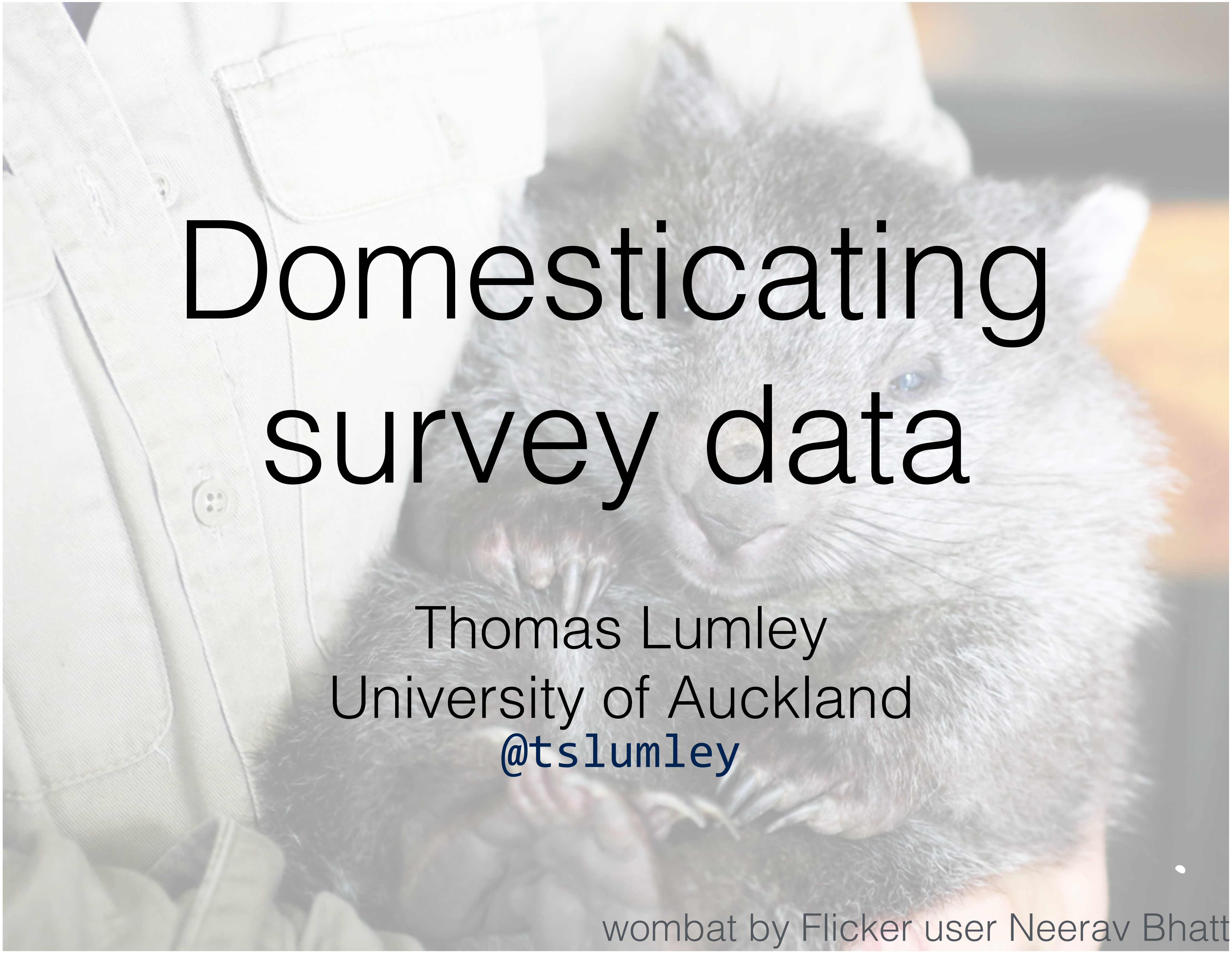
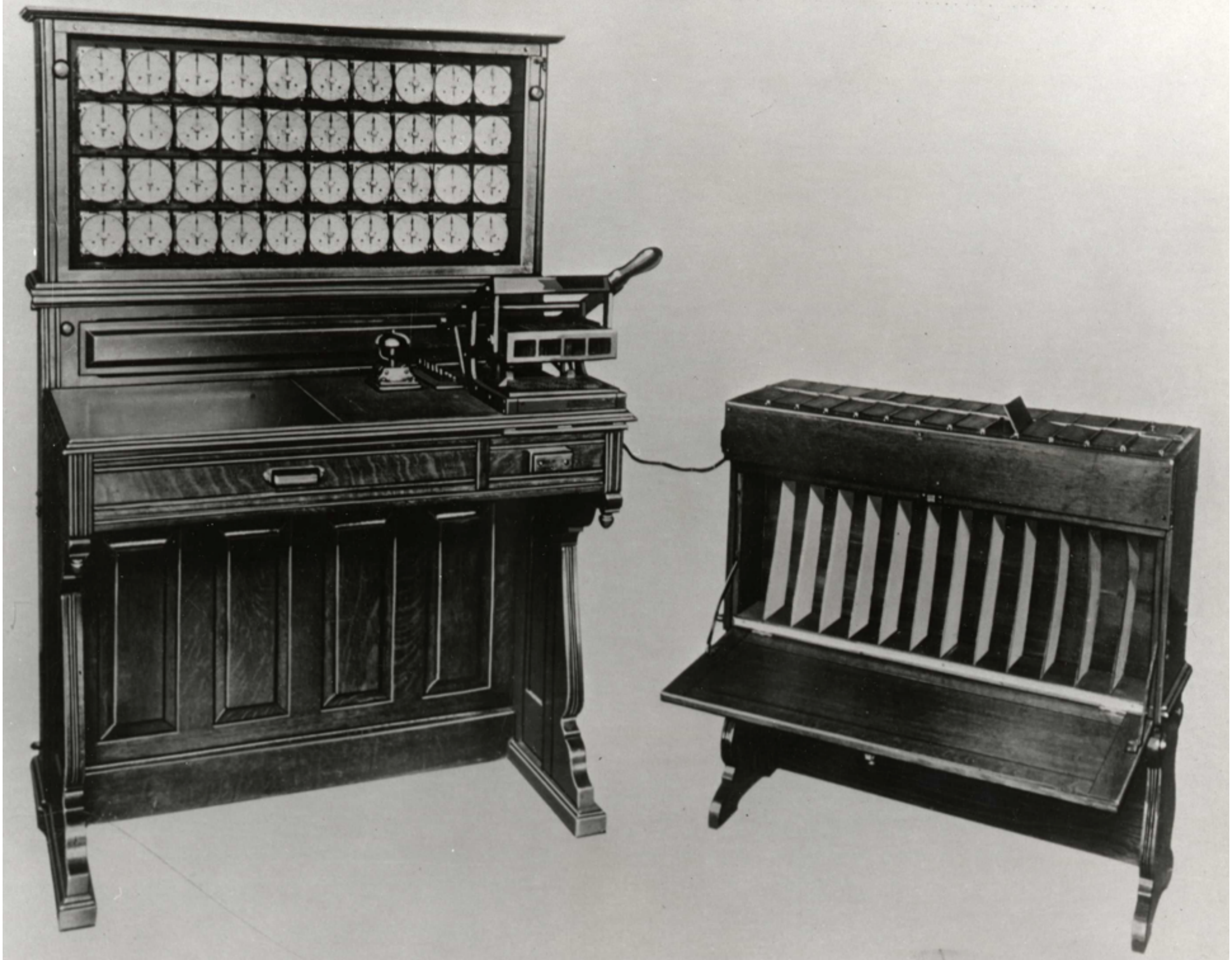
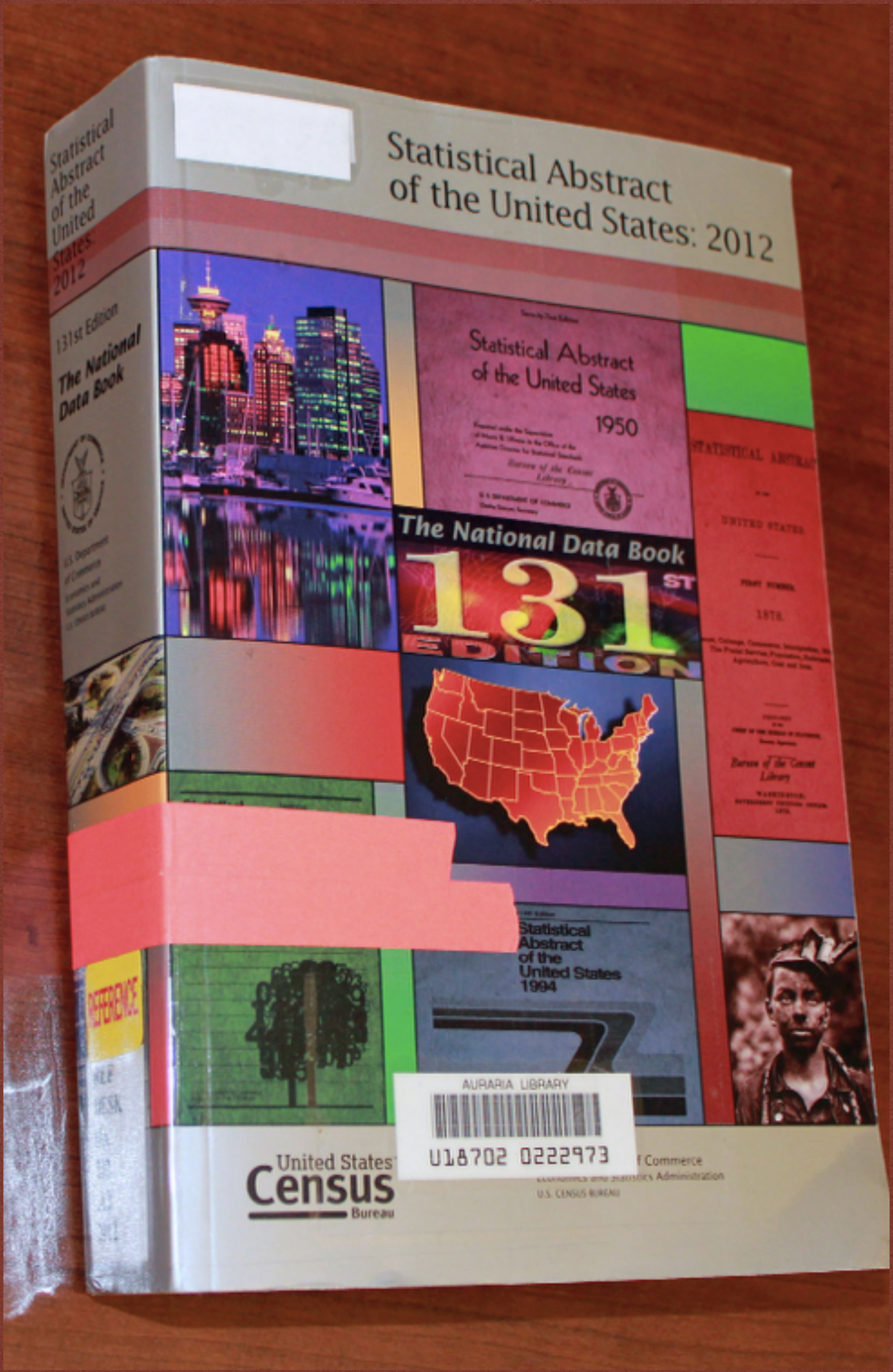
# Domesticating survey data

Thomas Lumley
University of Auckland
@tslumley



wombat by Flicker user Neerav Bhatt





Statistical Abstract of the United States: 2012

Statistical Abstract of the United States 2012

131st Edition

The National Data Book

The National Data Book 131ST EDITION

United States Census Bureau

...f Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

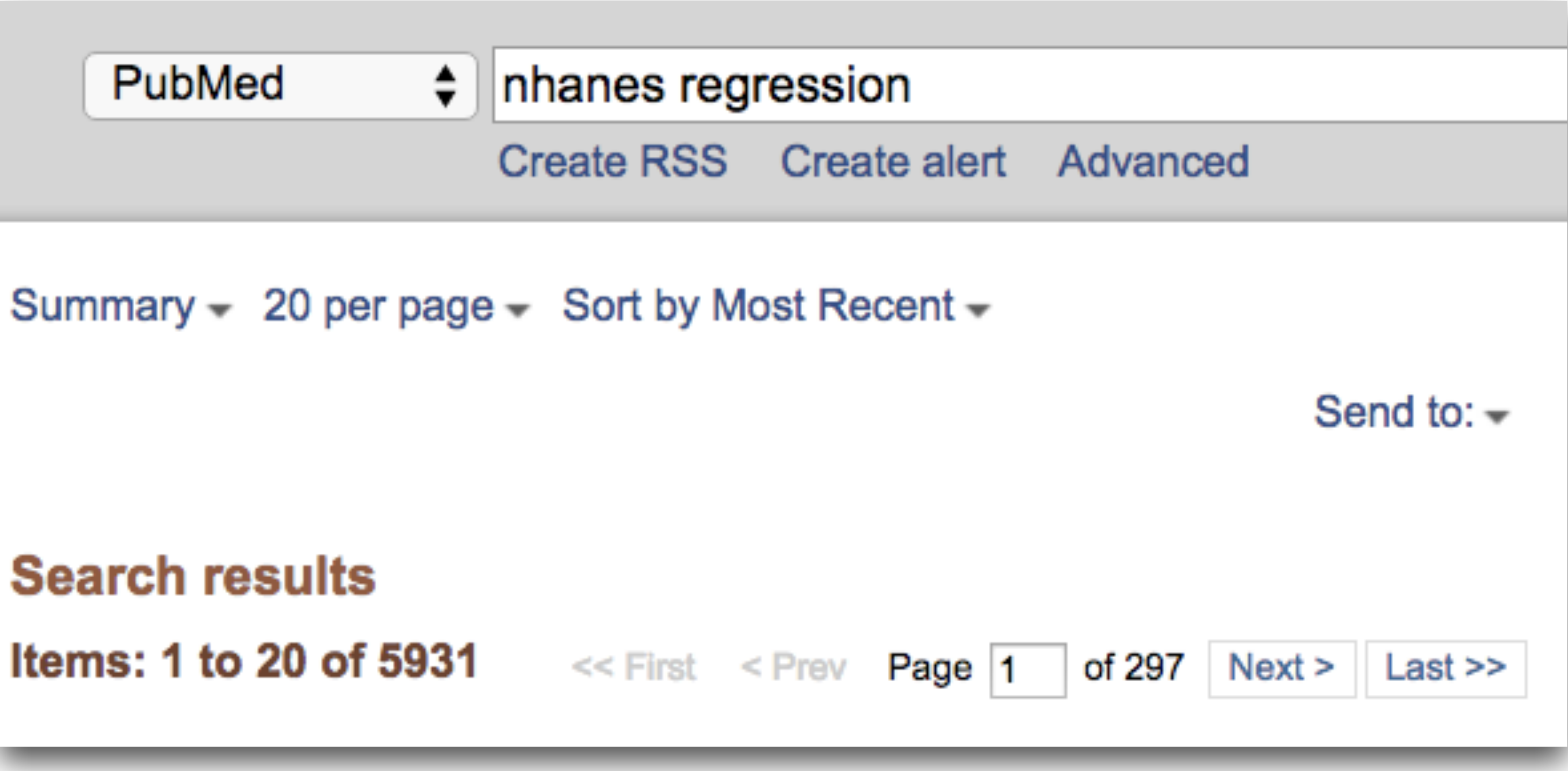
Statistical Abstract of the United States 1950
Bureau of the Census Library
STATISTICAL ABSTRACT
UNITED STATES
FIRST NUMBER
1878
Bureau of the Census Library
Statistical Abstract of the United States 1994

## Table 1159. Internet Activities of Adults by Geographic Community Type: 2011

[In percent. For Internet users 18 years old and over. Represents persons who have ever performed the activity. Based on telephone surveys of persons with land-line telephones and cell phones. See headnote, Table 1160]

| Activity | Survey date (month, year) | Total adults | Internet users performing activity | | | |
|---|---|---|---|---|---|---|
| | | | Total | Urban | Suburban | Rural |
| Buy a product online | May, 2011 | 55 | 71 | 73 | 72 | 70 |
| Buy or make a reservation for travel | May, 2011 | 51 | 65 | 66 | 66 | 60 |
| Categorize or tag online content like a photo, news story or blog post | Sept, 2010 | 24 | 33 | 37 | 32 | 25 |
| Create or work on your own online journal or blog | May, 2011 | 11 | 14 | 16 | 13 | 11 |
| Do any banking online | May, 2011 | 47 | 61 | 68 | 60 | 50 |
| Look for health or medical information online | May, 2011 | 55 | 71 | 72 | 69 | 81 |
| Look for news or information about politics | May, 2011 | 47 | 61 | 64 | 61 | 48 |
| Look online for info about a job | May, 2011 | 44 | 56 | 63 | 56 | 45 |
| Make a donation to a charity online | May, 2011 | 19 | 25 | 31 | 26 | 15 |
| Make a phone call online, using a service such as Skype or Vonage | May, 2011 | 18 | 24 | 25 | 27 | 13 |
| Pay bills online | Sept, 2010 | 42 | 57 | 55 | 62 | 45 |
| Pay to access or download digital content online (e.g. newspaper article) | Sept, 2010 | 32 | 43 | 47 | 43 | 35 |
| Play online games | Sept, 2010 | 27 | 36 | 36 | 38 | 34 |
| Post a comment or review online | Sept, 2010 | 24 | 32 | 34 | 35 | 24 |
| Research a product or service online | Sept, 2010 | 58 | 78 | 79 | 79 | 77 |
| Search online for a map or driving directions | Sept, 2010 | 60 | 82 | 84 | 83 | 79 |
| Send instant messages | Nov, 2010 | 34 | 46 | 49 | 47 | 42 |
| Send or read e-mail | Nov, 2010 | 68 | 92 | 93 | 93 | 90 |
| Take part in chat rooms or online discussions with other people | Sept, 2010 | 17 | 22 | 25 | 21 | 20 |
| Use a search engine to find information | May, 2011 | 71 | 92 | 90 | 93 | 89 |
| Use a social networking site like MySpace, Facebook or LinkedIn | May, 2011 | 50 | 65 | 67 | 65 | 61 |
| Use Twitter | May, 2011 | 10 | 13 | 15 | 14 | 7 |
| Visit a local, state, or federal government Web site | May, 2011 | 52 | 67 | 68 | 69 | 61 |
| Watch a video on a video-sharing site | May, 2011 | 55 | 71 | 72 | 71 | 68 |

Source: Pew Internet & American Life Project Surveys, <http://www.pewinternet.org>.



PubMed: nhanes regression

Summary · 20 per page · Sort by Most Recent

Search results

Items: 1 to 20 of 5931

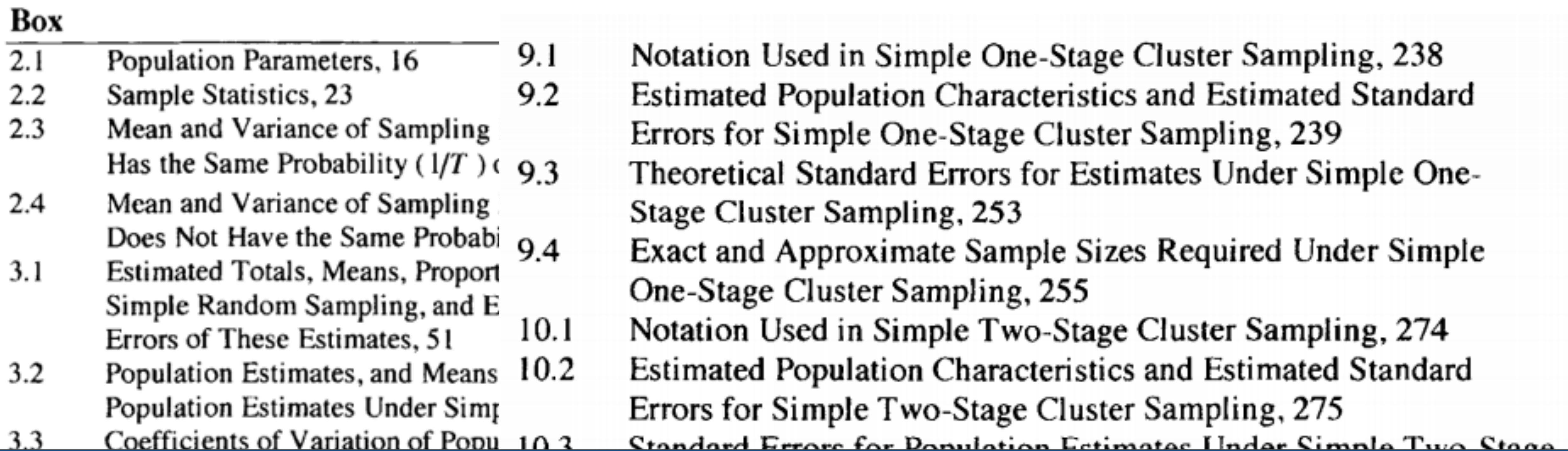
Page 1 of 297

1970s: 1-2/year

Now: ~1/day

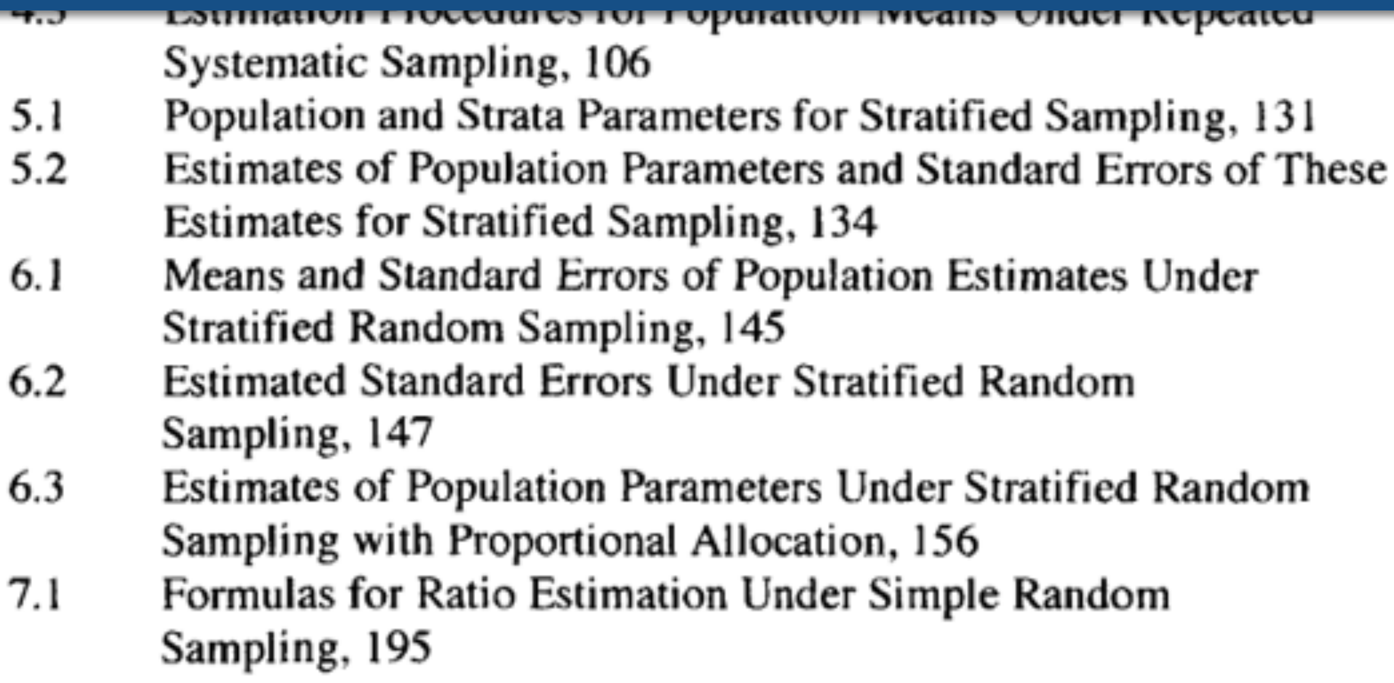
**Box**

**Box**



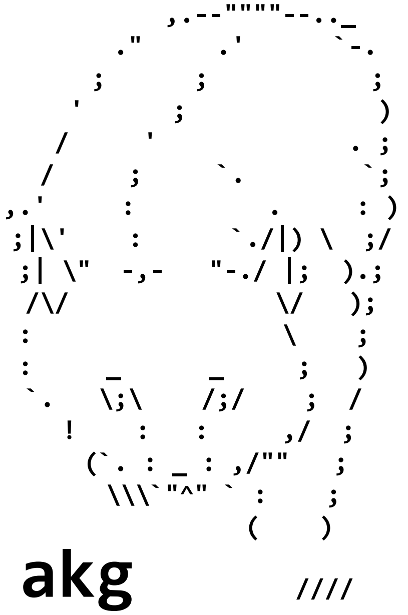
Box

…continues to provide a highly readable, practical treatment of the subject. Keeping mathematics to a minimum,…



“Often when an architecture deviates from a sane general design in some of its details that's because it's a bad design. So the same principles that make you write around the design specifics to achieve portability also make you write around the bad design features and stick to a more optimized general design.”

*– Linus Torvalds*



# Abstraction: data objects

- **Clusters**: what units did you sample?
- **Strata**: in what ways did you *force* the sampling to be representative.
- **Weights**: how many people in the population does this person represent?
- **Subsets**: can't just drop rows
- **Calibration**: what population information can we use to reduce bias and variance

```
des<-svydesign(id=~SDMVPSU, strat=~SDMVSTRA,
      weights=~fouryearwt, nest=TRUE,
      data=subset(nhanes, !is.na(WTDRD1)))

svyplot(BPXDAR~RIDAGEYR,style=“hex",design=des,
  legend=0,xlab="Age (yrs)”,
  ylab="Diastolic BP (mmHg)”)

des<-transform(des, age1=pmin(RIDAGEYR,50)/10,
   age2=pmin(pmax(RIDAGEYR,50),65)/10,
   age3=pmin(pmax(RIDAGEYR,65),90)/10)

ish3s<- svyglm(
    ish~(age1+age2+age3)*RIAGENDR+factor(RIDRETH1),
    design=des,family=quasibinomial)
anova(ish3s)
AIC(ish0s,ish1s,ish2s,ish3s)
```

```
des<-svydesign(id=~SDMVPSU, strat=~SDMVSTRA,
      weights=~fouryearwt, nest=TRUE,
      data=subset(nhanes, !is.na(WTDRD1)))

svyplot(BPXDAR~RIDAGEYR,style=“hex",design=des,
  legend=0,xlab="Age (yrs)”,
  ylab="Diastolic BP (mmHg)”)

des<-transform(des, age1=pmin(RIDAGEYR,50)/10,
   age2=pmin(pmax(RIDAGEYR,50),65)/10,
   age3=pmin(pmax(RIDAGEYR,65),90)/10)

ish3s<- svyglm(
    ish~(age1+age2+age3)*RIAGENDR+factor(RIDRETH1),
    design=des,family=quasibinomial)
anova(ish3s)
AIC(ish0s,ish1s,ish2s,ish3s)
```

```
des<-svydesign(id=~SDMVPSU, strat=~SDMVSTRA,
      weights=~fouryearwt, nest=TRUE,
      data=subset(nhanes, !is.na(WTDRD1)))

svyplot(BPXDAR~RIDAGEYR,style=“hex",design=des,
  legend=0,xlab="Age (yrs)”,
  ylab="Diastolic BP (mmHg)”)

des<-transform(des, age1=pmin(RIDAGEYR,50)/10,
   age2=pmin(pmax(RIDAGEYR,50),65)/10,
   age3=pmin(pmax(RIDAGEYR,65),90)/10)

ish3s<- svyglm(
    ish~(age1+age2+age3)*RIAGENDR+factor(RIDRETH1),
    design=des,family=quasibinomial)
anova(ish3s)
AIC(ish0s,ish1s,ish2s,ish3s)
```

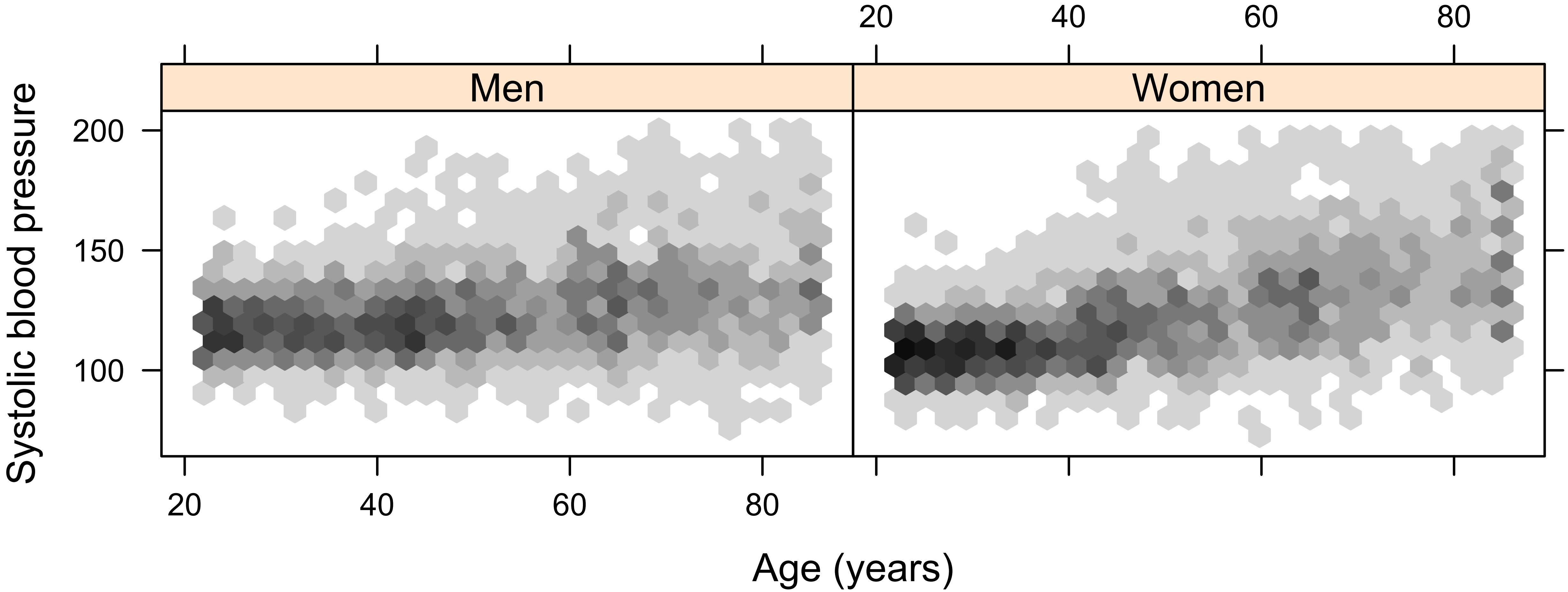
```
des<-svydesign(id=~SDMVPSU, strat=~SDMVSTRA,
      weights=~fouryearwt, nest=TRUE,
      data=subset(nhanes, !is.na(WTDRD1)))

svyplot(BPXDAR~RIDAGEYR,style=“hex",design=des,
  legend=0,xlab="Age (yrs)”,
  ylab="Diastolic BP (mmHg)”)

des<-transform(des, age1=pmin(RIDAGEYR,50)/10,
   age2=pmin(pmax(RIDAGEYR,50),65)/10,
   age3=pmin(pmax(RIDAGEYR,65),90)/10)

ish3s<- svyglm(
    ish~(age1+age2+age3)*RIAGENDR+factor(RIDRETH1),
    design=des,family=quasibinomial)
anova(ish3s)
AIC(ish0s,ish1s,ish2s,ish3s)
```

```
des<-svydesign(id=~SDMVPSU, strat=~SDMVSTRA,
      weights=~fouryearwt, nest=TRUE,
      data=subset(nhanes, !is.na(WTDRD1)))

svyplot(BPXDAR~RIDAGEYR,style=“hex",design=des,
  legend=0,xlab="Age (yrs)”,
  ylab="Diastolic BP (mmHg)”)

des<-transform(des, age1=pmin(RIDAGEYR,50)/10,
   age2=pmin(pmax(RIDAGEYR,50),65)/10,
   age3=pmin(pmax(RIDAGEYR,65),90)/10)

ish3s<- svyglm(
    ish~(age1+age2+age3)*RIAGENDR+factor(RIDRETH1),
    design=des,family=quasibinomial)
anova(ish3s)
AIC(ish0s,ish1s,ish2s,ish3s)
```

```
des<-svydesign(id=~SDMVPSU, strat=~SDMVSTRA,
      weights=~fouryearwt, nest=TRUE,
      data=subset(nhanes, !is.na(WTDRD1)))

svyplot(BPXDAR~RIDAGEYR,style=“hex",design=des,
  legend=0,xlab="Age (yrs)”,
  ylab="Diastolic BP (mmHg)”)

des<-transform(des, age1=pmin(RIDAGEYR,50)/10,
   age2=pmin(pmax(RIDAGEYR,50),65)/10,
   age3=pmin(pmax(RIDAGEYR,65),90)/10)

ish3s<- svyglm(
    ish~(age1+age2+age3)*RIAGENDR+factor(RIDRETH1),
    design=des,family=quasibinomial)
anova(ish3s)
AIC(ish0s,ish1s,ish2s,ish3s)
```

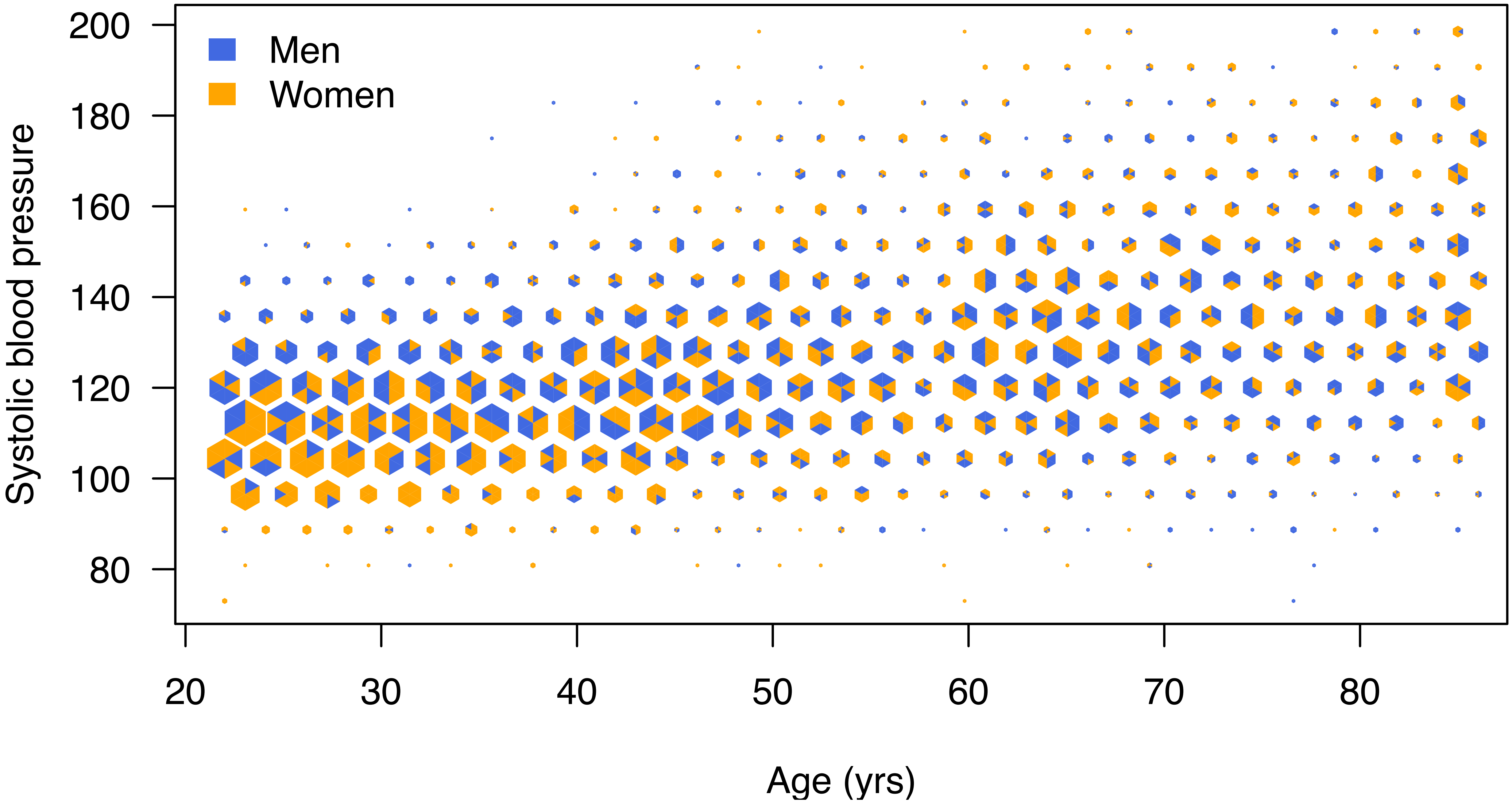
```
des<-svydesign(id=~SDMVPSU, strat=~SDMVSTRA,
      weights=~fouryearwt, nest=TRUE,
      data=subset(nhanes, !is.na(WTDRD1)))

svyplot(BPXDAR~RIDAGEYR,style=“hex",design=des,
  legend=0,xlab="Age (yrs)”,
  ylab="Diastolic BP (mmHg)”)

des<-transform(des, age1=pmin(RIDAGEYR,50)/10,
   age2=pmin(pmax(RIDAGEYR,50),65)/10,
   age3=pmin(pmax(RIDAGEYR,65),90)/10)

ish3s<- svyglm(
    ish~(age1+age2+age3)*RIAGENDR+factor(RIDRETH1),
    design=des,family=quasibinomial)
anova(ish3s)
AIC(ish0s,ish1s,ish2s,ish3s)
```

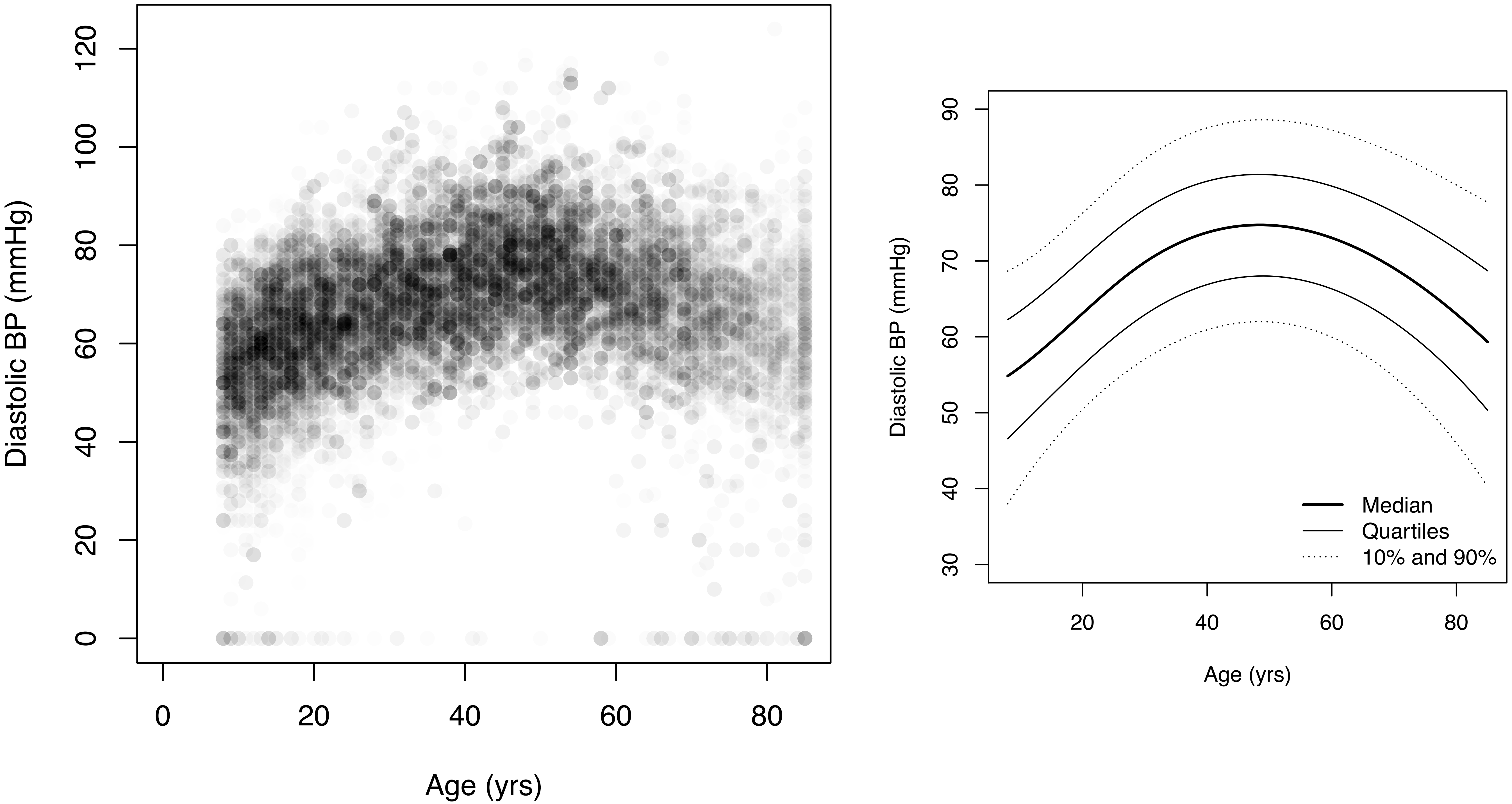
```
des<-svydesign(id=~SDMVPSU, strat=~SDMVSTRA,
      weights=~fouryearwt, nest=TRUE,
      data=subset(nhanes, !is.na(WTDRD1)))

svyplot(BPXDAR~RIDAGEYR,style=“hex",design=des,
  legend=0,xlab="Age (yrs)”,
  ylab="Diastolic BP (mmHg)”)

des<-transform(des, age1=pmin(RIDAGEYR,50)/10,
   age2=pmin(pmax(RIDAGEYR,50),65)/10,
   age3=pmin(pmax(RIDAGEYR,65),90)/10)

ish3s<- svyglm(
    ish~(age1+age2+age3)*RIAGENDR+factor(RIDRETH1),
    design=des,family=quasibinomial)
anova(ish3s)
AIC(ish0s,ish1s,ish2s,ish3s)
```



Weights used in graphics automatically (**population graphics**)

- alpha-blending
- hexagonal binning
- weighted smoothing


Men
Women
200
180
160
140
120
100
80
Systolic blood pressure
20
30
40
50
60
70
80
Age (yrs)


Diastolic BP (mmHg)
0
20
40
60
80
100
120
Age (yrs)
0
20
40
60
80
Diastolic BP (mmHg)
30
40
50
60
70
80
90
Age (yrs)
20
40
60
80
Median
Quartiles
10% and 90%

# Horvitz-Thompson estimator

- Involves $n^2$ operations
- Lots of computational special cases are faster
- Sparse matrices automate some more
- Users no longer need to know.

[also, resampling]

# Influence functions

- Classical sampling theory works for population totals

- Influence functions let (almost) anything look like a population total

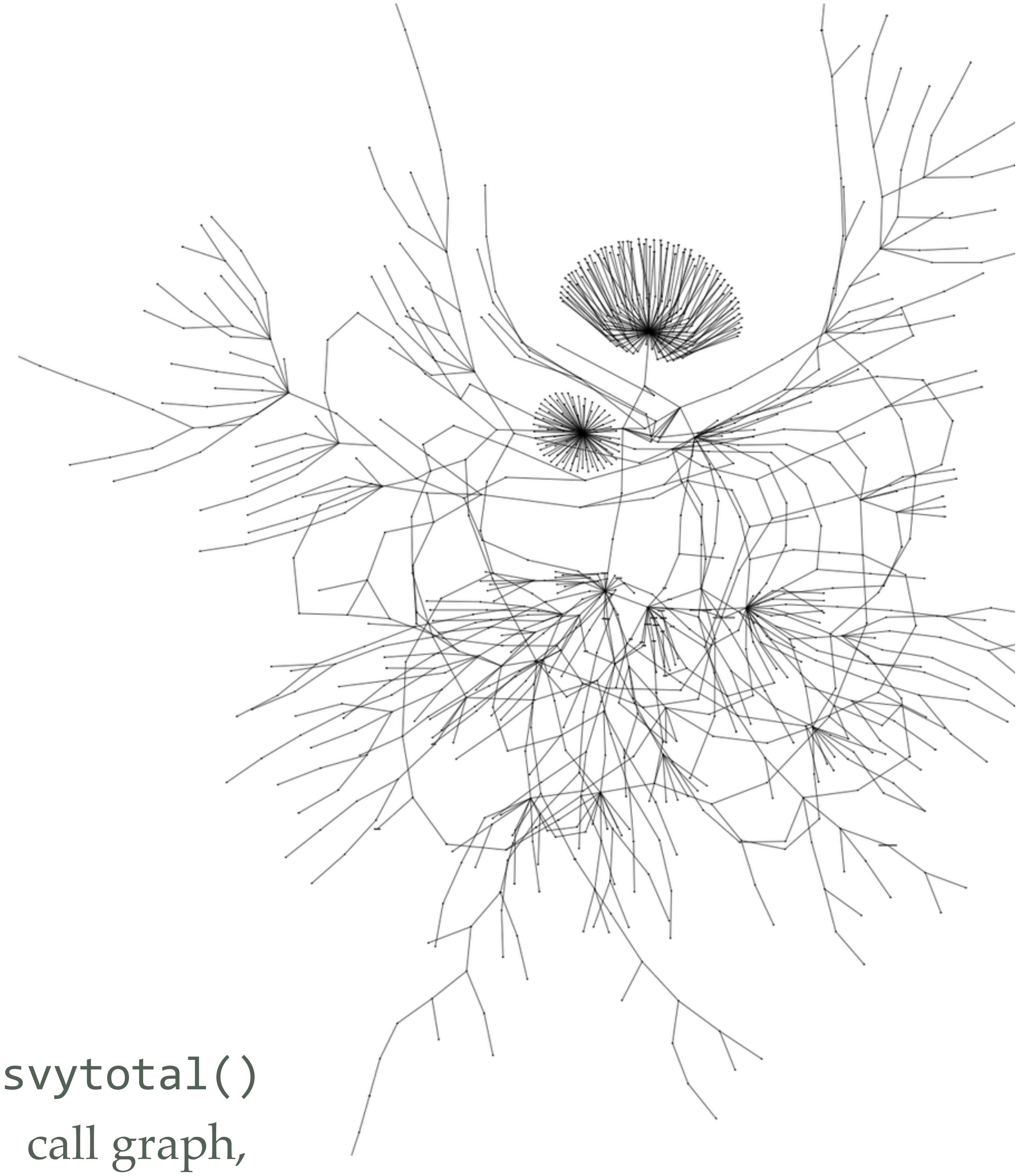
$$\sqrt{n}(\hat{\theta} - \theta) = \sum_{i=1}^{n} \mathbb{I}(\beta)_i + o_p(1)$$

# Regression

- svyglm, patterned on glm but with survey object instead of data frame
- same informal interface
  - coef,vcov, SE, AIC, BIC, anova methods
- residuals, diagnostic plots.
- also Cox model, loglinear model, ordinal models

# Easier, not faster

- Don’t worry much about efficiency until someone complains
- Profiling helps a lot
- Some parallel code



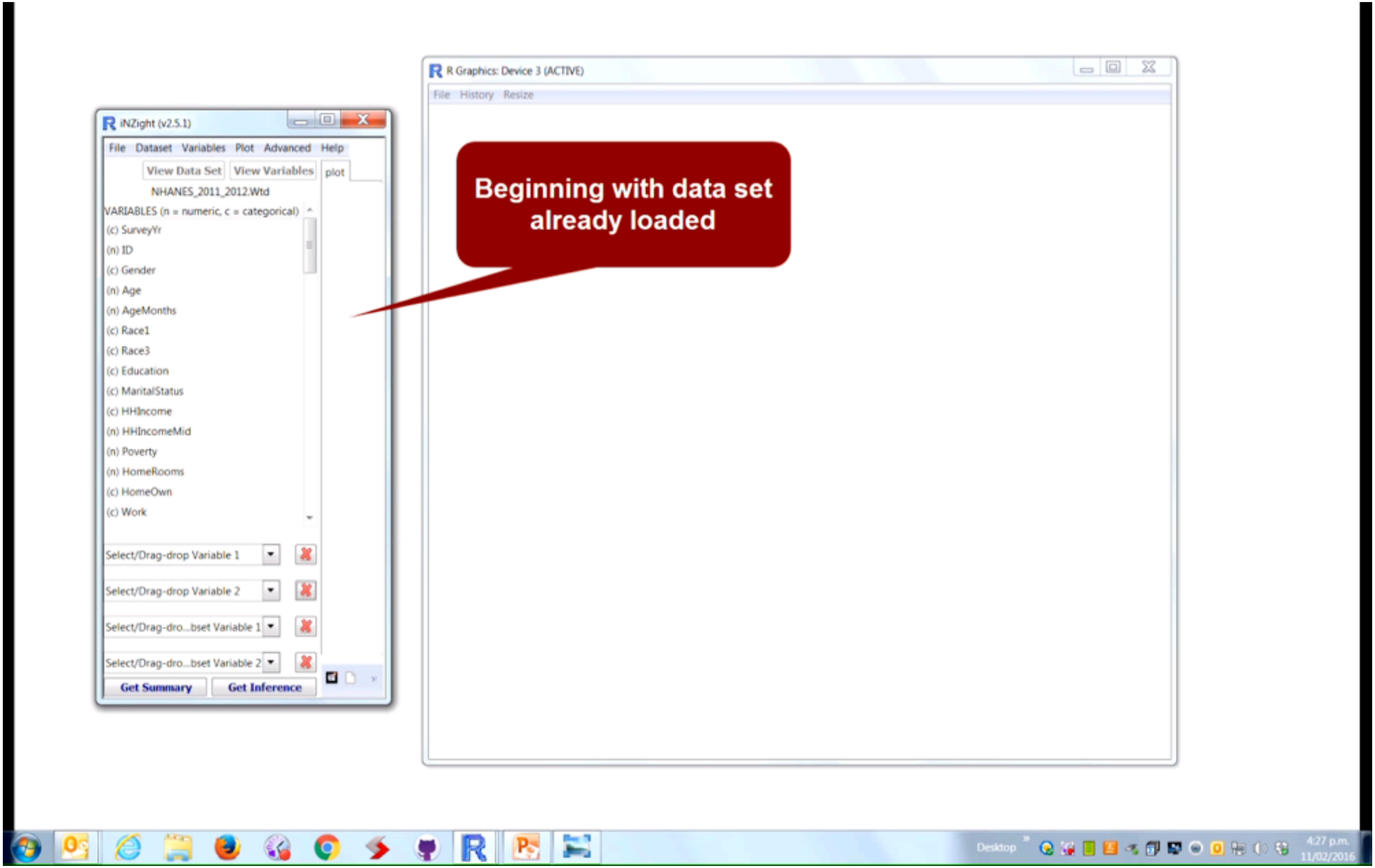
`svytotal()` call graph, from Renjin

# Data of Unusual Size

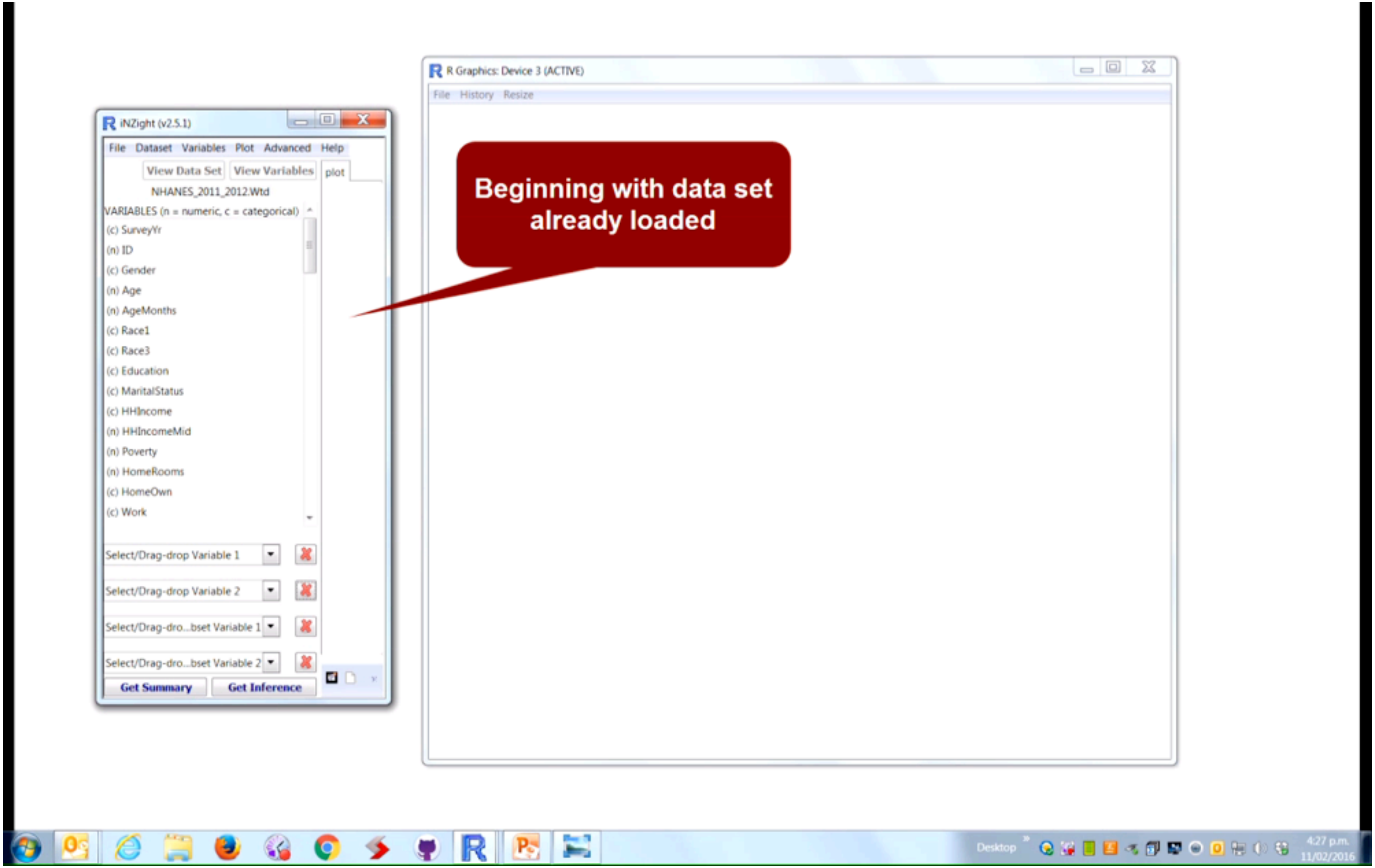
- 2007 “Data of Unusual Size? I don’t think they exist”
- 2011: use American Community Survey (“AAARGH”)
- sqlsurvey: backend computations in column-store MonetDB
- goal: convert to dplyr as middleware

# iNZight

- Interactive statistics package emphasising graphics
- Used in NZ high schools, Auckland Uni STATS 10x
- Complex survey support in early stage
    - weighted hexbin for scatterplots
    - weighted summary statistics
    - weights in generalised linear models



Beginning with data set already loaded



Beginning with data set already loaded

# Domestication

- Use objects to ensure survey data description stays part of the data (`Stata` has similar idea)

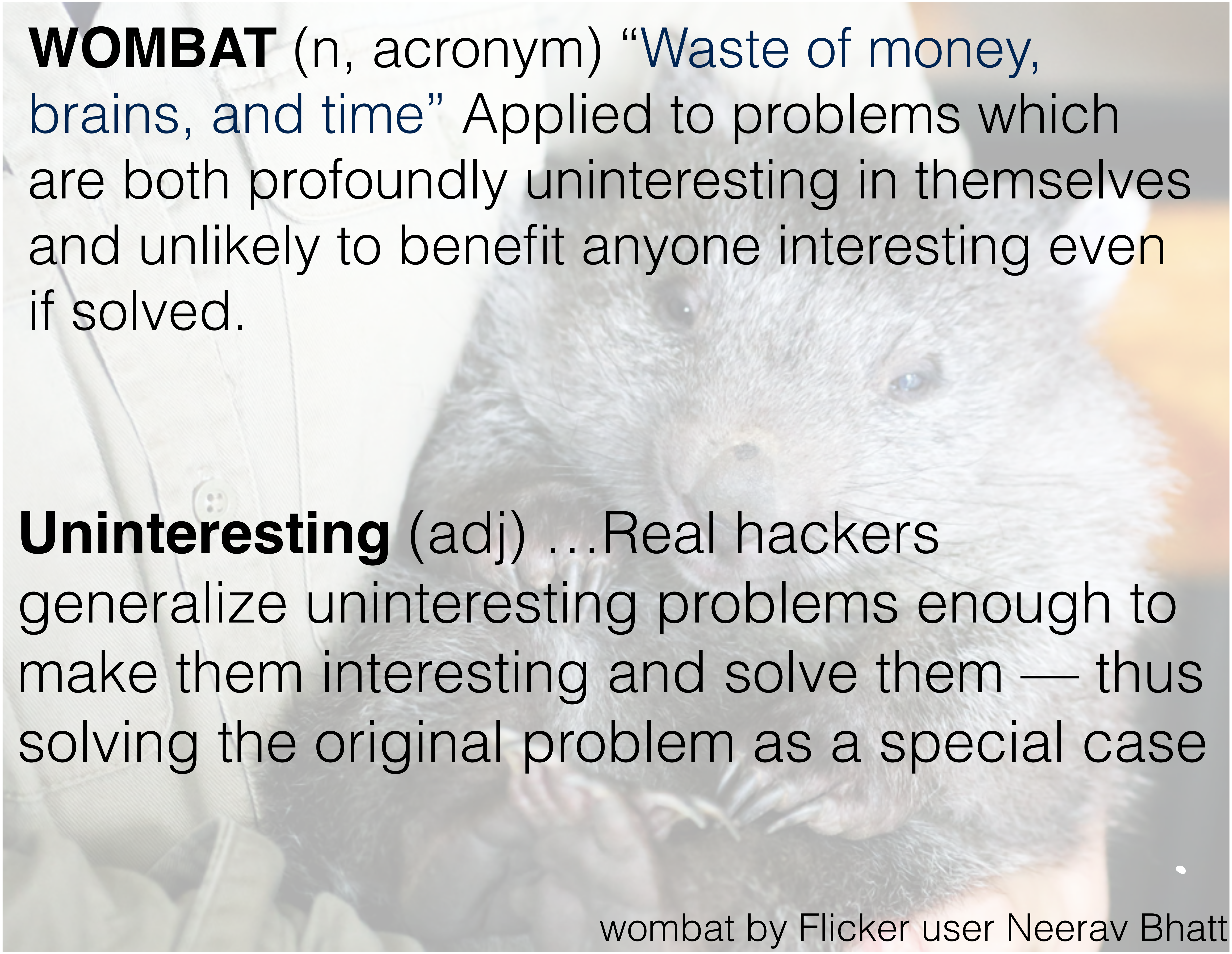
- Supply the same user interface as for cross-sectional data — including graphics

- Use influence functions and resampling as the common mathematical interface(s)

- Hide computational optimisations from the user

**WOMBAT** (n, acronym) "Waste of money, brains, and time" Applied to problems which are both profoundly uninteresting in themselves and unlikely to benefit anyone interesting even if solved.

wombat by Flicker user Neerav Bhatt



**WOMBAT** (n, acronym) “Waste of money, brains, and time” Applied to problems which are both profoundly uninteresting in themselves and unlikely to benefit anyone interesting even if solved.

**Uninteresting** (adj) …Real hackers generalize uninteresting problems enough to make them interesting and solve them — thus solving the original problem as a special case

wombat by Flicker user Neerav Bhatt

Superb fairywren by JJ Harrison, from Wikipedia

# Questions?