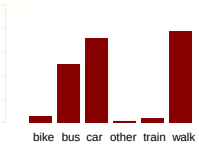
NZAMT 2013

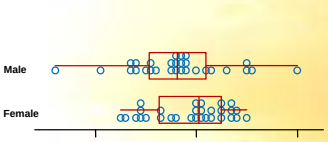
Caught in the Path of the Data Deluge

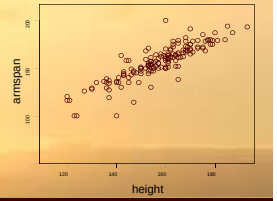
Chris Wild
Department of Statistics
University of Auckland, New Zealand

Laying story-telling foundations

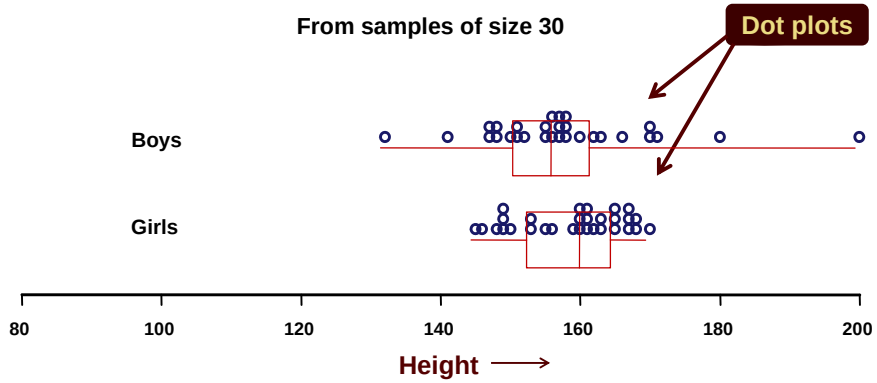
- Some very basic graphics



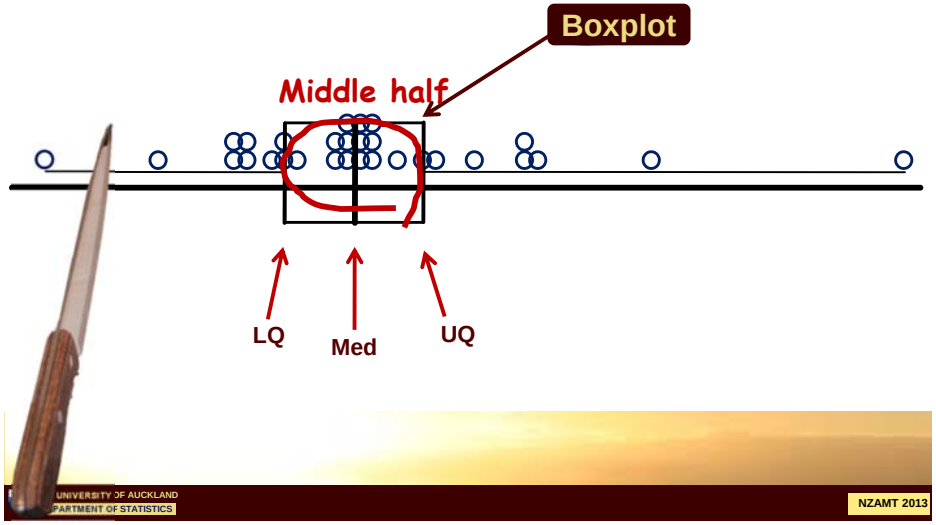




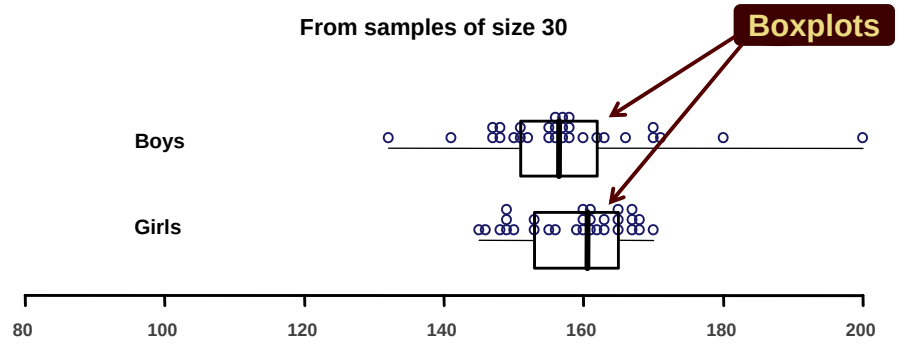
Comparing heights of boys and girls at age 12



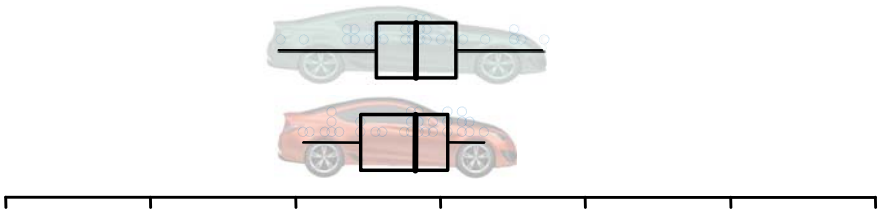
Aside: Marking dot plot up with with boxplot



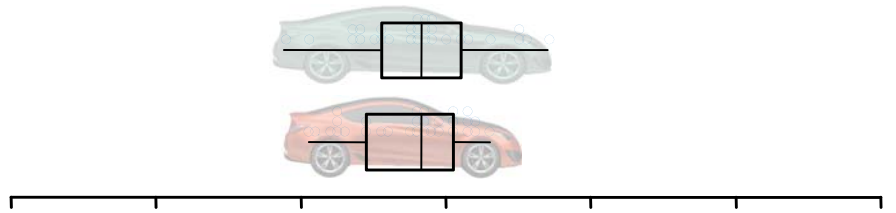
Comparing heights of boys and girls at age 12



Learning to see shifts

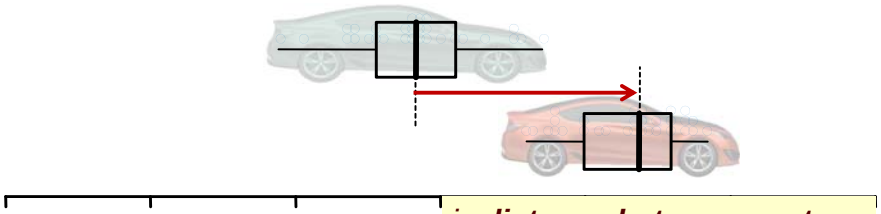


Learning to see shifts



Learning to see shifts

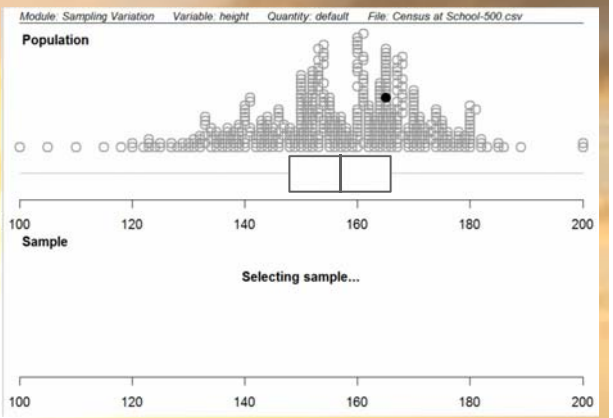
Obvious measure of ***extent of shift …***



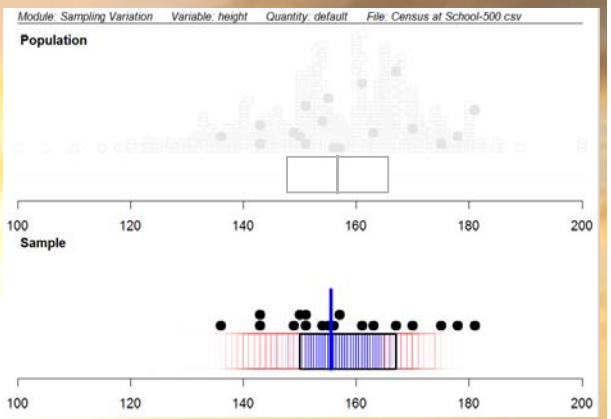
is ***distance between centres***

Experiencing Sampling Variation

Use ***boxplots***
Track
- medians
- quartiles



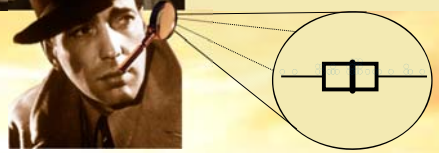
Use ***boxplots***
Track
- medians
- quartiles



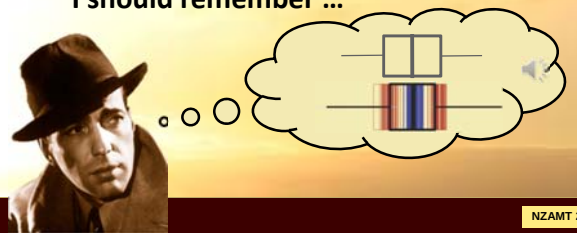
Makes …

Want to plant a reflex …

"Whenever I see …



"I should remember …



Desired reflex …

"Whenever I see …

"I should remember …

"Mine could be like this …"

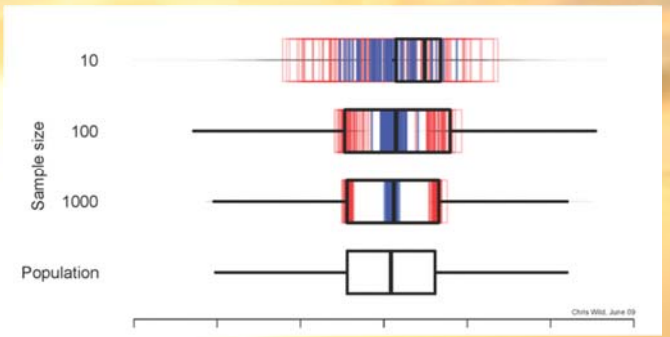
"Or even this …"

"Confidence Interval"

"I must take this uncertainty *about where it really should be* into account "

Very Useful

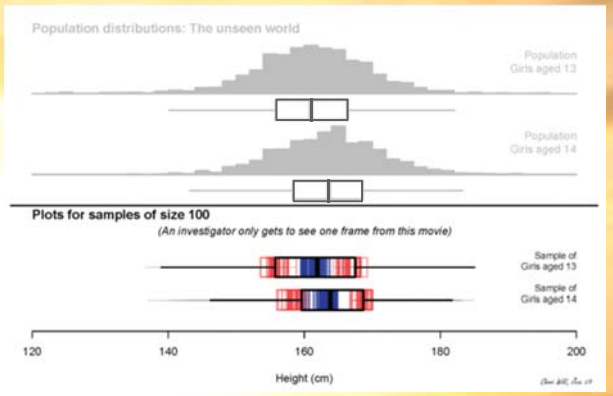
Effect of Sample Size



Towards Visual Inference
- Inference & data production
- Depicting Sampling Variation
- Uncertainty intervals

Very Useful

Comparisons



The Main Feature

Caught in the Path of the Data Deluge

Chris Wild

Department of Statistics
University of Auckland, New Zealand

Looking downwards

Looking downwards

NCEA Level 8 Stats 2013

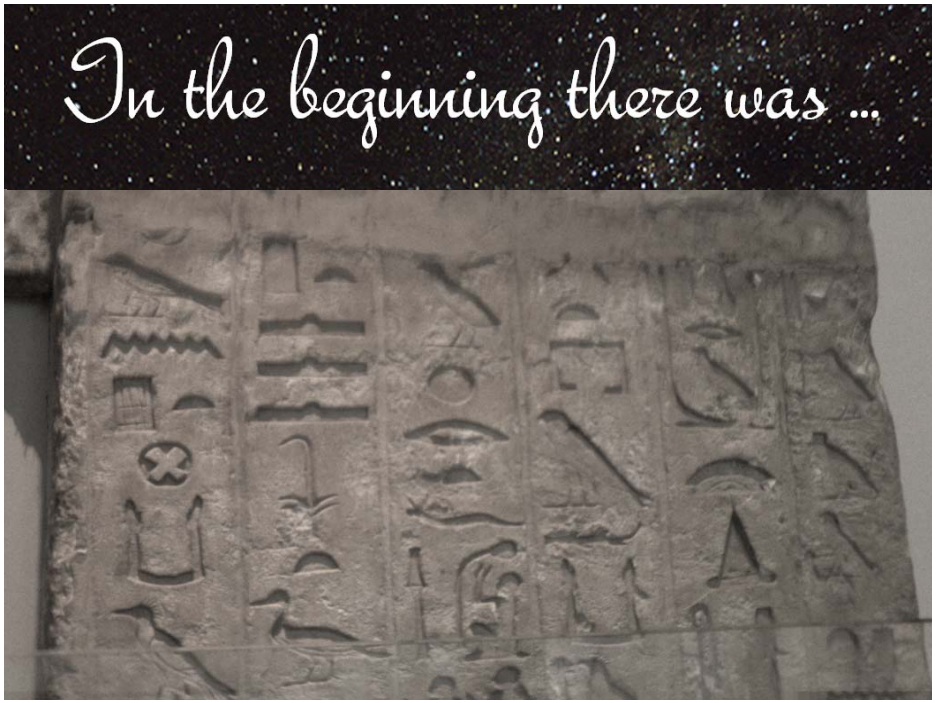
- **Getting my head around teaching all this new stuff – it's a lot of work**
- **How to play the new NCEA games well?** (for my kids)
- **What are the new game's rules??? Don't seem clear yet *!!***

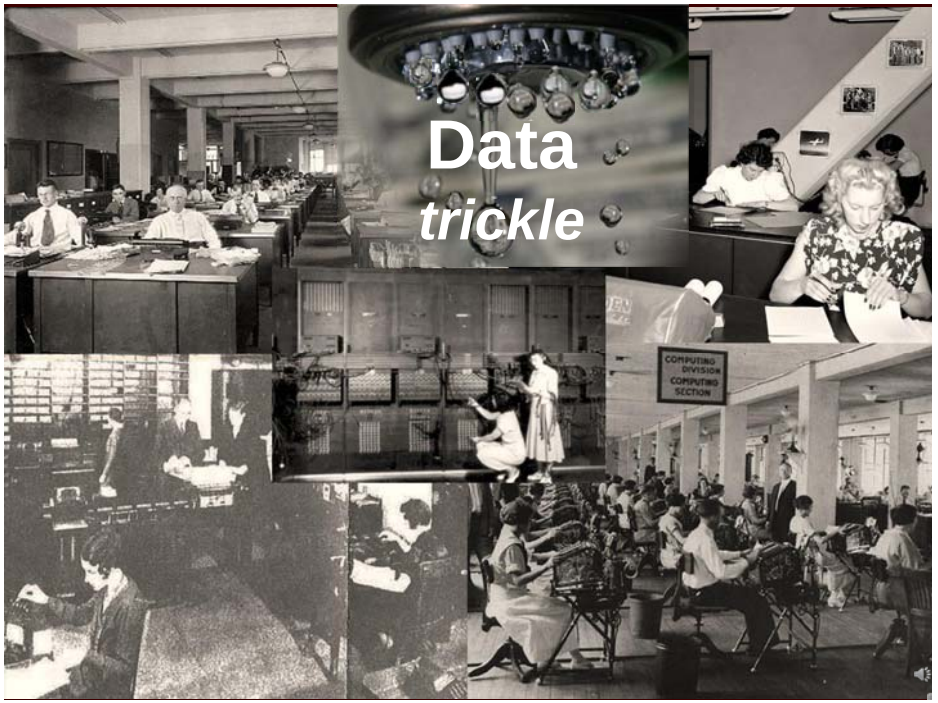
Looking downwards

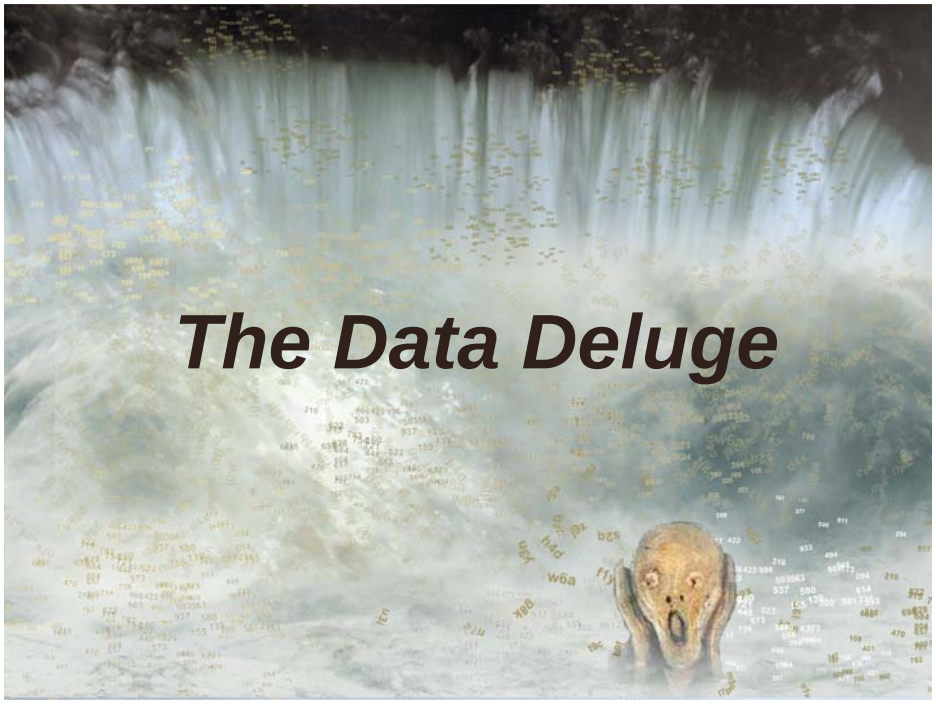
NCEA Level 8 Stats 2013

- **Let's drag ourselves out of the mire**
  - put aside teething problems
- **and lift up our eyes!**


A Brief History of
Data


In the beginning there was ...


Data
trickle


The Data Deluge


The New York Times
SundayReview | The Opinion Pages
WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEAL
NEWS ANALYSIS
The Age of Big Data
By STEVE LOHR
Published: February 11, 2012 | 82 Comments
GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.


THE GLOBE AND MAIL
Search: | News & Quotes | Jobs
Home News Commentary Business Investing Sports Life
Streetwise Economy International Industry News Small Business
Career Advice Management Leadership Business Education Top Employer
Home » Report on Business » Globe Careers » Business Education
ANALYTICS
Wanted: People who can crunch numbers and see the future
LYNN GREINER
Special to The Globe and Mail
Published Tuesday, Jun. 11 2013, 7:00 PM EDT


The Economist
World politics Business & finance Economics Science & technology Culture
Special report: Managing information
Data, data everywhere
Information has gone from scarce to superabundant. That brings huge new benefits, says Kenneth Cukier (interviewed here)—but also big headaches
Feb 25th 2010 | From the print edition
Like 420 Tweet 272

Data scientist shortage set to endure as demand grows

publication date: Jan 16, 2012

In December, IT service company EMC published its Data Scientist Study 2011, which forecast a shortage of people with suitable skills for at least the next five years.

The study – which EMC said was the largest-ever global survey of the data science community spanning the US, UK, France, Germany, India and China – comes at a time when the volume of data being generated is growing exponentially.

" We've watched as corporations turn the discussion from "spending money to save money", to "spending money to make money", and that discussion is underpinned by the Big Data opportunity at every company's doorstep. There is a critical demand for people with the ability to analyze and create value from Big Data, and …

Paul Zikopoulos
Director of Information Management
IBM

" For an organization to compete today, it has to embrace modern business complexity and harness the incredible power of data. A successful business now requires a proficient team of managers with both an in-depth knowledge of analytical tools and expertise in business.

Dr. Yuri Medvedev
Chief Mathematician
Bank of Montreal

So what's driving all of this?

- Cheap electronic capture of data
- Making money
- Saving money
- Saving Lives

There's gold in them thar data!

So the data world …
is getting a whole lot bigger

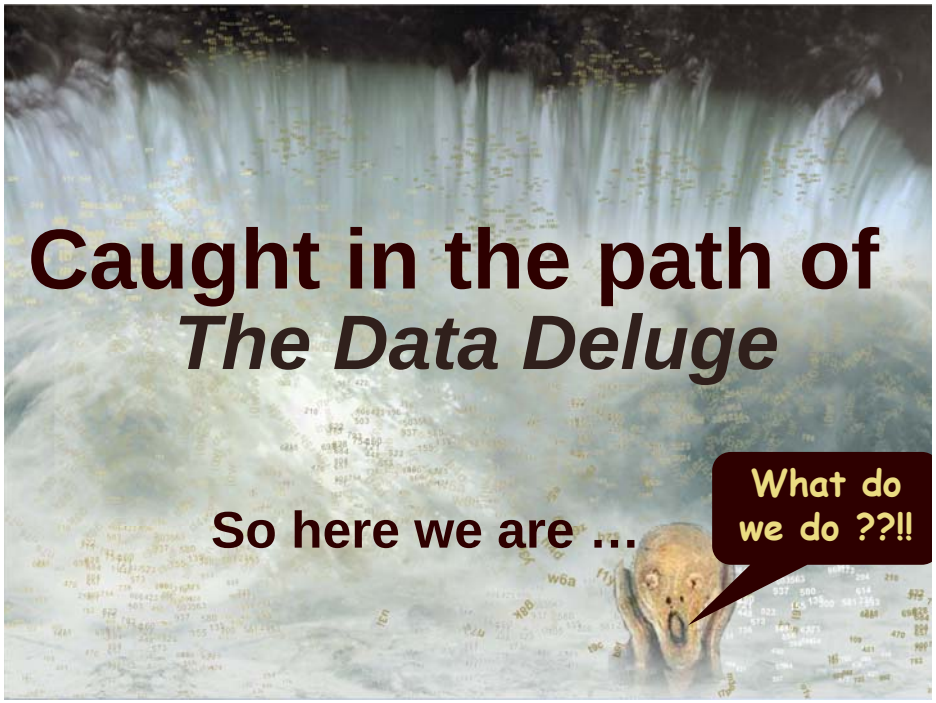
So the data world …
is getting a whole lot bigger

- There is an explosion in the …
  - quantities of data being collected
  - conceptions of what constitutes data
  - settings in which it can arise
  - ways of looking at it

Caught in the path of
The Data Deluge

So here we are …

What do we do ??!!



The democratisation of data

The democratisation of data

The democratisation of data

has 2 key ingredients:

"Power to the people!"

Competitive economies

1. **Access to data**
   - Open data movement …
2. **Ability to understand it**
   - *for most people*
     - » **And that's where we come in!**

With the data world … getting a whole lot bigger

Can't just keep illuminating same small patch

- **Need to get much …**
  - ***further***
  - ***faster***
  - & with ***better*** *comprehension*

Opening Up the World of data

wider, faster

VISION Statement for Early Statistics

- To ***create excitement*** about
  - **"What I can do with data &**
  - **What data can do for me"**

A educational maxim

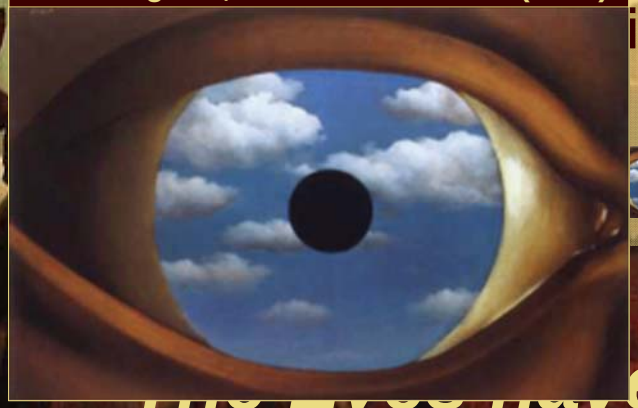
What can machines do?

and what thinking is essentially human?

Education should emphasise the latter

Further, Faster, Better
- Data world exploding
- Green shoots in Software
- Vision
- Future is visual
- Accelerators

René Magritte, "*The False Mirror*" (1928)



"the future of ical is

e it"

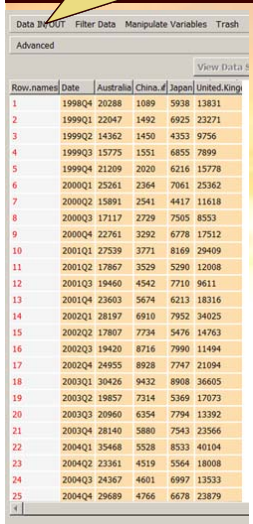
Further, Faster, Better
- Data world exploding
- Green shoots in Software
- Vision
- Future is visual
- Accelerators

"the future of improved statistical understanding is *visual*"

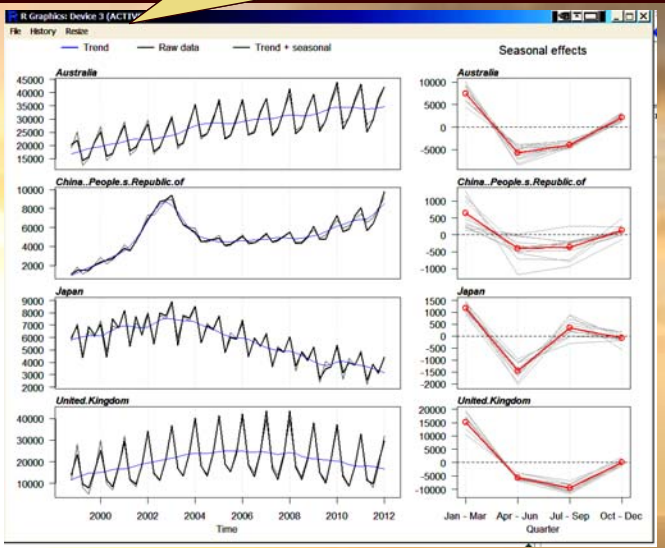
improved **understanding of ...**

- data
- inferential concepts

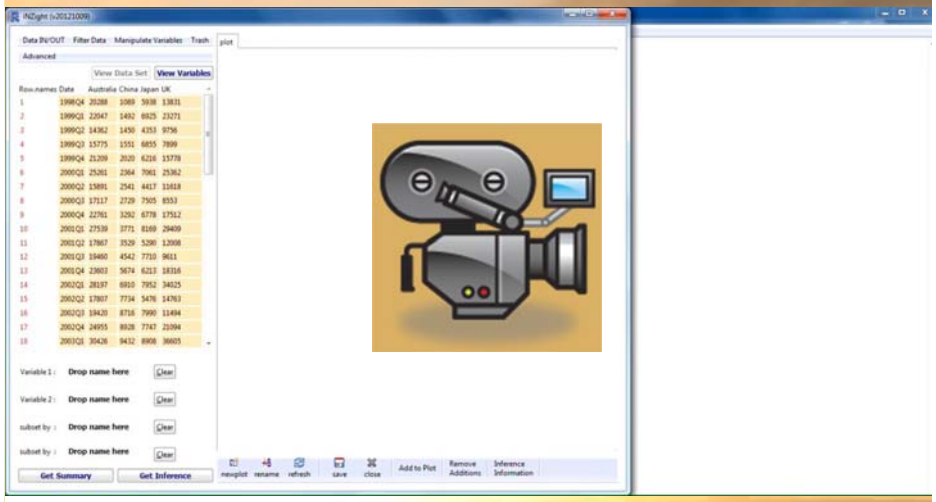
"I was blind" e s "but now I see"



Row.names	Date	Australia	China..P	Japan	United.King
1	1998Q4	20288	1089	5938	13831
2	1999Q1	22047	1492	6925	23271
3	1999Q2	14362	1450	4353	9756
4	1999Q3	15775	1551	6855	7899
5	1999Q4	21209	2020	6216	15778
6	2000Q1	25261	2364	7061	25362
7	2000Q2	15891	2541	4417	11618
8	2000Q3	17117	2729	7505	8553
9	2000Q4	22761	3292	6778	17512
10	2001Q1	27539	3771	8169	29409
11	2001Q2	17867	3529	5290	12008
12	2001Q3	19460	4542	7710	9611
13	2001Q4	23603	5674	6213	18316
14	2002Q1	28197	6910	7952	34025
15	2002Q2	17807	7734	5476	14763
16	2002Q3	19420	8716	7990	11494
17	2002Q4	24955	8928	7747	21094
18	2003Q1	30426	9432	8908	36605
19	2003Q2	19857	7314	5369	17073
20	2003Q3	20960	6354	7794	13392
21	2003Q4	28140	5880	7543	23566
22	2004Q1	35468	5528	8533	40104
23	2004Q2	23361	4519	5564	18008
24	2004Q3	24367	4601	6997	13533
25	2004Q4	29689	4766	6678	23879



Time Series



Further, Faster, Better

- Data world exploding
- Green shoots in Software
- Vision
- Future is visual
- Accelerators

Shift of focus

From:

- *"What are all the things I have to do to get the output I'm meant to produce?"*

To:

- *"What are the questions?"*
- *"What can I **see**?"*
- *"What does that tell me?"*

Using ***"enabling software"*** to **"cut out the middle man"**

Addressing the Need for Speed

- **iNZight-type software can facilitate a fast, accessible way in to understanding a much wider spectrum of data types**
- **VIT-type software can facilitate a fast , accessible way in to understanding basic inferential conceptions**

My "vision" is

- **initially create an appreciation of a very wide array of data types and what they can tell you**
  - **and only then back fill the details** (for those who need it)

Online Activities

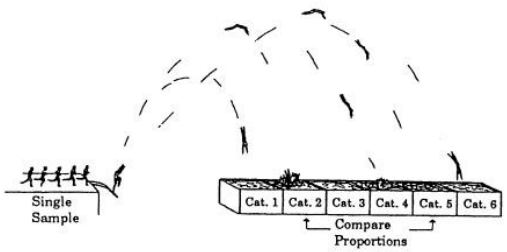
From Census At School NZ

27. In the last seven days, which of these online activities have you done? (You may tick more than one)

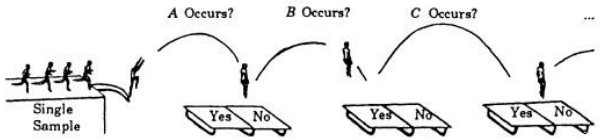
- ☐ Downloaded or listened online to music
- ☐ Downloaded or viewed online video (e.g. YouTube, TV shows, movies)
- ☐ Played online game(s)
- ☐ Kept in touch with friends (e.g. through instant messaging, Bebo or Facebook)
- ☐ Researched topics related to school work
- ☐ Have been online but did none of the above activities
- ☐ Have not been online in the last seven days

Comparison

- Single response

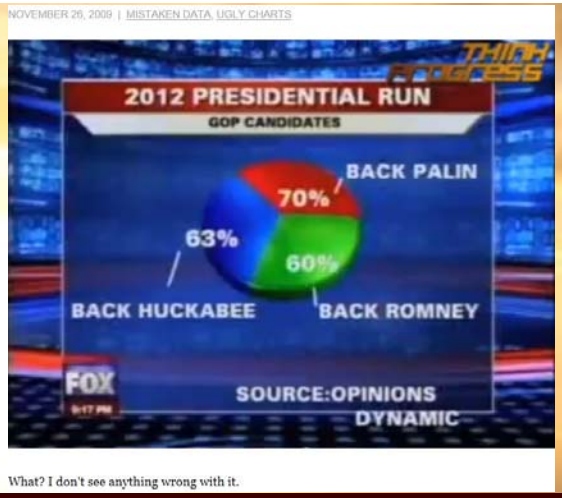


- Multiple response



How do people deal with MR questions?

Pie charts that add to more than 100%!

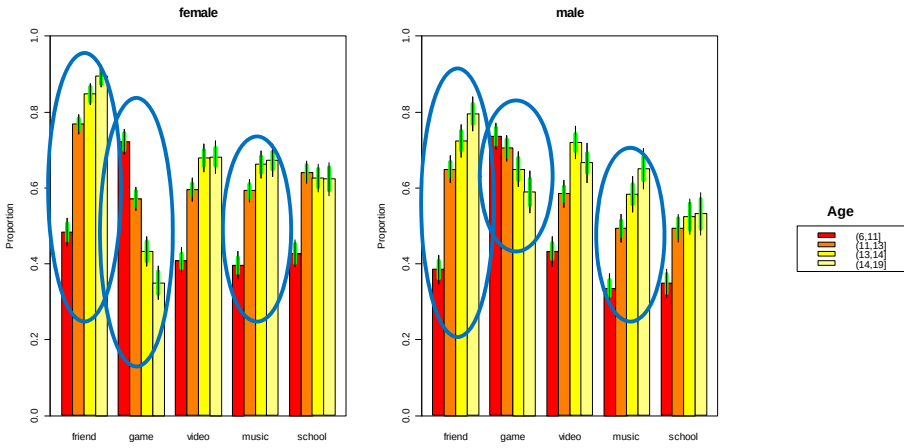


Data Structure

AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG
technone	cellmonths	onlinemusic	onlinevideo	onlinegame	onlinefriend	onlineschool	onlineother	onlinenone	bedtime
no	12	yes	yes	no	yes	yes	no	no	11:45:00
no	1	yes	yes	no	yes	no	no	no	12:30:00
no	NULL	yes	yes	yes	yes	yes	no	no	10:00:00
no	36	yes	yes	no	yes	no	no	no	10:00:00
no	10	no	no	no	no	no	no	no	11:30:00
no	12	no	no	yes	yes	yes	no	no	10:00:00
no	1	no	yes	yes	yes	no	no	no	9:00:00
no	17	no	yes	yes	yes	yes	no	no	10:00:00
no	13	no	yes	yes	yes	yes	no	no	8:45:00
no	2	no	no	no	no	no	no	no	8:45:00
no	3	no	yes	no	no	no	no	no	10:00:00
no	13	yes	yes	no	yes	yes	no	no	10:00:00
no	11	yes	yes	no	yes	no	no	no	10:30:00
yes	7	no	no	no	no	no	no	yes	8:00:00
yes	NULL	no	no	no	no	yes	no	no	12:00:00
no	12	yes	yes	yes	yes	yes	no	no	NULL
no	15	yes	yes	yes	yes	no	no	no	10:30:00
no	13	yes	yes	no	yes	yes	no	no	10:15:00
no	14	yes	no	yes	yes	yes	no	no	10:30:00
no	5	yes	yes	yes	yes	yes	no	no	10:30:00
no	5	no	no	yes	no	yes	no	no	8:00:00
no	8	yes	yes	no	yes	no	no	no	22:30:00
no	5	no	yes	yes	yes	no	no	no	10:30:00
no	NULL	no	no	yes	no	no	no	no	10:00:00
no	9	yes	yes	yes	no	yes	no	no	9:30:00
no	14	yes	yes	yes	yes	yes	no	no	11:00:00

"I was blind

"I was blind" teg "but now I see" nder



Some fundamentals of looking at data

With an eye to using much bigger data sets

Fundamental early-stat "big-data" question

- **Of the ways** in which we teach students ***to look*** **at data …**
  - **what still works?**

  and
  - **what needs to change?**
    (as datasets get bigger)

Big statistical lessons

" ***Data beats anecdote***"

– David Moore

Big statistical lessons

"***Garbage In … Garbage Out***"

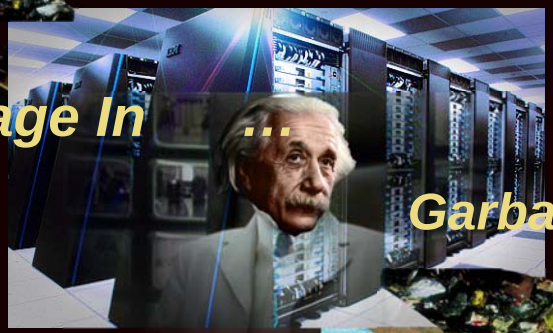
Big statistical lessons

"*Garbage In …*

… Garbage Out"

Big statistical lessons

"*Garbage In …*



… Garbage Out"

Big statistical lessons

Good

"*Data beats anecdote*"

– David Moore

Big statistical lessons

Looking at the world using data is like looking through a window with ripples in the glass

"*What I see …*
is not quite the way it really is"

We see **a blend of fact & artefact**

Big statistical lessons

Constantly struggle to separate

Fact

from

Artefact

Big statistical lessons

The *pleasures of discovery* in data need to be accompanied by…

- **statistical "safe sex" messages**

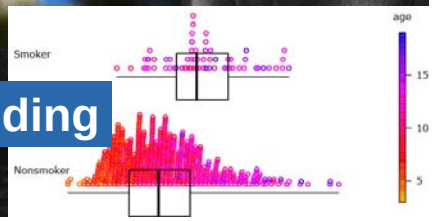
Main causes of artefact

The fundamental things apply …

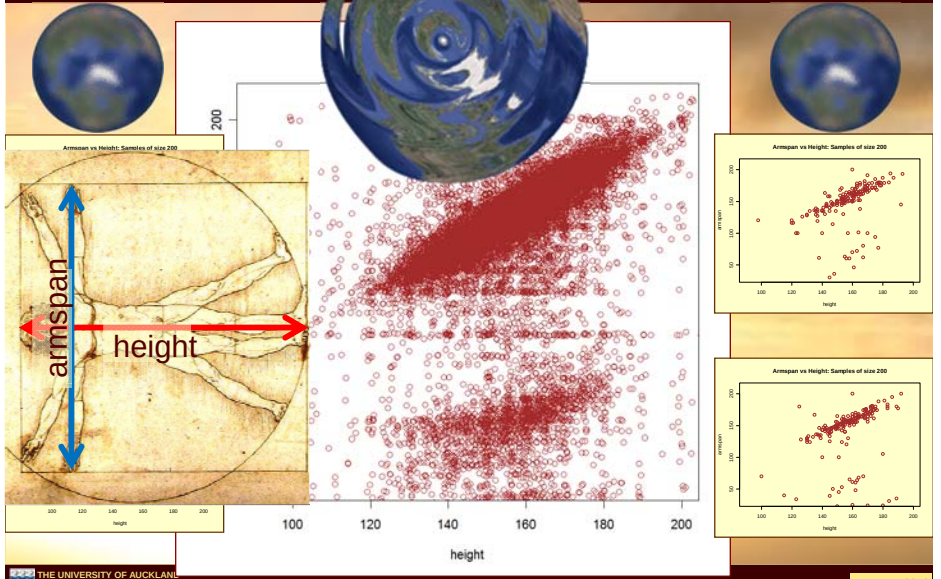
Bias

Random Error

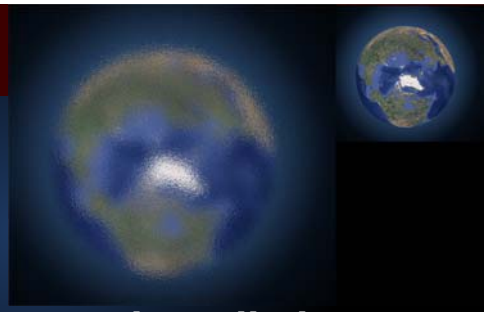
Confounding



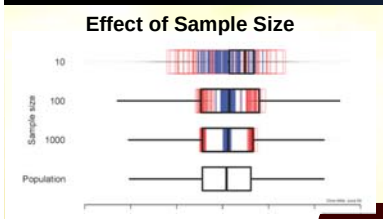
Arm … Height



Back to …



"What I see …
is not quite the way it really is"



Gathering more data (incr. sample size)**…**
is like sharpening the focus

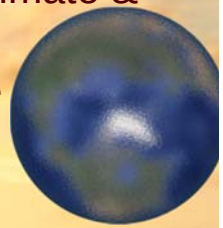
Fact versus Artefact

Fact versus Artefact

Fact versus Artefact

Biggest difference in "detailed features"

- ***With smaller data sets***, we only estimate & interpret gross/global features
  - *Realise* detailed features indistinguishable from randomness
    - *Lucky by-product*: deficiencies in data at a very detailed level may not hurt that much
  - **"Artefact filtering"** in terms of **"significance"**, etc
    - In statistical modelling, details swept up into "random error"
      - *Lucky by-product*: simple, comparatively understandable models
        - » e.g. no high-order interactions



Fact versus Artefact

Biggest difference in "detailed features"

With very big data sets …

- many detailed features are not "random"
  - Large numbers of effects are "significant"
  - but may not be real either
- **"Artefact detection"** has to consider distorting filters that may be "sieving" the streams of data we get to see
  - cf. staring through a distorting camera lens

Fact versus Artefact

Biggest difference in "detailed features"

With very big data sets …

- **Quite detailed features are not random**
  - **Large numbers of effects are "*significant*"**
  - **That which was swept into "random error" becomes structure** to be investigated/understood/interpreted
- **"Artefact detection"** has to consider distorting filters that may be "sieving" the streams of data we get to see
- **Models** likely have to **smooth over smaller features** we know are present **in order to be understandable**

Good news stories

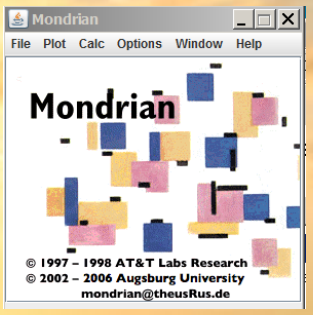
1. Most of our graphs "scale"

- **The huge conceptual leap comes** …
  - when move ***from representing individual points*** (concrete connections to data)
  - ***to depicting abstractions***
    - Summaries (e.g. counts and proportions), densities, …
- Once that hurdle has been jumped …
  - ***size doesn't matter***
    - Well not much anyway!
      - "2nd order effects" that are relatively easy to understand

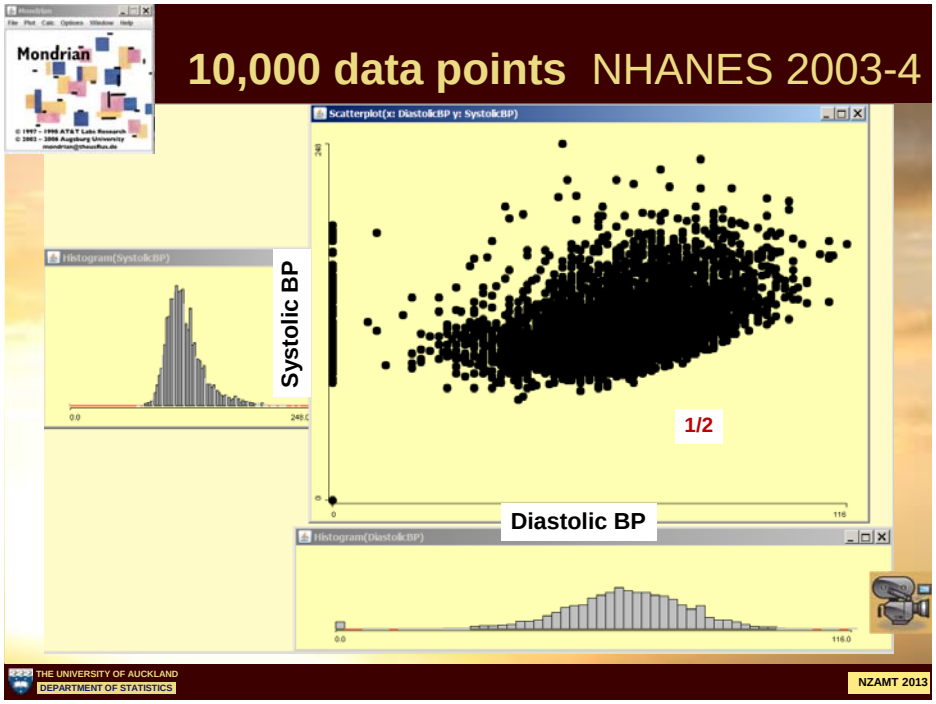
Good news stories

Even "individual-points" plots can be surprisingly useful with quite large data sets

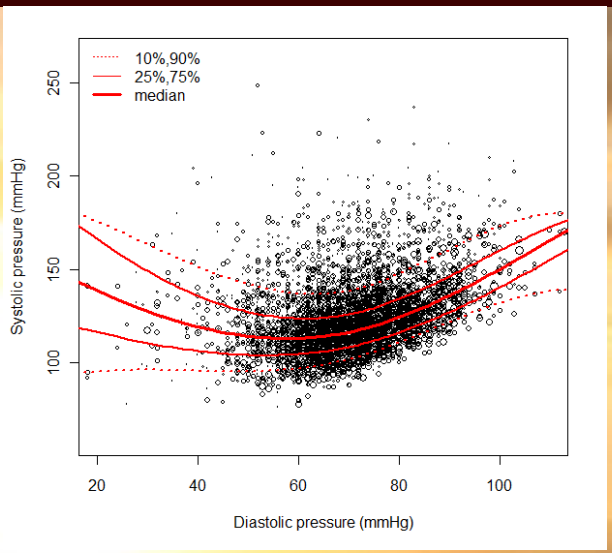
- **Will now use Mondrian …**
  (by Martin Theus)
  - **to look at some quite large data sets using the humble scatterplot**



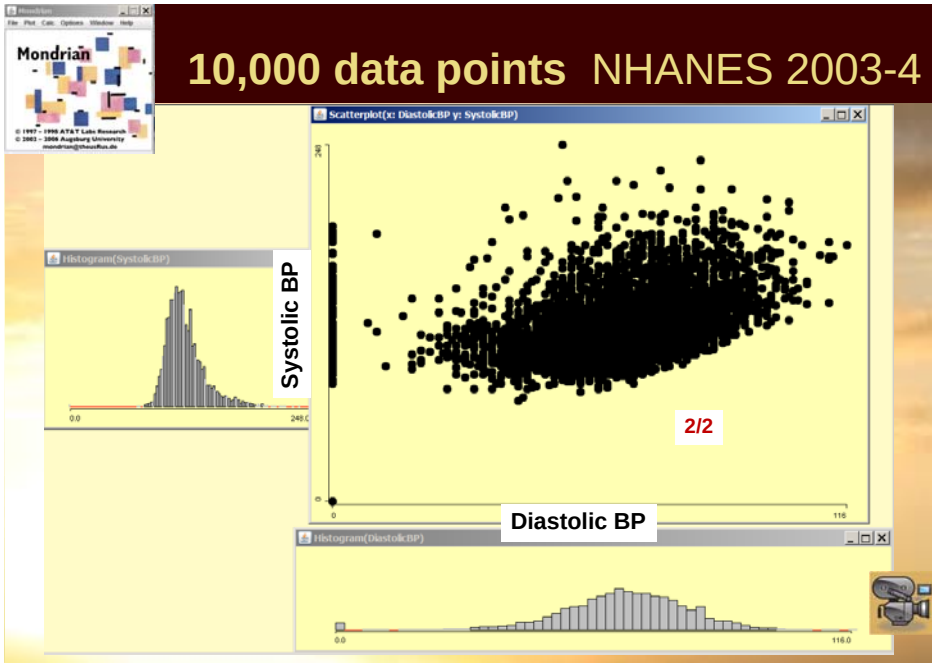
10,000 data points NHANES 2003-4



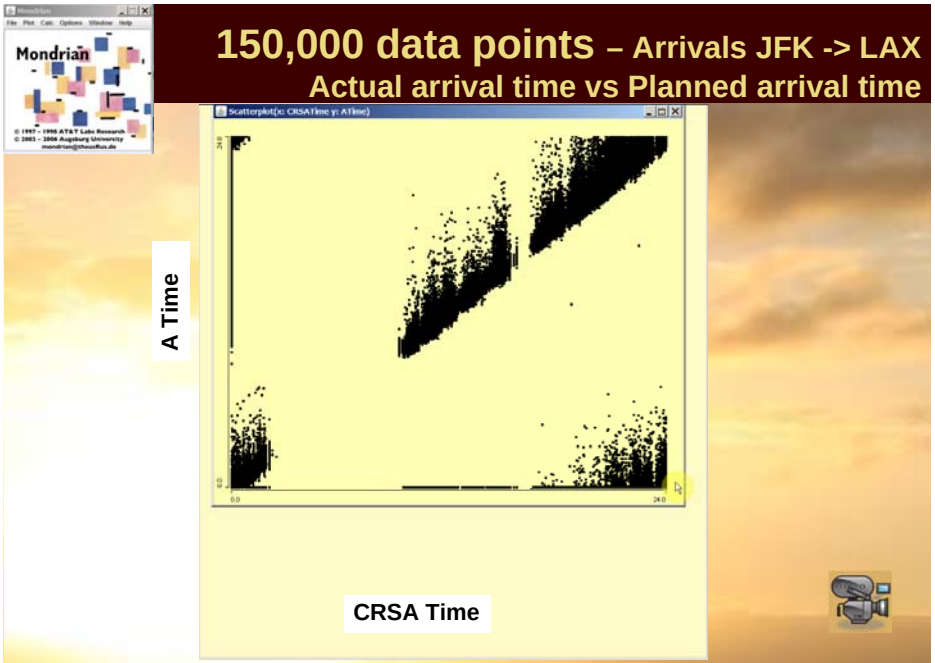
10,000 data points NHANES 2003-4



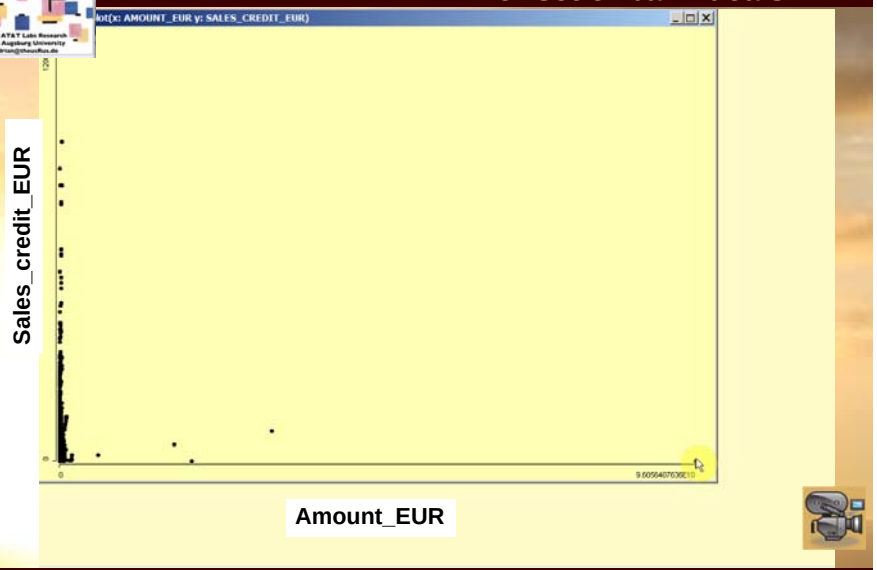
10,000 data points NHANES 2003-4



150,000 data points – Arrivals JFK -> LAX
Actual arrival time vs Planned arrival time

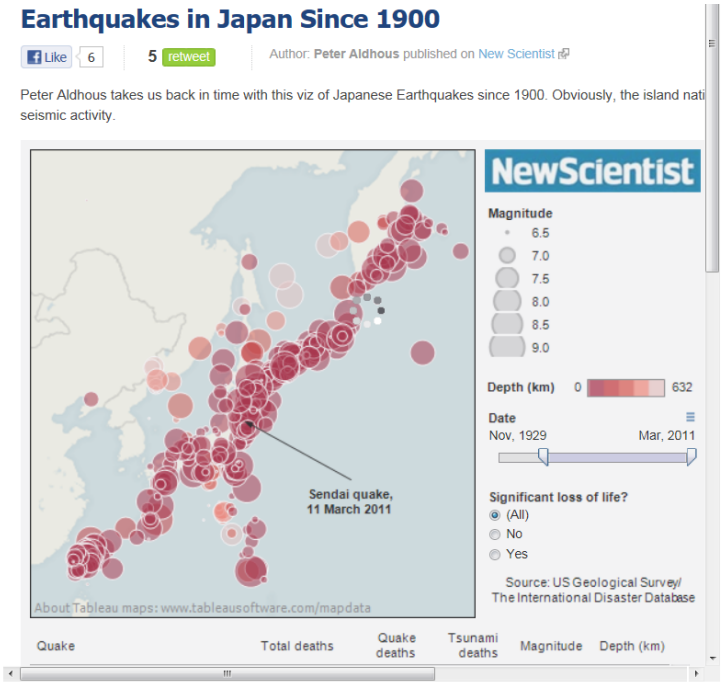


450,000 data points – Profit vs Amount for set of bank deals

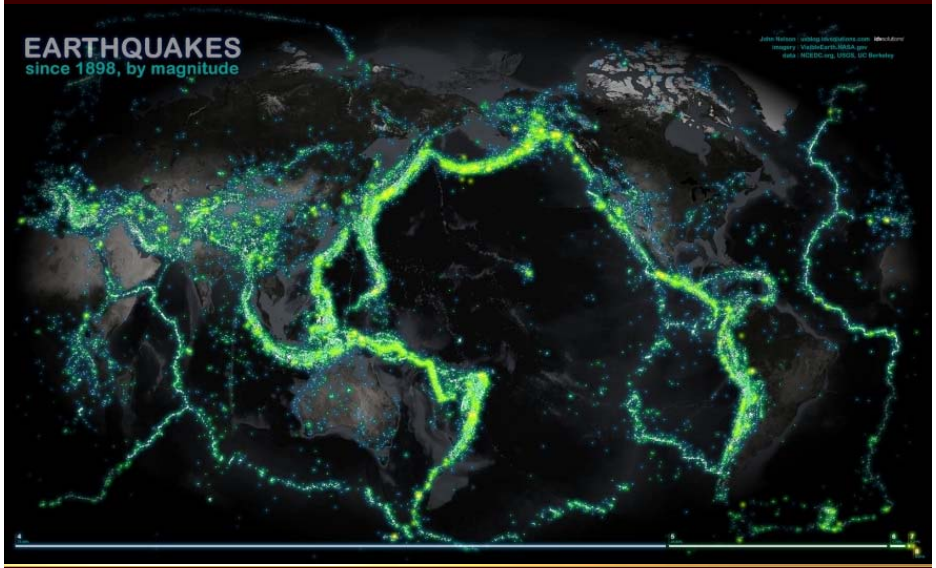


What have we used?

- The humble scatter plot with
  - Varying levels of transparency
  - Varying point size
  - Zooming
- Very simple ideas that help us see a lot
- Being able to vary them dynamically is important



Another scatterplot



Conclusions

Conclusions

- **The world is changing & so must we**
- **The data world is exploding**
  - **As is its impact on our lives**
  - Consequentially, the **wider populace** has to become better prepared to deal with it
  - **"*the democratisation of data*"**
    - **Ability *to understand*** as well as ***to access***

Conclusions

- **"*the democratisation of data*"**
  - **Ability *to understand*** as well as ***to access***
  - **& who** has the reach **to make the biggest impact** on the (widespread) ***ability to understand*?**
  - **So, who are the really key players here?**
  - **……**

Will change in statistics ever stop?

- We stand where **information technology meets human understanding**
- So the answer must be ….

  "Not if we are doing our jobs it won't"
- That said …
  - **we first need time …**
    - to get to grips with the changes already made
    - & to get the inevitable bugs out of our new systems
- Good news is that the new curriculum now allows …
  - for **on-going evolutionary advances**
  - rather than widely-spaced **step lurches**

And over time we can …

- Drag our attention from the day-to-day mire
- Lift up our eyes …

And find new & better ways to …

to open up the data world for our students

Thank you