With low power comes great responsibility

Challenges in modern spatial data analysis

Daniel Simpson

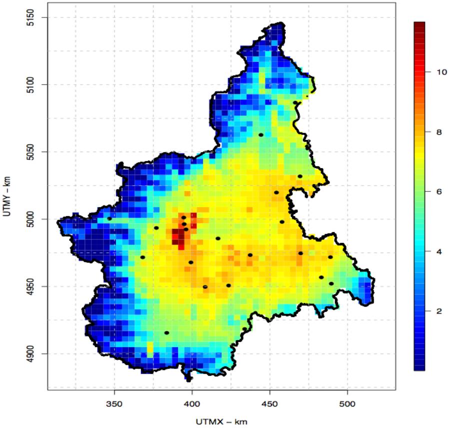
Joint with: Sigrunn Sørbye (Tromsø), Janine Illian (St Andrews / NTNU), Geir-Arne Fuglstad, Haakon Bakka, Håvard Rue (NTNU), Finn Lindgren (Bath)

Centre for Research in Statistical Methodology (CRiSM)
Department of Statistics
University of Warwick

Outline

Marlene Dietrich

On a clear day you can see forever



Daily PM-10 concentration in the Piemonte region, 10/05–03/06.

Gaussian random fields

Defn: Gaussian random fields

A random function $x(s)$ is a GRF iff there is a positive definite function $c(s, s')$ such that, for every finite collection of points $\{s_1, \ldots, s_p\}$,

$$\boldsymbol{x} \equiv (x(s_1), \ldots, x(s_p))^T \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\Sigma_{ij} = c(s_i, s_j)$.

- $\boldsymbol{\Sigma}$ will almost never be sparse or have any structure .
- It is typically very hard to find families of parameterised positive definite functions.
- This is hard for non-stationary, multivariate or spatiotemporal processes.

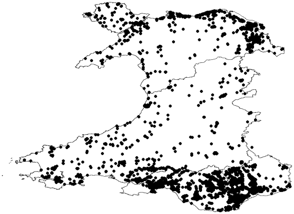
The challenge of big data

- GRFs are lovely models, but they do not scale with the size of a data set
- As data gets more complex, the models often grow as well
- Big data tends to be "observational"—we want to model the truth, not the sampling process
- Big data isn't just hard computationally. It's hard statistically!

Outline

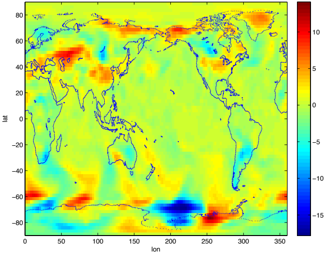
The minotaur justifies the labyrinth

Crime and Koalas

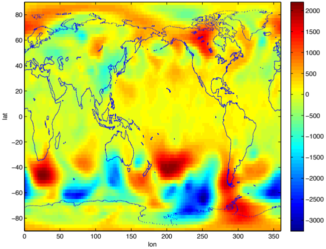


(Left: Antisocial behaviour in Wales. Right: Koalas in Australia)

There's power in a union



(a) Temperature



(b) Pressure

It's not the size of your data, it's how you use it

Key lesson: We cannot use classical models

So what do we give up?

- Point estimation?
- Small area estimation?
- Targeting inference towards quantities of interest?

We need to understand how to build models that answer our questions

A useful example: Log-Gaussian Cox processes

The likelihood *in the most boring case* is

$$\log(\pi(Y|x(s))) = |\Omega| - \int_{\Omega} \Lambda(s)\, ds + \sum_{s_i \in Y} \Lambda(s_i),$$

where Y is the set of observed locations and $\Lambda(s) = \exp(x(s))$, and $x(s)$ is a Gaussian random field.

The is very different from the Gaussian examples: it requires the field everywhere!

If you liked it then you should've put a grid on it

An approximate likelihood

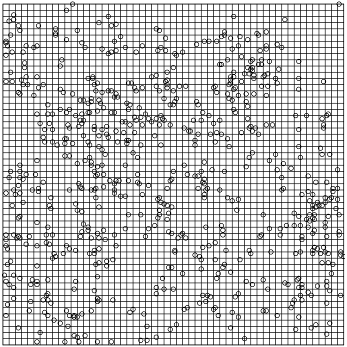
NB: *The number of points in a region* R *is Poisson distributed with mean* $\int_R \Lambda(s)\, ds$.

- Divide the 'observation window' into rectangles.
- Let y_i be the number of points in rectangle i.

$$y_i | x_i, \boldsymbol{\theta} \sim Po(e^{x_i}),$$

- The log-risk surface is replaced with

$$\mathbf{x} | \boldsymbol{\theta} \sim N(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})).$$



But does this lead to valid inference?

Yes—we have perturbation bounds.

- Loosely, the error in the likelihood is transferred exactly (order of magnitude) to the Hellinger distance between the true posterior and the computed posterior.
- This is conditional on parameters.
- For the LGCP example, it follows that, for smooth enough fields $x(s)$, the error is $\mathcal{O}(n^{-1})$

The approximation turns an impossible problem into a difficult, but still useful, problem.

Taking it to the world!

- Approximating the likelihood is not catastrophic
- Approximating the random field is not catastrophic
- Changes a “big data” (i.e. infinite dimensional datum) to a tractable problem
- Is there a lesson here?

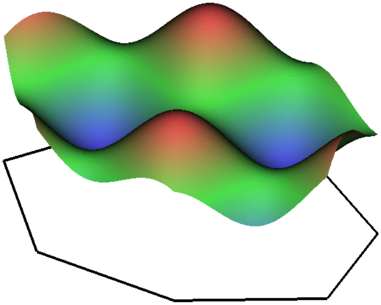
What's Loève but a second hand Karhunen?

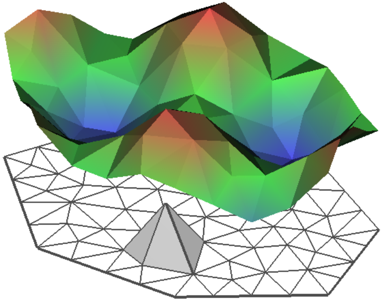
In order to exert some control over the computational cost of spatial problems, it has become common to replace the infinite dimensional GRF $x(s)$ with some finite dimensional version

$$x(s) \approx \sum_{i=1}^{n} w_i \phi_i(s),$$

where $\boldsymbol{w} \sim N(\mathbf{0}, \boldsymbol{Q}^{-1})$ is jointly Gaussian and $\phi_i(s)$ are a set of known deterministic functions.

Video games





NB: The basis functions have compact support.

I choo-choo-choose you!

Consider the Matérn covariance function

$$c(x, y) = \frac{C_\nu}{\kappa^{2\nu}\tau} \left(\kappa \left\| x - y \right\|\right)^\nu K_\nu \left(\kappa \left\| x - y \right\|\right),$$

are the stationary solutions to the SPDE

$$(\kappa^2 - \Delta)^{\frac{\nu + d/2}{2}} x(s) = \tau W(s),$$

where

- $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial s_i^2}$ is the Laplacian
- $W(s)$ is spatial white noise.
- The parameter ν controls the smoothness.
- The parameter κ controls the range.

Stochastic partial differential equation models

Idea

Find the best piecewise linear approximation to a Matérn field!

- This works very well
- You can even show (with effort) that posterior functionals converge like $\mathcal{O}(h^{1-\epsilon})$
- Everything can vary in space, you can add anisotropy, time, advection, etc
- Time is a challenge: all–at–once solvers are nice, but there are obvious problems

The advantage...

- Critically, this method produces a *sparse* $n \times n$ precision matrix, so the cost of a Cholesky goes from $\mathcal{O}(n^3)$ to, say, $\mathcal{O}(n^{3/2})$
- The basis functions have compact support, so evaluating the field at a point only costs $\mathcal{O}(1)$ flops
- This mean that using N data points to predict the field at m unobserved locations costs, for n piecewise linear basis functions is, in two dimensions,

$$\mathcal{O}(N + m + n^{3/2})$$

- If you use basis functions without compact support, this grows to $\mathcal{O}(Nn^2 + mn + n^3)$.

What do we give up?

- Fundamentally, we give up sub-grid variation

What do we give up?

- Fundamentally, we give up sub-grid variation
- Mathematically, this means that we can't get precise answers to "what is $x(s_i) \mid y$?"

What do we give up?

- Fundamentally, we give up sub-grid variation
- Mathematically, this means that we can't get precise answers to "what is $x(s_i) \mid y$?"
- But we can get answers to "How does the approximation affect $\ell(x(\cdot)) \mid y$?" for nice functionals (basically $\ell \in L^2$, not $\ell \in H^{-d/2-\epsilon}$)

What do we give up?

- Fundamentally, we give up sub-grid variation
- Mathematically, this means that we can't get precise answers to "what is $x(s_i) \mid y$?"
- But we can get answers to "How does the approximation affect $\ell(x(\cdot)) \mid y$?" for nice functionals (basically $\ell \in L^2$, not $\ell \in H^{-d/2-\epsilon}$)
- *Open Question:* Can we make a "sub-grid" process to capture this extra variation? (one case solved)

What do we give up?

- Fundamentally, we give up sub-grid variation
- Mathematically, this means that we can't get precise answers to "what is $x(s_i) \mid y$?"
- But we can get answers to "How does the approximation affect $\ell(x(\cdot)) \mid y$?" for nice functionals (basically $\ell \in L^2$, not $\ell \in H^{-d/2-\epsilon}$)
- *Open Question:* Can we make a "sub-grid" process to capture this extra variation? (one case solved)
- *Open Question:* How does the choice of basis function affect inference? (partial results)

Outline

Don't rain on my parade

Marlene Dietrich's career ended in 1975 when she fell off the stage in Sydney and broke her thigh

Information is power

- Spatial data typically only occurs once (i.e. there are no replications)

Information is power

- Spatial data typically only occurs once (i.e. there are no replications)
- Some observations (e.g. Gaussian observation noise) lead to an ergodic field under in-fill
  - This is good! Do what you want!
  - Most of the theory exists for this case

Information is power

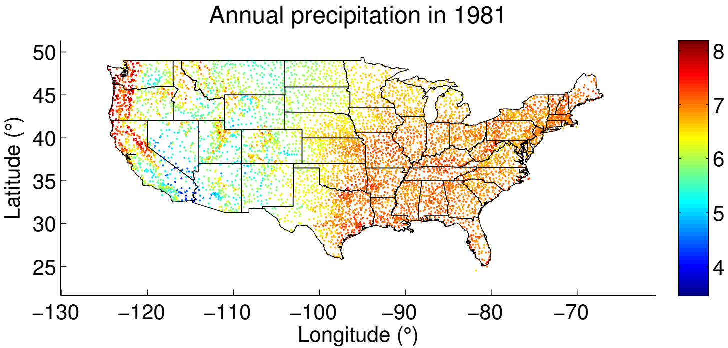
- Spatial data typically only occurs once (i.e. there are no replications)
- Some observations (e.g. Gaussian observation noise) lead to an ergodic field under in-fill
  - This is good! Do what you want!
  - Most of the theory exists for this case
- Some observation processes (e.g. LGCPs) are not ergodic in a fixed window
  - Serious problems!

Information is power

- Spatial data typically only occurs once (i.e. there are no replications)
- Some observations (e.g. Gaussian observation noise) lead to an ergodic field under in-fill
  - This is good! Do what you want!
  - Most of the theory exists for this case
- Some observation processes (e.g. LGCPs) are not ergodic in a fixed window
  - Serious problems!

Remember: *You're data will never overcome your prior!*

Blame it on the rain



A recent comment on Bayes (StatsLife Jan 2015)

Peter Diggle
(president of RSS)

... a lot of what's published, I think, has within it wrinkles that are hidden by the elegance and the simplicity of the Bayeisan formalism. So while people can easily check that their main conclusions are not heavily influenced by pretending to change their prior beliefs, there are subtle aspects that they can't check. I think it's too glib to say that because Bayesian methods are elegant and beautiful they're necessarily the right tools to use in all circumstances.

Failure isn't stationary!

- Real data often displays non-stationary aspects (different correlation structures in different regions)

Failure isn't stationary!

- Real data often displays non-stationary aspects (different correlation structures in different regions)
- A small industry has been built up around doing new, flexible models for this

Failure isn't stationary!

- Real data often displays non-stationary aspects (different correlation structures in different regions)
- A small industry has been built up around doing new, flexible models for this
- In the SPDE approach, we change the "linear filter"

$$(\kappa^2(s) - \nabla \cdot \boldsymbol{H}(s)\nabla)(\tau(s)x(s)) = W(s)$$

Failure isn't stationary!

- Real data often displays non-stationary aspects (different correlation structures in different regions)
- A small industry has been built up around doing new, flexible models for this
- In the SPDE approach, we change the "linear filter"

$$(\kappa^2(s) - \nabla \cdot \boldsymbol{H}(s)\nabla)(\tau(s)x(s)) = W(s)$$

- In theory, $\kappa^2(s)$ controls the local range, $\boldsymbol{H}(s)$ controls the local anisotropy, $\tau(s)$ controls the pointwise variance.

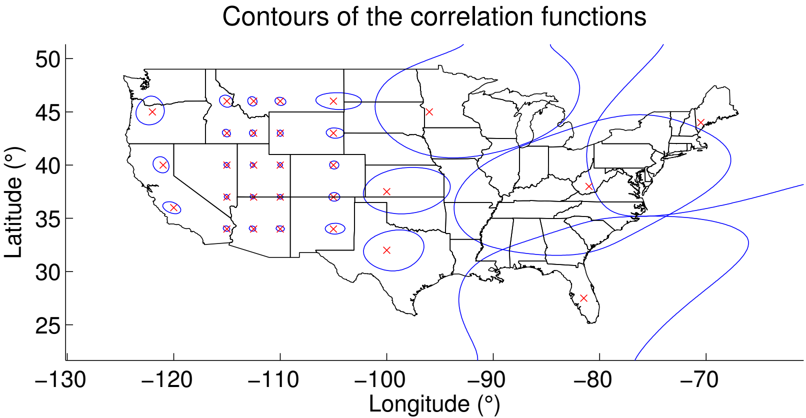
Failure isn't stationary!

- Real data often displays non-stationary aspects (different correlation structures in different regions)
- A small industry has been built up around doing new, flexible models for this
- In the SPDE approach, we change the "linear filter"

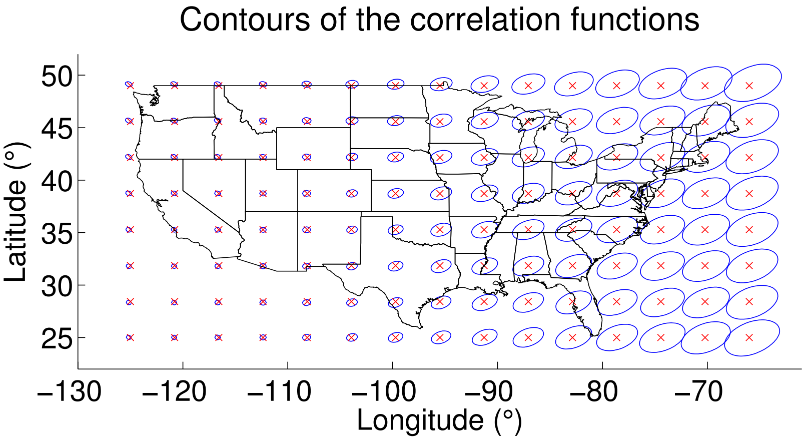
$$(\kappa^2(s) - \nabla \cdot \boldsymbol{H}(s)\nabla)(\tau(s)x(s)) = W(s)$$

- In theory, $\kappa^2(s)$ controls the local range, $\boldsymbol{H}(s)$ controls the local anisotropy, $\tau(s)$ controls the pointwise variance.
- This is not true!

If you open your mind too much, you're brain will fall out



Control


Contours of the correlation functions
Latitude (°)
50
45
40
35
30
25
–130
–120
–110
–100
–90
–80
–70
Longitude (°)

So what went wrong?

It was a bad parameterisation.

For simplicity, let's ignore anisotropy

- Range and variance are approximately separated with the following parameterisation

$$(\kappa^2(s) - \Delta)(\tau(s)x(s)) = \sqrt{4\pi}\kappa(s)W(s)$$

- With a transformation (and $\tau = 1$), this can be interpreted as the stationary random field $(1 - \Delta_E)x(s) = W_E(s)$, where E is $\mathbb{R}^2$ endowed with the Riemannian metric $g(s) = R^{-2}(s)\boldsymbol{I})$.
- Hence, you can view SPDE methods as an intrinsic version of the deformation method of Samson and Guttorp.

Implications for priors

- We can force the range and variance to only vary slowly using a prior
- Shrink towards a *base model* (constant range and variance)
- We couldn't do this without an interpretable parameterisation
- NB: (κ, τ) is more statistically relevant than $(\text{range}, \text{variance})$.

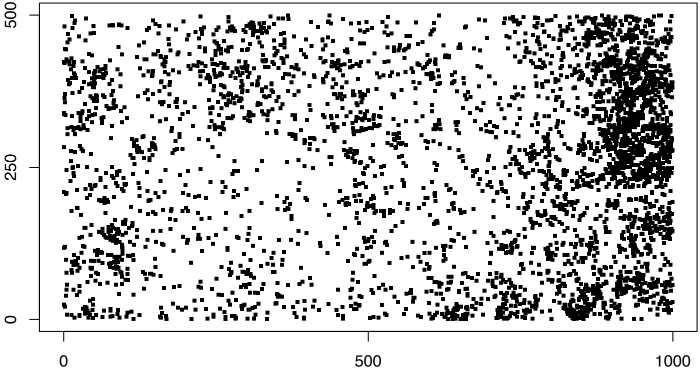
Lesson: Never use a prior that you cannot communicate!

Ronnie, talk to Russia

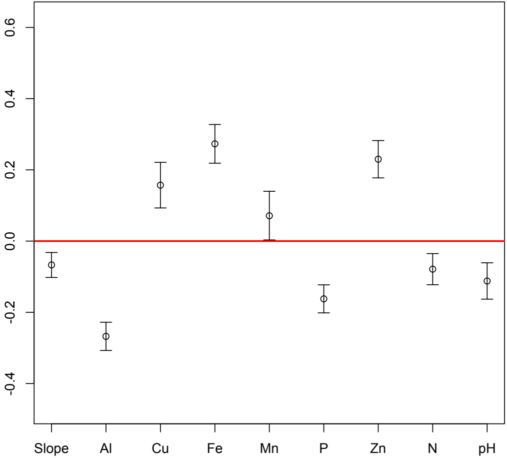
No Repliates, Mo' Problems

- Presence only data occurs frequently in ecology
- Simplest question to ask: How does covariate (xxx) change the local risk of a sighting?
- Basically, is a covariate effect "significant"?
- One big problem: No possibility of replicates.

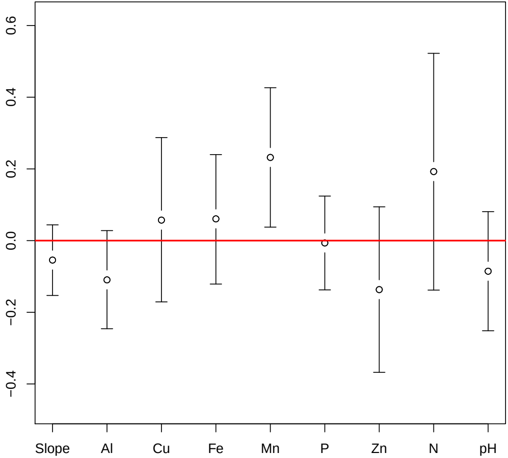
Protium Tenuifolium (4294 trees)



Covariate strength



Covariate strength (with spatial effect)



Oh dear!

- Adding a spatial random effect, which accounts for "un-modelled covariates" massively changes the scientific conclusions
- One solution: Make spatial effect orthogonal to covariates
  - Pro: Cannot "steal" significance
  - Cons: Interpretability, Poor coverage
- This is basically the "maximal covariate effect"
- Without replicates, we cannot calibrate the smoothing parameter to get coverage.

Subjective Bayes to the rescue!

- Key idea: If we can interpret the model, we can talk about the credible intervals as *updates of knowledge*

Subjective Bayes to the rescue!

- Key idea: If we can interpret the model, we can talk about the credible intervals as *updates of knowledge*
- The random field has two parameters: one controlling the range (unimportant) and one controlling the in-cell variance (IMPORTANT!)

Subjective Bayes to the rescue!

- Key idea: If we can interpret the model, we can talk about the credible intervals as *updates of knowledge*
- The random field has two parameters: one controlling the range (unimportant) and one controlling the in-cell variance (IMPORTANT!)
- A prior the variance can be constructed such that

$$\Pr(\text{std}(x_i) > U) < \alpha$$

Subjective Bayes to the rescue!

- Key idea: If we can interpret the model, we can talk about the credible intervals as *updates of knowledge*
- The random field has two parameters: one controlling the range (unimportant) and one controlling the in-cell variance (IMPORTANT!)
- A prior the variance can be constructed such that

$$\Pr(\text{std}(x_i) > U) < \alpha$$

- Changing U changes interpretation

Subjective Bayes to the rescue!

- Key idea: If we can interpret the model, we can talk about the credible intervals as *updates of knowledge*
- The random field has two parameters: one controlling the range (unimportant) and one controlling the in-cell variance (IMPORTANT!)
- A prior the variance can be constructed such that

$$\Pr(\text{std}(x_i) > U) < \alpha$$

- Changing U changes interpretation

  > *The effect of Aluminium is significantly negative when $U < 1$, but the credible crosses zero for all $U > 1$.*

Different random effect strengths

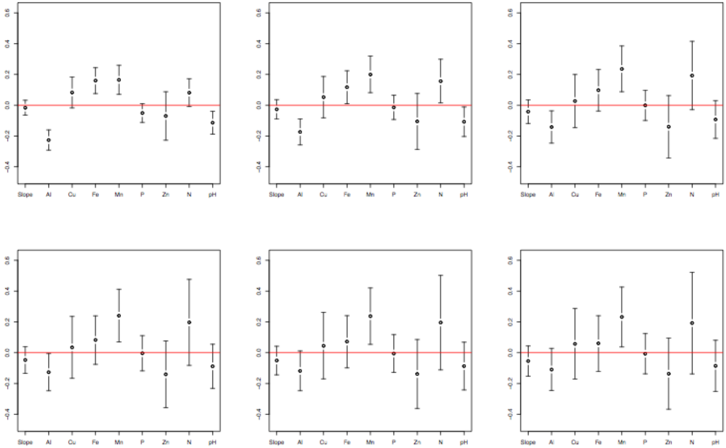


Figure 2: The estimated mean and 95% pointwise credible intervals for the effects of the included observed covariates. $U = (0.01, 0.05, 0.20, 0.50, 1.0, 5.0)$. Mixed model with unconstrained spatial effect.

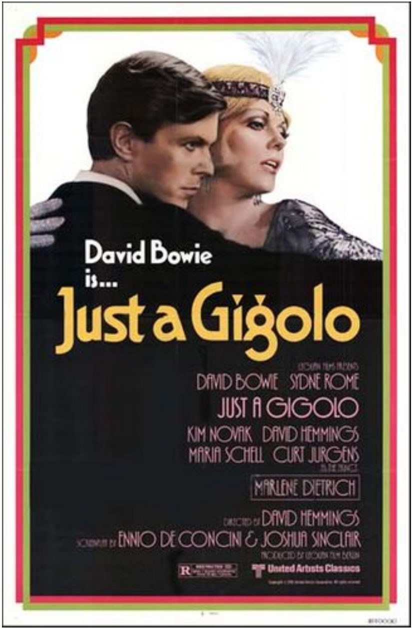
Advantages

- Once again, an interpretable prior allows us to control our inference in a sensible way
- We can talk about updating knowledge
- Explicitly conditioning on the prior allows us to communicate modelling assumptions
- Interpretation without appeals to asymptotics (but well behaved if more observations come)
- Prior and interpretation can/should be made independent of the lattice

Disadvantages

Outline

A final performance



Your models should be as big as elephants. Really simple elephants.

- The over-arching message of this talk is that big data requires us to take a closer look at our methods

Your models should be as big as elephants. Really simple elephants.

- The over-arching message of this talk is that big data requires us to take a closer look at our methods
- More data means that we can fit more flexible models

Your models should be as big as elephants. Really simple elephants.

- The over-arching message of this talk is that big data requires us to take a closer look at our methods
- More data means that we can fit more flexible models
- But we need to be careful not to *over-fit*. *Prefer simplicity!*

Your models should be as big as elephants. Really simple elephants.

- The over-arching message of this talk is that big data requires us to take a closer look at our methods
- More data means that we can fit more flexible models
- But we need to be careful not to *over-fit*. *Prefer simplicity!*
- It's important to think critically about what we want from our analysis and build models that can deal with it

Your models should be as big as elephants. Really simple elephants.

- The over-arching message of this talk is that big data requires us to take a closer look at our methods
- More data means that we can fit more flexible models
- But we need to be careful not to *over-fit*. *Prefer simplicity!*
- It's important to think critically about what we want from our analysis and build models that can deal with it
- When we're only seeing something once (or when we are making process assumptions), it is important to explicitly interpret the results in the light of those assumptions

Your models should be as big as elephants. Really simple elephants.

- The over-arching message of this talk is that big data requires us to take a closer look at our methods
- More data means that we can fit more flexible models
- But we need to be careful not to *over-fit*. *Prefer simplicity!*
- It's important to think critically about what we want from our analysis and build models that can deal with it
- When we're only seeing something once (or when we are making process assumptions), it is important to explicitly interpret the results in the light of those assumptions
- Subjective Bayes gives a formal framework for doing this

With low power comes great responsibility

- Under everything, this was a talk about setting prior distributions

With low power comes great responsibility

- Under everything, this was a talk about setting prior distributions
- This is hard.

With low power comes great responsibility

- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)

With low power comes great responsibility

- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- (Similar things for penalties!)

With low power comes great responsibility

- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- (Similar things for penalties!)
- We recently introduced a general framework for building Penalised Complexity Priors that encode

With low power comes great responsibility

- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- (Similar things for penalties!)
- We recently introduced a general framework for building Penalised Complexity Priors that encode
  - Knowledge about a simpler model

With low power comes great responsibility

- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- (Similar things for penalties!)
- We recently introduced a general framework for building Penalised Complexity Priors that encode
  - Knowledge about a simpler model
  - A penalty on increasing complexity

With low power comes great responsibility

- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- (Similar things for penalties!)
- We recently introduced a general framework for building Penalised Complexity Priors that encode
  - Knowledge about a simpler model
  - A penalty on increasing complexity
  - The graphical structure of the model

With low power comes great responsibility

- Under everything, this was a talk about setting prior distributions
- This is hard.
- Bayesian models should not be used / interpreted unless you can interpret all levels of your model (including your prior)
- (Similar things for penalties!)
- We recently introduced a general framework for building Penalised Complexity Priors that encode
  - Knowledge about a simpler model
  - A penalty on increasing complexity
  - The graphical structure of the model
- We believe that this is a good step in replacing *ad hoc* priors with more principled ones (See arXiv:1403.4630)