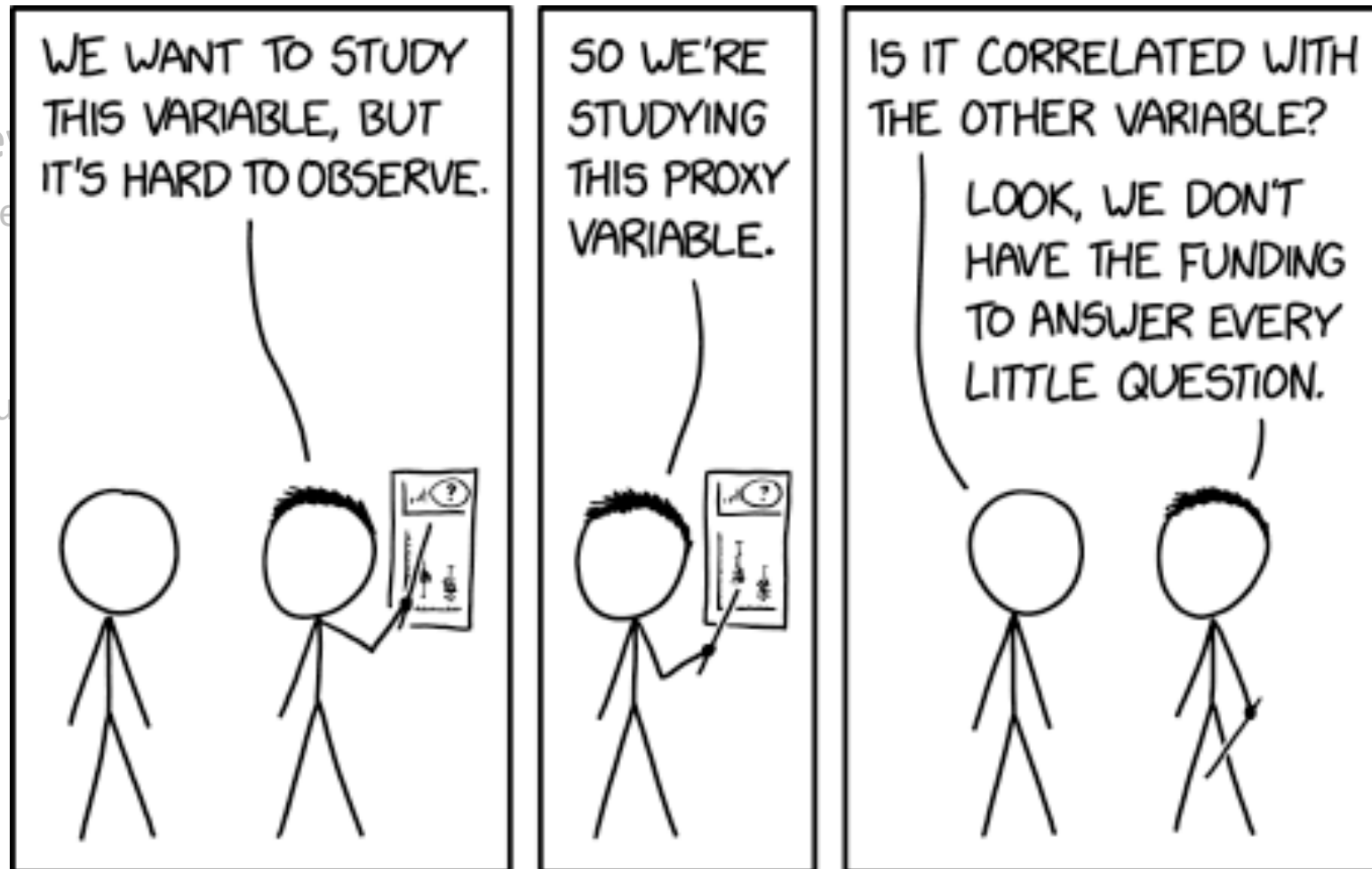# Measurement by Proxy:
On the Accuracy of Online Marketplace Measurements

Alejandro Cue

acuevasv@cs.cmu.e

Kyle Soska

soska2@illinois.edu

Carnegie
Mellon
University

https://xkcd.com/2652/

Measurement by Proxy:
On the Accuracy of Online Marketplace Measurements

Alejandro Cuevas
acuevasv@cs.cmu.edu

Fieke Miedema
f.e.g.miedema-1@tudelft.nl

Kyle Soska
soska2@illinois.edu

Nicolas Christin
nicolasc@cmu.edu

Rolf van Wegberg
r.s.vanwegberg@tudelft.nl

Drugs

Fraud
Tutorials
Services

Digital Goods

Darknet Marketplaces (DNMs) measurements influence **law enforcement**

FOR IMMEDIATE RELEASE   Thursday, July 20, 2017

AlphaBay, the Largest Online 'Dark Market,' Shut Down

'Dark Net' Site Was Major Source of Fentanyl and Heroin, Linked to Overdose Deaths, and Used By Hundreds of Thousands of People to Buy and Sell Illegal Goods and Services Anonymously over the Internet

Case 1:17-at-00557   Document 1   Filed 07/19/17   Page 8 of 27

FACTUAL OVERVIEW

4.   As detailed below, from December 2014, up to and including July 2017, the AlphaBay hidden website served as a marketplace for illegal goods—selling malware, controlled substances, chemicals, guns, stolen financial information and counterfeit documents to its users located all over the world. AlphaBay was the largest dark web market and its annual sales were estimated to be in the hundreds of millions of dollars, many times more than Silk Road, a predecessor dark web marketplace that primarily sold controlled substances and was shut down by the government in 2013.

DNM measurements influence **policy-making**

H.R.7300 - Dark Web Interdiction Act of 2022
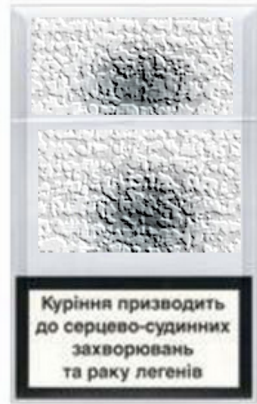117th Congress (2021-2022) | Get alerts

S.3782 - Dark Web Interdiction Act of 2022
117th Congress (2021-2022) | Get alerts

Congressional Research Service
Informing the legislative debate since 1914

The Dark Web: An Overview

Home / Drugs / ▬▬▬▬▬▬▬

[■■■■■] Silver x10 packs

(1)

USD 25.00
₿ 0.0338

999 in stock

Shipping options

Worldwide Express Shipping [8 days] [+ USD 15.00]

(2)

Vendor	■■■■■ Level 3 (10+) ★ Trusted Vendor
Class	Physical
Ships From	Worldwide
Ships To	Worldwide
⚠ Except	United States, Canada, Australia, France, Germany, New Zealand, United Kingdom

Quantity: 1

🛒 Buy Now

? Question   🚩 Report

hansamkt2rr6nfg3.onion

Куріння призводить до серцево-судинних захворювань та раку легенів

(3)

ⓘ Details   💬 Feedback

Listing Details

Price is per one carton (■■■■■■■■■)
Strength: ■■■■■■■■■■■■
See my profile page for other details and do not order before reading our terms and conditions!

Be free to PM me with any questions ;) Especially if you are searching brand, that isn't in my listings.

I DO NOT SEND cigarettes to USA, Germany, Australia, Canada, New Zealand and temporarily to France and UK.
If you order more than 1 carton - make different orders for each carton.
Not correct orders will be immediately refunded and you may lose funds on different exchange rates/commission etc..

I respect your part of deal and not asking for FE, respect my part of deal - don't forget to finalize orders after receiving.
Thank you!

Marketplaces encourage users to leave reviews

Reviews and listing price are used to estimate a lower bound of revenue

There's **substantial literature** on online anonymous marketplaces, but **little validation**

Background

We show the magnitude of **losses** and **biases** by comparing ground truth versus scraped data

Quantifying losses

We test the effectiveness of various **scraping designs** and coverage estimation techniques through **simulations**

Designing improvements

Previous work lacks methodological consistency and validation
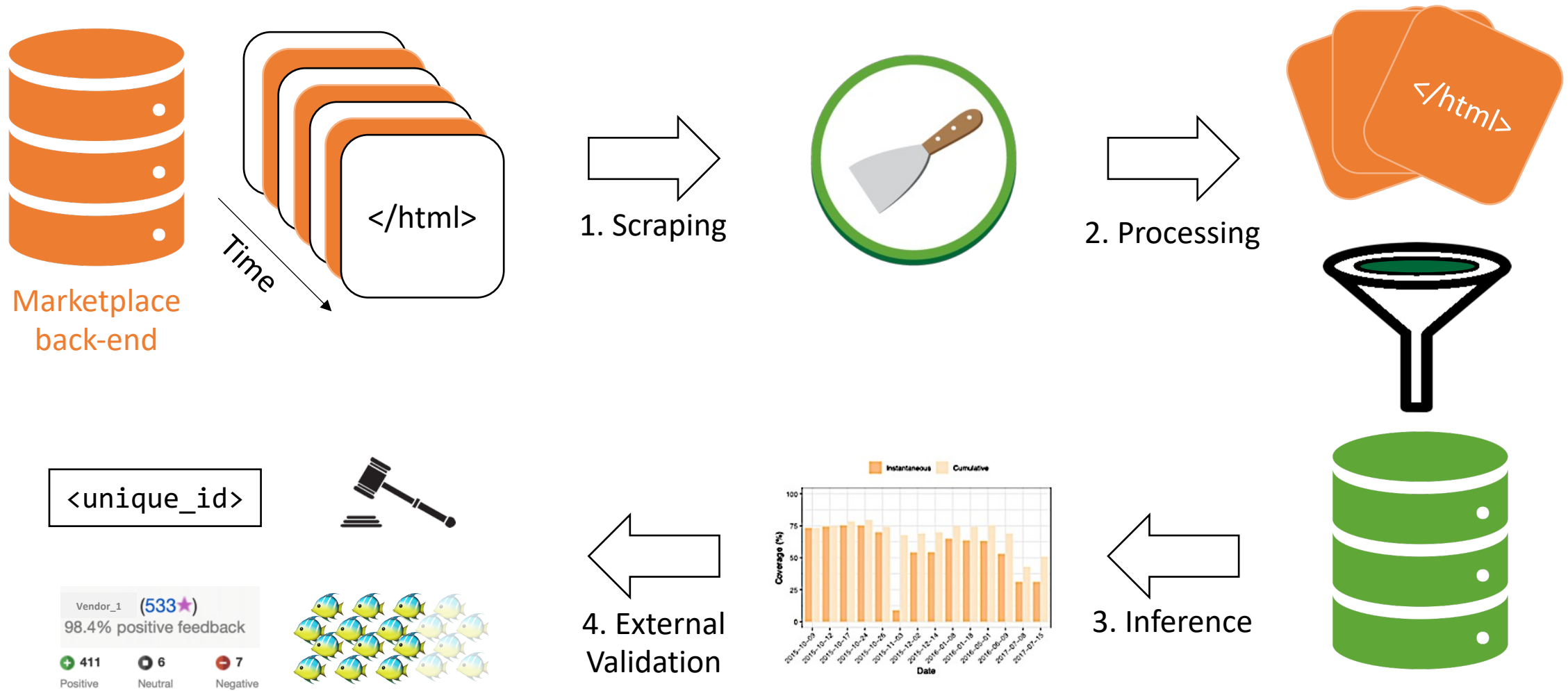
Marketplace back-end

Time

</html>

1. Scraping

2. Processing

</html>

3. Inference

4. External Validation

<unique_id>

Vendor_1 (533★)
98.4% positive feedback

⊕ 411   ⊖ 6   ⊖ 7
Positive Neutral Negative

Previous work lacks methodological consistency and validation

Marketplace back-end

Time

</html>

1. Scraping

Collection loss

2. Processing

</html>

<unique_id>

Vendor_1 (533★)
98.4% positive feedback

411 Positive   6 Neutral   7 Negative

4. External Validation

3. Inference

Previous work lacks methodological consistency and validation

Marketplace
back-end

Time

</html>

1. Scraping

2. Processing

</html>

Inference loss

3. Inference

4. External
Validation

<unique_id>

Vendor_1 (533★)
98.4% positive feedback

411 Positive   6 Neutral   7 Negative
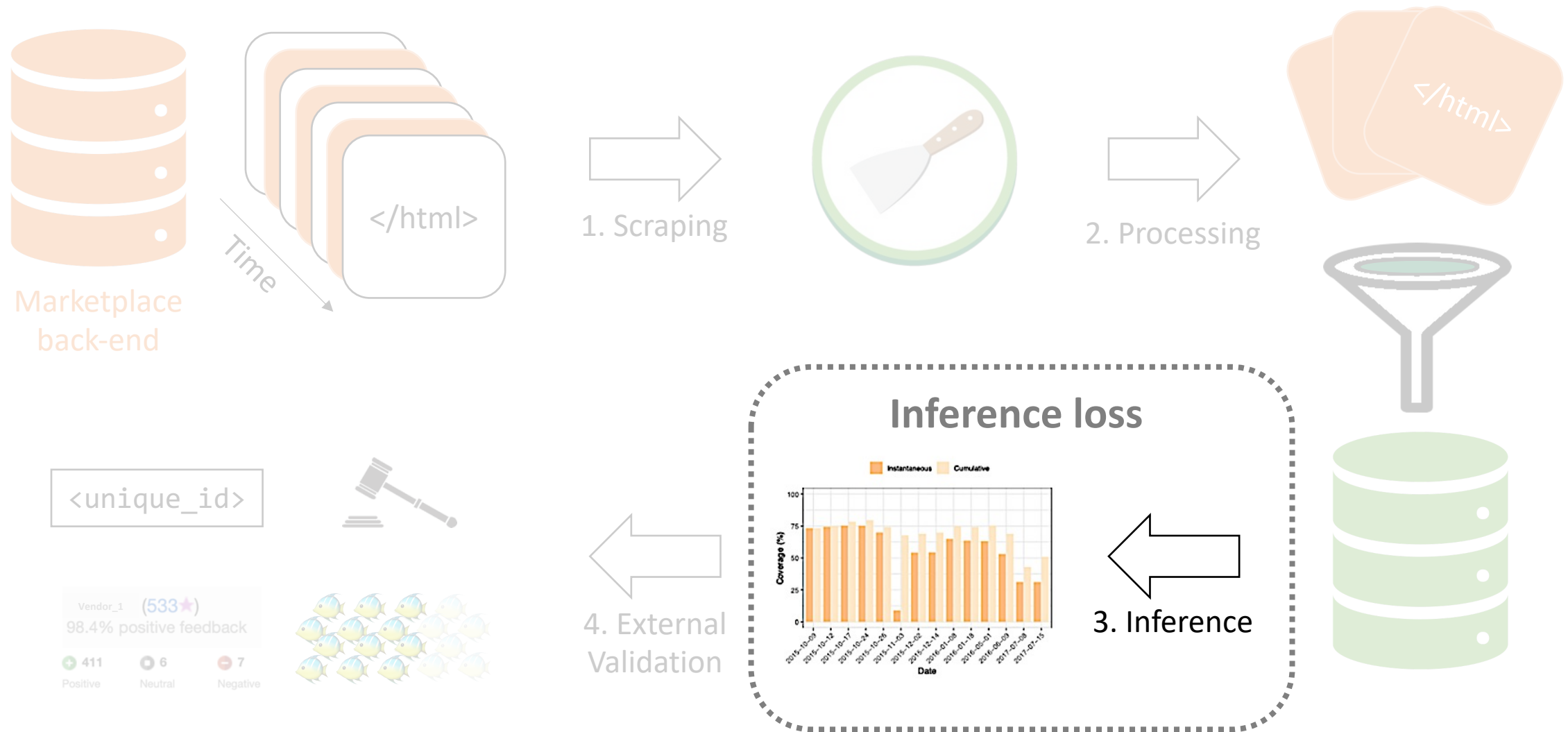
We leveraged scraped and back-end data sources…
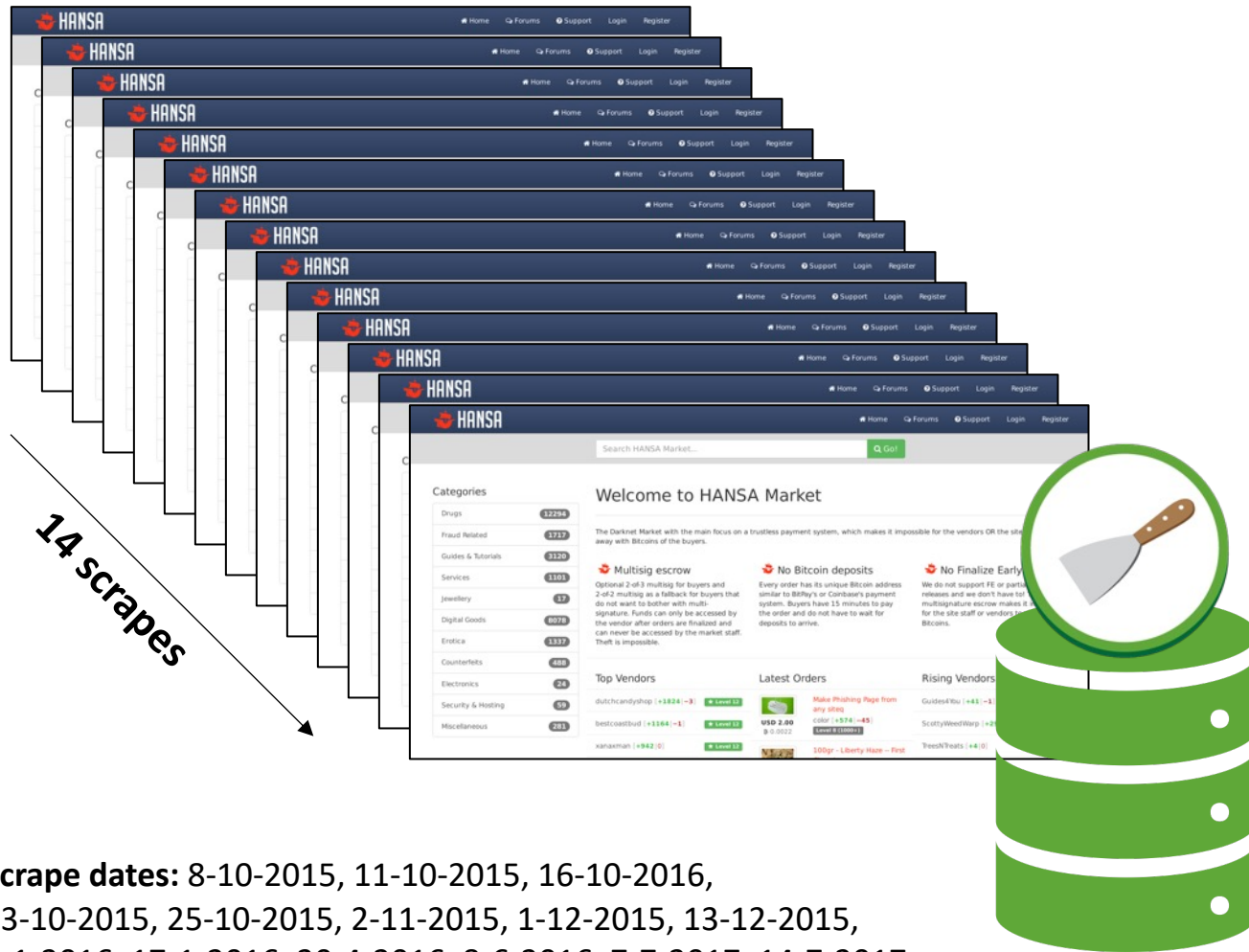
Seized database
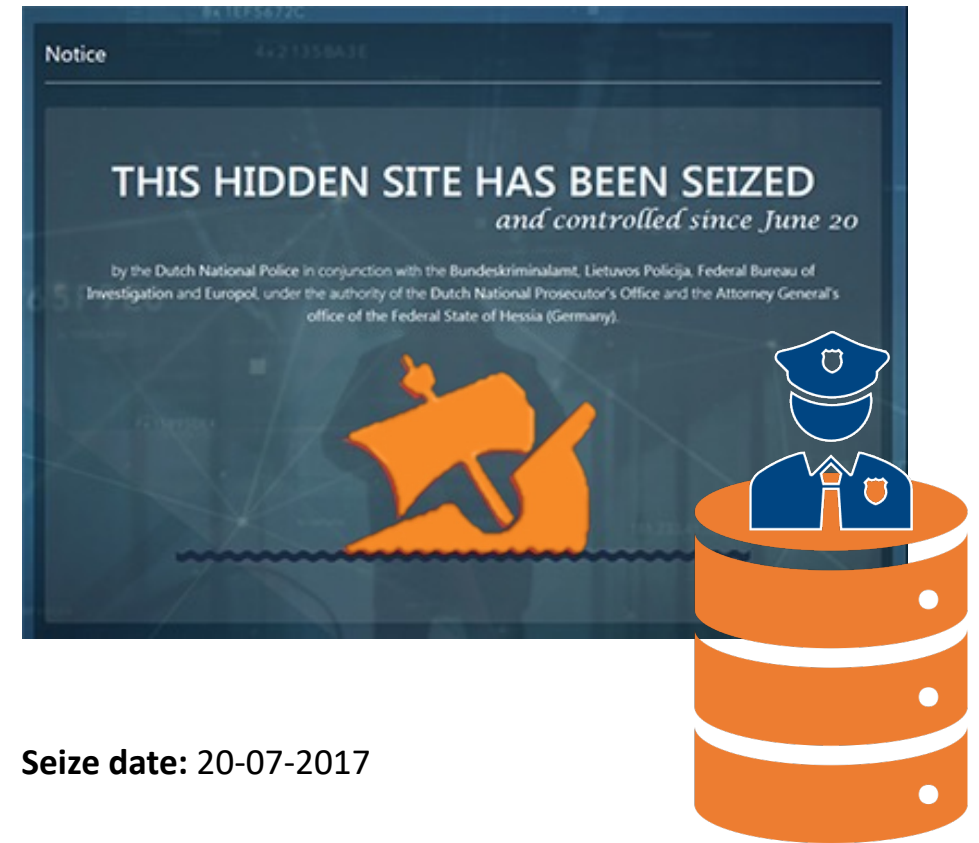
Scrape dates: 8-10-2015, 11-10-2015, 16-10-2016, 23-10-2015, 25-10-2015, 2-11-2015, 1-12-2015, 13-12-2015, 7-1-2016, 17-1-2016, 30-4-2016, 8-6-2016, 7-7-2017, 14-7-2017

Seize date: 20-07-2017

We leveraged scraped and back-end data sources… and calculated coverage

Listings

Reviews

Vendors

54%

Listings

Reviews

Vendors
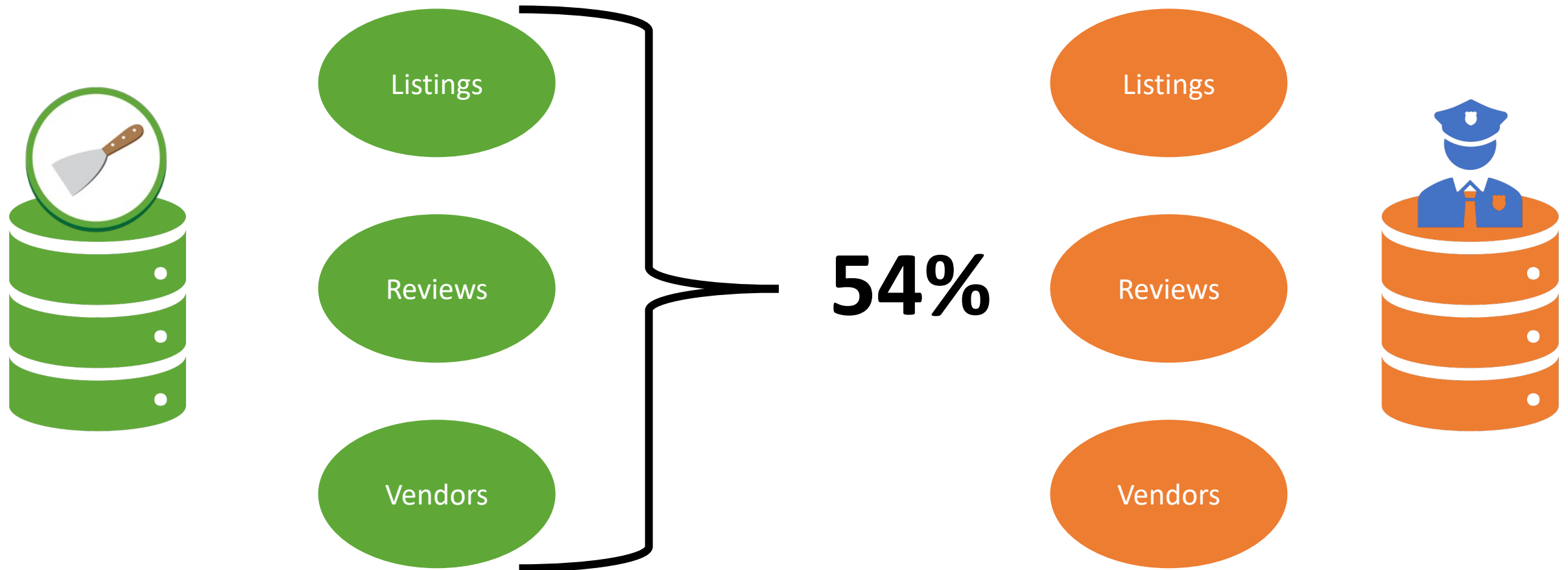
Listings are biased in price, reviews and category (among others)

SCRAPED LISTINGS

Lower priced
$

More reviews
A+++

More **digital goods** listings

NOT-SCRAPED LISTINGS

Higher priced
$$$

Less reviews
A+++

More **weed** listings

Only a quarter of revenue is captured due to collection and inference losses

$50.1 million

$50 million

$45 million

$40 million

$35 million

$30 million

$25 million

$20 million

$15 million

$10 million

$5 million

0

Total revenue

Only a quarter of revenue is captured due to collection and inference losses

$50.1 million

$50 million

$45 million

$40 million

$35 million

$30 million

$25 million

$20 million

$15 million

$ 13.2 million

$10 million

$5 million

Observed revenue

0

Reviews: scraped

Total revenue

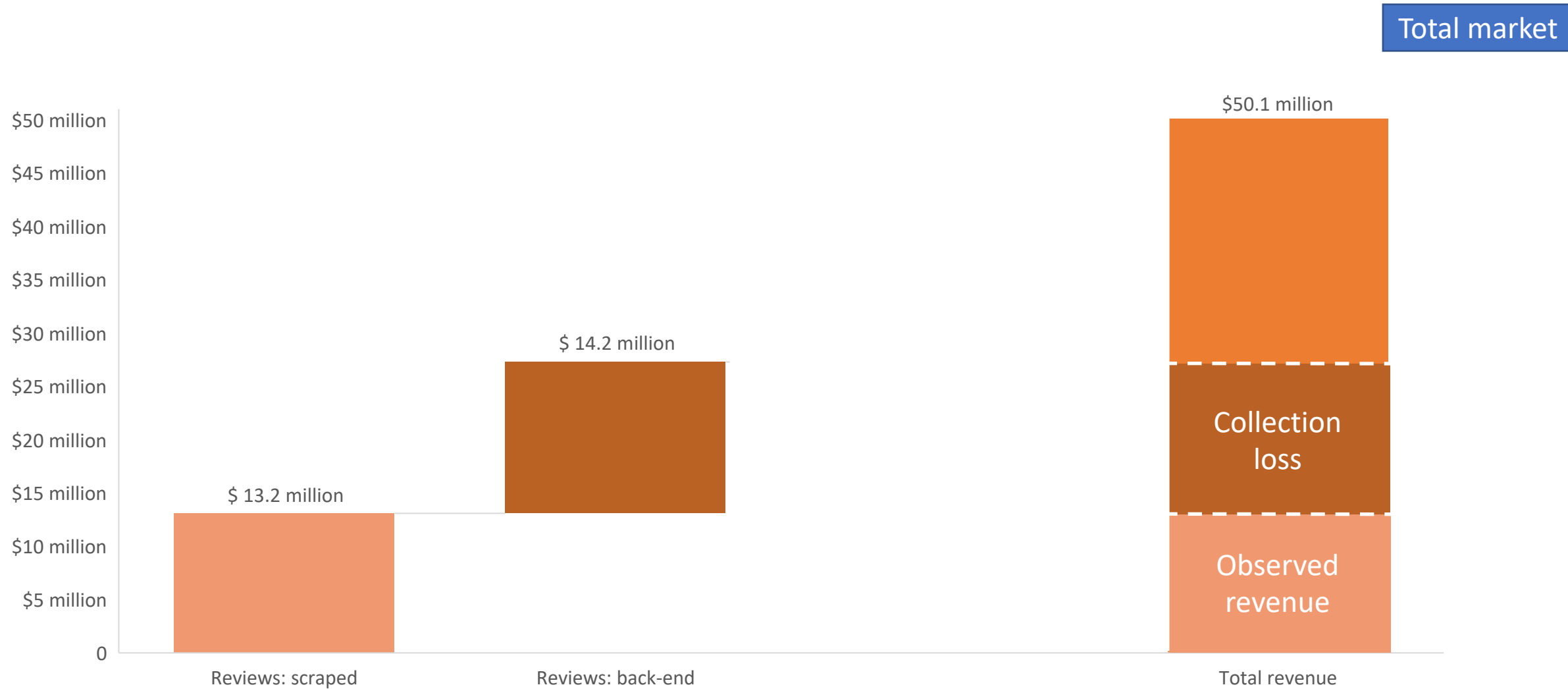# Only a quarter of revenue is captured due to collection and inference losses

Only a quarter of revenue is captured due to collection and inference losses
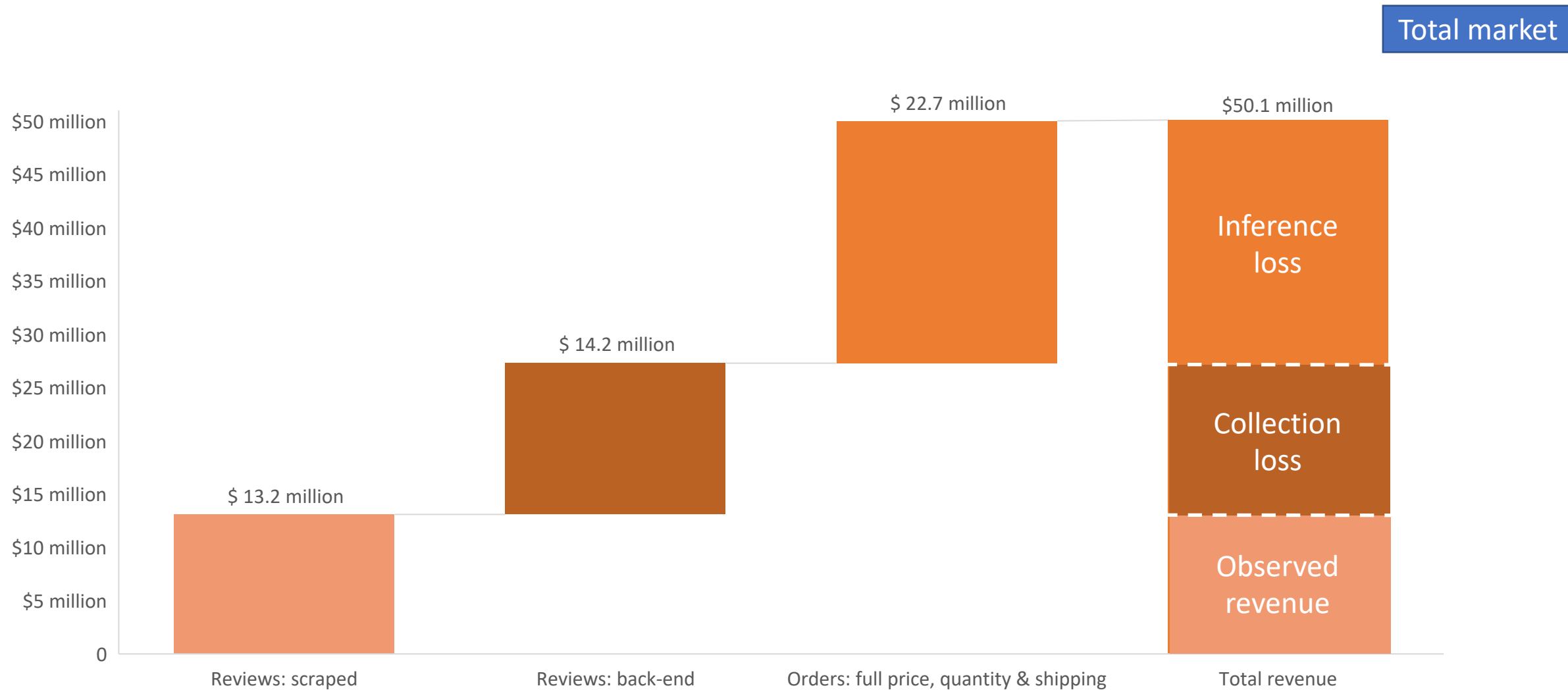
$50 million

$45 million

$40 million

$35 million

$30 million

$25 million

$20 million

$15 million

$10 million

$5 million

0

$ 13.2 million

$ 14.2 million

$ 22.7 million

$50.1 million

Inference loss

Collection loss

Observed revenue

Reviews: scraped

Reviews: back-end

Orders: full price, quantity & shipping

Total revenue

To understand how to improve our collection, we simulated artificial marketplaces and scrapers

of events

Day *n*

Events

New Listing

Probability Space at *n*

To understand how to improve our collection, we simulated artificial marketplaces and scrapers

# of events

Day *n*   Day *n+1*

Events

New Review | New Listing

Probability Space at *n+1*

Events

Days

Artificial Marketplace

Visible Pages

Listings

Reviews
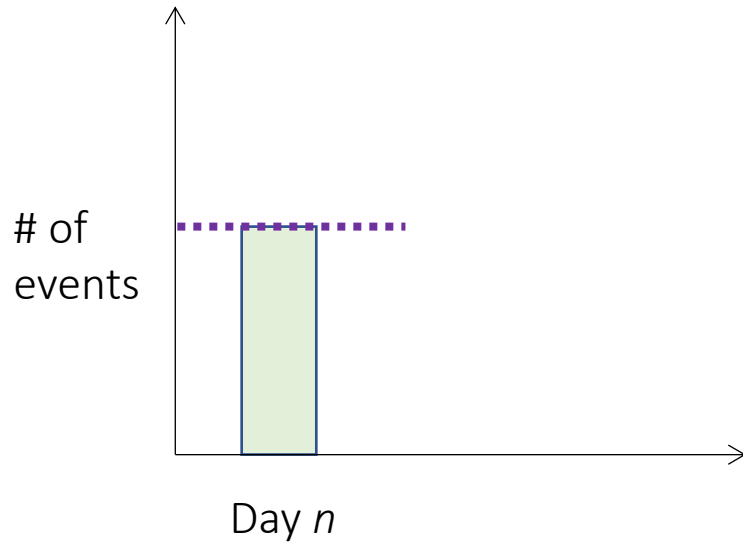
Scraper

Results
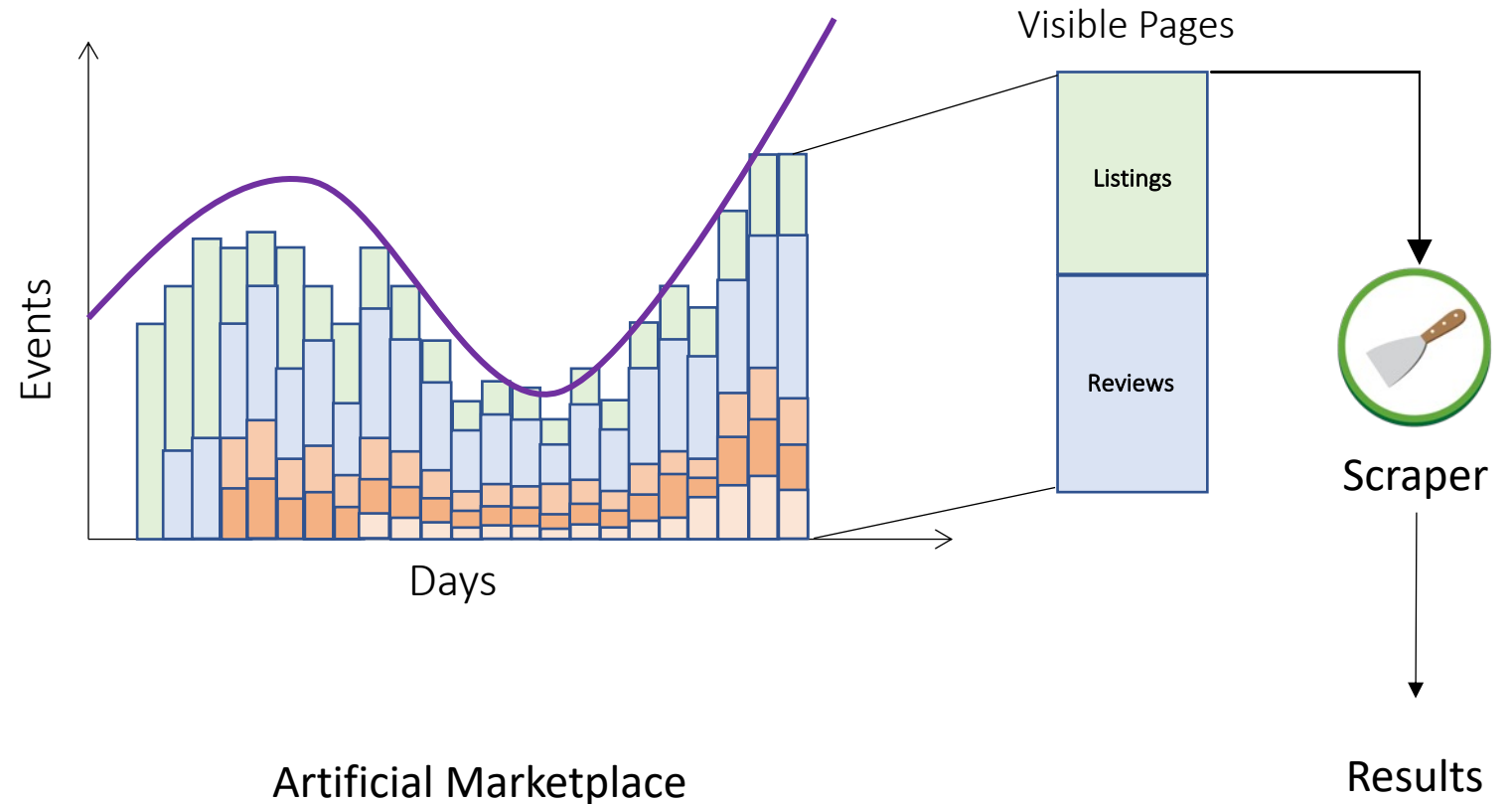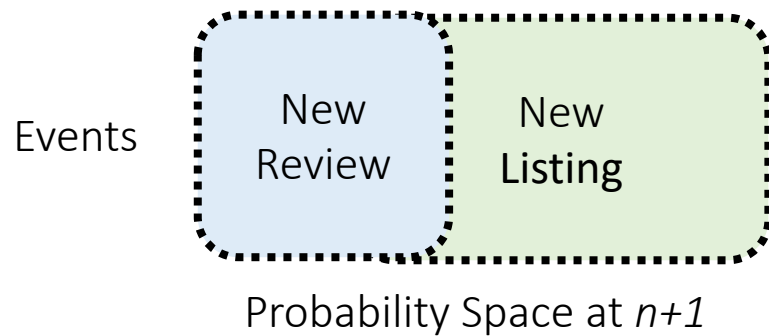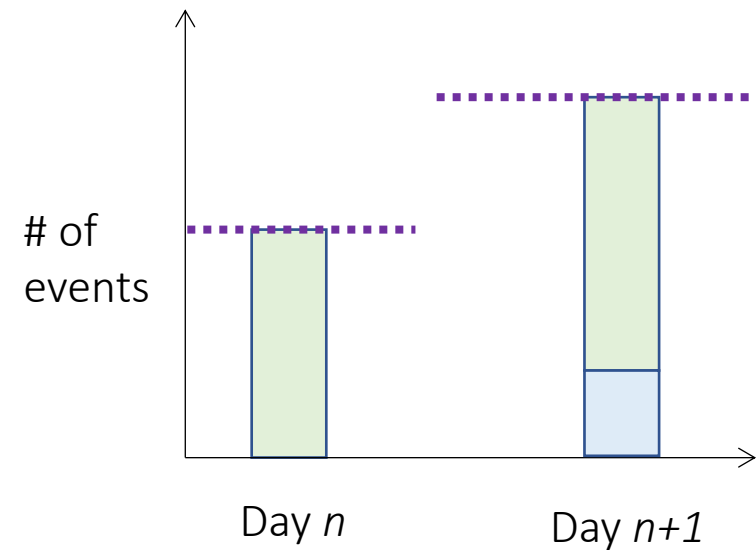
DNM measurement requires stealthy scraping;
good coverage and stealth require frequent scraping
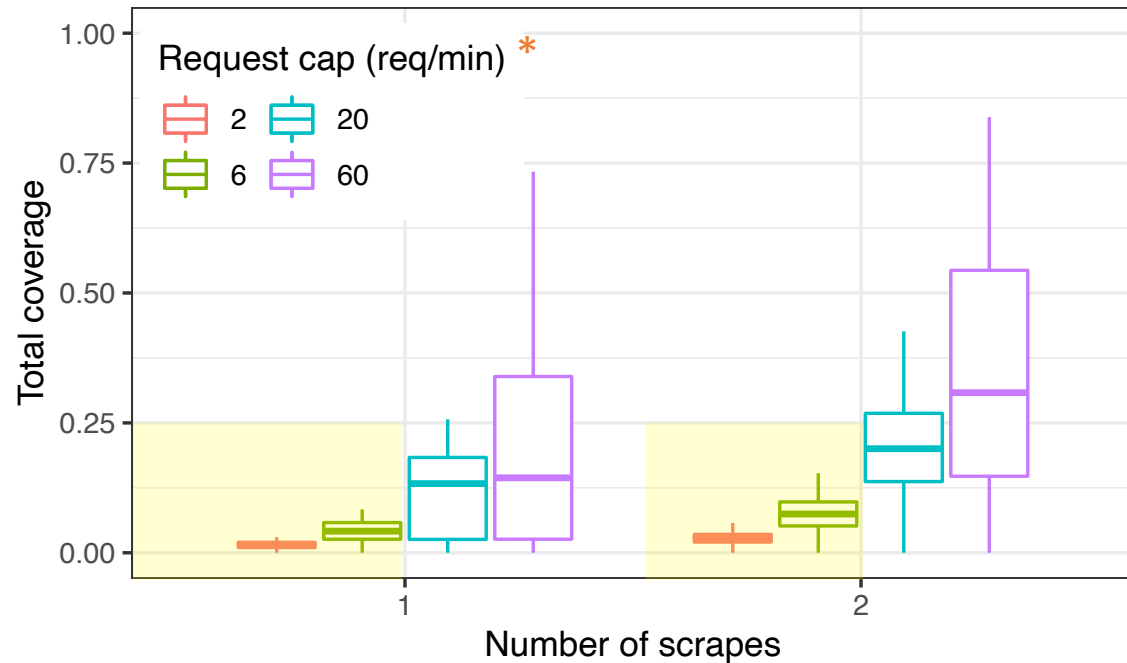
1 request per 30 seconds
1 request per 10 seconds
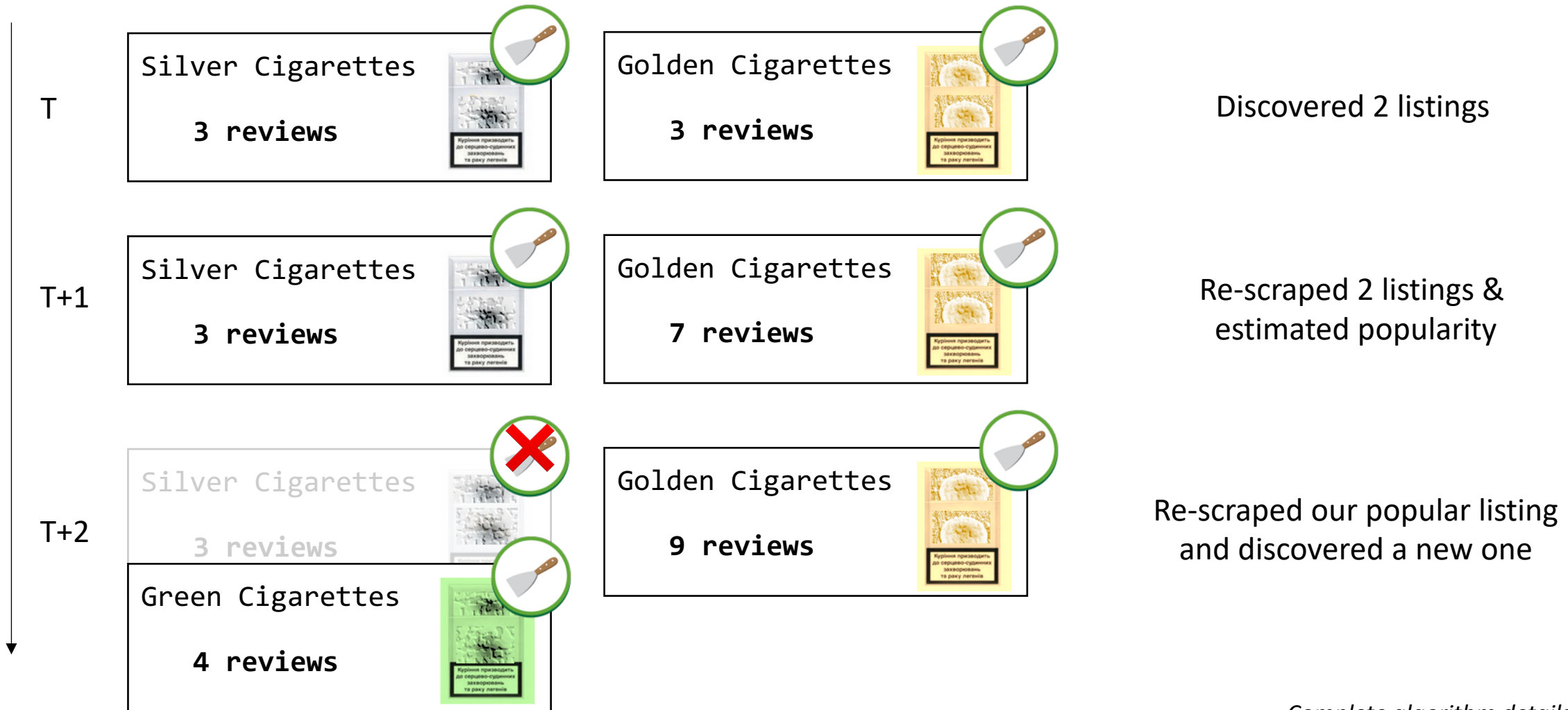1 request per 3 seconds
1 request per 1 second

Stealthy   Overt
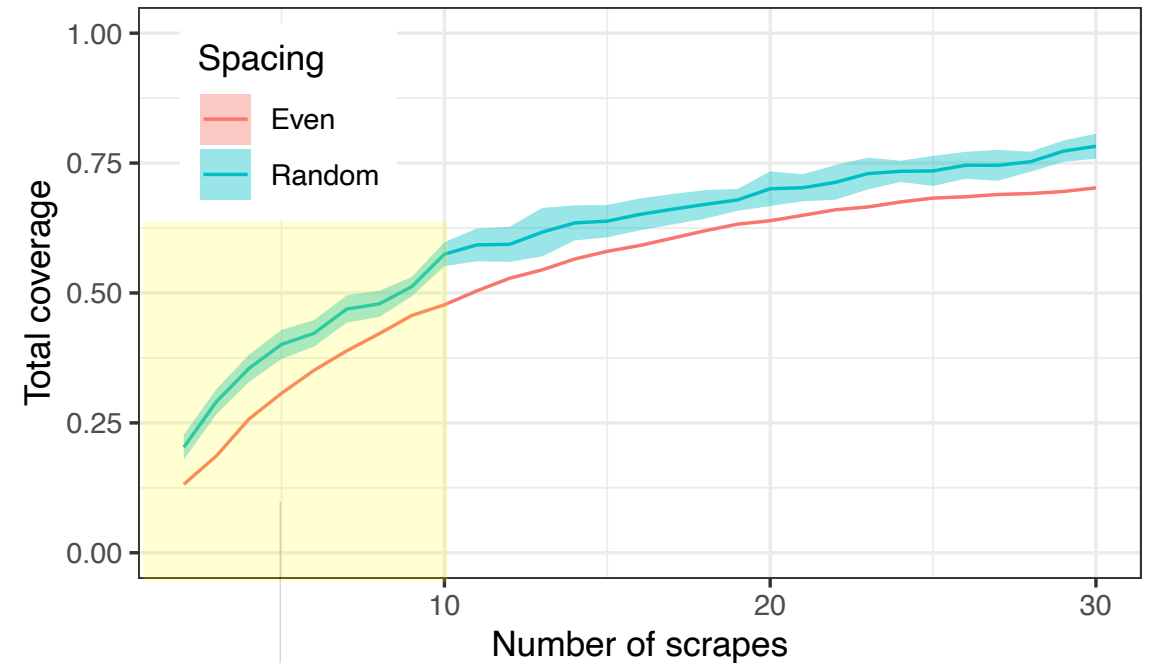
Over a 1-day period

10 stealthy scrapes yielded
53% avg. coverage

Guiding the scraper to more popular listings gained an avg. 9% of coverage

Time

T	Silver Cigarettes — 3 reviews	Golden Cigarettes — 3 reviews	Discovered 2 listings
T+1	Silver Cigarettes — 3 reviews	Golden Cigarettes — 7 reviews	Re-scraped 2 listings & estimated popularity
T+2	Silver Cigarettes — 3 reviews; Green Cigarettes — 4 reviews	Golden Cigarettes — 9 reviews	Re-scraped our popular listing and discovered a new one

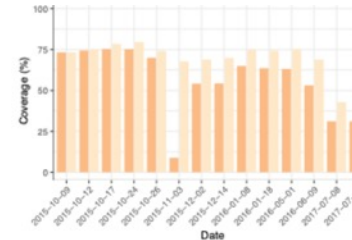*Complete algorithm details in the paper!*

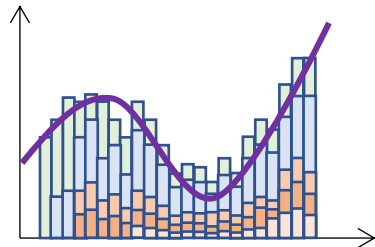# Our study provides a new lens for past studies, and a framework for future studies

(1)

Inform measurement studies

(2)

Better understanding of bias/losses

(3)

Use simulations for analysis

(4)

Better scraping design

(5)

Validation of validation techniques*

Details in the paper!

Alejandro Cuevas
acuevasv@cs.cmu.edu

Fieke Miedema
f.e.g.miedema-1@tudelft.nl

Kyle Soska
soska2@illinois.edu

Nicolas Christin
nicolasc@cmu.edu

Rolf van Wegberg
r.s.vanwegberg@tudelft.nl

Code: https://github.com/aledcuevas/dnm-simulation
Data: https://arima.cylab.cmu.edu/markets/index.php