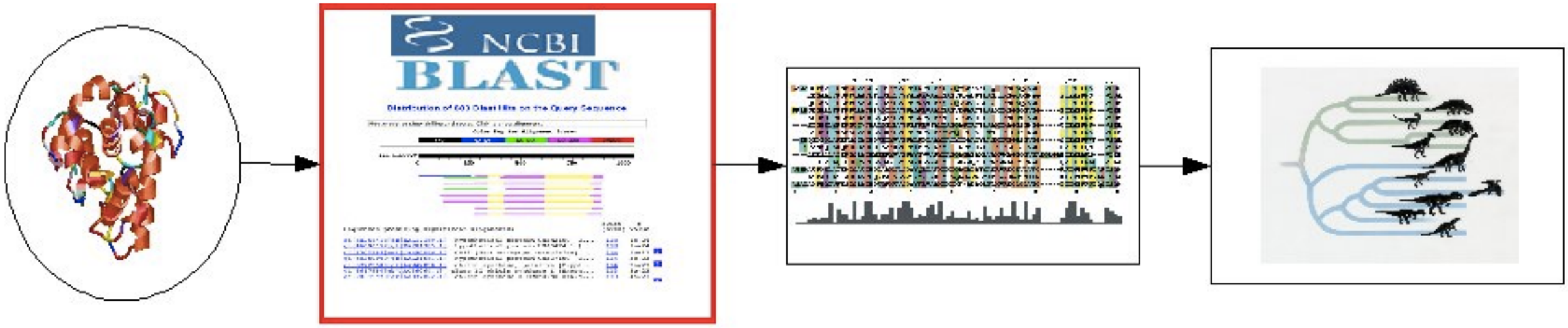
Probability

Random processes in biology

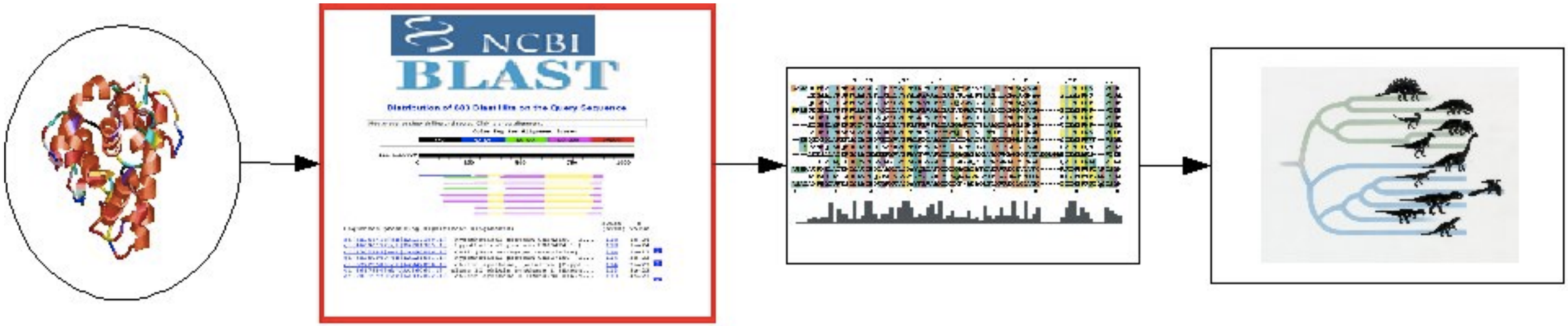
- Brownian motion and related processes
  - Birth-death process
- Molecular simulations and systems biology
- Population genetic models
- Evolution and ecology

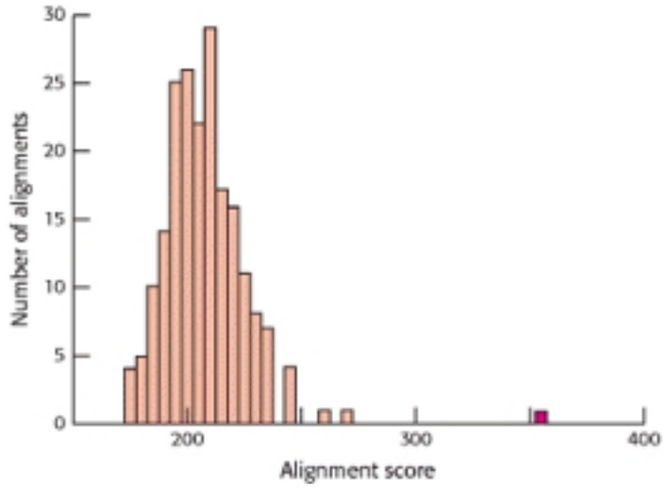
- Information theory
- Statistical machine learning

Characterizing a protein



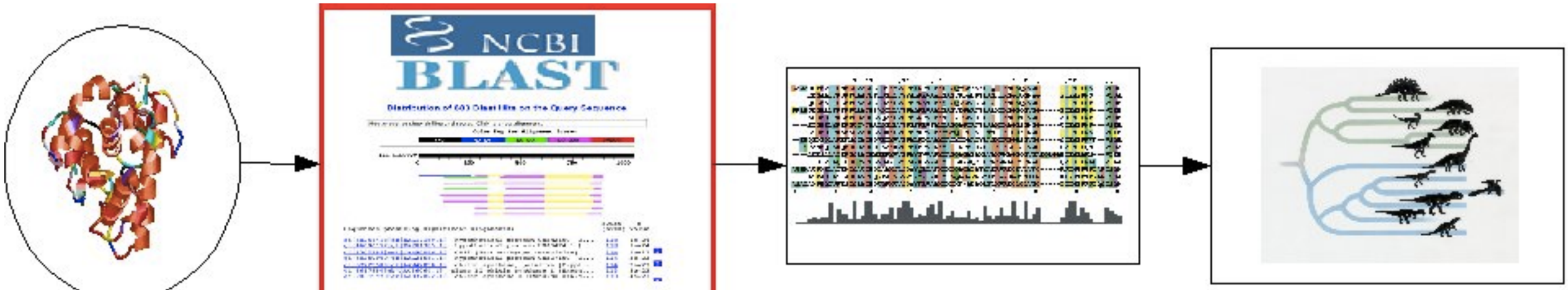
Characterizing a protein

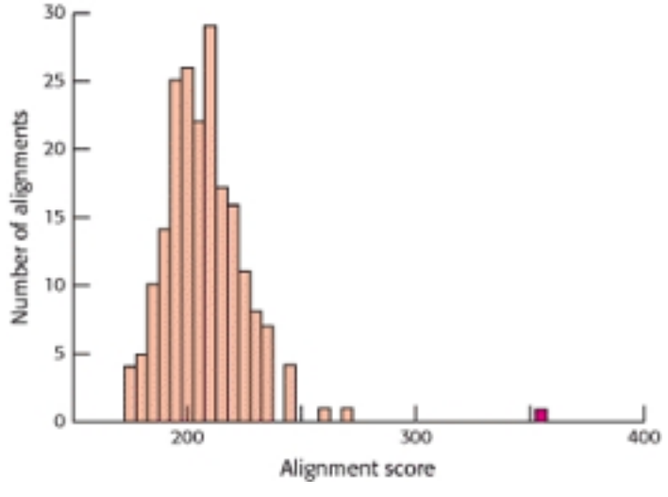




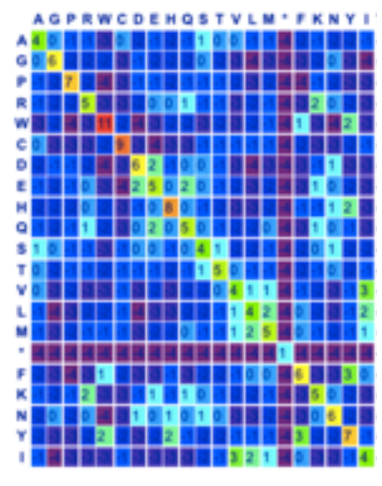
Extreme Value Statistics

Characterizing a protein



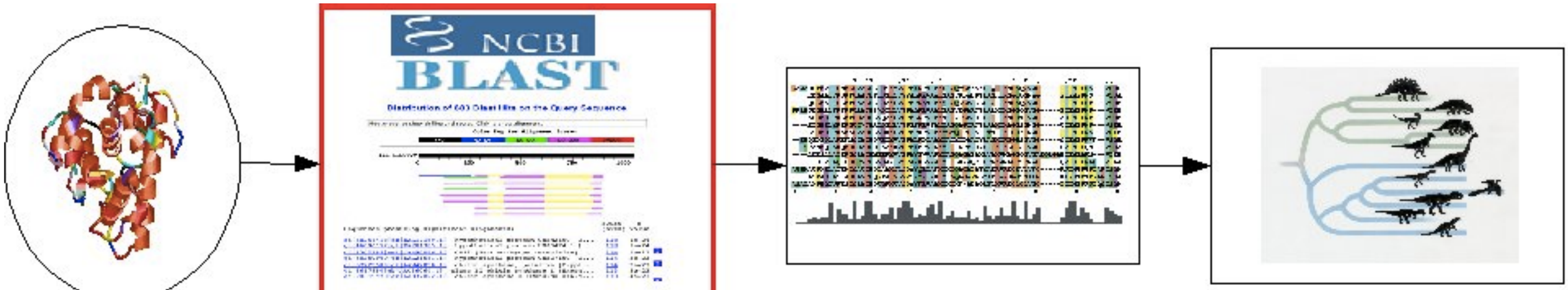


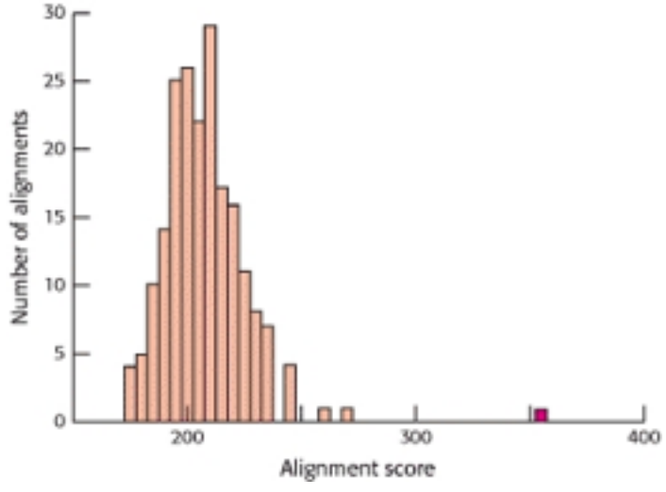
Extreme Value Statistics



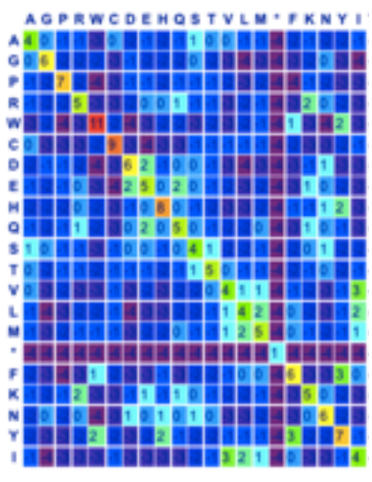
Information Theory

Characterizing a protein

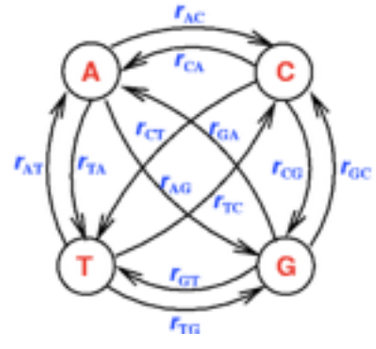




Extreme Value Statistics



Information Theory



Molecular Evolution

Just Generally Useful

- As a biomedical scientist/engineer
- As a [healthcare] consumer

- Consider the following question…

Genetic testing

A rare genetic disease is discovered. Although only one in a million people carry it, you consider getting screened. You are told that the genetic test is extremely good; it is 100% sensitive (it is always correct if you have the disease).

Would you take the test?

Genetic testing

A rare genetic disease is discovered. Although only one in a million people carry it, you consider getting screened. You are told that the genetic test is extremely good; it is 100% sensitive (it is always correct if you have the disease).

NEED MORE INFORMATION!

Genetic testing

A rare genetic disease is discovered. Although only one in a million people carry it, you consider getting screened. You are told that the genetic test is extremely good; it is 100% sensitive (it is always correct if you have the disease).

Let's assume the test is free in terms of time & money (it's covered by your plan, you're getting bloodwork done anyway) and the disease is treatable.

Would you take the test?

Genetic testing

A rare genetic disease is discovered. Although only one in a million people carry it, you consider getting screened. You are told that the genetic test is extremely good; it is 100% sensitive (it is always correct if you have the disease).

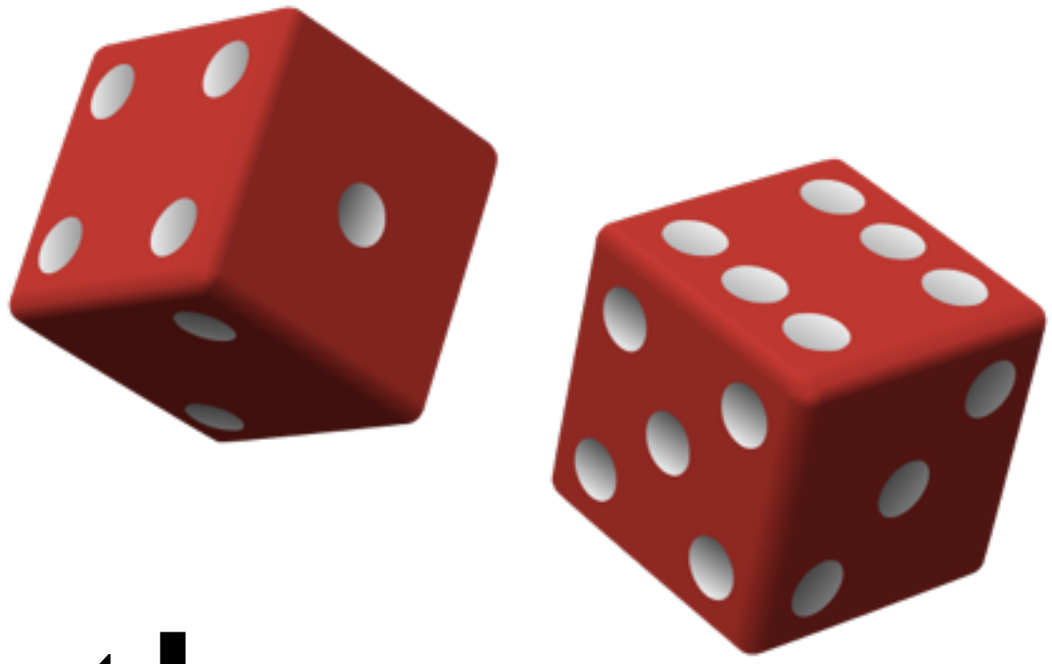
Let's assume the test is free in terms of time & money (it's covered by your plan, you're getting bloodwork done anyway) and the disease is treatable.

STILL NEED MORE INFORMATION!

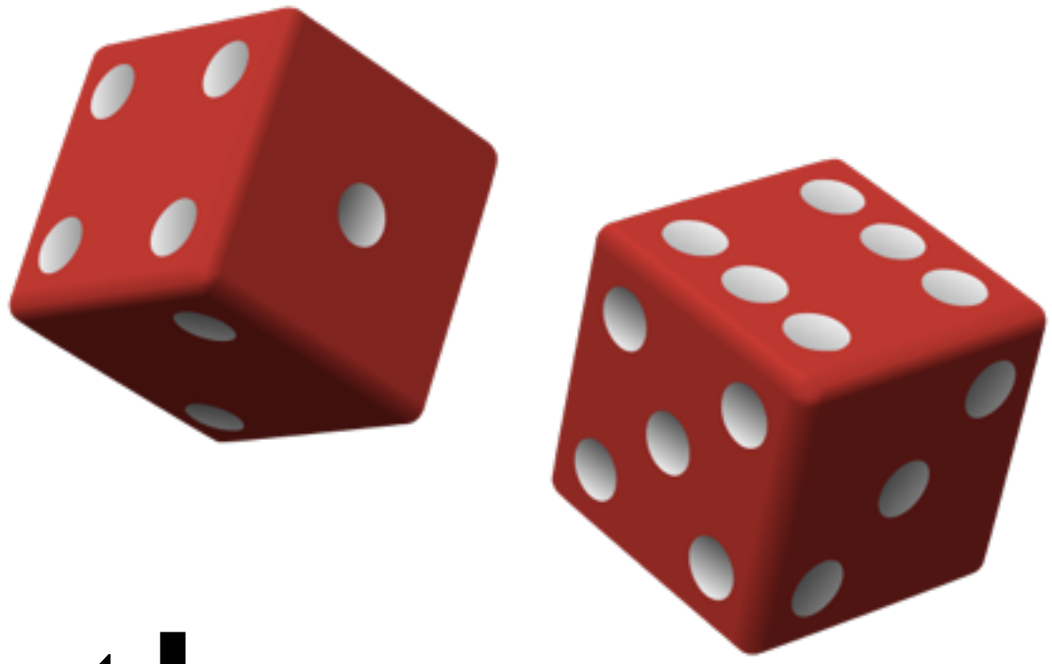
Genetic testing

A rare genetic disease is discovered. Although only one in a million people carry it, you consider getting screened. You are told that the genetic test is extremely good; it is 100% sensitive (it is always correct if you have the disease) **and 99.99% specific (it gives a false positive result only 0.01% of the time).**

Would you take the test?



To answer these questions, we need probability theory



To answer these questions, we need probability theory

- Basic defs: random variables & distributions
- Bayes' Theorem
- Random processes in biology

Basic definitions

Discrete and continuous distributions.

Let $\mathcal{X}, \mathcal{Y}$ = sets of possible values for respective *random variables* (**rv**'s) x, y.

Can often assume that x, y are numbers, e.g. integers or reals (discrete integers or continuous real numbers) or vectors of integers/reals.

Normalization constraints for probability distributions (discrete rv's) and density functions (continuous rv's):

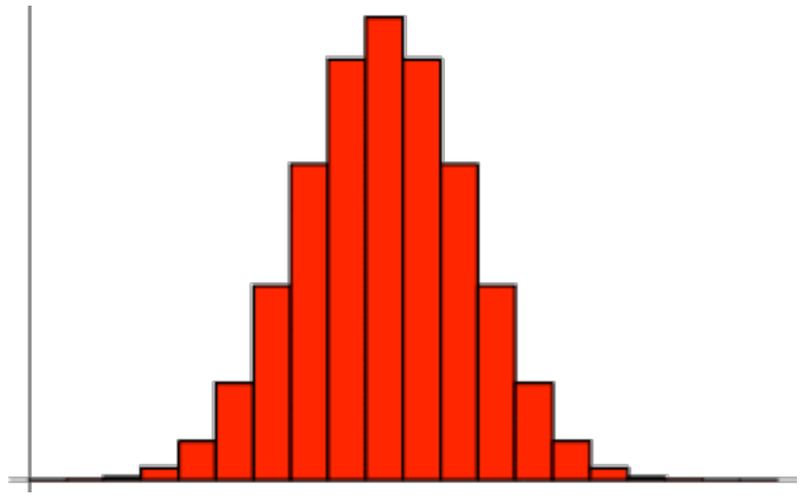
$$\sum_{x \in \mathcal{X}} P(x) = 1 \text{ (discrete; } P(x) \text{ is } \textit{probability distribution})$$

$$\int_{x=0}^{\infty} p(x)dx = 1 \text{ (continuous; } p(x) \text{ is } \textit{probability density function} \text{ or } \textbf{pdf})$$

NB $P(x), p(x) \geq 0$ but only $P(x) \leq 1$. $p(x)$ can be > 1.

IMPORTANT NOTE: probability measures the likelihood of a particular *event*. Thus, when we write $P(x)$, this is actually a shorthand for $P(x = a)$ where x is a random variable (that can take many values) and a is a particular, constant value.

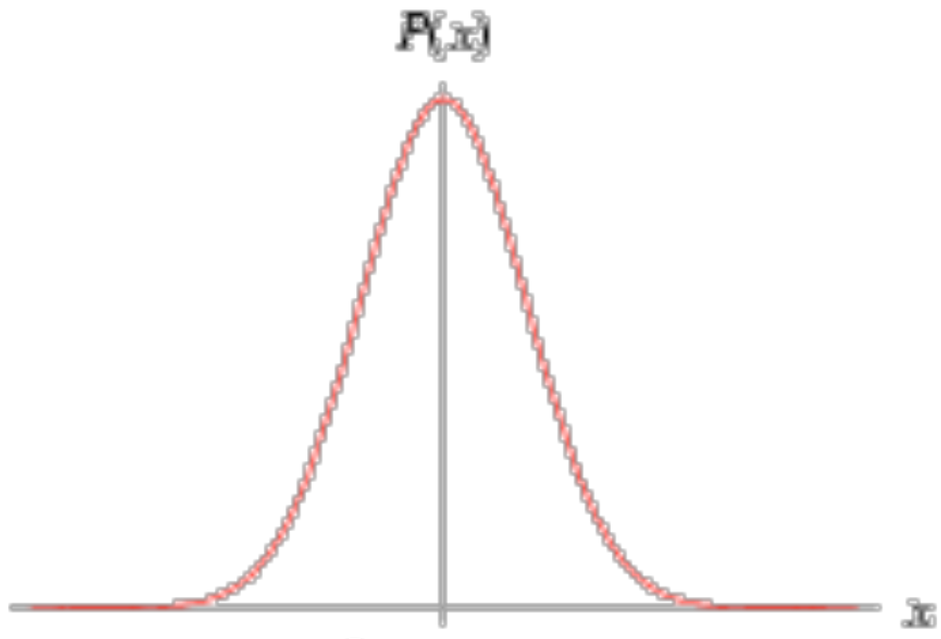
For example, if x is a dice roll then we can write $P(x = 1)$, $P(x = 2)$ and so on up to $P(x = 6)$. This notation can get cumbersome, so we will typically just write this as $P(x)$.



Binomial

$$f(k; n, p) = \Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$



Gaussian

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Cumulative distributions

- Density function:

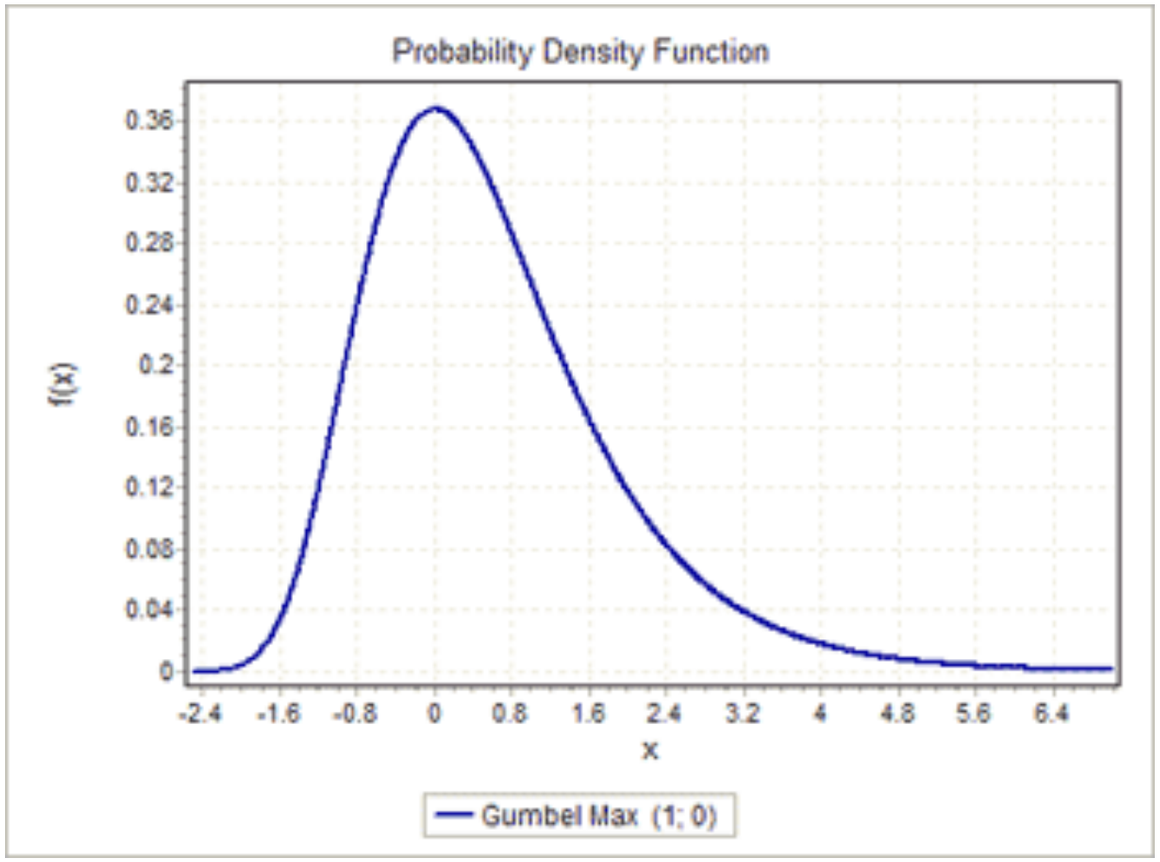
$$P(a \leq x \leq a + da) = p(a)da$$

- Cumulative distribution function:

$$P(x \leq a) = \int_{-\infty}^{a} p(x')dx'$$

EVD: Extreme Value (Gumbel) Distribution

- Distribution of the <u>maximum</u> of N random variables, e.g.
  - max of N normal rv's
- Arises in situations where a variable is random but (by construction) is also maximized somehow, e.g.
  - score of best alignment between two random protein sequences
  - score of best hybridization energy between two RNA sequences
  - etc.

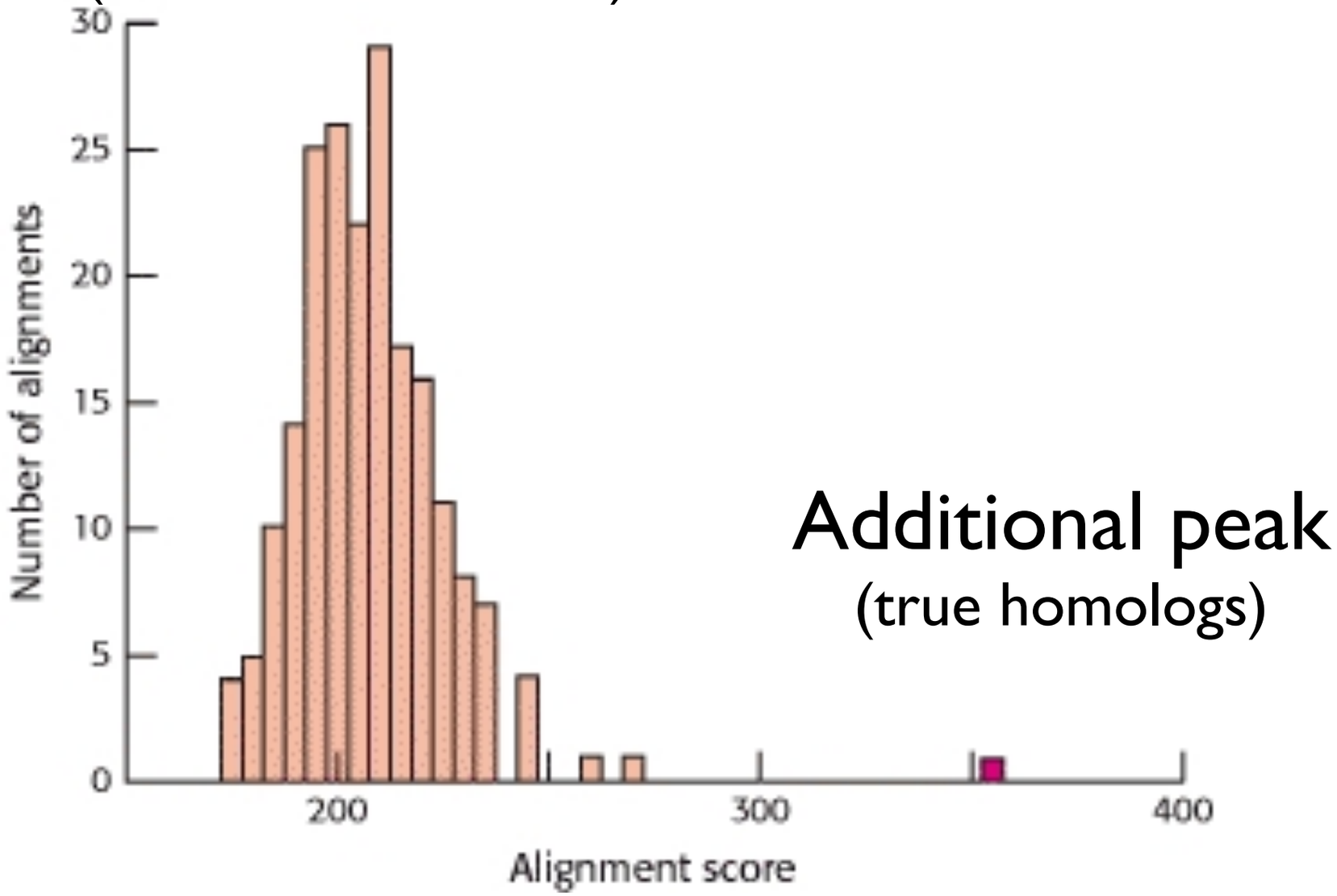


$$f(x) = e^{-x}e^{-e^{-x}}.$$

The "P-value" for a BLAST search is obtained from the integral of the EVD (i.e. the cumulative EVD).

Score distribution: Sequence homology search

Extreme Value Distribution
(chance similarities)



Additional peak
(true homologs)

More definitions

- Conditional, joint and marginal probability; independence.

$$
\begin{aligned}
P(x,y) &= \text{joint probability distribution for } x \text{ and } y \\
P(x) &= \sum_y P(x,y) \text{ (marginal probability distribution for } x) \\
P(y) &= \sum_x P(x,y) \text{ (marginal probability distribution for } y) \\
P(x|y) &= P(x,y)/P(y) \text{probability distribution for } x \text{ conditional on } y \\
P(y|x) &= P(x,y)/P(x) \text{probability distribution for } y \text{ conditional on } x \\
P(x,y) &= P(x|y)P(y)
\end{aligned}
$$

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

Normalization

$$\sum_x \sum_y P(x,y) = 1 \text{ (normalization of joint probability distribution)}$$

$$\sum_x P(x) = 1 \text{ (normalization of marginal probability distribution)}$$

$$\sum P(x|y) = 1 \; \forall y \text{ (normalization of conditional probability distribution)}$$

Similarly for probability density functions:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y) dx dy = 1$$

etc. (replace sums by integrals)

Independence

If $P(x, y) = P(x)P(y)$ then x and y are *independent.*

Example of rv's that may be assumed to be independent: pick any two nucleotides in a genome.

Example of rv's that are probably *not* independent: pick any two *adjacent* nucleotides in a genome. (Another example: pick any two *homologous* nucleotides in two related genomes.)

Why are these cases probably not independent?

"I.I.D."

If $x_1 \ldots x_N$ are a sequence of rv's that are all independent, and each one has the same probability distribution, then we say that they are "Independent and Identically Distributed" or "IID". (For example, rolling the same die 100 times yields 100 IID RV's.)

"Uniform"

Let $|\mathcal{X}|$ be the number of elements in $\mathcal{X}$ (assuming $\mathcal{X}$ is discrete & finite) and similarly for $\mathcal{Y}$.

If all possible values of x have exactly the same probability, $P(x) = 1/|\mathcal{X}|$, then $P(x)$ is often said to be a "flat" or "uniform" distribution. (As an example, consider a fair die, which should have probability $1/6$ of showing any given value.)

Question

- How many bases in the human genome?

Question

- How many bases in the human genome?
  - About 3 billion

Question

- How many bases in the human genome?
  - About 3 billion
- How many bases in a CRISPR guide RNA?

Question

- How many bases in the human genome?
  - About 3 billion
- How many bases in a CRISPR guide RNA?
  - About 18-25, let's say 20

Question

- How many bases in the human genome?
  - About 3 billion
- How many bases in a CRISPR guide RNA?
  - About 18-25, let's say 20
- Is a CRISPR guide RNA sufficient to specify a unique address in the human genome?

Question

- Is a (20nt) CRISPR guide RNA sufficient to specify a unique address in the (3Gb) human genome?

20 nucleotides = 4^{20} possibilities

$4^{20} = 2^{40} \simeq 10^{12} > 3*10^{9}$

Question

- Is a CRISPR guide RNA sufficient to specify a unique address in the human genome?

- We have assumed the human genome is IID; it's not.
- A given 20-mer may be repeated
- Some CRISPR protocols use two guides

Related question

- While examining a 10kb promoter you observe that a given 5-nucleotide motif occurs 8 times. Is this unusually frequent?

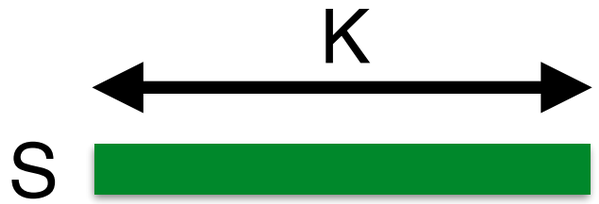
Related question

- While examining a 10kb promoter you observe that a given 5-nucleotide motif occurs 8 times. Is this unusually frequent?

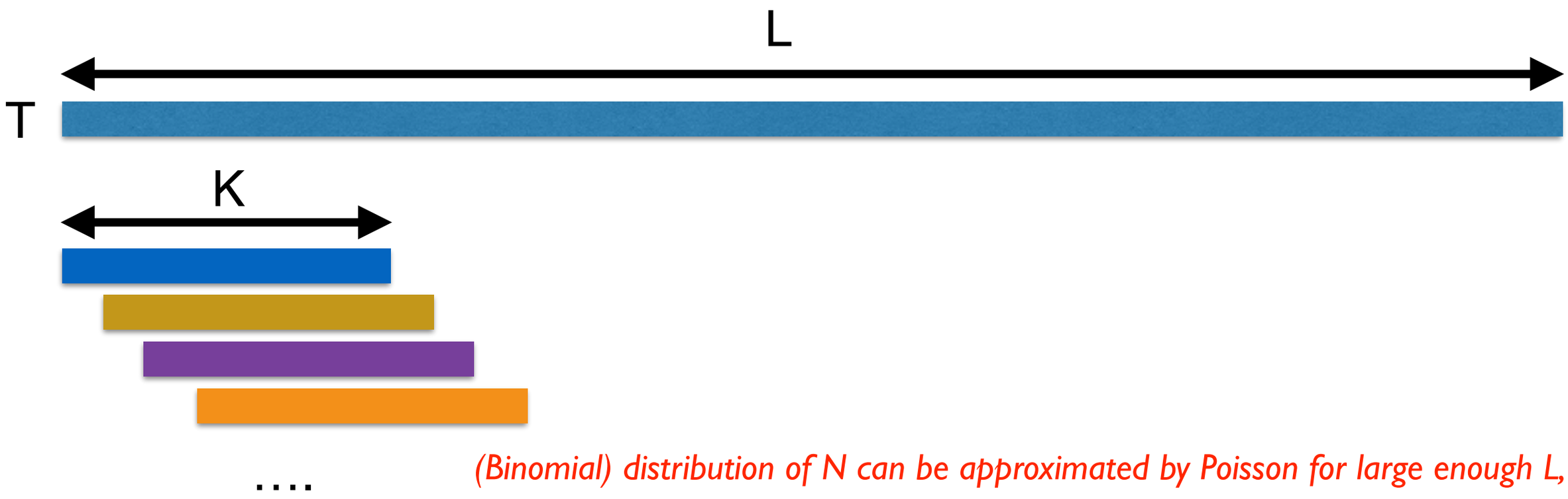
- A 5-nucleotide motif will occur on average once every 4^5 positions in an IID sequence, i.e. once every kilobase or so.
  So no, this is not significant. Bad luck

Subsequences

- Uniform I.I.D. DNA sequences S, T
  Respective lengths K, L≫K
- Probability of S is $(1/4)^K$
- P([n'th K-mer of T] = S) = $(1/4)^K$
- T contains N instances of S



- Approximate each overlapping K-mer subseq of T as independent
  Then N is binomially-distributed with mean $L*(1/4)^K$



(Binomial) distribution of N can be approximated by Poisson for large enough L, K

Expectation & variance

- Expectation, variance; moments.

$$\langle x \rangle \equiv E[x] = \sum_{x \in \mathcal{X}} xP(x) \text{ (expectation of } x\text{; two alternative notations)}$$

More generally, can take the expectation of some function $f(x)$

$$\langle f(x) \rangle \equiv E[f(x)] = \sum_{x \in \mathcal{X}} f(x)P(x) \text{ (expectation of } f(x)\text{)}$$

For two alternative probability distributions, $P(x)$ and $Q(x)$, write $\langle x \rangle_P$ and $\langle x \rangle_Q$ to distinguish.

Continuous distributions: substitute $\int p(x) \ldots dx$ for $\sum_{x \in \mathcal{X}}$.

$$\begin{aligned} \langle x^n \rangle &= \sum_{x \in \mathcal{X}} x^n P(x) \text{ (the ``}n\text{'th moment'' of } x\text{)} \\ V[x] &= \langle x^2 \rangle - (\langle x \rangle)^2 \text{ (variance of } x\text{)} \\ \sigma_x &= \sqrt{V[x]} \text{ (standard deviation of } x\text{)} \end{aligned}$$

Bayes' Theorem

- Bayes' theorem; prior and posterior distributions; "evidence".

$$P(x|y) = \frac{P(x,y)}{P(y)} = \frac{P(y|x)P(x)}{P(y)}$$

Common interpretation: x is "model" or "hypothesis"; y is "data". $P(x)$ is "prior probability", $P(y|x)$ is "likelihood", $P(y)$ is "evidence", $P(x|y)$ is "posterior probability".

Typically $P(x)$ and $P(y|x)$ are specified, and you need to calculate

$$\begin{aligned} P(x,y) &= P(x)P(y|x) \\ P(y) &= \sum_x P(x,y) \\ P(x|y) &= P(x,y)/P(y) \end{aligned}$$

Thomas Bayes(?), c.1702-1761

Example: d6 & d20 in a bag

Two dice in a bag

- If I tell you I picked one of these two dice at random, what's the chance it was the d6?
- If I tell you I rolled a 5 with it, does that probability change? (I hope so...)

Example: Fall ’05 admissions

	Accepted *(A=1)*	Rejected *(A=0)*	**Total**
California resident *(C=1)*	8,493	22,206	**30,699**
California nonresident *(C=0)*	1,162	5,098	**6,260**
Total	**9,655**	**27,304**	**36,959**

P(C=1)

P(A=1 | C=1)

P(A=1, C=1)

P(C=1 | A=1)

P(A=1)

Example: Fall ’05 admissions

	Accepted *(A=1)*	Rejected *(A=0)*	**Total**
California resident *(C=1)*	0.23	0.6	**0.83**
California nonresident *(C=0)*	0.03	0.14	**0.17**
Total	**0.26**	**0.74**	**1**

P(C=1)

P(A=1 | C=1)

P(A=1, C=1)

P(C=1 | A=1)

P(A=1)

Example: Fall '05 admissions

	Accepted *(A=1)*	Rejected *(A=0)*	**Total**
California resident *(C=1)*	0.23	0.6	**0.83**
California nonresident *(C=0)*	0.03	0.14	**0.17**
Total	**0.26**	**0.74**	**1**

P(C=1)

P(A=1 | C=1)

P(A=1, C=1)

P(C=1 | A=1)

P(A=1)

Example: Fall ’05 admissions

	Accepted *(A=1)*	Rejected *(A=0)*	**Total**
California resident *(C=1)*	0.23	0.6	**0.83**
California nonresident *(C=0)*	0.03	0.14	**0.17**
Total	**0.26**	**0.74**	**1**

P(C=1)

P(A=1 | C=1)

P(A=1, C=1)

P(C=1 | A=1)

P(A=1)

Example: Fall ’05 admissions

	Accepted *(A=1)*	Rejected *(A=0)*	**Total**
California resident *(C=1)*	0.23	0.6	**0.83**
California nonresident *(C=0)*	0.03	0.14	**0.17**
Total	**0.26**	**0.74**	**1**

P(C=1)

P(A=1 | C=1)

P(A=1, C=1)

P(C=1 | A=1)

P(A=1)

Example: Fall ’05 admissions

	Accepted *(A=1)*	Rejected *(A=0)*	**Total**
California resident *(C=1)*	0.23	0.6	**0.83**
California nonresident *(C=0)*	0.03	0.14	**0.17**
Total	**0.26**	**0.74**	**1**

P(C=1)

P(A=1 | C=1)

= 0.23 / 0.83

P(A=1, C=1)

P(C=1 | A=1)

P(A=1)

Example: Fall ’05 admissions

	Accepted *(A=1)*	Rejected *(A=0)*	**Total**
California resident *(C=1)*	0.23	0.6	**0.83**
California nonresident *(C=0)*	0.03	0.14	**0.17**
Total	**0.26**	**0.74**	**1**

P(C=1)

P(A=1 | C=1)
= 0.23 / 0.83

P(A=1, C=1)

P(C=1 | A=1)
= 0.23 / 0.26

P(A=1)

X=disease, Y=symptom

(or test result)

- 1% of women at age forty who participate in routine screening have breast cancer.
- 80% of women with breast cancer will get positive mammographies.
- 9.6% of women without breast cancer will also get positive mammographies.
- A woman in this age group had a positive mammography in a routine screening.
- What is the probability that she actually has breast cancer?

X=disease, Y=symptom

(or test result)

- 1% of women at age forty who participate in routine screening have breast cancer.
- 80% of women with breast cancer will get positive mammographies.
- 9.6% of women without breast cancer will also get positive mammographies.
- A woman in this age group had a positive mammography in a routine screening.
- What is the probability that she actually has breast cancer?

Scary fact: 85% of doctors get this wrong

(Casscells, Schoenberger, and Graboys 1978; Eddy 1982; Gigerenzer and Hoffrage 1995)

Alternative presentation

- 100 out of 10,000 women at age forty who participate in routine screening have breast cancer.
- 80 of every 100 women with breast cancer will get positive mammographies.
- 950 out of 9,900 women without breast cancer will also get positive mammographies.
- If 10,000 women in this age group have a positive mammography in a routine screening…
- About what fraction of them actually have breast cancer?

Alternative presentation

- 100 out of 10,000 women at age forty who participate in routine screening have breast cancer.
- 80 of every 100 women with breast cancer will get positive mammographies.
- 950 out of 9,900 women without breast cancer will also get positive mammographies.
- If 10,000 women in this age group have a positive mammography in a routine screening…
- About what fraction of them actually have breast cancer?

Equally scary fact: 54% of doctors still get it wrong

All relevant probabilities

p(cancer):	0.01	Group 1: 100 women with breast cancer
p(~cancer):	0.99	Group 2: 9900 women without breast cancer
p(positive|cancer):	80.0%	80% of women with breast cancer have positive mammographies
p(~positive|cancer):	20.0%	20% of women with breast cancer have negative mammographies
p(positive|~cancer):	9.6%	9.6% of women without breast cancer have positive mammographies
p(~positive|~cancer):	90.4%	90.4% of women without breast cancer have negative mammographies

p(cancer&positive):	0.008	Group A: 80 women with breast cancer and positive mammographies
p(cancer&~positive):	0.002	Group B: 20 women with breast cancer and negative mammographies
p(~cancer&positive):	0.095	Group C: 950 women without breast cancer and positive mammographies
p(~cancer&~positive):	0.895	Group D: 8950 women without breast cancer and negative mammographies
p(positive):	0.103	1030 women with positive results
p(~positive):	0.897	8970 women with negative results
p(cancer|positive):	7.80%	Chance you have breast cancer if mammography is positive: 7.8%
p(~cancer|positive):	92.20%	Chance you are healthy if mammography is positive: 92.2%
p(cancer|~positive):	0.22%	Chance you have breast cancer if mammography is negative: 0.22%
p(~cancer|~positive):	99.78%	Chance you are healthy if mammography is negative: 99.78%

Genetic testing

A rare genetic disease is discovered. Although only one in a million people carry it, you consider getting screened. You are told that the genetic test is extremely good; it is 100% sensitive (it is always correct if you have the disease) and 99.99% specific (it gives a false positive result only 0.01% of the time). Having recently learned Bayes' theorem, you decide not to take the test.

Why?

(From Durbin et.al. "Biological Sequence Analysis", Cambridge University Press, 1998)

Sometimes the problem is stated less clearly…

- Suppose you have a large barrel containing a number of plastic eggs.
- Some eggs contain pearls, the rest contain nothing.
- Some eggs are painted blue, the rest are painted red.
- Suppose that…
  - 40% of the eggs are painted blue
  - 5/13 of the eggs containing pearls are painted blue
  - 20% of the eggs are both empty and painted red.
- What is the probability that an egg painted blue contains a pearl?

Pearls and eggs

$$P(X=1) = 0.4$$
$$P(X=1|Y=1) = 5/13$$
$$P(X=0,Y=0) = 0.2$$

- X is egg color (0 for red, 1 for blue)
- Y is the pearl (0 if absent, 1 if present)

$$P(Y=1|X=1) = \frac{P(X=1,Y=1)}{P(X=1)} = \frac{P(Y=1)P(X=1|Y=1)}{P(X=1)}$$

$$P(Y=1) = \frac{P(X=0,Y=1)}{P(X=0|Y=1)}$$

$$P(X=0,Y=1) + P(X=0,Y=0) + P(X=1) = 1$$

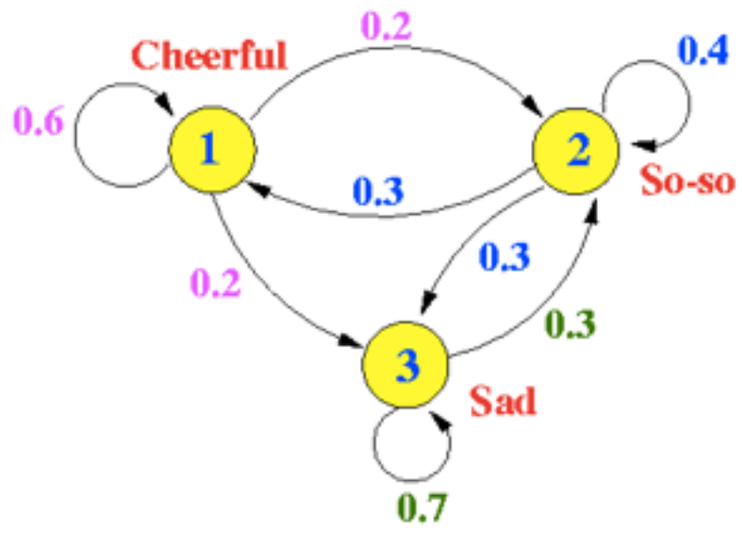
The Monty Hall problem

- We are presented with three doors - red, green, and blue - one of which has a prize.
- We choose the **red** door, but this door is not opened (yet), according to the rules.
- The rules are that the presenter *knows what door the prize is behind, and who must open a door, but is not permitted to open the door we have picked or the door with the prize*.
- The presenter opens the *green* door, revealing that there is **no prize** behind it, and subsequently asks if we wish to change our mind about our initial selection of **red**.
- What are the probabilities that the prize is behind (respectively) the blue and red doors?
- X=prize door, Y=presenter door…

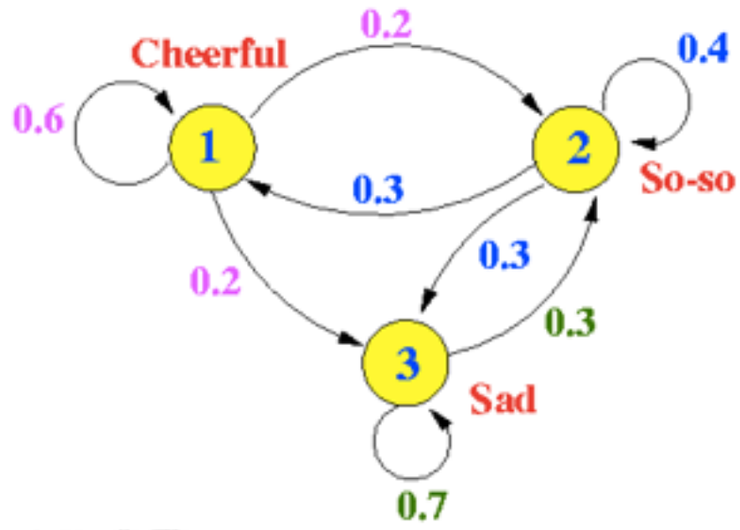
Random processes in biology

- Markov process
  - Random walk (Brownian motion)
  - Birth-death process
- Hardy-Weinberg equilibrium
- Wright-Fisher model
- Coalescent
- Unified Neutral Theory of Biodiversity
- Bayesian network

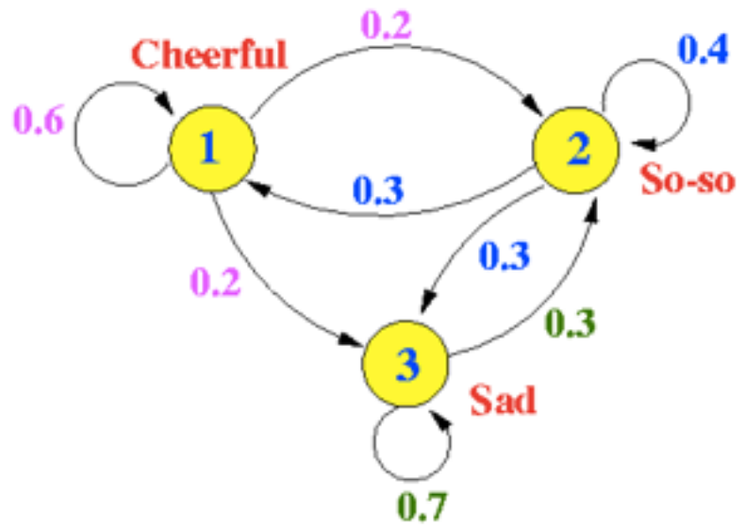
Markov process



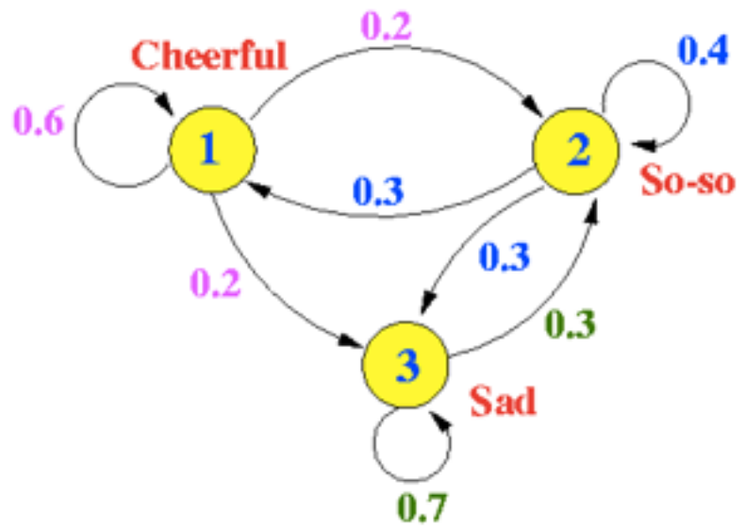
Markov process



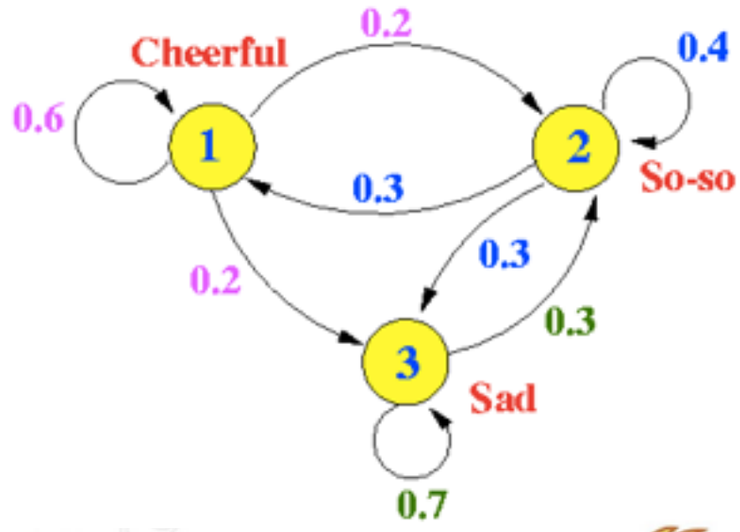
Markov process



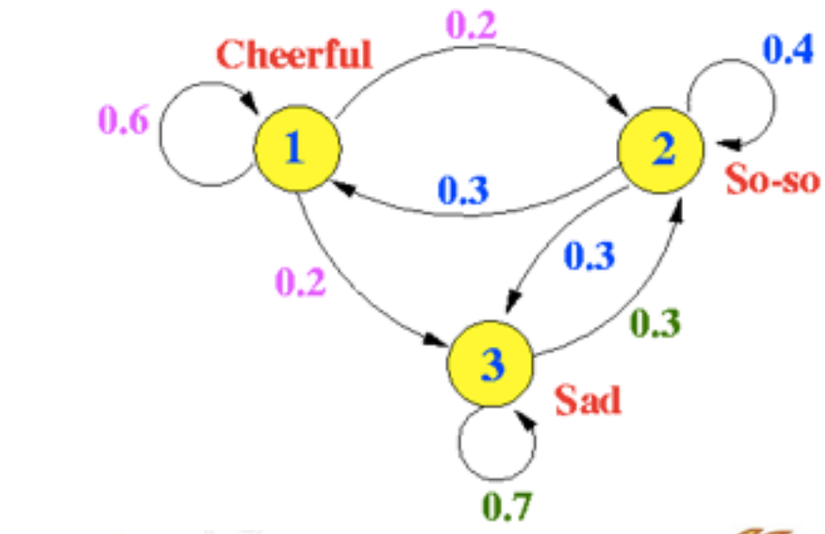
Markov process



Markov process

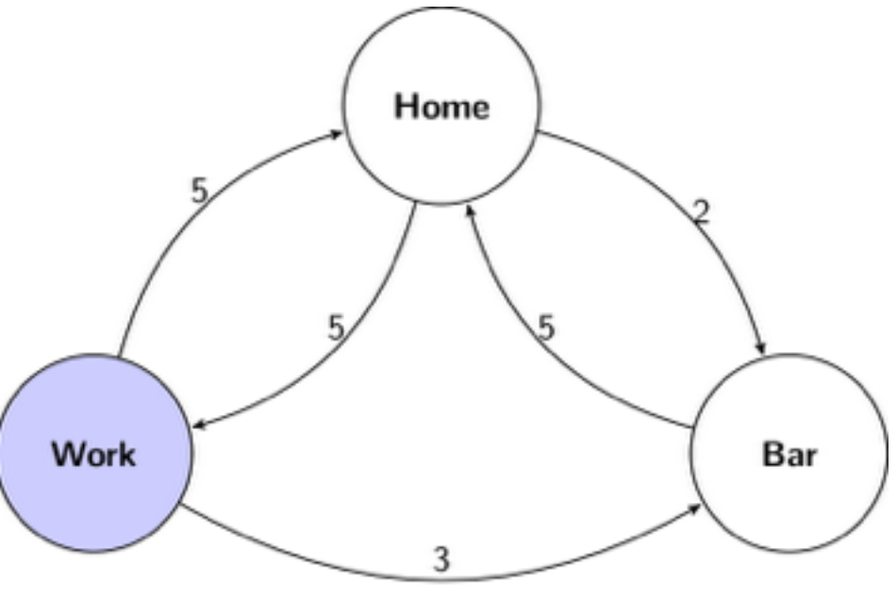


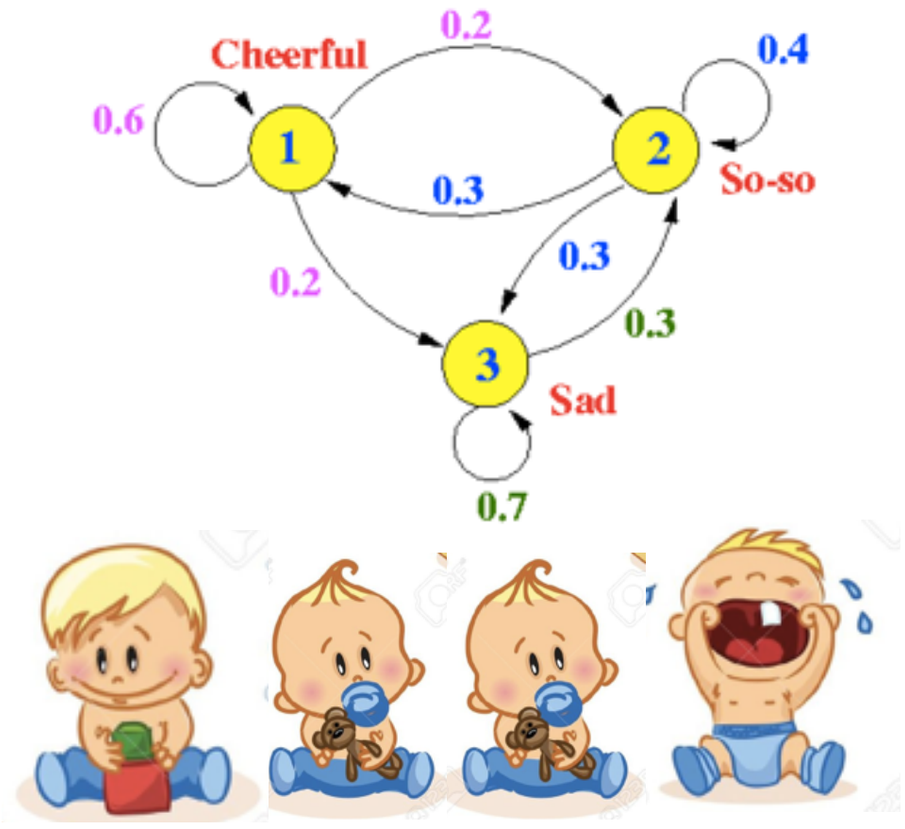
Markov process



etc.

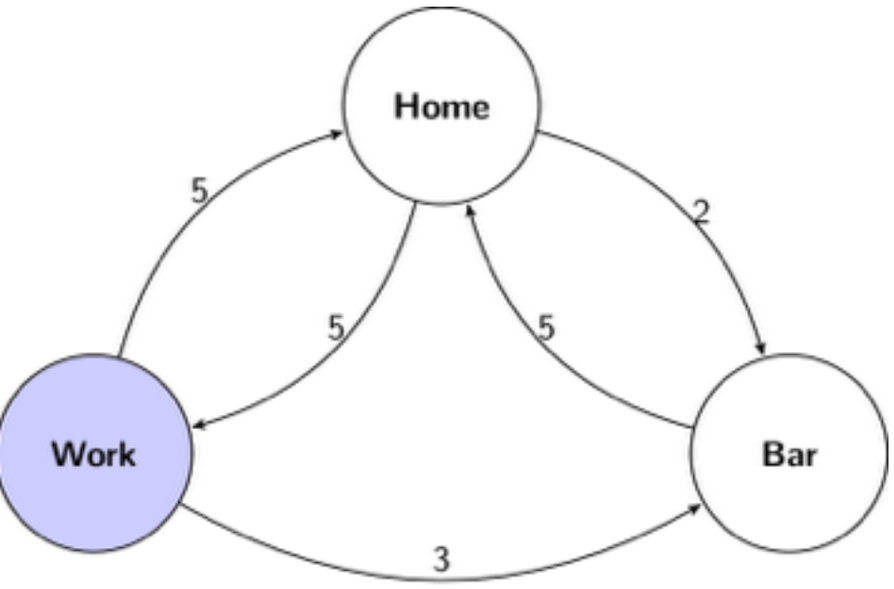
Markov process

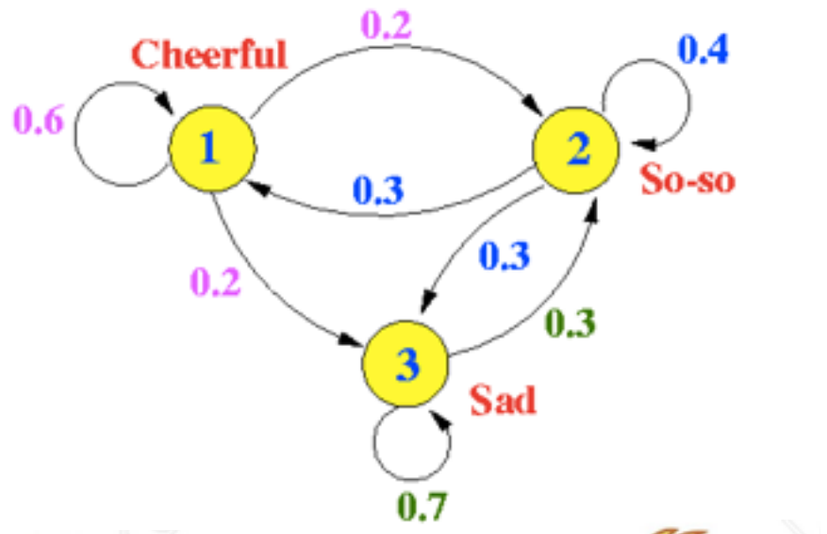




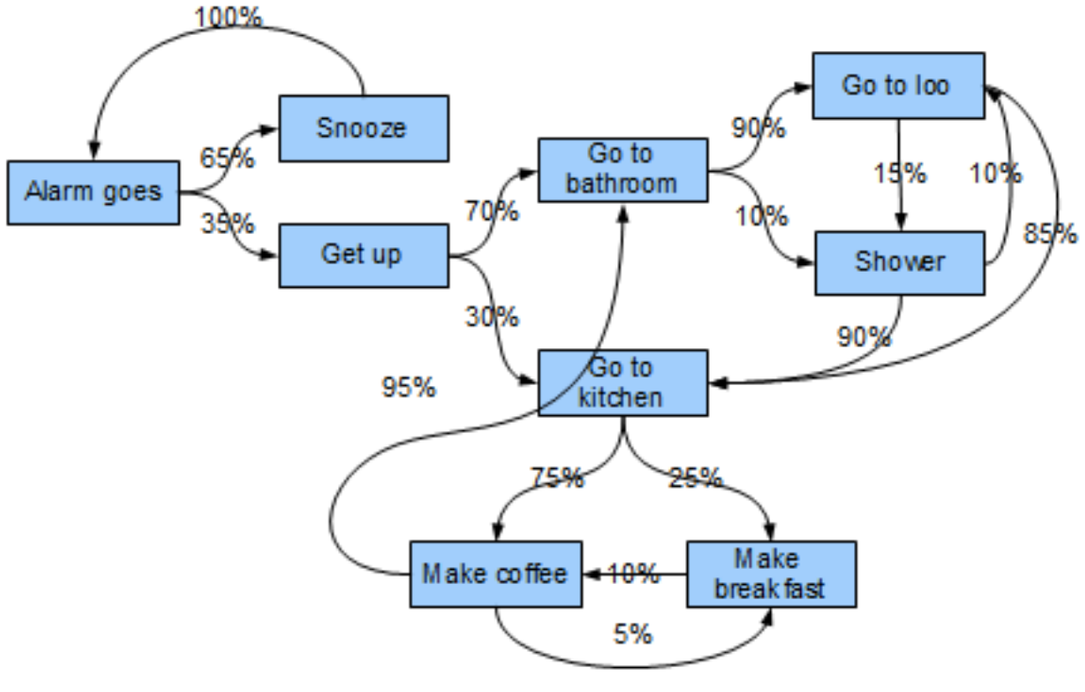
etc.

Markov process

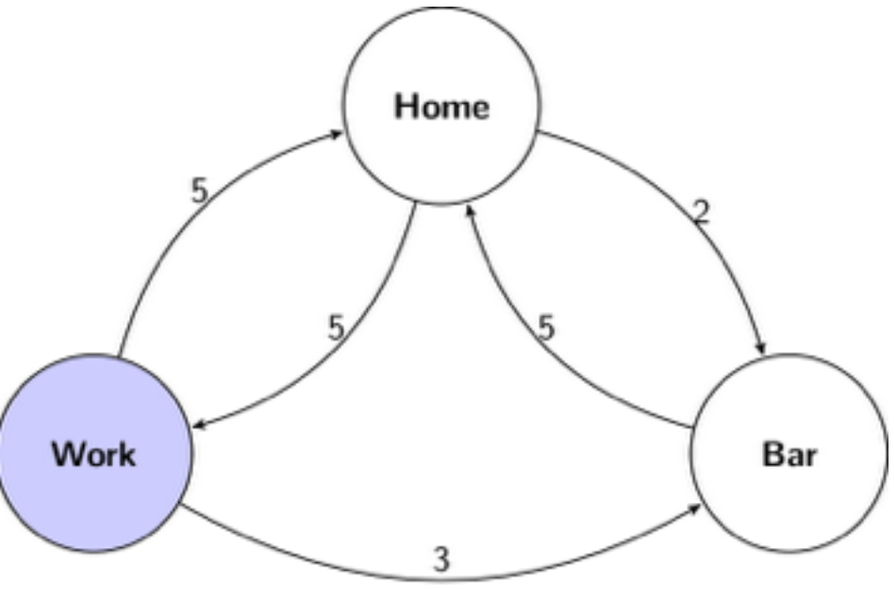


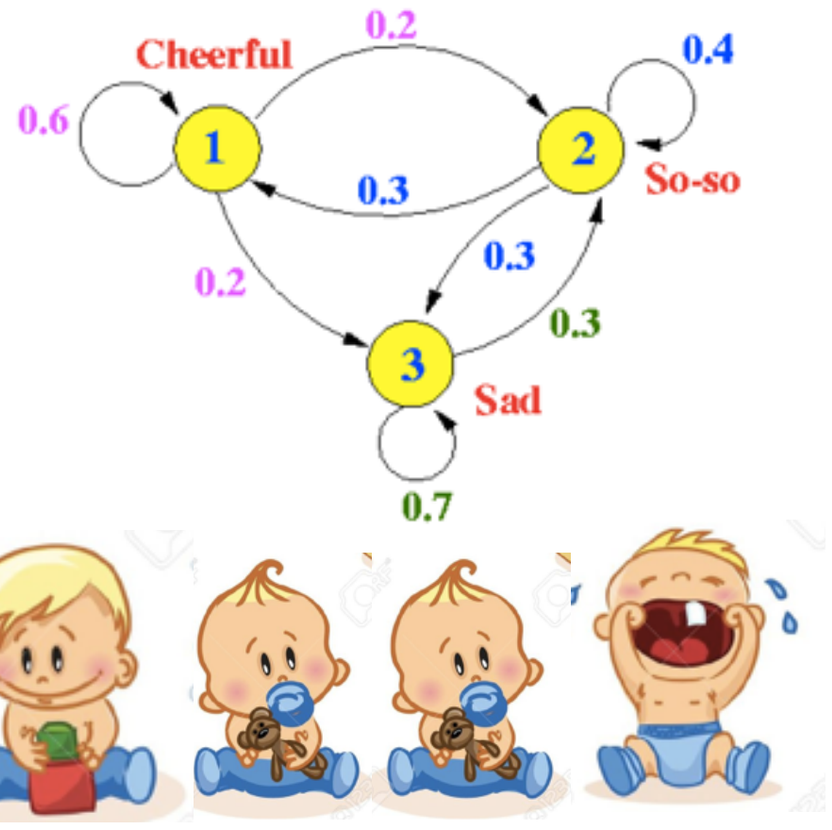


etc.

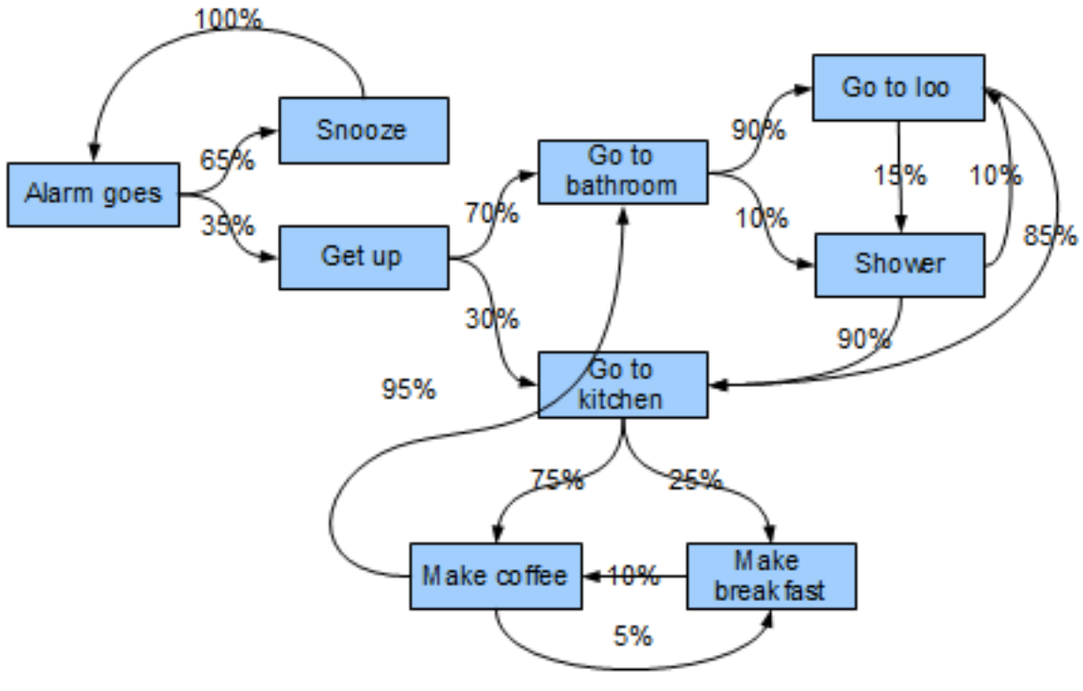


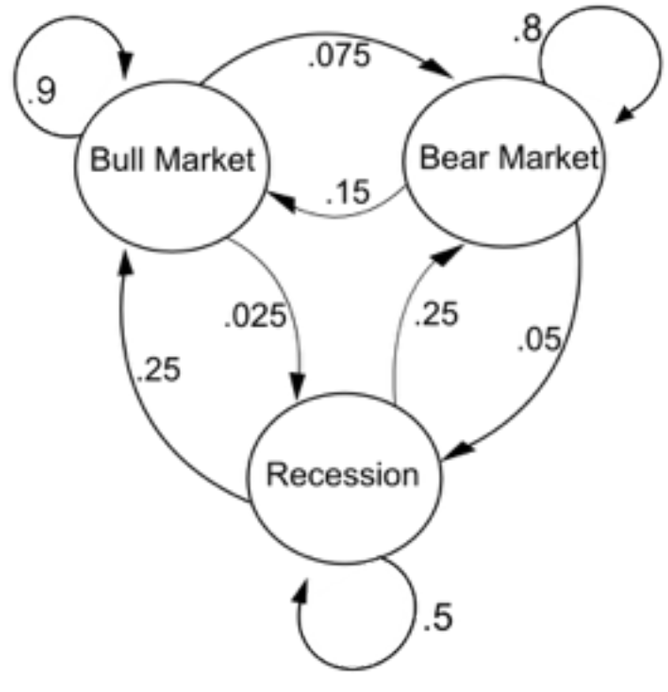
Markov process



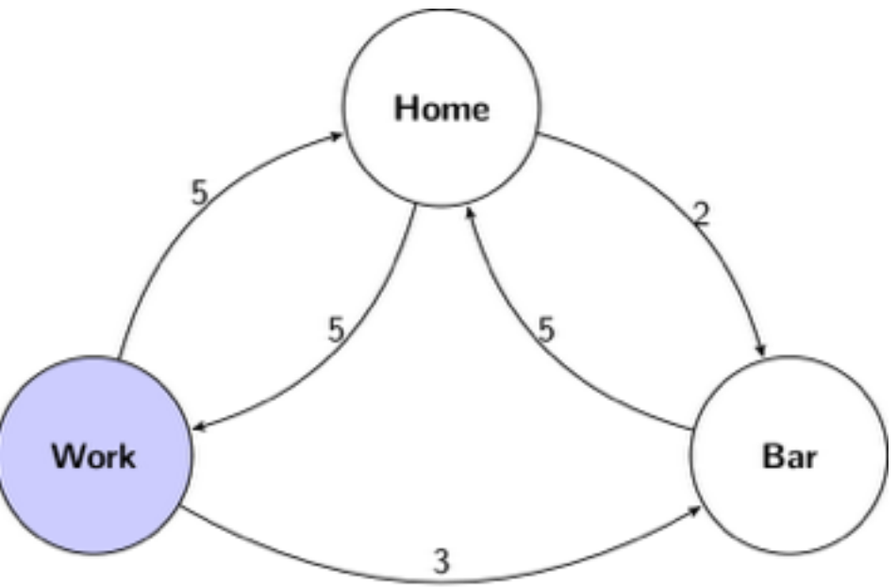


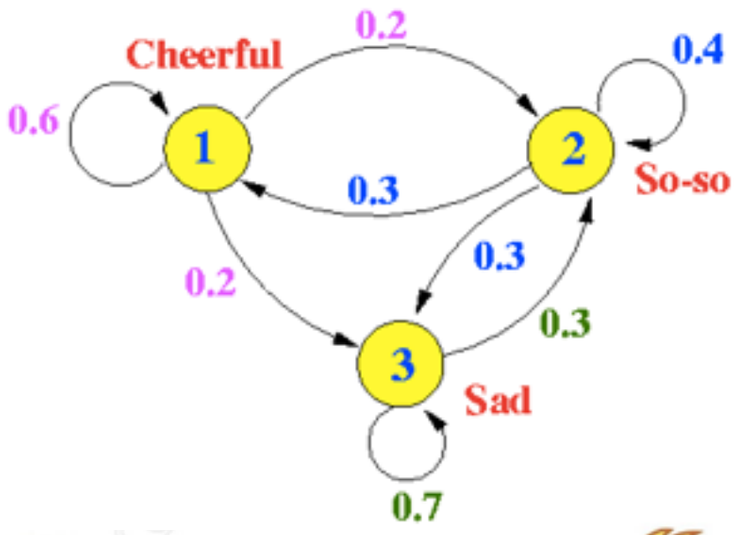
etc.

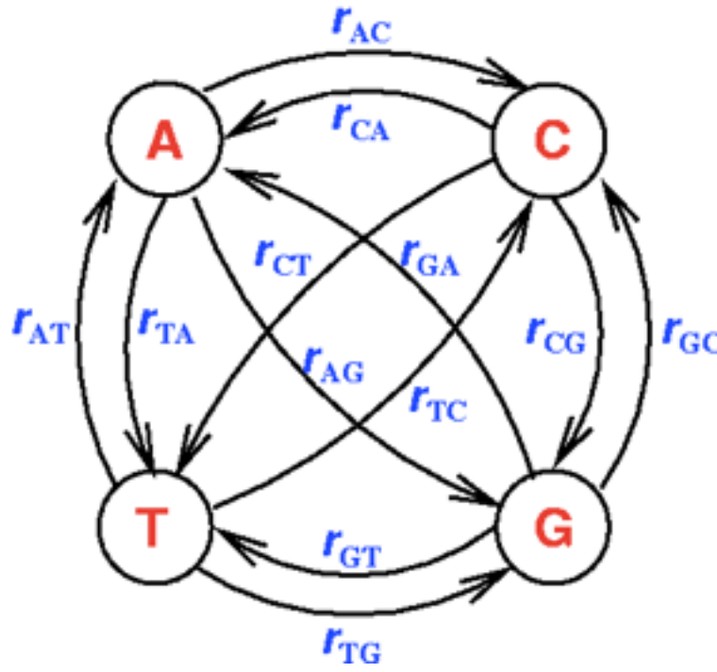




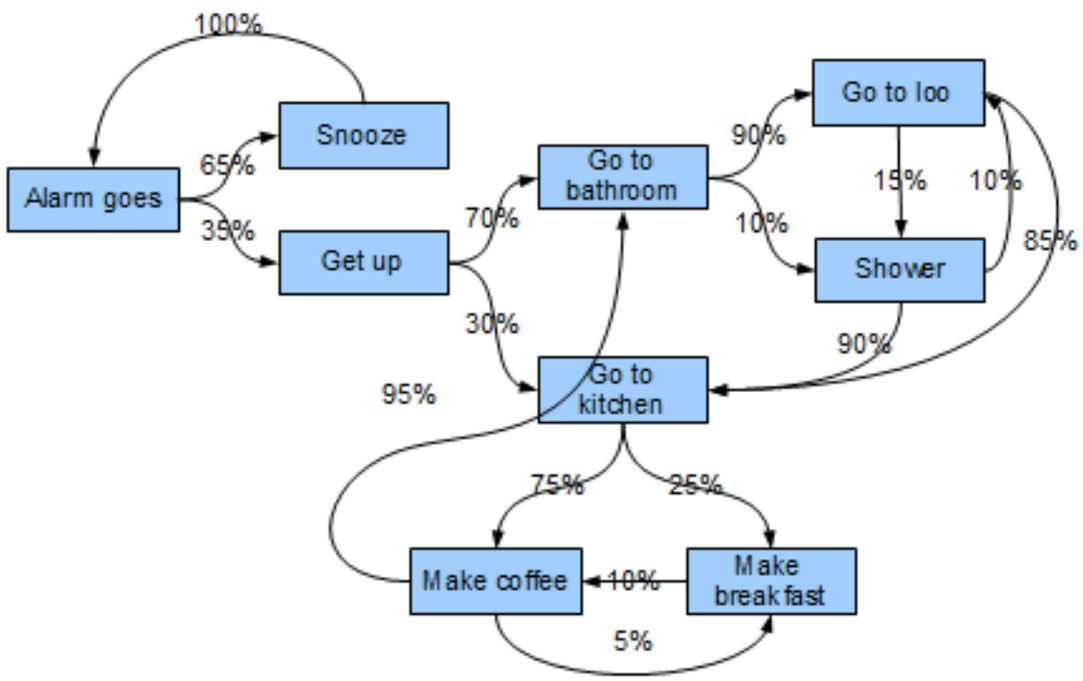
Markov process

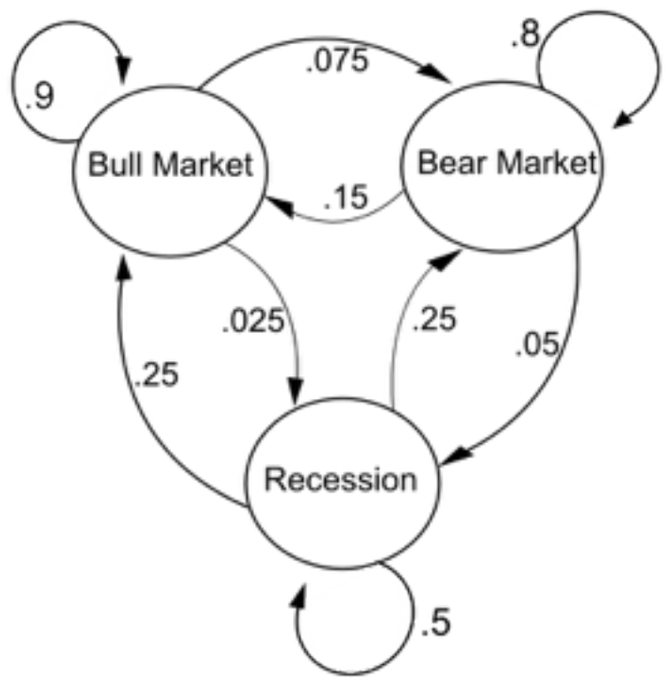




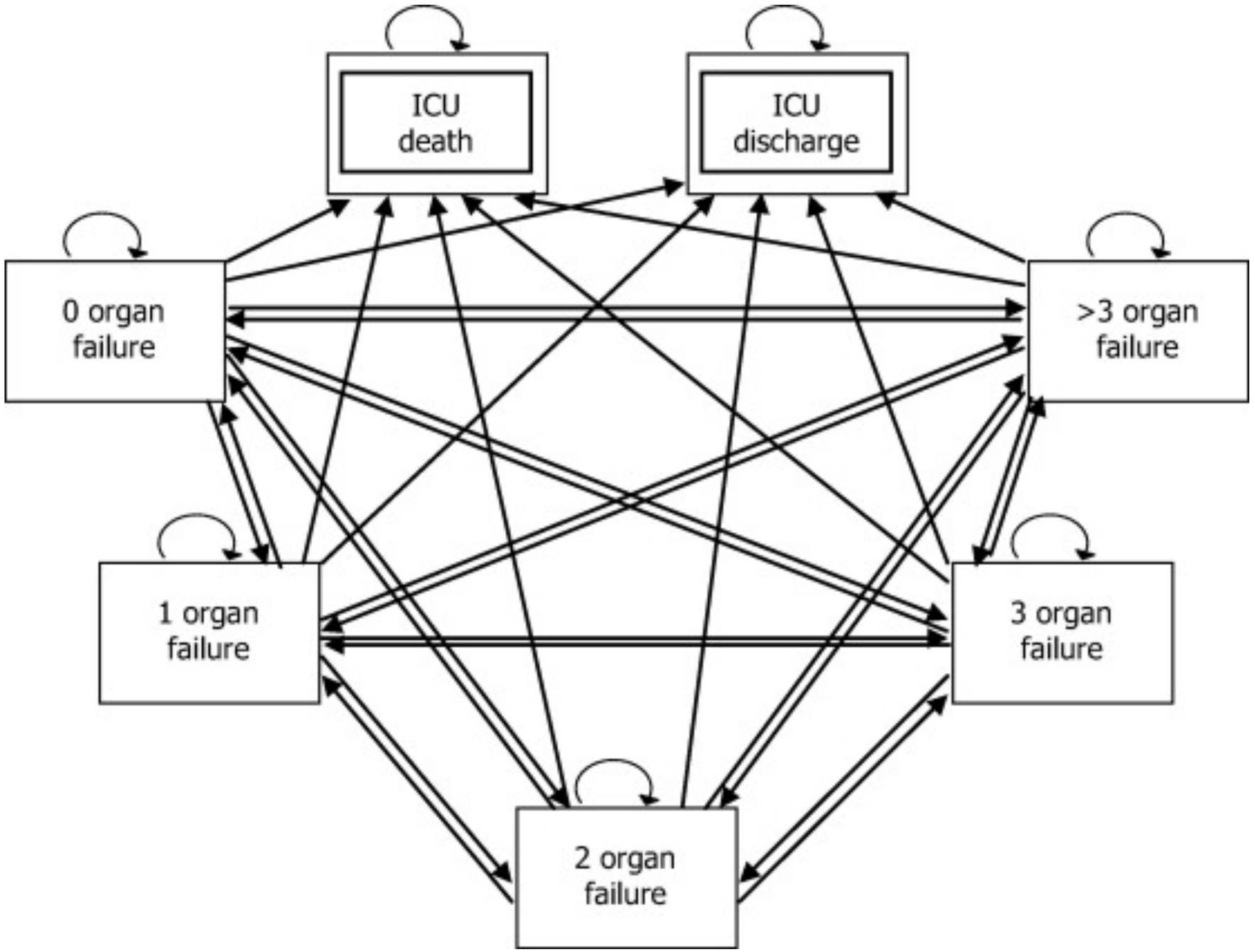


etc.

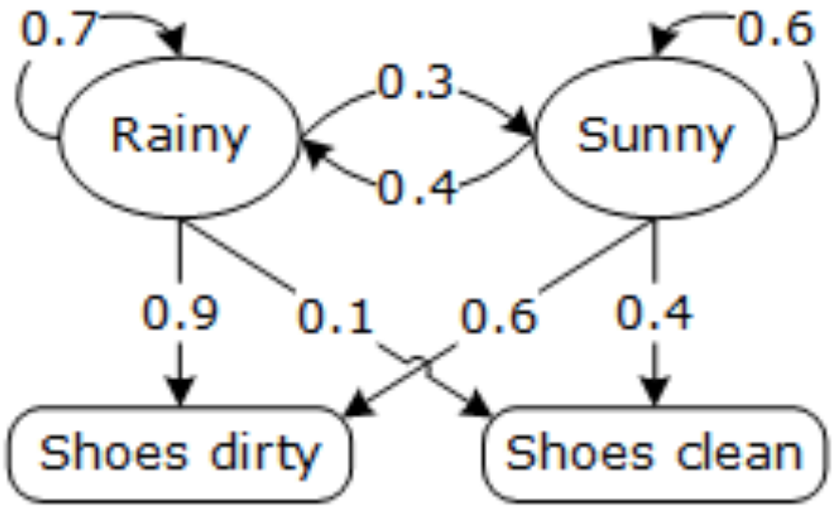




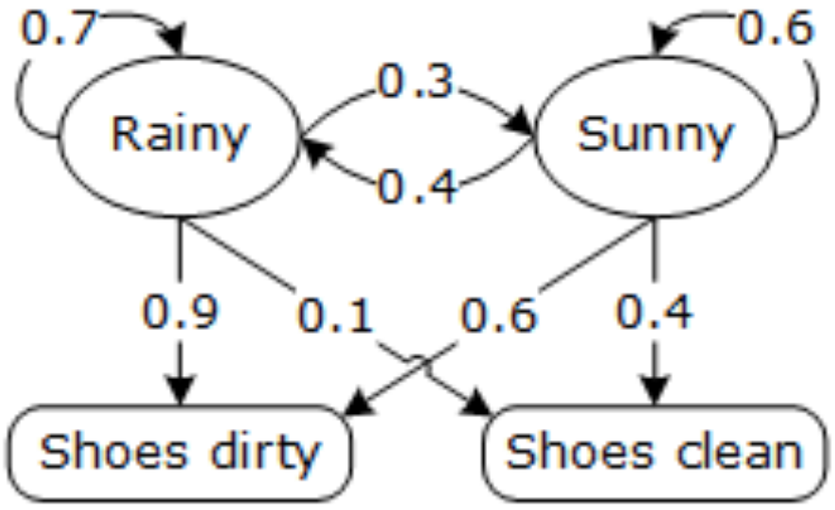
Markov process



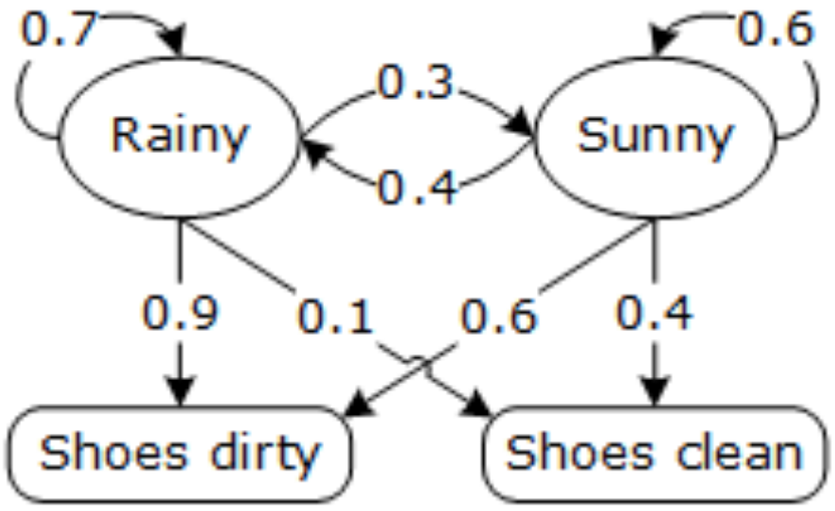
Hidden Markov process



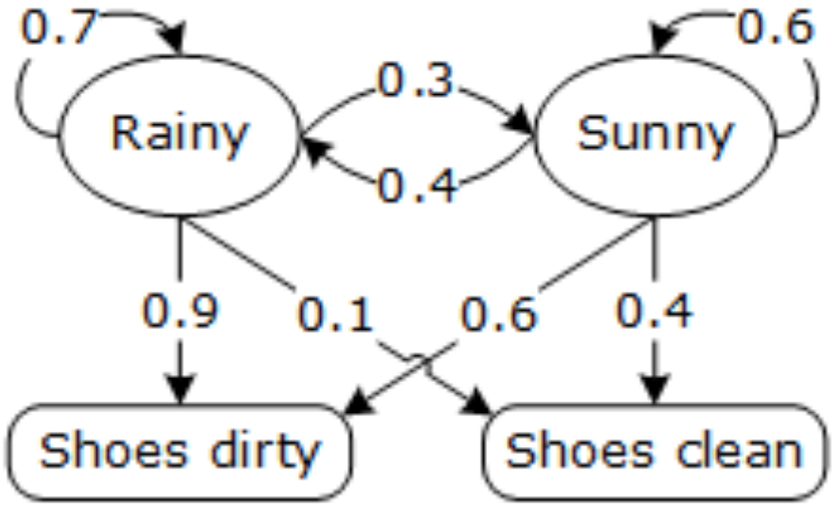
Hidden Markov process

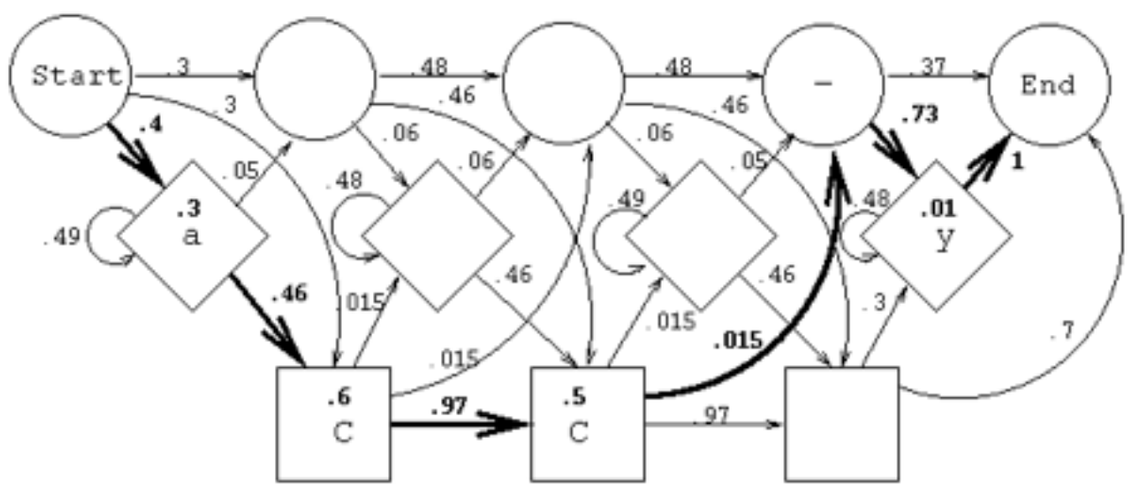


Hidden Markov process

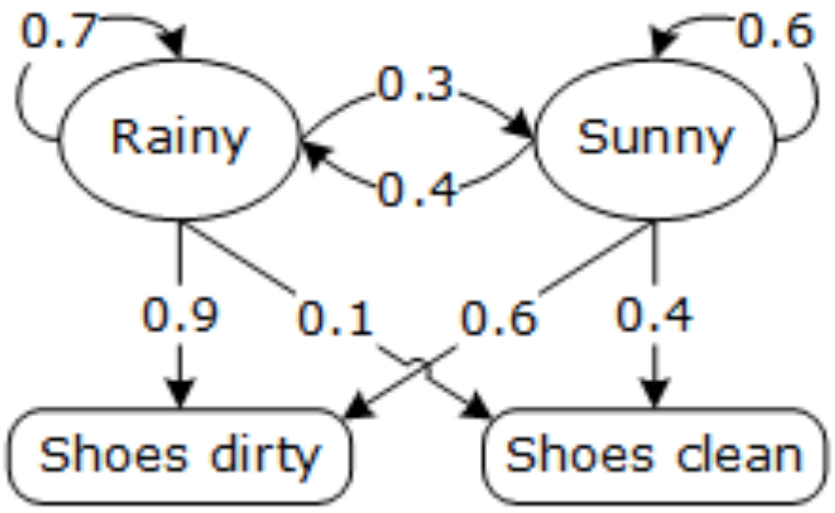


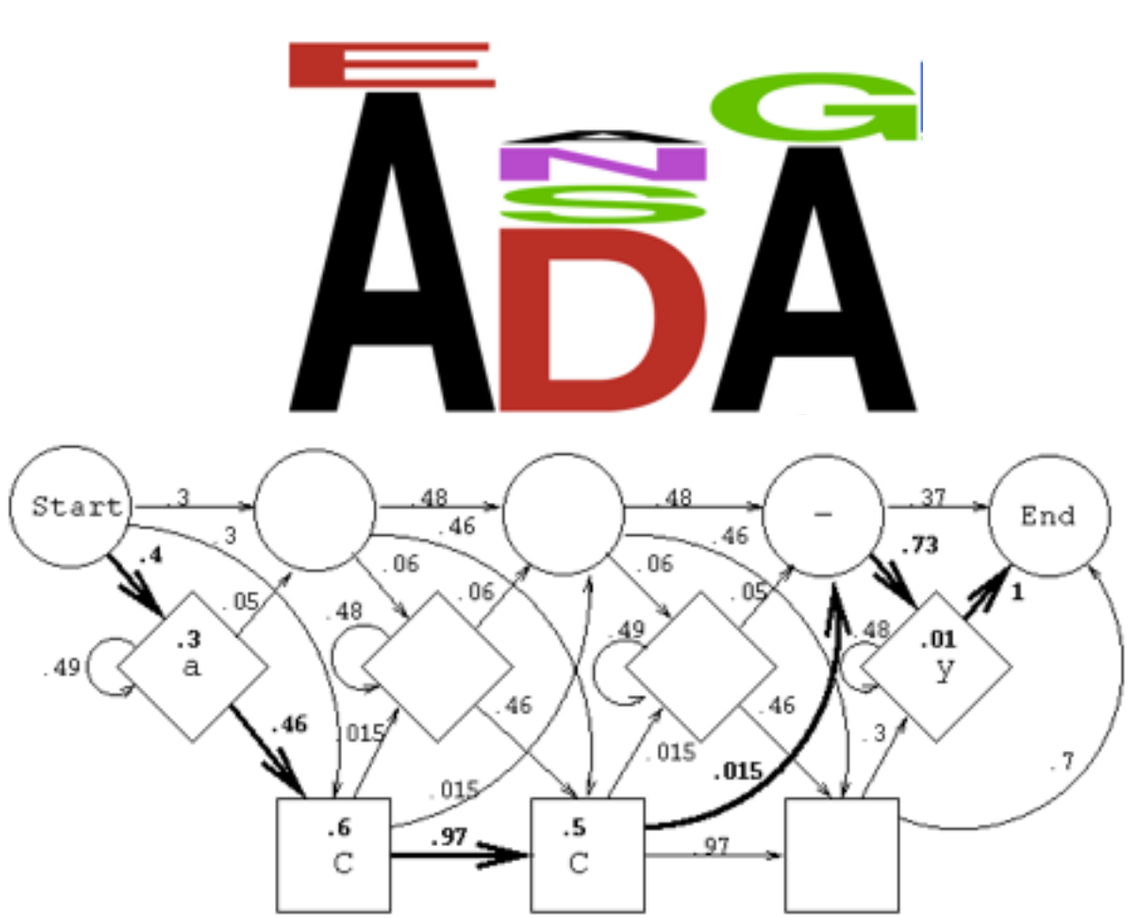
Hidden Markov process



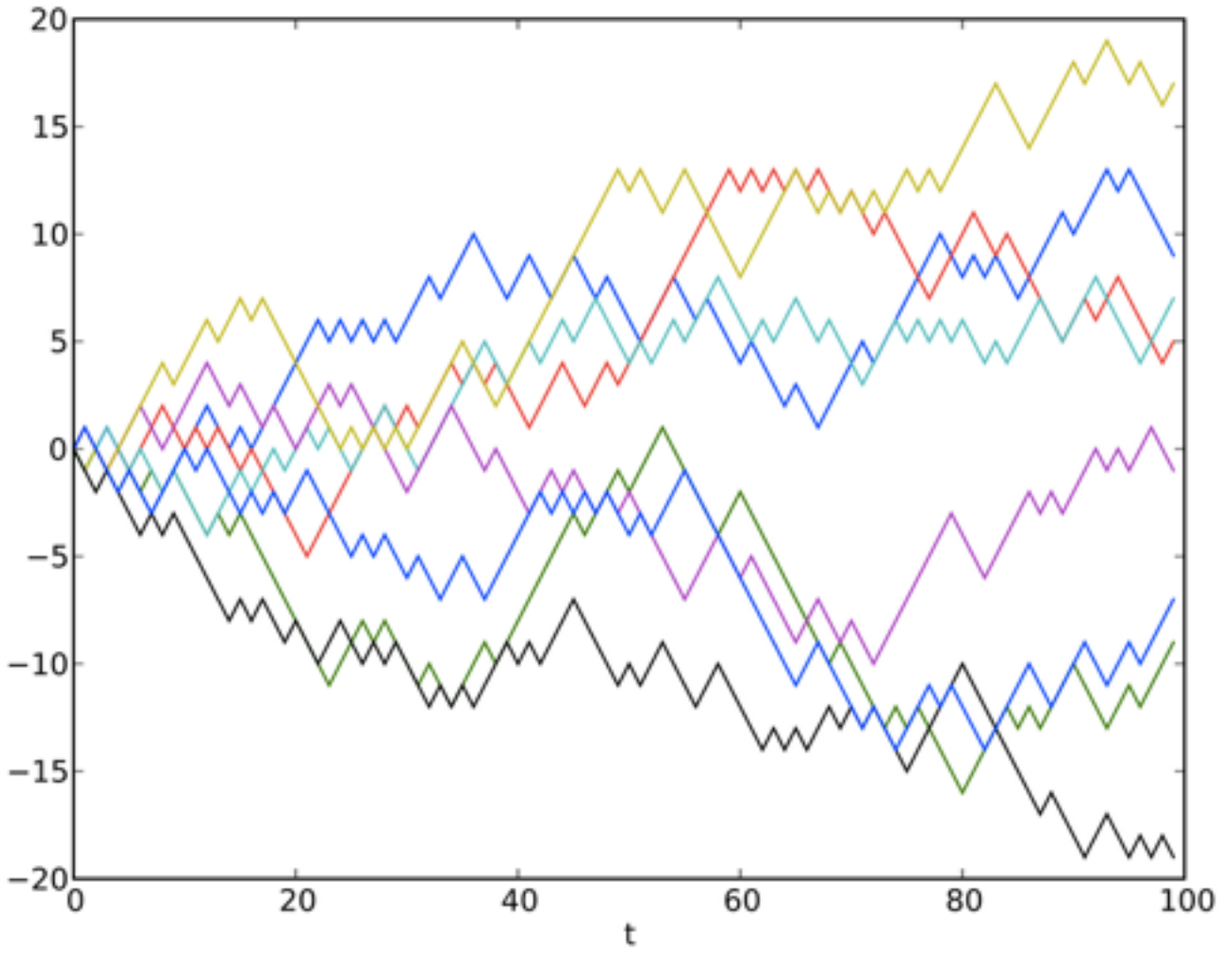


Hidden Markov process

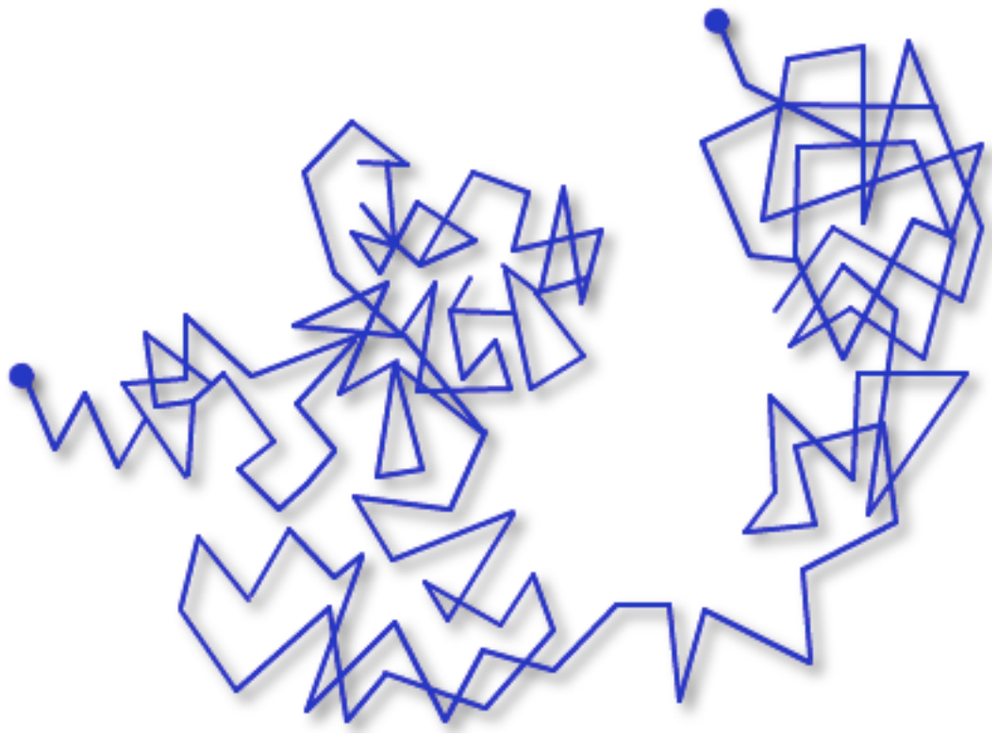




Random walk

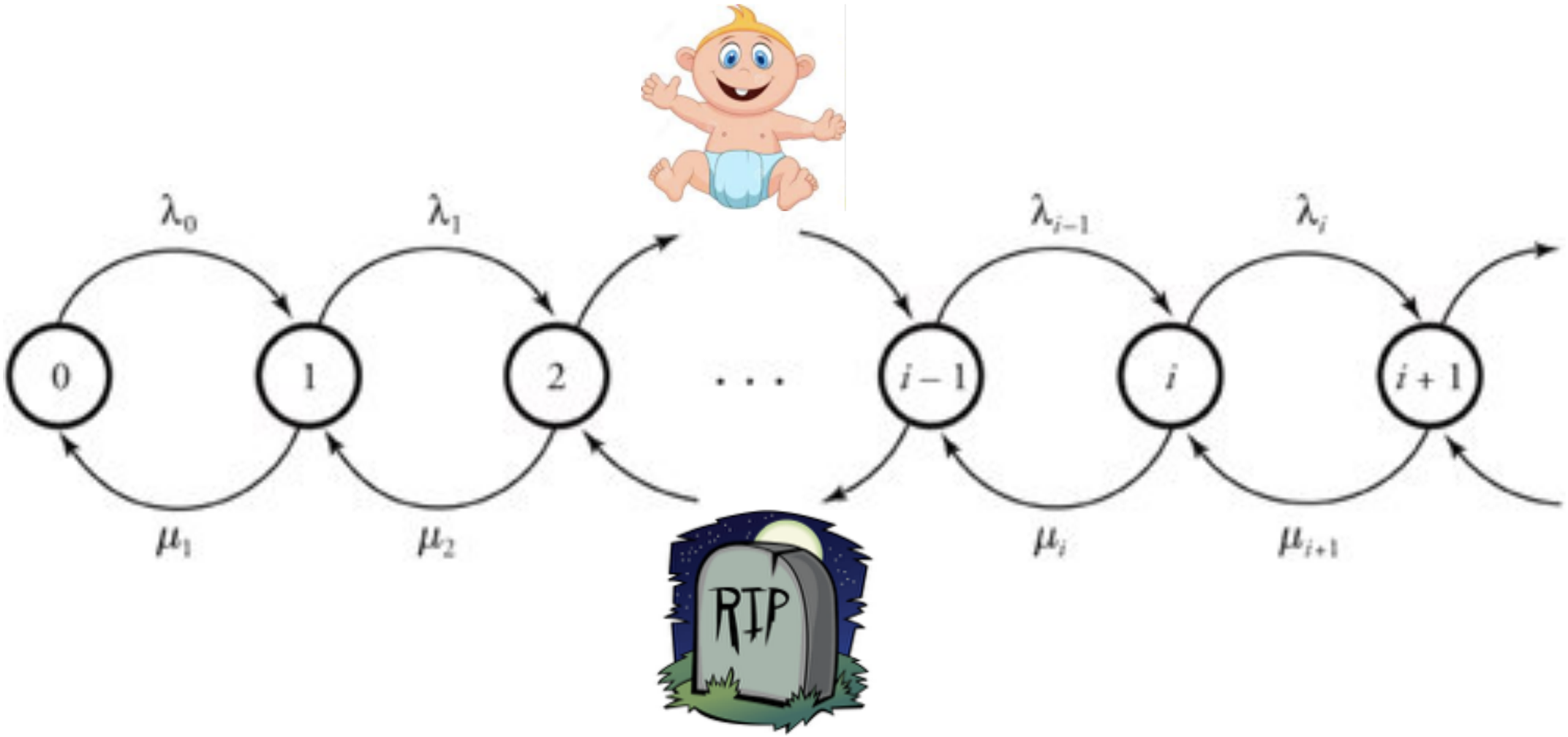


One-dimensional random walk trajectories



Brownian motion
(random impulses)

Birth-death process



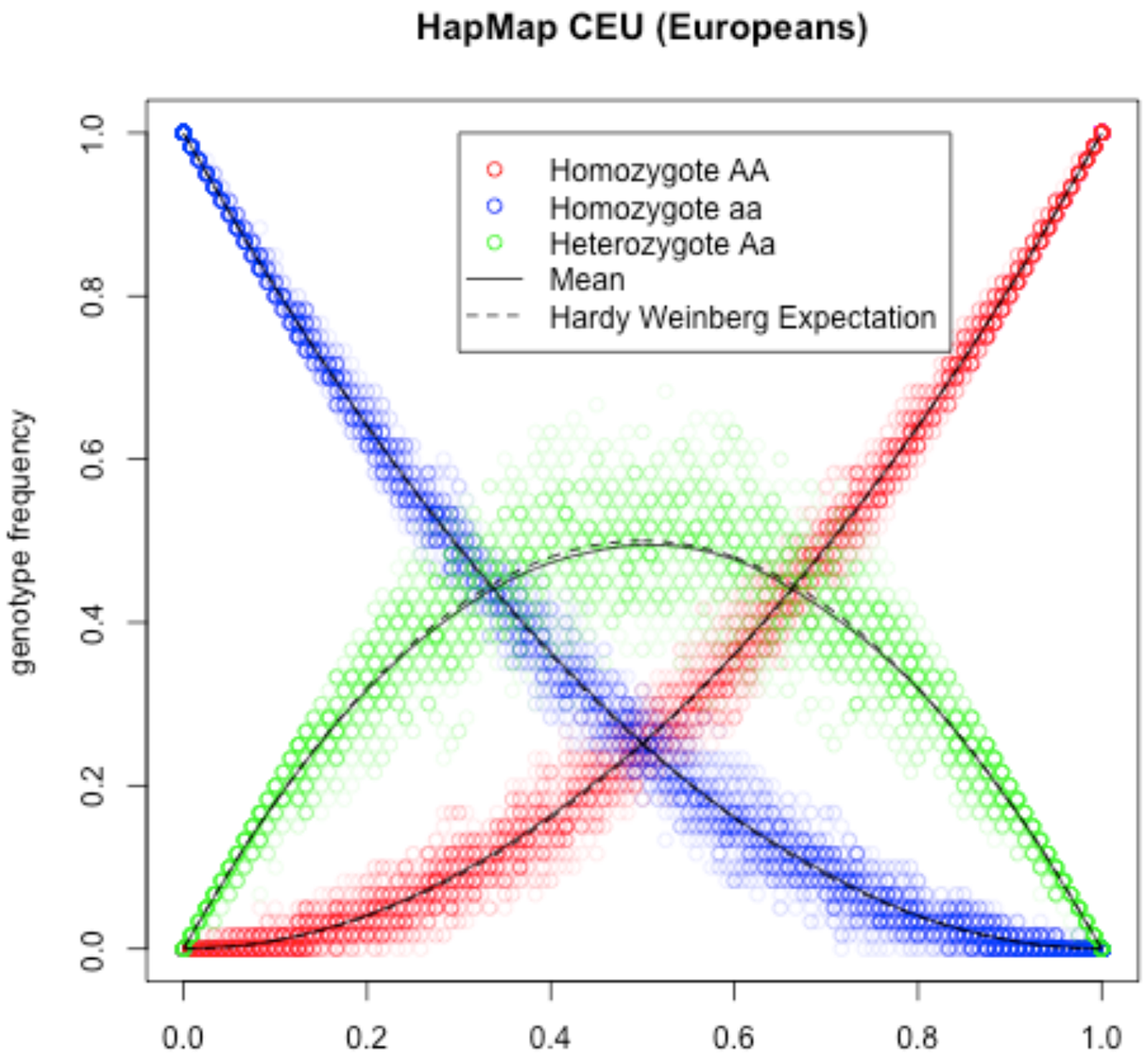
Hardy-Weinberg equilibrium

- Diploid genome (maternal + paternal chromosomes)
- Infinite population
- Random mating
- Biallelic locus (alleles: {A,a})

Hardy-Weinberg

- Allele frequencies
  - $P(A) = p, P(a) = q$
- Genotype frequencies after 1 generation
  - $P(AA)=p^2, P(Aa)=2pq, P(aa)=q^2$
- NB allele frequencies left unchanged:
  - $P(A) = P(AA) + P(Aa)/2 = p(p+q) = p$

Hardy-Weinberg & HapMap data



gcbias.org

Wright-Fisher model

- **Diploid genome**
- Finite population of constant size N
  - Effective population size 2N
- **Random mating**
- **Biallelic locus**
- Non-overlapping generations

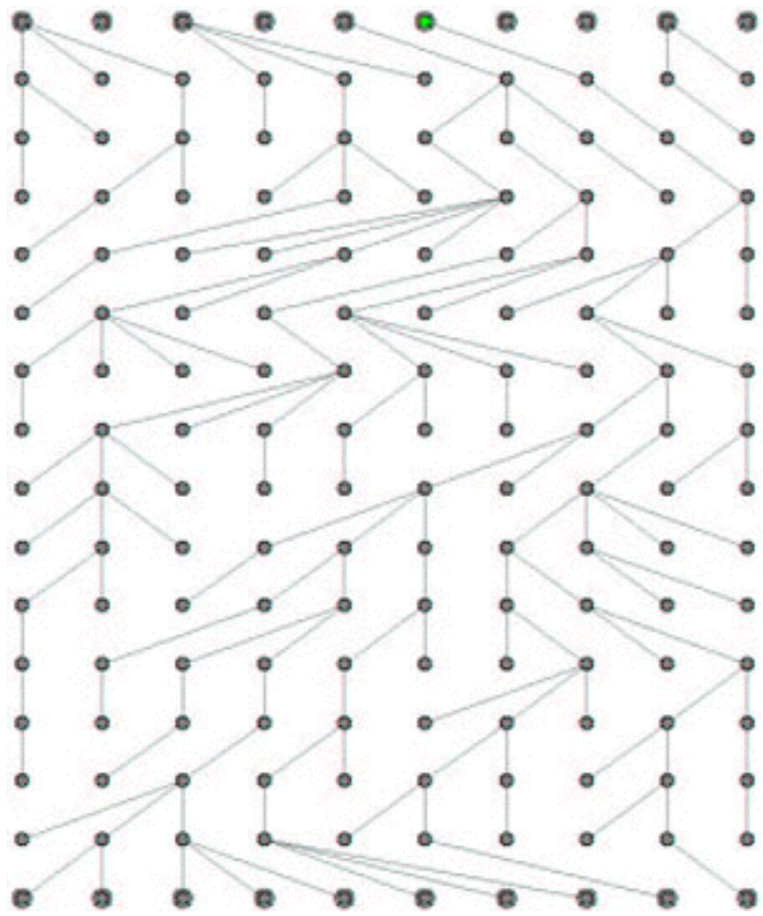
Bold assumptions are shared with Hardy-Weinberg

Wright-Fisher model

2 alleles, A_1 and A_2

X_t = number of A_2 alleles at time **t**

$p_{ij} = \Pr[X_{t+1} = j | X_t = i]$



Important framework for thinking about genealogy and populations

Wright-Fisher model

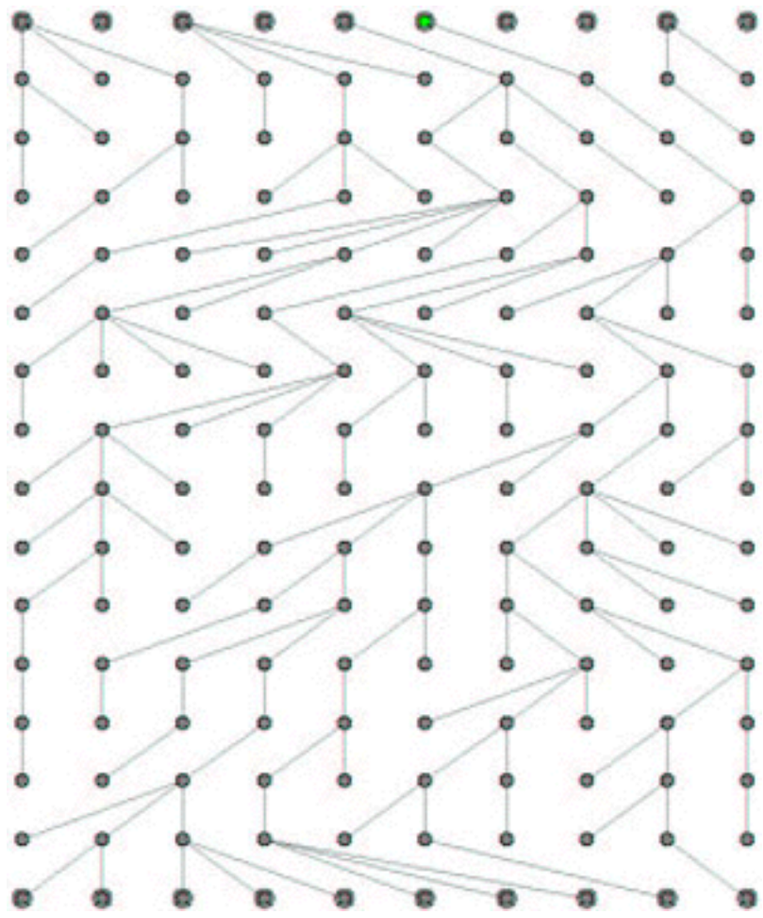
2 alleles, A_1 and A_2

X_t = number of A_2 alleles at time **t**

$p_{ij} = \Pr\left[X_{t+1} = j \middle| X_t = i\right]$

$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$

Binomial distribution



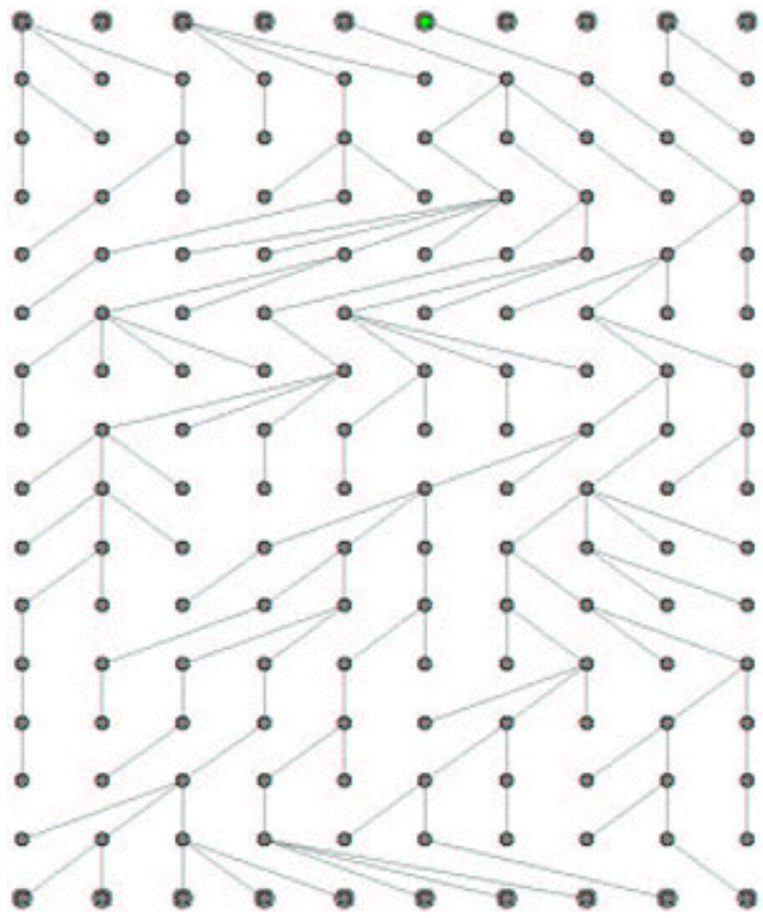
Important framework for thinking about genealogy and populations

Wright-Fisher model

2 alleles, A_1 and A_2

X_t = number of A_2 alleles at time **t**

$p_{ij} = \Pr\left[X_{t+1} = j \middle| X_t = i\right]$



$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}$$

Binomial distribution

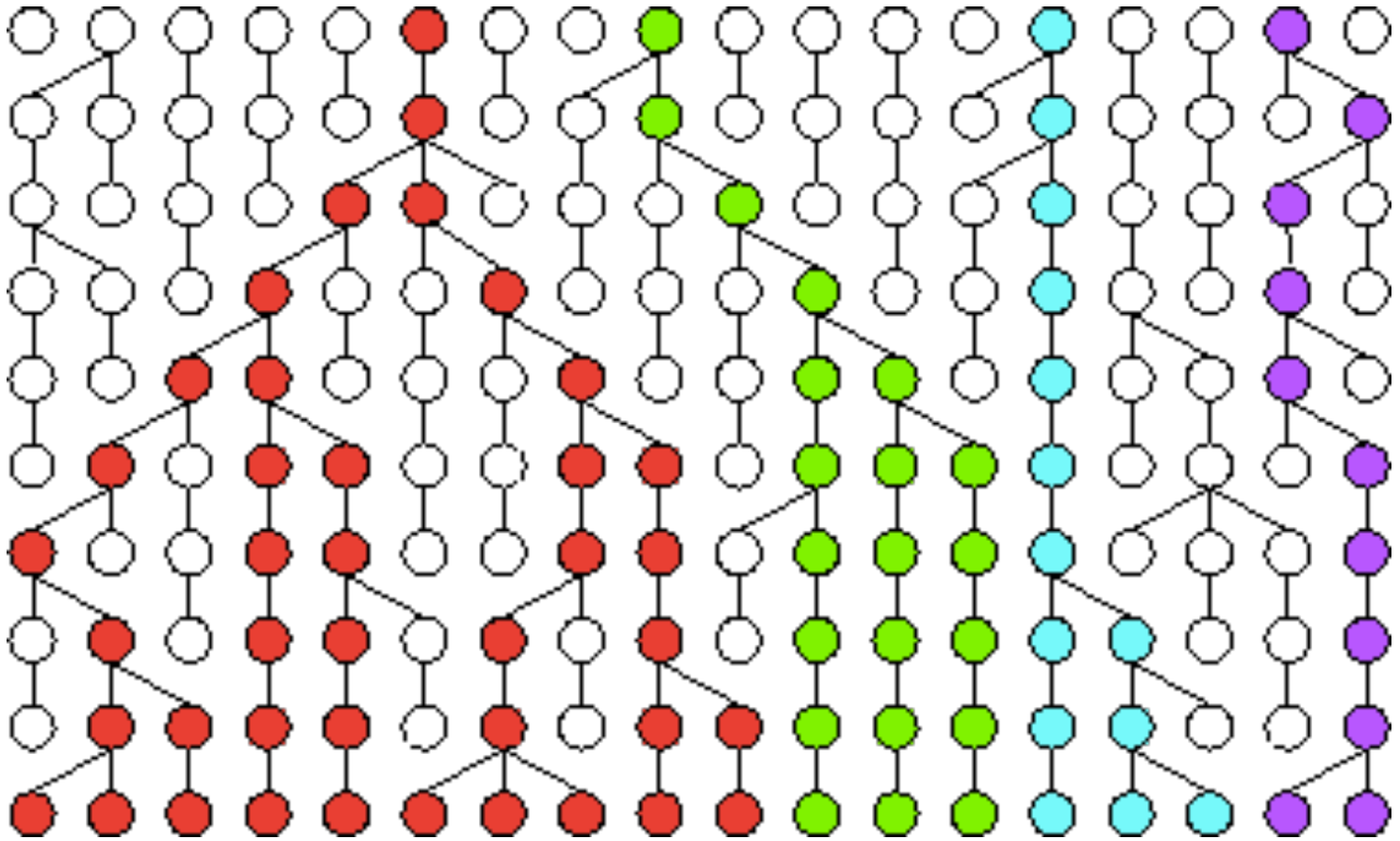
$$Pr\left[X_t \text{ absorbs at } \begin{cases} 0 \\ 2N \end{cases}\right] = \begin{cases} 1 - \frac{X_0}{2N} \\ \frac{X_0}{2N} \end{cases}$$

Fixation probability

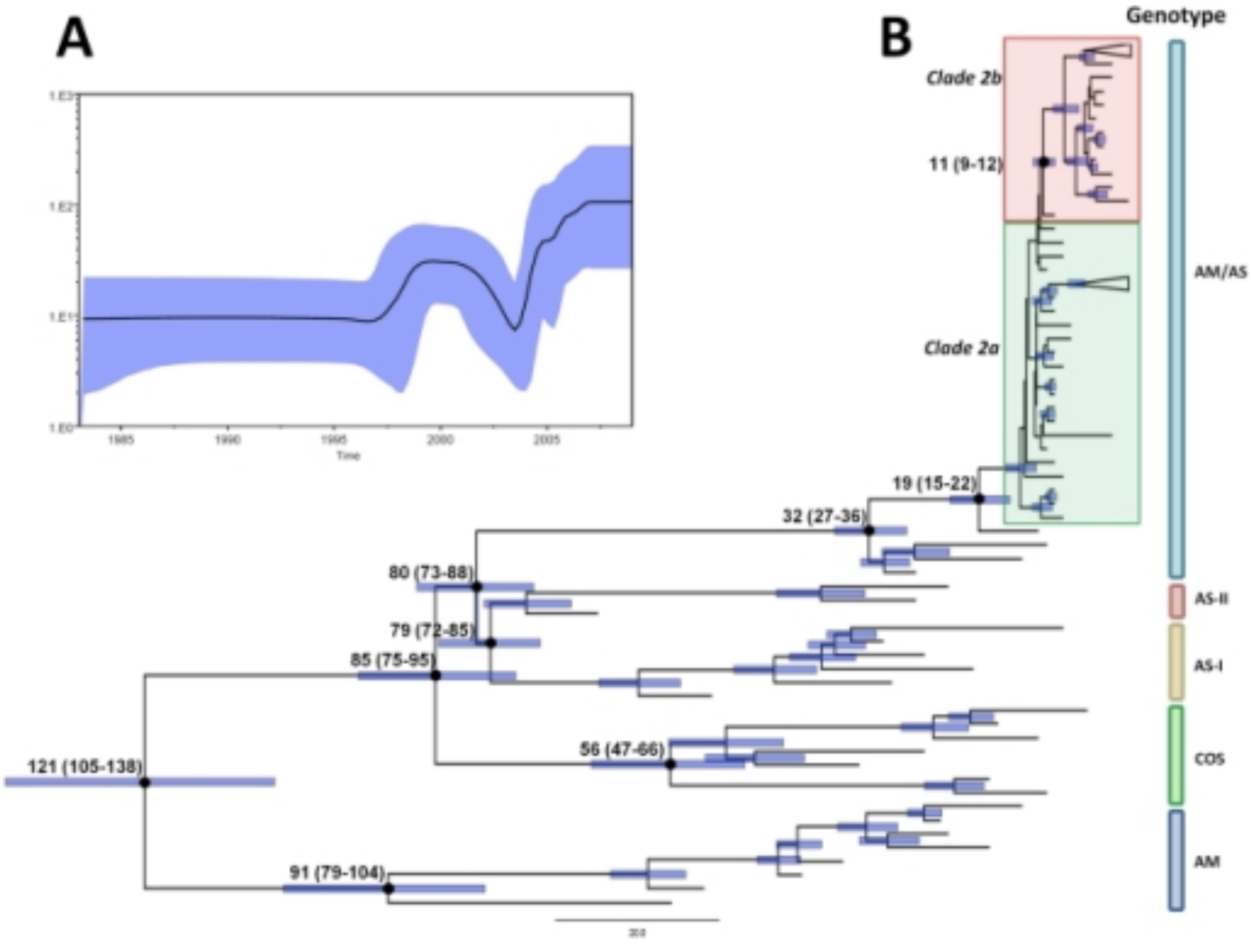
(via Doob's stopping theorem)

Important framework for thinking about genealogy and populations

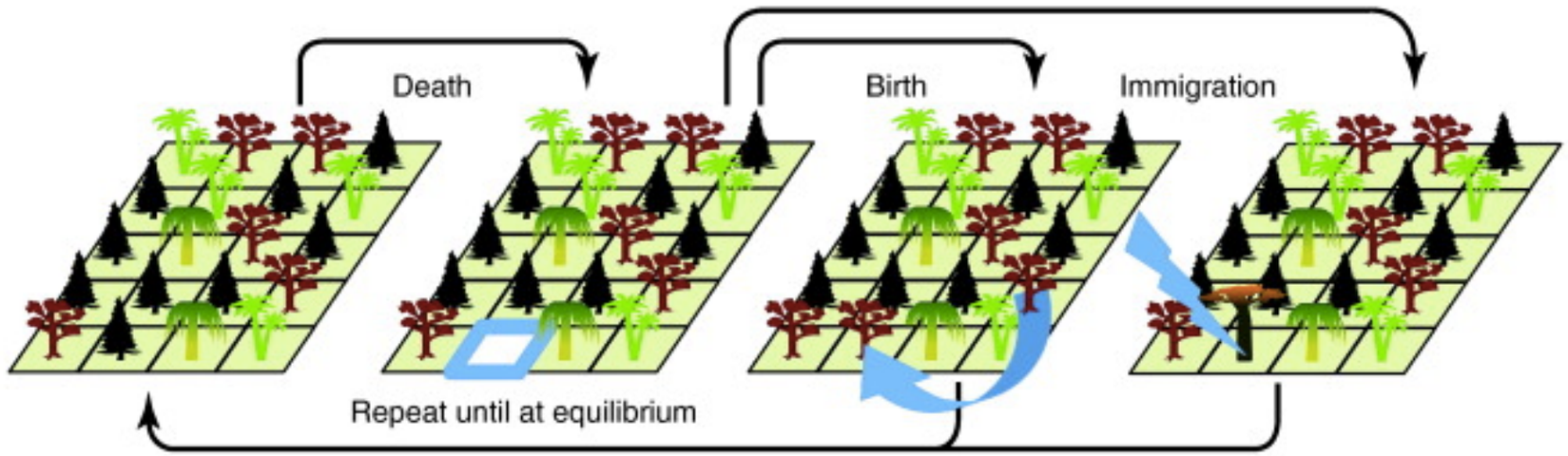
Coalescent process



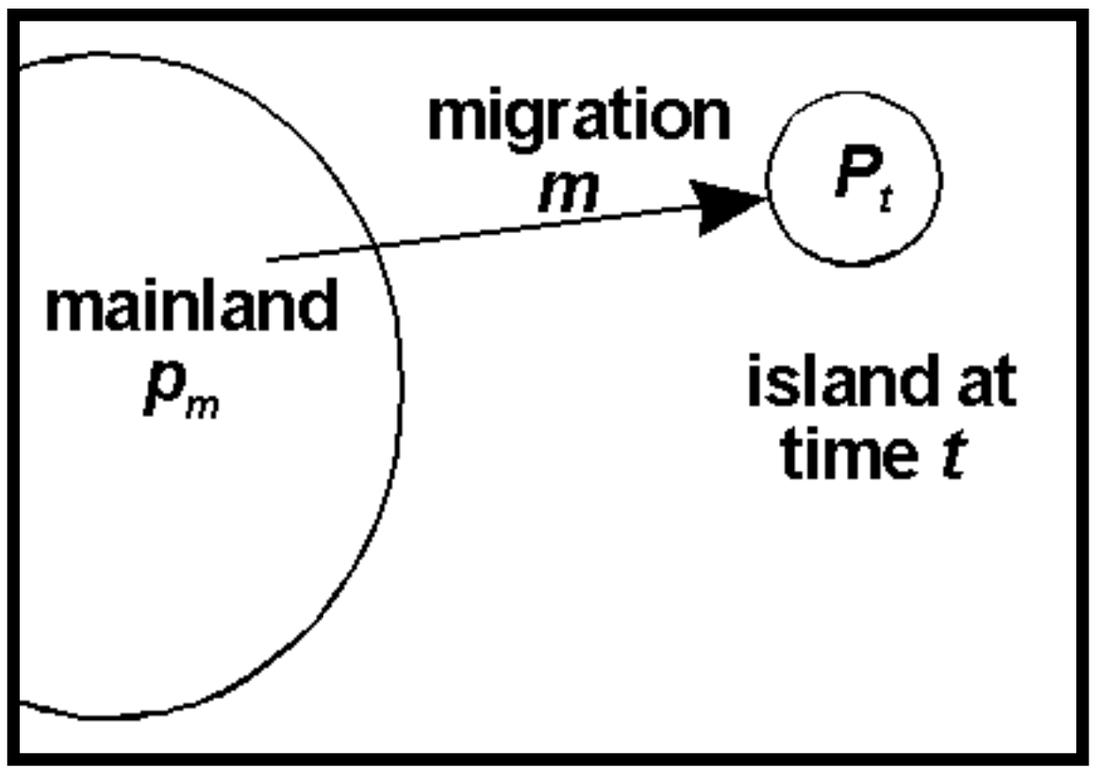
Skyline Plots



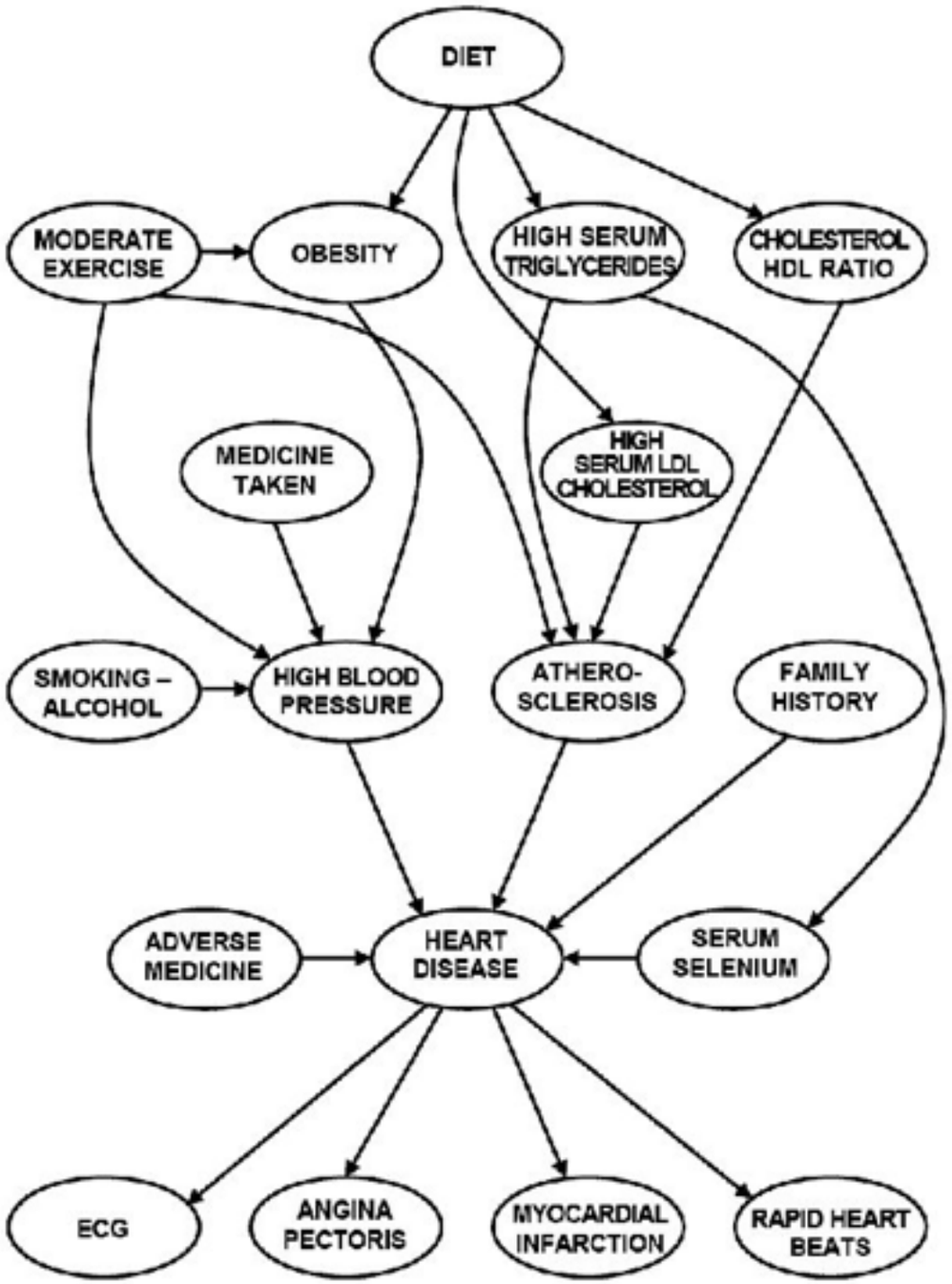
Unified Neutral Theory of Biodiversity

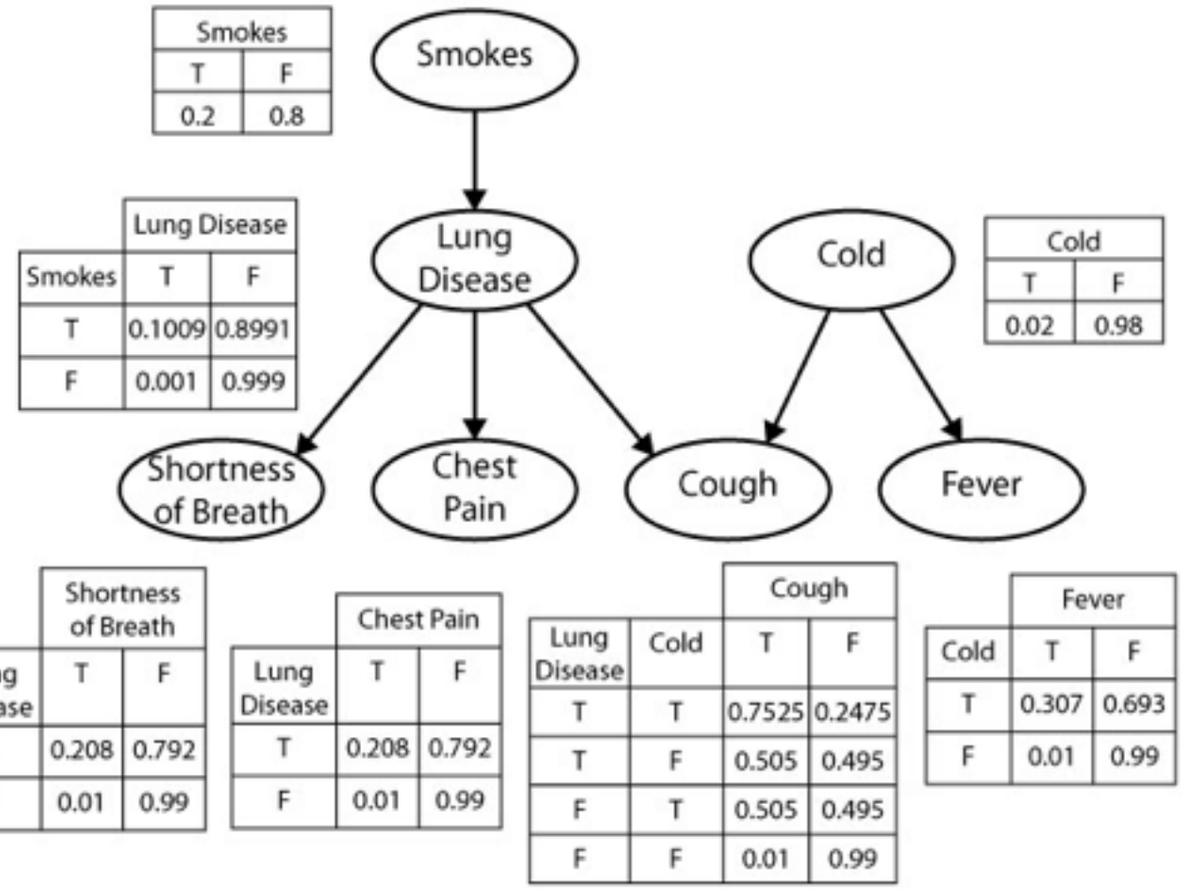


TRENDS in Ecology & Evolution



Bayesian networks





Smokes	
T	F
0.2	0.8

Smokes	Lung Disease T	Lung Disease F
T	0.1009	0.8991
F	0.001	0.999

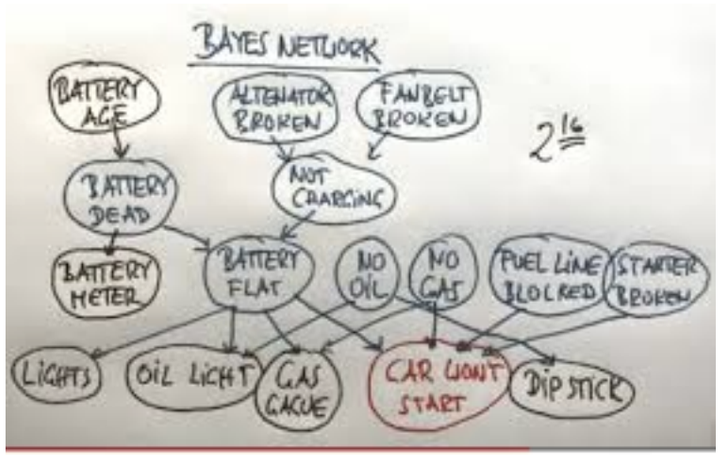
Cold	
T	F
0.02	0.98

Lung Disease	Shortness of Breath T	Shortness of Breath F
T	0.208	0.792
F	0.01	0.99

Lung Disease	Chest Pain T	Chest Pain F
T	0.208	0.792
F	0.01	0.99

Lung Disease	Cold	Cough T	Cough F
T	T	0.7525	0.2475
T	F	0.505	0.495
F	T	0.505	0.495
F	F	0.01	0.99

Cold	Fever T	Fever F
T	0.307	0.693
F	0.01	0.99



Probability summary

- Discrete & continuous random variables, normalization
- Binomial, Gaussian, Extreme Value distributions
- Joint, conditional, marginal, cumulative probability
- Uniform distribution, IID sequence
- Expectation & variance
- Bayes' Theorem
- Random processes in biology