Imperial College London

WINTON

But that’s not how I see it

“horse” constructs and what we want to know

David J. Hand

Imperial College London

and

Winton Group

HORSE 2017 Conference 20 Sept 2017

What’s the next number : 2, 4, 8, ... ?

16 ?

The sequence is $2^1, 2^2, 2^3, 2^4, \ldots$ that is, 2^i with i consecutive integers

32 ?

The sequence is $2^1, 2^2, 2^3, 2^5, \ldots$ that is, 2^i with i consecutive primes

Given a set of data, more than one model will fit it

But there are implications and subtleties

More than one model fits the data

As Sturm 2016 put it:

"There can be many ways to reproduce the ground truth of a dataset"

Choosing the best-fitting model may not be sensible, if there are other models which provide a fit which is almost as good

The *flat maximum* effect:

"Even within the context of classifiers defined in terms of simple linear combinations of the predictor variables, it has often been observed that the major gains are made by (for example) weighting the variables equally, with only little further gains to be had by careful optimization of the weights"

Hand (2006) Classifier technology and the illusion of progress.
Statistical Science, **21**, 2-14.

In particular, consider two weighted sums of *d* variables

$$w = \sum w_i x_i \quad \text{and} \quad v = \sum \left(1/d\right) x_i$$

where $w_i \geq 0 \quad \sum w_i = 1$

Then $r\left(v,w\right) \geq \sum_i r\left(x_i, x_k\right) \Big/ d$

where $k = \operatorname{argmin}_j r\left(x_i, x_j\right)$

"The correlation between an arbitrary weighted sum of the predictors (with weights summing to 1) and the simple combination using equal weights is bounded below by the average of the smallest entries in the rows of the correlation matrix of the predictors."

To be a horse it must use a model you did not consider; or answers a question you did not intend

We need to ***test*** our hypotheses:

Adversarial systems

The fundamental *falsifiability* principle of science

Beware confirmation bias:
the tendency to interpret new evidence as confirmation of one's existing beliefs or theories

We must be able to ***explain*** our thought processes

- not enough to get a prediction right
  - e.g. the person who feels lucky and wins the lottery

This is a problem for basic ML systems

because they don't have any thought processes

don't have any sequence of deductive steps

They use pattern matching

(though it might be disguised as a sophisticated neural net, random forest, SVM, etc)

They don't have an implicit causal theory

The classical perspective on statistical models

(S,P)

S – the sample space

P – a set of probability distributions on S, which contains some distribution which closely approximates the "true" distribution

Models are often parameterised by indexing P by a parameter set Θ

(A Bayesian model also requires a prior distribution on Θ)

Elaborate mathematical extensions of meaning of "model"

e.g. McCullagh, 2002

Extensions to nonparametric models

the number and nature of parameters not fixed in advance, but determined by the data

e.g. kernel regression

e.g. ensemble models

The important question is

what is the model for?

Main uses:

- to make inferences: about unobserved cases
  - to the future
  - to other cases drawn from the same population
- to make inferences: about mechanisms
  - about the underlying population
  - causal inferences
- summaries, for ease of comprehension
  - about the characteristics of a collection of data

Clearly the ***mathematical*** definition above is ***not adequate*** for all purposes

For example, sometimes,

instead of ***a sample*** from a population,

we have ***the entire population***

e.g. data on all the countries in the world

Does this mean we can't build statistical models?

That statistics, machine learning, data mining, etc, are irrelevant?

Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. ***Indeed, they don't have to settle for models at all***

Chris Anderson

Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough

Chris Anderson

That's right – ***as far as it goes***

But it does not go far enough

Need to distinguish between

- substantive models

- empirical models

HORSES are particularly a potential issue with ***empirical*** models

Substantive models:

theory-driven, mechanistic, ...

Based on some (simplified representation of a) theory about the underlying mechanism

Use statistical methods to determine precise form

- estimate parameters
- fit model to data
- ....

Models as *approximations to something* ***underlying***

- that is, to a "truth"

All models are wrong, but some are useful

George Box

[Can only be "wrong" if there's a "right" ?]

The unicorn, the normal curve, and other improbable creatures

Theodore Micceri, 1989

Normality is a myth; there never was, and never will be, a normal distribution

Geary, 1947

Then goes on to say "***This is an over-statement from the practical point of view***"

Empirical models:
data driven, descriptive, ...

- Aim to summarise, identify, extract the relationships in data

- No *substantive* theoretical base will have *statistical, mathematical, analytic* theoretical base

Search through variables, arbitrary transformations, optimisation, ...

Overfitting issues; regularisation; ...

Commonly used for *prediction*

Parametric and nonparametric methods

- kernel regression
- ANNs
- classification trees
- random forests
- ensemble methods
- deep learning
- ....

Example: ***Substantive***

Model the relationship between the height from which a stone is dropped and the time it takes to hit the ground

e.g. drop stones from various heights

→ Data $\left(H_i, t_i\right) \; i = 1, \ldots, n$

→ Model From *Newton's Laws*

$$t = \sqrt{2H/a} + \varepsilon \qquad H = a\left(t + \varepsilon\right)^2 \Big/ 2$$

Use statistical methods to estimate *a*

Example: ***Empirical***

Logistic regression model for the probability that someone will purchase a product

based on the values of their characteristics

$$x_1, x_2, \ldots, x_d$$

$$\log\left(\frac{p}{1-p}\right) = \sum_{j=1}^{d} \beta_j x_j$$

Models may start as empirical and then become substantive

This follows the classic picture of science in which

- we begin by collecting data and noticing regularities
- and then formulate theories which encapsulate those regularities

A comment on interpretability

Interpretability is reassuring:

the model “makes sense”, “is reasonable”, etc

Good for convincing people

- customers
- managers
- people who have to use the model

Substantive models, by definition, have an interpretation

Simple empirical ones also

And in many situations interpretability is a driver behind such models

e.g. weighted sums of predictors

But what is meant by 'interpretability'?

In many contexts: *how important is variable v_i to the model?*

For elaborate models (e.g. ANNs, RFs, etc) compare the predictive power of the complete model with the model without v_i

Not a sharp distinction

e.g. Regression

Empirical because you are aiming for a simple interpretable predictive model ?

"the increase in y for unit increase in x, holding z constant"

$$y = \alpha + \beta x + \gamma z + e$$

is not always meaningful:

$$y = \alpha + \beta x + \gamma x^2 + e$$

"How y responds, on average, to change in x1, after allowing for simultaneous linear change in x2, for the data at hand"

Terence Speed

Substantive:

- because you think it is linear
- or because you regard it as the first term in a Taylor series

"Treatment A is better than treatment B"

Could be a simple substantive model, or a mere empirical description

e.g. Cluster analysis

The horse problem is really one of finding a partition of the space which captures the concepts you are interested in, while not capturing other concepts.

There are many possible partitions given a finite data set

Dissection vs partitioning

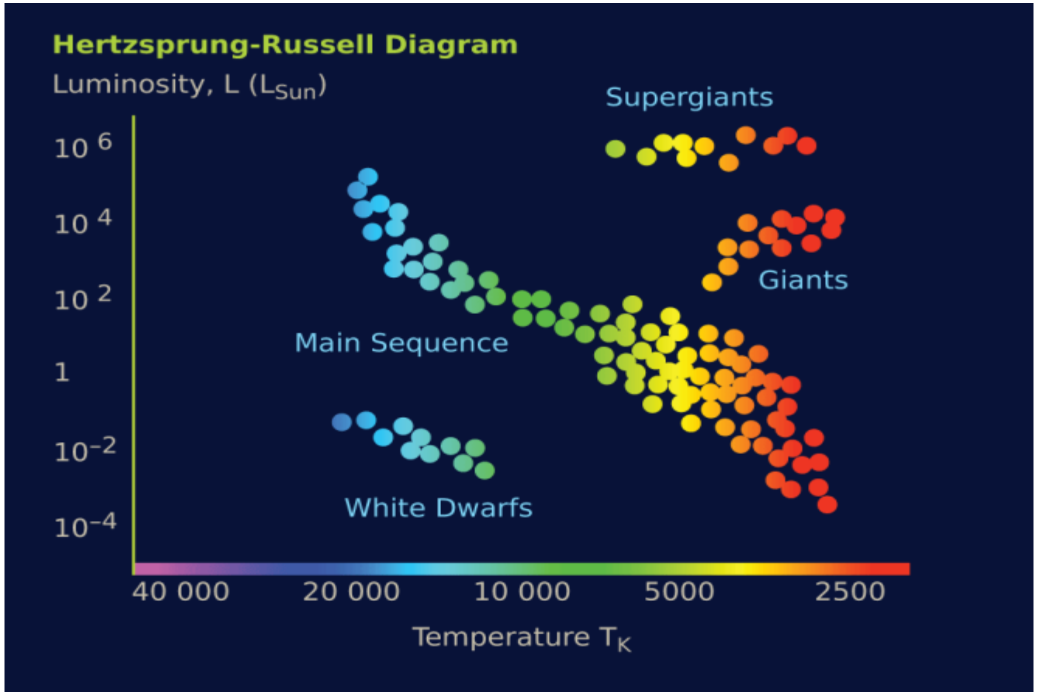
Substantive vs empirical

Dissection:

- carving nature at the joints
- finding natural clusters

  e.g. bipolar vs unipolar depression

  e.g. Herzsprung-Russell diagram for stars



Partitioning:

- finding a convenient way to split the data

e.g. 1: in an advertisement in the *Sunday Times* of 18th April 1999, James Meade Limited, a shirt manufacturer, gave a choice of sizes, as follows, where a **X** denotes sizes available, and **O** denotes standard size

			Sleeve length (inches)				
	31	32	33	34	35	36	37
Collar size							
14½	✓	✓	**O**	✓	✓	✓	✓
15	**X**	✓	**O**	✓	**X**	✓	✓
15½	**X**	**X**	**O**	**X**	**X**	**X**	**X**
16	✓	**X**	**X**	**O**	**X**	**X**	**X**
16½	✓	✓	**X**	**O**	**X**	**X**	**X**
17	✓	✓	**X**	✓	**O**	✓	**X**
17½	✓	✓	✓	✓	**O**	✓	**X**
18	✓	✓	✓	✓	✓	**O**	✓

Weaknesses *of substantive models*

Need a theory
OK in theory rich domains
Less so in other domains

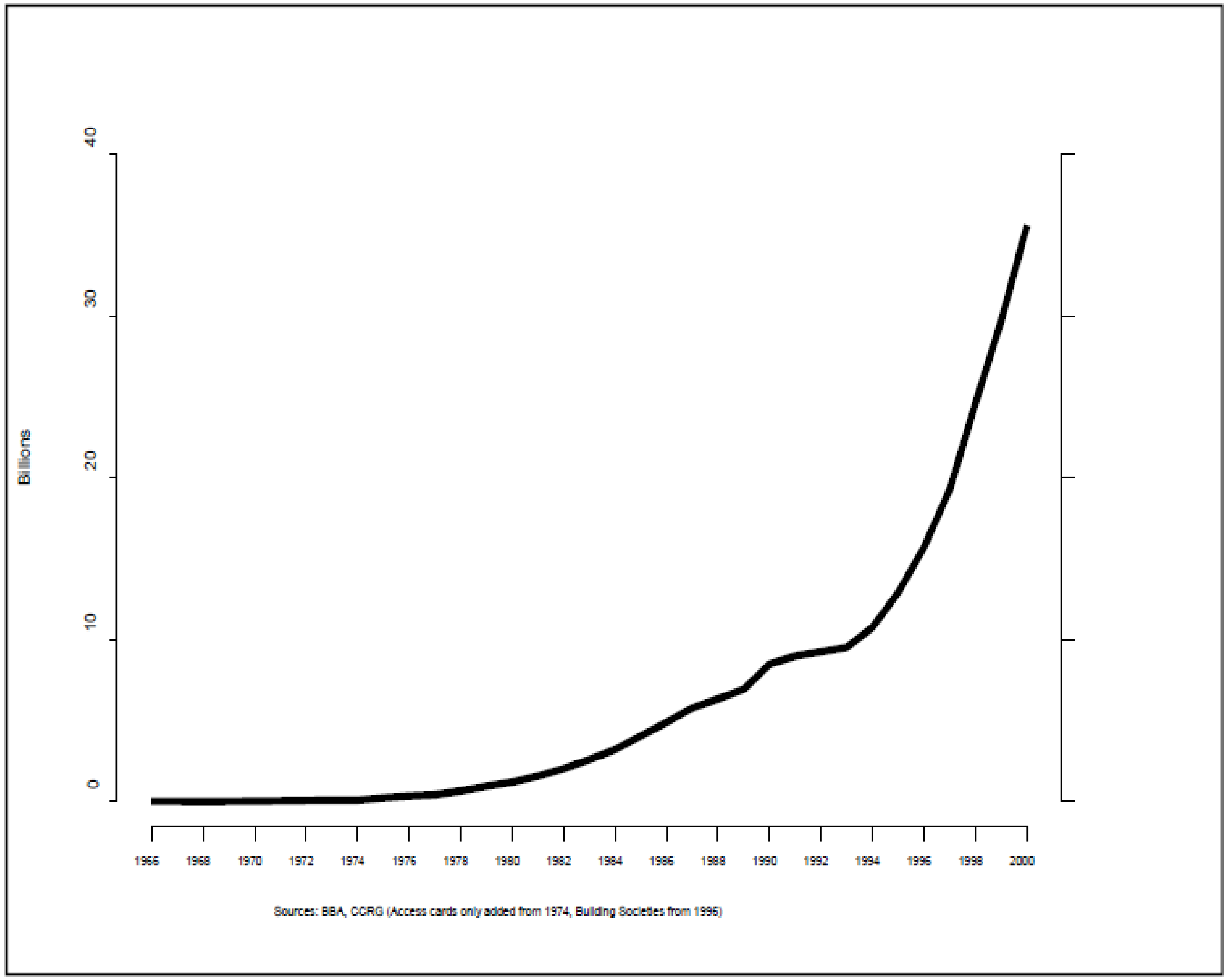
Bias, if you get theory wrong

Weaknesses *of empirical models*

Assume:

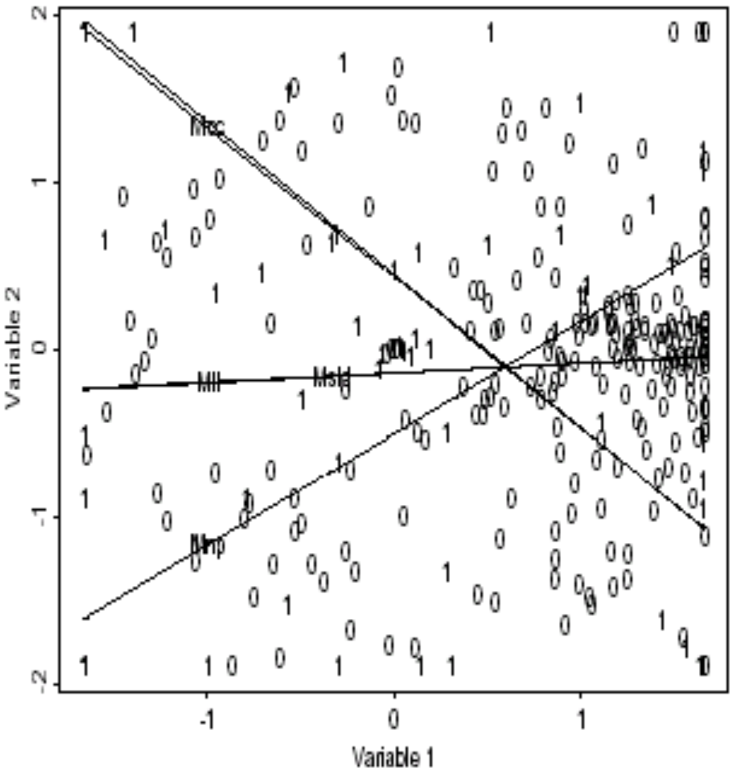
- *the future is like the past*
- *choose criterion to fit model to data*
- *good quality data*
- *no selection bias*
- *no gaming, feedback, etc*
- *...*

The future is like the past



Sources: BBA, CCRG (Access cards only added from 1974, Building Societies from 1996)

Choose criterion to fit model to data



Ionosphere data, Benton (2001)

Optimum error rate: top-left to bottom-right

Optimum Gini: bottom-left to top right

“a 'horse' is just a system that is not actually addressing the problem it appears to be solving”

Sturm, 2014

Example 1: I want to conveniently summarise a data set in terms of an "average" value (a "location")

Should I use the mean or the median?

Different statistics have different properties

e.g. it's possible for one of two groups to have a higher mean but a lower median than the other

	{1,3,5}	{1,2,9}
mean	3	4
median	3	2

Mean sensitive to values of handful of extreme points:

Can make the mean as large as you like by increasing the largest value, while the median stays unchanged

So you often see statements like:

The choice between mean and median should depend on how the data are distributed. The mean is a better measure of "location" than the median for symmetric distributions, but otherwise the median should be used

But this is ***wrong***

It makes no reference to the question being asked

That is, to what you want to know

Suppose I randomly choose the remuneration of a new recruit from a very skewed distribution of salaries

What interests ***me*** is the ***mean*** of the distribution:
my total wage bill is the product of the mean and the number of employees

What interests ***a potential new recruit*** is the ***median***:
to her the mean is of no interest, since she's almost certain to receive substantially less than that

The choice between mean and median depends on what you want to know

1994 Baseball players strike in US

Example 2: Simpson's paradox

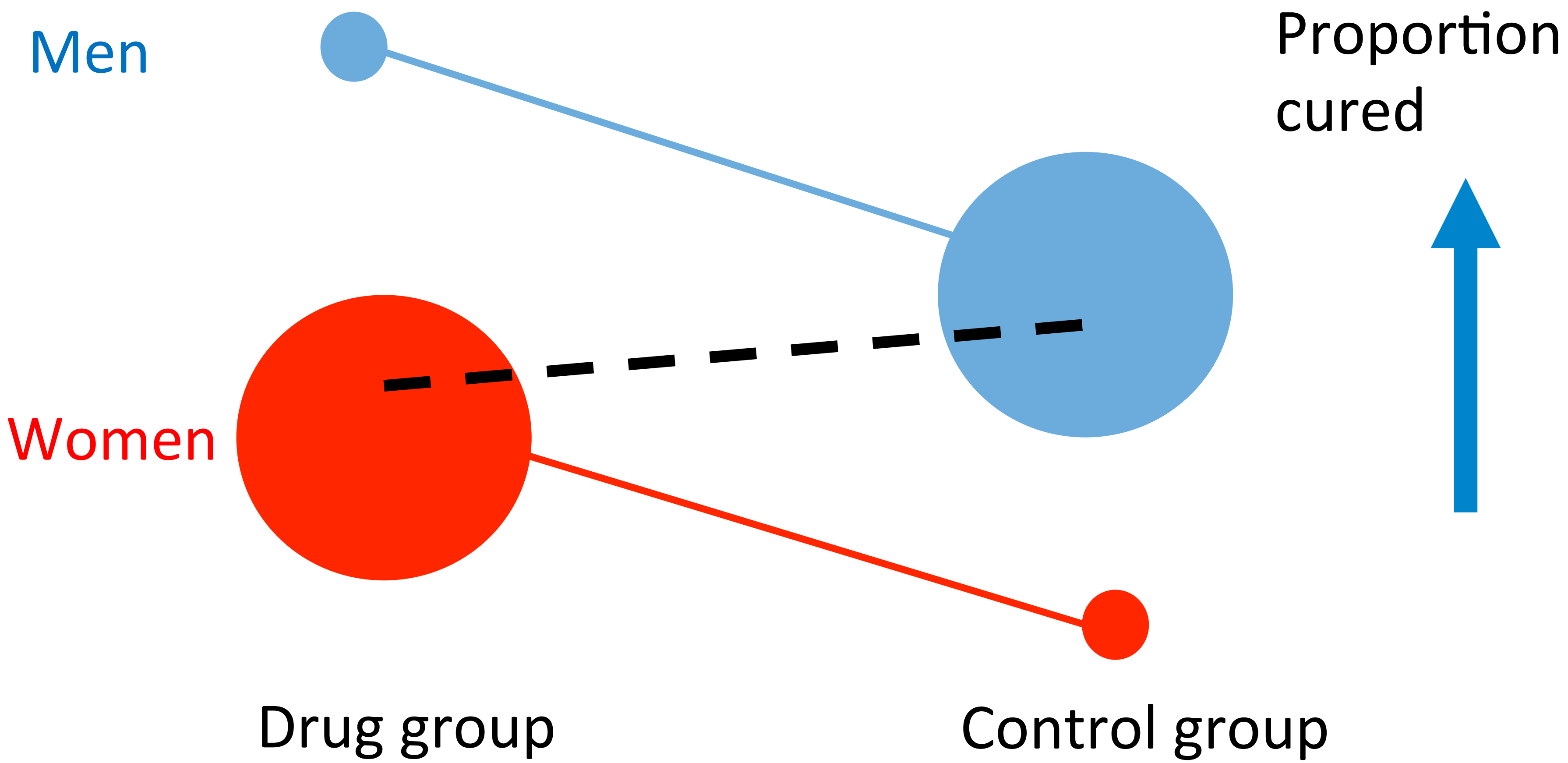
Proportions of patients who recover

	Drug			No drug	
	273/350	= 78%	<	289/350	= 83%

But split by gender

Men	81/87	= 93%	>	234/270	= 87%
Women	192/263	= 73%	>	55/80	= 69%

Disk size shows number in the group



Men

Women

Proportion cured

Drug group

Control group

Example 5: Diagnosis vs screening

Method: construct a predictive classification rule (e.g. logistic regression, ANN, random forest, …) to assign new cases to appropriate class

- ***diagnosis***: assign an individual to the class to which they have the greatest probability of belonging

- ***screening***: identify those members of the population most likely to belong to each of the classes

Diagnosis

- If $p\,(0|x)>p\,(1|x)$ assign to class 0
- Maybe *all* cases are assigned to class 0

Screening

- How many others are assigned to class 0 ?
- If $p\,(0|x)>p\,(1|x)$ do not necessarily assign to class 0
- Rank and assign top *q*% to class 0

“When I look at you as an individual, I think you have disease 0, but when I look at you in the context of others, independently sampled, I think you have disease 1”

Example 6: Discrimination

UK Equality Act, 2010:

> “***A person (A) discriminates against another (B) if, because of a protected characteristic*, A treats B less favourably than A treats or would treat others.***”

Informally this means treating people differently on the basis of their **group membership** rather than on the basis of ***their own merits***

e.g. Person A is female, therefore treated as lower risk than B, who is male

May be true on average, but perhaps not true for A and B in particular

[*Sex, race, religion, ……]

Credit scoring

Canonical example: construct statistical model to predict probability that someone will default on a loan

Predictors:

- demographic characteristics
- employment details
- living arrangements
- financial circumstances
- etc

But it would be illegal for the models to make decisions based on protected characteristics

Solution: omit the protected characteristics from the model

Problem: other variables will be correlated with the protected characteristics and can act as proxies

Solution: drop proxies as well

Problem: leads ultimately to dropping all predictive information

Insurance

EU Gen EU Gender Directive (2004/113/EC) sought to counter discrimination based on gender

But it included an opt-out clause in Article 5(2), which permitted "***proportionate differences in individuals' premiums and benefits where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data***."

> e.g. allowed different driving insurance premiums for males and females

All 26 countries took advantage of this clause

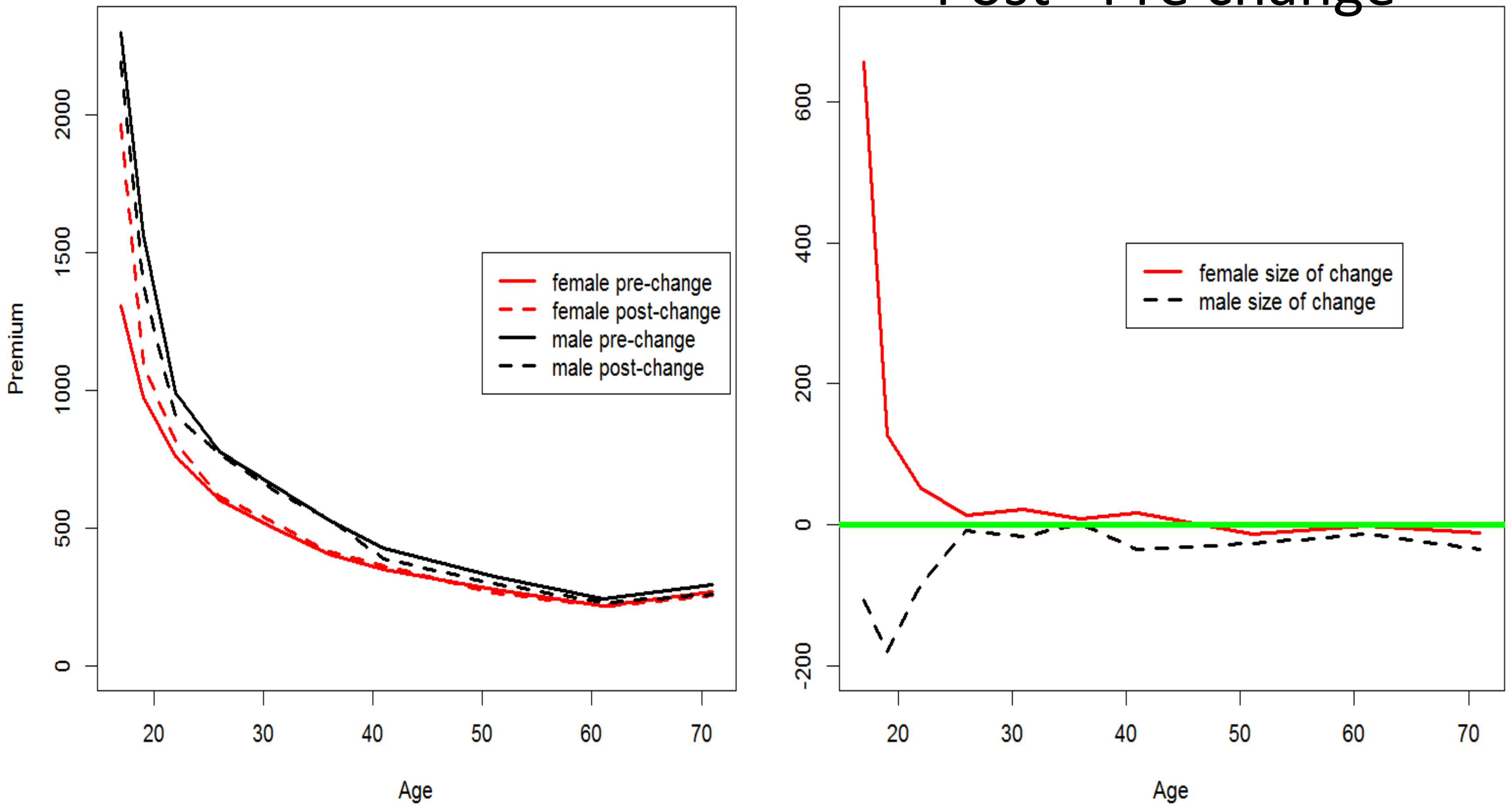
13 of them, including the UK, for all types of insurance

2008, law suit claiming that the opt-out was incompatible with the principle of equal treatment for men and women was brought to the Belgian Constitutional Court

The European Court of Justice judgement of 1st March 2011 concluded that the Article 5(2) ***opt out was to be treated as invalid*** from 21st December 2012

That is, from that date, it was illegal to have differential insurance premiums based on gender

e.g. females previously had lower motor insurance premiums, and these differences would no longer be allowed, ***even though the data showed risk females to be safer drivers***


Post - Pre change

£	Men	Women
Before	658	488
After	619	529

This drives a sample selection process

- more young men on the road
- since can now more readily afford the lower premiums
- ***i.e. more of the riskier drivers on the road***

- fewer young women
- since they cannot afford the higher premiums
- ***i.e. fewer of the safer drivers on the road***

Is this good for society?
Is it "fair" or "equal treatment"?

Leading to a death spiral:
More risky drivers, fewer safer drivers
Premiums have to go up
Safer drivers (not getting value for money) drop out
Accidents per driver increase
Premiums go up

Sometimes horses can be advantageous:

- *AlphaGo* seeing patterns which not spotted by humans

- Coming up with new scientific theories

Creativity and innovation are the other side of horses

Horses can arise from

- systems focusing on features you don't expect
- or from ambiguous definitions of the problem

Particular risk with ML systems and empirical models

But the flip side is discovery, innovation, and creativity

thank you