Welcome

Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer, IQSS
Harvard University's Research Data Officer, HUIT
@mercecrosas

About a year ago, reproducibility caught the eye of *even* Congress

“Concerns about reproducibility and replicability have been expressed in both scientific and popular media. As these concerns came to light, Congress requested that the National Academies of Sciences, Engineering, and Medicine conduct a study to assess the extent of issues related to reproducibility and replicability and to offer recommendations for improving rigor and transparency in scientific research.”

NASEM Consensus Study Report Highlights, Reproducibility and Replicability in Science

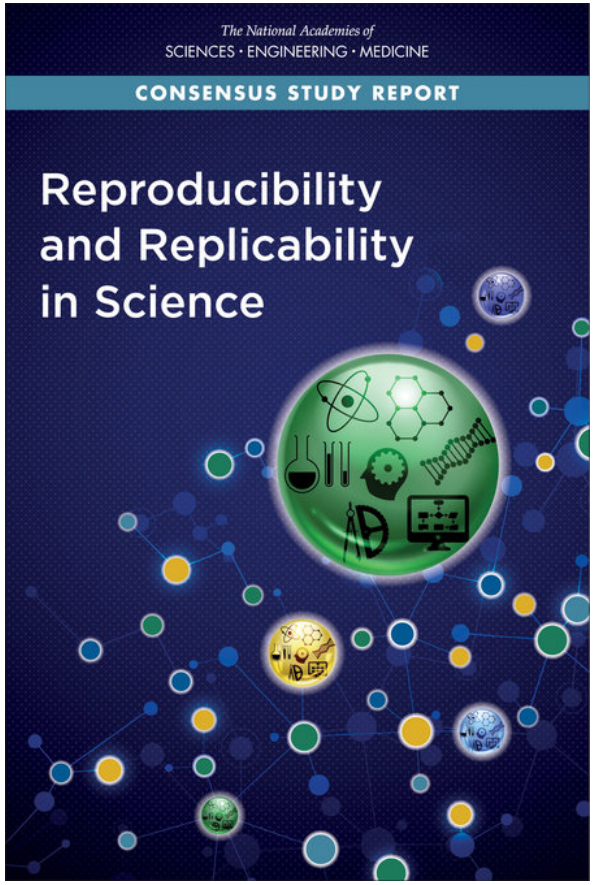
Reproducibility

Obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis. [computational reproducibility]

Replicability

Obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

Definitions from the National Academies of Sciences, Engineering, and Medicine Consensus Study Report Highlights, Reproducibility and Replicability in Science



Report Highlights include:

- **No crisis**, but we must do better
- Include a clear, specific, and complete description of how results are reached
- Promote use of open source tools
- Ensure computational reproducibility during peer review
- Facilitate transparent sharing and availability of digital artifacts, such as data and code

Published May 2019: http://sites.nationalacademies.org/sites/reproducibility-in-science/index.htm

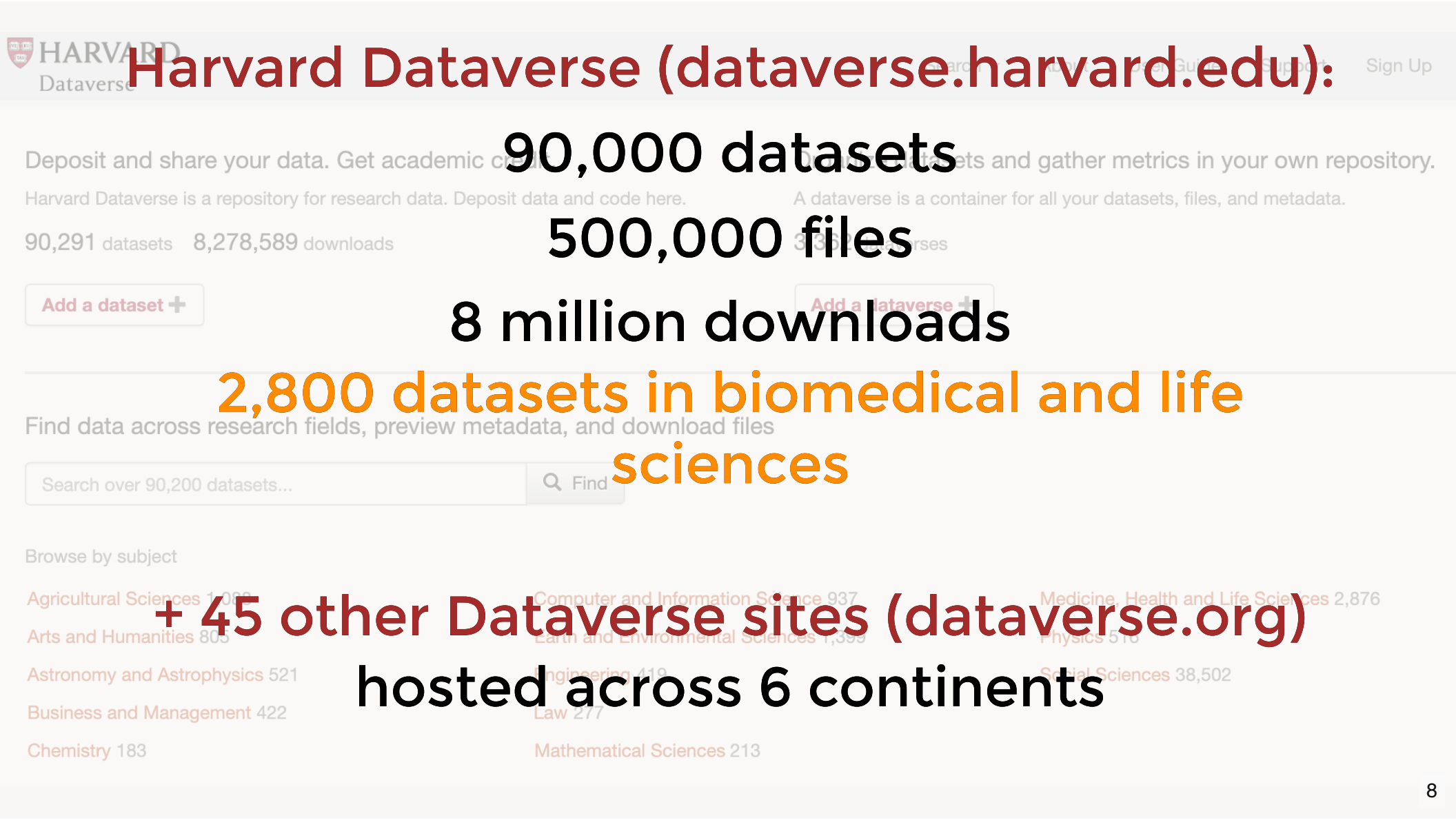
More than a decade ago, reproducibility had already caught the eye of IQSS.

We developed the Dataverse project, an open-source platform to:

“facilitate transparent sharing and availability of digital artifacts, such as data and code"

Data (+Code) Sharing with Dataverse

- Data citation with a persistent identifier (DOI)
- Standard metadata, plus custom metadata
- Tiered access to data:
  - Fully Open, CC0
  - Register to access; Guestbook
  - Restricted with DUA
- Multiple versions of a dataset
- FAIR principles (Findable, Accessible, Interoperable, Reusable)



Harvard Dataverse (dataverse.harvard.edu):

90,000 datasets

500,000 files

8 million downloads

2,800 datasets in biomedical and life sciences

+ 45 other Dataverse sites (dataverse.org)

hosted across 6 continents

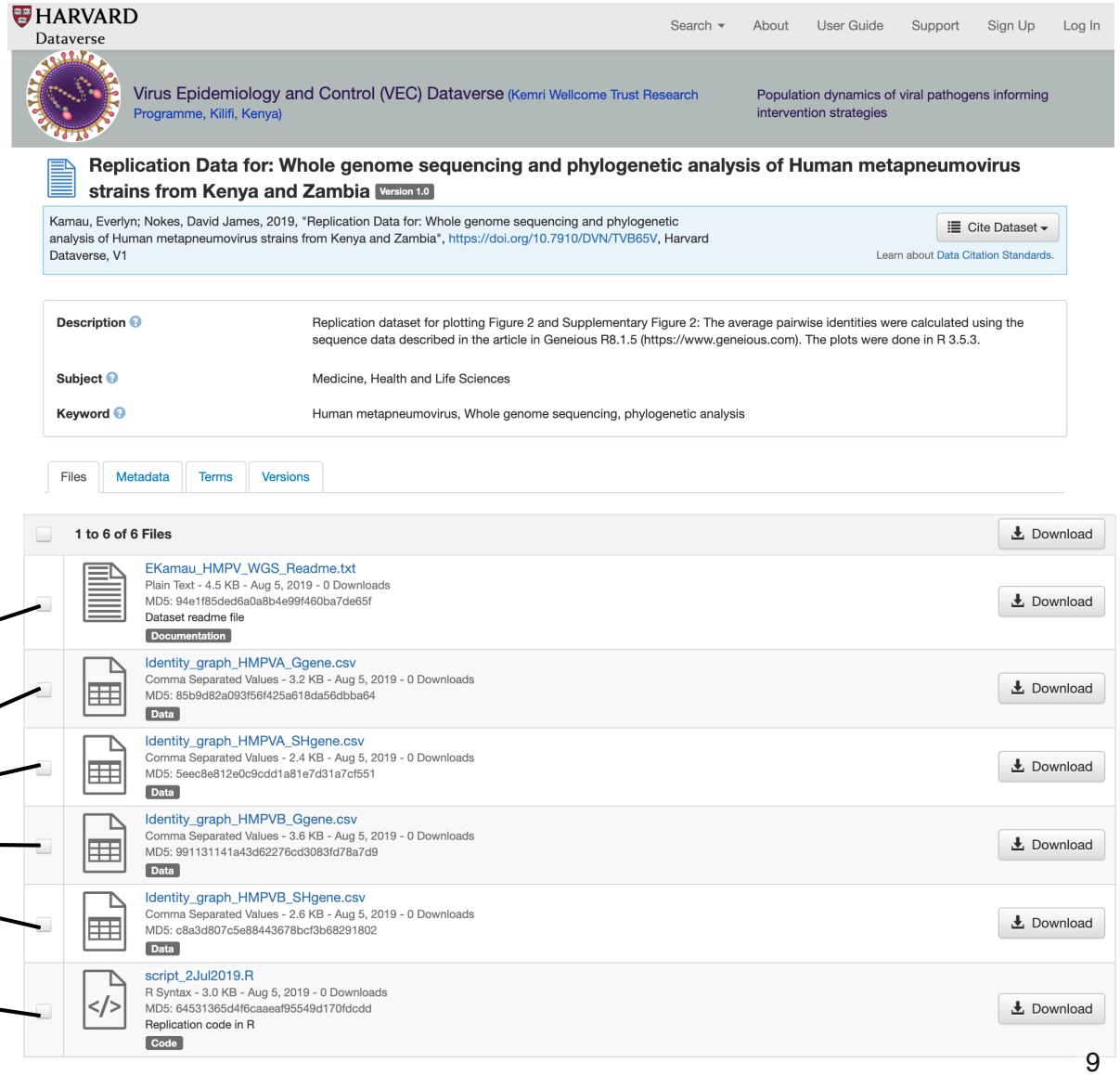
Reproducible?

8,000 of the 90,000 datasets in Harvard Dataverse contain the files to reproduce the published results.

Documentation

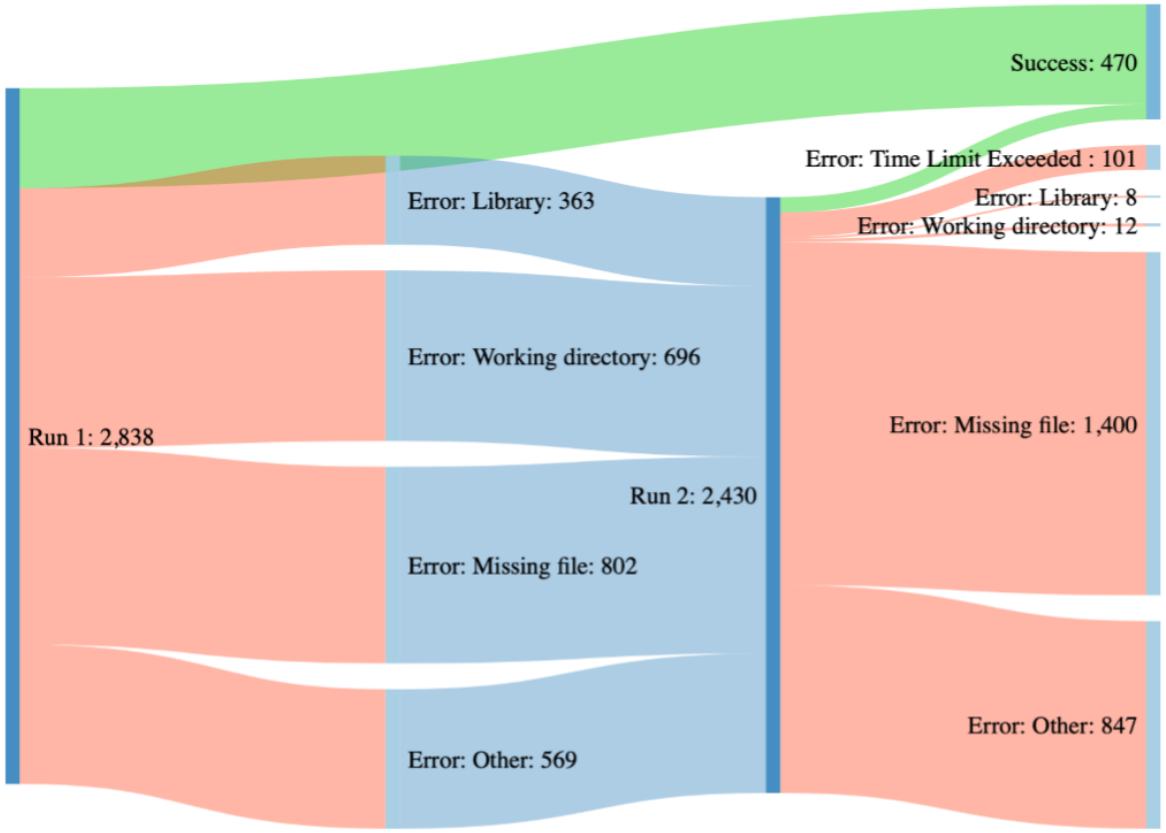
Data

R Code



But, it's complicated.

85.6% of archived R-based studies are not easily re-executable



From: Coding BetteR: Assessing and Improving the Reproducibility of R-Based Research with containR by Chris Chen

Can we do better?

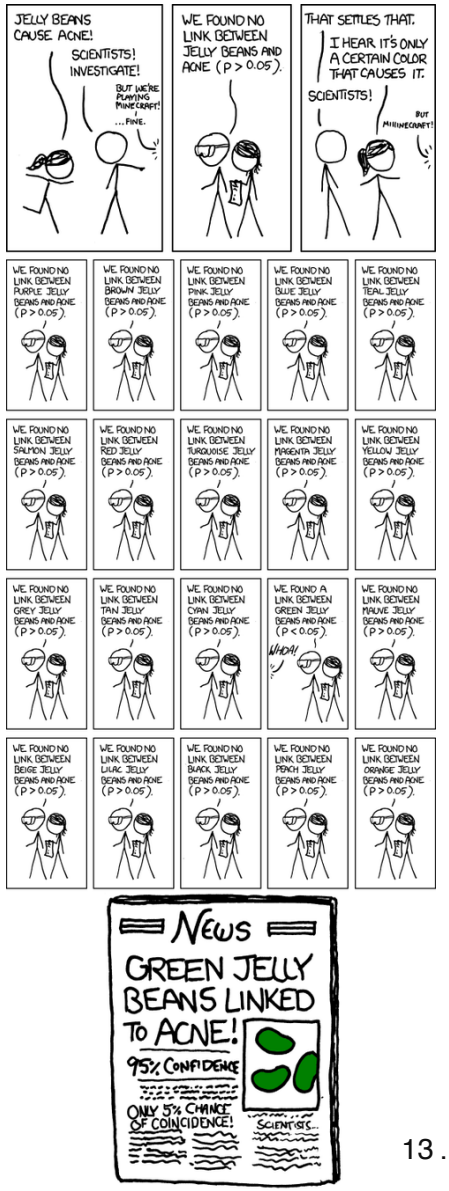
Current Dataverse projects to improve Computational Reproducibility

- Include reproducibility as part of the peer-review workflow in journal dataverses
- Integrate with reproducibility and computational web-based tools (e.g., Code Ocean) to facilitate code execution
- Publish a capsule (container with data and code) verified for reproducibility
- When possible, automate code execution upon depositing the data and code

The Dataverse® Project

Computational Reproducibility is not sufficient

- Be mindful of publication bias and specification searching (statistical power, p-values, effect size)
- Include a detailed description of the analysis:
  - all methods, instruments, materials, procedures;
  - decisions for the exclusion or inclusion of data;
  - the analytic decisions and when these decisions were;
  - a discussion of the expected constraints on generality
  - reporting of precision or statistical power; and
  - discussion of the uncertainty of the measurements, results, and inferences;
- Conduct meta-analysis for replicability



NASEM, 2019, Reproducibility and Replicability in Science

& Christinsen, Freese, Miguel, 2019, Transparent and Reproducible Social Science Research

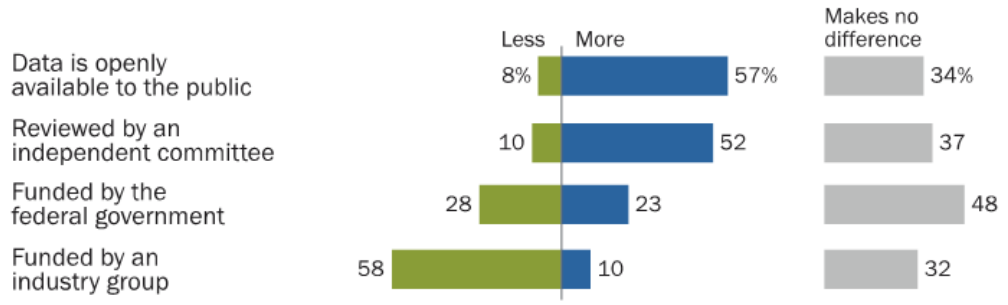
"Americans say open access to data (+ code) and independent review inspire more trust in research findings"

Trust and Mistrust of American Views on Scientific Experts, Pew Research Center, August 2, 2019

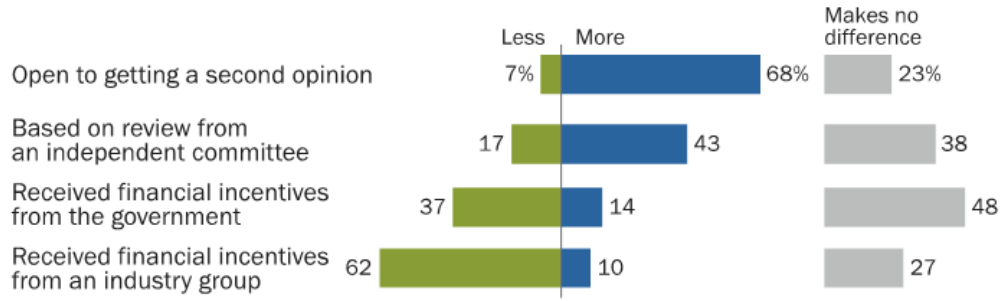
*(+ code) not in original statement

Majority of Americans say they are more apt to trust research when the data is openly available

% of U.S. adults who say when they hear each of the following, they trust scientific research findings ...



% of U.S. adults who say when they hear each of the following, they trust a science practitioner's recommendation ...



Note: Respondents who did not give an answer are not shown.
Source: Survey conducted Jan. 7-21, 2019.
"Trust and Mistrust in Americans' Views of Scientific Experts"

PEW RESEARCH CENTER