Working with Data Scientists

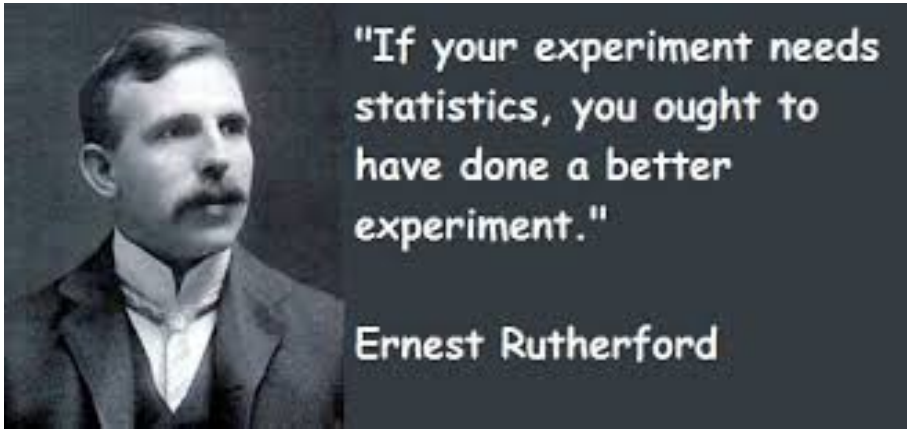
one biostatistician's experience

Peter J Diggle

CHICAS, Lancaster University Medical School

June 2019

Data science: more data = more information?

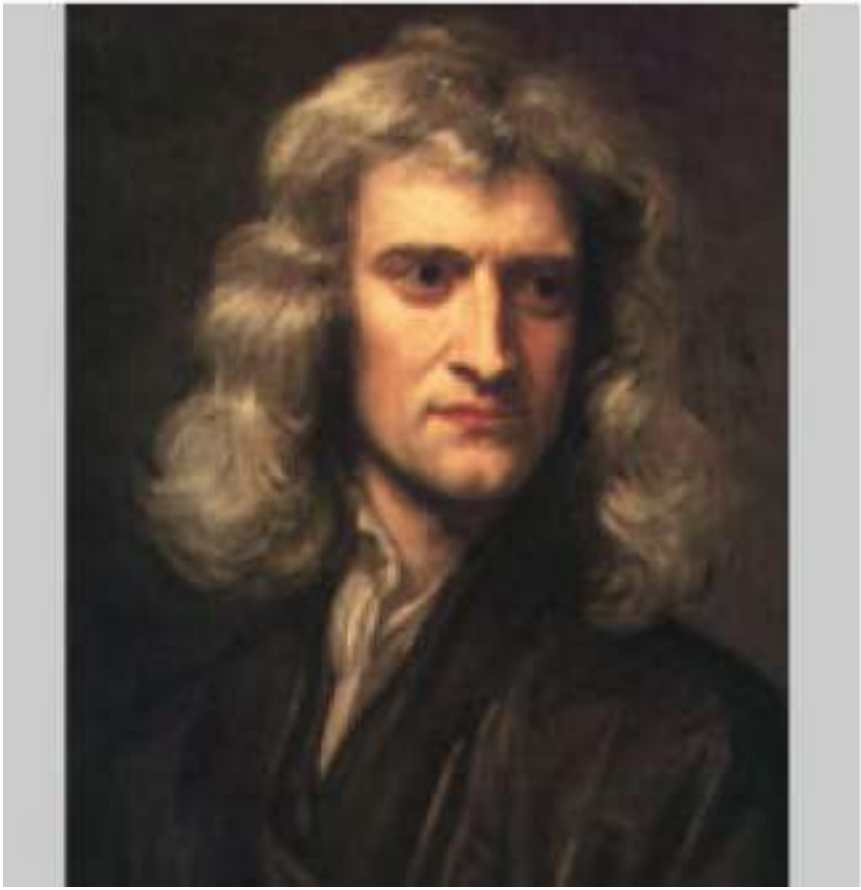


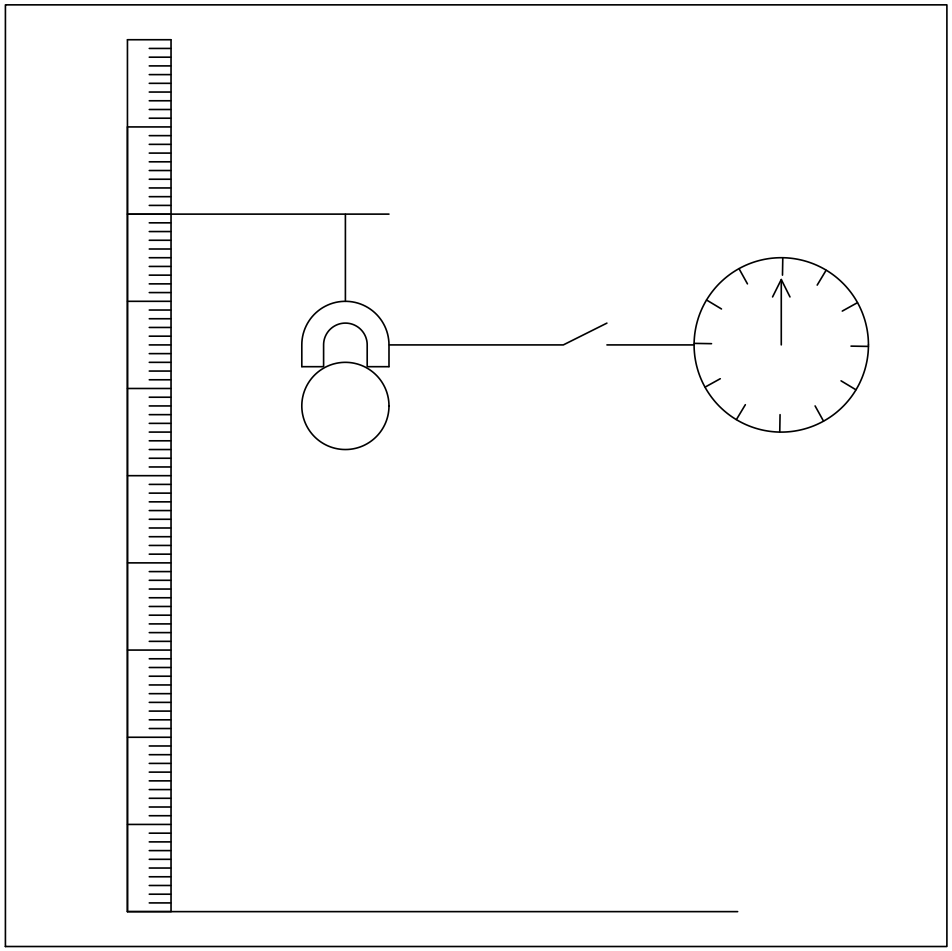
And who better to design that experiment than a statistician?'

"You are smarter than your data. Data do not understand causes and effects, humans do."

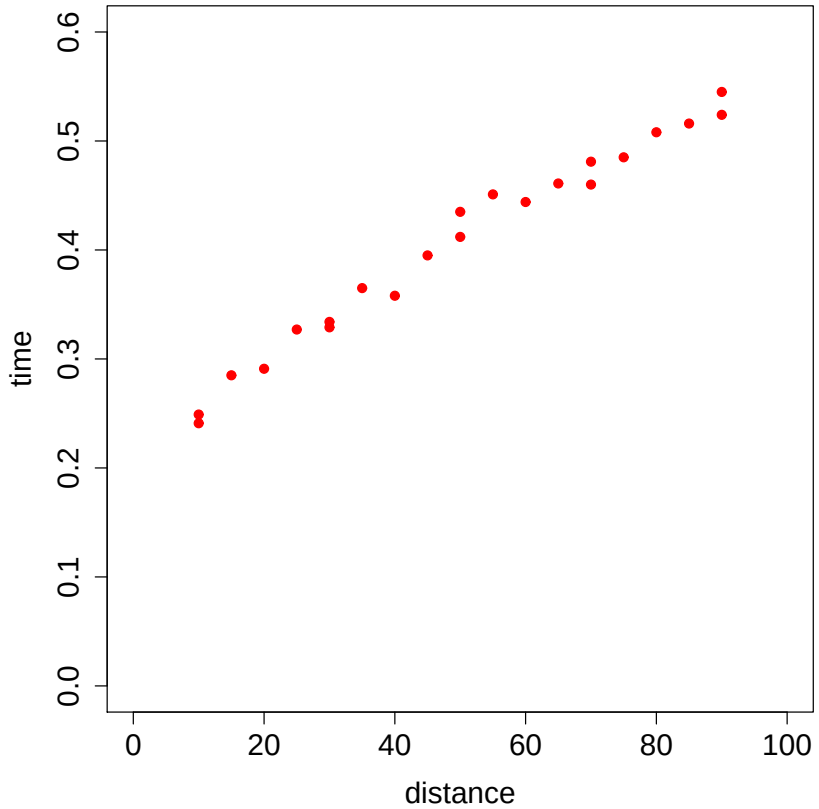
Judea Pearl

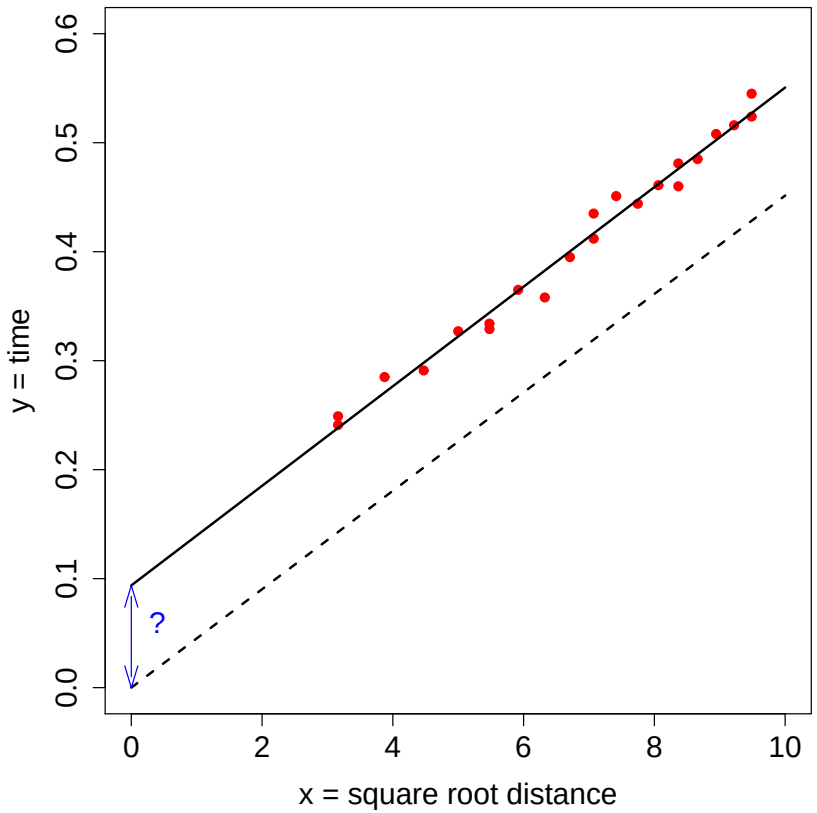
Isaac Newton (1643–1727)





One student's results





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

Statistical modelling principles

- models are devices to answer other people's questions
- models should:
  - be not demonstrably inconsistent with the data;
  - incorporate the underlying science, where this is well understood
  - be as simple as possible, within the above constraints

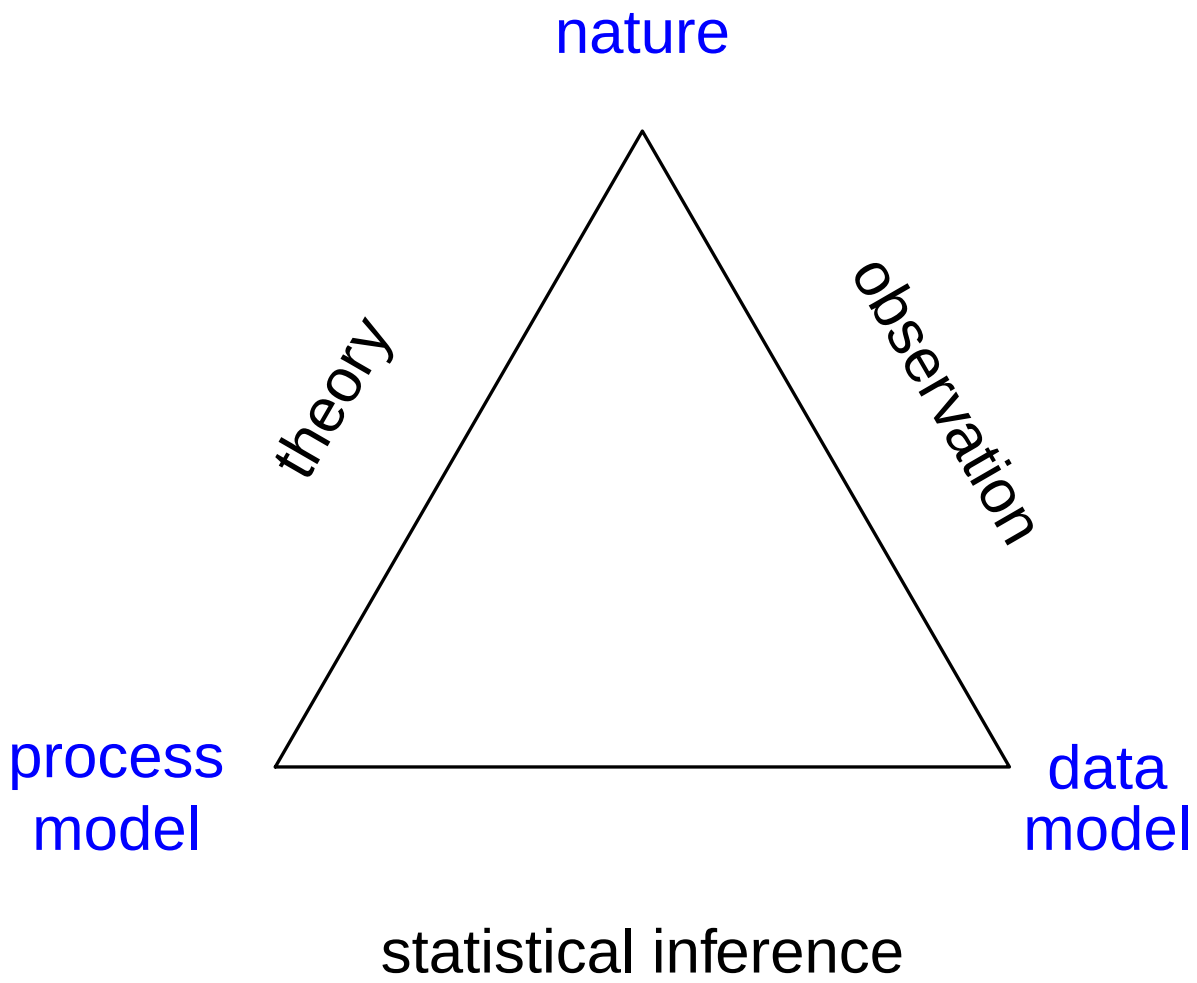
"Too many notes, Mozart"

Emperor Joseph II

"Only as many as there needed to be"

Mozart (apochryphal?)

The Science Triangle



Definitions

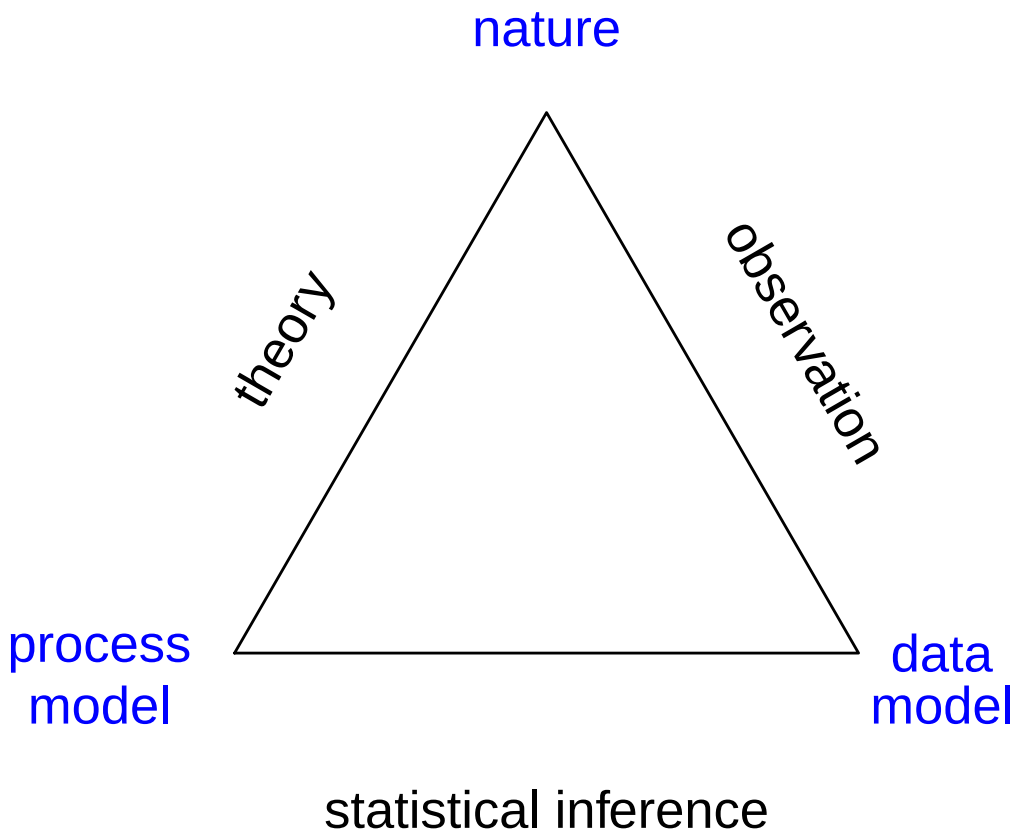
Wikipedia

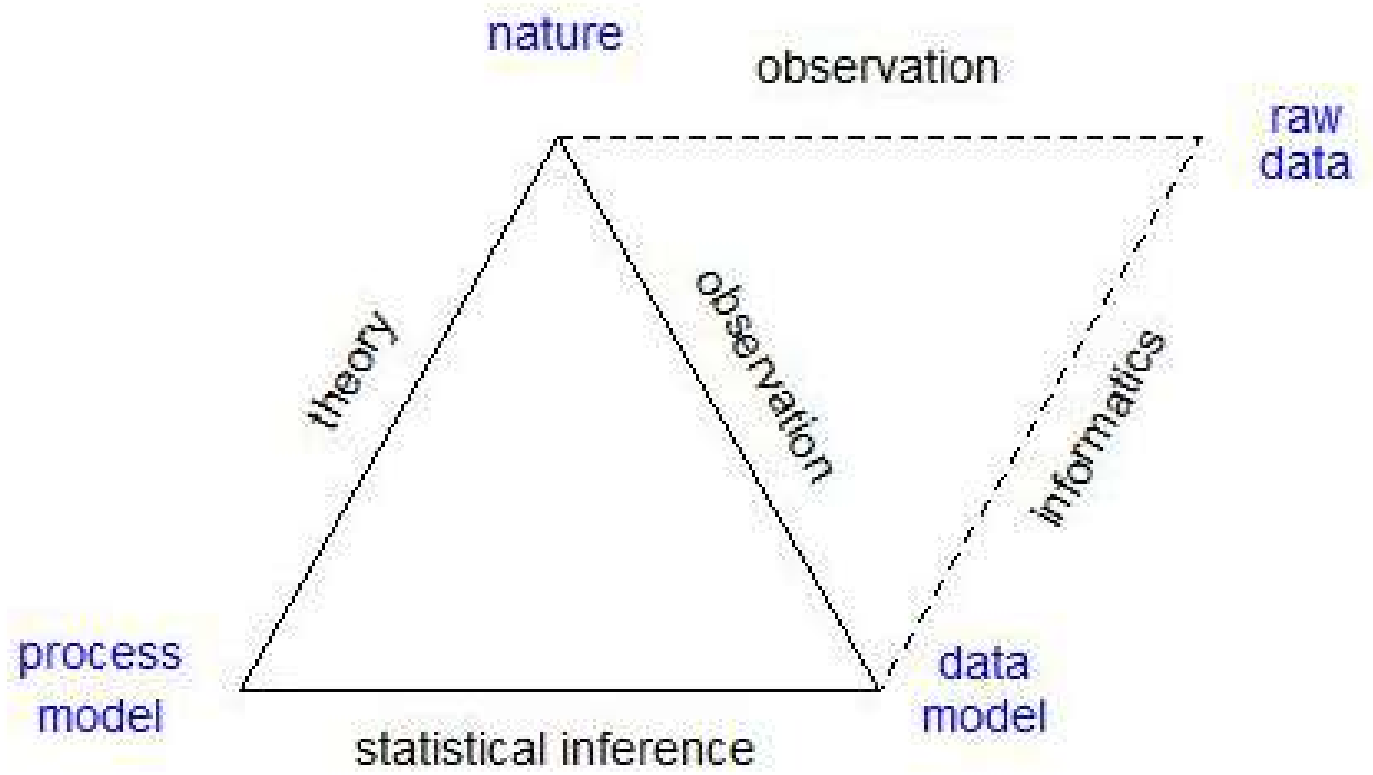
- **Data science is...the extraction of knowledge from data... It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, and information technology...**

PJD

- **Statistics is...the extraction of knowledge from data...**
- **Data science is Statistics + Informatics + Science**

The Science Triangle





"Informatics seeks to maximise the utility of data, statistics seeks to minimise the uncertainty associated with data"

Iain Buchan

What can we offer?

- **that probability theory is the correct way to deal with uncertainty**
  - **in our data ... stochastic models**
  - **in our conclusions ... probabilistic inference**
- **that design matters**
- **that context matters**

And what can we learn?

- **that a published article is not a complete solution to a practical problem.**
- **that reproducibility of computationally driven research findings should be a minimum standard**

Role of Statistical Method in Data Science

DATA = SIGNAL + NOISE

- **especially true of observational data**
- **stochasticity is the statistician's honesty box**

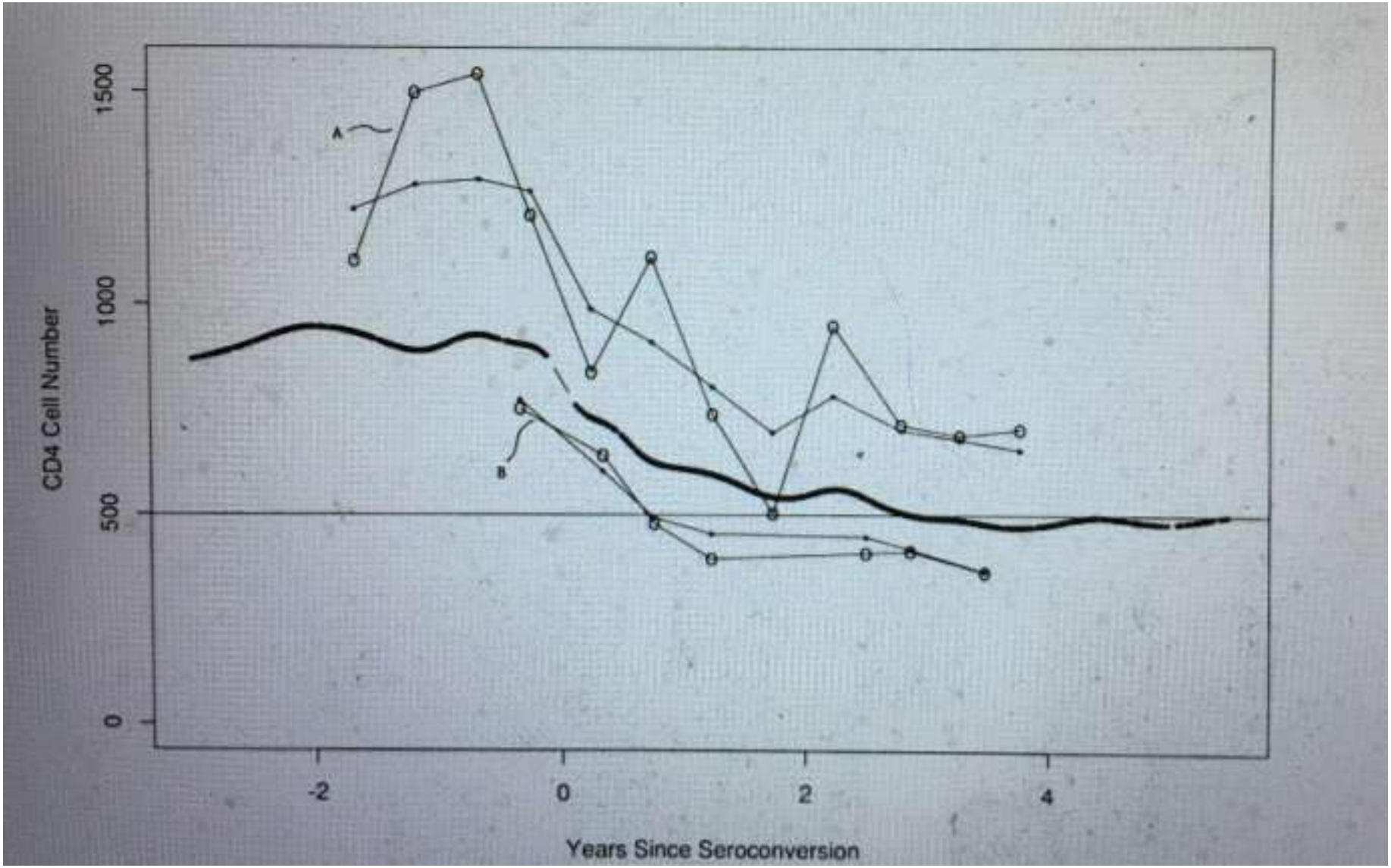
"Better an approximate answer to the right question than a precise answer to the wrong question"

John Tukey

"The answer to any prediction problem is a probability distribution"

Peter McCullagh

An old example...the early years of the AIDS epidemic



Clinical guideline ca 1989: **initiate AZT therapy when $CD4 < 500$**

What kind of statistics is needed?

- **modelling observational data**
- **describing and predicting variation in time and/or space**
- **real-time analysis to inform decisions**
- **off-line analysis to inform policies**

Early detection and treatment of kidney failure

Diagnosis

- Serum creatinine $\Rightarrow$ estimated glomerular filtration rate

$$\text{eGFR} = 186 \times \left(\frac{\text{SCr}}{88.4}\right)^{-1.154} \times \text{age}^{-0.203} (\times 0.742 \text{ if female})$$

- progression can be asymptomatic for many years
- SCr easy to measure from blood-sample, but noisy
- early diagnosis and intervention can slow rate of progression

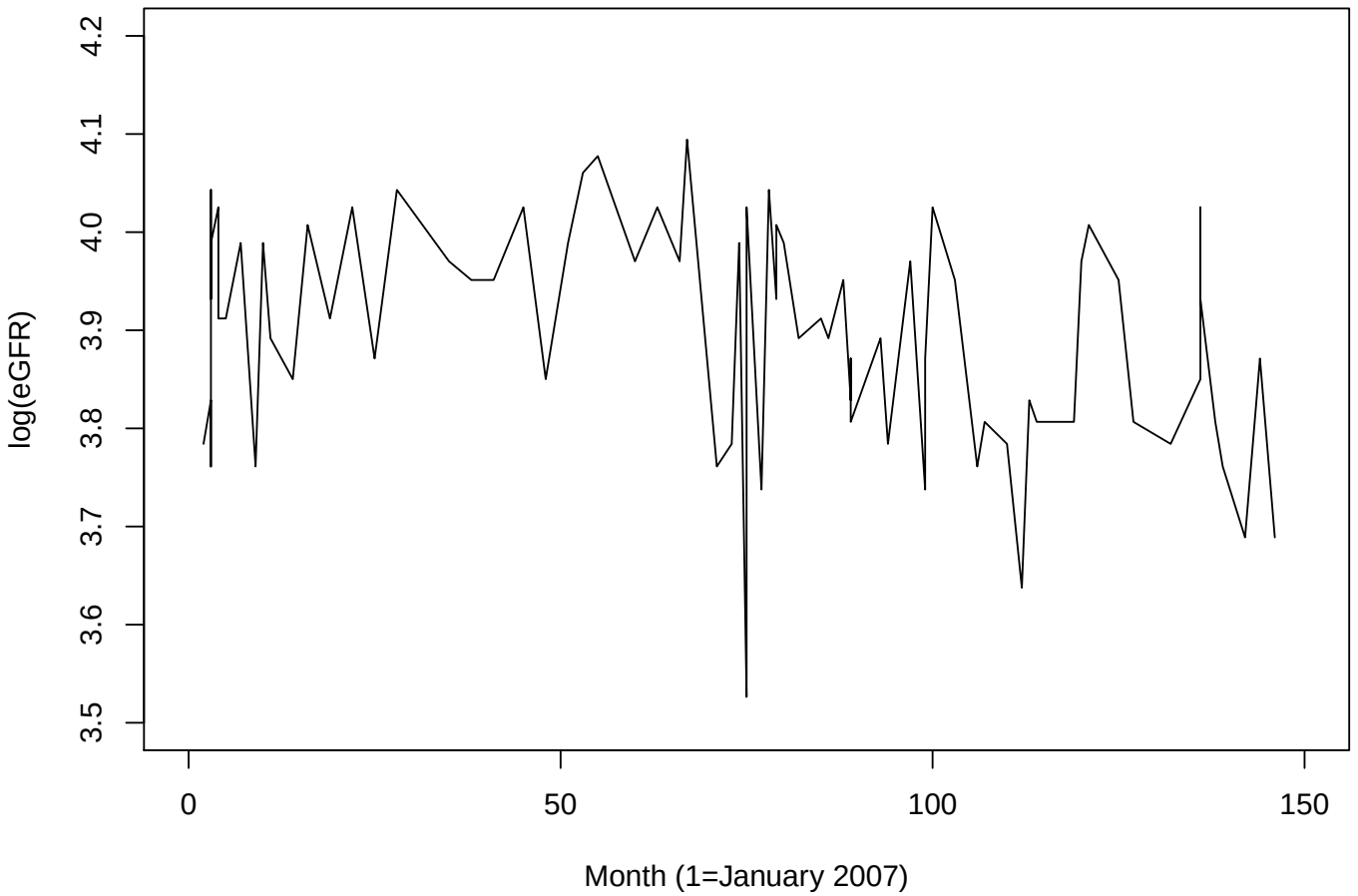
Clinical guideline

- Loss of kidney function $> 5\%$ per year $\Rightarrow$ consider referral to specialist secondary care

Predictive target

$$\frac{d \log GFR(t)}{dt} < -0.05$$

eGFR data from one patient



- **When did the patient meet the clinical guideline for referral?**

The Salford Integrated Record System

- **Pioneered in 2003**
- **Integrates information from primary and secondary care**
- **Updated every 24 hours.**
- **Anonymised research data repository also created.**

The clinician's question and the statistician's answer

Clinician

Is my patient losing more than $> 5\%$ of kidney function per year?

Data

- **measurements $Y_{ij} = \log \text{eGFR}$ at times t_{ij},**
  explanatory variables x_i (age, sex)
  - $i = 1, \ldots, m = 22,910$ "at-risk" primary care patients
  - $j = 1, \ldots, n_i \leq 305$ (median $n_i = 12$)
  - $0 \leq 10.02$ years follow-up (median 4.46)

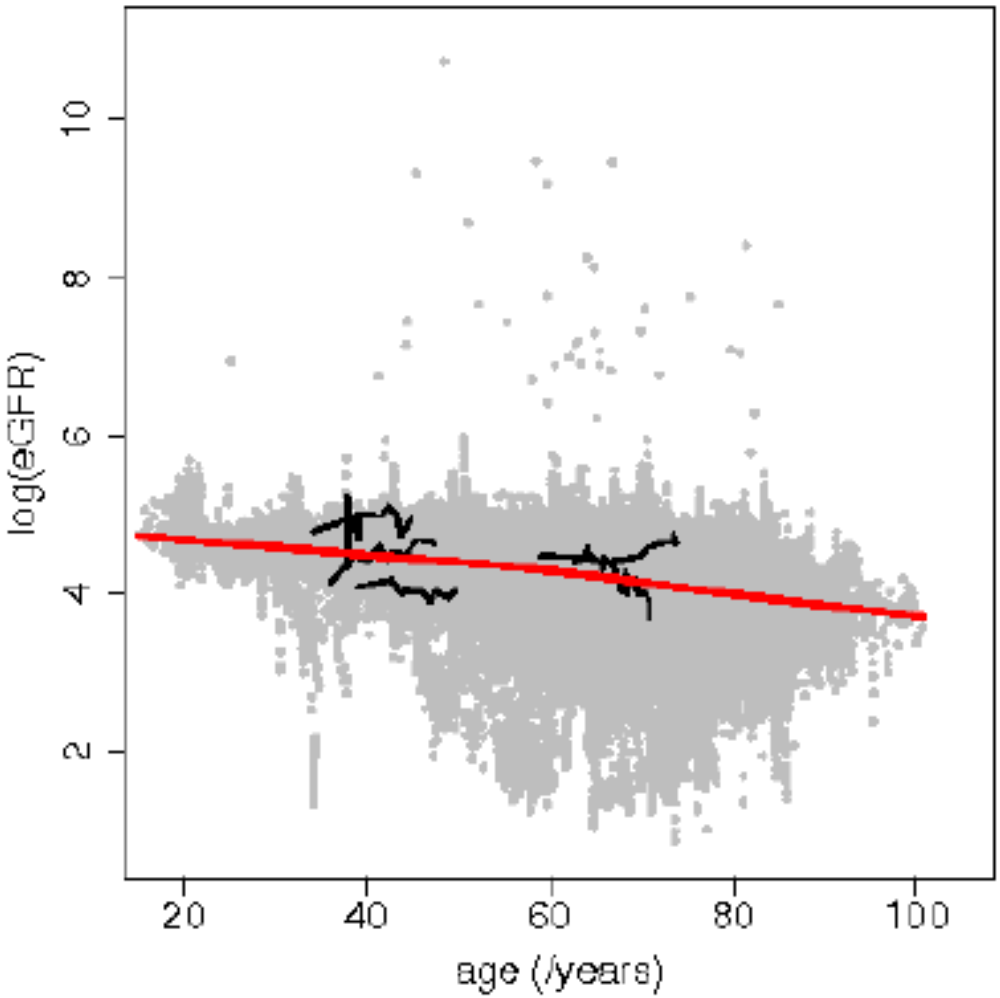
Statistician

Fit model to data, use model to calculate

$$\mathrm{P}\left(\frac{d}{dt} \log \text{GFR} < -0.05 | \mathcal{H}_t\right)$$

where $\mathcal{H}_t$ is patient data available at time t (nowcasting)

Data: all cross-sectional and selected longitudinal



Dynamic Regression Model

$$\begin{aligned} Y_{ij} &= \alpha_0 + \alpha_1 \times I(\text{female}) \\ &+ \beta_1 \times \text{age}_{i1} + \beta_2 \times (\text{age}_{ij} - \text{age}_{i1}) + \beta_3 \times \max(0, \text{age}_{ij} - 56.5) \\ &+ U_i + W_i(t_{ij}) + Z_{ij} \end{aligned}$$

- Z_{ij}: measurement error, $\mathrm{N}(0, \tau^2)$
- U_i: between-subject random intercept, $\mathrm{N}(0, \omega^2)$
- $W_i(t)$: within-subject stochastic process

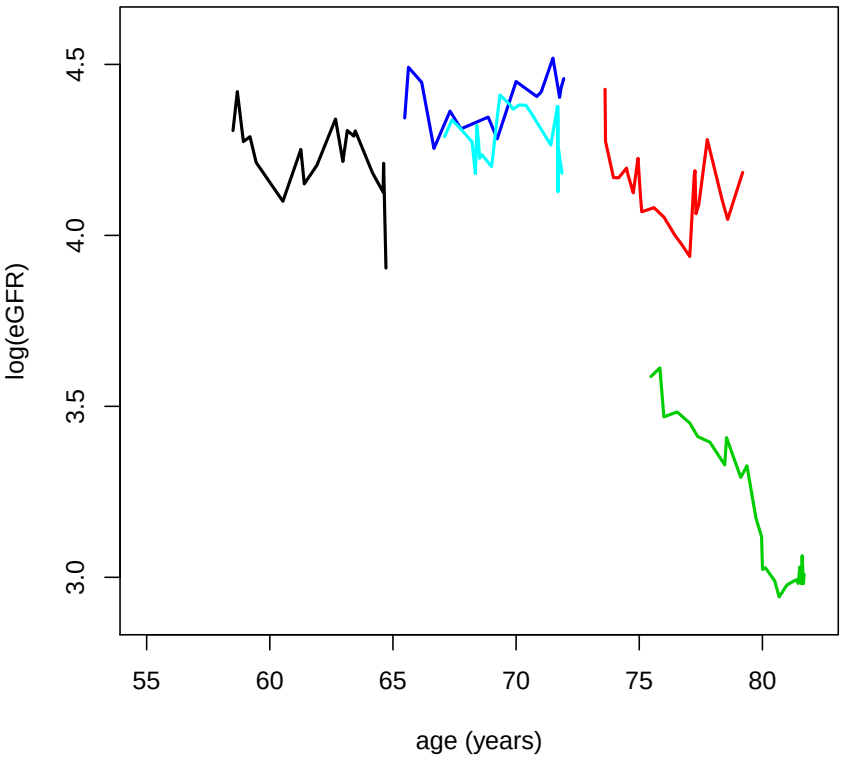
Model $W_i(t)$ as integrated Brownian motion

$$W_i(t) = \int_0^t B_i(u)du$$

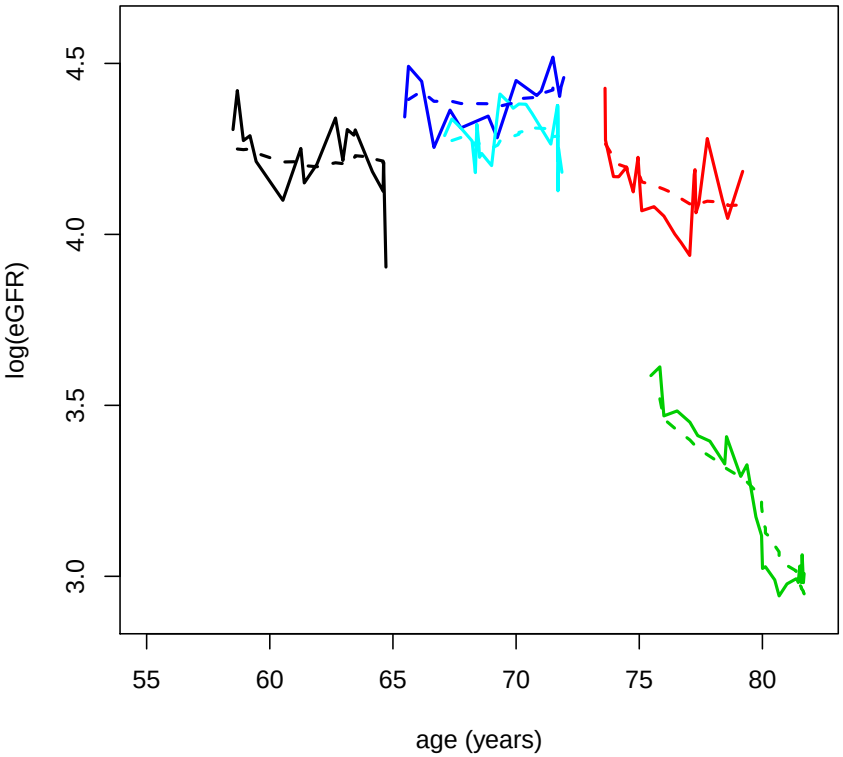
$$B_i(u)|B_i(s) \sim \mathrm{N}\left(B_i(u), (u-s)\sigma^2\right)$$

$B_i(u)$ is rate of progression for subject i at time t

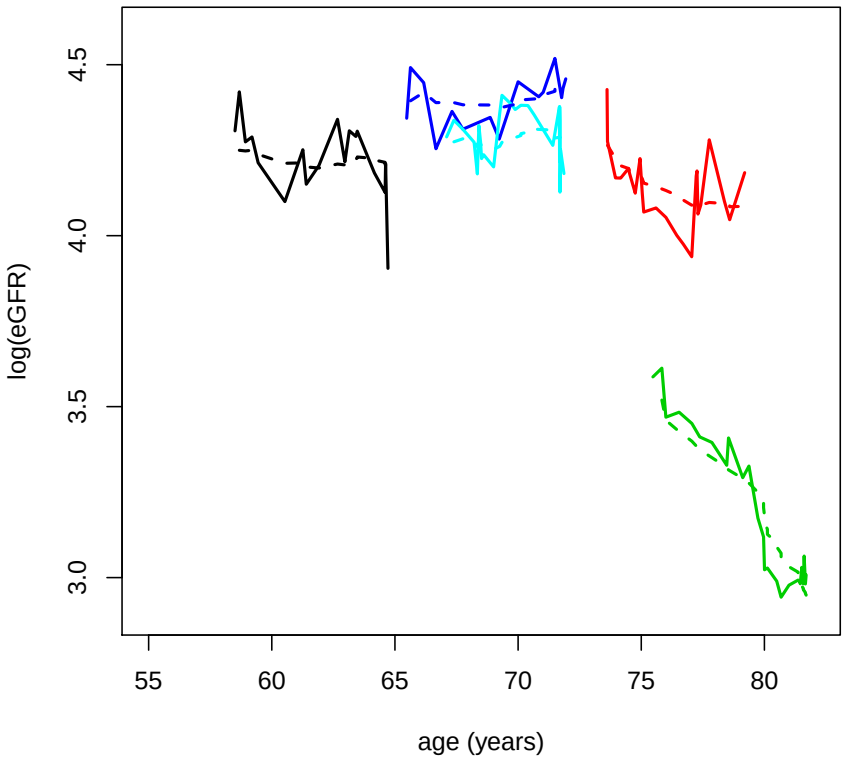
Modelling progression



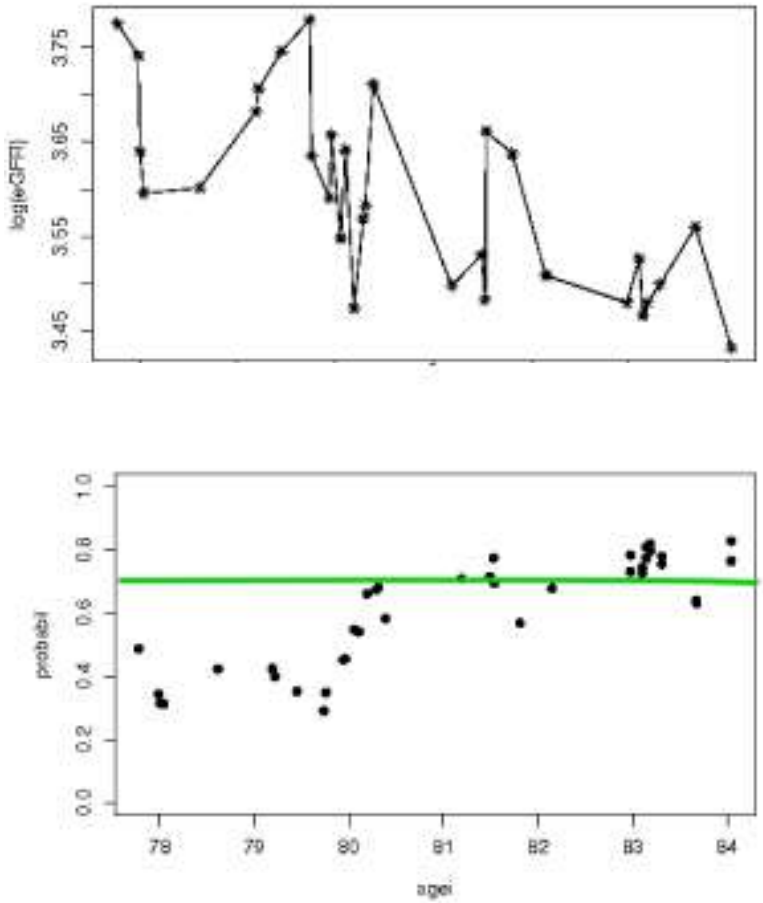
Modelling progression



Modelling progression



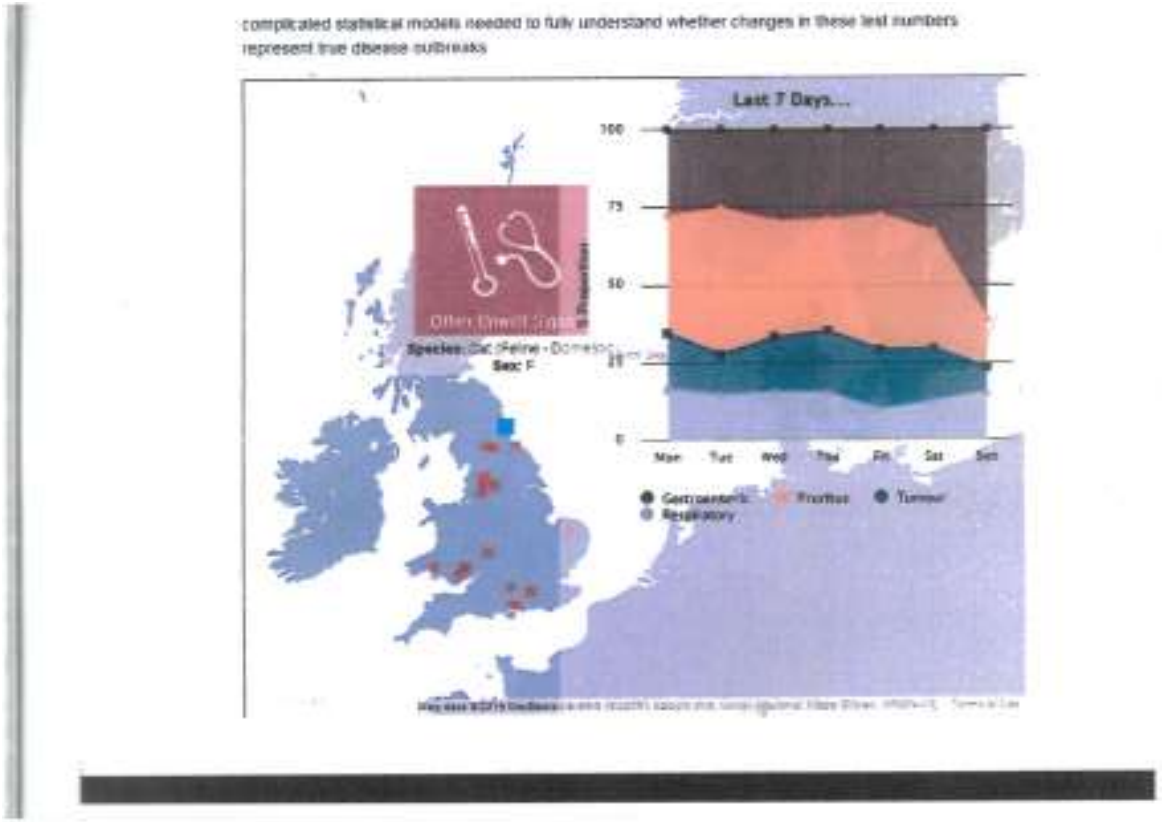
Rate of change in GFR?



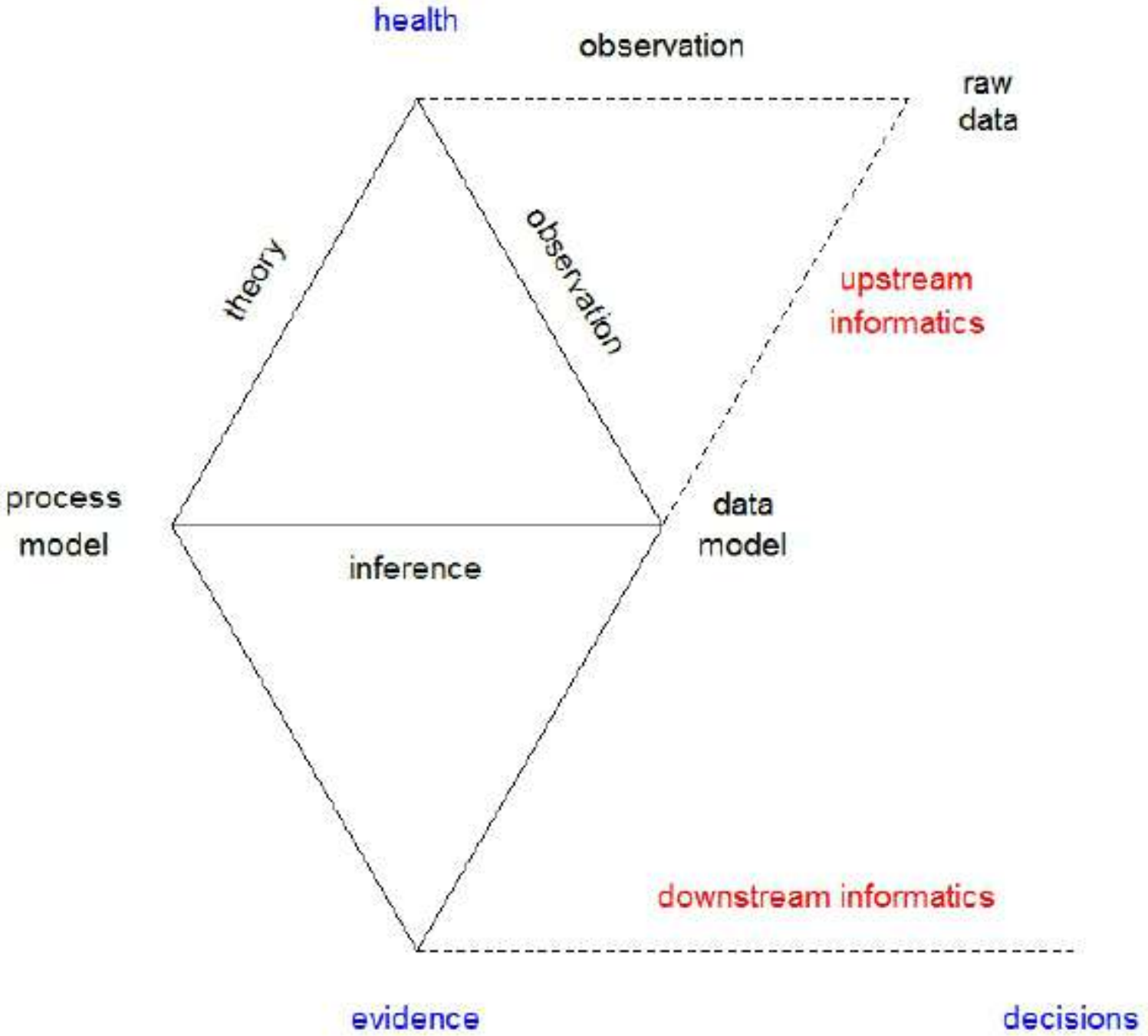
Small-animal veterinary surveillance

SAVSNET: real-time data-feed from network of small-animal veterinary practices:

- **practice location**
- **species (cat or dog)**
- **diagnosis**



http://www.savsnet.co.uk/realtimedata/


health
observation
raw
data
theory
observation
upstream
informatics
process
model
data
model
inference
downstream informatics
evidence
decisions

We are what we teach

Fewer lectures, more projects (problem-based learning?)

Builidng on a solid mathematical foundation

- **Design**
- **Probability and stochastic processes**
- **Likelihood-based inference**
- **Computation...numerical methods, programming**
- **Communication...scientific writing, including protocol/ethics**
- **Scientific method...core concepts in (biomedical) science**

Onchocerciasis: aka River Blindness

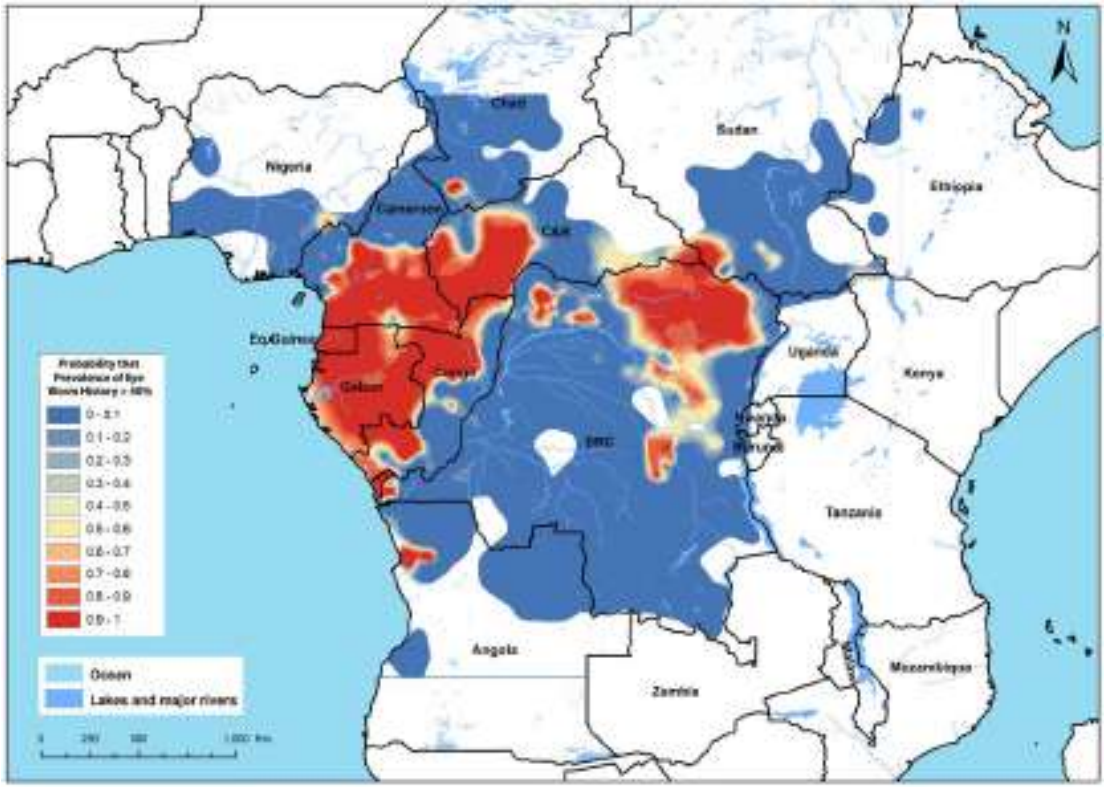
African Programme for Onchocerciasis Control (APOC)

- **Ivermectin (Mectizan): provides long-term protection if taken annually**
- **generally considered safe, with no serious side-effects**
- **mass distribution made possible by donation programme (Merck)**
- **multi-national programme coordinated by WHO**
- **recent decision to raise ambition from control to elimination**

The Loa loa problem. **People who are heavily co-infected with *Loa loa* parasites can experience serious (occasionally fatal) adverse reactions to Mectizan**

First solution: prevalence mapping with RAPLOA

A community is safe if its RAPLOA prevalence is less than 40%



Zoure, H., Wanji, S., Noma, M., Amazigo, U., Diggle, P.J., Tekle, A. and Remme, J.H. (2011). The geographic distribution of Loa loa in Africa: results of large-scale implementation of the Rapid Assessment Procedure for Loiasis (RAPLOA). *Public Library of Science: Neglected Tropical Diseases* **5, (6): e1210.doi:10.1371/journal.pntd.0001210**

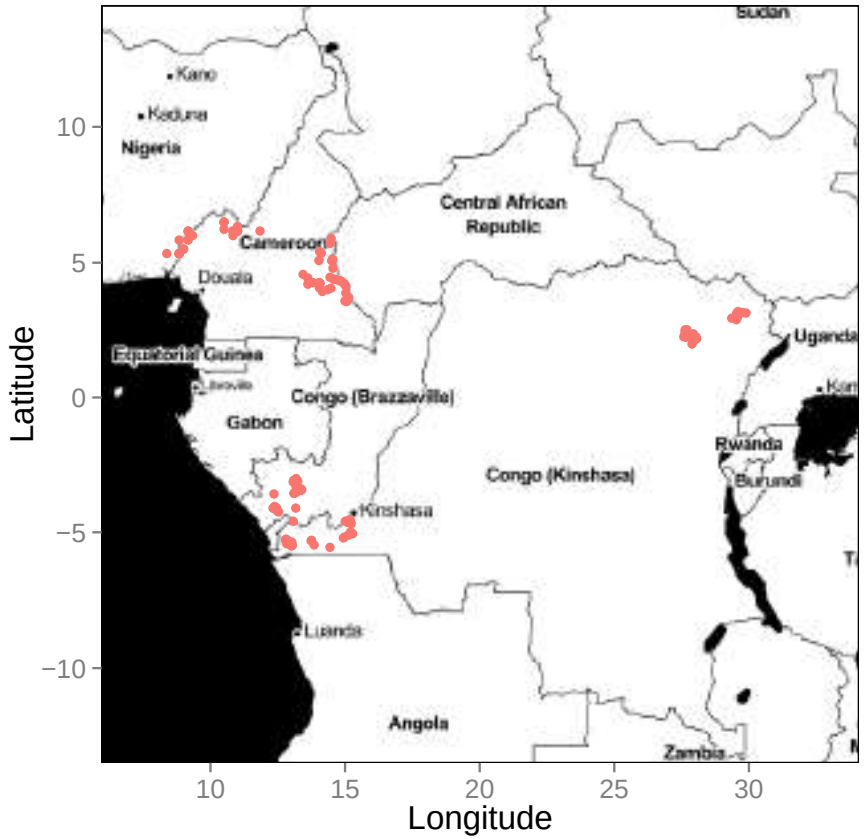
Second solution: collect individual-level data on levels of infection

A community is safe if the proportion of highly infected individuals is small

- **Infections levels $Y_i : i = 1, ..., n$ (parasites per ml blood)**
- **$Z = \#(Y_i > 20,000)$**

$$Z \sim \text{Bin}(n; p)$$

Building a predictive model for infection levels

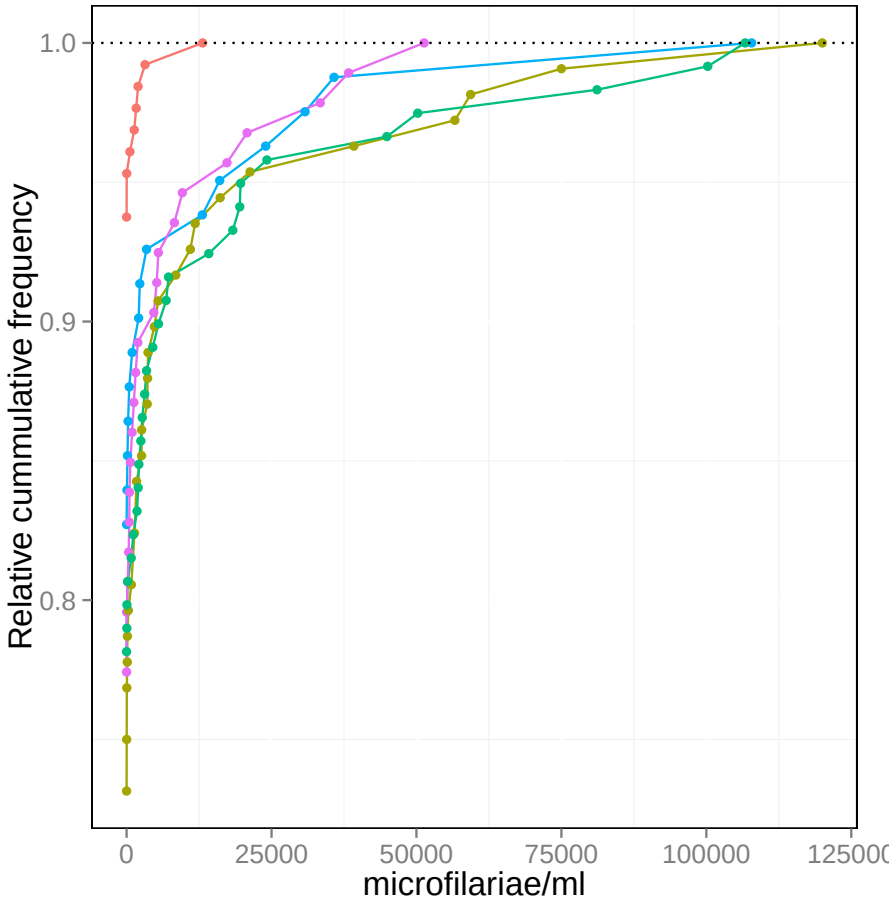


- 222 villages
- 24 to 229 individuals per village, total 19,128
- traditional microscopy
- Y = individual infection level/ml

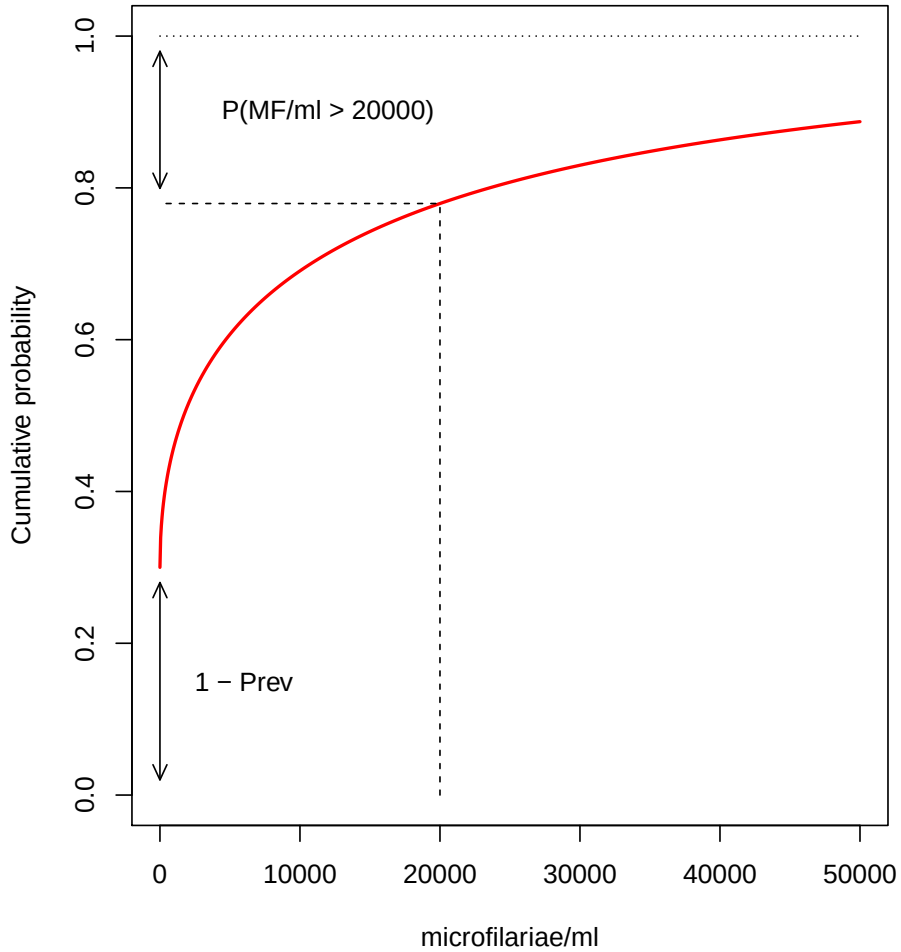
Schlüter, D.K., Ndeffo-Mbah, M.L., Takougang, I., Ukety, T., Wanji, S., Galvani, A.P. and Diggle, P.J. (2016). Using community-level prevalence of Loa loa infection to predict the proportion of highly-infected individuals: statistical modelling to support lymphatic filariasis elimination programs. *PLoS Neglected Tropical Diseases*, **10, 12,** e0005157. doi:10.1371/journal.pntd.0005157

Cumulative distribution of infection levels

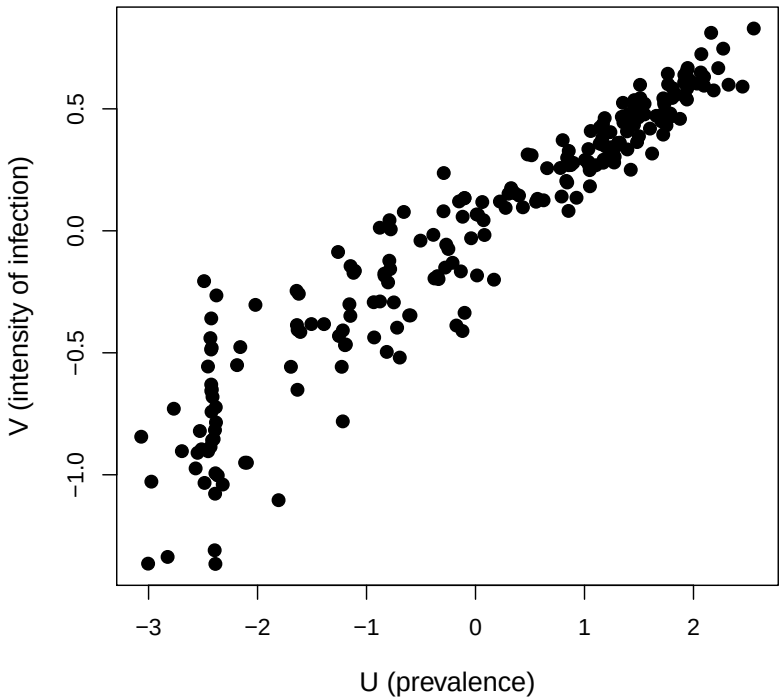
Data from 5 villages



Model schematic



Variation between villages: exploiting correlation between community-level prevalence and levels of infection



- **The proportion of a rare condition in a sample is an imprecise estimate of the proportion in the community**
- **Correlation between community-level prevalence and intensity of infection $\Rightarrow$ information gain (narrower prediction intervals)**

Both prevalence and intensity give important information and both are necessary to improve predictions

Problem statement: revisited

A community is safe for MDA if less than 1% of its population have infection level greater than 20,000 microfilariae/ml

Data

- **Sample size n from community of size N**
- **Estimated microfilariae/ml $Y_1, \ldots, Y_n$**
- **At-risk individual: $Y > 20{,}000$**

Required

- **Probability distribution of the proportion ($\Rightarrow$ number) of at-risk individuals in the community**

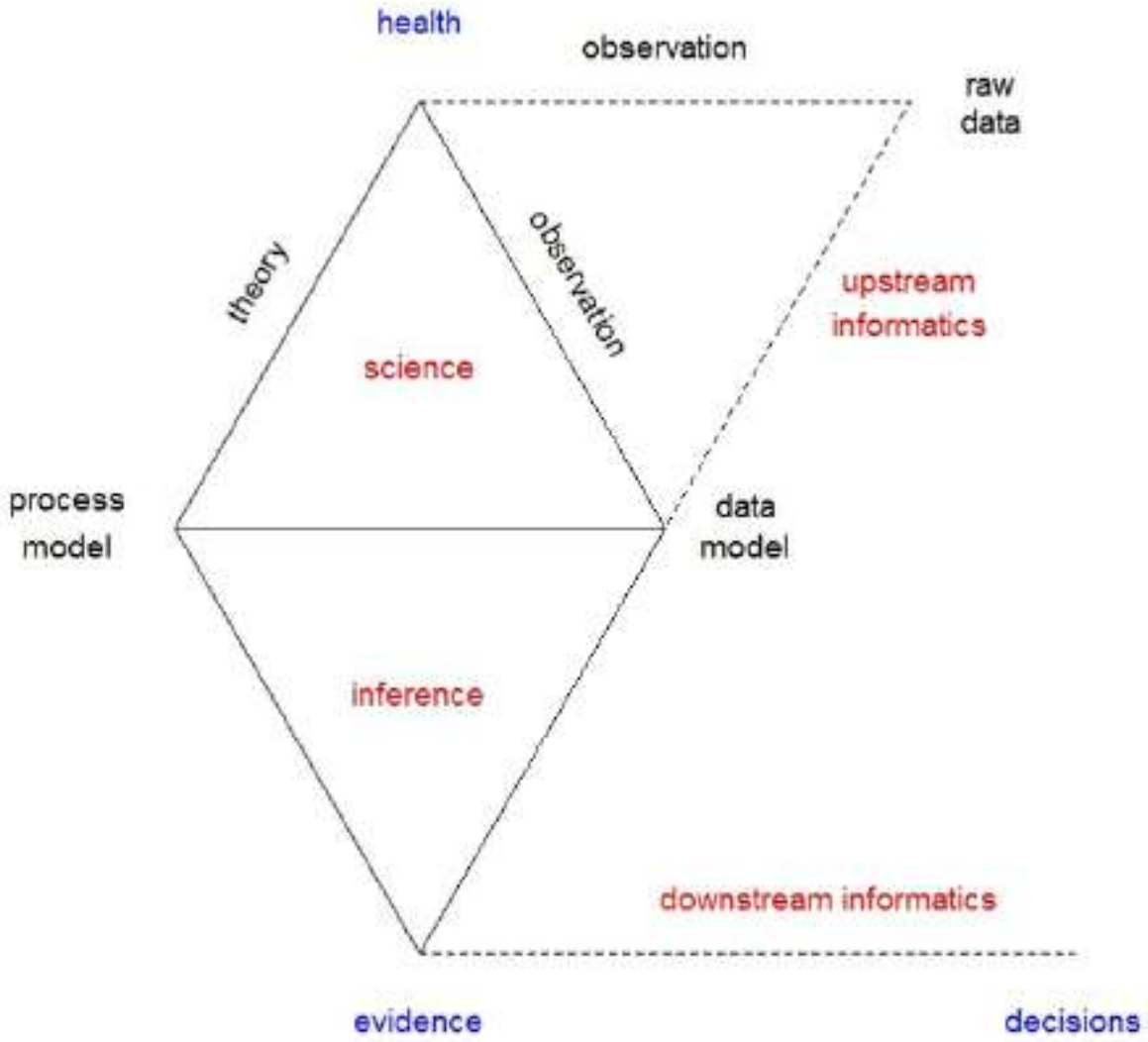
Model can be adjusted by the user to assess any policy-defined threshold and target

Scaling up: mobile microscopy

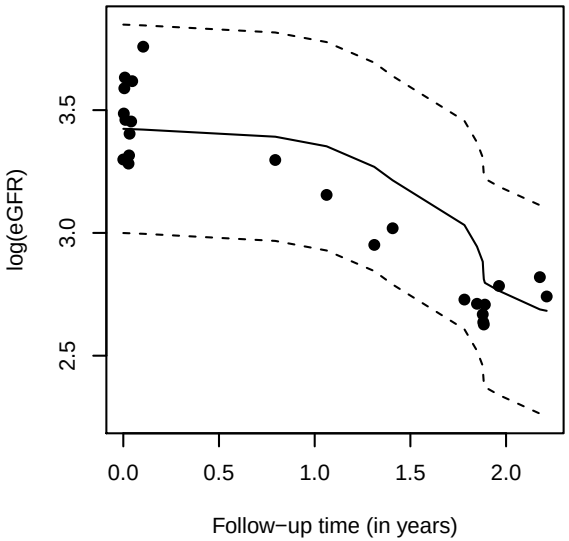
- **portable**
- **easy-to-use**
- **prediction algorithm can be hard-wired**

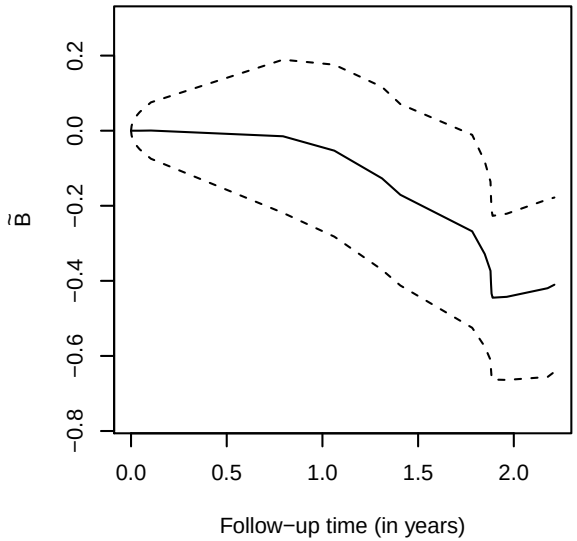
Ethical dilemma: **what is an acceptable risk threshold?**

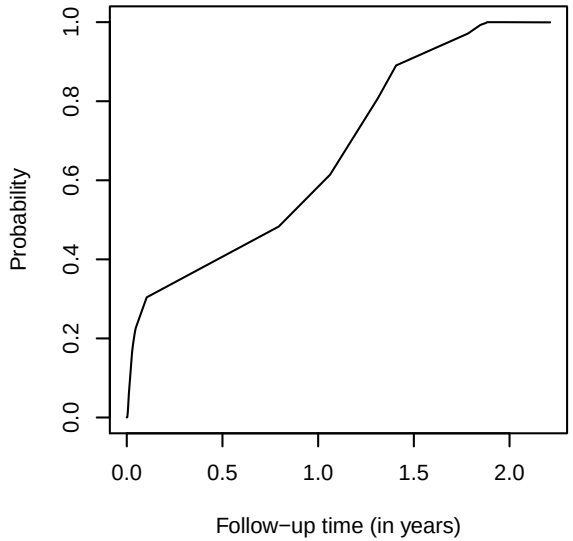
Health Data Science



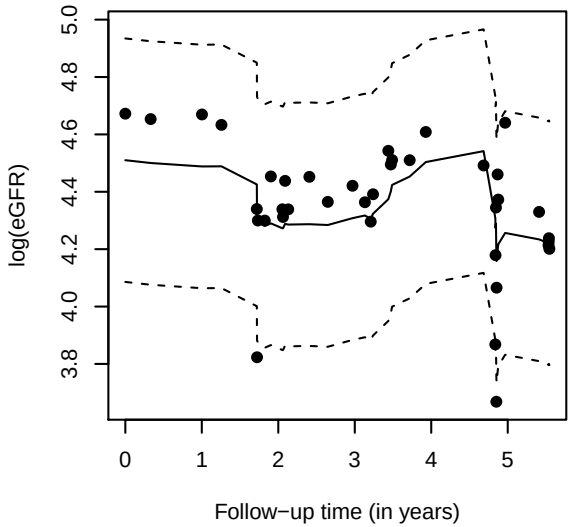
Prediction: classic progression pattern

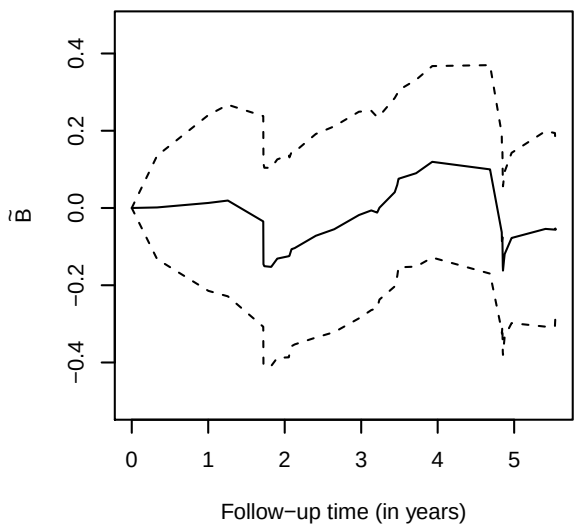


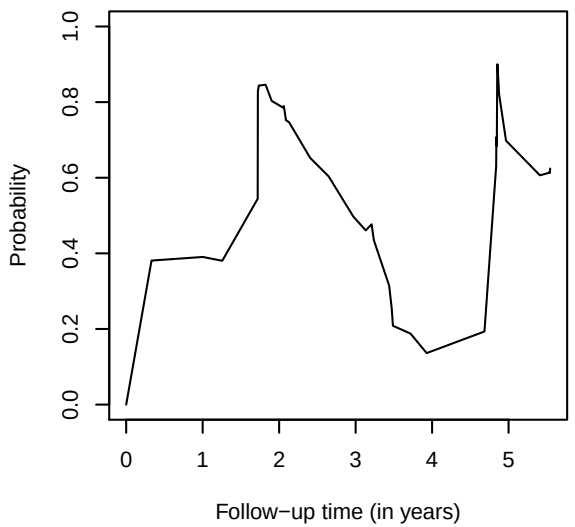




Prediction: AKI (Acute Kidney Injury) recovery







Prediction: non-recovery from AKI