

# Scraping Distributed, Hierarchical Web Data

with “Programming  
by Demonstration”!

Sarah E. Chasins<sup>1</sup>

Maria Mueller<sup>2</sup>

Rastislav Bodik<sup>2</sup>

<sup>1</sup>University of California, Berkeley

<sup>2</sup>University of Washington



# The web: a rich source of data!

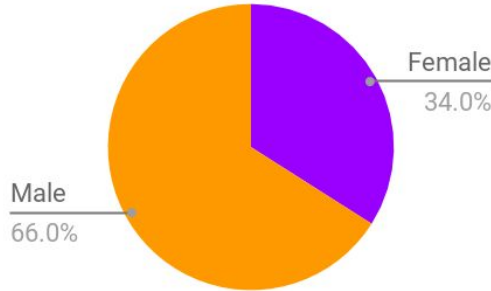
2008: Google indexed **1 trillion** pages

Now: indexes > **60 trillion** pages

→ lots of content out there

Have you written a scraper?

Percentages of Female and Male Speaking Characters - Top 100 Films of 2017



Woman director or writer: 42% female speaking roles  
Only male directors, writers: 32% female speaking roles

Martha M. Lauzen. 2018. It's a Man's (Celluloid) World: Portrayals of Female Characters in the 100 Top Films of 2017

Find Movies, TV shows, Celebrities and more.. All

IMDb

Movies, TV & Showtimes Celebs, Events & Photos Sign in with Facebook

### Top-US-Grossing Feature Films Released 2017-01-01 to 2017-12-31

1 to 50 of 11,605 titles | Next » View Mode: Compact Detailed

Sort by: Popularity Alphabetical IMDb Rating Number of Votes US Box Office Runtime Year Release Date

	1. <a href="#">Star Wars: The Last Jedi</a> (2017)	7.2	☆ Rate +
	2. <a href="#">Beauty and the Beast</a> (2017)	7.2	☆ Rate +
	3. <a href="#">Wonder Woman</a> (2017)	7.5	☆ Rate +
	4. <a href="#">Jumanji: Welcome to the Jungle</a> (2017)	7	☆ Rate +
	5. <a href="#">Guardians of the Galaxy Vol. 2</a> (2017)	7.7	☆ Rate +
	6. <a href="#">Spider-Man: Homecoming</a> (2017)	7.5	☆ Rate +
	7. <a href="#">It (I)</a> (2017)	7.4	☆ Rate +
	8. <a href="#">Thor: Ragnarok</a> (2017)	7.9	☆ Rate +
	9. <a href="#">Despicable Me 3</a> (2017)	6.3	☆ Rate +
	10. <a href="#">Justice League</a> (2017)	6.6	☆ Rate +
	11. <a href="#">Logan</a> (2017)	8.1	☆ Rate +
	12. <a href="#">The Fate of the Furious</a> (2017)	6.7	☆ Rate +
	13. <a href="#">Coco (I)</a> (2017)	8.4	☆ Rate +
	14. <a href="#">Dunkirk</a> (2017)	8	☆ Rate +
	15. <a href="#">Star Wars: The Last Jedi</a> (2017)	7.2	☆ Rate +

# Let's automate!

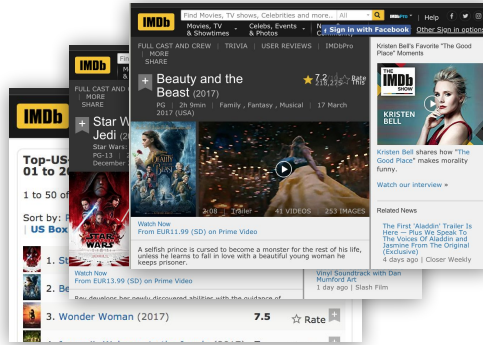


We've got some libraries...

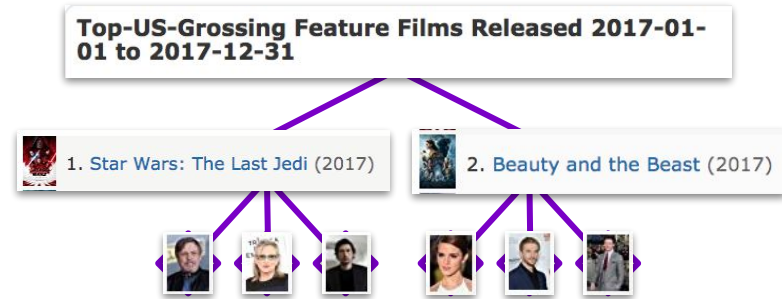
common thread: users  
must reverse engineer  
target webpages



# Formative Study: What kinds of web data?



distributed  
must navigate between pages -  
e.g., click, use forms + widgets



hierarchical  
must traverse and collect  
tree-structured data

# Formative Study: Can social scientists use...

Traditional  
programming?

Skills:

Basic programming

Web DSL

DOM

JavaScript

Server interaction



Manual  
collection?

Skills:

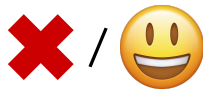
Browser use

But

Slow

Tedious

Small-scale data



Programming by  
demonstration?

Skills:

Browser use

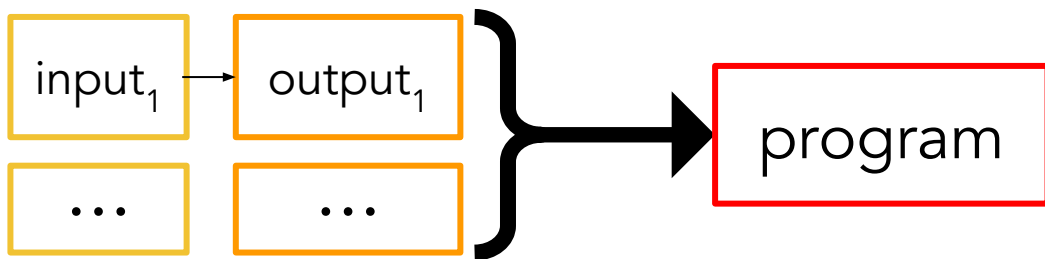
But

Can't collect  
distributed,  
hierarchical datasets

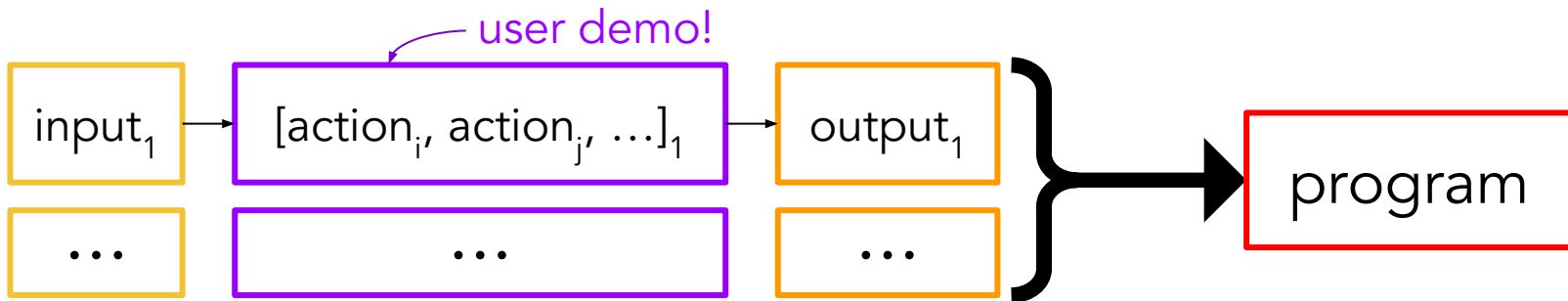


# What's Programming by Demonstration (PBD)?

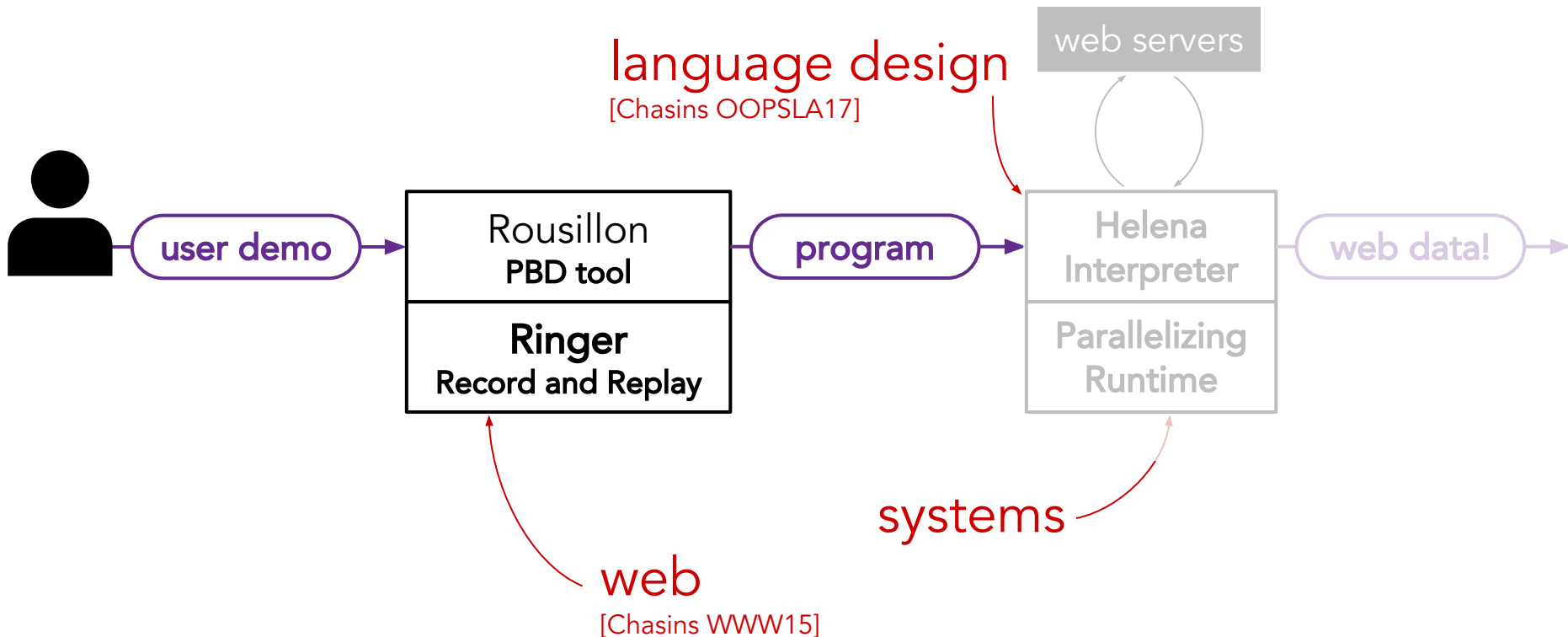
Closely related to Programming by Example (PBE) (e.g., FlashFill)



But PBD (e.g., SMARTedit) gets to see the input being transformed into the output:



# The Helena Ecosystem



# The Interaction Model

load `https://www.imdb.com/...`

user  
demonstrates  
how to collect  
one joined row

start recording

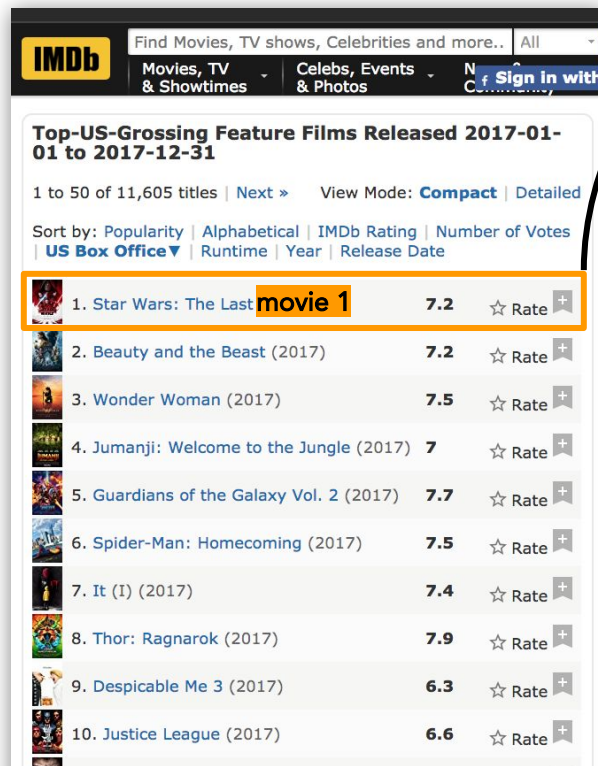
load `www.imdb.com...`

collect movie 1

click movie 1

collect actor 1

end recording








IMDb Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos Sign in with

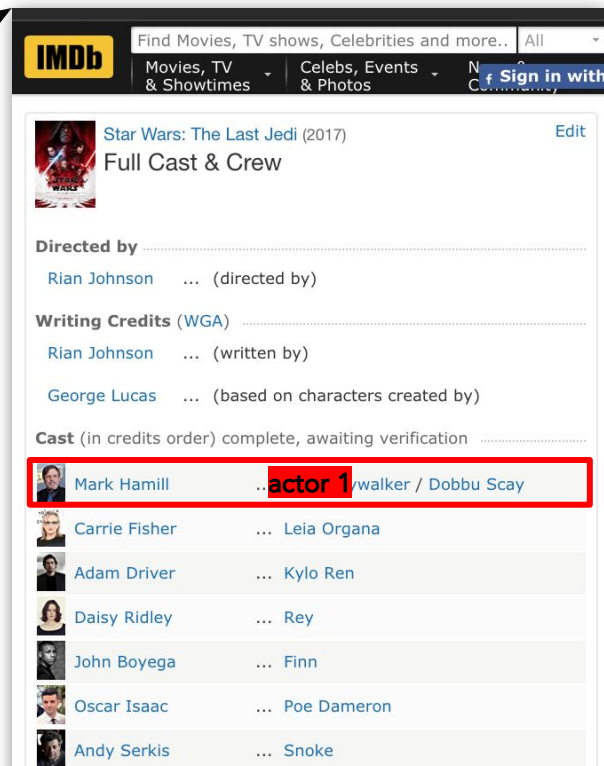
**Top-US-Grossing Feature Films Released 2017-01-01 to 2017-12-31**

1 to 50 of 11,605 titles | Next » View Mode: Compact Detailed

Sort by: Popularity Alphabetical IMDb Rating Number of Votes US Box Office Runtime Year Release Date

	1. Star Wars: The Last Jedi	movie 1	7.2	☆ Rate
	2. Beauty and the Beast (2017)		7.2	☆ Rate
	3. Wonder Woman (2017)		7.5	☆ Rate
	4. Jumanji: Welcome to the Jungle (2017)		7	☆ Rate
	5. Guardians of the Galaxy Vol. 2 (2017)		7.7	☆ Rate
	6. Spider-Man: Homecoming (2017)		7.5	☆ Rate
	7. It (I) (2017)		7.4	☆ Rate
	8. Thor: Ragnarok (2017)		7.9	☆ Rate
	9. Despicable Me 3 (2017)		6.3	☆ Rate
	10. Justice League (2017)		6.6	☆ Rate

click



IMDb Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos Sign in with







**Star Wars: The Last Jedi (2017)** Edit

**Full Cast & Crew**

**Directed by** Rian Johnson ... (directed by)

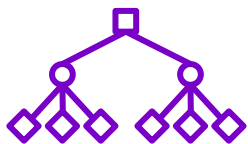
**Writing Credits (WGA)** Rian Johnson ... (written by) George Lucas ... (based on characters created by)

**Cast (in credits order) complete, awaiting verification**

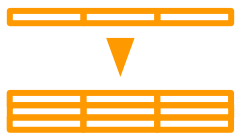
	Mark Hamill	.. actor 1	/walker / Dobbu Scay
	Carrie Fisher	... Leia Organa	
	Adam Driver	... Kylo Ren	
	Daisy Ridley	... Rey	
	John Boyega	... Finn	
	Oscar Isaac	... Poe Dameron	
	Andy Serkis	... Snoke	



# Can we even offer this interaction model?



**Hierarchical Data:** Synthesis of nested loops - needed for hierarchical data - is a long-standing open problem.



**Relation Ambiguity:** Single row is an ambiguous demo. Which relation did the user intend to select?



**Readability:** For robust automation, must run 100s of low-level, unreadable DOM events.



# Problem 1: Hierarchical Data

Top-US-Grossing Feature Films Released 2017-01-01 to 2017-12-31



hierarchical data → nested loops

## The issue:

Nested loop synthesis is an open problem.

```
for movie in movie_list:
    // scrape movie data
    for actor in actor_list:
        // scrape actor data
```

## Past solutions:

In web automation, none. In other domains, manually marking loop boundaries.

progs  
w/ no  
loops

progs w/  
single-level  
loops

progs w/  
nested loops

The space of possible programs is just too big. To pick among all these, our spec is ambiguous.



# Problem 1: Hierarchical Data

## Our solution:

Design user interaction to make search tractable

Contract w/ user: perform one iteration of each loop, ordered from outer to inner

## Label uses of relation cells

### movie relation

	1. Star Wars: The Last Jedi (2017)	7.2	☆ Rate	+
	2. Beauty and the Beast (2017)	7.2	☆ Rate	+
	3. Wonder Woman (2017)	7.5	☆ Rate	+
	4. Jumanji: Welcome to the Jungle (2017)	7	☆ Rate	+

### actor relation

	Mark Hamill
	Carrie Fisher
	Adam Driver
	Daisy Ridley
	John Boyega
...	Finn

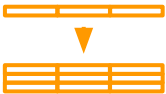
## PBD takeaway:

To add loops efficiently, first find objects that should be treated together.

One loop per relation, start before cell use

```
load https://www.imdb.com/se... into p1
scrape Star Wars: The Last Jedi in p1 and call it movie_title movie cell
scrape (2017) in p1 and call it movie_year movie cell
click Star Wars: The Last Jedi in p1 movie cell
scrape Mark Hamill in p2 and call it actor_name actor cell
scrape Luke Skywalker in p2 and call it actor_role actor cell
```

```
load https://www.imdb.com/se... into p1
for movie in movie_list:
  scrape Star Wars: The Last Jedi in p1 and call it movie_title movie cell
  scrape (2017) in p1 and call it movie_year movie cell
  click Star Wars: The Last Jedi in p1 movie cell
for actor in actor_list:
  scrape Mark Hamill in p2 and call it actor_name actor cell
  scrape Luke Skywalker in p2 and call it actor_role actor cell
```



## Problem 2: Relation Ambiguity

scrape

Top-US-Grossing Feature Films Released 2017-01-01 to 2017-12-31

1 to 50 of 11,607 titles | Next » | View Mode: Compact | Detailed

Sort by: Popularity | Alphabetical | IMDb Rating | Number of Votes  
US Box Office | Runtime | Year | Release Date

1. **Star Wars: The Force Awakens** (2017)  
PG-13 | 152 min | Action, Adventure, Fantasy  
★ 7.2 | Rate this | 85 Metascore  
Rey develops her newly discovered abilities with the guidance of Luke Skywalker, who is unsettled by the strength of her powers. Meanwhile, the Resistance prepares for battle with the First Order.  
Director: Rian Johnson | Stars: Daisy Ridley, John Boyega, Mark Hamill, Carrie Fisher  
Votes: 419,970 | Gross: \$620.18M

2. **Beauty and the Beast** (2017)  
PG | 129 min | Family, Fantasy, Musical  
★ 7.2 | Rate this | 65 Metascore  
A selfish prince is cursed to become a monster for the rest of his life, unless he learns to fall in love with a beautiful young woman he keeps prisoner.  
Director: Bill Condon | Stars: Emma Watson, Dan Stevens, Luke Evans, Josh Gad  
Votes: 218,261 | Gross: \$504.01M

3. **Wonder Woman** (2017)  
PG-13 | 141 min | Action, Adventure, Fantasy  
★ 7.5 | Rate this | 76 Metascore

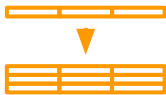
Given this demo, what's the right relation? Is node 1 included? If not, do we want purple or orange cells in rows 2 and 3? Maybe purple + orange + unhighlighted?

### The issue:

Can extract many relations from one page.  
Set of interacted nodes → 1 chosen relation?

### Past solutions:

Have user label multiple rows.



# Problem 2: Relation Ambiguity

IMDb Find Movies, TV shows, Celebrities and more.. All

Movies, TV & Showtimes Celebs, Events & Photos Sign in with

**Top-US-Grossing Feature Films Released 2017-01-01 to 2017-12-31**

1 to 50 of 11,607 titles | Next » View Mode: Compact Detailed

Sort by: Popularity | Alphabetical | IMDb Rating | Number of Votes  
US Box Office Runtime Year Release Date

1. **Star Wars: The Last Jedi** (2017)  
PG-13 | 152 min | Action, Adventure, Fantasy  
★ 7.2 Rate this 85 Metascore  
Rey develops her newly discovered abilities with the guidance of Luke Skywalker, who is unsettled by the strength of her powers. Meanwhile, the Resistance prepares for battle with the First Order.  
Director: Rian Johnson | Stars: Daisy Ridley, John Boyega, Mark Hamill, Carrie Fisher  
Votes: 419,970 | Gross: \$620.18M

2. **Beauty and the Beast** (2017)  
PG | 129 min | Family, Fantasy, Musical  
★ 7.2 Rate this 65 Metascore  
A selfish prince is cursed to become a monster for the rest of his life, unless he learns to fall in love with a beautiful young woman he keeps prisoner.  
Director: Bill Condon | Stars: Emma Watson, Dan Stevens, Luke Evans, Josh Gad  
Votes: 218,261 | Gross: \$504.01M

3. **Wonder Woman** (2017)  
PG-13 | 141 min | Action, Adventure, Fantasy  
★ 7.5 Rate this 76 Metascore

## Our solution:

$S$  = subsets of interacted nodes of size  $n \dots 1$   
for row1 in  $S$ :

```
shape = getSubtreeShape(row1)
row2 = siblingWithShape(row1, shape)
relation = extractRelation([row1, row2])
if relation:
    return relation
```

$\text{siblingWithShape}([n1, n2], s) \rightarrow \emptyset$

$\text{siblingWithShape}([n2], s) \rightarrow n3$

$\text{relation} \rightarrow [n2, n3, n4]$

## PBD takeaway:

Take advantage of domain-specific patterns (e.g, web design best practices) to find objects we should treat together



# Problem 3: Readability

...

```
[event45]type:dom
type:focus
xpath:HTML/BODY[1]/DIV[2]/DIV[1]/DIV[2]/DIV[5]/DIV[3]/DIV[6]/DIV[1]/A[1]
URL:https://www.imdb.com/title/tt2527336/?ref_=adv_li_tt
port:3
[event46]type:dom
type:mouseup
xpath:HTML/BODY[1]/DIV[2]/DIV[1]/DIV[2]/DIV[5]/DIV[3]/DIV[6]/DIV[1]/A[1]
URL:https://www.imdb.com/title/tt2527336/?ref_=adv_li_tt
port:3
[event47]type:dom
type:click
xpath:HTML/BODY[1]/DIV[2]/DIV[1]/DIV[2]/DIV[5]/DIV[3]/DIV[6]/DIV[1]/A[1]
URL:https://www.imdb.com/title/tt2527336/?ref_=adv_li_tt
port:3
[event48]type:dom
type:keyup
xpath:HTML/BODY[1]/DIV[2]/DIV[1]/DIV[2]/DIV[5]/DIV[3]/DIV[6]/DIV[1]/A[1]
URL:https://www.imdb.com/title/tt2527336/?ref_=adv_li_tt
port:3
[event53]type:dom
type:blur
xpath:HTML/BODY[1]/DIV[2]/DIV[1]/DIV[2]/DIV[5]/DIV[3]/DIV[6]/DIV[1]/A[1]
URL:https://www.imdb.com/title/tt2527336/?ref_=adv_li_tt
port:3
[event60]type:dom
type:keydown
xpath:HTML/BODY[1]
URL:https://www.imdb.com/title/tt2527336/fullcredits?ref_=tt_cl_sm#cast
port:6
```

Page allowed to react to any DOM event → prog must run low-level events like this to be robust on modern interactive DOM + JS + AJAX pages

**The issue:**  
It's not readable.

**Past solutions:**  
Actually, it's a new problem.

**Our solution:**  
Reverse compilation

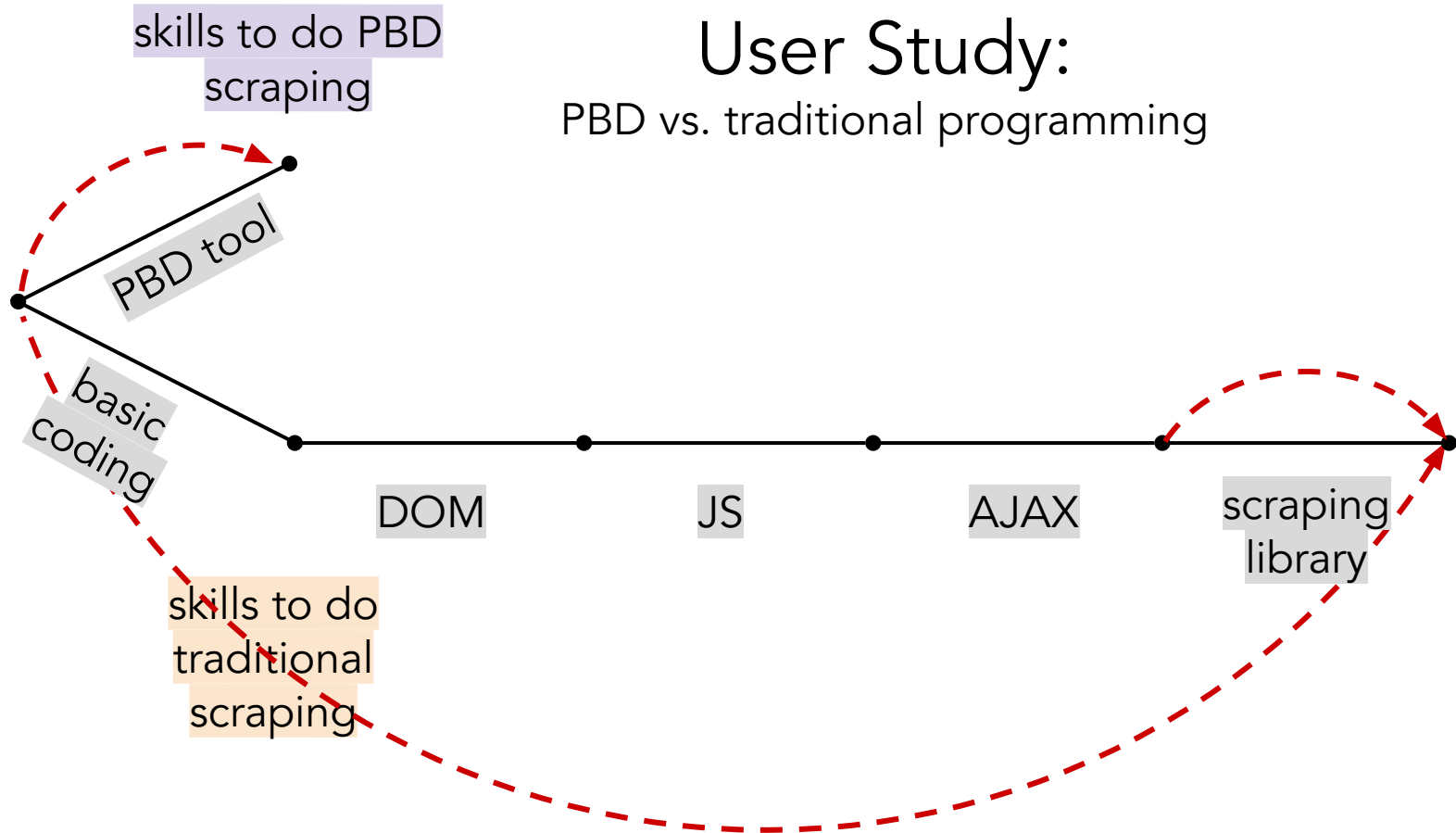
**PBD takeaway:**  
It's ok to record demo at one level, show program at another.

```
load www.imdb... into p1
scrape movie_title in p1
scrape movie_year in p1
click movie_title in p1
scrape actor_name in p2
scrape actor_role in p2
```

LET'S SCRAPE SOME STUFF!

# User Study:

PBD vs. traditional programming





# User Study:

PBD vs. traditional programming

## **Setup:**

Within-subject study, 15 CS PhD students

1 task, 2 tools; Helena then Selenium OR Selenium then Helena

9/15 prior scraping experience

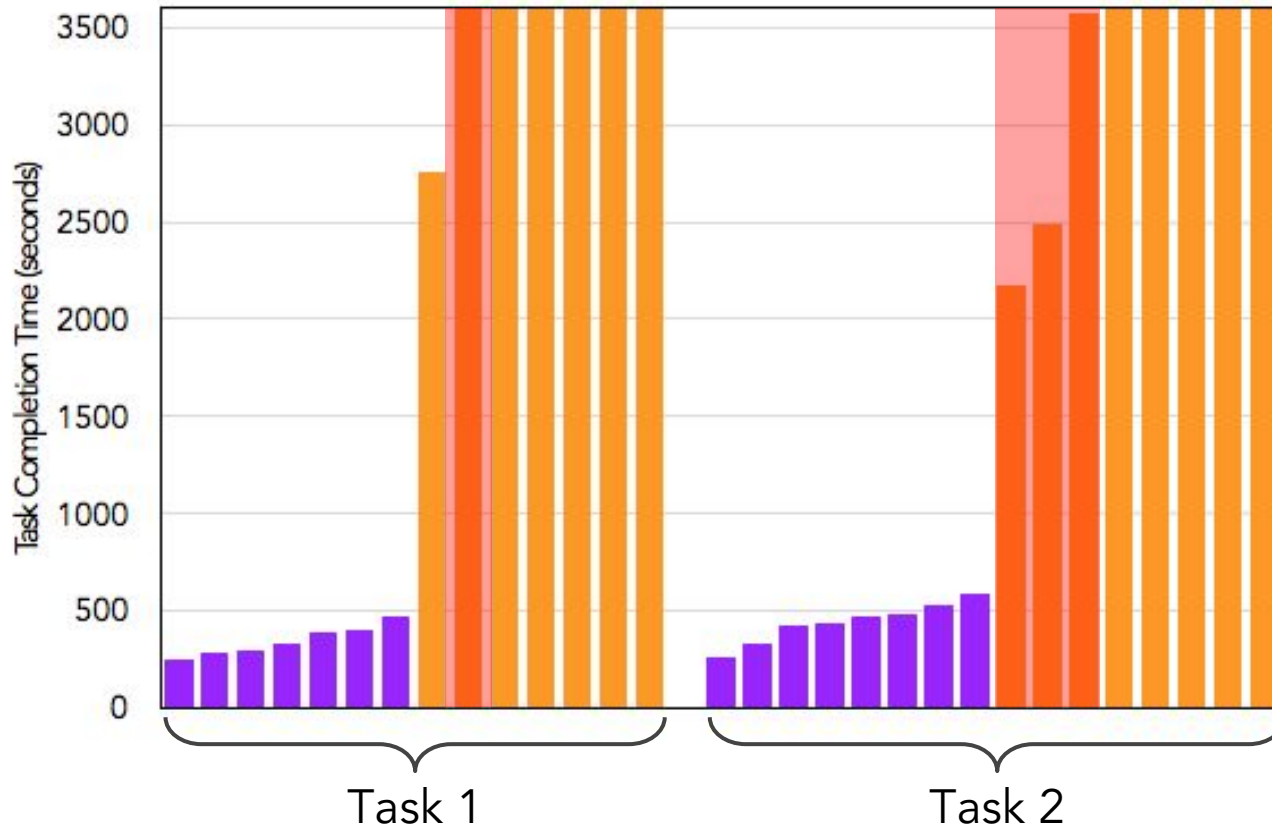
4/15 prior Selenium experience

## **Context:**

PBD vs. traditional programming eval is rare

To date, solid speedups, but only small tasks (best averaged 12 mins saved time)

# Q1: Can users learn PBD faster?



Helena

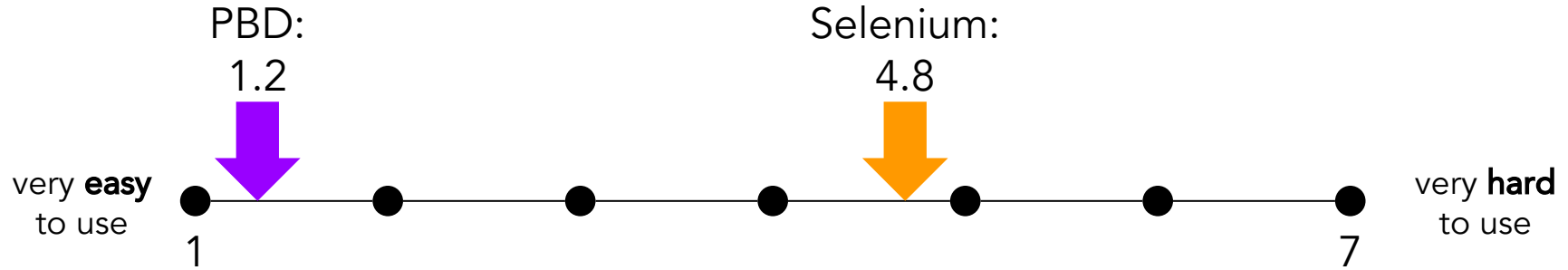
Selenium

Completion rate with Helena: 100%

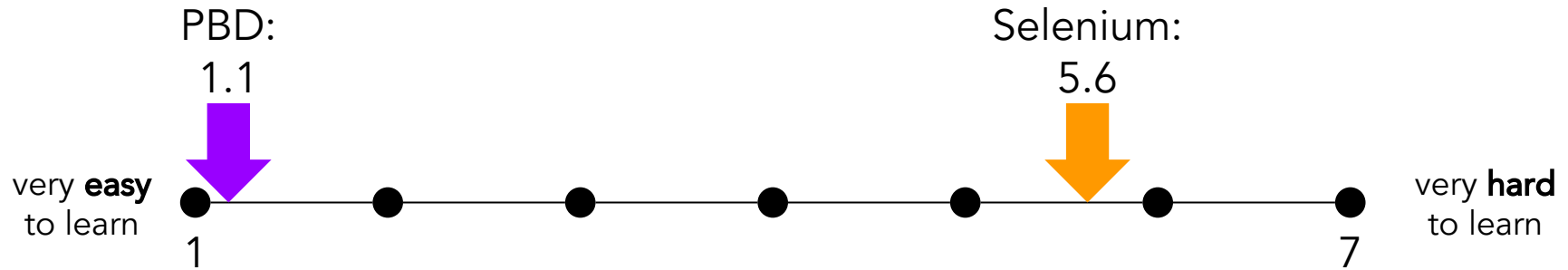
Completion rate with Selenium: 26.7%

Lower bound on time savings is 47 mins for task 1, 52 mins for task 2

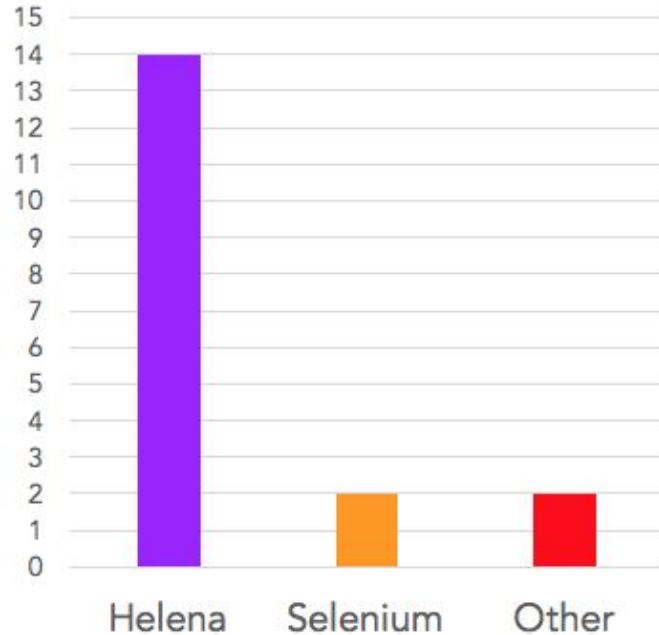
**Q2:** Do users perceive PBD as more usable?



**Q3:** Do users perceive PBD as more learnable?



**Q4:** Having already learned both tools, which tool would users want for future tasks?





[It] was very useful how it automatically inferred the nesting that I wanted when going to multiple pages so that I didn't have to write multiple loops.



Super easy to use... It felt like magic and for quick data collection tasks online I'd love to use it in the future.



Helena's way easier to use – point and click at what I wanted and it 'just worked' like magic. Selenium is more fully featured, but...pretty clumsy (inserting random sleeps into the script).

# The real test: social scientists and data scientists

DEPARTMENT OF SOCIOLOGY

UNIVERSITY of WASHINGTON

DEPARTMENT OF ECONOMICS

UNIVERSITY of WASHINGTON

15+ collaborations

TAL

6 different scrapers  
parallelized  
all run 24/7

N

BUBLIC  
CE

UNIVERSITY of WASHINGTON

Can we set housing voucher thresholds based on real-time neighborhood rents?



How is the minimum wage affecting Seattle restaurants?



Can we design a better carpool matching algorithm?

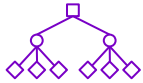


How do charitable foundations communicate with supporters?

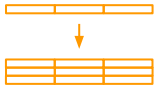


# Contributions

- A demonstration model that users love
- Solutions for key technical challenges:



**Hierarchical Data**



**Relation Ambiguity**




**Readability**



# Helena Scraper and Automator

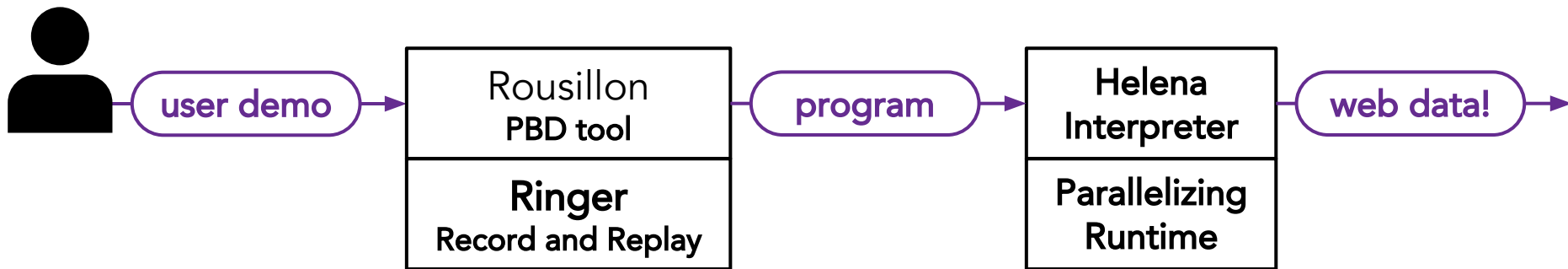
[helena-lang.org/install](https://helena-lang.org/install)

[github.com/schasins/helena](https://github.com/schasins/helena) 

Want to use the  
tool yourself?

Use it to write:

- Parallel and distributed scrapers
- Programs for non-scraping web automation tasks
- Voice automation 'skills'



[schasins@cs.berkeley.edu](mailto:schasins@cs.berkeley.edu)



[@sarahchasins](https://twitter.com/sarahchasins)

I'm on the academic job market!