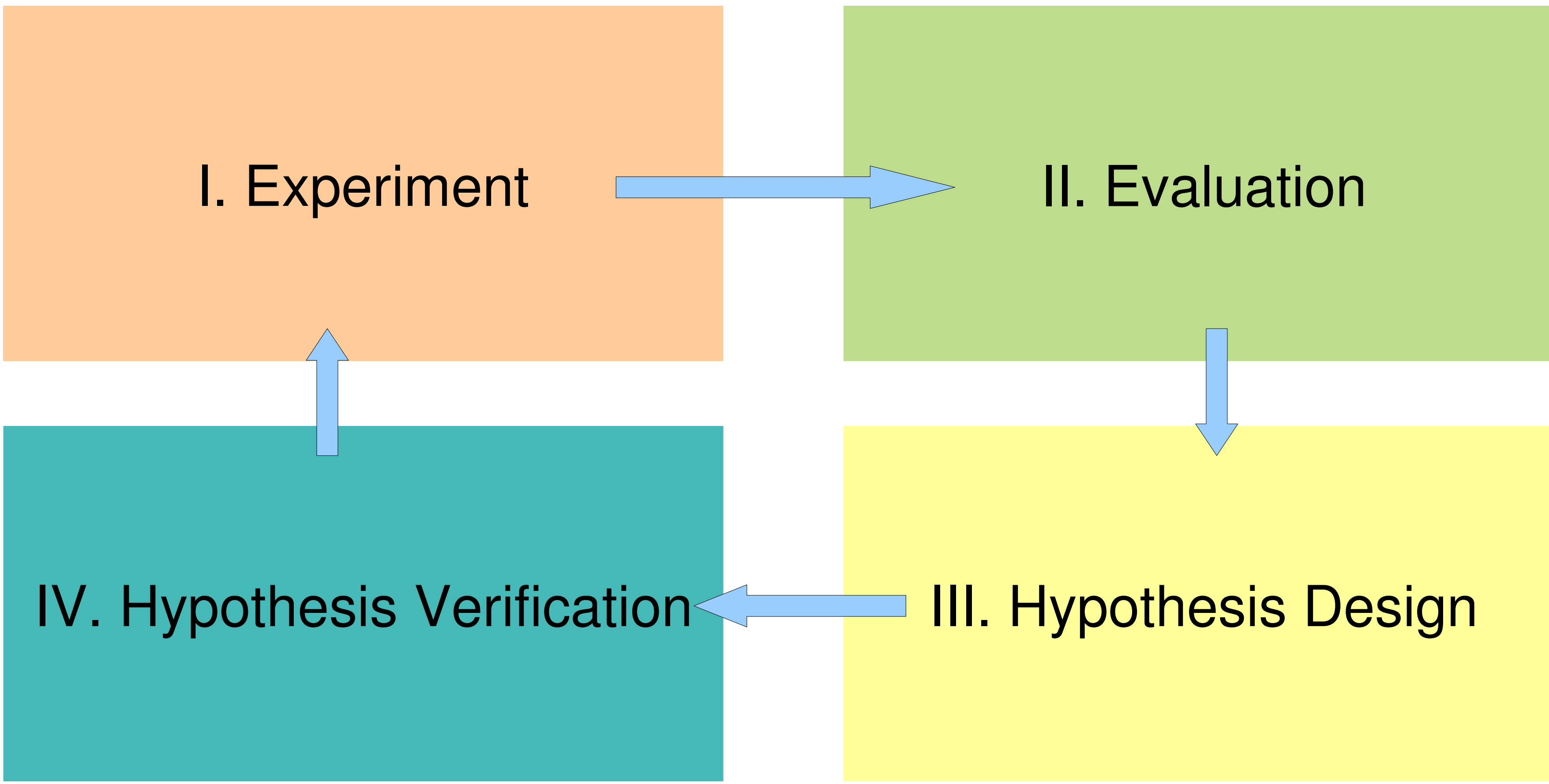
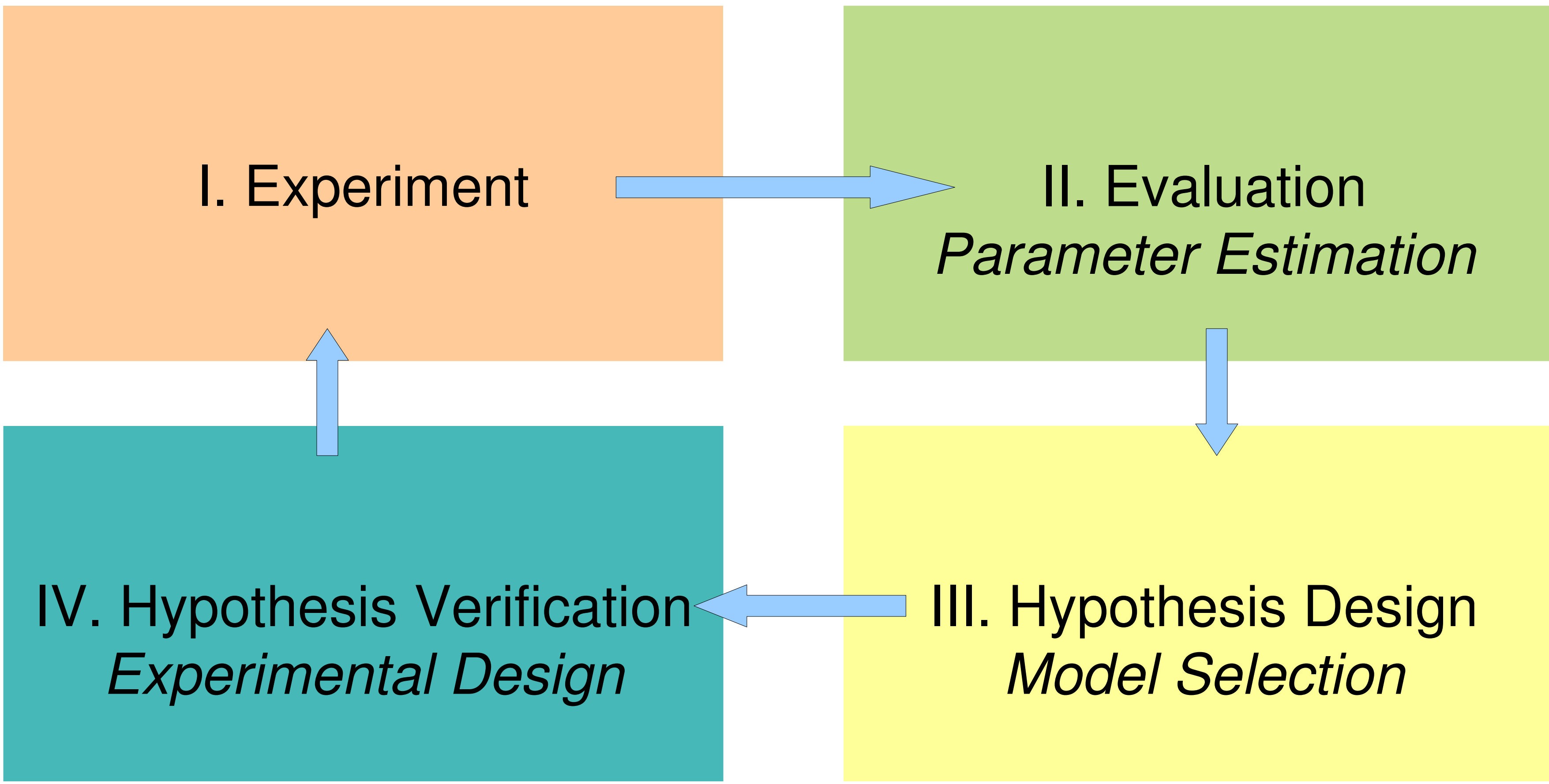
Bayesian Inference
in
Physics

U. von Toussaint

Scientific Inference Cycle



Scientific Inference Cycle



Prerequisite: Consistent reasoning...

Outline

I. Scientific Inference	Inference in Science Processing of Information
II. Model Comparison	Basic Concept Mass Spectroscopy
III. Experimental Design	Basic Concept Nuclear Reaction Analysis
IV. Numerical Interlude	Nested Sampling
V. Conclusion	Summary Outlook

Inference in Science

Science: prior information + new data → new knowledge

continous learning process

<u>Prior information:</u>

- old data
- calibration measurements, validation data
- theoretical considerations
- parameters
- model

<u>New data:</u>

- measurements
- calibration
- theories

But: prior information and data are uncertain

→ new knowledge and derived hypotheses are uncertain

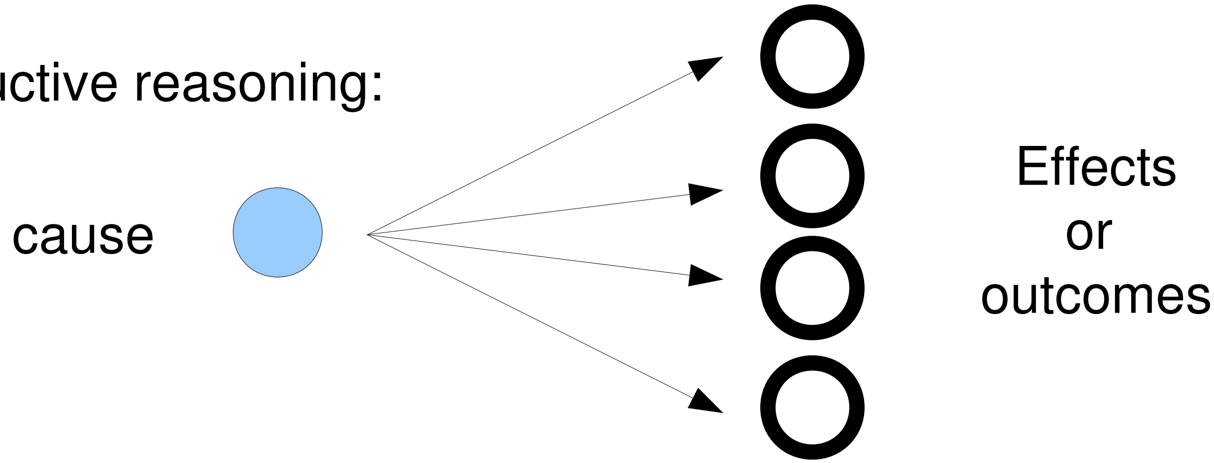
→ inductive (instead of deductive) reasoning required

→ **Calculus** for “uncertainties” needed (quantification)

Inference in Science

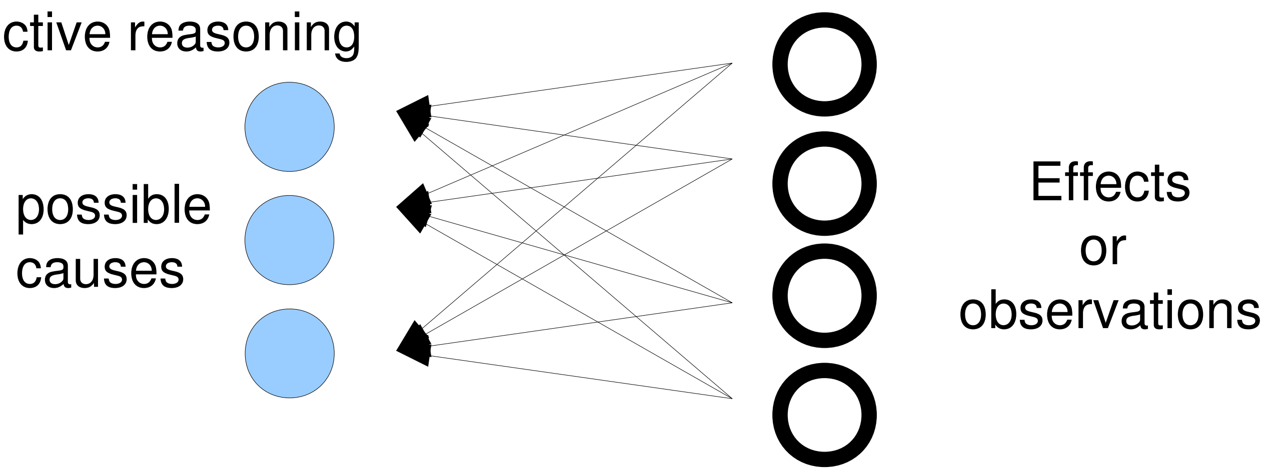
Science: prior information + new data → new knowledge

a) deductive reasoning:



Example: fair coin: p(7 heads | 10 tosses)=?

b) inductive reasoning



Example: 7 heads out of 10 tosses: p(fair coin)=?

Inference in Science

Calculus:

Cox (1946): Basic requirements (i.e. transitivity,consistency) single out *usual rules of probability theory* to handle uncertainty

Please note: probability here *not restricted* to frequency interpretation:

Degree of belief about a proposition

Data:
- statistical (counting statistics)
- measurement uncertainty (ruler)
- systematic (e.g. misalignment)
- outliers

Hypotheses:
- parameter of interest
- nuisance parameters
- physical models
- future data

Notation: p(x|I)

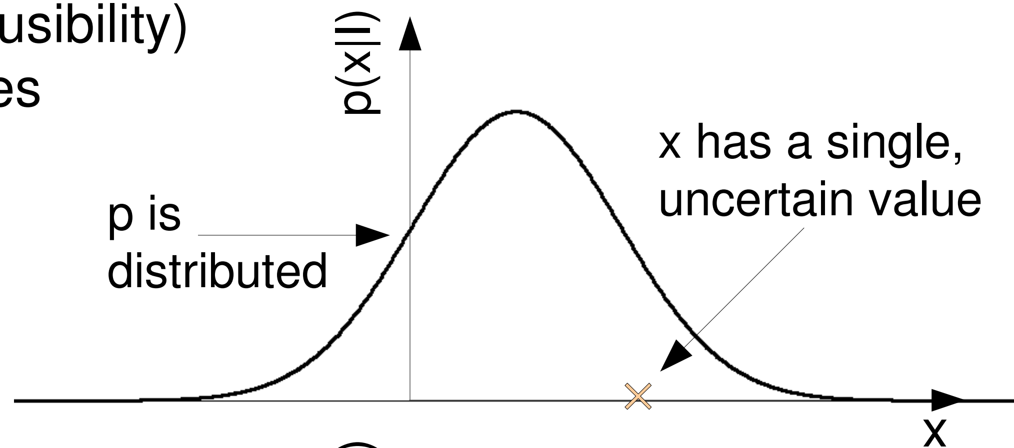
p(x|I) describes how probability (plausibility) is distributed among possible choices for x for the case at hand (information I)

Inference in Science

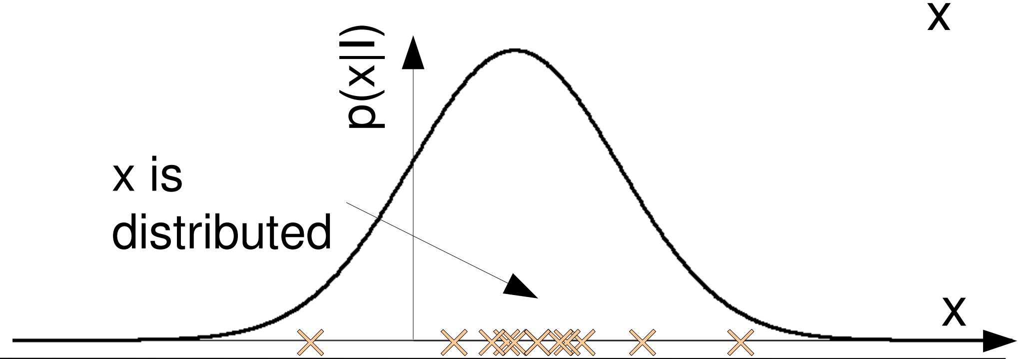
Calculus:

Cox (1946): Basic requirements (i.e. transitivity,consistency) single out *usual rules of probability theory* to handle uncertainty

Bayesian interpretation:
p(x|I) describes how probability (plausibility) is distributed among possible choices for x for the case at hand (information I)



Frequentist interpretation:
p(x|I) describes how x is distributed throughout an infinite (hypothetical) ensemble
probability = frequency



Processing of information

Processing of information: Combination of (cond.) probability distributions

Conditional Probabilities:

p(A| B) : probability of proposition A given truth of proposition B: quantification of uncertainty of A

Probability Theory Axioms:

- sum rule (OR): $p(H_1 + H_2 \mid I) = p(H_1 \mid I) + p(H_2 \mid I) - p(H_1, H_2 \mid I)$

- product rule (AND): $p(H_1, D \mid I) = p(D \mid H_1, I)\, p(H_1 \mid I)$

$$= p(H_1 \mid D, I)\, p(D \mid I)$$

$$\Rightarrow \quad p(H_1 \mid D, I) = \frac{p(D \mid H_1, I)\, p(H_1 \mid I)}{p(D \mid I)}$$

Bayes Theorem

Processing of information

Bayes' theorem:

Likelihood Prior

$$p\,(H_1 \mid D, I) = \frac{p\,(D \mid H_1, I)\, p\,(H_1 \mid I)}{p\,(D \mid I)}$$

Posterior Evidence

Prior: knowledge before experiment (logically)
Likelihood: Probability for data if the hypothesis was true
Posterior: Probability that the hypothesis is true given the data
Evidence: normalization; important for model comparison

Maximum Likelihood approach (parameters which maximise probability for data) ***does not*** yield most likely parameters* !

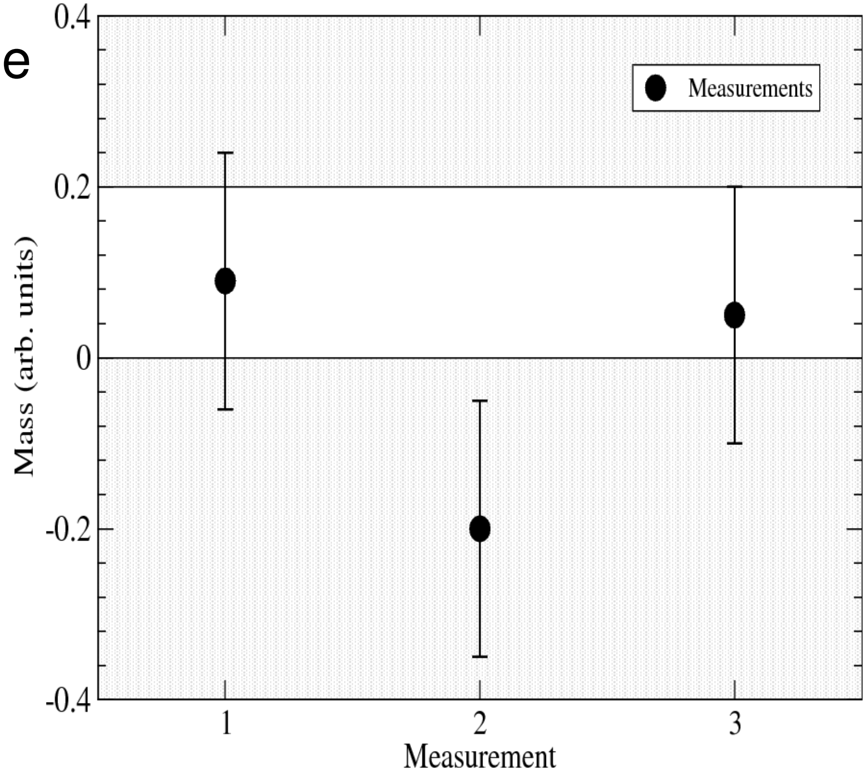
Examples: p(wet street | rain,I) ≠ p(rain | wet street,I)
p(female | pregnant,I)≠ p(pregnant | female,I)

*) in general, except e.g. flat, unbounded priors

Processing of information

Toy Example: Mass of elementary particle

- Prior knowledge: $0 \le m \le m_{upper}=0.2$
- measurement uncertainty $\sigma=0.15$
- measured data:
  $d_1=0.09$,
  $d_2=-0.2$,
  $d_3=0.05$



Maximum likelihood: m=-0.02 (<0!)

Bayes: 1) Assign prior:
$$p\,(m \mid I) = \begin{cases} 1/m_{\text{upper}}, & \text{if } 0 \le m \le m_{\text{upper}}; \\ 0, & \text{otherwise,} \end{cases}$$

2) Assign likelihood:
$$p\,(d \mid m, \sigma, I) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(d-m)^2}{\sigma^2}\right)$$

3) Compute posterior using Bayes' theorem:

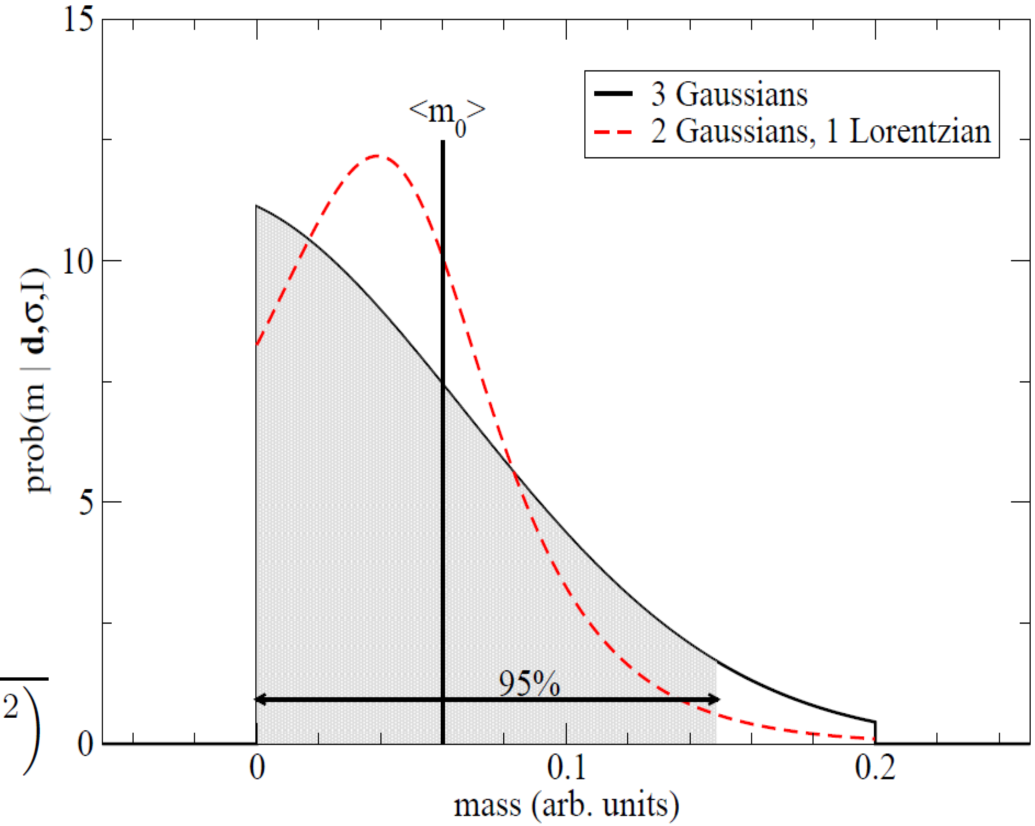
Processing of information

posterior : $p\left(m \mid \boldsymbol{d}, \sigma, I\right) = \dfrac{p\left(m \mid I\right) \prod_{i=1}^{N} p\left(d_i \mid m, \sigma, I\right)}{Z}$

contains complete information:

summarizing quantities:

- mode: m=0
- mean: $<m_0>$=0.06
- 95%-interval: [0;0.145]



Non-gaussian likelihoods:

E.g. $p\left(d \mid m, \beta, I\right) = \dfrac{\beta}{\pi\left(\beta^2 + (d-m)^2\right)}$

for d_3 with β=0.05

Processing of information

Reasoning about parameter *a*:

(uncertain) prior information	$p(a\|I)$	prior distribution
+ (uncertain) measured data	$d = D \pm \varepsilon$	$p(d\|a)$ likelihood distribution
+ physical model	$D = f(a)$	

+*Bayes* theorem

$$\boxed{p(a\,|\,d) = \frac{p(d\,|\,a) \times p(a)}{p(d)}}$$

posterior distribution

+an additional (nuisance) parameter *b*:

$$p(a\,|\,d) = \int db\; p(a,b\,|\,d) = \int db\, \frac{p(d\,|\,a,b)\; p(a,b)}{p(d)}$$

Marginalization: generalized error propagation

Clear recipe how to tackle a problem – possibly demanding mathematics/numerics

1) Quantify information at hand in probability distributions
2) Multiply probability distributions
3) Marginalize nuisance parameters
4) Analyze posterior distribution

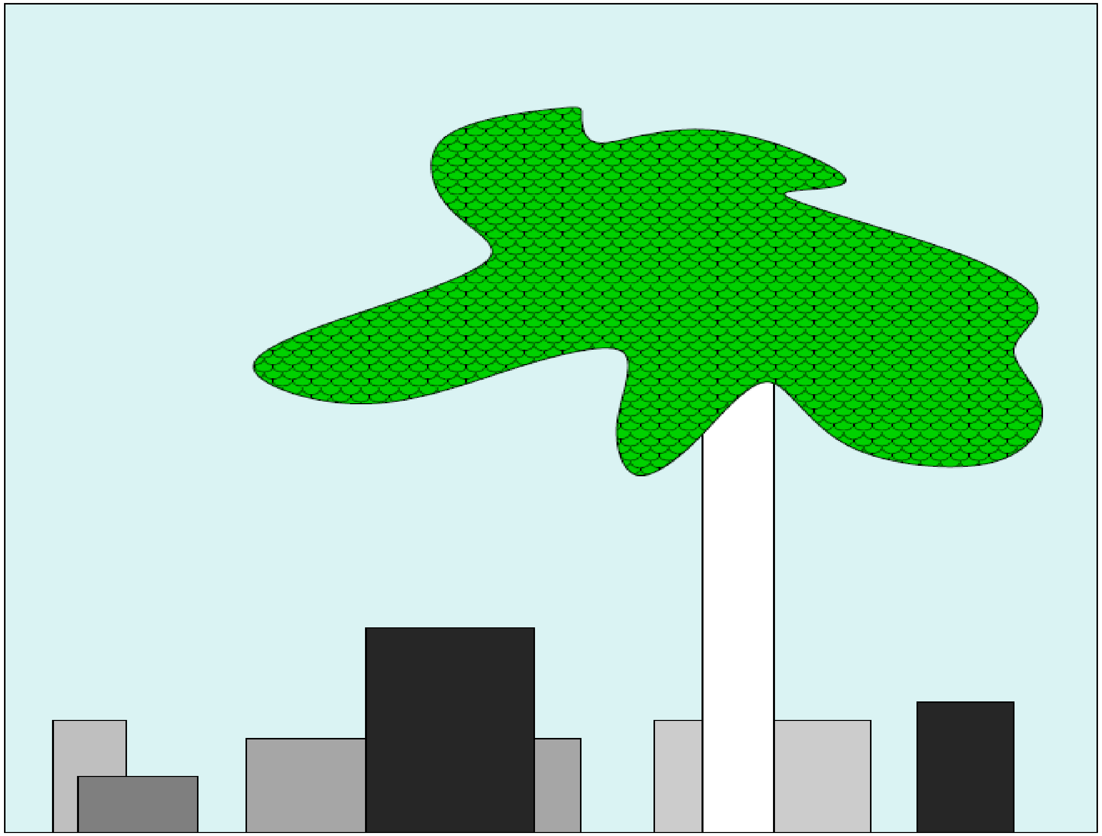
Outline

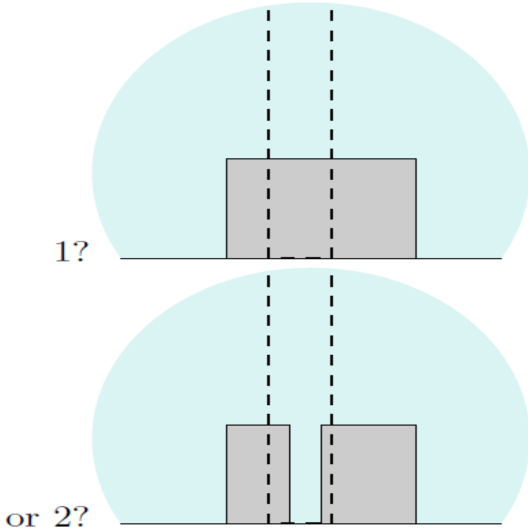
II. Model Comparison

Basic Concept
Mass Spectroscopy

Model Comparison

How many boxes are in the picture*)?





Desired: Occam's Razor (Prefer simpler models (that fit the data))

*)MacKay, Information theory

Model Comparison

Bayesian Approach:

$I = (M_1 + M_2 + \ldots)$ — Specify a set of models.
$H_i = M_i$ — Hypothesis chooses a model.

Posterior probability for a model:

$$p(M_i|D,I) = p(M_i|I)\frac{p(D|M_i,I)}{p(D|I)}$$
$$\propto p(M_i)\mathcal{L}(M_i)$$

Posterior for θ

But $\mathcal{L}(M_i) = p(D|M_i) = \int d\theta_i \, \underline{p(\theta_i|M_i)p(D|\theta_i, M_i)}$.

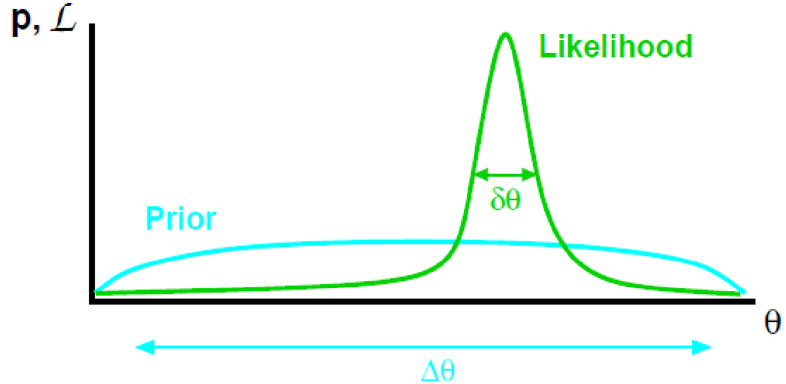
Likelihood for model = Average likelihood for its parameters

$$\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle$$

•Bayes Model Comparison requires always alternative models

Model Comparison

*The Occam Factor:**



$$p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i$$

$$\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i}$$

$$= \text{Maximum Likelihood} \times \text{Occam Factor}$$

Models with more parameters often make the data more probable— *for the best fit.*

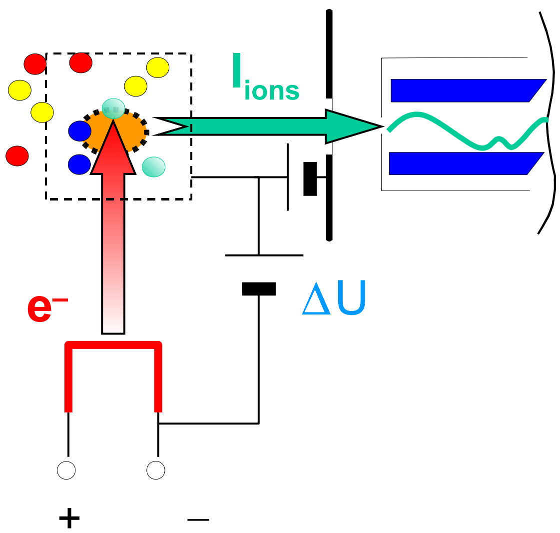
Occam factor penalizes models for "wasted" volume of parameter space.

*) T. Loredo, Garching, 2003

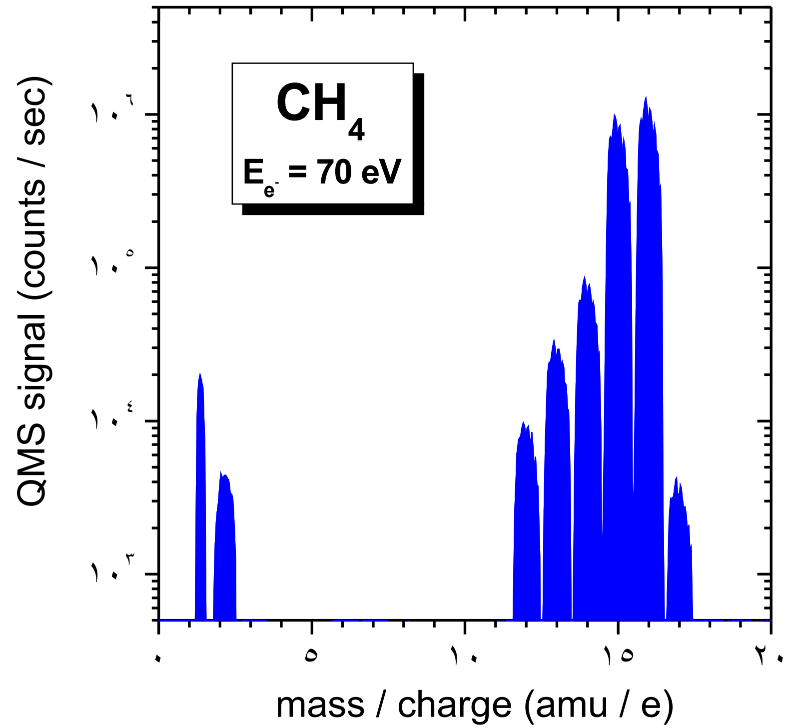
Mass Spectrometry

Quadrupole mass spectrometry with electron impact ionization:

$\mathbf{A} + e^- \Rightarrow \mathbf{A}^+ + 2e^-$



Electron impact ionization can lead to complex, **instrument** dependent fragmentation patterns:



Versatile tool for neutral gas analysis: fast, high dynamic range, flexible,...

$CH_4 + e^- \Rightarrow CH_3^+ + H + 2e^-$, $\quad CH_2^+ + 2H + 2e^-$

$CH^+ + 3H + 2e^-$, $\quad C^+ + 4H + 2e^-$

$H^+ + CH_3 + 2e^-$, $\quad$ ... (+^{13}C, D) ...

"a mass spectrometrist is someone, who figures out what something is, by smashing it with a hammer and looking at the pieces"

Quote from Th. Schwarz-Selinger

Mass Spectrometry

Model: $\mathbf{d}_j = \mathbf{C}\mathbf{x}_j + \epsilon_j$ with cracking matrix $\mathbf{C}$ ($\mathbf{C}$, $\mathbf{x}$ are uncertain/unknown)

Likelihood: $p\left(\mathbf{D}|\mathbf{C},\mathbf{X},\{\mathbf{S}\},E,I\right) =$

$$\prod_j \frac{1}{\prod_i \sqrt{2\pi s_{ij}}} \exp\left(-\frac{1}{2}\left(\mathbf{d}_j - \mathbf{C}\mathbf{x}_j\right)^T \mathbf{S}_j^{-1}\left(\mathbf{d}_j - \mathbf{C}\mathbf{x}_j\right)\right)$$

$\mathbf{S}$: measurement uncertainties

E : number of species

Prior terms:

Concentrations $\mathbf{x}$: depending on experiment (e.g. gas mixture)

Cracking matrix $\mathbf{C}$: exponential prior based on point estimates of Cornu&Massot (1979)

Probability for a set of species {E}: $p\left(E|\mathbf{D},\{\mathbf{S}\},I\right) = \dfrac{p\left(E|I\right)p\left(\mathbf{D}|\{\mathbf{S}\},E,I\right)}{p\left(\mathbf{D}|\{\mathbf{S}\},I\right)}$

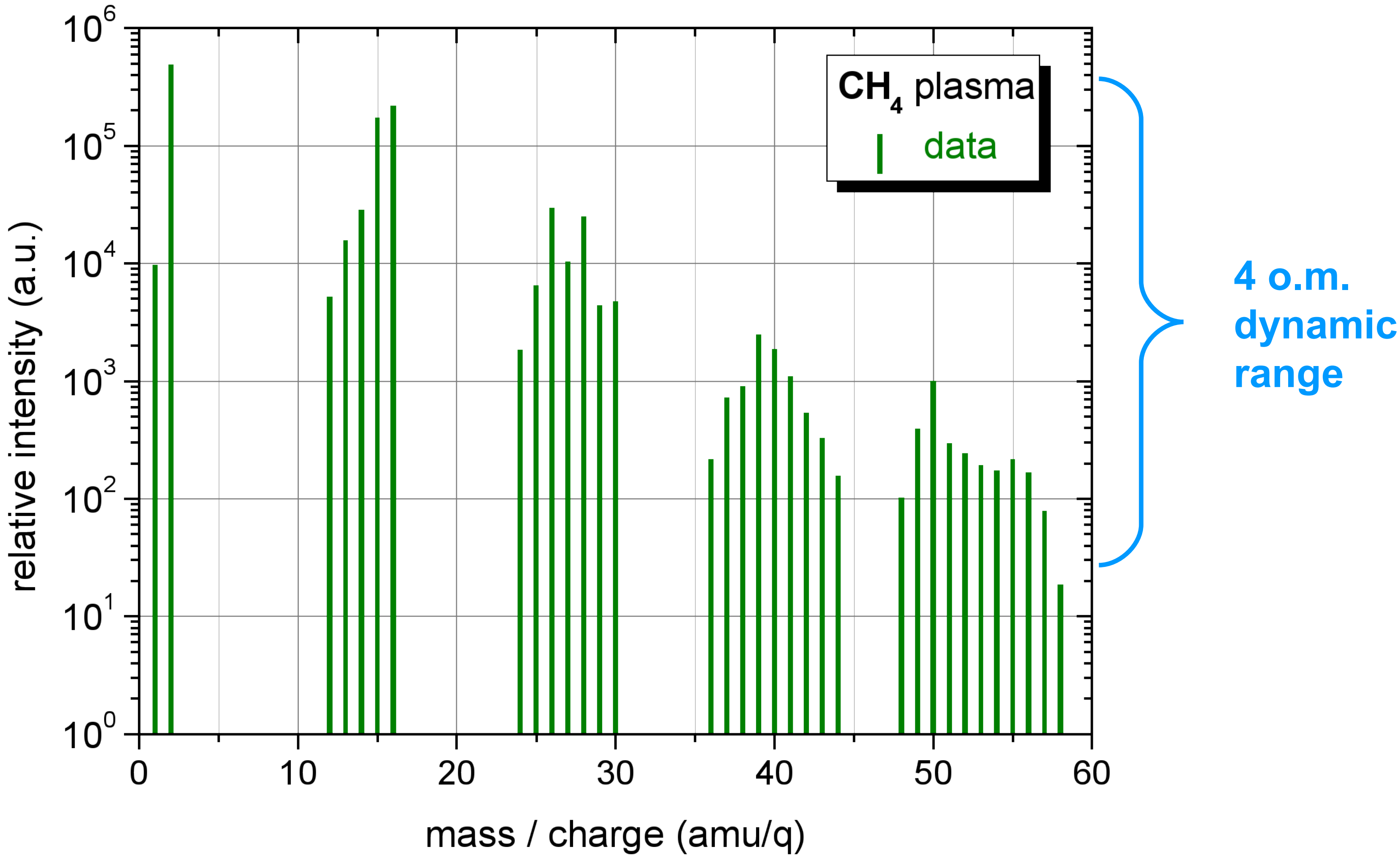
with $p\left(\mathbf{D}|\{\mathbf{S},E\},I\right) =$

$$\int \mathrm{d}\mathbf{C}\, d\mathbf{X}\, p\left(\mathbf{C}|E,I\right)p\left(\mathbf{X}|E,I\right)p\left(\mathbf{D}|\mathbf{C},\mathbf{X},\{\mathbf{S}\},E,I\right)$$

High-dimensional integrals!
MCMC-Integration

Mass Spectrometry

Measured data:



Mass Spectrometry

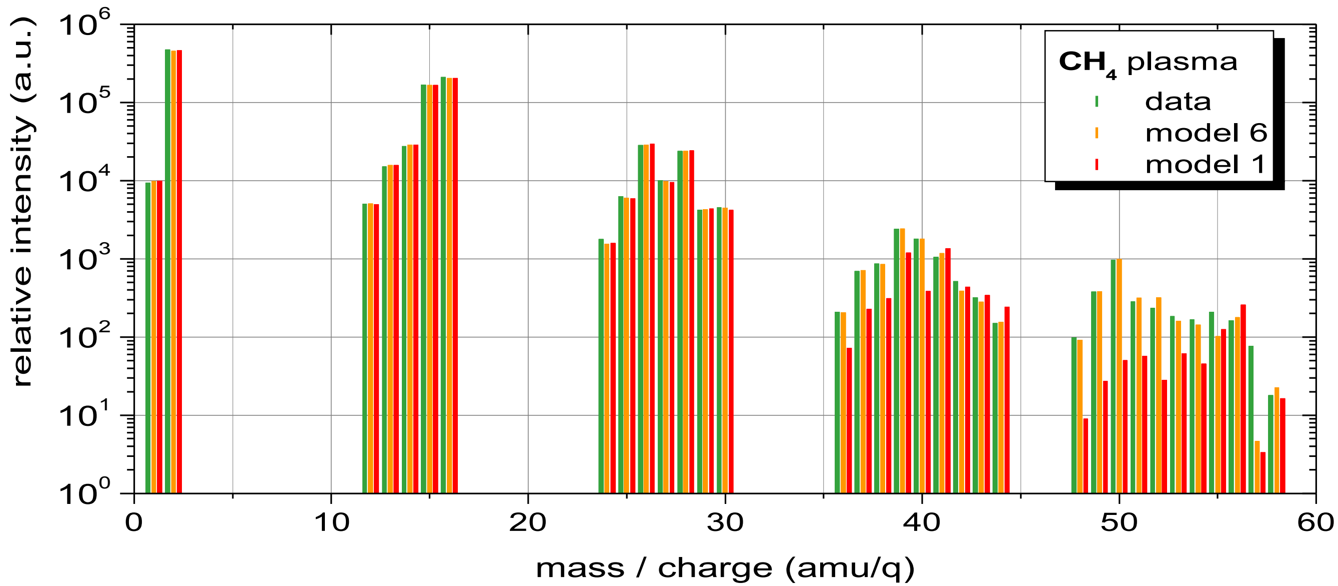
input:

data:
signal + error of 34 mass channels
for 27 different plasmas conditions

calibration measurements + error
for 11 species

prior:
cracking estimates for
14 species

ouput:
cracking pattern and
concentrations (+ errors!)
for 14 species
(e.g. C_4H_2 , C_3H_6)

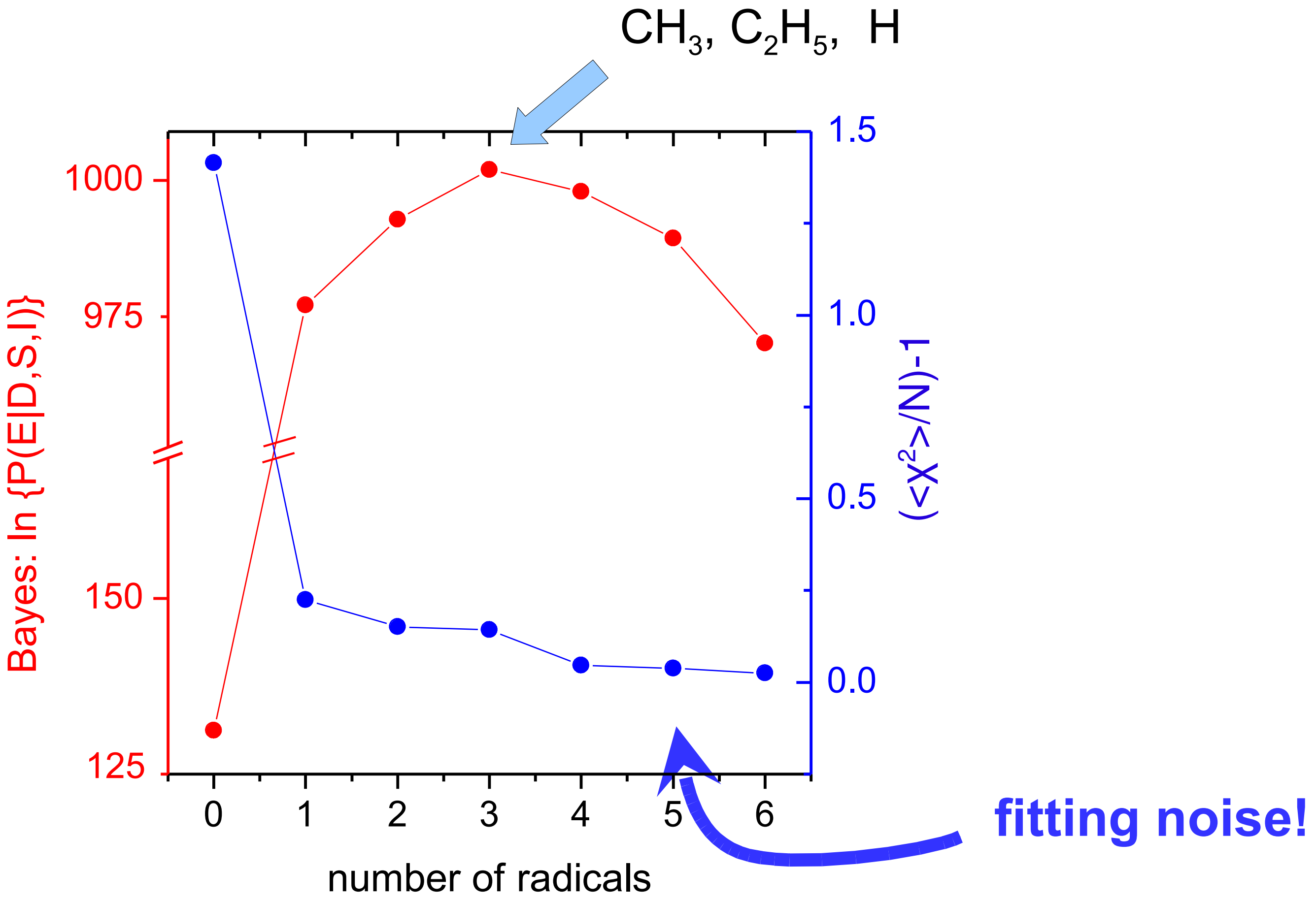


'shifted' cracking pattern for radical CH_3 very bad

*) Th. Schwarz-Selinger, H. Kang, U. von Toussaint et al, various publications (2001-2007)

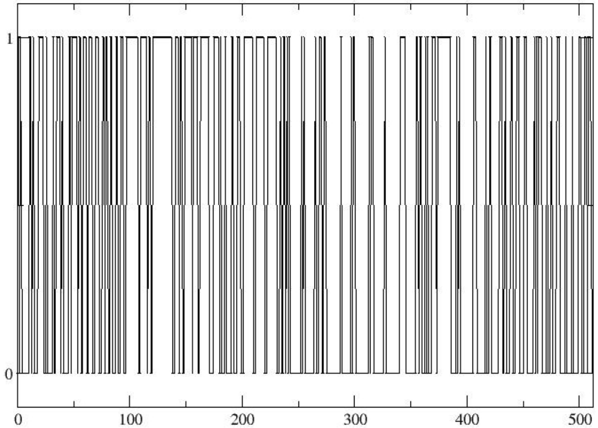
How many radicals are in the plasma?

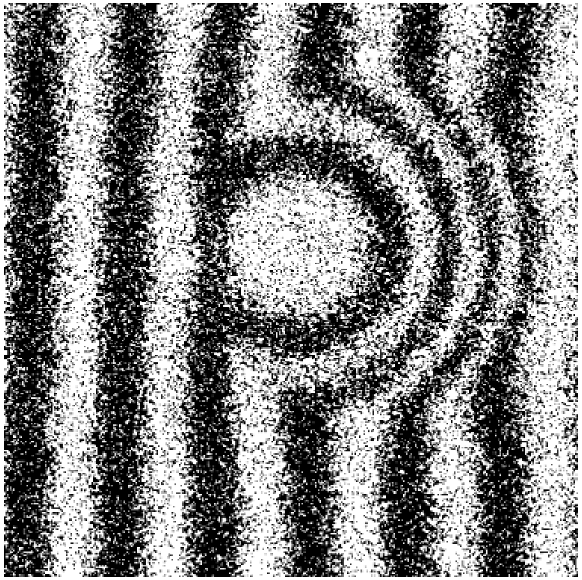
model comparison with „**Occams Razor**“

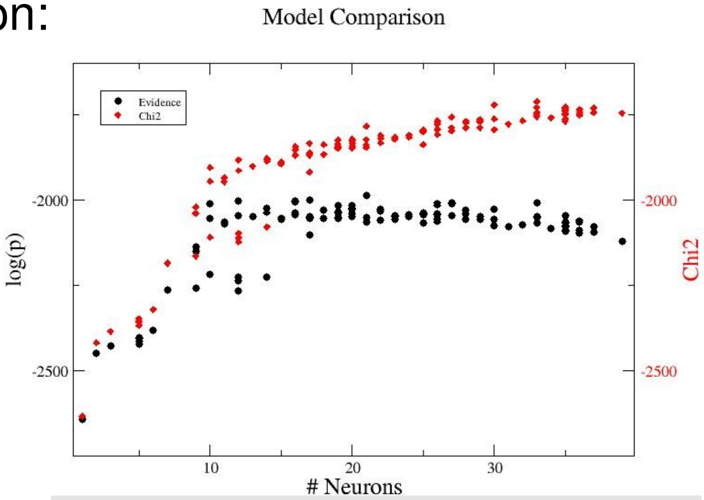


Model Comparison

An extreme example of model comparison:
Neural Networks without training data







U. von Toussaint et al, JAO, 2006

Outline

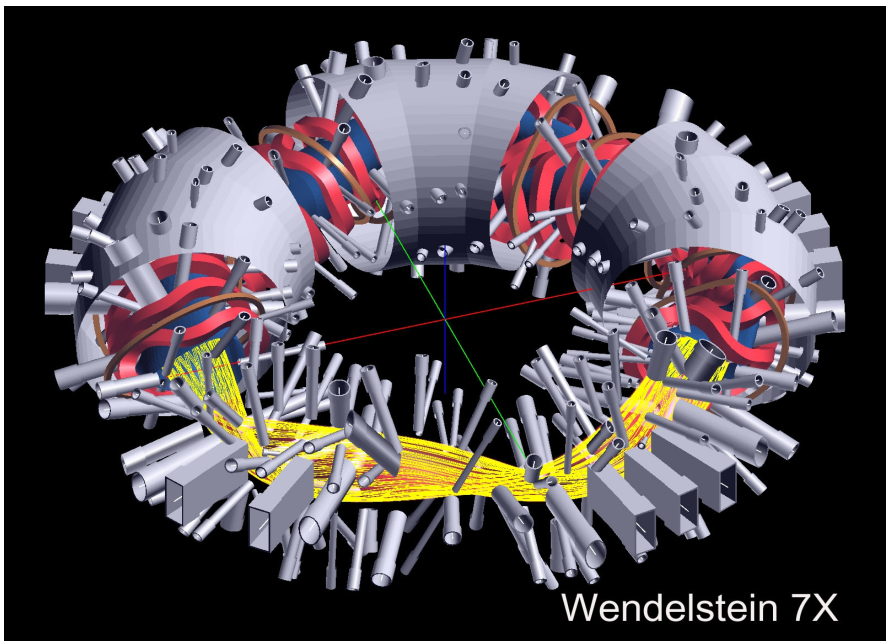
III. Experimental Design	Basic Concepts Nuclear Reaction Analysis

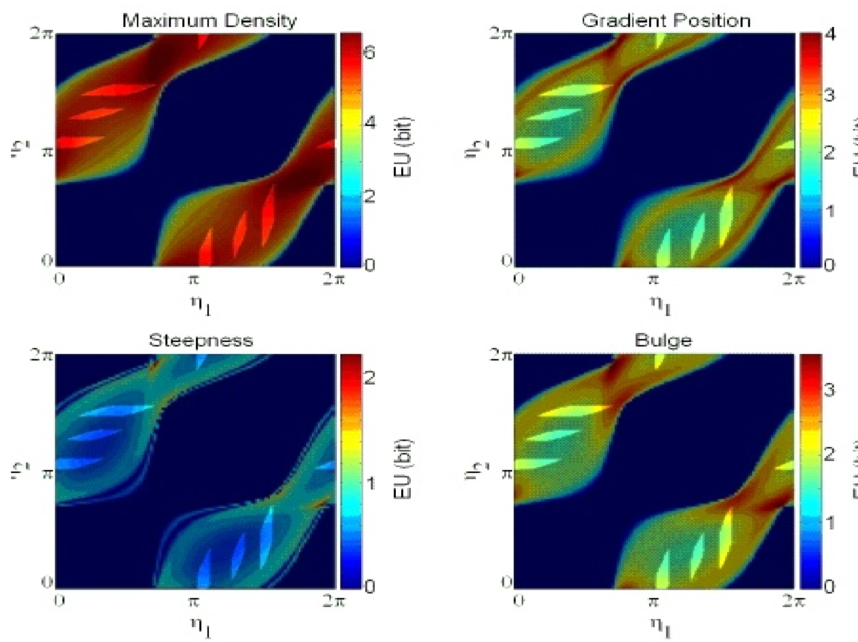
Experimental Design

Design of experiments:

Maximize information gain of planned experiments
Best performance for various physical scenarios,
hardware constraints, parameter constraints, financial budget

How to quantify the strength (weakness) of an experiment?
How to quantitatively chose experimental design parameter(s) ?
(spectral bands, number and position of line-of-sights, measurement times)





Bayesian Decision Theory

Decisions depend on consequences

Might bet on improbable outcome if payoff is large

Utility functions

Compare consequences via utility quantifying the benefits of a decision

Choice of action: c (eg next accelerator energy E_0)

Utility=U(c,o)

Outcome: o (eg next yield $d(E_0)$)

Deciding amidst uncertainty

We are uncertain of what the outcome will be: average possible results

Expected Utility: $$EU(c) = \sum_{\{outcomes\}} P(o|c)\, U(c,o)$$

Best choice maximizes EU: $c^* = \arg\max_{c} EU(c)$

Information as Utility:

Goal: minimize estimation uncertainty for parameter(s) **a**, maximize information gain

Utility function: Kullback-Leibler divergence K

(*generalizes* covariance-matrix)

$$Utility(d,c)=K(d,c)=\int da\, p(a|d,c)\ln\frac{p(a|d,c)}{p(a)}$$

(c=action, d=expected data)

Putting it all together:

$$EU(c)=\int dd \;\; p(d|c)K(d,c)=\int da \;\; p(a)\int dd \;\; p(d|a,c)K(d,c)$$

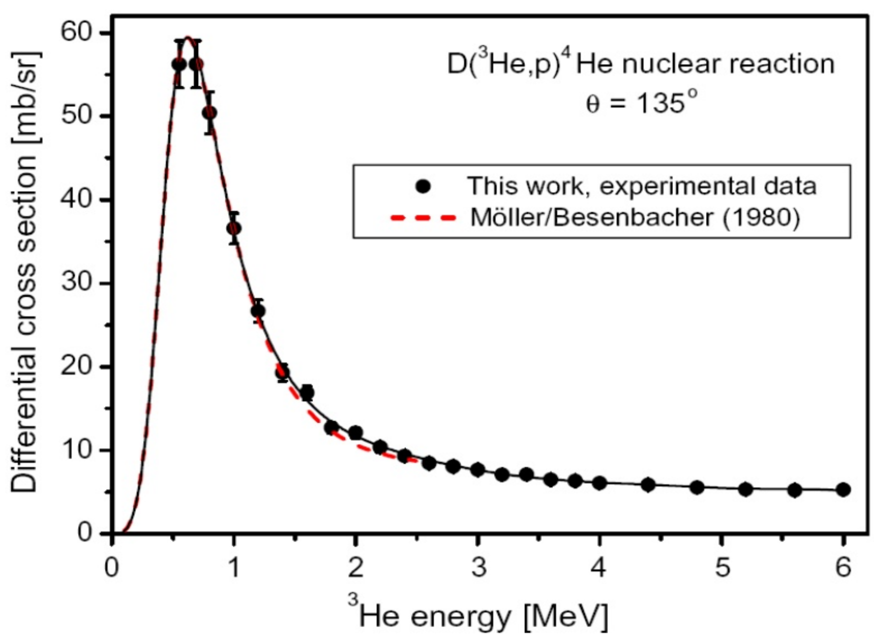
Best choice maximizes EU: c*=arg max EU(c)

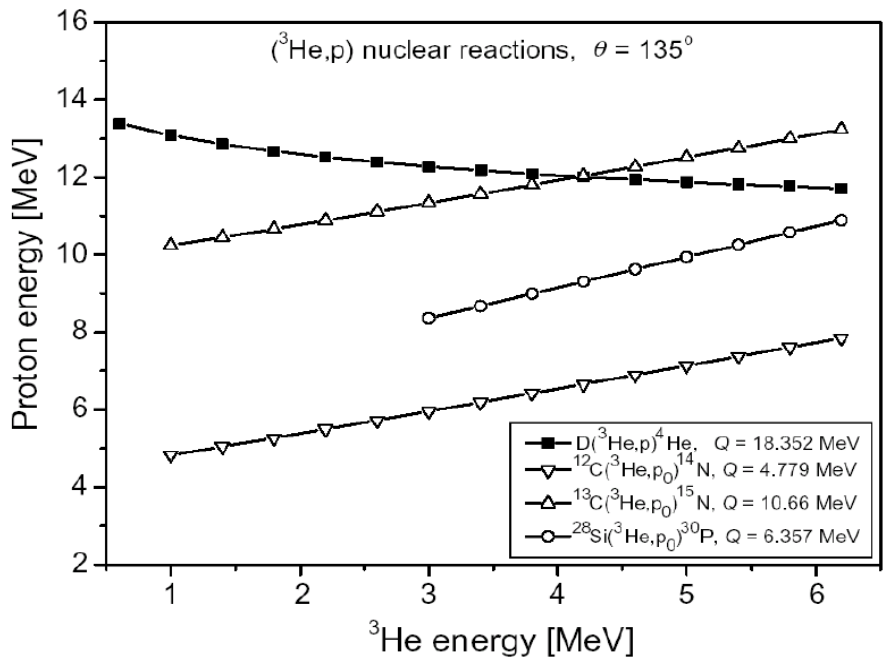
High- dimensional integration necessary: Markov Chain Monte Carlo-Methods

Experimental Design

- Hydrogen isotope depth profiling using NRA (not too many other ways)
- nuclear reaction: $d(^3He,p)^4He$: background free, range

$$Y_i \propto \int_0^\infty dx\, \sigma\left(E\left(\vec{c}\left(x\right), x, E_{i0}\right)\right)\left(\frac{dE}{dx}\right) \cdot c_j\left(x\right) + \epsilon_i$$





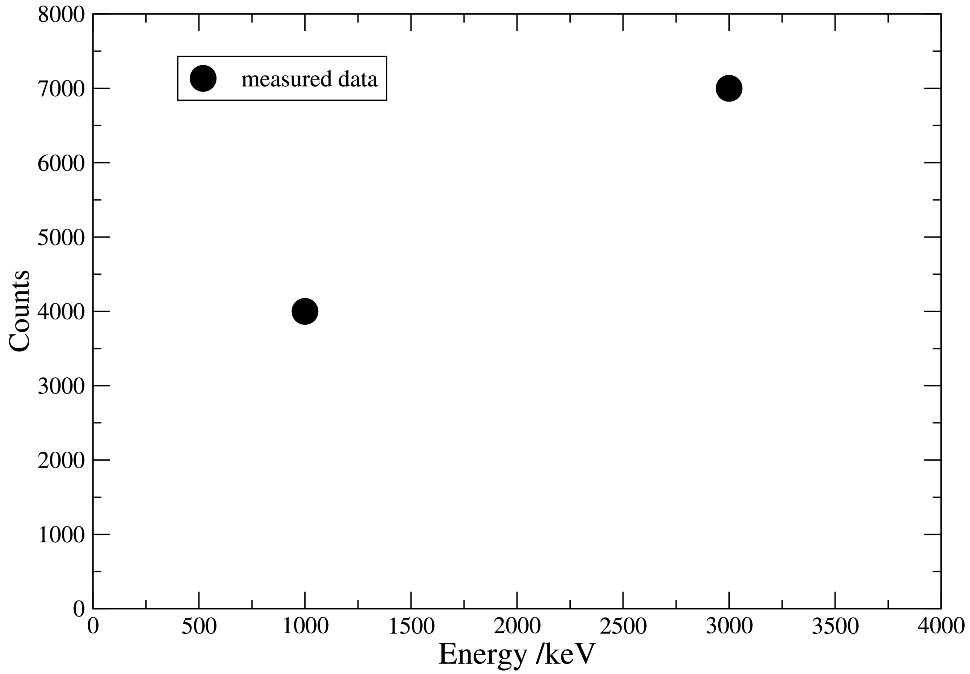
V. Alimov, J. Roth, NIMB, 2006

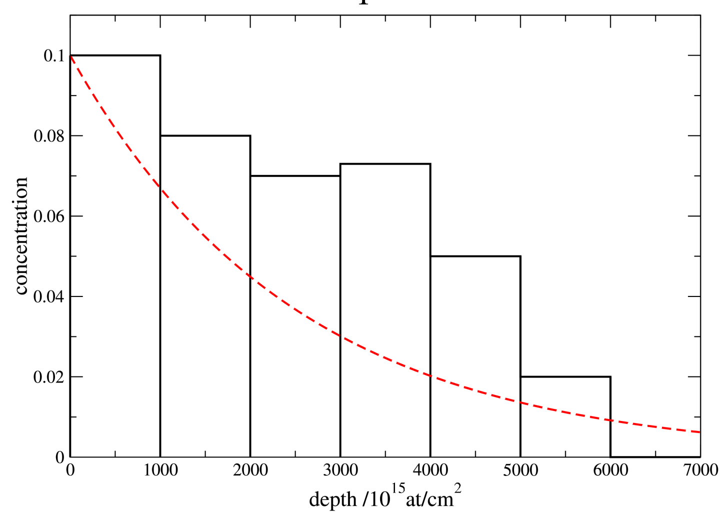
Experimental Design

$$Y_i \propto \int_0^\infty dx\, \sigma\left(E\left(\vec{c}\left(x\right), x, E_{i0}\right)\right)\left(\frac{dE}{dx}\right) \cdot c_j(x) + \epsilon_i$$

- depth profile has to be parametrized: piecewise constant or analytically:

$c(x)=\{a_0,a_1,a_2,a_3,\ldots\}$ or $c(x) = a \cdot \exp\left(-\frac{x}{a_1}\right) + a_2$



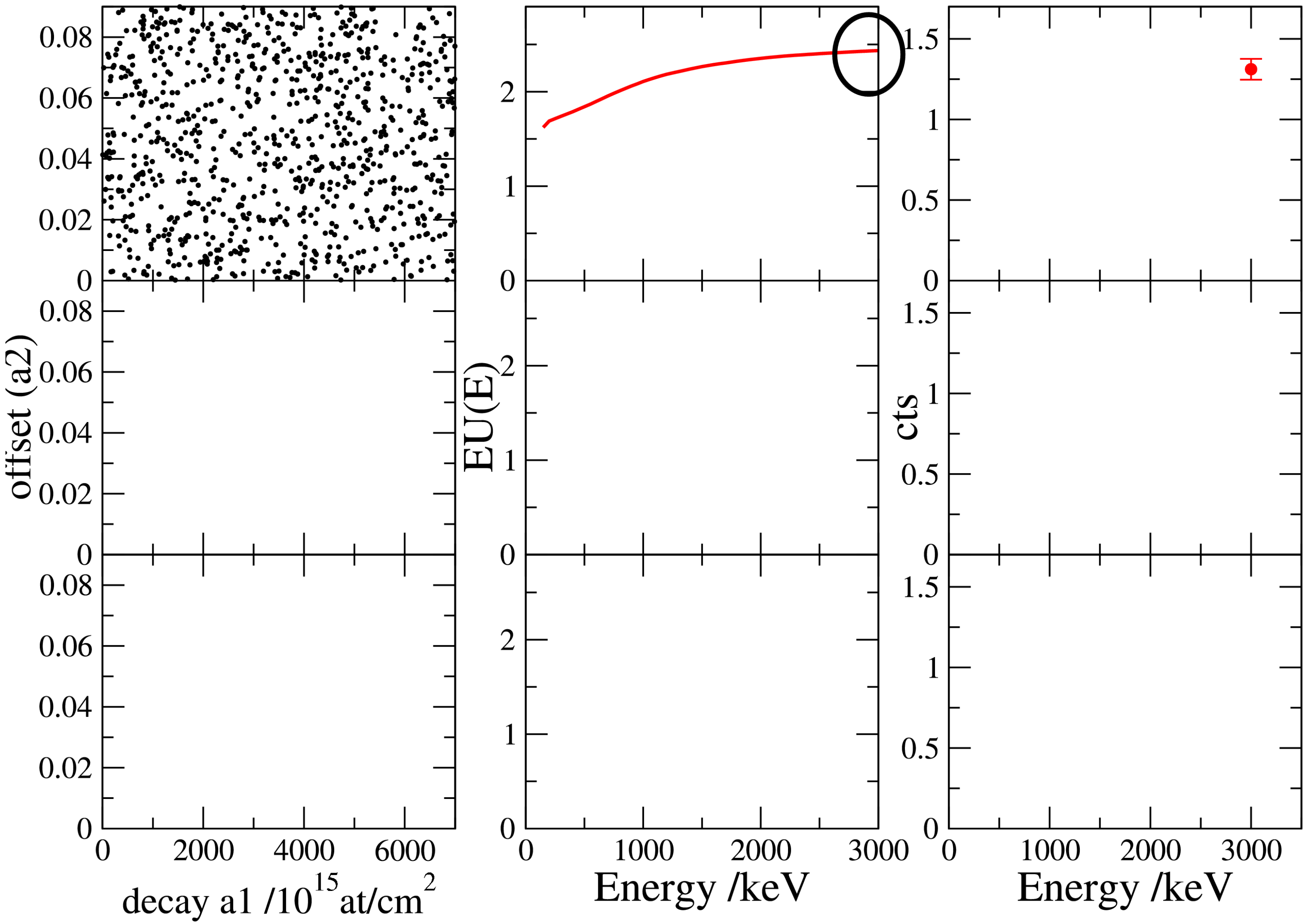


Sequential Design: Which energy E_0 next?

Experimental Design

Compute EU(E), and select best energy

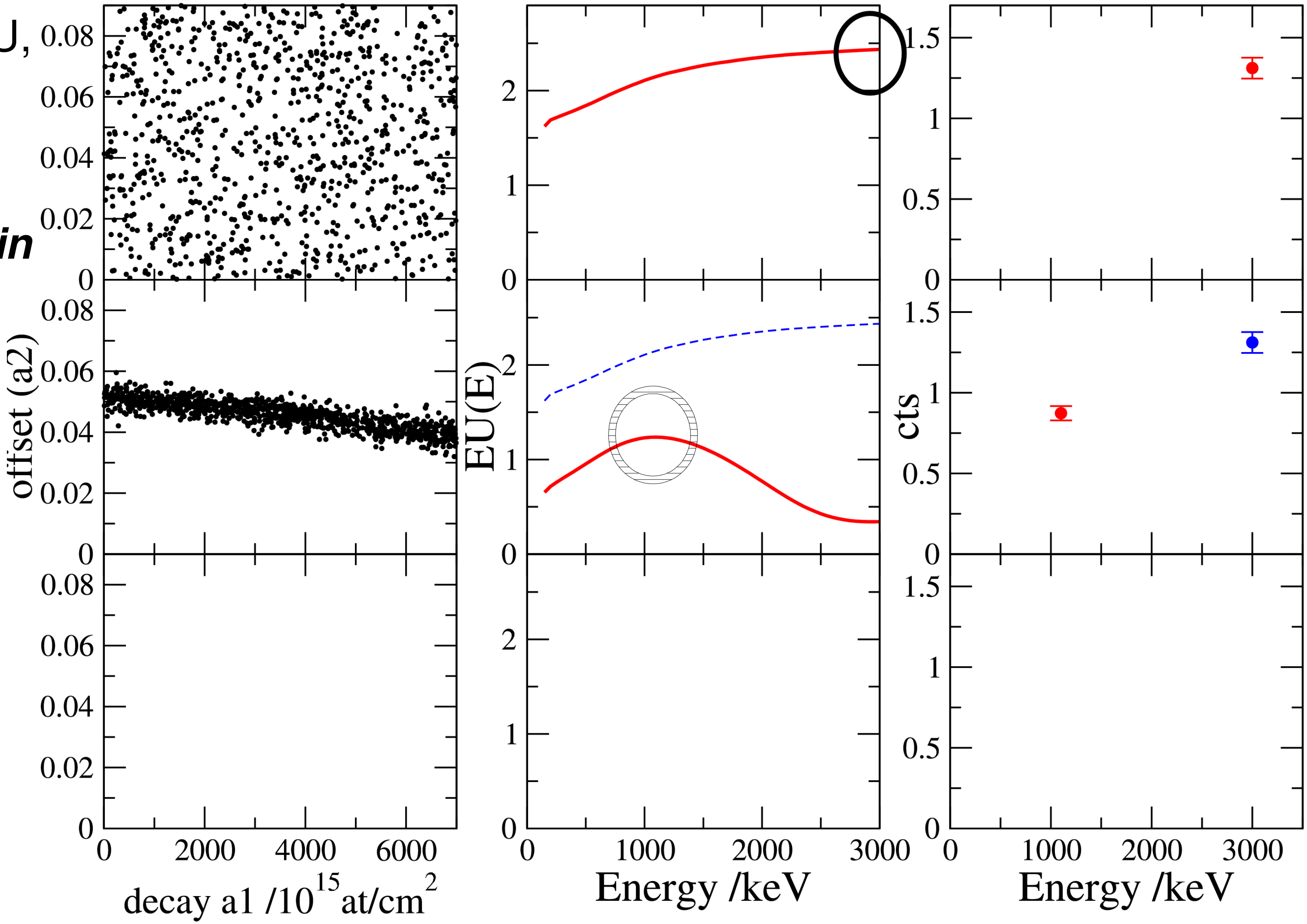
Integration by *posterior sampling*
*CPU:**1-4 min***



Experimental Design

Optimize EU, using *posterior sampling*
*CPU:**1-4 min***

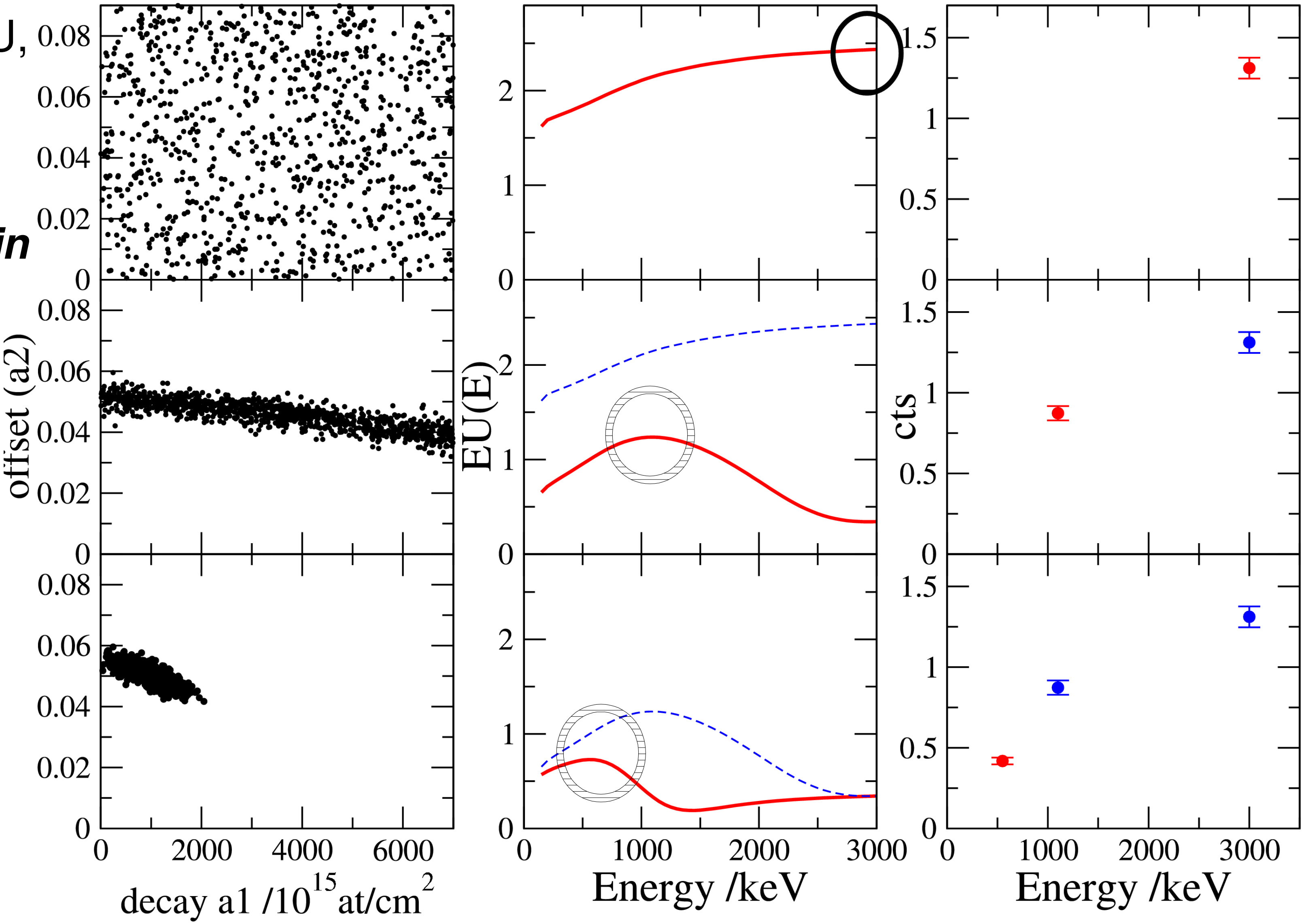
Cycle:
Prediction -
Verification



Experimental Design

Optimize EU, using *posterior sampling* *CPU:**1-4 min***

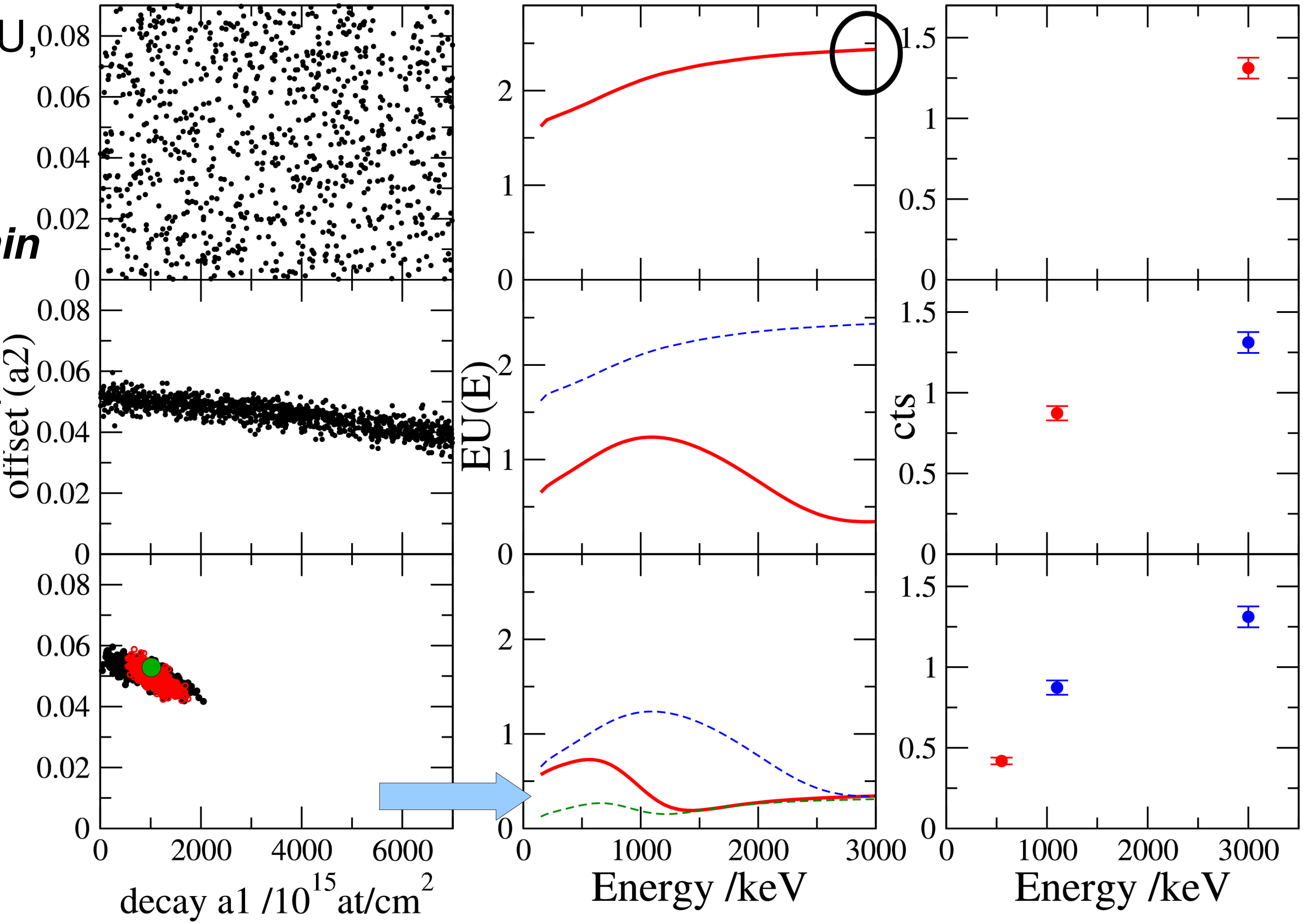
Cycle: Prediction - Verification



Experimental Design

Optimize EU, using *posterior sampling*
*CPU:**1-4 min***

Cycle: Prediction - Verification



Experimental Design

- Quantification of information

  Value of measurements /diagnostics / experimental setups accessible
  Optimized measurements protocols (on-line sequential design)

- Linear and **nonlinear** problems can be tackled

- Necessary integrations computationally demanding:

  High-dimensional integrals in data- and parameter space
  Anything beyond greedy algorithms unexplored (n-step ahead designs)

- Considerable gain in efficiency and/or accuracy has been demonstrated

- **Result-driven** automated measurement strategies (⇨ robotics) conceivable

Outline

IV. Numerical Interlude | Nested Sampling

Nested Sampling*

Bayesian inference depends on (high-dimensional) **integration**...

$$\underbrace{p(D\,|\,\theta,\mathcal{H})}_{\text{likelihood}}\,\underbrace{p(\theta\,|\,\mathcal{H})}_{\text{prior}} \;=\; \underbrace{p(\theta\,|\,D,\mathcal{H})}_{\text{posterior}}\,\underbrace{p(D\,|\,\mathcal{H})}_{\text{evidence}}$$

posterior

$$p(\theta\,|\,D,\mathcal{H}) \;=\; \frac{p(D\,|\,\theta,\mathcal{H})\,p(\theta\,|\,\mathcal{H})}{p(D\,|\,\mathcal{H})}$$

evidence

$$p(D\,|\,\mathcal{H}) = \int_{\theta} p(D\,|\,\theta,\mathcal{H})\,p(\theta\,|\,\mathcal{H})\,\mathrm{d}\theta$$

model comparison

$$\begin{aligned} p(\mathcal{H}_1\,|\,D) &\propto p(D\,|\,\mathcal{H}_1)\,p(\mathcal{H}_1) \\ p(\mathcal{H}_2\,|\,D) &\propto p(D\,|\,\mathcal{H}_2)\,p(\mathcal{H}_2) \end{aligned}$$

*J. Skilling 2004; MacKay 2006

Relation to statistical physics: partition function

Free energy

Probability of macrostate $\mathcal{H}$ is proportional to:

$$\mathcal{Z}(\beta, \mathcal{H}) = \int_\theta \exp(-\beta E(\theta, \mathcal{H}))\, \mathrm{d}\theta$$

Evidence

How well a model $\mathcal{H}$ predicted the data:

$$p(D \mid \mathcal{H}) = \int_\theta p(D \mid \theta, \mathcal{H})\, p(\theta \mid \mathcal{H})\, \mathrm{d}\theta$$

Common problem:

$$\mathcal{Z} = \int_\theta L(\theta) P(\theta)\, \mathrm{d}\theta$$

$L(\theta) \equiv \exp(-\beta E(\theta, \mathcal{H}))$

Boltzmann factor

$L(\theta) \equiv p(D \mid \theta, \mathcal{H})$

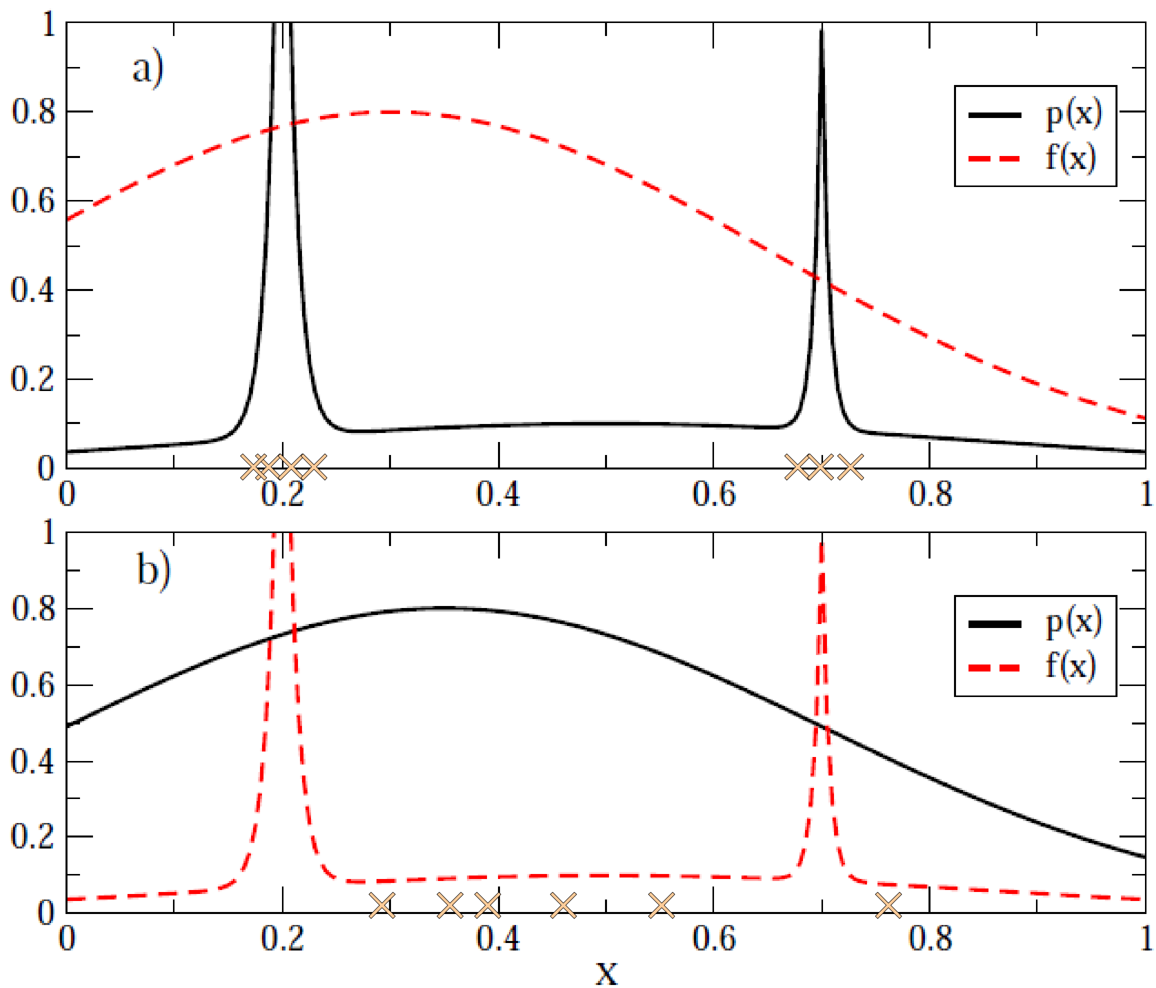
Likelihood function

Why is this integration difficult?

Nested Sampling

Typical situation for expectation values <f>:

Typical situation for evidence calculation:



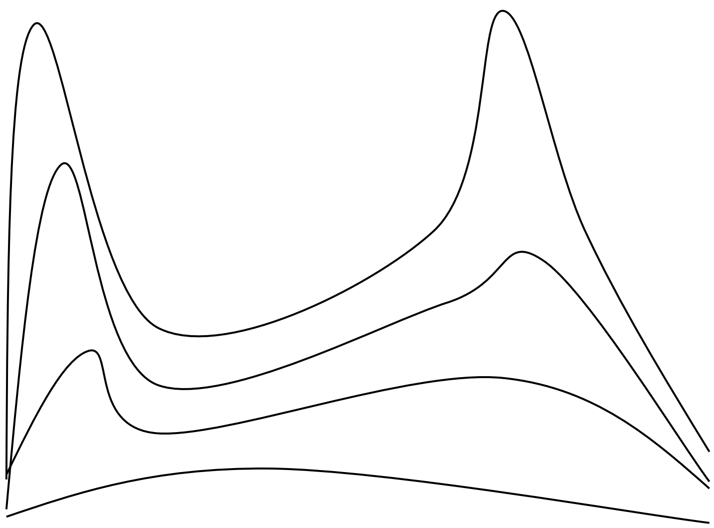
Nested Sampling

Thermodynamic Integration: Slowly introduce likelihood structure into integrand (similar: parallel tempering, perfect tempering):

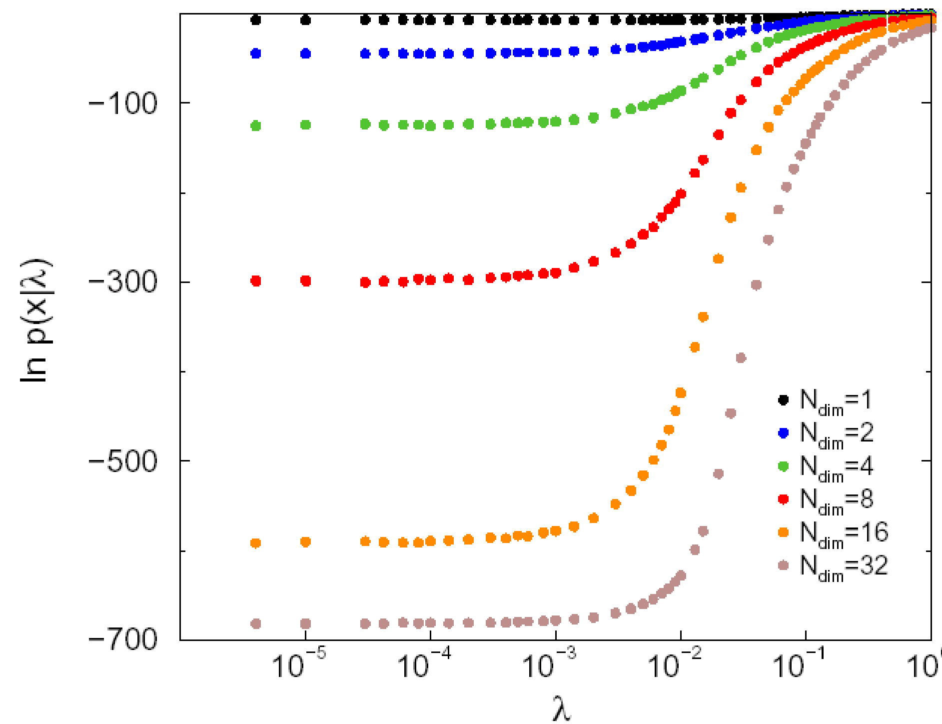
$$Z(\lambda) = \int d\underline{x}\, \Gamma^{\lambda}(\underline{x})\pi(\underline{x})$$

Z(0)=1 and Z(1)=Evidence

$$\ln(I) = \int_{\lambda=0}^{\lambda=1} d\lambda \, \frac{\partial \ln Z(\lambda)}{\partial \lambda} = \int_{\lambda=0}^{\lambda=1} d\lambda \int d\underline{x} \ln \Gamma(\underline{x}) \rho_{\lambda}(\underline{x})$$

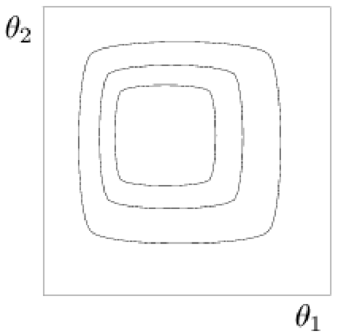


R. Preuss, U. von Toussaint, AIP, 2007



Nested Sampling

$$\mathcal{Z} = \int_\theta L(\theta) P(\theta)\, d\theta$$

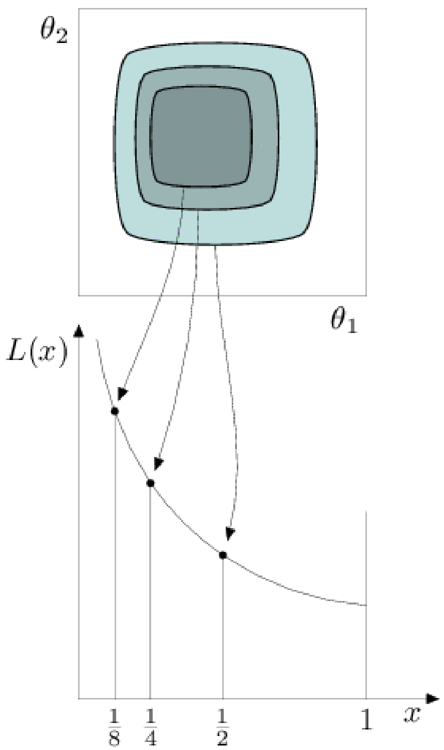


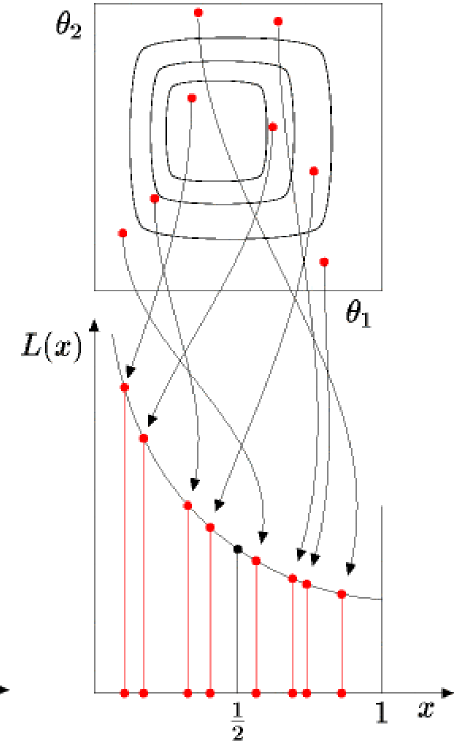
Contour plot of L

$$\mathcal{Z} = \int_\theta L(\theta) P(\theta)\, d\theta = \int_0^1 L(x)\, dx$$

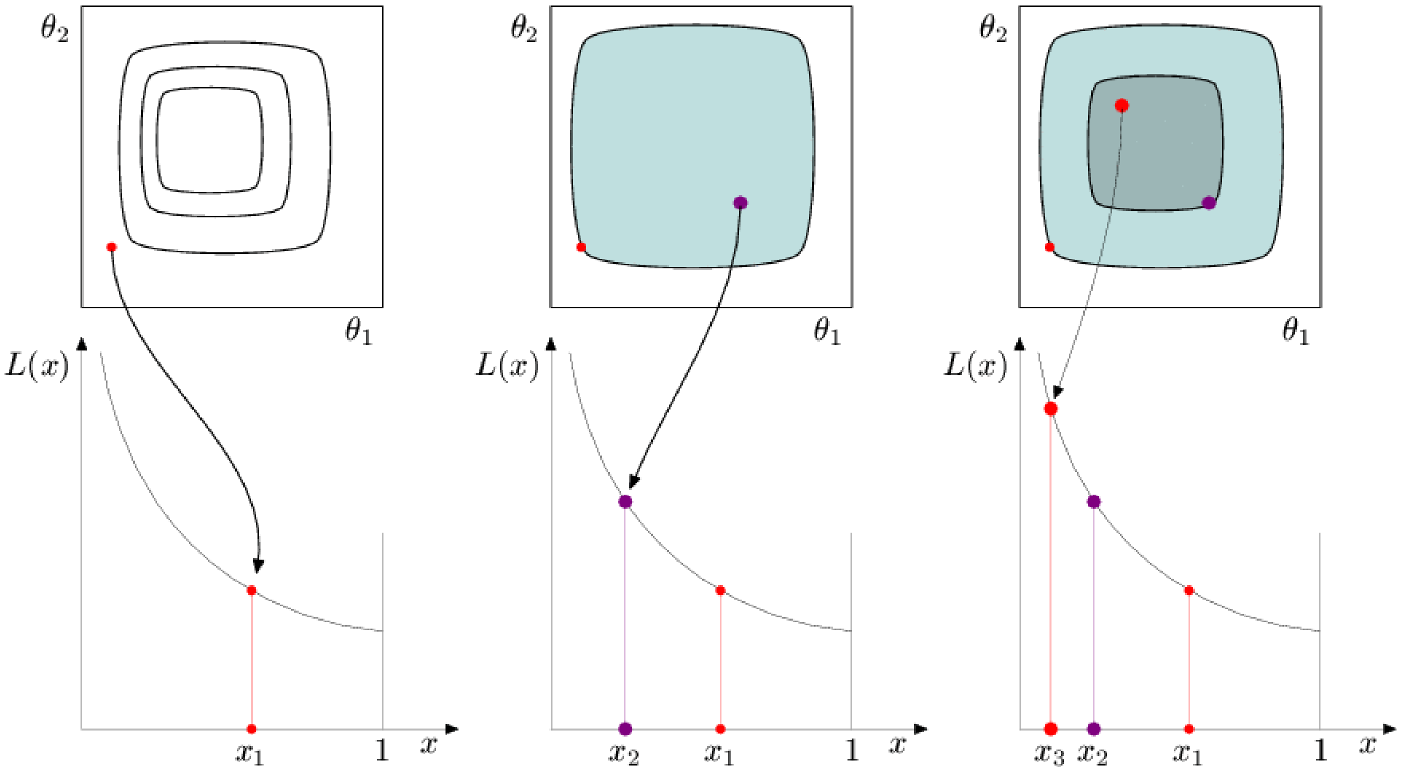
$$dx = P(\theta)\, d\theta$$

- Key concept:
  - Sort all points by L





Nested Sampling



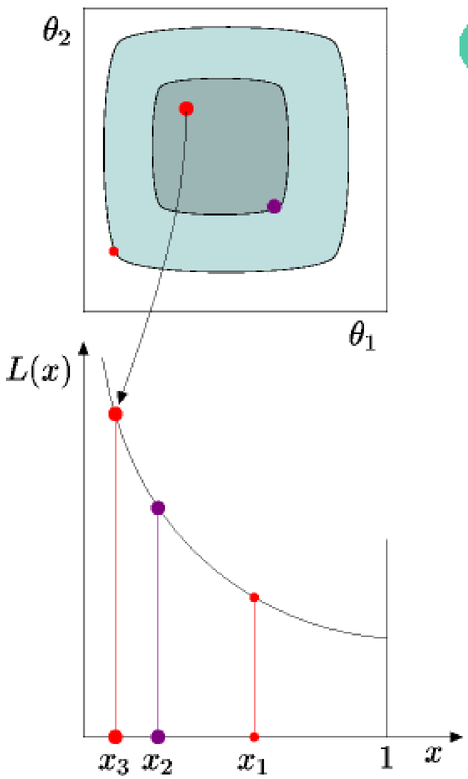
$$P(\theta^{(1)}) = P(\theta)$$

$$P(\theta^{(i+1)}) \propto \begin{cases} P(\theta) & L(\theta) > L_i \\ 0 & \text{otherwise} \end{cases}$$

$x_1 \sim \text{Uniform}(0, 1)$ $\quad$ $x_2 \sim \text{Uniform}(0, x_1)$ $\quad$ $x_3 \sim \text{Uniform}(0, x_2)$

$\langle x_1 \rangle = \frac{1}{2}$ $\quad$ $\langle x_2 \rangle = \frac{1}{2}\langle x_1 \rangle = \frac{1}{4}$ $\quad$ $\langle x_3 \rangle = \frac{1}{8}$

Nested Sampling



- Draw from

$$P(\theta^{(i+1)}) \propto \begin{cases} P(\theta) & L(\theta) > L_i \\ 0 & \text{otherwise} \end{cases}$$

- cf Annealing's intermediate distributions

$$P(\theta \mid \beta) \equiv \frac{1}{\mathcal{Z}(\beta)} L(\theta)^{\beta} P(\theta)$$

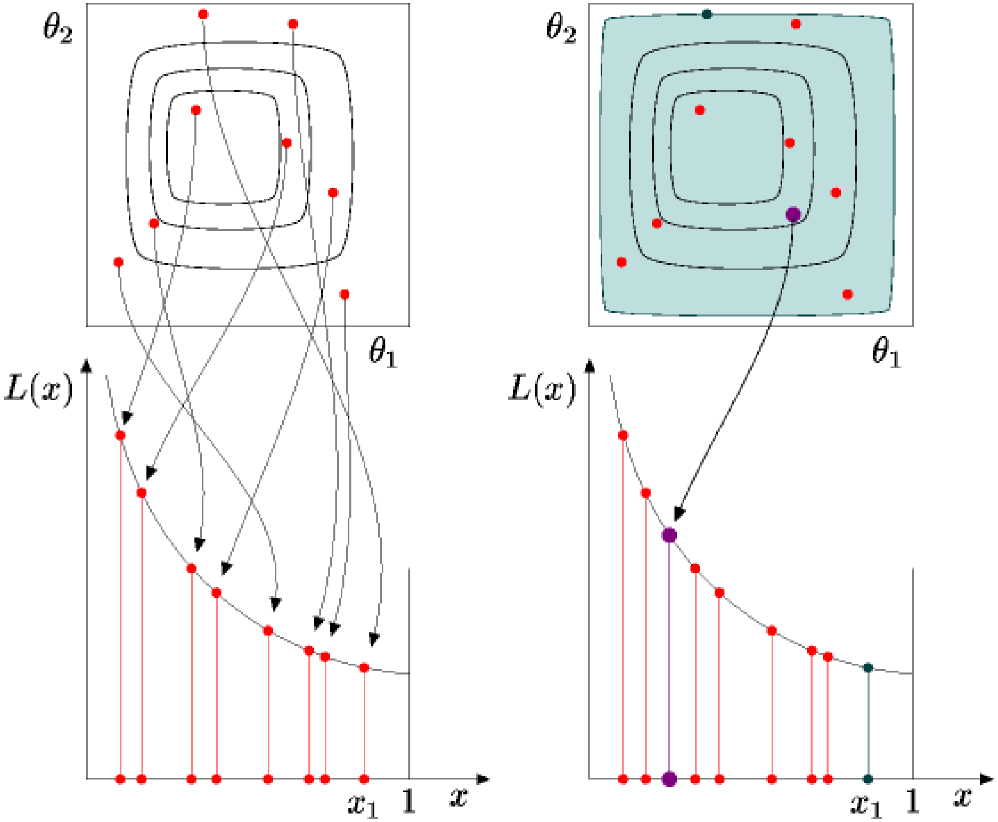
- Assuming uniform sampling subject to an energy constraint is possible, $P(\mathbf{X})$ is simple

$$\left\langle \log \frac{x_{i+1}}{x_i} \right\rangle = -1 \qquad \text{Order statistics for P(x)}$$

independent of L

- Nested sampling's behaviour is invariant under monotonic transformations of L

Nested Sampling



The result:

$$\hat{Z} = \sum_i \delta\widehat{x}_i \, L_i$$

$$\widehat{x}_i \equiv \exp(-i/N)$$

Pro: Quite different approach, very general, easy to implement

Con: Uniform sampling under constraint?

Outline

V. Conclusion	Summary Outlook

Summary

- Bayesian probability theory provides consistent and transparent approach to the cycle of scientific inference

- Incorporation of available prior information is straightforward

- Drawback of numerical complexity mitigated by
    - New algorithms
    - Increasing computing power

- State of the Art: parameter estimation, model comparison

- Coming soon: probabilistic combination of diagnostics (IDA)

- Still (largely) unexplored:
    - potential of Experimental Design, e.g. robotics, self-adapting 'measurement'-strategies on computer simulations (e.g. automated MD potential generation)
    - novelty detection in large scale simulations / experiments

Outlook

• Promising and/or unexplored research directions

Bayesian Experimental Design

Large data sets and complex (simulation based models) require consistent response-surface estimates (O'Hagan)

Estimation of stochastic partial differential equations or functionals from data (e.g. for turbulence)

Data based model design and model estimation, e.g. with respect to possible causation instead of correlation only (Pearl)

•32th Workshop on Bayesian Inference and Maximum Entropy Methods at IPP Garching (15.-20. July 2012)
See: **http://www.ipp.mpg.de/maxent2012**

32th Workshop
on
Bayesian Inference and Maximum Entropy Methods
at
IPP Garching (15.-20. July 2012)

See: **http://www.ipp.mpg.de/maxent2012**

Thank you
for
your attention