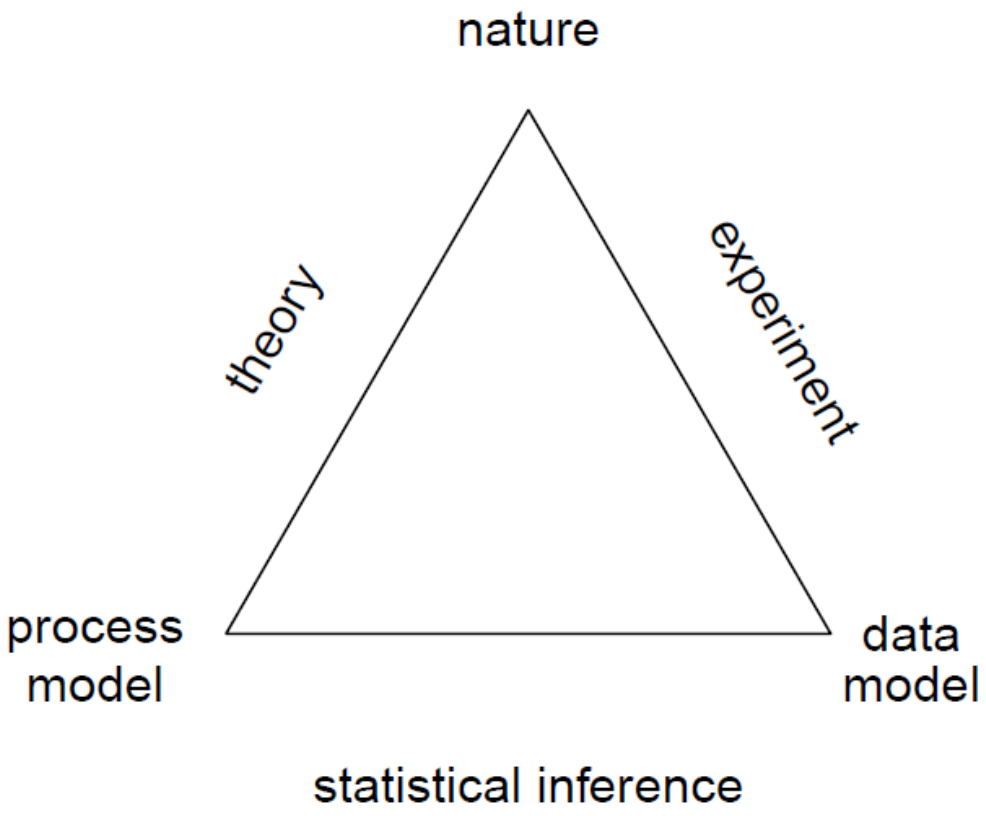
Statistics: a Data Science for the Twenty-first Century

Peter J Diggle

CHICAS, Lancaster University Medical School

November 2016

The science triangle



The rise of data science: threat or opportunity?

We've been here before:

- statistical packages ca 1970...the ubiquitous amateur statistician
- so why are we still here?
  1. wider appreciation of the added value of statistical thinking
  2. importance of design and context

"If your result needs a statistician, you should design a better experiment"

Rutherford?

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

PJD

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...
- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

"A rose by any other name would smell as sweet"

W. Shakespeare

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?

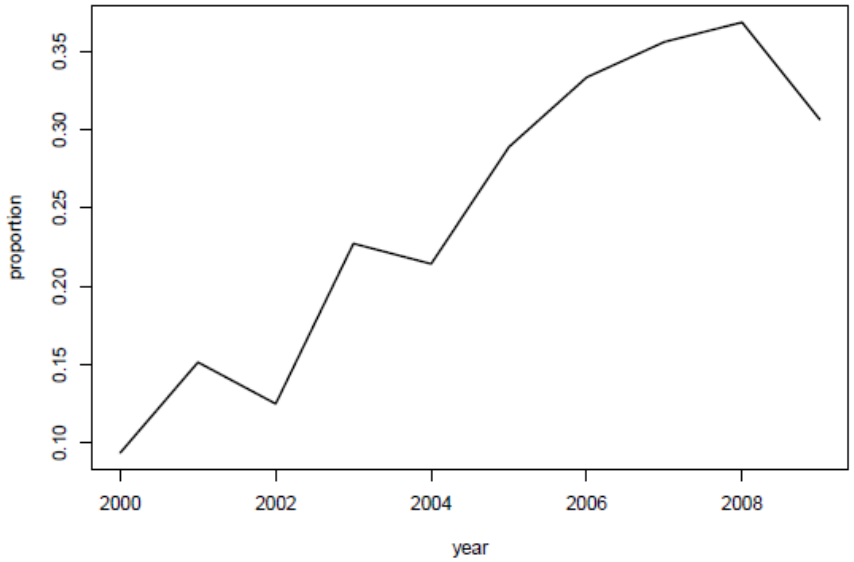
- that a published article is not a complete solution to a practical problem.
- that reproducibility of computationally driven research findings should be a minimum standard

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

Iain Buchan (Farr Institute)

The impact of modern biology on biostatistical research

- genetics/omics papers in the OUP journal "Biostatistics"



- bioinformatics $\rightarrow$ health informatics (e-health)

e-Health research... aka Health Informatics

"The wealth of electronic health data within the NHS ... to assess risks to public health and study the causes of diseases and disability."

MRC call for e-Health Research Centres

e-health research tools:

- linkage of electronic health records to other datasets including: research data; geo-spatial information; socio-economic records
- exploiting existing or emerging e-health records infrastructures
- new methods for data manipulation, linkage or analysis in key areas of statistics, computer science or informatics.

Health Informatics research: the role of statistical method

- low signal-to-noise ratio 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"

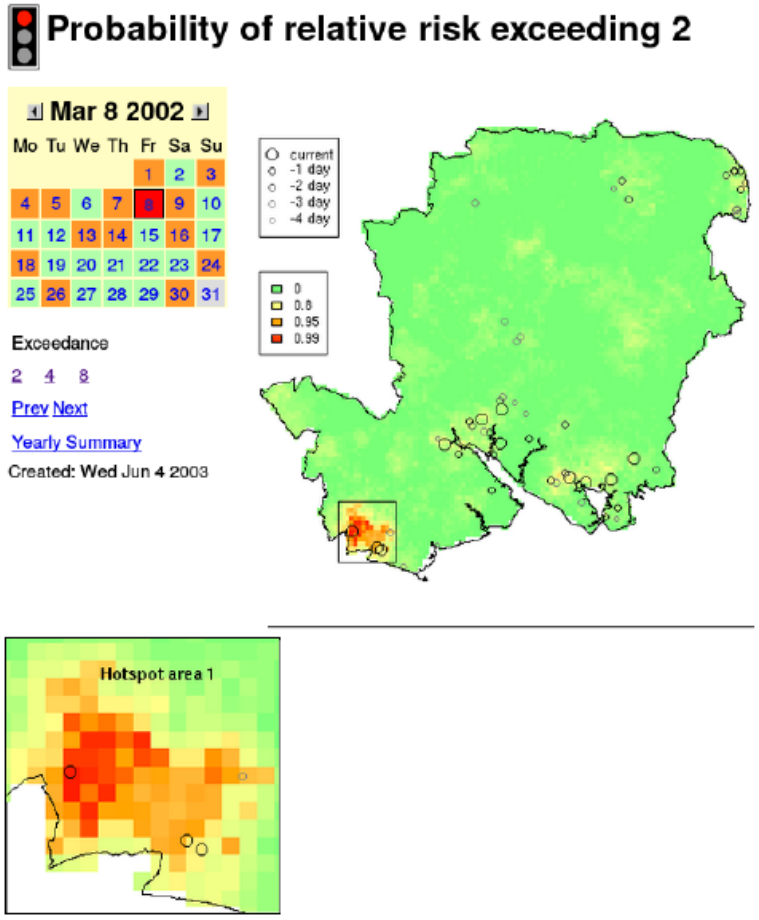
John Tukey

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

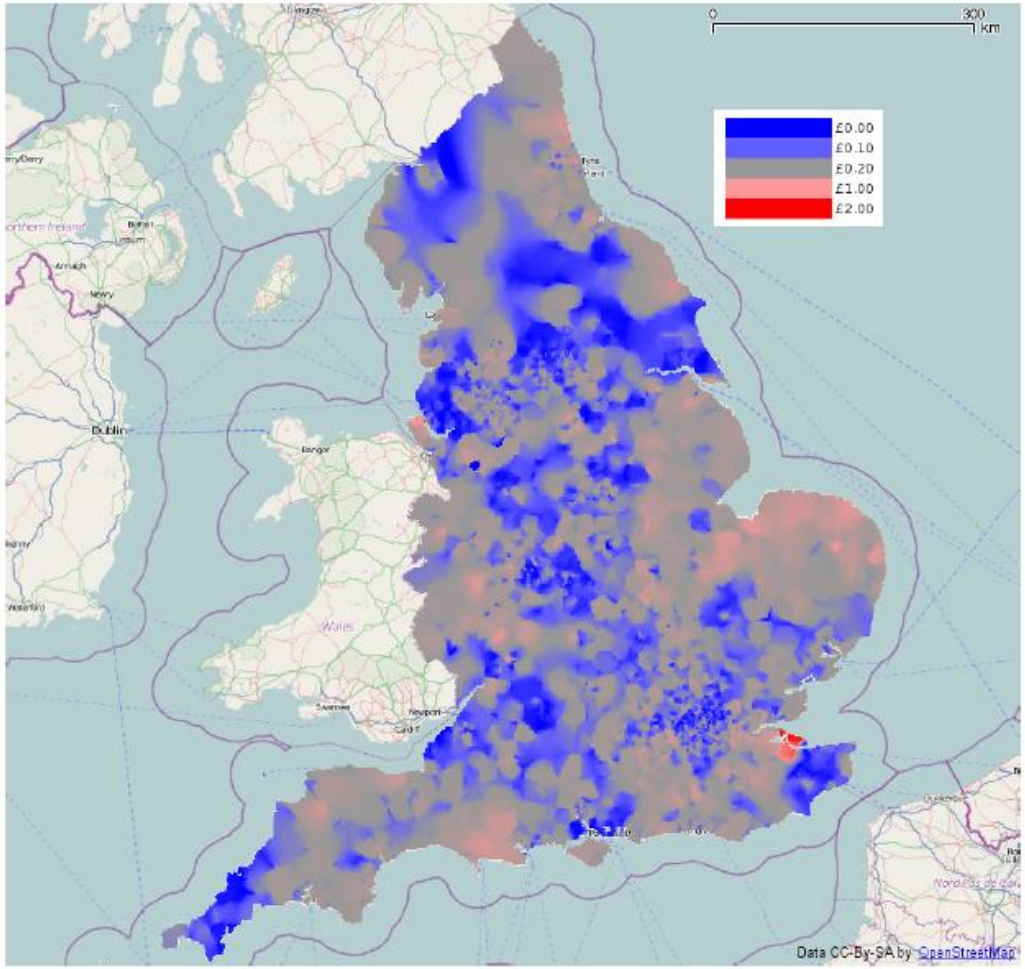
Peter McCullagh

Real-time spatial surveillance of gastro-entric illness

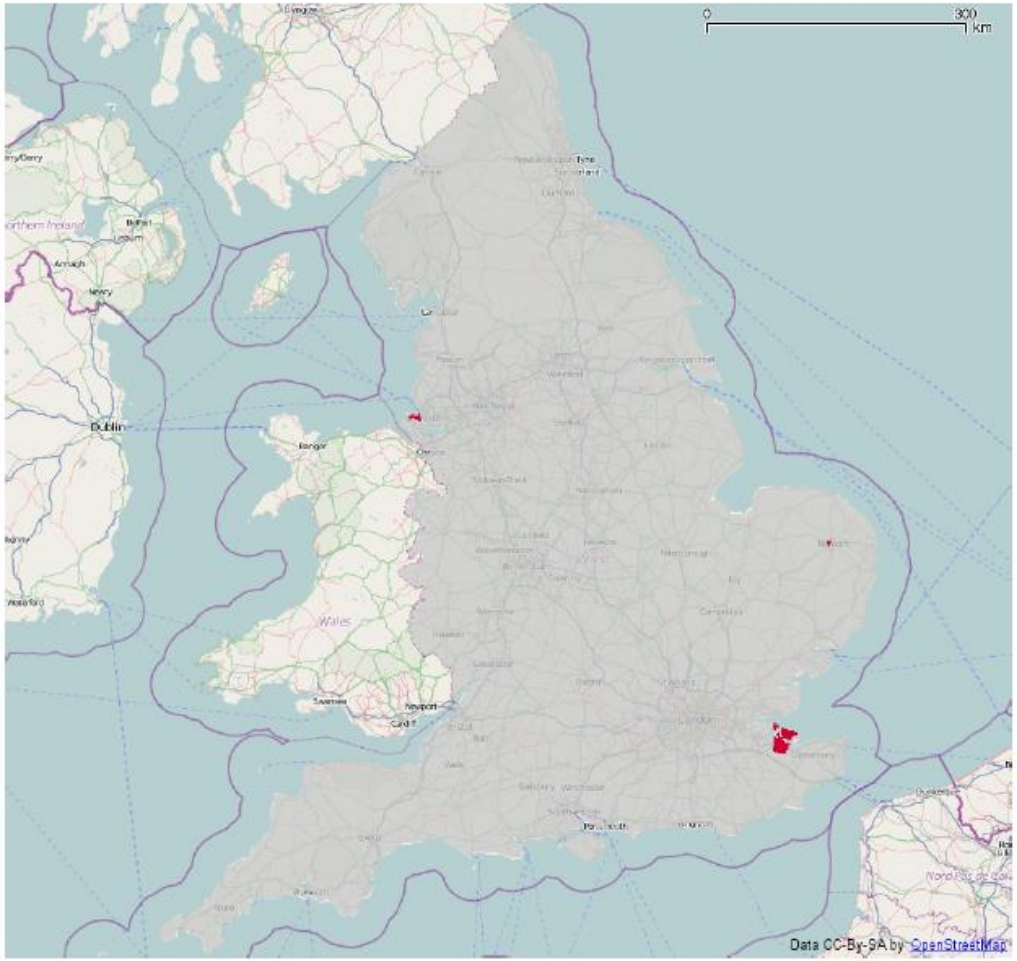
- early detection of anomalies in local incidence
- data on 3374 consecutive reports of non-specific gastro-intestinal illness
- log-Gaussian Cox process, space-time correlation $\rho(u, v)$



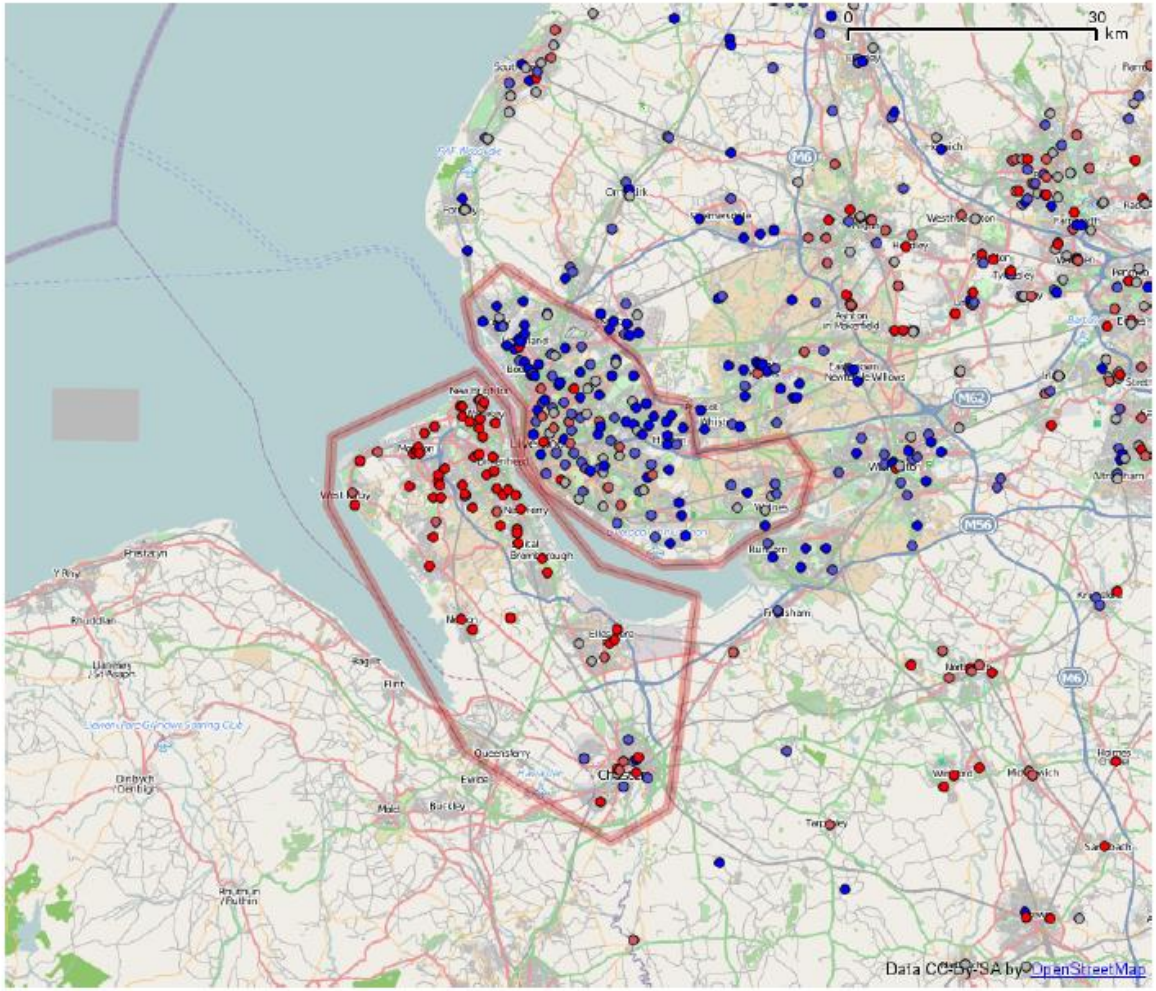
NHS Prescribing patterns: ritalin



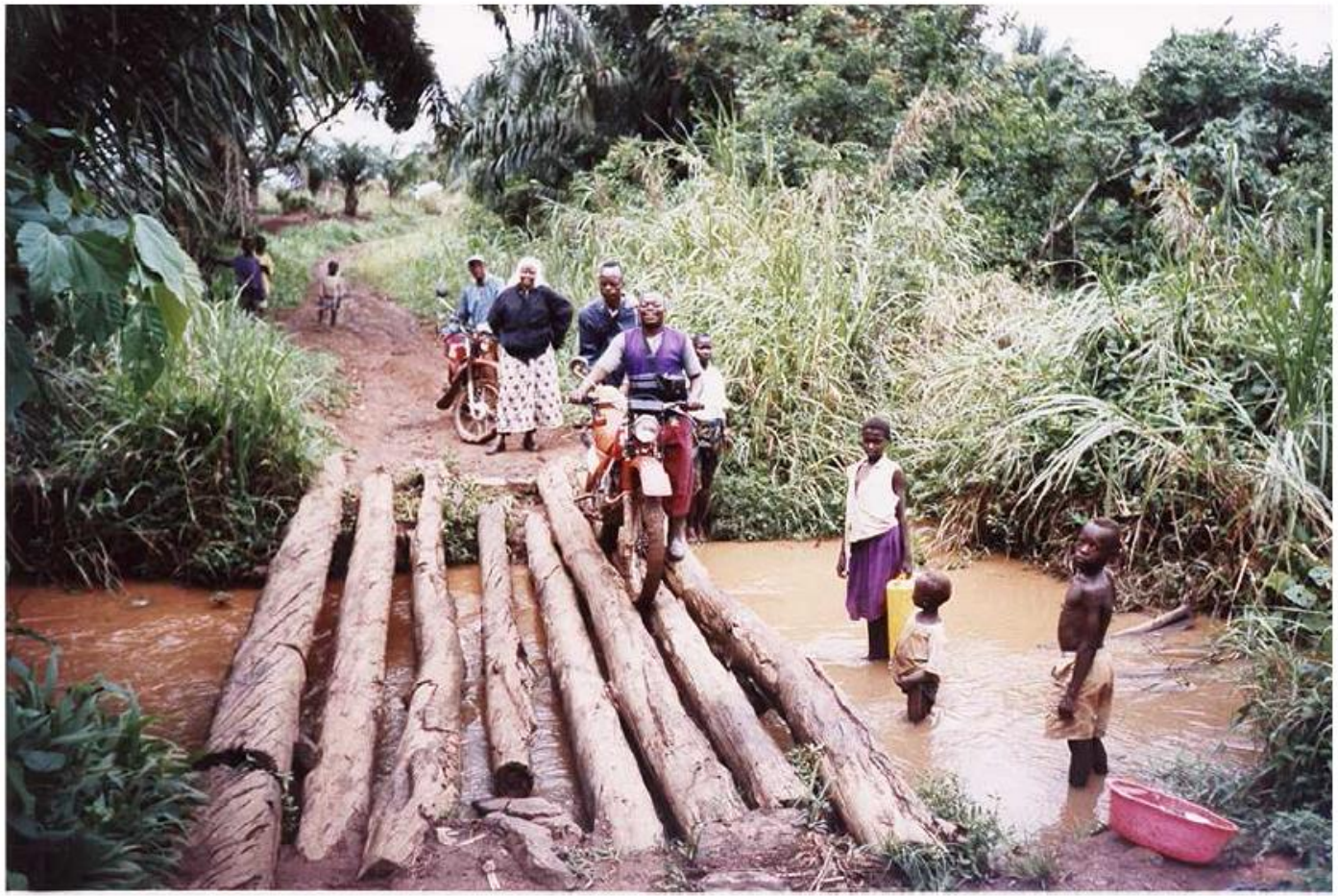
NHS Prescribing patterns: ritalin



NHS Prescribing patterns: ritalin



A developing country example



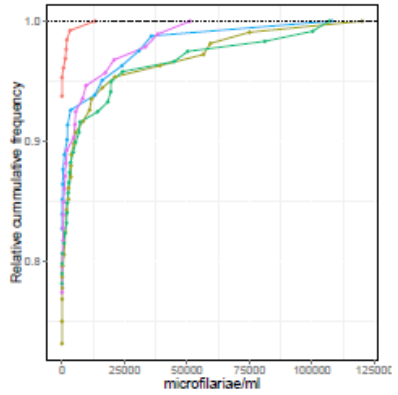
River blindness and eyeworm: a tale of two parasites

- multi-national programme of mass administration of medication to protect against river blindness
- risk of severe adverse reaction in individuals heavily co-infected with river blindness and eyeworm parasites
- Policy statement: a village is safe for mass treatment if, with probability at least 0.9, the proportion of individuals with more than 20,000 parasites/ml blood is at most 0.01

Statistical formulation

Problem. Can estimates of community-level eyeworm prevalence be used to predict proportion of heavily co-infected individuals?

Development data. Individual-level intensities,
Y_{ij} = number of parasites/ml blood,
for individual j in village i

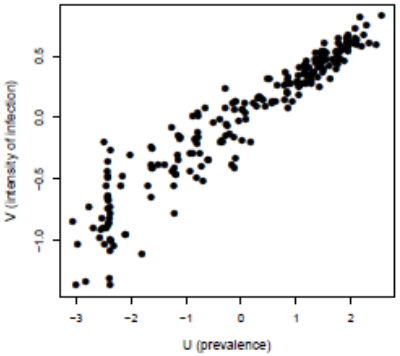


Model. $\mathrm{P}(Y_{ij} > y) = \rho_i \exp\{-(y/\lambda)^\kappa\}$

$\log\{\rho/(1-\rho)\} = \alpha + U$

$\log(\lambda) = \beta + V$

$(U, V) \sim \mathrm{BVN}(0, \Sigma)$



Predictive target.
$T(U, V) = \rho(U) \exp[-\{c/\lambda(V)\}]$

Solution (McCullagh). $[T(U, V)|\text{data}]$, where data is empirical prevalence in a newly sampled community

Embedding statistics in mobile technology

Organisational models

Where should statisticians sit?

- in a Department of Statistics?
- in a Department of Mathematics and Statistics?
- spread around Departments of X, Y, ...?

My ideal:

- a Statistics Institute
  - joint appointments between Institute and Departments X, Y, ...
  - dedicated time (and space) to meet and share ideas

And a pragmatic alternative?

- the beating heart of a Data Science Institute

We are what we teach

Fewer lectures, more projects

Emphasis on general principles rather than specific techniques

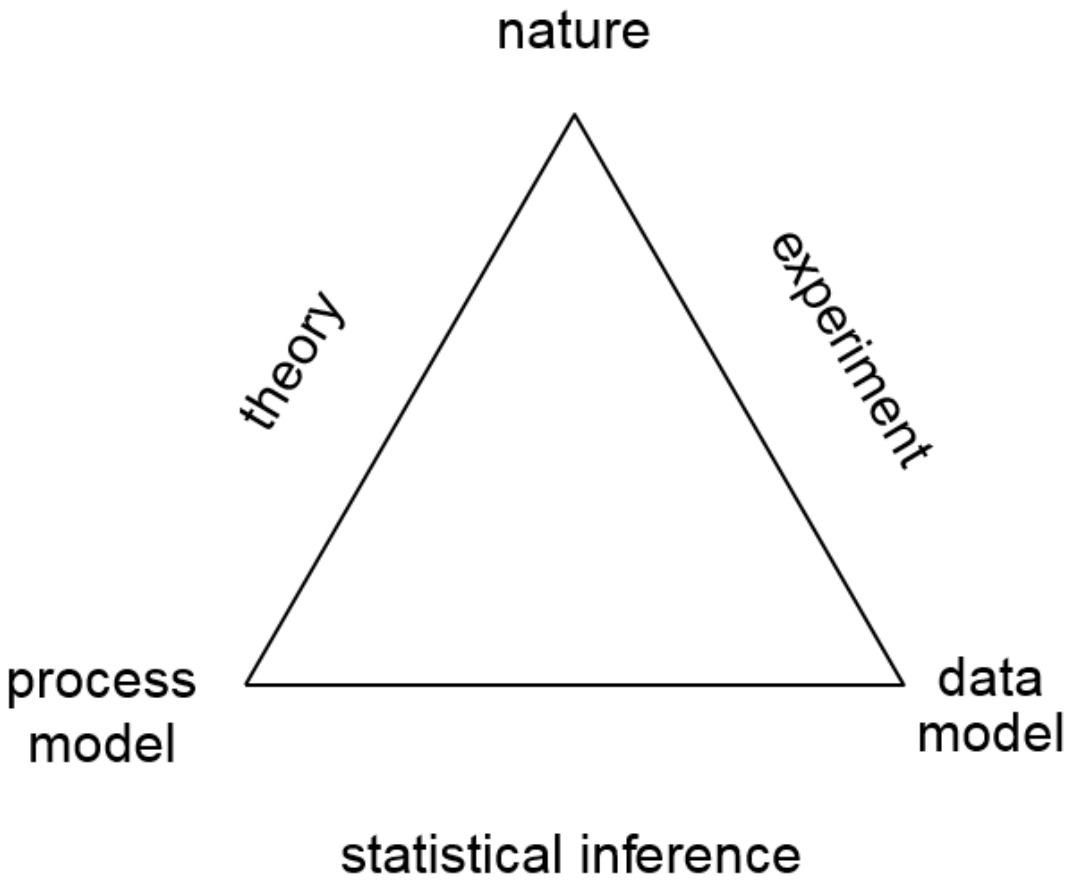
Built on a solid mathematical foundation

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

Meeting the need

- **School**...teach probability as part of maths, statistics as part of science (natural and social)
- **BSc**...focus on single disciplines, especially mathematics (including probability) and computing
- **MSc**...statistics as a postgraduate discipline: begin to develop multi-disciplinarity through project work within scientific teams
- **PhD**...encourage multi-disciplinarity ... team-based projects and co-authored theses
- **PostDoc**...focus on training fellowships, to entice candidates from other disciplines

The science triangle



The data science extension