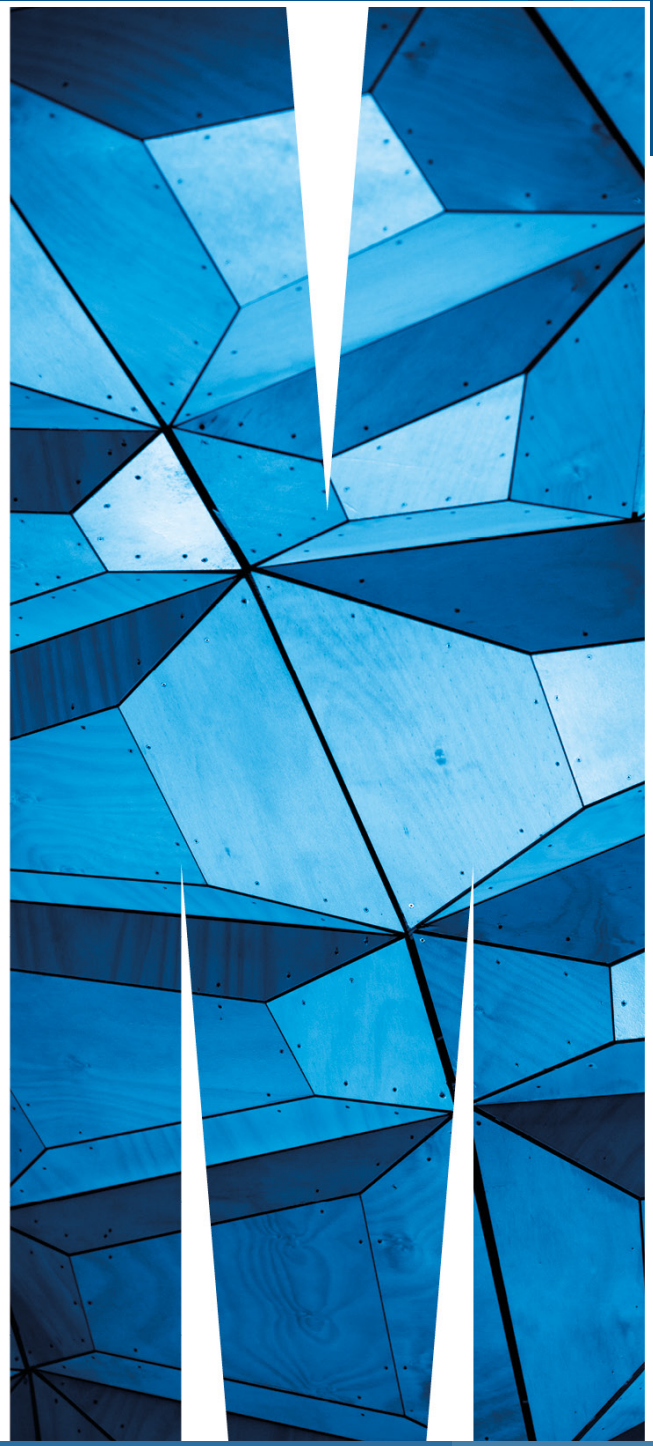
MONASH University

MONASH BUSINESS SCHOOL



Using Data Analytics for Quantitative Citizenship

Professor Di Cook, Econometrics and Business Statistics

Big Data Challenge Day: March 31, 2016

GROUP OF EIGHT

Power to the People

H. G. Wells (1903) Mankind in the Making

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write!"

Open data, open source

- Data is available everywhere today, publicly, free
- Software, very powerful software, for analysis of data is available publicly, free
- Combined with a knowledge of mathematics and statistics empowers each of us to contribute to understand and improve our world
-
- I'm going to show you two projects that I worked on using open data and open software, and how it helped me understand the world a little better.

Math Gender Gap

The Sydney Morning Herald
National

work relax eat sleep

Investigations Interactives Health Education Public Service News World War 1 Centenary Clique Photos Photo Galleries

You are here: Home » National » Education »

Search here... National Search

Numbers point to maths 'gap'

May 2, 2011

Read later

Caroline Milburn

Tweet Share 0 Share 0 Share Pin it submit

Email article Print Reprints & permissions

Australia is one of the few countries with a maths gap in favour of boys.

GIRLS are performing at much lower levels in maths than boys — and the gap is widening as students progress through school.

A study of NAPLAN numeracy test results for students in years 3, 5, 7 and 9, to be presented

"Girls are performing at much lower levels in maths than boys - and the gap is widening as students progress through school."
SMH, 2011

Education: OECD PISA

- OECD PISA survey "the world's global metric for quality, equity and efficiency in school education".
- Workforce readiness of 15-year old students
- 500,000 students were tested across 65 countries and 18,000 schools
- Math, reading and science
- Data available from `http://pisa2012.acer.edu.au`

What do I want to know?

Are boys better than girls at math?

How do we calculate this?

Difference the average math scores!

$$\bar{x}_{boys} - \bar{x}_{girls}$$

Are boys better than girls at math **on average?**

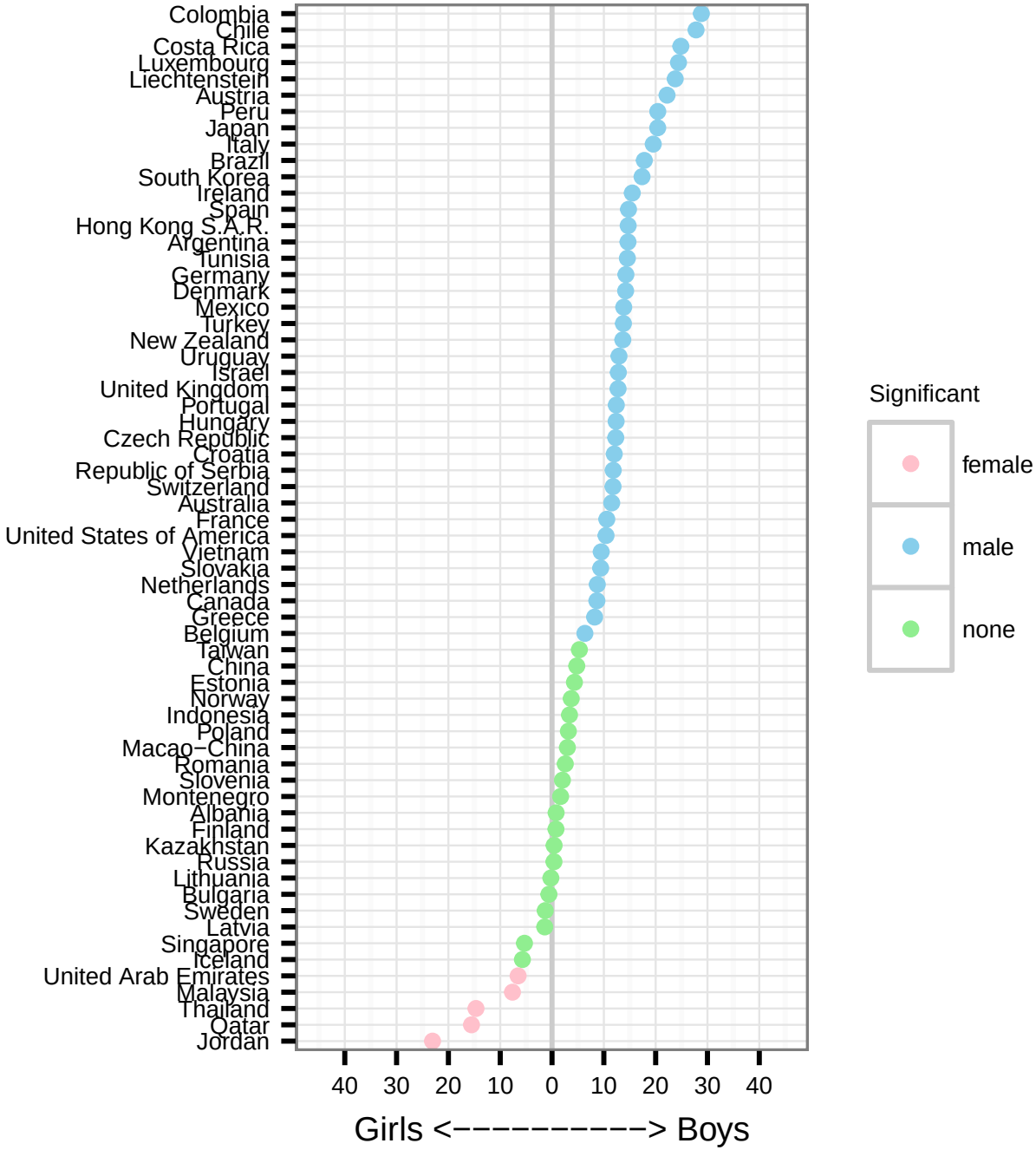
More complicated than that! Some other factors need to be considered

- Countries, need to allow for a difference from one country to another
- Scores have weights associated with them which reflect the demographic difference of students, and these need to be utilised in calculating the mean

A difference between weighted means is calculated for each country. A weighted mean is calculated by

$$\frac{1}{n}\sum_{i=1}^{n} w_i x_i$$

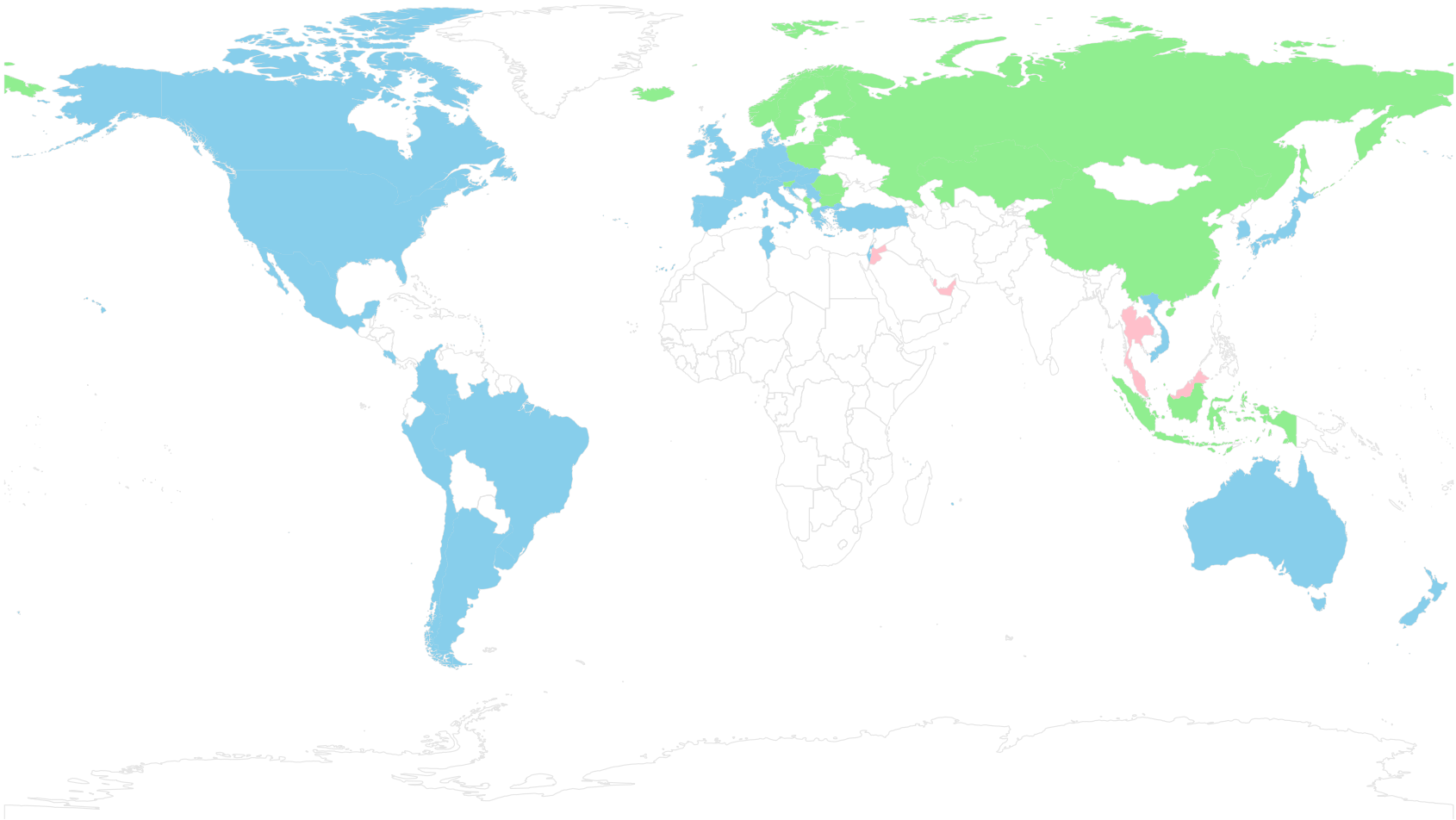
Then plot these by country, ordered from biggest difference to smallest.


Colombia
Chile
Costa Rica
Luxembourg
Liechtenstein
Austria
Peru
Japan
Italy
Brazil
South Korea
Ireland
Spain
Hong Kong S.A.R.
Argentina
Tunisia
Germany
Denmark
Mexico
Turkey
New Zealand
Uruguay
Israel
United Kingdom
Portugal
Hungary
Czech Republic
Croatia
Republic of Serbia
Switzerland
Australia
France
United States of America
Vietnam
Slovakia
Netherlands
Canada
Greece
Belgium
Taiwan
China
Estonia
Norway
Indonesia
Poland
Macao−China
Romania
Slovenia
Montenegro
Albania
Finland
Kazakhstan
Russia
Lithuania
Bulgaria
Sweden
Latvia
Singapore
Iceland
United Arab Emirates
Malaysia
Thailand
Qatar
Jordan
40
30
20
10
0
10
20
30
40
Girls <–––––––––> Boys
Significant
female
male
none

WHAT!

- The gender gap in math is not universal.
- It does exist in Australia
- In many countries there is no substantial gap
- In a few countries the gap is in favor of girls

- Can you see where those countries are?
- Its a bit hard, so lets combine the results with a map, and draw the map



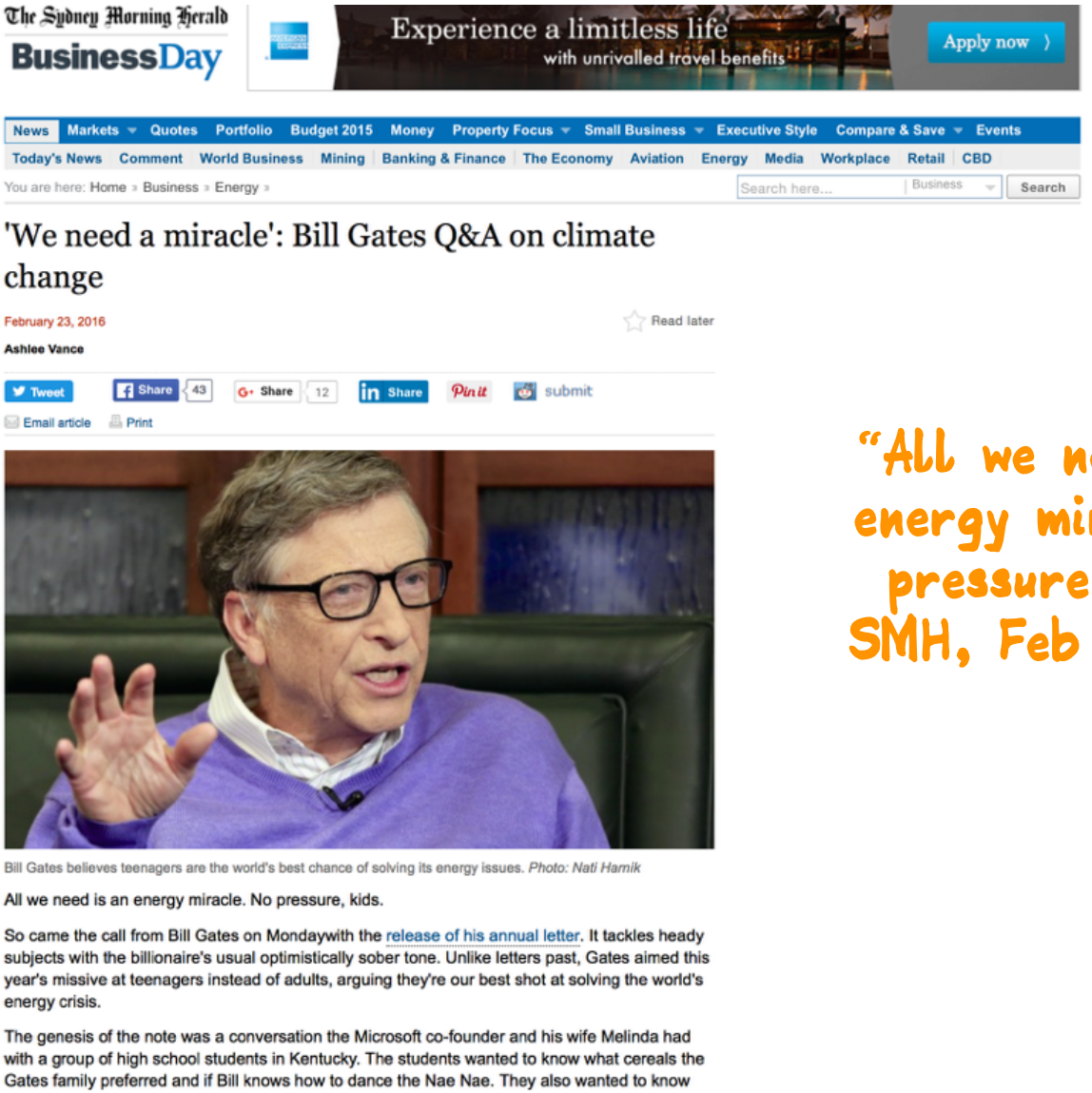
Statistical significance calculation:

$$t = \frac{\bar{x}_{boys} - \bar{x}_{girls}}{\sqrt{s^2_{boys}/n_{boys} + s^2_{girls}/n_{girls}}}$$

if $|t| > t_{1-\alpha/(2*p)}$ (approx 5 points), the difference is not significantly different from 0. If a new sample of boys and girls was tested the result may well be 0.

- What about reading and science scores?
- **The reading gap is universal, in favor of girls**
- Science gap is not universal

Climate change



"All we need is an energy miracle. No pressure, kids."
SMH, Feb 23, 2016

What's the deal about carbon dioxide?

The Sydney Morning Herald
Environment

DANDENONG
WORLD FARE

UN Climate Conference | Weather | Climate Change | Whale Watch | Animals | Conservation | Energy Sn

You are here: Home » Environment »

CSIRO team's study erodes credibility of key soil carbon model

November 1, 2015 Read later

Peter Hannam
Environment Editor, The Sydney Morning Herald
View more articles from Peter Hannam
Follow Peter on Twitter Follow Peter on Google+ Email Peter

Tweet Share 459 Share 11 Share Pin it 1 submit
Email article Print Reprints & permissions

Sydney Barrister Peter King argues a royal commission into banks' treatment of farmers is needed. *Photo: Jessica Shapiro*

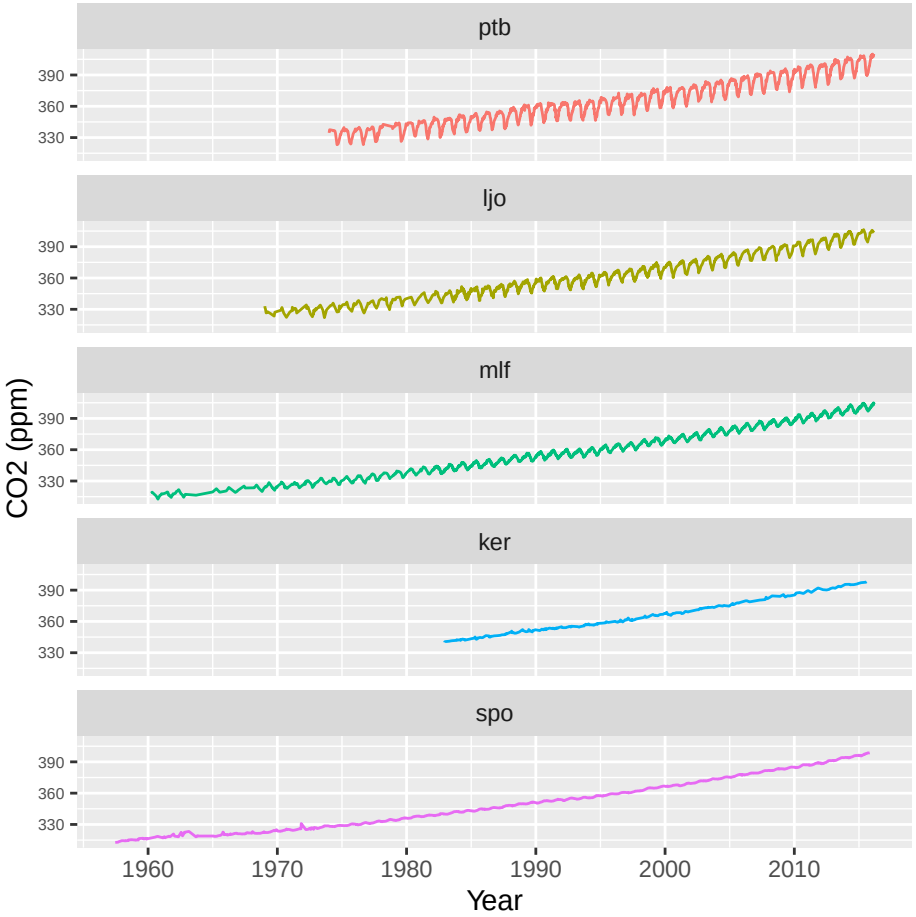
Australia's method of measuring how much carbon is being stored in its soil is flawed, undermining the credibility of government programs to pay farmers to sequester the climate change inducing element, a new study by CSIRO researchers has found.

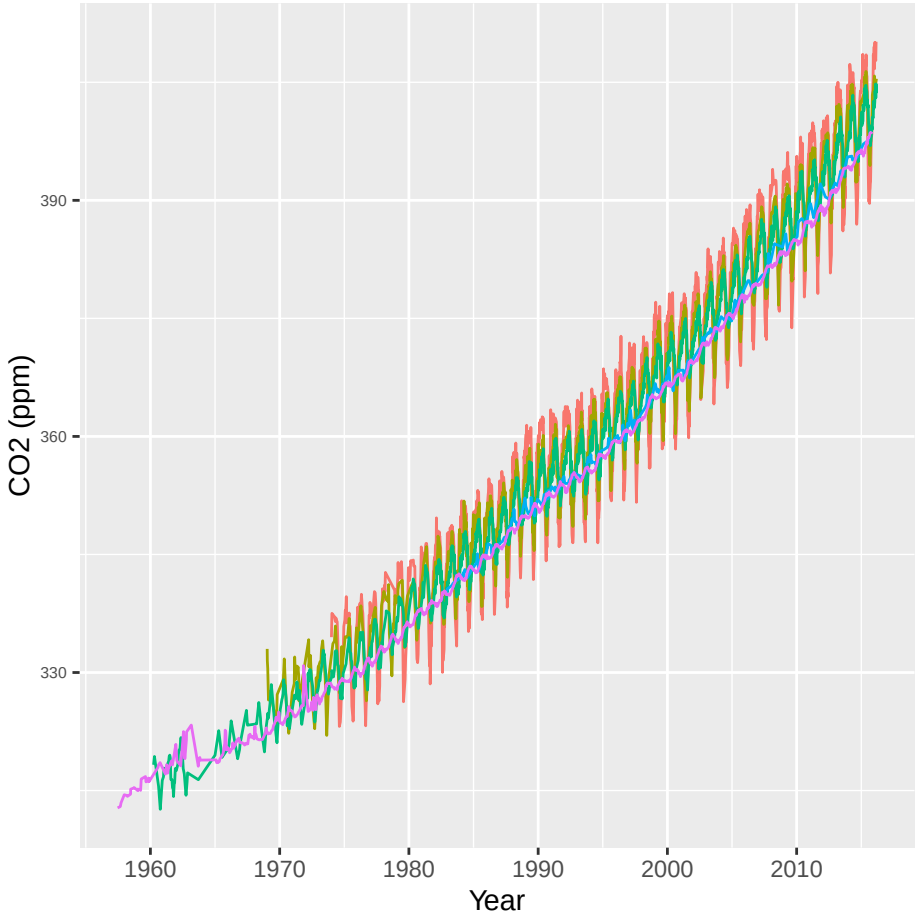
"Australia's method of measuring how much carbon is being stored in its soil is flawed, undermining the credibility of government programs to pay farmers to SEQUESTER the climate change inducing element."

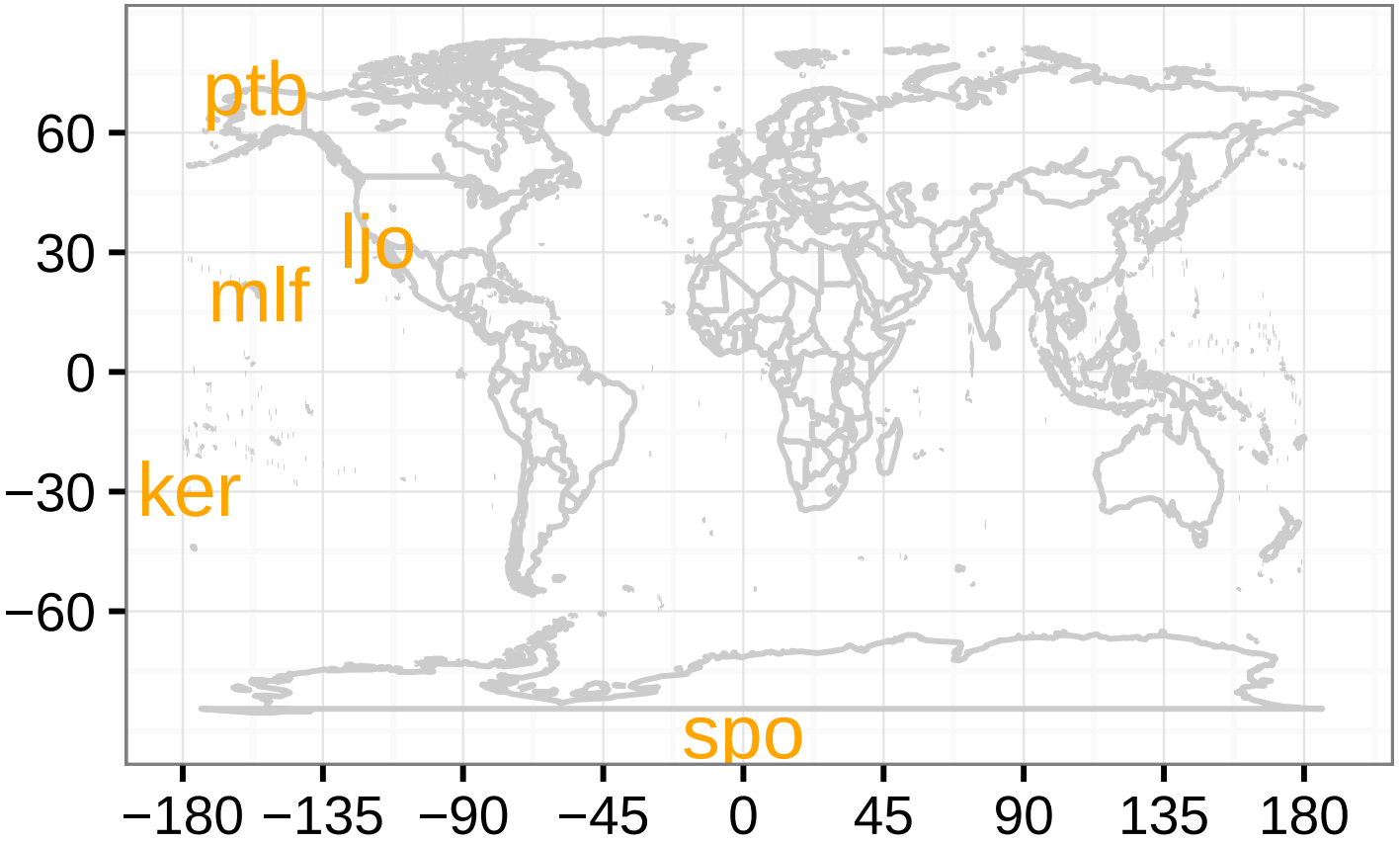
- "Scientific consensus states that carbon emissions must be reduced by 80% by 2050 to avoid temperature rise of more than 2°C." Carbon Neutral
- Carbon offsets: Carbon offsetting is the use of carbon credits to enable businesses to compensate for their emissions.
- Kyoto protocol in 1992, attempt to get international cooperation to reduce emissions.

Carbon dioxide data

- Data is collected at a number of locations world wide.
- See Scripps Inst. of Oceanography
- Let's pull the data from the web and take a look ...
-
- Recordings from South Pole (SPO), Kermadec Islands (KER), Mauna Loa Hawaii (MLF), La Jolla Pier, California (LJO), Point Barrow, Alaska (PTB).


ptb
ljo
mlf
ker
spo
390
360
330
CO2 (ppm)
1960
1970
1980
1990
2000
2010
Year

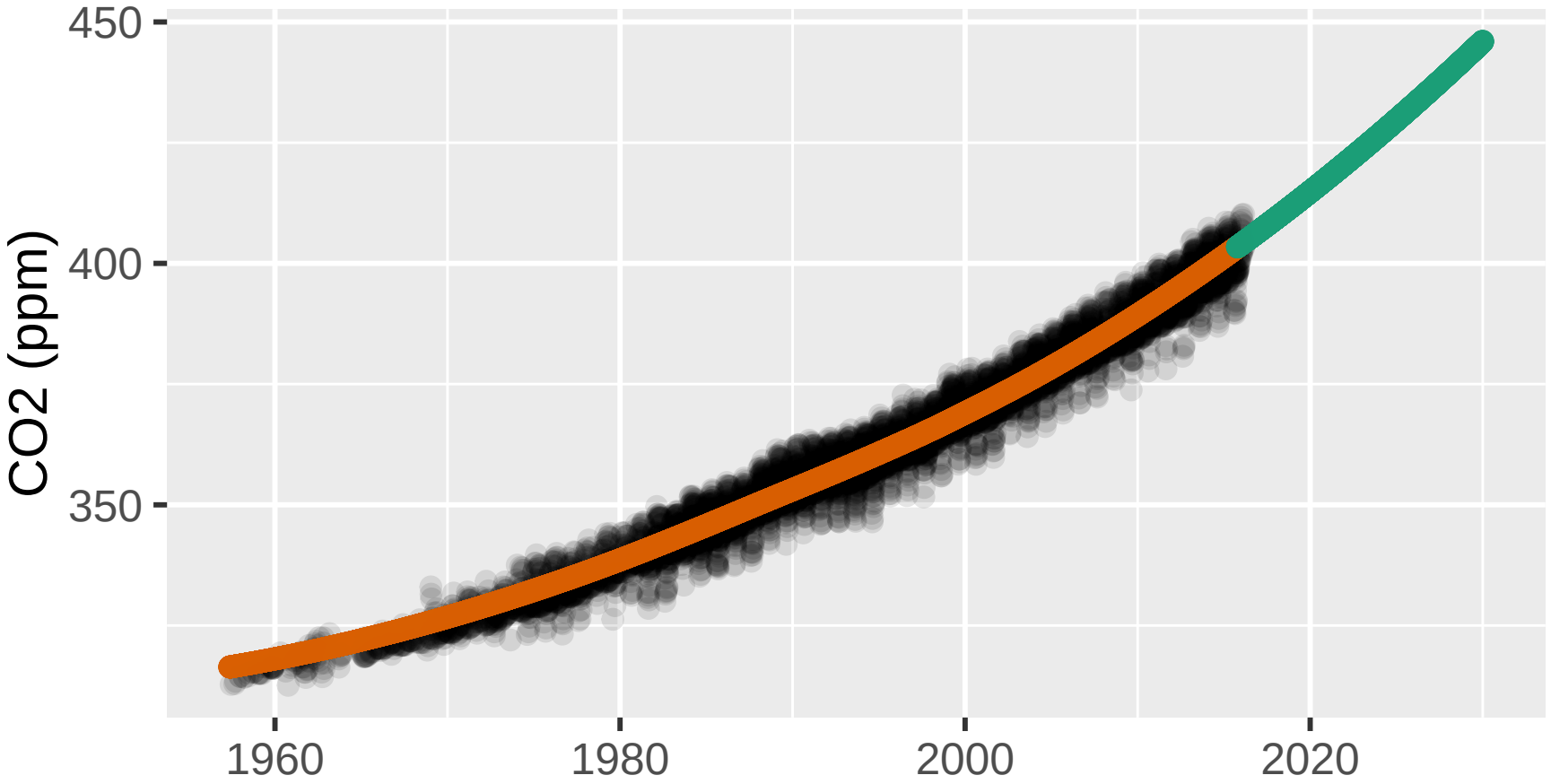

390
360
330
CO2 (ppm)
1960
1970
1980
1990
2000
2010
Year


ptb
ljo
mlf
ker
spo
60
30
0
−30
−60
−180
−135
−90
−45
0
45
90
135
180

What do we learn?

- CO_2 is increasing, and it looks like it is exponential increase. **I really expected that the concentration would have flattened out with all of the efforts to reduce carbon emissions.**
- The same trend is seen at every location - REALLY? Need some physics to understand this.
- Some stations show seasonal pattern - actually the more north the more seasonality - WHY?

Where will it be in 2030



In 2030 the CO_2 will be at 446 ppm.

Relationship between CO2 and temperature

- According to NASA: *Atmospheric carbon dioxide performs a role similar to that of the house thermostat in setting the equilibrium temperature of the Earth.* The global temperature has increased about 0.2^oC per decade, as CO_2 has increased around 10ppm per decade for the same period.
- Based on our model CO_2 is set to increase from 405ppm to 450ppm, an increase of 45. This would indicate a potential temperature increase of $4.5 \times 0.2 = 0.9^oC$ for this period.

Here's what the Guardian says. . .

The Sydney Morning Herald
Environment

DANDENONG
WORLD FARE

UN Climate Conference | Weather | Climate Change | Whale Watch | Animals | Conservation | Energy Sn

You are here: Home » Environment »

CSIRO team's study erodes credibility of key soil carbon model

November 1, 2015 Read later

Peter Hannam
Environment Editor, The Sydney Morning Herald
View more articles from Peter Hannam
Follow Peter on Twitter Follow Peter on Google+ Email Peter

Tweet Share 459 Share 11 Share Pin it 1 submit
Email article Print Reprints & permissions

Sydney Barrister Peter King argues a royal commission into banks' treatment of farmers is needed. *Photo: Jessica Shapiro*

Australia's method of measuring how much carbon is being stored in its soil is flawed, undermining the credibility of government programs to pay farmers to sequester the climate change inducing element, a new study by CSIRO researchers has found.

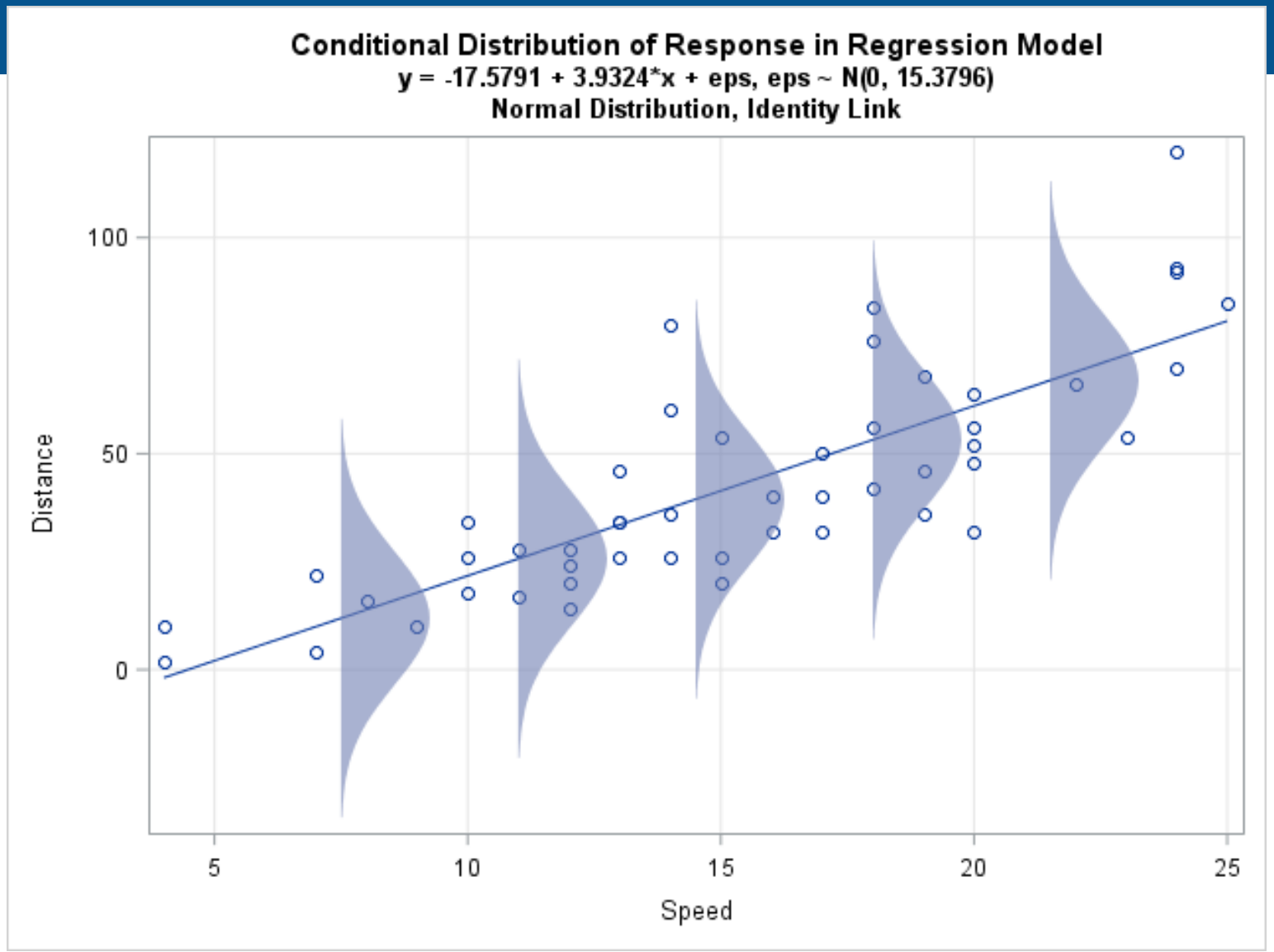
"Australia's method of measuring how much carbon is being stored in its soil is flawed, undermining the credibility of government programs to pay farmers to SEQUESTER the climate change inducing element."

How did we fit the model?

- A loess smoother was applied to the data, see wikipedia. (Explanation is ok, but could do with some more editing.)
- A great animation illustrating the loess method is available at this web site

Modeling data

- Non-parametric models, e.g. loess, are data centric
- Linear models, depend on parametric assumptions, e.g. regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where ε represents error from the model. Here the error is assumed to follow a bell curve.



Source: SAS Blog

Prediction from a linear model

- Estimate: If $x = 10$, then y would be predicted to be $-17.5791 + 3.9324 * 10 = 21.7449$.
- Variation: We would expect that 68% of observations with $x = 10$ to fall between $21.7449 - 15.3796 = 6.3653$ and $21.7449 + 15.3796 = 37.1245$.

How to be a data scientist

- Start by thinking of good questions to ask
- Then find what data is available
- How do we make calculations and plots which would help us answer the questions
- Open source software for data analytics: `http://www.r-project.org`
- You have to program a little, but its not too hard, and we will help

The analysis in these slides

were produced using the R software, with functions from a lot of user contributed packages, e.g.

- `readr`: read some data into the software
- `qplot`: make some pictures
- `summarise`: make some calculations
- `loess`: model the future

These slides

- This is a “live” document
- Code and explanations together
- Run the software to make the calculations on the data, and produce nice presentation, or Word or pdf or html document

Why?

"*R has become the most popular language for data science and an essential tool for Finance and analytics-driven companies such as Google, Facebook, and LinkedIn.*" Microsoft 2015

Top 10 Best Jobs

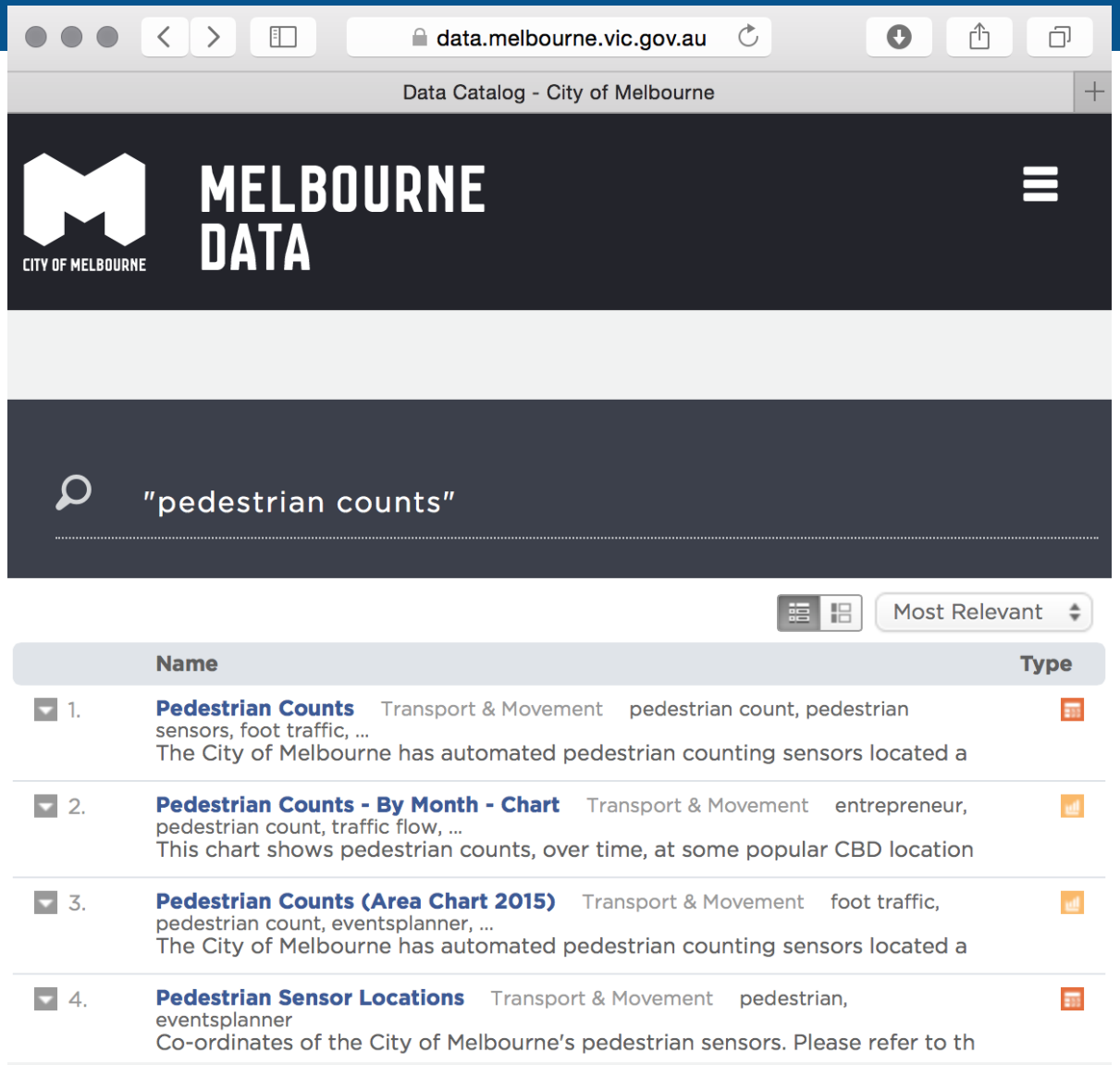
1. *Actuary*
2. Audiologist
3. *Mathematician*
4. *Statistician*
5. Biomedical Engineer
6. **Data Scientist**
7. Dental Hygienist
8. Software Engineer
9. Occupational Therapist
10. Computer Systems Analyst

2015 Career Best: `http://www.careercast.com`

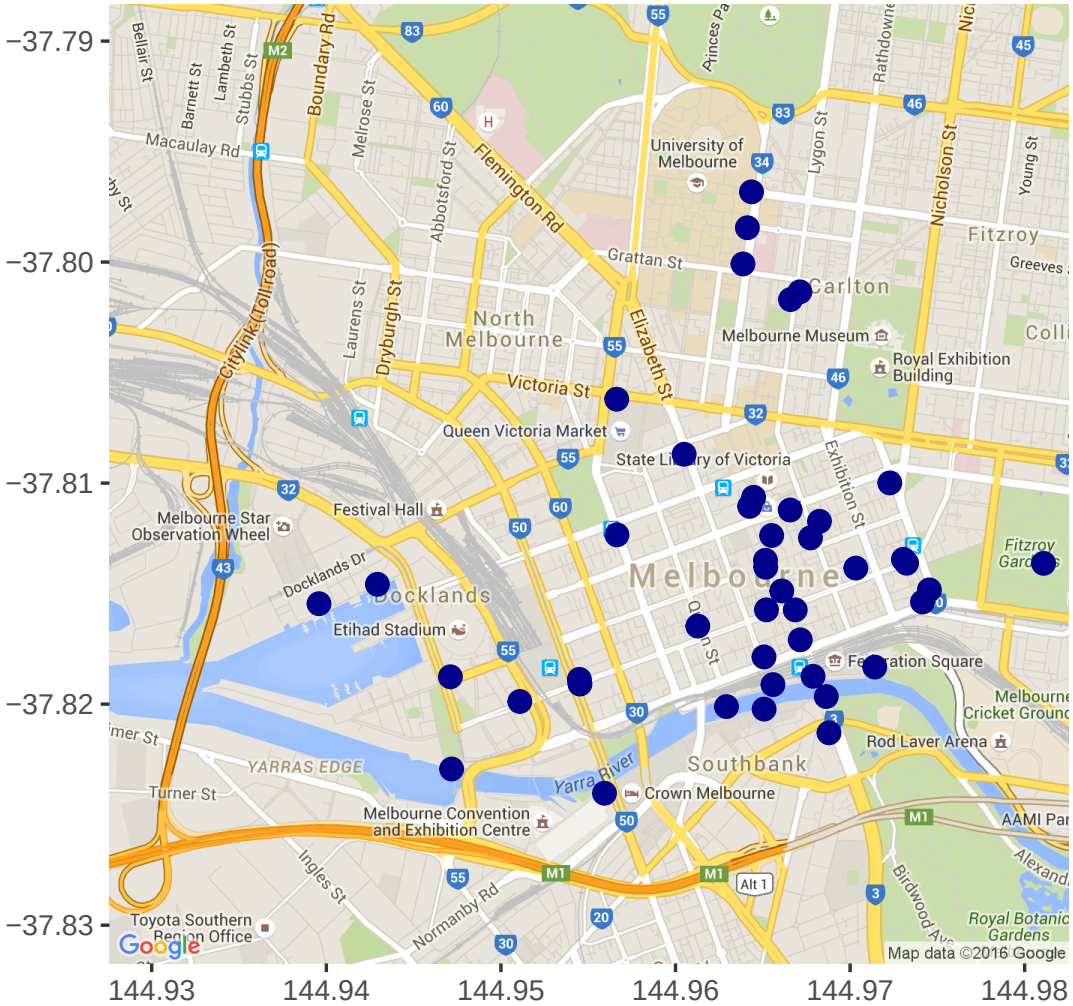
"*By 2018, the US could face a shortage of up to 190.000 workers with analytical skills*" McKinsey

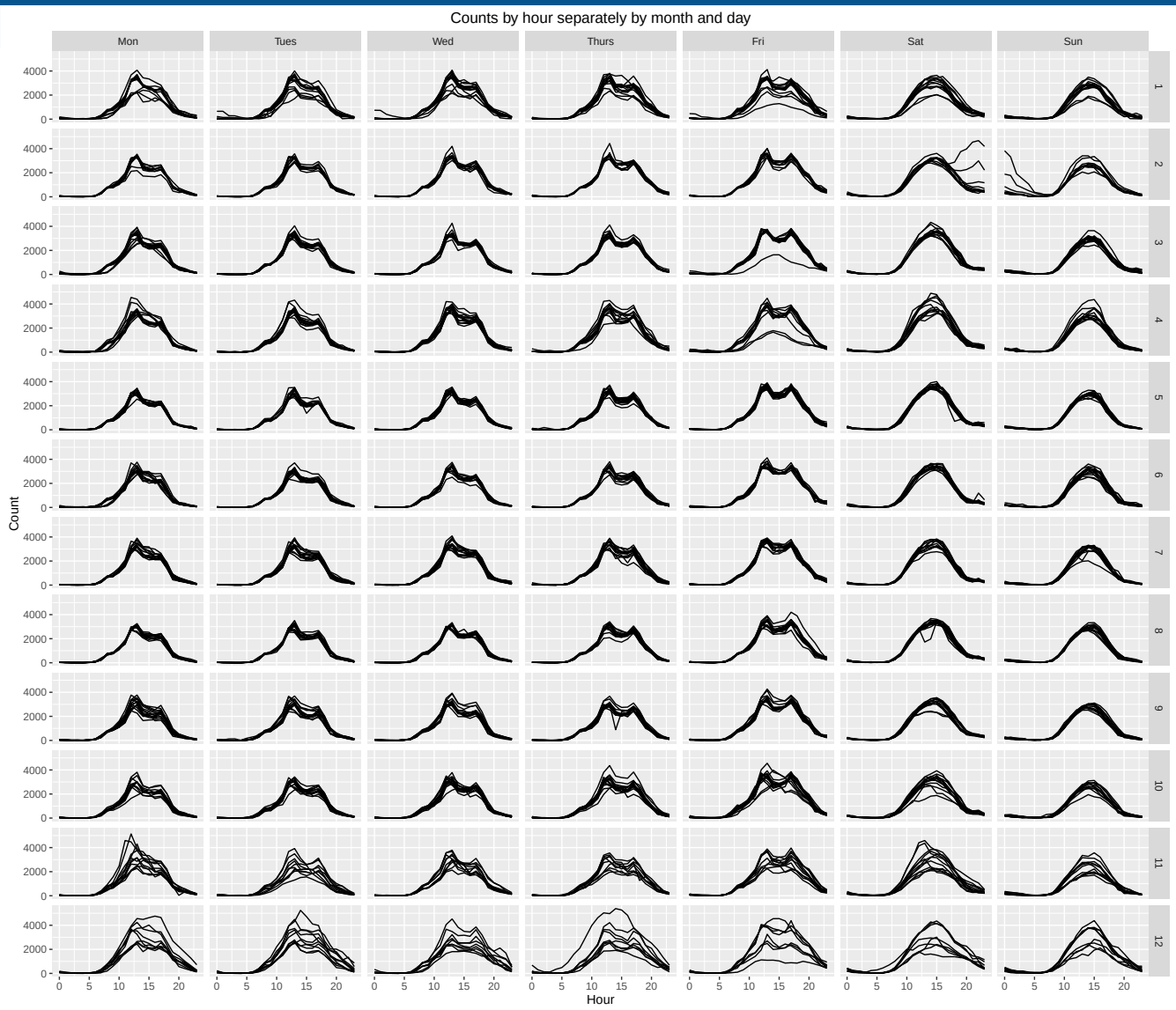
Today's lab

- A look at Melbourne's pedestrian traffic, by pulling data from sensors placed around the city
- Where are people walking? What times are there more people? Is the pattern different on weekends, or in different months, or at different locations?


data.melbourne.vic.gov.au
Data Catalog - City of Melbourne
CITY OF MELBOURNE
MELBOURNE DATA
"pedestrian counts"
Most Relevant
Name
Type
1. Pedestrian Counts Transport & Movement pedestrian count, pedestrian sensors, foot traffic, ...
The City of Melbourne has automated pedestrian counting sensors located a
2. Pedestrian Counts - By Month - Chart Transport & Movement entrepreneur, pedestrian count, traffic flow, ...
This chart shows pedestrian counts, over time, at some popular CBD location
3. Pedestrian Counts (Area Chart 2015) Transport & Movement foot traffic, pedestrian count, eventsplanner, ...
The City of Melbourne has automated pedestrian counting sensors located a
4. Pedestrian Sensor Locations Transport & Movement pedestrian, eventsplanner
Co-ordinates of the City of Melbourne's pedestrian sensors. Please refer to th

Sensor locations




Counts by hour separately by month and day
Mon
Tues
Wed
Thurs
Fri
Sat
Sun
1
2
3
4
5
6
7
8
9
10
11
12
Count
4000
2000
0
0
5
10
15
20
Hour

Lab help

- Carson (USA), Earo (China), Mitch (Australia) and Nathaniel (Australia)
- See you in the lab soon to get our hands on the data!