Finding Stories in Data
Before we start…

Please download the latest Chrome at:
http://tinyurl.com/install-google-chrome-now

Please set up a Google account if you don't have one already at: https://accounts.google.com/SignUp

The slides will be made available online after the course

David Tarrant · @davetaz

Introductions

Your name

What excites you most about data?

What do you want to do differently after the course?

Agenda - Today

1. Telling stories with data
2. Finding reliable data sources
3. *** Lunch ***
4. Exploring: is there a story in this data?
5. Presenting your story

WHAT IS DATA?

Discussion

In your groups discuss – what is data for you?

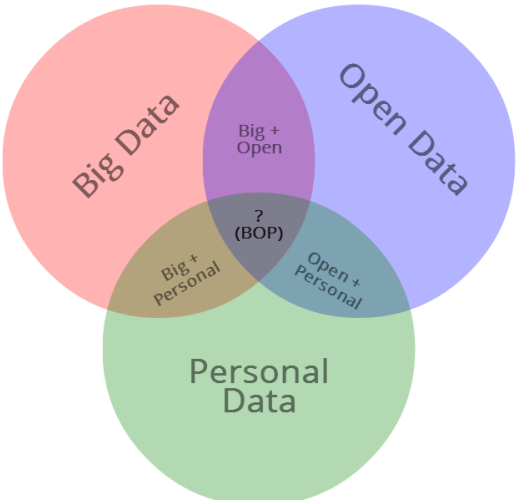
Exercise

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.

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

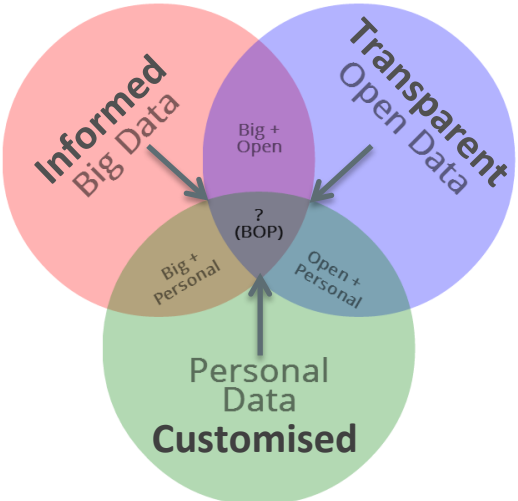
e.g. credit scores, energy and other consumption data.

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/

EXAMPLES OF USING DATA

The Upshot - Politics, Poll

www.nytimes.com/upshot/

SECTIONS HOME SEARCH

The New York Times

LOG IN

TheUpshot

EDITED BY DAVID LEONHARDT
FOLLOW US

Who Will Win the Senate? We give the Democrats a **55% chance** of keeping a majority. DETAILS

Interest Rates Are Falling. Thank Vladimir Putin.

Reuters

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/

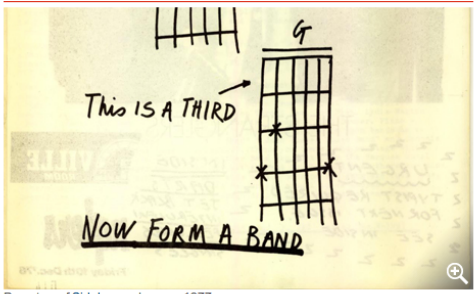
DATABLOG
Facts are sacred

Previous | Blog home | Next

Anyone can do it. Data journalism is the new punk

Can anyone be a data journalist? **Simon Rogers** on what we can learn from a 1977 diagram

• Another view: What data can and cannot do by Jonathan Gray



Page two of Sideburns, January 1977

This is a chord… this is another… this is a third. NOW FORM A BAND

Posted by
Simon Rogers
Thursday 24 May 2012
13.00 BST
theguardian.com
Jump to comments (8)

Article history

Media
Data journalism · Open journalism

Technology

http://www.theguardian.com/news/datablog/2012/may/24/data-journalism-punk

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

News: sources

Reactive	Proactive
UK National Statistics	FOI requests
Parliament	Surveys
Political groups	'Ideas journalism'
Businesses	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 Manag

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

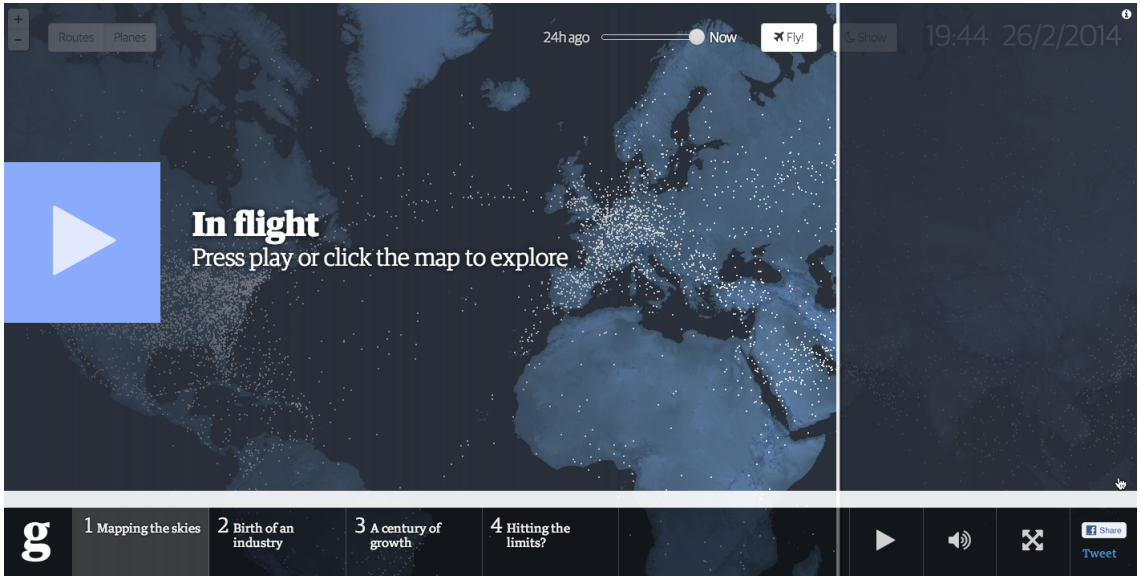
"That call cost how much?!" Photo: Indi.ca on Flickr

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

£60

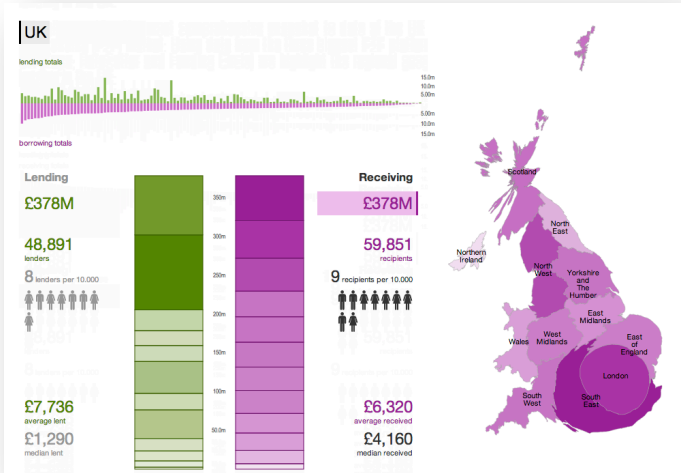
That's our estimation anyway. But we do use our phones *a lot*.

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



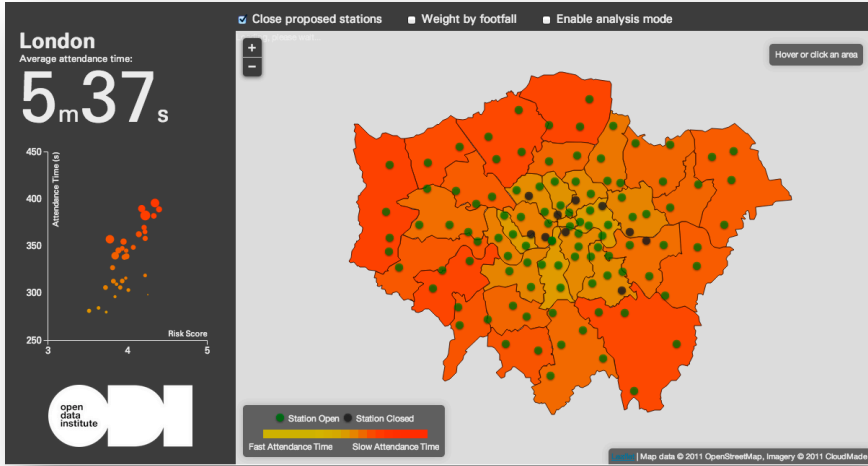
http://aviation.live.kiln.it/

Show me the money



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

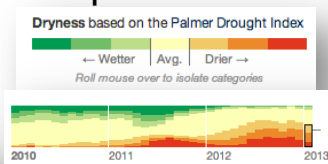
LFB Fire Station Closures

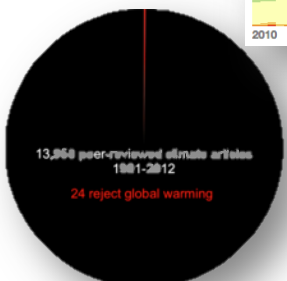


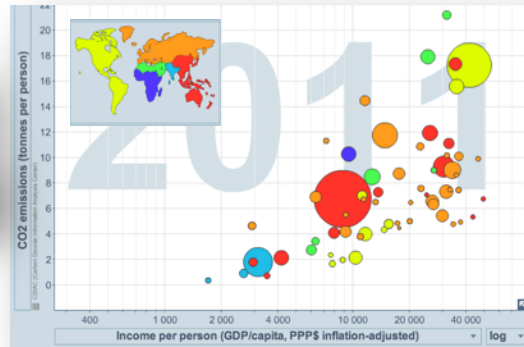
http://london-fire.labs.theodi.org/

Environmental Data

Open data shows the clear picture of our planet



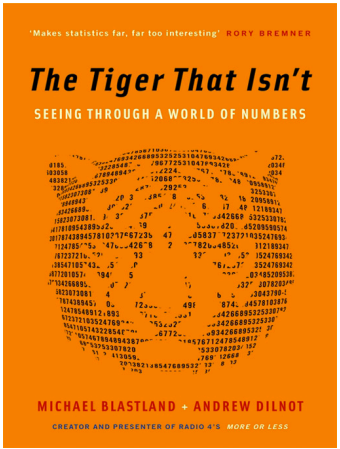




Open data makes us aware of the impact we have on our planet

DODGY STATISTICS

12 problems with numbers



1. Counting

Flickr: mattbrittaine

2. Big numbers

£300m

boost for childcare

1,000,000

new places

£1.15

per week
per child

3. Chance

Random events cluster

They do not evenly distribute

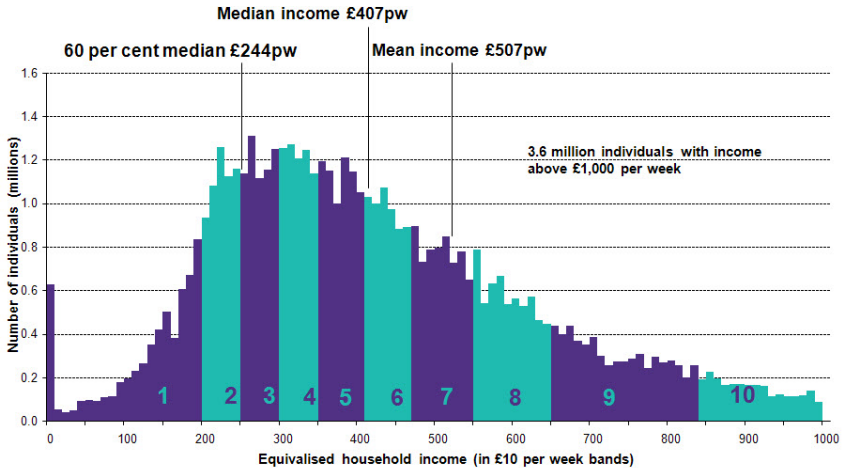
4. Fluctuation

Speed cameras installed on dangerous roads reactively

If you install on the top of a high wave, then the next wave is almost certainly going to be lower

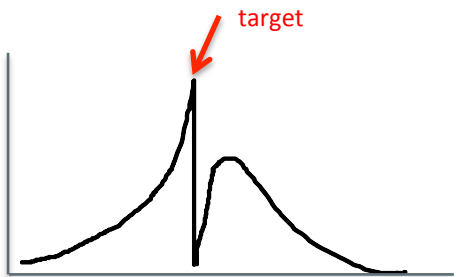
Very much related to correlation

5. Averages



6. Targets

Look at one aspect to measure an entire service.



Encourage gamification

7. Risk


SECTIONS
OTTAWA CITIZEN
HOME
NEWS
NATIONAL
LOCAL NEWS
Combined va
seizure risk ir
ELIZABETH PAYNE
More from
Published on: June 9, 2014 | Last Upda
CANCER RESEARCH UK
Donate
HOME
MENU
SEARCH
Home › About us › Cancer News › News report › Global cancer incidence predicted to increase by 75 per cent by 2030
Global cancer incidence predicted to increase by 75 per cent by 2030
News report
31 May 2012
In collaboration with the Press Association
The number of worldwide cancer cases is set increase by 75 per cent in the next two decades, according to researchers in France.
The scientists predict cancer cases will increase from 12.7 million in 2008 to 22.2 million by 2030.
Recent news
Investing in cancer research boosts economy as well as

8. Sampling

Between 2,000 and 5 million cases of norovirus in winter 2007/08.

Based upon 2,000 confirmed cases extrapolated from BMJ report based on sample size of …. **1**

The Telegraph

Home | News | World | Sport | World Cup | Finance | Comment | Cultu
Politics | Investigations | Obits | Education | Earth | Science | Defence

HOME » NEWS » UK NEWS

GPs urge millions hit by bug to stay at home

The NHS advises symptoms

BBC NEWS

LIVE | BBC NEWS CHANNEL

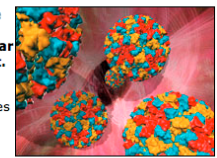
News Front Page
World
UK
England
Northern Ireland
Scotland
Wales
Business
Politics
Health
Medical notes
Education
Science & Environment
Technology

Last Updated: Friday, 11 January 2008, 12:02 GMT

E-mail this to a friend | Printable version

Vomiting bug 'hits three million'

Almost three million people have been affected by the norovirus stomach bug so far this winter, figures suggest.

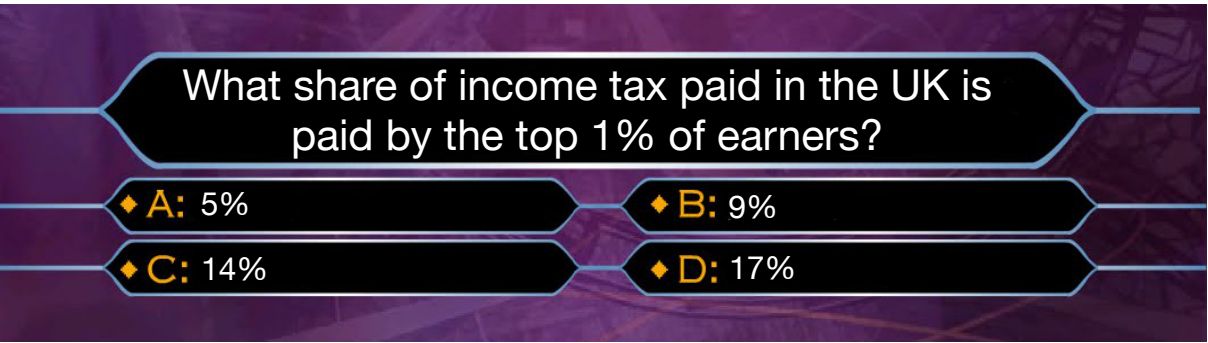


Surveillance from the Health Protection Agency shows cases in England and Wales are double those seen last year.

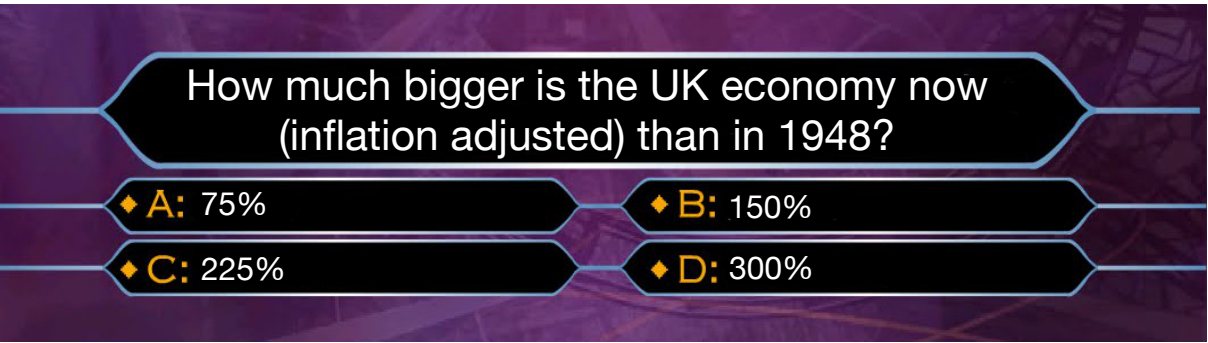
Norovirus causes sudden vomiting and diarrhoea

Doctors advise people to stay at home for 48 hours after

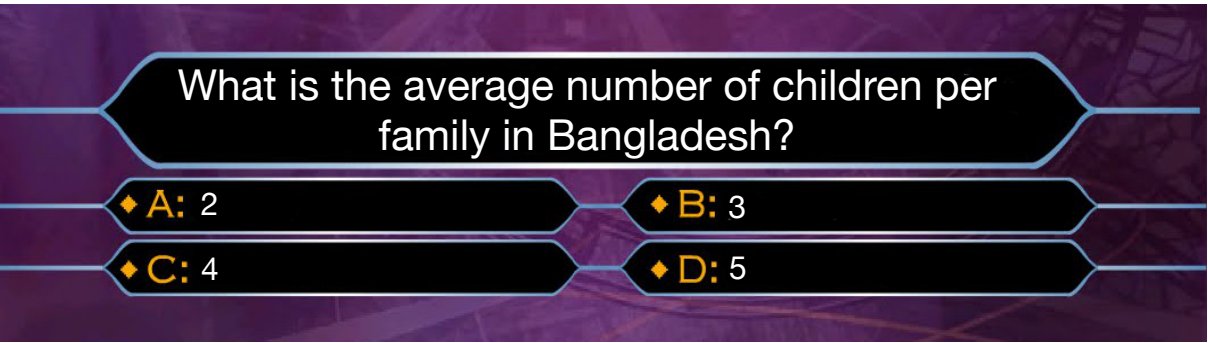
9. Data (known unknowns)



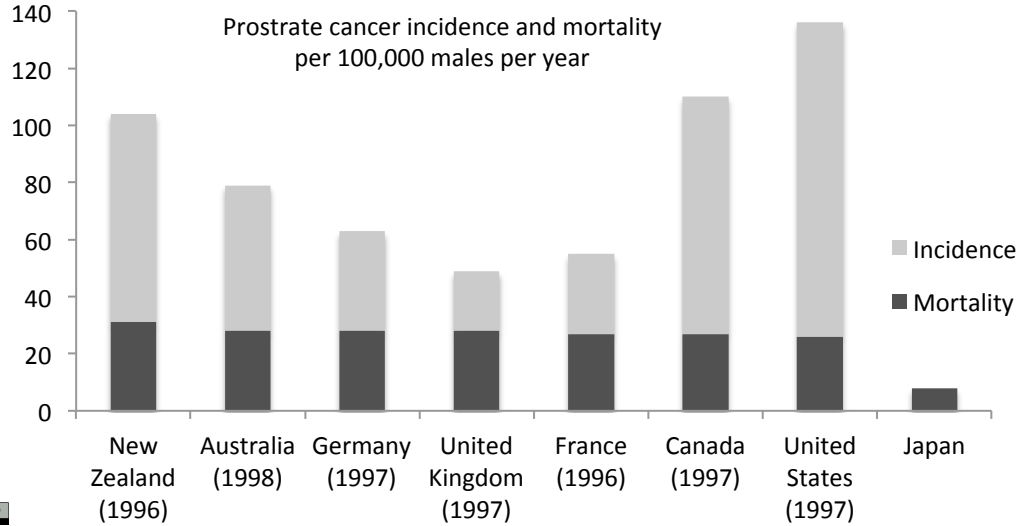
9. Data (known unknowns)



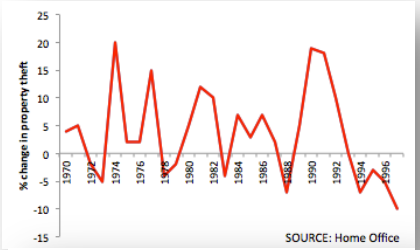
9. Data (known unknowns)

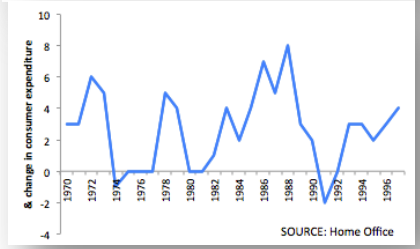


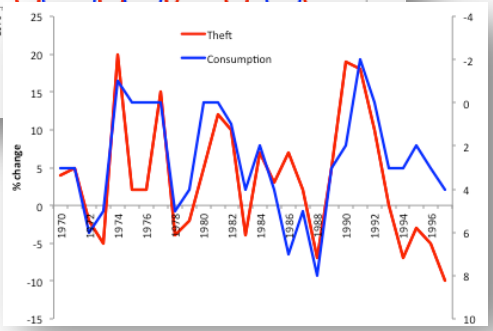
10: Comparison



11. Correlation







12. Percentages

Know the difference between a **percentage** and a **percentage point**.

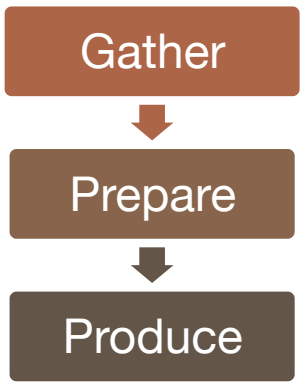
VAT increased from 17.5% to 20% on January 2011.

This is a rise of 2.5 percentage points **not** a rise of 2.5%.

How much would a rise in 2.5% actually be?

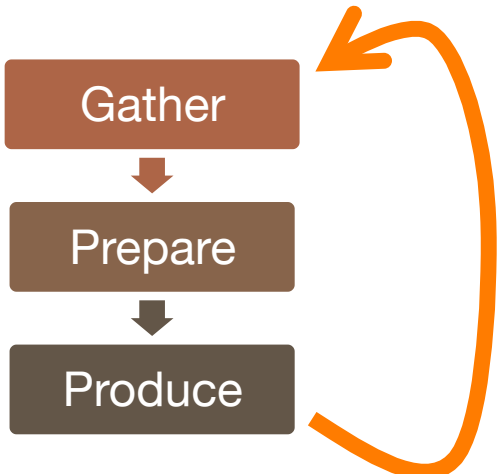
STORY TIME PLANNING

Data percolation: A model of data preparation and analysis

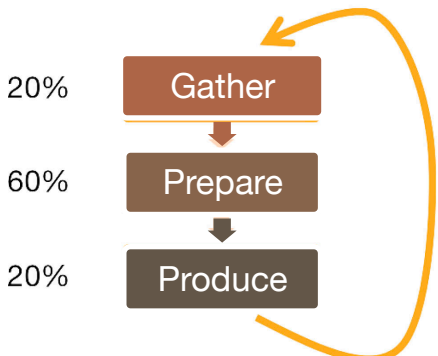


See also: the Data Journalism Handbook

Data percolation: A model of data preparation and analysis

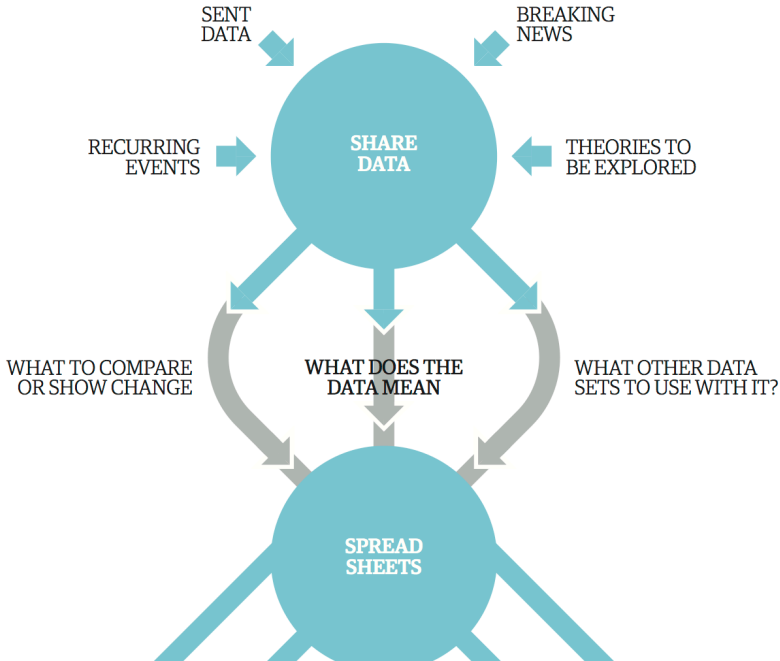


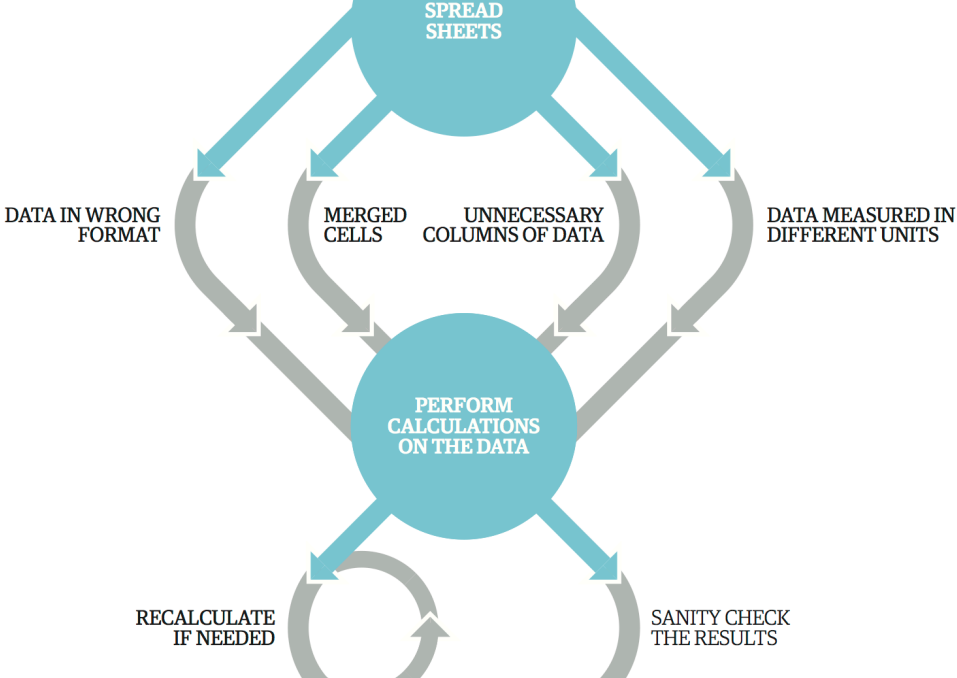
How should I budget my time?

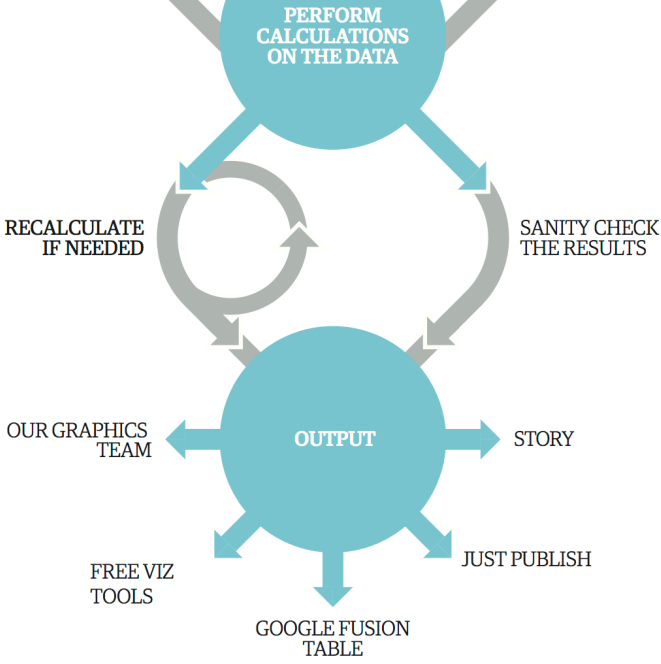


Finding a story is a creative process.

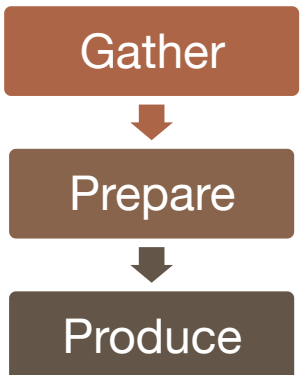
Let it percolate!


SENT DATA
BREAKING NEWS
RECURRING EVENTS
SHARE DATA
THEORIES TO BE EXPLORED
WHAT TO COMPARE OR SHOW CHANGE
WHAT DOES THE DATA MEAN
WHAT OTHER DATA SETS TO USE WITH IT?
SPREAD SHEETS


SPREAD SHEETS
DATA IN WRONG FORMAT
MERGED CELLS
UNNECESSARY COLUMNS OF DATA
DATA MEASURED IN DIFFERENT UNITS
PERFORM CALCULATIONS ON THE DATA
RECALCULATE IF NEEDED
SANITY CHECK THE RESULTS


PERFORM CALCULATIONS ON THE DATA
RECALCULATE IF NEEDED
SANITY CHECK THE RESULTS
OUTPUT
OUR GRAPHICS TEAM
STORY
FREE VIZ TOOLS
JUST PUBLISH
GOOGLE FUSION TABLE

How should I budget my time?



1.1 **FIND** reliable data sources

1.2 Understand its relevance and **LICENCE**

1.3 Visualise and **UNDERSTAND** your data

2.1 **CLEAN** your data

2.2 **TRANSFORM** it where useful

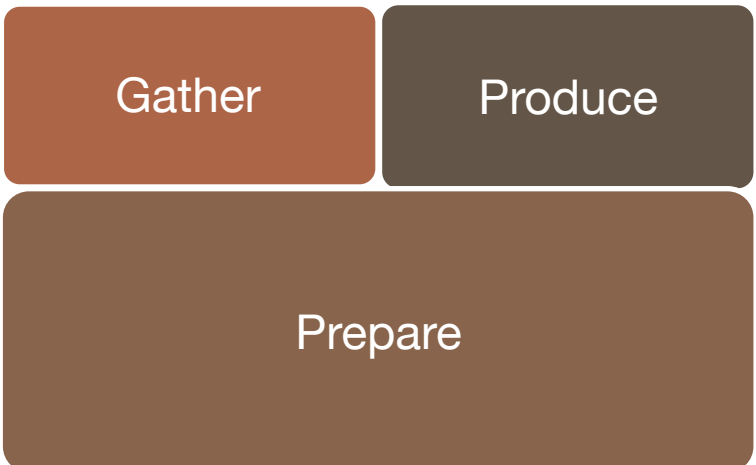
2.3 **COMBINE** it with other data sets

3.1 **REDUCE** and find the story

3.2 Think and understand the **CONTEXT**

3.3 Do your results pass a **SENSE-CHECK**?

Time planning



Gathering Data

Gather

What makes a trusted (data) source?

Plenty of data sources exist

17852 Results

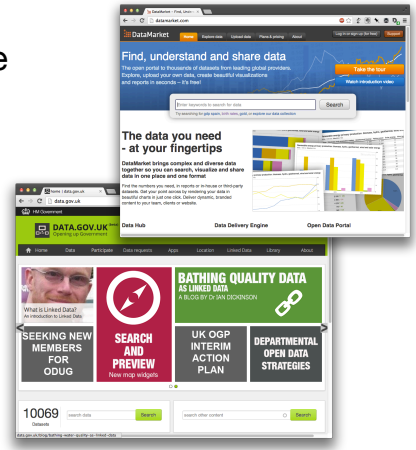
Office for National Statistics

Exercise: Sourcing reliable data

Starting with the links on the website, find a number of datasets that could be the source of a potential story.

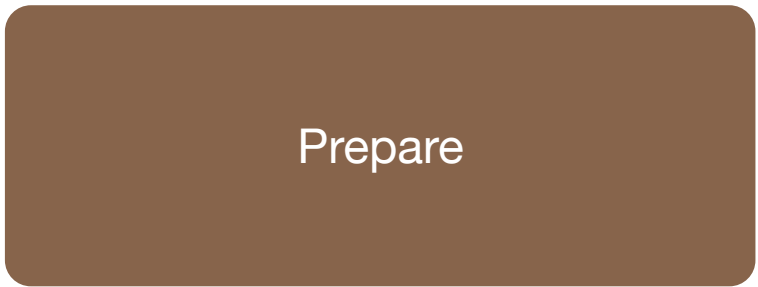
You could use the dataset you bought with you as a starting point.

What makes the data reliable and usable?
What makes the data relevant?
What quality is the data?
What processing is required?



EXPLORING: IS THERE A STORY?

Prepare (Stage 1)



2.1 **CLEAN**

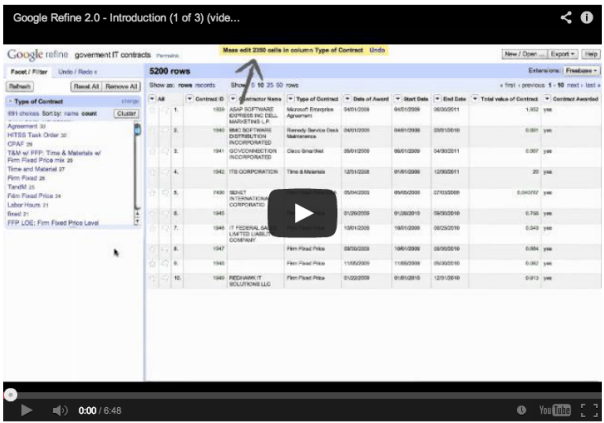
2.2 **TRANSFORM**

2.3 **COMBINE**

2.4 **ENRICH**

2.5 **ANALYSE**

Introducing Open Refine

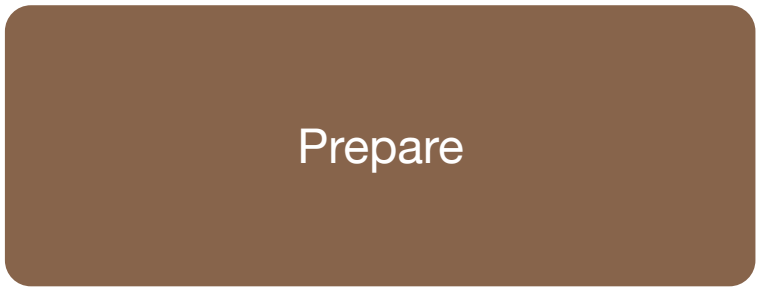


https://code.google.com/p/google-refine/

5 ideas for finding a story

1. **Create a ranking**
   Sort columns to explore top and bottom property prices
2. **Compare groups**
   Newly built properties to established residential building
3. **Use sum/average/min/max functions**
   Find the median price of a property
4. **Use pivot tables to compare groups**
   Compare the prices of different geographies in the UK
5. **Look for anomalies in the data**
   e.g. in the “date” or “postcode” columns

Prepare



2.1 **CLEAN**

2.2 **TRANSFORM**

2.3 **COMBINE**

2.4 **ENRICH**

2.5 **ANALYSE**

enigma.io Aggregator/Enabler

OFFICE OF FOREIGN LABOR CERTIFICATION

H-1B Visa Applications 2013

The H-1B is a non-immigrant visa in the United States under the Immigration and Nationality Act that allows U.S. employers to temporarily employ foreign workers in specialty occupations. Full Description

+ *See all nodes*
- UNITED STATES
- U.S. FEDERAL GOVERNMENT
- DEPARTMENT OF LABOR
- OFFICE OF FOREIGN LABOR CERTIFICATION
- H-1B VISA APPLICATIONS
- **H-1B VISA APPLICATIONS 2013**

H-1B VISA APPLICATIONS 2013 — 2,167 OF 442,277 ROWS — SHARE — EXPORT

Add filter… | WHOLE TABLE *contains* google ×

LCA Case Number	Status	Job Title	Employer Name	Employer Address	Employer City
I-200-12271-179543	DENIED	SOFTWARE ENGINE...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12251-455849	CERTIFIED	SOFTWARE ENGINE...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12265-043866	CERTIFIED	BUSINESS ANALYST	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12268-519668	CERTIFIED	SOFTWARE ENGINE...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12268-506660	CERTIFIED	WEB DEVELOPER	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12269-701878	CERTIFIED	SOFTWARE ENGINE...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW
I-200-12265-495825	CERTIFIED	INFORMATION SEC...	GOOGLE INC.	1600 AMPHITHEA...	MOUNTAIN VIEW

http://techcrunch.com/2013/05/01/and-the-winner-of-techcrunch-disrupt-ny-2013-is-enigma/

Prepare

opencorporates

The Open Database Of The Corporate World

Aggregator/Enabler

We have information on 70,597,888 companies

SEARCH

search companies search officers

Filter by jurisdiction

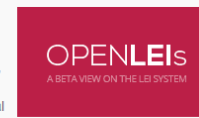
1,298	Abu Dhabi (UAE)
144,755	Alaska (US)
40,157	Albania
899,455	Arizona (US)
46,537	Aruba
165,582	Bahamas
99,185	Bahrain
88,563	Bangladesh

Just released: OpenCorporates API v0.3

Corporate network data, financial accounts, complex filters, and more. Read more

Announcing Open LEIs

Today, OpenCorporates announces a new sister website, Open LEIs, a user-friendly interface on the emerging Global Legal Entity Identifier System. Read more



New! Just added: Open corporate network data

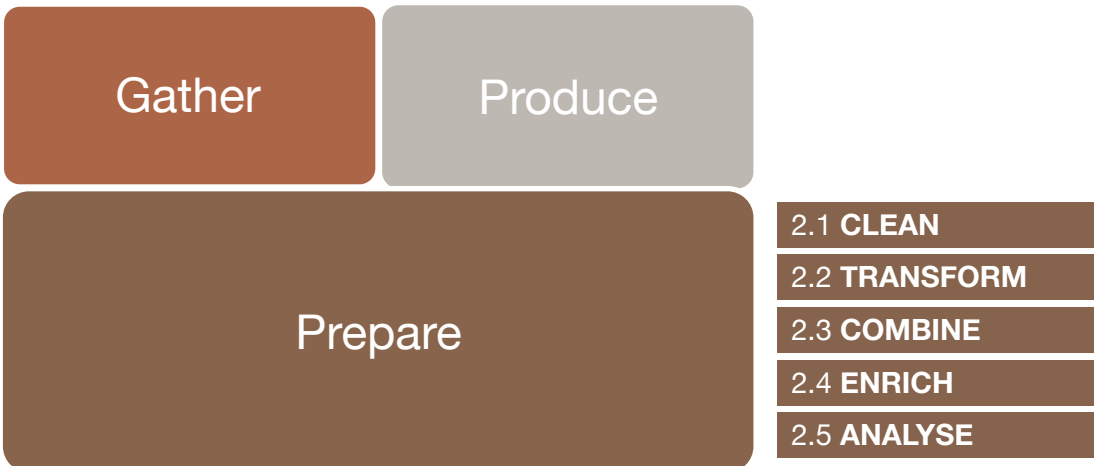
Read more about this important new feature

Sense-checking

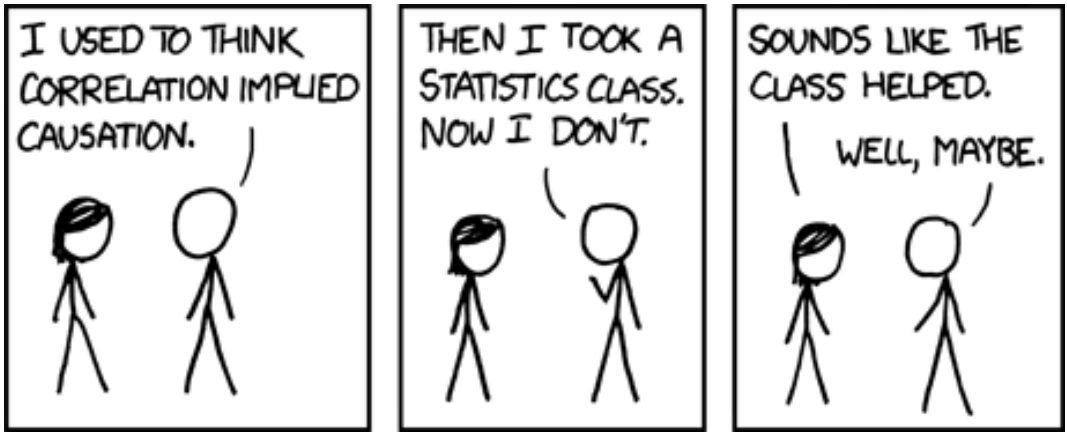
The best way to sense-check is to get a second pair of eyes to help you.

Any stories of common mistakes you'd like to share?

Recap



XKCD




Thank-you