



# Finding Stories in Data

---

David Tarrant · [@davetaz](https://twitter.com/davetaz)

What is data?

Your rights to use data to tell stories

How others tell stories with data

Open, big and personal data

Data discovery patterns

## Session 1

# Telling stories with data



## Discussion

In your groups discuss –  
what is data for you?



# 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.



# Your rights



Certain uses of content are prohibited under fair dealing  
(fair use in the US).

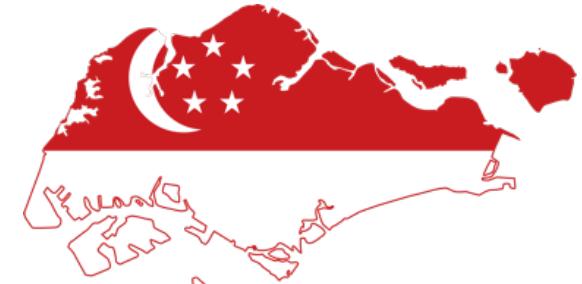
Typically these allow use of copyrighted material for:

- Commentary
- Search engines
- Criticism
- Parody
- News reporting
- Research
- Teaching
- Library archiving
- Scholarship



# Fair dealing

## Copyright act (Ch 63)



In deciding whether the use is a fair dealing, the following factors are considered:

- Purpose and character, including non-commercial, not for profit and educational usage
- Nature of the work or adaption
- Amount copied, relative to the whole work
- Affect on value of original work

In the case of reporting, commentary and review then sufficient acknowledgement of the work is required.



[http://en.wikipedia.org/wiki/Fair\\_dealing#Singapore](http://en.wikipedia.org/wiki/Fair_dealing#Singapore)

# Be careful

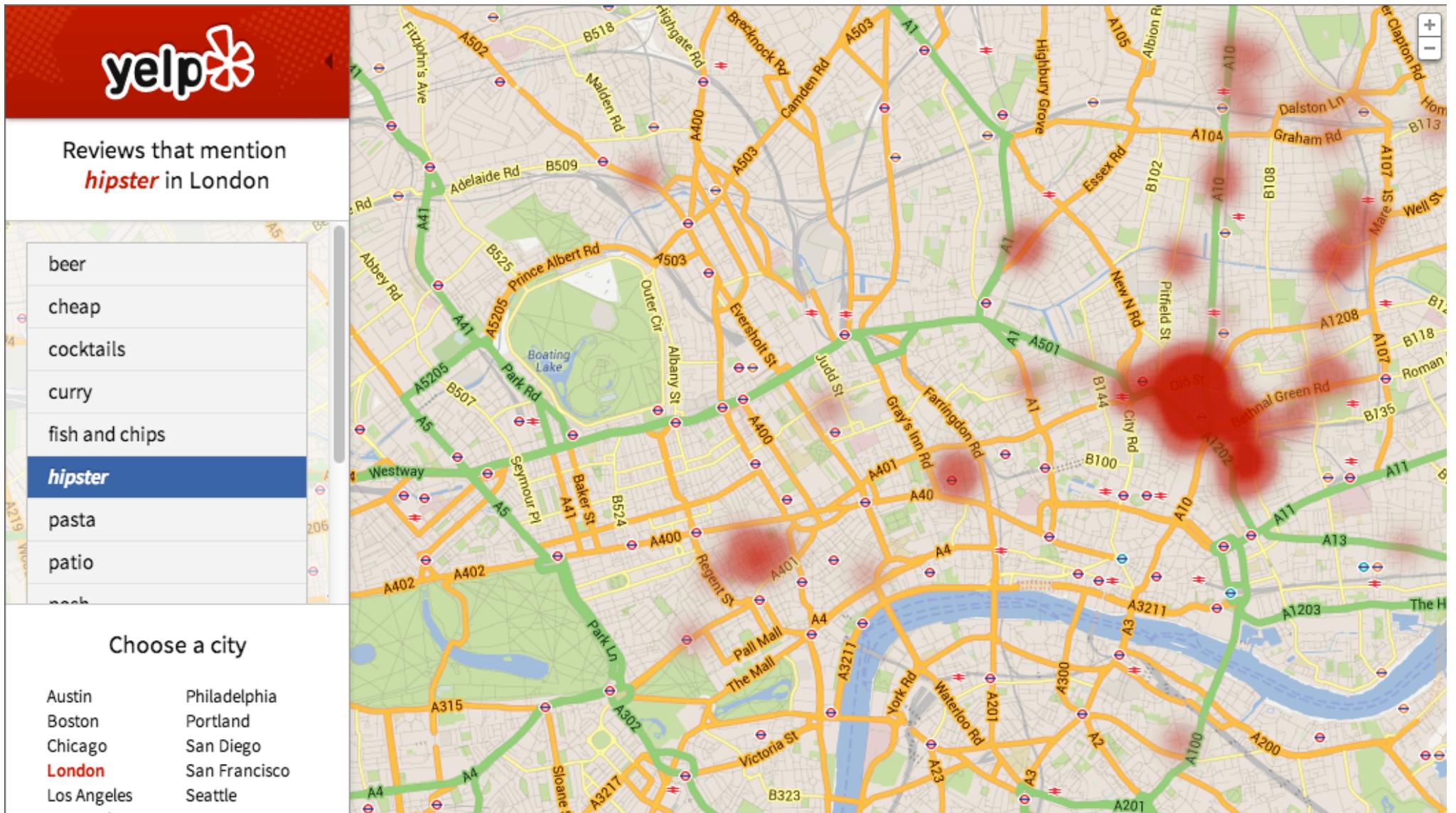


Just because something is available in the public domain, does not mean that there is a public license.



# Examples: Telling stories with data





**BBC**  Sign in News Sport Weather iPlayer TV Radio

**NEWS**  **HEALTH**

England | N. Ireland | Scotland | Wales | Business | Politics | Health | Education | Sci/Environment

the guardian

News | Sport | Comment | Culture | Business | Money | Life & style

updated at 16:15

News > Technology > Airbnb

## London's buy-to-let landlords look to move in on spare room website Airbnb

Data suggests investors with empty properties have carved out a huge presence on the site, leasing out their homes and flats

Winter brings extra pressures for the NHS, particularly in Accident and Emergency departments, as cold weather, flu and other winter bugs lead to falls, chest infections or heart problems.

**A&E tracker: The final week**

**WATCH LIVE** 7pm weekdays, week  
arest major A&E in

**4 News**



UK WORLD POLITICS BUSINESS SCIENCE TECHNOLOGY CULTURE

FRIDAY | 07 FEBRUARY 2014 | UK

## Why is government website carrying fake jobs?





The Upshot - Politics, Policy & Economics

www.nytimes.com/upshot/

SECTIONS HOME SEARCH

The New York Times

LOG IN

EDITED BY DAVID LEONHARDT

FOLLOW US [Facebook](#) [Twitter](#) [RSS](#)

# The Upshot

---

Who Will Win the Senate? We give the Democrats a 55% chance of keeping a majority. [DETAILS](#)

## Interest Rates Are Falling. Thank Vladimir Putin.



Risk aversion prompted by instability in Ukraine, among other things, has encouraged investors to seek safe havens like U.S. Treasury bonds.

---



A J Mast for The New York Times

### Another Opponent of Obamacare Starts to Soften

The Republican governor of Indiana still does not like the health care law, but he has now proposed a way to expand Medicaid.

---



Gavin Potenza

### Women and the 'I Don't Know' Problem

## Ampp3d's Sunday Paper Review

Posted 2 days ago by Federica Cocco in **HEALTH** | **POLITICS**



 Flickr/David McDermott

Some of the most important stories of the weekend, in numbers.

# 3,700

According to the [Sunday Times](#) 3,700 cancer patients waited more than 104 days for treatment in 2013. 30% of trusts have been breaching a government target of 85% of patients receiving treatment within 62 days of an urgent GP referral.

<http://ampp3d.mirror.co.uk/>



# Data is a source



Two main methods of using  
data as a source

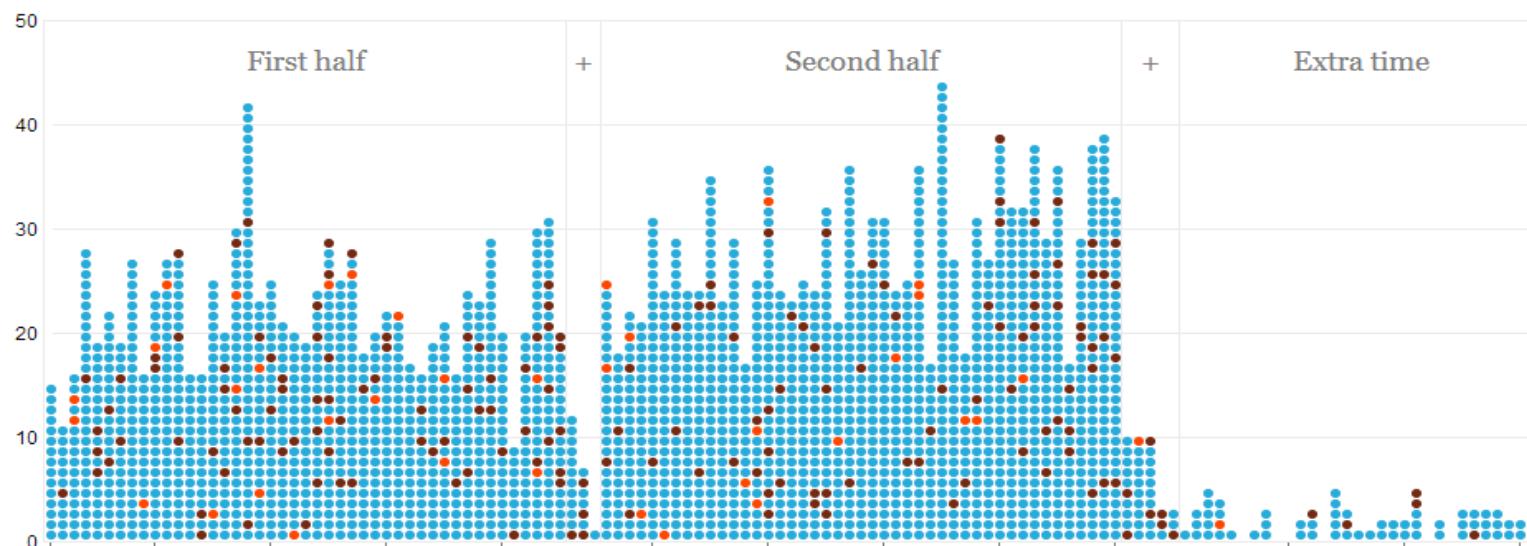
1. Story first – data used to enhance, fact check, dig deeper
2. Data first – story found/presented through data analysis



# Story, then data

World Cup goals  
1930-current

Total goals: **2,379** of which: Penalties\*: **174** | Own goals: **38**  
Average goals per game: **2.85**



# Data – then story



## For cops, no limit

Speeds reaching 90-130 mph are common among police  
Even when there's no emergency, even when they're off duty  
Punishment is rare, despite crashes and deaths

BY SALLY KESTIN AND JOHN MAINES | Staff writers



# **News: sources**

**Reactive**

**UK National Statistics**

**Parliament**

**Political groups**

**Businesses**

**Proactive**

**FOI requests**

**Surveys**

**'Ideas journalism'**

**Scraping**

Thanks to David Ottewell Head of Data Journalism Trinity Mirror (Regionals) for permission in using this slide



# Reactive news



Thanks to David Ottewell Head of Data Journalism Trinity Mirror (Regionals) for permission in using this slide



# Proactive news



Thanks to David Ottewell Head of Data Journalism Trinity Mirror (Regionals) for permission in using this slide



# Using data as a source ≠ must have visualisation

FINANCIAL TIMES

Welcome kcorrick

ft.com/globaleconomy

Search :

Home UK World Companies Markets Global Economy Lex Comment Manage

Economic Calendar Money Supply Americas China EU India Middle East UK US

September 18, 2013 2:51 pm

## Arctic sea ice melting faster than expected, UN report finds

By Pilita Clark, Environment Correspondent



The Arctic's summer sea ice is set to nearly vanish in less than 40 years, according to the final draft of a sweeping UN climate change report that sharply revises past estimates of how fast the icy north is melting.

"A nearly ice-free Arctic Ocean in September before mid-century is likely," says the draft seen by the Financial Times of the first large-scale study in six years by the Intergovernmental Panel on Climate Change.



<http://www.ft.com/cms/s/0/4b1a2f64-2048-11e3-9a9a-00144feab7de.html>

# Using data as a source ≠ (necessarily) big investigation

## The EU could ban roaming charges completely this year

Posted 6 hours ago by Anna Leach in MONEY



"That call cost how much?!" Photo: [Indi.ca](#) on Flickr

Here's how much that would might you on a 3 day holiday.

£60

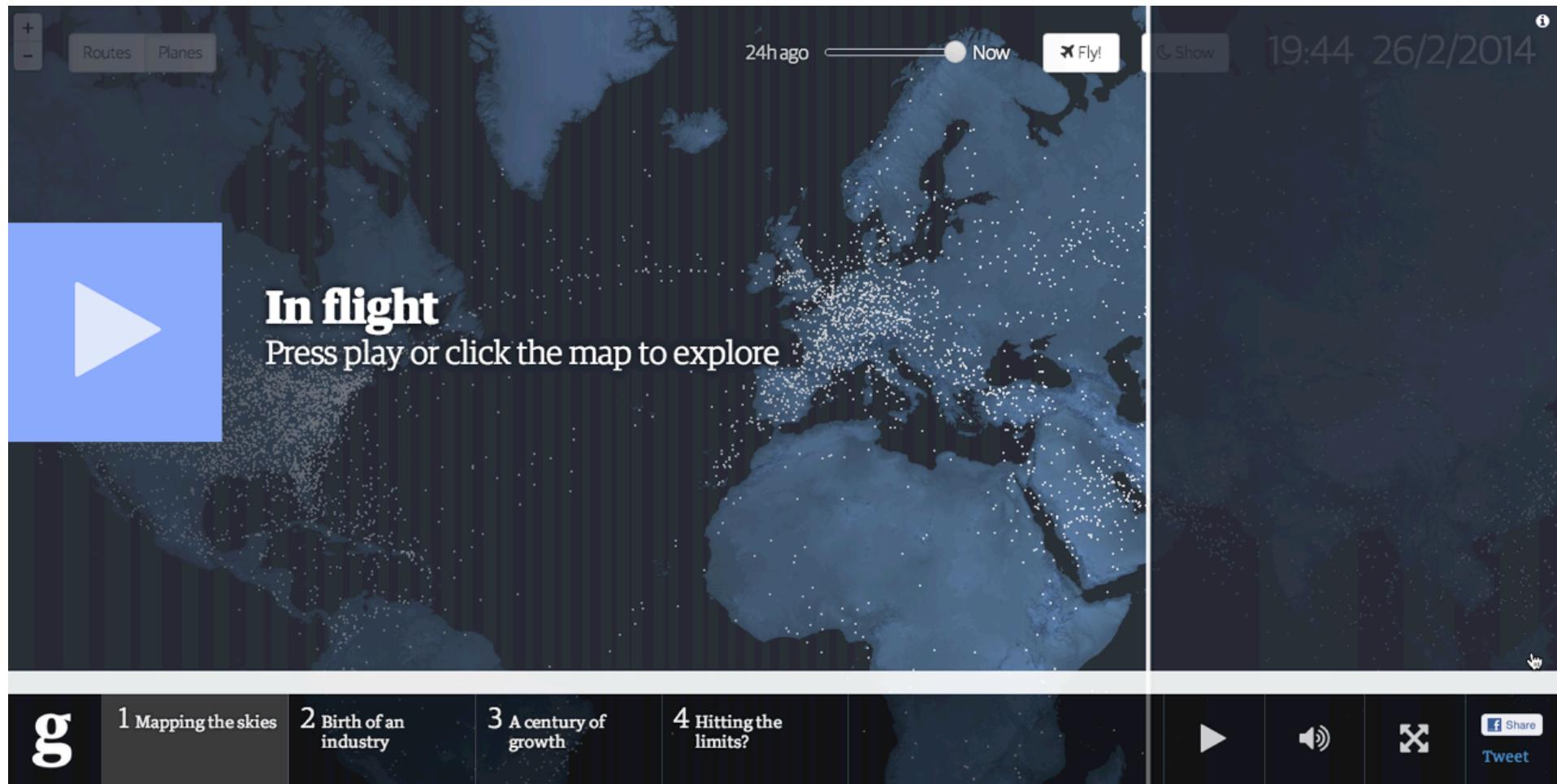
That's our estimation anyway. But we do use  
our phones *a lot*.



<http://ampp3d.mirror.co.uk/2014/02/26/the-eu-could-ban-roaming-charges-completely-this-year/>

# Data for education





<http://aviation.live.kiln.it/>

## A Boom in New Housing

In spite of a recession and foreclosure crisis, the mayor presided over a boom in residential construction, encompassing everything from new towers for the rich in Manhattan to disappearing vacant lots in the South Bronx. New York has added 40,000 new buildings since he took office, and the census counted an additional 170,000 housing units in 2010, up from 10 years earlier, more than any other city. Neighborhoods with the most growth: post-9/11 downtown; the West Side from Chelsea to Lincoln Square and Central Harlem in Manhattan; the Rockaways, Long Island City and Flushing, Queens; Williamsburg, Bushwick and Bedford-Stuyvesant, Brooklyn; the South Bronx.



New buildings constructed during Mr. Bloomberg's tenure as mayor.

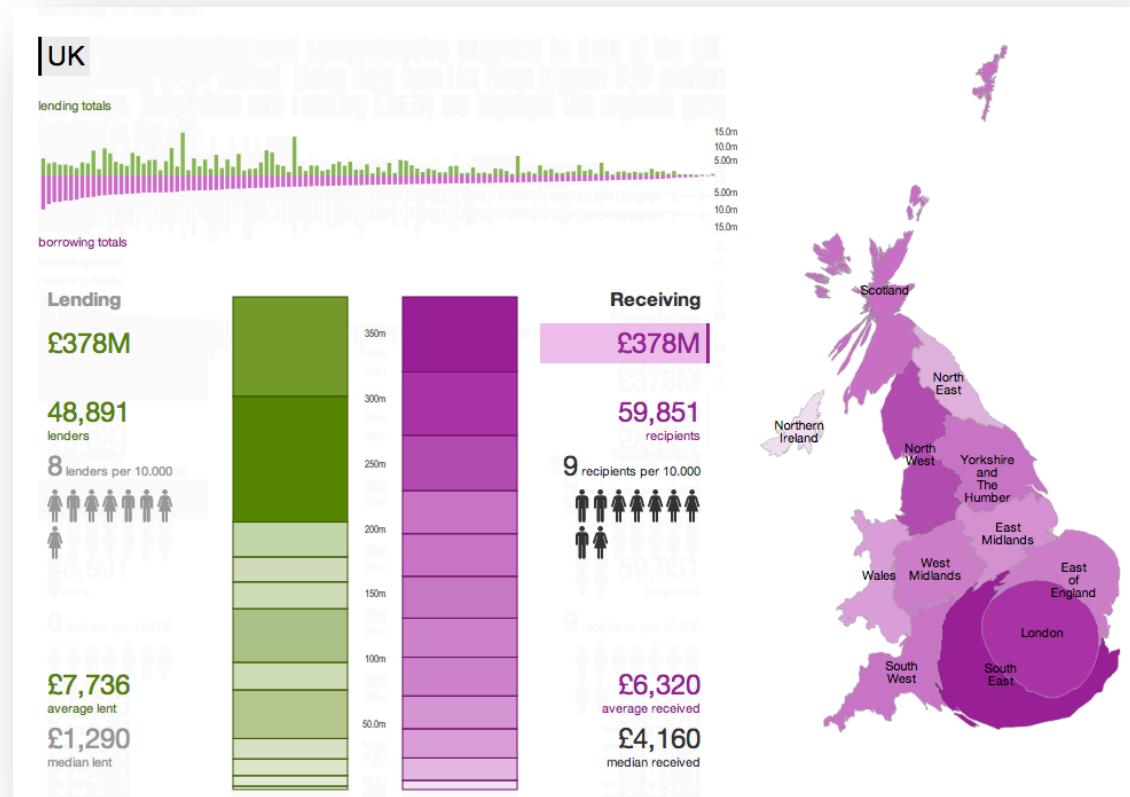


<http://www.nytimes.com/newsgraphics/2013/08/18/reshaping-new-york/>

# Self discovery



# Show me the money



<http://smtm.labs.theodi.org/>

# LFB Fire Station Closures

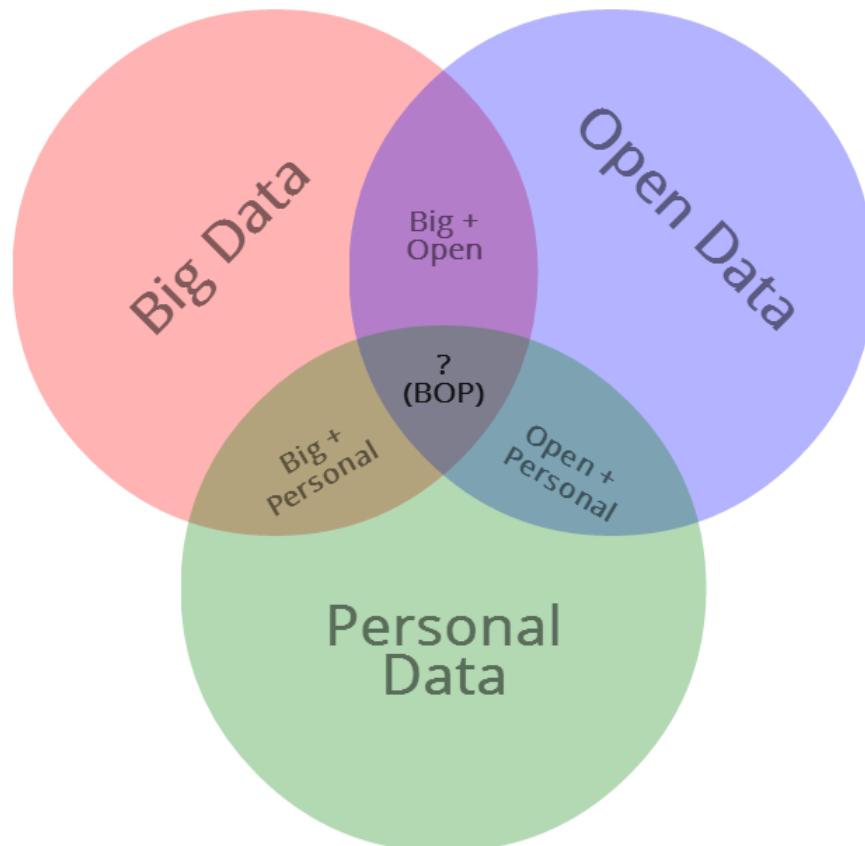


<http://london-fire.labs.theodi.org/>

# Open big and personal data



# Challenges and Risks



<http://theodi.github.io/data-definitions/>

# Types of personal data

## **Open** personal data

Data about people  
not a person

Available to anyone

Has been anonymised

e.g. number of people attending  
event, gender split, age ranges.  
(bigger numbers are better!)

## **Available** personal data

Data about a person  
Available to the person only!

Often known as MiData

e.g. credit scores, energy and other  
consumption data.

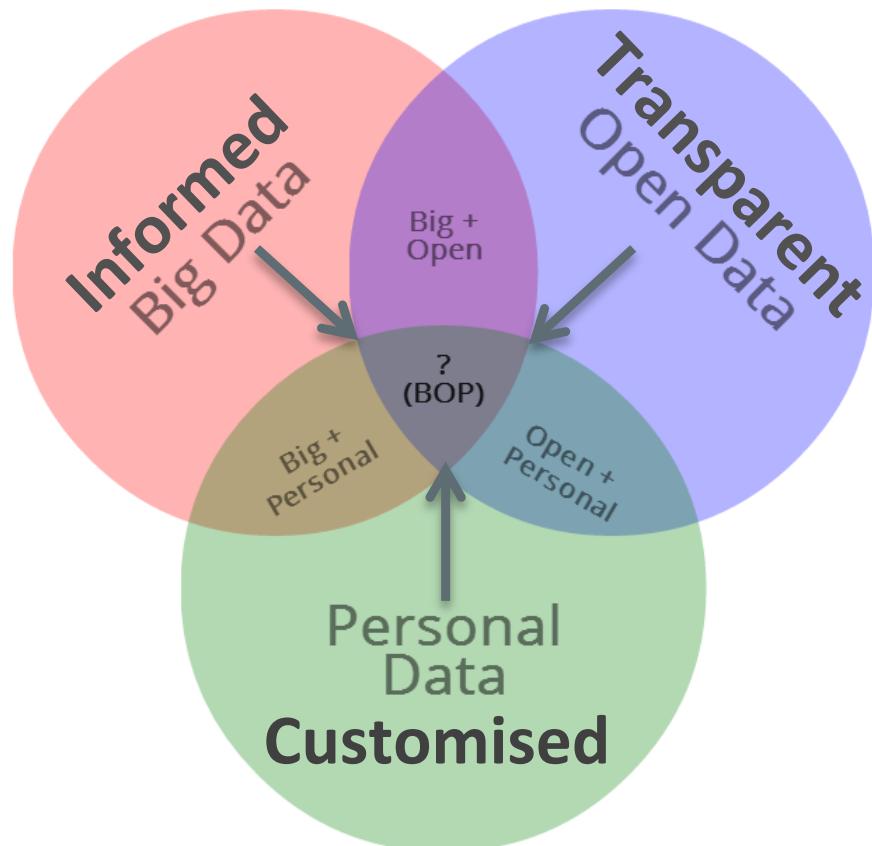
## **Personal** data

Data about a person  
which is neither open  
nor available.

Might belong to you or  
be collected by a  
company.



# Opportunity



<http://theodi.github.io/data-definitions/>

# Data discovery patterns

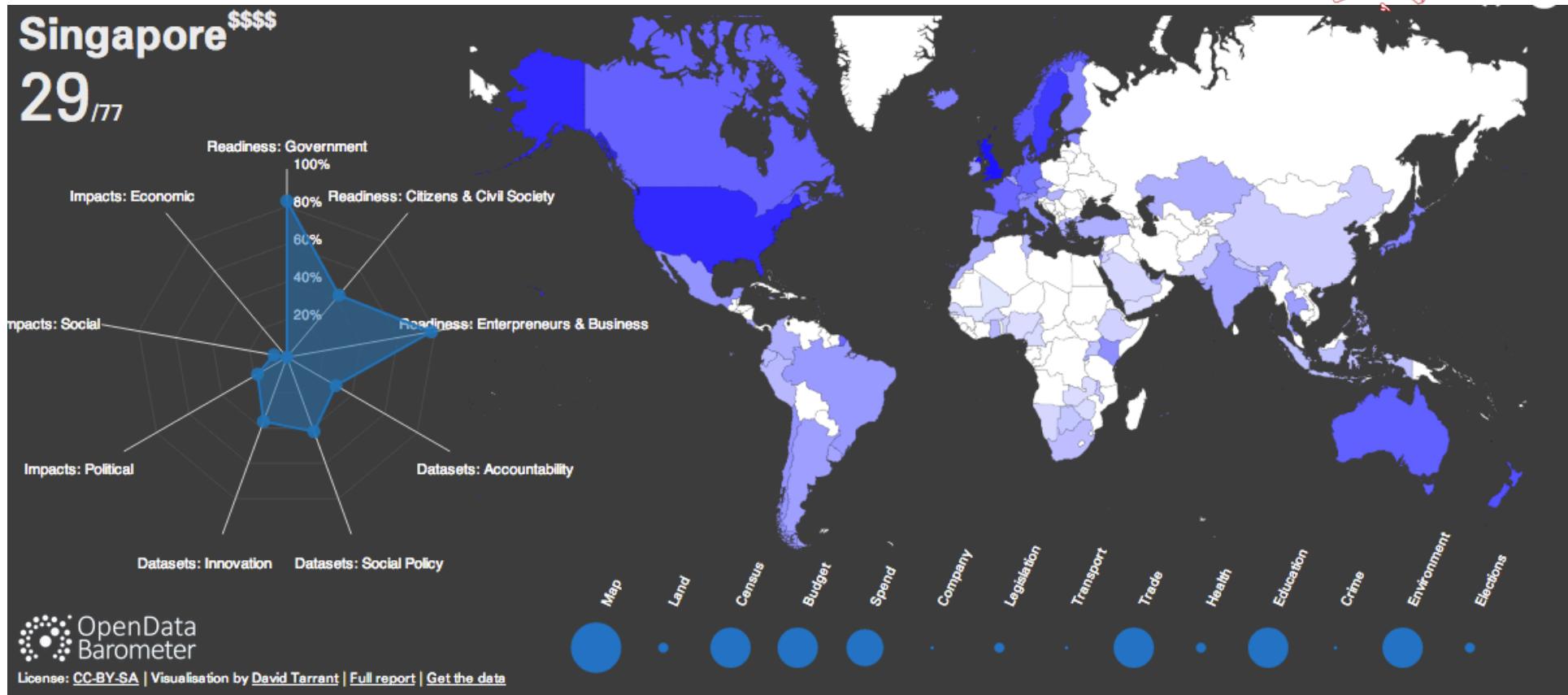


# Finding data on the web **(of documents)**

- Government data
- Google advanced
- Aggregators and portals
- Scraping



# Government data



# data.gov.XX



The screenshot shows the homepage of data.gov.sg. At the top left is the logo "data.gov.sg" with the tagline "discovering data, inspiring ideas". At the top right is the Singapore Government logo with the tagline "Integrity · Service · Excellence" and links for "Contact Us", "Sitemap", and "Feedback". A search bar is also at the top right. The main navigation menu includes "Home", "Data Sharing Principles", "Data Catalogue", "App Showcase", "For Developers", and "News & Events". Below the menu is a large banner featuring a smiling man and various icons related to government data, such as a lightbulb, a car, a DNA helix, a ship, and a train. The banner text reads "First-stop to Discover Government Data". A search bar labeled "Search Data Catalogue" is positioned below the banner. The page content area shows two tabs: "By Theme" (selected) and "By Government Agency". Below these tabs is a breadcrumb trail "Home > Data Catalogue". There are four thumbnail images representing different themes: "Business & Economy" (skyscrapers), "Education" (stacks of books), "Energy & Environment" (lightbulbs), and "Finance" (coins and banknotes). At the bottom left is a Creative Commons license logo.

# Google advanced

Google site:gov filetype:xls

Web Images Maps Shopping More ▾ Search tools

About 4,150,000 results (0.22 seconds)

[XLS] [Code List or Concept \(Acronym\)](#) ↗  
www.acquisition.gov/short\_codelistsTS.xls Share  
File Format: Microsoft Excel - View as HTML  
A, B, C, D, F, G, H, 1, Code List or Concept (Acronym), Definition, Authoritative or Maintenance Agency, Current Version, Update Frequency and/or ...

[XLS] [Approps - Foreign Assistance.gov](#) ↗  
www.foreignassistance.gov/Full\_ForeignAssistanceData.xls  
File Format: Microsoft Excel  
A, B, C, D, E, F, G, H, 1, Planning Data, 2, 3, 4, Fiscal Year, Fiscal Year Type Account Name, Agency Name, Operating Unit, Category, Sector, Amount ...

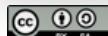
[XLS] [TSB Monthly Cash Flow Projection](#) ↗  
www.dia.iowa.gov/tsb/cashflow.xls

**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

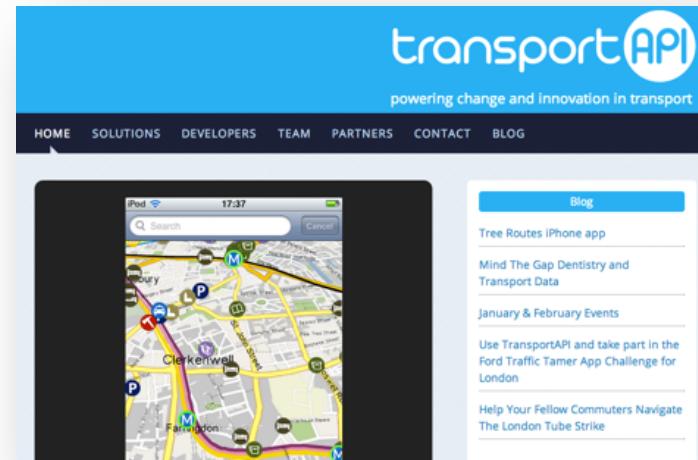


# Aggregators and portals

Collect together data from across the web into one place.



enigma.io

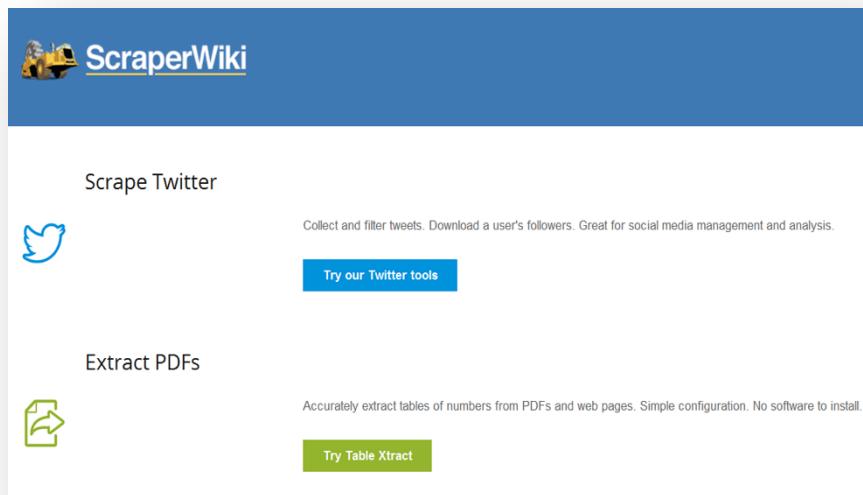


transportAPI



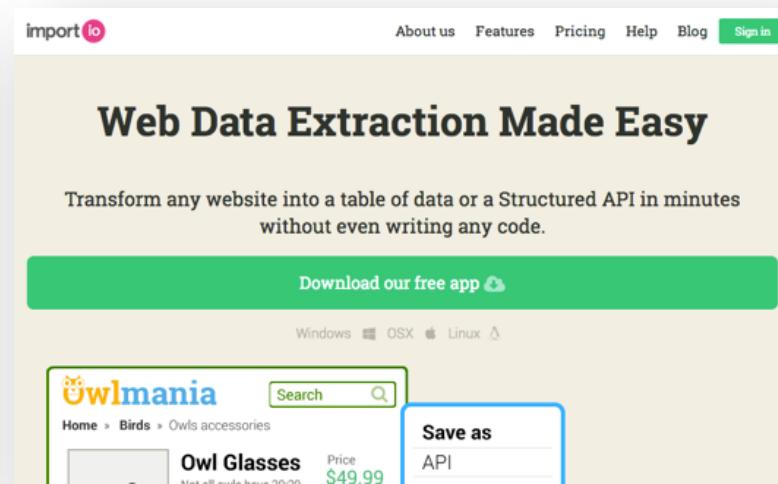
# Scraping

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



The ScraperWiki homepage features a blue header with the logo 'ScraperWiki' and a yellow bee icon. Below the header, there are two main sections: 'Scrape Twitter' and 'Extract PDFs'. The 'Scrape Twitter' section includes a Twitter icon, a brief description, and a 'Try our Twitter tools' button. The 'Extract PDFs' section includes a PDF icon, a brief description, and a 'Try Table Xtract' button.

[scraperwiki.com](http://scraperwiki.com)



The import.io homepage has a light beige background. At the top, it displays the 'import.io' logo and a navigation bar with links to 'About us', 'Features', 'Pricing', 'Help', 'Blog', and 'Sign in'. The main heading 'Web Data Extraction Made Easy' is prominently displayed. Below the heading, a subtext reads 'Transform any website into a table of data or a Structured API in minutes without even writing any code.' A large green button with the text 'Download our free app' and a download icon is centered. At the bottom, there's a screenshot of a web page from 'Owlmania' showing an 'Owl Glasses' product with a price of '\$49.99'. To the right of the screenshot, a 'Save as API' button is highlighted with a blue border.

[import.io](http://import.io)



# Finding data on the web (of data)

1. Add random extensions (.xml, .json, .csv etc)
2. Look for alternative links (rss feeds etc)
3. Look for embedded data
4. Do some content negotiation
5. Spot the API
6. Scrape (or search google again)



How the web should work,  
but people forgot that Tim  
put this in when he  
invented it!



# Duck typed data

If it looks like a duck  
and quacks like a duck,  
then it's probably a duck.

Basically, keep an eye out for tables,  
lists and other stuff that looks like data.



# 1. Adding random extensions

GOV.UK

Search

Home > Business and self-employed > Imports and exports

## Trade Tariff

Search the tariff  name or code

This tariff is for 6 August 2014 [change date](#)

View all sections [A-Z Index](#)

Trade between the UK and All countries [change country](#)

Section	Chapters	Title
I	1 to 5	Live animals; animal products
II	6 to 14	Vegetable products
III	15	Animal or vegetable fats and oils and their cleavage products; prepared edible fats; animal or vegetable waxes
IV	16 to 24	Prepared foodstuffs; beverages, spirits and vinegar; tobacco and manufactured tobacco substitutes
V	25 to 27	Mineral products
VI	28 to 38	Products of the chemical or allied industries
VII	39 to 40	Plastics and articles thereof; rubber and articles thereof
VIII	41 to 43	Raw hides and skins, leather, furskins and articles thereof; saddlery and harness; travel goods, handbags and similar containers; articles of animal gut (other than silkworm gut)
IX	44 to 46	Wood and articles of wood; wood charcoal; cork and articles of cork; manufactures of straw, of esparto or of other plaiting materials; basket-ware and wickerwork

UK Trade Tariff



Try using the following: [.csv](#) [.json](#) [.xml](#) [.rss](#) [.rdf](#)

one

## DOCTOR WHO

Home Episodes Clips Galleries Latest News Characters Monsters Fun and Games More

**On iPlayer**  
This programme will be available shortly after broadcast



**It's Tomorrow... Get the Latest on the Launch!**  
What's happening and how to follow the action during tomorrow's big launch in Cardiff.

**On TV**



**The Day of the Doctor**  
SATURDAY 19:00  
BBC THREE  
**All upcoming**  
(0 NEW AND 1 REPEAT)

BBC Music and Programmes

## 2. Look for alternative links



The screenshot shows a news website with a dark header. The header includes the logo for "Business Insight - NEWSASIA", navigation links for "NEWS", "TV", and "WATCH LIVE", and a date "Wed, Aug 06 2014". Below the header is a menu bar with categories: ASIA PACIFIC, SINGAPORE, WORLD, BUSINESS, SPORT, ENTERTAINMENT, TECHNOLOGY, HEALTH, LIFESTYLE, VIDEOS, WEATHER, and MORE. There are also three promotional banners: "CHANGELIVES", "LUMINARY AWARDS", and "START-UP". A large, bold text overlay "Scroll down!" is positioned over the main content area. The main content features a large image of two men shaking hands at a ceremony. The caption reads: "Raise of up to 12% for Home Team officers, with sign-on bonuses of up to S\$30,000". Below the image is a smaller text block: "Officers from the Home Team, both past and present were recognised at the Home Team National Day Observance Ceremony on Wednesday." A large black arrow points downwards from the "Scroll down!" text towards the bottom of the page. At the bottom left, there are links for "LIFESTYLE" and "VIDEOS". On the right side, there are two more news items: "Pay rise, special bonus for about 23,000 nurses" (posted 10 hours ago) and "50,000 openings on Jobs Bank for Singaporeans, PRs" (posted 1 hour ago). At the very bottom right, there is a news item about "NUS University Town identified as a high-risk dengue cluster" (posted 10 hours ago).



## 2. Look for alternative links



 <b>CHANNEL NEWSASIA</b>  MediaCorp News Group. © 2014 MediaCorp Pte Ltd. All Rights Reserved.  <a href="#">Terms and Conditions</a> <a href="#">Privacy Policy</a> <a href="#">About MediaCorp Pte Ltd</a>	<b>NEWS</b> <a href="#">Asia Pacific</a> <a href="#">Singapore</a> <a href="#">World</a> <a href="#">Business</a> <a href="#">Sport</a> <a href="#">Entertainment</a> <a href="#">Technology</a> <a href="#">Health</a> <a href="#">Lifestyle</a> <a href="#">Videos</a> <a href="#">Photos</a> <a href="#">Special Reports</a> <a href="#">Archives</a>	<b>TV</b> <a href="#">Live TV</a> <a href="#">TV Videos</a> <a href="#">TV Schedule</a>  <b>SERVICES</b> <a href="#">Weather</a>  <b>ADVERTISE WITH US</b> <a href="#">Online Advertising</a> <a href="#">Mobile Advertising</a> <a href="#">TV Advertising</a> <a href="#">Contact Sales</a>	<b>ABOUT US</b> <a href="#">About Channel NewsAsia</a> <a href="#">Our Logo</a> <a href="#">Our Coverage</a> <a href="#">Our Tagline</a> <a href="#">Presenters and Correspondents</a> <a href="#">Contact Us</a>  <b>GET OUR NEWS</b> 
---	---	---	---



RSS



# Finding data on the web

(  
Techniques 3-5 are not  
covered in this session. Please  
ask your trainer for more  
information if there is time.

1. Look for JSON (json, .csv etc)
2. Look for feeds (feeds etc)
3. Look for embedded data
4. Do some content negotiation
5. Spot the API
6. Scrape (or search google again)



# Exercise

Find a data set using one of the routes we've just looked at.....

Ask yourself – (and discuss in groups)  
Do you trust it? What makes it trustworthy?

A basic test with three questions:

1. Says who?
2. Compared to what?
3. Since when?

From the Statistical literacy guide  
How to spot spin and inappropriate use of statistics

<http://www.parliament.uk/briefing-papers/SN04446.pdf>





G'DAY

Thank-you