

Open Government Data

James Turk

jturk@uchicago.edu

2025-02

Today

- Open Government Data
- Web Scraping

James Turk

Assistant Clinical Professor

UChicago Computational Analysis & Public Policy

Previously: Civic Tech. Sunlight Foundation, PBS, Open States

Eight Principles of Open Government Data

December 2007

1. **Complete:** “all data not subject to valid privacy, security, or privilege limitations”
2. **Primary:** “highest possible level of granularity, not in aggregate”
3. **Timely:** “as quickly as necessary to preserve value”
4. **Accessible:** “to the widest range of users for the widest range of purposes”
5. **Machine processable:** “structured to allow automated processing”
6. **Non-discriminatory access:** “available to anyone”
7. **Non-proprietary:** “in a format over which no entity has exclusive control”
8. **License-free:** “not subject to any copyright, patent, trademark...”

Open Government EveryBlock

- Launched in January 2008
- From the creator of one of the first Google Maps/data mashups.
- Combined public data with hyperlocal news feeds.

The screenshot shows the EveryBlock Seattle website. At the top, there's a navigation bar with 'Browse: Public records', 'Articles', 'More', and 'Explore: Neighborhoods'. A search bar on the right says 'Find: Address, ZIP or neighborhood'. Below the navigation bar, the main header features the 'EveryBlock Seattle' logo and the tagline 'A news feed for your block.'.

The main content area is divided into several sections:

- Enter your address, find news nearby:** A section with a search input field labeled 'Address, ZIP or neighborhood' and a 'Search Seattle' button. Below it, a note says 'Seattle only, for now. Examples: 4000 Woodlawn Ave. N, 98133, or Capitol Hill'.
- Explore the city:** A section showing statistics: '78 Neighborhoods', '27 ZIP Codes', and '1,456 Streets'. It also includes a 'NEW HERE? READ OUR FAQ' link.
- FEATURED NEIGHBORHOOD:** A section titled 'Capitol Hill' with a small map icon. It lists recent articles and public records for that area. The first article is 'Broadway Grocery shuts down' from 'CHS Capitol Hill Seattle Blog' dated 'September 18', mentioning a store at the corner of Harrison and Broadway. The second article is 'CDNews Police Scanner - 9/17' from 'Central District News' dated 'September 18', reporting a disturbance at 321 Broadway E. The third article is '1-night stand for "Wizard," and a fantastic film fest' from 'The Seattle Times' dated 'September 18', mentioning a show at 805 E. Pine St. At the bottom of this section, it says '179 more articles in Capitol Hill in the last 30 days.'
- LOCAL MEDIA:** A section titled 'Which neighborhoods are in the news?' featuring a map of Seattle with red circles indicating active news locations. At the bottom, it says '2,077 more articles in the last 30 days.'

Open Data

pre-2009

- **1974** FEC Campaign Finance Data, digital storage, available via microfiche
- **1989** First public release of Census TIGER/Line data
- **2000** GPS made fully open for civilian use
- **2005-2007** First open data mashups: crime maps, legislative trackers, etc.
- **2007** DC Open Data Policy, precursor to federal effort

US Open Data

2009-2024

- **2009** Presidential Memorandum on Transparency & Open Government, data.gov, & Open Government Directive
- **2012** US & India launch Open Government Platform (portal for open data sites)
- **2013** GPS made fully open for civilian use
- **2019** OPEN Government Data Act signed into law (requires data.gov by statute)
- ~40 states have similar policies portal in place as of 2024.

US Open Data vs. International Data

Key Differences

- Many countries have their own open data initiatives,
- EU Open Data Directive (2019) (replaces 2012 PSI Directive)
- UK: data.gov.uk & Open Government License (2010)
- India: data.gov.in (2012)
- Mexico: General Law of Transparency (2015)

US Open Data vs. International Data

Key Differences

One major boon to US open data is that the data is not subject to any license restrictions. Data may generally be used for *any purpose*.

All US Federal data is in the **public domain**.

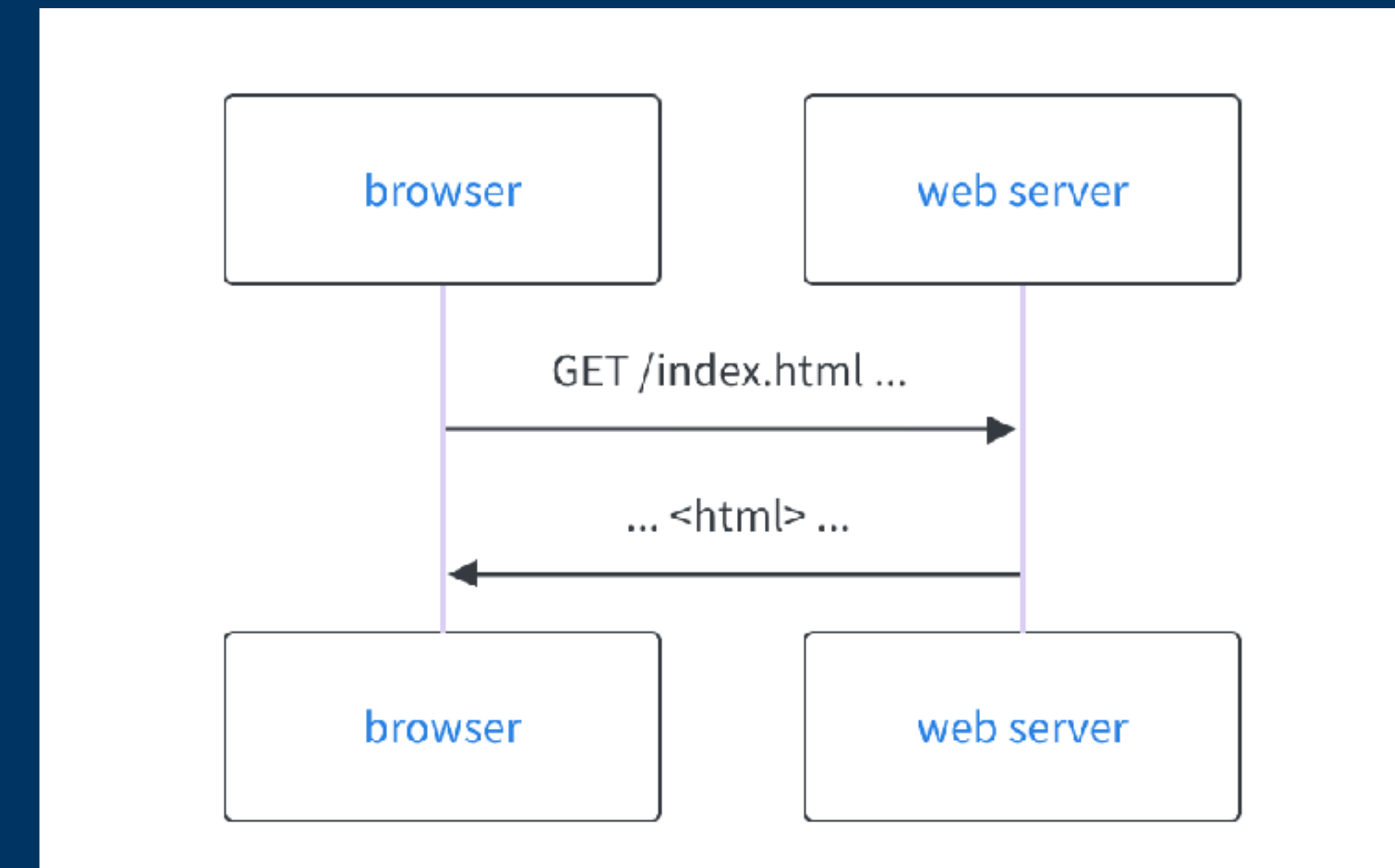
Further, copyright does not apply to factual data under US law.

Other countries however recognize **database rights** and often assert ownership over the data itself.

Be sure to familiarize yourself with the particular restrictions/limitations.

Web Scrapping

- The web is made up of HTTP requests & responses.
- Every page you access is your browser exchanging a message with a web server.
- Instead of making these requests with a web browser, we can make them with ***our code*** and extract the information we need.
- We'll need tools to:
 - make HTTP requests
 - parse HTML
 - extract specific information



Eight Principles of Open Government Data

December 2007

1. **Complete:** “all data not subject to valid privacy, security, or privilege limitations”
2. **Primary:** “highest possible level of granularity, not in aggregate”
3. **Timely:** “as quickly as necessary to preserve value”
4. **Accessible:** “to the widest range of users for the widest range of purposes”
5. **Machine processable:** “structured to allow automated processing”
6. **Non-discriminatory access:** “available to anyone”
7. **Non-proprietary:** “in a format over which no entity has exclusive control”
8. **License-free:** “not subject to any copyright, patent, trademark...”

Web Scrapping & Open Data

Data is often made available only on a webpage, PDF, or proprietary format.

We want **structured data**: CSV, JSON, GeoJSON, etc.

It is possible to scrape any information that is available over the web, *if the browser can access it, so can we.*



Web Scrapping Tools

Make HTTP Requests: `httpx`, `requests`


Parse HTML: `lxml.html`, `BeautifulSoup`

Extract Content: `CSS selectors` / `XPath`

Automate Complex Browser Interaction: `playwright`, `selenium`

Finding Data

Web Scrapping Example



UNITED STATES SENATE

SENATORS

COMMITTEES

LEGISLATION & RECORDS

ART & ARTIFACTS

ABOUT THE SENATE

Contact

Leadership & Officers

Former Senators

Qualifications & Terms of Service

Facts & Milestones

States

Senate Salaries (1789 to Present)

PDF

Search:

Years	Salary
1789–1815	\$6.00 per diem
1815–1817	\$1,500 per annum
1817–1855	\$8.00 per diem
1855–1865	\$3,000 per annum
1865–1871	\$5,000 per annum
1871–1873	\$7,500 per annum
1873–1907	\$5,000 per annum
1907–1925	\$7,500 per annum
1925–1932	\$10,000 per annum
1932–1933	\$9,000 per annum

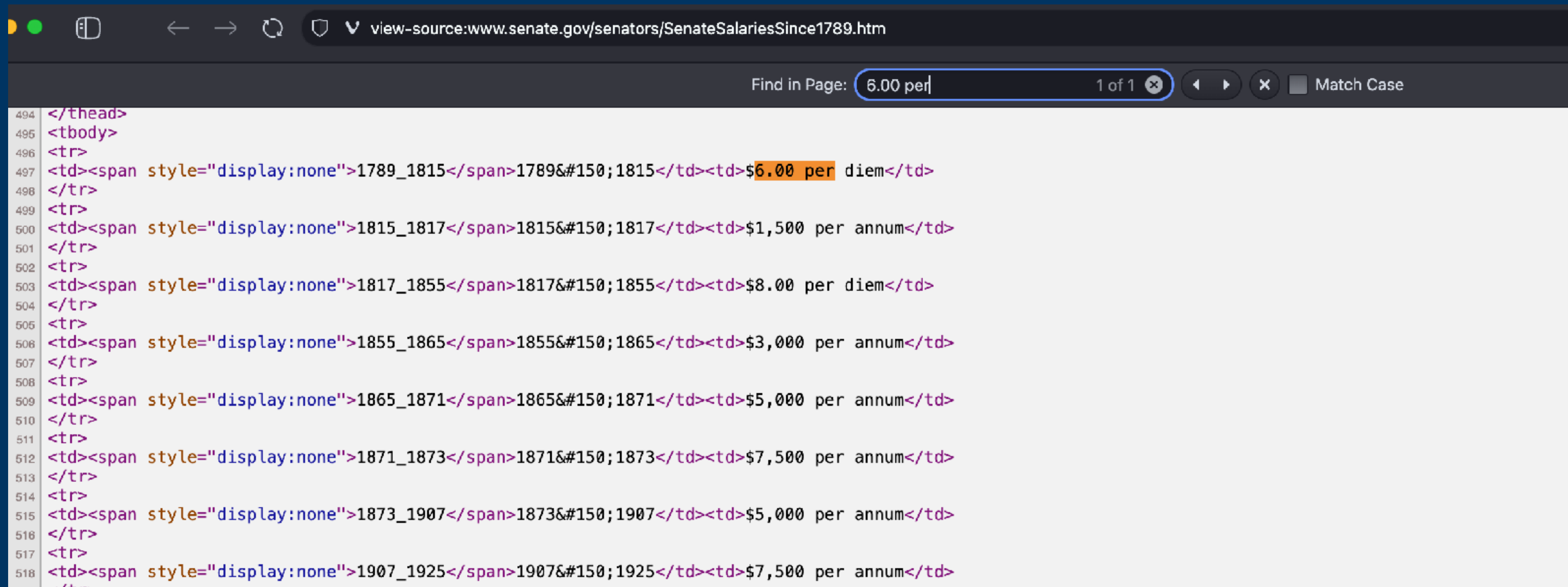
Related Reports

Congressional Salaries and Allowances (CRS) (pdf)

Retirement Benefits for Members of (CRS) (pdf)

Examining HTML

Web Scraping Example



```
494 </thead>
495 <tbody>
496 <tr>
497 <td><span style="display:none">1789_1815</span>1789&#150;1815</td><td>$6.00 per diem</td>
498 </tr>
499 <tr>
500 <td><span style="display:none">1815_1817</span>1815&#150;1817</td><td>$1,500 per annum</td>
501 </tr>
502 <tr>
503 <td><span style="display:none">1817_1855</span>1817&#150;1855</td><td>$8.00 per diem</td>
504 </tr>
505 <tr>
506 <td><span style="display:none">1855_1865</span>1855&#150;1865</td><td>$3,000 per annum</td>
507 </tr>
508 <tr>
509 <td><span style="display:none">1865_1871</span>1865&#150;1871</td><td>$5,000 per annum</td>
510 </tr>
511 <tr>
512 <td><span style="display:none">1871_1873</span>1871&#150;1873</td><td>$7,500 per annum</td>
513 </tr>
514 <tr>
515 <td><span style="display:none">1873_1907</span>1873&#150;1907</td><td>$5,000 per annum</td>
516 </tr>
517 <tr>
518 <td><span style="display:none">1907_1925</span>1907&#150;1925</td><td>$7,500 per annum</td>
519 </tr>
```

Inspector

Web Scraping Example

- [Live Demo](#)

Making a Request

Web Scraping Example

```
import httpx
```

```
response = httpx.get("https://www.senate.gov/senators/SenateSalariesSince1789.htm")
```

Parsing HTML to a tree

Web Scraping Example

```
import httpx
```

```
import lxml.html
```

```
response = httpx.get("https://www.senate.gov/senators/SenateSalariesSince1789.htm")
```

```
root = lxml.html.fromstring(response.text)
```

```
# root is now a tree structure representing the entire page
```

Navigating Information on the Page

Web Scraping Example

```
root = lxml.html.fromstring(response.text)
```

```
root.getchildren()[0].getchildren[0].getchildren()[...]
```


Navigating Information on the Page

Web Scraping Example

```
rows = root.cssselect("#SortableData_table tbody tr")
```

```
for row in rows:
```

```
    # this time we'll iterate over the <td> elements within
```

```
    # since we are starting the .cssselect with `row` instead of `root`
```

```
    # this only gets the <td>s within the current row
```

```
    year_td, salary_td = row.cssselect("td")
```

```
    # finally, we use .text_content() to extract the text nodes
```

```
    # which contain the data we're after
```

```
    year = year_td.text_content()
```

```
    salary = salary_td.text_content()
```

Web Scrapping: GIS

- Scrapping GIS data poses a particular challenge due to the file sizes.
- Approaches:
 - Access data visible in the underlying HTML & JSON. Often you will encounter web requests for GeoJSON.
 - May need to make thousands of requests with different geographic boundaries to scrape all points on a large map. (Example: create grid with squares of maximum map size, and iterate across entire region of interest.)

Legal & Ethical Reminders

- Within the US, the Computer Fraud and Abuse Act (CFAA) kicks in for **any circumvention of a security measure**. This can be interpreted very broadly but generally does not apply to information open to the public. [1]
- Be a good neighbor and do not harm anyone else's access. If you are writing a scraper that hits a site, do not scrape faster than a person would access pages. ~1/second is a good starting point.
- Remember: rules vary widely internationally.

[1] https://en.wikipedia.org/wiki/HiQ_Labs_v._LinkedIn