

From data to knowledge

Jim Ridgway
Durham University, UK

www.durham.ac.uk/smart.centre

Istanbul Declaration

“produce a broader, shared, public understanding of changing conditions, while highlighting the areas of significant change or inadequate knowledge”

Web 2.0: information explosion

- Search engines
- Government and NGO websites
- Communication tools
 - *YouTube* for video, *Clipmarks* to assemble webpages
- Interactive documents
- Discussion forums on data
 - *Many Eyes*, *Swivel*
- Wikis – on reasoning with evidence
 - interactive demonstrations, self test and diagnostics
- Mash-ups
 - e.g. *Google Maps* plus data on (war, water, mobility...) from 3G phones
- Information and misinformation explosion
 - astroturfing, *wikiscanner*

Web 2.0: opportunities

- Data driven, live updates of displays
- Interactivity
- Animation
- Sharing interface design e.g. [ONS population pyramid](#)

From problem to policy

- Problem definition
- Problem exploration:
 - data
 - what is relevant?
 - what is available?
 - is it reliable?
 - data exploration
 - modelling
 - theorising
 - informal
 - formal
- Policy formulation, implementation, revision

Implications...

You have to be able to DESCRIBE the phenomena before you begin (sometimes you might have to collect data)

DESCRIPTION brings you face to face with big statistical ideas – quality of data, experimental design, measurement error, interaction, effect size...

Challenge...

Every interesting problem in health, crime, poverty,
environment, education, personal well-being...

Challenge...

Every interesting problem in health, crime, poverty, environment, education, personal well-being...

- is multivariate

Challenge...

Every interesting problem in health, crime, poverty, environment, education, personal well-being...

- is multivariate
- has non-linear relationships

Challenge...

Every interesting problem in health, crime, poverty, environment, education, personal well-being...

- is multivariate
- has non-linear relationships
- has confounding variables

Challenge...

Every interesting problem in health, crime, poverty, environment, education, personal well-being...

- is multivariate
- has non-linear relationships
- has confounding variables

So we might have some problems developing 'public understanding'

And...

- Statistics is 'hard' (*is it?*)
- There is a need for a long 'statistical apprenticeship' learning to master difficult technique (*is this true?*)
- Understanding evidence is best left to experts, who are all nerds, so we can ignore them (!!!)

HOWEVER people like to make sense of things, and *do* use evidence...

Assertions

- Understanding multivariate data is not always hard
- Active exploration leads to better engagement, better understanding and better story telling

Two Strong Claims

Semi-qualitative descriptions can be very powerful, and can give more 'bangs per buck' than quantitative analyses of complex situations, for almost everyone

You get to big statistical ideas much sooner if you work

- top-down (look at data on teenage drinking)
- not bottom-up (practice t -tests on 2 strings of numbers)

SMART centre research

Understanding sense making based on evidence

- Rethinking 'statistics': *what? And where?*
- Interface design [STI](#), [calculating axes](#), [water fleas](#)
- Defining and describing 'new literacies':
 - what skills are critical for dealing with [\(mis\)information](#)?
 - are there hierarchies of knowledge?
 - what heuristics are useful?
- Misconceptions are?
 - diagnostic actions should be?
- Engagement with a variety of communities

Statistics to knowledge: examples

- [Alcohol mash-up](#)

Strong theoretical assumptions

- Millennium development goals

Conceptual and research issues

The development of both a semi-quantitative and semi-qualitative approach to understanding evidence

- For any situation:
 - what are the vices and virtues of applying quantitative methods?
 - what are the vices and virtues of applying qualitative methods?
 - how do we maximise understanding by using both approaches?

Promoting effective data visualization tools

- Creation and validation
- Common data formats
- Widespread use

Understanding 'new literacies'

- Tracking a moving target (PISA, PIAAC)
- Understanding user understanding:
 - what are the key heuristics?
 - can we 'teach' them?
- Revise conceptions of 'statistics':
 - media
 - NSOs
 - school and university
 - social science

Cultural impediments

- Disengaged citizens
 - Low educational attainment
 - Access to evidence:
 - very low web access in Asia, despite the hype
 - by STUDENTS – Pakistan (16%) Cambodia (24%) Sri Lanka (24%)
- source: Dhanarajan *et al* (2007)
- Misinformation:
 - astroturfing, *wikiscanner* on wiki editing

NSOs: valuable commentaries

- Descriptions of patterns in data
- Highlighting areas of significant change
- Highlighting areas of inadequate knowledge
- Saying when data are (are not) consistent with assertions by politicians, journalists...
- *Supporting particular models/theories/political positions is more problematic*

Time lines

Statistical offices	10-20 years
Ministers	1-2 years
Journalists	1-2 days (90 days)

Things we need to make progress

- Modelling systems (informally?)
 - players: politicians, journalists, NSOs, citizens...
- ‘Biological’ thinking – symbiosis, mutation, evolution
- Understanding ‘macrosystemic change’
 - revising practices
 - accepting plural time frames
- Communication with communicators
 - effective communication tools
- Progress indicators

Progress indicators

We are *FAILING* if...

- It is OK to say:
 - ‘My feeling about this is...’ (G. Bush, Senior)
 - ‘I can’t do mathematics’ (celebrity interviews)

Progress indicators

We are *SUCCESSING* if...

- MV data are used routinely:
 - in the media
 - in government policy documents
 - e.g. release of beta versions with interactive data displays
 - in one's private life

An example: visiting the doctor – evidence and access

Lifestyle change questions

- How big are the risks of doing nothing?
- If I want to change something, what should it be?

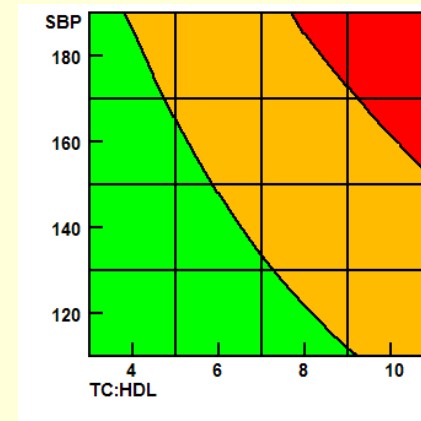
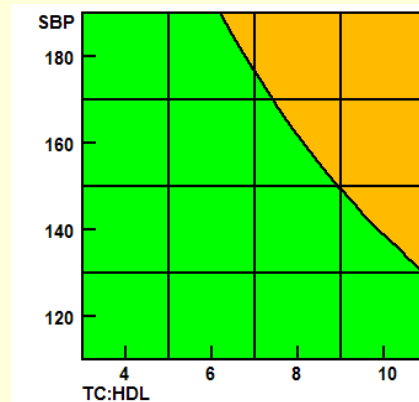
Male 'risk' as f (smoking, age, blood pressure, cholesterol)

Age

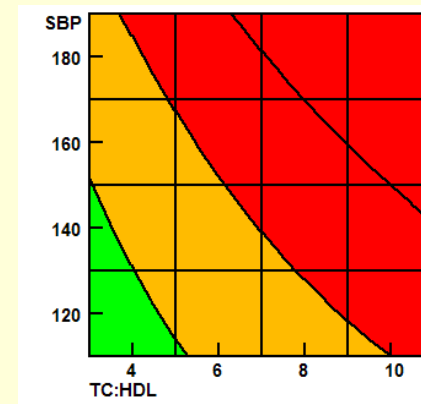
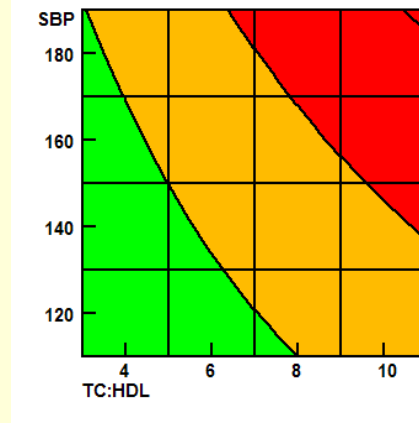
Non-Smoker

Smoker

35



45



Visiting the doctor – evidence and access

Coronary risk

Lifestyle change questions

- How big are the risks of doing nothing?
- If I want to change something, what should it be?

Access, empowerment

Conclusions

We CAN help people move from data to knowledge

ICT can support change

We are going to be redefining 'statistical literacy' for rather a long time...

From data to knowledge

Jim Ridgway
Durham University, UK

www.durham.ac.uk/smart.centre

Data references

- Alcohol data in *Drug use, smoking and drinking among young people in England in 2005*. Available at: <http://www.ic.nhs.uk/datasets>
- Pension annuity rates available at: <http://www.fsa.gov.uk/tables/>

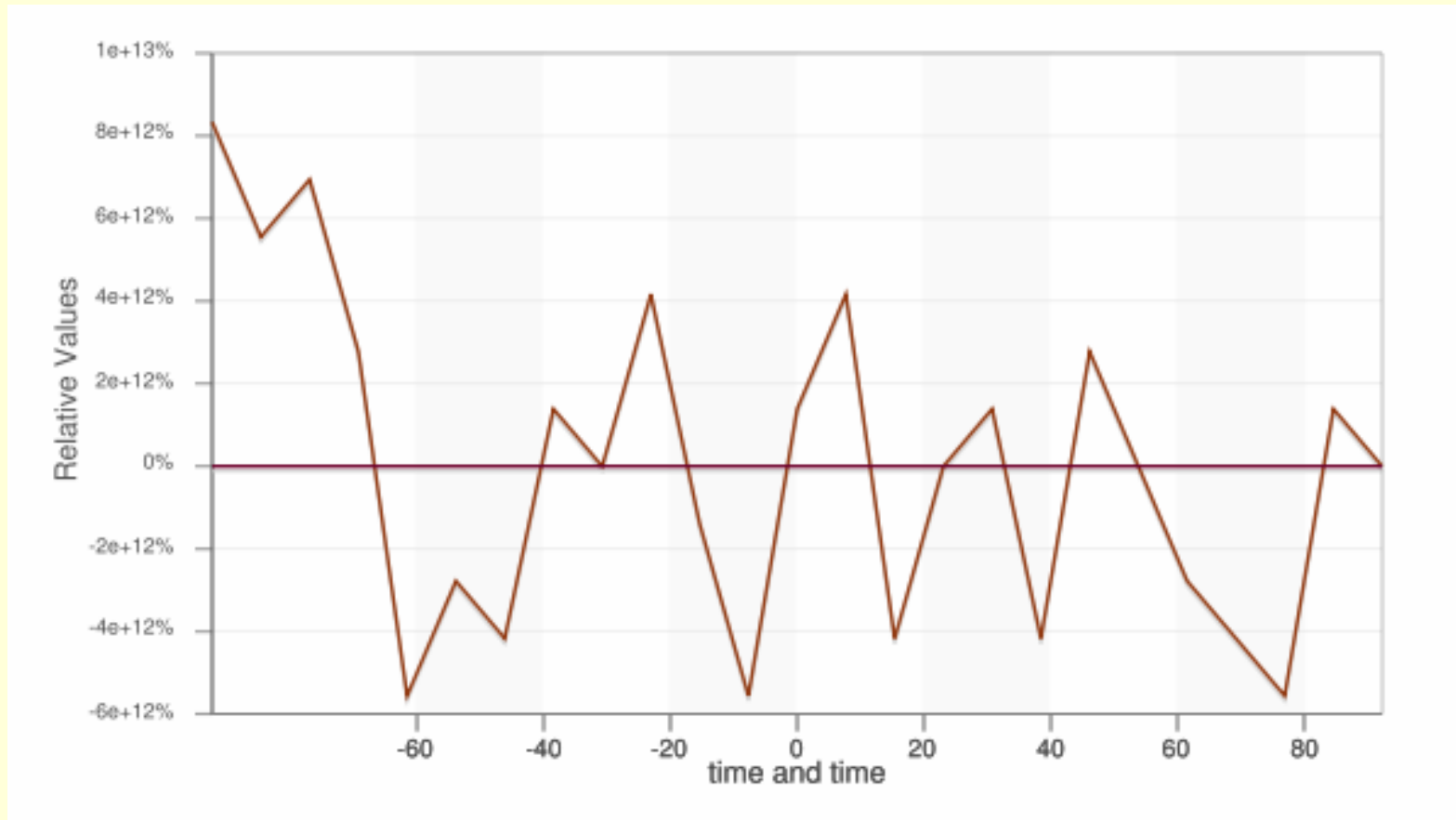
Data references

- STI data produced by the Health Protection Agency. Available at: <http://www.hpa.org.uk>
- Drugs data in report by ARK Northern Ireland. Available at: http://www.dhsspsni.gov.uk/drug_alcohol_use_among_young_people.pdf

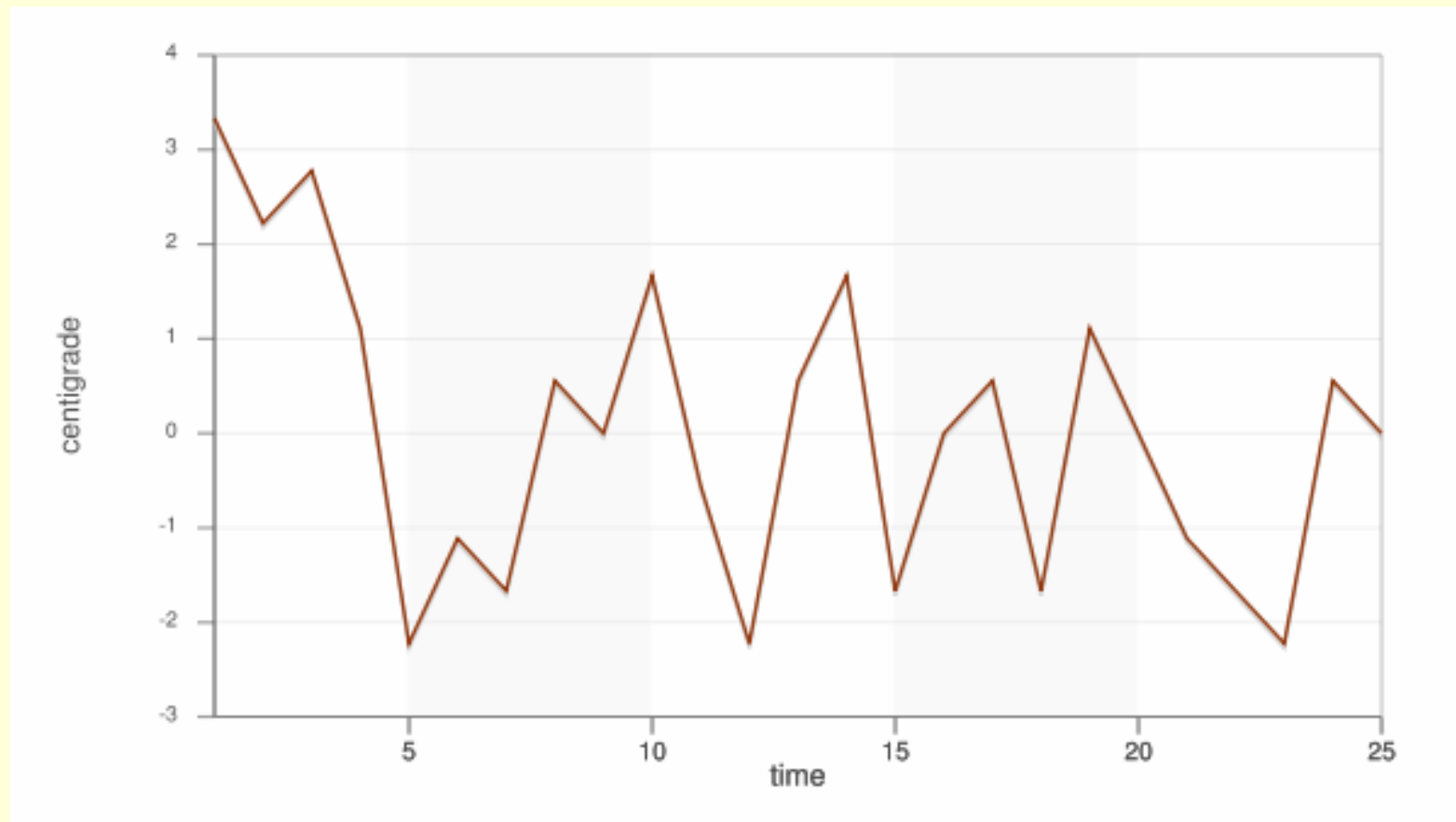
'All time' 12 most viewed graphs on swivel.com (12/5/2008)

	Title	Context	Duration	Hits	Comments
1	Atheists, Agnostics, Non-believers in God by Country	Part of population that are non-believers in God	~ 1 year	93347	1
2	净值 by 增幅	(Chinese text) context unknown	7 months	86598	0
3	Wine and Violent Crime	Wine consumption / U.S. violent crime	~ 1 year	55915	23
4	Growth of Creative Commons Photos on Flickr	Use creative commons photos on photo sharing site	~ 1 year	53504	40
5	<i>Atheists, Agnostics, Non-believers in God by Country</i>	<i>Same data as #1 but different format</i>	~ 1 year	49430	15
6	% Alcohol by Brand	U.S. domestic beer data	~ 1 year	37568	6
7	US GDP vs. Yearly Average Global Temp	U.S. GDP and global temperature	~ 1 year	32161	17
8	Cost of U.S. First Class Stamp (1885-2007 Est.)	U.S. postage prices	~ 1 year	28098	28
9	Apple Computer Daily Stock Price vs Temp	Stock price vs date (promoted as temperature)	~ 1 year	27740	8
10	U.S. Music Sales, 1975-2005: Vinyl, cassettes, CDs	U.S. music sales, by type	~ 1 year	26998	8
11	Sara McLachlan's \$150,000	Charitable donations by pop artist	9 Months	26048	1
12	Extreme Temperatures Make For Electricity Revenues	Changes in electricity cost, with temperature against date	~ 1 year	25985	6

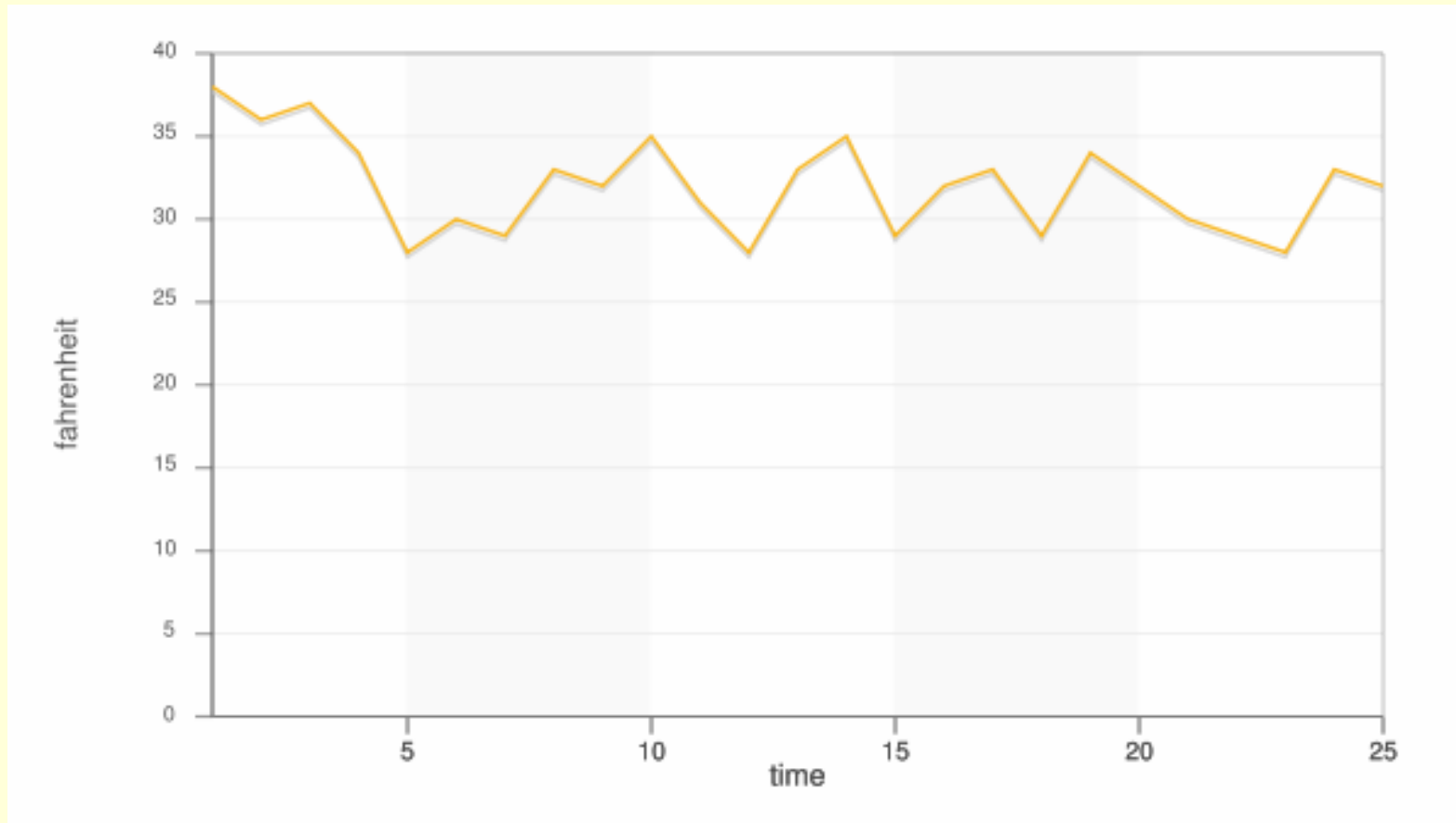
Temperature: where are the data from?



Centigrade: raw data



Fahrenheit: raw data



Heuristics

- Critique the quality and source of the data
- Describe and explore phenomena before you try to explain things
- Separate analysis and interpretation – especially in observational data

Heuristics

- Focus on effect size not significance level
- Check that the effect size is a lot bigger than the likely error of measurement
- Identify variables that have the strongest effects
- Look at absolute levels – are they big enough to be worth worrying about?
- look for the ‘dog that didn’t bark’ – were there things you expected to see, but didn’t?

Heuristics

- Explore the effects over different values of each variable: look for different functional relationships over different values of a variable (if it is cold, nothing happens...then as it *gets warmer...*)
- Look for non-linear relationships
- Look for interactions, and think about 'data surfaces'
- Disaggregate the data, and see if the patterns of relationships stay the same as in the aggregated data
- Think about possible confounding variables outside the variables being analysed