Data + Analytics (+ Visualization) = Information

Louise Ryan,
Chief of CSIRO Mathematical and Information Sciences
Adjunct Professor, Harvard Biostatistics Department
Adjunct Professor, Macquarie Statistics Department

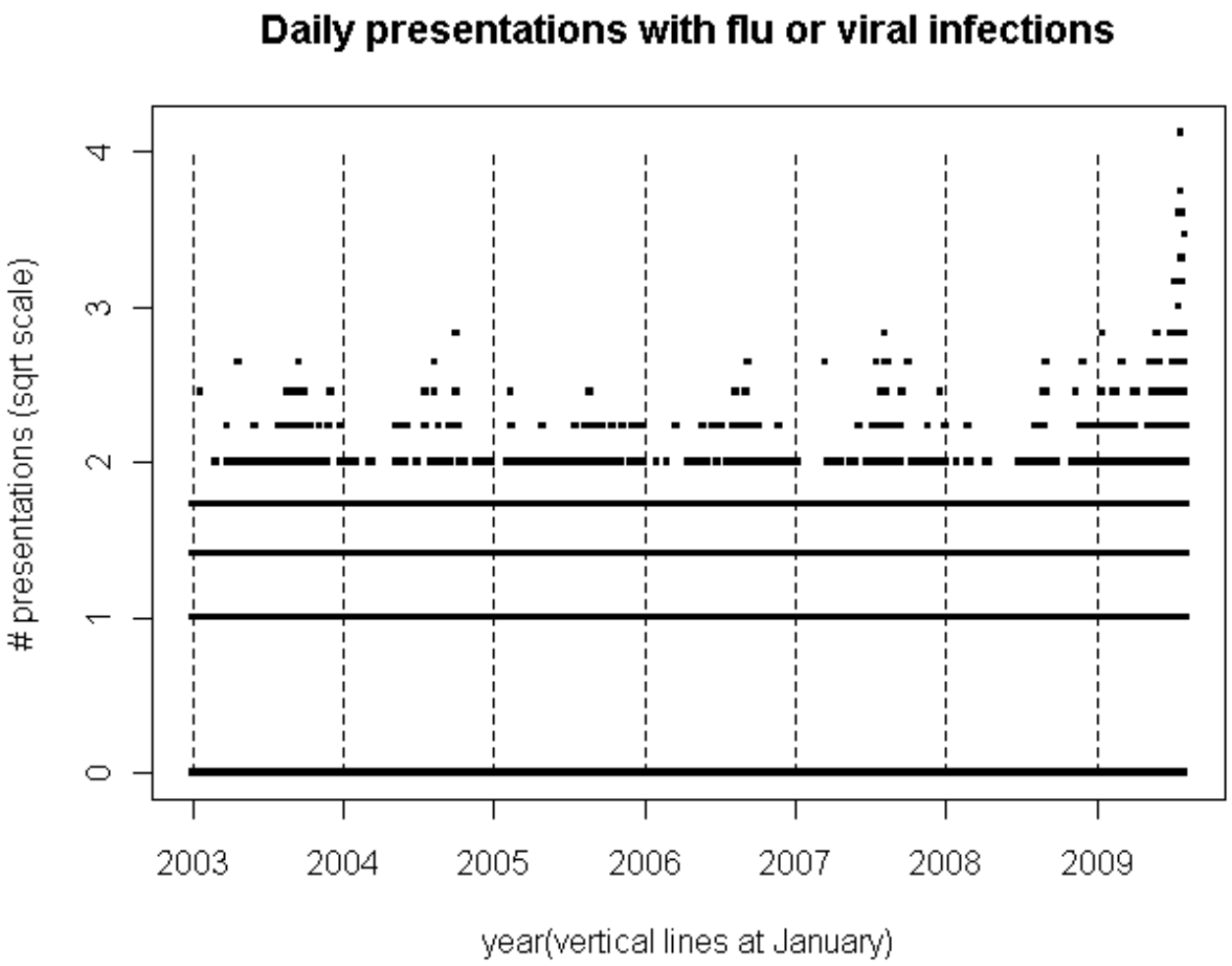
CSIRO

H1N1 Flu patterns for a QLD Hospital

- Australia E-Health Centre – joint venture with CSIRO and QLD Health
  - improve the quality and safety of healthcare for individuals and communities through an ICT research program

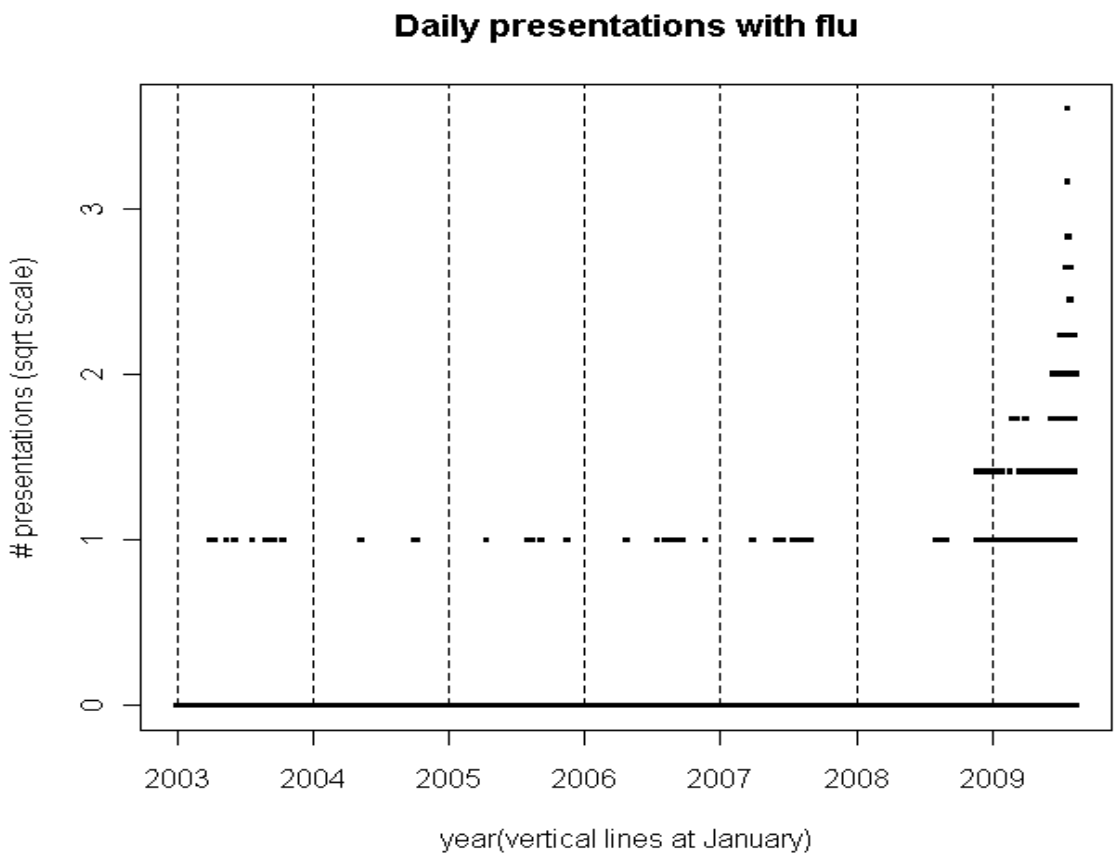
- Patient Admission Prediction Tool (PAPT) uses EDIS data to predict ED patient loads, hospital admissions, length of stay etc.
  - Swine flu event very unusual
  - New questions

Challenges



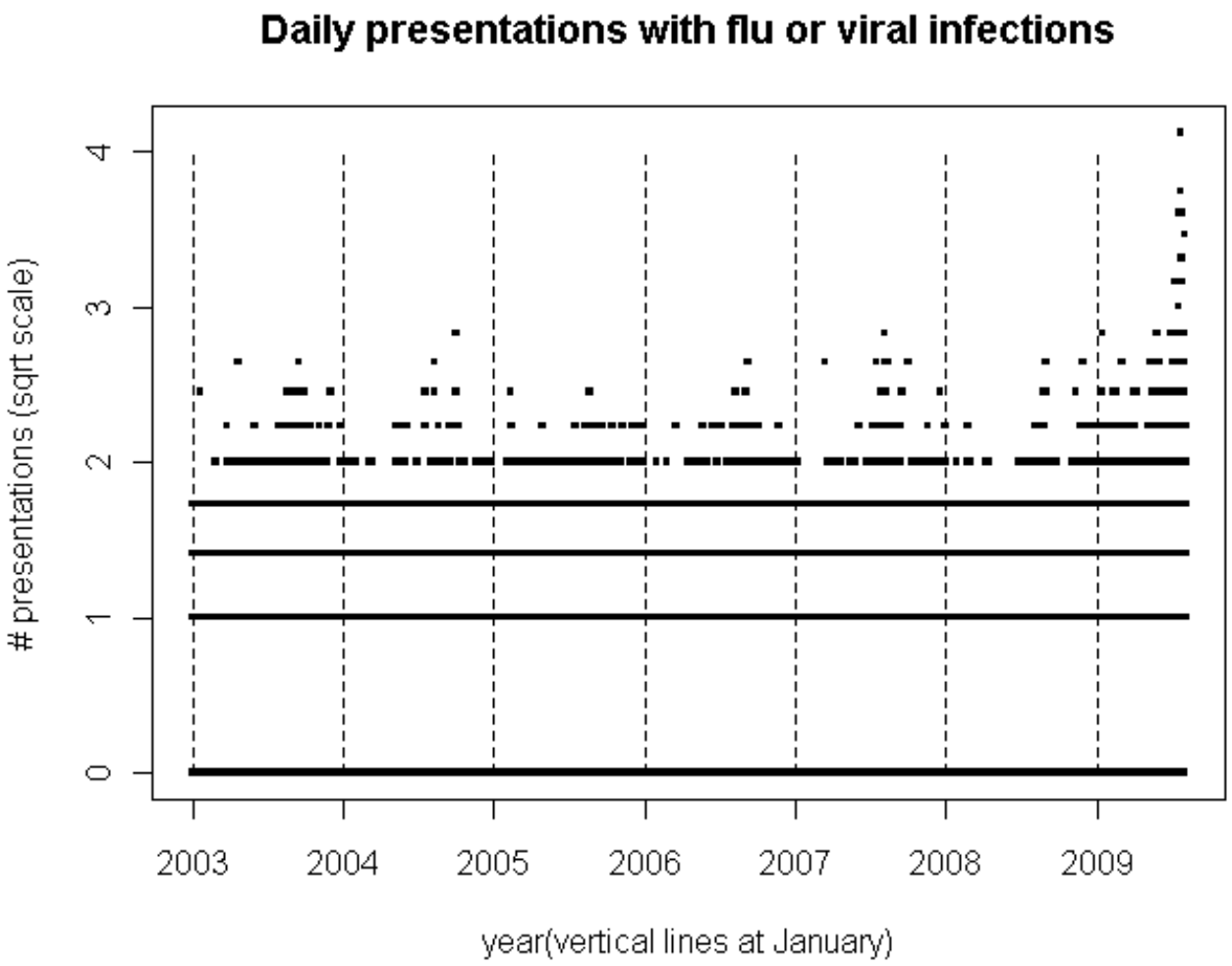
- Which ICD10 codes?
- Signal vs noise – e.g. Swine Flu vs "regular" seasonal pattern?

Challenges



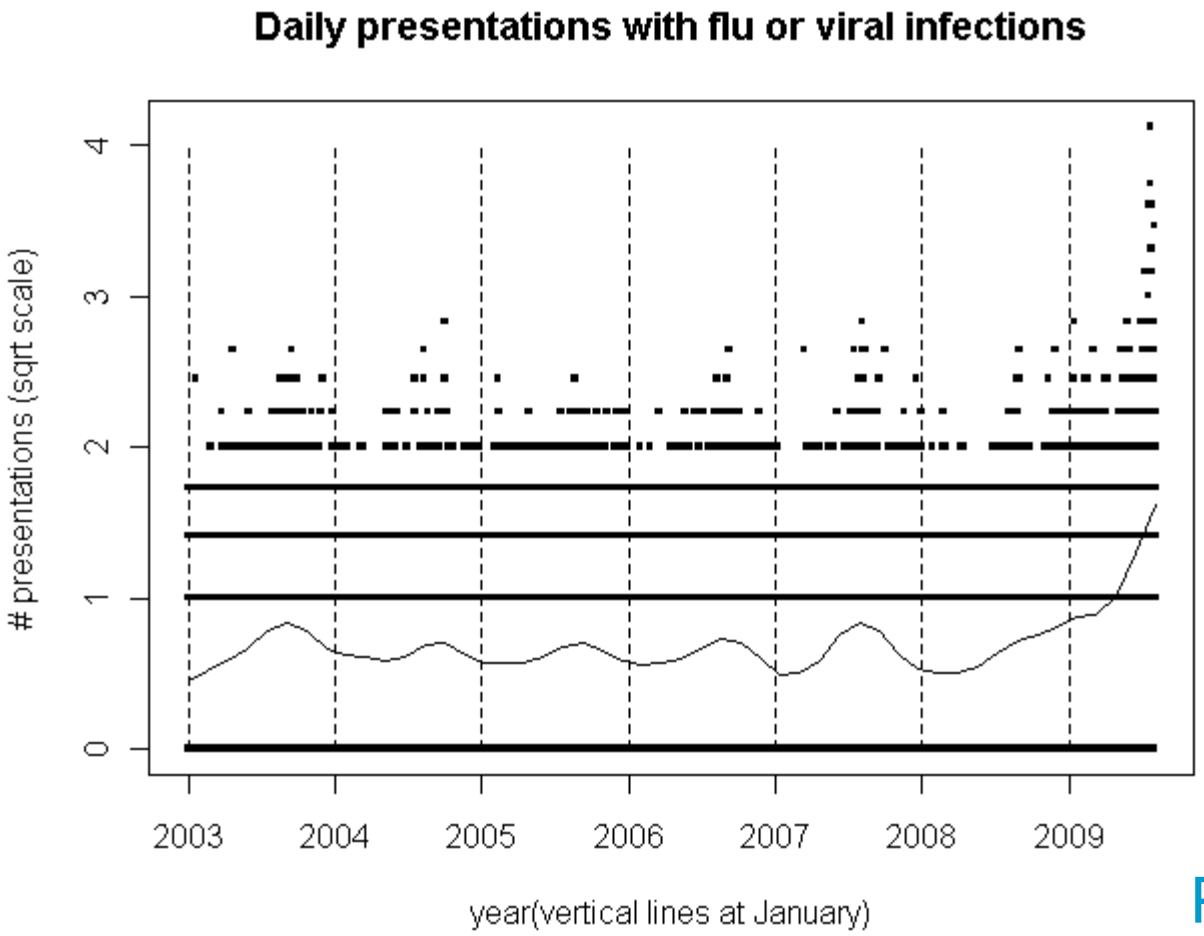
- Which ICD10 codes?
- Signal vs noise – e.g. Swine Flu vs “regular” seasonal pattern?

Challenges



- Which ICD10 codes?
- Signal vs noise – e.g. Swine Flu vs "regular" seasonal pattern?

A simple scatterplot smooth does wonders!



What do we see?

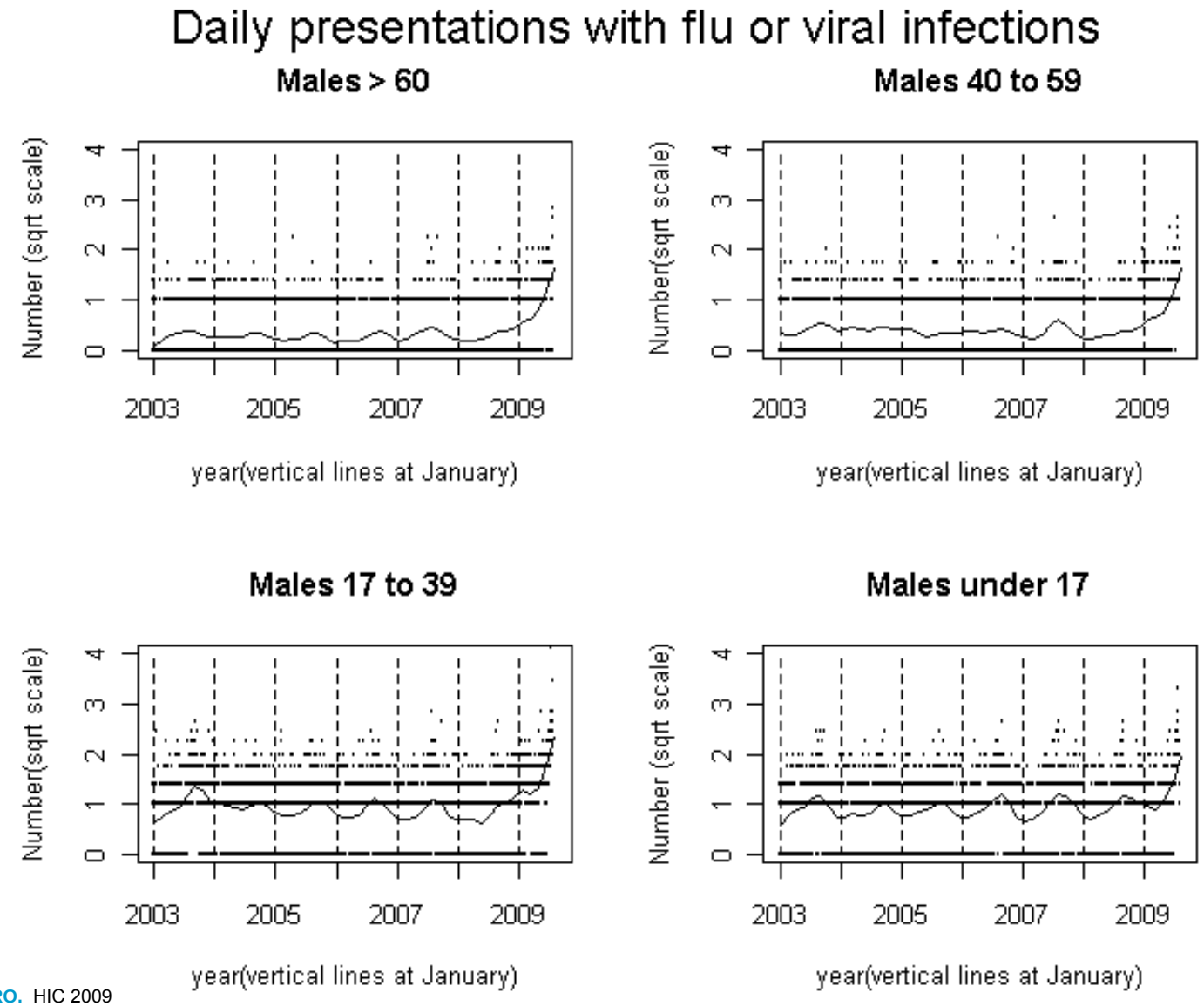
- Annual patterns
- Something unusual in late 2008?

What else?

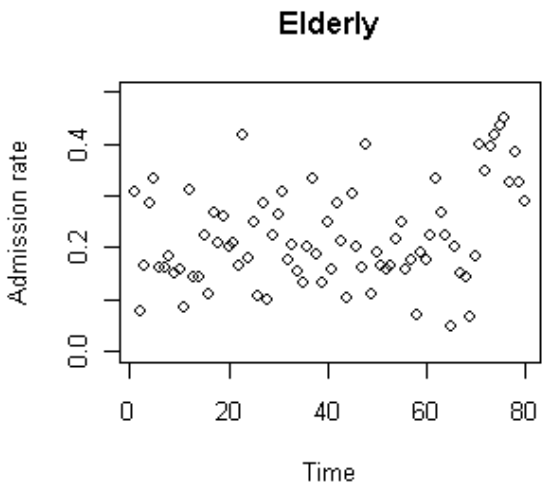
- Subgroups?
- Severity?
- Direction?
- How early might we have detected it?

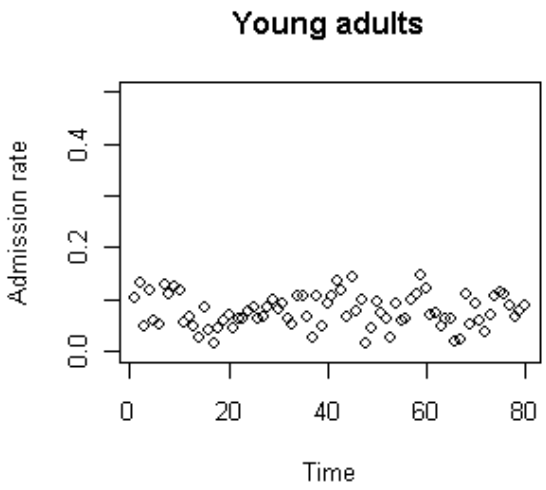
Poisson Regression

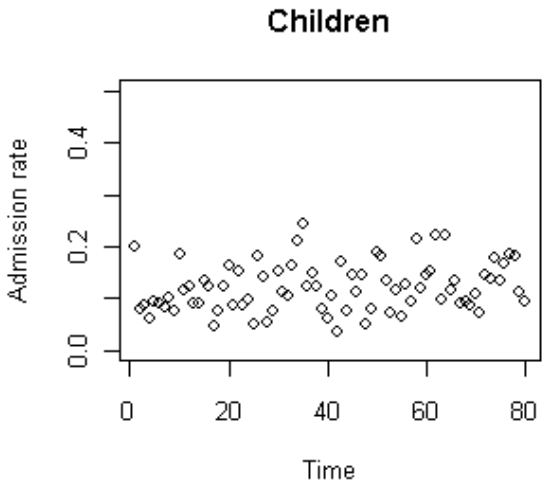
Patterns by age and gender?

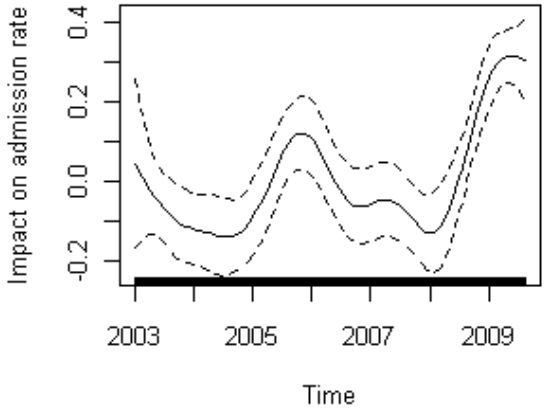


How do admission rates look?





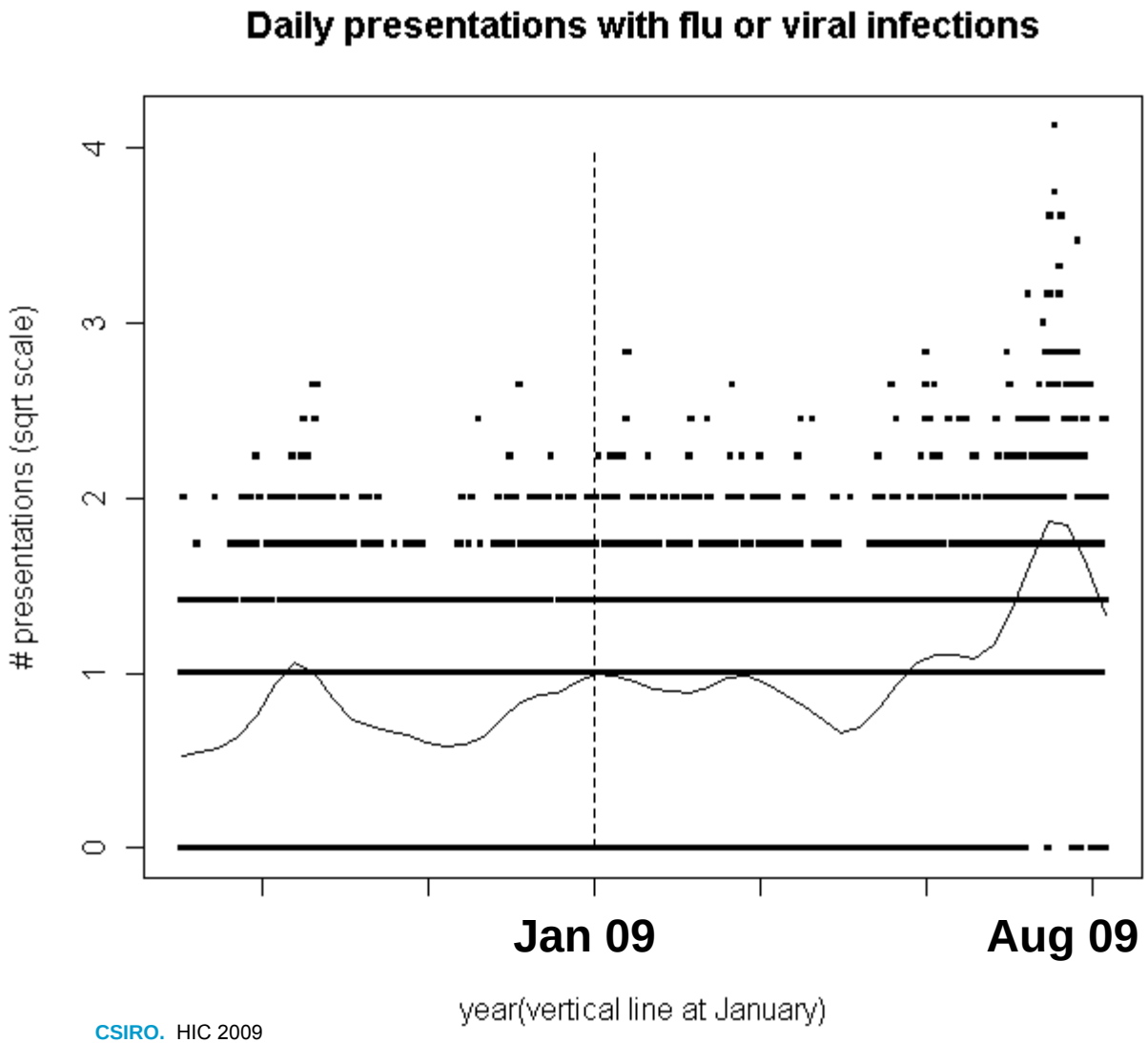




Generalized Additive Model reveals

- Strong age effects
- Gender effect
- Time effect

More recent patterns... Some good news ☺

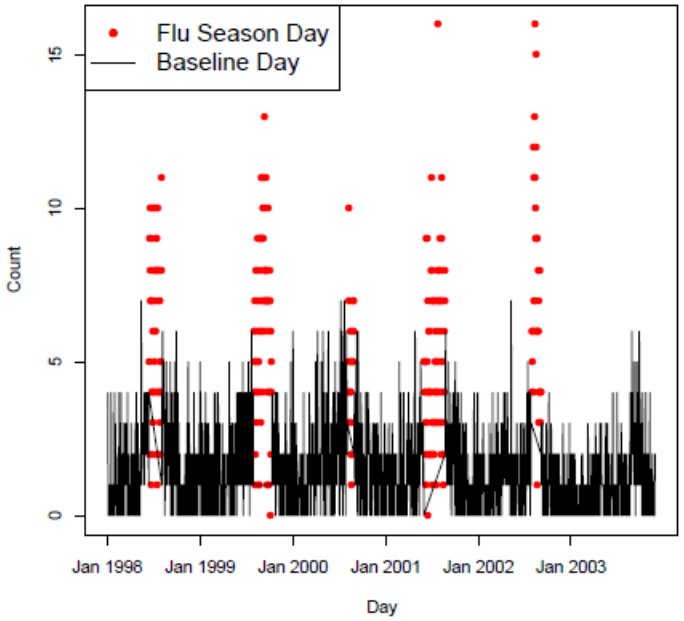


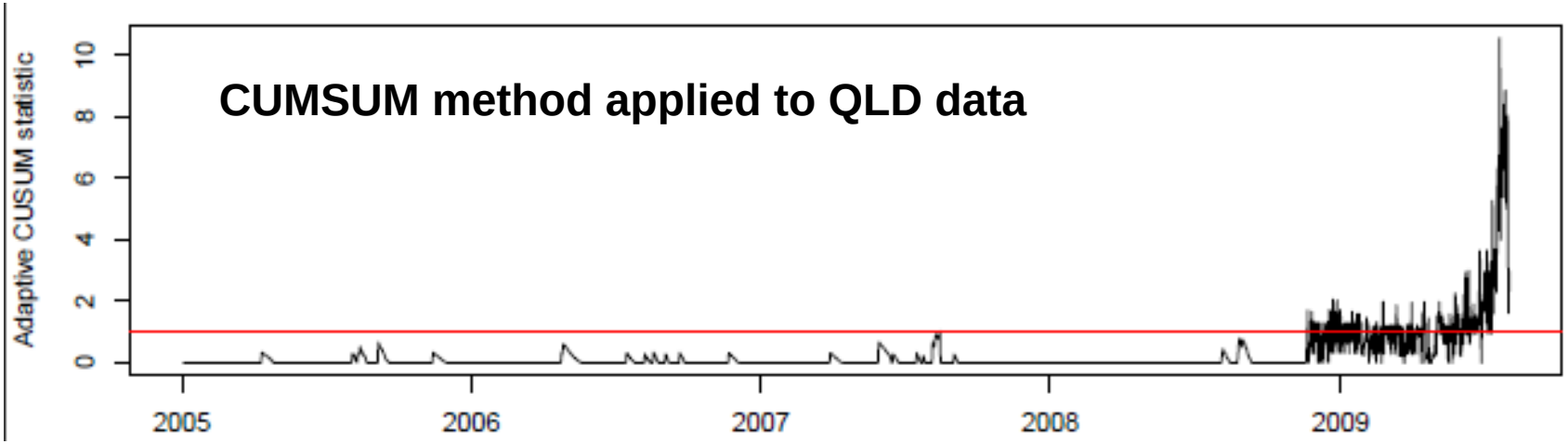
- Consistent with reported very high infection rate
- Confirm via agent-based infectious disease modelling as well as theoretical models,

Epidemic surveillance tool ....

Many approaches possible

- SATSCAN etc
- Hidden Markov Models
- CUMSUM methods

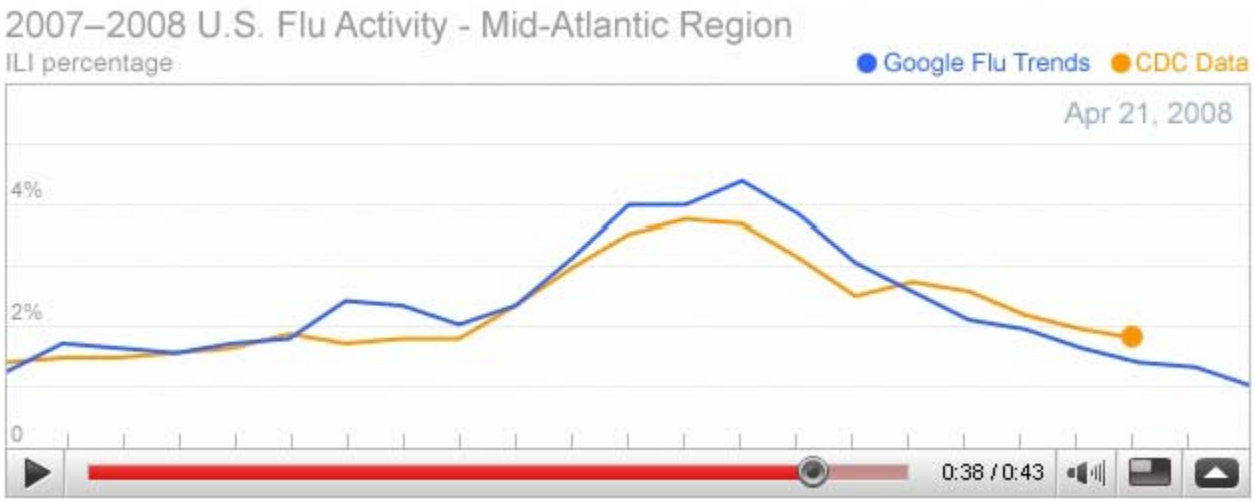




Let's step - what have we learned?

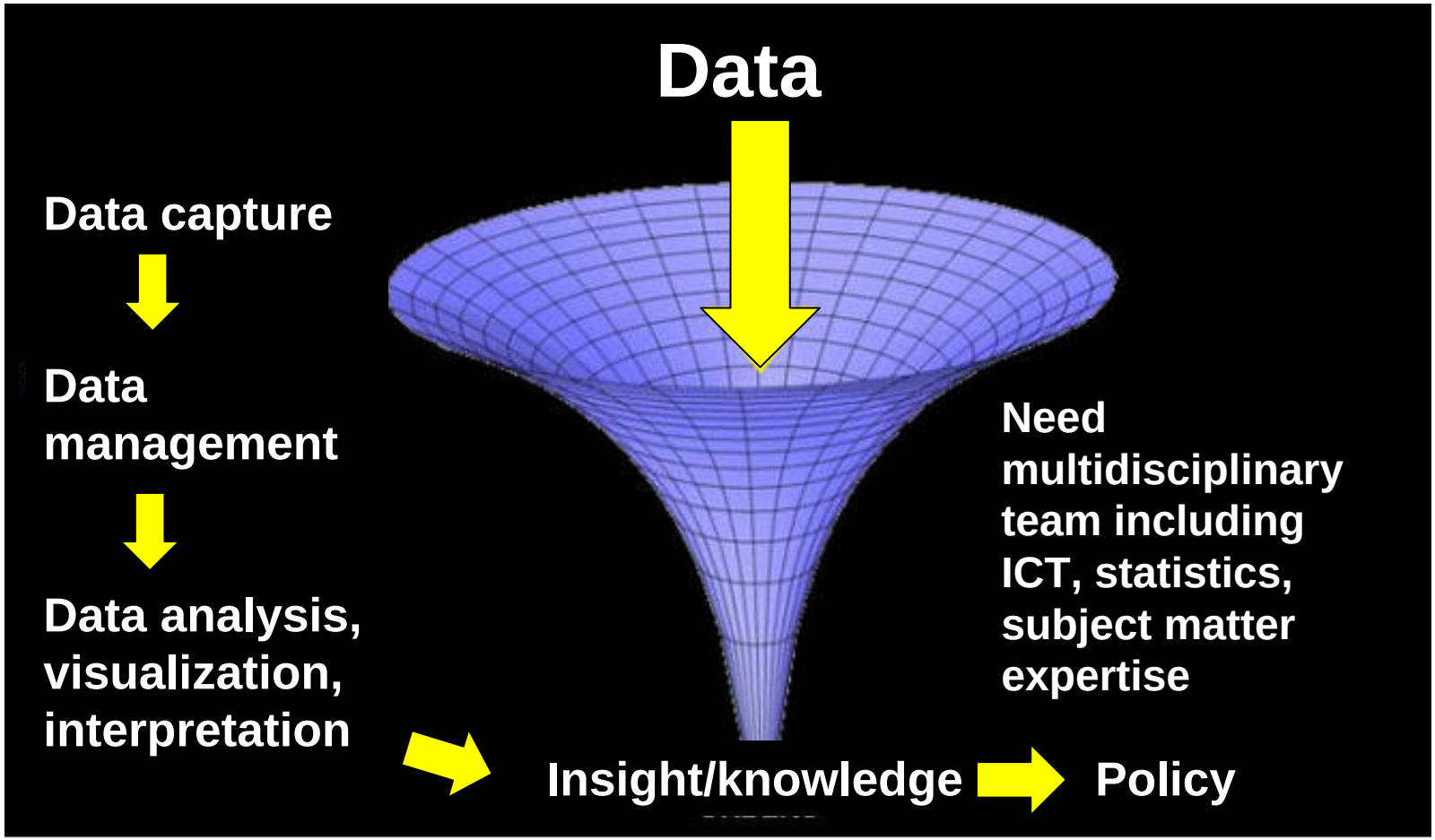
1) Unprecedented volumes of data - fantastic opportunities

- Analysis of hospital databases for
  - Management
  - Research
- Unusual data sources – e.g. Google Flu Trends



- Record linkage – e.g. DVT and air travel

But, data alone is not information



The need for careful analytics

1. Sometimes need specialized solutions to cope with massive databases

   Eg - SEE Australia – An ARC funded project involving Sydney Uni and NSW Health to explore use of routine databases for health research

   - Database development and linking
   - Methodological explorations

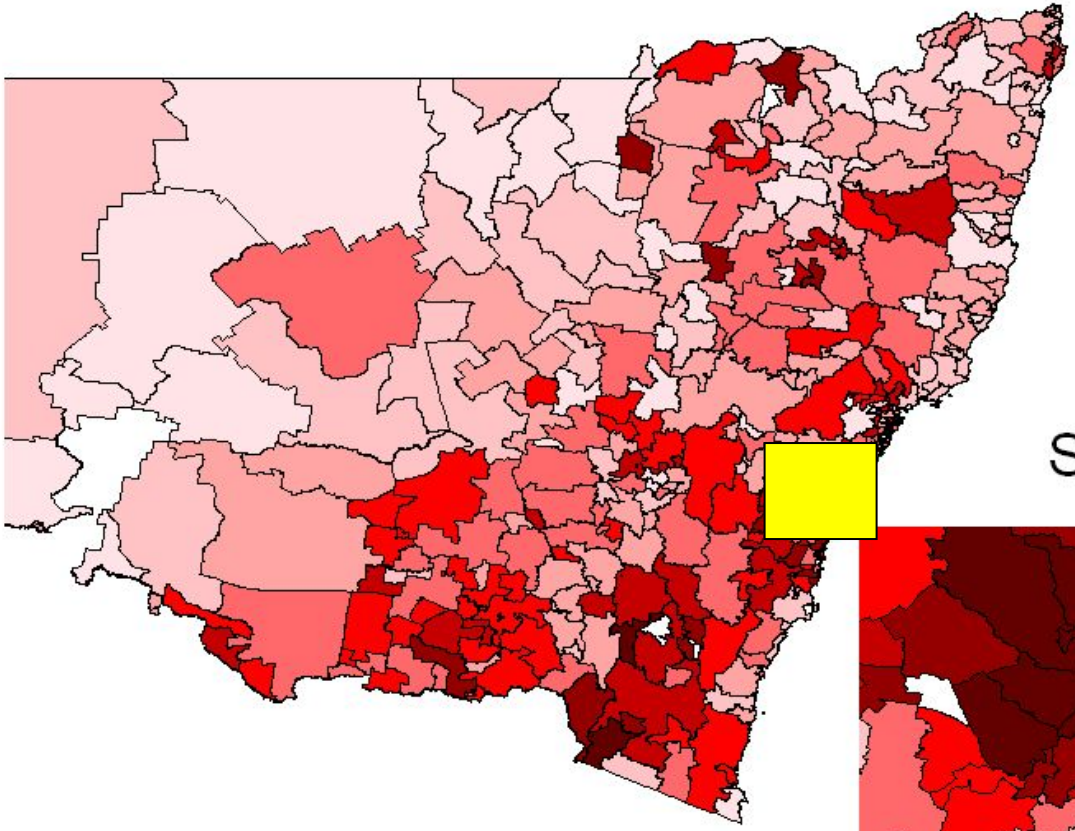
Case study – spatio-temporal trends in ischemic heart disease as well as association with social disadvantage

The Data (from HOIST)

- Number of daily hospital discharges (Y) with Ischemic Heart Disease (IHD) where admission had been via emergency room for
  - 591 postcodes in NSW
  - Every day from July 1, 1996 to June 30, 2001
  - Males and females
  - 5-year age increments
- Denominator (N) obtained from census
- Social disadvantage measured at postal area level using the census-derived SEIFA index

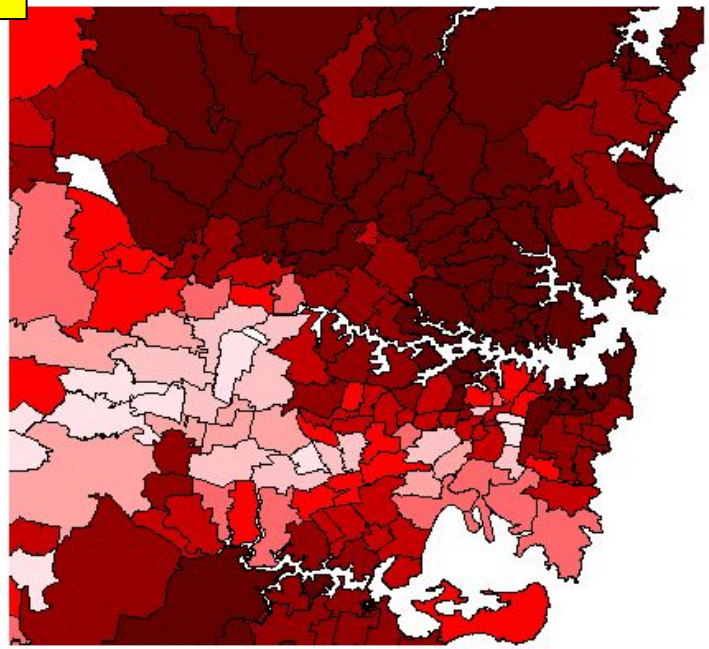
CSIRO

SEIFA distribution in NSW



High values (dark red) indicate social advantage

Sydney Metro

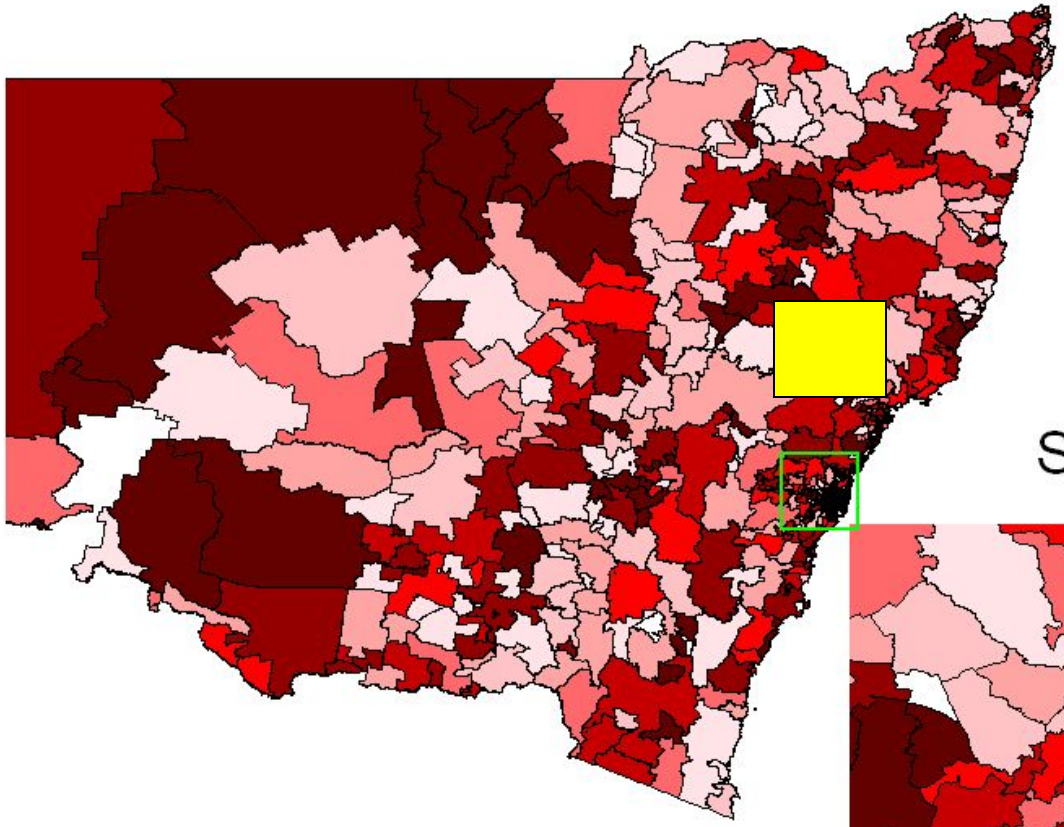


SED (SEIFA) by Postcode, NSW 94/95

- 602.3 - 921.2
- 921.2 - 947.7
- 947.7 - 968.5
- 968.5 - 991.5
- 991.5 - 1013.4
- 1013.4 - 1046.7
- 1046.7 - 1091.8
- 1091.8 - 1179.7

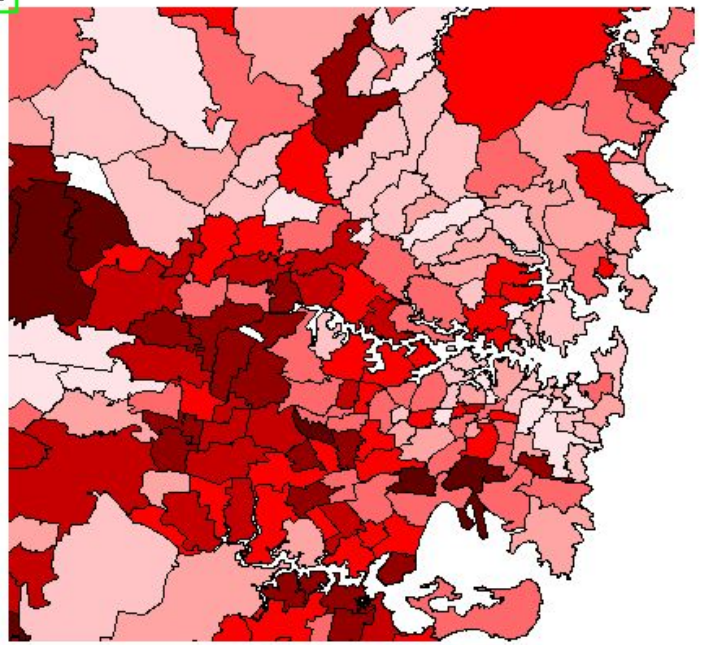
CSIRO

NSW IHD rates



Dark red indicates high rates of IHD

Sydney Metro



ISIR for IHD by Postcode, NSW 94/95

- 0 - 41.5
- 41.5 - 61
- 61 - 76.9
- 76.9 - 89.2
- 89.2 - 104.1
- 104.1 - 124
- 124 - 151
- 151 - 359.4

CSIRO

SES and Heart disease

Well known relationship

- 25% – 50% of observed gradient due to risk factors like smoking, hypertension and diabetes in lower socio-economic groups (Marmot et al.,1997)
- Access to healthcare (Bosma et al., 2005)
- Imbalance between workplace demands and economic reward (Lynch et al.,1997)
- Poor education, lower levels of health literacy, low birth weight (Marmot, 2000)

Relationship may vary with gender with the association thought to be stronger in males (Thurston, 2005)

CSIRO

Analysis challenges

Would like to fit a Poisson regression model that characterizes postcode-specific IHD rate as a function of time, age, gender, SEIFA, other variables. But

- Dataset has ~33 million records (591 areas x 5 years x 365 days x 18 age groups x 2 genders).
- Standard approach (standardization) still leaves dataset of ~1 million, plus loses chance to explore age/sex effects.
- Need to account for spatial correlation

Even with a powerful computer, standard methods would not run

CSIRO

Key insight

A property of the Poisson distribution:

if $Z_1 \sim \text{Poisson}(\lambda_1)$ and $Z_2 \sim \text{Poisson}(\lambda_2)$, then

$Z_1 + Z_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$ β

Algorithm:

i) initialize parameters

ii) update one parameter at a time, treating others as known. Can drastically reduce data dimension by collapsing over all the variables, except for the one whose parameter is being updated

Keep iterating step ii until model has converged

Algorithm worked spectacularly well with our data

CSIRO

The algorithm in more detail

Model, $Y_{ijk} \sim \text{Poisson}(N_{ijk}\,\lambda_{ijk})$ where

$$\log(\lambda_{ijk}) = \delta_k + \alpha_{W+} \beta_0 + \beta_1 SEIFA_i + \beta_{2j} + b_i$$

Suppose δ_k, $\alpha_{W,}$ $\beta_{0,}$ β_{2j} and b_i are all known so that we just need to estimate β_1

Using Poisson property,

$$Y_{i..} = \sum Y_{ijk} \sim \text{Poisson}(o_i \exp(\beta_1 SEIFA_i)) \text{ where}$$

$$o_i = \sum N_{ijk} \exp(\delta_k + \alpha_{W+} \beta_0 + \beta_{2j} + b_i)$$

This is an easy Poisson regression with $\log(o_i)$ as an offset.

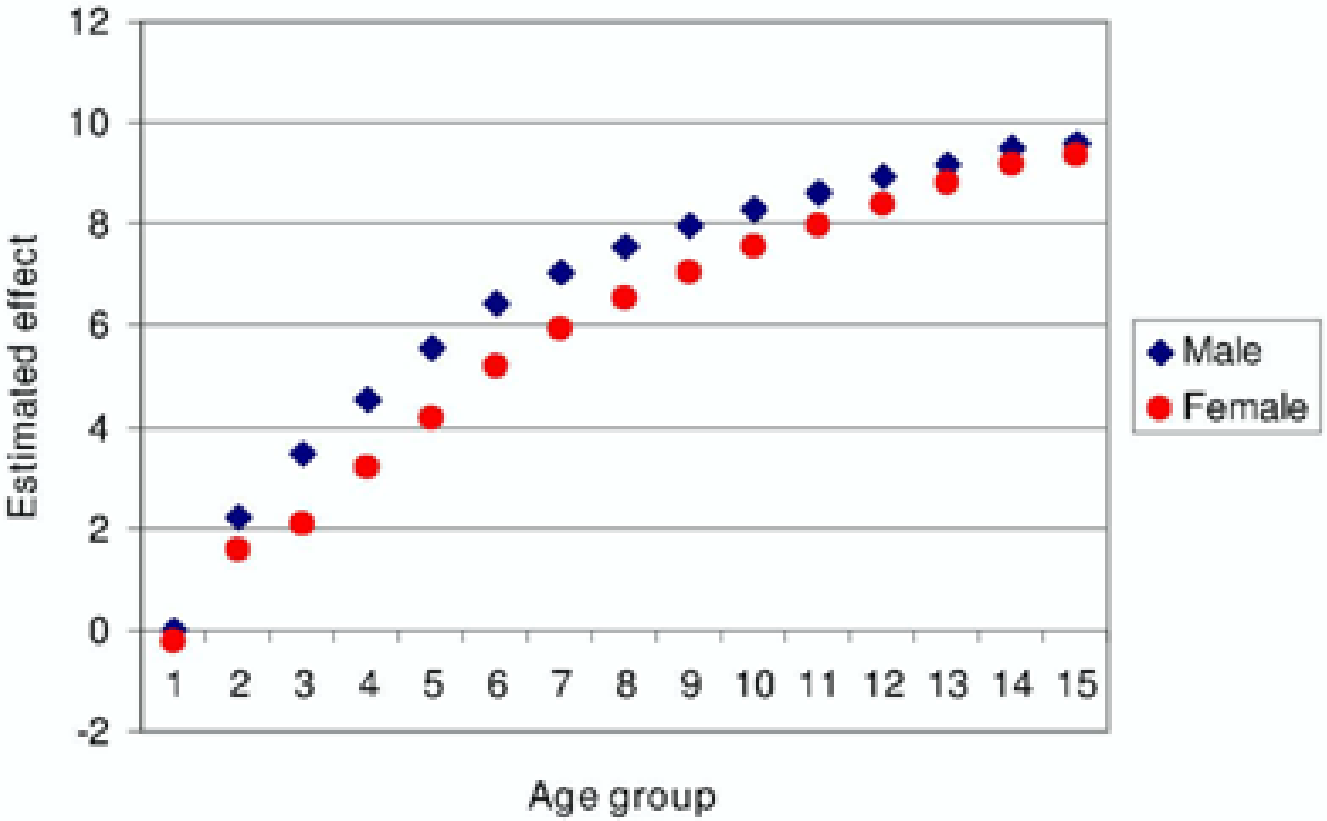
Full algorithm:

1. Initialize parameter values
2. Update one parameter at a time, treating others as fixed and known, collapsing over redundant dimensions until convergence

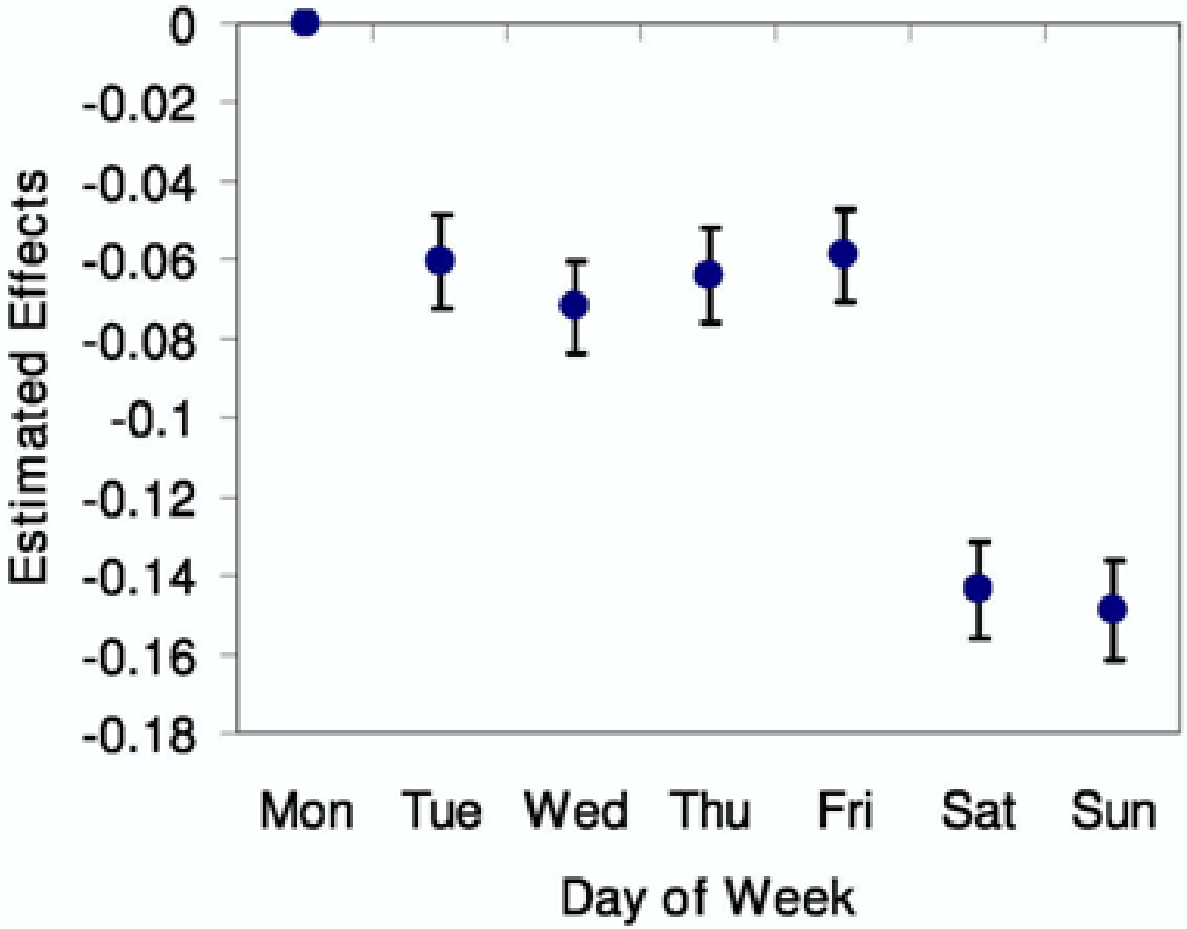
Simple example of Gauss-Seidel algorithm.

CSIRO

Age and Gender effects

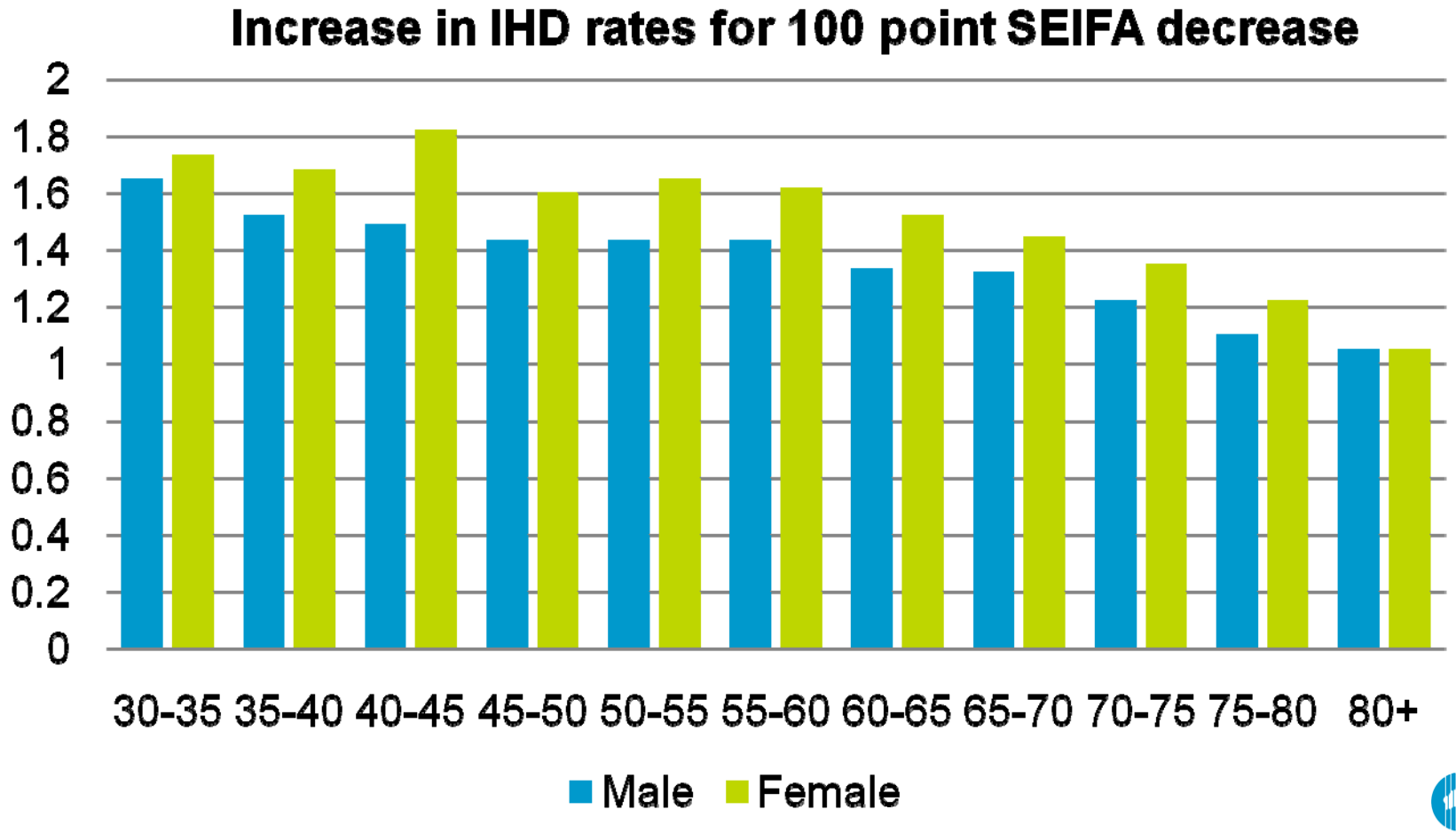


Day of Week effects



SEIFA effect

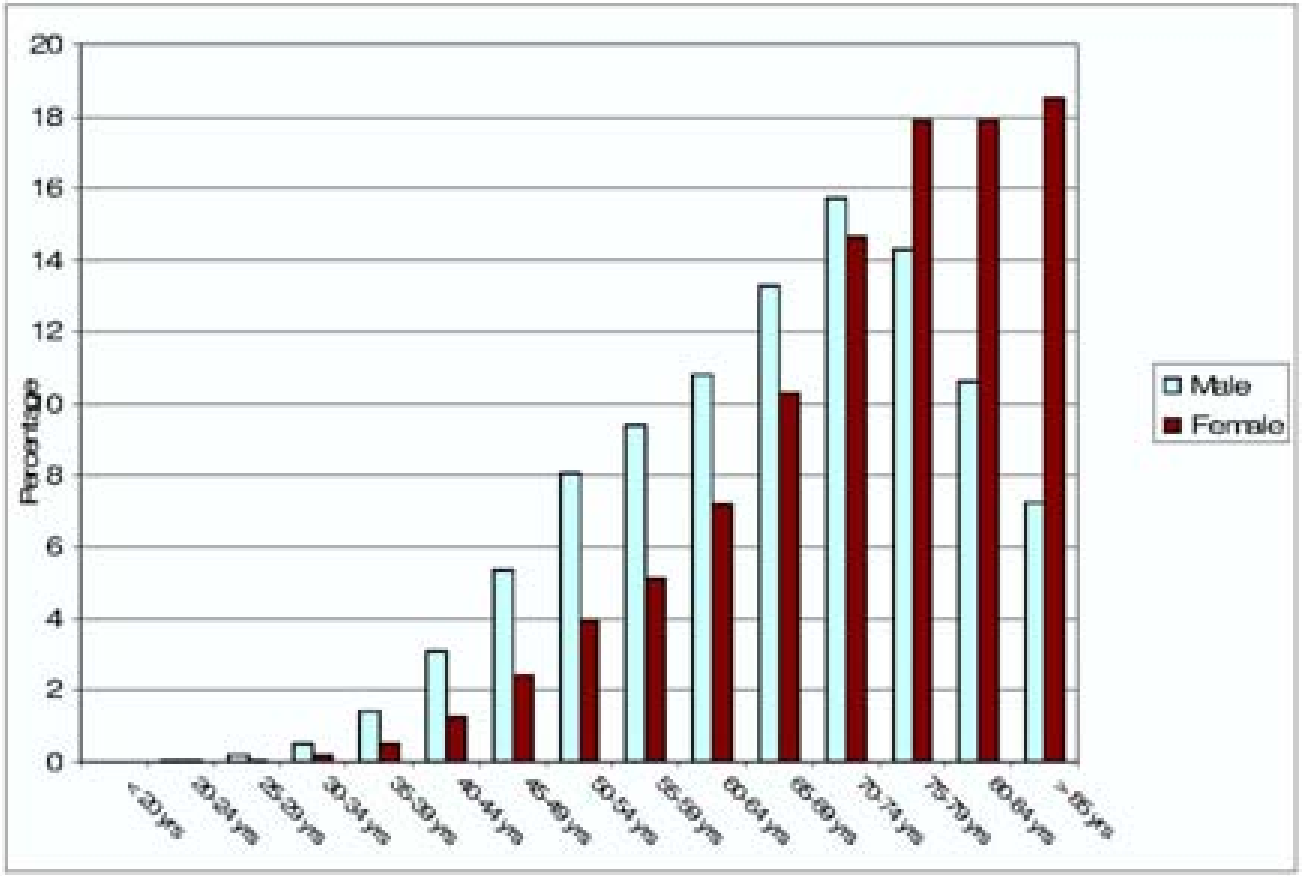
Estimated overall SEIFA effect highly significant, with a 20% increase in IHD rates associated with a 100% decrease in SEIFA levels



Interpretation

- Impact of social disadvantage greater at younger ages
- For a given age, the impact of social disadvantage is stronger for women than for men
- Curious contradiction? Model with just a gender/ SEIFA interaction suggests that men are more sensitive to social disadvantage. Consistent with the published literature. What's happening?

Simpson's paradox in interaction space!



The need for careful analytics (cont'd)

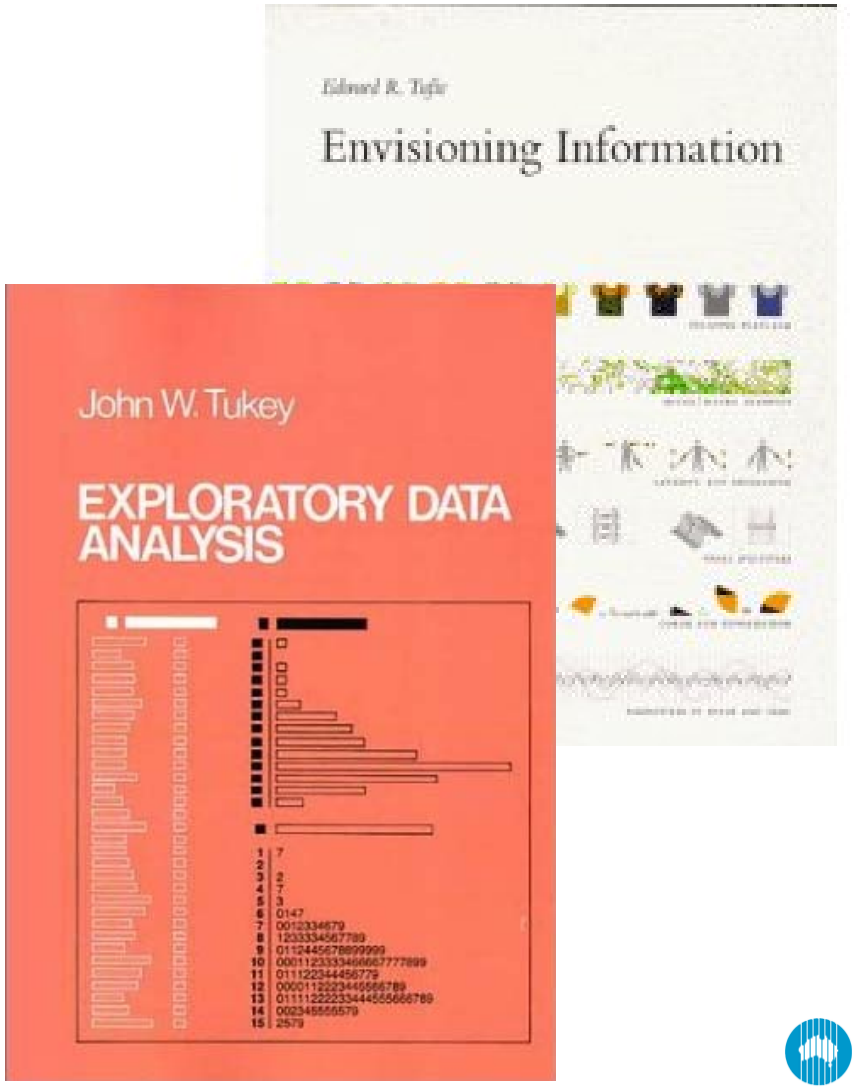
1. Sometimes need specialized solutions needed to cope with data volume

1. Administratively collected data don't always satisfy criteria for good quality epidemiology:
   - Non-scientific sampling
   - Presence in sampled may be informative
   - Important data may be missing
2. Inherent complexities in health data
   - Censoring
   - Competing risks (e.g. Length of stay data)

CSIRO

Visualization is often neglected

Need to follow lead of visionaries like Edward Tufte and John Tukey who developed new ways to communicate data and information



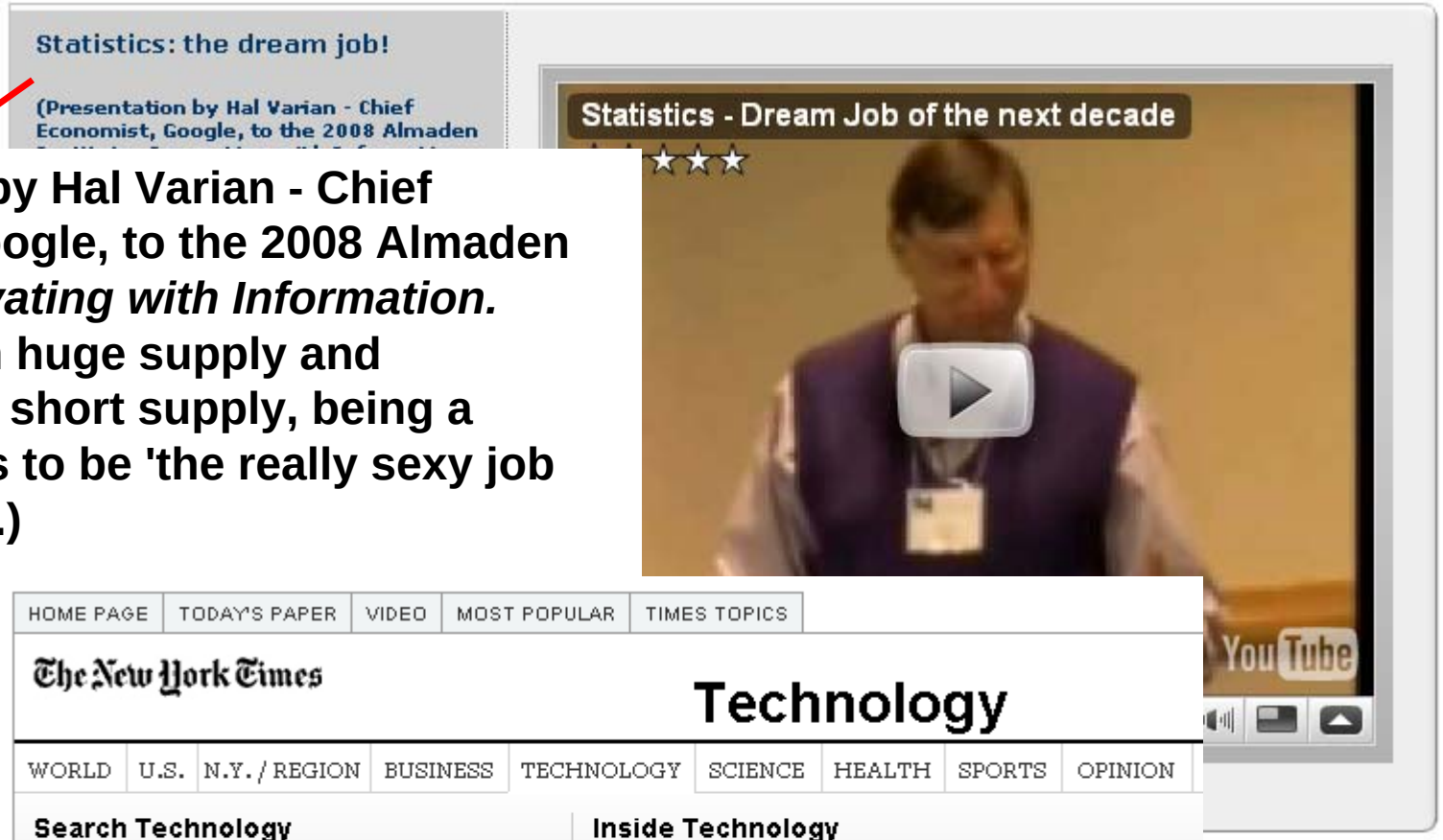
Conclusion

- Exciting times, amazing opportunities.
- Australia is well positioned compared to many countries because of
  - Nationalized health care system
  - Successful efforts such as in Western Australia
- Better and often specialized analytics needed
- Visualization needs more attention
- Need strong interdisciplinary teams combining
  - ICT
  - Statistics
  - Subject matter experts
- Maths and statistics education suffering at present – tragic given the need and opportunity

CSIRO

Statistics – the dream job!

**(Presentation by Hal Varian - Chief Economist, Google, to the 2008 Almaden Institute, *Innovating with Information.*
"... with data in huge supply and statisticians in short supply, being a statistician has to be 'the really sexy job for the 2010s'".)**



CSIRO

Louise Ryan
Phone: +61 2 9325 3203
Email: Louise.Ryan@csiro.au

www.csiro.au

Thank you

Contact Us
Phone: 1300 363 400 or +61 3 9545 2176
Email: enquiries@csiro.au Web: www.csiro.au

CSIRO