Scalable Synthesis of Regular Expressions from Only Positive Examples

a{5}[a-z]{2,}

(+)?\d+

\d{2}.\d{2}.\d{4}

[a-z]+

Mark Barbone* and Elizaveta Pertseva*

Advisors: Nadia Polikarpova and Taylor Berg-Kirkpatrick

Defining the Problem

Motivating Example

- Alex is searching for Brazilian CNPJs in a large file
- Quickly finds a couple examples
- What now? Regex+

02.916.265/0001-60

60.701.190/0001-04

→ **Regex+** →

`\d{2}\.\d{3}\.\d{3}/0001-\d{2}`

Challenges

1. Underspecification

"Correct" but too **specific**:
(02\.916\.265/0001-60) OR (60\.701\.190/0001-04)

"Correct" but too **simple**: *.**

2. Search

Contribution 1: Pragmatic Ranking Function

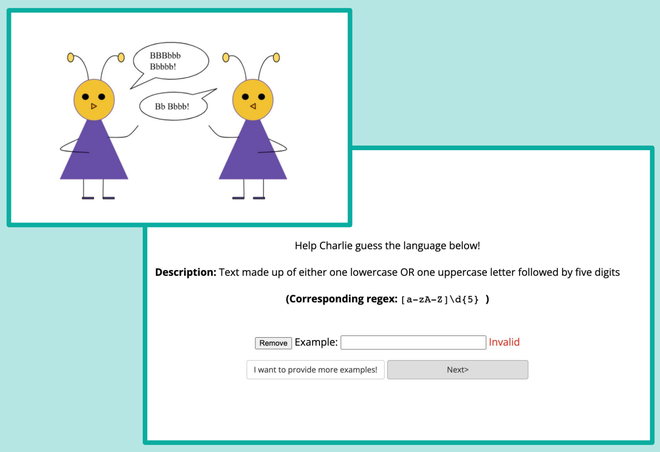
Model

specificity

$$P(regex|input) \sim P(input|regex) \cdot P(regex)$$

simplicity

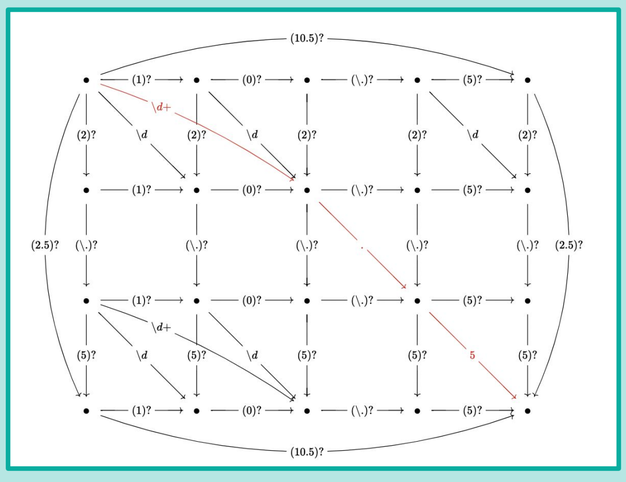
Studying Reality



- Human study to evaluate hypothesis posed as a game
- **412** new data points

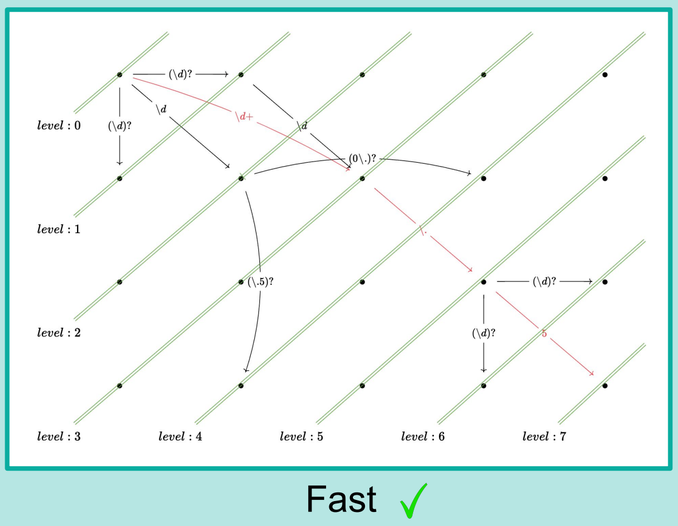
Contribution 2: Search Algorithms

VSA



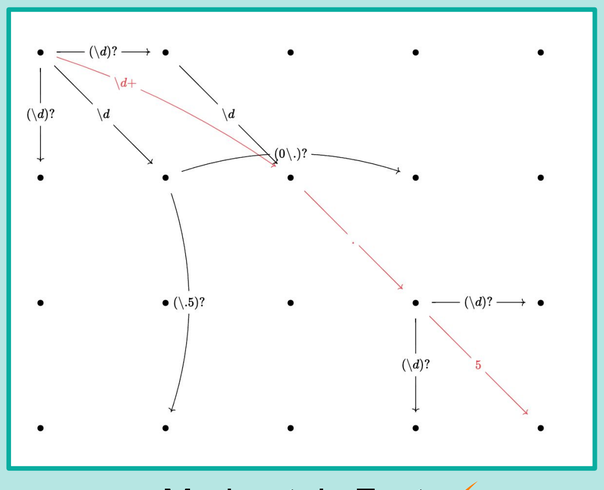
Slow, memory intensive ✗
Completeness Guarantees ✓

Beam



Fast ✓
Completeness Guarantees ✗

A*



Moderately Fast ✓
Completeness Guarantees ✓

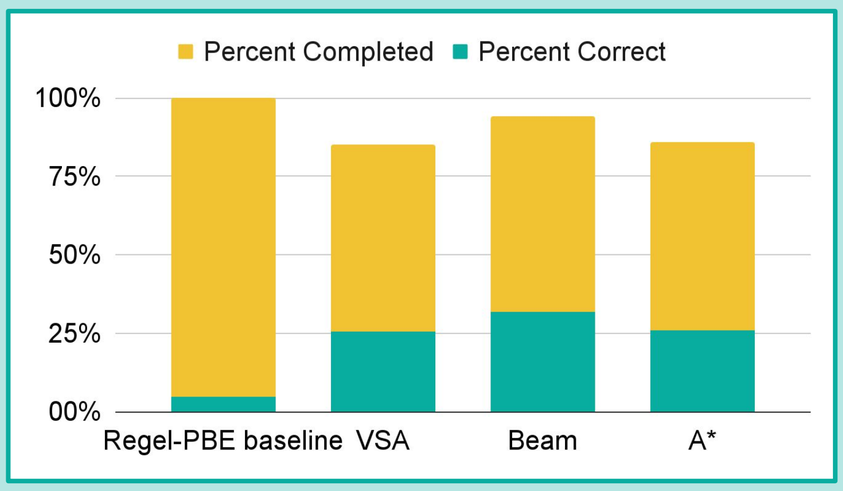
admissible heuristic:

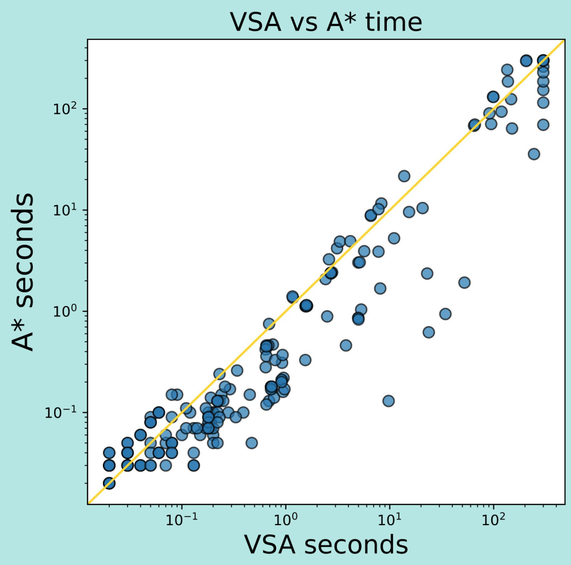
$$\min_{\text{regexes } R}\left[\textsc{Simpl}(R) + \sum_{i=1}^{N}\textsc{Spec}(R, e_i)\right]$$

$$\geq \max_{\text{examples } e_i}\left[\min_{\text{regexes } R}\left[\textsc{Simpl}(R) + \textsc{Spec}(R, e_i)\right]\right]$$

Evaluation

Comparison





Specific Example

111-324-2344
276-943-3044
584-143-2455
000-000-0000
249-394-1232

Correct	\d{3}-\d{3}-\d{4}
Regel-PBE	([0-9]).*
VSA	MEMORY LIMIT
A*	\d{3}-\d{3}-\d{4}
Beam	\d{3}-\d{3}-\d{4}

*Equal contribution. Authors contact: {mbarbone, epertsev}@ucsd.edu