Data Sets

General Data Sets

- UCSF Industry Documents
  Massive digital archive of documents created by industries which influence public health, such as the tobacco, chemical, drug and fossil fuel industries.
- LuminDatabase
  A searchable database that collects and analyzes legal complaints and requests for removal of online materials (DMCA take-downs), helping Internet users to know their rights and understand the law.
- National Center for Biotechnology Information
  Allows you to search 39 different scientific databases such as Pubmed, SRA, OMIN, MedGen and more from a single page.
- Afrobarometer
  Large archive of sociological surveys conducted in African countries over the last ~20 years.
- Arabbarometer
  Large archive of sociological surveys conducted in the Arab countries of Africa and the Middle East from 2007 to 2021.
- CensoredPlanet
  A censorship measurement platform that collects data using multiple remote measurement techniques in more than 200 countries. Provides reports and offers their raw data sets which are available for download.
- DomainsProject
  Massive dataset of over 600 million domains. Total size is ~16 GB.
- Face Recognition Datasets
  A large collection of face datasets for training facial recognition systems and other things of that nature.
- Kaggle
  Offers over 50,000 public datasets for all kinds of various things.
- Common Crawl
  An open repository of web crawl data that can be accessed and analyzed by anyone.
- OCCRP Catalogue of Research Databases
  A massive collection of public data sources compiled by OCCRP researchers that are the most useful for investigative reporting.
- CORE Research Paper Database
  CORE currently contains 207,255,818 searchable open access articles and research papers collected from 10,286 data providers around the world, which you can search using keywords.
- Public Intelligence
  An international, collaborative research project aimed at aggregating the collective work of independent researchers around the globe who wish to defend the public's right to access information.
- GIJC21 Resources
  A large list of investigative resources from the GIJN 2021 conference. There is some very useful content in here.
- Google Dataset Search
  A search engine for all kinds of data sets provided by Google.

- MartinDale
  Search for attorneys and related articles.
- HuggingFace
  Offers models based on transformers for PyTorch and TensorFlow 2.0. There are thousands of pre-trained models to perform tasks such as text classification, extraction, question answering, and more.
- Information Operations Archive
  An archive of publicly available and attributed data from known online information operations. The archive currently consists of over 10 million messages from Russian and Iranian state-sponsored influence operations on Twitter and Reddit, and will be updated on an ongoing basis.
- Bloopbase Searchable Data Dumps
  A collection of searchable data dumps. Includes law enforcement, government, extremist, conventional, commercial, Slack user and exploit databases.
- AI Incident Database
  A collection of AI deployment harms or near harms across all disciplines, geographies, and use cases.

Government Data Sets

- FBI Vault
  The US FOIA library. Contains over 6,700 scanned FOIA documents.
- CIA Reading Room
  Search and view documents released through the FOIA and other CIA release programs.
- US Department of State Records
  Search through 221,373 documents reviewed and released to the public.
- US National Archives
  An independent agency of the United States government charged with the preservation and documentation of government and historical records. It is also tasked with increasing public access to those documents which make up the National Archive.
- US Library of Congress
  The Library of Congress is the research library that officially serves the United States Congress and is the national library of the United States. It is the oldest federal cultural institution in the U.S.
- UK National Archives
  One of the worlds largest archives, containing over 11 million historical government and public records. From Domesday Book to modern government files. Includes paper records, digital records, websites, photographs, posters, maps, drawings and paintings.
- Canada Declassified
  A digital repository of government records declassified under the Canadian Access to Information Act. Spans from 1945 through 1991.
- Archives Canada
  A gateway to over 800 archival repositories across Canada.
- Australian National Archives Search
  A tool to search the national archives of Australia. Includes an advanced search function.
- ICO Search
  The Information Commissioner's Office (ICO) upholds information rights in the public interest, promoting openness by public bodies and data privacy for individuals. ICO is an executive non-departmental public body, sponsored by the Department for Digital, Culture, Media and Sport.
- David McKie Open Data Portals
  This is a great collection of Canadian open data portals, both federal and provincial. This site

also provides some other useful non-Canadian data sets.
- SpatialHub Scotland Datasets
  A collection of various datasets for Scotland.
- Netronline Public Records
  An online directory and portal to those Tax Assessors', Treasurers' and Recorders' offices that have developed websites for the retrieval of available public records for the U.S.
- BlackBookOnline
  A large database and search tool for locating U.S. public records. Find everything from parking tickets to property records.

Leaked Data Sets

- WikiLeaks
  An international non-profit organization that publishes leaks and classified media from governments, companies and organizations alike. All data is provided by anonymous sources. #FreeAssange.
- CryptoMe
  Publishes documents that are prohibited by governments worldwide. Particularly material on freedom of expression, privacy, cryptology, dual-use technologies, national security, intelligence, secret governance, open, secret and/or classified documents.
- ICIJ Offshore Leaks
  Data from more than 785,000 offshore companies, foundations and trusts from the Panama Papers, Offshore Leaks, Bahamas Leaks, and the Paradise Papers.
- ICIJ Luxembourg Leaks
  Also known as the "LuxLeaks", is a collection of over 350 documents about Luxembourg's tax rulings set up by PricewaterhouseCoopers from 2002 to 2010 to the benefits of its clients.
- Distributed Denial of Secrets
  A journalist 501(c)(3) non-profit devoted to enabling the free transmission of data in the public interest. Aims to avoid political, corporate or personal leanings, to act as a beacon of available information.