Bayesian statistics and Information Theory

Lecture 1: Aspects of probability theory

… a.k.a. *why am I not allowed to "change the prior" or "cut the data"?*

Florent Leclercq

www.florent-leclercq.eu

Imperial Centre for Inference and Cosmology

Imperial College London

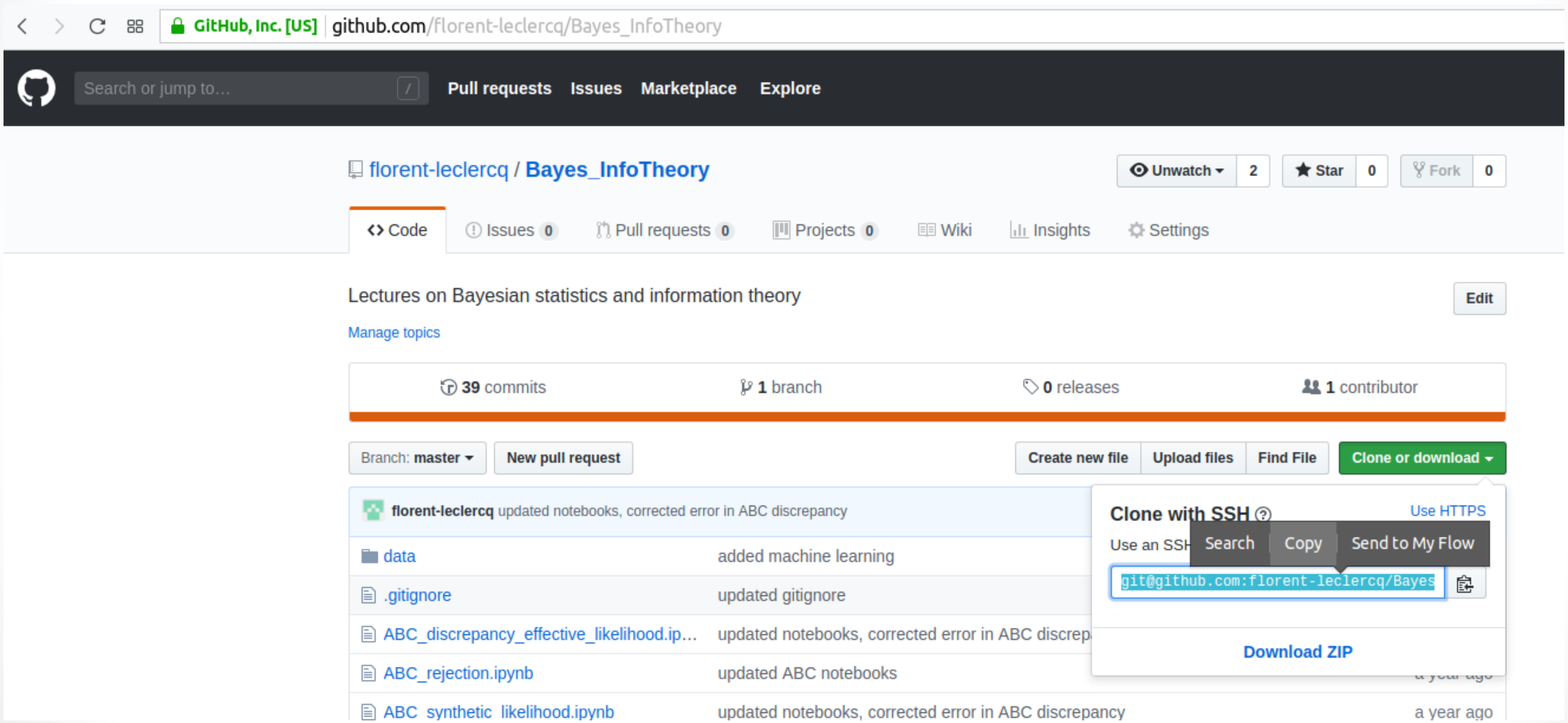
May 14th, 2019

ICIC Imperial Centre for Inference & Cosmology

Imperial College London

The github repository

- https://github.com/florent-leclercq/Bayes_InfoTheory

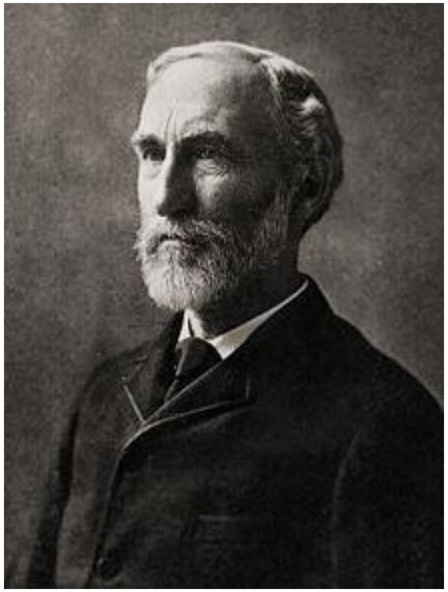


git clone https://github.com/florent-leclercq/Bayes_InfoTheory.git (or with SSH)

- Course website: http://florent-leclercq.eu/teaching.php

Introduction: why proper statistics matter

An historical example: the Gibbs paradox



J. Willard Gibbs (1839-1903)

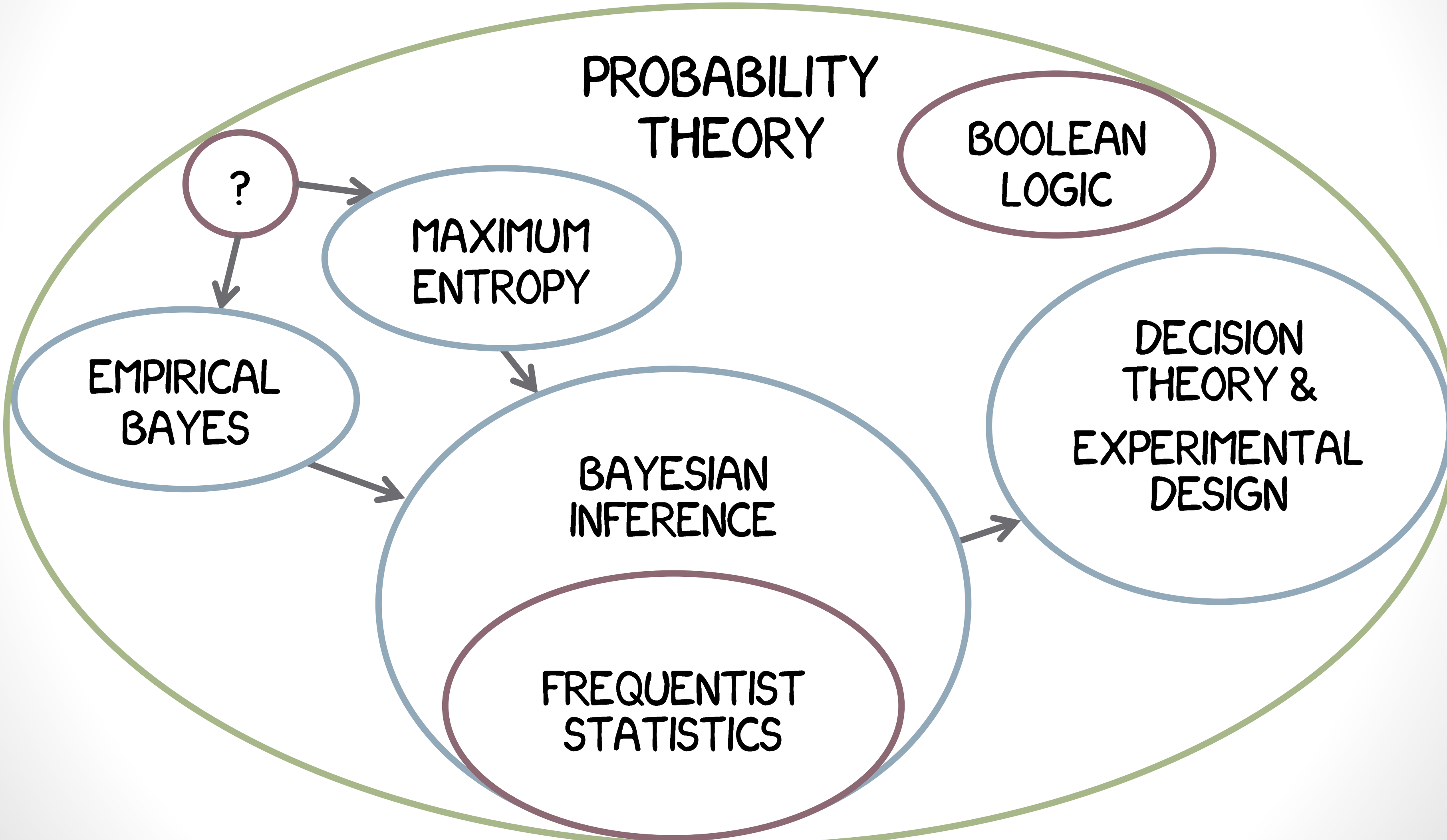
- Gibbs's canonical ensemble and grand canonical ensembles, derived from the maximum entropy principle, ***fail to correctly predict thermodynamic properties*** of real physical systems.
- The predicted entropies are always larger than the observed ones… there must exist ***additional microphysical constraints***:
  - Discreteness of energy levels: radiation: Planck (1900), solids: Einstein (1907), Debye (1912), Ising (1925), individual atoms : Bohr (1913)…
  - …Quantum mechanics: Heisenberg, Schrödinger (1927)

The first clues indicating the need for quantum physics were uncovered by seemingly "unsuccessful" application of statistics.

Outline: Lecture 1

- Probability theory and Bayesian statistics: reminders
- Ignorance priors and the maximum entropy principle
- Gaussian random fields (and a digression on non-Gaussianity)
- Bayesian signal processing and reconstruction:
  - Bayesian de-noising
  - Bayesian de-blending
- Bayesian decision theory and Bayesian experimental design
- Bayesian networks, Bayesian hierarchical models and Empirical Bayes
- (time permitting) Hypothesis testing beyond the Bayes factor:
  - Model selection as a decision analysis
  - Model averaging
  - Model selection with insufficient summary statistics

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



Reminders

- A tribute to my PhD supervisor (Benjamin Wandelt):

  Ben's summary of Bayesian statistics:

  "Whatever is uncertain gets a pdf."

- Product rule: $p(AB|C) = p(A|BC)\,p(B|C)$
- Sum rule: $p(A+B|C) = p(A|C) + p(B|C) - p(AB|C)$
- Bayes's formula: $$p(s|d) = \frac{p(d|s)\,p(s)}{p(d)}$$
  (posterior: $p(s|d)$; likelihood: $p(d|s)$; prior: $p(s)$; evidence: $p(d)$)
- Bayesian model comparison: $$\mathcal{B}_{12} = \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_2)}$$

Ignorance priors and the maximum entropy principle

Notebook 1: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/LighthouseProblem.ipynb
Notebook 2: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/MaximumEntropy.ipynb

Gaussian random fields

Notebook 3: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/GRF_and_fNL.ipynb

Bayesian signal processing and reconstruction

Notebook 4: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising.ipynb
Notebook 4bis: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_denoising_CMB.ipynb
Notebook 5: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/WienerFilter_deblending.ipynb

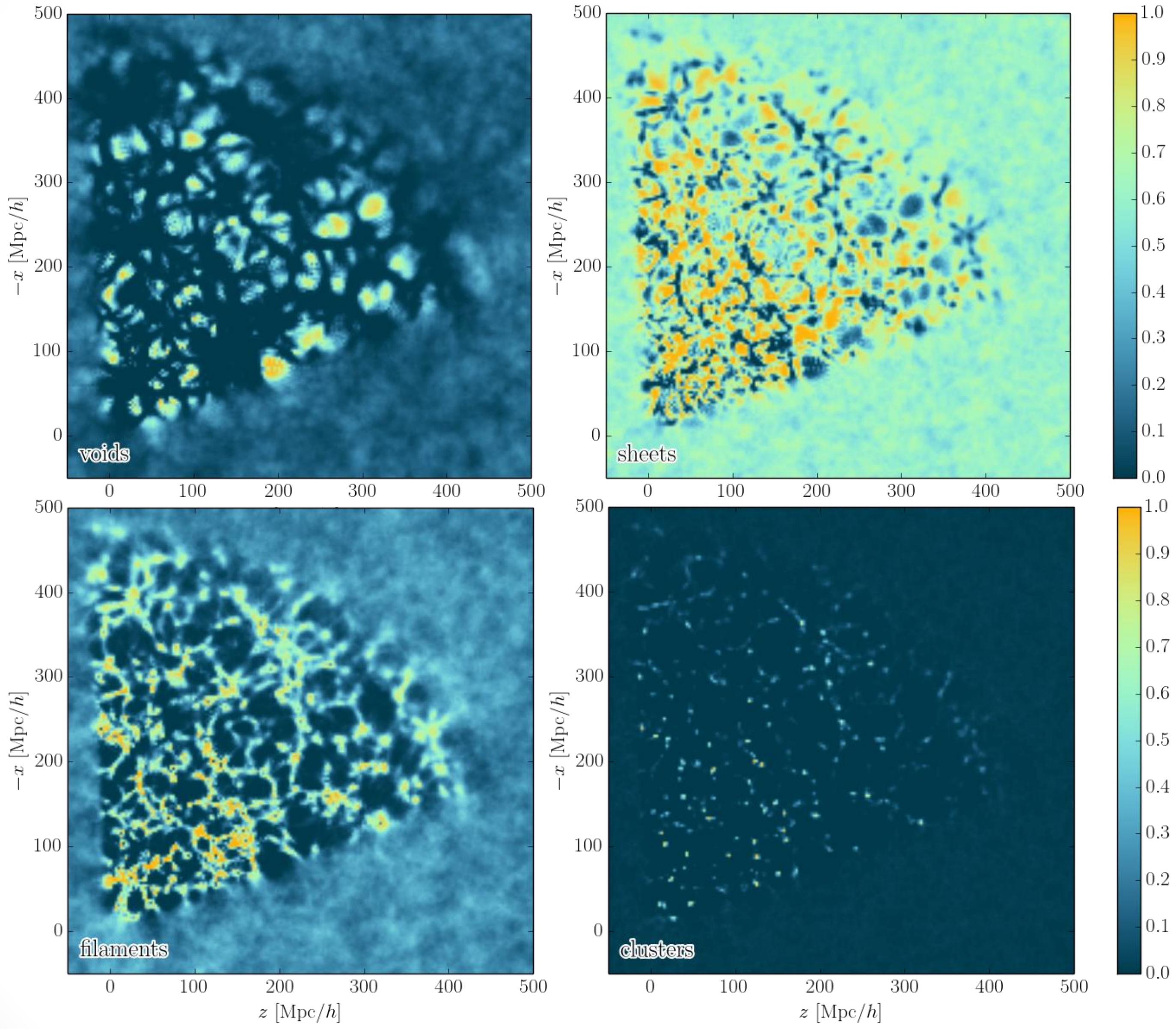
Bayesian decision theory

Notebook 6: https://github.com/florent-leclercq/Bayes_InfoTheory/blob/master/DecisionTheory.ipynb

Bayesian experimental design

(more about that in lecture 3)

Structures in the cosmic web



FL, Jasche & Wandelt 2015a, arXiv:1502.02690

A decision rule for structure classification

- Space of "input features":

$$\{\mathrm{T}_0 = \text{void}, \mathrm{T}_1 = \text{sheet}, \mathrm{T}_2 = \text{filament}, \mathrm{T}_3 = \text{cluster}\}$$

- Space of "actions":

$$\{a_0 = \text{“decide void”}, a_1 = \text{“decide sheet”}, a_2 = \text{“decide filament”}, a_3 = \text{“decide cluster”}, a_{-1} = \text{“do not decide”}\}$$

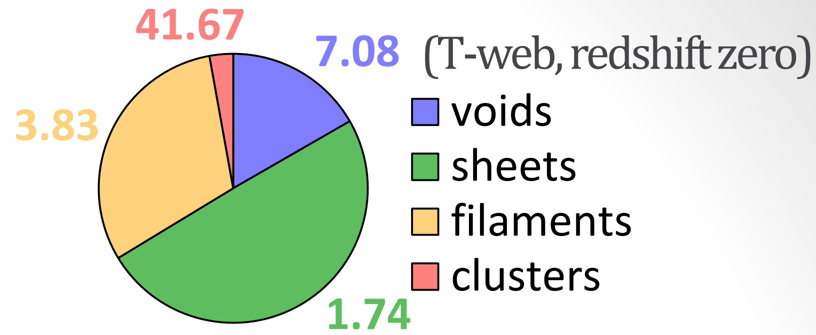
⇨ A problem of **Bayesian decision theory**:
one should take the action that maximizes the utility

$$U(a_j(\vec{x}_k)|d) = \sum_{i=0}^{3} G(a_j|\mathrm{T}_i)\, \mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)$$

- How to write down the gain functions?

FL, Jasche & Wandelt 2015b, arXiv:1503.00730

Gambling with the Universe



- One proposal:

$$G(a_j|\mathrm{T}_i) = \begin{cases} \frac{1}{\mathcal{P}(\mathrm{T}_i)} - \alpha & \text{if } j \in [\![0,3]\!] \text{ and } i = j \\ -\alpha & \text{if } j \in [\![0,3]\!] \text{ and } i \neq j \\ 0 & \text{if } j = -1. \end{cases}$$

"Winning"
"Losing"
"Not playing"

- Without data, the expected utility is

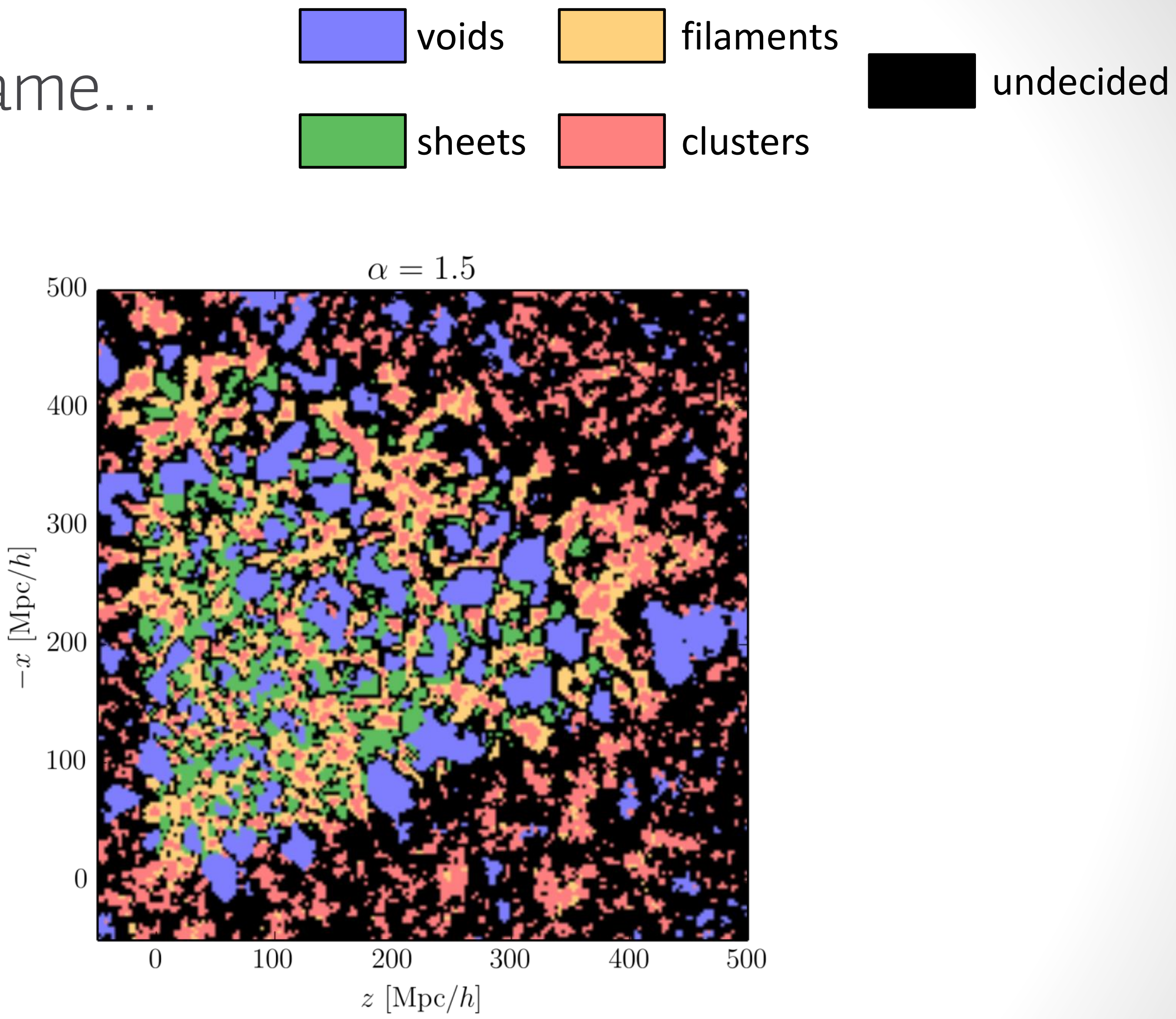
$$U(a_j) = 1 - \alpha \quad \text{if } j \neq 1$$ "Playing the game"

$$U(a_{-1}) = 0$$ "Not playing the game"

- With $\alpha = 1$, it's a *fair game* ⇨ always play
⇨ "**speculative map**" of the LSS
- Values $\alpha > 1$ represent an *aversion for risk*
⇨ increasingly "**conservative maps**" of the LSS

FL, Jasche & Wandelt 2015b, arXiv:1503.00730

Playing the game…



FL, Jasche & Wandelt 2015b, arXiv:1503.00730

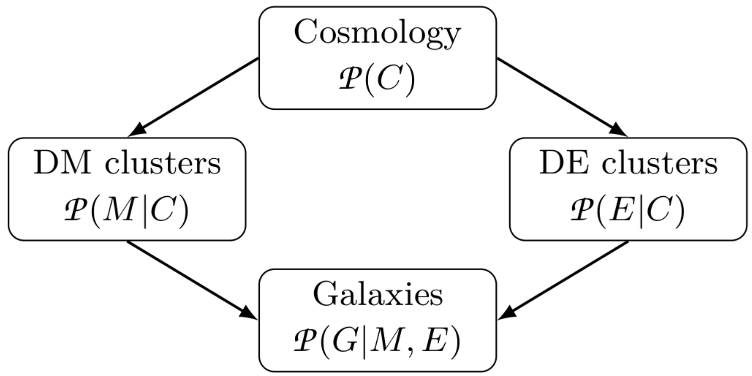
Bayesian networks

Bayesian hierarchical models

and Empirical Bayes

Bayesian networks

$\mathcal{P}(C)$	$\mathcal{P}(\bar{C})$
0.5	0.5



c	$\mathcal{P}(M|C=c)$	$\mathcal{P}(\bar{M}|C=c)$
0	0.2	0.8
1	0.7	0.3

c	$\mathcal{P}(E|C=c)$	$\mathcal{P}(\bar{E}|C=c)$
0	0.6	0.4
1	0.1	0.9

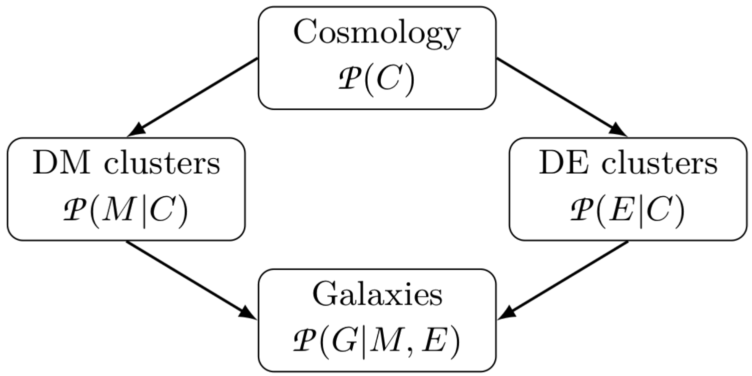
m	e	$\mathcal{P}(G|M=m,E=e)$	$\mathcal{P}(\bar{G}|M=m,E=e)$
0	0	0.10	0.90
1	0	0.95	0.05
0	1	0.95	0.05
1	1	0.99	0.01

Bayesian networks are probabilistic graphical models consisting of:

- A **directed acyclic graph**
- At each node, **conditional probabilities distributions**

Bayesian networks

$\mathcal{P}(C)$	$\mathcal{P}(\bar{C})$
0.5	0.5



c	$\mathcal{P}(M\|C=c)$	$\mathcal{P}(\bar{M}\|C=c)$
0	0.2	0.8
1	0.7	0.3

c	$\mathcal{P}(E\|C=c)$	$\mathcal{P}(\bar{E}\|C=c)$
0	0.6	0.4
1	0.1	0.9

m	e	$\mathcal{P}(G\|M=m,E=e)$	$\mathcal{P}(\bar{G}\|M=m,E=e)$
0	0	0.10	0.90
1	0	0.95	0.05
0	1	0.95	0.05
1	1	0.99	0.01

$$p(C,M,E,G) = p(C)\,p(E|C)\,p(M|C,\cancel{E})\,p(G|\cancel{C},M,E)$$

$$p(C,M,E,G) = p(C)\,p(E|C)\,p(M|C)\,p(G|M,E)$$

Bayesian networks
inference and prediction

- Inference:

$$p(M|G) = \frac{p(M,G)}{p(G)} = \frac{\sum_{c,e} p(C{=}c,M{=}1,E{=}e,G{=}1)}{\sum_{c,m,e} p(C{=}c,M{=}m,E{=}e,G{=}1)} = \frac{0.4313}{0.70305} \approx 0.6135$$

$$p(E|G) = \frac{p(E,G)}{p(G)} = \frac{\sum_{c,m} p(C{=}c,M{=}m,E{=}1,G{=}1)}{\sum_{c,m,e} p(C{=}c,M{=}m,E{=}e,G{=}1)} = \frac{0.3363}{0.70305} \approx 0.4783$$

$$p(\bar{M},\bar{E}|G) = \frac{p(\bar{M},\bar{E},G)}{p(G)} = \frac{\sum_{c} p(C{=}c,M{=}0,E{=}0,G{=}1)}{\sum_{c,m,e} p(C{=}c,M{=}m,E{=}e,G{=}1)} = \frac{0.0295}{0.70305} \approx 0.0420$$

- Prediction:

$$p(G|C) = \frac{p(G,C)}{p(C)} = \frac{\sum_{m,e} p(C{=}1,M{=}m,E{=}e,G{=}1)}{p(C{=}1)} = 0.7233$$

Bayesian networks

the "explaining away" phenomenon

$$p(E|M,G) = \frac{p(E,M,G)}{p(M,G)} = \frac{\sum_c p(C{=}c,M{=}1,E{=}1,G{=}1)}{\sum_{c,e} p(C{=}c,M{=}1,E{=}e,G{=}1)} = \frac{0.09405}{0.4313} \approx 0.2181$$

$$p(E|G) = \frac{p(E,G)}{p(G)} = \frac{\sum_{c,m} p(C{=}c,M{=}m,E{=}1,G{=}1)}{\sum_{c,m,e} p(C{=}c,M{=}m,E{=}e,G{=}1)} = \frac{0.3363}{0.70305} \approx 0.4783$$

- So we have both:

$$p(E|M) = p(E)$$
$$p(E|M,G) < p(E|,G)$$

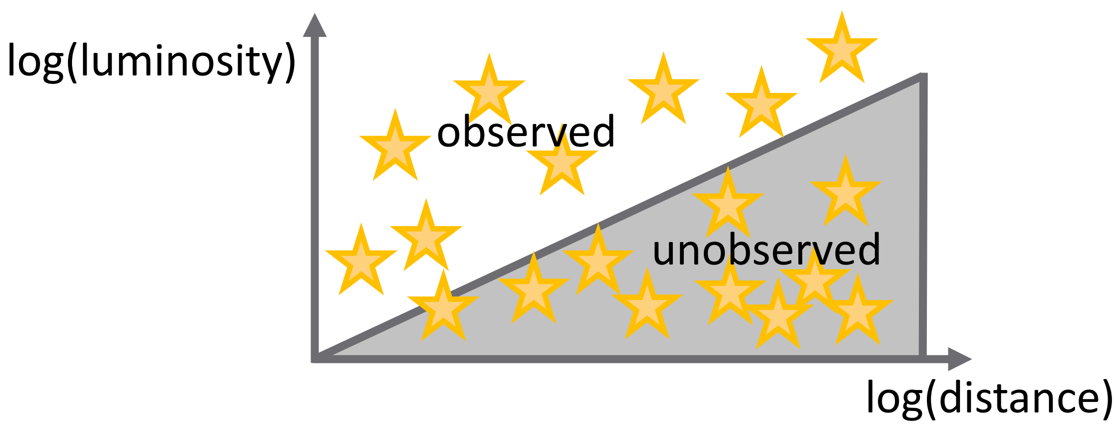
- This is "**collider bias**" or the "**explaining away**" phenomenon: two causes collide to explain the same effect.
- Particular case: "**selection bias**" or "**Berkson's paradox**"

$$0 < p(A) < 1; \quad 0 < p(B) < 1; \quad p(A|B) = p(A)$$

⇒ $$\begin{array}{l} p(A|B,C) < p(A|C) \\ p(A|\bar{B},C) = 1 > p(A|C) \end{array} \qquad C = A + B$$

Malmquist bias

- Malmquist (1925) bias: in magnitude-limited surveys, far objects are preferentially detected if they are intrinsically bright.



$$0 < p(A) < 1; \quad 0 < p(B) < 1; \quad p(A|B) = p(A)$$

$$C = A + B \qquad \Rightarrow \qquad p(A|\bar{B}, C) = 1 > p(A|C)$$

detected (C) bright (A) close (B)

Bayesian hierarchical models

- Simple inference:

prior

$$p(\theta|d) \propto p(d|\theta)\, p(\theta)$$

- Adaptive prior:

prior hyperprior

$$p(\theta|d) \propto p(d|\theta)\, p(\theta|\eta)\, p(\eta)$$

- … or a full hierarchy of hyperpriors.

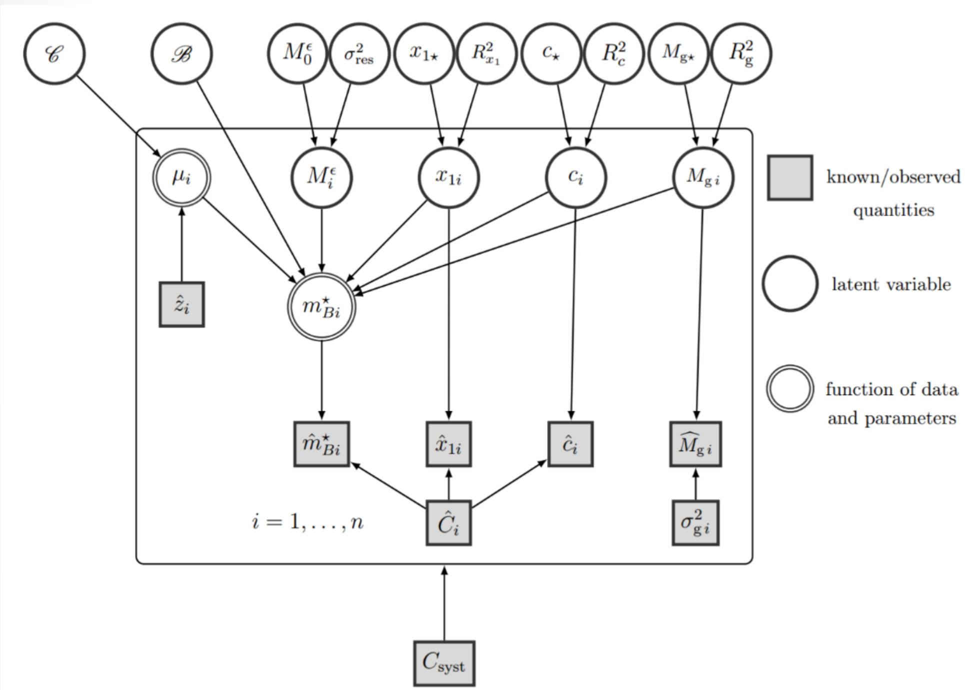
- Examples:
  - Cosmic microwave background:

$$p(\{\Omega\}, \{C_\ell\}, s|d) \propto p(d|s)\, p(s|\{C_\ell\})\, p(\{C_\ell\}|\{\Omega\})\, p(\{\Omega\})$$

  - Large-scale structure:

$$p(\{\Omega\}, \phi, g|d) \propto p(d|g)\, p(g|\phi)\, p(\phi|\{\Omega\})\, p(\{\Omega\})$$

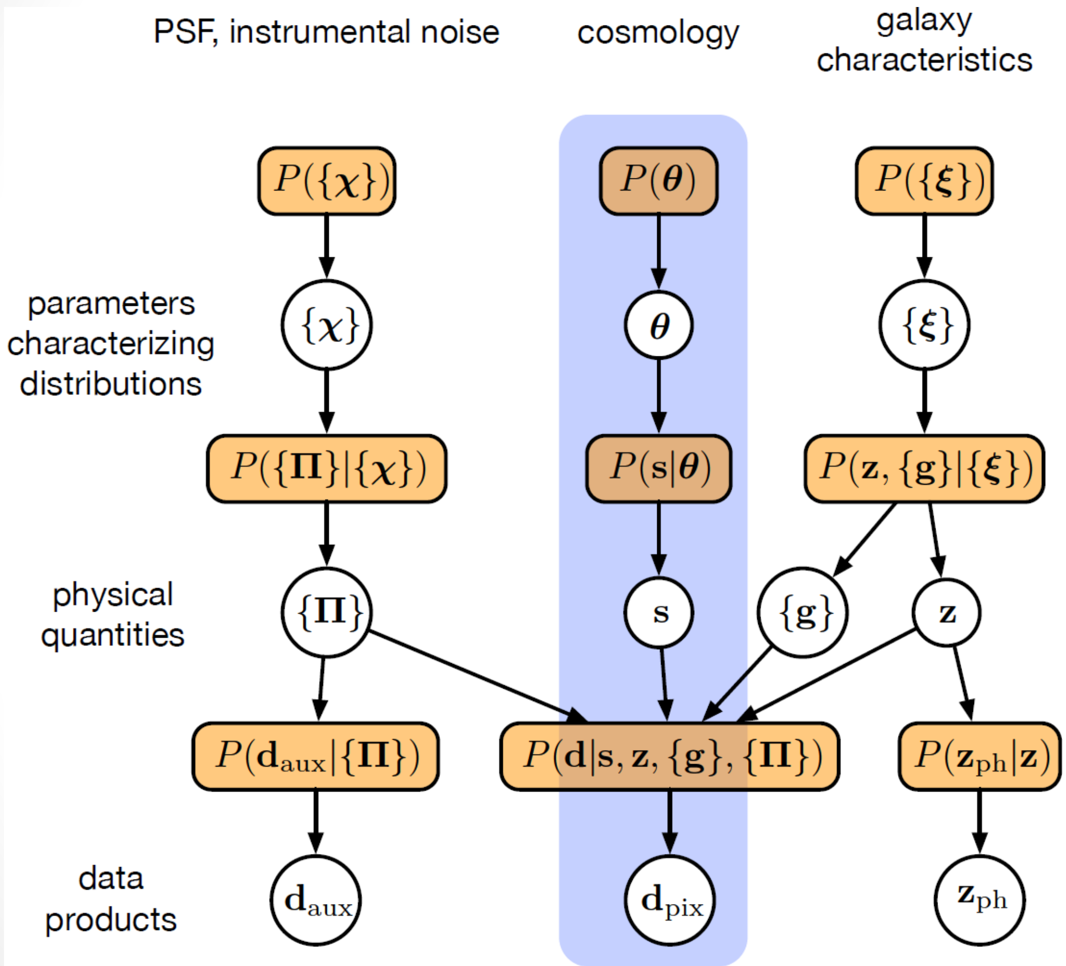
BHM example: supernovae (BAHAMAS)



Parameter	Notation and Prior Distribution
Cosmological parameters	
Matter density parameter	$\Omega_m \sim \textsc{Uniform}(0, 2)$
Cosmological constant density parameter	$\Omega_\Lambda \sim \textsc{Uniform}(0, 2)$
Dark energy EOS	$w \sim \textsc{Uniform}(-2, 0)$
Hubble parameter	$H_0/\text{km/s/Mpc} = 67.3$
Covariates	
Coefficient of stretch covariate	$\alpha \sim \textsc{Uniform}(0, 1)$
Coefficient of color covariate	β (or β_0) $\sim \textsc{Uniform}(0, 4)$
Coefficient of interaction of color correction and z	$\beta_1 \sim \textsc{Uniform}(-4, 4)$
Jump in coefficient of color covariate	$\Delta\beta \sim \textsc{Uniform}(-1.5, 1.5)$
Redshift of jump in color covariate	$z_t \sim \textsc{Uniform}(0.2, 1)$
Coefficient of host galaxy mass covariate	$\gamma \sim \textsc{Uniform}(-4, 4)$
Population-level distributions	
Mean of absolute magnitude	$M_0^\epsilon \sim \mathcal{N}(-19.3, 2^2)$
Residual scatter after corrections	$\sigma^2_{\text{res}} \sim \textsc{InvGamma}(0.003, 0.003)$
Mean of absolute magnitude, low galaxy mass	$M_0^{\text{lo}} \sim \mathcal{N}(-19.3, 2^2)$
SD of absolute magnitude, low galaxy mass	${\sigma^{\text{lo}}_{\text{res}}}^2 \sim \textsc{InvGamma}(0.003, 0.003)$
Mean of absolute magnitude, high galaxy mass	$M_0^{\text{hi}} \sim \mathcal{N}(-19.3, 2^2)$
SD of absolute magnitude, high galaxy mass	${\sigma^{\text{hi}}_{\text{res}}}^2 \sim \textsc{InvGamma}(0.003, 0.003)$
Mean of stretch	$x_{1\star} \sim \mathcal{N}(0, 10^2)$
SD of stretch	$R_{x_1} \sim \textsc{LogUniform}(-5, 2)$
Mean of color	$c_\star \sim \mathcal{N}(0, 1^2)$
SD of color	$R_c \sim \textsc{LogUniform}(-5, 2)$
Mean of host galaxy mass	$M_{g\star} \sim \mathcal{N}(10, 100^2)$
SD of host galaxy mass	$R_g \sim \textsc{LogUniform}(-5, 2)$

Shariff *et al.* 2015, arXiv:1510.05954

BHM example: weak lensing



Can include:

Mask
Intrinsic alignments
Baryon feedback
Shape measurement
Photometric redshifts

Alsing *et al.* 2015, arXiv:1505.07840

Empirical Bayes

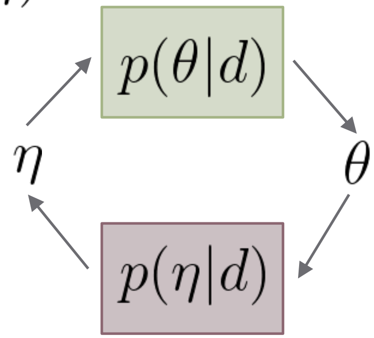
an alternative to maximum entropy for choosing priors

prior hyperprior

$$p(\theta|d) \propto p(d|\theta)\, p(\theta|\eta)\, p(\eta)$$

$$\underline{p(\theta|d)} = \int p(\theta|\eta, d)\, p(\eta|d)\, \mathrm{d}\eta = \int \frac{p(d|\theta)\, p(\theta|\eta)}{p(d|\eta)}\, \underline{p(\eta|d)}\, \mathrm{d}\eta$$

$$\underline{p(\eta|d)} = \int p(\eta|\theta)\, \underline{p(\theta|d)}\, \mathrm{d}\theta$$



⇨ Iterative scheme (“Gibbs” sampler)

- **Empirical Bayes** is a truncation of this scheme after a few steps (often just one).
- Particular case: $p(\eta|d) \approx \delta_\mathrm{D}(\eta - \eta^\star(d))$ ⇨ $\underline{p(\theta|d)} \approx \dfrac{p(d|\theta)\, p(\theta|\eta^\star)}{p(d|\eta^\star)}$

⇨ the **Expectation-Maximization** (EM) algorithm (machine learning, data mining).

Hypothesis testing beyond the Bayes factor