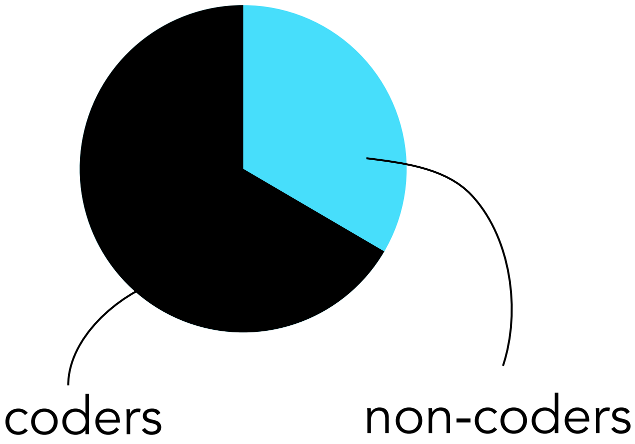
Skip Blocks: Reusing Execution History to Accelerate Web Scripts

Sarah Chasins
University of California, Berkeley

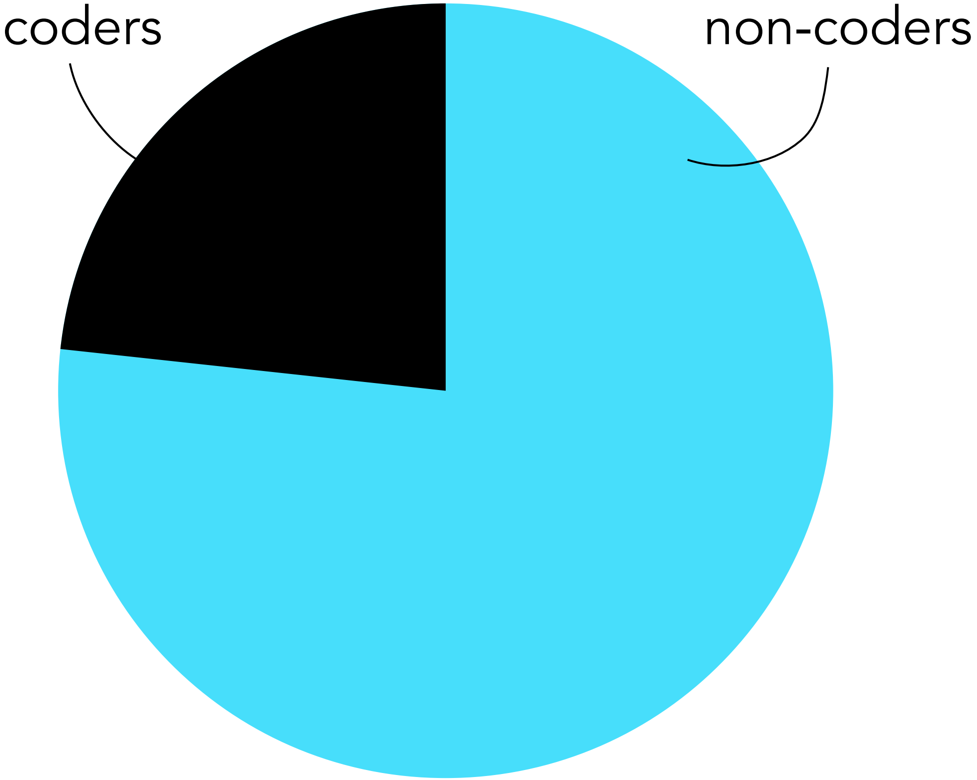
Rastislav Bodik
University of Washington

OOPSLA Oct 25, 2017 **Vancouver**

care about data today



care about data tomorrow



What web data collection tools do we have?

tools that require users to reverse engineer target webpages

```
    <link href="asset
  </head>
    <body>
      <div id="main">

          lass="h
        <div    ss="pull-right">
          <a hr  ="https://eecs.berkeley.edu/" title="  C   keley EECS"><img
src="./index files uc-berkeley-logo.png" height="21px"        ="_blank"></a> &nbsp;
          <a hre  "https://uwplse.org/" title="UW PLSE"><im  src="./index_files/uw-plse-
lo   ng"     ight="2 px" target="_blank"></a>
          </d
        <div c    "mark">
          <sp   lass="uw">Helena</span>
        </div
              id="title">Web  utom   on for En  Us  </div>

              ontent">

              "mission">
              s a high-level programming la       for web automation.
              omes with Rousillon, a programmi     demonstration (PBD) tool      iting
Helena programs.
               the Rousillon browser ext      n, demonstrate how to collect    firs
              n let Rousillon write a p      n for collecting the remaining


              "header">
              s="lead">Tools</div>

              class="paper project">

          <div class="thumbnail">
            <a class="thumbnail-wrapper" href="https://                       /helena"
target="_blank"><div class="thumbnail-img" style="background-im
```

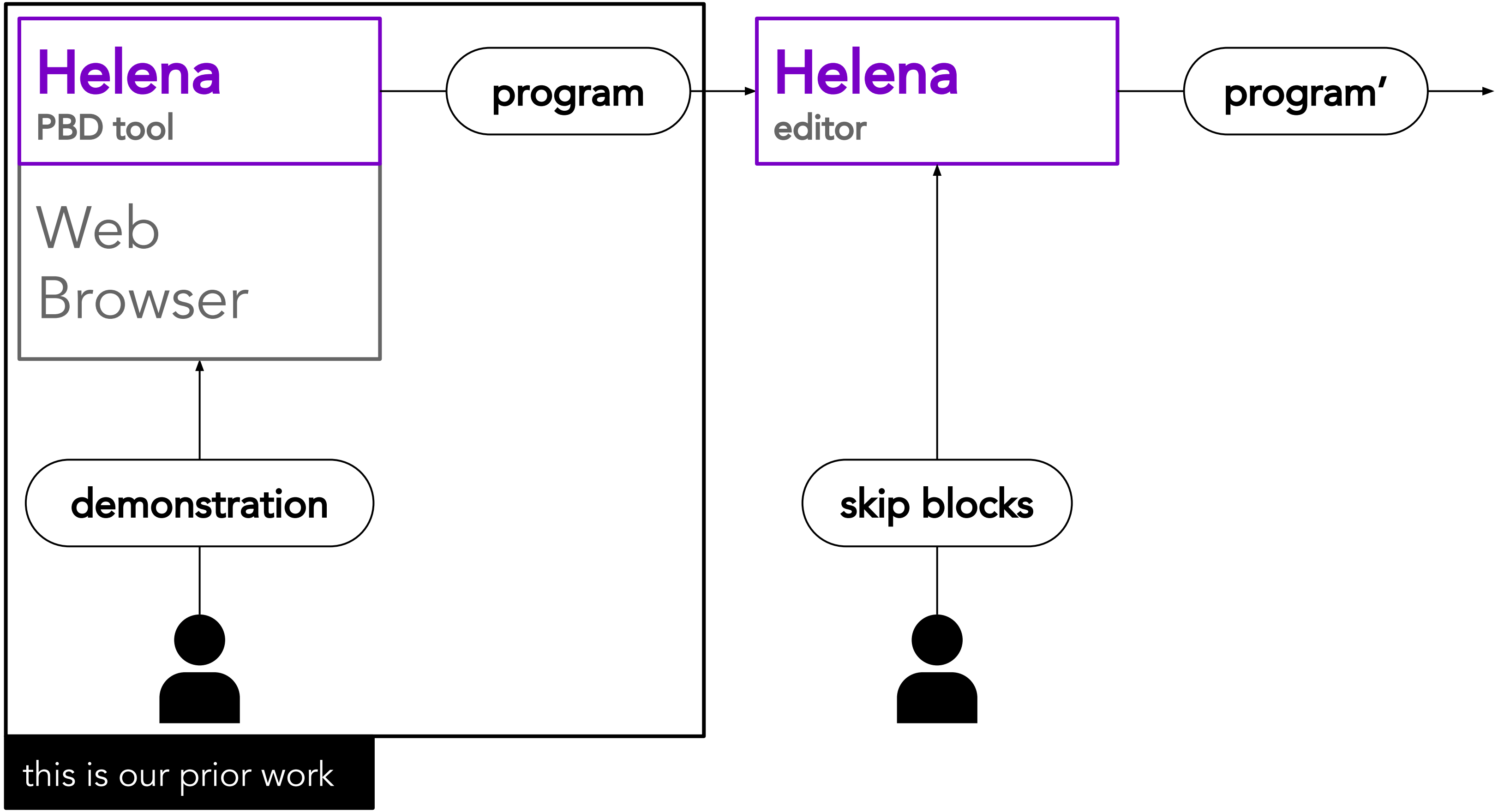
- hire a human to copy & paste
- hire a coder to use one of these

- **Helena**
  WEB AUTOMATION
  FOR END USERS

our tool!

coders

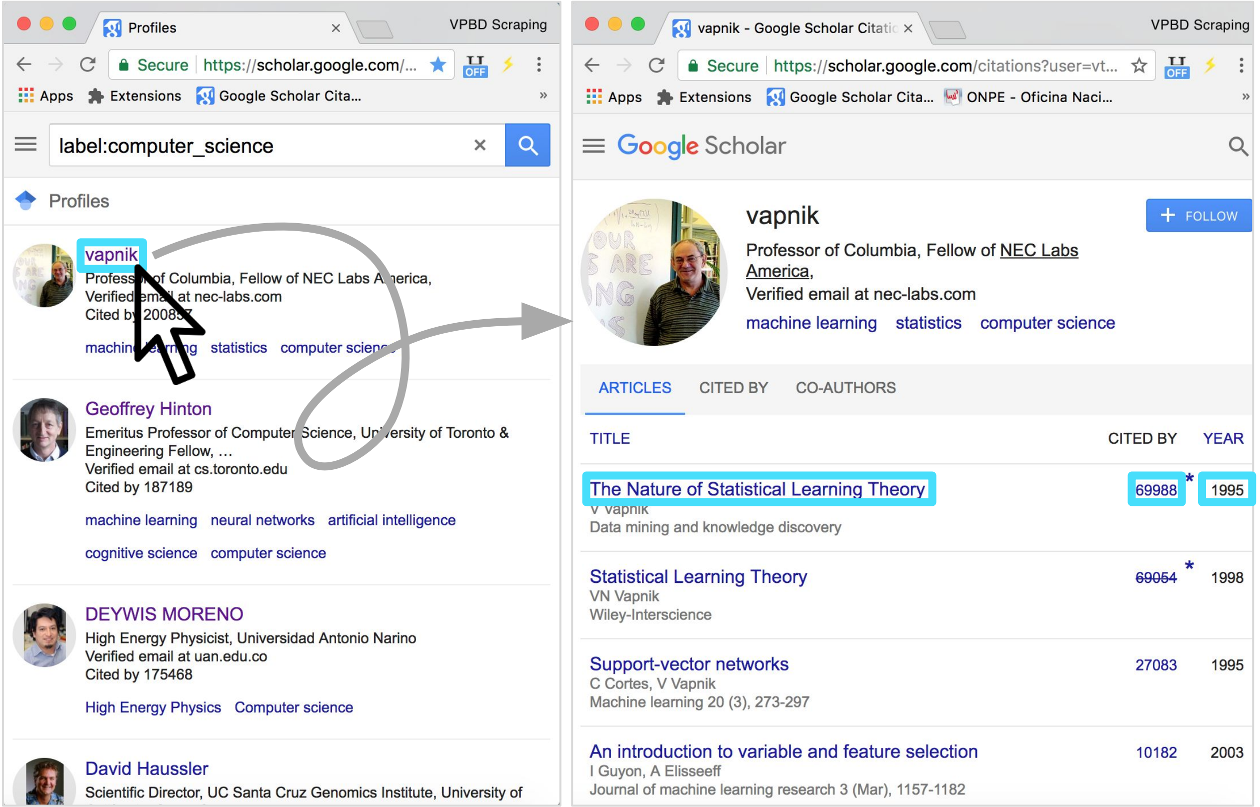
non-coders


Helena
PBD tool
Web
Browser
program
Helena
editor
program'
demonstration
skip blocks
this is our prior work

Let’s PBD a web automation script!

Goal: scrape all papers by top 10,000 CS authors from Google Scholar

input: a demonstration



Profiles × VPBD Scraping

Secure https://scholar.google.com/...

Apps Extensions Google Scholar Cita...

label:computer_science

Profiles

vapnik
Professor of Columbia, Fellow of NEC Labs America,
Verified email at nec-labs.com
Cited by 200857
machine learning statistics computer science

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto & Engineering Fellow, ...
Verified email at cs.toronto.edu
Cited by 187189
machine learning neural networks artificial intelligence cognitive science computer science

DEYWIS MORENO
High Energy Physicist, Universidad Antonio Narino
Verified email at uan.edu.co
Cited by 175468
High Energy Physics Computer science

David Haussler
Scientific Director, UC Santa Cruz Genomics Institute, University of

vapnik - Google Scholar Citatic × VPBD Scraping

Secure https://scholar.google.com/citations?user=vt...

Apps Extensions Google Scholar Cita... ONPE - Oficina Naci...

Google Scholar

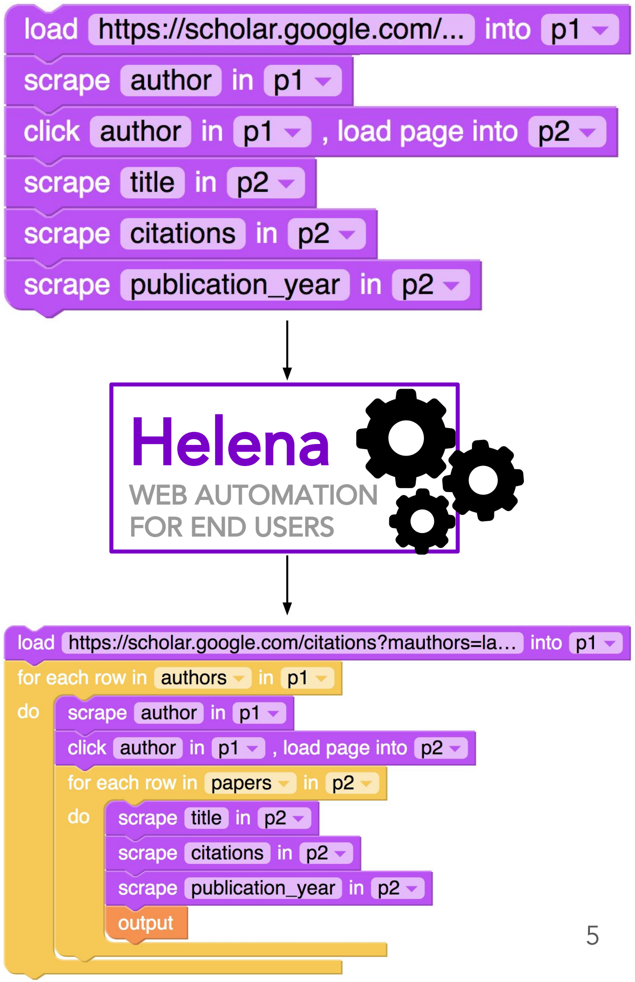
vapnik
Professor of Columbia, Fellow of NEC Labs America,
Verified email at nec-labs.com
machine learning statistics computer science

FOLLOW

ARTICLES CITED BY CO-AUTHORS

TITLE	CITED BY	YEAR
The Nature of Statistical Learning Theory V Vapnik Data mining and knowledge discovery	69988 *	1995
Statistical Learning Theory VN Vapnik Wiley-Interscience	69054 *	1998
Support-vector networks C Cortes, V Vapnik Machine learning 20 (3), 273-297	27083	1995
An introduction to variable and feature selection I Guyon, A Elisseeff Journal of machine learning research 3 (Mar), 1157-1182	10182	2003

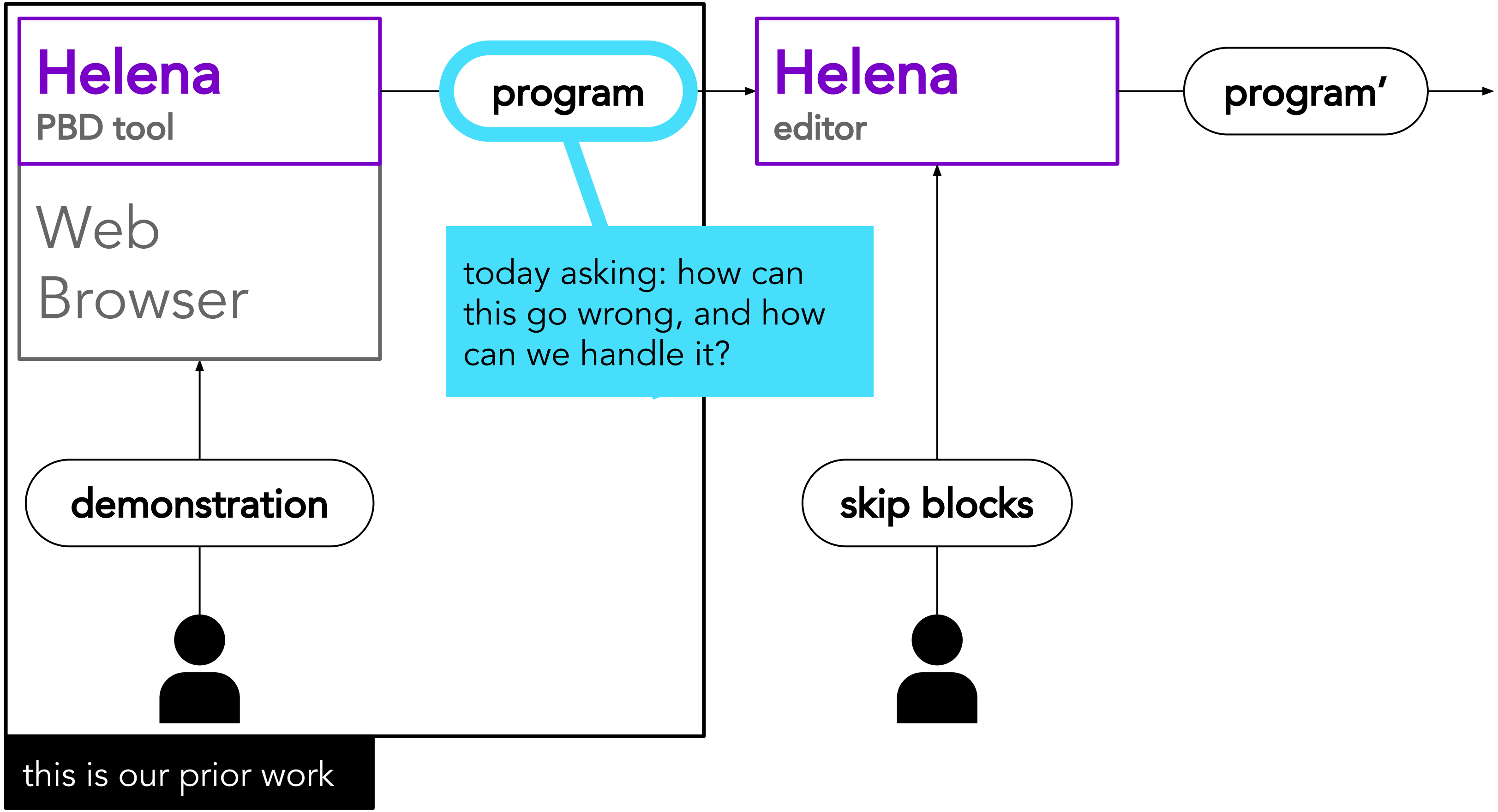
output: a script



```
load https://scholar.google.com/... into p1
scrape author in p1
click author in p1 , load page into p2
scrape title in p2
scrape citations in p2
scrape publication_year in p2
```

Helena
WEB AUTOMATION FOR END USERS

```
load https://scholar.google.com/citations?mauthors=la... into p1
for each row in authors in p1
do
    scrape author in p1
    click author in p1 , load page into p2
    for each row in papers in p2
    do
        scrape title in p2
        scrape citations in p2
        scrape publication_year in p2
        output
```



today asking: how can this go wrong, and how can we handle it?

this is our prior work

DEPARTMENT OF SOCIOLOGY
UNIVERSITY *of* WASHINGTON

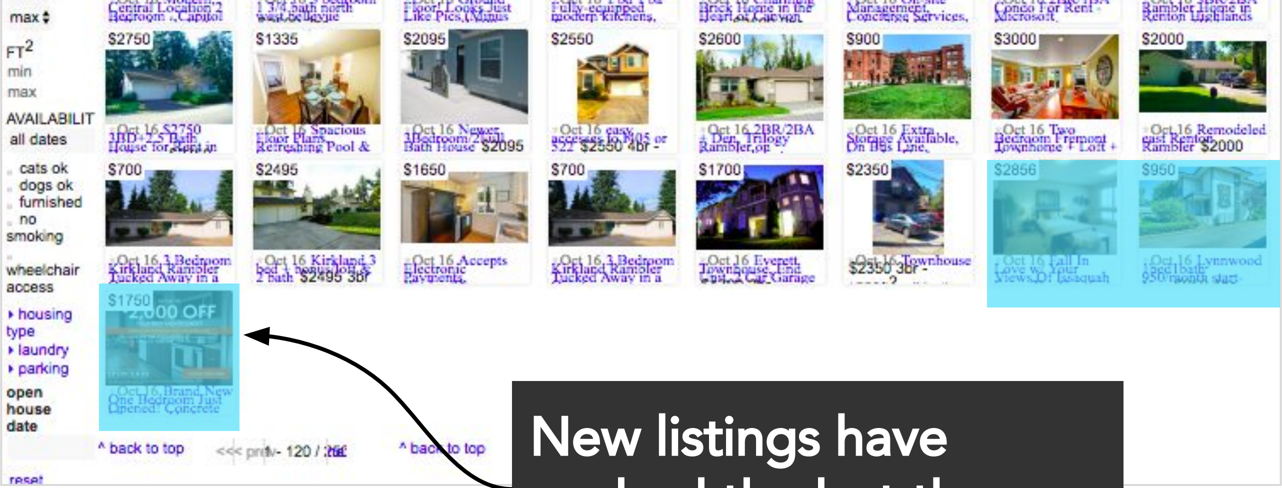
+

City of Seattle

How is rent changing across Seattle neighborhoods?

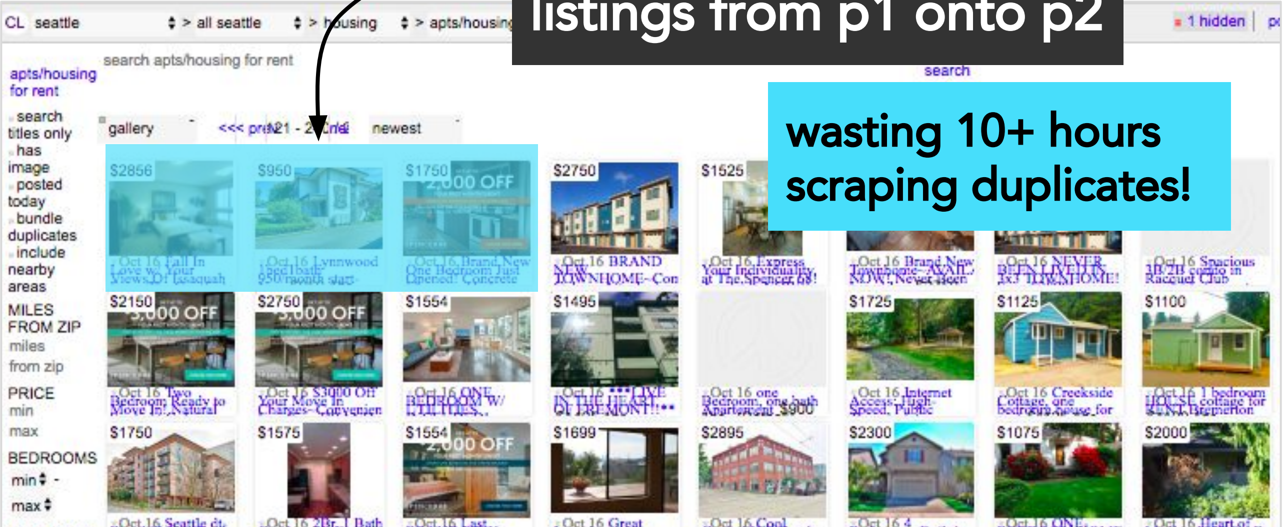
Kept losing network connection

page 1



New listings have pushed the last three listings from p1 onto p2

page 2



wasting 10+ hours scraping duplicates!

DEPARTMENT OF ECONOMICS

UNIVERSITY *of* WASHINGTON

How is the minimum wage affecting Seattle restaurants?

yelp

CIVIL & ENVIRONMENTAL ENGINEERING

UNIVERSITY *of* WASHINGTON

Can we design a better carpool matching algorithm?

zimride

by enterprise

EVANS SCHOOL OF PUBLIC POLICY & GOVERNANCE

UNIVERSITY *of* WASHINGTON

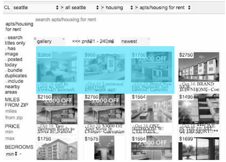
How do charitable foundations communicate with supporters?

Community Foundation

Atlas

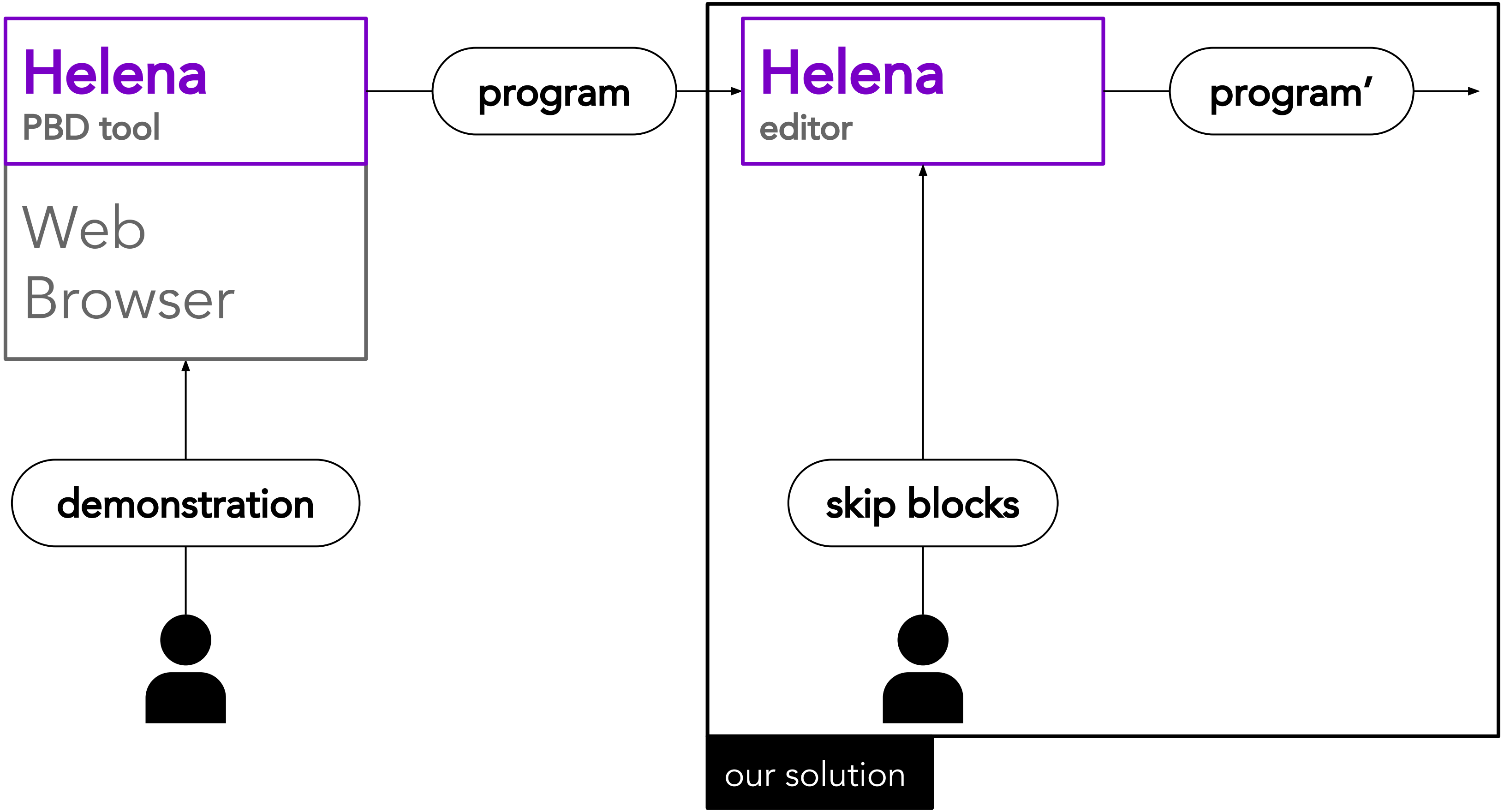
Problem Statement

(1) **Failures**: What happens when the network fails, the server fails, the computer fails? When we lose our session with the server and have to start over?



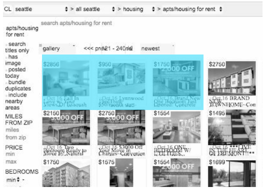
(2) **Data changes**: What happens when the server gives the client pages produced from different (potentially conflicting) reads of the underlying data store?

not client side problems → scraping script can't prevent them, must handle them


Helena
PBD tool
Web Browser
program
Helena
editor
program’
demonstration
skip blocks
our solution

Solution

on the surface, seem like very different problems


CL seattle > all seattle > housing > apts/housing for rent
search apts/housing for rent
apts/housing for rent
search titles only
has image
posted today
bundle duplicates
include nearby areas
MILES FROM ZIP
miles
from zip
PRICE
min
max
BEDROOMS
min -
gallery
<<< prev 121 - 240 next
newest
$2856
$950
$1750
2,000 OFF
$2750
Oct 16 Fall In Love w/ Your Views Of Issaquah
Oct 16 Lynnwood 1bed1bath 950/month start
Oct 16 Brand New One Bedroom Just Opened: Concrete
Oct 16 BRAND NEW TOWNHOME~Con
$2150
3,000 OFF
$2750
3,000 OFF
$1554
$1495
Oct 16 Two Bedroom Ready to Move In! Natural
Oct 16 $3000 Off Your Move In Charges~Convenien
Oct 16 ONE BEDROOM W/ UTILITIES
Oct 16 ***LIVE IN THE HEART OF FREMONT!!**
$1750
$1575
$1554
2,000 OFF
$1699

failures

data changes

"Just don't redo the same work you've already done!"

But what's the 'same' work? After all, data changes...

Our answer: the skip block! User can

- tell us what makes objects the same
- associate the code that operates on an object

- If object already committed (memoized), skip block; else, run block
- No reverse engineering! Reasoning about output data

Recovering from Failures

text-ify-ed representation of the block language

```
for (aRow in p1.authors){

        scrape aRow.author_name
        scrape aRow.author_institution
        p2 = click aRow.author_name
        for (pRow in p2.papers){
            scrape pRow.title
            scrape pRow.citations
            output([aRow.author_name, pRow.title, pRow.citations])
        }

}
```

scrape stuff about the author, click the author

for the author’s papers, scrape paper stuff

add a row of output with the author and paper info

Recovering from Failures

key attributes: is the current author the same as another we've already seen?

```
for (aRow in p1.authors){
     skipBlock(Author(aRow.author_name, aRow.author_institution)){
          scrape aRow.author_name
          scrape aRow.author_institution
          p2 = click aRow.author_name
          for (pRow in p2.papers){
               scrape pRow.title
               scrape pRow.citations
               output([aRow.author_name, pRow.title, pRow.citations])
          }
     }
}
```

block: the code that operates on the author object

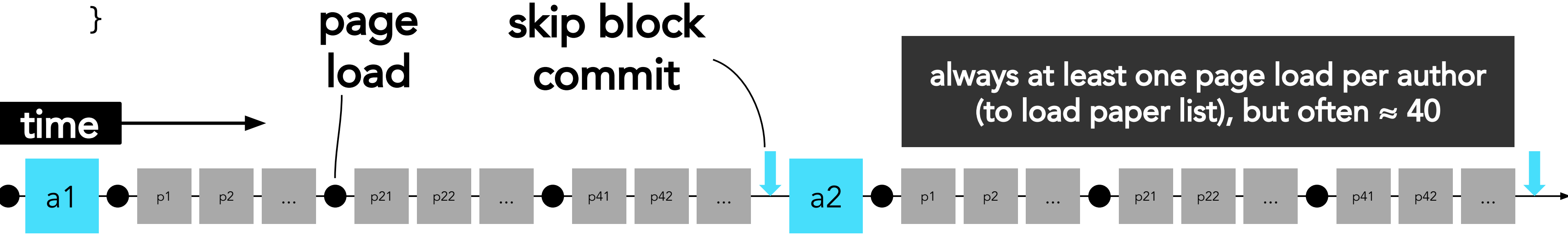
if ever, **in any run**, script has committed an object with the same key attributes, skips the block

Recovering from Failures

```
for (aRow in p1.authors){
    skipBlock(Author(aRow.author_name, aRow.author_institution)){
        scrape aRow.author_name
        scrape aRow.author_institution
        p2 = click aRow.author_name
        for (pRow in p2.papers){
            scrape pRow.title
            scrape pRow.citations
            output([aRow.author_name, pRow.title, pRow.citations])
        }
    }
}
```

AUTHOR

PAPER



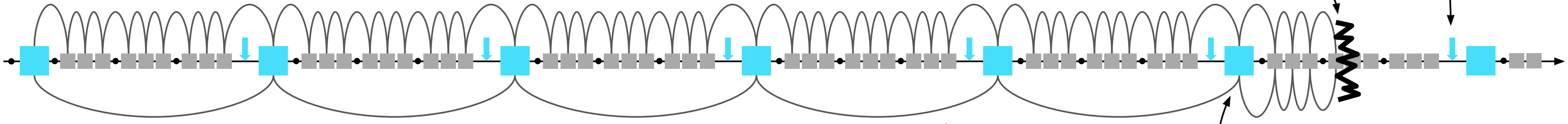
always at least one page load per author (to load paper list), but often ≈ 40

Recovering from Failures

recovery **without** the author skip block

external failure point

didn’t reach this commit



recovery **with** the author skip block

“fast-forwarding” over prior work

skips 40 page loads

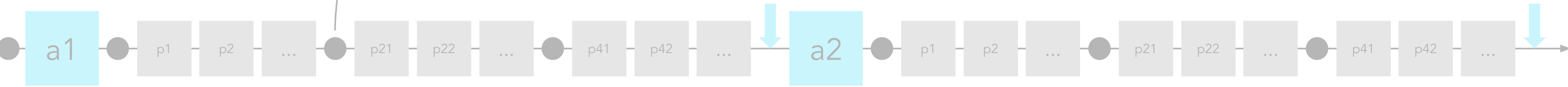
10 authors per page, so just 1 page load by this point, 200 skipped

page load

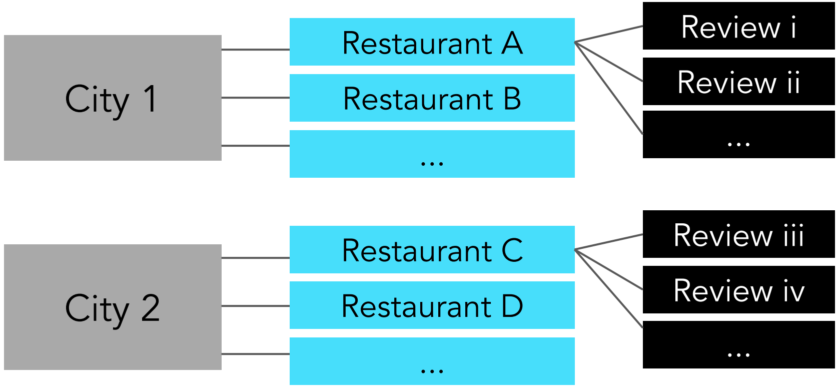
skip block commit

time

always at least one page load per author (to load paper list), but often ≈ 40



Nested Skip Blocks



In authors vs. papers, authors is clearly the right level for the skip block. But here?

skip block only at **city** → scraping a whole city takes many hours, so scraping half a city also takes hours

skip block only at **restaurant** → iterating through a city's restaurant list takes a long time, and now we have to go through all of Seattle, San Francisco before we can resume in the middle of Vancouver

skip block at **city** & **restaurant** → adjustable granularity skipping

Nested Skip Blocks

```
for (aRow in p1.authors){
    skipBlock(Author(aRow.author_name, aRow.author_institution)){
        scrape aRow.author_name
        scrape aRow.author_institution
        p2 = click aRow.author_name
        for (pRow in p2.papers){
            skipBlock(Paper(pRow.title, pRow.year)){
                scrape pRow.title
                scrape pRow.citations
                output([aRow.author_name, pRow.title, pRow.citations])
            }
        }
    }
}
```

and the inner block may commit even if the outer doesn’t - like a nested open transaction

Refreshing a Dataset

```
for (aRow in p1.authors){
    skipBlock(Author(aRow.author_name, aRow.author_institution), -∞)){
        scrape aRow.author_name
        scrape aRow.author_institution
        p2 = click aRow.author_name
        for (pRow in p2.papers){
            skipBlock(Paper(pRow.title, pRow.year)){
                scrape pRow.title
                scrape pRow.citations
                output([aRow.author_name, pRow.title, pRow.citations])
            }
        }
    }
}
```

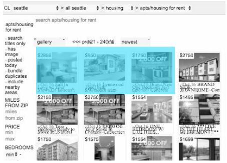
this is the default **staleness threshold**

-∞ means skip any duplicate we've seen ever

If we're scraping once a week, we don't want to revisit each author. But after a year, maybe we should see what's new.

Refreshing a Dataset

```
for (aRow in p1.authors){
    skipBlock(Author(aRow.author_name, aRow.author_institution), now - 365*24*60)){
        scrape aRow.author_name
        scrape aRow.author_institution
        p2 = click aRow.author_name
        for (pRow in p2.papers){
            skipBlock(Paper(pRow.title, pRow.year)){
                scrape pRow.title
                scrape pRow.citations
                output([aRow.author_name, pRow.title, pRow.citations])
            }
        }
    }
}
```



Also have logical time (ex: last 3 runs)

Bonus! In addition to failure recovery and data redundancy handling, get incremental/longitudinal scraping!

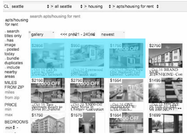
Demo time!

Benchmark Suite

Need web data? → Urban@UW → benchmark suite: 7 long-running web scraping tasks

Ex: for 50 top foundations, scrape the last 1,000 tweets they tweeted

Ex: scrape all Seattle apartment listings from Craigslist



Data Change

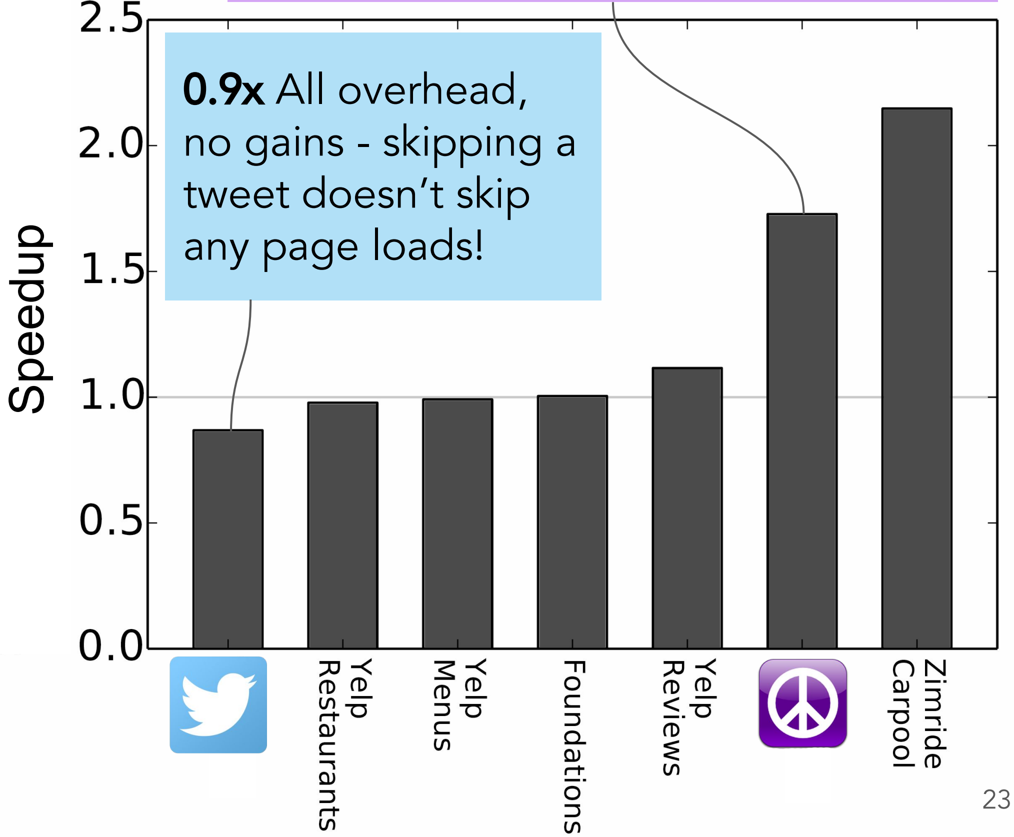
within one run

Measured full execution time of:

- Script with skip blocks
- Script without skip blocks

Chart shows speedup from using skip blocks

higher is better



1.7x Skipping one ad skips one page load, and pagination gives us so many duplicate ads!

0.9x All overhead, no gains - skipping a tweet doesn't skip any page loads!

Data Re-Scraping

within multiple runs

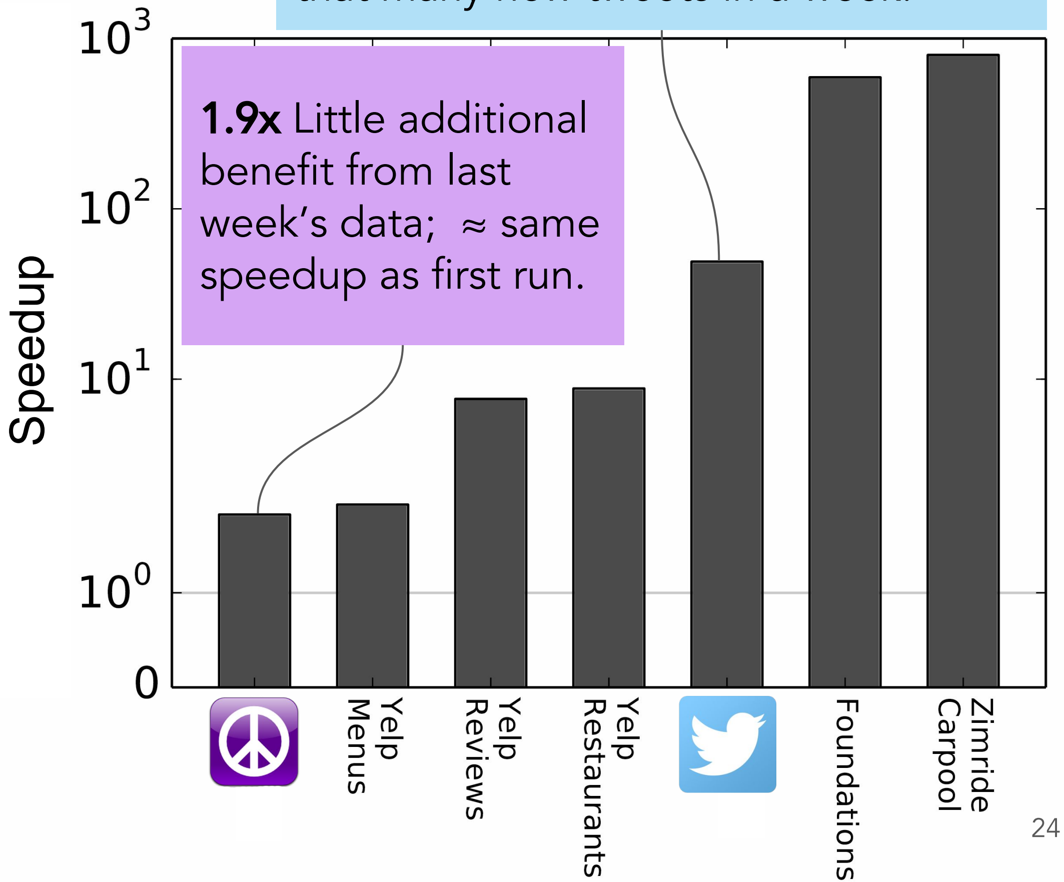
Executed script with skip blocks

One week later, measured full execution time of:

- Script with skip blocks
- Script without skip blocks

Chart shows speedup from using skip blocks

higher is better



49x Lots of benefit from last week's data. Gates Foundation doesn't post that many new tweets in a week!

1.9x Little additional benefit from last week's data; ≈ same speedup as first run.

Failure Recovery

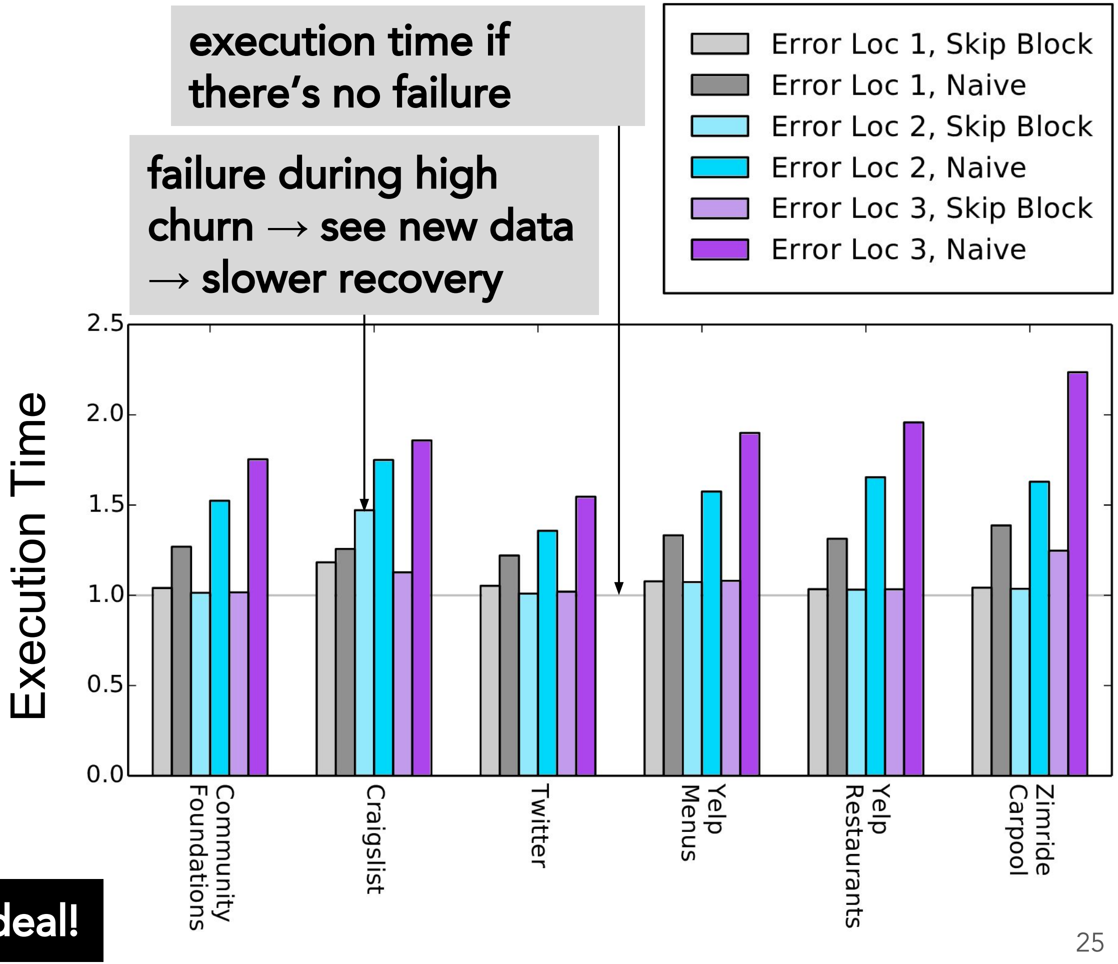
with skip block fast-forwarding

For each benchmark, for three failure locations, the execution time of:

- Script that recovers by naive restarting
- Script that recovers by skip block fastforwarding

Normalized by execution time of a script that doesn't encounter failures

lower is better



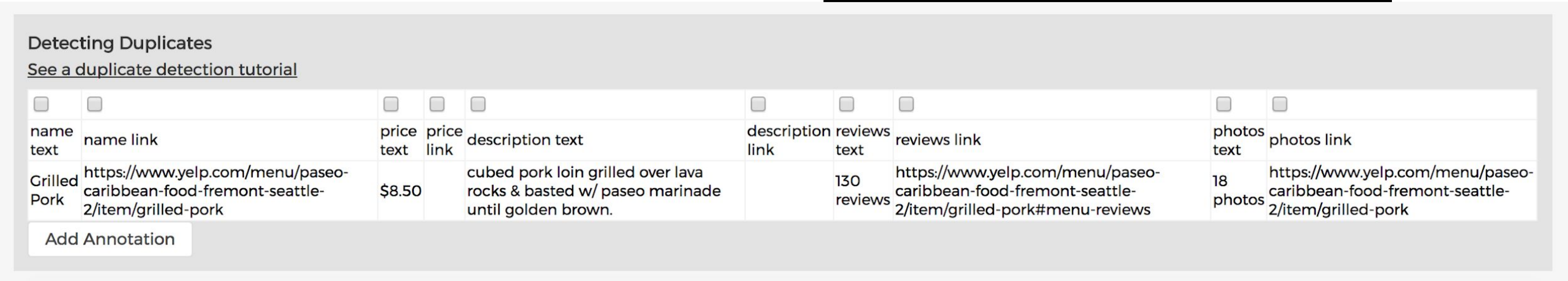
execution time if there's no failure

failure during high churn → see new data → slower recovery

overall, performance close to ideal!

User Study

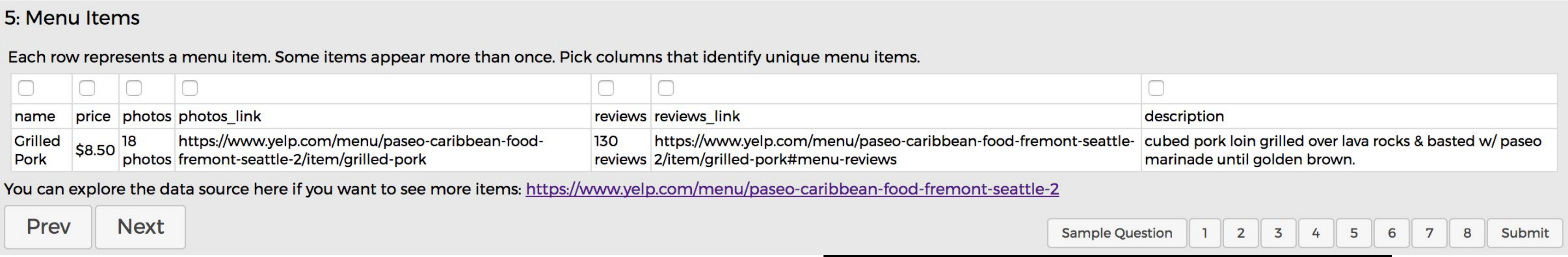
the UI in the Helena tool



Detecting Duplicates

See a duplicate detection tutorial

□	□	□	□	□	□	□	□	□	□
name text	name link	price text	price link	description text	description link	reviews text	reviews link	photos text	photos link
Grilled Pork	https://www.yelp.com/menu/paseo-caribbean-food-fremont-seattle-2/item/grilled-pork	$8.50		cubed pork loin grilled over lava rocks & basted w/ paseo marinade until golden brown.		130 reviews	https://www.yelp.com/menu/paseo-caribbean-food-fremont-seattle-2/item/grilled-pork#menu-reviews	18 photos	https://www.yelp.com/menu/paseo-caribbean-food-fremont-seattle-2/item/grilled-pork



5: Menu Items

Each row represents a menu item. Some items appear more than once. Pick columns that identify unique menu items.

□	□	□	□	□	□	□
name	price	photos	photos_link	reviews	reviews_link	description
Grilled Pork	$8.50	18 photos	https://www.yelp.com/menu/paseo-caribbean-food-fremont-seattle-2/item/grilled-pork	130 reviews	https://www.yelp.com/menu/paseo-caribbean-food-fremont-seattle-2/item/grilled-pork#menu-reviews	cubed pork loin grilled over lava rocks & basted w/ paseo marinade until golden brown.

You can explore the data source here if you want to see more items: https://www.yelp.com/menu/paseo-caribbean-food-fremont-seattle-2

the UI in the online survey

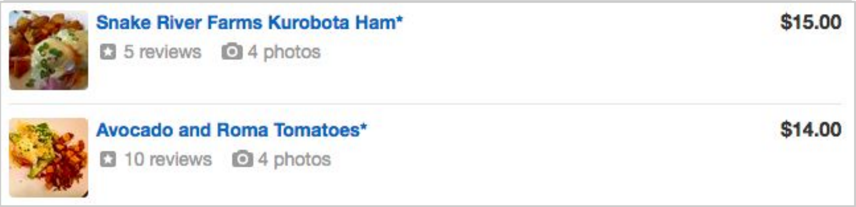
User Study

If a participant uses the Helena UI to add a skip block, doesn't adjust the default skip block parameters, how many rows of output data are wrong?

	coders	non-coders
% of rows **kept** that should have been skipped	0%	0%
% of rows **skipped** that should have been kept	1.3%	2.3%

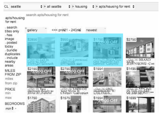
difference not statistically significant

time to write each skip block:
coders: 52 seconds
non-coders: 61 seconds



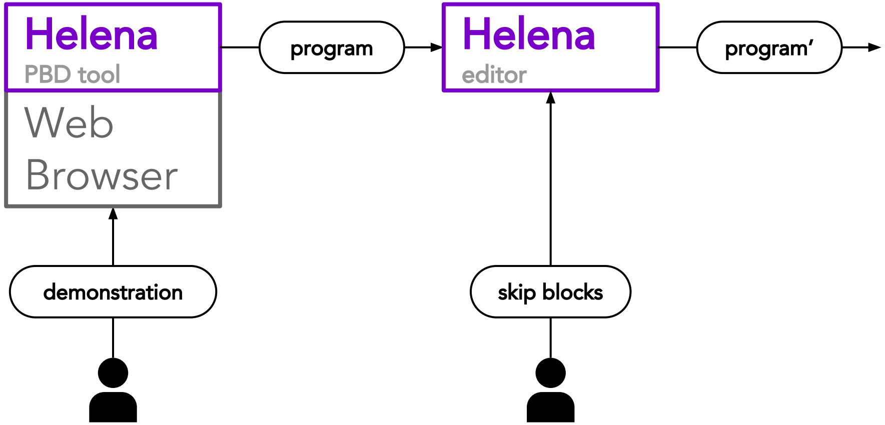
Snake River Farms Kurobota Ham* $15.00
5 reviews 4 photos

Avocado and Roma Tomatoes* $14.00
10 reviews 4 photos



Unified handling of three apparently disparate challenges with a **single language construct.**

By keeping reasoning at the level of target output data, made skip blocks **usable by non-programmers.**



helena-lang.org

contact: schasins@cs.berkeley.edu

github.com/schasins/helena