# Evidence in favor of a scientific theory
## With great complexity comes great honesty

Gustavo Landfried
@GALandfried

MSc in Anthropological Sciences
PhD student in Computer Sciences

DEPARTAMENTO DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Laboratorio de Inteligencia Artificial Aplicada

## Why Bayesian inference?

Allows us to optimally update a priori beliefs given a model and data.

# Where comes from?

| | Not infected | Infected | |
|---|---|---|---|
| Not vaccinated | 4 | 2 | 6 |
| Vaccinated | 76 | 18 | 94 |
| | 80 | 20 | 100 |

From conditional probability

# Where comes from?

| | Not infected | Infected | |
|---|---|---|---|
| Not vaccinated | 4 | 2 | 6 |
| Vaccinated | 76 | 18 | 94 |
| | 80 | 20 | 100 |

## From conditional probability

$$P(\text{Not infected}|\text{Vaccinated}) = \frac{P(\text{Vaccinated} \cap \text{Not infected})}{P(\text{Vaccinated})}$$

# Where comes from?

| | Not infected | Infected | |
|---|---|---|---|
| Not vaccinated | 4 | 2 | 6 |
| Vaccinated | 76 | 18 | 94 |
| | 80 | 20 | 100 |

## From conditional probability

$$P(\text{Not infected}|\text{Vaccinated}) = \frac{P(\text{Vaccinated} \cap \text{Not infected})}{P(\text{Vaccinated})}$$

Bayes theorem:

$$P(A_1|B_1) = \frac{P(B_1 \cap A_1)}{P(B_1)} = \frac{P(B_1|A_1)P(A_1)}{P(B_1)} \tag{1}$$

# Scientific test example

There is a test that correctly detects zombies $95\%$ of the time.

- $P(\text{positive}|\text{zombie}) = 0.95$

# Scientific test example

There is a test that correctly detects zombies $95\%$ of the time.

- $P(\text{positive}|\text{zombie}) = 0.95$

One percent of the time it incorrectly detect normal persons as zombies.

- $P(\text{positive}|\text{mortal}) = 0.01$

# Scientific test example

There is a test that correctly detects zombies $95\%$ of the time.

- $P(\text{positive}|\text{zombie}) = 0.95$

One percent of the time it incorrectly detect normal persons as zombies.

- $P(\text{positive}|\text{mortal}) = 0.01$

We know that zombies are only $0.1\%$ of the population.

- $P(\text{zombie}) = 0.001$

# Scientific test example

There is a test that correctly detects zombies $95\%$ of the time.

- $P(\text{positive}|\text{zombie}) = 0.95$

One percent of the time it incorrectly detect normal persons as zombies.

- $P(\text{positive}|\text{mortal}) = 0.01$

We know that zombies are only $0.1\%$ of the population.

- $P(\text{zombie}) = 0.001$

Someone receive a positive test:

# Scientific test example

There is a test that correctly detects zombies $95\%$ of the time.
- $P(\text{positive}|\text{zombie}) = 0.95$

One percent of the time it incorrectly detect normal persons as zombies.
- $P(\text{positive}|\text{mortal}) = 0.01$

We know that zombies are only $0.1\%$ of the population.
- $P(\text{zombie}) = 0.001$

Someone receive a positive test:
She has **only 8.7% chance** to actually be a zombie!?

$$P(\text{zombie}|\text{positive}) = \frac{P(\text{positive}|\text{zombie})P(\text{zombie})}{P(\text{positive})}$$

# Scientific test example

There is a test that correctly detects zombies $95\%$ of the time.

- $P(\text{positive}|\text{zombie}) = 0.95$

One percent of the time it incorrectly detect normal persons as zombies.

- $P(\text{positive}|\text{mortal}) = 0.01$

We know that zombies are only $0.1\%$ of the population.

- $P(\text{zombie}) = 0.001$

Someone receive a positive test:
She has **only 8.7% chance** to actually be a zombie!?

$$P(\text{zombie}|\text{positive}) = \frac{P(\text{positive}|\text{zombie})P(\text{zombie})}{P(\text{positive})}$$

In this example all frequencies were observables

# The inferential jump

**Bayesian inference is about hidden variables**

About our **belief distributions** of those hidden variables!

# The inferential jump

**Bayesian inference is about hidden variables**
About our **belief distributions** of those hidden variables!

$$\underbrace{P(\text{Belief}|\text{Data})}_{\text{Posterior}} = \frac{\overbrace{P(\text{Data}|\text{Belief})}^{\text{Likelihood}}\,\overbrace{P(\text{Belief})}^{\text{Prior}}}{\underbrace{P(\text{Data})}_{\substack{\text{Evidence or}\\ \text{Average likelihood}}}}$$

# The inferential jump

**Bayesian inference is about hidden variables**
About our **belief distributions** of those hidden variables!

$$\underbrace{P(\text{Belief}|\text{Data})}_{\text{Posterior}} = \frac{\overbrace{P(\text{Data}|\text{Belief})}^{\text{Likelihood}}\,\overbrace{P(\text{Belief})}^{\text{Prior}}}{\underbrace{P(\text{Data})}_{\substack{\text{Evidence or}\\ \text{Average likelihood}}}}$$

A model is always there!

$$\underbrace{P(\text{Belief}|\text{Data},\text{Model})}_{\text{Posterior}} = \frac{\overbrace{P(\text{Data}|\text{Belief},\text{Model})}^{\text{Likelihood}}\,\overbrace{P(\text{Belief}|\text{Model})}^{\text{Prior}}}{\underbrace{P(\text{Data}|\text{Model})}_{\substack{\text{Evidence or}\\ \text{Average likelihood}}}}$$

- **Prior** belief (distribution):

$$P(B|M) = \frac{1}{\#\text{Beliefs}} \qquad \forall B \in \text{Beliefs}$$

- **Prior** belief (distribution):

$$P(B|M) = \frac{1}{\#\text{Beliefs}} \qquad \forall B \in \text{Beliefs}$$

- **Likelihood** or ways in which data may have been generated (distribution):

$$P(D|B,M) = \frac{\text{Ways to produce } D \text{ given } B \text{ and } M}{\text{Total ways given } B \text{ and } M} \qquad \forall B \in \text{Beliefs}$$

- **Prior** belief (distribution):

$$P(B|M) = \frac{1}{\#\text{Beliefs}} \qquad \forall B \in \text{Beliefs}$$

- **Likelihood** or ways in which data may have been generated (distribution):

$$P(D|B,M) = \frac{\text{Ways to produce } D \text{ given } B \text{ and } M}{\text{Total ways given } B \text{ and } M} \qquad \forall B \in \text{Beliefs}$$

- **Evidence** or Average likelihood (scalar):

$$P(D|M) = \sum_{B \in \text{Beliefs}} \underbrace{P(D|B,M)}_{\text{likelihood}} \underbrace{P(B|M)}_{\text{prior}}$$

- **Prior** belief (distribution):

$$P(B|M) = \frac{1}{\#\text{Beliefs}} \qquad \forall B \in \text{Beliefs}$$

- **Likelihood** or ways in which data may have been generated (distribution):

$$P(D|B, M) = \frac{\text{Ways to produce } D \text{ given } B \text{ and } M}{\text{Total ways given } B \text{ and } M} \qquad \forall B \in \text{Beliefs}$$

- **Evidence** or Average likelihood (scalar):

$$P(D|M) = \sum_{B \in \text{Beliefs}} \underbrace{P(D|B, M)}_{\text{likelihood}} \underbrace{P(B|M)}_{\text{prior}}$$

- **Posterior** belief (distribution):

$$P(B|D, M) = \frac{P(D|B, M)P(B|M)}{P(D|M)} \qquad \forall B \in \text{Beliefs}$$

# The garden of forking paths

To update our beliefs (posterior), we need to consider every possible path in the model that could have lead us to the observed data (likelihood).

# The garden of forking paths

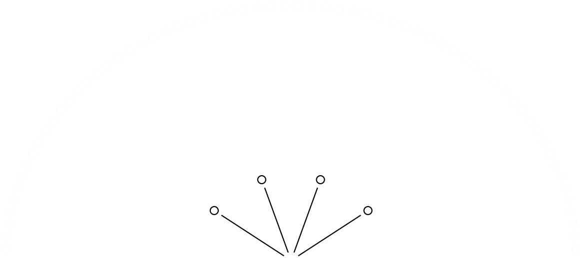
Data (D): ● ○ ● Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$

# The garden of forking paths

Data (D): ● ○ ●    Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$

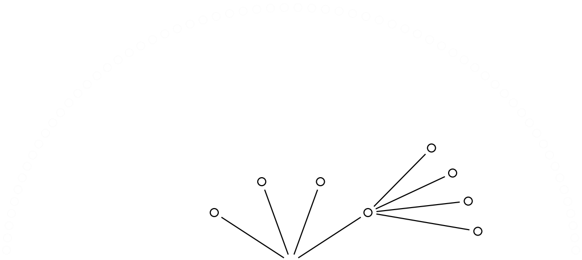


Ways given $M$ and $B =$ ○○○○    (First marbel)

# The garden of forking paths

Data (D): ● ○ ●    Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$

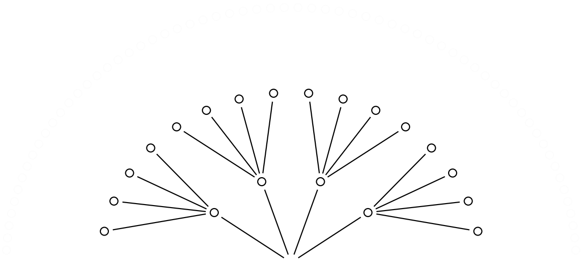


Ways given $M$ and $B = \circ\circ\circ\circ$    (Second marbel)

# The garden of forking paths

Data (D): ● ○ ●    Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data ∼ Binomial$(n, p)$

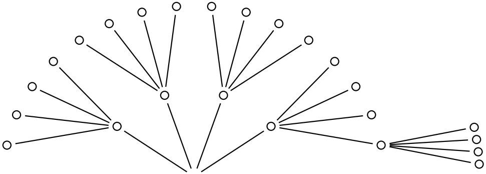


Ways given $M$ and $B =$ ○○○○    (Second marbel)

# The garden of forking paths

Data (D): ● ○ ● Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$

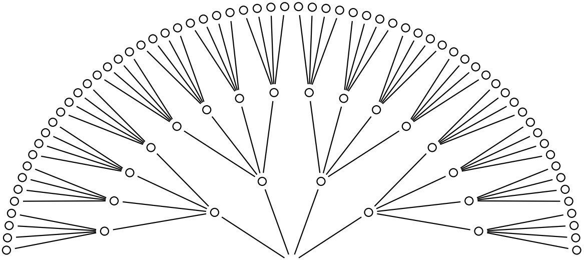


Ways given $M$ and $B = \circ\circ\circ\circ$ (Third marbel)

# The garden of forking paths

Data (D): ● ○ ●    Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$

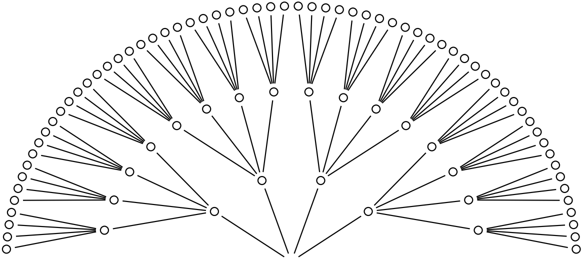


Ways given $M$ and $B =$ ○○○○    (Third marbel)

# The garden of forking paths

Data (D): ● ○ ●  Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



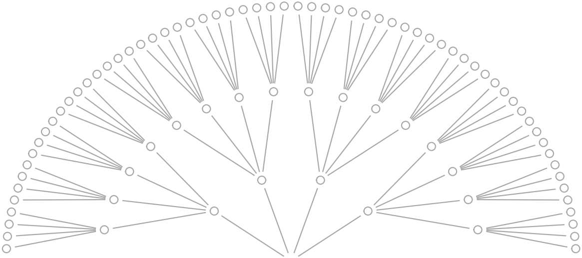
Ways given $M$ and $B =$ ○○○○

| Belief | Ways to produce ● ○ ● |
|---|---|
| ○○○○ | |

# The garden of forking paths

Data (D): ● ○ ● Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



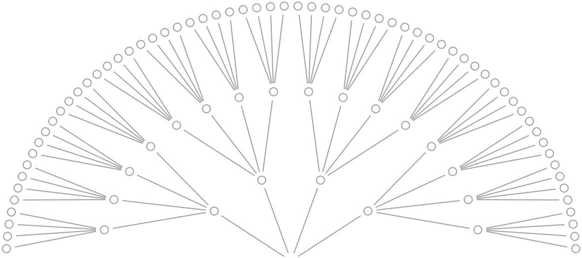
Ways given $M$ and $B =$ ○○○○

| Belief | Ways to produce ● ○ ● |
|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ |

# The garden of forking paths

Data (D): ● ○ ● Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



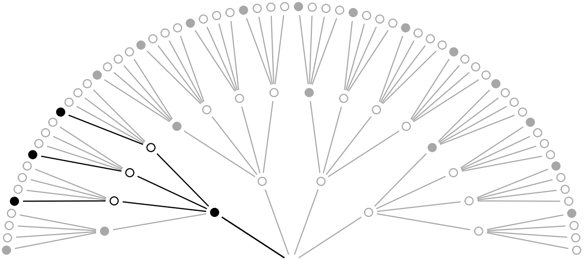
Ways given $M$ and $B =$ ○○○○

| Belief | Ways to produce ● ○ ● | Likelihood | Prior | Posterior $\propto$ |
|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0\times4\times0}{4\times4\times4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ |

# The garden of forking paths

Data (D): ● ○ ●  Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



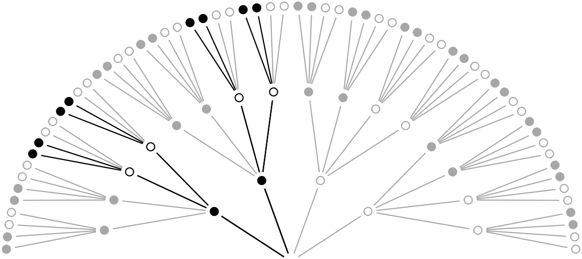
Ways given $M$ and $B =$ ●○○○

| Belief | Ways to produce ● ○ ● | Likelihood | Prior | Posterior $\propto$ |
|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0\times4\times0}{4\times4\times4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ |
| ●○○○ | $1 \times 3 \times 1 = 3$ | $3/64$ | $1/5$ | $\frac{3}{64}\,\frac{1}{5}$ |

# The garden of forking paths

Data (D): ● ○ ● Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



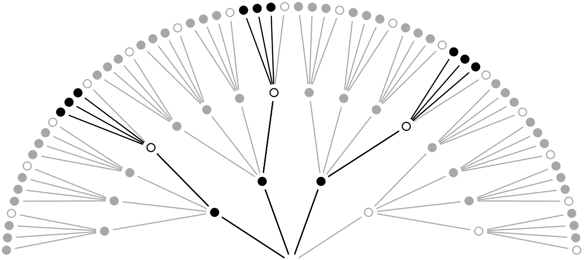
Ways given $M$ and $B =$ ●●○○

| Belief | Ways to produce ● ○ ● | Likelihood | Prior | Posterior $\propto$ |
|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0\times4\times0}{4\times4\times4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ |
| ●○○○ | $1 \times 3 \times 1 = 3$ | $3/64$ | $1/5$ | $\frac{3}{64}\,\frac{1}{5}$ |
| ●●○○ | $2 \times 2 \times 2 = 8$ | $8/64$ | $1/5$ | $\frac{8}{64}\,\frac{1}{5}$ |

# The garden of forking paths

Data (D): ● ○ ●  Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



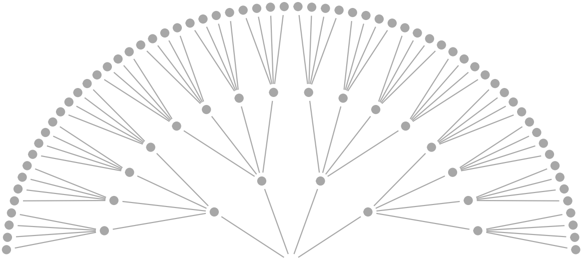
Ways given $M$ and $B =$ ●●●○

| Belief | Ways to produce ● ○ ● | Likelihood | Prior | Posterior $\propto$ |
|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0\times4\times0}{4\times4\times4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ |
| ●○○○ | $1 \times 3 \times 1 = 3$ | $3/64$ | $1/5$ | $\frac{3}{64}\,\frac{1}{5}$ |
| ●●○○ | $2 \times 2 \times 2 = 8$ | $8/64$ | $1/5$ | $\frac{8}{64}\,\frac{1}{5}$ |
| ●●●○ | $3 \times 1 \times 3 = 9$ | $9/64$ | $1/5$ | $\frac{9}{64}\,\frac{1}{5}$ |

# The garden of forking paths

Data (D): ●○●  Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



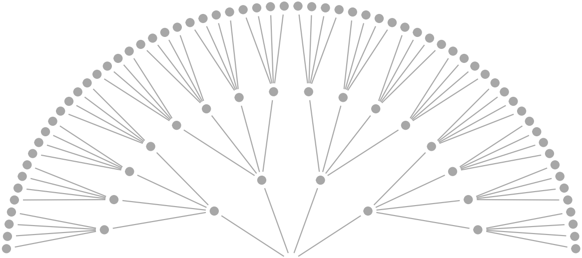
Ways given $M$ and $B =$ ●●●●

| Belief | Ways to produce ●○● | Likelihood | Prior | Posterior $\propto$ |
|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0\times4\times0}{4\times4\times4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ |
| ●○○○ | $1 \times 3 \times 1 = 3$ | $3/64$ | $1/5$ | $\frac{3}{64}\,\frac{1}{5}$ |
| ●●○○ | $2 \times 2 \times 2 = 8$ | $8/64$ | $1/5$ | $\frac{8}{64}\,\frac{1}{5}$ |
| ●●●○ | $3 \times 1 \times 3 = 9$ | $9/64$ | $1/5$ | $\frac{9}{64}\,\frac{1}{5}$ |
| ●●●● | $4 \times 0 \times 4 = 0$ | $0/64$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ |

# The garden of forking paths

Data (D): ● ○ ●    Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



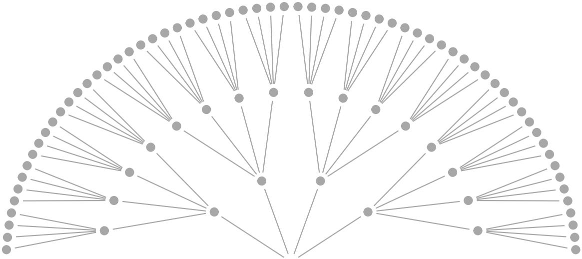
Ways given $M$ and $B =$ ●●●●

| Belief | Ways to produce ● ○ ● | Likelihood | Prior | Posterior $\propto$ |
|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0\times4\times0}{4\times4\times4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ |
| ●○○○ | $1 \times 3 \times 1 = 3$ | $3/64$ | $1/5$ | $\frac{3}{64}\,\frac{1}{5}$ |
| ●●○○ | $2 \times 2 \times 2 = 8$ | $8/64$ | $1/5$ | $\frac{8}{64}\,\frac{1}{5}$ |
| ●●●○ | $3 \times 1 \times 3 = 9$ | $9/64$ | $1/5$ | $\frac{9}{64}\,\frac{1}{5}$ |
| ●●●● | $4 \times 0 \times 4 = 0$ | $0/64$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ |
| | | | | $P(D\|M)$ |

# The garden of forking paths

Data (D): ● ○ ●  Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



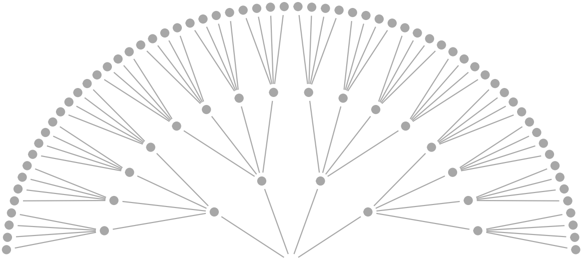
Ways given $M$ and $B =$ ●●●●

| Belief | Ways to produce ● ○ ● | Likelihood | Prior | Posterior $\propto$ |
|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0 \times 4 \times 0}{4 \times 4 \times 4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64}\frac{1}{5}$ |
| ●○○○ | $1 \times 3 \times 1 = 3$ | $3/64$ | $1/5$ | $\frac{3}{64}\frac{1}{5}$ |
| ●●○○ | $2 \times 2 \times 2 = 8$ | $8/64$ | $1/5$ | $\frac{8}{64}\frac{1}{5}$ |
| ●●●○ | $3 \times 1 \times 3 = 9$ | $9/64$ | $1/5$ | $\frac{9}{64}\frac{1}{5}$ |
| ●●●● | $4 \times 0 \times 4 = 0$ | $0/64$ | $1/5$ | $\frac{0}{64}\frac{1}{5}$ |
| | | | | $\frac{3+8+9}{64 \cdot 5}$ |

# The garden of forking paths

Data (D): ● ○ ●  Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



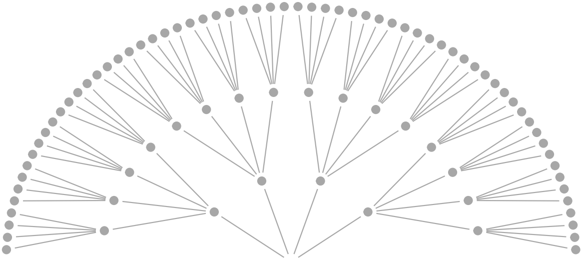
Ways given $M$ and $B =$ ●●●●

| Belief | Ways to produce ● ○ ● | Likelihood | Prior | Posterior $\propto$ | Posterior |
|---|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0\times4\times0}{4\times4\times4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ | $\frac{0}{64}\,\frac{1}{5}\,\frac{64\cdot5}{3+8+9}$ |
| ●○○○ | $1 \times 3 \times 1 = 3$ | $3/64$ | $1/5$ | $\frac{3}{64}\,\frac{1}{5}$ | |
| ●●○○ | $2 \times 2 \times 2 = 8$ | $8/64$ | $1/5$ | $\frac{8}{64}\,\frac{1}{5}$ | |
| ●●●○ | $3 \times 1 \times 3 = 9$ | $9/64$ | $1/5$ | $\frac{9}{64}\,\frac{1}{5}$ | |
| ●●●● | $4 \times 0 \times 4 = 0$ | $0/64$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ | |
| | | | | $\frac{3+8+9}{64\cdot5}$ | |

# The garden of forking paths

Data (D): ●○●   Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



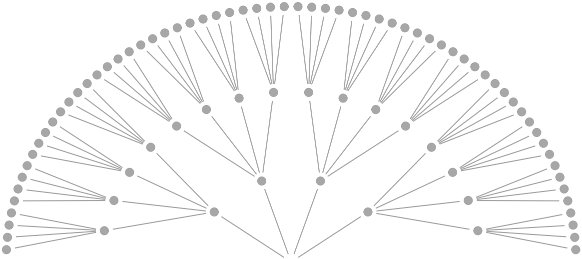
Ways given $M$ and $B =$ ●●●●

| Belief | Ways to produce ●○● | Likelihood | Prior | Posterior $\propto$ | Posterior |
|---|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0\times4\times0}{4\times4\times4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ | $\frac{0}{3+8+9} = 0.00$ |
| ●○○○ | $1 \times 3 \times 1 = 3$ | $3/64$ | $1/5$ | $\frac{3}{64}\,\frac{1}{5}$ | |
| ●●○○ | $2 \times 2 \times 2 = 8$ | $8/64$ | $1/5$ | $\frac{8}{64}\,\frac{1}{5}$ | |
| ●●●○ | $3 \times 1 \times 3 = 9$ | $9/64$ | $1/5$ | $\frac{9}{64}\,\frac{1}{5}$ | |
| ●●●● | $4 \times 0 \times 4 = 0$ | $0/64$ | $1/5$ | $\frac{0}{64}\,\frac{1}{5}$ | |
| | | | | $\frac{3+8+9}{64\cdot5}$ | |

# The garden of forking paths

Data (D): ● ○ ●　　Beliefs (B): ○○○○, ●○○○, ●●○○, ●●●○, ●●●●

Model (M): Data $\sim$ Binomial$(n, p)$



Ways given $M$ and $B =$ ●●●●

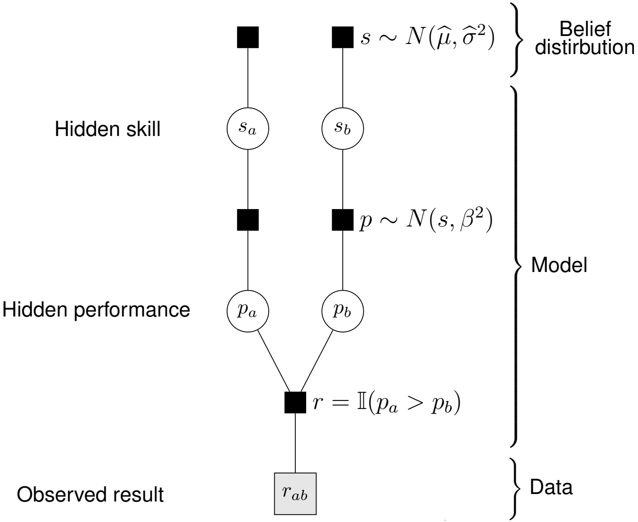
| Belief | Ways to produce ● ○ ● | Likelihood | Prior | Posterior $\propto$ | Posterior |
|---|---|---|---|---|---|
| ○○○○ | $0 \times 4 \times 0 = 0$ | $\frac{0 \times 4 \times 0}{4 \times 4 \times 4} = \frac{0}{64}$ | $1/5$ | $\frac{0}{64} \frac{1}{5}$ | $\frac{0}{3+8+9} = 0.00$ |
| ●○○○ | $1 \times 3 \times 1 = 3$ | $3/64$ | $1/5$ | $\frac{3}{64} \frac{1}{5}$ | $\frac{3}{3+8+9} = 0.15$ |
| ●●○○ | $2 \times 2 \times 2 = 8$ | $8/64$ | $1/5$ | $\frac{8}{64} \frac{1}{5}$ | $\frac{8}{3+8+9} = 0.40$ |
| ●●●○ | $3 \times 1 \times 3 = 9$ | $9/64$ | $1/5$ | $\frac{9}{64} \frac{1}{5}$ | $\frac{9}{3+8+9} = 0.45$ |
| ●●●● | $4 \times 0 \times 4 = 0$ | $0/64$ | $1/5$ | $\frac{0}{64} \frac{1}{5}$ | $\frac{0}{3+8+9} = 0.00$ |
| | | | | $\frac{3+8+9}{64 \cdot 5}$ | |

# Bayesian skill estimator

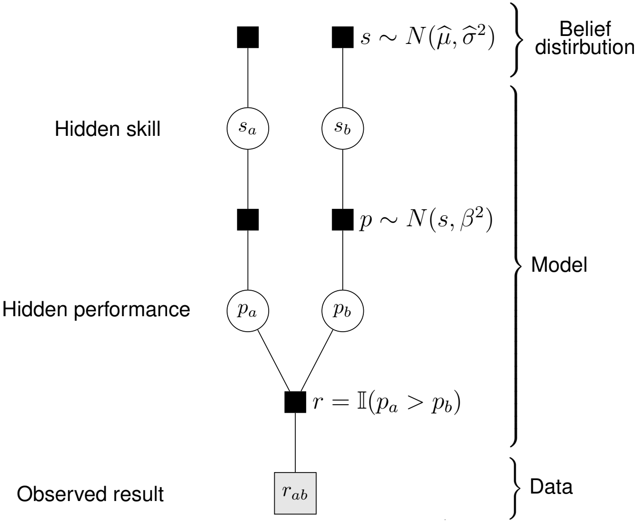
How to estimate skill of players?

Arpad Elo

# Bayesian Elo factor graph



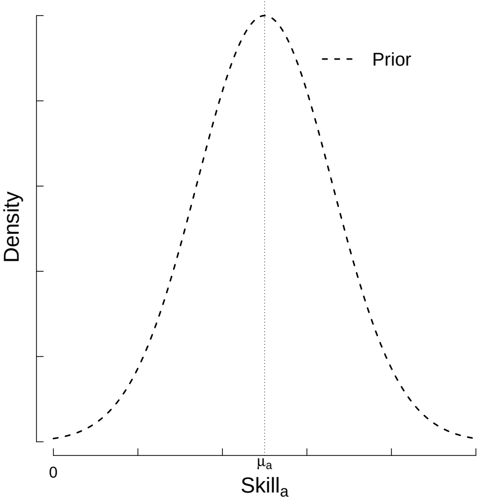
# Bayesian Elo factor graph



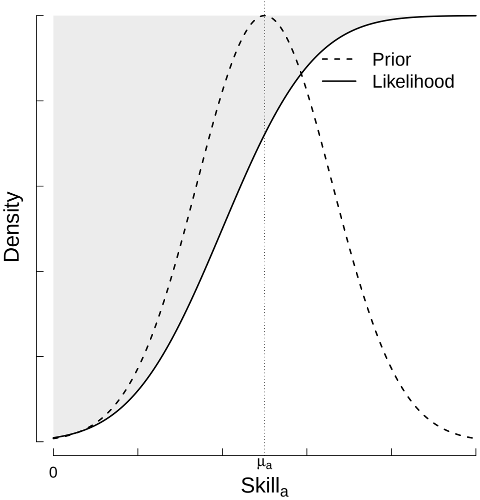

The factor graphs specifies the way to compute the posterior, likelihood, and evidence.
Kschischang FR, Frey BJ, Loeliger HA. Factor graphs and the sum-product algorithm. 2001

$$\overbrace{P(s_a \mid r_{ab}, \text{Elo model})}^{\text{Posterior}} \propto \overbrace{N(s_a \,|\, \widehat{\mu}_a, \widehat{\sigma}_a^2)}^{\text{Prior}} \; \overbrace{1 - \Phi(s_a \,|\widehat{\mu}_b, 2\beta^2 + \widehat{\sigma}_b^2)}^{\text{Likelihood}} \qquad \text{Win case}$$

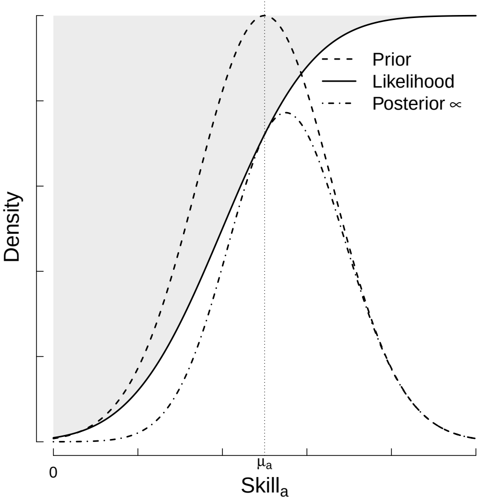
$$\overbrace{P(s_a \mid r_{ab}, \text{Elo model})}^{\text{Posterior}} \propto \overbrace{N(s_a \,|\, \widehat{\mu}_a, \widehat{\sigma}_a^2)}^{\text{Prior}} \; \overbrace{1 - \Phi(s_a \,|\widehat{\mu}_b, 2\beta^2 + \widehat{\sigma}_b^2)}^{\text{Likelihood}} \qquad \text{Win case}$$



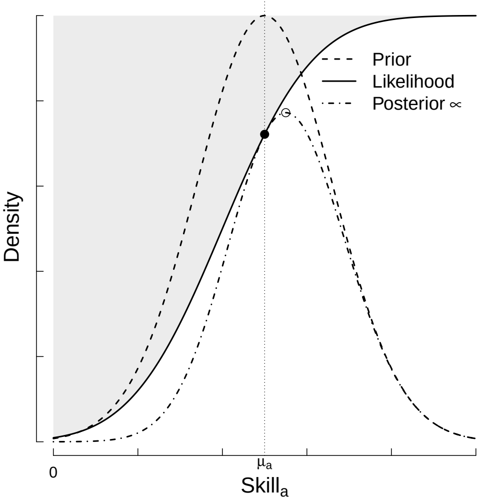
$$\overbrace{P(s_a \mid r_{ab}, \text{Elo model})}^{\text{Posterior}} \propto \overbrace{N(s_a \,|\, \widehat{\mu}_a, \widehat{\sigma}_a^2)}^{\text{Prior}} \; \overbrace{1 - \Phi(s_a \,|\widehat{\mu}_b, 2\beta^2 + \widehat{\sigma}_b^2)}^{\text{Likelihood}} \qquad \text{Win case}$$



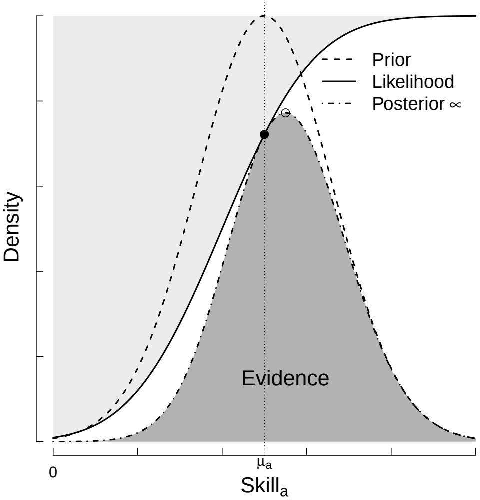
$$\overbrace{P(s_a \mid r_{ab}, \text{Elo model})}^{\text{Posterior}} \propto \overbrace{N(s_a \,|\, \widehat{\mu}_a, \widehat{\sigma}_a^2)}^{\text{Prior}} \; \overbrace{1 - \Phi(s_a \,|\widehat{\mu}_b, 2\beta^2 + \widehat{\sigma}_b^2)}^{\text{Likelihood}} \qquad \text{Win case}$$



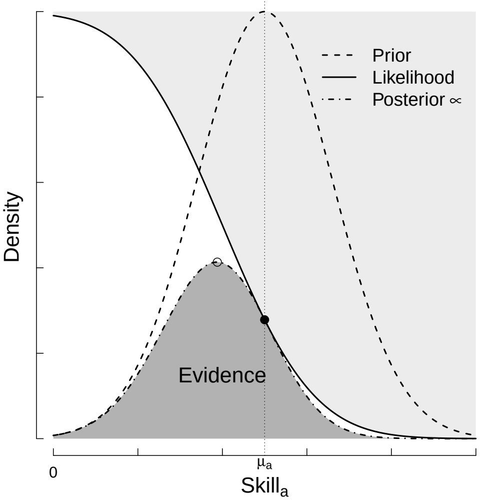
$$\overbrace{P(s_a \mid r_{ab}, \text{Elo model})}^{\text{Posterior}} \propto \overbrace{N(s_a \,|\, \widehat{\mu}_a, \widehat{\sigma}_a^2)}^{\text{Prior}} \; \overbrace{1 - \Phi(s_a \,|\widehat{\mu}_b, 2\beta^2 + \widehat{\sigma}_b^2)}^{\text{Likelihood}} \qquad \text{Win case}$$



$$\overbrace{P(s_a \mid r_{ab}, \text{Elo model})}^{\text{Posterior}} \propto \overbrace{N(s_a \,|\, \widehat{\mu}_a, \widehat{\sigma}_a^2)}^{\text{Prior}} \; \overbrace{1 - \Phi(s_a \,|\widehat{\mu}_b, 2\beta^2 + \widehat{\sigma}_b^2)}^{\text{Likelihood}} \qquad \text{Win case}$$

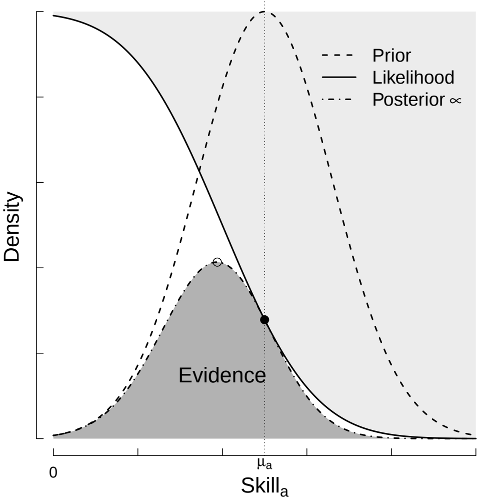


$$\overbrace{P(s_a \mid r_{ab}, \text{Elo model})}^{\text{Posterior}} \propto \overbrace{N(s_a \,|\, \widehat{\mu}_a, \widehat{\sigma}_a^2)}^{\text{Prior}} \; \overbrace{\Phi(s_a \,|\widehat{\mu}_b, 2\beta^2 + \widehat{\sigma}_b^2)}^{\text{Likelihood}} \qquad \text{Loose case}$$



$$\overbrace{P(s_a \mid r_{ab}, \text{Elo model})}^{\text{Posterior}} \propto \overbrace{N(s_a \,|\, \widehat{\mu}_a, \widehat{\sigma}_a^2)}^{\text{Prior}} \; \overbrace{\Phi(s_a \,|\widehat{\mu}_b, 2\beta^2 + \widehat{\sigma}_b^2)}^{\text{Likelihood}}$$

Loose case




For a detailed demostration, see Landfried. TrueSkill: Technical Report. 2019

# Bayesian model inference

- Which are our beliefs about different hidden models $M$?

# Bayesian model inference

- Which are our beliefs about different hidden models $M$?

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{n} P(D|M_i)P(M_i)}$$

# Bayesian model inference

- Which are our beliefs about different hidden models $M$?

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{n} P(D|M_i)P(M_i)}$$

- To compare models we can compute their ratio probability,

# Bayesian model inference

- Which are our beliefs about different hidden models $M$?

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{n} P(D|M_i)P(M_i)}$$

- To compare models we can compute their ratio probability,

$$\text{Bayes factor}(q,r) = \frac{P(M_q|D)}{P(M_r|D)} = \frac{P(D|M_q)P(M_q)}{P(D|M_r)P(M_r)}$$

# Bayesian model inference

- Which are our beliefs about different hidden models $M$?

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{n} P(D|M_i)P(M_i)}$$

- To compare models we can compute their ratio probability,

$$\text{Bayes factor}(q,r) = \frac{P(M_q|D)}{P(M_r|D)} = \frac{P(D|M_q)P(M_q)}{P(D|M_r)P(M_r)}$$

$$\stackrel{*}{=} \underbrace{\frac{P(D|M_q)}{P(D|M_r)}}_{\text{Evidence!}}$$

$*$ With no prior preferences $P(M_q)=P(M_r)$

# Bayesian model inference

- Which are our beliefs about different hidden models $M$?

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{n} P(D|M_i)P(M_i)}$$

- To compare models we can compute their ratio probability,

$$\text{Bayes factor}(q,r) = \frac{P(M_q|D)}{P(M_r|D)} = \frac{P(D|M_q)P(M_q)}{P(D|M_r)P(M_r)}$$

$$\stackrel{*}{=} \underbrace{\frac{P(D|M_q)}{P(D|M_r)}}_{\text{Evidence!}}$$

$*$ With no prior preferences $P(M_q)=P(M_r)$

$$\boxed{P(M_q|D) > P(M_r|D) \stackrel{*}{\iff} P(D|M_q) > P(D|M_r)}$$

# Bayesian model inference

- Which are our beliefs about different hidden models $M$?

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{n} P(D|M_i)P(M_i)}$$

- To compare models we can compute their ratio probability,

$$\text{Bayes factor}(q,r) = \frac{P(M_q|D)}{P(M_r|D)} = \frac{P(D|M_q)P(M_q)}{P(D|M_r)P(M_r)}$$
$$\stackrel{*}{=} \underbrace{\frac{P(D|M_q)}{P(D|M_r)}}_{\text{Evidence!}}$$

$*$ With no prior preferences $P(M_q)=P(M_r)$

$$\boxed{P(M_q|D) > P(M_r|D) \stackrel{*}{\Longleftrightarrow} P(D|M_q) > P(D|M_r)}$$

All you need is evidence

# Bayesian model inference

- Which are our beliefs about different hidden models $M$?

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{i=1}^{n} P(D|M_i)P(M_i)}$$

- To compare models we can compute their ratio probability,

$$\text{Bayes factor}(q,r) = \frac{P(M_q|D)}{P(M_r|D)} = \frac{P(D|M_q)P(M_q)}{P(D|M_r)P(M_r)}$$
$$\stackrel{*}{=} \underbrace{\frac{P(D|M_q)}{P(D|M_r)}}_{\text{Evidence!}}$$

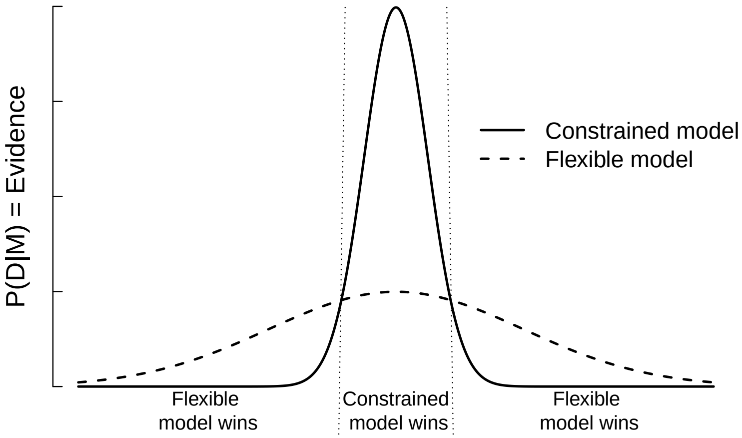
$*$ With no prior preferences $P(M_q)=P(M_r)$

$$P(M_q|D) > P(M_r|D) \stackrel{*}{\iff} P(D|M_q) > P(D|M_r)$$

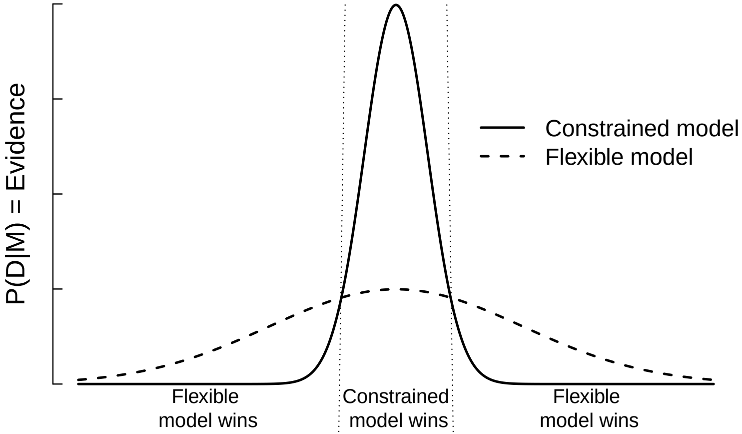
## All you need is evidence

For a dicussion of bayes factor see Kass & Raftery. Bayes factors. 1995.

# Evidence

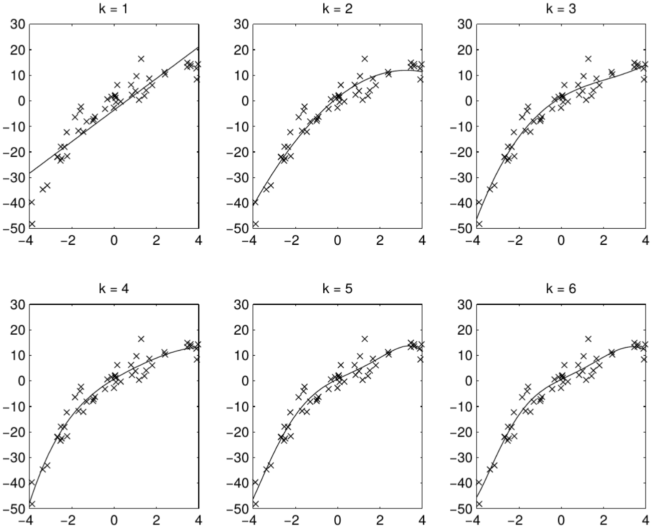

P(D|M) = Evidence
Constrained model
Flexible model
Flexible
model wins
Constrained
model wins
Flexible
model wins

# Evidence

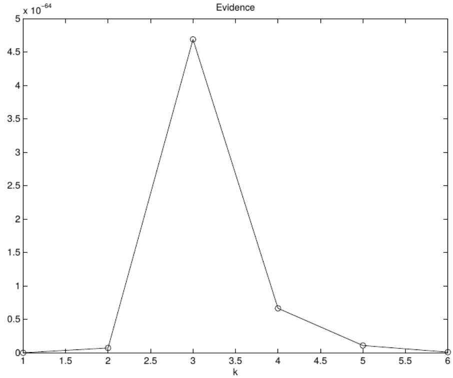


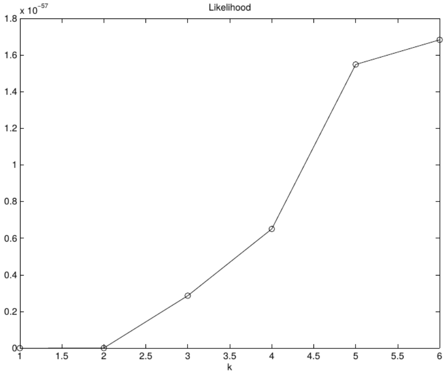

Evience encode a trade-off between complexity and prediction.

# Evidence vs maximum likelihood

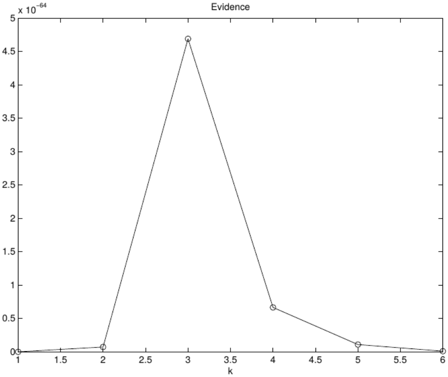


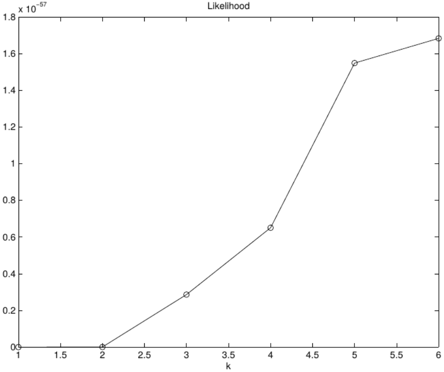
# Evidence vs maximum likelihood



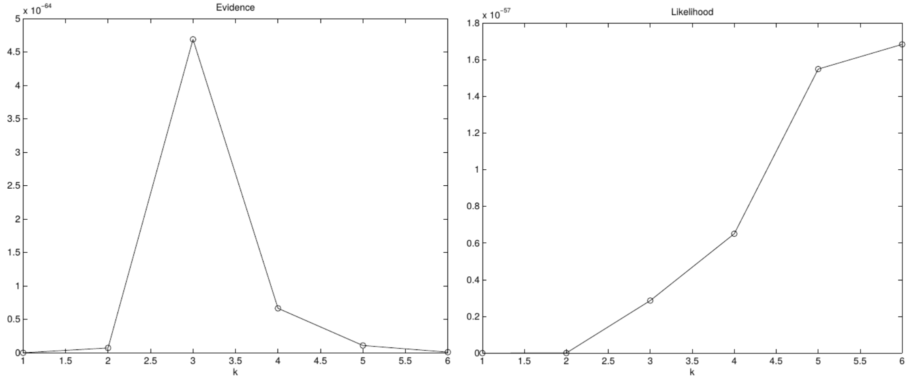
# Evidence vs maximum likelihood




With evidence there is no need for regularization

# Evidence vs maximum likelihood




With evidence there is no need for regularization

For more examples see Tom Minka

# Why evidence is not widely used in machine learning?

# Why evidence is not widely used in machine learning?

First let's take a look at Bayesian no-doubt case

# Bayesian no-doubt case

Fixed beliefs, even with infinite new data

$$\#\text{Beliefs} = 1 \Longrightarrow \underbrace{P(B)}_{\text{Prior}} = \underbrace{P(B|D)}_{\text{Posterior}} \qquad \substack{\forall D \in \text{Data} \\ \forall B \in \text{Beliefs}}$$

# Bayesian no-doubt case

Fixed beliefs, even with infinite new data

$$\#\text{Beliefs} = 1 \implies \underbrace{P(B)}_{\text{Prior}} = \underbrace{P(B|D)}_{\text{Posterior}} \qquad \substack{\forall D \in \text{Data} \\ \forall B \in \text{Beliefs}}$$

Likelihood is just the Evidence

$$\#\text{Beliefs} = 1 \iff \text{Likelihood} = \text{Evidence}$$

Who has no doubt? Who has only one belief?

Who has no doubt? Who has only one belief?

- God (if exists)

Who has no doubt? Who has only one belief?

- God (if exists)
- Mathematicians (and other non-empricial sciences)

Who has no doubt? Who has only one belief?

- God (if exists)
- Mathematicians (and other non-empricial sciences)
- Maybe some extremists

Who has no doubt? Who has only one belief?

- God (if exists)
- Mathematicians (and other non-empricial sciences)
- Maybe some extremists
- **All non-bayesian machine learning**

Who has no doubt? Who has only one belief?

- God (if exists)
- Mathematicians (and other non-empricial sciences)
- Maybe some extremists
- **All non-bayesian machine learning** (the hacked-belief approach)

# The hacked-belief approach

$$\overset{\text{The best belief after seeing the data}}{\cancel{\text{maximum likelihood estimator}}} = \underset{B}{\text{argmax}} P(D|B, M) = \widehat{B}$$

# The hacked-belief approach

$$\overbrace{\cancel{\text{maximum likelihood estimator}}}^{\text{The best belief after seeing the data}} = \underset{B}{\operatorname{argmax}} P(D|B,M) = \widehat{B}$$

$$\overbrace{P(D|M)}^{\text{Hacked }\textbf{evidence}} \quad = \quad \overbrace{P(D|\widehat{B},M)}^{\text{Hacked }\textbf{likelihood}}$$

# The hacked-belief approach

$$\overset{\text{The best belief after seeing the data}}{\cancel{\text{maximum likelihood estimator}}} = \underset{B}{\text{argmax}} P(D|B,M) = \widehat{B}$$

$$\overbrace{P(D|M)}^{\text{Hacked } \textbf{evidence}} = \overbrace{P(D|\widehat{B},M)}^{\text{Hacked } \textbf{likelihood}} = \overbrace{P(\underset{\uparrow}{D}|\underset{B}{\text{argmax}} P(\underset{\uparrow}{D}|B,M),M)}^{\text{Maximum likelihood}}$$

Data appears back and forth!!

# The hacked-belief approach

$$\overset{\text{The best belief after seeing the data}}{\cancel{\text{maximum likelihood estimator}}} = \underset{B}{\text{argmax}} P(D|B,M) = \widehat{B}$$

$$\overbrace{P(D|M)}^{\text{Hacked }\textbf{evidence}} = \overbrace{P(D|\widehat{B},M)}^{\text{Hacked }\textbf{likelihood}} = \overbrace{P(\underset{\uparrow}{D}|\underset{B}{\text{argmax}} P(\underset{\uparrow}{D}|B,M),M)}^{\text{Maximum likelihood}}$$

Data appears back and forth!!

$$\boxed{\text{Hacked evidence (with MLE)} = \text{Maximum likelihood}}$$

# The hacked-belief approach

$$\overset{\text{The best belief after seeing the data}}{\cancel{\text{maximum likelihood estimator}}} = \underset{B}{\text{argmax}} P(D|B,M) = \widehat{B}$$

$$\overbrace{P(D|M)}^{\text{Hacked } \textbf{evidence}} = \overbrace{P(D|\widehat{B},M)}^{\text{Hacked } \textbf{likelihood}} = \overbrace{P(\underset{\uparrow}{D}|\underset{B}{\text{argmax}} P(\underset{\uparrow}{D}|B,M),M)}^{\text{Maximum likelihood}}$$

Data appears back and forth!!

$$\boxed{\text{Hacked evidence (with MLE)} = \text{Maximum likelihood}}$$

With great hacked-belief approach comes great overfitting!

# With great overfitting comes great regularization!

$$\overset{\text{The best belief after seeing the data}}{\cancel{\text{maximum a posteriori (estimator)}}} = \underset{B}{\text{argmax}} P(D|B,M)P(B|M) = \widehat{B}$$

# With great overfitting comes great regularization!

$$\overbrace{\cancel{\text{maximum a posteriori (estimator)}}}^{\text{The best belief after seeing the data}} = \underset{B}{\text{argmax}} P(D|B,M)P(B|M) = \widehat{B}$$

$$\overbrace{P(D|M)}^{\text{Hacked } \textbf{evidence}} = \overbrace{P(D|\widehat{B},M)}^{\text{Hacked } \textbf{likelihood}}$$

## With great overfitting comes great regularization!

$$\overbrace{\cancel{\text{maximum a posteriori (estimator)}}}^{\text{The best belief after seeing the data}} = \underset{B}{\text{argmax}} P(D|B,M)P(B|M) = \widehat{B}$$

$$\overbrace{P(D|M)}^{\text{Hacked } \textbf{evidence}} = \overbrace{P(D|\widehat{B},M)}^{\text{Hacked } \textbf{likelihood}} = \overbrace{P(\underset{\uparrow}{D}|\underset{B}{\text{argmax}} P(\underset{\uparrow}{D}|B,M)P(B|M),M)}^{\text{Likelihood at maximum a posteriori}}$$

Again data appears back and forth!!

# With great overfitting comes great regularization!

$$\overbrace{\cancel{\text{maximum a posteriori (estimator)}}}^{\text{The best belief after seeing the data}} = \underset{B}{\text{argmax}} P(D|B,M)P(B|M) = \widehat{B}$$

$$\overbrace{P(D|M)}^{\text{Hacked } \textbf{evidence}} = \overbrace{P(D|\widehat{B},M)}^{\text{Hacked } \textbf{likelihood}} = \overbrace{P(\underset{\uparrow}{D}|\underset{B}{\text{argmax}} P(\underset{\uparrow}{D}|B,M)P(B|M),M)}^{\text{Likelihood at maximum a posteriori}}$$

Again data appears back and forth!!

$$\boxed{\text{Hacked evidence (with MAP)} = \text{L2 or L1 regularization}}$$

# With great overfitting comes great regularization!

$$\overbrace{\cancel{\text{maximum a posteriori (estimator)}}}^{\text{The best belief after seeing the data}} = \underset{B}{\text{argmax}} P(D|B,M)P(B|M) = \widehat{B}$$

$$\overbrace{P(D|M)}^{\text{Hacked } \textbf{evidence}} = \overbrace{P(D|\widehat{B},M)}^{\text{Hacked } \textbf{likelihood}} = \overbrace{P(\underset{\uparrow}{D}|\underset{B}{\text{argmax}} P(\underset{\uparrow}{D}|B,M)P(B|M),M)}^{\text{Likelihood at maximum a posteriori}}$$

Again data appears back and forth!!

$$\boxed{\text{Hacked evidence (with MAP)} = \text{L2 or L1 regularization}}$$

For more regularization techniques for hacked-belief approach see
Zivik talk. With great complexity comes great regularization

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}}$$

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}} \propto \log P(D|M)$$

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}} \propto \log P(D|M) = \log\left(\prod_{i=1}^{|D|} P(D_i|M)\right)$$

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}} \propto \log P(D|M) = \log\left(\prod_{i=1}^{|D|} P(D_i|M)\right) = \sum_{i=1}^{|D|} \log P(D_i|M)$$

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}} \propto \log P(D|M) = \log \left( \prod_{i=1}^{|D|} P(D_i|M) \right) = \sum_{i=1}^{|D|} \log P(D_i|M)$$

$$\propto \frac{1}{|D|} \sum_{i=1}^{|D|} \log P(D_i|M)$$

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}} \propto \log P(D|M) = \log\left(\prod_{i=1}^{|D|} P(D_i|M)\right) = \sum_{i=1}^{|D|} \log P(D_i|M)$$

$$\propto \frac{1}{|D|}\sum_{i=1}^{|D|} \log P(D_i|M) = E_p\left[\log P(D|M)\right]$$

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}} \propto \log P(D|M) = \log\left(\prod_{i=1}^{|D|} P(D_i|M)\right) = \sum_{i=1}^{|D|} \log P(D_i|M)$$

$$\propto \frac{1}{|D|}\sum_{i=1}^{|D|} \log P(D_i|M) = E_p\left[\log P(D|M)\right]$$

$$= E_p\left[\log q\right]$$

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}} \propto \log P(D|M) = \log\left(\prod_{i=1}^{|D|} P(D_i|M)\right) = \sum_{i=1}^{|D|} \log P(D_i|M)$$

$$\propto \frac{1}{|D|}\sum_{i=1}^{|D|} \log P(D_i|M) = E_p\left[\log P(D|M)\right]$$

$$= E_p\left[\log q\right] = -\underbrace{H(p,q)}_{\text{Cross entropy}}$$

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}} \propto \log P(D|M) = \log\left(\prod_{i=1}^{|D|} P(D_i|M)\right) = \sum_{i=1}^{|D|} \log P(D_i|M)$$

$$\propto \frac{1}{|D|}\sum_{i=1}^{|D|} \log P(D_i|M) = E_p\left[\log P(D|M)\right]$$

$$= E_p\left[\log q\right] = - \underbrace{H(p,q)}_{\text{Cross entropy}}$$

Evidence $\propto$ Cross entropy

# Evidence and data science metrics

Is there any data science metrics equivalent to evidence?

$$\overbrace{P(D|M)}^{\text{Evidence}} \propto \log P(D|M) = \log\left(\prod_{i=1}^{|D|} P(D_i|M)\right) = \sum_{i=1}^{|D|} \log P(D_i|M)$$

$$\propto \frac{1}{|D|}\sum_{i=1}^{|D|} \log P(D_i|M) = E_p\left[\log P(D|M)\right]$$

$$= E_p\left[\log q\right] = -\underbrace{H(p,q)}_{\text{Cross entropy}}$$

Evidence $\propto$ Cross entropy
(at validation data set, if hacked-belief approach)