Spatial Web scraping (aka web harvesting, webcrawling, web data extraction)

GEOG 384

RE Sieber, McGill University

Three ways to find a website's geospatial data

- Downloading a dataset
- Using an API (we'll see that later with streaming data)
- Parsing the contents of a web page(s) using code
- Easy to do by hand but what if you have thousands, millions of pages.

Web Scraping

- Extracts the information you want – whatever you consider to be data – from one or more webpages.
- Involves computer software/code to extract information from websites.
- Is (mostly) automated
- Integrates web-based information into another applications
- Transforms
  - unstructured data into structured data or
  - data that's structured in a way different from the way you need it.

How does it work?

- Set up search criteria -- seed
- Find data (often one record /webpage) on 3rd party website
  - Crawl (use a spider) – request copy of page
  - Follow links to other pages / resources

- Scrape data / page
- Store data

What is Seeding

- You MUST seed the web harvesting software with general search query terms (e.g., “Nova scotia”, “Halifax nova scotia”).
- A search on these broad terms would produce reviews for destinations other than Nova Scotia that had been made by users from Nova Scotia or Halifax, the provincial capital.
- This meant that a substantial amount of manual editing is required to remove reviews for other destinations.

Johnson et al. 2011

What's a Spider/Crawler?

- A program designed to automatically identify and find content on webpages.
- Google Search Engine
- E.g., automatically download all instances of flu and where (geography) they were mentioned – without manually clicking on every one, cutting and pasting
- BTW, NOT FIXED BY AI

Trip Advisor Study

- Use VGI to assist in tourism and economic development
- Mine/ harvest a popular travel VGI site
- Determine the nature (type, assertions) of reviews of Nova Scotia
  - Get a sense of the spatial distribution
  - Sort reviews by themes

Peter A. Johnson, Renee E. Sieber, Nicolas Magnien & Joseph Ariwi. 2011. Automated web harvesting to collect and analyse user-generated content for tourism. *Current Issues in Tourism*

Touristy and Hyped, but Still Lovely

Peggy's Cove Lighthouse

Save Review

JohnTheBear 4,348 contributions
Toronto, Canada

Dec 23, 2008



More photos

Probably the most famous location in Nova Scotia. The lighthouse itself is not especially unusual, but it setting on bare rocks and view of the charmng but also touristy hamlet of Peggy's Cove make it worthwhile. Ground floor contains a post office.

This review is the subjective opinion of a Trip... TripAdvisor LLC.

Was this review helpful? Yes | No

View profile | Send message | Co...

Report Inappropriate Content

Wonderful,whimsical pottery

Lucky Rabbit Pottery

Save Review

lucydogns 1 contribution
Dartmouth,NS

May 3, 2009

I try to make Lucky Rabbit Pottery a destination at least once a year, as I am smitten with their wonderful,whimsical pottery.Lovely bowls,teapots,vases,etc colorfully glazed with nature themes-you may find a chickadee perched on a butter dish,or a sweet little mouse curled up to be a handle for your teapot lid. This is a family-run studio ,and the owners are very friendly and welcoming.In summer you will see some of the most luscious bouquets featuring locally-grown flowers in their own vases,and their garden is well-worth a visit!

...ive opinion of a TripAdvisor member and not of

Yes | No

...d message | Compliment reviewer

...ntent

Surprisingly uncommercial

Peggy's Cove Lighthouse

Save Review

In_the_Know017 18 contributions
Red Deer, AB

Jul 24, 2009



Though it was difficult to get a picture without hordes of tourists scrambling around the lighthouse, the trip was worth it. Tour buses start arriving early and it is often foggy. A scenic town that has tourists shops yet remained colorful and authentic. Some shops and an excellent art gallery.

This review is the subjective opinion of a TripAdvisor member and not of TripAdvisor LLC.

Was this review helpful? Yes | No

View profile | Send message | Compliment reviewer

Report Inappropriate Content

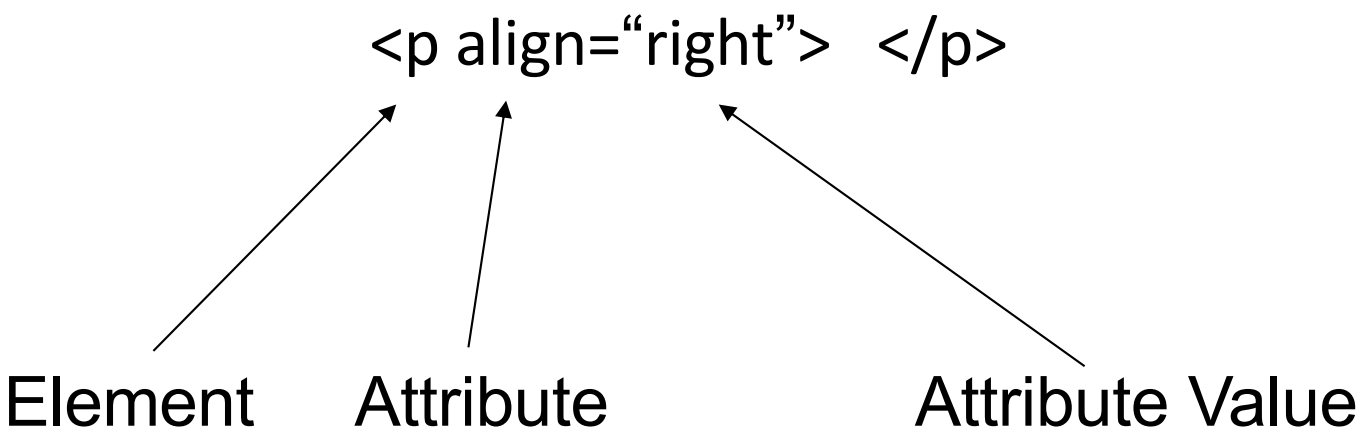
With XPATH we're simulating the seed and using a modified idea of a crawler

Seed assumptions same in XPATHL you're coding exactly what you search for. It breaks if each data source ("webpage") is different (in terms of words used)

You'll see lots of this

Expression	Description
nodename	Selects all nodes with the name "*nodename*"
/	Selects from the root node
//	Selects nodes in the document from the current node that match the selection no matter where they are
.	Selects the current node
..	Selects the parent of the current node
@	Selects attributes

Remember



//p[@class="row"]/a/@href

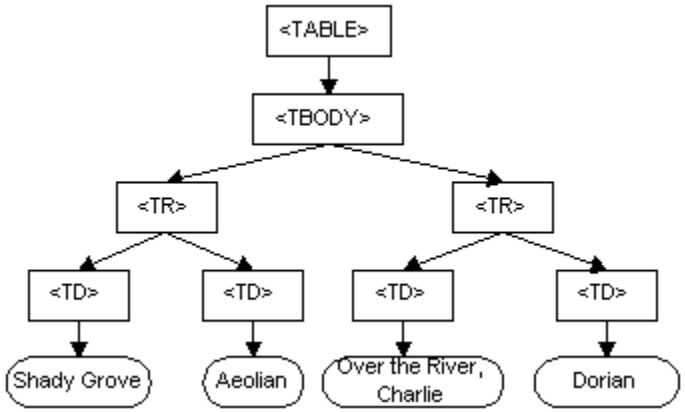
You are working with the DOM

The Document Object Model (DOM) defines the logical structure of documents and the way a document is accessed and manipulated.

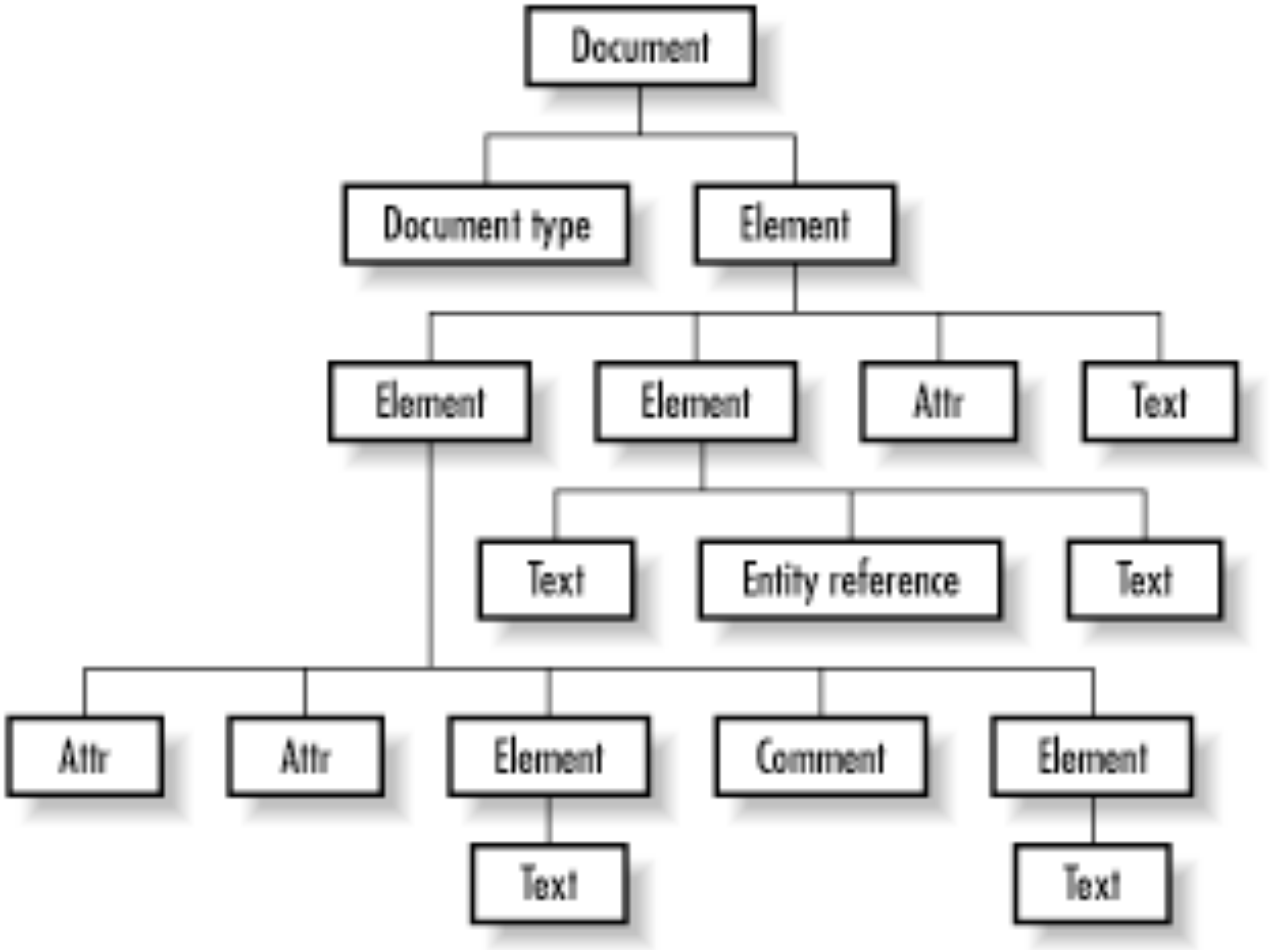
Dom dom dom dom de dom dom de dom

DOM Structure: You traverse a tree of nodes

```
<TABLE>
<TBODY>
<TR>
<TD>Shady Grove</TD>
<TD>Aeolian</TD>
</TR>
<TR>
<TD>Over the River, Charlie</TD>
<TD>Dorian</TD>
</TR>
</TBODY>
</TABLE>
```



XML DOM


Document
Document type
Element
Element
Element
Attr
Text
Text
Entity reference
Text
Attr
Attr
Element
Comment
Element
Text
Text

Issues with Spiders

- Spiders (and other similar types of programs – “robots”, “crawlers”) can be nefarious:
  - appropriating copyrighted materials
  - extracting email addresses for spammers
  - overwhelming servers to create “denial of service”
  - generally violating a site’s “terms of service”
- If you are not careful to use legal and ethical good practices, you can
  - be denied access to a website altogether
  - get yourself or the university sued or even subjected to criminal penalties

What Platforms do in Response

- Specify Terms of Service→
- Protect database rights
- Retaliate against copyright infringement
- Enact technical remedies
  - Embedding data in images, nesting clicks
  - Cloaking
  - Spoofing

- Also, what's legal to scrape and what you're allowed to claim IP
- Check the terms of service

Should Scraping be Legal?

- No
  - RyanAir v. Bravofly (think Expedia)
- Yes
  - Wild west. All data is/should be free!

Should Scraping be Legal?

“I believe that we should be able to decide who is allowed to use our content – permissions are ours to give not other people’s to take.”

“I believe that search engines are essential to our business, in enabling customers to find our content.”

“I am concerned that third parties may seek to monetize exploitation of our digital content without our permission.”

Should Scraping be Legal?

“I believe that a significant element of our revenues will be derived from digital exploitation of our content by 2012.”

“I recognise that an ability to express permissions in a machine-readable form is an essential element of infrastructure for managing relationships in the online supply chain.”

Should Scraping be Legal?

- No
  - RyanAir v. Bravofly
- Yes
  - Wild west. All data is/should be free!

“Scraping Is Just Automated Access, and Everyone Does It”

The *San Francisco Chronicle* used automated web browsing to gather data on Airbnb properties to assess the impact of Airbnb listings on the San Francisco rental market. *ProPublica* used automated web browsing to uncover that Amazon’s pricing algorithm was hiding the best deals from its customers.

https://www.eff.org/deeplinks/2018/04/scraping-just-automated-access-and-everyone-does-it

Techniques to Scrape (a big list from Wikipedia)

- Human copy-and-paste
- Expression matching
- HTML parsers like XPATH
- Web scraping software (e.g., Beautiful Soup)
- Semantic annotation (e.g., metadata)

Criminalizing Web Scraping

- Linkedin v. Doe Defendants http://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=2261&context=historical
- Linkedin sued between 1-100 people who anonymously scraped their website. For
  - Violation of the Computer Fraud and Abuse Act (CFAA).
  - Violation of California Penal Code.
  - Violation of the Digital Millennium Copyright Act (DMCA).
  - Breach of contract.
  - Trespass.
  - Misappropriation.
- In 2017, California fed court disagreed
- https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data

Criminalizing Web Scraping

In 2019 *hiQ Labs, Inc. v. LinkedIn Corp*., the Ninth Circuit Court of Appeals ruled that automated scraping of publicly accessible data likely does not violate the Computer Fraud and Abuse Act (CFAA).

https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data

LinkedIn, however, wasn't done. LinkedIn took the case to the US Supreme Court. The high ruled that since its 2021 decision in Van Buren v. United States showed that the federal computer crime law doesn't criminalize scraping publicly available internet information, the LinkedIn case needed another look. So, SCOTUS sent the case back to the Ninth Circuit. Ninth Circuit reaffirmed in April 2022

“A win for academics, archivists, journalists, researchers, and companies like hiQ that use data that's been made publicly available.” https://www.zdnet.com/article/court-rules-that-data-scraping-is-legal-in-linkedin-appeal/

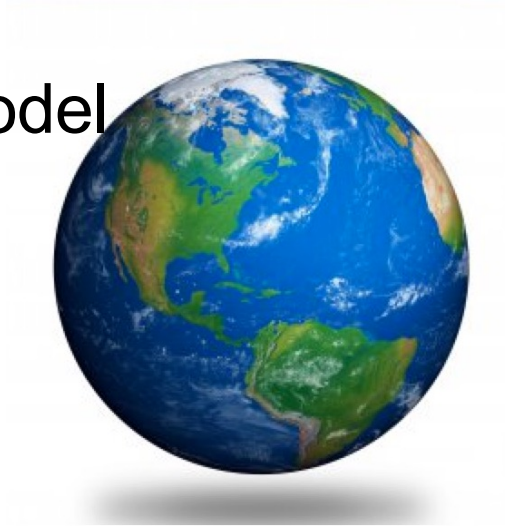
Criminalizing Web Scraping

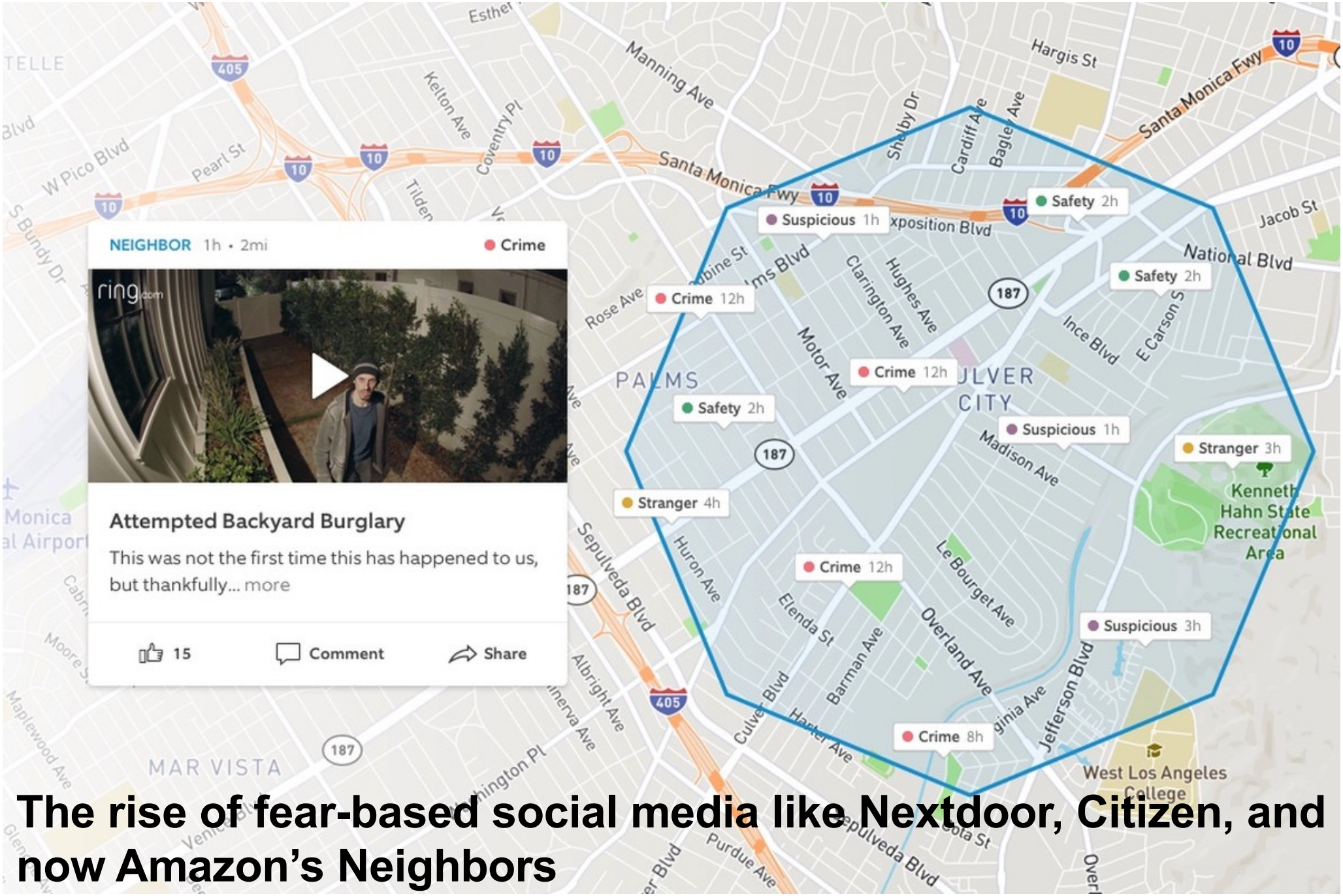
The Van Buren case used a "gates-up-or-down" analogy. Either data is open and the gate is up, or it's not open, and the gate is down. HiQ argued that --on a publicly available website -- that there is no gate to begin with, or at the very least, the gate is up. The Ninth Circuit agreed, ruling that "the concept of 'without authorization does not apply to public websites."

https://www.zdnet.com/article/court-rules-that-data-scraping-is-legal-in-linkedin-appeal/

Geography

- Lat/longs
- Addresses (mult components)
- Points (in screen scraping)
- Attributes
  - placenames
  - characteristics
- Also a question of a viable business model





The rise of fear-based social media like Nextdoor, Citizen, and now Amazon's Neighbors

https://www.vox.com/recode/2019/5/7/18528014/fear-social-media-nextdoor-citizen-amazon-ring-neighbors