Unambiguity Makes Context Free Grammars Big

Stefan Mengel

Univ. Artois, CNRS, CRIL, Lens **Harry Vinall-Smeeth**

TU Ilmenau

FMT Workshop 2025, Les Houches

 $S \rightarrow Aa \mid aB$

 $S \rightarrow Aa \mid aB$ $G \qquad A \rightarrow a \mid c$

 $S \rightarrow Aa \mid aB$ $A \rightarrow a \mid c$ $B \rightarrow a \mid b$

 $S \rightarrow Aa \mid aB$ $A \rightarrow a \mid c$ $B \rightarrow a \mid b$

Derivations

$$G \qquad \begin{array}{c|c} S \rightarrow Aa & aB \\ \hline A \rightarrow a & c \\ \hline B \rightarrow a & b \end{array}$$

 $S \rightarrow Aa$

Derivations

$$S \rightarrow Aa \mid aB$$

$$A \rightarrow a \mid c$$

$$B \rightarrow a \mid b$$

Derivations

$$S \rightarrow Aa \rightarrow aa$$

Derivations
$$S oup Aa oup aa$$
 $S oup Aa oup aa$ $S oup Aa oup aa$ $S oup aB oup aa$ $S oup aB oup ab$

$$S \rightarrow Aa \mid aB$$

$$A \rightarrow a \mid c$$

$$B \rightarrow a \mid b$$

$$L(G) = \{aa, ab, ca\}$$

Derivations
$$S o Aa o aa$$
 $S o Aa o ca$ $S o aB o aa$ $S o aB o ab$

$$S \rightarrow Aa \mid aB$$

$$A \rightarrow a \mid c$$

$$B \rightarrow a \mid b$$

$$L(G) = \{aa, ab, ca\}$$

Derivations
$$S \rightarrow Aa \rightarrow aa$$

 $S \rightarrow Aa \rightarrow ca$
 $S \rightarrow aB \rightarrow aa$
 $S \rightarrow aB \rightarrow ab$

$$|G| = 8 =$$
 $|Aa| + |aB| + |a| + |c| + |a| + |b|$

CFG for Finite Languages

$$S \rightarrow Aa \mid aB$$
 $G \qquad A \rightarrow a \mid c$
 $B \rightarrow a \mid b$

$$L(G) = \{aa, ab, ca\}$$

Derivations
$$S \rightarrow Aa \rightarrow aa$$

 $S \rightarrow Aa \rightarrow ca$
 $S \rightarrow aB \rightarrow aa$
 $S \rightarrow aB \rightarrow ab$

$$|G| = 8 =$$
 $|Aa| + |aB| + |a| + |c| + |a| + |b|$

$$S \rightarrow Aa \mid aB$$

$$A \rightarrow a \mid c$$

$$B \rightarrow a \mid b$$

$$L(G) = \{aa, ab, ca\}$$

Derivations
$$S o Aa o aa$$
 $S o Aa o ca$ $S o aB o aa$ $S o aB o ab$

$$|G| = 8 =$$
 $|Aa| + |aB| + |a| + |c| + |a| + |b|$

$$S \rightarrow Aa \mid aB$$

$$A \rightarrow a \mid c$$

$$B \rightarrow a \mid b$$

$$L(G) = \{aa, ab, ca\}$$

Derivations
$$S \rightarrow Aa \rightarrow aa$$

 $S \rightarrow Aa \rightarrow ca$
 $S \rightarrow aB \rightarrow aa$
 $S \rightarrow aB \rightarrow ab$

$$|G| = 8 =$$
 $|Aa| + |aB| + |a| + |c| + |a| + |b|$

$$S \rightarrow Aa \mid aB$$

$$A \rightarrow a \mid c$$

$$B \rightarrow a \mid b$$

$$L(G) = \{aa, ab, ca\}$$

Derivations
$$S \rightarrow Aa \rightarrow aa$$

 $S \rightarrow Aa \rightarrow ca$
 $S \rightarrow aB \rightarrow aa$
 $S \rightarrow aB \rightarrow ab$

$$|G| = 8 =$$
 $|Aa| + |aB| + |a| + |c| + |a| + |b|$

$$S \rightarrow Aa \mid aB$$
 $G \qquad A \rightarrow a \mid c$
 $B \rightarrow a \mid b$

$$L(G) = \{aa, ab, ca\}$$

Derivations
$$S \rightarrow Aa \rightarrow aa$$

 $S \rightarrow Aa \rightarrow ca$
 $S \rightarrow aB \rightarrow aa$
 $S \rightarrow aB \rightarrow ab$

$$|G| = 8 =$$
 $|Aa| + |aB| + |a| + |c| + |a| + |b|$

G is ambiguous: two derivations of the same word!

$$G'$$

$$S \rightarrow Aa \mid aB$$

$$A \rightarrow a \mid c$$

$$B \rightarrow b$$

$$L(G') = \{aa, ab, ca\}$$

$$S \to Aa \to aa$$
 S \to Aa \to ca
$$S \to Aa \to ca$$
 Derivations
$$S \to aB \to ab$$

$$|G'| = 7 =$$
 $|Aa| + |aB| + |a| + |c| + |b|$

G' is unambiguous: every word has a unique derivation.



Large Object



Large Object



Large Object

Smaller Object



Large Object (Encoded as a set of strings)

Smaller Object



Large Object (Encoded as a set of strings)

CFG



Large Object (Encoded as a set of strings)

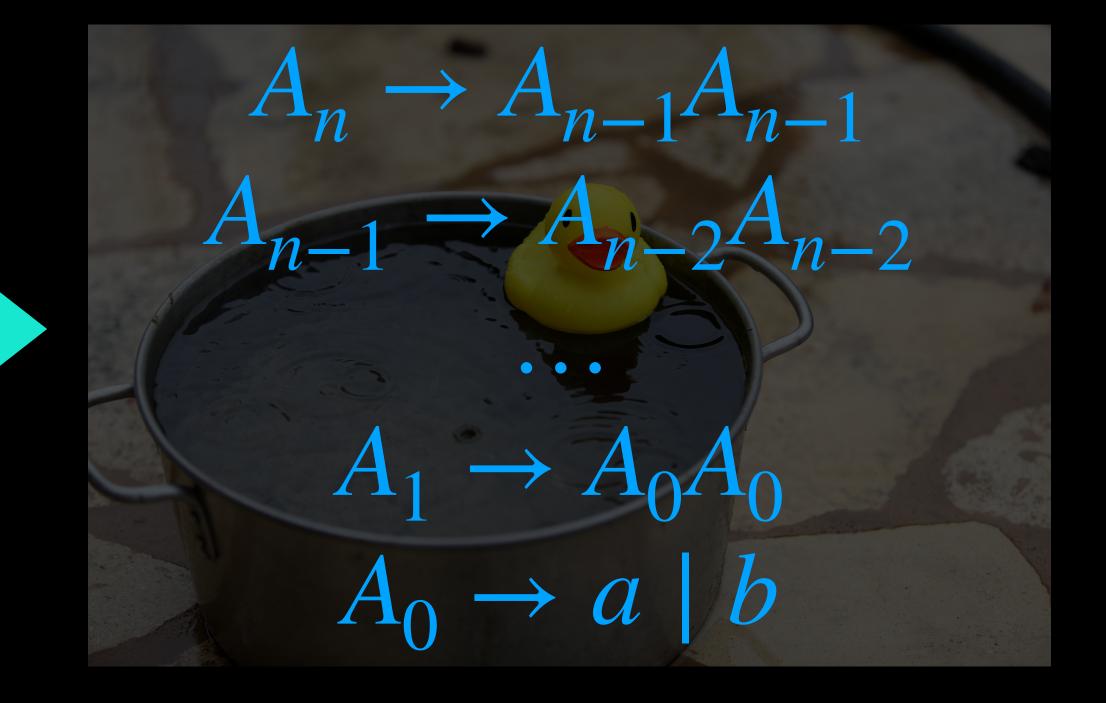


Large Object (Encoded as a set of strings)

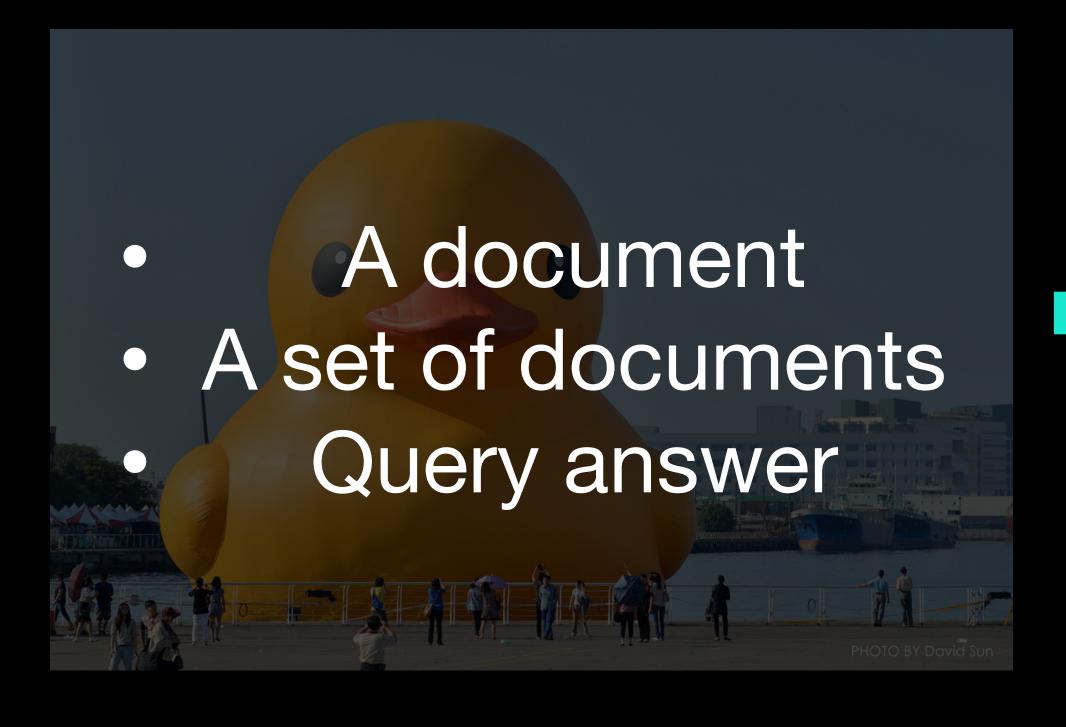


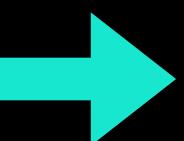
Large Object (Encoded as a set of strings)

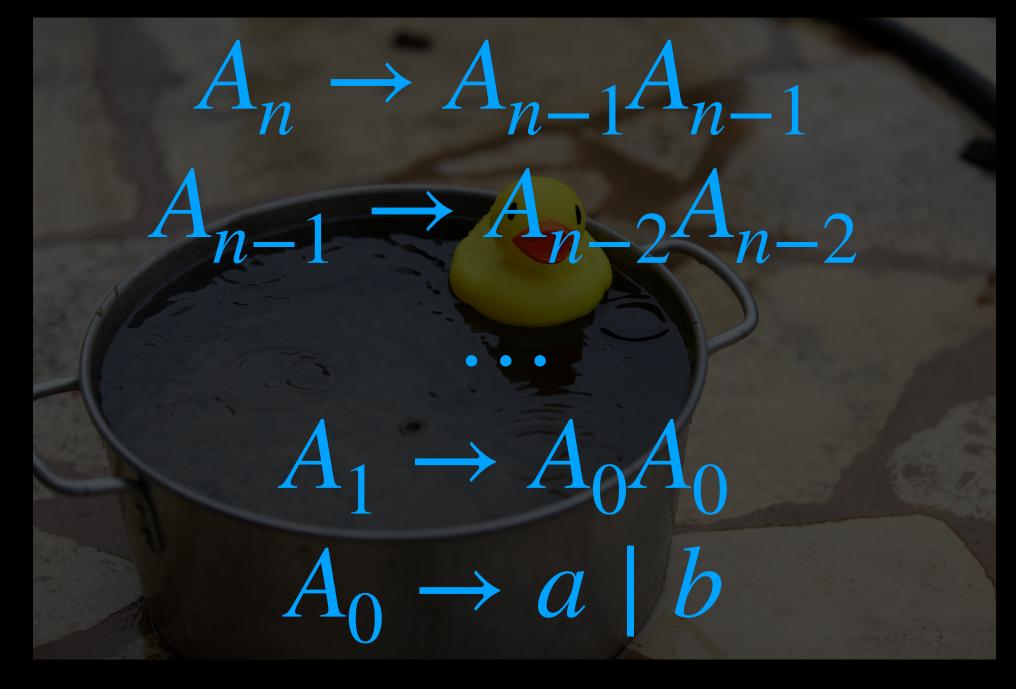




Large Object (Encoded as a set of strings)







Large Object (Encoded as a set of strings)

Motivation

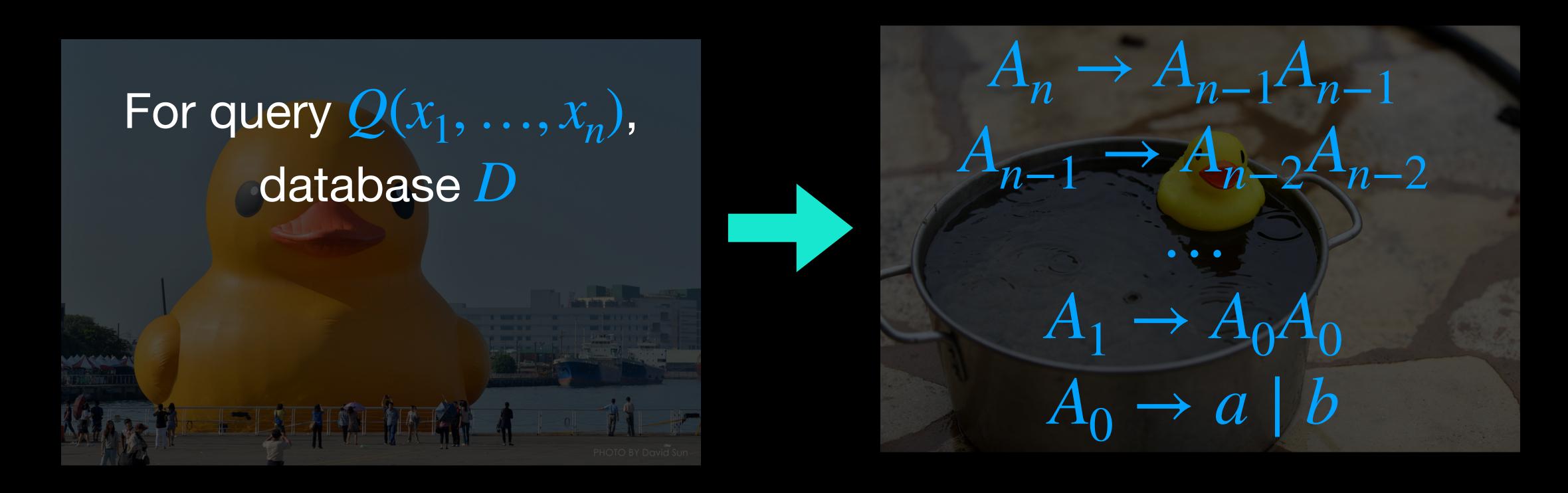
CFGs to Represent Query Answers



Large Object = Q(D)(Encoded as a set of strings)

Motivation

CFGs to Represent Query Answers

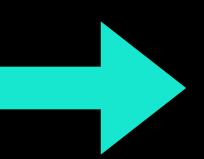


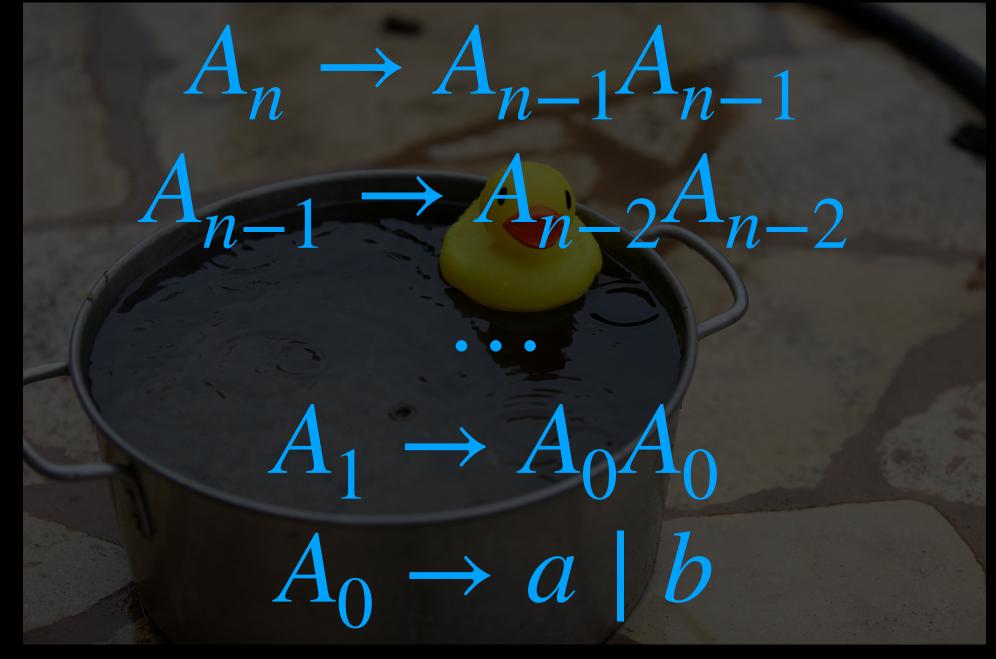
Large Object = Q(D)(Encoded as a set of strings)

Motivation

CFGs to Represent Query Answers

For query $Q(x_1, ..., x_n)$, database D identify $(a_1, ..., a_n) \in Q(D)$ with word $a_1a_2...a_n$





Large Object = Q(D)(Encoded as a set of strings)

Motivation CFGs to Represent Query Answers

• This idea recently proposed by Kimelfeld, Martens and Niewerth.

Kimelfeld, Martens and Niewerth: A Formal Language Perspective on Factorized Representations. ICDT (2025)

Motivation CFGs to Represent Query Answers

- This idea recently proposed by Kimelfeld, Martens and Niewerth.
- For bounded treewidth CQs can compute 'small' CFG representations directly from the query and the database.

Motivation CFGs to Represent Query Answers

- This idea recently proposed by Kimelfeld, Martens and Niewerth.
- For bounded treewidth CQs can compute 'small' CFG representations directly from the query and the database.
- Then can perform algorithms directly on CFG, e.g. recent result shows you can do approximate counting!

Meel and de Colnet: #CFG and #DNNF admit FPRAS. Preprint (2024)

Motivation uCFGs to Represent Query Answers

- This idea recently proposed by Kimelfeld, Martens and Niewerth.
- For bounded treewidth CQs can compute 'small' uCFG representations directly from the query and the database.

Motivation uCFGs to Represent Query Answers

- This idea recently proposed by Kimelfeld, Martens and Niewerth.
- For bounded treewidth CQs can compute 'small' uCFG representations' directly from the query and the database.
- More algorithmically powerful! Now can do exact counting as well as more efficient enumeration.

Motivation uCFGs to Represent Query Answers

This idea recently proposed by Kimelfeld Martens and Niewerth.

For bounded treewing directly from the que

But are there situations where unambiguity comes at a price?

representations

 More algorithmically powerful! Now can do exact counting as well as more efficient enumeration.

Our Contribution Background

Background

Let $L_n \subseteq \{a, b\}^{2n}$ contain all words containing two a symbols at distance n from each other.

Background

Let $L_n \subseteq \{a,b\}^{2n}$ contain all words containing two a symbols at distance n from each other.

 $aaaa \in L_2$

Background

Let $L_n \subseteq \{a,b\}^{2n}$ contain all words containing two a symbols at distance n from each other.

$$aaaa \in L_2$$

$$abab \in L_2$$

Background

Let $L_n \subseteq \{a,b\}^{2n}$ contain all words containing two a symbols at distance n from each other.

$$aaaa \in L_2$$

$$abab \in L_2$$

$$abba \not\in L_2$$

Background

Let $L_n \subseteq \{a, b\}^{2n}$ contain all words containing two a symbols at distance n from each other.

Then L_n admits a CFG of size $\Theta(\log(n))$.

Background

Let $L_n \subseteq \{a, b\}^{2n}$ contain all words containing two a symbols at distance n from each other.

Then L_n admits a CFG of size $\Theta(\log(n))$.

Kimelfeld, Martens and Niewerth: A Formal Language Perspective on Factorized Representations. ICDT (2025)

Background

Let $L_n \subseteq \{a,b\}^{2n}$ contain all words containing two a symbols at distance n from each other.

Then L_n admits a CFG of size $\Theta(\log(n))$.

Best possible compression.

Kimelfeld, Martens and Niewerth: A Formal Language Perspective on Factorized Representations. ICDT (2025)

Background

Let $L_n \subseteq \{a,b\}^{2n}$ contain all words containing two a symbols at distance n from each other.

Conjecture: Every uCFG for L_n has size $2^{\Omega(n)}$.

Kimelfeld, Martens and Niewerth: A Formal Language Perspective on Factorized Representations. ICDT (2025)

Let $L_n \subseteq \{a,b\}^{2n}$ contain all words containing two a symbols at distance n from each other.

Theorem: Every uCFG for L_n has size $2^{\Omega(n)}$.

Theorem: Every uCFG for L_n has size $2^{\Omega(n)}$.

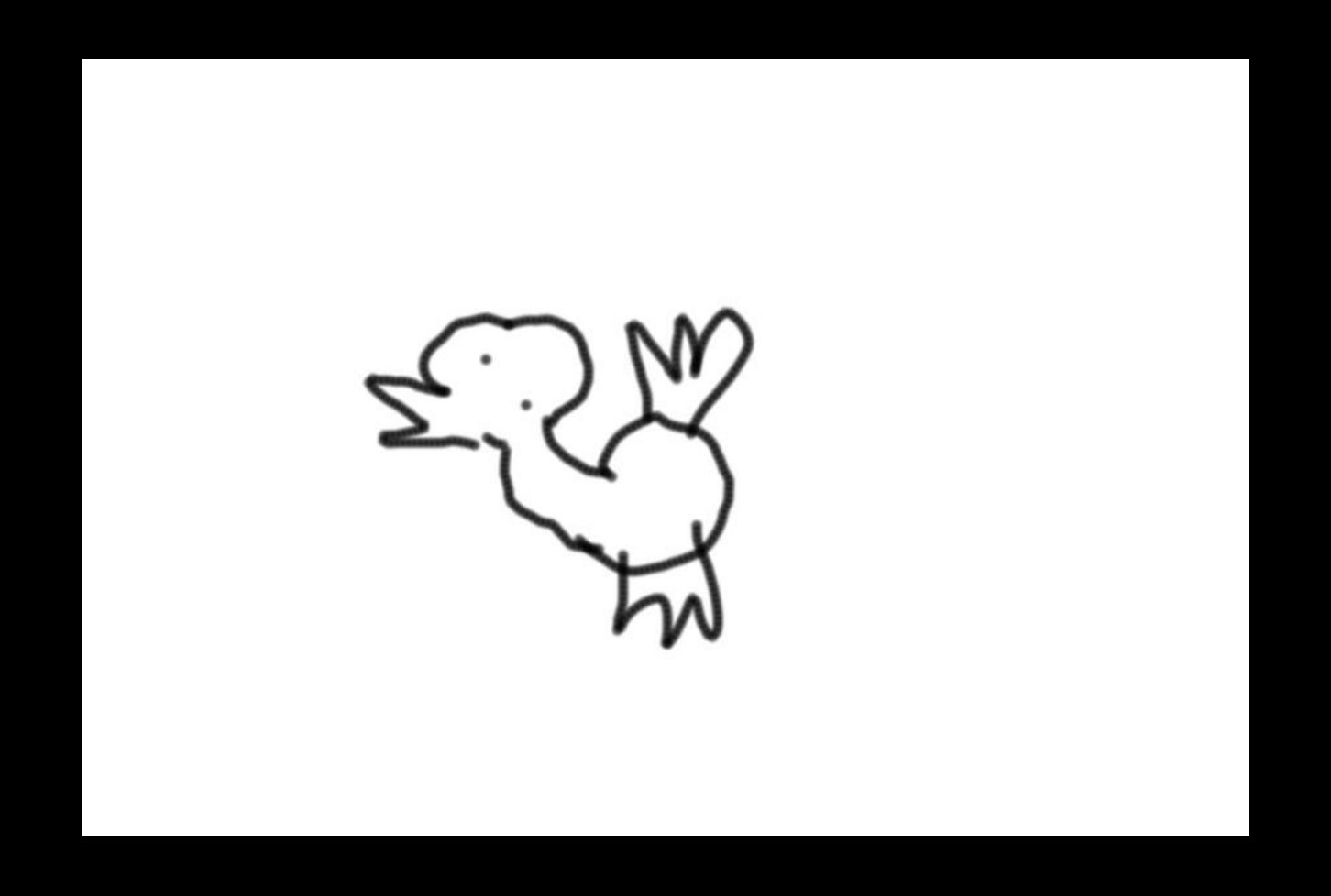
Theorem: Every uCFG for L_n has size $2^{\Omega(n)}$.

⇒ double exponential succinctness separation CFG vs uCFG for finite languages.

Theorem: Every uCFG for L_n has size $2^{\Omega(n)}$.

⇒ double exponential succinctness separation CFG vs uCFG for finite languages.

⇒ exponential succinctness separation NFA vs uCFG for finite languages.



From (u)CFG to Rectangle Covers

Let $L \subseteq \Sigma^n$, G be a CFG accepting L.

From (u)CFG to Rectangle Covers

Let $L \subseteq \Sigma^n$, G be a CFG accepting L.

Lemma: $L = \bigcup_{i=1}^{\ell} R_i, \text{ where each } R_i \text{ is a }$ rectangle and $\ell \leq n \mid G \mid$.

From (u)CFG to Rectangle Covers

Let $L \subseteq \Sigma^n$, G be a CFG accepting L.

Lemma: $L = \bigcup_{i=1}^{\ell} R_i, \text{ where each } R_i \text{ is a } i=1$ rectangle and $\ell \leq n \mid G \mid$.

Rectangle:

From (u)CFG to Rectangle Covers

Let $L \subseteq \Sigma^n$, G be a CFG accepting L.

Lemma: $L = \bigcup_{i=1}^{\ell} R_i, \text{ where each } R_i \text{ is a}$ rectangle and $\ell \leq n |G|$.

Rectangle: $R = \{xy \mid x \in L_1, y \in L_2\}$

From (u)CFG to Rectangle Covers

Let $L \subseteq \Sigma^n$, G be a CFG accepting L.

Lemma: $L = \bigcup_{i=1}^{\ell} R_i, \text{ where each } R_i \text{ is a } i=1$ rectangle and $\ell \leq n |G|$.

Rectangle: $R = \{xy \mid x \in L_1, y \in L_2\}$ $L_1 \subseteq \Sigma^{n/2}$

From (u)CFG to Rectangle Covers

Let $L \subseteq \Sigma^n$, G be a CFG accepting L.

Lemma: $L = \bigcup_{i=1}^{\ell} R_i, \text{ where each } R_i \text{ is a } i=1$ rectangle and $\ell \leq n |G|$.

Rectangle: $R = \{xy \mid x \in L_1, y \in L_2\}$ $L_1 \subseteq \Sigma^{n/2}$ $L_2 \subseteq \Sigma^{n/2}$

From (u)CFG to Rectangle Covers

Let $L \subseteq \Sigma^n$, G be a CFG accepting L.

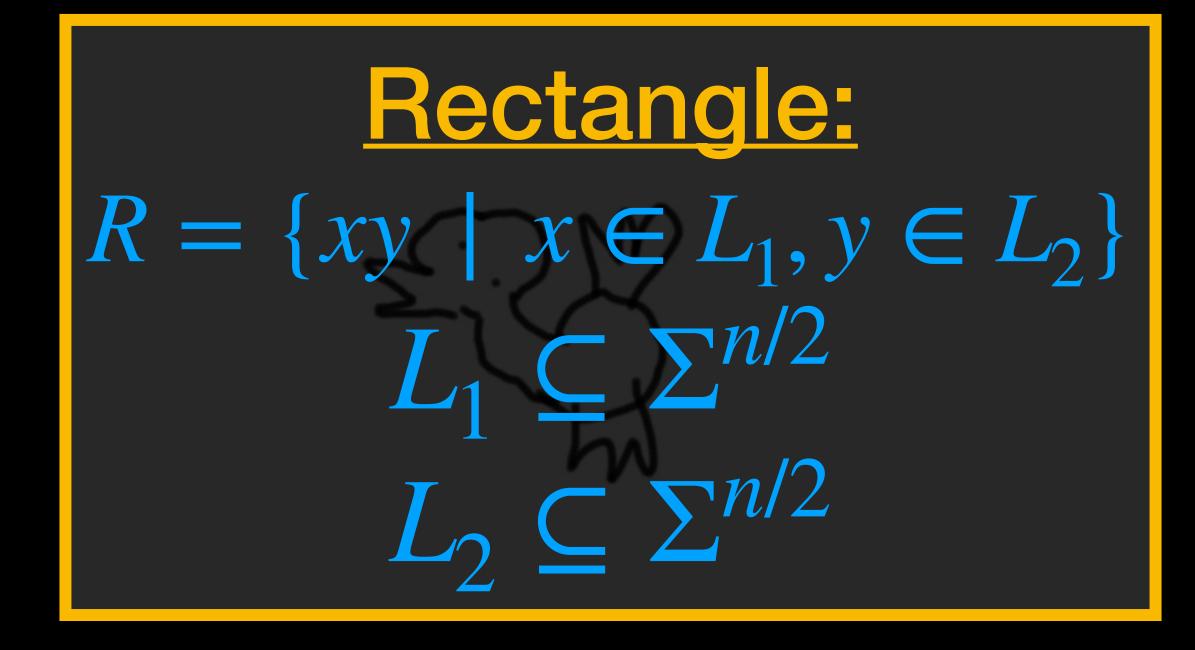
Lemma: $L = \bigcup_{i=1}^{\ell} R_i, \text{ where each } R_i \text{ is a }$ rectangle and $\ell \leq n \mid G \mid$.

```
Rectangle:
R = \{xy \mid x \in L_1, y \in L_2\}
L_1 \subseteq \Sigma^{n/2}
L_2 \subseteq \Sigma^{n/2}
```

From (u)CFG to Rectangle Covers

Let $L \subseteq \Sigma^n$, G be a uCFG accepting L.

Lemma: $L = \biguplus R_i, \text{ where each } R_i \text{ is a}$ = 1 rectangle and $\ell \leq n \mid G \mid$.



From (u)CFG to Rectangle Covers

Let $L \subseteq \Sigma^n$, G be a uCFG accepting L.

Lemma: $L = \biguplus R_i, \text{ where each } R_i \text{ is a}$ $\underset{i=1}{i=1}$ rectangle and $\ell \leq n \mid G \mid$.

ldea:

Show every disjoint rectangle cover of L_n has size $2^{\Omega(n)}$.

From (u)CFG to F

Let L

Difficulty:

how to make use of disjointness?

epting L.

Lemma: $L = \biguplus R_i, \text{ where each } R_i \text{ is a}$ i=1 rectangle and $\ell \leq n \mid G \mid$.

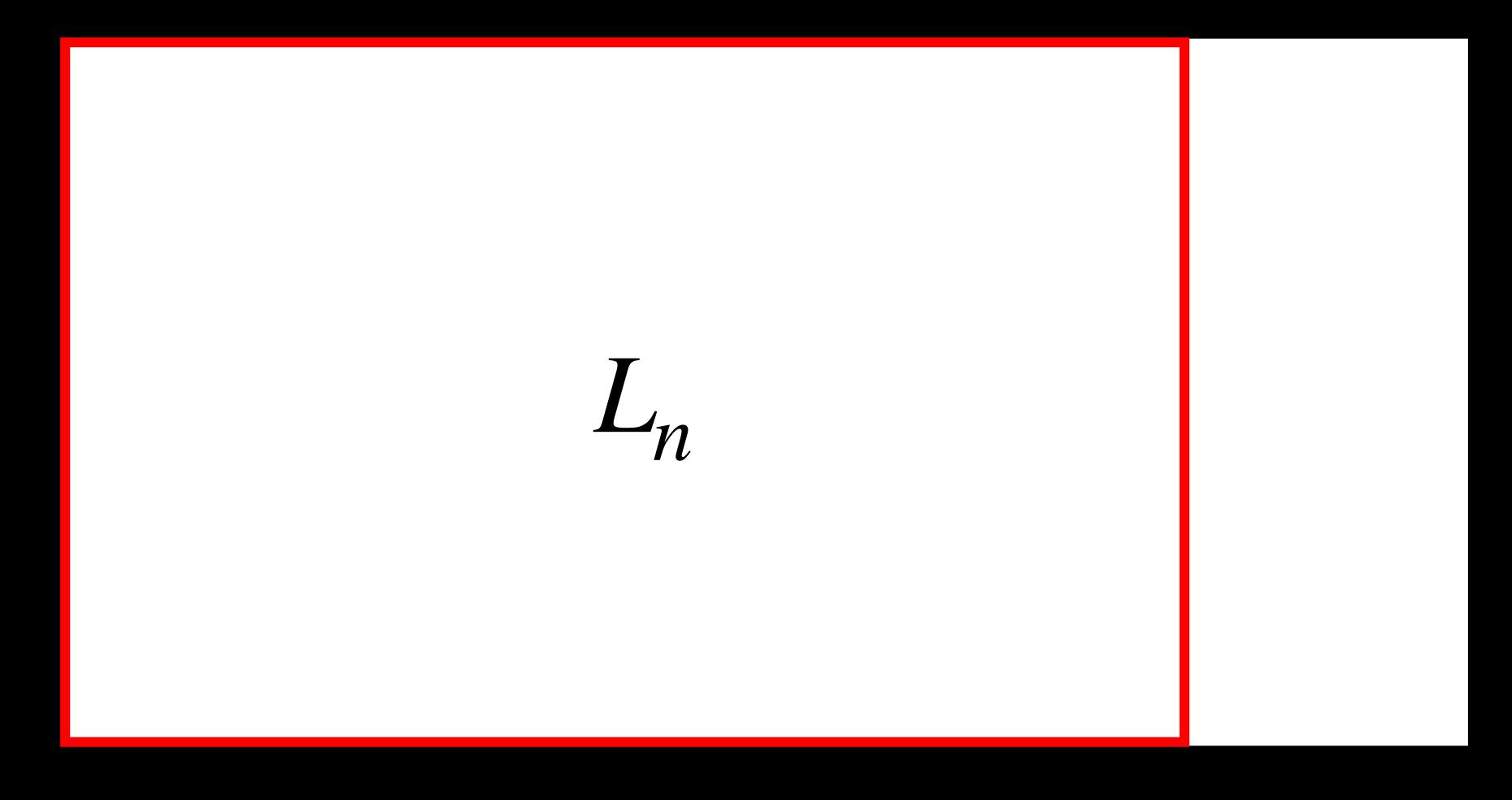
<u>ldea:</u>

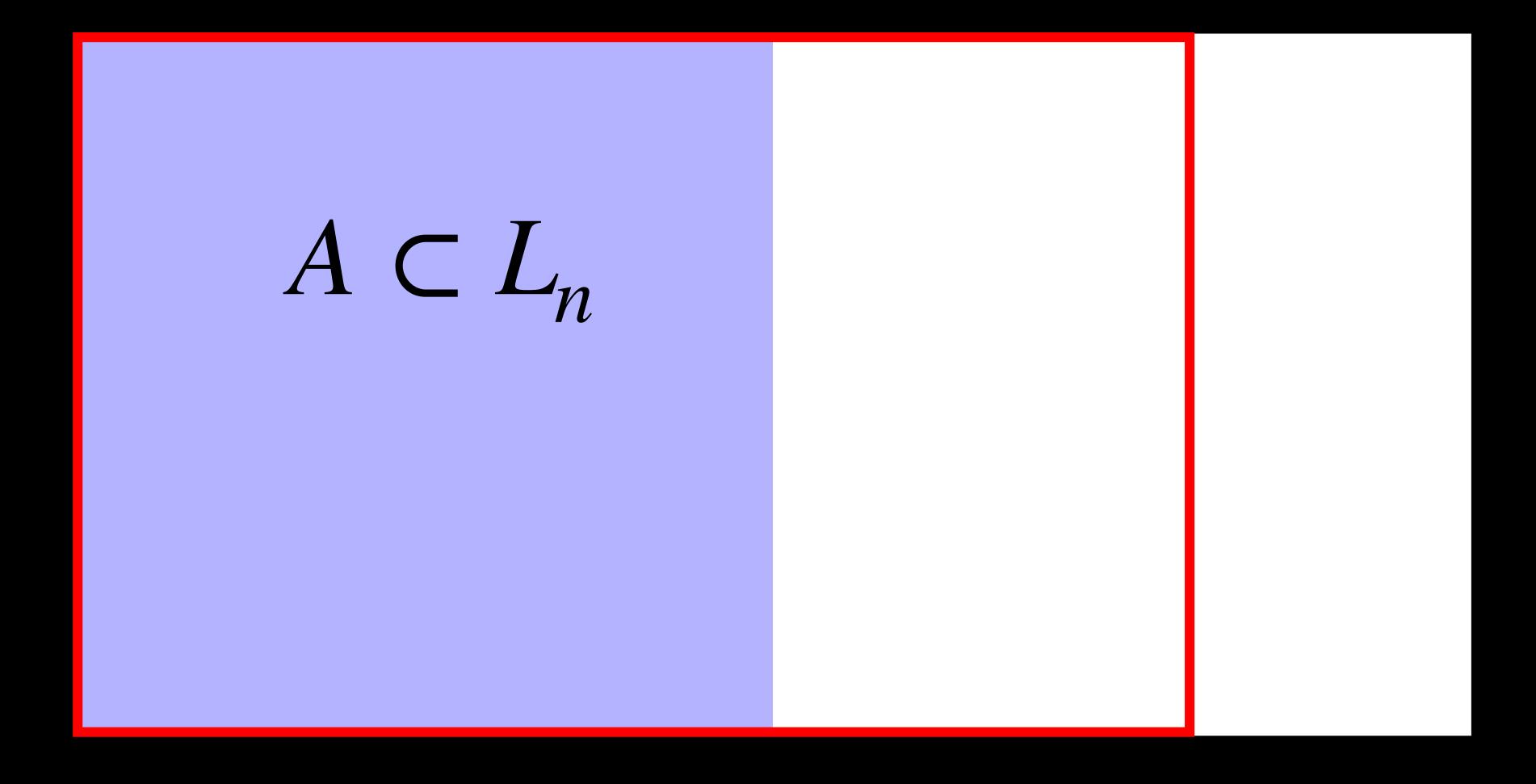
Show every disjoint rectangle cover of L_n has size $2^{\Omega(n)}$.

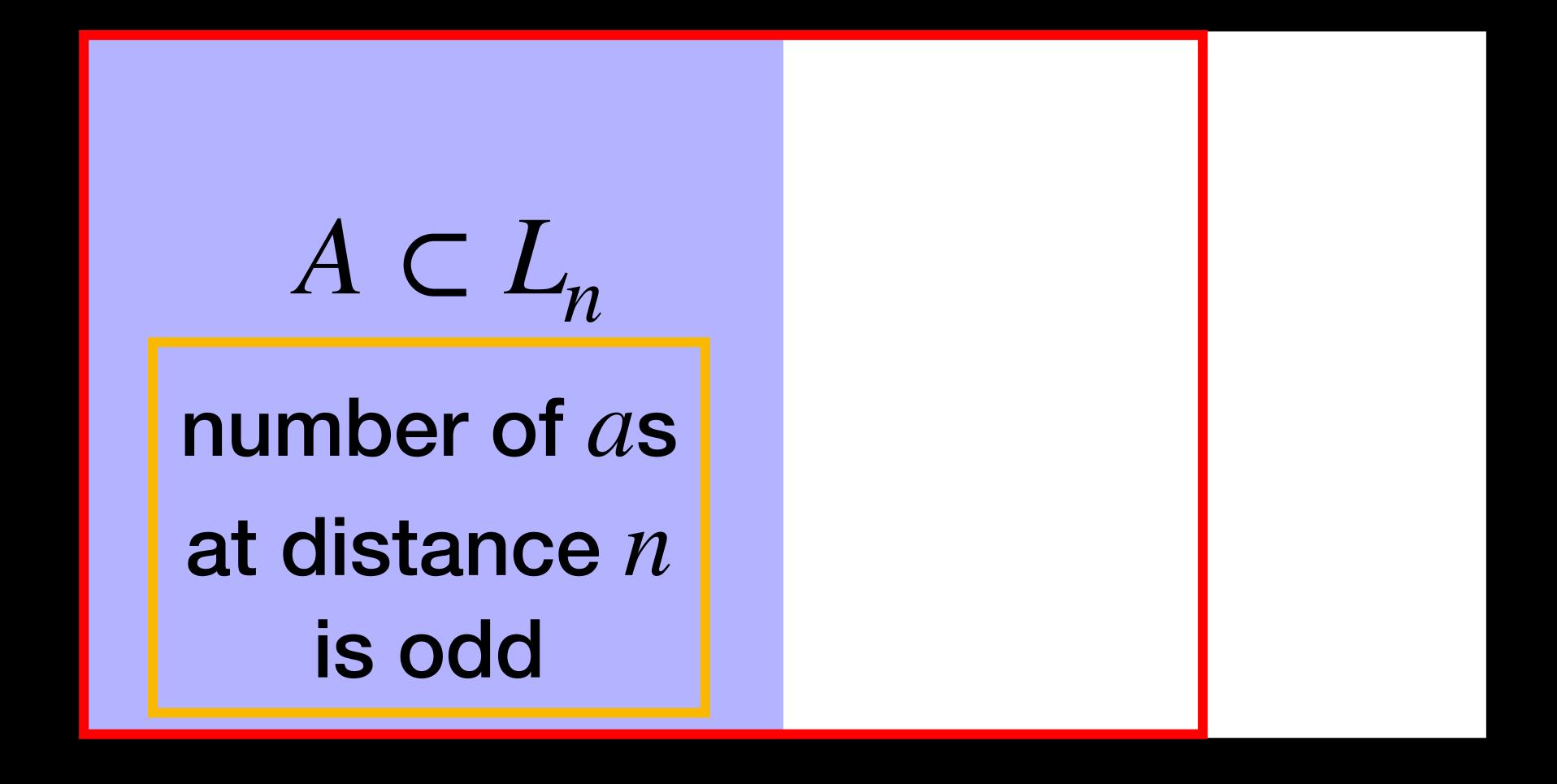
Discrepancy Argument

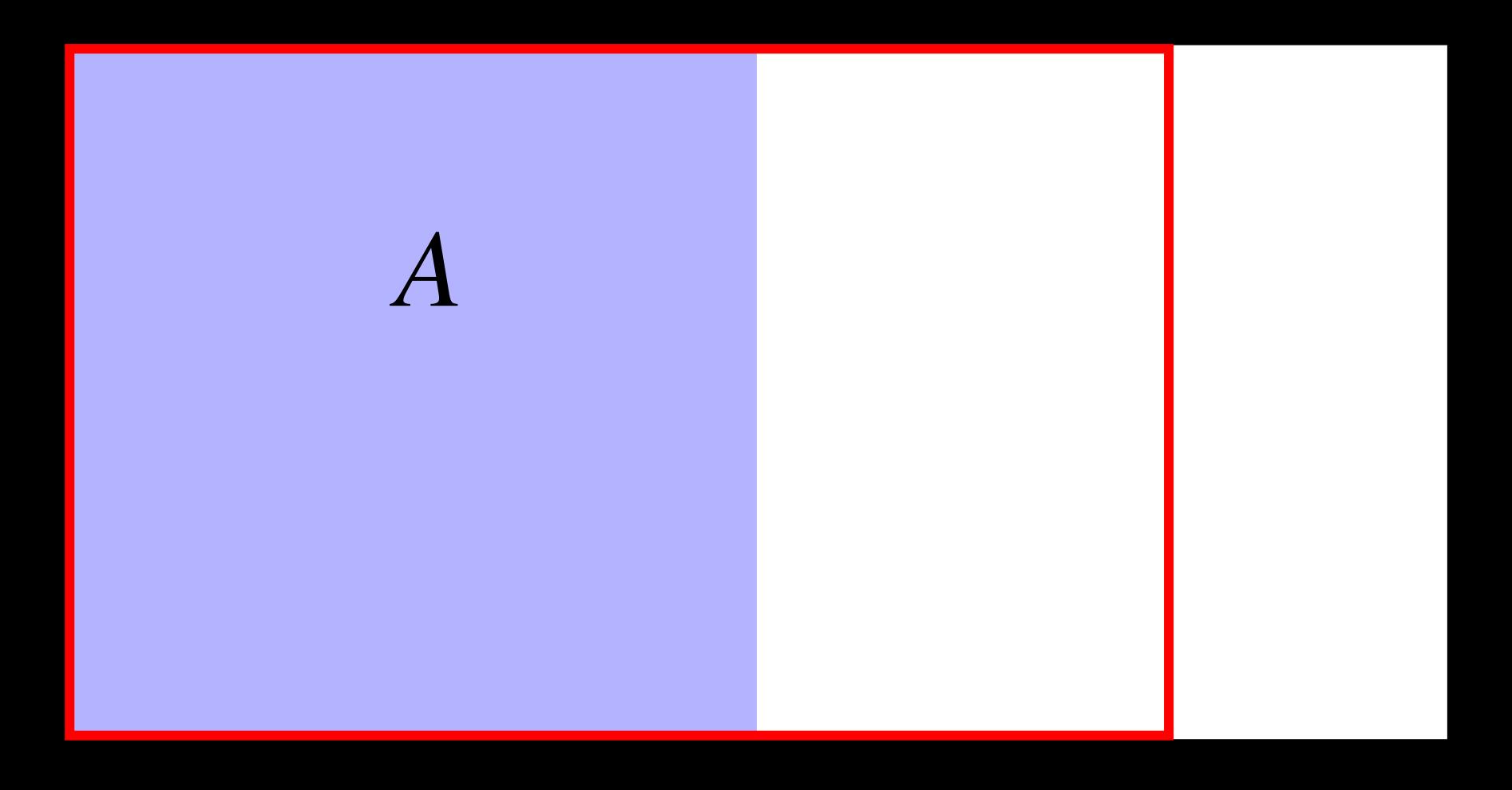
All words over $\{a, b\}$ of length n

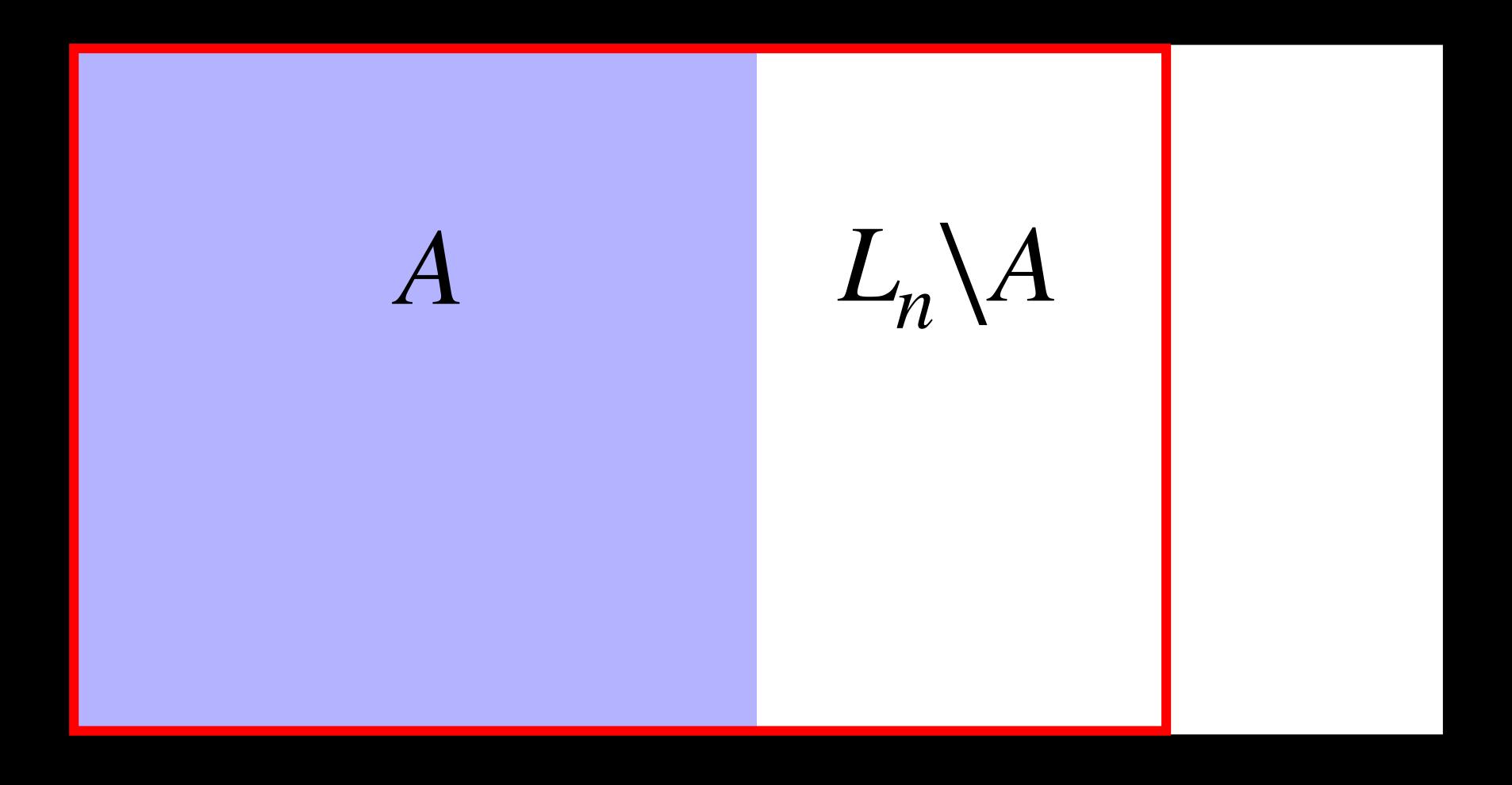
Discrepancy Argument

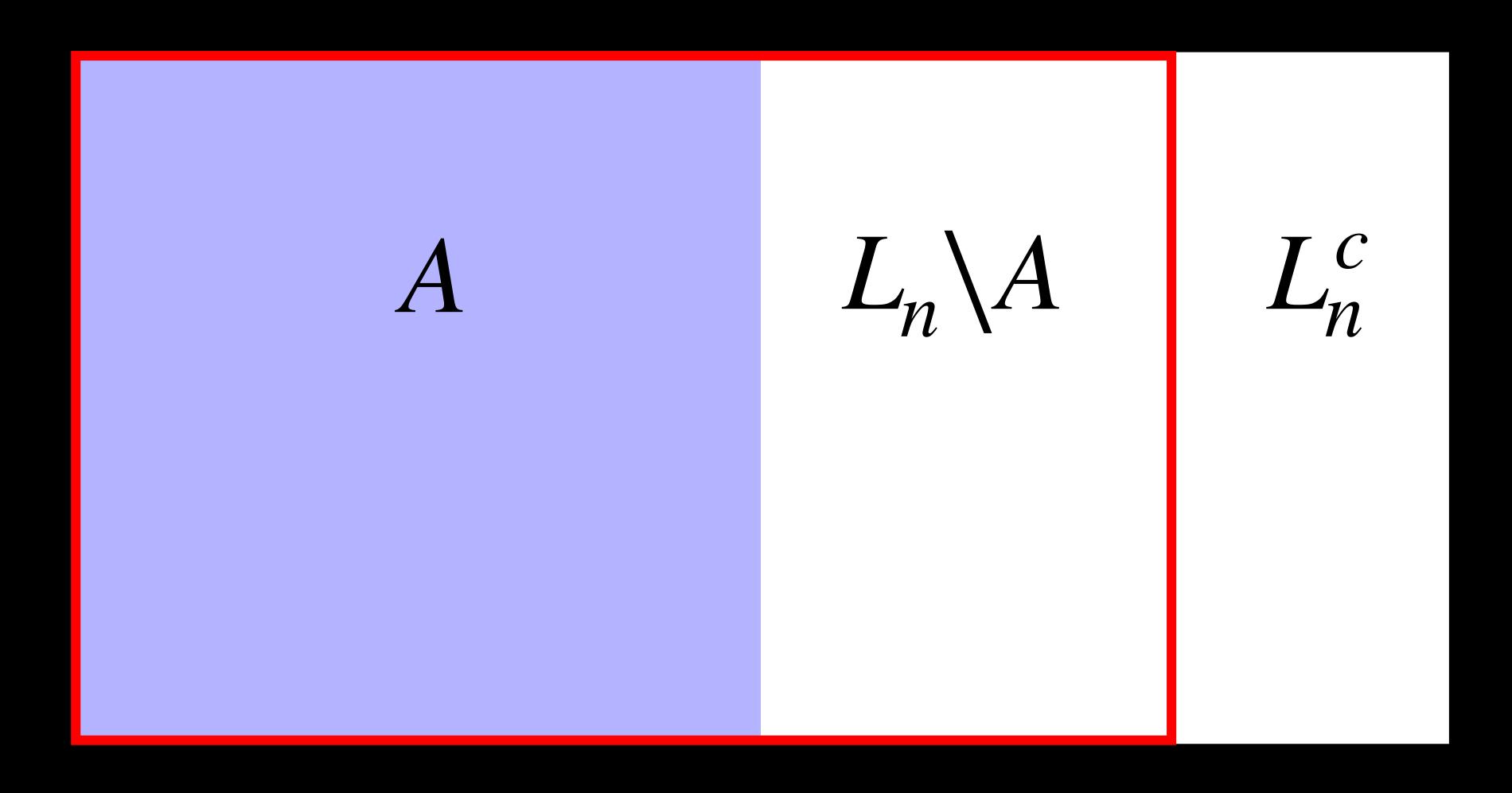




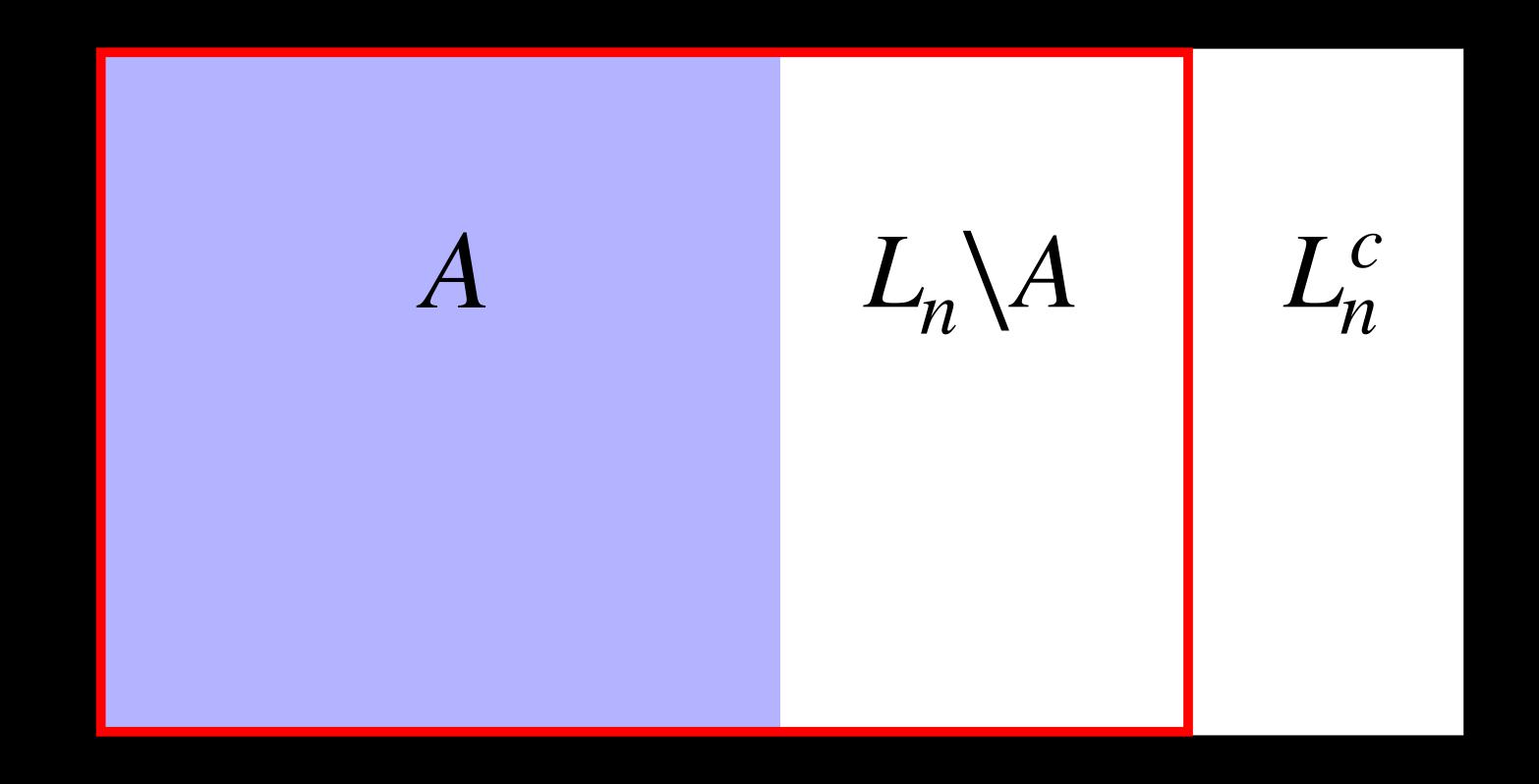




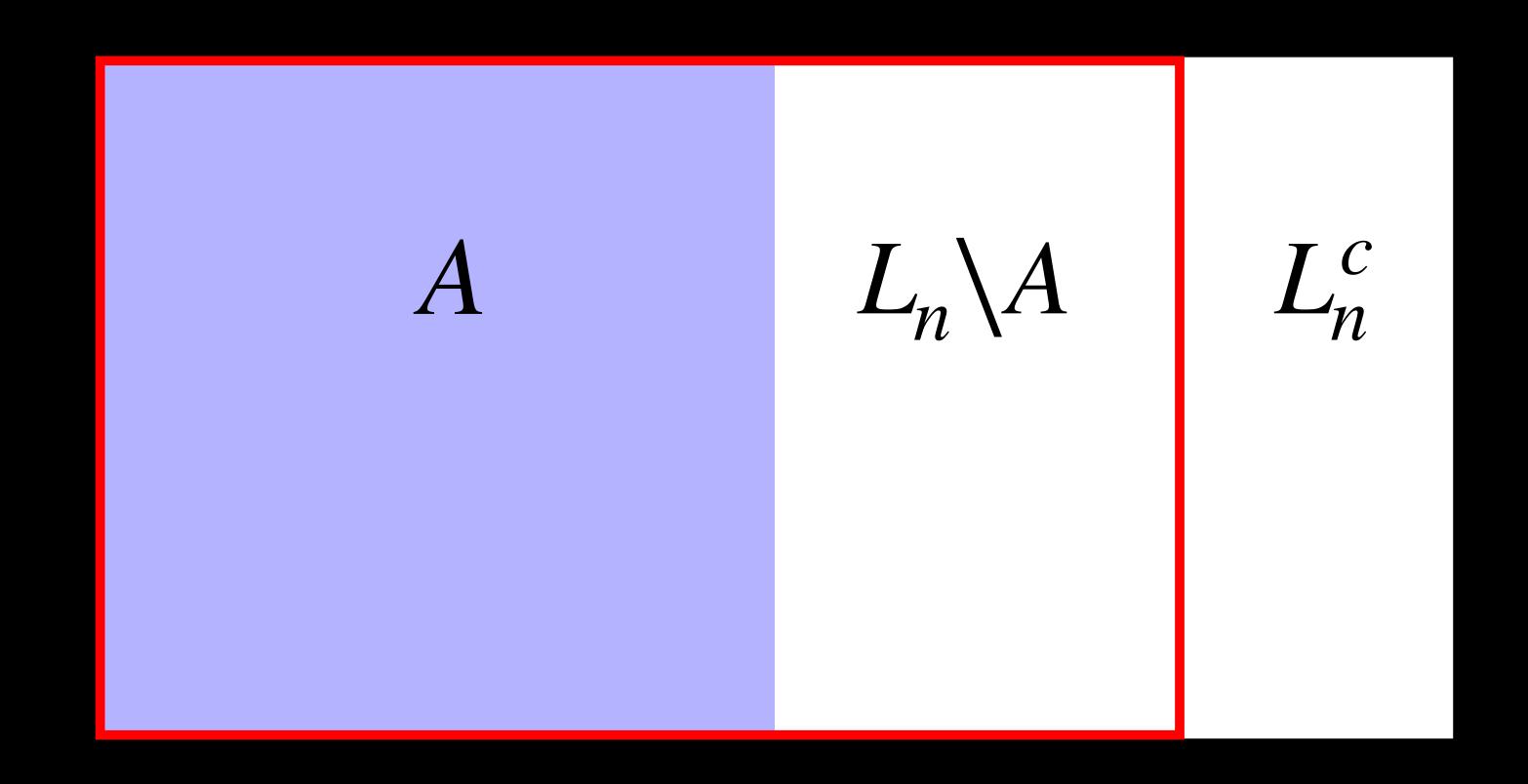




Facts:



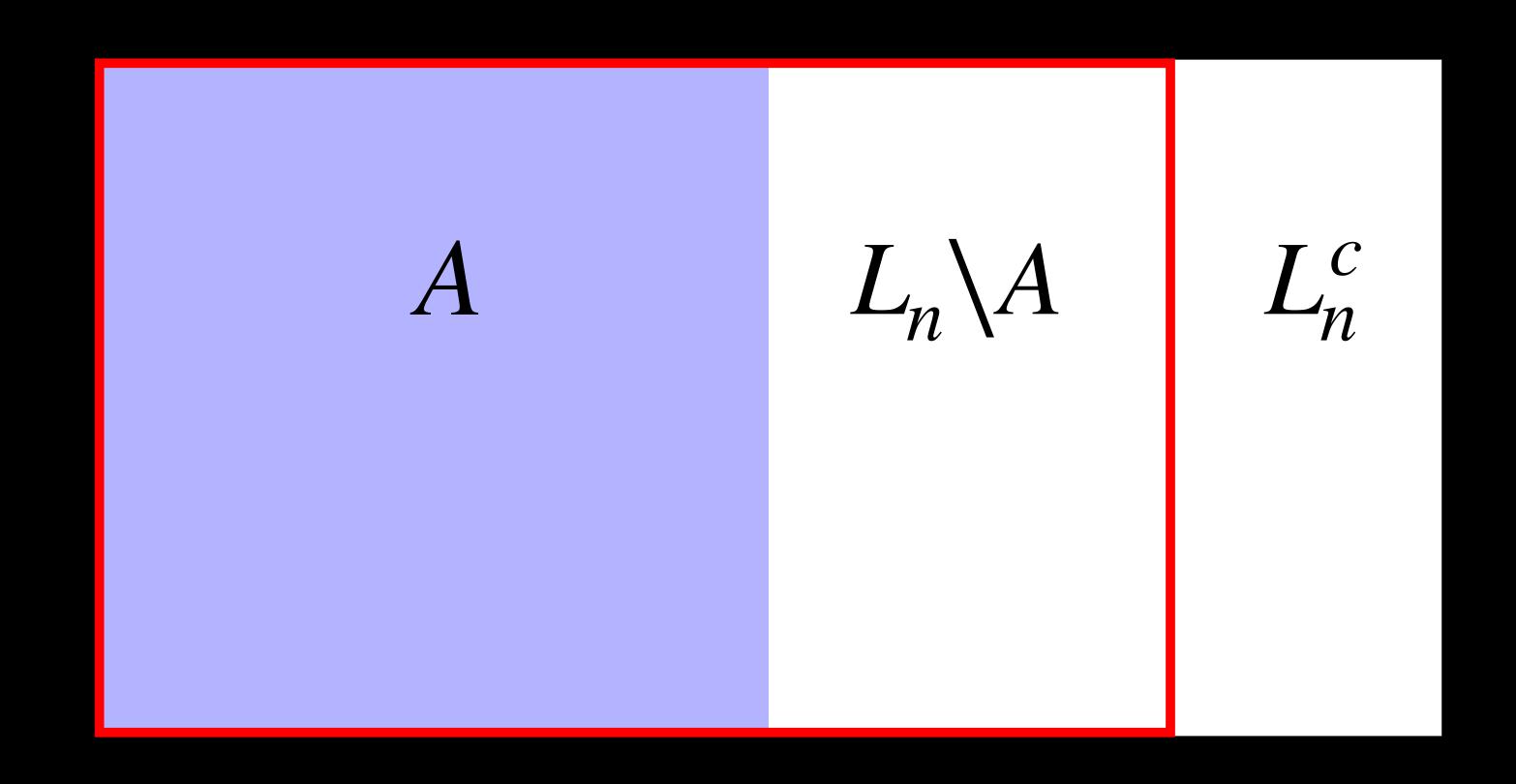
$$\frac{\text{Facts:}}{\text{(1)} |A| \gg |L_n \backslash A|}.$$



Facts:

$$(1) |A| \gg |L_n \backslash A|.$$

(2) $R \cap A \approx R \setminus A$

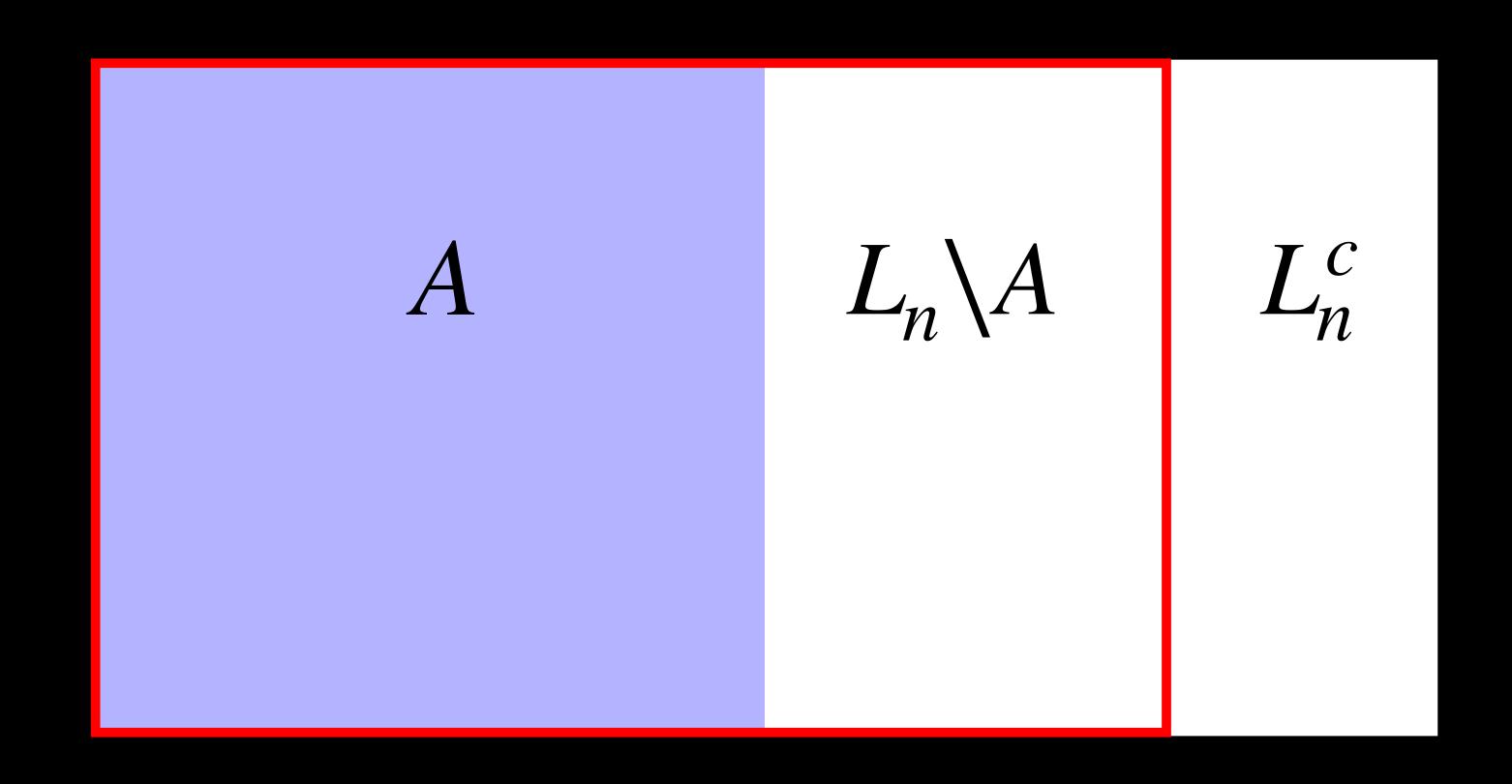


Discrepancy Argument

Facts:

$$(1) |A| \gg |L_n \backslash A|.$$

(2) $|R \cap A| \approx |R \setminus A|$ for all *big* rectangles R.

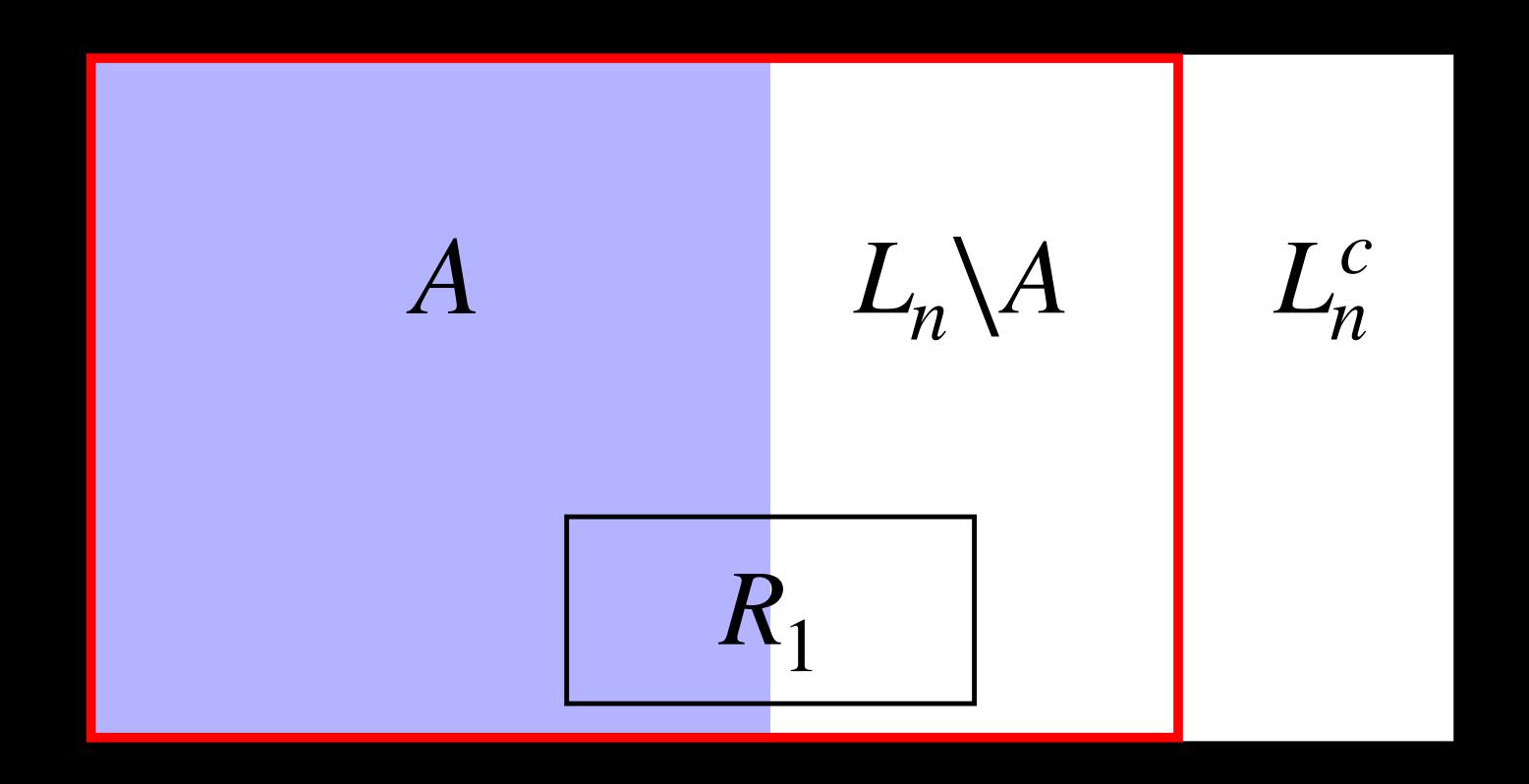


Discrepancy Argument

Facts:

$$(1) |A| \gg |L_n \backslash A|.$$

(2) $|R \cap A| \approx |R \setminus A|$ for all *big* rectangles R.



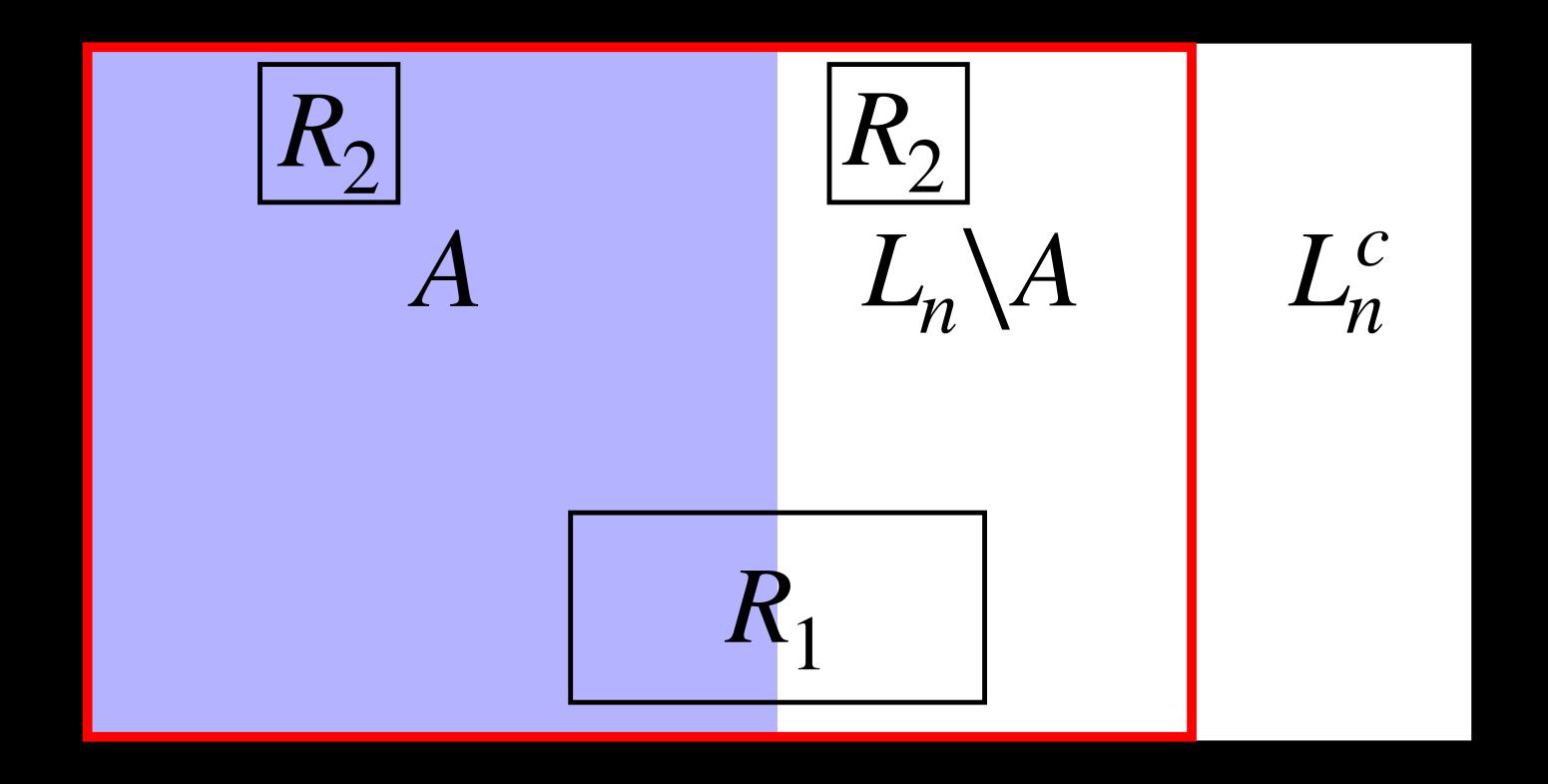
Proof Sketch

Discrepancy Argument

Facts:

$$(1) |A| \gg |L_n \backslash A|.$$

(2) $|R \cap A| \approx |R \setminus A|$ for all *big* rectangles R.



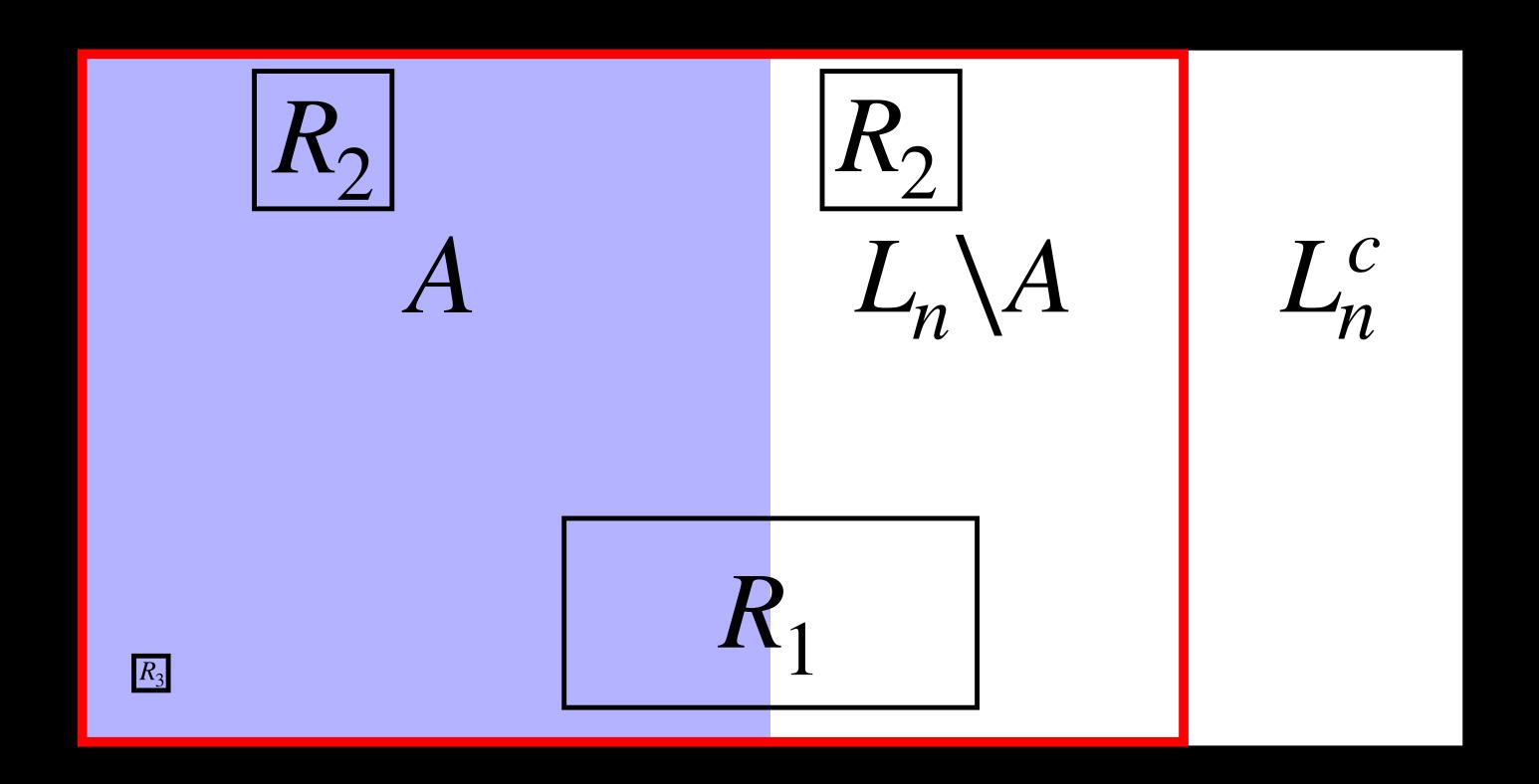
Proof Sketch

Discrepancy Argument

Facts:

(1)
$$|A| \gg |L_n \backslash A|$$
.

(2) $|R \cap A| \approx |R \setminus A|$ for all *big* rectangles R.



Proof Sketch

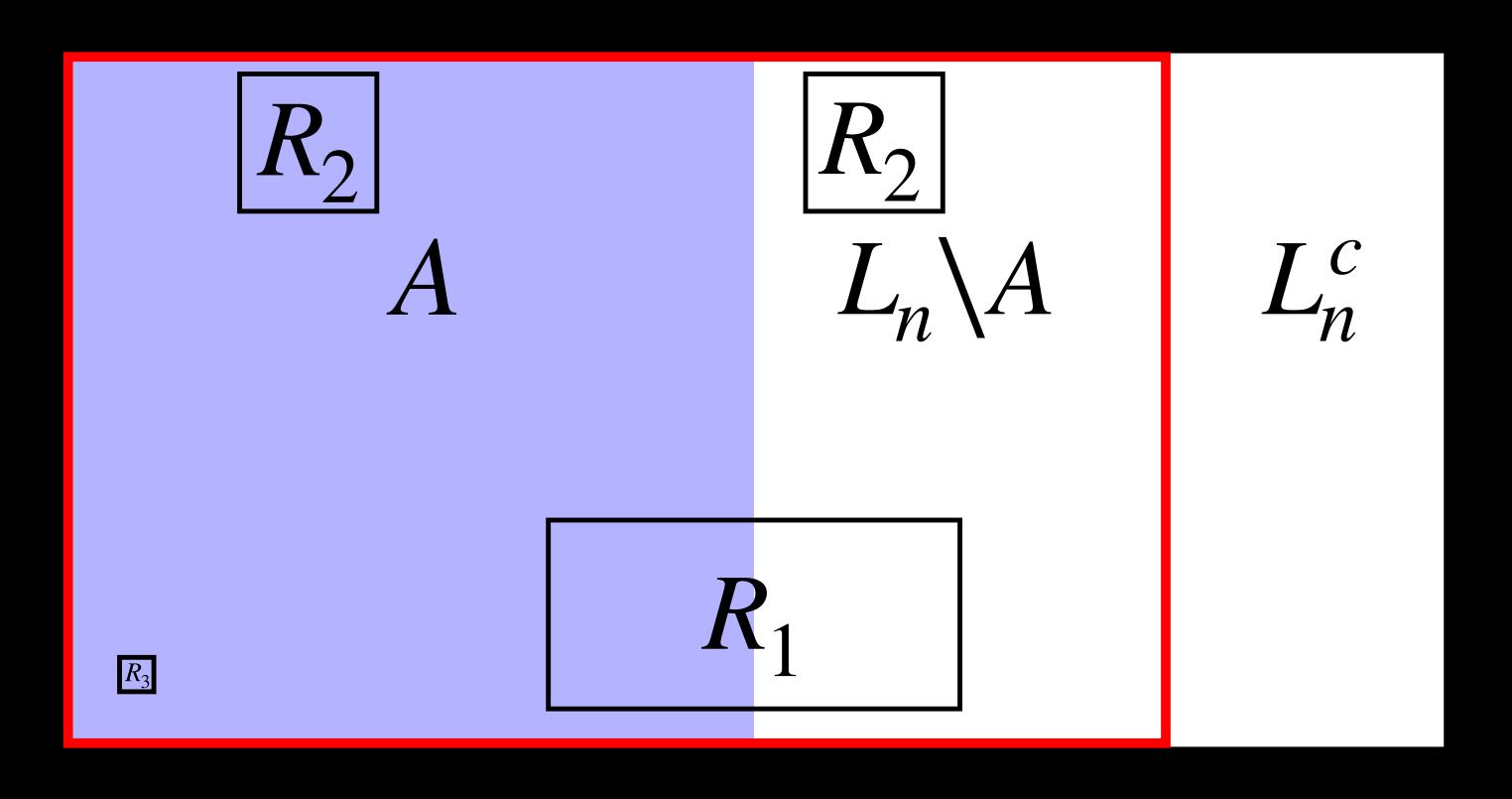
Discrepancy Argument

Facts:

$$(1) |A| \gg |L_n \backslash A|.$$

(2) $|R \cap A| \approx |R \setminus A|$ for all *big* rectangles R.

So a disjoint rectangle cover of L_n must use a lot of small rectangles to cover A.



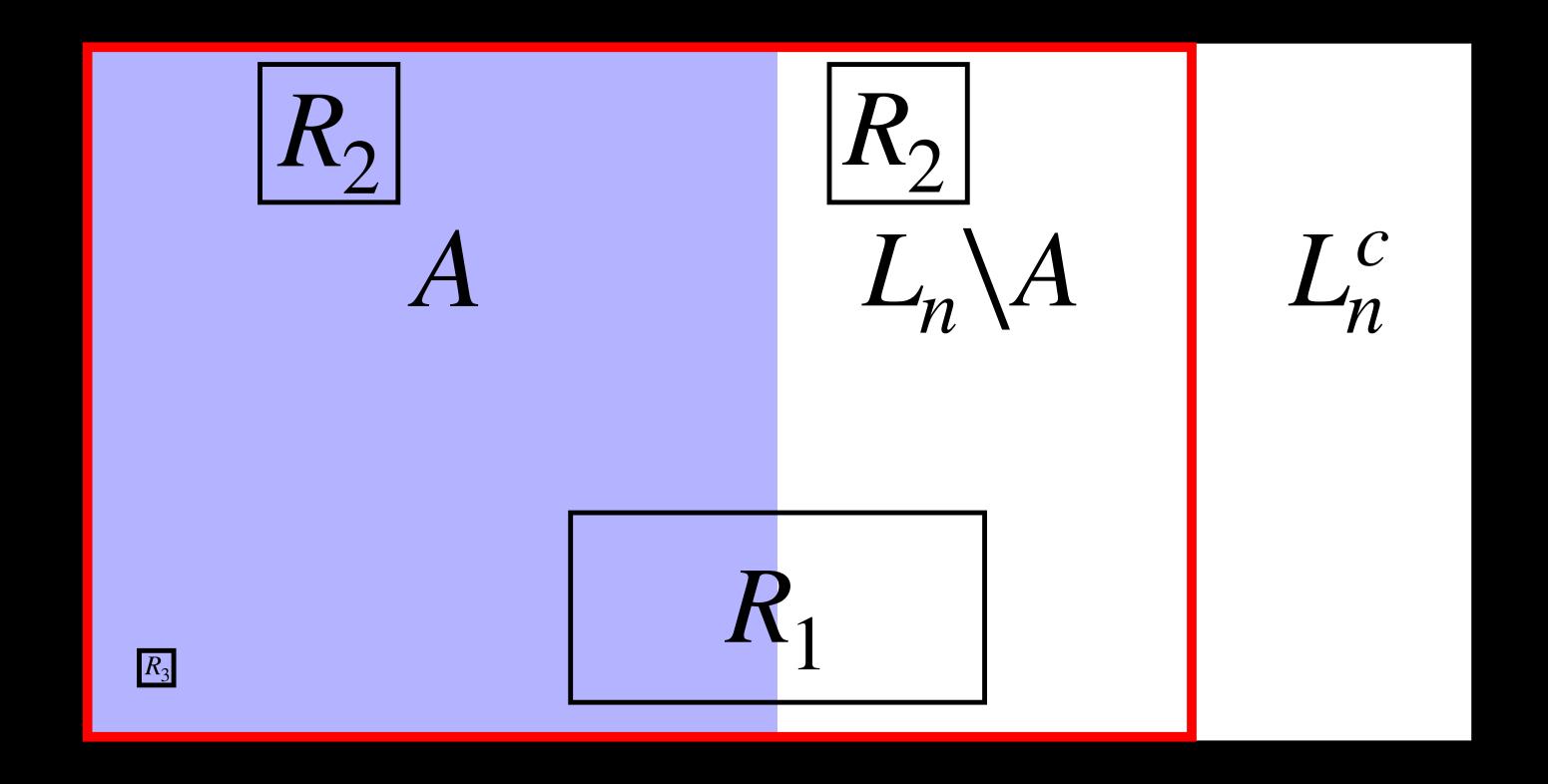
Proof Sketch Discrepancy Argument

So every disjoint rectangle cover of L_n must be big.

Facts:

$$(1) |A| \gg |L_n \backslash A|.$$

(2) $|R \cap A| \approx |R \setminus A|$ for all *big* rectangles R.



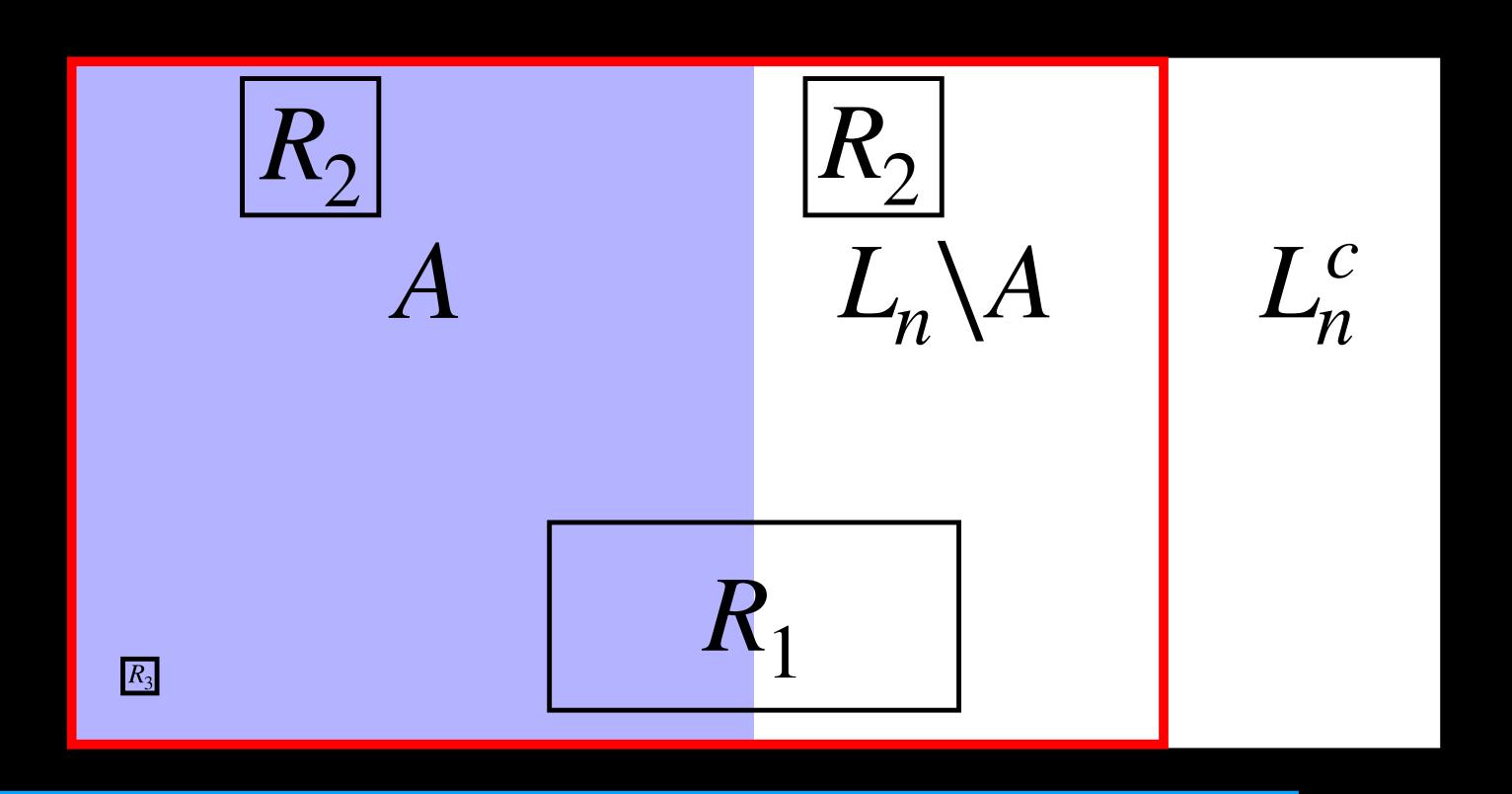
Proof Sketch Discrepancy Argument

So every uCFG accepting L_n must be big.

Facts:

$$(1) |A| \gg |L_n \backslash A|.$$

(2) $|R \cap A| \approx |R \setminus A|$ for all *big* rectangles R.



Sherstov: The multiparty communication complexity of set disjointness. SIAM Journal on Computing (2016)

• Showed a optimal **double-exponential** succinctness separation of CFGs from uCFGs.

- Showed a optimal **double-exponential** succinctness separation of CFGs from uCFGs.
- Lower bound on a natural language: could be basis for other lower bounds.

- Showed a optimal double-exponential succinctness separation of CFGs from uCFGs.
- Lower bound on a natural language: could be basis for other lower bounds.
- Can we use similar techniques to answer other questions on ambiguity?

- Showed a optimal **double-exponential** succinctness separation of CFGs from uCFGs.
- Lower bound on a natural language: could be basis for other lower bounds.
- Can we use similar techniques to answer other questions on ambiguity?
 - Particularly intriguing: understanding negation!

- Showed a optimal double-exponential succinctness separation of CFGs from uCFGs.
- Lower bound on a natural language: could be basis for other lower bounds.
- Can we use similar techniques to answer other questions on ambiguity?
 - Particularly intriguing: understanding negation!
 - Recent quasi-poly lower bounds for UFA and structured d-DNNF.

Göös, Kiefer, and Yuan. "Lower Bounds for Unambiguous Automata via Communication Complexity." ICALP (2022)

H.V.S.: "Structured d-DNNF is Not Closed Under Negation." IJCAI (2024).

- Showed a optimal double-exponential succinctness separation of CFGs from uCFGs.
- Lower bound on a natural language: could be basis for other lower bounds.
- Can we use similar techniques to answer other questions on ambiguity?
 - Particularly intriguing: understanding negation!
 - Recent quasi-poly lower bounds for UFA and structured d-DNNF.
 - What about e.g. uCFG or d-DNNF?

Göös, Kiefer, and Yuan. "Lower Bounds for Unambiguous Automata via Communication Complexity." ICALP (2022)

H.V.S.: "Structured d-DNNF is Not Closed Under Negation." IJCAI (2024).

$$R = \{x_1yx_2 \mid x_1x_2 \in L_1, y \in L_2, |x_1| = a_1, |x_2| = a_2\}$$

$$R = \{x_1 y x_2 \mid x_1 x_2 \in L_1, y \in L_2, |x_1| = a_1, |x_2| = a_2\}$$

$$L_1 \subseteq \Sigma^{a_1 + a_2}$$

$$R = \{x_1 y x_2 \mid x_1 x_2 \in L_1, y \in L_2, |x_1| = a_1, |x_2| = a_2\}$$

$$L_1 \subseteq \Sigma^{a_1 + a_2}$$

$$L_2 \subseteq \Sigma^b$$

$$R = \{x_1 y x_2 \mid x_1 x_2 \in L_1, y \in L_2, |x_1| = a_1, |x_2| = a_2\}$$

$$L_1 \subseteq \Sigma^{a_1 + a_2}$$

$$L_2 \subseteq \Sigma^b$$

$$\min(a_1 + a_2, b) \ge n/3$$

$$R = \{x_1 y x_2 \mid x_1 x_2 \in L_1, y \in L_2, |x_1| = a_1, |x_2| = a_2\}$$

$$L_1 \subseteq \Sigma^{a_1 + a_2}$$

$$L_2 \subseteq \Sigma^b$$

$$\min(a_1 + a_2, b) \ge n/3$$

$$a_1 + a_2 + b = n$$

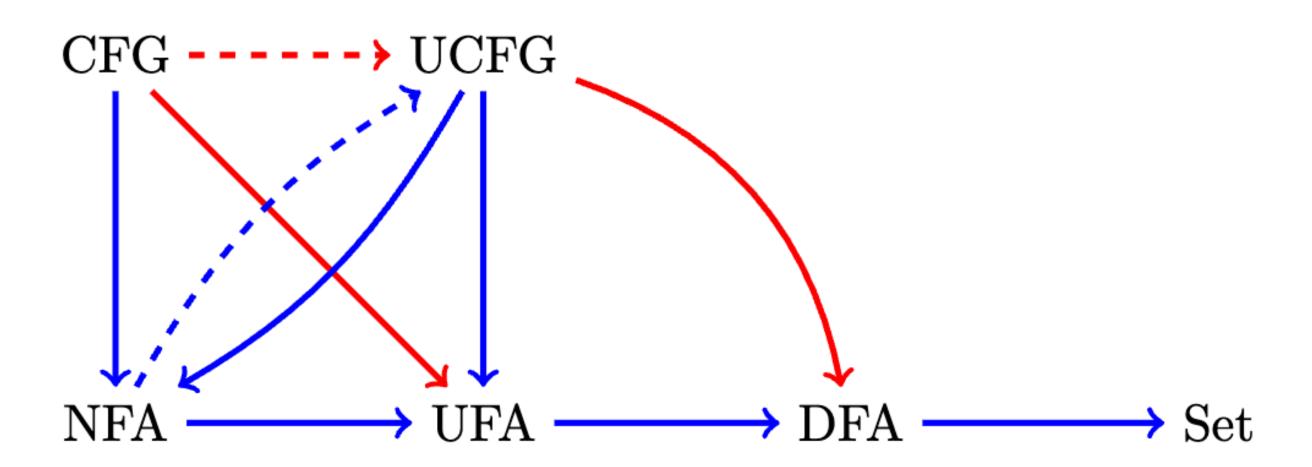


Figure 3 Worst-case unavoidable blow-ups for succinct representations of uniform length relations. Every path that consists of only blue edges represents an unavoidable exponential blow-up and every path that contains at least one red (solid) edge represents an unavoidable double exponential blow-up. If there is no path, then there exists a linear translation. For the dashed edges, we only prove an upper bound. The corresponding lower bounds are conditional on Conjecture 5.7.

Kimelfeld, Martens and Niewerth: A Formal Language Perspective on Factorized Representations. ICDT (2025)

Image Credit

- Big duck: Jaykov https://universe.roboflow.com/jaykov-l5qjv/ducks-eqnsf
 (https://creativecommons.org/licenses/by/4.0/)
- Small Duck: Steve Miller https://www.flickr.com/photos/
 smiller999/29012477655 (https://creativecommons.org/licenses/by-nc-sa/2.0/)
- Bad duck drawing: Bella454578 https://www.deviantart.com/bella454578/ art/A-BAD-DRAWING-OF-A-BIRD-627093890 (https:// creativecommons.org/licenses/by-nd/3.0/)