Lecture 02:

Bayesian Methods

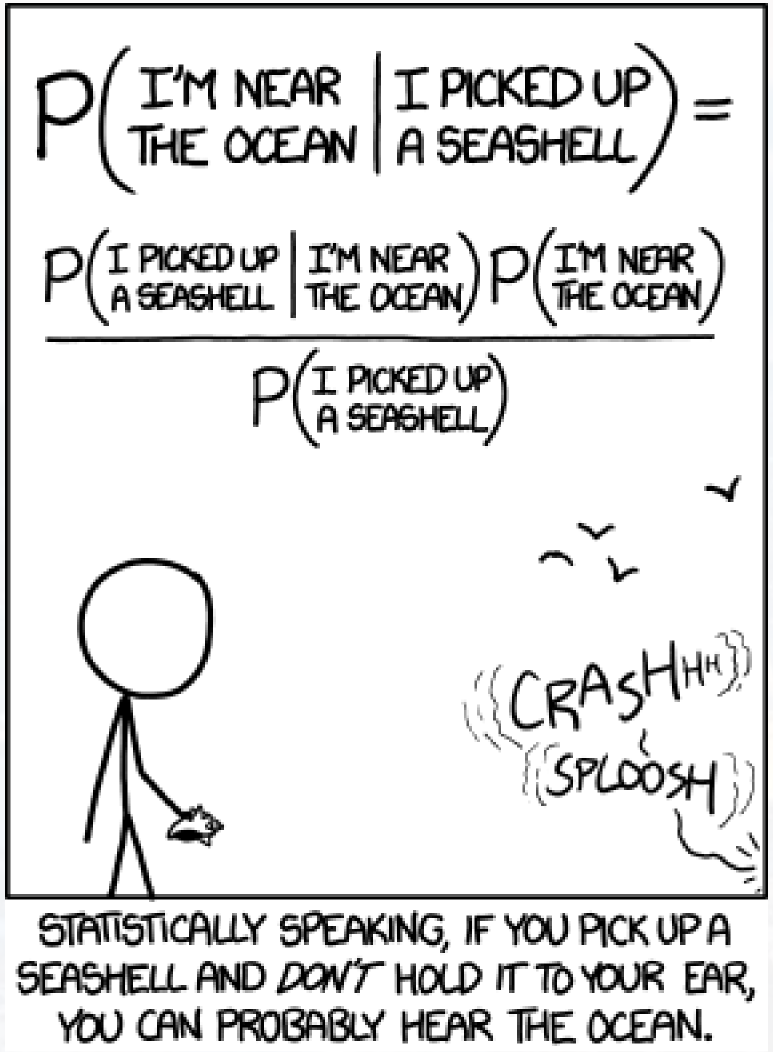
Markus Hohle

University California, Berkeley

Machine Learning Algorithms

MSSE 277B, 3 Units

Fall 2024

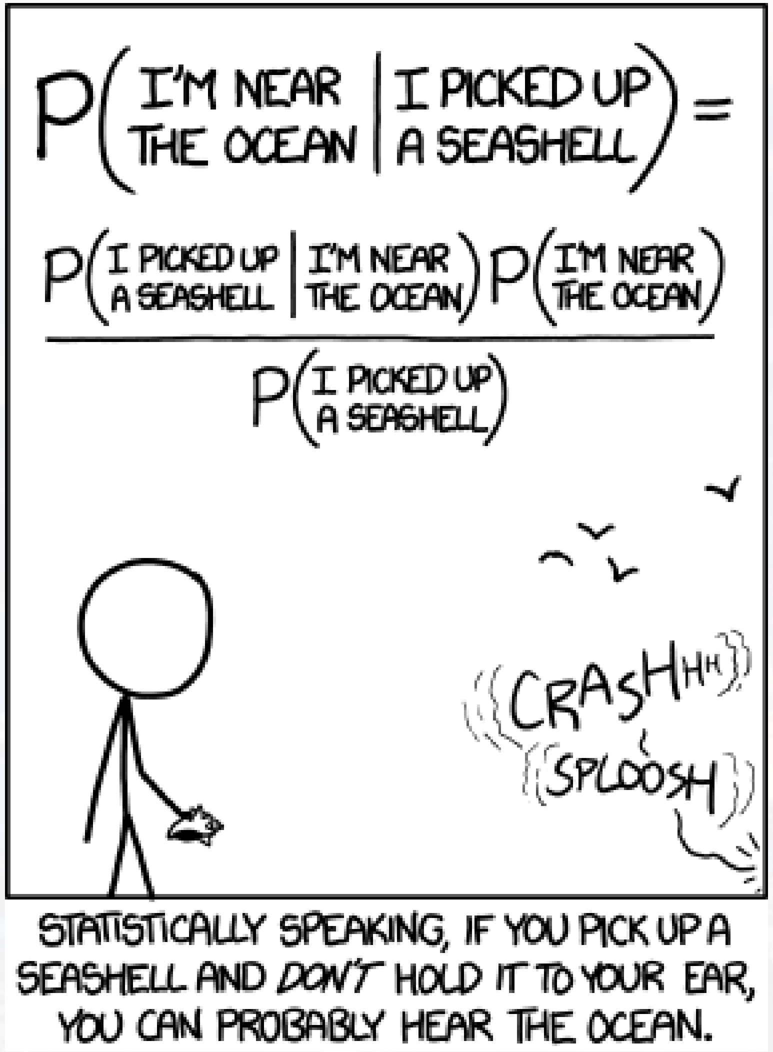


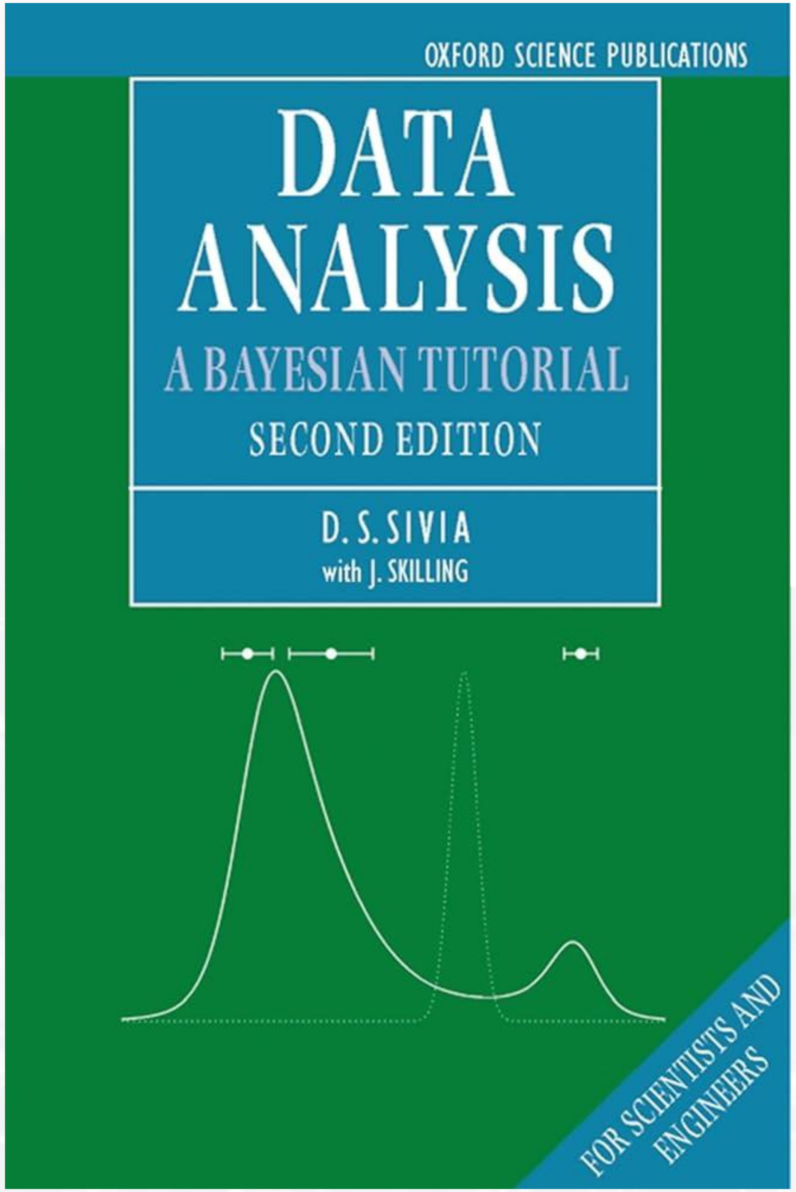
Outline

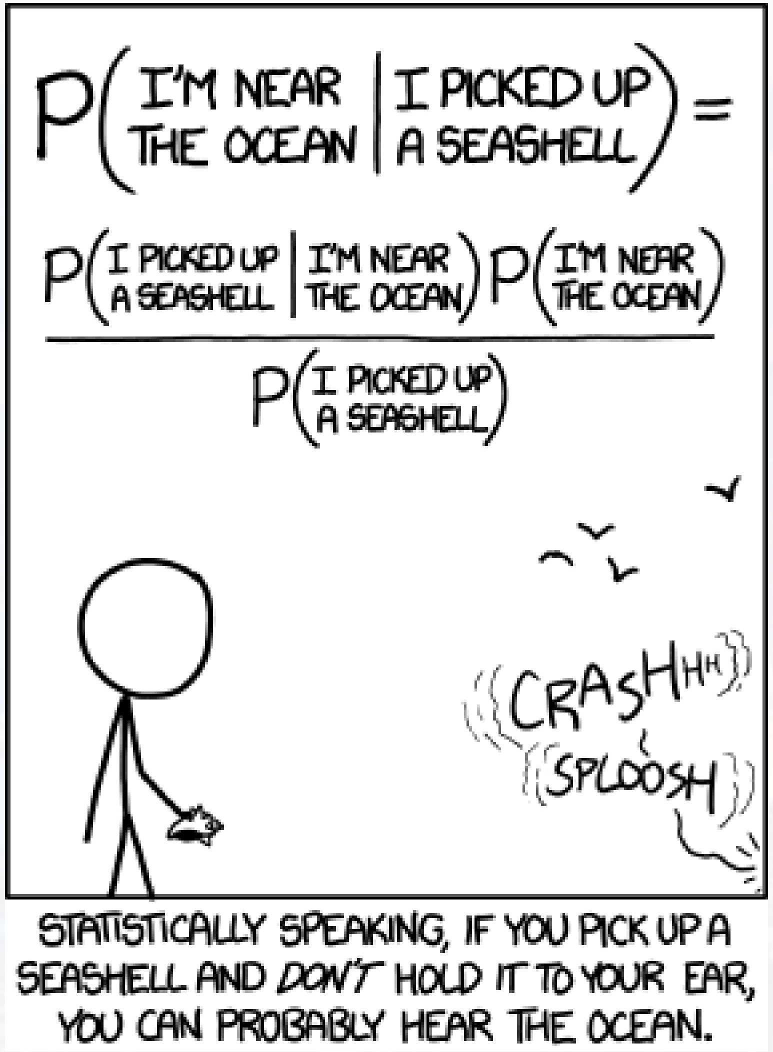
- The Idea and Bayes Theorem
- Naïve Bayes
- Parameter Estimation
- Model Selection

FYI

- Bayesian Networks (Graphs)
- Variational Bayes


P(I'M NEAR THE OCEAN | I PICKED UP A SEASHELL) =
P(I PICKED UP A SEASHELL | I'M NEAR THE OCEAN) P(I'M NEAR THE OCEAN)
P(I PICKED UP A SEASHELL)
CRASHHH
SPLOOSH
STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND DON'T HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.


OXFORD SCIENCE PUBLICATIONS
DATA ANALYSIS
A BAYESIAN TUTORIAL
SECOND EDITION
D. S. SIVIA
with J. SKILLING
FOR SCIENTISTS AND ENGINEERS



Outline

- The Idea and Bayes Theorem
- Naïve Bayes
- Parameter Estimation
- Model Selection

FYI

- Bayesian Networks (Graphs)
- Variational Bayes

Why Bayesian Statistics?

- frequentist: assuming sample is infinite **(even tough there are corrections for small n)**

vs:

- bayesian:
  → taking the **exact amount** of information into account that's available
  → system "learns" by adding more data
  → is based on information theory & links to quantum mechanics
  → **maximum entropy, given constrains** (prior knowledge)
    → variational calculus
      → EM algorithm (GMM, HMM etc)
      → **V**ariational **A**uto **E**ncoder
    → non-parametric (e. g. in contrast to MLE)

….and more

Bayesian Methods:

The Idea and Bayes Theorem

Thomas Bayes
(1701 - 1761)

$P(A \cap B)$ probability ***P*** that the events ***A* and *B*** occur

so far: A and B were independent $P(A \cap B) = P(A)P(B) = P(B)P(A)$

now: **conditional probabilities** | "given" or "under the condition"

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$\rightarrow \boxed{P(A|B)P(B) = P(B|A)P(A)}$$

Bayes Theorem

posterior

$$\boldsymbol{P(A|B)} = \frac{P(B|A)\boldsymbol{P(A)}}{P(B)}$$

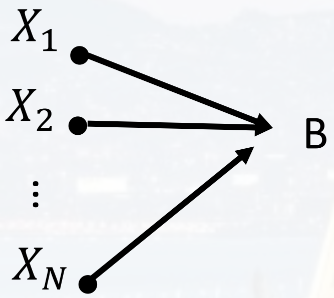
prior

Thomas Bayes
(1701 - 1761)

$$\boxed{P(A|B)P(B) = P(B|A)\mathrm{P(A)}}$$

Bayes Theorem

posterior $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ **prior**



$$P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$$

$$P(B) = \int P(B|X)P(X)\, dX$$

marginalization

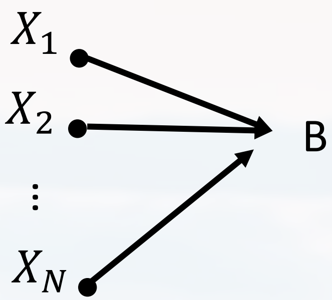
example:

model: M
data: D

$$P(D|M) = \int P(D|all\ model\ param, M)\ P(all\ model\ param|M)\ \ d\ all\ model\ param$$

Bayesian Methods:

The Idea and Bayes Theorem



$$P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$$

$$P(B) = \int P(B|X)P(X)\,dX$$

marginalization

Thomas Bayes (1701 - 1761)

example:

model: M
data: D

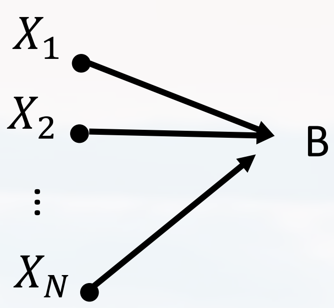
$$P(D|M) = \int P(D|all\ model\ param, M)\ P(all\ model\ param|M)\ \ d\ all\ model\ param$$

for a normal distribution $\mathcal{N}(\mu,\sigma)$ $$P(D|\mathcal{N}(\mu,\sigma)) = \int P(D|\mathcal{N}(\mu,\sigma))\ P(\mu,\sigma|\mathcal{N}(\mu,\sigma))\ d\ \Omega_{\mu,\sigma}$$

for a Poisson distribution $p(\lambda)$ $$P(D|p(\lambda)) = \int P(D|p(\lambda))\ P(\lambda|p(\lambda))\ d\ \lambda$$

and so on…

Bayesian Methods:



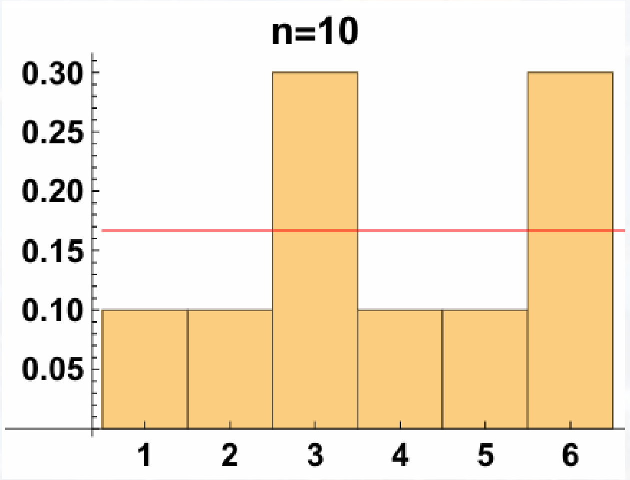
$$P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$$

$$P(B) = \int P(B|X)P(X)\,dX$$

marginalization

Thomas Bayes (1701 - 1761)

for a normal distribution $\mathcal{N}(\mu,\sigma)$ $\quad P(D|\mathcal{N}(\mu,\sigma)) = \int P(D|\mathcal{N}(\mu,\sigma))\,P(\mu,\sigma|\mathcal{N}(\mu,\sigma))\,d\,\Omega_{\mu,\sigma}$

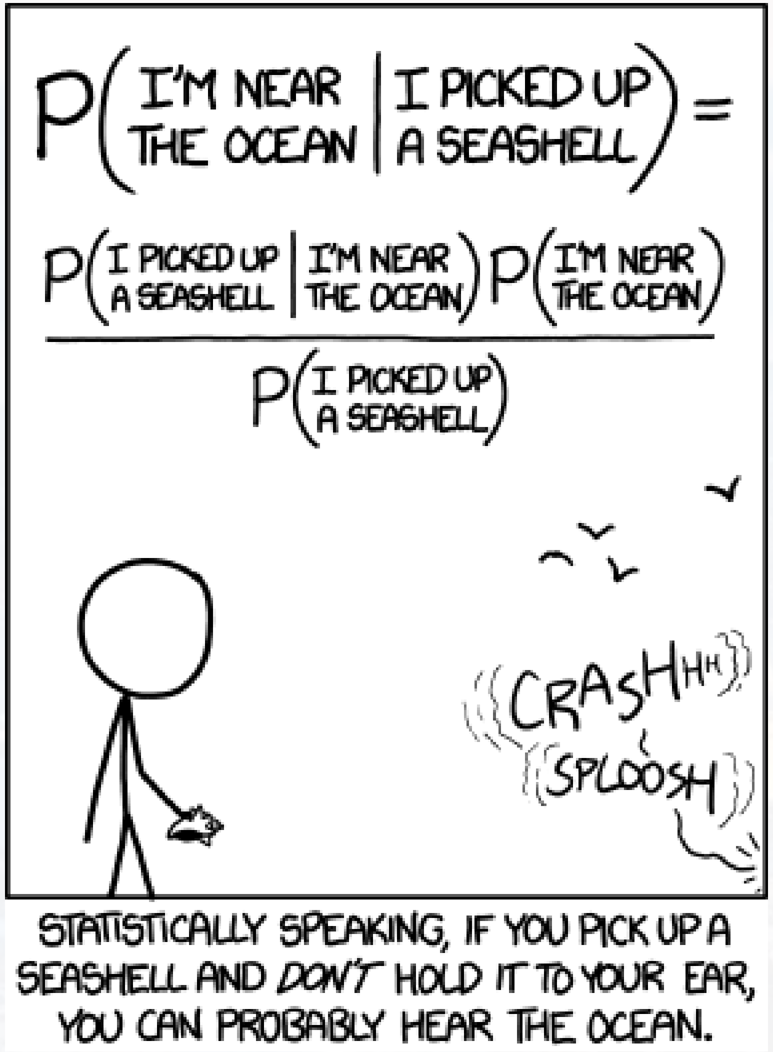


$\sigma = 2$, $\mu = 3.5$

$\sigma = 2$, $\mu = 5.0$

$\sigma = 1.5$, $\mu = 3.5$

$\sigma = 7.0$, $\mu = 1.0$



<u>Outline</u>

- The Idea and Bayes Theorem
- Naïve Bayes
- Parameter Estimation
- Model Selection

<u>FYI</u>

- Bayesian Networks (Graphs)
- Variational Bayes

Bayesian Methods:

model: M
data: D

$$P(D|M) = \int P(D|all\ model\ param, M)\ \boxed{P(all\ model\ param|M)}\ d\ all\ model\ param$$

for a normal distribution $\mathcal{N}(\mu, \sigma)$ $$P(D|\mathcal{N}(\mu,\sigma)) = \int P(D|\mathcal{N}(\mu,\sigma))\ \boxed{P(\mu,\sigma|\mathcal{N}(\mu,\sigma))}\ d\ \Omega_{\mu,\sigma}$$

Naïve Bayes:

- all model parameters are **mutually independent**
- i. e.: no **correlation** between model parameters

→ $\vec{x}$: vector with all model parameters (or features)

$$P(M|\vec{x}) = P(M)\ \prod_{i=1}^{I} P(x_i|M)$$

again, for a normal distribution:

$$P(\mu,\sigma|\mathcal{N}(\mu,\sigma)) = P\big(\mathcal{N}(\mu,\sigma)\big)\ \cdot P(\sigma|\ \mathcal{N}(\mu,\sigma))\ P(\mu|\ \mathcal{N}(\mu,\sigma))$$

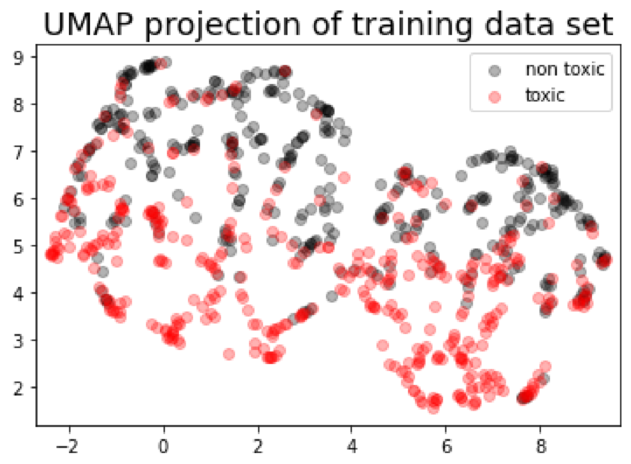
Bayesian Methods:

Naïve Bayes

$\vec{x}$: vector with all model parameters (or features)

Index	molecular_weight	electronegativity	bond_lengths	num_hydrogen_bonds	logP	label
0	413.228	2.94416	3.41991	1	10.4335	Toxic
1	447.945	3.55371	3.66831	7	10.3475	Toxic
2	309.199	3.19761	2.84841	0	7.88825	Non-Toxic
3	382.554	3.8653	3.46237	8	9.59041	Toxic
4	310.904	3.18141	2.87774	6	7.85477	Non-Toxic
5	353.857	3.12105	3.32724	6	8.58887	Non-Toxic

K **different classes**



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Theorem

Bayesian Methods:

$\vec{x}$: vector with all model parameters (or features)

K different classes

Index	molecular_weight	electronegativity	bond_lengths	num_hydrogen_bonds	logP	label
0	413.228	2.94416	3.41991	1	10.4335	Toxic
1	447.945	3.55371	3.66831	7	10.3475	Toxic
2	309.199	3.19761	2.84841	0	7.88825	Non-Toxic
3	382.554	3.8653	3.46237	8	9.59041	Toxic
4	310.904	3.18141	2.87774	6	7.85477	Non-Toxic
5	353.857	3.12105	3.32724	6	8.58887	Non-Toxic

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Theorem

$$P(C_k|\vec{x}) = \frac{\boxed{P(\vec{x}|C_k)}P(C_k)}{P(\vec{x})} = \frac{P(C_k)}{P(\vec{x})}\boxed{\prod_{i=1}^{I} P(x_i|C_k)} \sim P(C_k)\prod_{i=1}^{I} P(x_i|C_k)$$

$$\sum_{k=1}^{K} P(C_k|\vec{x}) = 1$$

Naïve Bayes

Bayesian Methods:

$$\boldsymbol{P(A|B)} = \frac{P(B|A)\boldsymbol{P(A)}}{\boldsymbol{P(B)}}$$

Bayes Theorem

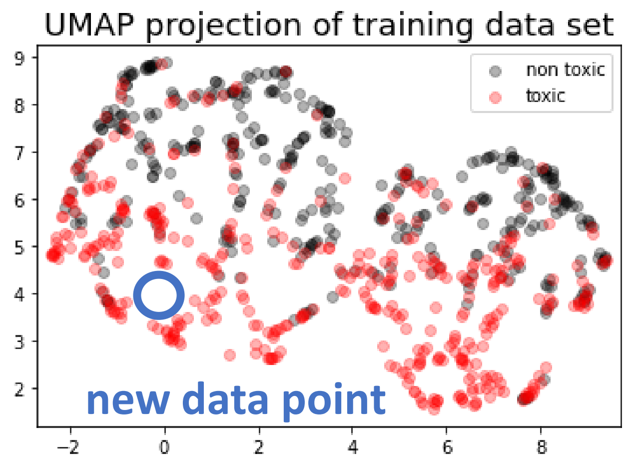
$$P(C_k|\vec{x}) = \frac{P(\vec{x}|C_k)P(C_k)}{P(\vec{x})} = \frac{P(C_k)}{P(\vec{x})}\prod_{i=1}^{I} P(x_i|C_k) \sim P(C_k)\prod_{i=1}^{I} P(x_i|C_k)$$

$$\sum_{k=1}^{K} P(C_k|\vec{x}) = 1$$

finding the ***k***, that maximizes $P(C_k|\vec{x})$

$$k_{new} = \underset{k}{argmax}\left\{P(C_k)\prod_{i=1}^{I}\boxed{P(x_i|C_k)}\right\}$$

from the training data



different models for $P(x_i|C_k)$

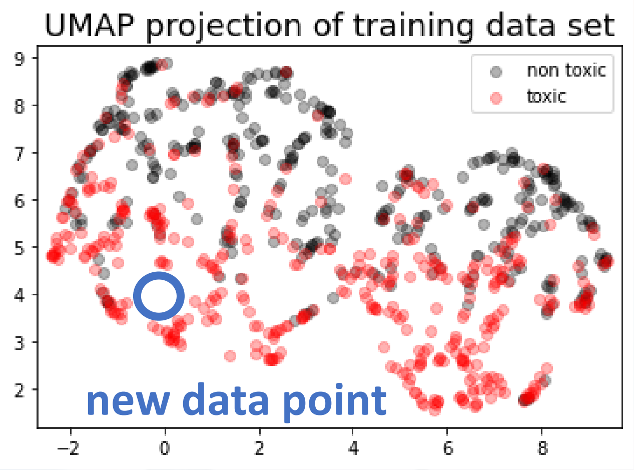
- multinomial
- Gaussian
- …

Bayesian Methods:

finding the ***k***, that maximizes $P(C_k|\vec{x})$

$$k_{new} = \underset{k}{argmax}\left\{P(C_k)\prod_{i=1}^{I}\boxed{P(x_i|C_k)}\right\}$$

from the training data



different models for $P(x_i|C_k)$

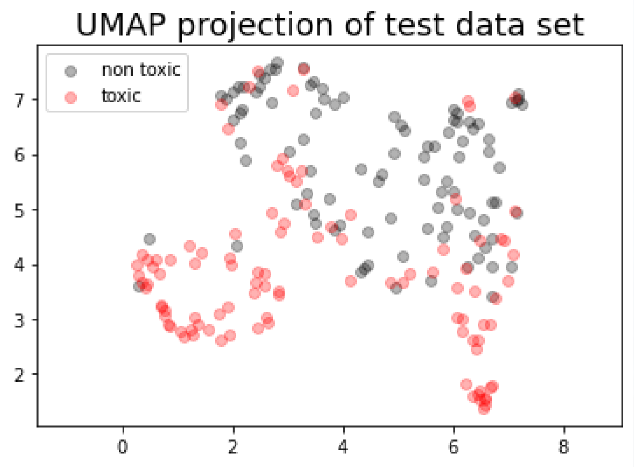
- multinomial
- Gaussian
- …

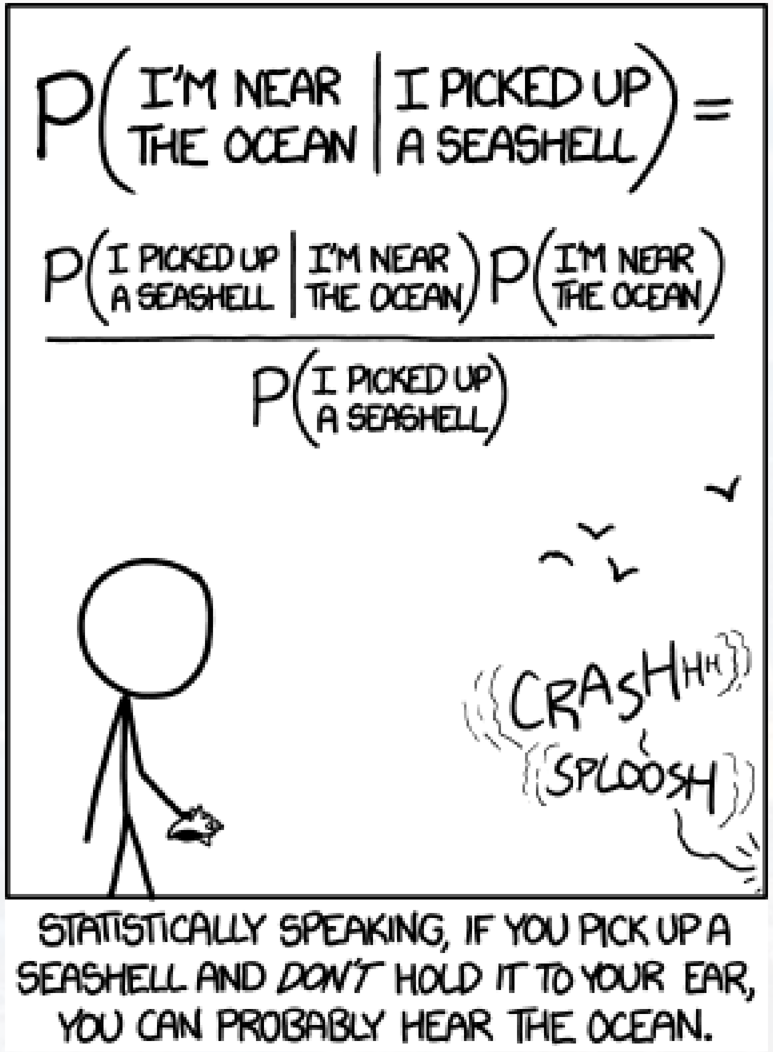
Python:

```
from sklearn.naive_bayes import *

gnb    = GaussianNB()
k_pred = gnb.fit(TrainX, Traink).predict(TestX)


mnb    = MultinomialNB()
k_pred = mnb.fit(TrainX, Traink).predict(TestX)
```





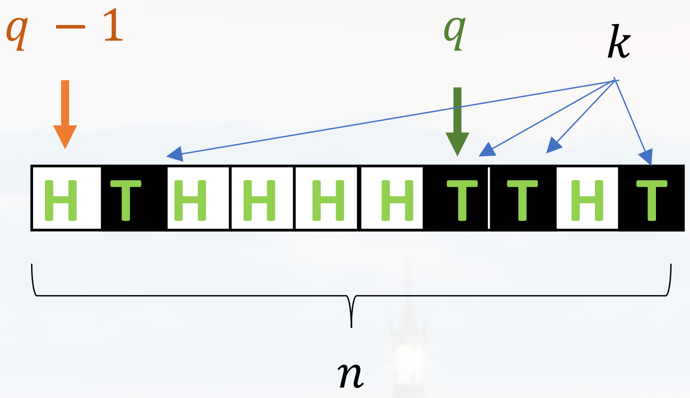
Outline

- The Idea and Bayes Theorem
- Naïve Bayes
- Parameter Estimation
- Model Selection

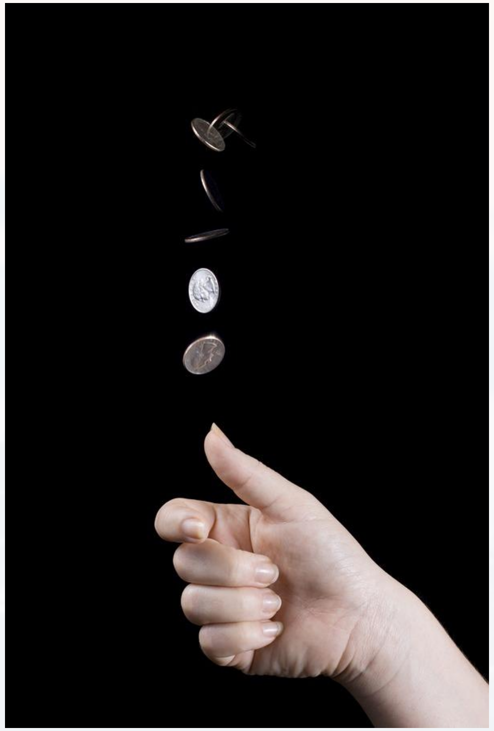
FYI

- Bayesian Networks (Graphs)
- Variational Bayes

Bayesian Methods:



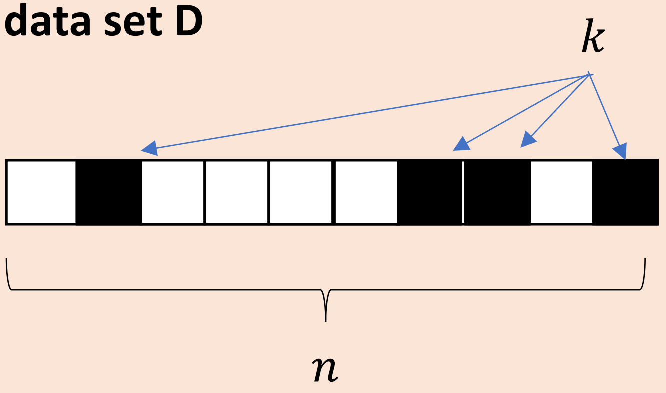
$\boldsymbol{q} = ?$



fair coin? q = 0.5 ???

mutation q = ??

data set D

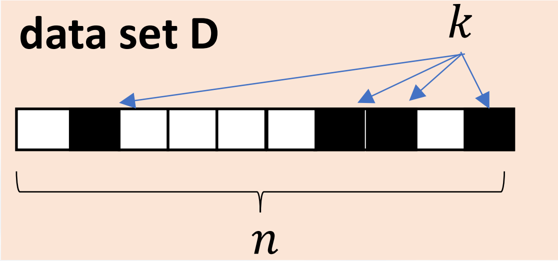


$q = ?$

goal:

- **P(q|D)**
- the larger ***D***, the more certain ***q***

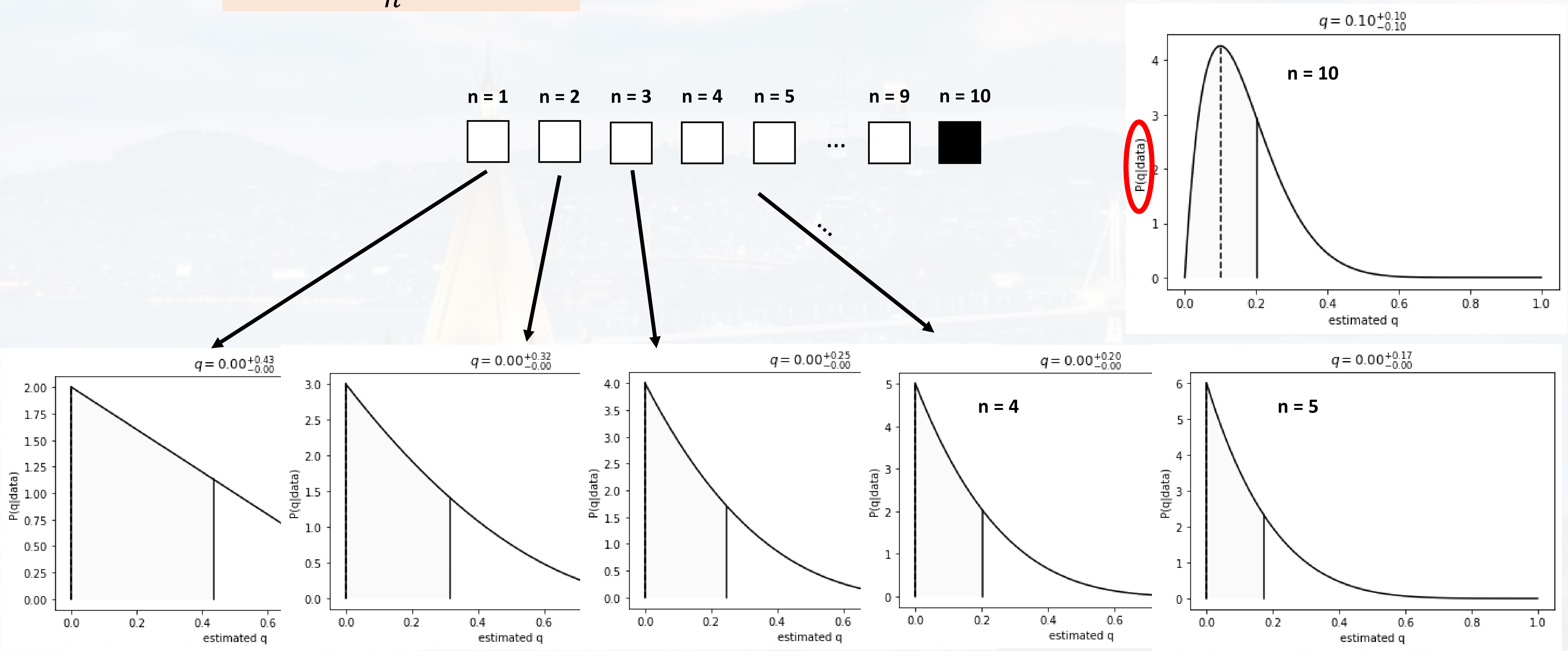
→ learning



$q = ?$

goal:

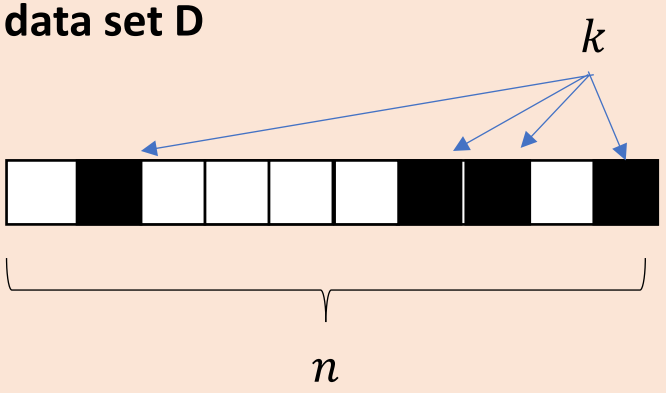
- **P(q|D)**
- the larger ***D***, the more certain ***q***
  → learning



Bayesian Methods:

Parameter Estimation

data set D



$\boldsymbol{q} = ?$

goal:
- **P(q|D)**
- the larger ***D***, the more certain ***q***
  → learning

$$P(k|n,q) = \binom{n}{k} q^k (1-q)^{n-k}$$

Bayes' theorem:

likelihood function **(here: binomial)**

prior

evidence **(const wrt q)**

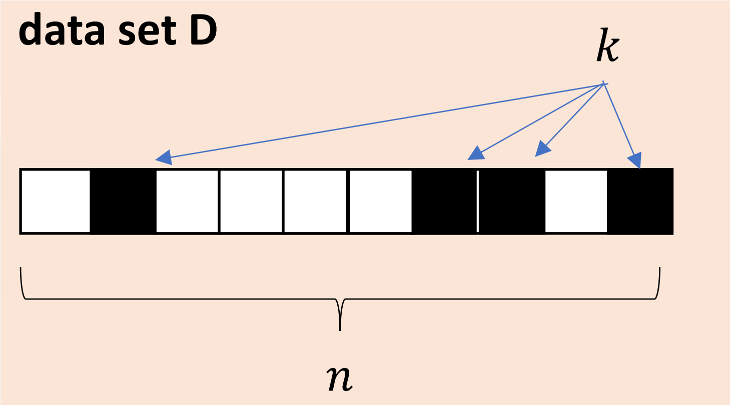
$$P(q|data\ set) = \frac{\boxed{P(data\ set|q)}\boxed{P(q)}}{\boxed{P(data\ set)}}$$

$$= \frac{\binom{n}{k} q^k (1-q)^{n-k}}{P(D)} P(q)$$

$$\sim q^k (1-q)^{n-k}\, P(q)$$

$P(D)$ and $\binom{n}{k}$ are no functions of q

Bayesian Methods:



$\boldsymbol{q} = ?$

goal:
- **P(q|D)**
- the larger ***D***, the more certain ***q***
  → learning

$$P(k|n,q) = \binom{n}{k} q^k (1-q)^{n-k}$$

$$P(q|data\ set) = \frac{P(data\ set|q)P(q)}{P(data\ set)}$$

$$= \frac{\binom{n}{k} q^k (1-q)^{n-k}}{P(D)} P(q)$$

$$\sim q^k (1-q)^{n-k}\ P(q)$$

max. entropy: $\mathrm{P}(q) = const$
if no prior information about q

$$\sim q^k (1-q)^{n-k}$$

$$\boxed{P(q|data\ set) = \frac{q^k (1-q)^{n-k}}{\int_0^1 q^k (1-q)^{n-k}\ dq}}$$

Bayesian Methods:

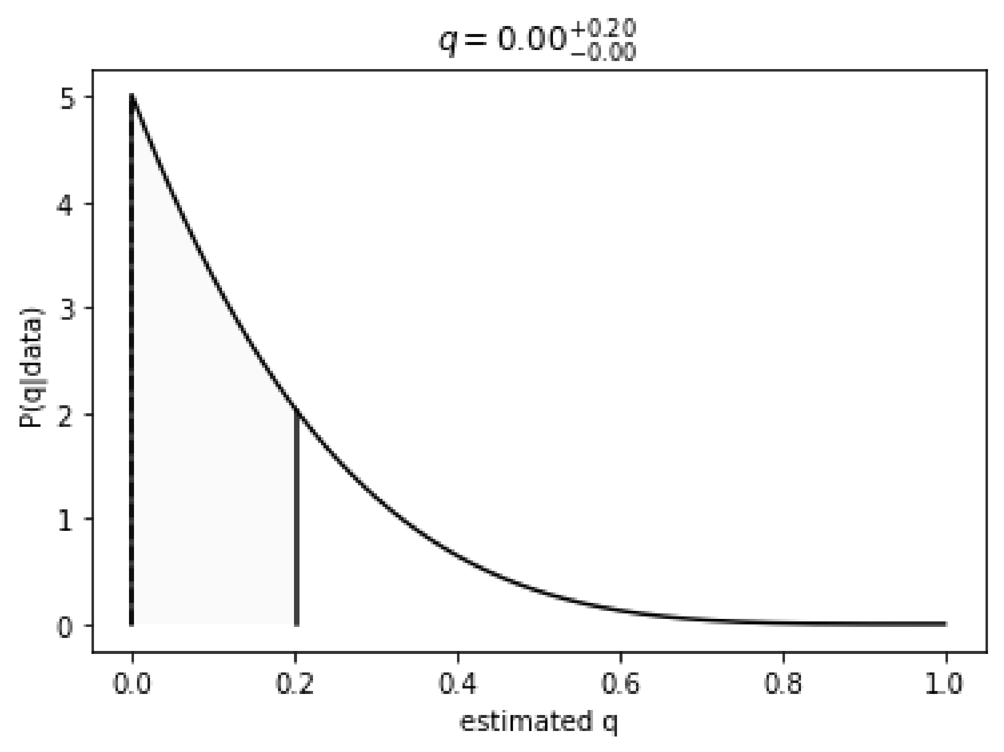
$$P(q|data\ set) = \frac{q^k(1-q)^{n-k}}{\int_0^1 q^k(1-q)^{n-k}\ dq}$$

check out bayesian_bino.py

```
n1 = 4
k1 = np.random.binomial(n1, 0.25)
```

creating artificial data set
note: in reality ***q*** is unknown!

```
[q1, b, _] = bayesian_bino(n1, k1)
```

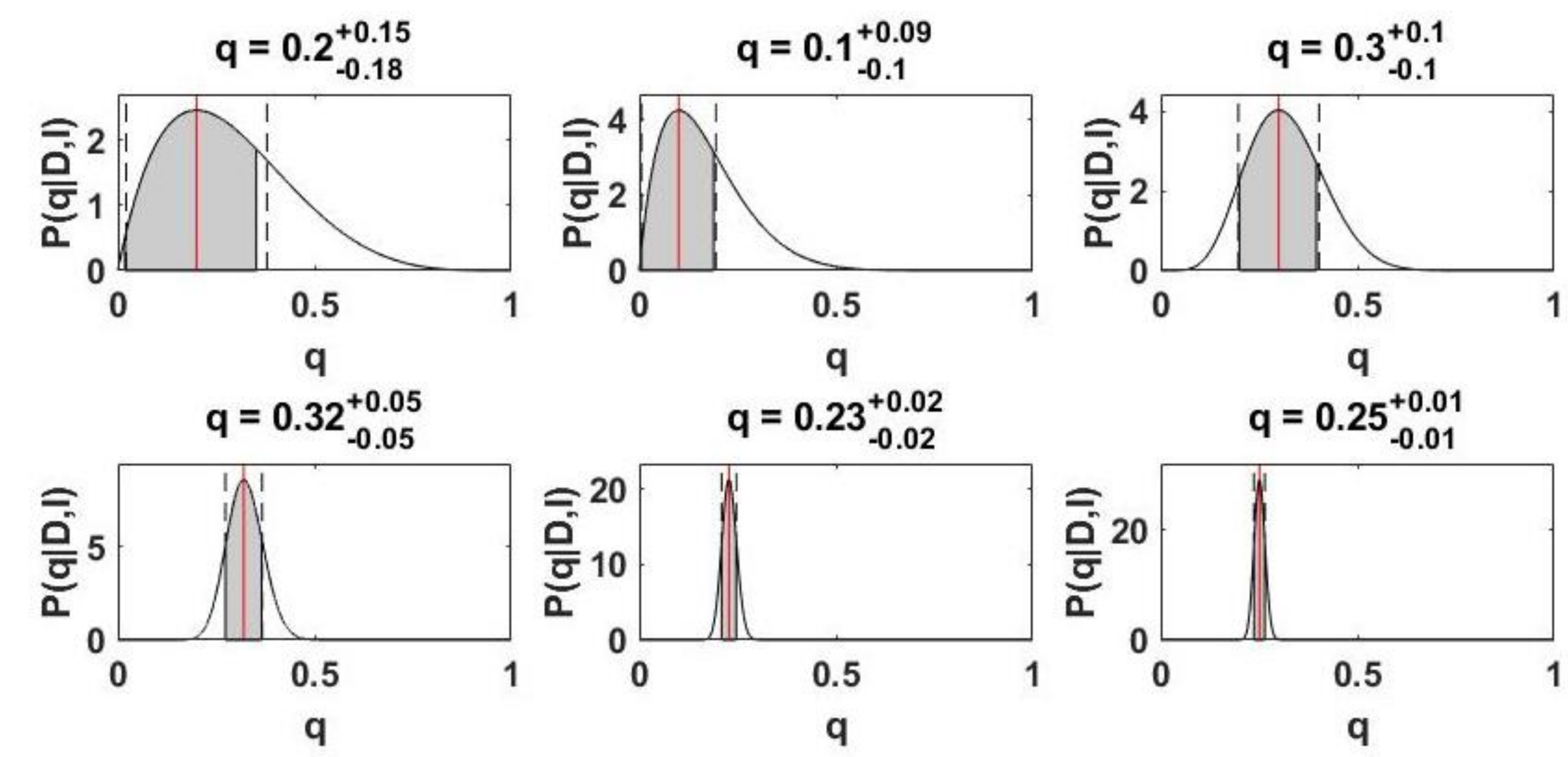


Bayesian Methods:

Parameter Estimation

check out `bayesian_bino.py`

$$P(q|data\ set) = \frac{q^k(1-q)^{n-k}}{\int_0^1 q^k(1-q)^{n-k}\ dq}$$



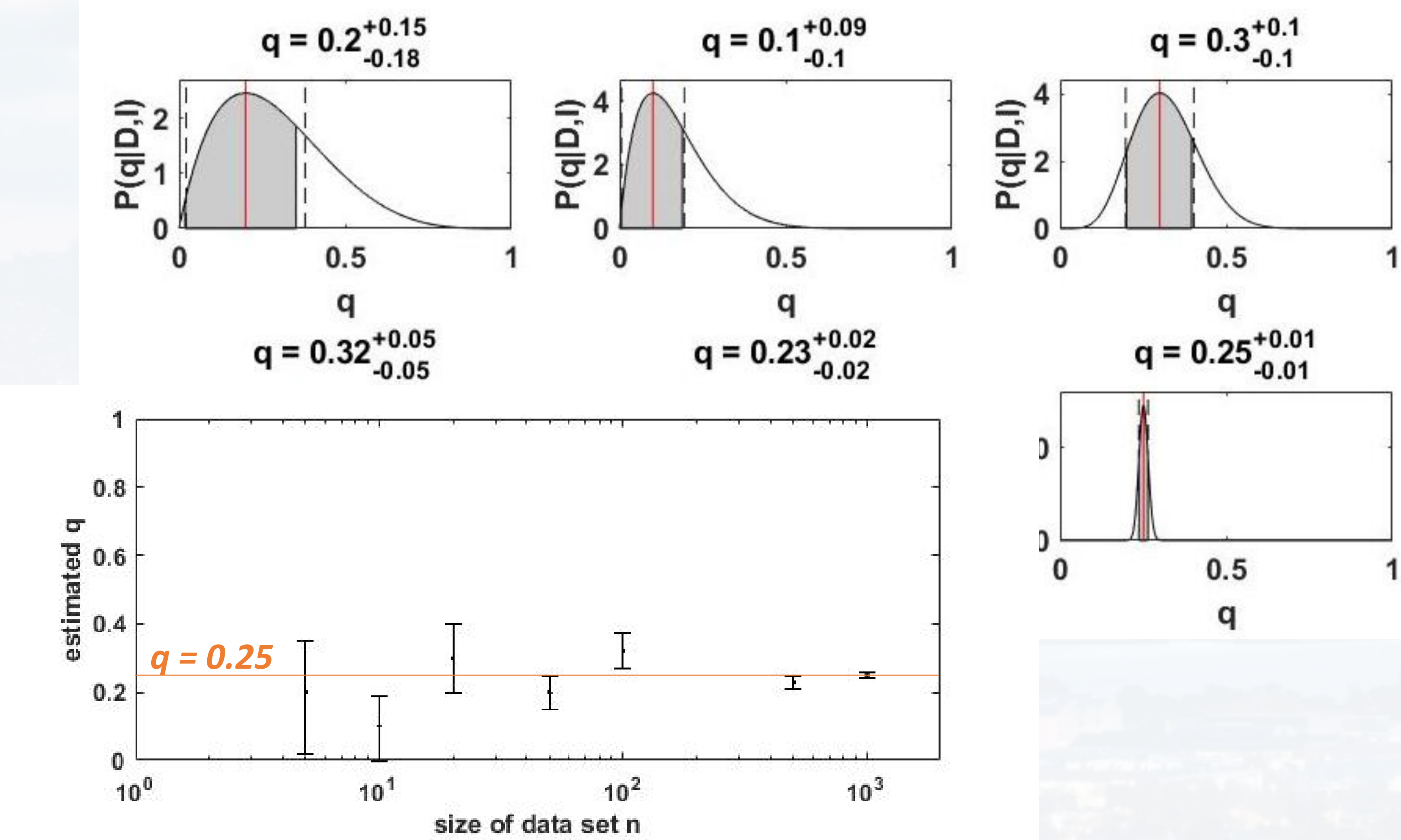
n	estimated q
5	$0.2^{+0.15}_{-0.18}$
10	$0.1^{+0.09}_{-0.1}$
20	$0.3^{+0.1}_{-0.1}$
50	$0.2^{+0.05}_{-0.06}$
100	$0.32^{+0.05}_{-0.05}$
500	$0.23^{+0.02}_{-0.02}$
1,000	$0.25^{+0.01}_{-0.01}$
infinity	0.25

Bayesian Methods:

Parameter Estimation

check out `bayesian_bino.py`

$$P(q|data\ set) = \frac{q^k(1-q)^{n-k}}{\int_0^1 q^k(1-q)^{n-k}\ dq}$$



n	estimated q
5	$0.2^{+0.15}_{-0.18}$
10	$0.1^{+0.09}_{-0.1}$
20	$0.3^{+0.1}_{-0.1}$
50	$0.2^{+0.05}_{-0.06}$
100	$0.32^{+0.05}_{-0.05}$
500	$0.23^{+0.02}_{-0.02}$
1,000	$0.25^{+0.01}_{-0.01}$
infinity	0.25

Of course, Bayesian Parameter Estimation works with **any other pdf**

goal: - **P(q|D)**
- the larger ***D***, the more certain ***q***
→ learning

likelihood function

$$P(q|data\ set) = \frac{\boxed{P(data\ set|q)}\boxed{P(q)}}{\boxed{P(data\ set)}}$$

prior

evidence **(const wrt q)**

<u>What is the average number of WhatsUp messages I get every day?</u>

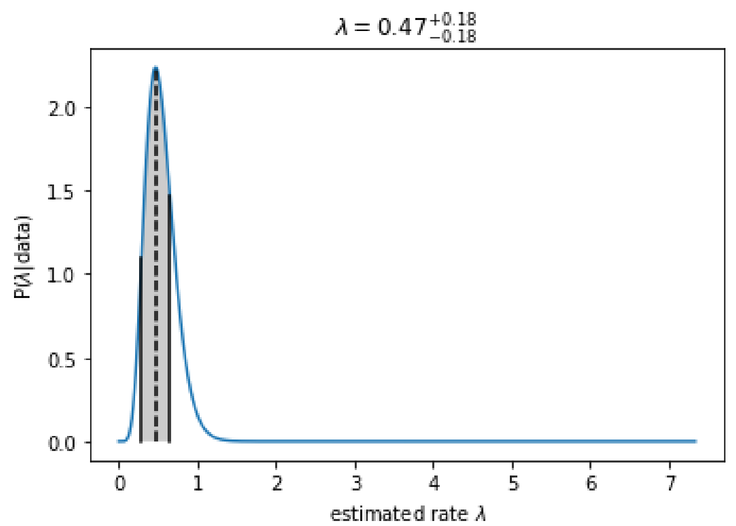
Mon:	5
Tue:	7
Wed:	1
Thu:	3
Fri:	9
Sat:	2
Sun:	5

event - has no duration
- is rare

→ Poissonian

$$P(k|\lambda) = \frac{\lambda^k}{k!}\, e^{-\lambda}$$

```
data = np.random.poisson(lam = 0.4, 15)
poissfit(data)
```



Bayesian Methods:

Of course, Bayesian Parameter Estimation works with **any other pdf**

goal: - **P(q|D)**
- the larger ***D***, the more certain ***q***
→ learning

<u>What is the average number of WhatsUp messages I get every day?</u>

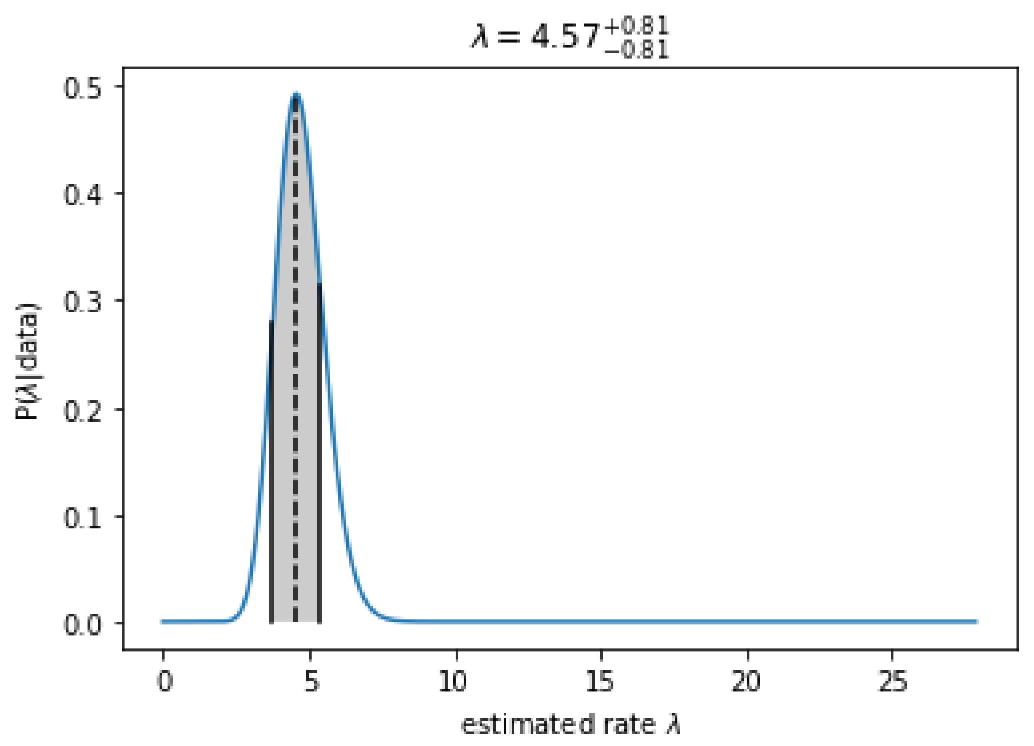
Mon:	5
Tue:	7
Wed:	1
Thu:	3
Fri:	9
Sat:	2
Sun:	5

event
- has no duration
- is rare

→ Poissonian

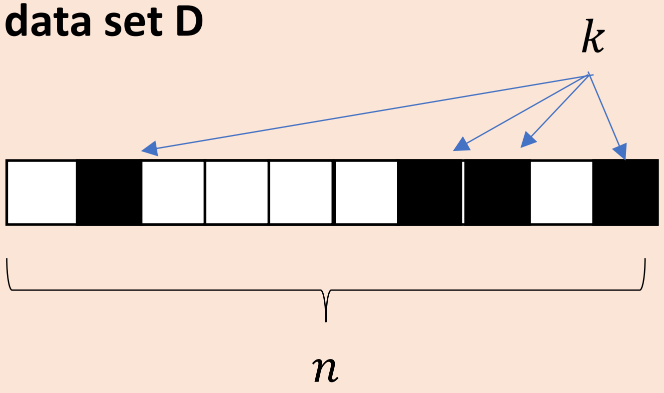
$$P(k|\lambda) = \frac{\lambda^k}{k!}\, e^{-\lambda}$$

```
poissfit([5, 7, 1, 3, 9, 2, 5])
```



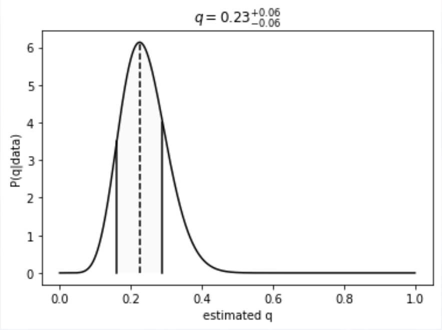
What if there is new data?

data set D

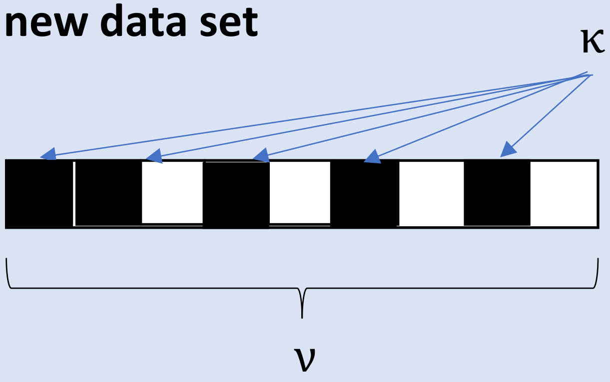


$\cancel{\boldsymbol{q} = ?}$

$$\boxed{P(q|data\ set) = \frac{q^k(1-q)^{n-k}}{\int_0^1 q^k(1-q)^{n-k}\ dq}}$$



new data set



if there **is** prior information ***I*** about q:

$$P(q|new\ data\ set, I) = \frac{P(new\ data\ set|q, I)\ \boldsymbol{P(q, I)}}{P(new\ data\ set)}$$

What if there is new data?

$$P(q|data\ set) = \frac{q^k(1-q)^{n-k}}{\int_0^1 q^k(1-q)^{n-k}\ dq}$$

$$P(q|new\ data\ set, I) = \frac{P(new\ data\ set|q,I)\ \boldsymbol{P(q,I)}}{P(new\ data\ set)}$$

$$= \frac{q^{\kappa}(1-q)^{\nu-\kappa}\ q^k(1-q)^{n-k}}{\int_0^1 q^{\kappa}(1-q)^{\nu-\kappa}\ q^k(1-q)^{n-k}\ dq}$$

$$= \frac{q^{k+\kappa}(1-q)^{\nu-\kappa+n-k}}{\int_0^1 q^{k+\kappa}(1-q)^{\nu-\kappa+n-k}\ dq}$$

often: $\kappa = \alpha - 1$

$\beta = \nu - \kappa - 1$

$$= \frac{q^{k+\alpha-1}(1-q)^{n-k+\beta-1}}{\boxed{\int_0^1 q^{k+\alpha-1}(1-q)^{n-k+\beta-1}\ dq}}$$

Beta function

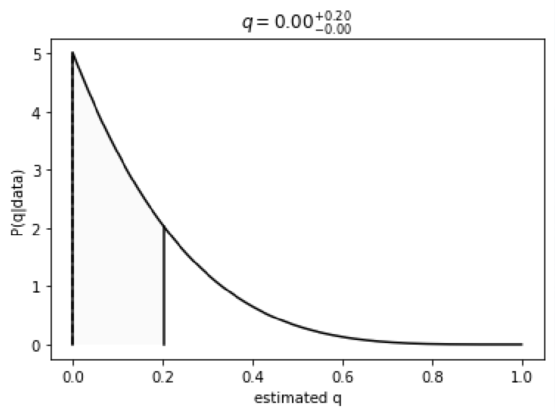
Bayesian Methods:

What if there is new data?

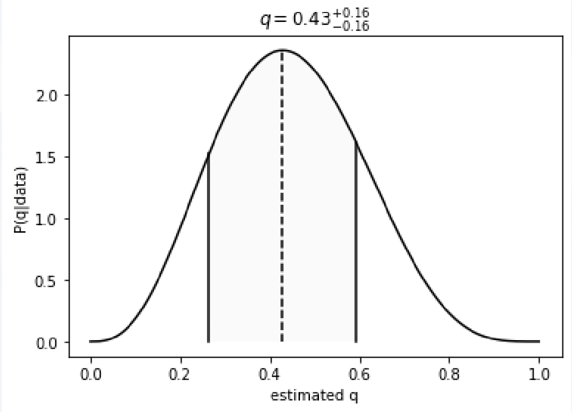
$$P(q|new\ data\ set, I) = \frac{q^{\kappa}(1-q)^{\nu-\kappa}\ q^{k}(1-q)^{n-k}}{\int_0^1 q^{\kappa}(1-q)^{\nu-\kappa}\ q^{k}(1-q)^{n-k}\ dq}$$

```
n1 = 4
k1 = np.random.binomial(n1, q = 0.2)
[_, _, Prior] = bayesian_bino(n1, k1)
```

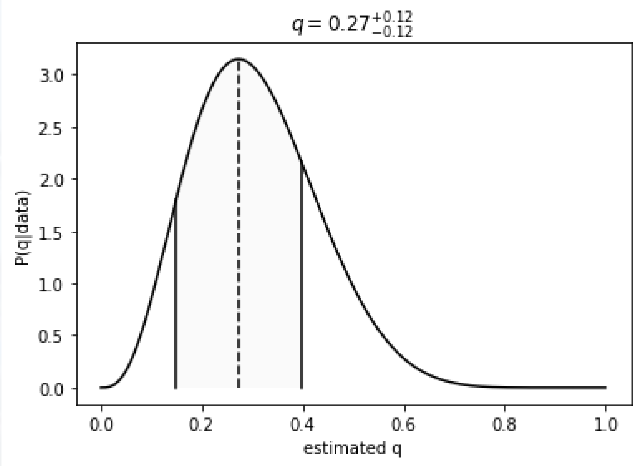
```
n2 = 7
k2 = np.random.binomial(n2, q = 0.2)
[_, _, _] = bayesian_bino(n2, k2)
```



$$P(q, I) = \frac{q^{k}(1-q)^{n-k}}{\int_0^1 q^{k}(1-q)^{n-k}\ dq}$$




```
[_ , _, _] = bayesian_bino(n2, k2, Prior = Prior)
```

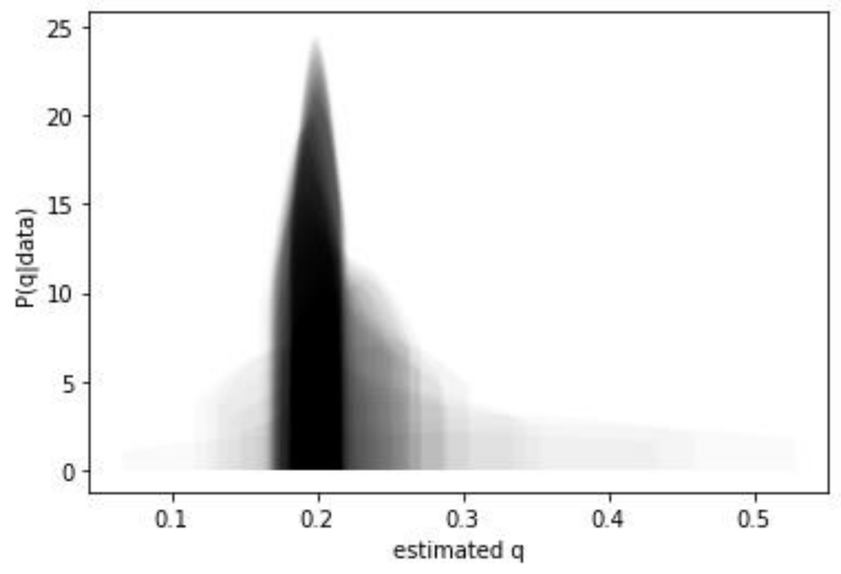


Bayesian Methods:

What if there is new data?

$$P(q|new\ data\ set, I) = \frac{q^{\kappa}(1-q)^{\nu-\kappa}\ q^{k}(1-q)^{n-k}}{\int_0^1 q^{\kappa}(1-q)^{\nu-\kappa}\ q^{k}(1-q)^{n-k}\ dq}$$

The **posterior** from the **previous** experiment is the **prior** of the **next** experiment



→ we become more certain about the model parameters
→ learning!

→ see e.g. **V**ariational **A**uto **E**ncoders

2D images → 3D objects

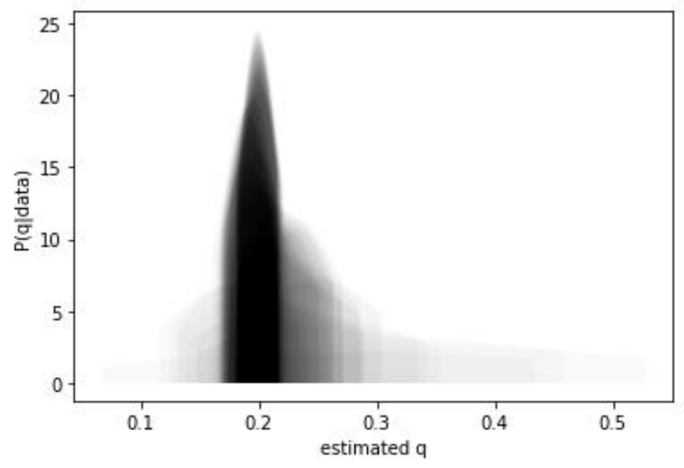
credit: StableAI

Bayesian Methods:

What if there is new data?

$$P(q|new\ data\ set, I) = \frac{q^{\kappa}(1-q)^{\nu-\kappa}\ q^{k}(1-q)^{n-k}}{\int_0^1 q^{\kappa}(1-q)^{\nu-\kappa}\ q^{k}(1-q)^{n-k}\ dq}$$

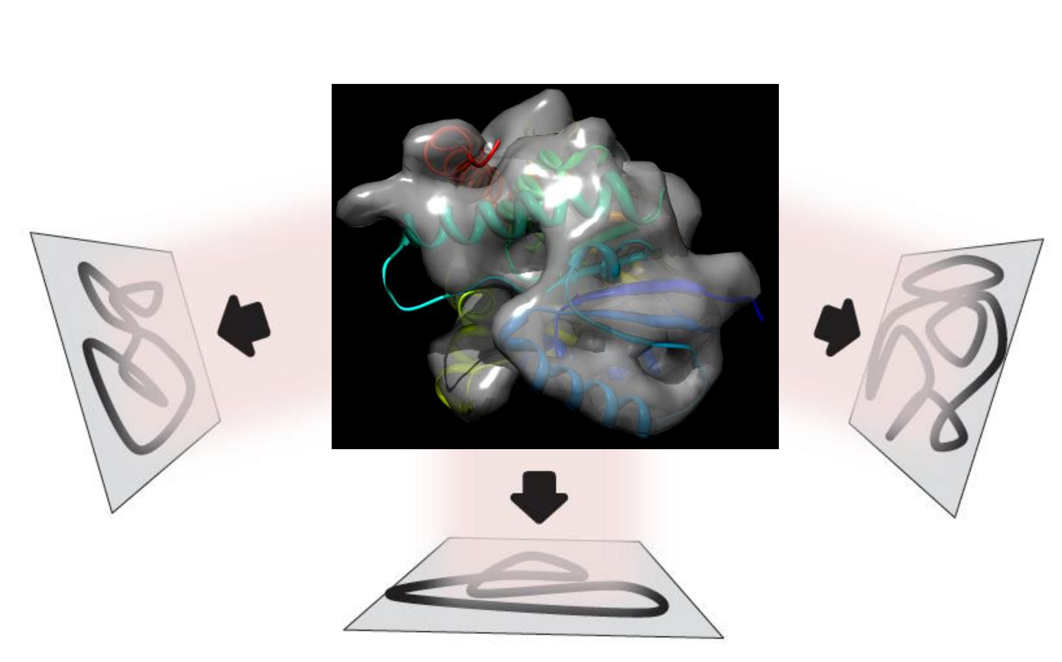
The posterior from the previous experiment is the prior of the next experiment



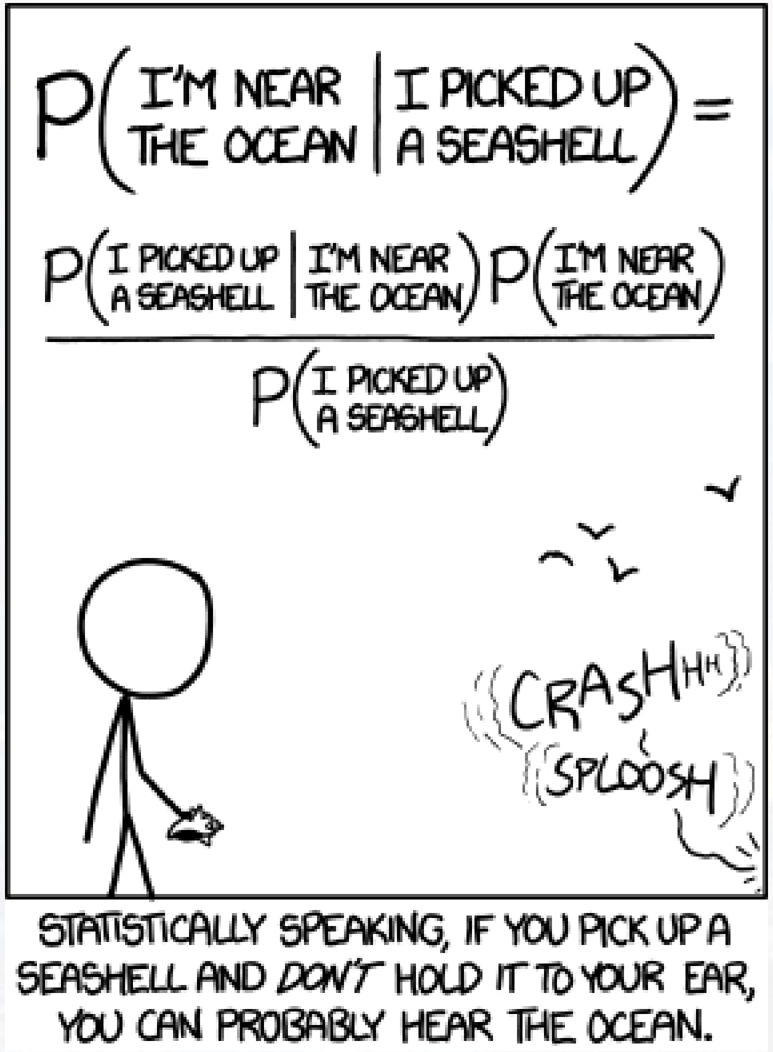
→ we become more certain about the model parameters
→ learning!

→ see e.g. **V**ariational **A**uto **E**ncoders   2D images → 3D objects

Cryo – EM: 3D structure from 2D images/projections



(image courtesy: Thomas Becker, GC LMU Munich)



Outline

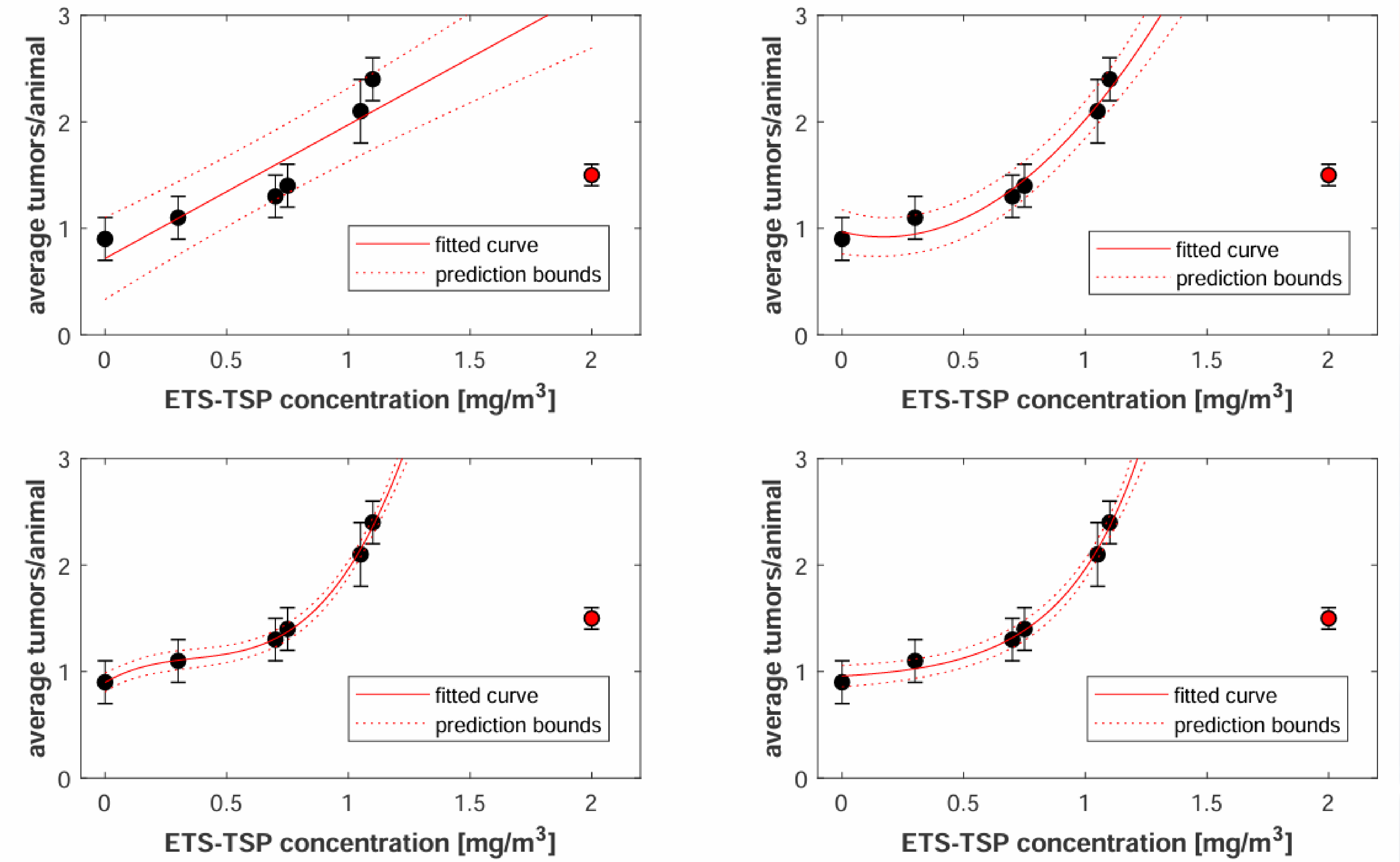
- The Idea and Bayes Theorem
- Naïve Bayes
- Parameter Estimation
- Model Selection

FYI

- Bayesian Networks (Graphs)
- Variational Bayes

Bayesian Methods:

often, we have many competing models → assigning probabilities if a model is correct



DOI: 10.1093/carcin/23.3.511
Source: PubMed

Bayesian Methods:

often, we have many competing models → assigning probabilities if a model is correct

D : data
$\mathrm{M_A}$: model A
$\mathrm{M_B}$: model B

goal: $$\rho = \frac{P(\mathrm{M_A}|D)}{P(\mathrm{M_B}|D)}$$

Bayes' theorem

$$= \frac{\boxed{P(D|M_A)}\,P(M_A)}{P(D)} \cdot \frac{P(D)}{\boxed{P(D|M_B)}\,P(M_B)}$$

marginalization:

$\{\alpha\}_i$: all parameter of model M_i

$$\boldsymbol{P(D|M_i) = \int P(D|\{\alpha\}_i\, M_i)\; P(\{\alpha\}_i|\, M_i)\;\, d\Omega_{\{\alpha\}_i}}$$

$$= \int P(D|\{\alpha\}_i, M_i) \prod_j P(\alpha_{ij}|\, M_i)\;\, d\alpha_{ij}$$

assuming all α_{ij} are mutually independent
(Naïve Bayes)

Bayesian Methods:

goal:

$$\rho = \frac{P(\mathrm{M_A}|D)}{P(\mathrm{M_B}|D)} = \frac{\boxed{P(D|M_A)}\ P(M_A)}{P(D)} \cdot \frac{P(D)}{\boxed{P(D|M_B)}P(M_B)}$$

D : data
$\mathbf{M_A}$: model A
$\boldsymbol{M_B}$: model B
$\{\alpha\}_i$: all parameter of model $\boldsymbol{M_i}$

marginalization:

$$\boldsymbol{P(D|M_i) = \int P(D|\{\alpha\}_i\, M_i)\, P(\{\alpha\}_i|\, M_i)\ d\Omega_{\{\alpha\}_i}}$$

$$= \int \underbrace{P(D|\{\alpha\}_i, M_i)} \prod_j \underbrace{P(\alpha_{ij}|\, M_i)}\ d\alpha_{ij}$$

assuming all α_{ij} are mutually independent **(Naïve Bayes)**

likelihood function
→ the actual model

prior of $\boldsymbol{\alpha_{ij}}$ BEVORE(!) measurement
Maximum Entropy without prior knowledge:

$$\frac{1}{\boldsymbol{\alpha_{ij}}(max) - \boldsymbol{\alpha_{ij}}(end)}$$

Bayesian Methods:

goal:

$$\rho = \frac{P(\mathrm{M_A}|D)}{P(\mathrm{M_B}|D)} = \frac{\boxed{P(D|M_A)}\; P(M_A)}{P(D)} \cdot \frac{P(D)}{\boxed{P(D|M_B)}P(M_B)}$$

- $\mathbf{D}$: data
- $\mathbf{M_A}$: model A
- $\boldsymbol{M_B}$: model B
- $\{\boldsymbol{\alpha}\}_i$: all parameter of model $\boldsymbol{M_i}$

marginalization:

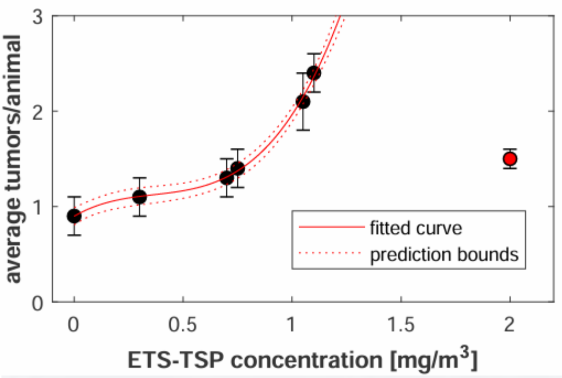
$$\boldsymbol{P(D|M_i) = \int P(D|\{\alpha\}_i\, M_i)\, P(\{\alpha\}_i|\, M_i)\;\, d\Omega_{\{\alpha\}_i}}$$

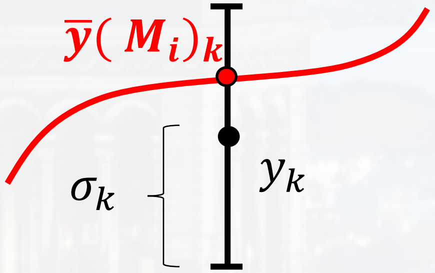
$$= \int P(D|\{\alpha\}_i, M_i) \prod_j P(\alpha_{ij}|\, M_i)\;\, d\alpha_{ij}$$

$$= \prod_j \frac{1}{\boldsymbol{\alpha_{ij}}(max) - \boldsymbol{\alpha_{ij}}(min)} \cdot \int P(D|\{\alpha\}_i\,, M_i)\;\;\, d\alpha_{ij}$$

assuming all α_{ij} are mutually independent **(Naïve Bayes)**

likelihood function
→ the actual model





- $\boldsymbol{y_k}$: measured value
- $\boldsymbol{\sigma_k}$: error
- $\bar{\mathbf{y}}(\mathbf{M_i})_\mathbf{k}$: model value (after fit)

Bayesian Methods:

goal:

$$\rho = \frac{P(\mathrm{M_A}|D)}{P(\mathrm{M_B}|D)} = \frac{\boxed{P(D|M_A)}\, P(M_A)}{P(D)} \cdot \frac{P(D)}{\boxed{P(D|M_B)}\, P(M_B)}$$

D	**: data**
$\mathbf{M_A}$	**: model A**
$\boldsymbol{M_B}$	**: model B**
$\{\alpha\}_i$	**: all parameter of model $\boldsymbol{M_i}$**
$\boldsymbol{y_k}$	**: measured value**
$\boldsymbol{\sigma_k}$	**: error**
$\bar{\mathrm{y}}(\mathbf{M_i})_\mathrm{k}$	**: model value (after fit)**

marginalization:

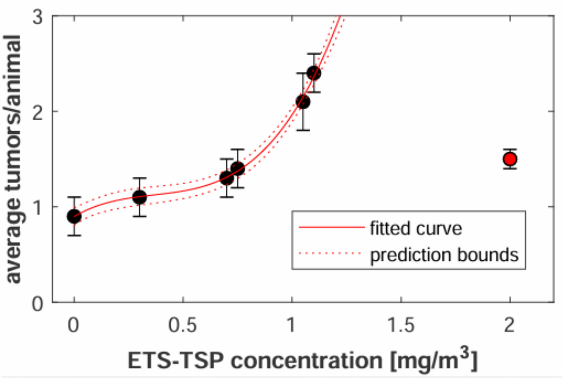
$$\boldsymbol{P(D|M_i) = \int P(D|\{\alpha\}_i\; M_i)\; P(\{\alpha\}_i|\; M_i)\;\; d\Omega_{\{\alpha\}_i}}$$

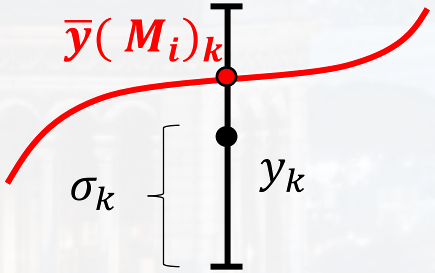
$$= \int P(D|\{\alpha\}_i, M_i) \prod_j P(\alpha_{ij}|\, M_i)\;\, d\alpha_{ij}$$

$$= \prod_j \frac{1}{\boldsymbol{\alpha_{ij}}(max) - \boldsymbol{\alpha_{ij}}(min)} \cdot \int P(D|\{\alpha\}_i\, , M_i)\;\; d\alpha_{ij}$$

assuming all α_{ij} are mutually independent **(Naïve Bayes)**

likelihood function
→ the actual model





$$P\big(y_k|\alpha_{ij}, M_i\big) \approx \frac{1}{\sqrt{2\pi\sigma_k^2}}\; e^{-\frac{1}{2}\frac{(\bar{\boldsymbol{y}}(\boldsymbol{M_i})_{\boldsymbol{k}} - y_k)^2}{\sigma_k^2}}$$

for $\sigma_k \ll |y_k|$

Bayesian Methods:

D	: data
$\mathbf{M_A}$	: model A
$\boldsymbol{M_B}$	: model B
$\{\alpha\}_i$	: all parameter of model $\boldsymbol{M_i}$
$\boldsymbol{y_k}$	: measured value
$\boldsymbol{\sigma_k}$	: error
$\bar{\mathrm{y}}(\mathrm{M_i})_\mathrm{k}$	: model value (after fit)

marginalization:

$$\boldsymbol{P(D|M_i) = \int P(D|\{\alpha\}_i\, M_i)\, P(\{\alpha\}_i|\, M_i)\;\, d\Omega_{\{\alpha\}_i}}$$

$$= \int P(D|\{\alpha\}_i, M_i) \prod_j P(\alpha_{ij}|\, M_i)\;\, d\alpha_{ij}$$

$$= \prod\nolimits_j \frac{1}{\boldsymbol{\alpha_{ij}}(max)-\boldsymbol{\alpha_{ij}}(min)} \cdot \int P(D|\{\alpha\}_i\,, M_i)\;\; d\alpha_{ij}$$

likelihood function

→ the actual model

$$P(D|\{\alpha\}_i\; M_i) = \prod_k P(y_k|\alpha_{ij}, M_i) \;\; = \prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2}\frac{(\bar{y}(\boldsymbol{M_i})_{\boldsymbol{k}} - y_k)^2}{\sigma_k^2}}$$

$$= \left(\prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}}\right) \cdot e^{-\frac{1}{2}\;\sum_k \frac{(\bar{y}(\boldsymbol{M_i})_{\boldsymbol{k}} - \boldsymbol{y_k})^2}{\boldsymbol{\sigma_k^2}}} \qquad = \left(\prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}}\right) \cdot e^{-\frac{1}{2}\chi_i^2}$$

Bayesian Methods:

$$\rho = \frac{P(D|M_A)\, P(M_A)}{P(D|M_B)\, P(M_B)}$$

$$\boxed{= \underbrace{\frac{P(M_A)}{P(M_B)}}_{} \cdot \underbrace{\frac{\int e^{-\frac{1}{2}\chi_A^2}\, d\Omega_{\{\alpha\}_A}}{\int e^{-\frac{1}{2}\chi_B^2}\, d\Omega_{\{\alpha\}_B}}}_{} \cdot \underbrace{\frac{\prod_j \boldsymbol{\alpha_{iB}}(max) - \boldsymbol{\alpha_{iB}}(min)}{\prod_j \boldsymbol{\alpha_{iA}}(max) - \boldsymbol{\alpha_{iA}}(min)}}_{}}$$

D	**: data**
$\mathbf{M_A}$	**: model A**
$\boldsymbol{M_B}$	**: model B**
$\{\alpha\}_i$	**: all parameter of model** $\boldsymbol{M_i}$
$\boldsymbol{y_k}$	**: measured value**
$\boldsymbol{\sigma_k}$	**: error**
$\bar{\mathbf{y}}(\mathbf{M_i})_\mathbf{k}$	**: model value (after fit)**

prior probability of each model: maximum entropy → 1:1

fit quality: integral over χ^2

Occam's Razor: simple models are prefered

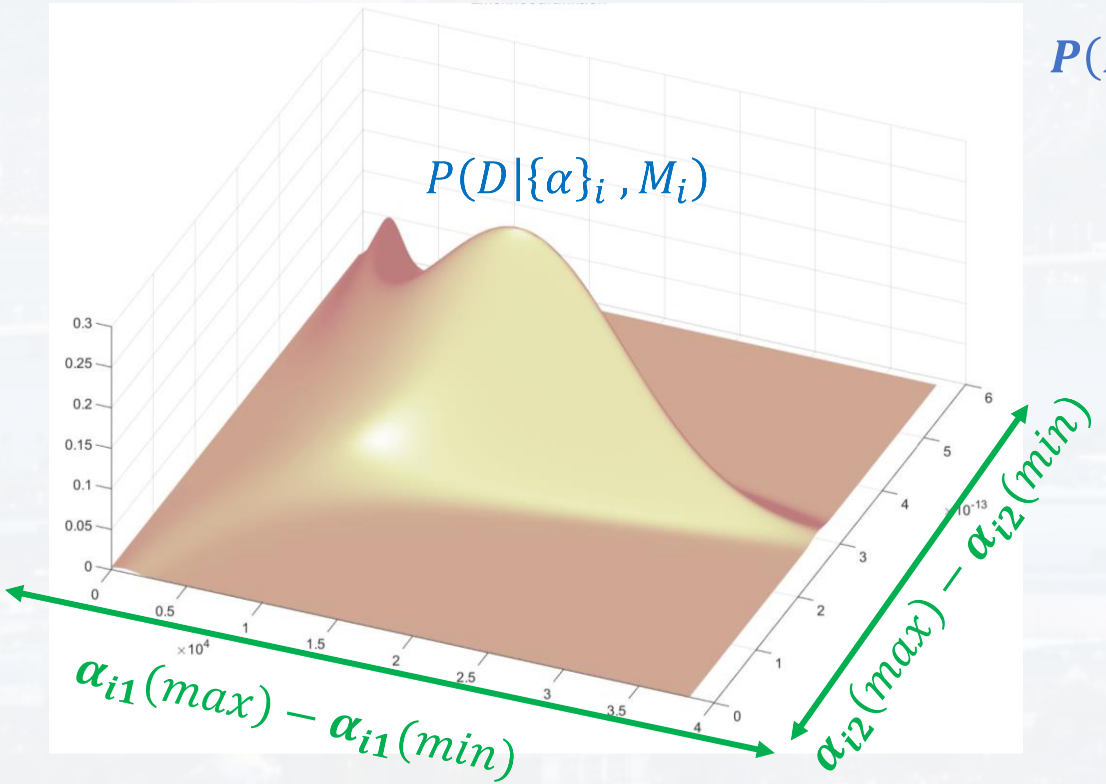
Bayesian Methods:

Model Selection

$$\rho = \frac{P(D|M_A)\,P(M_A)}{P(D|M_B)\,P(M_B)}$$

$$= \frac{P(M_A)}{P(M_B)} \cdot \boxed{\frac{\int e^{-\frac{1}{2}\chi_A^2}\,d\Omega_{\{\alpha\}_A}}{\int e^{-\frac{1}{2}\chi_B^2}\,d\Omega_{\{\alpha\}_B}} \cdot \frac{\prod_j \boldsymbol{\alpha_{iB}}(max) - \boldsymbol{\alpha_{iB}}(min)}{\prod_j \boldsymbol{\alpha_{iA}}(max) - \boldsymbol{\alpha_{iA}}(min)}}$$

D	**: data**
$\mathbf{M_A}$	**: model A**
$\boldsymbol{M_B}$	**: model B**
$\{\alpha\}_i$	**: all parameter of model $\boldsymbol{M_i}$**
$\boldsymbol{y_k}$	**: measured value**
$\boldsymbol{\sigma_k}$	**: error**
$\bar{y}(\mathrm{M_i})_k$	**: model value (after fit)**



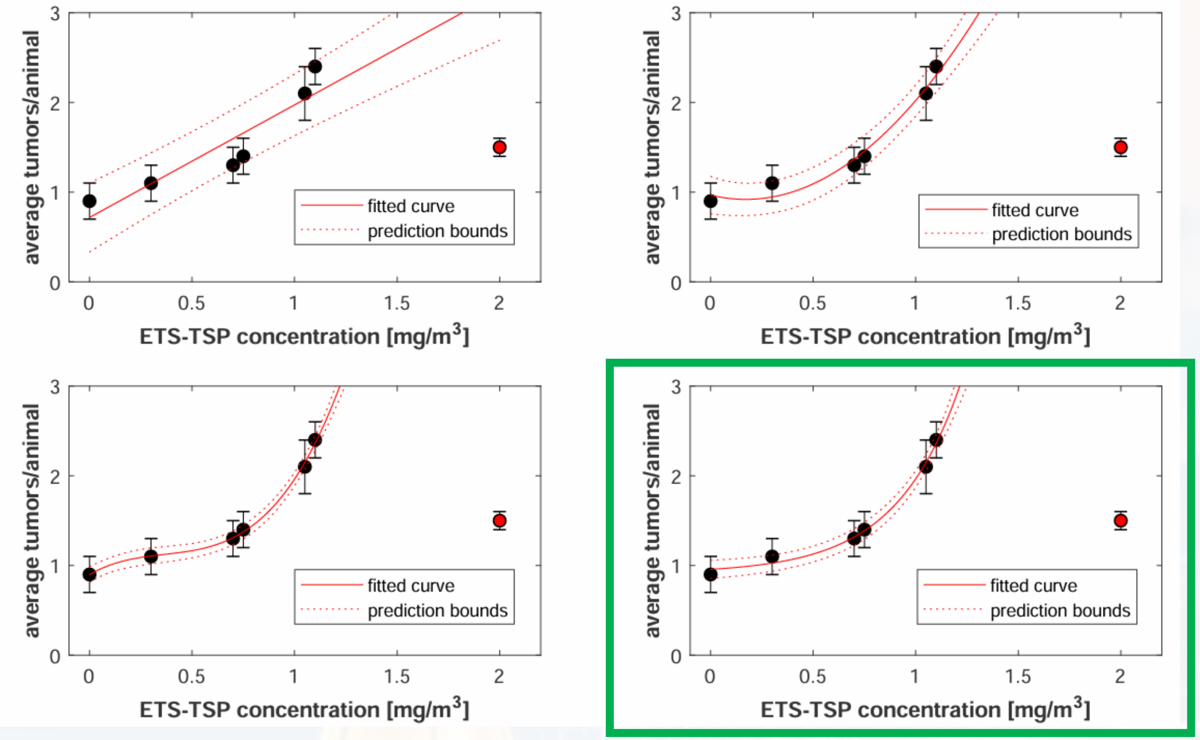
$$\boldsymbol{P(D|M_i) = \int P(D|\{\alpha\}_i\, M_i)\; P(\{\alpha\}_i|\, M_i)\;\; d\Omega_{\{\alpha\}_i}}$$

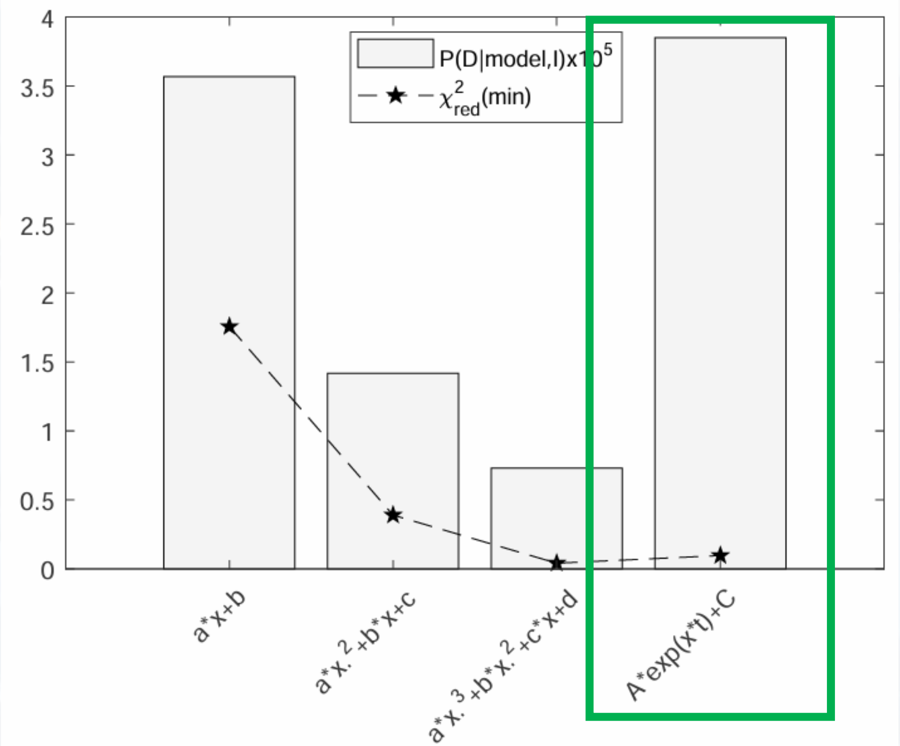
$$= \int P(D|\{\alpha\}_i, M_i) \prod_j P(\alpha_{ij}|\, M_i)\;\; d\alpha_{ij}$$

$$= \prod_j \frac{1}{\boldsymbol{\alpha_{ij}}(max) - \boldsymbol{\alpha_{ij}}(min)} \cdot \int P(D|\{\alpha\}_i\,, M_i)\;\; d\alpha_{ij}$$

Bayesian Methods:

Model Selection





Thank you for your attention