O'REILLY®
Strata
CONFERENCE
Making Data Work

11–13 Nov 2013
LONDON, ENGLAND

#strataconf
strataconf.com/london

DataKind UK
USING DATA IN THE SERVICE OF HUMANITY

USING DATA FOR EVIL

Fran Bennett, CEO Mastodon C
@fhr

Duncan Ross, Director Data Science, Teradata
@duncan3ross

Our hypothesis

- For a data scientist
  - Doing good deliberately is hard
  - Doing evil deliberately is hard
  - Doing evil accidentally is easy

- Hypothesis:
  - Every data scientist has the capability to do good by thinking about what they do

- Null hypothesis:
  - Every data scientist has the capability to do evil by not thinking about what they do

Your quest for Global Data Dominance begins here

- Doing evil doesn't need to be about the big things
  - We all have the ability to make the world a little worse
  - Where would Khan have got to without his minions?
- We will highlight some of the biggest mistakes you can make
- We will give you some excellent hints on how to maximise evil with the least possible effort

O'REILLY®

Strata
CONFERENCE
Making Data Work

11–13 Nov 2013
LONDON, ENGLAND

#strataconf
strataconf.com/london

Who do *you* work for?

Where can you do most evil?

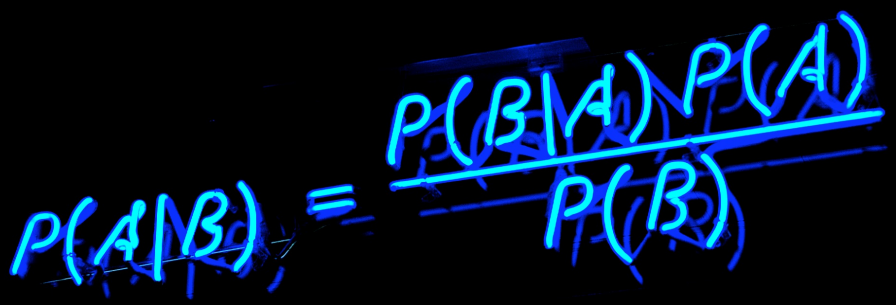
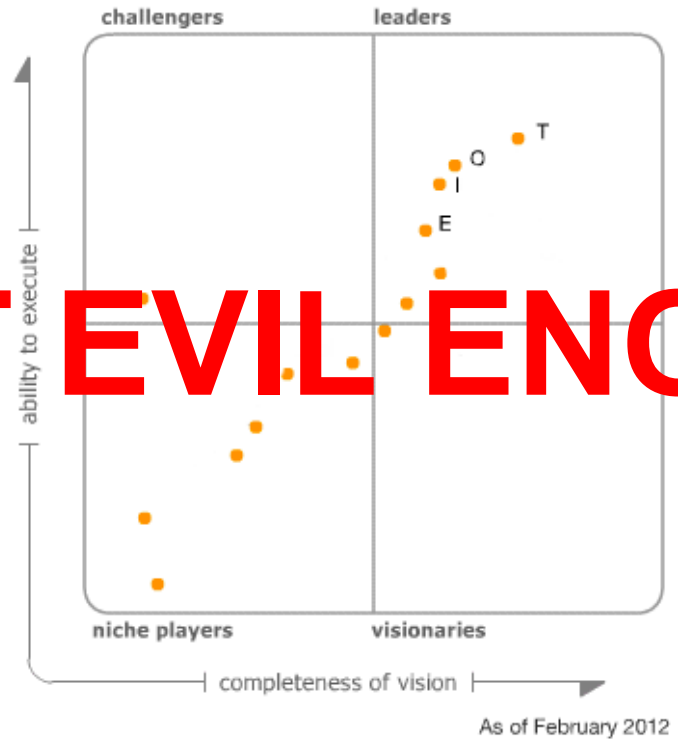


Image: mattbuck

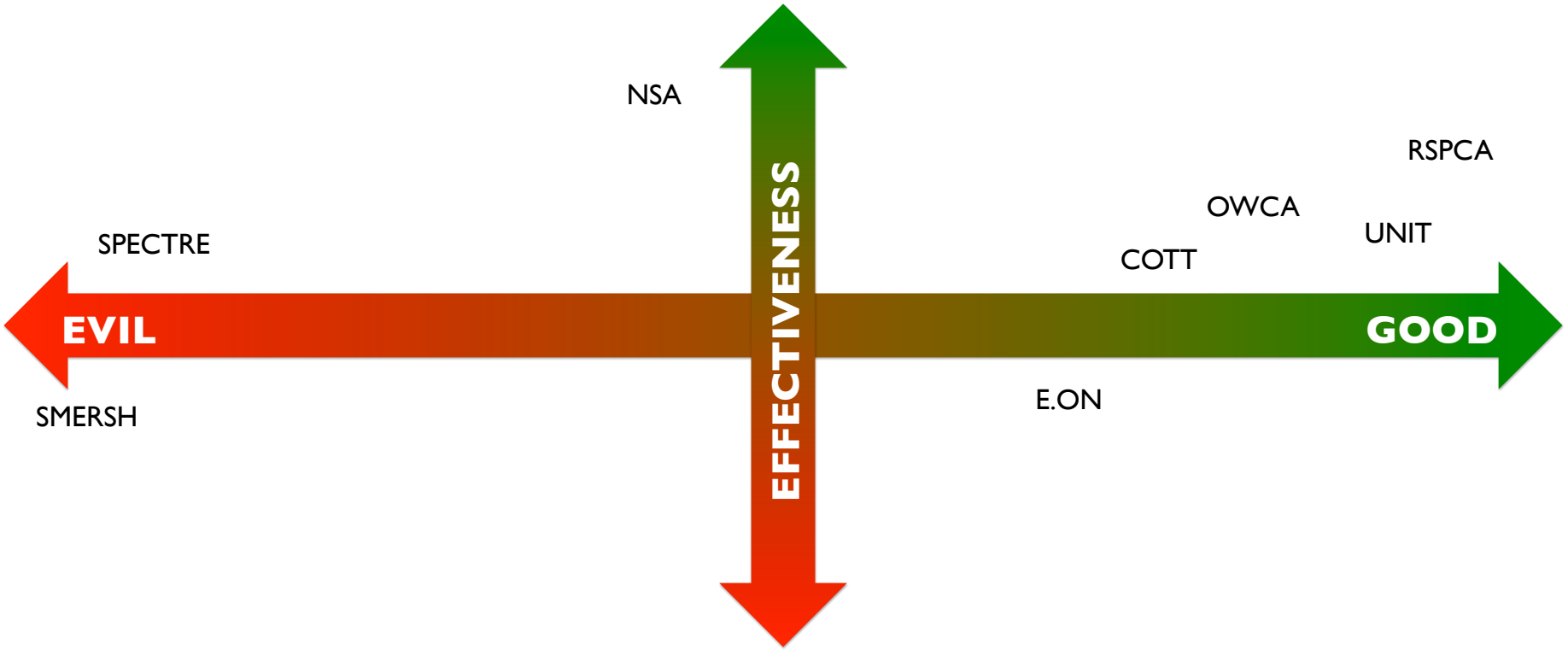
- In an organisation committed to evil?
- In an organisation committed to good?
- In an organisation committed to shareholder value?
- In an organisation that isn't sure?

Spectrum of evil



NOT EVIL ENOUGH

Spectrum of evil



O'REILLY®

Strata
CONFERENCE
Making Data Work

11–13 Nov 2013
LONDON, ENGLAND

#strataconf
strataconf.com/london

The work that you do

Quiz: which of these is most exciting?

① Analysing call ‘meta data’ to find out when people are entering (and leaving) a relationship

② Finding out how different people react to different drugs using medical data?

③ Predicting and changing behaviours through examining purchasing behaviour

Key mistake 1: caring about impact

- Analysis is cool, and powerful
- Your analysis can have really great effect on people
- You can make them do things they wouldn't normally want to
  - Never measure the results
  - Don't worry about what might happen
- UK Border Agency

Home Office
UK Border Agency

How can you avoid caring?

- Make sure that the data scientist is kept away from the business
- Think about numbers, not people
- Don’t look at differential impact
  - Type II error

    Don’t send message to someone who should go
  - Type I error

    Send “get out” message to someone who is allowed to stay

Type I terrorists

P(+ | bad guy) = 0.99

P(bad guy | +) =??

P(bad guy) = 1/1,000,000

P(+ | good guy) = 0.01

Then:

P(bad guy | +) = 1/10,102

http://bayesianbiologist.com/2013/06/06/how-likely-is-the-nsa-prism-program-to-catch-a-terrorist/

O'REILLY®

Strata
CONFERENCE
Making Data Work

11–13 Nov 2013
LONDON, ENGLAND

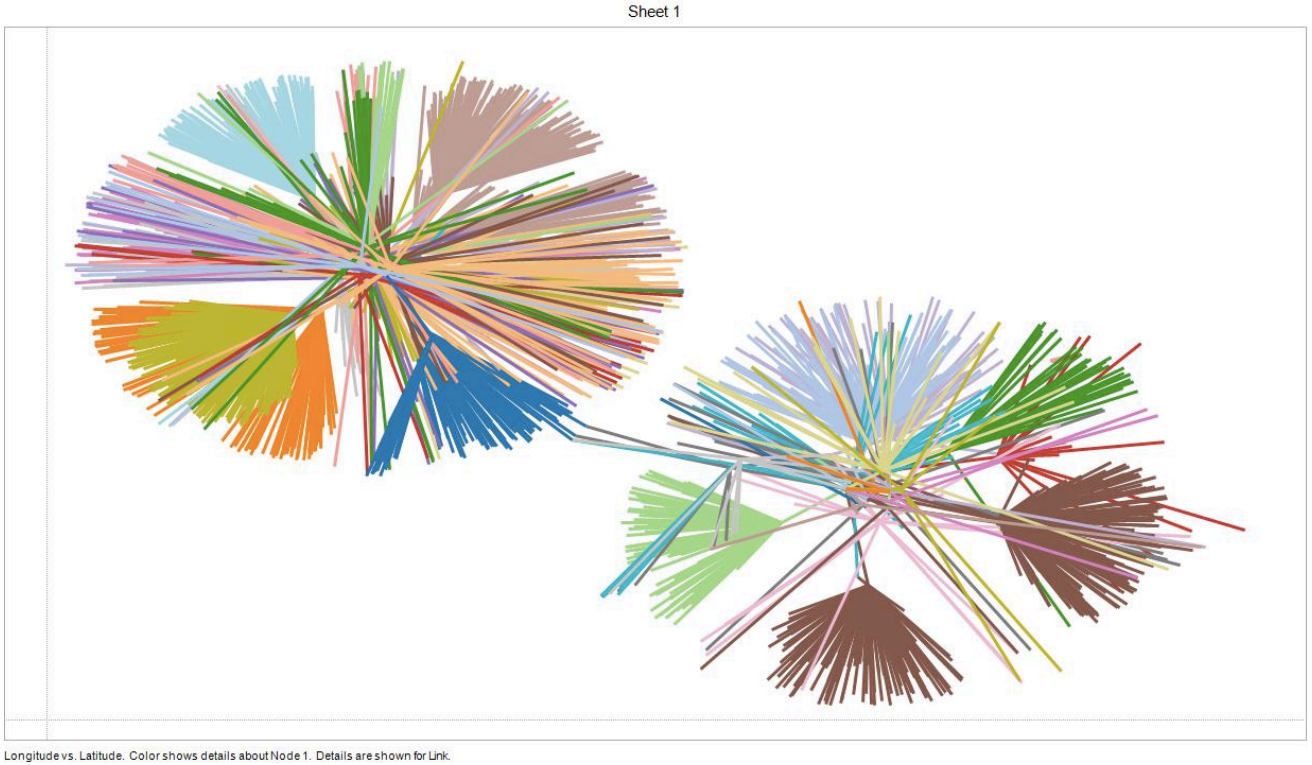
#strataconf
strataconf.com/london

The way you work

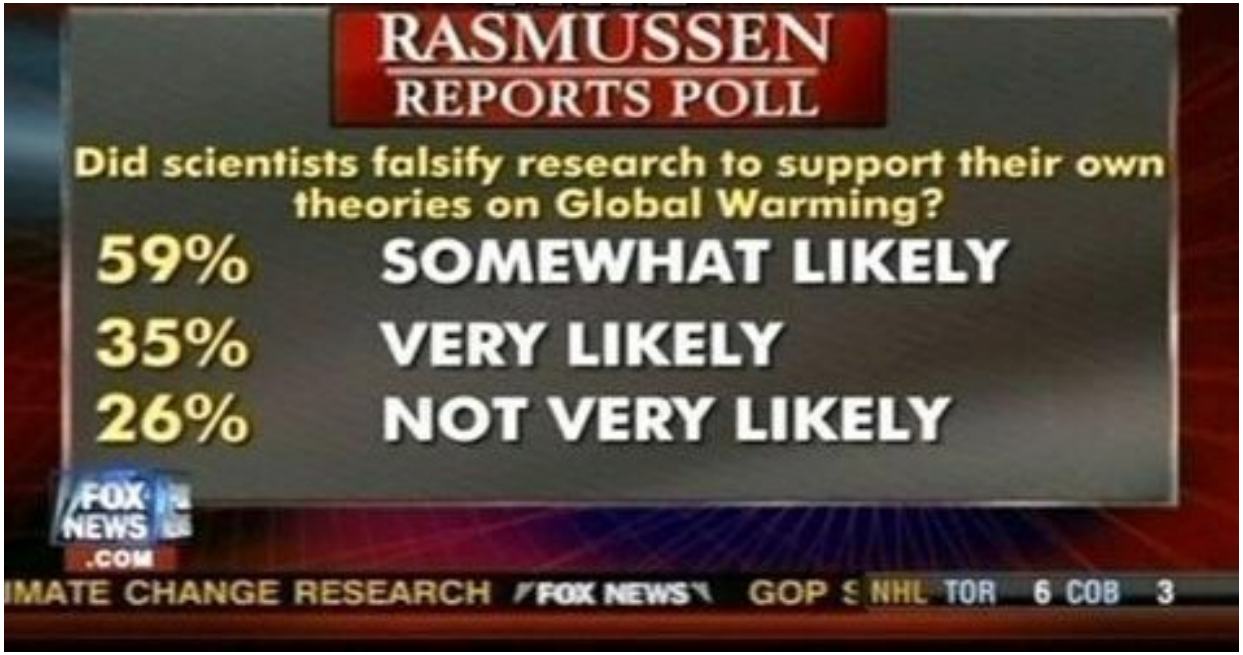
There are simple steps you can take…
… to make your organisation more evil

- Never check your work with anyone else
- What you do is complex, make sure they are fully aware of it
- Remember, the HiPPO is always right

A bad visualisation is worth a thousand swear words



If in doubt… lie



Correlation, causation, blah, blah, blah

http://dailymailoncology.tumblr.com/

O'REILLY®
Strata
CONFERENCE
Making Data Work

#strataconf
strataconf.com/london

Big data problems

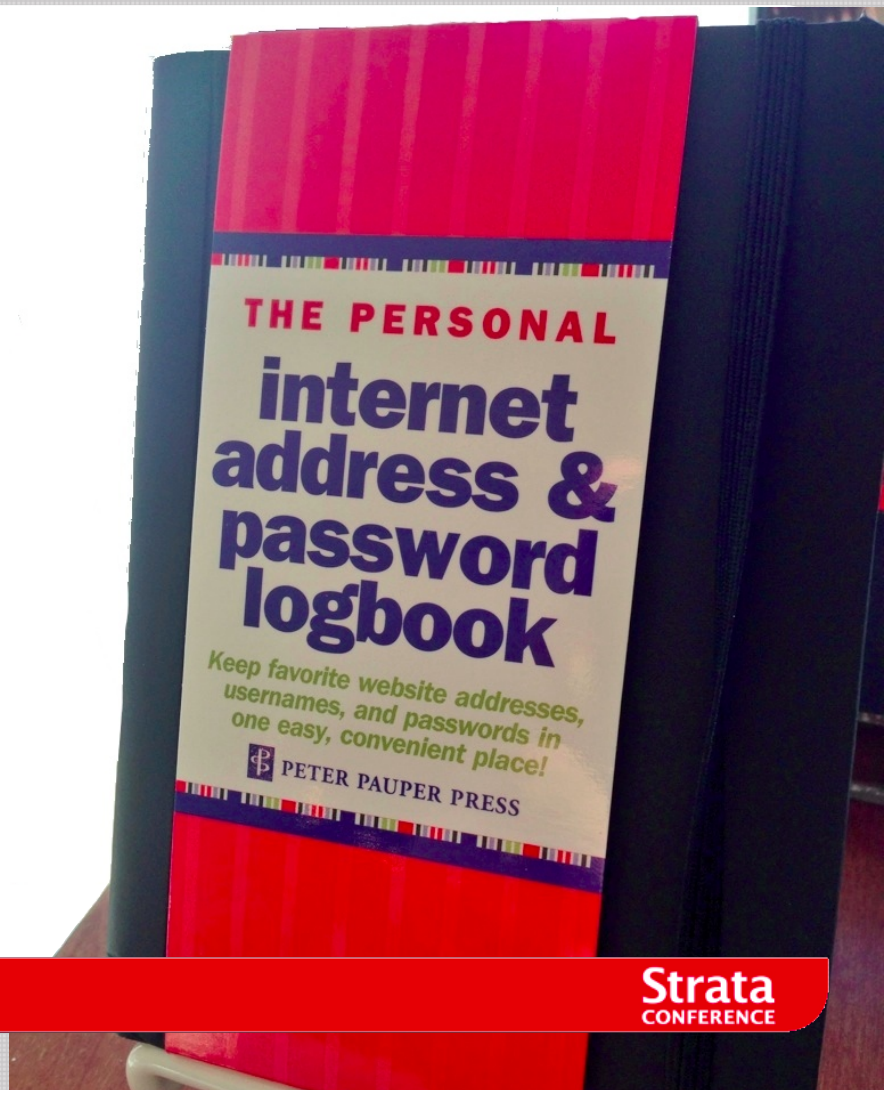
Worse than getting involved in a land war in Asia

If data is so valuable…

- Why don't we see more of these?

If data is so valuable…

- Why do we see this at all?

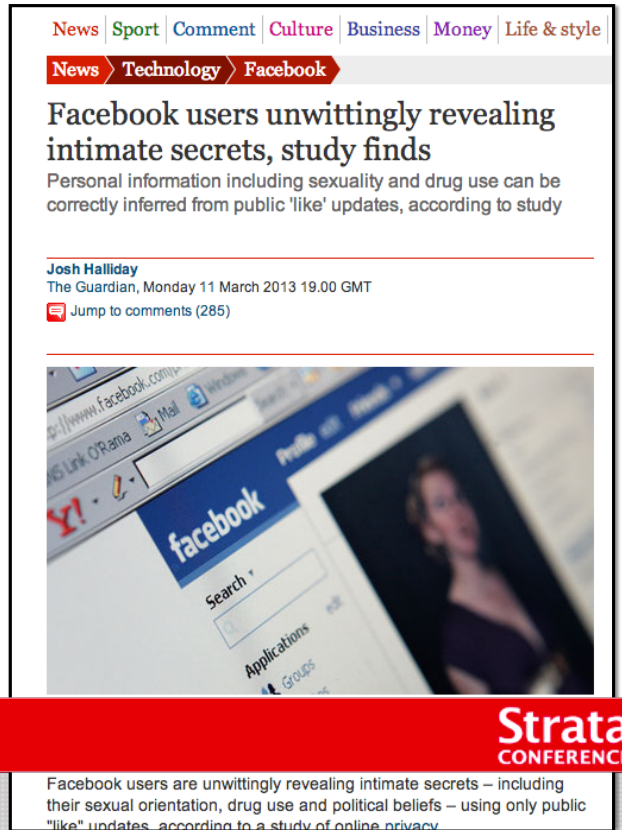


Key mistake 2: anonymization

- There is nothing better than revealing people's dirty secrets
- MIT "Project Gaydar"
- Logistic regression – identifies gay men with 78% accuracy using friend lists

"That pulls the rug out from a whole policy and technology perspective that the point is to give you control over your information – because you don't have control over your information."

Hal Abelson, MIT Prof Comp Sci



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

News › Technology › Facebook

Facebook users unwittingly revealing intimate secrets, study finds

Personal information including sexuality and drug use can be correctly inferred from public 'like' updates, according to study

Josh Halliday
The Guardian, Monday 11 March 2013 19.00 GMT
Jump to comments (285)

Facebook users are unwittingly revealing intimate secrets – including their sexual orientation, drug use and political beliefs – using only public "like" updates, according to a study of online privacy

How can you avoid anonymization?

- Good news! It's really hard to guarantee anonymity
  - The more data you link together the easier it is to work backwards
  - A second rate anonymisation approach is worth its weight in gold

- Example: the Netflix Prize
  - Winning the prize? $1 million
  - De-anonymizing the data using IMDb? Priceless
  - Narayanan and Shmatikov,
    University of Texas

De-anonymisation made easy

- 1.5 million mobile phone users over 15 months
- 4 points of reference (at low resolution) identifies 95% of users
- i.e. if you know time and location at 4x in a year, you can extract that person's info
- Impressively evil!

http://web.mit.edu/newsoffice/2013/de-anonymize-cellphone-data-0327.html

O'REILLY®
Strata
CONFERENCE
Making Data Work

11–13 Nov 2013
LONDON, ENGLAND

#strataconf
strataconf.com/london

The wider world

Key mistake 3: Volcanoes

- It may seem cool to build your evil data centre in a volcano…
- But geothermal energy is clean energy
- 60% of a data centre's CO_2 emission is down to the local grid, not server efficiency
- You will actually be helping to stop global warming
- Studies show that volcanoes are surprisingly vulnerable to attack

Image ©The Simpsons
(as shown by the "Billionaire Tyrant Rupert Murdoch" also © The Simpsons)

How can you avoid volcanoes?

- Come on, this is the easy one…
- Don't think about where the cloud is
  - If senior management ask, just look at them as if they're two years old
  - Distract them with talk about the green credentials of the machines that are being used

"The **** is extremely energy efficient — save up to 75 percent on energy costs compared to x86 alternatives; and the more you grow your workload inside a **** machine, the greater your energy savings… up to 60% more performance at 35% lower price delivering a virtual Linux server for under $1 day"

Reminder: Key mistakes

- Key mistake 1: caring about impact
- Key mistake 2: anonymization
- Key mistake 3: volcanoes

But the opportunities to do good are greater than ever

DataKind UK

USING DATA IN THE SERVICE OF HUMANITY

Thanks

- To villains everywhere, especially
  - Hank Scorpio
  - Dr Doofenshmirtz
  - Alistair Croll*
  - The Daily Mail

- Boo to the good guys
  - Bond, James Bond
  - DataKind UK (http://datakind.org.uk)
  - UN Global Pulse (http://unglobalpulse.org)

*We don't know that Alistair is actually evil, but he has a good name for it