CS 364 COMPUTATIONAL BIOLOGY

Sara Mathieson
Haverford College

Outline

Lab 7 posted
- Due next Tues
- Shorter coding
- Includes project proposal

- Deep learning in genetics
- Coalescent Theory
- Putting it all together: Tajima's D for natural selection

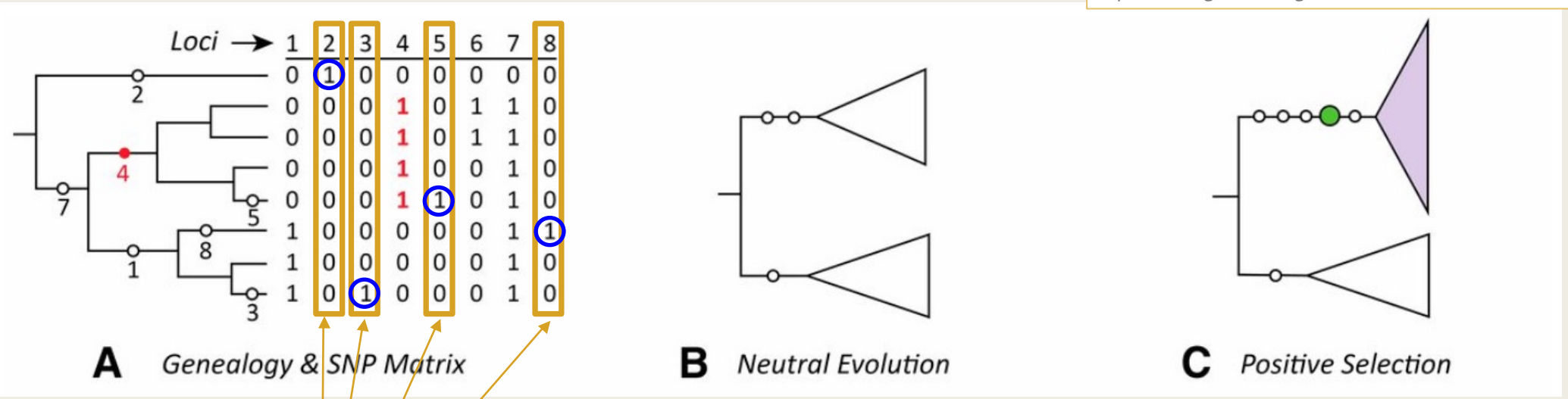
Deep learning in population genetics

2013: Using machine learning to infer selection

Learning Natural Selection from the Site Frequency Spectrum

Roy Ronen, Nitin Udpa, Eran Halperin and Vineet Bafna

GENETICS *September 1, 2013 vol. 195 no. 1 181-193; https://doi.org/10.1534/genetics.113.152587*



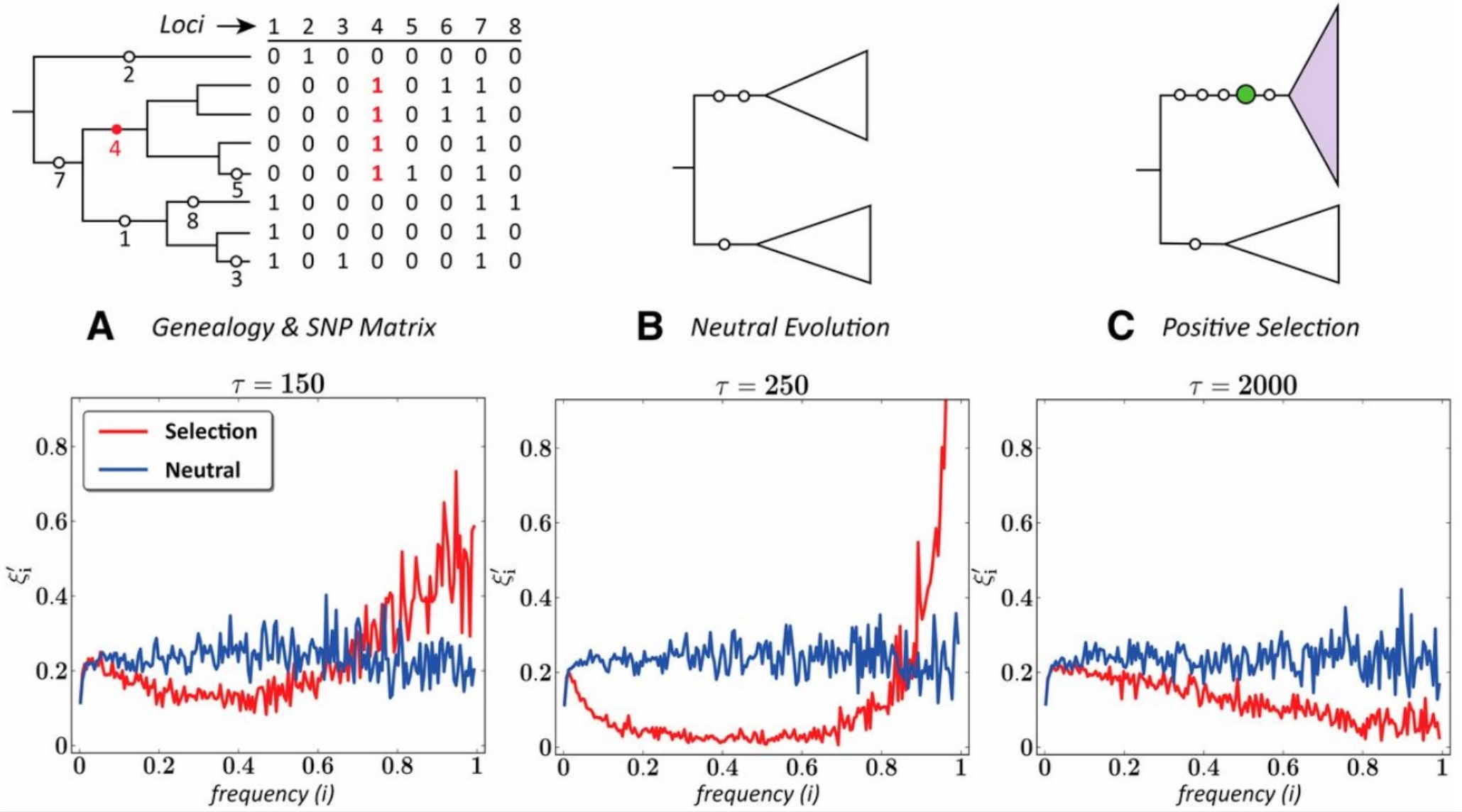
SFS = [4, 1, 1, 1, 0, 0, 1]

2013: Using machine learning to infer selection

Learning Natural Selection from the Site Frequency Spectrum

Roy Ronen, Nitin Udpa, Eran Halperin and Vineet Bafna

GENETICS *September 1, 2013 vol. 195 no. 1 181-193; https://doi.org/10.1534/genetics.113.152587*



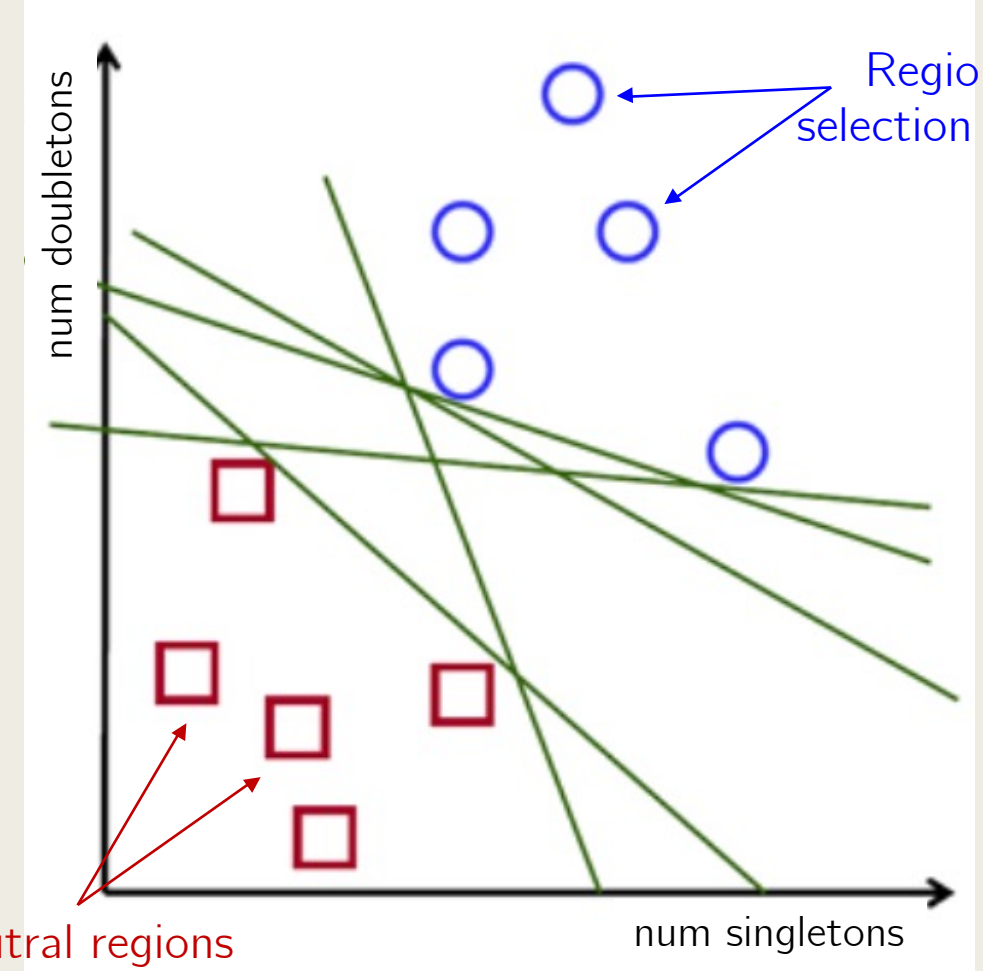
2013: Using machine learning to infer selection

Learning Natural Selection from the Site Frequency Spectrum

Roy Ronen, Nitin Udpa, Eran Halperin and Vineet Bafna

GENETICS *September 1, 2013 vol. 195 no. 1 181-193;*
https://doi.org/10.1534/genetics.113.152587

Method: support vector machine (SVM)



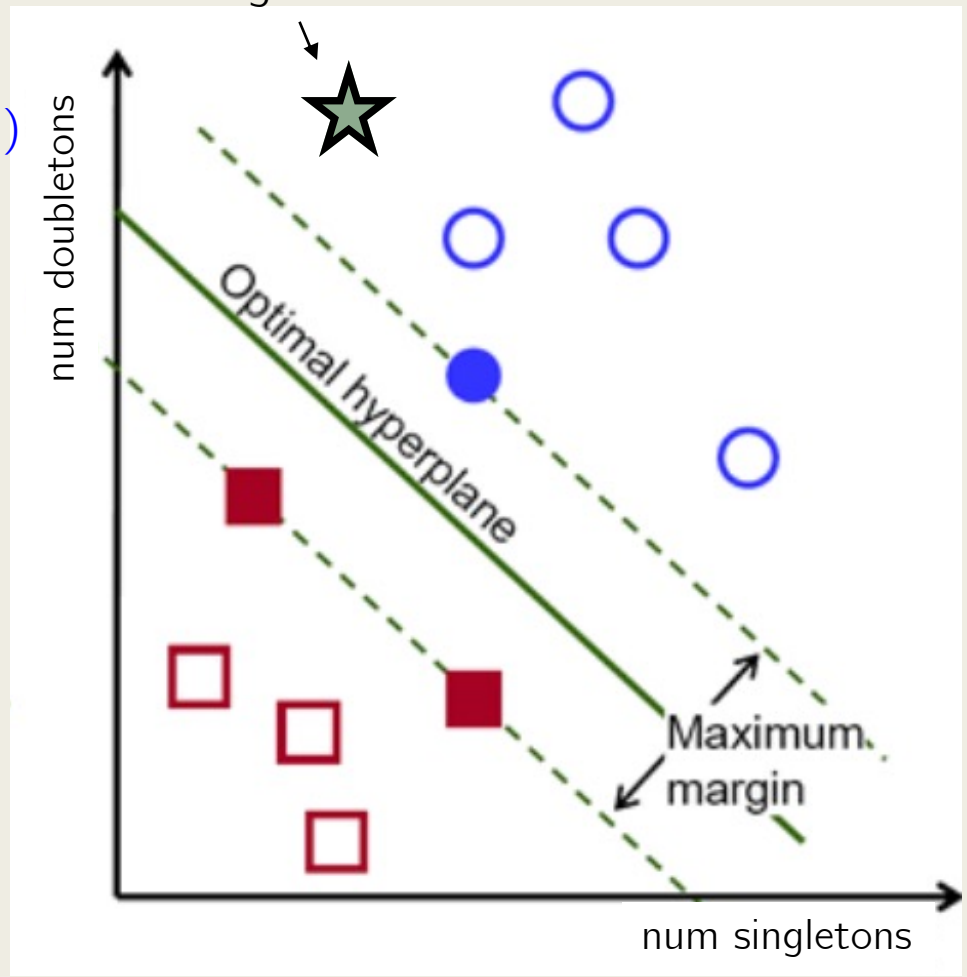


Image from: "Towards Data Science"

Multi-layer (deep) neural network will distill information

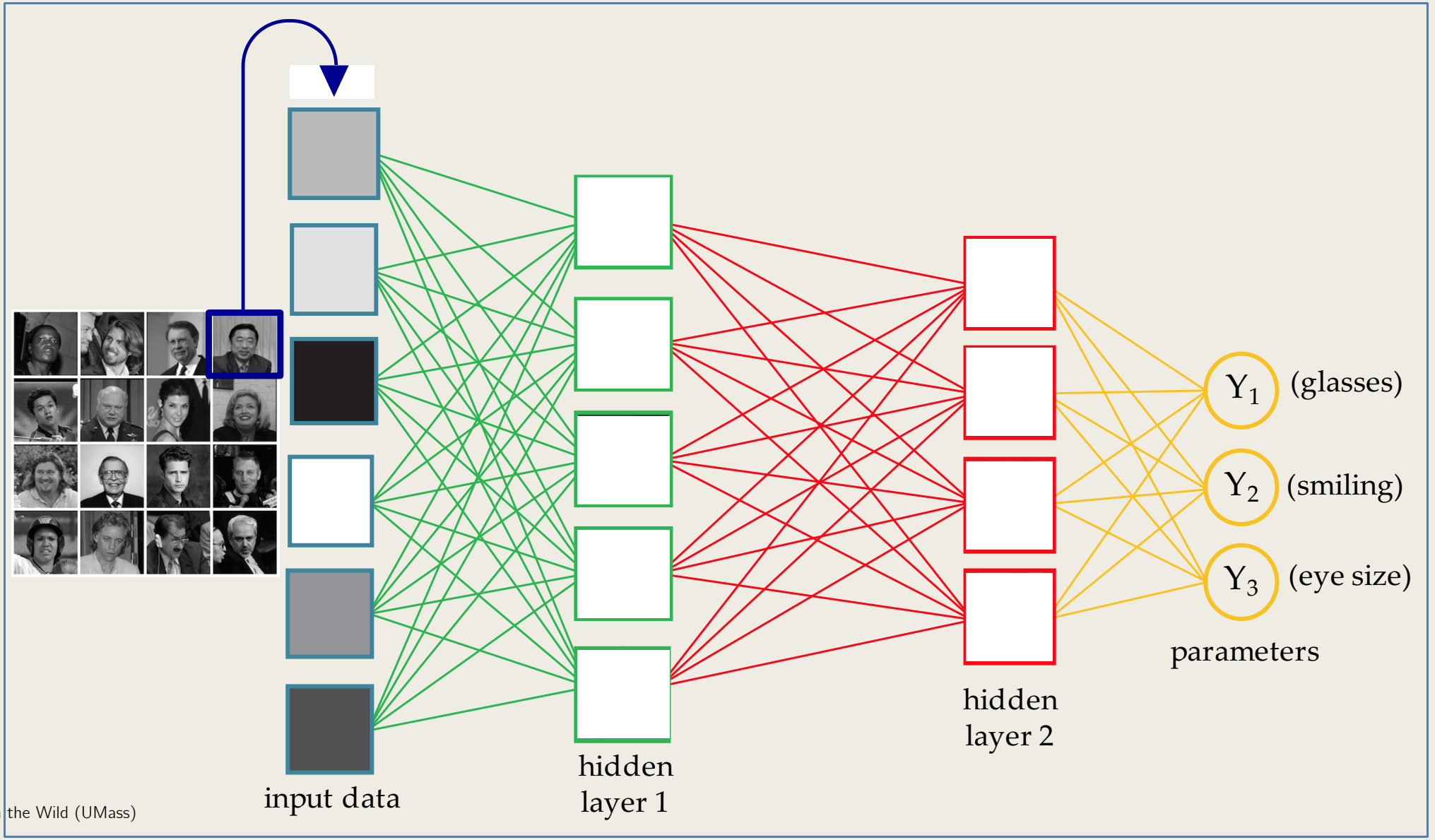
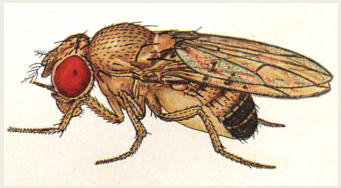
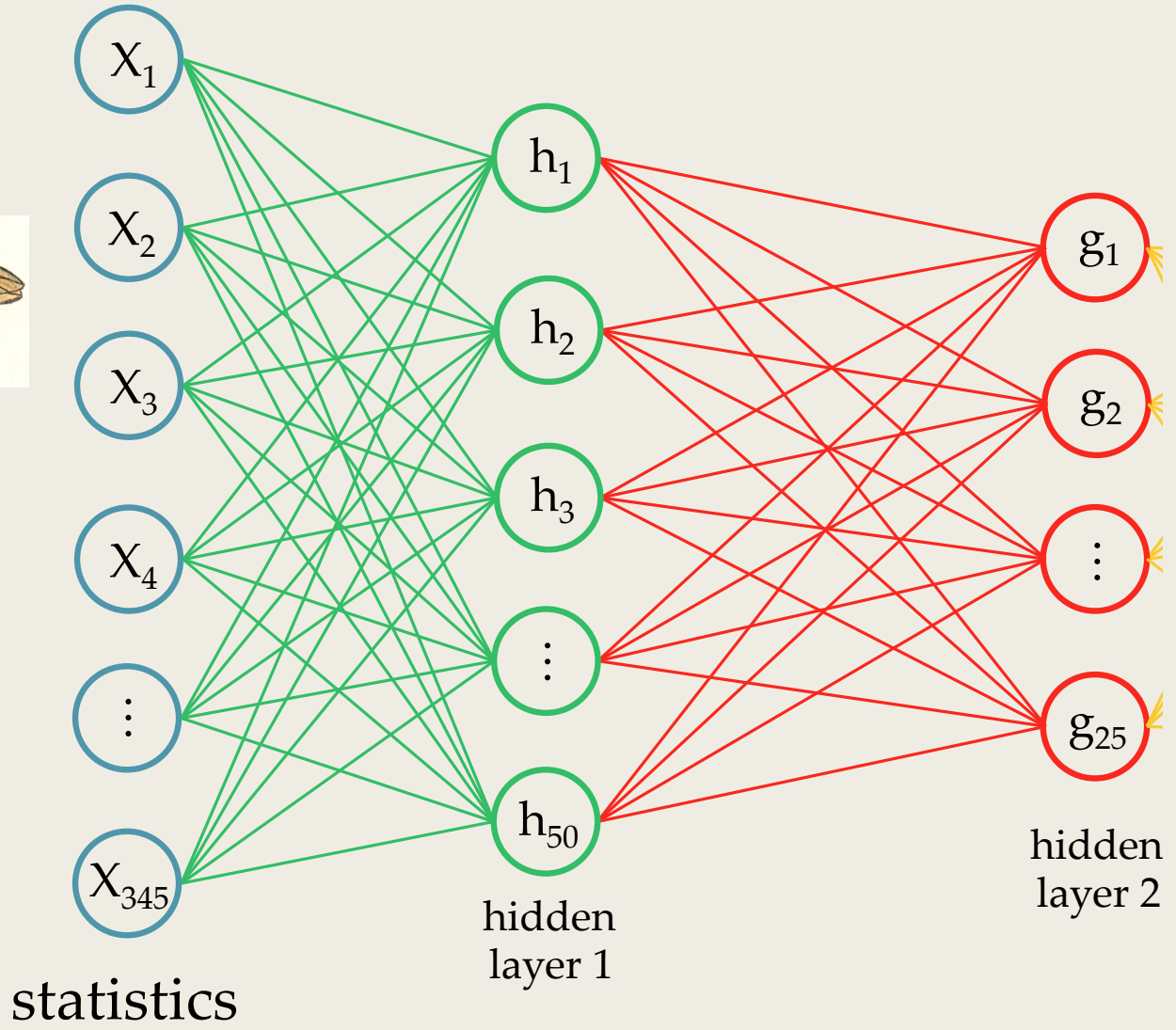


Image: Labeled Faces in the Wild (UMass)

2016: **evoNet** – deep learning with summary statistics

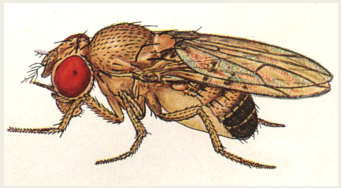


GACTGGCTA
AGCTAGCTT
TAATCCGCA

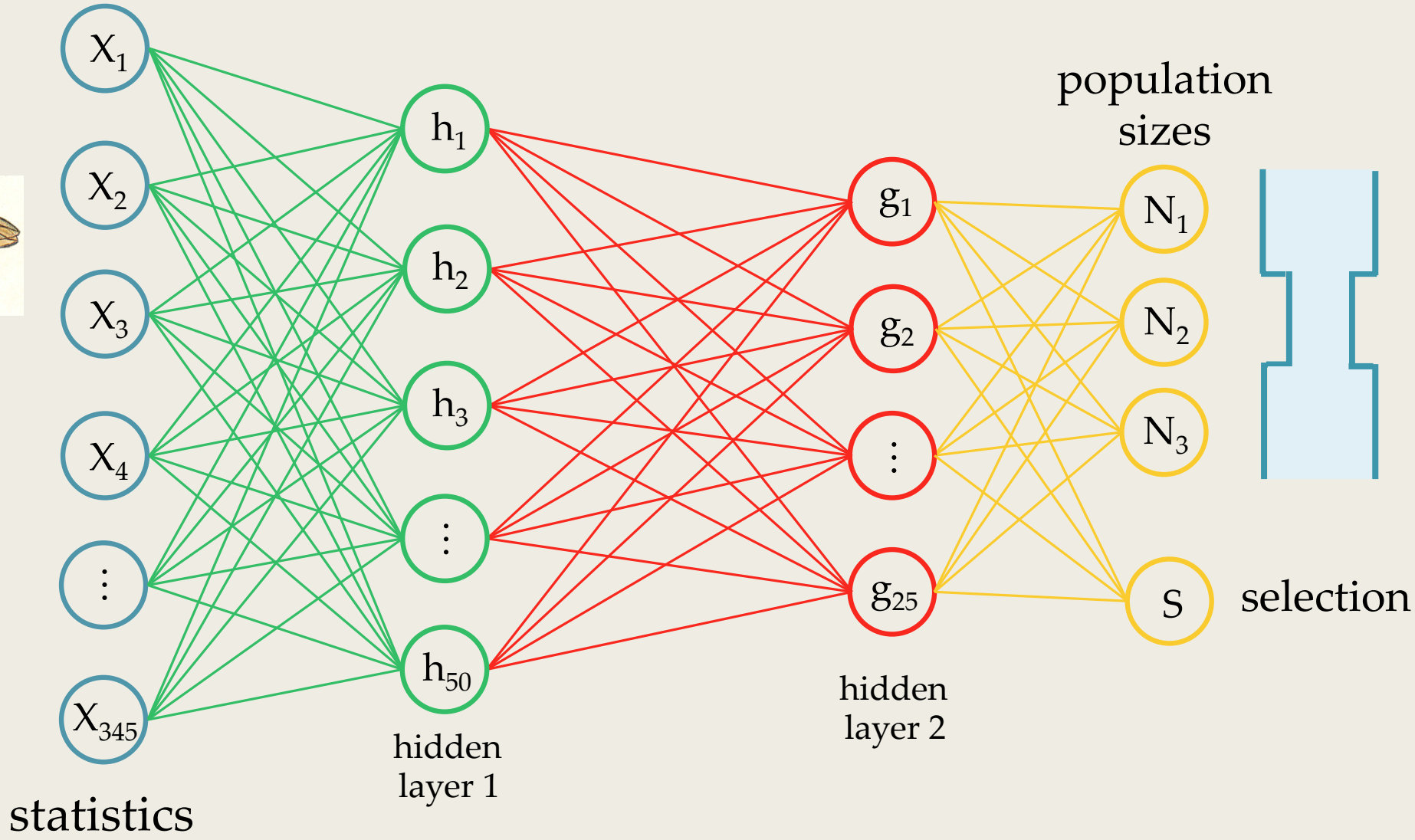


"Deep learning for population genetic inference", *PLOS Comp Bio*, 2016

2016: **evoNet** – deep learning with summary statistics

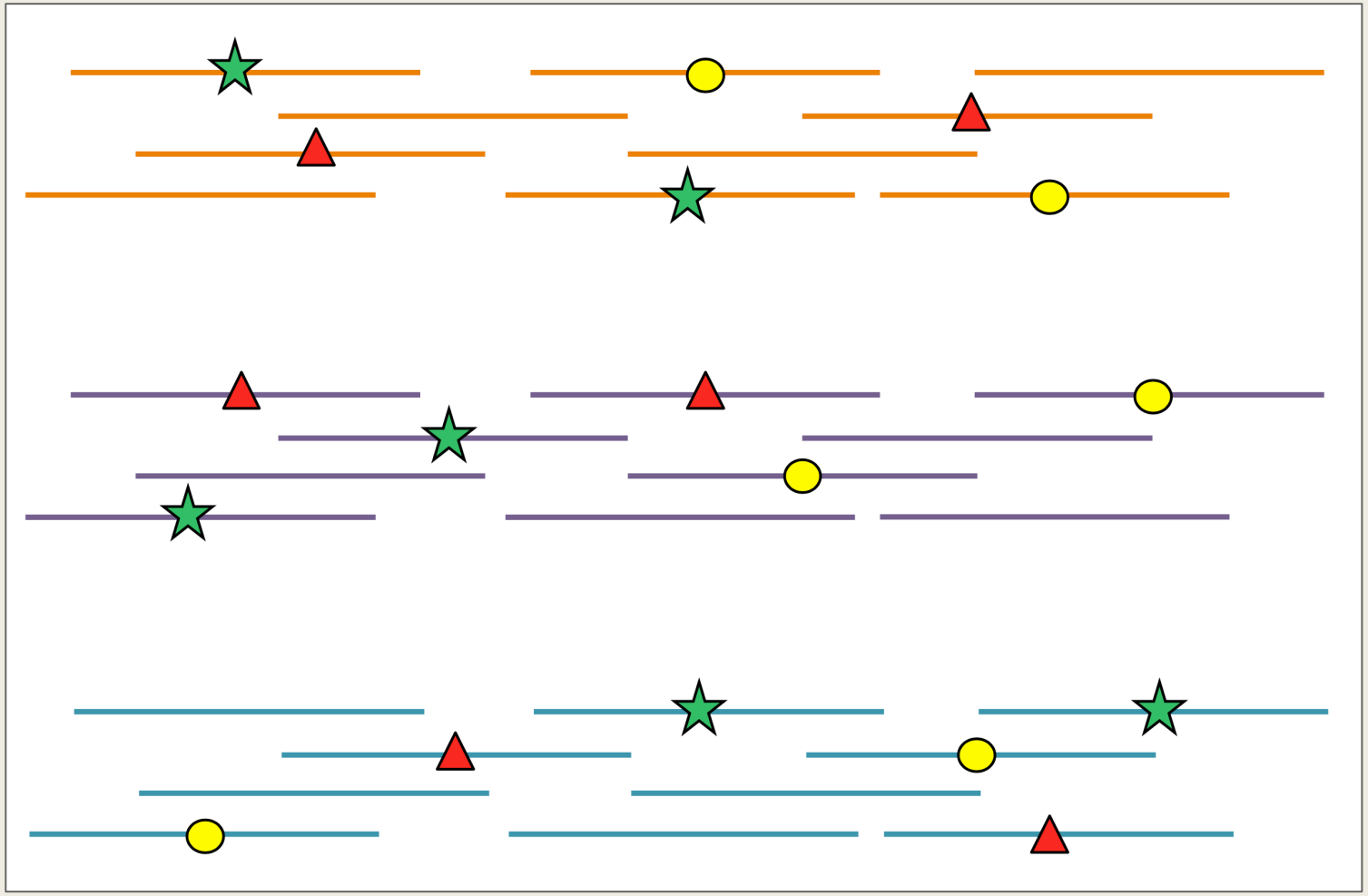


GACTGGCTA
AGCTAGCTT
TAATCCGCA



"Deep learning for population genetic inference", *PLOS Comp Bio*, 2016

Training Data: simulated data under different modes of selection



Results: confusion matrices for natural selection

Predicted Class

True Class	No selection	New mutation	Existing mutation	Balancing
No selection	1.000	0.000	0.000	0.000
New mutation	0.978	0.007	0.000	0.015
Existing mutation	1.000	0.000	0.000	0.000
Balancing	1.000	0.000	0.000	0.000

Random initialization

Layer-by-layer training

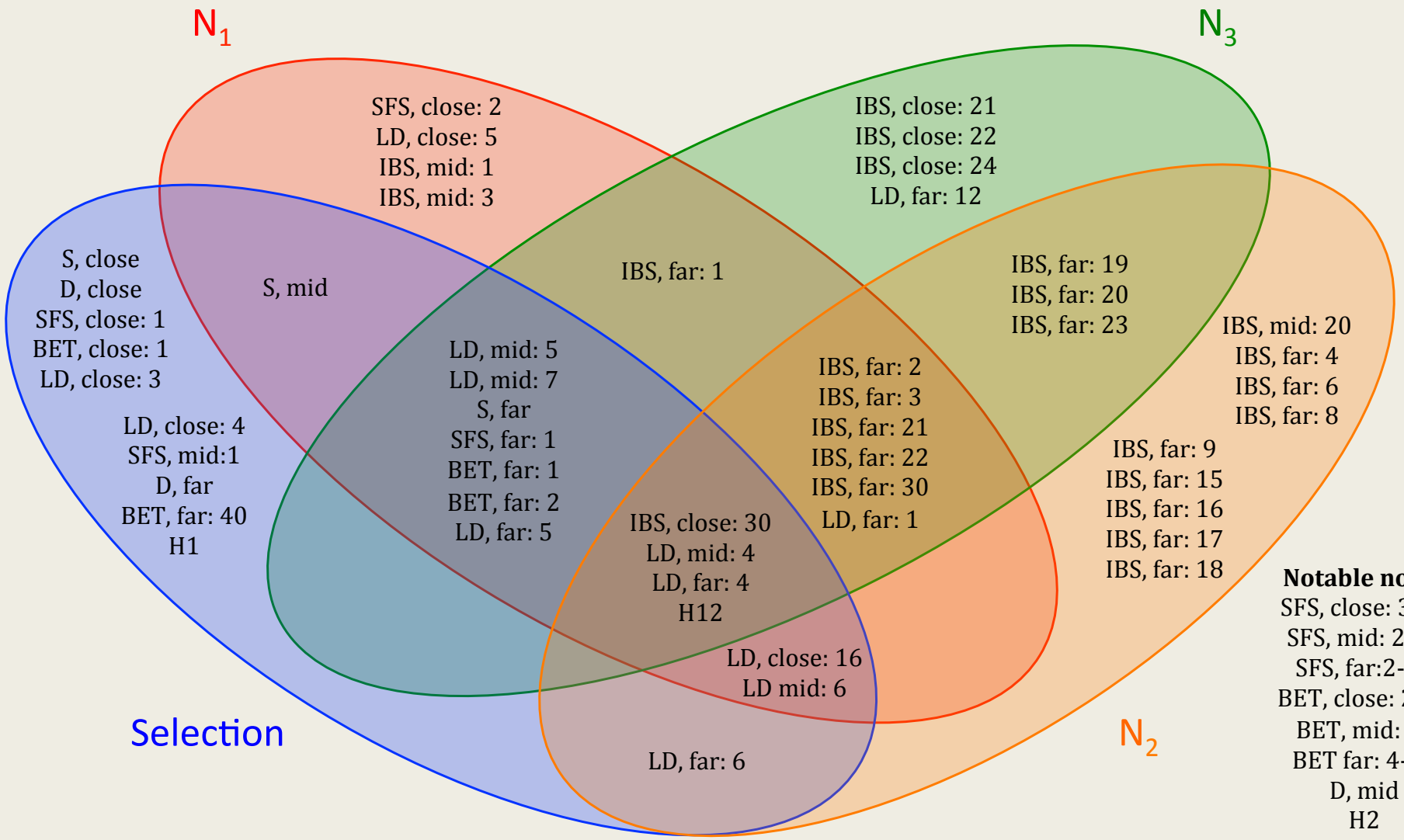
Predicted Class

True Class	No selection	New mutation	Existing mutation	Balancing
No selection	1.000	0.000	0.000	0.000
New mutation	0.145	0.831	0.004	0.021
Existing mutation	0.011	0.001	0.987	0.000
Balancing	0.030	0.028	0.001	0.941

Results: real *Drosophila* data

Selection class	Number of regions
No selection	1191
New mutation	2572
Existing mutation	429
Balancing	637

Feature selection: “best” statistics via perturbation of the inputs



Can we do better? Convolutional neural networks (CNNs)

Motivation

1. Need different summary statistics for each application

2. Computationally intensive

Flagel, Brandvain, Schrider. "The unreasonable effectiveness of convolutional neural networks in population genetic inference."
Molecular biology and evolution, 2018

Chan, Perrone, Spence, Jenkins, **Mathieson**, Song. "A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks"
NeurIPS, 2018, https://github.com/popgenmethods/defiNETti

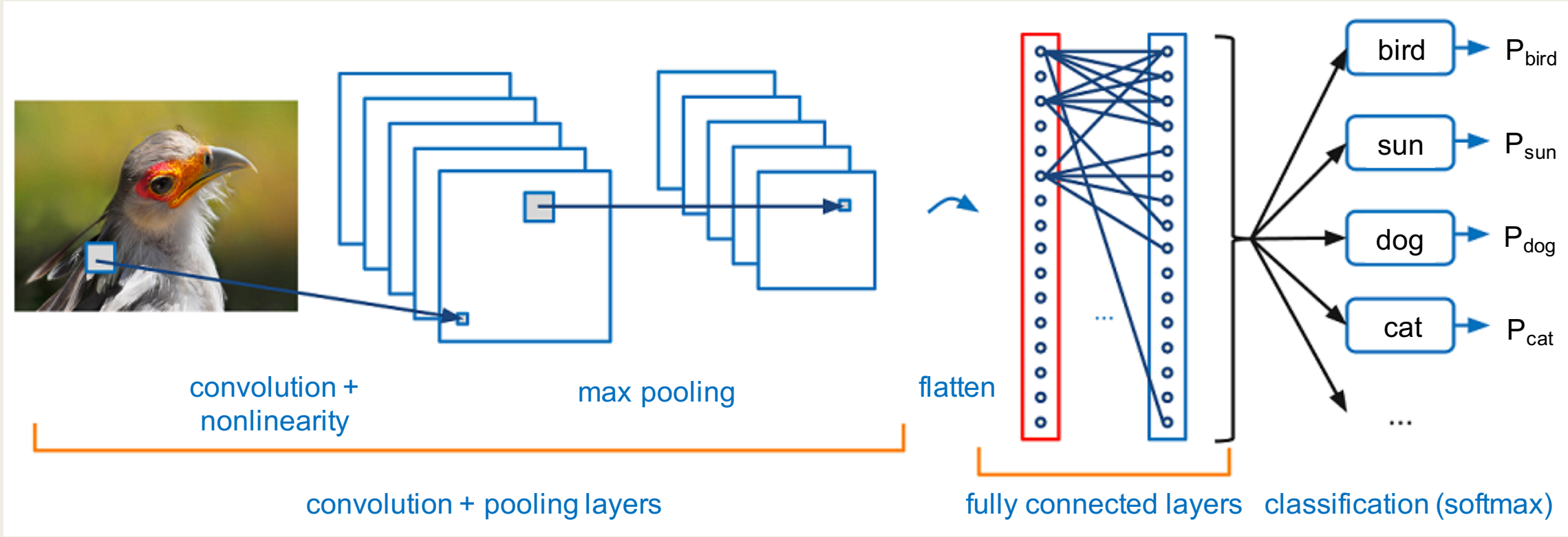
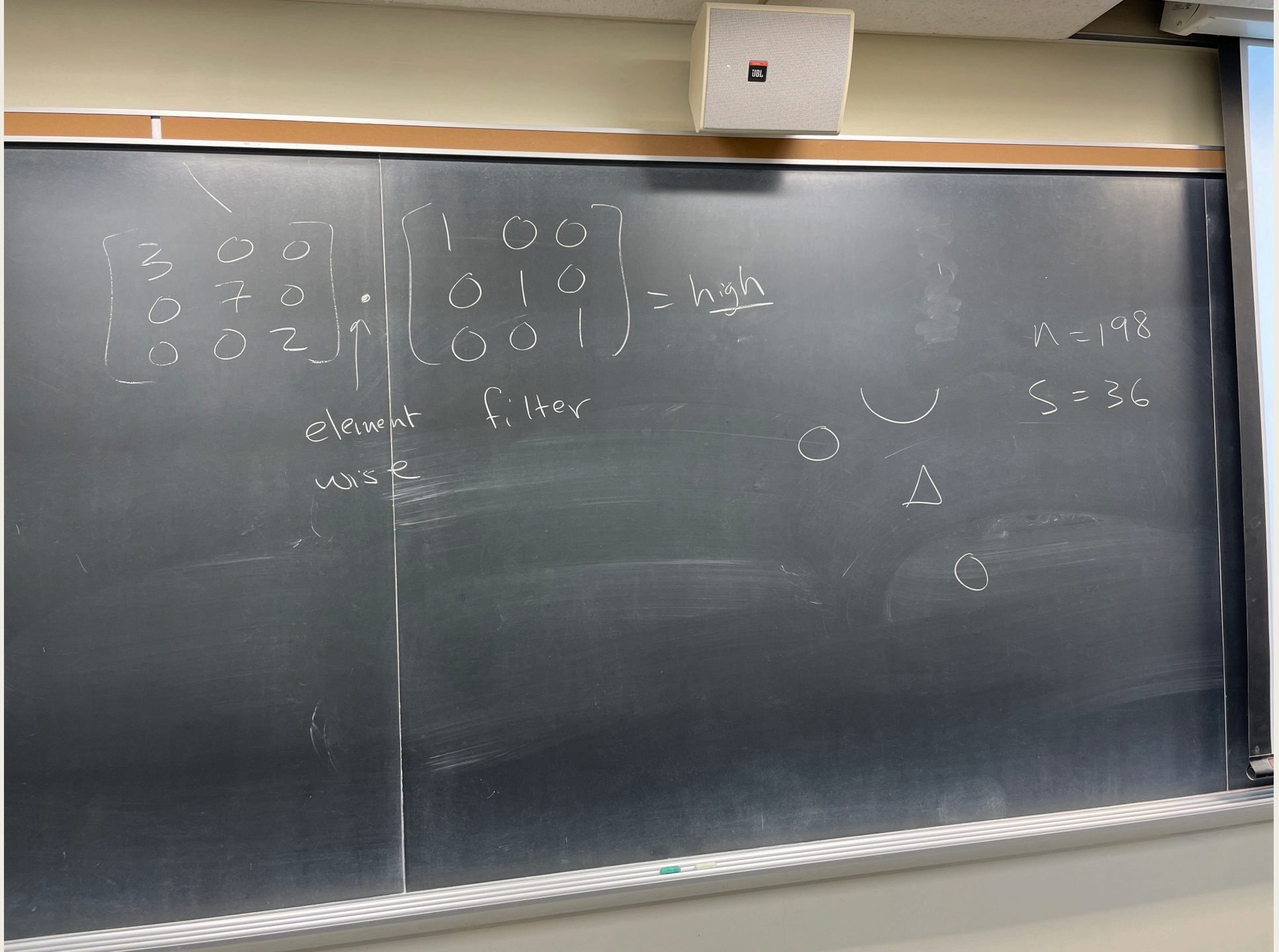
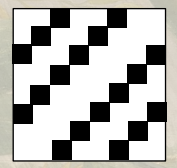


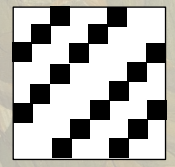
Figure: Adit Deshpande

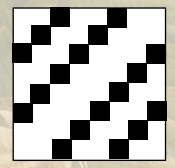

$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ = high
element wise
filter
n = 198
S = 36

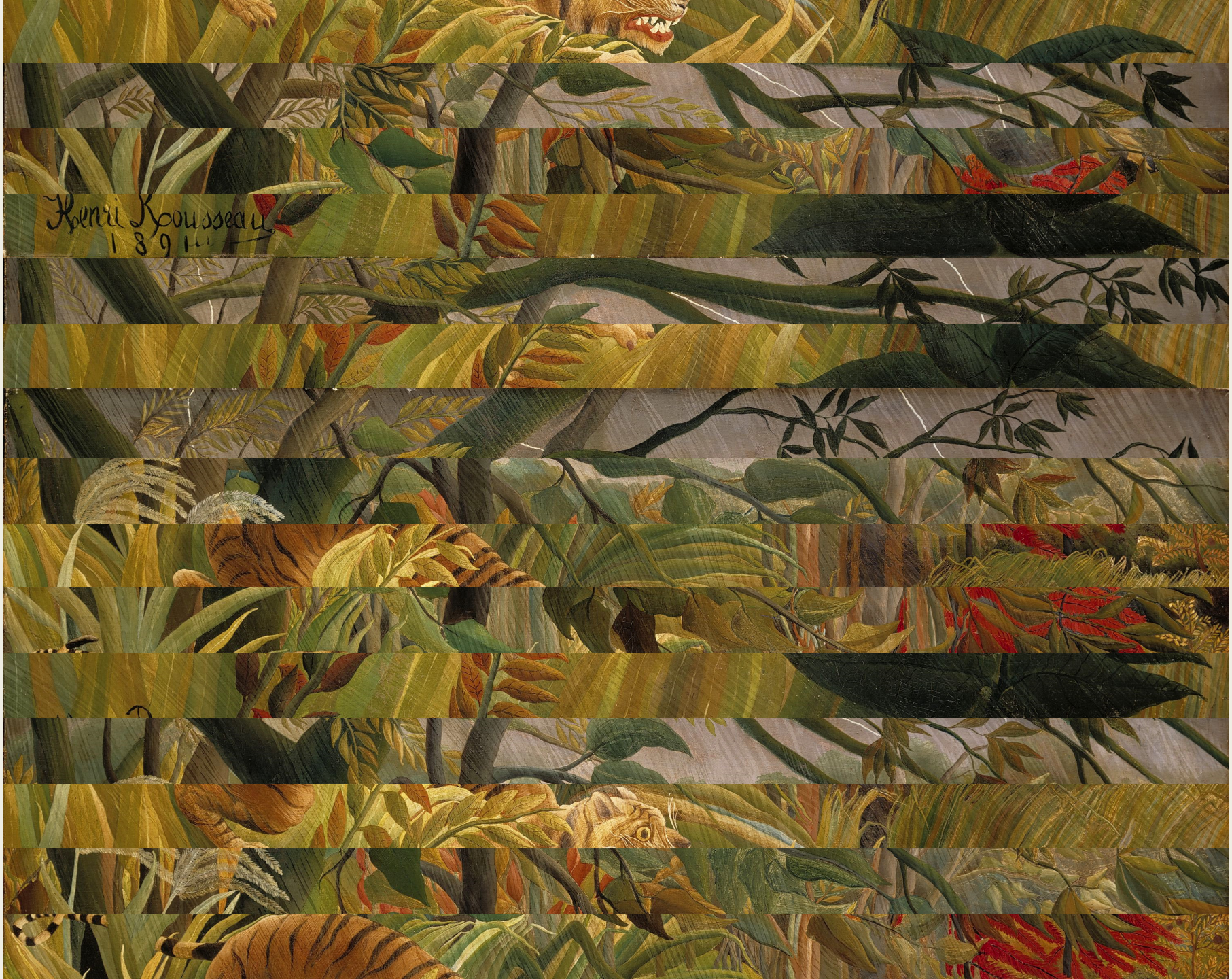

Henri Rousseau
1891

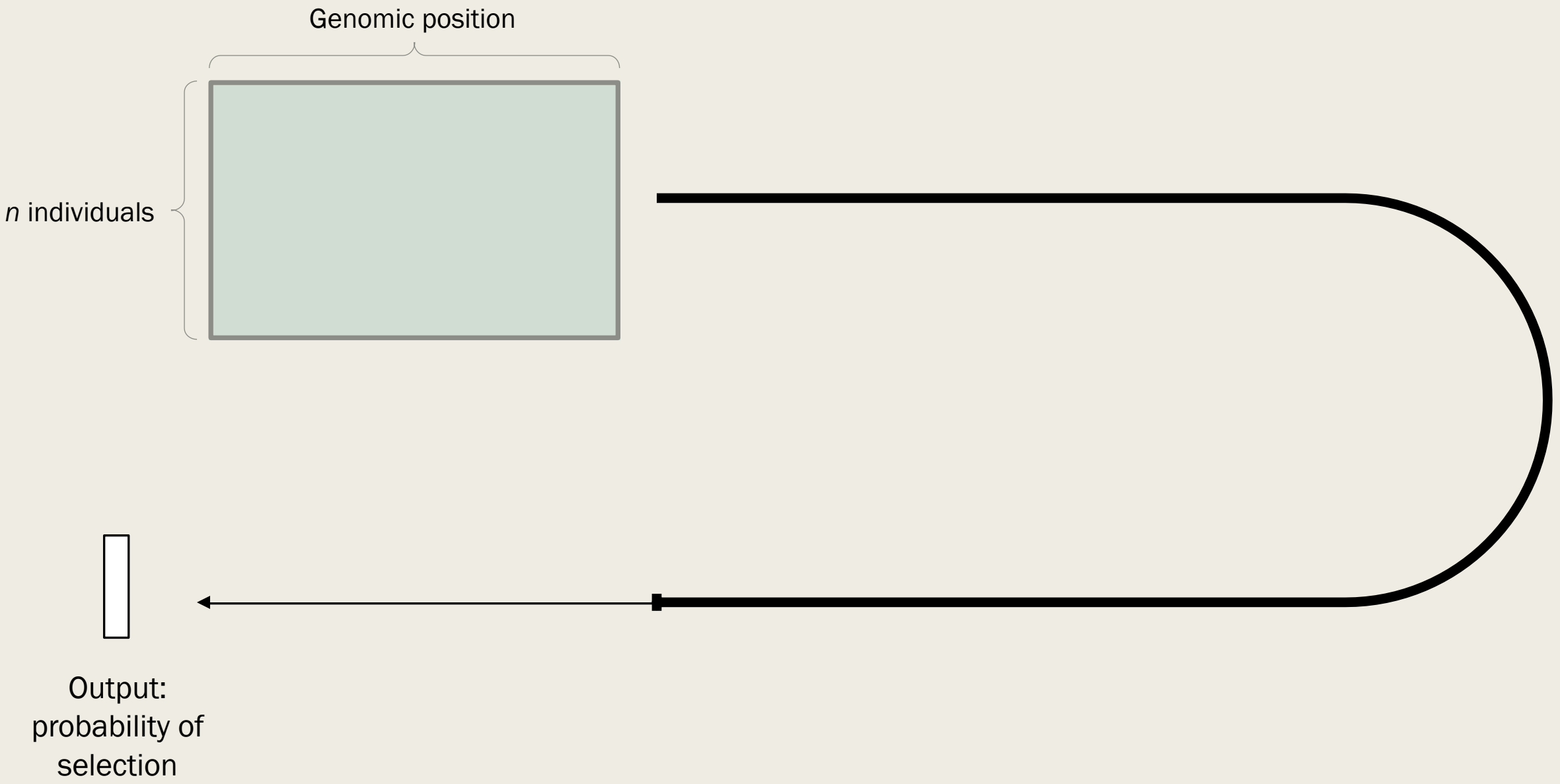

Henri Rousseau
1891

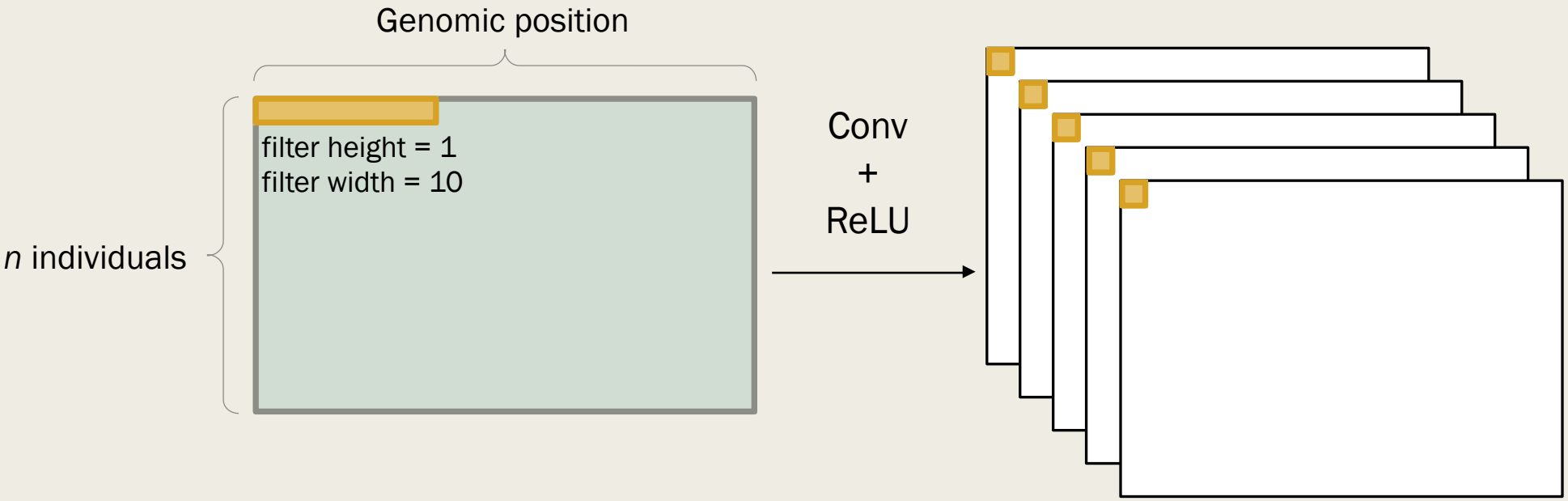
 ·  = low

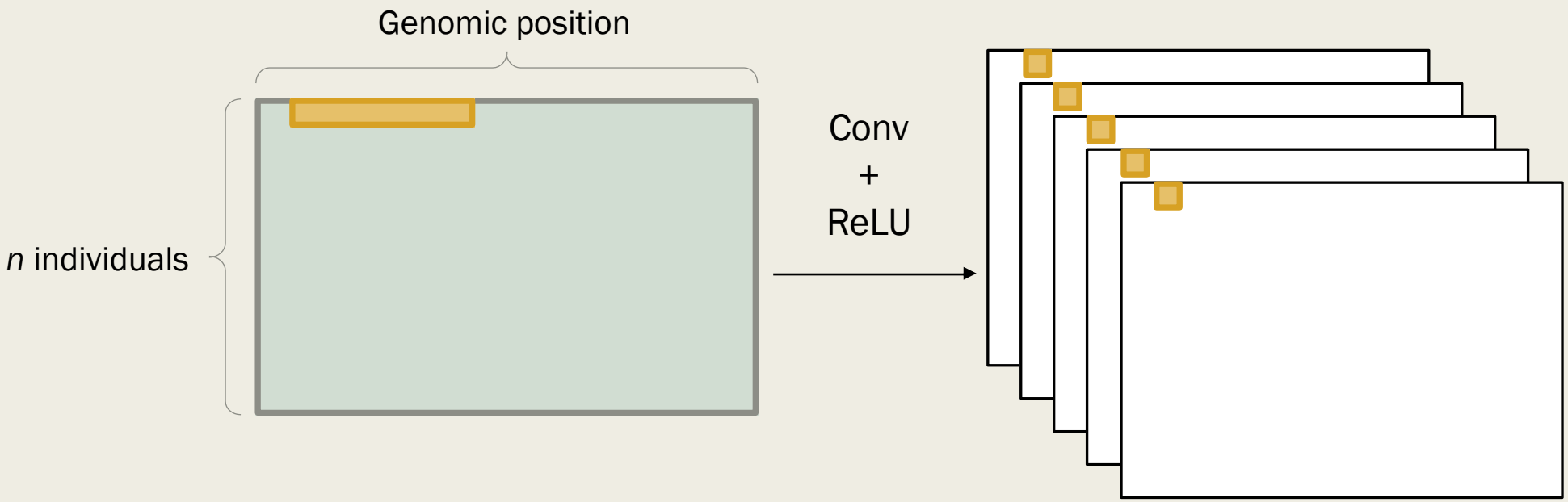
 ·  = low

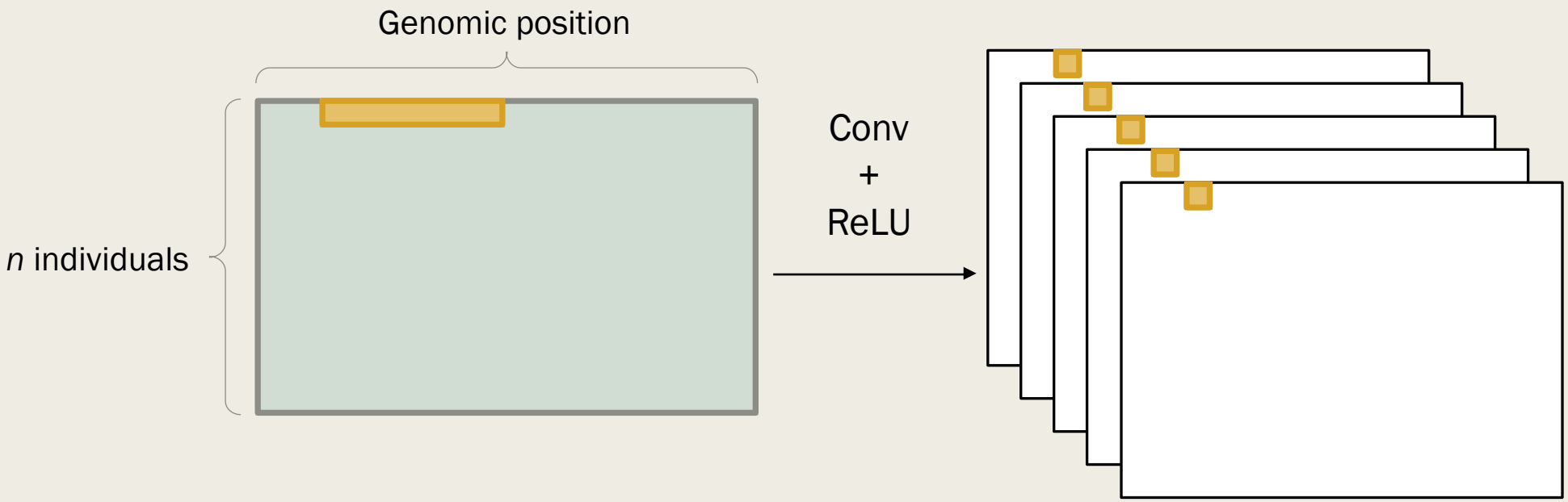
 ·  = HIGH

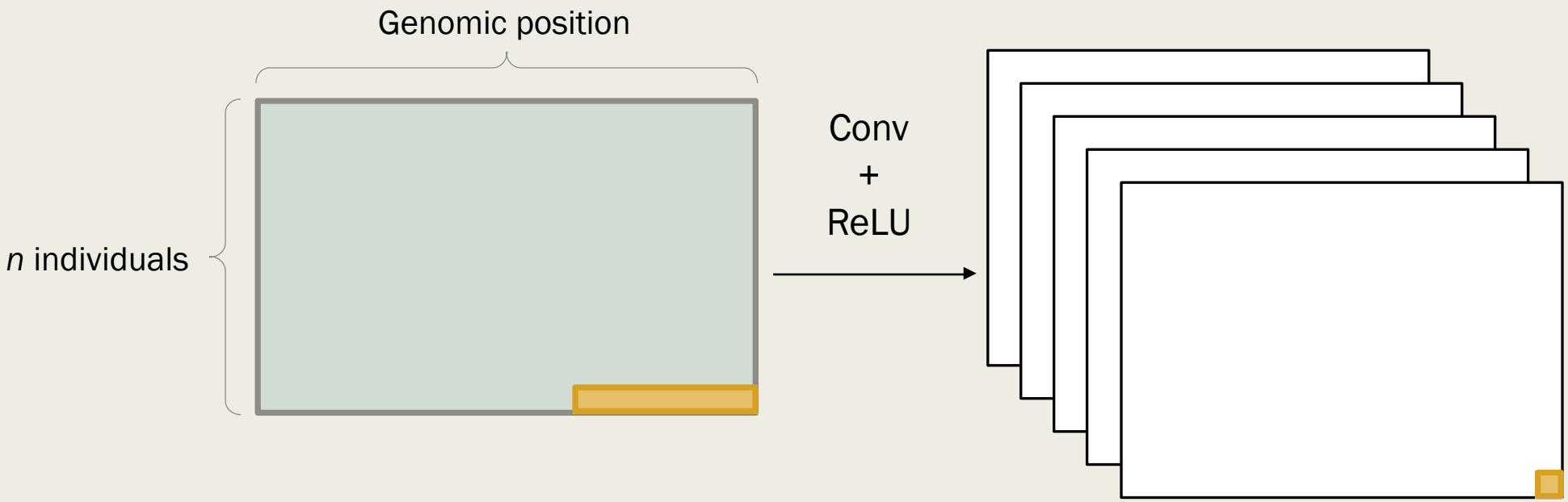

Henri Rousseau
1891

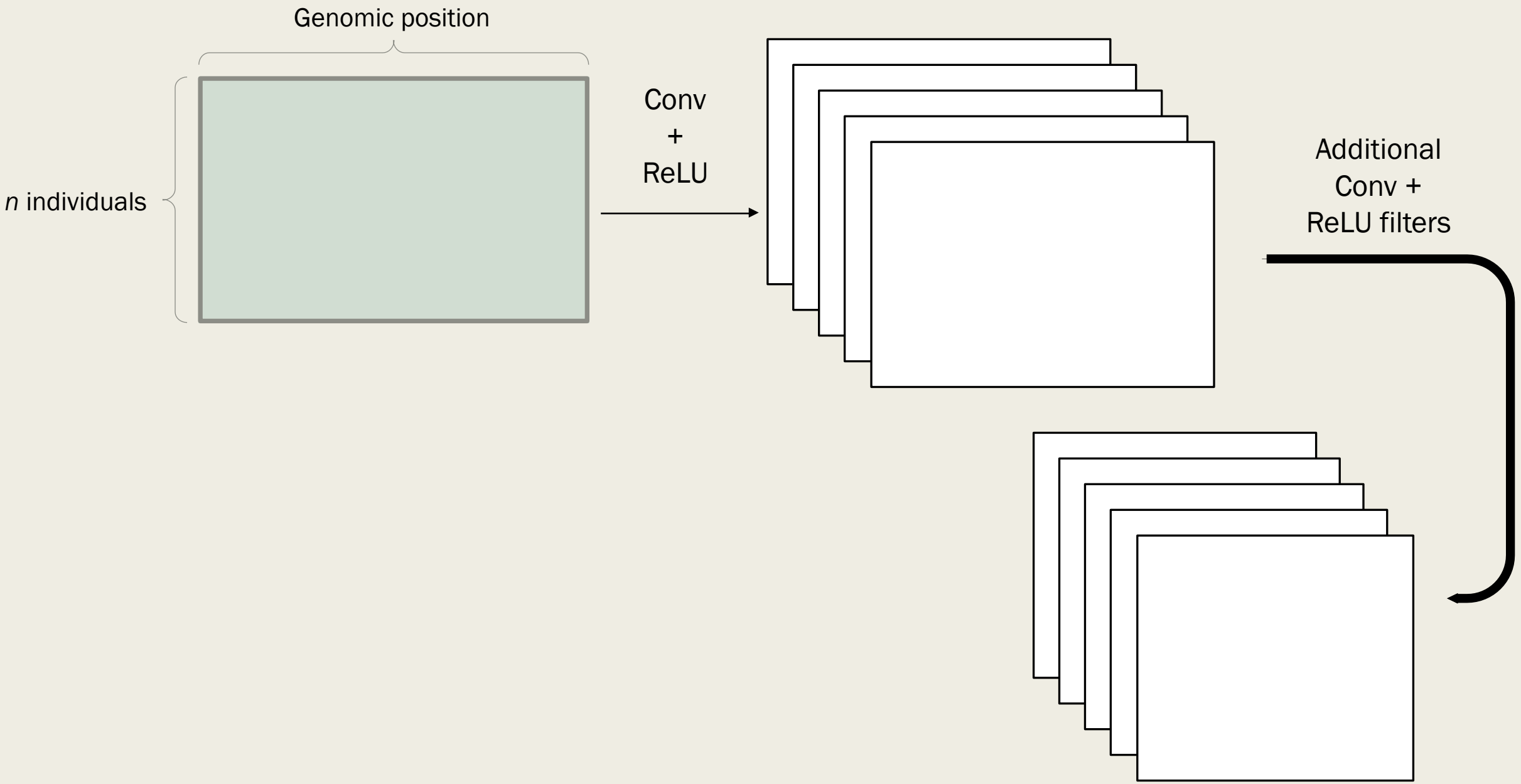

Genomic position
n individuals
Output:
probability of
selection

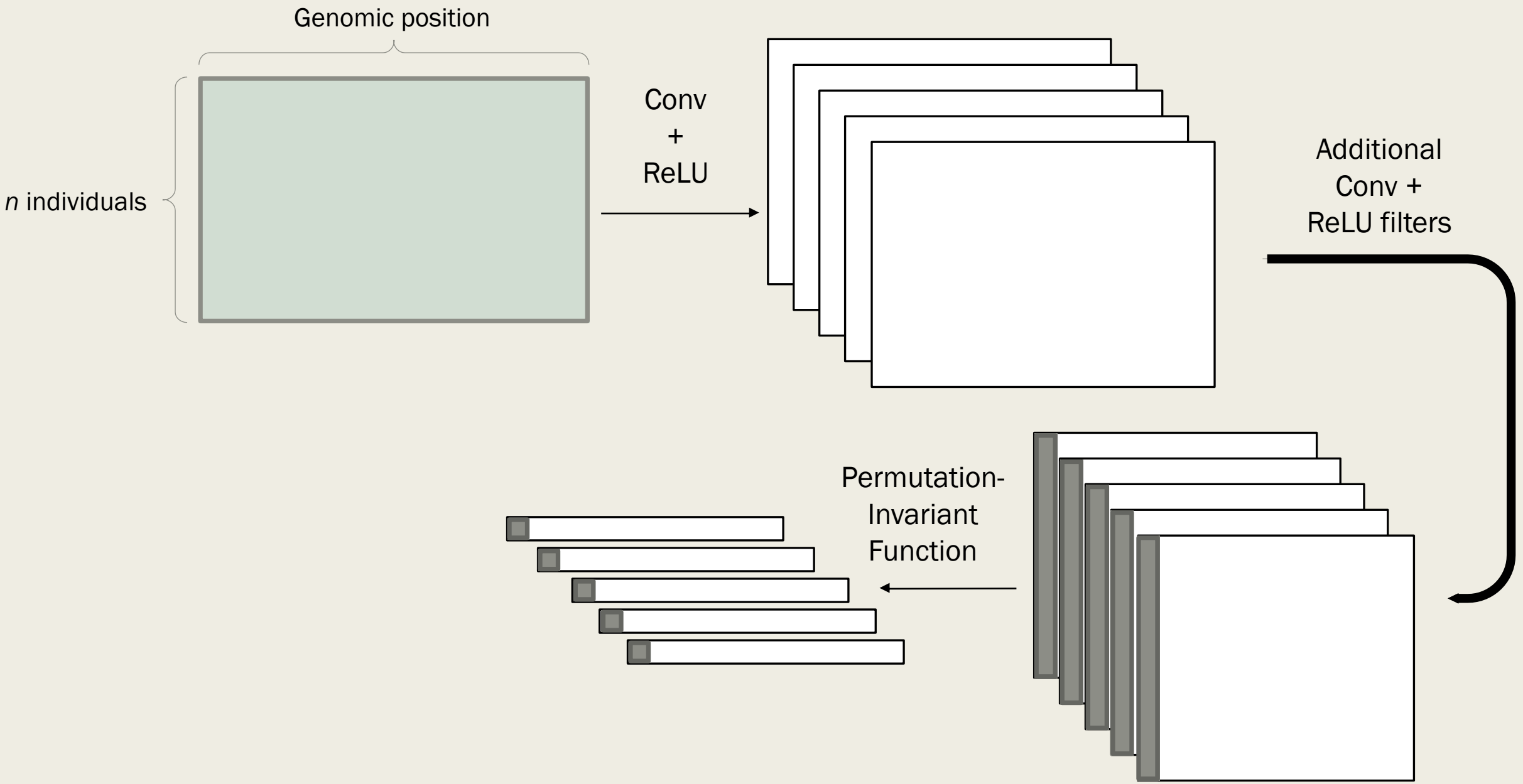

Genomic position
filter height = 1
filter width = 10
n individuals
Conv
+
ReLU

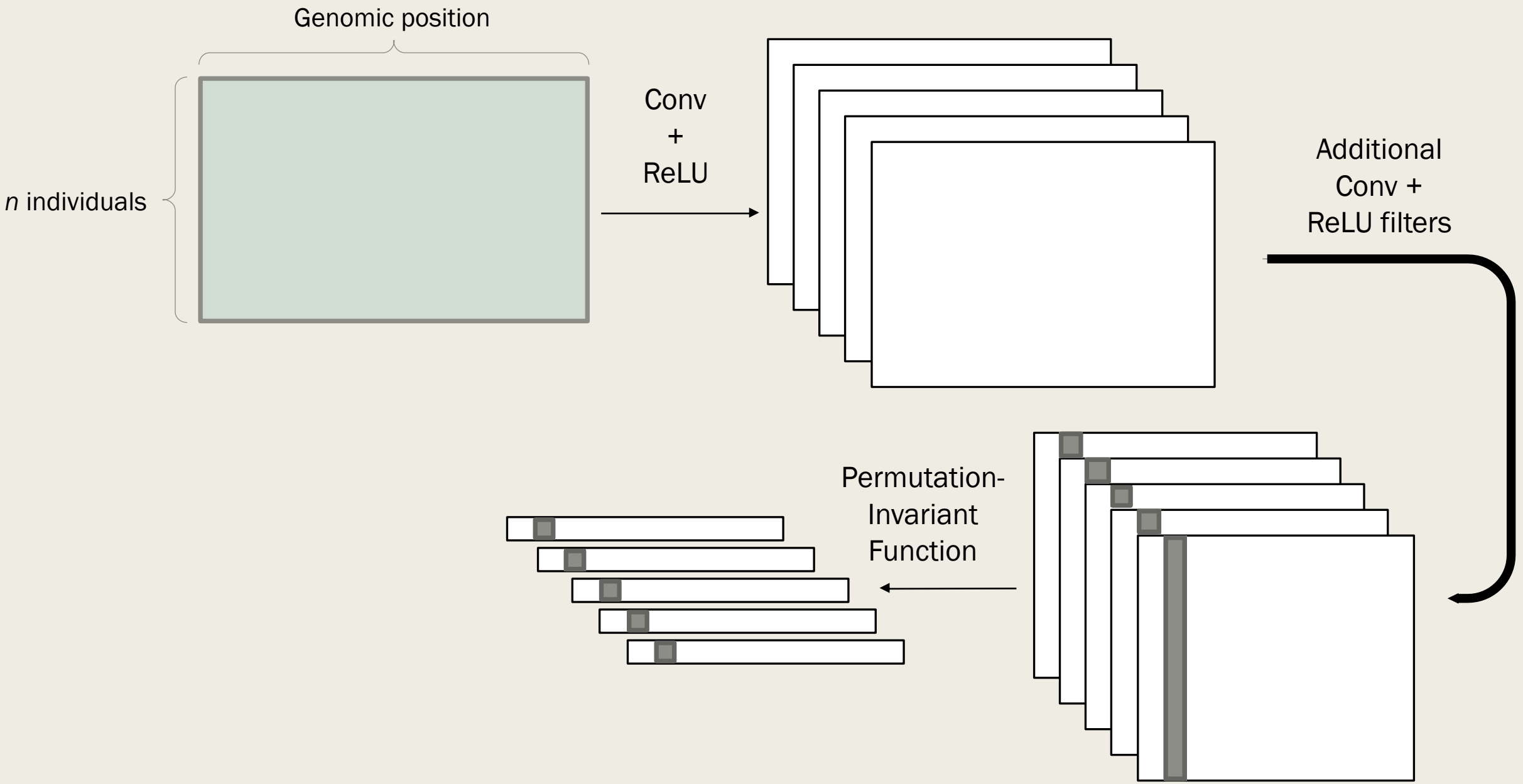

Genomic position
n individuals
Conv
+
ReLU

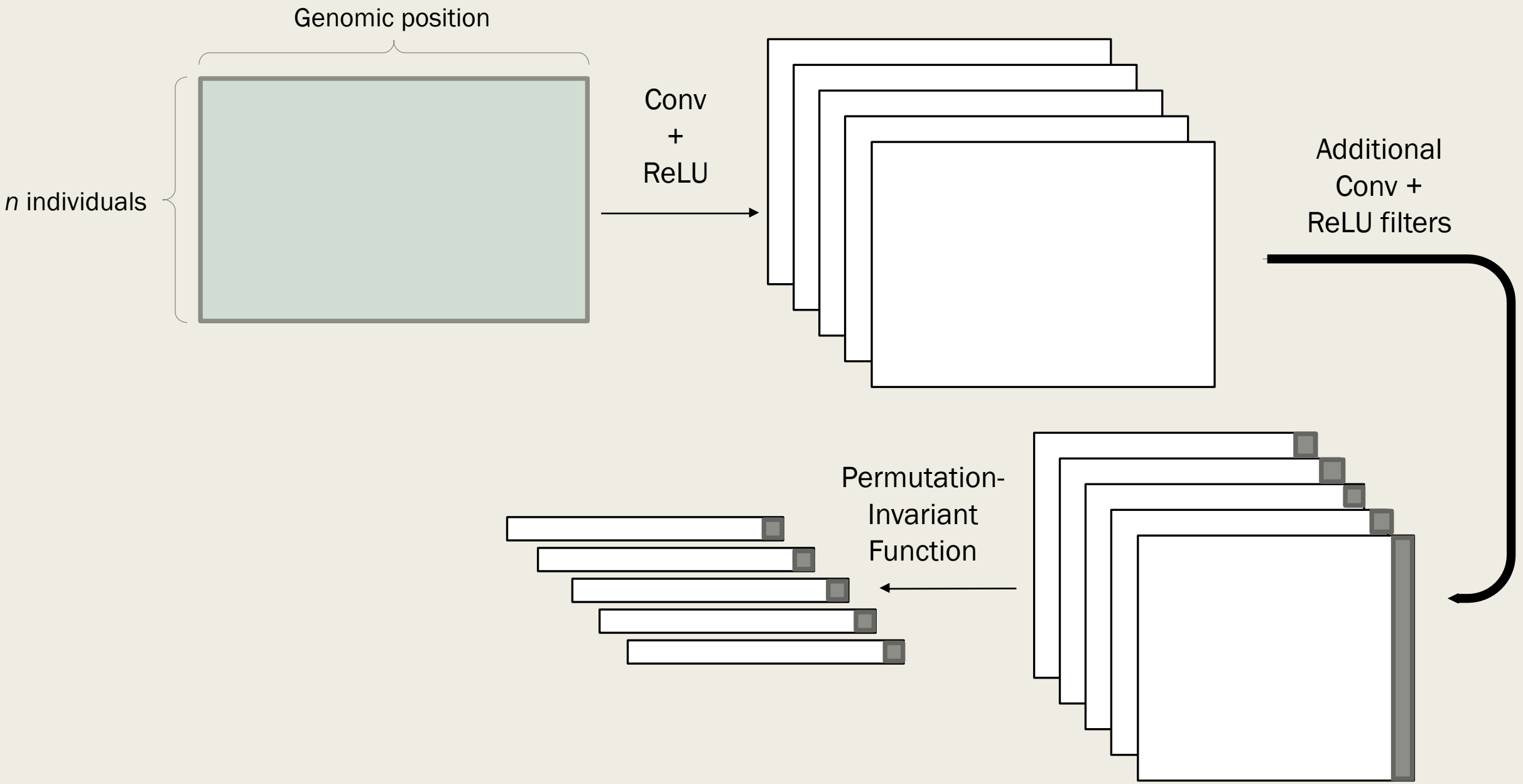

Genomic position
n individuals
Conv
+
ReLU

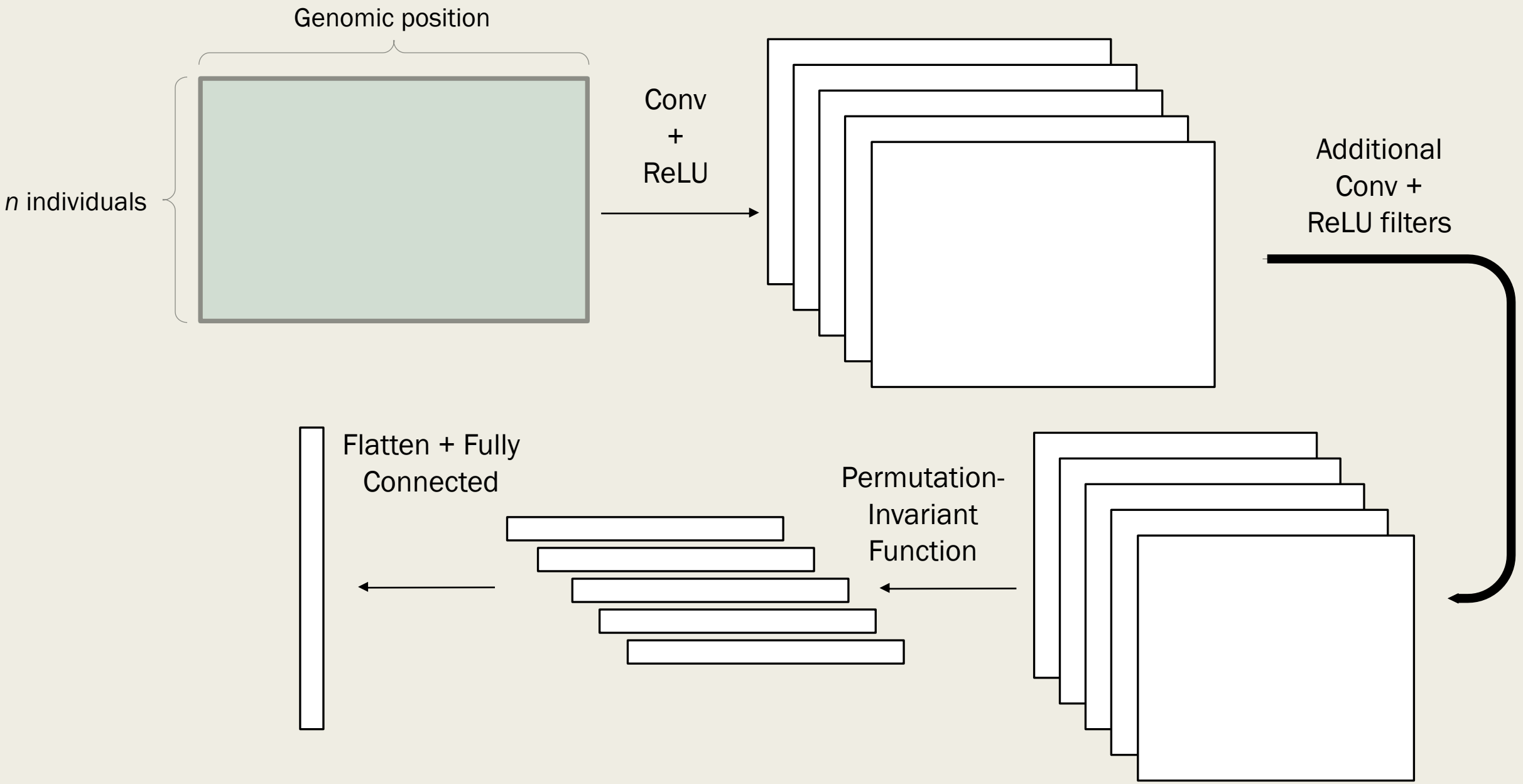

Genomic position
n individuals
Conv
+
ReLU

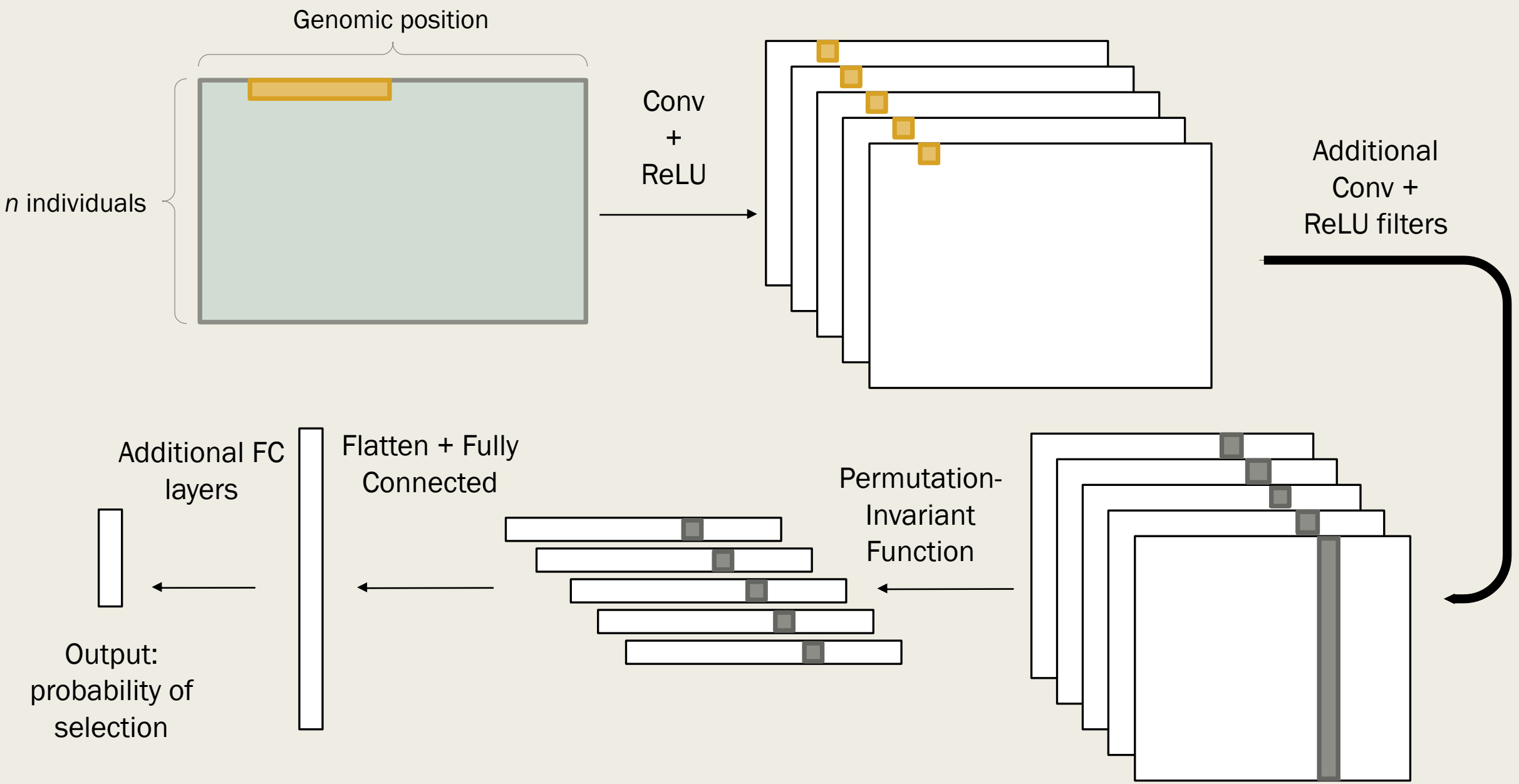

Genomic position
n individuals
Conv
+
ReLU
Additional
Conv +
ReLU filters


Genomic position
n individuals
Conv
+
ReLU
Additional
Conv +
ReLU filters
Permutation-
Invariant
Function

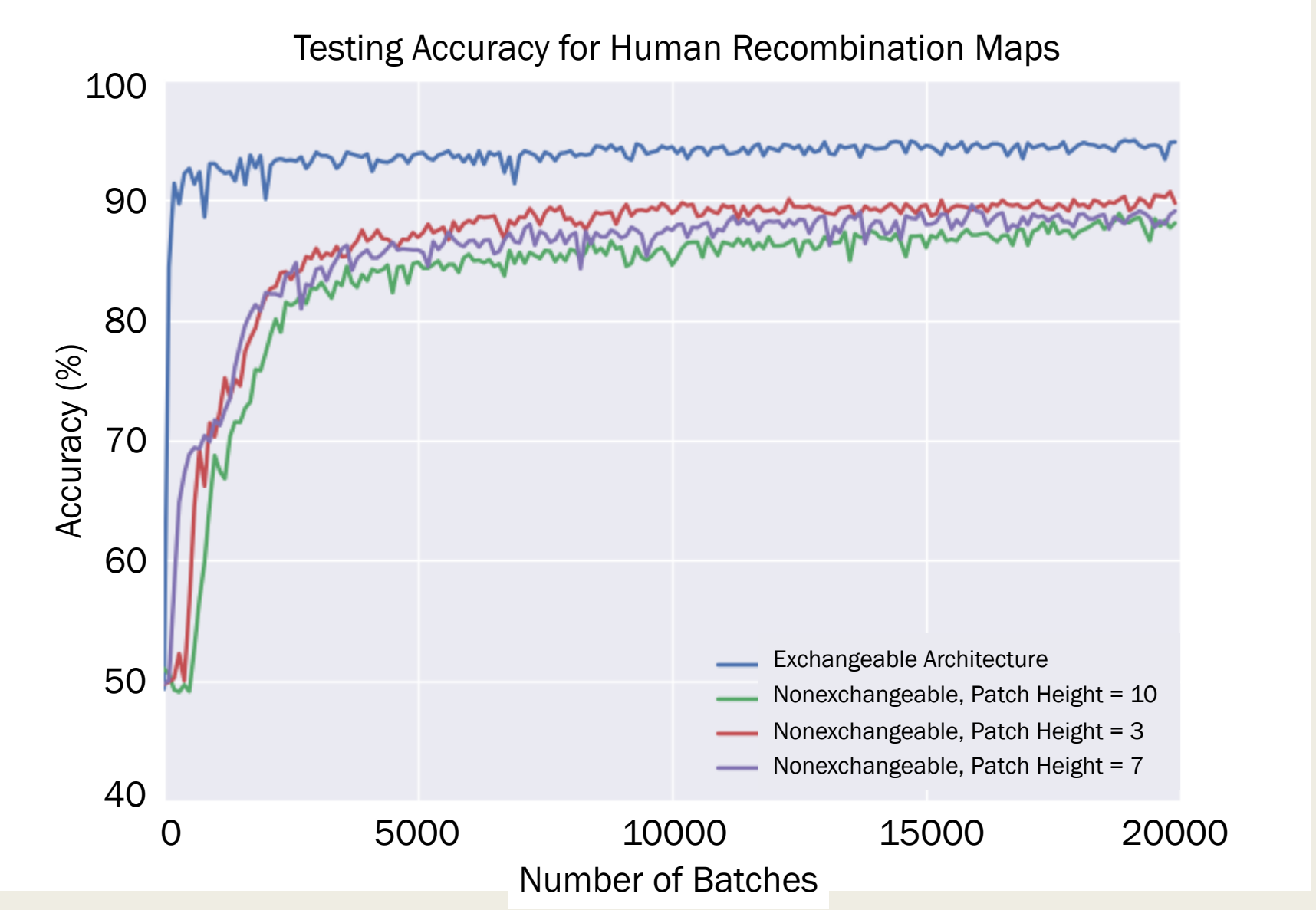

Genomic position
n individuals
Conv
+
ReLU
Additional
Conv +
ReLU filters
Permutation-
Invariant
Function


Genomic position
n individuals
Conv
+
ReLU
Additional
Conv +
ReLU filters
Permutation-
Invariant
Function


Genomic position
n individuals
Conv
+
ReLU
Additional
Conv +
ReLU filters
Flatten + Fully
Connected
Permutation-
Invariant
Function


Genomic position
n individuals
Conv
+
ReLU
Additional
Conv +
ReLU filters
Permutation-
Invariant
Function
Flatten + Fully
Connected
Additional FC
layers
Output:
probability of
selection

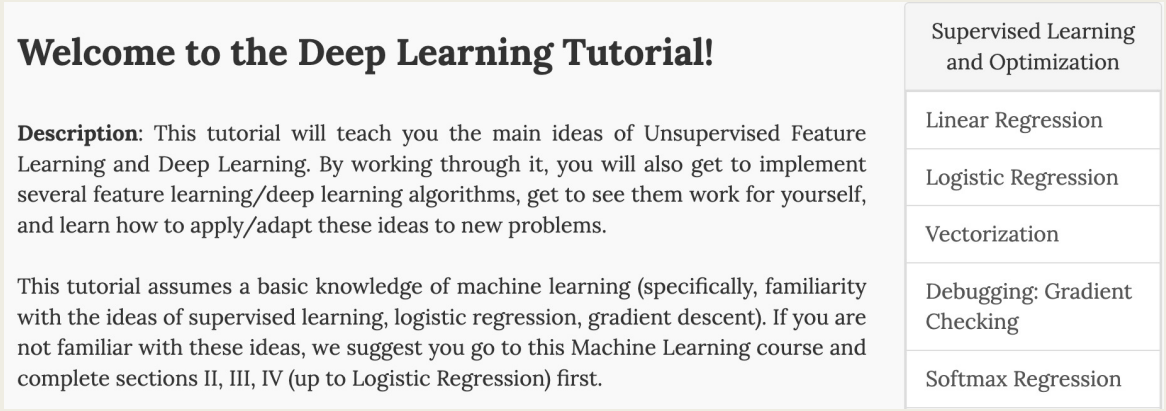
Impact of exchangeable architecture



Deep learning and CNN resources

- Deep learning tutorial

http://ufldl.stanford.edu/tutorial/



Welcome to the Deep Learning Tutorial!

Description: This tutorial will teach you the main ideas of Unsupervised Feature Learning and Deep Learning. By working through it, you will also get to implement several feature learning/deep learning algorithms, get to see them work for yourself, and learn how to apply/adapt these ideas to new problems.

This tutorial assumes a basic knowledge of machine learning (specifically, familiarity with the ideas of supervised learning, logistic regression, gradient descent). If you are not familiar with these ideas, we suggest you go to this Machine Learning course and complete sections II, III, IV (up to Logistic Regression) first.

- Tensorflow CNN tutorial https://www.tensorflow.org/tutorials/images/cnn

- Tensorflow advanced quickstart

https://www.tensorflow.org/tutorials/quickstart/advanced

Brief detour to Hardy-Weinberg + recap expected value

Discrete probability distribution

- Let X be a random variable that can take on values $x_1, x_2, \ldots, x_k$
- Example: a die that can take on values 1,2,3,4,5,6
- If we rolled the die many times and took the average, we would have an estimate of the expected value
- Let p_i = the probability of observing value x_i
- Example: p_1=0, p_2=1/6, p_3=1/6, p_4=1/6, p_5=1/6, p_6 = 1/3
- We should check that the sum of the probabilities of all possible values is 1

$$\sum_{i=1}^{k} p_i = 1$$

- Compute expectation:

$$E[X] = p_1 x_1 + p_2 x_2 + \cdots + p_k x_k = \sum_{i=1}^{k} p_i x_i$$

$$0 \cdot 1 + \frac{1}{6}(2+3+4+5) + \frac{1}{3} \cdot 6 = 4\frac{1}{3}$$

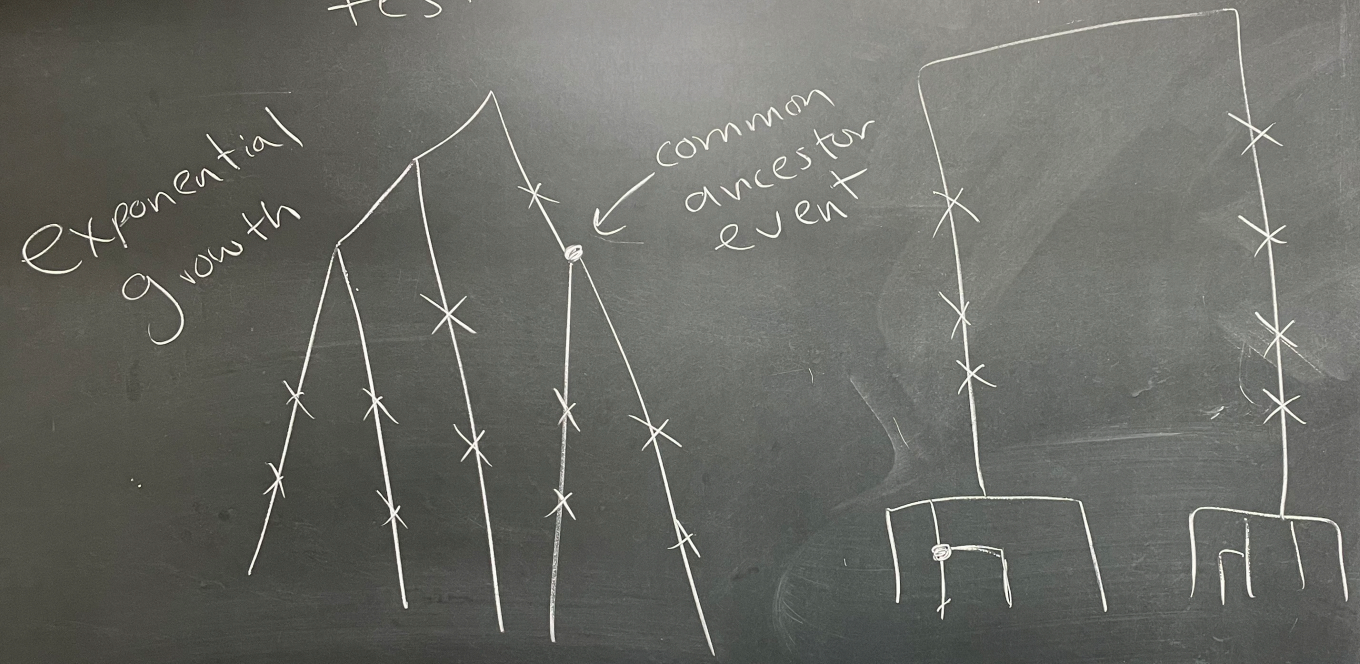
Hardy-Weinberg expectations

- If we have two alleles, A and a, then each individual can have *genotype* AA, Aa, or aa
- We say that AA and aa are *homozygous* and Aa is *heterozygous*
- If the genotype at this *locus* (site) is responsibly for a *Mendelian* (think: binary) *phenotype* and A is *dominant*, then AA and Aa will have the same phenotype
- In that case we would call aa recessive
- If aa is disease causing or *deleterious*, this can reduce the frequency of a through selection
- If most alleles either become fixed or die out, that means eventually everyone will either be aa or AA. This is called the *loss of heterozygosity*

Wright-Fisher Model (discrete) and Coalescent (continuous)

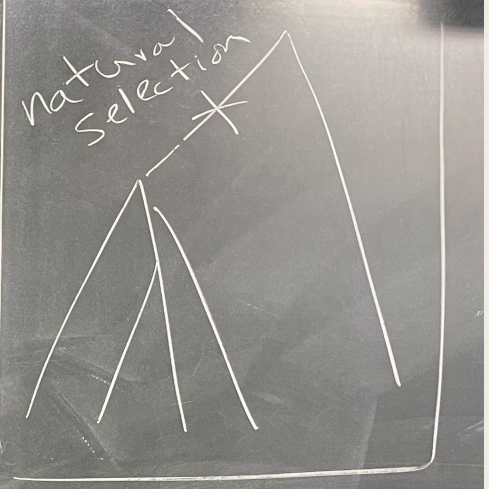
Tajima's D

test for natural selection

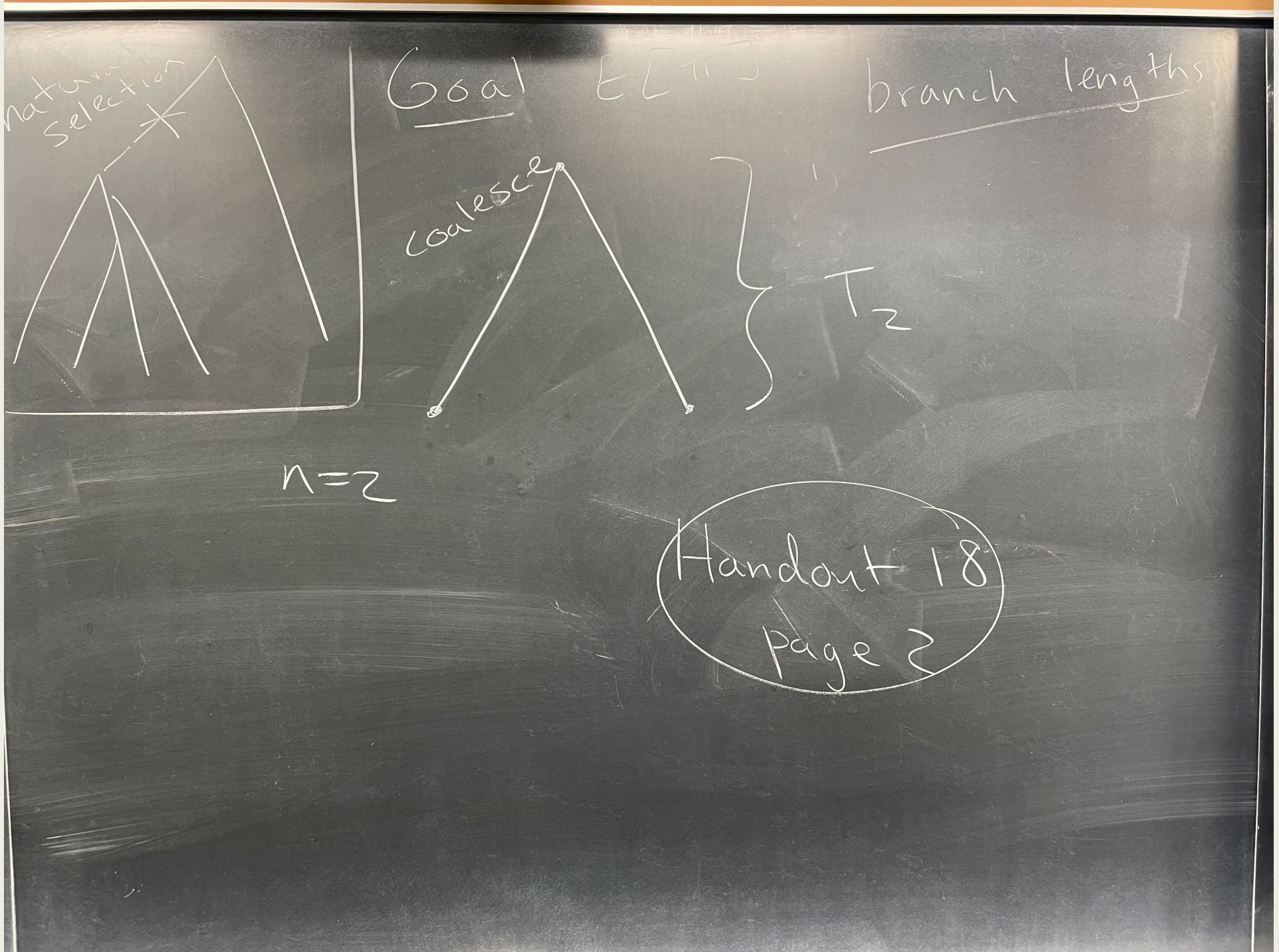


n = 5 Sample Size

Population Structure

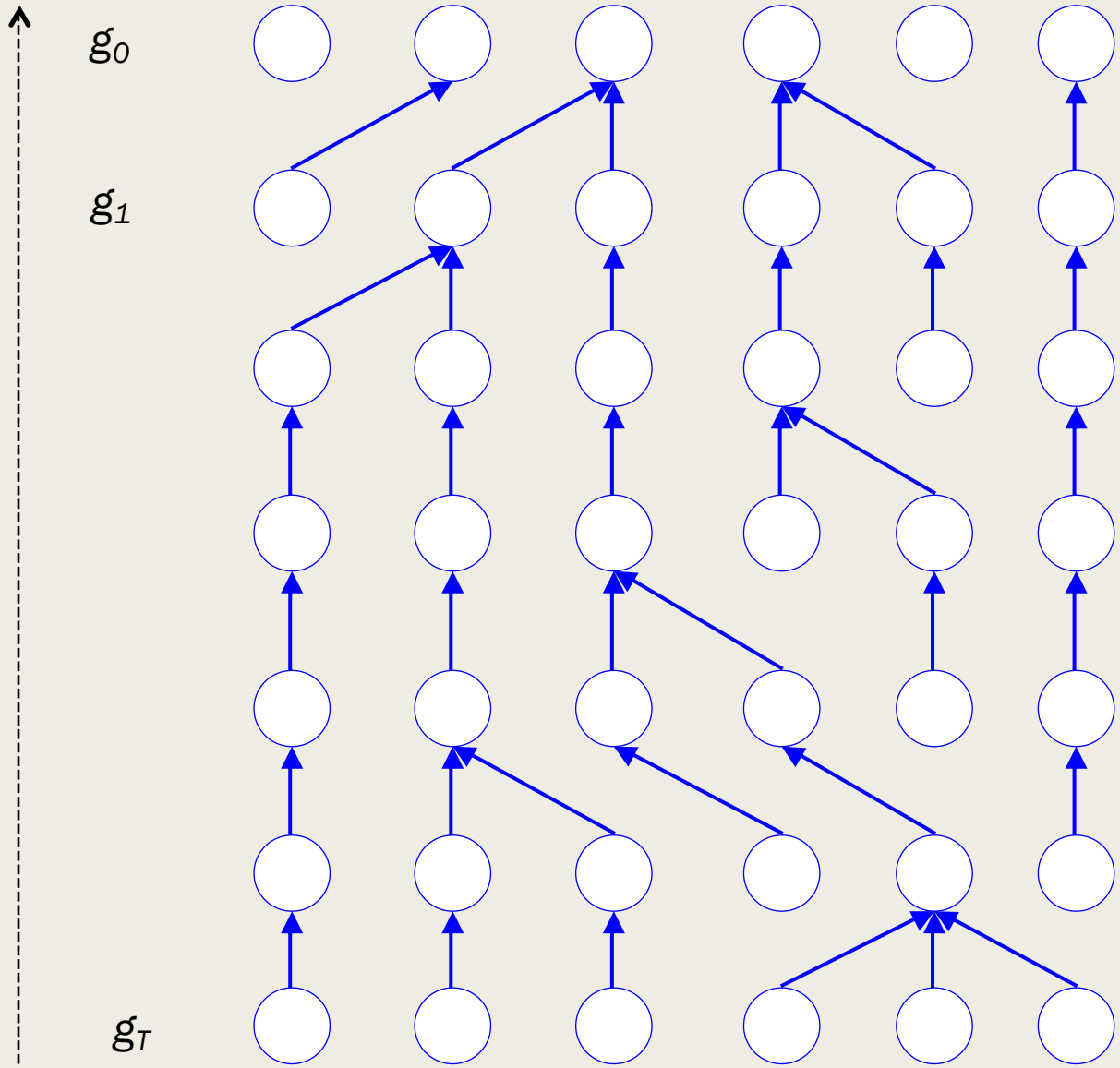


n = 2


natural selection
Goal
branch lengths
coalesce
T_2
n=2
Handout 18
page 2

Wright-Fisher Model

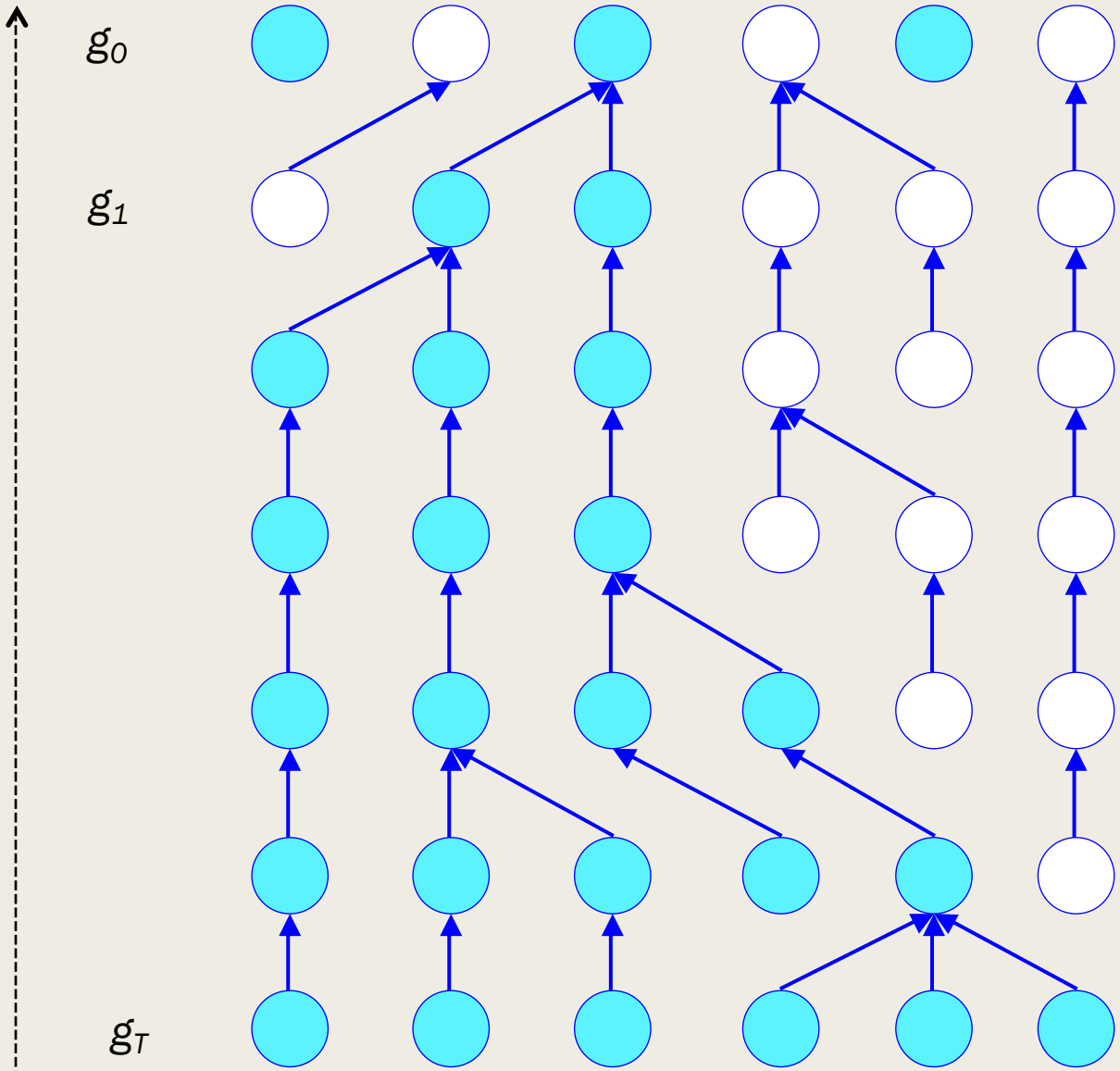
- Imagine each child choosing their parent at random
- When two descendants choose the same parent, they “coalesce”
- From then on, they have the same ancestry and follow the same lineage



Constant population size: 2N

Wright-Fisher Model

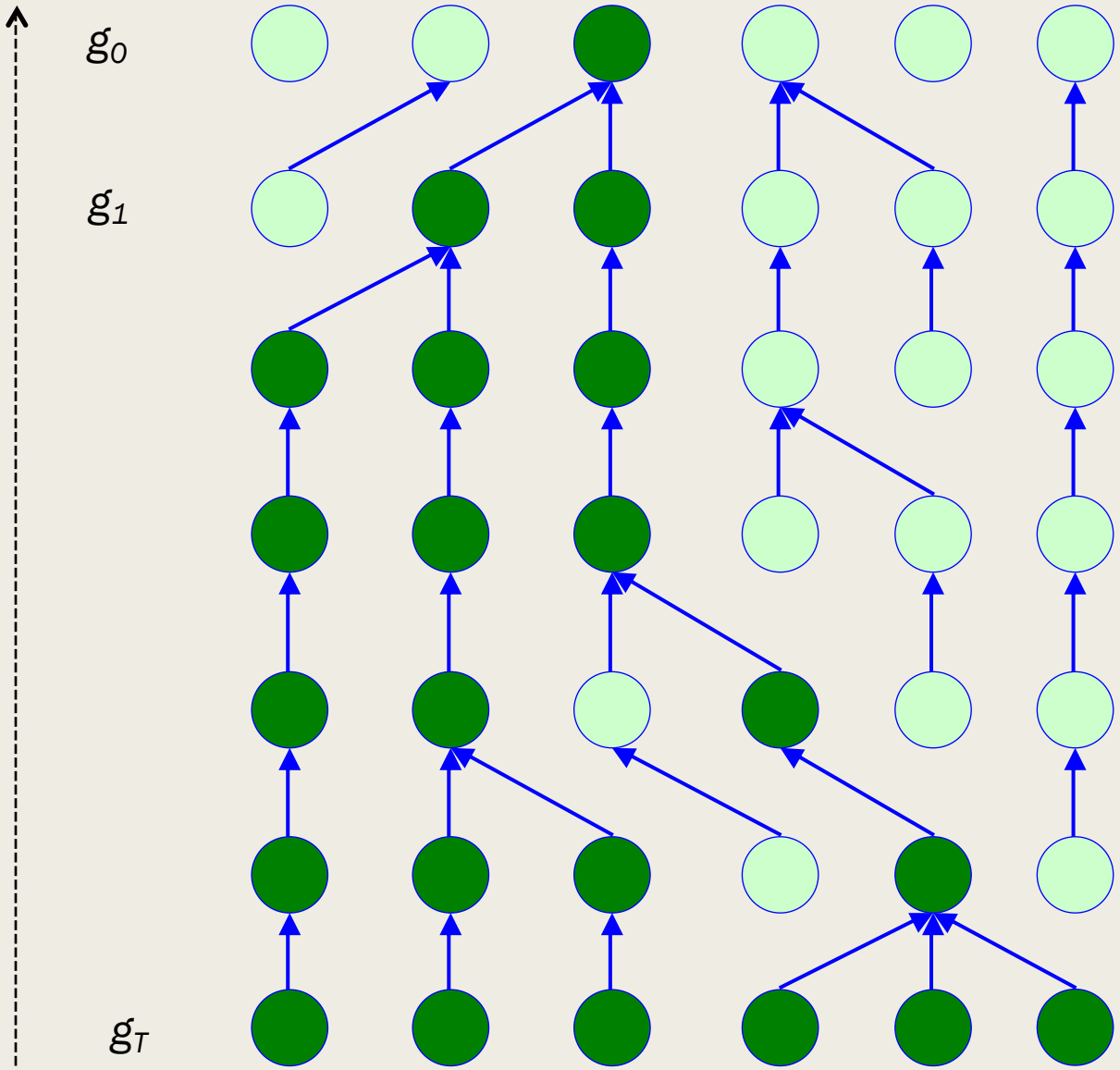
- First: model variation but not any new mutations
- Blue is the “A” allele
- White is the “a” allele



Constant population size: 2N

Wright-Fisher Model

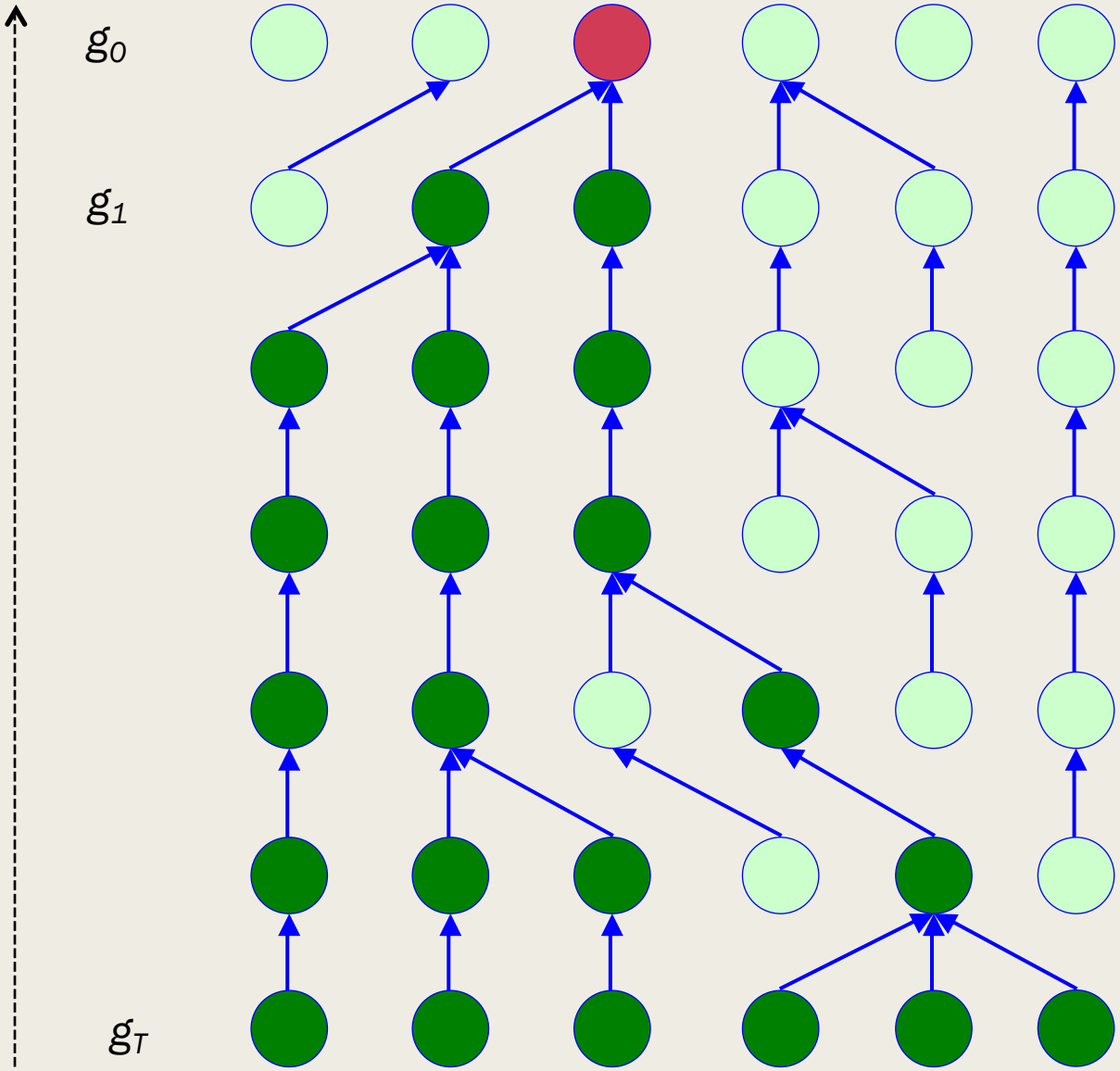
- Viewed another way, track which individuals pass on genetic material that is observable at the present
- Dark green: contributes to genetic material at present
- Light green: does not contribute genetic material to the present



Constant population size: 2N

Wright-Fisher Model

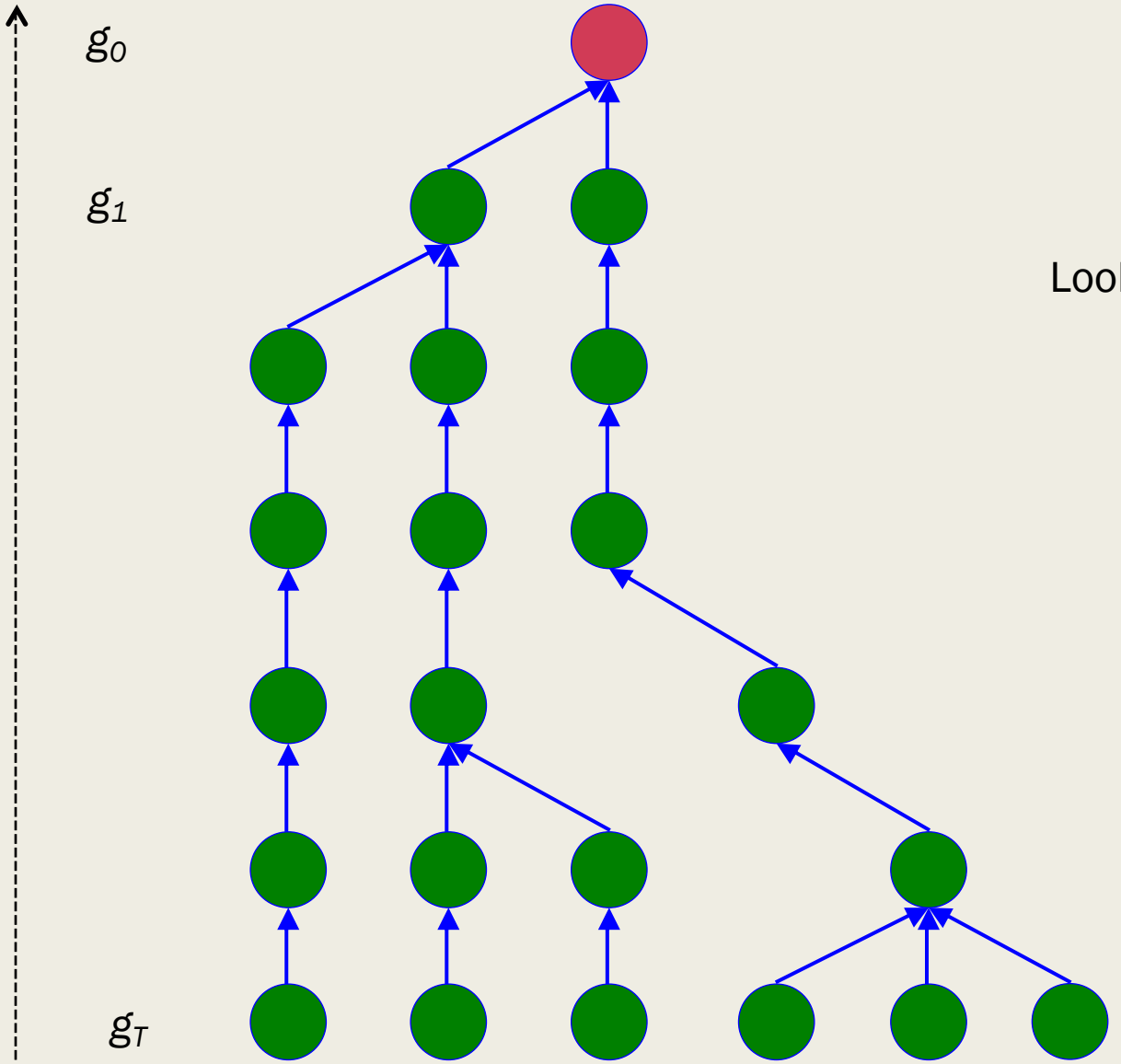
- Eventually, all the present-day individuals will “coalesce” and share one common ancestor
- Common ancestor in red



Constant population size: 2N

Wright-Fisher Model

- Eventually, all the present-day individuals will “coalesce” and share one common ancestor
- Common ancestor in red



Looks like a tree!

Constant population size: 2N

Wright-Fisher model

- Wright-Fisher model of evolution; discrete time (measured in generations)

- Assumptions (for now):
  - *constant population size*
  - *random mating*
  - *the two chromosomes for each individual choose their parents independently*
  - *mutations are neutral (i.e. not selectively advantageous or deleterious)*

- Genetic drift: changes in allele frequencies are due to random chance, not selection

Wright-Fisher model

- All neutral genetic variation will eventually die out or become fixed in the population **(question: so why do we observe variation?)**

Intermediate frequencies can persist for many generations, selection, admixture, any deviations from neutrality

- The probability of **fixation** for a new mutation is $1/(2N)$ where N is the population size

- In general the fixation probability is f_0, the initial frequency of the mutation in generation 0

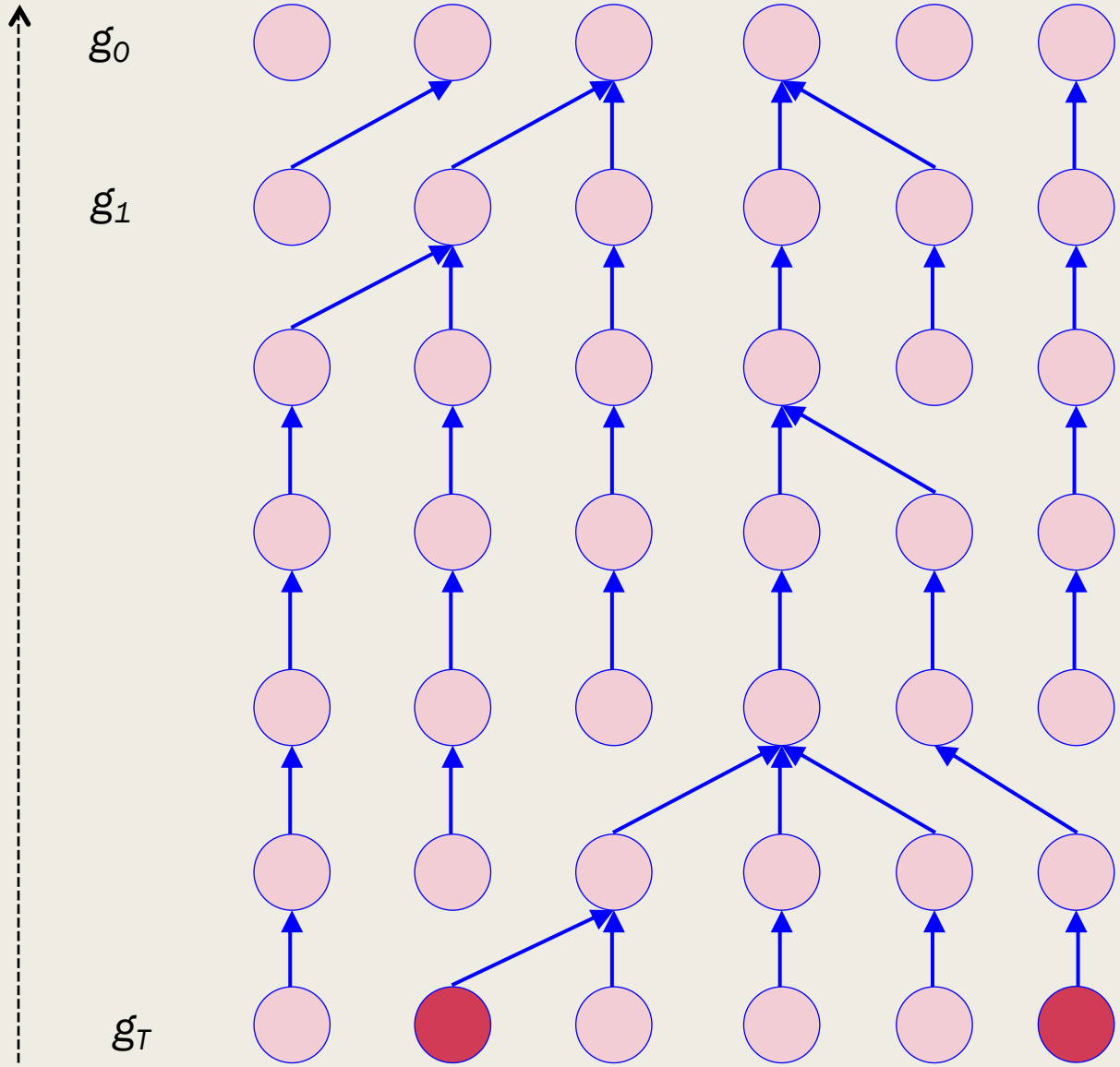
- **Question: how is genetic drift affected by the population size N? What consequences might this affect have?**

The lower the population size, the greater the chance new mutations will fix, even weakly deleterious ones. This can lead to what would typically be rare traits reaching high frequency.

Wright-Fisher Model

Generations back in time

Question: how long will it take two randomly chosen individuals to coalesce (i.e. find a common ancestor?)



Constant population size: 2N

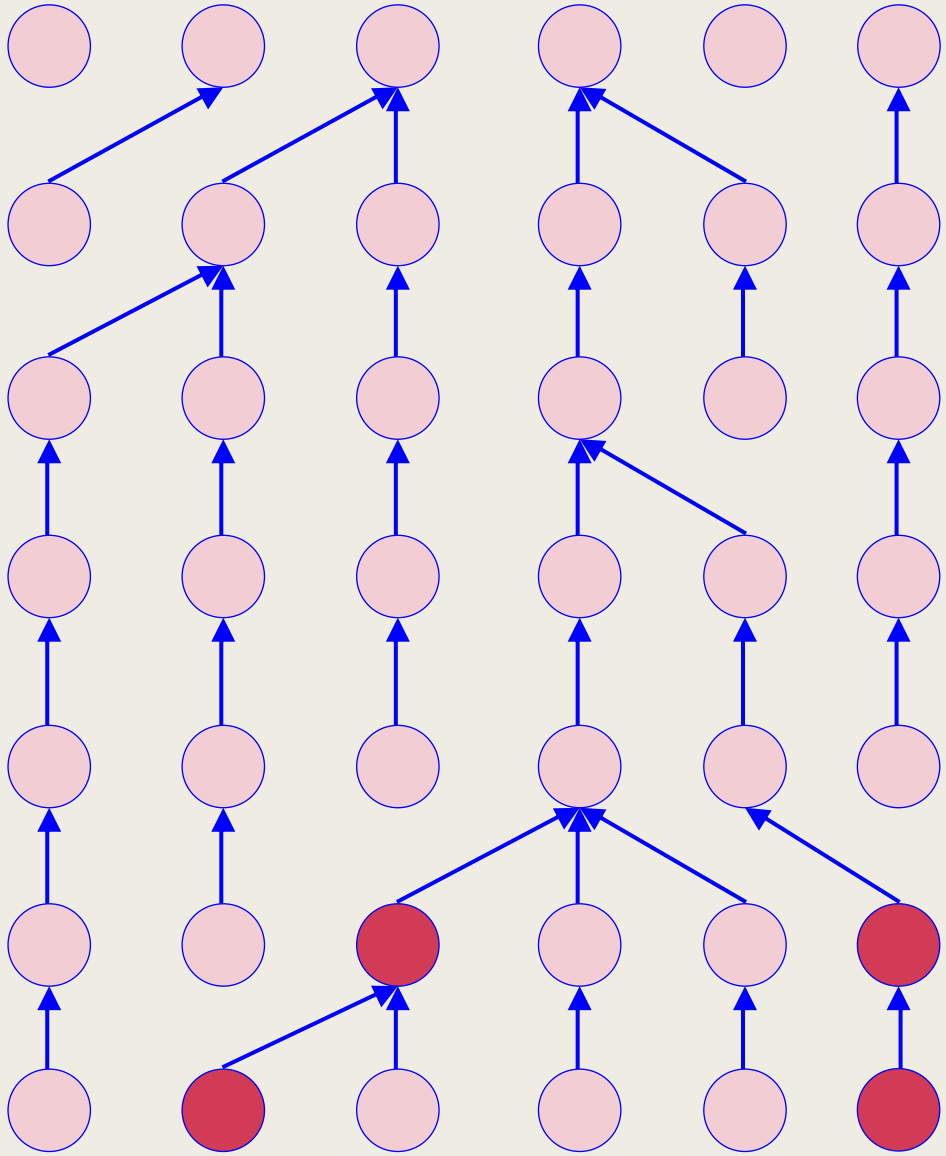
Wright-Fisher Model

Generations back in time

g_0

g_1

g_T



Question: how long will it take two randomly chosen individuals to coalesce (i.e. find a common ancestor?)

Probability they don't choose the same parent in the previous generation: 1-1/(2N)

Constant population size: 2N

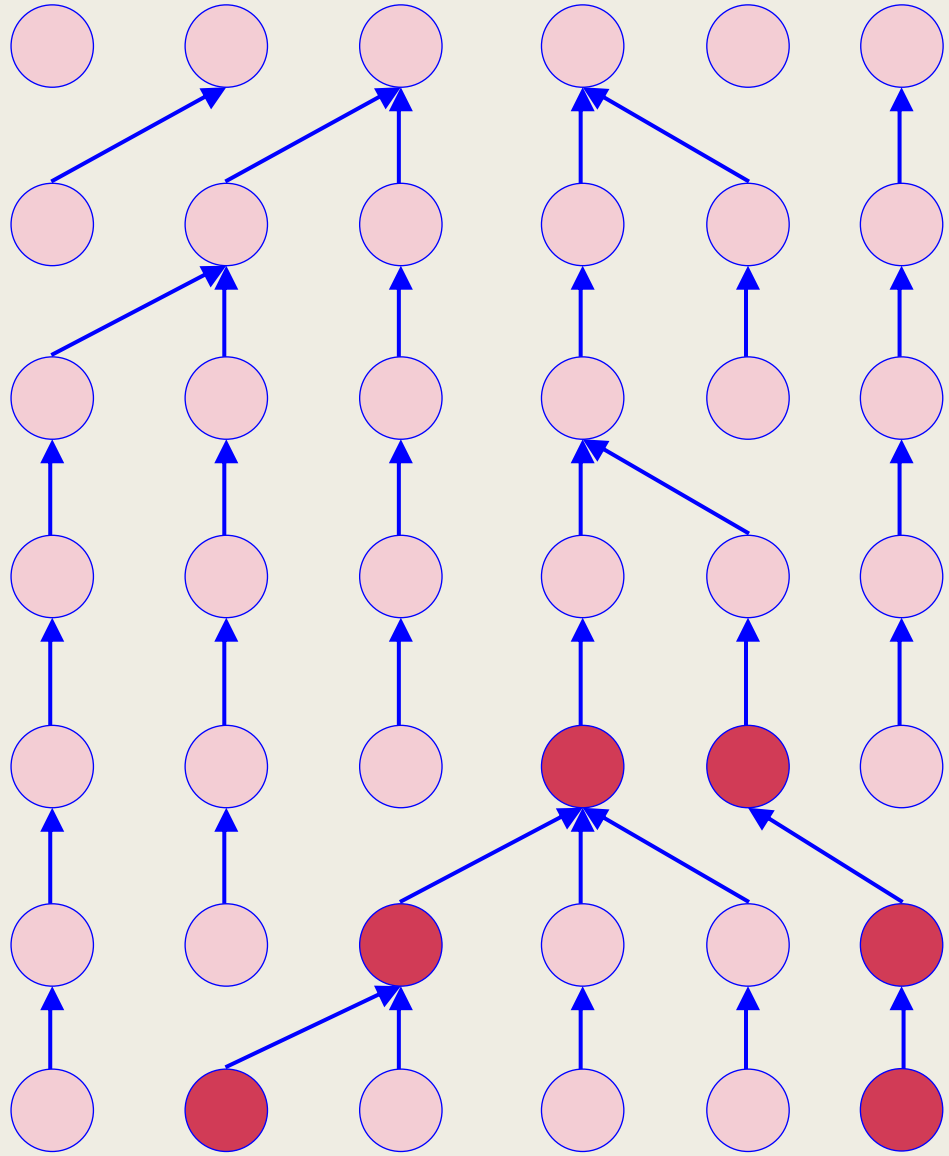
Wright-Fisher Model

Generations back in time

g_0

g_1

g_T



Question: how long will it take two randomly chosen individuals to coalesce (i.e. find a common ancestor?)

Probability they don't choose the same parent in the previous generation: 1-1/(2N)

Constant population size: 2N

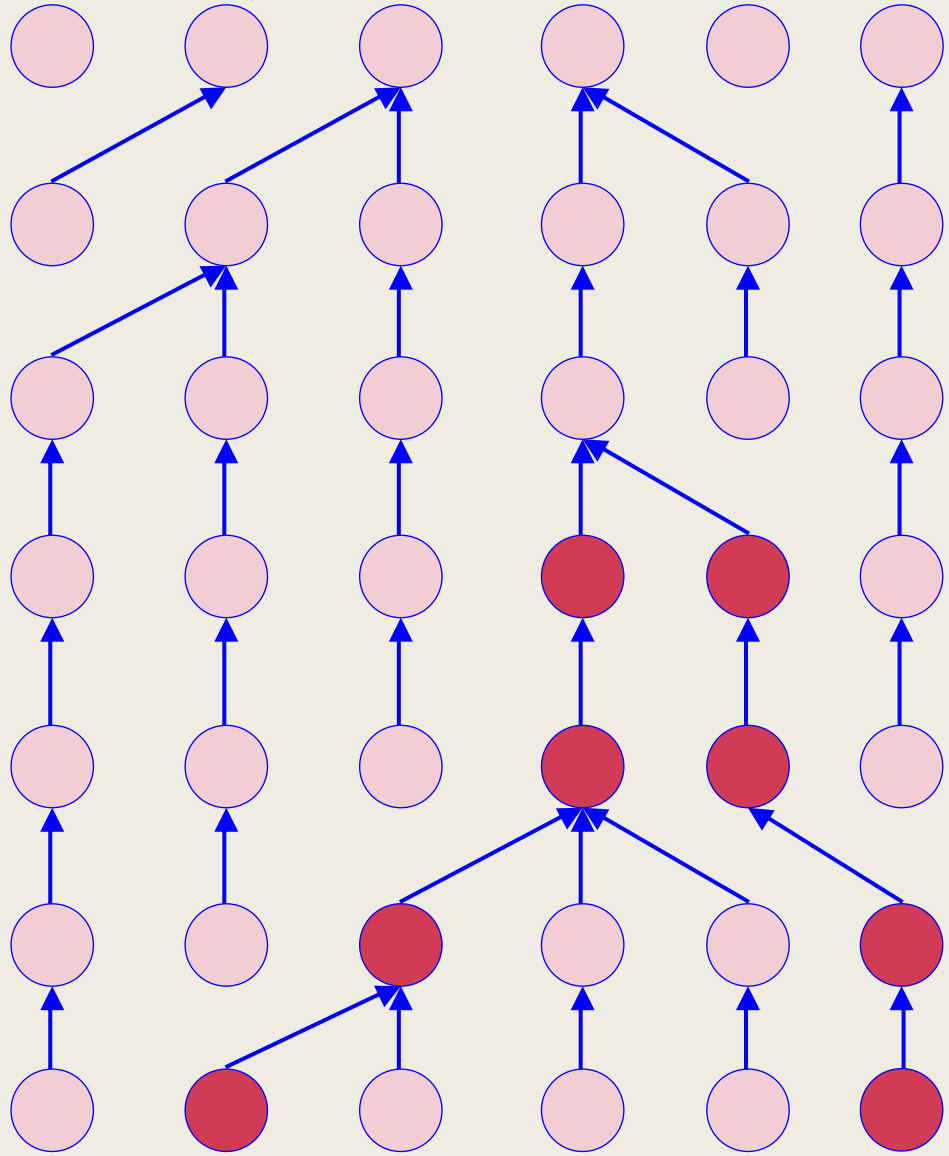
Wright-Fisher Model

Generations back in time

g_0

g_1

g_T



Question: how long will it take two randomly chosen individuals to coalesce (i.e. find a common ancestor?)

Probability they don't choose the same parent in the previous generation: 1-1/(2N)

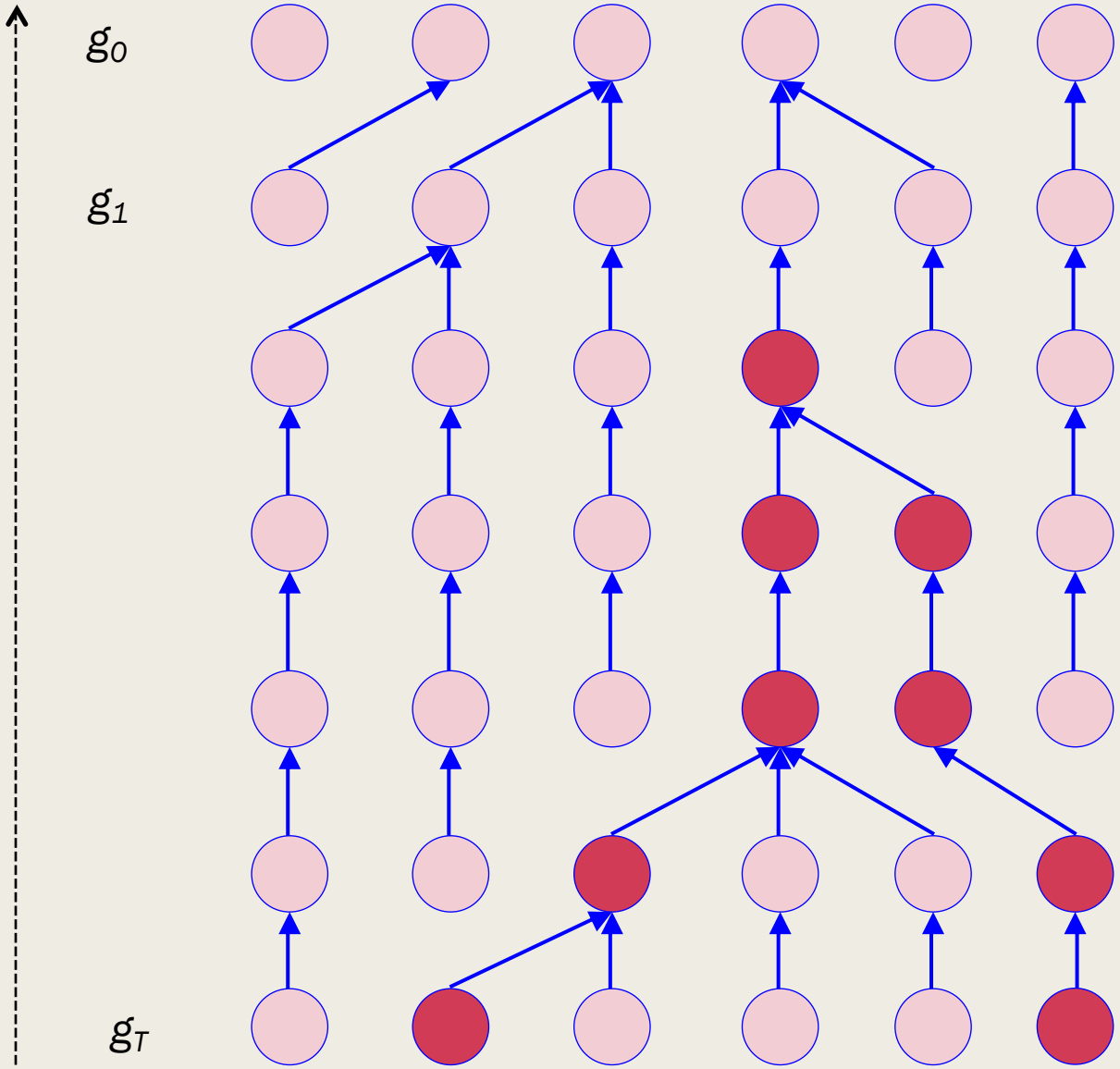
Constant population size: 2N

Wright-Fisher Model

Generations back in time

Question: how long will it take two randomly chosen individuals to coalesce (i.e. find a common ancestor?)

Probability they do choose the same parent: 1/(2N)



Constant population size: 2N

Coalescent Theory

- The Coalescent (usually attributed to Kingman, 1982) is a mathematical model for the evolution and genealogical history of a population
- The Coalescent can be derived from the Wright-Fisher model, but also several other discrete-time models (i.e. the Moran model)
- We assume the population size N is large
- We rescale time where 1 unit in coalescent time = $2N$ generations
- Rescaling time allows us to work with numbers that are on order 1 (avoiding numerical issues that arise with very small numbers) and we also avoid a factor of $2N$ in every formula

Coalescent derivation from the Wright-Fisher model

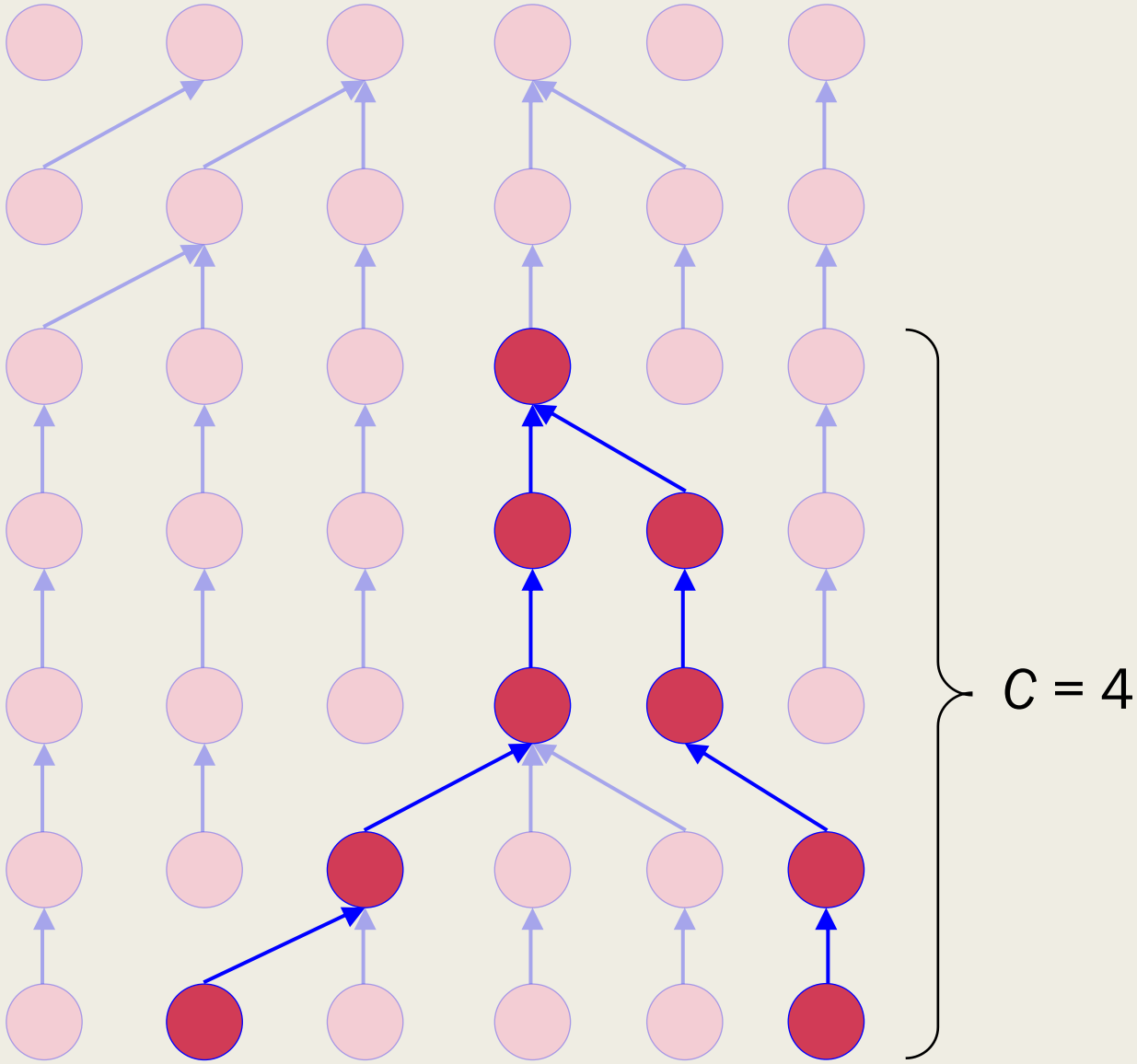
Probability two samples *coalesce* after *g* generations:

$$P_C(g) = \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N}$$

Don't choose the same parent for *g*-1 generations

Choose same parent in the g^{th} generation

[Geometric distribution]



Population size $2N$=6, sample size n = 2

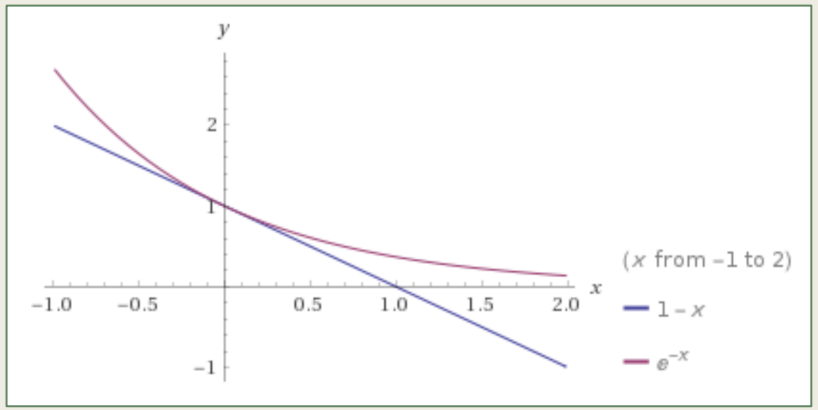
Coalescent derivation from the Wright-Fisher model

- We will make use of the Taylor series for e^{-x} around $x = 0$:

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

- We will only use the first 2 terms:

$$e^{-x} \approx 1 - x$$



Created using WolframAlpha

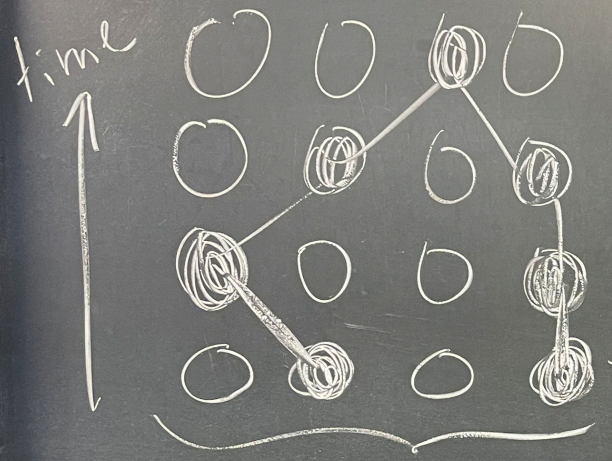
- This allows us to rewrite our geometric coalescent probability

$$P_C(g) = \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N}$$

- as (drop the -1 since g is large):

$$P_C(g) \approx \frac{1}{2N} e^{-\frac{g}{2N}}$$

step back to WF model



3 generations to find a common ancestor

$n = 2$

$2N$

↑

individuals in population

haplotypes

large pop size

$N \to \infty$

$$p(t) = xe^{-xt}$$

↑

time (continuous)

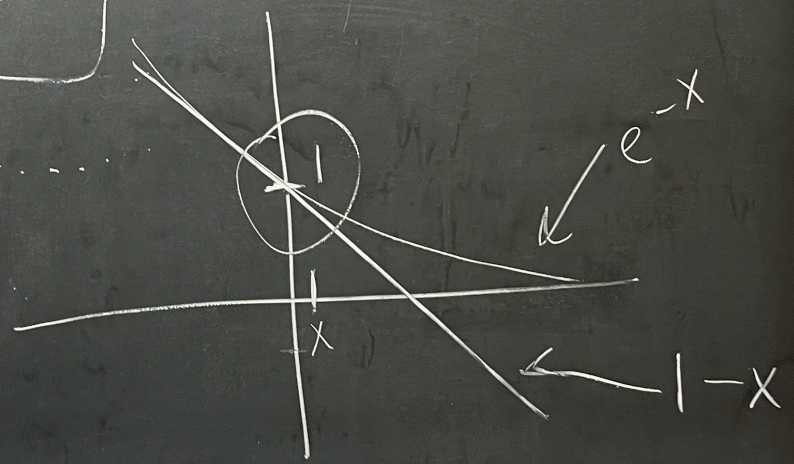
prob they do choose same parent? $\frac{1}{2N}$

$$P_c(g) = \left(1 - \frac{1}{2N}\right)^{g-1} \frac{1}{2N}$$

geometric distribution

$P_c(g)$: coalesce; g: generations; $\left(1 - \frac{1}{2N}\right)^{g-1}$: don't choose same parent; $\frac{1}{2N}$: do choose same parent

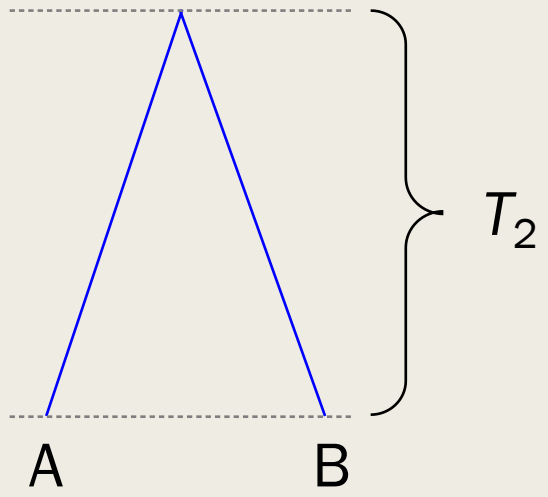
$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} \cdots$$



Coalescent for $n = 2$

- We let 1 coalescent unit = $2N$ generations, and let our new variable be t

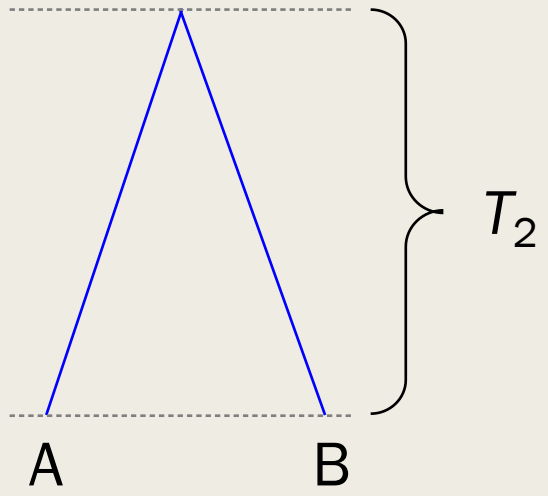
- We let T_i be a random variable representing the time when there are i lineages



Coalescent for $n = 2$

- We let 1 coalescent unit = $2N$ generations, and let our new variable be t
- We let T_i be a random variable representing the time when there are i lineages
- For n=2, this gives us an exponential distribution with parameter 1

$$P_{T_2}(t) = e^{-t}$$

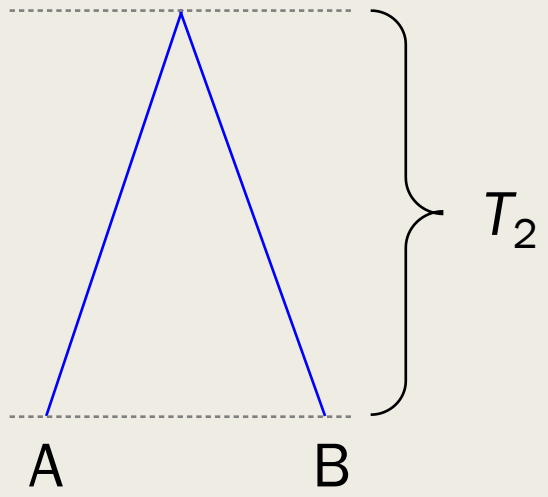


Coalescent for *n* = 2

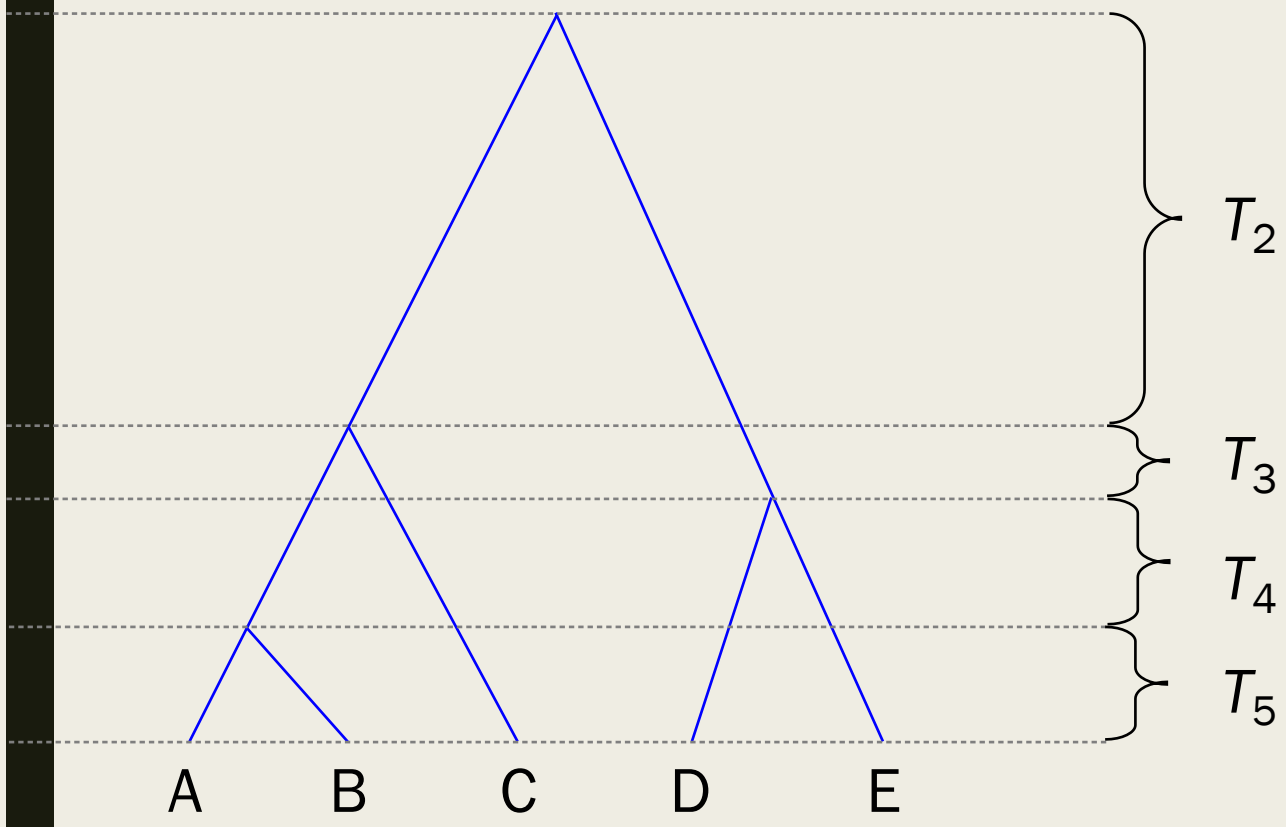
- We let 1 coalescent unit = 2*N* generations, and let our new variable be *t*
- We let T_i be a random variable representing the time when there are *i* lineages
- For *n*=2, this gives us an exponential distribution with parameter 1
- The expected time for 2 lineages to coalesce is 1 coalescent unit of time => 2N generations

$$P_{T_2}(t) = e^{-t}$$

$$E[T_2] = \int_0^\infty te^{-t}dt = 1$$



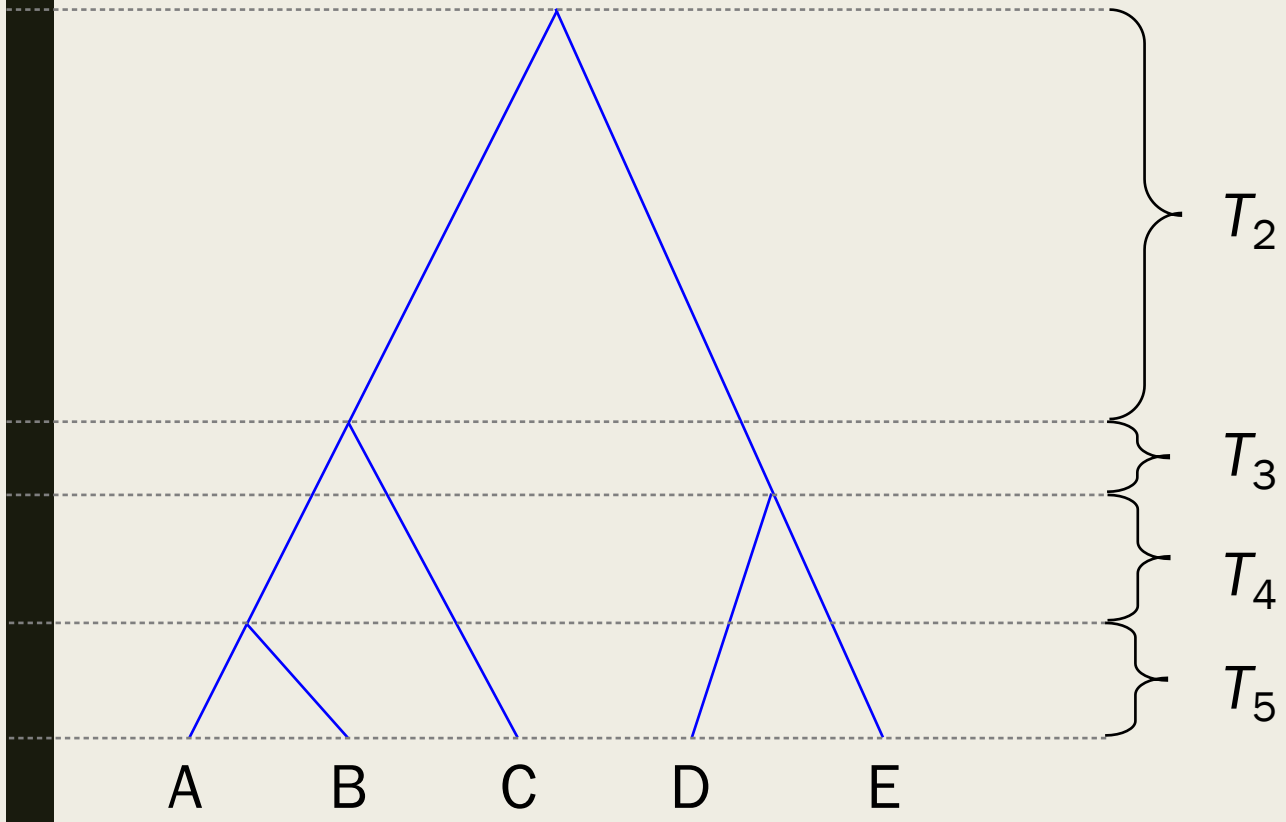
The Coalescent



- The larger our sample size n, the more pairs we have that can coalesce right away
- In general, the time when there are *i* lineages is also exponentially distributed with parameter *i*(*i*-1)/2 (*i* "choose" 2)

$$P_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2}t}$$

The Coalescent



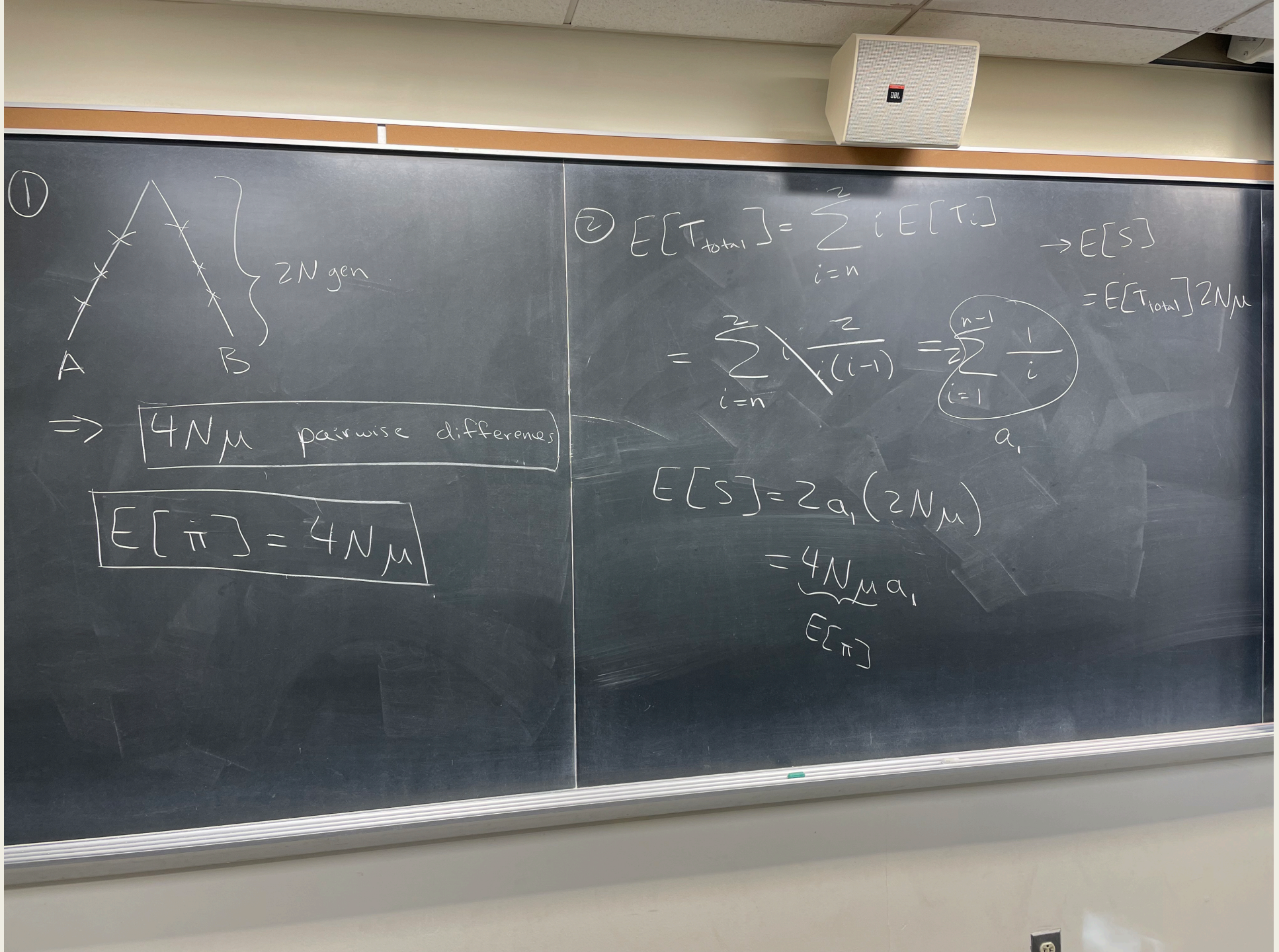
- The larger our sample size n, the more pairs we have that can coalesce right away
- In general, the time when there are *i* lineages is also exponentially distributed with parameter *i*(*i*-1)/2 (*i* "choose" 2)

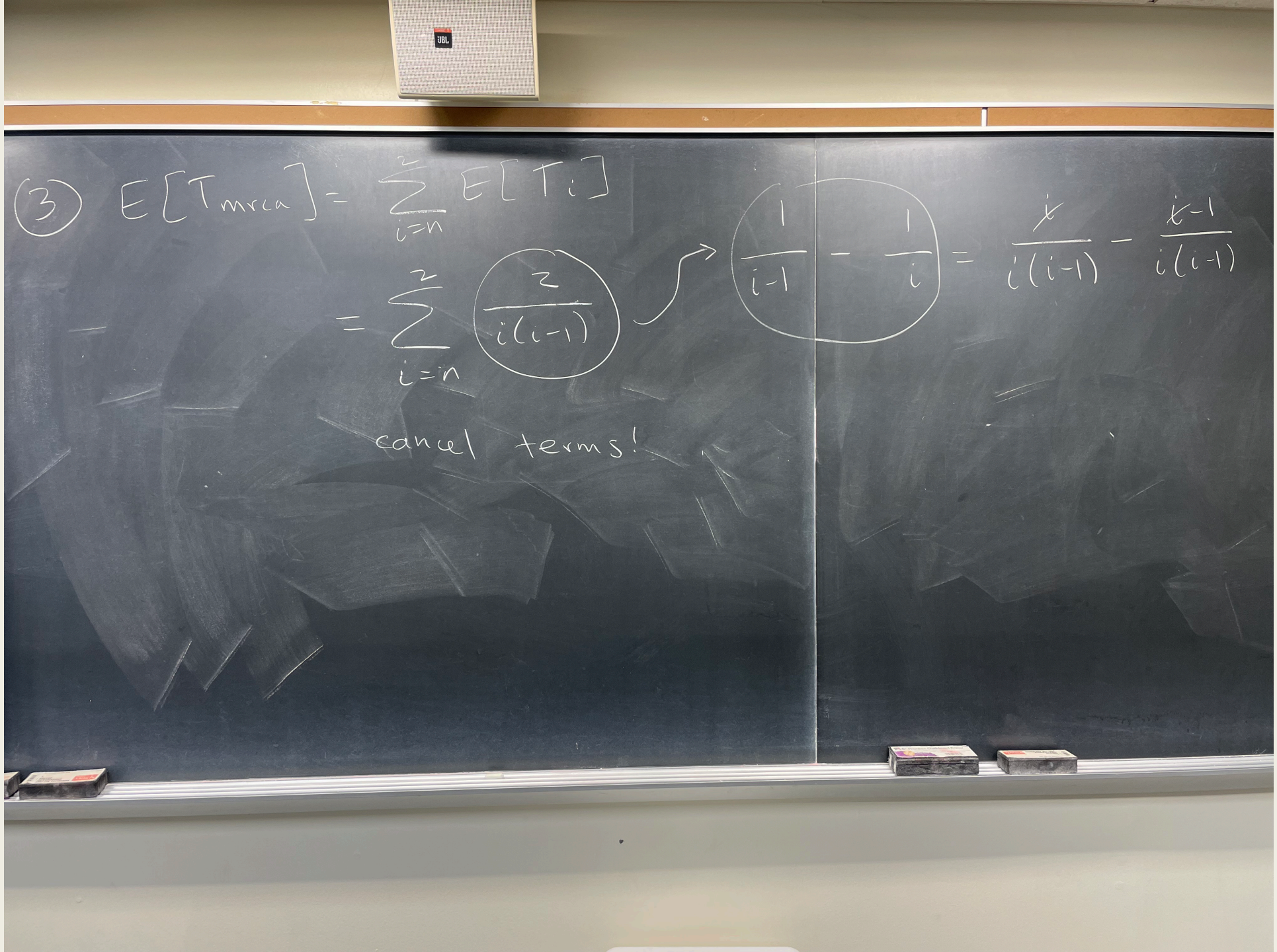
$$P_{T_i}(t) = \binom{i}{2} e^{-\binom{i}{2}t}$$

- Expected value (think: weighted average, mean)

$$E[T_i] = \int_0^\infty t \binom{i}{2} e^{-\binom{i}{2}t} dt = \frac{1}{\binom{i}{2}}$$

Handout 18, pg 2


①
A
B
2N gen
⇒ 4Nμ pairwise differences
E[π] = 4Nμ
② E[T_total] = Σ_{i=n}^{2} i E[T_i]
→ E[S] = E[T_total] 2Nμ
= Σ_{i=n}^{2} i · 2/(i(i-1)) = 2 Σ_{i=1}^{n-1} 1/i
a_1
E[S] = 2a_1(2Nμ)
= 4Nμ a_1
E[π]


(3) $E[T_{mrca}] = \sum_{i=n}^{2} E[T_i]$
$= \sum_{i=n}^{2} \frac{2}{i(i-1)}$
$\frac{1}{i-1} - \frac{1}{i} = \frac{i}{i(i-1)} - \frac{i-1}{i(i-1)}$
cancel terms!