Lecture 4: Forecasts, perspectives, simulations

Data Science and Information Theory, ED127 course (2025)

Florent Leclercq

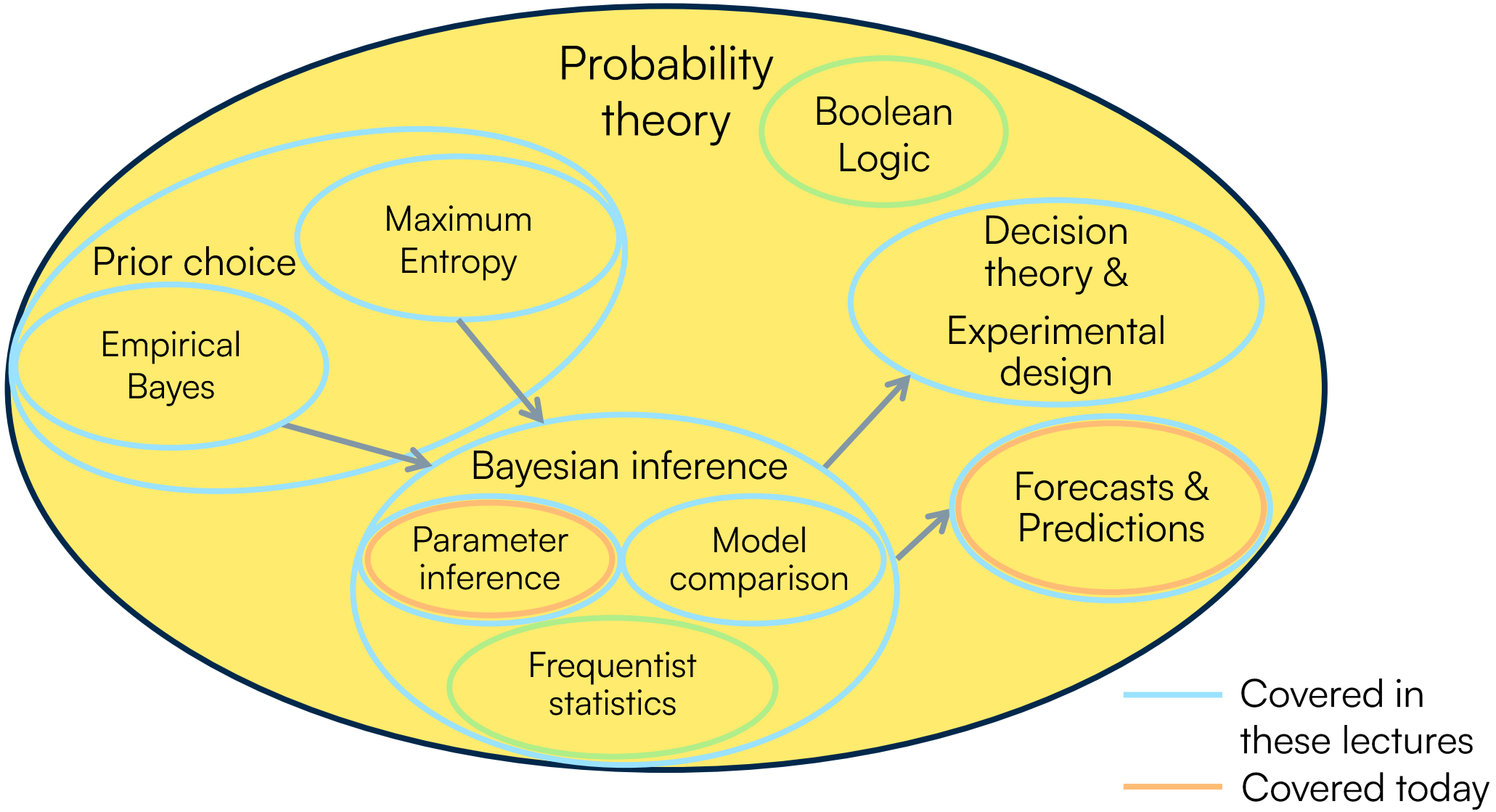
www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

7 April 2025

Kenai Fjords National Park, Alaska

Jaynes's "probability theory": an extension of ordinary Boolean logic



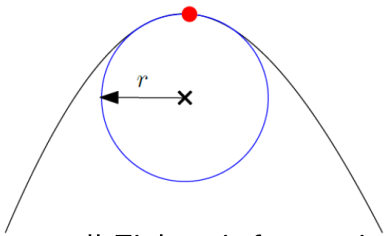
04 FORECASTS

CONDITIONAL AND MARGINAL ERRORS

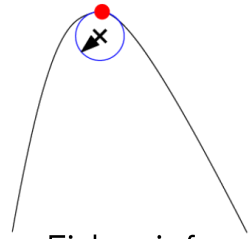
Conditional and marginal errors

- Usual estimators:
  - Maximum a posterior estimator:
    $\theta_{\mathrm{MAP}} = \mathrm{argmax}_\theta \, p(\theta|d)$
  - Posterior mean:
    $\langle\theta\rangle = \int \theta \, p(\theta|d) \, \mathrm{d}\theta$
- An estimator $\hat{\theta}$ is unbiased if its expectation value is the true value θ_0:
  $\left\langle \hat{\theta} \right\rangle = \theta_0$
- To estimate θ, we generally can try to construct an estimator that is:
  - Unbiased
  - With small error, i.e. minimising $\Delta\theta_\alpha \equiv \sqrt{\langle\theta_\alpha^2\rangle - \langle\theta_\alpha\rangle^2}$
  - (in statistics jargon, we want the best unbiased estimator, BUE)

- With Bayes' theorem we know *how* to learn. But how much *can* we learn? The Fisher information (1922) measures the amount of information that a random variable contains about an unknown parameter.

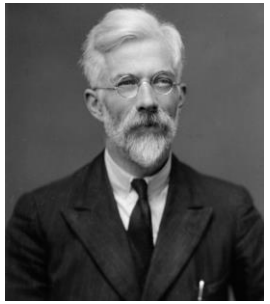


small Fisher information
flat likelihood peak
large variance, low accuracy



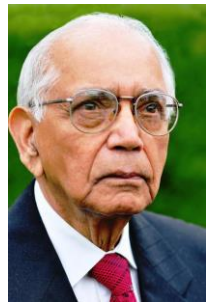
large Fisher information
sharp likelihood peak
small variance, high accuracy

Nielsen, 1808.08271



Ronald Aylmer Fisher (1890-1962)

Harald Cramér (1893-1985)



Calyampudi Radhakrishna Rao (1920-2023)

Conditional and marginal errors

- Assume flat priors and Taylor-expand the log-likelihood around its peak (if it is unique):

(Einstein summation implied)

$$\ln L(\theta) = \underbrace{\ln L(\theta_0) + \frac{1}{2}(\theta_\alpha - \theta_{0\alpha})(\theta_\beta - \theta_{0\beta})\frac{\partial^2 \ln L(\theta)}{\partial\theta_\alpha \partial\theta_\beta}}_{\text{Laplace approximation}} + \dots$$

or $$L(\theta) \approx L(\theta_0) \exp\left[-\frac{1}{2}(\theta_\alpha - \theta_{0\alpha}) H_{\alpha\beta} (\theta_\beta - \theta_{0\beta})\right]$$

- The Hessian matrix element $H_{\alpha\beta} = -\frac{\partial^2 \ln L(\theta)}{\partial\theta_\alpha \partial\theta_\beta}$ controls whether the estimates of parameters θ_α and θ_β are correlated or not.
  - If the Hessian matrix is diagonal the estimates are uncorrelated.
  - Note: this is a statement about the estimates of the quantities, not the quantities themselves, which may be entirely independent, but if they have a similar effect on the data, their estimates may be correlated.

- Conditional errors: if we fix all the parameters but θ_α: $\sigma_{\text{cond},\alpha} = \frac{1}{\sqrt{H_{\alpha\alpha}}}$
  - It is the minimum error bar attainable on θ_α if all other parameters were known. It is rarely relevant and should almost never be quoted.
- Marginal errors: the marginal error on parameter θ_α (marginalising all other parameters) is: $\sigma_\alpha = \sqrt{(H^{-1})_{\alpha\alpha}}$
  - This is normally the relevant error to quote.
  - Proof: it uses the characteristic function to perform "integration by differentiation". See e.g. Heavens (2009) for a demonstration.
- Conditional and marginal errors coincide if H is diagonal. If not, the estimates of the parameters are correlated (even if the parameters themselves are uncorrelated).

Heavens, 0906.0664

FISHER MATRIX ANALYSES

Fisher information matrices

- We try to forecast the result of an experiment, using the same assumptions (flat priors, Gaussian likelihood or Laplace approximation around a single peak), but we haven't got the data yet.
- So we replace the (true, realised) Hessian by its expectation, the Fisher information matrix:

$$F_{\alpha\beta} \equiv \langle H_{\alpha\beta} \rangle = \left\langle -\frac{\partial^2 \ln L(\theta)}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle$$

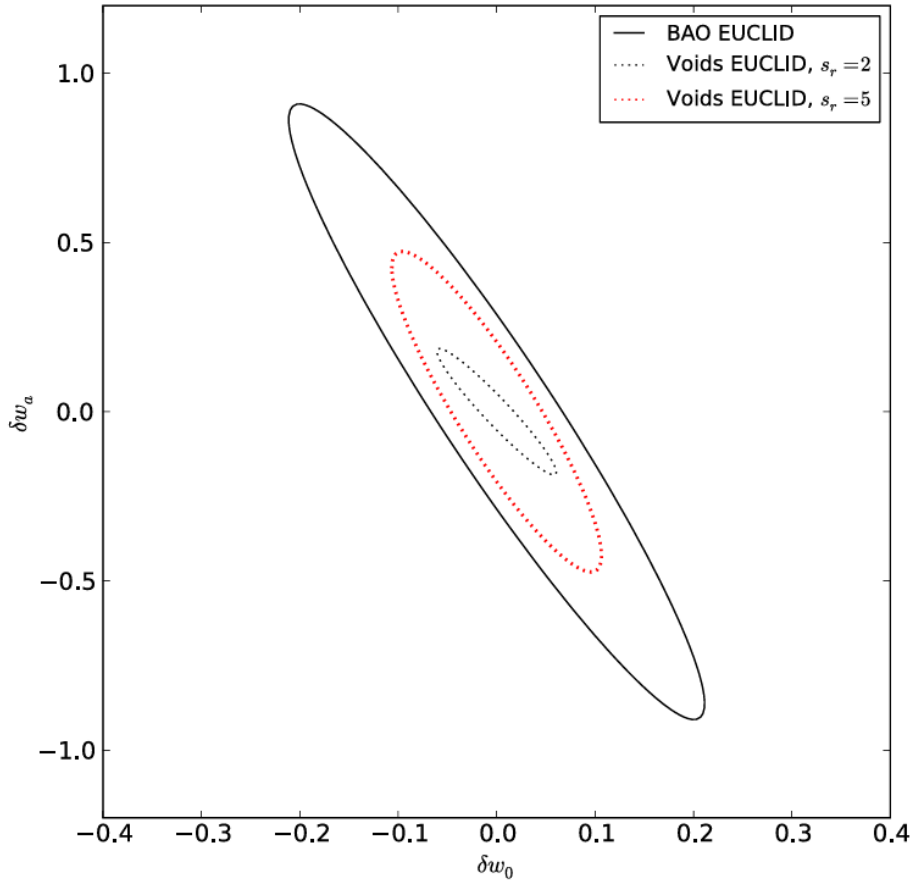
  - Expected conditional error:

$$\sigma_{\mathrm{exp,cond},\alpha} = \frac{1}{\sqrt{F_{\alpha\alpha}}}$$

  - Expected marginal error:

$$\sigma_{\mathrm{exp},\alpha} = \sqrt{(F^{-1})_{\alpha\alpha}}$$

- An example (among very many):



Lavaux & Wandelt, 1110.0345

Cramér-Rao bound

- Let θ_{MLE} be the maximum likelihood estimator (MLE).
- Theorem:
  1. For any unbiased estimator,

$$\Delta\theta_\alpha \geq \frac{1}{\sqrt{F_{\alpha\alpha}}} = \sigma_{\mathrm{exp,cond},\alpha}$$

     (Cramér-Rao inequality or information inequality).
  2. If an estimator, attaining the Cramér-Rao bound ("saturating the information inequality") exists, it is the MLE (or a function thereof).
  3. The MLE is asymptotically the best unbiased estimator (BUE).
- Demonstration: see e.g. Heavens (2009)

- Interpretation:
  1. The Cramér-Rao bound is a lower limit on the error bars achievable by the experiment. You won't do better, but you might do worse.
  2. If there is a best method, the MLE is the one.
  3. In the limit of large data sets, the MLE is the best estimate for all practical purposes.
     - It is these properties that have made maximum likelihood estimators so popular.

Heavens, 0906.0664

The Gaussian likelihood case

Exercise: Fisher information matrix for a Gaussian likelihood

- Assume Gaussian-distributed data,

$$-2\ln L(\theta) = \ln|2\pi C| + (x-\mu)^{\mathsf{T}} C^{-1}(x-\mu)$$

where both μ and C depend on θ in general.

- Then the Fisher information matrix is

$$F_{\alpha\beta} = \frac{1}{2}\mathrm{Tr}\left[C^{-1}C_{,\alpha}C^{-1}C_{,\beta} + C^{-1}(\mu_{,\alpha}\mu_{,\beta}^{\mathsf{T}} + \mu_{,\beta}\mu_{,\alpha}^{\mathsf{T}})\right]$$

(denoting $X_{,\alpha} \equiv \dfrac{\partial X}{\partial\theta_\alpha}$).

- This is a very powerful result: if you know how μ and C depend on the parameters, you can calculate the Fisher information matrix before you do the experiment and get expected errors (in the best case).

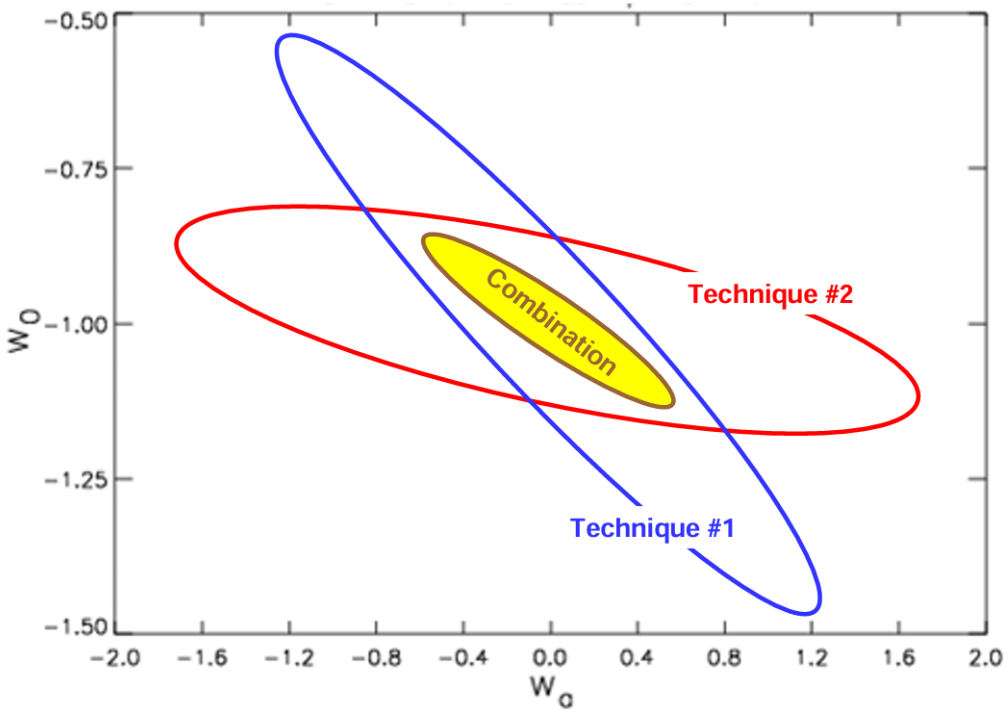
Tegmark, Taylor, Heavens, astro-ph/9603021

- Exercise: demonstrate this result.
- Hints:
  - Define the data matrix $D \equiv (x-\mu)(x-\mu)^{\mathsf{T}}$
  - You may need the following identities:

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$$
$$\ln\det C = \mathrm{Tr}(\ln C)$$
$$(A^{-1})_{,\alpha} = -A^{-1}A_{,\alpha}A^{-1}$$
$$(\ln A)_{,\alpha} = A^{-1}A_{,\alpha}$$

Fisher forecasts in practice

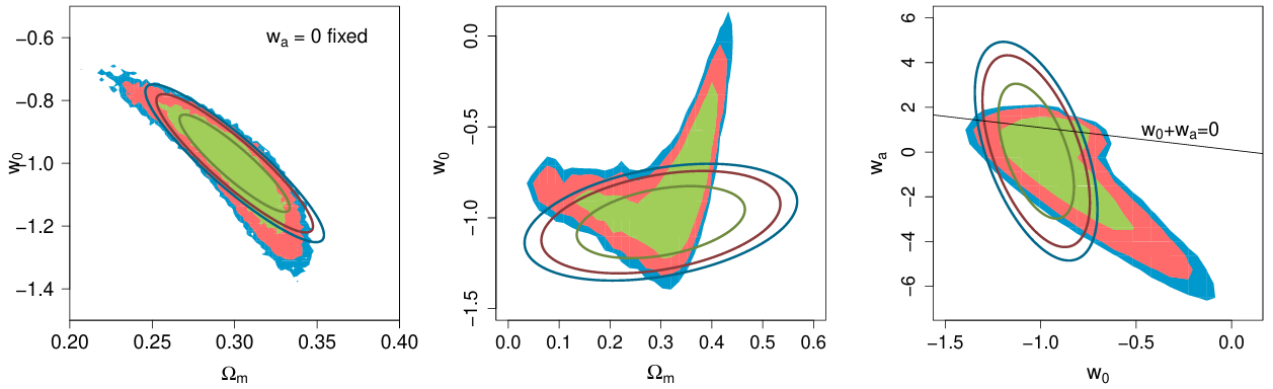
- Fisher forecasts require no data.
- Visualising joint parameter constraints (marginalising over all but two variables):
  - The shape and orientation of ellipses in parameter space indicate correlations.
  - The area of the ellipse is proportional to the volume of parameter space allowed by the data.
- Combining independent experiments (or adding a Gaussian prior): you can simply add the Fisher matrices (why?)
  - Marginal error ellipses in the combined dataset can be smaller than you might expect, given the marginal error ellipses for the individual experiments (because the operations of adding experimental data and marginalising do not commute)



Report of the Dark Energy Task Force (2006), astro-ph/0609591

Advantages and limitations of Fisher forecasts

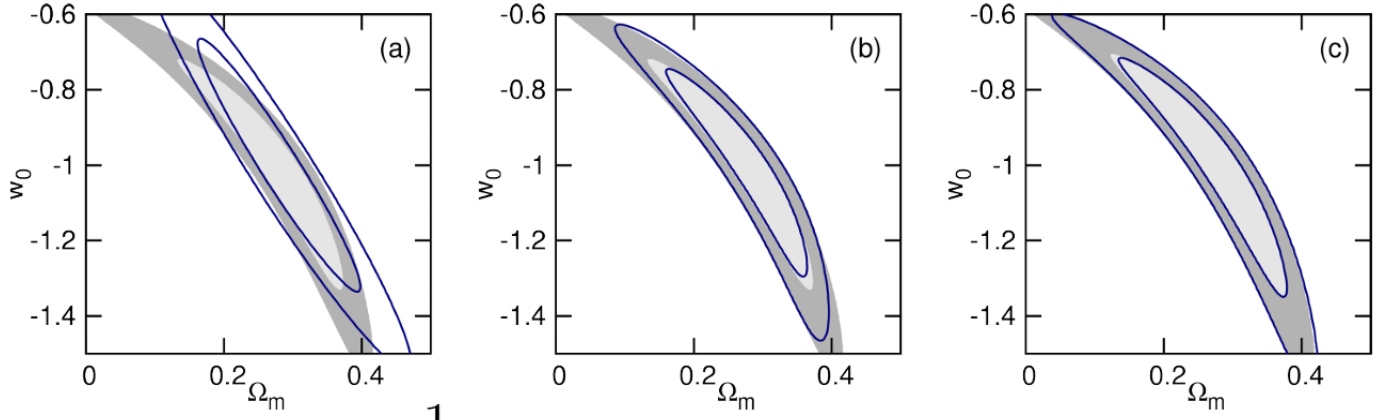
- Advantages:
  - Efficiency: Provides quick, first-order estimates of parameter uncertainties without running full simulations.
  - Simplicity: Offers analytic insight into how different parameters affect the observational signal.
  - Optimisation: Useful tool for experimental design and survey optimisation.
- Limitations:
  - Gaussian assumption: Accuracy depends on the likelihood being well-approximated by a Gaussian near its maximum.
  - Non-linearities: Can underestimate errors for strongly non-linear models or in the presence of significant parameter degeneracies.
  - Systematics: Often does not account for systematic uncertainties unless explicitly modelled.



Wolz *et al.*, 1205.3984

Generalisation of Fisher matrices

- A generalisation of the Fisher matrix is DALI (Derivative Approximation for Likelihoods), which expands the likelihood surface to include higher-order derivatives than the second:



$$\ln L \approx \ln L_0 + \frac{1}{2!} F_{\alpha\beta} \Delta\theta_\alpha \Delta\theta_\beta + \frac{1}{3!} S_{\alpha\beta\gamma} \Delta\theta_\alpha \Delta\theta_\beta \Delta\theta_\gamma + \frac{1}{4!} Q_{\alpha\beta\gamma\delta} \Delta\theta_\alpha \Delta\theta_\beta \Delta\theta_\gamma \Delta\theta_\delta$$

- Other generalisation of Fisher matrices exist, which have been motivated by cosmology:
  - Situations where the data have error bars in both x and y;
  - Expected Bayesian Evidence, generalising Fisher matrices to model selection.

Sellentin, Quartin, Amendola, 1401.6892; Heavens, 1606.06455

BAYESIAN PREDICTION

Predictions: the Bayesian perspective

- In the Bayesian framework, we can use present-day knowledge to produce probabilistic forecasts for the outcome of a future measurement.
- This is *not* limited to assuming a model and parameter value and to determine future errors.

- Extending the power of forecasts: thanks to posterior predictive probabilities, we can extend the scope and power of forecasts:

LEVEL 0:

Assume a model $\mathcal{M}^*$ and a fiducial value for its parameters, θ^*. Produce a forecast for a future experiment assuming $\mathcal{M}^*$ and θ^* are correct.

LEVEL 1:

Average over current model uncertainty within $\mathcal{M}^*$ to forecast future outcomes.

LEVEL 2:

Average over current model uncertainty ($\mathcal{M}_1, \mathcal{M}_2$, ...) to forecast future outcomes.

The posterior predictive distribution

- Applying the usual product rule: $p(y|d) = \iint p(y|d,\theta,\mathcal{M})p(\theta|d,\mathcal{M})p(\mathcal{M}|d)\,\mathrm{d}\theta\mathrm{d}\mathcal{M}$
- **LEVEL 0:** $\mathcal{M} = \mathcal{M}^*$ fixed, $\theta \approx \theta^*$:

$$p(y|d) \approx p(y|d,\theta^*,\mathcal{M}^*)$$

- **LEVEL 1:** average over current parameter uncertainty within $\mathcal{M}^*$

$$p(y|d) \approx p(y|d,\mathcal{M}^*)$$
$$= \int p(y|d,\theta,\mathcal{M}^*)p(\theta|d,\mathcal{M}^*)\,\mathrm{d}\theta$$

Posterior predictive probability = future likelihood weighted by current posterior

- **LEVEL 2:** average over current model uncertainty

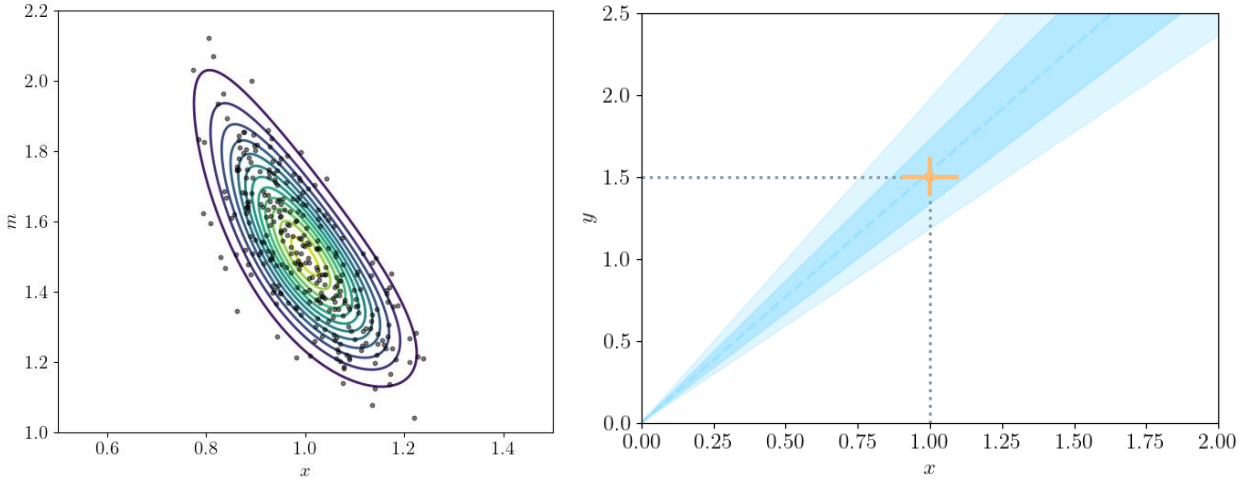
$$p(y|d) = \sum_i p(y|d,\mathcal{M}_i)p(\mathcal{M}_i|d)$$
$$= \sum_i \left(\int p(y|d,\theta_i,\mathcal{M}_i)p(\theta_i|d,\mathcal{M}_i)\,\mathrm{d}\theta_i \right) p(\mathcal{M}_i|d)$$

Posterior predictive probability = average of future likelihoods weighted by current posteriors, weighted by model probabilities

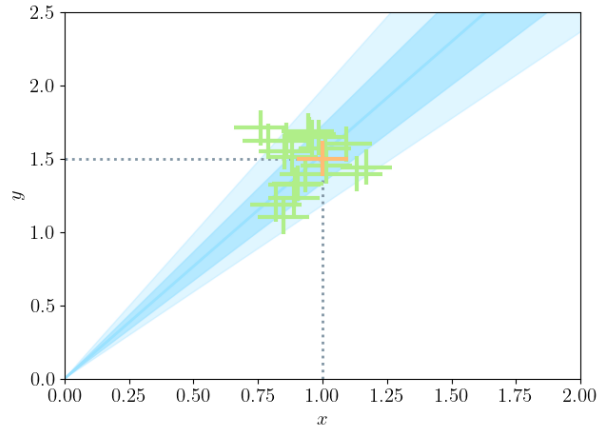
Posterior predictive tests

Exercise: Bayesian linear model – Posterior predictive test

- Let's go back to generalised linear regression (with error bars both in x and y):



- To generate posterior predictive simulations (or constrained simulations), for any sample (x, m):
  - $\tilde{X} \curvearrowleft \mathcal{G}(x, \sigma_x)$
  - $\tilde{Y} \curvearrowleft \mathcal{G}(mx, \sigma_y)$



- The posterior predictive distribution for a new data point $(\tilde{X}, \tilde{Y})$ is:

$$p(\tilde{X}, \tilde{Y}|X, Y) = \iint p(\tilde{X}, \tilde{Y}|m, x) p(m, x|X, Y)\, \mathrm{d}m\, \mathrm{d}x$$

$$= \iint \underbrace{p(\tilde{X}|x)}_{\mathcal{G}(x, \sigma_x)} \underbrace{p(\tilde{Y}|m, x)}_{\mathcal{G}(mx, \sigma_y)} p(m, x|X, Y)\, \mathrm{d}m\, \mathrm{d}x$$

- The currently observed data (X, Y) should look like a typical sample of the posterior predictive test. This is a good check for the internal consistency of the inferred model, and a diagnostic for potential model misspecification.

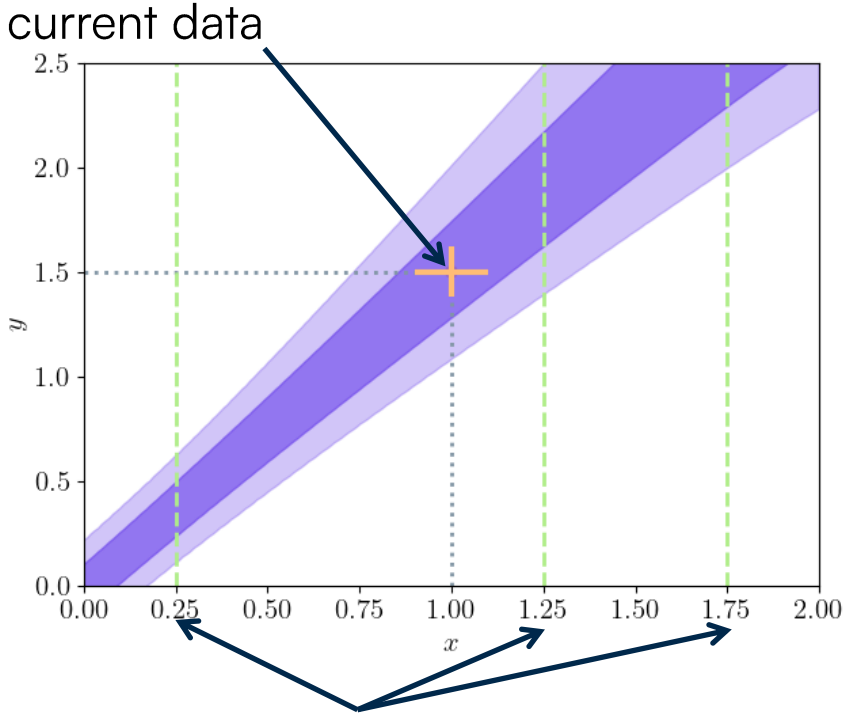
Bayesian prediction of new data

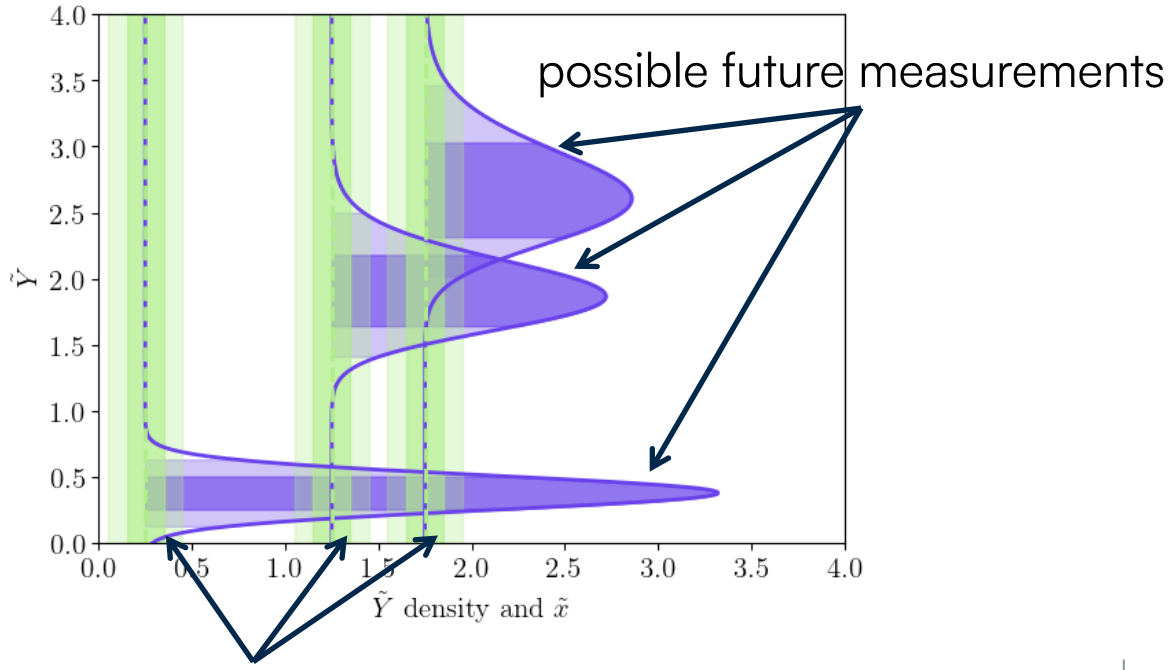
- Assuming a true $\tilde{x}$, the posterior predictive distributions for $(\tilde{X}, \tilde{Y})$ are:

$$p(\tilde{X}|\tilde{x}, X, Y) = p(\tilde{X}|\tilde{x}) = \mathcal{G}(\tilde{x}, \sigma_x)$$

$$p(\tilde{Y}|\tilde{x}, X, Y) = \int p(\tilde{Y}|\tilde{x}, m, X, Y) p(m|\tilde{x}, X, Y)\, \mathrm{d}m$$

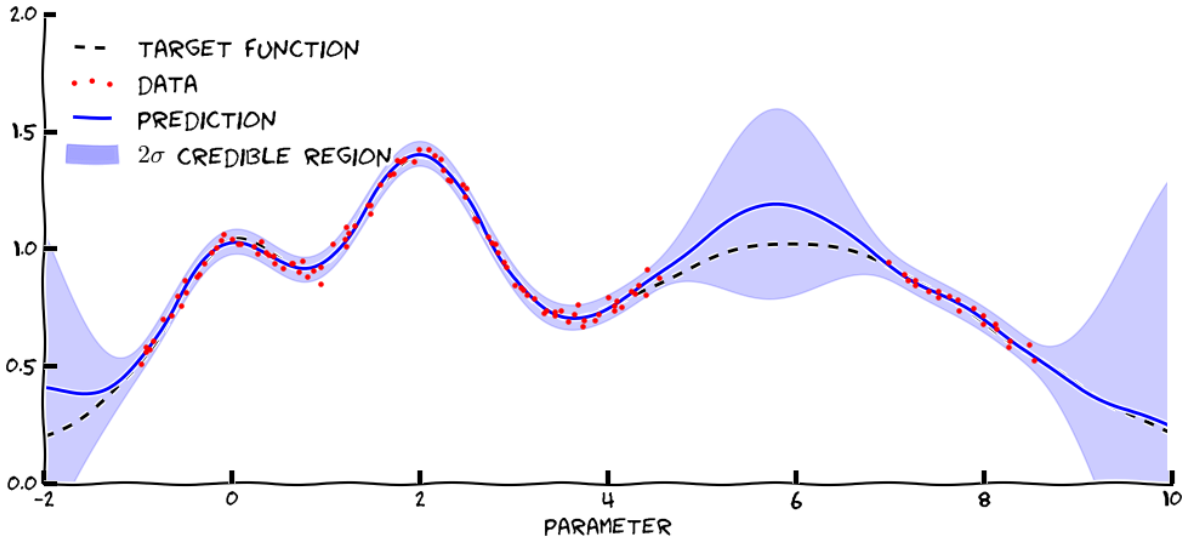
$$= \int p(\tilde{Y}|\tilde{x}, m) p(m|X, Y)\, \mathrm{d}m \quad \propto \int \exp\left[-\frac{1}{2}\frac{(\tilde{Y} - m\tilde{x})^2}{\sigma_y^2}\right] \frac{\sigma_x \sigma_y}{\sqrt{\sigma_y^2 + m^2\sigma_x^2}} \exp\left[-\frac{1}{2}\frac{(Y - mX)^2}{\sigma_y^2 + m^2\sigma_x^2}\right] \mathrm{d}m$$





GAUSSIAN PROCESSES

Gaussian process regression (a.k.a. kriging)



- Why?
  - It is a general-purpose regressor: capable of handling a wide variety of complex and non-linear features of functions.
  - It provides not only a prediction, but also the uncertainty of the regression.
  - It facilitates extrapolation in regions where no data points are available.

Rasmussen & Williams (2006)

- Assume the training set is a Gaussian random field:

$$-2\ln p(\mathbf{f}|\mathbf{X}) = (\mathbf{f}-\boldsymbol{\mu})^\intercal K^{-1}(\mathbf{f}-\boldsymbol{\mu}) + \text{const}$$

$$p(\mathbf{f}|\mathbf{X}) \propto \exp\left[-\frac{1}{2}\sum_{mn}(f(\mathbf{x}_m)-\mu(\mathbf{x}_m)) \times K^{-1}(\mathbf{x}_m,\mathbf{x}_n)(f(\mathbf{x}_n)-\mu(\mathbf{x}_n))\right]$$

- The prediction $f_\star$ at $x_\star$ and the training set form a joint Gaussian random field:

$$p(f_\star|\mathbf{x}_\star,\mathbf{X},\mathbf{f}) \propto \exp\left[-\frac{1}{2}\left(\frac{f_\star-\alpha(\mathbf{x}_\star)}{\sigma(\mathbf{x}_\star)}\right)^2\right]$$

$$\alpha(\mathbf{x}_\star) = \mu(\mathbf{x}_\star) + K(\mathbf{x}_\star,\mathbf{x}_m)^\intercal K^{-1}(\mathbf{x}_m,\mathbf{x}_n)(\mathbf{f}-\mu(\mathbf{X}))_n$$

$$\sigma(\mathbf{x}_\star)^2 = K(\mathbf{x}_\star,\mathbf{x}_\star) - K(\mathbf{x}_\star,\mathbf{x}_m)^\intercal K(\mathbf{x}_m,\mathbf{x}_n)^{-1}K(\mathbf{x}_\star,\mathbf{x}_n)$$

Mean and kernel selection

- There exists a vast literature on mean and kernel selection for Gaussian Processes, depending on the properties and regularity of the target function (see Rasmussen & Williams 2006).
- For the mean, $\boldsymbol{\mu} = \text{const}$ (often zero, simplest choice), or $\mu(x) = \sum_j a_j x_j^2 + b_j x_j + c$ are typical choices.
- For the kernel, a typical choice is:

$$K(\mathbf{x}_m, \mathbf{x}_n) = C_1 \times \exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}_m - \mathbf{x}_n}{C_2}\right)^2\right] + C_3 \delta_{\mathrm{K}}^{mn}$$

with:

  - $K_{\mathrm{C}}(C_1) \equiv C_1$: constant kernel,
  - $K_{\mathrm{RBF}}(C_2) \equiv \exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}_m - \mathbf{x}_n}{C_2}\right)^2\right]$: radial basis function kernel,
  - $K_{\mathrm{GN}}(C_3) \equiv C_3 \delta_{\mathrm{K}}^{mn}$: Gaussian noise kernel

- Interpretation:
  - $\sigma_f \equiv K_{\mathrm{C}}(C_1) \times K_{\mathrm{RBF}}(C_2)$ is the signal variance (marginal variance at x if the observation noise was zero),
  - $\sigma_n \equiv K_{\mathrm{GN}}(C_3)$ is the noise variance.
- There exists optimisation techniques to automatically adjust hyperparameters C_1, C_2, C_3 during the regression.

Rasmussen & Williams (2006)

04

IMPLICIT LIKELIHOOD INFERENCE

Simulations and implicit likelihood

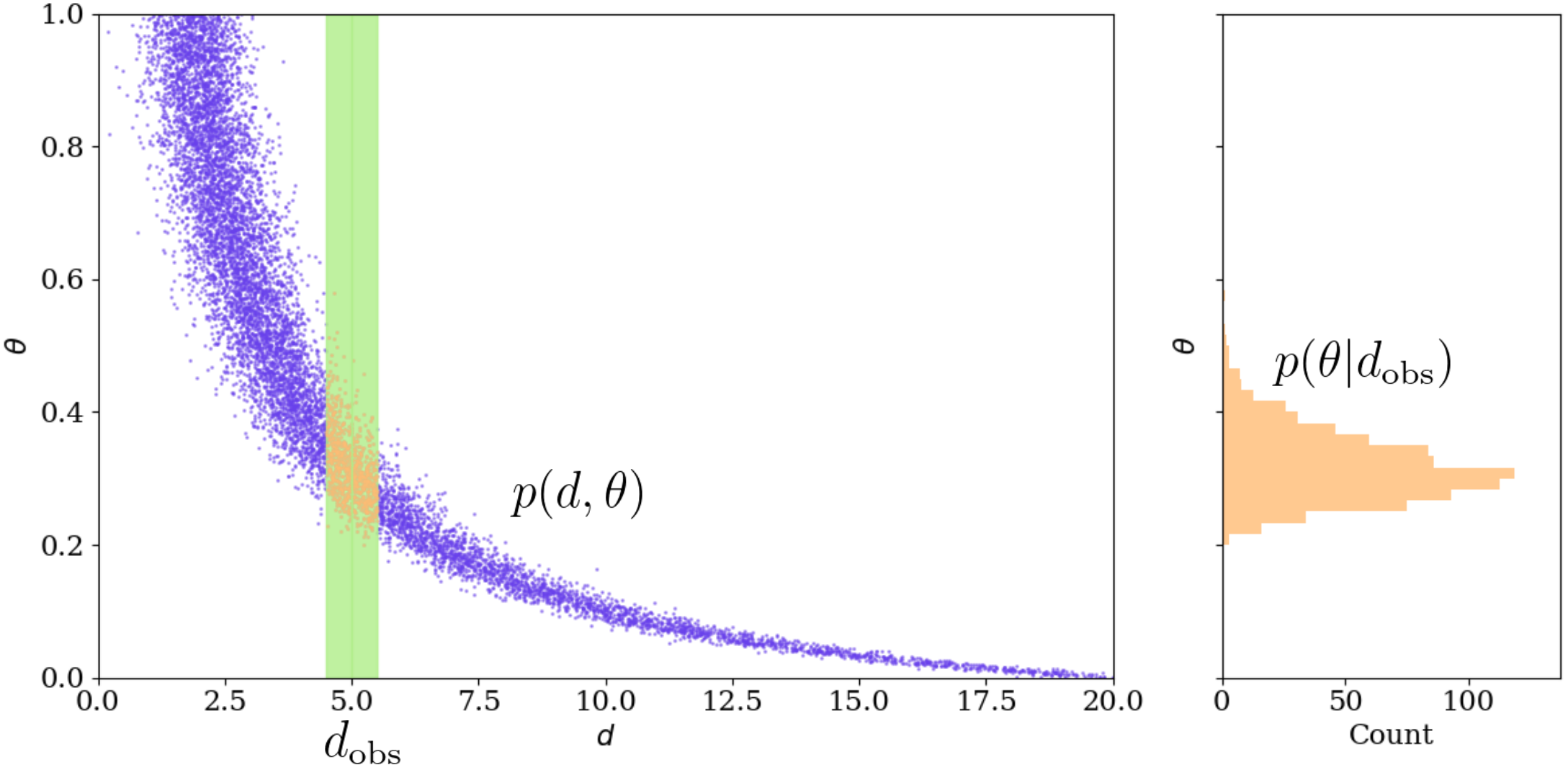
- Simulated data are nothing other than draws from the likelihood!

$$\mathbf{d}_{\text{sim}} \curvearrowleft p(\mathbf{d}|\boldsymbol{\theta})$$

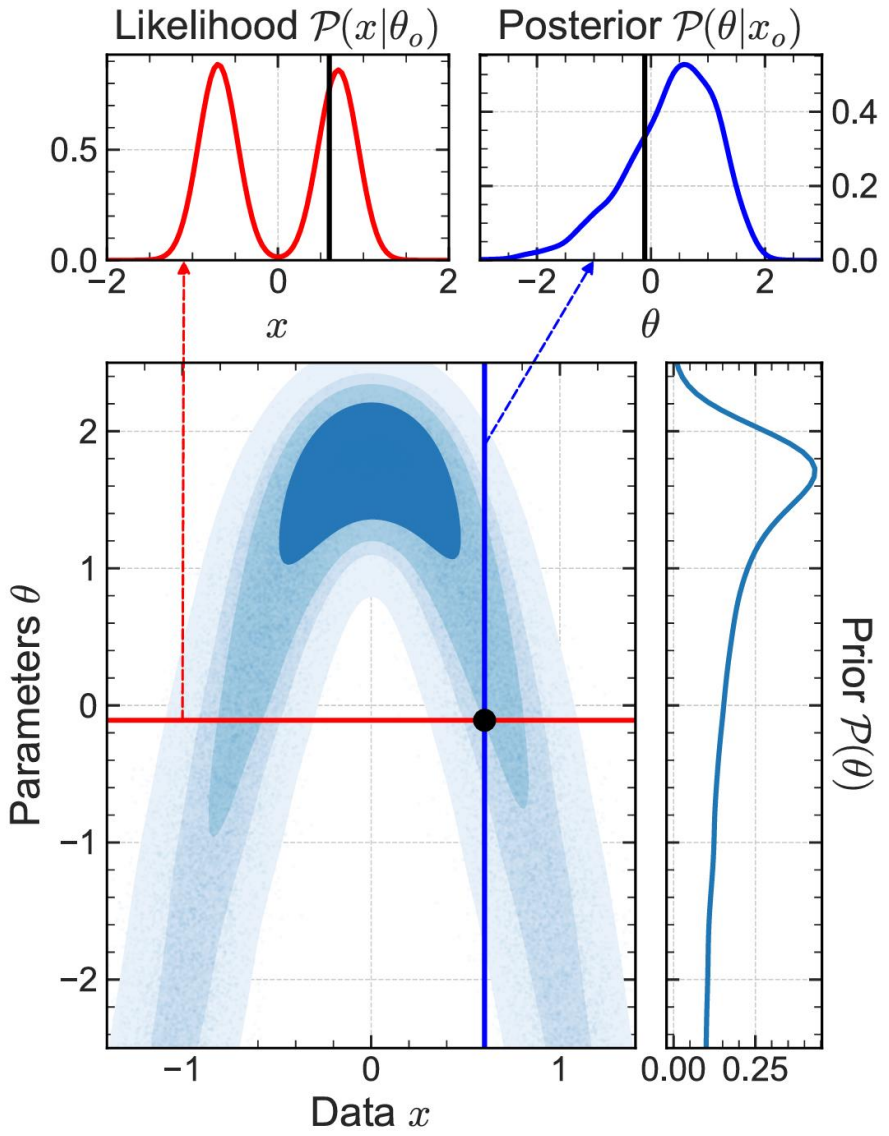
- You may not be able to score variables with your likelihood (an “intractable likelihood”), but you are always able to sample it (i.e. draw samples = simulations).
  - Because if you can’t even simulate the data, you’re in big trouble… The only way forward is to make simplifying assumptions about the data-generating process.

> What I cannot create, I do not understand.
> R. Feynman

Everything is included in the joint probability of parameters and data



Everything is included in the joint probability of parameters and data



Ho *et al.*, 2402.05137

Implicit (likelihood) inference

- We consider the Bayesian problem of inferring parameters, using Bayes' theorem, when the some of the pdfs are implicitly defined by a simulator:

$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{d})}$$

- This is implicit inference:
  - When the likelihood and/or the prior are not explicitly specified, but implicit in simulations, generative models, labelled data, calibration data, etc.
  - Usually it is the likelihood that is implicit, and we talk about implicit likelihood inference (ILI)
- There are various forms of ILI also known as
  - Approximate Bayesian Computation (ABC)
  - Likelihood-Free inference
  - Simulation-based inference

Implicit (likelihood) inference

- Some vocabulary considerations:
  - "Likelihood-free inference" needs to be understood as "free of an *explicit* likelihood that I can score". The opposite would be "likelihood-based inference" (via, e.g. MCMC).
  - "Simulation-based inference" needs to be understood as "inference based *only* on simulations". Otherwise, essentially any Bayesian inference problem is simulation-based (scoring from the likelihood also involves running a simulation).
  - In "implicit likelihood inference"...
    - It's not necessarily the likelihood that is implicit (but usually, it is). In this case we can say "implicit inference".
    - And it's not necessarily inference that we want to do, but anything involving probabilities, e.g. prediction or decision-making. See for example "Optimal simulation-based Bayesian decisions" (Alsing et al., 2311.05742)

APPROXIMATE BAYESIAN COMPUTATION

Approximate Bayesian Computation (ABC)

- Statistical inference for models where:
  - The likelihood function is intractable
  - Simulating data is possible
- General idea: find parameter values for which the distance between simulated data and observed data is small.
- The exact posterior is replaced by an approximate posterior:

$$p(\theta|d) \Longrightarrow p(\theta|\tilde{d}) \quad \text{where} \quad \mathrm{d}(\tilde{d}(\theta), d) \quad \text{is small}$$

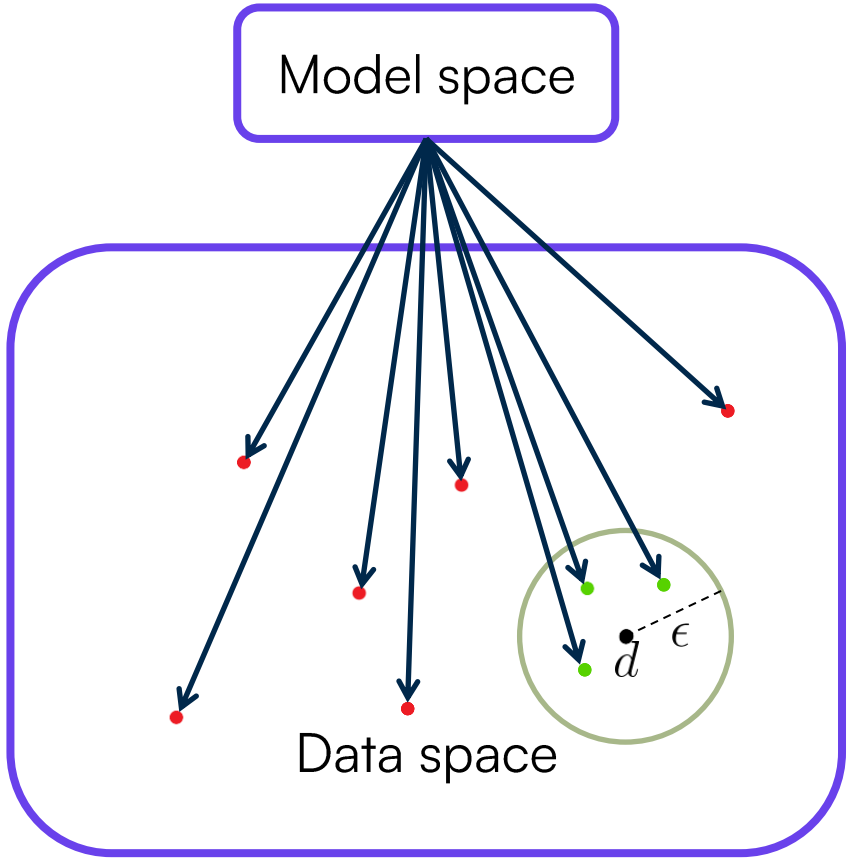
- Assumptions:
  - Only a small number of parameters are of interest
  - But the process generating the data is a very general “black box”:
    it can be a noisy non-linear dynamical system with an unrestricted number of hidden variables

Likelihood-free rejection sampling

- Iterate many times:
  - Sample θ from a proposal distribution $q(\theta)$
  - Simulate $\tilde{d}(\theta)$ according to the data model
  - Compute distance $\mathrm{d}(\tilde{d}(\theta), d)$ between simulated and observed data
  - Retain θ if $\mathrm{d}(\tilde{d}(\theta), d) \leq \epsilon$, otherwise reject

- Effective likelihood approximation:

$$L(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\mathrm{d}(\tilde{d}(\theta), d) \leq \epsilon\right)$$



ADVANCED IMPLICIT LIKELIHOOD INFERENCE

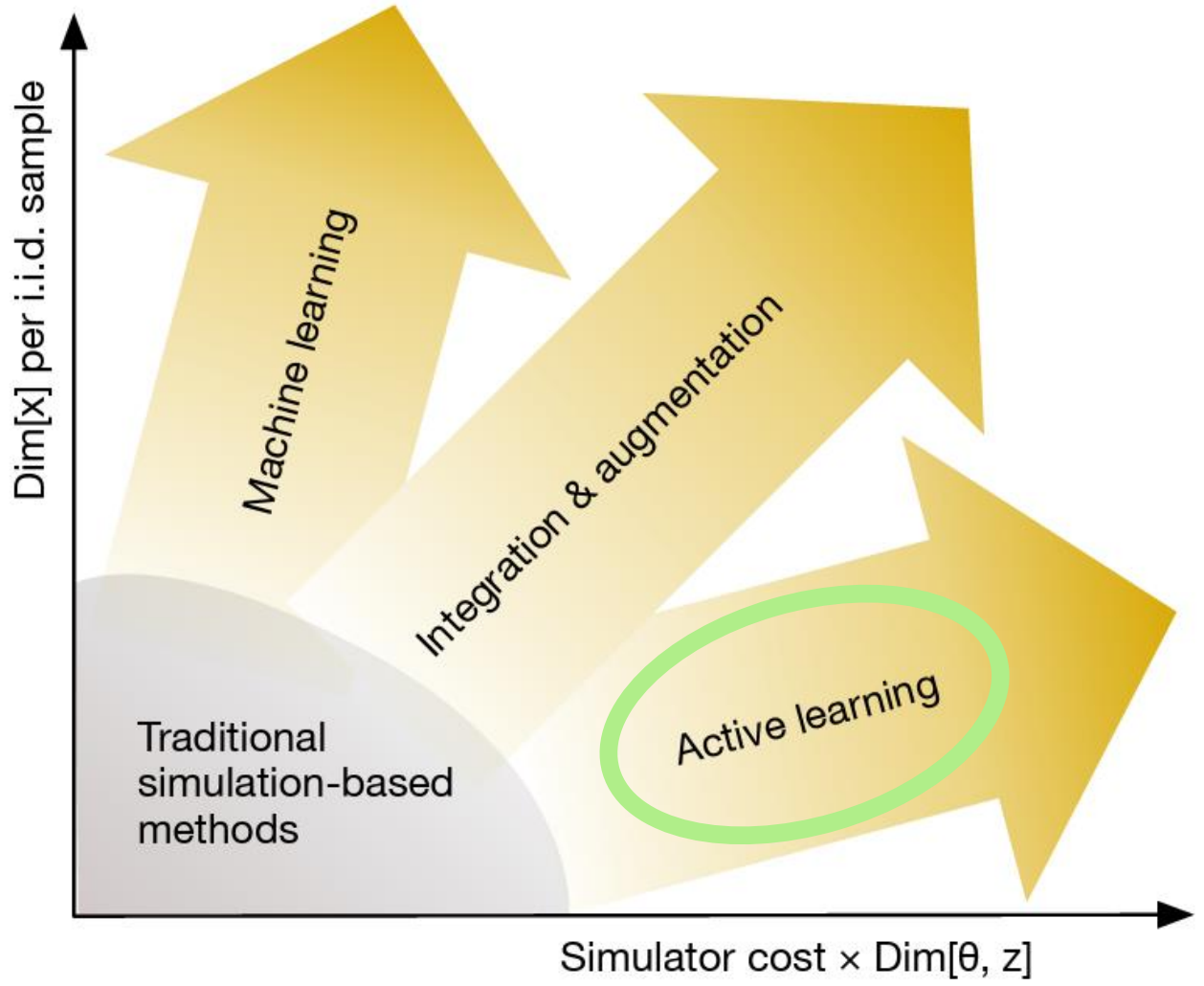
Challenges for implicit inference

1. Curse of dimensionality: the simplest version of implicit inference, ABC, becomes exponentially difficult when the dimension is high.

2. Retainment of the information: implicit inference requires (massive) data compression, so we need to find informative summaries of the data.

3. Model misspecification: any uncertainty on the data-generating process can lead to misspecified models, harder to deal with than with explicit likelihood techniques — where additional parameters can be introduced and marginalised over (but see Leclercq 2022 for an approach).

Avenues beyond traditional implicit inference techniques



Cranmer, Brehmer, Louppe, 1911.01429

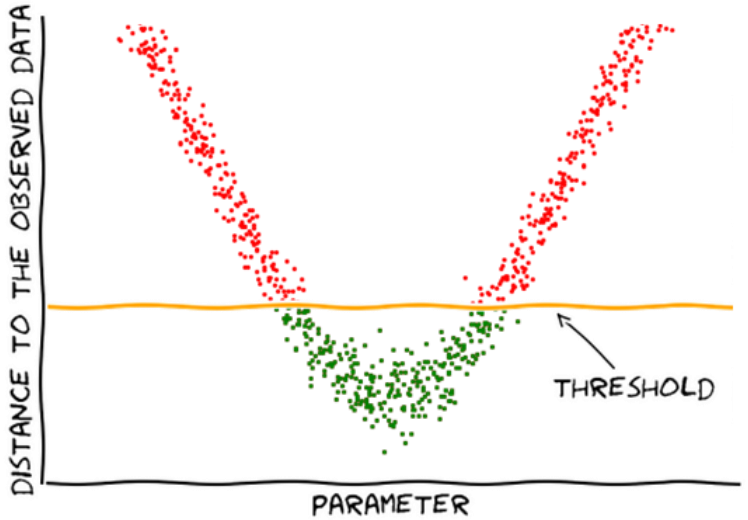
Why is likelihood-free rejection sampling so expensive?

1. It rejects most samples when ϵ is small

2. It does not make assumptions about the shape of $L(\theta)$

3. It uses only a fixed proposal distribution, not all information available

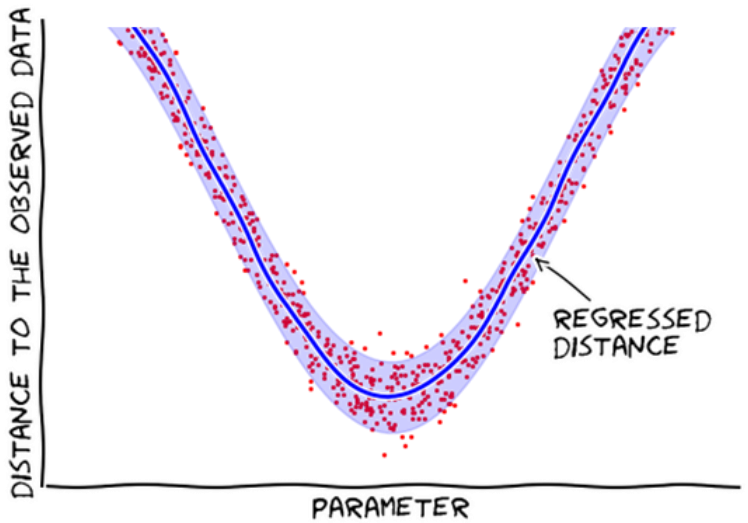
4. It aims at equal accuracy for all regions in parameter space



$$L(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\mathrm{d}(\tilde{d}(\theta), d) \leq \epsilon\right)$$

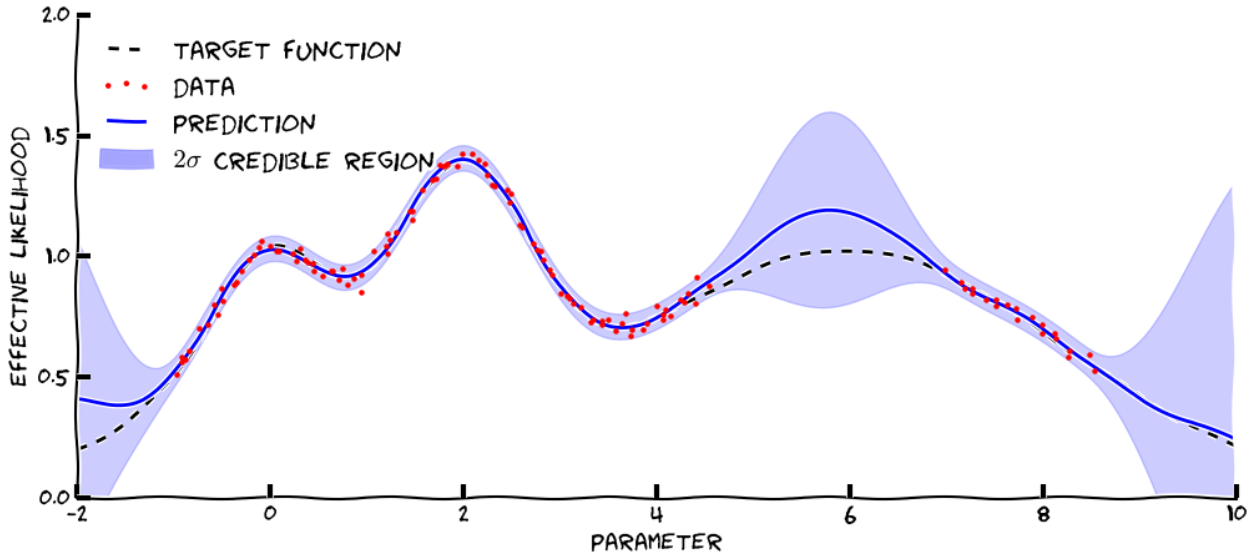
Proposed solution: Bayesian optimisation for likelihood-free inference (BOLFI)

1. It rejects most samples when ϵ is small
   ➡ Don't reject samples: learn from them!
2. It does not make assumptions about the shape of $L(\theta)$
   ➡ Model the distances, assuming the average distance is smooth
3. It uses only a fixed proposal distribution, not all information available
   ➡ Use Bayes' theorem to update the proposal of new points
4. It aims at equal accuracy for all regions in parameter space
   ➡ Prioritize parameter regions with small distances to the observed data



Gutmann & Corander, 1501.03291; FL, 1805.07152

Regressing the effective likelihood (points 1 & 2)

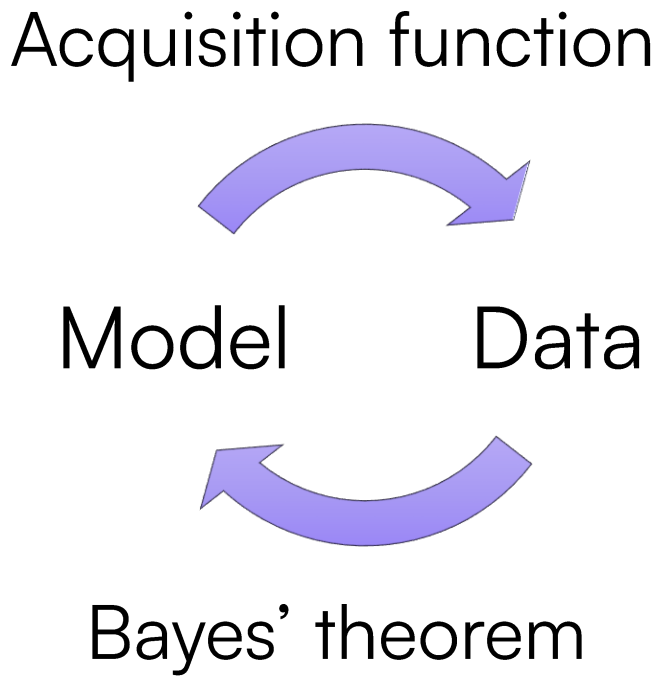


1. "It rejects most samples when ϵ is small"
   - Keep all values $(\theta_i, \mathrm{d}_i) \qquad \mathrm{d}_i = \mathrm{d}(\tilde{d}(\theta_i), d)$
2. "It does not make assumptions about the shape of $L(\theta)$"
   - Model the conditional distribution of distances given this training set, using Gaussian process regression

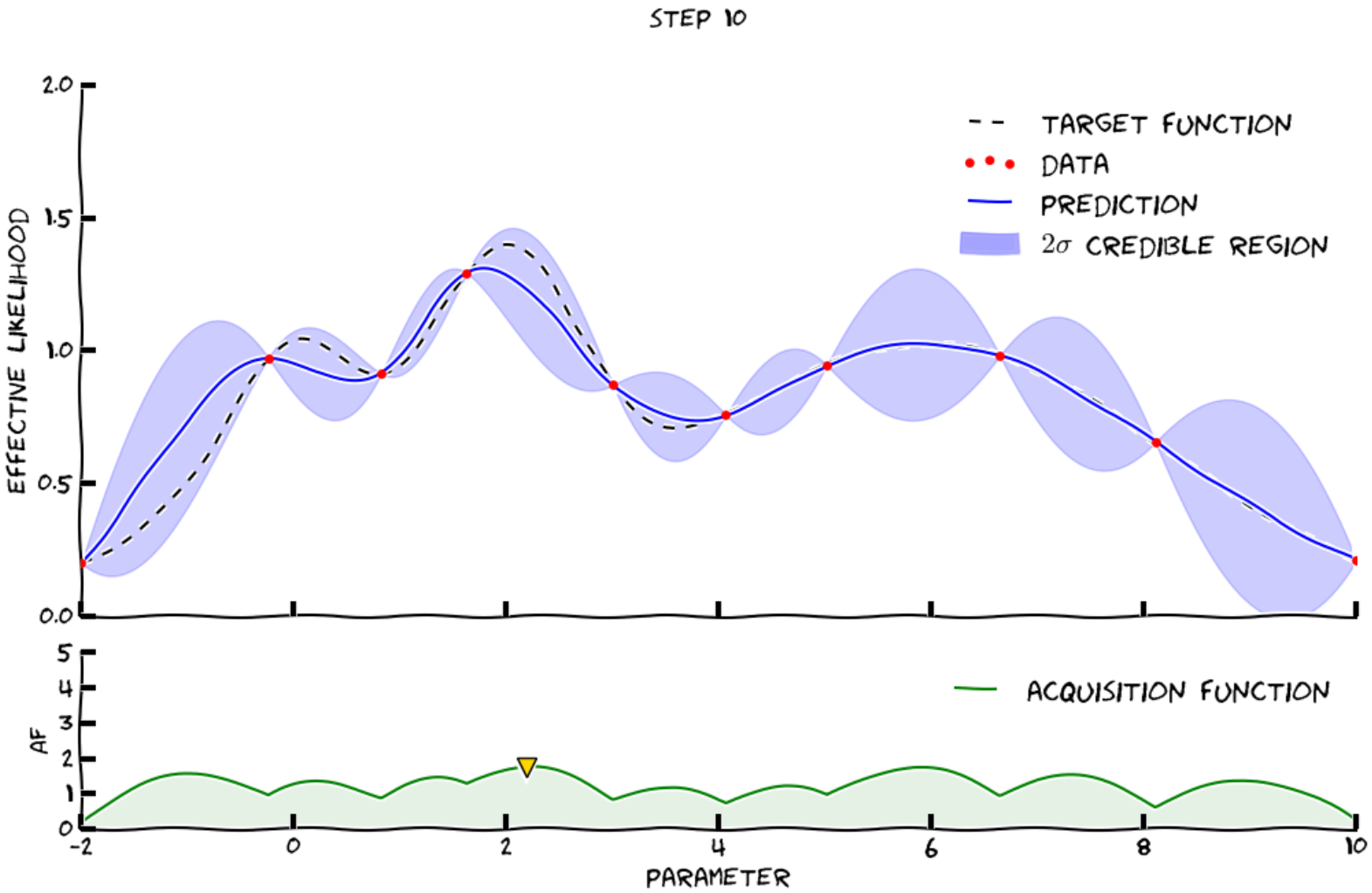
Data acquisition (points 3 & 4)

3. “It uses only a fixed proposal distribution, not all information available”
   - Samples are obtained from sampling an adaptively-constructed proposal distribution, using the regressed effective likelihood

4. “It aims at equal accuracy for all regions in parameter space”
   - The acquisition function finds a compromise between exploration (trying to find new high-likelihood regions) & exploitation (giving priority to regions where the distance to the observed data is already known to be small)
   - Bayesian optimisation (decision making under uncertainty) can then be used



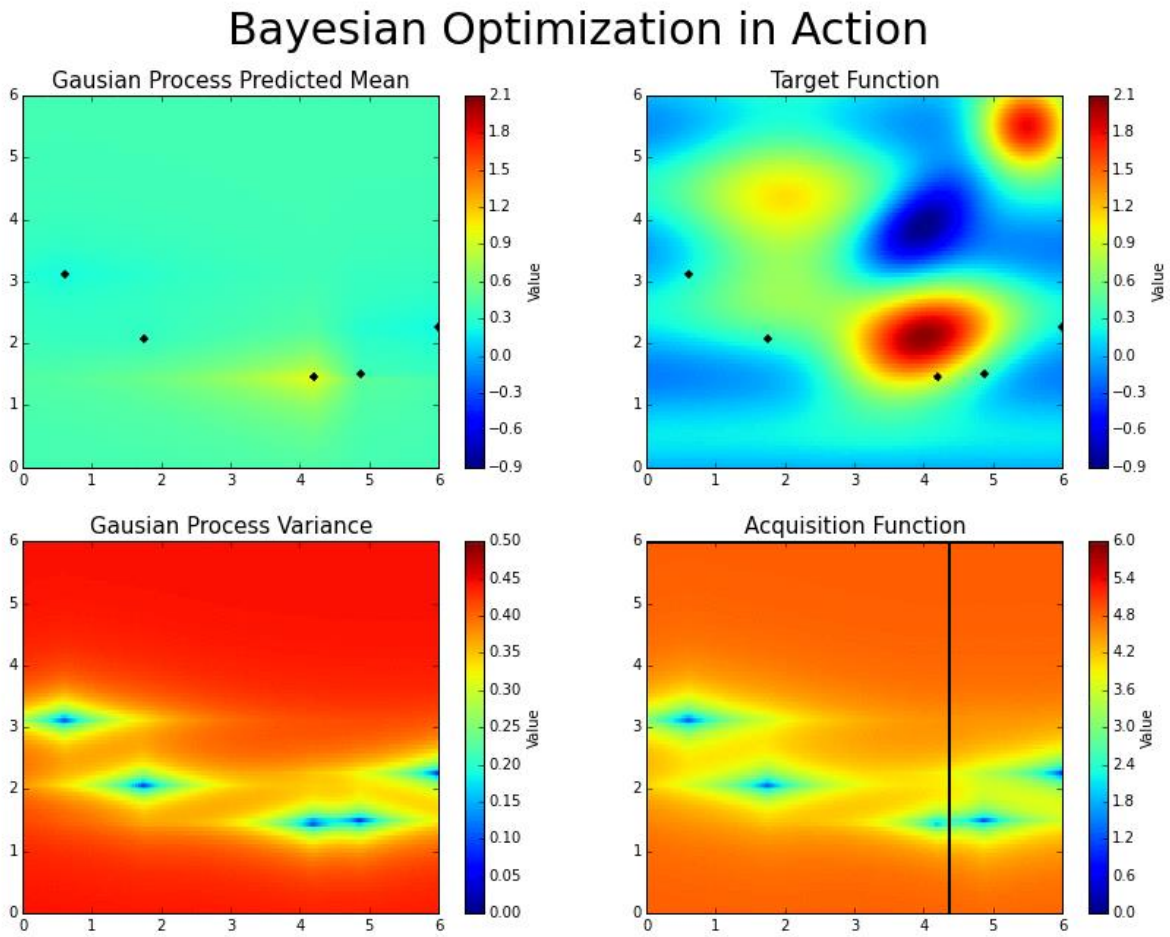
Data acquisition: example



FL, 1805.07152

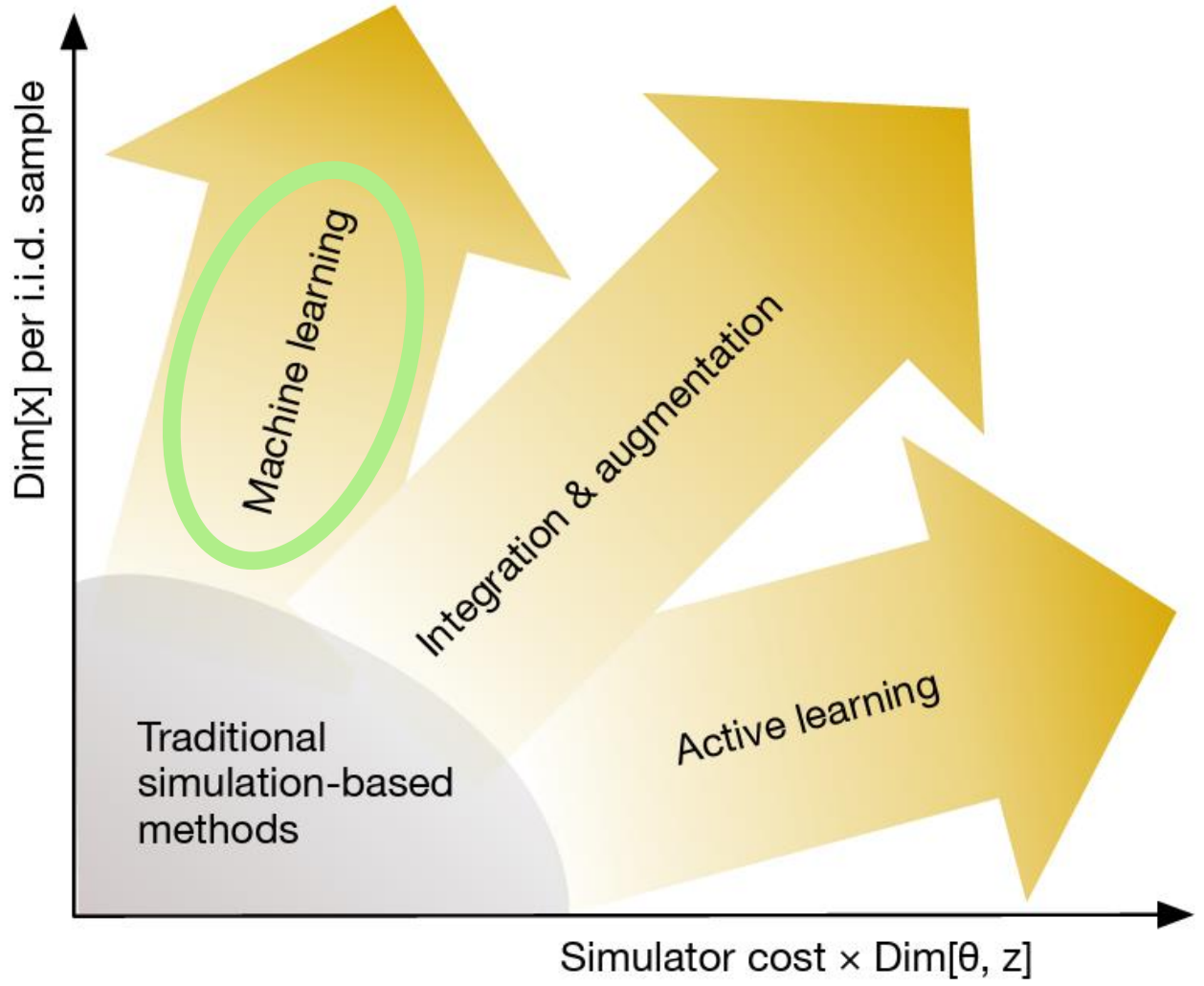
Data acquisition: example

- In higher dimension…



F. Nogueira, https://github.com/fmfn/BayesianOptimization

Avenues beyond traditional implicit inference techniques

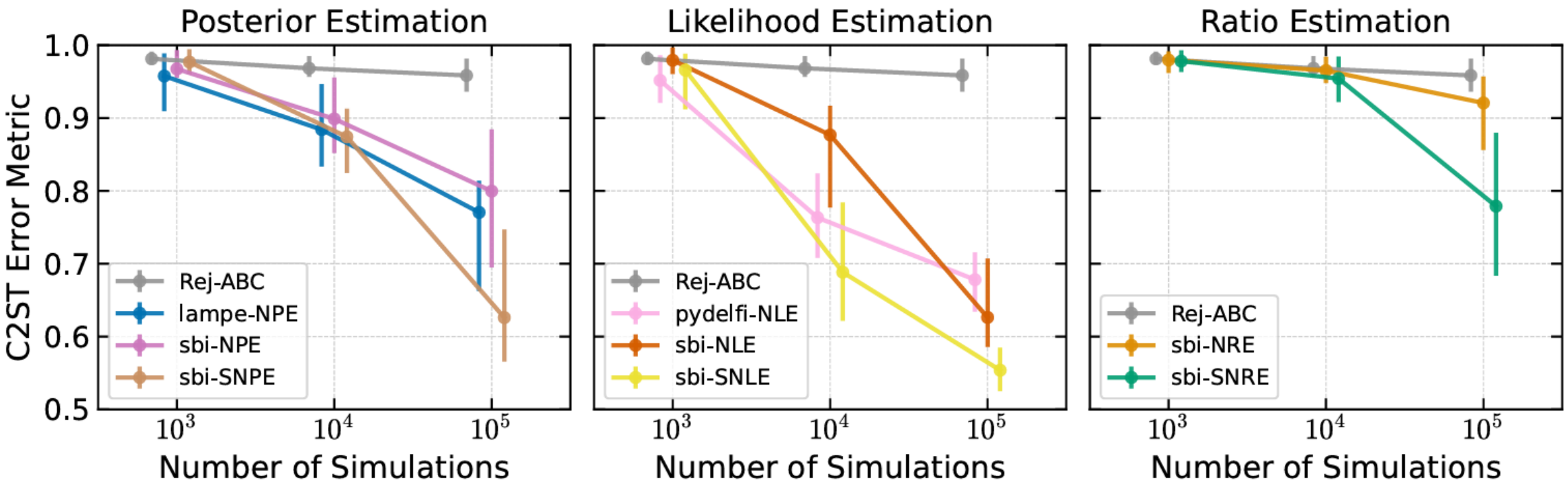


Cranmer, Brehmer, Louppe, 1911.01429

Machine-learning enhanced implicit inference

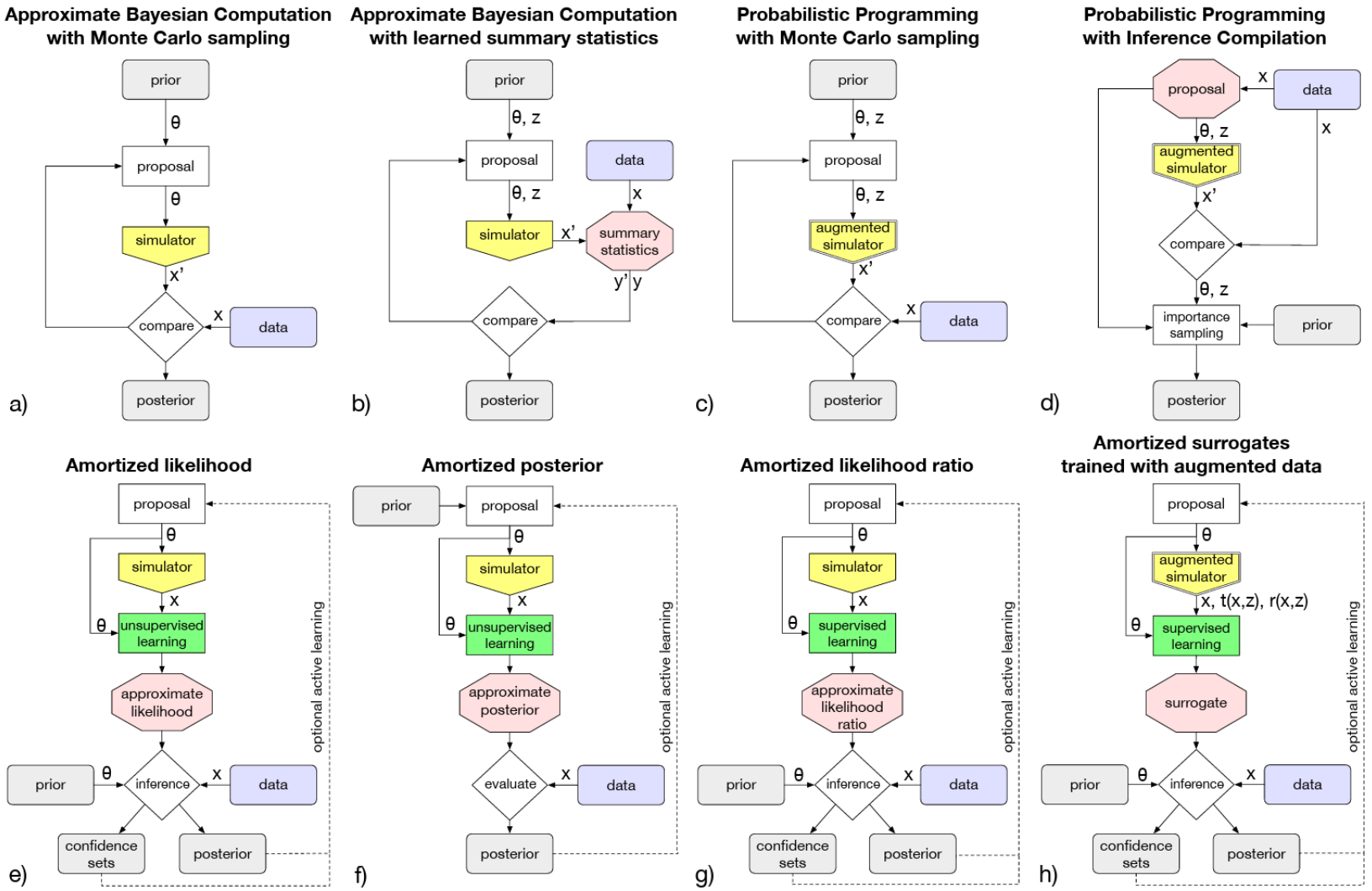
- For “cheap” simulators, machine learning takes us the rest of the way:
  - Recast inference problems as optimisation problems.
    - Parametrise the solution using a neural network.
    - Define its architecture.
  - Write down a loss function that defines the problem.
    - Generate training simulations.
    - Minimise the loss function.
    - Validate using another set of simulations.
- There are typically three classes of neural-enhanced implicit inference techniques:
  - Neural likelihood estimation (NLE, see also DELFI)
  - Neural posterior estimation (NPE)
  - Neural ratio estimation (NRE) — evaluates the ratio $\frac{p(\mathbf{d}|\boldsymbol{\theta})}{p(\mathbf{d})} = \frac{p(\boldsymbol{\theta}|\mathbf{d})}{p(\boldsymbol{\theta})} = \frac{p(\boldsymbol{\theta},\mathbf{d})}{p(\boldsymbol{\theta})p(\mathbf{d})}$

Accuracy comparison of different neural implicit inference techniques



Ho *et al.*, 2402.05137

Many flavours of implicit inference



Cranmer, Brehmer, Louppe, 1911.01429

DATA COMPRESSION

Score compression

- Originally introduced under the name MOPED (Multiple Optimised Parameter Estimation and Data compression algorithm, Heavens *et al*. 1999). Later generalised by Alsing & Wandelt (2018).

$$y_\alpha = \mathbf{b}_\alpha^T \mathbf{x}$$

$$\mathbf{b}_1 = \frac{\mathsf{C}^{-1}\boldsymbol{\mu}_{,1}}{\sqrt{\boldsymbol{\mu}_{,1}^T \mathsf{C}^{-1}\boldsymbol{\mu}_{,1}}}$$

$$\mathbf{b}_\alpha = \frac{\mathsf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{\mu}_{,\alpha}^T \mathbf{b}_\beta)\mathbf{b}_\beta}{\sqrt{\boldsymbol{\mu}_{,\alpha}^T \mathsf{C}^{-1}\boldsymbol{\mu}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{\mu}_{,\alpha}^T \mathbf{b}_\beta)^2}} \qquad 1 < \alpha \leq m,$$

$$\mathbf{t} = \nabla\boldsymbol{\mu}_*^{\mathrm{T}}\mathbf{C}_*^{-1}(\mathbf{d}-\boldsymbol{\mu}_*) + \frac{1}{2}(\mathbf{d}-\boldsymbol{\mu}_*)^{\mathrm{T}}\mathbf{C}_*^{-1}\nabla\mathbf{C}_*\,\mathbf{C}_*^{-1}(\mathbf{d}-\boldsymbol{\mu}_*),$$

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_* + \mathbf{F}_*^{-1}\Big[\nabla\boldsymbol{\mu}_*^{\mathrm{T}}\mathbf{C}_*^{-1}(\mathbf{d}-\boldsymbol{\mu}_*) + \frac{1}{2}(\mathbf{d}-\boldsymbol{\mu}_*)^{\mathrm{T}}\mathbf{C}_*^{-1}\nabla\mathbf{C}_*\mathbf{C}_*^{-1}(\mathbf{d}-\boldsymbol{\mu}_*) - \frac{1}{2}\mathrm{tr}(\mathbf{C}_*^{-1}\nabla\mathbf{C}_*)\Big],$$

- This is the optimal linear transformation to conserve the Fisher information of a Gaussian likelihood. This compression is lossless for a Gaussian likelihood (it can be lossy otherwise).

$$F_{\alpha\beta} = \frac{1}{2}\mathrm{Tr}\left[C^{-1}C_{,\alpha}C^{-1}C_{,\beta} + C^{-1}(\mu_{,\alpha}\mu_{,\beta}^{\mathsf{T}} + \mu_{,\beta}\mu_{,\alpha}^{\mathsf{T}})\right]$$

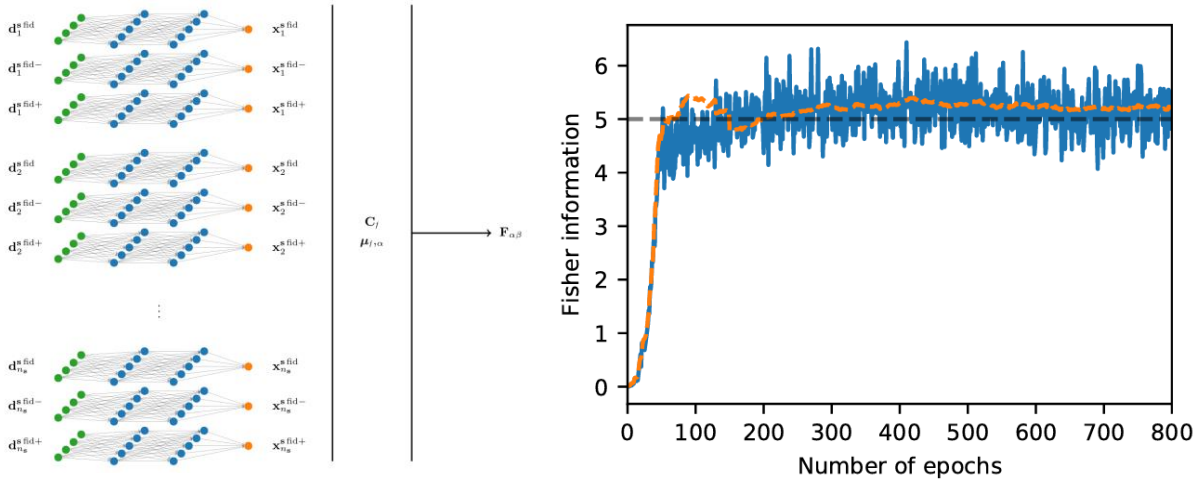
  - Note: in implicit inference, a lossy compression of the data will not introduce a bias, but only make the result suboptimal.
  - See also related work: Heavens *et al*., 1707.06529; Alsing & Wandelt, 1903.01473

Heavens *et al*., astro-ph/9911102; Alsing & Wandelt, 1712.00012

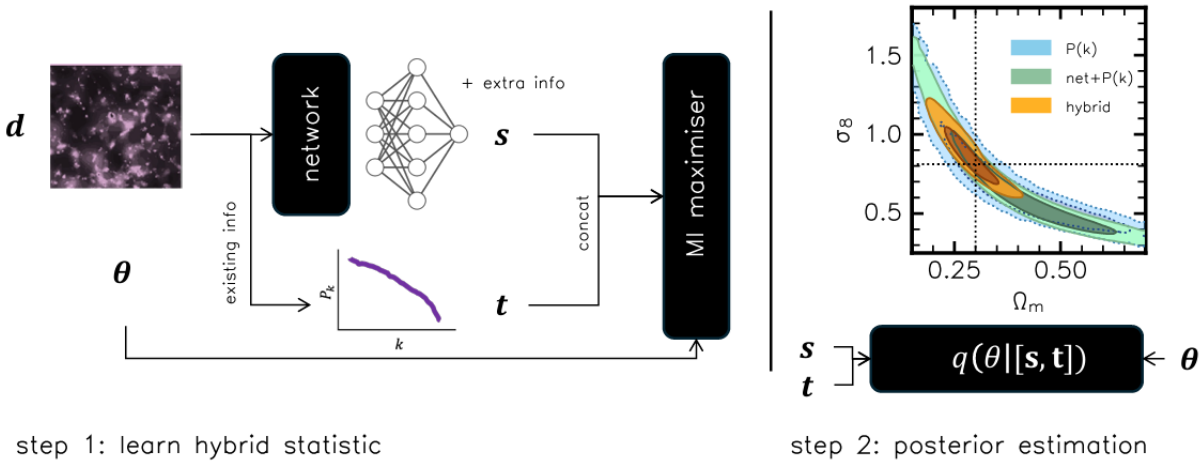
Neural data compression

- Beyond score compression, it is possible to train neural network to build informative summaries:
  - Information-maximising neural networks (IMNN) maximise the Fisher information of the summary
  - Hybrid summary statistics boost information extraction beyond existing summaries by maximising the mutual information (MI) between an existing summary of the data and the parameters of interest.

Information-maximising neural networks



Hybrid summary statistics

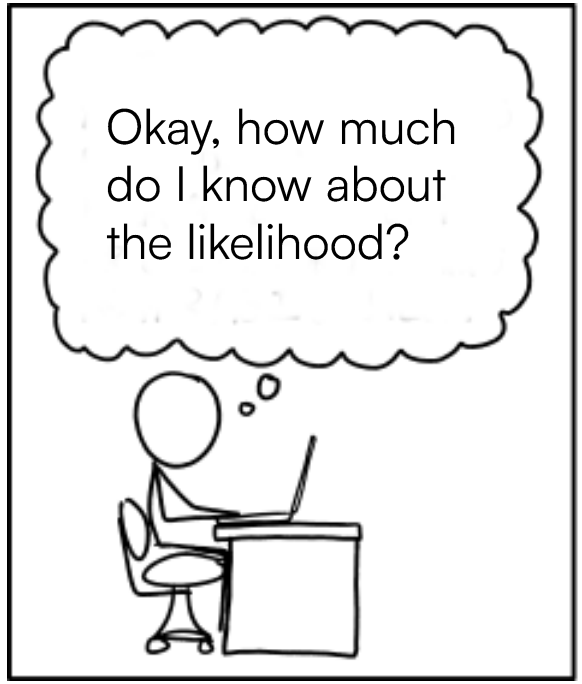


Charnock *et al.*, 1802.03537; Makinen *et al.* 2107.07405; Makinen *et al.* 2410.07548

CHOOSING THE METHOD

Probabilistic computations: two approaches

(a personal point of view)



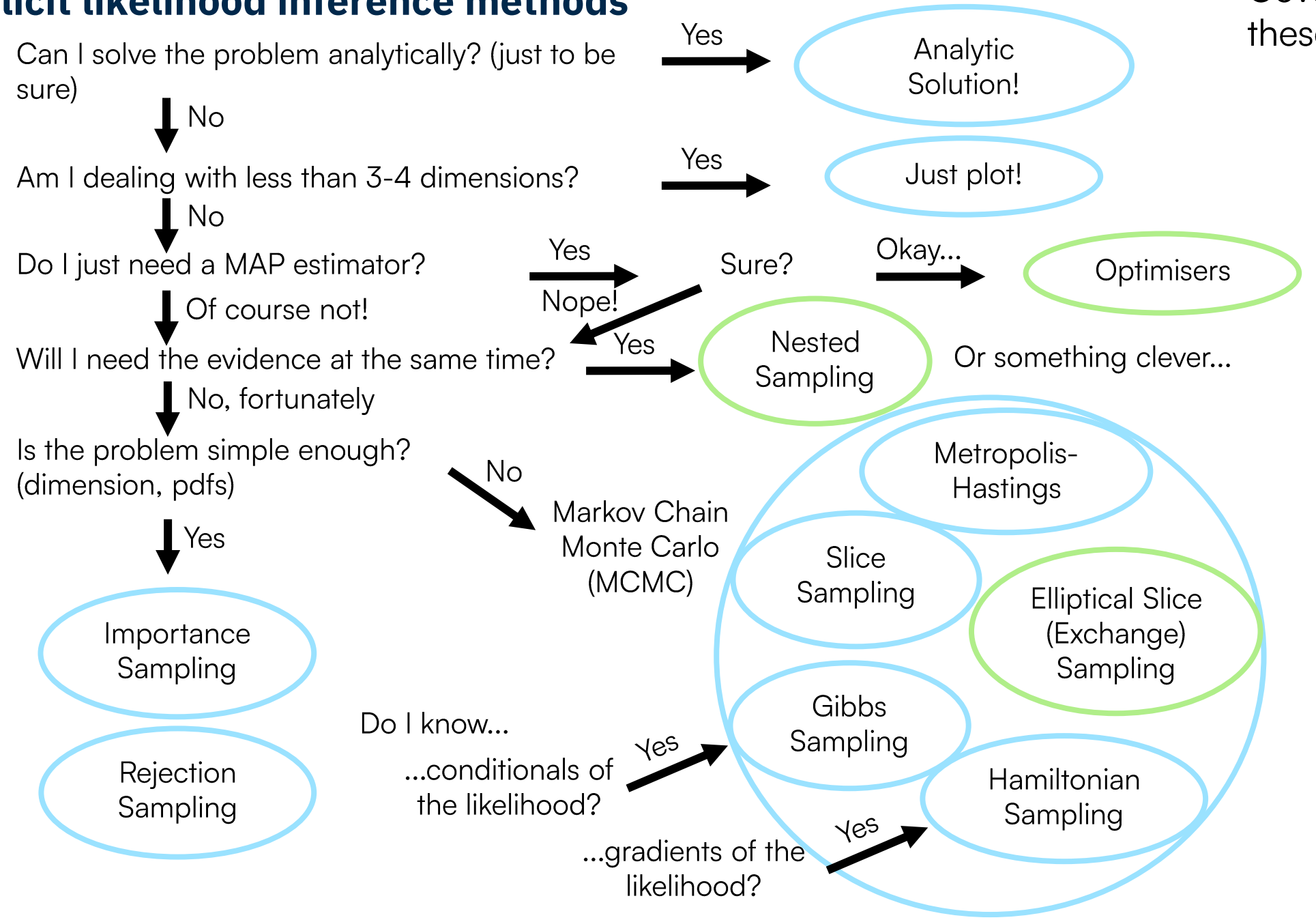
Quite a bit (score and sample)

Explicit likelihood inference (a.k.a. "likelihood-based" methods):
"Exact" Bayesian Inference

Not much (sample only, i.e. run simulations)

Implicit likelihood inference (a.k.a. "likelihood-free" or "simulation-based" methods):
Approximate Bayesian Computation (ABC), ...

Explicit likelihood inference methods



Implicit likelihood inference methods

Covered in these lectures

Is it really impossible to score from my likelihood?

No, I'm just lazy → ILI is not carte blanche replacement for explicit likelihood inference methods. I go back to the previous slide.

Yes ↓

Do I know a sufficient summary statistic of my data?

Yes → Great, I'm using this summary in the following!

No ↓

Then, I use score compression and/or neural data compression

Do I have a reasonable chance to hit the data "by chance" with my generative model? (strong prior, low dimension, high tolerance)

Yes →

- Likelihood-Free Rejection Sampling
- Likelihood-Free Population MC (PMC), Sequential MC (SMC), Particle Filters

No, but I know ML →

- Machine learning-enhanced ILI
  - Neural likelihood estimation (NLE)
  - Neural posterior estimation (NPE)
  - Neural ratio estimation (NRE)

No, it's hopeless ↓

Can I build a synthetic likelihood?

Yes →

- Synthetic likelihood methods
  - Likelihood-Free MCMC
    - Hamiltonian ABC
  - Bayesian Optimisation BOLFI

No ↓

Hopeless problem. I take a break, then think more.

... and many hybrids of all these

References and acknowledgements

References:

- A. Heavens (2009), 0906.0664, *Statistical techniques in cosmology*
- C. E. Rasmussen, C. K. I. Williams (2006), *Gaussian Processes for Machine Learning*
- K. Cranmer, J. Brehmer, G. Louppe (2019), 1911.01429, *The frontier of simulation-based inference*

- For their lectures, thanks to: Andrew Jaffe, Elena Sellentin, Roberto Trotta

https://florent-leclercq.eu/teaching.php