Perspectives from a data center with multiple repositories: NCBI

Valerie Schneider, Ph.D.

U.S. National Library of Medicine
National Center for Biotechnology Information

Acknowledgements

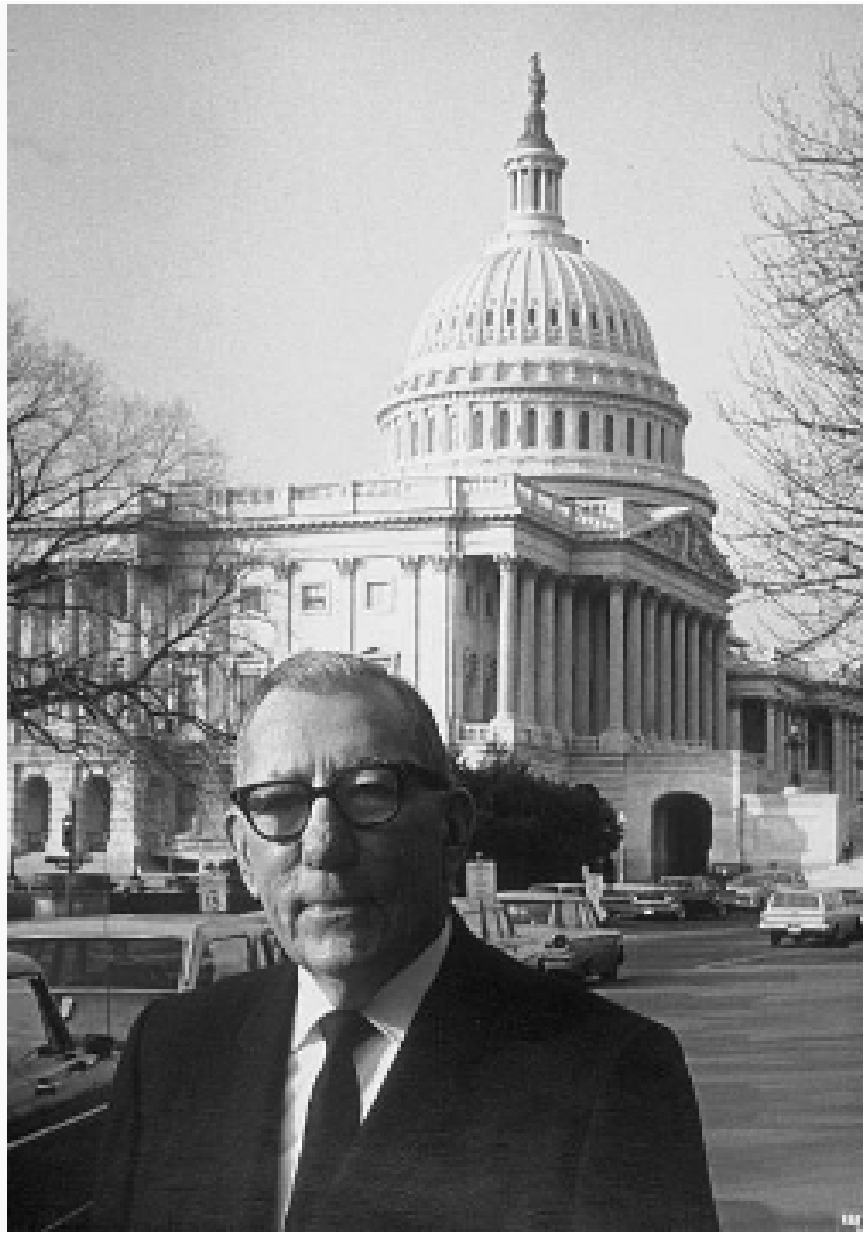
- Hundreds of talented and dedicated scientists, programmers, data wranglers, project managers, product managers who:
  - Process data submissions
  - Validate, QC, or curate submitted data
  - Generate new data through analysis, curation, and collaboration
  - Develop and maintain databases, process flows, analysis methods
  - Design web sites, tools, and displays
  - Store all of the data and provide safe, reliable access to it
- ~30 years of support from NIH, NLM, and NCBI Senior Management

Overview

- Who we are
- Metrics 101
- Metrics-Based Resource Management
  - PubMed
  - GEO
  - dbGaP

November 4, 1988

To develop new information technologies to aid in the understanding of fundamental molecular and genetic processes that control health and disease.



In 1987, Senator Pepper introduced H.R. 393, the National Biotechnology Information Act, to establish the NCBI.

Re-introduced in 1988 as part of the NIH authorization process, it was signed into law by President Ronald Reagan on November 4, 1988 as part of the Health Omnibus Extension Act P.L.100-607.

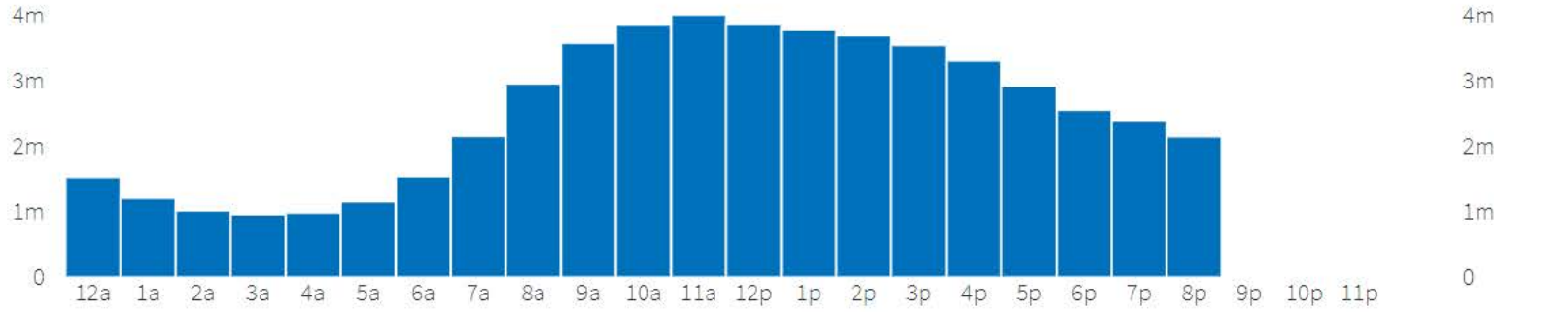
analytics.usa.gov | All Participating Websites | About this site | Data | API

196,366

people on government websites now

Visits Today

Eastern Time



Visits in the Past 90 Days

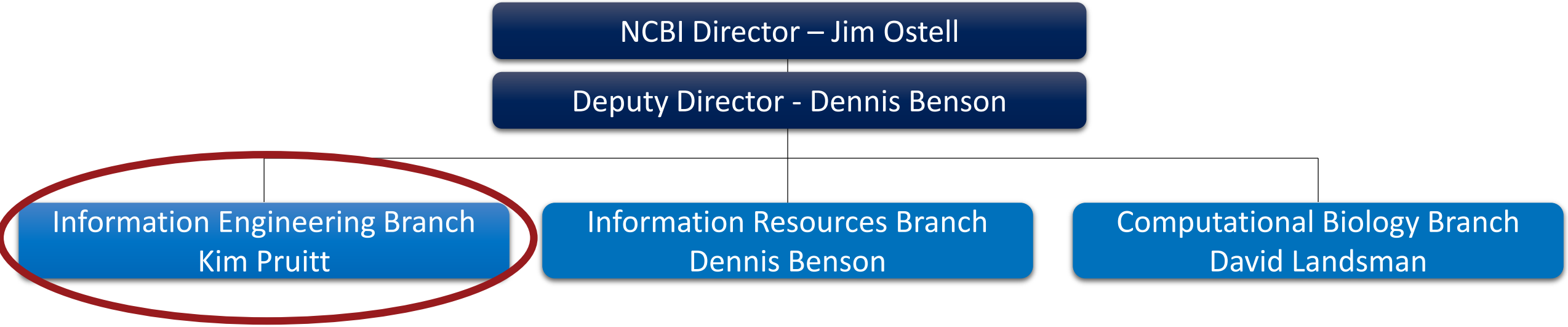
There were **3.72 billion** visits over the past 90 days.

Top Domains

Now | 7 Days | 30 Days

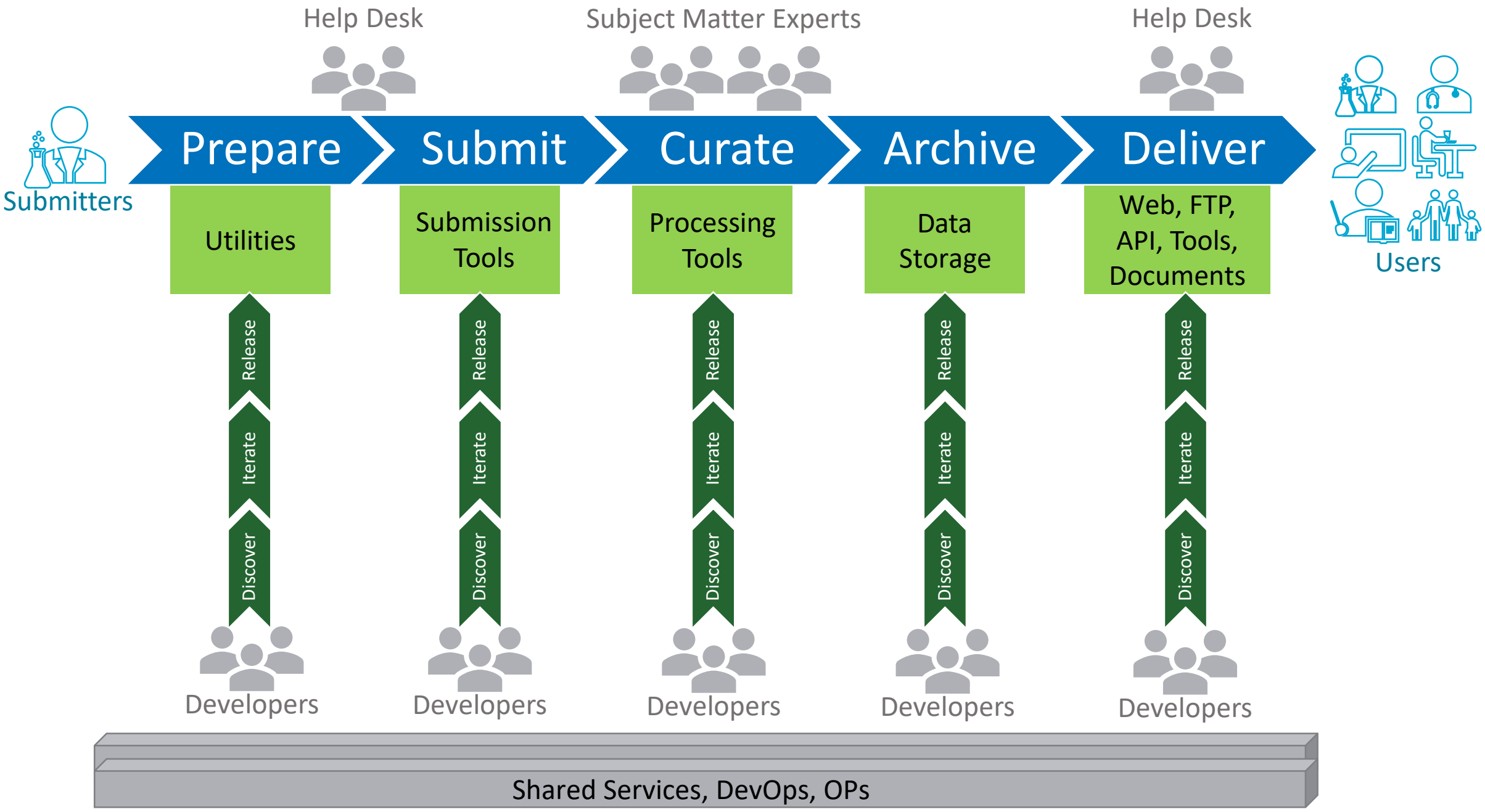
*Visits over the last week to **domains**, including traffic to all pages within that domain.*

ncbi.nlm.nih.gov	41,779,148
tools.usps.com	38,014,955
irs.gov	35,246,080
sa.www4.irs.gov	30,424,188
cdc.gov	17,413,203
forecast.weather.gov	13,933,533
medlineplus.gov	12,861,685
reg.usps.com	8,302,373
usps.com	7,545,972
weather.gov	6,356,164
ssa.gov	5,016,329
informeddelivery.usps.com	4,894,997
usajobs.gov	4,650,048

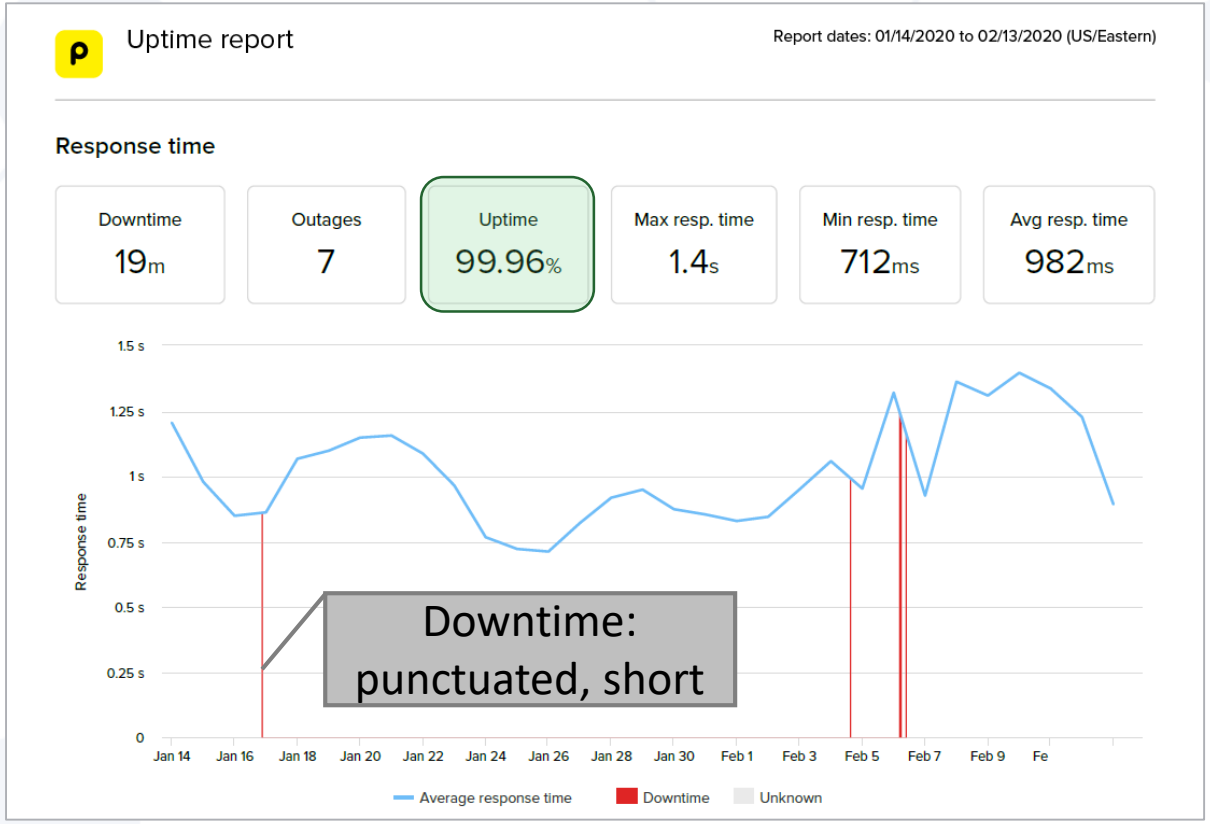


IEB Mission: Accrue, **enhance**, and **deliver** biomedical data and information to support learning, discovery, research, and medical advances.

IEB Offering	Research	Policy	Public Health	Patient Care
offerings comprise multiple products that serve similar or related customer constituencies				
ClinicalTrials.gov	✓	✓	✓	✓
Clinical Variation (ClinVar)	✓			✓
Genotype and Phenotype (dbGaP)	✓	✓		
Human Genetic Variation (dbSNP)	✓			
GenBank	✓	✓	✓	✓
Gene Expression Omnibus (GEO)	✓	✓		
Genetic Testing Registry (GTR)		✓		✓
Pathogens	✓		✓	
Protein domains	✓			
PubChem	✓		✓	
PubMed	✓		✓	✓
PubMed Central	✓	✓	✓	✓
Reference Sequence Standards (RefSeq)	✓		✓	
Sequence Analysis Tools (BLAST, more)	✓		✓	✓
Sequence Read Archive	✓	✓	✓	


Help Desk
Subject Matter Experts
Help Desk
Submitters
Prepare
Submit
Curate
Archive
Deliver
Users
Utilities
Submission Tools
Processing Tools
Data Storage
Web, FTP, API, Tools, Documents
Release
Iterate
Discover
Developers
Shared Services, DevOps, OPs

Performance and Errors Monitoring



Usage and Engagement

Basic User Behavior

- Page views, downloads, events, time on page

Visit Characteristics

- Referrer, entry/exit pages, page path, pages/clicks/events per visit

User Characteristics

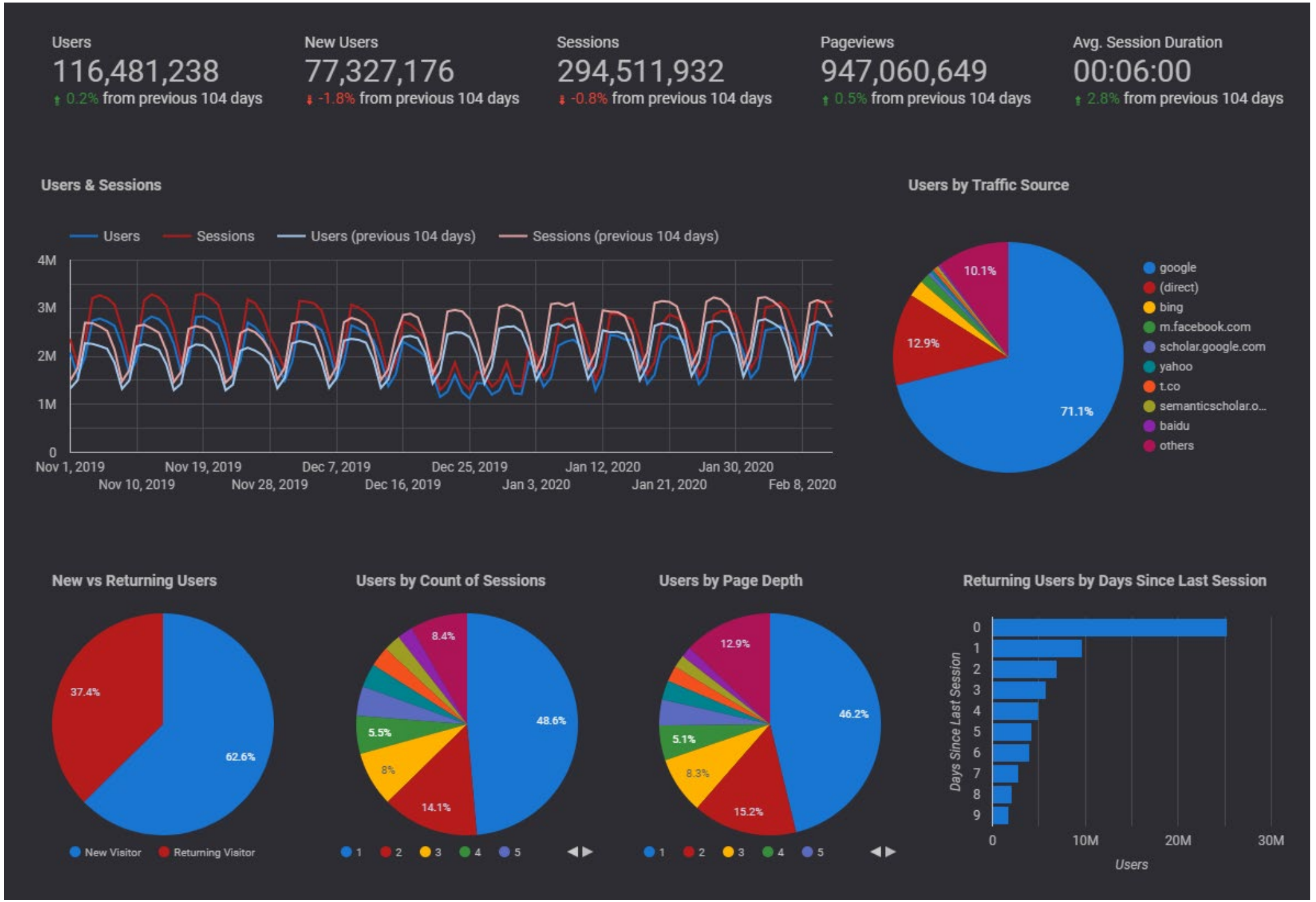
- Count, geography, OS, browser, new vs. returning, visit frequency

Customer Feedback

- Net Promoter Score (NPS), surveys, interviews, help desk reports

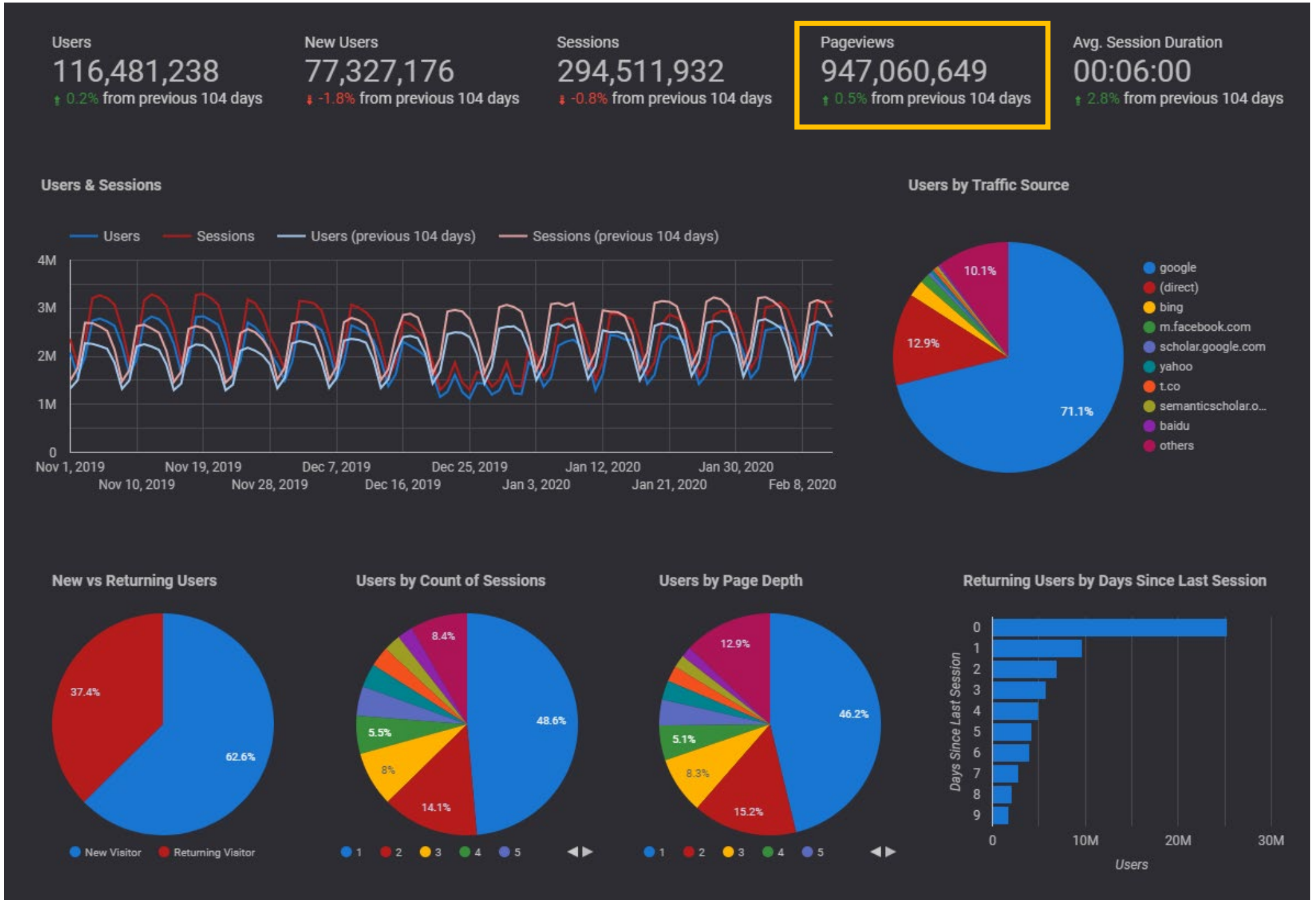
Dashboards

- Use detailed analytics to understand user behavior
- Facilitate monitoring
- Reveal patterns



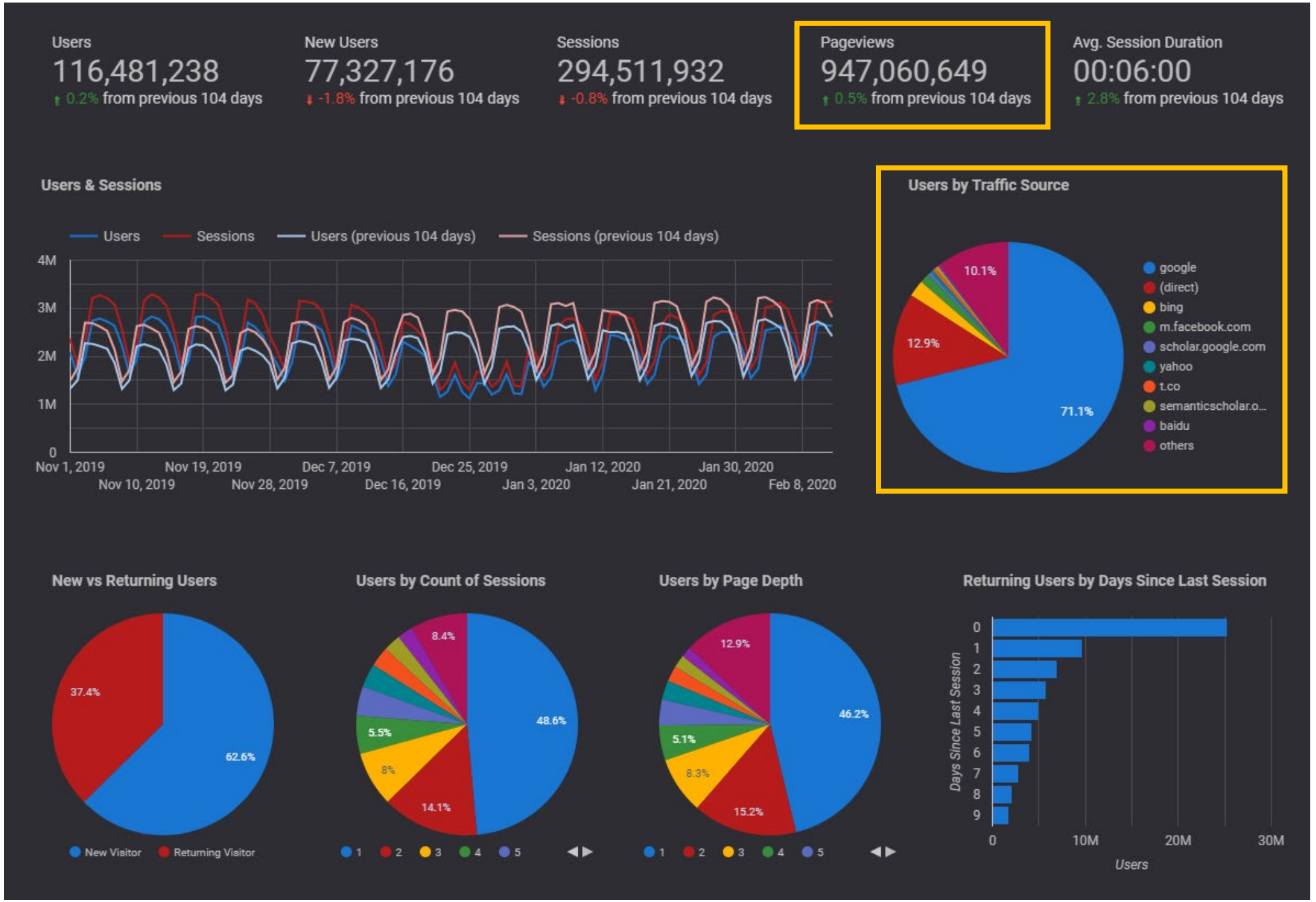
Dashboards

- Use detailed analytics to understand user behavior
- Facilitate monitoring
- Reveal patterns



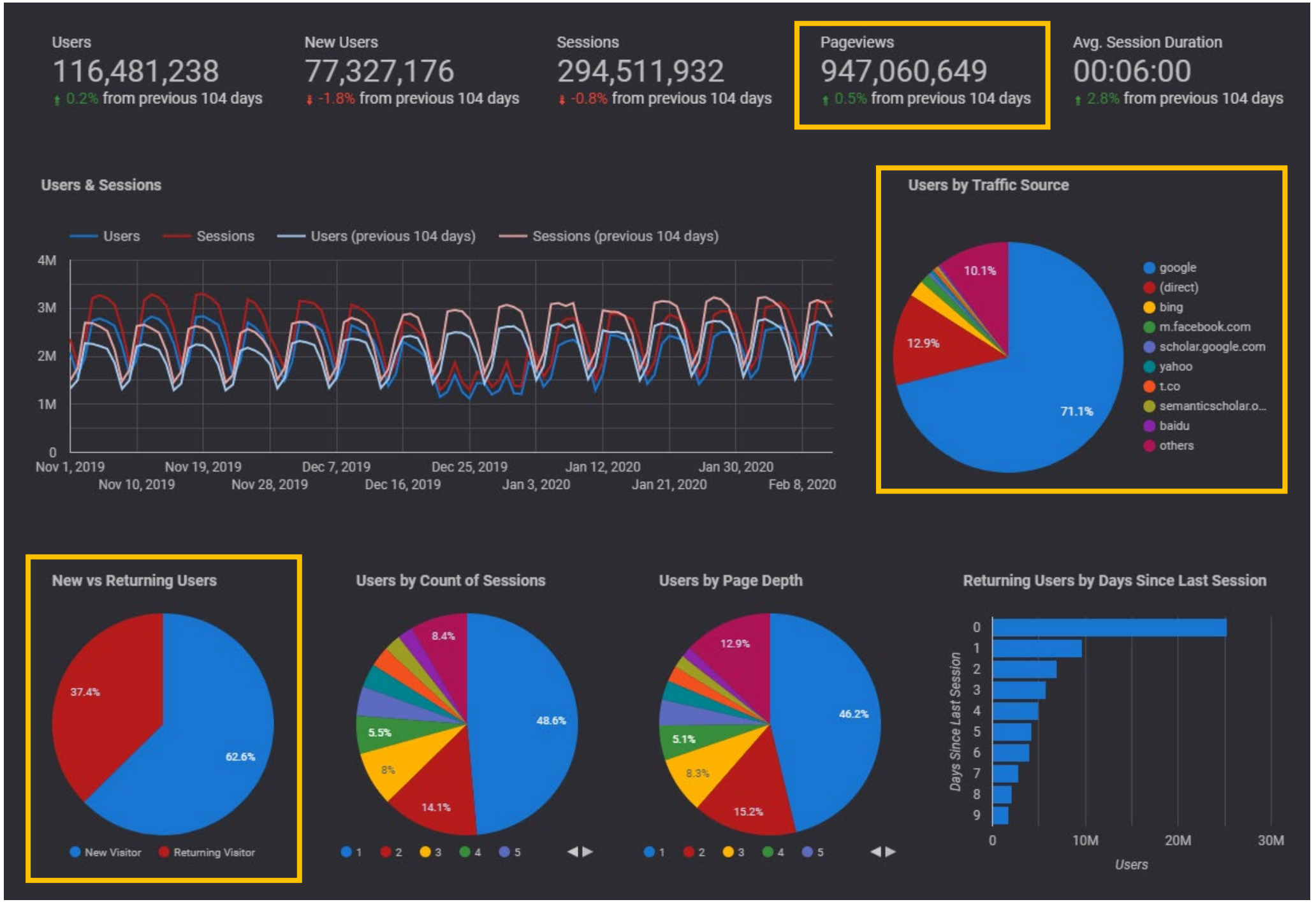
Dashboards

- Use detailed analytics to understand user behavior
- Facilitate monitoring
- Reveal patterns



Dashboards

- Use detailed analytics to understand user behavior
- Facilitate monitoring
- Reveal patterns



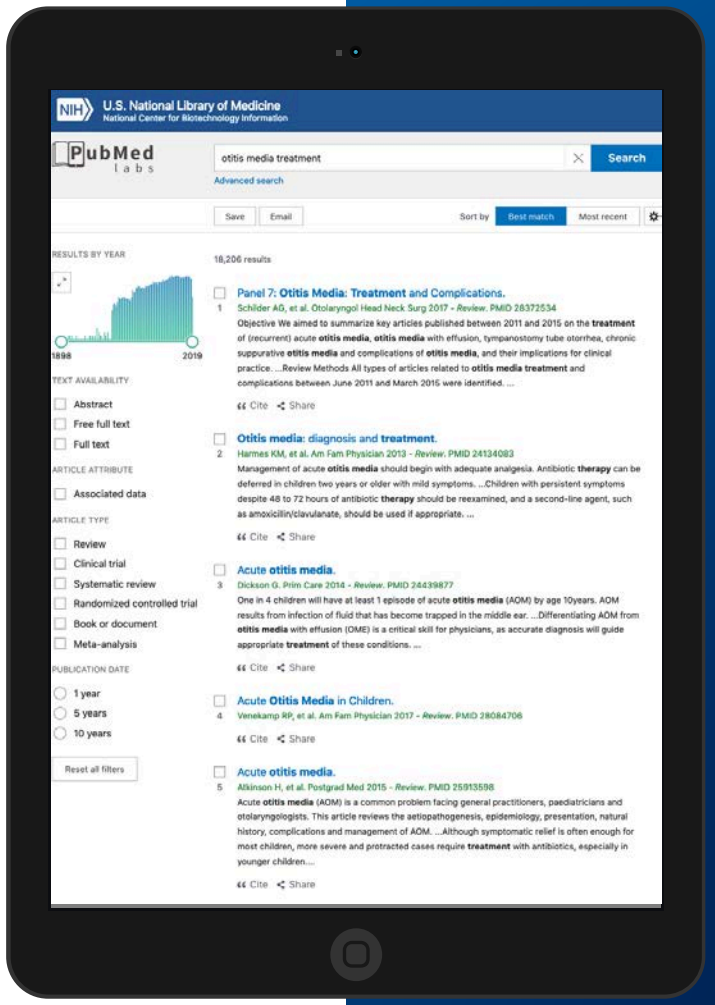
Using metrics to improve PubMed

PubMed

An essential, free **citation** resource that connects researchers, clinicians, healthcare providers, and the public to biomedical literature and data.

Essential service “the interruption of which would endanger. . .the whole or part of the population.”

First offering to be completely rebuilt in cloud environment – a modern model to ensure reliability.



30.5M

Records

53.4M

Interactive Monthly Users

67.1TB

Monthly Bytes Delivered

Legacy PubMed User Survey: NPS

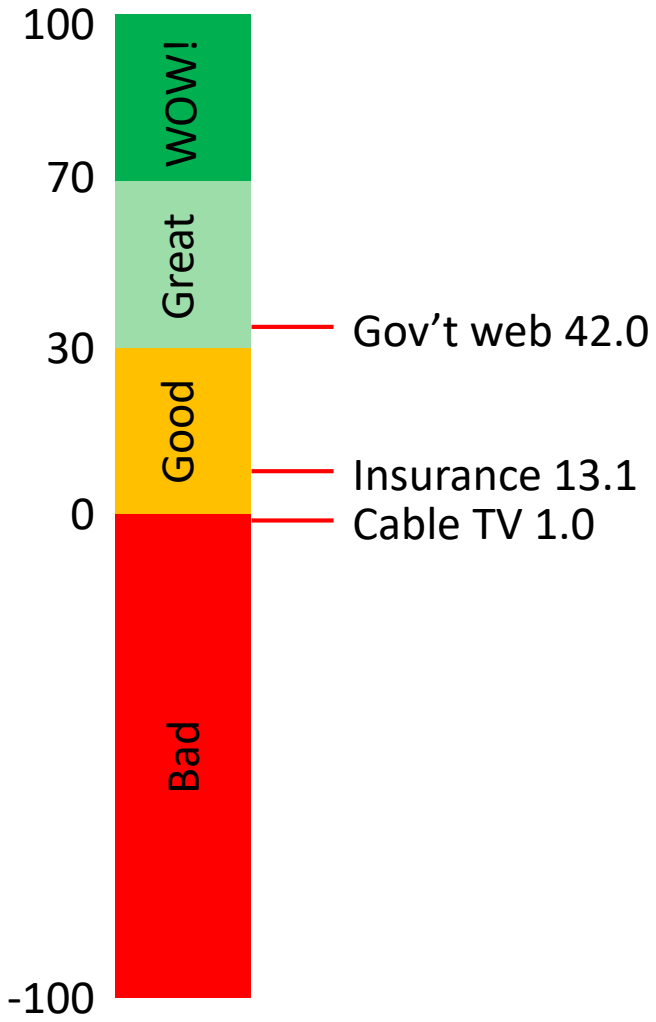
On a scale from 0-10, how likely are you to recommend this site to a colleague?

NPS

=

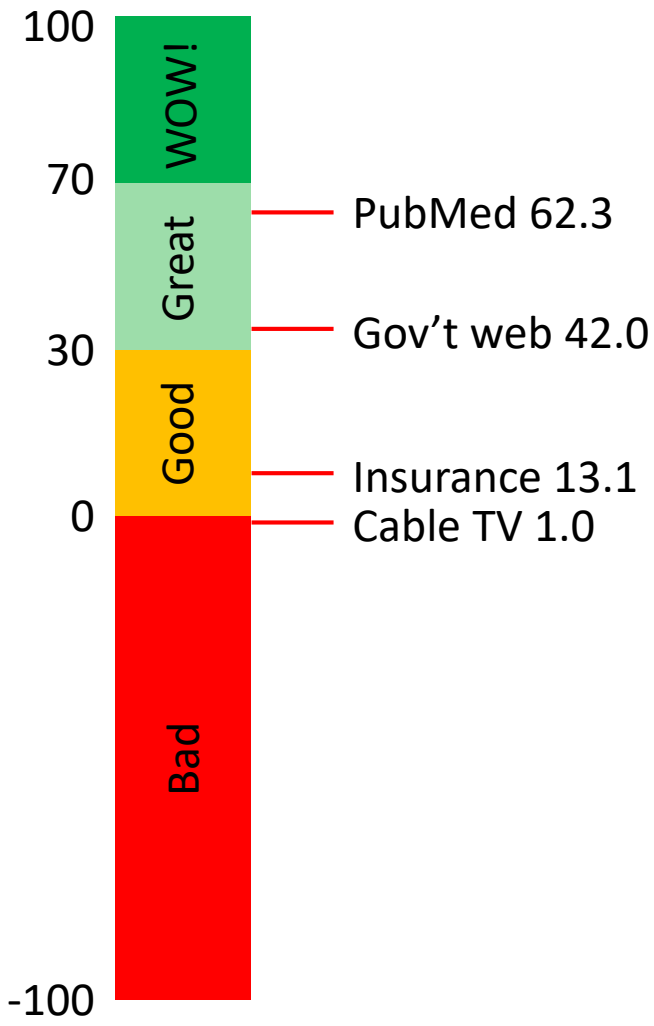
%Promoters (9-10) - %Detractors (0-6)

Legacy PubMed User Survey: NPS

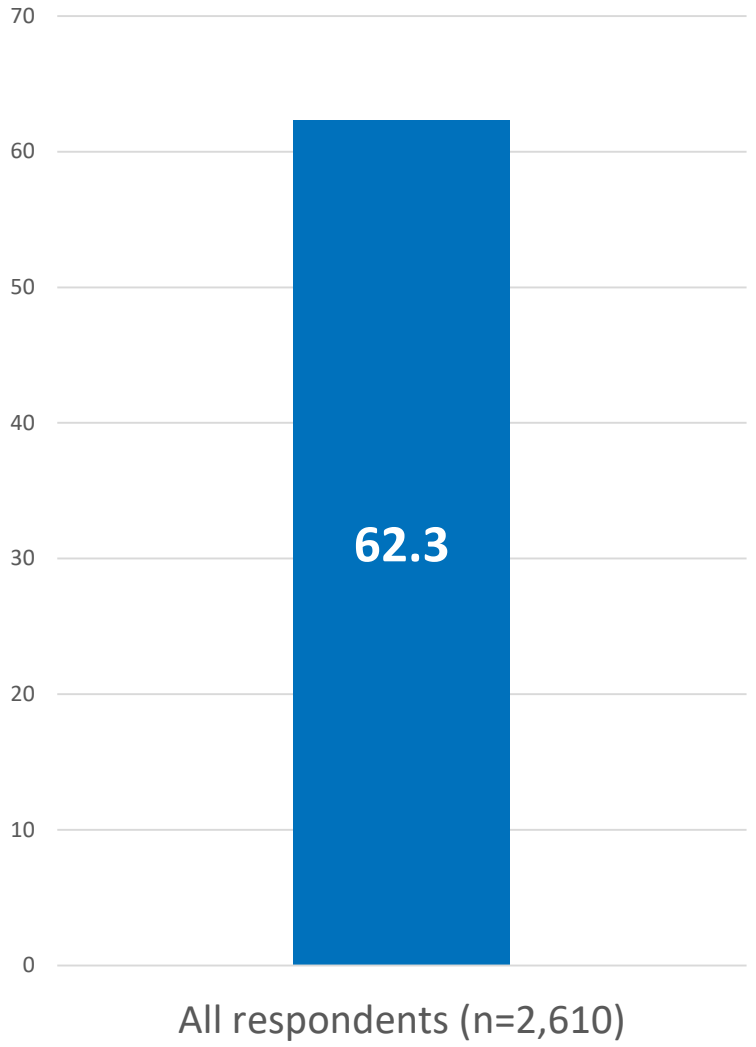


What's a good NPS?

Legacy PubMed User Survey: NPS

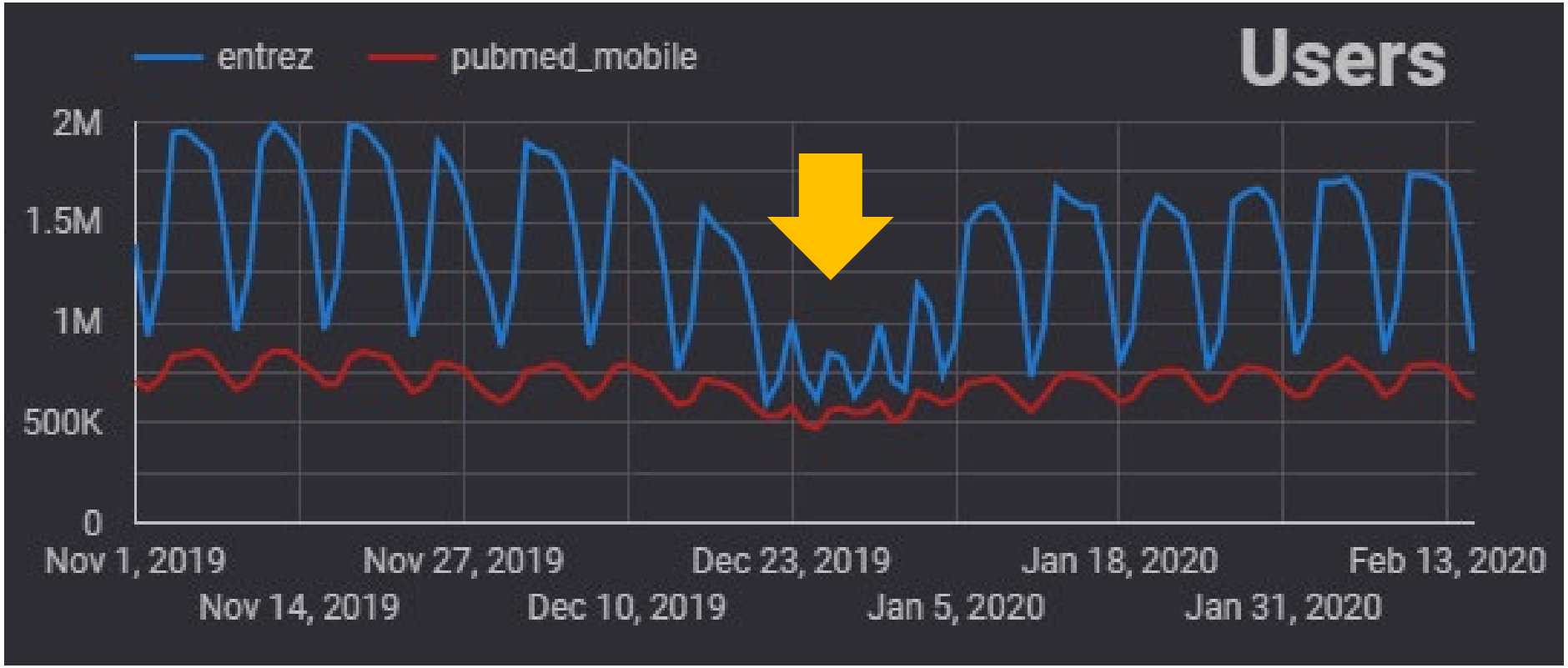


What's a good NPS?



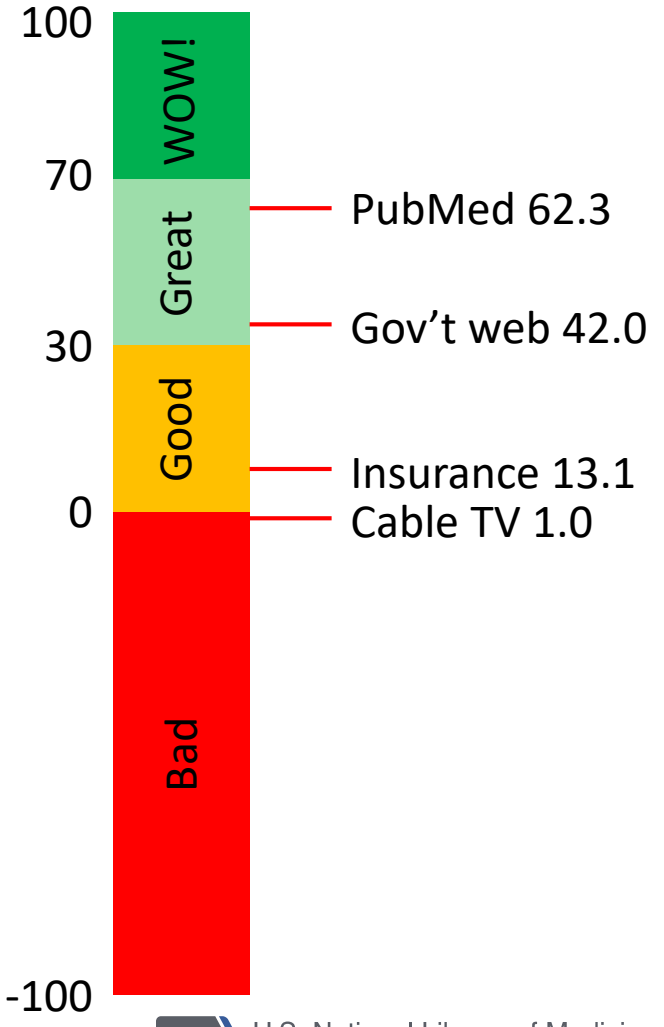
PubMed NPS

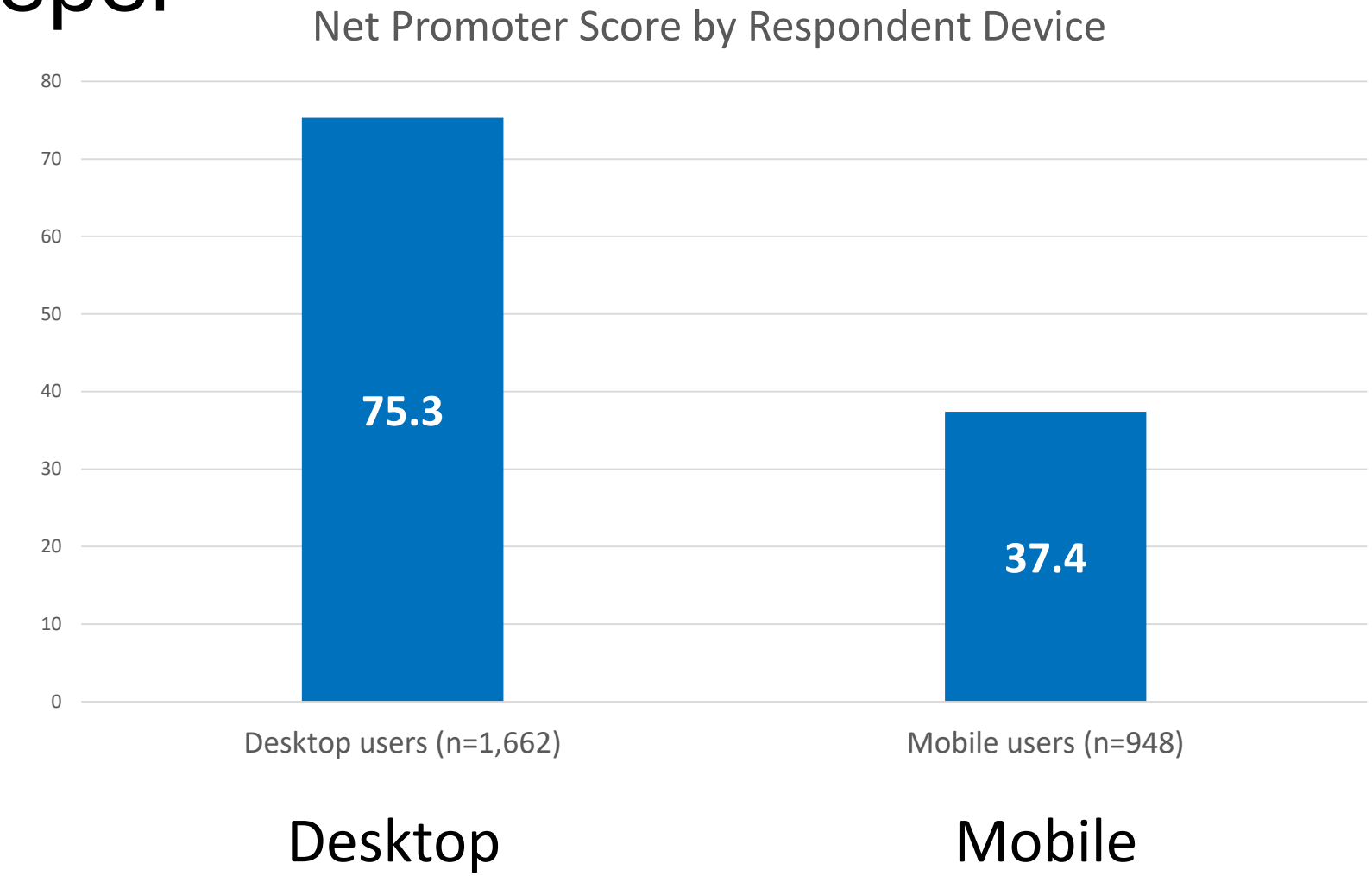
Multiple metrics matter



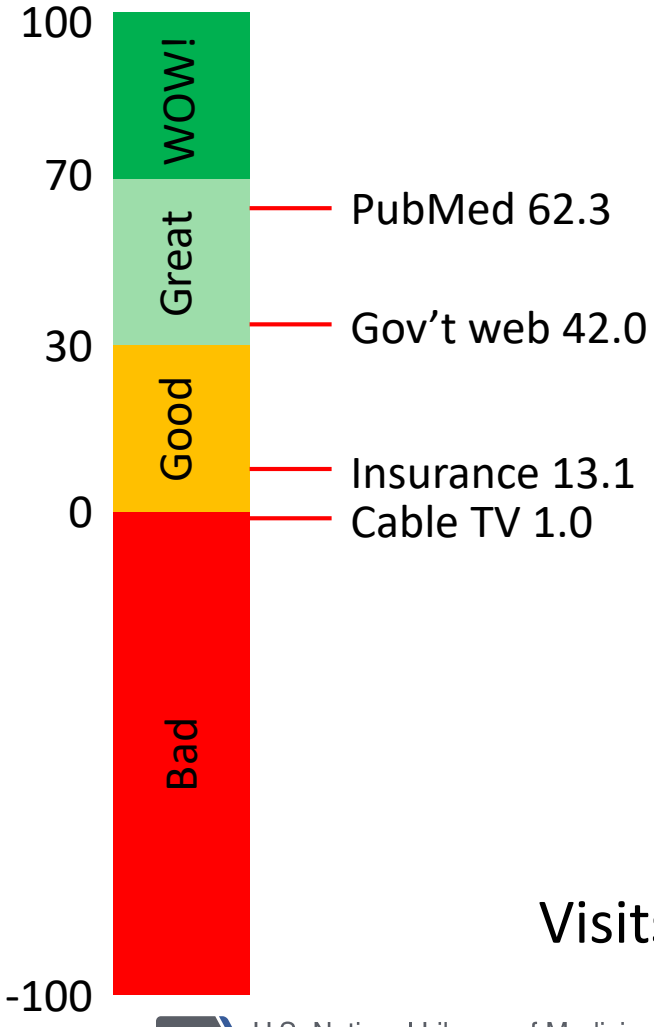
Take-home: Desktop and mobile users exhibit different behaviors

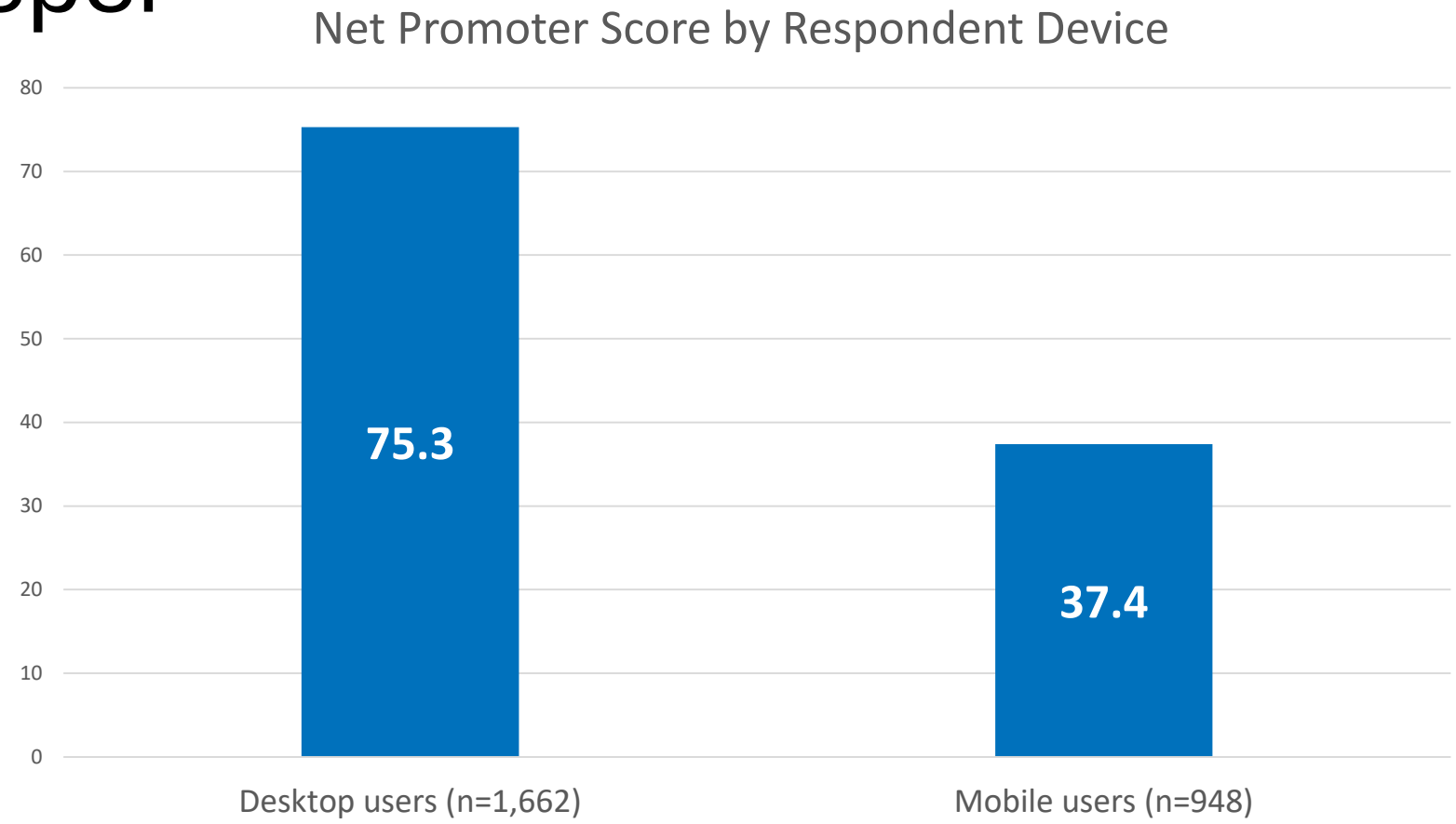
NPS: Digging deeper





NPS: Digging deeper



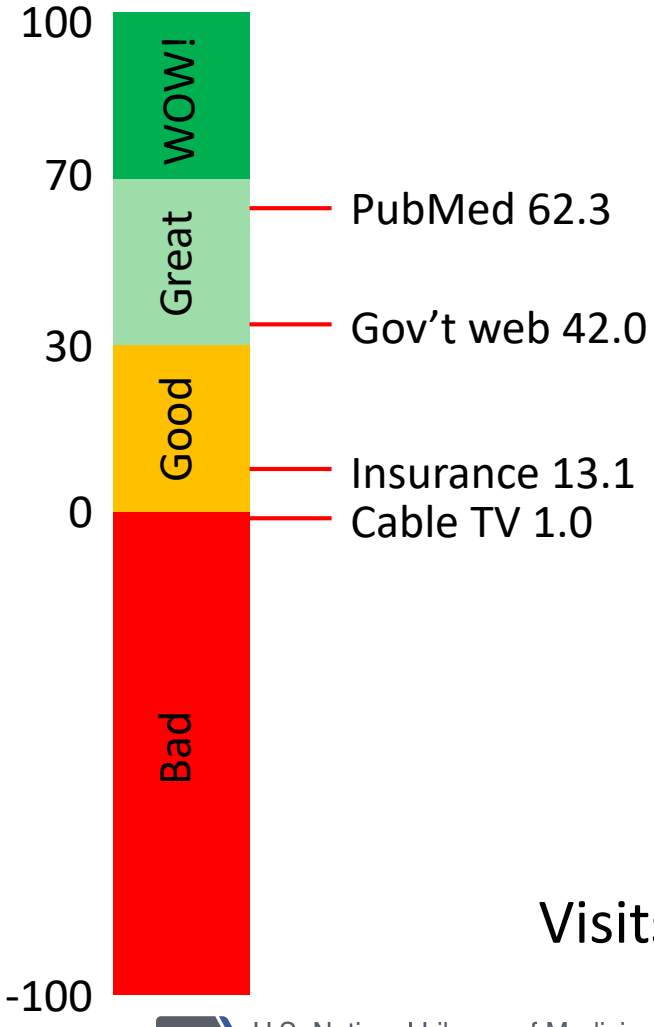


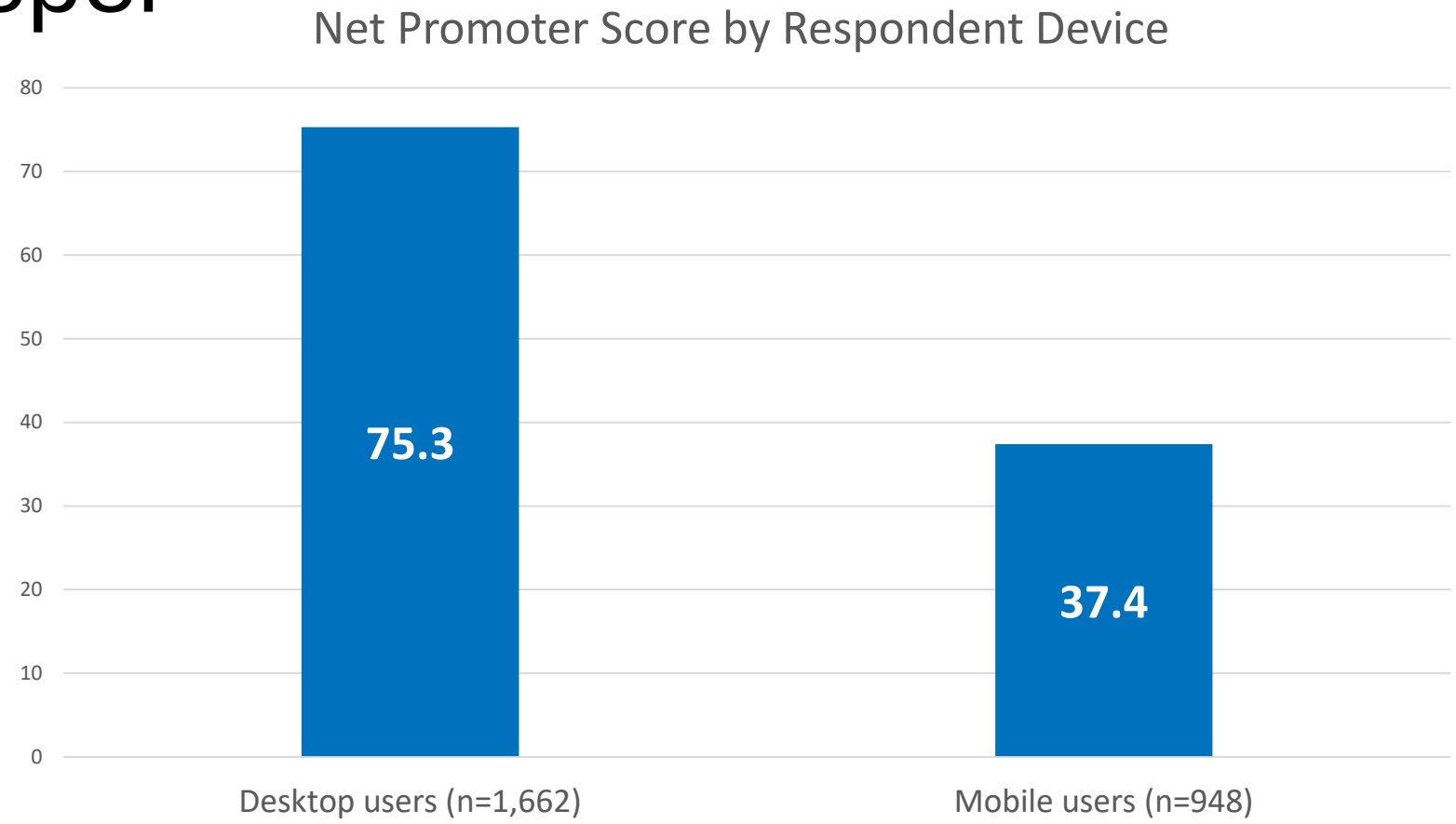
Desktop

Mobile

Visits with Site Search: 36.8%

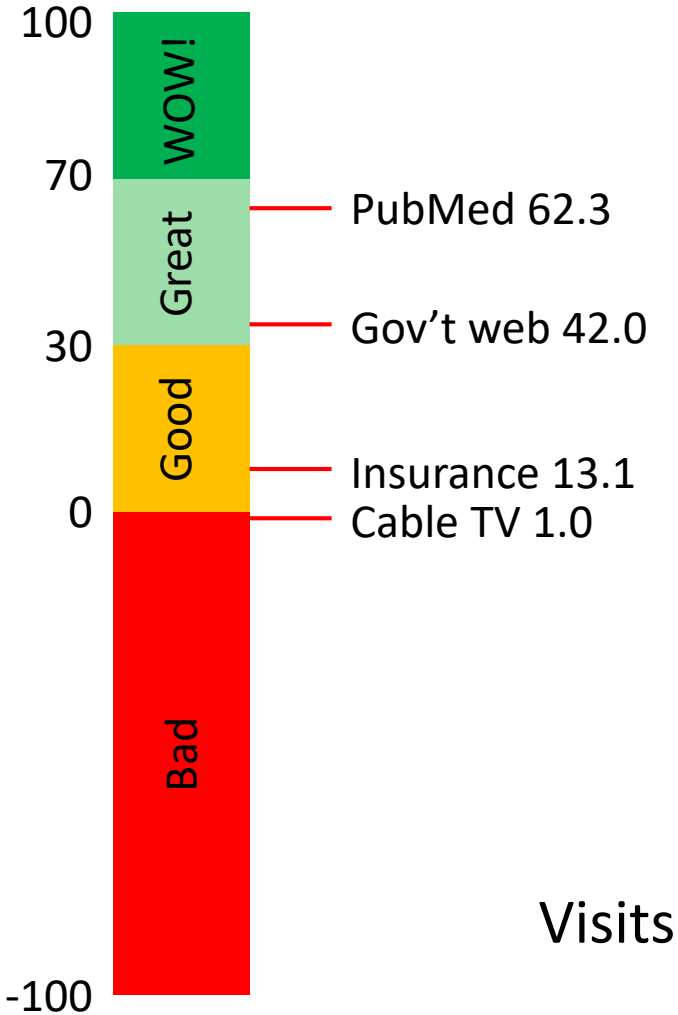
NPS: Digging deeper

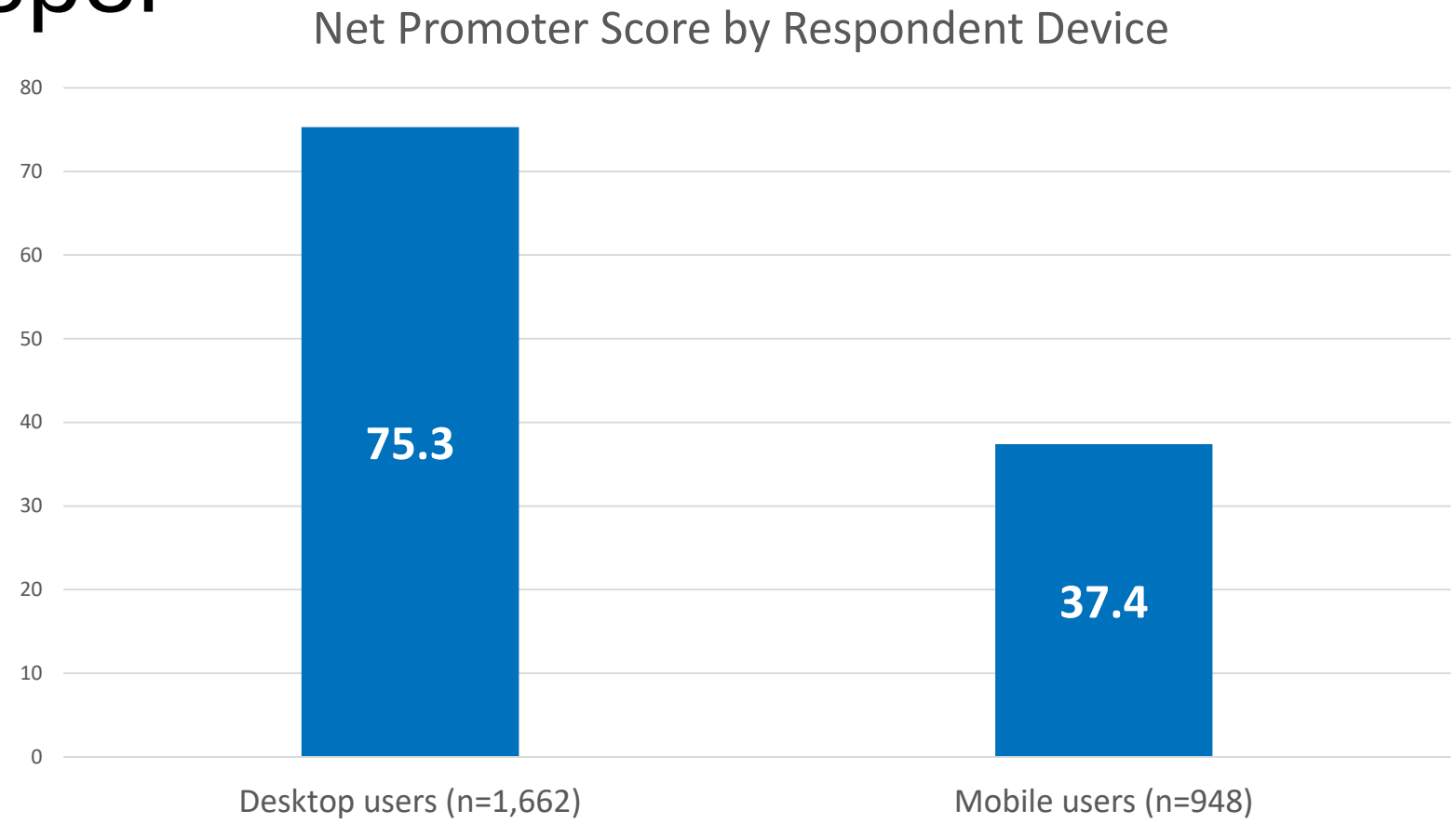




	Desktop	Mobile
Visits with Site Search:	36.8%	8%

NPS: Digging deeper

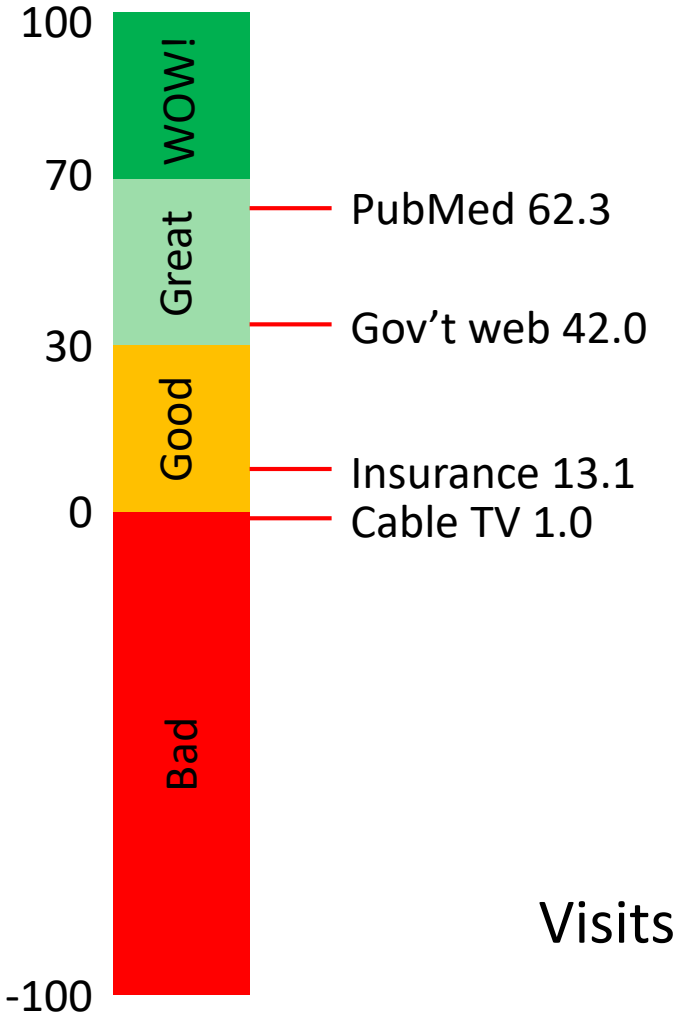


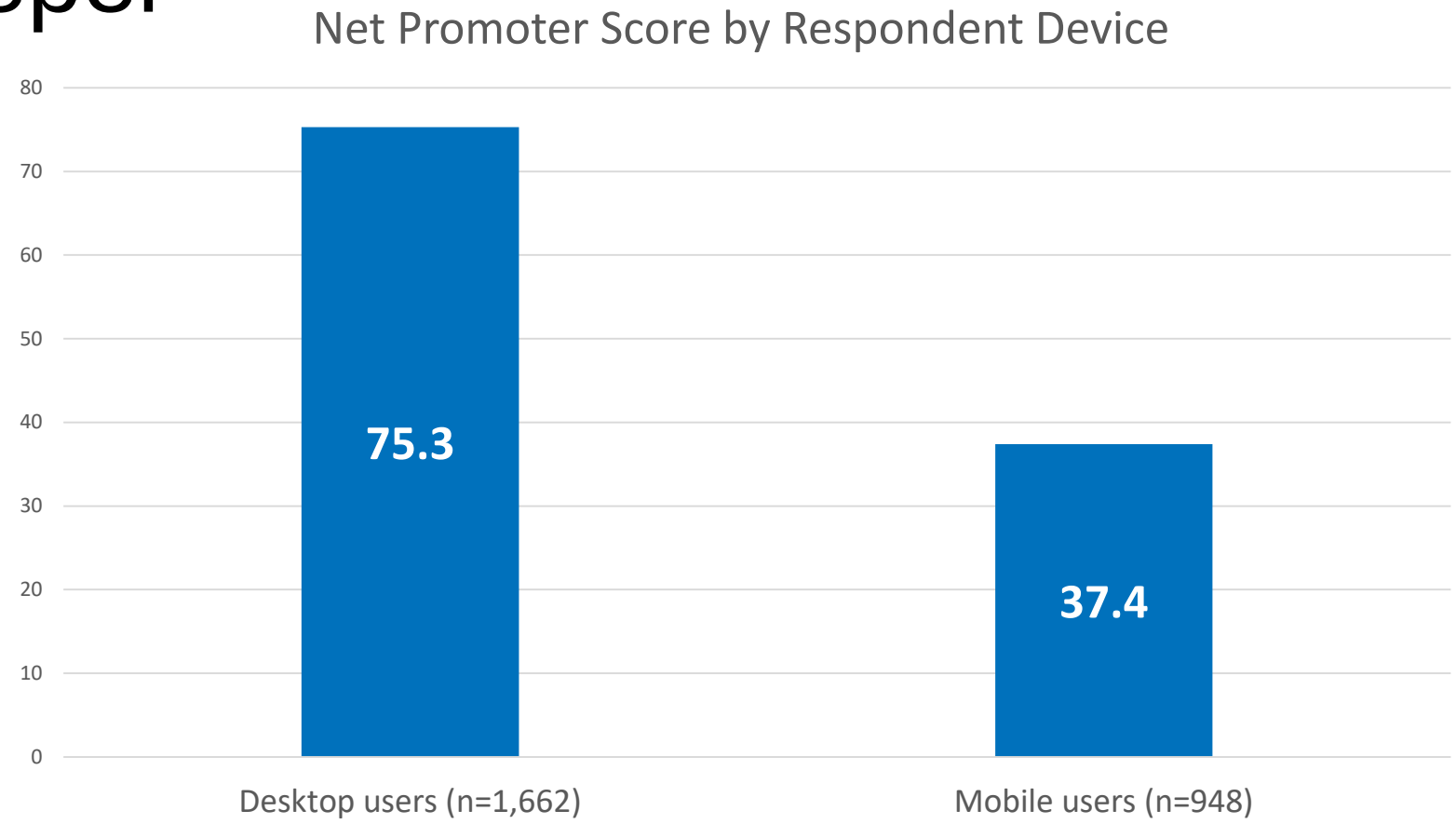


	Desktop	Mobile
Visits with Site Search:	36.8%	8%

10%

NPS: Digging deeper



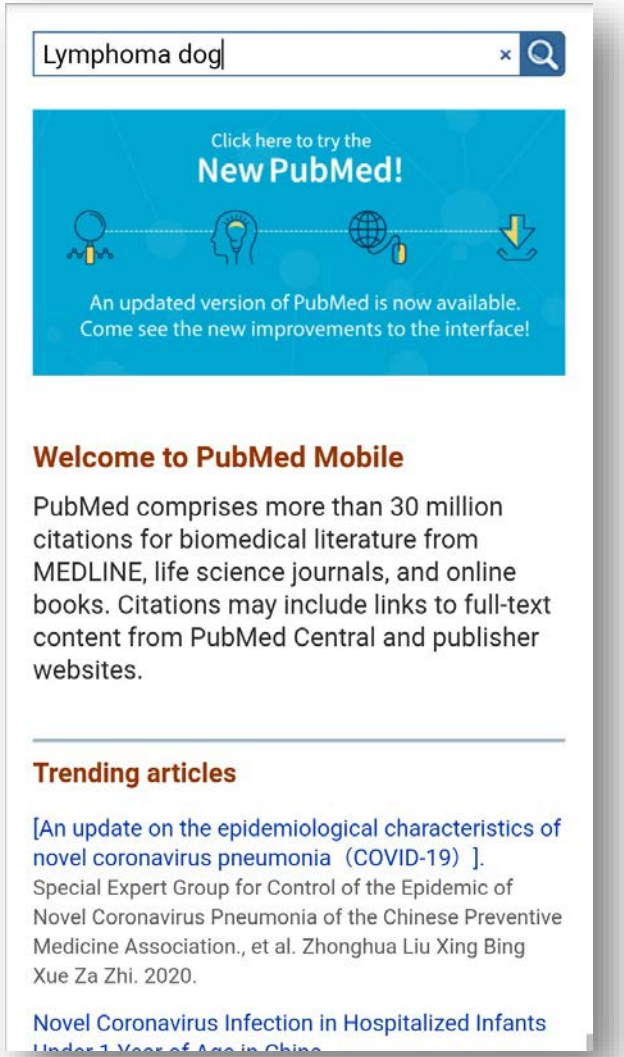


	Desktop	Mobile
Visits with Site Search:	36.8%	8%
	20.6%	

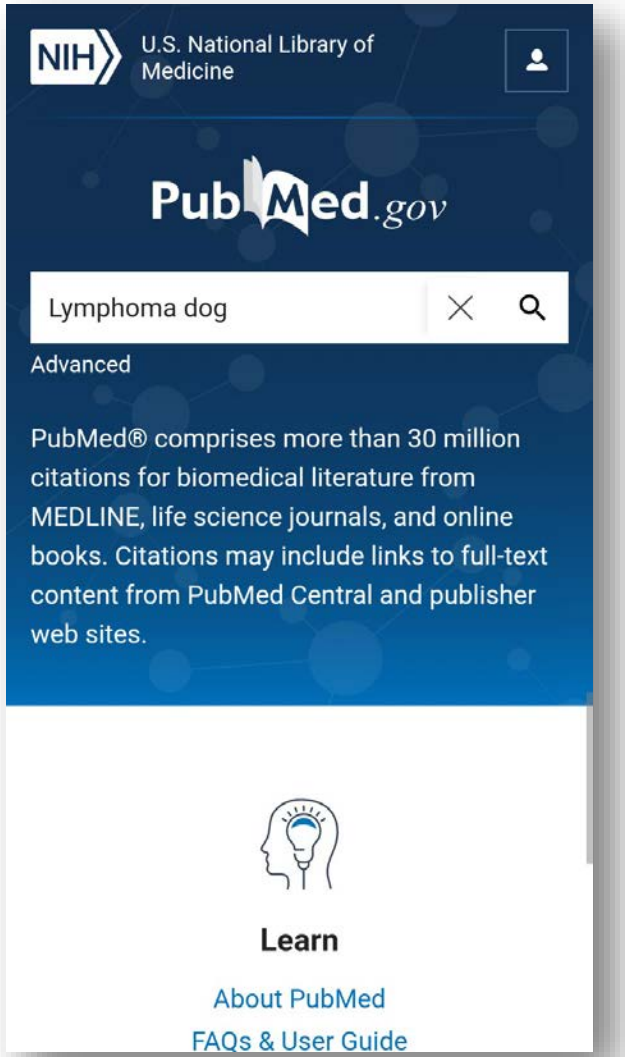
10%

Goal: Improve the mobile PubMed experience!

Legacy:

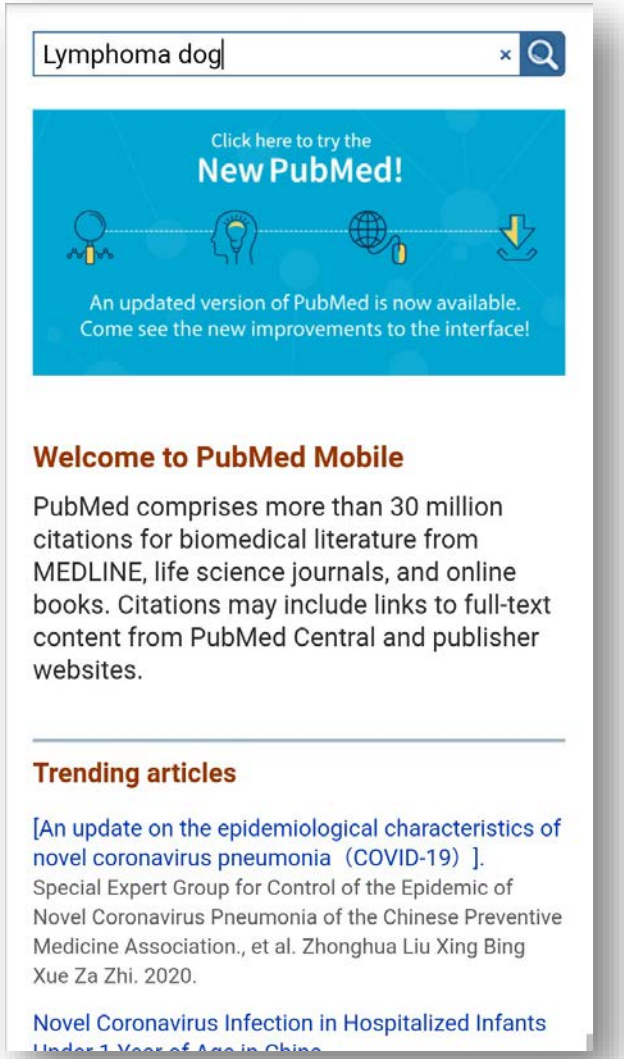


New:

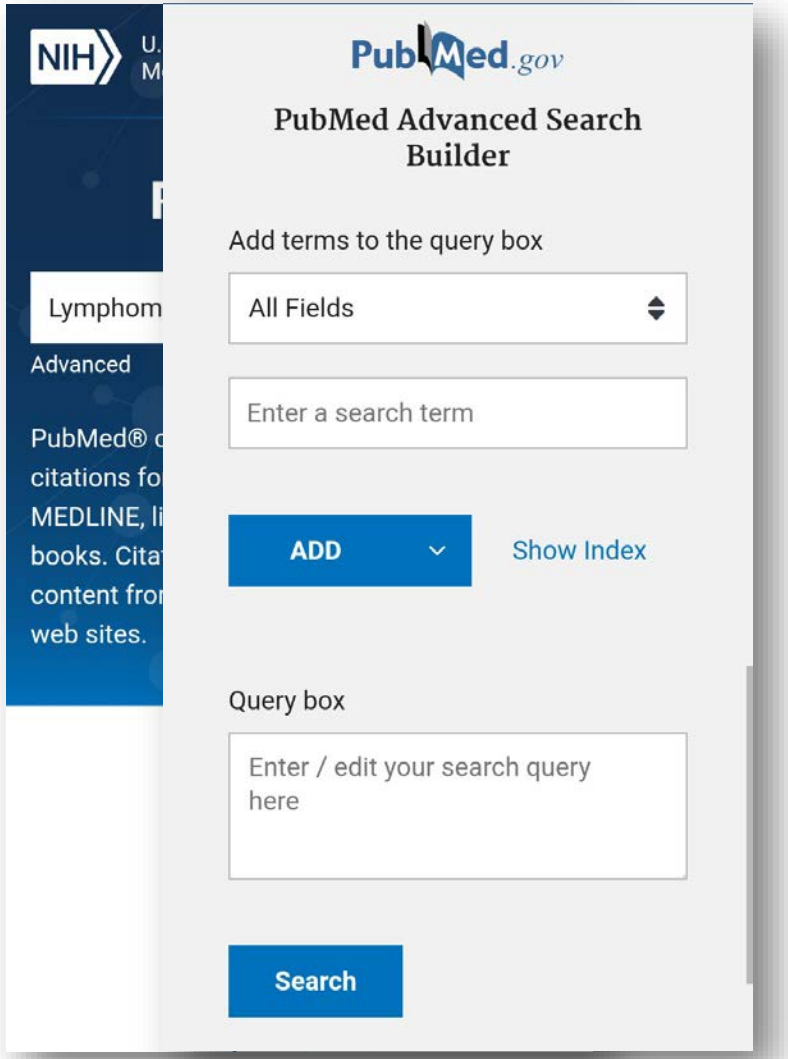


Goal: Improve the mobile PubMed experience!

Legacy:

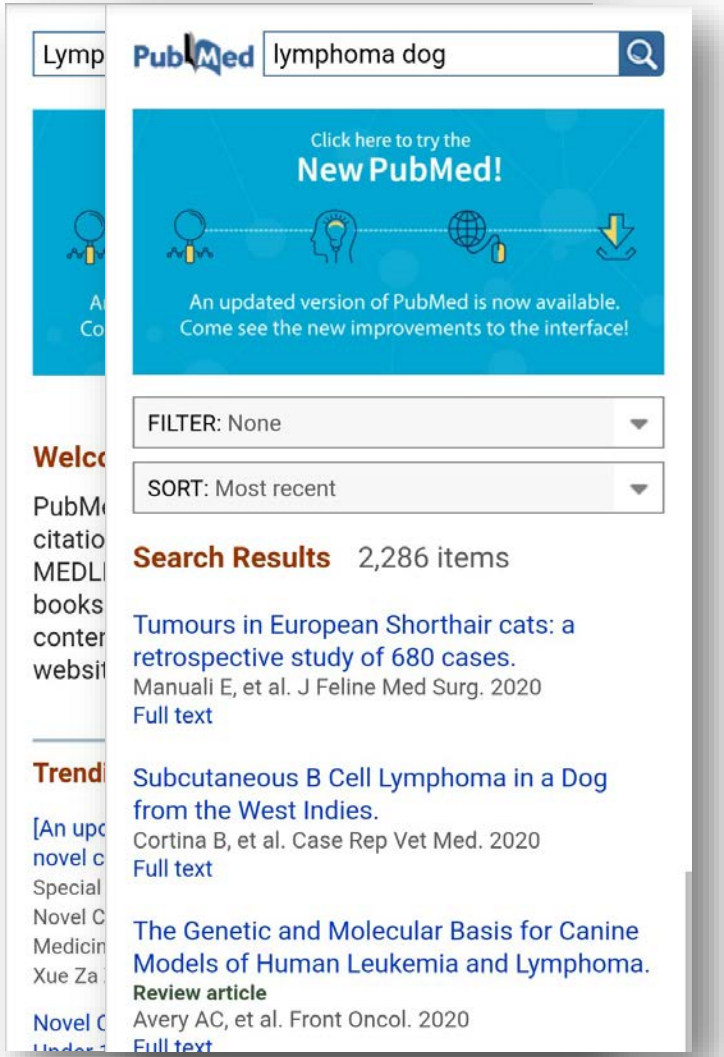


New:

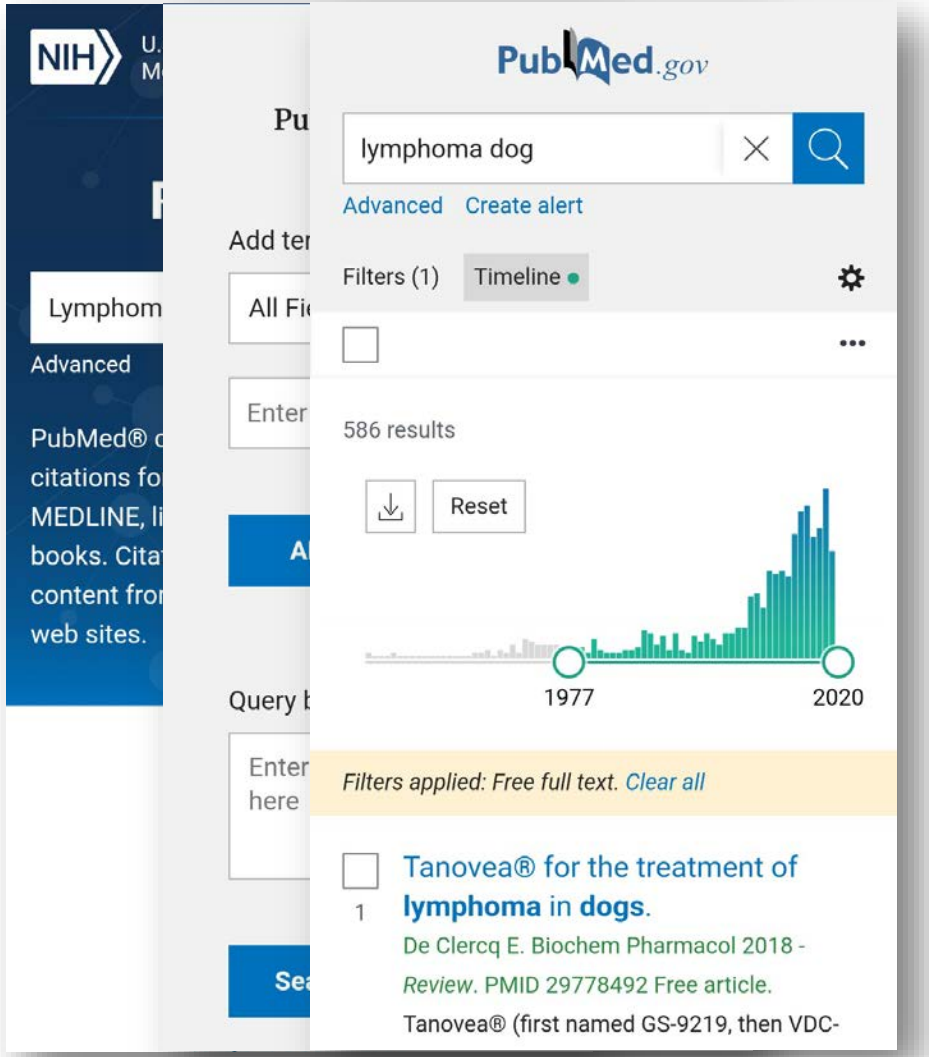


Goal: Improve the mobile PubMed experience!

Legacy:

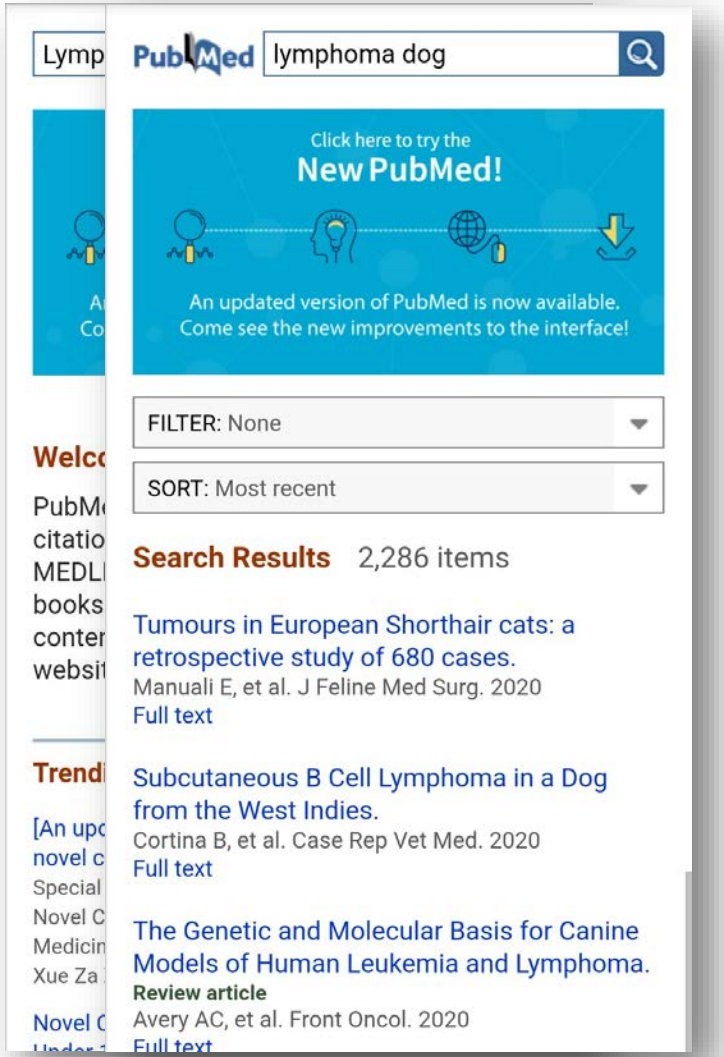


New:

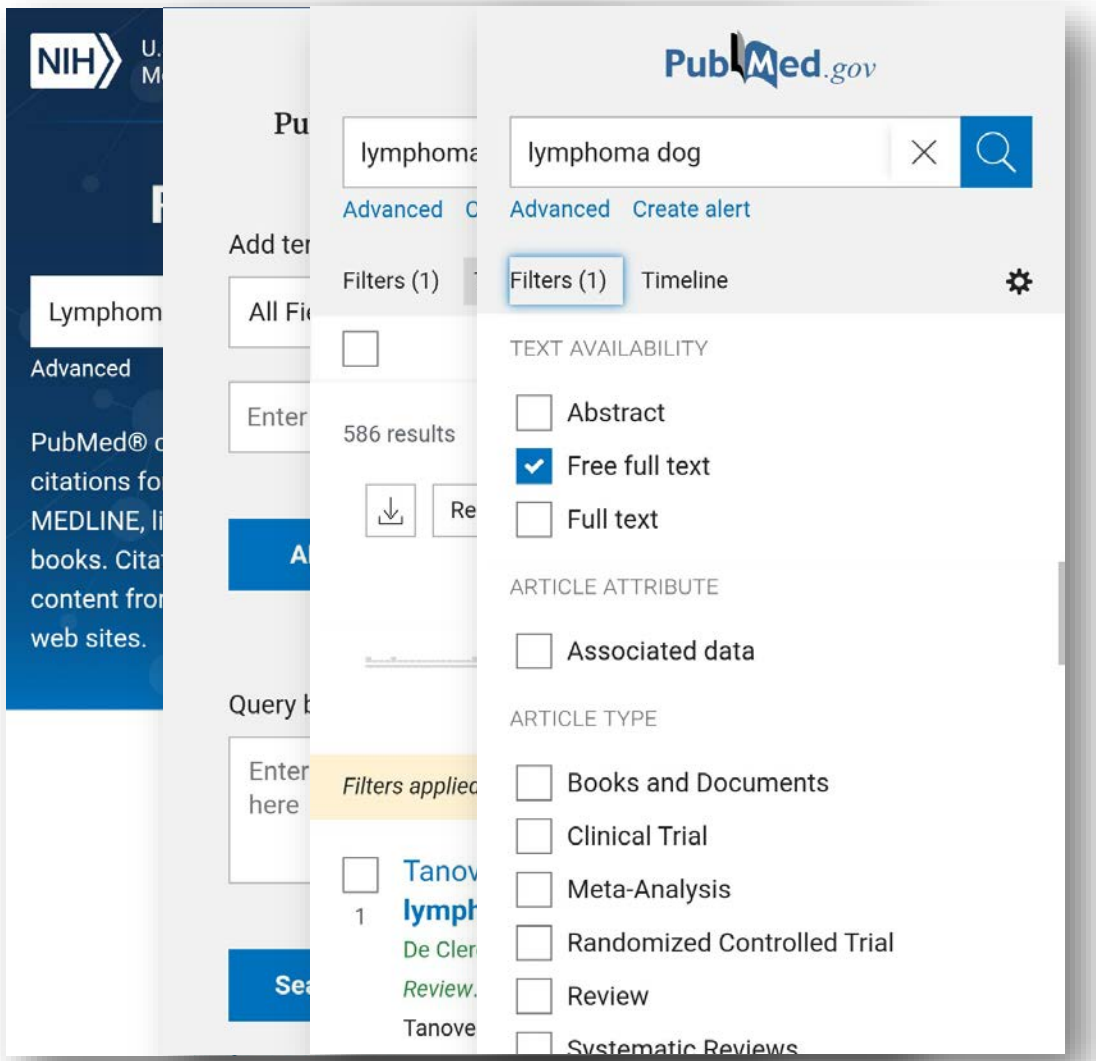


Goal: Improve the mobile PubMed experience!

Legacy:



New:



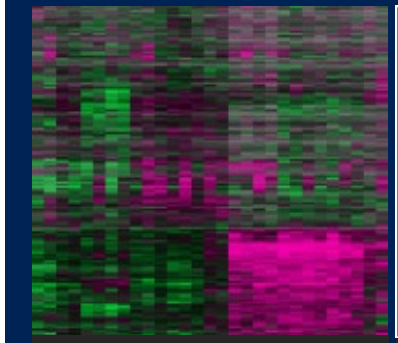
GEO

Measuring repository value through dataset usage metrics

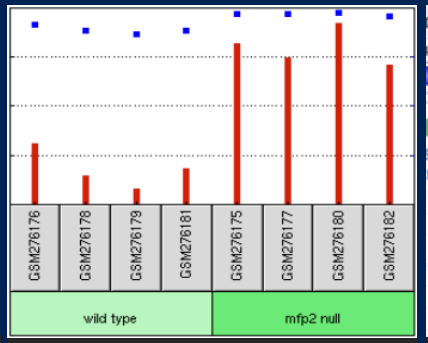
Gene Expression Omnibus

The largest collection of richly-annotated, open-access **gene expression and epigenomics datasets** from all branches of life.

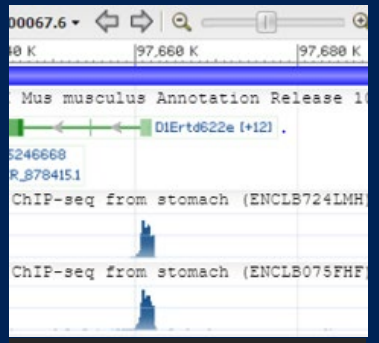
Data analysis and visualization tools



heatmaps



expression profiles



genome tracks

3.5M
Records

160,000
Interactive Monthly Users

1.5 TB
Monthly Bytes Delivered

- Submission rates doubled over last 5 years.
- Working to deliver more processed data for users (e.g., gene expression counts and genome visualization tracks).
- >68,000 depositor citations
- >10,000 third-party usage citations

NCBI


My question is:
Are we making an impact???
S.GROSS

GEO third-party usage citations

- Dataset usage: lagging indicator of value
- Inform curation activities (invest more effort in most used studies)
- Inform development of tools or search strategies
- Measurement not yet automated
- Metric for which further efforts are planned

NCBI

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO » Info » Citation listings

User: tanya | My submissions | Logout

Citation listings: deposit and third-party usage

GEO third-party usage citations | GEO deposit citations

The following list represents third-party publications that cite GEO data as evidence to support c
complement independent studies, or use GEO data as the basis of statistical/analytical hypotheses
or tools. Please report omitted publications to geo@ncbi.nlm.nih.gov

Total number of citations: 10,426

Wang P, Liu J, Song Y, Liu Q et al.
Screening of immunosuppressive factors for biomarkers of breast cancer malignancy phenotypes and subtype-specific targeted therapy.
PeerJ 2019;7:e7197. PMID: 31293831

Zhang T, Guo J, Gu J, Chen K et al.
KIAA0101 is a novel transcriptional target of FoxM1 and is involved in the regulation of hepatocellular carcinoma microvascular invasion by regulating epithelial-mesenchymal transition.
J Cancer 2019;10(15):3501-3516. PMID: 31293655

Li Y, Wu Y, Zhang X, Bai Y et al.
SCIA: A Novel Gene Set Analysis Applicable to Data With Different Characteristics.
Front Genet 2019;10:598. PMID: 31293623

Cortesi M, Pasini A, Furini S, Giordano E.
Identification via Numerical Computation of Transcriptional Determinants of a Cell Phenotype Decision Making.
Front Genet 2019;10:575. PMID: 31293614

Hu D, Kan G, Hu W, Li Y et al.
Identification of Loci and Candidate Genes Responsible for Pod Dehiscence in Soybean via Genome-Wide Association Analysis Across Multiple Environments.
Front Plant Sci 2019;10:811. PMID: 31293609

Bao M, Jiang G.
Differential expression and functional analysis of lung cancer gene expression datasets: A systems biology perspective.
Oncol Lett 2019 Jul;18(1):776-782. PMID: 31289554

Yu Z, Xu Q, Wang G, Rowe M et al.
DNA topoisomerase IIα and RAD21 cohesin complex component are predicted as potential therapeutic targets in bladder cancer.
Oncol Lett 2019 Jul;18(1):518-528. PMID: 31289523

Yan Z, Yang J, Fan L, Xu D et al.
31 gene expression-based signatures serve as indicators of prognosis for patients with glioma.
Oncol Lett 2019 Jul;18(1):291-297. PMID: 31289499

dbGaP

Metrics at multiple points in the data life-cycle

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

NCBI

Genotype and Phenotype

A controlled access public archive to store and disseminate **human genotype and associated phenotype data**.

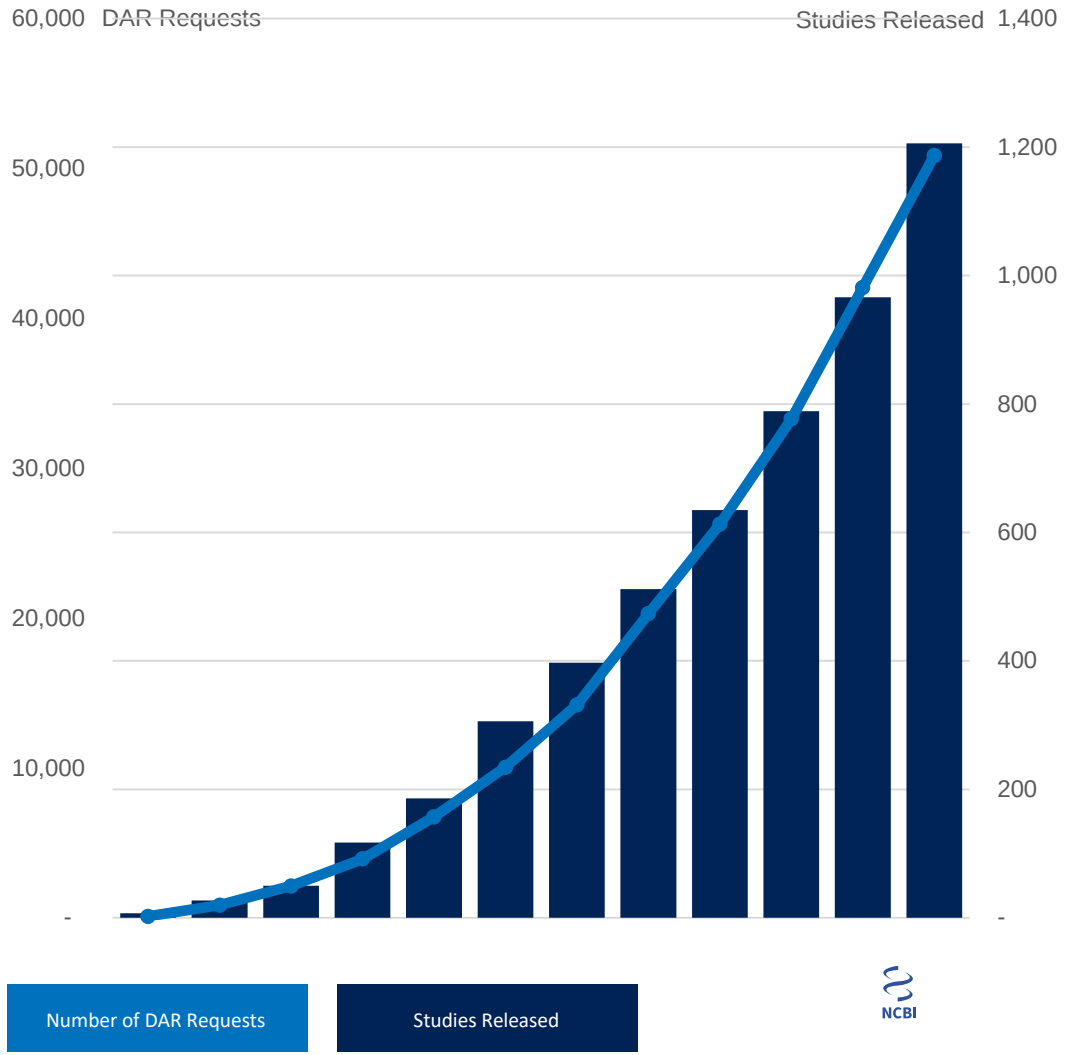
- Phenotype
- Molecular Data
- Documents
- Images
- Association Results

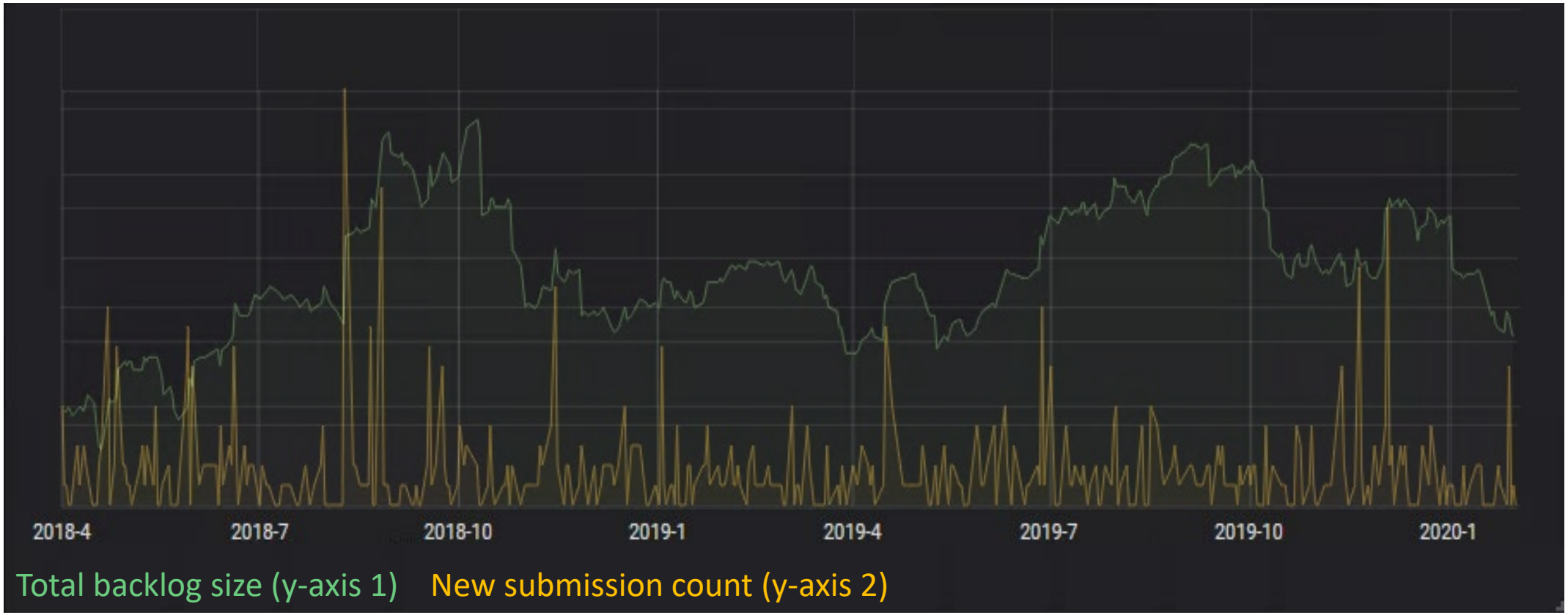
>1,400
Records

>2.3M
People

93,000
Interactive Monthly Users

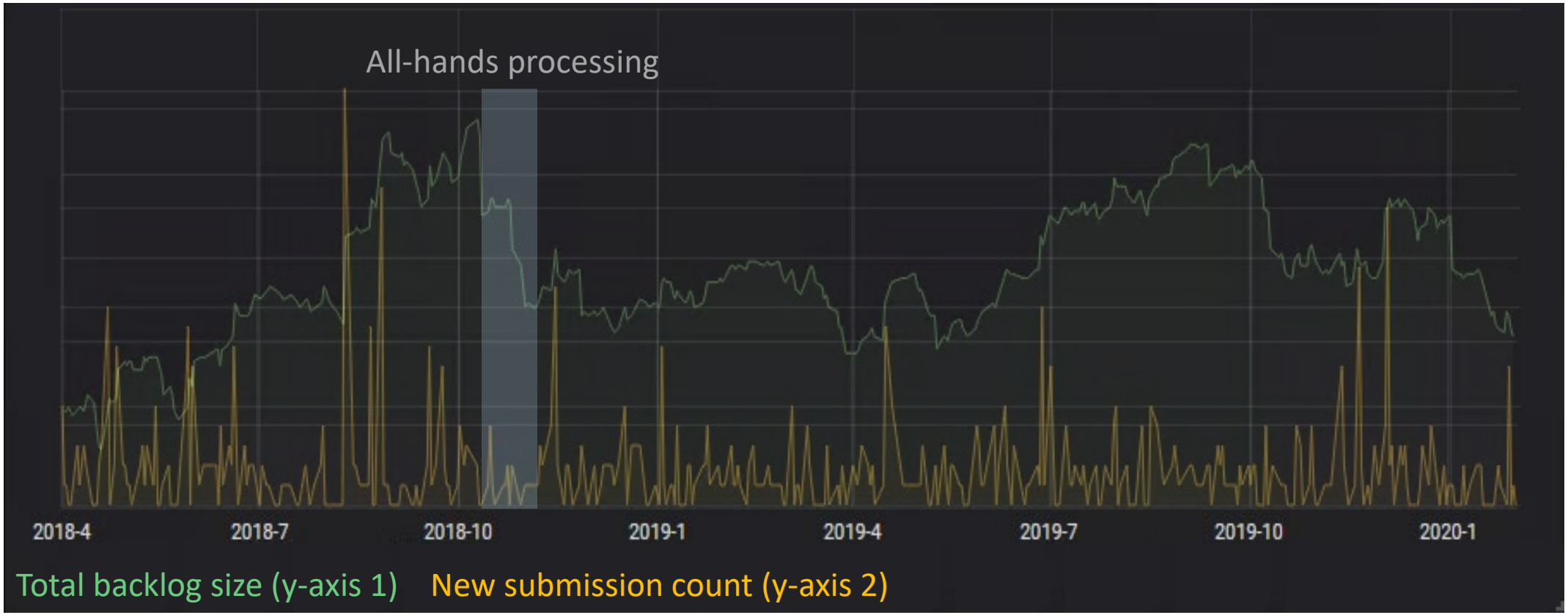
182TB
Monthly Bytes Delivered





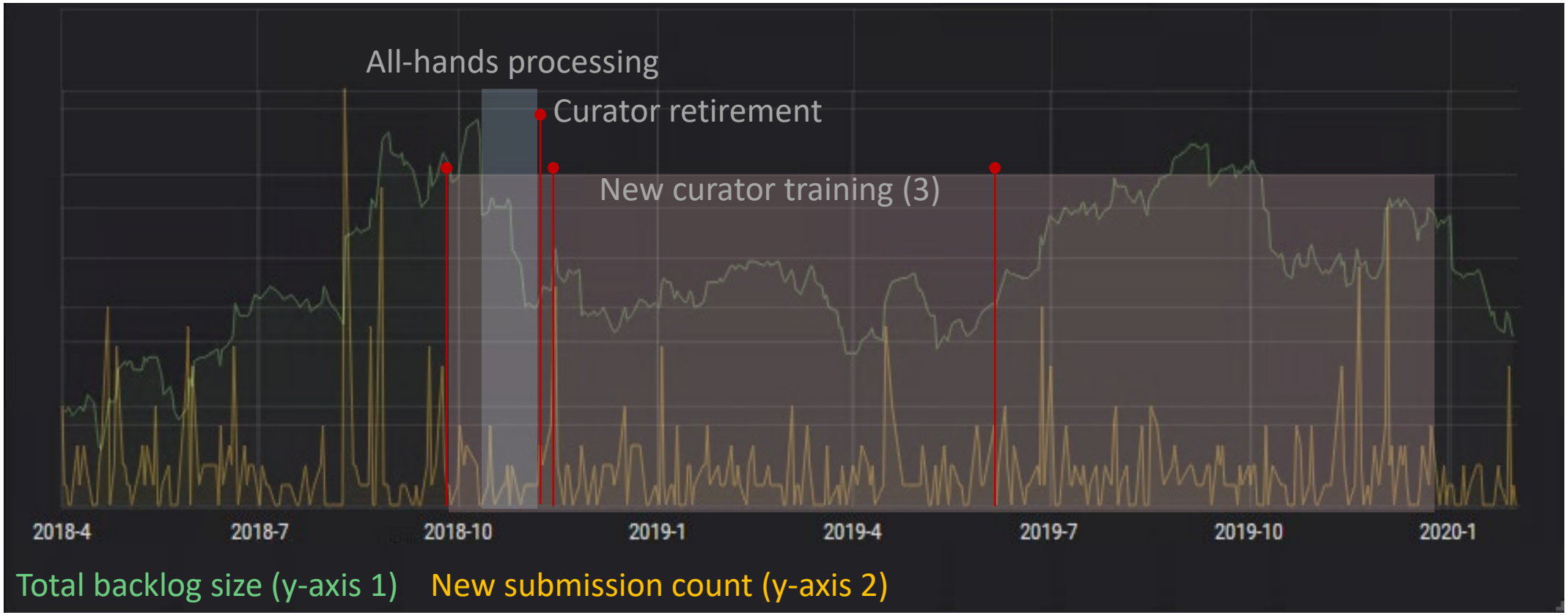
Process metrics: informing progress against goals

- Reduce backlog
- Faster data releases



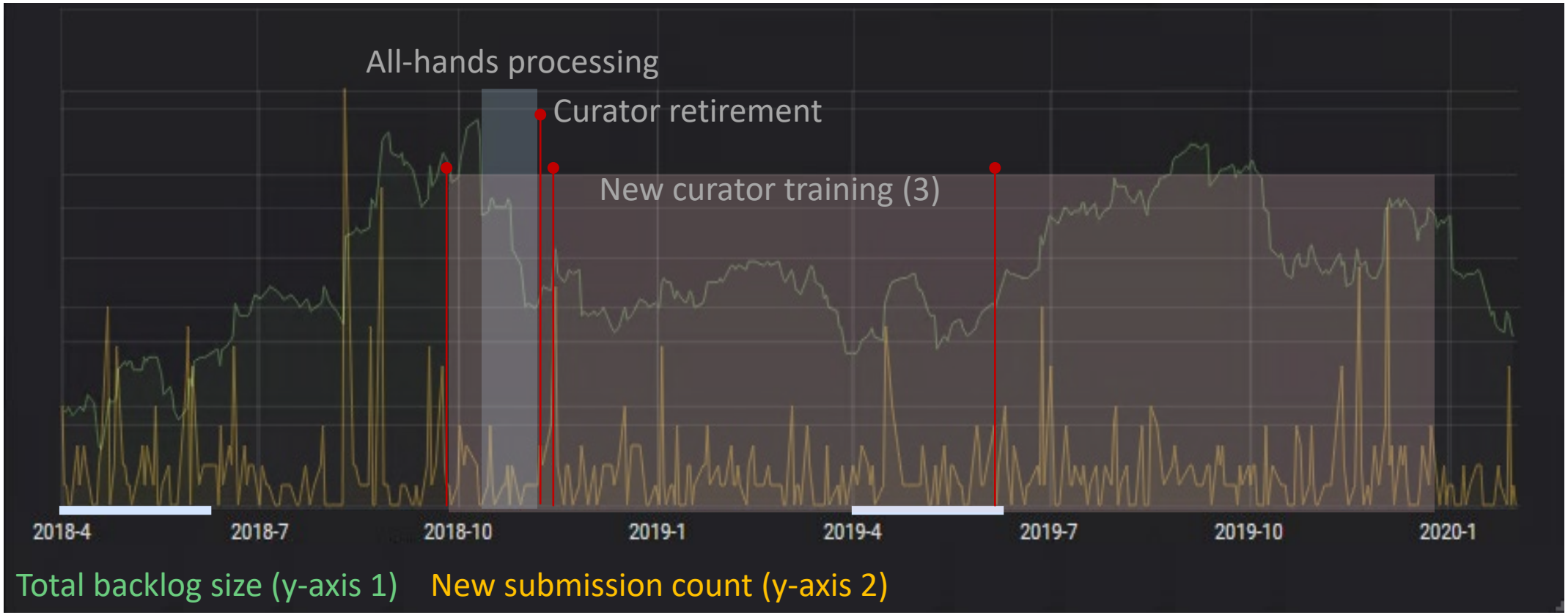
Process metrics: informing progress against goals

- Reduce backlog
- Faster data releases



Process metrics: informing progress against goals

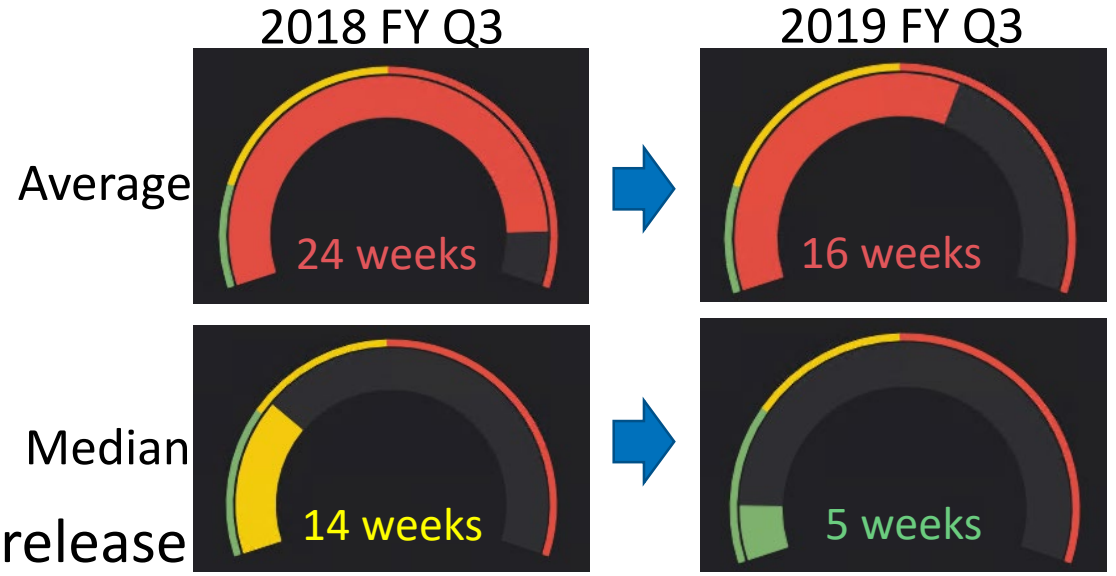
- Reduce backlog
- Faster data releases

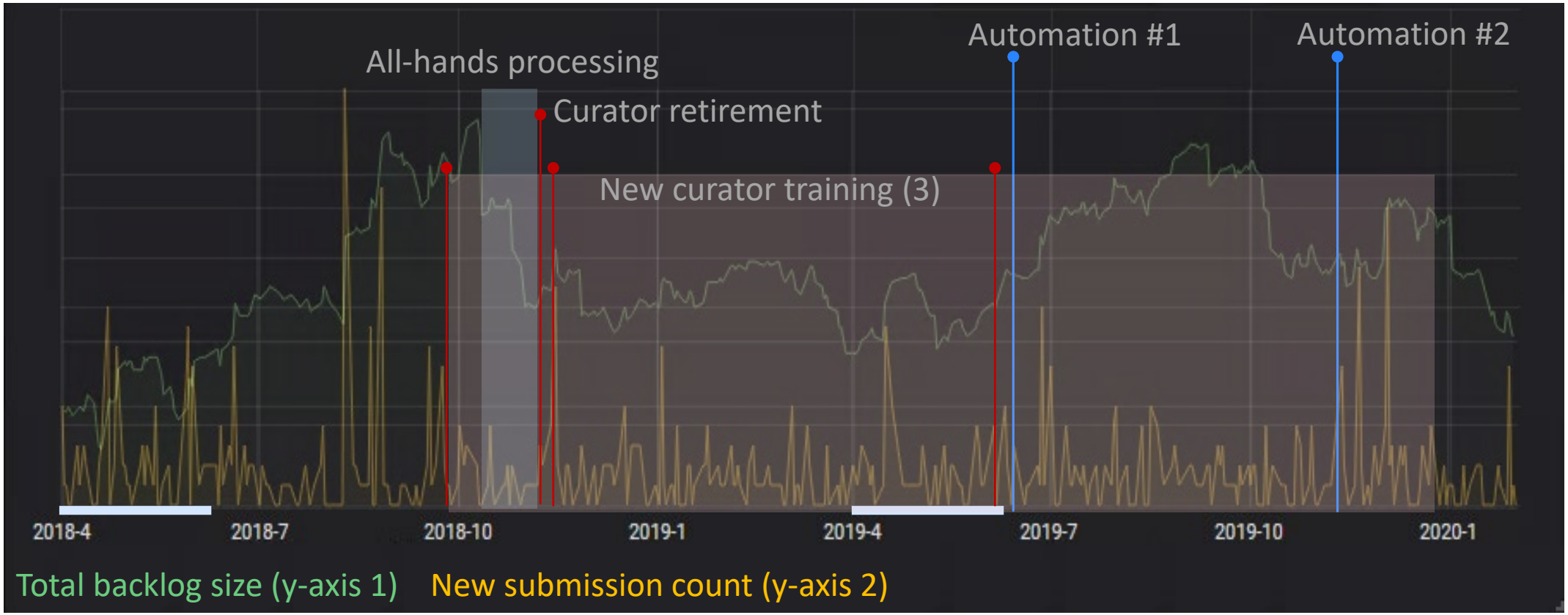


Process metrics: informing progress against goals

- Reduce backlog
- Faster data releases

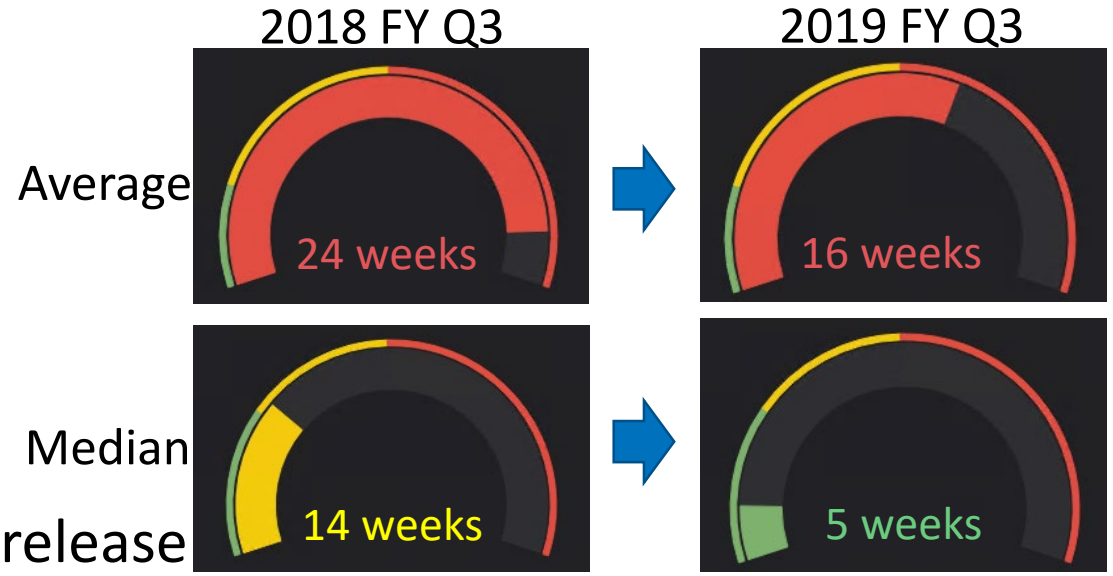
Time of last submission to release





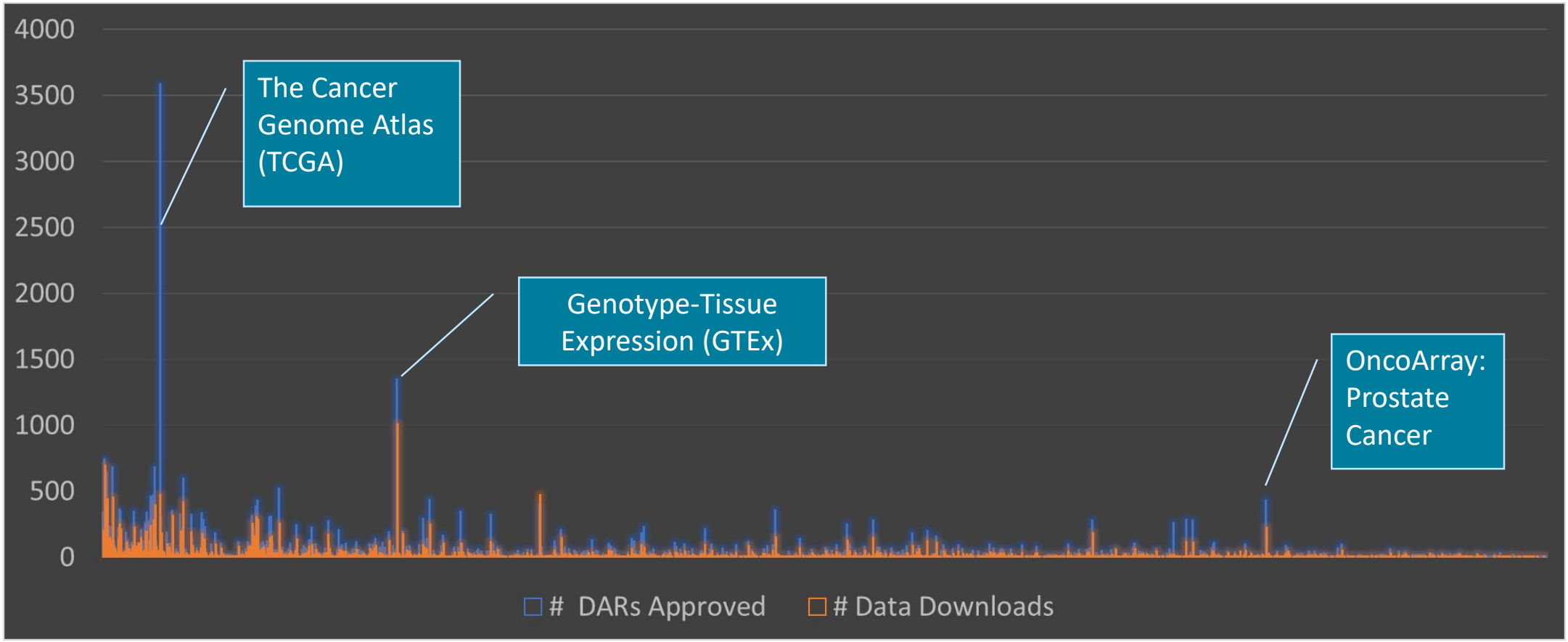
Process metrics: informing progress against goals

- Reduce backlog
- Faster data releases



dbGaP dataset usage: access requests & downloads

Study age: Old → New



Summary

- Who we are
- Metrics 101
- Metrics-Based Resource Management
  - PubMed
  - GEO
  - dbGaP