Towards an Artificial Scientist

Intelligence as the ability to build functioning models of the world

Antonio Norelli

III y. PhD student in CS at GLADIA lab, Sapienza - Rome, IT
Currently Applied Scientist Intern at Amazon - Tubingen, DE
norelli@di.uniroma1.it

Dipartimento di Informatica
Sapienza Università di Roma

GLADIA

Plan

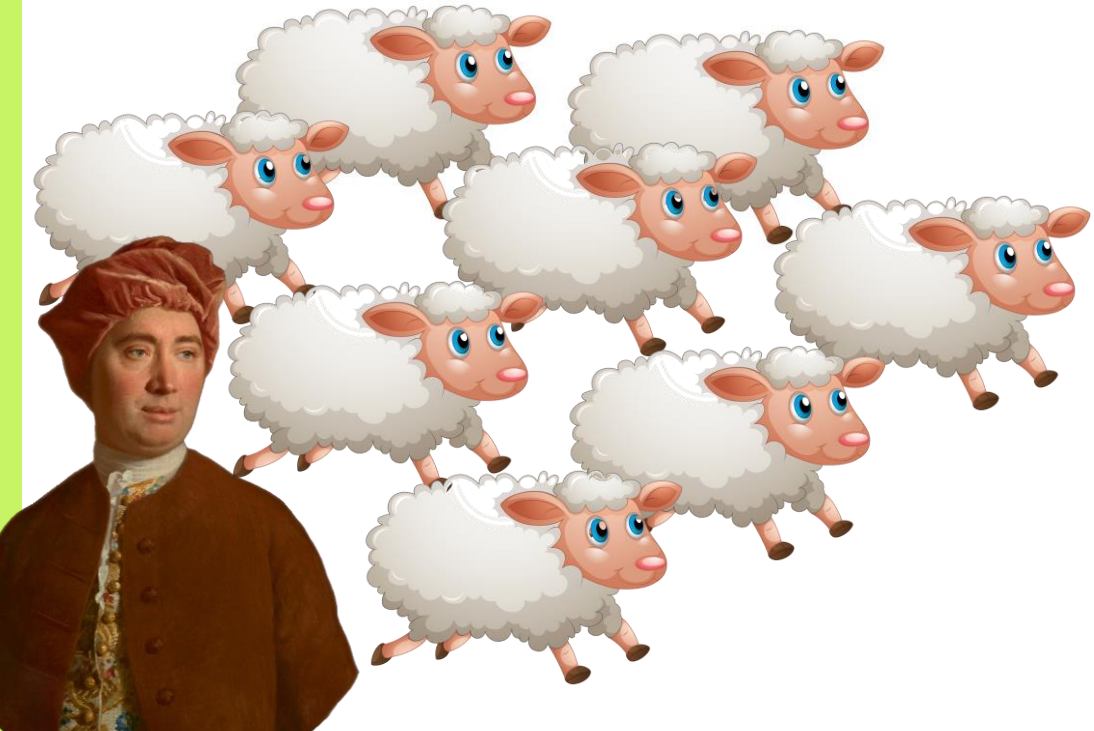
- Prologue
  - Science is possible? A theoretical result
  - Three key ingredients for [a scientific] AI
- **Connect**: *Explanatory Learning (2022)*
- **Reduce**: *LIMP (2020)*
- **Search**: *OLIVAW (2021)*

Immanuel Kant

«All metaphysicians (AI researchers) *are therefore solemnly and legally suspended from their occupations till they shall have answered in a satisfactory manner the question, how are synthetic cognitions a priori possible* (is it possible to learn from observations)*?»*

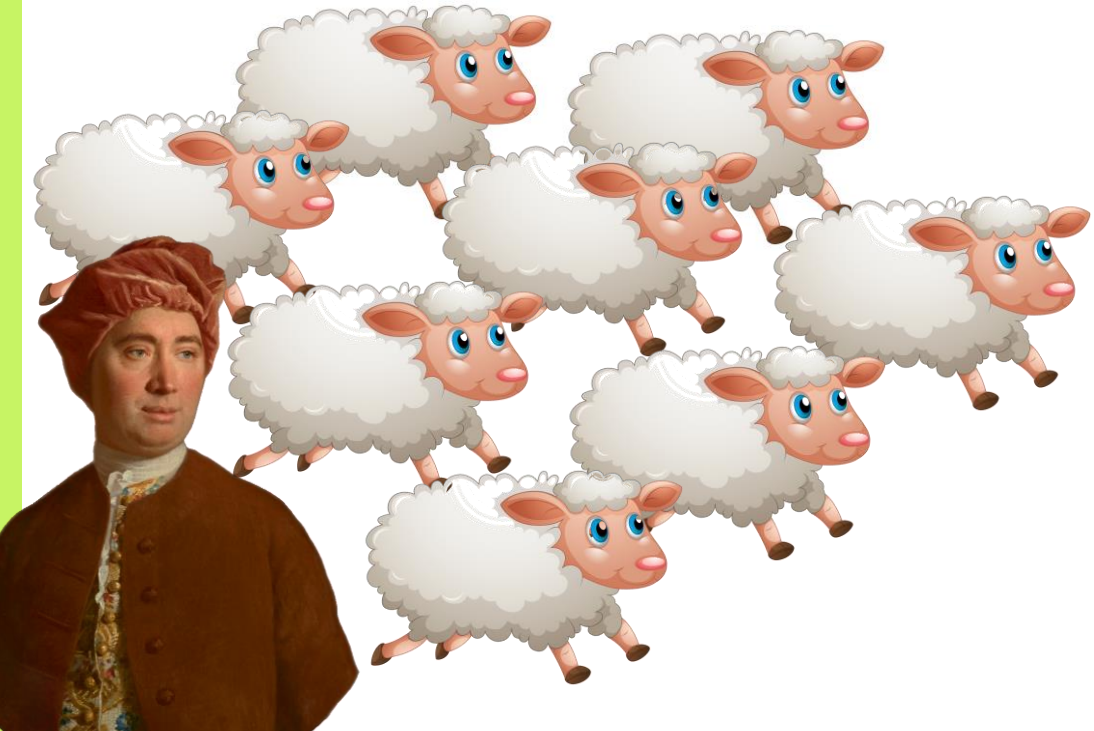
Prolegomena to any future metaphysics, 1783

Hume: Learning from the past is logically impossible



Will the 1001st sheep be white?

Hume: Learning from the past is logically impossible



Will the 1001st sheep be white?

A solution: Occam’s razor

Simpler hypotheses consistent with the data are more likely to be correct.

All sheeps are white

Most sheeps are black or lilla, only the 1000 we happened to see were white

Occam’s razor

Simpler hypotheses consistent with the data are more likely to be correct. But:

- What wo we mean by *simpler*?

Occam’s razor

Simpler hypotheses consistent with the data are more likely to be correct. But:

- What wo we mean by *simpler*?
- What properties must reality have for Occam's Razor to work?

Occam’s razor

Simpler hypotheses consistent with the data are more likely to be correct. But:

- What wo we mean by *simpler*?
- What properties must reality have for Occam's Razor to work?
- How much data is needed to find a predictive simple hypothesis ?

Occam’s razor

Simpler hypotheses consistent with the data are more likely to be correct. But:

- What wo we mean by *simpler*?
- What properties must reality have for Occam's Razor to work?
- How much data is needed to find a predictive simple hypothesis ?
- How do we go about finding it?

Central result of PAC learning

Theorem *Consider a finite hypothesis class* $\mathcal{H}$*, a Boolean function* $f : \mathcal{S} \to \{0,1\}$ *in* $\mathcal{H}$*, and a sample distribution* $\mathcal{D}$ *over* $\mathcal{S}$*, as well as an error rate* $\varepsilon > 0$ *and failure probability* $\delta > 0$ *that the learner is willing to tolerate. Call a hypothesis* $h : \mathcal{S} \to \{0,1\}$ *"good" if*

$$\Pr_{x \sim \mathcal{D}} [h(x) = f(x)] \geq 1 - \varepsilon.$$

Also, call sample points $x_1, \ldots, x_m$ *"reliable" if any hypothesis* $h \in \mathcal{H}$ *that satisfies* $h(x_i) = f(x_i)$ *for all* $i \in \{1, \ldots, m\}$ *is good. Then*

$$m = \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

sample points $x_1, \ldots, x_m$ *drawn independently from* $\mathcal{D}$ *will be reliable with probability at least* $1 - \delta$.

L. G. Valiant. "A theory of the learnable" Communications of the ACM, 27 (1984)

Aaronson, Scott. "Why philosophers should care about computational complexity." Computability: Turing, Gödel, Church, and Beyond (2013)

Theorem *Consider a finite hypothesis class* $\mathcal{H}$*, a Boolean function* $f : \mathcal{S} \to \{0, 1\}$ *in* $\mathcal{H}$*, and a sample distribution* $\mathcal{D}$ *over* $\mathcal{S}$*, as well as an error rate* $\varepsilon > 0$ *and failure probability* $\delta > 0$ *that the learner is willing to tolerate. Call a hypothesis* $h : \mathcal{S} \to \{0, 1\}$ *"good" if*

$$\Pr_{x \sim \mathcal{D}} [h(x) = f(x)] \geq 1 - \varepsilon.$$

Also, call sample points $x_1, \ldots, x_m$ *"reliable" if any hypothesis* $h \in \mathcal{H}$ *that satisfies* $h(x_i) = f(x_i)$ *for all* $i \in \{1, \ldots, m\}$ *is good. Then*

$$m = \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

sample points $x_1, \ldots, x_m$ *drawn independently from* $\mathcal{D}$ *will be reliable with probability at least* $1 - \delta$.

- What wo we mean by *simpler*? ✓

Theorem *Consider a finite hypothesis class $\mathcal{H}$, a Boolean function $f : \mathcal{S} \to \{0,1\}$ in $\mathcal{H}$, and a sample distribution $\mathcal{D}$ over $\mathcal{S}$, as well as an error rate $\varepsilon > 0$ and failure probability $\delta > 0$ that the learner is willing to tolerate. Call a hypothesis $h : \mathcal{S} \to \{0,1\}$ "good" if*

$$\Pr_{x \sim \mathcal{D}} [h(x) = f(x)] \geq 1 - \varepsilon.$$

Also, call sample points $x_1, \ldots, x_m$ "reliable" if any hypothesis $h \in \mathcal{H}$ that satisfies $h(x_i) = f(x_i)$ for all $i \in \{1, \ldots, m\}$ is good. Then

$$m = \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

sample points $x_1, \ldots, x_m$ drawn independently from $\mathcal{D}$ will be reliable with probability at least $1 - \delta$.

- What wo we mean by *simpler*? ✓
- What properties must reality have for Occam's Razor to work? ✓

Theorem *Consider a finite hypothesis class $\mathcal{H}$, a Boolean function $f : \mathcal{S} \to \{0,1\}$ in $\mathcal{H}$, and a sample distribution $\mathcal{D}$ over $\mathcal{S}$, as well as an error rate $\varepsilon > 0$ and failure probability $\delta > 0$ that the learner is willing to tolerate. Call a hypothesis $h : \mathcal{S} \to \{0,1\}$ "good" if*

$$\Pr_{x \sim \mathcal{D}} [h(x) = f(x)] \geq 1 - \varepsilon.$$

Also, call sample points $x_1, \ldots, x_m$ "reliable" if any hypothesis $h \in \mathcal{H}$ that satisfies $h(x_i) = f(x_i)$ for all $i \in \{1, \ldots, m\}$ is good. Then

$$m = \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

sample points $x_1, \ldots, x_m$ drawn independently from $\mathcal{D}$ will be reliable with probability at least $1 - \delta$.

- What wo we mean by *simpler*? ✓
- What properties must reality have for Occam's Razor to work? ✓
- How much data is needed to find a predictive simple hypothesis ? ✓

Theorem *Consider a finite hypothesis class $\mathcal{H}$, a Boolean function $f : \mathcal{S} \to \{0,1\}$ in $\mathcal{H}$, and a sample distribution $\mathcal{D}$ over $\mathcal{S}$, as well as an error rate $\varepsilon > 0$ and failure probability $\delta > 0$ that the learner is willing to tolerate. Call a hypothesis $h : \mathcal{S} \to \{0,1\}$ "good" if*

$$\Pr_{x \sim \mathcal{D}} [h(x) = f(x)] \geq 1 - \varepsilon.$$

Also, call sample points $x_1, \ldots, x_m$ "reliable" if any hypothesis $h \in \mathcal{H}$ that satisfies $h(x_i) = f(x_i)$ for all $i \in \{1, \ldots, m\}$ is good. Then

$$m = \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

sample points $x_1, \ldots, x_m$ drawn independently from $\mathcal{D}$ will be reliable with probability at least $1 - \delta$.

- What wo we mean by *simpler*? ✓
- What properties must reality have for Occam's Razor to work? ✓
- How much data is needed to find a predictive simple hypothesis ? ✓
- How do we go about finding it? ✗

Theorem *Consider a finite hypothesis class $\mathcal{H}$, a Boolean function $f : \mathcal{S} \to \{0,1\}$ in $\mathcal{H}$, and a sample distribution $\mathcal{D}$ over $\mathcal{S}$, as well as an error rate $\varepsilon > 0$ and failure probability $\delta > 0$ that the learner is willing to tolerate. Call a hypothesis $h : \mathcal{S} \to \{0,1\}$ "good" if*

$$\Pr_{x \sim \mathcal{D}} [h(x) = f(x)] \geq 1 - \varepsilon.$$

Also, call sample points $x_1, \ldots, x_m$ "reliable" if any hypothesis $h \in \mathcal{H}$ that satisfies $h(x_i) = f(x_i)$ for all $i \in \{1, \ldots, m\}$ is good. Then

$$m = \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

sample points $x_1, \ldots, x_m$ drawn independently from $\mathcal{D}$ will be reliable with probability at least $1 - \delta$.

- What wo we mean by *simpler*? ✓
- What properties must reality have for Occam's Razor to work? ✓
- How much data is needed to find a predictive simple hypothesis ? ✓
- How do we go about finding it? ✗ **Scientist**

Theorem *Consider a finite hypothesis class $\mathcal{H}$, a Boolean function $f : \mathcal{S} \to \{0, 1\}$ in $\mathcal{H}$, and a sample distribution $\mathcal{D}$ over $\mathcal{S}$, as well as an error rate $\varepsilon > 0$ and failure probability $\delta > 0$ that the learner is willing to tolerate. Call a hypothesis $h : \mathcal{S} \to \{0, 1\}$ "good" if*

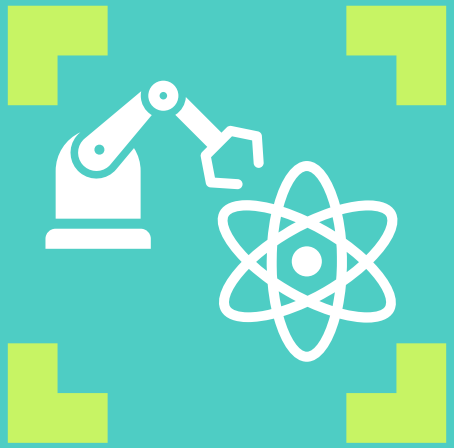
$$\Pr_{x \sim \mathcal{D}} [h(x) = f(x)] \geq 1 - \varepsilon.$$

Also, call sample points $x_1, \ldots, x_m$ "reliable" if any hypothesis $h \in \mathcal{H}$ that satisfies $h(x_i) = f(x_i)$ for all $i \in \{1, \ldots, m\}$ is good. Then

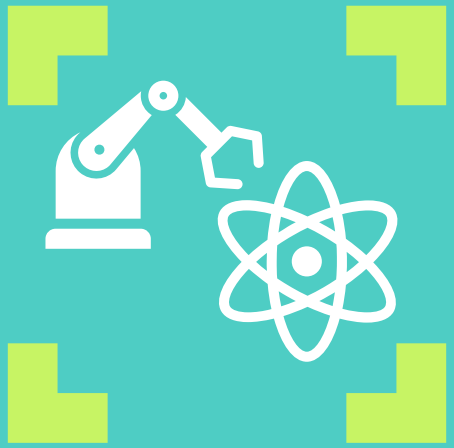
$$m = \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

sample points $x_1, \ldots, x_m$ drawn independently from $\mathcal{D}$ will be reliable with probability at least $1 - \delta$.

- What wo we mean by *simpler*? ✓
- What properties must reality have for Occam's Razor to work? ✓
- How much data is needed to find a predictive simple hypothesis ? ✓
- How do we go about finding it? ✗ **AI research!**



AI as Automatic Science



AI as Automatic Science

Three key ingredients

Reduce
the set of admissible hypotheses

Search
the best hypothesis

Connect
it to the real world

Reduce
the set of admissible hypotheses

Search
the best hypothesis

Symbolic AI
(1960-2000)

Reduce

Deep Learning (2010-2020)

Connect

it to the real world

Reduce
the set of admissible hypotheses

Search
the best hypothesis

Today trend (2020-?)

Connect
it to the real world

Reduce

LIMP: Learning Latent Shape Representations with Metric Preservation Priors. *ECCV, 2020*

Search

OLIVAW: Mastering Othello without Human Knowledge, nor a Penny. *IEEE Transactions on Games, 2021*

Connect

Explanatory Learning: Beyond Empiricism in Neural Networks. *Under review, arXiv, 2022*

1.

Connect

hypotheses to the real world

Explanatory Learning: Beyond Empiricism in Neural Networks

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

Giorgio Parisi

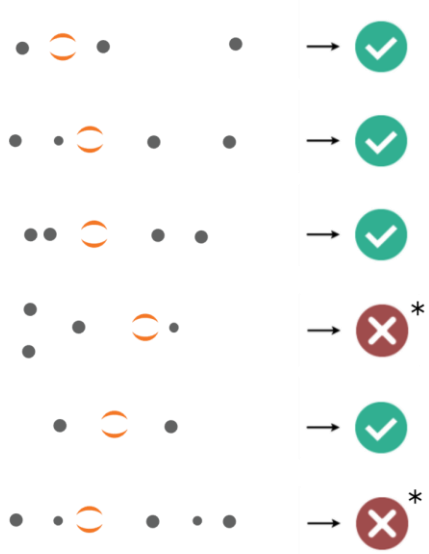
2021 Nobel laureate in Physics

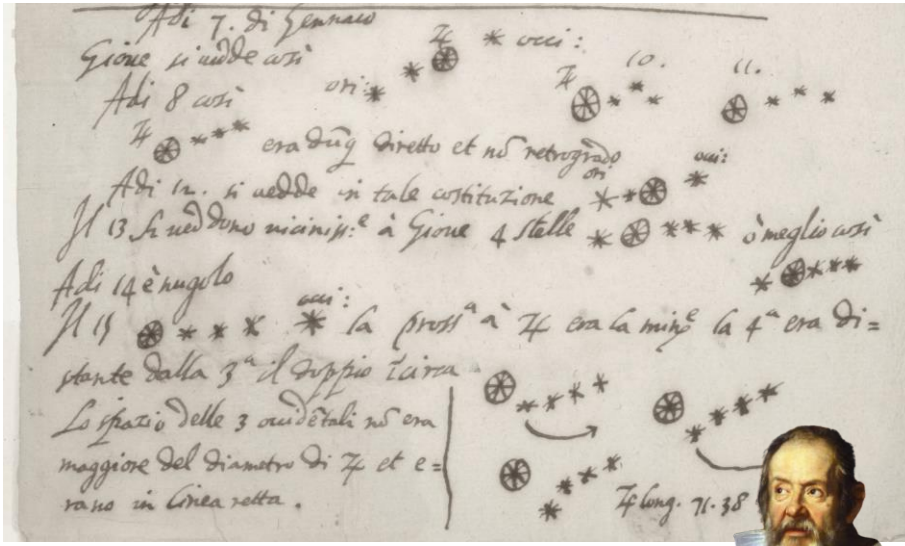
«So many people spend their time doing puzzles, that's it, research is like putting together pieces that do not seem to be linked with each other. Except that if one solves them, they become heritage of mankind.»

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

Scientific puzzles

Sketches of Jupiter moons made by Galileo

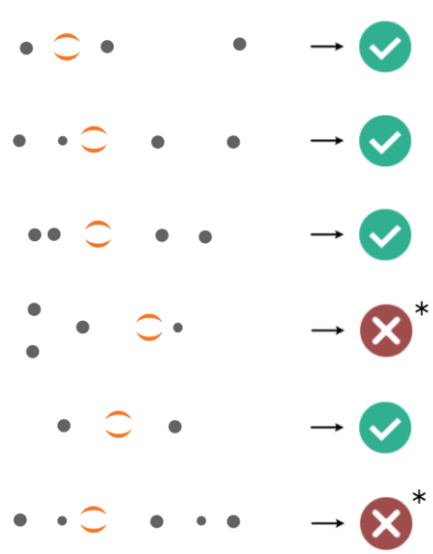


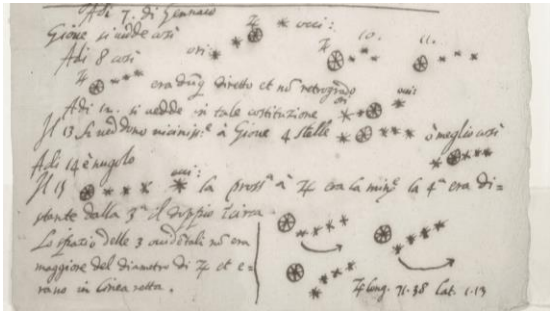


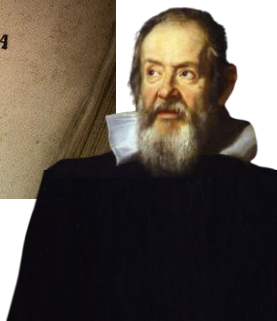
Scientific puzzles: finding an explanation

Sketches of Jupiter moons made by Galileo

Four wandering stars having their period around a principal star

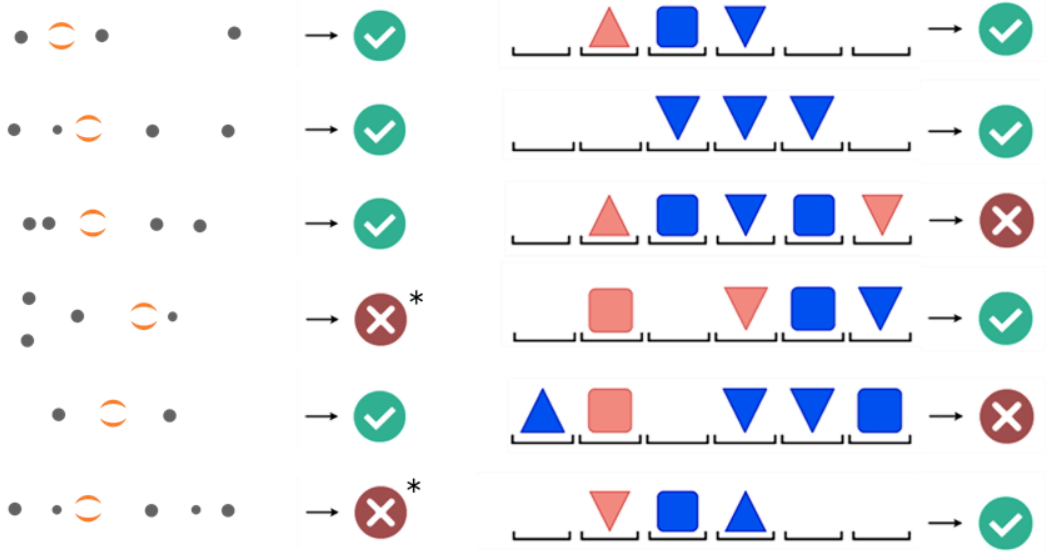






Scientific puzzles: Odeen

Sketches of Jupiter moons made by Galileo

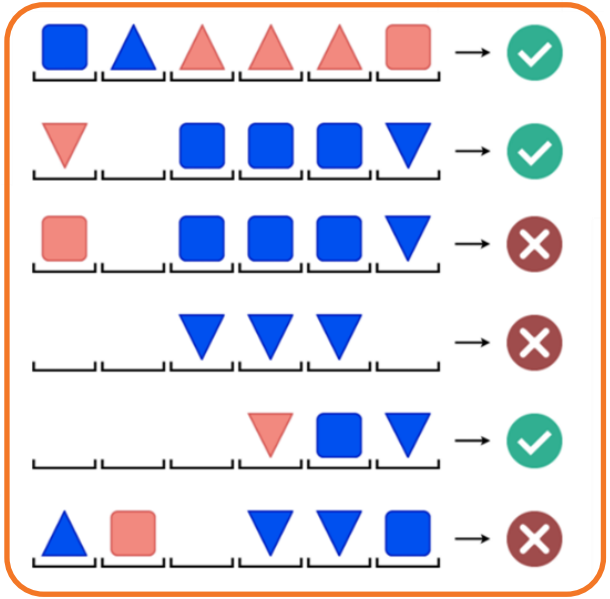


Observations of a phenomenon in Odeen

Four wandering stars having their period around a principal star

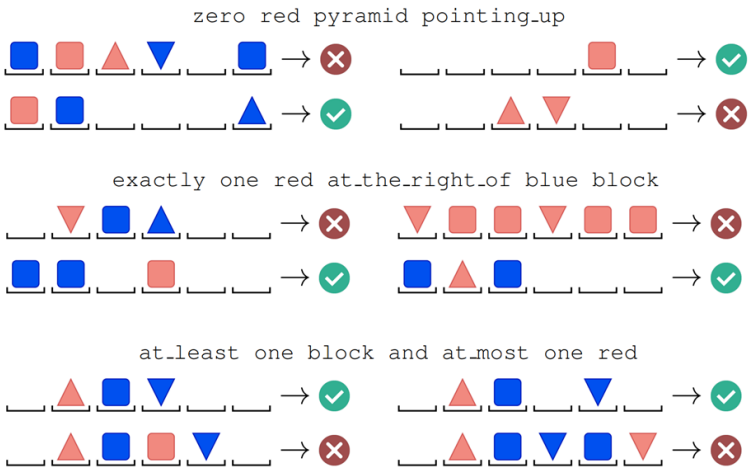
Exactly one blue surrounded by triangles

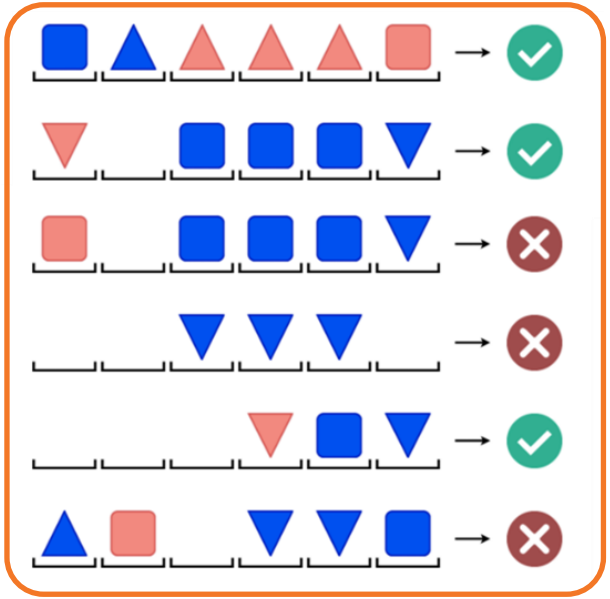
Scientific puzzles: Odeen



???

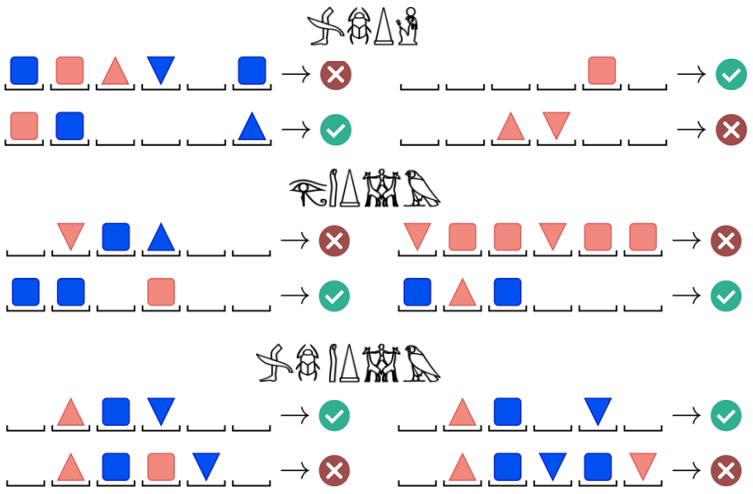
Scientific puzzles: Odeen

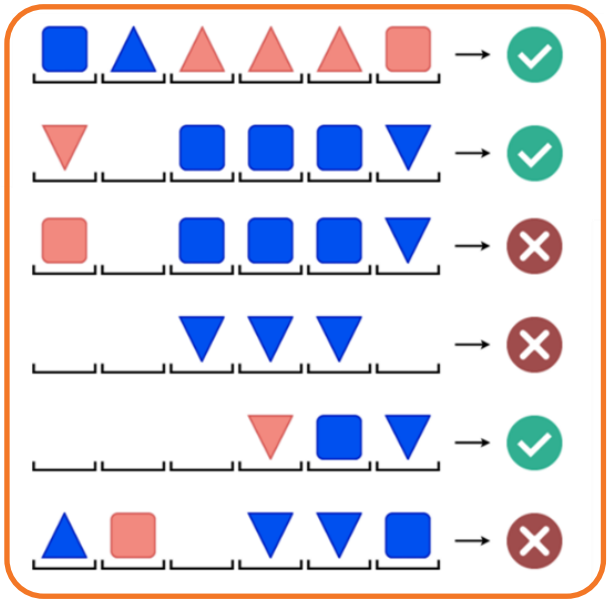




???

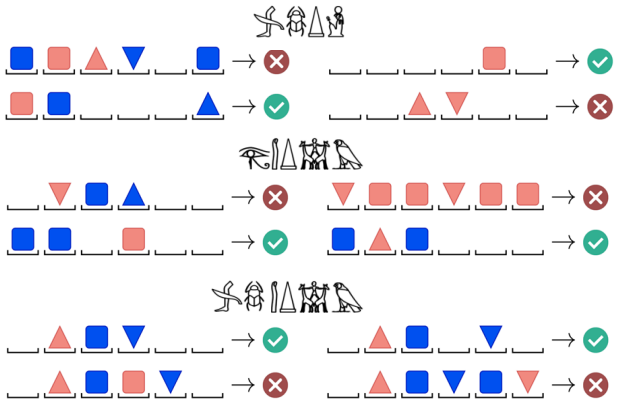
Scientific puzzles: Odeen

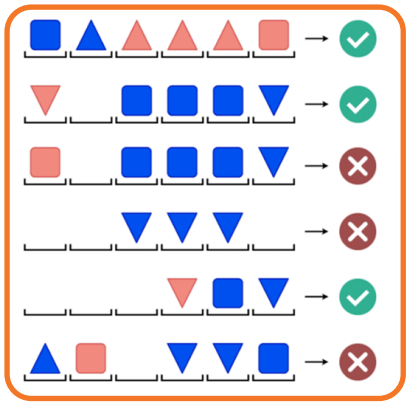


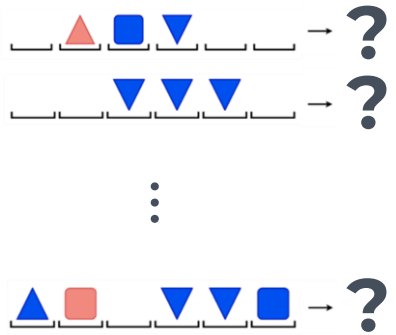


???

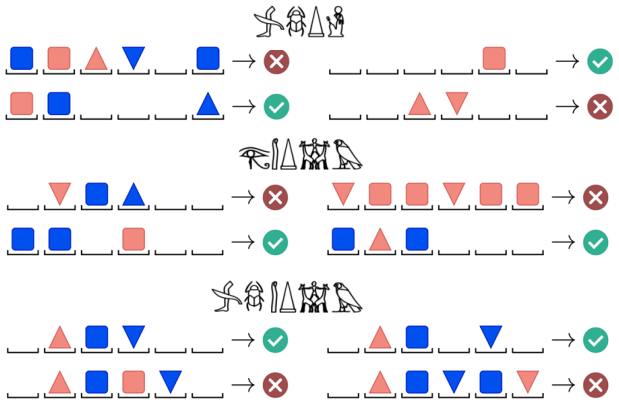
Scientific puzzles: Odeen

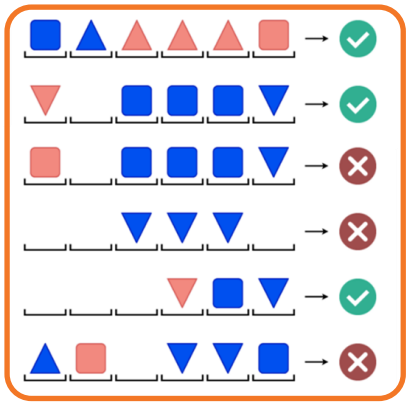


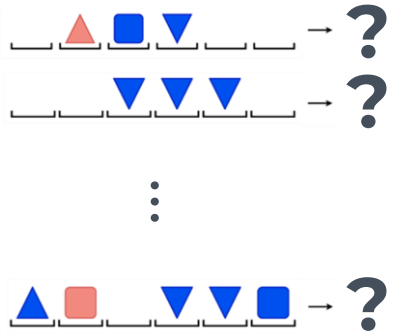


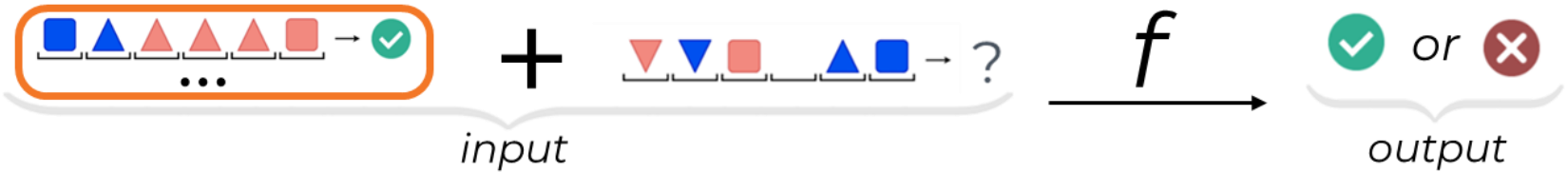


Explanatory Learning problem: find f









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.

Generalization gaps

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

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*

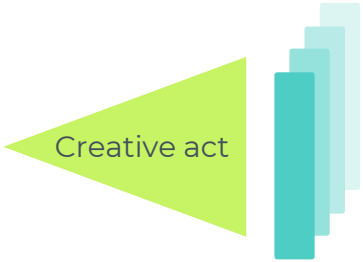
Data ⟶ Theory

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*

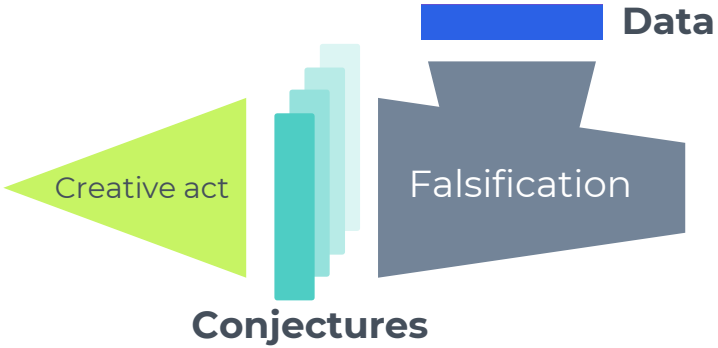
Data ⟶ Theory

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*

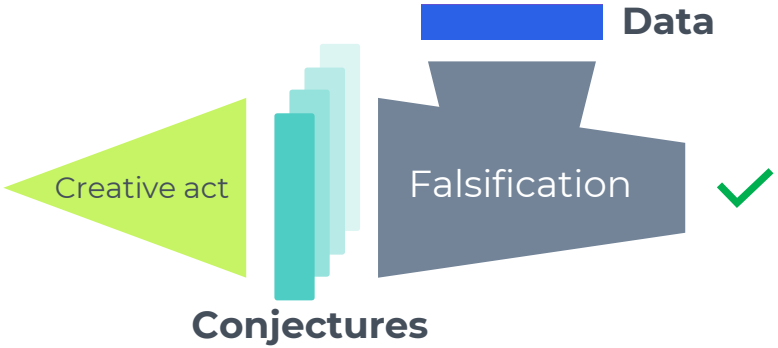
Data ⟶ Theory

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*

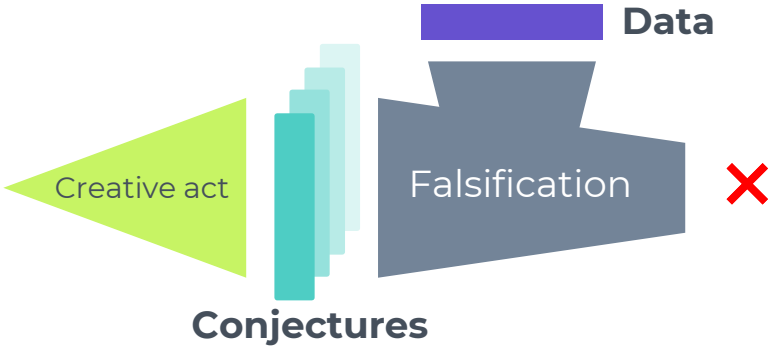
Data → Theory

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*

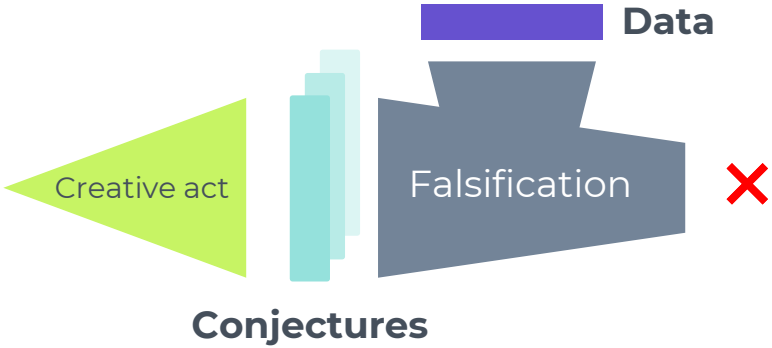
Data ⟶ Theory

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

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

Rationalist perspective shift

Theory ⟶ **Data**



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

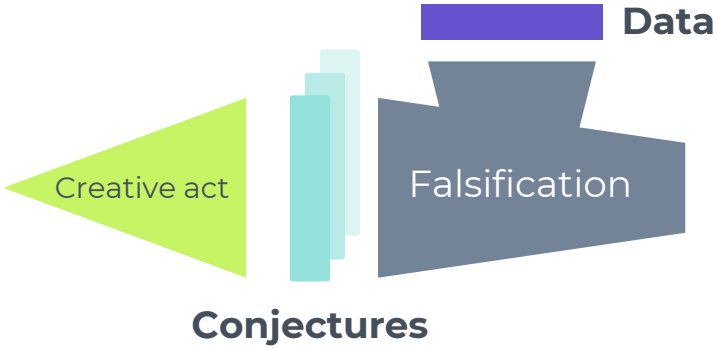
Data ⟶ Theory

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

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

Rationalist perspective shift

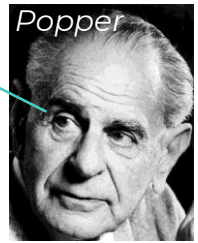
Theory → **Data**



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

Data → Theory

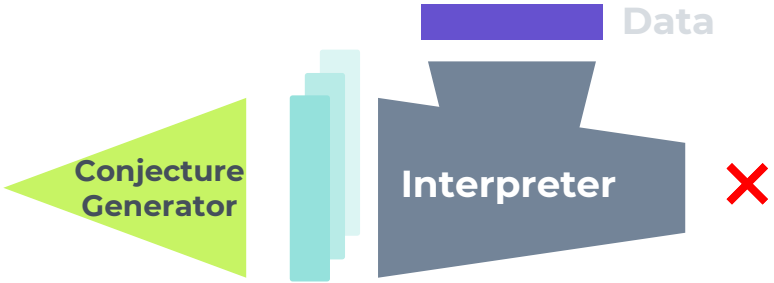
- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time



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

Critical Rationalist Networks

Theory ⟶ **Data**



Conjectures

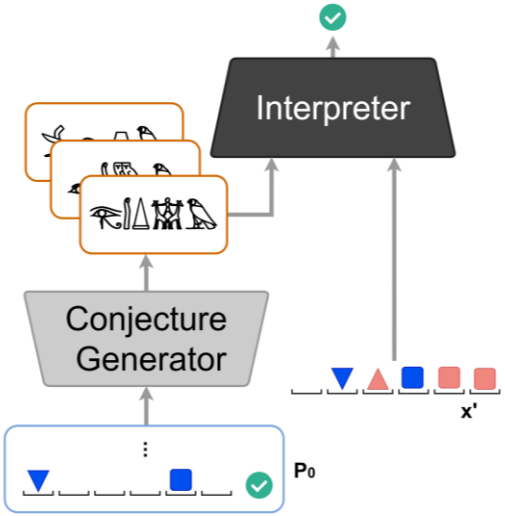
Hypothesis *as a* ***language proposition*** *which can only be accepted or refused in toto*

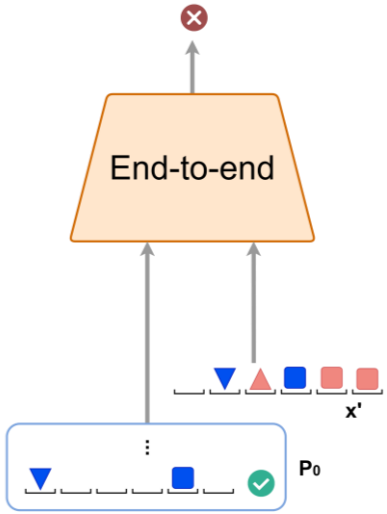
Data ⟶ Theory

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

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

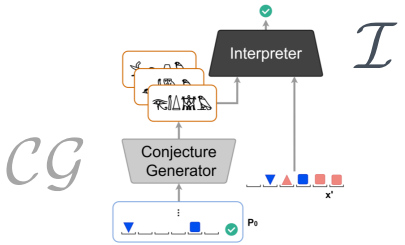
Critical Rationalist Networks

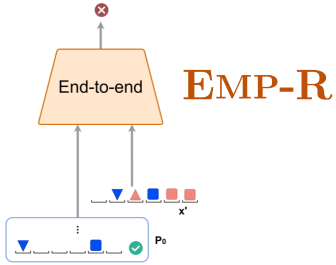


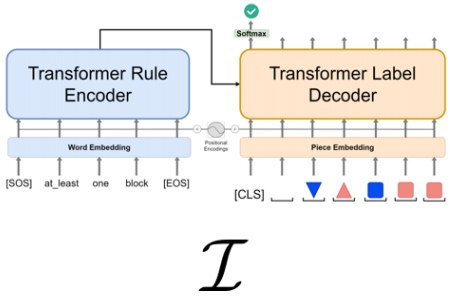


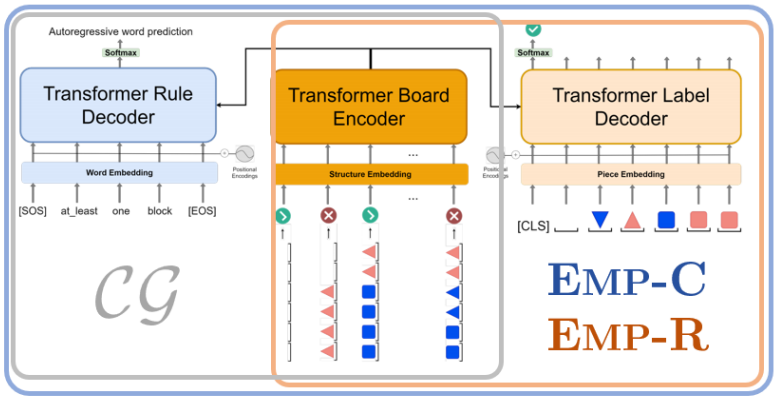
- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Critical Rationalist Networks





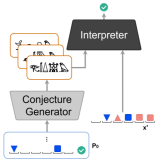


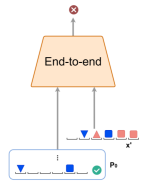


- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time

Results: CRNs vs empiricist models

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

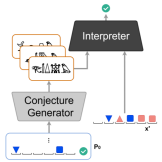




Results: CRNs vs empiricist models

The CRN can discover the correct explanation 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	-

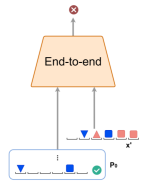


Generalization gaps

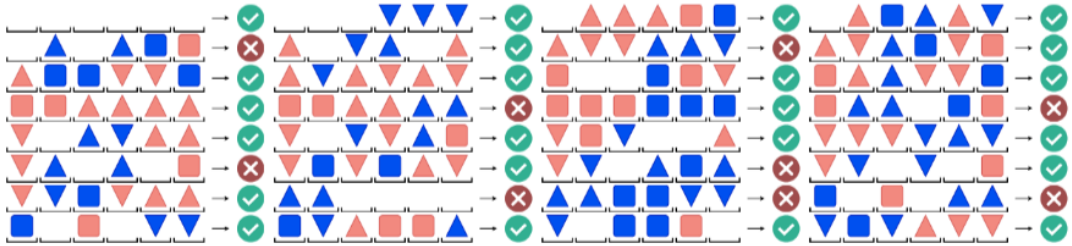
Unexplainable predictions

Unreliable predictions

Fixed thinking time

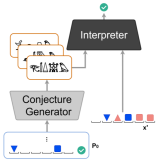


Results: CRNs vs empiricist models

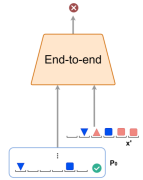


Board 04
Golden Rule: “at_most 1 blue pyramid pointing_up”
CRN: “zero blue or at_most 1 blue pyramid pointing_up”; T-acc 1.0 ✓
EMP-C: “zero 1 blue touching or or”; T-acc: 0.89 ✗

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	-

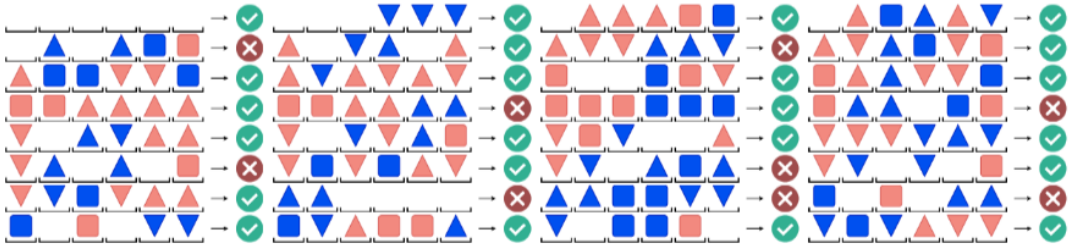


- Generalization gaps
- Unexplainable predictions
- Unreliable predictions
- Fixed thinking time



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 ✗

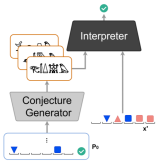
MODEL	NRS	T-ACC	R-ACC
CRN	**0.777**	**0.980**	**0.737**
EMP-C	0.225	0.905	0.035
EMP-R	0.156	0.898	-

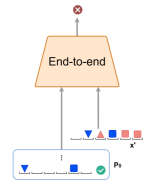
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time





Results: CRNs vs empiricist models

Generalization power

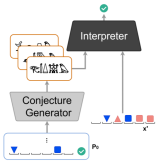
The bank ML algorithm spoke: "Loan denied"; explanation: "Two not paid loan in the past and resident in a district with a high rate of insolvents".
With a CRN, we can naturally discard this explanation and compute a new prediction for just "Two not paid loan in the past".

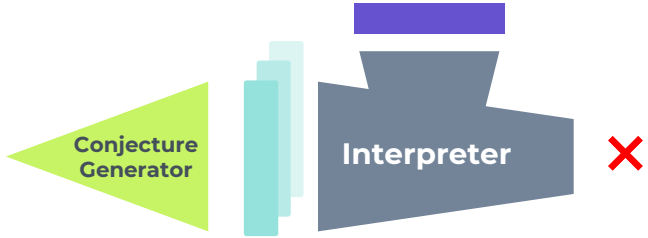
Generalization gaps

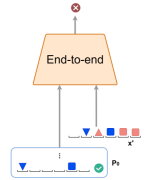
Unexplainable predictions

Unreliable predictions

Fixed thinking time







Results: CRNs vs empiricist models

Generalization power

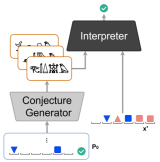
The bank ML algorithm spoke: "Loan denied"; explanation: "Two not paid loan in the past and resident in a district with a high rate of insolvents".
With a CRN, we can naturally discard this explanation and compute a new prediction for just "Two not paid loan in the past".

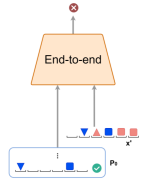
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time





Results: CRNs vs empiricist models

Generalization power

Truly Explainable predictions

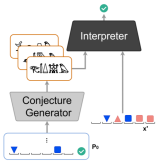
The bank ML algorithm spoke: "Loan denied"; explanation: "Two not paid loan in the past and resident in a district with a high rate of insolvents".
With a CRN, we can naturally discard this explanation and compute a new prediction for just "Two not paid loan in the past".

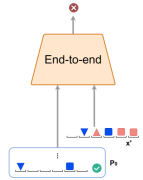
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time





Results: CRNs vs empiricist models

Generalization power

Truly Explainable predictions

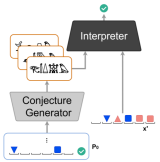
If no conjecture is compatible with data?
A CRN returns “unknown explanation” rather than a random prediction

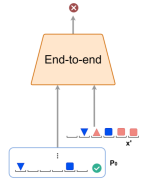
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time





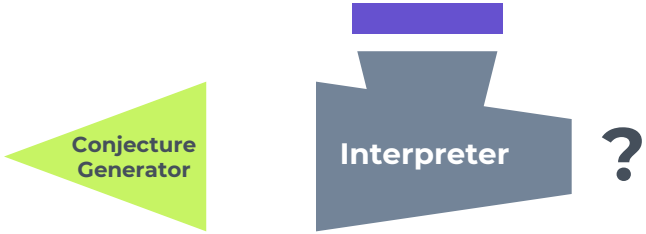
Results: CRNs vs empiricist models

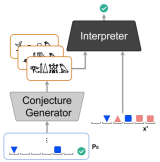
Generalization power

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



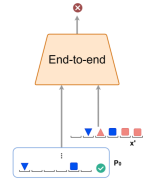


Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time



Results: CRNs vs empiricist models

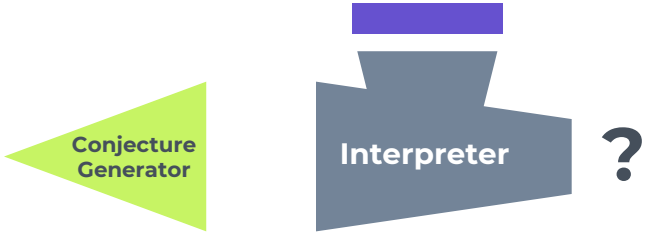
Generalization power

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

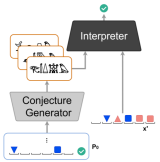


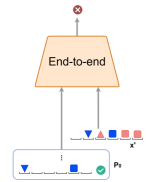
Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time





Results: CRNs vs empiricist models

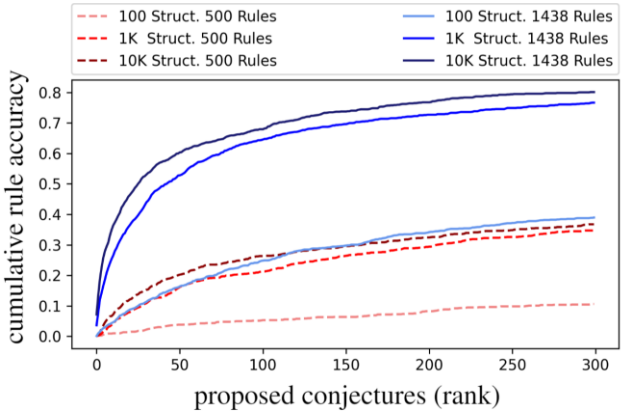
Generalization power

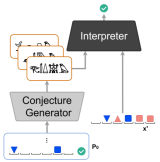
Truly Explainable predictions

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



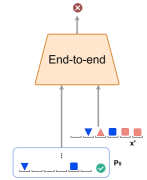


Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time



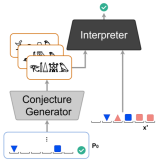
Results: CRNs vs empiricist models

Generalization power

Truly Explainable predictions

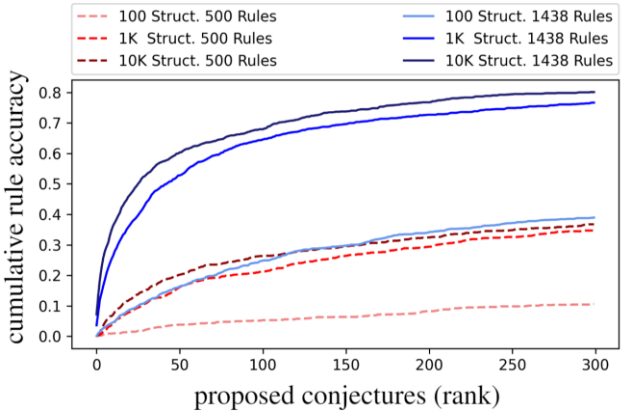
Reliable predictions

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

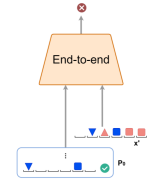


Generalization gaps

Unexplainable predictions

Unreliable predictions

Fixed thinking time



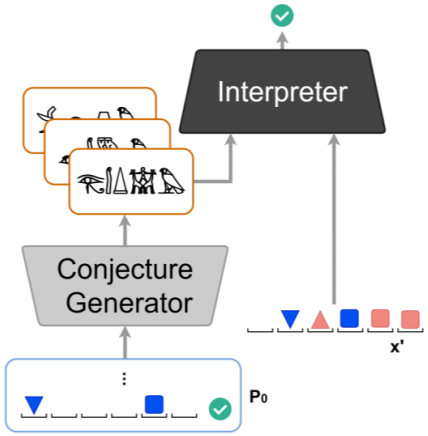
Results: CRNs vs empiricist models

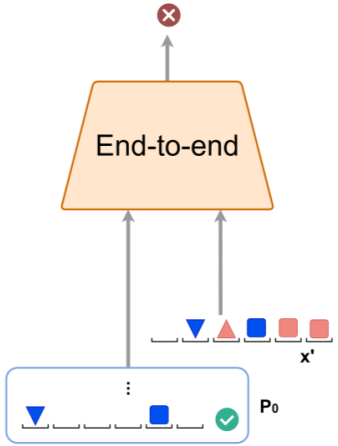
Generalization power

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

Bonus: A surprising result

Handling ambiguity and contradiction. One may reasonably expect that a CRN equipped with the ground-truth interpreter used to generate the dataset, would perform better than a CRN with a learned interpreter. Remarkably, this is not always the case, as reported in Table 1.

Train Data		NRS Fully-learned CRN	NRS Hardcoded $\mathcal{I}$ CRN
10K struct.	1438 rules	**0.813**	0.801
1K struct.	1438 rules	**0.777**	0.754
100 struct.	1438 rules	0.402	**0.406**
10K struct.	500 rules	0.354	**0.377**
1K struct.	500 rules	0.319	**0.336**
100 struct.	500 rules	**0.109**	0.101

Bonus: Explanatory Learning formalism

Problem setup. Formally, let phenomena $P_1, P_2, P_3, \ldots$ be subsets of a universe U, which is a large set with no special structure (i.e., all the possible observations $U = \{x_1, \ldots, x_z\}$). Over a universe U, one can define a language L as a pair $(\Sigma_L, \mathcal{I}_L)$, where Σ_L is a finite collection of short strings over some alphabet A, with $|\Sigma_L| \gg |A|$, and $\mathcal{I}_L$ is a binary function $\mathcal{I}_L : U \times \Sigma_L \to \{0, 1\}$, which we call *interpreter*. We say that a phenomenon P_i is *explainable* in a language L if there exists a string $e \in \Sigma_L$ such that, for any $x \in U$, it occurs $\mathcal{I}_L(x, e) = \mathbf{1}_{P_i}(x)$, where $\mathbf{1}_{P_i}(x)$ is the indicator function of P_i. We call the string e an explanation, in the language L, for the phenomenon P_i.

Explainability definition

Our first contribution is the introduction of a new class of machine learning problems, which we refer to as *Explanatory Learning* (EL).

Consider the general problem of making a new prediction for a phenomenon $P_0 \subset U$. In our setting, this is phrased as a binary classification task: given a sample $x' \in U$, establish whether $x' \in P_0$ or not. We are interested in two instances of this problem, with different underlying assumptions:

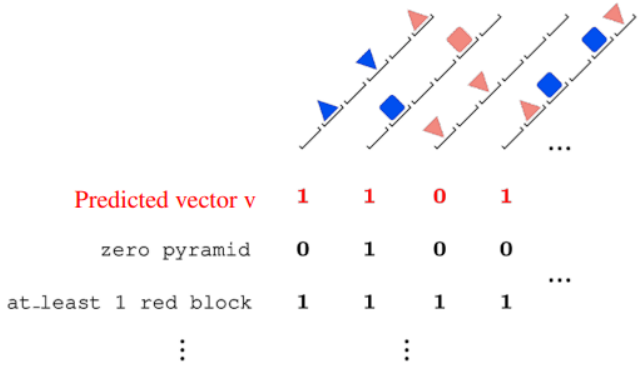
Bonus: Explanatory Learning formalism

- **The communication problem: we have an explanation**. We are given an explanation e_0 for P_0, in an unknown language L. This means that we do not have access to an interpreter $\mathcal{I}_L$; e_0 looks like Japanese to a non-Japanese speaker. Instead, we are also given other explanations $\{e_1, \ldots, e_n\}$, in the same language, for other phenomena $P_1, \ldots, P_n$, as well as observations of them, i.e., datasets $\{D_1, \ldots, D_n\}$ in the form $D_i = \{(x_1, \mathbf{1}_{P_i}(x_1)), \ldots, (x_m, \mathbf{1}_{P_i}(x_m))\}$, with $m \ll |U|$. Intuitively, here we expect the learner to use the explanations paired with the observations to build an approximated interpreter $\hat{\mathcal{I}}_L$, and then use it to make the proper prediction for x' by evaluating $\hat{\mathcal{I}}_L(x', e_0)$.
- **The scientist problem: we do not have an explanation**. We are given explanations $\{e_1, \ldots, e_n\}$ in an unknown language L for other phenomena $P_1, \ldots, P_n$ and observations of them $\{D_1, \ldots, D_n\}$. However, we do not have an explanation for P_0; instead, we are given just a small set of observations $D_0 = \{(x_1, \mathbf{1}_{P_0}(x_1)), \ldots, (x_k, \mathbf{1}_{P_0}(x_k))\}$ and two guarantees, namely that P_0 is explainable in L, and that D_0 is *representative* for P_0 in L. That is, for every phenomenon $P \neq P_0$ explainable in L there should exist at least a $x_i \in D_0$ such that $\mathbf{1}_{P_0}(x_i) \neq \mathbf{1}_P(x_i)$. Again, we expect the learner to build the interpreter $\hat{\mathcal{I}}_L$, which should first guide the search for the missing explanation e_0 based on the clues D_0, and then provide the final prediction through $\hat{\mathcal{I}}_L(x', e_0)$.

Representativity definition

Bonus: Metrics

Metrics. As described above, the task is to tag ℓ new structures for each of s unexplained games. An EL algorithm addressing this task encodes the predicted rule as an ℓ-dimensional binary vector $\mathbf{v}$ per game (predicted vector), where $v_i = 1$ means that the i-th structure satisfies the predicted rule, and $v_i = 0$ otherwise (see inset). Let $\mathbf{w}^*$ be the ground-truth vector, obtained by tagging the ℓ structures according to the correct secret rule. Then, the Hamming distance $d_H(\mathbf{v}, \mathbf{w}^*)$ measures the number of wrong tags assigned by the EL algorithm; if $d_H(\mathbf{v}, \mathbf{w}^*) < d_H(\mathbf{v}, \mathbf{w}_i)$, where $\mathbf{w}_i \neq \mathbf{w}^*$ ranges over all the possible ≈25k rules, then the predicted rule $\mathbf{v}$ made by the algorithm is deemed correct.



According to this, the *Nearest Rule Score* (NRS) is the number of correctly predicted rules over a total of s games. A second score, the *Tagging Accuracy* (T-Acc), directly counts the number of correct tags averaged over s games; this is more permissive in the following sense. Consider two different rules A and B sharing 99% of the taggings, and let A be the correct one; if an EL model tags all the structures according to the *wrong* rule B, it still reaches a T-Acc of 99%, but the NRS would be 0. An EL algorithm with these scores would be good at making predictions, but would be based on a wrong explanation.

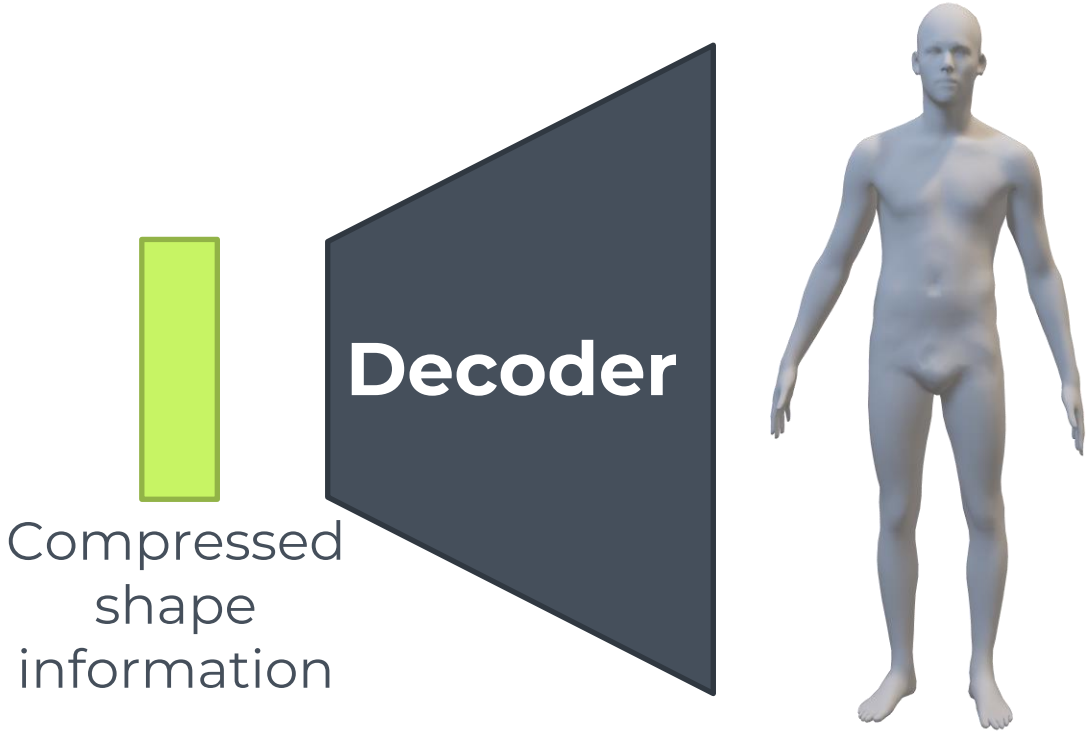
2.

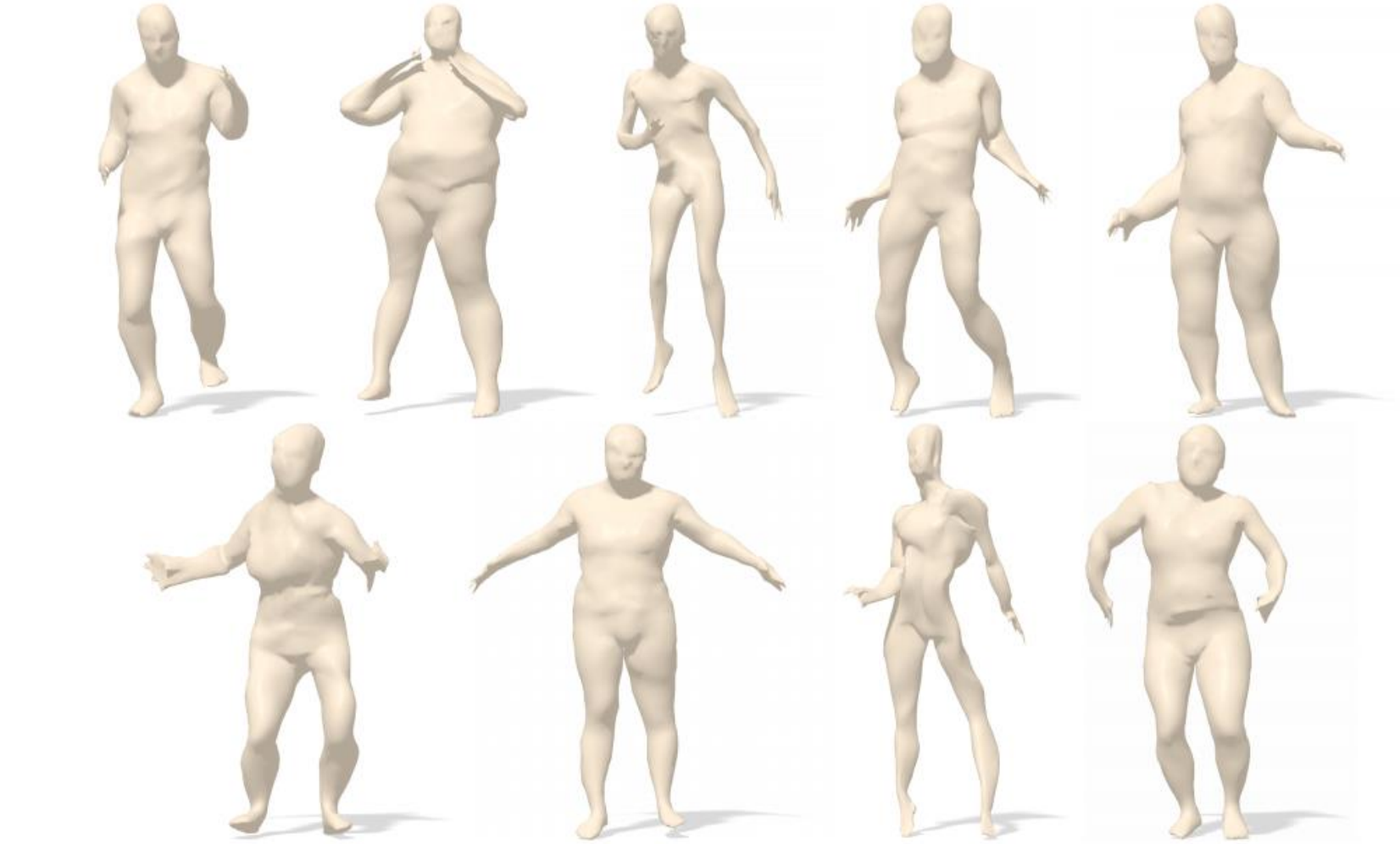
Reduce

the set of admissible hypotheses

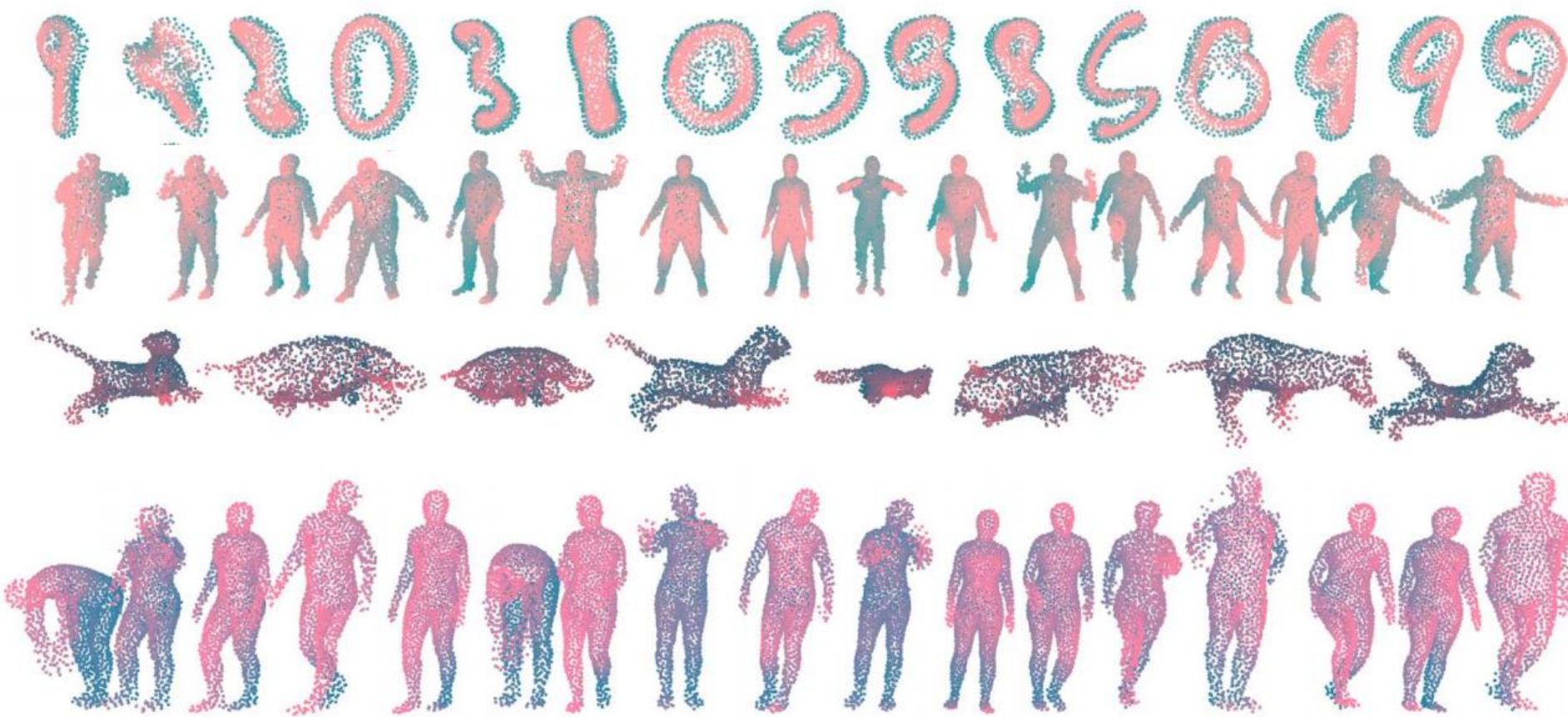
LIMP: Learning Latent Shape Representations with Metric Preservation Priors

Shape synthesis through a NN

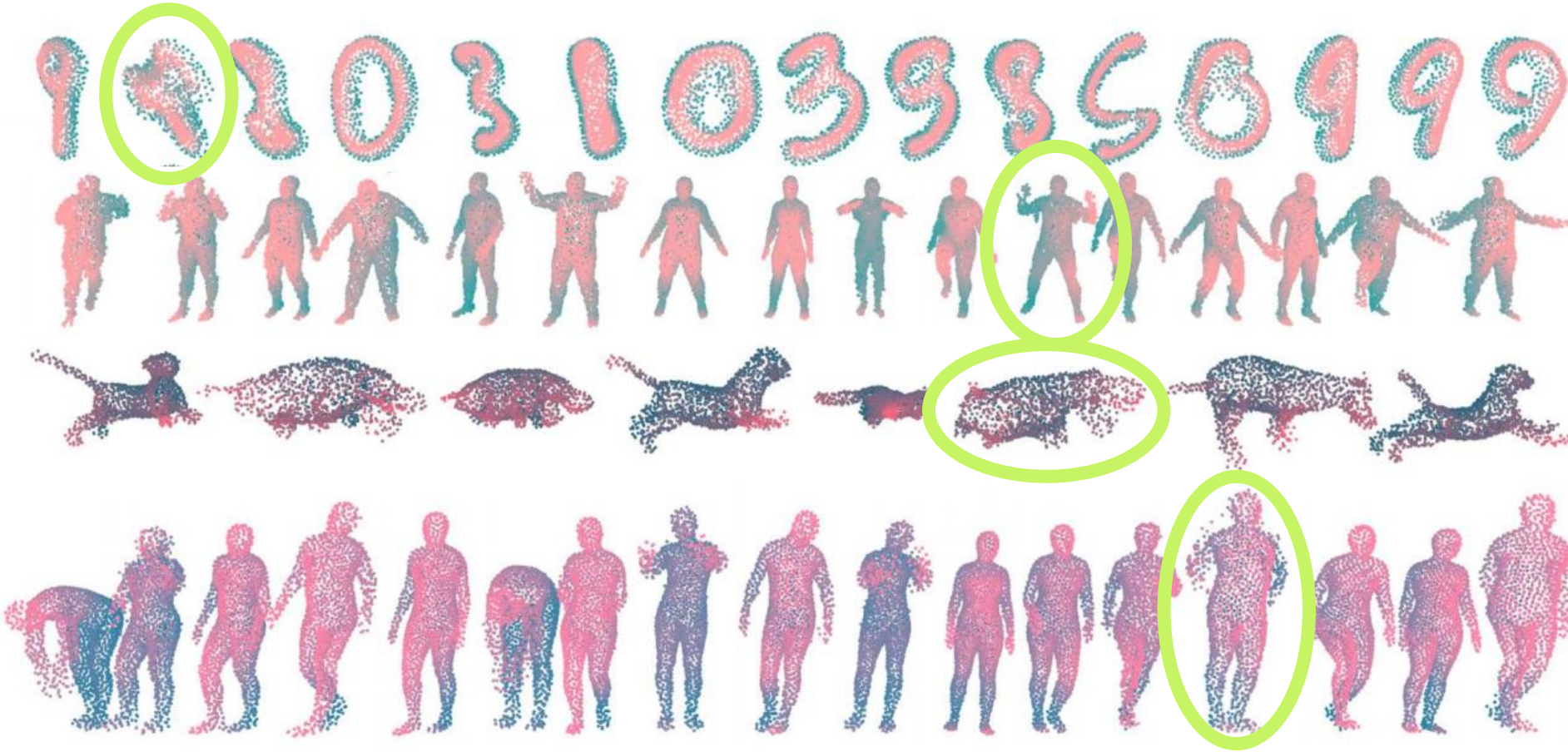




[Litany et al. 2018]

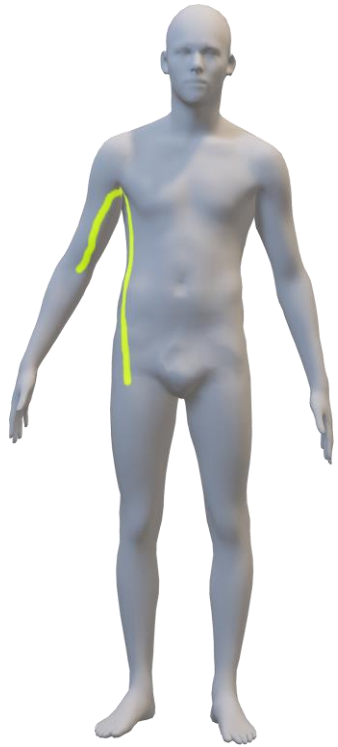


[Aumentado-Armstrong et al. 2019]



[Aumentado-Armstrong et al. 2019]

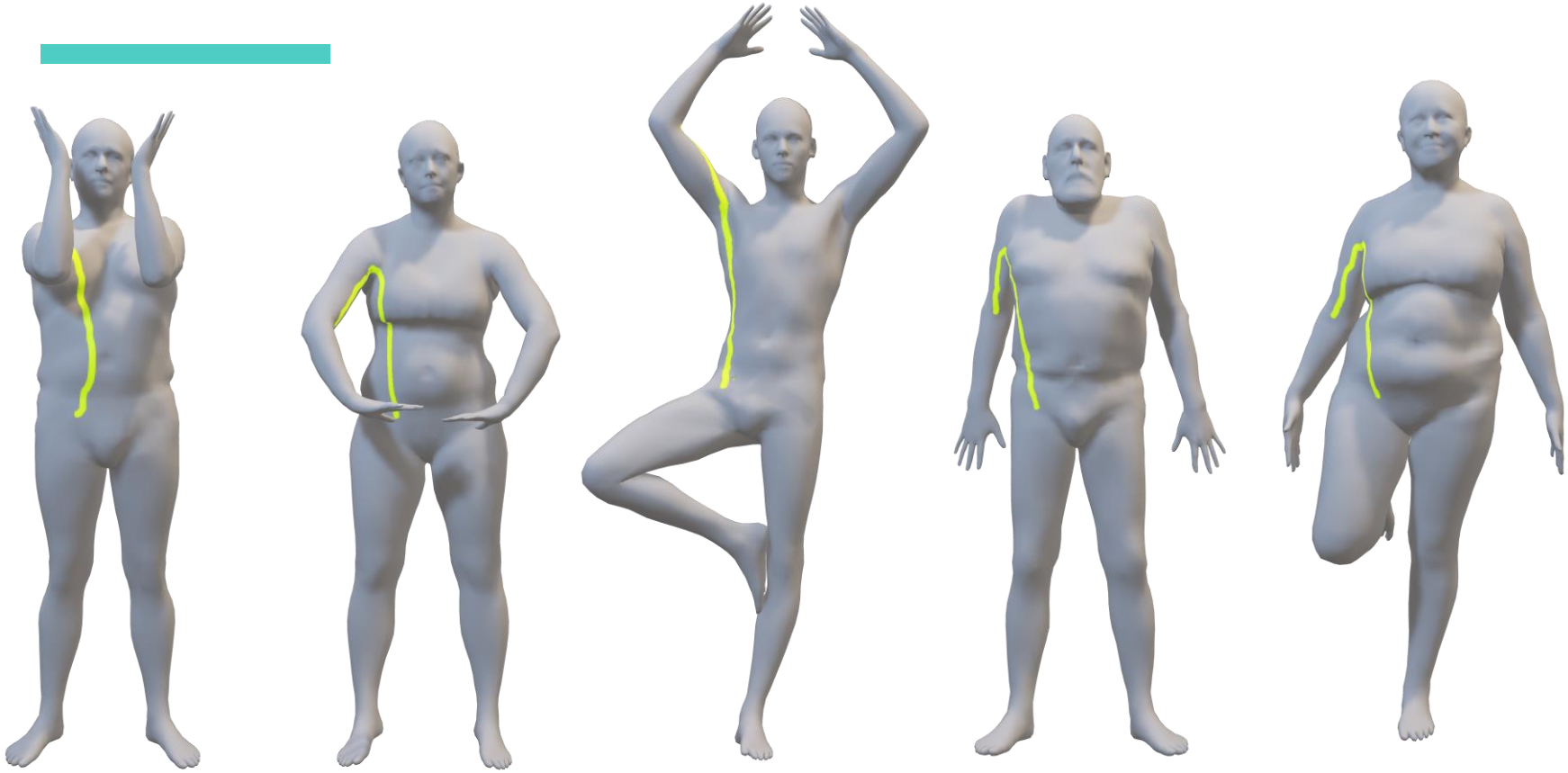
Beyond the data prior



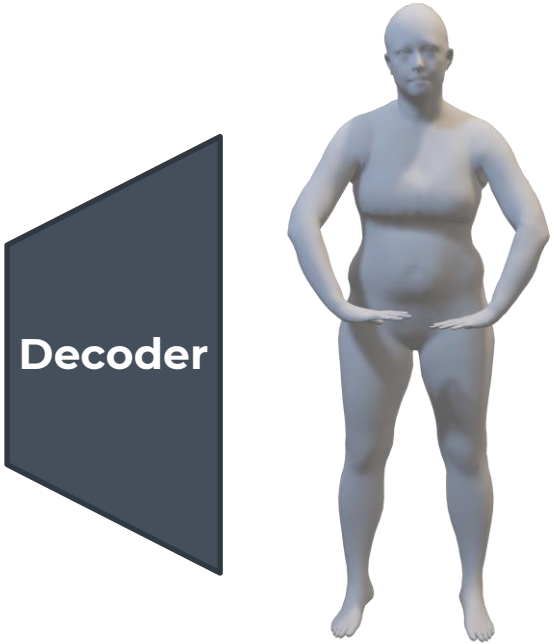
We should have near-isometries

$$d_{geo1} \approx d_{geo2}$$

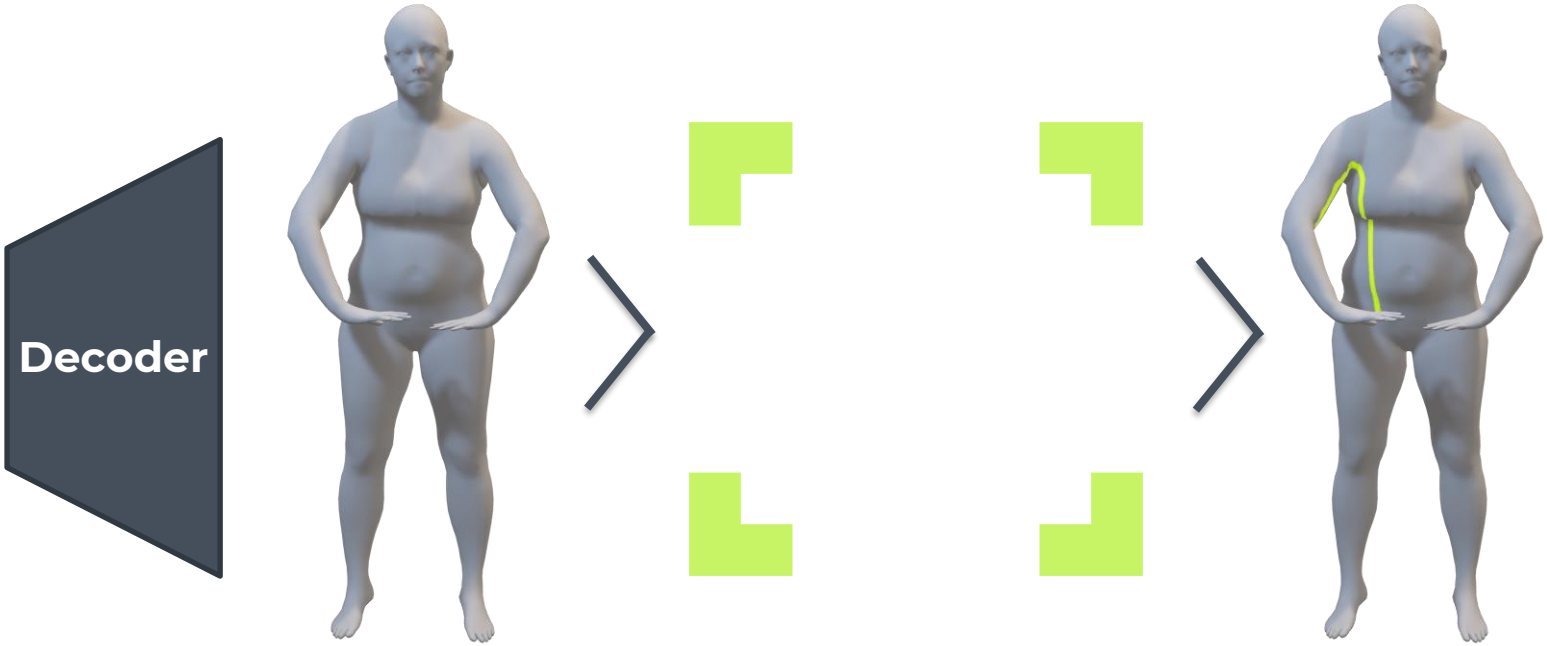
We should have near-near-isometries



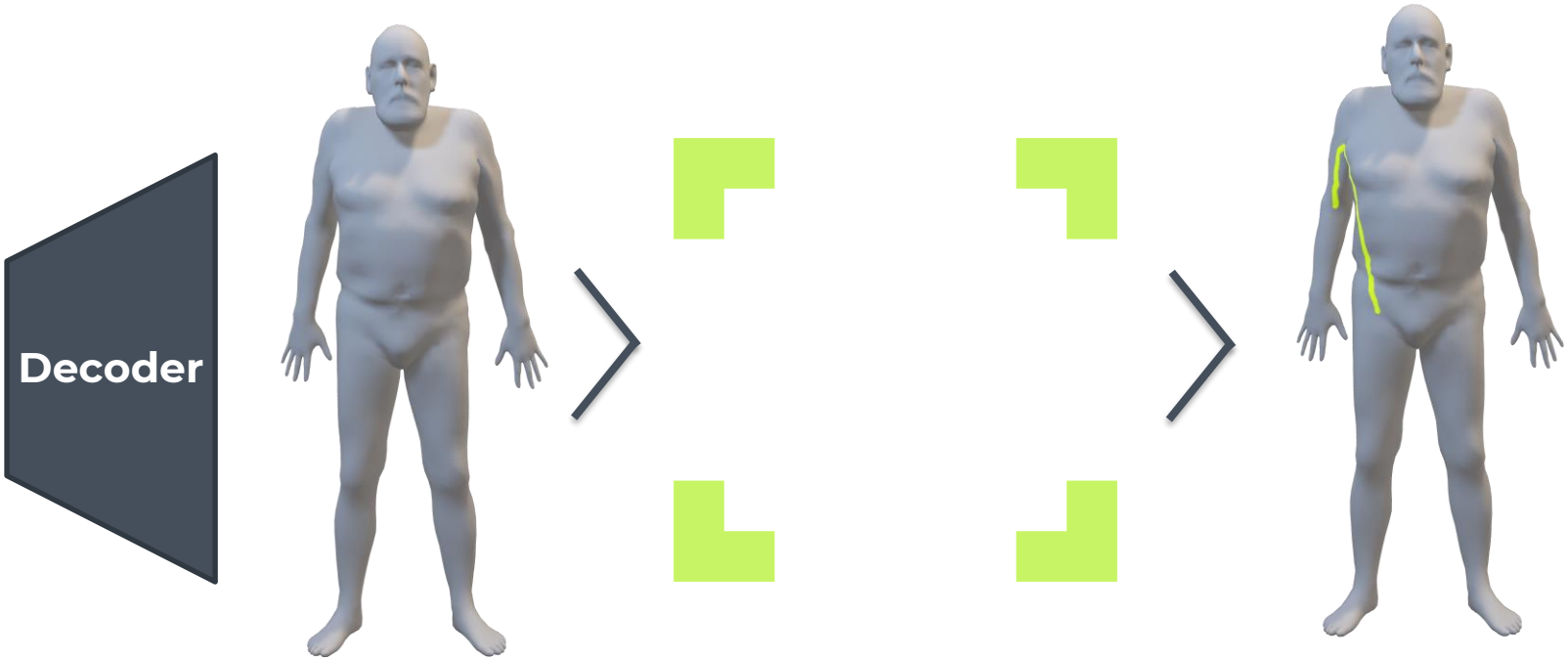
Compute geodesics on the fly



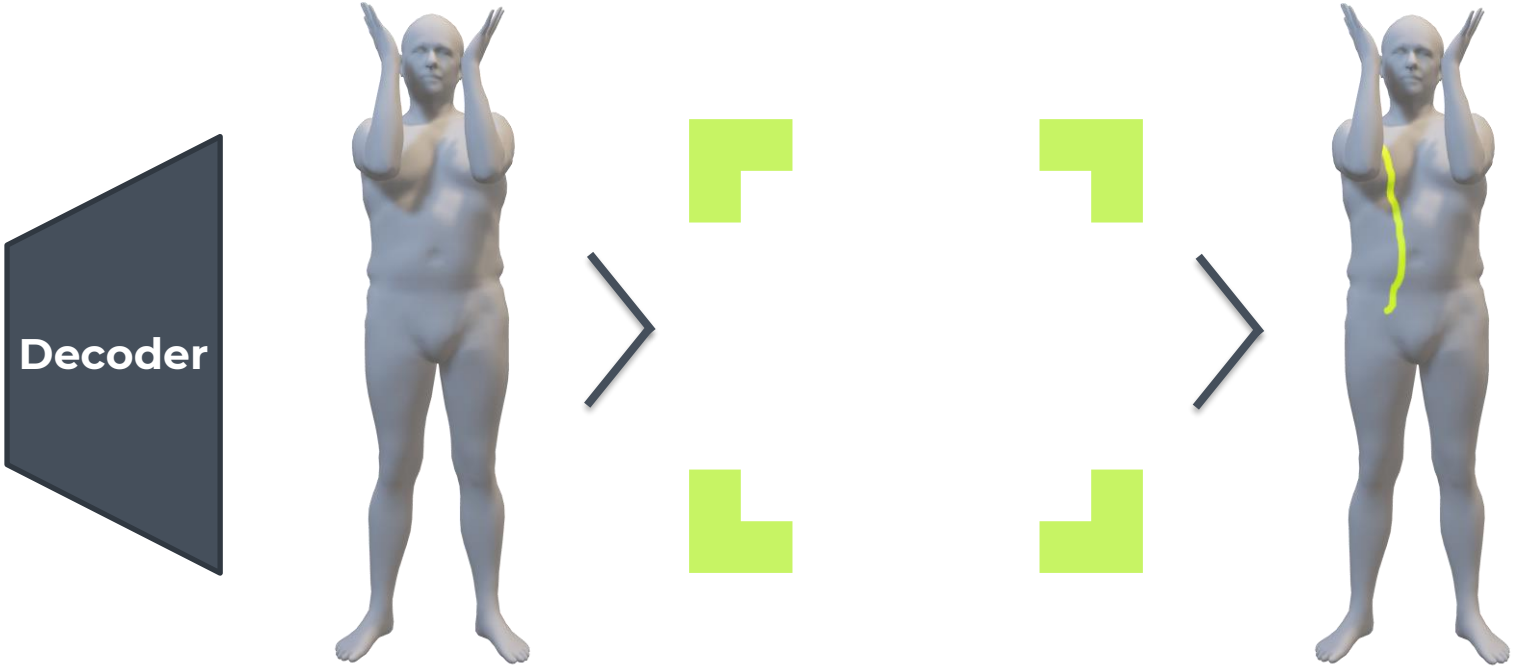
Compute geodesics on the fly



Compute geodesics on the fly

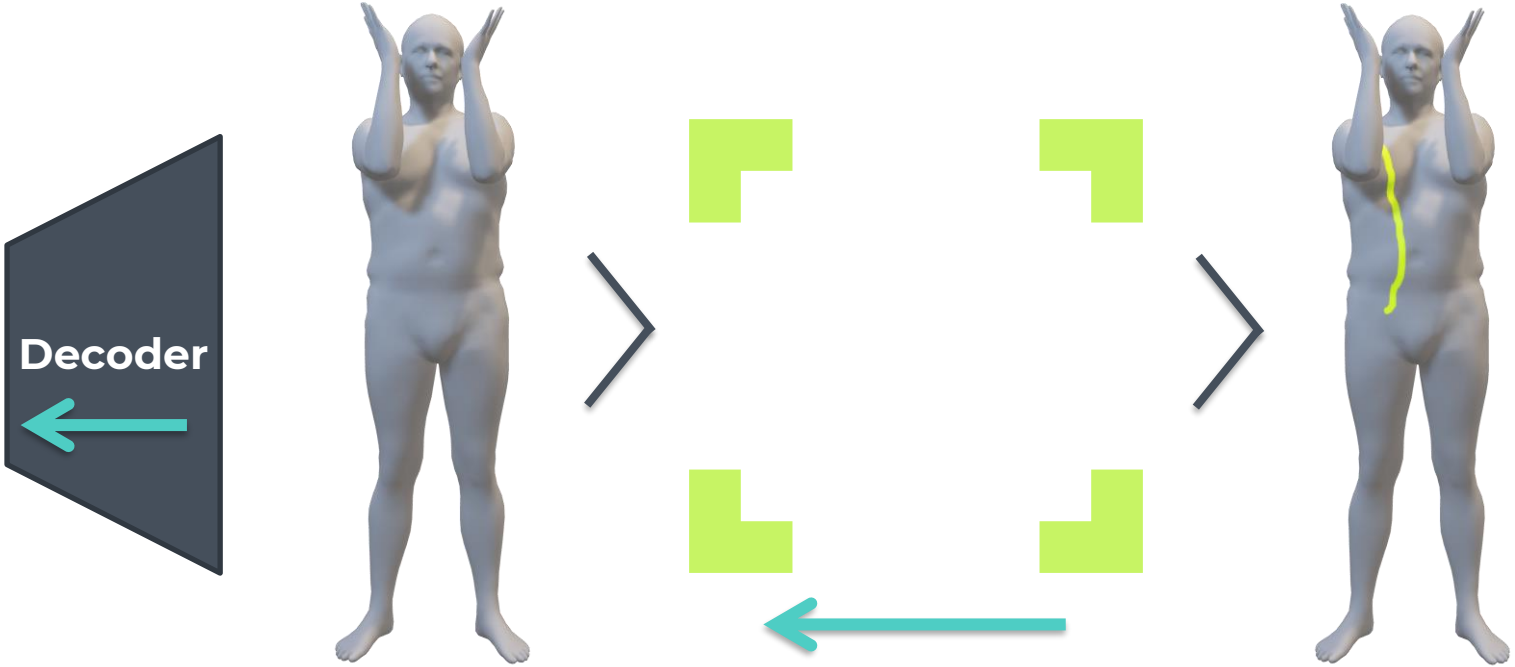


Compute geodesics on the fly



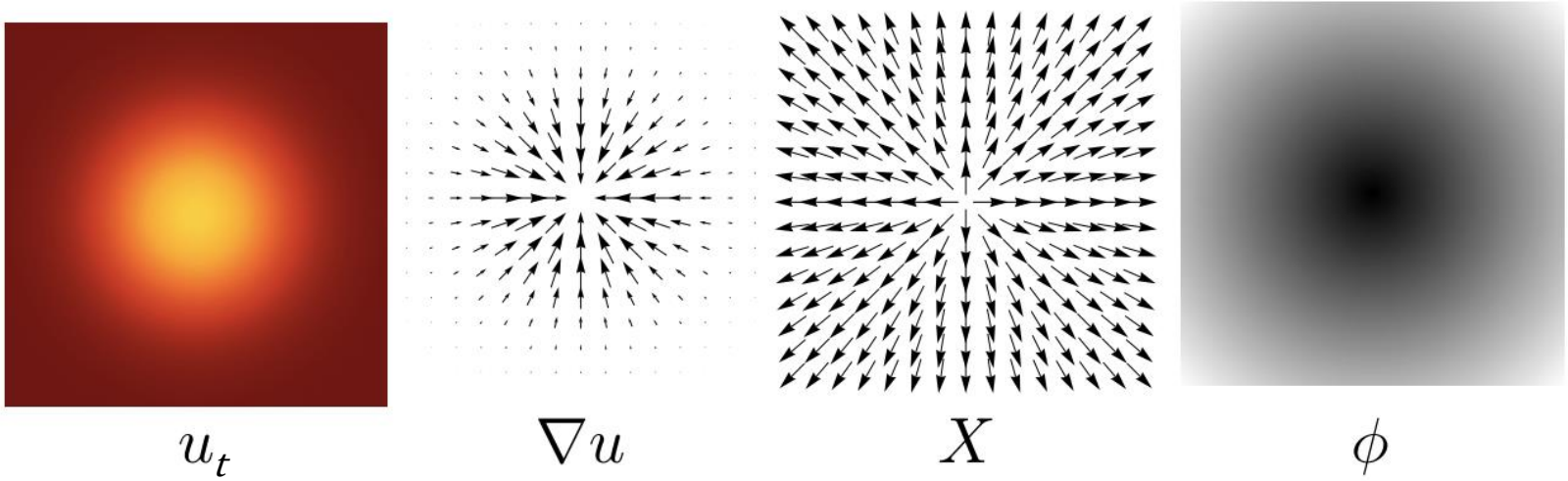
Compute geodesics on the fly

In a **differentiable** way



Geodesics in heat

Geodesics in heat



[Crane et al. 2012]

Geodesics in heat - differentiable

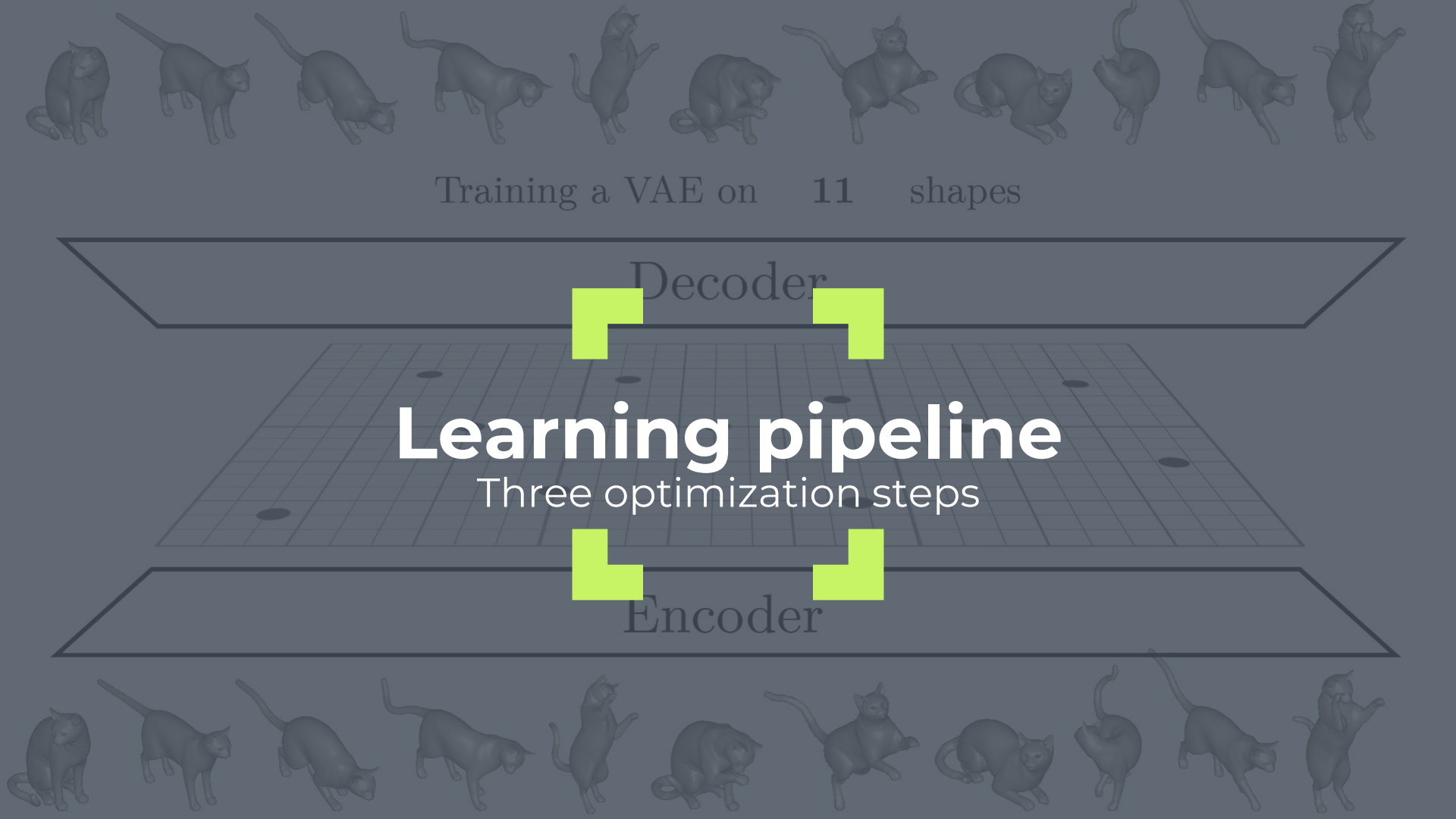
Algorithm 1 The Heat Method

I. Integrate the heat flow $\dot{u} = \Delta u$ for some fixed time t.

II. Evaluate the vector field $X = -\nabla u / |\nabla u|$.

III. Solve the Poisson equation $\Delta \phi = \nabla \cdot X$.

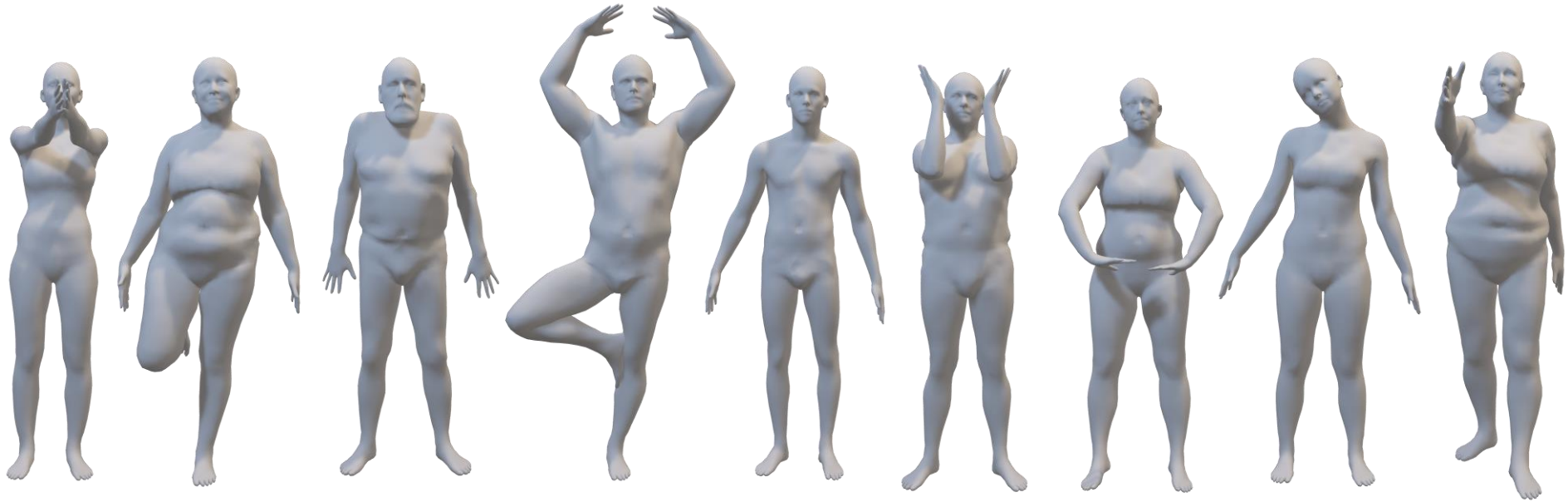
[Crane et al. 2012]



Learning pipeline

Three optimization steps

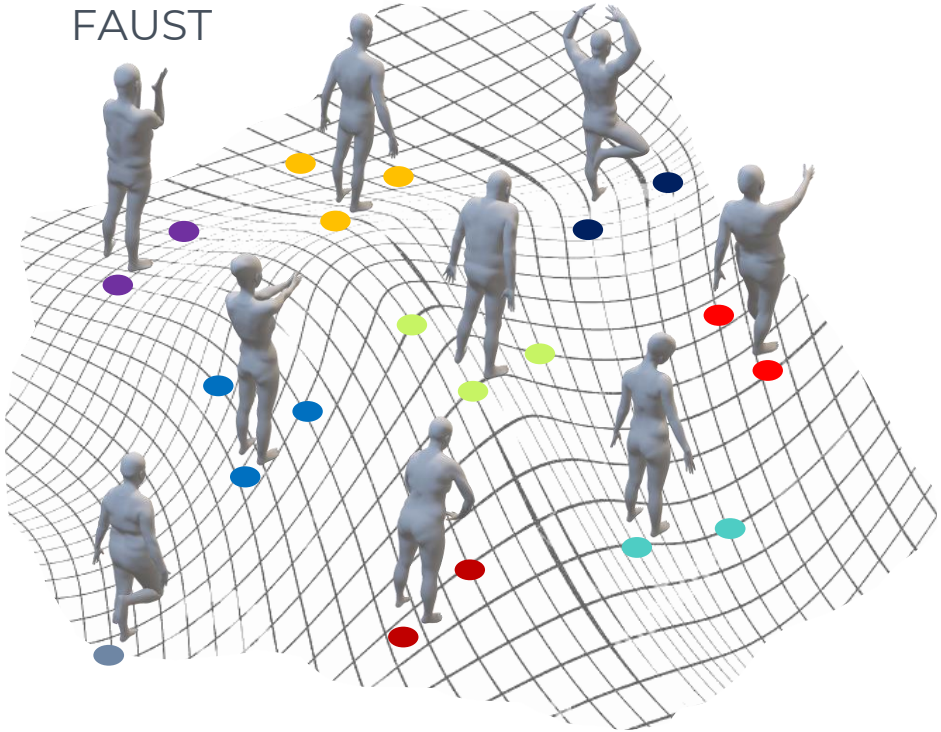
Faust reduced to 2100 vertices



80 shapes for training (10 poses x 8 subjects)

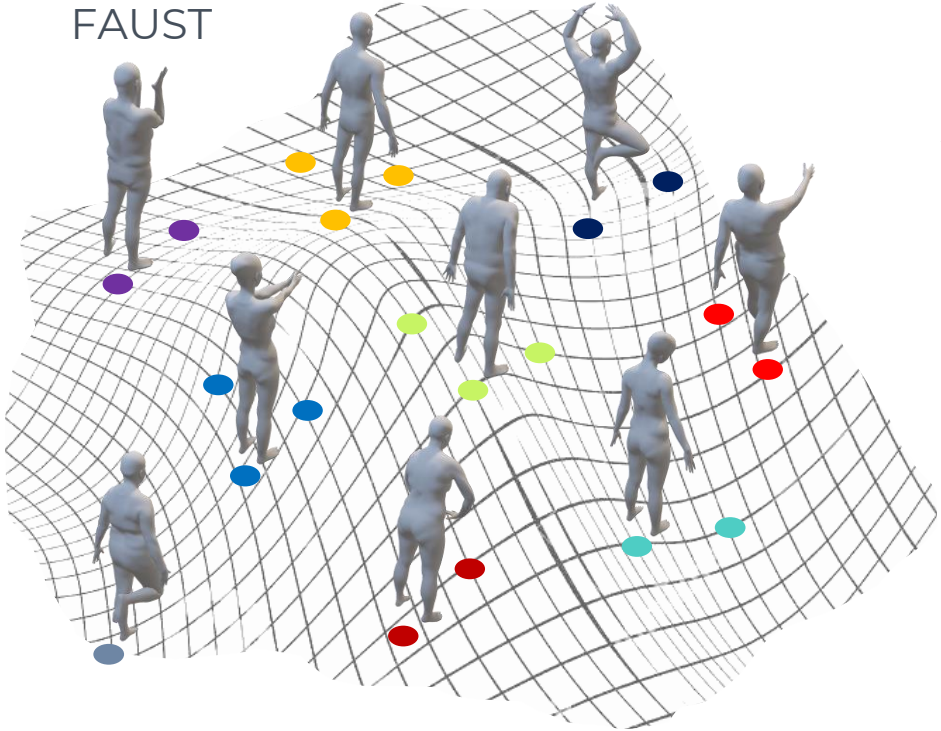
A sparse and small dataset

FAUST



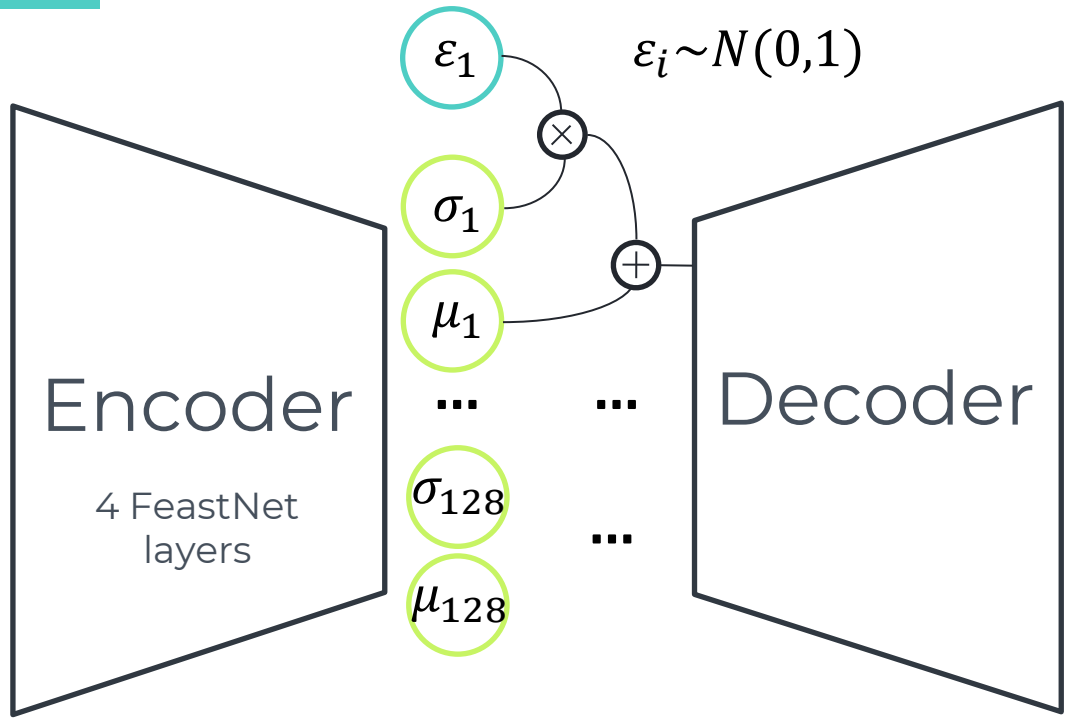
A sparse and small dataset

FAUST



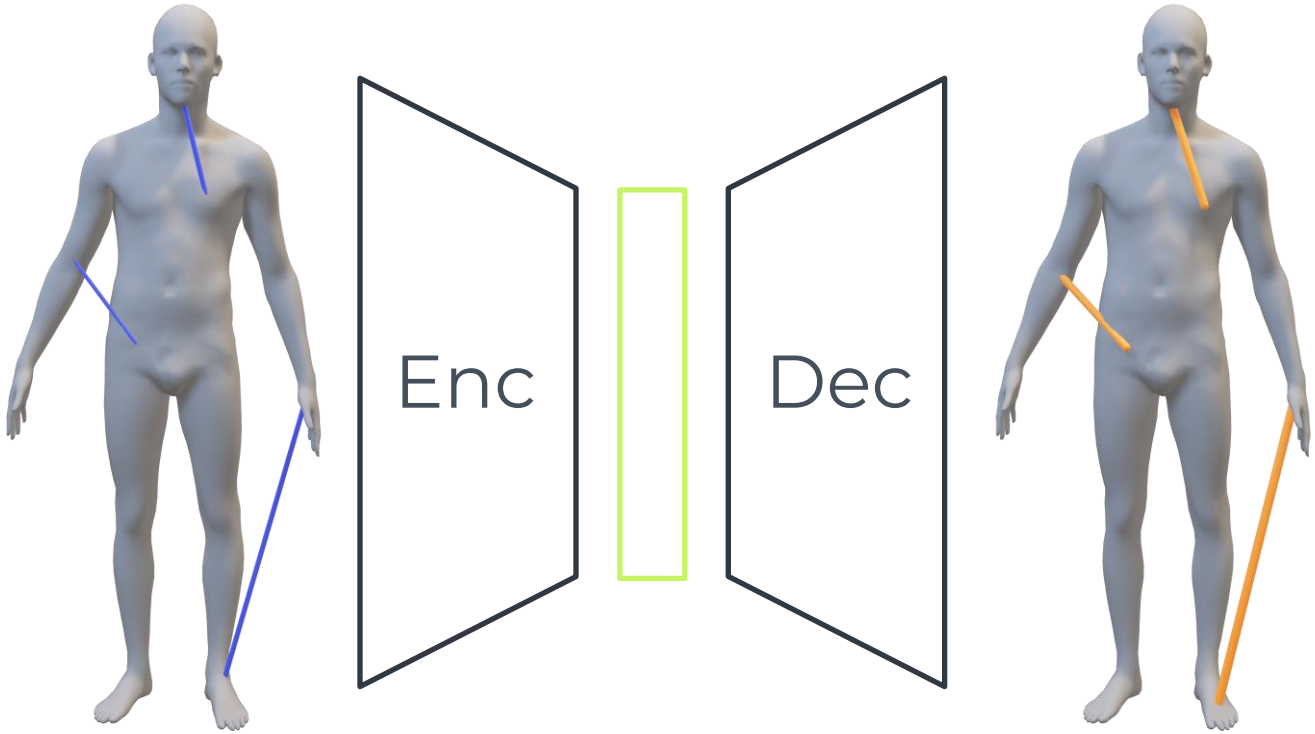
DFAUST
SMPL
AMASS
Dyna

Variational AutoEncoder



1° optimization

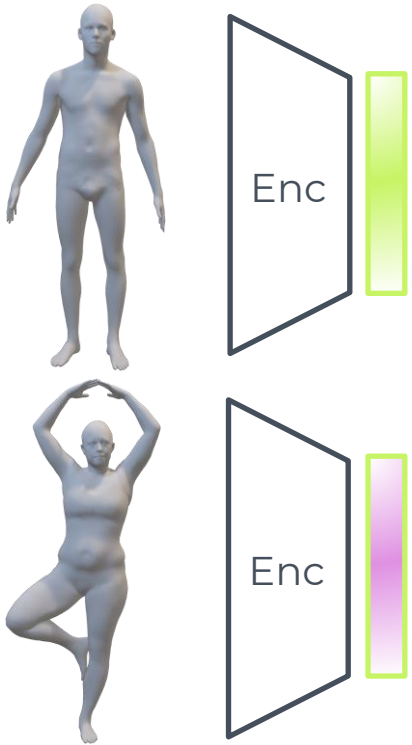
$$L = \alpha|D_{eucl} - \widehat{D}_{eucl}| + \beta\left(\frac{D_{eucl} - \widehat{D}_{eucl}}{D_{eucl}}\right)^2$$



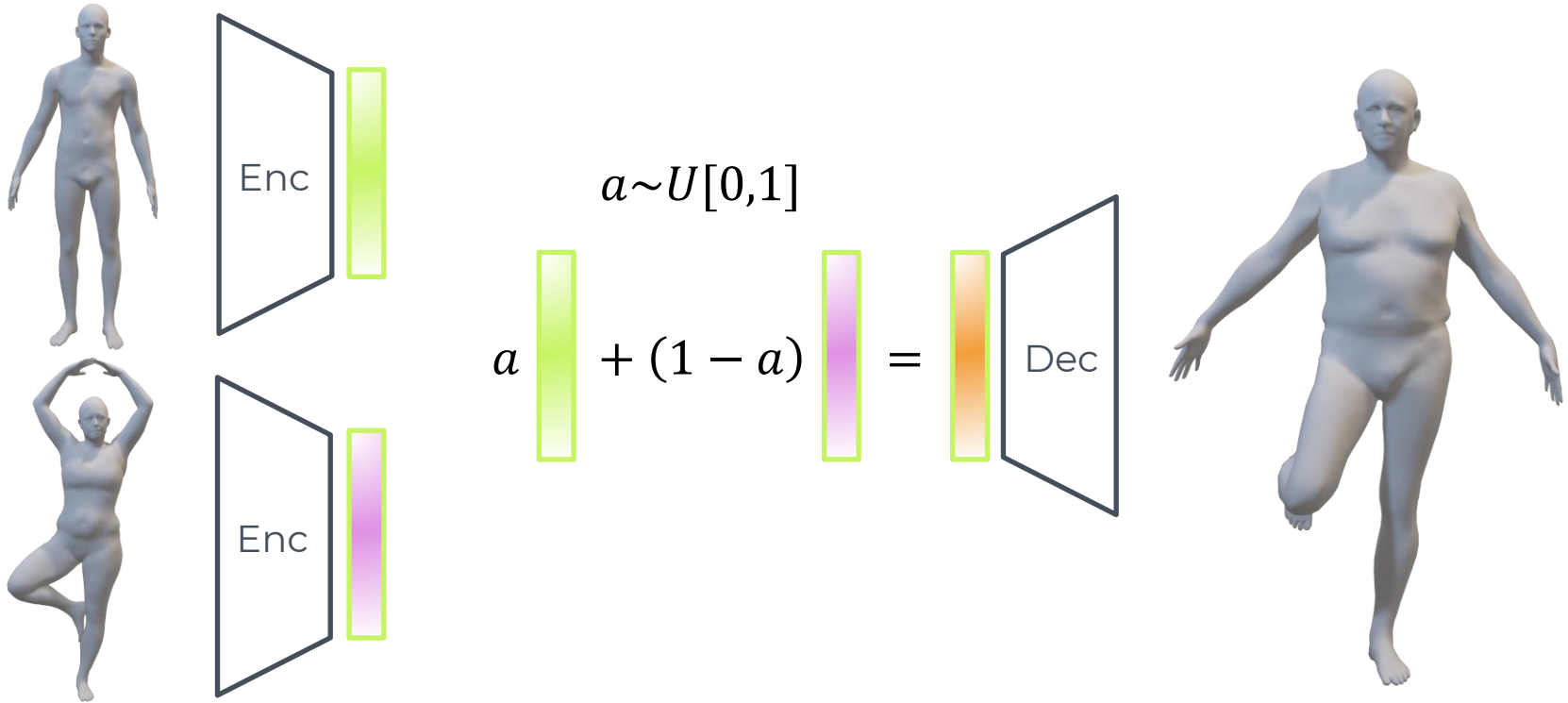
$$D_{eucl} = \frac{1}{n}\sum_{i}^{n} d_i$$

$$\widehat{D}_{eucl} = \frac{1}{n}\sum_{i}^{n} d_i$$

2° optimization

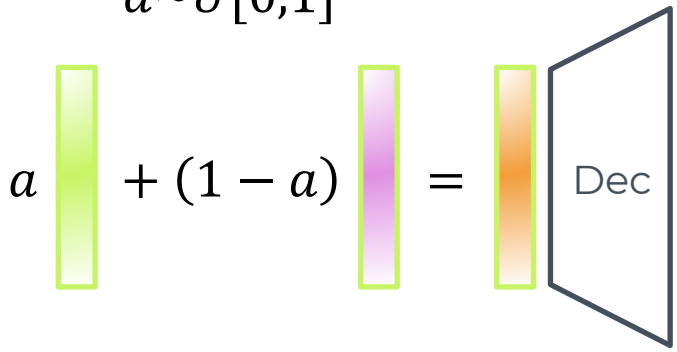


2° optimization



2° optimization

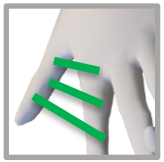
$$a \sim U[0,1]$$



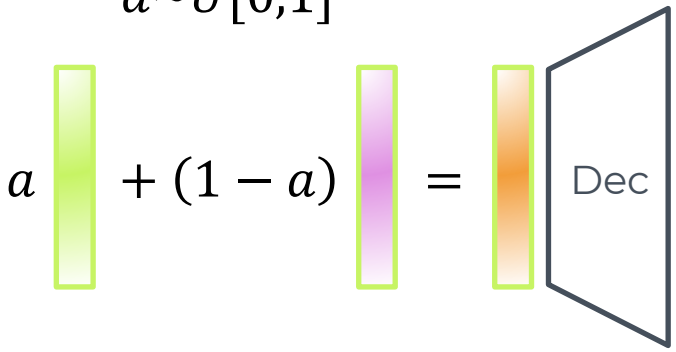
2° optimization

$$D_{loc1} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i$$

$$D_{loc2} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i$$



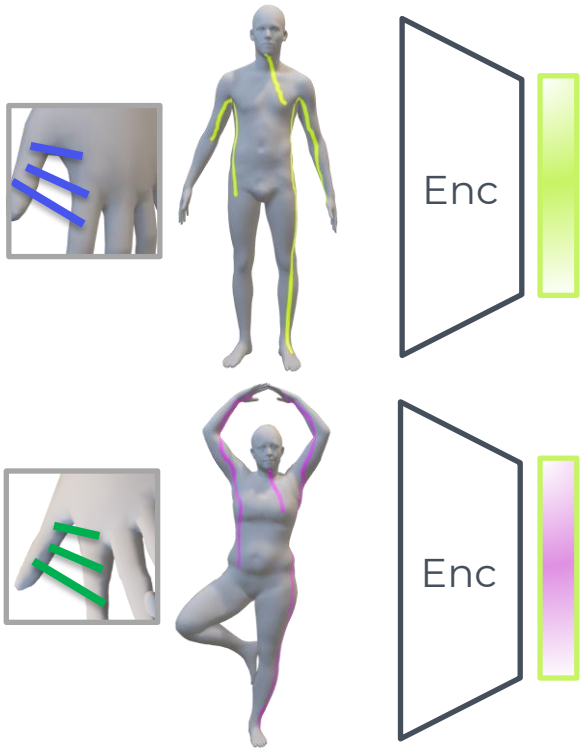
$$a \sim U[0,1]$$



2° optimization

$$D_{loc1} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i \qquad D_{geo1} = \frac{1}{n}\sum_{i}^{n} d_i$$

$$D_{loc2} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i \qquad D_{geo2} = \frac{1}{n}\sum_{i}^{n} d_i$$

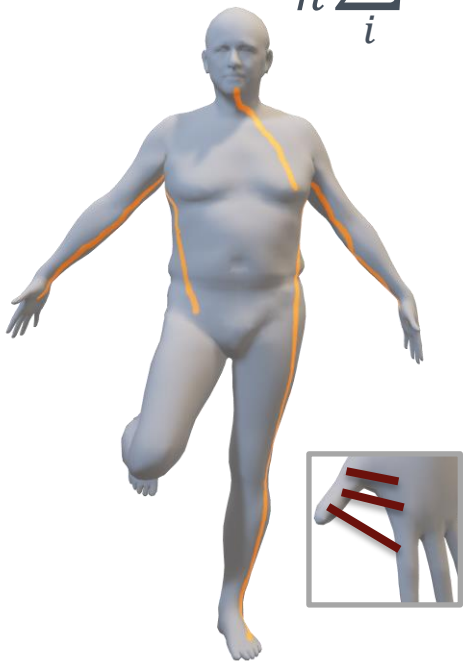


$$a \sim U[0,1]$$

$$a + (1-a) = \text{Dec}$$

$$D_{loc} = aD_{loc1} + (1-a)D_{loc2}$$

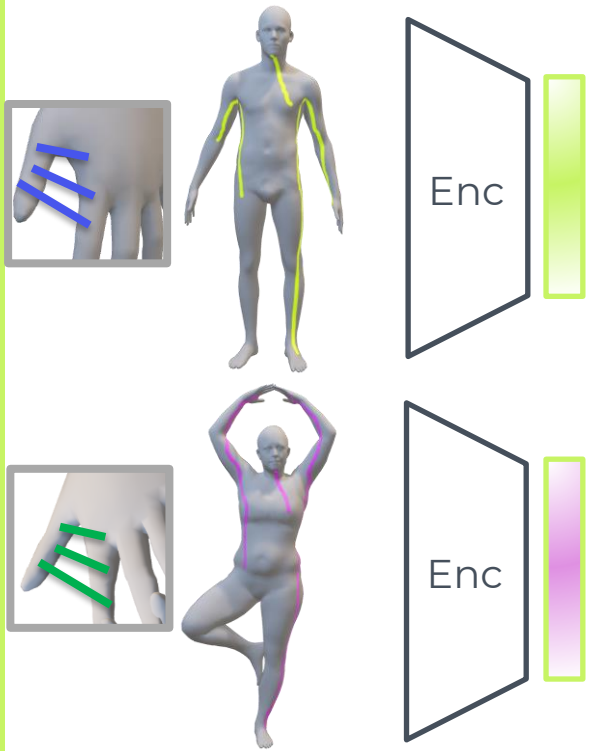
$$D_{geo} = aD_{geo1} + (1-a)D_{geo2}$$



2° optimization

$$L = \alpha\left(\frac{D_{loc} - \hat{D}_{loc}}{D_{loc}}\right)^2 + \beta\left(\frac{D_{geo} - \hat{D}_{geo}}{D_{geo}}\right)^2$$

$$\hat{D}_{loc} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i \qquad \hat{D}_{geo} = \frac{1}{n}\sum_{i}^{n} d_i$$

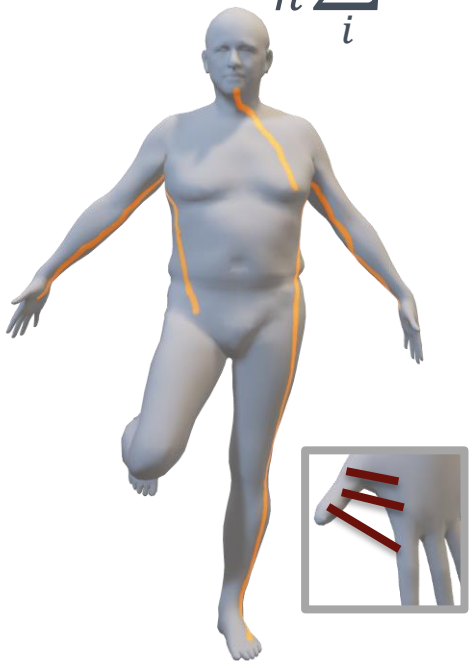


$a \sim U[0,1]$

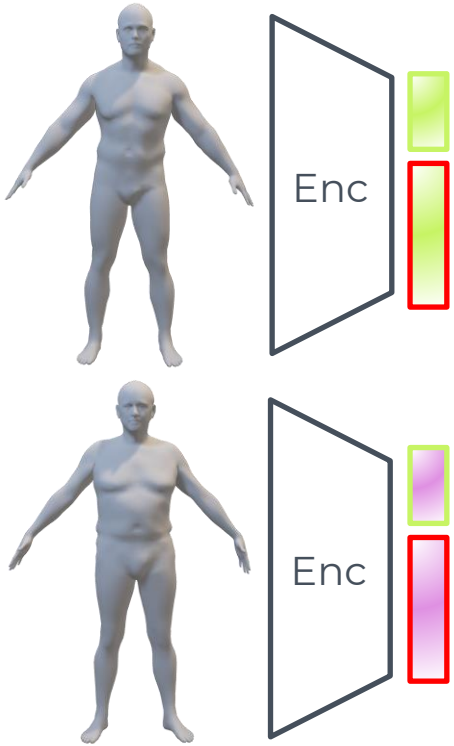
$$a + (1-a) = \text{Dec}$$

$$D_{loc} = aD_{loc1} + (1-a)D_{loc2}$$

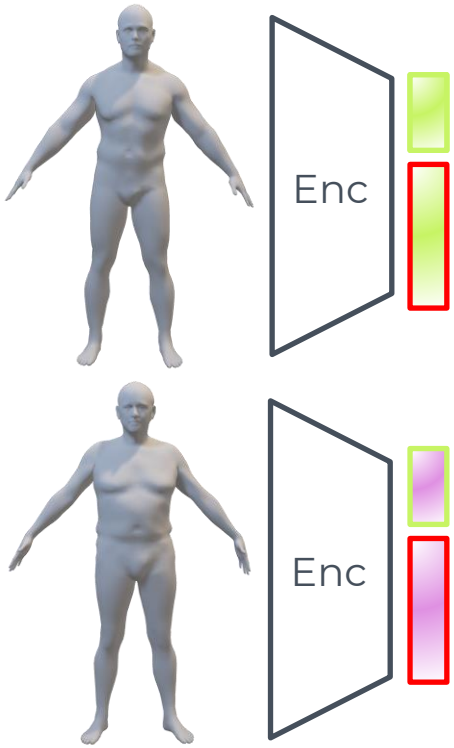
$$D_{geo} = aD_{geo1} + (1-a)D_{geo2}$$



3° optimization



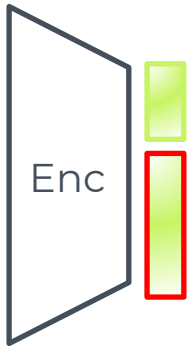
3° optimization



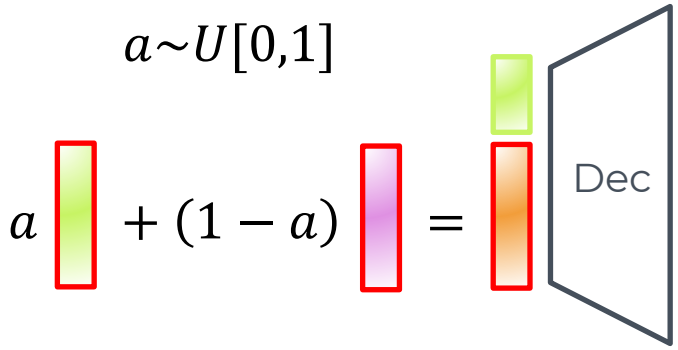
$$a \sim U[0,1]$$

$$a + (1-a) = \text{Dec}$$

3° optimization



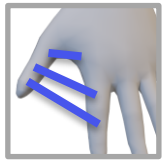
$$a \sim U[0,1]$$

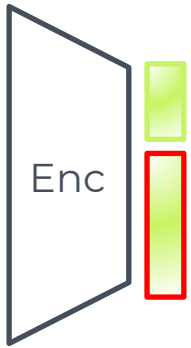


3° optimization

$$D_{loc1} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i$$

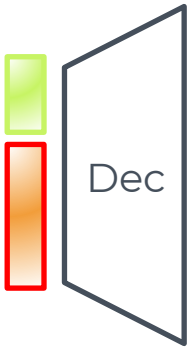
$$D_{loc2} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i$$





$$a \sim U[0,1]$$

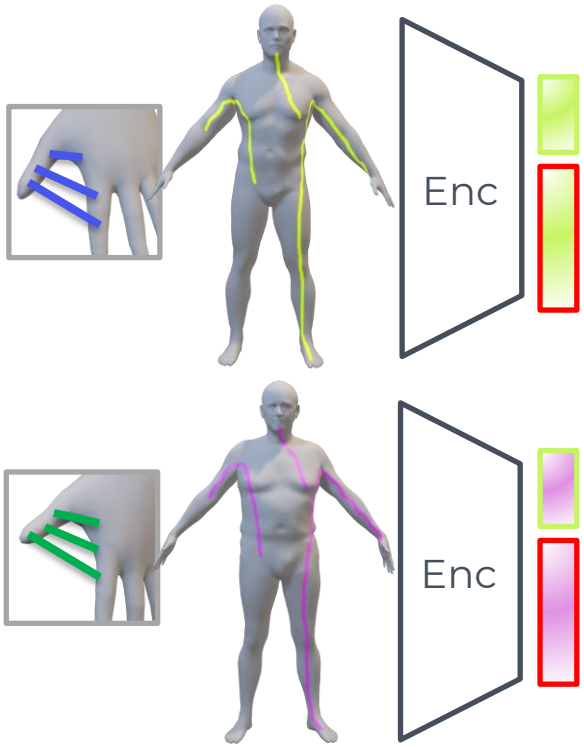
$$a\,[\;] + (1-a)\,[\;] = [\;]$$



3° optimization

$$D_{loc1} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i \qquad D_{geo1} = \frac{1}{n}\sum_{i}^{n} d_i$$

$$D_{loc2} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i \qquad D_{geo2} = \frac{1}{n}\sum_{i}^{n} d_i$$

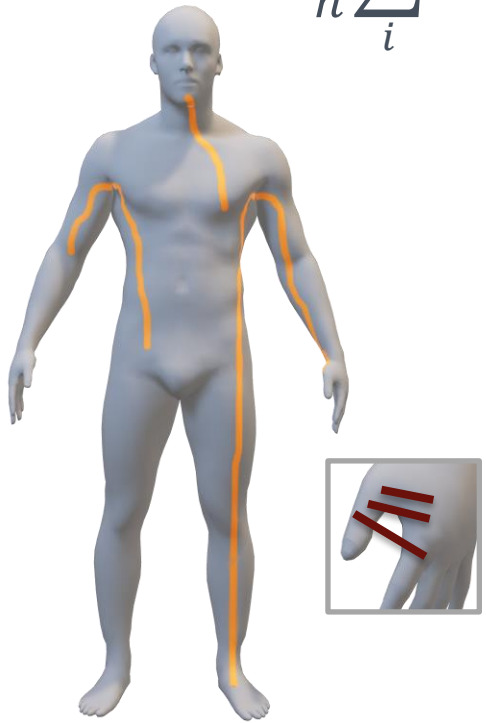


$$a \sim U[0,1]$$

$$a + (1-a) = \text{Dec}$$

$$D_{loc} = aD_{loc1} + (1-a)D_{loc2}$$

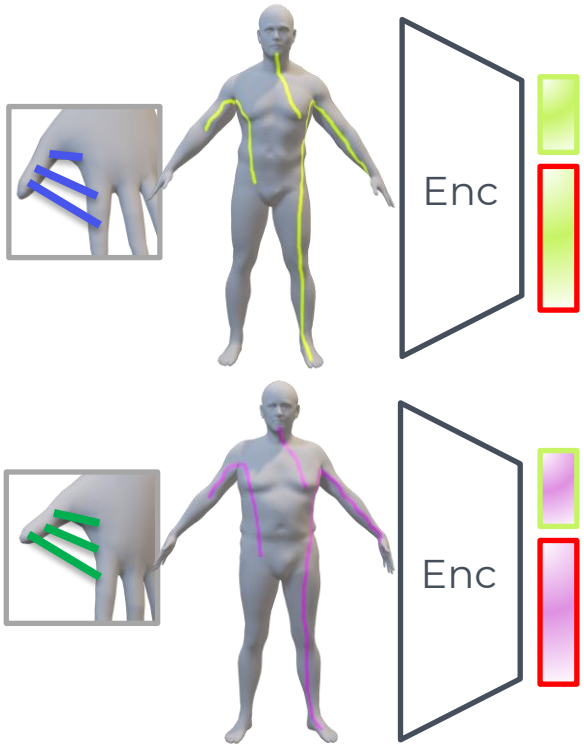
$$D_{geo} = aD_{geo1} + (1-a)D_{geo2}$$



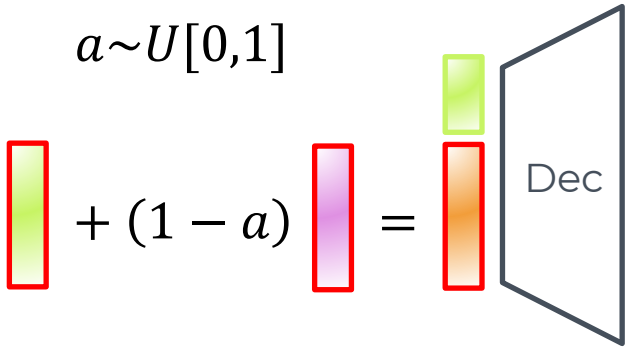
3° optimization

$$L = \alpha\left(\frac{D_{loc} - \hat{D}_{loc}}{D_{loc}}\right)^2 + \beta\left(\frac{D_{geo} - \hat{D}_{geo}}{D_{geo}}\right)^2$$

$$\hat{D}_{loc} = \frac{1}{m}\sum_{k}^{m}\sum_{i} d_i \qquad \hat{D}_{geo} = \frac{1}{n}\sum_{i}^{n} d_i$$



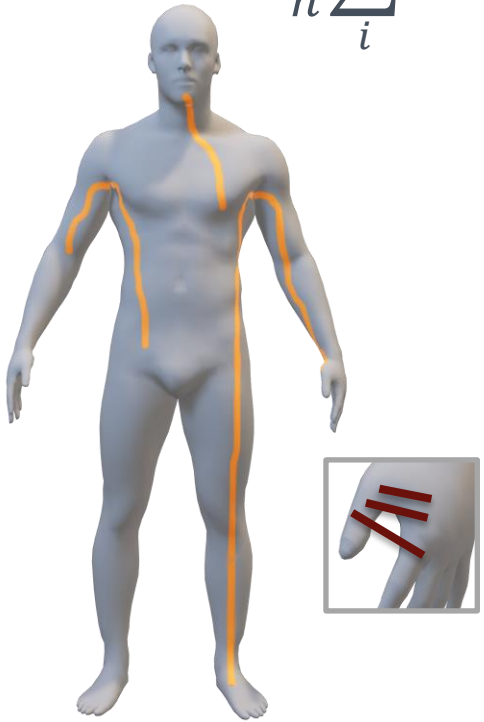
$a \sim U[0,1]$



$$a + (1-a) = $$

$$D_{loc} = aD_{loc1} + (1-a)D_{loc2}$$

$$D_{geo} = aD_{geo1} + (1-a)D_{geo2}$$



Recap

1. **Global euclidean** distortion
2. Linear **interpolation** of data points
3. Pose-Style **disentanglement**

Recap

1. **Global euclidean** distortion
2. Linear **interpolation** of data points
3. Pose-Style **disentanglement**

Recap

1. **Global euclidean** distortion
2. Linear **interpolation** of data points
3. Pose-Style **disentanglement**

Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodolà

"LIMP: Learning Latent Shape Representations with Metric Preservation Priors" Oral at ECCV 2020

Synthesizing **novel shapes**

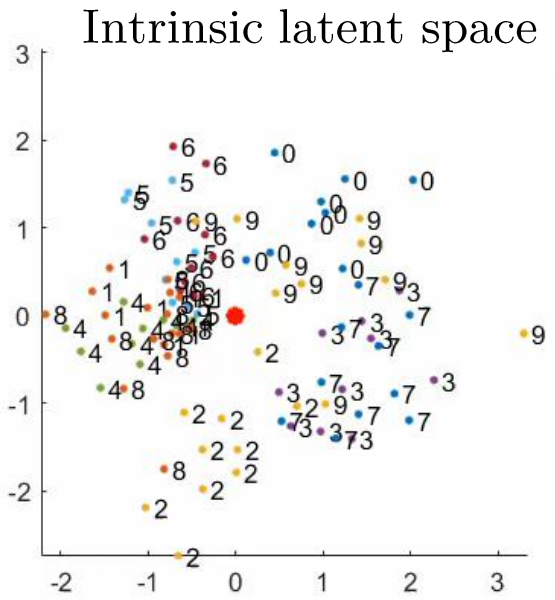
Results

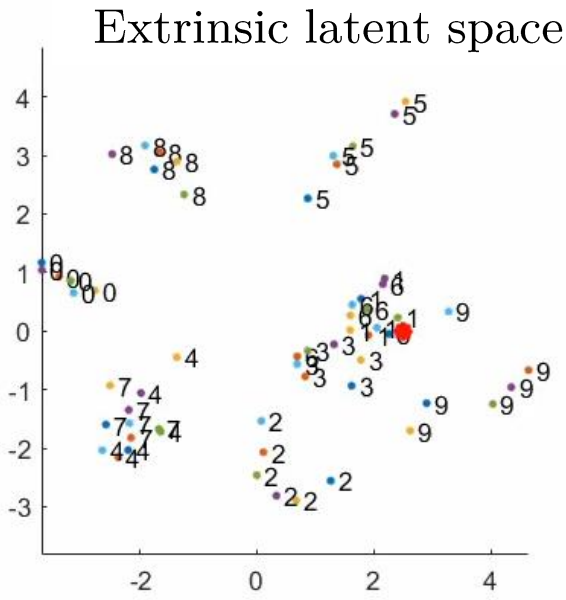
Using metric priors during training reduces distorsion on the generated shapes

Using metric priors during training reduces distorsion on the generated shapes

Pose and style are disentangled, yielding a more expressive generative model

Pose and style are disentangled, yielding a more expressive generative model





3.

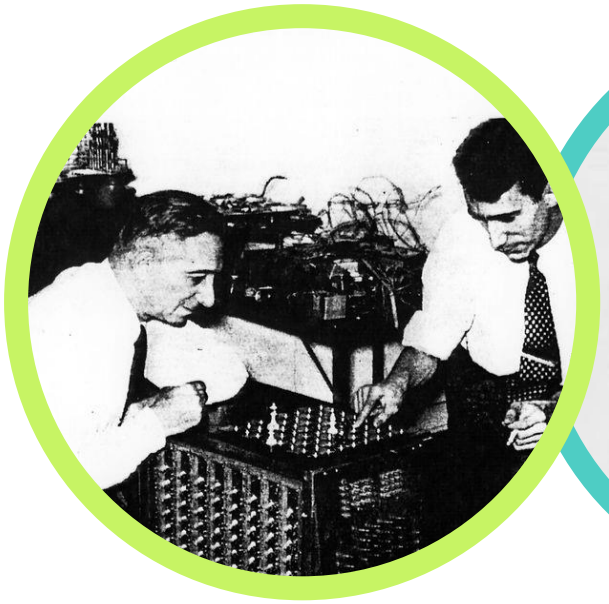
Search

The best hypothesis

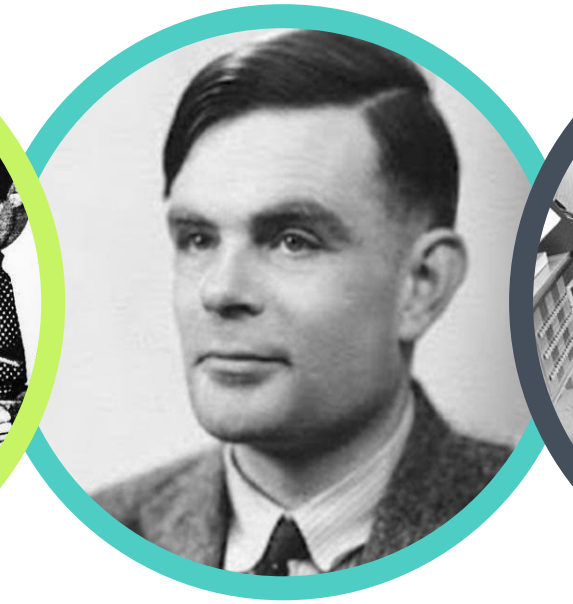
OLIVAW:
Mastering Othello without Human Knowledge, nor a Penny.


ACCESS ALL AREAS
AlphaGo
Lee Sedol

Why games?



Shannon
1949



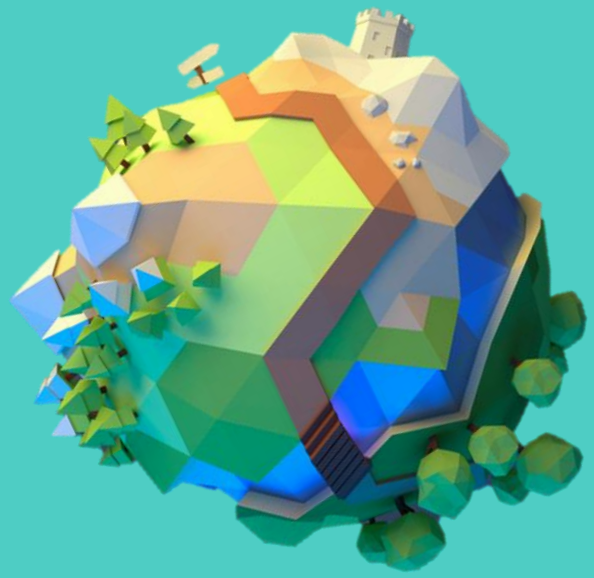
Turing
1950

McCarthy
1956

Shannon
1949

A satisfactory solution of [chess] *will act as a wedge in attacking other problems of a similar nature and of greater significance.*

Programming a Computer for Playing Chess
Philosophical Magazine, Vol 41, No. 314

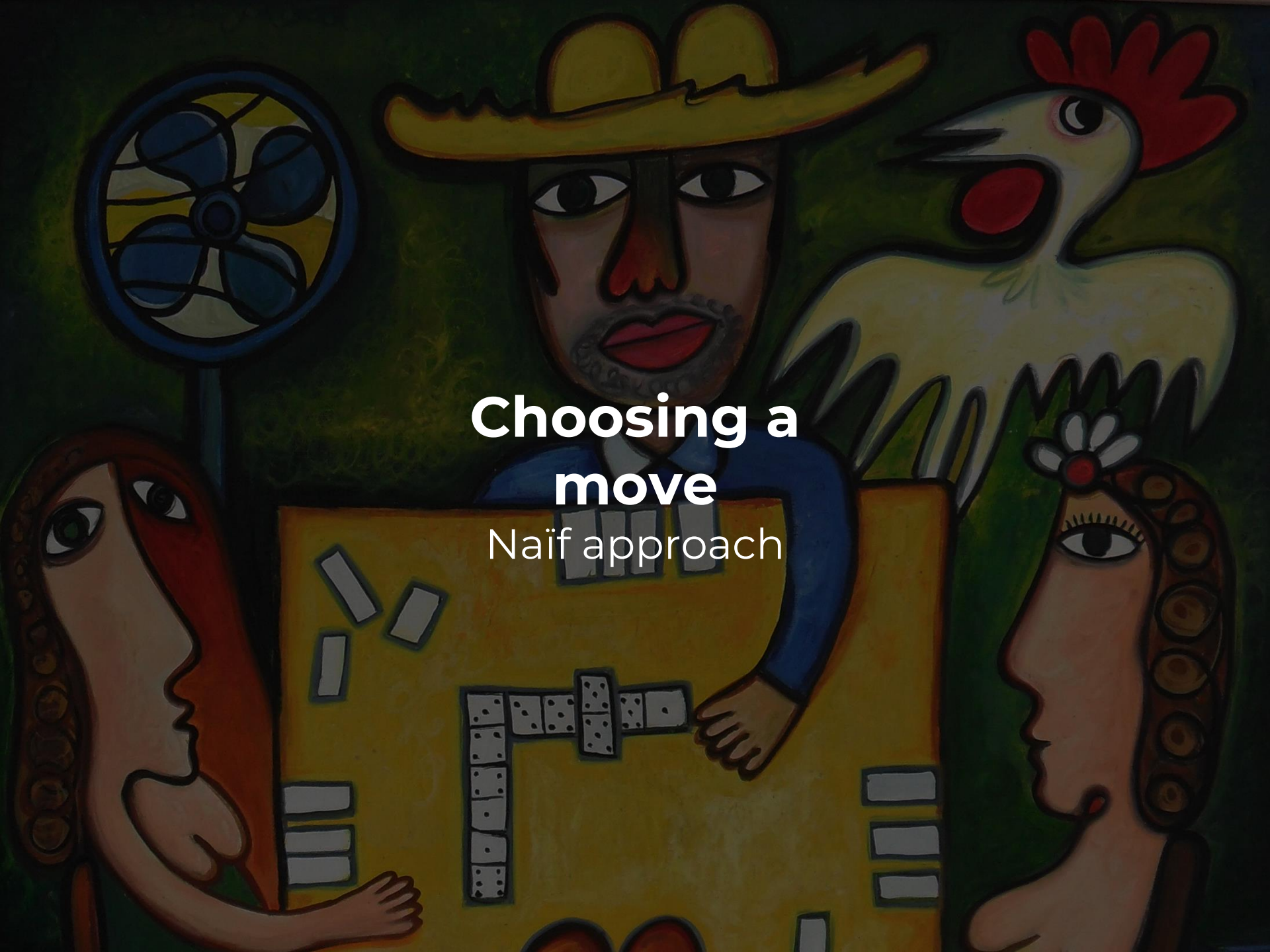


Games as a micro world

Small set of rules

Clear objectives

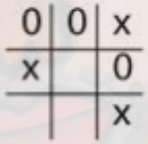
Still interesting complexity



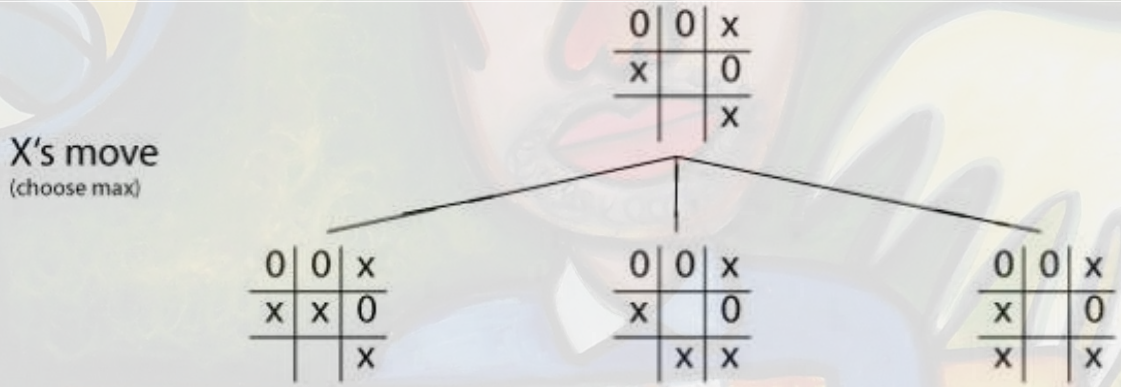
Choosing a move

Naïf approach

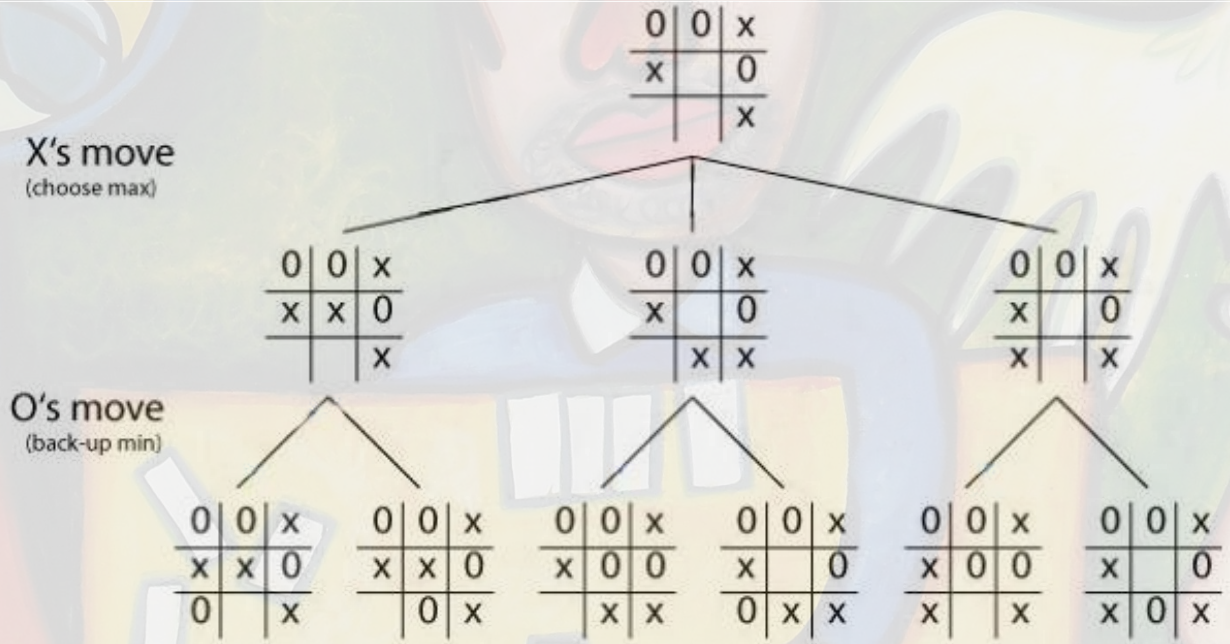
The Minimax algorithm



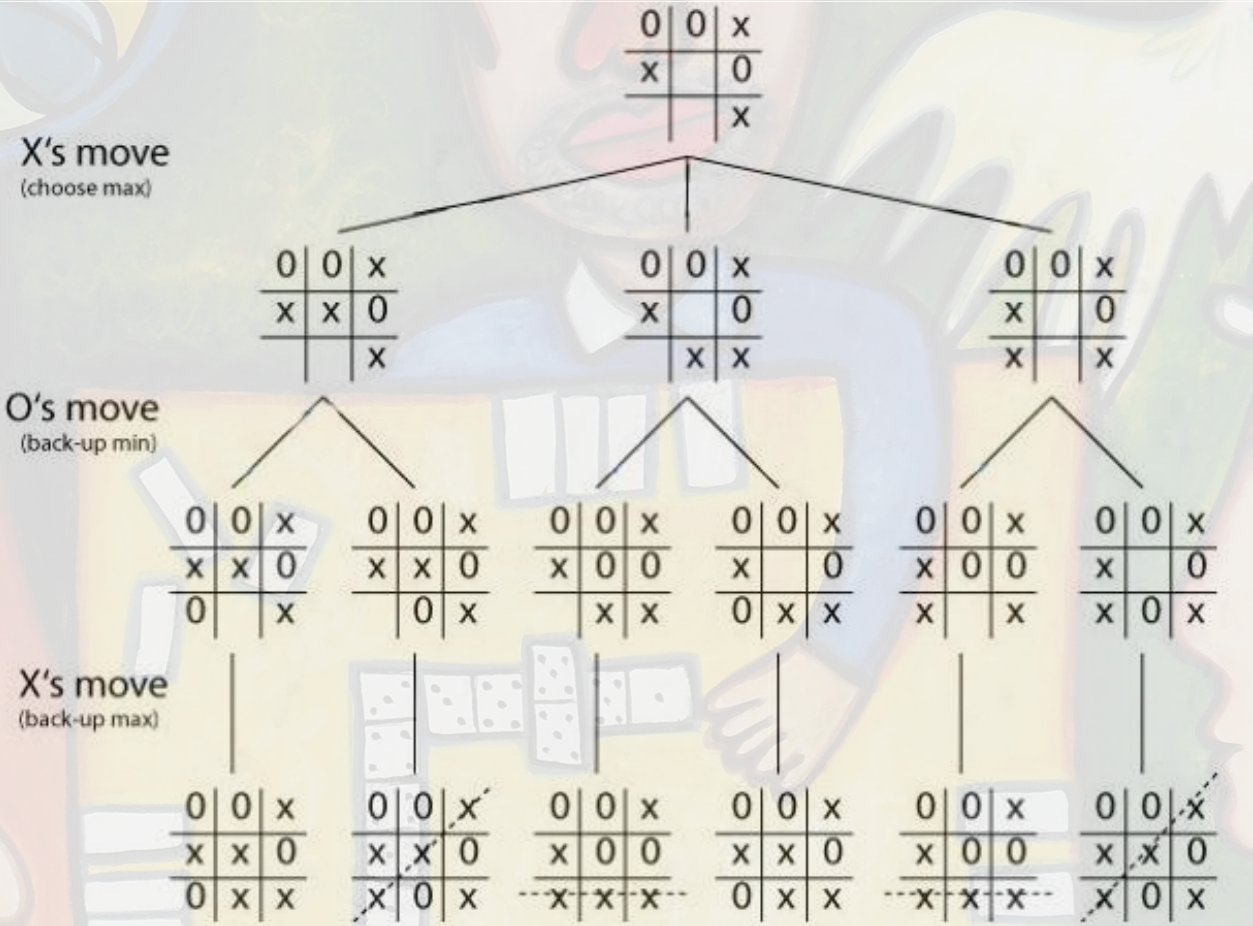
Minimax algorithm



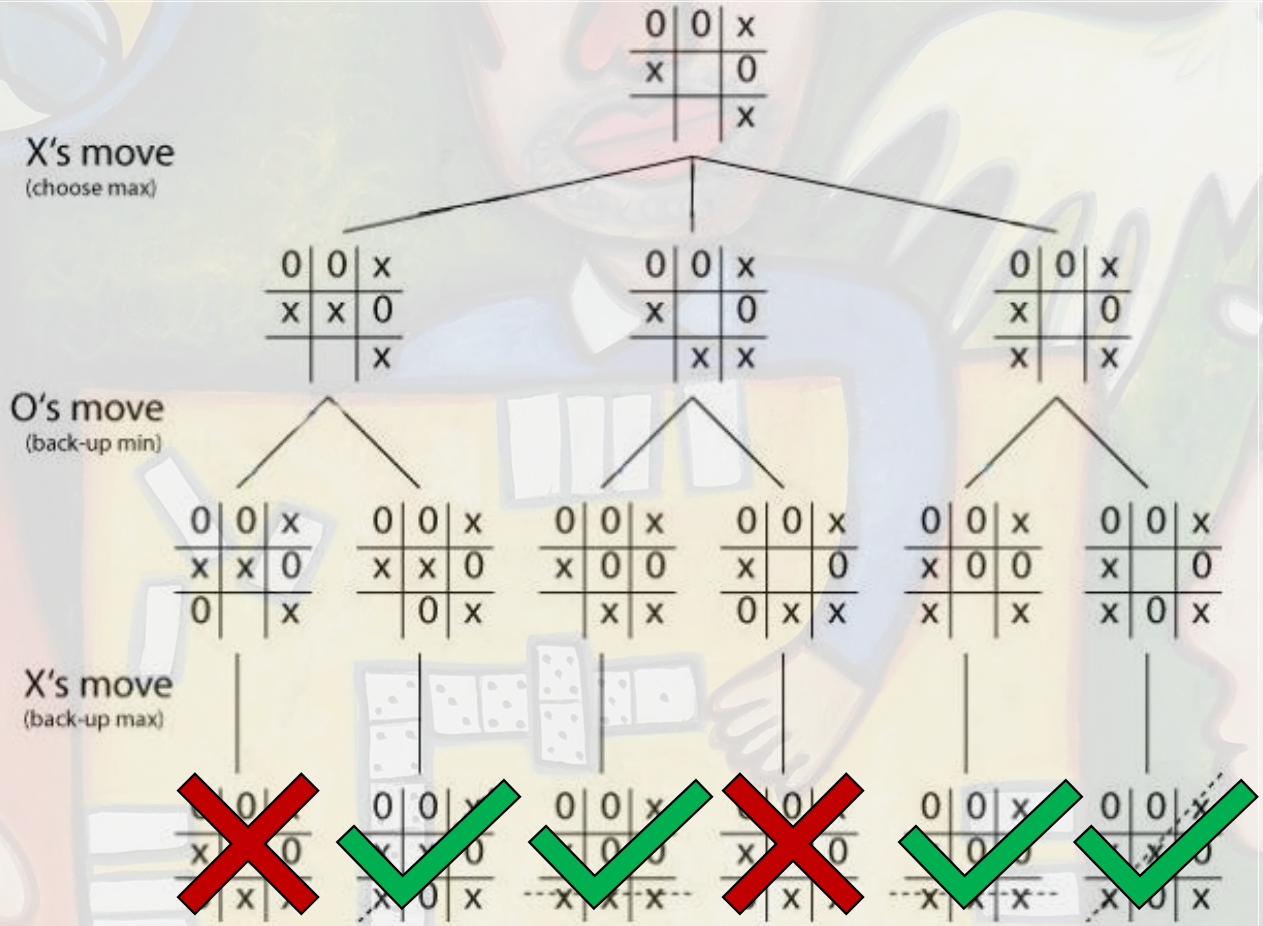
Minimax algorithm



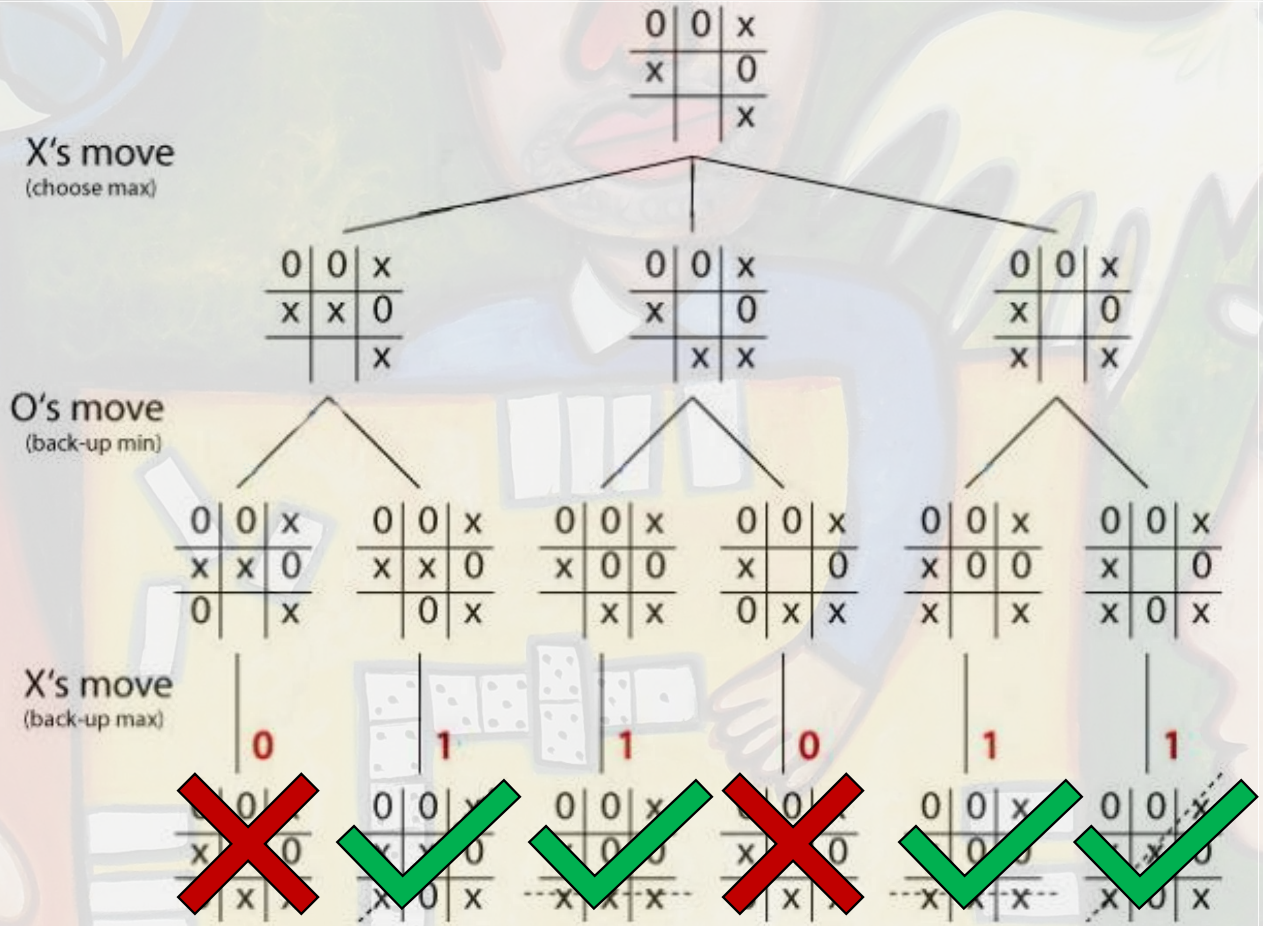
Minimax algorithm



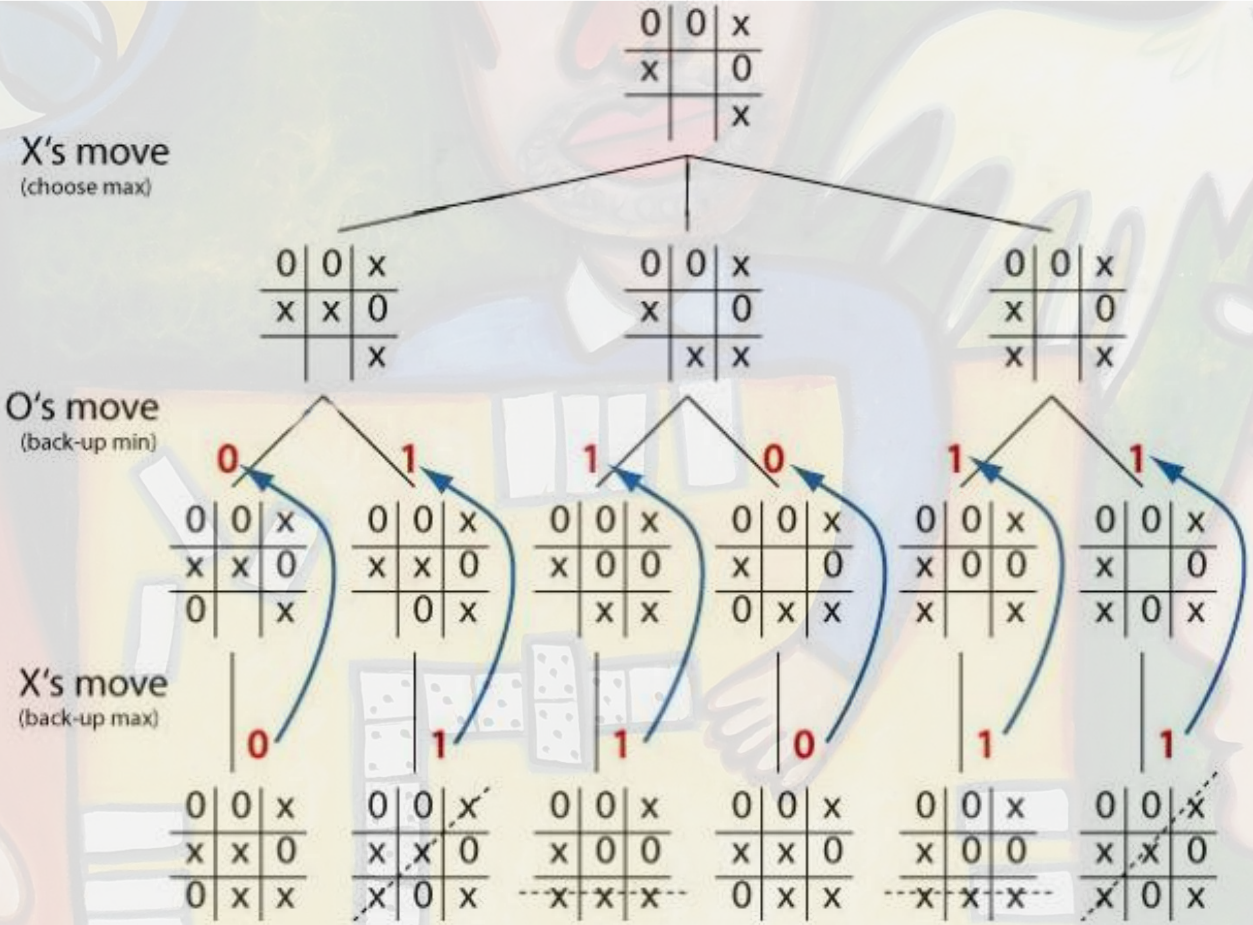
Minimax algorithm



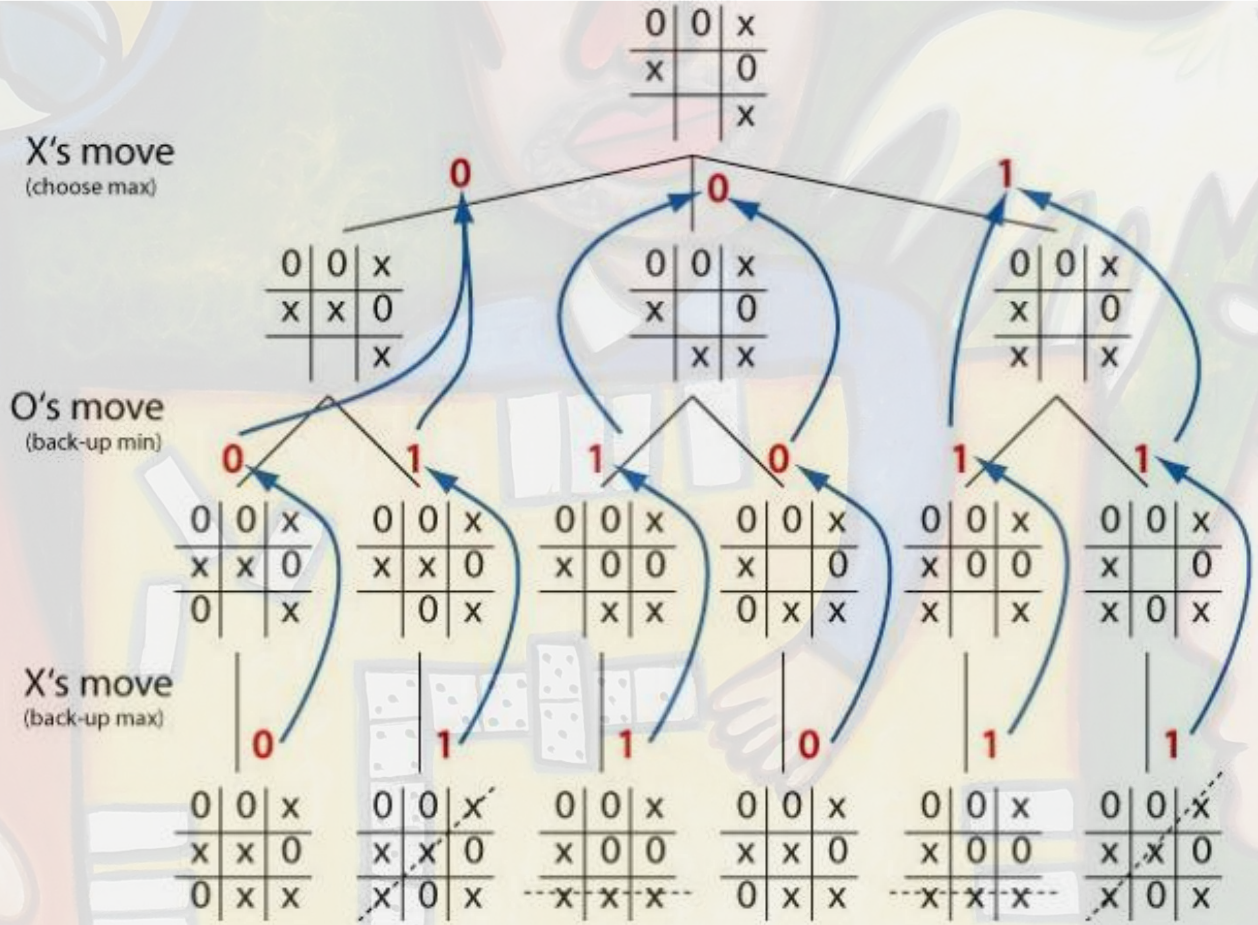
Minimax algorithm



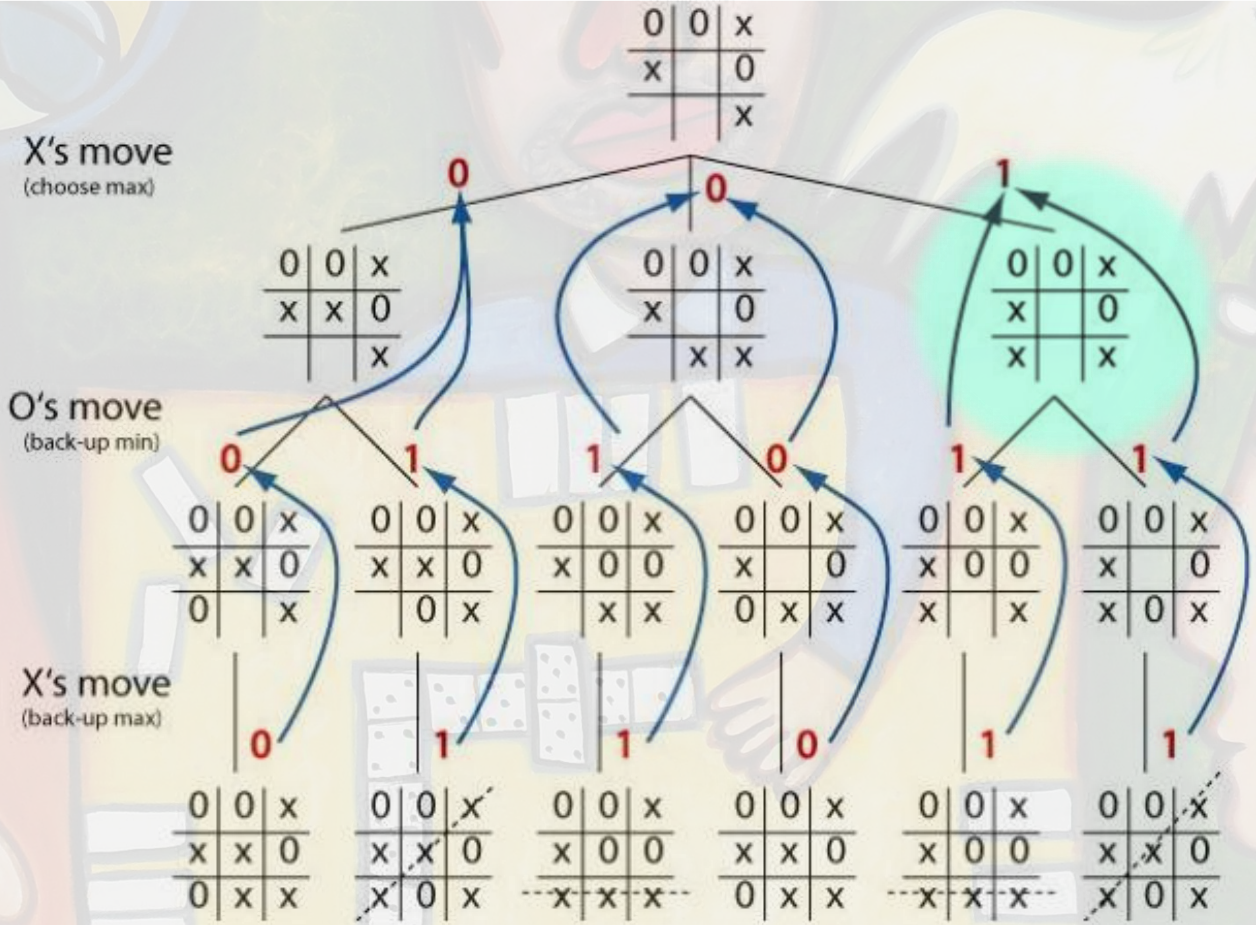
Minimax algorithm



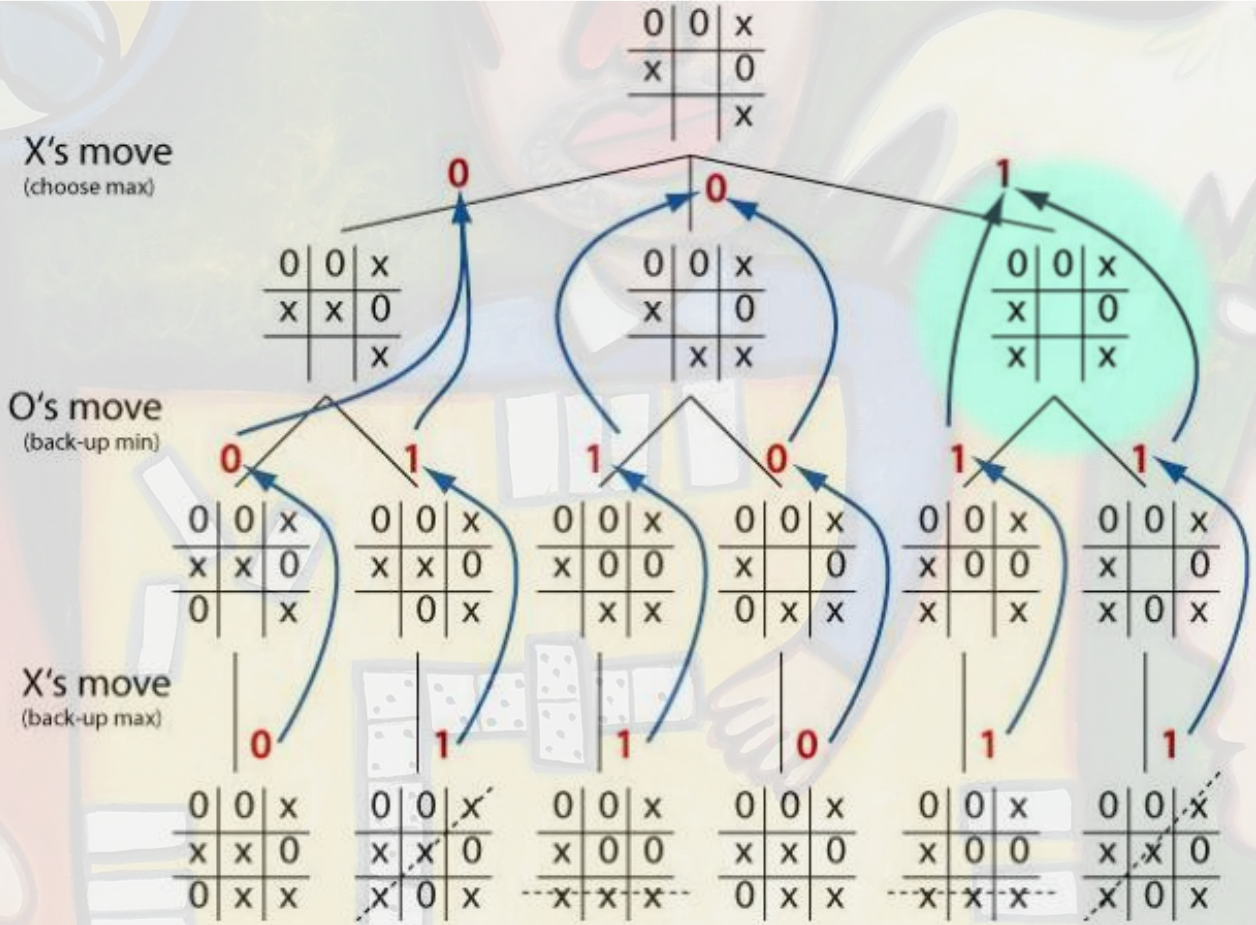
Minimax algorithm



Minimax algorithm



Minimax algorithm $O(b^d)$

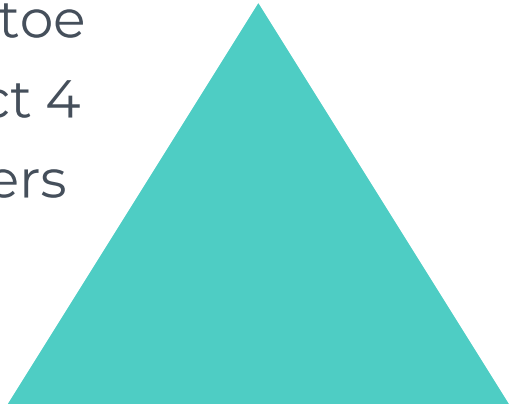


Looking into the future

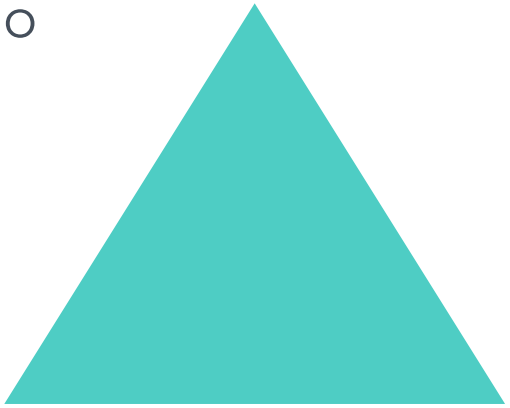
Tic-tac-toe

Connect 4

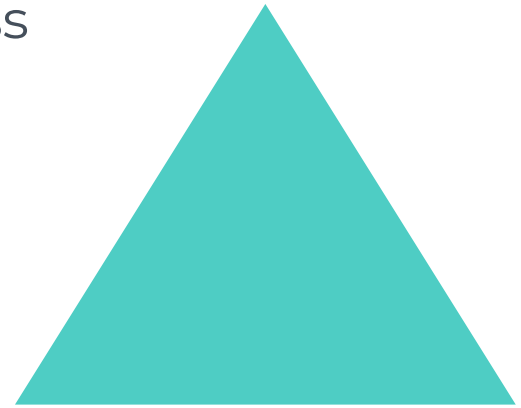
Checkers



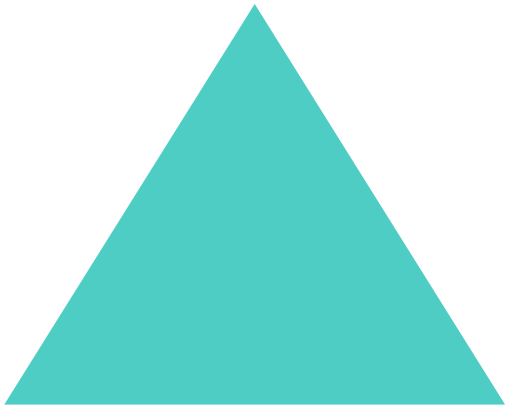
Othello



Chess

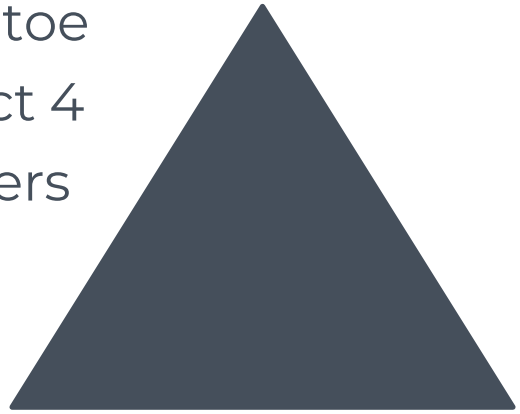


Go



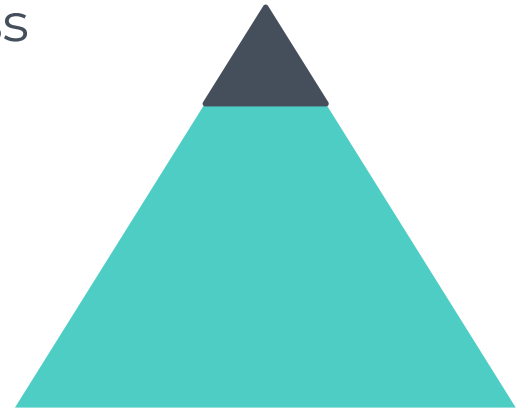
Looking into the future

Tic-tac-toe
Connect 4
Checkers

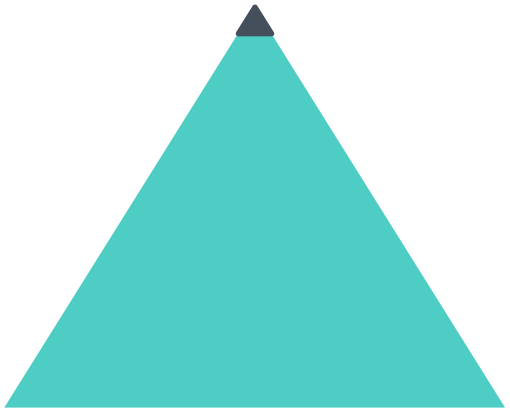


Othello

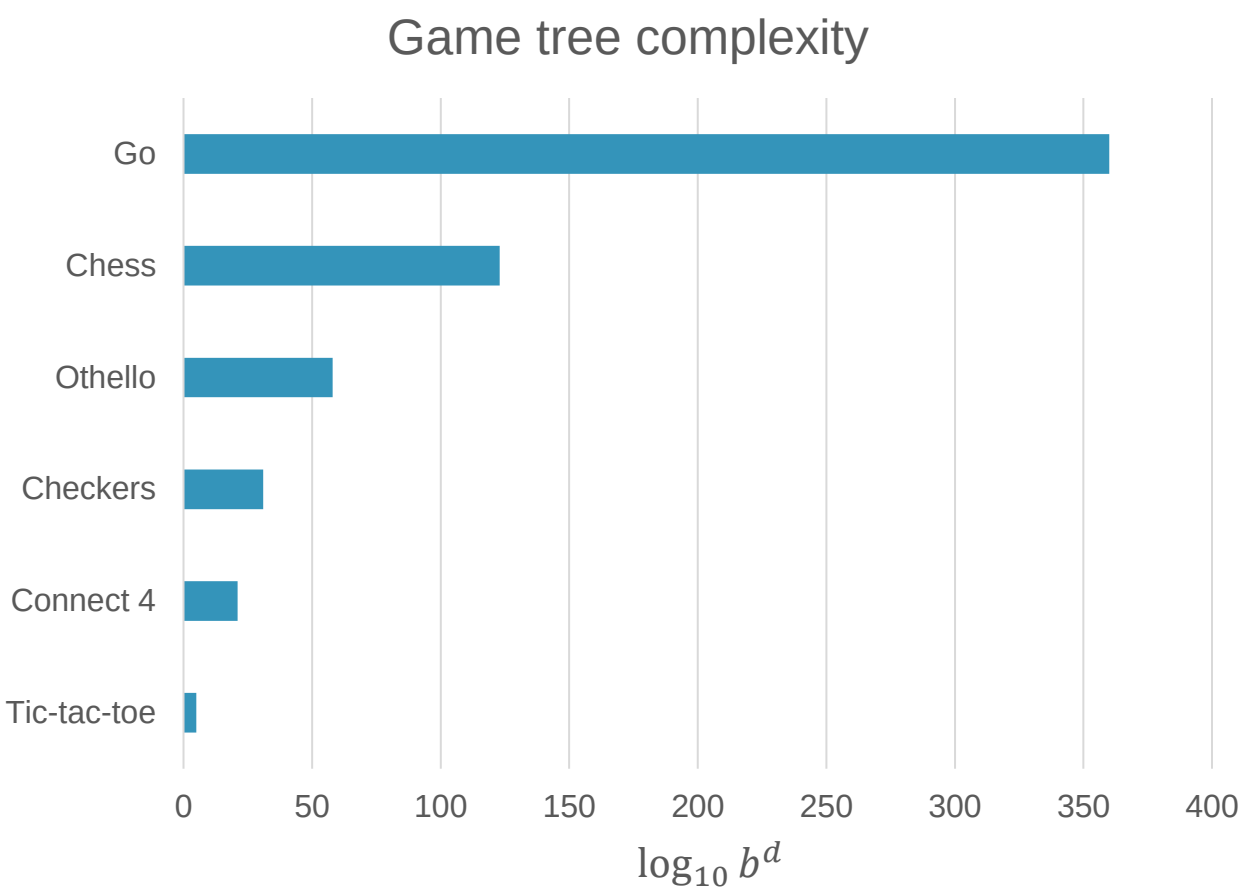
Chess



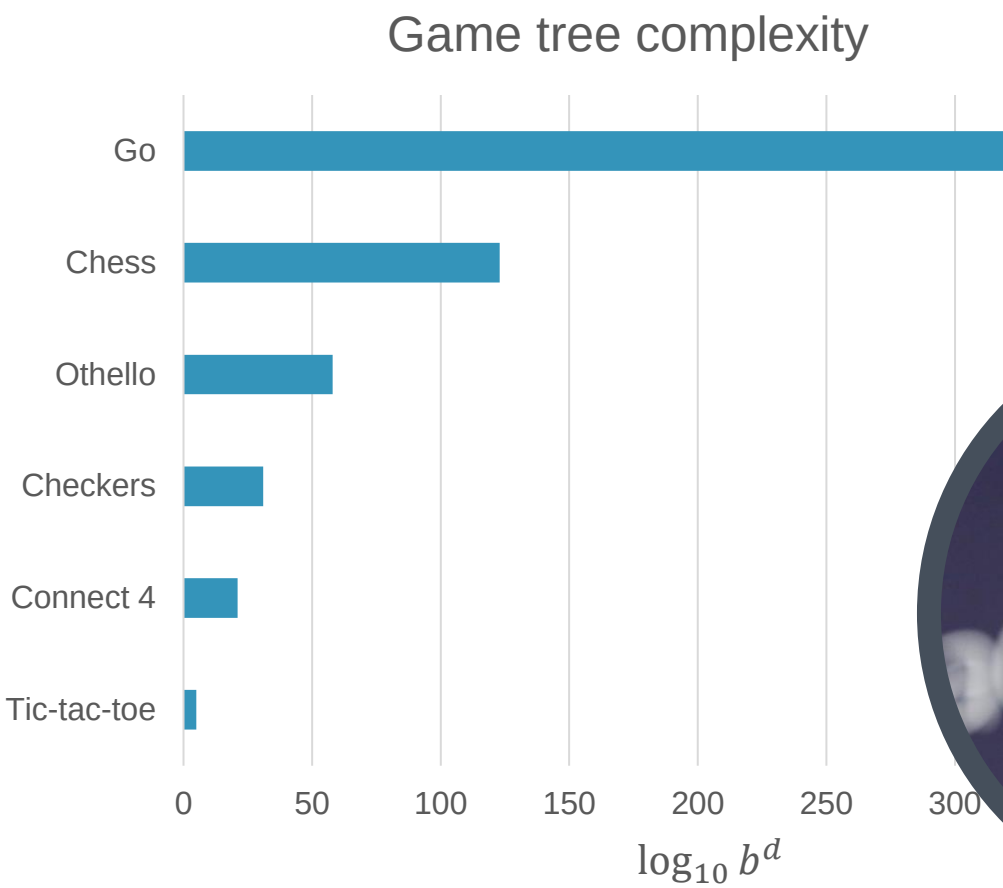
Go



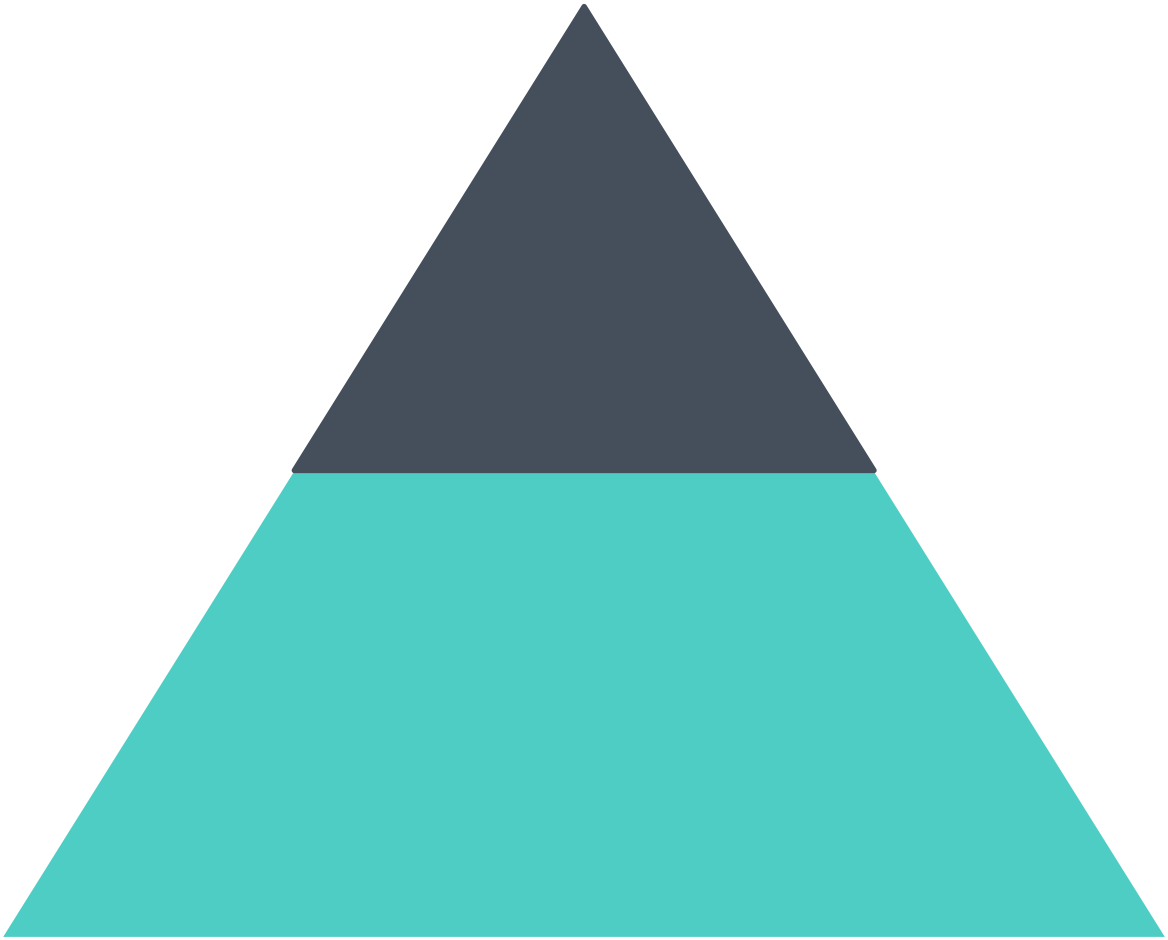
Exploring the full tree is impossible



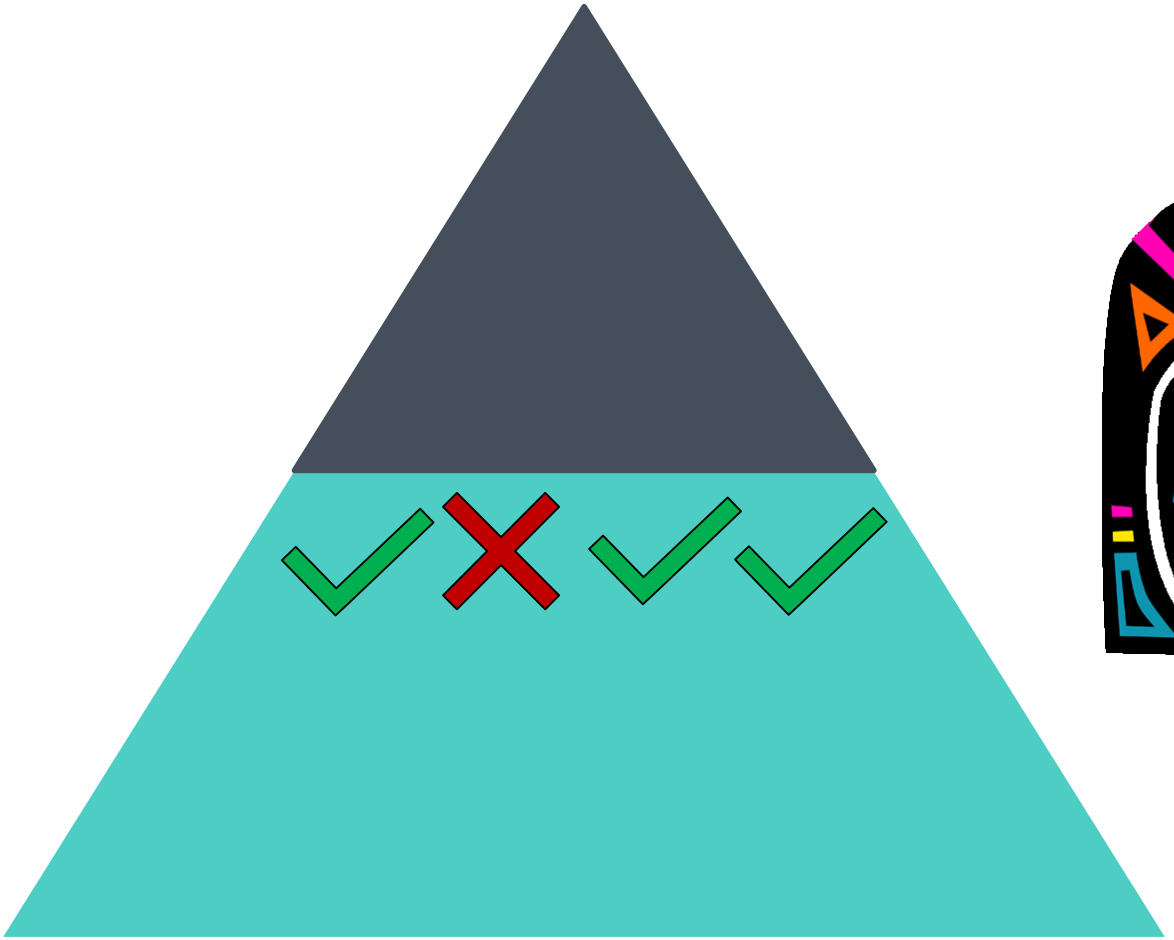
Exploring the full tree is impossible



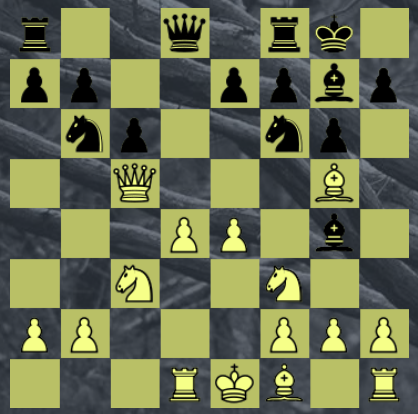
What does the future hold?



What does the future hold?




Oracle
to evaluate
intermediate positions



Oracle

to evaluate
intermediate positions

-0,42


Oracle
using knowledge of the game

DeepBlue

1997

AlphaGo

2016

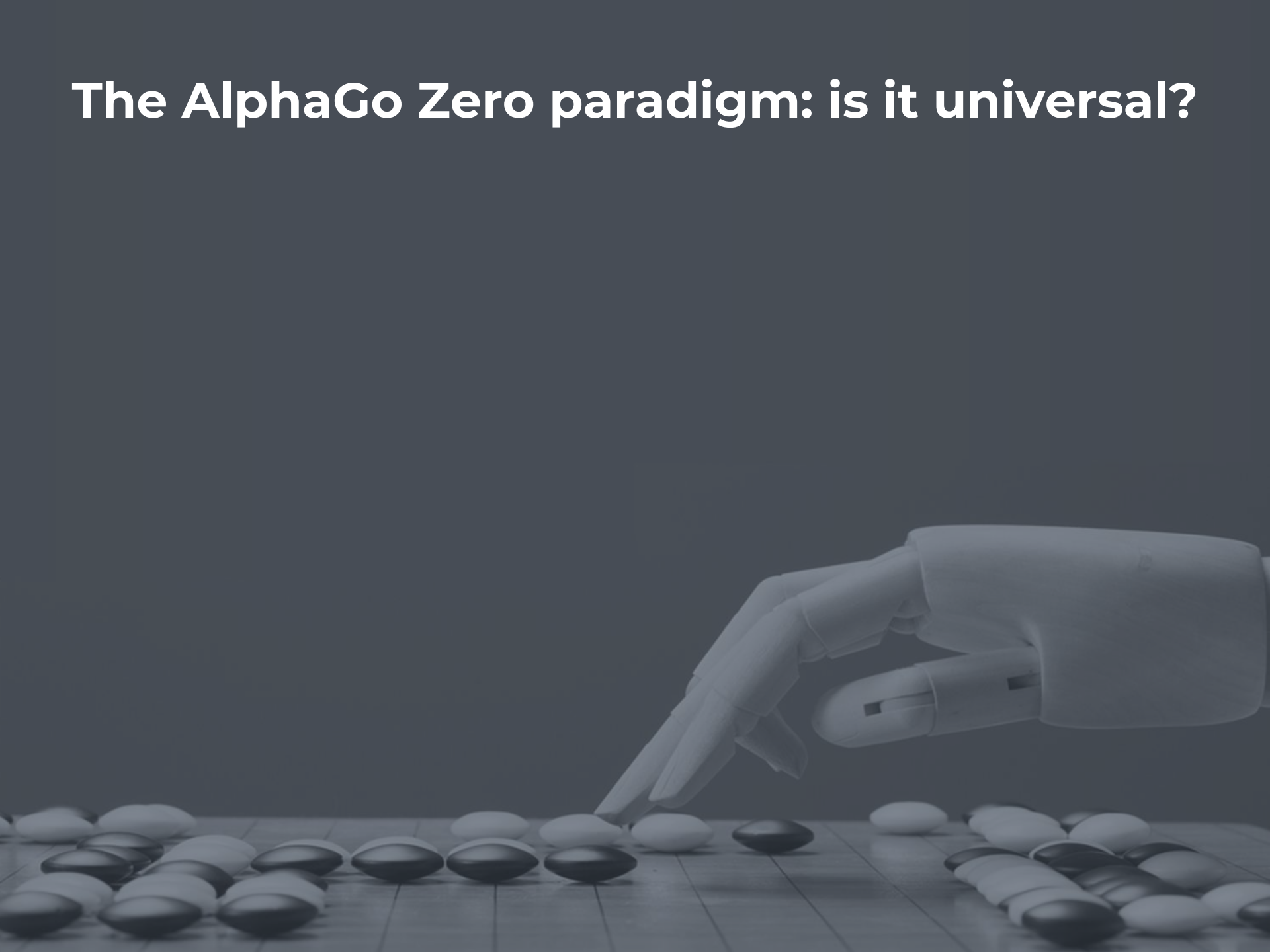
AlphaGo Zero

2017

AlphaGo Zero

2017

The AlphaGo Zero paradigm: is it universal?



The AlphaGo Zero paradigm: is it universal?

Does it scale DOWN in terms of resources?

The AlphaGo Zero paradigm: is it universal?

Does it scale DOWN in terms of resources?

Can we reach superhuman strength at Othello with the same paradigm and "normal" computing power?

AlphaGo Zero		Our solution
5000	TPUs	**≈1**

AlphaGo Zero		Our solution
5000	TPUs	**≈1**
40 (3)	Days of training	**30**

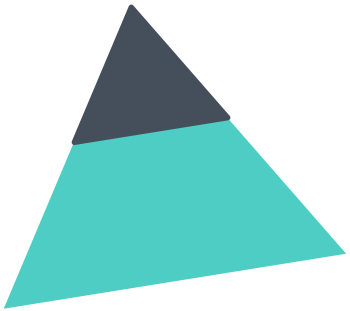
AlphaGo Zero		Our solution
5000	TPUs	**≈1**
40 (3)	Days of training	**30**
$ 25 Million	Estimated Hardware cost	**$ 500**

AlphaGo Zero		Our solution
5000	TPUs	**≈1**
40 (3)	Days of training	**30**
$ 25 Million	Estimated Hardware cost	**$ 500**
	Budget	**0**

AlphaGo Zero		Our solution
5000	TPUs	**≈1**
40 (3)	Days of training	**30**
$ 25 Million	Estimated Hardware cost	**$ ~~[illegible]~~0**
	Budget	**0**

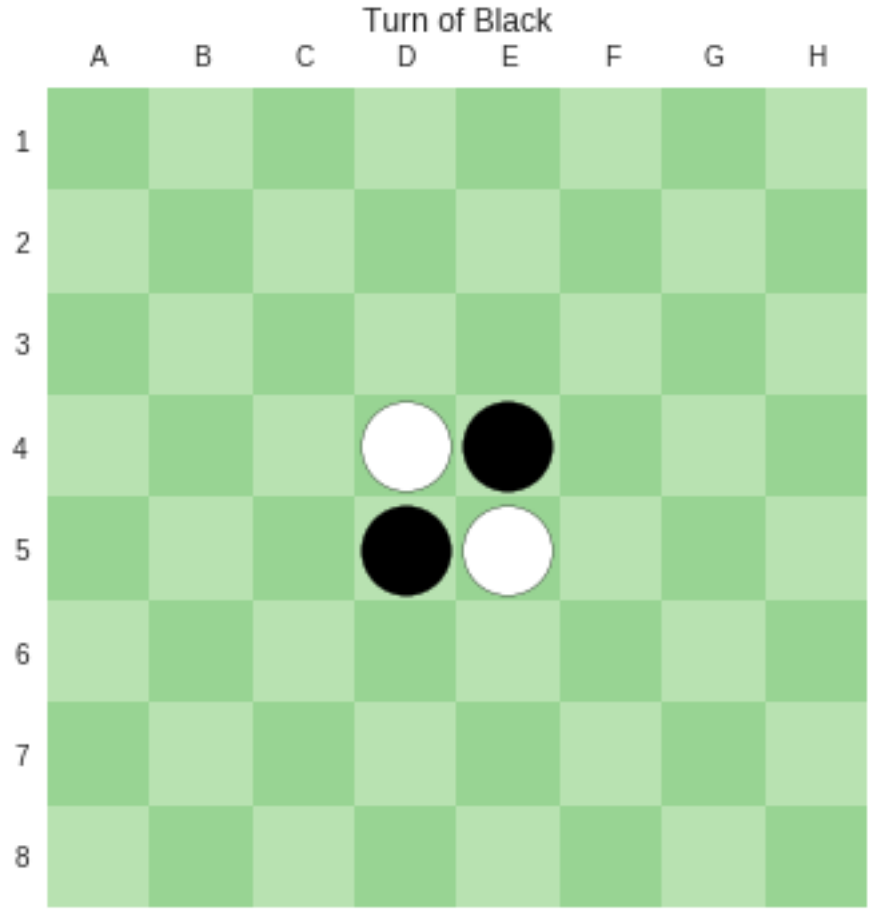
Why Othello?

Simple
but still
interesting



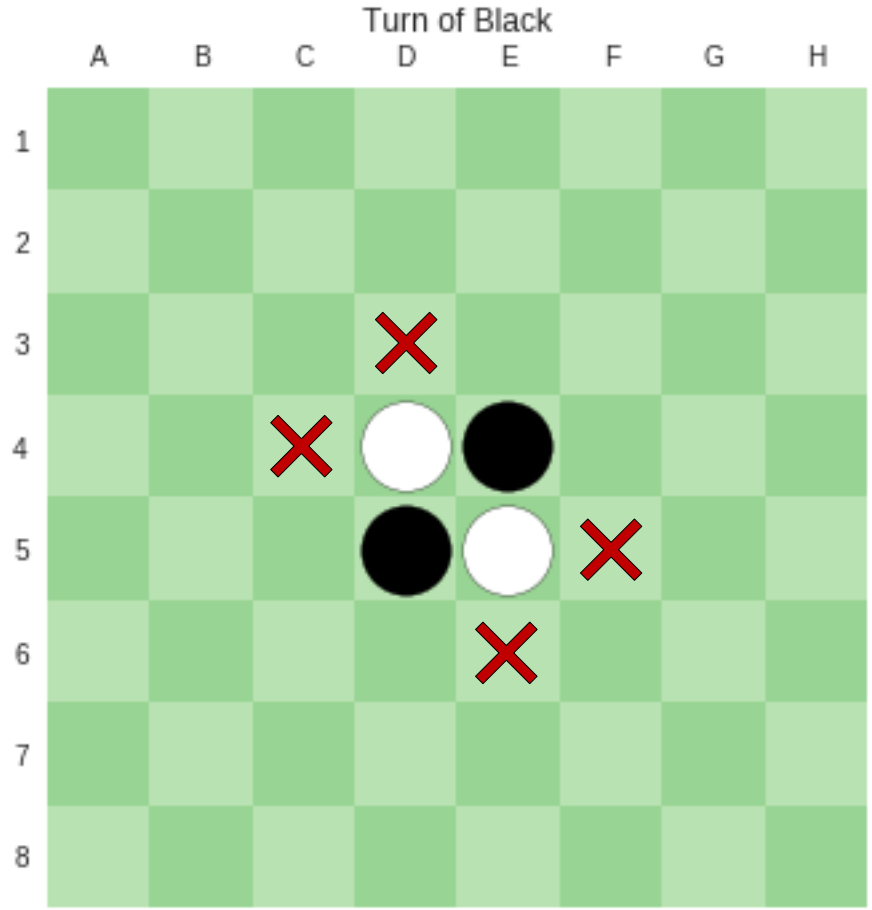
Why Othello?

Simple
but still
interesting



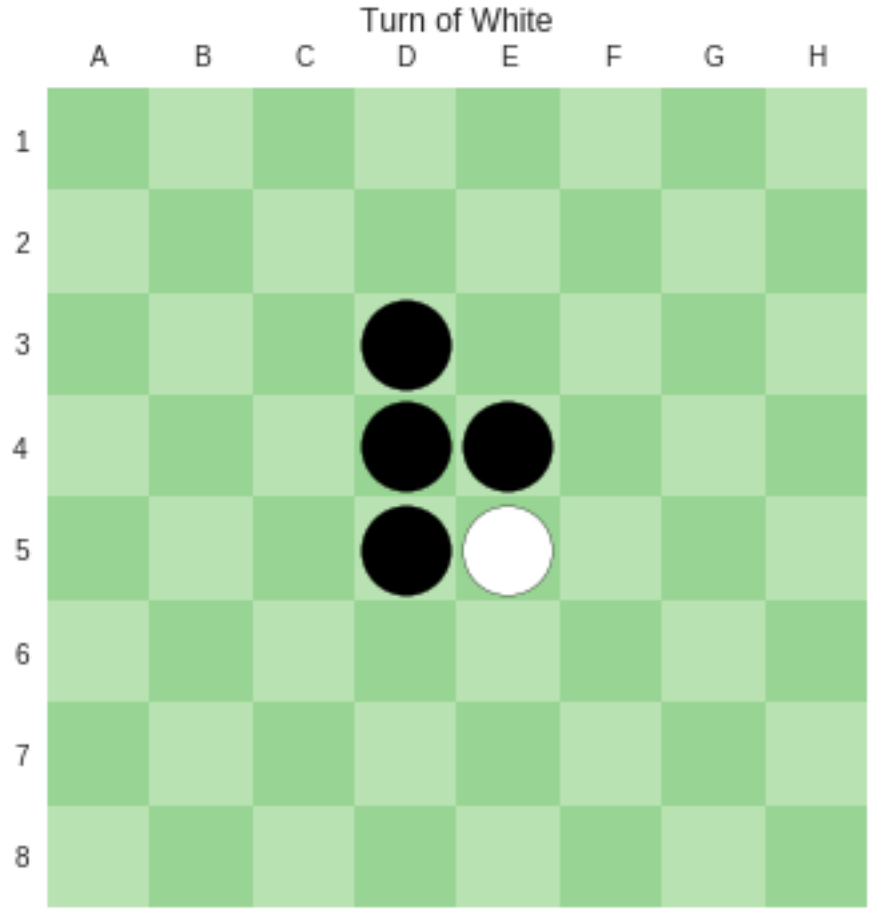
Why Othello?

Simple
but still
interesting



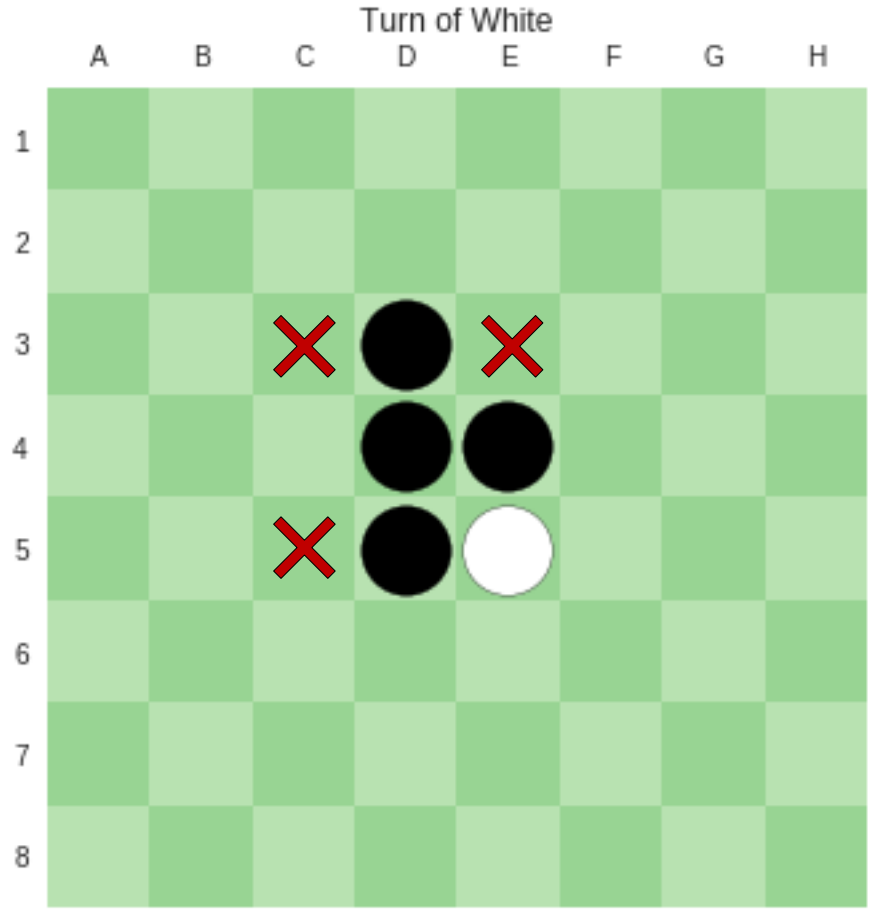
Why Othello?

Simple but still interesting



Why Othello?

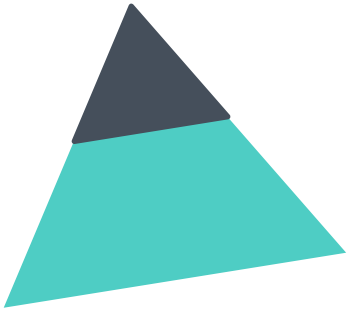
Simple but still interesting

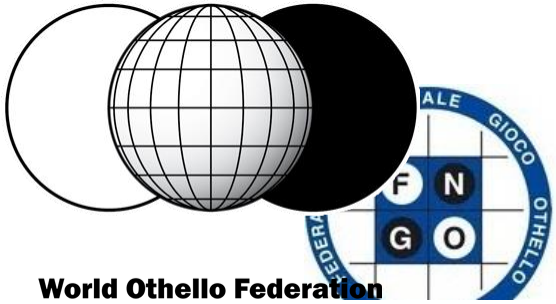


Why Othello?

Simple but still interesting

Well known



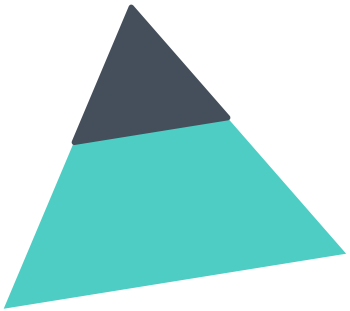


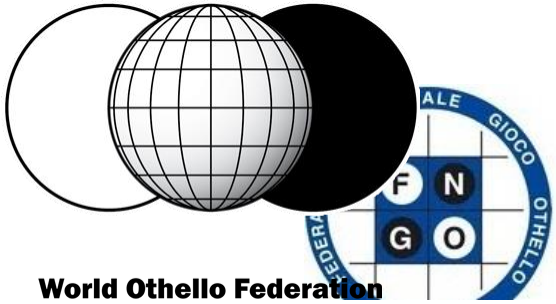
Why Othello?

Simple but still interesting

Well known

Easy to implement





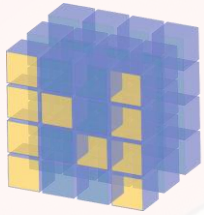
World Othello Federation

```
13  class OthelloGame():
14      """ Rules of the Othello ga
15
16      Implementation derived from
17      A game state contains the p
18
19      """
20      def __init__(self, n=8, mem
21          """ Generates rules for
```


OLIVAW
Othello learning in very anthropic way

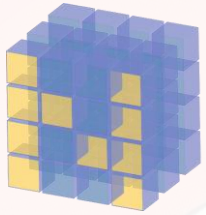
OLIVAW

Othello learning in very anthropic way



OLIVAW

Othello learning in very anthropic way



value head

policy head

residual layer

residual layer

residual layer

residual layer

residual layer

residual layer

residual layer

residual layer

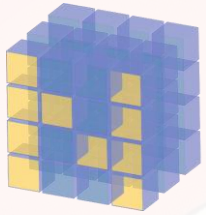
residual layer

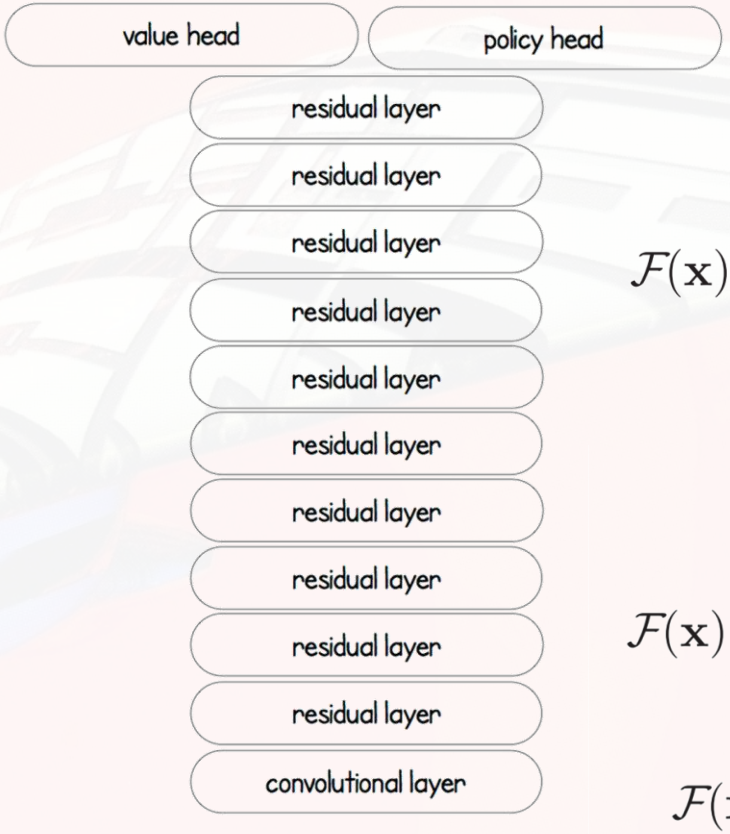
residual layer

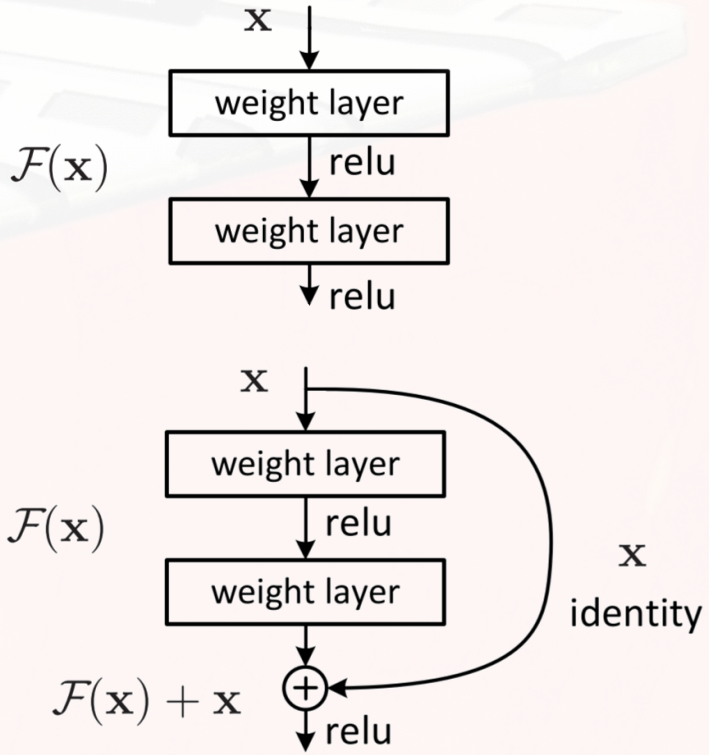
convolutional layer

OLIVAW

Othello learning in very anthropic way

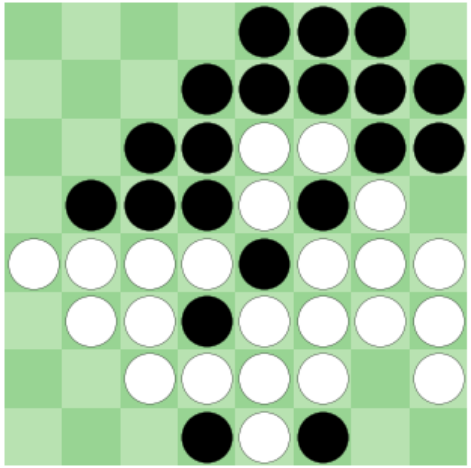







OLIVAW
the training process

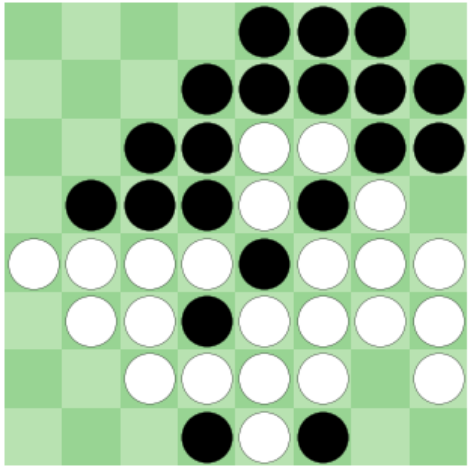
OLIVAW: the training process

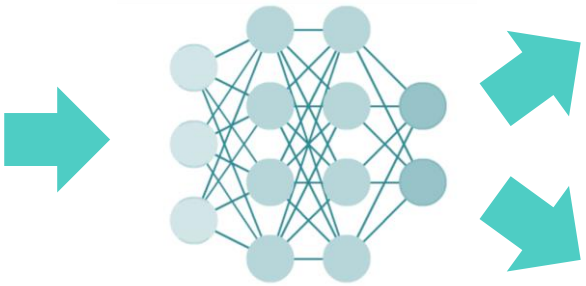


+0,24

```
[[0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.03 0.   0.   0.   0.   0.  ]
 [0.15 0.   0.   0.   0.   0.   0.   0.  ]
 [0.64 0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.18 0.   0.   0.   0.   0.  ]]
```

OLIVAW training process

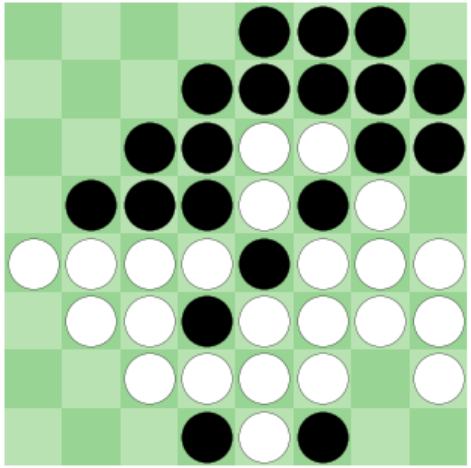




+0,24

```
[[0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.03 0.   0.   0.   0.   0.  ]
 [0.15 0.   0.   0.   0.   0.   0.   0.  ]
 [0.64 0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.18 0.   0.   0.   0.   0.  ]]
```

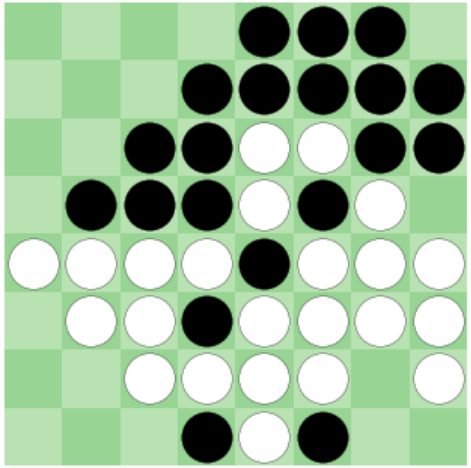
OLIVAW training process

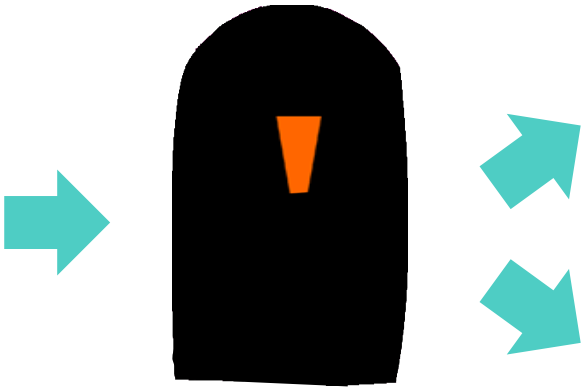


+0,24

```
[[0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.03 0.   0.   0.   0.   0.  ]
 [0.15 0.   0.   0.   0.   0.   0.   0.  ]
 [0.64 0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.18 0.   0.   0.   0.   0.  ]]
```

OLIVAW training process

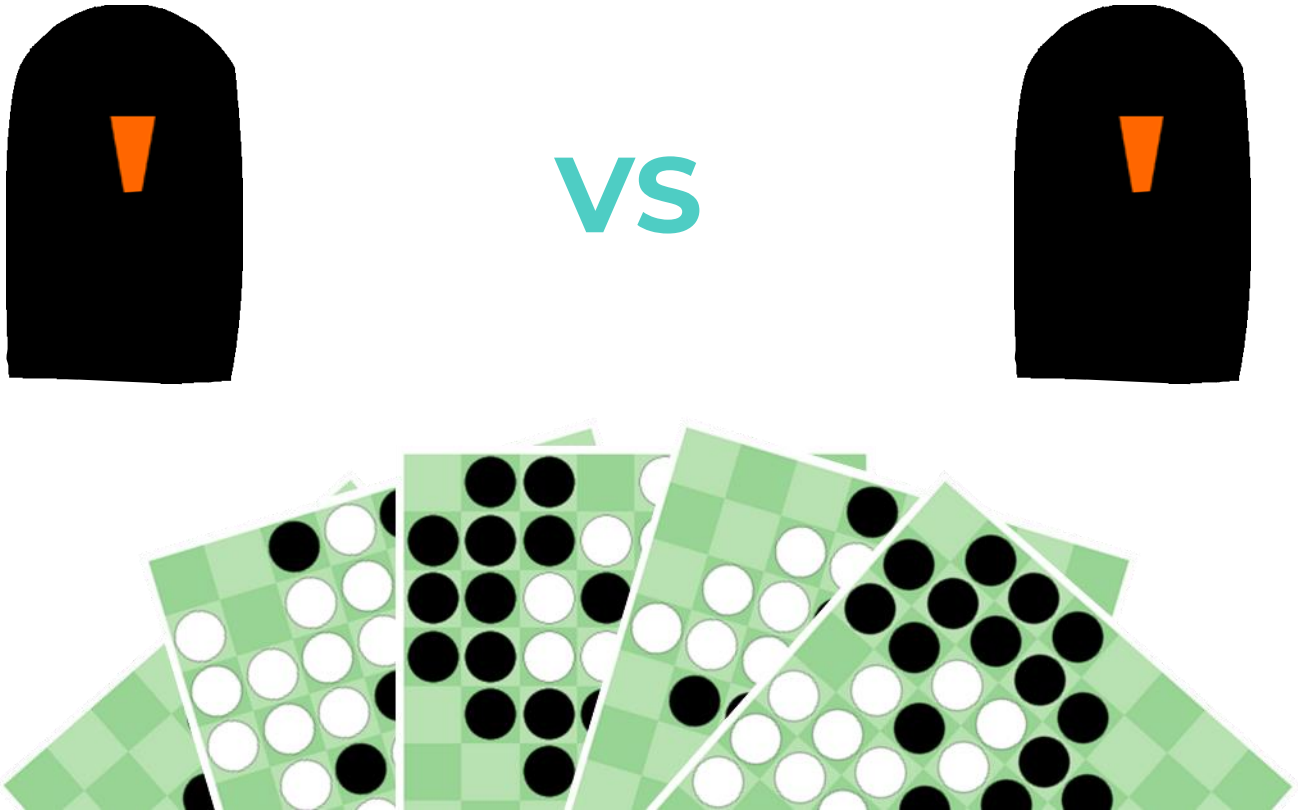




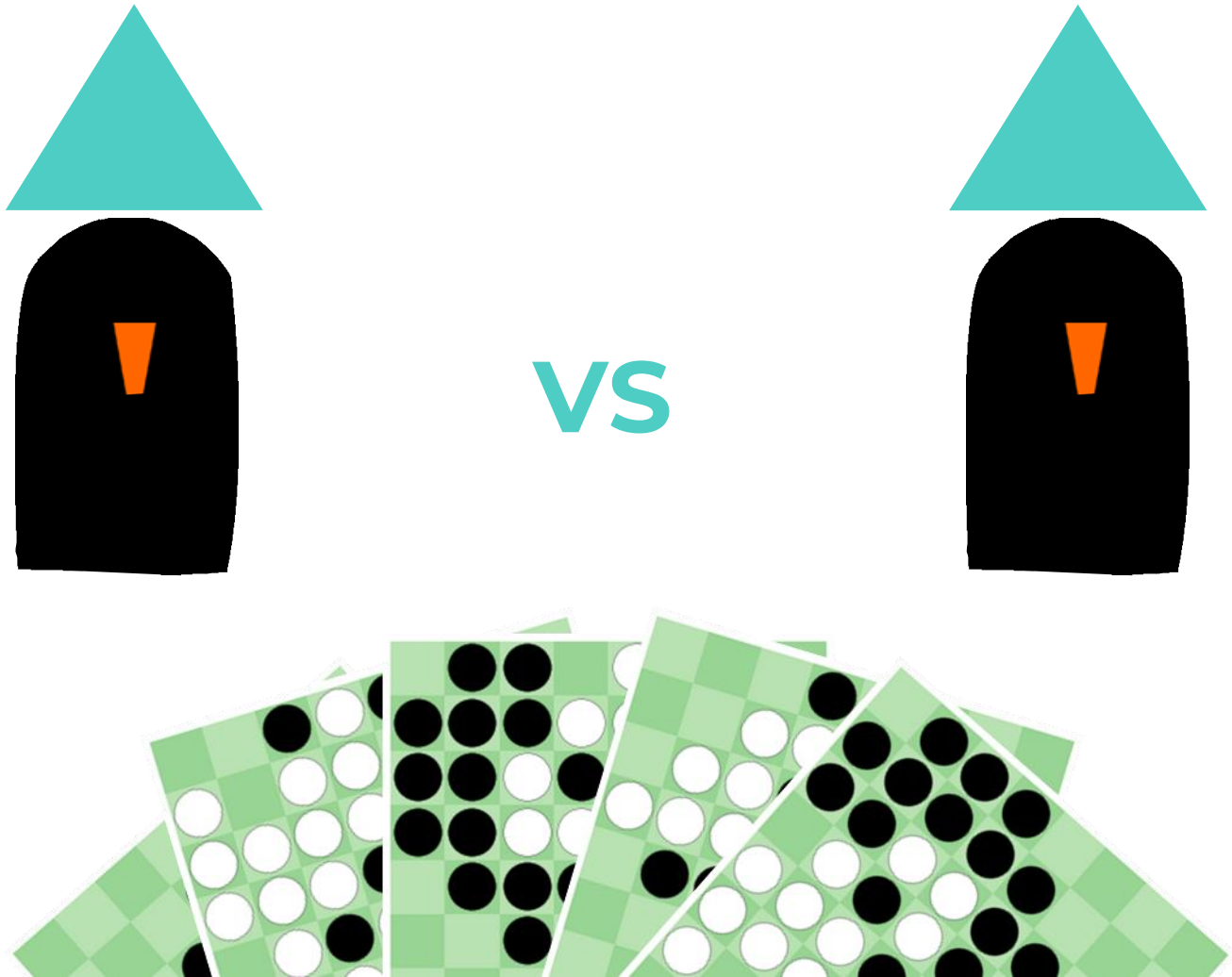
-0,03

```
[[0.   0.   0. 11 0.11   0.   0.   0.   0.12  ]
 [0.   0.13   0.9 0.   0.   0.   0.   0.  ]
 [0.10 0.11   0.   0.   0.   0.   0.   0.  ]
 [0.12 0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.   0.   0.   0.   0.   0.  ]
 [0.   0.   0.11 0.   0.   0.   0.11   0.  ]]
```

OLIVAW training process

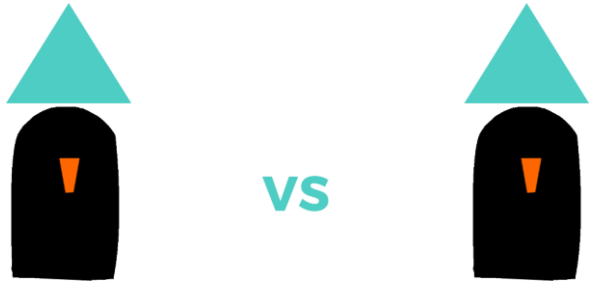


OLIVAW training process



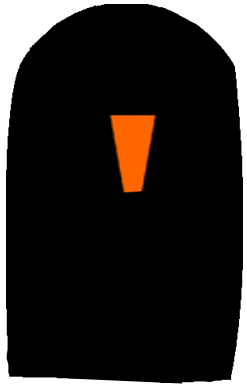
OLIVAW training process

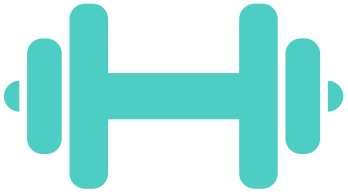
1. Self-play games generation

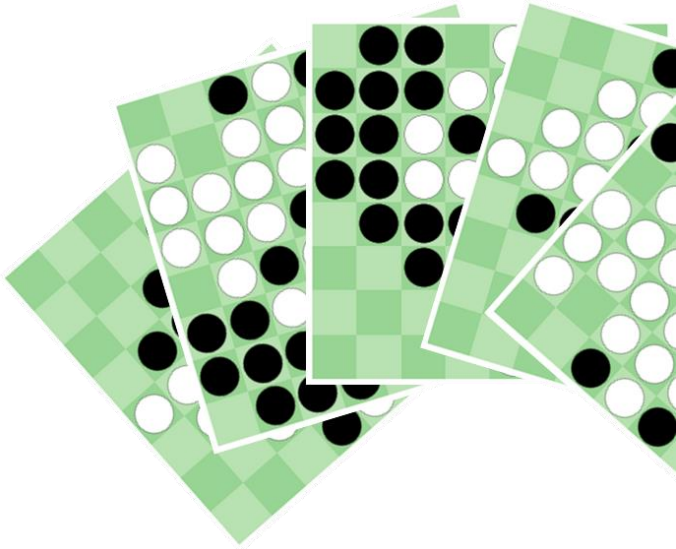


OLIVAW training process

1. Self-play games generation





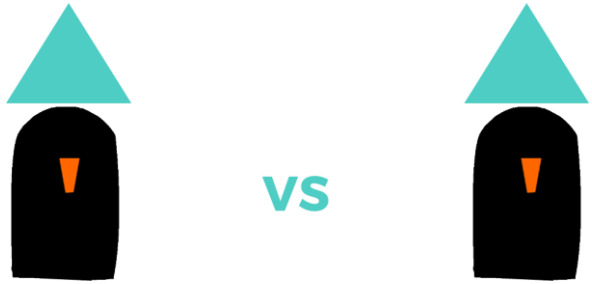


OLIVAW training process

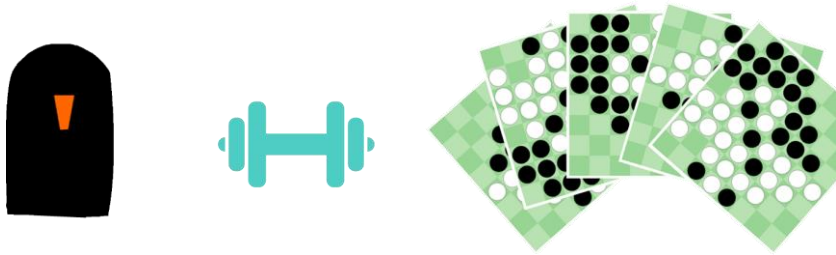
1. Self-play games generation

OLIVAW training process

1. Self-play games generation



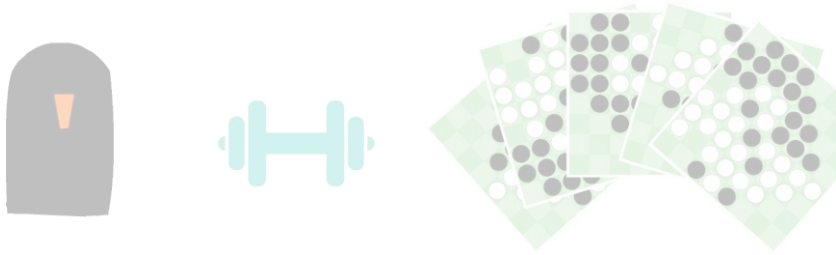
2. Neural Net training



OLIVAW training process

1. Self-play games generation

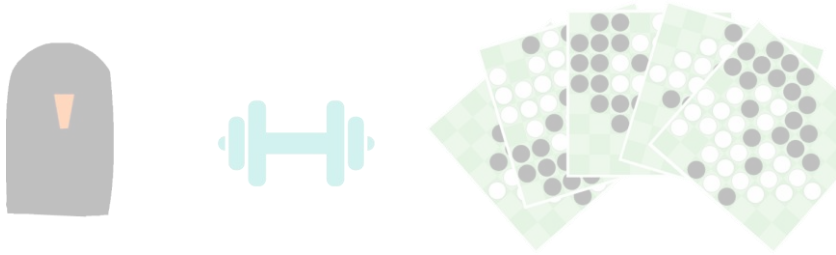
2. Neural Net training



OLIVAW training process

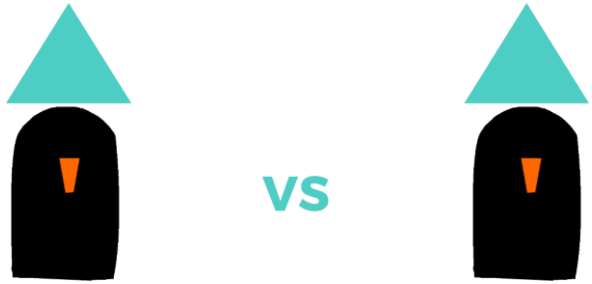
1. Self-play games generation

2. Neural Net training

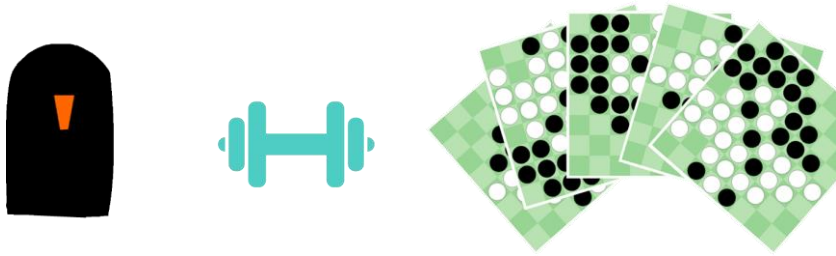


OLIVAW training process

1. Self-play games generation



2. Neural Net training



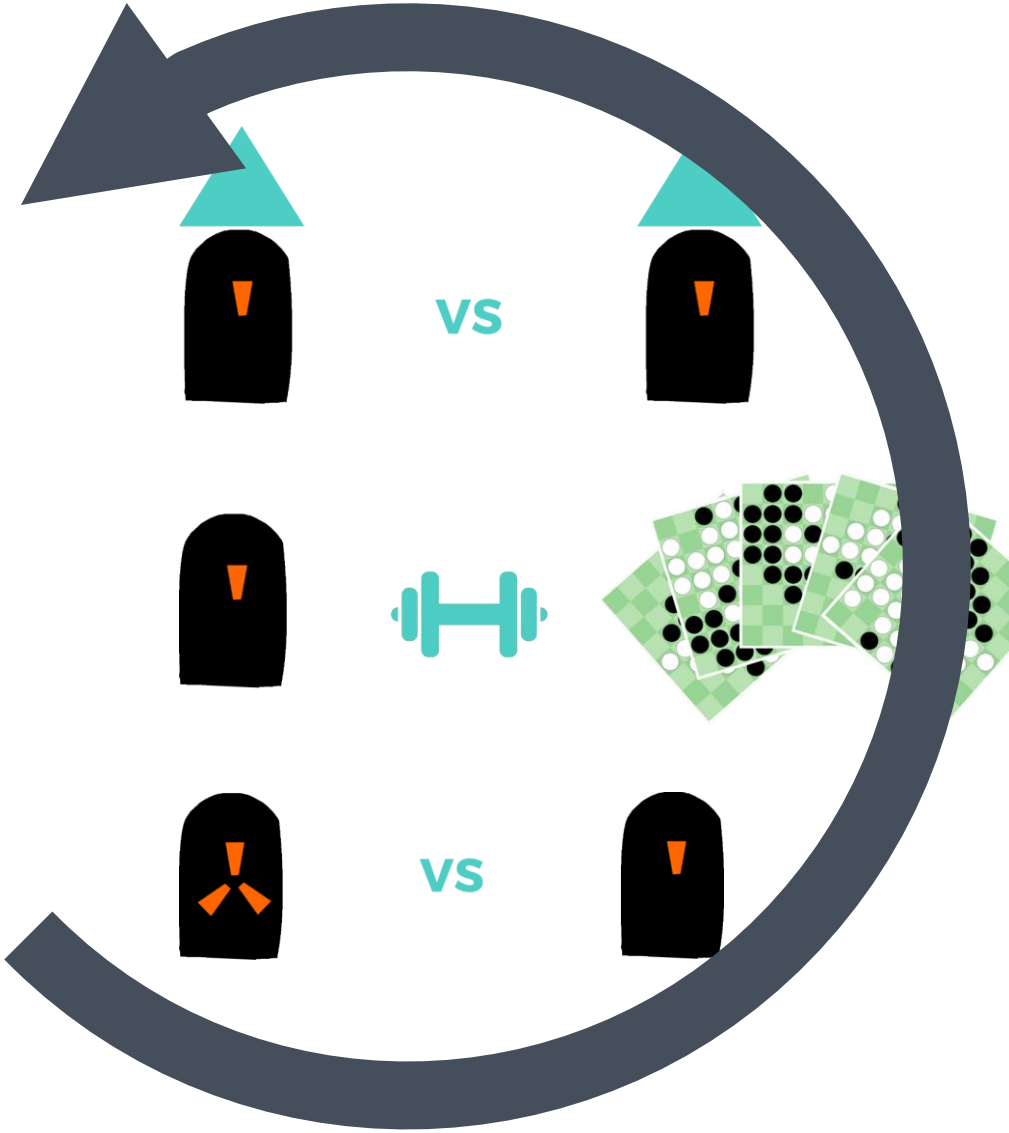
3. Evaluation

OLIVAW training process

1. Self-play games generation

2. Neural Net training

3. Evaluation



Reaching superhuman strength

OLIVAW

vs

Alessandro Di Mattei

2016-2017 Italian champion

27-11-2018

OLIVAW

vs

Alessandro Di Mattei

2016-2017 Italian champion

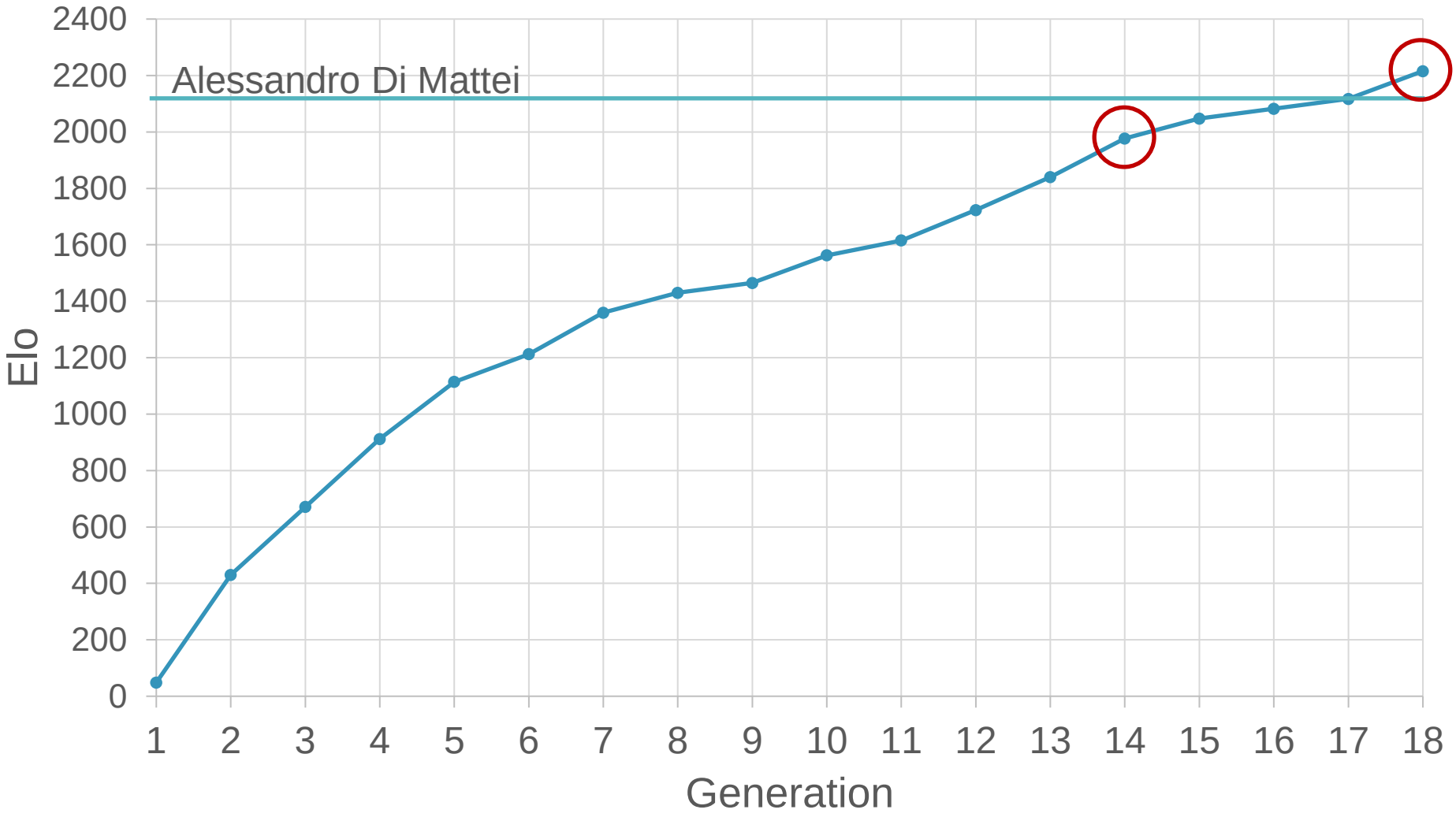
2-3

Draw – Draw – Defeat – Victory - Defeat

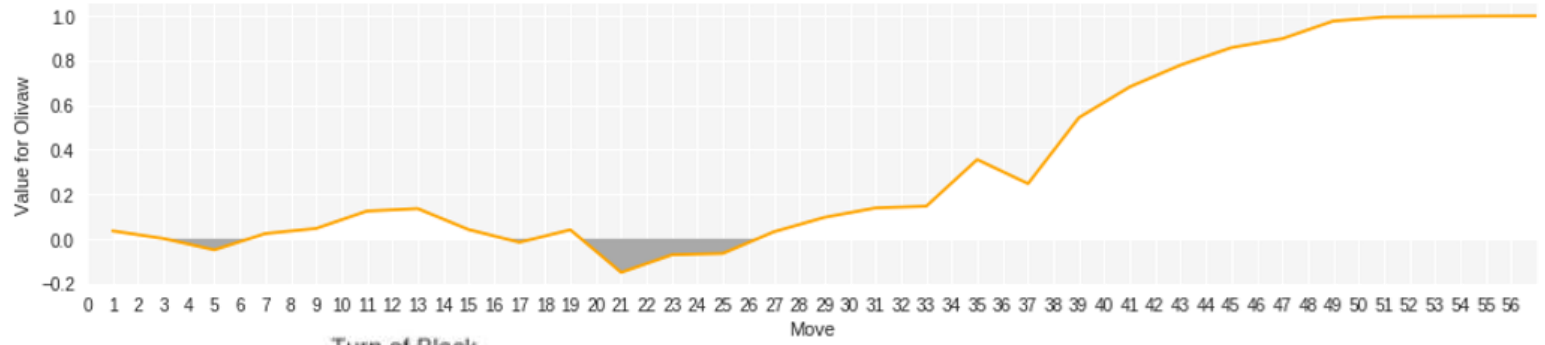

OLIVAW
vs
Alessandro Di Mattei

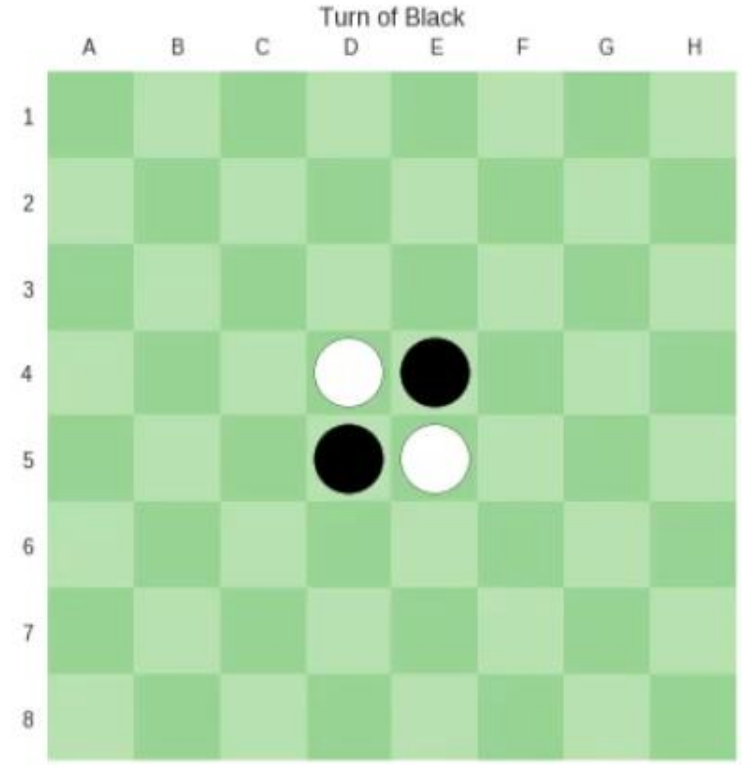

3-12-2018
OLIVAW
vs
Alessandro Di Mattei
4-0

OLIVAW's strength



OLIVAW vs Di Mattei 03-12 Game 4





A. Di Mattei (**B**) vs Olivaw gen 18 (**W**) – **27 - 37**

---

1. C4, E3
2. F6, E6
3. F5, C5
4. C3, C6
5. D3, D2
6. E2, B3
7. C1, C2
8. B4, A3
9. A5, B5
10. A6, B6
11. A4, A7
12. E7, E1
13. D6, D7
14. C8, F3
15. C7, F8
16. F7, G5
17. H4, G6
18. H5, D1
19. F1, F2
20. B1, H6
21. H7, G4
22. H3, F4
23. G3, E8
24. B7, D8
25. G8, B2
26. B8, G7
27. A2, A1
28. PASS, G1
29. G2, H2
30. H1


OLIVAW
vs
Michele Borassi
2008 World Othello
champion

HUMANS VS MACHINES

Round III

Borassi

2008 World Othello Champion

VS

OLIVAW

19-01-2019

h 14.30

Odeion - Museo di lettere classiche
Piazzale Aldo Moro, 5

Dipartimento di Informatica
Sapienza Università di Roma

OLIVAW

vs

Michele Borassi

Best of 3

19 January 2019

16,30 – Dipartimento di Matematica Guido Castelnuovo
Sapienza, piazzale Aldo Moro, 5

OLIVAW

vs

Michele Borassi

Game 1

1-0

OLIVAW wins with black

OLIVAW

vs

Michele Borassi

Game 2

1-1

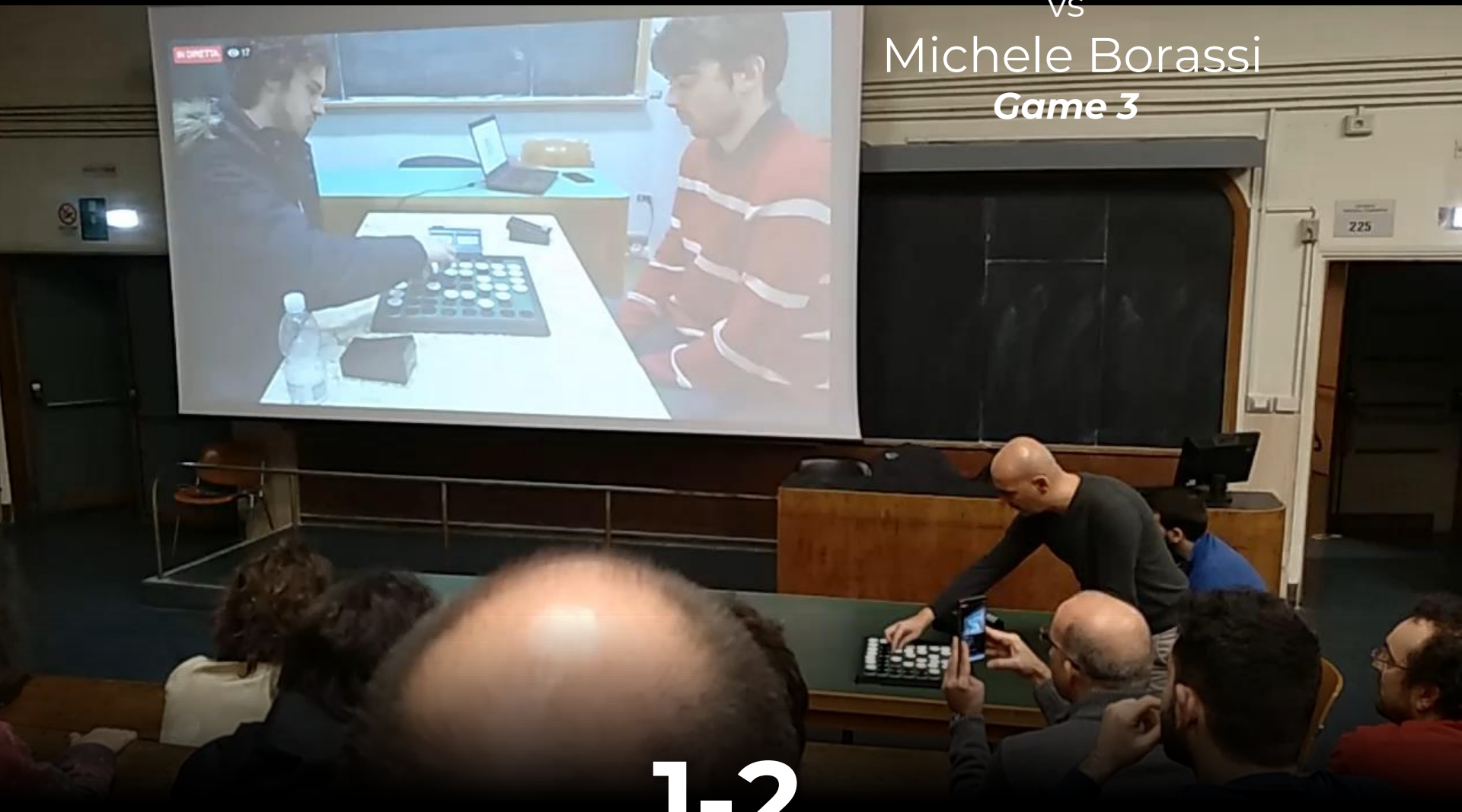
Michele Borassi wins with black

OLIVAW

vs

Michele Borassi

Game 3



1-2

Michele Borassi wins with black

Thanks!

Any questions?

You can find me at norelli@di.uniroma1.it