# Data Sets and Resources
# Characterizing
# Large Numbers of Observations
# And
# Rare Events

This document is provided as a public service by Lone Star Analysis. It is not copyrighted, and users can feel free to modify, copy and re-use as they see fit. Additions to the list are welcomed. Please make suggestions by email to info@Lone-Star.com.

This documents summarizes datasets referenced by Wikipedia, and documents from other sources which may be of use. Some of the links are to unstructured feeds (e.g., Digg) and some to tools, which have may not be data sets per se, but are associated with data sets, data structures, and data bases. In addition, some resources and references related to understanding information across many orders of magnitude and/or rare events are provided.

Multidisciplinary applications for models based on data of this type include uses in:

- Corporate Strategy
- Government Policy
- Risk Assessments
- Investment Strategy

Organizations seeking this kind of understanding include:

- Corporations with large fixed capital projects at risk (e.g., Energy Exploration)
- Organizations seeking to understand risks (e.g., COSO based risk assessments in compliance with Sarbanes-Oxley requirements)
- Government Agencies
- NGO's
- Investment Banks

This resource set contains three lists.

- **List 1:** Summary of the Wikipedia "Rare Event" data resources (some are large observation sets, including both rare and non-rare events); starts on Page 2.
- **List 2:** General Purpose Large Data Sets and Rare Event Sources and Resources; starts on page 4.
- **List 3:** List 3: Narrower but Useful Data Sets and Resources; Starts on page 5. Some of these are not "narrow" but are somewhat narrower than List 2.

| List 1: Wikipedia Rare Event Data Sets (https://en.wikipedia.org/wiki/Rare_events) | | |
|---|---|---|
| **Advanced National Seismic System (ANSS) Comprehensive Earthquake Catalog (ComCat)** | http://earthquake.usgs.gov/earthquakes/search/ | The ANSS Comprehensive Catalog (ComCat) contains earthquake source parameters (e.g. hypocenters, magnitudes, phase picks and amplitudes) and other products (e.g. moment tensor solutions, macroseismic information, tectonic summaries, maps) produced by contributing seismic networks. |
| **Armed Conflict Database** | https://acd.iiss.org/ | The Armed Conflict Database (ACD) monitors armed conflicts worldwide, focusing on political, military and humanitarian trends in current conflicts, whether they are local rebellions, long-term insurgencies, civil wars or inter-state conflicts. In addition to the comprehensive historical background for each conflict, the weekly timelines and the monthly updates, the statistics, data and reports in the ACD date back to 1997. |
| **Armed Conflict Location & Event Data Project** | http://www.acleddata.com/data/ | The Armed Conflict data set covers events occurring in Africa from 1997 to present. This data set includes the event date, longitude, latitude, and fatality magnitude scale. |
| **Aviation Safety Database** | http://aviation-safety.net/database/ | The Aviation Safety Database covers aviation safety incidents around the world. Every incident reports the location of the incident, the departing and arriving airports, number of fatalities and type of Airplane involved in the incident. |
| **Dartmouth Flood Observatory** | http://floodobservatory.colorado.edu/ | Dartmouth Flood Observatory uses “Space-based Measurement and Modeling of Surface Water” to track floods and uses news reporting to validate the results. This data set includes the country, start date, end date, affected square km, and cause of the flood. Additionally, this data set includes many magnitude scales, such as: dead, displaced, severity, damage, and flood magnitude. |
| **Database of Radiological Incidents and Related Events** | http://www.johnstonsarchive.net/nuclear/radevents/ | The Database of Radiological Incidents and Related Events covers events that resulted in acute radiation exposures to humans sufficient enough to cause casualties. The database includes the date, location, number of deaths, number of injuries and highest radiation dose recorded. |
| **Dow Jones Averages** | http://www.djaverages.com/?go=industrial-index-data | Dow Jones Averages includes data and information on some of the worlds most renowned and widely-cited market indexes. Here you'll find rich historical data, robust analytical tools and exclusive educational content on the Dow Jones Industrial Average and a host of related indices. |

| | | |
|---|---|---|
| **FluView** | http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html | FluView is produced by the U.S. Centers for Disease Control (CDC) and provides weekly influenza surveillance information in the United States by census area and includes the number of people tested and number of positive cases. |
| **FAOSTAT** (Famine) | http://faostat.fao.org/ | The FAOSTAT data set was developed by the Statistics Division of the Food and Agricultural Organization of the United Nations (FAO). It is an active, global data set that covers famine events from 1990-2013. |
| **Global Health Atlas** | http://apps.who.int/globalatlas/default.asp | The Global Health Atlas contains data on four communicable diseases: Cholera, Influenza, Polio, and Yellow Fever. It is an active, global data set that covers number of cases and fatalities due to these infectious diseases. |
| **Global Volcanism Program** | http://www.volcano.si.edu/search_eruption.cfm | "Volcanoes of the World is a database describing the physical characteristics of volcanoes and their eruptions." The data contain a start date, end date, volcano name (which can be used to look up the location) and VEI magnitude scale. |
| **International Disaster Database** | http://www.emdat.be/ | EM-DAT contains essential core data on the occurrence and effects of over 18,000 mass disasters in the world from 1900 to present. The database is compiled from various sources, including UN agencies, non-governmental organizations, insurance companies, research institutes and press agencies. |
| **Major Episodes of Political Violence** | http://www.systemicpeace.org/inscrdata.html | The Major Episodes of Political Violence data set is part of a larger armed conflict database produced by the Center for Systemic Peace. Political Violence data include annual, cross-national, time-series data on interstate, societal, and communal warfare magnitude scores (independence, interstate, ethnic, and civil; violence and warfare) for all countries. |
| **Militarized Interstate Disputes** | http://www.correlatesofwar.org/COW2%20Data/MIDs/MID40.html | The Militarized Interstate Disputes (MID) data set "provides information about conflicts in which one or more states threaten, display, or use force against one or more other states between 1816 and 2010." |
| **NOAA Natural Hazards** | http://www.ngdc.noaa.gov/hazard/ | The Natural Hazards dataset is part of the National Geophysical Data Center run by the U.S. National Oceanic and Atmospheric Administration (NOAA). The National Geophysical Data Center archives and assimilates tsunami, earthquake and volcano data to support research, planning, response and mitigation. Long-term data, including photographs, can be used to establish the history of natural hazard occurrences and help mitigate against future events. |
| **Political Instability Task Force (PITF) State Failure Problem Set, 1955-2013** | http://www.systemicpeace.org/inscrdata.htm | The Political Instability Task Force (PITF), State Failure Problem Set is part of a larger armed conflict database produced by the Center for Systemic Peace from open source data. Data in PITF are available on various subsets: ethnic war, revolutionary war, adverse regime change, and genocide/politicide. |

| | | |
|---|---|---|
| **Rand Database of Worldwide Terrorism Incidents** | https://www.rand.org/nsrd/projects/terrorism-incidents.html | The Rand Database of Worldwide Terrorism Incidents data set covers terrorism incidents worldwide from 1968 through 2009 but is not currently active. The data set includes a date, location (city, country), perpetrator, detailed description, and number of injuries and fatalities. |
| **U.S. National Flood Insurance Program** | http://www.fema.gov/policy-claim-statistics-flood-insurance/policy-claim-statistics-flood-insurance/policy-claim-13 | The U.S. National Flood Insurance Program data set contains a data table detailing flooding events with 1,500 or more paid losses from 1978 to the current month and year. The table includes the name and year of the event, the number of paid losses, the total amount paid and the average payment per loss. |
| **List 2: General Purpose Large Data Sets and Rare Event Sources and Resources** | | |
| **World Bank Datasets** | http://api.worldbank.org/v2/datacatalog/downloadfile | The World Bank produces more than 1000 data products. This link downloads a spreadsheet catalog of available data. The next citation is an example. |
| **World Bank Development Indicators** | http://data.worldbank.org/products/data-books/WDI-2007<br><br>http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators | WDI includes more than 900 indicators in more than 80 tables.<br><br>The two links shown are for the current, and for comparison, a historical version<br><br>2007 Data are shown for 152 economies with populations of more than 1 million, as well as for Taiwan, China, in selected tables. |
| **A New Heuristic Measure of Fragility and Tail Risks: Application to Stress Testing** | http://www.imf.org/external/pubs/ft/wp/2012/wp12216.pdf | IMF working paper by Nassim N. Taleb, Elie Canetti, Tidiane Kinda, Elena Loukoianova, and Christian Schmieder |
| **Underestimating extreme events in power-law behavior due to machine-dependent cutoffs** | http://homes.soic.indiana.edu/filiradi/Mypapers/PhysRevE.90.050801.pdf | A paper published in Physical Review. Implementing Power Law Distributions (e.g., Ziph's law) is a preferred method for some events with both probabilities and impacts spanning orders of magnitude. Unfortunately, these methods are vulnerable to limitations in sampling methods and other computational implementations. |

| List 3: Narrower But Useful Data Sets and Resources | | |
|---|---|---|
| **Big, Allied and Dangerous (BAAD)** | http://www.start.umd.edu/news/database-spotlight-big-allied-and-dangerous-baad | Terror group data set at SUNY Albany – Dr. Asal and Dr. Rethemeyer compiled the Monterey WMD Terrorism Database and the Tactical Terrorism Dataset into BAAD. |
| **CIA World Factbook** | https://www.cia.gov/library/publications/the-world-factbook/ | The World Factbook, produced for US policymakers and coordinated throughout the US Intelligence Community; facts on every country, dependency, and geographic entity in the world. |
| **CIA's Intellipedia** | Secure Logon Required | Intellipedia is an online wiki system for collaborative data sharing used by the United States Intelligence Community (IC). |
| **Digg** | http://digg.com/ | Digg is a news aggregator from which several social computing datasets have been constructed.<br><br>Digg can be useful for unstructured data feeds |
| **Defence Science and Technology Organization (DSTO)** | http://www.dsto.defence.gov.au/discover-dsto#sthash.jzuG3btd.dpuf | DSTO is the Australian Government's lead agency responsible for applying science and technology to safeguard Australia.<br><br>DSTO provides one of the largest open source bodies of text for defense related analysis. |
| **Gallup World Poll** | http://www.gallup.com/services/170945/world-poll.aspx | Gallup conducts nationally representative surveys in over 160 countries and over 140 languages, covering the emerging and developed world. |
| **US Energy Information Agency** | http://www.eia.gov/ | The EIA "collects, analyzes, and disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment." The EIA produces a wide range of energy related data sets (coal, oil, gas, fuels, consumption…) |
| **Global Financial Data** | https://www.globalfinancialdata.com | Records from 1209 AD for some financial measures, and wide range of measures from 1500 AD forward. |
| **Global Terrorism Database (GTD)** | http://www.start.umd.edu/gtd/ | National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland – "The Global Terrorism Database (GTD) is an open-source database including information on terrorist events around the world from 1970 through 2014 (with annual updates planned for the future). |

| | | |
|---|---|---|
| **Global U.S. Troop Deployment** | http://www.heritage.org/research/reports/2004/10/global-us-troop-deployment-1950-2003 | The Heritage Foundation – A study detailing U.S. troop deployment locations and levels around the world from 1950-2003. A download is available. This data is useful for any study of the military force, deployments and distributions. |
| **Historicalstatistics.org** | http://www.historicalstatistics.org/ | A portal for historical statistics; macroeconomic data on Sweden and other European nations in the 19th and 20th centuries: GDP, inflation, employment, interest rates, exchange rates, population, money supply, capital stocks, worked hours, wages, profit rates and business cycle indicators. |
| **Human Relations Area Files (HRAF)** | http://hraf.yale.edu/ | Work of a consortium of over 300 educational and research institutions from 30 countries, consists of over a million pages of information about nearly 400 cultures and is organized according to the Outline of cultural Materials (OCM). The OCM contains a list of more than 700 subject codes or searchable descriptors. Collection such as HRAF are composed of 'unformatted text' or 'unformatted data'. |
| **International Atomic Energy Agency** | https://www.iaea.org/ | The IAEA site contains several factsheets and other publications related to nuclear materials, activities, and policies. |
| **KOF Globalization Scores** | http://globalization.kof.ethz.ch/ | The KOF Index of Globalization measures the three main dimensions of globalization: economic, social, and political |
| **Open Source Center** | https://www.opensource.gov/public/content/login/login.fcc | Large Scale Internet Exploitation (LSIE). The OSC is available to government employees and contractors.<br><br>Some related and similar tools can be found at http://www.casos.cs.cmu.edu/computational_tools/tools.html |
| **Measuring Progress in Conflict Environments (MPICE)** | http://www.usip.org/publications/measuring-progress-conflict-environments-mpice | A project sponsored by the United States Institute of Peace, "The purpose of this project is to establish a system of metrics that will assist in formulating policy and implementing strategic plans to transform conflict and bring stability to war-torn societies. These metrics provide both a baseline assessment tool for policymakers to diagnose potential obstacles to stabilization prior to an intervention and an instrument for practitioners to track progress from the point of intervention through stabilization and ultimately to a self-sustaining peace" |

| | | |
|---|---|---|
| **Monterey WMD Terrorism Database** | http://montrep.miis.edu/databases.html | Monterey Institute for International Studies - "The Monterey Terrorism Research and Education Program (MonTREP) has three main goals: to research topics related to terrorism studies and extremist movements; to educate students about the history and trends of these groups; and to generate policy recommendations that can guide professionals in counterterrorism and related fields." |
| **New Horizons Report** | http://www.nmc.org/nmc-horizon/ | The NMC Horizon Project charts the landscape of emerging technologies for teaching, learning, and creative inquiry. Downloads characterizing global higher education, K-12 education, libraries, and museums. The regional- and sector-focused NMC Technology Outlook series has examined STEM+ education, community colleges. |
| **Profiler Plus** | http://socialscience.net/tech/profilerplus.aspx | "Profiler Plus is a general purpose text analytics (Natural Language Processing) system that has been refined and extended during more than 10 years of development and use." |
| **Project Civil Strife** | http://www.planetinform.com/Analytic/default.aspx | Uses Text Analysis By Augmented Replacement Instruction (TABARI) BBC Monitor<br>Moscow Times, Moscow News, Russia & CIS Statistics Online Database compiled by Planet Inform |
| **Quality of Government (QoG)** | http://qog.pol.gu.se/data | Five datasets supporting research on the causes, consequences and nature of Good Governance and the Quality of Government (QoG) - that is, trustworthy, reliable, impartial, uncorrupted and competent government institutions. |
| **Russia & CIS Statistics Online Database compiled by Planet Inform** | http://www.planetinform.com/pdf/Russia_CIS_Exporters-Importers/Russia_CIS_Exporters-Importers/assets/basic-html/page1.html | 2013 RUSSIA/CIS Exporters-Importers Directory Export-oriented Enterprises of Russia, Belarus, Kazakhstan and Ukraine |
| **Tactical Terrorism Dataset** | http://www.isvg.org/research-focus-violent-groups.php | Compiled by the Institute for the Study of Violent Groups (ISVG) at University of New Haven. |
| **Translingual Automatic Language Exploitation System (TALES)** | http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2011-10/ibm-tales-2/ | IEEE summary of an effort based on work at IBM – Data and Information Ingest and Characterization of foreign language news broadcasts and websites. |

| | | |
|---|---|---|
| **USGS Mineral Commodity Summaries** | http://minerals.usgs.gov/minerals/pubs/mcs/ | One Example of a USGS data set (see the next entry for their catalog). This data compliments the EIA data sets for energy products. It provides data for over 90 individual minerals and materials. |
| **USGS Survey Data Catalog** | http://data.usgs.gov/datacatalog/#q=*%3A* | The U.S. Geological Survey has thousands of data compilations, some of them incorporating data from more than a million site surveys. This catalog is extensive and searchable. |