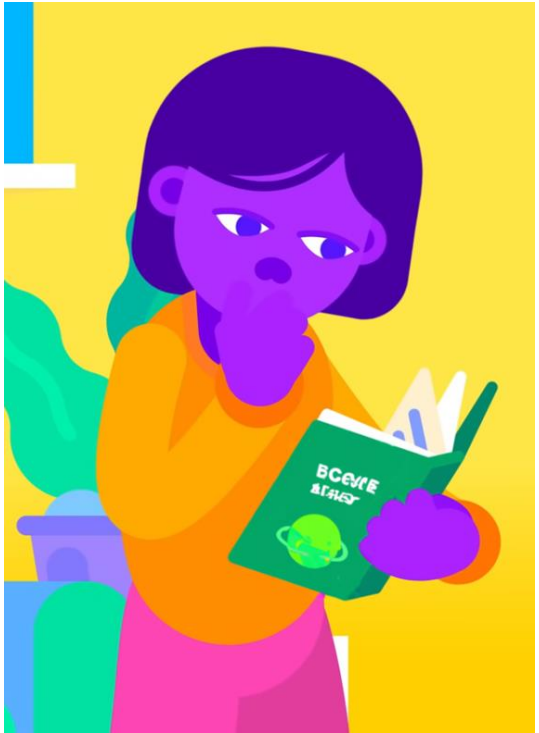
Human unique stystem

Four wandering stars having their period around a principal star



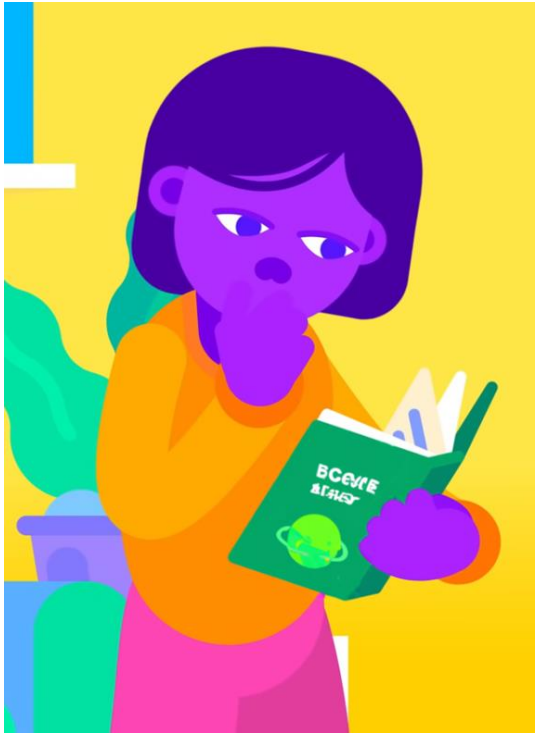
Human unique stystem

Four wandering stars having their period around a principal star



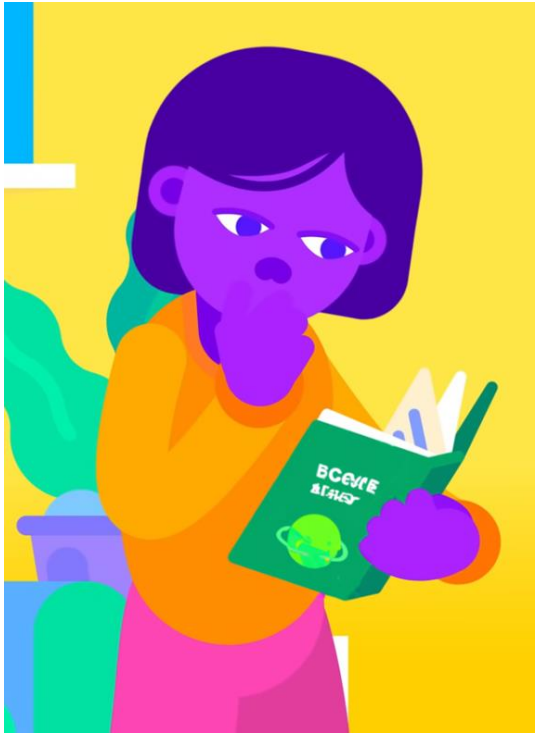
Human unique stystem

Four wandering stars having their period around a principal star



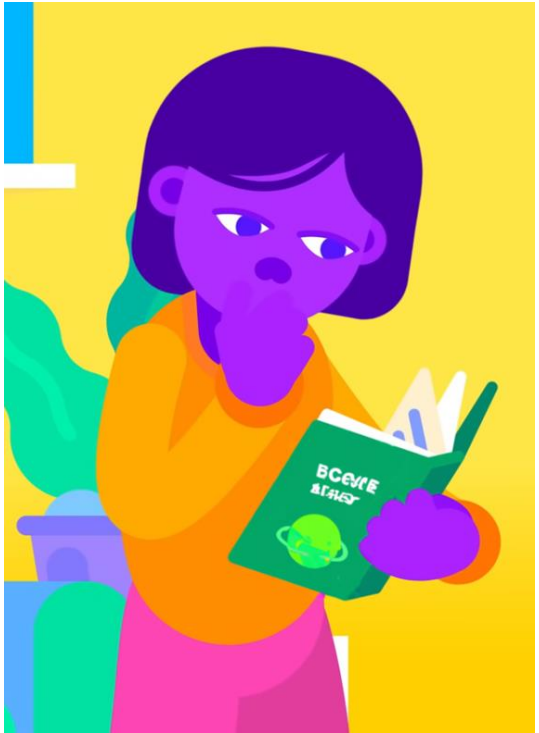
Human unique stystem

Four wandering stars having their period around a principal star



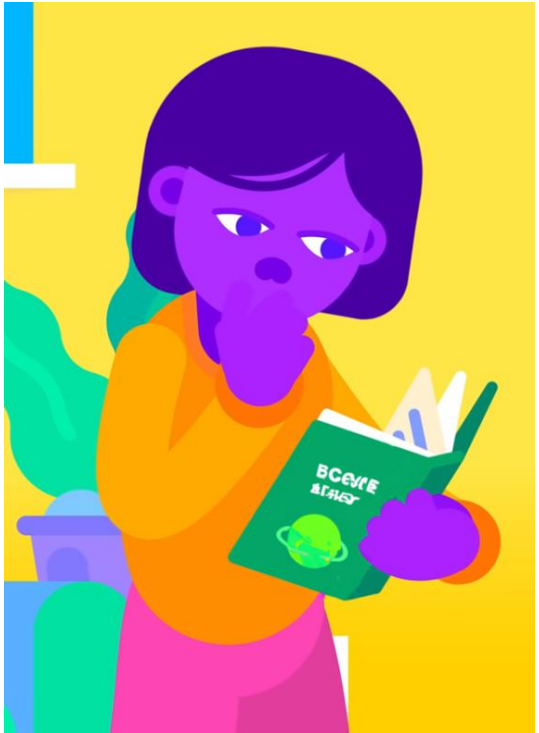
Human unique stystem

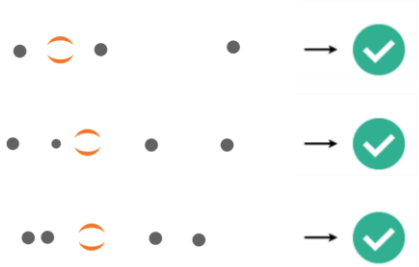
Four wandering stars having their period around a principal star



Human unique stystem

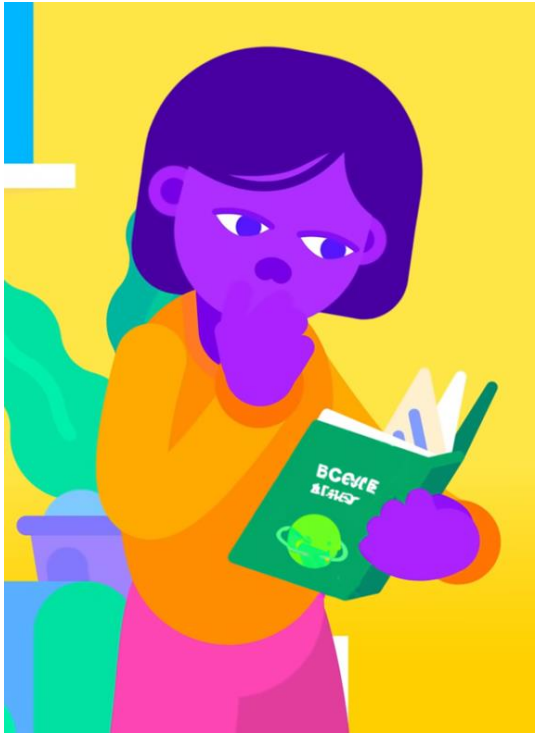
Four wandering stars having their period around a principal star

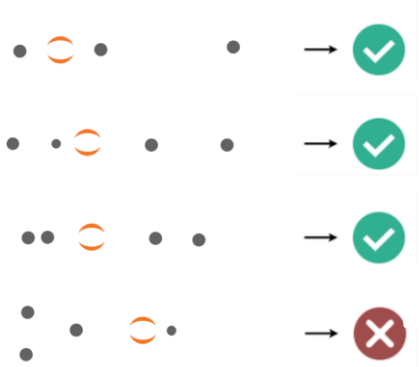




Human unique stystem

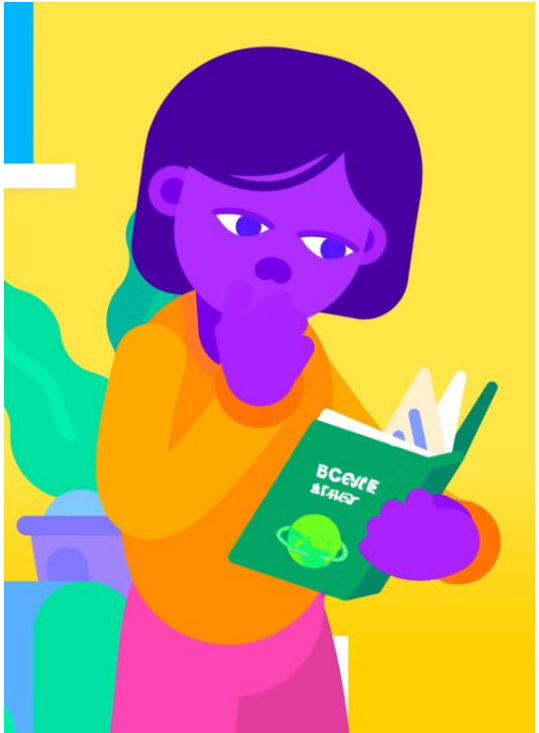
Four wandering stars having their period around a principal star

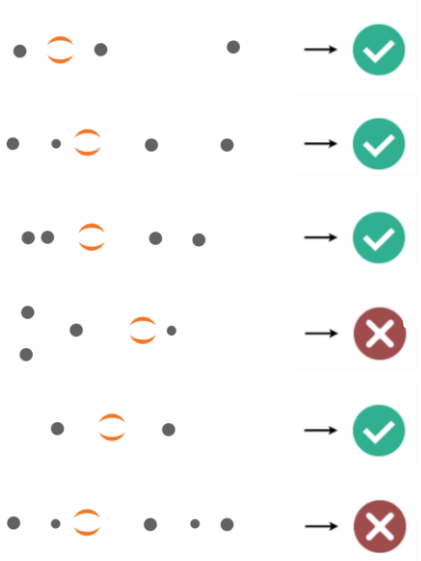




Human unique stystem

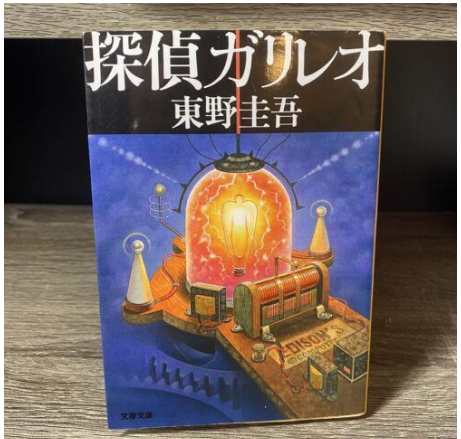
Four wandering stars having their period around a principal star

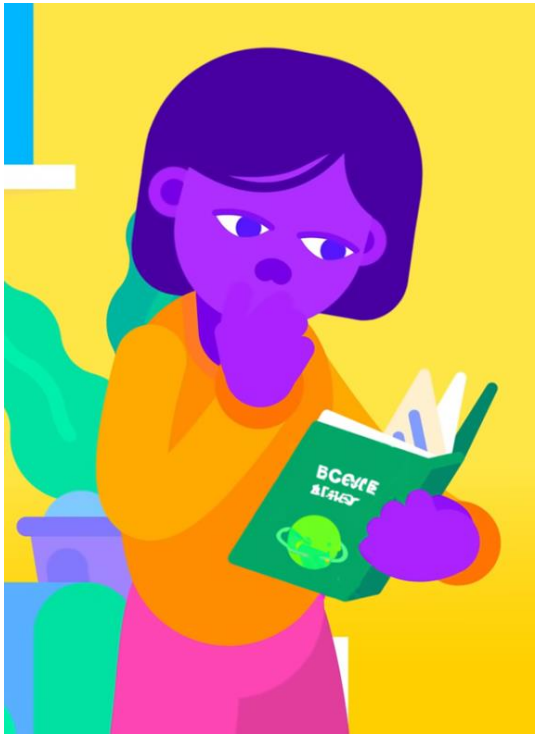


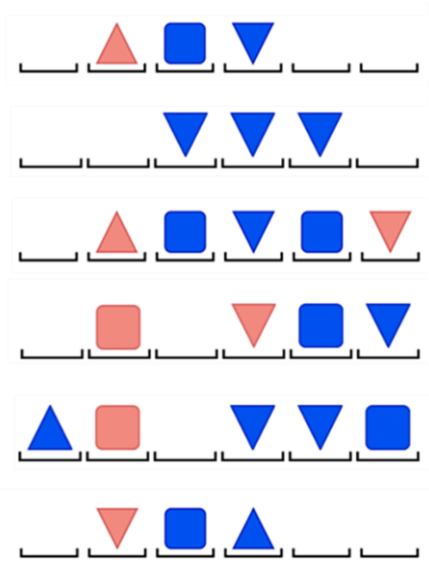


Human unique stystem: Mastering a language

三角形で囲まれた正確に1つの青

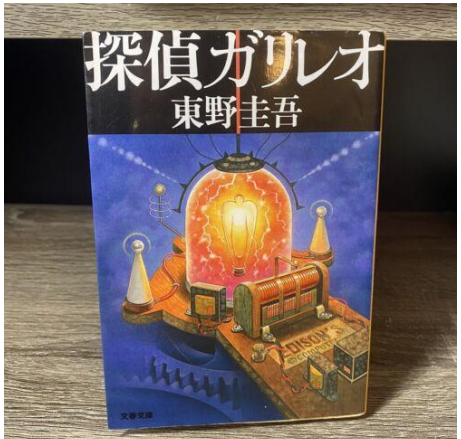


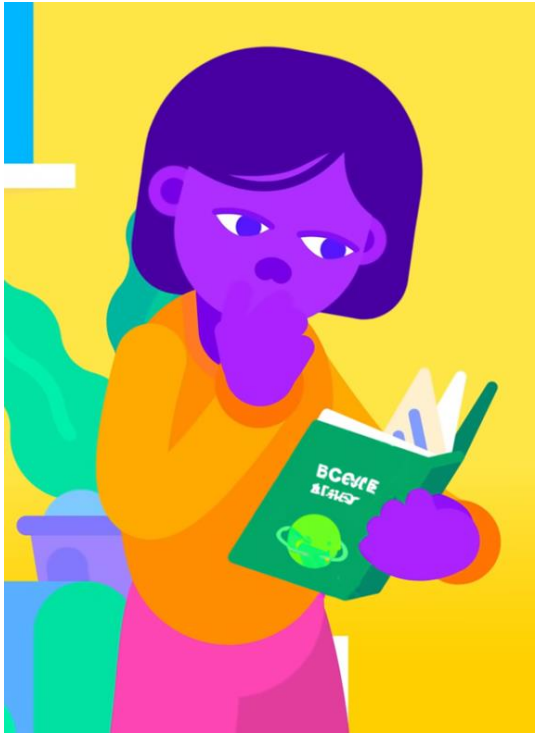


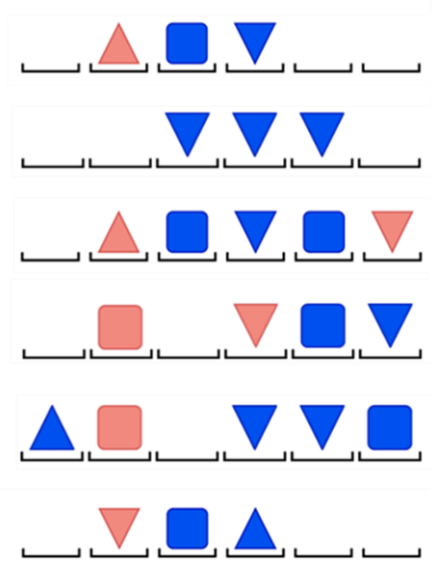


Human unique stystem: Mastering a language

Exactly one blue surrounded by triangles

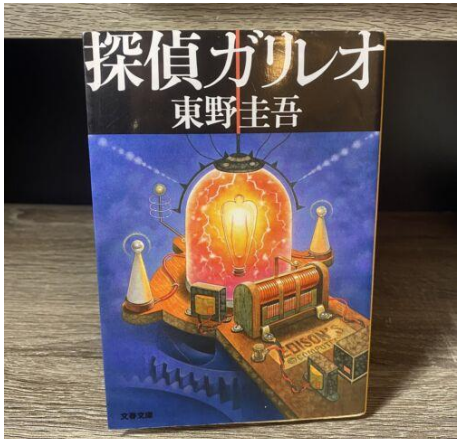


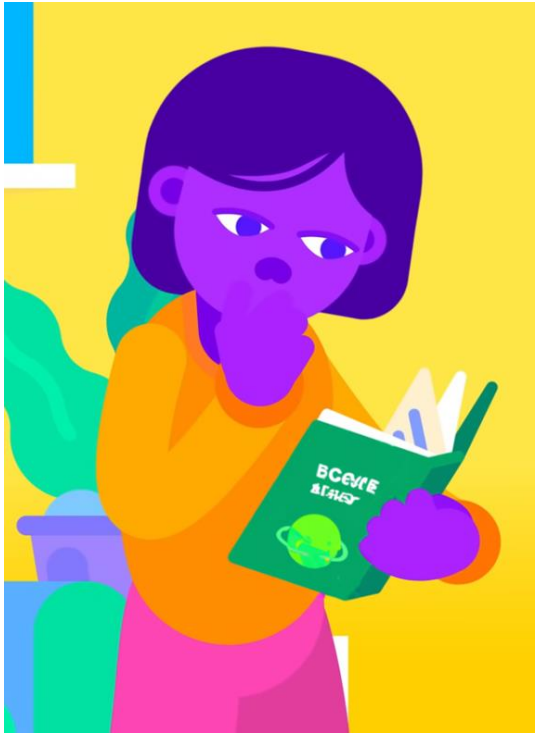


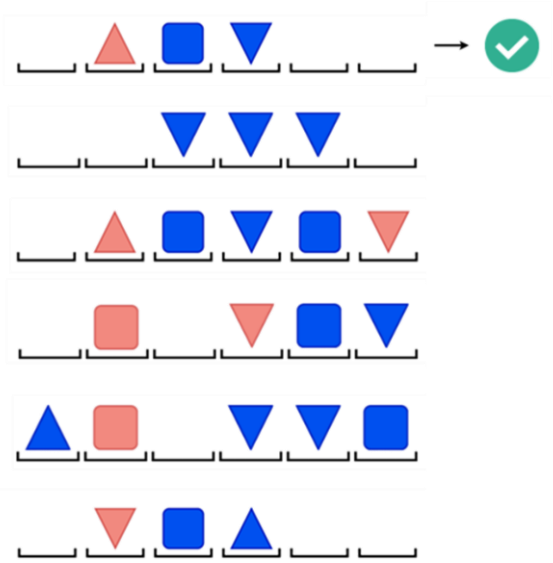


Human unique stystem: Mastering a language

Exactly one blue surrounded by triangles

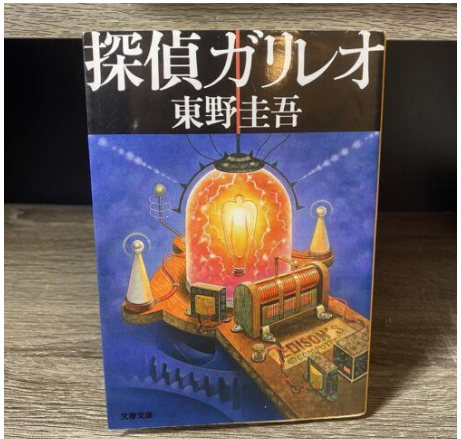


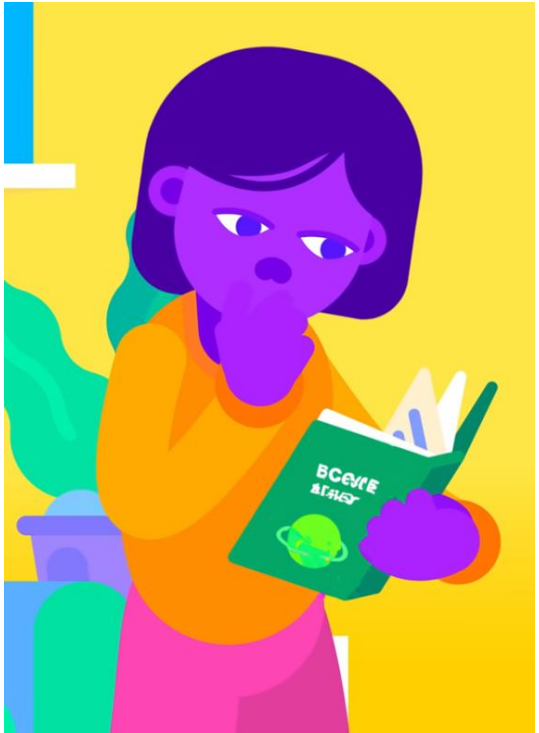


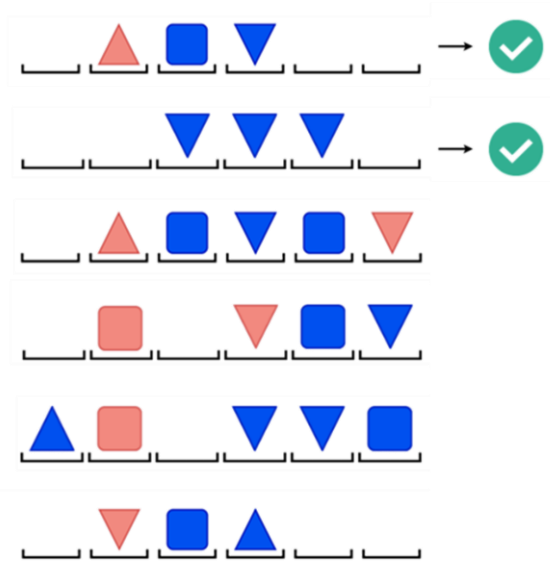


Human unique stystem: Mastering a language

Exactly one blue surrounded by triangles

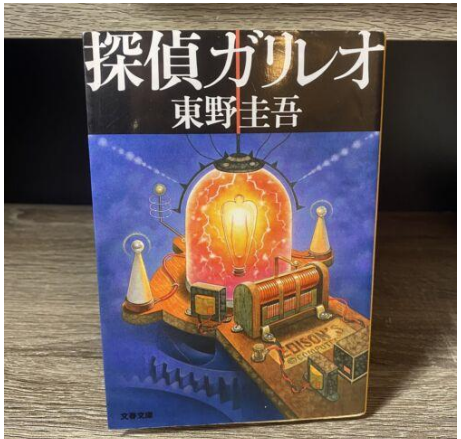


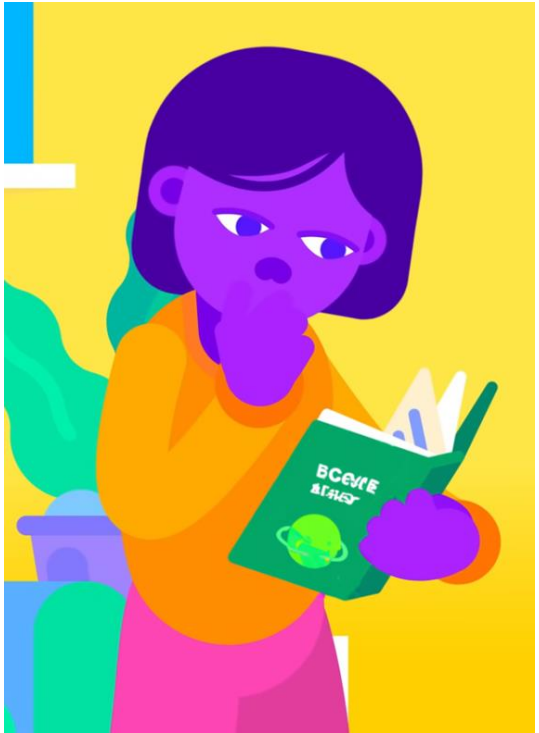


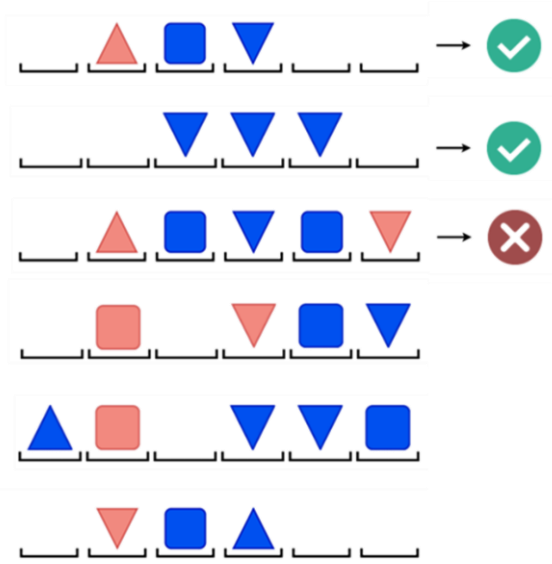


Human unique stystem: Mastering a language

Exactly one blue surrounded by triangles

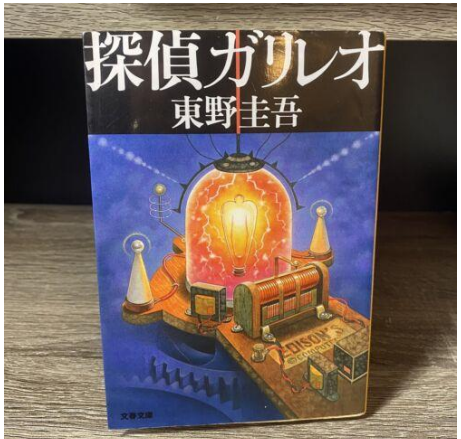


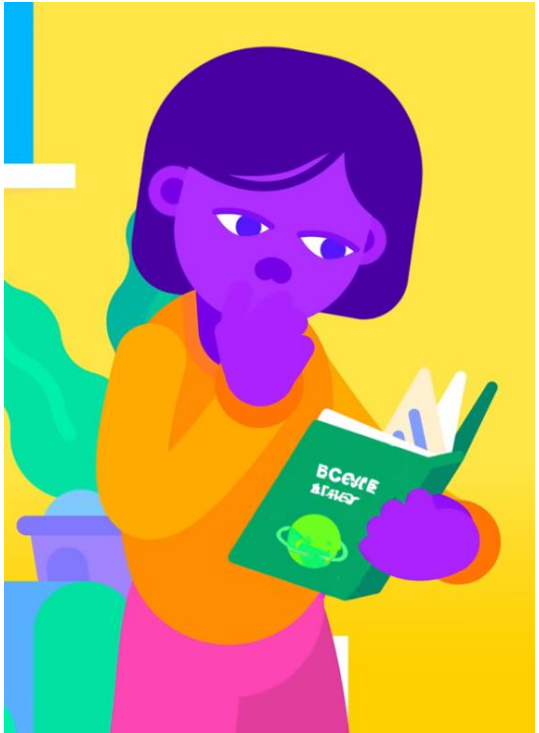


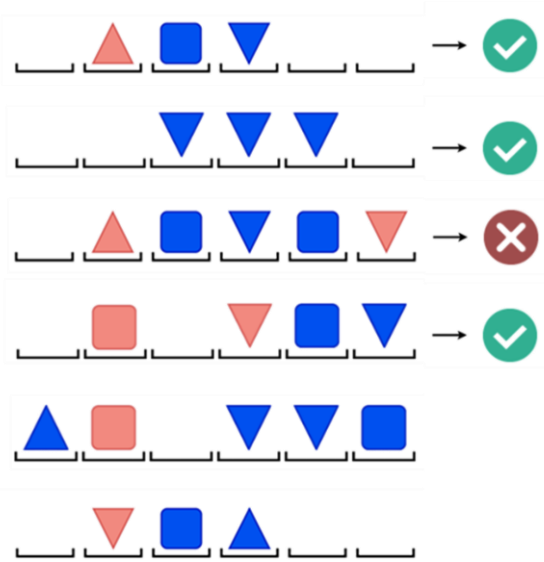


Human unique stystem: Mastering a language

Exactly one blue surrounded by triangles

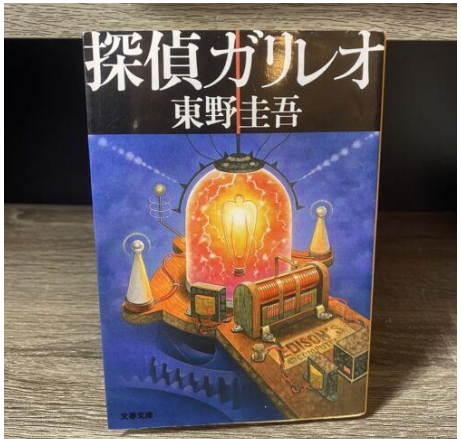


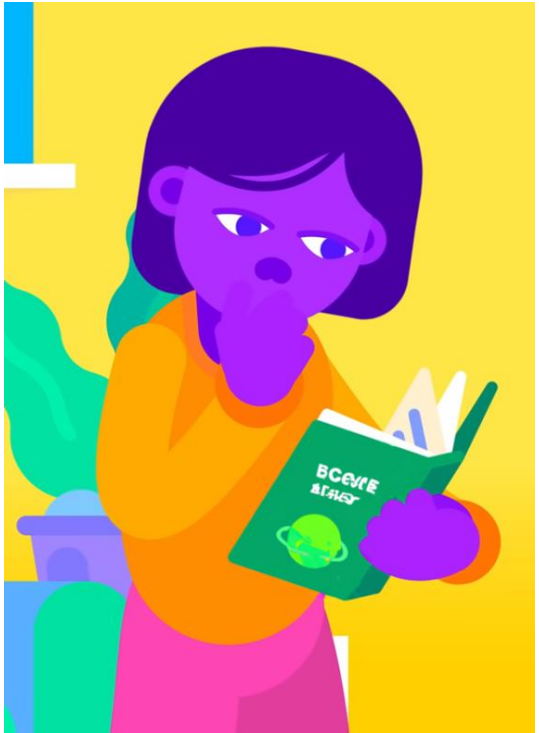


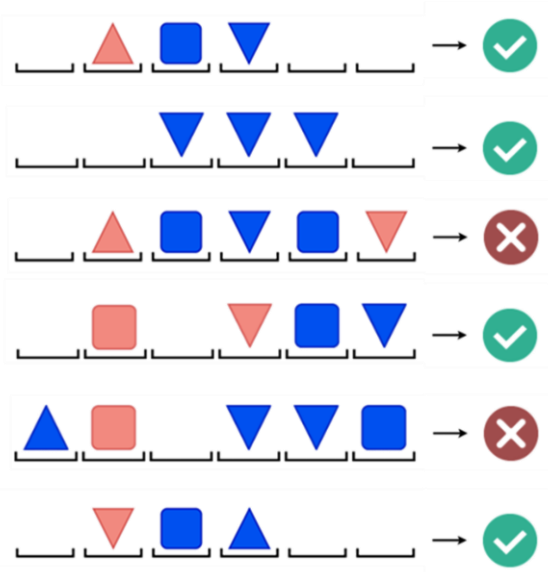


Human unique stystem: Mastering a language

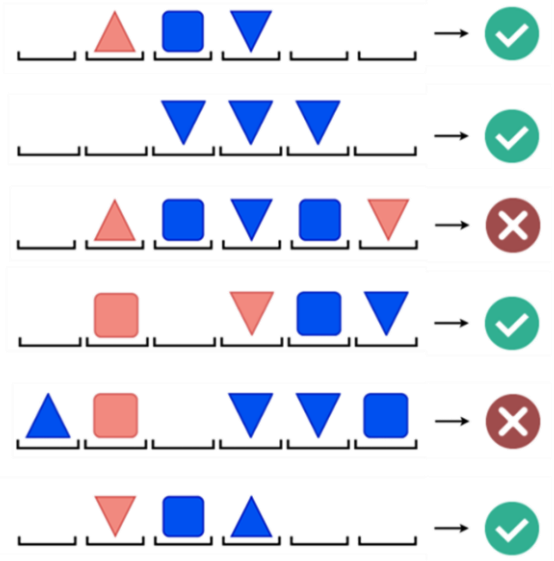
Exactly one blue surrounded by triangles

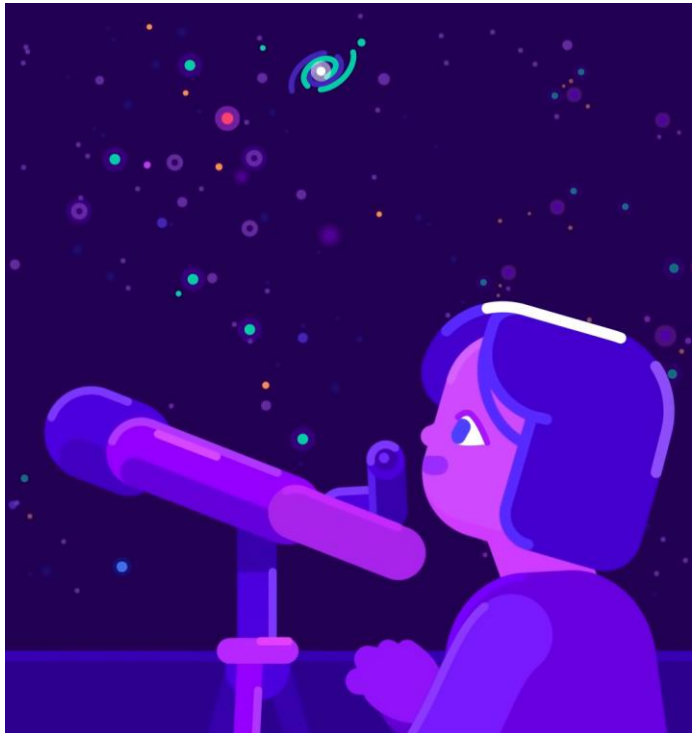




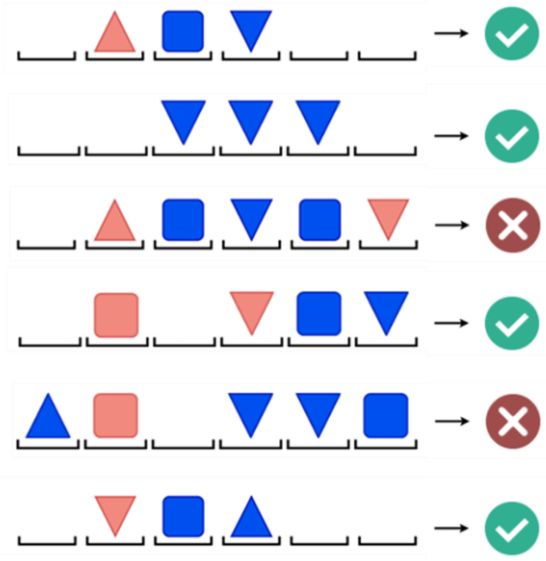


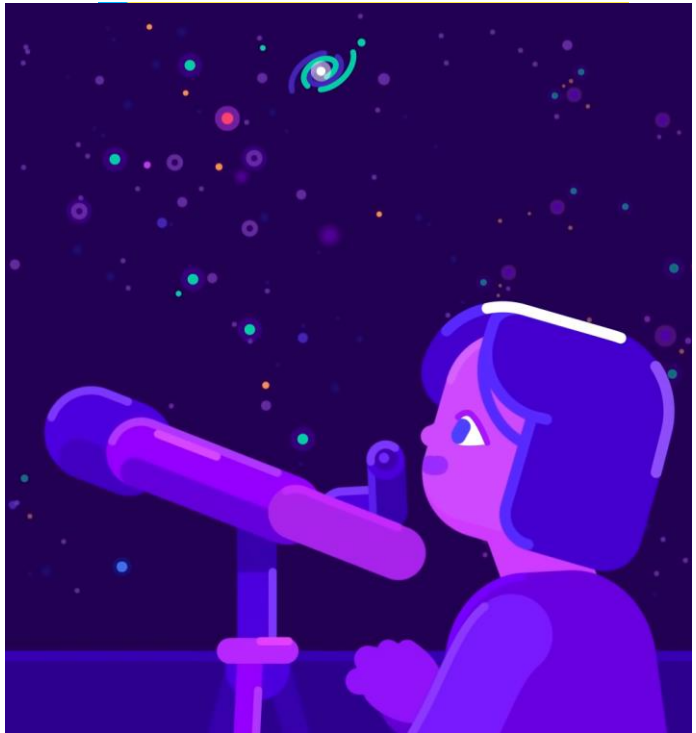
Human unique stystem: Mastering a language





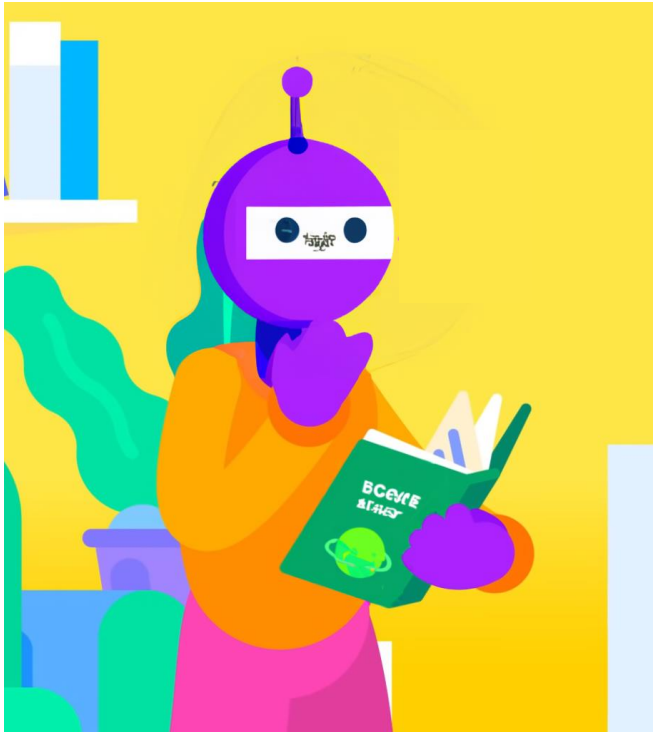
Human unique stystem: Mastering a language

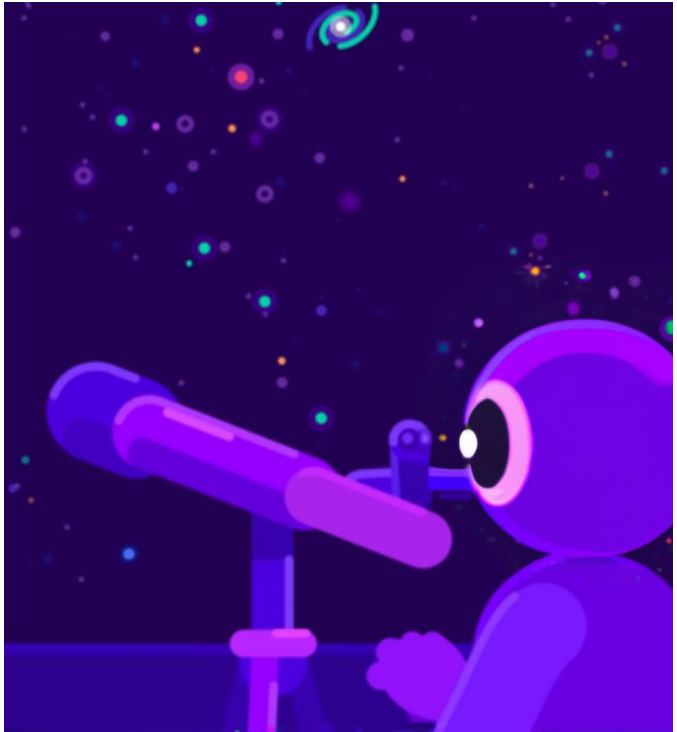




Exactly one blue surrounded by triangles

How can we make machines take part in this orchestra?





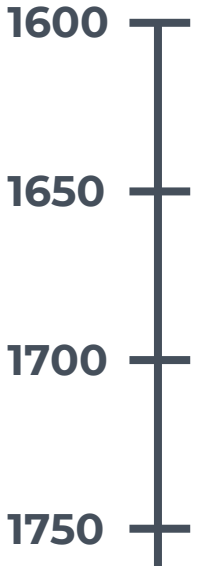
How can we build machines that creatively invent entirely new theories from data, like scientists do?

1. Prologue
2. Explanatory Learning

invent new theories from data

Epistemology:
invent new theories from data

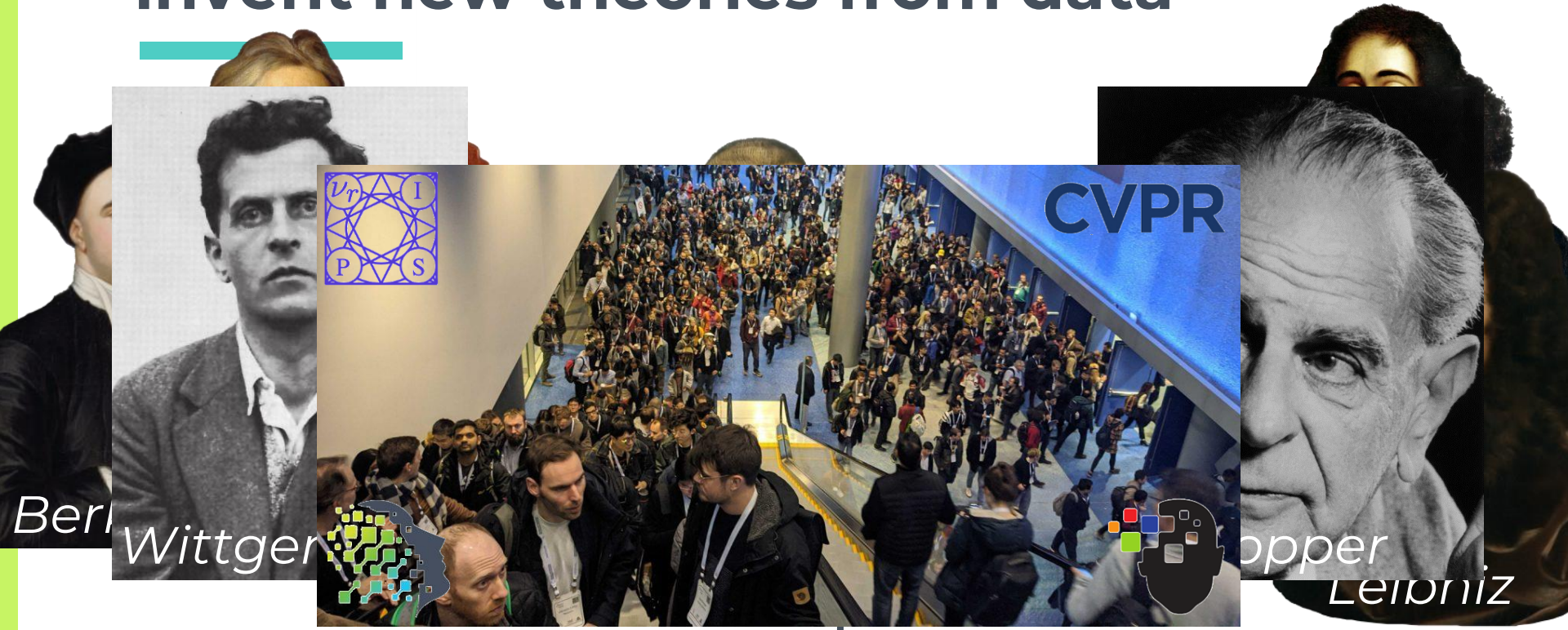
Epistemology:
invent new theories from data



Empiricists

Rationalists

Epistemology:
invent new theories from data



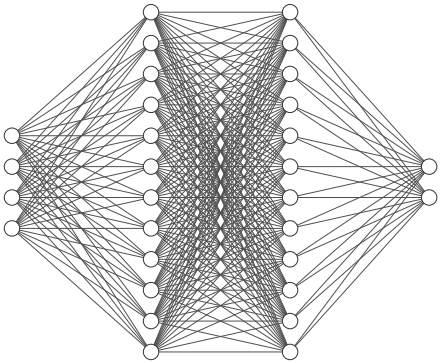
Empiricists

1750

Rationalists

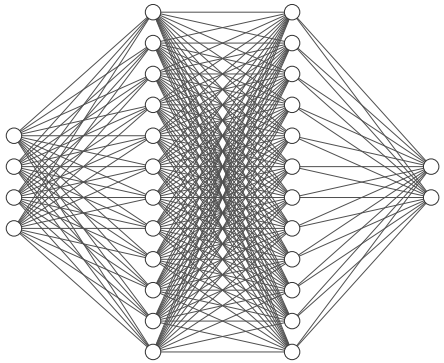
Deep learning
is aligned with the empiricist view

```
--1----11--1---11----------1---1--1--11---1-1---111-1--11-11--------1---1
-11---1----1-1--1-1--11------1-----1-1---11--1-----------11-------1------
---1-1-11---1---1--------11--1-1-1---1-1-1--1--------1---1--11-11--------
-111-1--------------1--1-------1--1-1---------11---1----1----------11----
-1111---1--11--11--11--------1--11-1-1----1-------1---1-1-11-----1-------
-----1-------1----1-1----1---11-----1----1---1--1-1---------------11-----
-1--1--11-1-----1-------------11-----1-------1-------1---11-------------1
--1--11-1----------1---1-------1--1---------1-11-1--1------1---1-------1-
```



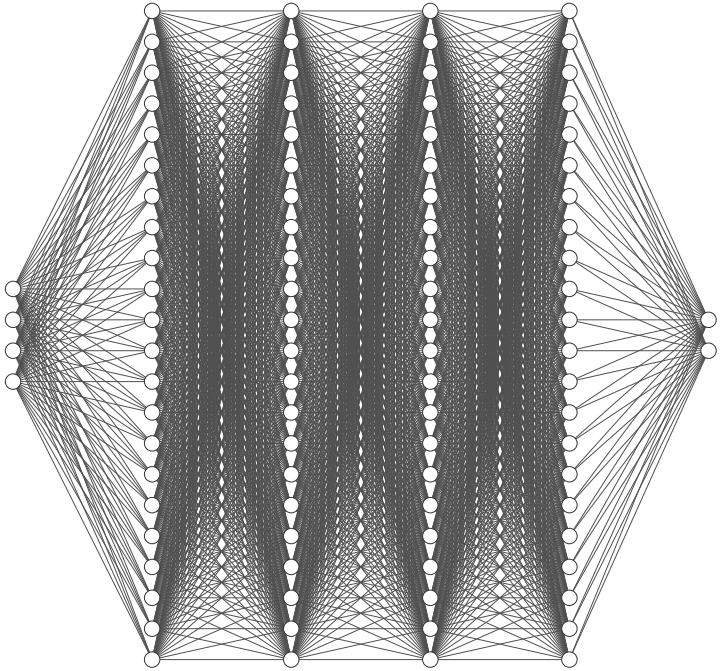
Deep learning
is aligned with the empiricist view

```
-----------------------------------------------------------------------
11111111----11----11----1111111-------1111-------1111111------11----11-
---11-------11----11----11----------11----11-----11----11-----11----11-
---11-------11----11----1111-------11------11----11----11------11--11--
---11-------11111111----11---------11------11----1111111---------11----
---11-------11----11----11----------11----11-----11----11--------11----
---11-------11----11----1111111-------1111-------11-----11-------11----
-----------------------------------------------------------------------
```

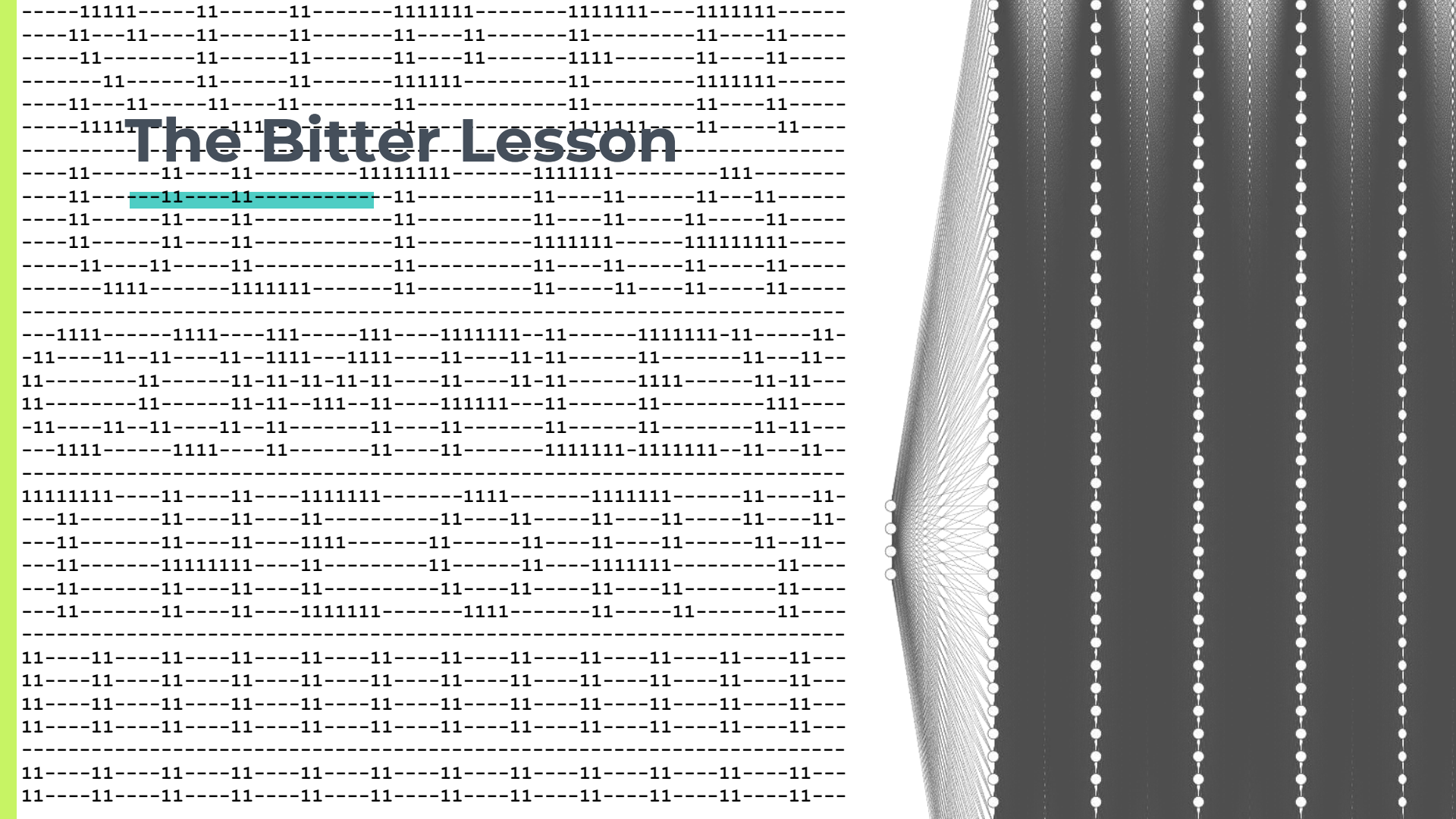


Deep learning
is aligned with the empiricist view

```
-----------------------------------------------------------------------
---1111------1111----111-----111----1111111--11------1111111-11-----11-
-11----11--11----11--1111---1111----11----11-11------11-------11---11--
11--------11------11-11-11-11-11----11----11-11------1111------11-11---
11--------11------11-11--111--11----111111---11------11---------111----
-11----11--11----11--11-------11----11-------11------11--------11-11---
---1111------1111----11-------11----11-------1111111-1111111--11---11--
-----------------------------------------------------------------------
11111111----11----11----1111111-------1111-------1111111------11----11-
---11-------11----11----11----------11----11-----11----11-----11----11-
---11-------11----11----1111-------11------11----11----11------11--11--
---11-------11111111----11---------11------11----1111111---------11----
---11-------11----11----11----------11----11-----11----11--------11----
---11-------11----11----1111111-------1111-------11-----11-------11----
-----------------------------------------------------------------------
```

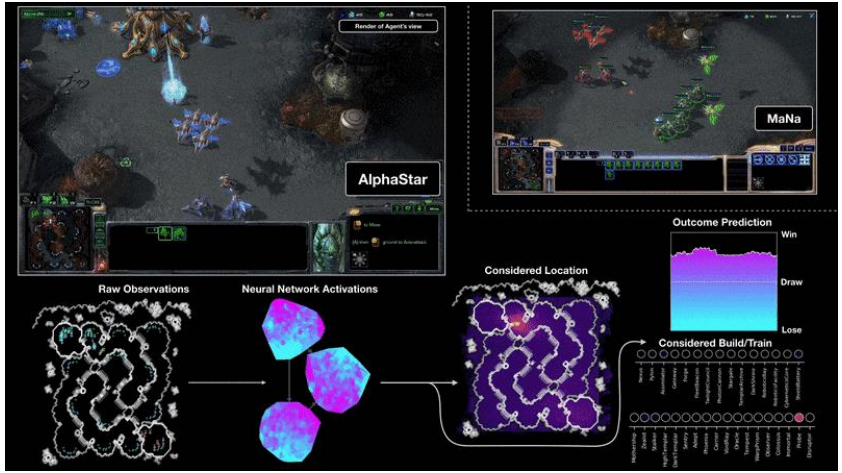


The Bitter Lesson



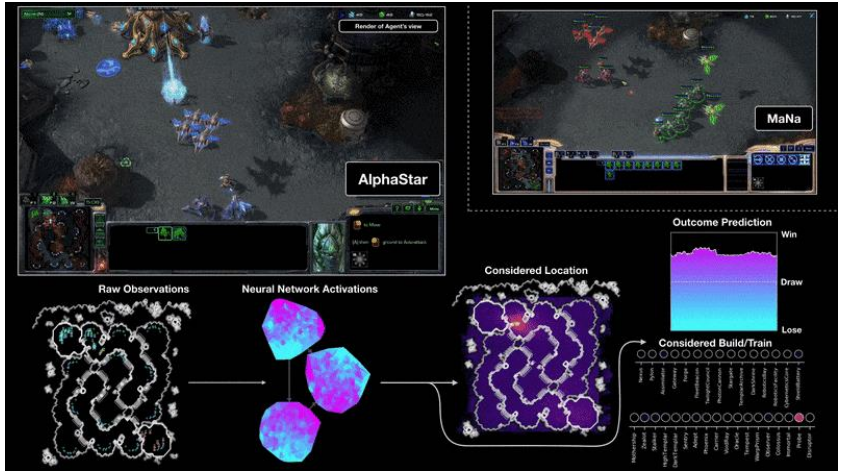
The Bitter Lesson achievements

The Bitter Lesson achievements



Vinyals Oriol, et al. "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II" (2019)

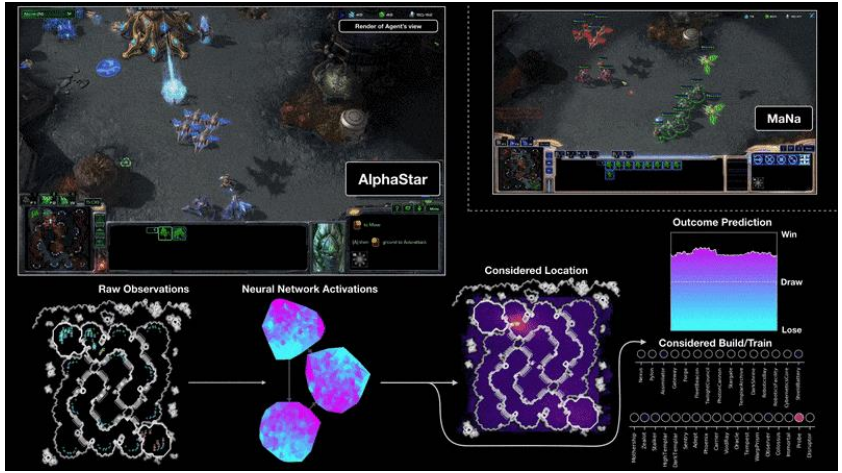
The Bitter Lesson achievements



Vinyals Oriol, et al. "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II" (2019)

Brown Tom B., et al. "Language models are few-shot learners" (2020)

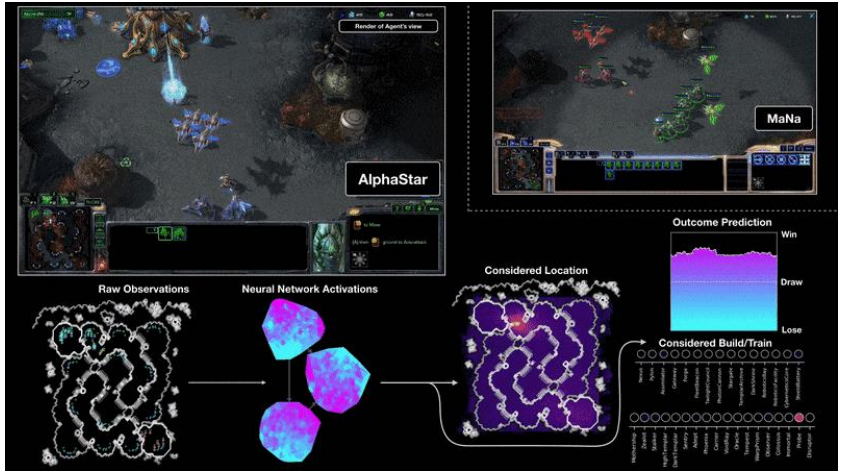
The Bitter Lesson achievements



Vinyals Oriol, et al. "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II" (2019)

Brown Tom B., et al. "Language models are few-shot learners" (2020)

The Bitter Lesson achievements

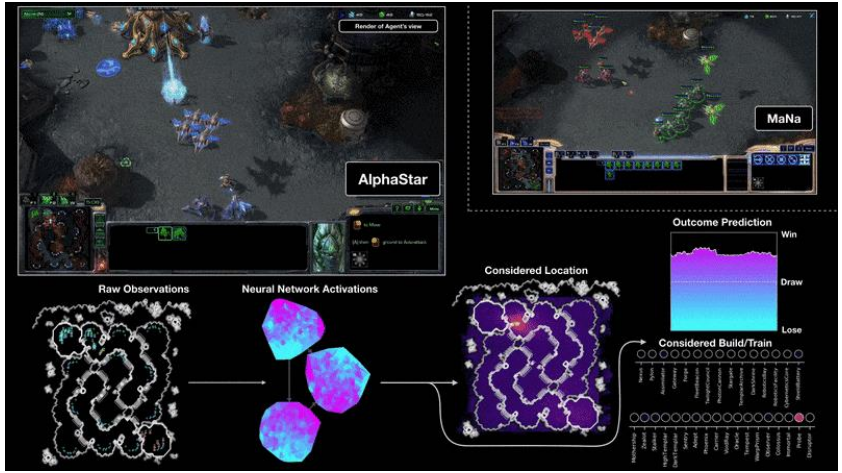


> $ 1,000,000

Vinyals Oriol, et al. "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II" (2019)

Brown Tom B., et al. "Language models are few-shot learners" (2020)

The Bitter Lesson achievements



> $ 1,000,000

BIG Data

Vinyals Oriol, et al. "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II" (2019)

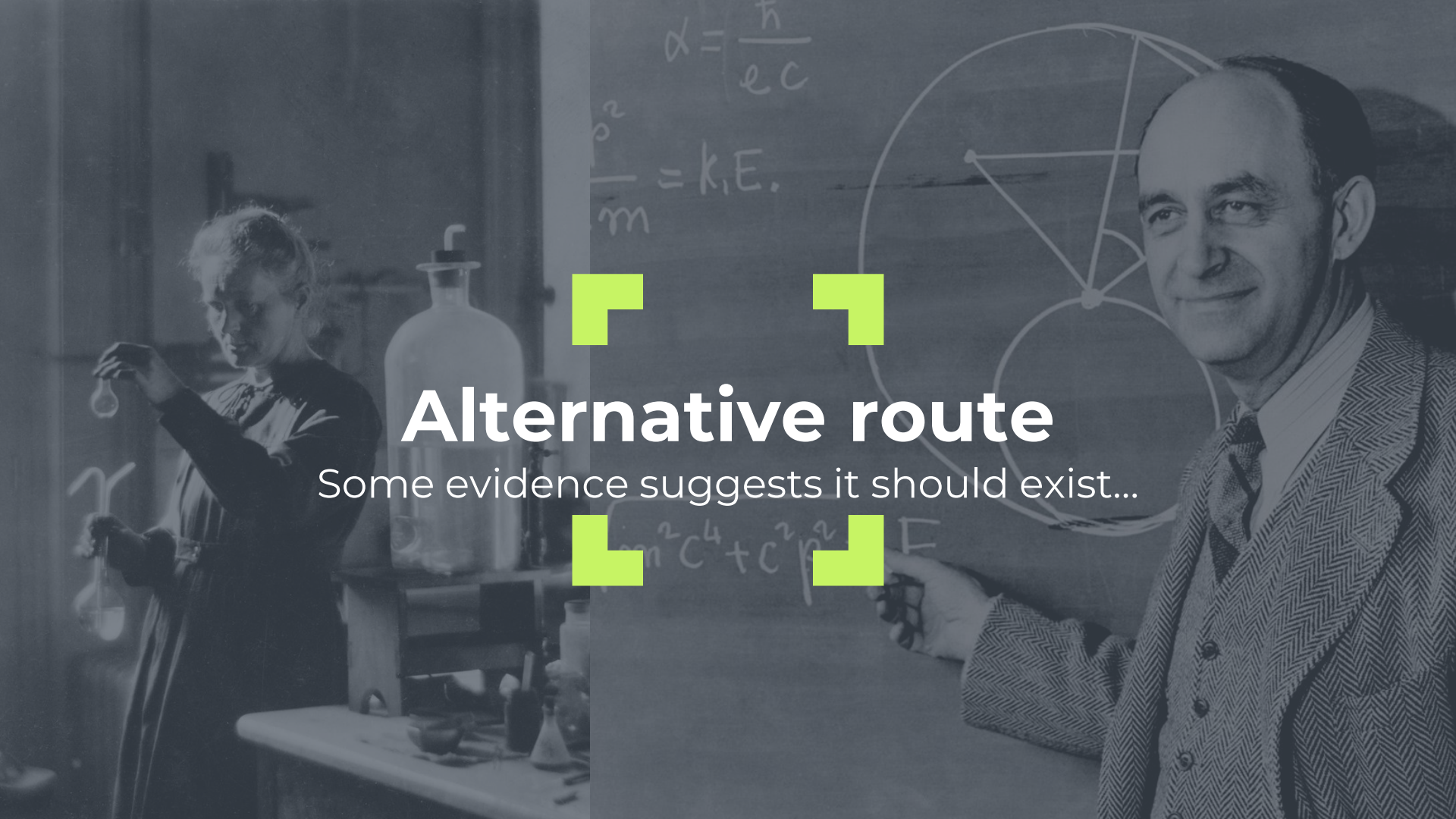
Brown Tom B., et al. "Language models are few-shot learners" (2020)

Alternative route?

Some evidence suggests it should exist...

Alternative route

Some evidence suggests it should exist...

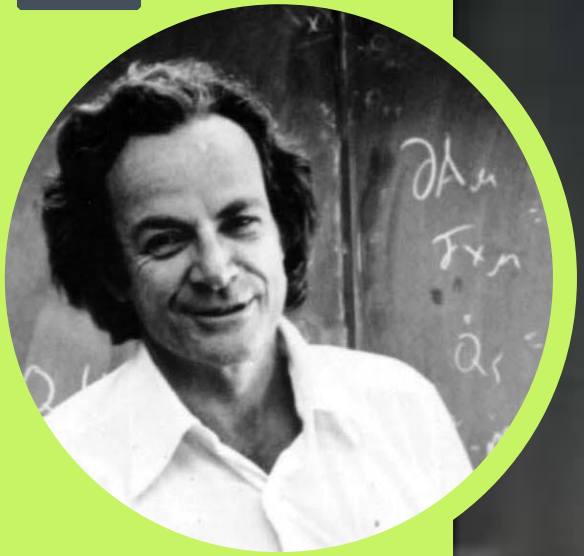


Alternative route

Some evidence suggests it should exist...

Alternative route

But how does a scientist work?



Feynman

Lecture on the scientific method, 1964

Nature has allowed us to discover how we have arrived at our present knowledge by the following process.

Don't Hastily address really true!

Data → Theory

Rationalist perspective shift

Theory → Data

Rationalist perspective shift

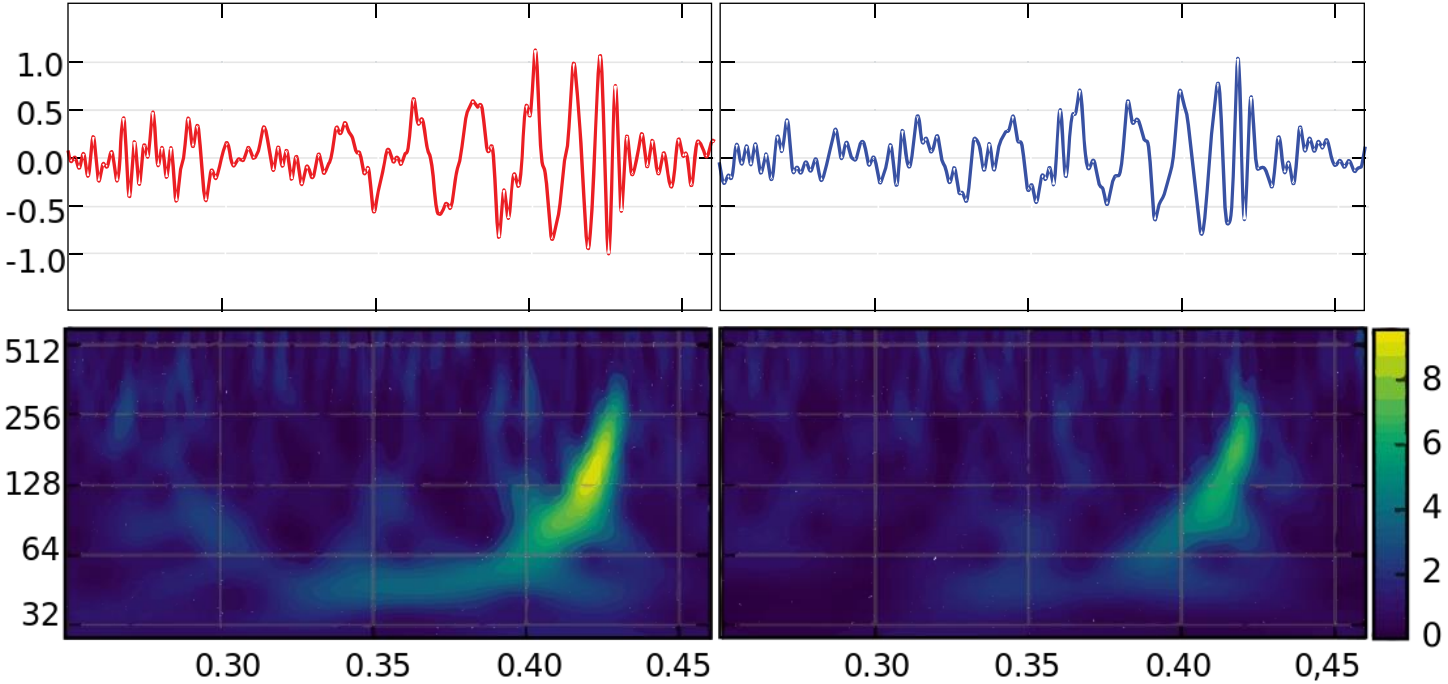
Conjectures → Data

Rationalist perspective shift

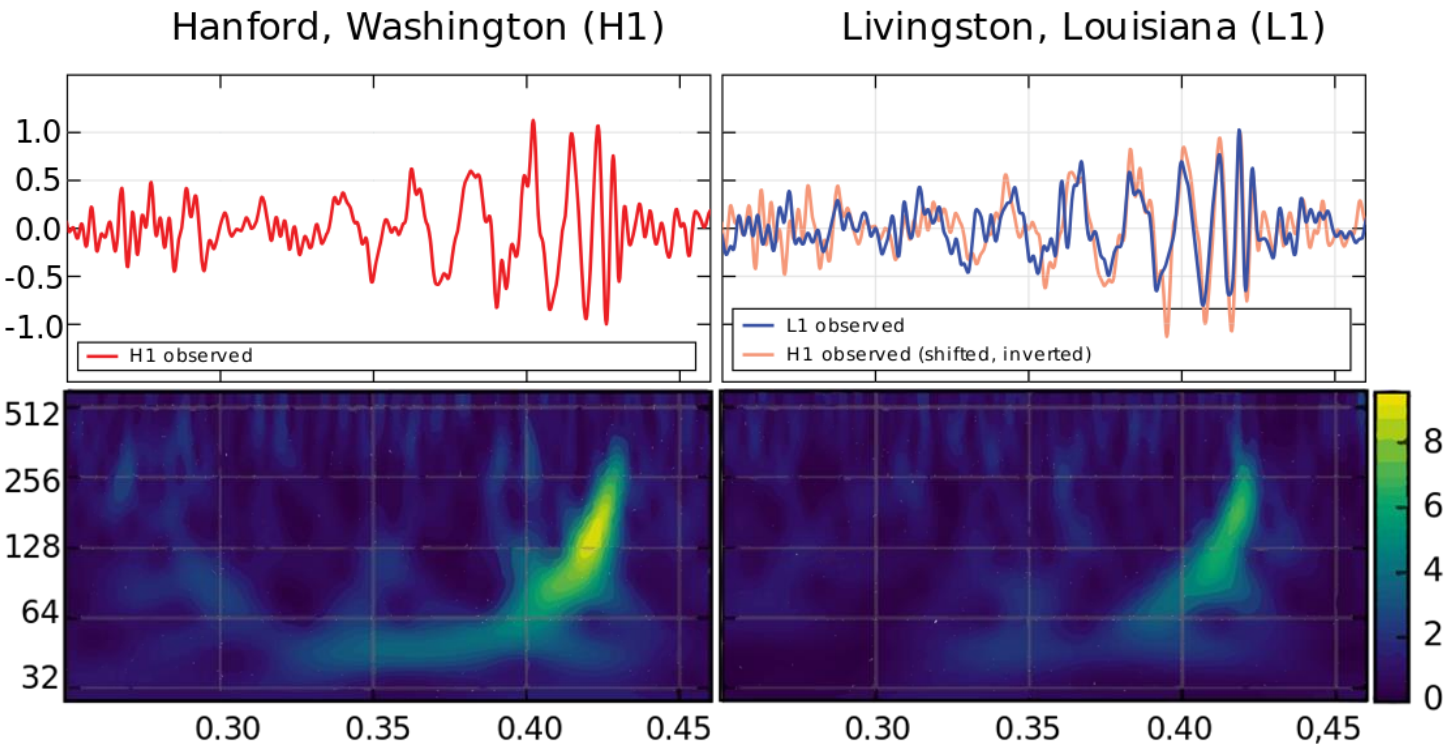
Falsifiable Conjectures → Data

Rationalist perspective shift

Data

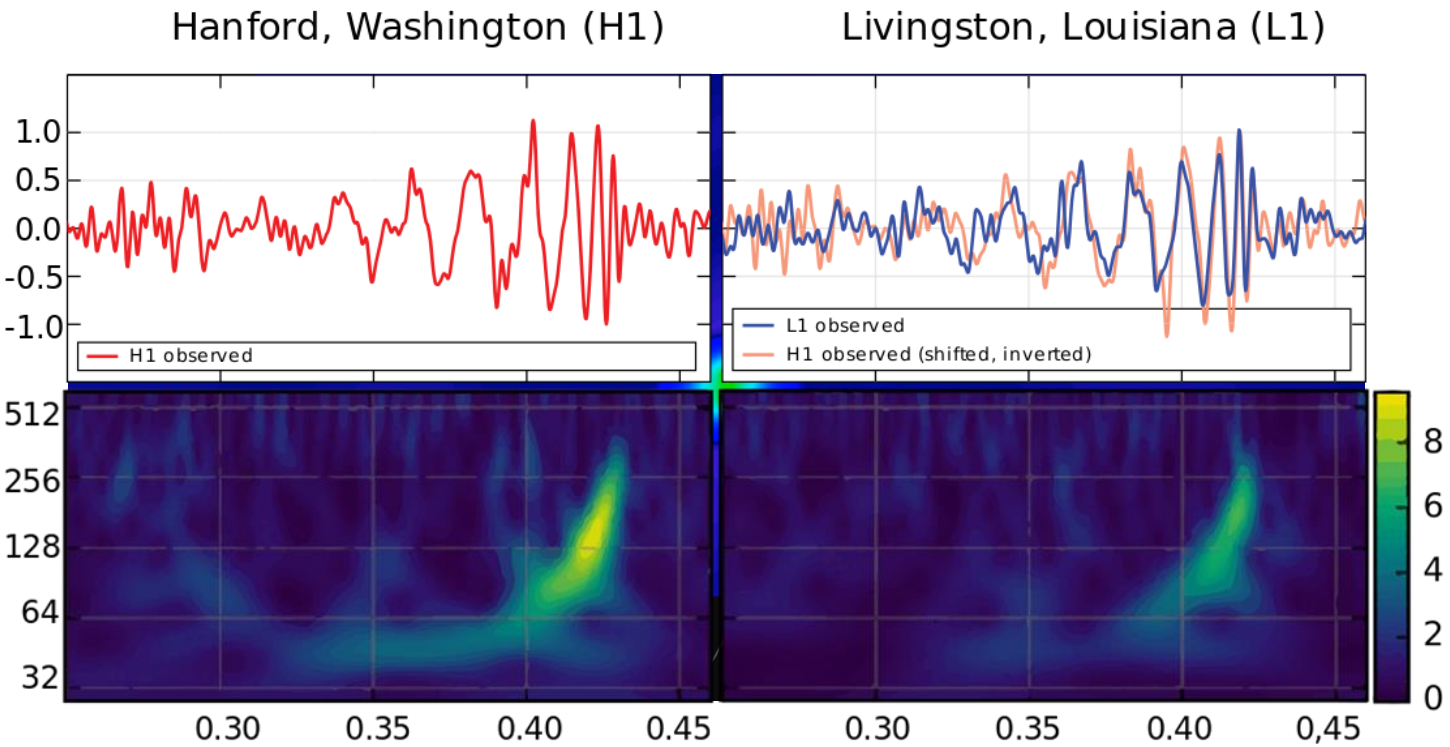

1.0
0.5
0.0
-0.5
-1.0
512
256
128
64
32
0.30
0.35
0.40
0.45
0.30
0.35
0.40
0,45
8
6
4
2
0

Data as theory-laden observations



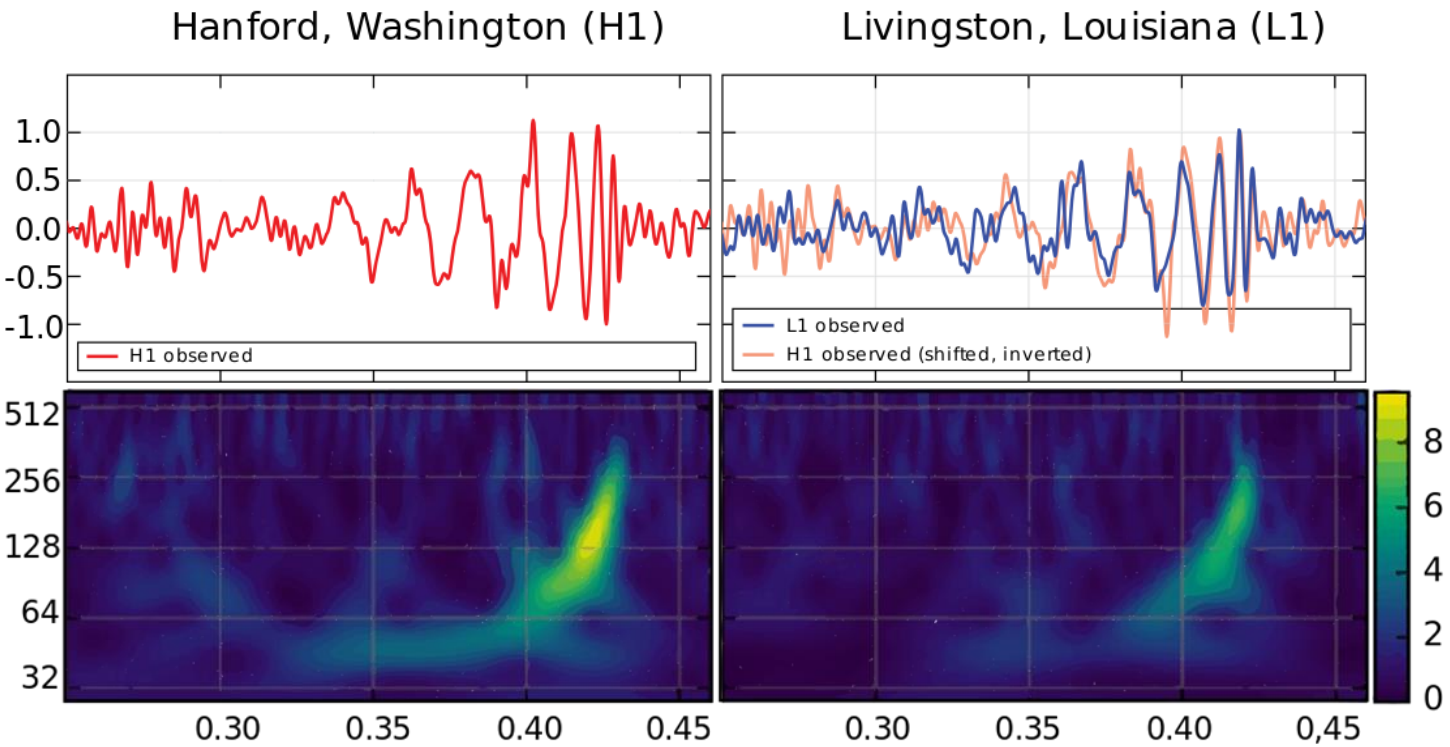
LIGO Scientific Collaboration and Virgo Collaboration, *"Observation of Gravitational Waves from a Binary Black Hole Merger" (2016)*

Data as theory-laden observations



LIGO Scientific Collaboration and Virgo Collaboration, *"Observation of Gravitational Waves from a Binary Black Hole Merger" (2016)*

Data as theory-laden observations



LIGO Scientific Collaboration and Virgo Collaboration, *"Observation of Gravitational Waves from a Binary Black Hole Merger" (2016)*

Data as theory-laden observations

THE FOUNDATION OF THE GENERAL THEORY OF RELATIVITY

By A. EINSTEIN

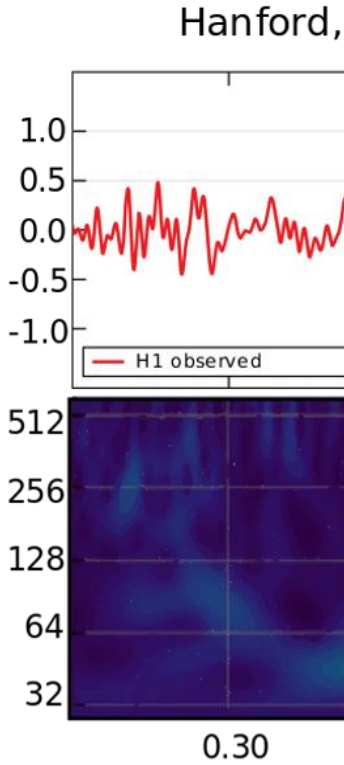
A. Fundamental Considerations on the Postulate of Relativity

§ 1. **Observations on the Special Theory of Relativity**

THE special theory of relativity is based on the following postulate, which is also satisfied by the mechanics of Galileo and Newton.

If a system of co-ordinates K is chosen so that, in relation to it, physical laws hold good in their simplest form, the *same* laws also hold good in relation to any other system of co-ordinates K′ moving in uniform translation relatively to K. This postulate we call the "special principle of relativity." The word "special" is meant to intimate

Albert Einstein, *"The Foundation of the General Theory of Relativity" (1916)*



LIGO Scientific Collabo

As AI researchers, what can we learn from this?

THE FOUNDATION OF THE GENERAL THEORY OF RELATIVITY

By A. EINSTEIN

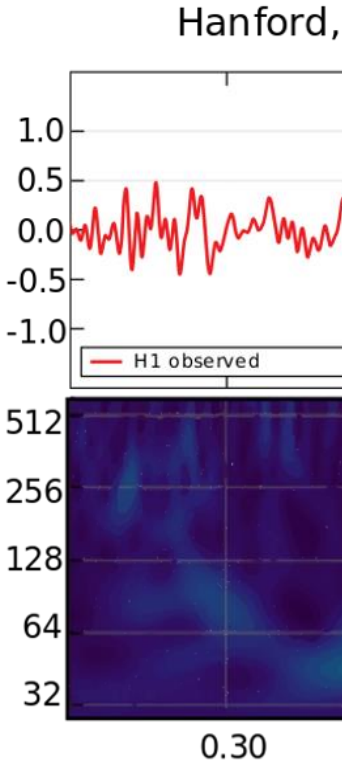
A. Fundamental Considerations on the Postulate of Relativity

§ 1. **Observations on the Special Theory of Relativity**

THE special theory of relativity is based on the following postulate, which is also satisfied by the mechanics of Galileo and Newton.

If a system of co-ordinates K is chosen so that, in relation to it, physical laws hold good in their simplest form, the *same* laws also hold good in relation to any other system of co-ordinates K′ moving in uniform translation relatively to K. This postulate we call the "special principle of relativity." The word "special" is meant to intimate

Albert Einstein, *"The Foundation of the General Theory of Relativity" (1916)*



LIGO Scientific Collabo

How can we build machines that creatively invent entirely new theories from data, like scientists do?

1. Prologue

2. Explanatory Learning

Giorgio Parisi
2021 Nobel laureate in Physics

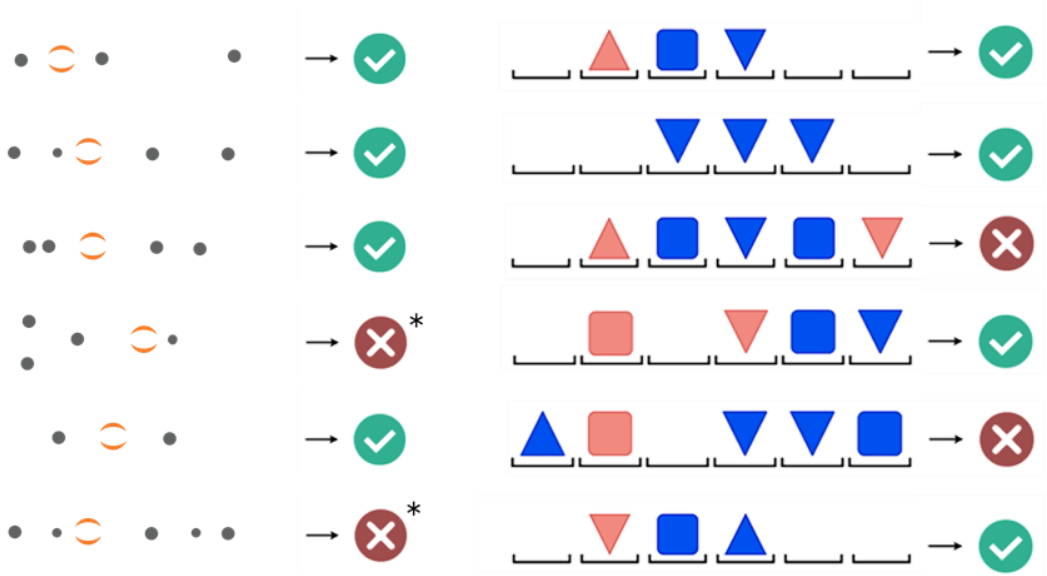
«Tanta gente passa il tempo a fare i puzzle, ecco, la ricerca è come mettere insieme dei pezzi che sembrano non essere connessi l'uno con l'altro e che se uno risolve diventano patrimonio dell'umanità»

Interview with Paolo Tarvisi, Il Messaggero, 15/02/2021

Scientific puzzles: Odeen

Sketches of Jupiter moons made by Galileo

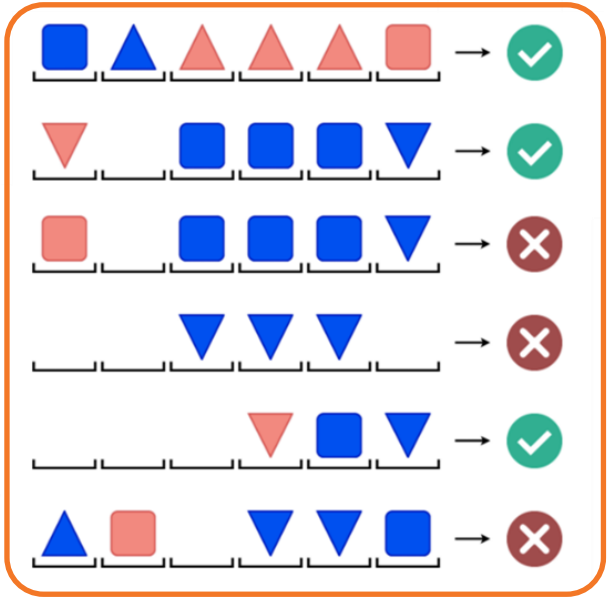
Four wandering stars having their period around a principal star



Observations of a phenomenon in Odeen

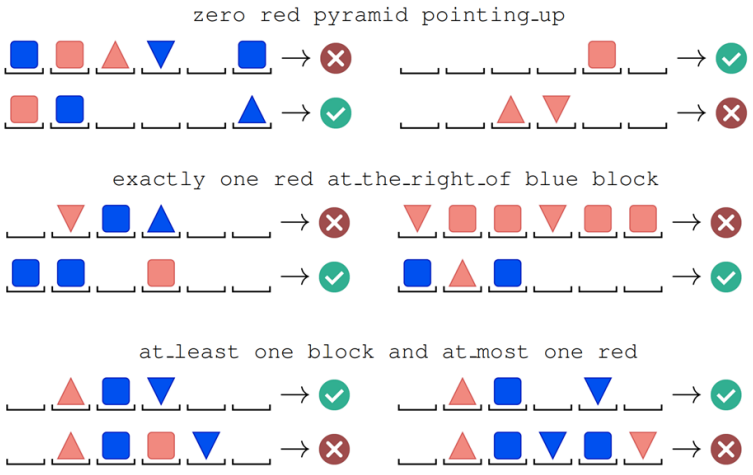
Exactly one blue surrounded by triangles

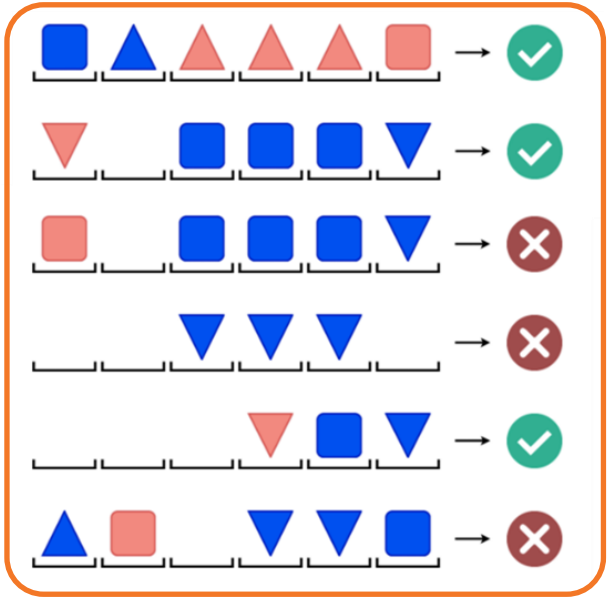
Scientific puzzles: Odeen



???

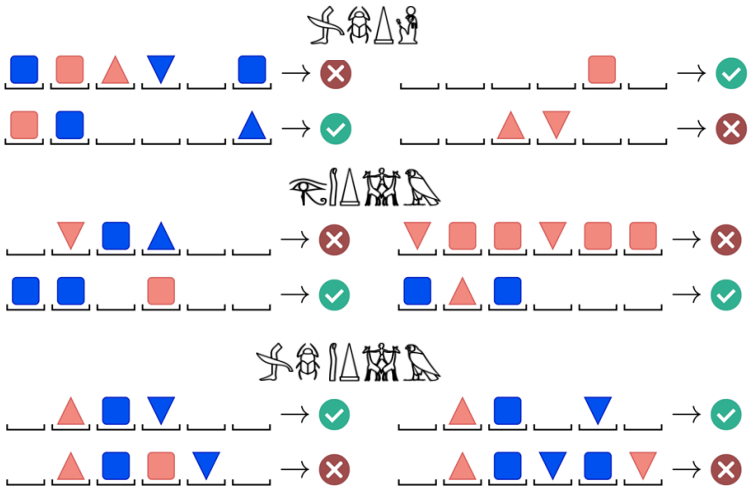
Scientific puzzles: Odeen

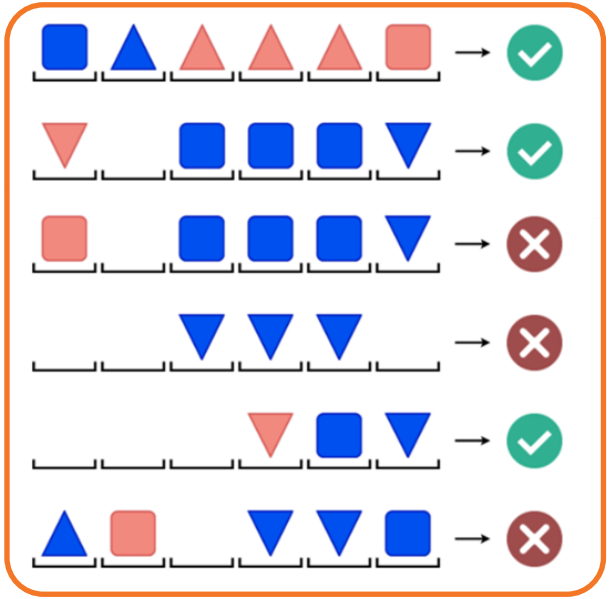




???

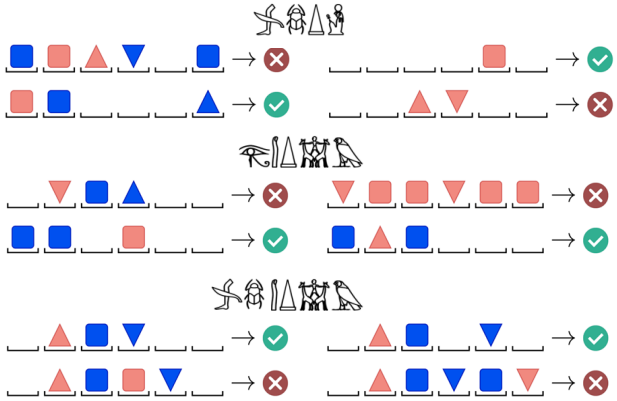
Scientific puzzles: Odeen

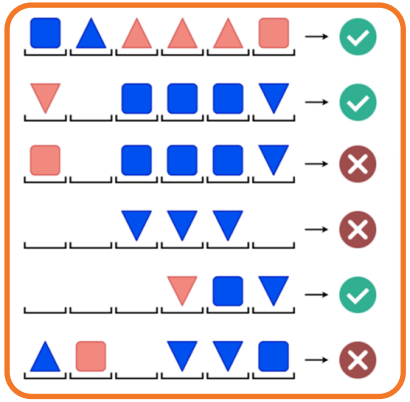


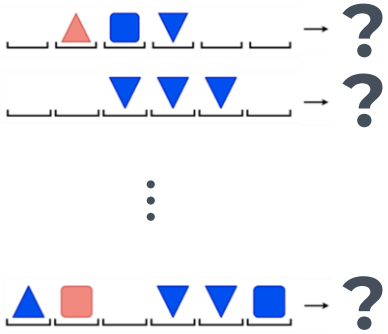


???

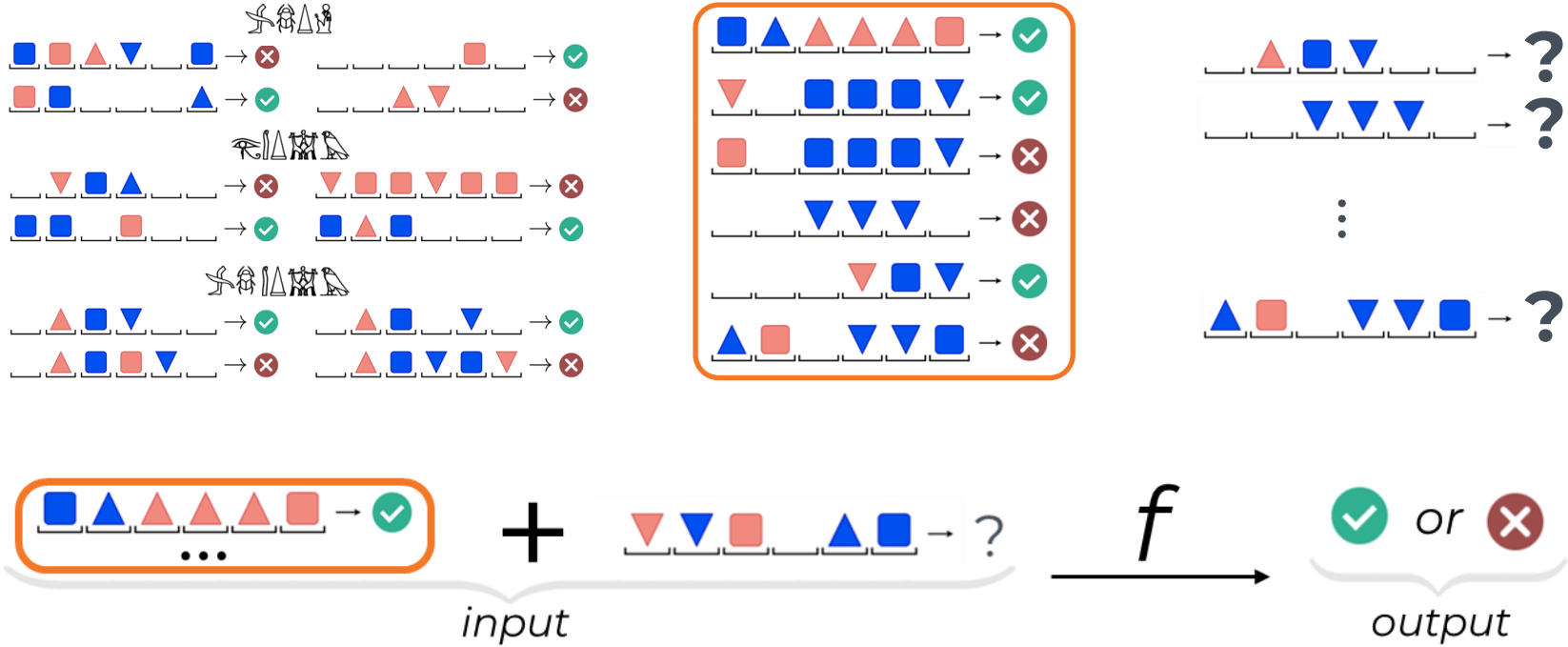
Scientific puzzles: Odeen





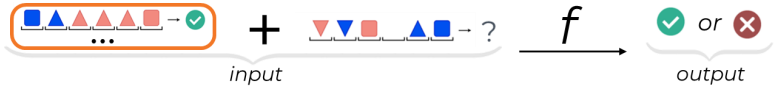


Explanatory Learning problem: find f



Empiricism: read the book of nature

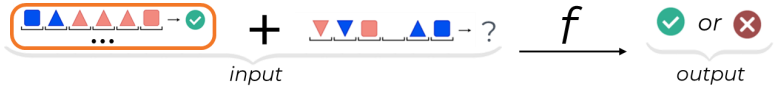
End-to-end deep learning is aligned with the empiricist view on the acquisition of knowledge.



Data → Theory

Empiricism: read the book of nature

End-to-end deep learning is aligned with the empiricist view on the acquisition of knowledge.

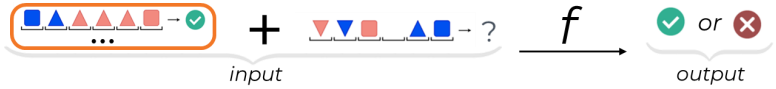


Data → Theory

Parametric hypothesis
continuously updated based on each new data sample

Empiricism: read the book of nature

End-to-end deep learning is aligned with the empiricist view on the acquisition of knowledge.



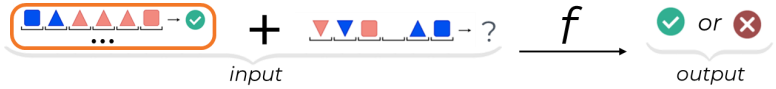
Data → Theory

Generalization gaps

Parametric hypothesis
continuously updated based on each new data sample

Empiricism: read the book of nature

End-to-end deep learning is aligned with the empiricist view on the acquisition of knowledge.



Data → Theory

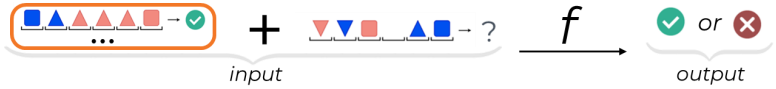
Generalization gaps

Unexplainable predictions

Parametric hypothesis
continuously updated based on each new data sample

Empiricism: read the book of nature

End-to-end deep learning is aligned with the empiricist view on the acquisition of knowledge.



Data → Theory

Generalization gaps

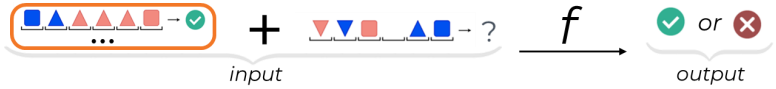
Unexplainable predictions

Unreliable predictions

Parametric hypothesis
continuously updated based on each new data sample

Empiricism: read the book of nature

End-to-end deep learning is aligned with the empiricist view on the acquisition of knowledge.



Data → Theory

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Parametric hypothesis *continuously updated based on each new data sample*

Rationalist perspective shift

Theory → **Data**

Data → Theory

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*

Parametric hypothesis *continuously updated based on each new data sample*

Rationalist perspective shift

Theory → **Data**

Conjecture

Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*

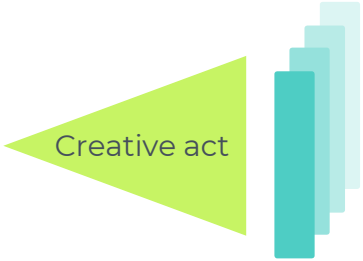
Data → Theory

Parametric hypothesis *continuously updated based on each new data sample*

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Rationalist perspective shift

Theory → **Data**



Conjectures

Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*

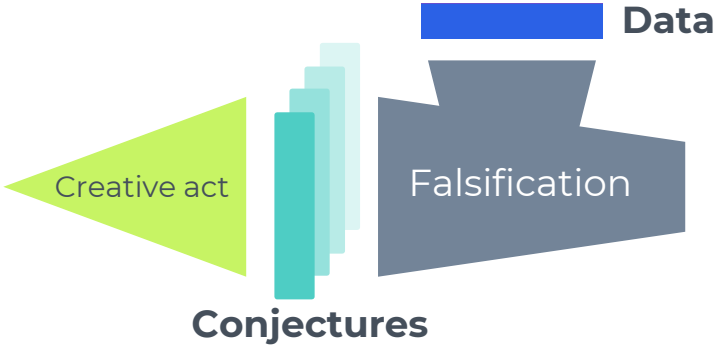
Data → Theory

Parametric hypothesis *continuously updated based on each new data sample*

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Rationalist perspective shift

Theory → **Data**



Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*

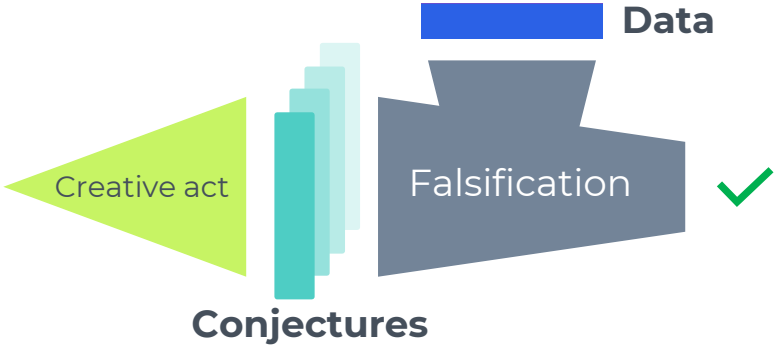
Data → Theory

Parametric hypothesis *continuously updated based on each new data sample*

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Rationalist perspective shift

Theory → **Data**



Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*

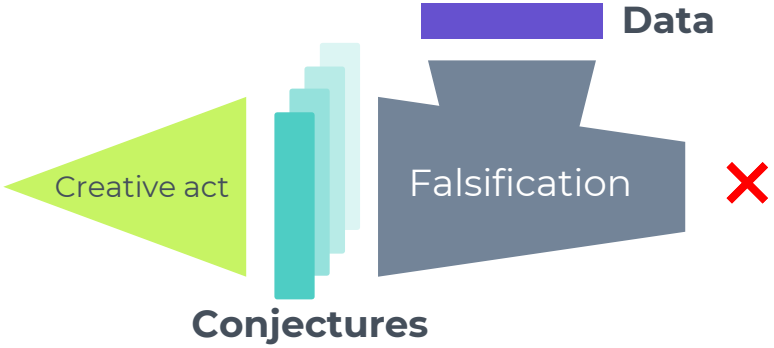
Data → Theory

Parametric hypothesis *continuously updated based on each new data sample*

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Rationalist perspective shift

Theory → **Data**



Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*

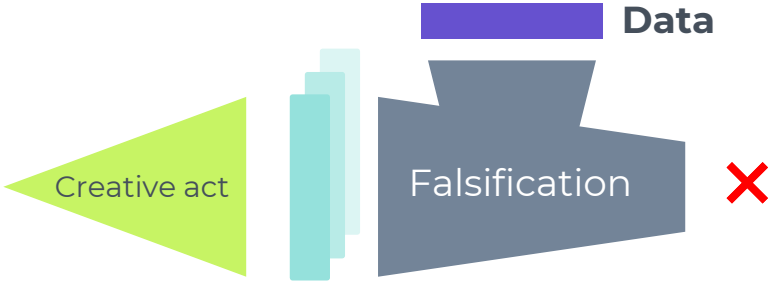
Data → Theory

Parametric hypothesis *continuously updated based on each new data sample*

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Rationalist perspective shift

Theory → **Data**



Conjectures

Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*

Data → Theory

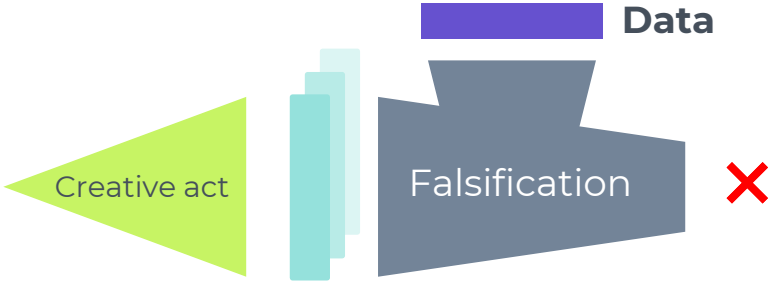
Parametric hypothesis *continuously updated based on each new data sample*

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Rationalist perspective shift

Theory → **Data**

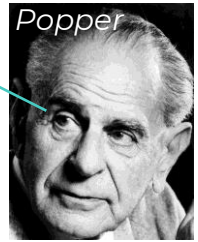
Data → Theory



Conjectures

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

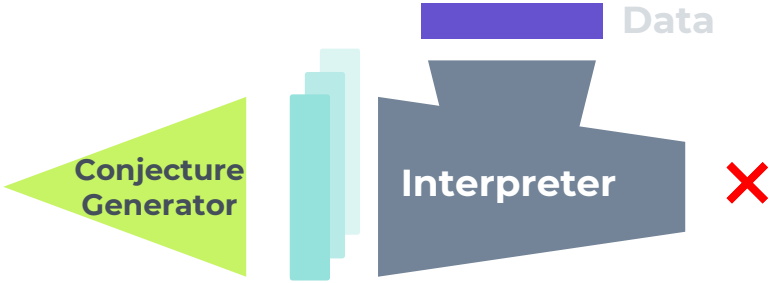
Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*



A scientist works like this!
Data as theory-laden observations, useful to falsify concurrent conjectures

Critical Rationalist Networks

Theory → **Data**



Conjectures

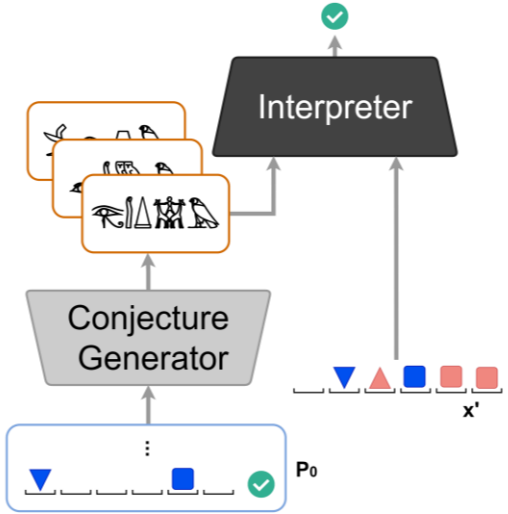
Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*

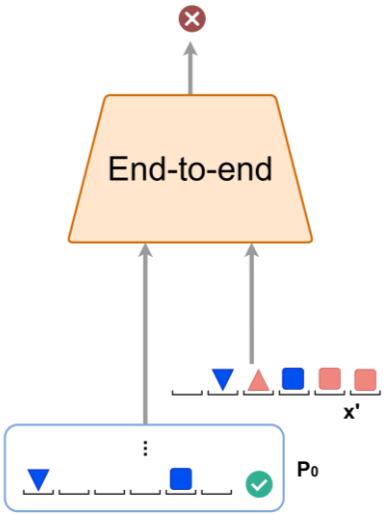
Data → Theory

Parametric hypothesis *continuously updated based on each new data sample*

- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

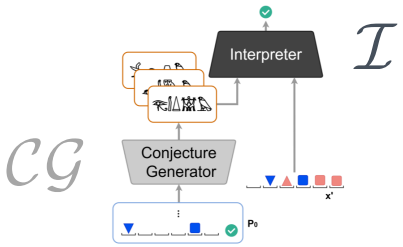
Critical Rationalist Networks

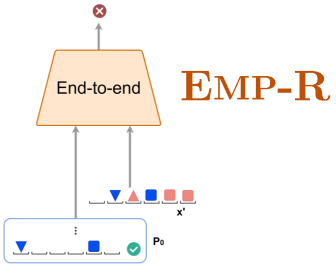


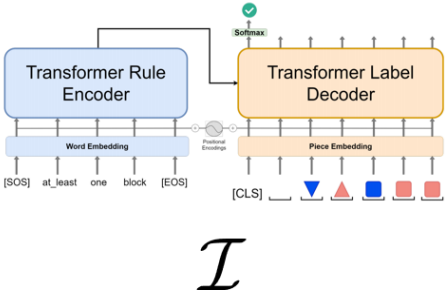


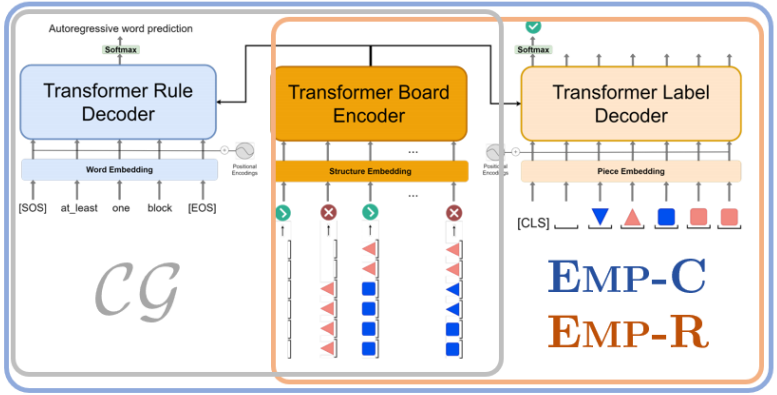
- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Critical Rationalist Networks









- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

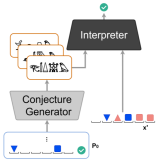
Results: CRNs vs empiricist models

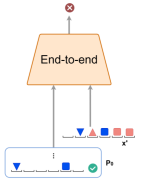
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time

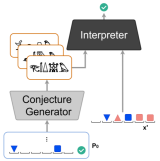




Results: CRNs vs empiricist models

The CRN can discover the correct explaantion of 777 out of 1000 new phenomena. Using the same data and ~ the same number of learnable parameters the empiricists do not go over 225.

MODEL	NRS	T-ACC	R-ACC
CRN	**0.777**	**0.980**	**0.737**
EMP-C	0.225	0.905	0.035
EMP-R	0.156	0.898	-

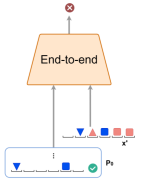


Generalization gaps

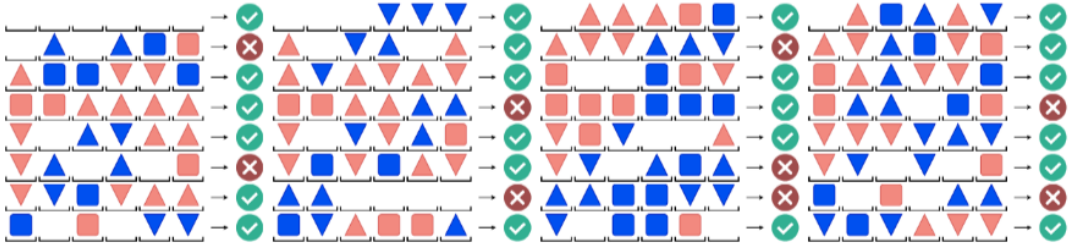
Unexplainable predictions

Unreliable predictions

Fixed thinking time



Results: CRNs vs empiricist models



Board 04
Golden Rule: “at_most 1 blue pyramid pointing_up”
CRN: “zero blue or at_most 1 blue pyramid pointing_up”; T-acc 1.0 ✓
EMP-C: “zero 1 blue touching or or”; T-acc: 0.89 ✗

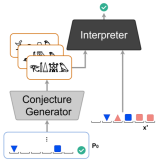
MODEL	NRS	T-ACC	R-ACC
CRN	**0.777**	**0.980**	**0.737**
EMP-C	0.225	0.905	0.035
EMP-R	0.156	0.898	-

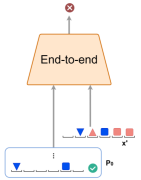
Generalization gaps

Unexplainable predictions

Unreliable predictions

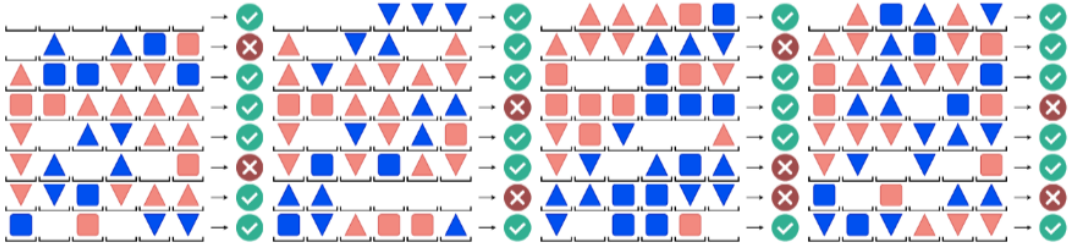
Fixed thinking time





Results: CRNs vs empiricist models

Generalization power



Board 04
Golden Rule: “at_most 1 blue pyramid pointing_up”
CRN: “zero blue or at_most 1 blue pyramid pointing_up”; T-acc 1.0 ✓
EMP-C: “zero 1 blue touching or or”; T-acc: 0.89 ✗

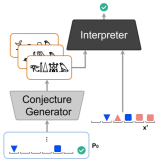
MODEL	NRS	T-ACC	R-ACC
CRN	**0.777**	**0.980**	**0.737**
EMP-C	0.225	0.905	0.035
EMP-R	0.156	0.898	-

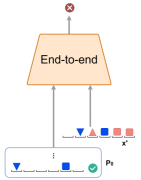
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time





Results: CRNs vs empiricist models

Generalization power

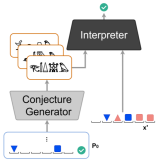
The bank ML algorithm spoke: "Loan denied"; explanation: "Two not paid loan in the past and resident in a district with a high rate of insolvents".
With a CRN, we can naturally discard this explanation and compute a new prediction for just "Two not paid loan in the past".

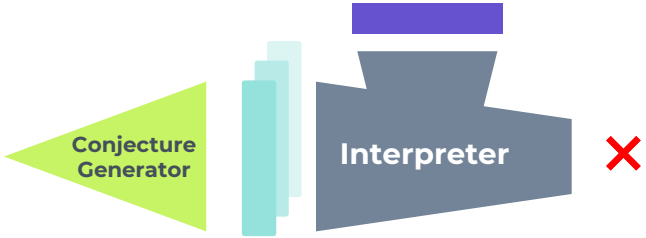
Generalization gaps

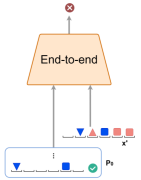
Unexplainable predictions

Unreliable predictions

Fixed thinking time







Results: CRNs vs empiricist models

Generalization power

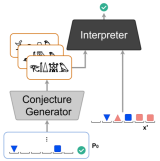
The bank ML algorithm spoke: "Loan denied"; explanation: "Two not paid loan in the past and resident in a district with a high rate of insolvents".
With a CRN, we can naturally discard this explanation and compute a new prediction for just "Two not paid loan in the past".

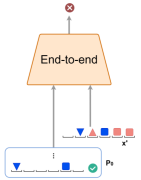
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time





Results: CRNs vs empiricist models

Generalization power

Truly Explainable predictions

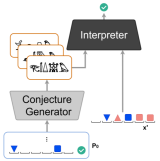
The bank ML algorithm spoke: "Loan denied"; explanation: "Two not paid loan in the past and resident in a district with a high rate of insolvents".
With a CRN, we can naturally discard this explanation and compute a new prediction for just "Two not paid loan in the past".

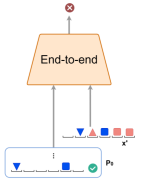
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time





Results: CRNs vs empiricist models

Generalization power

Truly Explainable predictions

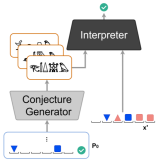
If no conjecture is compatible with data?
A CRN returns "unknown explanation" rather than a random prediction

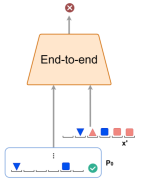
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time





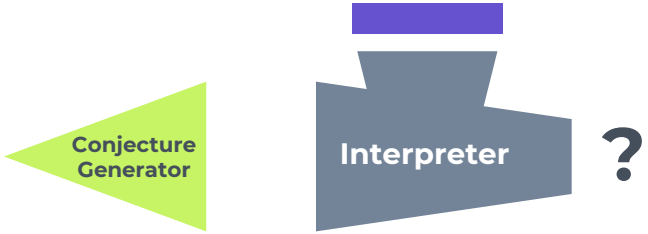
Results: CRNs vs empiricist models

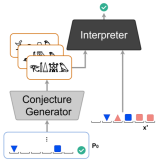
Generalization power

Truly Explainable predictions

If no conjecture is compatible with data?
A CRN returns "unknown explanation" rather than a random prediction

Model	Guesses	Unkn	Wrong
CRN	**0.760**	**0.240**	**0**
Emp-C	0.225	0	0.775
Emp-R	0.156	0	0.844



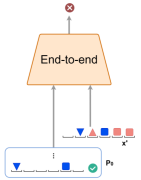


Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time



Results: CRNs vs empiricist models

Generalization power

Truly Explainable predictions

Reliable predictions

If no conjecture is compatible with data?
A CRN returns "unknown explanation" rather than a random prediction

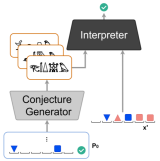
MODEL	GUESSES	UNKN	WRONG
CRN	**0.760**	**0.240**	**0**
EMP-C	0.225	0	0.775
EMP-R	0.156	0	0.844

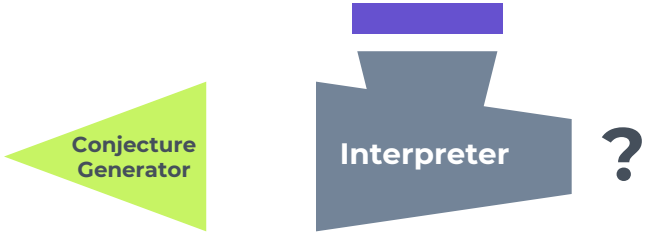
Generalization gaps

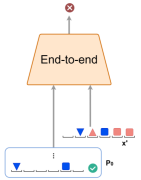
Unexplainable predictions

Unreliable predictions

Fixed thinking time







Results: CRNs vs empiricist models

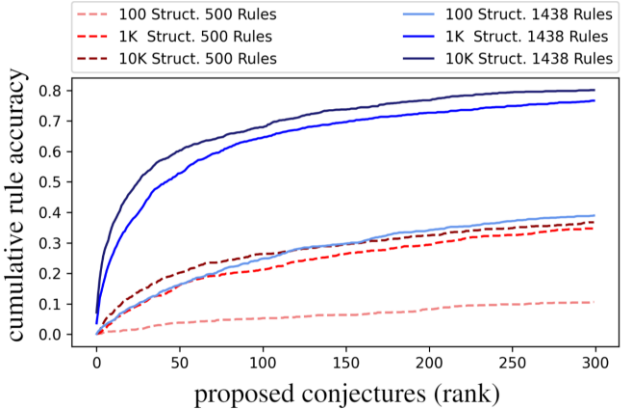
Generalization power

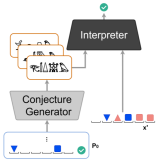
Truly Explainable predictions

Reliable predictions

CRNs exhibit a parameter at test time to adjust their processing to the complexity of the incoming prediction.

t = number of conjectures generated



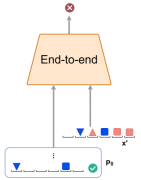


Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time



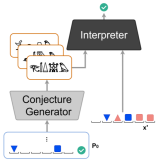
Results: CRNs vs empiricist models

Generalization power

Truly Explainable predictions

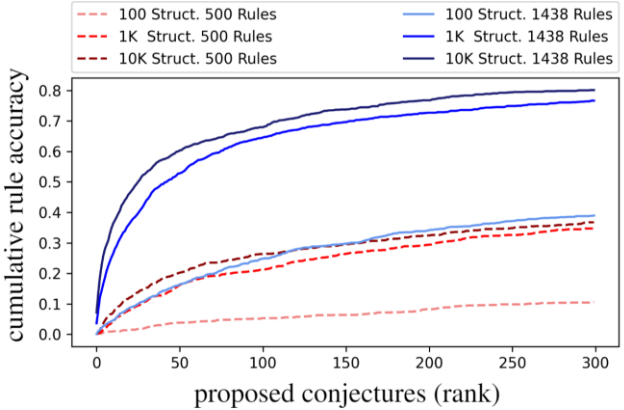
Reliable predictions

Adjustable thinking time



CRNs exhibit a parameter at test time to adjust their processing to the complexity of the incoming prediction.

t = number of conjectures generated

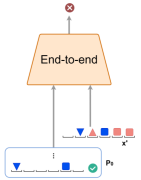


Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time



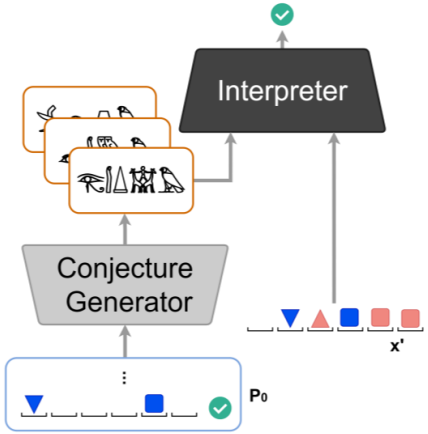
Results: CRNs vs empiricist models

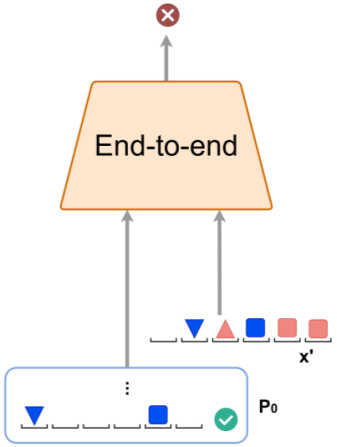
Generalization power

Truly Explainable predictions

Reliable predictions

Adjustable thinking time





Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time

Antonio Norelli, Giorgio Mariani, Luca Moschella, Andrea Santilli, Giambattista Parascandolo, Simone Melzi, Emanuele Rodolà

"Explanatory Learning: Beyond Empiricism in Neural Networks" under review

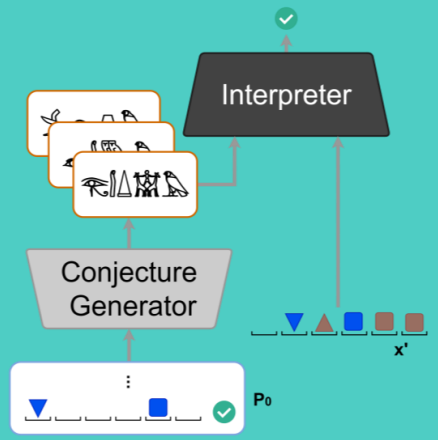
Thanks!

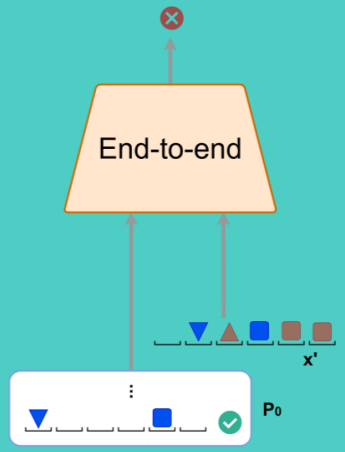
Generalization power

Truly Explainable predictions

Reliable predictions

Adjustable thinking time





Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time

Antonio Norelli - norelli@di.uniroma1.it