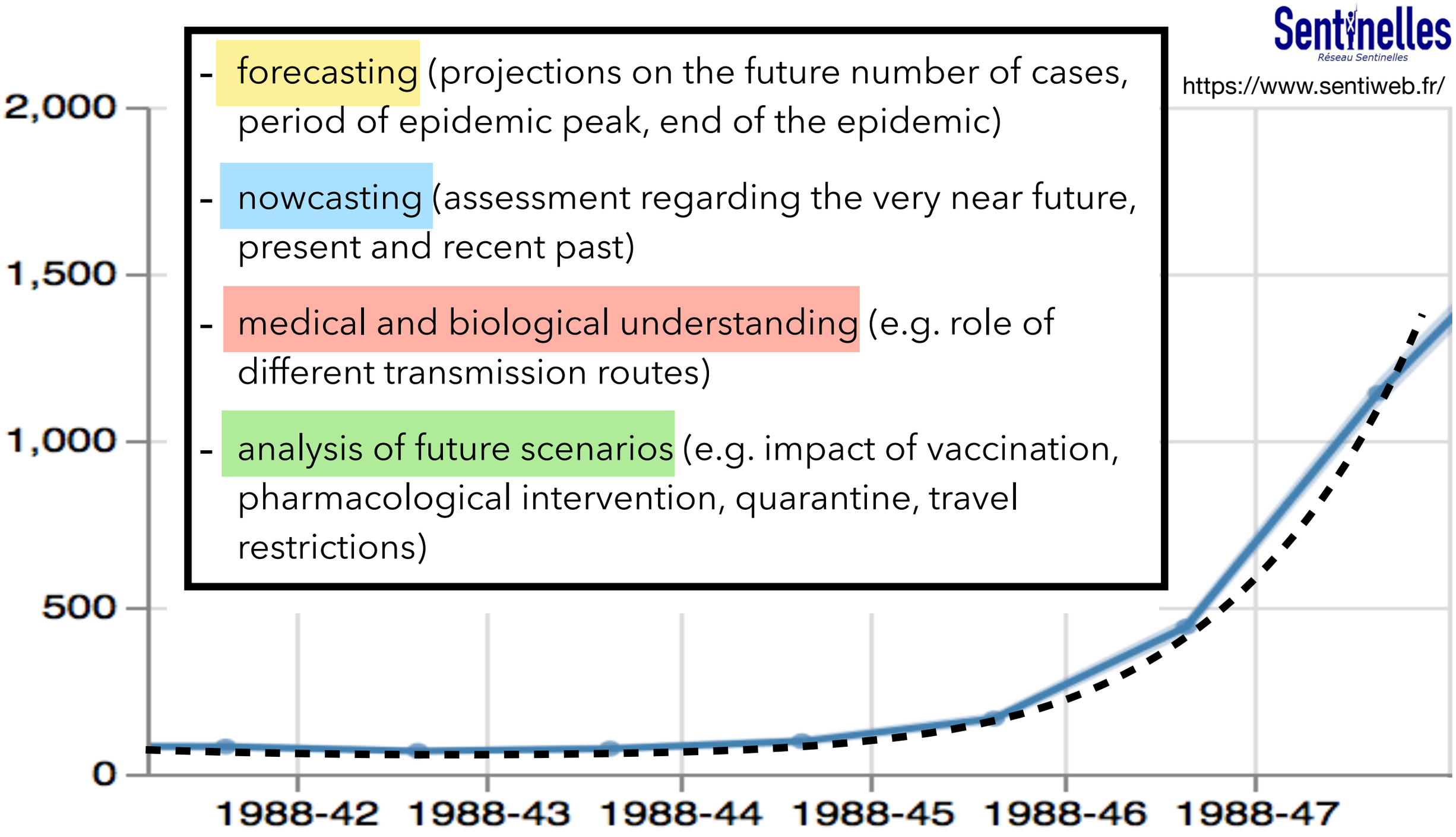
# LIFE DATA EPIDEMIOLOGY

*lect. 6: Model fitting*

Chiara Poletto
polettoc@gmail.com

# epidemiology: making sense of data

- forecasting (projections on the future number of cases, period of epidemic peak, end of the epidemic)
- nowcasting (assessment regarding the very near future, present and recent past)
- medical and biological understanding (e.g. role of different transmission routes)
- analysis of future scenarios (e.g. impact of vaccination, pharmacological intervention, quarantine, travel restrictions)



# Epidemic modeling

Model design

Model implementation

decide the model ingredients that synthesise available medical, biological, etc., information. We can consider different models, ingredients

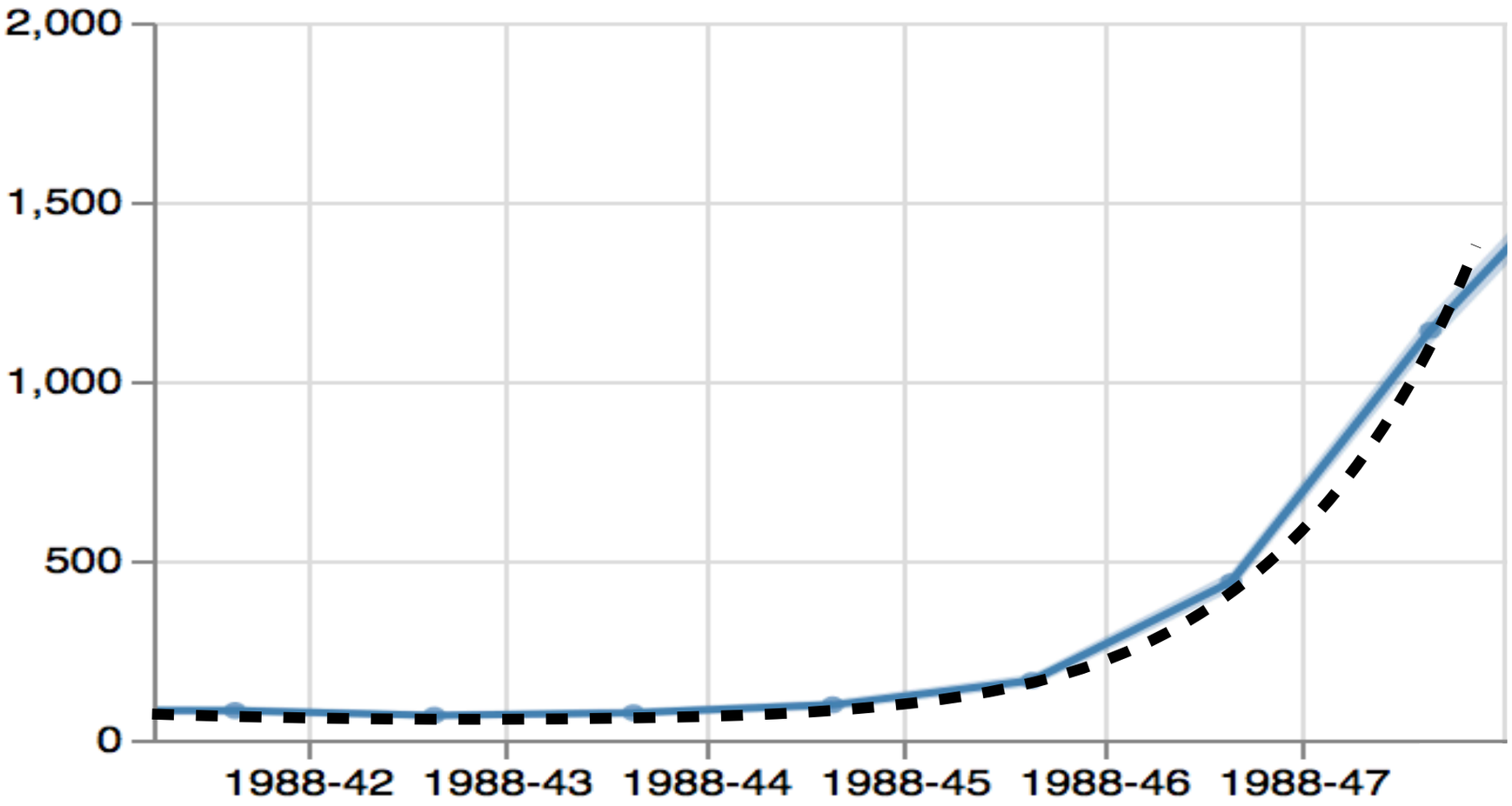
Model calibration

estimate model parameters from available data

Model validation

confirming that model output is sufficiently accurate in reproducing the data

# Model calibration




Sentinelles
Réseau Sentinelles

https://www.sentiweb.fr/

▬ ▬ modelled trajectory = $F(M, \theta, I_0)$

$M$ = epidemic model — e.g. SIR

$\theta$ = vector of parameters — e.g. $(\beta, \mu)$ = (transmission rate, recovery rate)

$I_0$ = initial conditions — e.g. initial number of infectious

# Maximum likelihood approach

probabilistic formulation:

- relation between model and data is probabilistic

- we want that identify the trajectory and thus the parameters, $\theta$, that are more probable given the data

**Bayesian framework: probability as a measure of uncertainty**

# The very basics of probability theory

**Basic definitions**

*univariate probability*

- $A$ : random variable
- $p(A = a) = p(a)$ : probability that $A$ takes value $a$
- normalisation : $\sum_a p(a) = 1$

*multivariate probability*

- $A$ and $B$ : random variables
- $p(A = a, B = b) = p(a, b)$ : joint probability that $A$ takes value $a$ and $B$ takes value $b$
- marginal probability: $p(a) = \sum_b p(a, b)$

# The very basics of probability theory

**Basic properties**

- conditional probability of *a* from random variable *A*, given that the outcome of random variable *B* was *b* :
  $p(A = a \mid B = b) = p(a \mid b)$
- Bayes Theorem : $p(a \mid b) = \frac{p(a, b)}{p(b)}$
- Chain rule: $p(a, b, c) = p(a \mid b, c)p(b \mid c)p(c)$

# The very basics of probability theory

**Continuous variables**

- Normalization : $\int p(a)\, da = 1$

- Marginal probability : $p(a) = \int p(a, b)\, db$

# basic probability distributions

## Poisson distribution

The events occur with a known constant rate $r$ and independently each other. The probability of having $y$ occurrence in an interval of time $\Delta$ follows a Poisson

$$p(y \mid \Delta, r) = \frac{(r\,\Delta)^y e^{-r\,\Delta}}{y!}$$

average number of occurrence: $\lambda = r\,\Delta$

The Poisson is parametrised as a function of $\lambda$:

$$p(y \mid \lambda) = \text{Poisson}(y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

# basic probability distributions

## Bernoulli trials and Binomial distribution

Ingredients:

- $n$ exchangeable trials
- two possible outcomes: failure or success
- $y$ success
- $\theta$ probability of success

$$p(y \mid \theta) = \text{Bin}(y \mid n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

# Bayesian inference

drawing conclusion from numerical data about quantities that are not observed

Unobserved quantities for which statistical inferences are made:

- $\tilde{y}$: potentially observable quantities such as future observation of a process
- $\theta$: quantities that are not directly observable such as parameters that govern the hypothetical process

**Bayesian statistical conclusions about a parameter $\theta$ or unobserved data $\tilde{y}$ are made in terms of *probability* statements. These are conditional on the observed values of $y$: $p(\theta \mid y)$**

# Bayesian inference

We want to obtain a distribution for $\theta$ conditioned to $y$: $p(\theta \mid y)$

1) we need a *model* ($M$) that provides us the joint probability distribution of $\theta$ and $y$ : $p(\theta, y)$

2) Given the $M$, we write $p(\theta, y) = p(y \mid \theta)p(\theta)$, with $p(\theta)$ *prior distribution* and $p(y \mid \theta)$ *sampling distribution*

3) we use the Bayes rule to condition on the known value of the data $y$

$$p(\theta, y) = p(\theta \mid y)p(y) = p(y \mid \theta)p(\theta)$$

$$p(\theta \mid y) \propto p(\theta)p(y \mid \theta) \text{ (}\textit{unnormalized posterior density}\text{)}$$

The data affect the posterior only through $p(y \mid \theta)$. Regarded as a function of $\theta$ and fixing $y$ this is the Likelihood function $\mathcal{L}(\theta) = p(y \mid \theta)$

# example 1

**man** X-chromosome, Y-chromosome
**woman** X-chromosome, X-chromosome

Hemophilia: hereditary disease associated to a gene of the chromosome X

This is recessive inheritance: a man who inherits the gene is affected, a woman who inherits the gene on only one X is not affected

# example 1

A woman has an affected brother and a father not affected ⇒ she can be a carrier of the gene on *one* X. Is she a carrier?

*Parameter we want to estimate*

$\theta = 1$ carrier

$\theta = 0$ not a carrier

*Prior*: $p(\theta = 1) = p(\theta = 0) = 0.5$ (a priori we say fifty-fifty)

*Data*: she has two sons, neither of the two affected, $y_1 = 0,\ y_2 = 0$

*Likelihood*:

$p(y_1 = 0, y_2 = 0 \,|\, \theta = 1) = 0.5^2$

$p(y_1 = 0, y_2 = 0 \,|\, \theta = 0) = 1$

*Posterior*:

$$p(\theta = 1 \,|\, y_1, y_2) = \frac{p(y_1, y_2 \,|\, \theta = 1)p(\theta = 1)}{p(y_1, y_2 \,|\, \theta = 1)p(\theta = 1) + p(y_1, y_2 \,|\, \theta = 0)p(\theta = 0)} = \frac{0.25\ 0.5}{0.25\ 0.5 + 0.5} = 0.2$$

# example 1

## adding more data

*More data*: she has a third son, not affected, $y_3 = 0$

P*rior*: $p(\theta = 1) = 0.2 \quad p(\theta = 0) = 0.8$ (the posterior of before)

*Likelihood*:

$p(y_3 = 0 \,|\, \theta = 1) = 0.5$

*Posterior*:

$$p(\theta = 1 \,|\, y_3) = \frac{p(y_3 \,|\, \theta = 1)p(\theta = 1)}{p(y_3 \,|\, \theta = 1)p(\theta = 1) + p(y_3 \,|\, \theta = 0)p(\theta = 0)} = \frac{0.5\ 0.2}{0.5\ 0.2 + 0.8} = 0.111$$

# example 2

## Bernoulli trials

Ingredients:

- $n$ exchangeable trials
- two possible outcomes: failure or success
- $y$ success
- $\theta$ probability of success

# example 2

Bernoulli trial

P*rior*: uniform in [0,1]

*Likelihood*:

$$p(y\,|\,\theta) = \text{Bin}(y\,|\,n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

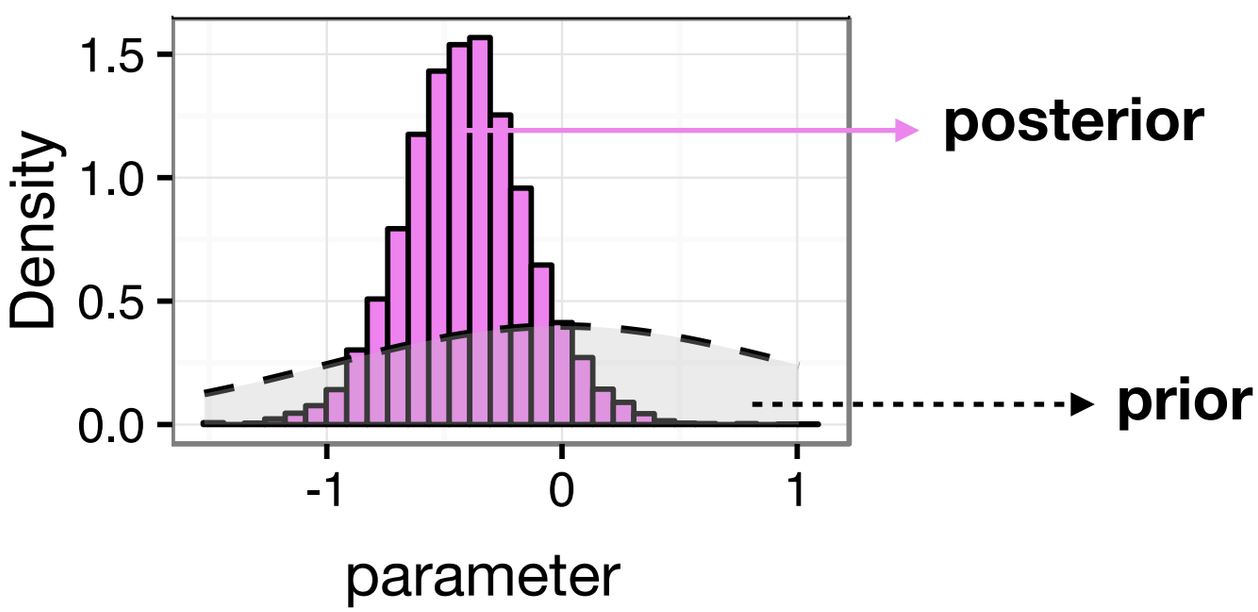
(we don't write the dependence on $n$ on the left side because is part of the experimental design and considered fixed. All probabilities will be conditional on $n$)

*Posterior:*

$p(\theta\,|\,y) \propto \theta^y(1-\theta)^{n-y}$ → Beta distribution: $\theta\,|\,y \sim \text{Beta}(y+1, n-y+1)$

$\binom{n}{y}$ treated does not depend on $\theta$ so it can be disregarded

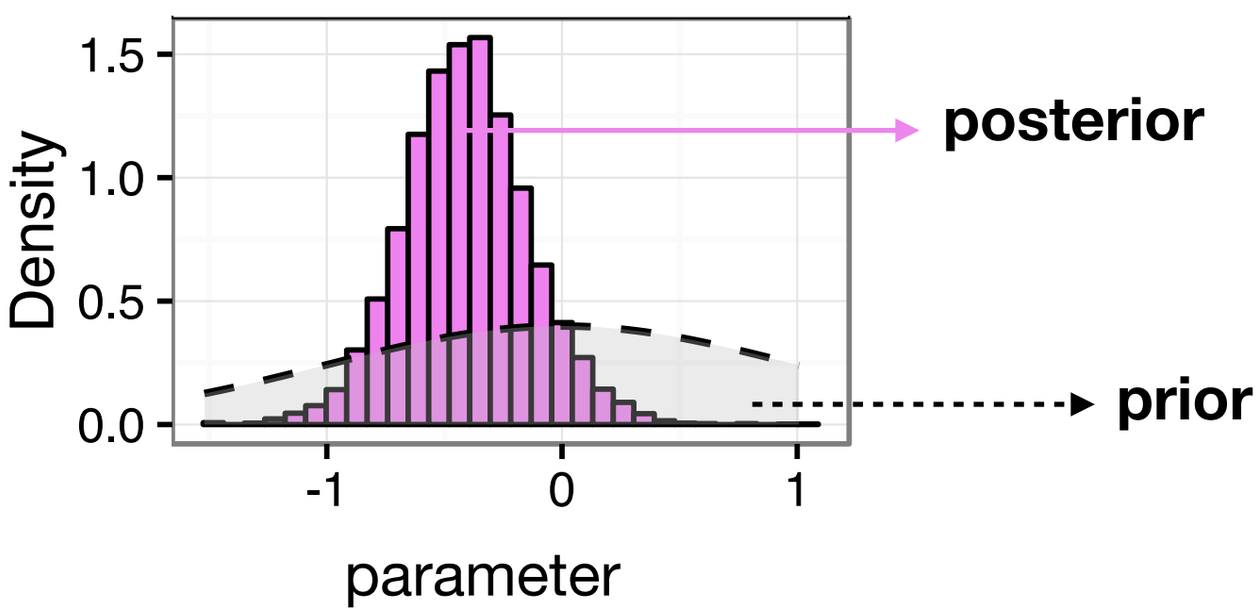
# Bayesian analysis: some thoughts




prior distribution summarises my a priori knowledge about the parameters. E.g. it may be defined based on the literature (e.g. if we are analysing an outbreak of flu and we want the estimate $R_0$ we may look at previous $R_0$ estimates).

If no priory knowledge is available, the best is to use a vague, or flat, or *noninformative* prior

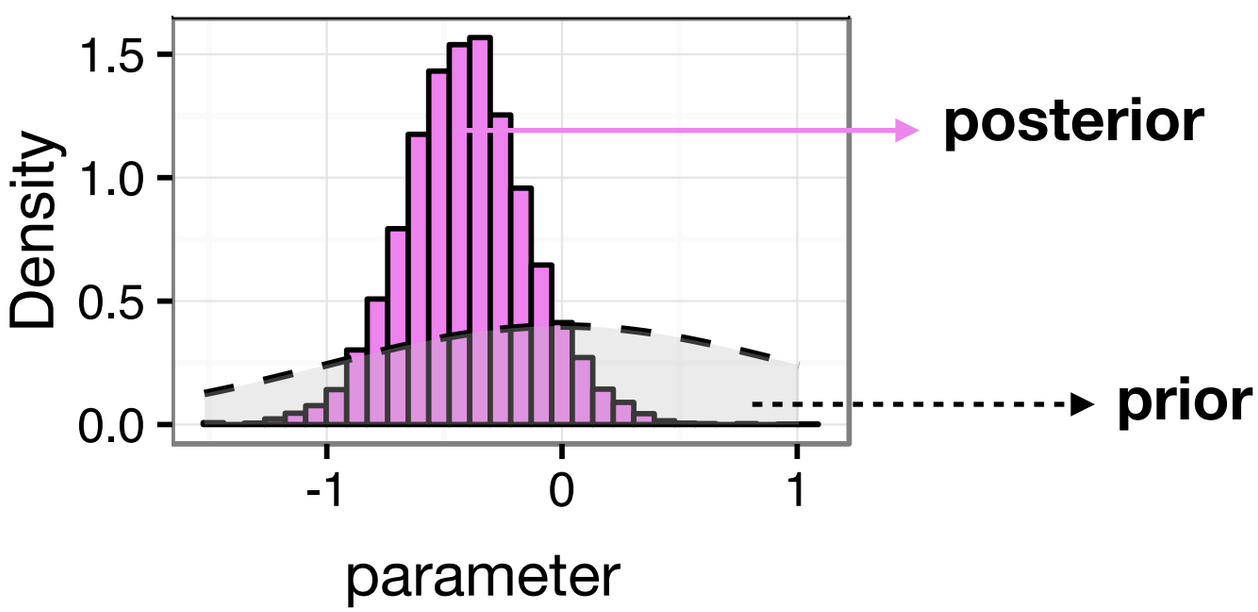
# Bayesian analysis: some thoughts




posterior distribution is a compromise between data and prior information. The compromise is increasingly controlled by the data as the sample size increases

posterior variance on average smaller than prior variance (it can be larger, but often this indicates a conflict or inconsistency between the sampling model and the prior distribution)

# Bayesian analysis: some thoughts



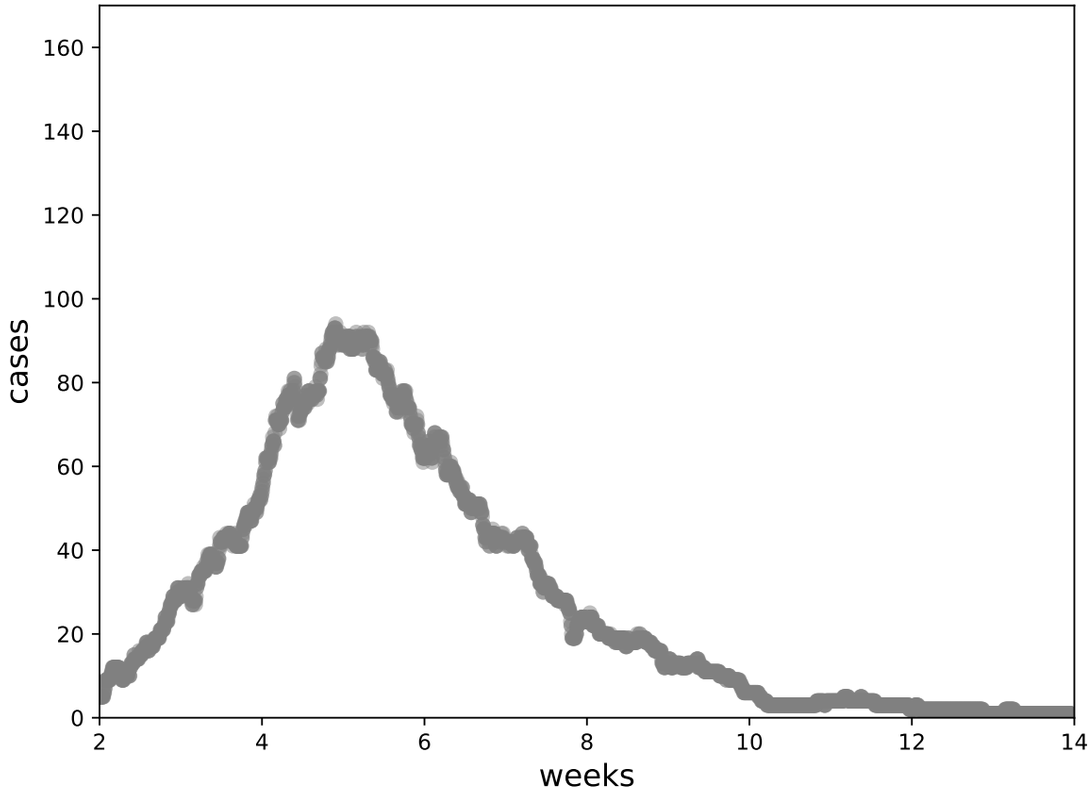

posterior contain the whole information.

key information we are interest in:

- the most likely parameter value given the data: the mode of the posterior
- the uncertainty associated to our estimate: quantiles of the posterior, i.e. $95\,\%$ interval, that goes from $2.5\,\%$ to $97.5\,\%$

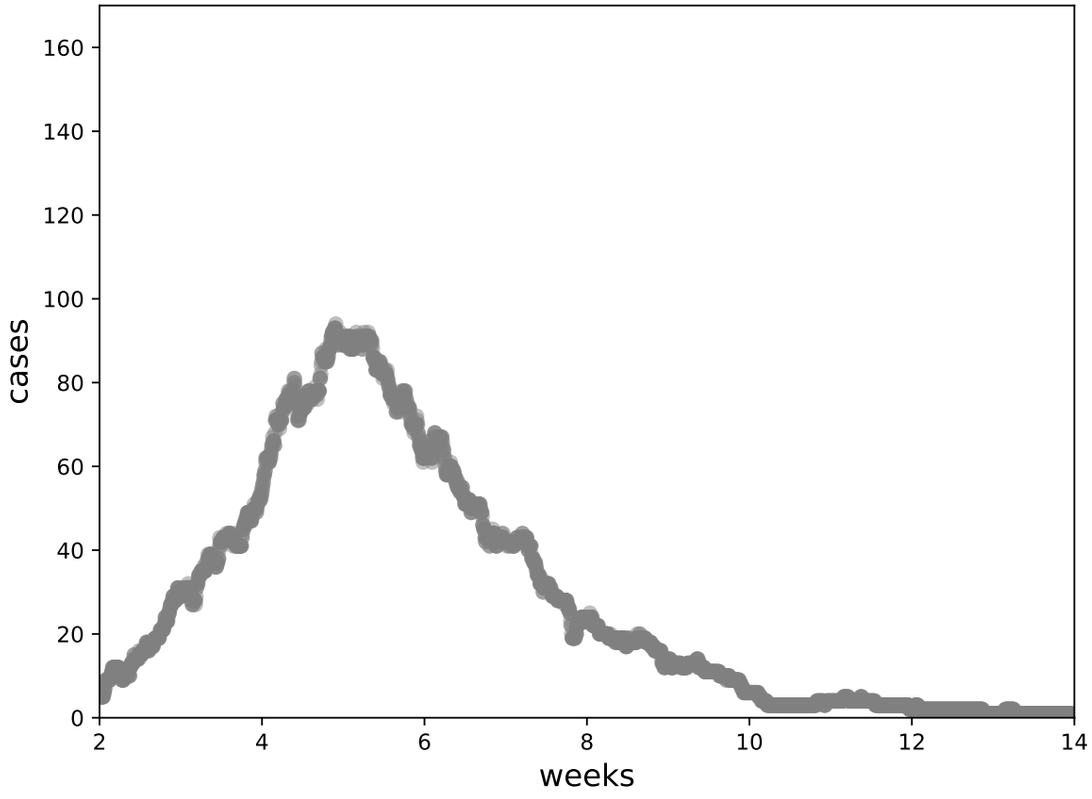
# fitting an incidence curve

**data:** cases are detected independently each day with probability $d = 50\,\%$.
Independent observations $y_t$



# fitting an incidence curve

**data:** cases are detected independently each day with probability $d = 50\,\%$.
Independent observations $y_t$
**model:** SIR, we know the average infections duration $\mu^{-1} = 5.5$ days
**parameter:** $\theta = \beta$
**prior**: uniform in [0,1]
**Likelihood**: $\mathscr{L}(\theta) = p(y_1, \ldots, y_t, \ldots, y_{t_M} | \theta) = \prod_t p(y_t | \theta)$



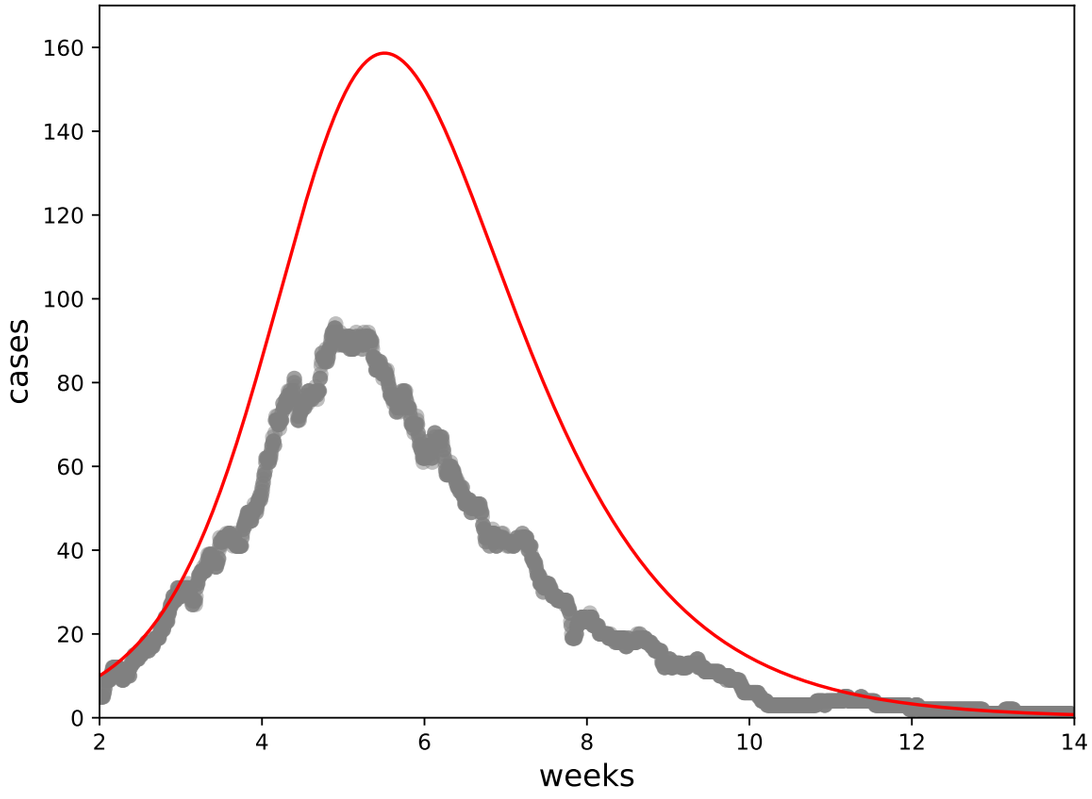
# fitting an incidence curve

**In practice:**

for each value of $\beta$ we simulate the trajectory of the SIR, fixing $\mu$ and $I_0$ based on available knowledge

Observation model: observed data are distributed as a Poisson

$y_t \,|\, \theta \sim \text{Poisson}(\lambda); \quad \lambda = I(t)\, d$



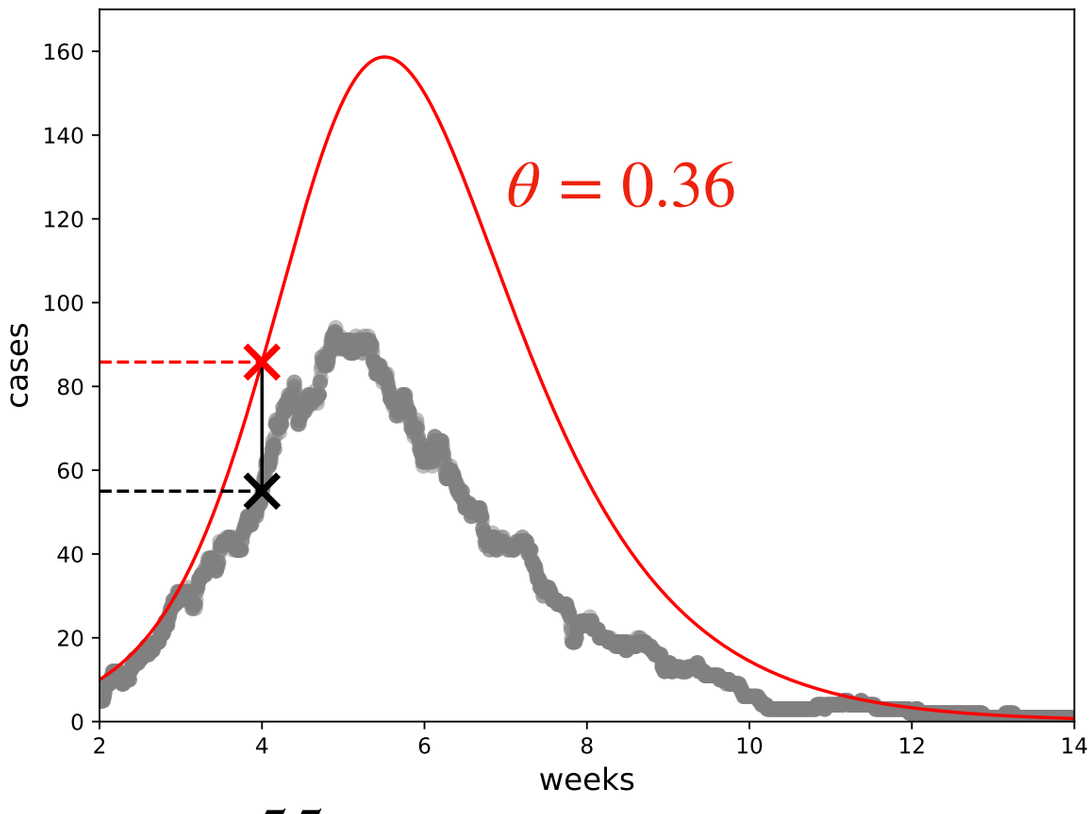
# fitting an incidence curve

**In practice:**

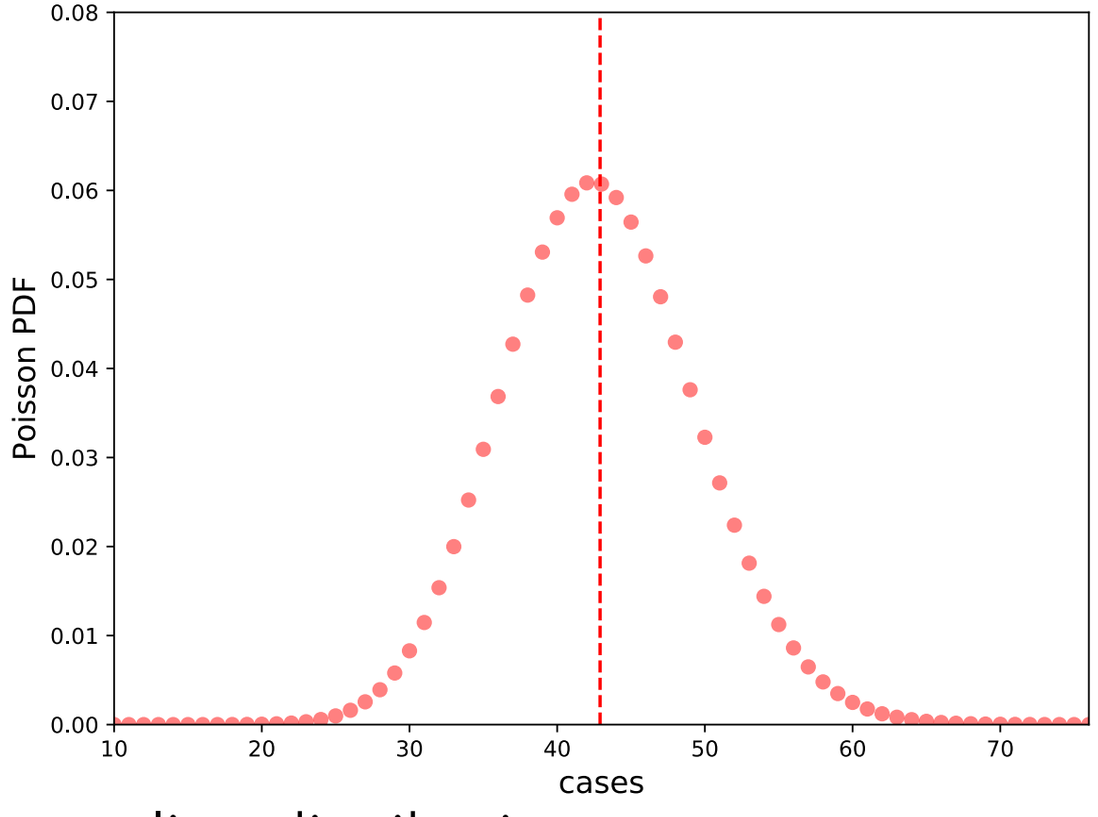
for each value of $\beta$ we simulate the trajectory of the SIR, fixing $\mu$ and $I_0$ based on available knowledge

Observation model : observed data are distributed as a Poisson

$y_t | \theta \sim \text{Poisson}(\lambda); \quad \lambda = I(t)\, d$




$y_{t=4} = 55$




sampling distribution

$y_{t=4} | \theta \sim \text{Poisson}(\lambda); \quad \lambda = I(t)\, \lambda = 86\ 0.5$

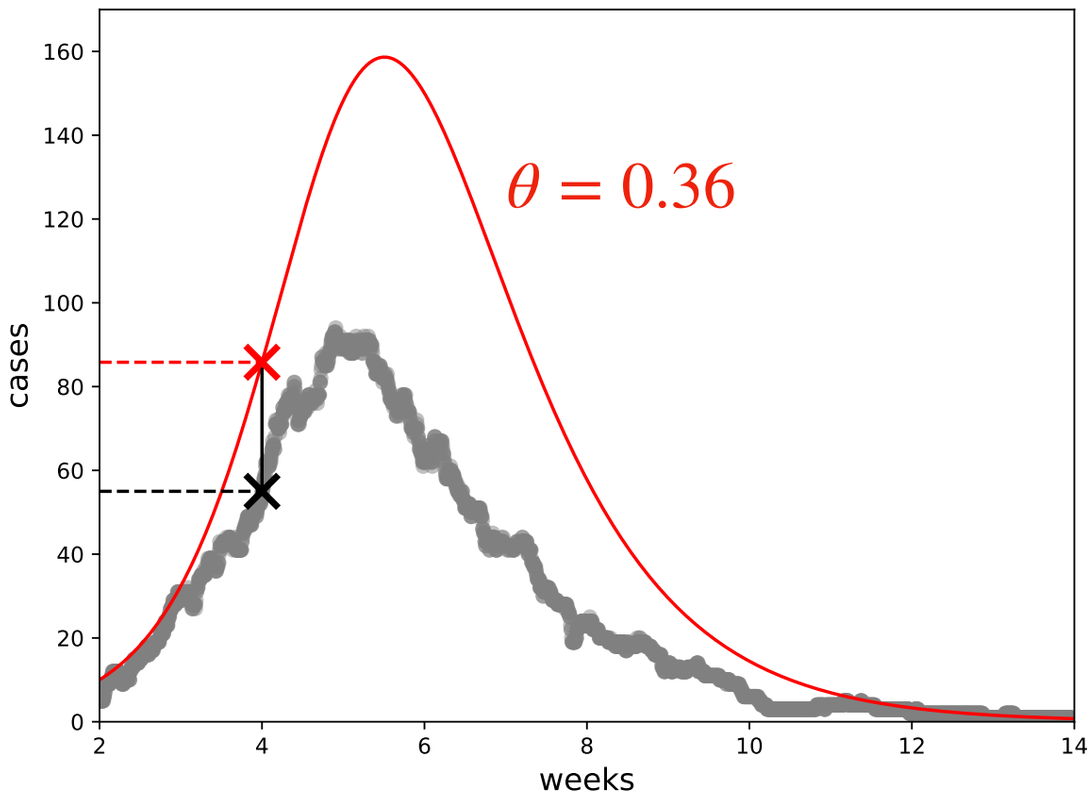
# fitting an incidence curve

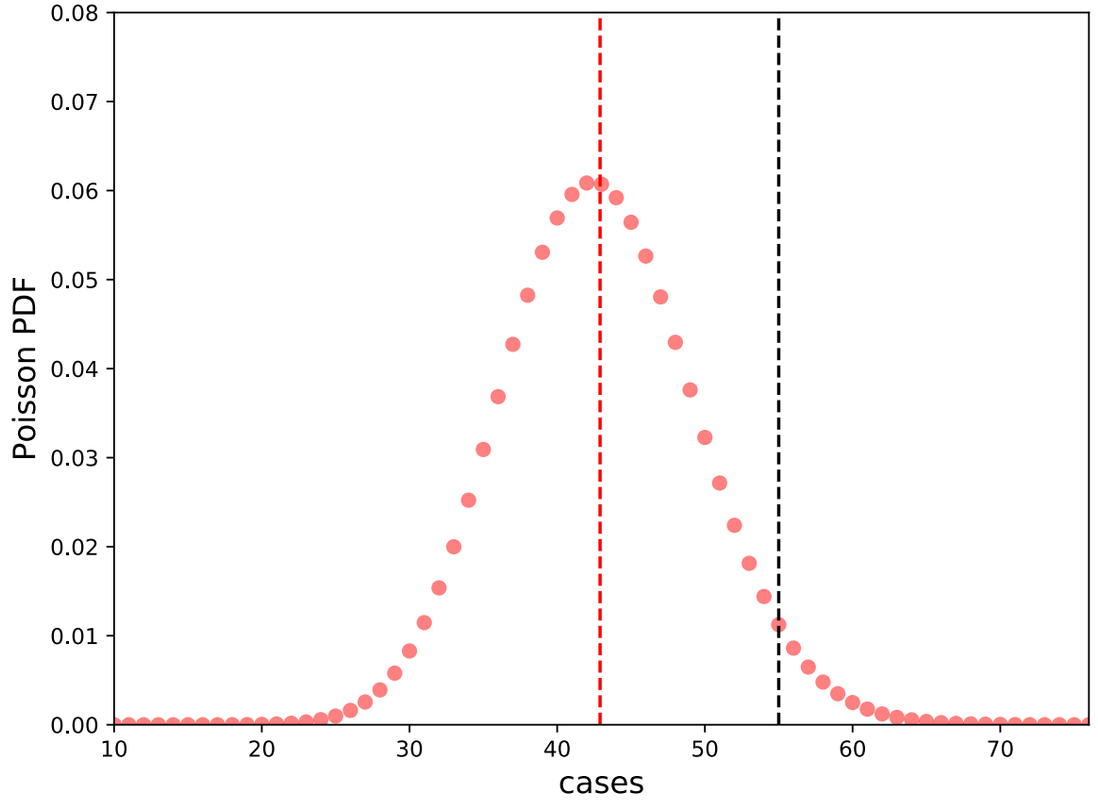
**In practice:**

for each value of $\beta$ we simulate the trajectory of the SIR, fixing $\mu$ and $I_0$ based on available knowledge

Observation model : observed data are distributed as a Poisson

$y_t \,|\, \theta \sim \text{Poisson}(\lambda); \quad \lambda = I(t)\, d$




$\mathscr{L}(\theta = 0.36) = \text{Poisson}(55 \,|\, \lambda); \quad \lambda = 86\ 0.5$

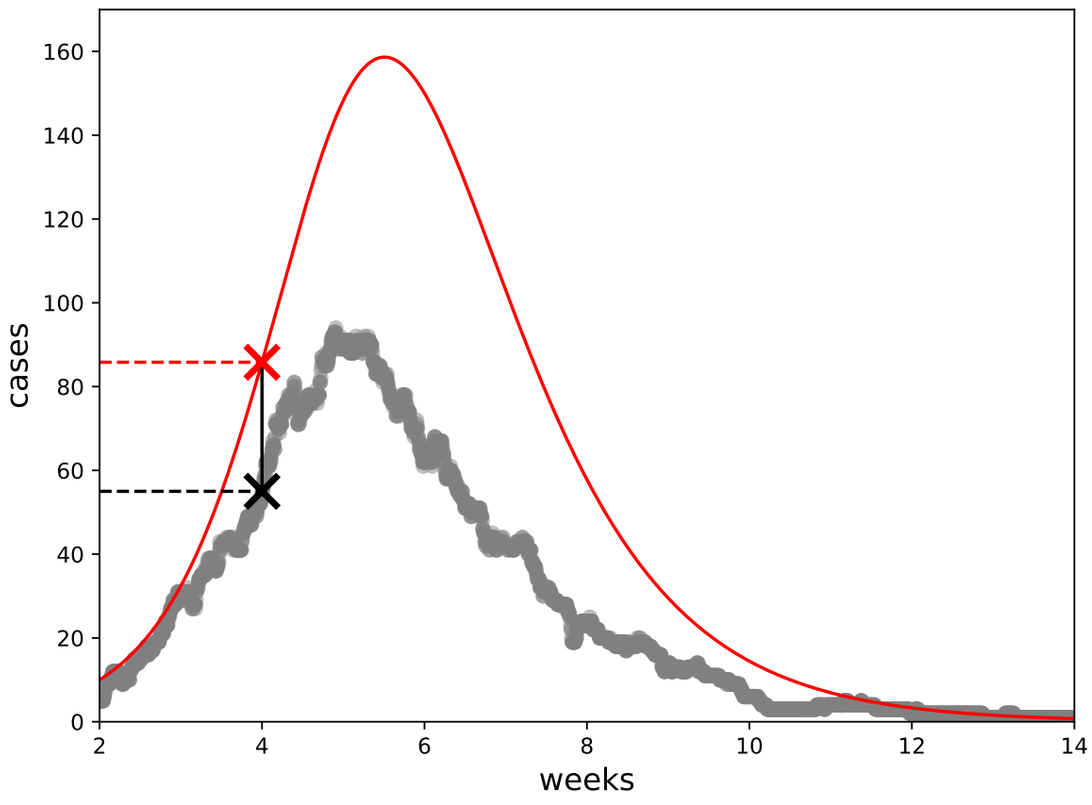
# fitting an incidence curve

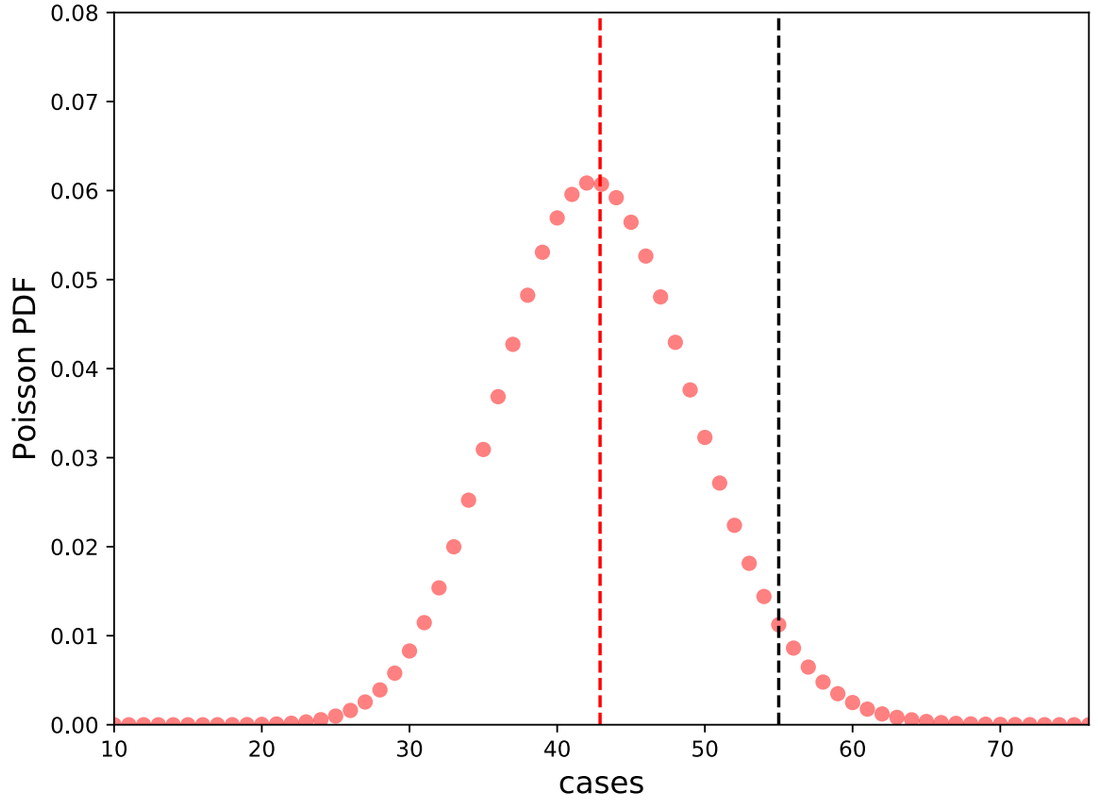
**In practice:**

for each value of $\beta$ we simulate the trajectory of the SIR, fixing $\mu$ and $I_0$ based on available knowledge

Observation model : observed data are distributed as a Poisson

$y_t \,|\, \theta \sim \text{Poisson}(\lambda); \quad \lambda = I(t)\, d$




$$\mathcal{L}(\theta) = p(y_1, \ldots, y_t, \ldots, y_{t_M} \,|\, \theta) = \prod_t p(y_t \,|\, \theta)$$

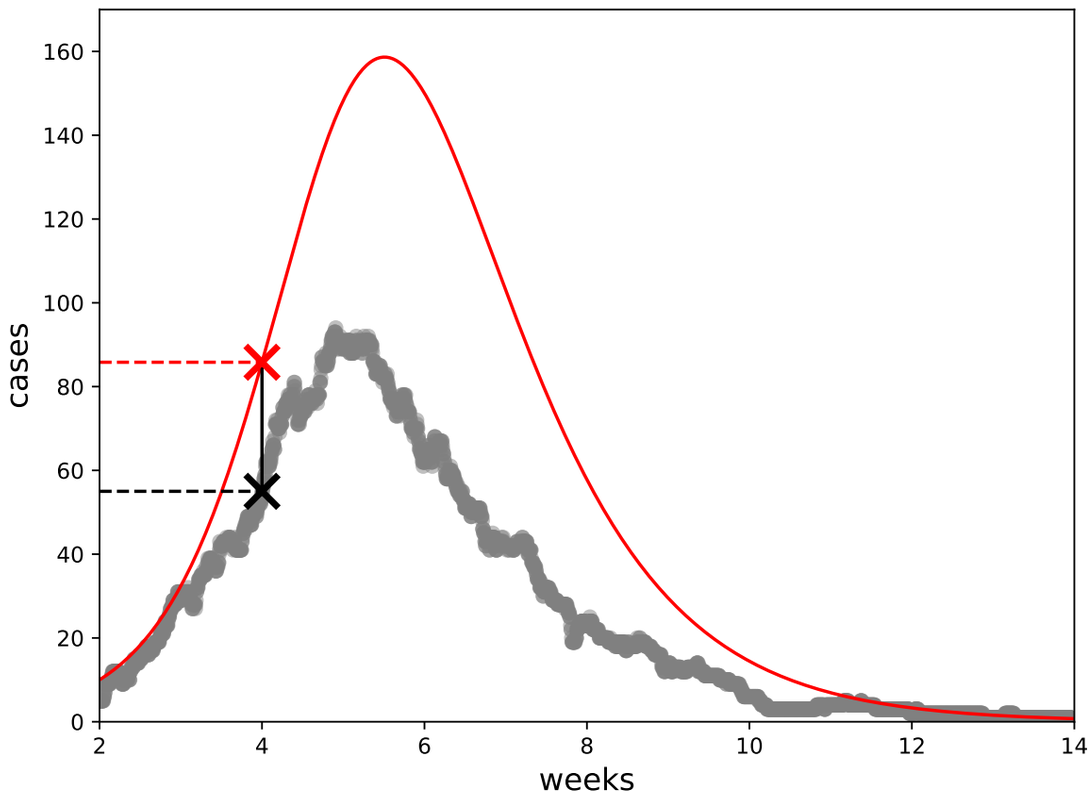
# fitting an incidence curve

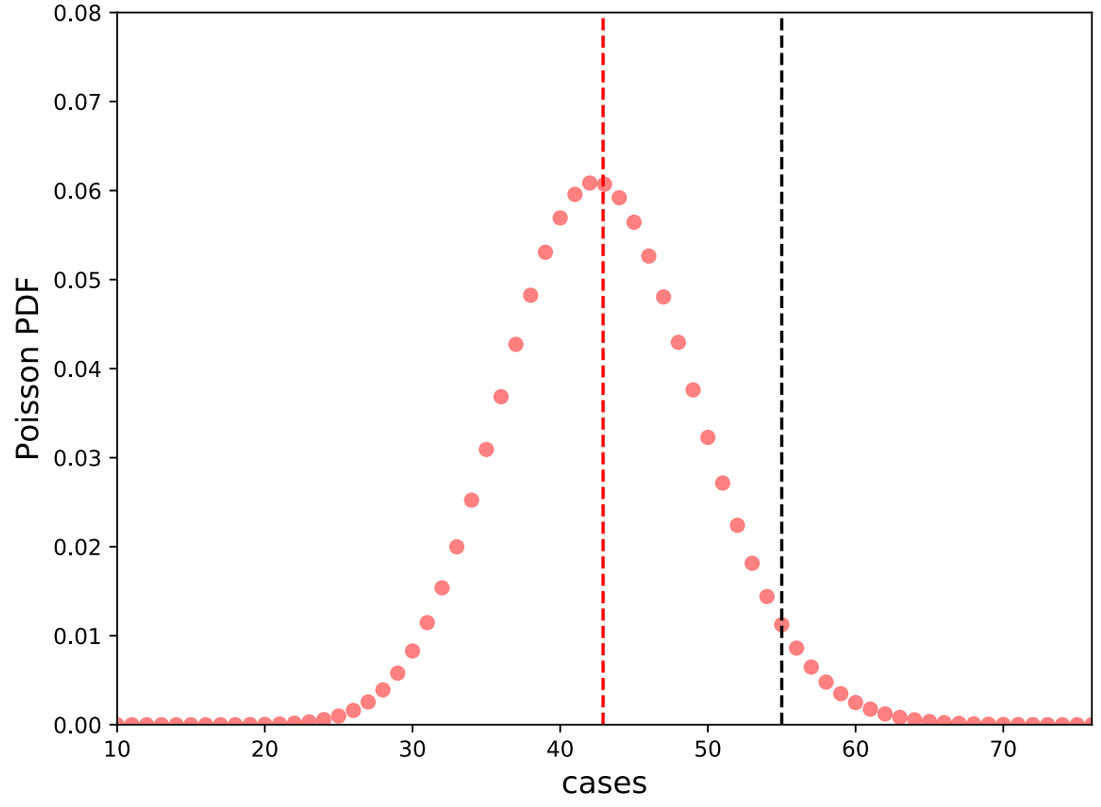
**In practice:**

for each value of $\beta$ we simulate the trajectory of the SIR, fixing $\mu$ and $I_0$ based on available knowledge

Observation model : observed data are distributed as a Poisson

$y_t \,|\, \theta \sim \text{Poisson}(\lambda); \quad \lambda = I(t)\, d$



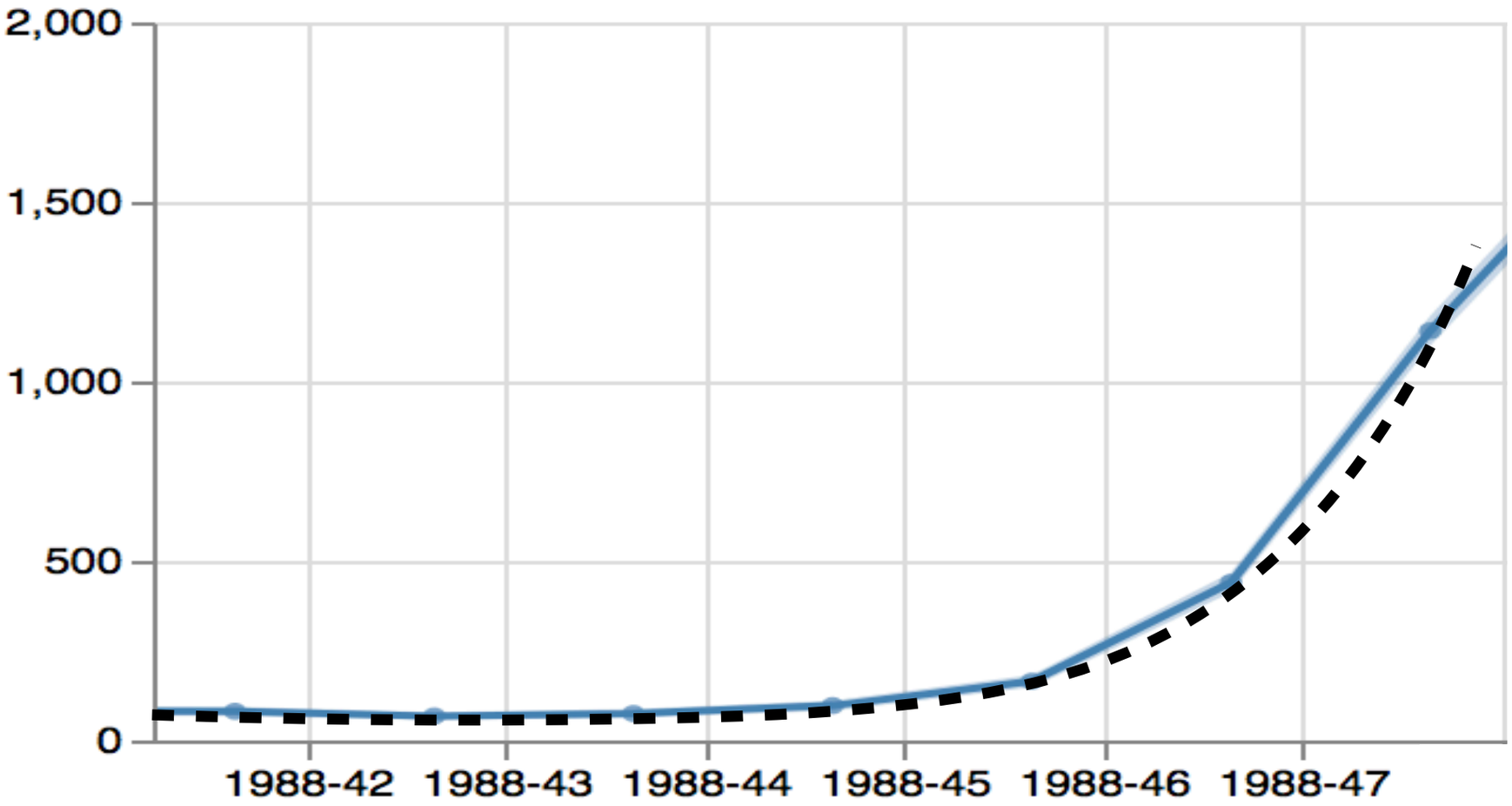




$$\log \mathscr{L}(\theta) = \sum_t \log p(y_t \,|\, \theta)$$

# recap

- Basic idea behind likelihood computation: evaluate the probability of the data given the model and the parameters
- To estimate $\theta$ we keep $M$ and $x_0$ fixed and vary $\theta$ to compute the probability of $p(y \mid \theta)$
- Likelihood function: $\mathscr{L}(\theta) = p(y \mid \theta)$
- likelihoods can span a wide range of orders of magnitude, which can lead to numerical problems. In practice better to work with the log-likelihood $\log \mathscr{L}(\theta) = \log p(y_1, \ldots, y_n \mid \theta) = \sum_i \log p(y_i \mid \theta)$

# likelihood computation in practice




Sentinelles
Réseau Sentinelles

https://www.sentiweb.fr/

- in general, the posterior distribution is difficult to obtain
- it requires numerical integration

# likelihood computation in practice

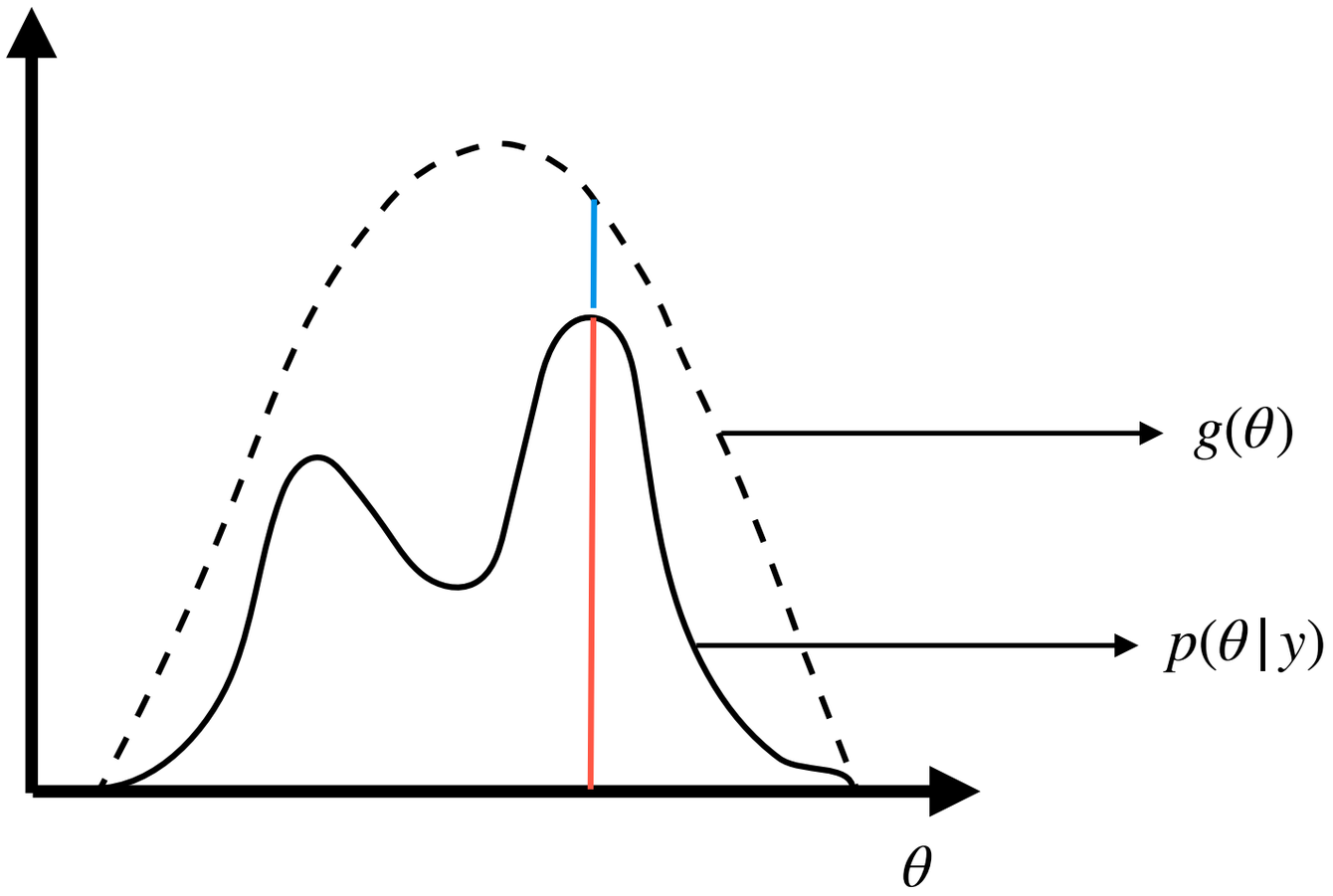
## basic numerical methods

- **Grid of points:** compute $p(\theta \mid y)$ for $\theta_1, \ldots, \theta_n$ equally spaced. We approximate the continuous density function $p(\theta \mid y)$ with the discrete density function $p(\theta_i \mid y) / \sum_i p(\theta_i \mid y)$

  Limitations: this become rapidly unfeasible as the dimensionality of the parameter space increases

- **Trapezoidal approximation:** after computing $p(\theta \mid y)$ for a discrete set of points $\theta_1, \ldots, \theta_{n'}$ we can approximate $p(\theta \mid y)$ with a piecewise-linear function, connecting the $p(\theta_i \mid y)$ points with liner segments.

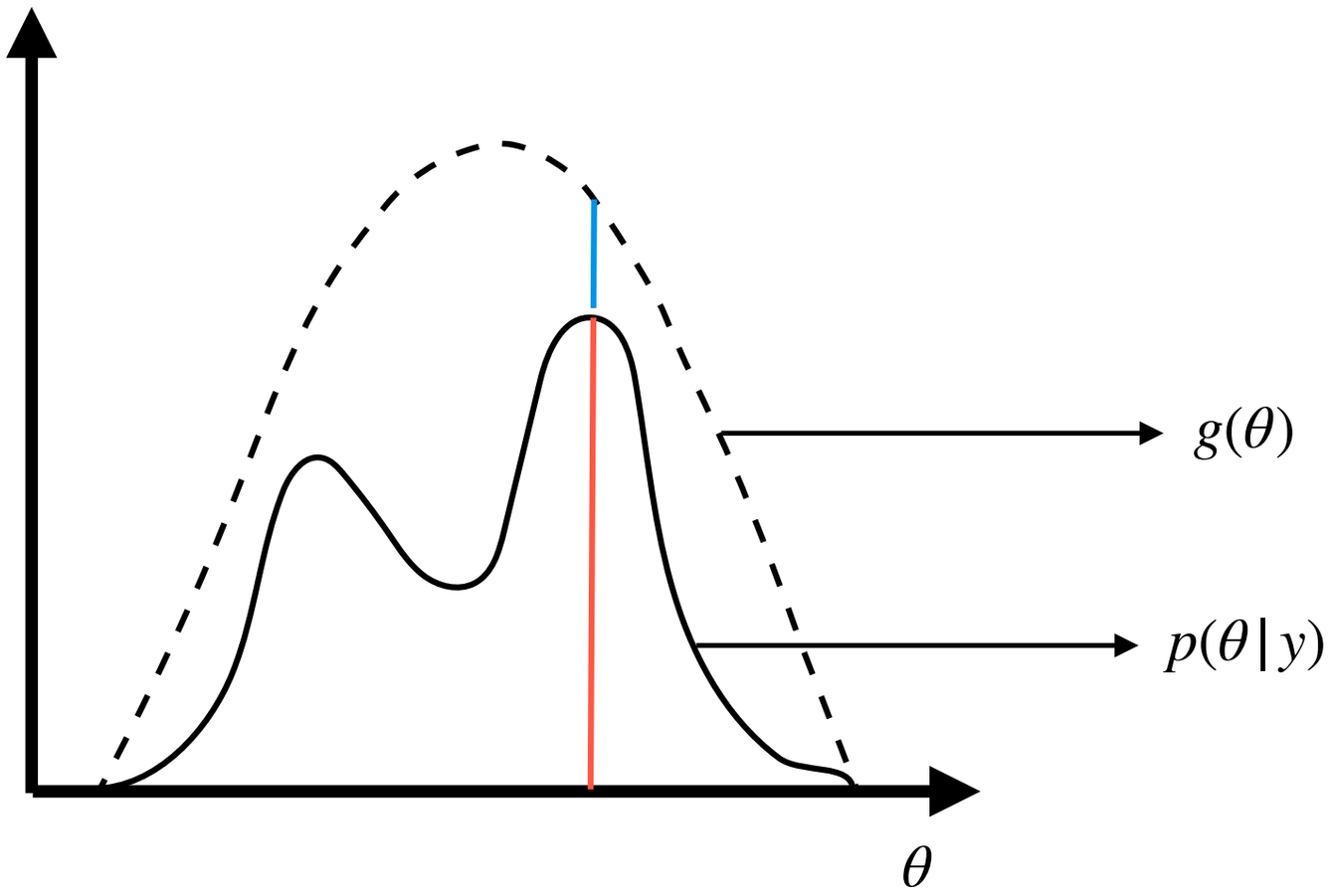
# rejection sampling




we want to numerically sample $p(\theta|y)$. We use a positive function $g(\theta)$ defined for all $\theta$ for which $p(\theta|y) > 0$, such that:

- we are able to draw sample from $g(\theta)$
- the importance ratio $p(\theta|y)/g(\theta) \leq M$, for some constant M
- $g(\theta)$ must have a finite integral

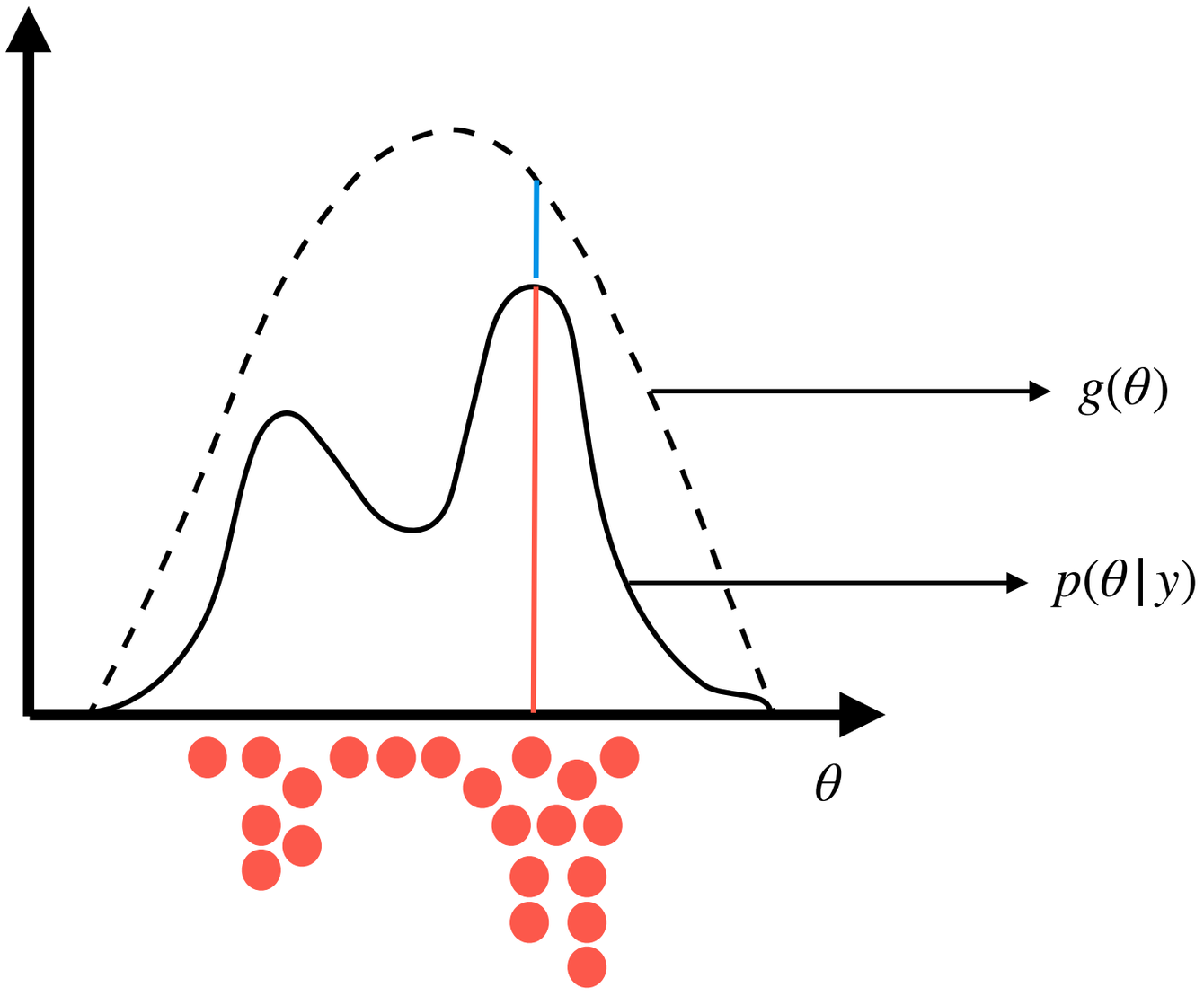
# rejection sampling




I iterate the two following steps:

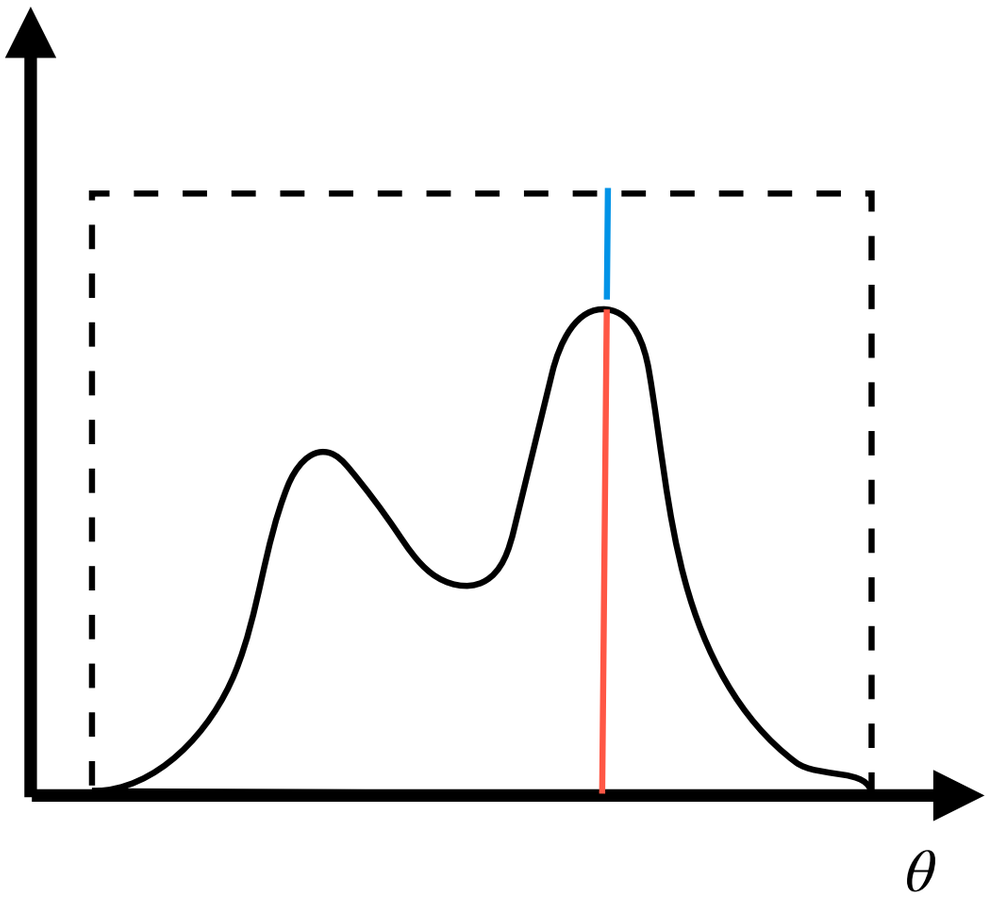
- sampling $\theta$ at random from the probability density proportional to $g(\theta)$
- with probability $\frac{p(\theta|y)}{Mg(\theta)}$ accept $\theta$ as draw from $p(\theta|y)$

# rejection sampling

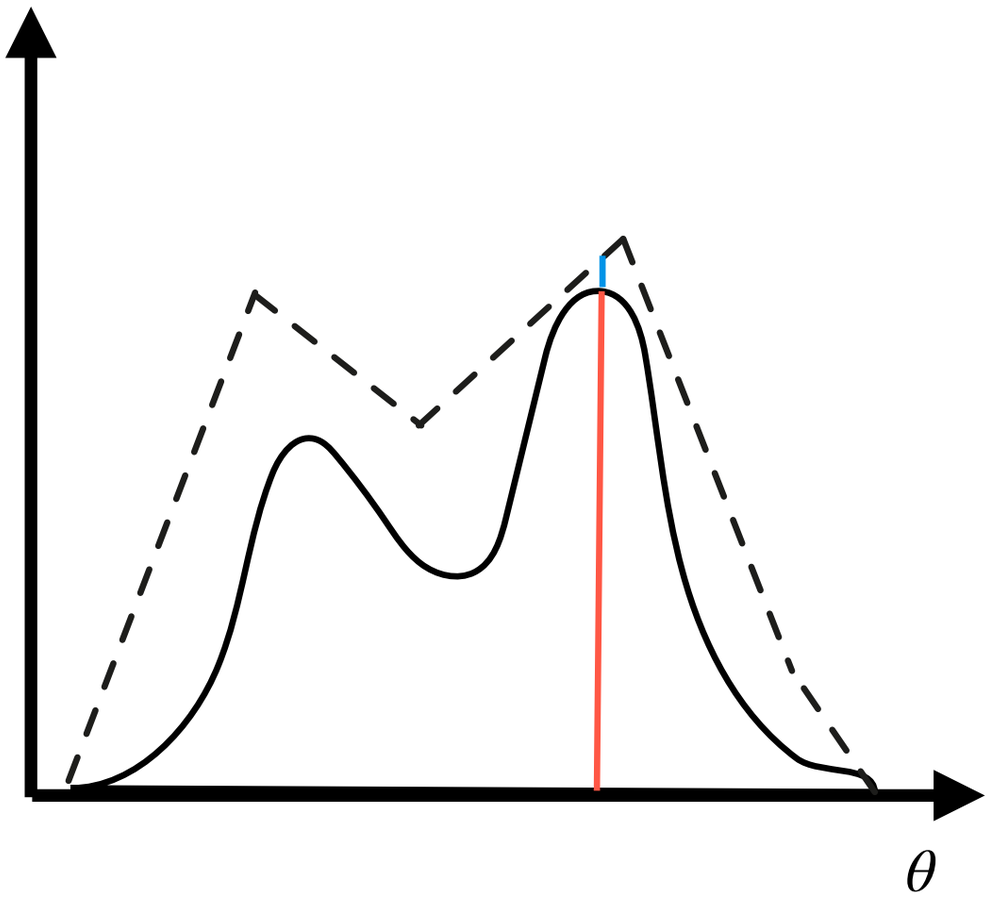


# rejection sampling: limitations

choice of $g(\theta)$ difficult




A simple choice with no a priory knowledge of $p(\theta \mid y)$ yields high rejection rate (a lot of computation)




use trapezoidal approximation to define $g(\theta)$

# sampling from a distribution

with a computer we can easily draw random numbers uniformly distributed in [0,1]. How to sample with a distribution $p(v)$?

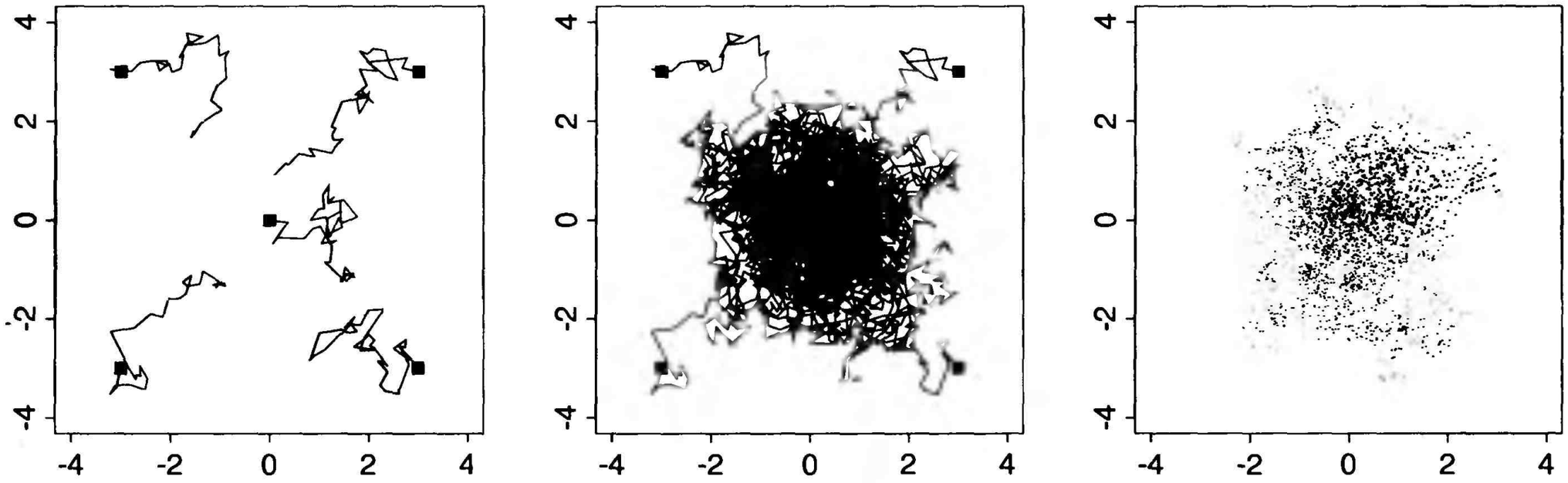
- $p(v)$ probability density function
- $F(v_*)$ cumulative density function : $F(v_*) = \Pr(v \leq v_*)$
- draw $U$ from a uniform distribution and compute $v = F^{-1}(U)$
- $v$ will be draw with probability $p(v)$

**Example:**

$p(v) = \beta e^{-\lambda v} \Rightarrow F(v) = 1 - e^{-\lambda v}$

$v = F^{-1}(U) = -\log(1 - U)/\lambda$

# Markov chain Monte Carlo




General idea:

I star from $\theta_0$ and create a random walk: a sequence $\theta_0, \theta_1, \ldots, \theta_t$ where each $\theta_t$ is drawn from a given *transition distribution*, built such that the random walk converges to $p(\theta \mid y)$

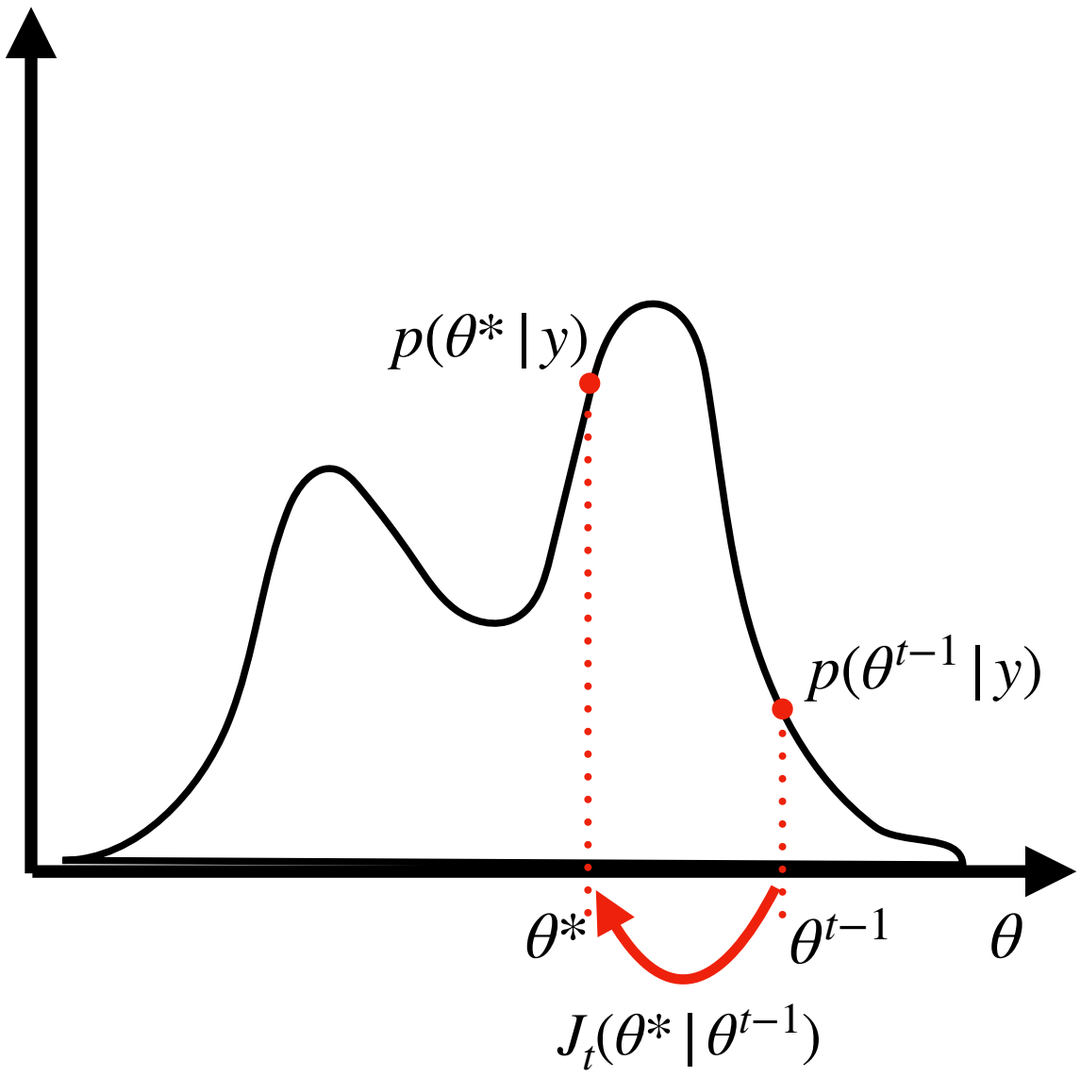
run the simulation long enough that the distribution of the current draws is close enough to the stationary distribution

# Metropolis algorithm

1) draw a starting point $\theta_0$ from a starting distribution

2) for $t = 1, 2, \ldots$:

   A. sample a candidate point $\theta^*$ from a jumping distribution $J_t(\theta^* \mid \theta^{t-1})$. *The distribution must be symmetric* $J_t(\theta_a \mid \theta_b) = J_t(\theta_b \mid \theta_a)$ *for all* $a, b, t$



# Metropolis algorithm

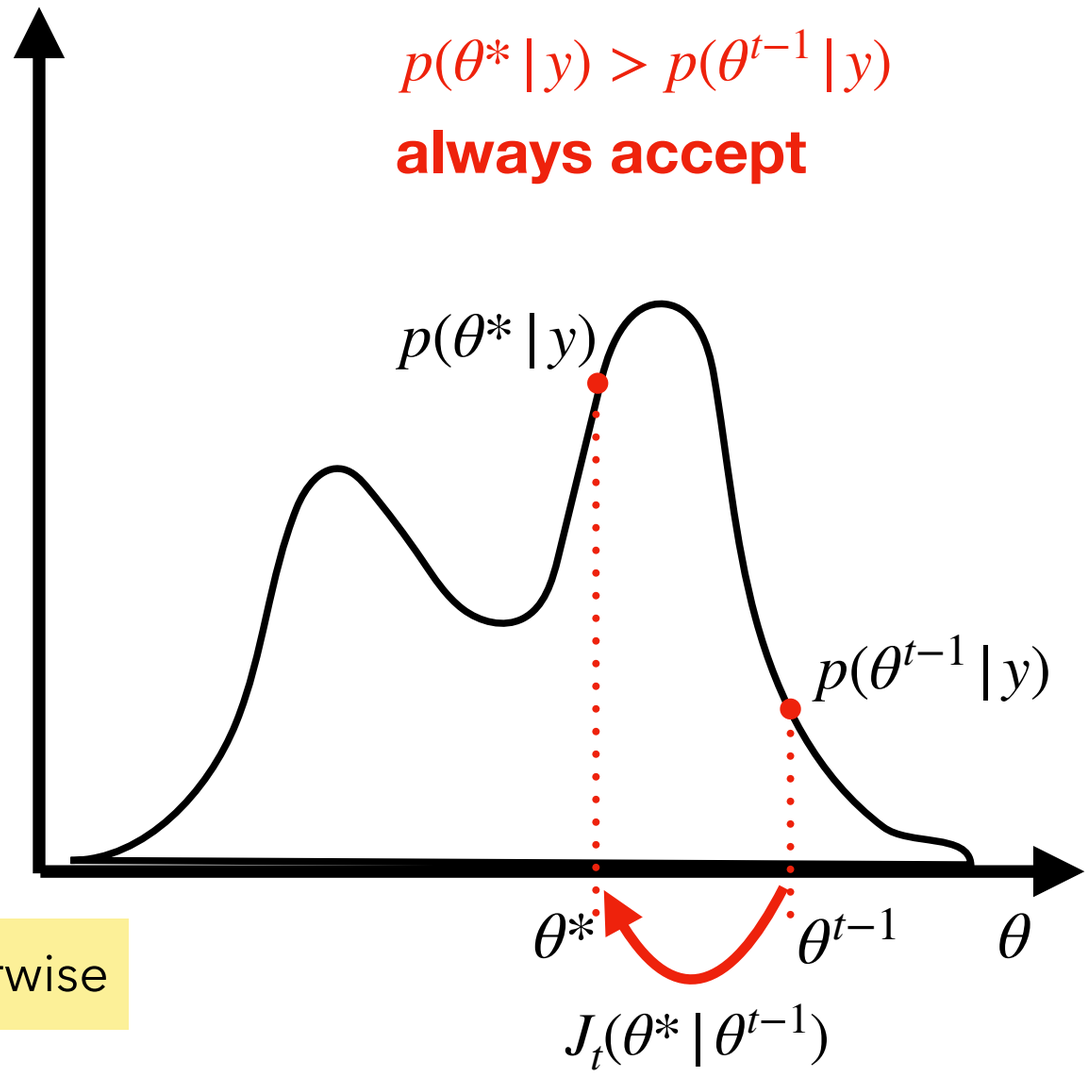
1) draw a starting point $\theta_0$ from a starting distribution

2) for $t = 1, 2, \ldots$:

   A. sample a candidate point $\theta^*$ from a jumping distribution $J_t(\theta^* \mid \theta^{t-1})$.
   *The distribution must be symmetric*
   $J_t(\theta_a \mid \theta_b) = J_t(\theta_b \mid \theta_a)$ *for all* $a, b, t$

   B. Calculate the ratio of the densities
   $$r = \frac{p(\theta^* \mid y)}{p(\theta^{t-1} \mid y)}$$

   C. set:
   $\theta^t = \theta^*$ with probability $\min(r,1)$; $\theta^{t-1}$ otherwise



# Metropolis algorithm

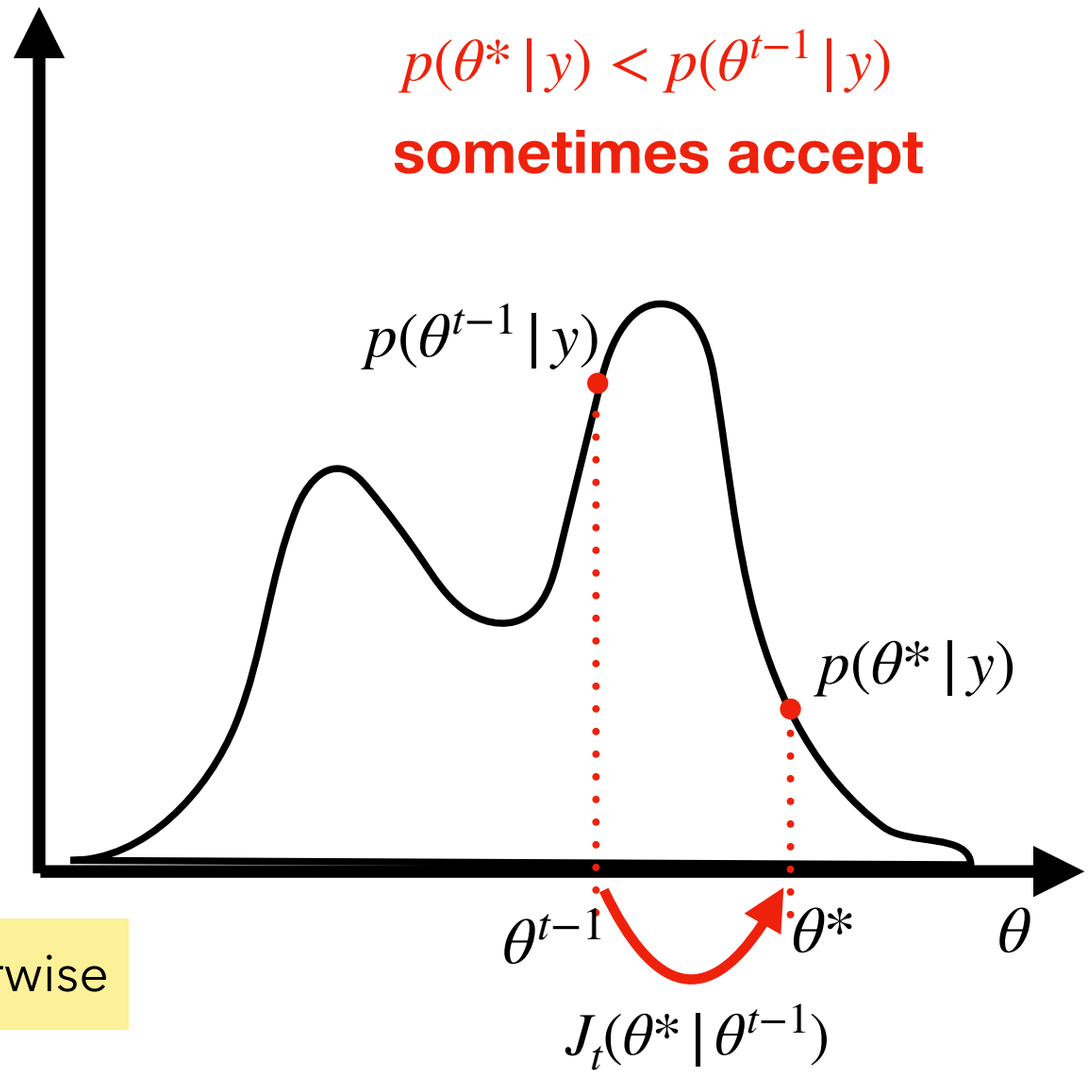
1) draw a starting point $\theta_0$ from a starting distribution

2) for $t = 1, 2, \ldots$:

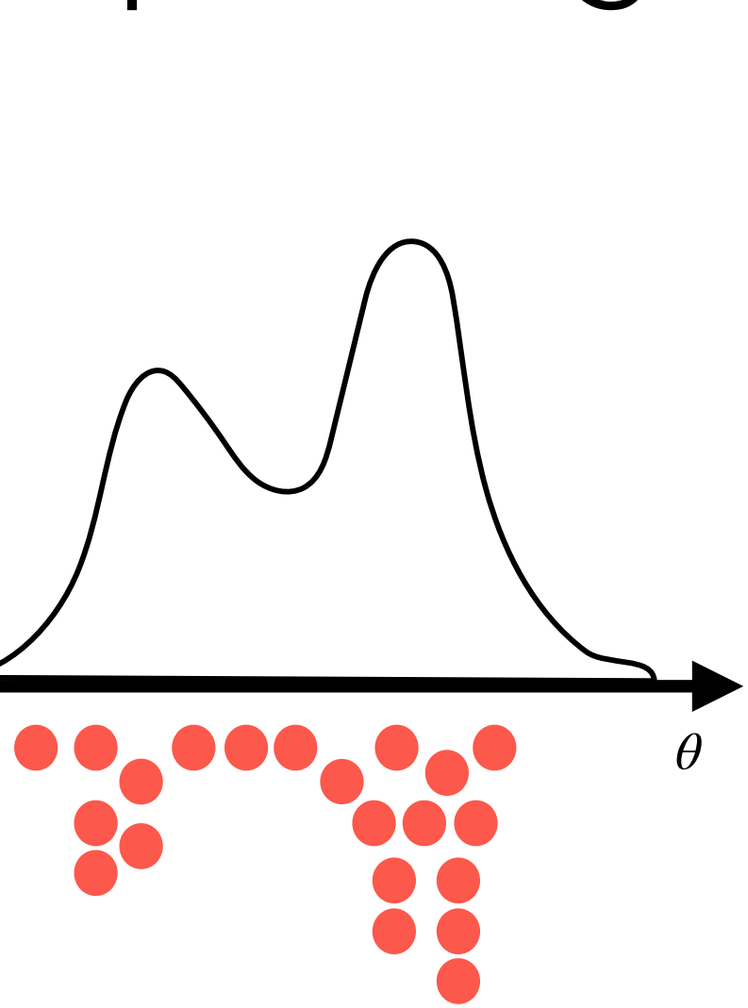
   A. sample a candidate point $\theta^*$ from a jumping distribution $J_t(\theta^* \mid \theta^{t-1})$.
   *The distribution must be symmetric*
   $J_t(\theta_a \mid \theta_b) = J_t(\theta_b \mid \theta_a)$ *for all* $a, b, t$

   B. Calculate the ratio of the densities
   $$r = \frac{p(\theta^* \mid y)}{p(\theta^{t-1} \mid y)}$$

   C. set:
   $\theta^t = \theta^*$ with probability $\min(r,1)$; $\theta^{t-1}$ otherwise



# Metropolis algorithm




Metropolis:

It is possible to show that

1) the Markov chain converges to a stationary distribution

2) the stationary distribution is $p(\theta \mid y)$

# Metropolis algorithm

**The Markov chain has a unique stationary distribution:**

The Markov chain is aperiodic, not transient irreducible. The random walk has a positive probability to eventually reach any state from any other state.

**The stationary distribution is $p(\theta \mid y)$:**

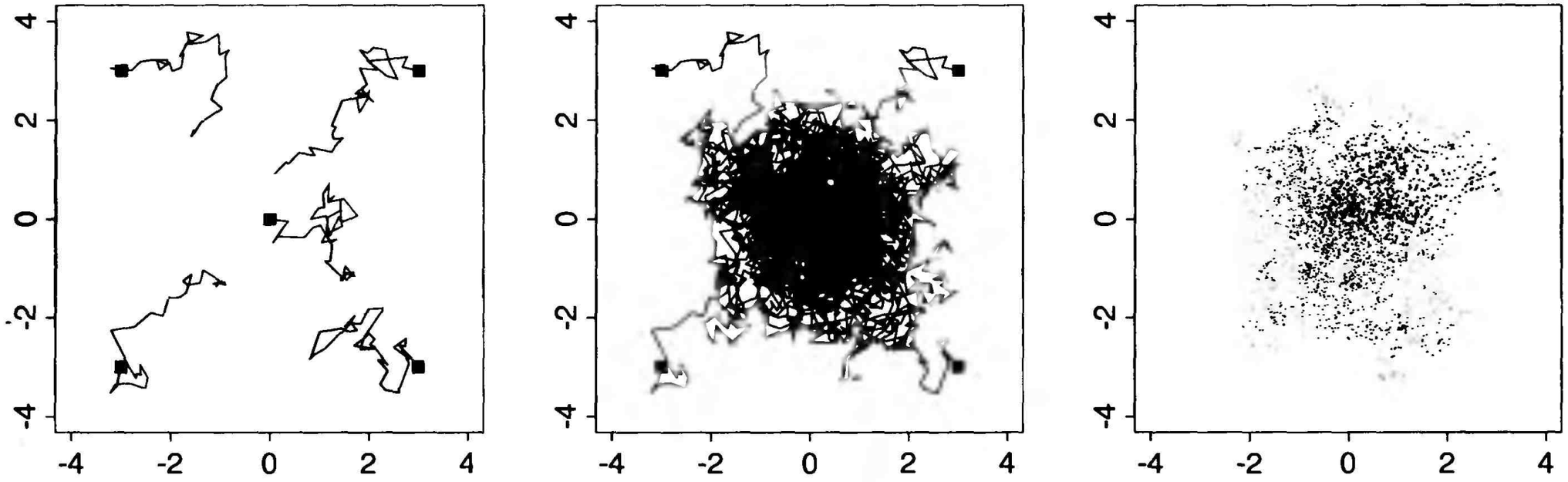
consider $\theta_a$ and $\theta_b$ such that $p(\theta_b \mid y) \geq p(\theta_a \mid y)$

$$p(\theta^{t-1} = \theta_a, \theta^t = \theta_b) = p(\theta_a \mid y)\, J_t(\theta_b \mid \theta_a)$$

$$\begin{aligned} p(\theta^t = \theta_a, \theta^{t-1} = \theta_b) &= p(\theta_b \mid y)\, J_t(\theta_a \mid \theta_b) \frac{p(\theta_a \mid y)}{p(\theta_b \mid y)} \\ &= p(\theta_a \mid y)\, J_t(\theta_a \mid \theta_b) \end{aligned}$$

Since their joint distribution is symmetric $\theta^t$ and $\theta^{t-1}$ have the same marginal distributions, thus $p(\theta \mid y)$ is stationary

# Metropolis algorithm

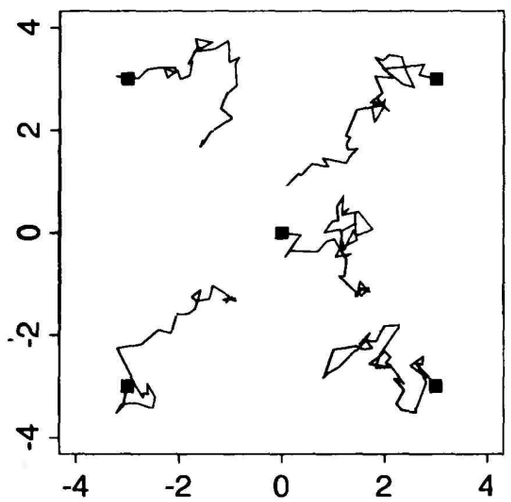


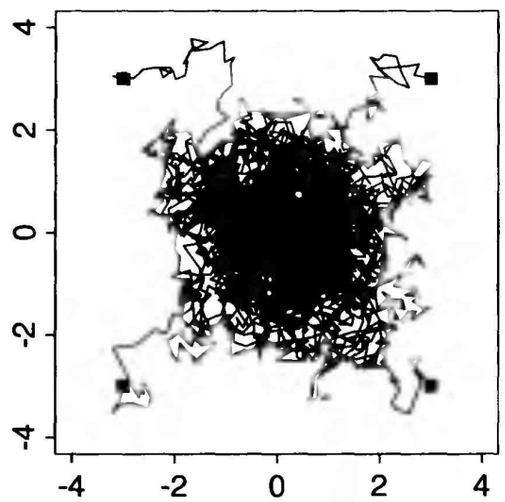

We simulate multiple chain simultaneously with starting point dispersed in the parameter space

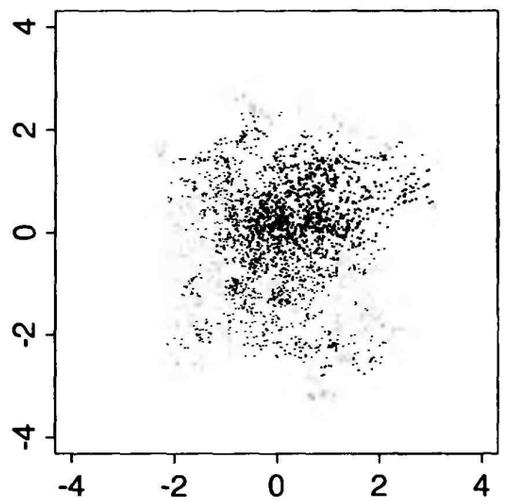
We monitor quantity of interest and measure variation between and within the different sequences, until “within” variation roughly equal “between” variation

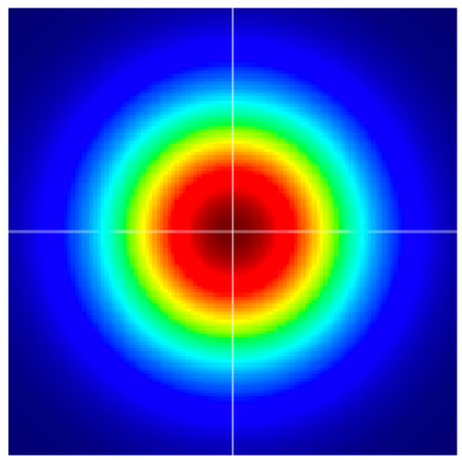
We want the distribution of each sequence = distribution all sequence mixed together

# Metropolis algorithm



consider the case of posterior density **bivariate unit normal**

$p(\theta_1, \theta_2 \,|\, y) = N(\theta_1, \theta_2 \,|\, 0, \mathbf{1})$ with $\mathbf{1}$ is the 2x2 identity matrix:

$$p(\theta_1, \theta_2 \,|\, y) = \frac{1}{2\pi} \exp\left[-\frac{\theta_1^2 + \theta_2^2}{2}\right]$$

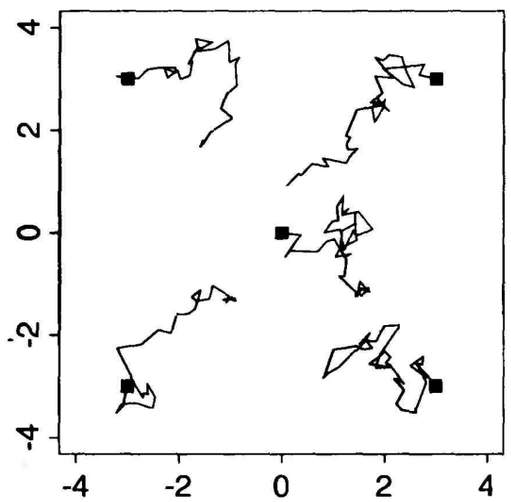
Convenient in this case to take the jumping distribution also bivariate normal

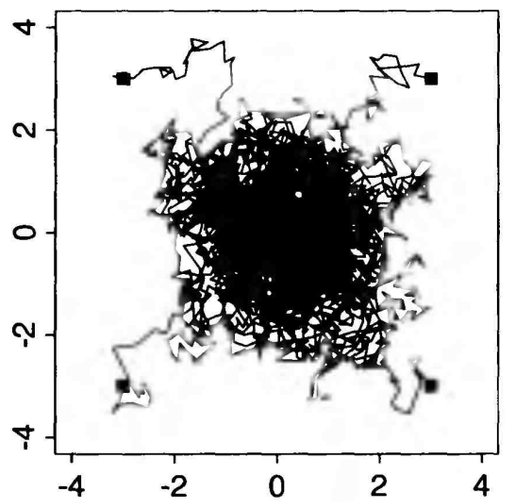
$J_t(\theta^* \,|\, \theta^{t-1}) = N(\theta^* \,|\, \theta^{t-1}, \sigma^2\, \mathbf{1})$:

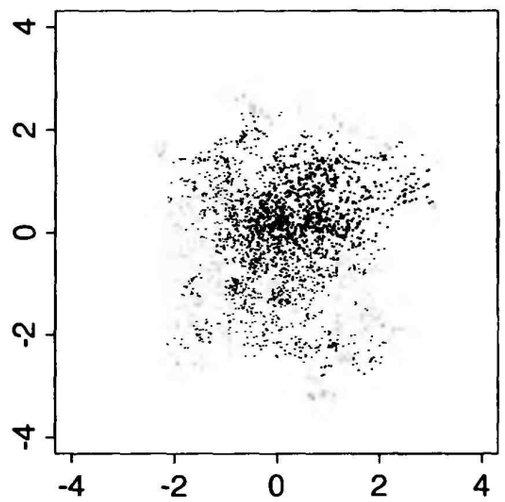
$$J(\theta_1^*, \theta_2^* \,|\, \theta_1^{t-1}, \theta_2^{t-1}) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(\theta_1^* - \theta_1^{t-1})^2 + (\theta_2^* - \theta_2^{t-1})^2}{2\sigma^2}\right]$$

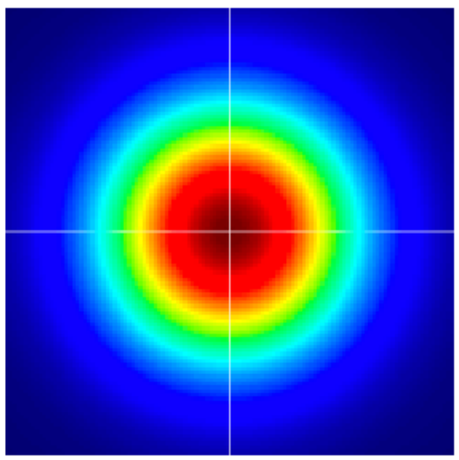
**(this is symmetric!)**

# Metropolis algorithm



$$J(\theta_1^*, \theta_2^* \mid \theta_1^{t-1}, \theta_2^{t-1}) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(\theta_1^* - \theta_1^{t-1})^2 + (\theta_2^* - \theta_2^{t-1})^2}{2\sigma^2}\right]$$

how do I choose $\sigma$?

- too big: I will go too far and I will reject almost all proposals and get stuck
- too small: the chain will be too slow and the algorithm will be inefficient
- rule of thumbs: acceptance rate from 0.23 to 0.44