Introduction and EPIC Data Lab Vision

Speaking: Sarah E. Chasins, Aditya Parameswaran, Joe Hellerstein

Sarah Chasins

Faculty *Co-Director*

Joe Hellerstein

Faculty *Co-Director*

Aditya Parameswaran

Faculty *Co-Director*

Joseph Gonzalez

Faculty

Björn Hartmann

Faculty

Marti Hearst

Faculty

Anthony Joseph

Faculty

Michael Mahoney

Faculty

Niloufar Salehi

Faculty

Koushik Sen

Faculty

Dawn Song

Faculty

- Quick refresher on lab scope, mission
- Whirlwind tour through prior projects that led us to this lab's mission
  - 
  - 
  - 
- Summary of themes from projects, how they form lab's foundation, preview of today

- **Quick refresher on lab scope, mission**

- Whirlwind tour through prior projects that led us to this lab's mission
  - 
  - 
  - 
- Summary of themes from projects, how they form lab's foundation, preview of today

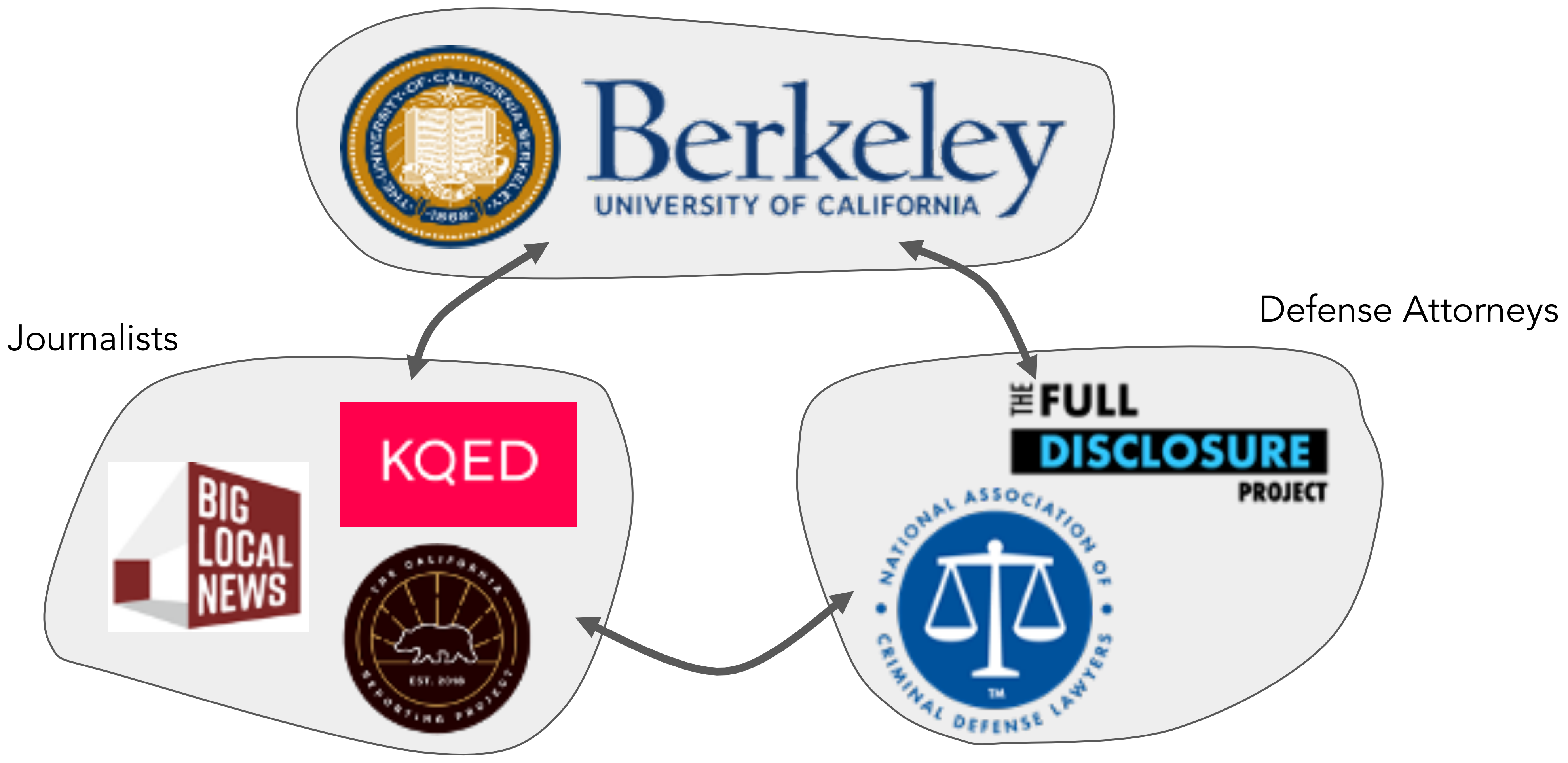
Huge thank you to our sponsors who make this work possible.

and !

Some familiar images…

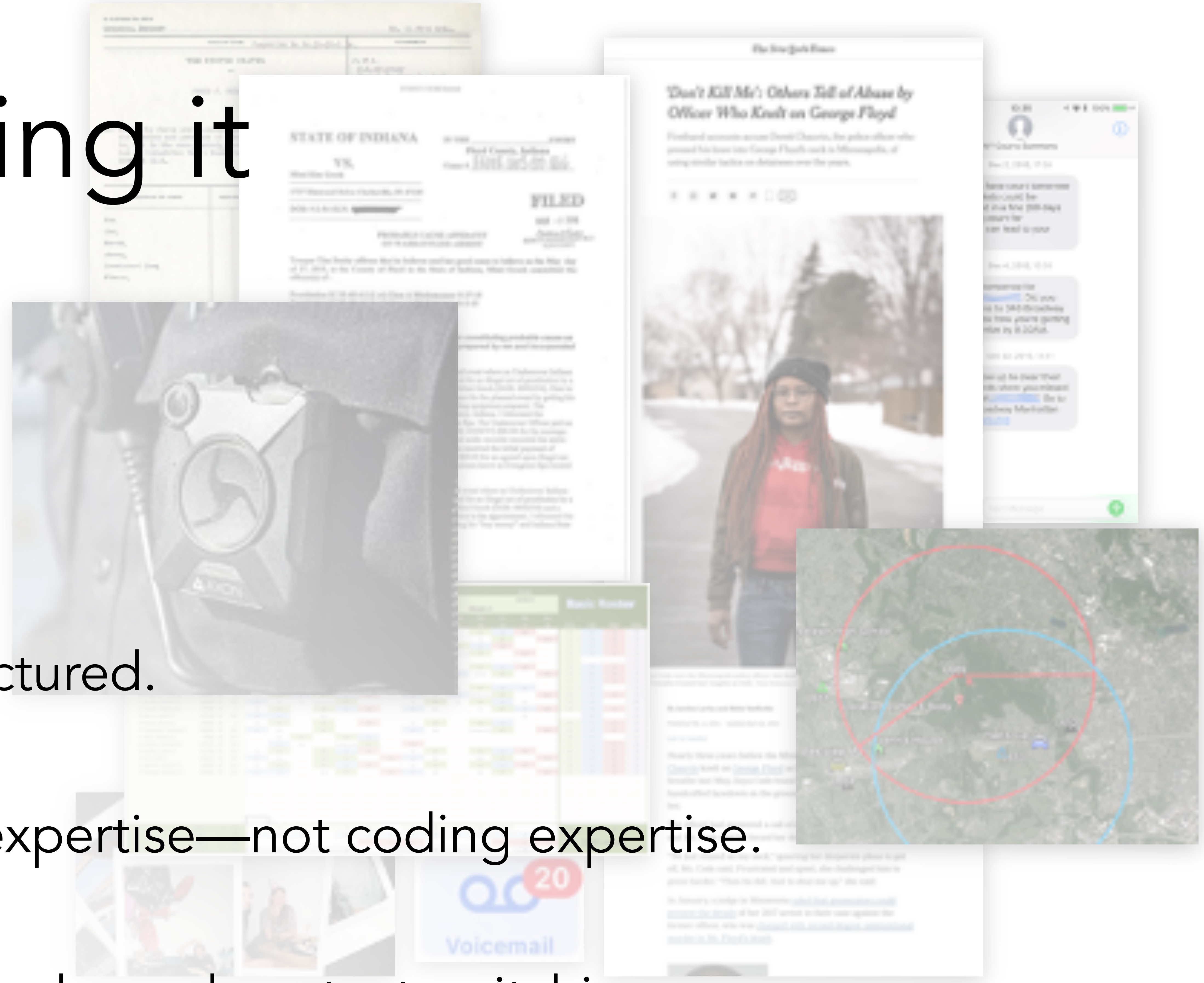
people who have data + ~~4 years data science training~~ learnable data science/AI tools! = evidence, insights, models

What we're making!



How we're making it!

Why existing tools aren't cutting it



- They demand **unrealistic data**.
  - Real data is often messy, poorly formatted, even unstructured.
- They demand **unrealistic expertise**.
  - Need a 4-year CS degree. Tools should need domain expertise—not coding expertise.
- They demand **unrealistic processes.**
  - Tools require lots of manual and mental effort, lines of code, and context-switching
- They demand **unrealistic teams.**
  - They assume one expert programmer, working alone—no team, organization, or diversity of roles.

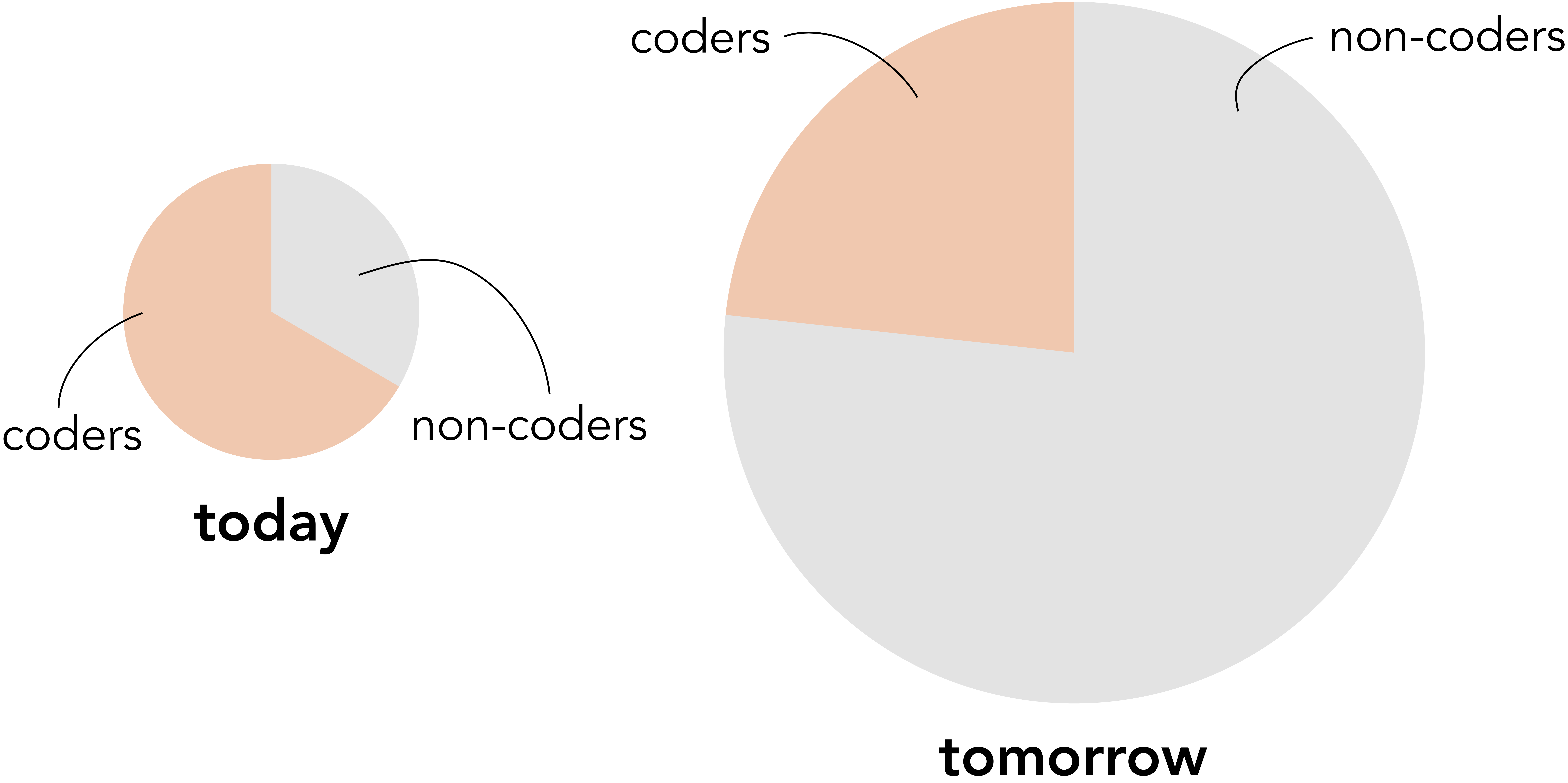
EPIC DATA lab
UC Berkeley

Mission: To develop no-code and low-code tools for data science/AI work shaped by the needs of heterogeneous teams.

EPIC Data Lab Intro Talk

- Quick refresher on lab scope, mission
- **Whirlwind tour through prior projects that led us to this lab's mission**
    - 
    - 
    - 
- Summary of themes from projects, how they form lab's foundation, preview of today

People who care about (web) data...

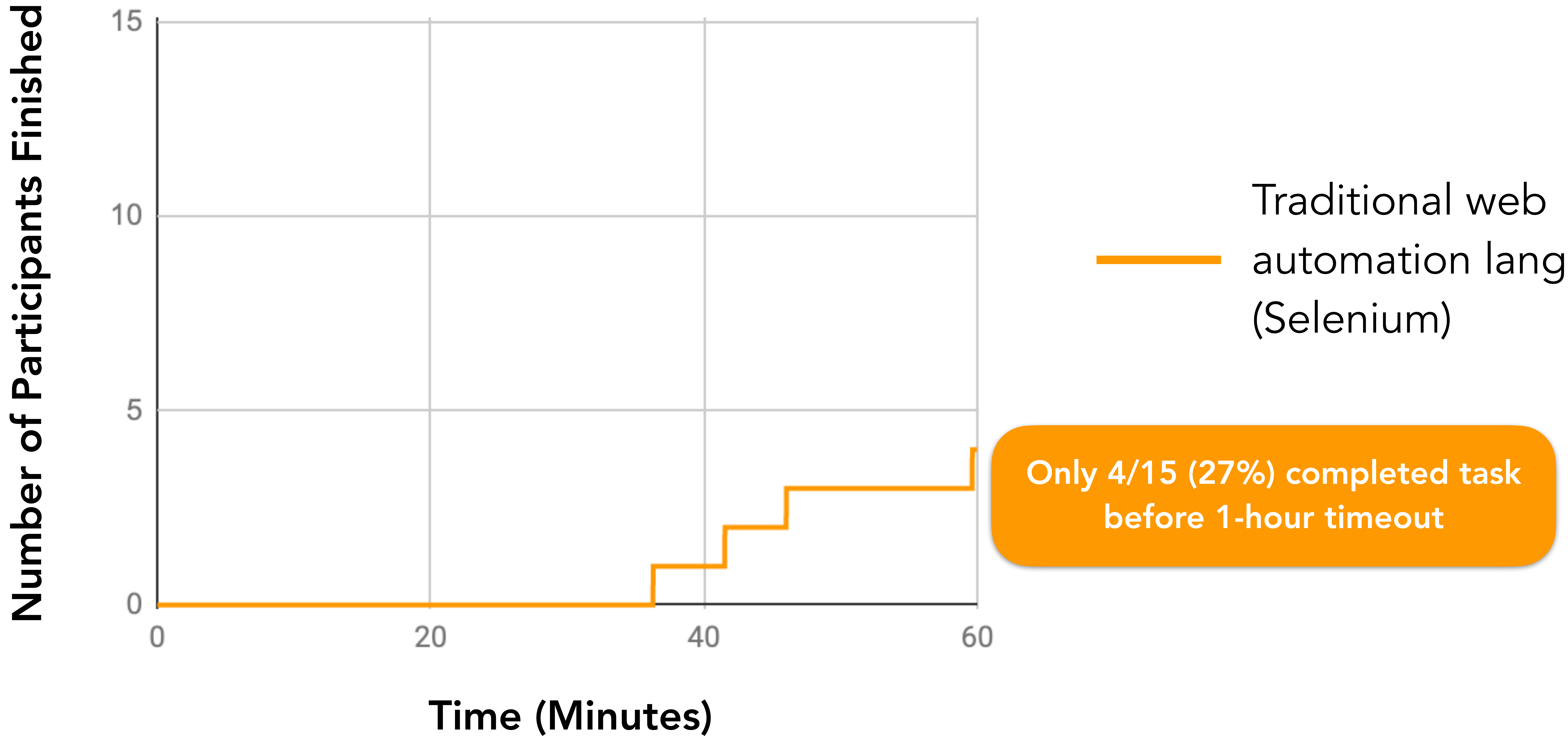


Web Data → Policy Action, Social Change

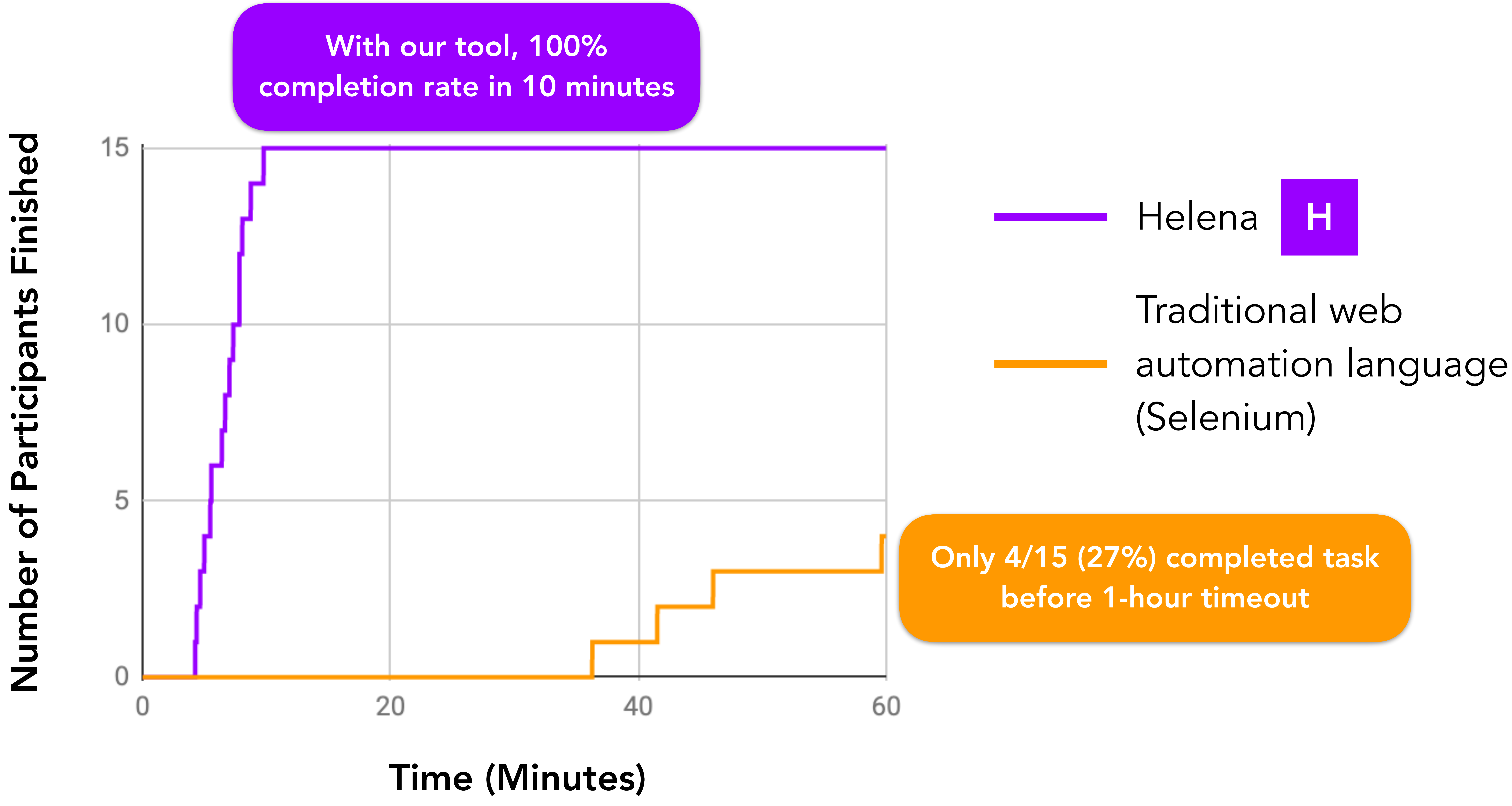
Or: How are our collaborators transforming society with web data right now?

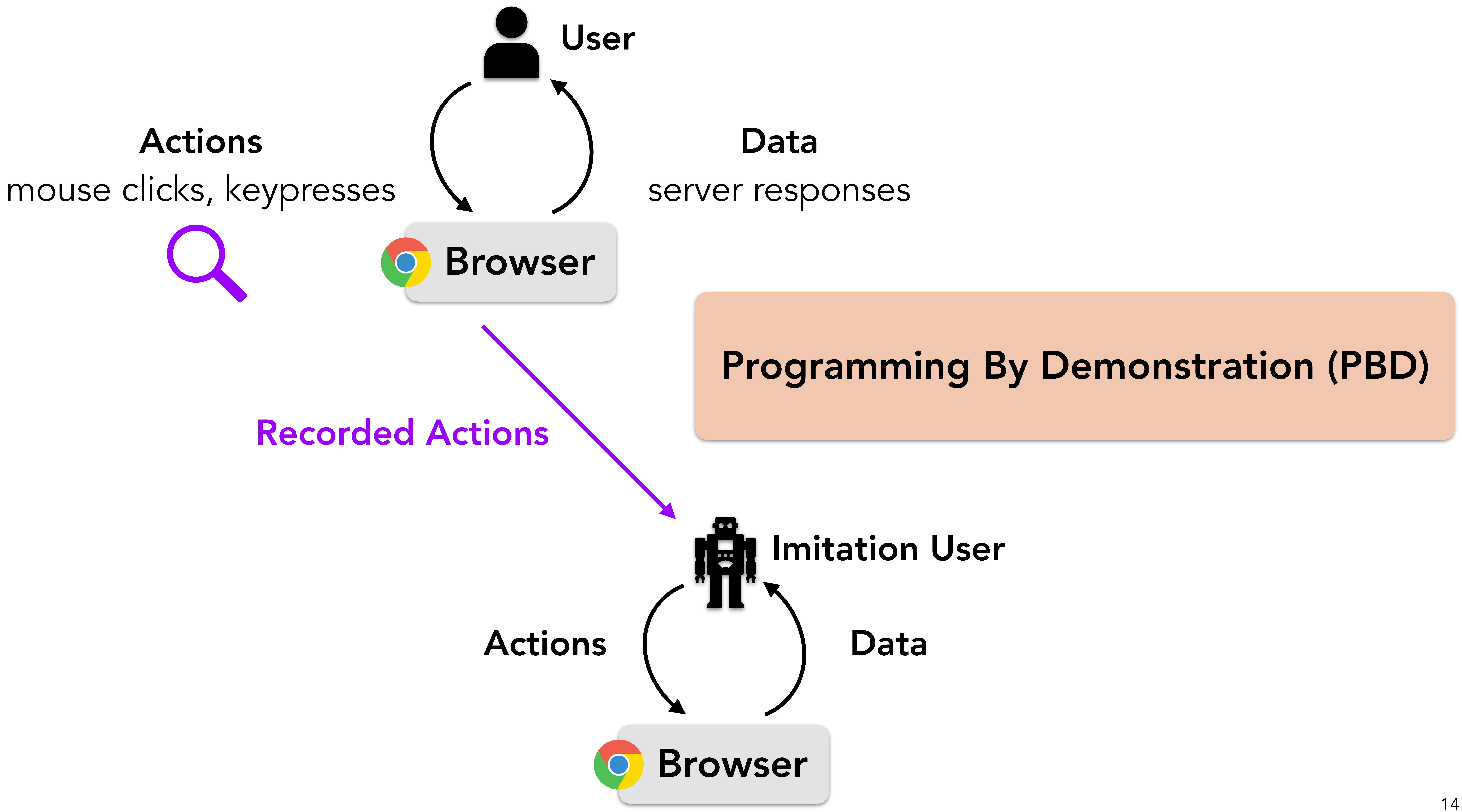
Sociology	helping low-income families move to high-opportunity neighborhoods	craigslist, Apartments.com
Nursing	reducing effects of perceived race on medical crowdfunding outcomes	YouCaring, gofundme
Political Science	increasing transparency of government agencies, governing bodies	ONPE, JNE, infogob
Civil Engineering	reducing effects of natural disasters on infrastructure and human mobility	The City of Houston Official Site for Houston, Texas

Web automation programming is hard.



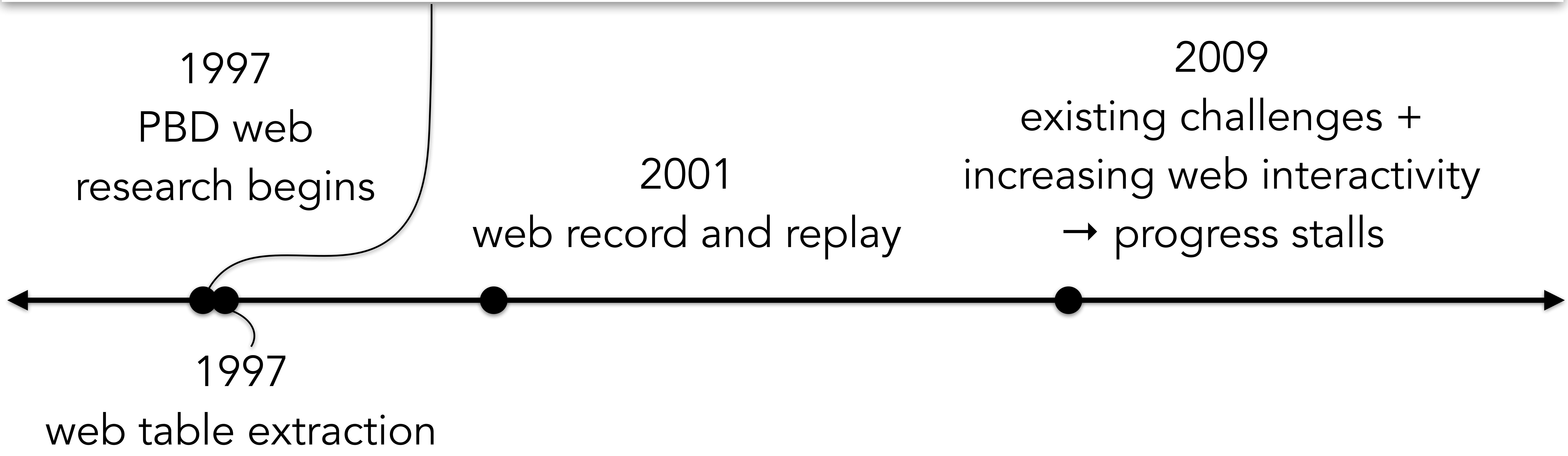
But we can make it easy.




User
Actions
mouse clicks, keypresses
Data
server responses
Browser
Programming By Demonstration (PBD)
Recorded Actions
Imitation User
Actions
Data
Browser

PBD web automation is a long-standing dream

An important design aspect is that Scrapbook is designed so that Web data can be copied directly from the most commonly used Web browsers—Netscape Navigator 3.0, Microsoft Internet Explorer 3.0, and their newer versions for Windows95/NT3.51—rather than forcing users to use a special



Demo!

What data should we collect to learn when CS researchers peak?


Profiles
scholar.google.com/citations?hl=en&view_op=search_authors&mauthors=label%3Acomputer_science&btnG=
Apps
Extensions
Helena Scraper and Automator
Has access to this site
Google Scholar
label:computer_science
SIGN IN
Profiles
My profile
My library
We open the browser extension.
Geoffrey Hinton
Emeritus Prof. Comp Sci, U.Toronto & Engineering Fellow, Google
Verified email at cs.toronto.edu
machine learning psychology artificial intelligence cognitive science computer science
Cited by 362360
David S. Johnson
Visiting Professor, Columbia University Computer Science Department
Verified email at research.att.com
Algorithms computer science optimization traveling salesman problem bin packing
Cited by 238863
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
Cited by 216544
Freeman Hu
Shandong University
Verified email at mail.sdu.edu.cn
Computer Science
Cited by 215347
vapnik
Professor of Columbia, Fellow of NEC Labs America,
Verified email at nec-labs.com
machine learning statistics computer science
Cited by 212828

We demonstrate how to find information that goes in the first row of our target dataset.

We continue on another page (and another table).


chrome-extension://bcajeffgcbmkndhonbkfmahpckenfoib/pages/mainpanel.html?start...
Current Script
Saved Scripts
Scheduled Runs
We're recording! Remember, collect ONLY the FIRST ROW of data. When you're ready to add a new cell, hover over the text you want, then press ALT + click. We'll show the data you've collected right here:
Cancel Recording
Geoffrey Hinton
Imagenet classification with deep convolutional neural networks
66206
2012
Stop Recording
We're done demonstrating.
Profiles
Geoffrey Hinton - Google Scho
scholar.google.com/citations?hl=en&user=JicYPdAAAAAJ
Apps
Extensions
Google Scholar
Geoffrey Hinton
FOLLOW
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
ARTICLES
CITED BY
CO-AUTHORS
TITLE
CITED BY
YEAR
Imagenet classification with deep convolutional neural networks
A Krizhevsky, I Sutskever, GE Hinton
Advances in neural information processing systems, 1097-1105
66206
2012
Deep learning
Y LeCun, Y Bengio, G Hinton
Nature 521 (7553), 436-444
28294
2015
Learning internal representations by error-propagation
DE Rumelhart, GE Hinton, RJ Williams
Parallel Distributed Processing: Explorations in the Microstructure of …
26773
1986
Learning internal representations by error propagation
DE Rumelhart, GE Hinton, RJ Wlliams
Learning internal representations by error propagation
26468
1986
Learning internal representations by error propagation
DE Rumelhart, GE Hinton, RJ Williams
MIT Press, Cambridge, MA 1 (318)
26422
1986


chrome-extension://bcajeffgcbmkndhonbkfmahpckenfoib/pages/mainpanel.html?start...
Current Script
Saved Scripts
Scheduled Runs
Save and Run Script
program_name
Save Script
Advanced Options
Start New Script
text
numbers
other
load https://scholar.google.com/citations?hl=en&view_... into p
for each row in list_1 in page1 (✓ for all rows, for the first 2
do
scrape list_1_item_1 in page1
click list_1_item_1 in page1 , load page into page2
for each row in list_3 in page2 (✓ for all rows, for the f
do
scrape title in page2
scrape cited_by in page2
scrape year in page2
add dataset row that includes: list_1_item_1 TE
Relevant Tables
Troubleshooting
What kind of problem are you having?
New Recording Window
Profiles
Geoffrey Hinton - Google Sc
scholar.google.com/citations?hl=en&user=JicYPdAAAAAJ
Apps
Extensions
Google Scholar
Geoffrey Hinton
FOLLOW
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
ARTICLES
CITED BY
CO-AUTHORS
TITLE
CITED BY
YEAR
Imagenet classification with deep convolutional neural networks
A Krizhevsky, I Sutskever, GE Hinton
Advances in neural information processing systems, 1097-1105
66206
2012
Deep learning
28294
2015
earning internal representations by error-propagation
DE Rumelhart, GE Hinton, RJ Williams
Parallel Distributed Processing: Explorations in the Microstructure of …
26773
1986
Learning internal representations by error propagation
DE Rumelhart, GE Hinton, RJ Wlliams
Learning internal representations by error propagation
26468
1986
Learning internal representations by error propagation
DE Rumelhart, GE Hinton, RJ Williams
MIT Press, Cambridge, MA 1 (318)
26422
1986
The synthesizer writes our program.

chrome-extension://bcajeffgcbmkndhonbkfmahpckenfoib/pages/mainpanel.html?start...
Current Script
Saved Scripts
Scheduled Runs
Script Run 1
Pause Script
Resume Script
Restart From Beginning
Cancel Script Run
Download Data (This Scrape)
Download Data (All Scrapes)
Note: the downloaded dataset may be slightly out of date if we haven't saved all data yet.
Rows so far: 40
Geoffrey Hinton Imagenet classification with deep convolutional neural networks 66206 2012 1
Geoffrey Hinton Deep learning 28294 2015 2
Geoffrey Hinton Learning internal representations by error-propagation 26773 1986 3
Geoffrey Hinton Learning internal representations by error propagation 26468 1986 4
Geoffrey Hinton Learning internal representations by error propagation 26422 1986 5
Geoffrey Hinton Learning representations by back-propagating errors 21859 1986 6
Geoffrey Hinton Dropout: a simple way to prevent neural networks from overfitting 21365 2014 7
Geoffrey Hinton Visualizing data using t-SNE 14439 2008 8
Geoffrey Hinton A fast learning algorithm for deep belief nets 13397 2006 9
Geoffrey Hinton Reducing the dimensionality of data with neural networks 12573 2006 10
Geoffrey Hinton Rectified linear units improve restricted boltzmann machines 10069 2010 11
Geoffrey Hinton Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups 8233 2012 12
Geoffrey Hinton Learning multiple layers of features from tiny images 8034 2009 13
Geoffrey Hinton Speech recognition with deep recurrent neural networks 5975 2013 14
Geoffrey Hinton Improving neural networks by preventing co-adaptation of feature detectors 5308 2012 15
Geoffrey Hinton Training products of experts by minimizing contrastive divergence 4574 2002 16
Geoffrey Hinton Adaptive mixtures of local experts 4263 1991 17
Geoffrey Hinton A learning algorithm for Boltzmann machines
Geoffrey Hinton Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude 3924 2012 19
Geoffrey Hinton Distilling the knowledge in a neural network 3887 2015 20
New Recording Window
Profiles
David Haussler - Google Sch
scholar.google.com/citations?hl=en&user=S2OjOvYAAAAJ
Apps
Extensions
This page is being controlled by Helena.
If you want to interact with this page anyway, click here to remove the overlay. Keep in mind that navigating away from the current page may disrupt the Helena process.
David Haussler
FOLLOW
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
ARTICLES
CITED BY
TITLE
CITED BY
YEAR
Initial sequencing and analysis of the human genome
ES Lander, LM Linton, B Birren, C Nusbaum, MC Zody, J Baldwin, ...
Macmillan Publishers Ltd.
19484
2001
An integrated encyclopedia of DNA elements in the human genome
ENCODE Project Consortium
Nature 489 (7414), 57-74
10539 *
2012
The human genome browser at UCSC
WJ Kent, CW Sugnet, TS Furey, KM Roskin, TH Pringle, AM Zahler, ...
Genome research 12 (6), 996-1006
8020
2002
Initial sequencing and comparative analysis of the mouse genome
RH Waterston, K Lindblad-Toh, E Birney, J Rogers, JF Abril, P Agarwal, ...
7069
2002
A map of human genome variation from population-scale sequencing
1000 Genomes Project Consortium
Nature 467 (7319), 1061
7053
2010
The program collects our data.

Extensions ×   Profiles ×   +

https://scholar.google.com/citations?hl=en&view_op=search_authors&mauthors=label%3Acomputer_science&btnG=

Apps   Extensions   Other Bookmarks

Google Scholar   label:computer_science

Profiles   My profile   My library

Geoffrey Hinton — Cited by 246012
Emeritus Prof. Comp Sci, U.Toronto & Engineering Fellow, Google
Verified email at cs.toronto.edu
machine learning   neural networks   artificial intelligence   cognitive science   computer science

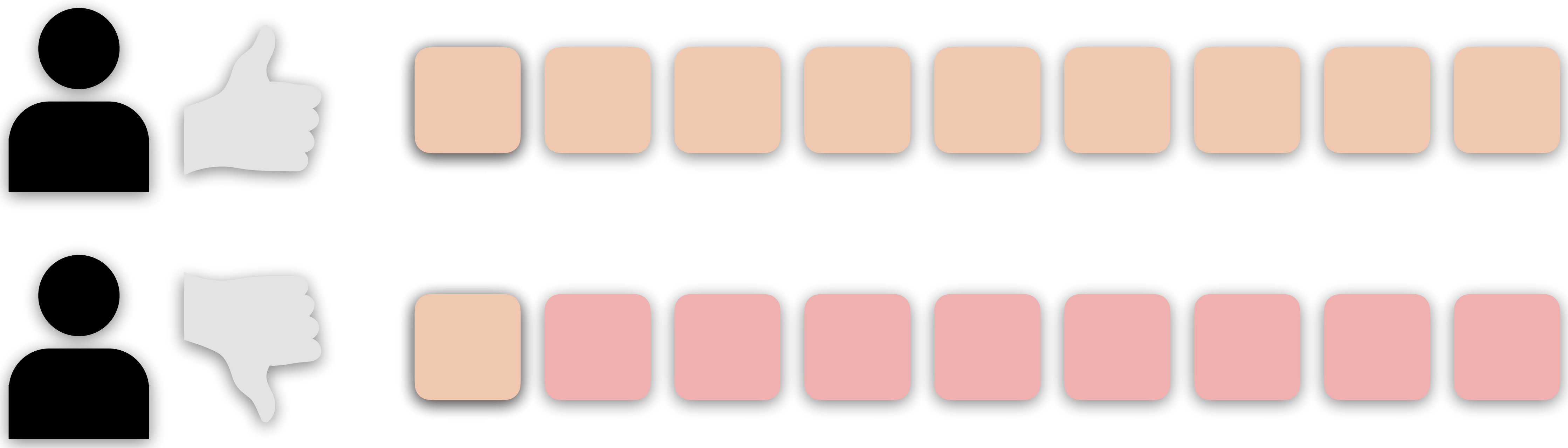
DEYWIS MORENO — Cited by 206009
High Energy Physicist, Universidad Antonio Narino
Verified email at uan.edu.co
High Energy Physics   Computer science

David S. Johnson — Cited by 176731
Visiting Professor, Columbia University Computer Science Department
Verified email at research.att.com
Algorithms   computer science   optimization   traveling salesman problem   bin packing

David Haussler — Cited by 174202
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

vapnik — Cited by 170728

End User Program Synthesis



actual reactions

over the years, learned what users *don’t* like, but not much about what they *do* like

this talk: what users do like

This design was a reaction to prior synthesis issues

Similarly, SMARTedit's users complained that they **wanted to be able to directly modify the generated hypotheses (e.g., "set the font size to 12") without having to retrain the system with additional examples.**

Tessa Lau, 2008

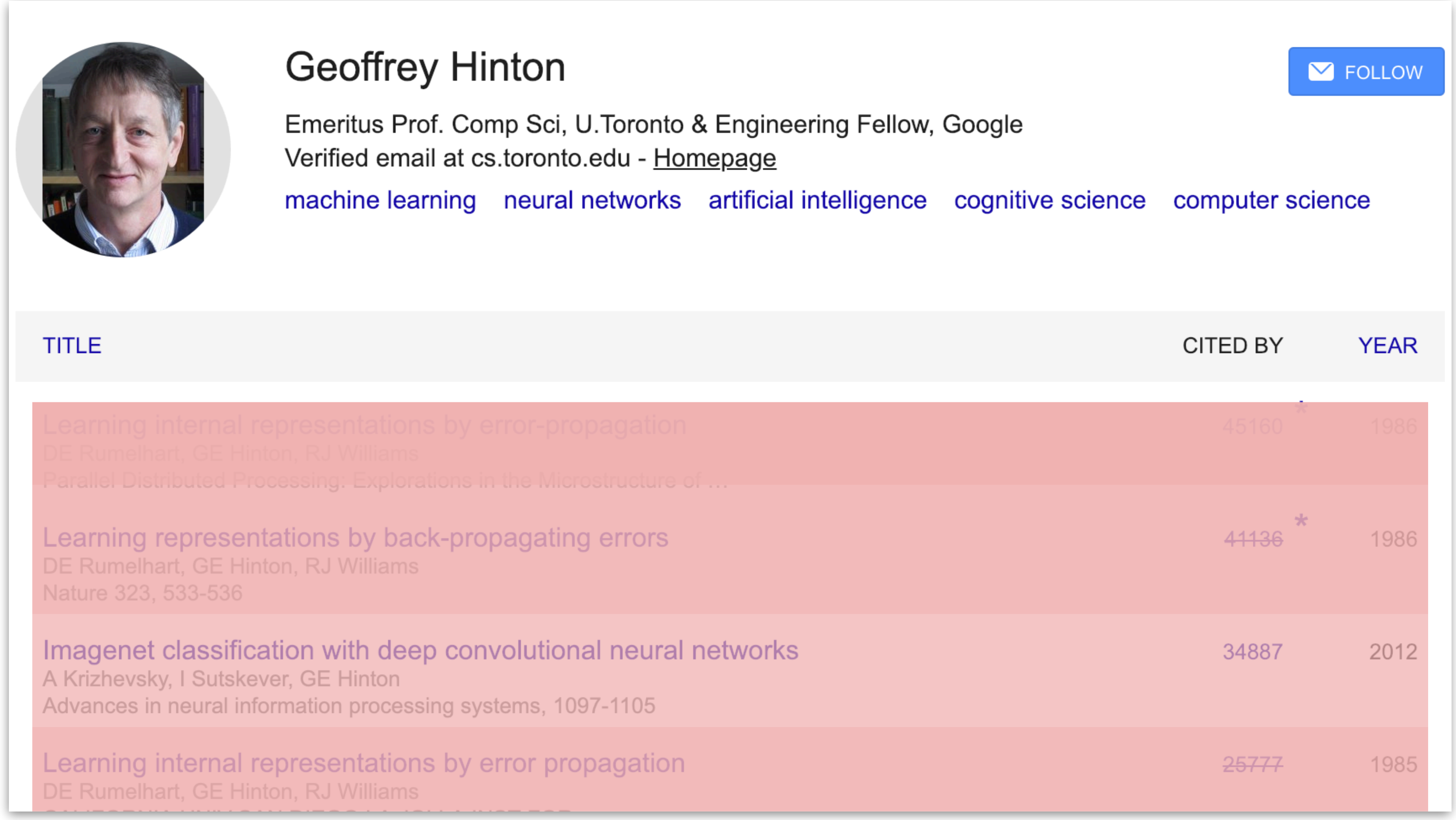
Several users found the sequence of steps needed to construct a mashup **overly complicated.** One user stated "If I had been given the tool without any instruction, I could not have figured out how to use it. It needs to be more 'discoverable.'" Another user said that it was **confusing to use one technique to create the initial table, and another technique to add information to a new column.**

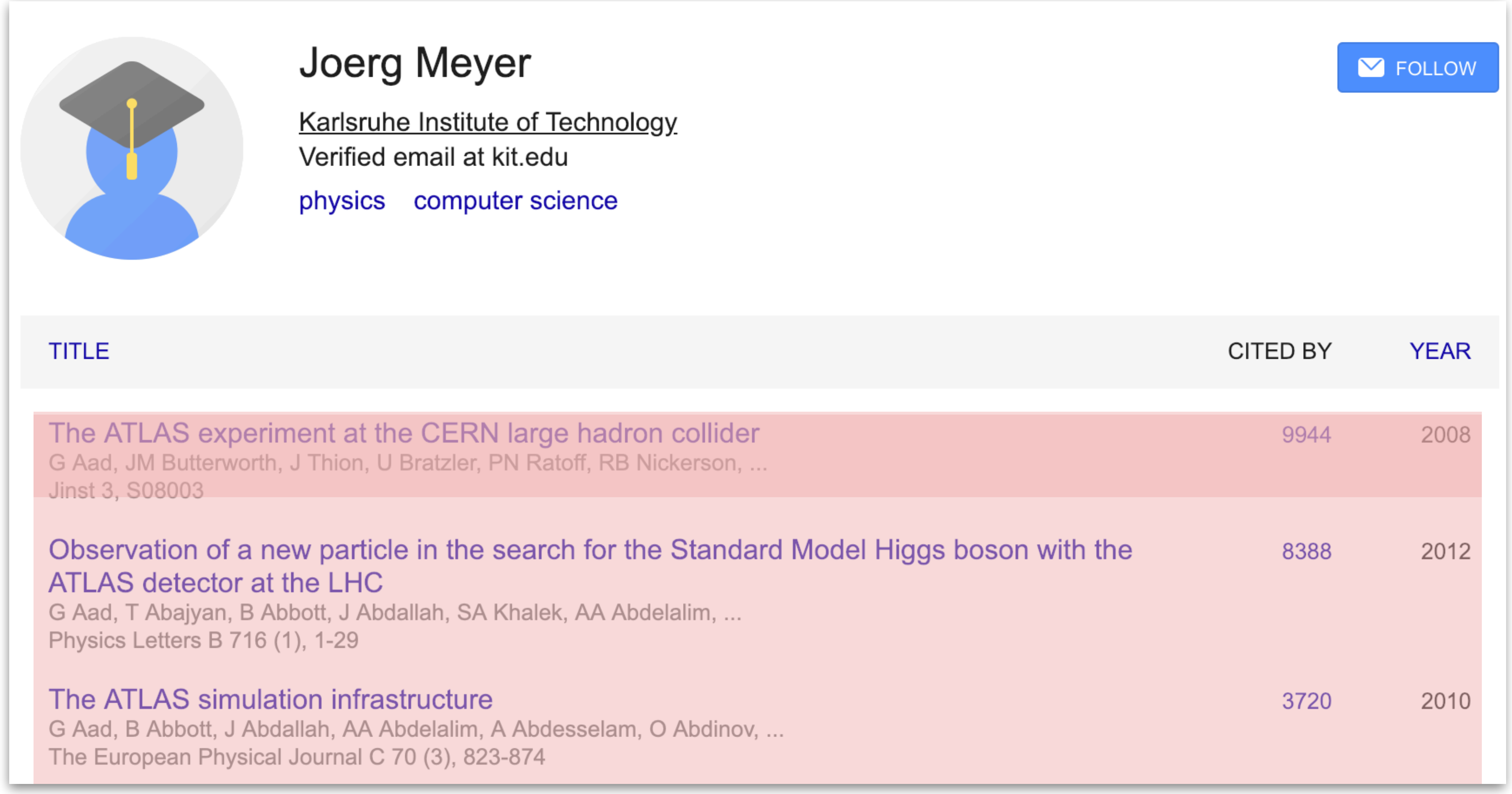
End-User Programming of Mashups with Vegemite, 2009

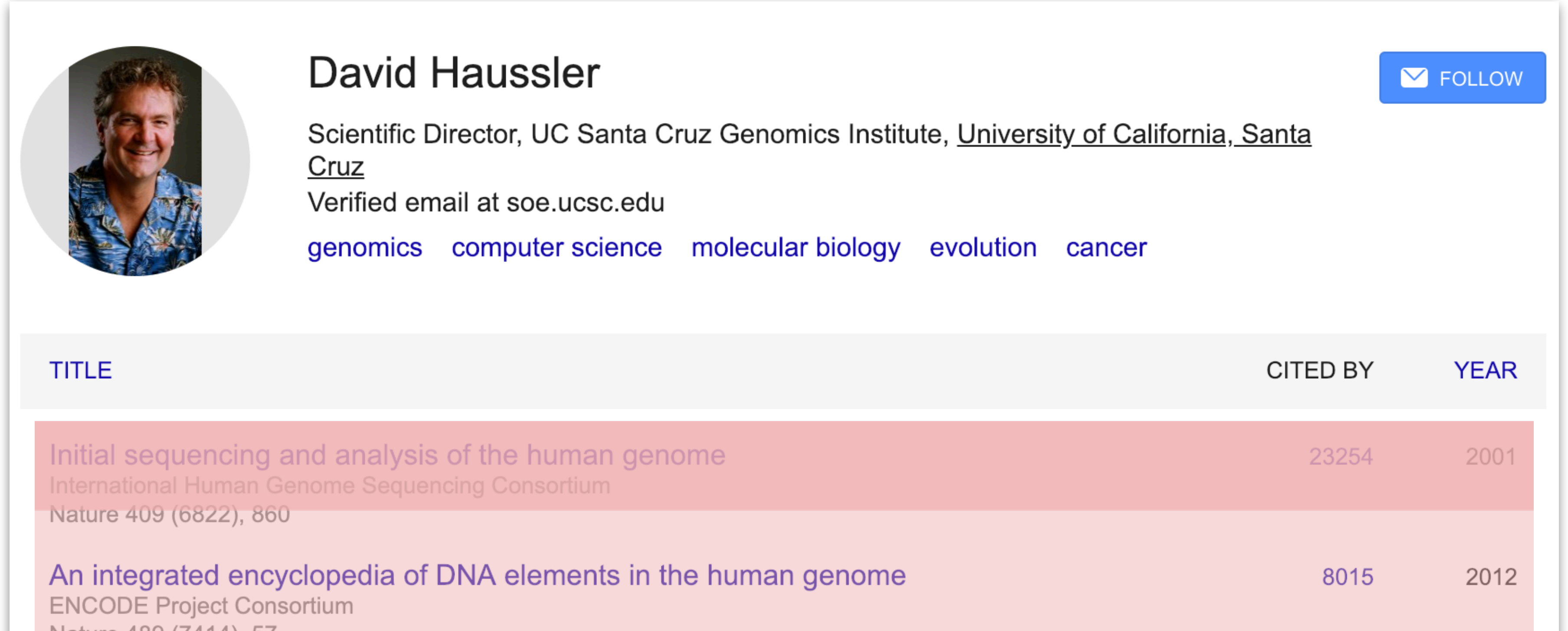
Excel Flash Fill

could we do PBD with one demo?

-our research team







# of course, single-demo is crazy…

synthesis person's first instinct is to discard this idea immediately

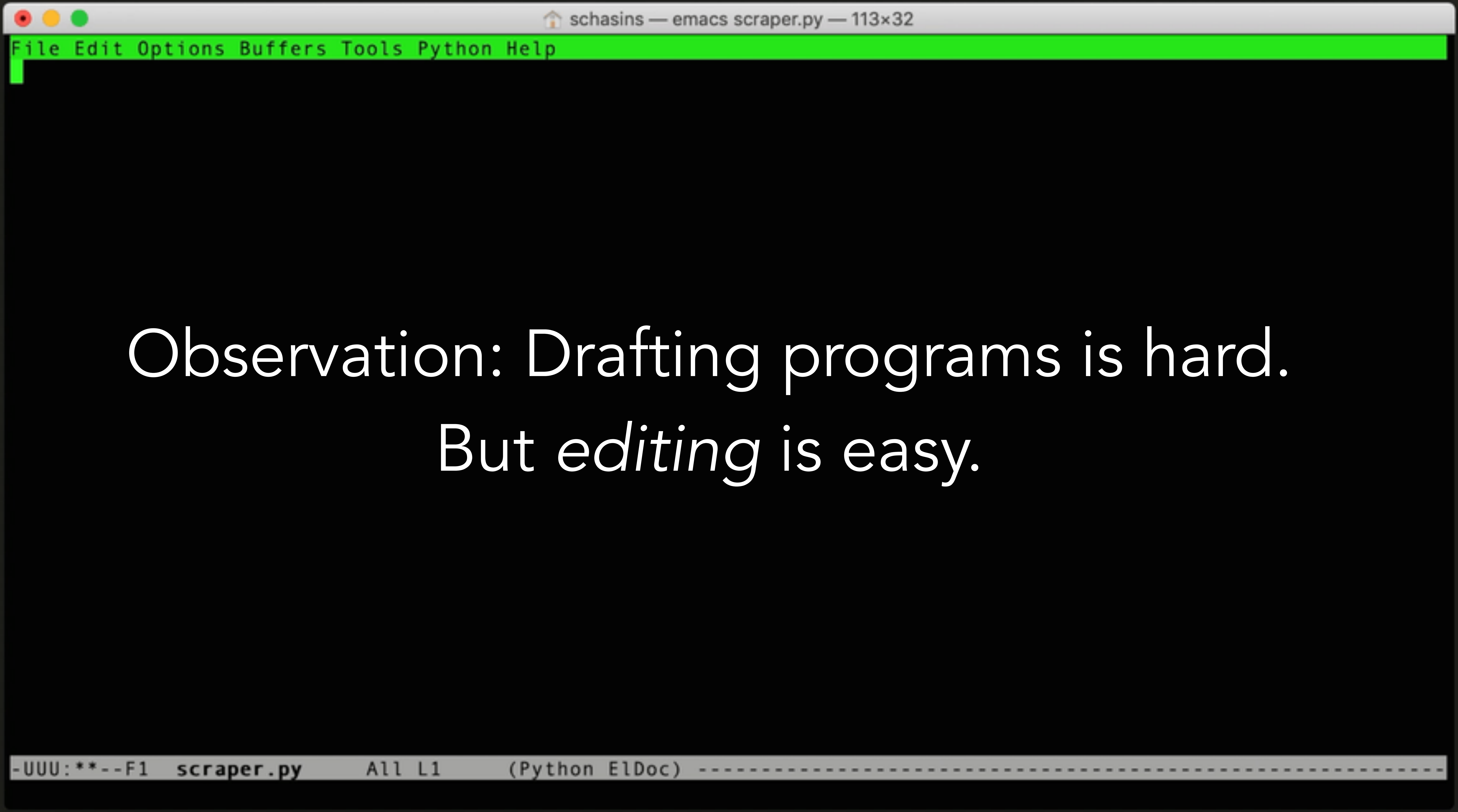
…because one demo is very ambiguous

first paper by each author?

all papers by all authors?

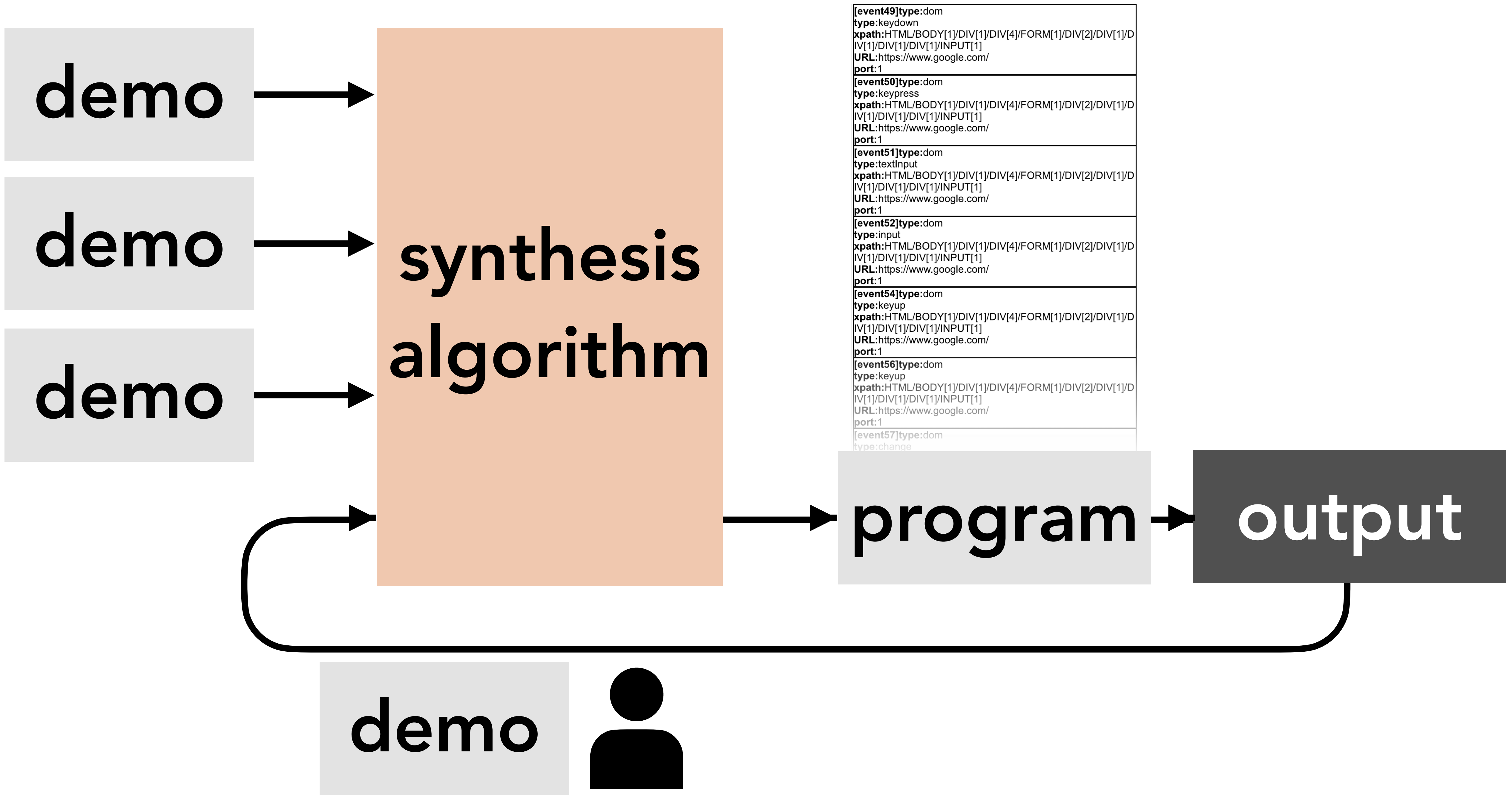
all papers with more than x citations?

all papers that mention a given word in the title?

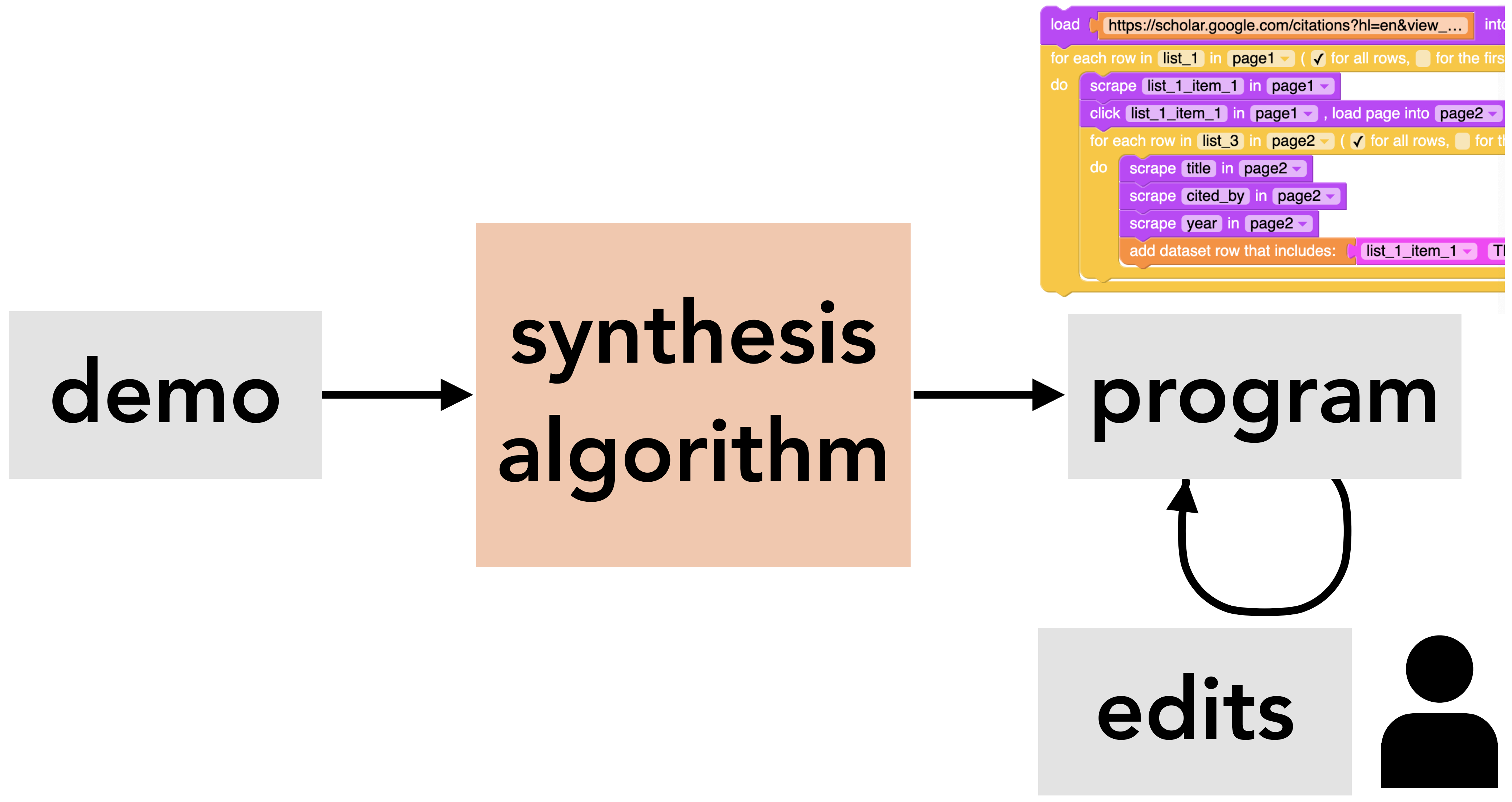


Observation: Drafting programs is hard.

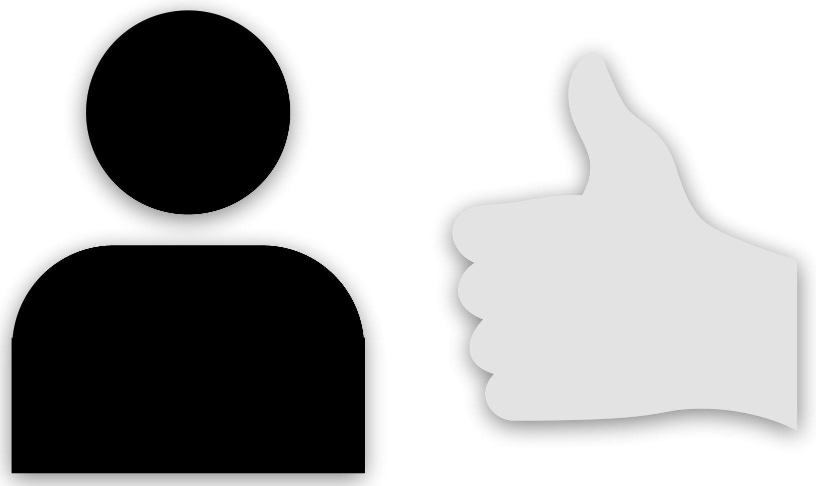
But *editing* is easy.



Traditional PBD

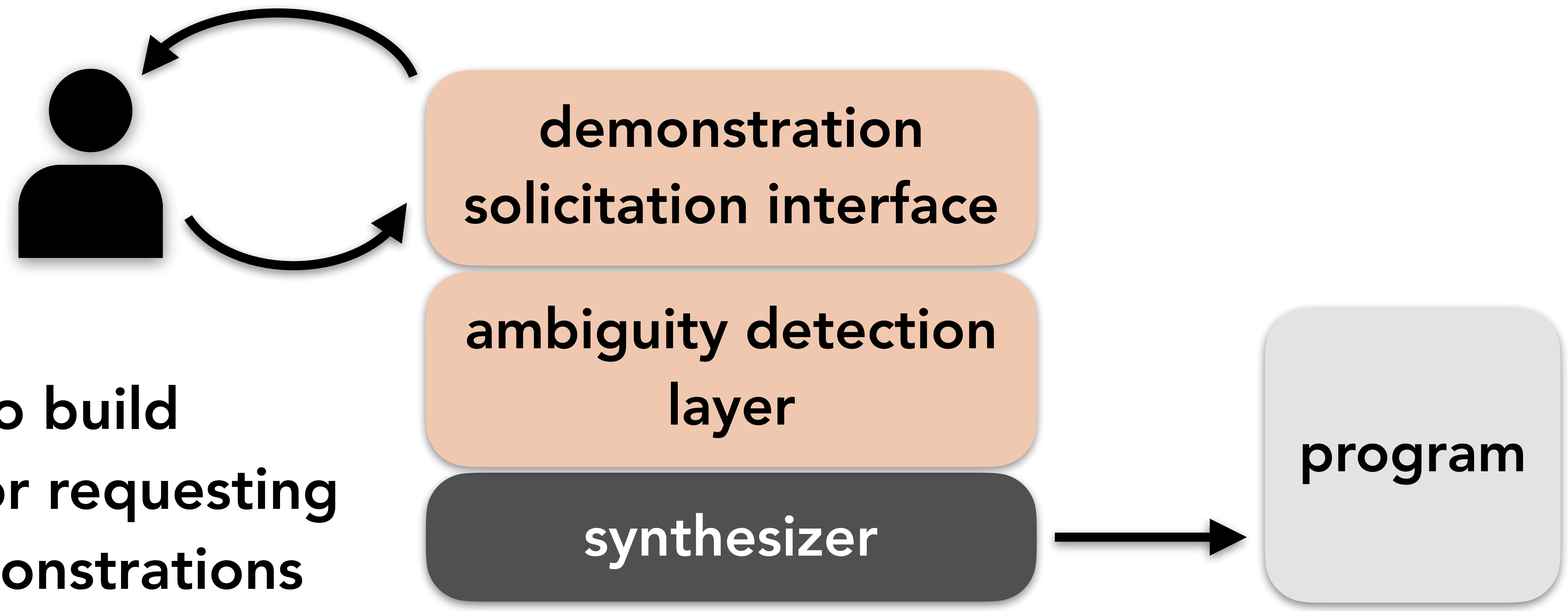


With learnable languages

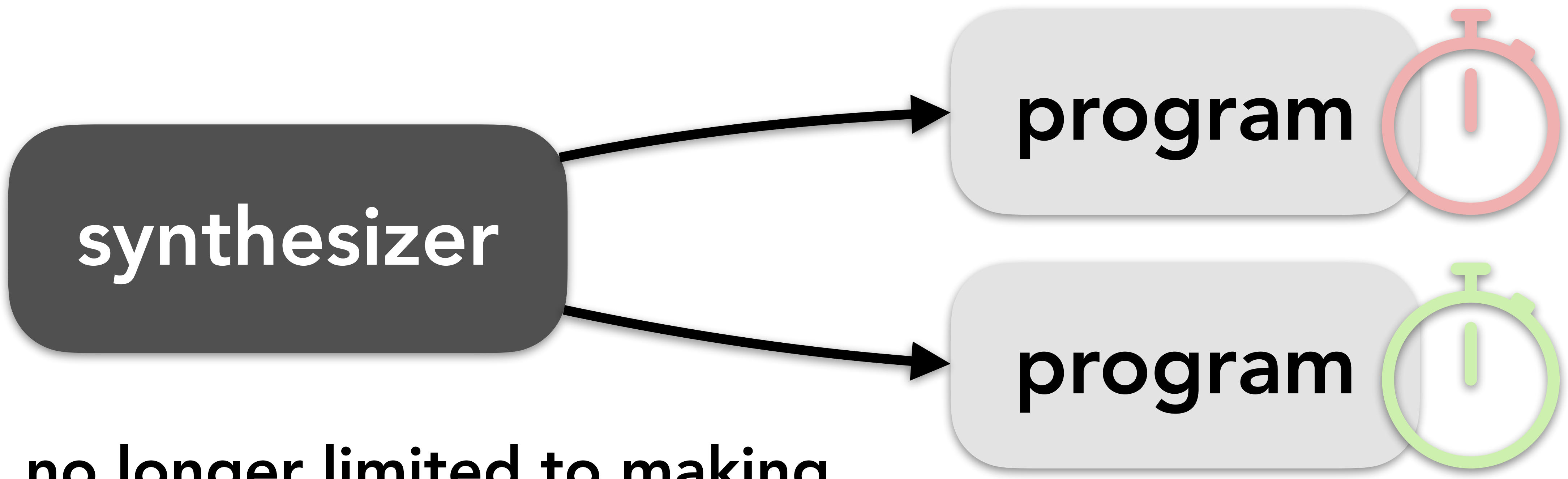


happy, successful users

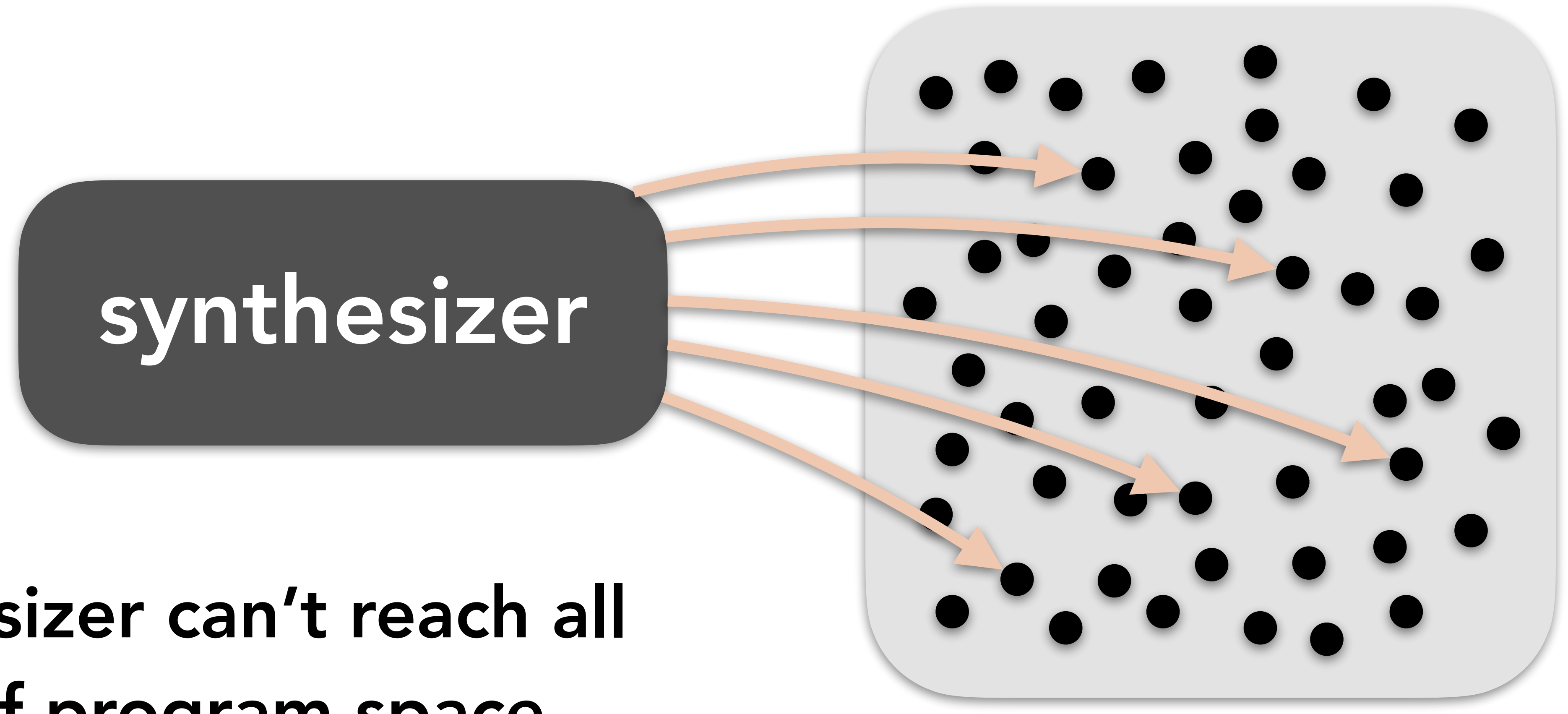
no need to build infrastructure for requesting particular demonstrations



Why this mixed-modality (demo + program edits) input was so much more successful



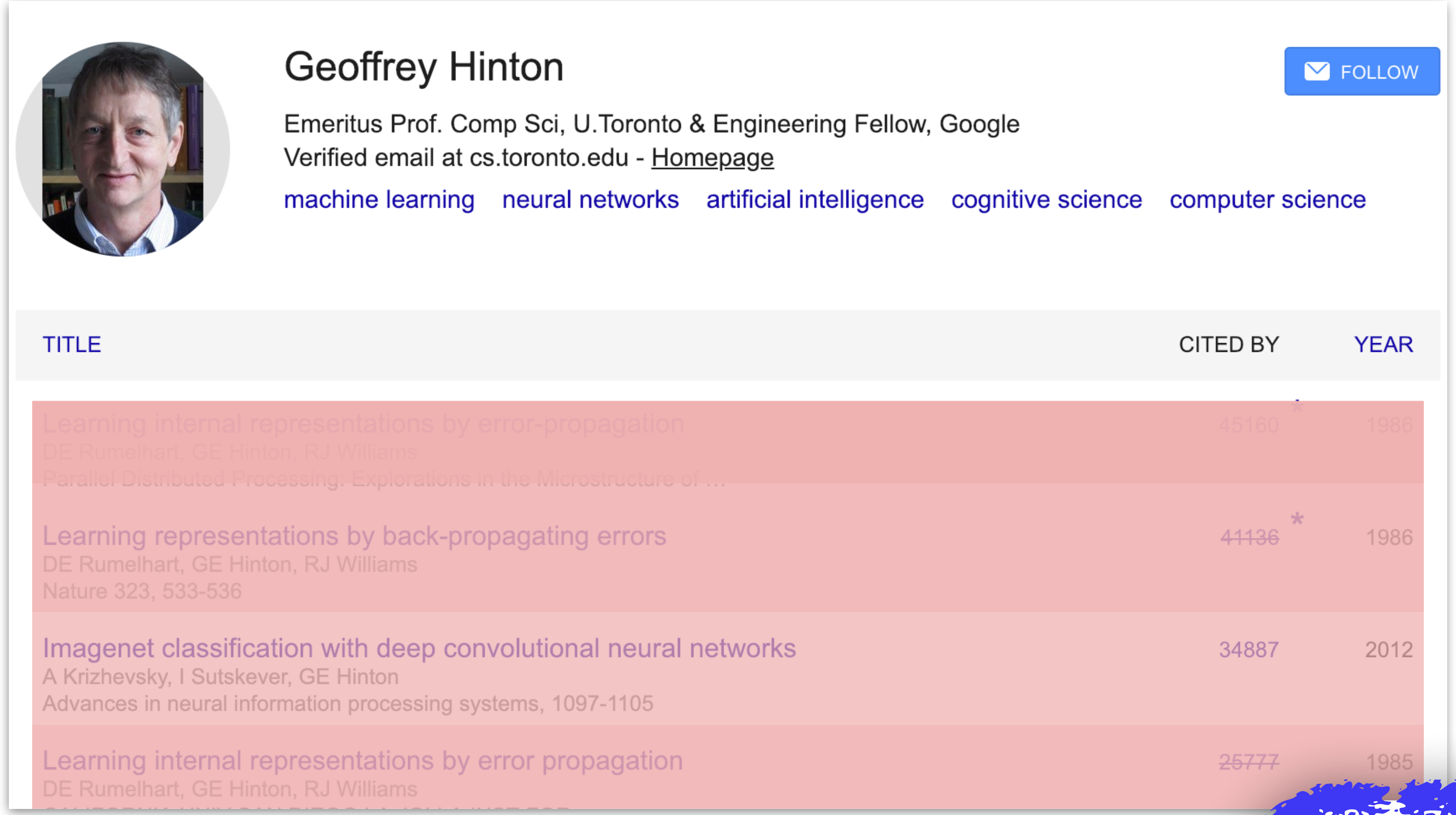
no longer limited to making programs that can run in interactive time

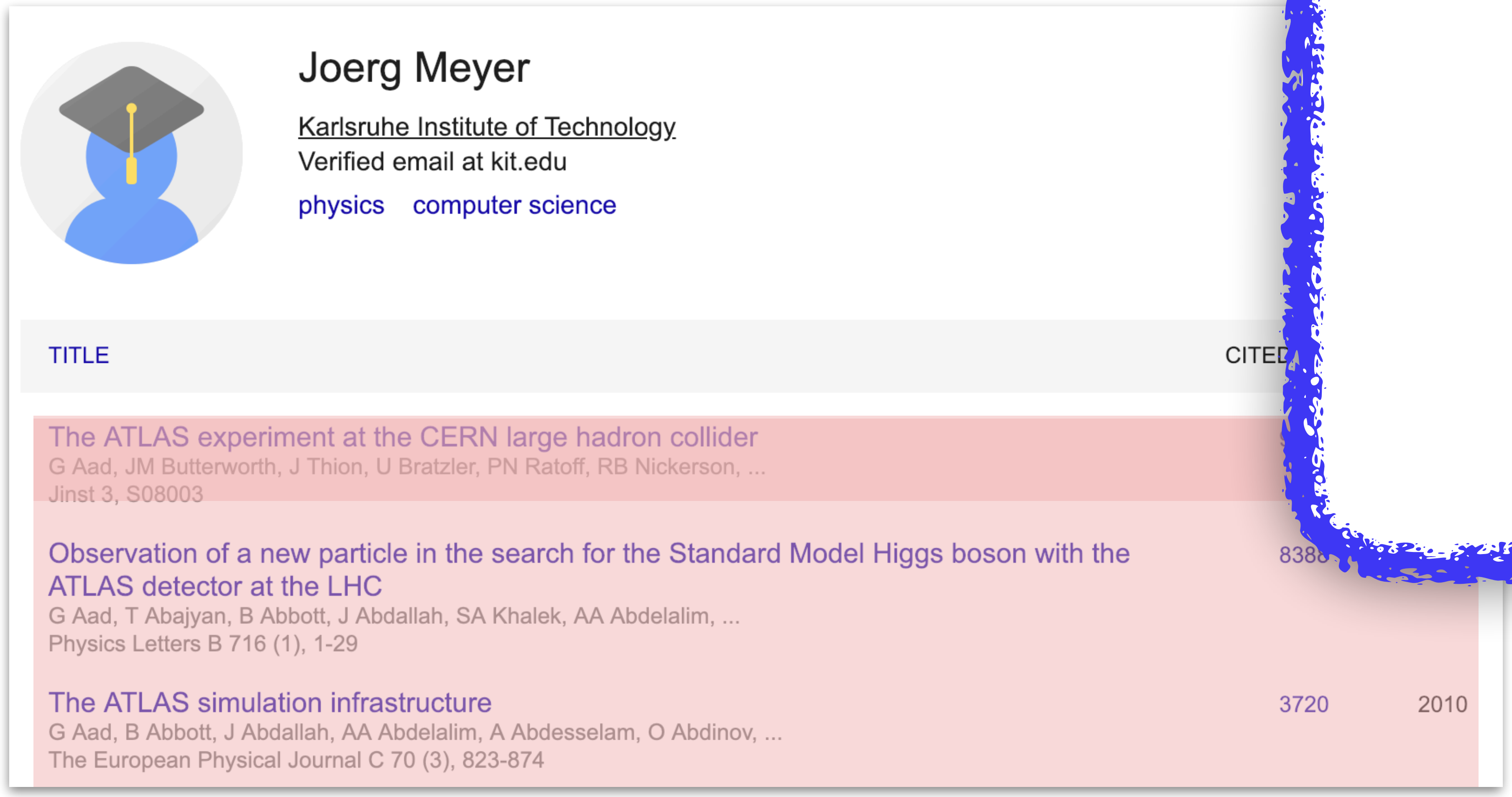


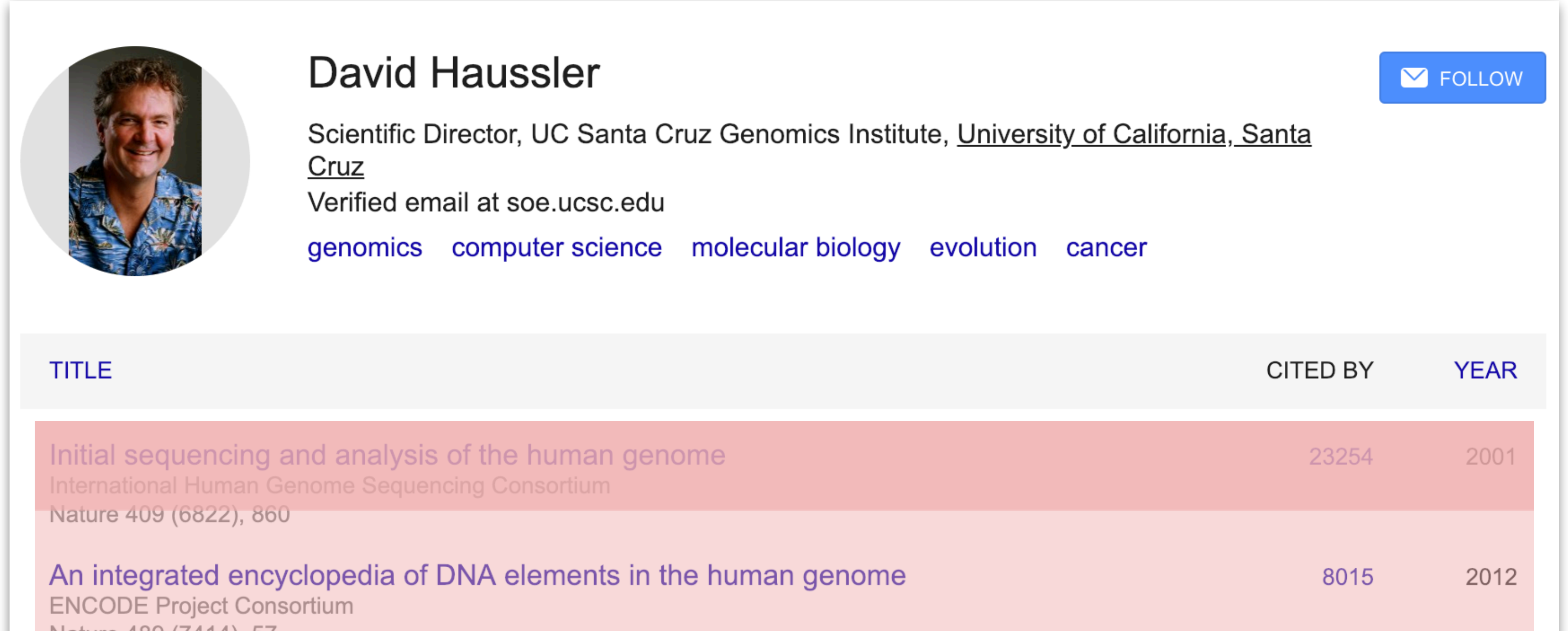
ok if synthesizer can't reach all corners of program space

allowable programs

Does this design exhibit those key themes?







# of course, single-demo is crazy...

synthesis person's first instinct is to discard this idea immediately

no is very ambiguous

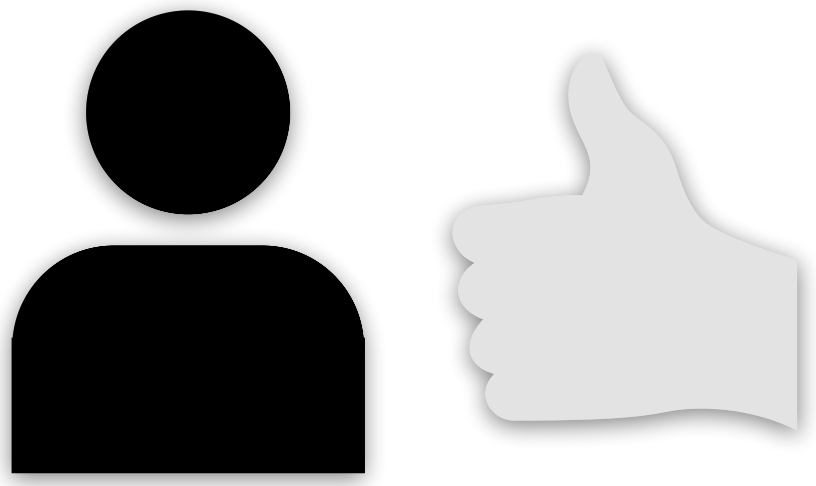
Imprecision tolerated

y each author?

all papers by all authors?

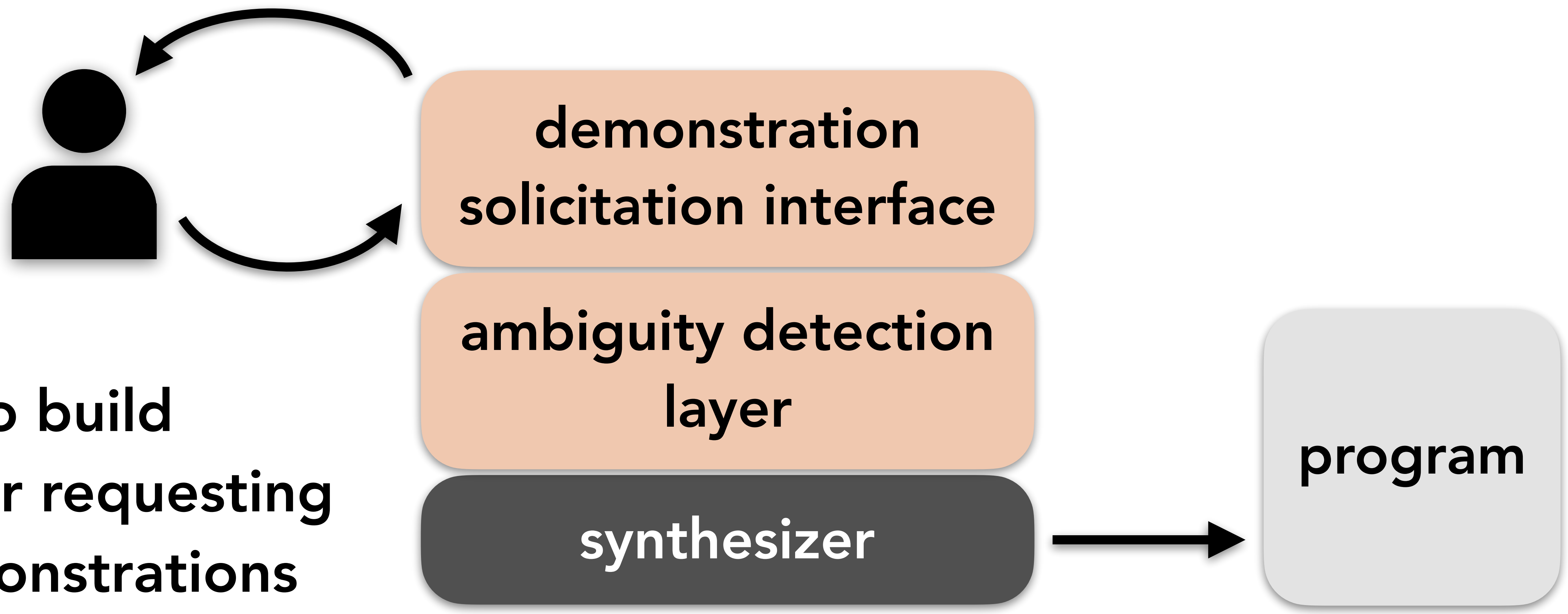
all papers with more than x citations?

all papers that mention a given word in the title?

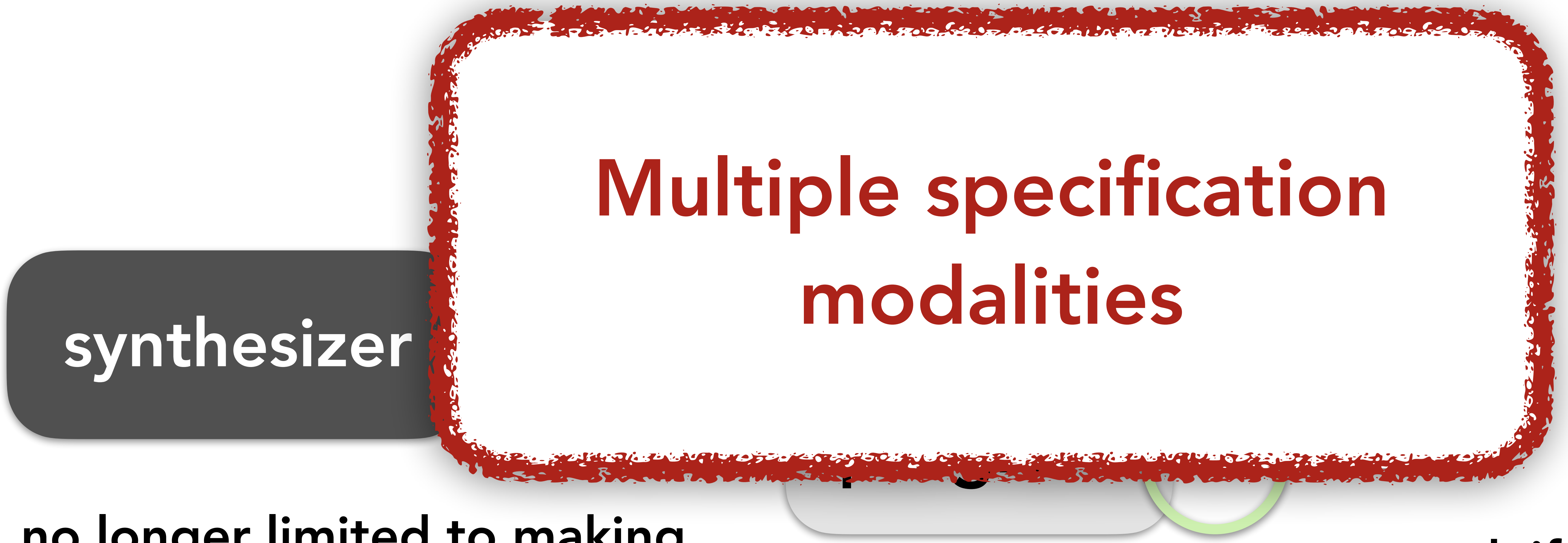


happy, successful users

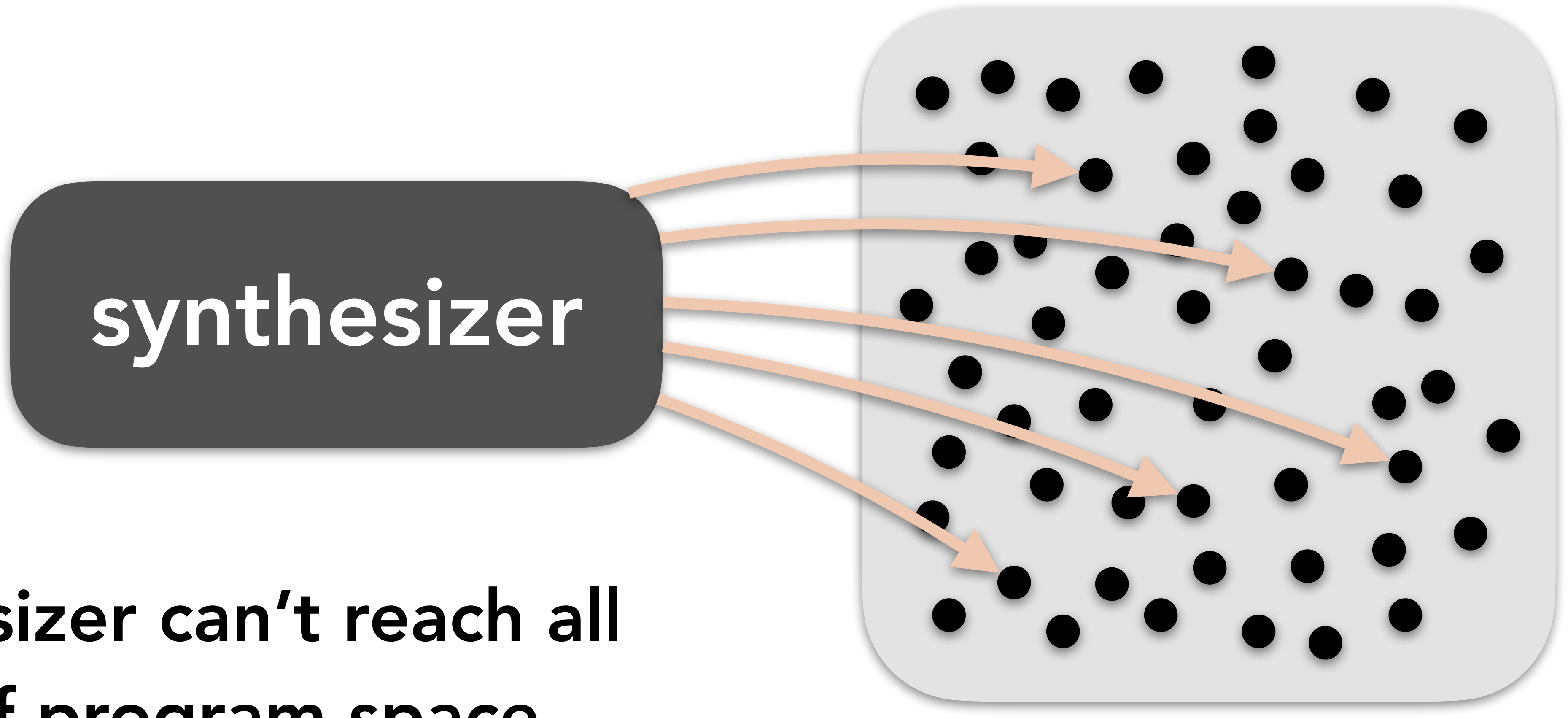
no need to build infrastructure for requesting particular demonstrations



Why this mixed-modality (demo + program edits) input was so much more successful



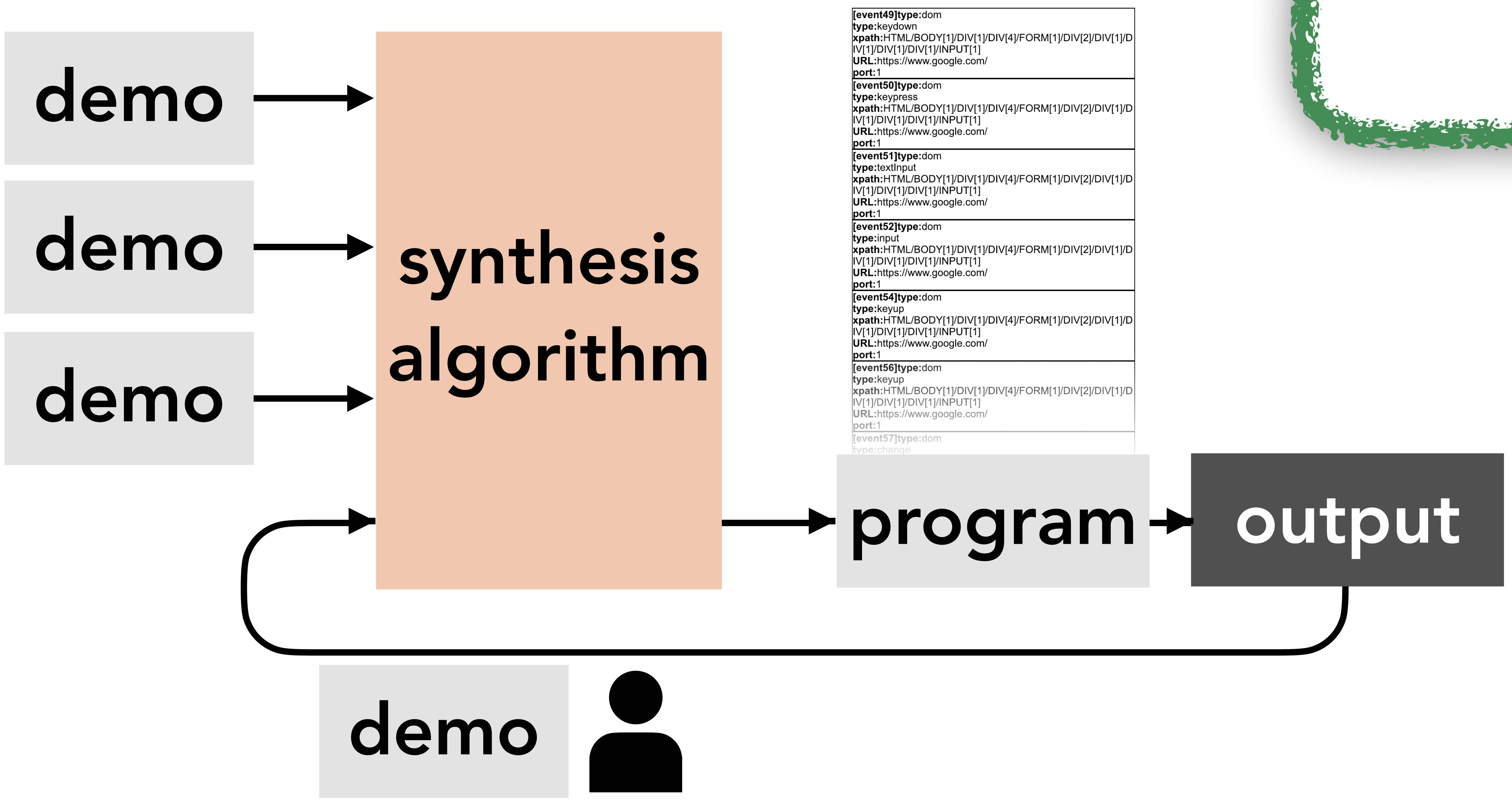
no longer limited to making programs that can run in interactive time



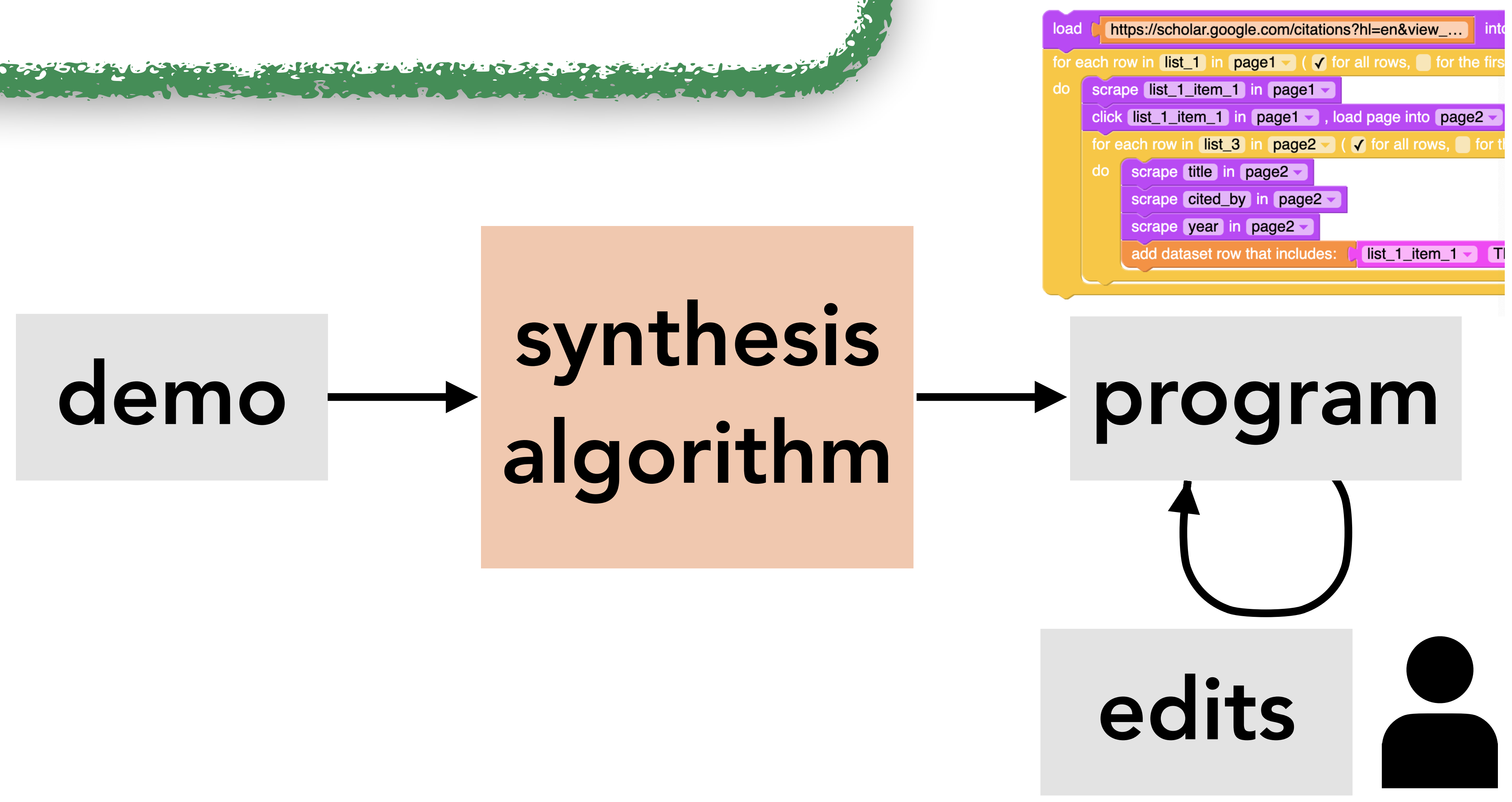
ok if synthesizer can't reach all corners of program space

allowable programs

Human in a dialog with automation



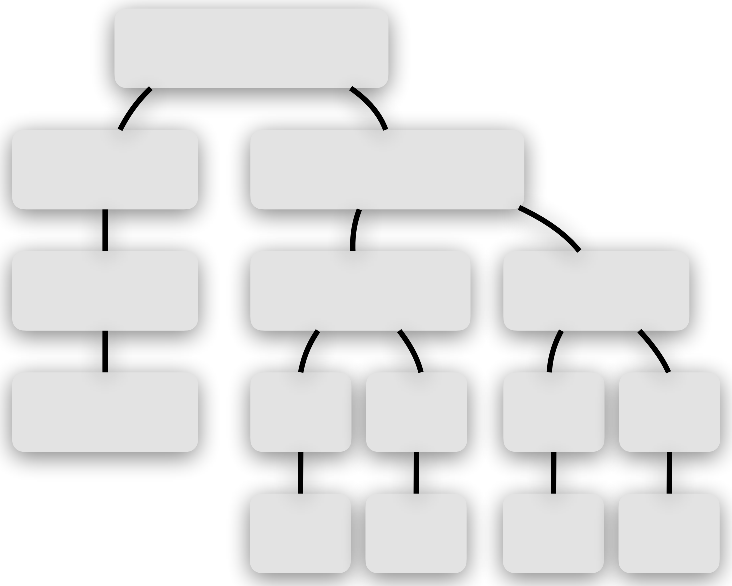
Traditional PBD



With learnable languages

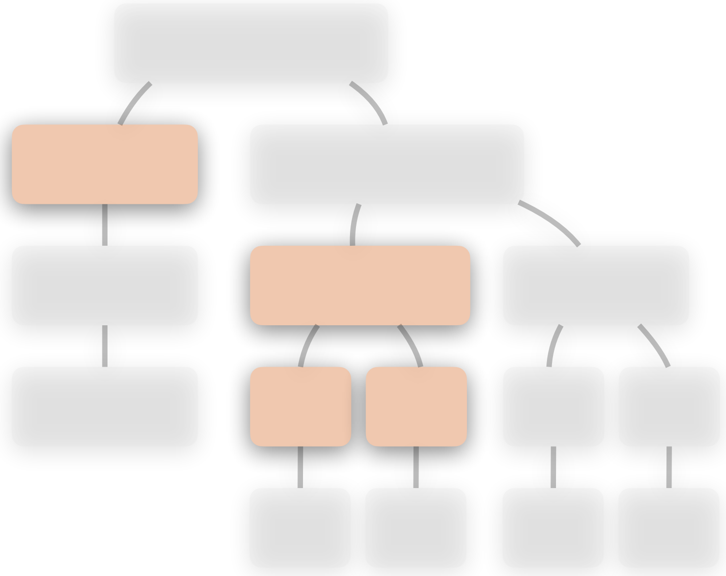
Table Selector Synthesis Problem

Record-Time Webpage W



Interacted Nodes

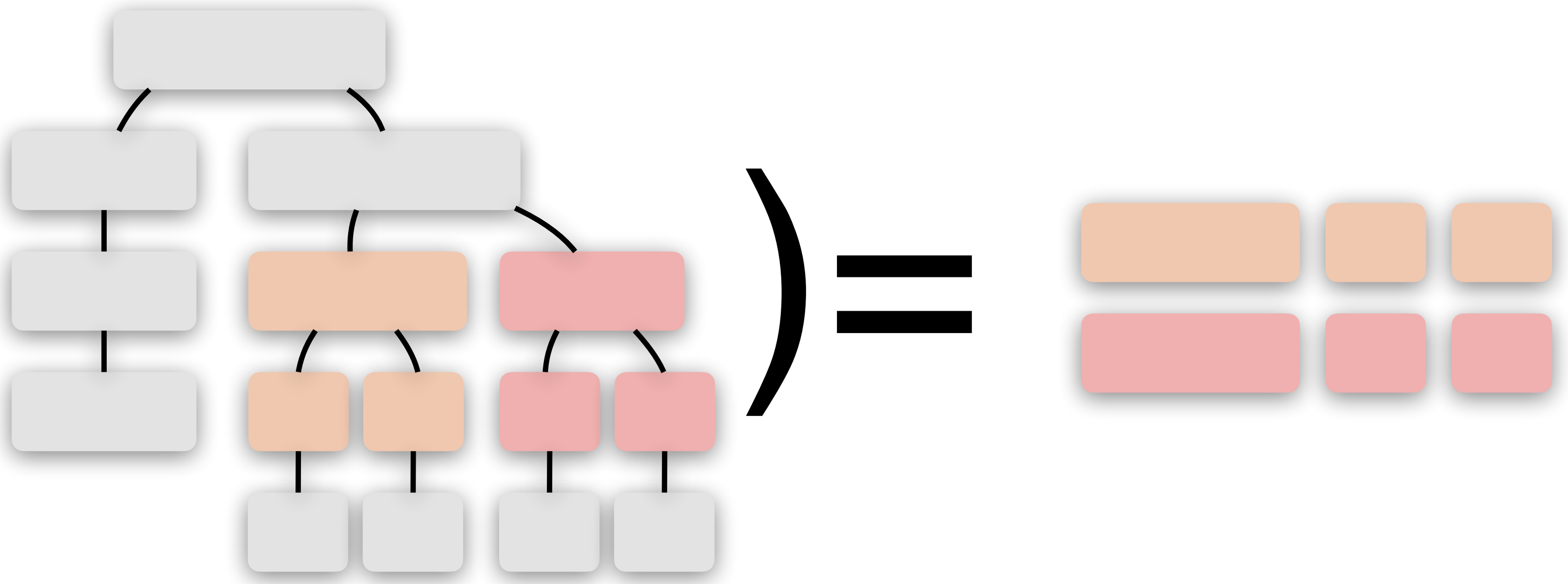
$\mathrm{I}\ \{\mathrm{n} \in \mathrm{W} \mid \mathrm{interacted(n)}\}$

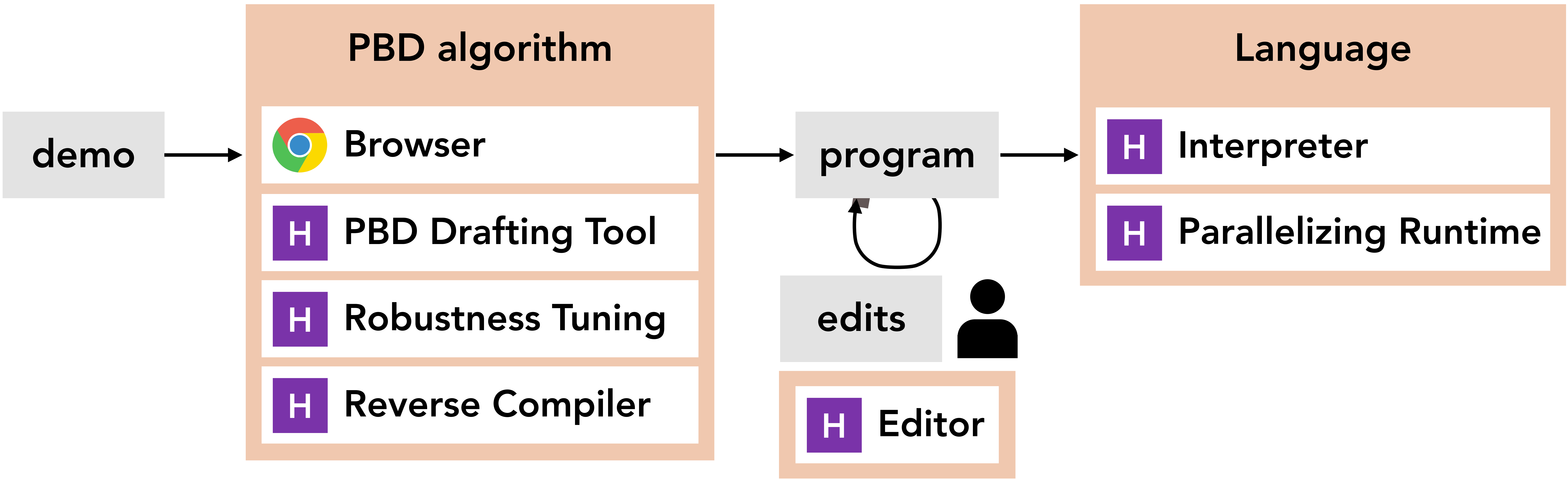


synthesizer

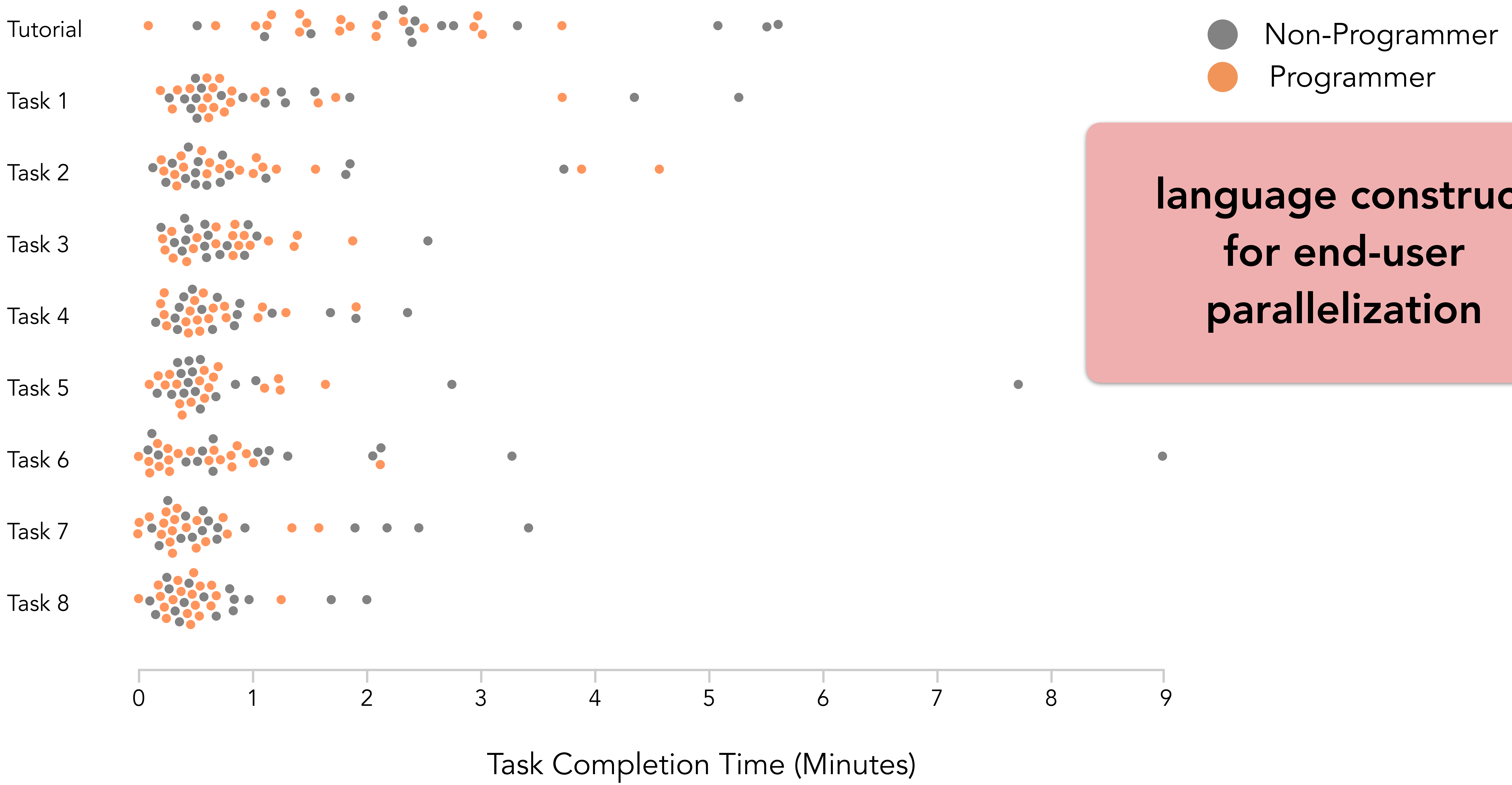
Table Selector

$s.ls(W)_{0,*} \cap I$ maximized $\wedge$ $rows(s(W)) > 1$

s()=


demo
PBD algorithm
Browser
PBD Drafting Tool
Robustness Tuning
Reverse Compiler
program
edits
Editor
Language
Interpreter
Parallelizing Runtime

Non-Programmers Can Parallelize!



language construct for end-user parallelization

Adding a "skip block"

scrape year_published in page2

add dataset row that includes: name TEXT title TEXT citations TEXT year_published TEXT

▸ Relevant Tables

Troubleshooting

What kind of problem are you having?

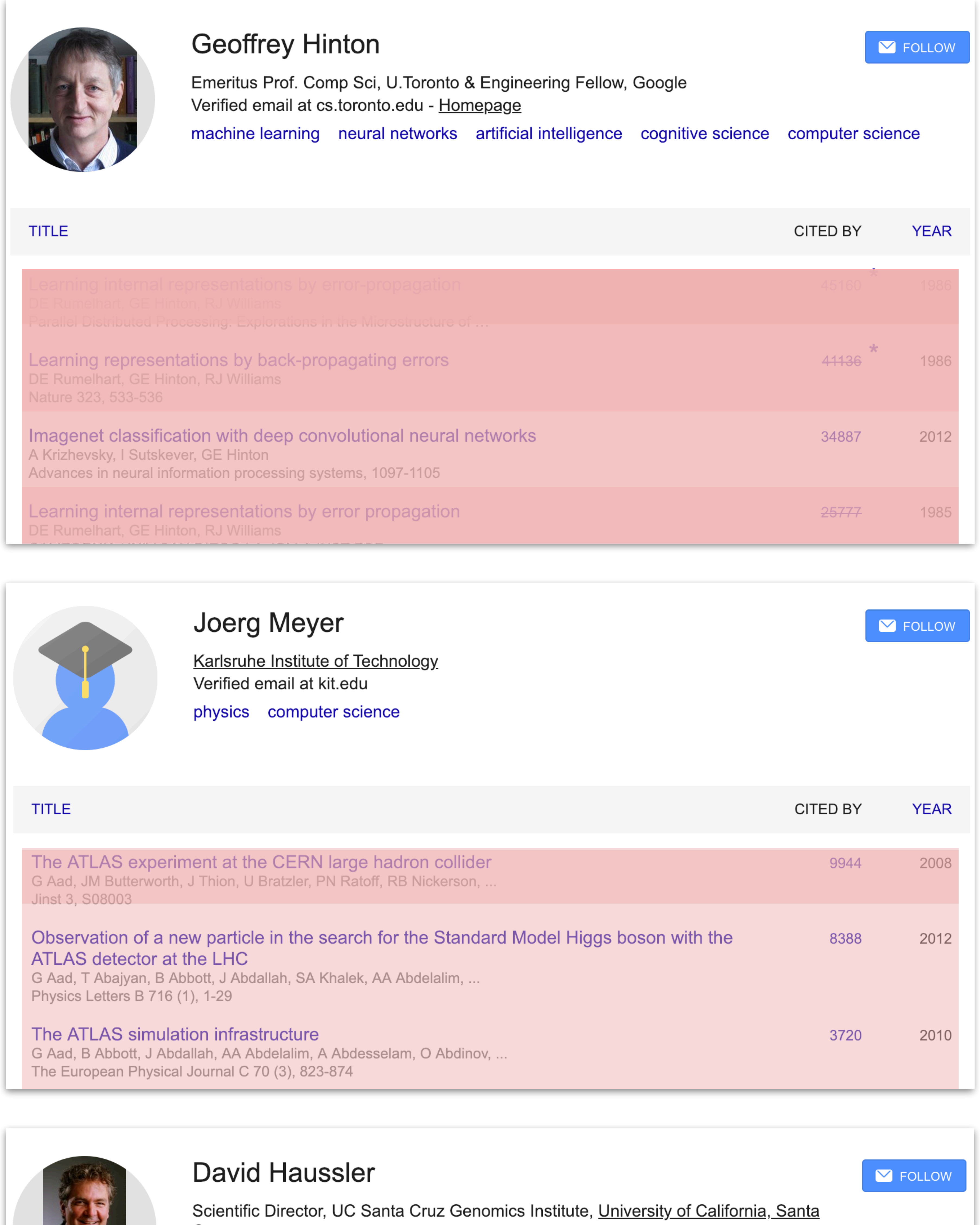
My script wastes time scraping the same stuff it's already scraped.

Detecting Duplicates

institution text	institution link	citations_for_author text	citations_for_author link	text text	text link	pic text	pic link	name text	name link
Emeritus Prof. Comp Sci, U.Toronto & Engineering Fellow, Google		Cited by 266452		Geoffrey Hinton image(https://scholar.google.com/citations?view_op=small_photo&user=JicYPdAAAAAJ&citpid=2) Geoffrey Hinton Emeritus Prof. Comp Sci, U.Toronto & Engineering Fellow, Google Verified email at cs.toronto.edu Cited by 266452 machine learning neural networks artificial intelligence cognitive science computer science		Geoffrey Hinton image(https://scholar.google.com/citations?view_op=small_photo&user=JicYPdAAAAAJ&citpid=2)		Geoffrey Hinton	

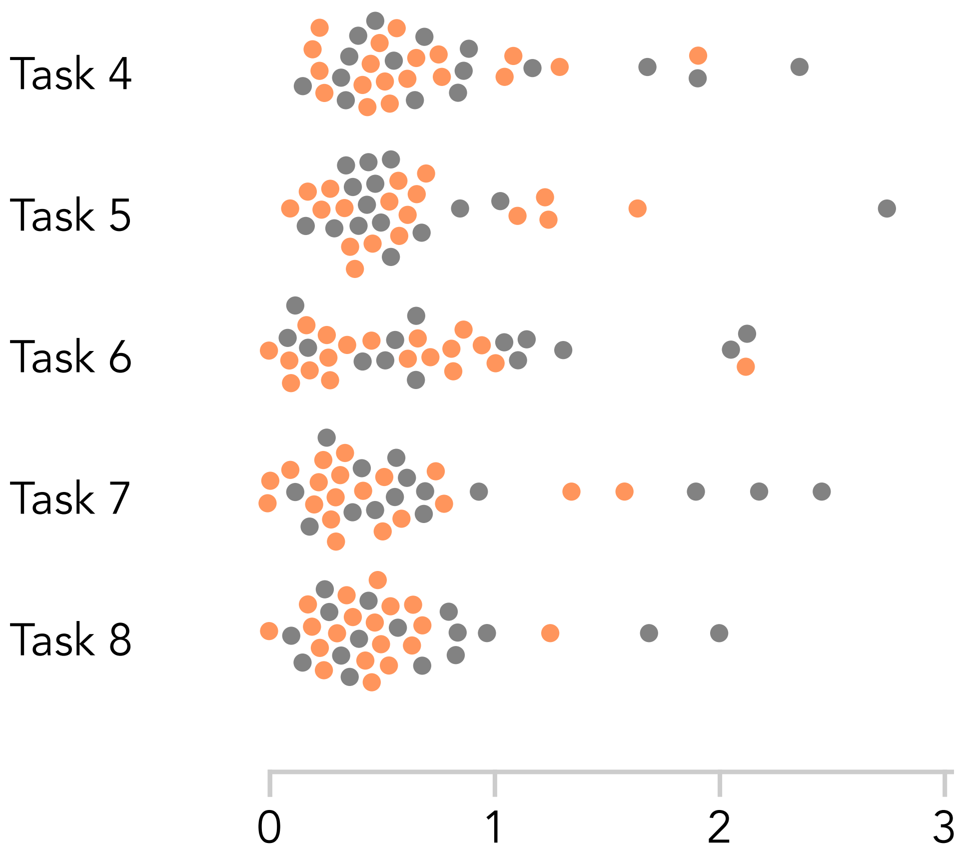
Add Skip Block

Imprecision tolerated

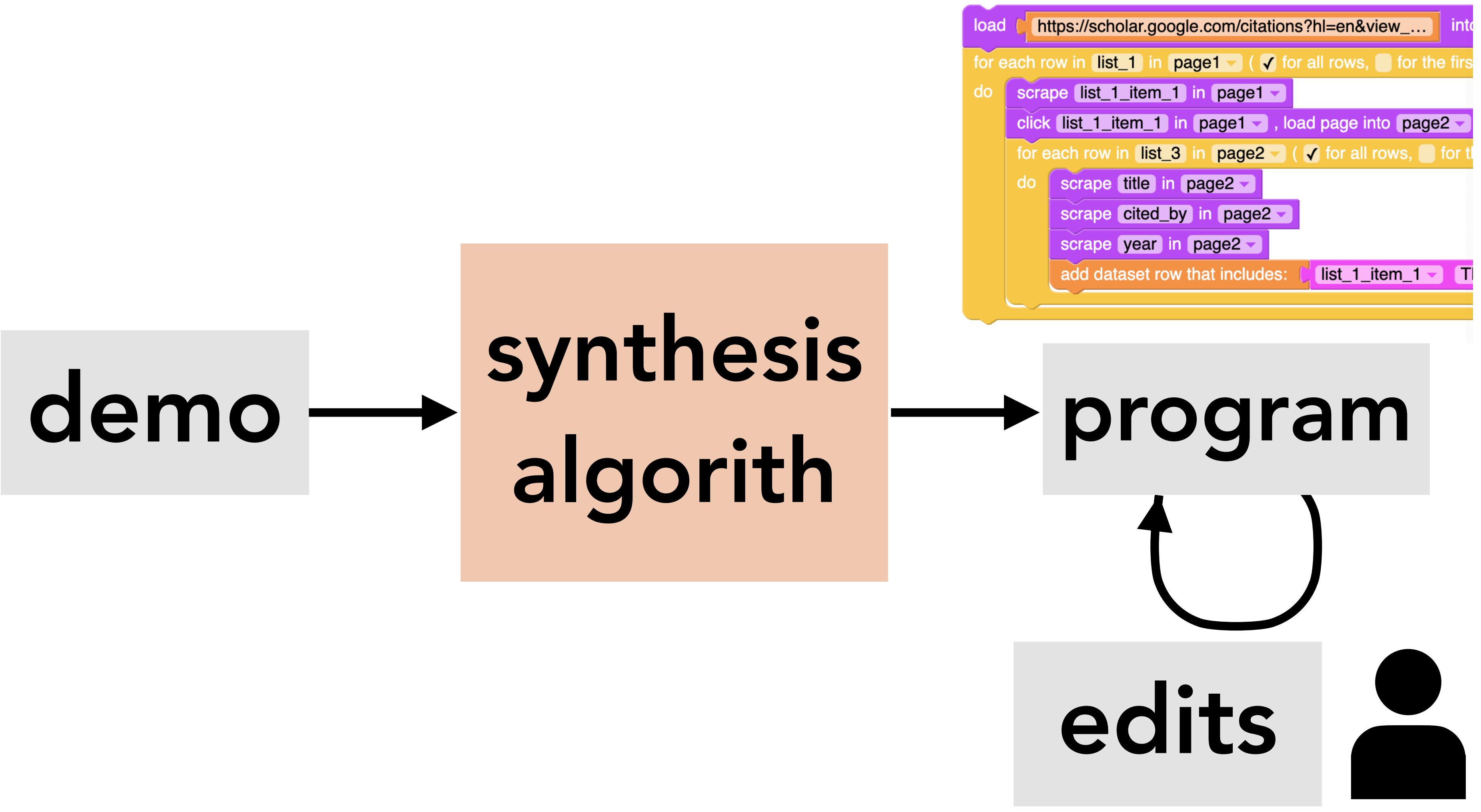


Multiple specification modalities

Why this **mixed-modality (demo + program edits)** input was so much more successful…



Human in a dialog with automation



With learnable languages

- Quick refresher on lab scope, mission

- Whirlwind tour through prior projects that led us to this lab's mission
  - 
  - 
  - 

- **Summary of themes from projects, how they form lab's foundation, preview of today**

EPIC DATA lab

UC Berkeley

Mission: To develop no-code and low-code tools for data science/AI work shaped by the needs of heterogeneous teams.

Key Themes from Prior Work

Imprecision tolerated

Human in a dialog with automation

Multiple specification modalities

people who have data

+

~~4 years data science training~~

learnable data science/AI tools!

=

evidence, insights, models

The building blocks inside our tools

learnable data science/ AI tools

Imprecision tolerated

Human in a dialog with automation

Multiple specification modalities

This Morning

- 9:00am This talk!
- 9:45am Remarks by Jennifer Chayes
- 10:00am Talks
  - Justin Lubin: Exploring the Learnability of Program Synthesizers by Novice Programmers
  - Rachel Warren: Data Munging for Justice
- 10:40am Break
- 11:15am Talks
  - Samantha Robertson, Human-Centered Tools for Reliable Use of Machine Translation
  - Shreya Shankar, Operationalizing Machine Learning: An Interview Study
  - Dixin Tang, Lux: Always-on Visualization Recommendations
  - Hellina Hailu Nigatu, Document Organization Three Ways
- 12:30am Lunch

This Afternoon

- 2:00pm Talks
  - Rebecca Brown, Challenges Facing Nonprofits in Justice Reform
  - Çağatay Demiralp, Research Problems at Sigma Computing
- 3:00pm Break
- 3:30pm Poster Session Preview
- 4:15pm Poster Session/Reception
- 6:30pm Dinner (Offsite)

Tomorrow

Centered on conversations, two-way communication

- 8:30am Breakfast available in Room 511 Soda Hall
- 9:00am Small Group Discussions
  - Research theme discussion rooms in 4th-floor lab and 5th-floor lab
  - Scheduled meetings with faculty group
- 12:00pm Lunch
- 1:30pm Wrap up session in 510 Soda
- 2:00pm End