

# Data Resources

Jeffrey Leek, Assistant Professor of Biostatistics  
Johns Hopkins Bloomberg School of Public Health

# Open Government Data (U.S.)



The screenshot shows the Data.gov website homepage. The browser address bar displays "www.data.gov". The page features a navigation bar with links: HOME, ABOUT, DATA, METRICS, OPEN GOVERNMENT, BLOGS, and COMMUNITIES. A search bar is located in the top right corner. The main content area includes a large banner for the "AMERICAN COMMUNITY SURVEY 2007-2011" with a photo of diverse people. To the right, a "Latest Datasets" section lists several datasets, including "Combined Federal Campaign, CFC, 2009" and "Gravesite locations of Veterans and...". Below the banner, there are three columns: "DATA AND TOOLS" (showing a map and listing 378,529 raw and geospatial datasets, 1,264 data tools, and 236 citizen-developed data tools), "COMMUNITIES" (with a world map and text about exploring and discussing data), and "OPEN GOVERNMENT DATA" (with an American flag and text about open source code released for the Open Government Platform).

<http://www.data.gov/>

2/11

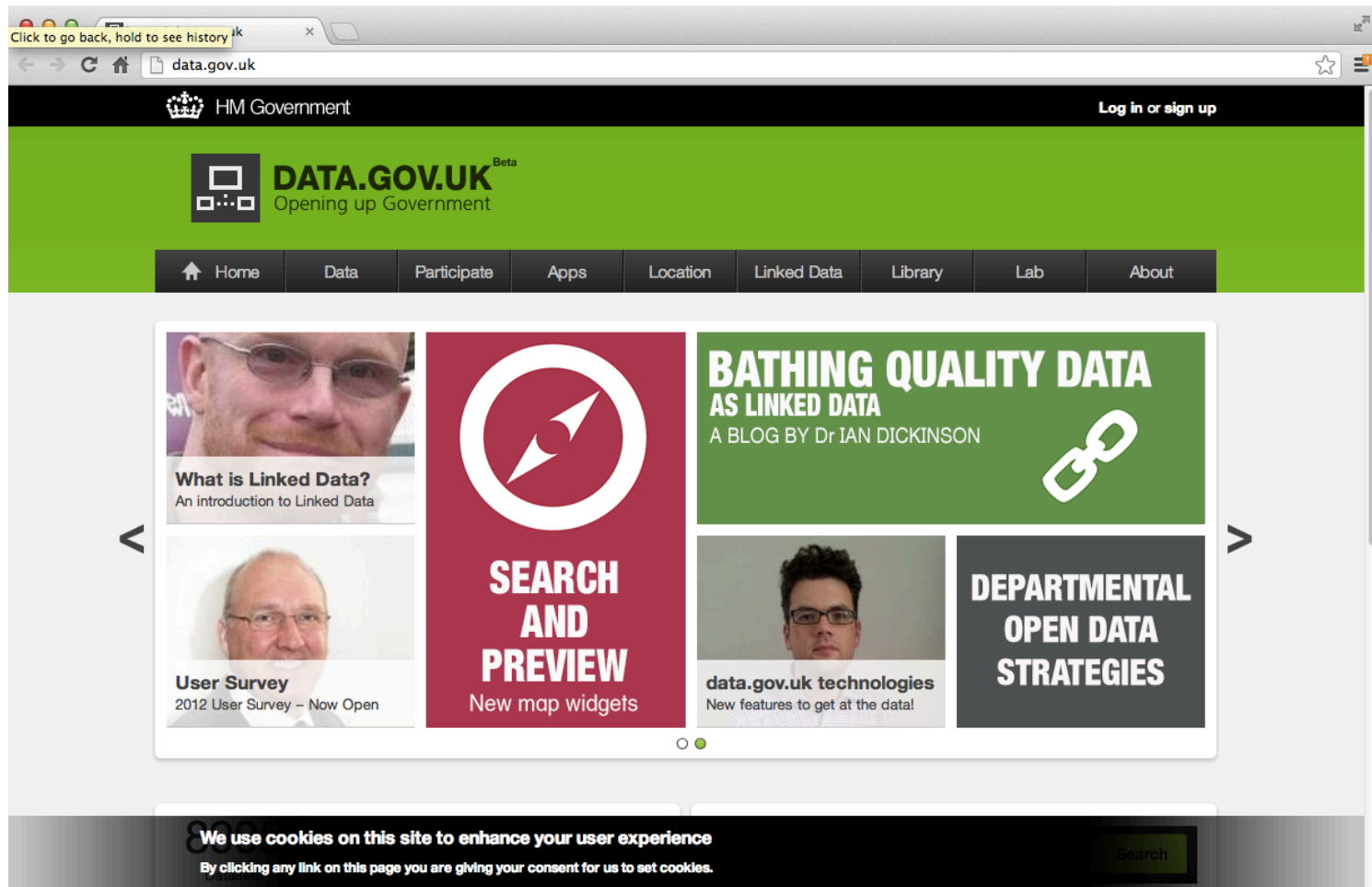
# Open Government Data (France)

The screenshot shows the homepage of data.gouv.fr, the French Open Data platform. The browser address bar displays 'www.data.gouv.fr'. The navigation bar includes links for ACCUEIL, DONNÉES, PRODUCTEURS, ARTICLES, and LICENCE OUVERTE. A 'COMMUNAUTÉ' section with a user icon and links for 'S'IDENTIFIER' and 'A propos' is also present. The main header features the 'data.gouv.fr' logo with the tagline 'INNOVATION TRANSPARENCE OUVERTURE' and a 'BETA' label. Below the header is a search bar with the placeholder text 'Rechercher une donnée' and a 'RECHERCHER' button. A 'RECHERCHE AVANCÉE' link is also visible. The main content area is divided into three columns. The left column features a large image of various data visualizations (charts, graphs, maps) and a text box stating 'LA MISE À DISPOSITION DES DONNÉES PUBLIQUES SUR LA PLATEFORME DATA.GOUV.FR' and 'Etalab, mission interministérielle placée sous l'autorité du Premier ministre, coordonne depuis [...]'. Below this image are social media icons and a link to 'TOUTES LES ACTUS'. The middle column is titled 'SUGGESTION DE RECHERCHE' and 'LES PLUS RECHERCHÉS', listing various data categories such as 'Principales infractions en vigueur', 'Résultats des élections européennes', 'Statistiques du commerce extérieur', 'Financement et dépenses de la sécurité sociale', 'Liste des associations subventionnées', 'Effort financier de l'Etat en faveur des PME', 'Pôles de compétitivité', 'Avis de rappel de produits', 'Adresse des événements culturels français', and 'Adresses diplomatiques et consulaires'. The right column features a 'REJOIGNEZ LA COMMUNAUTÉ' button, a 'REPÈRES' section showing '353 226 jeux de données publiques et plus sur data.gouv.fr', an 'ETALAB' section with the 'etalab' logo and a link to 'Suivez notre actualité sur le blog', and a Twitter follow button for '@etalab' with '4 012 abonnés'.

<http://www.data.gouv.fr/>

3/11

# Open Government Data (UK)



<http://data.gov.uk/>

4/11

# Gapminder

The screenshot shows the Gapminder website's 'Data' section. The browser address bar displays 'www.gapminder.org/data/'. The website header includes the Gapminder logo, navigation links (Blog, FAQ, About, Contact, Donate), and a search bar. A secondary navigation bar contains links for HOME, GAPMINDER WORLD, DATA (selected), VIDEOS, DOWNLOADS, FOR TEACHERS, and LABS. The main content area is titled 'Data in Gapminder World' and includes a breadcrumb trail 'Browse: Home / Data'. Below this, there are links for 'List of indicators', 'About countries & territories', 'Documentation', and 'Data blog'. A paragraph explains that the table lists all indicators and provides instructions on how to use the data. The 'List of indicators in Gapminder World' section features a table with columns for Indicator name, Data provider, Category, Subcategory, and Download/View/Visualize options. A search bar and a 'Show 25 indicators' dropdown are also present.

Browse: [Home](#) / [Data](#)

## Data in Gapminder World

[List of indicators](#) [About countries & territories](#) [Documentation](#) [Data blog](#)

The table below lists all indicators displayed in Gapminder World. Click the name of the indicator or the data provider to access information about the indicator and a link to the data provider.  
Indicators labeled "Various sources" are compiled by Gapminder. They can be reused freely but please attribute Gapminder.

### List of indicators in Gapminder World

Show  indicators Search:

Indicator name	Data provider	Category	Subcategory	Download	View	Visualize
Adults with HIV (% , age 15-49)	Based on UNAIDS	Health	HIV			
Age at 1st marriage (women)	Various sources	Population				
Aged 15+ employment rate (%)	International Labour Organization	Work	Employment rate			
Aged 15+ labour force participation rate (%)	International Labour Organization	Work	Labour force participation			
Aged 15+ unemployment rate (%)	International Labour Organization	Work	Unemployment			
Aged 15-24 employment rate (%)	International Labour Organization	Work	Employment rate			

<http://www.gapminder.org/>

# More open government data (possibly overlapping)

- <http://opengovernmentdata.org/data/catalogues/>
- <http://wiki.civiccommons.org/Initiatives>
- [List of cities/states with open data](#)

# Survey data from the United States

The screenshot shows a web browser window with the address bar displaying "www.asdfree.com". The page content includes a navigation bar with links: "about / faq", "main code repository", "latest releases", "rss", "ajdamico@gmail.com", and "twotutorials". Below this is a section titled "analyze survey data for free" with a sub-link "r-bloggers". The main content area is divided into two columns. The left column, titled "reproducible survey analysis syntax from a website that's easy to type.", lists "AVAILABLE DATA" (including ACS, ARF, BSAFUF, BRFSS, CPS, GSS, HRS, MEPS, NHANES, NHIS, and NSDUH) and "METHODS" (including a guide to installing monetdb with R on Windows). The right column, titled "analyze the health and retirement study (hrs) with r", contains two paragraphs of text about the HRS dataset and its analysis. At the bottom, there is a form to enter an email address for updates and a link to "1992 - 2010 download HRS microdata.R".

analyze survey data for free

about / faq   main code repository   latest releases   rss   ajdamico@gmail.com   twotutorials

r-bloggers

reproducible survey analysis syntax  
from a website that's easy to type.

**AVAILABLE DATA**

- [american community survey \(acs\)](#)
- [area resource file \(arf\)](#)
- [basic stand alone medicare claims public use files \(bsapufs\)](#)
- [behavioral risk factor surveillance system \(brfss\)](#)
- [consumer expenditure survey \(ce\)](#)
- [current population survey \(cps\)](#)
- [general social survey \(gss\)](#)
- [health and retirement study \(hrs\)](#)
- [medical expenditure panel survey \(meps\)](#)
- [national health and nutrition examination survey \(nhanes\)](#)
- [national health interview survey \(nhis\)](#)
- [national study on drug use and health \(nsduh\)](#)

**METHODS**

- [why and how to install monetdb with r on windows](#)

**analyze the health and retirement study (hrs) with r**

the hrs is the one and only longitudinal survey of american seniors. with a panel starting its third decade, the current pool of respondents includes older folks who have been interviewed every two years as far back as 1992. unlike [cross-sectional](#) or shorter panel surveys, respondents keep responding until, well, death do us part. paid for by [the national institute on aging](#) and administered by the university of michigan's [institute for social research](#), if you apply for an interviewer job with them, i hope you like werther's original.

figuring out how to analyze this data set might trigger your [fight-or-flight](#) synapses if you just start clicking around on michigan's website. instead, read pages numbered 10-17 (pdf pages 12-19) of [this introduction pdf](#) and don't touch the data until you understand figure a-3 on that last page. if you start enjoying yourself, here's [the whole book](#). after that, it's time to [register](#) for access to the (free) data. keep your username and password handy, you'll need it for the top of the download automation r script. next, look at this [data flowchart](#) to get an idea of why the [data download](#) page is such a righteous jungle. but wait, good news: umich recently farmed out its data management to [the rand corporation](#), who promptly constructed [a giant consolidated file](#) with one record per respondent across the whole panel. oh so beautiful. the rand hrs files make much of the older data and syntax examples obsolete, so when you come across stuff like [instructions on how to merge years](#), you can happily ignore them - rand has done it for you.

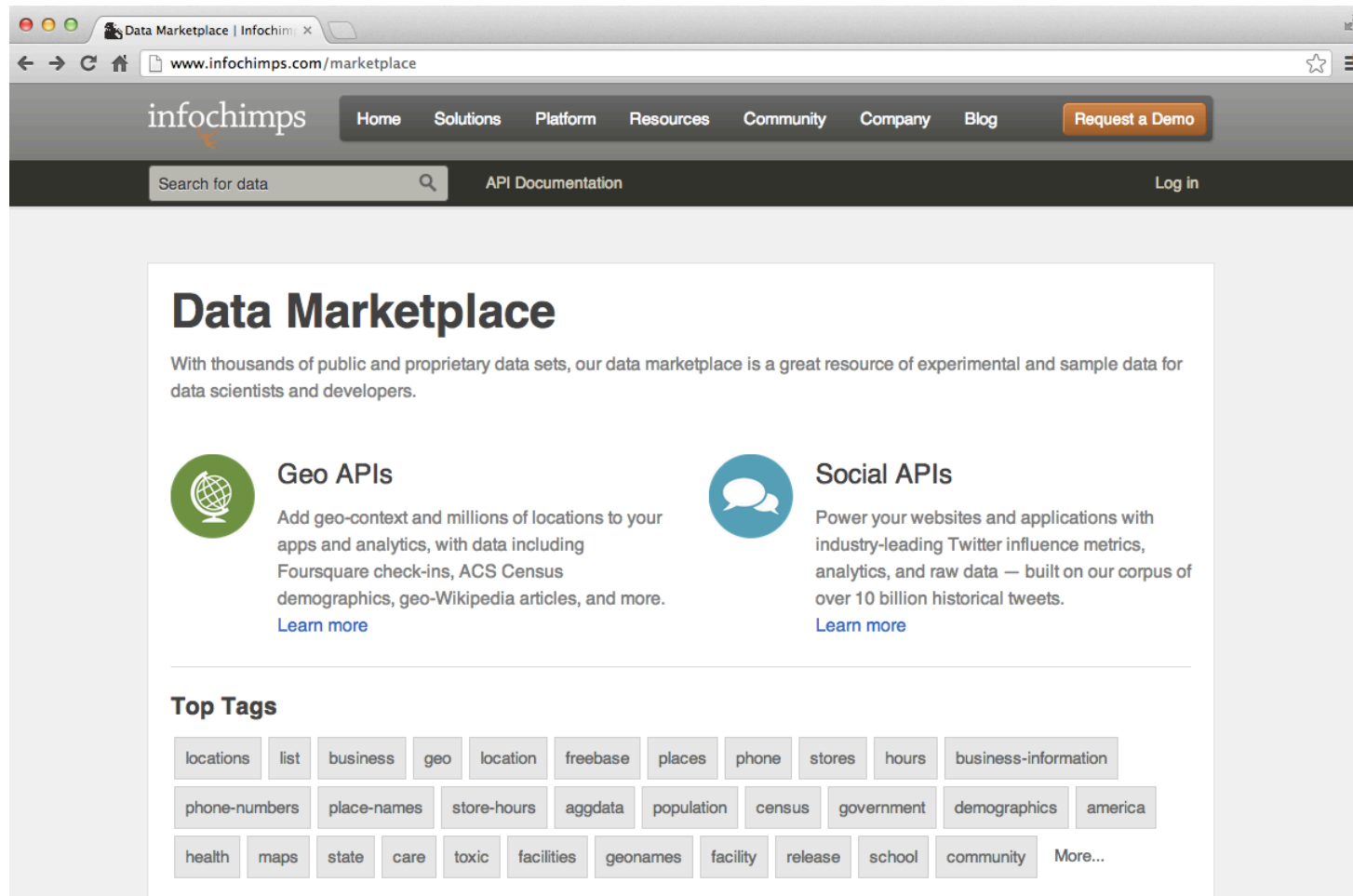
the health and retirement study only includes noninstitutionalized adults when new respondents get added to the panel (as they were in [1992](#), [1993](#), [1998](#), [2004](#), and [2010](#)) but once they're in, they're in - respondents have a weight of zero for interview waves when they were nursing home residents; but they're still responding and will continue to contribute to your statistics so long as you're generalizing about a population from a previous wave (for example: it's possible to compute "among all americans who were 50+ years old in 1998, x% lived in nursing homes by 2010"). my source for that 411? [page 13 of the design doc](#). wicked. this new github repository contains five scripts:

enter your email address for updates:  
[www.asdfree.com/2013/01/analyze-health-and-retirement-study-hrs.html](http://www.asdfree.com/2013/01/analyze-health-and-retirement-study-hrs.html)   1992 - 2010 download HRS microdata.R

<http://www.asdfree.com/>

7/11

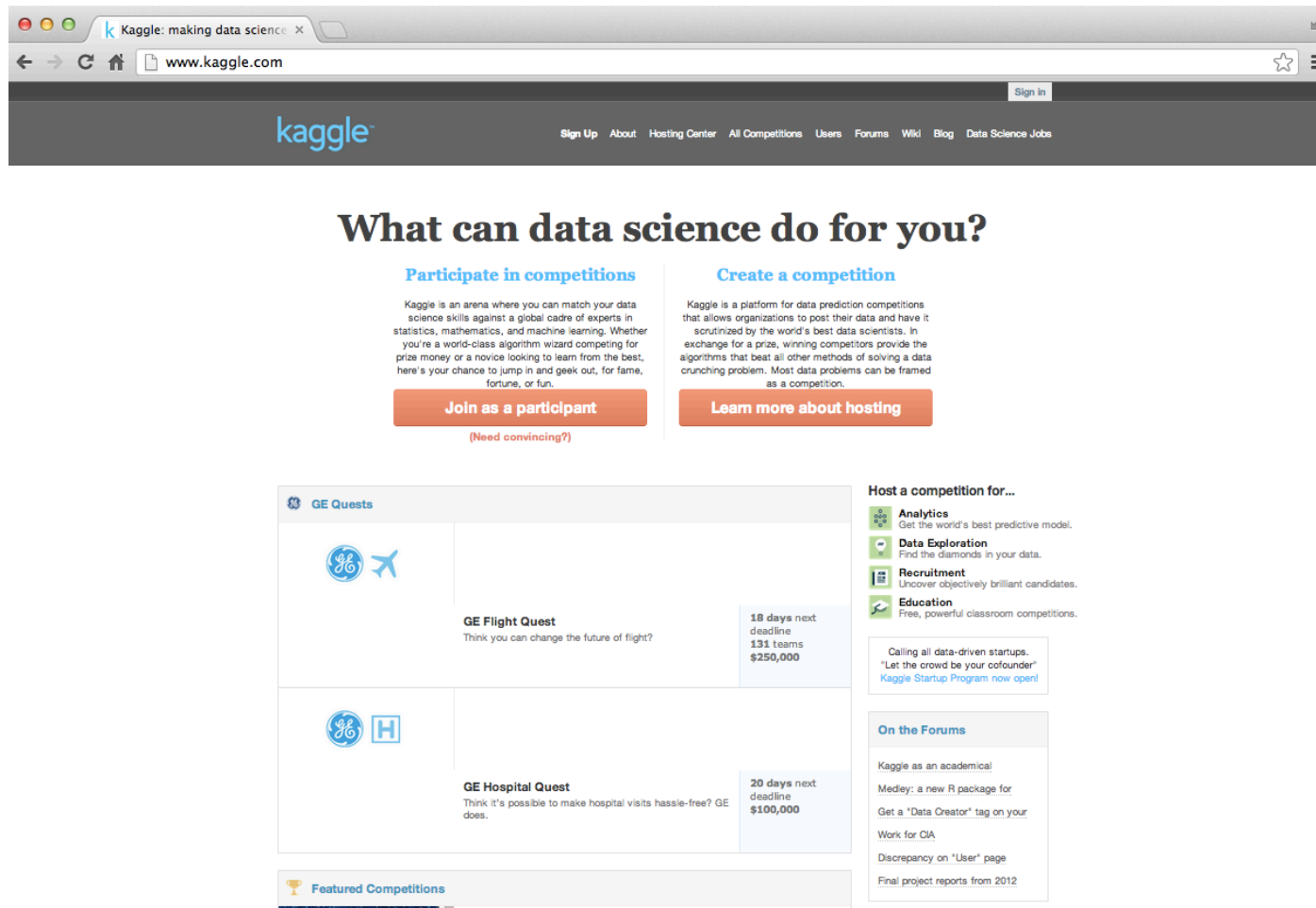
# Infochimps Marketplace



<http://www.infochimps.com/marketplace>



# Kaggle



The screenshot shows the Kaggle website homepage in a web browser. The browser's address bar displays "www.kaggle.com". The page features a dark header with the Kaggle logo and navigation links: Sign Up, About, Hosting Center, All Competitions, Users, Forums, Wiki, Blog, and Data Science Jobs. A "Sign In" button is also present. The main content area is titled "What can data science do for you?" and is divided into two columns. The left column, "Participate in competitions", describes Kaggle as an arena for data science skills and includes a "Join as a participant" button with a "(Need convincing?)" link. The right column, "Create a competition", describes Kaggle as a platform for data prediction competitions and includes a "Learn more about hosting" button. Below these columns, there are two featured competitions: "GE Flight Quest" (18 days next deadline, 131 teams, \$250,000) and "GE Hospital Quest" (20 days next deadline, \$100,000). To the right of these, there is a section "Host a competition for..." with categories: Analytics, Data Exploration, Recruitment, and Education. At the bottom right, there is a section "On the Forums" with links to various forum topics.

**What can data science do for you?**

**Participate in competitions**

Kaggle is an arena where you can match your data science skills against a global cadre of experts in statistics, mathematics, and machine learning. Whether you're a world-class algorithm wizard competing for prize money or a novice looking to learn from the best, here's your chance to jump in and geek out, for fame, fortune, or fun.

**Join as a participant**

(Need convincing?)

**Create a competition**

Kaggle is a platform for data prediction competitions that allows organizations to post their data and have it scrutinized by the world's best data scientists. In exchange for a prize, winning competitors provide the algorithms that beat all other methods of solving a data crunching problem. Most data problems can be framed as a competition.

**Learn more about hosting**

**GE Quests**

**GE Flight Quest**  
Think you can change the future of flight?

18 days next deadline  
131 teams  
\$250,000

**GE Hospital Quest**  
Think it's possible to make hospital visits hassle-free? GE does.

20 days next deadline  
\$100,000

**Host a competition for...**

- Analytics**  
Get the world's best predictive model.
- Data Exploration**  
Find the diamonds in your data.
- Recruitment**  
Uncover objectively brilliant candidates.
- Education**  
Free, powerful classroom competitions.

Calling all data-driven startups.  
"Let the crowd be your cofounder"  
[Kaggle Startup Program now open!](#)

**On the Forums**

- [Kaggle as an academic](#)
- [Medley: a new R package for](#)
- [Get a "Data Creator" tag on your](#)
- [Work for CIA](#)
- [Discrepancy on "User" page](#)
- [Final project reports from 2012](#)

**Featured Competitions**

<http://www.kaggle.com/>

# More specialized collections

- [Hilary Mason's research data](#)
- [Stanford Large Newtork Data](#)
- [UCI Machine Learning](#)
- [KDD Nugets Datasets](#)
- [CMU Statlib](#)
- [Gene expression omnibus](#)
- [ArXiv Data](#)

# Some API's

- [twitter](#) and [twitter](#) package
- [figshare](#) and [rfigshare](#)
- [PLOS](#) and [rplos](#)
- [rOpenSci](#)