Open Data in Practice

http://training.theodi.org/InPractice

David Tarrant · @davetaz

Schedule

Day 1: Publishing open data

Day 2: Business, the law and open data

Day 3: Using, enriching and visualising data

Agenda - Today

The characteristics of data

Data discovery patterns

*** Lunch ***

Data publication platforms

Quick big data break

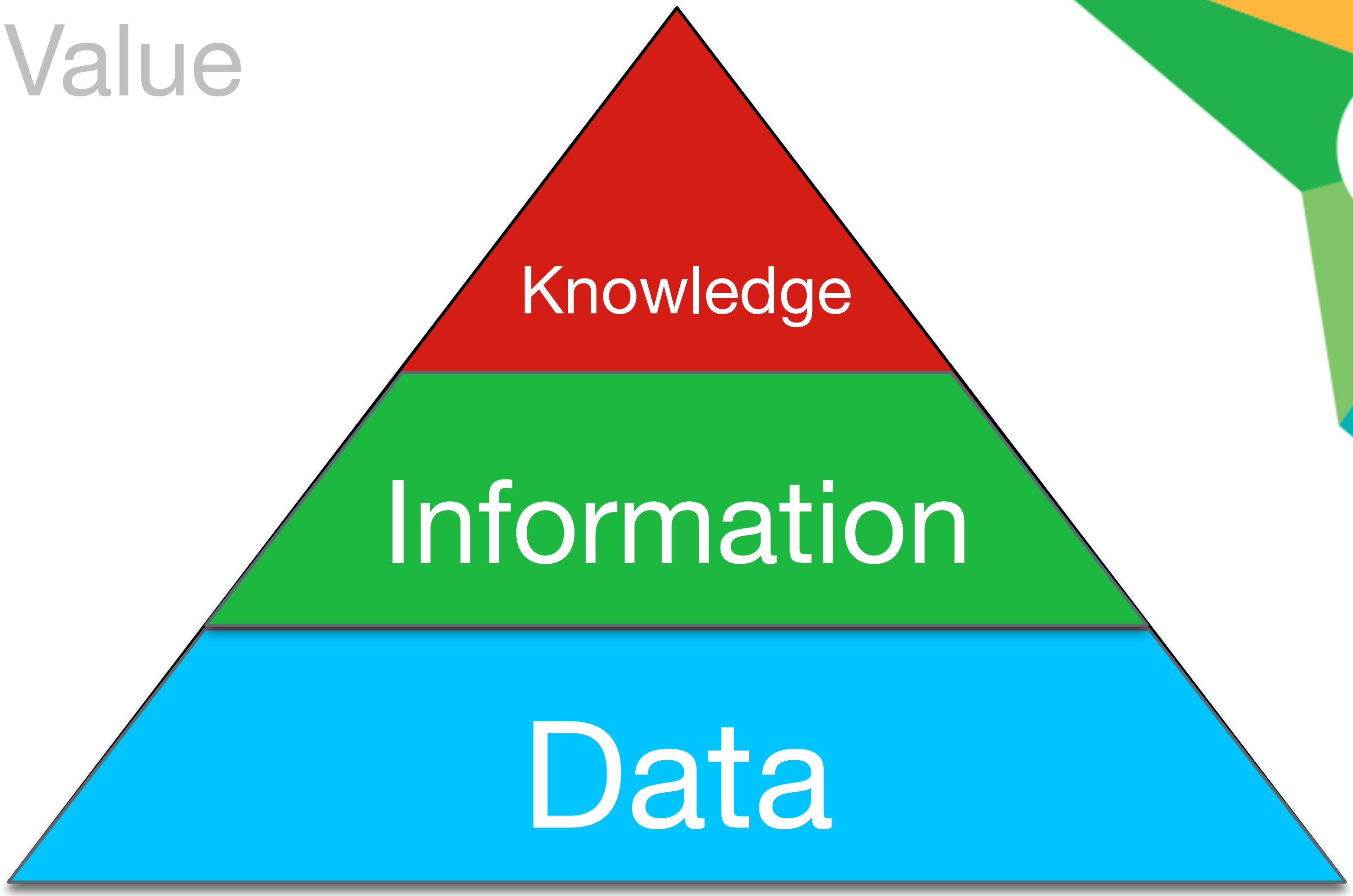
Practical publication hands-on

Recap session

Exercise

What is Data?

Value



Knowledge

Information

Data

Exercise

What is Open Data?

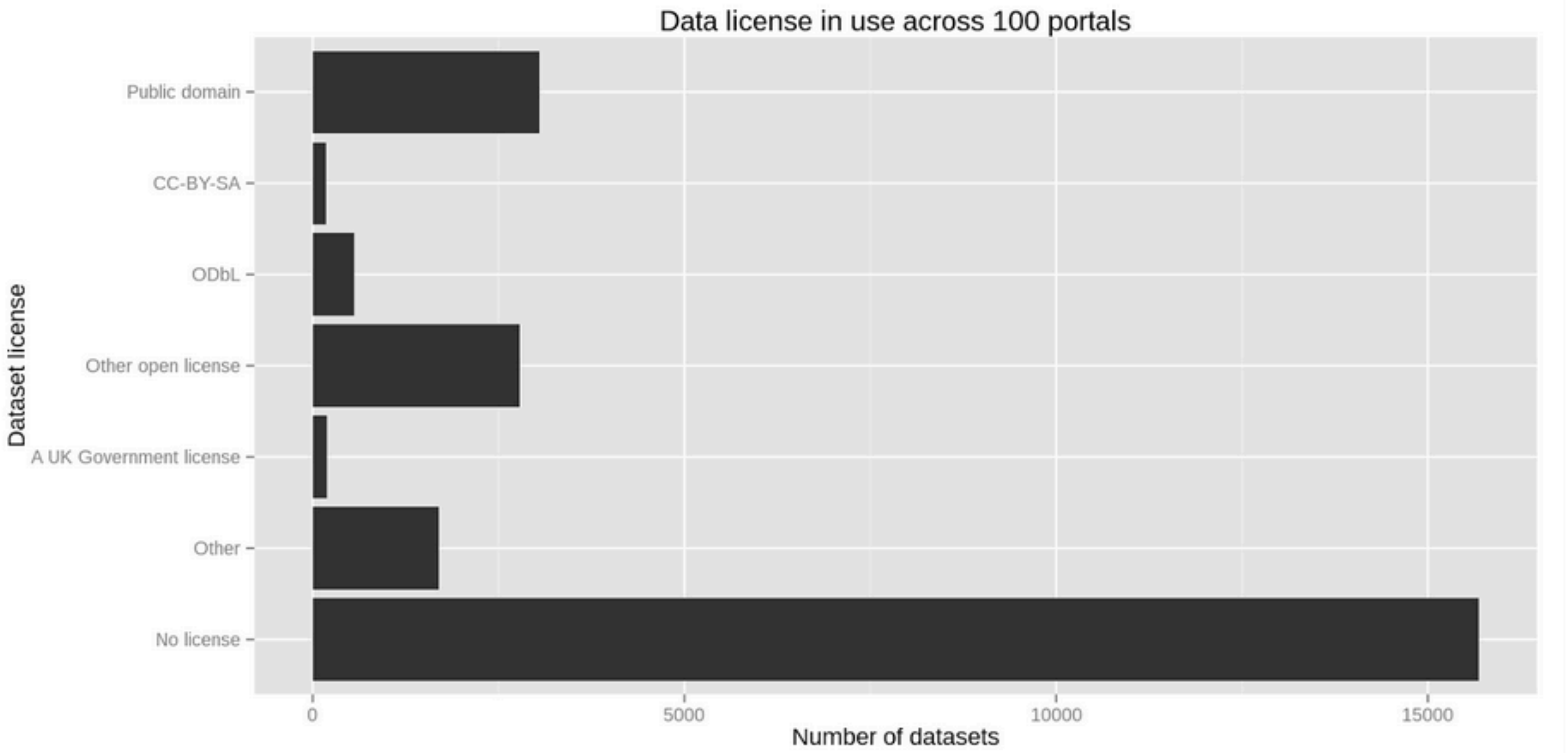
Definition of Open (OKF)

A piece of data or content is open if **anyone** is **free to use, reuse, and redistribute** it — subject only, at most, to the requirement to attribute and/or share-alike.

	Open Definition *Open Knowledge Foundation*	**OMB Memo, 2013** *The White House* *Sylvia Burwell et al.*	**Data.Gov.UK** *Antonio Acuña*	**“DBpedia: A Nucleus for a Web of Open Data”** *Sören Auer et al.*	**Open Data Institute (ODI)** *Open Data Institute*	**LinkedGov** *LinkedGov*	**McKinsey** *James Manyika et al.*	**Open Data Now** *Joel Gurin*	**Open Data Barometer** *Tim Davies*	**The World Bank** *The World Bank*
Free	✓	✓		✓	✓		✓			
Negligible Cost							✓			
Publicly Available	✓	✓			✓		✓	✓		
Re-usable	✓		✓		✓					✓
Can be Redistributed	✓			✓						✓
Non-exclusive (No Restrictions from copyright, patents, etc.)	✓			✓	✓				✓	✓
Structured for Usability		✓	✓				✓		✓	✓
Requires “Open” License			✓		✓	✓			✓	✓
Non Personally Identifiable						✓				
Produced during business operation						✓				
Belongs to the Taxpayer (when not in violation of laws/privacy)						✓				
Accessible in Bulk									✓	

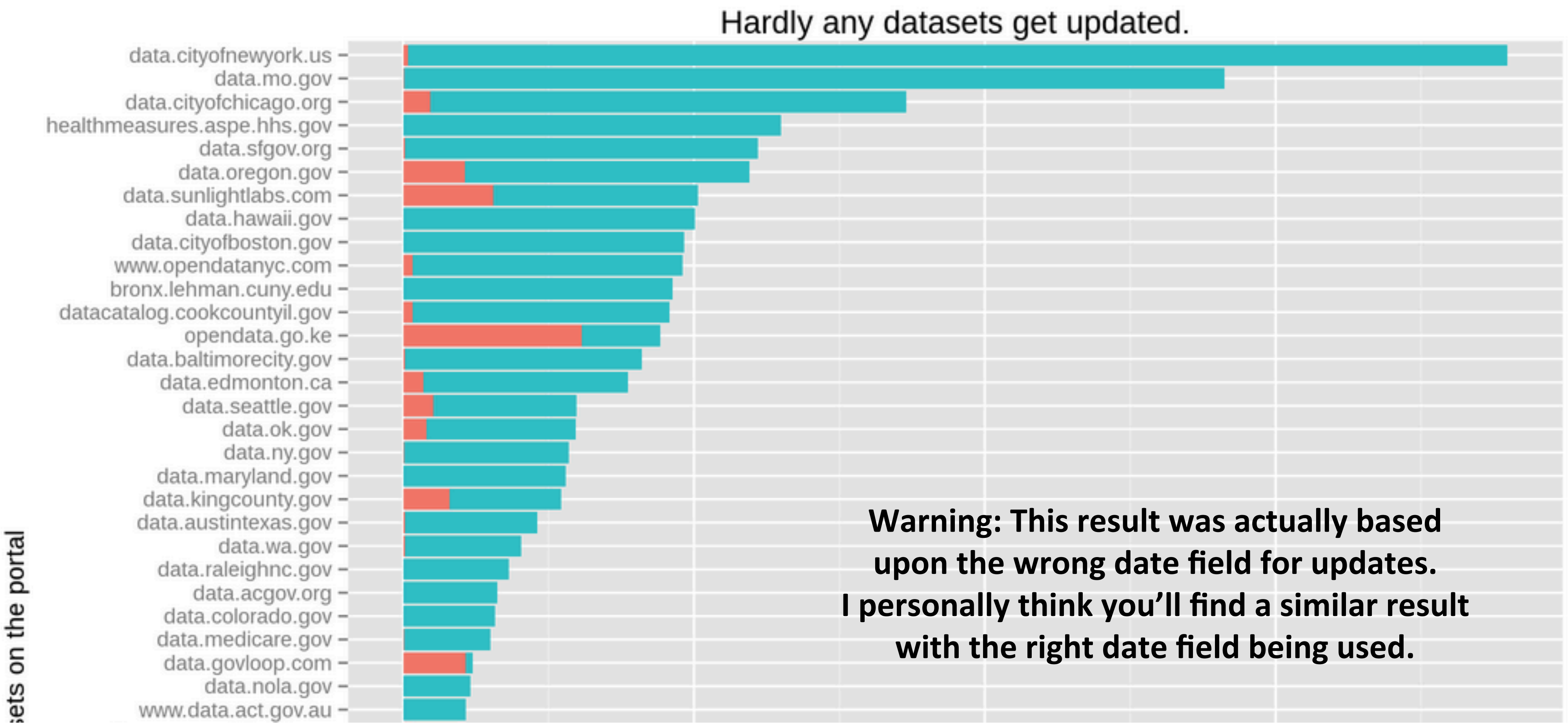
GOVLAB

Open data is hardly ever appropriately licensed.



Source: Thomas Levine

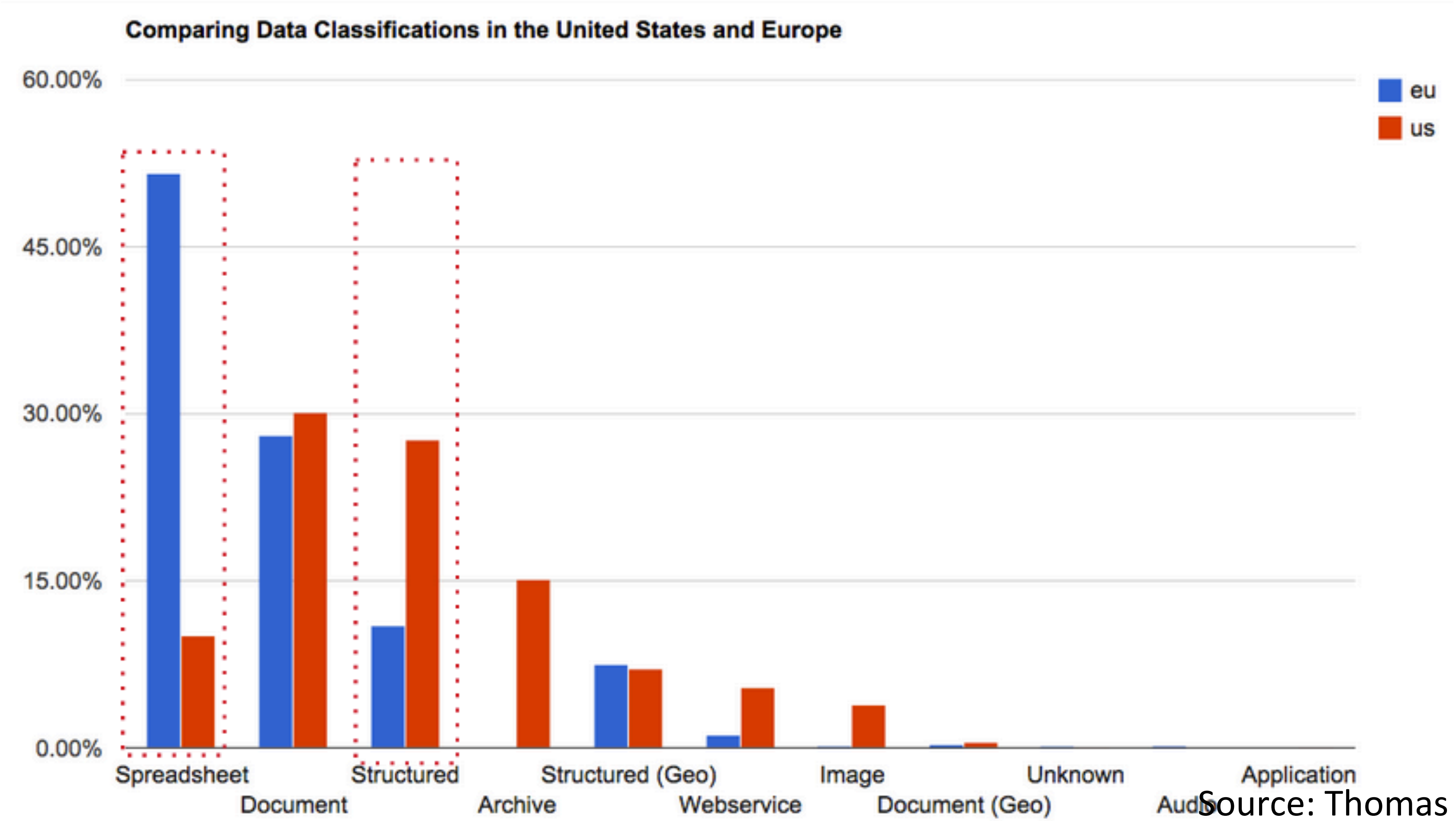
Source: Thomas Levine



Warning: This result was actually based upon the wrong date field for updates. I personally think you'll find a similar result with the right date field being used.

Source: Thomas Levine

Open data is rarely structured.



Source: Thomas Levine

Today’s mission

To move to phase 2 of publishing open data and solve some of the phase 1 problems.

Publication phases

Phase 1: Get the data online, in some form. This will help with the trust and transparency and community building.

Phase 2: Increase the usability of the data by potentially publishing differently and keeping it up to date.

Session 1

The characteristics of data

Exercise (part 1)

In your pre-training exercise, you were all asked to identify a dataset.

In your groups briefly discuss each others datasets and write down some key characteristics of each.

Also write the dataset title on a post-it, one per post-it.

Types of Data

Reference data

"things"

Transaction data

"stats involving things"

Exercise

Categorize your data into reference and transactional data.

If they are all in one category you have 2 minutes to add some new datasets to the empty category.

When done, put a “T” or and “R” on each dataset post-it.

Types of Data

Reference data

“things”

People Facilities Places
Books Buildings

Transaction data

“stats involving things”

Expenditure
Weather Consumption
Observation

Update frequency

Static

In frequent updates

Frequent updates

Live

Exercise

Categorize your data into **frequency of updates**

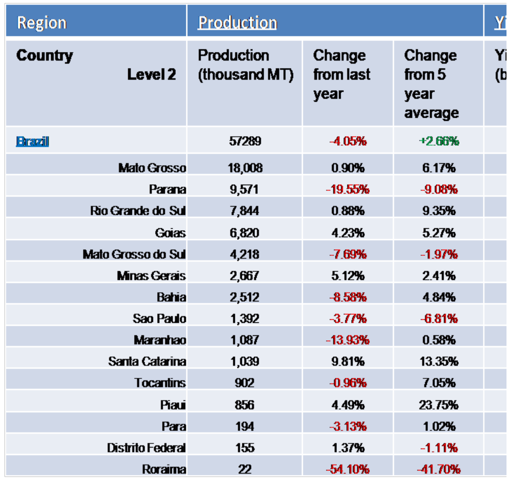
If they are all in one category you have 2 minutes to add some new datasets to the empty category/ies

Put a number on your post-its representing the frequency of updates.

0 = static, 1= In frequent, 2 = Frequent, 3 = Live

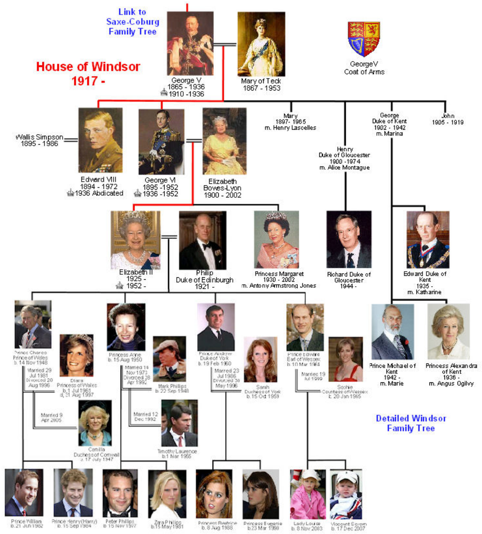
Data Representations

Tabular

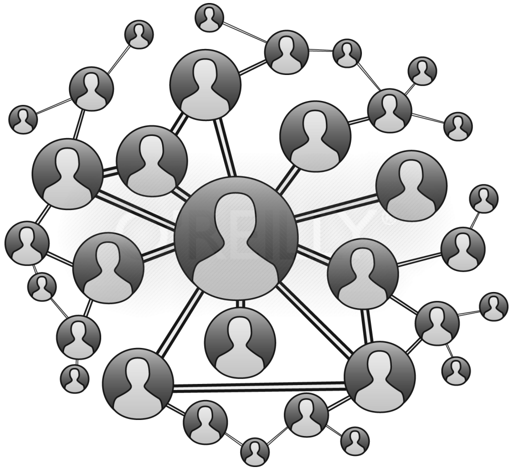


Region	Production			Yi
Country Level 2	Production (thousand MT)	Change from last year	Change from 5 year average	Yi (b
Brazil	57289	-4.05%	+2.66%	
Mato Grosso	18,008	0.90%	6.17%	
Parana	9,571	-19.55%	-9.08%	
Rio Grande do Sul	7,844	0.88%	9.35%	
Goias	6,820	4.23%	5.27%	
Mato Grosso do Sul	4,218	-7.69%	-1.97%	
Minas Gerais	2,667	5.12%	2.41%	
Bahia	2,512	-8.58%	4.84%	
Sao Paulo	1,392	-3.77%	-6.81%	
Maranhao	1,087	-13.93%	0.58%	
Santa Catarina	1,039	9.81%	13.35%	
Tocantins	902	-0.96%	7.05%	
Piaui	856	4.49%	23.75%	
Para	194	-3.13%	1.02%	
Distrito Federal	155	1.37%	-1.11%	
Roraima	22	-54.10%	-41.70%	

Hierarchical


Link to Saxe-Coburg Family Tree
House of Windsor 1917 -
George V 1865 - 1936 1910 -1936
Mary of Teck 1867 - 1953
GeorgeV Coat of Arms
Wallis Simpson 1895 - 1986
Edward VIII 1894 - 1972 1936 Abdicated
George VI 1895 -1952 1936 -1952
Elizabeth Bowes-Lyon 1900 - 2002
Mary 1897- 1965 m. Henry Lascelles
Henry Duke of Gloucester 1900 -1974 m. Alice Montague
George Duke of Kent 1902 - 1942 m. Marina
John 1905 - 1919
Elizabeth II 1925 - 1952 -
Philip Duke of Edinburgh 1921 -
Princess Margaret 1930 - 2002 m. Antony Armstrong Jones
Richard Duke of Gloucester 1944 -
Edward Duke of Kent 1935 - m. Katharine
Prince Charles Prince of Wales b. 14 Nov 1948
Married 29 Jul 1981 Divorced 28 Aug 1996
Diana Princess of Wales b. 1 Jul 1961 d. 31 Aug 1997
Married 9 Apr 2005
Camilla Duchess of Cornwall b. 17 July 1947
Princess Anne b. 15 Aug 1950
Married 14 Nov 1973 Divorced 28 Apr 1992
Mark Phillips b. 22 Sep 1948
Married 12 Dec 1992
Timothy Laurence b.1 Mar 1955
Prince Andrew Duke of York b. 19 Feb 1960
Married 23 Jul 1986 Divorced 30 May 1996
Sarah Duchess of York b. 15 Oct 1959
Prince Edward Earl of Wessex b. 10 Mar 1964
Married 19 Jul 1999
Sophie Countess of Wessex b. 20 Jan 1965
Prince Michael of Kent 1942 - m. Marie
Princess Alexandra of Kent 1936 - m. Angus Ogilvy
Detailed Windsor Family Tree
Prince William b. 21 Jun 1982
Prince Henry(Harry) b. 15 Sep 1984
Peter Phillips b. 15 Nov 1977
Zara Phillips b.15 May 1981
Princess Beatrice b. 8 Aug 1988
Princess Eugenie b.23 Mar 1990
Lady Louise b. 8 Nov 2003
Viscount Severn b. 17 Dec 2007

Network/Graph



Exercise

Categorize your data into **tabular, hierarchical (tree) and graph (network)**

If they are all in one category you have 2 minutes to add some new datasets to the empty category.

Add the word “**tab**”, “**tree**” or “**net**” to your post-its to represent the different structures.

Justifications

Trust and Transparency

Enabling the economy

One more

Categorize your data into **transparent** and **enabling**.

Summing up

Do you have any obvious grouping of your datasets?

Is this reflective of the whole open data ecosystem?

Exercise

Pick one “group” of datasets that share similar colours and come up with a data publication strategy for getting these datasets online and usable.

What are the publication requirements on the human publisher?

What are the requirements on potential users?

Thank-you