How to tell the world about data you cannot show them

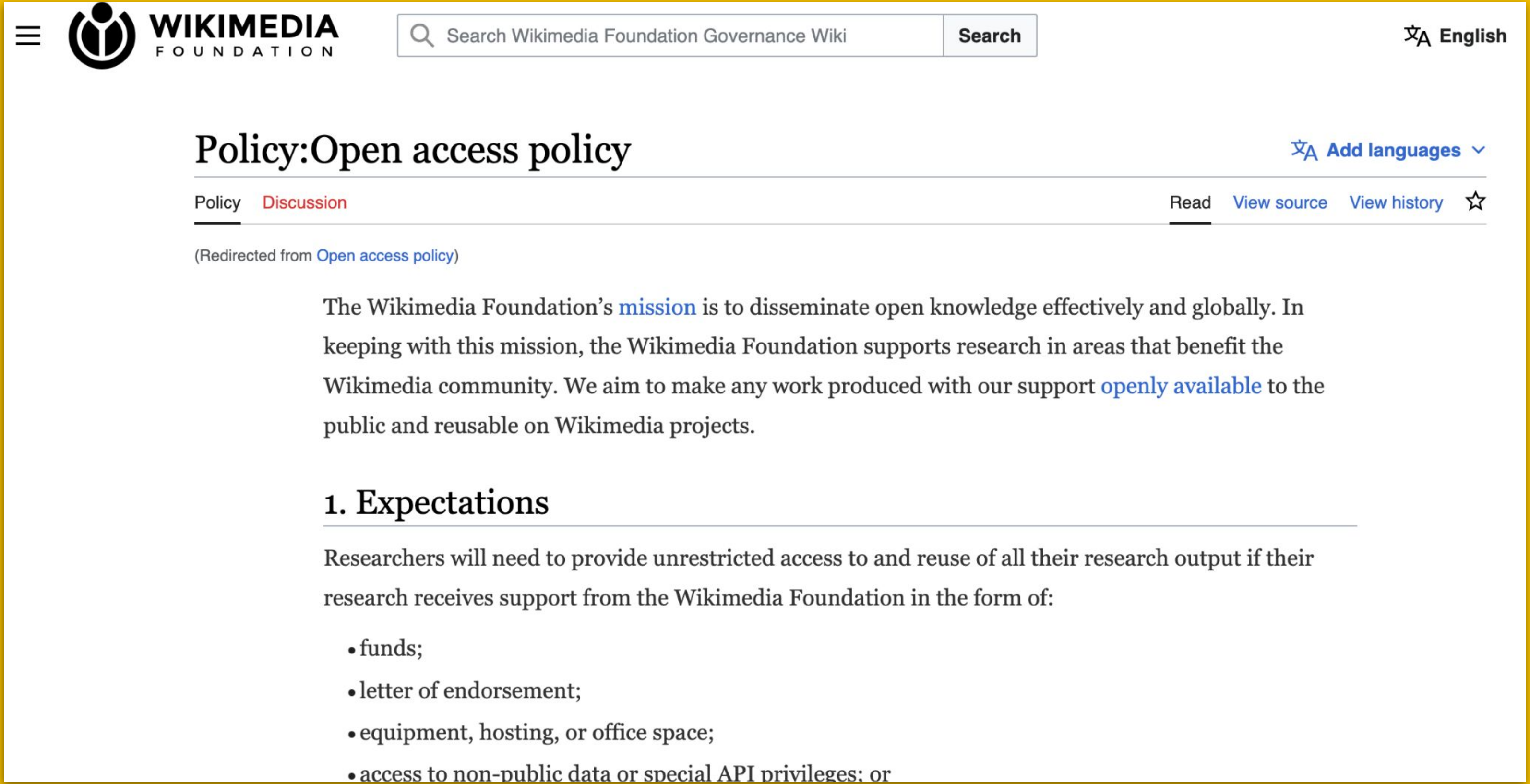
Differential privacy at the Wikimedia Foundation

Hal Triedman, Senior Privacy Engineer, WMF

18 October 2023

The Wikimedia Foundation (WMF)

WIKIPEDIA
WIKIMEDIA ENTERPRISE
WIKISOURCE
WIKIMEDIA COMMONS
WIKIMEDIA
WIKIDATA
WIKINEWS
MediaWiki
WIKIVERSITY



Policy:Open access policy

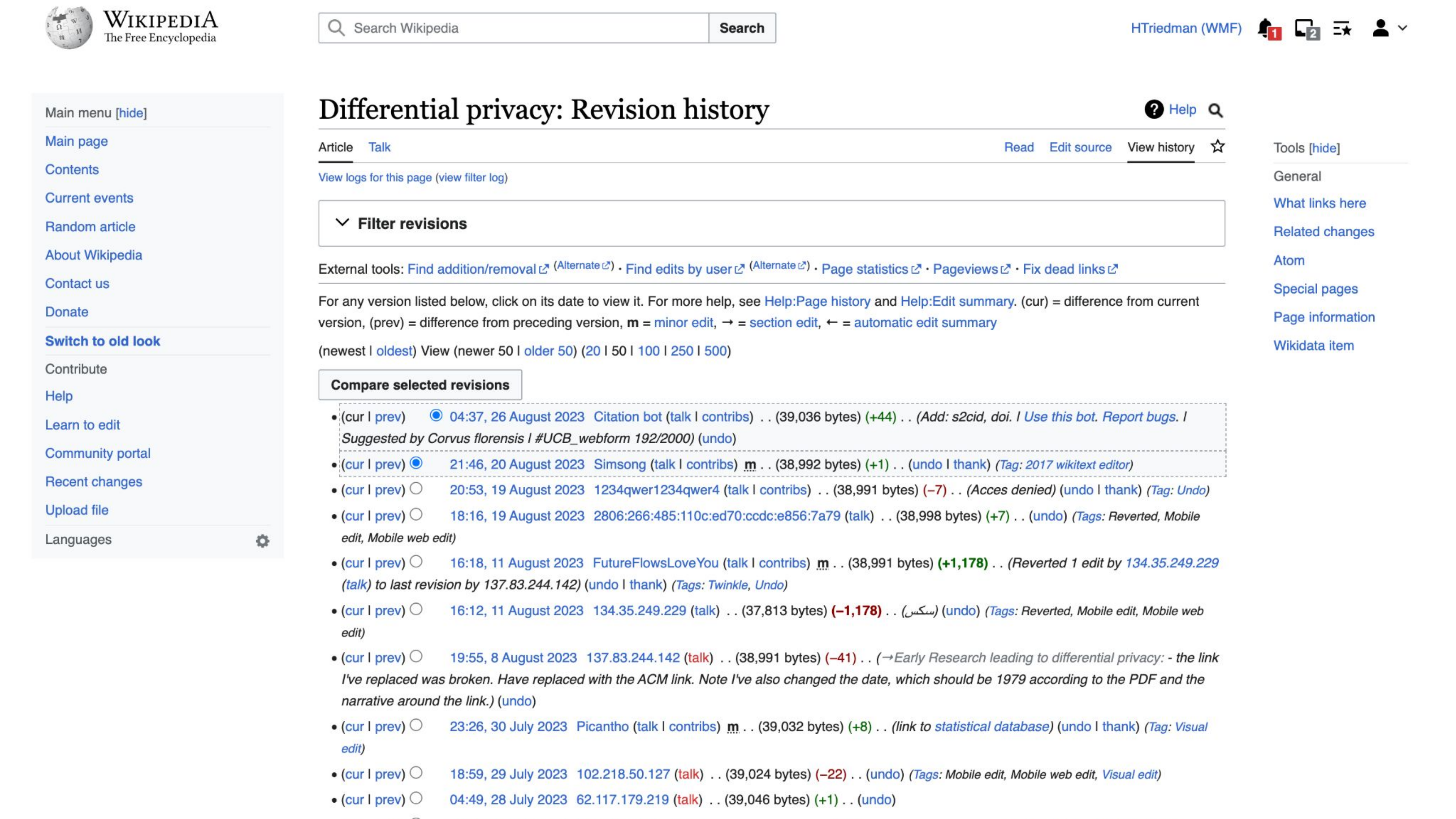
(Redirected from Open access policy)

The Wikimedia Foundation's mission is to disseminate open knowledge effectively and globally. In keeping with this mission, the Wikimedia Foundation supports research in areas that benefit the Wikimedia community. We aim to make any work produced with our support openly available to the public and reusable on Wikimedia projects.

1. Expectations

Researchers will need to provide unrestricted access to and reuse of all their research output if their research receives support from the Wikimedia Foundation in the form of:

- funds;
- letter of endorsement;
- equipment, hosting, or office space;
- access to non-public data or special API privileges; or



Differential privacy: Revision history

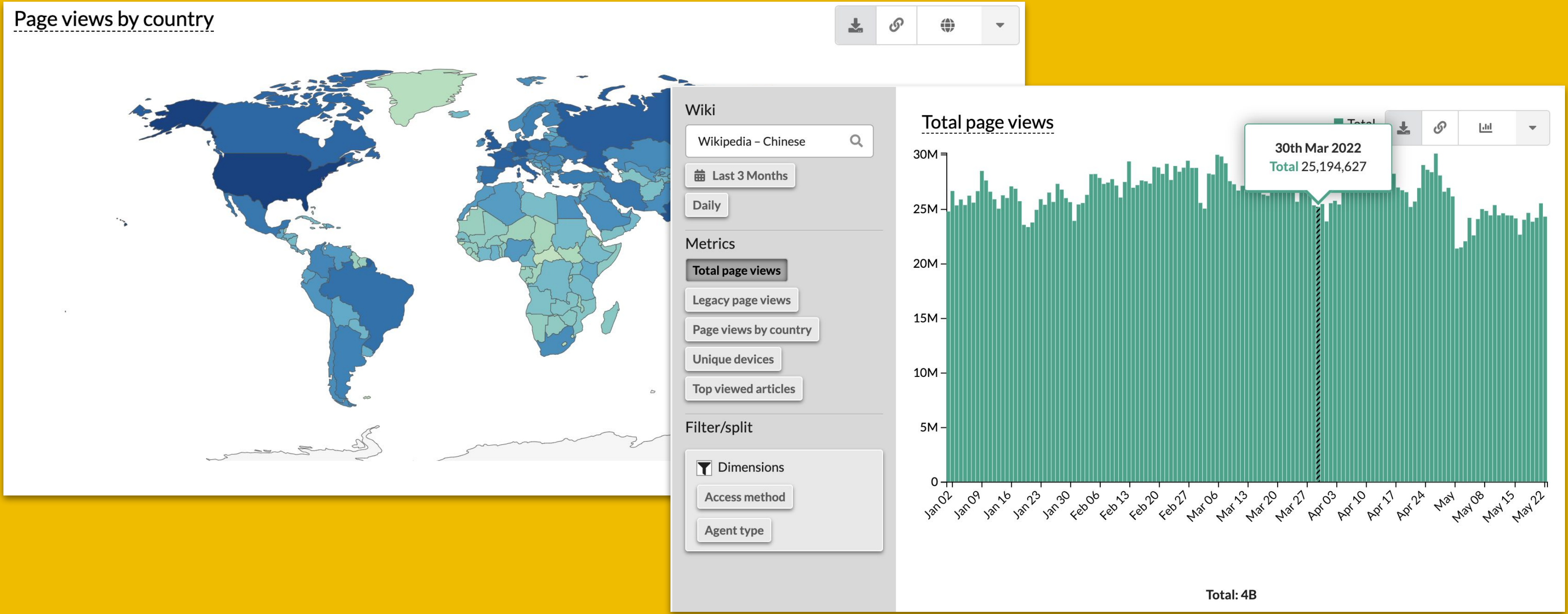
View logs for this page (view filter log)

External tools: Find addition/removal (Alternate) · Find edits by user (Alternate) · Page statistics · Pageviews · Fix dead links

For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary. (cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary

(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

- (cur | prev) ● 04:37, 26 August 2023 Citation bot (talk | contribs) . . (39,036 bytes) (+44) . . *(Add: s2cid, doi. | Use this bot. Report bugs. | Suggested by Corvus florensis | #UCB_webform 192/2000)* (undo)
- (cur | prev) ● 21:46, 20 August 2023 Simsong (talk | contribs) **m** . . (38,992 bytes) (+1) . . (undo | thank) *(Tag: 2017 wikitext editor)*
- (cur | prev) ○ 20:53, 19 August 2023 1234qwer1234qwer4 (talk | contribs) . . (38,991 bytes) (−7) . . *(Acces denied)* (undo | thank) *(Tag: Undo)*
- (cur | prev) ○ 18:16, 19 August 2023 2806:266:485:110c:ed70:ccdc:e856:7a79 (talk) . . (38,998 bytes) (+7) . . (undo) *(Tags: Reverted, Mobile edit, Mobile web edit)*
- (cur | prev) ○ 16:18, 11 August 2023 FutureFlowsLoveYou (talk | contribs) **m** . . (38,991 bytes) **(+1,178)** . . *(Reverted 1 edit by 134.35.249.229 (talk) to last revision by 137.83.244.142)* (undo | thank) *(Tags: Twinkle, Undo)*
- (cur | prev) ○ 16:12, 11 August 2023 134.35.249.229 (talk) . . (37,813 bytes) **(−1,178)** . . *(سکس)* (undo) *(Tags: Reverted, Mobile edit, Mobile web edit)*
- (cur | prev) ○ 19:55, 8 August 2023 137.83.244.142 (talk) . . (38,991 bytes) (−41) . . *(→Early Research leading to differential privacy: - the link I've replaced was broken. Have replaced with the ACM link. Note I've also changed the date, which should be 1979 according to the PDF and the narrative around the link.)* (undo)
- (cur | prev) ○ 23:26, 30 July 2023 Picantho (talk | contribs) **m** . . (39,032 bytes) (+8) . . *(link to statistical database)* (undo | thank) *(Tag: Visual edit)*
- (cur | prev) ○ 18:59, 29 July 2023 102.218.50.127 (talk) . . (39,024 bytes) (−22) . . (undo) *(Tags: Mobile edit, Mobile web edit, Visual edit)*
- (cur | prev) ○ 04:49, 28 July 2023 62.117.179.219 (talk) . . (39,046 bytes) (+1) . . (undo)



Page views by country

Wiki

Wikipedia – Chinese

Last 3 Months

Daily

Metrics

Total page views

Legacy page views

Page views by country

Unique devices

Top viewed articles

Filter/split

Dimensions

Access method

Agent type

Total page views

30th Mar 2022
Total 25,194,627

Total: 4B

WMF's Lean Data Diet

Defined by our Privacy Policy and Data Retention Guidelines:

No first-party tracking cookies

No account needed

90 days until aggregation + deletion

(images from Wikimedia Commons)

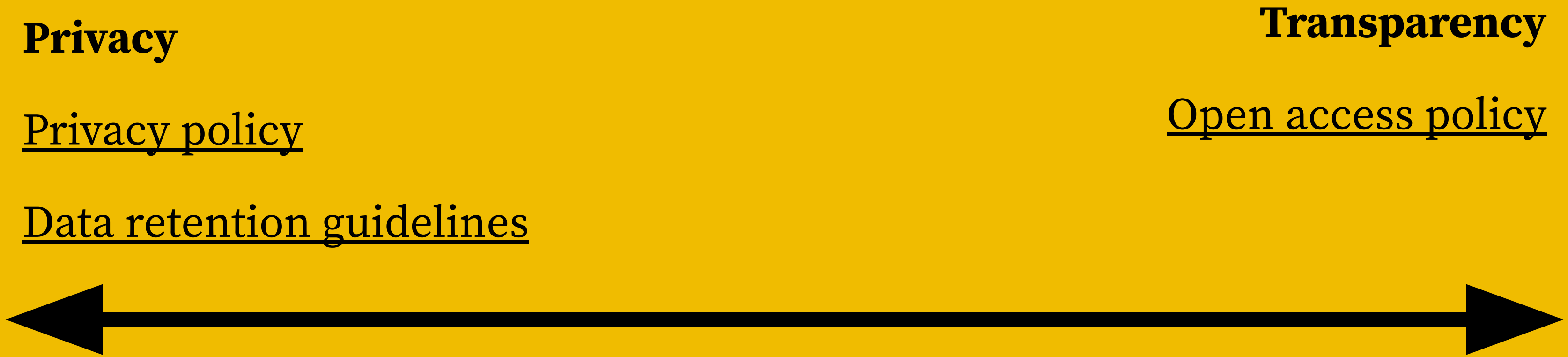
In 2020, community members request WMF release pageviews by country *and* project

(known as the “pageview data release”)

Pageview data release privacy concerns

- Both pageviews by country and pageviews by project are made up of user data
- Lean data diet constrains the kinds of actions WMF can take

This data release illuminates a tension between privacy and transparency



Privacy

Privacy policy

Data retention guidelines

Transparency

Open access policy

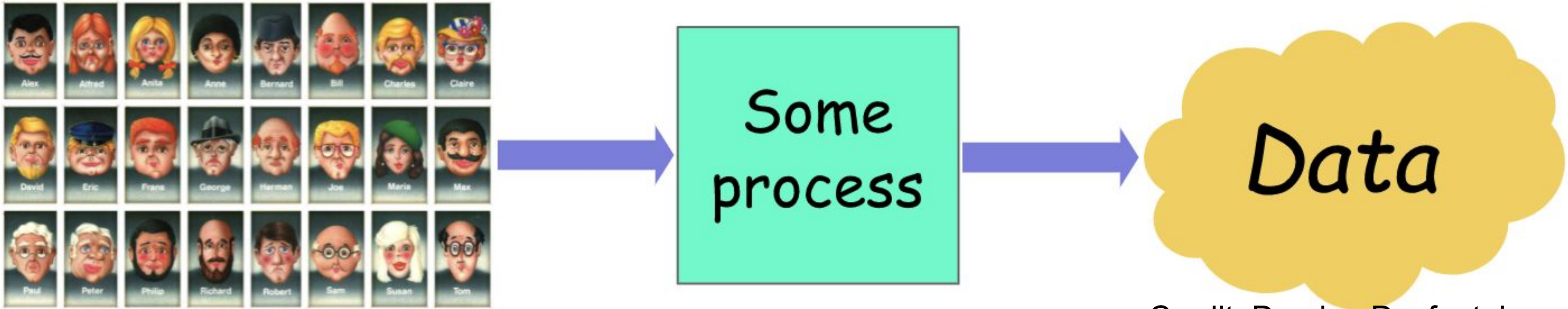
The stakes are high, because Wikipedia is inherently political — users and editors are pseudonymous for a very good reason

Tension → DP could be useful

Wait… so what is differential privacy?

What is DP?

A **process** takes a **database** in as input and returns some data as output



Credit: Damien Desfontaines

What is DP?

Add **random noise** (ignore for now how much, what type) to the process

For now we’ll call that magic

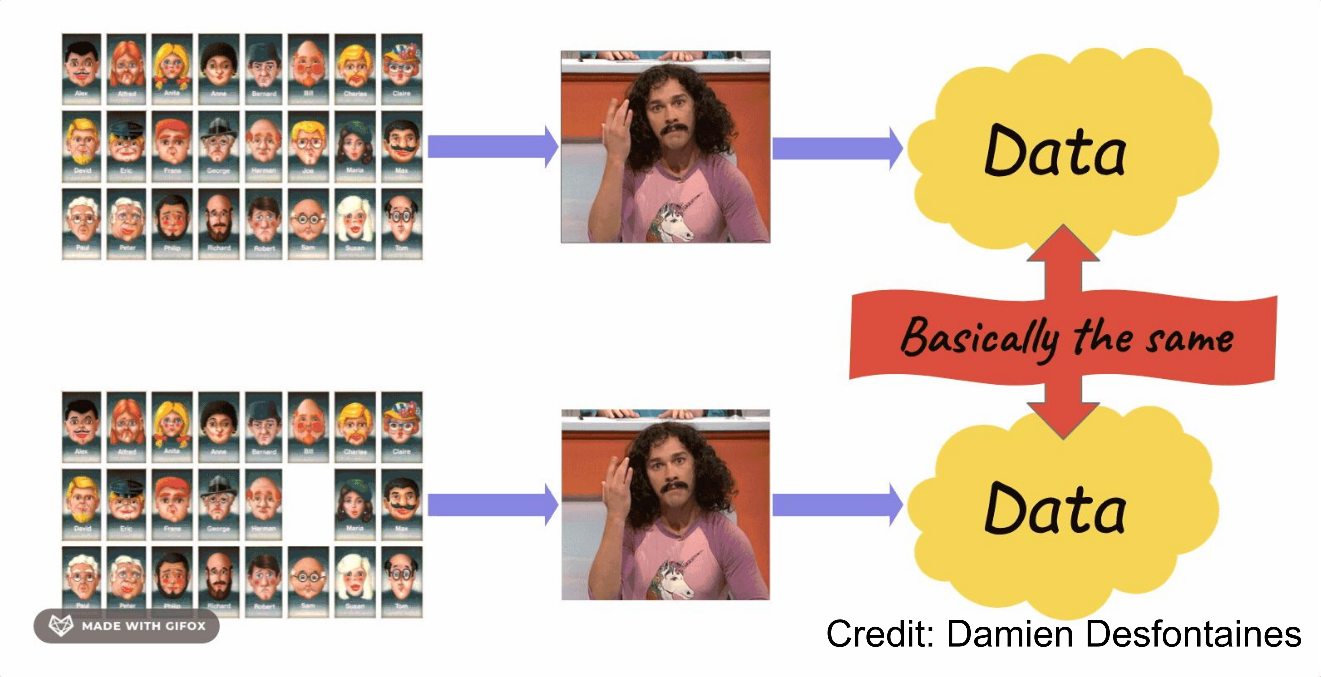
Alex
Alfred
Anita
Anne
Bernard
Bill
Charles
Claire
David
Eric
Frans
George
Herman
Joe
Maria
Max
Paul
Peter
Philip
Richard
Robert
Sam
Susan
Tom
MAGIC
Data
MADE WITH GIFOX

Credit: Damien Desfontaines

What is DP?

Remove one person from the database and re-run the process with magic

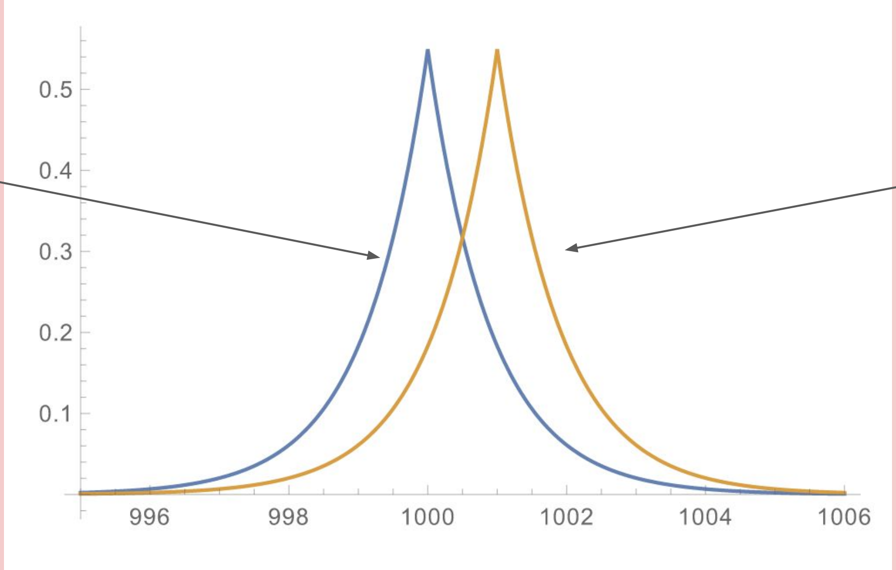
Outputs should be **basically the same**



Credit: Damien Desfontaines

What is DP?

Basically the same: Exact same outputs are possible with similar likelihood



Probability distribution **without** person in the database

Probability distribution **with** person in the database

Credit: Damien Desfontaines

What is DP?

Differential privacy is a **promise** WMF can make to the readers and editors who contribute to our public releases:

> From the perspective of someone looking at this data release, your contribution to this database will be hidden. High-level trends about the data will be visible, but no one will be able to infer your presence or absence in the data (even if you're an outlier).

Why is DP nice?

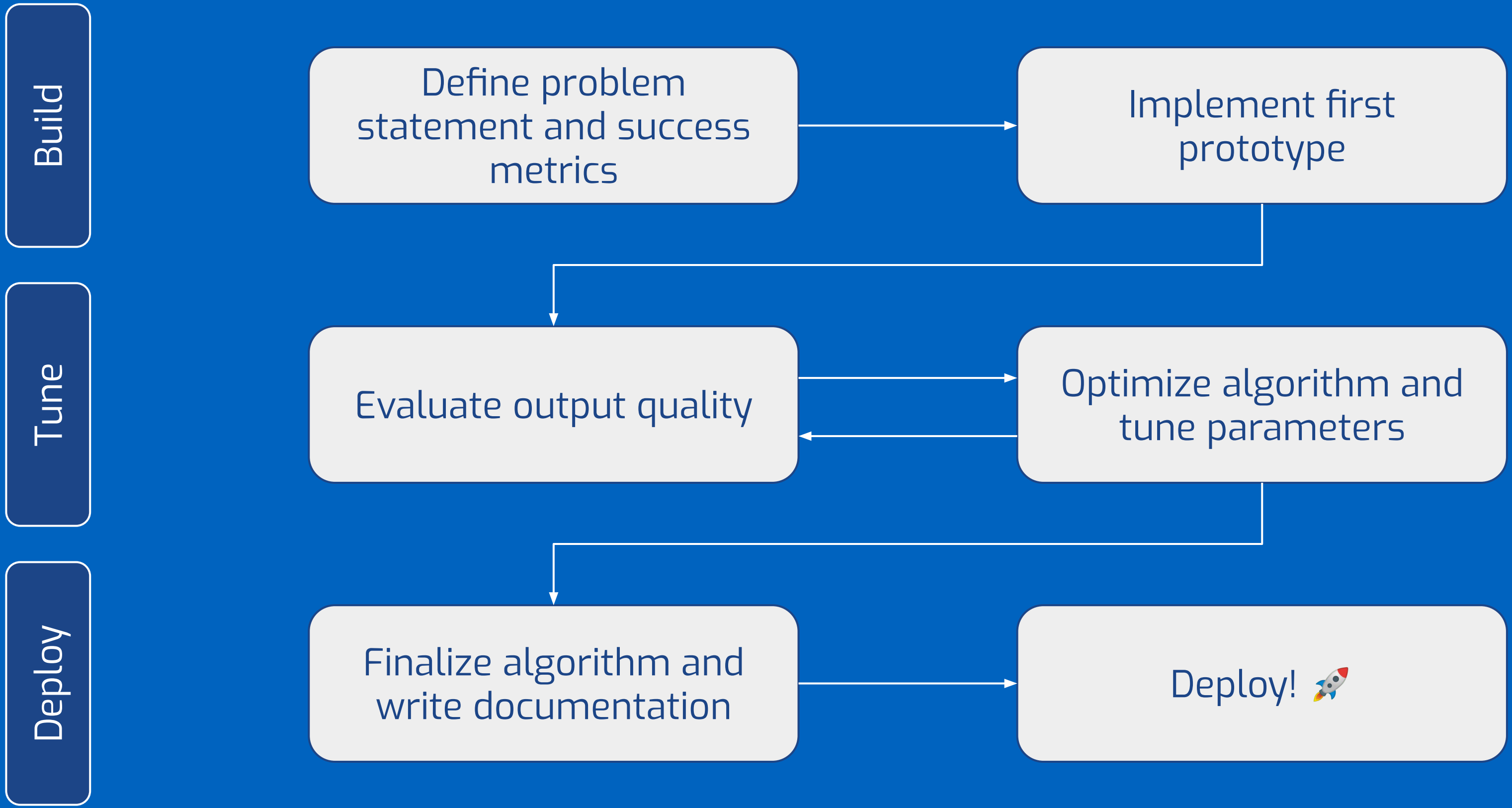
- Magic noise is configurable using a parameter called **epsilon (ϵ)**, which represents the **privacy budget**
  - Privacy budget is an worse-case bound on how much info can be gleaned from a data release
  - Smaller epsilon → more noise; larger epsilon → less noise
- Noise is **randomly generated**, so it’s impossible for DP data to be subject to re-identification attacks
- Any post-processing with DP data (modeling, sharing, combining with other data) is covered by these guarantees

Pageview data release

TUMULT

LABS

Tumult Labs' approach



Build
Tune
Deploy
Define problem statement and success metrics
Implement first prototype
Evaluate output quality
Optimize algorithm and tune parameters
Finalize algorithm and write documentation
Deploy! 🚀

Define problem and success metrics

What problem are we trying to solve?

- release **as much data as possible** about reading activity
- partition by **country, project, and page**
- release **every day**

What does success look like? (broadly)

- Privacy protected at a **user-day level**
- Data is **more plentiful and granular** than baseline
- Output is **equitable, accurate, and trustworthy for data consumers**

Implement prototype (conceptually)

country	project	page ID
US	es.wikipedia	1234
DE	de.wikipedia	5678
…	…	…
AR	wikidata	9012

group-by
and count

country	project	page ID	views
US	es.wikipedia	1234	109,283
DE	de.wikipedia	5678	4,756
…	…	…	…
AR	wikidata	9012	134

add noise
to views

country	project	page ID	noisy views
US	es.wikipedia	1234	110,170
DE	de.wikipedia	5678	4,704
…	…	…	…
AR	wikidata	9012	138

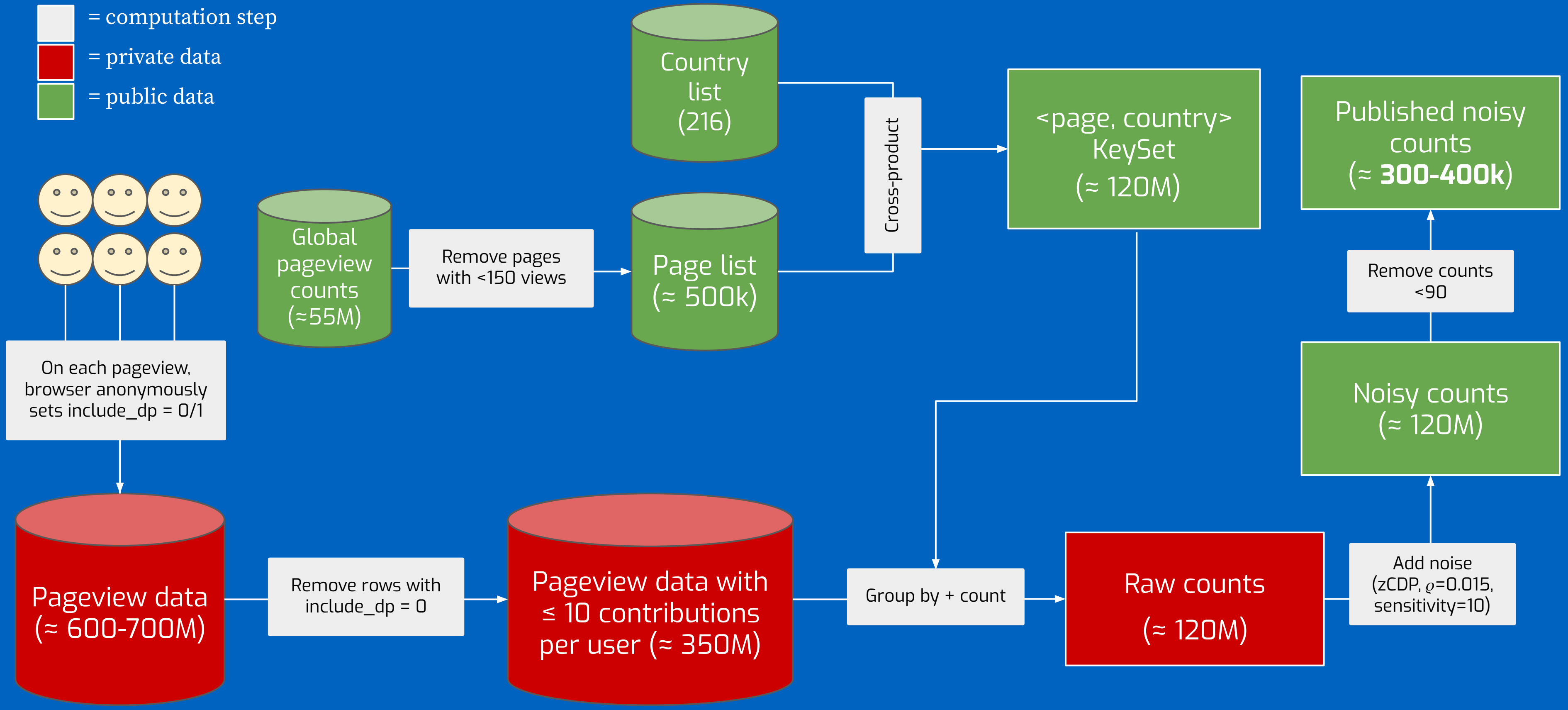
Implement prototype (in reality)

Legend

⬜ = computation step

🟥 = private data

🟩 = public data



On each pageview, browser anonymously sets include_dp = 0/1

Pageview data (≈ 600-700M)

Remove rows with include_dp = 0

Pageview data with ≤ 10 contributions per user (≈ 350M)

Global pageview counts (≈55M)

Remove pages with <150 views

Page list (≈ 500k)

Country list (216)

Cross-product

<page, country> KeySet (≈ 120M)

Group by + count

Raw counts (≈ 120M)

Add noise (zCDP, ρ=0.015, sensitivity=10)

Noisy counts (≈ 120M)

Remove counts <90

Published noisy counts (≈ **300-400k**)

Implement prototype (Historical data)

Similar approach for historical data (pre-DP cookie), with some tweaks:

- different kind of noise
- larger noise scale
- weaker privacy guarantee

Evaluate output quality

Success metric	Met?	Notes
Data is more plentiful and granular than baseline	✅	n/a

Evaluate output quality

Principle error metrics:

- **Median relative error <6%**
- **Drop rate <1%** (*similar to FNR: percentage of above-threshold true values not published*)
- **Spurious rate <1%** (*similar to FPR: percentage of published values with true count of 0*)
- **Equitable regional error rates**

Why are drop rate and spurious rate important? **Data is sparse and has a long tail**

Meeting goals for equity, accuracy, and trust requires optimizing for these metrics

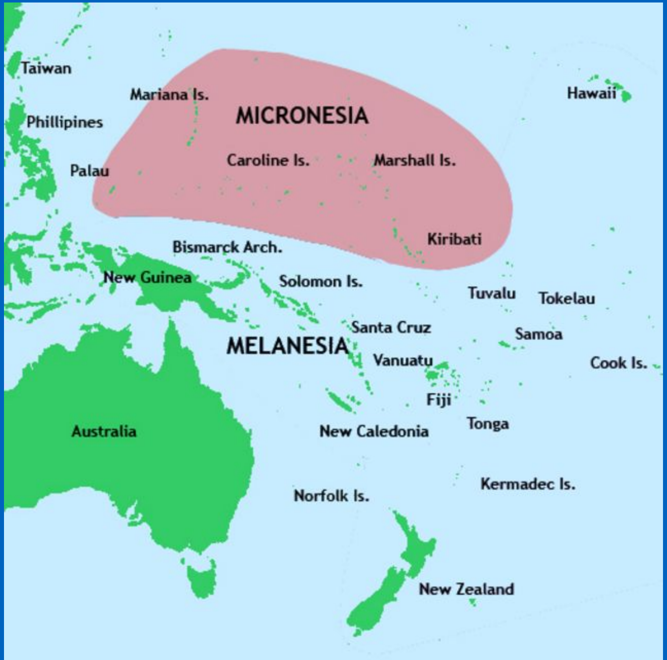
Evaluate output quality

Metric (global)	Goal value	Met?
Median relative error	<6%	✅
Drop rate	<1%	✅
Spurious rate	<1%	✅

What about if you look at sub-global metrics?

Optimize algorithm

The “Micronesia problem”



(image from Wikimedia Commons)

- Seven Pacific Island nations
- Very little traffic to WMF
- Naive first implementation
  - **>99% of published data is spurious**
  - **9 out of 23 subcontinental regions have spurious rate of >25%**
  - **Africa, Oceania, Central Asia, and the Caribbean**

Lesson: Global metrics can conceal local inequities

Solution: Change the kind of DP noise to solve this problem

Evaluate output quality

Success metric	Met?	Notes
Data is more plentiful and granular than baseline	✅	n/a
Output is equitable, accurate, and trustworthy for data consumers	✅	spurious rate ≤1% both globally and for 21/23 subcontinental regions

Optimize algorithm

Bounding user contributions

Recall: no first-party tracking cookies. So how to bound user contributions?

- Can look at hash of IP + UA, but that often fails
- Our solution? Client-side filtering:
  - Client-side cookie sends server a boolean to include only first k unique pageviews in a day

Lesson: Data minimization and strong privacy guarantees can be in conflict with each other

Solution: Build new privacy-preserving infrastructure

Evaluate output quality

Success metric	Met?	Notes
Data is more plentiful and granular than baseline	✅	n/a
Output is equitable, accurate, and trustworthy for data consumers	✅	spurious rate ≤1% both globally and for 21/23 subcontinental regions
Privacy protection at a user-day level	✅	client-side filtering has fewer failure modes than hash of IP + UA

Evaluate output quality

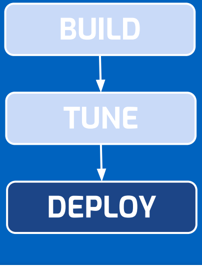
Our latest attempt meets our equity, accuracy, and trustworthiness goals…

Metric	Goal	Actual
Spurious rate	<1%	<0.01%
Drop rate	<1%	<0.1%
Median relative error	<6%	<6%
Geographic equity	✅	✅

…while also significantly improving on a baseline non-DP data release.

Metric	Before DP	After DP	Percent change
Median # data points released / day	9,000	360,000	**+4,000%**
Median # pageviews released / day	50M	120M	**+240%**

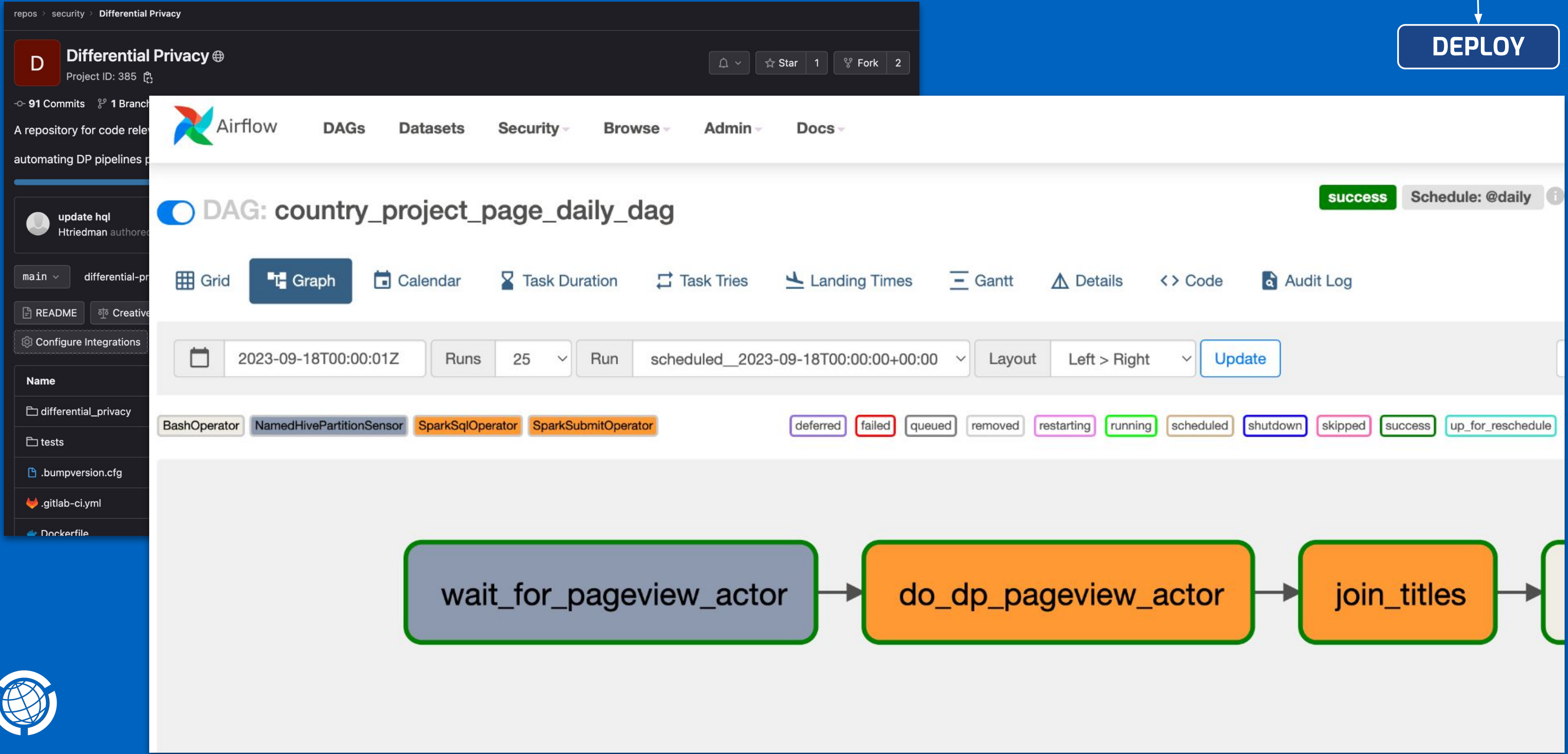
Finalize algorithm...



BUILD

TUNE

DEPLOY



repos › security › Differential Privacy

D Differential Privacy

Project ID: 385

Star 1 Fork 2

91 Commits 1 Branch

A repository for code rele

automating DP pipelines p

update hql

Htriedman authored

main differential-pr

README Creative

Configure Integrations

Name
differential_privacy
tests
.bumpversion.cfg
.gitlab-ci.yml
Dockerfile

Airflow DAGs Datasets Security Browse Admin Docs

DAG: country_project_page_daily_dag

success Schedule: @daily

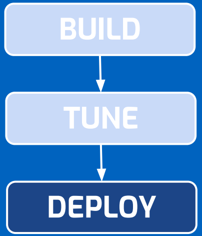
Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

2023-09-18T00:00:01Z Runs 25 Run scheduled__2023-09-18T00:00:00+00:00 Layout Left > Right Update

BashOperator NamedHivePartitionSensor SparkSqlOperator SparkSubmitOperator

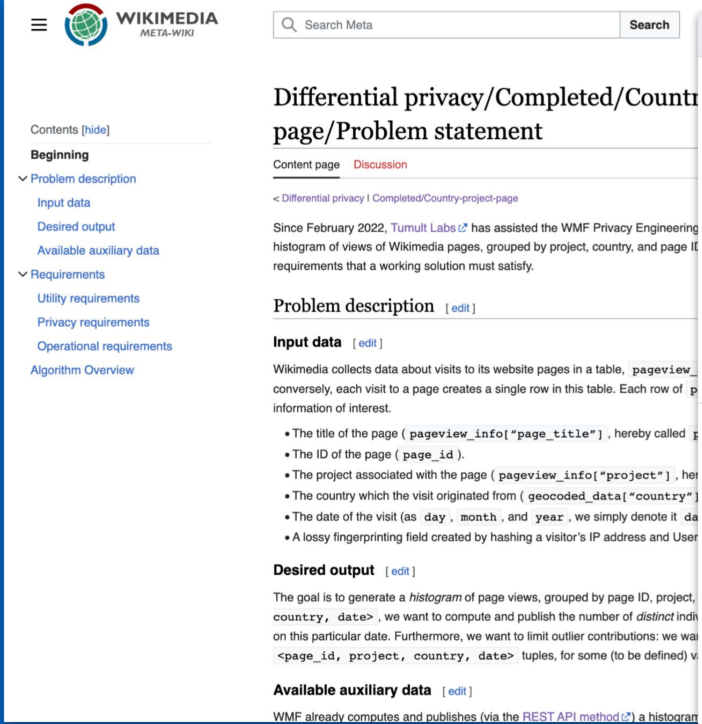
deferred failed queued removed restarting running scheduled shutdown skipped success up_for_reschedule

…and write documentation


BUILD

TUNE

DEPLOY



Differential privacy/Completed/Countr page/Problem statement

< Differential privacy | Completed/Country-project-page

Since February 2022, Tumult Labs has assisted the WMF Privacy Engineering histogram of views of Wikimedia pages, grouped by project, country, and page ID requirements that a working solution must satisfy.

Problem description [edit]

Input data [edit]

Wikimedia collects data about visits to its website pages in a table, `pageview_` conversely, each visit to a page creates a single row in this table. Each row of `p` information of interest.

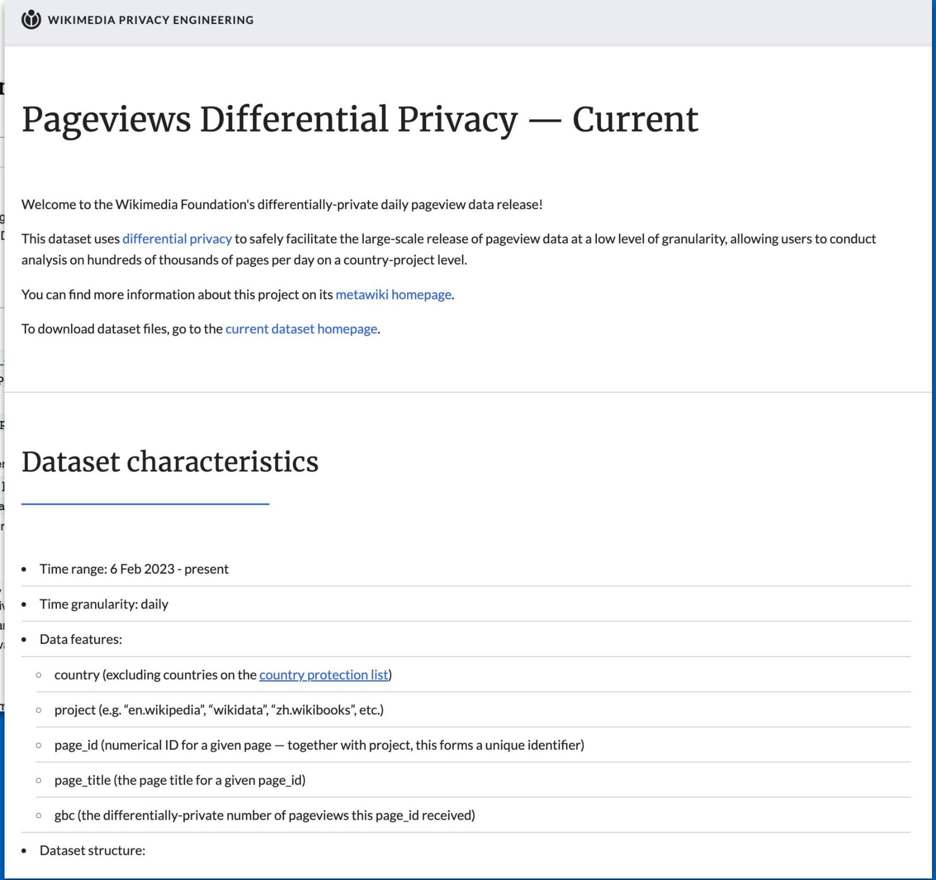
- The title of the page (`pageview_info["page_title"]` , hereby called `p`
- The ID of the page (`page_id`).
- The project associated with the page (`pageview_info["project"]` , her
- The country which the visit originated from (`geocoded_data["country"]`
- The date of the visit (as `day` , `month` , and `year` , we simply denote it `da`
- A lossy fingerprinting field created by hashing a visitor's IP address and User

Desired output [edit]

The goal is to generate a *histogram* of page views, grouped by page ID, project, `country, date>` , we want to compute and publish the number of *distinct* indiv on this particular date. Furthermore, we want to limit outlier contributions: we wa `<page_id, project, country, date>` tuples, for some (to be defined) v

Available auxiliary data [edit]

WMF already computes and publishes (via the REST API method) a histogram



Pageviews Differential Privacy — Current

Welcome to the Wikimedia Foundation's differentially-private daily pageview data release!

This dataset uses differential privacy to safely facilitate the large-scale release of pageview data at a low level of granularity, allowing users to conduct analysis on hundreds of thousands of pages per day on a country-project level.

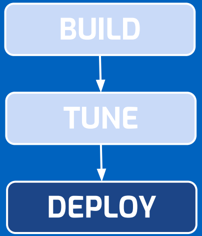
You can find more information about this project on its metawiki homepage.

To download dataset files, go to the current dataset homepage.

Dataset characteristics

- Time range: 6 Feb 2023 - present
- Time granularity: daily
- Data features:
  - country (excluding countries on the country protection list)
  - project (e.g. "en.wikipedia", "wikidata", "zh.wikibooks", etc.)
  - page_id (numerical ID for a given page — together with project, this forms a unique identifier)
  - page_title (the page title for a given page_id)
  - gbc (the differentially-private number of pageviews this page_id received)
- Dataset structure:

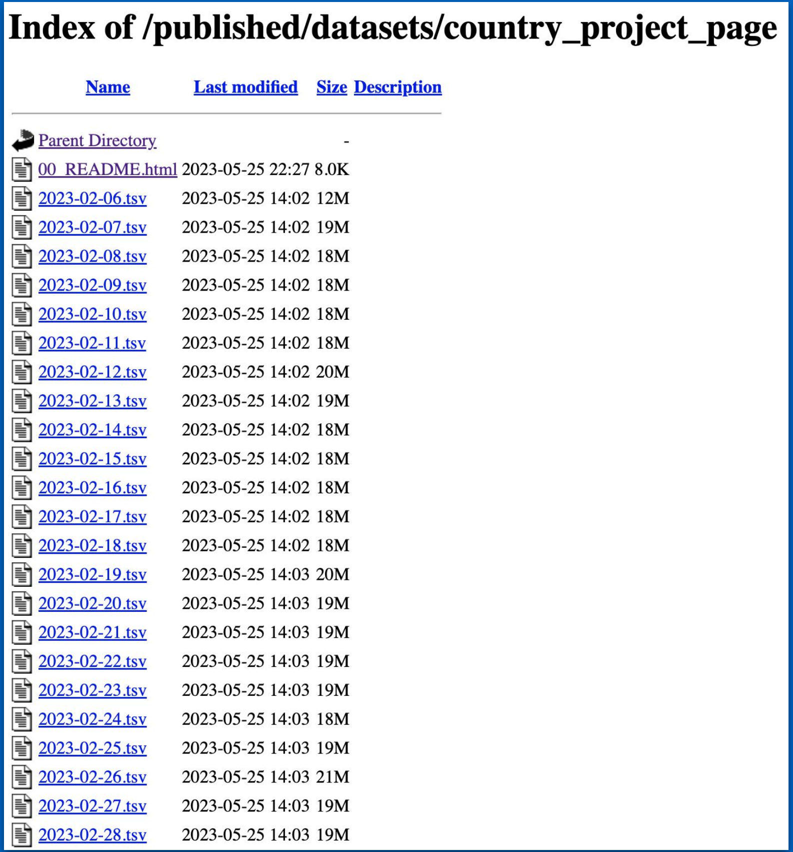
Deploy! 🚀



BUILD

TUNE

DEPLOY

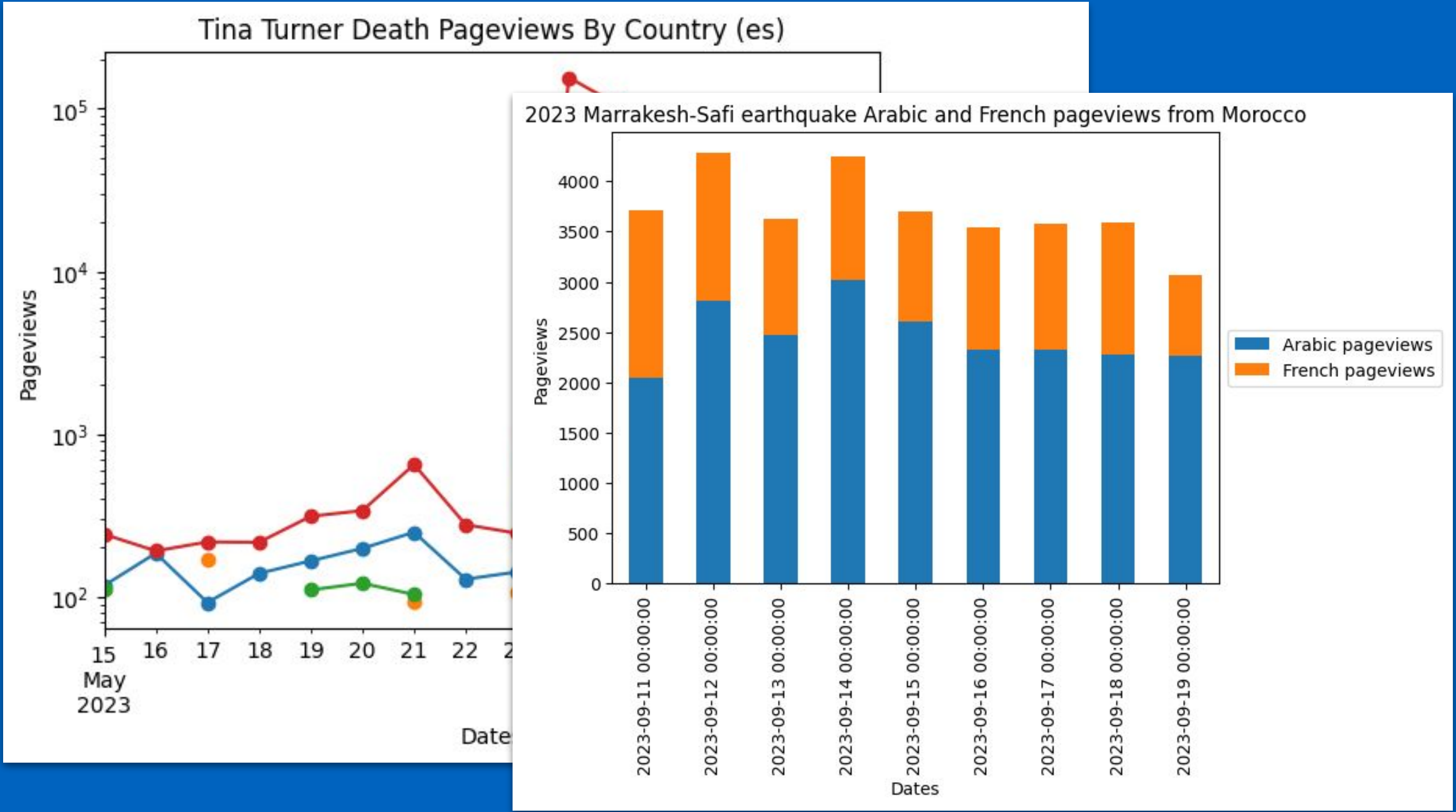


Index of /published/datasets/country_project_page

Name	Last modified	Size	Description
Parent Directory		-	
00_README.html	2023-05-25 22:27	8.0K	
2023-02-06.tsv	2023-05-25 14:02	12M	
2023-02-07.tsv	2023-05-25 14:02	19M	
2023-02-08.tsv	2023-05-25 14:02	18M	
2023-02-09.tsv	2023-05-25 14:02	18M	
2023-02-10.tsv	2023-05-25 14:02	18M	
2023-02-11.tsv	2023-05-25 14:02	18M	
2023-02-12.tsv	2023-05-25 14:02	20M	
2023-02-13.tsv	2023-05-25 14:02	19M	
2023-02-14.tsv	2023-05-25 14:02	18M	
2023-02-15.tsv	2023-05-25 14:02	18M	
2023-02-16.tsv	2023-05-25 14:02	18M	
2023-02-17.tsv	2023-05-25 14:02	18M	
2023-02-18.tsv	2023-05-25 14:02	18M	
2023-02-19.tsv	2023-05-25 14:03	20M	
2023-02-20.tsv	2023-05-25 14:03	19M	
2023-02-21.tsv	2023-05-25 14:03	19M	
2023-02-22.tsv	2023-05-25 14:03	19M	
2023-02-23.tsv	2023-05-25 14:03	19M	
2023-02-24.tsv	2023-05-25 14:03	18M	
2023-02-25.tsv	2023-05-25 14:03	19M	
2023-02-26.tsv	2023-05-25 14:03	21M	
2023-02-27.tsv	2023-05-25 14:03	19M	
2023-02-28.tsv	2023-05-25 14:03	19M	

Download data: https://w.wiki/754L

Outcomes


Tina Turner Death Pageviews By Country (es)
Pageviews
10^5
10^4
10^3
10^2
15
May
2023
16
17
18
19
20
21
22
2
Date
2023 Marrakesh-Safi earthquake Arabic and French pageviews from Morocco
Pageviews
4000
3500
3000
2500
2000
1500
1000
500
0
2023-09-11 00:00:00
2023-09-12 00:00:00
2023-09-13 00:00:00
2023-09-14 00:00:00
2023-09-15 00:00:00
2023-09-16 00:00:00
2023-09-17 00:00:00
2023-09-18 00:00:00
2023-09-19 00:00:00
Dates
Arabic pageviews
French pageviews

Outcomes

In total:

- 8 years of safer, more granular data, ~300M rows of data, ~350B source data points
- Publicly accessible and openly licensed
- Safe for post-processing (currently trying to use it to do country-level trend modeling)

Future work

Dataset	Status
Geolocated editor activity	✅
WMF grant data	✅
Banner views / clicks	🔁
Search data	🔁
Chains of pageviews	🔜
Geolocated edit activity	🔜
Global pageviews (hourly)	🔜

For more information…

- **For a beginner-friendly introduction:** Damien Desfontaines' privacy blog
  - (I worked closely with Damien and his company, Tumult Labs, on this project)
- **For a theoretically-sound foundation:** Dwork and Roth, *Algorithmic Foundations of Differential Privacy* (2014)
- **For keeping up with my work:** Wikimedia's differential privacy homepage

Thank you!

And my deepest gratitude to collaborators at WMF (Isaac Johnson, Gabriele Modena, Temi Adeleye, Nuria Ruiz, Cléo Lemoisson) **and Tumult Labs** (Damien Desfontaines, Daniel Simmons-Marengo, Skye Berghel, David Pujol, Tom Magerlein, Ashwin Machanavajjhala, Michael Hay)

Q+A