Interactive applications written in R to accelerate statistical learning

Chris Wild

Department of Statistics
University of Auckland, New Zealand

https://www.stat.auckland.ac.nz/~wild/

What are we doing?

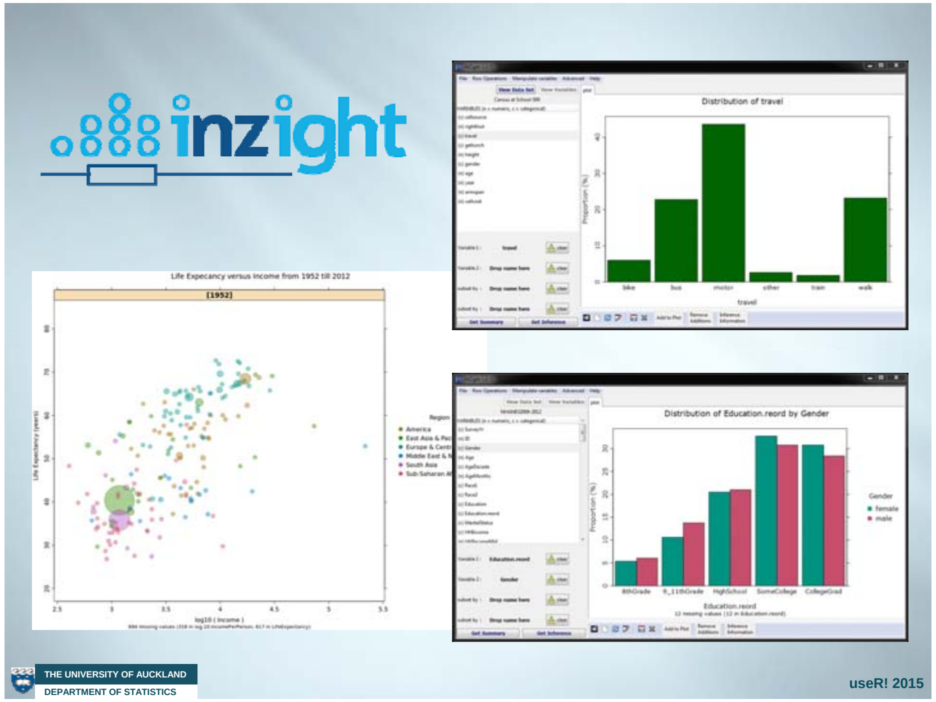
Applications

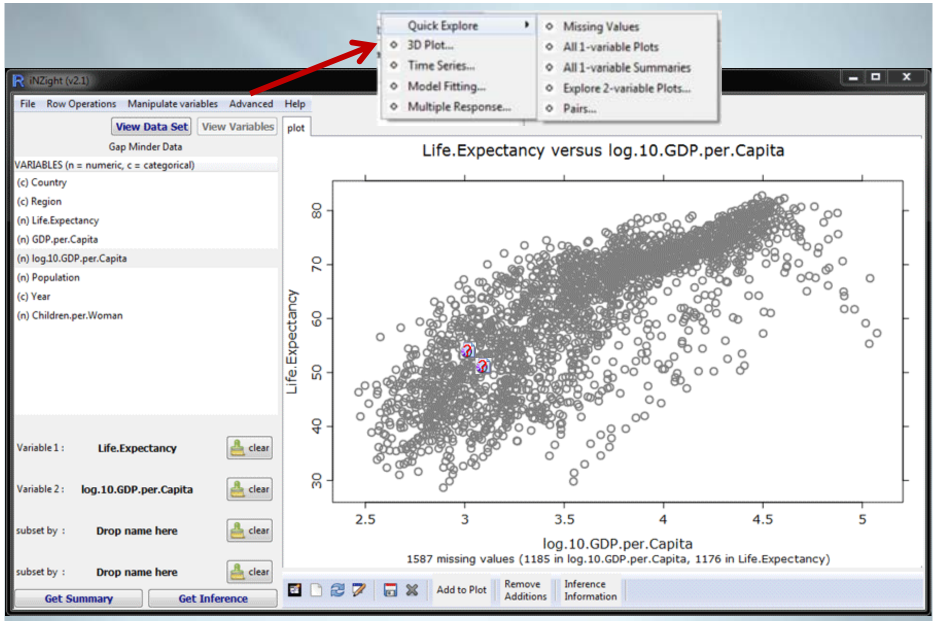
- iNZight
- iNZight Lite
- VIT: Visual Inference Tools
- Mortality Calculator
- Table Maker
- Others

inzight

- An interactive data analysis system that has R “unseen under the hood”

https://www.stat.auckland.ac.nz/~wild/iNZight/





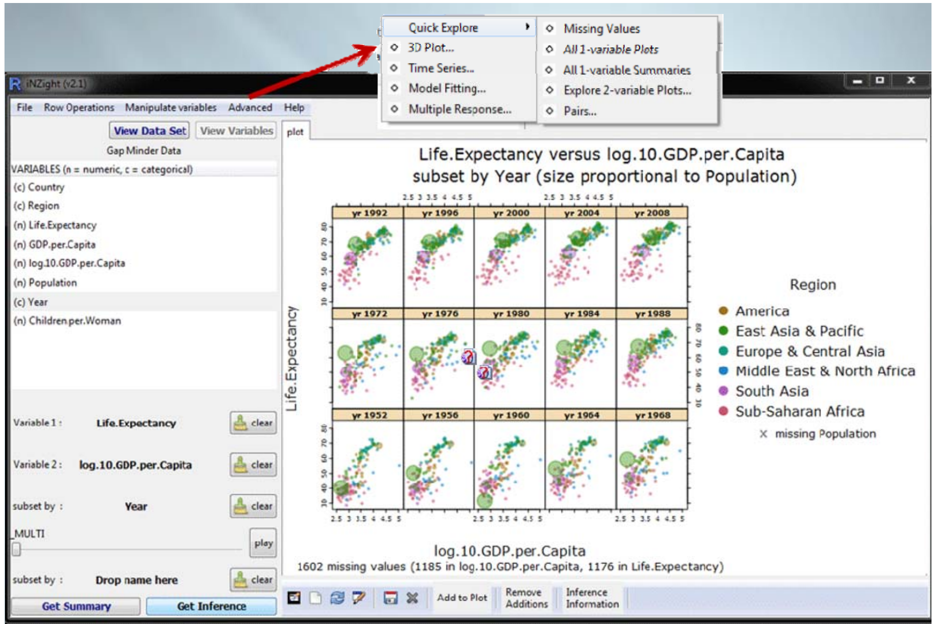
Variable 1 : Life.Expectancy

Variable 2 : log.10.GDP.per.Capita

subset by : Drop name here

subset by : Drop name here

1587 missing values (1185 in log.10.GDP.per.Capita, 1176 in Life.Expectancy)



Variable 1 : Life.Expectancy

Variable 2 : log.10.GDP.per.Capita

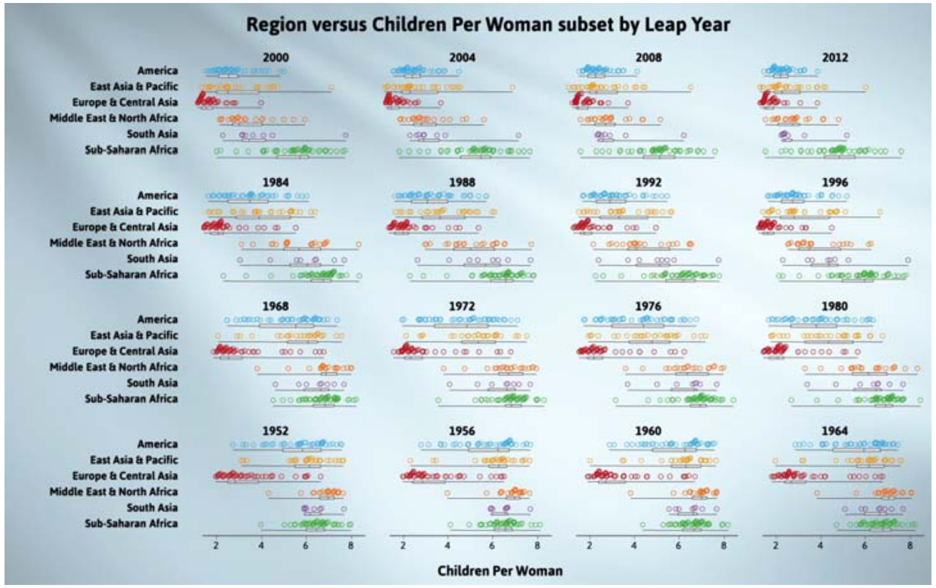
subset by : Year

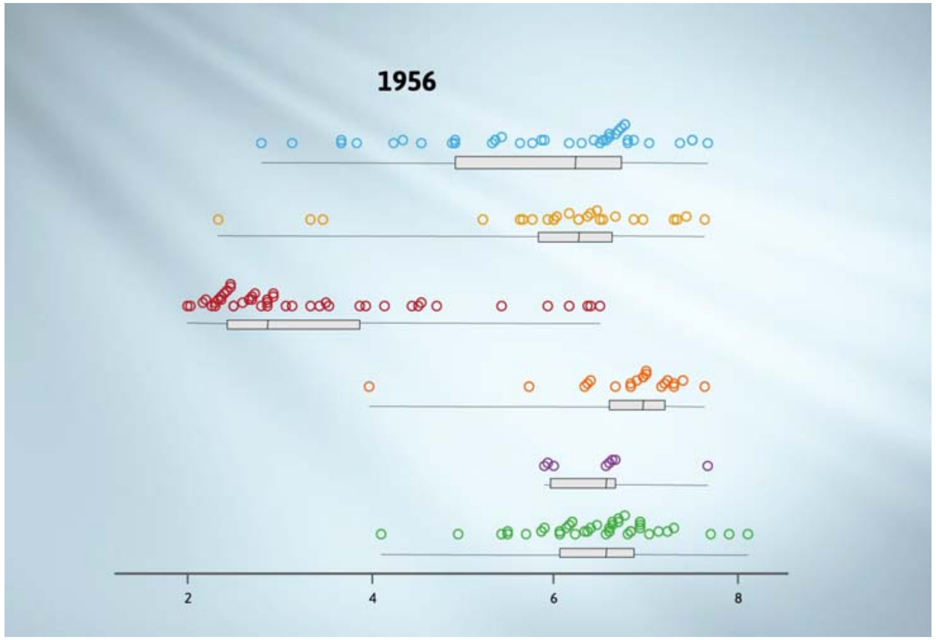
_MULTI

subset by : Drop name here

1602 missing values (1185 in log.10.GDP.per.Capita, 1176 in Life.Expectancy)

https://www.stat.auckland.ac.nz/~wild/iNZight/user_guides/interface

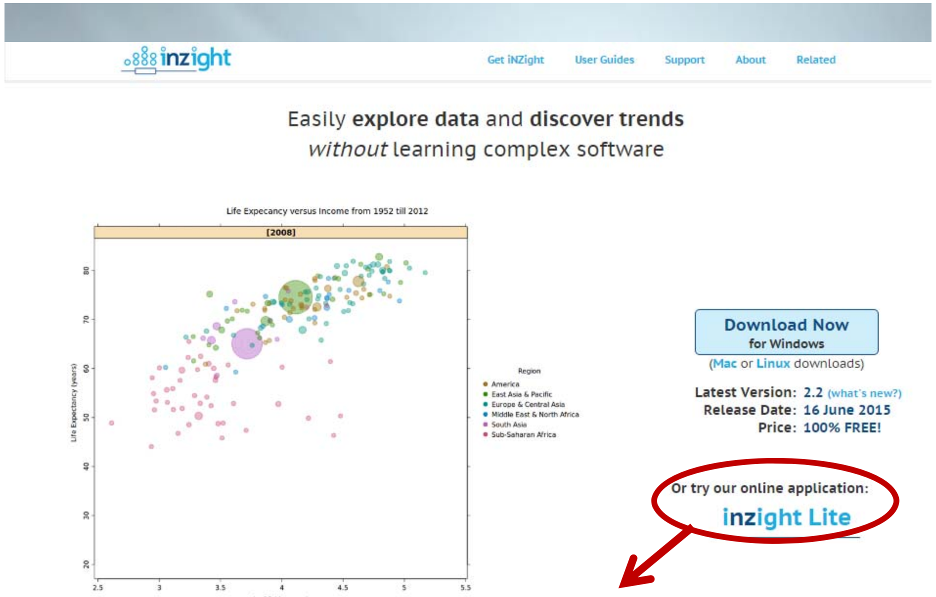




(playing across time)

inzight

- An interactive data analysis system that has R unseen “under the hood”
- Interactivity from John Verzani’s gwidgets2
  - High-level uses RGtk2 which uses Gtk+
- It has R inside it but users do not see R
  - Windows version also packages Gtk+ inside it
- Caters for beginners through to quite advanced modelling
  - Youngest users in early levels of high school



Easily **explore data** and **discover trends**
without learning complex software

(Mac or Linux downloads)

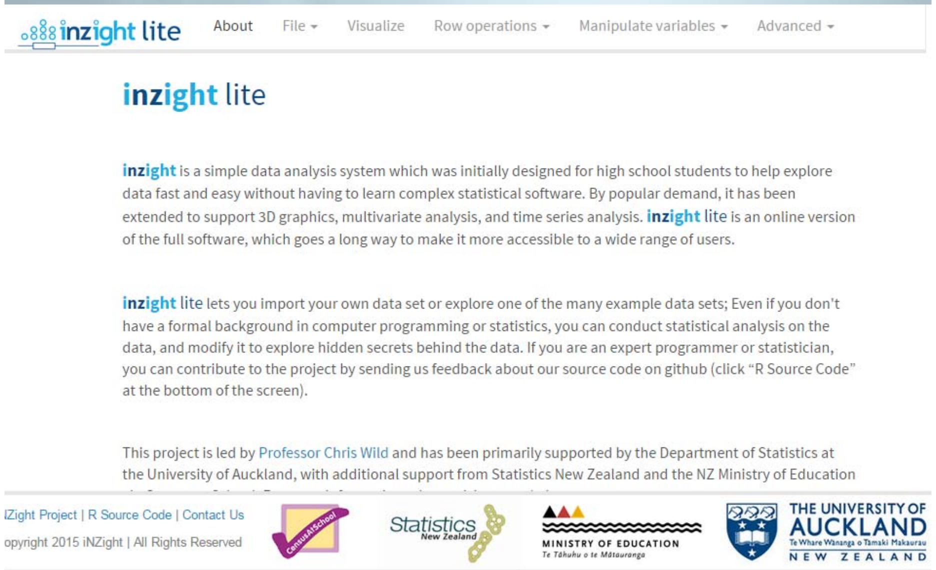
Latest Version: 2.2 (what's new?)
Release Date: 16 June 2015
Price: 100% FREE!

Or try our online application:
inzight Lite

Why??

BYOD

("Bring your own device"-driven need for tools to work on tablets and even phones)



inzight lite

inzight is a simple data analysis system which was initially designed for high school students to help explore data fast and easy without having to learn complex statistical software. By popular demand, it has been extended to support 3D graphics, multivariate analysis, and time series analysis. inzight lite is an online version of the full software, which goes a long way to make it more accessible to a wide range of users.

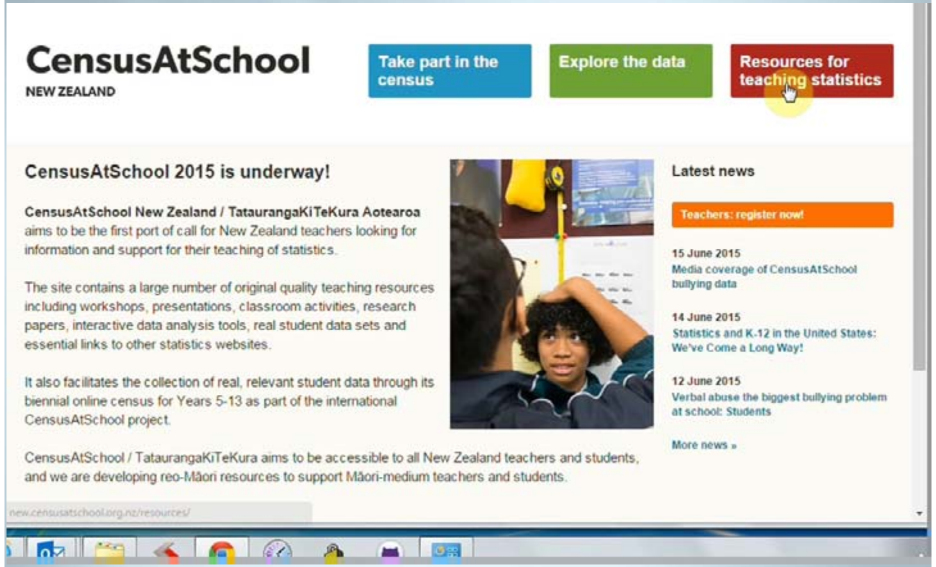
inzight lite lets you import your own data set or explore one of the many example data sets; Even if you don't have a formal background in computer programming or statistics, you can conduct statistical analysis on the data, and modify it to explore hidden secrets behind the data. If you are an expert programmer or statistician, you can contribute to the project by sending us feedback about our source code on github (click “R Source Code” at the bottom of the screen).

This project is led by Professor Chris Wild and has been primarily supported by the Department of Statistics at the University of Auckland, with additional support from Statistics New Zealand and the NZ Ministry of Education

iZight Project | R Source Code | Contact Us
opyright 2015 iNZight | All Rights Reserved

CensusAtSchool
Statistics New Zealand
MINISTRY OF EDUCATION
Te Tāhuhu o te Mātauranga
THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

http://new.censusatschool.org.nz/
(demo in which CaS data tools call up iNZight Lite for analysis)



CensusAtSchool
NEW ZEALAND

CensusAtSchool 2015 is underway!

CensusAtSchool New Zealand / TataurangaKiTeKura Aotearoa aims to be the first port of call for New Zealand teachers looking for information and support for their teaching of statistics.

The site contains a large number of original quality teaching resources including workshops, presentations, classroom activities, research papers, interactive data analysis tools, real student data sets and essential links to other statistics websites.

It also facilitates the collection of real, relevant student data through its biennial online census for Years 5-13 as part of the international CensusAtSchool project.

CensusAtSchool / TataurangaKiTeKura aims to be accessible to all New Zealand teachers and students, and we are developing reo-Māori resources to support Māori-medium teachers and students.

Latest news

Teachers: register now!

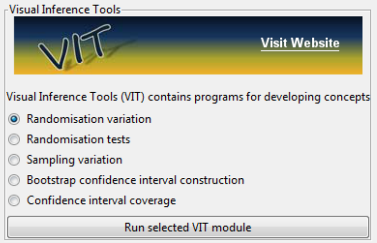
15 June 2015
Media coverage of CensusAtSchool bullying data

14 June 2015
Statistics and K-12 in the United States: We've Come a Long Way!

12 June 2015
Verbal abuse the biggest bullying problem at school: Students

More news »

new.censusatschool.org.nz/resources/



Visual Inference Tools

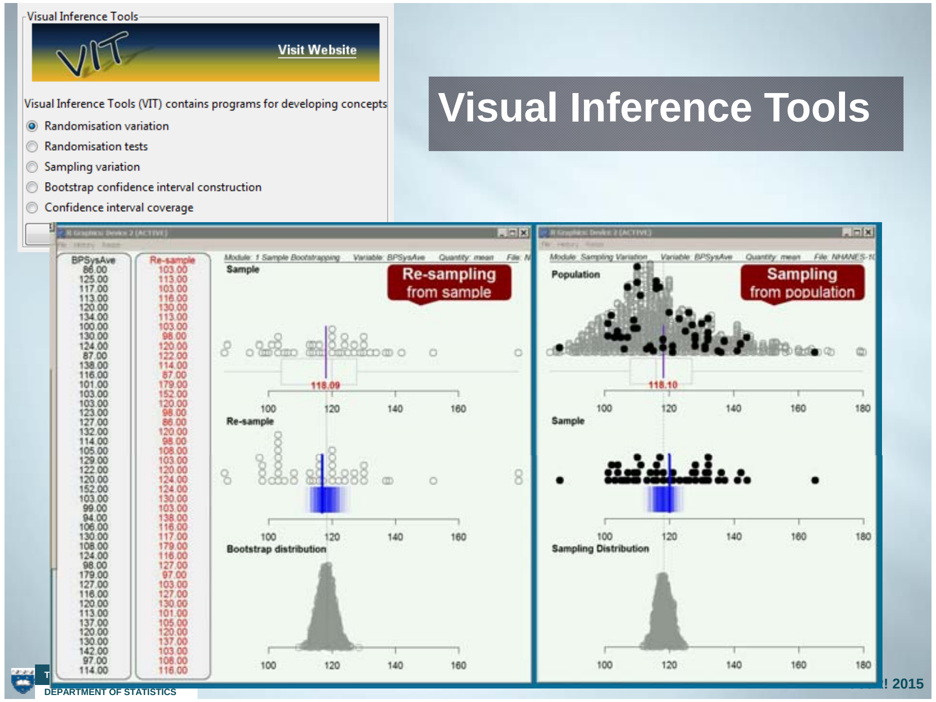
Visual Inference Tools (VIT) contains programs for developing concepts

- ● Randomisation variation
- ○ Randomisation tests
- ○ Sampling variation
- ○ Bootstrap confidence interval construction
- ○ Confidence interval coverage

Visual Inference Tools

- VIT **(Visual Inference Tools)** is about conceptual development

https://www.stat.auckland.ac.nz/~wild/VIT/



Visual Inference Tools

Visual Inference Tools (VIT) contains programs for developing concepts

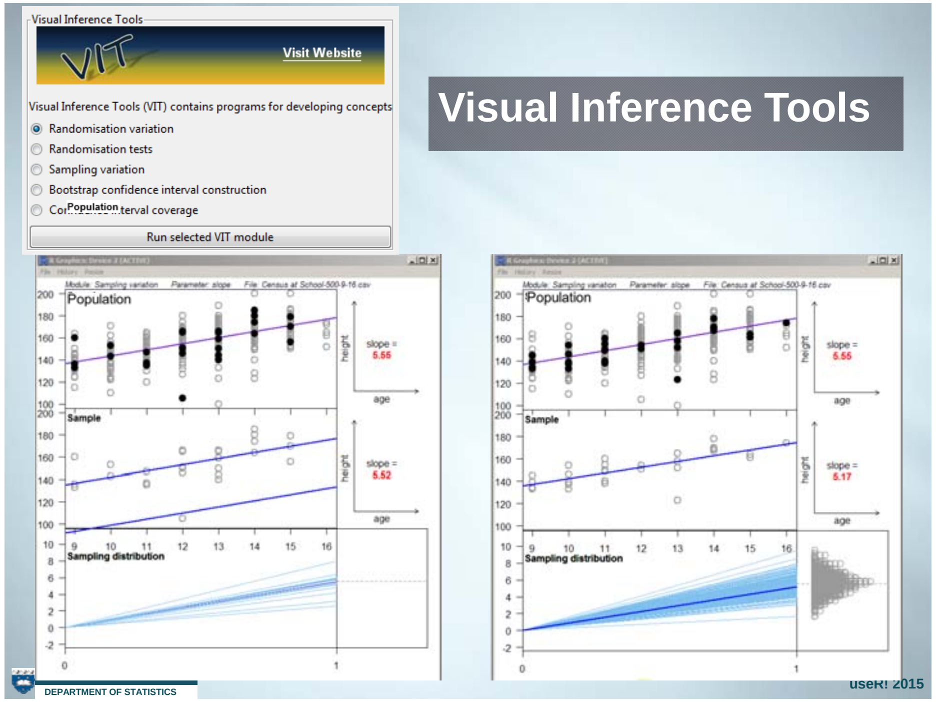
- ● Randomisation variation
- ○ Randomisation tests
- ○ Sampling variation
- ○ Bootstrap confidence interval construction
- ○ Confidence interval coverage

BPSysAve	Re-sample
86.00	103.00
125.00	113.00
117.00	103.00
113.00	116.00
120.00	130.00
134.00	113.00
100.00	103.00
130.00	98.00
124.00	120.00
87.00	122.00
138.00	114.00
116.00	87.00
101.00	179.00
103.00	152.00
103.00	120.00
123.00	98.00
127.00	86.00
132.00	120.00
114.00	98.00
105.00	108.00
129.00	103.00
122.00	120.00
120.00	124.00
152.00	124.00
103.00	130.00
99.00	103.00
94.00	138.00
106.00	116.00
130.00	117.00
108.00	179.00
124.00	116.00
98.00	127.00
179.00	97.00
127.00	103.00
116.00	127.00
120.00	130.00
113.00	101.00
137.00	105.00
120.00	120.00
130.00	137.00
142.00	103.00
97.00	108.00
114.00	116.00

Module: 1 Sample Bootstrapping Variable: BPSysAve Quantity: mean

Module: Sampling Variation Variable: BPSysAve Quantity: mean File: NHANES-1

Visual Inference Tools

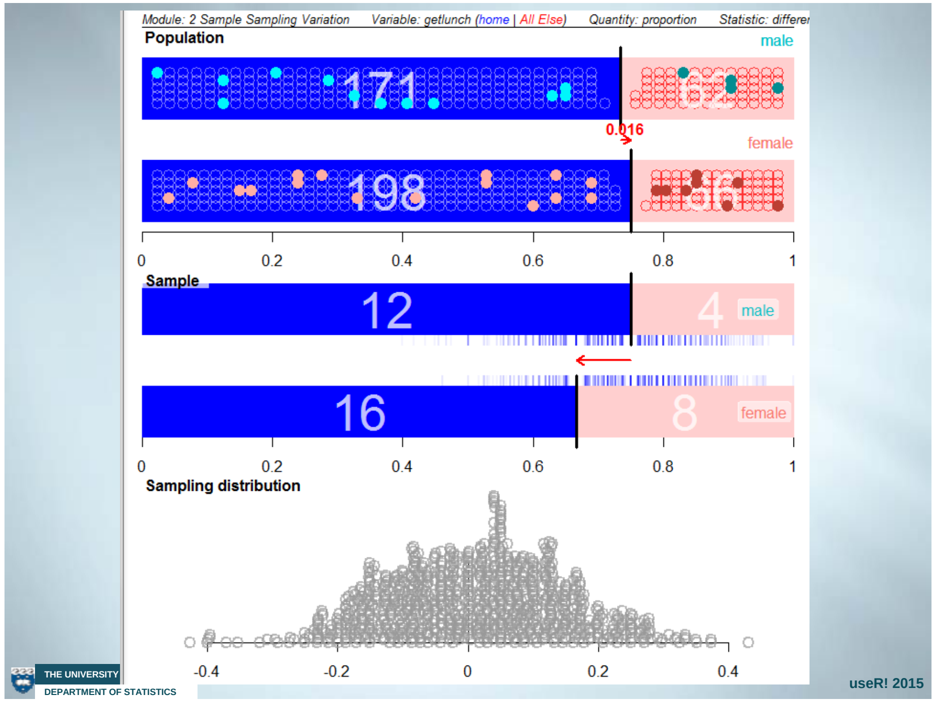


Visual Inference Tools

Visual Inference Tools (VIT) contains programs for developing concepts

- ● Randomisation variation
- ○ Randomisation tests
- ○ Sampling variation
- ○ Bootstrap confidence interval construction
- ○ Confidence interval coverage

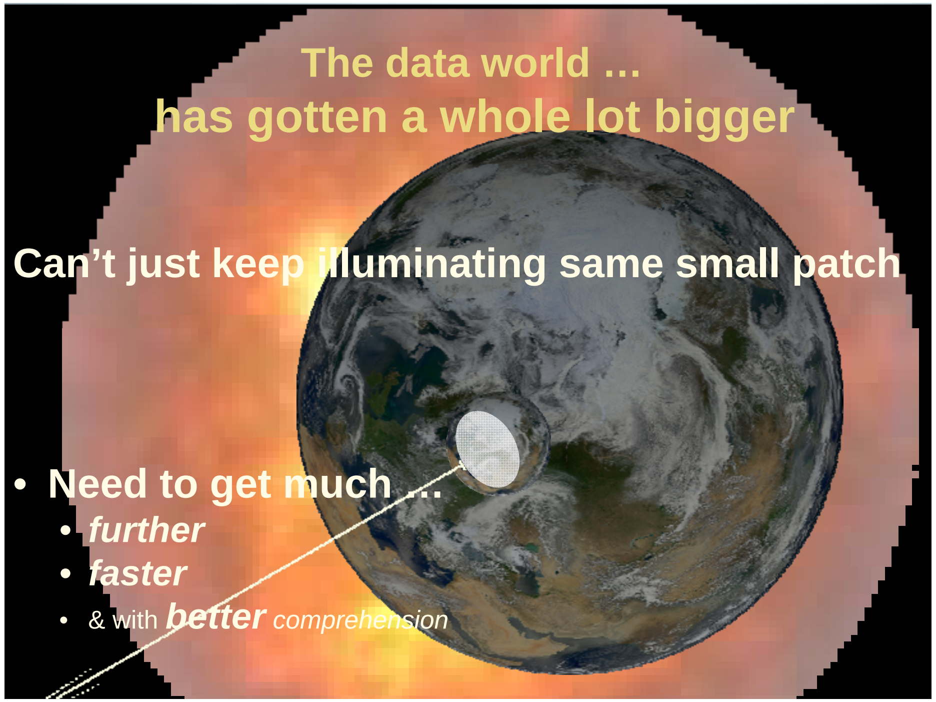
Module: Sampling variation Parameter: slope File: Census at School-500-9-16.csv



Module: 2 Sample Sampling Variation   Variable: getlunch (home | All Else)   Quantity: proportion   Statistic: differer

Why are we doing it?

The data world …
has gotten a whole lot bigger



The data world …
has gotten a whole lot bigger

Can’t just keep illuminating same small patch

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

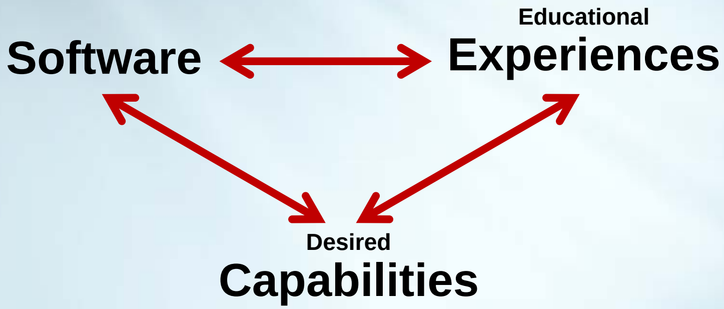
“Middleware”

(Not in the technical sense)

- software aimed at …
  - allowing student to experience
    - as much as possible of “discovery in the data world”
    - in the least possible time
  - Minimal learning curves, everything happens instantly & you don’t have to remember anything

Initial data analysis experiences should feel like this!

Developing in concert …



Educational Experiences

Software

Desired Capabilities

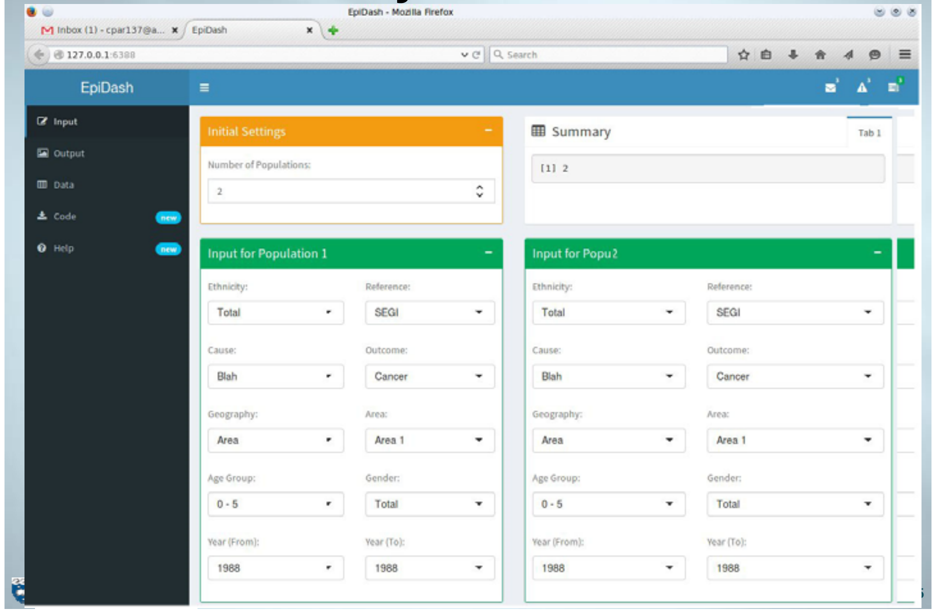
(Trade-offs everywhere)

How are we doing it?

How are we doing it?

- Sitting in behind iNZight and VIT are sets of R packages
    - (Can be run directly from R)
  - User interfaces use John Verzani's **gwidgets**
    - which uses Gtk+ via RGtk2
  - Each module has a "receiver function" which manages the mapping of user choices to R calls
- iNZight Lite is same with R Shiny user interface
  - Connecting to exactly the same receiver-functions

Mortality "Calculator"



EpiDash - Mozilla Firefox

Inbox (1) - cpar137@a... | EpiDash

127.0.0.1:6388

Initial Settings

Number of Populations: 2

Summary Tab 1

```
[1] 2
```

Input for Population 1

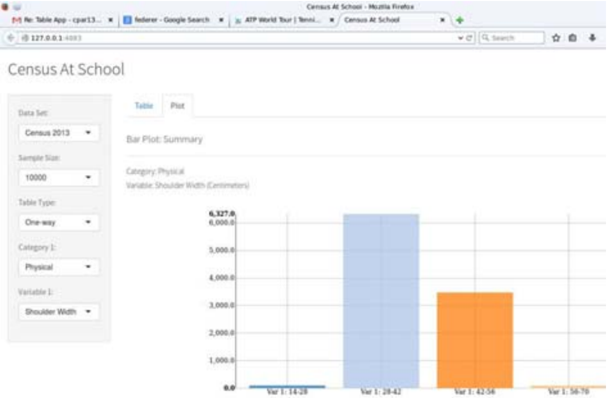
Ethnicity: Total
Reference: SEGI
Cause: Blah
Outcome: Cancer
Geography: Area
Area: Area 1
Age Group: 0 - 5
Gender: Total
Year (From): 1988
Year (To): 1988

Input for Popu2

Ethnicity: Total
Reference: SEGI
Cause: Blah
Outcome: Cancer
Geography: Area
Area: Area 1
Age Group: 0 - 5
Gender: Total
Year (From): 1988
Year (To): 1988

Census At School Table Maker

(1 & 2-way tables)



Census At School - Mozilla Firefox

Re: Table App - cpar13... | federer - Google Search | ATP World Tour | Tenni... | Census At School

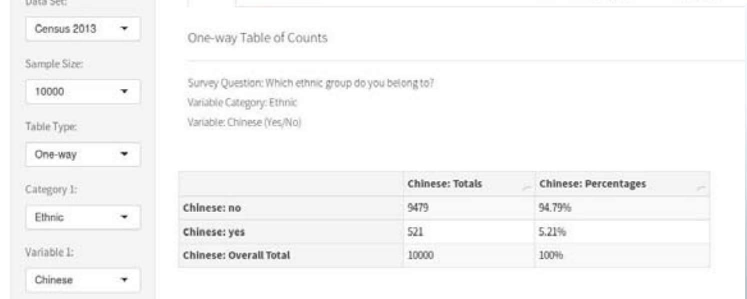
127.0.0.1:4083

Census At School

Data Set: Census 2013
Sample Size: 10000
Table Type: One-way
Category 1: Physical
Variable 1: Shoulder Width

Bar Plot: Summary

Category: Physical
Variable: Shoulder Width (Centimeters)



Re: Table App - cpar13... | Census At School

127.0.0.1:4624

Census At School

Data Set: Census 2013
Sample Size: 10000
Table Type: One-way
Category 1: Ethnic
Variable 1: Chinese

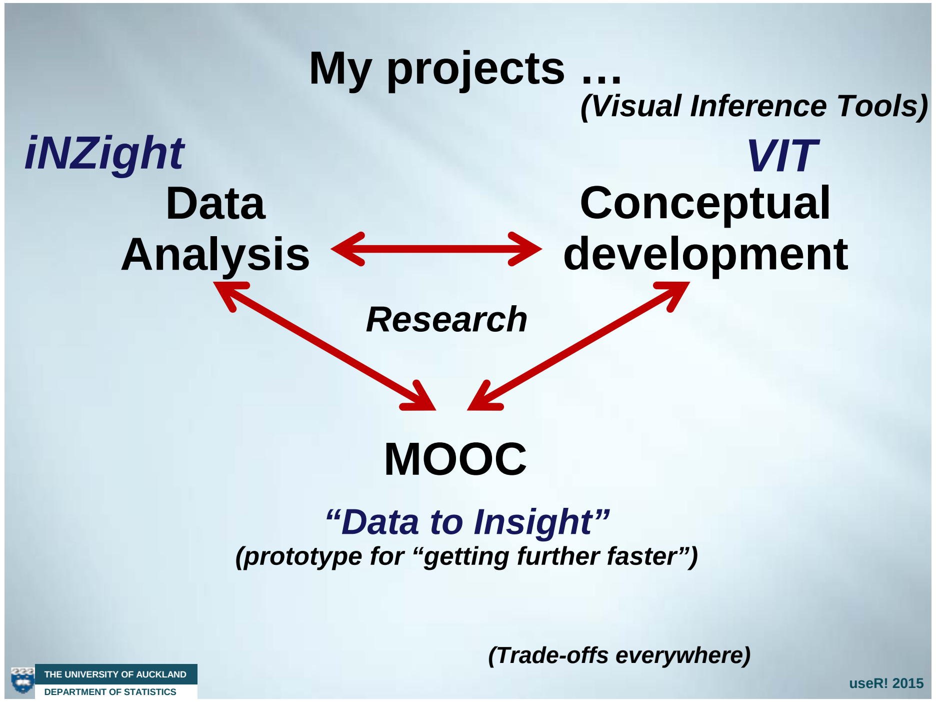
One-way Table of Counts

Survey Question: Which ethnic group do you belong to?
Variable Category: Ethnic
Variable: Chinese (Yes/No)

	Chinese: Totals	Chinese: Percentages
Chinese: no	9479	94.79%
Chinese: yes	521	5.21%
Chinese: Overall Total	10000	100%

Shiny applications

  - iNZight Lite (versions), mortality calculator, probability visualisations, …
- Each application
  - Is in its own docker container with R shiny server (the free one) with R and relevant libraries.
  - Docker container inside a virtual machine
  - Replicate virtual machines if needed to meet demand
- We want others to be able …
  - just pick these things up and put down on own servers with a minimum of effort and knowledge
  - **Volunteers anyone???**



My projects …

iNZight

Data Analysis

(Visual Inference Tools)

VIT

Conceptual development

Research

MOOC

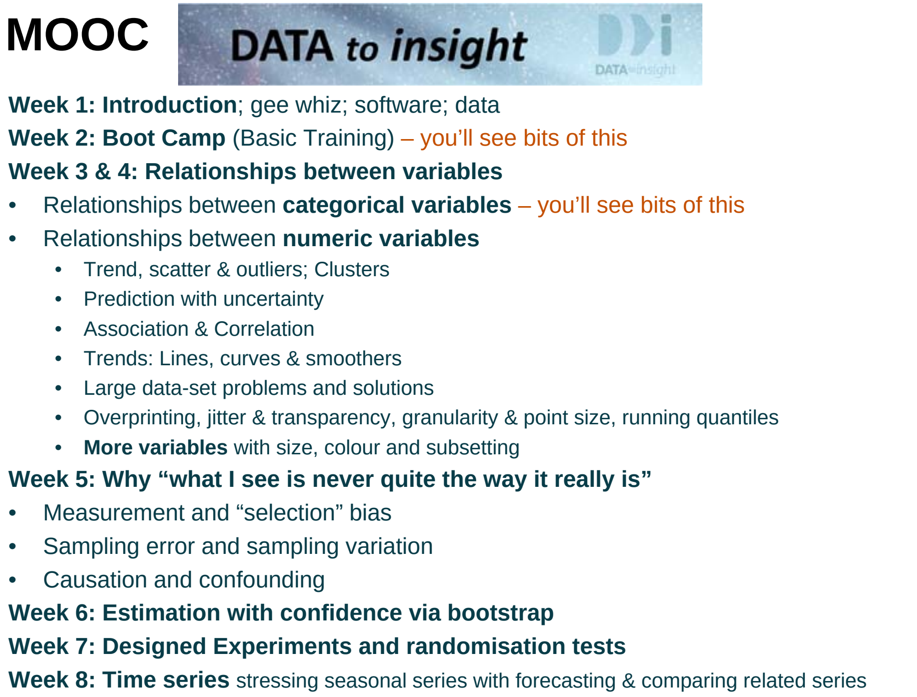
"Data to Insight"

(prototype for "getting further faster")

(Trade-offs everywhere)

THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS

useR! 2015



MOOC DATA to insight

Week 1: Introduction; gee whiz; software; data

Week 2: Boot Camp (Basic Training) – you'll see bits of this

Week 3 & 4: Relationships between variables

- Relationships between **categorical variables** – you'll see bits of this
- Relationships between **numeric variables**
  - Trend, scatter & outliers; Clusters
  - Prediction with uncertainty
  - Association & Correlation
  - Trends: Lines, curves & smoothers
  - Large data-set problems and solutions
  - Overprinting, jitter & transparency, granularity & point size, running quantiles
  - **More variables** with size, colour and subsetting

Week 5: Why "what I see is never quite the way it really is"

- Measurement and "selection" bias
- Sampling error and sampling variation
- Causation and confounding

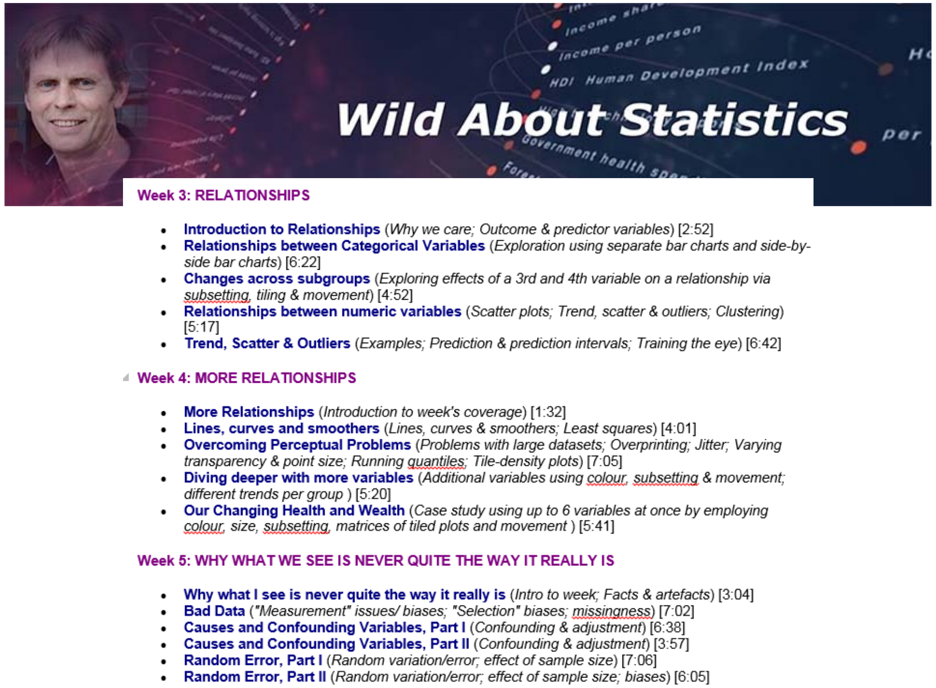
Week 6: Estimation with confidence via bootstrap

Week 7: Designed Experiments and randomisation tests

Week 8: Time series stressing seasonal series with forecasting & comparing related series

https://www.stat.auckland.ac.nz/~wild/d2i/4StatEducators/

https://www.stat.auckland.ac.nz/~wild/wildaboutstatistics/ (index to youTube channel)



Week 3: RELATIONSHIPS

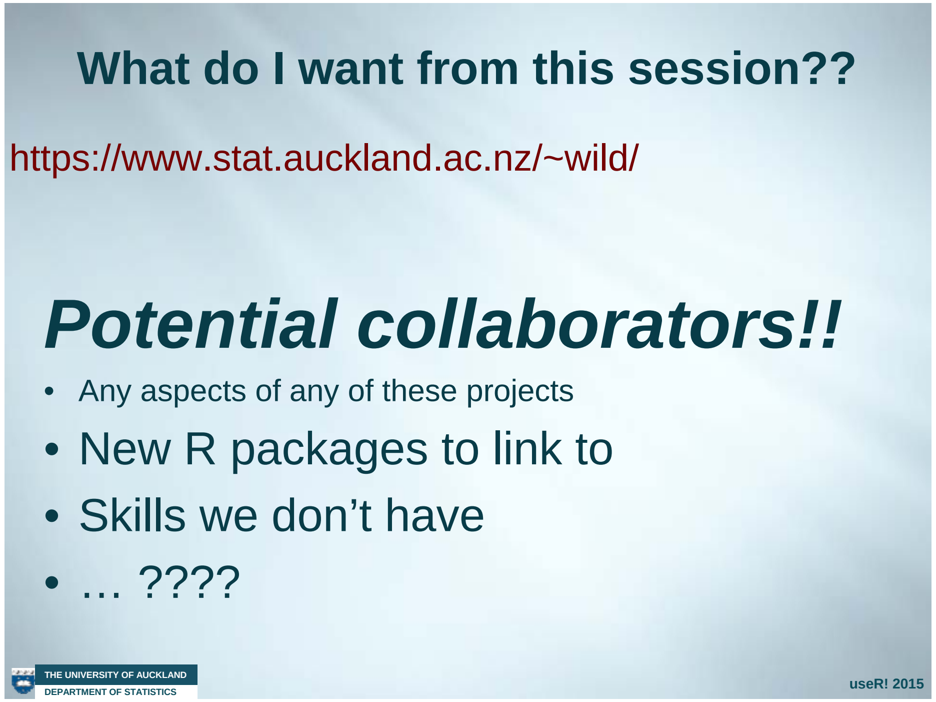
- **Introduction to Relationships** (*Why we care; Outcome & predictor variables*) [2:52]
- **Relationships between Categorical Variables** (*Exploration using separate bar charts and side-by-side bar charts*) [6:22]
- **Changes across subgroups** (*Exploring effects of a 3rd and 4th variable on a relationship via subsetting, tiling & movement*) [4:52]
- **Relationships between numeric variables** (*Scatter plots; Trend, scatter & outliers; Clustering*) [5:17]
- **Trend, Scatter & Outliers** (*Examples; Prediction & prediction intervals; Training the eye*) [6:42]

Week 4: MORE RELATIONSHIPS

- **More Relationships** (*Introduction to week's coverage*) [1:32]
- **Lines, curves and smoothers** (*Lines, curves & smoothers; Least squares*) [4:01]
- **Overcoming Perceptual Problems** (*Problems with large datasets; Overprinting; Jitter; Varying transparency & point size; Running quantiles; Tile-density plots*) [7:05]
- **Diving deeper with more variables** (*Additional variables using colour, subsetting & movement; different trends per group*) [5:20]
- **Our Changing Health and Wealth** (*Case study using up to 6 variables at once by employing colour, size, subsetting, matrices of tiled plots and movement*) [5:41]

Week 5: WHY WHAT WE SEE IS NEVER QUITE THE WAY IT REALLY IS

- **Why what I see is never quite the way it really is** (*Intro to week; Facts & artefacts*) [3:04]
- **Bad Data** (*"Measurement" issues/ biases; "Selection" biases; missingness*) [7:02]
- **Causes and Confounding Variables, Part I** (*Confounding & adjustment*) [6:38]
- **Causes and Confounding Variables, Part II** (*Confounding & adjustment*) [3:57]
- **Random Error, Part I** (*Random variation/error; effect of sample size*) [7:06]
- **Random Error, Part II** (*Random variation/error; effect of sample size; biases*) [6:05]



What do I want from this session??

https://www.stat.auckland.ac.nz/~wild/

Potential collaborators!!

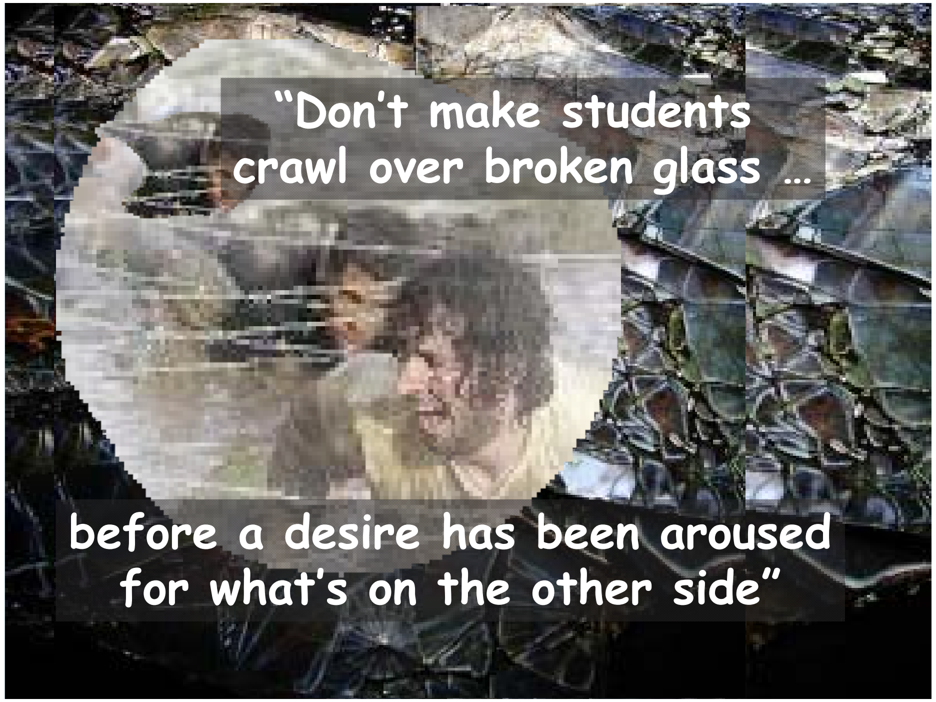
- Any aspects of any of these projects
- New R packages to link to
- Skills we don't have
- … ????

THE UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS

useR! 2015

Initial data analysis experiences should feel like this!

Thank you



"Don't make students crawl over broken glass …

before a desire has been aroused for what's on the other side"