



Using Open Data

Dr David Tarrant | @davetaz | The Open Data Institute



Agenda

Discovering open data

Quality and provenance

Data analysis and visualisation

Open data in policy cycles

Referencing data

Agenda

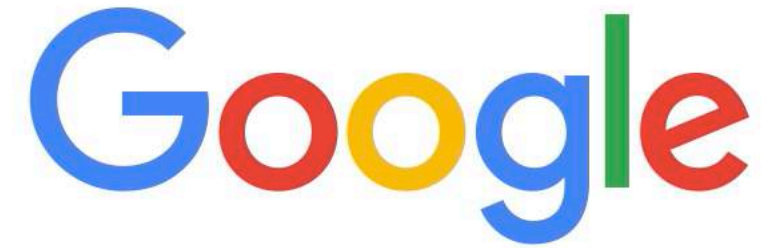
Discovering open data

Quality and provenance

Data analysis and visualisation

Open data in policy cycles

Referencing data

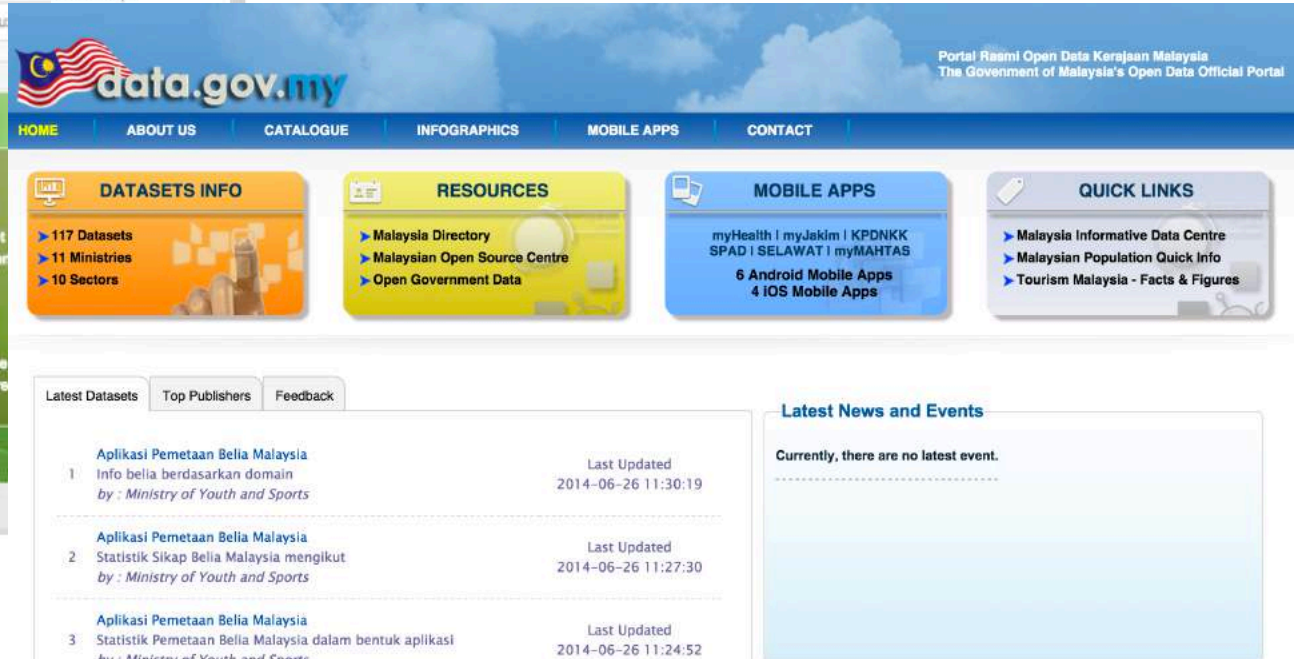
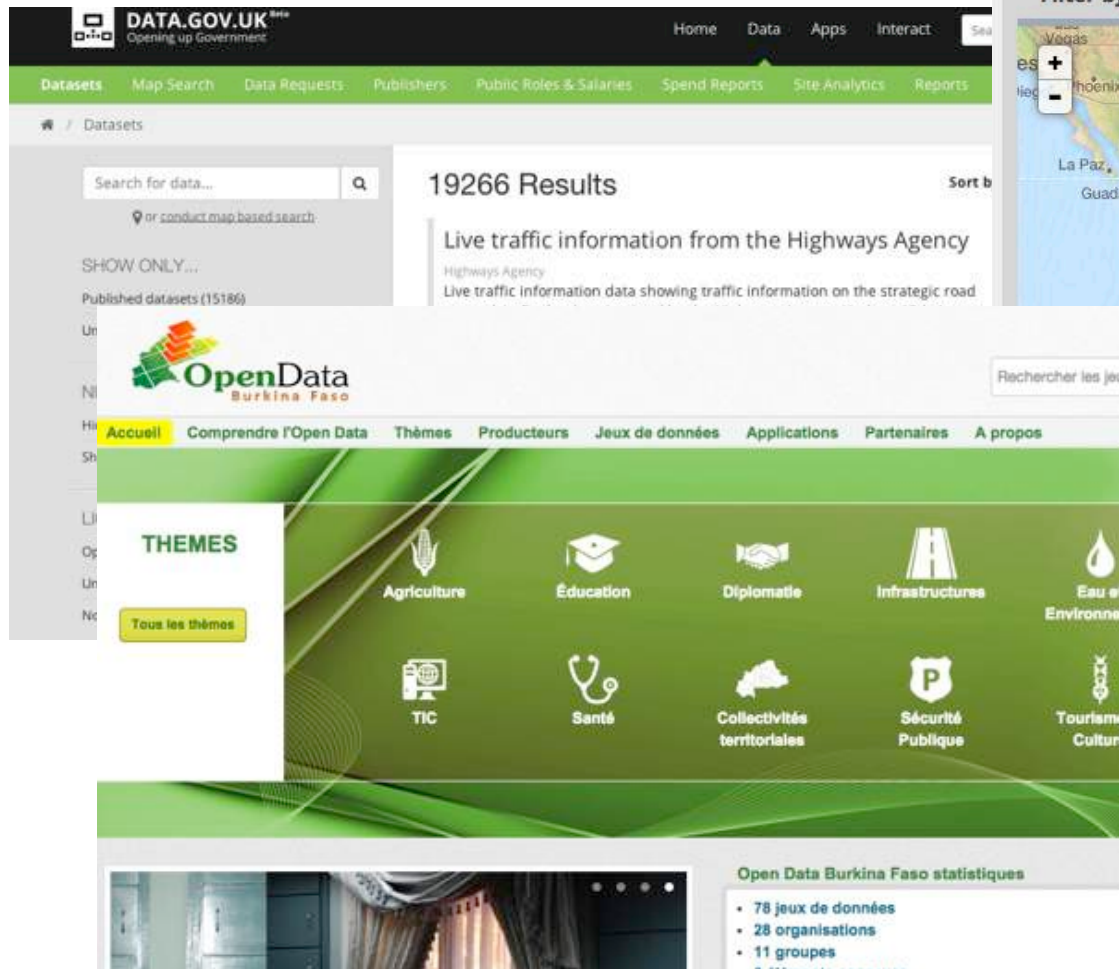
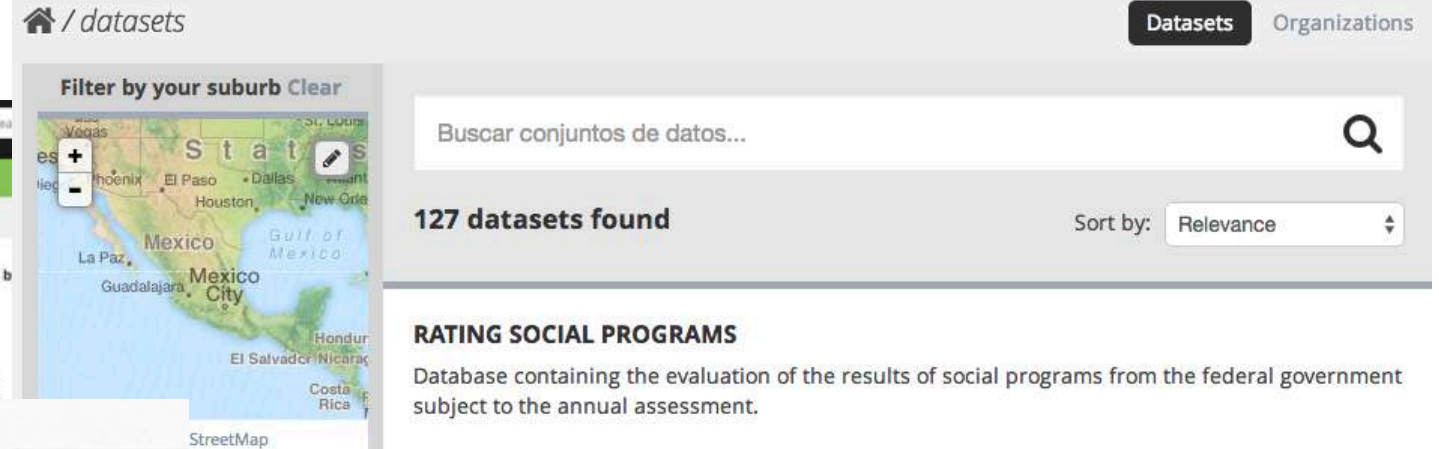
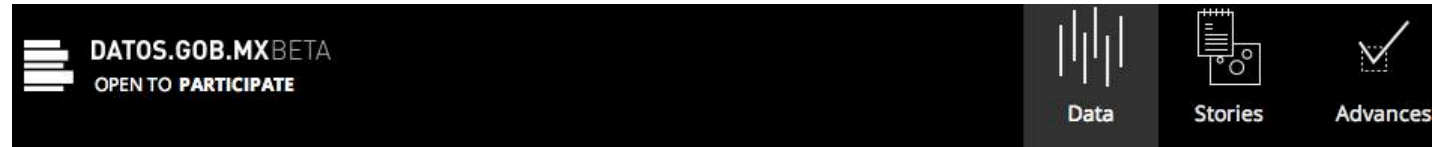


How do you find open data on the web?

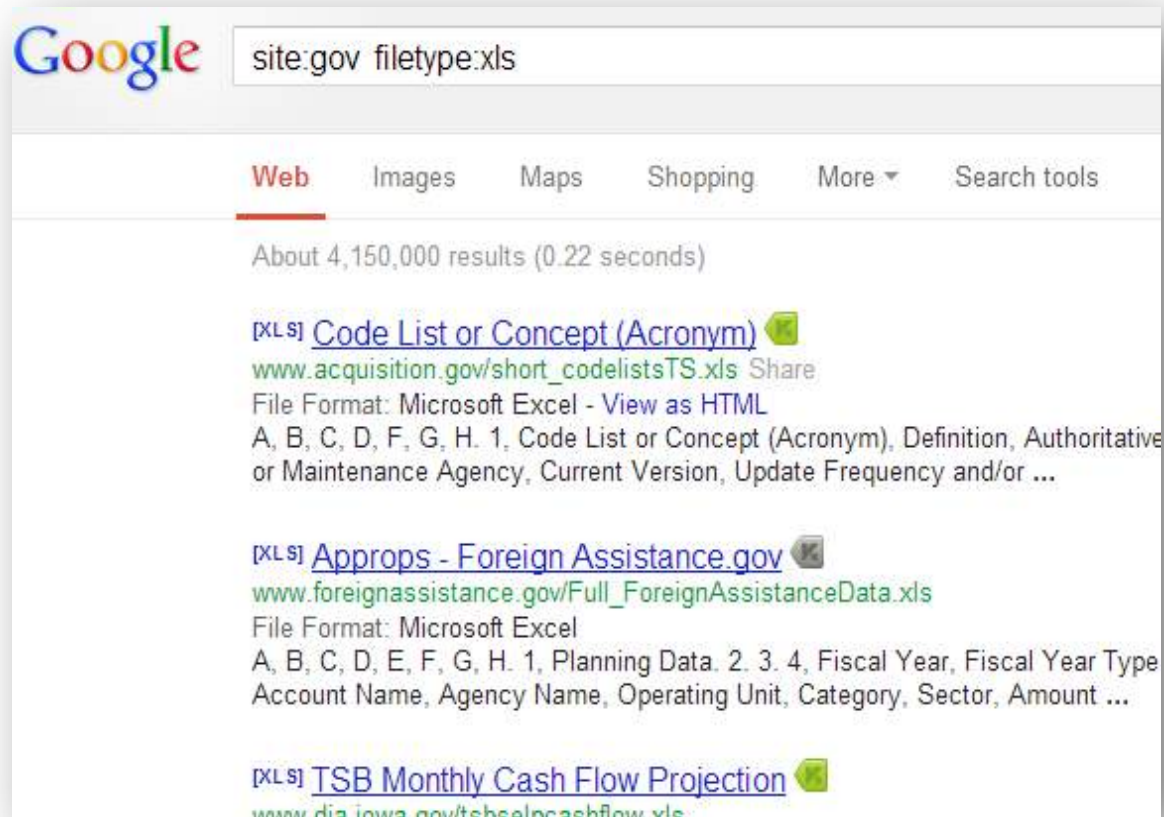
Google Search

I'm Feeling Lucky

data.gov.XX



Google advanced



site: Get results only from certain sites or domains

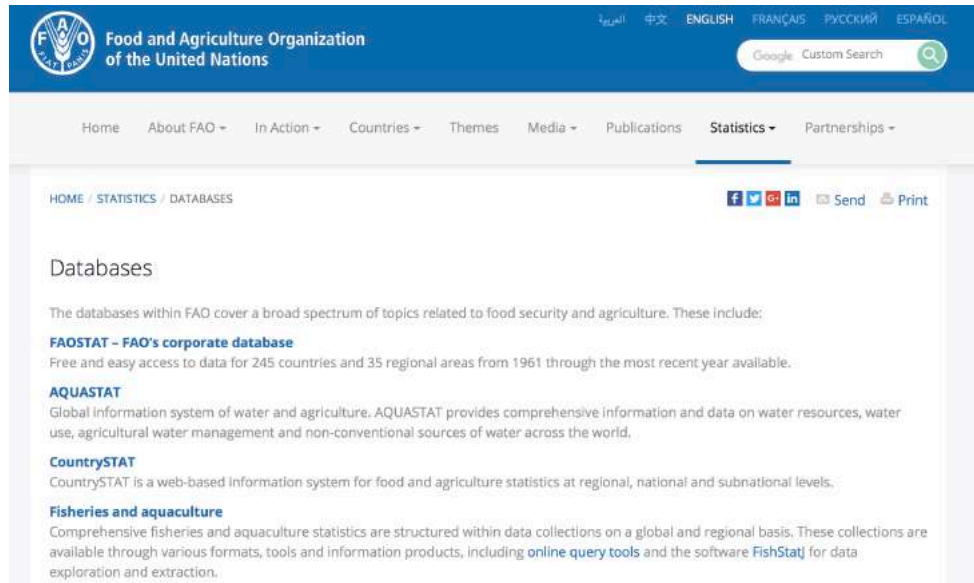
link: Find pages that link to a certain page

related: Find sites similar to one you already know

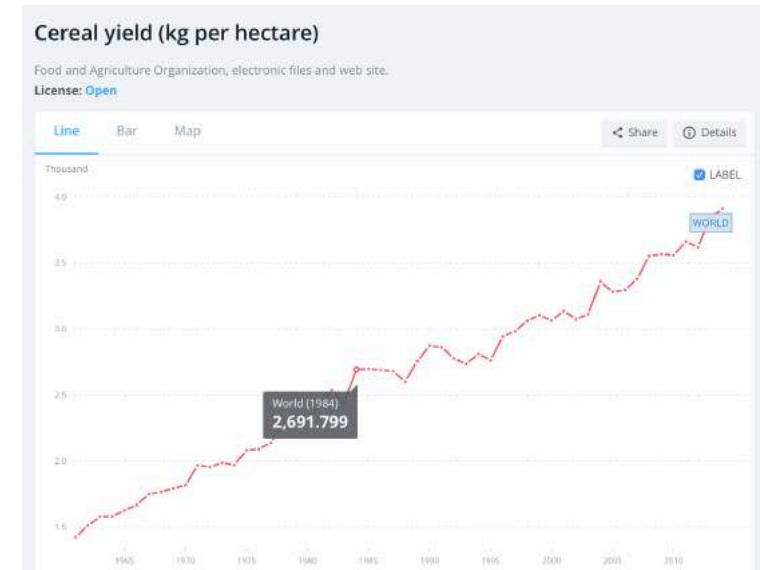
filetype: Find certain file types only

Aggregators and portals

Collect together data from across the web into one place.



FAO



World Bank

Scraping

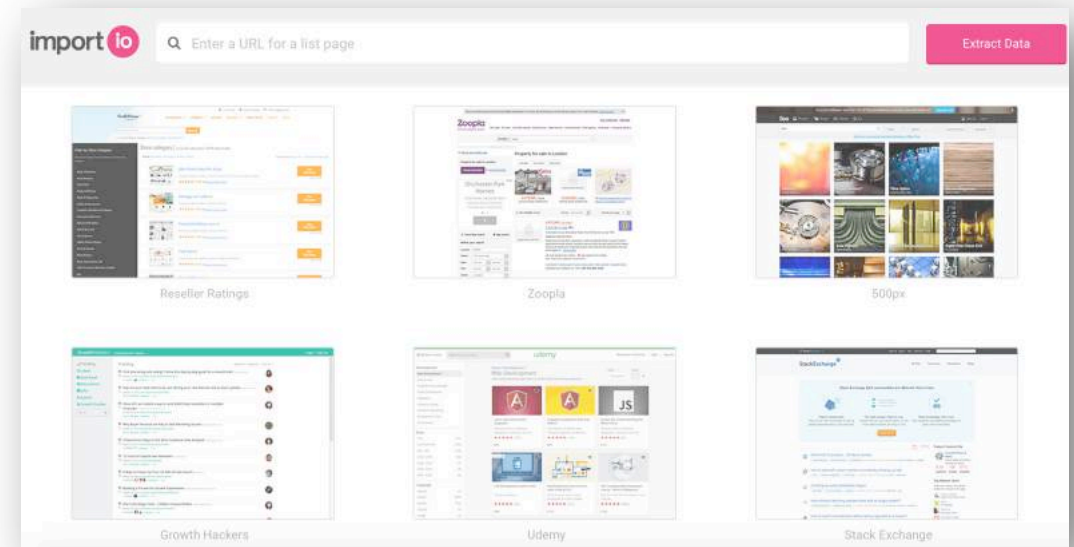
If you can't obtain usable data (csv, xls) then you may have to resort to scraping.



	A	B	C	D	E
1	Country Name	Total Receipts	Total Receipts	Total Invoices	Receipts
2	Algeria	49,043,476	48,903,632	560,807	2.61%
3	Aruba	6,896,444	6,812,239	196,276	0.87%
4	Burkina Faso	144,895,523	147,382,233	1,206,402	7.70%
5	Burkina Faso	253,647,898	253,329,006	1,335,351	11.47%
6	Cameroon	171,482,607	170,846,761	1,537,020	9.19%
7	Cameroon	158,436	158,436	1,800	0.00%
8	France	135,175,477	135,346,890	1,854,841	1.66%
9	Germany	36,907	36,907	1,191	0.00%
10	Guatemala	147,616	148,480	1,812	0.00%
11	La Grand	9,797	9,797	201	0.00%
12	Luxembourg	7,164,450	7,165,706	170,219	0.39%
13	Malawi	438,442	438,442	9,706	0.00%
14	Morocco	57,600	57,768	846	0.00%
15	Mozambique	17,365,270	17,174,884	70,947	0.92%

“excellent, so excited beyond description”
George Ofosu, Doctoral Student, UCLA

pdftables.com



magic.import.io



TANZANIA NATIONAL BUREAU OF STATISTICS

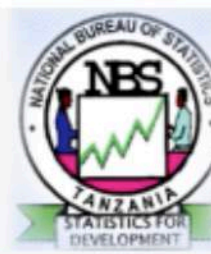
STATISTICS FOR DEVELOPMENT



English



Swahili



[Home](#) [About Us](#) [Standards](#) [Data Portal](#) [TSMP](#) [Census & Surveys](#) [Publications](#) [Data Visualization](#) [Statistics by Subject](#)

eGDDS

Agriculture Statistics

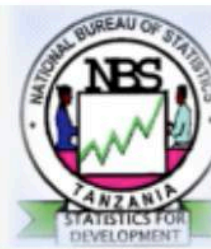
Title	Hits
2014/15 Annual Agricultural Sample Survey (Crops and Livestock Report)	3138
2014/15 Annual Agricultural Sample Survey - Crops Report	2608
Tanzania Agriculture Statistics Strategic Plan (2014/15– 2018/19)	2635
Large Scale Farms Survey 2012/13 Report	1412
The Regional Report(S) For 2007/2008 Agriculture Sample Census	1528
Large Scale Farms - 2007/08 Agriculture Sample Census	596
Crop Sector - National Report - Census of Agriculture 2007/2008	958
Zanzibar Crop Sector Report - 2007/08 Agriculture Sample Census	408

www.nbs.go.tz/nbstz/index.php/english/statistics-by-subject/agriculture-statistics/803-2014-15-annual-agricultural-sample-survey-crops-and-livestock-report



TANZANIA NATIONAL BUREAU OF STATISTICS

STATISTICS FOR DEVELOPMENT



[Home](#) [About Us](#) [Standards](#) [Data Portal](#) [TSMP](#) [Census & Surveys](#) [Publications](#) [Data Visualization](#) [Statistics by Subject](#)

eGDDS

2014/15 Annual Agricultural Sample Survey (Crops and Livestock Report)

The National Bureau of Statistics (NBS) in collaboration with the Ministry of Agriculture, Livestock and Fisheries, Ministry Industry, Trade and Investment; the President's Office, Regional Administration and Local Government (PO/RALG) and the Office of the Chief Government Statistician, (OCGS), Ministry of Agriculture and Natural Resources Zanzibar; Ministry of Livestock and Fisheries Zanzibar; conducted the Annual Agricultural Sample Survey (AASS) in 2014/15 Agricultural Year.

The survey collected detailed information on area and production for major crops, potentially livestock and fish farming. This is the first survey of its kind conducted in Tanzania using Point Sample Area Frame Methodology, therefore it limits the comparisons with the previous agriculture censuses and surveys due to different methodologies applied.

The extensive nature of the survey in relation to its scope and coverage is a result of the increasing demand for more timely, accurate estimates of area and production for major crops, livestock and fish farming.

It is expected that, the survey report will provide new insights for planners, policy makers, researchers and others involved in the agricultural sector in order to improve the prevailing conditions faced by agricultural operators in the country.

[Click Here to download the Crops and Livestock Report.](#)

Table 13: Total Number of Cattle by Category and Region as of 1st October 2015

Region	Total Number of Cattle							Total
	Oxen	Bulls	Cows	Steers	Heifers	Male Calves	Female Calves	
Dodoma	212,902	199,182	406,984	78,435	299,142	176,716	140,346	1,513,707
Arusha	108,716	163,772	386,588	151,942	156,490	214,719	239,736	1,421,963
Kilimanjaro	66,436	143,538	244,257	38,414	100,490	78,739	99,017	770,891
Tanga	22,409	71,327	300,792	42,790	95,006	89,511	107,859	729,694
Morogoro	103,925	76,252	179,370	28,785	53,078	64,380	100,498	606,288
Pwani	74,802	85,471	165,944	16,344	71,592	76,456	101,711	592,320
Dar-es-salaam	1,141	786	7,154	S	S	3,137	4,248	21,096
Lindi	2,632	20,927	96,631	4,987	42,032	28,234	37,674	233,117
Mtwara	S ³⁴	19,174	37,805	S	29,314	1,660	7,750	96,486
Ruvuma	S	12,267	62,611	1,563	S	10,734	31,563	120,135
Iringa	97,205	71,754	136,413	3,855	123,043	83,635	97,366	613,271
Mbeya	246,747	298,917	544,114	30,887	177,901	154,566	188,263	1,641,395
Singida	328,721	156,119	411,093	56,665	164,557	169,702	177,942	1,464,799
Tabora	503,345	309,471	799,395	180,069	222,236	358,382	371,772	2,744,670
Rukwa	174,612	83,577	169,682	18,195	46,116	63,193	77,016	632,391
Kigoma	39,620	67,030	260,946	9,416	78,685	97,073	67,367	620,137
Shinyanga	357,220	247,089	474,631	184,273	234,882	195,250	191,234	1,884,579
Kagera	42,222	72,515	313,977	40,394	127,264	106,704	139,767	842,843
Mwanza	440,126	256,976	542,032	106,439	273,614	244,862	224,670	2,088,719
Mara	192,838	378,512	491,193	82,719	316,602	209,559	216,169	1,887,592
Manyara	250,770	388,729	657,582	114,823	256,761	260,326	232,727	2,161,718
Njombe	34,081	47,947	50,353	1,039	22,524	30,833	30,690	217,467
Katavi	166,118	39,063	50,851	977	36,441	18,944	18,282	330,676
Simiyu	127,489	140,166	203,230	27,126	118,468	84,387	86,240	787,106
Geita	305,363	198,815	450,585	84,254	237,161	196,041	158,903	1,631,122
MAINLAND	3,901,084	3,549,376	7,444,213	1,305,073	3,287,883	3,017,743	3,148,810	25,654,182
Kaskazini Unguja	S	5,823	10,938	S	3,560	2,896	4,565	28,073
Kusini Unguja	655	3,891	11,656	1,500	2,888	3,078	3,312	26,980
Mjini Magharibi	431	2,876	6,817	273	4,796	2,183	2,975	20,350
Kaskazini Pemba	1,112	7,993	18,537	476	10,011	6,984	5,840	50,954
Kusini Pemba	S	4,833	11,638	S	5,373	4,043	4,472	31,664
ZANZIBAR	2,628	25,416	59,585	3,416	26,628	19,184	21,163	158,021
TANZANIA	3,903,712	3,574,792	7,503,798	1,308,489	3,314,511	3,036,927	3,169,973	25,812,203

2014_15_AASS_%20report_with_Livestock_mod

How did we do?


[DOWNLOAD AS EXCEL](#)


Page 1

Table 13: Total Number of Cattle by Category and Region as of 1st October 2015

				Total Number of Cattle					
	Region								
		Oxen	Bulls	Cows	Steers	Heifers	Male Calves	Female Calves	Total
	Dodoma	212,902	199,182	406,984	78,435	299,142	176,716	140,346	1,513,707
	Arusha	108,716	163,772	386,588	151,942	156,490	214,719	239,736	1,421,963
	Kilimanjaro	66,436	143,538	244,257	38,414	100,490	78,739	99,017	770,891
	Tanga	22,409	71,327	300,792	42,790	95,006	89,511	107,859	729,694
	Morogoro	103,925	76,252	179,370	28,785	53,078	64,380	100,498	606,288

Agenda

Discovering open data

Quality and provenance

Data analysis and visualisation

Open data in policy cycles


Referencing data

Guidelines



5 - S t a r s

Open Data Certificate



Bronze I
self certified

This data has achieved **Bronze level** on 29 Mar 2014. A great start at the basics of publishing open data.

National Public Transport Access Nodes (NaPTAN)

What's this data called? This is required for **Bronze level certification** and above

National Public Transport Access Nodes (NaPTAN)

Who publishes this data? This is required for **Bronze level certification** and above

Department for Transport

What is the website of the publisher?

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

This URL has been verified

General Information

This data is described at
<http://data.hounslow.gov.uk/View/loc>

Legal Information

This data was originally created or generated by its curator

The rights statement is at
<http://www.hounslow.gov.uk/index/c...>

This data is available under
UK Open Government Licence

There are
no rights in the content of the data

Summary

Type of release
ongoing release of a series of related datasets

Data Licence
UK Open Government Licence

Content Licence
Not Applicable

Verification
self certified

Community verification

Other people can verify whether the answers on this

Certificate level key

This shows your progress towards each data certificate. [Read how it works.](#)

[How can I improve my open data?](#)

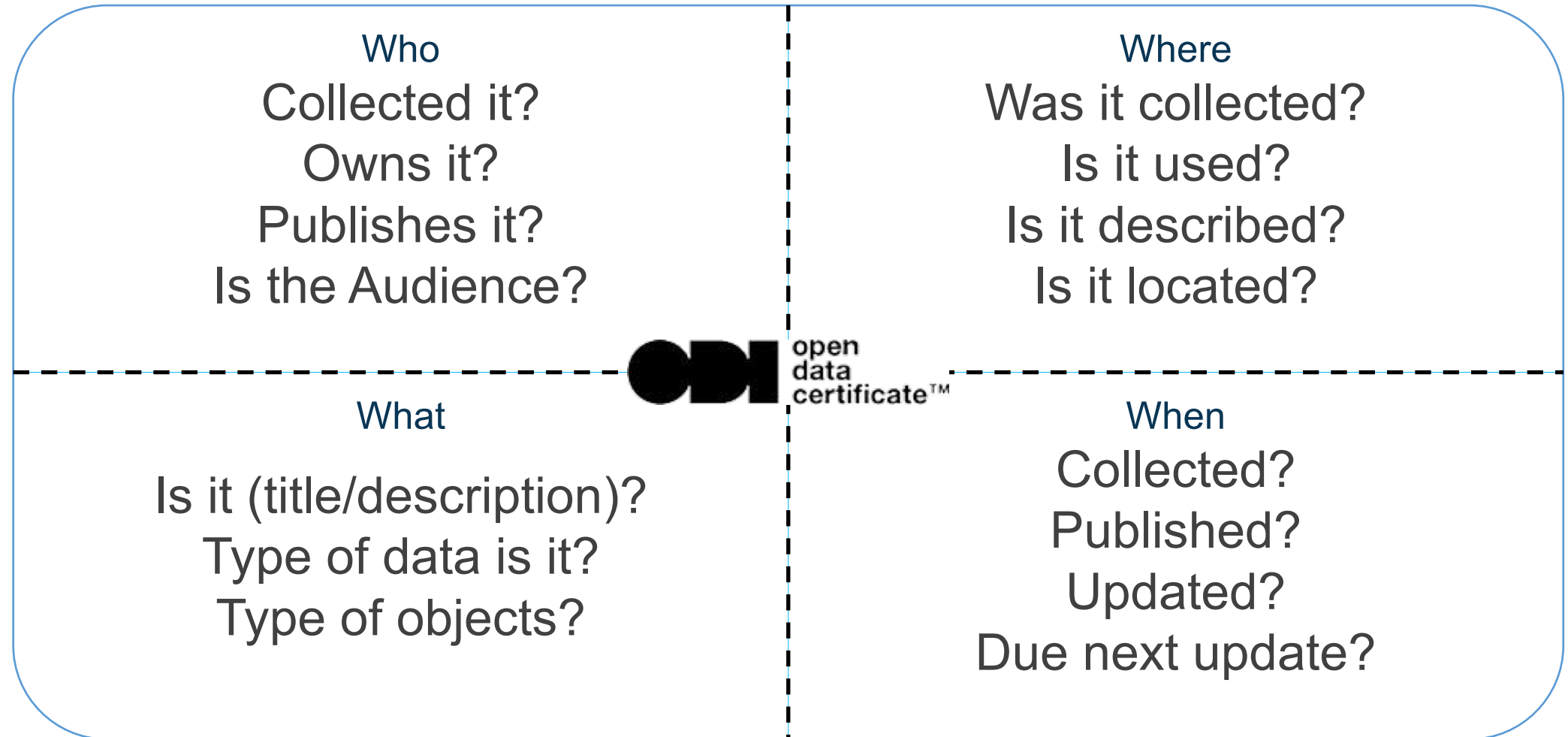
Bronze: this data makes a great start at the basics of publishing open data.

Silver: extra effort went in to support and encourage feedback from people who use this open data.

Gold: this is regularly published open data with robust support that people can rely on.

Platinum: this open data is an exceptional example of information infrastructure.

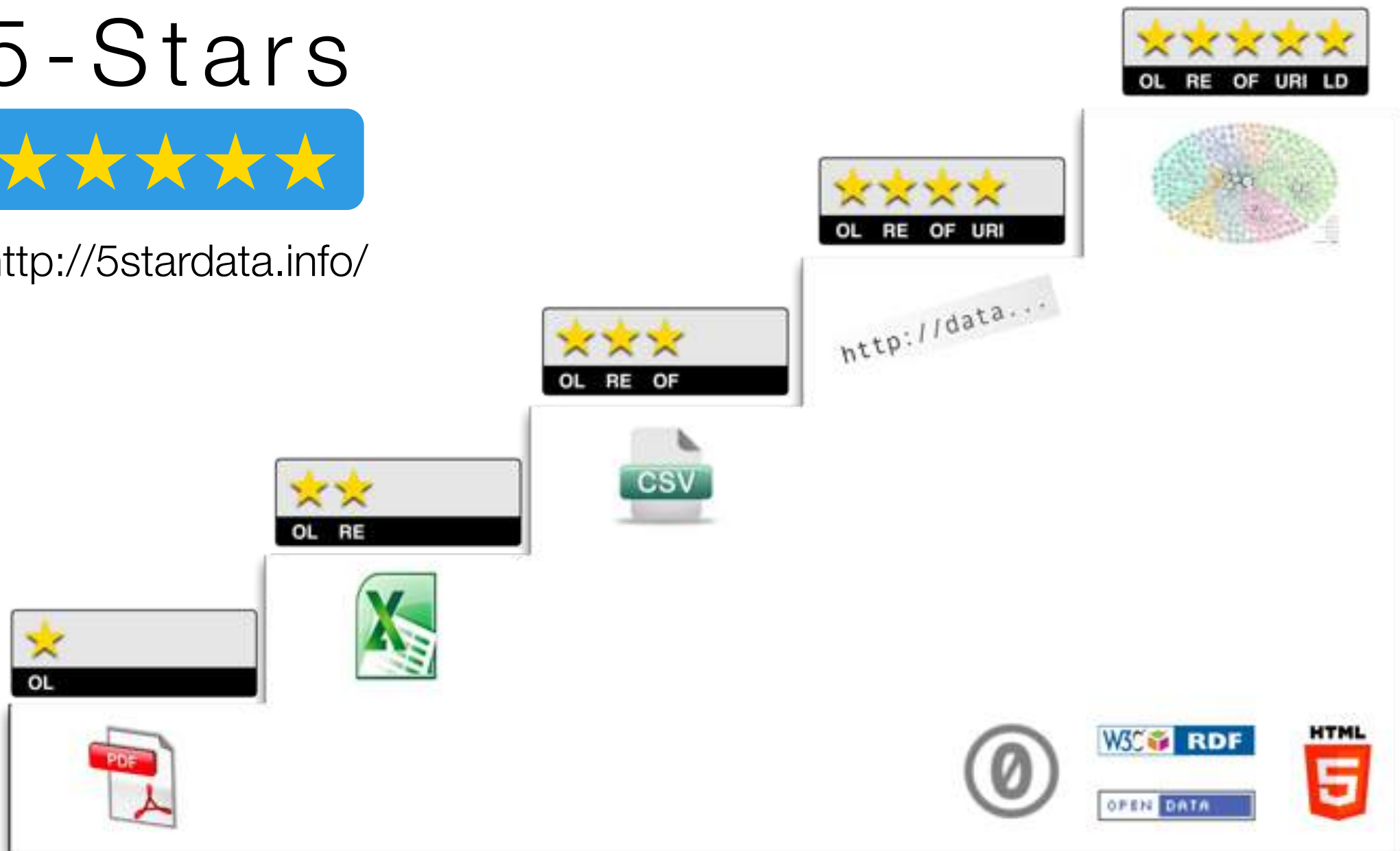
Establishing trust in data



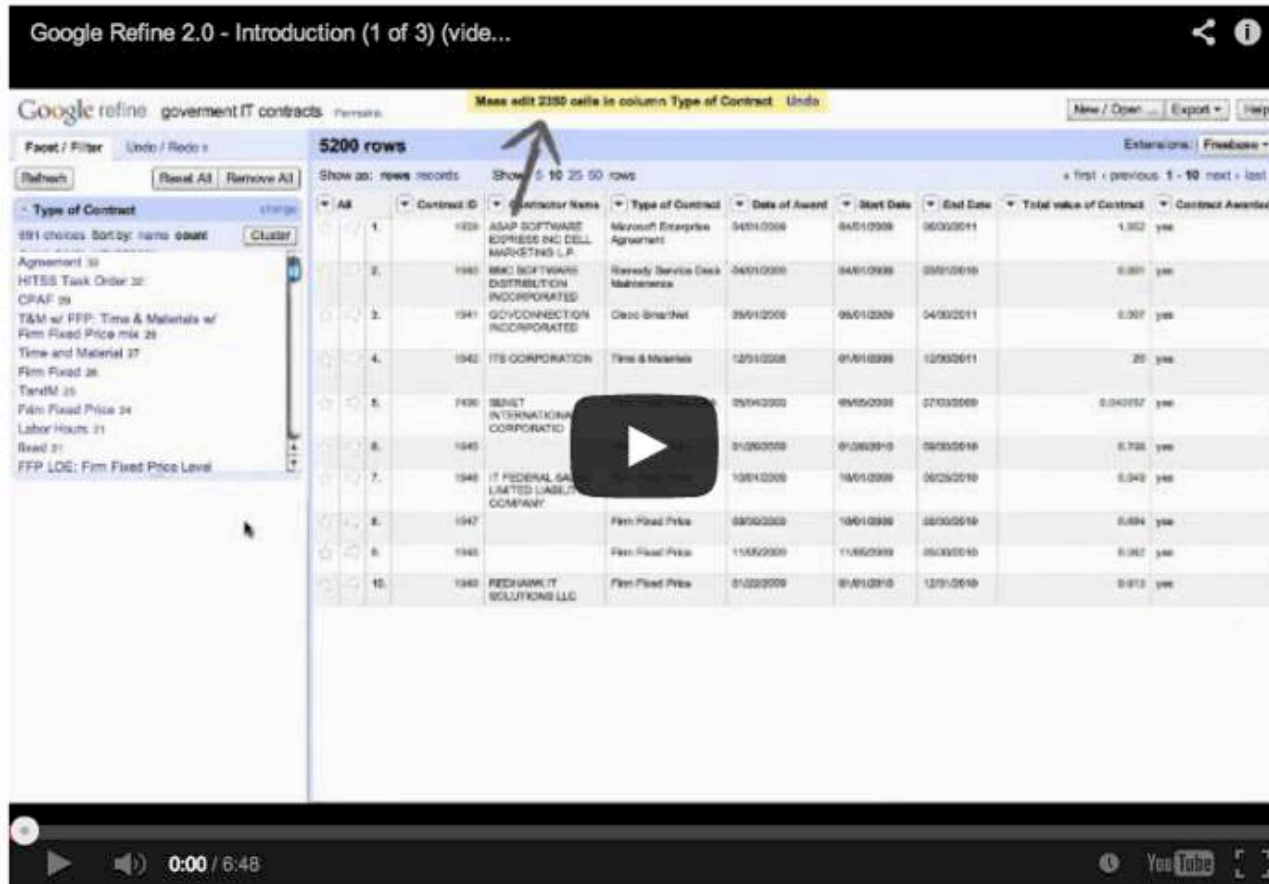
5-Stars



<http://5stardata.info/>



Open Refine



A free power tool for cleaning messy data

<http://openrefine.org>

Agenda

Discovering open data

Quality and provenance

Data analysis and visualisation

Open data in policy cycles

Referencing data

Data analysis



Quantitative

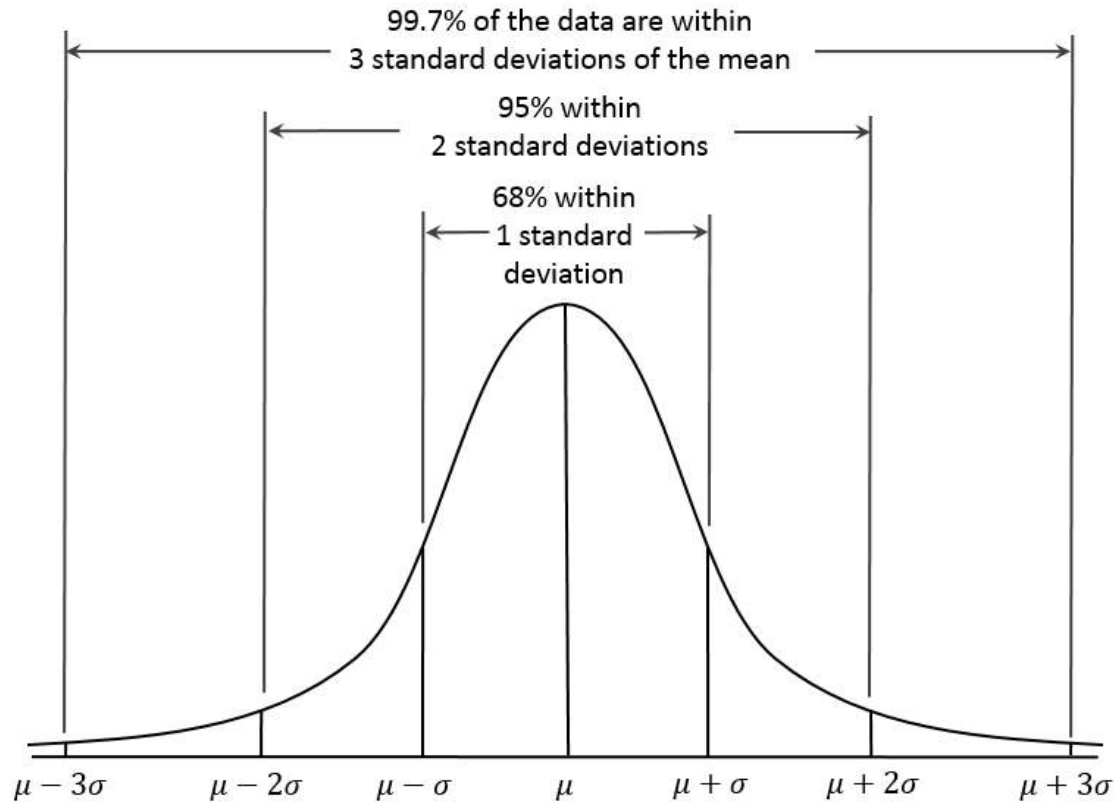
*the whole of which must be the most till the
did not of which must be the most till the
leaving you you must be the most till the
aid & do, have of this most till the
my real feeling of commence till the
ards you my approach, if all to
to offer. Caroline I am sufficiently be
tears from perceived not and must of
on that I have nervous are I have no of
in Caroline My dearest Caroline, love you
which the proof off tears, which you
the whole of I am not apt to shed, if the
not of which must have nervous mom
the whole of which must have nervous mom*

Qualitative

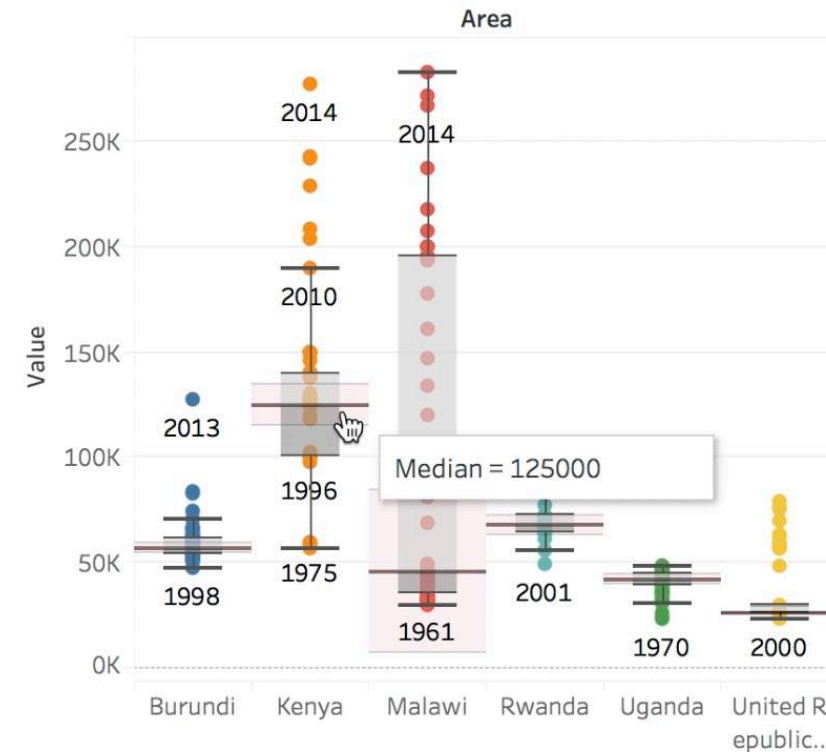
Remember

- Not all data is structured
- Not all numeric data is structured
- Some text data is structured

Analysing quantitative data

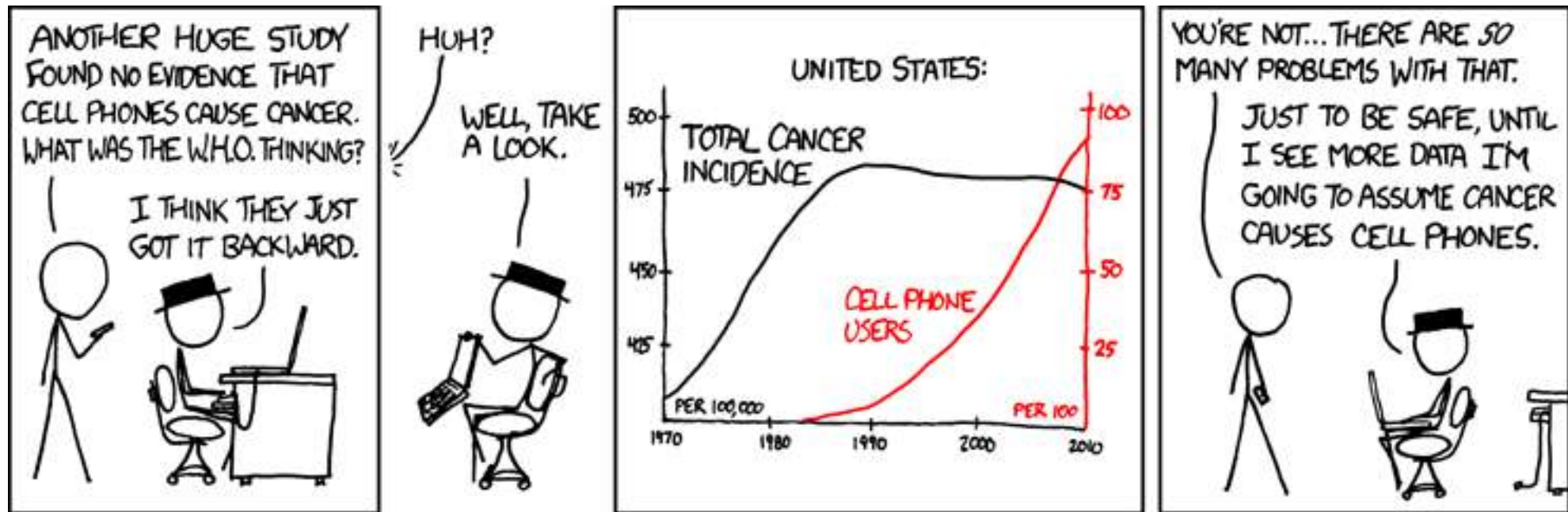


Banana yield



Beware!

- Targets
- Chance
- Fluctuation
- Correlation \neq Causation



<https://xkcd.com/925/>

Analysing qualitative data

The screenshot displays the Open Calais interface, which is used for analysing qualitative data. It is divided into two main sections: 'FOUND IN DOCUMENT' on the left and 'DOCUMENT VIEW' on the right.

FOUND IN DOCUMENT: This section shows a list of entities identified in the document, categorized by type. Each category has a checkbox and a list of specific entities with their respective relevance percentages.

Category	Entity	Relevance
ENTITIES	Irish Republic	20%
	Northern Ireland	80%
	Scotland	20%
	United Kingdom	80%
	Wales	20%
Industry Term	agri-food sector	20%
	Organization	
Organization	Department for Env...	20%
	European Union	20%
Person	Brexit	80%

DOCUMENT VIEW: This section shows the full text of the document with entities highlighted in yellow. The text is as follows:

Farmers in Northern Ireland can expect to get roughly the same amount of financial support after Brexit, the UK agriculture minister has said.

Michael Gove was speaking on a visit to a farm near Doagh in County Antrim.

Mr Gove said there would be a move away from a system where farmers were paid for the amount of land they had.

It would shift to a scheme where they were rewarded for the environmental benefits of their work, and supported to access new export markets.

Northern Ireland currently gets a larger share of EU subsidy money compared to other UK regions.

That is a reflection of the economy's reliance on the agri-food sector.

The right thing to do

Mr Gove said he had been in discussions with the authorities in Scotland and Wales, and there had been a "broad feeling that proceeding on that basis" was the right thing to do.

But he said a final UK agreement on the issue would have to be negotiated.

The Department for Environment, Food and Rural Affairs secretary said he hoped a new Northern Ireland Executive would soon be in place to facilitate that discussion.

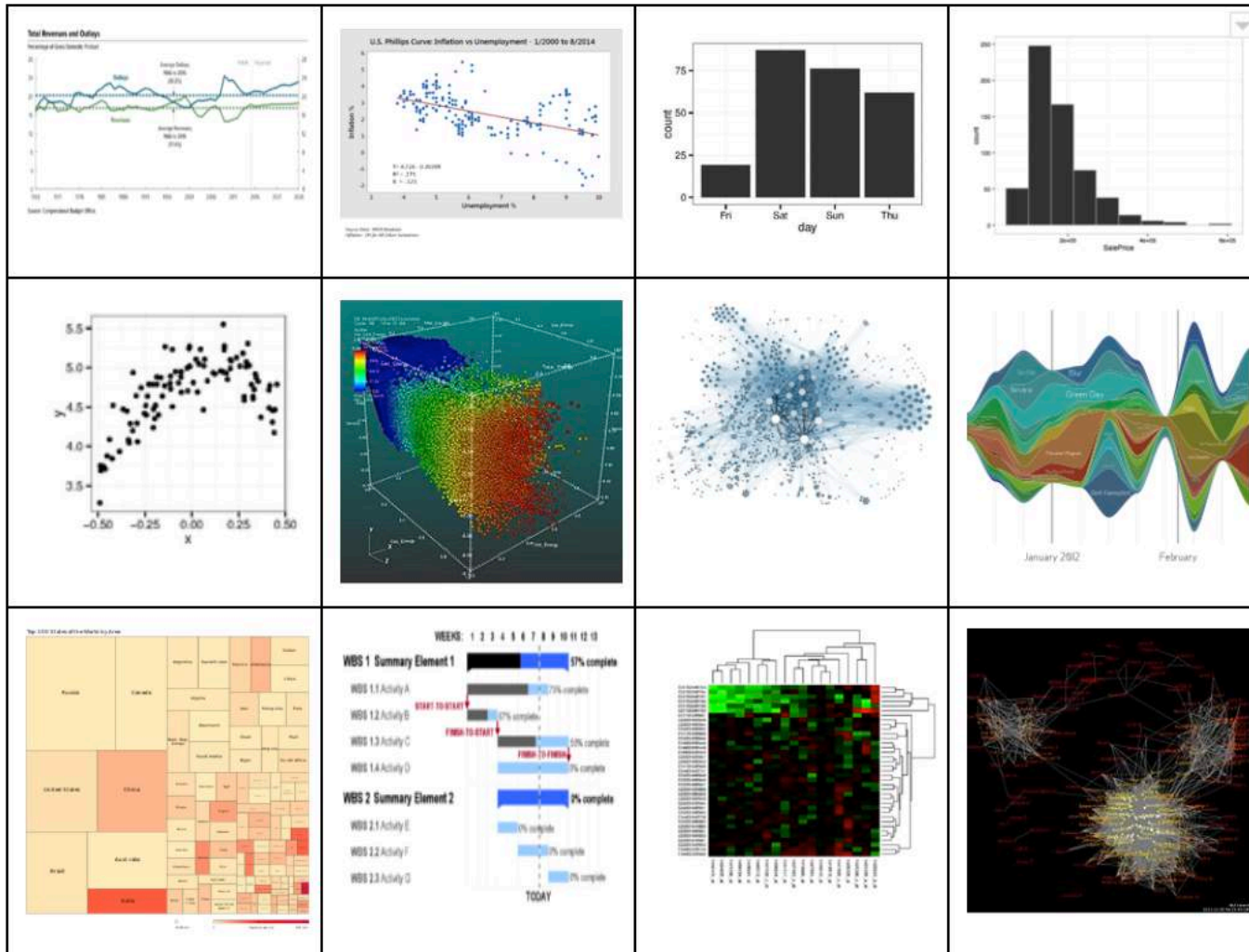
Mr Gove said his second visit to Northern Ireland in two months was aimed at hearing from those at "the sharp end" of Brexit.

That included those who traded across the border, sourcing or supplying goods to and from

Entity recognition can help with coding and thematic network analysis.

Try Open Calais
Search: open calais

Visualisaion



Not all data
visualisations are
good!

Picking the right visulisation

1) Audience

- Who are your audience and what do they expect?

2) Purpose

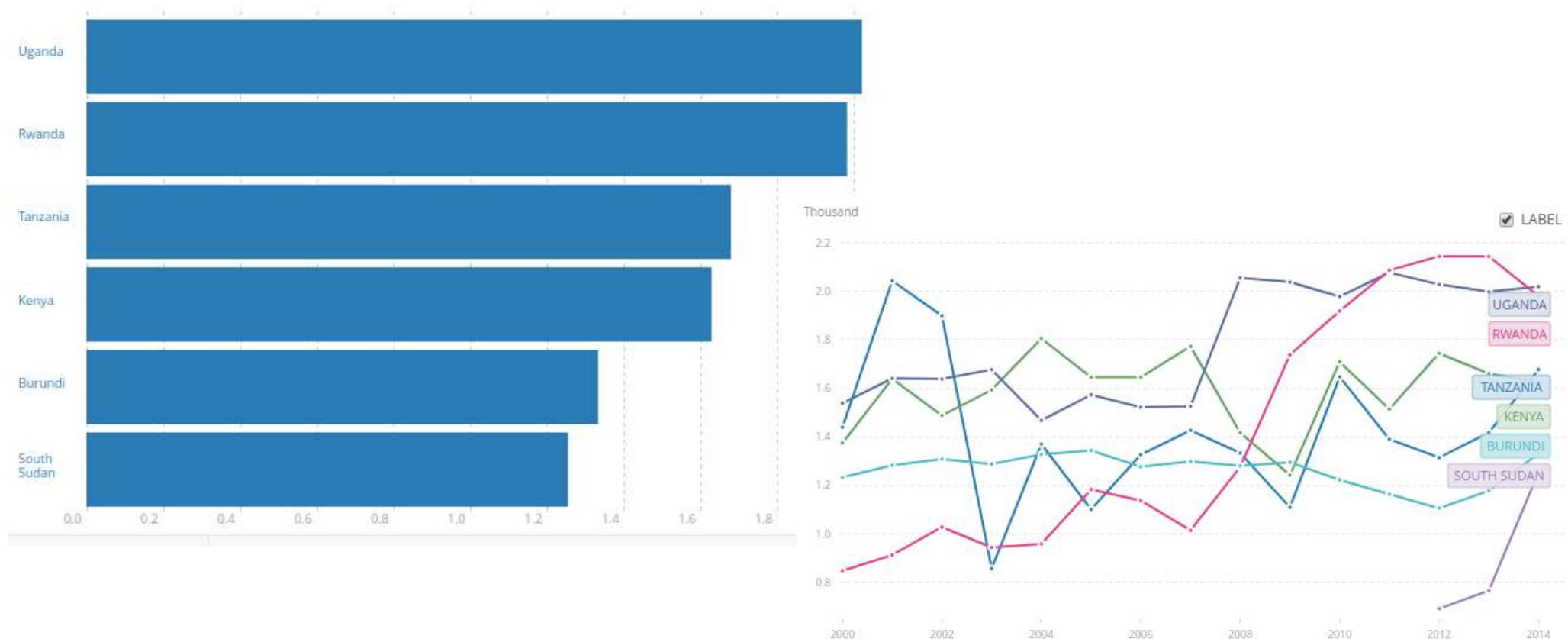
- What story are you trying to tell.

3) Data

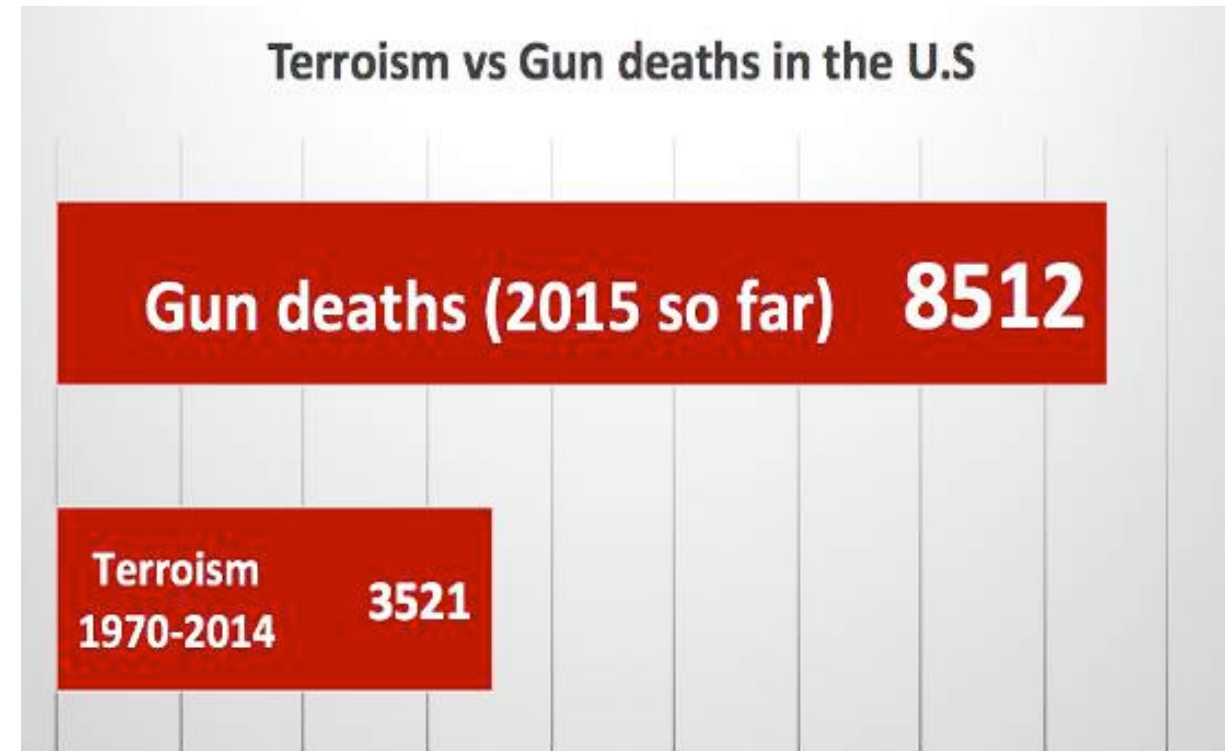
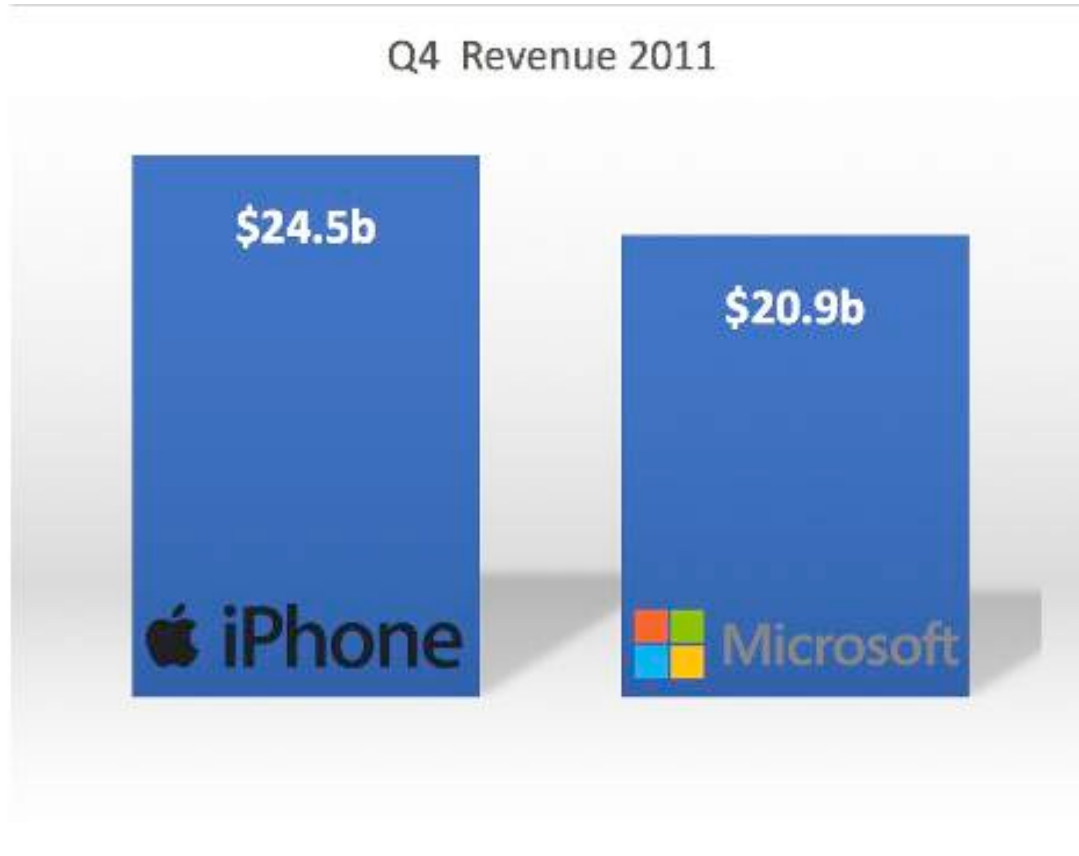
- What types of visulisation suit the data

Keep it simple!

Which country achieved the greatest crop yield in 2014?

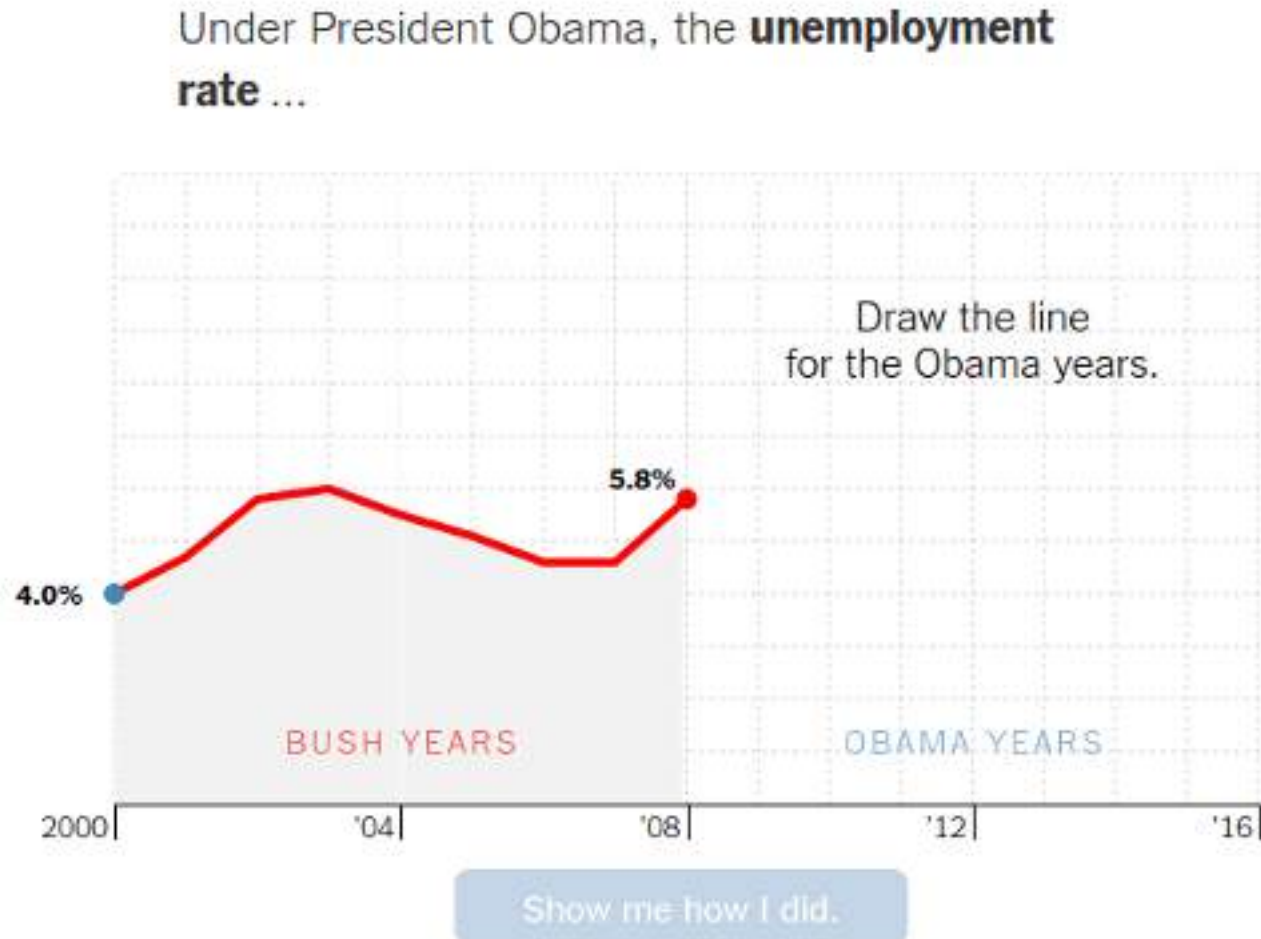


Nothing wrong with a bar chart



Observe how you don't need unnecessary clutter like axis and labels you can't read

Simple lines and interactivity



https://www.nytimes.com/interactive/2017/01/15/us/politics/you-draw-obama-legacy.html?_r=0

Agenda

Discovering open data

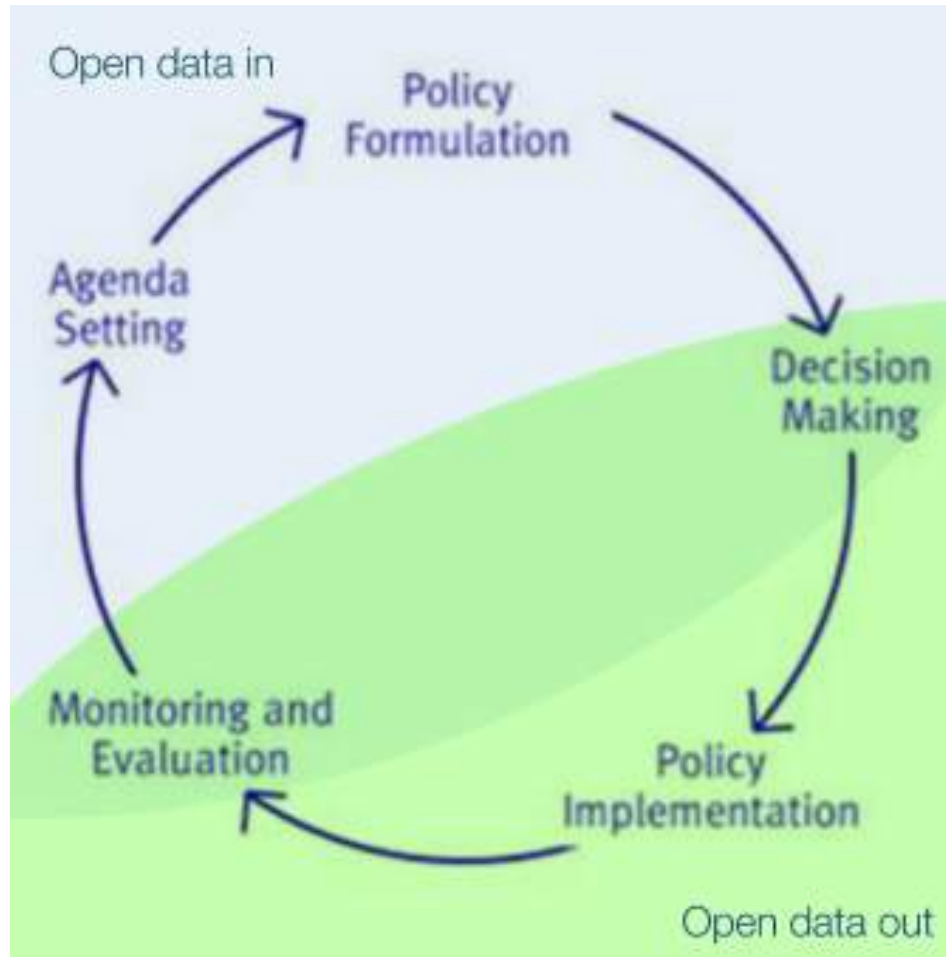
Quality and provenance

Data analysis and visualisation

Open data in policy cycles

Referencing data

The policy cycle



Open data helps at every stage of the policy cycle!

Example policy

Agenda: To publish more open data from Universities on Agriculture.

Why? To increase the benefit from this data to improve agriculture worldwide.

But what is the benefit to those who already hold the data?



Understanding researchers



Universities are ranked on the quality of their research which is linked to publication.

Therefore if data publication can hold the same value and benefit then we should see more data.

How research creates impact

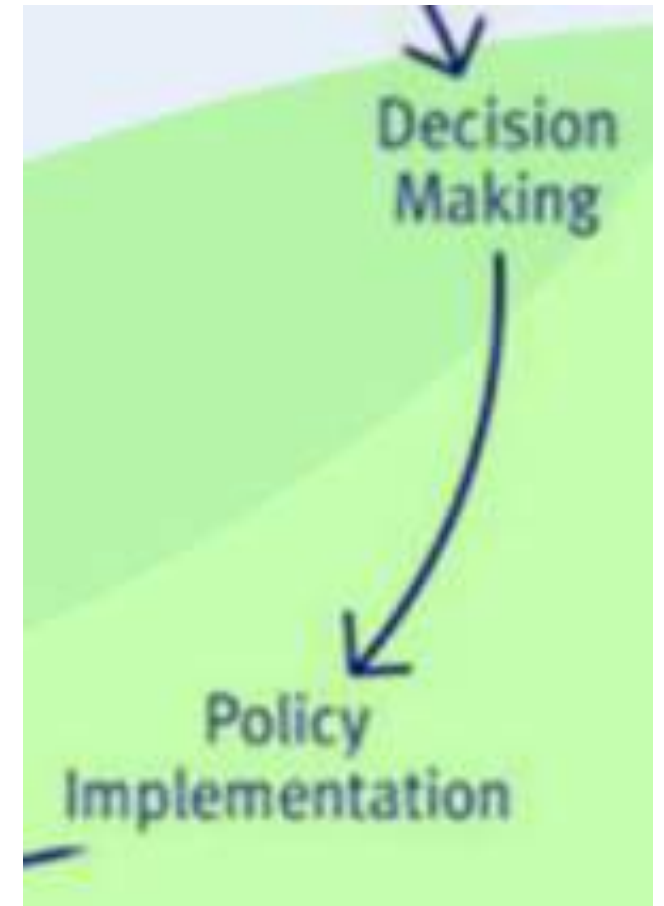


1) The journal of publication

2) The number of citations the paper has

Doing the same for research data

- 1) Create reputable places to share data
- 2) Create a way to link/reference the data, including an index
- 3) Mandate the publication of research data



DATAcite SYNOPSIS



27 DataCite Members

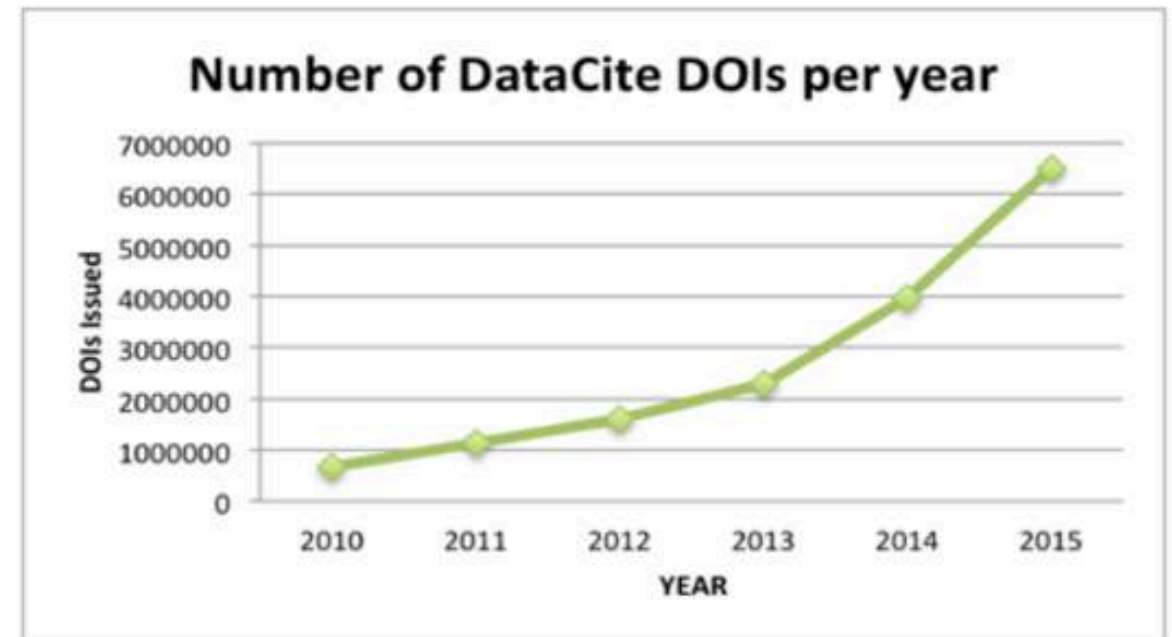
New data centers in 31-Dec-2015: 209

Total data centers by 31-Dec-2015: 635

DOIs created in 2015: 2,571,770

Total number of DOIs by 31-Dec-2015:
6,168,989

208,000 DOIs with and ORCID iDs



Monitoring and
Evaluation

Recap

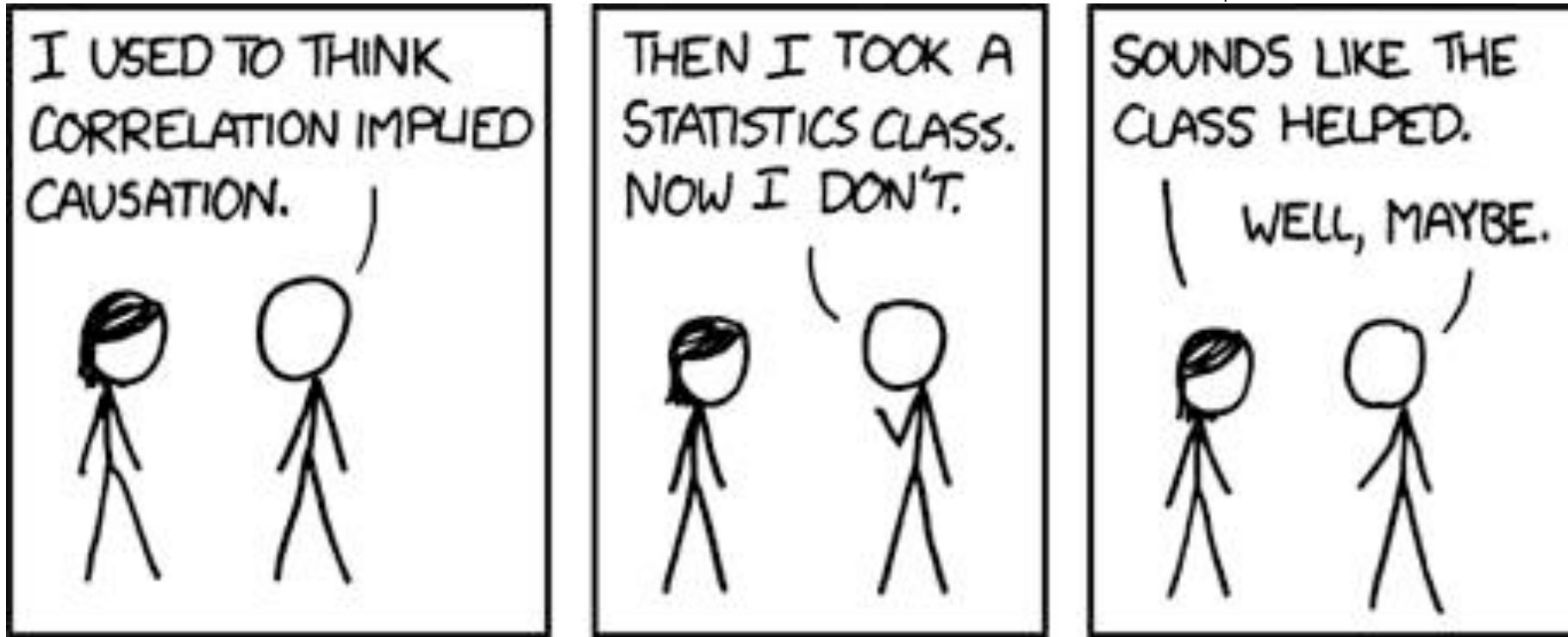
Discovering open data

Quality and provenance

Data analysis and visualisation

Open data in policy cycles

Referencing data



Thank-you

Dr David Tarrant | @davetaz | The Open Data Institute