ESPON

EUROPEAN UNION

Co-financed by the European Regional Development Fund

Inspire Policy Making with Territorial Evidence

// Open data:
A data journalist's best friend

Doug Dowson, *The Economist*

What is Data Journalism?

“Data-driven journalism is the future.”

— **Tim Berners-Lee**,

inventor of the World Wide Web

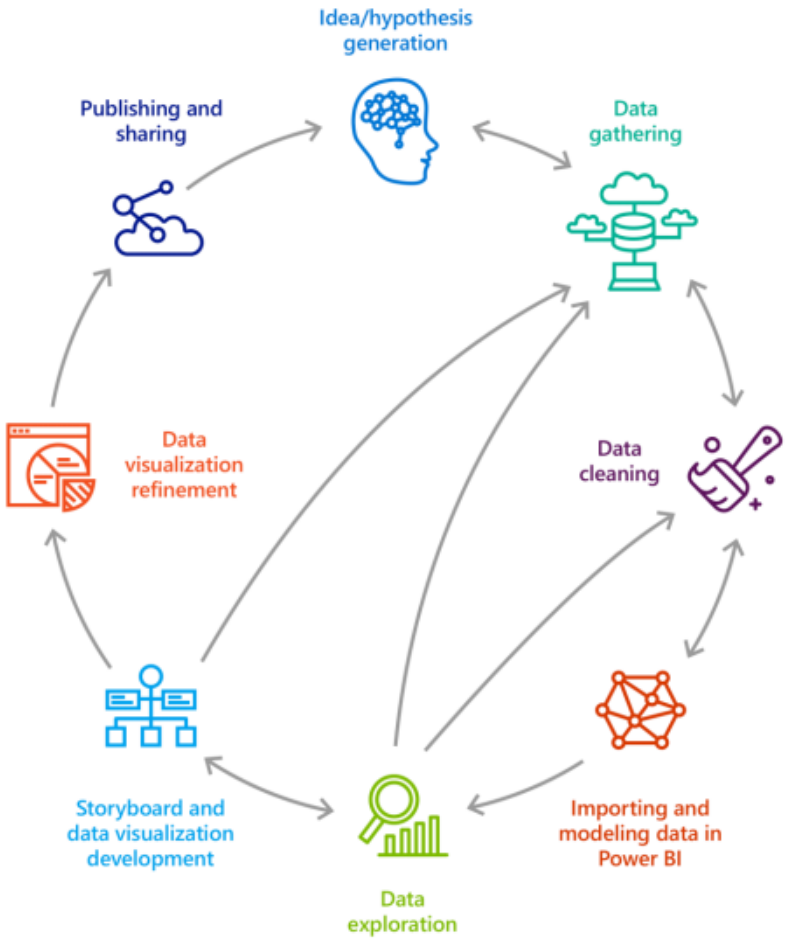
What is Data Journalism?

- 1970s – **“Precision Journalism”**: applying social science research methods to journalism
- 1980s – **“Computer Assisted Reporting (CAR)”**: using computers to gather, clean and analyse data for use in news stories

- Today – **“Data journalism”**:
  - 1) news stories that require the use of computers and software to collect, clean and analyse data
  - 2) apply social science research methods to draw original conclusions and
  - 3) whose results are often presented visually

What is Data Journalism?

- Data → Analysis → Conclusions
- Sometimes, a data-driven story **begins with a dataset**
- Other times, a story **begins with a question** and the data follow



Data Journalism Examples

The Washington Post: Fatal Force

- In 2014, after a series of civilian killings by American police officers, *Washington Post* staff writers learned that there were no official statistics about such fatalities.

- So they decided to collect the data themselves…

Data Journalism Examples

- The result was “Fatal Force”, a database of all fatal shootings by American police officers in the line of duty:

Data Journalism Examples

- The data showed:
  - About one-quarter of those fatally shot had a history of mental illness
  - 55 officers involved in fatal shootings in 2015 had previously been involved in a deadly incident while on duty
  - Most people (74%) killed by police were armed with guns or were killed after attacking police officers or civilians
- In 2016, the series was awarded the **Pulitzer Prize**

Data Journalism Examples

The New York Times: Nike Vaporflys

- In 2018, the *New York Times* wanted to test Nike's claim that its $250-a-pair Zoom Vaporfly running shoes were significantly better than the competition

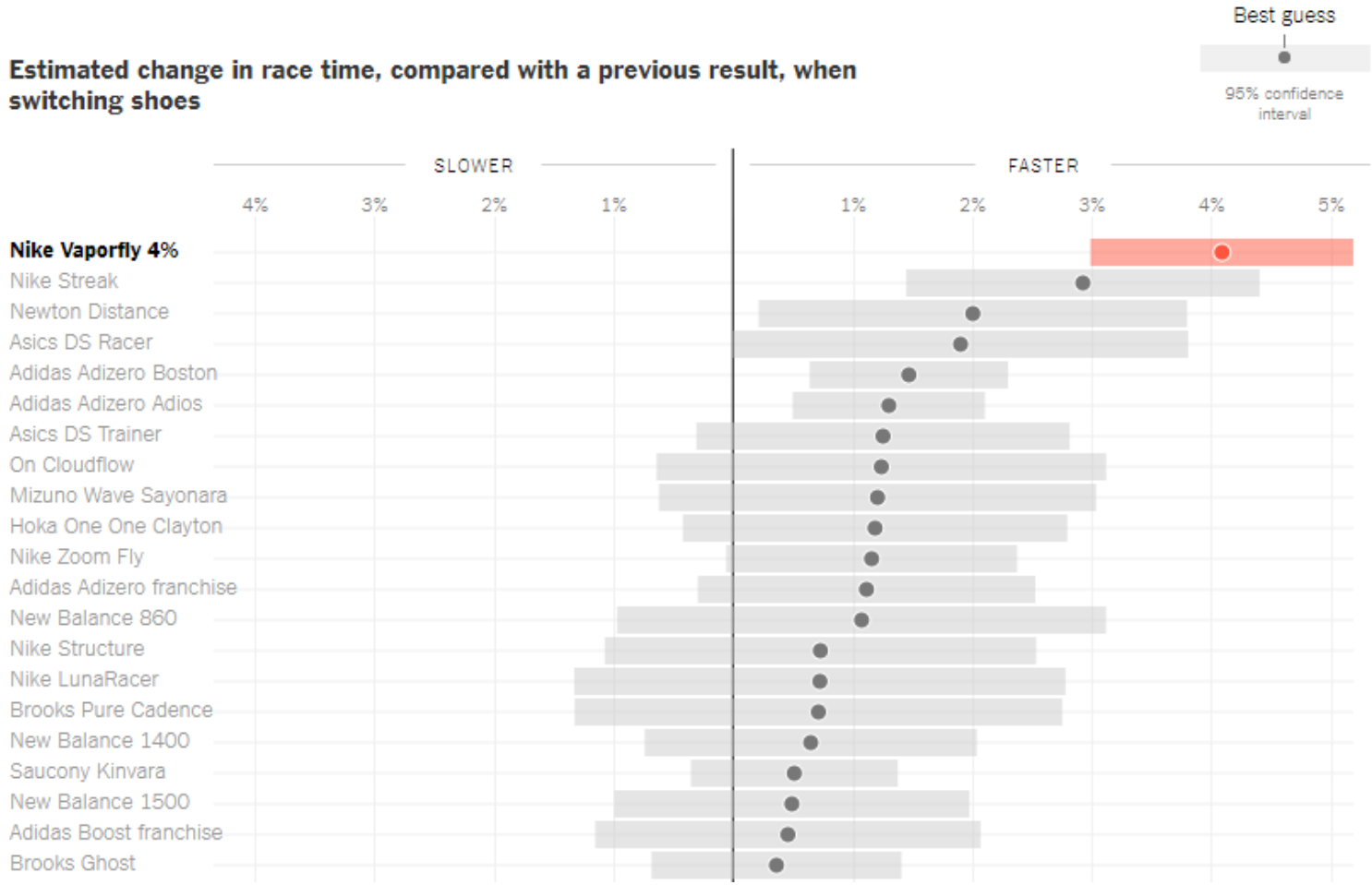
Data Journalism Examples

- The Times collected **public data** on some **500,000** marathon and half marathon **races** from Strava, a fitness app

- After running a statistical model that controlled for variables such as age, gender, and weather, they concluded…

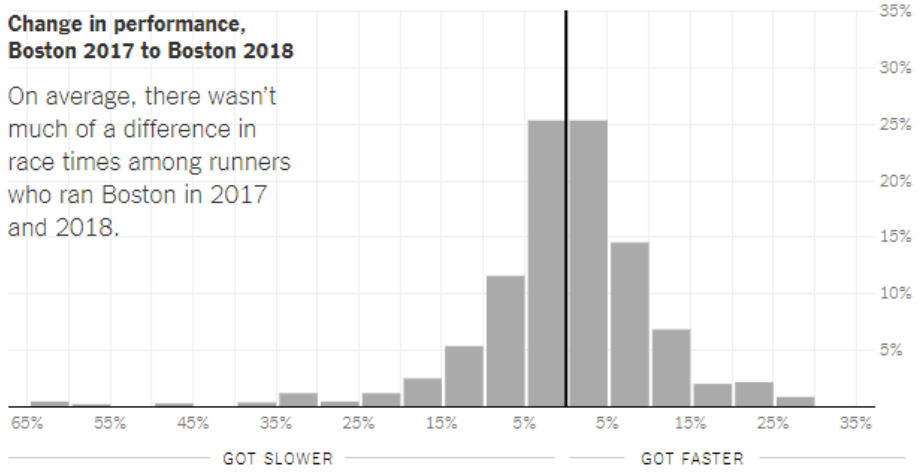
Data Journalism Examples

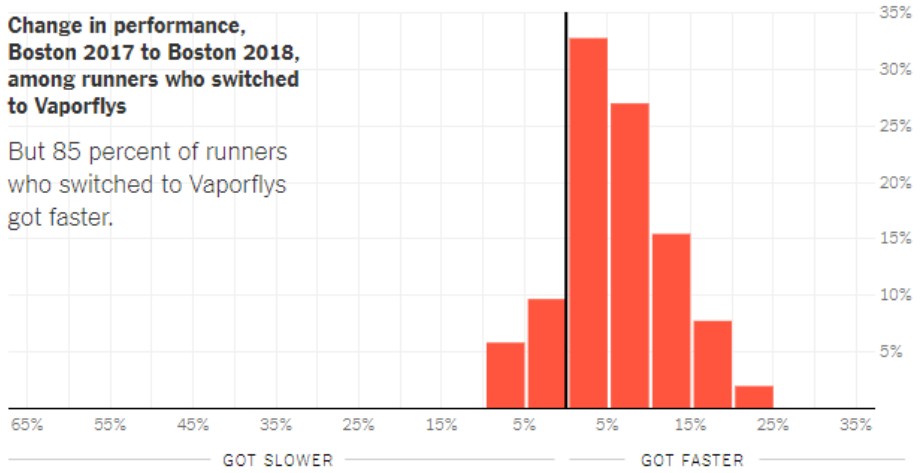
- …that Vaporflys boost performance by **3-4%**



Data Journalism Examples

- **85%** of runners who switched to Vaporflys between the 2017 and 2018 Boston marathons got faster





Open Data

“Open data is data that can be freely used, re-used and redistributed by anyone”

— **Open Data Handbook**

Open Data

- Available to anyone, **free** of charge
- Available **without license restrictions** to use, reuse, and redistribute
- Available in a useful format:
  - **Electronic and machine-readable**: structured data in a standardised format that can be read and processed by a computer, such as CSV, JSON, XML, etc.
- Available to **download in bulk**

Case Study: CEO Pay

- In May 2018 I wrote about CEO pay

- America's Dodd–Frank Act, a financial-reform law signed in 2010, included a provision requiring publicly-listed firms to report:
  - The annual compensation of their **bosses**
  - The compensation of their **median employees**
  - The **ratio** of these two numbers

Case Study: CEO Pay

- But collecting these data is difficult because each figure is hosted on a separate page of the SEC's website
- Here's **Apple's** CEO pay ratio data for 2018:

CEO Pay Ratio—2018

The 2018 annual total compensation of our CEO was $15,682,219, the 2018 annual total compensation of our median compensated employee was $55,426, and the ratio of these amounts is 283 to 1.

We determined our median compensated employee by using base salary, bonuses, commissions, and grant date fair value of equity awards granted to employees in 2018. We applied this measure to our global employee population as of the last day of our 2018 fiscal year and annualized base salaries for permanent full-time and part-time employees that did not work the full year. Once we determined our median compensated employee using these measures, we calculated the employee's 2018 annual total compensation using the same methodology that is used to calculate our CEO's annual total compensation in the table entitled "Summary Compensation Table—2018, 2017, and 2016."

Apple Inc. | 2019 Proxy Statement | 46

Case Study: CEO Pay

How I collected the data:

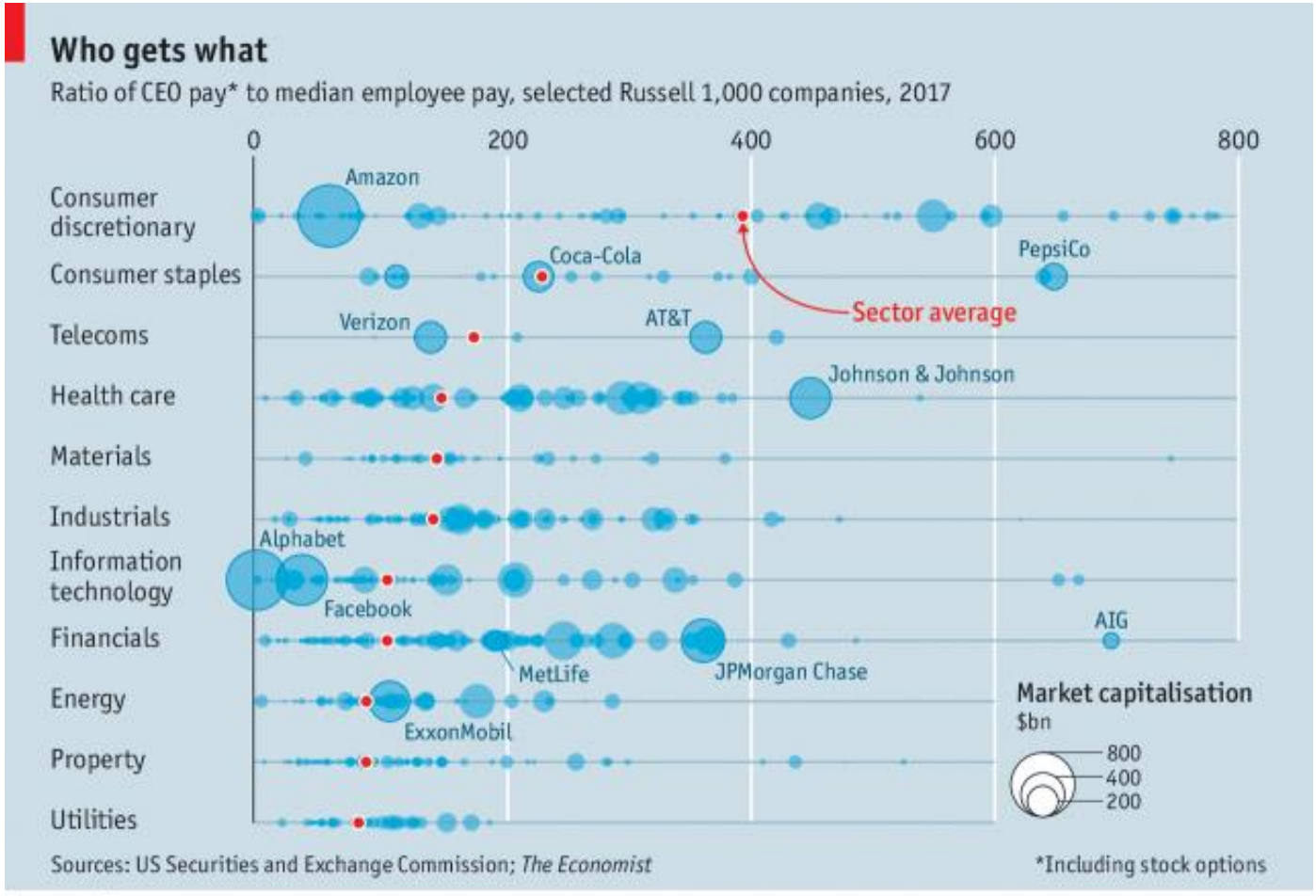
- I used **Python**, a programming language, to scrape the SEC's website for hundreds of URLs
- I hired two data-entry professionals to **download** each SEC filing, manually search for the relevant data, and **copy and paste** them into a spreadsheet

Case Study: CEO Pay

Findings:

- America’s largest publicly listed firms (those worth at least $1bn) on average paid their chief executives **130 times** more than their typical workers in 2017
- CEO pay ratios are influenced by company size, industry and the share of employees that are part-time or temporary

Case Study: CEO Pay



- Yet there are still **large disparities** in pay among similar firms

Case Study: CEO Pay

How open are the data?

- Free of charge? **YES**
- Without license restrictions? **YES**
- Electronic and machine-readable? **NO**
- Available to download in bulk? **NO**

Case Study: Amazon

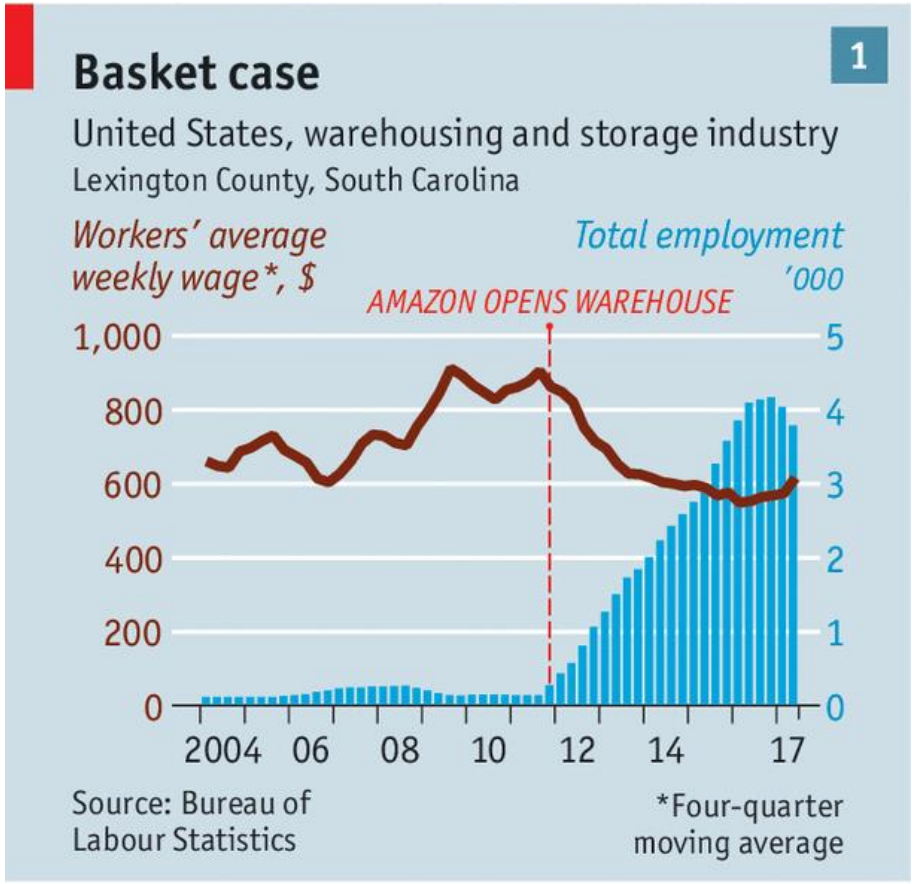
- In January 2018, I wrote about how **Amazon** pays its warehouse employees

- Using **official figures** from America's Bureau of Labour Statistics, I found…

Case Study: Amazon

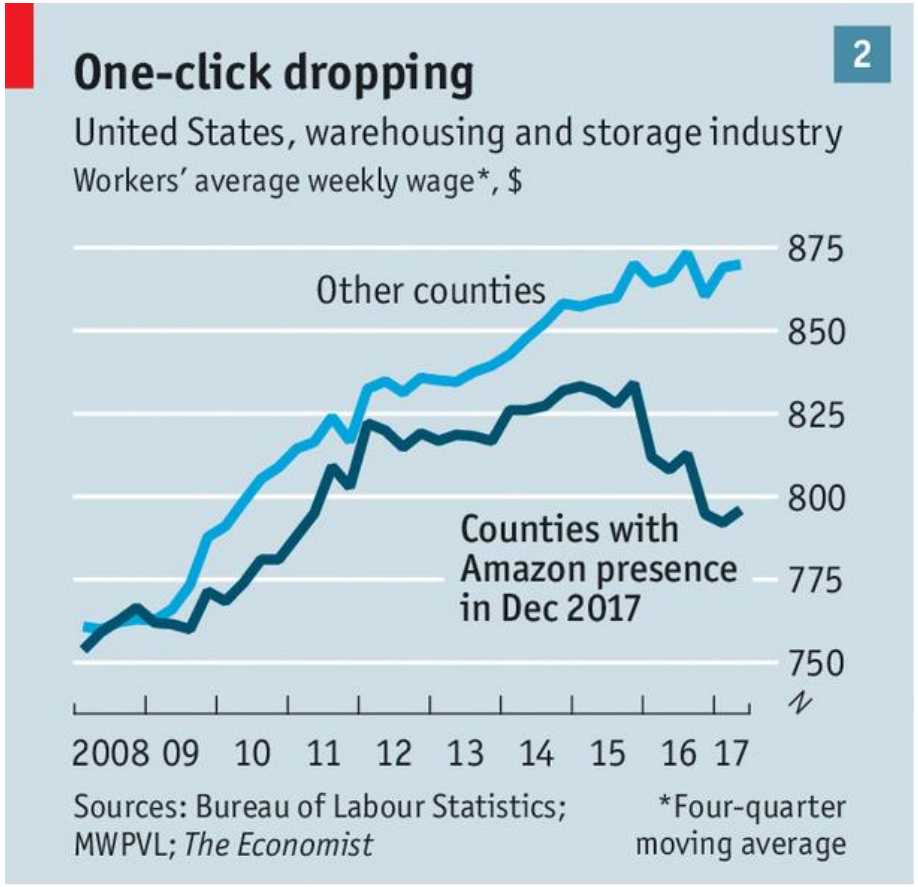
- After Amazon opens a warehouse, local wages for warehouse workers fall by an average of **3%**

- In Lexington County, SC earnings fell by over **30%**



Case Study: Amazon

- In places where Amazon operates, warehouse workers earn about **10% less** than similar workers employed elsewhere



Economist.com

Case Study: Amazon

- The data for this story were all available on the BLS website

QCEW NAICS-Based Data Files (1975 - most recent)

Excel Files	CSVs By Area		CSVs By Industry		CSVs Single Files		CSVs By Size	Legacy Flat Files
County High-Level	Quarterly	Annual Averages	Quarterly	Annual Averages	Quarterly	Annual Averages	First Quarter	All ENB/END
File Layout	File Layout	File Layout	File Layout	File Layout	File Layout	File Layout	File Layout	File Layouts
2018	2018	N/A	2018	N/A	2018	N/A	2018	2018
2017	2017	2017	2017	2017	2017	2017	2017	2017
2016	2016	2016	2016	2016	2016	2016	2016	2016
2015	2015	2015	2015	2015	2015	2015	2015	2015
2014	2014	2014	2014	2014	2014	2014	2014	2014
2013	2013	2013	2013	2013	2013	2013	2013	2013
2012	2012	2012	2012	2012	2012	2012	2012	2012
2011	2011	2011	2011	2011	2011	2011	2011	2011
2010	2010	2010	2010	2010	2010	2010	2010	2010
2009	2009	2009	2009	2009	2009	2009	2009	2009
2008	2008	2008	2008	2008	2008	2008	2008	2008
2007	2007	2007	2007	2007	2007	2007	2007	2007
2006	2006	2006	2006	2006	2006	2006	2006	2006
2005	2005	2005	2005	2005	2005	2005	2005	2005
2004	2004	2004	2004	2004	2004	2004	2004	2004
2003	2003	2003	2003	2003	2003	2003	2003	2003
2002	2002	2002	2002	2002	2002	2002	2002	2002
2001	2001	2001	2001	2001	2001	2001	2001	2001
2000	2000	2000	2000	2000	2000	2000	2000	2000
1999	1999	1999	1999	1999	1999	1999	1999	1999
1998	1998	1998	1998	1998	1998	1998	1998	1998
1997	1997	1997	1997	1997	1997	1997	1997	1997
1996	1996	1996	1996	1996	1996	1996	1996	1996
1995	1995	1995	1995	1995	1995	1995	1995	1995
1994	1994	1994	1994	1994	1994	1994	1994	1994

Case Study: CEO Pay

How open are the data?

- Free of charge? **YES**
- Without license restrictions? **YES**
- Electronic and machine-readable? **YES**
- Available to download in bulk? **YES**

Conclusions

- In summary, data journalism is challenging
- Collecting, cleaning, and analysing data using social science research methods, and then visualising the results is difficult
- Having access to open data makes things MUCH easier
- Open data are a data journalist’s best friend

ESPON EUROPEAN UNION

Co-financed by the European Regional Development Fund

Inspire Policy Making with Territorial Evidence

// Thank you

Doug Dowson, The Economist

@doug_dowson

This presentation will be made available at: www.espon.eu/open-data-training