

OPEN DATA

FROM DATA POINTS TO DATA LAKES

DR J ROGEL-SALAZAR
IMPERIAL COLLEGE LONDON
AND UNIVERSITY OF HERTFORDSHIRE

J.ROGEL@PHYSICS.ORG

@QUANTUM_TUNNEL / @HIDDEN_NODE

SOCIAL MEDIA

USE

#DIALOGO_OPENDATA



LET'S START AT THE BEGINNING

DATA?



INTERCONNECTED
KNOWLEDGE

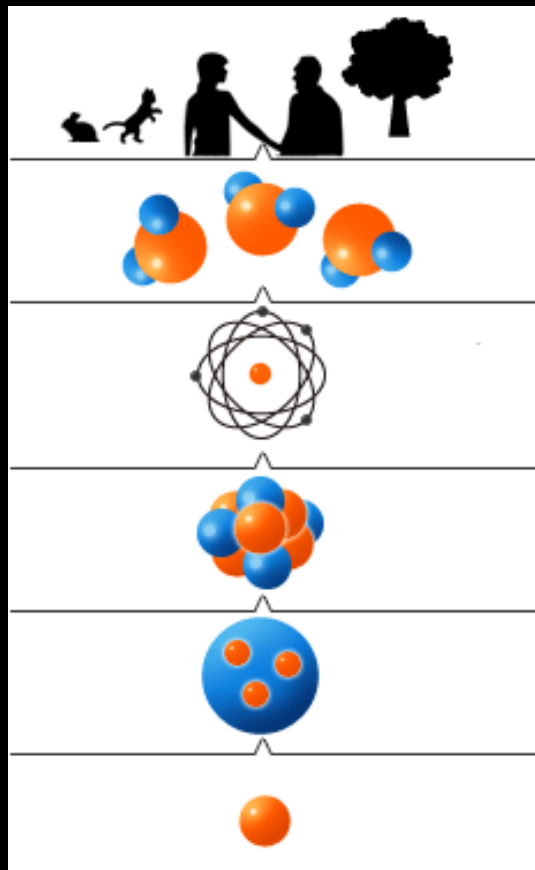
KNOWLEDGE

LINKED
INFORMATION

INFORMATION

STRUCTURED DATA

DATA



DATA EVERYWHERE!

- Lots of data is being collected and warehoused
- Scientific studies
- Web data, e-commerce
- Purchases at department/grocery stores
- Bank/Credit card transactions
- Social network



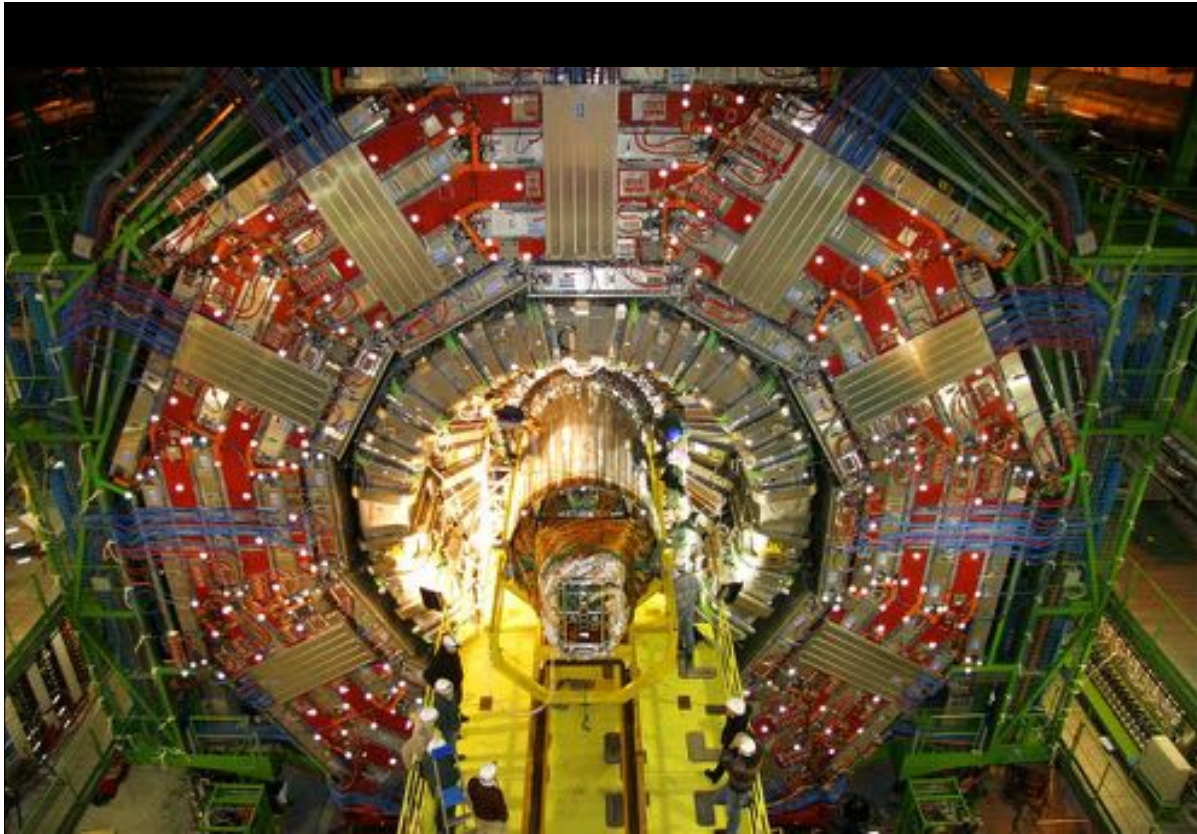
HOW MUCH DATA?

- Google processes 100 PB a day (2014)
- Facebook 600 TB/day (2014)
- Twitter 100 TB/day (2013/14)
- CERN's Large Hydron Collider (LHC) generates 15 PB a year



640K ought to be enough for anybody.

Source: <https://followthedata.wordpress.com/2014/06/24/data-size-estimates/>



Maximilien Brice, © CERN

THE EARTHSCOPE

- The Earthscope is also a large science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyses seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more.



http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--ul

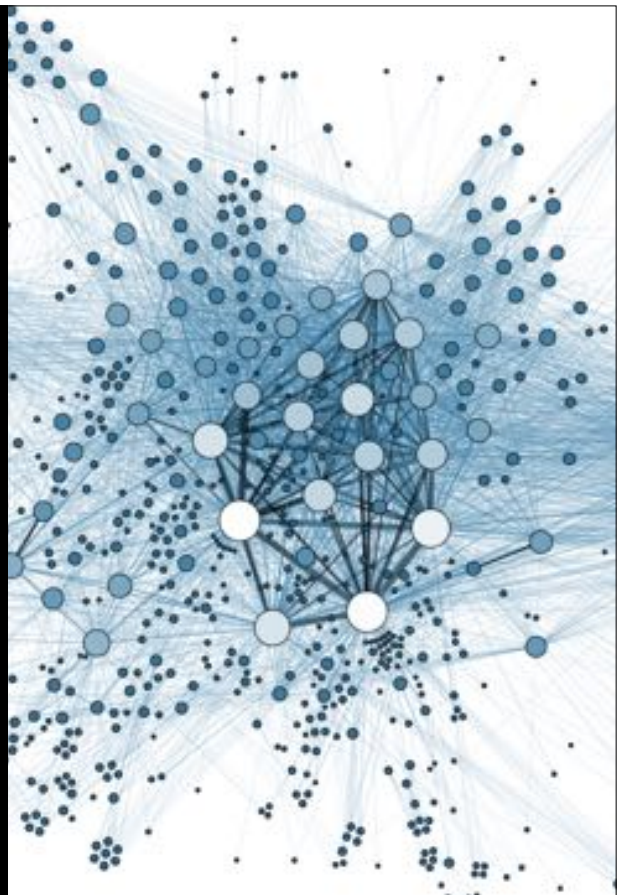
TYPE OF DATA

- Relational Data (Tables/ Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
 - Social Network, Semantic Web (RDF), ...
- Streaming Data
 - You can only scan the data once



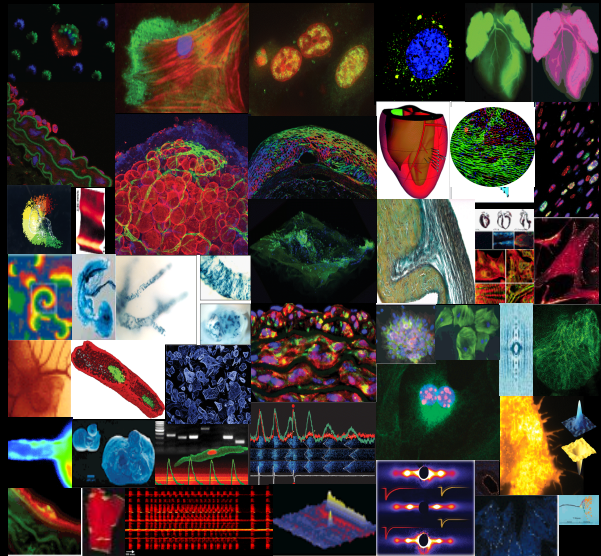
WHAT TO DO WITH THESE DATA?

- Aggregation and Statistics
 - Data warehouse and OLAP
- Indexing, Searching, and Querying
 - Keyword based search
 - Pattern matching (XML/ RDF)
- Knowledge discovery
 - Data Mining
 - Statistical Modeling



THE DATA

- Fundamental to research
- Basis for writing papers
- Important for experiment replication
- Meet contractual/funding requirements
- Settle intellectual property claims
- Defense against a charge of fraud



Images from the front covers of Circulation Research – S. Elliott (Van Eyk Lab)

INDIVIDUAL RESPONSIBILITY DATA MANAGEMENT

Some aspects to consider:

- Ownership
- Collection
- Storage/protection of confidentiality/sharing
- Interpretation and publication



COPYRIGHT

WHAT IS



?

"To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries."

-US CONSTITUTION

NOT A TOOL
TO CONTROL
ALL CONTENT
FOREVER IN
ALL MEDIA



A SET OF RIGHTS

- The right to reproduce the work
- The right to prepare derivative works
- The right to distribute the work
- The right to perform the work
- The right to display the work
- The right to license any of the above to third parties

HOW?

First, it must meet some basic requirements:

- It must be **original**.
- It must have some level of **creativity**.
- It must be in a **fixed medium**.

In the old-days, you would use this symbol:
Provide a date and register it.



NOWADAYS

IT'S INSTANT!





Copyright protects...

Writing
Choreography
Music
Visual art
Film
Architectural works

Copyright doesn't protect...

Ideas
Facts
Data (mostly)
Useful articles (that's patent)

HOW LONG
DOES IT LAST?

The life of the author
plus 70 years

FOR NOW...

And then?

THE PUBLIC DOMAIN

GENERAL RULES FOR STATUS

Works No Longer Protected by Copyright

- Published before 1923
- Published between '23 and '63, but it depends.
- Authored by the Federal Government (US)

VERBOSE MODE...

- All works published in the United States before 1923 are in the public domain.
- Works published after 1922, but before 1978 are protected for 95 years from the date of publication. If the work was created, but not published, before 1978, the copyright lasts for the life of the author plus 70 years. However, even if the author died over 70 years ago, the copyright in an unpublished work lasts until December 31, 2002.
- For works published after 1977, the copyright lasts for the life of the author plus 70 years. However, if the work is a work for hire (that is, the work is done in the course of employment or has been specifically commissioned) or is published anonymously or under a pseudonym, the copyright lasts between 95 and 120 years, depending on the date the work is published.
- Lastly, if the work was published between 1923 and 1963, you must check with the U.S. Copyright Office to see whether the copyright was properly renewed. If the author failed to renew the copyright, the work has fallen into the public domain and you may use it.

CONFUSED?



Hard to share

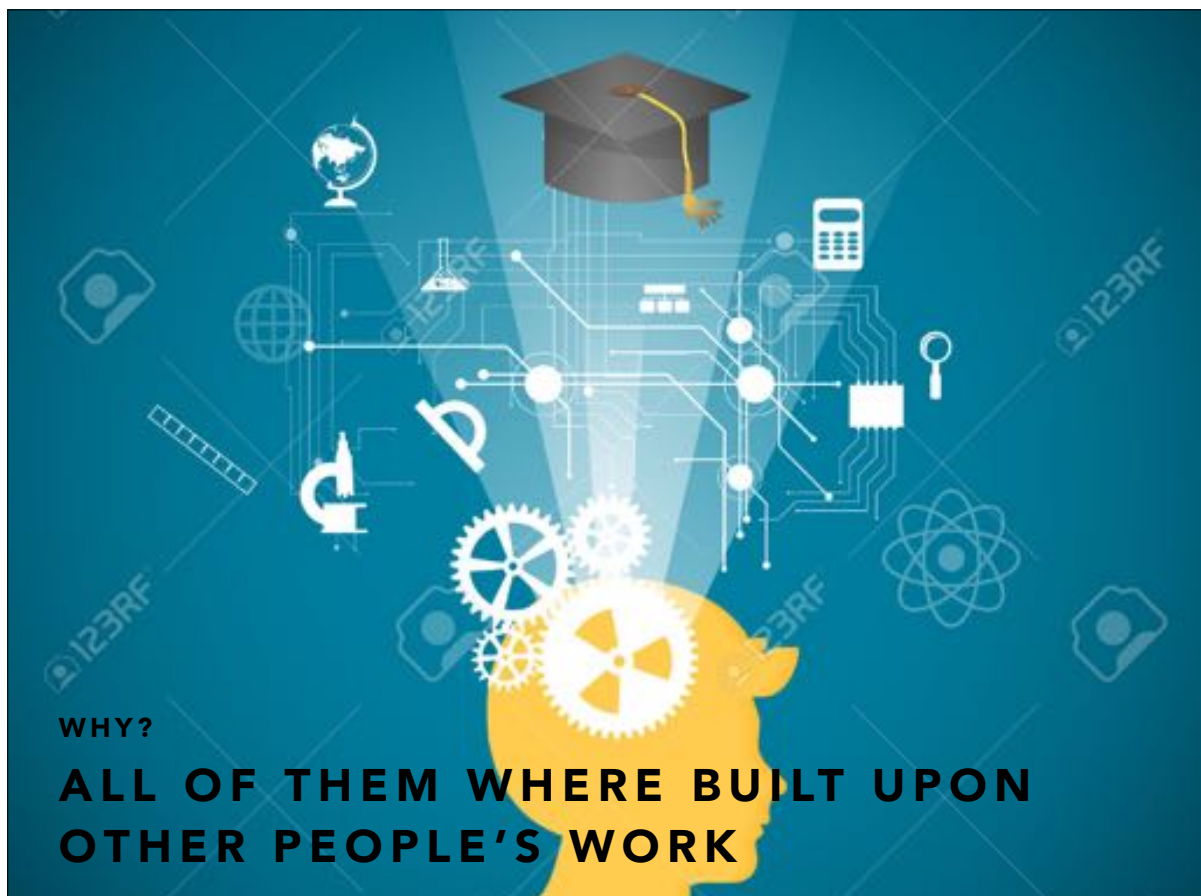


WHY SHARE?



ALL OF THEM CAN...

AND SHOULD BE
SHARED!



[HTTPS://TERRYTAO.WORDPRESS.COM](https://terrytao.wordpress.com)

TERRENCE TAO BLOG



What's new

Updates on my research and expository papers, discussion of open problems, and other maths-related topics. By Terence Tao

 [Subscribe to feed](#)

[Home](#) [About](#) [Career advice](#) [On writing](#) [Books](#) [Applets](#)

RECENT COMMENTS



Terence Tao on Heath-Brown's theorem on...



Sergei on Heath-Brown's theorem on...



Anonymous on Heath-Brown's theorem on...



Terence Tao on Heath-Brown's theorem on...



Sergei on Heath-Brown's theorem on...

Heath-Brown's theorem on prime twins and Siegel zeroes

26 August, 2015 in [expository](#), [math.NT](#) | [Tags](#): [prime numbers](#), [Roger Heath-Brown](#), [Siegel zero](#), [twin primes](#) | [by Terence Tao](#) | [16 comments](#)

The twin prime conjecture is one of the oldest unsolved problems in analytic number theory. There are several reasons why this conjecture remains out of reach of current techniques, but the most important obstacle is the parity problem which prevents purely sieve-theoretic methods (or many other popular methods in analytic number theory, such as the circle method) from detecting pairs of prime twins in a way that can distinguish them from other twins of almost primes.

BRITTANY WENGER FROM FLORIDA:



The Grand Prize winner of the science fair, for good reason, was a 17-year-old from Lakewood Ranch, Florida. Combining the fields of biology and computer science, Wenger wrote an app that helps doctors diagnose breast cancer, [according to the description of her project on Google](#).

The type of computer program, called a "neural network," was designed by Wenger to mimic the human brain: Give it a massive amount of information (in this case, 7.6 million trials), and the artificial "brain" will learn to detect complex patterns and make diagnostic calls on breast cancer. Her program used data from "fine needle aspirates," a minimally invasive procedure that, unfortunately, is often one of the least precise diagnosis processes, [according to Fox News](#). But Wenger is helping change that, as her program correctly identifies 99 percent of malignant tumors.

"I think it might be hospital ready," she [told WWSB](#). "I'd love to get different data from doctors. Right now, I have 700 test samples."

[HTTPS://WWW.KAGGLE.COM/COMPETITIONS](https://www.kaggle.com/competitions)

KAGGLE

The screenshot shows the Kaggle website's 'Active Competitions' page. The header includes navigation links: Host, Competitions, Scripts, Jobs, Community, Sign up, and Login. The main content area is divided into three columns: 'Download' (Choose a competition & download the training data), 'Build' (Build a model using whatever methods and tools you prefer), and 'Submit' (Upload your predictions. Kaggle scores your solution and shows your score on the leaderboard). Below this, the 'Active Competitions' section lists several competitions:

Competition	Description	Duration	Teams	Scripts	Prize
Springleaf Marketing Response	Determine whether to send a direct mail piece to a customer	42 days	1104 teams	591 scripts	\$100,000
Western Australia Rental Prices	Predict rental prices for properties across Western Australia	2 months	18 teams	100,000	
Coupon Purchase Prediction	Predict which coupons a customer will buy	23 days	757 teams	480 scripts	\$50,000

A SPECTRUM OF RIGHTS



least
restrictive



most
restrictive



SO...
WHAT IS OPEN
DATA?



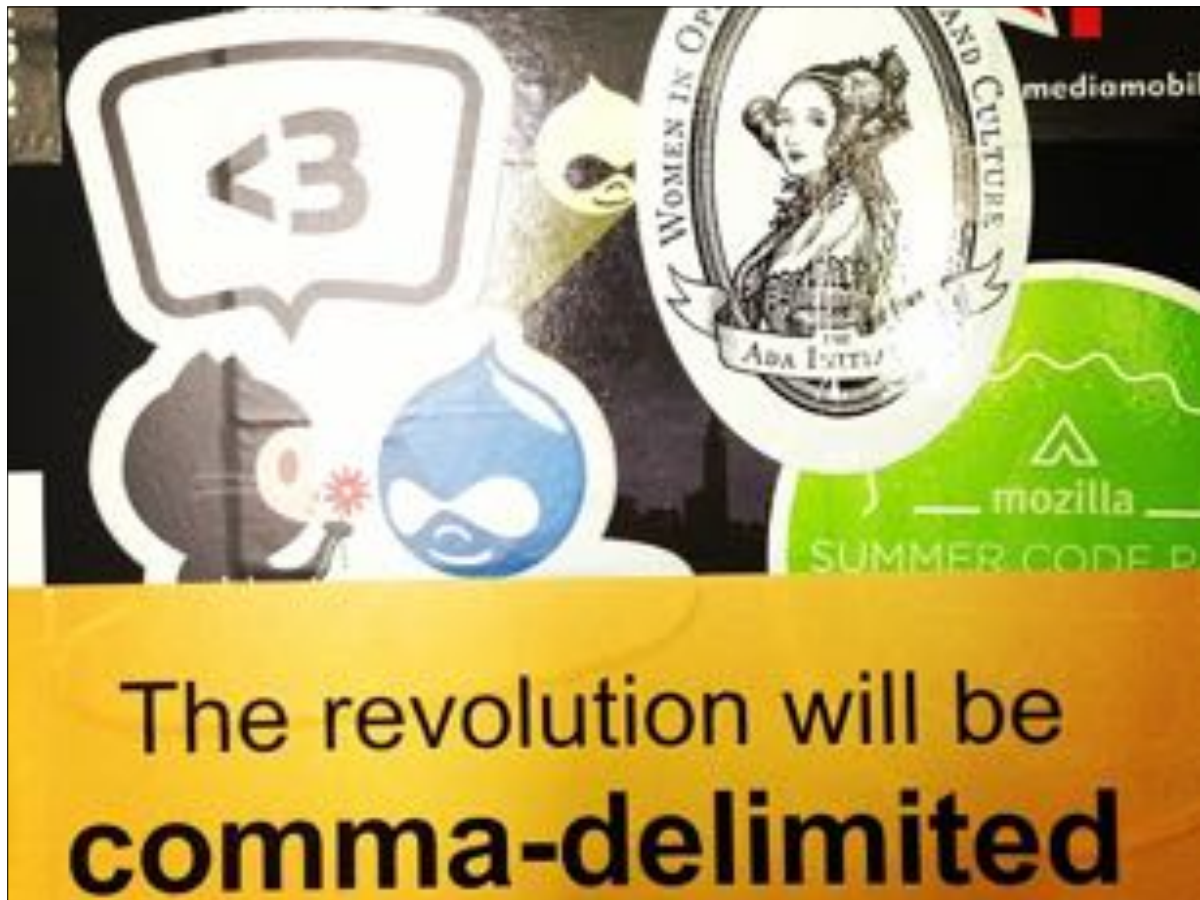
Open data is information that is available for anyone to use, for any purpose, at no cost.

The logo for 'Open Data' is displayed in a pixelated, monospace font. The word 'OPEN' is in blue on a white background, and the word 'DATA' is in white on a blue background. The entire logo is enclosed in a thin white border.

OPEN DATA

GOOD OPEN DATA

- Can be linked: shared more easily
- Available in a standard format: easily processed
- Guaranteed availability and consistency: easily reliable
- Traceable: easily trusted



[HTTPS://WWW.OVERLEAF.COM](https://www.overleaf.com)

OVERLEAF



Online Collaborative Rich Text Editing



The convenience of an easy-to-use WYSIWYG manuscript editor, with real-time collaboration, and structured, fully typeset output produced automatically in the background as you type.

Prefer to edit directly in LaTeX? Overleaf provides a full collaborative online LaTeX editor you can switch to at any time.



[HTTP://DATADRYAD.ORG](http://datadryad.org)

DRYAD



About - For researchers - For organizations - Contact us Log in Sign up

DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad has integrated data submission for a growing list of journals; submission of data from other publications is also welcome.



Submit data now

How and why?

Search for data

Enter keyword, author, title, DOI, etc. [Go](#)

[Advanced search](#)

Browse for data

Recently published

Popular

By author

By journal

Recently published data

Koch H, Jeschke A, Becks L (2015) Data from: Use of ddPCR in experimental evolution studies. *Methods in Ecology and Evolution* <http://dx.doi.org/10.5061/dryad.8ky2s>

Latest from @datadryad

Tweets

[Follow](#)

Dryad
@datadryad

15h

"These are wild west days for data citation" says @jenniferin15 blog.datacite.org/when-counting-...

[HTTPS://DATAVERSE.HARVARD.EDU](https://dataverse.harvard.edu)

DATAVERSE

The screenshot shows the Harvard DataVerse homepage. At the top left is the Harvard crest logo. To its right, the text 'Harvard DataVerse' is displayed, followed by a subtitle: 'A collaboration with Harvard Library, Harvard University IT, and IQSS'. Below this, a 'Metrics' bar shows '1,374,055 Downloads'. To the right of the metrics are icons for email and social media. A main heading reads: 'Share, publish, and archive your data. Find and cite data across all research fields.' Below this heading is a row of four featured data archives, each with a logo and name: 'World Agroforestry Centre - ICRAF DataVerse', 'Population Services International (PSI) DataVerse', 'International Food Policy Research Institute (IFPRI) DataVerse', and 'Henry A. Murray Research Archive at Harvard University - Murray Research Archive DataVerse'. Below the featured archives is a search bar with the placeholder text 'Search this dataverse...'. To the right of the search bar are buttons for 'Find', 'Advanced Search', and '+ Add Data'. Below the search bar, there is a summary of search results: 'Dataverses (1,206)' and 'Datasets (59,048)'. To the right of this summary is a pagination control showing '1 to 10 of 60,254 Results', a 'Sort' button, and a series of numbered links (1, 2, 3, 4, 5) with '< Previous' and 'Next >' buttons.

[HTTP://DATA.GOV.UK](http://data.gov.uk)

DATA GOV UK

The screenshot shows the Data.gov.uk homepage. At the top is the 'DATA.GOV.UK' logo with the tagline 'Opening up Government'. To the right of the logo are navigation links: 'Home', 'Data', 'Apps', and 'Interact'. Below these links is a search bar with the placeholder text 'Search for data...'. Below the search bar is a horizontal menu with links: 'Datasets', 'Map Search', 'Data Requests', 'Publishers', 'Organograms', 'Spend Reports', 'Site Analytics', 'Reports', and 'Contracts'. Below the menu is a grid of six featured articles. The first article on the left is titled 'NII' and has the subtitle 'The National Information Infrastructure exemplars work'. The second article in the middle is titled 'Contracts Finder Archive' and has the subtitle 'Contracts Finder Archive'. The third article on the right is titled 'RELEASE OF DATA FUND' and has the subtitle 'One small step for local...'. Below these three articles are three more articles: 'Defra announces major open data', 'Kicking off the Open Government Project', and 'Companies House new free information service'.

[HTTP://WWW.DATA.GOV](http://www.data.gov)

DATA GOV US



[DATA](#) [TOPICS](#) - [IMPACT](#) [APPLICATIONS](#) [DEVELOPERS](#) [CONTACT](#)

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

SEARCH OVER 147,148 DATASETS



[HTTP://DATOS.GOB.MX](http://datos.gob.mx)

DATOS GOB MEX

 [datos.gob.mx](#)

[Datos](#) [Guía](#) [Historias](#) [Apps](#) [Herramientas](#) [Avances](#) [Acerca](#)



CONOCE LAS HISTORIAS



EXPLORA LOS DATOS



UTILIZA LAS HERRAMIENTAS

En **datos.gob.mx** puedes descargar y utilizar libremente datos públicos que el Gobierno de la República genera y recolecta.

ANYONE ELSE?

ANY BIG INSTITUTION COULD PUBLISH OPEN DATA

[HTTP://WWW.THEGUARDIAN.COM/NEWS/DATABLOG/](http://www.theguardian.com/news/datablog/)

THE GUARDIAN



The screenshot shows the Guardian Datablog homepage. At the top is the Guardian logo and a navigation bar with links to various sections like Home, UK, World, Sport, Football, Opinion, Culture, Economy, Lifestyle, Fashion, Environment, Tech, Travel, and Money. Below this is a 'News' tab and a 'Datablog' tab. The main heading is 'DATA BLOG' with the tagline 'Facts are sacred'. A red horizontal line separates the header from the main content. Below the line, there are 'Previous', 'Blog home', and 'Next' links. The main article is titled 'All our datasets: the complete index' and discusses the hundreds of datasets published by the Guardian Datablog since 2009. It mentions that the table below is live and updated every day. To the right of the article text are social media sharing buttons for Facebook, Twitter, Google+, LinkedIn, and Email, each with a count of shares or tweets.

the guardian

Google Custom Search

Home UK World Sport Football Opinion Culture Economy Lifestyle Fashion Environment Tech Travel Money

News Datablog

DATA BLOG

Facts are sacred

Previous Blog home Next

All our datasets: the complete index

Lost track of the hundreds of datasets published by the Guardian Datablog since it began in 2009? Thanks to [ScraperWiki](#), this is the ultimate list and resource. The table below is live and updated every day - if you're still looking for that ultimate dataset, the chance is we've already done it. Click below to find out

- DATA: download as a CSV from [ScraperWiki](#)

Facebook Share 28
Twitter Tweet 22
Google+ 33
LinkedIn Share 7
Email

[HTTP://WWW.OPENDATA500.COM](http://www.opendata500.com)

OPEN DATA 500



OD500 US

ABOUT

FULL LIST

SURVEY

ROUNDTABLES

RESOURCES

About the Open Data 500

The Open Data 500 is the first comprehensive study of U.S. companies that use open government data to generate new business and develop new products and services. Open Data is free, public data that can be used to launch commercial and nonprofit ventures, do research, make data-driven decisions, and solve complex problems.

[HTTP://FIGSHARE.COM](http://figshare.com)

FIGSHARE



search figshare (titles, tags, authors, etc.)



Browse

Upload

Sign up

Login

store, share, discover **research**

manage your research in the cloud and control who you share it with
or make it publicly available and citable

About figshare

Browse research

See how we support data management for Institutions >

See how we partner with Publishers >

sign up for free

first name

last name

email

confirm email

password

☐ Accept Terms & Conditions

Sign up

[HTTP://ARCHIVE.ICS.UCI.EDU/ML/](http://archive.ics.uci.edu/ml/)

UCI MACHINE LEARNING

The screenshot shows the homepage of the UCI Machine Learning Repository. At the top, there is a navigation bar with links for 'About', 'Citation Policy', 'Donate a Data Set', and 'Contact'. Below this is the UCI logo, which includes a stylized animal, and the text 'Machine Learning Repository' and 'Center for Machine Learning and Intelligent Systems'. A search bar is also present. The main content area features a welcome message and a paragraph explaining the repository's mission. It lists the number of data sets (332) and provides links to view all data sets, the old web site, about page, citation policy, donation policy, and contact information. There are also logos for 'Supported By' (a circular logo) and 'In Collaboration With' (Rexa.info). The bottom section is divided into three columns: 'Latest News' with two news items dated 2013-04-04 and 2010-03-01; 'Newest Data Sets' with two entries dated 2015-08-04 and 2015-07-29; and 'Most Popular Data Sets (hits since 2007)' with two entries: 'Iris' (767906 hits) and 'Adult' (538619 hits).

[HTTPS://TFL.GOV.UK/INFO-FOR/OPEN-DATA-USERS/](https://tfl.gov.uk/info-for/open-data-users/)

TRANSPORT FOR LONDON

The screenshot shows the 'Open Data Users' page on the Transport for London website. The top navigation bar includes the TfL logo, 'Plan a Journey', 'Status updates', 'Maps', 'Fares & payments', and 'More...'. A search bar is also present. The main heading is 'OPEN DATA USERS'. Below this, there is a paragraph explaining that all public TfL data (or 'open data') is released here for developers to use in their own software and services. To the right, there is a 'Sign in or register for data feeds' section with a 'Sign in' button and a link to 'See detailed developer documentation'. At the bottom, there is a section titled 'OPEN DATA USERS' with two sub-sections: 'Our open data' and 'Data feeds'. The background of the bottom section features images of a TfL bus stop display and a person using a mobile device.



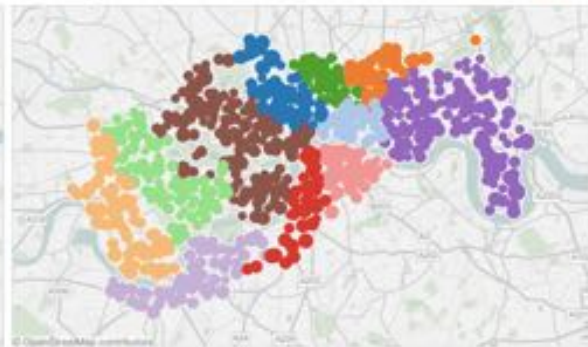
WHAT CAN WE DO WITH IT?

A map of the Boris bikes of London.

Taking a closer look at the 700+ bike stations in the city:



Click a borough to see the stations it contains & select a station to see the street it is on.



A map of the bikes of London

Using Tableau and a pinch of Python to look at ~20,000 bikes and 700+ bike stations.



Eric Hannell

When I moved to London I exchanged my Vespa for a Boris bike. It was tough going at first but I have come to love these bikes. I ride one almost every day and I could not imagine living in London without them.

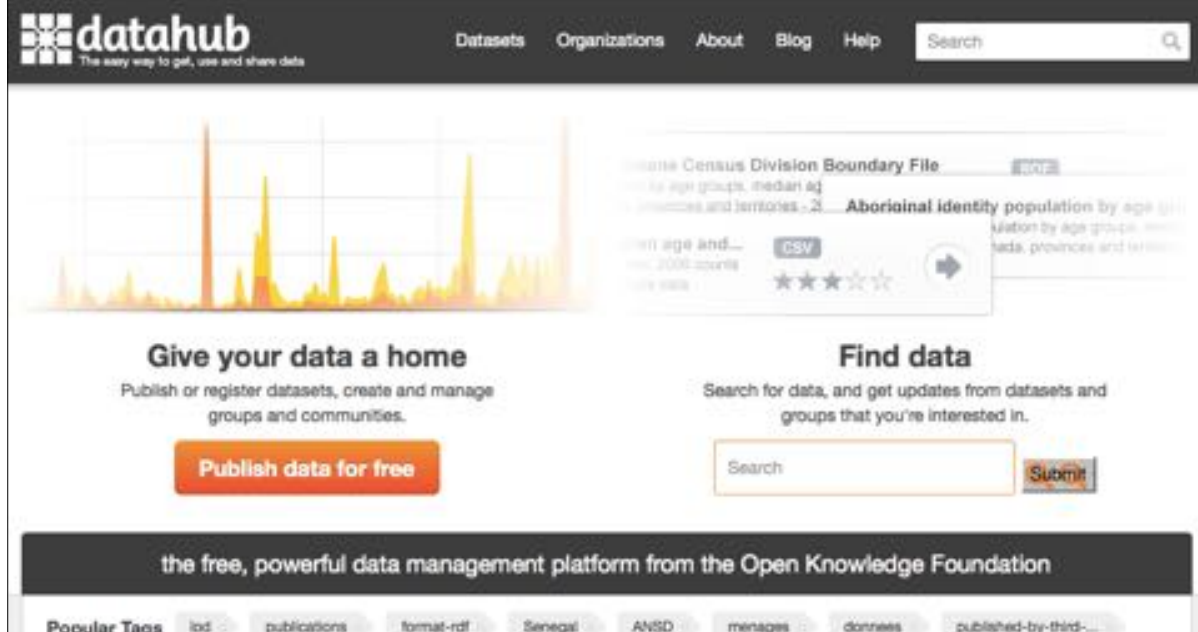
Being a "data person" it did not take too long before I started searching for data about the bikes. The program, alternately referred to as "Boris Bikes" / "Barclays Bikes" /

WHERE TO FIND OPEN DATA?



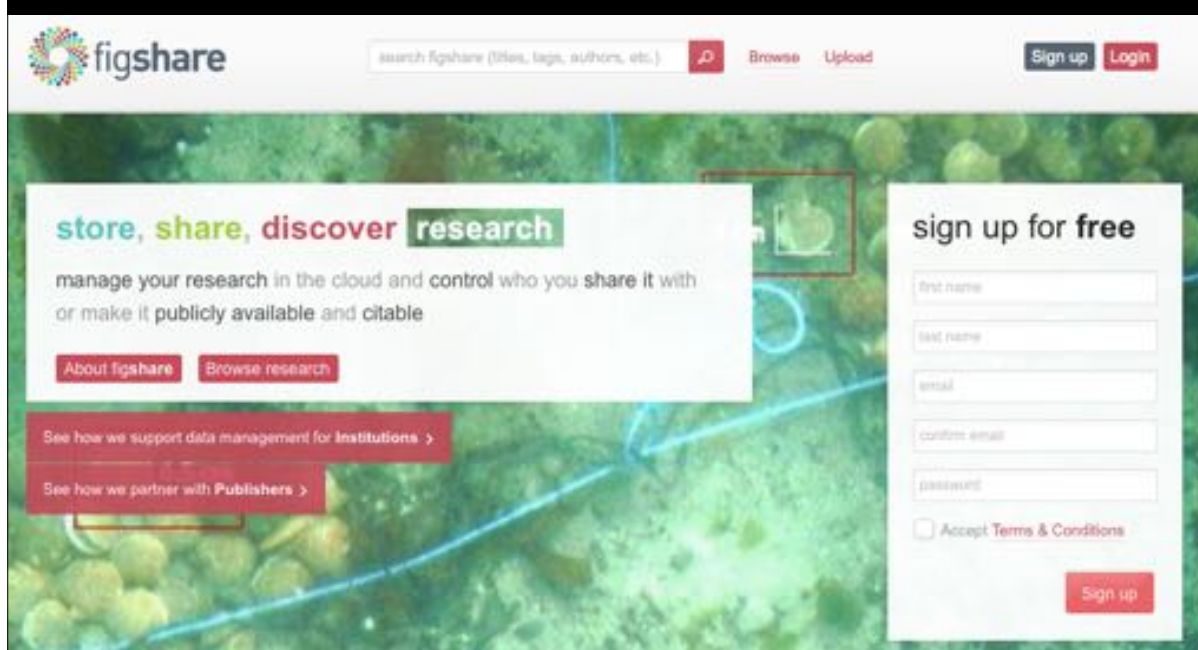
[HTTP://DATAHUB.IO](http://datahub.io)

DATAHUB




[HTTP://FIGSHARE.COM](http://figshare.com)

FIGSHARE



[HTTP://WWW.RE3DATA.ORG](http://www.re3data.org)

REGISTER OF DATA REPOS



The screenshot shows the re3data.org website. At the top is the logo "re3data.org" with the tagline "REGISTRY OF RESEARCH DATA REPOSITORIES". Below the logo is a navigation bar with links: Home, Search, Browse, Suggest, FAQ, About, Schema, API, Contact, and Imprint. The main content area features a post titled "Introduction of the re3data.org persistent identifier" dated August 26, 2015, by the re3data.org team. The text explains that every record is now persistently accessible and citable via its own persistent identifier, a result of cooperation with DataCite. To the right of the text is a search bar with the placeholder "re3data.org search" and a "PARTNERS" section featuring the logo of GFZ Helmholtz Centre Potsdam.

[HTTP://DATABIB.ORG](http://databib.org)

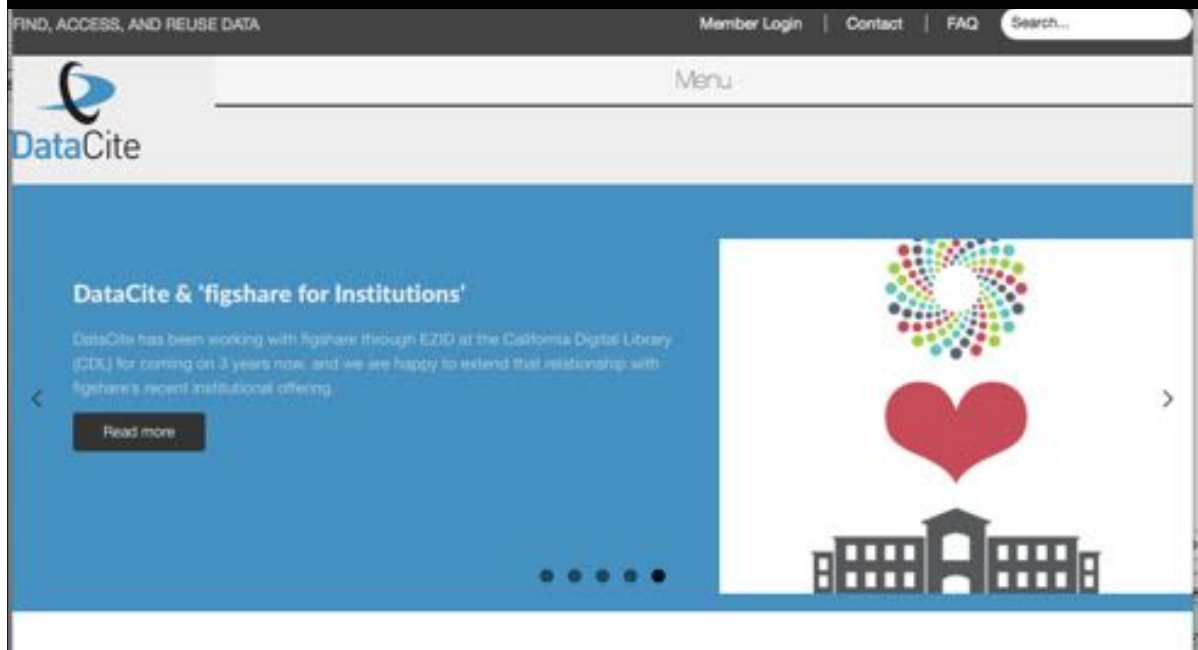
DATABIB



The screenshot shows the Databib website. At the top is the logo "Databib" with the tagline "Find Repositories | Submit | Connect | About" and a "Login/Register" link. The main content area features a section titled "About Databib" which states that Databib is a tool for helping people identify and locate online repositories of research data. It lists three key questions: "What repositories are appropriate for a researcher to submit his or her data to?", "How do users find appropriate data repositories and discover datasets that meet their needs?", and "How can librarians help patrons locate and integrate data into their research or learning?". It then states that Databib attempts to address these needs for the research community, including data users, data producers, publishers and professional societies, librarians, and research funding agencies. Below this is a section titled "Databib Advisory Board" which lists several members and their affiliations, including Adrián Bonilla Soria (FLACSO - Ecuador), Alma Swan (SPARC Europe - United Kingdom), Andrew Kaniki (National Research Foundation - South Africa), Andrew Trelor (ANDS - Australia), Frank Scholze (KIT Library - Germany), Martin Donnelly (DCC - United Kingdom), Patricia Cruise (California Digital Library - United States), and Paul Uhlir (National Academy of Sciences - United States).

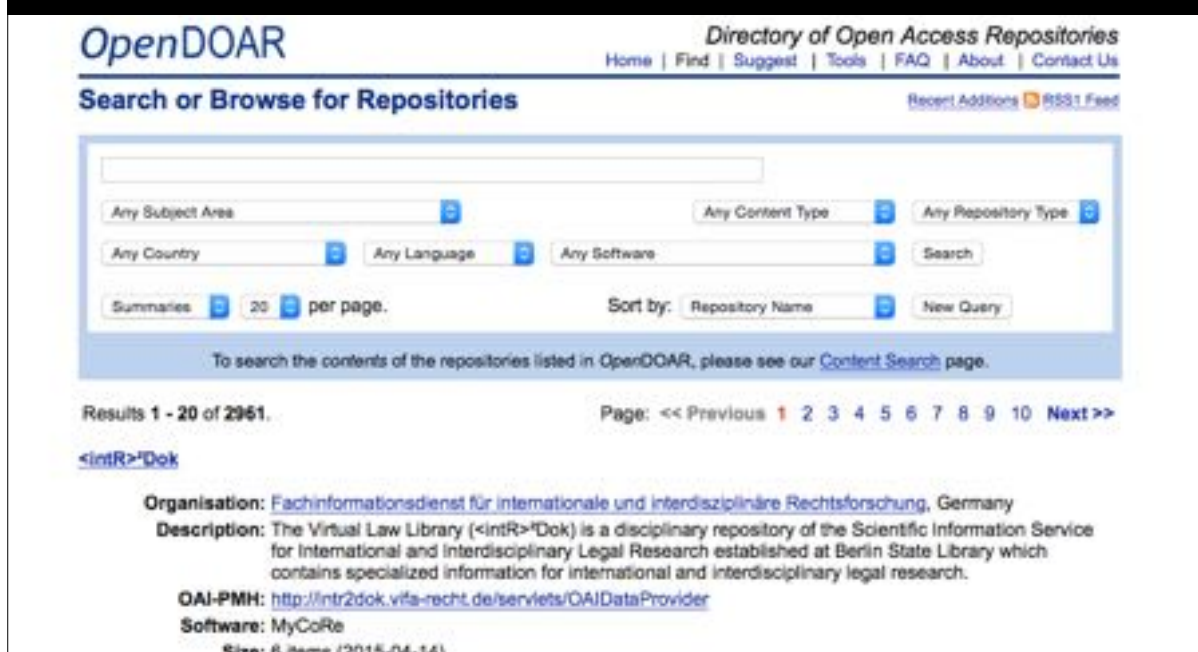
[HTTPS://WWW.DATACITE.ORG](https://www.datacite.org)

DATA CITE



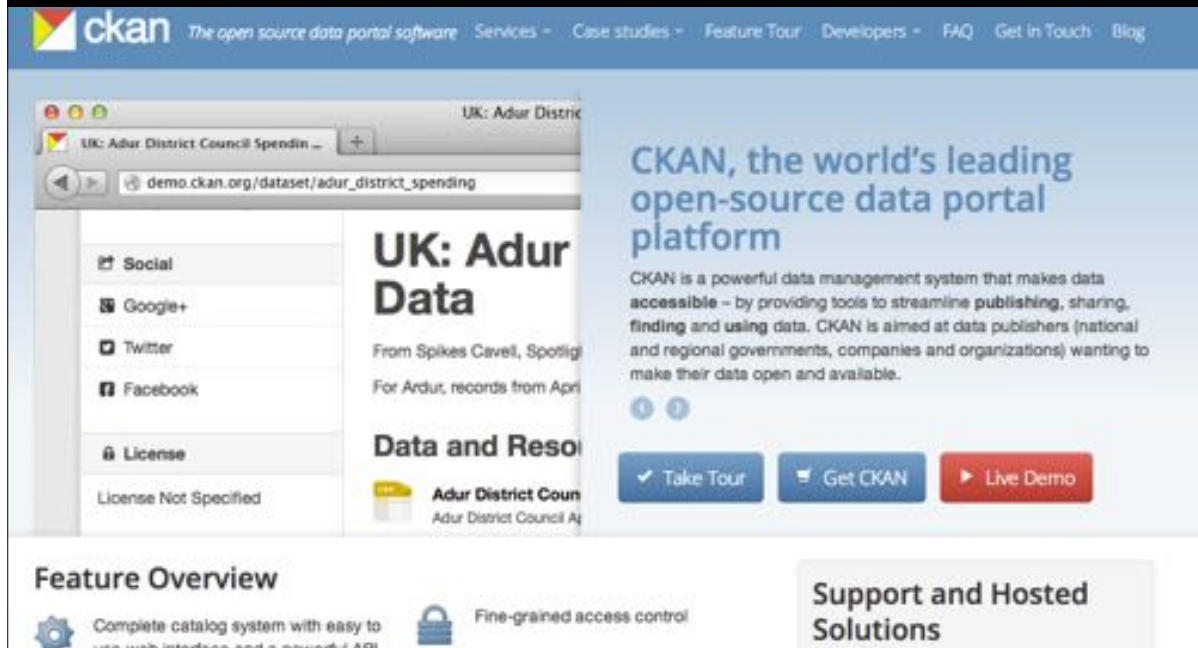
[HTTP://WWW.OPENDOAR.ORG](http://www.opendoar.org)

OPENDOAR



[HTTP://CKAN.ORG](http://ckan.org)

CKAN



The screenshot shows the CKAN website homepage. At the top, the CKAN logo is followed by the tagline "The open source data portal software" and a navigation menu with links: Services, Case studies, Feature Tour, Developers, FAQ, Get In Touch, and Blog. The main content area features a large header with the title "CKAN, the world's leading open-source data portal platform". Below this, a paragraph describes CKAN as a powerful data management system that makes data accessible by providing tools to streamline publishing, sharing, finding, and using data. It is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available. To the left of this text is a screenshot of a web browser displaying a dataset titled "UK: Adur District Council Spending". Below the main text are three buttons: "Take Tour", "Get CKAN", and "Live Demo". At the bottom, there is a "Feature Overview" section with icons and text describing features like "Complete catalog system with easy to use web interface and a powerful API" and "Fine-grained access control". To the right of this is a "Support and Hosted Solutions" section.

ckan The open source data portal software Services - Case studies - Feature Tour Developers - FAQ Get In Touch Blog

UK: Adur District Council Spending demo.ckan.org/dataset/adur_district_spending

UK: Adur Data

From Spikes Caveil, Spotlight For Adur, records from April

Data and Resources

Adur District Council Adur District Council Agency

CKAN, the world's leading open-source data portal platform

CKAN is a powerful data management system that makes data **accessible** – by providing tools to streamline **publishing**, **sharing**, **finding** and **using** data. CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available.

1 2

✓ Take Tour Get CKAN ▶ Live Demo

Feature Overview

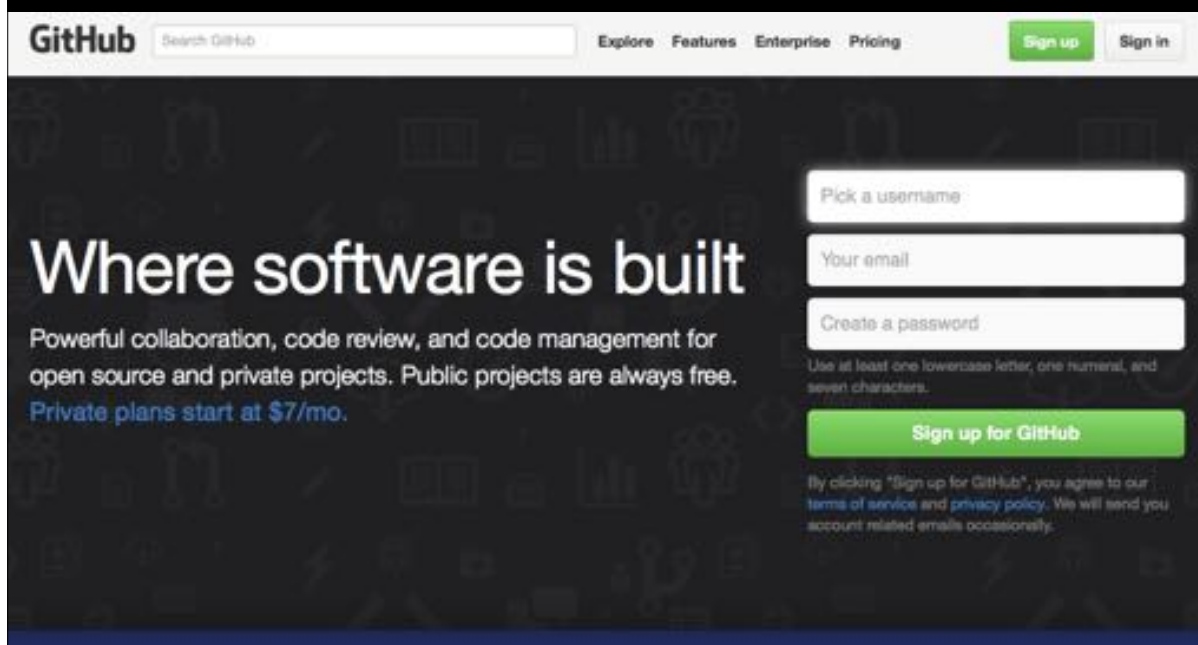
Complete catalog system with easy to use web interface and a powerful API

Fine-grained access control

Support and Hosted Solutions

[HTTPS://GITHUB.COM](https://github.com)

GITHUB



The screenshot shows the GitHub homepage. At the top, the GitHub logo is followed by a search bar and navigation links: Explore, Features, Enterprise, Pricing, Sign up, and Sign in. The main content area features a large header with the title "Where software is built". Below this, a paragraph describes GitHub as a powerful collaboration, code review, and code management platform for open source and private projects. It states that public projects are always free and that private plans start at \$7/mo. To the right of this text is a sign-up form with fields for "Pick a username", "Your email", and "Create a password". Below these fields is a green button labeled "Sign up for GitHub". At the bottom, a small disclaimer states that by clicking "Sign up for GitHub", the user agrees to the terms of service and privacy policy, and that they will receive account-related emails occasionally.

GitHub Search GitHub Explore Features Enterprise Pricing Sign up Sign in

Where software is built

Powerful collaboration, code review, and code management for open source and private projects. Public projects are always free. Private plans start at \$7/mo.

Pick a username

Your email

Create a password

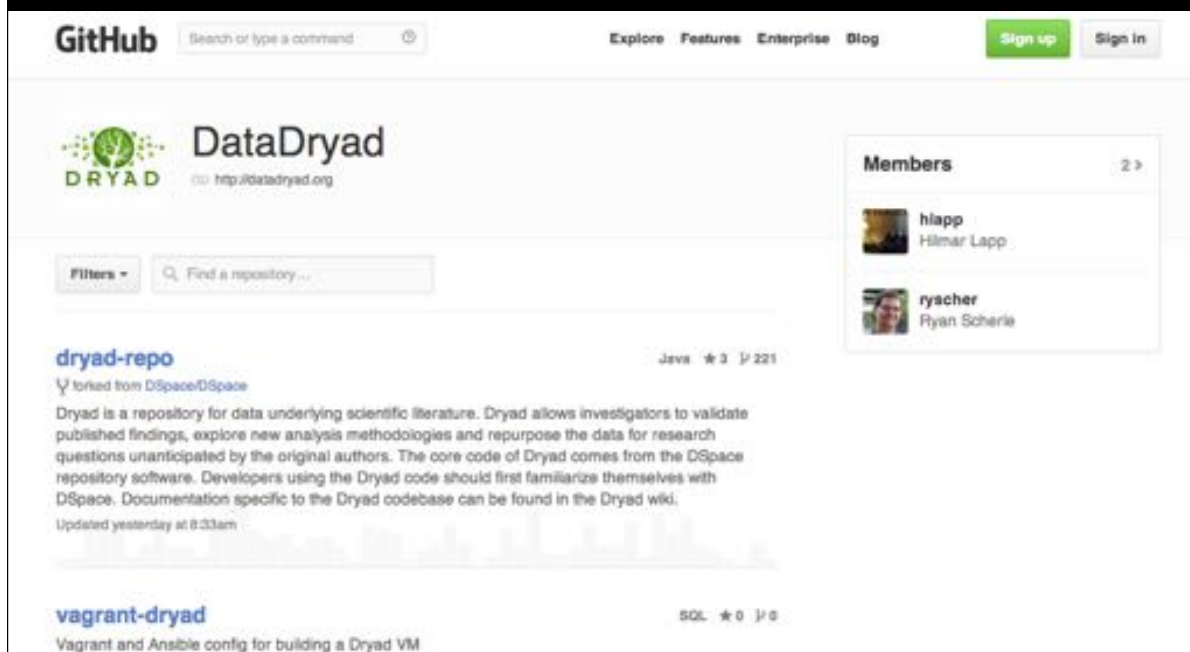
Use at least one lowercase letter, one numeral, and seven characters.

Sign up for GitHub

By clicking "Sign up for GitHub", you agree to our terms of service and privacy policy. We will send you account related emails occasionally.

[HTTPS://GITHUB.COM/DATADRYAD](https://github.com/DataDryad)

GITHUB - DATA DRYAD



The screenshot shows the GitHub repository page for DataDryad. At the top, the GitHub logo is on the left, and navigation links for Explore, Features, Enterprise, and Blog are on the right. A search bar and Sign up / Sign in buttons are also present. The repository name "DataDryad" is prominently displayed with its logo and URL. Below this, there's a search bar and a list of repositories. The first repository, "dryad-repo", is highlighted, showing its description, language (Java), and star count (3). The second repository, "vagrant-dryad", is also visible, showing its description and language (SQL).

GitHub Search or type a command

Explore Features Enterprise Blog Sign up Sign in

DataDryad <http://datadryad.org>

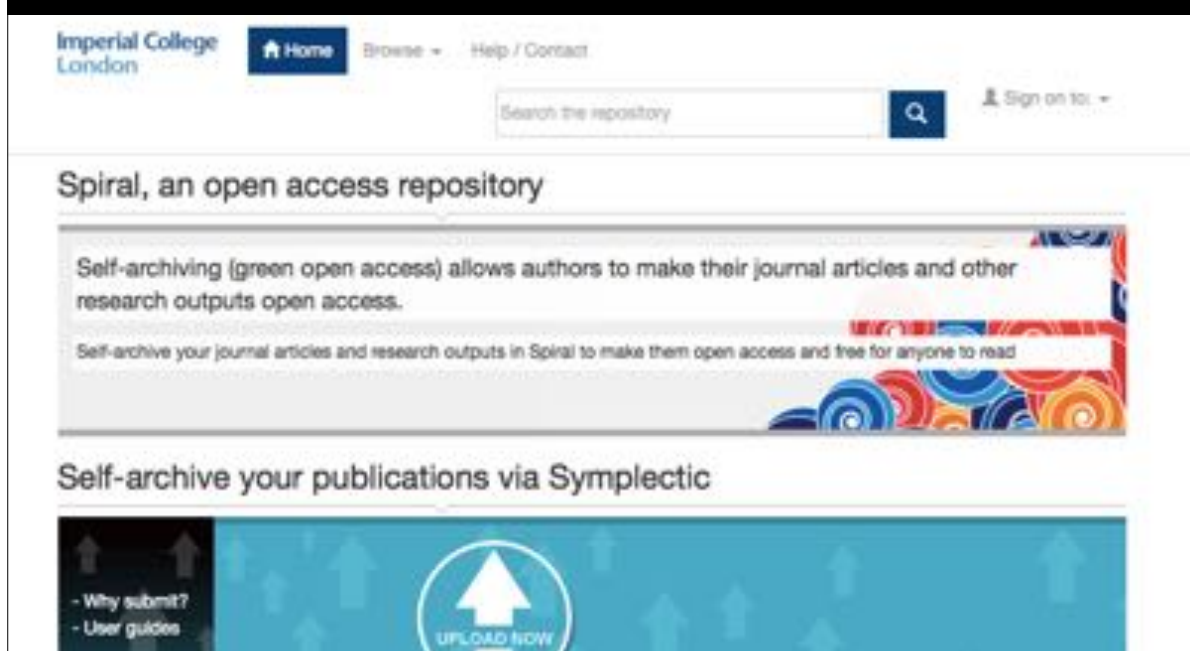
Filters Find a repository...

dryad-repo Java ★ 3 221
Forked from DSpace/DSpace
Dryad is a repository for data underlying scientific literature. Dryad allows investigators to validate published findings, explore new analysis methodologies and repurpose the data for research questions unanticipated by the original authors. The core code of Dryad comes from the DSpace repository software. Developers using the Dryad code should first familiarize themselves with DSpace. Documentation specific to the Dryad codebase can be found in the Dryad wiki.
Updated yesterday at 8:33am

vagrant-dryad SQL ★ 0 0
Vagrant and Ansible config for building a Dryad VM

[HTTPS://SPIRAL.IMPERIAL.AC.UK](https://spiral.imperial.ac.uk)

SPIRAL - IMPERIAL COLLEGE



The screenshot shows the Imperial College Spiral website. At the top, the Imperial College London logo is on the left, and navigation links for Home, Browse, and Help / Contact are on the right. A search bar and a Sign on to: button are also present. The main heading is "Spiral, an open access repository". Below this, there's a section about self-archiving (green open access) and a section about self-archiving via Symplectic. The bottom section features a large blue button labeled "UPLOAD NOW" with a white arrow icon.

Imperial College London Home Browse Help / Contact

Search the repository Sign on to:

Spiral, an open access repository

Self-archiving (green open access) allows authors to make their journal articles and other research outputs open access.

Self-archive your journal articles and research outputs in Spiral to make them open access and free for anyone to read

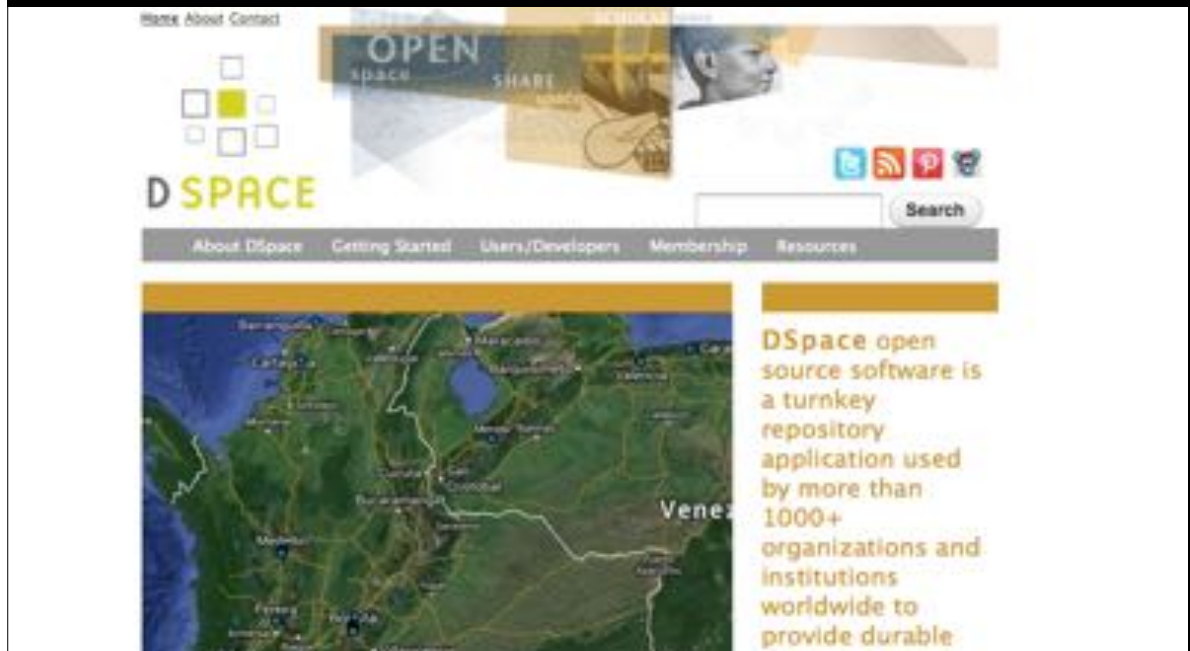
Self-archive your publications via Symplectic

Why submit? User guides

UPLOAD NOW

[HTTP://WWW.DSPACE.ORG](http://www.dspace.org)

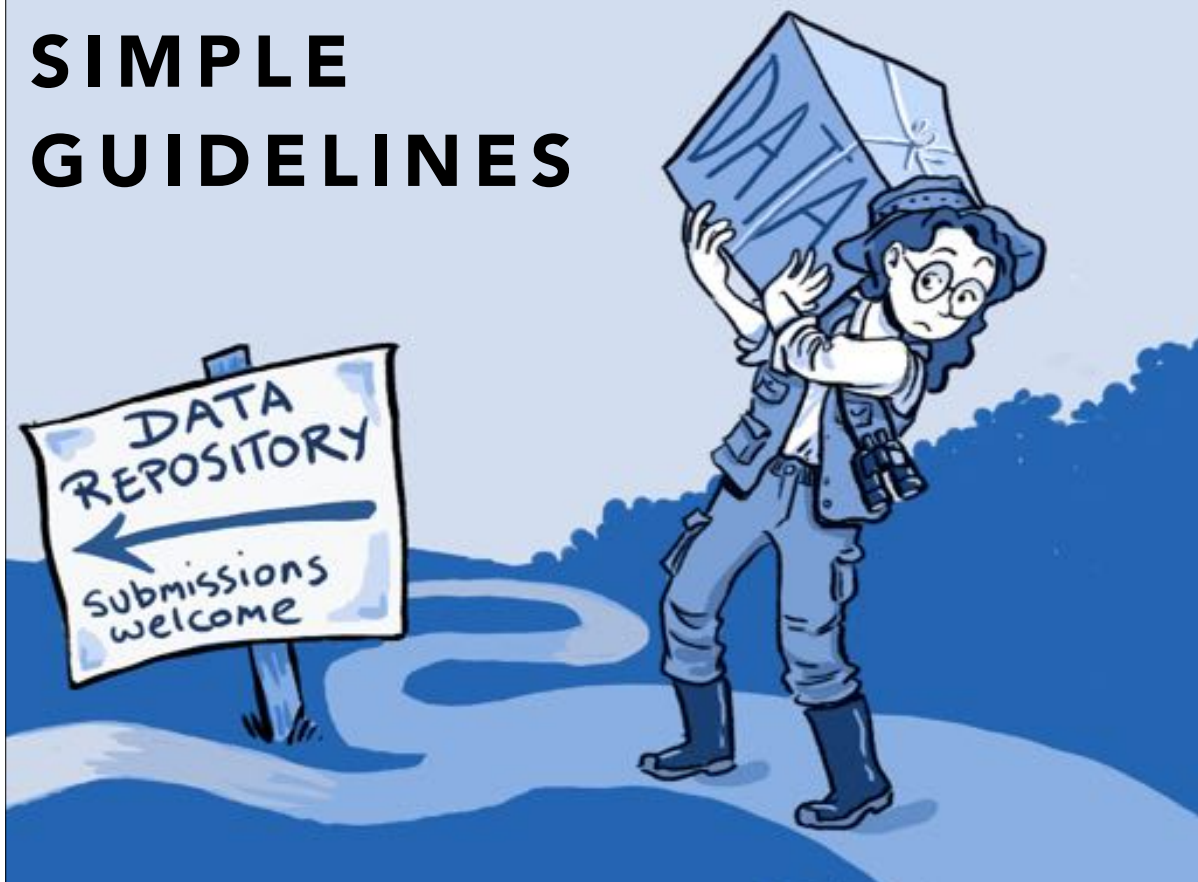
DSPACE



**SOUNDS
GOOD...
NOW
WHAT?**



SIMPLE GUIDELINES



3 THINGS

- Keep it simple
- Engage early and often
- Address common fears and misunderstandings



4 STEPS

- Choose your dataset(s)
- Licensing
- Make the data available
- Make it discoverable



DATA SETS

- Asking the community
- Cost basis
- Ease of release
- Observe peers



Data that doesn't explicitly have an open license is
NOT open data

LICENSING



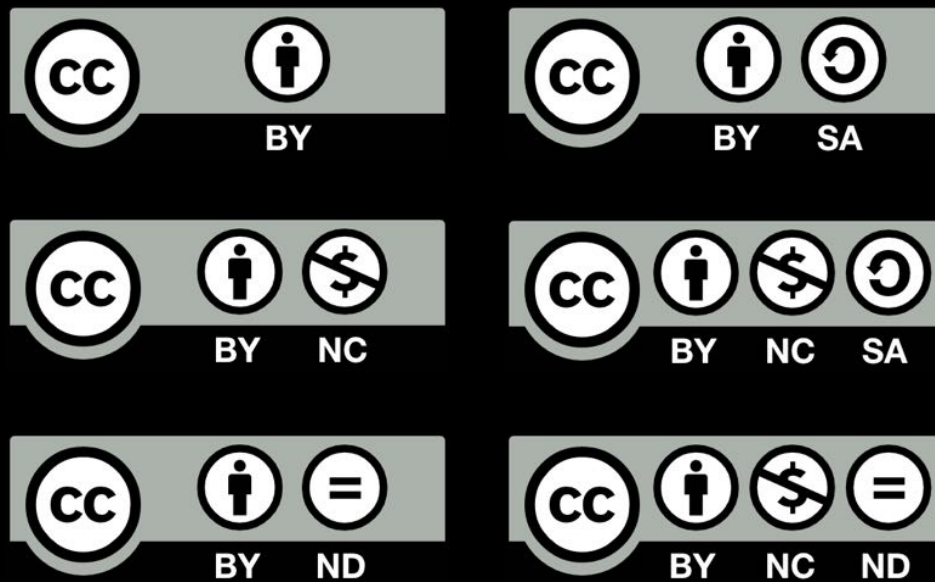
OWNERSHIP

COPYRIGHT OVER
WORKS YOU CREATE
AND ARE ORIGINAL
TO YOU.

DATABASE RIGHT
OVER COLLECTIONS
OF DATA YOU HAVE
PUT A SUBSTANTIAL
EFFORT INTO
OBTAINING,
VERIFYING OR
PRESENTING (**ONLY**
EU, MEXICO, BRAZIL)

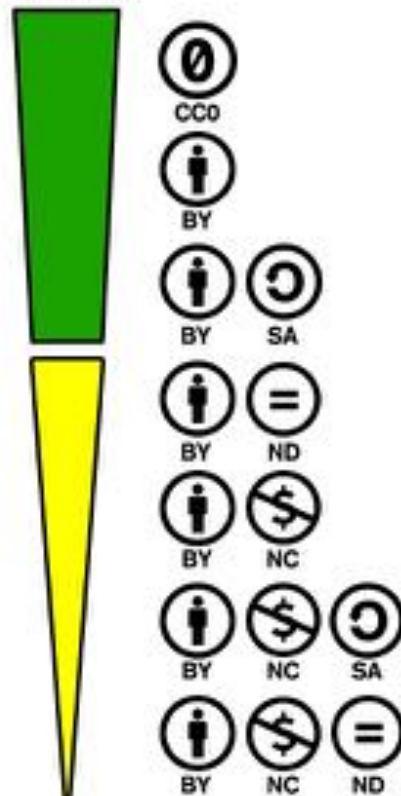


CREATIVE COMMONS LICENSING



KNOW THE
TYPES!!

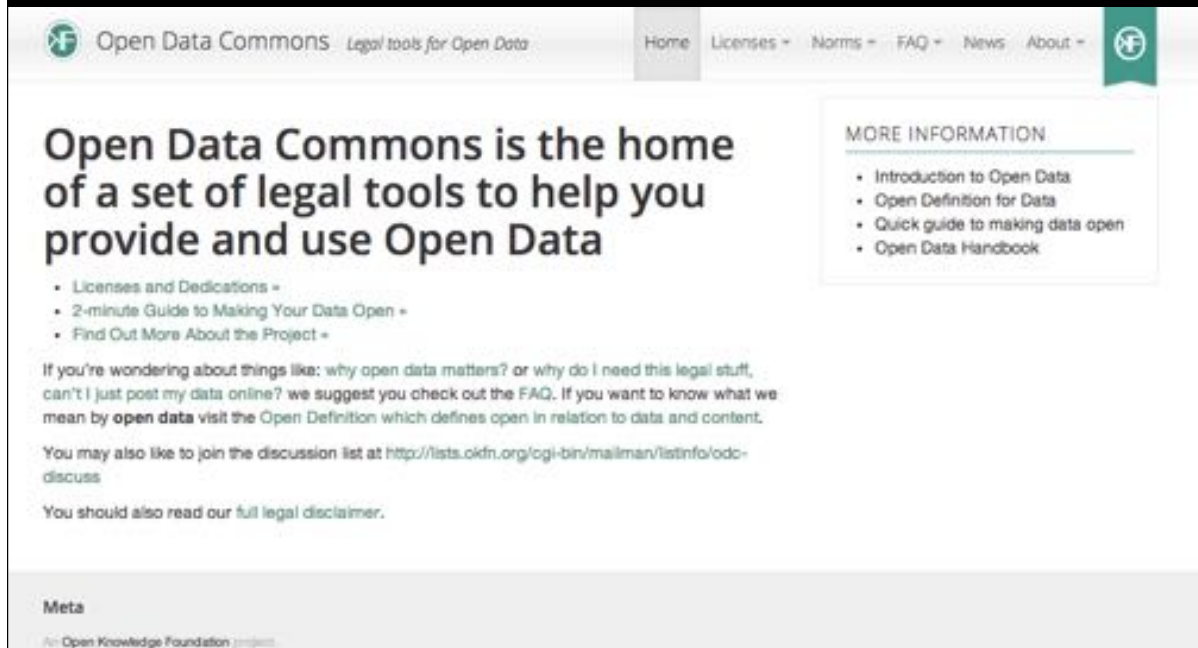
MOST OPEN



LEAST OPEN

[HTTP://OPENDATACOMMONS.ORG](http://opendatacommons.org)

OPEN DATA COMMONS



The screenshot shows the Open Data Commons website. At the top, there's a navigation bar with links: Home, Licenses, Norms, FAQ, News, and About. The main heading reads "Open Data Commons is the home of a set of legal tools to help you provide and use Open Data". Below this, there are three bullet points: "Licenses and Dedications", "2-minute Guide to Making Your Data Open", and "Find Out More About the Project". A paragraph follows, explaining that if you're wondering about things like why open data matters or why you need legal stuff, you should check out the FAQ. It also mentions the Open Definition and provides a link to a discussion list. At the bottom, there's a "Meta" section with a link to the Open Knowledge Foundation project.

Open Data Commons Legal tools for Open Data

Home Licenses Norms FAQ News About

Open Data Commons is the home of a set of legal tools to help you provide and use Open Data

- Licenses and Dedications
- 2-minute Guide to Making Your Data Open
- Find Out More About the Project

If you're wondering about things like: why open data matters? or why do I need this legal stuff, can't I just post my data online? we suggest you check out the FAQ. If you want to know what we mean by **open data** visit the Open Definition which defines open in relation to data and content.

You may also like to join the discussion list at <http://lists.okfn.org/cgi-bin/mailman/listinfo/odc-discuss>

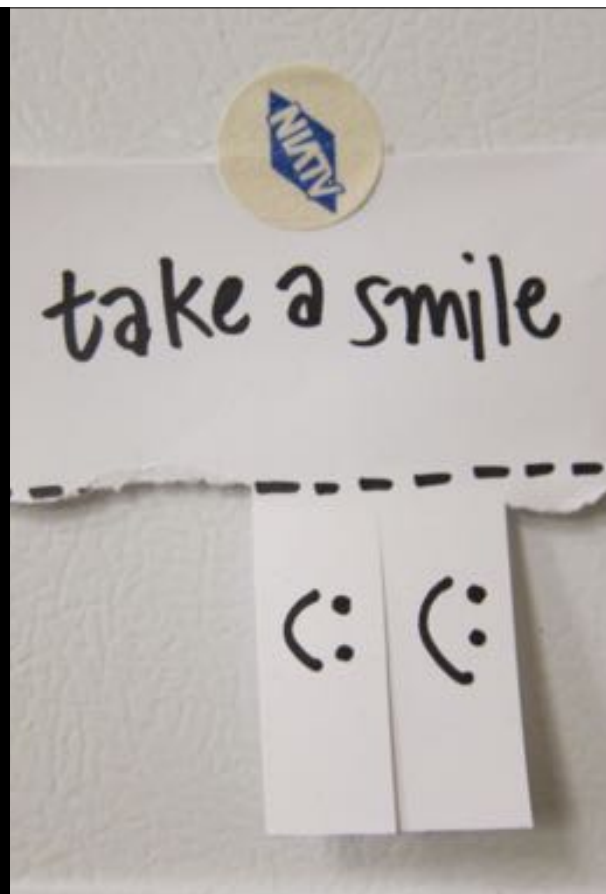
You should also read our full legal disclaimer.

Meta

As an Open Knowledge Foundation project

AVAILABILITY

- Data should be complete
- In a (open) machine-readable format
- It should contain metadata



HOW?

- Your website
- Existing repositories
- Creating your own repository



MAKE IT DISCOVERABLE

- Publish it in Public services (Datahub)
- Index it in Catalog (Databib)
- Promote it in your community
- Engage with users



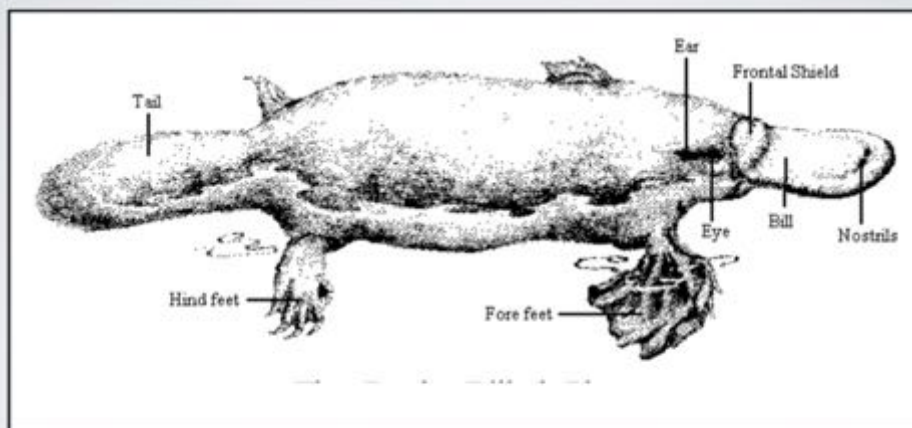


WHAT IS DATA SCIENCE

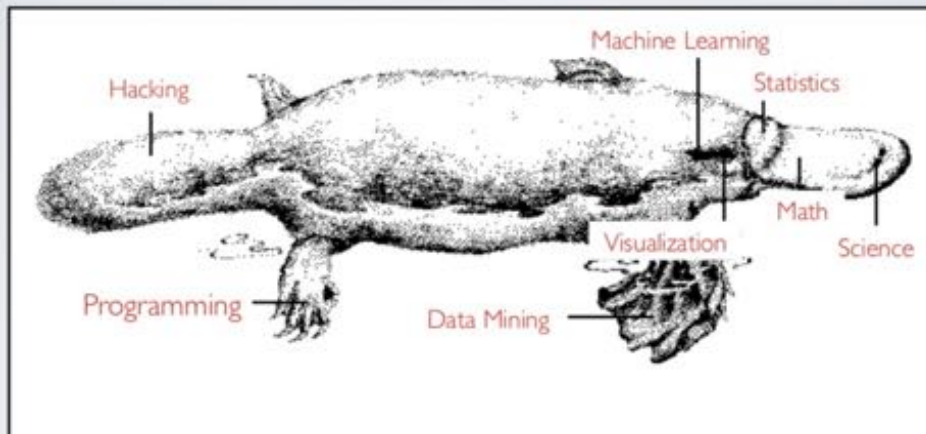
A set of tools and techniques used to extract useful information from data.

An interdisciplinary, problem-oriented subject.

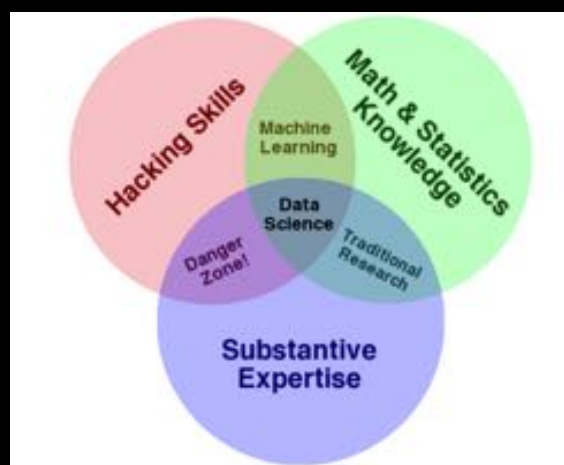
The Duck – Billed Platypus



The Platypus – Billed Data Scientist



THE INGREDIENTS OF A DATA SCIENTIST



ONE MORE
THING!

COMMUNICATION
SKILLS

STATISTICIAN



What my friends think I do



What my mom thinks I do



What society thinks I do



What my boss thinks I do



What I think I do



What I actually do

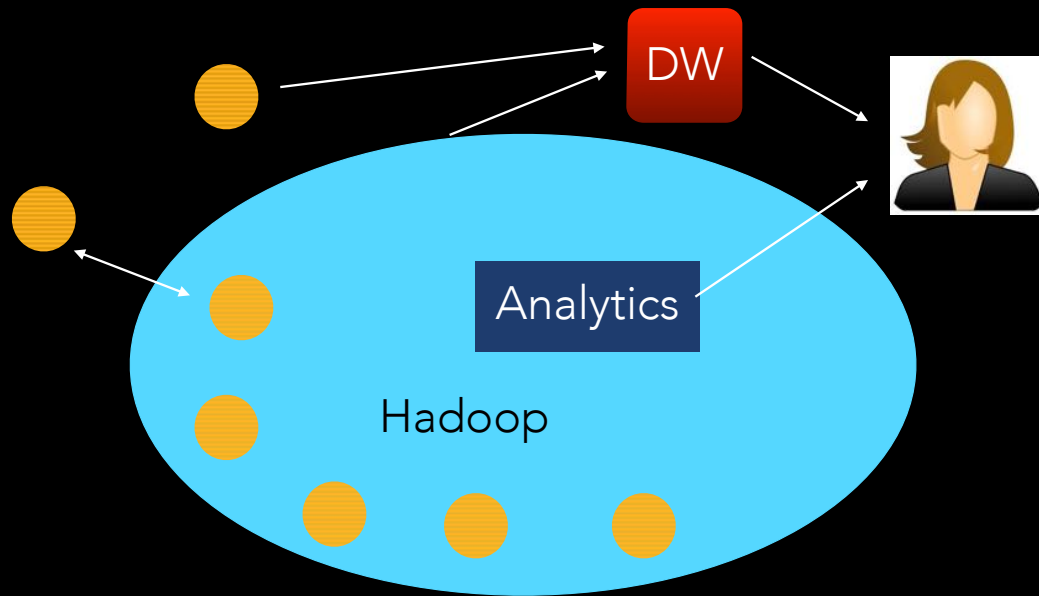
Class *DataScientist* {

- Is skeptical, curious. Has inquisitive mind
- Knows Machine Learning, Statistics, Probability
- Applies Scientific Method. Runs Experiments
- Is good at Coding & Hacking
- Able to deal with IT Data Engineering
- Knows how to build data products
- Able to find answers to *known unknowns*
- Tells relevant business stories from data
- Has Domain Knowledge

}

DATA LAKE

WHAT IS IT?



THANKS!

YOUR THOUGHTS?

DR J ROGEL-SALAZAR

J.ROGEL@PHYSICS.ORG

@QUANTUM_TUNNEL / @HIDDEN_NODE

