Digital Trace Data

Jacob Habinek

LINKÖPING UNIVERSITY

Institute for Analytical Sociology

Day 2: digital trace data

- Morning lecture
  - What is digital trace data?
  - Strengths and weaknesses
  - Collecting digital trace data
  - Screen-scraping
  - APIs

- Afternoon workshop
  - Screen-scraping
  - APIs

What is digital trace data?

An interdisciplinary field that advances theories of human behavior by applying computational techniques to large datasets from social media sites, the Internet, or other digitized archives such as administrative records."

From Edelmann *et al.* (2020: 15)

"CSS and sociology" https://doi.org/10.1146/annurev-soc-121919-054621

Article

The scales of human mobility

https://doi.org/10.1038/s41586-020-2909-1

Received: 3 February 2020

Accepted: 25 September 2020

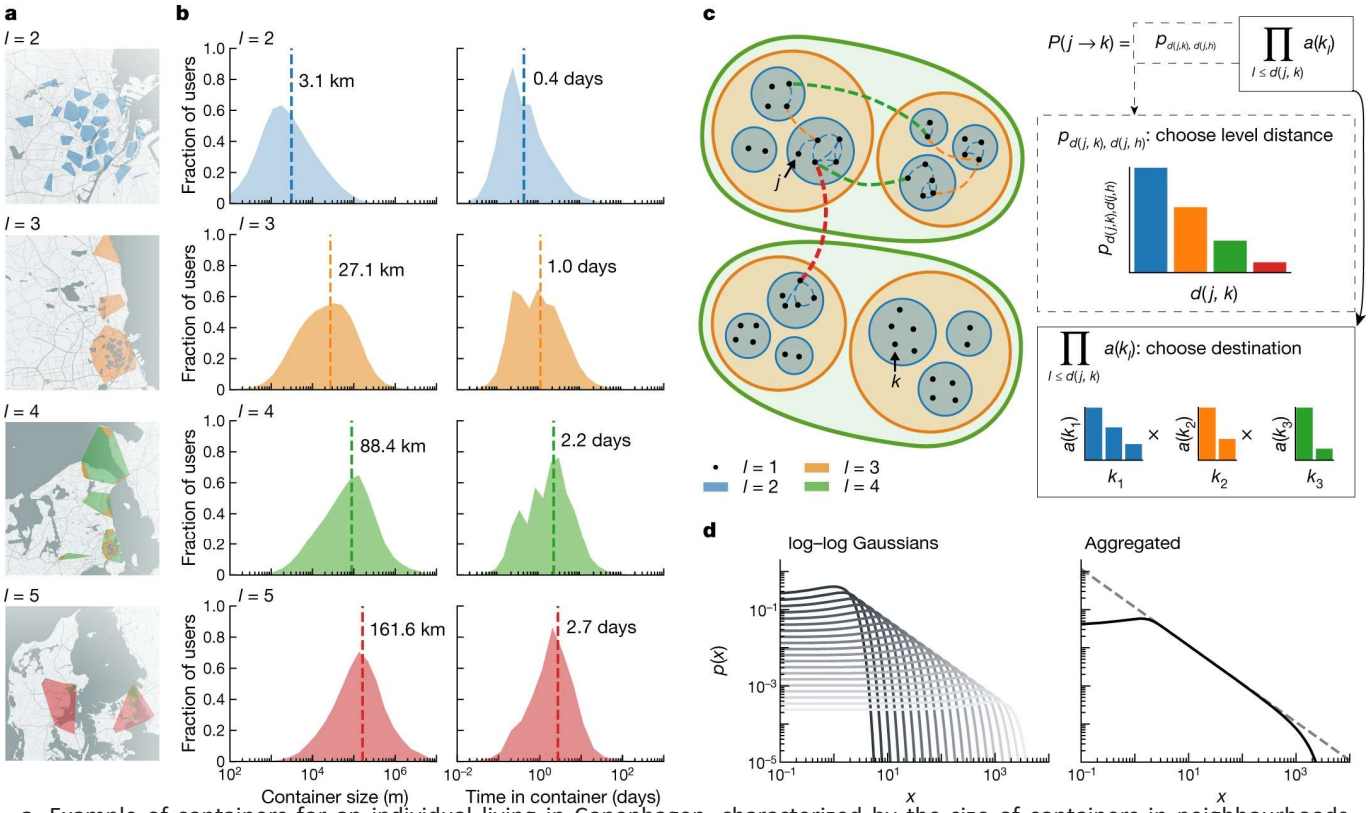
Published online: 18 November 2020

Check for updates

Laura Alessandretti[1,2,3], Ulf Aslak[1,2,3] & Sune Lehmann[1,2]✉

There is a contradiction at the heart of our current understanding of individual and collective mobility patterns. On the one hand, a highly influential body of literature on human mobility driven by analyses of massive empirical datasets finds that human movements show no evidence of characteristic spatial scales. There, human mobility is described as scale free[1–3]. On the other hand, geographically, the concept of scale—referring to meaningful levels of description from individual buildings to neighbourhoods, cities, regions and countries—is central for the description of various aspects of human behaviour, such as socioeconomic interactions, or political and cultural dynamics[4,5]. Here we resolve this apparent paradox by showing that day-to-day human mobility does indeed contain meaningful scales, corresponding to spatial 'containers' that restrict mobility behaviour. The scale-free results arise from aggregating displacements across containers. We present a simple model—which given a person's trajectory—infers their neighbourhood, city and so on, as well as the sizes of these geographical containers. We find that the containers—characterizing the trajectories of more than 700,000 individuals—do indeed have typical sizes. We show that our model is also able to generate highly realistic trajectories and provides a way to understand the differences in mobility behaviour across countries, gender groups and urban–rural areas.

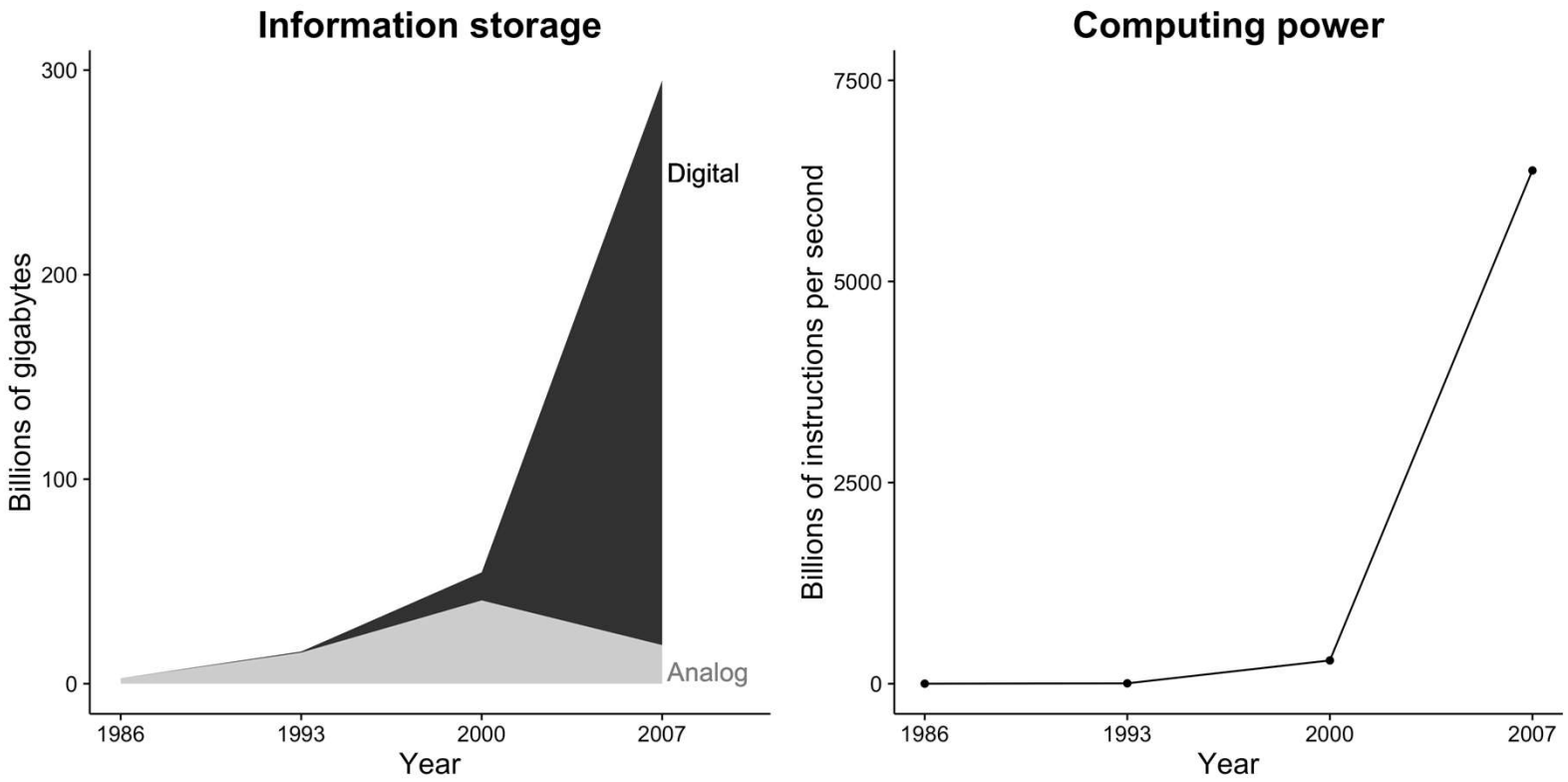
https://doi.org/10.1038/s41586-020-2909-1



a, Example of containers for an individual living in Copenhagen, characterized by the size of containers in neighbourhoods (blue), cities (orange), urban agglomerations (green) and regions (red). Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org. **b**, Distribution of container sizes (left) and median time spent in the same container (right) across individuals. Dashed lines correspond to medians. Results, shown here for containers at different hierarchical levels, are obtained by fitting the container model to the D1 dataset, consisting of approximately 700,000 anonymized GPS traces of individuals distributed across the globe (see Extended Data Fig. 2 for dataset D2). **c**, Schematic representation of the container model. Individuals move between locations (black dots) inside a nested set of containers. The probability of transitioning between two locations j and k is the product of two factors, corresponding to choosing level distance and destination (see main text). **d**, Gaussian distributions with different variances (left) and their mixture (right) on a log–log scale. The dashed line (right) is a power law $P(x) \approx x^{-\beta}$ with variable of interest x and exponent $\beta = 1$ to guide the eye.

Strengths and weaknesses

Big



Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital. These changes create incredible opportunities for social researchers. Adapted from Hilbert and López (2011), figures 2 and 5.

Always on

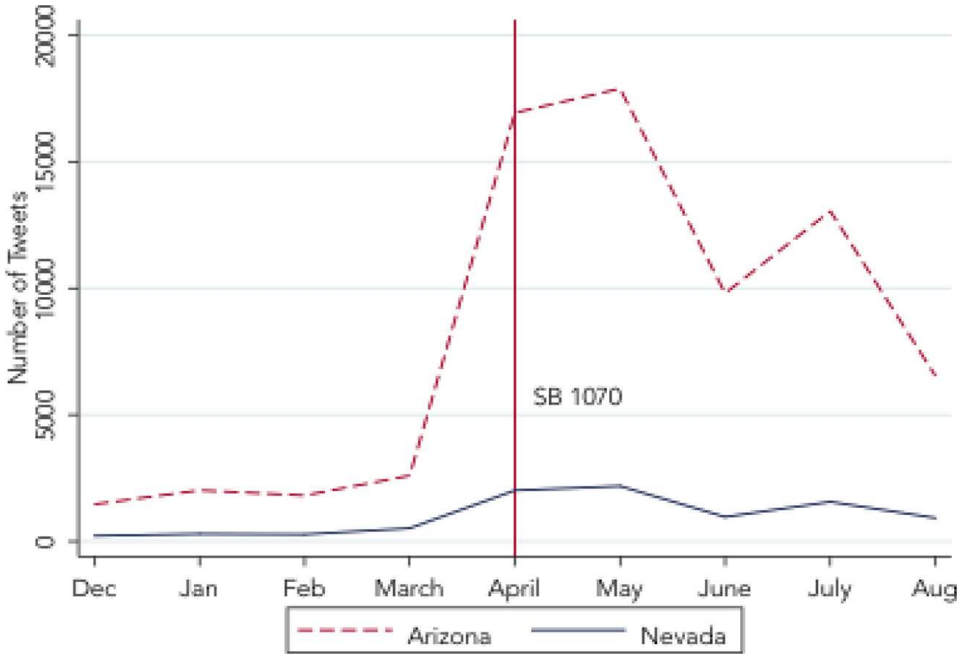


Fig. 2. Number of Twitter messages related to immigrants per month in Arizona and Nevada (December 2010–August 2011). The vertical line on April 2010 indicates when the Arizona governor approved SB 1070.

Non-reactive

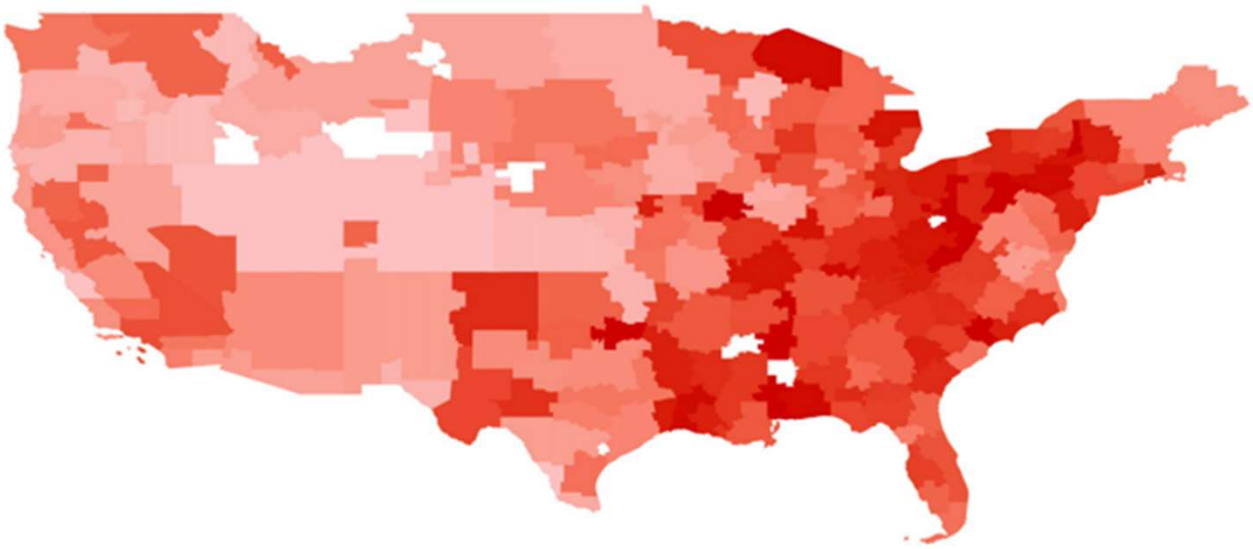


Fig. 2. Racially charged search rate, media market. Notes: This maps search volume for "[Word 1](s)," from 2004 to 2007, at the media market level. Darker areas signify higher search volume. White areas signify media markets without data. Alaska and Hawaii, for which data are available, are not shown

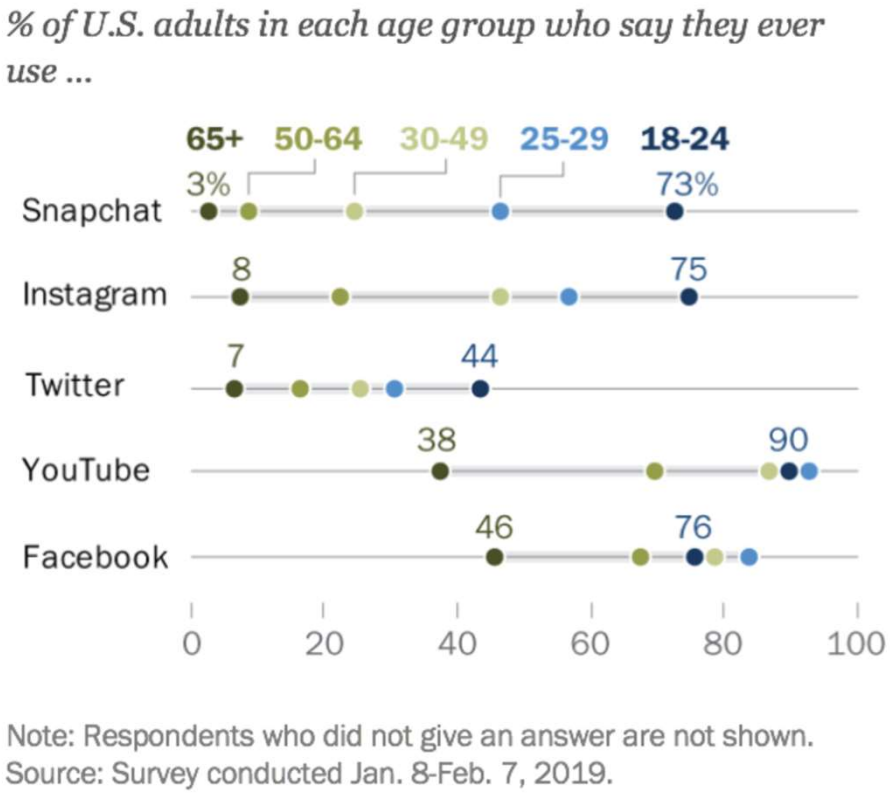
"The cost of racial animus on a black candidate" https://doi.org/10.1016/j.jpubeco.2014.04.010

Incomplete


Like · Comment · Share
7 people like this.
This comment has been removed. You can Undo this, Report it as abusive, or Ban [redacted].
Write a comment...

Inaccessible

Non-representative



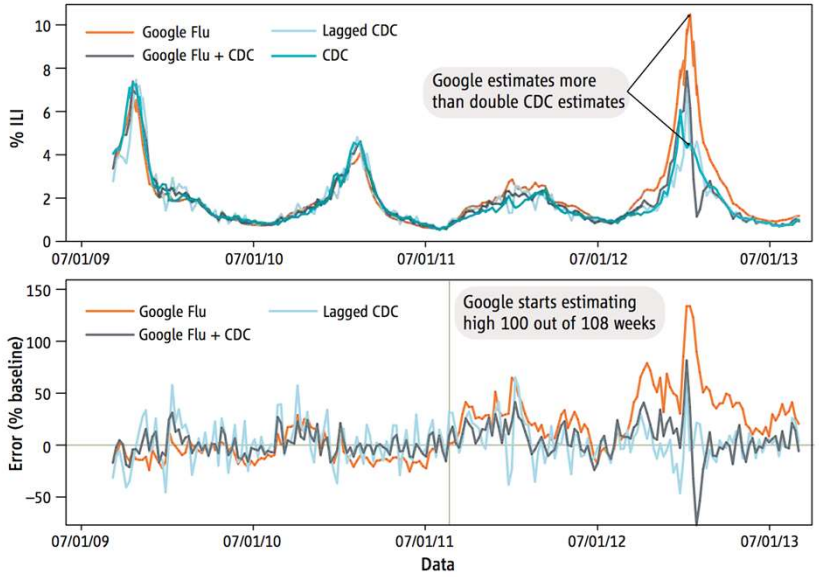
Note: Respondents who did not give an answer are not shown.
Source: Survey conducted Jan. 8-Feb. 7, 2019.

Drifting

Drifting

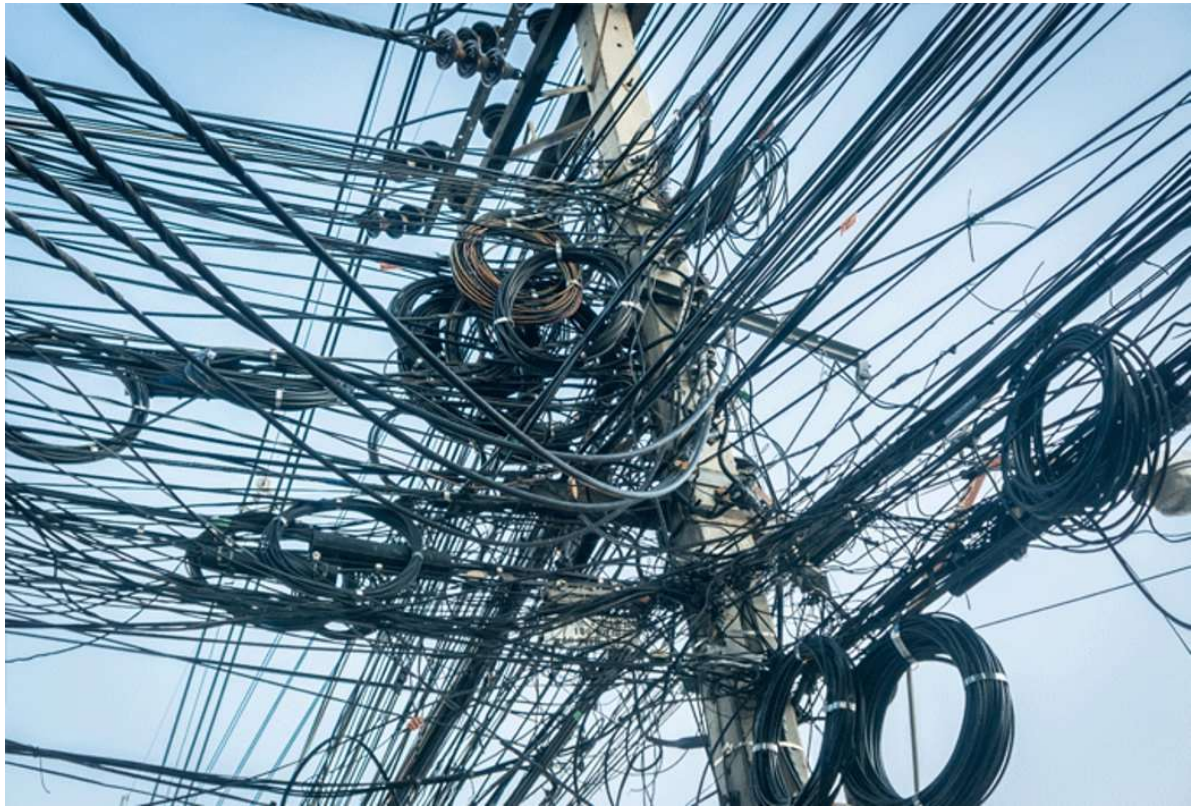
Drifting

Algorithmically confounded



GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage [[Non-CDC estmate)−(CDC estimate)]/(CDC) estimate)]. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

Unstructured



Sensitive

OKCUPID

Researchers Caused an Uproar By Publishing Data From 70,000 OkCupid Users

Robert Hackett
May 18, 2016

https://fortune.com/2016/05/18/okcupid-data-research/

Biased

https://www.rsph.org.uk/about-us/news/instagram-ranked-worst-for-young-people-s-mental-health.html

Biased

CAREERS

IMAGES.COM/CORBIS

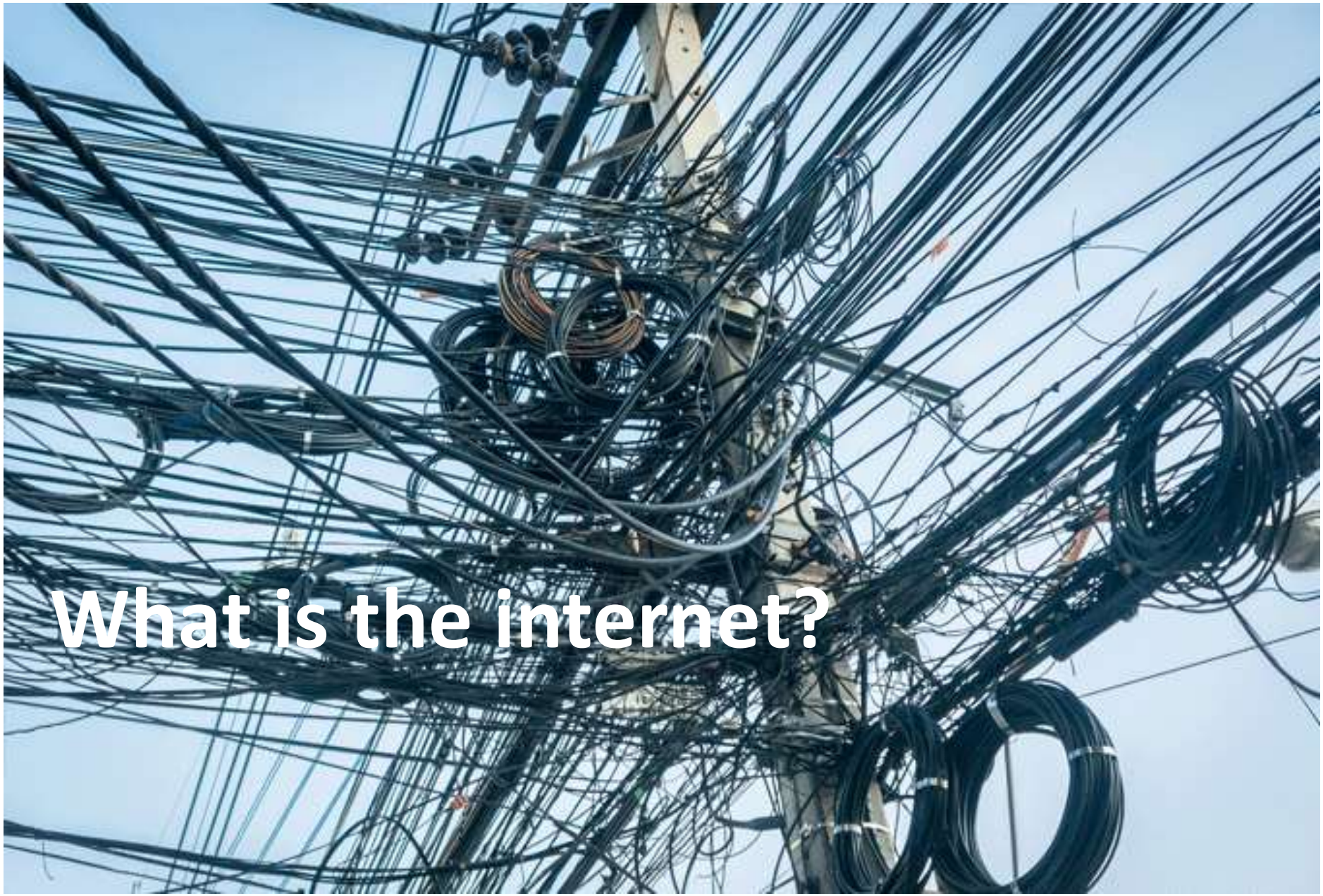
COLUMN

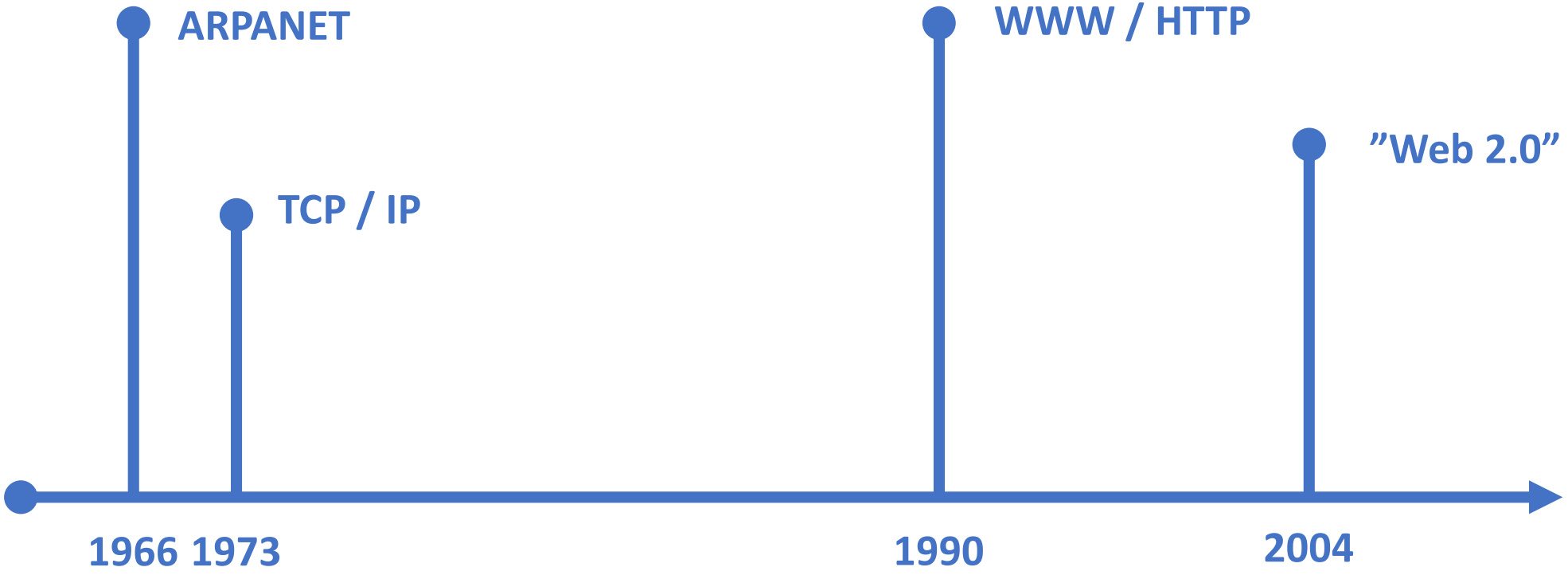
A CV of failures

Keeping a visible record of your rejected applications can help others to deal with setbacks, says **Melanie Stefan**.

https://www.nature.com/articles/nj7322-467a

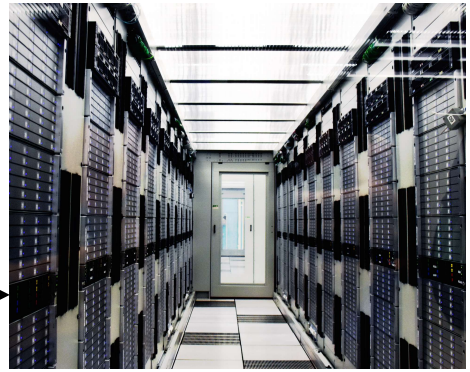
Collecting digital trace data


What is the internet?

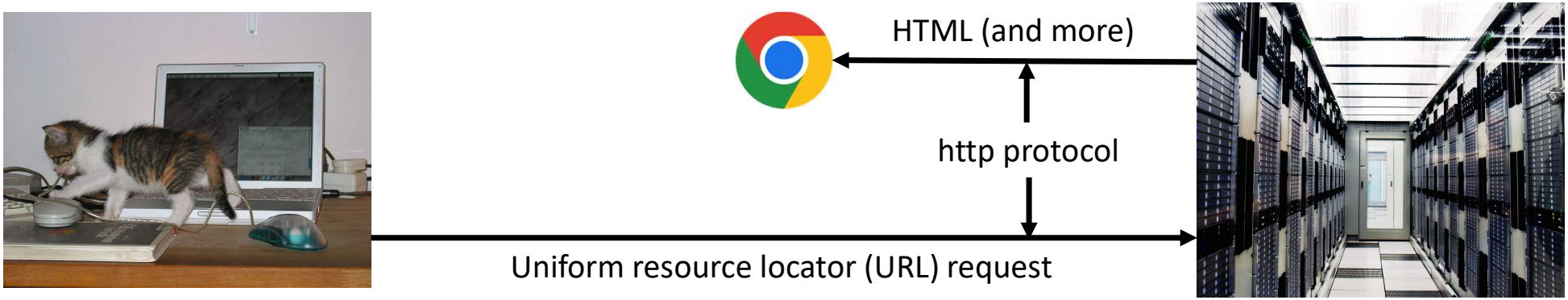

ARPANET
TCP / IP
WWW / HTTP
”Web 2.0”
1966 1973
1990
2004

https://commons.wikimedia.org/wiki/File:Cat-and-computer.JPG

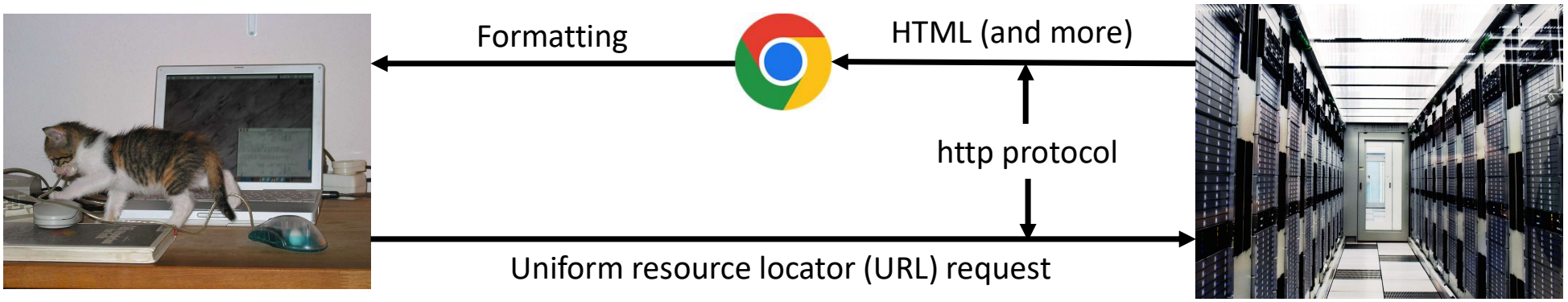
Uniform resource locator (URL) request



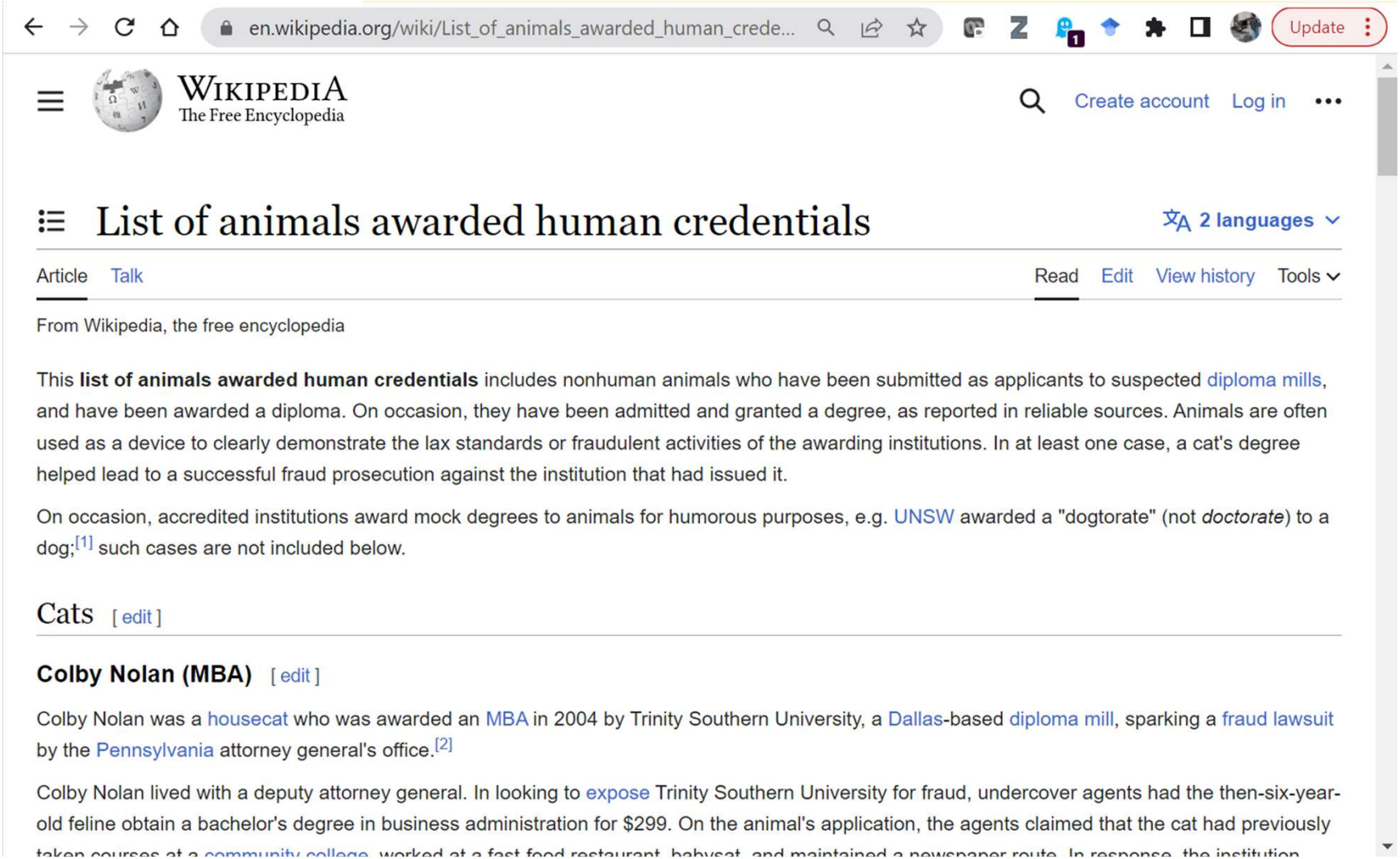
https://commons.wikimedia.org/wiki/File:Cat-and-computer.JPG
https://commons.wikimedia.org/wiki/Category:Server_rooms#/media/File:CERN_Computer_Center_07.jpg



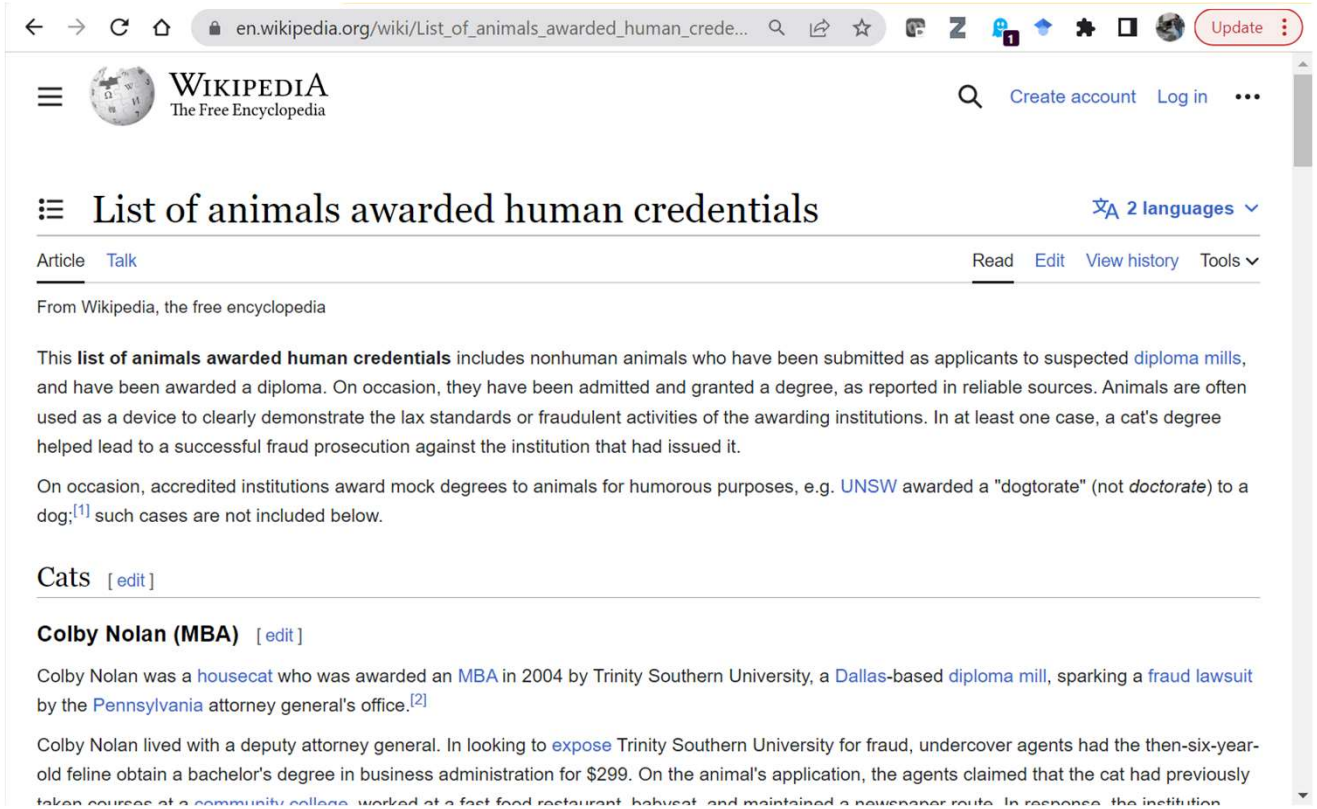
https://commons.wikimedia.org/wiki/File:Cat-and-computer.JPG
https://commons.wikimedia.org/wiki/Category:Server_rooms#/media/File:CERN_Computer_Center_07.jpg



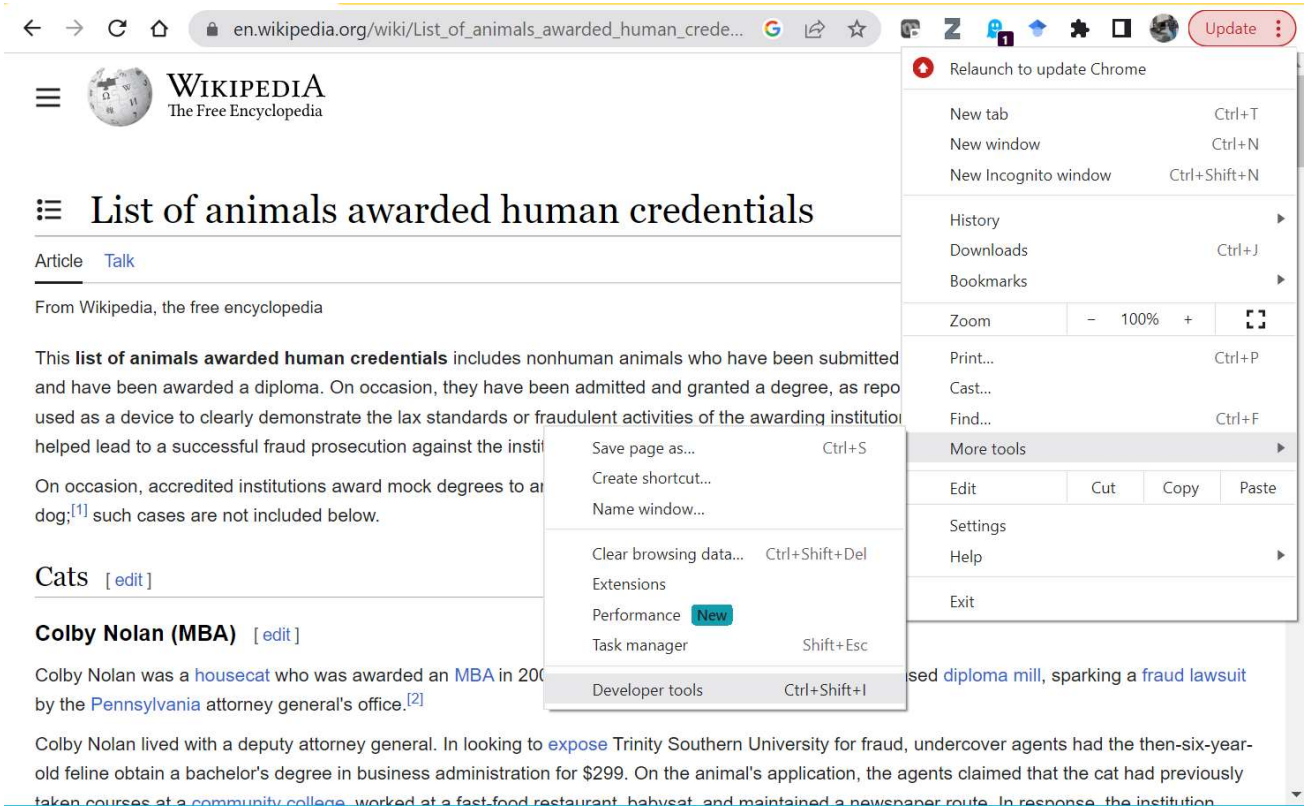
https://commons.wikimedia.org/wiki/File:Cat-and-computer.JPG
https://commons.wikimedia.org/wiki/Category:Server_rooms#/media/File:CERN_Computer_Center_07.jpg


en.wikipedia.org/wiki/List_of_animals_awarded_human_crede...
Update
WIKIPEDIA
The Free Encyclopedia
Create account
Log in
List of animals awarded human credentials
2 languages
Article
Talk
Read
Edit
View history
Tools
From Wikipedia, the free encyclopedia
This **list of animals awarded human credentials** includes nonhuman animals who have been submitted as applicants to suspected diploma mills, and have been awarded a diploma. On occasion, they have been admitted and granted a degree, as reported in reliable sources. Animals are often used as a device to clearly demonstrate the lax standards or fraudulent activities of the awarding institutions. In at least one case, a cat's degree helped lead to a successful fraud prosecution against the institution that had issued it.
On occasion, accredited institutions award mock degrees to animals for humorous purposes, e.g. UNSW awarded a "dogtorate" (not *doctorate*) to a dog;[1] such cases are not included below.
Cats [edit]
Colby Nolan (MBA) [edit]
Colby Nolan was a housecat who was awarded an MBA in 2004 by Trinity Southern University, a Dallas-based diploma mill, sparking a fraud lawsuit by the Pennsylvania attorney general's office.[2]
Colby Nolan lived with a deputy attorney general. In looking to expose Trinity Southern University for fraud, undercover agents had the then-six-year-old feline obtain a bachelor's degree in business administration for $299. On the animal's application, the agents claimed that the cat had previously

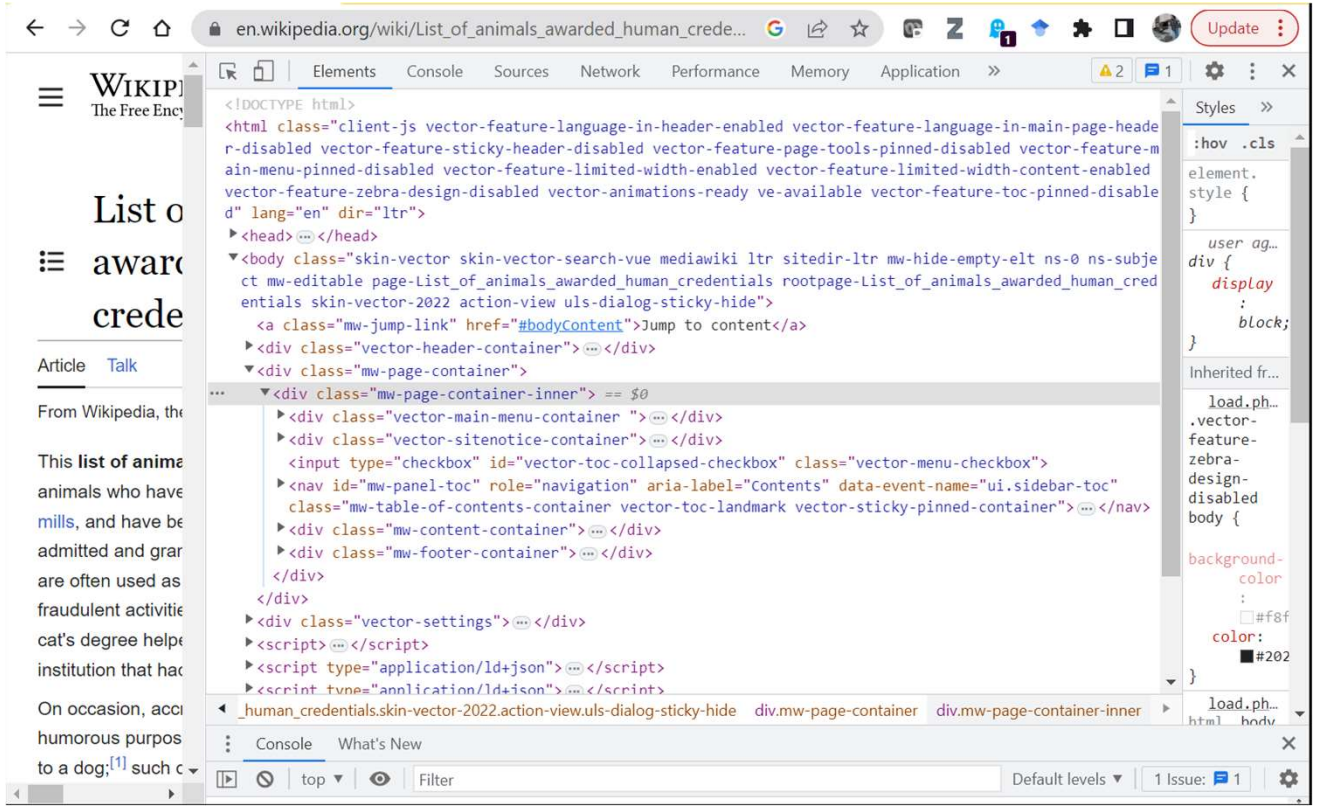
A web page (to you)



A web page (to you)



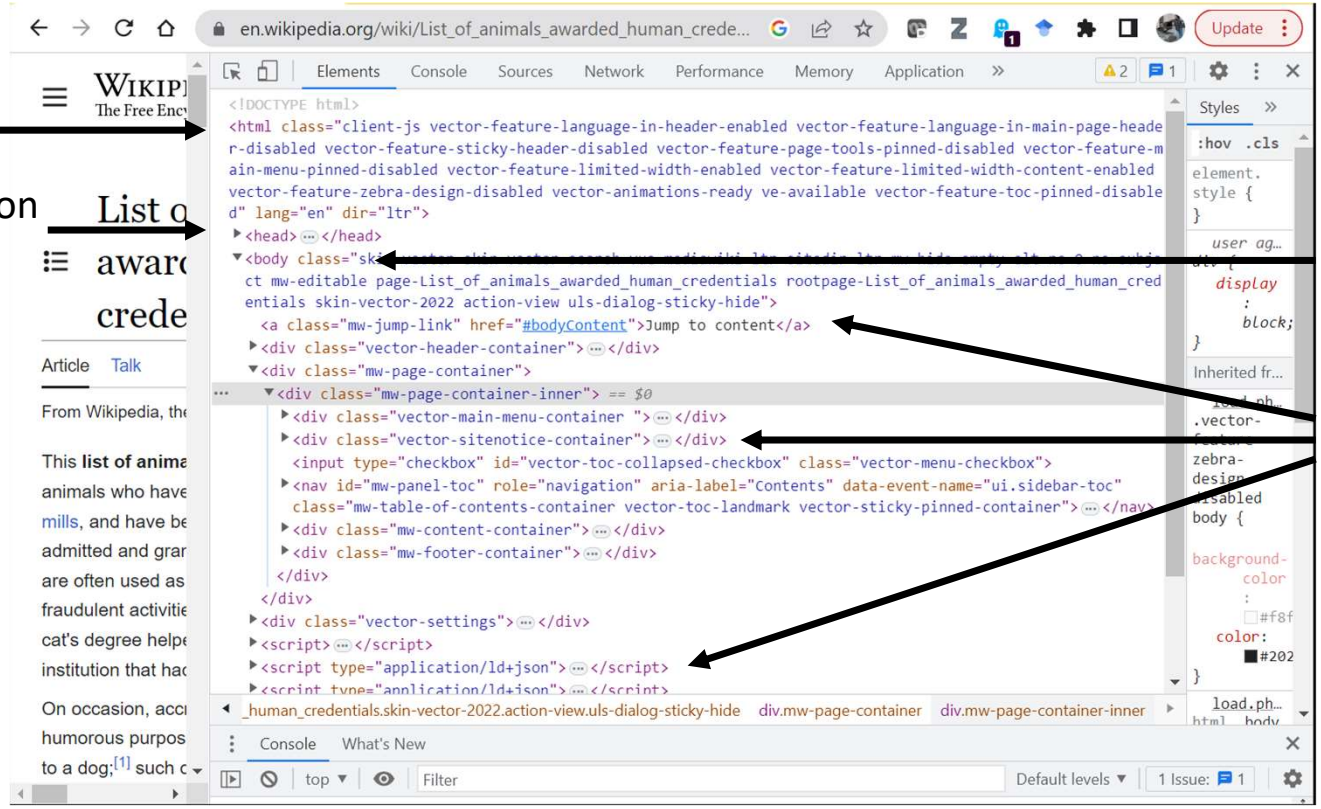
A web page (to your browser)



A web page (to your browser)

Type of document (declares format)

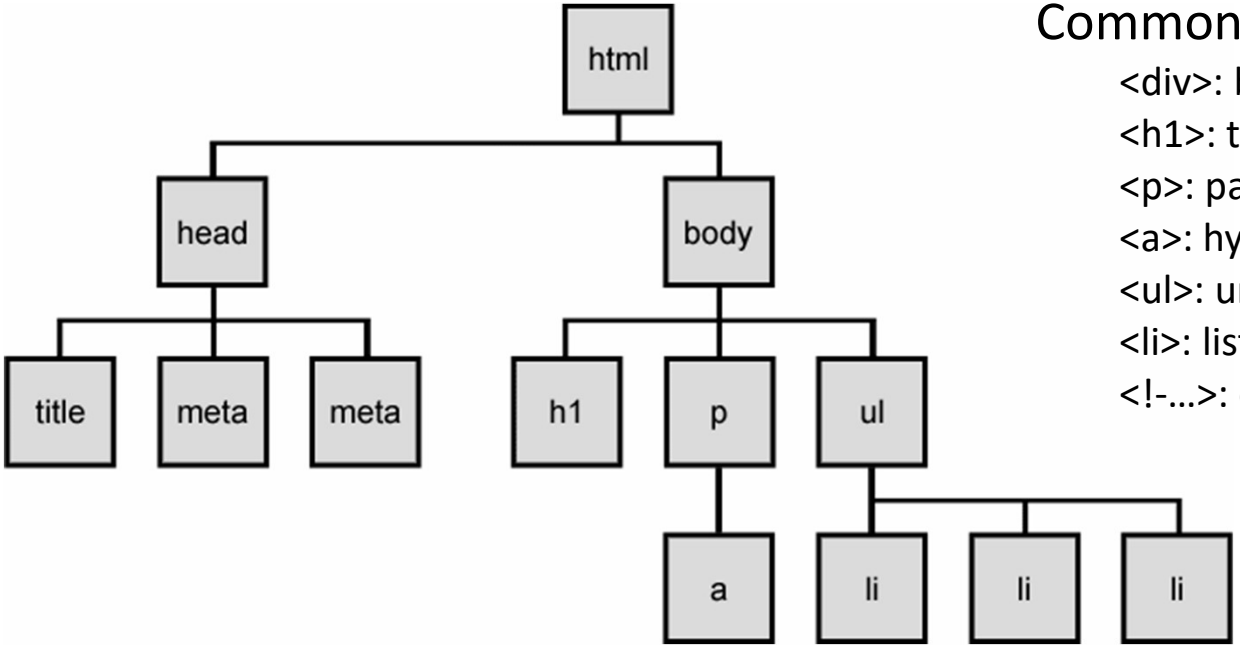
Head (meta information about the document)



Body (most content is somewhere in here)

Other tags

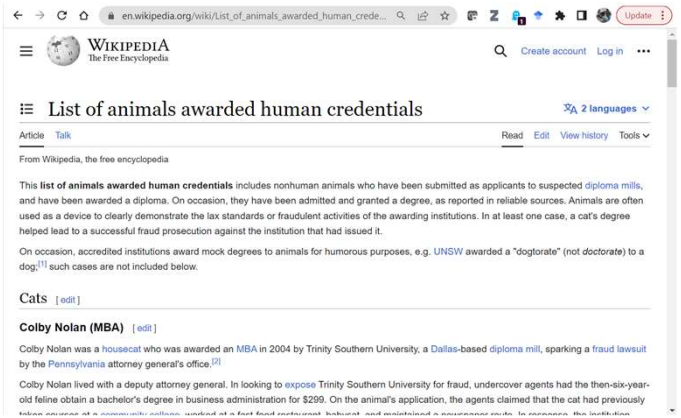
A web page has a tree structure

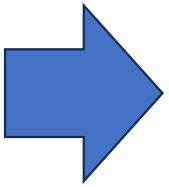


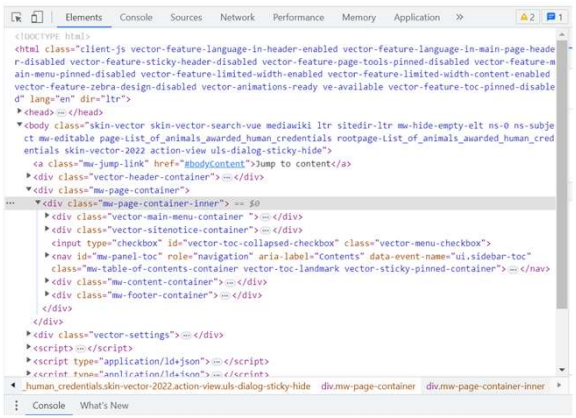
Common tags

- : block of text
- <h1>: title 1
- <p>: paragraph
- <a>: hypertext link
- : unordered list
- : list element
- <!-…>: comment

Our goal

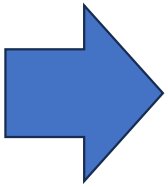

en.wikipedia.org/wiki/List_of_animals_awarded_human_crede...
Update
WIKIPEDIA
The Free Encyclopedia
Create account
Log in
List of animals awarded human credentials
2 languages
Article
Talk
Read
Edit
View history
Tools
From Wikipedia, the free encyclopedia
This list of animals awarded human credentials includes nonhuman animals who have been submitted as applicants to suspected diploma mills, and have been awarded a diploma. On occasion, they have been admitted and granted a degree, as reported in reliable sources. Animals are often used as a device to clearly demonstrate the lax standards or fraudulent activities of the awarding institutions. In at least one case, a cat's degree helped lead to a successful fraud prosecution against the institution that had issued it.
On occasion, accredited institutions award mock degrees to animals for humorous purposes, e.g. UNSW awarded a "dogtorate" (not doctorate) to a dog;[1] such cases are not included below.
Cats [edit]
Colby Nolan (MBA) [edit]
Colby Nolan was a housecat who was awarded an MBA in 2004 by Trinity Southern University, a Dallas-based diploma mill, sparking a fraud lawsuit by the Pennsylvania attorney general's office.[2]
Colby Nolan lived with a deputy attorney general. In looking to expose Trinity Southern University for fraud, undercover agents had the then-six-year-old feline obtain a bachelor's degree in business administration for $299. On the animal's application, the agents claimed that the cat had previously




Elements
Console
Sources
Network
Performance
Memory
Application
<!DOCTYPE html>
<html class="client-js vector-feature-language-in-header-enabled vector-feature-language-in-main-page-header-disabled vector-feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-main-menu-pinned-disabled vector-feature-limited-width-enabled vector-feature-limited-width-content-enabled vector-feature-zebra-design-disabled vector-animations-ready ve-available vector-feature-toc-pinned-disabled" lang="en" dir="ltr">
<head>…</head>
<body class="skin-vector skin-vector-search-vue mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable page-List_of_animals_awarded_human_credentials rootpage-List_of_animals_awarded_human_credentials skin-vector-2022 action-view uls-dialog-sticky-hide">
Jump to content
…

…
…
<input type="checkbox" id="vector-toc-collapsed-checkbox" class="vector-menu-checkbox">
<nav id="mw-panel-toc" role="navigation" aria-label="Contents" data-event-name="ui.sidebar-toc" class="mw-table-of-contents-container vector-toc-landmark vector-sticky-pinned-container">…</nav>
…
…

…
<script>…</script>
<script type="application/ld+json">…</script>
_human_credentials.skin-vector-2022.action-view.uls-dialog-sticky-hide
div.mw-page-container
div.mw-page-container-inner
Console
What's New



Screen-scraping

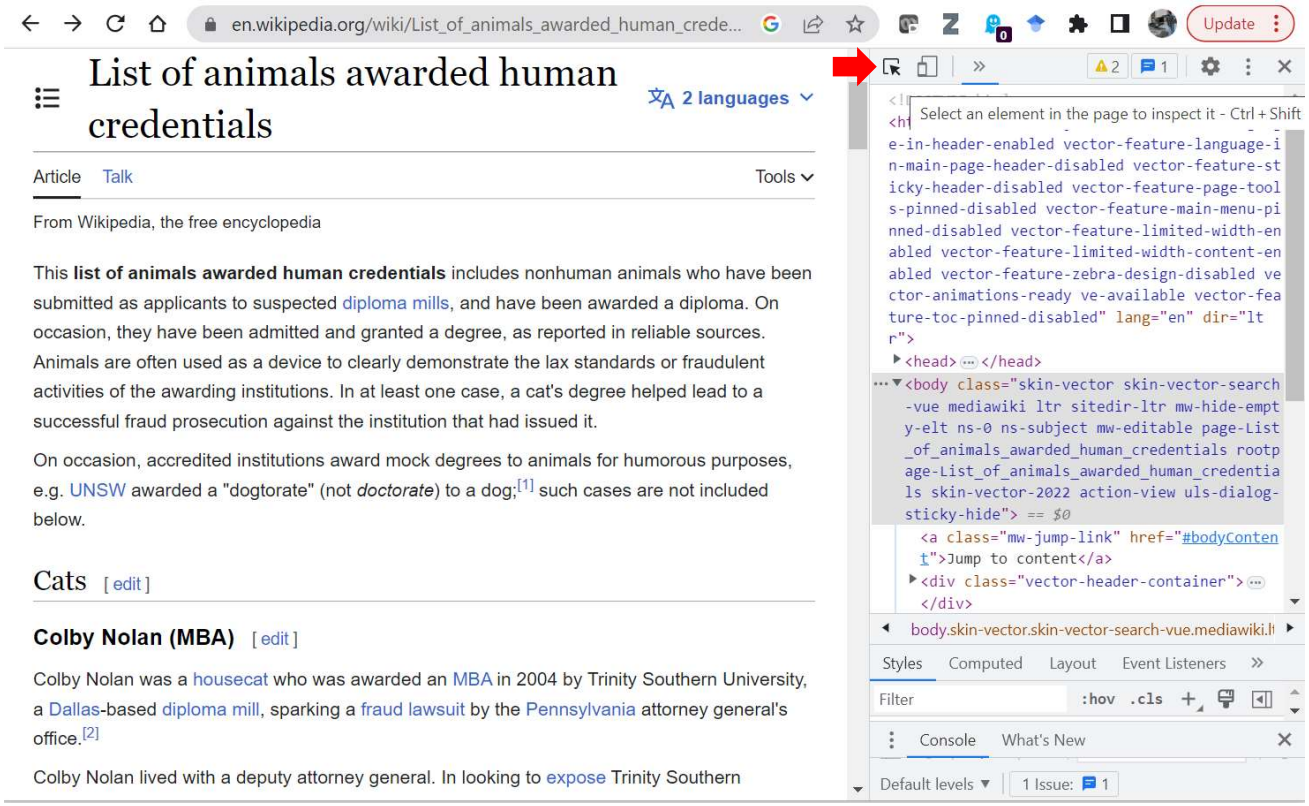
But first, back to yesterday!

- Respect for persons: participants decide.
- Beneficence: minimize risk, maximize benefits.
- Justice: distribute the burdens and benefits of research.
- Respect for law and public interest: compliance and transparency.

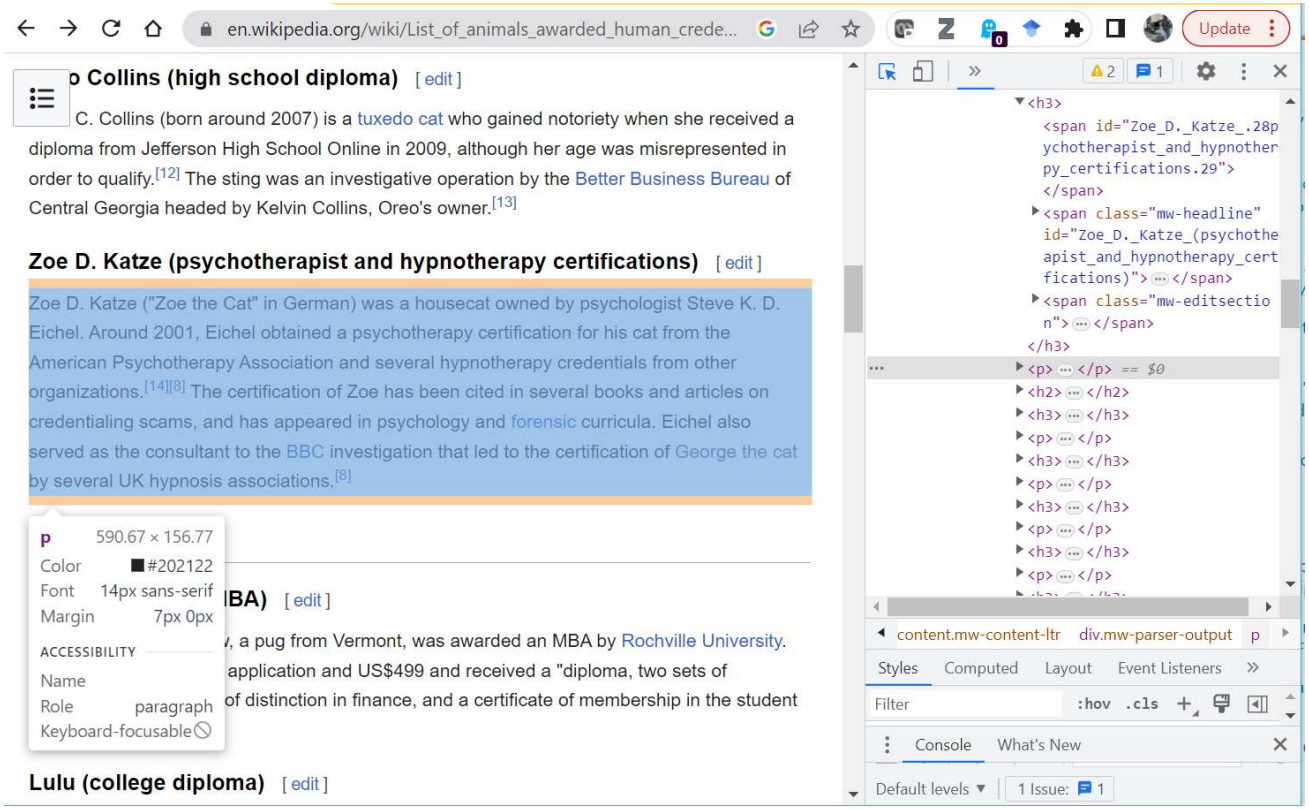
Is screen-scraping legal?

- Always check for:
  - Terms of Service (usually at the bottom of the page).
  - A robots.txt file (eg. https://en.wikipedia.org/robots.txt).

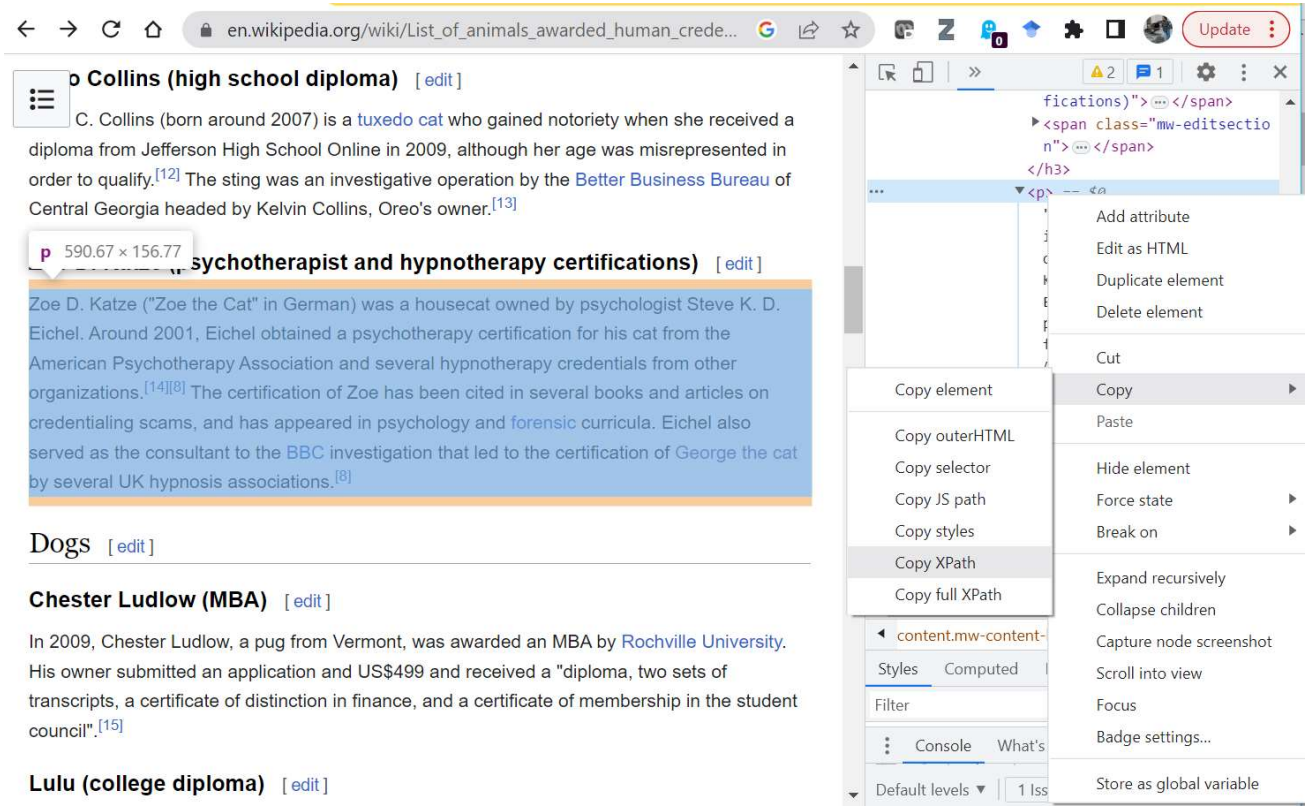
Parsing HTML



Parsing HTML

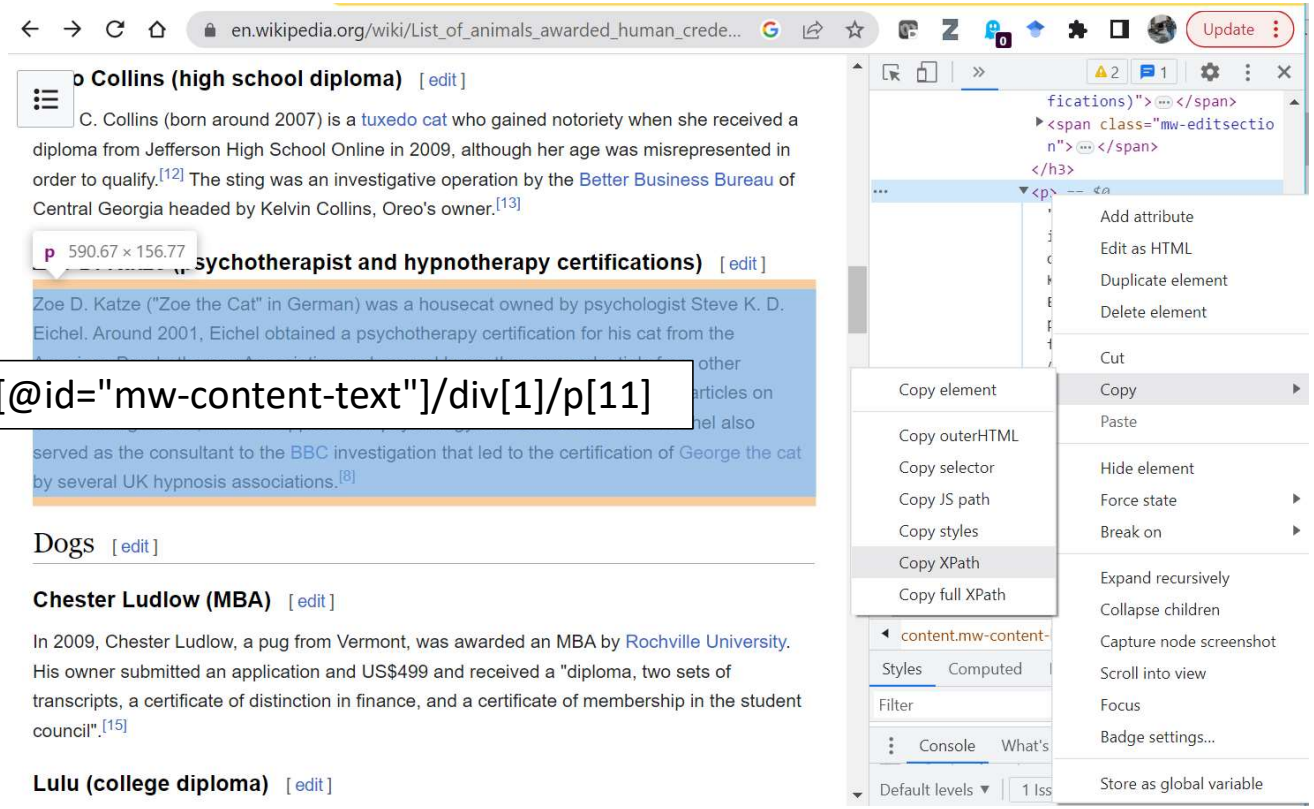


Parsing HTML

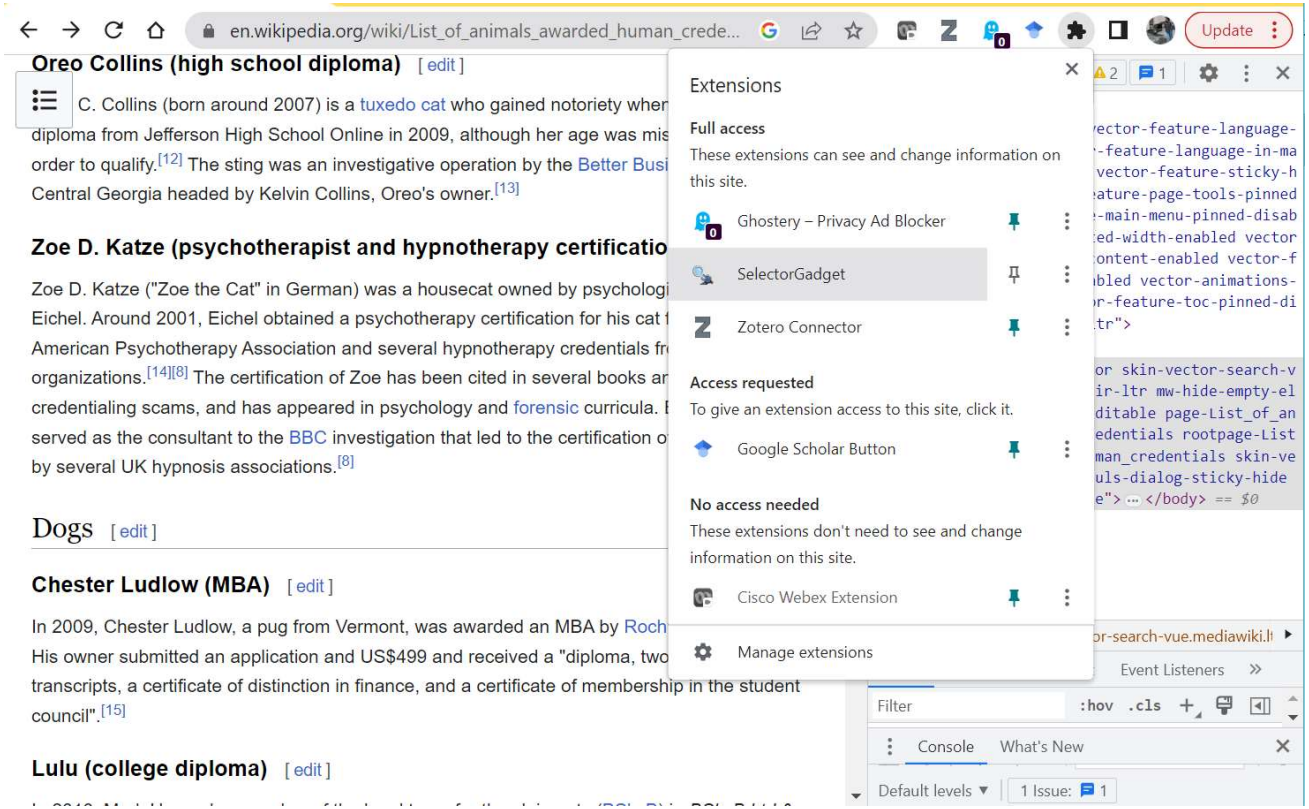


Parsing HTML

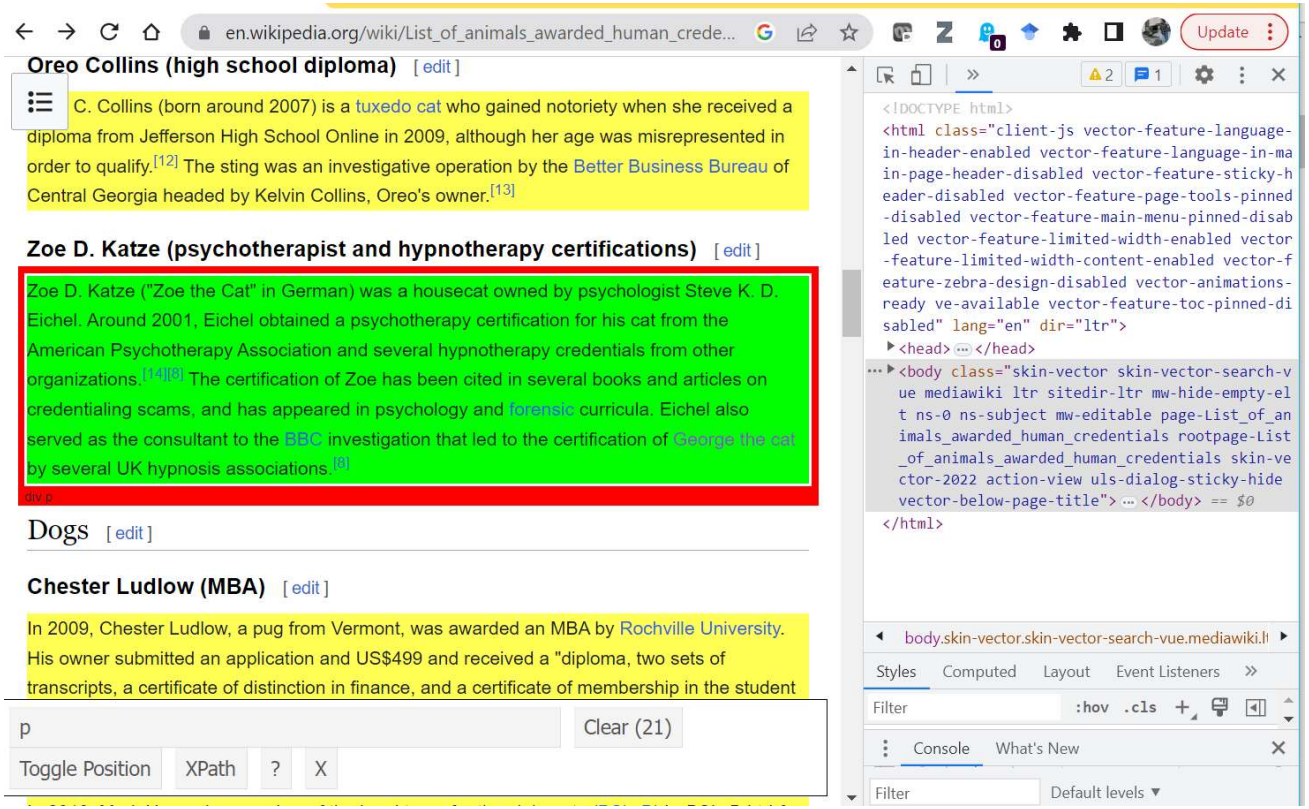
The XPath: //*[@id="mw-content-text"]/div[1]/p[11]



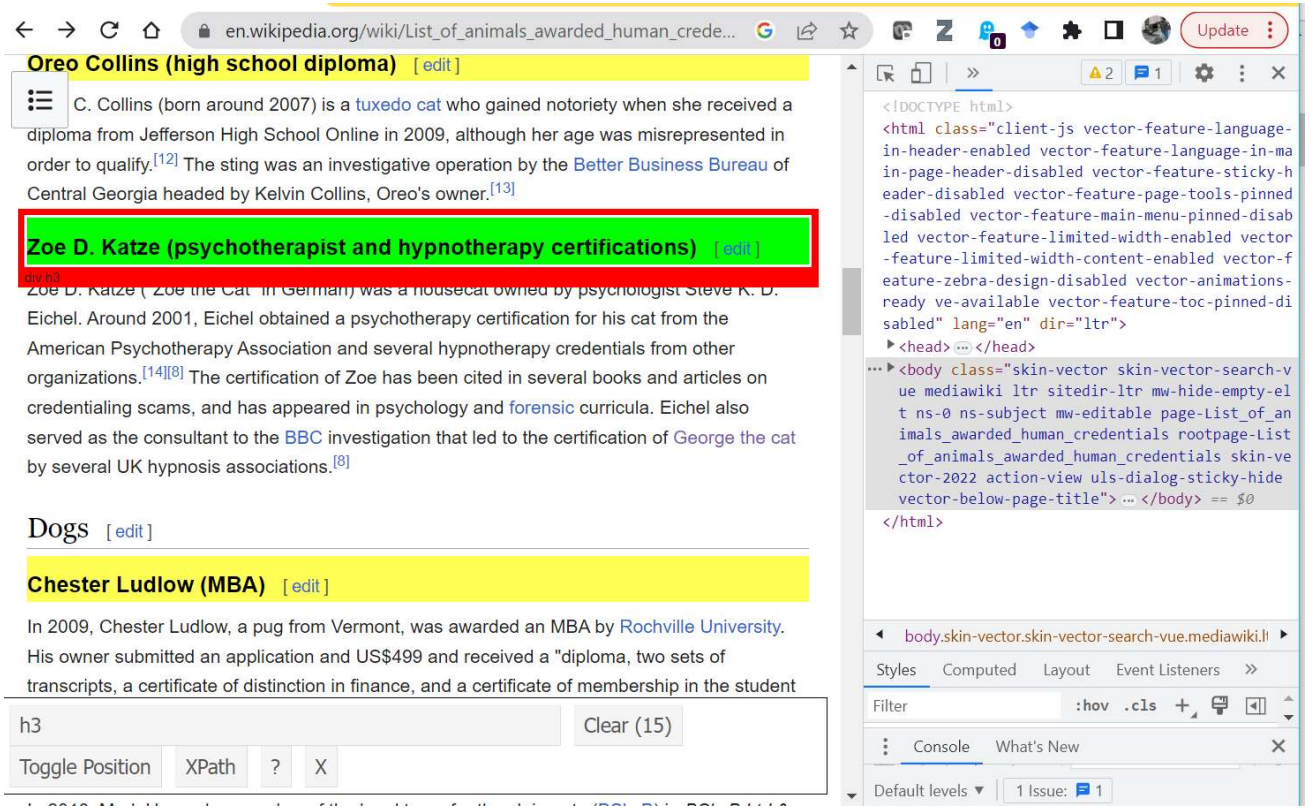
Parsing with the CSS Selector



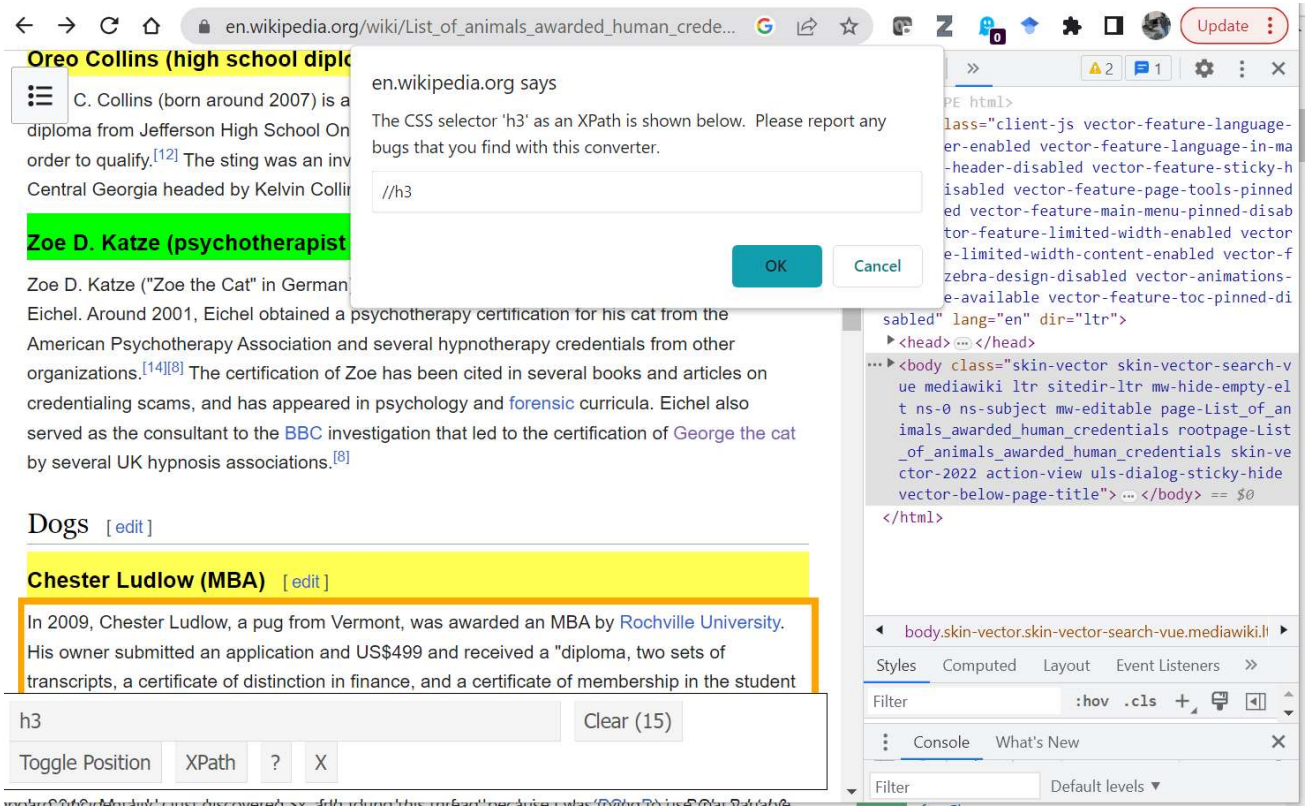
Parsing with the CSS Selector



Parsing with the CSS Selector

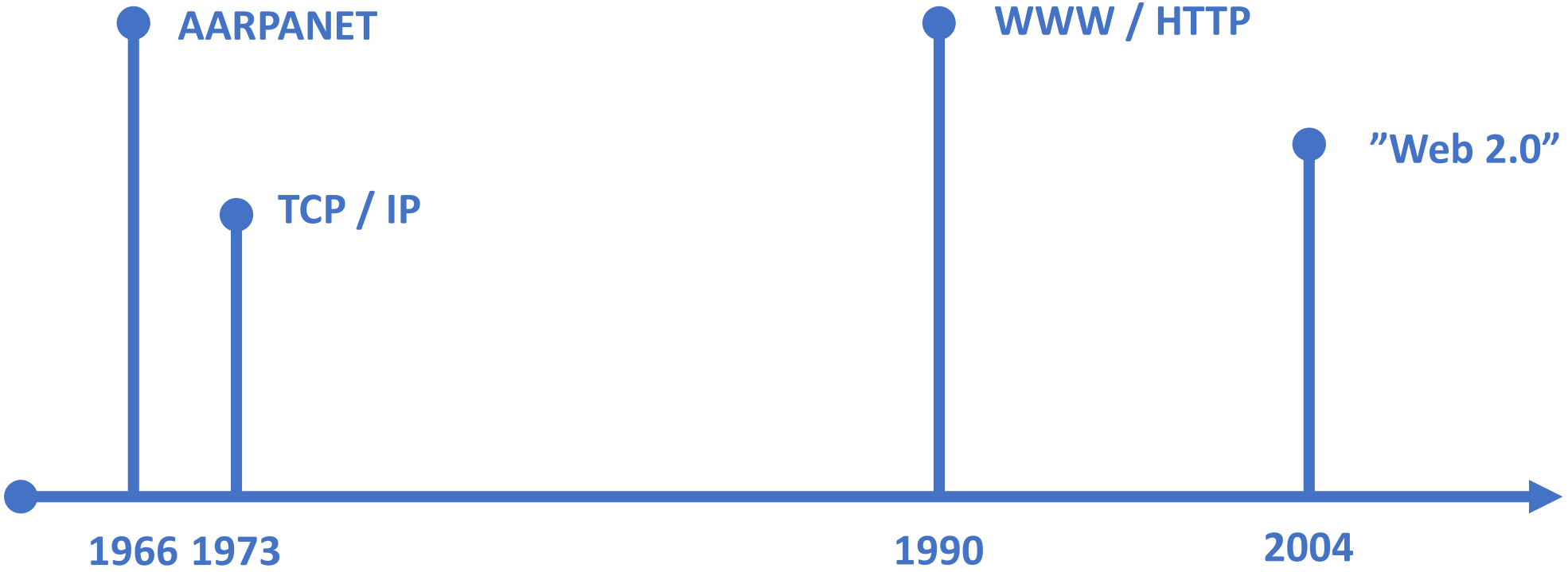


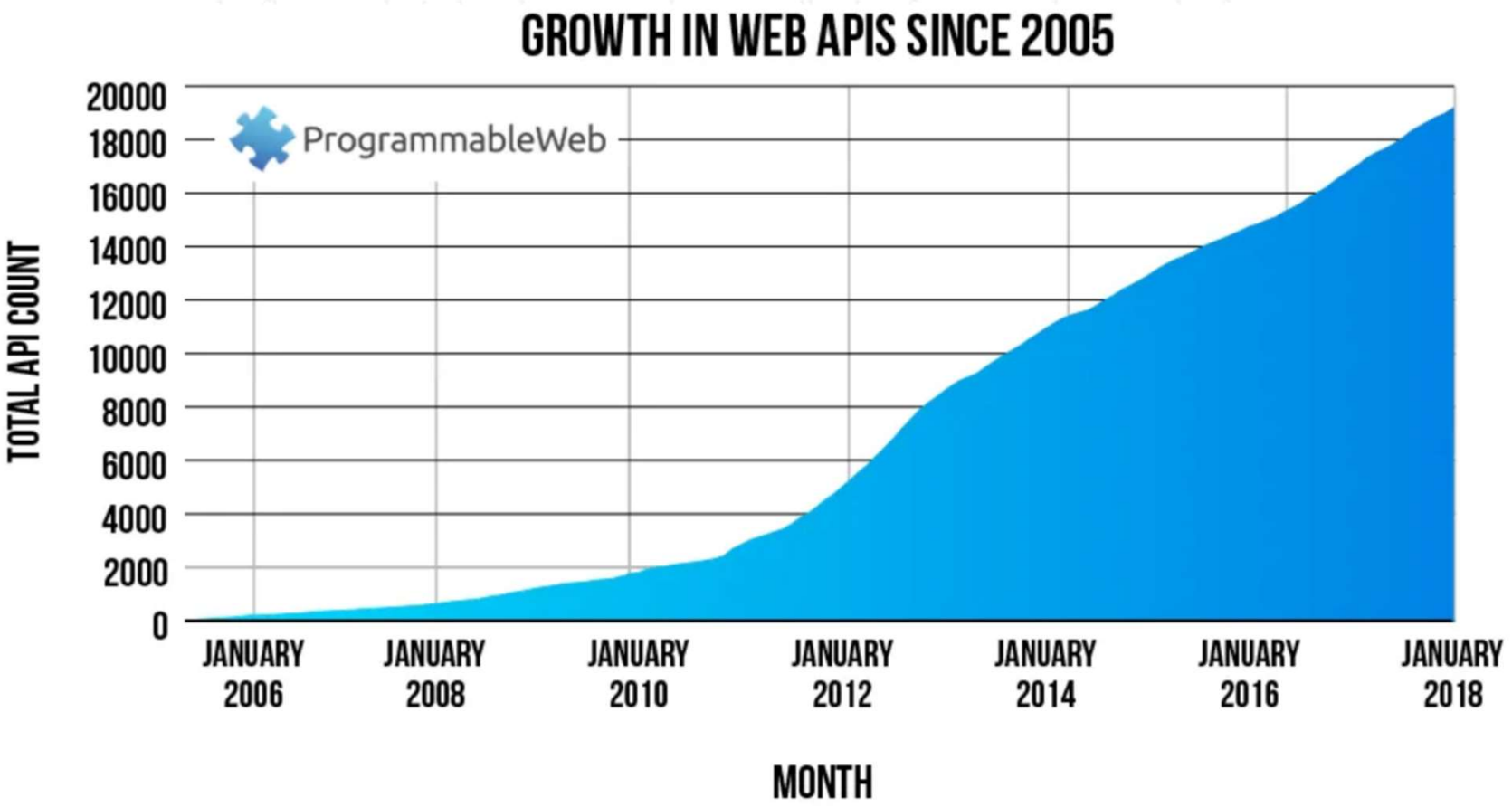
Parsing with the CSS Selector



Application programming interfaces

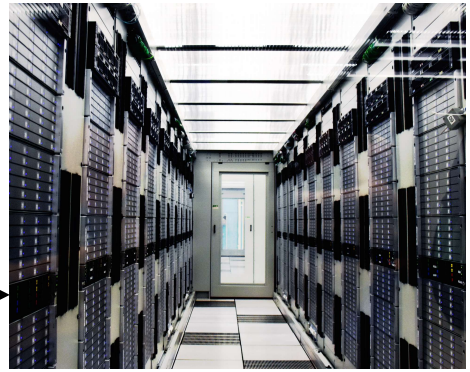
(APIs)


AARPANET
TCP / IP
WWW / HTTP
”Web 2.0”
1966 1973
1990
2004



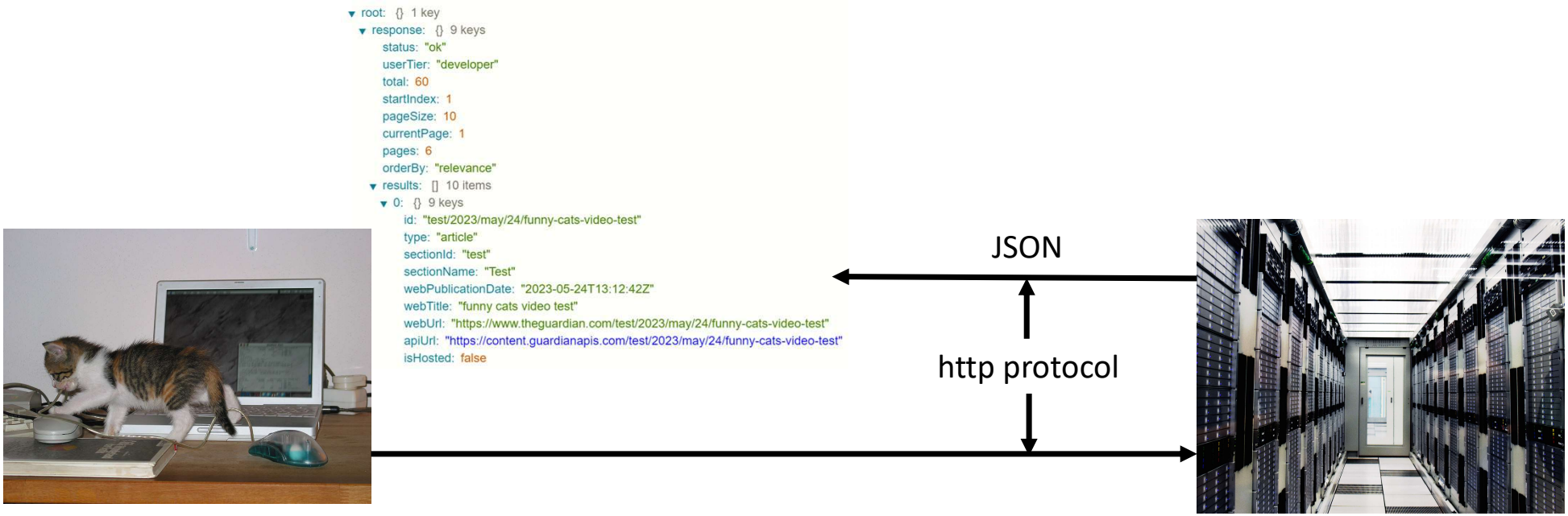
https://medium.com/@programmableweb/research-shows-interest-in-providing-apis-still-high-c9cb5c680c09

https://commons.wikimedia.org/wiki/File:Cat-and-computer.JPG



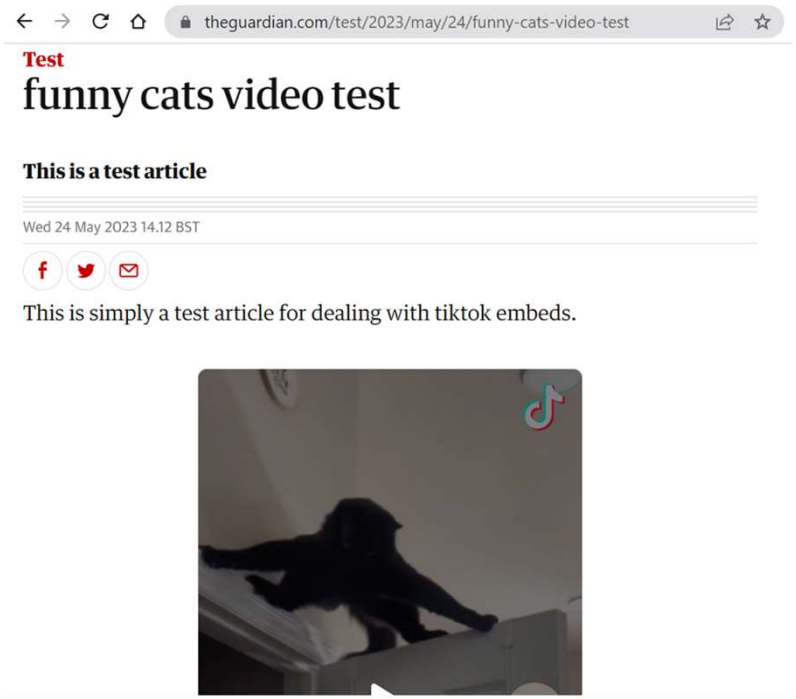
https://content.guardianapis.com/search?from-date=2023-05-12&q=cats&api-key= ...

https://commons.wikimedia.org/wiki/File:Cat-and-computer.JPG
https://commons.wikimedia.org/wiki/Category:Server_rooms#/media/File:CERN_Computer_Center_07.jpg



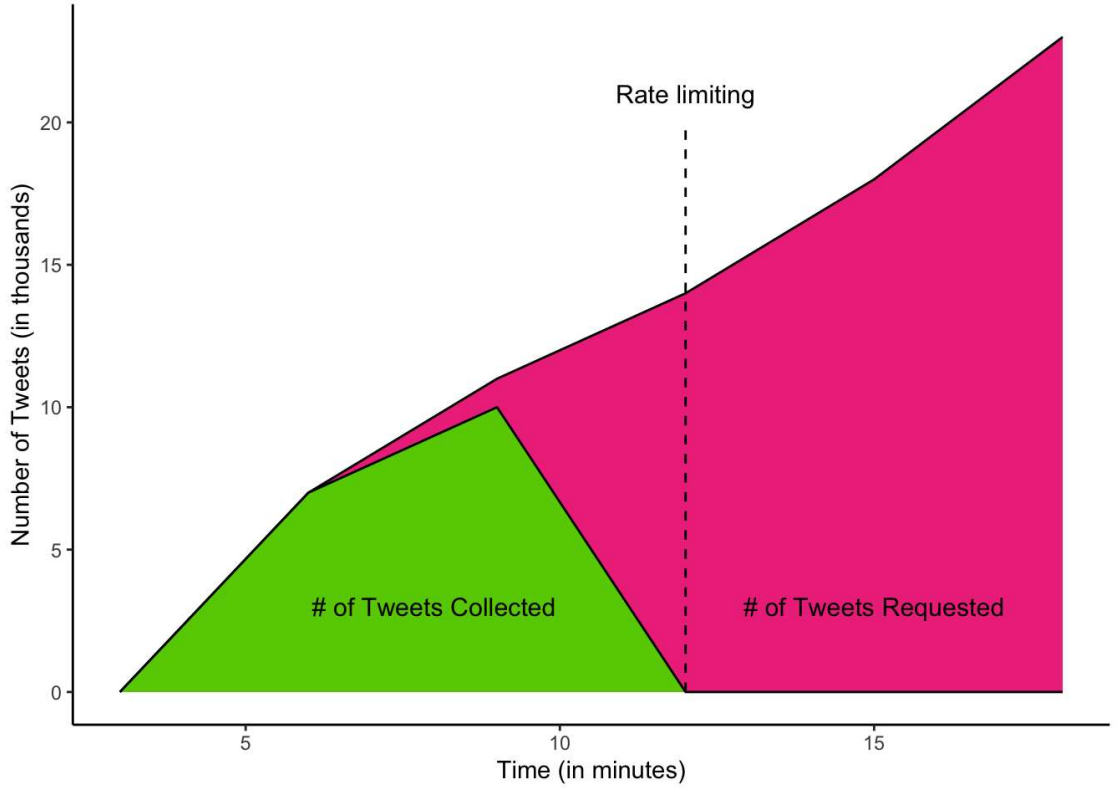
https://content.guardianapis.com/search?from-date=2023-05-12&q=cats&api-key= ...

https://commons.wikimedia.org/wiki/File:Cat-and-computer.JPG
https://commons.wikimedia.org/wiki/Category:Server_rooms#/media/File:CERN_Computer_Center_07.jpg

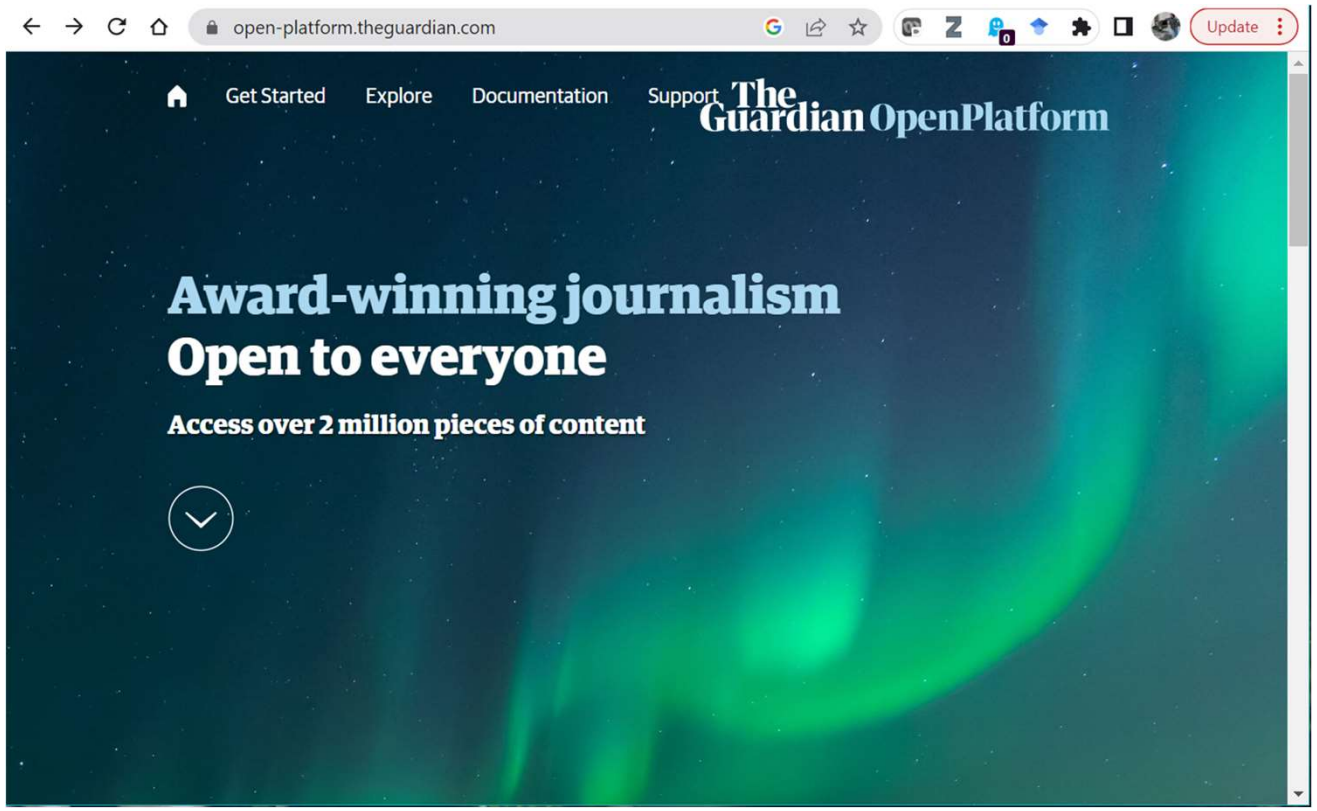


https://www.theguardian.com/test/2023/may/24/funny-cats-video-test
https://open-platform.theguardian.com/explore/

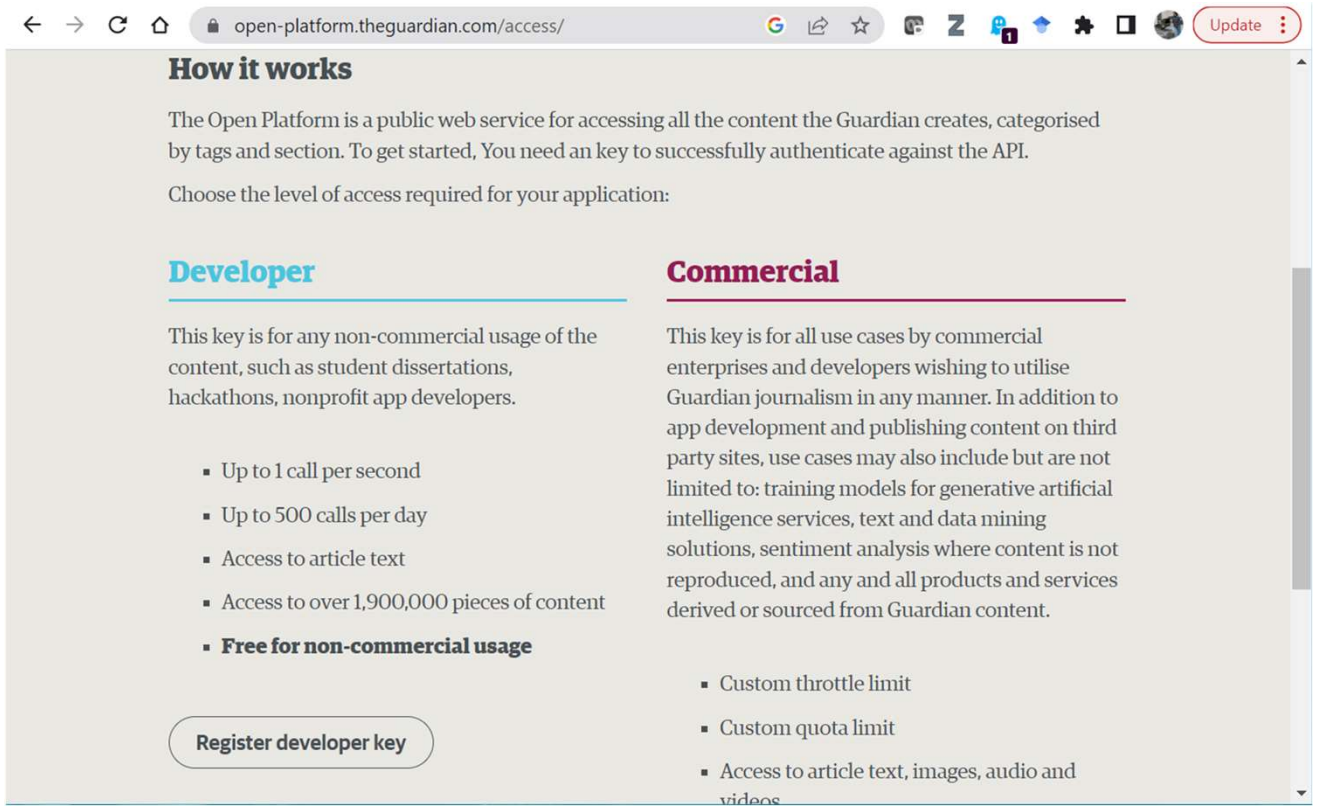
API restrictions



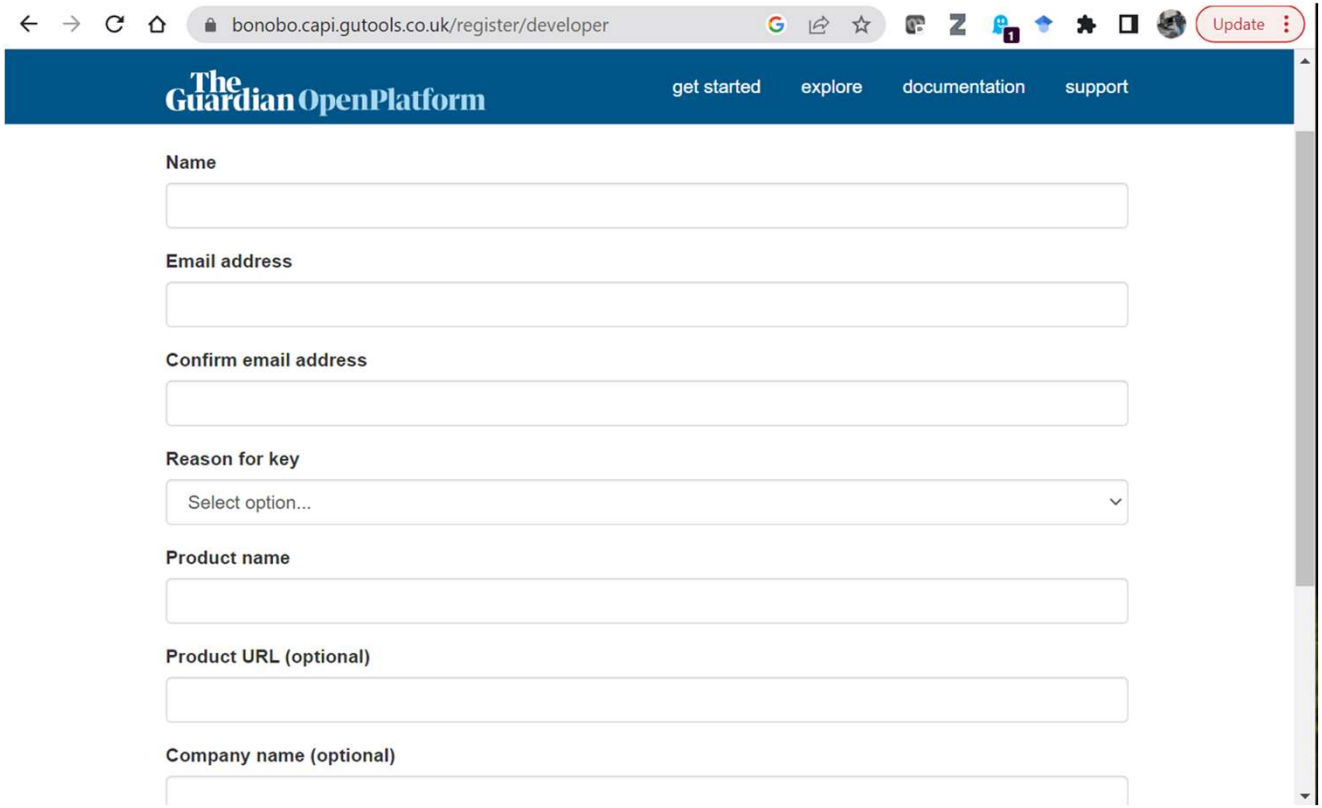
API registration



API registration



API registration



More APIs

- Lists of APIs
  - https://docs.google.com/spreadsheets/d/1ZEr3okdlb0zctmX0MZKo-gZKPsq5WGn1nJOxPV7al-Q/edit?usp=sharing
  - https://github.com/toddmotto/public-apis
  - https://apilist.fun/
  - R OpenScience has a list of R packages that work with APIs: https://ropensci.org/packages/

- More APIs
  - The SEC: https://www.sec.gov/edgar/sec-api-documentation
  - For the Science of Science: https://doi.org/10.1038/s41597-023-02198-9

For the workshop

- Install Chrome and the Selector Gadget extension

- Register for a Guardian developer API key:
  https://bonobo.capi.gutools.co.uk/register/developer

More APIs

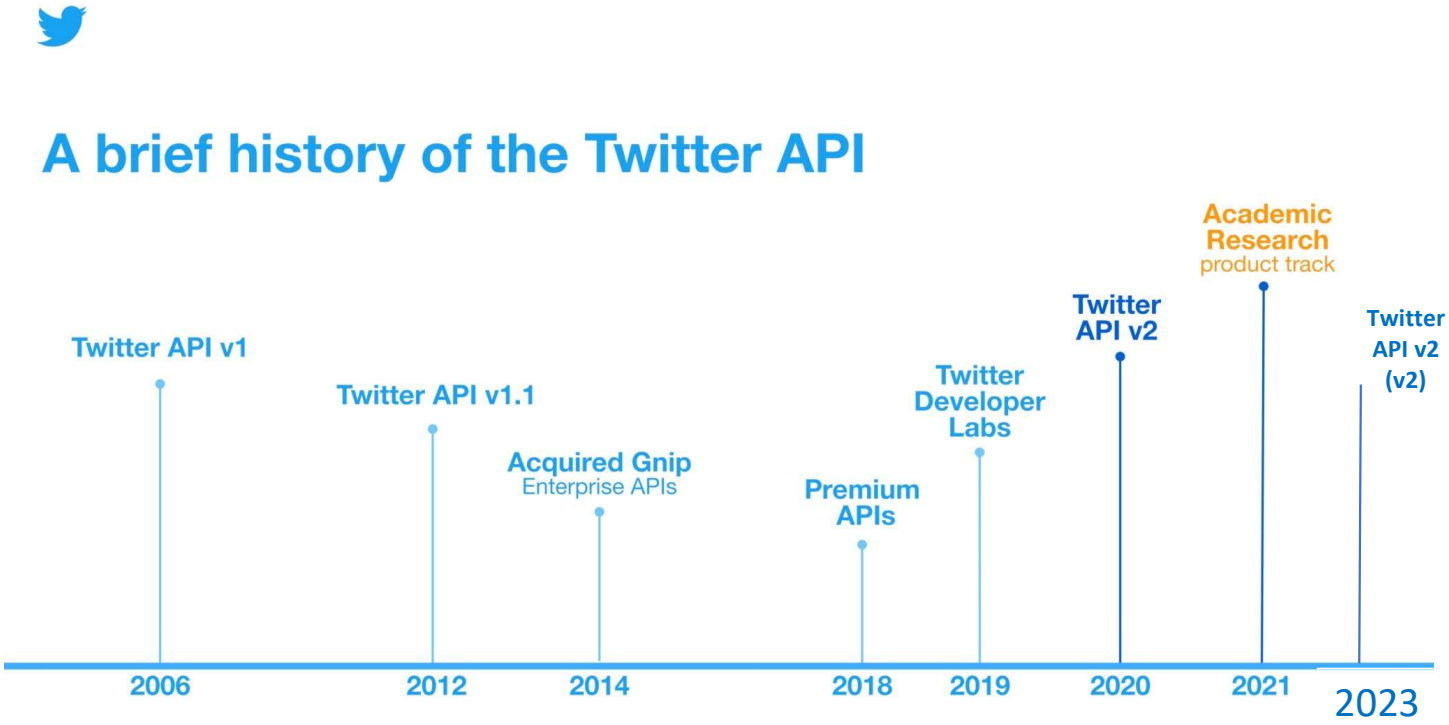
Data source	Highlights	API	Data dump
Crossref	Data on publications with DOIs registered in Crossref.	✓	✓
OpenAlex	Data connecting publications, authors, institutions, and concepts.	✓	✓
Dimensions	Data connecting publications, grants, datasets, trials, and patents.	—	—
Overton	Policy documents and their citations to science and policy.	—	—
OpenCitations	DOI-DOI open citation links.	✓	✓
AMiner	Advanced information generated through data mining techniques.	✓	✓
CiteSeerX	Full-text publications, one of the earliest digital library search engines.	✓	—
ORCID	Data on researchers with ORCID IDs (funding, works, peer review, etc.).	✓	✓
ROR	Data on research organizations with ROR IDs, seeded by GRID.	✓	✓
Retraction Watch	Data on retracted papers and reasons for retraction.	✗	—
Semantic Scholar	Publication dataset featuring AI-derived products (e.g., embeddings).	✓	—
Web of Science	Curated by in-house experts, basis for Journal Citation Reports.	—	—
PubMed	Biomedical literature with PubMed IDs, linked to NIH projects, clinical trials, and other biomedical entities.	✓	✓
NIH RePORTER	Data on NIH-funded projects, with linkages to publications, patents, and clinical studies.	✓	✓
NSF Awards	Data on NSF-funded projects, with linkages to publications.	✓	✓
Clinical Trials	Information on clinical studies and linkages to references worldwide.	✓	✓
PatentsView	Data on USPTO patents (citations, classifications, inventors, etc.).	✓	✓
Patent Citation to Science	Patent-science citations extracted from USPTO and EPO patents.	✗	✓
Publications of Nobel laureates	Publication records and prize-winning papers of Nobel laureates.	✗	✓
Altmetric	Data on online attention (e.g., mainstream and social media).	✓	—
CORE	Metadata and full-text information of 87 M + papers.	✓	✓
Unpaywall	Publication metadata and open-access related information.	✓	✓
DOAJ	Community-curated data on open-access journals and papers.	✓	✓
OpenAIRE Research Graph	Data connecting scientific products, organizations, funded projects, etc. from 70 K + sources.	✓	✓
Faculty Opinions with Gender	Metadata of authors from Faculty Opinions with gender classification from Faculty Opinions and Web of Science.	—	✓
Scopus	Documents selected by an independent review board of experts.	—	—
Lens	Citation relationships within and across papers and patents.	—	—
Springer Nature SciGraph	Triples connecting multiple entities in the research landscape, including publications, funders, and affiliations.	✓	✓
Google Scholar	Large-scale data on publications, citations, and disambiguated scholar profiles indexed by Google.	✗	✗

Table 1. Brief summary of major data sources commonly used in the science of science literature. ✓: publicly available, —: available upon application or subscription, ✗: not available to the best of our knowledge (a more detailed summary is given in Table S1).

What's up with Twitter?

What's up with Twitter?

What's up with Twitter?



https://techcrunch.com/2021/01/26/twitters-new-api-platform-now-opened-to-academic-researchers/

What's up with Twitter?

2021

Basic access

- 500,000 Tweets / month
- Free

Academic access

- 10 million Tweets / month
- Free

2023

Free access

- 1,500 Tweets / month
- Free

Basic access

- 3,000 – 10,000 Tweets / month
- $100.00 / month