



The Future of Statistics

London, September, 2013 · ulrich atz · @statshero



A classic example

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

- Linda is a banker.
- Linda is a banker and is active in the feminist movement.

Tversky, A. and Kahneman, D. (1982) "Judgments of and by representativeness". In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press



The image features a solid blue background. In the center, the letters 'ODI' are rendered in a large, white, stylized font. The 'O' is a simple circle, the 'D' is a rounded rectangle, and the 'I' is a vertical rectangle. The text 'What is the ODI?' is centered within the white space of the 'D' in the 'ODI' graphic.

What is the ODI?

**ODI is a catalyst for
unlocking value**



ODI

10 Startups

MASTODON C



BIG DATA DONE BETTER

opencorporates



CARBON CULTURE.
Join in, team up, make a difference

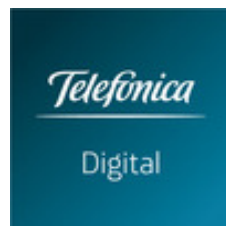


PROVENANCE

Spend Network



36 Members



<http://certificates.theODI.org>

The first **robust quality badge** for open data

- Helps publishers certify their data
- Helps users find and use it
- Helps policy makers benchmark



<http://certificates.theODI.org>



Expert

An exceptional example of information infrastructure.



Standard

Regularly published open data with robust support that people can rely on.



Pilot

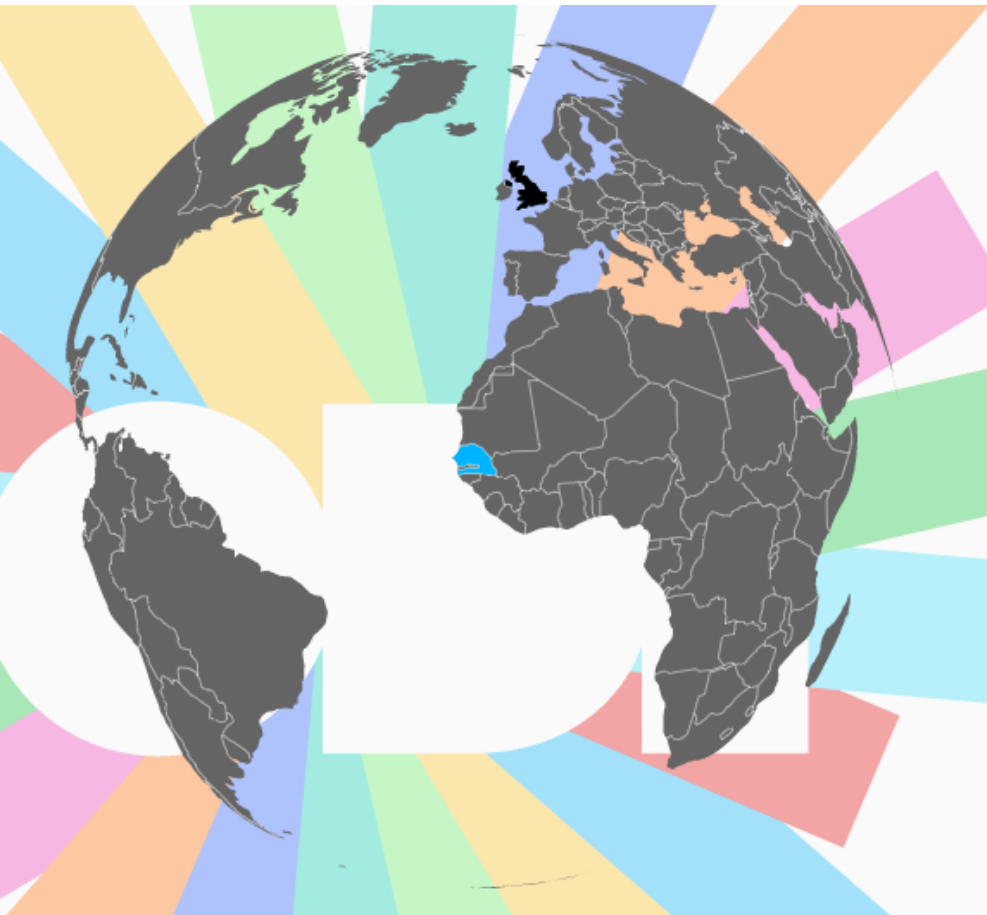
Data users receive extra support from, and can provide feedback to, the publisher.



Raw

A great start at the basics of publishing open data.





Where can you create certificates?

Different countries have different laws, so their certificates are slightly different too. Choose your country to create the right certificate.

Create a certificate

SN alpha

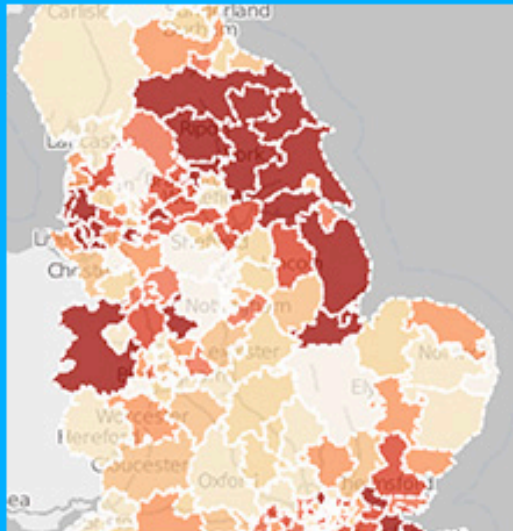
This certificate hasn't been properly localised yet, and we need your help! Please contact certificate@theodi.org

Legend

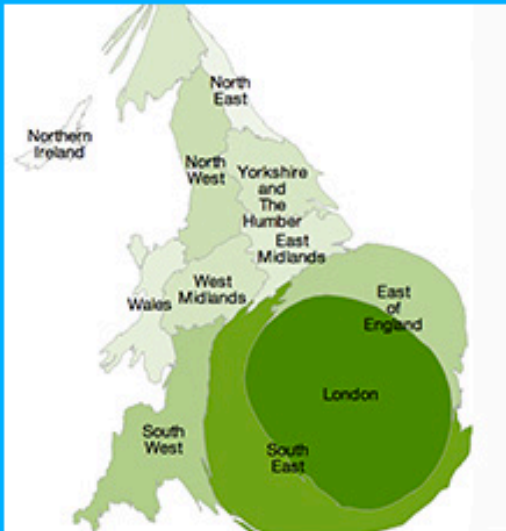
- Certificate in alpha stage
- Certificate in beta stage
- Certificate in final stage



Open Data Stories



**Prescribing
Analytics**



**Show me
the money**



**Open
Corporates**

A collaboration with the World Bank



THE WORLD BANK
Working for a World Free of Poverty



News

This page in: [English](#)

PRESS RELEASE

New Partnership Seeks to Bring Benefits of Open Data to Developing Countries



WHAT IS DATA?



Data is the raw material of the
information age.



Data is the New Oil



kenhodge13 (40132991@N07) on flickr.com



But data is not oil!

- data is a non-rival good
- marginal cost of distribution
- falling cost of analysis (“refinement”)



More like gold than oil?




<http://www.flickr.com/photos/48806909@N00/497511638/>



A new gold rush?

A NEW AND MAGNIFICENT CLIPPER FOR SAN FRANCISCO.
MERCHANTS' EXPRESS LINE OF CLIPPER SHIPS!
Loading none but First-Class Vessels and Regularly Dispatching the greatest number.
THE SELENDID NEW OUT-AND-OUT CLIPPER SHIP



CALIFORNIA
HENRY BARREN, Commander, AT PIER 13 EAST RIVER.

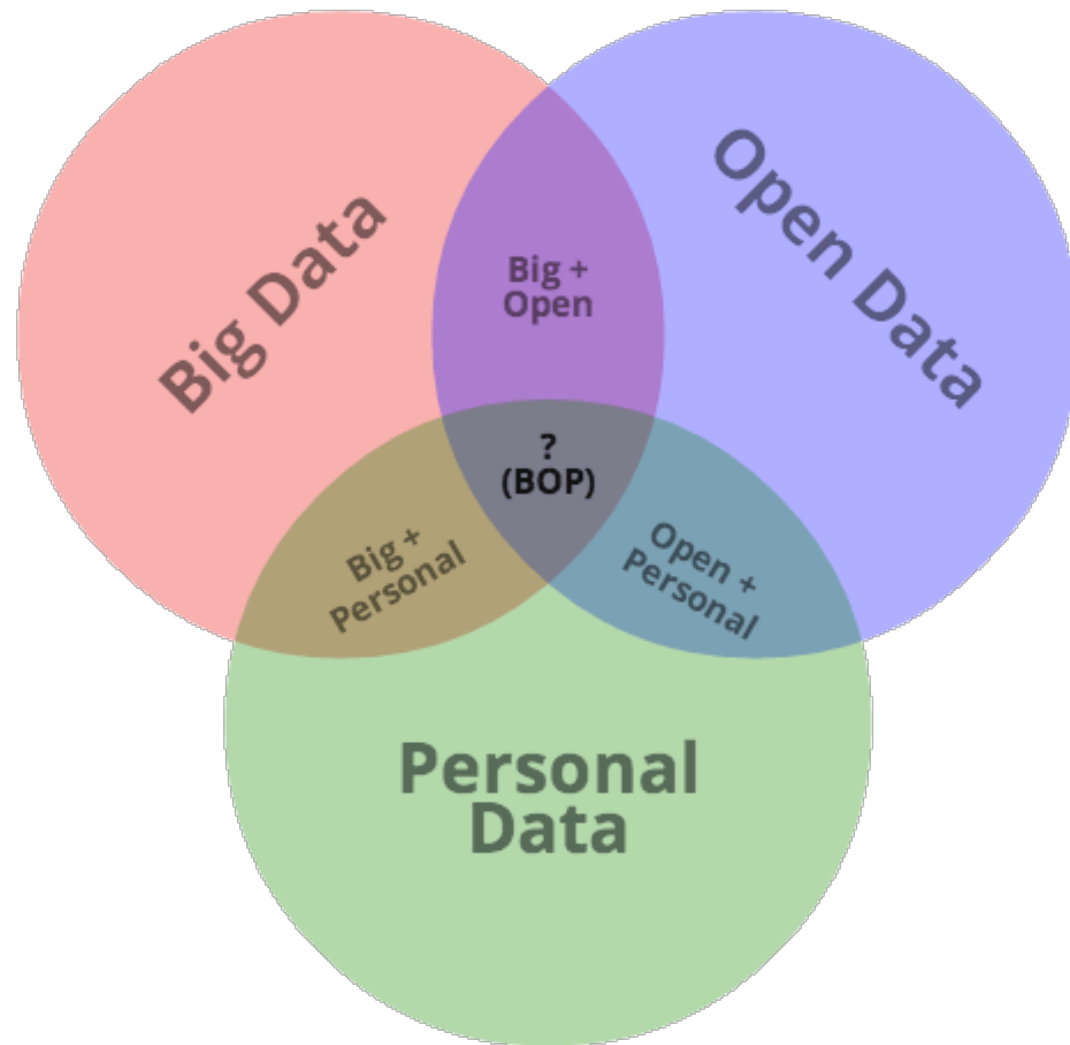
This elegant Clipper Ship was built expressly for this trade by Samuel Hall, Esq., of East Boston, the builder of the celebrated Clippers "Swiftness," "Hambrook," "John Gilroy," and others. She will fully equal them in speed! Unusually prompt dispatch and a very quick stop may be relied upon. Engagements should be completed at once.

Agents in New York,
ROBERT W. BENTLEY & CO.

RANDOLPH M. COOLY, 88 Wall Street, Tenth Building.

WALTON & CO., CHICAGO





Open data supports
collaborative ecosystems



THE OFFICIAL HOME OF REVISED
ENACTED UK LEGISLATION
1267-PRESENT
CHANGES OVER TIME

Search All Legislation

Title:

Year: Number:

Type:

[Advanced Search >](#)

New Legislation

-  [The Town and Country Planning \(Section 62A Applications\) \(Written Representations and Miscellaneous Provisions\) Regulations 2013 >](#)
-  [The Town and Country Planning \(Section 62A Applications\) \(Hearings\) Rules 2013 >](#)
-  [The Town and Country Planning \(Section 62A Applications\) \(Procedure and Consequential Amendments\) Order 2013 >](#)
-  [The Planning \(Listed Buildings and Conservation Areas\) \(Amendment No. 2\) \(England\) Regulations 2013 >](#)
-  [The Town and Country Planning \(Appeals\) \(Written Representations Procedure and Advertisements\) \(England\) \(Amendment\) Regulations 2013 >](#)

Frequently Asked Questions

- [What legislation is held on legislation.gov.uk? >](#)
- [Will I find new legislation on legislation.gov.uk? >](#)
- [What legislation is available as revised? >](#)
- [How up to date is the revised content on this website? >](#)

[View more >](#)

Most requested Acts

- [Data Protection Act 1998 >](#)
- [Disability Discrimination Act 1995 >](#)
- [Consumer Credit Act 1974 >](#)
- [Health and Safety at work etc. 1974 >](#)
- [Children Act 2004 >](#)
- [Employment Rights Act 1996 >](#)
- [Environmental Protection Act 1990 >](#)

© Crown copyright

You may use and re-use the information featured on this website (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence



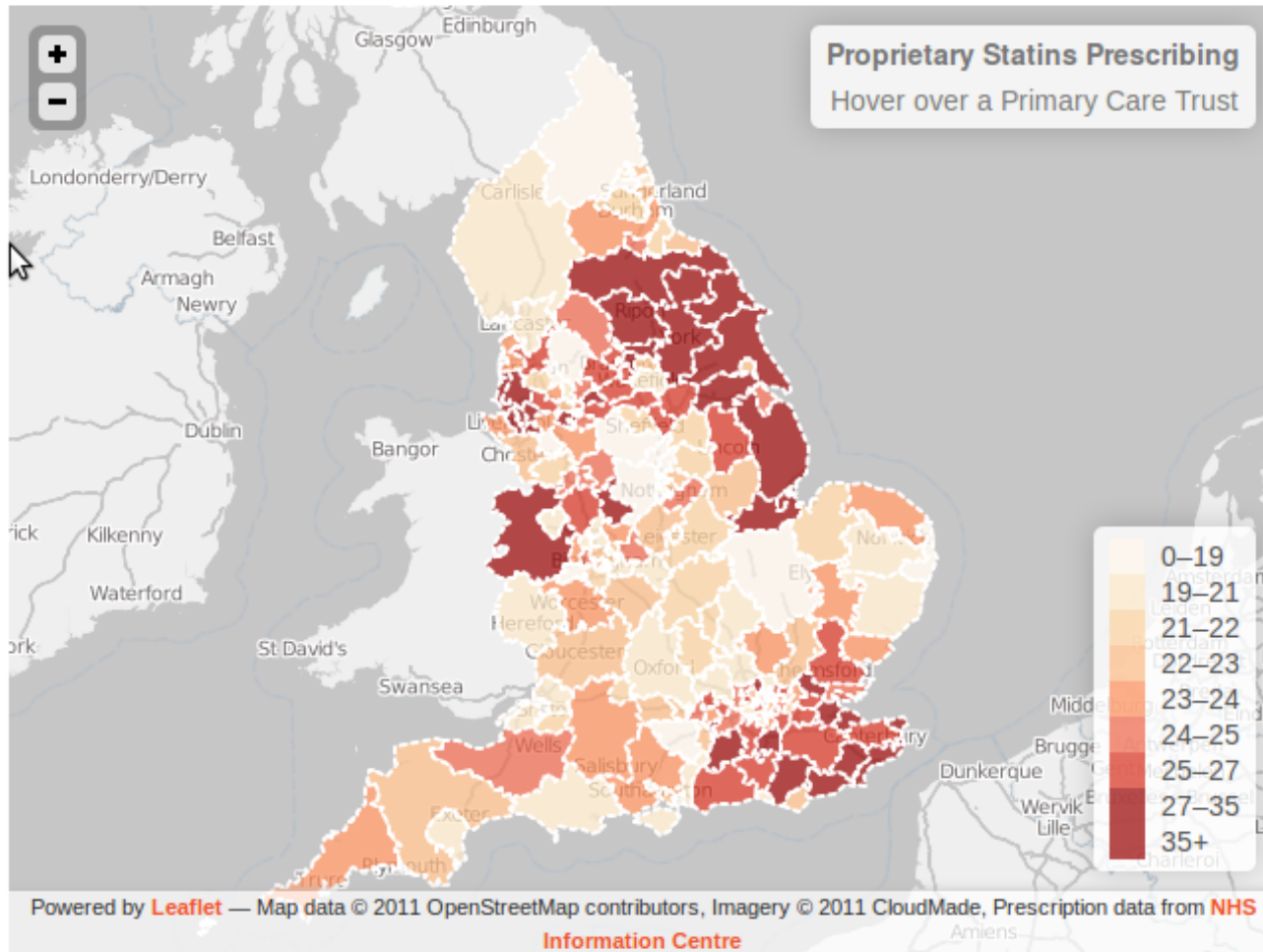
WHAT IS DATA SCIENCE?



Open Data found a £200m saving in the NHS budget



Percentage of proprietary statin prescribing by CCG Sep 2011 - May 2012



[Show PCT data](#)



What is data science?

“often defined in terms of attributes of the data scientist”

A naming problem:

- the science of data?
- science with data?

More: <http://theodi.github.io/presentations/2013-09-oxford-real-data-scientist.html>



What is Statistics?

- The science of learning from (or making sense out of) data
- The theory and methods of extracting information from observational data for solving real-world problems
- The science of uncertainty
- The quintessential interdisciplinary science
- The art of telling a story with [numerical] data



What is Statistics?

1. Understanding uncertainty
2. Making data useful (“analysing”)



Stereotypes

Statistics	Data Science
Sampling	Use all the data
Slow and steady	Fast and dirty
“It depends”	The fallacy of certainty
B-W bar charts	Infographics
Institutions	Startups



Stereotypes

Statistics	Data Science
Causal	Predictive
Small	Big Large
Report	Data product
Asking	Doing
?	HBR Sexiest job of the 21 st century



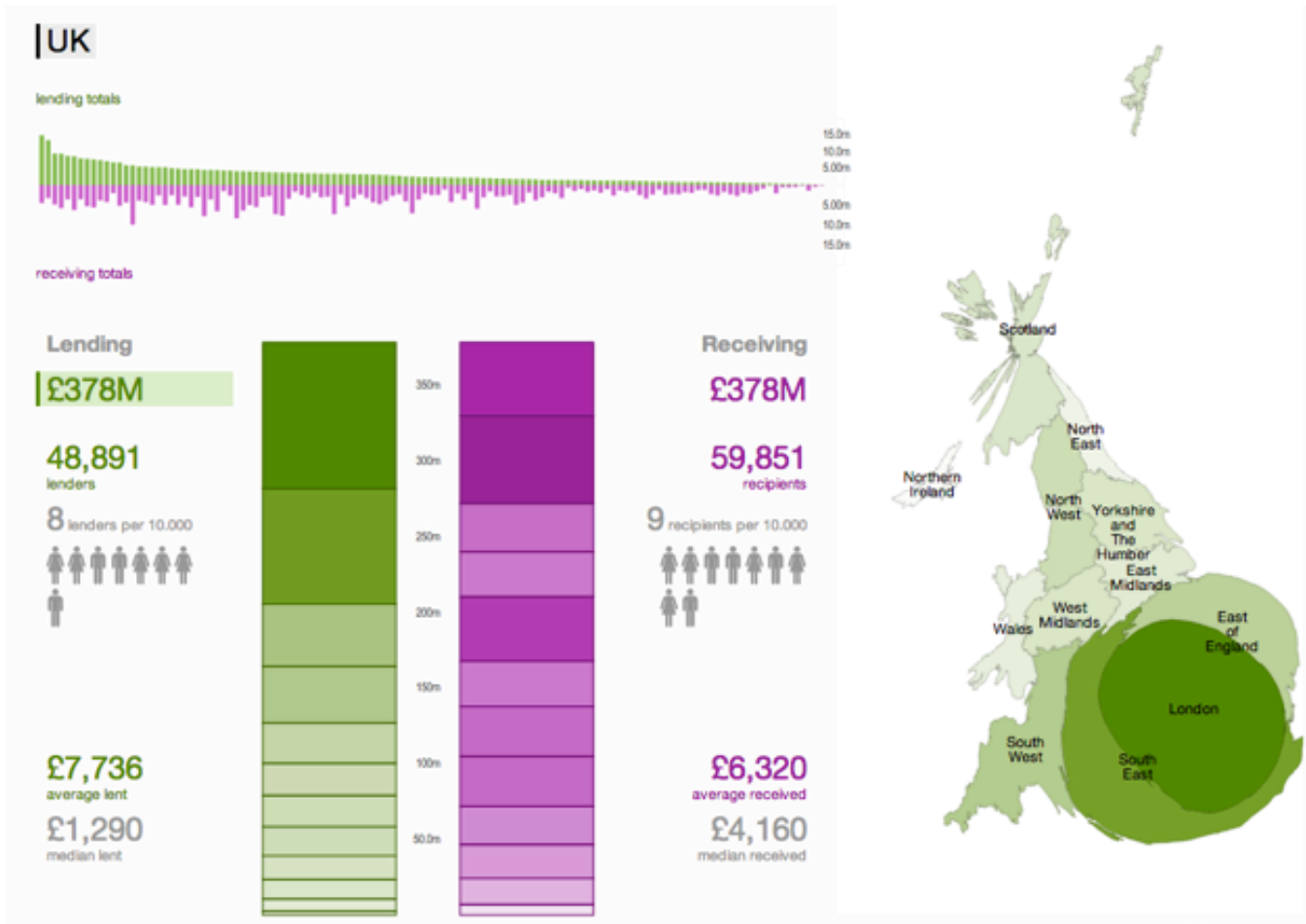
“I’m looking for a data scientist...”
(intentions are good)



<http://www.flickr.com/photos/waitingfortheword/5546445871/>



Mapping £378m of peer-to-peer lending across the UK



Helped convene domain-experts

- + *p2p lenders*
- + *banking professionals*
- + *data analytics (ODI)*
- + *communications (ODI)*

Analysed 14m records

- + *all the data*
- + *anonymised and analysed*
- + *ODI research*

National & international reach

- + *Front-page Financial Times*

Long-term

- + *Create real-time view*
- + *Stimulate market*



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

TRENDS



Where are we now?

Data, data everywhere
5 trends

*“Predictions are difficult; especially
about the future.”*



1. Commoditisation of data

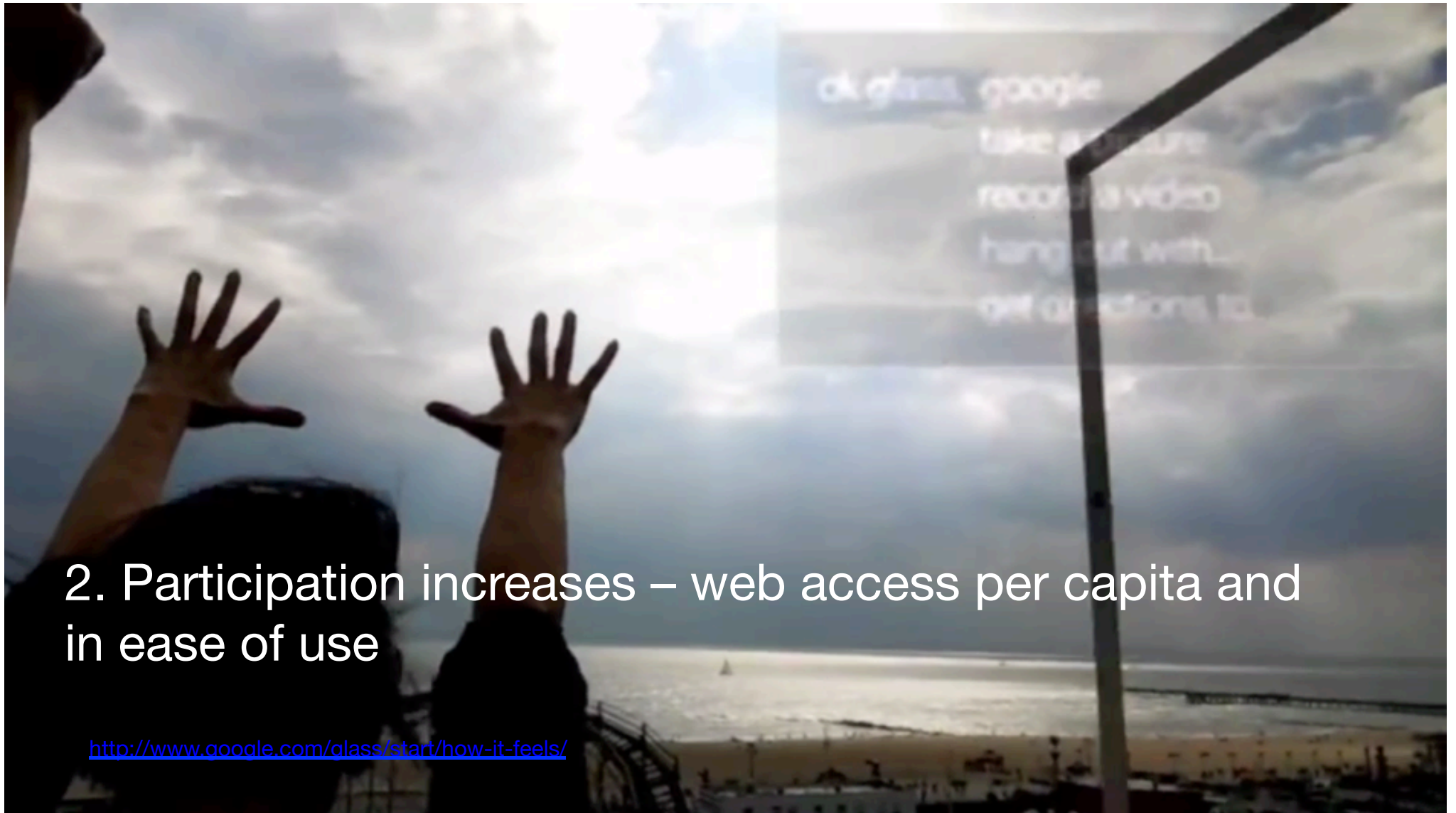
The cost of working with data has lowered and is continuing to fall



The screenshot shows the Amazon S3 product page. At the top left is the Amazon Web Services logo. To the right are links for 'Sign Up', 'My Account / Console', and 'English'. Below the logo is a navigation bar with 'AWS Products & Solutions', a search bar, and 'Developers' and 'Support' links. The main content area is titled 'Amazon Simple Storage Service (Amazon S3)'. It includes a sidebar with 'Amazon S3 Overview', 'FAQs', 'Pricing', and 'Amazon S3 SLA'. The main text describes S3 as storage for the Internet, designed for web-scale computing. A 'Get Started with AWS for Free' section offers a 'Create Free Account' button and lists the AWS Free Tier benefits: 5GB storage, 20,000 Get Requests, and 2,000 Put Requests. Below this is a 'View AWS Free Tier Details' link. At the bottom, a section titled 'This page contains the following categories of information. Click to jump down:' lists various topics like 'Amazon S3 Functionality', 'Protecting Your Data', 'Managing Your Data', 'Pricing', 'Getting Started with Amazon S3', 'Transferring Large Amounts of Data', 'Common Use Cases', 'Resources', 'Amazon S3 Design Requirements', and 'Intended Usage and Restrictions'.

<http://aws.amazon.com/s3/>





2. Participation increases – web access per capita and in ease of use

<http://www.google.com/glass/start/how-it-feels/>

Tools are getting easier to use – without the need for coding skills



[Login](#) [About Us](#) [Features](#) [Pricing](#) [Users](#) [Support](#)



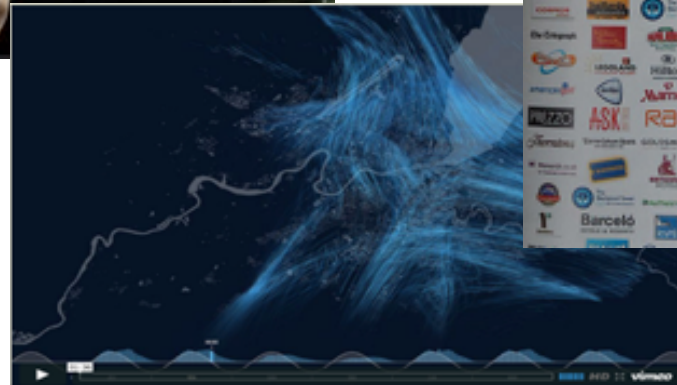
Hello, **data**

Data for everyone

 Download



3. Pervasive data collection



Sensors and Internet of Things
Mobile technologies
Shopping and behaviour monitoring



<http://www.londoncityairport.com/News/ReadArticle/93/>
<http://www.villevivante.ch/#animation>

4. Quantifying of personal data



\$1,664,375

Raised of \$100,000 Goal

0 time left

Flexible Funding
This campaign has ended and will receive all funds raised. Funding duration: May 22, 2013 - July 20, 2013 (11:59pm PT).

<http://quantifiedself.com/2013/01/future-normal-quantified-self-tools-at-the-apple-store/>



5. Demands for greater transparency



Publish WhatYouFund
The Global Campaign for Aid Transparency

Index Resources Updates

The Issue About Us Contact Us

Aid is a precious resource
To get the most out of it we need more and better aid information. Working with organisations from around the world, we call on donors to publish what they fund.

See how we're doing

MAKE AID TRANSPARENT
Make aid transparent - sign the petition

How do you rank?
Aid Transparency Index 2012
72 donors in 2012 - Browse Aid Transparency Index

Publish To IATI Standard
Find out more about IATI and how to publish

Towards Climate Finance Transparency
View study on aid transparency & climate

Major Donors Latest News Latest Tweets

WHERE DOES MY MONEY GO?

Showing you where your taxes get spent

The Daily Bread Country & Regional Analysis Departmental Spending About

How is your tax money spent?

The Daily Bread



See how your daily taxes are divided between the different parts of government.



How much is spent on the various functions of government in total — and where?

Country Regional Analysis

Where Does My Money Go? is part of **OpenSpending**, where you can find information about government finance from countries across the world.

OpenSpending

<http://www.flickr.com/photos/kcorrick/7609944890/> <http://www.flickr.com/photos/brostad/5841297115>



ANALYSING THE ANALYSERS

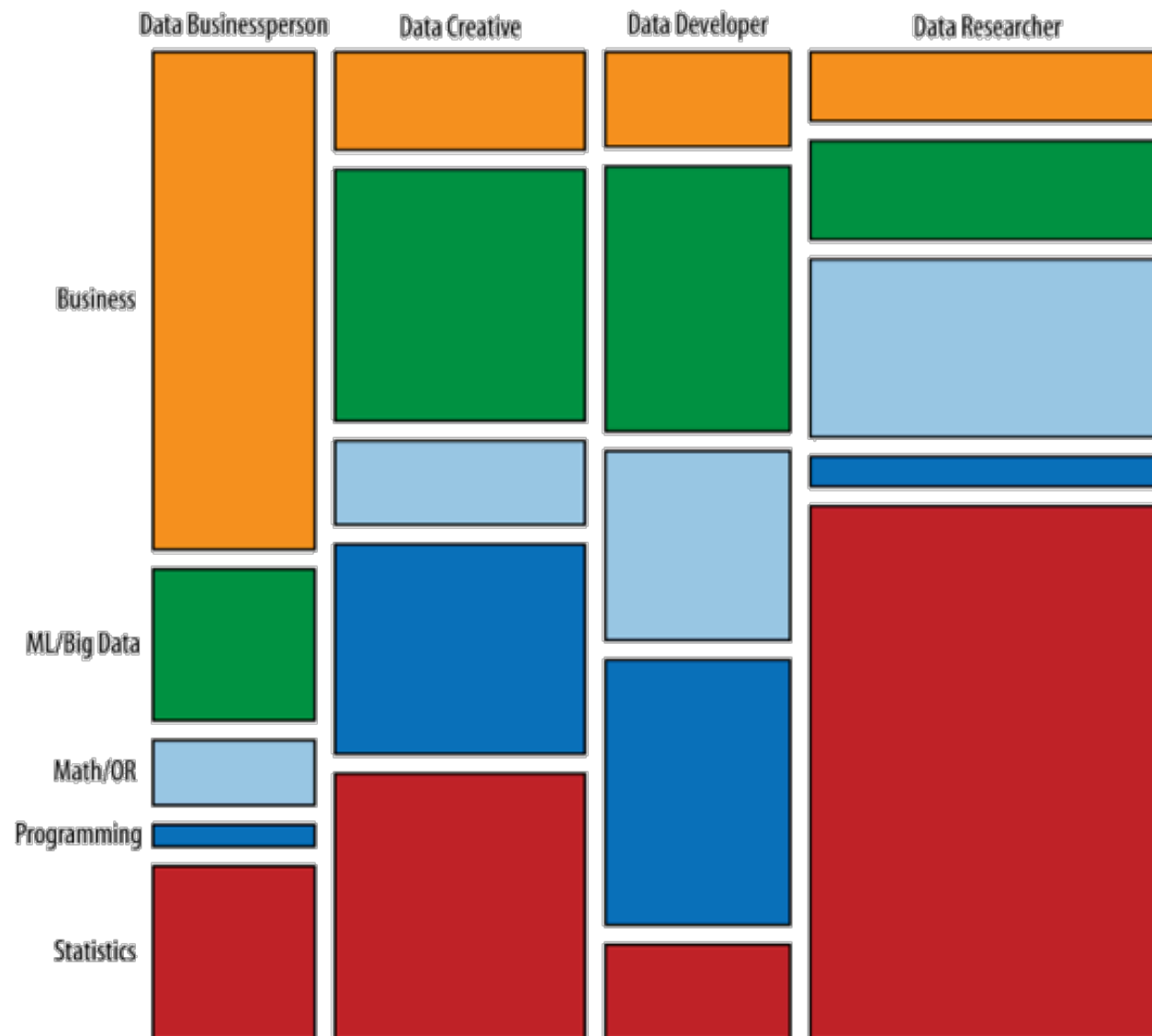


Analyzing the Analyzers – a recent study in the data science community

Harlan D. Harris, Sean Patrick Murphy, and Marck Vaisman

A skills-sorting task





Again we see

Data Science defined by a set of skill.



THE FUTURE OF STATISTICS



A look back

19th century:
large data
sets, simple
questions

21st century:
large data
sets, complex
questions

20th century:
small data
sets, complex
questions



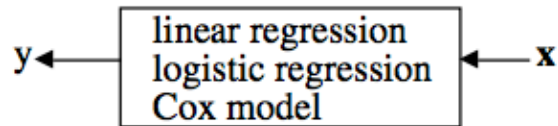
Statistical Modeling: The Two Cultures

Leo Breiman:

“There are two cultures in the use of statistical modeling to reach conclusions from data.”

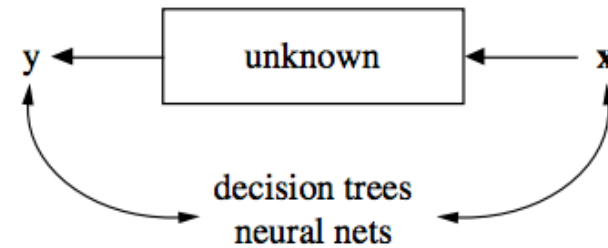


The *Two Cultures*



Model validation. Yes–no using goodness-of-fit tests and residual examination.

Estimated culture population. 98% of all statisticians.



Model validation. Measured by predictive accuracy.

Estimated culture population. 2% of statisticians, many in other fields.

(Tukey spoke of algorithmic and algebraic models.)



Convergence



Why use R?

Welcome to RStudio

Software, education, and services for
the R community



Open source / free

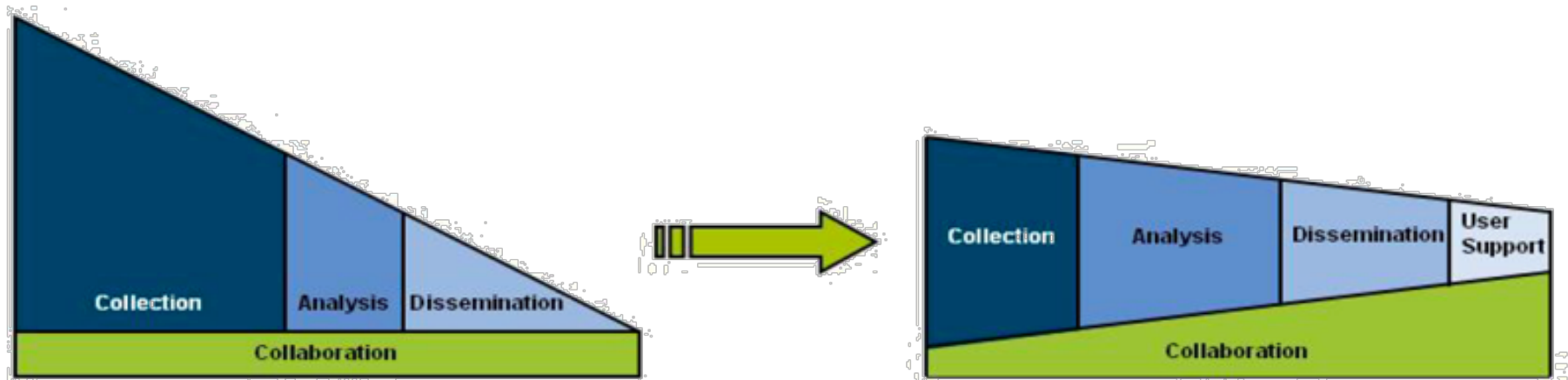
Widely used (> 2 million and growing)

R has an incredible community (> 4,000 packages)

Used by Facebook, Google (apparently 500+ users),
weather forecasts, finance industry



UK GSS 2020 strategy: “Collaboration will become the norm”

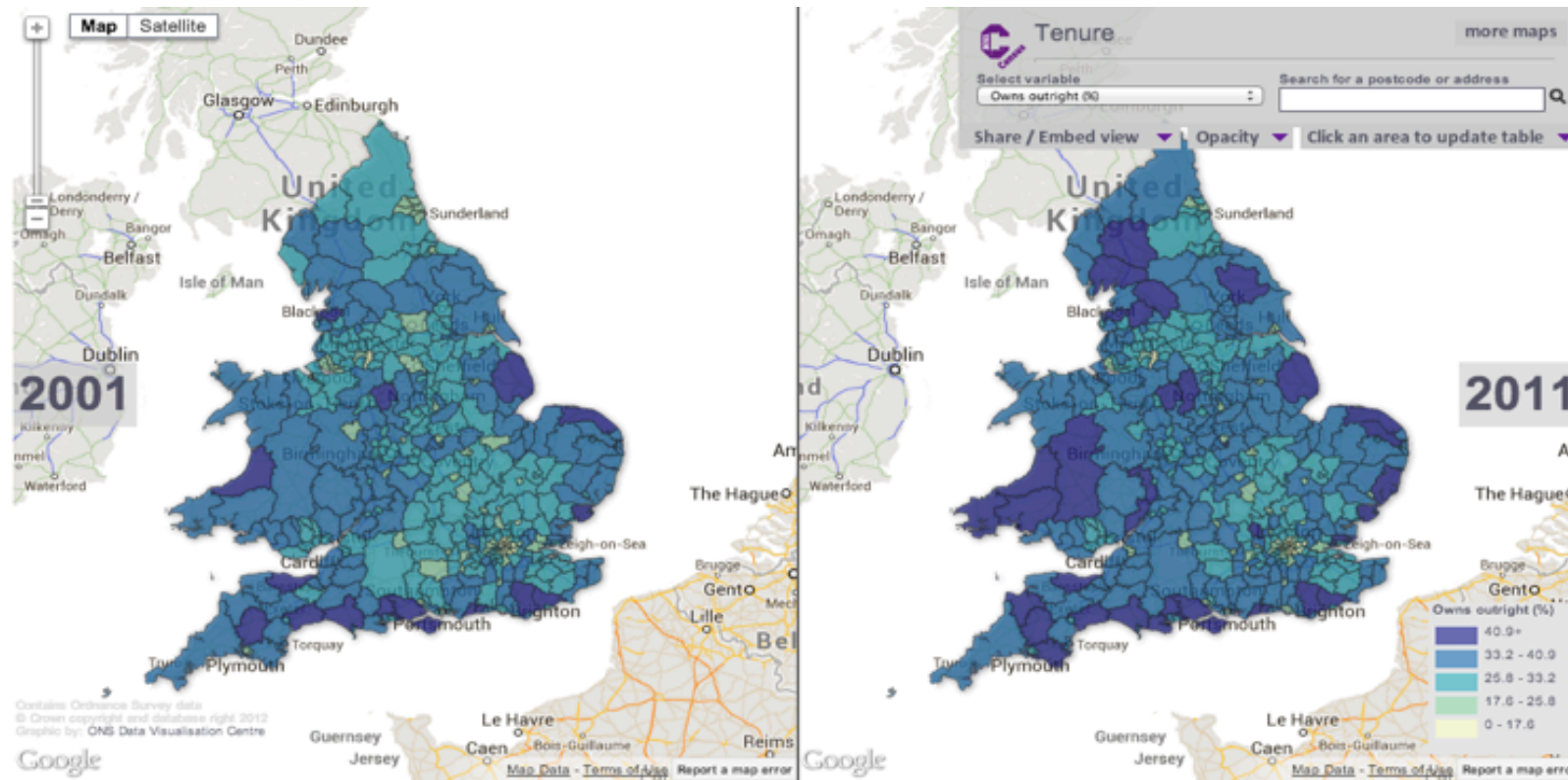




There are rock stars, and then there are rock bands: statisticians frequently work in teams.

Marie Davidian, <http://www.independent.co.uk/news/world/americas/heroes-of-zeroes-nate-silver-his-rivals-and-the-big-electoral-data-revolution-8734380.html>
<http://www.flickr.com/photos/11055761@N04/1842696307/>

Mapping the Census

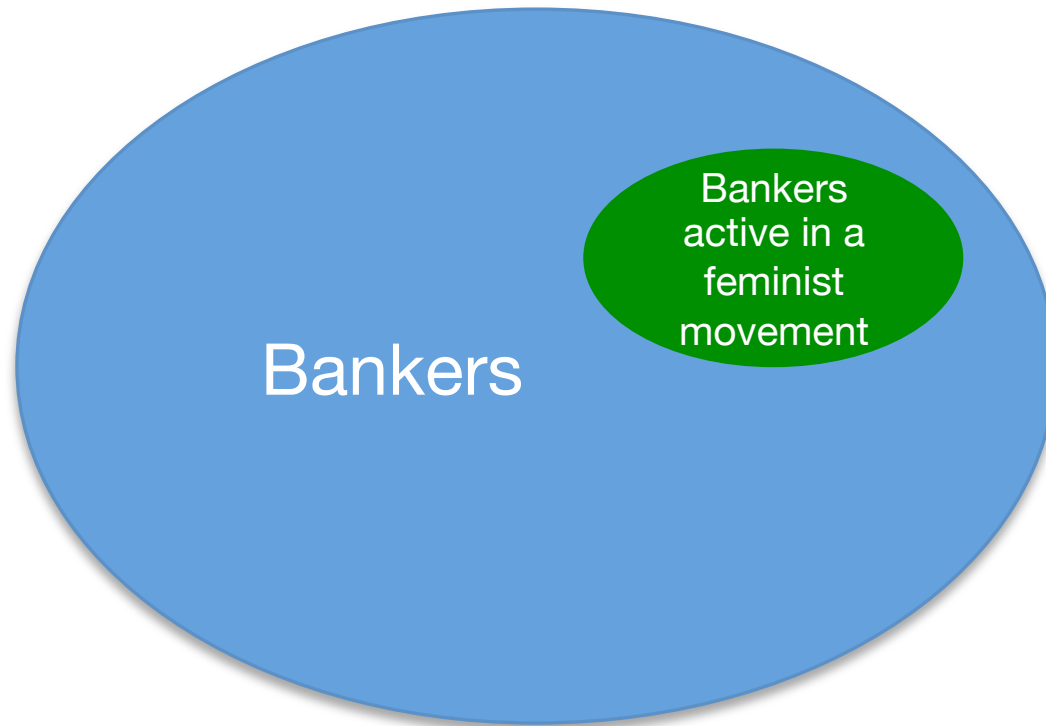


<http://www.ons.gov.uk/ons/interactive/census-map-2-1---tenure/index.html>

<http://bit.ly/1232hJb>



Remember Linda, 31, bright, activist...?



We learn

Even statisticians are bad
intuitive statisticians.



Statistics is about
uncertainty

More biases: *We seek overconfidence.*

But understanding uncertainty should
not be our only domain – combine it
with data analysis and data-driven
decisions.



The danger: Define our limits in terms of familiar tools and familiar problems.



A graphic featuring the word "OPEN" in a large, white, stylized font against a solid blue background. The letter "O" is a circle, "P" is a rounded rectangle, "E" is a tall rectangle, and "N" is a tall, narrow rectangle. The phrase "BE OPEN" is written in a smaller, blue, sans-serif font across the middle of the "O".

BE OPEN

Thank you!



Email: ulrich@theodi.org · Twitter: [@statshero](https://twitter.com/statshero)



GOLDMAN SACHS (INDIA) SECURITIES PRIVATE LIMITED

One of 27 subsidiaries registered in India

CONTROL CHAIN: GOLDMAN SACHS GROUP, INC., THE > GS INDIA HOLDINGS L.P. > GOLDMAN SACHS INVESTMENTS (MAURITIUS) I LIMITED > GOLDMAN SACHS (MAURITIUS) LLC. > GOLDMAN SACHS (INDIA) SECURITIES PRIVATE LIMITED

