

Statistics: a Data Science for the Twenty-first Century

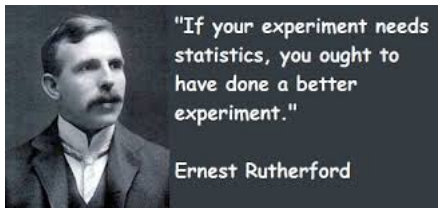
Peter J Diggle

CHICAS, Lancaster University Medical School

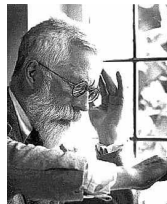
November 2019



Data science: more data = more information?



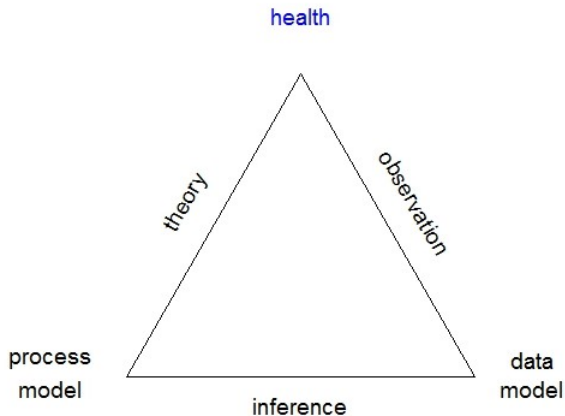
And who better to design that experiment than a statistician?'



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

Judea Pearl

The (Health) Science Triangle



- 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?)

Data Science: threat or (missed?) opportunity?

Threat?

- we've been here before...statistical packages ca 1970
- the ubiquitous amateur statistician
- but we're still here

Missed opportunity?

Tukey, J. W. (1962), The Future of Data Analysis. *Annals of Mathematical Statistics*, 33, 1–67.

Wu, C. F. J. (1997). Statistics = Data Science?

<http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf>

Donoho, D. (2017) 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26, 745–766.

- **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...

Wikipedia

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

Iain Buchan

- **Data science** is **Statistics + Informatics + Science**

PJD

What can we offer?

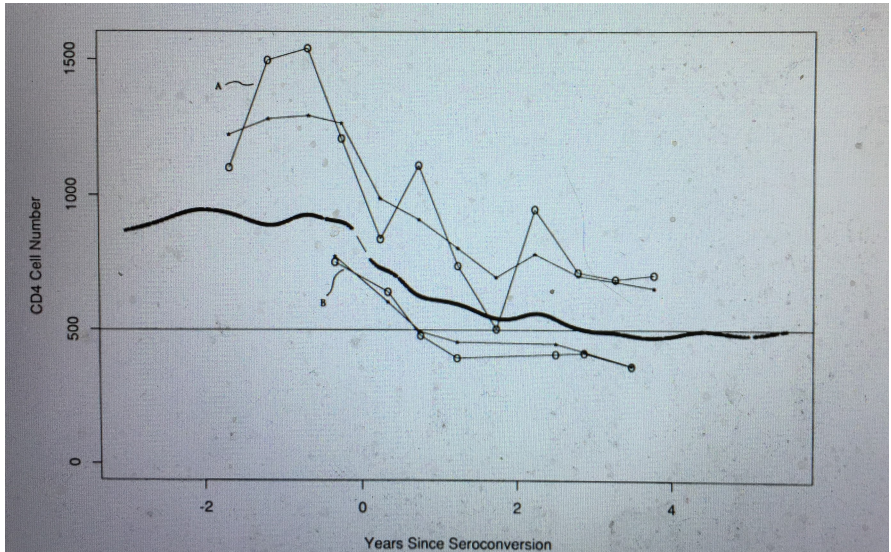
- that probability theory is the correct way to deal with uncertainty
 - in our data ... stochastic models
 - in our conclusions ... probabilistic inference
- that design matters
- that context matters

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

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

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



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

$$\text{DATA} = \text{SIGNAL} + \text{NOISE}$$

$$[S][D|S] \Rightarrow [S|D]$$

"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



Kidney failure

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**

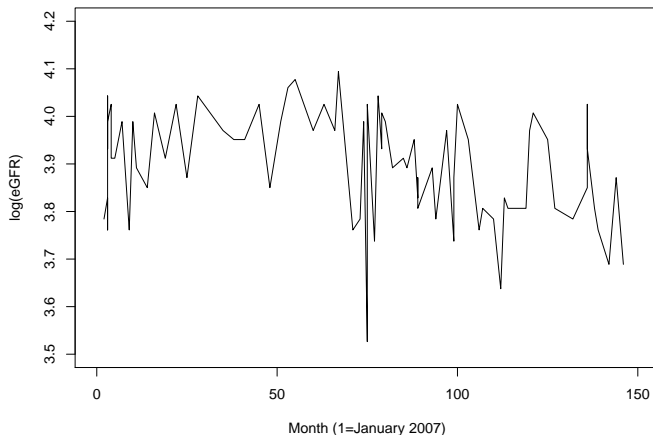
Clinical guideline

- 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?

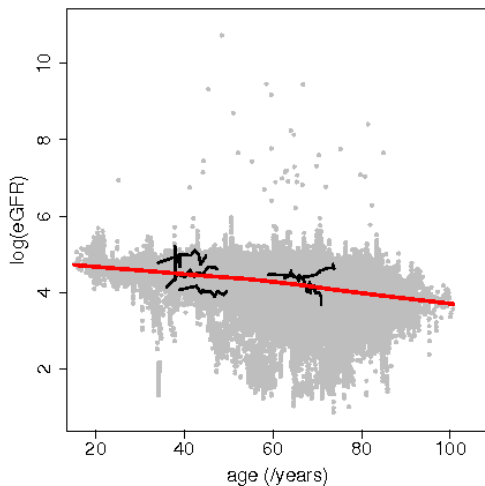
Data

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

Statistician

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

Data: all cross-sectional and selected longitudinal



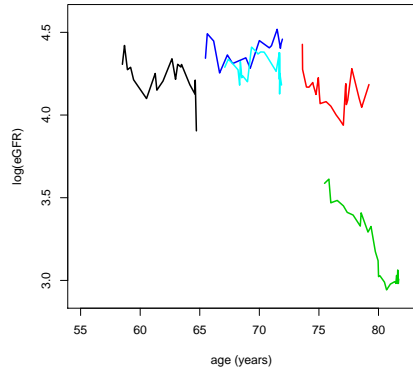
$$Y_{ij} = x'_{ij}\beta + U_i + W_i(t_{ij}) + Z_{ij}$$

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

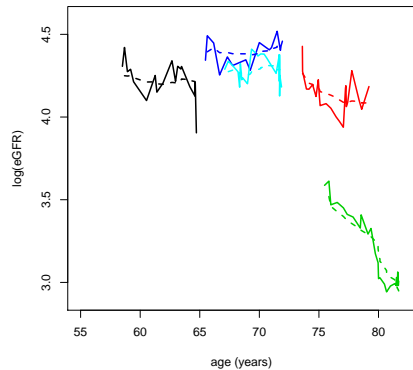
$$B_i(u)|B_i(s) \sim N(B_i(u), (u-s)\sigma^2)$$

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

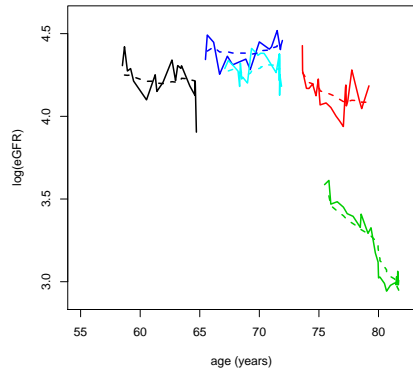
Modelling progression



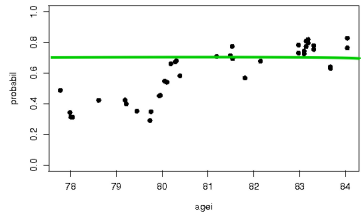
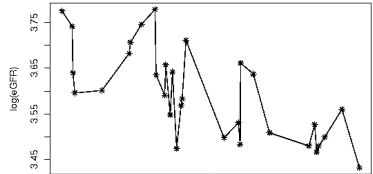
Modelling progression



Modelling progression



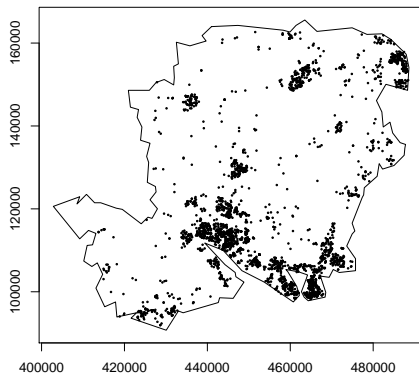
Rate of change in GFR?



Food-poisoning

The AEGISS project

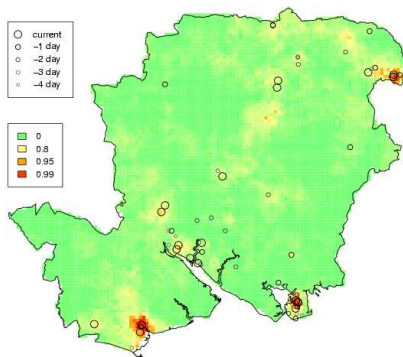
Calls to NHS Direct, ca 2001–2003, probable food-poisoning



<http://www.lancaster.ac.uk/staff/diggle/aegiss/>

The AEGISS project

- **Goal:** early detection of anomalies in local incidence
- **Data:** locations of 3374 consecutive reports of probable food-poisoning
- **Model:** log-Gaussian Cox process



Cox process: a Poisson process whose intensity function is a realisation of a real-valued stochastic process

- a **natural modelling framework** when the observed point process is the product of **unobserved environmental variation**
- analogous to a **mixed effects regression model**: stochastic process as proxy for unmeasured explanatory variables

Log-Gaussian Cox process:

$$\Lambda(x, t) = \exp\{z(x, t)' \beta + S(x, t)\}$$

- $z(x, t)$: observed covariates
- $S(x, t)$: unobserved Gaussian process

Log-Gaussian Cox processes: likelihood function

Data

$$\mathcal{X} = \{(x_i, t_i) \in \mathbf{A} \times [0, T] : i = 1, \dots, n\}$$

Poisson process likelihood

$$\ell_{\lambda}(\theta; \mathcal{X}) = \prod \lambda(x_i, t_i) \exp \left\{ - \int_0^T \int_{\mathbf{A}} \lambda(x, t) dx dt \right\}$$

Cox process likelihood

$$\ell(\theta; \mathcal{X}) = \mathbb{E}_{\Lambda} [\ell_{\lambda}(\theta; \mathcal{X})]$$

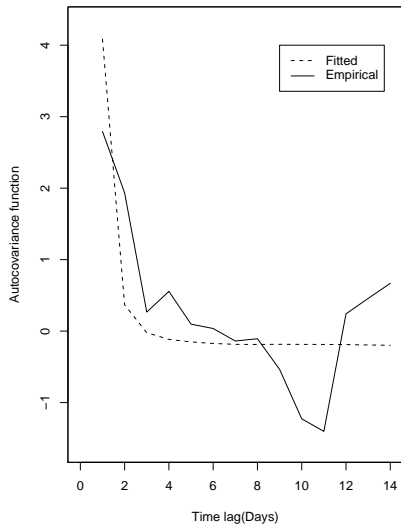
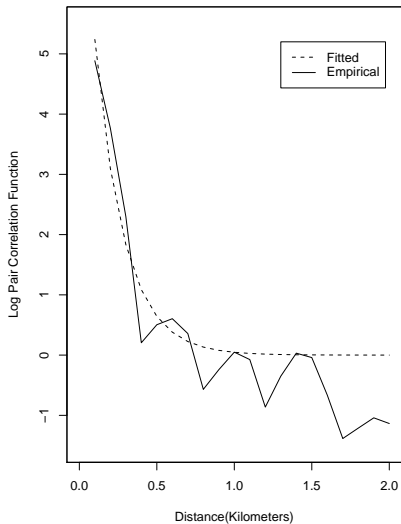
Cox process inference

Monte Carlo maximum likelihood or Bayes according to taste

$$\lambda(x, t) = \lambda_0(x) \mu_0(t) \exp\{S(x, t)\}$$

- $\lambda_0(x)$ = non-parametric – adaptive kernel estimate
- $\mu_0(t)$ = Poisson log-linear model with terms for:
 - trend
 - seasonal
 - day-of-week
- $\text{Cov}\{S(x, t), S(x - u, t - v)\} = \sigma^2 \exp(-u/\phi) \exp(-v/\theta)$
- Use incident data up to time t to construct predictive distribution for $\exp\{S(x, t)\}$, hence identify potential **anomalies** in incidence pattern

Estimated spatial and temporal correlation functions



Spatial prediction (cf kidney failure example)

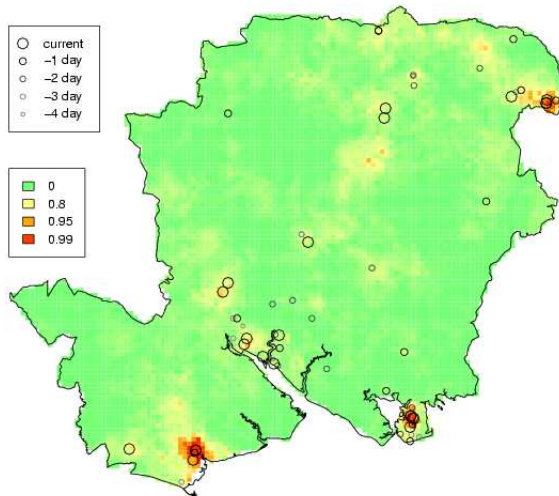
- choose **critical threshold** $c > 1$
- map **exceedance probabilities**,

$$p_t(x) = P(\exp\{S(x, t)\} > c | \text{data})$$

- web-reporting with **daily updates**

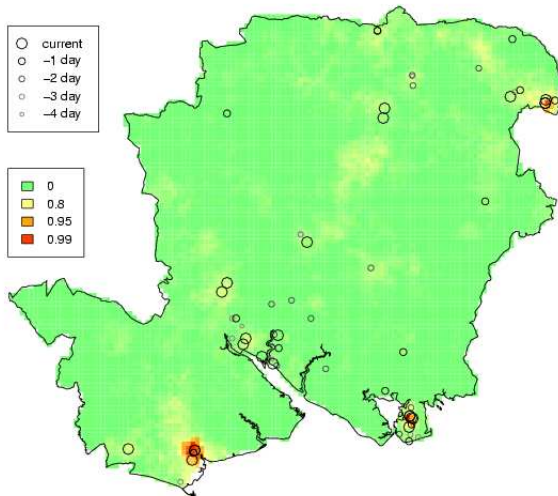
<http://www.lancaster.ac.uk/staff/diggle/>

Spatial prediction: results for 6 March 2003



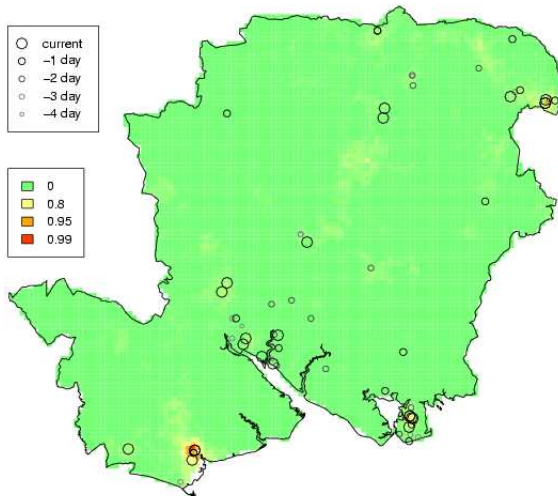
$$c = 2$$

Spatial prediction: results for 6 March 2003



$c = 4$

Spatial prediction: results for 6 March 2003

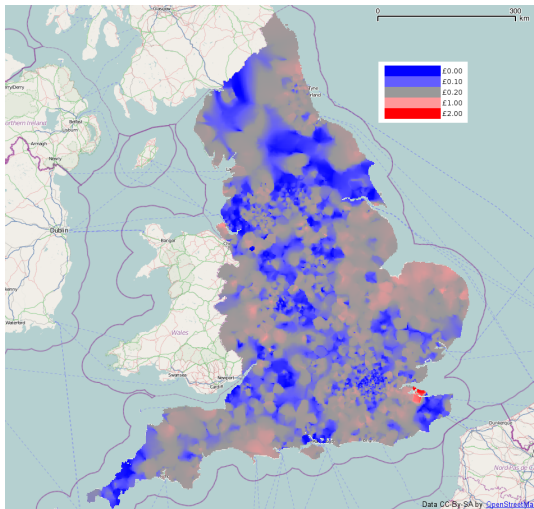


$c = 8$

Ritalin prescribing

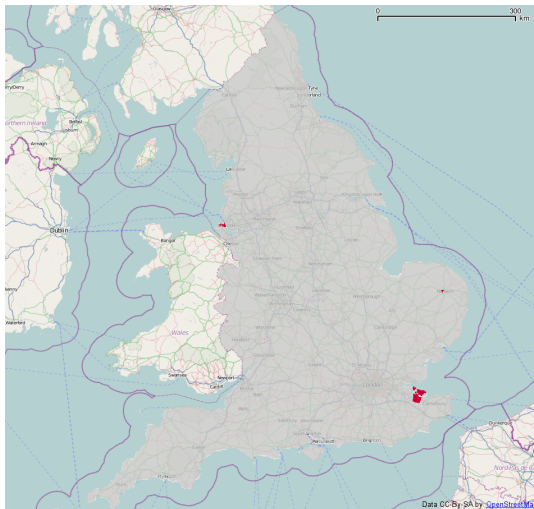
NHS Prescribing patterns: ritalin, October 2011

Wide geographical variation in monthly spend per person

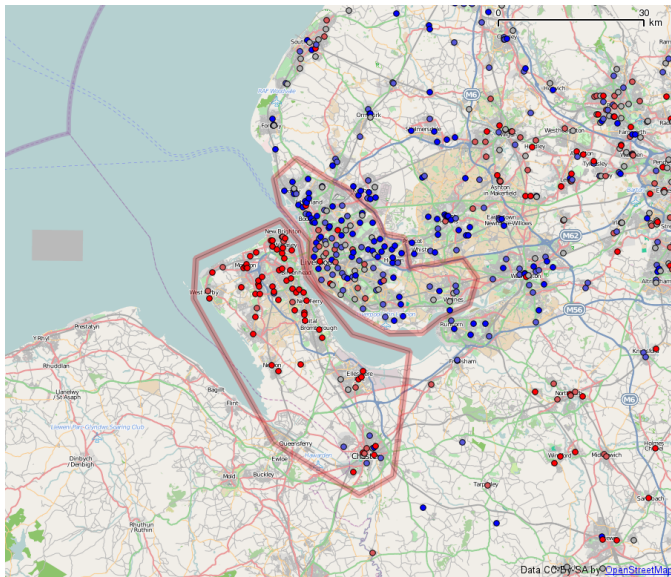


NHS Prescribing patterns: ritalin, October 2011

Local rates at least four times national average?



NHS Prescribing patterns: ritalin, October 2011

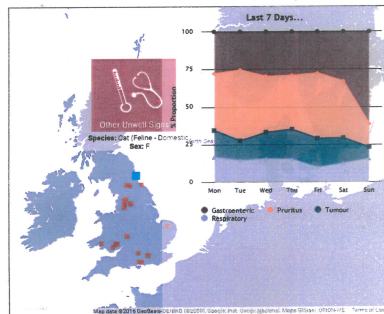


Veterinary surveillance

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

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

complicated statistical models needed to fully understand whether changes in these test numbers represent true disease outbreaks



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

Loa loa

Tropical disease prevalence mapping

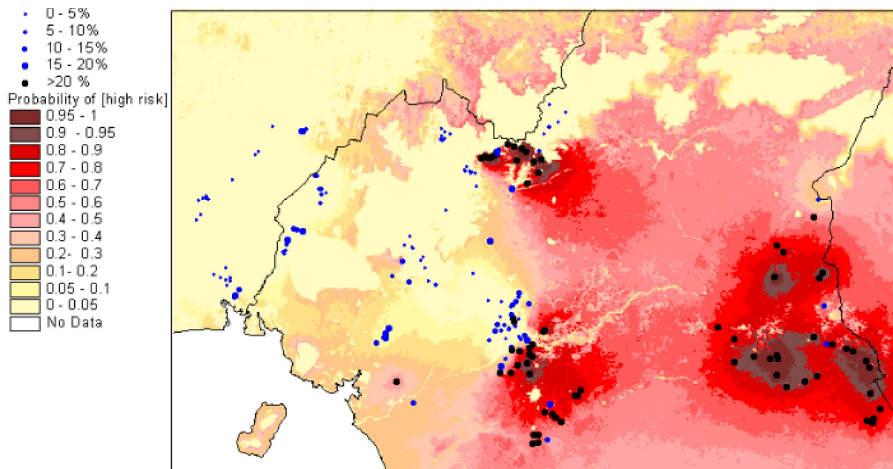


Figure 6: PCM for [high risk] in Cameroon based on ERM with ground truth data.

Mobile microscopy



- portable
- easy-to-use
- local prediction algorithm can be hard-wired
- data can be sent to the cloud for off-line analysis

Where should statisticians sit?

“...the importance of making contact with the best research workers in other subjects and aiming over a period to establish genuine involvement and collaboration in their activities.”

Sir David Cox (b 1924)



My ideal:

- in a core Data Science Unit
- with links to substantive disciplines through:
 - dedicated time (and space) to meet and share ideas
 - joint appointments

Principles

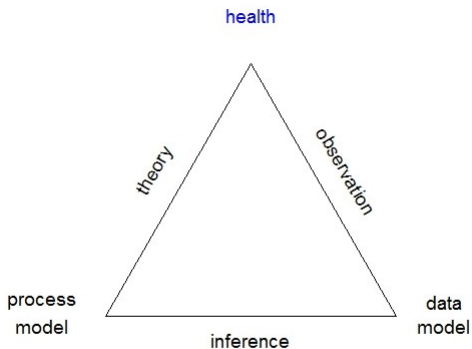
- Fewer lectures, more projects/placements
- Team science ... co-authored MSc/PhD Theses
- Less emphasis on techniques, more on core concepts

Content

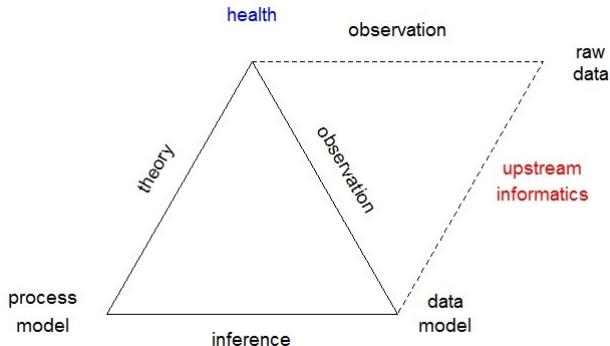
- Design
- Probability and stochastic processes
- Likelihood-based inference

- Computation...numerical methods, programming
- Communication...scientific writing, including protocol/ethics
- Core concepts in (biomedical) science

The (Health) Science Triangle



The (Health) Informatics Extension



(Health) Data Science

