# The coalescent made fun and easy

Andy Clark and Ida Moltke

EMBO 2017 Population Genomics

Napoli – 22 May 2017

# What is the coalescent?

- A mathematical construct of the genealogical process describing a random sample of alleles from a population.
- A way to model drift and mutation.
- Only parameters are:
  - Population size (N)
  - Sample size (n)
  - Mutation rate (µ)

# What good is the coalescent?

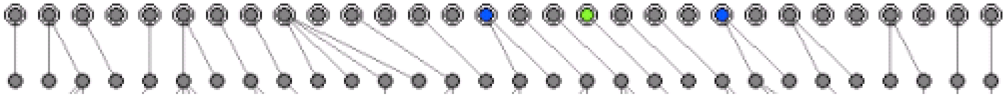
- Makes the mathematics to derive many aspects of population genetics much easier.
- Super-fast to simulate.

# Objectives

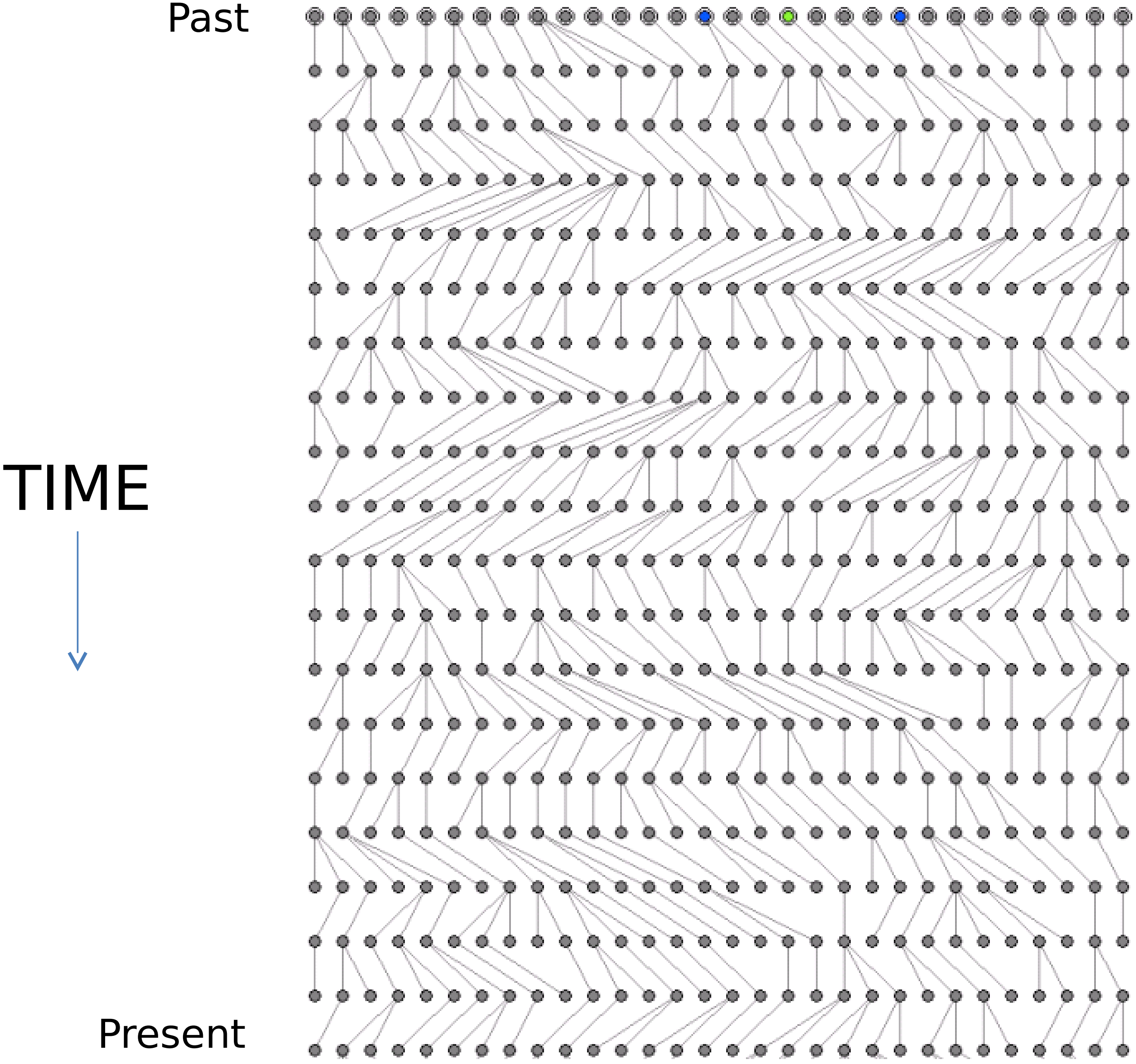
- To thoroughly understand why the coalescent is useful.
- To be able to explain it to your grandmother.
- To be able to code up a simulations, even if trapped in an elevator on a cruise ship with no internet.
- To have FUN

Past

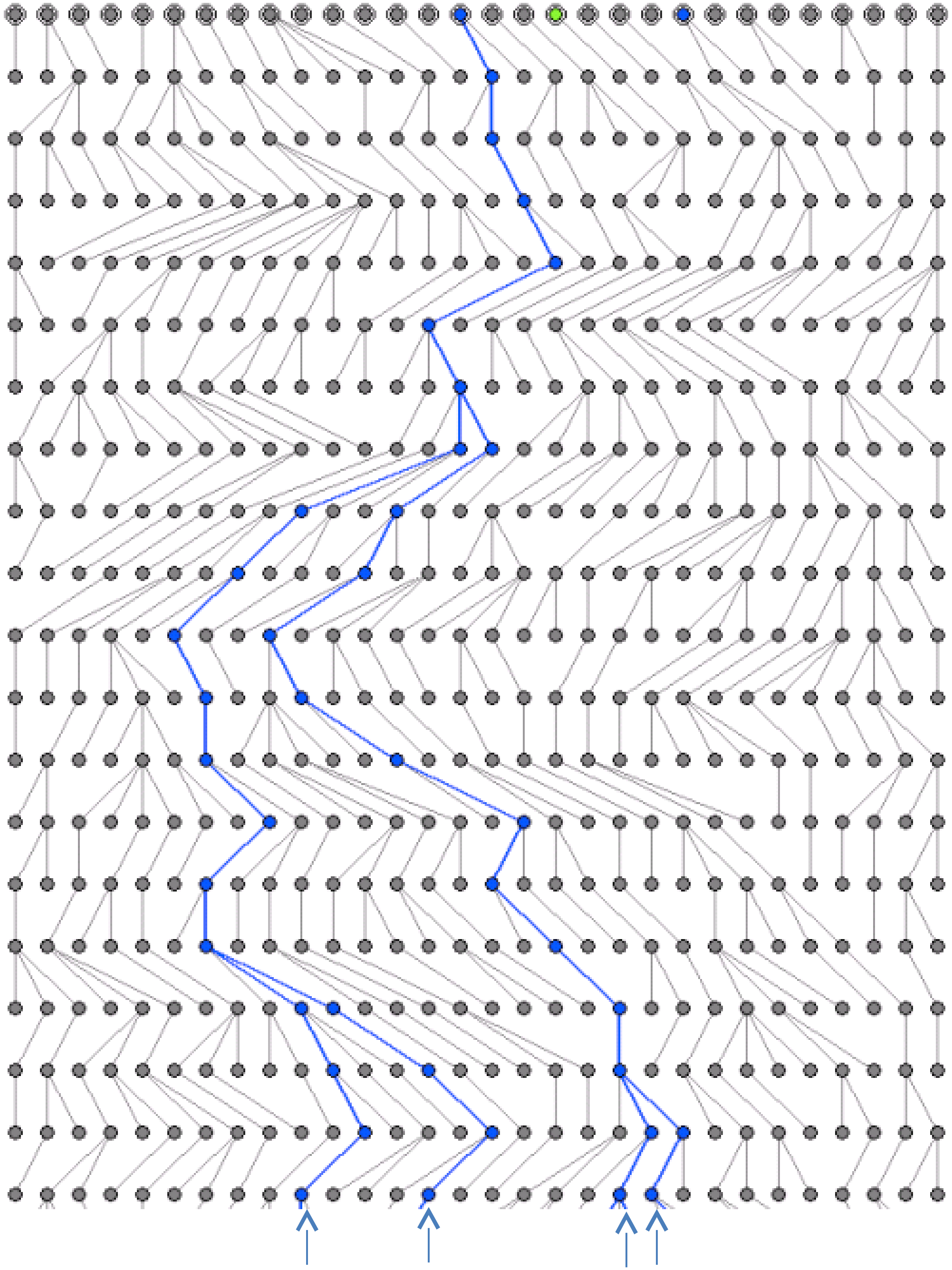
Past

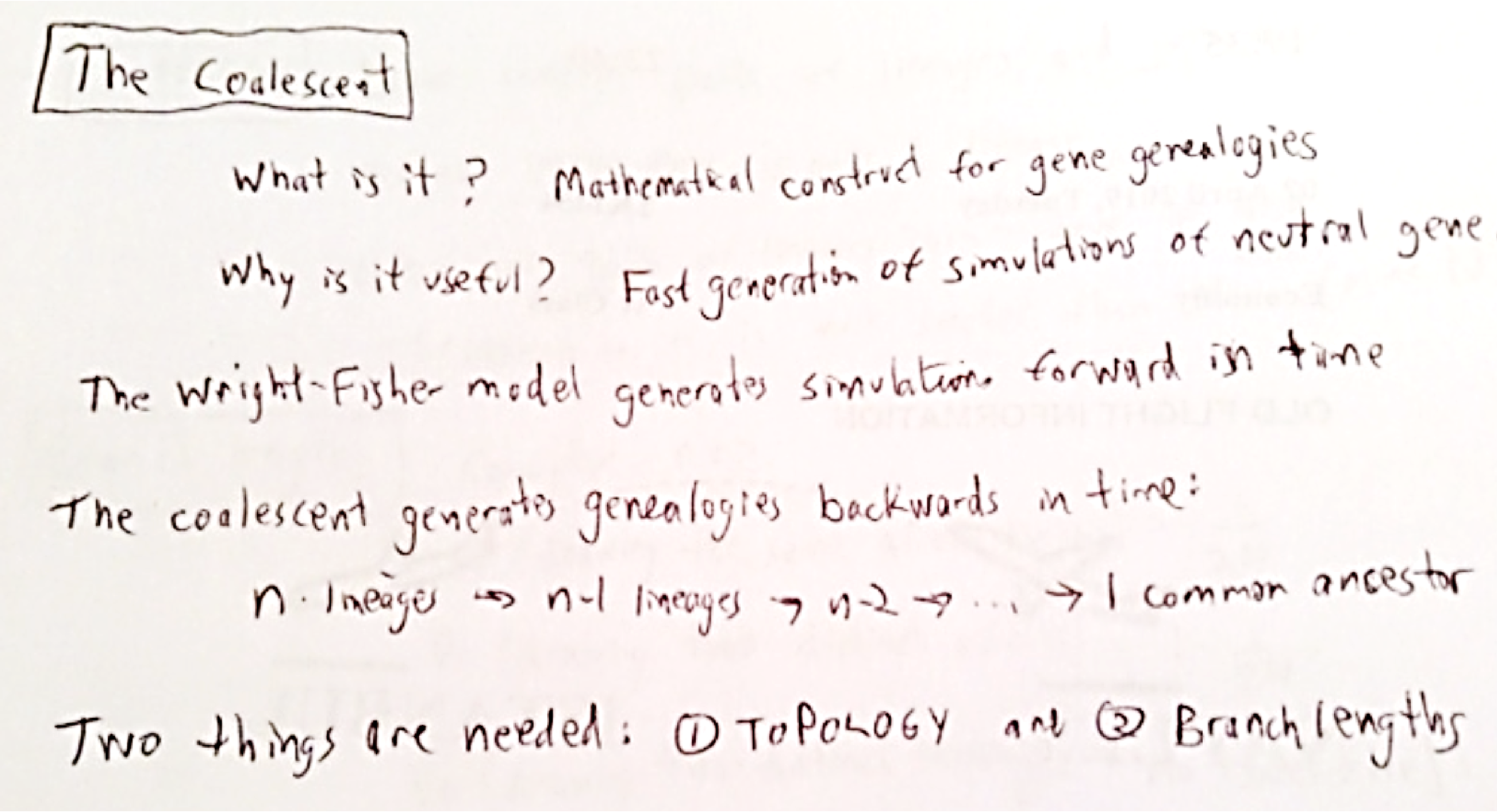


TIME


Past
TIME
Present

The coalescent considers a collection of genes observed in the present, and asks about their past genealogy.





# The Coalescent

What is it? Mathematical construct for gene genealogies

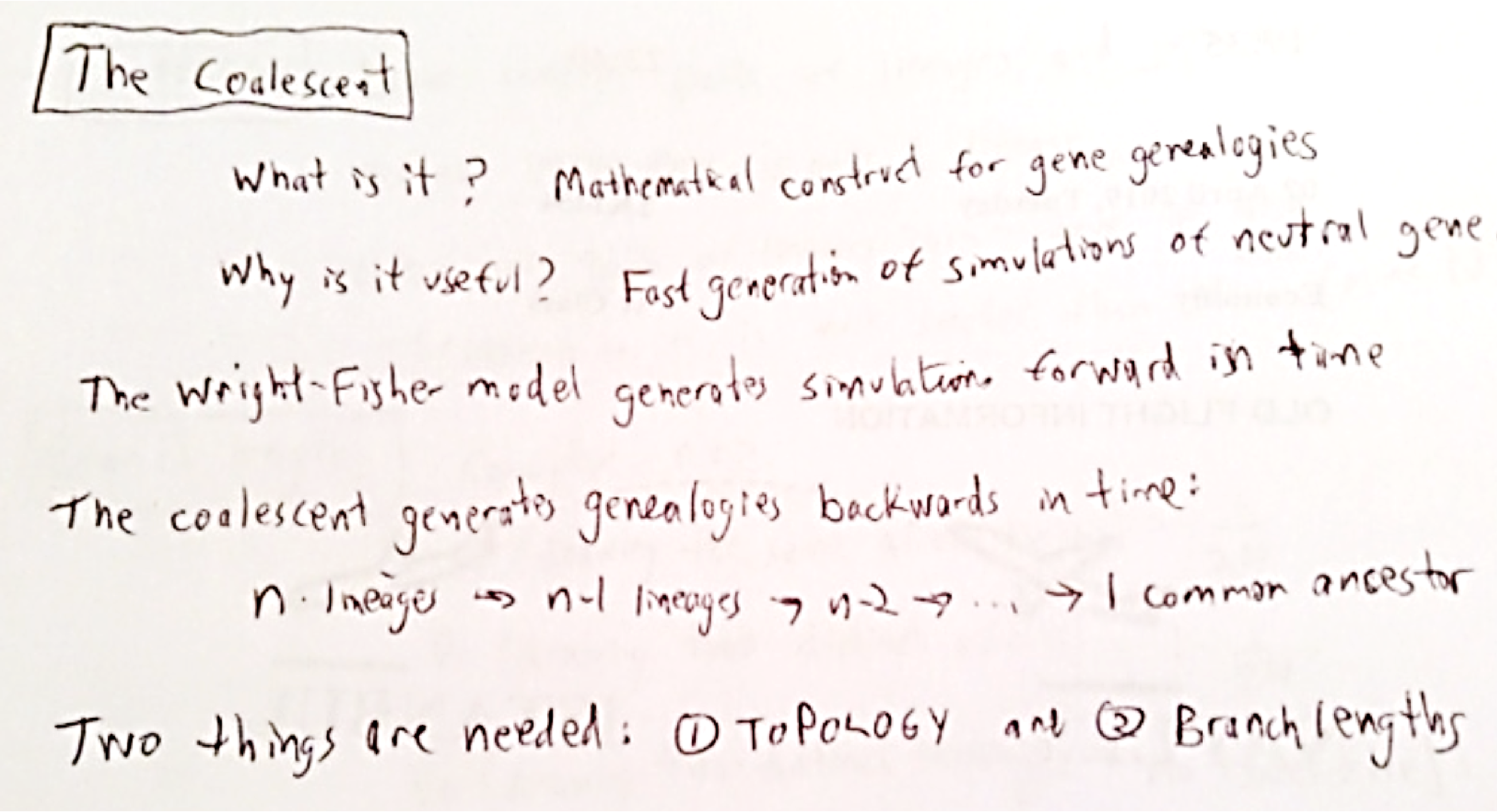
Why is it useful? Fast generation of simulations of neutral gene

The Wright-Fisher model generates simulations forward in time

The coalescent generates genealogies backwards in time:

n lineages → n-1 lineages → n-2 → ... → 1 common ancestor

Two things are needed: ① TOPOLOGY and ② Branch lengths



# The Coalescent

What is it? Mathematical construct for gene genealogies

Why is it useful? Fast generation of simulations of neutral gene

The Wright-Fisher model generates simulations forward in time

The coalescent generates genealogies backwards in time:

n · lineages → n-1 lineages → n-2 → ... → 1 common ancestor

Two things are needed: ① TOPOLOGY and ② Branch lengths

# The Coalescent

What is it? Mathematical construct for gene genealogies

Why is it useful? Fast generation of simulations of neutral gene

The Wright-Fisher model generates simulations forward in time

The coalescent generates genealogies backwards in time:

n · lineages → n-1 lineages → n-2 → ... → 1 common ancestor

Two things are needed: ① TOPOLOGY and ② Branch lengths

# The Coalescent

What is it ? Mathematical construct for gene genealogies

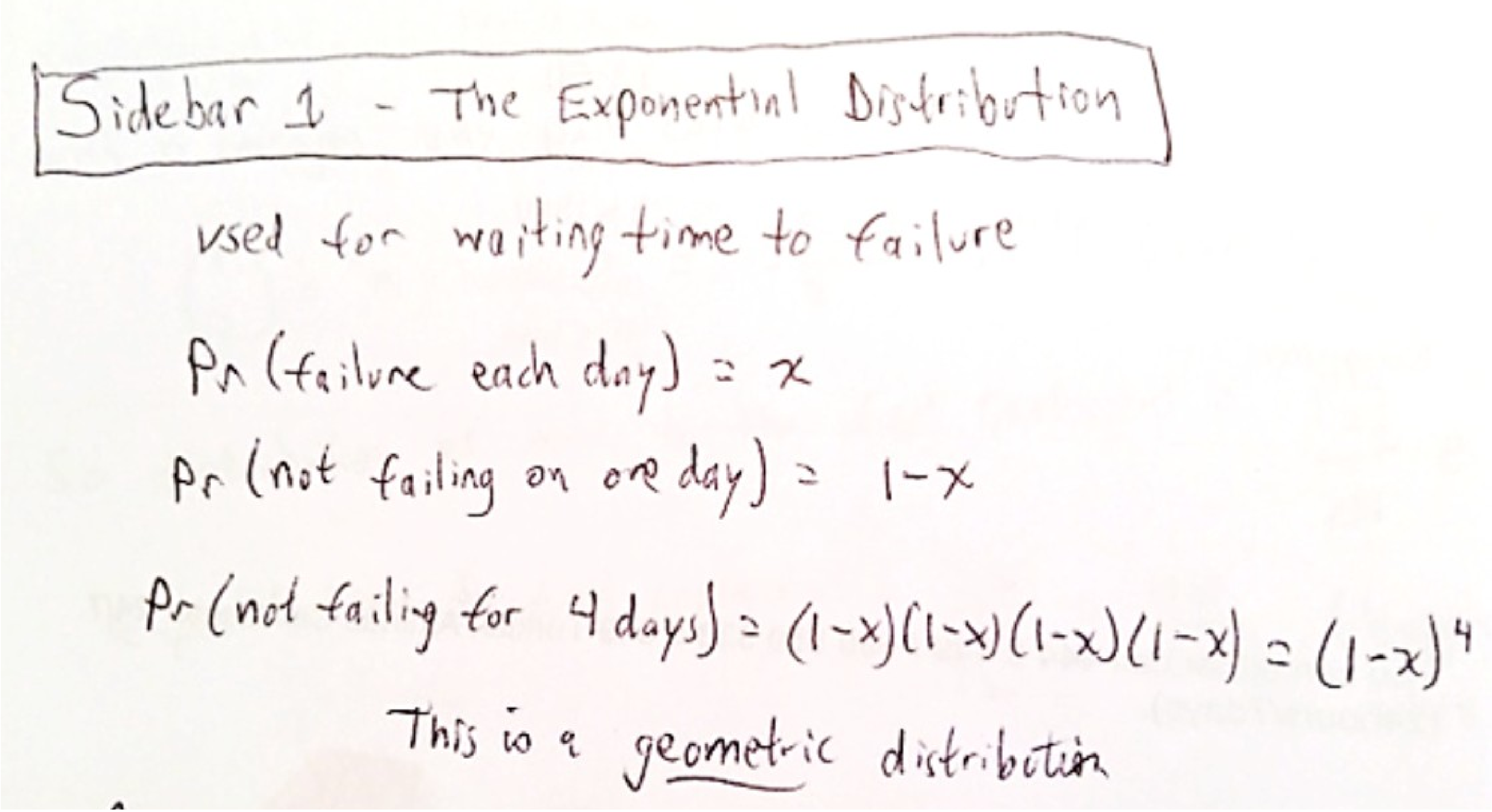
Why is it useful? Fast generation of simulations of neutral gene

The Wright-Fisher model generates simulations forward in time

The coalescent generates genealogies backwards in time:

n lineages → n-1 lineages → n-2 → ... → 1 common ancestor

Two things are needed: ① TOPOLOGY and ② Branch lengths



## Sidebar 1 - The Exponential Distribution

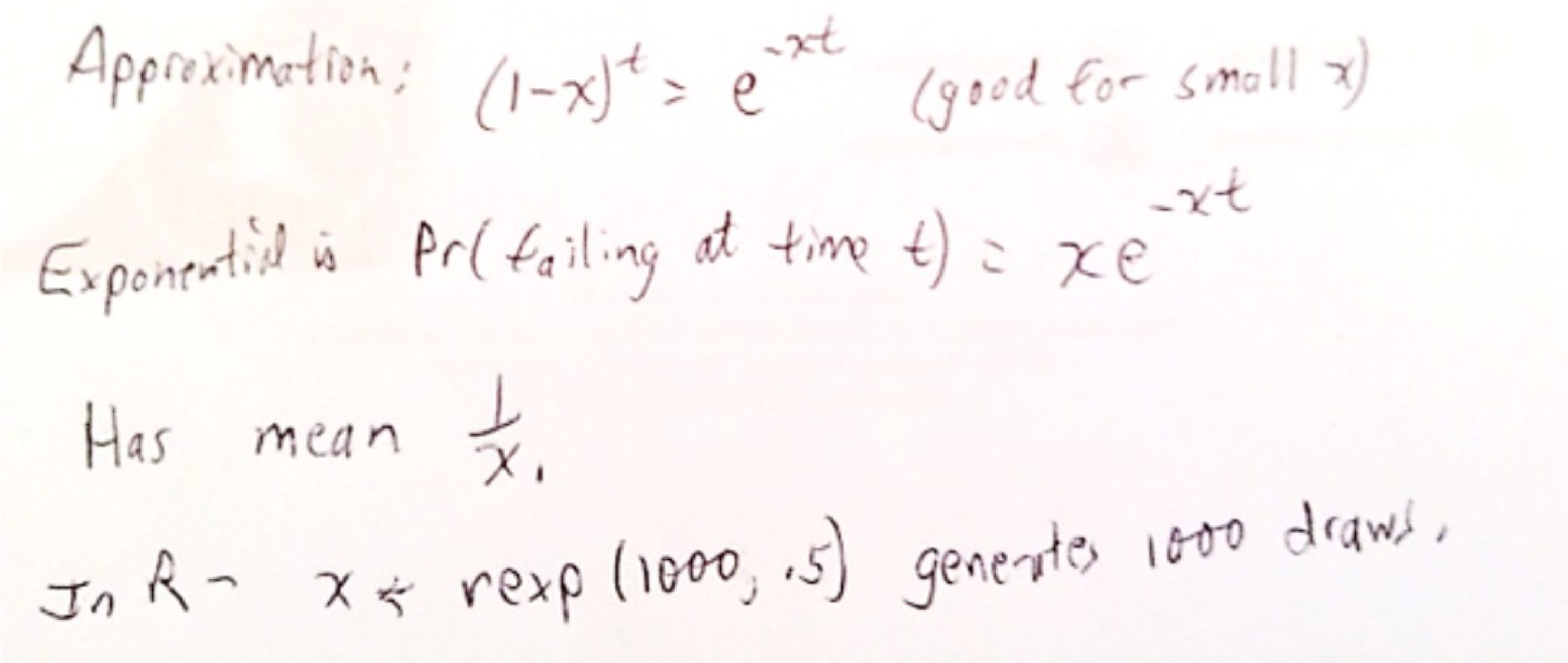
used for waiting time to failure

Pr (failure each day) = $x$

Pr (not failing on one day) = $1-x$

Pr (not failing for 4 days) = $(1-x)(1-x)(1-x)(1-x) = (1-x)^4$

This is a <u>geometric</u> distribution

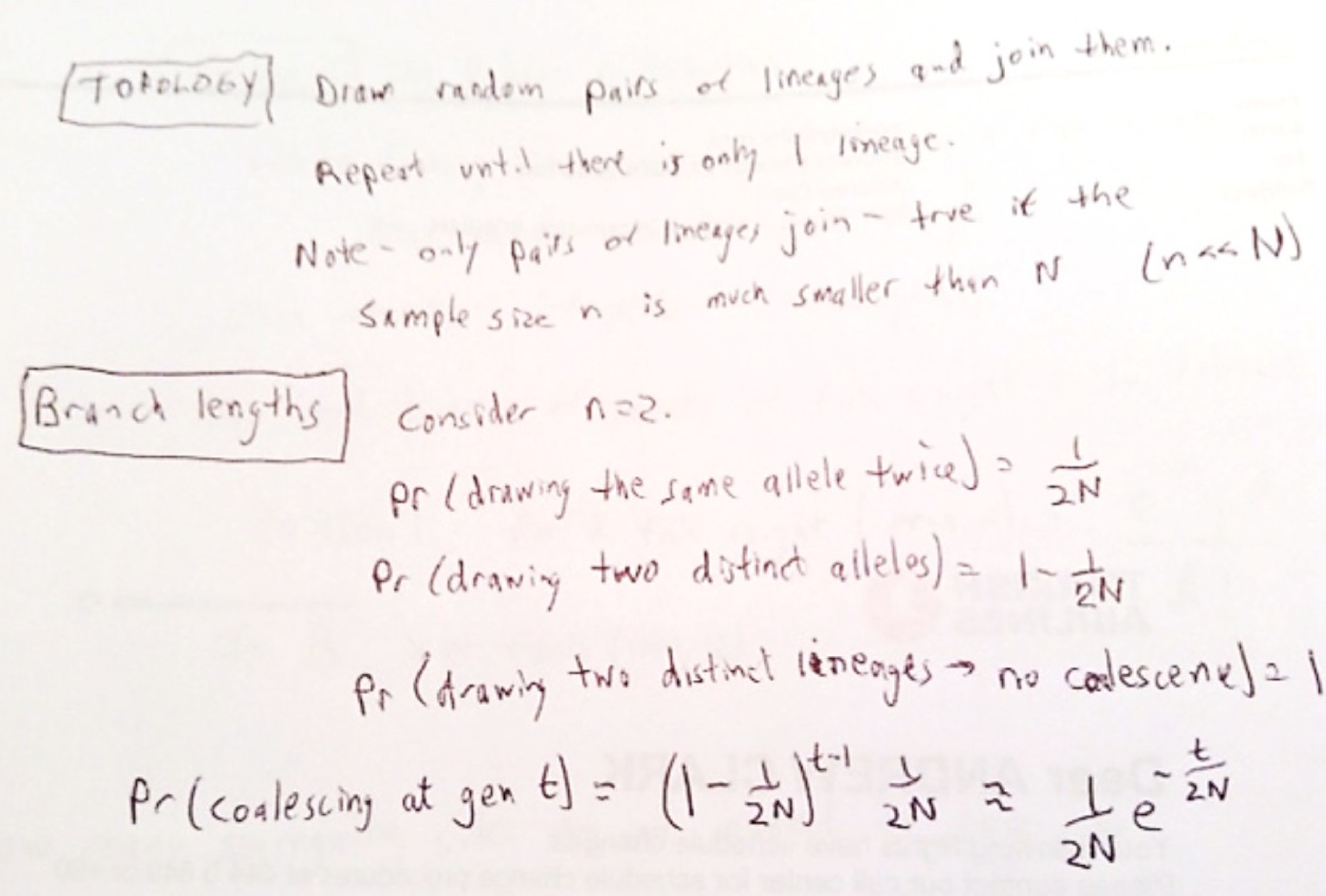


Approximation: $(1-x)^t \approx e^{-xt}$ (good for small $x$)

Exponential is $Pr(\text{failing at time } t) = xe^{-xt}$

Has mean $\frac{1}{x}$.

In R – `x <- rexp(1000, .5)` generates 1000 draws.



TOPOLOGY

Draw random pairs of lineages and join them.

Repeat until there is only 1 lineage.

Note - only pairs of lineages join - true if the sample size n is much smaller than N $(n \ll N)$

Branch lengths

Consider $n=2$.

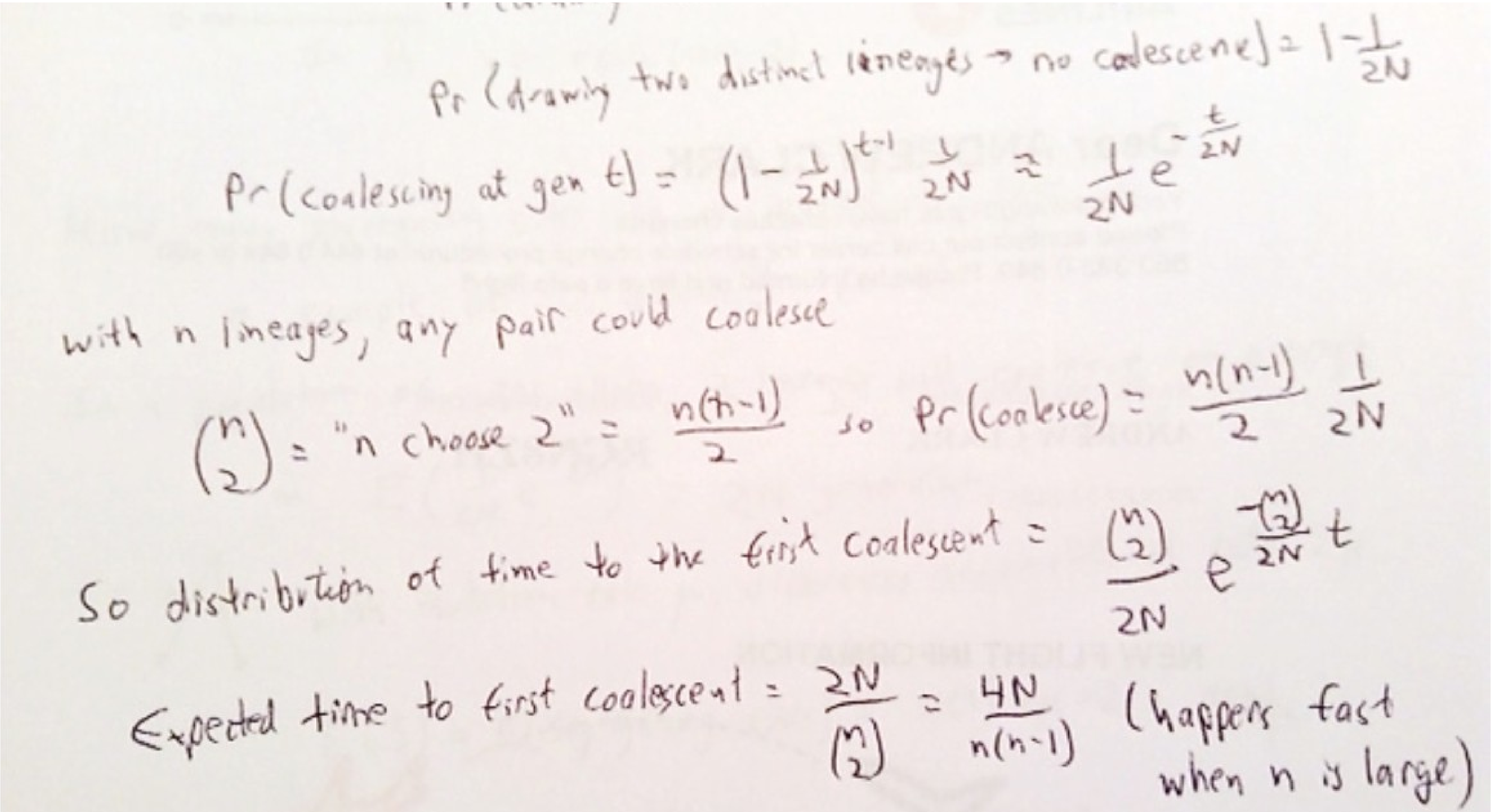
$Pr(\text{drawing the same allele twice}) = \frac{1}{2N}$

$Pr(\text{drawing two distinct alleles}) = 1 - \frac{1}{2N}$

$Pr(\text{drawing two distinct lineages} \rightarrow \text{no coalescene}) = 1$

$$Pr(\text{coalescing at gen } t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N} \approx \frac{1}{2N} e^{-\frac{t}{2N}}$$

Probability of drawing the same allele twice  1/(2N)



Pr (drawing two distinct lineages → no coalescene] = $1-\frac{1}{2N}$

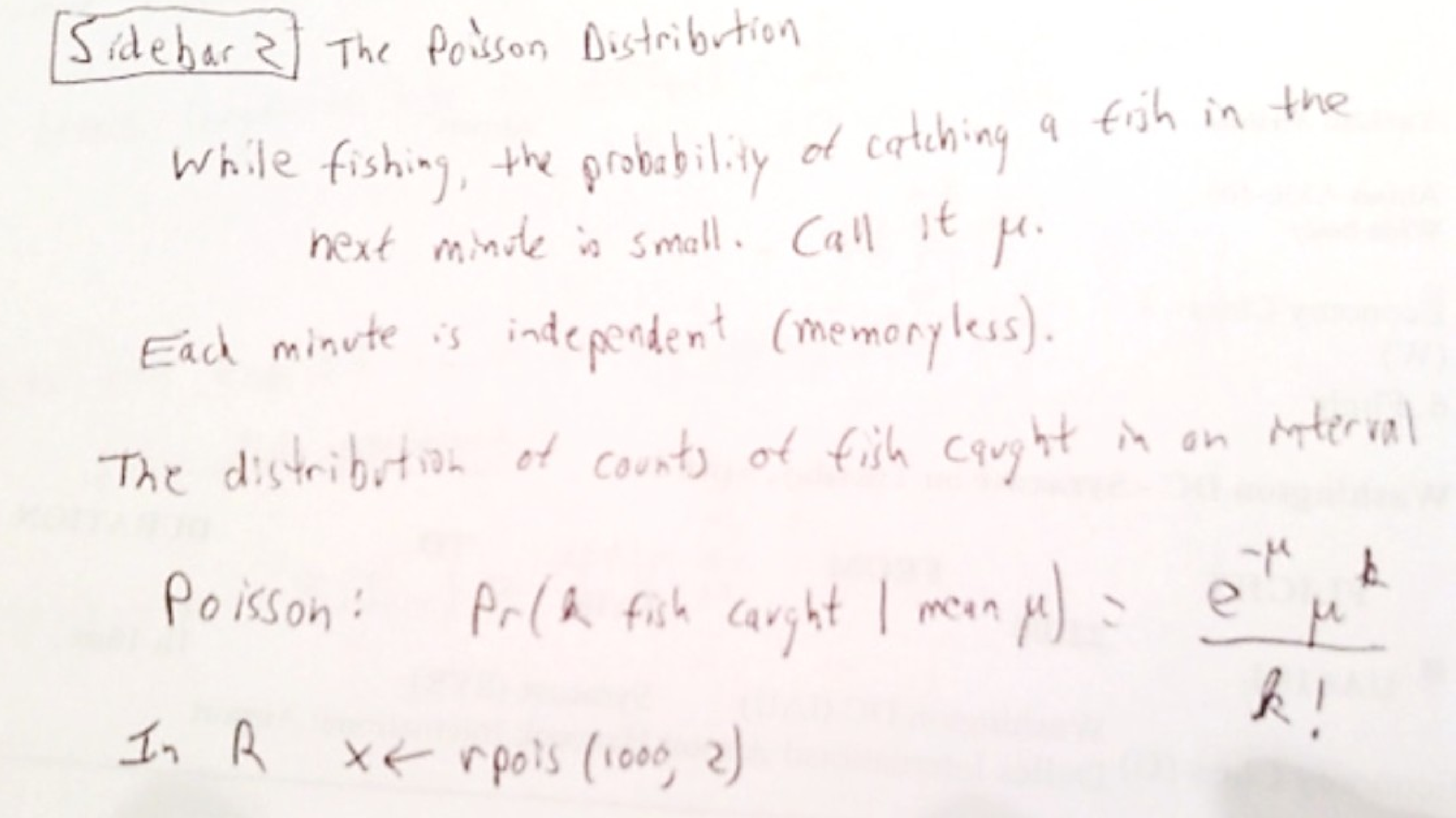
Pr (coalescing at gen t] = $\left(1-\frac{1}{2N}\right)^{t-1}\frac{1}{2N} \approx \frac{1}{2N}e^{-\frac{t}{2N}}$

with n lineages, any pair could coalesce

$\binom{n}{2}$ = "n choose 2" = $\frac{n(n-1)}{2}$ so Pr (coalesce) = $\frac{n(n-1)}{2}\,\frac{1}{2N}$

So distribution of time to the first coalescent = $\frac{\binom{n}{2}}{2N}\,e^{-\frac{\binom{n}{2}}{2N}t}$

Expected time to first coalescent = $\frac{2N}{\binom{n}{2}} = \frac{4N}{n(n-1)}$ (happens fast when n is large)



## Sidebar 2 The Poisson Distribution

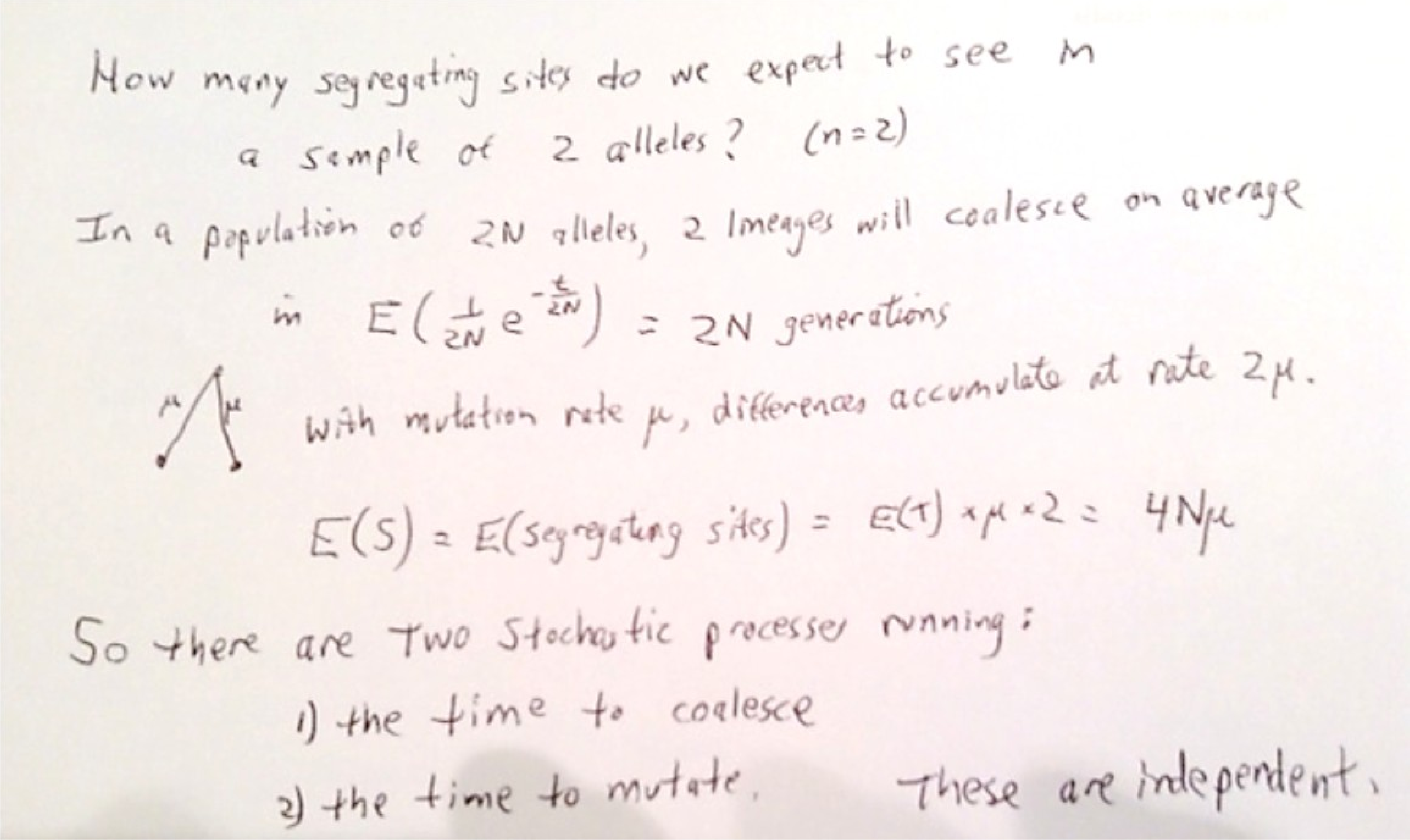
While fishing, the probability of catching a fish in the next minute is small. Call it $\mu$.

Each minute is independent (memoryless).

The distribution of counts of fish caught in an interval

Poisson: $$\Pr(k \text{ fish caught} \mid \text{mean } \mu) = \frac{e^{-\mu}\mu^{k}}{k!}$$

In R x ← rpois(1000, 2)



How many segregating sites do we expect to see in a sample of 2 alleles? $(n=2)$

In a population of $2N$ alleles, 2 lineages will coalesce on average in $E(\frac{1}{2N}e^{-\frac{t}{2N}}) = 2N$ generations

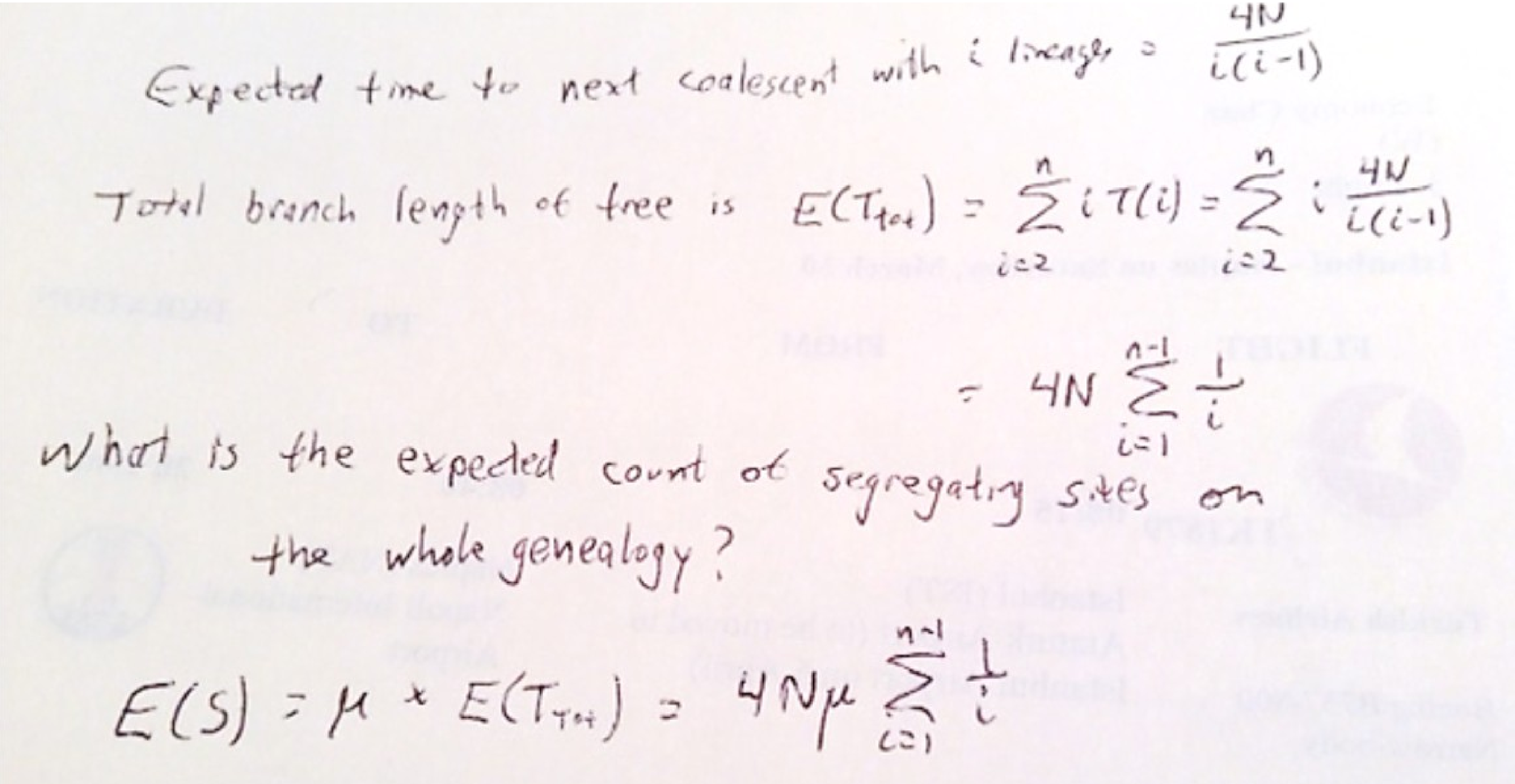
With mutation rate $\mu$, differences accumulate at rate $2\mu$.

$$E(S) = E(\text{segregating sites}) = E(T) \times \mu \times 2 = 4N\mu$$

So there are Two Stochastic processes running:

1) the time to coalesce

2) the time to mutate. These are independent.



Expected time to next coalescent with $i$ lineages = $\frac{4N}{i(i-1)}$

Total branch length of tree is $E(T_{tot}) = \sum_{i=2}^{n} i\,T(i) = \sum_{i=2}^{n} i\,\frac{4N}{i(i-1)}$

$$= 4N \sum_{i=1}^{n-1} \frac{1}{i}$$

What is the expected count of segregating sites on the whole genealogy?

$$E(S) = \mu * E(T_{Tot}) = 4N\mu \sum_{i=1}^{n-1} \frac{1}{i}$$