Making spreadsheets machine-readable

With Dave from ScraperWiki

About Me / Databaker

David McKee
dragon@scraperwiki.com
@dragondave on Twitter

ScraperWiki

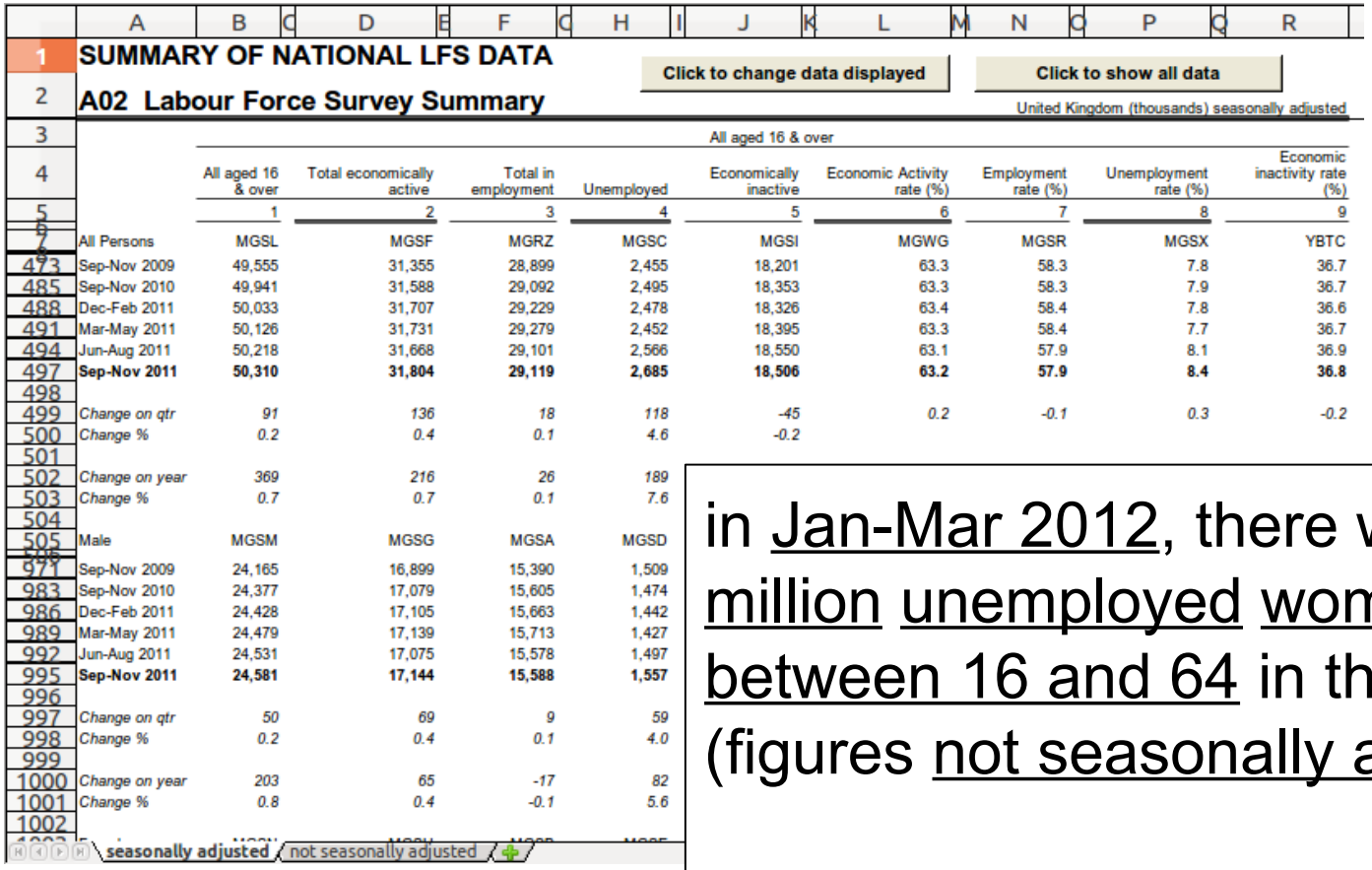
Background

Databaker takes complicated published spreadsheets, and converts them into tabular CSV.

By way of example we will feature a project we have delivered for the UK's Office of National Statistics (ONS).

ScraperWiki

The Problem



	A	B	D	F	H	J	L	N	P	R
1	SUMMARY OF NATIONAL LFS DATA		Click to change data displayed				Click to show all data			
2	A02 Labour Force Survey Summary								United Kingdom (thousands) seasonally adjusted	
3						All aged 16 & over				
4		All aged 16 & over	Total economically active	Total in employment	Unemployed	Economically inactive	Economic Activity rate (%)	Employment rate (%)	Unemployment rate (%)	Economic inactivity rate (%)
5		1	2	3	4	5	6	7	8	9
7	All Persons	MGSL	MGSF	MGRZ	MGSC	MGSI	MGWG	MGSR	MGSX	YBTC
473	Sep-Nov 2009	49,555	31,355	28,899	2,455	18,201	63.3	58.3	7.8	36.7
485	Sep-Nov 2010	49,941	31,588	29,092	2,495	18,353	63.3	58.3	7.9	36.7
488	Dec-Feb 2011	50,033	31,707	29,229	2,478	18,326	63.4	58.4	7.8	36.6
491	Mar-May 2011	50,126	31,731	29,279	2,452	18,395	63.3	58.4	7.7	36.7
494	Jun-Aug 2011	50,218	31,668	29,101	2,566	18,550	63.1	57.9	8.1	36.9
497	**Sep-Nov 2011**	**50,310**	**31,804**	**29,119**	**2,685**	**18,506**	**63.2**	**57.9**	**8.4**	**36.8**
498										
499	*Change on qtr*	*91*	*136*	*18*	*118*	*-45*	*0.2*	*-0.1*	*0.3*	*-0.2*
500	*Change %*	*0.2*	*0.4*	*0.1*	*4.6*	*-0.2*				
501										
502	*Change on year*	*369*	*216*	*26*	*189*					
503	*Change %*	*0.7*	*0.7*	*0.1*	*7.6*					
504										
505	Male	MGSM	MGSG	MGSA	MGSD					
971	Sep-Nov 2009	24,165	16,899	15,390	1,509					
983	Sep-Nov 2010	24,377	17,079	15,605	1,474					
986	Dec-Feb 2011	24,428	17,105	15,663	1,442					
989	Mar-May 2011	24,479	17,139	15,713	1,427					
992	Jun-Aug 2011	24,531	17,075	15,578	1,497					
995	**Sep-Nov 2011**	**24,581**	**17,144**	**15,588**	**1,557**					
996										
997	*Change on qtr*	*50*	*69*	*9*	*59*					
998	*Change %*	*0.2*	*0.4*	*0.1*	*4.0*					
999										
1000	*Change on year*	*203*	*65*	*-17*	*82*					
1001	*Change %*	*0.8*	*0.4*	*-0.1*	*5.6*					
1002										

seasonally adjusted | not seasonally adjusted

in Jan-Mar 2012, there were 1.09 million unemployed women aged between 16 and 64 in the UK (figures not seasonally adjusted)

How do you use this data?

ScraperWiki

How do we use this data?

- Too difficult! Keep the complicated shape

- Laborious, manual process to convert
- Complex, error-prone manual workflow

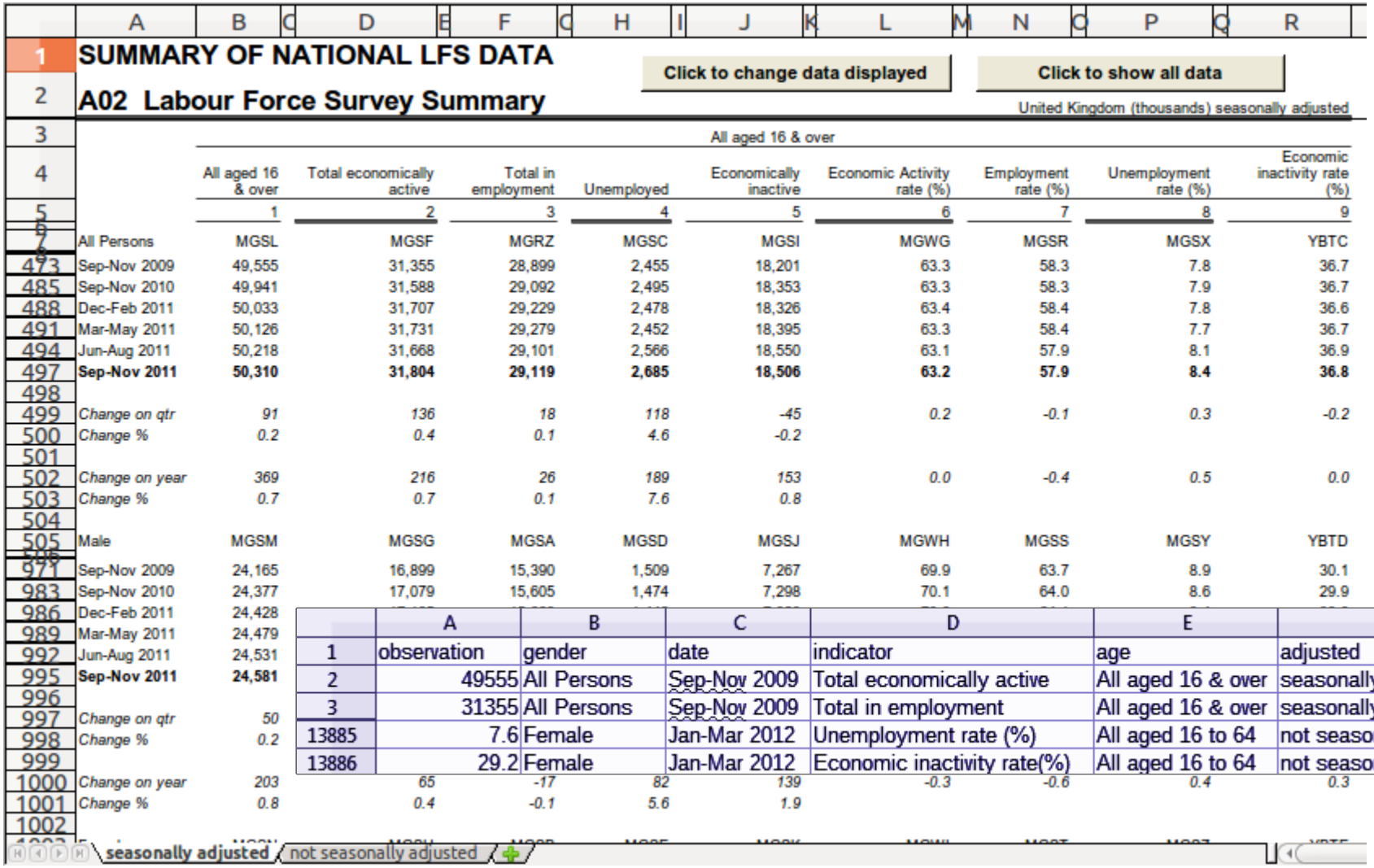
- Maybe it shouldn't have been in a spreadsheet to start with… but too late now!

ScraperWiki

What goes wrong?

- Ad-hoc downstream mangling of data vs. approved data in readily usable format
- Analysis based on what's easy, rather than what's important
- Not easy to import into existing tools to inform decisions - slows decision making

ScraperWiki



SUMMARY OF NATIONAL LFS DATA

Click to change data displayed | Click to show all data

A02 Labour Force Survey Summary

United Kingdom (thousands) seasonally adjusted

	All aged 16 & over								
	All aged 16 & over	Total economically active	Total in employment	Unemployed	Economically inactive	Economic Activity rate (%)	Employment rate (%)	Unemployment rate (%)	Economic inactivity rate (%)
	1	2	3	4	5	6	7	8	9
All Persons	MGSL	MGSF	MGRZ	MGSC	MGSI	MGWG	MGSR	MGSX	YBTC
Sep-Nov 2009	49,555	31,355	28,899	2,455	18,201	63.3	58.3	7.8	36.7
Sep-Nov 2010	49,941	31,588	29,092	2,495	18,353	63.3	58.3	7.9	36.7
Dec-Feb 2011	50,033	31,707	29,229	2,478	18,326	63.4	58.4	7.8	36.6
Mar-May 2011	50,126	31,731	29,279	2,452	18,395	63.3	58.4	7.7	36.7
Jun-Aug 2011	50,218	31,668	29,101	2,566	18,550	63.1	57.9	8.1	36.9
Sep-Nov 2011	**50,310**	**31,804**	**29,119**	**2,685**	**18,506**	**63.2**	**57.9**	**8.4**	**36.8**
Change on qtr	*91*	*136*	*18*	*118*	*-45*	*0.2*	*-0.1*	*0.3*	*-0.2*
Change %	*0.2*	*0.4*	*0.1*	*4.6*	*-0.2*				
Change on year	*369*	*216*	*26*	*189*	*153*	*0.0*	*-0.4*	*0.5*	*0.0*
Change %	*0.7*	*0.7*	*0.1*	*7.6*	*0.8*				
Male	MGSM	MGSG	MGSA	MGSD	MGSJ	MGWH	MGSS	MGSY	YBTD
Sep-Nov 2009	24,165	16,899	15,390	1,509	7,267	69.9	63.7	8.9	30.1
Sep-Nov 2010	24,377	17,079	15,605	1,474	7,298	70.1	64.0	8.6	29.9
Dec-Feb 2011	24,428								
Mar-May 2011	24,479								
Jun-Aug 2011	24,531								
Sep-Nov 2011	**24,581**								
Change on qtr	*50*								
Change %	*0.2*								
Change on year	*203*	*65*	*-17*	*82*	*139*	*-0.3*	*-0.6*	*0.4*	*0.3*
Change %	*0.8*	*0.4*	*-0.1*	*5.6*	*1.9*				

seasonally adjusted | not seasonally adjusted

	A	B	C	D	E	F
1	observation	gender	date	indicator	age	adjusted
2	49555	All Persons	Sep-Nov 2009	Total economically active	All aged 16 & over	seasonally adjusted
3	31355	All Persons	Sep-Nov 2009	Total in employment	All aged 16 & over	seasonally adjusted
13885	7.6	Female	Jan-Mar 2012	Unemployment rate (%)	All aged 16 to 64	not seasonally adjusted
13886	29.2	Female	Jan-Mar 2012	Economic inactivity rate(%)	All aged 16 to 64	not seasonally adjusted

ScraperWiki

Aim: convert human-readable spreadsheets into something truly machine readable

Beautiful, dangerous spreadsheets

- Focussed on the printed page, not data reuse
- Need to migrate away to something else - genuinely tabular CSV is portable
- Specifically, portable to http://www.ons.gov.uk/ons/data/web/explorer

ScraperWiki

Break the problem down

- Identify cells which act as headers
  - Want to be robust against minor changes
    - especially inserted rows
  - Select each dimension's headers in turn
- Identify cells which act as values
  - May want to do this from the intersection of headers
- Determine correct set of header cells for each value
  - Needs a simple but robust rule

ScraperWiki

Identifying cells

- ‘Bags’ of cells
  - non-contiguous
  - limited to a single sheet

- How to describe where they are?
  - Gender: all the cells which are ‘All Persons’, ‘Male’ or ‘Female’
  - Date: all the cells below ‘All Persons’, except blank, italic cells or the Gender cells we already described

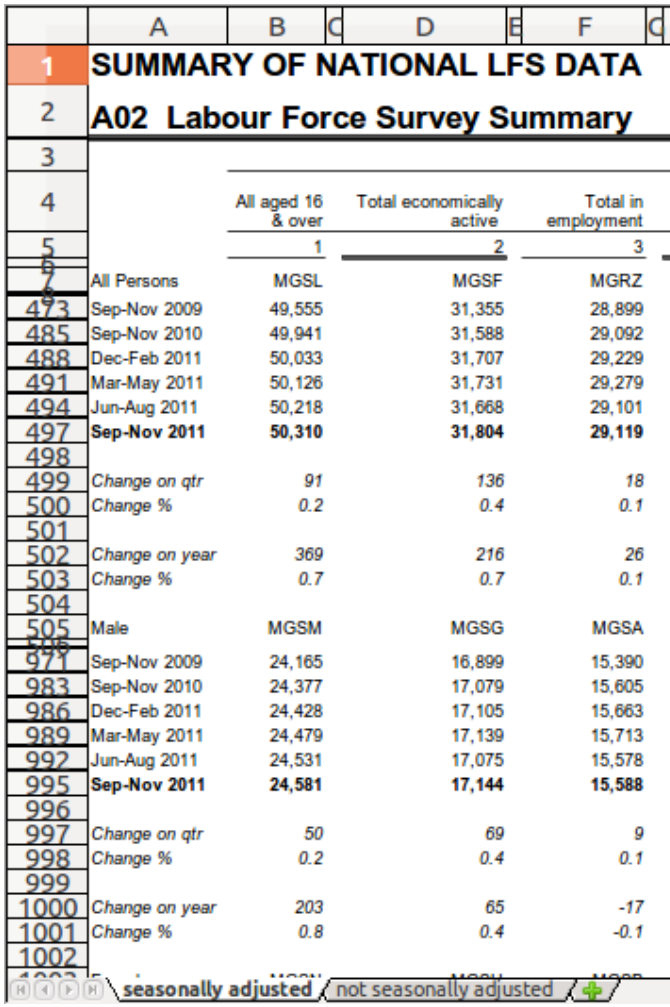
```
gender = tab.filter(one_of(['Male', 'Female', 'All Persons'])
date = tab.filter(‘All Persons’).fill(DOWN)
       .not_blank().not_italic().difference(gender)
```

ScraperWiki

What about the values?

- Non-contiguous
- Contains unwanted numbers

- Below indicators, right of dates
  - `values = indicators.waffle(dates)`



	A	B	D	F
1	SUMMARY OF NATIONAL LFS DATA			
2	A02 Labour Force Survey Summary			
3				
4		All aged 16 & over	Total economically active	Total in employment
5		1	2	3
7	All Persons	MGSL	MGSF	MGRZ
473	Sep-Nov 2009	49,555	31,355	28,899
485	Sep-Nov 2010	49,941	31,588	29,092
488	Dec-Feb 2011	50,033	31,707	29,229
491	Mar-May 2011	50,126	31,731	29,279
494	Jun-Aug 2011	50,218	31,668	29,101
497	**Sep-Nov 2011**	**50,310**	**31,804**	**29,119**
498				
499	*Change on qtr*	*91*	*136*	*18*
500	*Change %*	*0.2*	*0.4*	*0.1*
501				
502	*Change on year*	*369*	*216*	*26*
503	*Change %*	*0.7*	*0.7*	*0.1*
504				
505	Male	MGSM	MGSG	MGSA
971	Sep-Nov 2009	24,165	16,899	15,390
983	Sep-Nov 2010	24,377	17,079	15,605
986	Dec-Feb 2011	24,428	17,105	15,663
989	Mar-May 2011	24,479	17,139	15,713
992	Jun-Aug 2011	24,531	17,075	15,578
995	**Sep-Nov 2011**	**24,581**	**17,144**	**15,588**
996				
997	*Change on qtr*	*50*	*69*	*9*
998	*Change %*	*0.2*	*0.4*	*0.1*
999				
1000	*Change on year*	*203*	*65*	*-17*
1001	*Change %*	*0.8*	*0.4*	*-0.1*
1002				

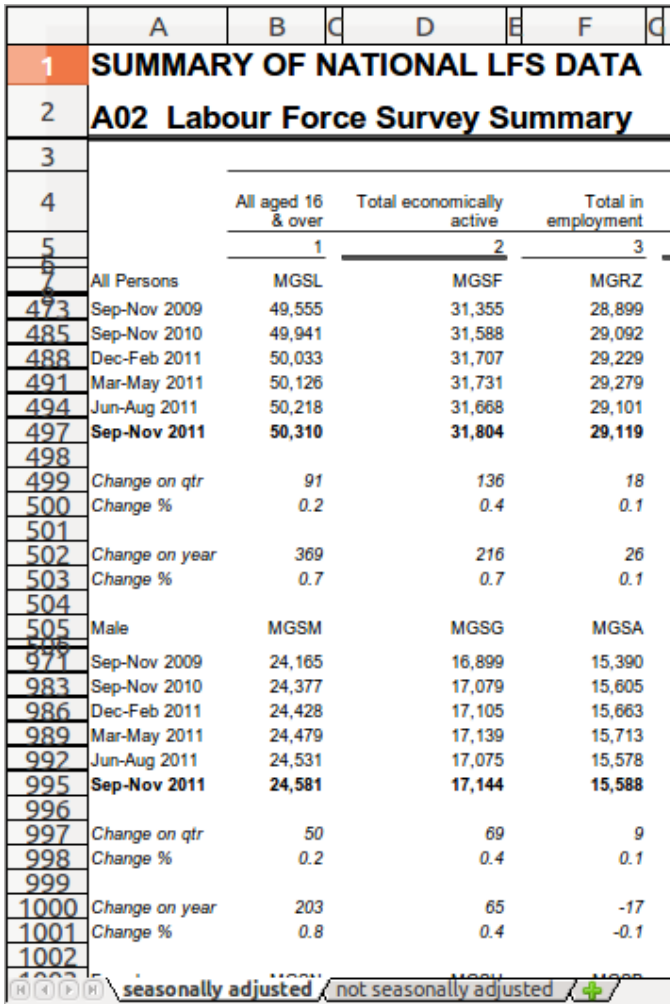
seasonally adjusted | not seasonally adjusted

ScraperWiki

Matching values to dimension cells

- Good practice - **above** or to the **left**!

- Header either **directly** up/left of value or the **closest**
  - `date.dimension("date", DIRECTLY, LEFT)`
  - `indicator.dimension("indicator", DIRECTLY, ABOVE)`
  - `gender.dimension("gender", CLOSEST, ABOVE)`



	A	B	D	F
1	SUMMARY OF NATIONAL LFS DATA			
2	A02 Labour Force Survey Summary			
3				
4		All aged 16 & over	Total economically active	Total in employment
5		1	2	3
473	All Persons	MGSL	MGSF	MGRZ
485	Sep-Nov 2009	49,555	31,355	28,899
488	Sep-Nov 2010	49,941	31,588	29,092
491	Dec-Feb 2011	50,033	31,707	29,229
494	Mar-May 2011	50,126	31,731	29,279
497	Jun-Aug 2011	50,218	31,668	29,101
498	**Sep-Nov 2011**	**50,310**	**31,804**	**29,119**
499				
500	*Change on qtr*	*91*	*136*	*18*
501	*Change %*	*0.2*	*0.4*	*0.1*
502				
503	*Change on year*	*369*	*216*	*26*
504	*Change %*	*0.7*	*0.7*	*0.1*
505				
971	Male	MGSM	MGSG	MGSA
983	Sep-Nov 2009	24,165	16,899	15,390
986	Sep-Nov 2010	24,377	17,079	15,605
989	Dec-Feb 2011	24,428	17,105	15,663
992	Mar-May 2011	24,479	17,139	15,713
995	Jun-Aug 2011	24,531	17,075	15,578
996	**Sep-Nov 2011**	**24,581**	**17,144**	**15,588**
997				
998	*Change on qtr*	*50*	*69*	*9*
999	*Change %*	*0.2*	*0.4*	*0.1*
1000				
1001	*Change on year*	*203*	*65*	*-17*
1002	*Change %*	*0.8*	*0.4*	*-0.1*

seasonally adjusted | not seasonally adjusted

ScraperWiki

Selectors

- You can only select what you can describe
- Select by
  - text, formatting, borders, font, font size, cell ref.
- Combine groups of cells
  - set operators (union, difference, intersection)
  - get cells below this bag and right of another bag
- Get other cells nearby
  - relative movement (the cells one beneath these)
  - fill (all the cells to the right)
- All these transform a bag of cells into another bag of cells

Previewing

- Check your work!

SUMMARY OF NATIONAL LFS DATA

A02 Labour Force Survey Summary — United Kingdom (thousands) seasonally adjusted

	All aged 16 & over								
	All aged 16 &	Total economically a	Total in employ	Unemployed	Economically ina	Economic Activity ra	Employment rat	Unemployment rate	Economic inactivity rate (%)
	1	2	3	4	5	6	7	8	9
All Persons	MGSL	MGSF	MGRZ	MGSC	MGSI	MGWG	MGSR	MGSX	YBTC
Jan-Mar 2010	49,687	31335244.6564892	28824552.563	2510692.093	18351663.3435	63.0653947243	58.012369301	8.0123583544	36.9346052757
Jan-Mar 2011	50,064	31695579.6674312	29240308.249	2455271.418	18368545.3326	63.3099643056	58.405711174	7.7464158855	36.6900356944
Apr-Jun 2011	50,157	31758847.7742467	29265215.88	2493631.894	18397863.2258	63.3192391228	58.347557678	7.8517706681	36.6807608772
Jul-Sep 2011	50,248	31690970.8108037	29068520.974	2622449.837	18557304.1892	63.0687736262	57.849788822	8.2750694269	36.9312263738
Oct-Dec 2011	50,339	31799559.6244686	29128872.387	2670687.237	18539628.3755	63.1705851602	57.865201137	8.3985038439	36.8294148398
Jan-Mar 2012	**50,431**	31858843.3296458	29233403.676	2625439.653	18572001.6704	63.1733284058	57.967308849	8.2408505115	36.8266715942
Change on qtr	*92*	*59*	*105*	*-45*	*32*	*0.0*	*0.1*	*-0.2*	*0.0*
Change %	*0.2*	*0.2*	*0.4*	*-1.7*	*0.2*				
Change on year	*367*	*163*	*-7*	*170*	*203*	*-0.1*	*-0.4*	*0.5*	*0.1*
Change %	*0.7*	*0.5*	*0.0*	*6.9*	*1.1*				
Male	MGSM	MGSG	MGSA	MGSD	MGSJ	MGWH	MGSS	MGSY	YBTD
Jan-Mar 2010	24,238	16903241.1945317	15362508.233	1540732.962	7334418.80547	69.7395754975	63.382802765	9.11501495	30.2604245025
Jan-Mar 2011	24,445	17079751.7811117	15651748.363	1428003.418	7365462.21889	69.8695122125	64.027863955	8.3607972534	30.1304877875
Apr-Jun 2011	24,497	17146290.7188694	15700433.294	1445857.425	7350376.28113	69.9943821699	64.092120343	8.4324793567	30.0056178301
Jul-Sep 2011	24,547	17079371.4597125	15547372.581	1531998.879	7467597.54029	69.5783314824	63.337239643	8.9698785621	30.4216685176
Oct-Dec 2011	24,597	17138427.1629446	15590246.704	1548180.459	7458413.83706	69.6773506929	63.383125922	9.0333870447	30.3226493071
Jan-Mar 2012	**24,647**	17176013.16831	15670284.125	1505729.043	7471439.83169	69.6867671005	63.577701618	8.7664641888	30.3132328995
Change on qtr	*51*	*38*	*80*	*-42*	*13*	*0.0*	*0.2*	*-0.3*	*0.0*
Change %	*0.2*	*0.2*	*0.5*	*-2.7*	*0.2*				
Change on year	*202*	*96*	*19*	*78*	*106*	*-0.2*	*-0.5*	*0.4*	*0.2*
Change %	*0.8*	*0.6*	*0.1*	*5.4*	*1.4*				

Complications

- Headers in the wrong place - fake spans

		Top Header					Top Header		
Sub-header	Sub-header	Sub-header	Sub-header	Sub-header	Sub-header	Sub-header	Sub-header	Sub-header	Sub-header
22	33	44	55	66	77	88	99	111	222

  - ‘66’ needs to look left, ‘77’ right

- Real spans are not a problem
  - can just use the top-left cell
  - can also get the whole span as a bag of cells

ScraperWiki

Recipes

- Scripts to convert spreadsheets
  - Technical users who don't consider themselves programmers can write them
  - All the power of Python

```
from databaker.constants import *

def per_file(tableset):
    return "*"

def per_tab(tab):
    obs = tab.filter("MGSL").assert_one().shift(DOWN).fill(RIGHT).fill(DOWN).is_number().is_not_italic()

    tab.col('A').one_of(['Male', 'Female', 'All Persons']).dimension('gender', CLOSEST, ABOVE)
    tab.col('A').is_date().dimension(TIME, DIRECTLY, LEFT)
    tab.regex("All aged .*").dimension('ages', CLOSEST, UP)
    tab.filter("Total economically active").fill(LEFT).fill(RIGHT).is_not_blank().dimension('indicator', DIRECTLY, ABOVE)

    tab.dimension('adjusted_yn', tab.name)
    return obs
```

ScraperWiki

Bigger Picture

- Critical task of correctly allocating headers done
- Training critical - they need to write their own recipes
- Lots of post processing to get identifiers consistent and correct
- Currently ONS-specific output format, intend to change

ScraperWiki

More information

- Databaker is open source
- Databaker documentation / downloads
  - https://scraperwiki.github.io/eot-docs/

- dragon@scraperwiki.com
- @dragondave on Twitter

ScraperWiki

Thank you!

- Wouldn’t have happened without:
  - UN OCHA & ONS - spreadsheet problems, £
  - Open Source
  - OKFN - libraries, organising csv,conf

- and **you!** for listening

ScraperWiki