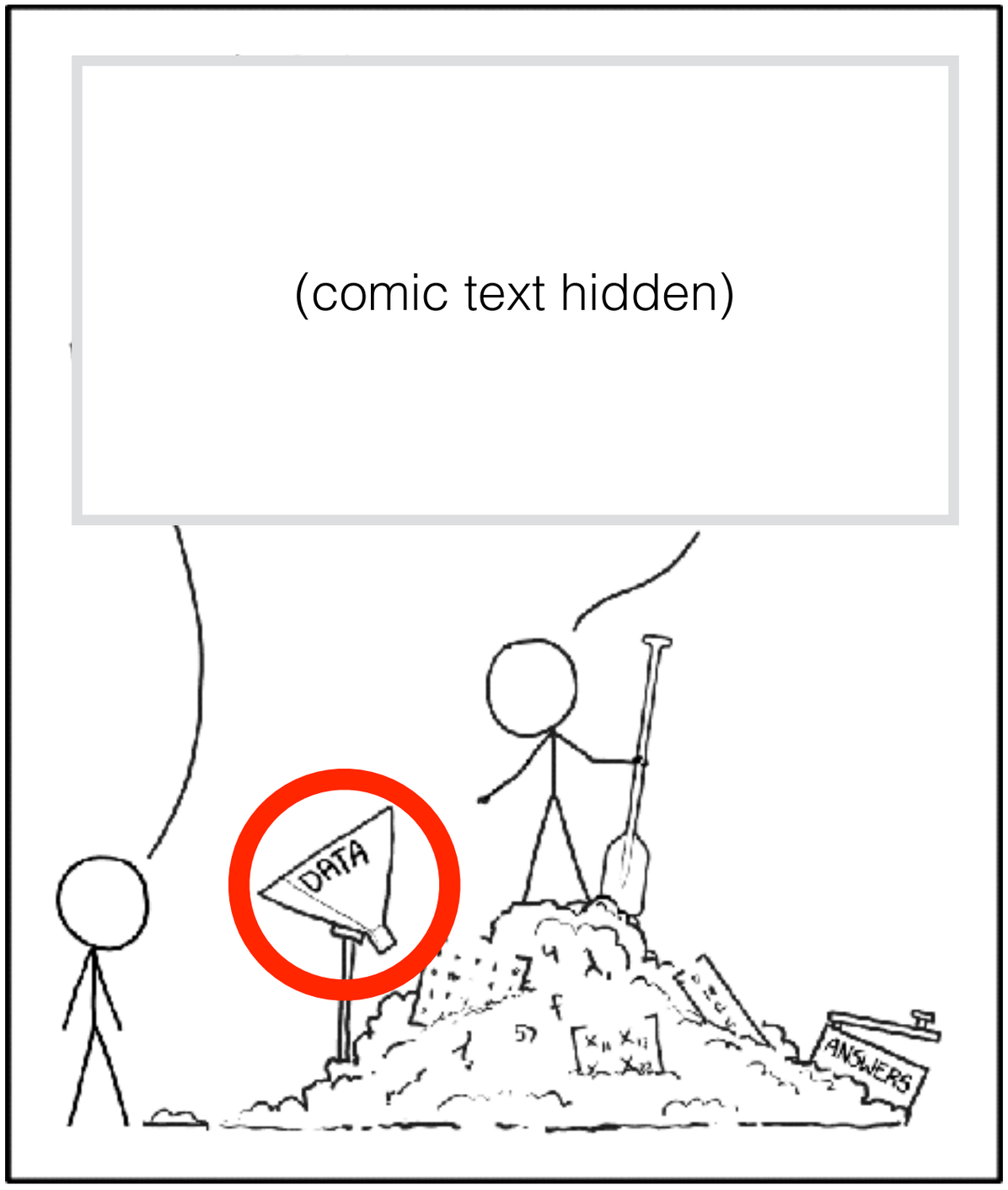
NEXT: Crowdsourcing, machine learning and cartoons

Scott Sievert
@stsievert

Link to slides and proceedings:
tinyurl.com/scipy-next

Problem



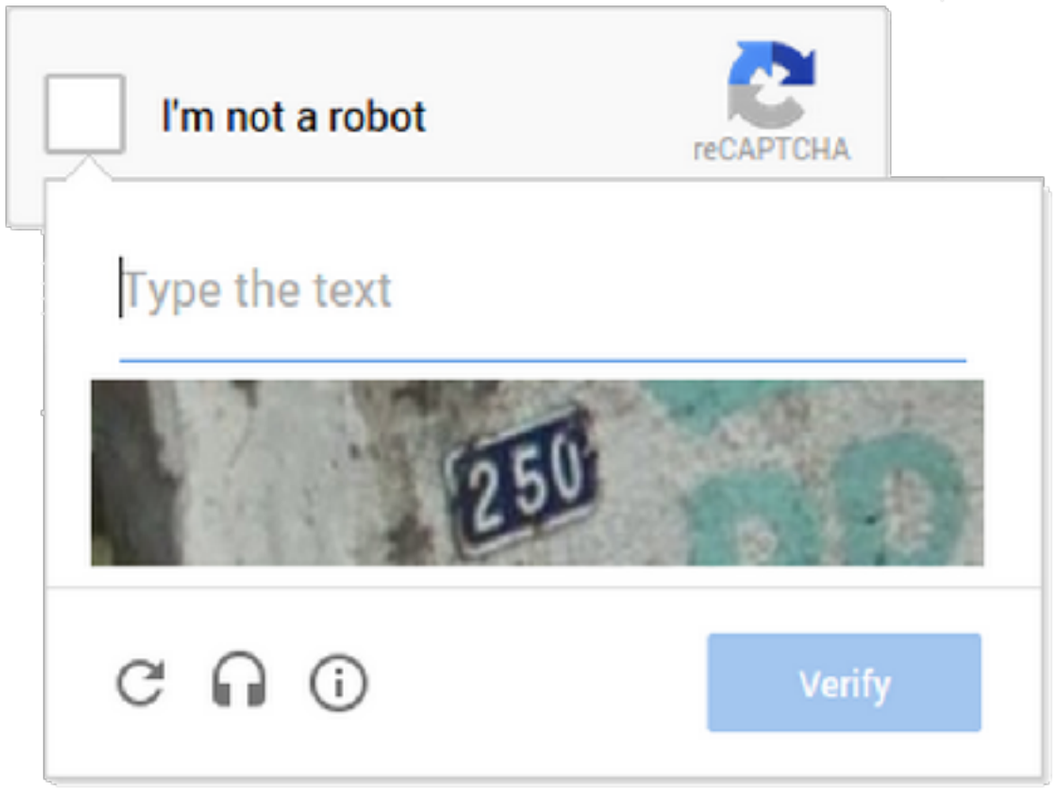
Data collection can be costly

https://xkcd.com/1838/

tinyurl.com/scipy-next

Example

Data collection done with
crowdsourcing can be expensive

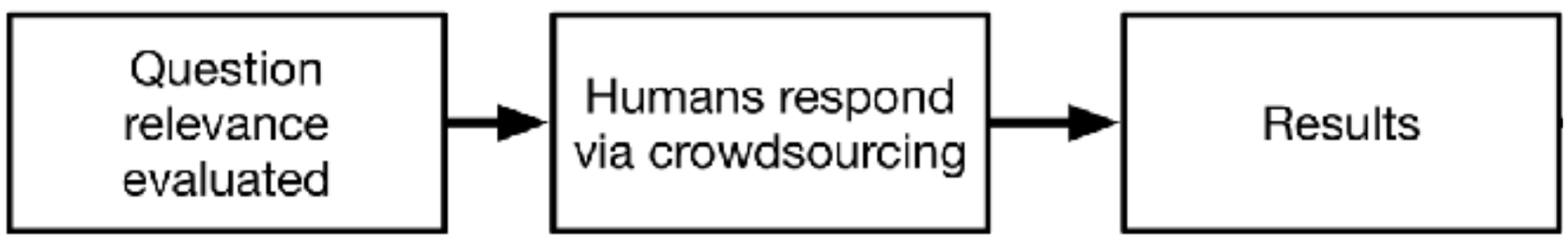


Goal: achieve goal with minimal
responses

tinyurl.com/scipy-next

One solution

Existing crowdsourcing systems are *passive*

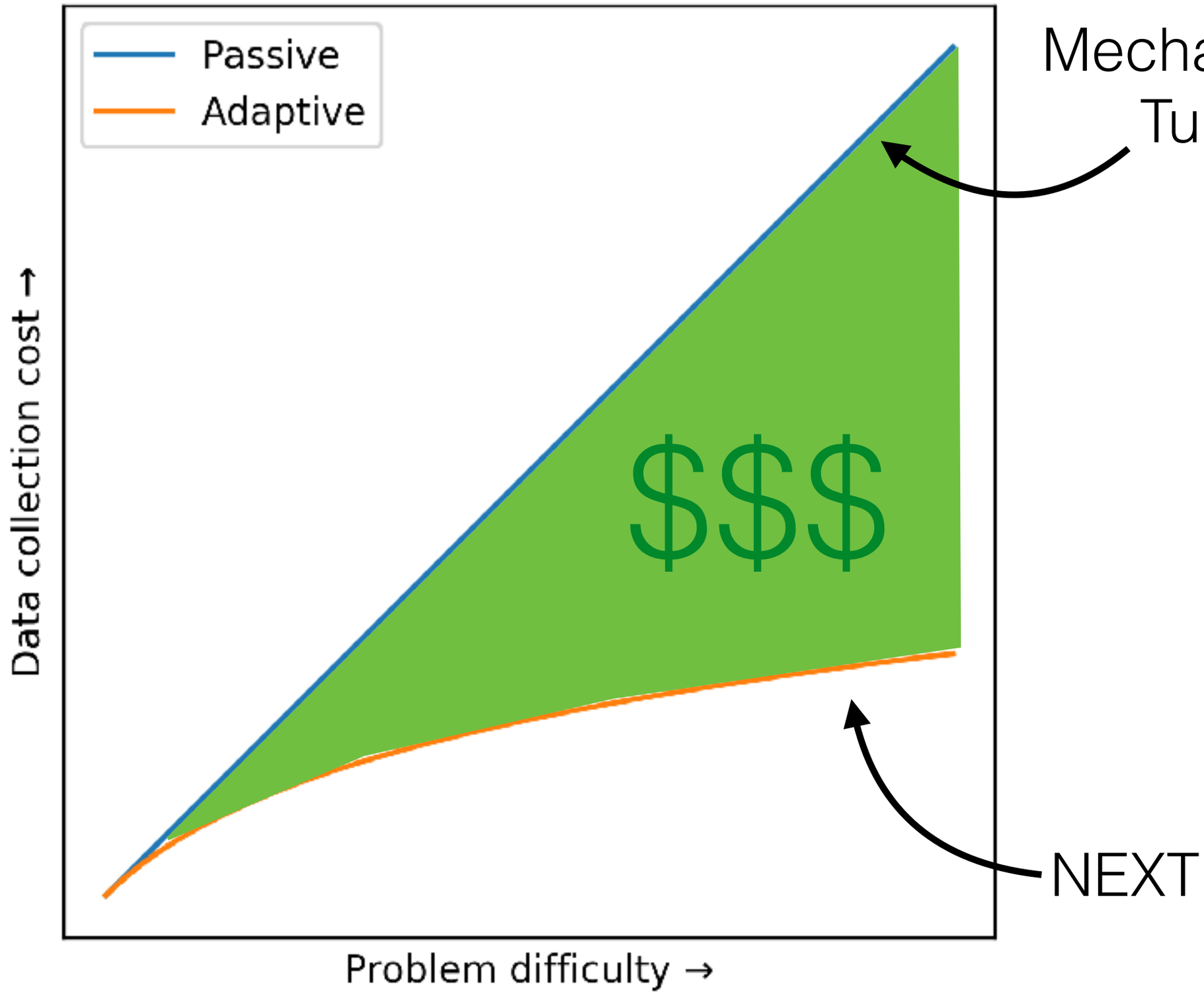


Adapting to previous responses requires fewer data

Goal: adapt to previously collected responses

tinyurl.com/scipy-next

Benefits

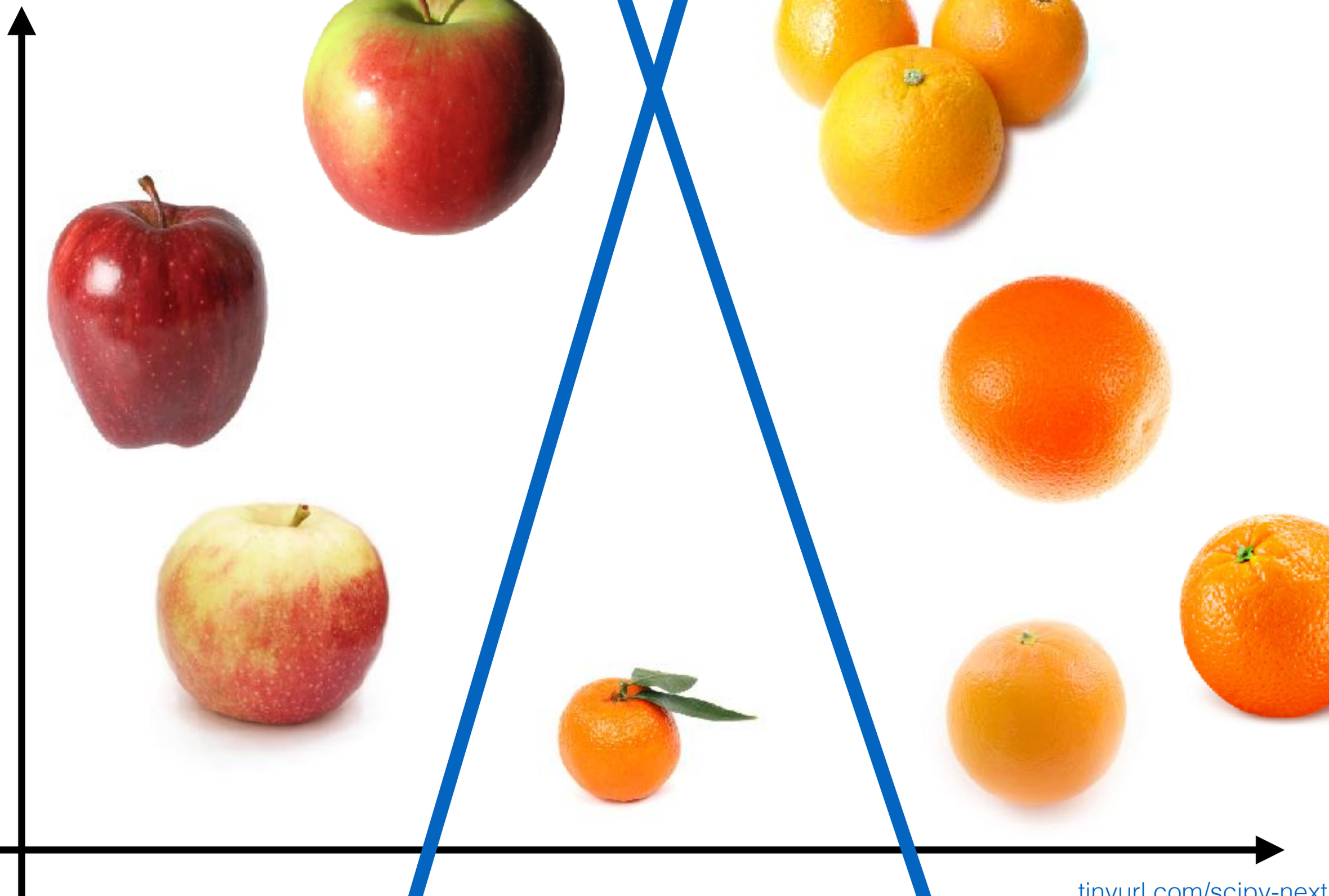


Adaptive sampling can have large benefits

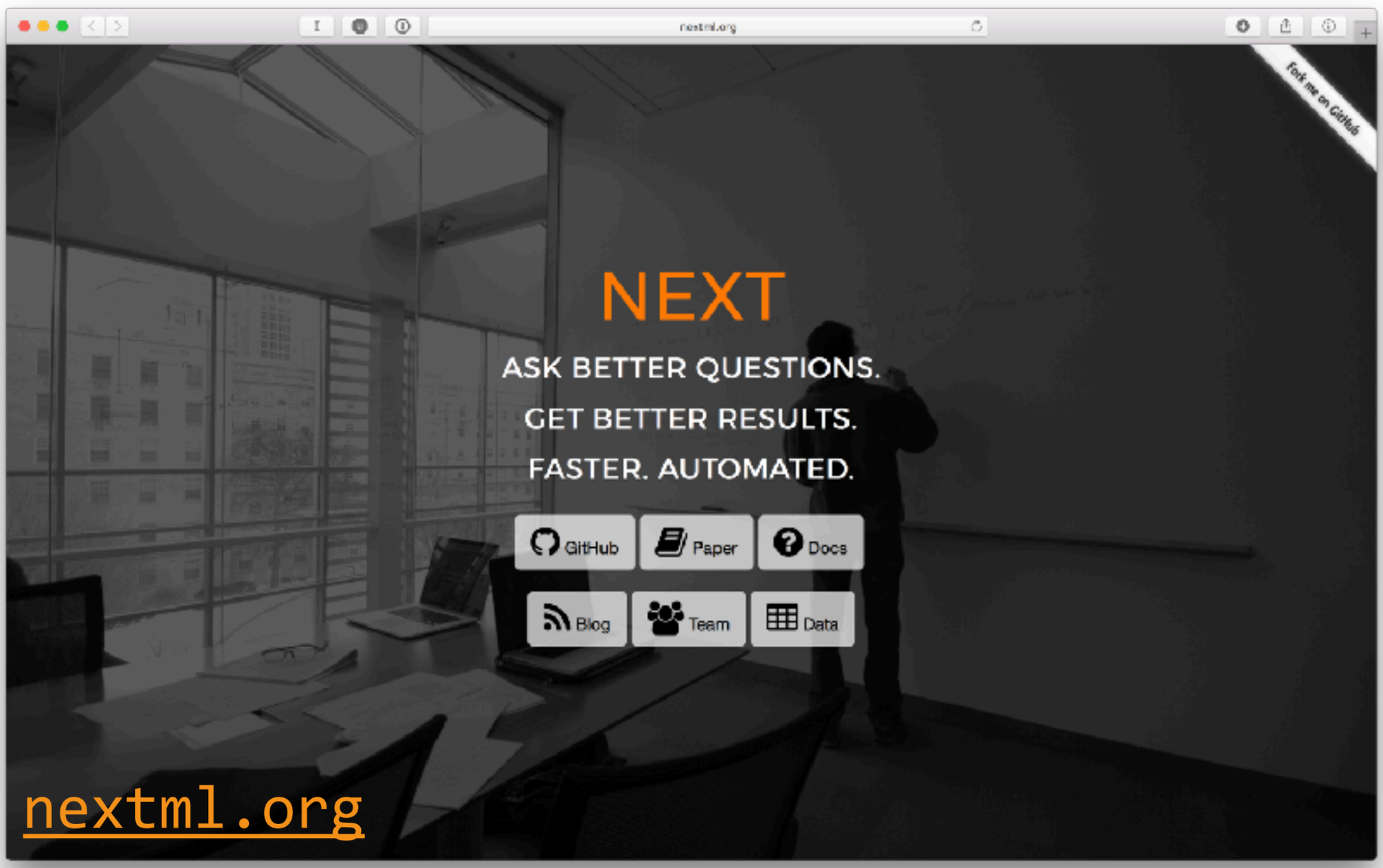
tinyurl.com/scipy-next

Example solution

Adapting to previous responses yields better results



tinyurl.com/scipy-next



Lalit Jain

Daniel Ross

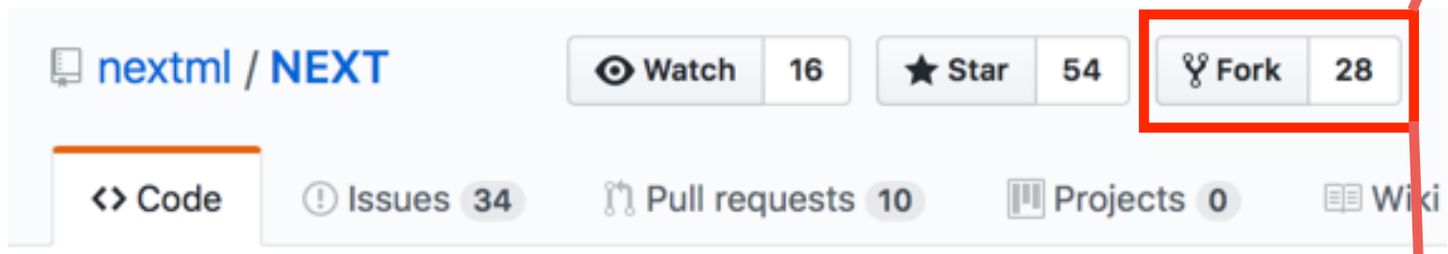
Homepage: http://nextml.org
Source: https://github.com/nextml/NEXT
Documentation: https://github.com/nextml/NEXT/wiki

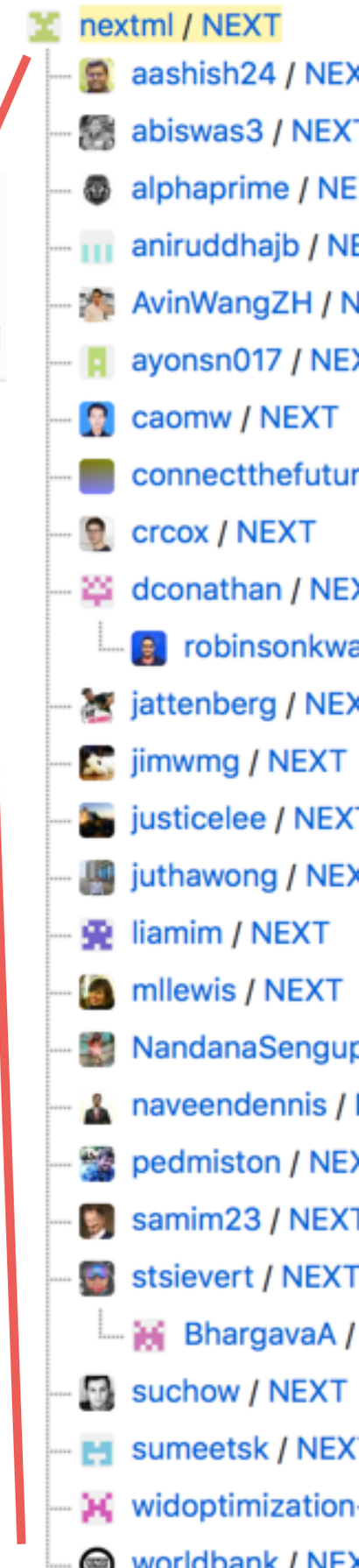
Kevin Jamieson

Rob Nowak

tinyurl.com/scipy-next

NEXT users

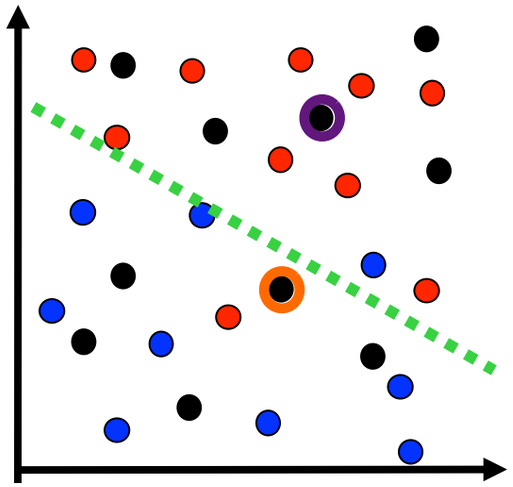




Theory ← → Practice

ML Researchers

Air Force Research Lab uses NEXT for active image classification.



Experimentalists

UW Psychology uses NEXT to find the best algorithms for adaptive data collection in cognitive science.

Practitioners

The New Yorker uses NEXT to crowd-source the weekly cartoon caption contest.

tinyurl.com/scipy-next

Example problem

[Bob Mankoff]

YOUR CAPTION

Enter your caption (250 characters or fewer):

The New Yorker has to find the funniest caption from ~5,000 captions

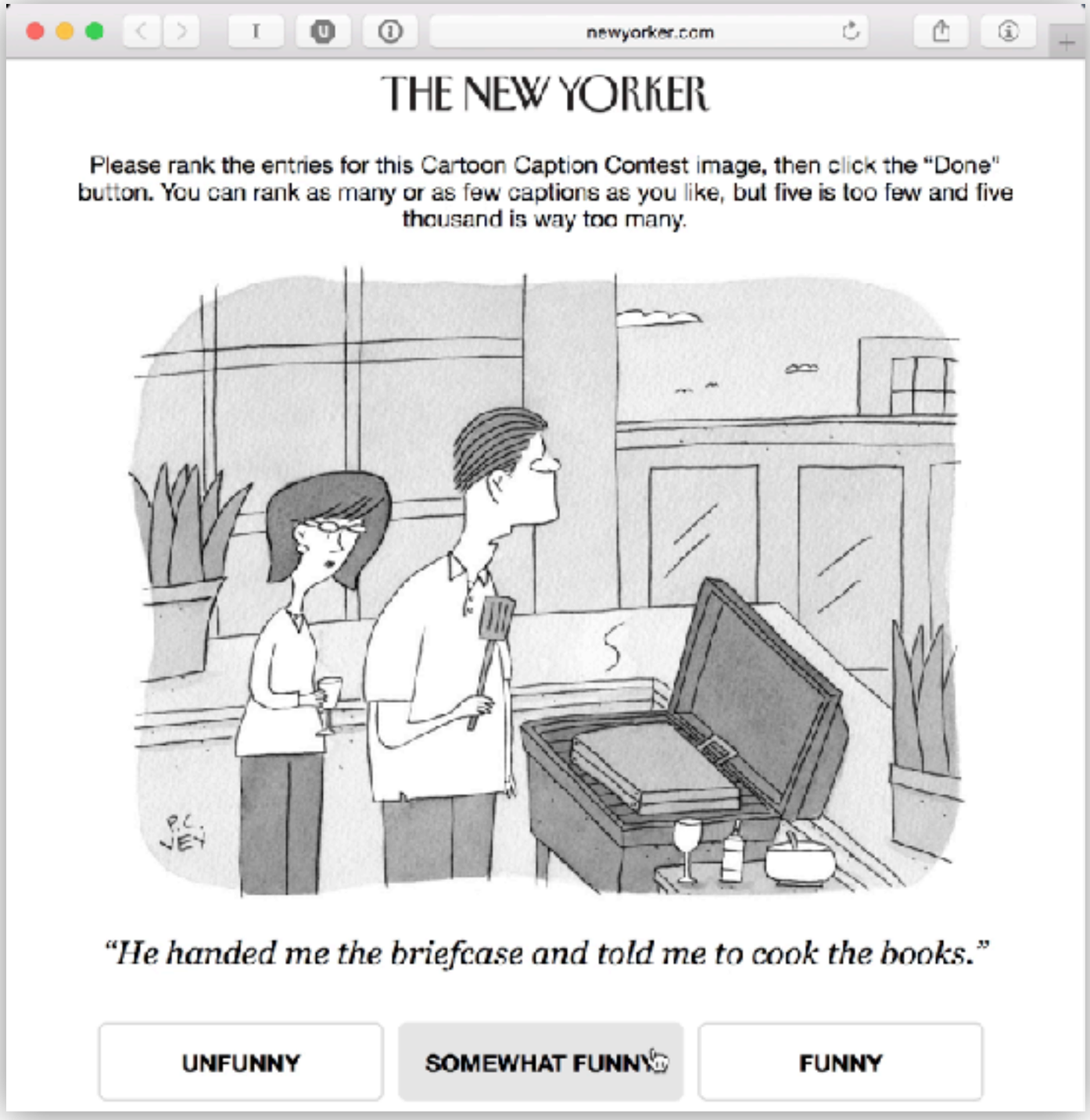
Comic by [P. C. Vey]

tinyurl.com/scipy-next

Interface

http://www.newyorker.com/cartoons/vote

http://nextml.org/captioncontest



Comic by P. C. Vey

Dashboard

ec2-34-208-101-150.us-west-2.compute.amazonaws.com

NEXT Experiment Dashboard | NEXT - b85faed0ebbc9c6256bdcaa5f2951a

NEXT

EXPERIMENT SYSTEM

Histogram of time responses received

Experiment Info

Start date: 2017-06-26 14:39:13.745959 UTC
Dashboard data generated: <1 minute ago
Number of participants: 14111
Number of answers: 423367

Rankings

KLUCB

Download CSV

Rank	Target	Score	Precision	Count	Unfunny	Somewhat Funny	Funny
1	I'd better give it a little longer. It's a really tough case.	2.0921	0.0135	3107	751	1319	1037
2	He told me to cook the books, but didn't say for how long.	1.9946	0.0113	4458	1280	1922	1256
3	Hopefully cooking the books works better than laundering the money.	1.9777	0.0177	1880	577	768	535
4	Something t s					1568	1336
5	No, I won't be going back to work on Monday. Why	1.9679	0.0123	3864	1193	1602	1069

Data from contests:
https://github.com/nextml/caption-contest-data

tinyurl.com/scipy-next

Goal

Adaptive sampling algorithms

fewer responses
more accurate models

real-world data
participant fatigue
algorithm delays
participant label quality

Crowdsourcing

Goal: let both parties *easily* use NEXT

tinyurl.com/scipy-next

Software uses

By default, NEXT can be applied to 3 problems

Select face on the bottom most similar to the face on top

Pool based triplets

Cardinal Bandits

comic by P. C. Vey

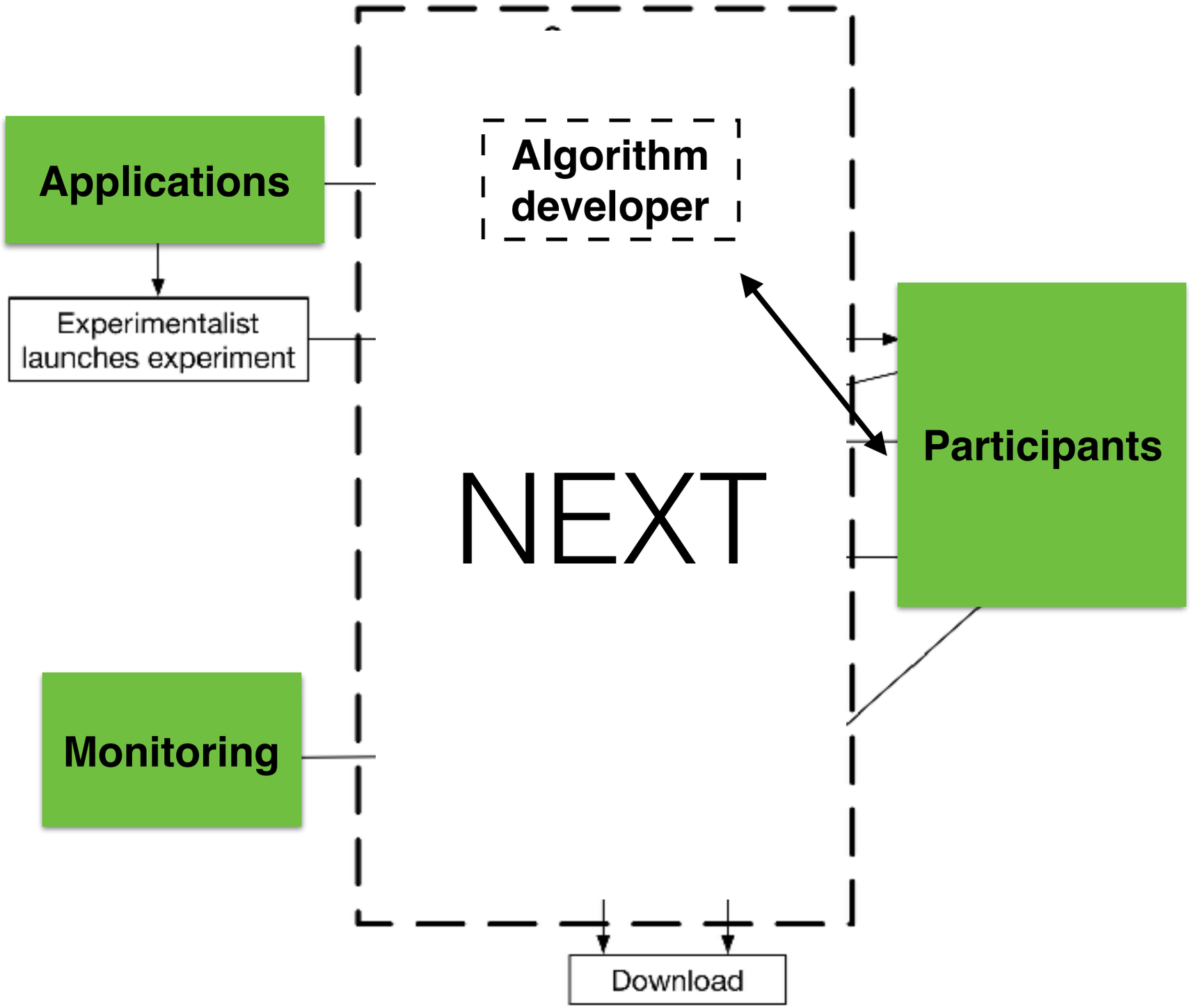
Dueling Bandits

Select the street that looks safer

NEXT can also be used with REST API

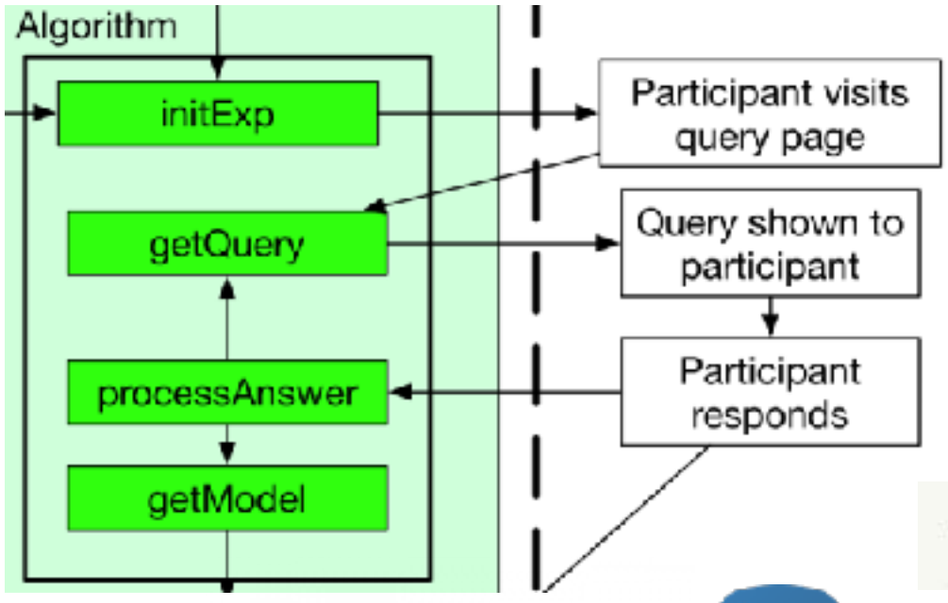
tinyurl.com/scipy-next

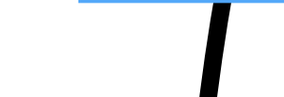
NEXT use

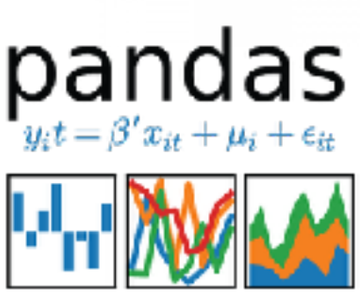


tinyurl.com/scipy-next

Algorithm developer use







initExp

getQuery

processAnswer

getModel

$$\Pr\left(|y - \hat{y}| < \epsilon\right) \leq 1 - \delta$$

Algorithm developer and mathematician

tinyurl.com/scipy-next

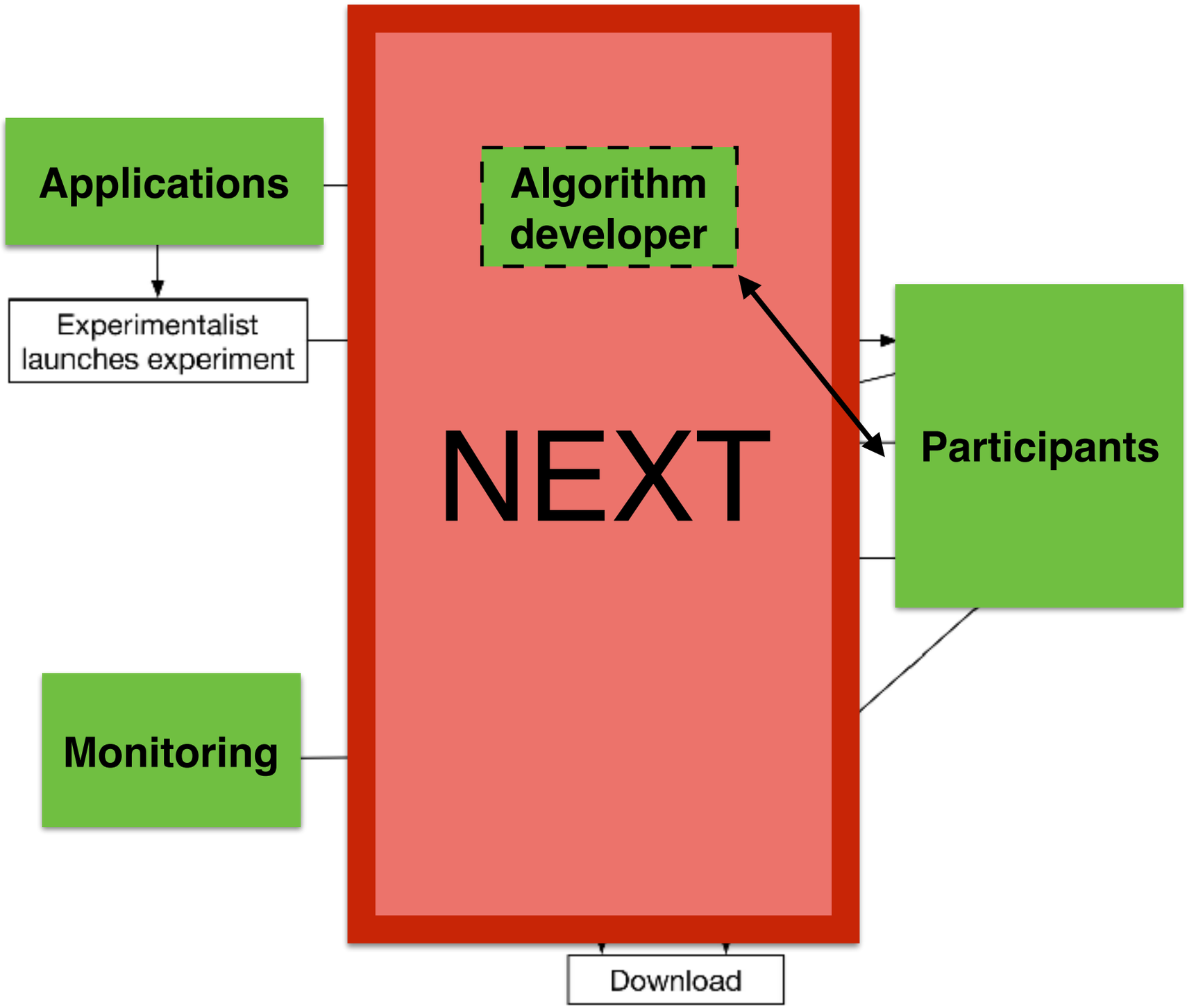
Algorithm design decisions

1. Treat algorithms as black boxes
    - (for each function, inputs and outputs are documented and type-checked)
2. Use wrapper to allow *easy* access to experiment information and background jobs
3. Objects are abstracted to integers (i.e., object `42`, not `{'filename': foo.png, 'url': …}`)

(more detail in proceedings and on docs)

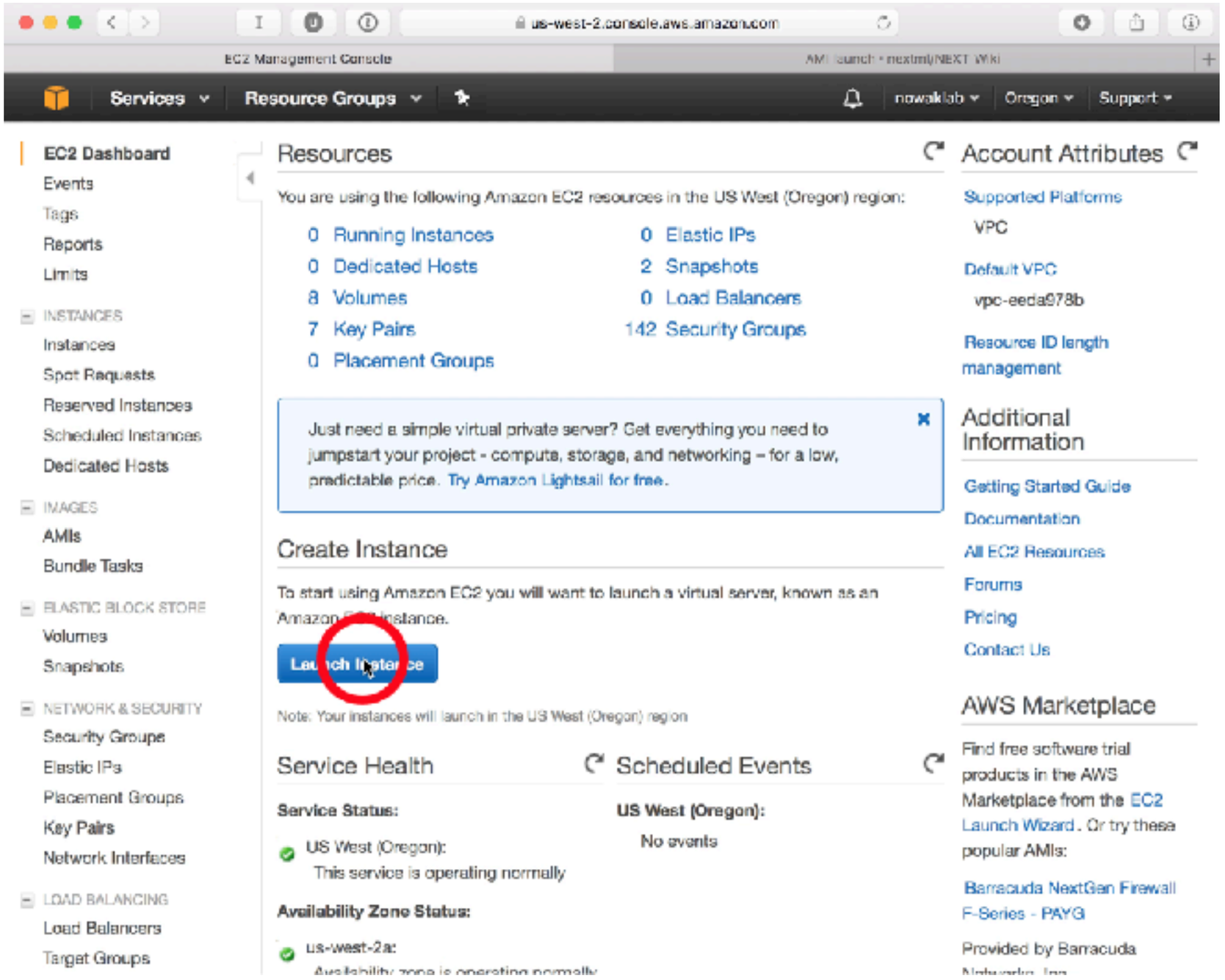
tinyurl.com/scipy-next

Algorithm use



tinyurl.com/scipy-next

Launching NEXT via Amazon EC2 AMI

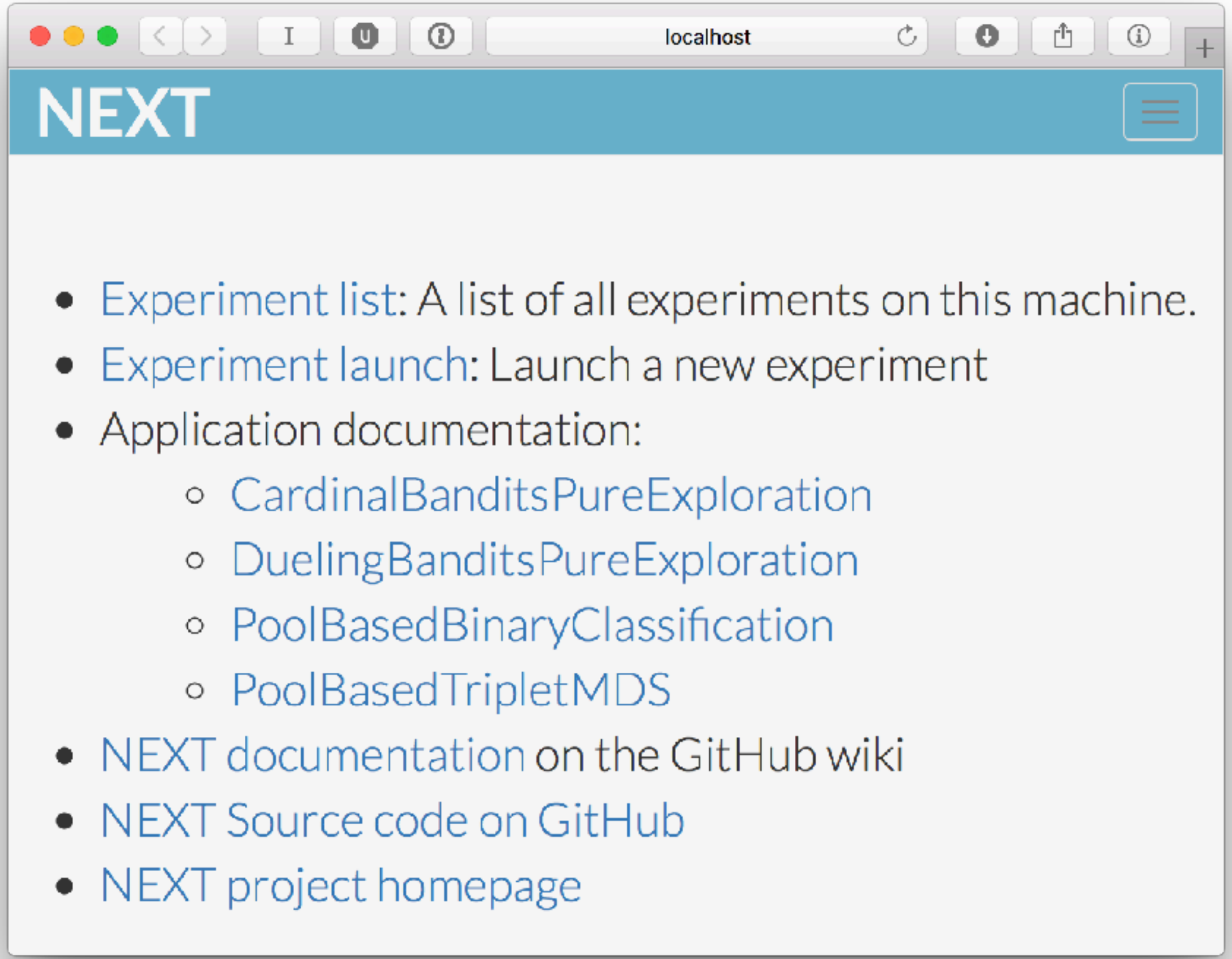


(more detail in proceedings and on docs)

See https://github.com/nextml/NEXT/wiki for details and more launching options

tinyurl.com/scipy-next

NEXT startup page



tinyurl.com/scipy-next

Key messages

1. Adaptive sampling reduces data collection cost.

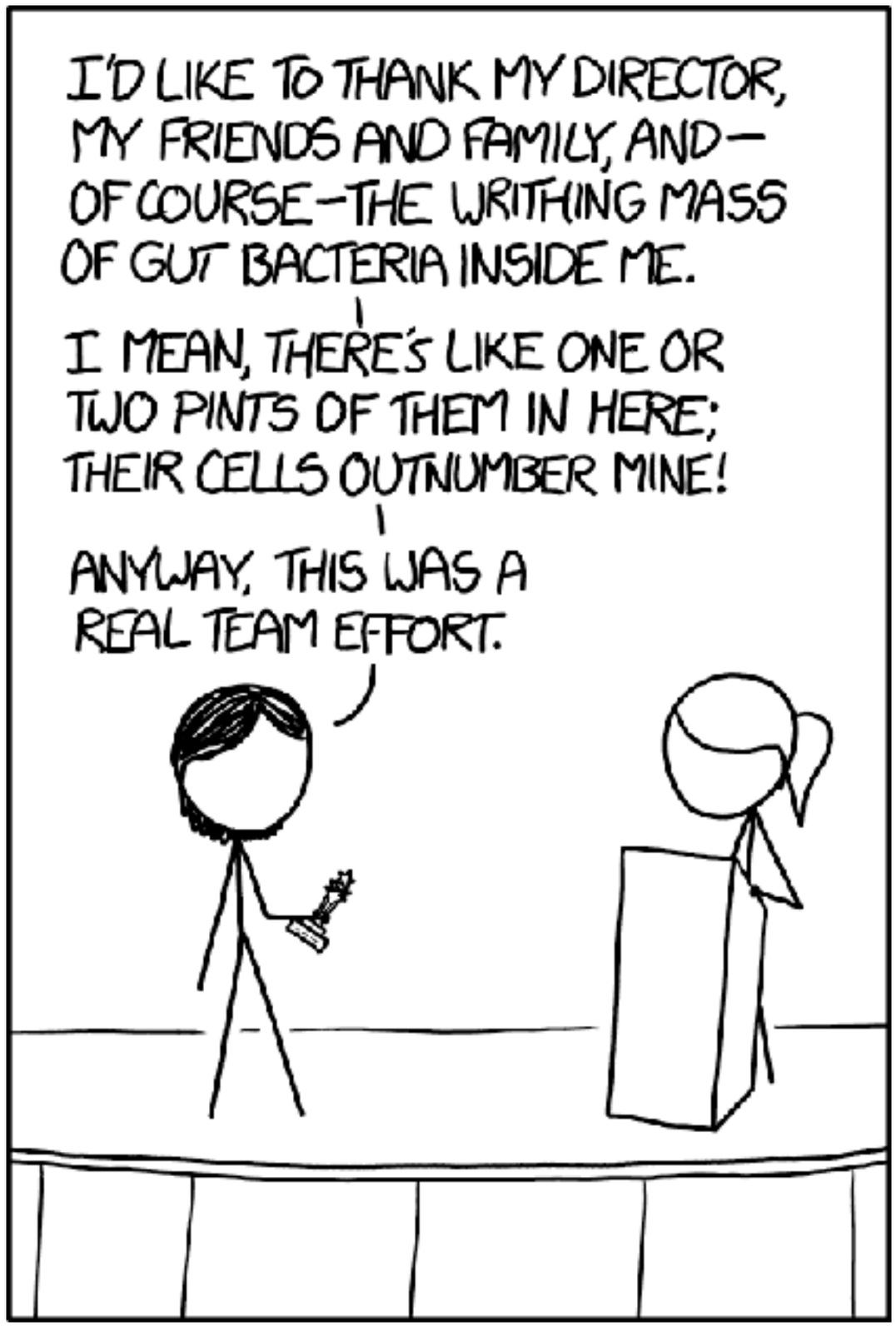
2. NEXT is a crowdsourcing data collection tool that can use adaptive sampling techniques

3. NEXT is easy* to use by experimentalists, algorithm developers and practitioners, and a mathematical background is not required.

4. NEXT developers experimentalist engagement to aid research and to gain feedback to improve the software

* NEXT has been created by an academic research group for collaboration with psychologists

tinyurl.com/scipy-next



https://xkcd.com/1543/

tinyurl.com/scipy-next

Extras…

Algorithm inputs and outputs

- Documented exactly in `apps/[app-id]/algs/Algs.yaml`

```
getQuery:
  args:
    participant_uid:
      type: string
      description: ID of the participant answering the query
  rets:
    description: The index of the target to ask about
    type: num
```

- Function implementation

```
import random

def getQuery(self, butler, participant_uid):
    n = butler.algortithms.get(key='n')
    return random.choice(n)
```

Depends on a library we developed:
https://github.com/daniel3735928559/pijemont

Adaptive data flow

- Fundamentally requires 4 functions:
  - **initExp**: experiment initialization
  - **getQuery**: selects query to present to participant
  - **processAnswer**: process participants response
  - **getModel**: provides experiment monitoring