Data Sets (1/2)

These are some public data sets that might prove interesting for experimentation:

- Wikimedia Data Dumps
- DBPedia
- ArXiv Articles - Bulk Data Access
- Million Song Dataset
- GeoNames Data Dump
- IP Geolocation Database
- NHTSA Office of Defects Investigation data
- Federal Election Commission Downloadable Data
- BioID Face Database
- 100,000+ Official Crossword Words
- 350,000+ Simple English Words
- Google Ngram Corpora
- Amex Exchange Daily Data: 1970-2010
- NASDAQ Exchange Daily Data: 1970-2010
- NYSE Exchange Daily Data: 1970-2010
- SEC EDGAR Filings Data
- IRS Tax data

Data Sets (2/2)

These are some more public data sets that might prove interesting for experimentation:

- Amazon Product Review and other Sentiment Data
- US Census 2000 Tract Level Planning Database
- US Census TIGER Geographic Data
- US Census Frequently Occurring Names
- Movie Review Data from Cornell
- GroupLens / MovieLens Data Download
- Music Recommendation Datasets from Last.FM
- MusicBrainz Database Download
- MetaFilter Infodump
- Fair Market Rents Data from HUD
- Climate Data from Univ. of Delaware
- NOAA Daily Climate Data
- USGS Earthquake Data
- Supreme Court Database
- Enron e-mail Corpus
- Common Crawl data set
- Reliability data for 41,000 hard drives

Data Collections (1/2)

These are some sites with multiple datasets:

- AdMob (their "Easy API" can return consistent XML from an HTTP POST request)
- U.S. Census Bureau
- Freebase
- Amazon Web Services
- InfoChimps
- Machine Learning Repository at University of California – Irvine
- Metacatalog of Open Data
- Open Knowledge Foundation
- Federal Reserve Board Statistics Data
- Hilary Mason's curated list of data sets
- US Census 2000 Data Sets
- Stanford Large Network Dataset Collection
- KONECT (the Koblenz Network Collection)
- Carnegie Mellon StatLib Archive
- National Space Science Data Center

Data Collections (2/2)

These are some additional sites with multiple datasets:

- Department of Energy Data Explorer
- US Housing and Urban Development Datasets
- US Government-related Data Sets on GitHub
- A huge collection of data collection links
- Many datasets linked from this Quora answer
- Combined Online Information System - COINS (Finance data for UK Government)
- UK National Health Service Data Sets
- Catalog of social services data sets from Washington University in St. Louis
- http://academictorrents.com (more than 200GB of data via BitTorrent)
- OpenFDA (data APIs for the U.S. Food and Drug Administration)
- City of Chicago Data Portal
- Data Streams from SparkFun (e.g., sensor data from hobbyists)