Domain Specific Data Sets

Description

This page will help you find a domain-specific data set. While you do not need to complete your final project on a data set that is specific to your domain, it may be a good experience to do so! If you have not read the final project description please do this first, you can find the [description here](https://elearning.mines.edu/courses/52392/pages/semester-project-introduction).  You may also find your own data if you would like and if you find a cool one let us know!

Using Kaggle

Kaggle is a great source for data science and machine learning. With tons of data sets in many different domains! In order to download and use the data sets, you must set up a Kaggle account using the Kaggle [register/sign-in pageLinks to an external site.](https://www.kaggle.com/account/login?ReturnUrl=%2Faccount%2Flogoff).  Once your account is set up start exploring some datasets!

## Domains

### **General**

#### **Used in Class:**

**Boston Housing**

The Boston Housing Data Set is information collected by the U.S. Census service in the area of Boston Massachusetts. You can find the data set here: [Boston Housing Data Set Links to an external site.](https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html)

**Colorado Data**

This website includes various data sets from around Colorado. You can find one that interests you an try it out! The website can be found here: [https://gocode.colorado.gov/Links to an external site.](https://gocode.colorado.gov/)

**WHO Data**

Here you can find some data sets provided by the World Health Organization. This could be a fun project to deal with something like COVID-19 predictions. The data sets can be found here: [WHO Data SetsLinks to an external site.](https://www.who.int/data)

### **Mathematics**

### **Business/Econ**

**BigMart**

The BigMart Sales Data is sales data from 10 different stores and 1559 different products. Use this to build a prediction model to determine the sales of each of the 10 stores. You can find the data set here: [Bigmart Sales DataLinks to an external site.](https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data" \t "_blank)

**World Bank**

The World Bank Development Indicators data set contains indications of development in different countries around the world. You can find the data set here: [World Bank Development IndicatorsLinks to an external site.](https://www.kaggle.com/worldbank/world-development-indicators)

**NYC Property Sales**

This data set contains information on property sales in NYC over a 12-month span. Use the data to see if you can predict the sale values in the future! You can find the data set here: [NYC Property SalesLinks to an external site.](https://www.kaggle.com/new-york-city/nyc-property-sales).

**Fraud Detection**

This data set contains generated mobile money transactions, some are fraudulent and some are not. See if you can classify which are fraudulent and which are not based on the other columns! You can find the data set here: [Fraudulent ActivityLinks to an external site.](https://www.kaggle.com/ntnu-testimon/paysim1).

**Emergency Management Data Sources**

[https://catalog.data.gov/dataset?groups=disasters#topic=disasters\_navigationLinks to an external site.](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fcatalog.data.gov%2Fdataset%3Fgroups%3Ddisasters%23topic%3Ddisasters_navigation&data=04%7C01%7Cmdcox%40mymail.mines.edu%7C2f81d1e34ba5472dff9b08d8be37f810%7C997209e009b346239a4d76afa44a675c%7C0%7C0%7C637468494871298703%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=dNNYIErBviRl0L3xgPgCM3HGwOqKXG%2FOiagho%2FmrOE4%3D&reserved=0)

[https://www.fema.gov/about/openfema/data-setsLinks to an external site.](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.fema.gov%2Fabout%2Fopenfema%2Fdata-sets&data=04%7C01%7Cmdcox%40mymail.mines.edu%7C2f81d1e34ba5472dff9b08d8be37f810%7C997209e009b346239a4d76afa44a675c%7C0%7C0%7C637468494871308704%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=xGFQ83qgJsZzS6EIR%2BqTHmPnWtVa3YdL33vDkhFI6CA%3D&reserved=0)

[https://data.world/datasets/disastersLinks to an external site.](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdata.world%2Fdatasets%2Fdisasters&data=04%7C01%7Cmdcox%40mymail.mines.edu%7C2f81d1e34ba5472dff9b08d8be37f810%7C997209e009b346239a4d76afa44a675c%7C0%7C0%7C637468494871318696%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=7xoordwH%2FaEWrxZmKPSwGgwb9ScW85dnEM529Cq2nHw%3D&reserved=0)

### **Geology**

**Earthquakes**

This data from the National Earthquake Information Center includes information about each reported earthquake 5.5 magnitudes or higher since 1965. The data set can be found here: [EarthquakesLinks to an external site.](https://www.kaggle.com/usgs/earthquake-database" \t "_blank)

**Volcanoes**

This data set contains information about volcanoes that have been active for the last 10,000 years. The data set can be found here: [Volcanic EruptionsLinks to an external site.](https://www.kaggle.com/smithsonian/volcanic-eruptions)

**Geo Thermal Data Repository**

Note: This data may be messy!

https://gdr.openei.org/submissions/all

**Other sources provided by Mines:**

* Wyoming well logs, tops, etc [https://sites.google.com/a/wyo.gov/oil-and-gas/prb-studyLinks to an external site.](https://sites.google.com/a/wyo.gov/oil-and-gas/prb-study)
* DZ data from Magallanes Basin [https://doi.org/10.1130/B31757.1Links to an external site.](https://doi.org/10.1130/B31757.1) - see "Supplemental Information 2" down at the bottom. 2 tabs in that xls are high throughput and low throughput ages from Tres Pasos that could be compared. Warning, messy data!
* USGS core data [https://my.usgs.gov/crcwc/Links to an external site.](https://my.usgs.gov/crcwc/)
  + also the full csv for cores (similar dataset can be created for cuttings) [https://docs.google.com/spreadsheets/d/1fX8ZyF2Pmx7apcBftvWVVOUlPTCRl014cIASIowAVNE/edit?usp=sharingLinks to an external site.](https://docs.google.com/spreadsheets/d/1fX8ZyF2Pmx7apcBftvWVVOUlPTCRl014cIASIowAVNE/edit?usp=sharing)
* LOTS of stuff in here [https://github.com/yohanesnuwara/open-geoscience-repositoryLinks to an external site.](https://github.com/yohanesnuwara/open-geoscience-repository)
* LOTS of stuff in here too [https://dataunderground.orgLinks to an external site.](https://dataunderground.org/)
  + e.g., - macrostrat api [https://dataunderground.org/dataset/macrostrat-geologic-maps/resource/e412b05e-a45e-4809-82ab-5095ac383ddcLinks to an external site.](https://dataunderground.org/dataset/macrostrat-geologic-maps/resource/e412b05e-a45e-4809-82ab-5095ac383ddc)
* LOTS of GOM data here (wells, production, etc.) [https://www.data.boem.gov/Links to an external site.](https://www.data.boem.gov/)
* [https://hackmd.io/iAyNR\_TOSguKnlnkrY2AuwLinks to an external site.](https://hackmd.io/iAyNR_TOSguKnlnkrY2Auw)
* Netherlands data [https://nlog.nl/en/dataLinks to an external site.](https://nlog.nl/en/data)
* earthquakes [https://earthquake.usgs.gov/earthquakes/search/Links to an external site.](https://earthquake.usgs.gov/earthquakes/search/)
* volcanoes [https://volcano.si.edu/list\_volcano\_holocene.cfmLinks to an external site.](https://volcano.si.edu/list_volcano_holocene.cfm)
* water and snow
  + Water For the Nation [https://waterdata.usgs.gov/nwis/rtLinks to an external site.](https://waterdata.usgs.gov/nwis/rt) and API [https://help.waterdata.usgs.gov/faq/automated-retrievalsLinks to an external site.](https://help.waterdata.usgs.gov/faq/automated-retrievals)
  + SNOTEL [https://www.wcc.nrcs.usda.gov/snow/Links to an external site.](https://www.wcc.nrcs.usda.gov/snow/)
* Earth Engine Data on deltas, waves, and tides [https://jhnienhuis.users.earthengine.appLinks to an external site.](https://jhnienhuis.users.earthengine.app/)
* oxygen isotope data [https://doi.pangaea.de/10.1594/PANGAEA.917503Links to an external site.](https://doi.pangaea.de/10.1594/PANGAEA.917503) and the associated paper [https://science.sciencemag.org/content/369/6509/1383Links to an external site.](https://science.sciencemag.org/content/369/6509/1383)
* geothermal data 1 [http://repository.stategeothermaldata.org/repository/browse/Links to an external site.](http://repository.stategeothermaldata.org/repository/browse/)
* geothermal data 2 [http://geothermal.smu.edu/gtda/Links to an external site.](http://geothermal.smu.edu/gtda/)
* paleobiology [https://paleobiodb.org/#/Links to an external site.](https://paleobiodb.org/#/)
* space / exoplanets [https://exoplanetarchive.ipac.caltech.edu/docs/data.htmlLinks to an external site.](https://exoplanetarchive.ipac.caltech.edu/docs/data.html)
* Kaggle
  + USGS earthquakes [https://www.kaggle.com/usgs/earthquake-databaseLinks to an external site.](https://www.kaggle.com/usgs/earthquake-database)
  + Volcanic eruptions [https://www.kaggle.com/smithsonian/volcanic-eruptionsLinks to an external site.](https://www.kaggle.com/smithsonian/volcanic-eruptions)
  + Landslides after rain [https://www.kaggle.com/nasa/landslide-eventsLinks to an external site.](https://www.kaggle.com/nasa/landslide-events)
    - parsing dates from these three datasets [https://www.kaggle.com/rtatman/data-cleaning-challenge-parsing-datesLinks to an external site.](https://www.kaggle.com/rtatman/data-cleaning-challenge-parsing-dates)

### **Physics**

**Particle Tracking**

This data set contains particle collision data collected using the Large Hadron Collider.  The goal is to properly group the measurements for each event into sets that belong to the same particle. The data set can be found here: [Particle TrackingLinks to an external site.](https://www.kaggle.com/c/trackml-particle-identification/data)

**Sun Spots**

This data set contains information on sun sports over the last few years and provides the date of occurrence and the monthly mean total sunspots. The data set can be found here:  [Sun SpotsLinks to an external site.](https://www.kaggle.com/robervalt/sunspots)

**Solar Flares**

This data set contains information on Solar Flares and can be used to predict future solar flares! The data set can be found here: [Solar FlaresLinks to an external site.](https://www.kaggle.com/khsamaha/solar-flares-rhessi)

**Detection of 'Elves'**

Information about the data and data use:

[https://eos.org/science-updates/catching-elves-in-argentinaLinks to an external site.](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Feos.org%2Fscience-updates%2Fcatching-elves-in-argentina&data=04%7C01%7Cmdcox%40mymail.mines.edu%7C0ad562fddd6f40c39df608d8be37e559%7C997209e009b346239a4d76afa44a675c%7C0%7C0%7C637468494556751427%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=97b3FOh8kUSl5skekcTjXtS9P5mjrzdQfb%2BajISgXz8%3D&reserved=0)

Data set:

[https://agupubs.onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1029%2F2019EA000582&file=ess2484-sup-0002-Data\_Set\_SI-S01.txtLinks to an external site.](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fagupubs.onlinelibrary.wiley.com%2Faction%2FdownloadSupplement%3Fdoi%3D10.1029%252F2019EA000582%26file%3Dess2484-sup-0002-Data_Set_SI-S01.txt&data=04%7C01%7Cmdcox%40mymail.mines.edu%7C0ad562fddd6f40c39df608d8be37e559%7C997209e009b346239a4d76afa44a675c%7C0%7C0%7C637468494556741428%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=YqRey2J8Ap%2BXVzQPXwBIJOxARJkoX4JDE81V0VJ9UUM%3D&reserved=0)

### **Medical**

**Patient Falling Classification**

This data provided by Özdemir, Ahmet Turan, and Billur Barshan. “Detecting Falls with Wearable Sensors Using Machine Learning Techniques.” Sensors (Basel, Switzerland) 14.6 (2014): 10691–10708. PMC. Web. 23 Apr. 2017 shows data about elderly patients who fall, see if you can classify them by the features and predict who is more likely to fall! You can find the data set here: [Patient FallingLinks to an external site.](https://www.kaggle.com/pitasr/falldata).

**Orthopedic Patients**

This data gives you information about patients and if they are normal or abnormal. Use the data provided to predict if a patient is normal or abnormal. You can find the data set here: [Ortho PatientsLinks to an external site.](https://www.kaggle.com/uciml/biomechanical-features-of-orthopedic-patients).

**Heart Disease**

This data provides a lot of patient information and if they are at risk for heart disease. Use the features to determine which are more effective for predicting who is at risk. You can find the data set here: [Heart DiseaseLinks to an external site.](https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility).