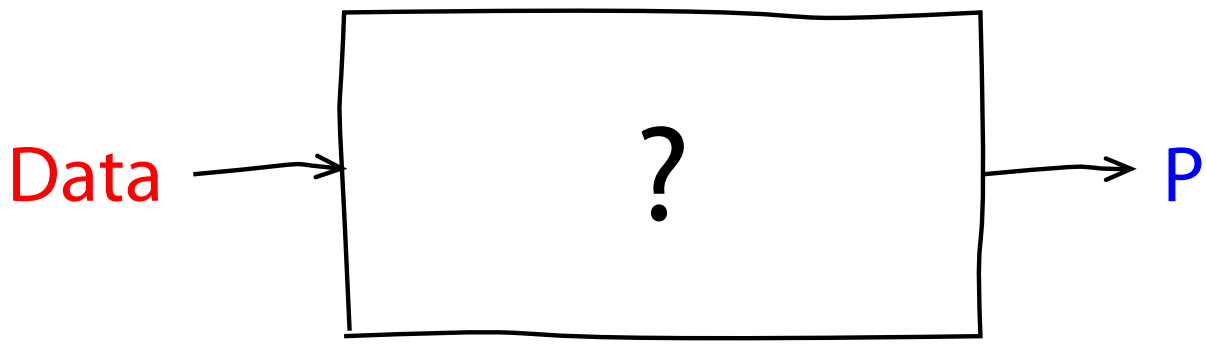
How to Learn Probability Without Learning



Young-Han Kim, UCSD

Munich Workshop on Causal Inference and Information Theory
May 23, 2016

Laplace's rule of succession (1812):

$$\mathrm{P}(☼ \mid n \times ☼) = \frac{n+1}{n+2}$$

●●○○○ AT&T 6:02 PM 57%

google.com

WEB IMAGES Y

Google

how to learn p

how to learn **piano**

how to learn **parkour**

how to learn **programming**

how to learn **python**

how to learn **patience**

Unknown - Use precise location

Settings Privacy Terms

Advertising Business About

●●○○○ AT&T 6:02 PM 57%

google.com

WEB IMAGES Y

Google

how to learn p

how to learn piano

how to learn parkour

how to learn programming

how to learn python

how to learn patience

Unknown - Use precise location

Settings Privacy Terms

Advertising Business About

Politics TOTALLY SUBJECTIVE ODDS: FIRST-TIER REPUBLICANS

Analysts	J. Bush	S. Walker	M. Rubio	R. Paul	C. Christie
NATE	25%	25%	18%	5%	3%
HARRY	30	26	15	2	2
MICAH	17	22	15	9	3

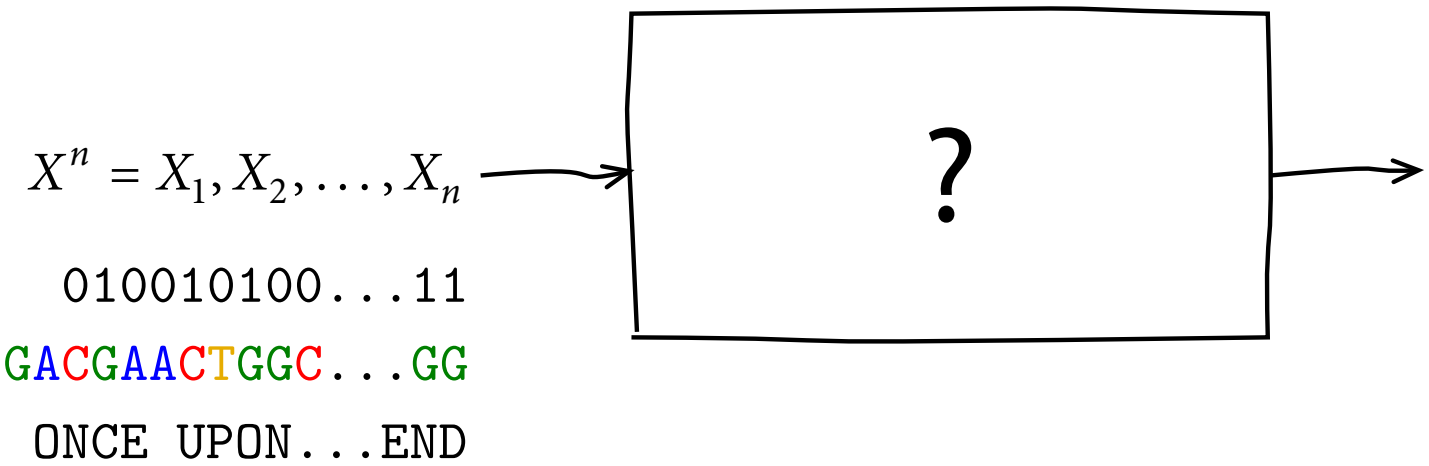
Politics TOTALLY SUBJECTIVE ODDS: REPUBLICANS

Analysts	T. Cruz	R. Perry	M. Huckabee	B. Jindal	B. Carson
NATE	2%	4%	3%	2%	2%
HARRY	3	3	3	3	1
MICAH	2	12	3	2	1

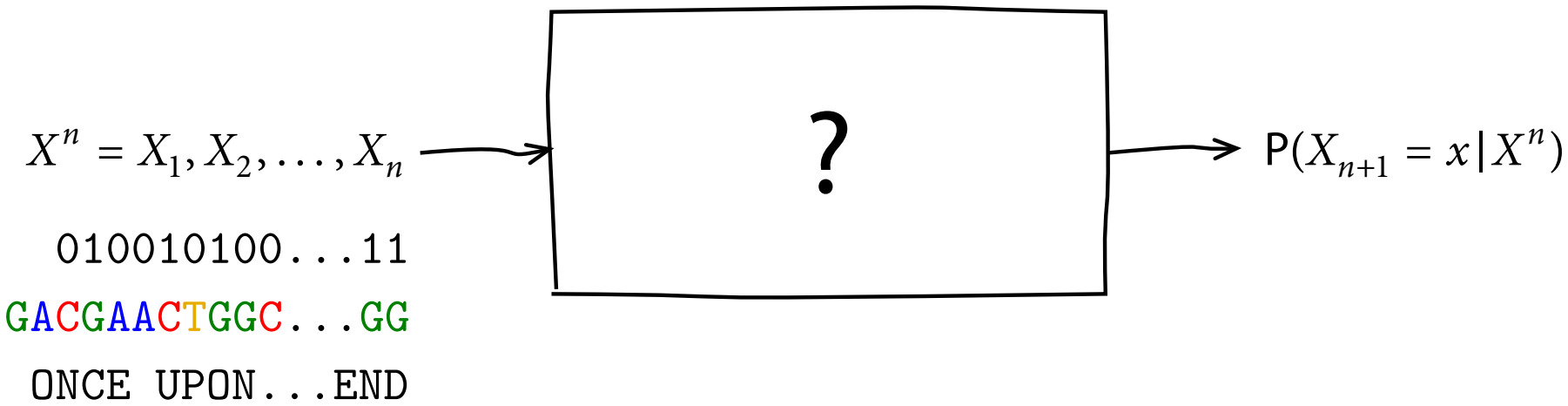
Politics TOTALLY SUBJECTIVE ODDS: REPUBLICANS

Analysts	R. Santorum	M. Pence	L. Graham	C. Fiorina	Other
NATE	1%	3%	0%	0%	8%
HARRY	1	2	0	2	7
MICAH	2	1	1	1	9

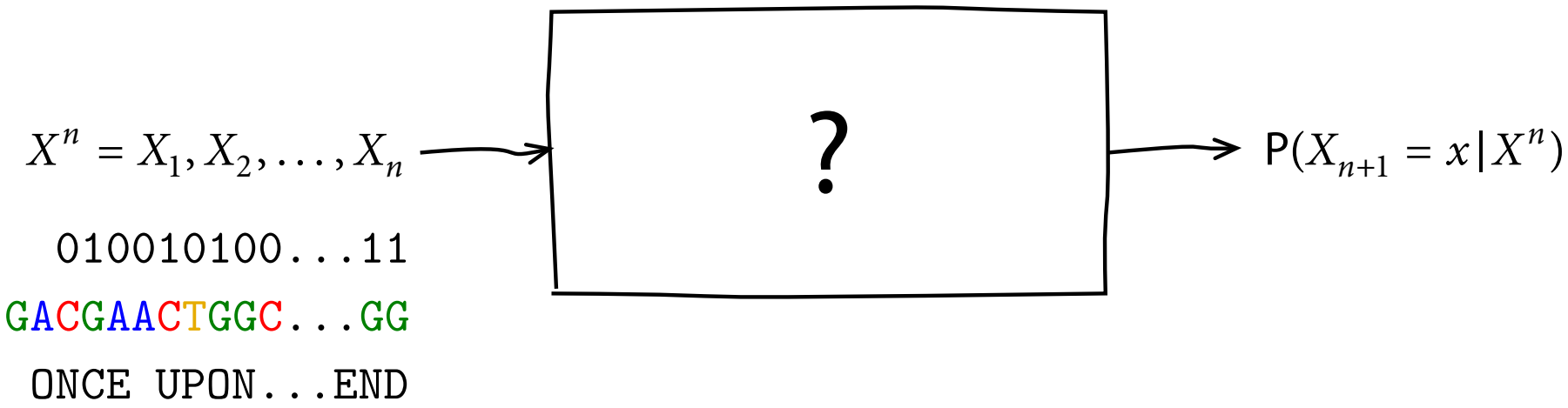
How to learn probability



How to learn probability



How to learn probability



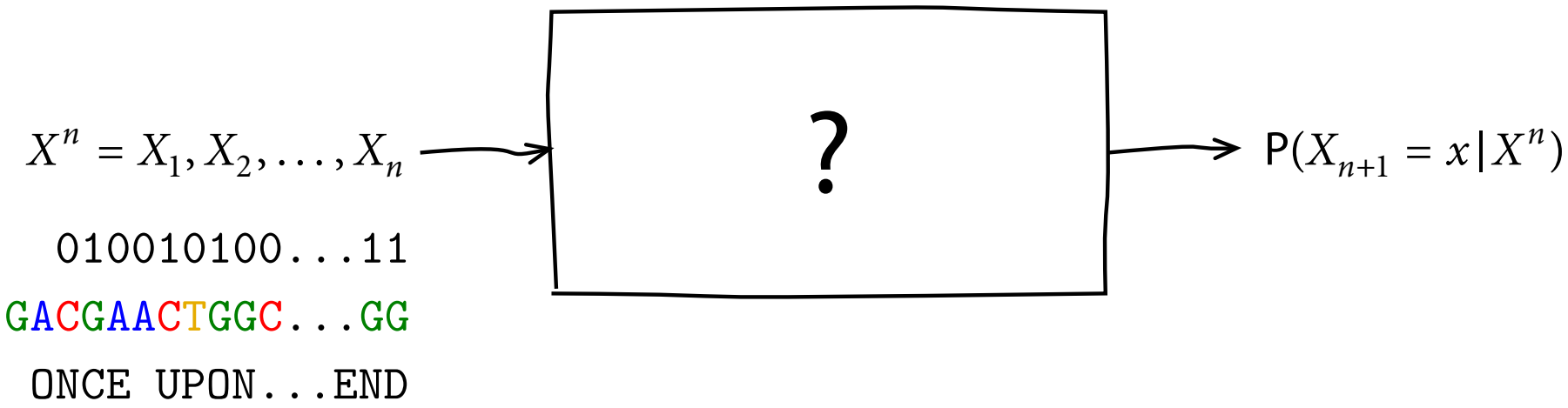
Forward estimation

Is there f_n such that

$$\left| f_n(X_1^n) - \mathsf{P}(X_{n+1} = x \mid X_1^n) \right| \to 0 \quad \text{a.s.}$$

for every stationary ergodic $\{X_n\}$?

How to learn probability



Forward estimation

Is there f_n such that

$$\left| f_n(X_1^n) - \mathsf{P}(X_{n+1} = x \mid X_1^n) \right| \to 0 \quad \text{a.s.}$$

for every stationary ergodic $\{X_n\}$?

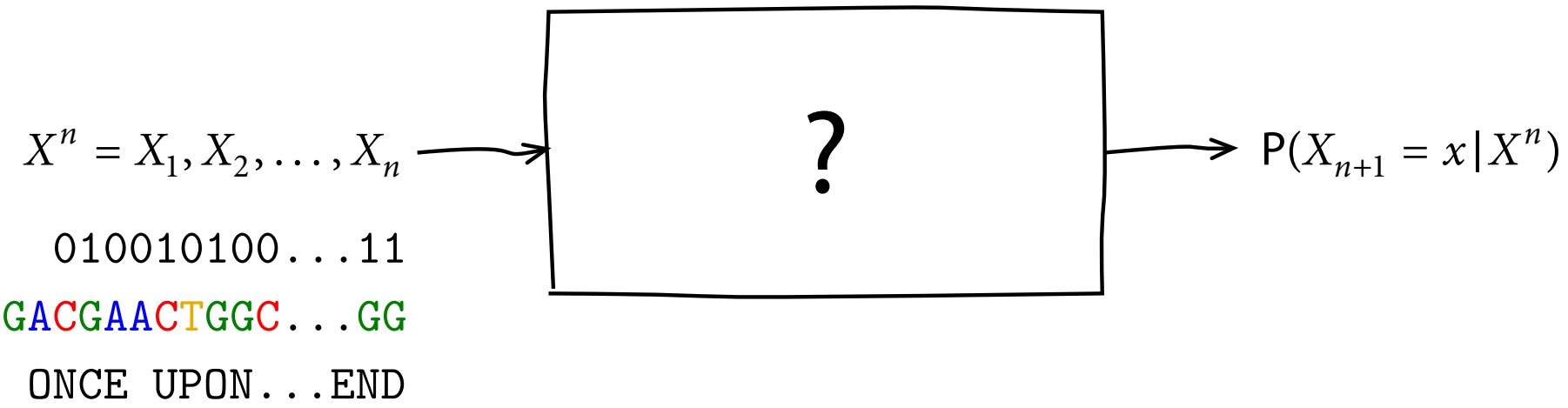
Backward estimation (Cover 1975)

Is there g_n such that

$$\left| g_n(X_{-n}^{-1}) - \mathsf{P}(X_0 = x \mid X_{-n}^{-1}) \right| \to 0 \quad \text{a.s.}$$

for every stationary ergodic $\{X_n\}$?

How to learn probability



Forward estimation

Is there f_n such that

$$\left| f_n(X_1^n) - \mathsf{P}(X_{n+1} = x \mid X_1^n) \right| \to 0 \quad \text{a.s.}$$

for every stationary ergodic $\{X_n\}$?

Answer: No! (Bailey 1976)

Getting older doesn't make you wiser

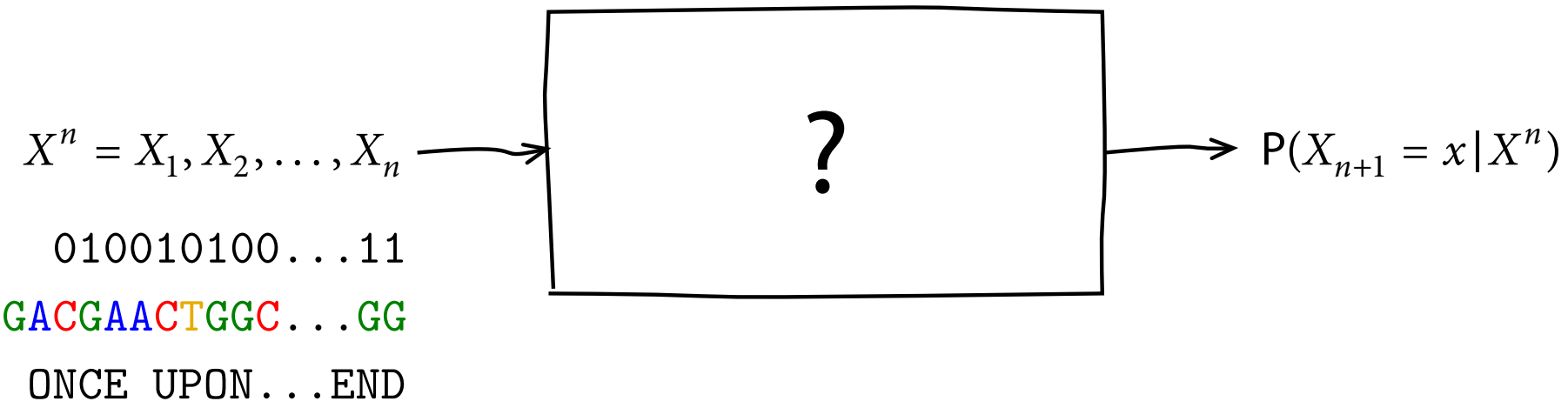
Backward estimation (Cover 1975)

Is there g_n such that

$$\left| g_n(X_{-n}^{-1}) - \mathsf{P}(X_0 = x \mid X_{-n}^{-1}) \right| \to 0 \quad \text{a.s.}$$

for every stationary ergodic $\{X_n\}$?

How to learn probability



Forward estimation

Is there f_n such that

$$\left| f_n(X_1^n) - \mathsf{P}(X_{n+1} = x \mid X_1^n) \right| \to 0 \quad \text{a.s.}$$

for every stationary ergodic $\{X_n\}$?

Answer: No! (Bailey 1976)

Getting older doesn't make you wiser

Backward estimation (Cover 1975)

Is there g_n such that

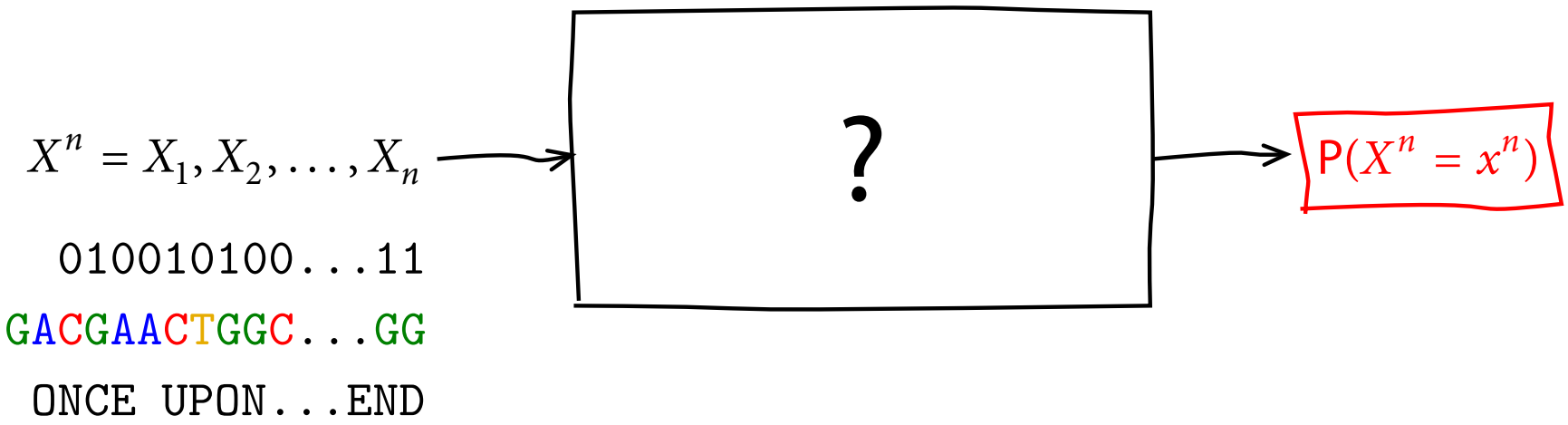
$$\left| g_n(X_{-n}^{-1}) - \mathsf{P}(X_0 = x \mid X_{-n}^{-1}) \right| \to 0 \quad \text{a.s.}$$

for every stationary ergodic $\{X_n\}$?

Answer: Yes! (Ornstein 1978)

Learning more history does

How to learn probability



Forward estimation

Is there f_n such that

$$\left| f_n(X_1^n) - \mathrm{P}(X_{n+1} = x \mid X_1^n) \right| \to 0 \text{ a.s.}$$

for every stationary ergodic $\{X_n\}$?

Answer: No! (Bailey 1976)

Getting older doesn't make you wiser

Backward estimation (Cover 1975)

Is there g_n such that

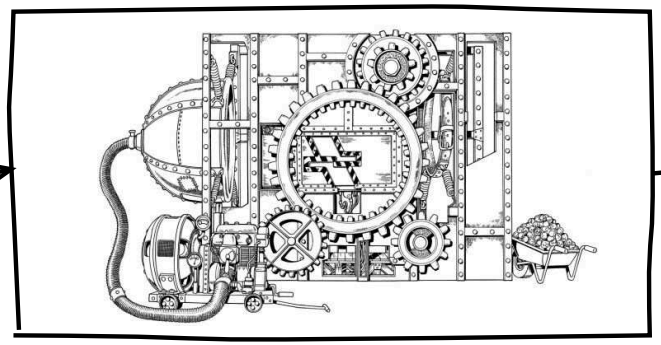
$$\left| g_n(X_{-n}^{-1}) - \mathrm{P}(X_0 = x \mid X_{-n}^{-1}) \right| \to 0 \text{ a.s.}$$

for every stationary ergodic $\{X_n\}$?

Answer: Yes! (Ornstein 1978)

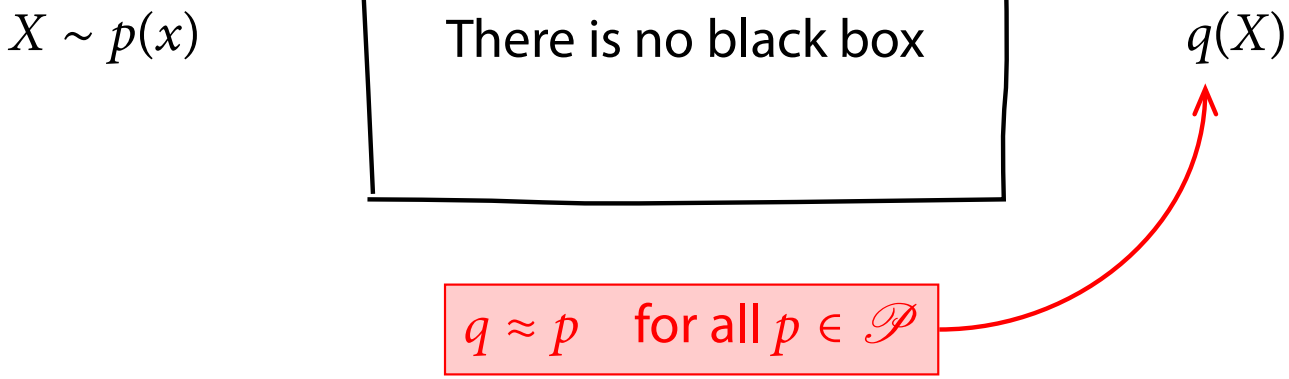
Learning more history does

How to learn probability without learning

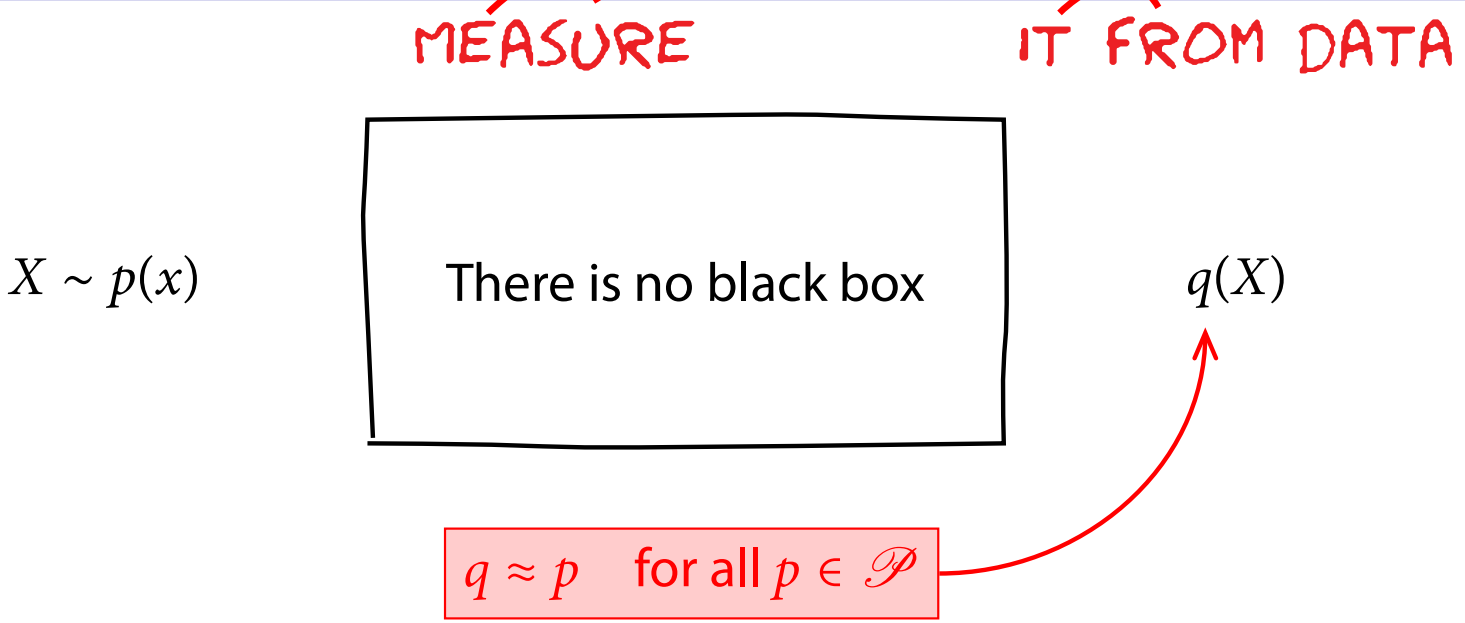
$X \sim p(x)$ →  → $\hat{p}(\cdot \mid X, p) \;\to\; \hat{p}(X)$

How to learn probability without learning

MEASURE IT FROM DATA

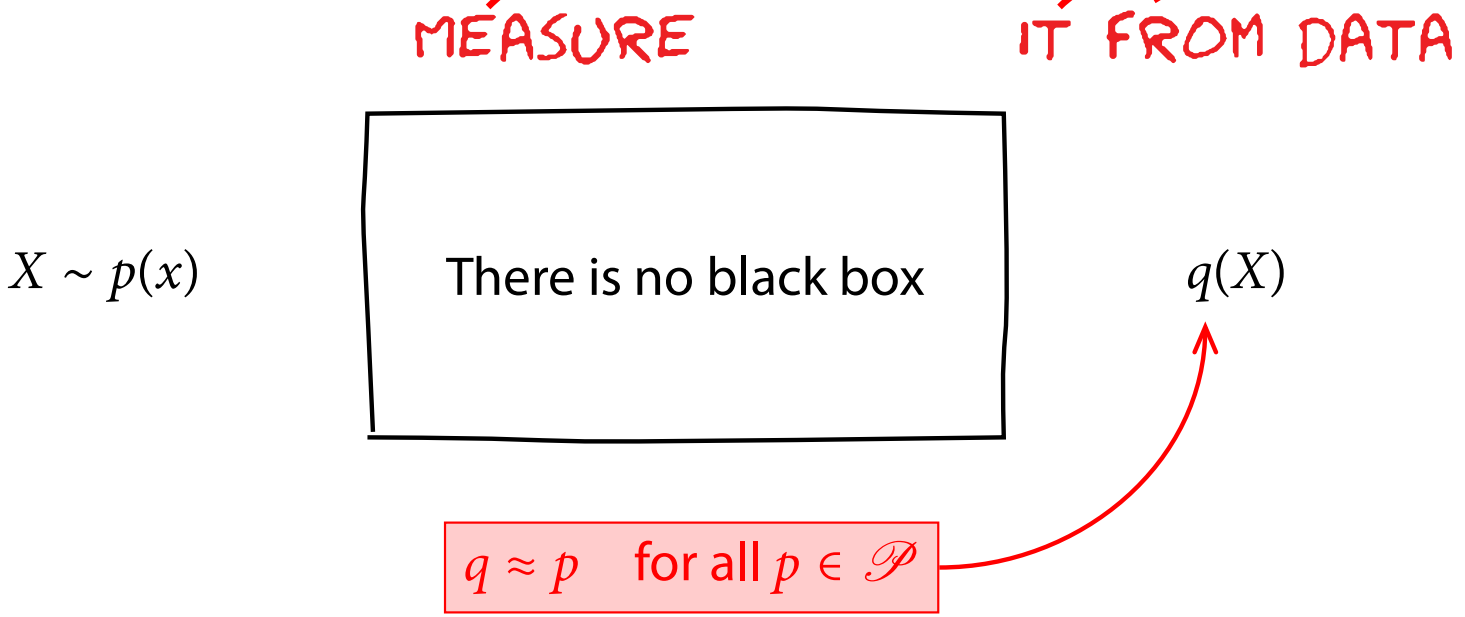


How to learn probability without learning



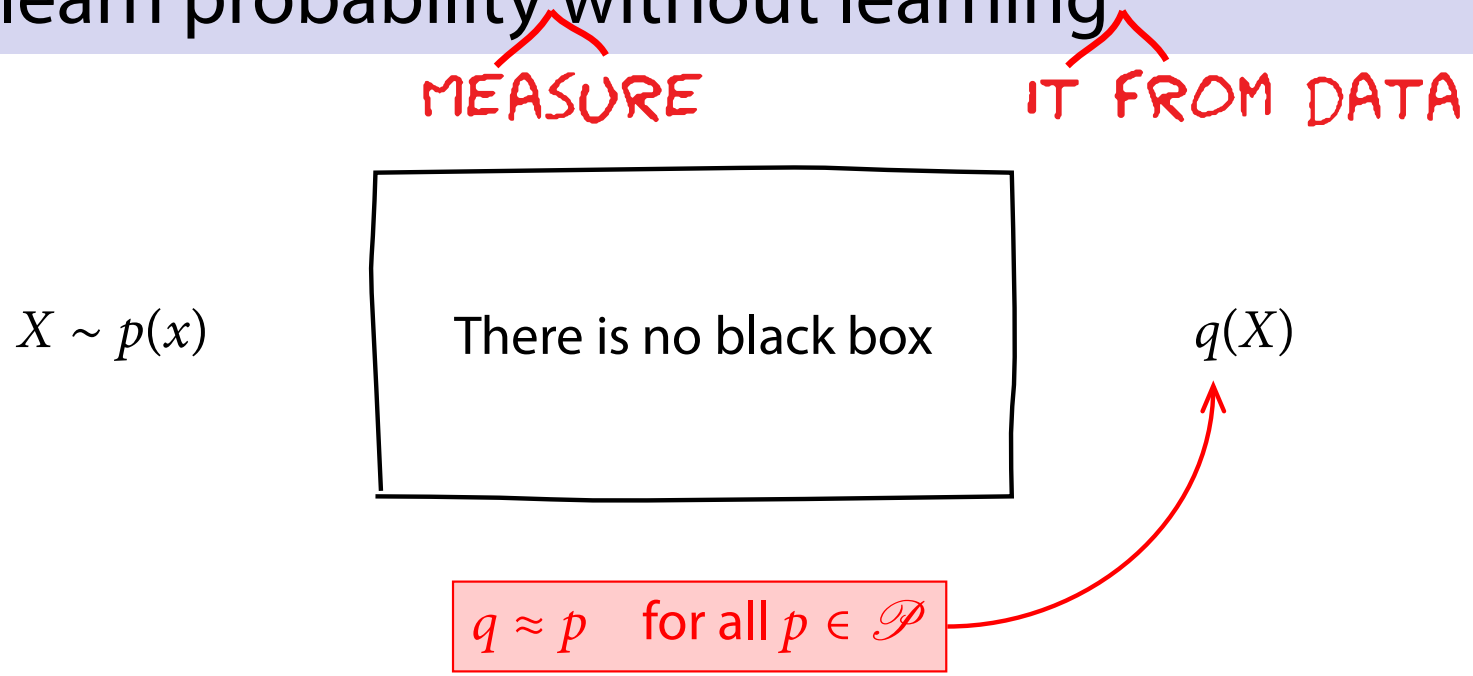
- Universal
  - $\mathscr{P}$: parametric, IID, Markov, VMM, HMM, FSM, stationary ergodic, …

How to learn probability without learning



- Universal
  - $\mathscr{P}$: parametric, IID, Markov, VMM, HMM, FSM, stationary ergodic, …
- Quick and ~~dirty~~ CLEAN
  - Compression, prediction, filtering, denoising, portfolio, entropy estimation, classification

How to learn probability without learning



- Universal
  - $\mathscr{P}$: parametric, IID, Markov, VMM, HMM, FSM, stationary ergodic, …
- Quick and ~~dirty~~ CLEAN
  - Compression, prediction, filtering, denoising, portfolio, entropy estimation, classification
- Avoids overfitting (built-in regularization)

Outline of the talk

- Brief overview of universal probability assignment
  1. Definition, existence, and construction
  2. Convergence control
  3. Well-known applications

Outline of the talk

- Brief overview of universal probability assignment
  1. Definition, existence, and construction
  2. Convergence control
  3. Well-known applications

- Directed information and its application to causality inference

Outline of the talk

- Brief overview of universal probability assignment
  1. Definition, existence, and construction
  2. Convergence control
  3. Well-known applications

- Directed information and its application to causality inference

- Classification of DNA/RNA sequences using universal probability

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n =$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = 0$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = 0\ 00$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = \texttt{0 00 1}$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = \texttt{0 00 1 10}$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = \texttt{0 00 1 10 101}$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = \texttt{0 00 1 10 101 1011}$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = \texttt{0 00 1 10 101 1011 000}$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = \underbrace{0 \;\; 00 \;\; 1 \;\; 10 \;\; 101 \;\; 1011 \;\; 000}_{c(x^n)} \quad \Rightarrow \quad q(x^n) = \frac{1}{(1 + c(x^n))!}$$

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = \underbrace{0 \;\; 00 \;\; 1 \;\; 10 \;\; 101 \;\; 1011 \;\; 000}_{c(x^n)} \quad \Rightarrow \quad q(x^n) = \frac{1}{(1 + c(x^n))!}$$

- Works “well” with deterministic (=real-life) sequences

Universal probability for stationary ergodic processes

q is mean universal if

$$\frac{1}{n} D(p(x^n) \| q(x^n)) \to 0 \quad \forall p \in \mathscr{P}$$

q is pointwise universal if

$$\frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \to 0 \quad p\text{-a.s. } \forall p \in \mathscr{P}$$

- Relative entropy (Kullback–Leibler divergence)

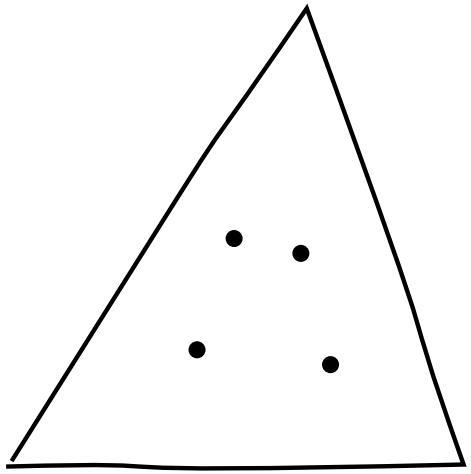
$$D(p(x^n) \| q(x^n)) = \sum_{x^n} p(x^n) \log \frac{p(x^n)}{q(x^n)} = \mathsf{E}_p \log \frac{p(X^n)}{q(X^n)}$$

- Simple construction using LZ78 incremental parsing (Ziv–Lempel 1978)

$$x^n = \underbrace{0 \;\; 00 \;\; 1 \;\; 10 \;\; 101 \;\; 1011 \;\; 000}_{c(x^n)} \quad \Rightarrow \quad q(x^n) = \frac{1}{(1 + c(x^n))!}$$

- Works "well" with deterministic (=real-life) sequences
- Faster convergence (minimax) for smaller classes $\mathscr{P}$ (IID, Markov, CTM, …)

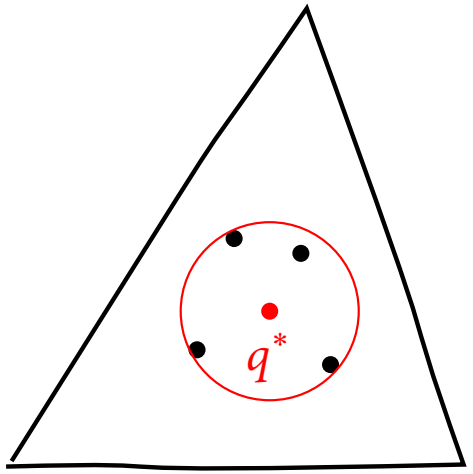
Minimax probability assignment



Minimax probability assignment

Minimax redundancy (Gallager 1974)

$$R^* = \min_{q} \max_{p \in \mathscr{P}} D(p(x) \| q(x)) = \max_{F(p)} I(P; X)$$

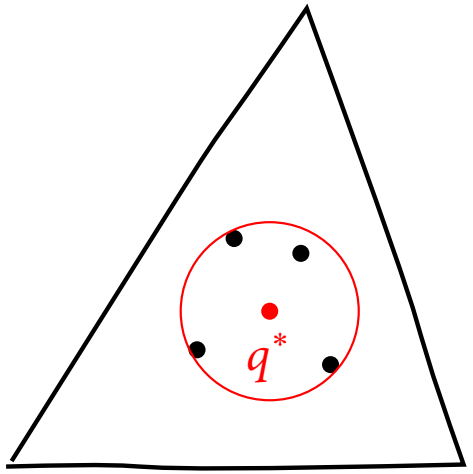


Minimax probability assignment

Minimax redundancy (Gallager 1974)

$$R^* = \min_{q} \max_{p \in \mathscr{P}} D(p(x) \| q(x)) = \max_{F(p)} I(P; X)$$

$$q^*(x) = \int p(x) dF^*(p)$$

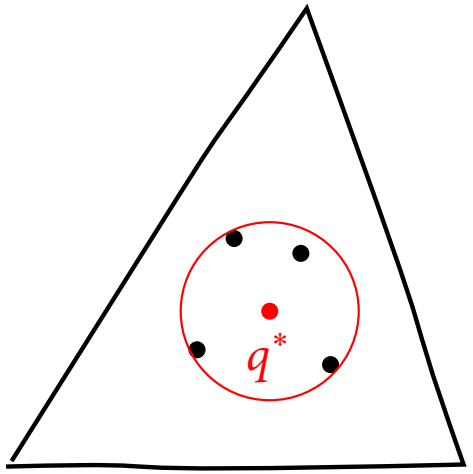


Minimax probability assignment

Minimax redundancy (Gallager 1974)

$$R^* = \min_q \max_{p \in \mathscr{P}} D(p(x) \| q(x)) = \max_{F(p)} I(P; X)$$

$$q^*(x) = \int p(x) dF^*(p)$$



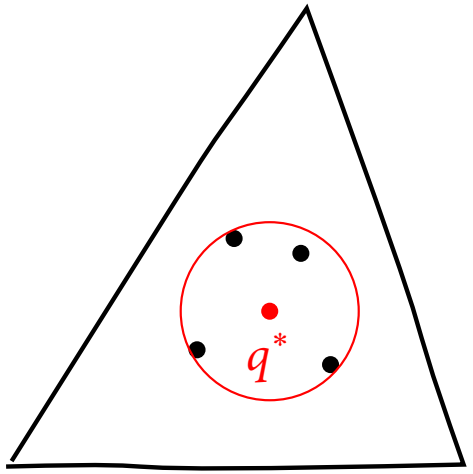
- Mixture probability $F(p)$ can upper and lower bound R^*

Minimax probability assignment

Minimax redundancy (Gallager 1974)

$$R^* = \min_q \max_{p \in \mathscr{P}} D(p(x) \| q(x)) = \max_{F(p)} I(P; X)$$

$$q^*(x) = \int p(x) dF^*(p)$$



- Mixture probability $F(p)$ can upper and lower bound R^*
- For the deterministic setting,

$$R^* = \min_q \max_{p \in \mathscr{P}} \max_x \log \frac{p(x)}{q(x)}$$
$$= \log \sum_x \max_{p \in \mathscr{P}} p(x)$$

$$q^*(x) \propto \max_{p \in \mathscr{P}} p(x) \quad \text{(normalized ML)}$$

Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \mathrm{Bern}(\theta)$, $\theta \in [0,1]$ unknown

Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \mathrm{Bern}(\theta)$, $\theta \in [0,1]$ unknown
- Uniform mixture (Laplace 1812): $R \sim \log n$ (universal for Bernoulli sources!)

$$q_{\mathrm{L}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n+1)} \quad \text{and} \quad q_{\mathrm{L}}(1 \mid x^n) = \frac{k+1}{n+2}$$

Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \mathrm{Bern}(\theta)$, $\theta \in [0,1]$ unknown
- Uniform mixture (Laplace 1812): $R \sim \log n$ (universal for Bernoulli sources!)

$$q_{\mathrm{L}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n+1)} \quad \text{and} \quad q_{\mathrm{L}}(1 \,|\, x^n) = \frac{k+1}{n+2}$$

- Jeffreys mixture (Krichevsky–Trofimov 1981): $R \sim \frac{1}{2} \log n \sim R^*$

$$q_{\mathrm{KT}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} \frac{1}{\sqrt{\theta(1-\theta)}} d\theta \quad \text{and} \quad q_{\mathrm{KT}}(1 \,|\, x^n) = \frac{k+1/2}{n+1}$$

Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \mathrm{Bern}(\theta)$, $\theta \in [0,1]$ unknown
- Uniform mixture (Laplace 1812): $R \sim \log n$ (universal for Bernoulli sources!)

$$q_{\mathrm{L}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n+1)} \quad \text{and} \quad q_{\mathrm{L}}(1 \,|\, x^n) = \frac{k+1}{n+2}$$

- Jeffreys mixture (Krichevsky–Trofimov 1981): $R \sim \frac{1}{2} \log n \sim R^*$

$$q_{\mathrm{KT}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} \frac{1}{\sqrt{\theta(1-\theta)}} d\theta \quad \text{and} \quad q_{\mathrm{KT}}(1 \,|\, x^n) = \frac{k+1/2}{n+1}$$

- For m-ary sources, $R^* \sim \frac{(m-1)}{2} \log n$ (both stochastic and deterministic)

Bernoulli sources

- Let $X_1, X_2, \ldots$ be i.i.d. $\sim \mathrm{Bern}(\theta)$, $\theta \in [0,1]$ unknown
- Uniform mixture (Laplace 1812): $R \sim \log n$ (universal for Bernoulli sources!)

$$q_{\mathrm{L}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} d\theta = \frac{1}{\binom{n}{k}(n+1)} \quad \text{and} \quad q_{\mathrm{L}}(1 \,|\, x^n) = \frac{k+1}{n+2}$$

- Jeffreys mixture (Krichevsky–Trofimov 1981): $R \sim \frac{1}{2} \log n \sim R^*$

$$q_{\mathrm{KT}}(x^n) = \int_0^1 \theta^k (1-\theta)^{n-k} \frac{1}{\sqrt{\theta(1-\theta)}} d\theta \quad \text{and} \quad q_{\mathrm{KT}}(1 \,|\, x^n) = \frac{k+1/2}{n+1}$$

- For m-ary sources, $R^* \sim \frac{(m-1)}{2} \log n$ (both stochastic and deterministic)
- Can be generalized to Markov and tree (CTM) sources (Willems et al. 1995)

Applications

- Compression: Compress to the entropy rate using arithmetic coding

Applications

- Compression: Compress to the entropy rate using arithmetic coding

- Prediction: Take an action $a(X^n)$ for X_{n+1}
  - Bayes response: $a^*(p) = \arg\min_a \mathsf{E}_p[l(X, a)]$
  - Merhav–Feder (1998): Choose action $a^*(q(x_{n+1}|X_1^n))$

Applications

- Compression: Compress to the entropy rate using arithmetic coding
- Prediction: Take an action $a(X^n)$ for X_{n+1}
  - Bayes response: $a^*(p) = \arg\min_a \mathsf{E}_p[l(X, a)]$
  - Merhav–Feder (1998): Choose action $a^*(q(x_{n+1}|X_1^n))$
- Portfolio selection: Choose asset allocation $b_x(Y^n)$ for stock x
  - Fund of funds: Multi-period asset allocation using $q(x^n)$
  - Cover (1991): Minimax performance against constant-rebalanced portfolios

Applications

- Compression: Compress to the entropy rate using arithmetic coding
- Prediction: Take an action $a(X^n)$ for X_{n+1}
  - Bayes response: $a^*(p) = \arg\min_a \mathsf{E}_p[l(X, a)]$
  - Merhav–Feder (1998): Choose action $a^*(q(x_{n+1}|X_1^n))$
- Portfolio selection: Choose asset allocation $b_x(Y^n)$ for stock x
  - Fund of funds: Multi-period asset allocation using $q(x^n)$
  - Cover (1991): Minimax performance against constant-rebalanced portfolios
- Entropy estimation: Estimate the entropy rate of $\{X_n\}$
  - Shannon–McMillan–Breiman theorem: $\frac{1}{n} \log \frac{1}{p(X^n)} \to \bar{H}(X)$
  - Plug-in strategy: Use q in place of p

Outline of the talk

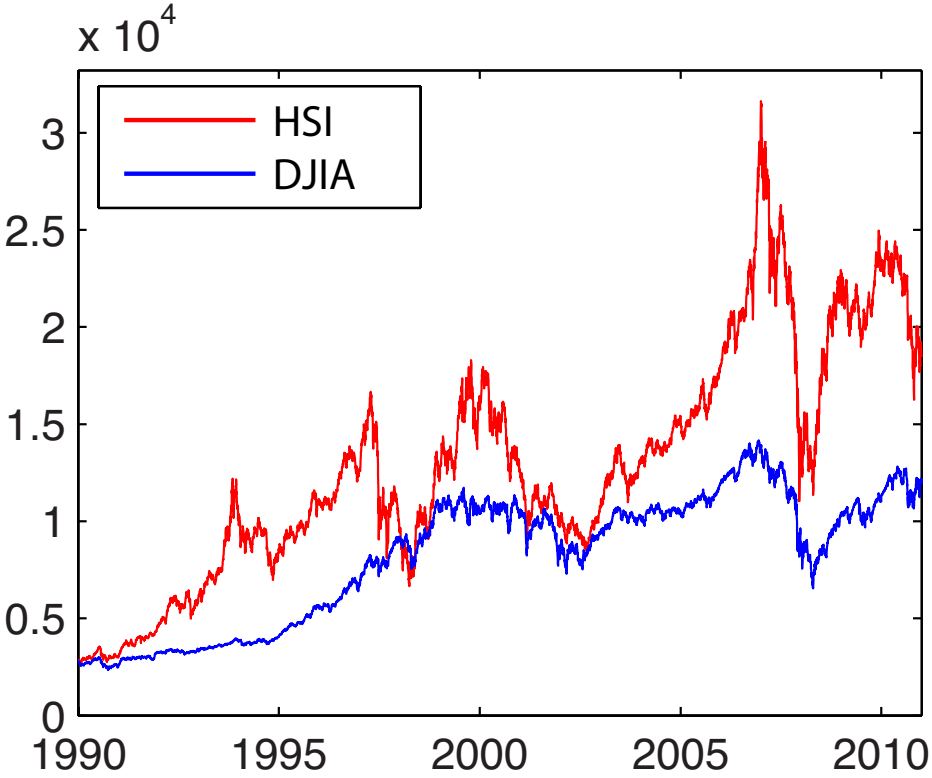
- Brief overview of universal probability assignment

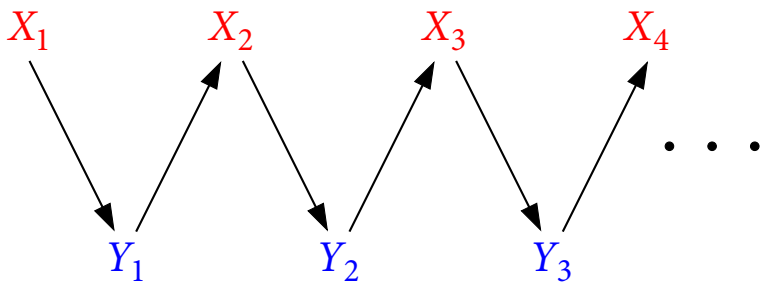
- Directed information and its application to causality inference



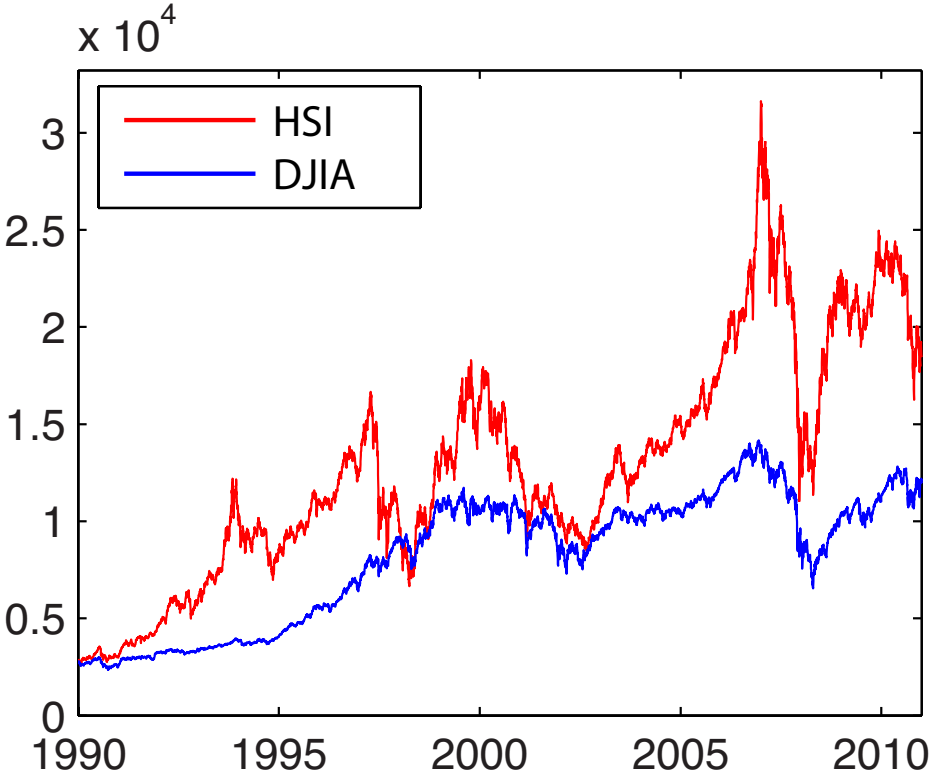
- Classification of DNA/RNA sequences using universal probability

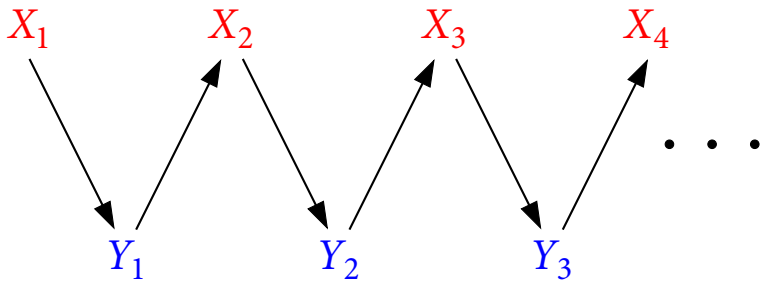
Correlation and causation between time series





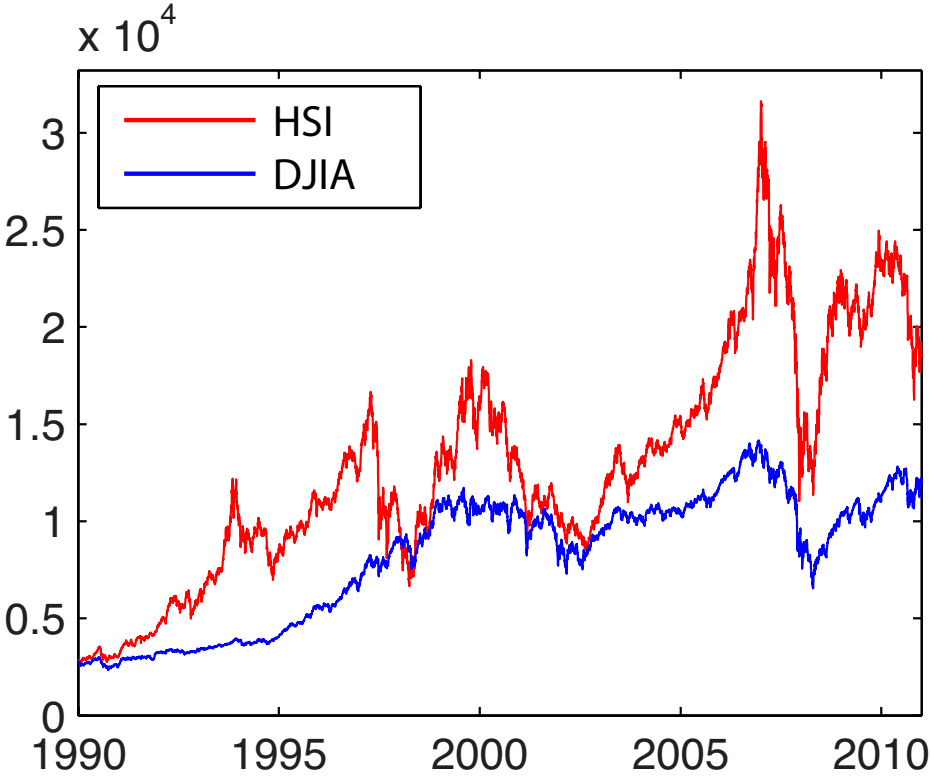
Correlation and causation between time series

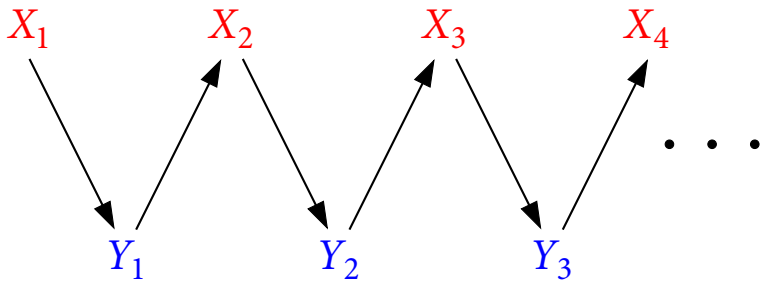




Are they "correlated"?

Correlation and causation between time series

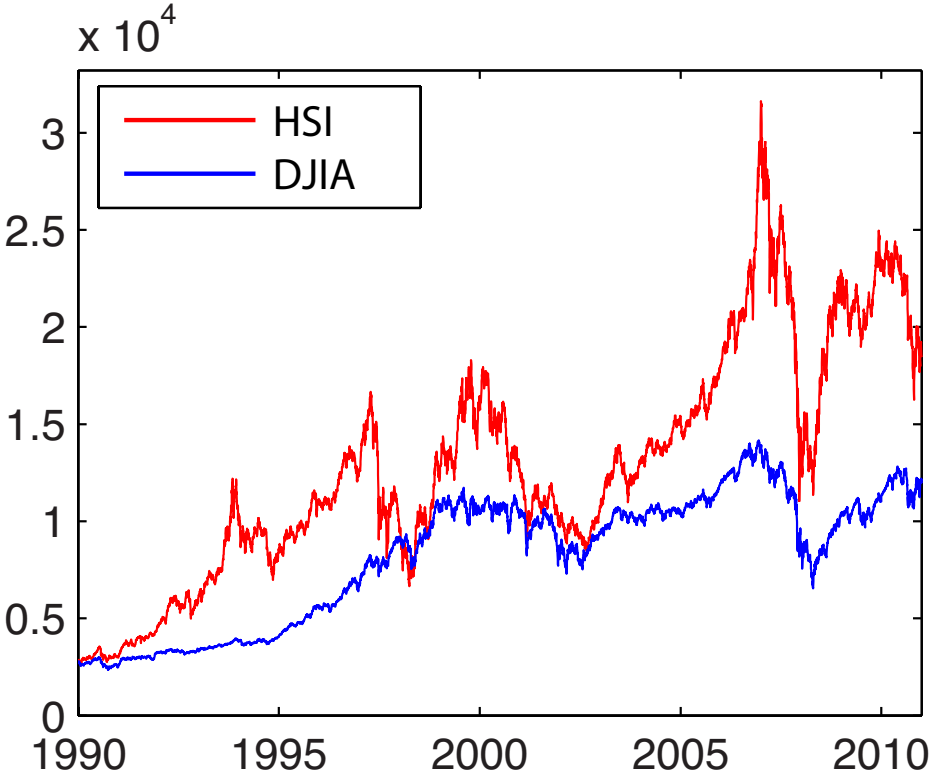


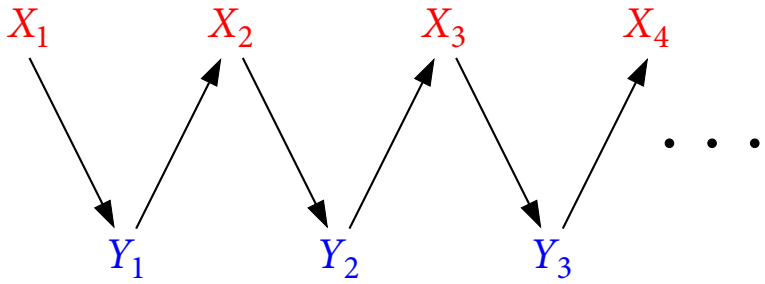


Are they "correlated"?

Yes if $I(X;Y) \gg 0$

Correlation and causation between time series



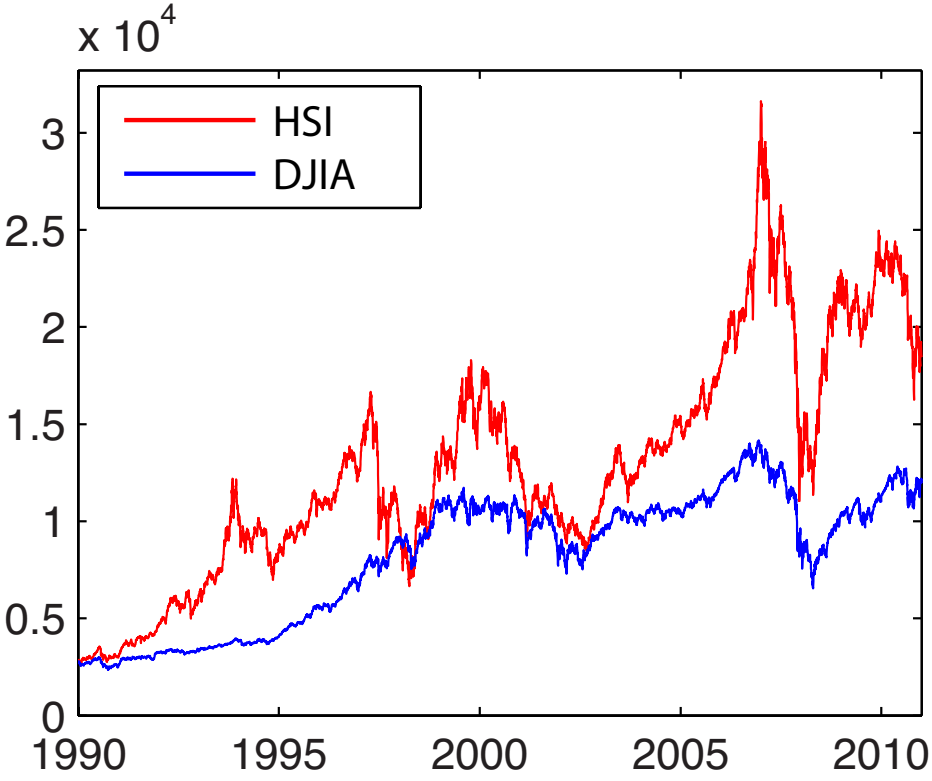


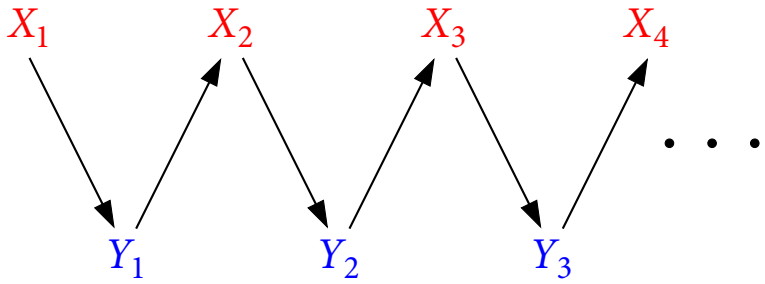
Are they "correlated"?

Yes if $I(X;Y) \gg 0$

Which "leads" the other?

Correlation and causation between time series





Are they "correlated"?

Yes if $I(X;Y) \gg 0$

Which "leads" the other?

X if $I(X \to Y) \gg I(Y \to X)$

Y if $I(Y \to X) \gg I(X \to Y)$

Directed information

$I(X \to Y)$

$$H(Y) - H(Y\|X)$$
$$= \sum H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i)$$

Directed information

$$I(X \to Y)$$

$$H(Y) - H(Y\|X)$$
$$= \sum H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i)$$

- Causal information from X to Y

Marko (1966, 1973)
Massey (1990)

- Cybernetics, feedback comm., …
(Kramer 1998, Permuter 2008)

Directed information

$$I(X \to Y)$$

$$H(Y) - H(Y\|X)$$
$$= \sum H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i)$$

$$G(X \to Y)$$

$$\sum \log \frac{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1})}{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1}, X_{i-p}^{i})}$$

- Causal information from X to Y

Marko (1966, 1973)
Massey (1990)

- Cybernetics, feedback comm., …
(Kramer 1998, Permuter 2008)

Directed information

$$I(X \to Y)$$

$$H(Y) - H(Y\|X) = \sum H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i)$$

- Causal information from X to Y

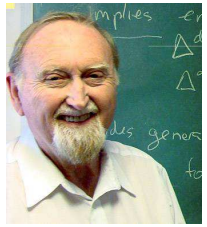
Marko (1966, 1973)
Massey (1990)

- Cybernetics, feedback comm., …
(Kramer 1998, Permuter 2008)

$$G(X \to Y)$$

$$\sum \log \frac{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1})}{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1}, X_{i-p}^{i})}$$

- Causal influence of X on Y



Granger (1969)
Geweke (1982)

- Econometrics, neuroscience, …
(Sims 1972, Quinn et al. 2011)

Directed information

$$I(X \to Y)$$

$$H(Y) - H(Y\|X) = \sum H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i)$$

- Causal information from X to Y

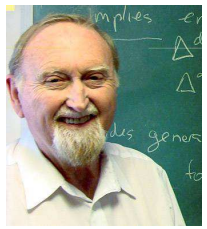
Marko (1966, 1973)
Massey (1990)

- Cybernetics, feedback comm., …
(Kramer 1998, Permuter 2008)

$$G(X \to Y)$$

$$\sum \log \frac{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1})}{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1}, X_{i-p}^{i})}$$

- Causal influence of X on Y



Granger (1969)
Geweke (1982)

- Econometrics, neuroscience, …
(Sims 1972, Quinn et al. 2011)

- Other interpretations (Permuter–Kim–Weissman 2011, Kamath–Kim 2014)

Directed information

$$I(X \to Y)$$

$$H(Y) - H(Y\|X)$$
$$= \sum H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i)$$

- Causal information from X to Y

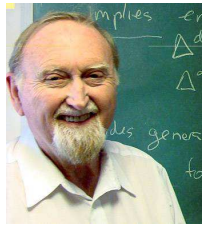
Marko (1966, 1973)
Massey (1990)

- Cybernetics, feedback comm., …
(Kramer 1998, Permuter 2008)

$$G(X \to Y)$$

$$\sum \log \frac{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1})}{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1}, X_{i-p}^{i})}$$

- Causal influence of X on Y



Granger (1969)
Geweke (1982)

- Econometrics, neuroscience, …
(Sims 1972, Quinn et al. 2011)

- Other interpretations (Permuter–Kim–Weissman 2011, Kamath–Kim 2014)
- Can be generalized to continuous time (Weissman–Kim–Permuter 2013)

Directed information

$$I(X \to Y)$$

$$H(Y) - H(Y\|X) = \sum H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i)$$

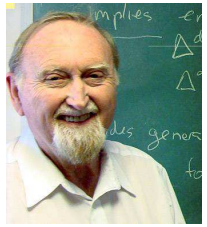
- Causal information from X to Y

Marko (1966, 1973)
Massey (1990)

$$G(X \to Y)$$

$$\sum \log \frac{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1})}{\mathrm{MSE}(Y_i|Y_{i-p}^{i-1}, X_{i-p}^{i})}$$

- Causal influence of X on Y



Granger (1969)
Geweke (1982)

- Cybernetics, feedback comm., …
  (Kramer 1998, Permuter 2008)
- Econometrics, neuroscience, …
  (Sims 1972, Quinn et al. 2011)
- Other interpretations (Permuter–Kim–Weissman 2011, Kamath–Kim 2014)
- Can be generalized to continuous time (Weissman–Kim–Permuter 2013)
- Conservation law: $I(X;Y) = I(X \to Y) + I(Y \to X)$

Directed information estimation (Jiao et al. 2013)

Algorithm 1

$$\hat{I}_1(X \to Y) = \hat{H}_1(Y) - \hat{H}_1(Y\|X)$$

$$\frac{1}{n} \log \frac{1}{q(Y^n)}$$

- ☺ Very good convergence (a.s. & L_1)
- ☹ Erratic for small n
- ☹ Unbounded support

Algorithm 2

$$\hat{I}_2(X \to Y) = \hat{H}_2(Y) - \hat{H}_2(Y\|X)$$

$$\frac{1}{n} \sum_{i=1}^{n} H(q(y_i \mid Y^{i-1}))$$

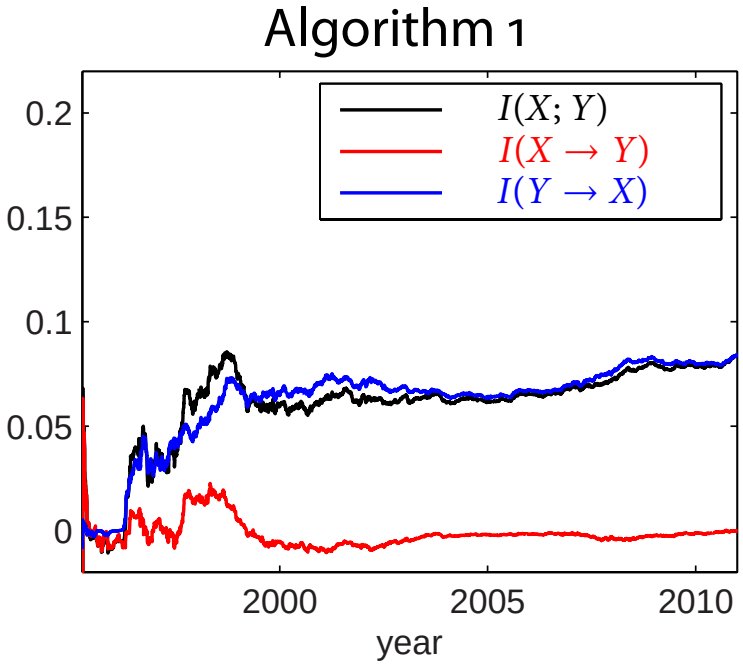
- ☺ Similar convergence rate
- ☺ Smooth and bounded support
- ☹ Can be negative

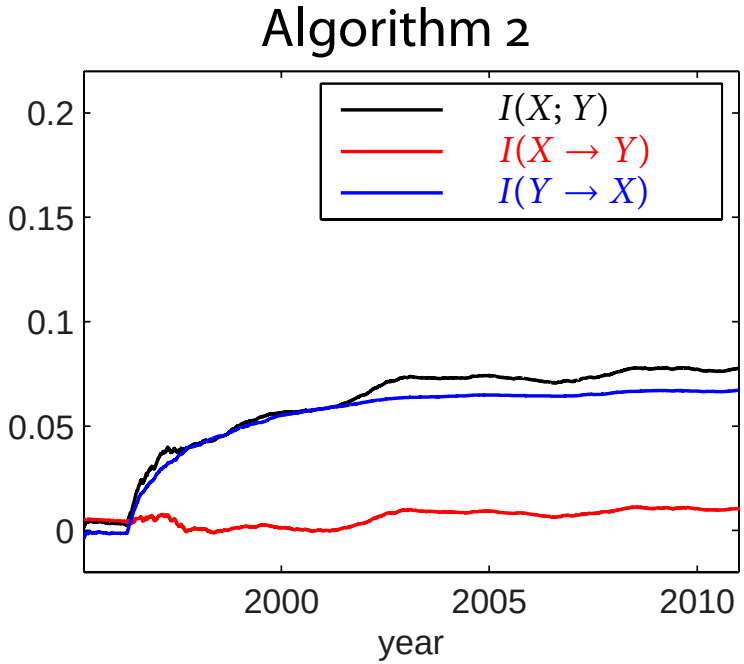
Algorithms 3 & 4

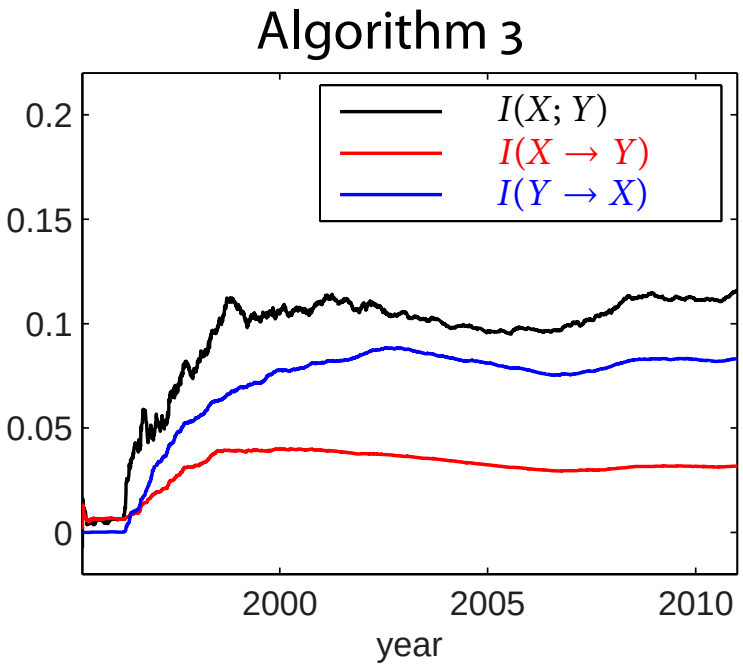
$$\hat{I}_3(X \to Y) = \frac{1}{n} \sum_{i=1}^{n} D(q(y_i|X^i, Y^{i-1}) \| q(y_i|Y^{i-1}))$$

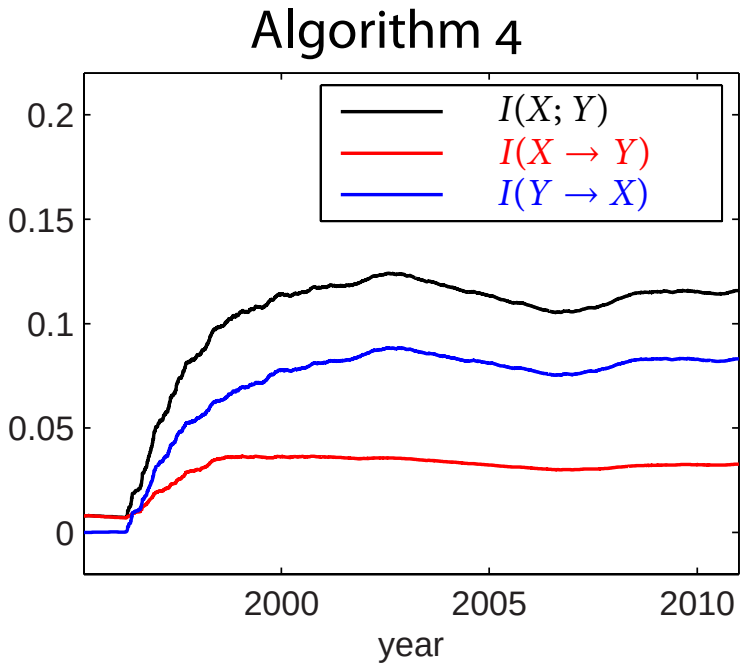
$$\hat{I}_4(X \to Y) = \frac{1}{n} \sum_{i=1}^{n} D(q(x_i, y_i|X^i, Y^{i-1}) \| q(y_i|Y^{i-1}) q(x_i|X^i, Y^i))$$

HSI (X) versus DJIA (Y)









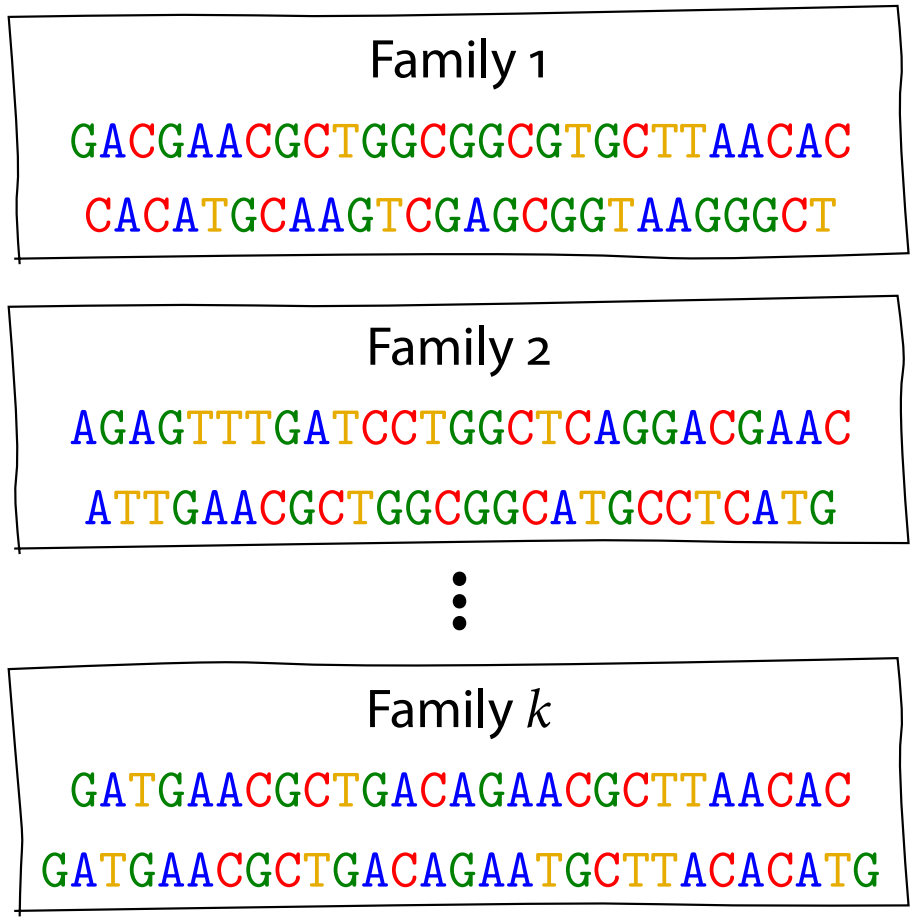
Outline of the talk

- Brief overview of universal probability assignment
- Directed information and its application to causality inference
- Classification of DNA/RNA sequences using universal probability

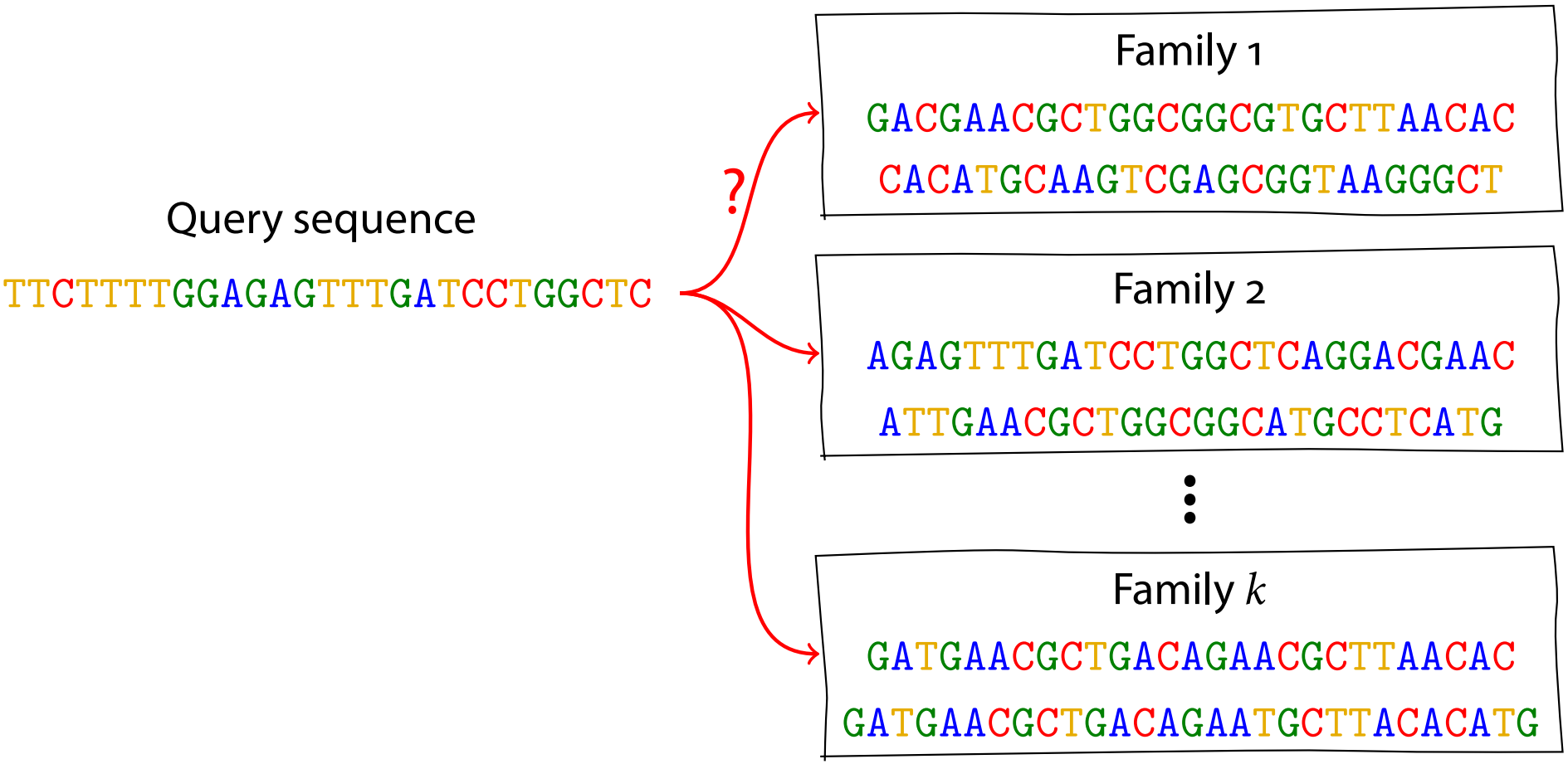
Classification of nucleic acid sequences

Query sequence

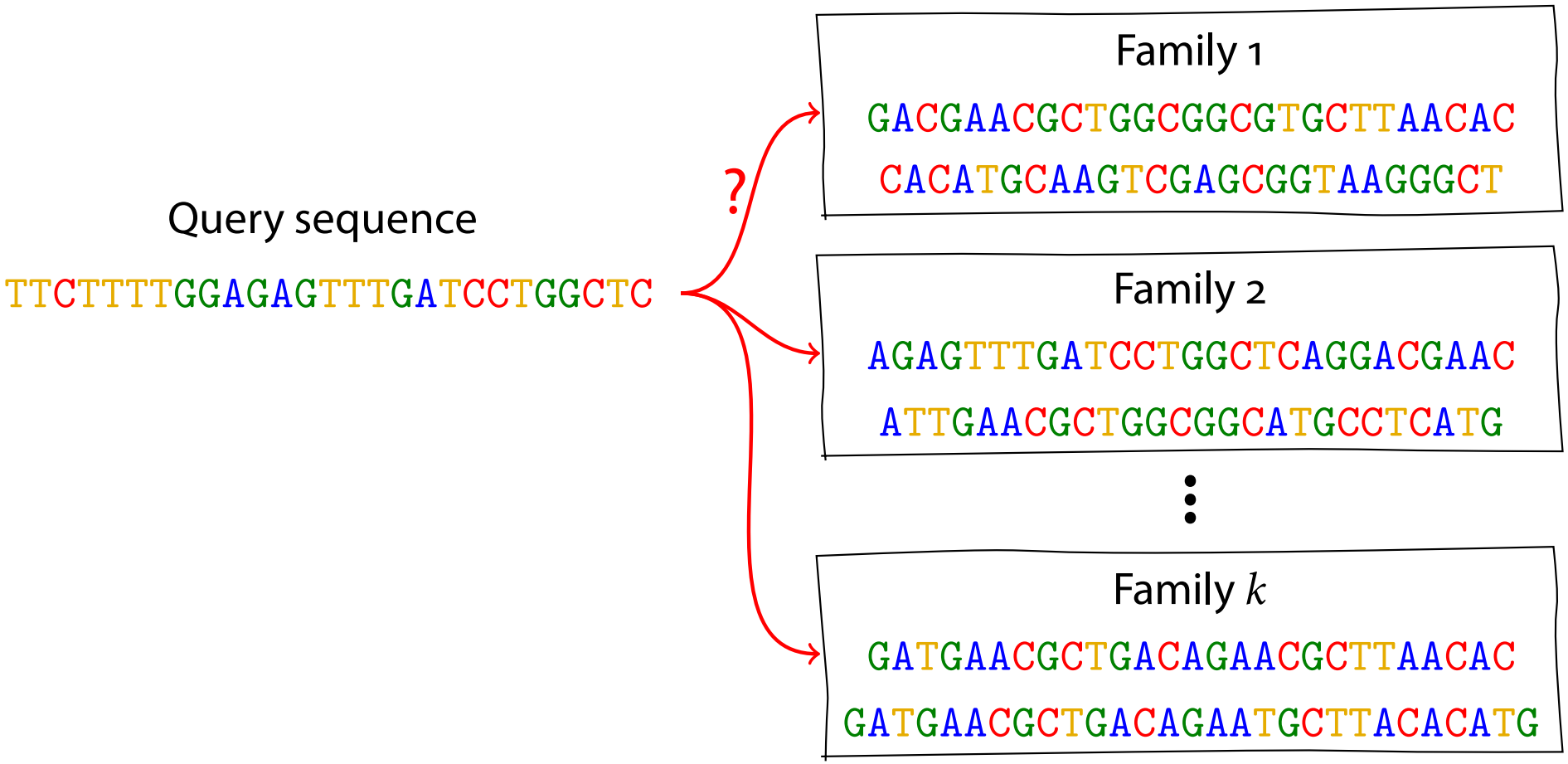
TTCTTTTGGAGAGTTTGATCCTGGCTC



Classification of nucleic acid sequences



Classification of nucleic acid sequences

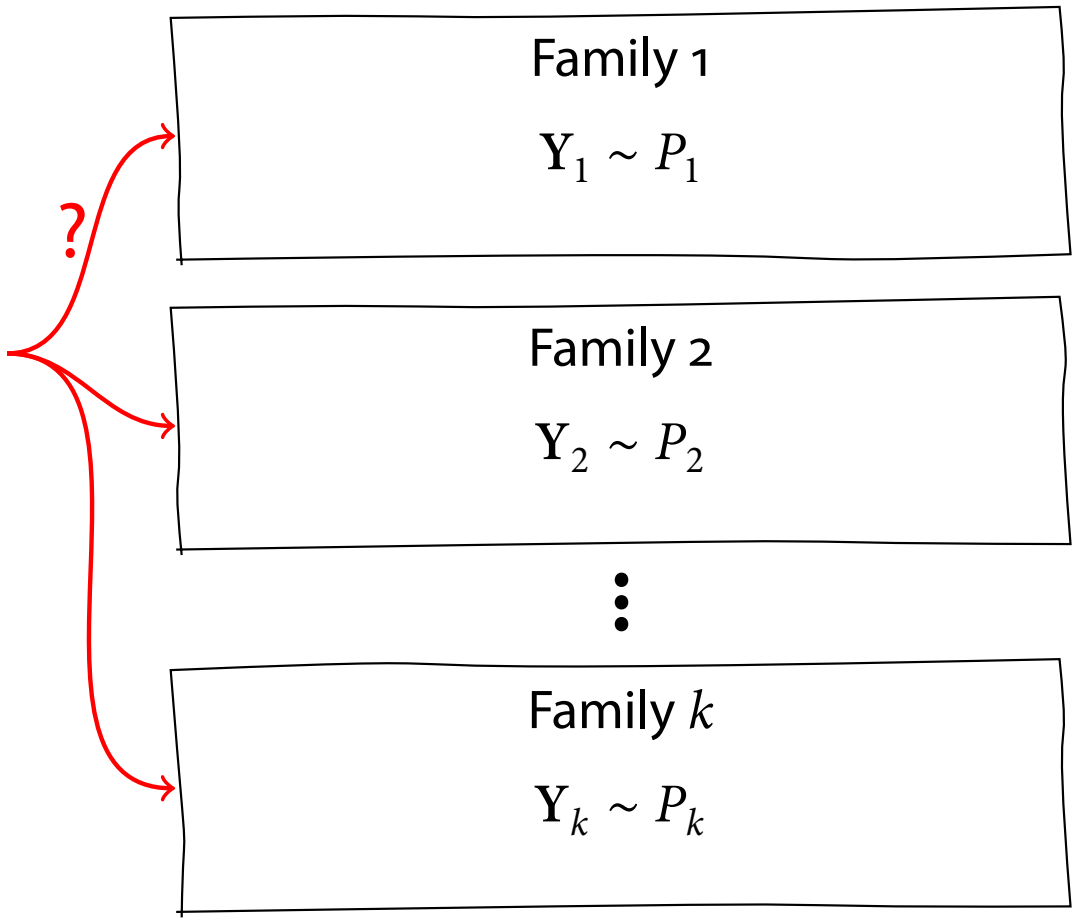


- Alignment-based methods: BLAST, USEARCH, UBLAST, caBLAST, BLAT, …
- Model/feature-based methods: nhmmer, ICM, RDP, …

Classification of nucleic acid sequences

Query sequence

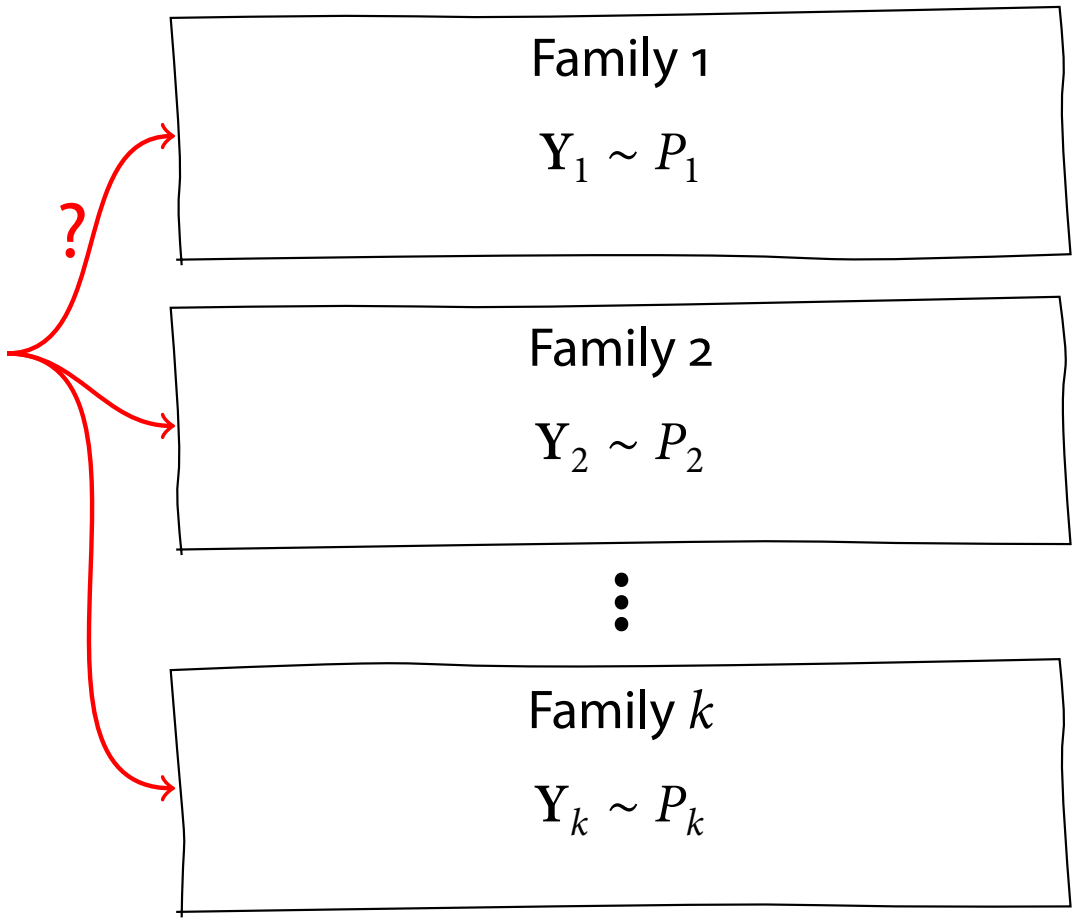
$\mathbf{X}$



Classification of nucleic acid sequences

Query sequence

$\mathbf{X}$

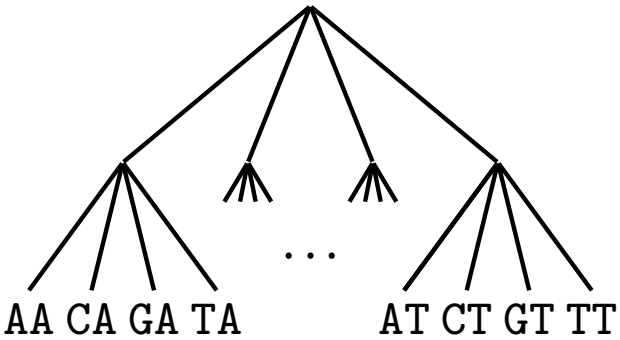


Were $P_1, \ldots, P_k$ known ...

$$j^* = \arg\max_j P_j(\mathbf{X})$$

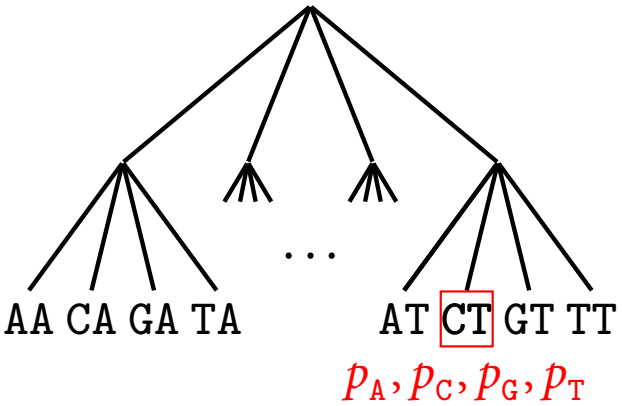
Method

- Context tree models



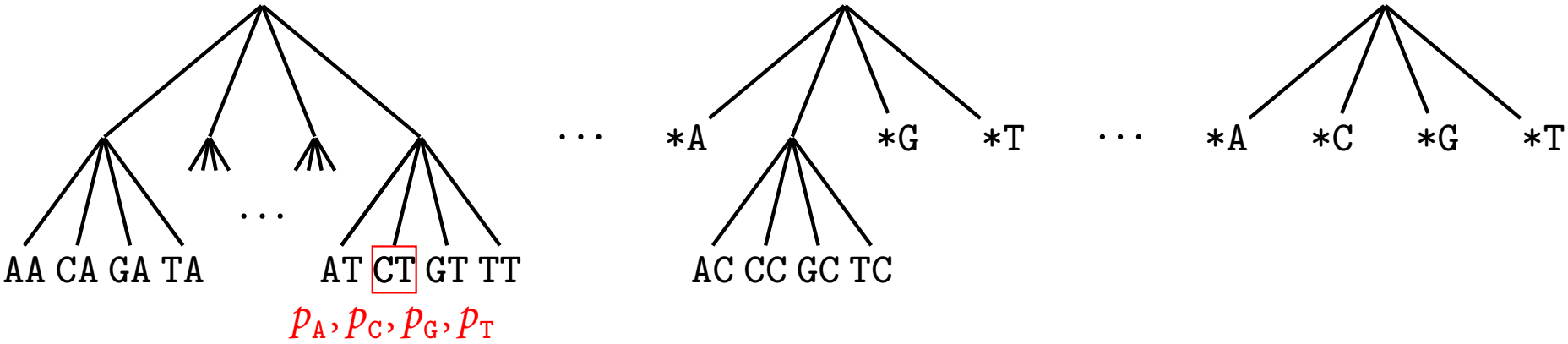
Method

- Context tree models



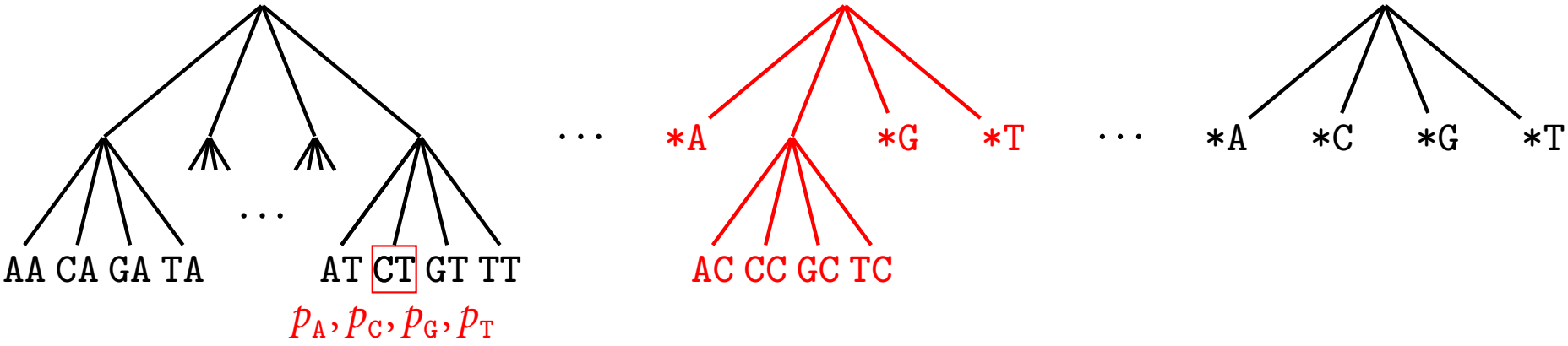
Method

- Context tree models



Method

- Context tree models



Modeling

For each family j and its sequence $\mathbf{Y}_j$ find the best context tree model

$$M_j^* = \arg\max_M Q_M(\mathbf{Y}_j)$$

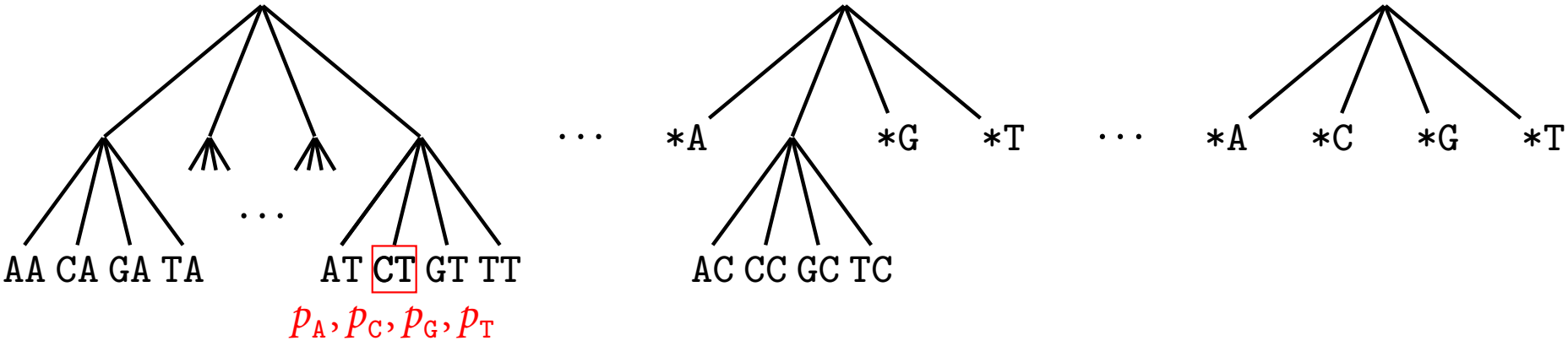
- Q_M: Universal prob. for model M

$$Q_{M_j^*} \approx P_j$$

- Simple recursive maximization

Method

- Context tree models



Modeling

For each family j and its sequence $\mathbf{Y}_j$ find the best context tree model

$$M_j^* = \arg\max_M Q_M(\mathbf{Y}_j)$$

- Q_M: Universal prob. for model M

$$Q_{M_j^*} \approx P_j$$

- Simple recursive maximization

Classification

Given a query sequence $\mathbf{X}$ find the best family

$$j^* = \arg\max_j Q_{M_j^*}(\mathbf{X}|\mathbf{Y}_j)$$

- Close approximation of ML

$$Q_{M_j^*}(\mathbf{X}|\mathbf{Y}_j) \approx P_j(\mathbf{X}|\mathbf{Y}_j) \approx P_j(\mathbf{X})$$

- Simple Bayesian update (Dirichlet)

Performance highlights

- Nine RNA datasets of different types (including large pyrosequencing databases)

classification category of the dataset		ID	dataset name (version)	AIFD‡	# families*	# total† sequences	sequence length	ground truth
functional non-coding RNA		RF	Rfam (11.0)	0.33	1,320	170,881	20–1,875	accession
microbial taxonomy	rRNA database (16S, 18S, 23S/28S)	RD	RDP (10.0)	0.08	134	3,838	320–1,833	taxonomy (genus level)
		GG	Greengenes (13.5)	0.12	464	23,142	1,254–2,146	
		SS	SILVA-SSU (119.1)	0.15	313	17,625	902–3,749	
		SL	SILVA-LSU (119)	0.21	107	4,593	1,900–4,954	
	pyrosequencing data (16S rRNA)	AR	Artificial	0.18	60	44,407	40–294	reference sequences
		DV	Divergent	0.14	23	55,466	38–521	
coding/non-coding RNA		CN	RefSeq,Rfam	0.60	2	103,136	22–9,993	specified
		HS	Ensembl (human)	0.67	2	112,180	20–15,945	

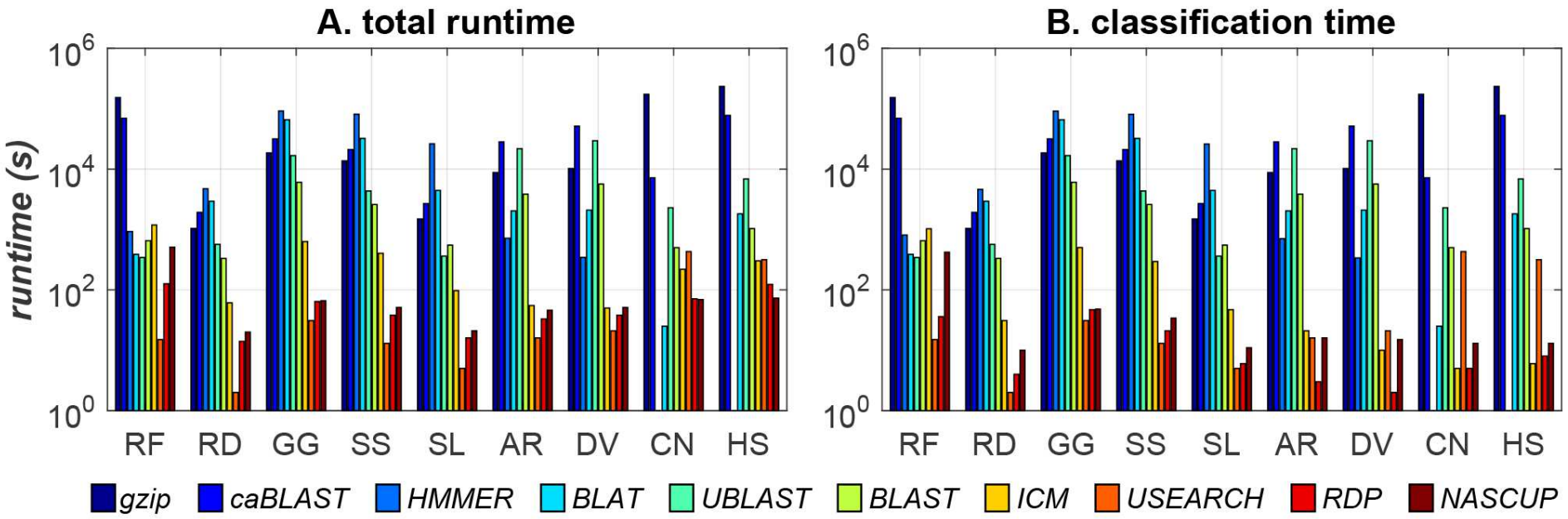
* the number of families with more than 10 sequences

† the total number of sequences after the preprocessing

‡ average intra-family distance (the normalized pairwise distance between the sequences within a family)

Performance highlights

- Nine RNA datasets of different types (including large pyrosequencing databases)
- Comparison to 9 existing methods (BLAST, RDP, USEARCH, HMMER, ICM, …)



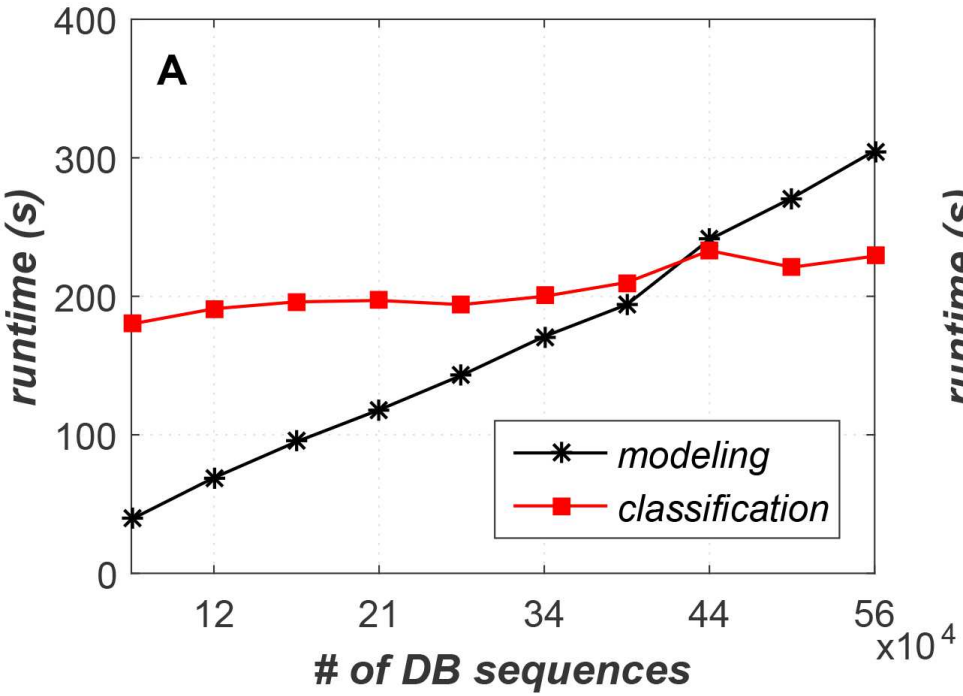
Performance highlights

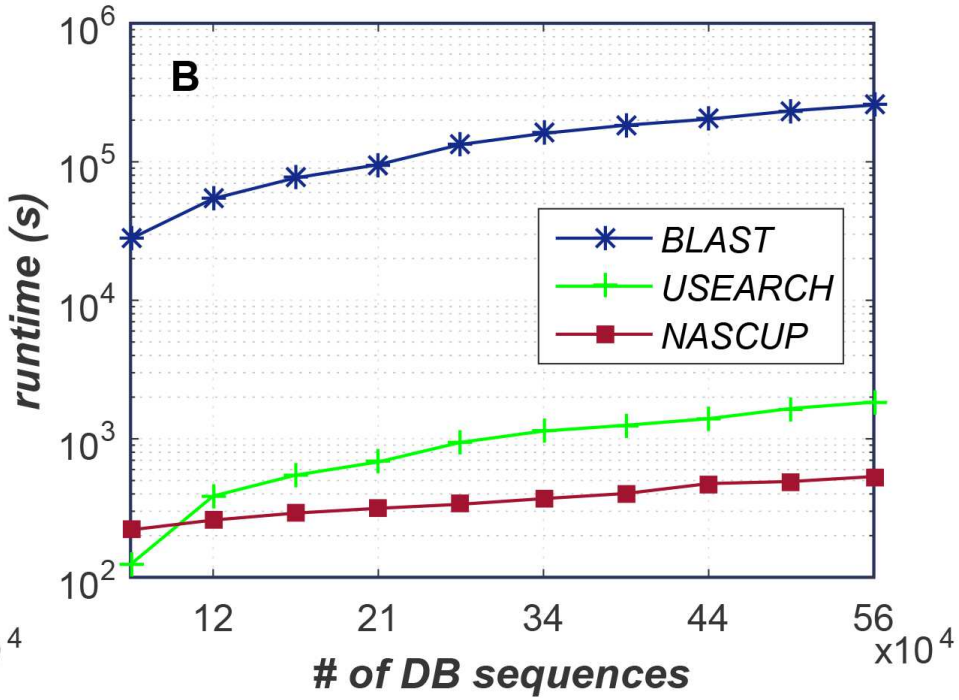
- Nine RNA datasets of different types (including large pyrosequencing databases)
- Comparison to 9 existing methods (BLAST, RDP, USEARCH, HMMER, ICM, …)
- Accuracy of 95.2% (next to 96.5% of BLAST)

method/data	RF	RD	GG	SS	SL	AR	DV	CN	HS	average	geomean
BLAST	95.8%	98.5%	98.4%	96.3%	**97.4%**	96.5%	98.8%	**92.9%**	**94.2%**	**96.5%**	**96.5%**
NASCUP	95.8%	**99.0%**	98.2%	**96.8%**	96.5%	**97.6%**	99.0%	89.3%	84.5%	**95.2%**	**95.1%**
USEARCH	**96.5%**	98.6%	**98.6%**	96.7%	97.4%	89.8%	98.8%	26.7%	84.3%	87.5%	82.5%
UBLAST	79.9%	98.5%	97.9%	95.9%	97.1%	96.4%	98.7%	23.1%	88.2%	86.2%	80.3%
RDP	52.6%	99.0%	98.3%	96.5%	96.9%	97.1%	**99.1%**	60.6%	70.7%	85.6%	83.5%
BLAT	79.1%	97.2%	92.1%	92.3%	95.2%	94.7%	98.9%	18.7%	87.8%	84.0%	77.1%
ICM	93.6%	77.5%	76.7%	39.5%	93.2%	95.0%	98.9%	92.8%	87.3%	83.8%	81.4%
gzip	62.7%	96.3%	90.3%	80.1%	77.6%	80.9%	96.3%	59.5%	69.1%	79.2%	78.2%
caBLAST	39.4%	97.1%	86.9%	90.5%	93.9%	95.5%	97.0%	18.8%	82.8%	78.0%	70.1%
HMMER	96.1%	98.4%	80.1%	14.9%	80.3%	41.9%	64.0%	♯	♯	68.0%	58.8%

Performance highlights

- Nine RNA datasets of different types (including large pyrosequencing databases)
- Comparison to 9 existing methods (BLAST, RDP, USEARCH, HMMER, ICM, …)
- Accuracy of 95.2% (next to 96.5% of BLAST)
- Scalability





Concluding remarks

Maslow's axiom (1966)

If all you have is a hammer,

everything looks like a nail.

Concluding remarks

Maslow's axiom (1966)

If all you have is a hammer,

everything looks like a nail.

- Our hammer: Universal probability q

Concluding remarks

Maslow's axiom (1966)

If all you have is a hammer,

everything looks like a nail.

- Our hammer: Universal probability q
- Versatile and often on par with custom tools

Concluding remarks

Maslow's axiom (1966)

If all you have is a hammer,

everything looks like a nail.

- Our hammer: Universal probability q
- Versatile and often on par with custom tools
- Many classical results, but still more to explore

Concluding remarks

Maslow's axiom (1966)

If all you have is a hammer,

everything looks like a nail.

- Our hammer: Universal probability q
- Versatile and often on par with custom tools
- Many classical results, but still more to explore

Towards information-theoretic data science

References

Bailey, D. H. (1976). *Sequential schemes for classifying and predicting ergodic processes*. Ph.D. thesis, Stanford University.

Cover, T. M. (1975). Open problems in information theory. In *IEEE Joint Works. Inf. Theory*, pp. 35–36.

Cover, T. M. (1991). On the competitive optimality of Huffman codes. *IEEE Trans. Inf. Theory*, IT-37(1), 172–174.

Gallager, R. G. (1974). Capacity and coding for degraded broadcast channels. *Probl. Inf. Transm.*, 10(3), 3–14.

Geweke, J. F. (1982). Measurement of linear dependence and feedback between multiple time series. *J. Amer. Statist. Assoc.*, 77(378), 304–324. With discussion and with a reply by the author.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 424–438.

Jiao, J., Permuter, H. H., Zhao, L., Kim, Y.-H., and Weissman, T. (2013). Universal estimation of directed information. *IEEE Trans. Inf. Theory*, 59(10), 6220–6242.

Kamath, S. and Kim, Y.-H. (2014). Chop and roll: Improving the cutset bound. In *Proc. 52nd Ann. Allerton Conf. Comm. Control Comput.*, Monticello, IL, pp. 921–927.

Kramer, G. (1998). *Directed Information for Channels with Feedback*. Hartung-Gorre Verlag, Konstanz. Dr. sc. thchn. Dissertation, Swiss Federal Institute of Technology (ETH) Zurich.

Krichevsky, R. E. and Trofimov, V. K. (1981). The performance of universal encoding. *IEEE Trans. Inf. Theory*, 27(2), 199–207.

References (cont.)

Laplace, P. S. (1812). *Théorie analytique des probabilités*. Mme Ve Courcier.

Marko, H. (1966). Die Theorie der bidirektionalen Kommunikation und ihre Anwendung auf die Nachrichtenübermittlung zwischen Menschen (Subjektive Information). *Kybernetik*, 3(3), 128–136.

Marko, H. (1973). The bidirectional communication theory: A generalization of information theory. *IEEE Trans. Comm.*, 21(12), 1345–1351.

Maslow, A. H. (1966). *The Psychology of Science: A Reconnaissance*. Harper & Row, New York.

Massey, J. L. (1990). Causality, feedback, and directed information. In *Proc. Int. Symp. Inf. Theory Appl.*, Honolulu, HI, pp. 303–305.

Merhav, N. and Feder, M. (1998). Universal prediction. *IEEE Trans. Inf. Theory*, IT-44(6), 2124–2147.

Ornstein, D. (1978). Guessing the next output of a stationary process. *Israel J. Math.*, 30, 292–296.

Permuter, H. H. (2008). *Capacity of finite state channels with time-invariant feedback*. Ph.D. thesis, Stanford University, Stanford, CA.

Permuter, H. H., Kim, Y.-H., and Weissman, T. (2011). Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. *IEEE Trans. Inf. Theory*, 57(3), 3248–3259.

Quinn, C. J., Coleman, T. P., Kiyavash, N., and Hatsopoulos, N. G. (2011). Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J. Comput. Neurosci.*, 30, 17–44.

Sims, C. A. (1972). Money, income, and causality. *Am. Econ. Rev.*, 62(4), 540–552.

References (cont.)

Weissman, T., Kim, Y.-H., and Permuter, H. H. (2013). Directed information, causal estimation, and communication in continuous time. *IEEE Trans. Inf. Theory*, 59(3), 1271–1287.

Willems, F. M. J., Shtarkov, Y. M., and Tjalkens, T. J. (1995). The context-tree weighting method: Basic properties. *IEEE Trans. Inf. Theory*, 41(3), 653–664.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory*, IT-24(5), 530–536.