

# Web Question Answering with Neurosymbolic Program Synthesis

Jocelyn Chen<sup>1</sup>, Aaron Lamoreaux<sup>1</sup>, Xinyu Wang<sup>2</sup>, Greg Durrett<sup>1</sup>, Osbert Bastani<sup>3</sup>, Isil Dillig<sup>1</sup>

<sup>1</sup>The University of Texas at Austin

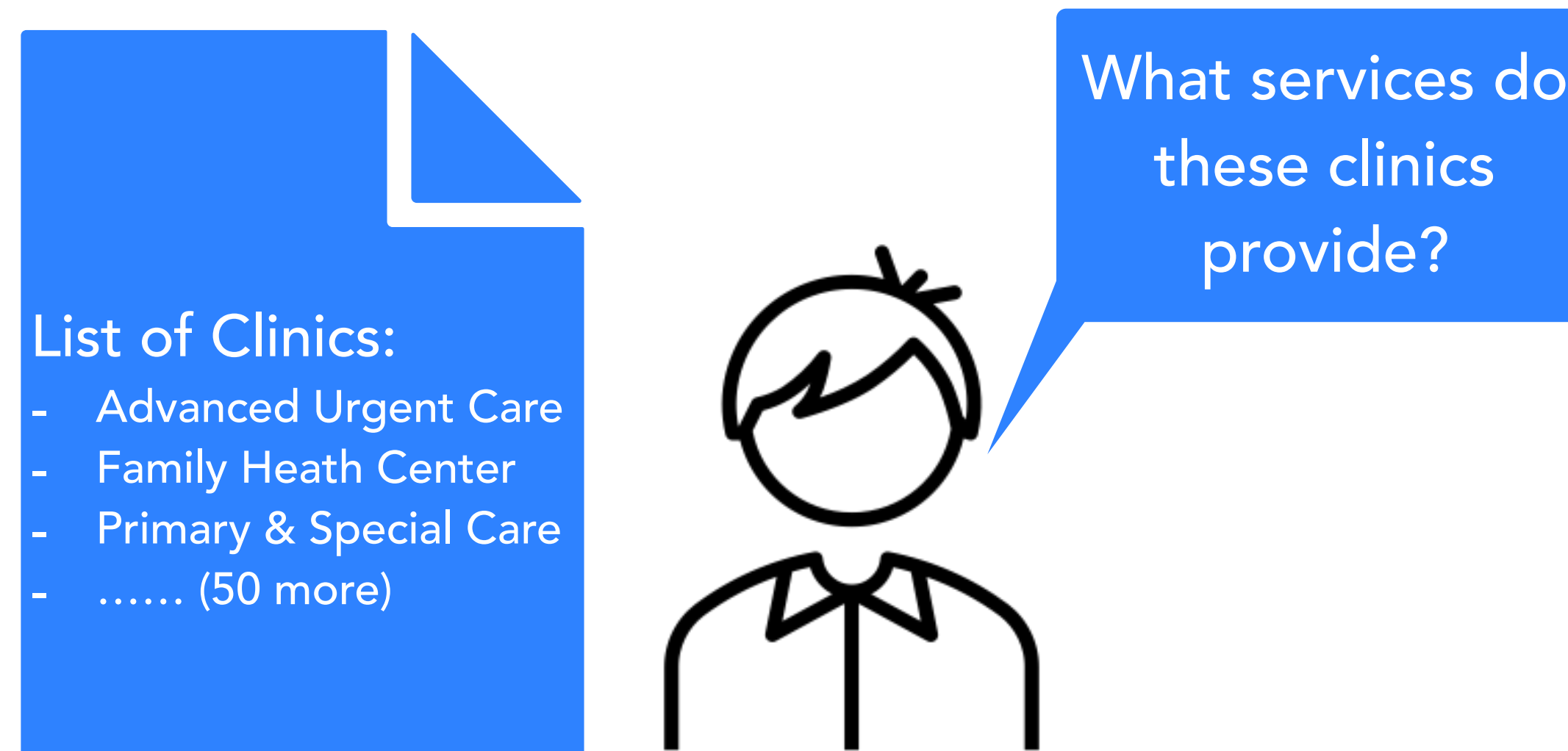
<sup>2</sup>University of Michigan

<sup>3</sup>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



**Geoffrey Hinton**  
Emeritus Prof. Comp Sci, U.Toronto & Engineering Fellow, Google  
Verified email at cs.toronto.edu · [Homepage](#)  
[machine learning](#) [psychology](#) [artificial intelligence](#) [cognitive science](#) [computer science](#)

[FOLLOW](#) [GET MY OWN PROFILE](#)

Cited by [VIEW ALL](#)

All Since 2016

**David Haussler**  
Scientific Director, UC Santa Cruz Genomics Institute, [University of California, Santa Cruz](#)  
Verified email at soe.ucsc.edu  
[genomics](#) [computer science](#) [molecular biology](#) [evolution](#) [cancer](#)

[FOLLOW](#) [GET MY OWN PROFILE](#)

Cited by [VIEW ALL](#)

All Since 2016

**Michael I. Jordan**  
Professor of Electrical Engineering and Computer Sciences and Professor of Statistics, [UC Berkeley](#)  
Verified email at cs.berkeley.edu · [Homepage](#)  
[machine learning](#) [computer science](#) [statistics](#) [artificial intelligence](#) [optimization](#)

[FOLLOW](#) [GET MY OWN PROFILE](#)

Cited by [VIEW ALL](#)

All Since 2016

**These webpages are structurally similar**

**Imaginet classification**  
A Krizhevsky, I Sutskever  
Advances in neural information processing systems 25 (2012), 1099-1107

**Deep learning**  
Y Lecun, Y Bengio, G Hinton  
Nature 521 (7553), 436-444 (2015)

**Learning internal representations by using distributed representations of sentences**  
DE Rumelhart, GE Hinton  
Parallel Distributed Processing 1 (1986), 318-361

**Learning internal representations by using distributed representations of sentences**  
DE Rumelhart, GE Hinton  
Learning internal representations by using distributed representations of sentences 1 (1986), 318-361

**Learning internal representations by using distributed representations of sentences**  
DE Rumelhart, GE Hinton  
MIT Press, Cambridge, MA (1986)

**Dropout: a simple way to prevent neural networks from overfitting**  
N Srivastava, G Hinton, A Krizhevsky, I Sutskever  
Journal of machine learning research 15 (2014), 1923-1930

**Learning internal representations by using distributed representations of sentences**  
DE Rumelhart, GE Hinton  
Nature 323 (6086), 681-686 (1995)

**Visualizing the structure of learned representations**  
L van der Maaten, G Hinton  
Journal of machine learning research 9 (2008), 1224-1232

**A fast learning algorithm for training a neural network**  
GE Hinton, T J Sejnowski  
Neural computation 1 (1989), 1-17

**Reducing the dimensionality of data by using a generative stochastic autoencoder**  
GE Hinton, RR Salakhutdinov  
Science 313 (5786), 504-507 (2005)

**A map of human genome organization from nucleotide to chromosome scale**  
1000 Genomes Project Consortium  
Nature 467 (7319), 1061-1073 (2010)

**An integrated map of genetic variation from 1000 Genomes Project Consortium**  
Nature 491 (7422), 56-65 (2012)

**Initial sequencing and analysis of the human genome**  
Nature 409 (6822), 860-921 (2001)

**Hierarchical dirichlet processes**  
YW Teh, MJ Jordan, MJ Beal, DM Blei  
Journal of the american statistical association 101 (478), 1566-1581 (2006)

**Graphical models, exponential families, and variational inference**  
MJ Wainwright, MJ Jordan  
Now Publishers Inc (2008)

**An introduction to variational methods for graphical models**  
MJ Jordan, Z Ghahramani, TS Jaakkola, LK Saul  
Machine learning 37 (2), 183-233 (1999)

**Hierarchical mixtures of experts and the EM algorithm**  
MJ Jordan, RA Jacobs  
Neural computation 6 (2), 181-214 (1994)

**Distance metric learning with application to clustering with side-information**  
EP Xing, AY Ng, MJ Jordan, S Russell  
NIPS 15 (605-612), 12 (2002)

**An internal model for sensorimotor integration**  
DM Wolpert, Z Ghahramani, MJ Jordan  
Science 269 (5232), 1880-1882 (1995)

HOME ABOUT US SERVICES LOCATIONS PROVIDERS PATIENT RESOURCES TESTIMONIALS COVID-19

**Family Health Center**  
**Our Team**

Person A, MD  
[LEARN MORE](#)

Person D, NP-C  
[LEARN MORE](#)

**Advanced Urgent Care**

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

**Hours and Locations**

**Plac**  
**Addr**  
**A, St**  
**Hour**

**Healthcare Services**

- Medical
- COVID
- COVID
- Flu Shot
- Physical
- Lacerati
- Hydrati
- On-site

- Primary & Specialty Care** : 1000 Some Way, Place A, State 00000
- Primary & Specialty Care**: 1000 Some Blvd., Place B, State 00000
- Cancer Care** 1000 Some Dr., Place C, State 00000

Book an appointment at one of our 1000 Some Way locations below:

**We are proud to offer direct appointment scheduling for in-person and video visits at this location.**

**Book an Appointment**

- Immunizations
- Work/School/Travel Exams
- Minor Ambulatory Procedures
- Chronic Disease Management
- COPD & asthma care

[Request medical records >](#)

**Appointment Reminders by Text**  
Skip the phone calls and receive your appointment reminders by text.  
[Learn how to sign up >](#)

# 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

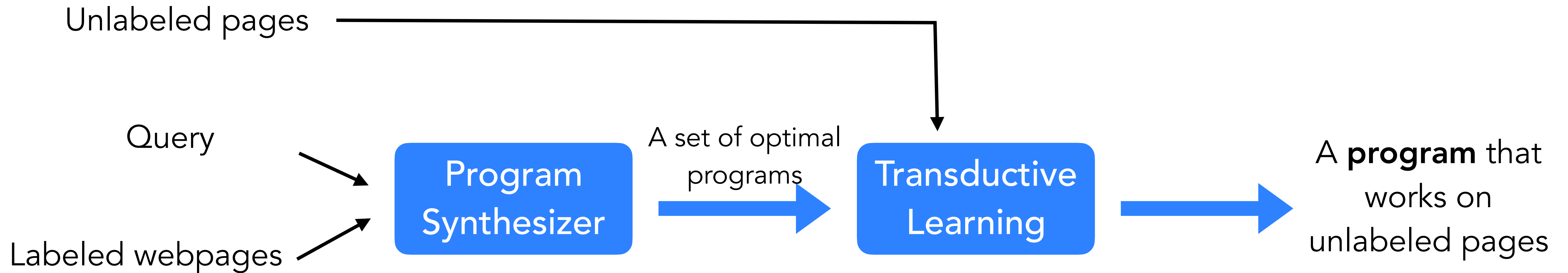
Neurosymbolic Program Synthesis



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.

---

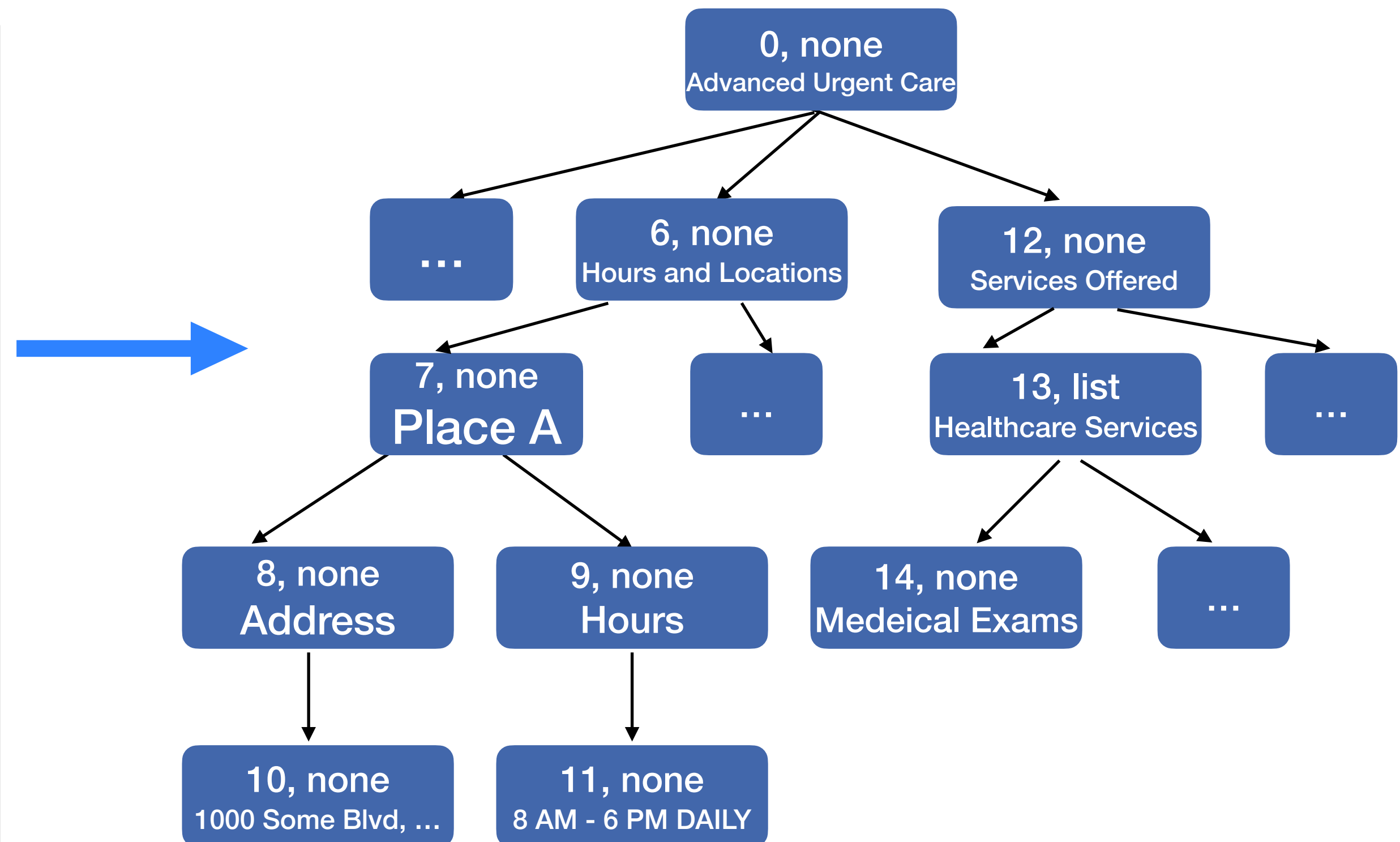
**Hours and Locations**

Place A	Place B	Place C
<b>Address:</b> 1000 Some Blvd. Place A, State, 00000	<b>Address:</b> 2000 Some St., Place B, State, 00000	<b>Address:</b> 1000 Some Hwy. Place C, State, 00000
<b>Hours:</b> 8 AM - 6 PM DAILY	<b>Hours:</b> 9 AM - 5 PM MON - FRI. 9 AM - 3 PM SAT.	<b>Hours:</b> 9 AM - 5 PM MON - FRI. 9 AM - 3 PM SAT.

---

**Services Offered**

Healthcare Services	Diagnostic Testing
<ul style="list-style-type: none"><li>Medical Exams</li><li>COVID-19 Testing</li><li>COVID-19 Ivermectin Therapy</li><li>Flu Shots</li><li>Physicals (work, school, etc.)</li><li>Laceration Repair</li><li>Hydration Therapy</li><li>On-site Mini Pharmacy (Place C only)</li></ul>	<ul style="list-style-type: none"><li>CT Scan (Place A &amp; Place B only)</li><li>Digital X-ray</li><li>Ultrasound</li><li>Lab Testing</li><li>Lung Cancer Screening (Place B only)</li><li>Cardiac Calcium Scoring (Place B only)</li></ul>





# WebQA Program Workflow

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

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

**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)

## Tree Navigation

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

Substring(t, p)  
Split(t, c)

## String Processing

Given a text t, get its

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

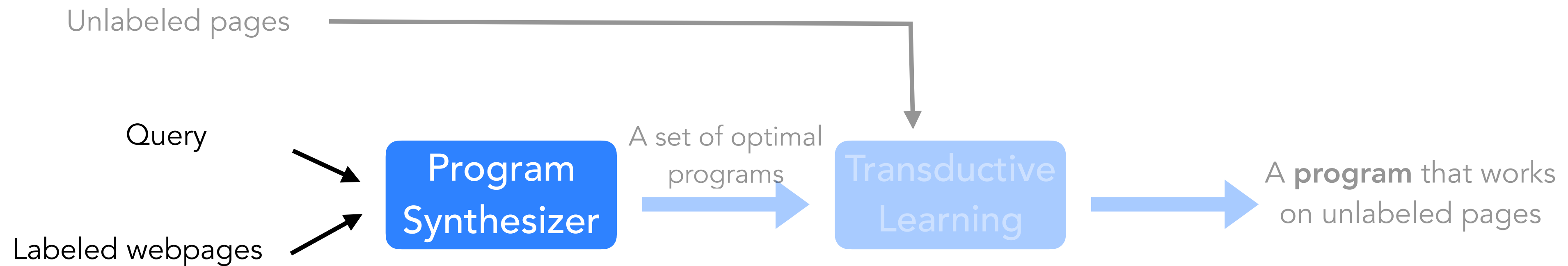
containsKeyword(t, K)  
hasAnswer(t, Q)  
hasEntity(t, PERSON)

## 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



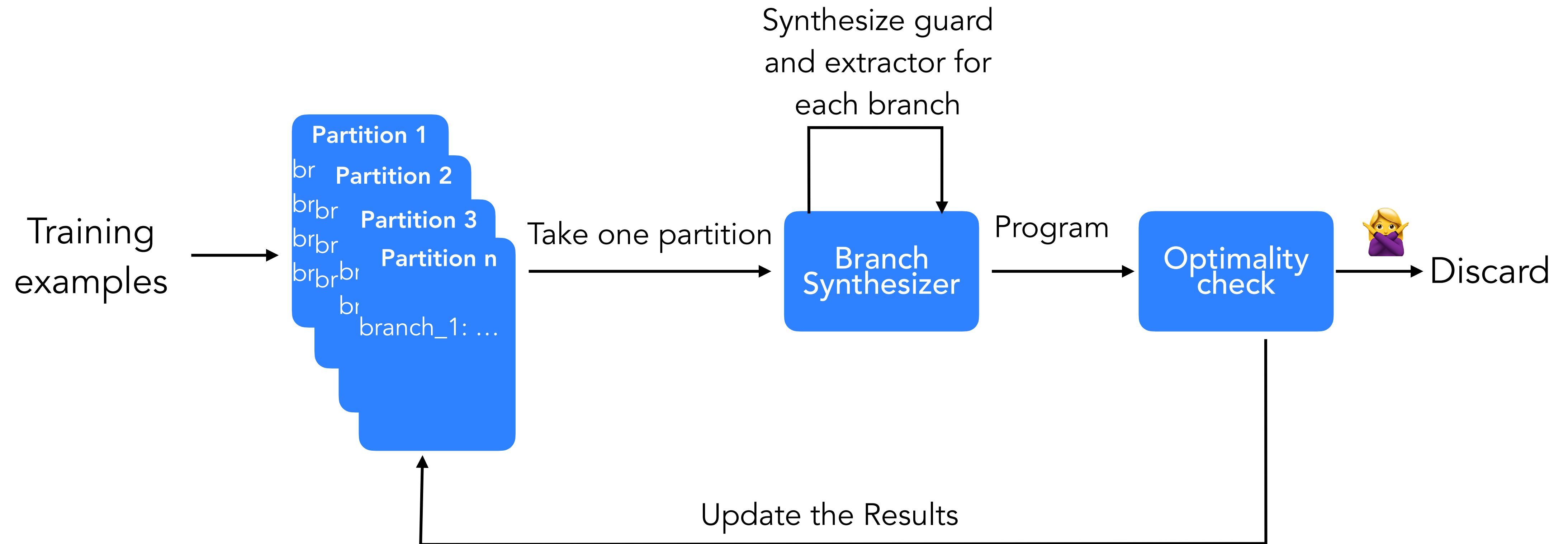
Generate a **set of programs** that **extract correct information** from labeled webpage

Difficult to get exact match

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

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

# Optimal Synthesis Workflow



Yields programs with highest F1 score on labeled training examples

# Highlight of our Synthesis Technique

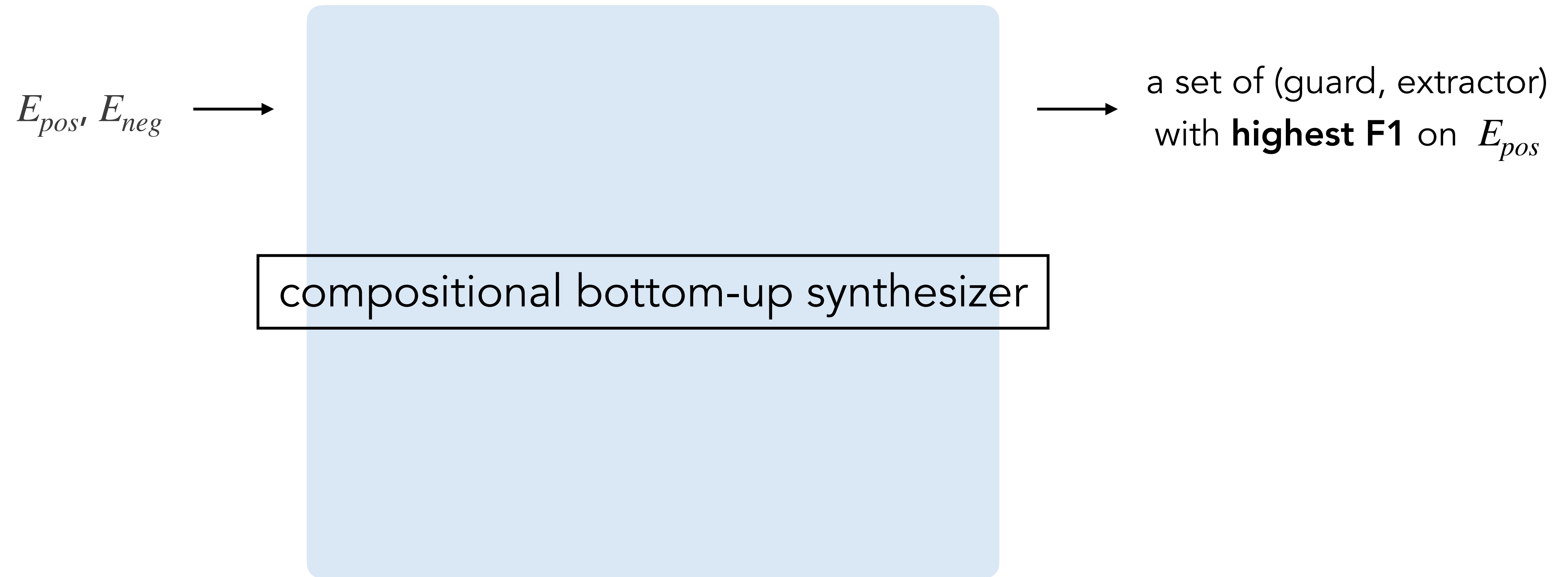
## Decomposition

separate extractor synthesis  
and guard synthesis

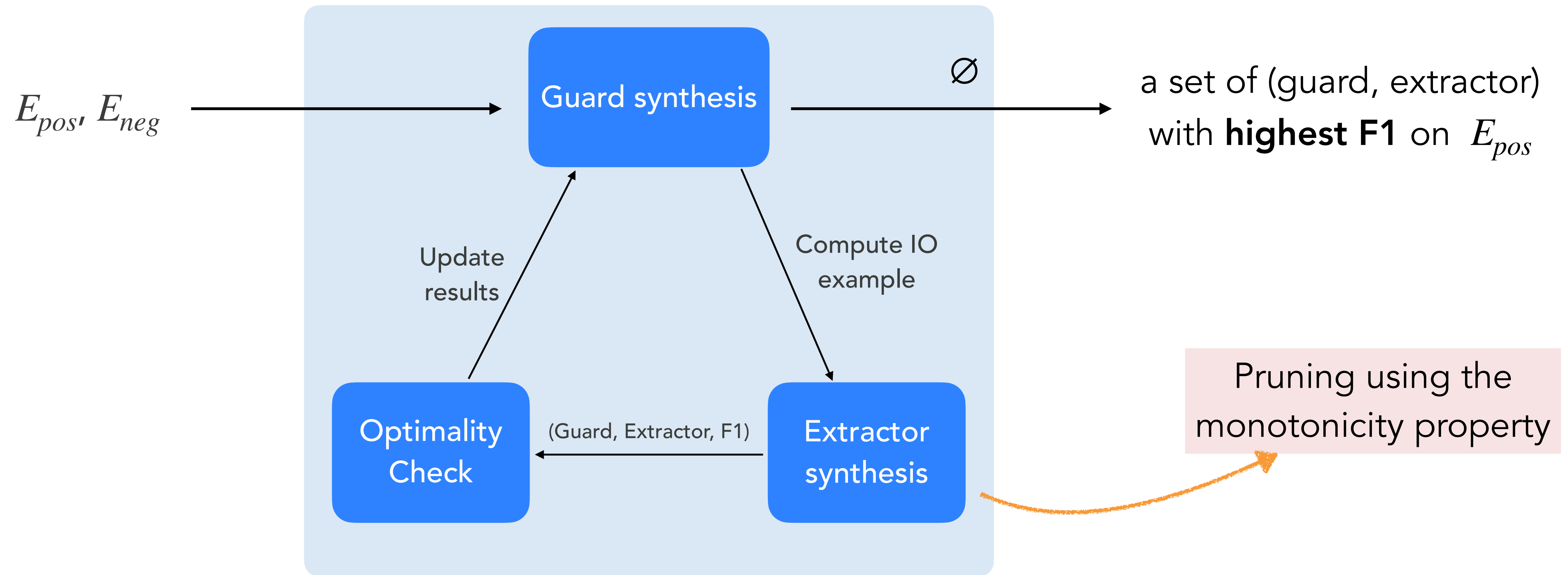
## Pruning

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  $\text{Recall}(e) \leq \text{Recall}(e')$

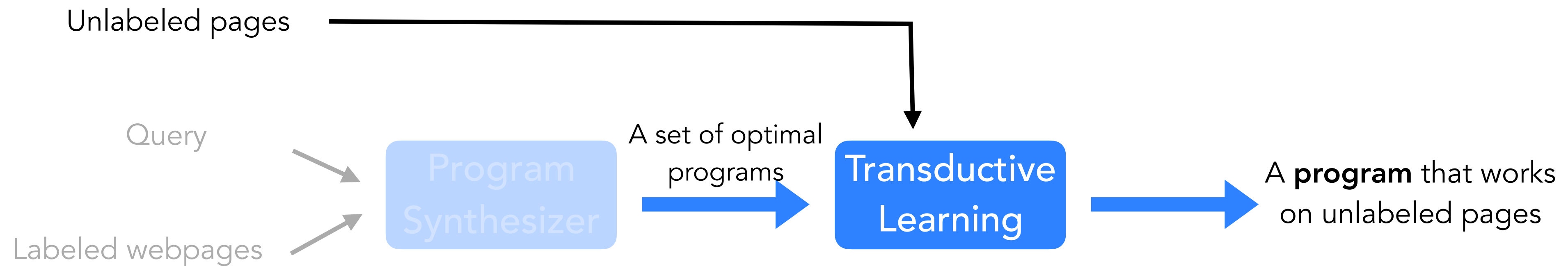
$F_1$  Upper bound for any program uses  $e$  as sub-expression on example  $E$ :

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

precision = 1  
(best case)

We prune out any expression  $e$  if  $\text{UB}(e, E) < \text{current\_best\_F1}$

# 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

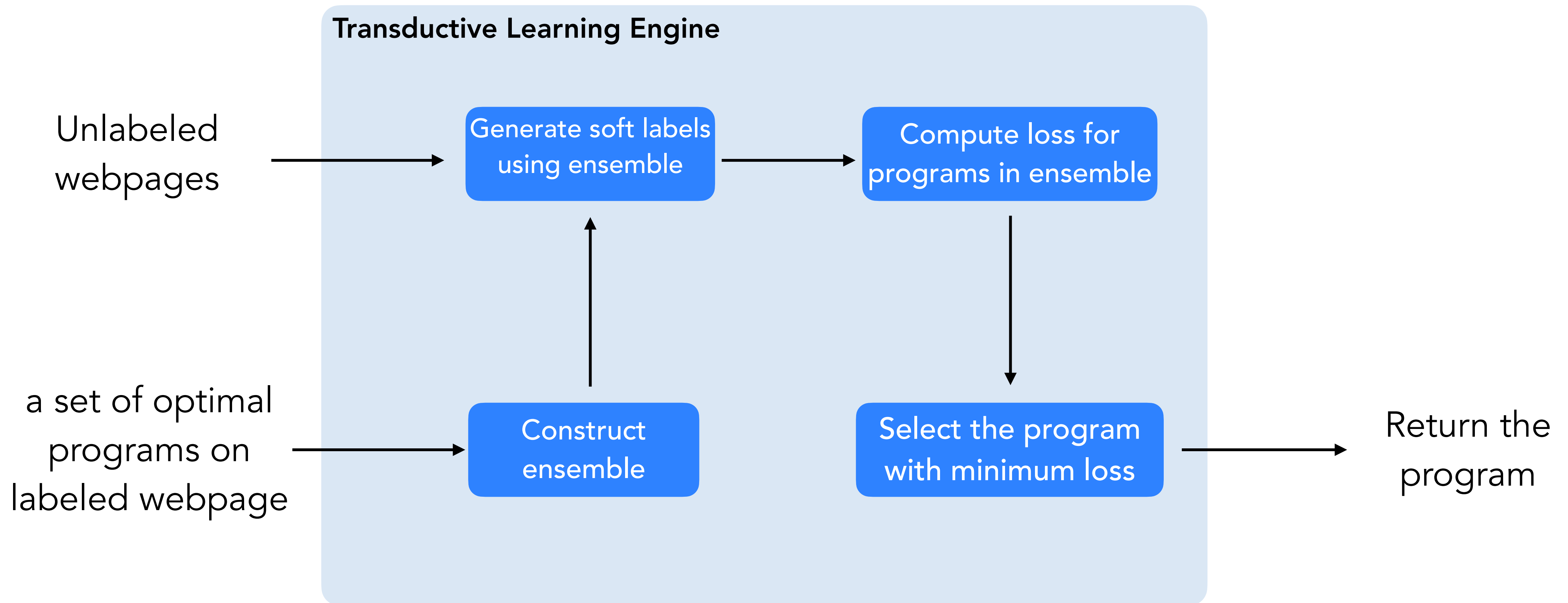
- Aggregate predication over samples of the optimal programs
- Generalize better than a single program with lower variance

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

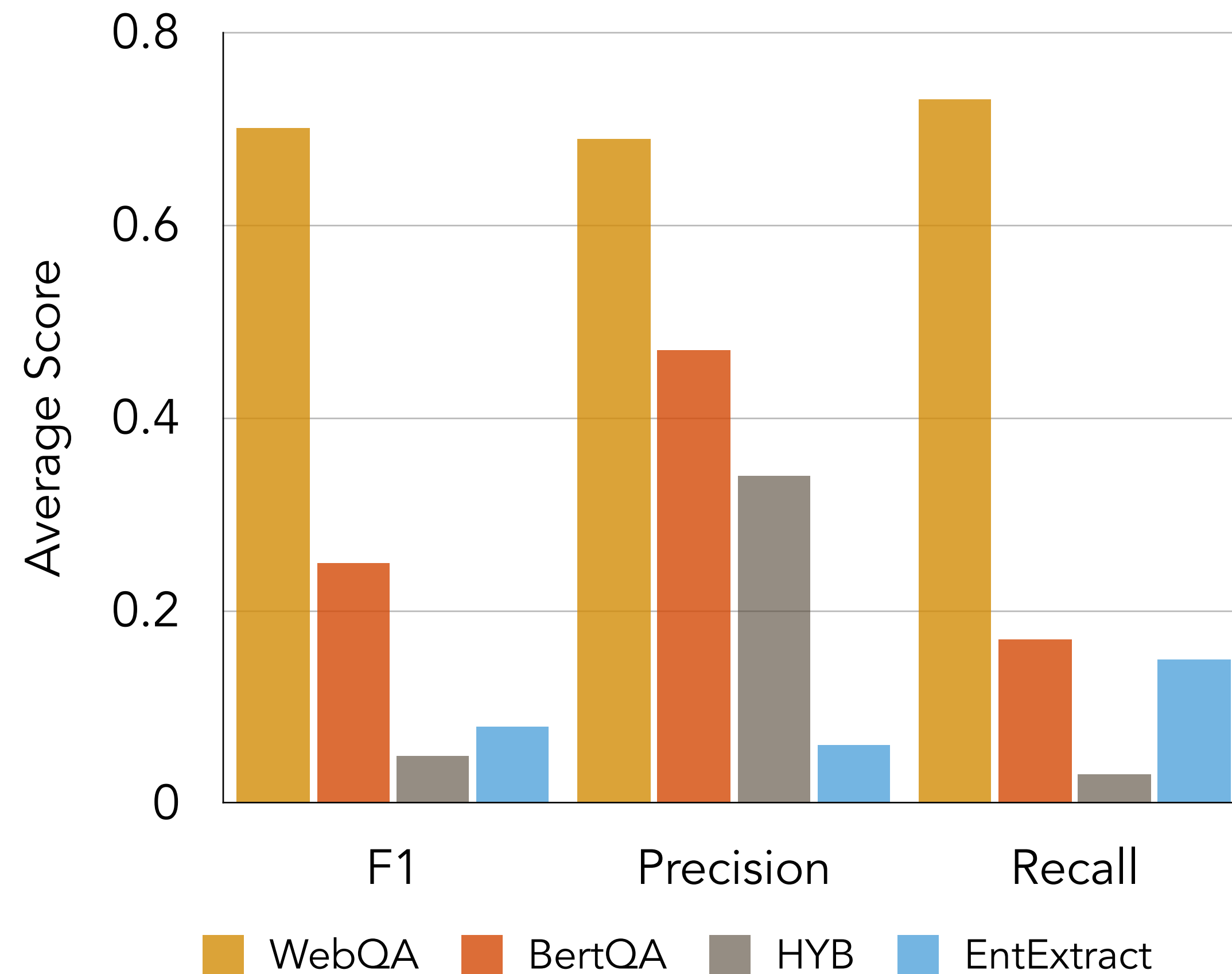


# Evaluation: dataset

- 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	<ul style="list-style-type: none"><li>- Extract current PhD students</li><li>- Extract papers published in 2012</li></ul>
Conference	<ul style="list-style-type: none"><li>- Extract program committee members</li><li>- Extract paper submission deadlines</li></ul>
Class	<ul style="list-style-type: none"><li>- Extract the time of the lectures</li><li>- Extract name of the TAs</li></ul>
Clinic	<ul style="list-style-type: none"><li>- Extract the list of doctors or providers</li><li>- Extract the provided services</li></ul>

# 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

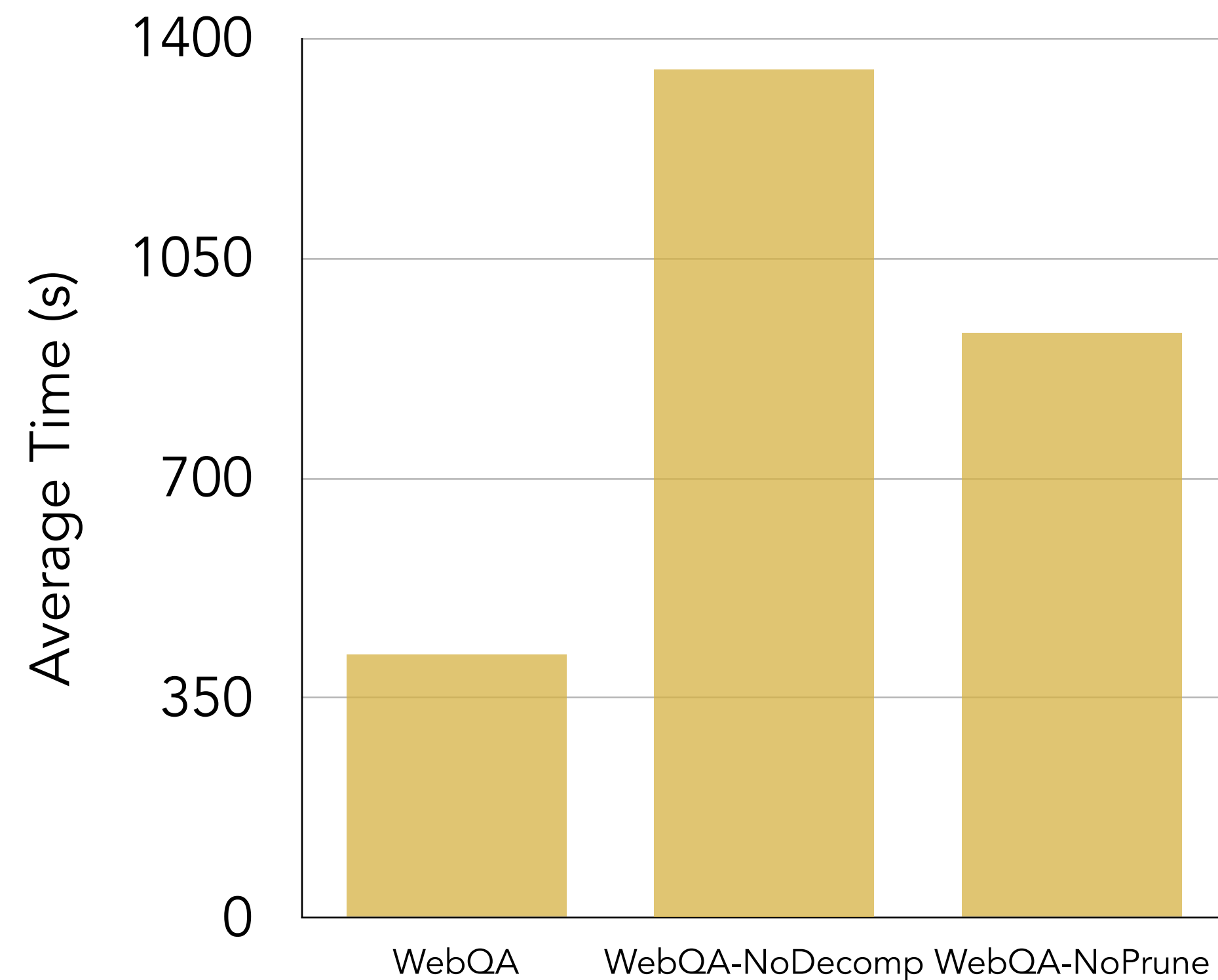
# Evaluation: ablation study

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>