

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>

Laura Alessandretti^{1,2,3}, Ulf Aslak^{1,2,3} & Sune Lehmann^{1,2,3,4}

Received: 3 February 2020

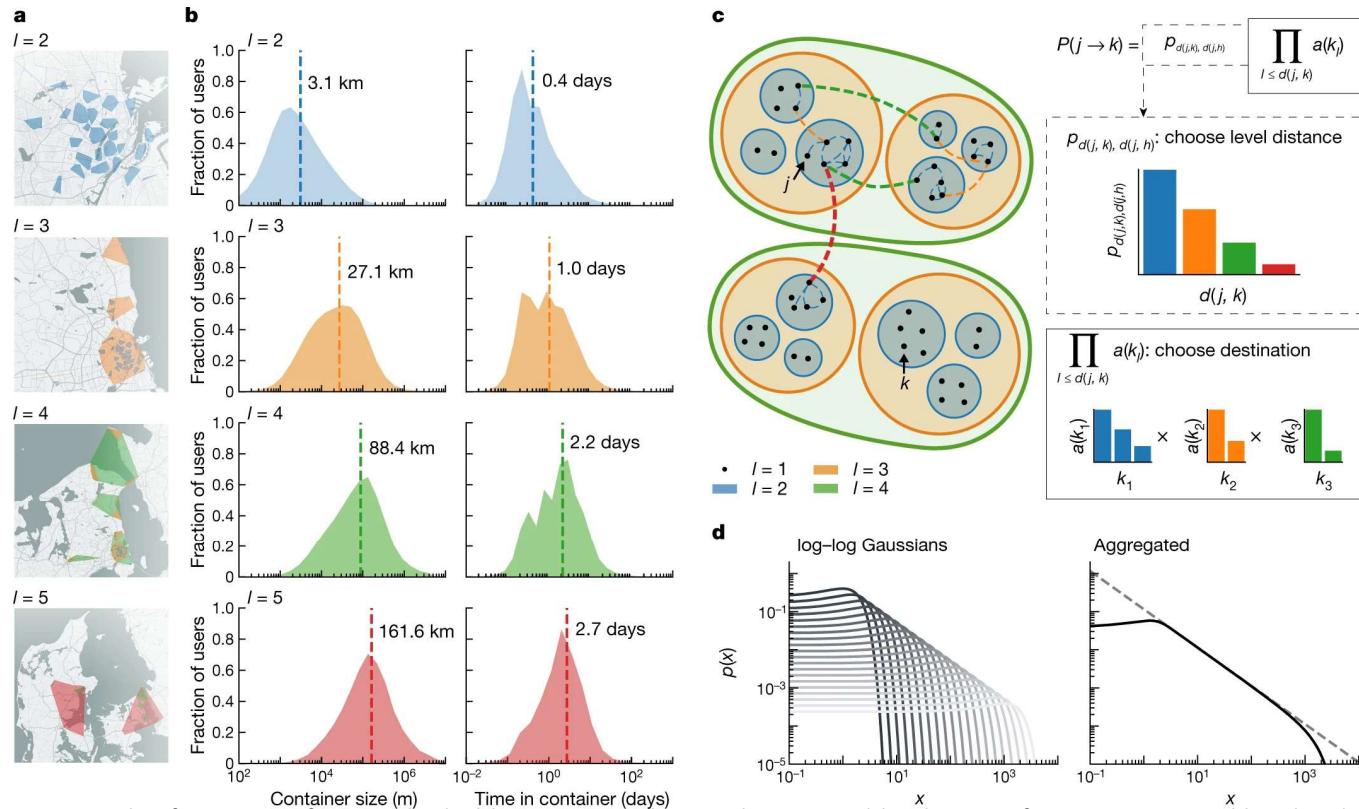
Accepted: 25 September 2020

Published online: 18 November 2020

 Check for updates

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.

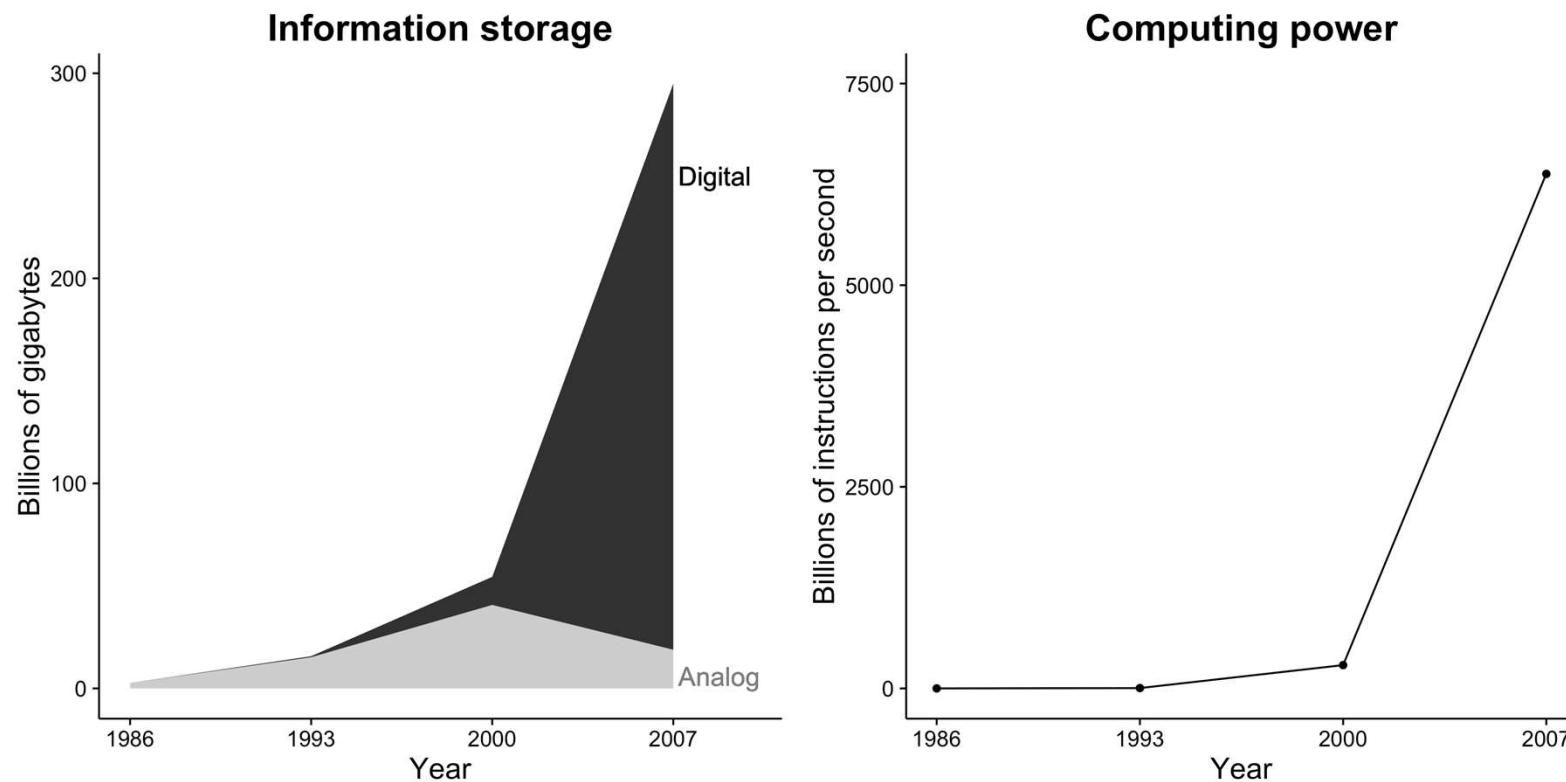
<https://doi.org/10.1038/s41586-020-2909-1>



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

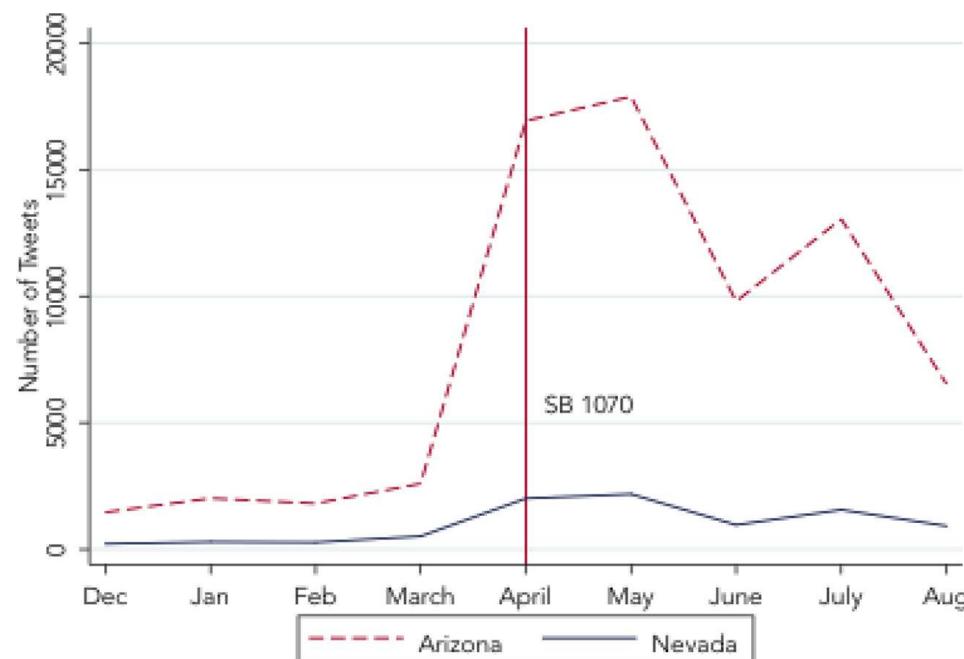


Fig. 2. Number of Twitter messages related to immigrants per month in Arizona and Nevada (December 2010–August 2011). The vertical line on April 2010 indicates when the Arizona governor approved SB 1070.

Non-reactive

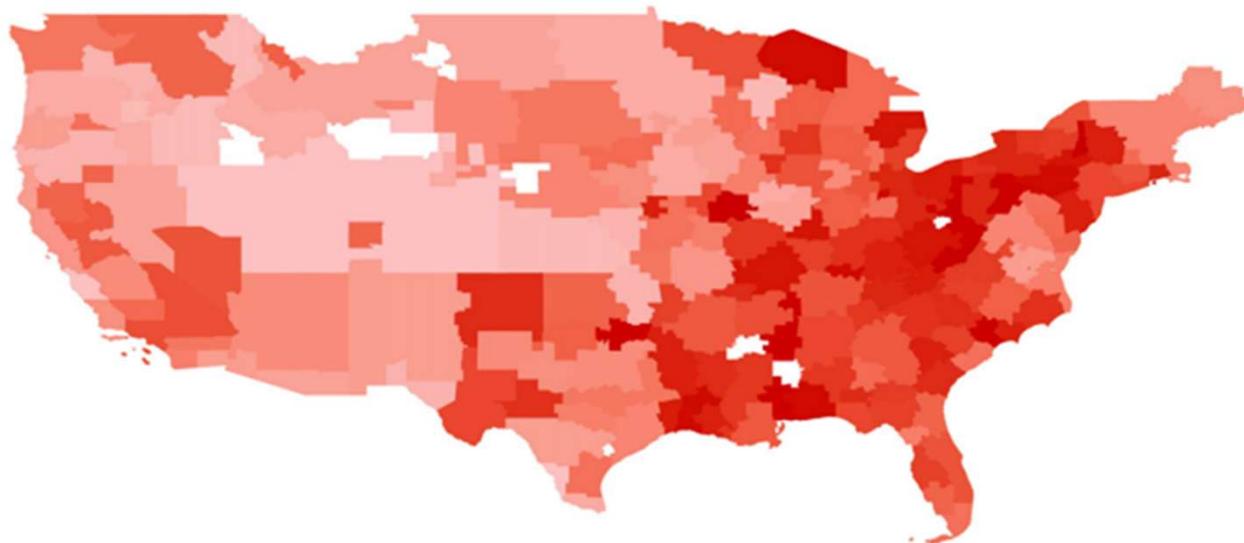


Fig. 2. Racially charged search rate, media market. Notes: This maps search volume for “[Word 1](s),” from 2004 to 2007, at the media market level. Darker areas signify higher search volume. White areas signify media markets without data. Alaska and Hawaii, for which data are available, are not shown

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

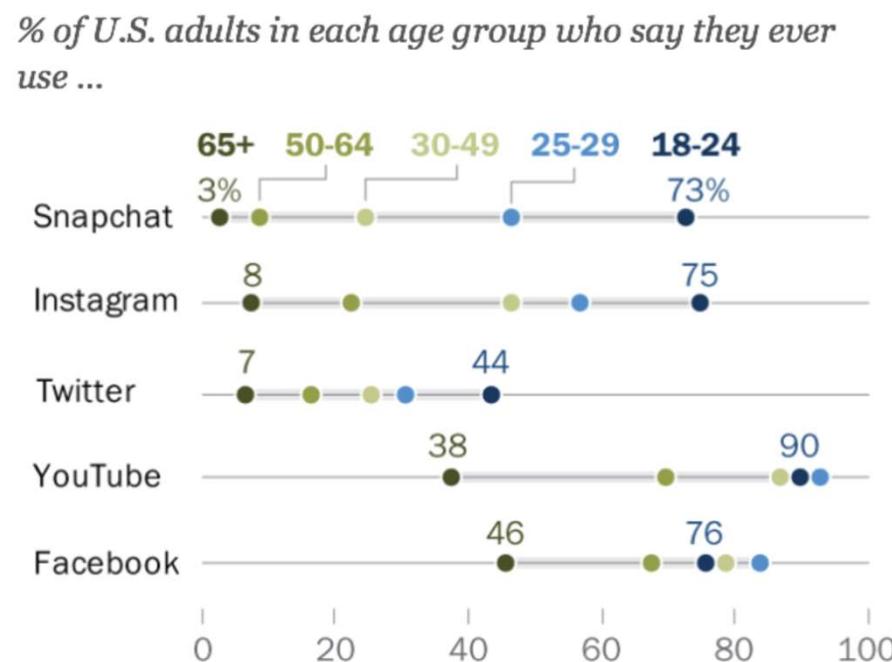
Incomplete



Inaccessible



Non-representative



Note: Respondents who did not give an answer are not shown.

Source: Survey conducted Jan. 8-Feb. 7, 2019.

Drifting



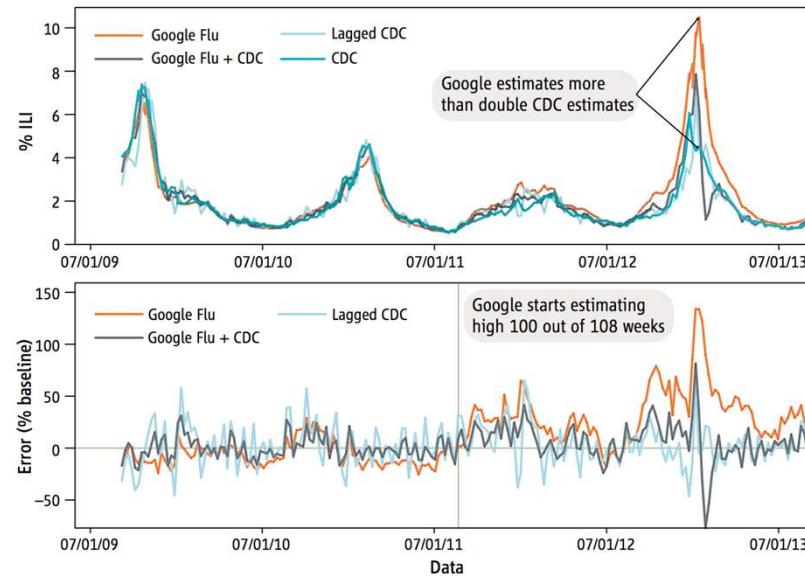
Drifting



Drifting

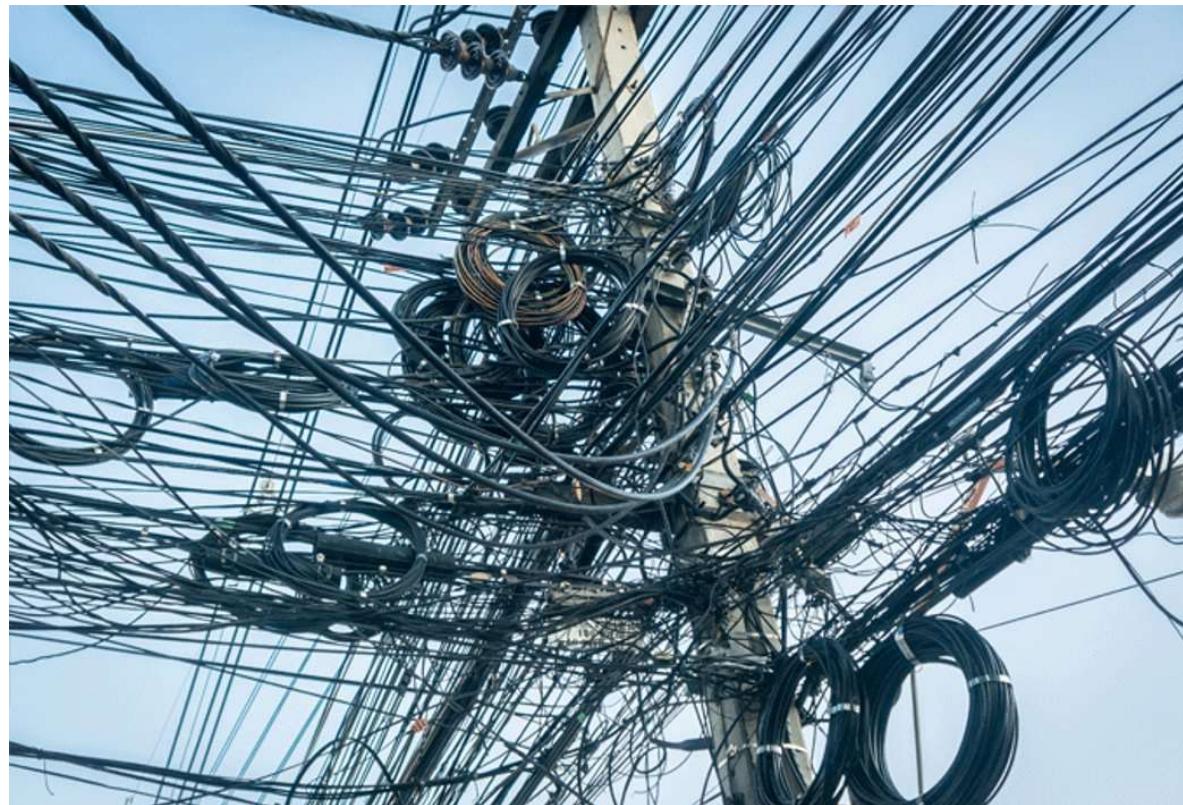


Algorithmically confounded



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 ILL. “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 $\{(\text{Non-CDC estimate}) - (\text{CDC estimate})\}/(\text{CDC estimate})\}$. 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

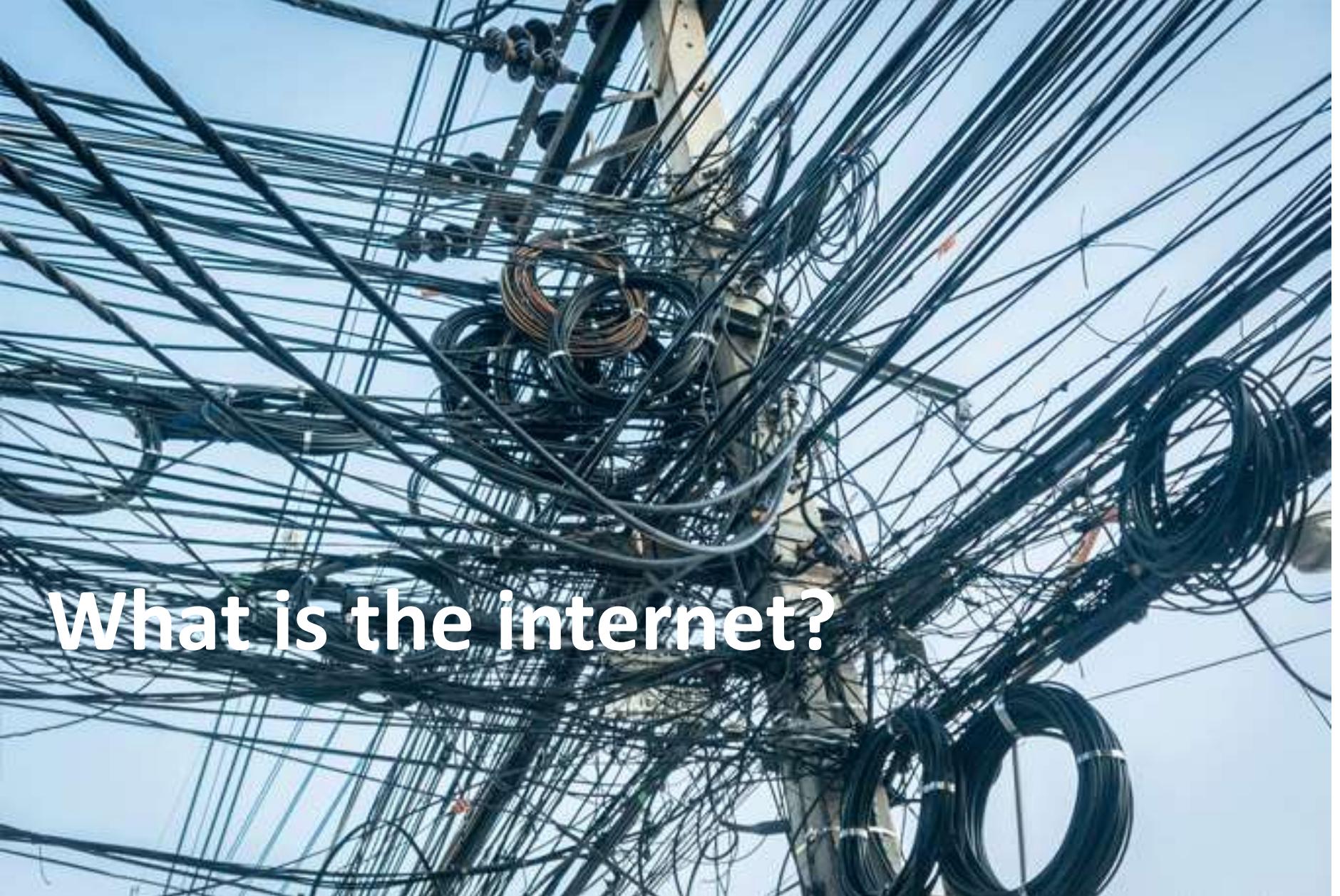
COLUMN

A CV of failures

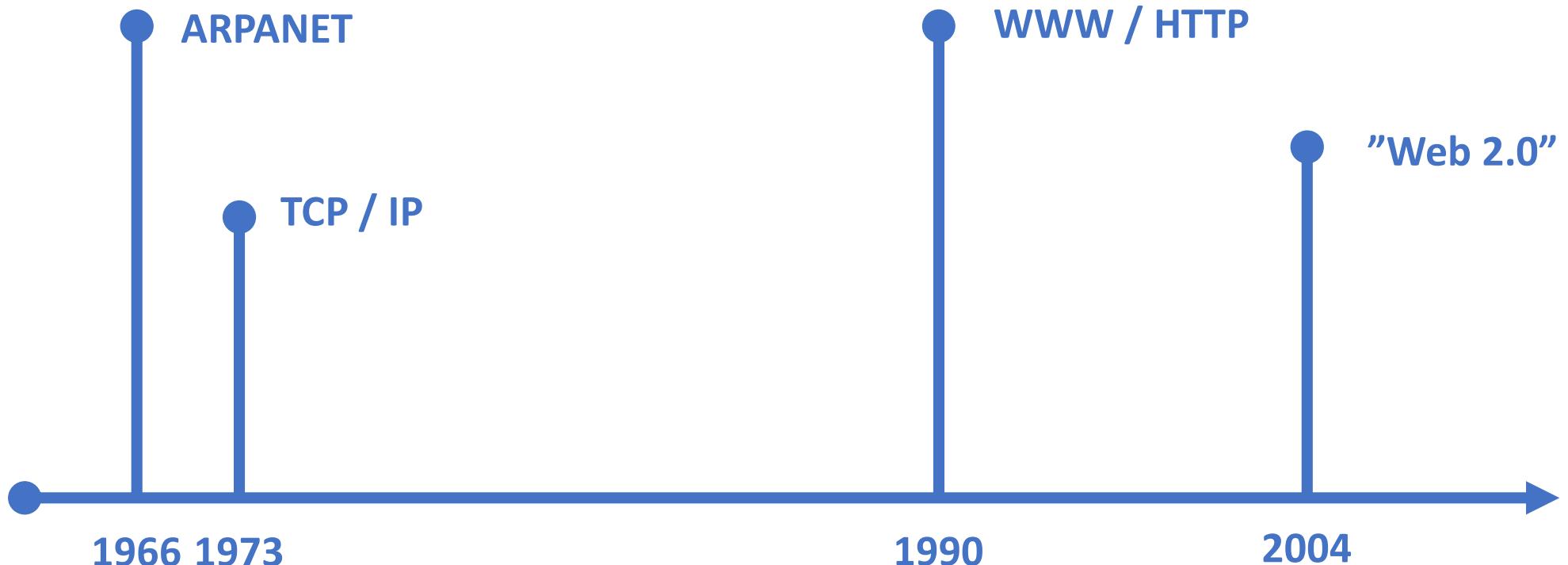
Keeping a visible record of your rejected applications can help others to deal with setbacks, says **Melanie Stefan**.

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

Collecting digital trace data

A photograph showing a dense, chaotic web of electrical wires against a clear blue sky. The wires are primarily black, with some copper-colored ones visible, all crisscrossing and overlapping in a complex, sprawling pattern.

What is the internet?





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

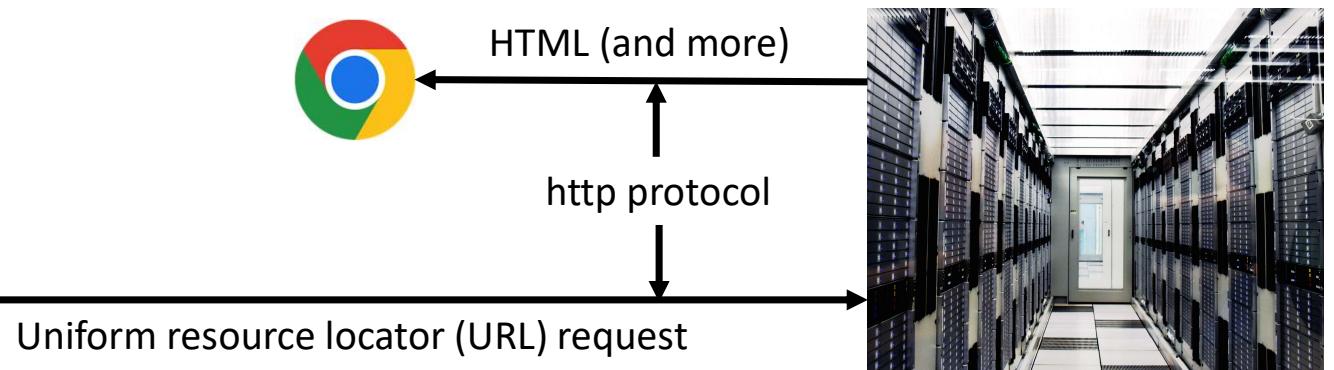


Uniform resource locator (URL) request



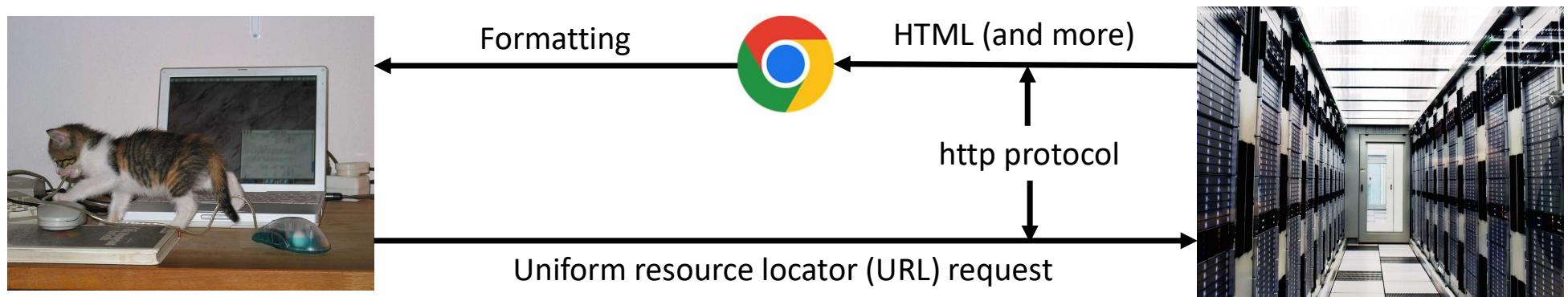
<https://commons.wikimedia.org/wiki/File:Cat-and-computer.JPG>

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

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.

A web page (to you)

The screenshot shows a web browser window with the URL en.wikipedia.org/wiki/List_of_animals_awarded_human_credentials in the address bar. The page title is "List of animals awarded human credentials". The Wikipedia logo is visible on the left. The top right corner shows user options like "Create account" and "Log in". Below the title, there are tabs for "Article" (which is selected) and "Talk". A horizontal menu bar includes "Read", "Edit", "View history", and "Tools". The main content area starts with a paragraph about the list including nonhuman animals who have been awarded diplomas. It then mentions a specific case of a dog named Colby Nolan who was awarded an MBA. The text is presented in a standard Wikipedia article format with links to other pages.

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 taken courses at a [community college](#), worked at a fast food restaurant, [brought](#), and maintained a [newspaper route](#). In response, the institution

A web page (to you)

The screenshot shows a web browser window displaying the Wikipedia article "List of animals awarded human credentials". The page content includes a brief introduction, a section on "Cats", and a detailed paragraph about "Colby Nolan (MBA)". A context menu is open from the top right corner of the browser window, listing various browser functions such as Relaunch to update Chrome, New tab, History, and Developer tools.

en.wikipedia.org/wiki/List_of_animals_awarded_human_credential... Update

WIKIPEDIA
The Free Encyclopedia

List of animals awarded human credentials

Article Talk

From Wikipedia, the free encyclopedia

This **list of animals awarded human credentials** includes nonhuman animals who have been submitted and have been awarded a diploma. On occasion, they have been admitted and granted a degree, as reported by the [University of Pennsylvania](#), used as a device to clearly demonstrate the lax standards or fraudulent activities of the awarding institution. In one case, a dog was granted a degree, which helped lead to a successful fraud prosecution against the institution.

On occasion, accredited institutions award mock degrees to animals, such as 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 2009 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 taken courses at a [community college](#), worked at a [fast food restaurant](#), [babysat](#), and maintained a [newspaper route](#). In response, the institution

Relaunch to update Chrome

New tab Ctrl+T

New window Ctrl+N

New Incognito window Ctrl+Shift+N

History

Downloads Ctrl+J

Bookmarks

Zoom - 100% +

Print... Ctrl+P

Cast...

Find... Ctrl+F

More tools

Edit Cut Copy Paste

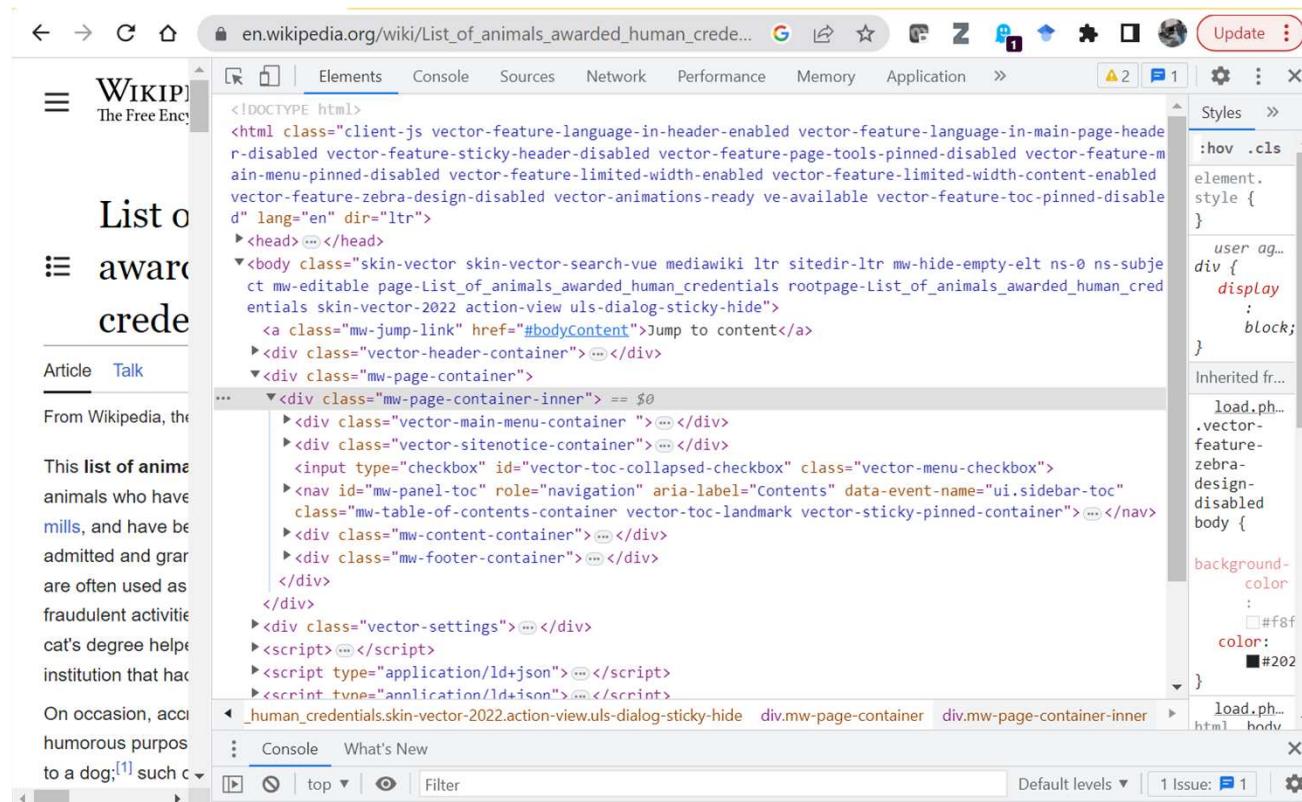
Settings

Help

Exit

Developer tools Ctrl+Shift+I

A web page (to your browser)



A web page (to your browser)

Type of document
(declares format)

Head (meta information
about the document)

WIKIPEDIA

List of animals awarded human credentials

Article Talk

From Wikipedia, the free encyclopedia

This list of animals awarded human credentials includes many species of mammals, and have been admitted and granted awards for various reasons. Some are often used as models in scientific research, while others have been granted honorary degrees by universities.

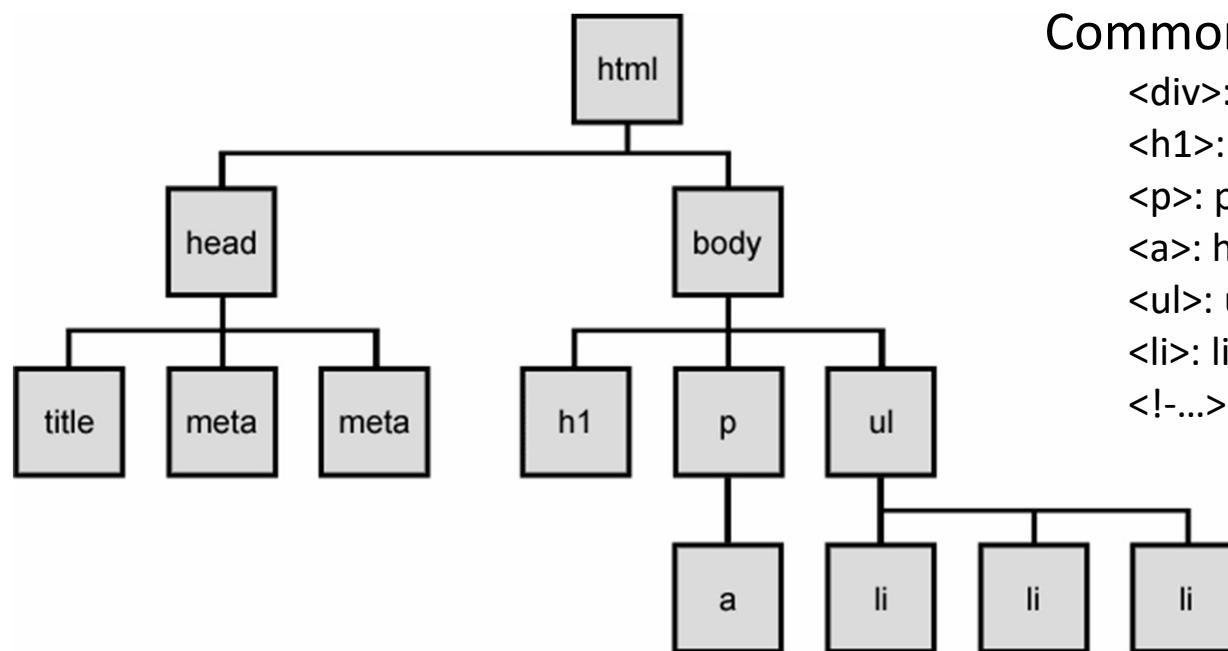
On occasion, animals have been granted awards for humorous purposes, such as giving a dog a

Body (most content is somewhere in here)

Other tags

The screenshot shows a browser window displaying the URL en.wikipedia.org/wiki/List_of_animals_awarded_human_credentials. The developer tools are open, specifically the 'Elements' tab, which displays the HTML source code of the page. The code is heavily minified, with many class names and IDs being long strings of characters. The 'Styles' panel on the right shows some CSS rules, including a rule for the `:hover` pseudo-class. Several arrows point from the explanatory text on the left to specific parts of the DOM tree in the 'Elements' panel. One arrow points to the `<head>` tag, another to the `<body>` tag, and another to a `<script>` tag at the bottom of the tree. The overall structure is complex, reflecting the dynamic nature of the Wikipedia page.

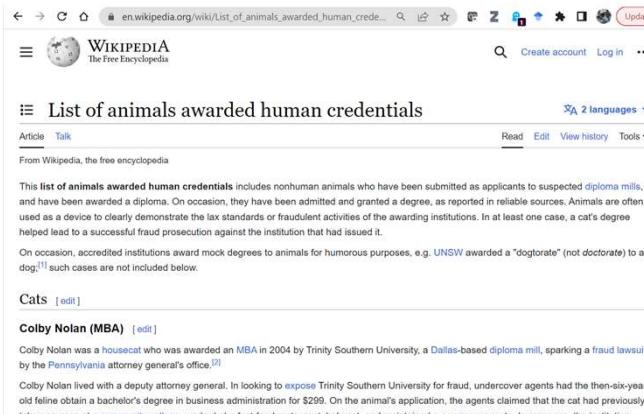
A web page has a tree structure



Common tags

- <div>: 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_cred...

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.

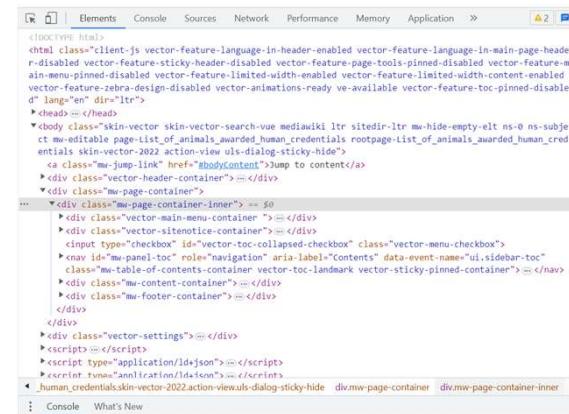
On occasion, accredited institutions award mock degrees to animals for humorous purposes, e.g. UNSW awarded a "dogtorate" (not doctorate) to a dog.^[1] such cases are not included below.

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

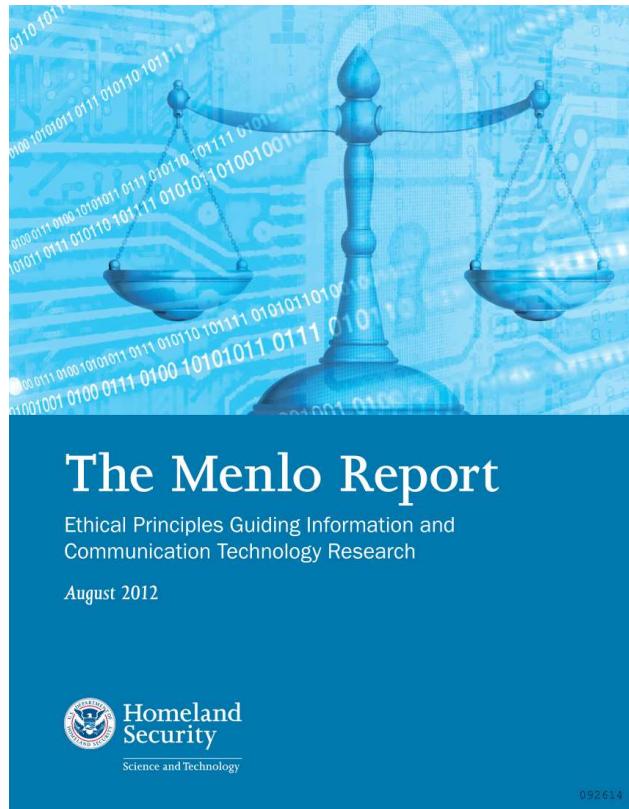


```
<!DOCTYPE html>
<html class="client-js vector-feature-language-in-header-enabled vector-feature-language-in-main-page-headr-disabled vector-feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-main-menu-pinned-disabled vector-feature-limited-width-enabled vector-feature-limited-width-content-enabled vector-feature-zebra-design-disabled vector-animations-ready ve-available vector-feature-toe-pinned-disabled" lang="en" dir="ltr">
  > head
    > body
      > div 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"
        > a class="mw-jump-link" href="#bodyContent">Jump to content</a>
        > div class="vector-header-container"> ... </div>
      > div class="mw-page-container">
        > div class="mw-page-container-inner"> ...
          > div class="vector-main-menu-container"> ... </div>
          > div class="vector-sitetemplate-container"> ... </div>
          > input type="checkbox" id="vector-toe-collapsed-checkbox" class="vector-menu-checkbox"
          > knav id="mw-panel-toc" role="navigation" aria-label="Contents" data-event-name="ui.sidebar-toc"
          > div class="mw-table-of-contents-container vector-toe-landmark vector-sticky-pinned-container"> ... </div>
          > div class="mw-content-container"> ... </div>
          > div class="mw-footer-container"> ... </div>
        > /div>
      > div class="vector-settings"> ... </div>
      > script> ... </script>
      > script type="application/json"> ... </script>
      > script type="application/javascript"> ... </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
```

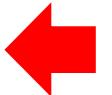


Screen-scraping

But first, back to yesterday!



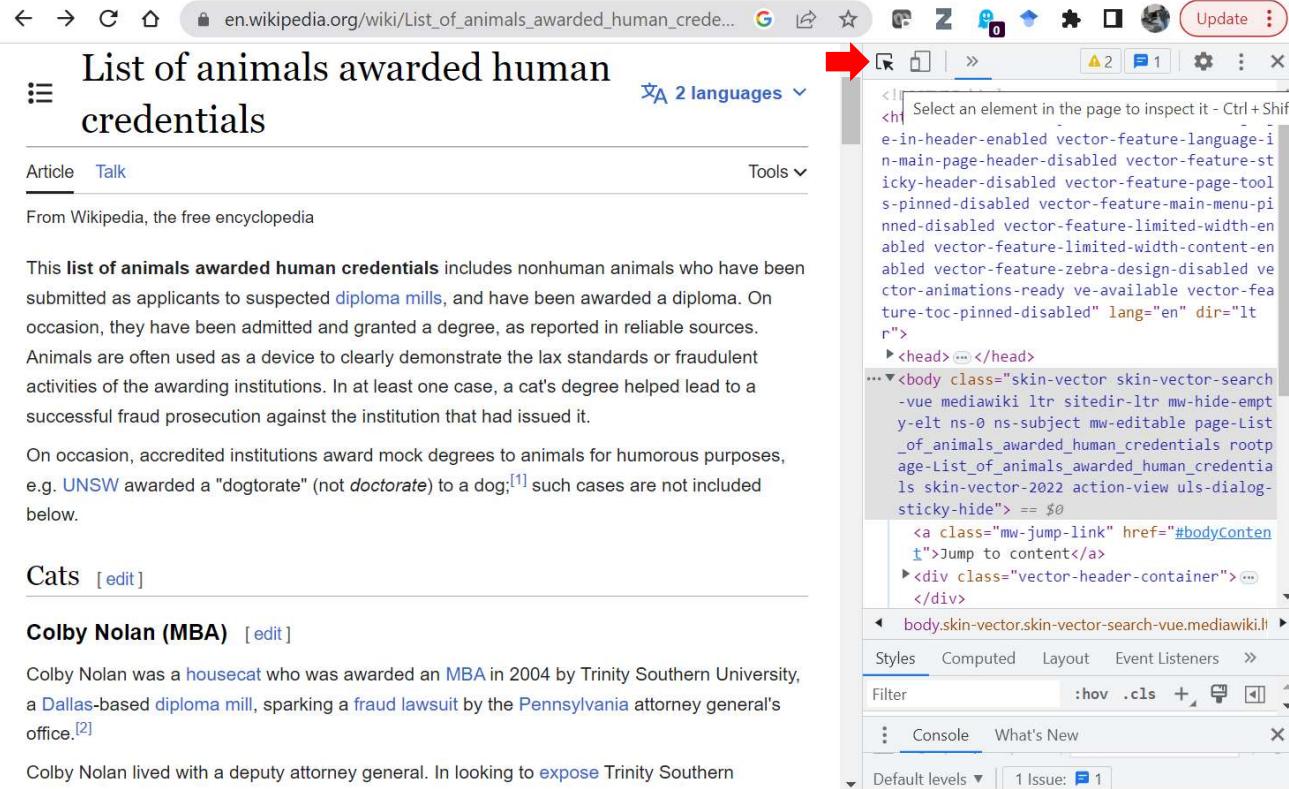
- Respect for persons: participants decide.
- Beneficence: minimize risk, maximize benefits.
- Justice: distribute the burdens and benefits of research.
- Respect for law and public interest: compliance and transparency.



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



The screenshot shows a web browser window displaying the Wikipedia page for "List of animals awarded human credentials". The browser's address bar shows the URL: en.wikipedia.org/wiki/List_of_animals_awarded_human_credential... . The page title is "List of animals awarded human credentials". Below the title, there are links for "Article" and "Talk", and a "Tools" dropdown menu. The main content area describes the list of animals awarded human credentials, mentioning nonhuman animals who have been granted diplomas by diploma mills. It also notes that animals are often used as devices to demonstrate lax standards or fraudulent activities. Below this, a section titled "Cats" lists "Colby Nolan (MBA)". The "Colby Nolan (MBA)" section includes a brief biography and a note about her living with a deputy attorney general. On the right side of the browser, the developer tools are open, specifically the element inspector. A red arrow points to the "Select an element in the page to inspect it - Ctrl + Shift e-in-header-enabled vector-feature-language-i n-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" text in the element list. The element inspector interface includes tabs for "Styles", "Computed", "Layout", and "Event Listeners", along with a "Filter" bar and a "Console" tab at the bottom.

Parsing HTML

The screenshot shows a browser window displaying a Wikipedia page about animals awarded human credentials. The page lists several cats and their achievements. The developer tools are open, specifically the Element tab of the Inspector panel, which shows the HTML structure of the page. The DOM tree is visible on the right side, with various elements like `<h3>`, ``, and `<p>` nodes. The element `<p> ... </p> == $0` is highlighted in orange, indicating it's the current selection. On the left, the element inspector shows detailed information for the selected element, including its bounding box (590.67 x 156.77), color (#202122), font (14px sans-serif), and margin (7px 0px). It also includes an Accessibility section with fields for Name, Role (paragraph), and Keyboard-focusable.

Collins (high school diploma) [edit]

C. Collins (born around 2007) is a [tuxedo cat](#) who gained notoriety when she received a diploma from Jefferson High School Online in 2009, although her age was misrepresented in order to qualify.^[12] The sting was an investigative operation by the [Better Business Bureau](#) of Central Georgia headed by Kelvin Collins, Oreo's owner.^[13]

Zoe D. Katze (psychotherapist and hypnotherapy certifications) [edit]

Zoe D. Katze ("Zoe the Cat" in German) was a housecat owned by psychologist Steve K. D. Eichel. Around 2001, Eichel obtained a psychotherapy certification for his cat from the American Psychotherapy Association and several hypnotherapy credentials from other organizations.^{[14][8]} The certification of Zoe has been cited in several books and articles on credentialing scams, and has appeared in psychology and forensic curricula. Eichel also served as the consultant to the BBC investigation that led to the certification of George the cat by several UK hypnosis associations.^[8]

p 590.67 x 156.77
Color #202122
Font 14px sans-serif
Margin 7px 0px

ACCESSIBILITY
Name
Role paragraph
Keyboard-focusable

MBA) [edit]

A, a pug from Vermont, was awarded an MBA by [Rochville University](#). application and US\$499 and received a "diploma, two sets of of distinction in finance, and a certificate of membership in the student

Lulu (college diploma) [edit]

Styles Computed Layout Event Listeners >
Filter :hov .cls + □
Console What's New
Default levels ▾ | 1 Issue: 1

Parsing HTML

The screenshot shows a web browser window displaying the Wikipedia page "List of animals awarded human credentials". The URL in the address bar is en.wikipedia.org/wiki/List_of_animals_awarded_human_credentials. A context menu is open over a paragraph of text, specifically over the word "certifications". The menu is a developer tool from Firebug, showing the DOM structure and various options like Copy element, Copy XPath, and Copy full XPath.

Collins (high school diploma) [edit]
C. Collins (born around 2007) is a [tuxedo cat](#) who gained notoriety when she received a diploma from Jefferson High School Online in 2009, although her age was misrepresented in order to qualify.^[12] The sting was an investigative operation by the [Better Business Bureau](#) of Central Georgia headed by Kelvin Collins, Oreo's owner.^[13]

Zoe D. Katze ("Zoe the Cat" in German) was a housecat owned by psychologist Steve K. D. Eichel. Around 2001, Eichel obtained a psychotherapy certification for his cat from the American Psychotherapy Association and several hypnotherapy credentials from other organizations.^{[14][8]} The certification of Zoe has been cited in several books and articles on credentialing scams, and has appeared in psychology and forensic curricula. Eichel also served as the consultant to the BBC investigation that led to the certification of George the cat by several UK hypnosis associations.^[8]

Dogs [edit]

Chester Ludlow (MBA) [edit]
In 2009, Chester Ludlow, a pug from Vermont, was awarded an MBA by [Rochville University](#). His owner submitted an application and US\$499 and received a "diploma, two sets of transcripts, a certificate of distinction in finance, and a certificate of membership in the student council".^[15]

Lulu (college diploma) [edit]

Parsing HTML

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

C. Collins (high school diploma) [edit]

C. Collins (born around 2007) is a [tuxedo cat](#) who gained notoriety when she received a diploma from Jefferson High School Online in 2009, although her age was misrepresented in order to qualify.^[12] The sting was an investigative operation by the [Better Business Bureau](#) of Central Georgia headed by Kelvin Collins, Oreo's owner.^[13]

p 590.67 × 156.77 **psychotherapist and hypnotherapy certifications** [edit]

Zoe D. Katze ("Zoe the Cat" in German) was a housecat owned by psychologist Steve K. D. Eichel. Around 2001, Eichel obtained a psychotherapy certification for his cat from the

Dogs [edit]

Chester Ludlow (MBA) [edit]

In 2009, Chester Ludlow, a pug from Vermont, was awarded an MBA by [Rochville University](#). His owner submitted an application and US\$499 and received a "diploma, two sets of transcripts, a certificate of distinction in finance, and a certificate of membership in the student council".^[15]

Lulu (college diploma) [edit]

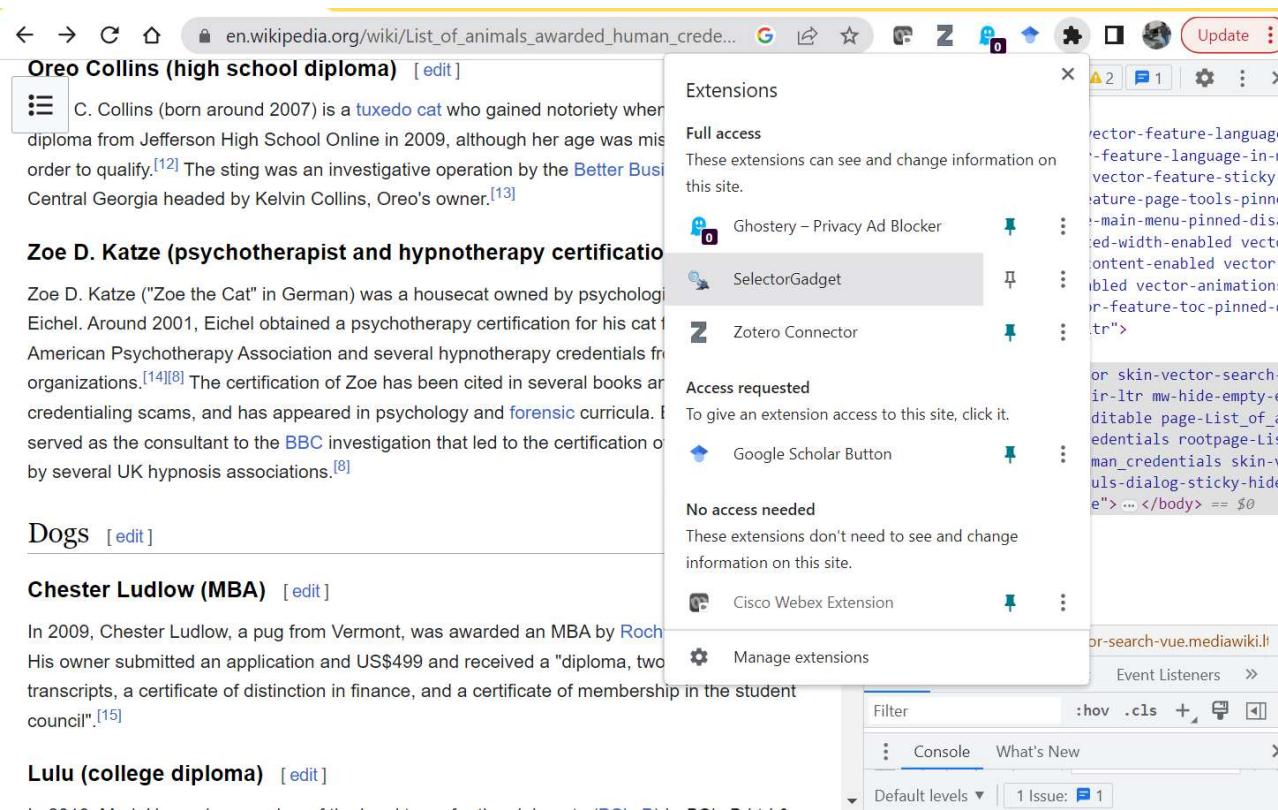
content.mw-content-

- Copy element
- Copy outerHTML
- Copy selector
- Copy JS path
- Copy styles
- Copy XPath
- Copy full XPath
- Expand recursively
- Hide element
- Force state
- Break on
- Store as global variable

- Styles Computed
- Filter
- Console What's

Default levels | 1 Iss

Parsing with the CSS Selector



Parsing with the CSS Selector

The screenshot shows a browser window displaying a Wikipedia page titled "List of animals awarded human credentials". The page lists several animals, each with a brief description. The developer tools panel is open on the right side, showing the DOM tree and various inspection and editing tools.

Oreo Collins (high school diploma) [edit]
C. Collins (born around 2007) is a [tuxedo cat](#) who gained notoriety when she received a diploma from Jefferson High School Online in 2009, although her age was misrepresented in order to qualify.^[12] The sting was an investigative operation by the [Better Business Bureau](#) of Central Georgia headed by Kelvin Collins, Oreo's owner.^[13]

Zoe D. Katze (psychotherapist and hypnotherapy certifications) [edit]
Zoe D. Katze ("Zoe the Cat" in German) was a housecat owned by psychologist Steve K. D. Eichel. Around 2001, Eichel obtained a psychotherapy certification for his cat from the American Psychotherapy Association and several hypnotherapy credentials from other organizations.^{[14][15]} The certification of Zoe has been cited in several books and articles on credentialing scams, and has appeared in psychology and [forensic](#) curricula. Eichel also served as the consultant to the [BBC](#) investigation that led to the certification of [George the cat](#) by several UK hypnosis associations.^[8]

Dogs [edit]

Chester Ludlow (MBA) [edit]
In 2009, Chester Ludlow, a pug from Vermont, was awarded an MBA by [Rochville University](#). His owner submitted an application and US\$499 and received a "diploma, two sets of transcripts, a certificate of distinction in finance, and a certificate of membership in the student

Filter :
p
Clear (21)
Toggle Position XPath ? X

Styles Computed Layout Event Listeners ▾
body.skin-vector.skin-vector-search-vue.mediawiki.ltr
Filter :
Console What's New
Filter Default levels ▾

```
<!DOCTYPE html>
<html lang="en" dir="ltr">
  <head> ... </head>
  <body class="skin-vector skin-vector-search-vue mediawiki ltr sitedir-ltr mw-hide-empty-el et ns-0 ns-subject mw-editable page-List_of_an imals_awarded_human_credentials rootpage-List _of_animals_awarded_human_credentials skin-ve ctor-2022 action-view uts-dialog-sticky-hide vector-below-page-title"> ... </body> == $0
</html>
```

Parsing with the CSS Selector

The screenshot shows a browser window with the URL en.wikipedia.org/wiki/List_of_animals_awarded_human_credentials. The page content is displayed with several sections highlighted by yellow and green boxes. A red box highlights the section about **Zoe D. Katze**. A green box highlights the section about **Dogs**. The developer tools panel on the right shows the DOM tree with various classes and IDs. The bottom left of the developer tools panel shows a search bar with the text "h3" and a "Clear (15)" button, along with tabs for "Styles", "Computed", "Layout", and "Event Listeners".

Oreo Collins (high school diploma) [edit]

C. Collins (born around 2007) is a tuxedo cat who gained notoriety when she received a diploma from Jefferson High School Online in 2009, although her age was misrepresented in order to qualify.^[12] The sting was an investigative operation by the Better Business Bureau of Central Georgia headed by Kelvin Collins, Oreo's owner.^[13]

Zoe D. Katze (psychotherapist and hypnotherapy certifications) [edit]

Zoe D. Katze ("Zoe the Cat" in German) was a housecat owned by psychologist Steve K. D. Eichel. Around 2001, Eichel obtained a psychotherapy certification for his cat from the American Psychotherapy Association and several hypnotherapy credentials from other organizations.^{[14][8]} The certification of Zoe has been cited in several books and articles on credentialing scams, and has appeared in psychology and forensic curricula. Eichel also served as the consultant to the BBC investigation that led to the certification of George the cat by several UK hypnosis associations.^[8]

Dogs [edit]

Chester Ludlow (MBA) [edit]

In 2009, Chester Ludlow, a pug from Vermont, was awarded an MBA by [Rochville University](#). His owner submitted an application and US\$499 and received a "diploma, two sets of transcripts, a certificate of distinction in finance, and a certificate of membership in the student

```
<!DOCTYPE html>
<html class="client-js vector-feature-language-in-header-enabled vector-feature-language-in-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-el 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 vector-below-page-title"> ... </body> == $0
</html>
```

h3

Toggle Position XPath ? X

Clear (15)

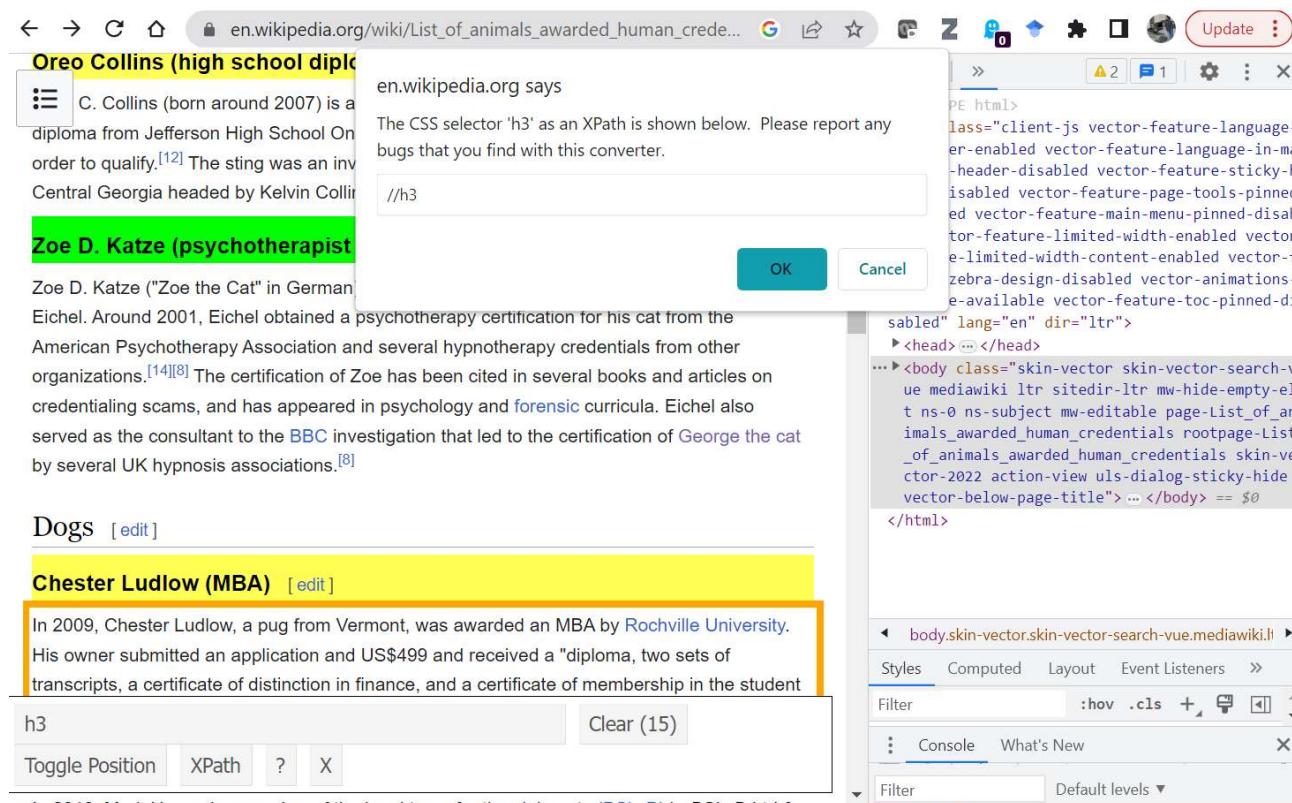
Styles Computed Layout Event Listeners >

Filter :hover .cls + ↻

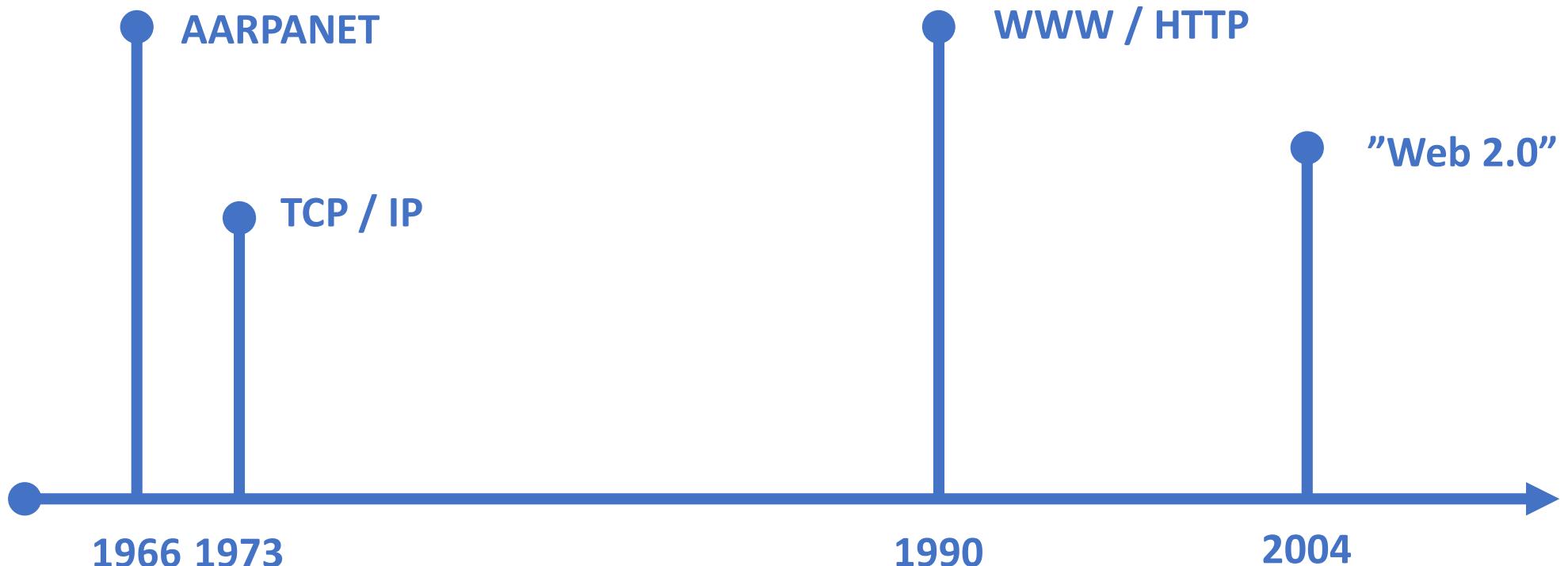
Console What's New

Default levels | 1 Issue: 1

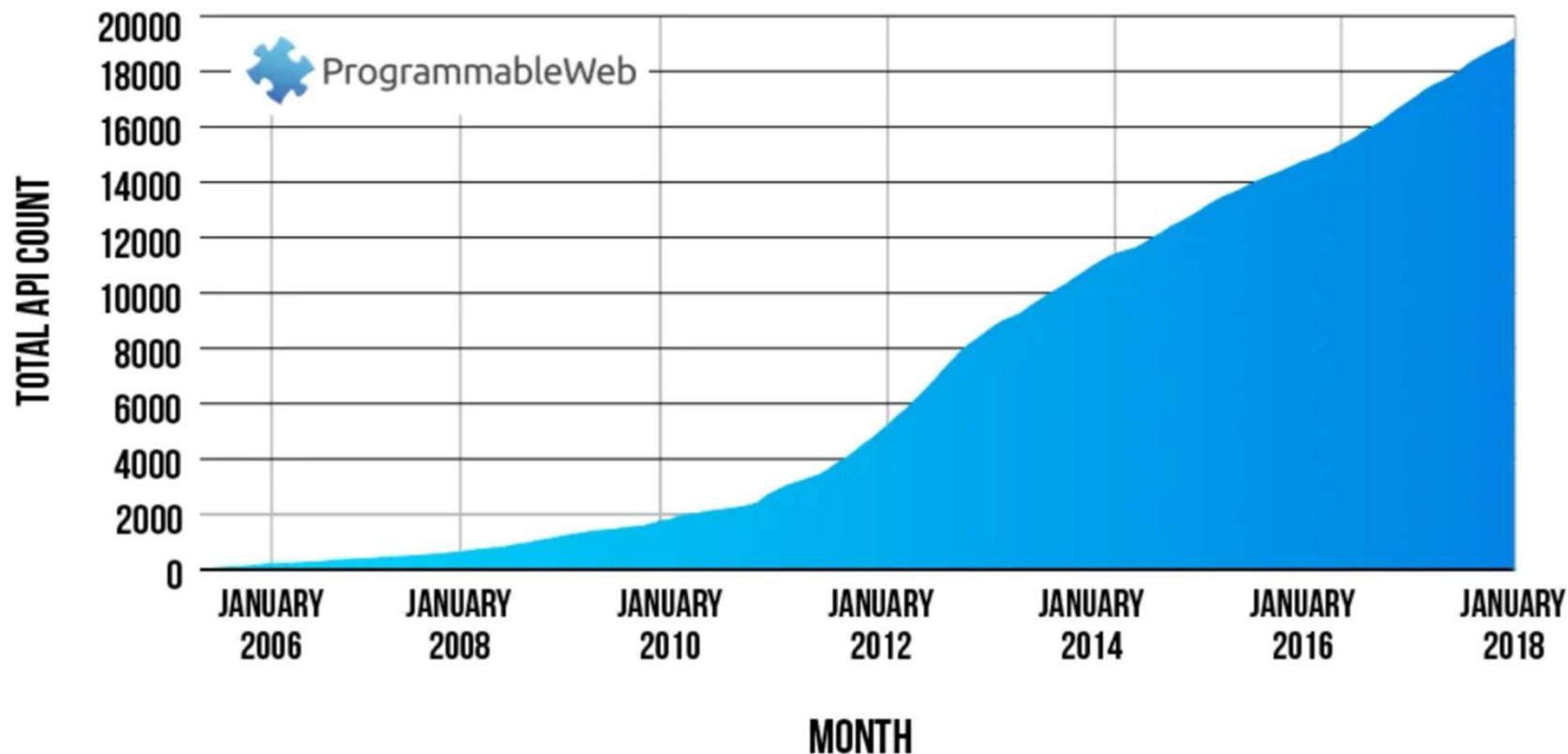
Parsing with the CSS Selector



Application programming interfaces (APIs)



GROWTH IN WEB APIs SINCE 2005



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



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



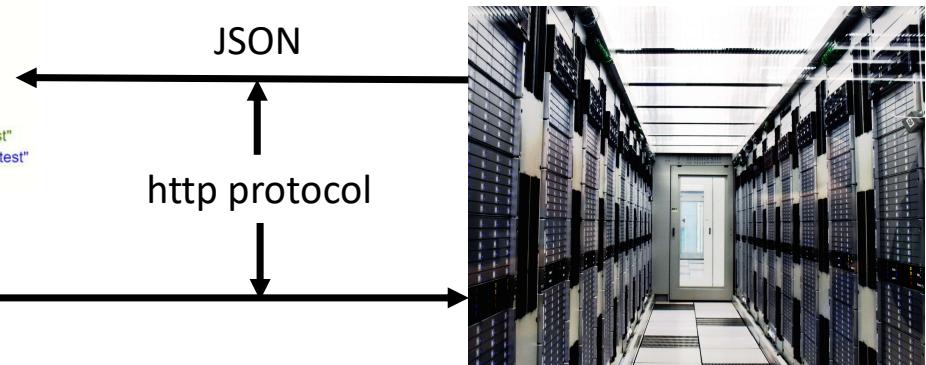
<https://content.guardianapis.com/search?from-date=2023-05-12&q=cats&api-key=...>

<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



```
▼ root: {} 1 key
  ▼ response: {} 9 keys
    status: "ok"
    userTier: "developer"
    total: 60
    startIndex: 1
    pageSize: 10
    currentPage: 1
    pages: 6
    orderBy: "relevance"
  ▼ results: [] 10 items
    ▼ 0: {} 9 keys
      id: "test/2023/may/24/funny-cats-video-test"
      type: "article"
      sectionId: "test"
      sectionName: "Test"
      webPublicationDate: "2023-05-24T13:12:42Z"
      webTitle: "funny cats video test"
      webUrl: "https://www.theguardian.com/test/2023/may/24/funny-cats-video-test"
      apiUrl: "https://content.guardianapis.com/test/2023/may/24/funny-cats-video-test"
      isHosted: false
```



<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

Test

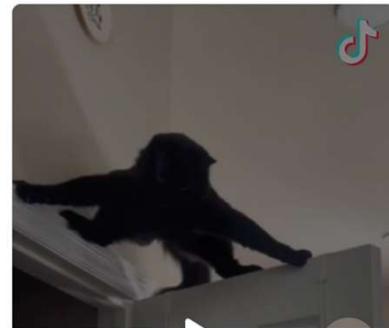
funny cats video test

This is a test article

Wed 24 May 2023 14.12 BST

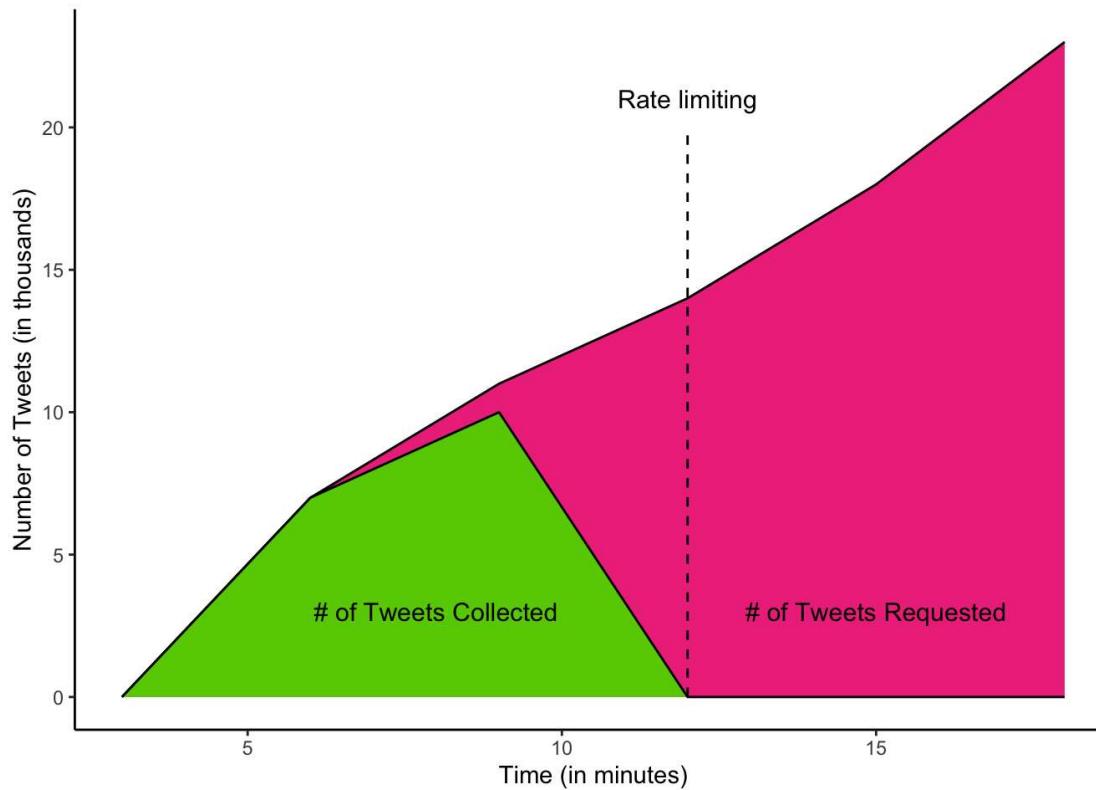


This is simply a test article for dealing with tiktok embeds.

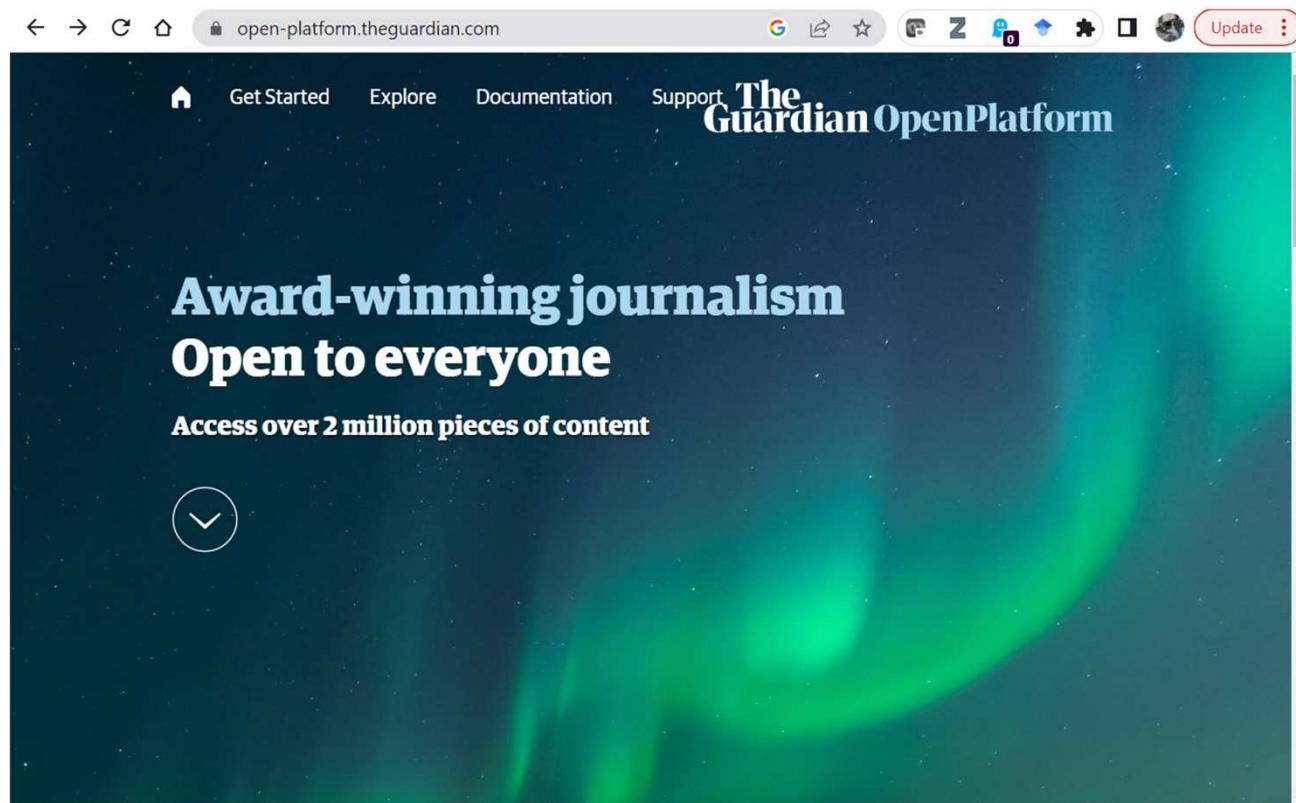


<https://www.theguardian.com/test/2023/may/24/funny-cats-video-test>
<https://open-platform.theguardian.com/explore/>

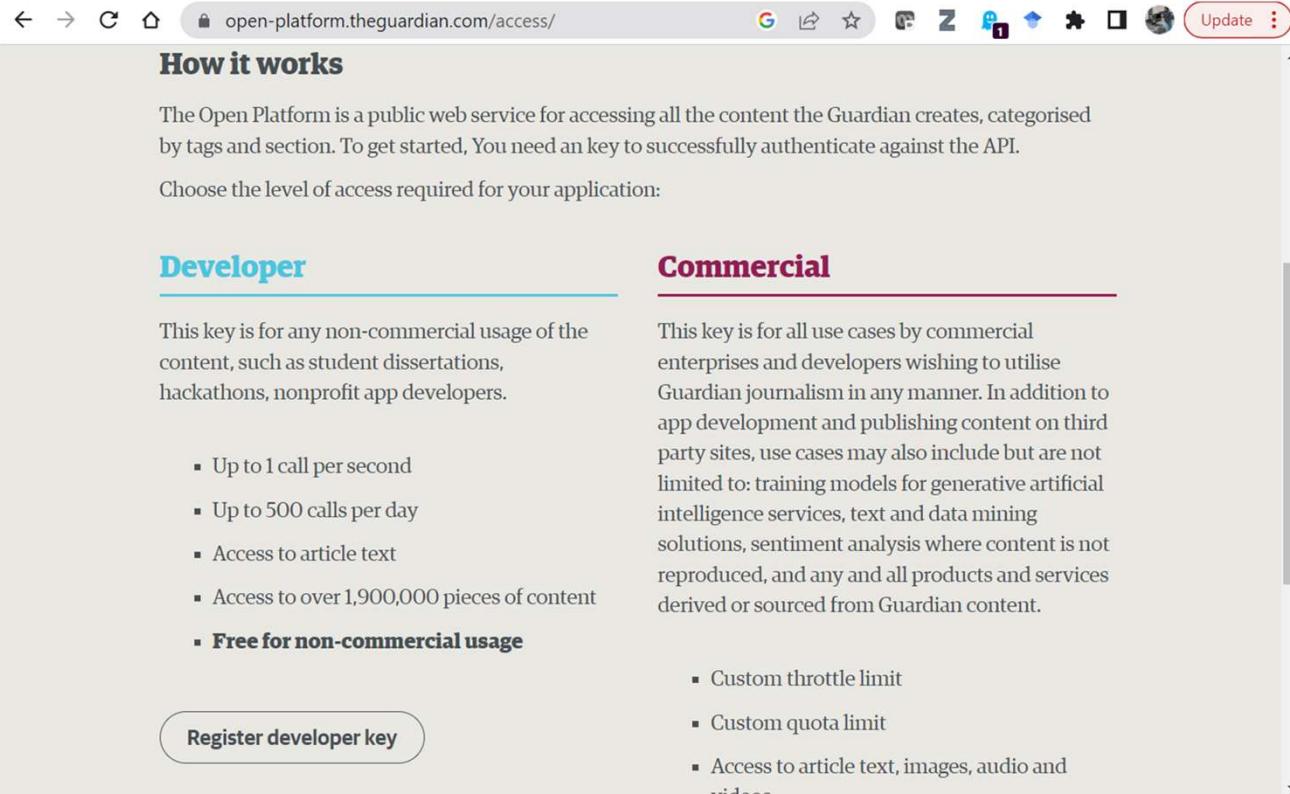
API restrictions



API registration



API registration



The screenshot shows a web browser window with the URL open-platform.theguardian.com/access/. The page title is "How it works". The content explains that the Open Platform is a public web service for accessing all the content the Guardian creates, categorized by tags and section. It requires an API key for authentication. Below this, there are two sections: "Developer" and "Commercial".

Developer

This key is for any non-commercial usage of the content, such as student dissertations, hackathons, nonprofit app developers.

- Up to 1 call per second
- Up to 500 calls per day
- Access to article text
- Access to over 1,900,000 pieces of content

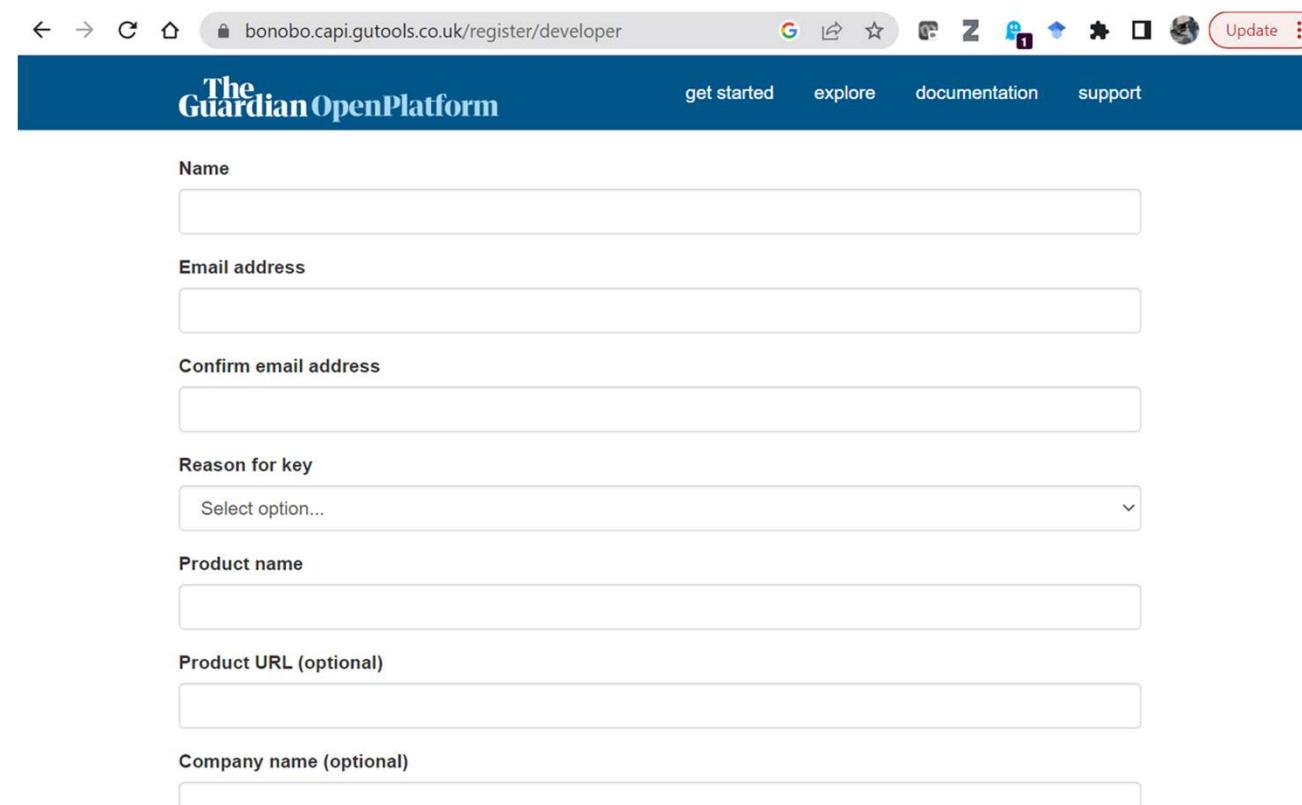
Commercial

This key is for all use cases by commercial enterprises and developers wishing to utilise Guardian journalism in any manner. In addition to app development and publishing content on third party sites, use cases may also include but are not limited to: training models for generative artificial intelligence services, text and data mining solutions, sentiment analysis where content is not reproduced, and any and all products and services derived or sourced from Guardian content.

- Custom throttle limit
- Custom quota limit
- Access to article text, images, audio and videos

[Register developer key](#)

API registration



The screenshot shows a web browser window with the URL `bonobo.capi.gutools.co.uk/register/developer` in the address bar. The page has a dark blue header with the text "The Guardian OpenPlatform". Below the header are navigation links: "get started", "explore", "documentation", and "support". On the right side of the header is a red "Update" button with three dots. The main content area contains several input fields:

- Name**: An input field with a placeholder text area below it.
- Email address**: An input field with a placeholder text area below it.
- Confirm email address**: An input field with a placeholder text area below it.
- Reason for key**: A dropdown menu with the placeholder "Select option...".
- Product name**: An input field with a placeholder text area below it.
- Product URL (optional)**: An input field with a placeholder text area below it.
- Company name (optional)**: An input field with a placeholder text area below it.

More APIs

- Lists of APIs
 - <https://docs.google.com/spreadsheets/d/1ZErl3okdlb0zctmX0MZKo-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?

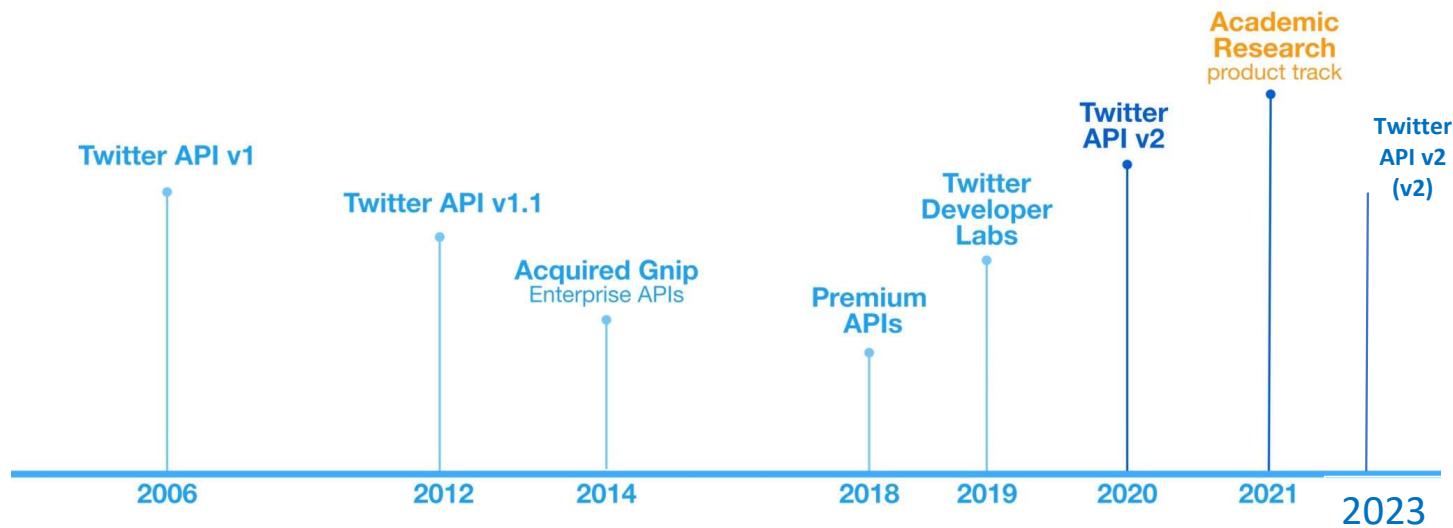
What's up with Twitter?



What's up with Twitter?



A brief history of the Twitter API



<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