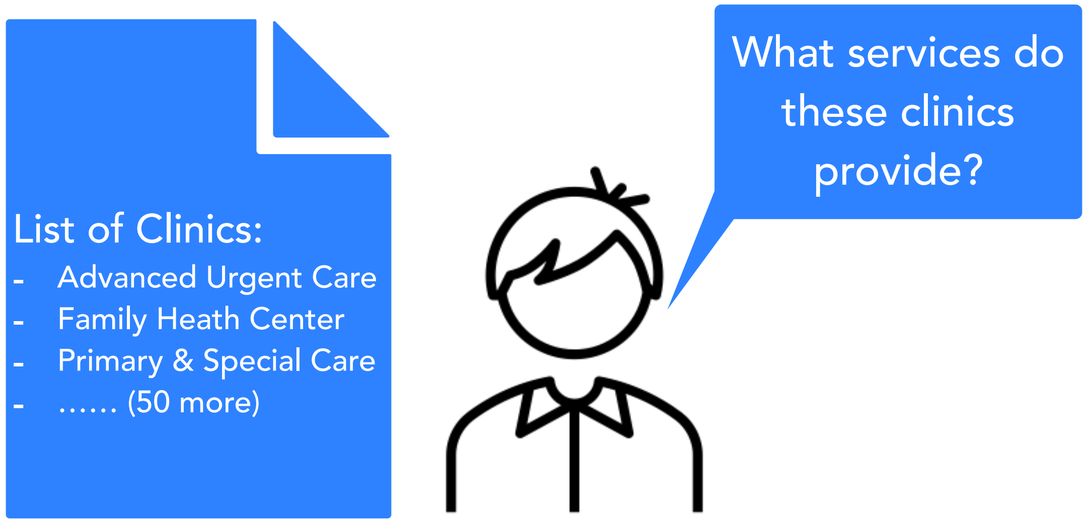
Web Question Answering with Neurosymbolic Program Synthesis

Jocelyn Chen[1], Aaron Lamoreaux[1], Xinyu Wang[2], Greg Durrett[1], Osbert Bastani[3], Isil Dillig[1]
[1]The University of Texas at Austin
[2]University of Michigan
[3]University of Pennsylvania

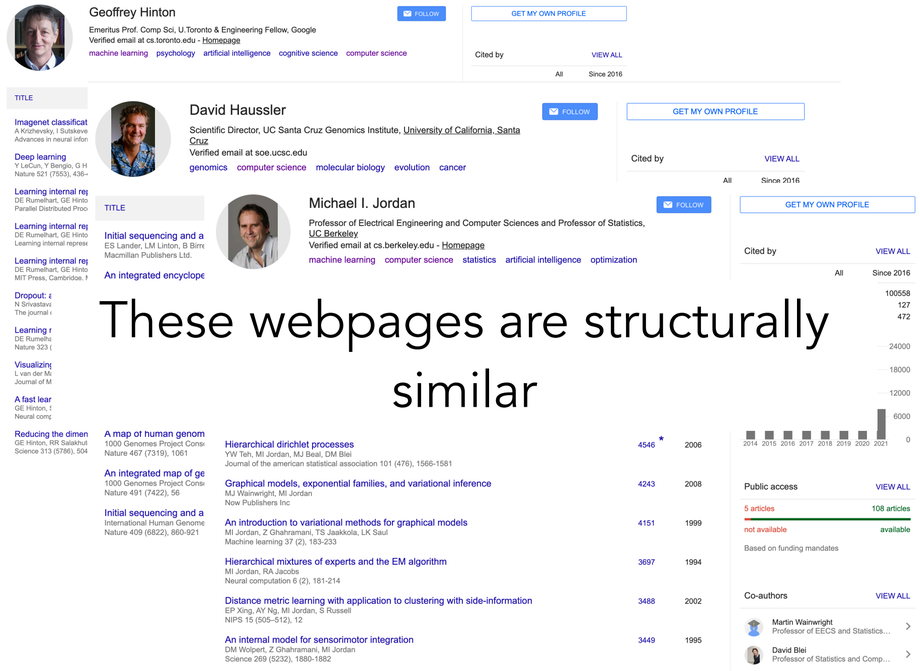
Motivation

This work: automatically extract information from a large set of websites.

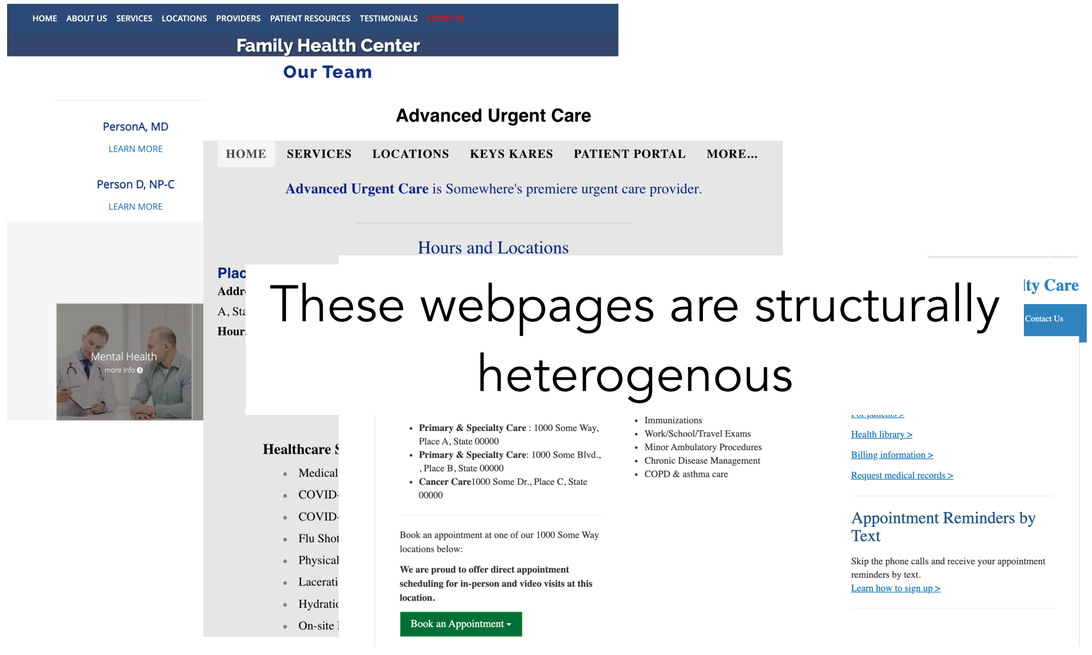


Possible Solution: Wrapper Induction/automatic scraping

Extract information from structurally similar webpages



These webpages are structurally similar



These webpages are structurally heterogenous

Possible Solution: Question Answering

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity

Question

What causes precipitation to fall?

Answer Candidate

Gravity

Answer questions from plain-text documents

Passage Sentence

Advanced Urgent Care. Home Service Locations … 9AM - 3PM SAT. Service offered Healthcare Services Medical Exams Covid Testing…

Question

What services does clinic provides?

Answer Candidate

Healthcare Services

Websites are not plain-text documents

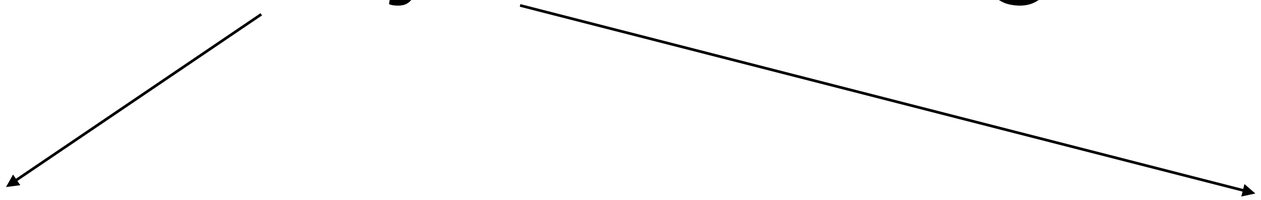
Our Solution: WebQA

Targets at structurally heterogeneous websites with no global schemas

Handles diverse schemas while able to reason about the content

Key idea

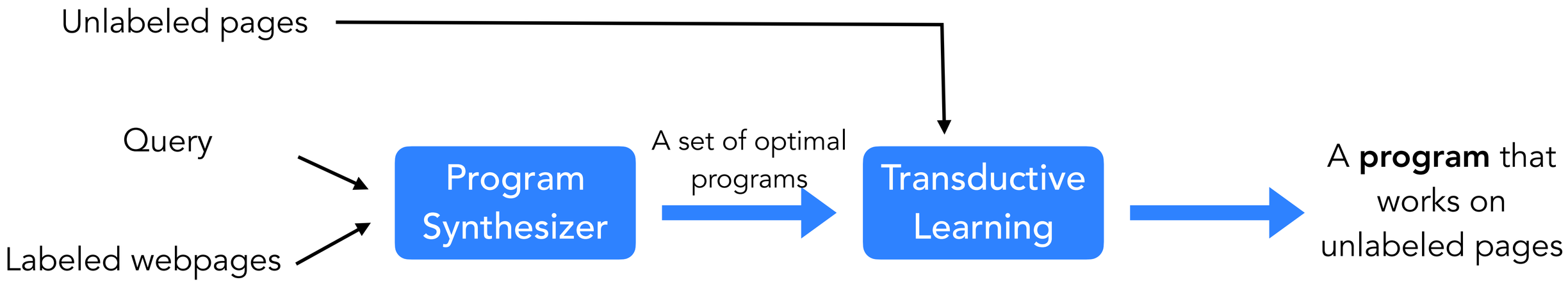
Neurosymbolic Program Synthesis



Better reasoning about the content

Handle the tree structure of the webpage

WebQA Workflow



Webpage

A tree that captures the relationship between text elements on the rendered webpage

Advanced Urgent Care

HOME SERVICES LOCATIONS KEYS KARES PATIENT PORTAL MORE...

Advanced Urgent Care is Somewhere's premiere urgent care provider.

Hours and Locations

Place A
Address: 1000 Some Blvd. Place A, State, 00000
Hours: 8 AM - 6 PM DAILY

Place B
Address: 2000 Some St., Place B, State, 00000
Hours: 9 AM - 5 PM MON - FRI. 9 AM - 3 PM SAT.

Place C
Address: 1000 Some Hwy. Place C, State, 00000
Hours: 9 AM - 5 PM MON - FRI. 9 AM - 3 PM SAT.

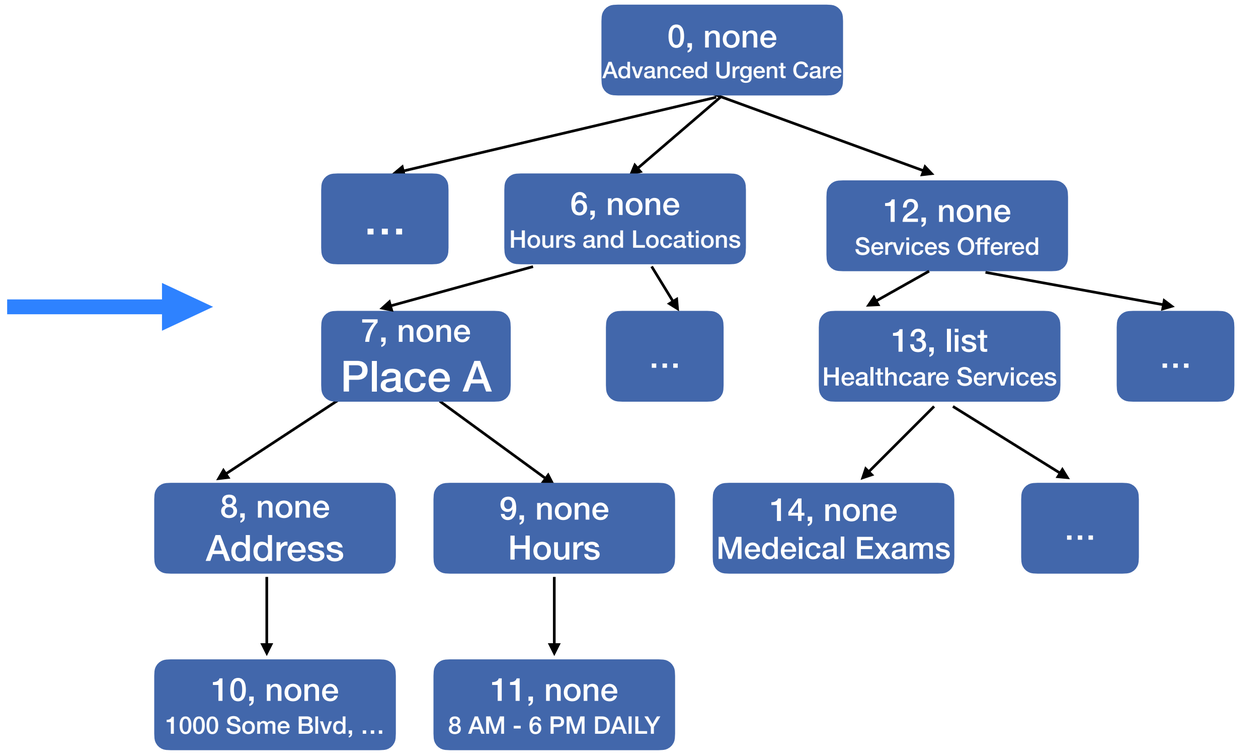
Services Offered

Healthcare Services

- Medical Exams
- COVID-19 Testing
- COVID-19 Ivermectin Therapy
- Flu Shots
- Physicals (work, school, etc.)
- Laceration Repair
- Hydration Therapy
- On-site Mini Pharmacy (Place C only)

Diagnostic Testing

- CT Scan (Place A & Place B only)
- Digital X-ray
- Ultrasound
- Lab Testing
- Lung Cancer Screening (Place B only)
- Cardiac Calcium Scoring (Place B only)



WebQA Program Workflow

Input: Q (question), K (keyword), W (webpage)

Top-level Program: $\lambda Q, K, W \,.\, \{\psi_1 \rightarrow \lambda x \,.\, e_1, \ldots, \psi_n \rightarrow \lambda x \,.\, e_n\}$

Guard

Identify schema and locate relevant section

Extractor

Extract information in the located section under this schema

Constructs in Guards and Extractors

GetChildren(v, p)

GetDescendents(v, p)

Substring(t, p)

Split(t, c)

containsKeyword(t, K)

hasAnswer(t, Q)

hasEntity(t, PERSON)

Tree Navigation

Given a node v, get its children/descendent that satisfy a predicate p

String Processing

Given a text t, get its

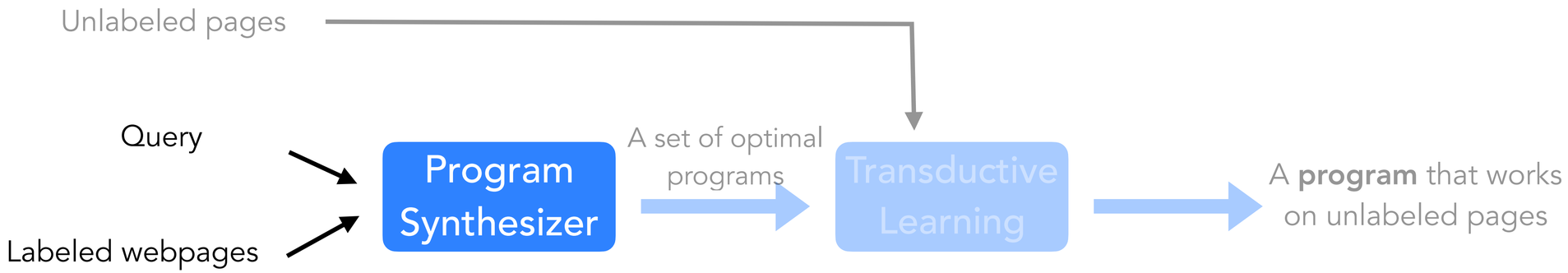
- substrings that satisfy some predicate p
- split substrings using delimiter c

Neural Components

Given a text t, check if it

- contains answers to questions Q
- contains words similar to keywords K
- contains entity such as person

WebQA Workflow



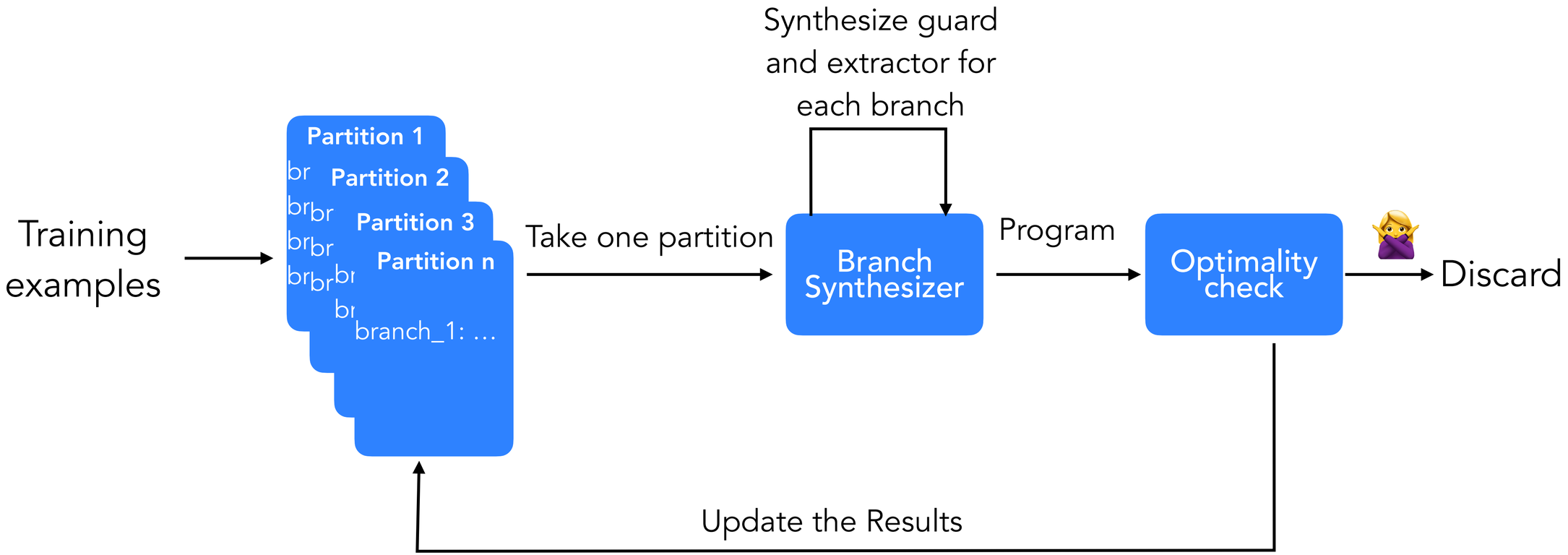
Generate **a set of programs** that **extract correct information** from labeled webpage

Difficult to get exact match

Generate **a set of programs** achieve the **highest F1 score** on the labeled webpage

$$F_1 = \frac{2 \cdot \textbf{Precision} \cdot \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

Optimal Synthesis Workflow



Yields programs with highest F1 score on labeled training examples

Highlight of our Synthesis Technique

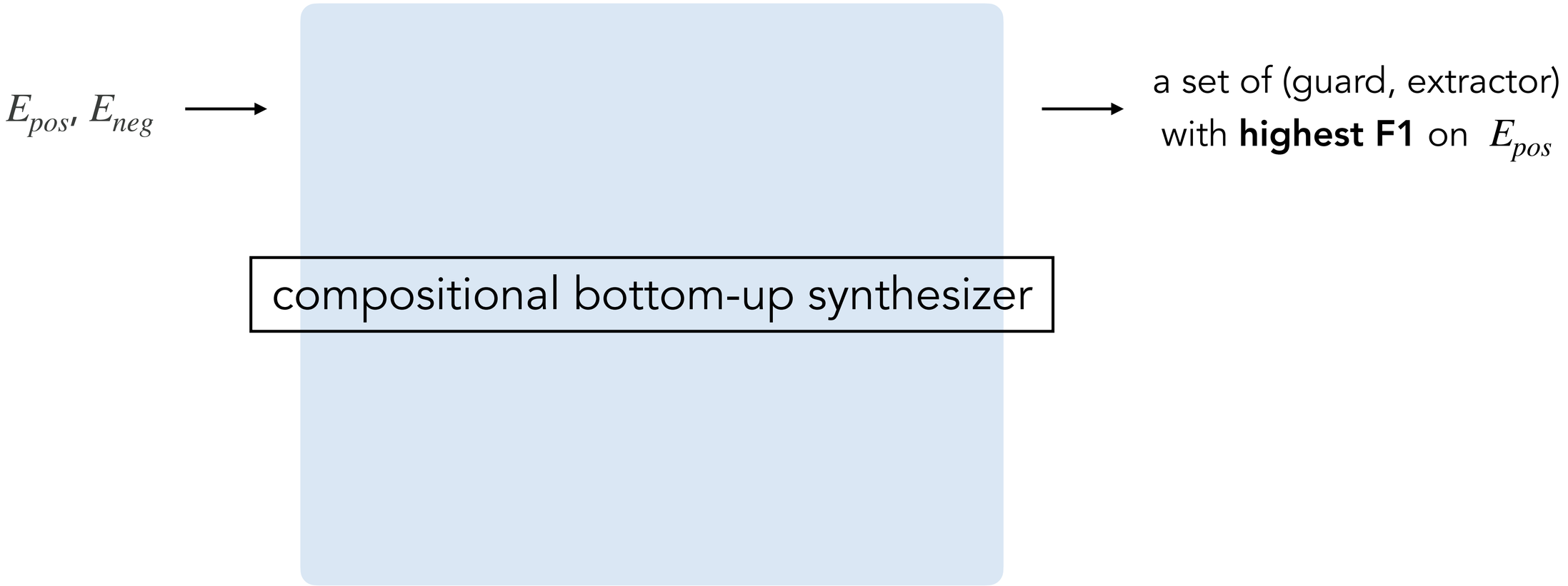
Decomposition

separate extractor synthesis
and guard synthesis

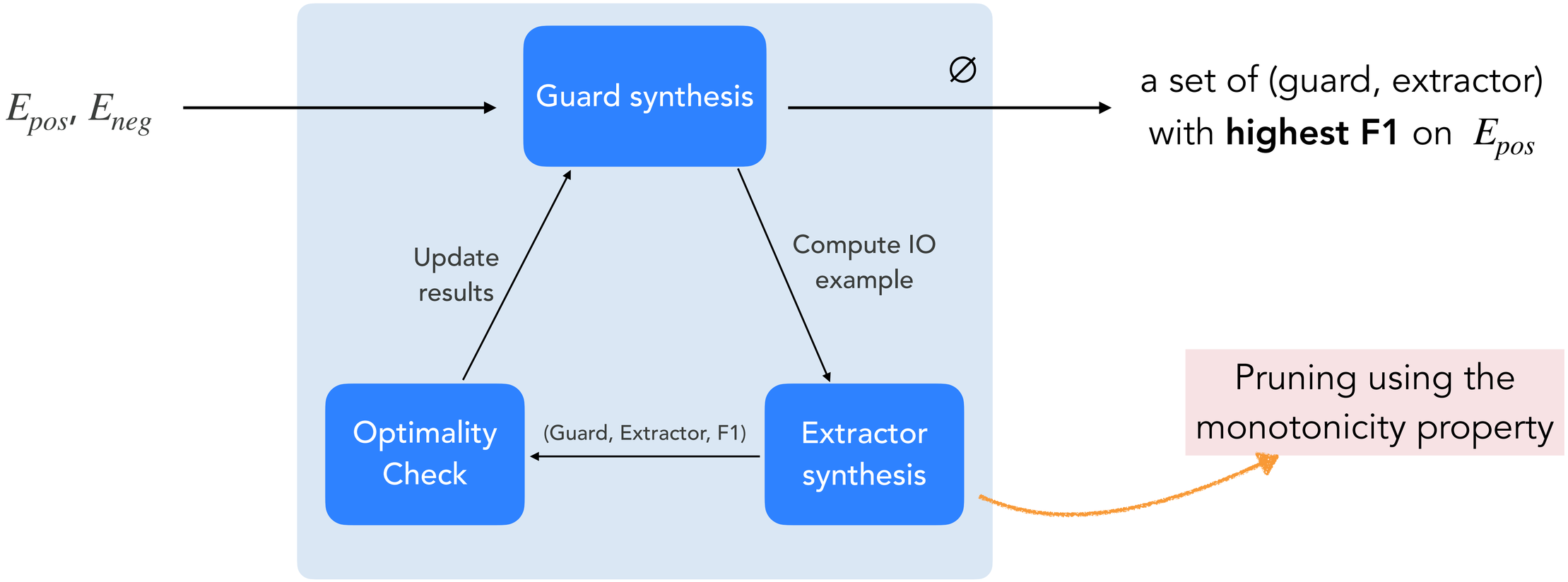
Pruning

using a monotonicity property
of the DSL with respect to recall

Interface of Branch Synthesis



Branch Synthesis Workflow



Decomposition reduces the number of (guard, extractor) pairs!

Pruning

Monotonicity property of the DSL:

Let e, e' be two extractors, if e' is a sub-expression of e, then Recall(e) $\leq$ Recall(e')

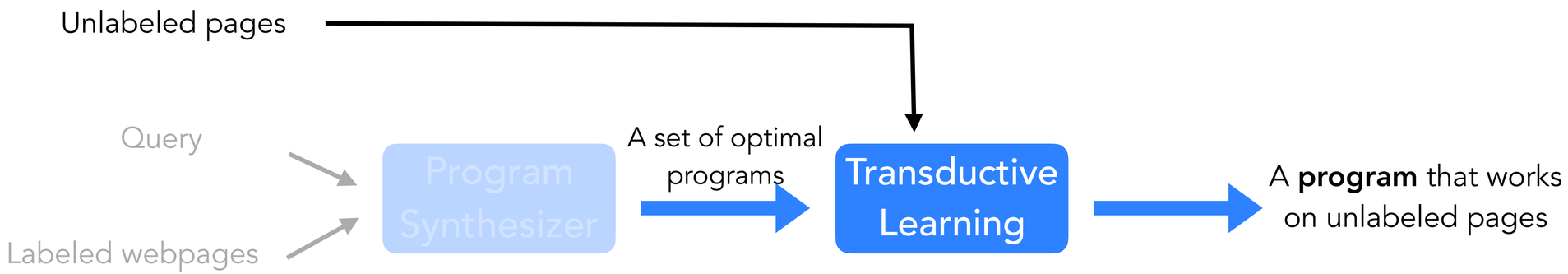
F_1 Upper bound for any program uses e as sub-expression on example E:

$$\text{UB}(e, E) = \frac{2 \cdot 1 \cdot \text{Recall}(e, E)}{1 + \text{Recall}(e, E)}$$

precision = 1
(best case)

We prune out any expression e if $\text{UB}(e, E) <$ current_best_F1

WebQA Workflow



- Working with websites of *diverse* schemas: **not all** optimal programs on the label webpages **work well** on unlabeled webpages
- Performance of randomly chosen program **varies a lot**

Given a **large** set of **optimal programs**, use the **unlabeled data** to select **a program** that **works well to the unlabeled data** with **low variance**

Program selection via transductive learning

Transductive learning: use unlabeled data to help obtain predictions

Key idea: Ensemble of optimal programs

- Aggregate predication over samples of the optimal programs
- Generalize better than a single program with lower variance

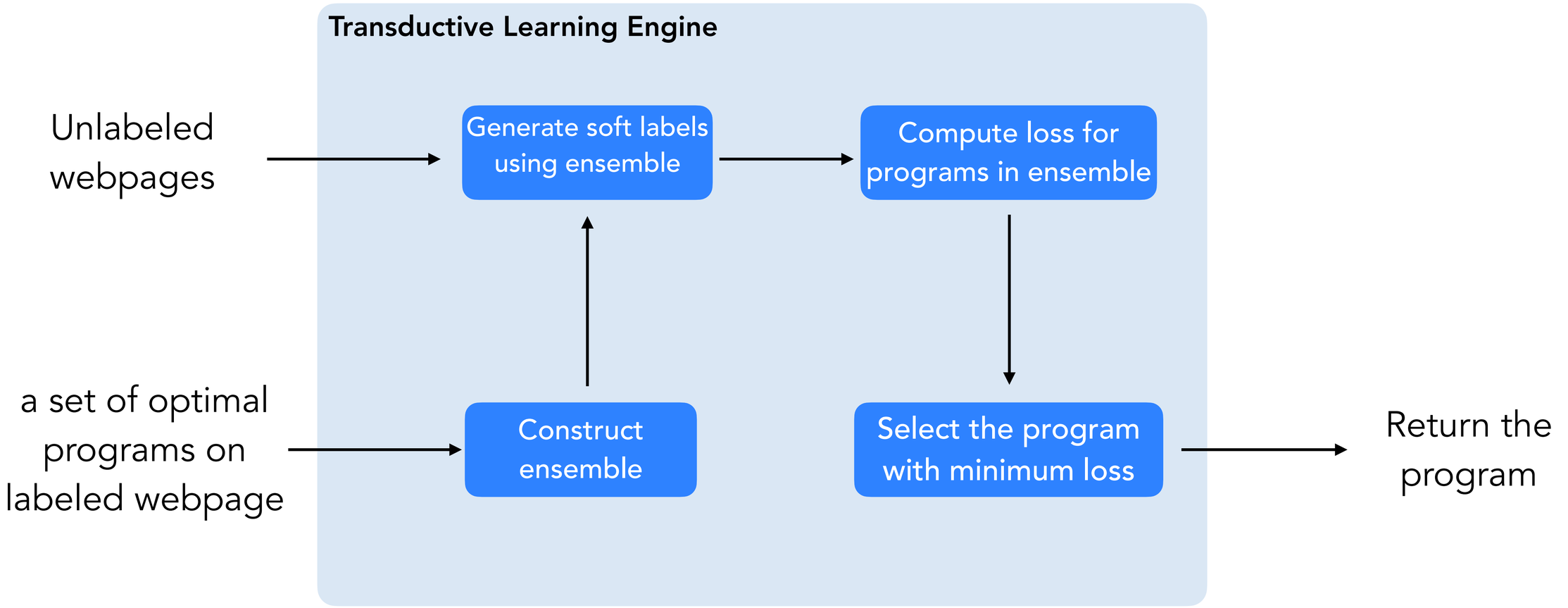
Return the ensemble

- Ensemble is less interpretable
- Ensemble is expensive to compute

→

Select **one program** whose outputs are **most similar** to the ensemble

Program selection via transductive learning

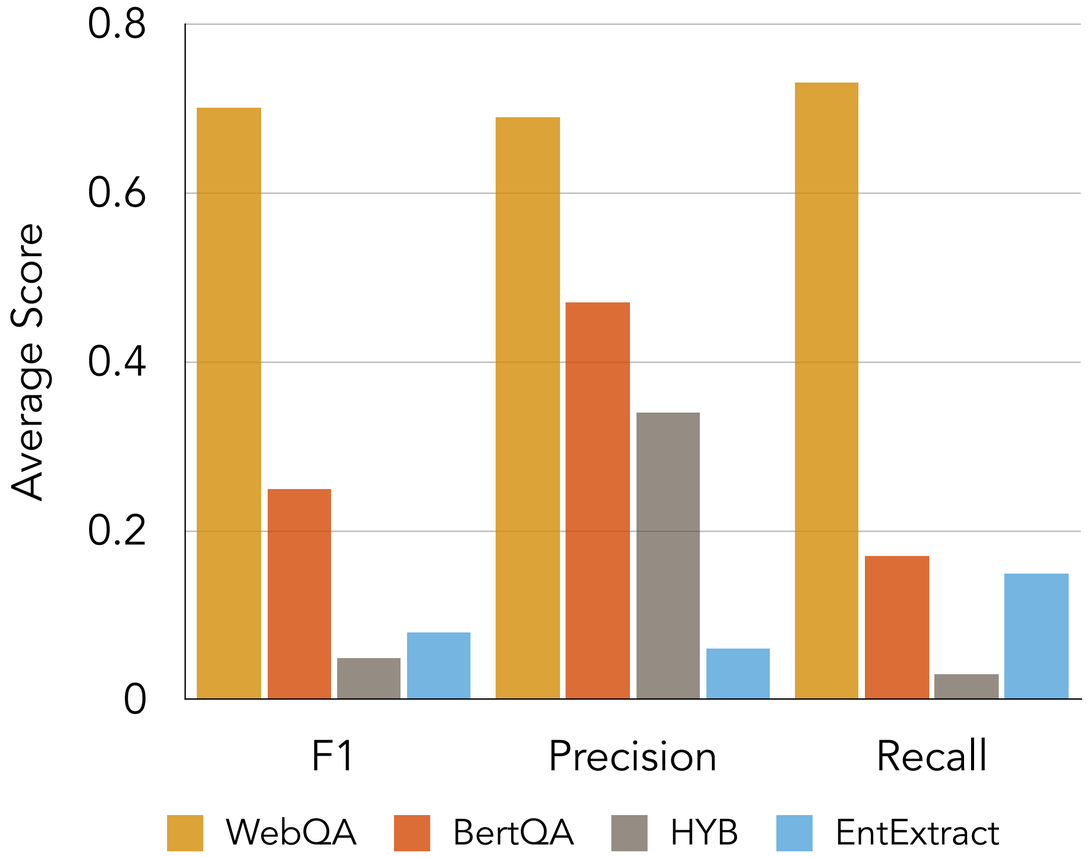


Evaluation: dataset

- We collect webpages:
  - 4 domains (Faculty profile, Conference, Class, Clinic)
  - 40 webpages for each domain
  - ~6 tasks per domain
  - 5 for training and remaining for testing

Domain	Sample tasks
Faculty	- Extract current PhD students - Extract papers published in 2012
Conference	- Extract program committee members - Extract paper submission deadlines
Class	- Extract the time of the lectures - Extract name of the TAs
Clinic	- Extract the list of doctors or providers - Extract the provided services

Evaluation: Benefit of Neurosymbolic Approach



- Baselines:
  - BERTQA: **Question answering system** takes a webpage and a question as input and outputs the answers [Zero shot]
  - HYB: **PBE system** that takes a set of webpages as example inputs and synthesize **a Xpath program**
  - EntExtract: **Entity extraction tool** for webpages using a natural language query as input [Zero shot]

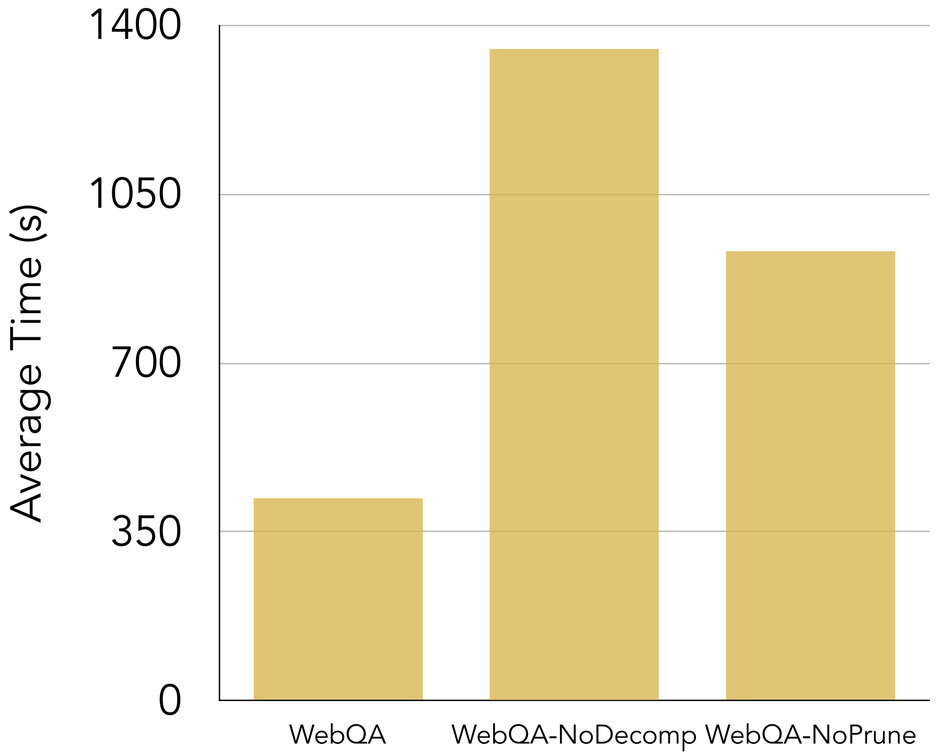
WebQA outperforms prior works for our task

Evaluation: ablation study

Three key ideas:

- **Decomposition** between guard and extractor synthesis
- **Pruning using the monotonicity property** of the DSL
- Reduce variance of the output program using **transductive learning**

Evaluation: Effectiveness of Decomposition and Pruning



- Baselines :
  - WebQA-NoDecomp: Synthesizes guards and extractors jointly
  - WebQA-NoPrune: Does not compute upper bound of partial programs for pruning

WebQA achieves 3.6x speedup compare to WebQA-NoPrune and 2.4x speedup compare to WebQA-NoDecomp.

Evaluation: Effectiveness of the Transductive Learning

Technique	% Improvement in F1	Reduction in Variance
Random	6.0%	1550X
Shortest	6.3%	1570X

- Two Baselines :
  - Random: randomly chooses one of the optimal programs
  - Shortest: chooses randomly one program with the smallest AST size among the optimal programs
- We measured the % of improvement and reduction in variance that WebQA achieved over the baseline

WebQA achieves more stable performance in obtaining high quality synthesized programs

Conclusion

- Extract information from websites with globally different schemas using **neurosymbolic program synthesis**
- Performs better than techniques for wrapper induction and question answering for such structurally heterogeneous websites

Try WebQA at https://github.com/utopia-group/WebQA