Topoi of automata (1/n)
Automata and Regular languages

Ryuya Hora

University of Tokyo

April 30, 2025

Based on (the first half of)^2 *Topoi of automata I* [Hora 2024]

Table of Contents

Table of Contents

Automata theory is multi-faceted ... just like topos theory

Automata theory is known for its relationship with

Logic	first-order, and higer (mainly second-order) logic, ...
Topology	Stone duality, fibration and profinite topological monoids, ...
Algebra	Monoids, and Grothendieck (semi-)Galois theory, ...
Category	Coalgebra, fibration, factorization systems, bicategory...

Question

Does this reflect the multi-facetedness of some (automaton-theoretic) Grothendieck topos?

(Quasi-)answer:

Automata theory is multi-faceted ... just like topos theory

Automata theory is known for its relationship with

Logic	first-order, and higer (mainly second-order) logic, ...
Topology	Stone duality, fibration and profinite topological monoids, ...
Algebra	Monoids, and Grothendieck (semi-)Galois theory, ...
Category	Coalgebra, fibration, factorization systems, bicategory...

Question

Does this reflect the multi-facetedness of some (automaton-theoretic) Grothendieck topos?

(Quasi-)answer: in next talk
cf. Toposes of topological monoid actions [Rogers 2023]

Ambitious motivation: New geometric invariant

There remain elementary-stated, discrete, and finite **open problems in automata theory**, one of which is *the generalized star-height problem.*

Ambitious motivation: New geometric invariant

There remain elementary-stated, discrete, and finite **open problems in automata theory**, one of which is *the generalized star-height problem.*

"Isn't an essentially new invariant necessary?"

Ambitious motivation: New geometric invariant

There remain elementary-stated, discrete, and finite **open problems in automata theory**, one of which is *the generalized star-height problem.*

“Isn’t an essentially new invariant necessary?”

Later I will provide a construction:

$$\mathcal{T} \colon (\text{a language class } C \subset \mathcal{P}(\Sigma^*)) \mapsto (\text{a Grothendieck topos } \mathcal{T}_C).$$

The cohomology of $\mathcal{E}_L$ contains information about the “star-height,” in a very simple case (so that even I can calculate)!
This situation reminds me of what Grothendieck says in ReS.

Today's talk in one slide (30th April)

1. **Regular language** is a language **recognize**d by a finite **automaton**.
2. We can check regularity of a language with the **Myhill-Nerode theorem.**
3. We can obtain the above notions/theorem from **the canonical point**

$$p\colon \mathbf{Set} \to \Sigma\text{-}\mathbf{Set}.$$

I suggest some related research topics that I can mention already.

Table of Contents

Notations and terminologies

- We fix a finite set Σ, say $\Sigma = \{a, b\}$, which we call an **alphabet**.
- Σ^* denotes the (free) monoid of finite words.
- A **language** is a subset of Σ^*.
    - The set of all languages is denoted by **Lan** $:= \mathcal{P}(\Sigma^*)$.

Σ-set

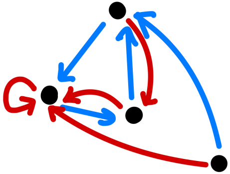
Definition (Σ-set)

A Σ-**set** is a pair (Q, δ) of

- a set Q (of states) and
- a (transition) function $\delta \colon Q \times \Sigma \to Q$.

Σ-set = right Σ^*-action = presheaf over the one-object category Σ^*.

$\Sigma = \{a, b\}$



: Σ- set

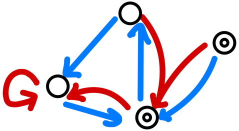
Automaton

Definition (Automaton)

A (Σ-)**automaton** is a triple (Q, δ, F) of

- a set Q (of states),
- a (transition) function $\delta \colon Q \times \Sigma \to Q$, and
- a subset $F \subset Q$ of (final states).

$\Sigma = \{a, b\}$



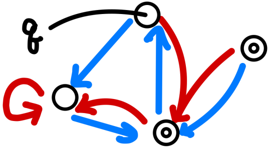
: automaton

Language recognition

Definition (Language recognition)

The language **recognised** by an automaton (Q, δ, F) and a state $q \in Q$ is

$$\{w \in \Sigma^* \mid qw \in F\} \subset \Sigma^*.$$



recognises $\{b, aa, aba, bab, bba, aaab, \dots\}$

$\subseteq \Sigma^*$

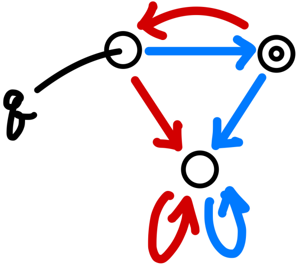
Regular languages

Definition

A language $L \subset \Sigma^*$ is **regular** if L is recognised by a finite automaton (Q, δ, F) (and a state $q \in Q$) $|Q| < \infty$.

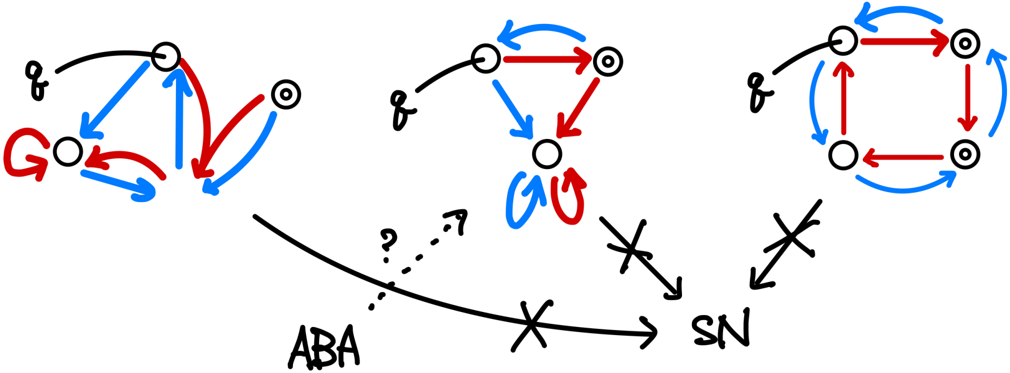
- $\mathrm{ABA} = \{a, aba, ababa, abababa, \ldots\}$ is regular.
- $\mathrm{SN} =$ $\{\varepsilon, ab, ba, aabb, abab, abba, baab, baba, bbaa, \ldots\}$ is not regular.

How do we know a given language L is non-regular?



Myhill-Nerode Theorem (1/3): Motivating Questions

- if L is regular, how can we find (small) automaton?
- if L is not regular, how can we prove it?



Myhill-Nerode Theorem (2/3): Nerode congruence

Definition (Nerode-congruence)

For a language L, its **Nerode congruence** $\sim_L \subset \Sigma^* \times \Sigma^*$ is defined by

$$u \sim_L v \iff (\forall w \in \Sigma^*, uw \in L \iff vw \in L)$$

Example

For $\mathrm{SN} = \{w \in \Sigma^* \mid |w|_a = |w|_b\}$, we have

$$u \sim_{\mathrm{SN}} v \iff |u|_a - |u|_b = |v|_a - |v|_b.$$

Therefore, $\Sigma^* / {\sim_{\mathrm{SN}}} \cong \mathbb{Z}$.

Myhill-Nerode Theorem (3/3): the statement without proof

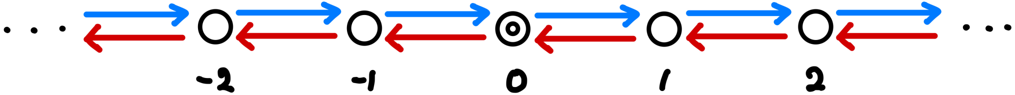
Theorem (Myhill-Nerode theorem)

For any language L, the set $Q := \Sigma^*/\sim_L$ (with a canonical choice of δ, F, q) provides ***the minimal automaton*** *that recognizes L. Therefore,*

$$L \text{ is regular} \iff \Sigma^*/\sim_L \text{ is finite.}$$

Example

SN is not regular, since $\Sigma^*/\sim_{\mathrm{SN}} \cong \mathbb{Z}$ is infinite.



What to remember

- We fix an **alphabet** Σ.
- A **Σ-set** is a right Σ^*-action $(Q, \delta\colon Q \times \Sigma \to Q)$.
- An **automaton** $(Q, \delta\colon Q \times \Sigma \to Q, F \subset Q)$ together with a(n initial) state $q \in Q$ recognises a **language** $\{w \in \Sigma^* \mid qw \in F\}$.
- A language L is said to be **regular** if it is recognized by a finite automaton.
- (Non-)regularity of a language L can be checked by the **Myhill-Nerode theorem**, which utilizes the **Nerode congruence**

$$u \sim_L v \iff (\forall w \in \Sigma^*, uw \in L \iff vw \in L)\,,$$

and **the minimal automaton** $\Sigma^*/{\sim_L}$.

Today’s talk in one slide (30th April)

1. **Regular language** is a language **recognize**d by a finite **automaton**.
2. We can check regularity of a language with the **Myhill-Nerode theorem.**
3. We can obtain the above notions/theorem from **the canonical point**

$$p\colon \mathbf{Set} \to \Sigma\text{-}\mathbf{Set}.$$

I suggest some related research topics that I can mention already.

Table of Contents

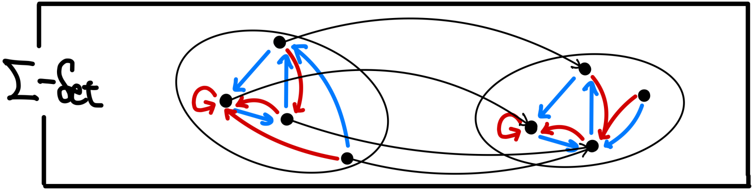
Topos of Σ-sets

Definition

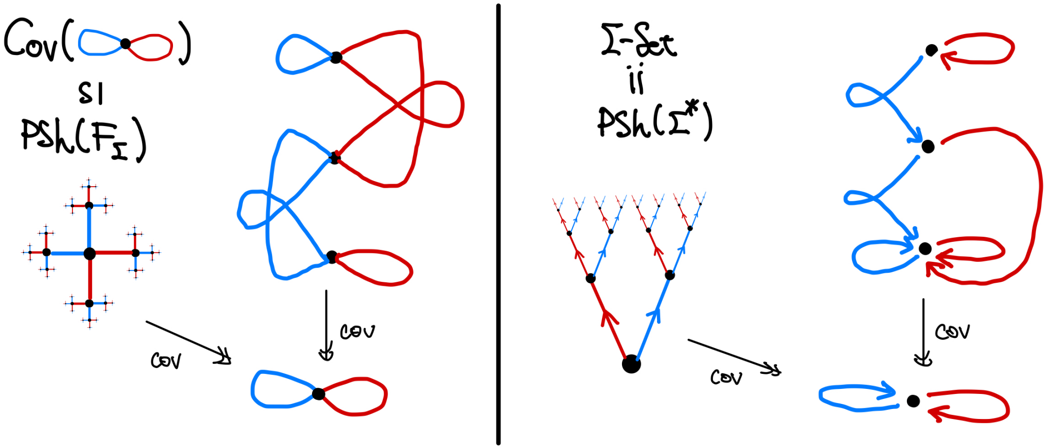
Σ-**Set** is the category of Σ-sets defined by $\Sigma\text{-}\mathbf{Set} \coloneqq \mathbf{PSh}(\Sigma^*)$.

Σ-**Set** is a (presheaf) topos.

$\Sigma = \{a, b\}$



Geometric idea: Σ-**Set** = directed bouquet



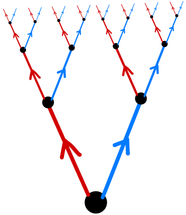
(cf. Topological proof of Nielsen-Schreier thm, Semi-Galois theory)

Two objects of Σ-sets (1/2): Words

The topos Σ-**Set** has **four important objects**:

1. The Σ-set of words Σ^*,
2. The Σ-set of languages **Lan**.
3. The Σ-set of right congruences Ξ.
4. The Σ-set of right ideals Ω.

Today, we will see the first two.



Example (The Σ-set of words Σ^*.)

Elements	Words.
Σ-action	Concatenation $v * w = vw$.
Universality	Yoneda lemma $\Sigma\text{-}\mathbf{Set}(\Sigma^*, (Q, \delta)) \cong Q$.

Three objects of Σ-sets (2/2): Languages

The topos Σ-**Set** has three important objects:

Example (The Σ-set of languages Lan)

Elements	Languages $\in \mathbf{Lan} := \mathcal{P}(\Sigma^*)$[1]
Σ-action	$L * w := \{v \in \Sigma^* \mid wv \in L\}$ (Brzozowski derivative)
Universality	We will see soon!

Remark (Nerode congruence is just…)

$$u \sim_L v \iff (\forall w \in \Sigma^*, uw \in L \iff vw \in L) \iff L * u = L * v$$

---

[1] This is not the powerset object of Σ^* in Σ-**Set**.

The canonical point of Σ-**Set**

Example (The canonical point[2])

The forgetful-cofree adjunction defines a point of Σ-**Set**.

$$\begin{tikzcd} p_* : \mathbf{Set} \arrow[r, bend right, "S \mapsto S^{\Sigma^*}"'] \arrow[r, phantom, "\perp"] & \Sigma\text{-}\mathbf{Set} : p^* \arrow[l, bend right, "Q \mapsfrom (Q,\delta)"'] \end{tikzcd}$$

We have $p_*(S) = S^{\Sigma^*}$. In particular, we have $p_*(\{\top, \bot\}) = \mathbf{Lan}$

Lemma (Universality of Lan)

$\Sigma\text{-}\mathbf{Set}((Q,\delta), \mathbf{Lan}) \cong \mathbf{Set}(Q, \{\top, \bot\}) \cong \mathcal{P}(Q)$.

---

[2] This terminology is borrowed from [Rogers 2023]

Language recognition = composition

Lemma

These data are in bijective correspondence:

1. *An automaton* (Q, δ, F) *together with an initial state* $q \in Q$.
2. *A diagram* $\Sigma^* \to (Q, \delta) \to \mathbf{Lan}$ *in the topos* Σ-$\mathbf{Set}$.

Thus, an automaton with an initial state:

$$\begin{tikzcd} \Sigma^* \arrow[r, "\text{pick } q"] & (Q, \delta) \arrow[r, "\chi_F^\sharp"] & \mathbf{Lan} \end{tikzcd}$$

Language recognition = composition

Lemma

These data are in bijective correspondence:

1. *An automaton* (Q, δ, F) *together with an initial state* $q \in Q$.
2. *A diagram* $\Sigma^* \to (Q, \delta) \to \mathbf{Lan}$ *in the topos* Σ-**Set**.

Thus, an automaton with an initial state:

$$\begin{tikzcd} \Sigma^* \arrow[r, "\text{pick } q"] & (Q, \delta) \arrow[r, "\chi_F^\sharp"] & \mathbf{Lan} \end{tikzcd}$$

recognizes a language:

$$\begin{tikzcd} \Sigma^* \arrow[r, "\text{the recognised language}"] & \mathbf{Lan} \end{tikzcd}$$

Myhill-Nerode theorem = epi-mono factorization

Conversely, for any language L

$$\begin{tikzcd}\Sigma^* \arrow[rr, "\lceil L \rceil"] & & \mathbf{Lan},\end{tikzcd}$$

Myhill-Nerode theorem = epi-mono factorization

Conversely, for any language L

$$\begin{tikzcd}\Sigma^* \arrow[rr, "\lceil L \rceil"] & & \mathbf{Lan},\end{tikzcd}$$

we obtain the minimal automaton with an initial state

$$\begin{tikzcd}\Sigma^* \arrow[r, two heads] & \mathrm{Im}(\lceil L \rceil) \arrow[r, tail] & \mathbf{Lan}.\end{tikzcd}$$

This is minimal in the sense that it is subquotient of any other factorization of $\lceil L \rceil$.
(cf. So far, this is quite similar to Functorial approach [Colcombet and Petrişan 2020]. But this will be done in other pointed topoi later.)

Nerode congruence = Kernel pair of $\lceil L \rceil$

For each language $\lceil L \rceil \colon \Sigma^* \to \mathbf{Lan}$, we have

$$\begin{tikzcd}
\sim_L \arrow[r] \arrow[d] \arrow[dr, phantom, "\lrcorner", very near start] & \Sigma^* \arrow[d, "\lceil L \rceil"] \\
\Sigma^* \arrow[r, "\lceil L \rceil"'] & \mathbf{Lan},
\end{tikzcd}$$

since $\lceil L \rceil$ sends u to $L * u$, and hence

$$\lceil L \rceil(u) = \lceil L \rceil(v) \iff L * u = L * v \iff u \sim_L v.$$

This provides the construction of the minimal automaton $\mathrm{Im}(\lceil L \rceil) = \Sigma^*/{\sim_L}$ and the proof of the Myhill-Nerode theorem.

Digression: Comparison with the Coalgebraic method

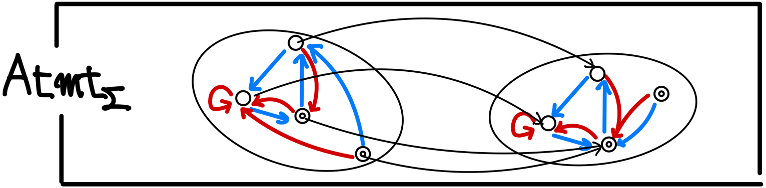
Proposition

The category of automata $\mathbf{Atmt}_{\Sigma} = \mathbf{Coalg}_{2\times\Sigma}$ *is a presheaf topos.*

Proof.

$\mathbf{Atmt}_{\Sigma} \simeq \Sigma\text{-}\mathbf{Set}/\mathbf{Lan} \simeq \mathbf{PSh}\left(\int \mathbf{Lan}\right)$ □

$\Sigma = \{a, b\}$



It'll be interesting for variants of Σ-**Set**: automata form an étale cover.

Today's talk in one slide (30th April)

1. **Regular language** is a language **recognized** by a finite **automaton**.
2. We can check regularity of a language with the **Myhill-Nerode theorem.**
3. We can obtain the above notions/theorem from **the canonical point**

$$p\colon \mathbf{Set} \to \Sigma\text{-}\mathbf{Set}.$$

I suggest some related research topics that I can mention already.

Overview of what follows

Today From the pointed topos $p\colon \mathbf{Set} \to \Sigma\text{-}\mathbf{Set}$, we obtain the object $p_*(\top, \bot) = \mathbf{Lan}$, which is the Boolean algebra of all languages **Lan**.

ToA I We construct a pointed topos $p\colon \mathbf{Set} \to \Sigma\text{-}\mathbf{Set}_{\mathrm{o.f.}}$ such that $p_*(\{\top, \bot\})$ is the Boolean algebra of regular languages **Reg**. Furthermore, the equivalence between four characterizations of regular languages is enriched into **the equivalence between four Boolean ringed topoi**, all of which turn out to be $(\Sigma\text{-}\mathbf{Set}_{\mathrm{o.f.}}, \mathbf{Reg})$.

ToA II There is a construction $\mathcal{T}$

$$\mathbf{Lan} \supset C \mapsto \mathcal{T}_C\colon \text{ Grothendieck topos}$$

such that $\mathcal{T}_{\mathbf{Lan}} = \Sigma\text{-}\mathbf{Set}$, $\mathcal{T}_{\mathbf{Reg}} = \Sigma\text{-}\mathbf{Set}_{\mathrm{o.f.}}$ (and $\mathcal{T}$ subsumes ...).

Table of Contents

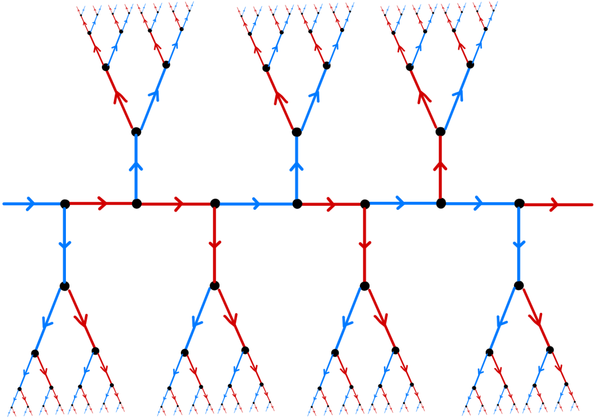
Other points of Σ-**Set** = infinite words recognition!? (1/2)

We can apply the same argument to all other points as well.

Proposition

Non-canonical points of Σ-**Set** *are in bijective correspondence with* ***infinite words*** *up to finite edit distance.*

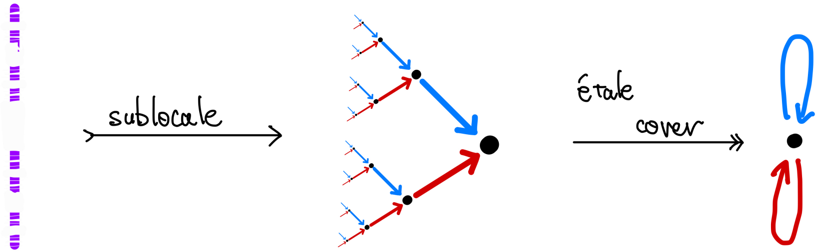
(cf. Büchi automata)



Other points of Σ-**Set** = infinite words recognition!? (2/2)

There is a canonical geometric morphism:

$$\gamma : \mathbf{Sh}(\Sigma^\omega) \xhookrightarrow{\text{sublocale}} \mathbf{Sh}(\Sigma^*, \leq) \simeq \Sigma\text{-}\mathbf{Set}/\Sigma^* \xrightarrow{\text{étale cover}} \Sigma\text{-}\mathbf{Set}$$



The stalk of $\gamma^*(Q, \delta)$ at a point $p \in \Sigma^\omega$ captures the eventual behavior of (Q, δ) with the infinite input p. (cf. [Leinster 2024])

Subtopoi (1/2)

Proposition (j.w.w. Morgan Rogers [Hora and Rogers n.d.])

The topos Σ-**Set** *has the maximum non-trivial subtopos* $\mathbf{JT}_\Sigma$. *Furthermore, there is a one-to-one correspondence between*

- *non-trivial subtopoi of* Σ-**Set**, *and*
- *self-similar sublocales of the Cantor space* Σ^ω.

Subtopoi (1/2)

Proposition (j.w.w. Morgan Rogers [Hora and Rogers n.d.])

The topos Σ-**Set** *has the maximum non-trivial subtopos* $\mathbf{JT}_\Sigma$. *Furthermore, there is a one-to-one correspondence between*

- *non-trivial subtopoi of* Σ-**Set**, *and*
- *self-similar sublocales of the Cantor space* Σ^ω.

Q. How is this related to automata theory?

Subtopoi (2/2)

The structure of the complete lattice of hyperconnected quotients $\mathrm{HQ}(\mathcal{E})$ is automata-theoretically important (next talk).

Proposition

For any Grothendieck topos $\mathcal{E}$, the complete lattice of hyperconnected quotients $\mathrm{HQ}(\mathcal{E})$ is a left module of the semiring $\mathrm{End}_{\wedge}(\Omega)$.

The lattice of subtopoi is the lattice of idempotents of $\mathrm{End}_{\wedge}(\Omega)$. (cf. Pierce spectrum)

Example (Σ-Set when $\Sigma = \{*\}$)

For **PSh**$(\mathbb{N})$, the semiring $\mathrm{End}_{\wedge}(\Omega)$ is the min-plus algebra $\{0 < 1 < \cdots < \infty < \infty'\}$ acting on $\mathrm{HQ}(\mathbf{PSh}(\mathbb{N})) \cong \overline{\mathbb{N}}^{\mathrm{Spec}(\mathbb{Z})} \sqcup \{\infty\}$.

Quotient topoi (1/2)

Definition

A geometric morphism $f : \mathcal{E} \to \mathcal{F}$ is said to be **connected** if $f^* : \mathcal{F} \to \mathcal{E}$ is fully faithful.

A **quotient topos** of Σ-**Set** is a connected geometric morphism from Σ-**Set**.
"Folding a space by restricting coverings."

Example (Covering spaces over Σ-bouquet.)

The topos $\mathbf{PSh}(F_\Sigma)$ is a quotient topos $\Sigma\text{-}\mathbf{Set} \twoheadrightarrow \mathbf{PSh}(F_\Sigma) \simeq \mathbf{Cov}(\Sigma\text{-bouquet})$.

Example (Quotienting Σ-bouquet into a circle)

$\Sigma\text{-}\mathbf{Set} \twoheadrightarrow \mathbf{PSh}(\mathbb{N})$

Quotient topoi (2/2)

How many quotient topoi does Σ-**Set** have?

- When $|\Sigma| = 1$, there are $2^{\aleph_0}$-many quotients that reflect the (possibly **non-periodic/global**) behaviors of dynamical systems. [Hora and Kamio 2024]
- When $|\Sigma| = \infty$, there are proper class many quotients. I and Yuhi Kamio solves the first of Lawvere's open problems with this example [Kamio and Hora 2024].
- **When** $1 < |\Sigma| < \infty$**, even the number of quotients are unknown.**

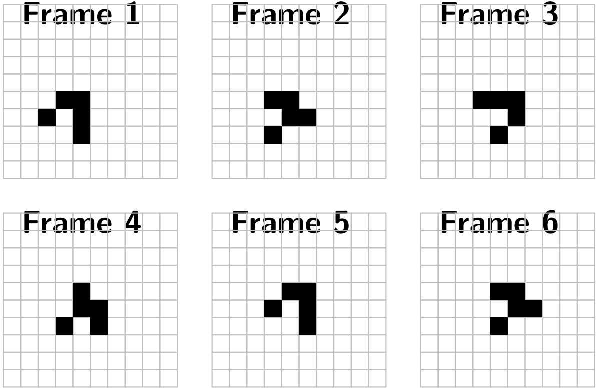
Question (Non-periodic behavior of Σ-sets)

Assuming $1 < |\Sigma| < \infty$*, classify all full subcategories of* Σ-**Set** *that are closed under finite limits and small colimits.*

(cf. Hyperconnected quotients of Σ-**Set** play a crucial role in [Hora n.d.].)

Cellular automata live "above" Σ-**Set** (1/3)

How can we study more complicated systems, like **Conway's game of life**?



Cellular automata live "above" Σ-**Set** (2/3)

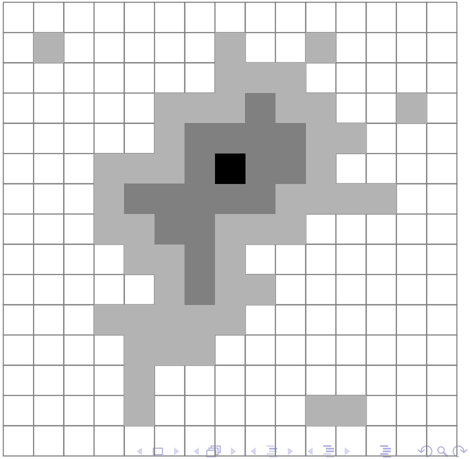
A **topology** on a set X is a lex-comonad

$$\mathrm{Int}\colon \mathcal{P}(X) \to \mathcal{P}(X).$$

To deal with “dynamical/discrete geometry", we consider the **pretopology**

$$\mathrm{Int}\colon \mathcal{P}(\mathbb{Z}^2) \to \mathcal{P}(\mathbb{Z}^2)$$

with $\mathrm{Int}^2 \subsetneq \mathrm{Int}$, which is an internal poset in **PSh**$(\mathbb{N})$.



Cellular automata live "above" Σ-**Set** (3/3)

(Informal) Theorem

Conway's game of life is naturally an object of

$$\mathbf{PSh}(\mathcal{P}(\mathbb{Z}^2) \rtimes_{\mathrm{Int}} \mathbb{N}) \to \mathbf{PSh}(\mathbb{N}) = \{*\}\text{-}\mathbf{Set}.$$

This is one (unsatisfying[3]) example of how **more complicated dynamical systems can be treated topos theoretically**.
(If we allow Σ-inputs, then this will become a relative topos over Σ-**Set**.)

---

[3] We should put a topology in it.

Other problems I'm interested in

References

- Colcombet, Thomas and Daniela Petrişan (2020). "Automata minimization: a functorial approach". In: *Logical Methods in Computer Science* 16.
- Hora, Ryuya (2024). "Topoi of automata I: Four topoi of automata and regular languages". In: *arXiv:2411.06358*.
- — (n.d.). *Topoi of automata II: Hyperconnected geometric morphisms, syntactic monoids, and language classes (ongoing)*.
- Hora, Ryuya and Yuhi Kamio (2024). "Quotient toposes of discrete dynamical systems". In: *Journal of Pure and Applied Algebra*, p. 107657.
- Hora, Ryuya and Morgan Rogers (n.d.). *Subtoposes of actions of free monoids (ongoing)*.
- Kamio, Yuhi and Ryuya Hora (2024). "A solution to the first Lawvere's problem A Grothendieck topos that has a proper class many quotient topoi". In: *arXiv:2407.17105*.
- Leinster, Tom (2024). "The eventual image". In: *Theory and Applications of Categories* 42.9, pp. 180–221.
- Rogers, Morgan (2023). "Toposes of topological monoid actions". In: *Compositionality: the open-access journal for the mathematics of composition* 5.