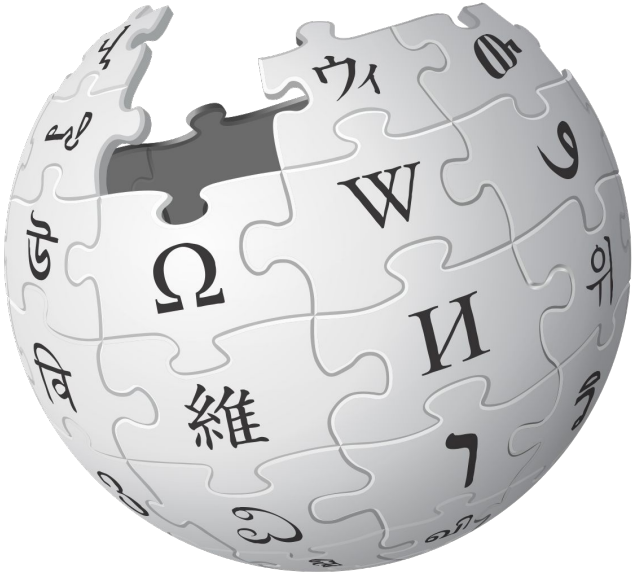
Traffic Stitching

Wikipedia

Project dedicated to the building of free encyclopedias in all languages of the world.

As of December 2019, there are **309 language** editions of Wikipedia, which collectively have more than **50 million articles** that have been collectively **viewed 15 billion times**.



It **big.**

WIKIMEDIA FOUNDATION

Wikimedia Foundation

~380 people working together to:

- **keep knowledge accessible** by everyone, everywhere, all the time
- **keep knowledge free** from advertising
- actively **improve the diversity** of voices
- **make it easier to participate** in the free knowledge movement

https://commons.wikimedia.org/wiki/File:Wikimedia_All_Hands_2019_Group_Photo.jpg

Photo by Myleen Hollero [CC BY-SA 3.0]

WIKIMEDIA
FOUNDATION

Product Analytics team

- Empowering others to make data-informed decisions through education & self-service analytics tools
- Extracting insights from the Foundation's data repositories
- Crafting Key Performance Indicators (KPIs)* and other metrics
- Building dashboards for tracking success
- Design and analysis of experiments (A/B tests)
- Ad-hoc analyses and machine learning projects
- Develop tools & software for working with data

WIKIMEDIA FOUNDATION

*** Tip:** learn this term, the industry **loves** this term

2 main categories of questions

editing (editors, content)
or **readership** (traffic, time spent reading)

WIKIMEDIA
FOUNDATION

Editing questions

- How many active editors are there? How many new editors have registered?
- How many articles are there? How many are new?
- How many edits from new editors have been reverted vs not reverted?

Pretty easy to find out yourself by going to stats.wikimedia.org/v2/

or querying the public versions of databases directly at quarry.wmflabs.org/

WIKIMEDIA FOUNDATION

Readership questions

- **Long term trends and changes in patterns of traffic**
  - Desktop vs mobile
  - Topics of interest (celebrities, world events)
  - Sources of traffic (countries, search engines)

WIKIMEDIA
FOUNDATION

Unfortunately...

WIKIMEDIA
FOUNDATION

Answering that is **hard**

But **not** for the reasons you expect!

2007-2016: legacy pagecounts

2015-present: modern counts

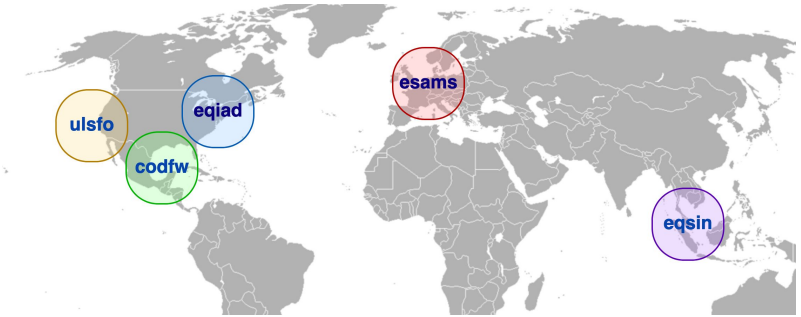
Visiting Wikipedia.org

1. Browser uses a DNS server to translate “https://wikipedia.org” into an IP address
2. Browser makes **HTTP GET** request to this IP address at **/** (root, defaults to /index.html)
3. Server responds to browser with an HTTP response
    - **Status code** & **headers** like Content-Type (“text/html”)
    - Response **body** contains HTML for the page
4. This HTML contains marked-up text and links to resources like images, JavaScript files, CSS files
5. Browser requests each linked resource via individual HTTP GET requests to same server
6. Server responds with the resources, browser displays & executes them as they are loaded

Visiting Wikipedia.org

2. Browser makes **HTTP GET** request to this IP address at **/** (root, defaults to /index.html)

- Request goes to nearest of 5 our data centers
- If requested page is cached, cache is returned
- If not cached, maybe forwarded to another, “application” data center for rendering from PHP
- **Request details (& full path it took) are logged**
- Response is returned

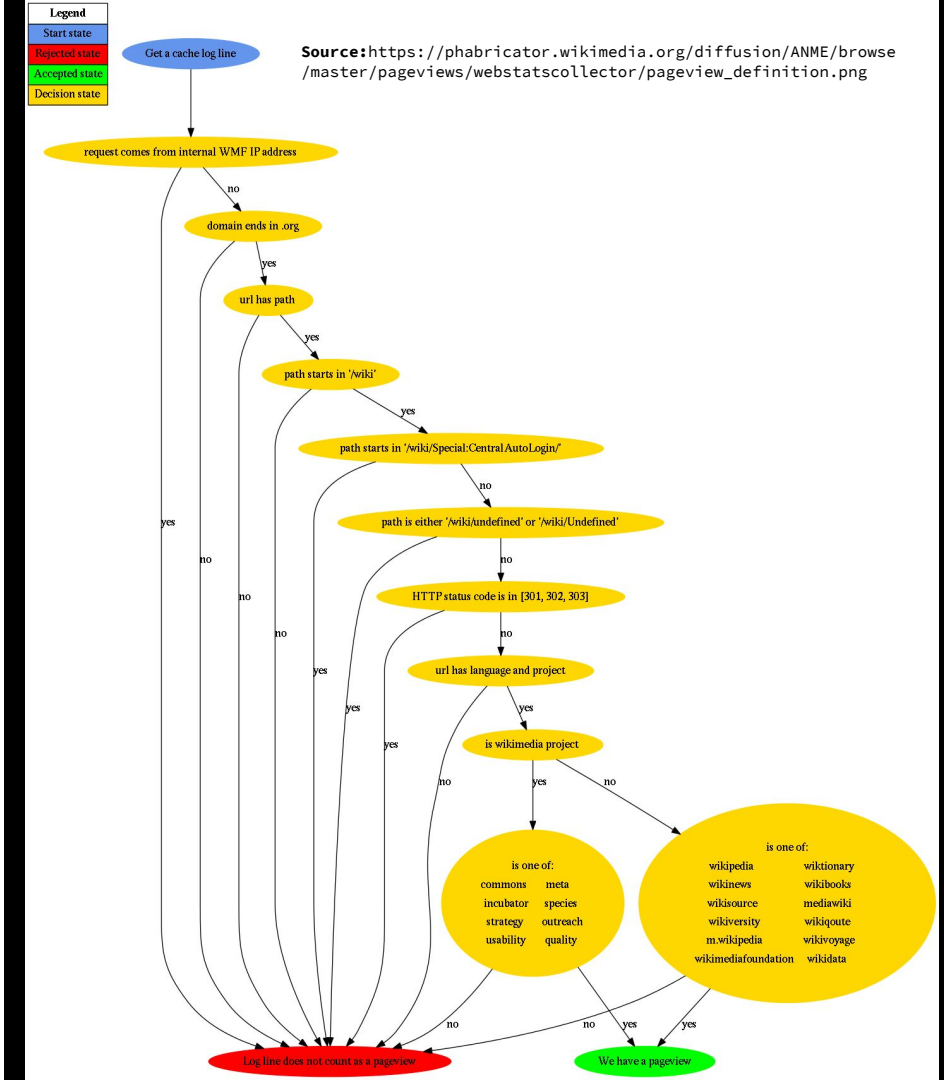


3. Server responds to browser with an HTTP response
    - **Status code** & **headers** like Content-Type (“text/html”)
    - Response **body** contains HTML for the page

WIKIMEDIA FOUNDATION

2013 flowchart of whether a logged web request is a pageview

WIKIMEDIA FOUNDATION



What is a pageview?

- HTTP status `200 OK` or `304 Not Modified`
- `MIME` type is text/html
- …and a few other conditions mostly involving the URL; see meta.wikimedia.org/wiki/Research:Page_view for more info

WIKIMEDIA FOUNDATION

main difference among legacy pagecounts and the current pageview data is lack of filtering of self-reported bots, ***thus automated and human traffic are reported together***

WIKIMEDIA
FOUNDATION

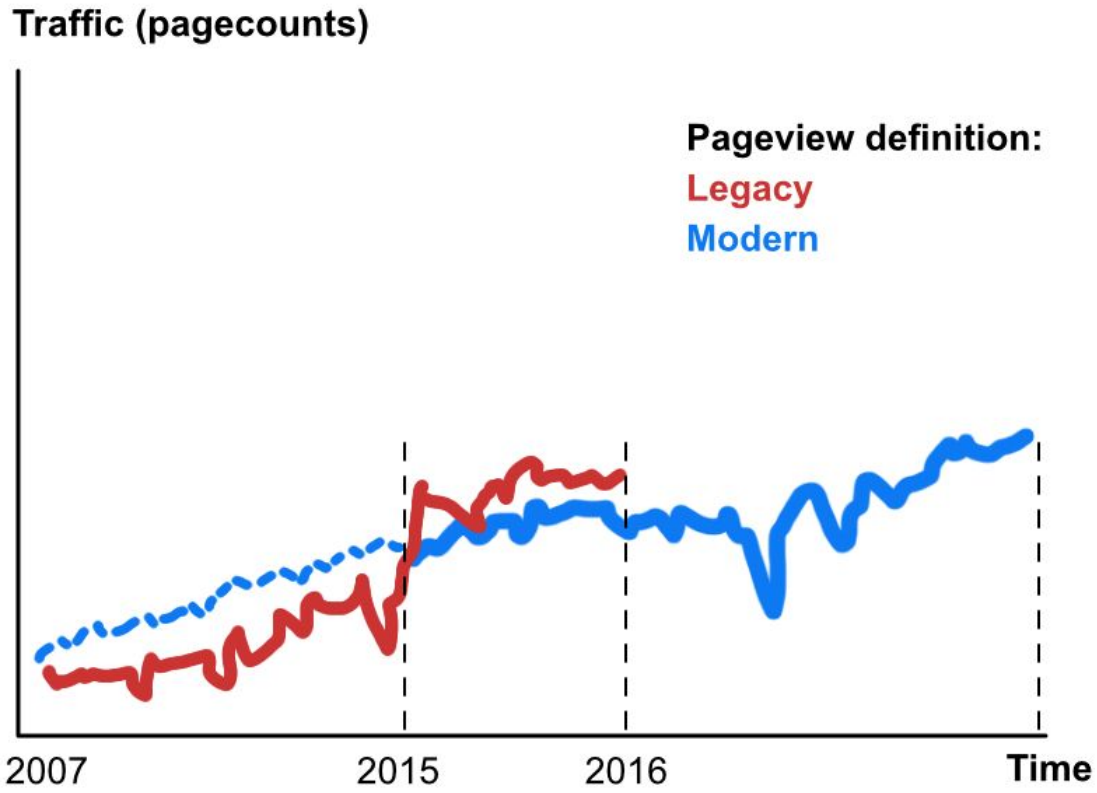
How do we answer questions about long term trends when we have two different measurements of traffic for the first ~10 years and other ~5 years?

Statistical modeling & machine learning, probably???

Project description

Using the traffic under both definitions & counting processes **from May 2015** (start of modern counts) **to August 2016** (end of legacy counts), be able to **estimate what 2007-2015 traffic would have looked like under the modern counting process.**

WIKIMEDIA
FOUNDATION



WIKIMEDIA
FOUNDATION

Size informs granularity

In May 2015 there were	
815	wikis (including Wikipedia, Wikivoyage, Wiktionary, etc.)
293	languages of Wikipedia
~108 million	articles across all wikis
~35 million	articles across Wikipedias
~4.8 million	articles on English Wikipedia

Source: stats.wikimedia.org/v2/#/all-projects/content/pages-to-date/normal|line|2015-03-26~2015-10-01|page_type~content|monthly

WIKIMEDIA
FOUNDATION

THANK YOU