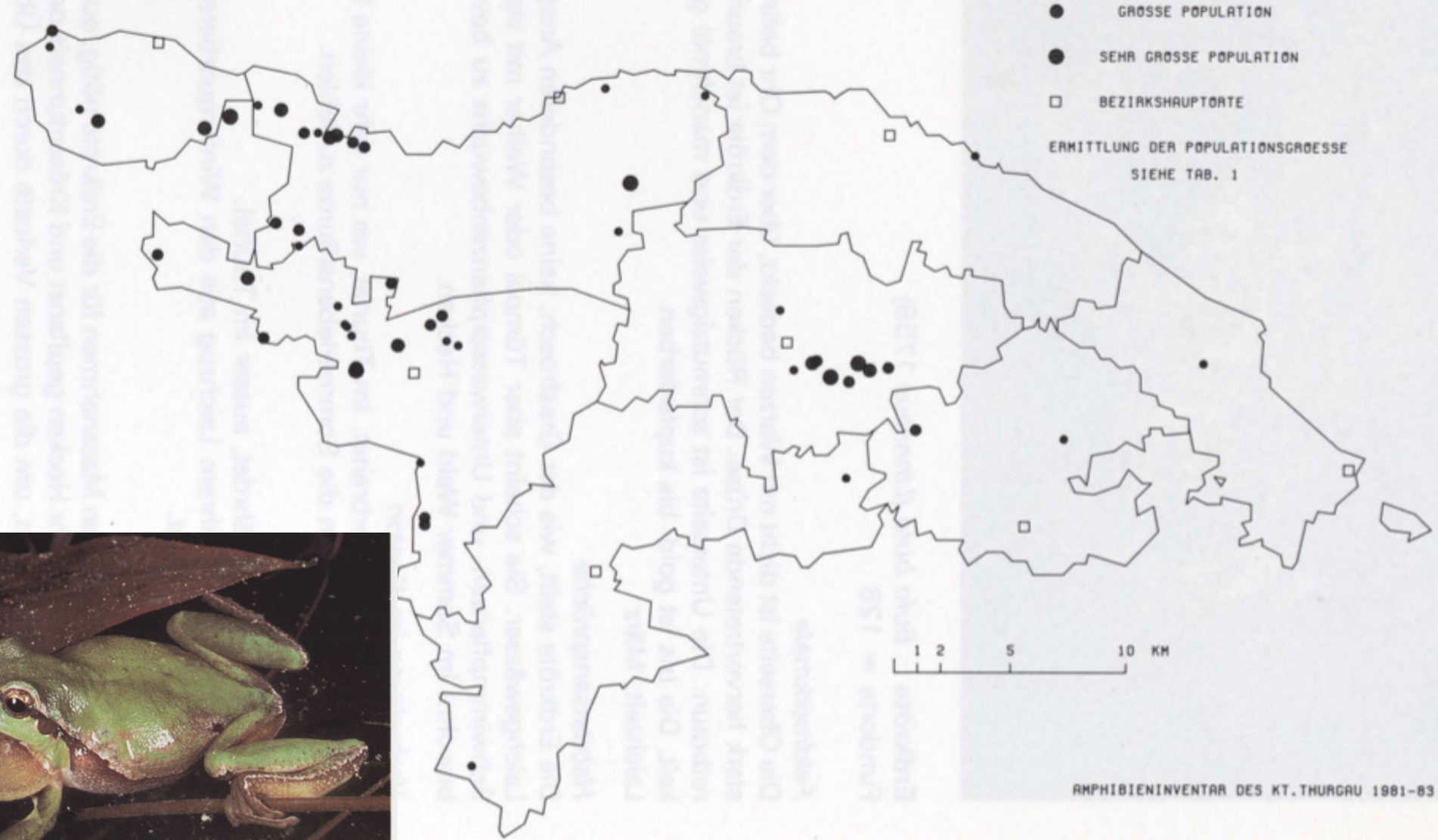
From frogs to theory

Peter Beerli
Department of Scientific Computing
Florida State University


Bodensee

VERBREITUNG DES **LAUBFROSCHES** IM KANTON THURGAU
(HYLA ARBOREA)



CREDIT SUISSE

Kaden & Partner

Ihr zuverlässiger Partner für komplexe Projekte

Ökologie

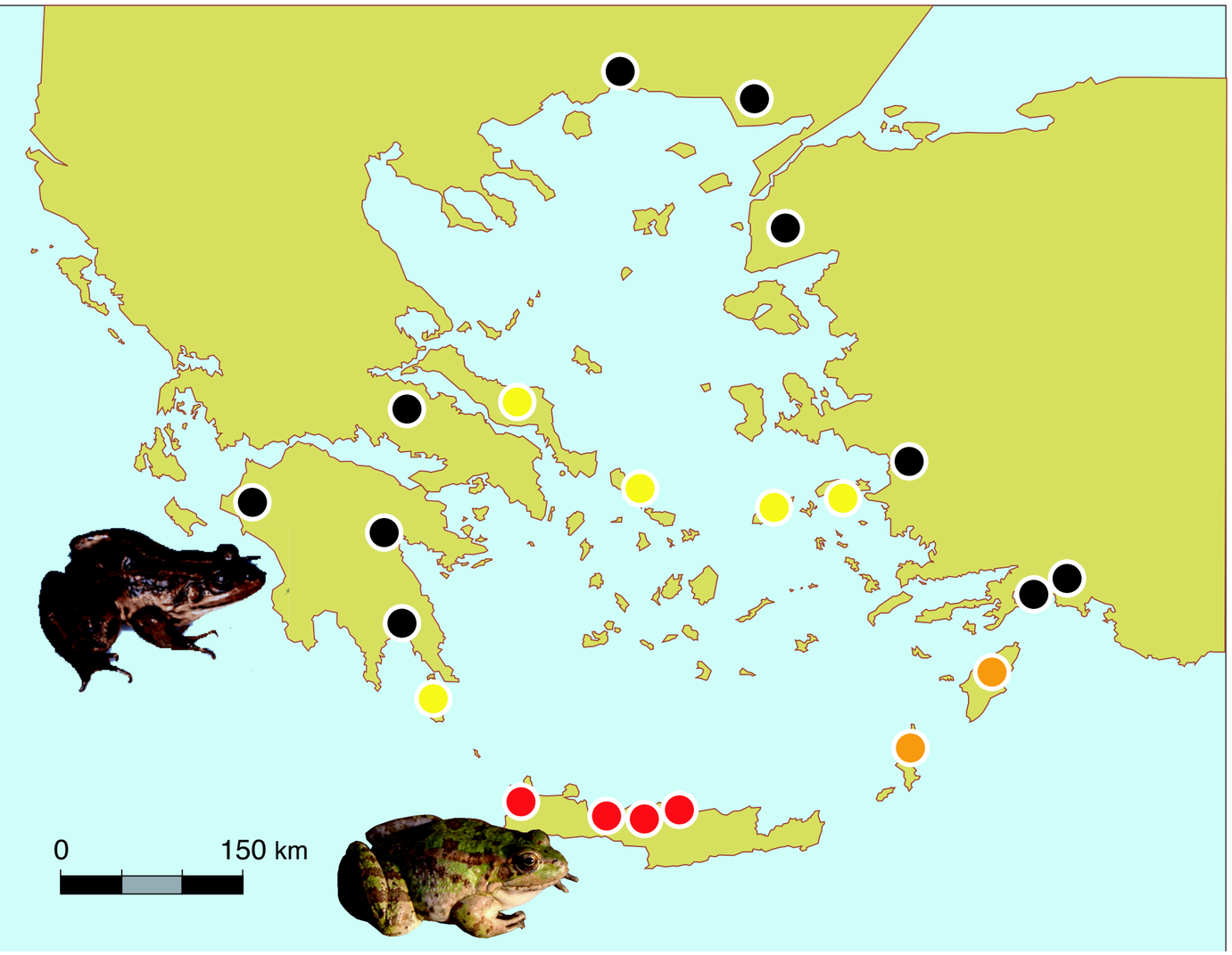
Ökologische Fachberatung

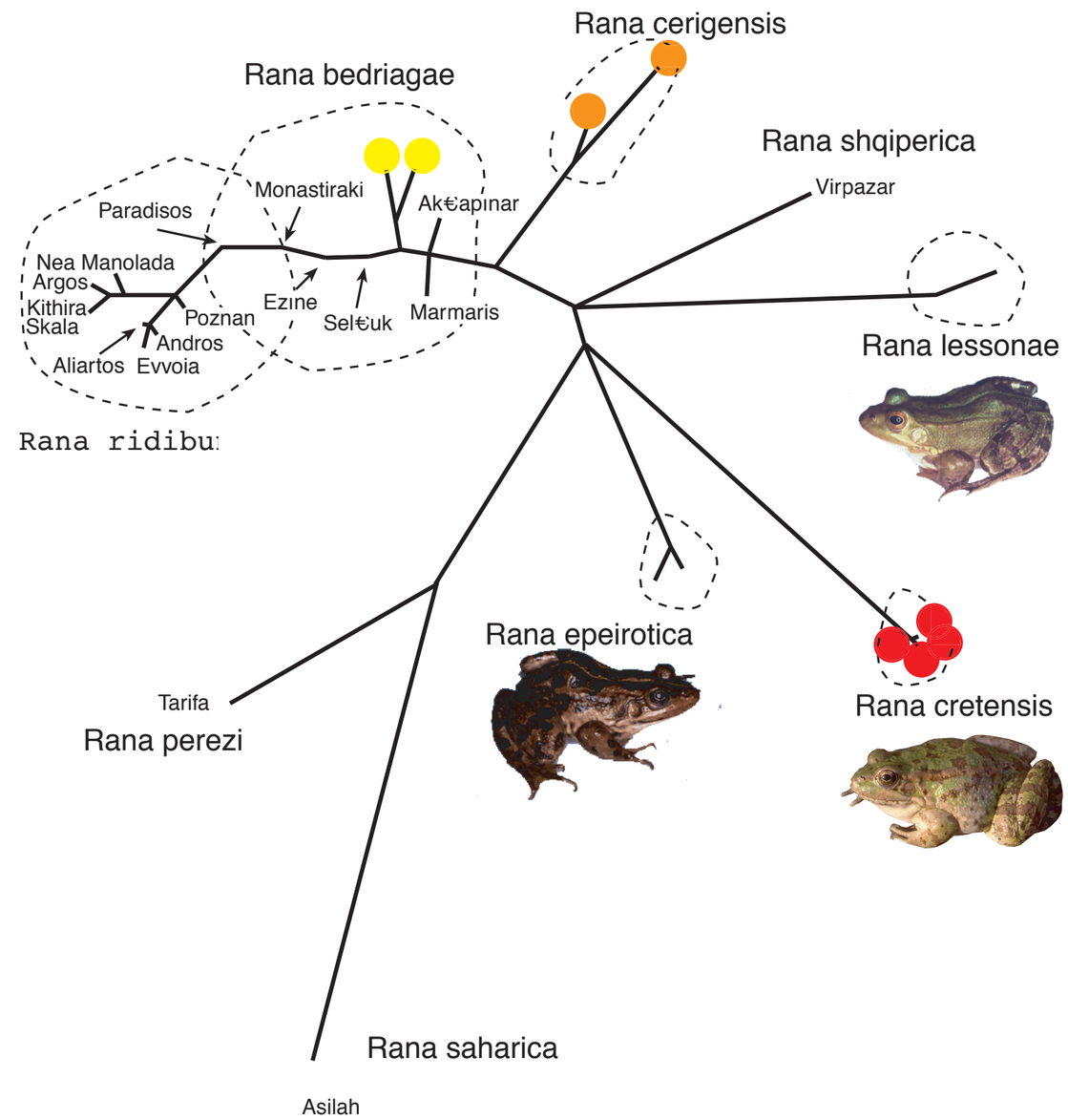
Software

Web- Mobile- und GIS-Apps

Luftbilder

Drohnenaufnahmen der Landschaft

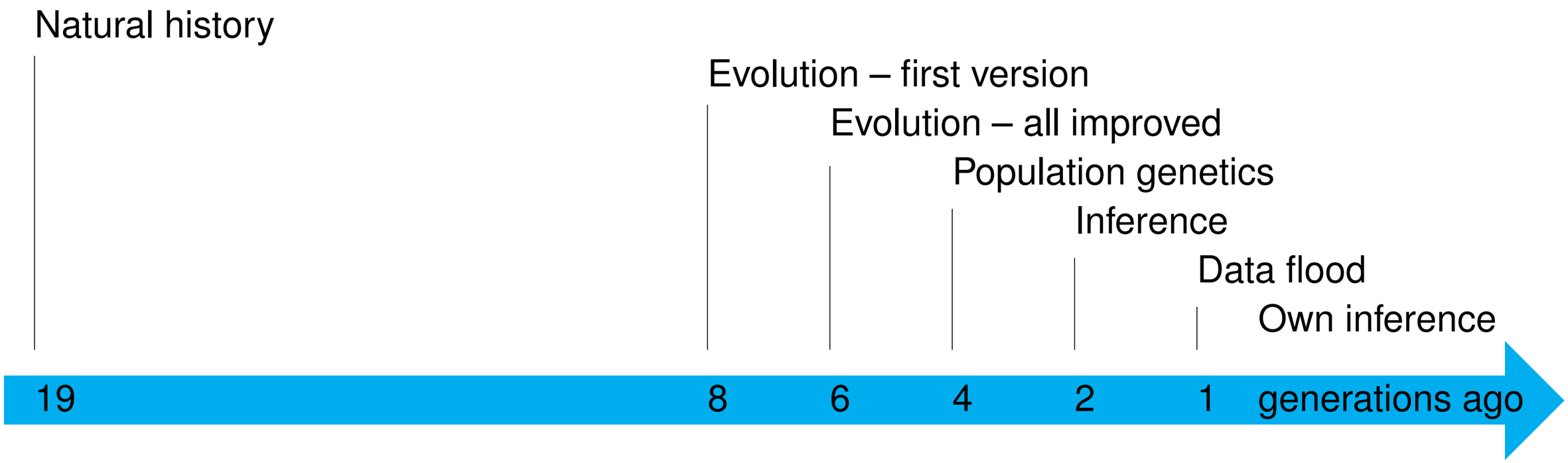

0
150 km


Rana cerigensis
Rana bedriagae
Rana shqiperica
Virpazar
Monastiraki
Akçapınar
Paradisos
Nea Manolada
Argos
Kithira
Skala
Poznan
Ezine
Selçuk
Marmaris
Andros
Aliartos
Evvoia
Rana ridibu
Rana lessonae
Rana epeirotica
Tarifa
Rana perezi
Rana cretensis
Rana saharica
Asilah

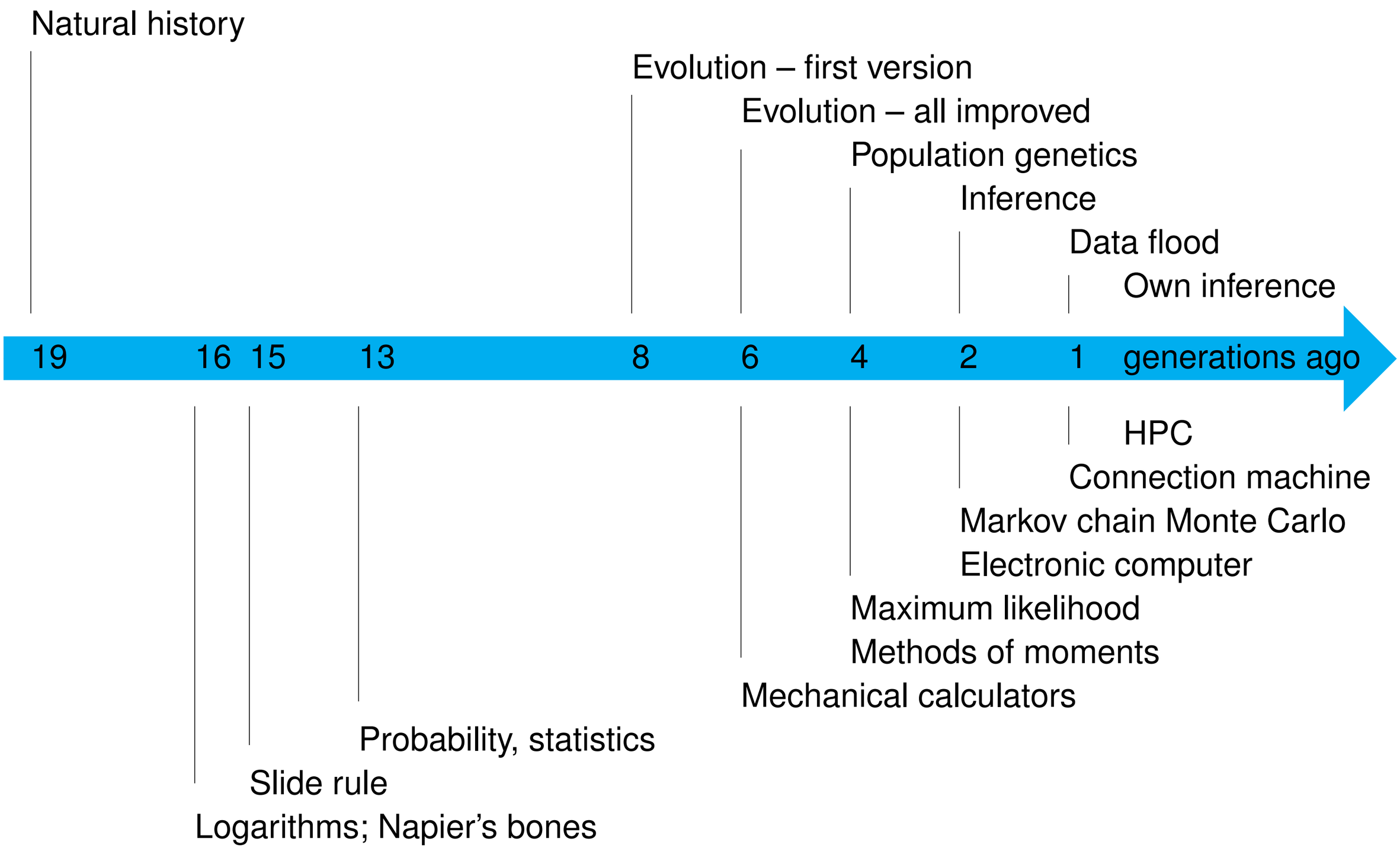
Computational population genetics

Peter Beerli
Department of Scientific Computing
Florida State University

History of population genetics

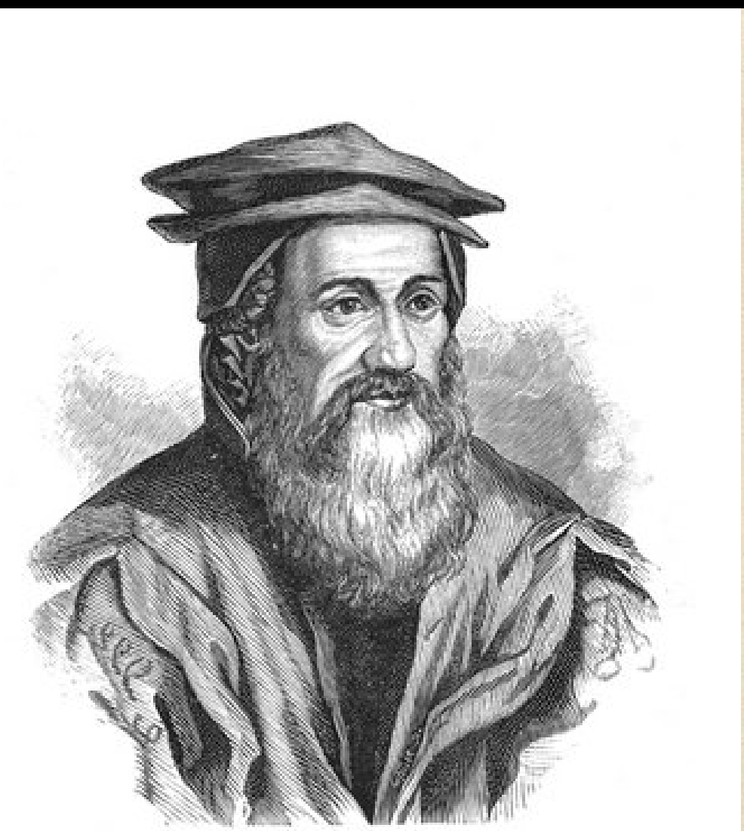


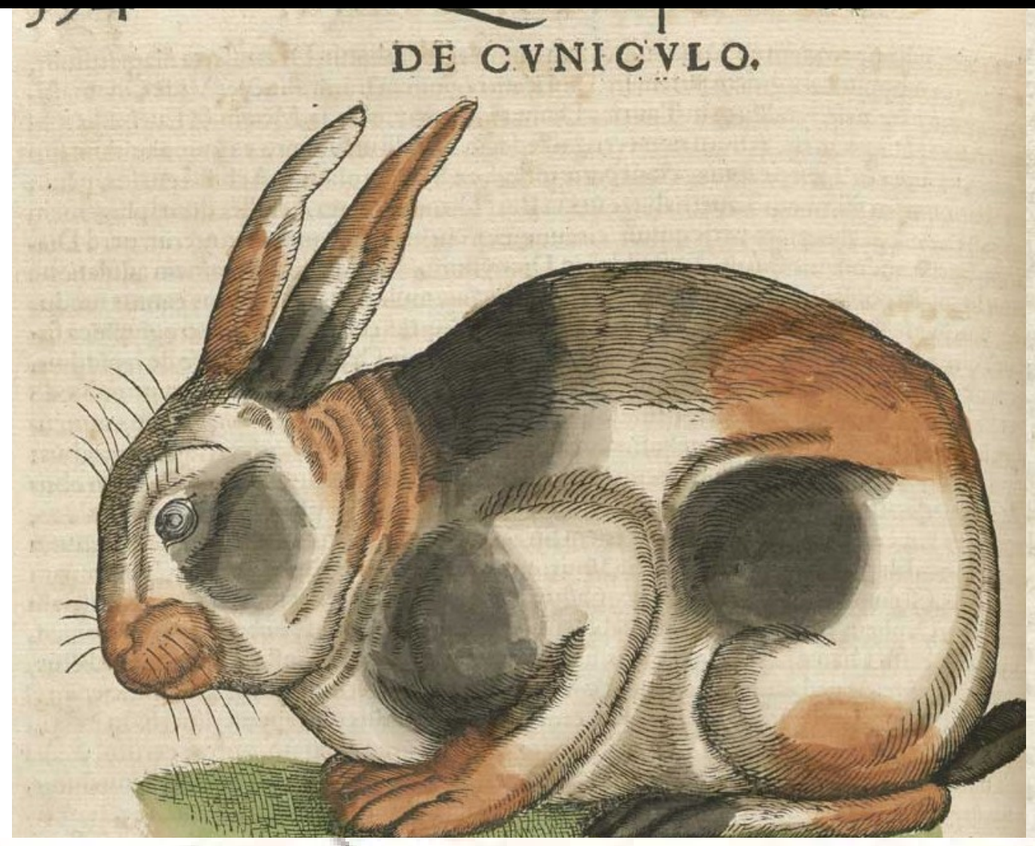
History of population genetics



Pre-dawn of population genetics

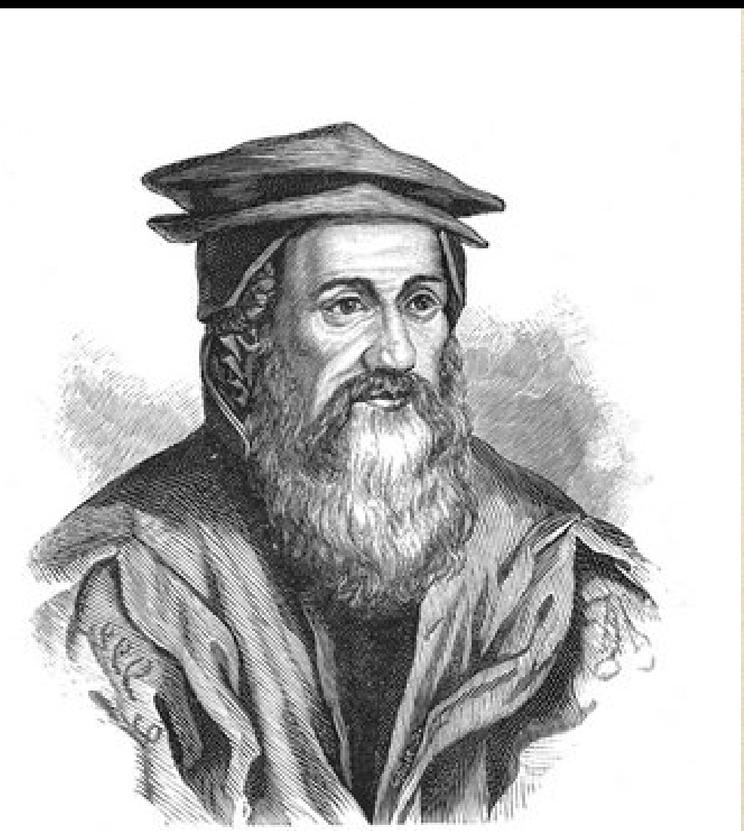
19 generations ago

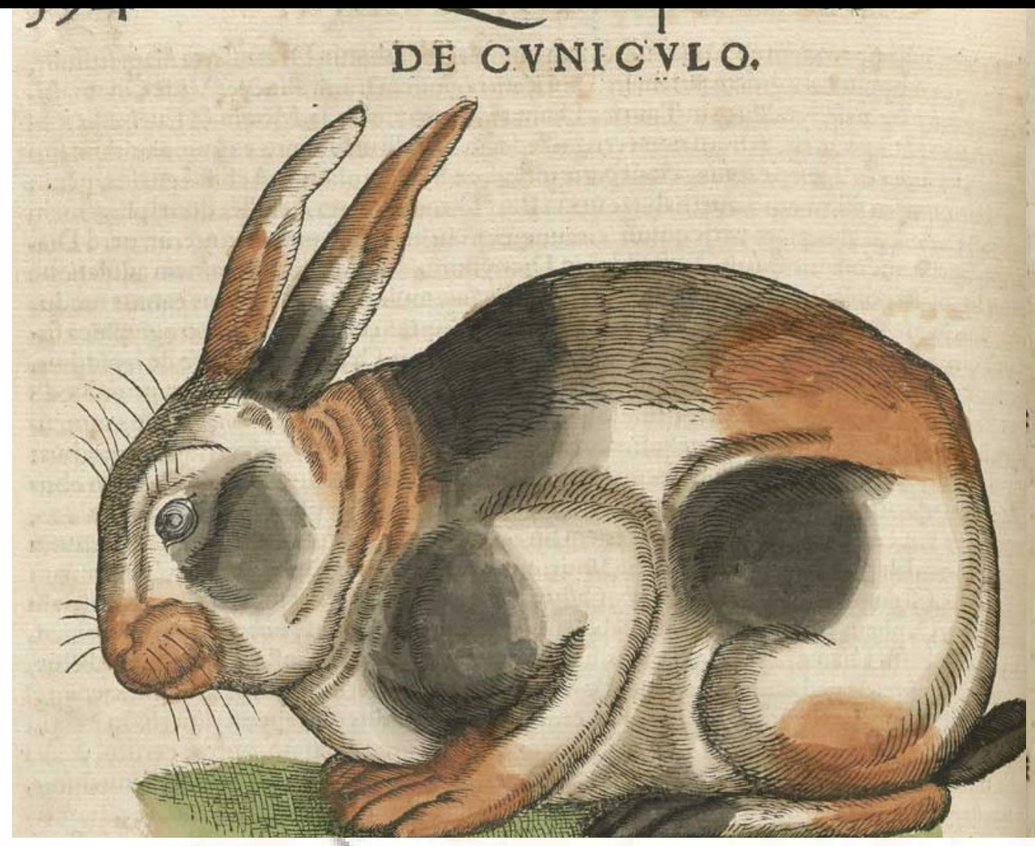




Pre-dawn of population genetics

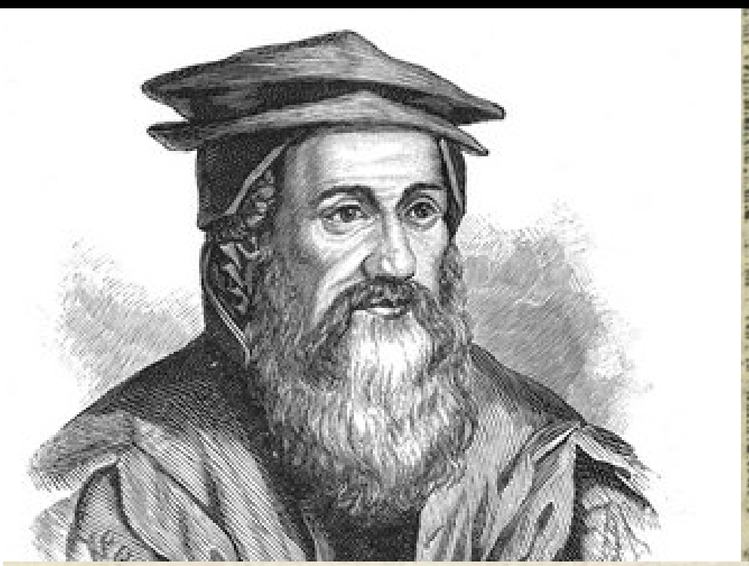
19 generations ago

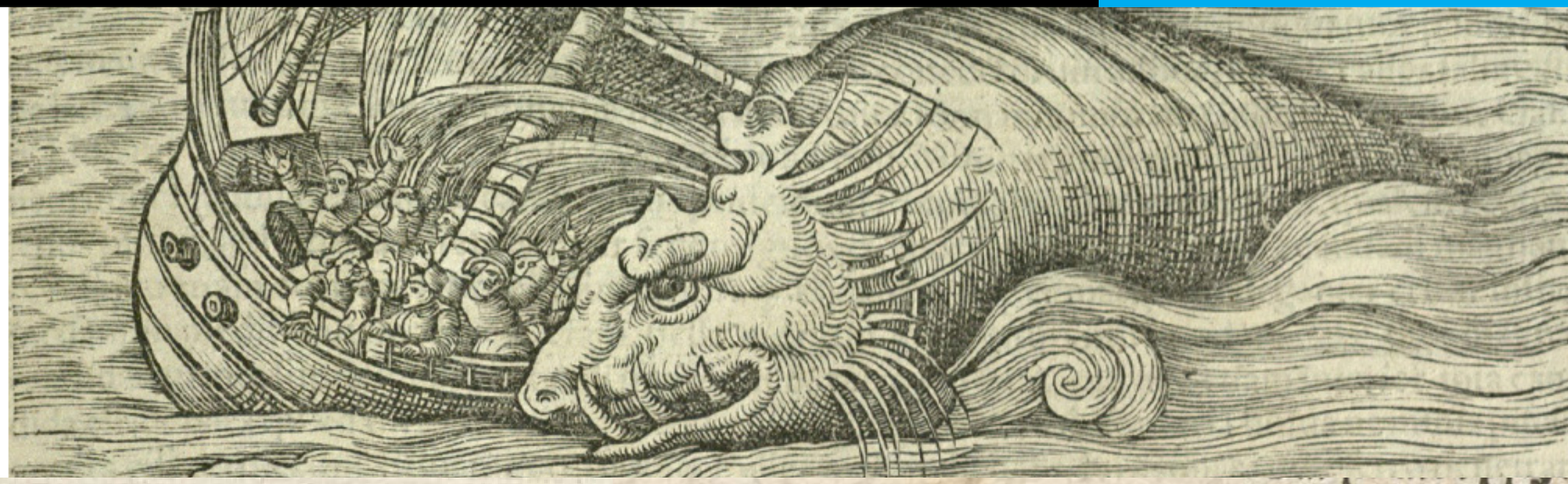




Pre-dawn of population genetics

19 generations ago



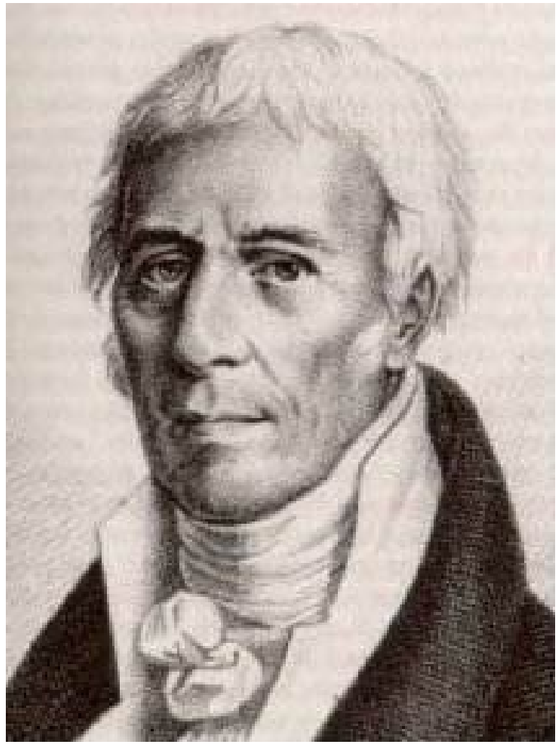


Computing devices

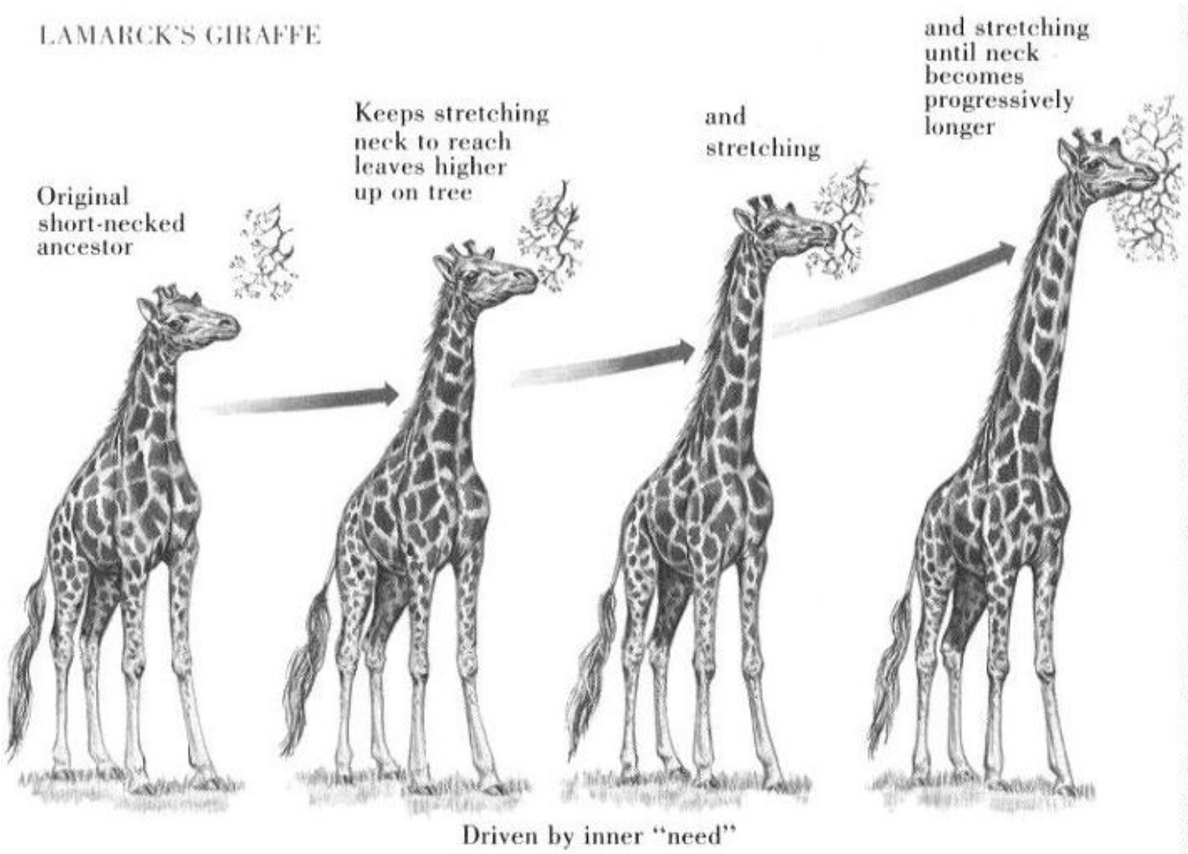
15 generations ago

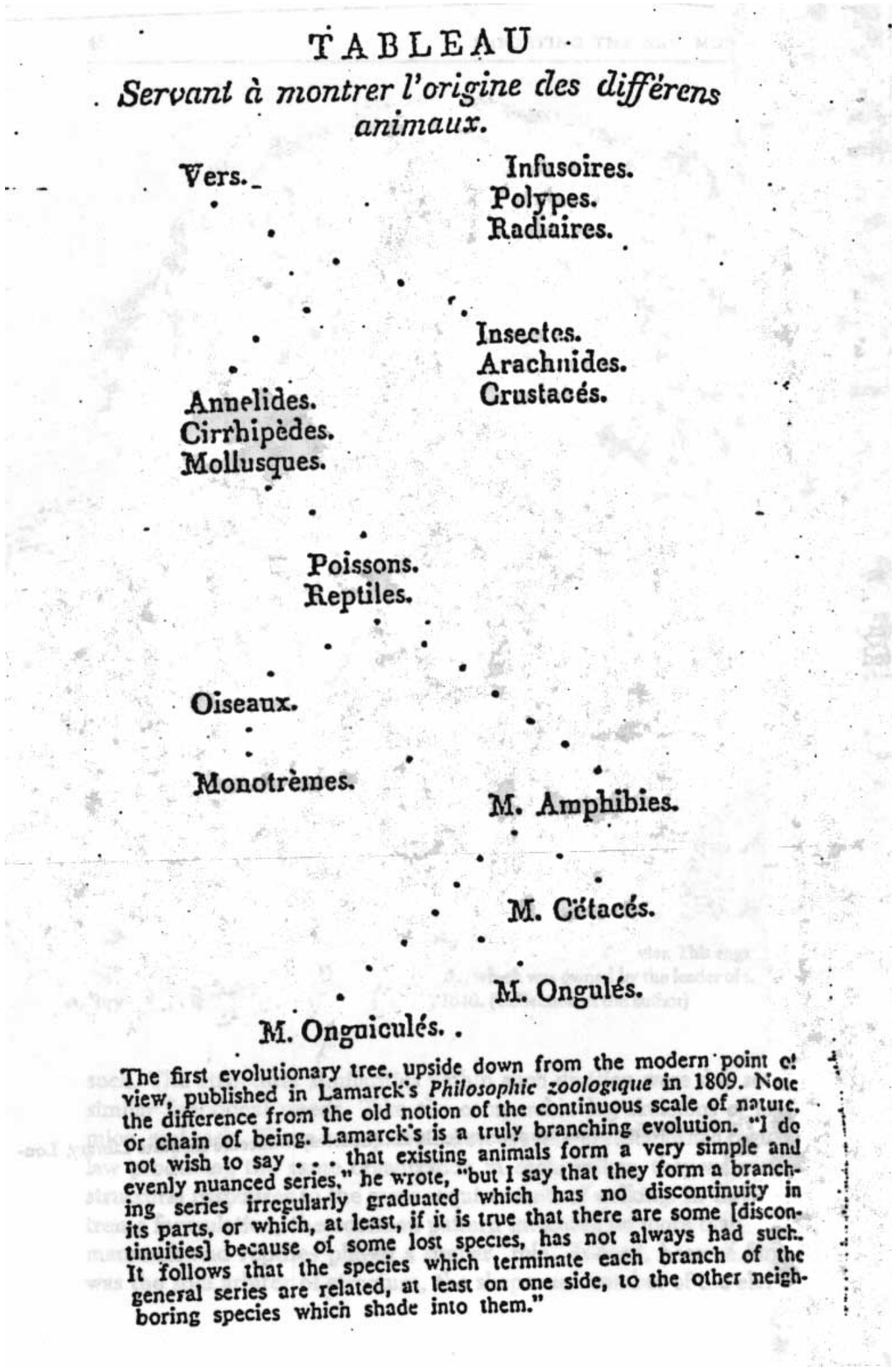
Dawn of population genetics

8 generations ago



Lamarck



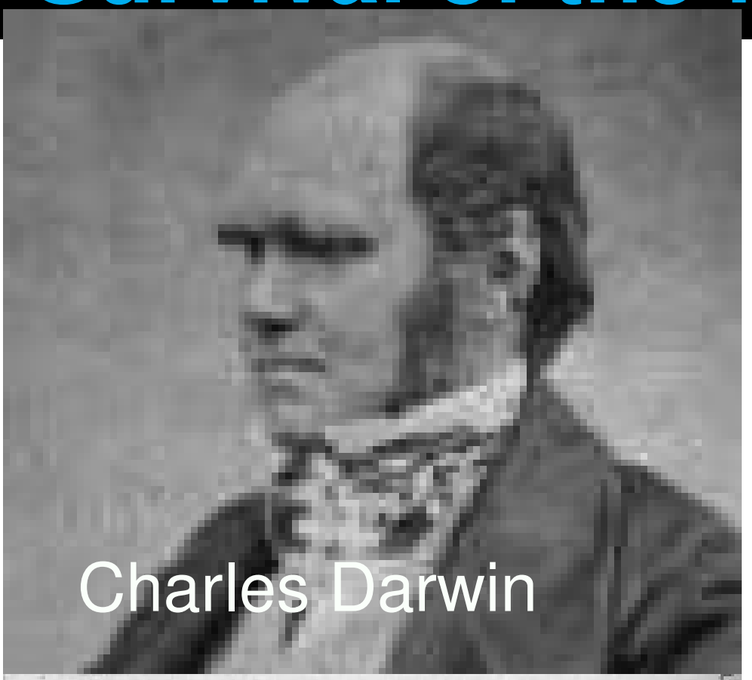


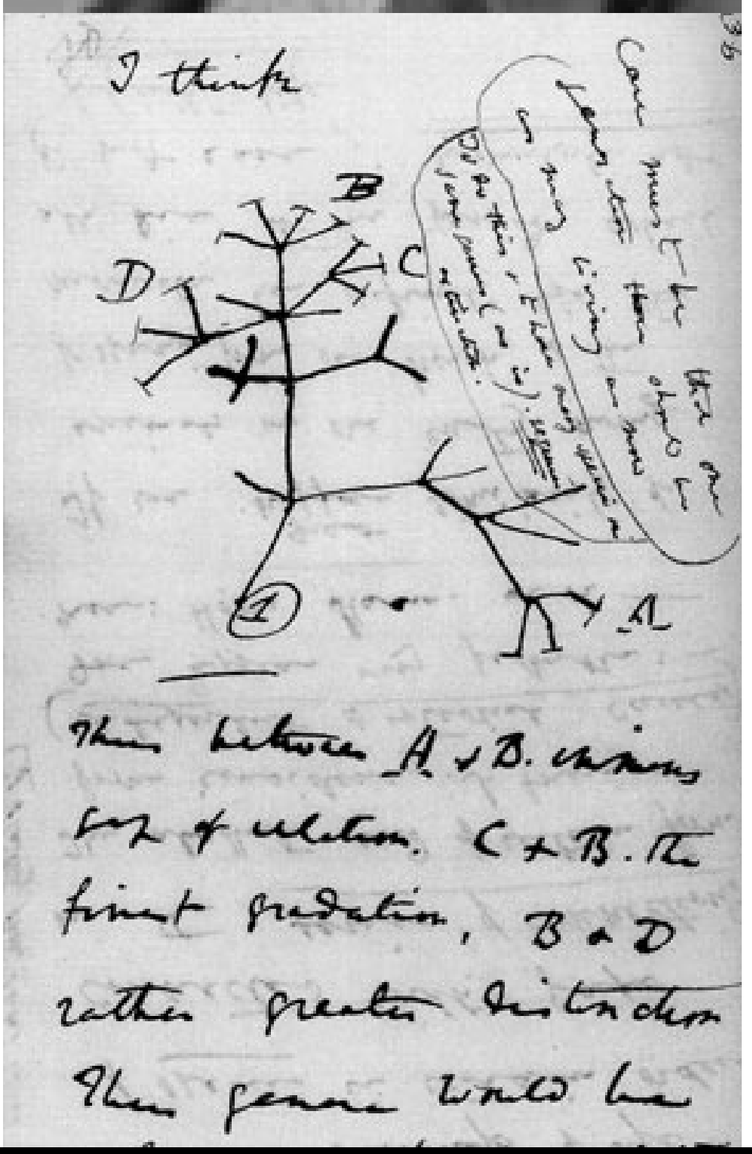
The first evolutionary tree, upside down from the modern point of view, published in Lamarck's *Philosophie zoologique* in 1809. Note the difference from the old notion of the continuous scale of nature, or chain of being. Lamarck's is a truly branching evolution. "I do not wish to say . . . that existing animals form a very simple and evenly nuanced series," he wrote, "but I say that they form a branching series irregularly graduated which has no discontinuity in its parts, or which, at least, if it is true that there are some [discontinuities] because of some lost species, has not always had such. It follows that the species which terminate each branch of the general series are related, at least on one side, to the other neighboring species which shade into them."

(c) http://www.princessleia.com/images/MyImages/essays/giraffe_lamark.jpg

Survival of the fittest

6 generations ago

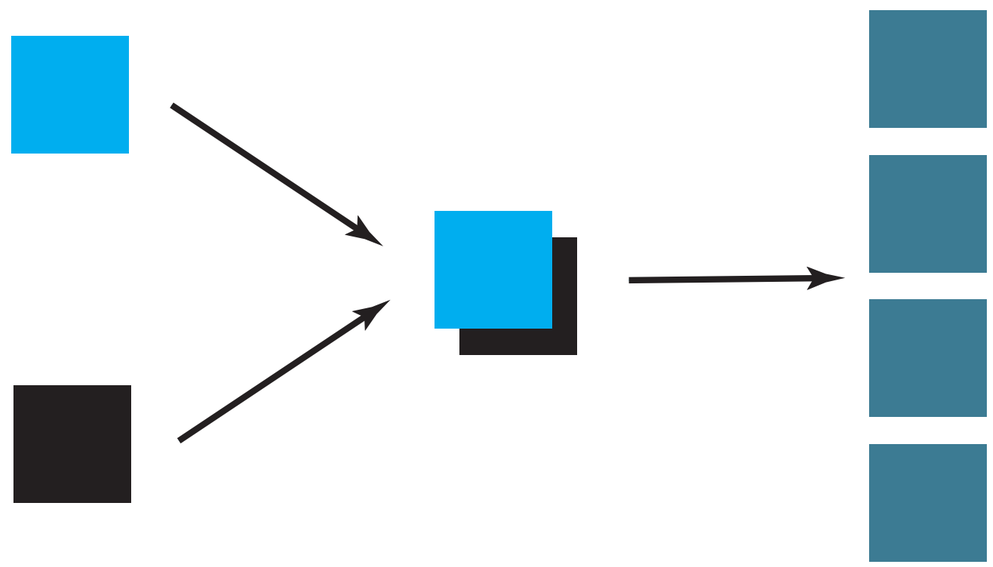




Again, it may be asked, how is it that varieties, which I have called incipient species, become ultimately converted into good and distinct species, which in most cases obviously differ from each other far more than do the varieties of the same species? **How do those groups of species, which constitute what are called distinct genera and which differ from each other more than do the species of the same genus, arise?** All these results, as we shall more fully see in the next chapter, follow from the struggle for life. Owing to this struggle, variations, however slight and from whatever cause proceeding, if they be in any degree profitable to the individuals of a species, in their infinitely complex relations to other organic beings and to their physical conditions of life, will tend to the preservation of such individuals, and will generally be inherited by the offspring. The offspring, also, will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive. I have called this principle, by which each slight variation, if useful, is preserved, by the term **natural selection**, in order to mark its relation to man's power of selection. But the expression often used by Mr. Herbert Spencer, of the **Survival of the Fittest**, is more accurate, and is sometimes equally convenient. We have seen that man by selection can certainly produce great results, and can adapt organic beings to his own uses, through the accumulation of slight but useful variations, given to him by the hand of Nature. But Natural Selection, we shall hereafter see, is a power incessantly ready for action, and is as immeasurably superior to man's feeble efforts, as the works of Nature are to those of Art. [**Origin of Species 6th ed.**]

Blending inheritance

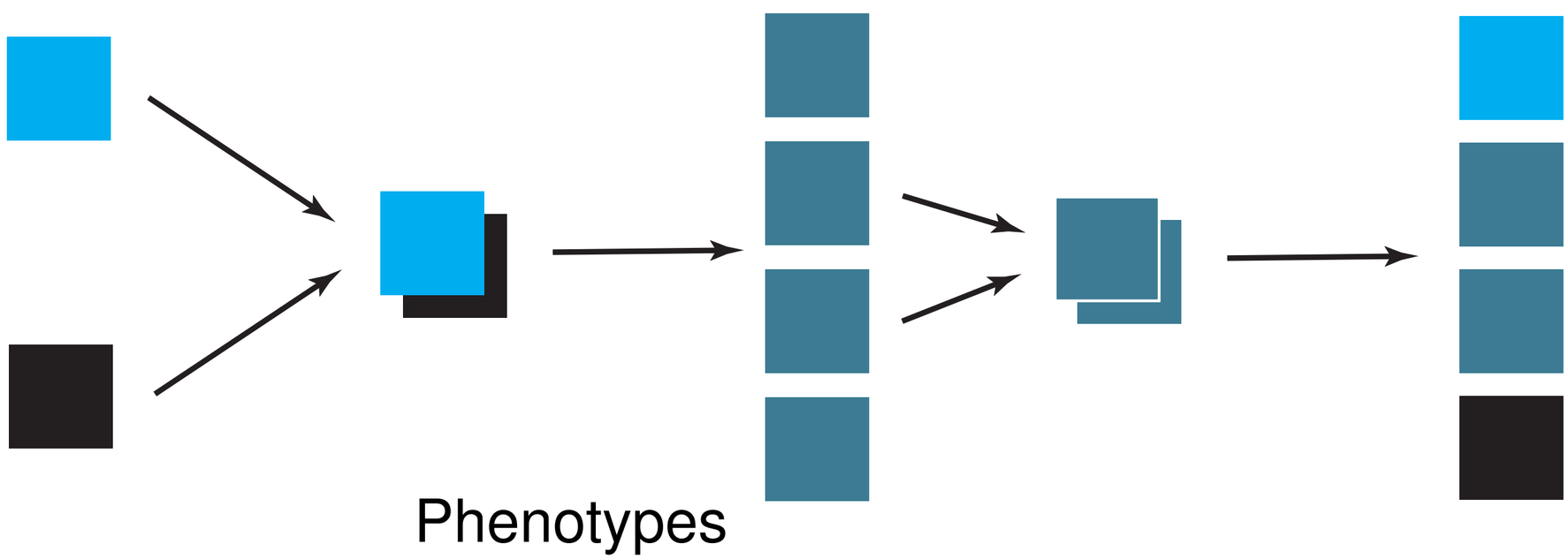
Common genetics mechanism at the time could not explain natural selection because a ”favored” trait blended with other less favored traits.



Lost Genetics

Gregor Mendel

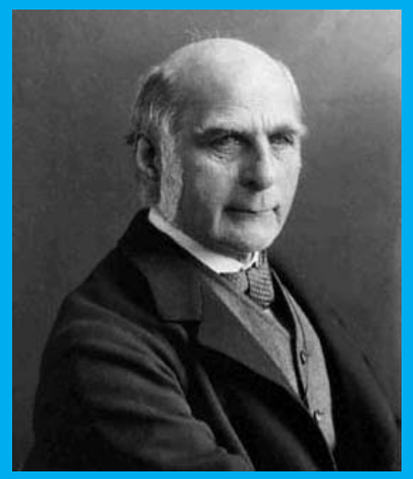
Versuche über Pflanzen-Hybriden was the result after years spent studying genetic traits in pea plants. Mendel read his paper to the Natural History Society of Brunn (Brno) on February 8 and March 8, 1865. [wikipedia]

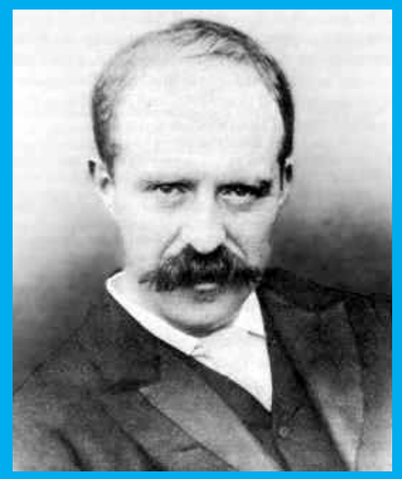


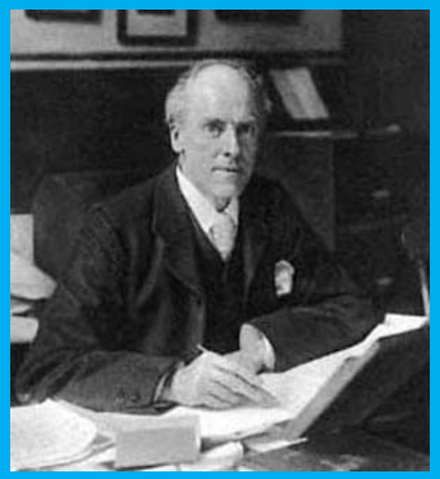
Mutation wars

Gregor Mendel's detection of genetic inheritance (and its re-detection) caused a great controversy because it seemed that at the time Mendelian genetics was only associated with large discrete changes whereas natural selection was described as working with small steady changes.

Biometricians





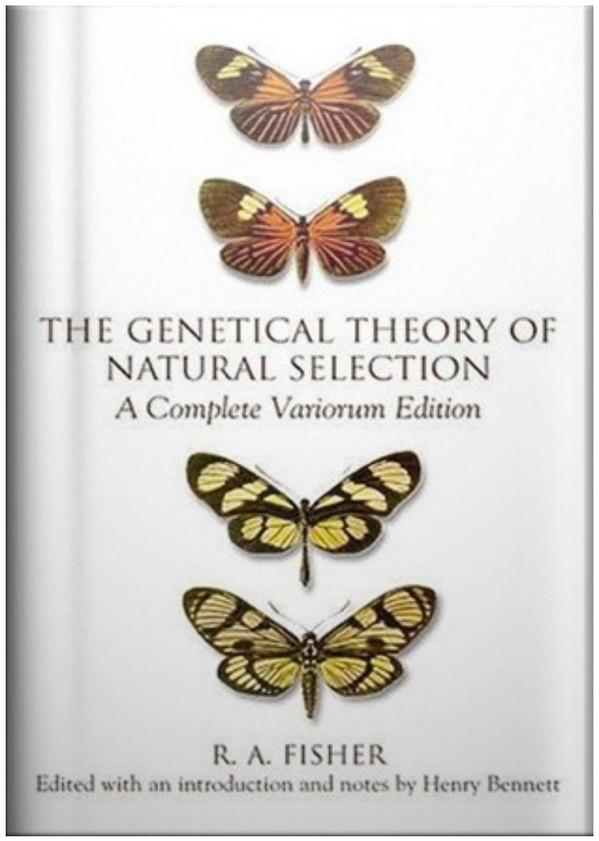


Galton, Weldon, Pearson

Mendelians

Bateson, De Vries

R. A. Fisher

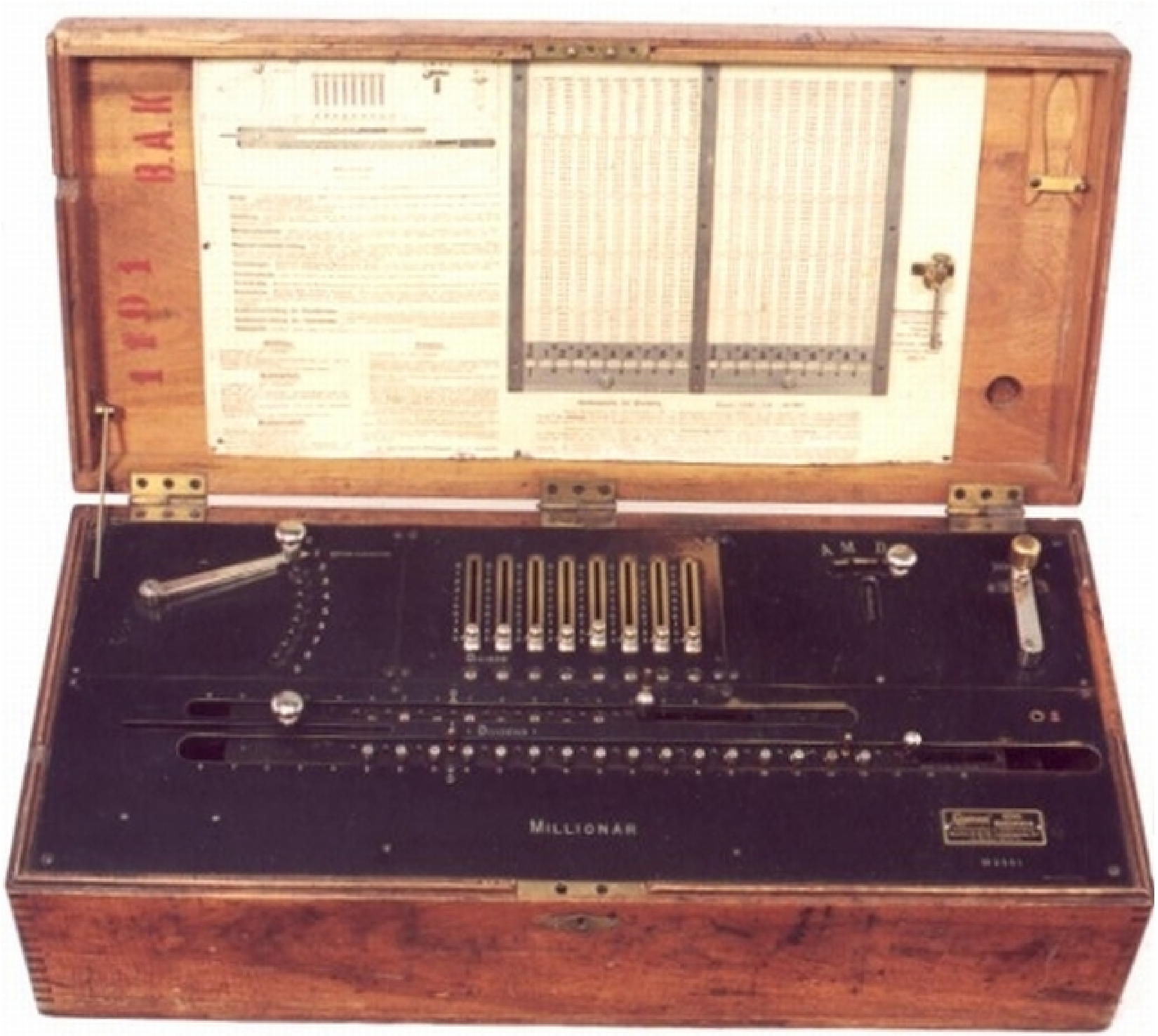


Ronald Aylmer Fisher not only concluded the fight between the Mendelians and the Biometricians and founded population genetics,
but he also invented the concept of
likelihood,
and variance analysis
among others.

Fisher about Bateson: "Unfortunately [Bateson] was unprepared to recognize the mathematical and statistical aspects of biology, and from this and other causes he was not only incapable of framing an evolutionary theory himself, but entirely failed to see how Mendelism supplied the missing parts of the structure first erected by Darwin. His interpretation of Mendelian facts was from the first too exclusively coloured by his earlier belief in the discontinuous origin of specific forms. Though his influence upon evolutionary theory was thus chiefly retrogressive, the mighty body of Mendelian researches throughout the world has evidently outgrown the fallacies with which it was first fostered. As a pioneer of genetics he has done more than enough to expiate the rash polemics of his earlier writings."

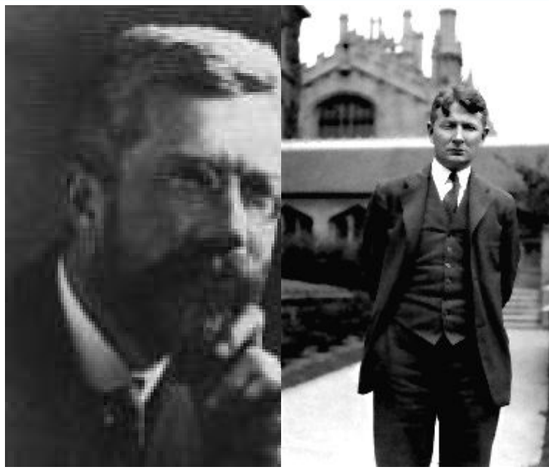
Computing devices

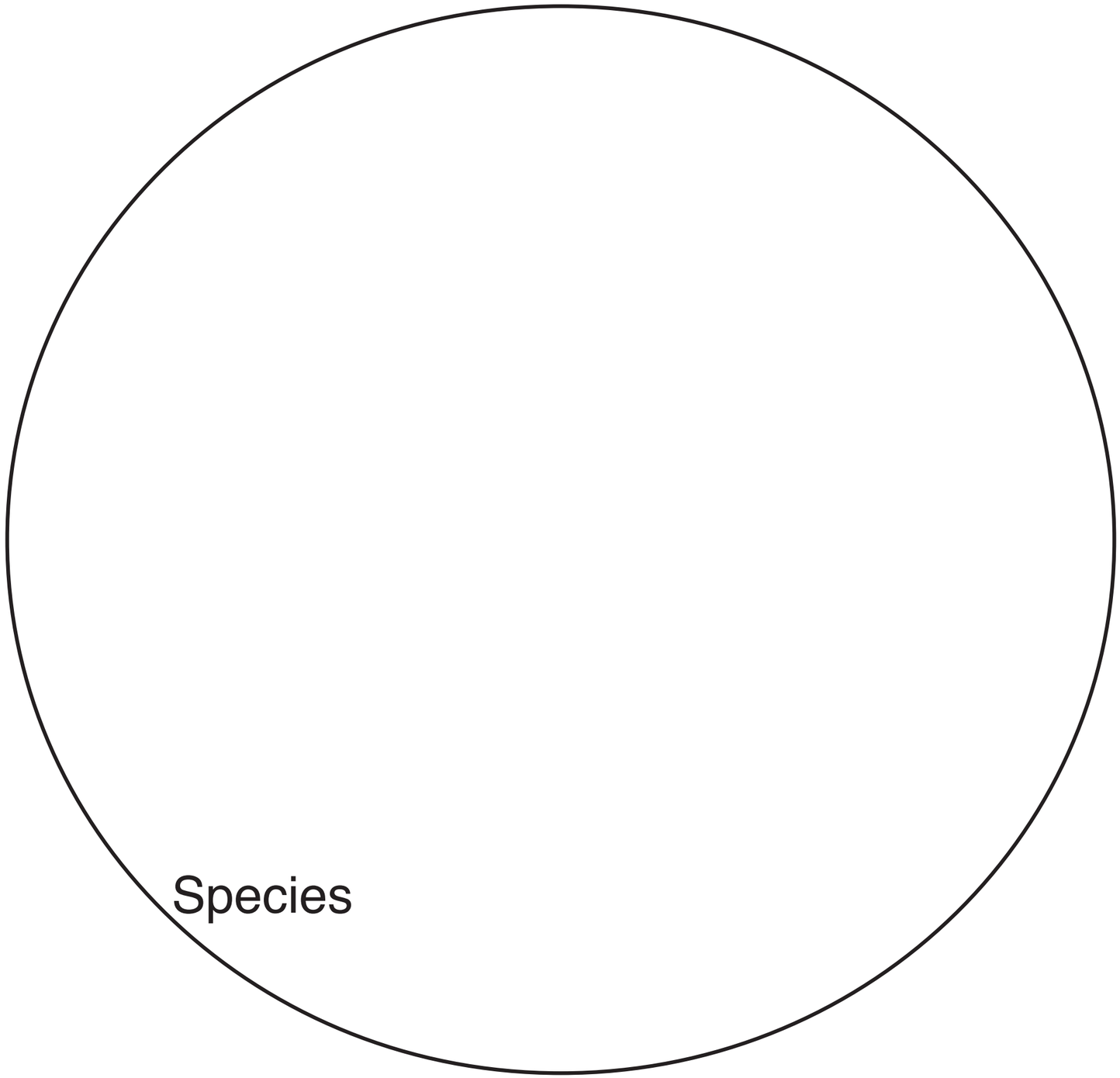
Fisher and colleagu



Population models

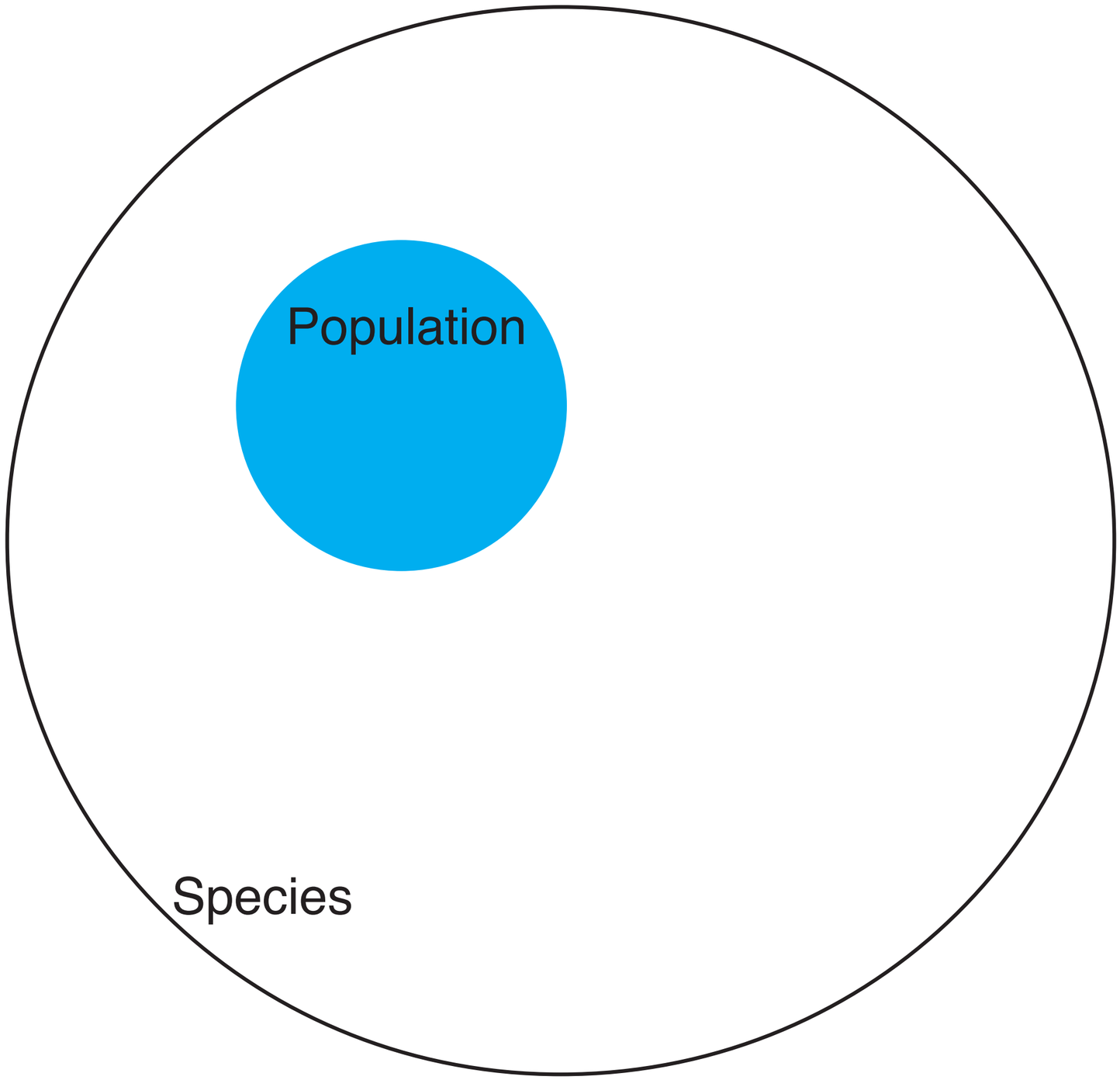
Fisher, Wright





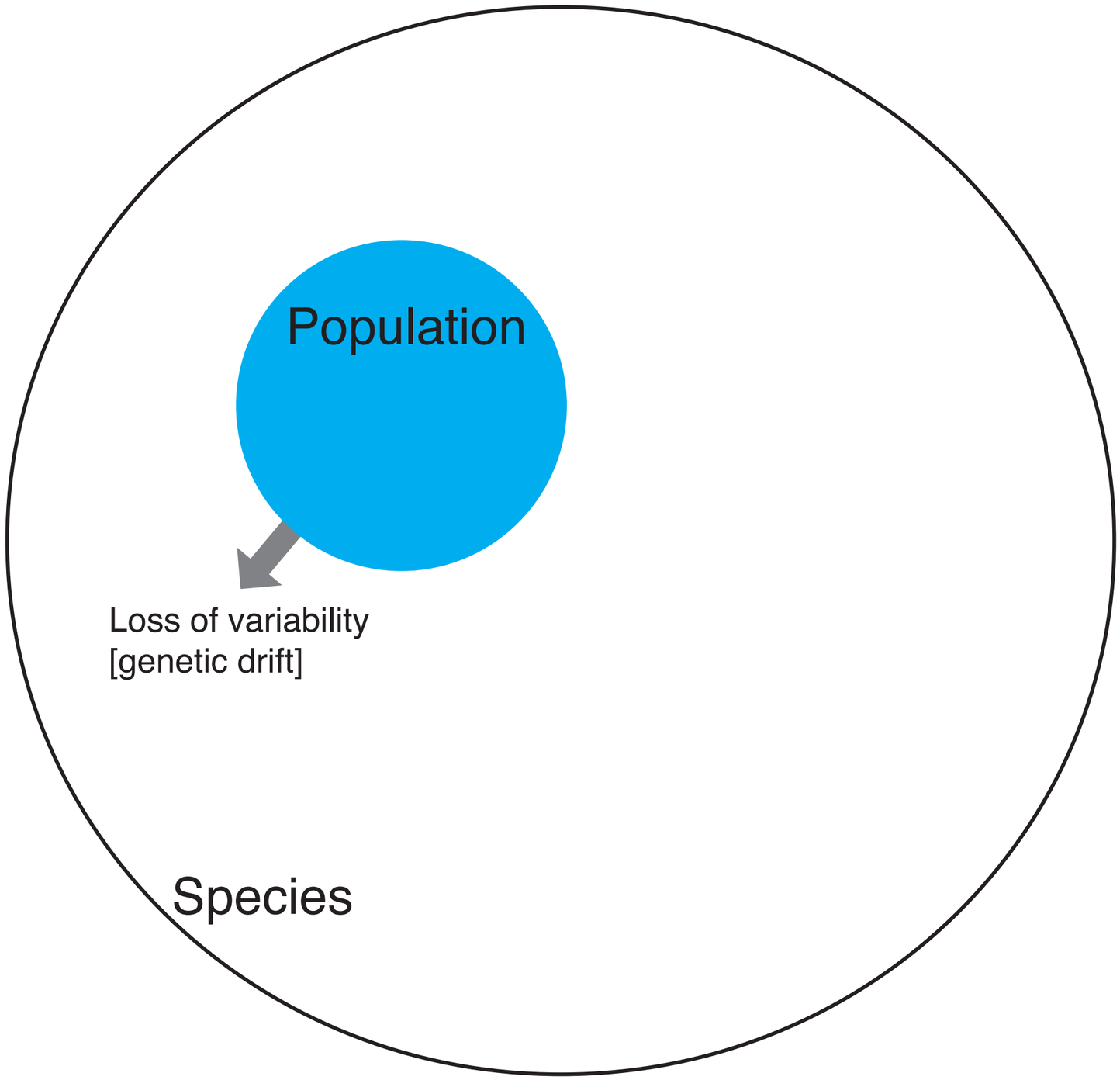
Population models

Fisher, Wright



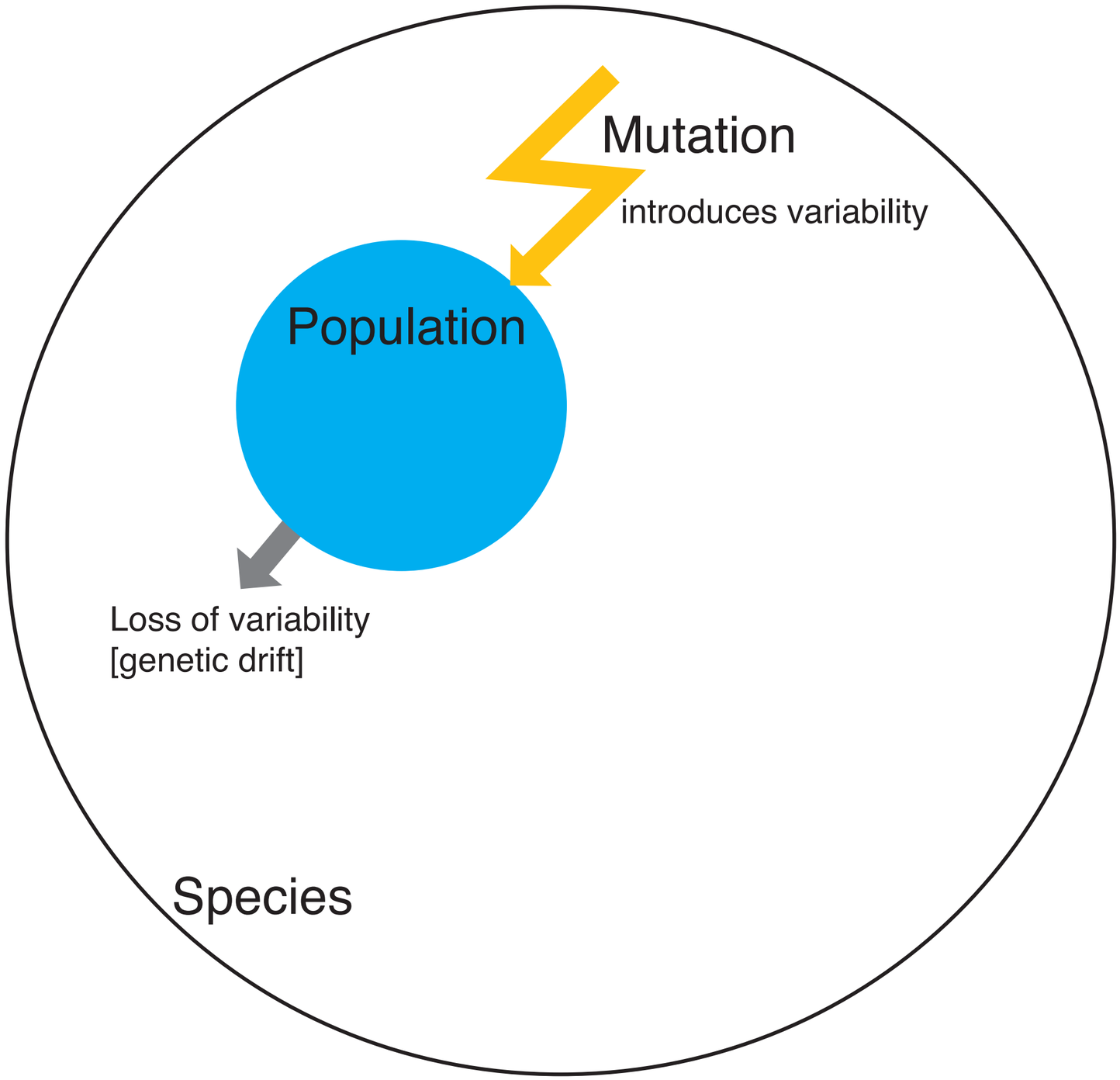
Population models

Fisher, Wright



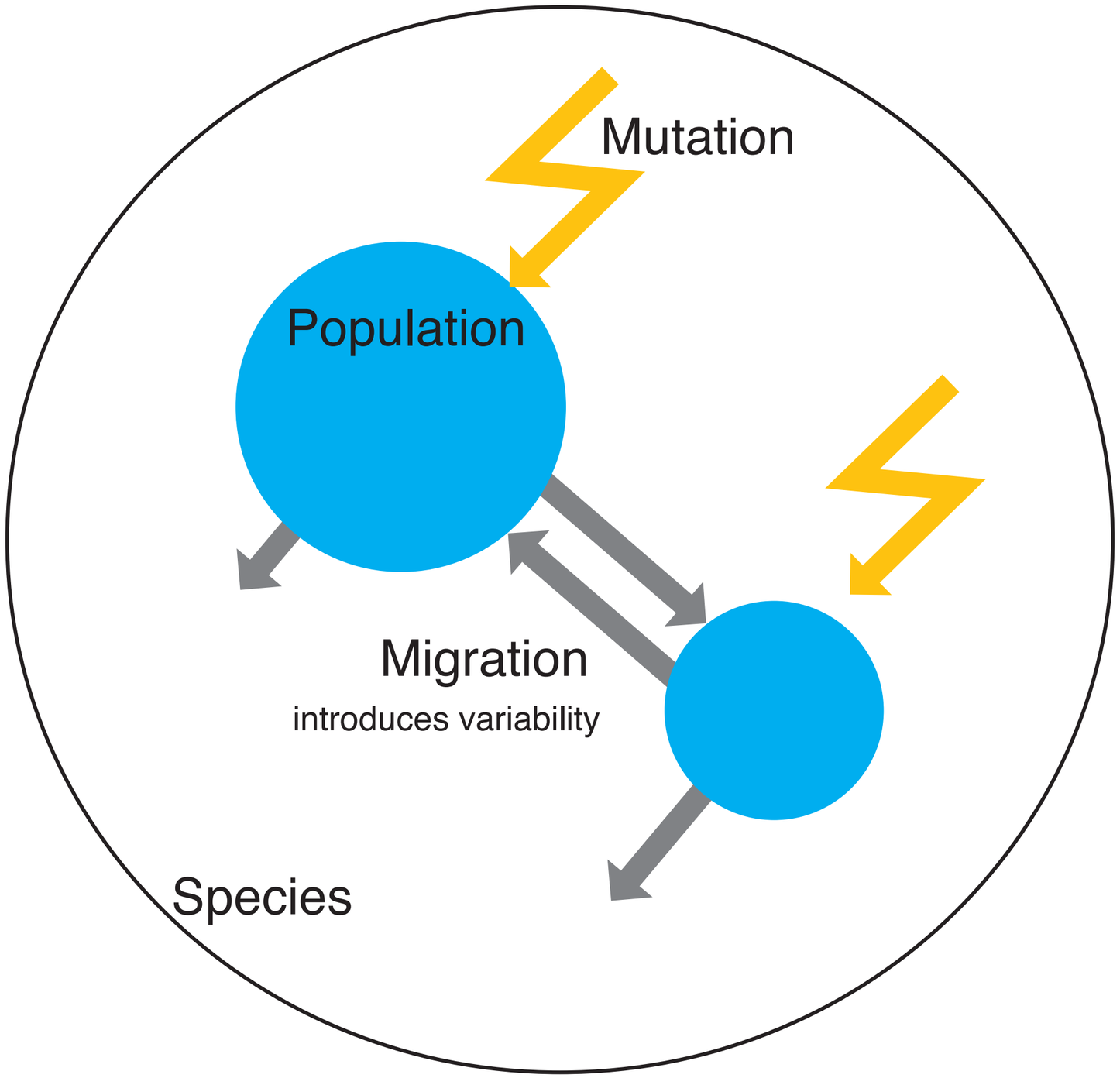
Population models

Fisher, Wright



Population models

Fisher, Wright



Population models

$$\text{Population size} = f(\text{Alleles}, \text{Mutation}, \text{Migration}, \text{population size in last generation})$$

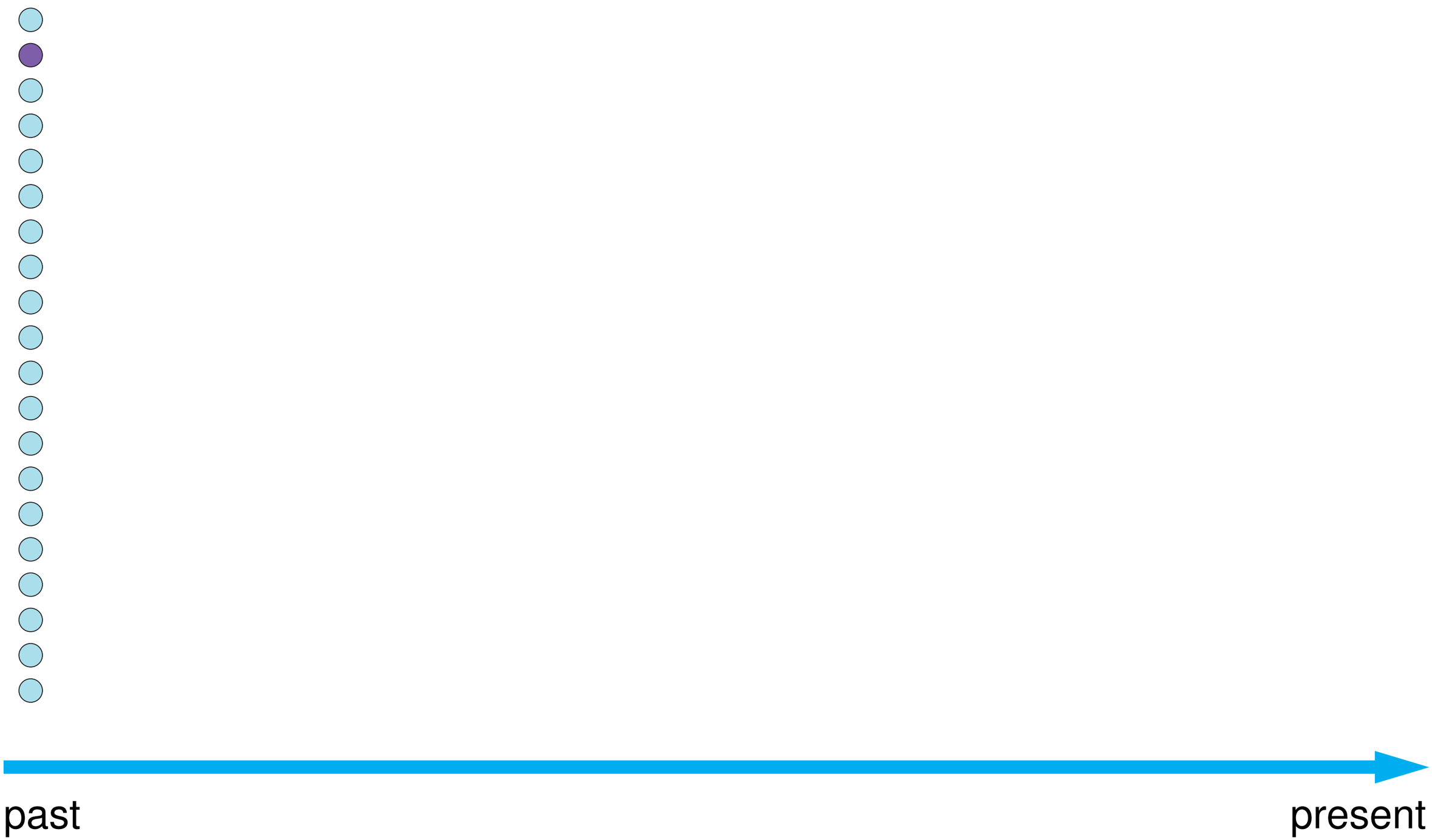
$$N_t = f(X, \mu, m, N_{t-1})$$

Simply looking only at a single population this is

$$N_t = f(X, \mu, N_{t-1})$$

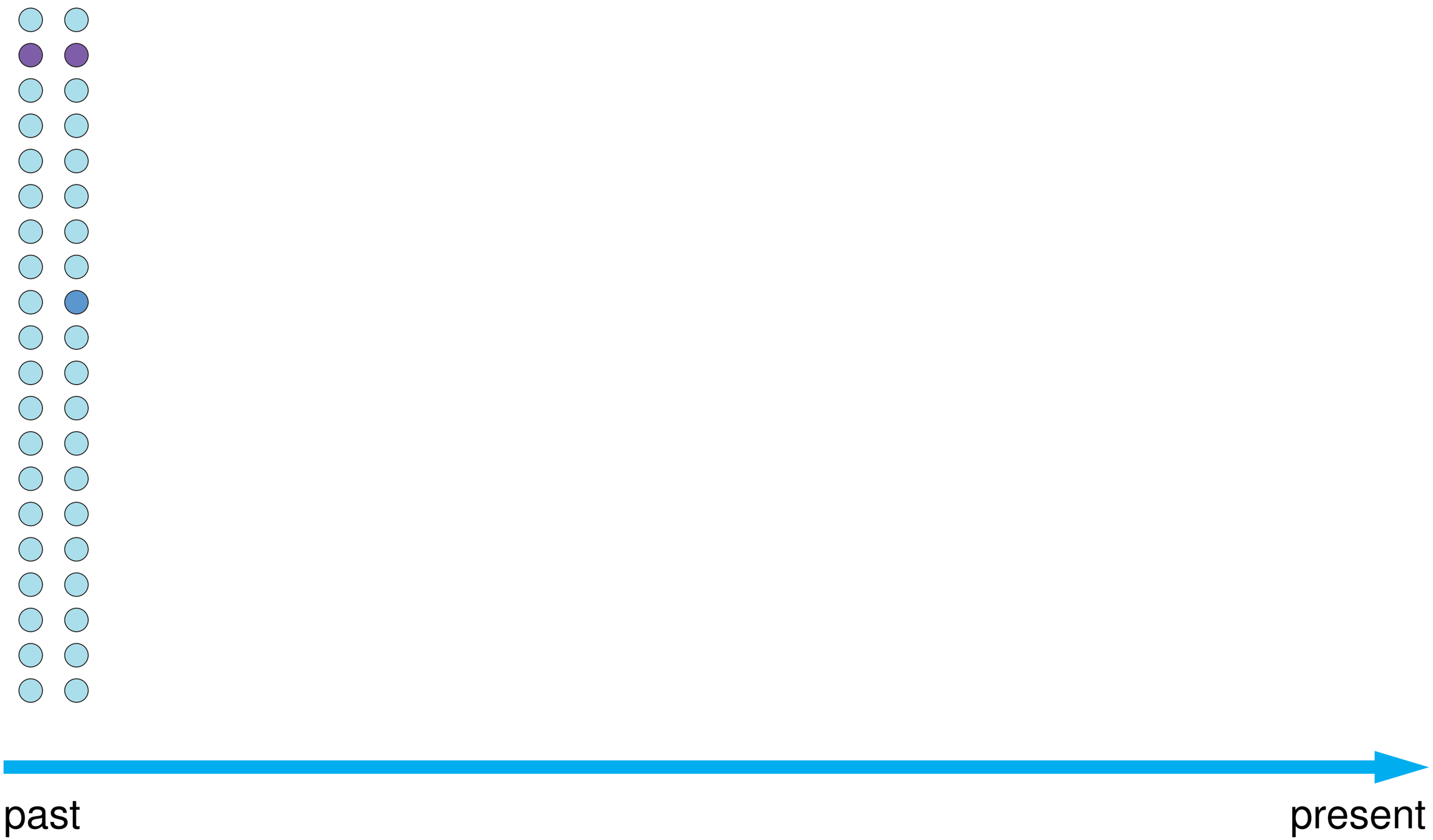
Population models

Fisher, Wright



Population models

Fisher, Wright

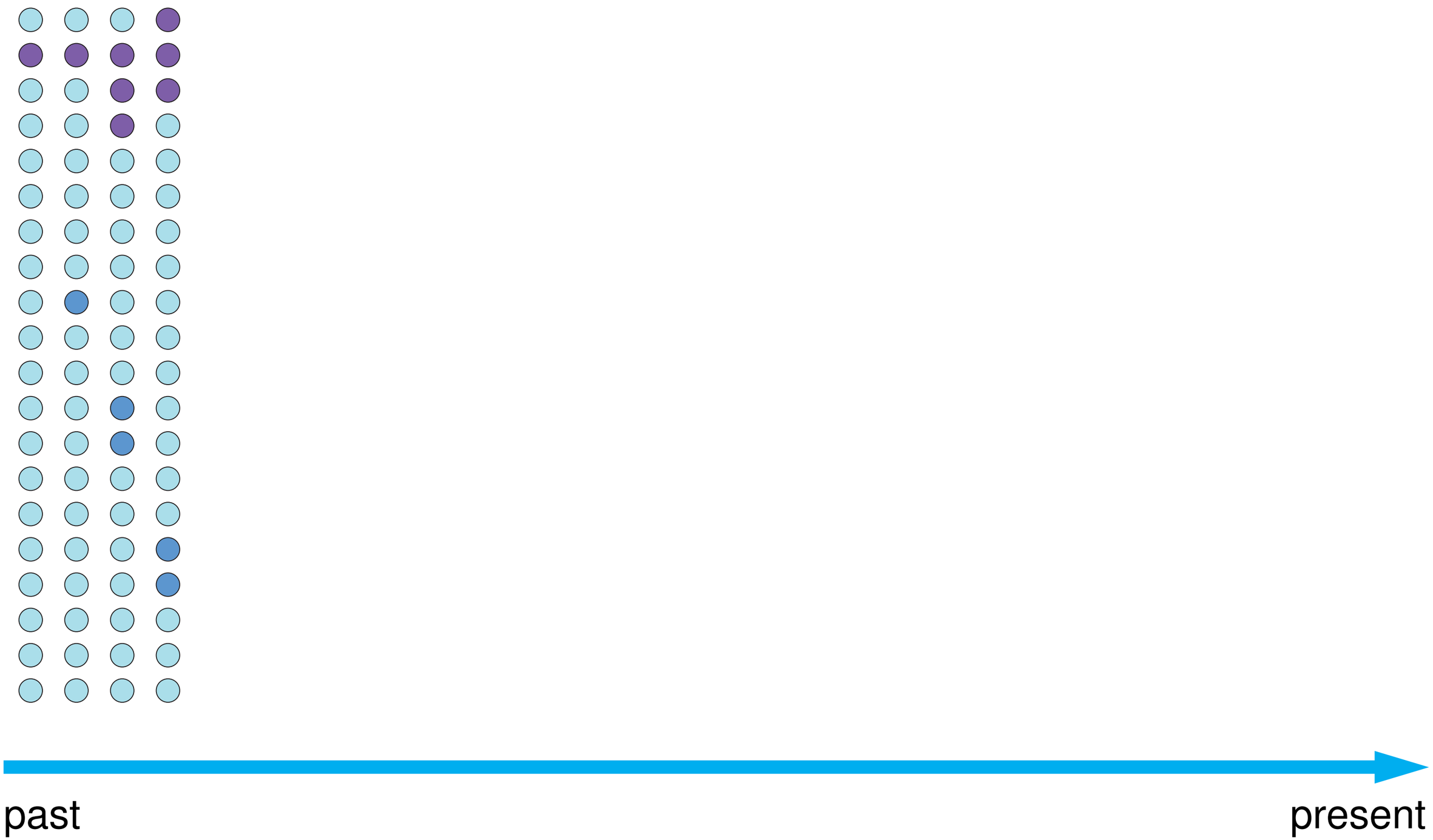


Population models

Fisher, Wright

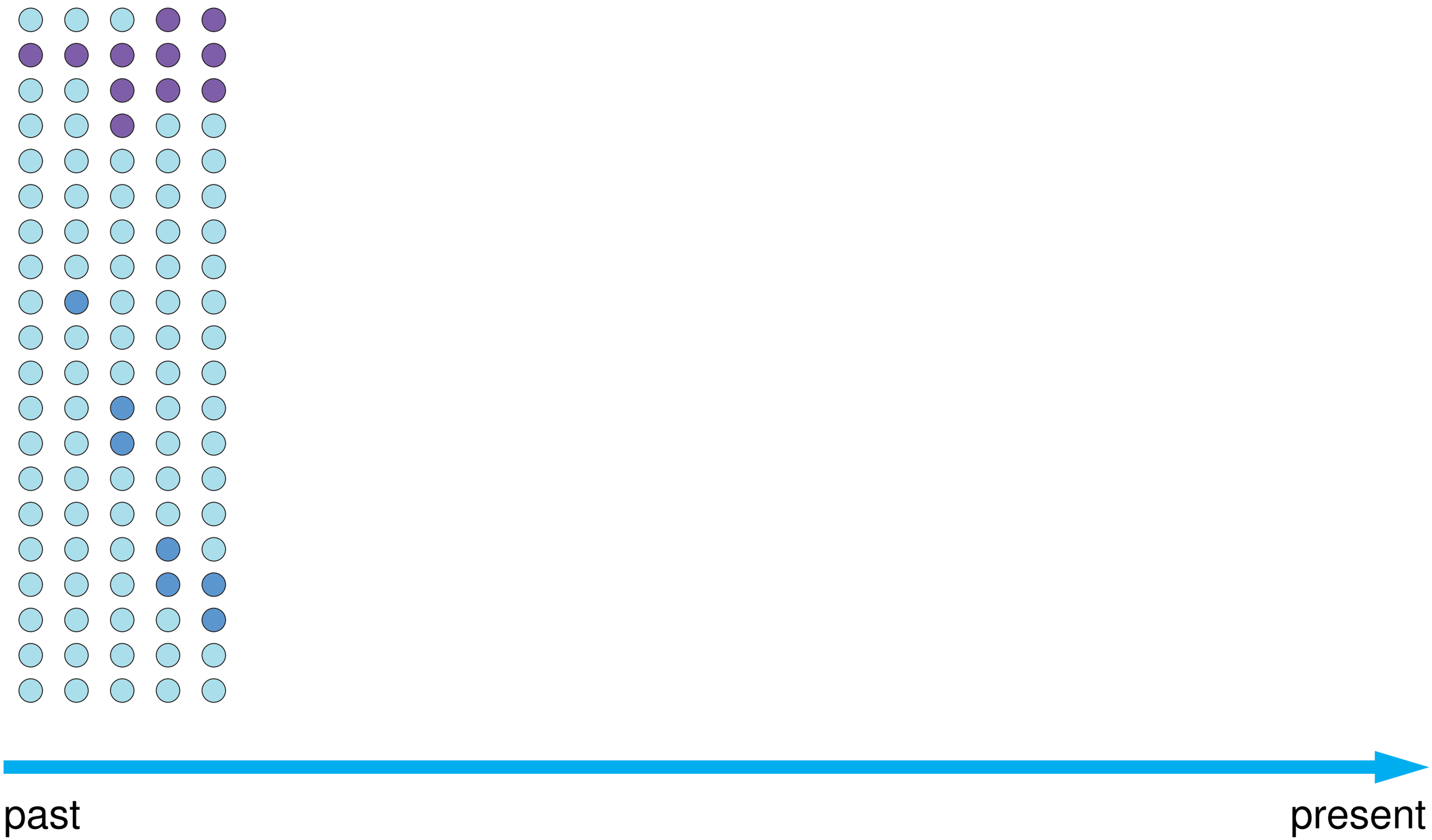
Population models

Fisher, Wright



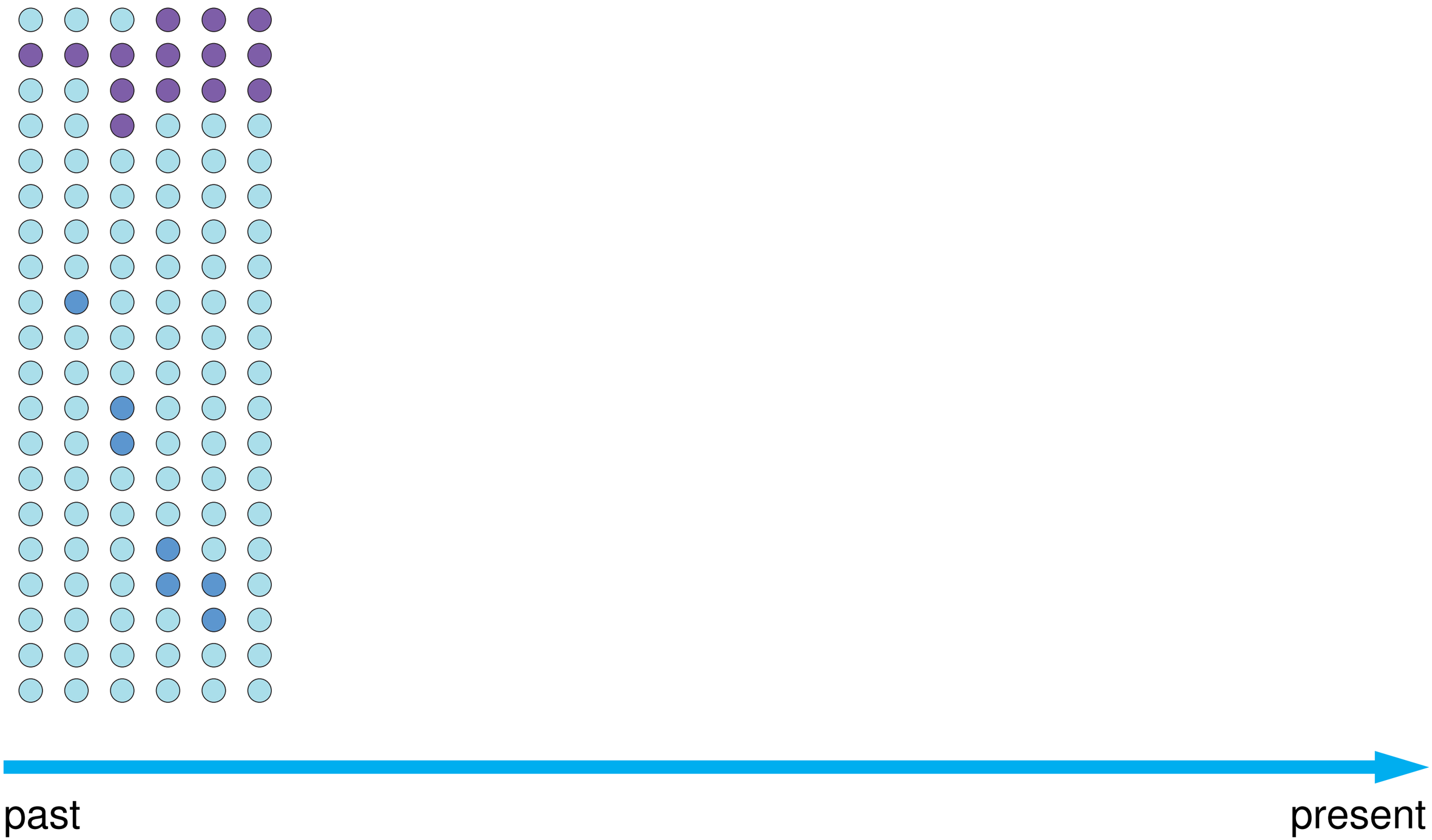
Population models

Fisher, Wright



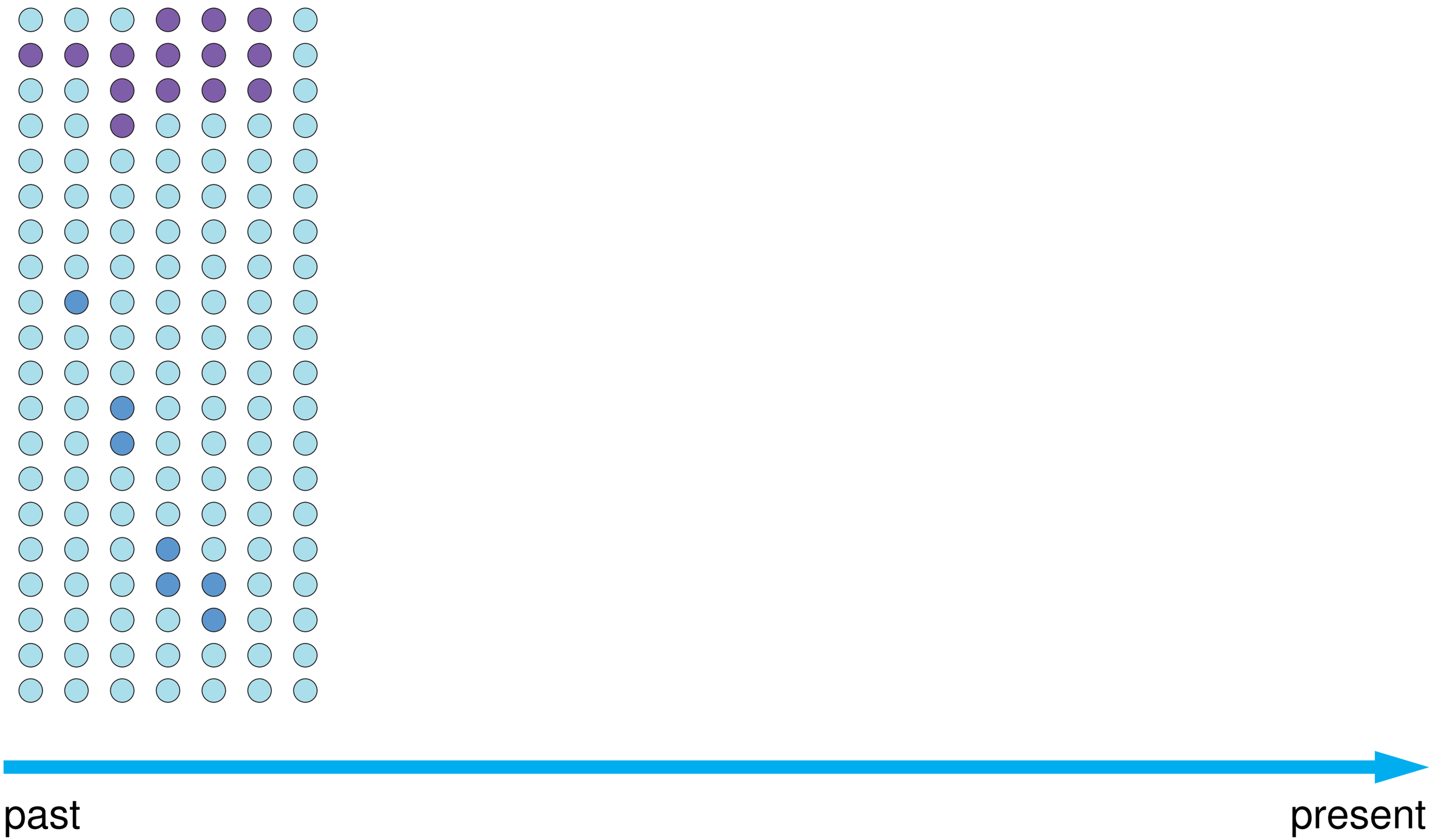
Population models

Fisher, Wright



Population models

Fisher, Wright

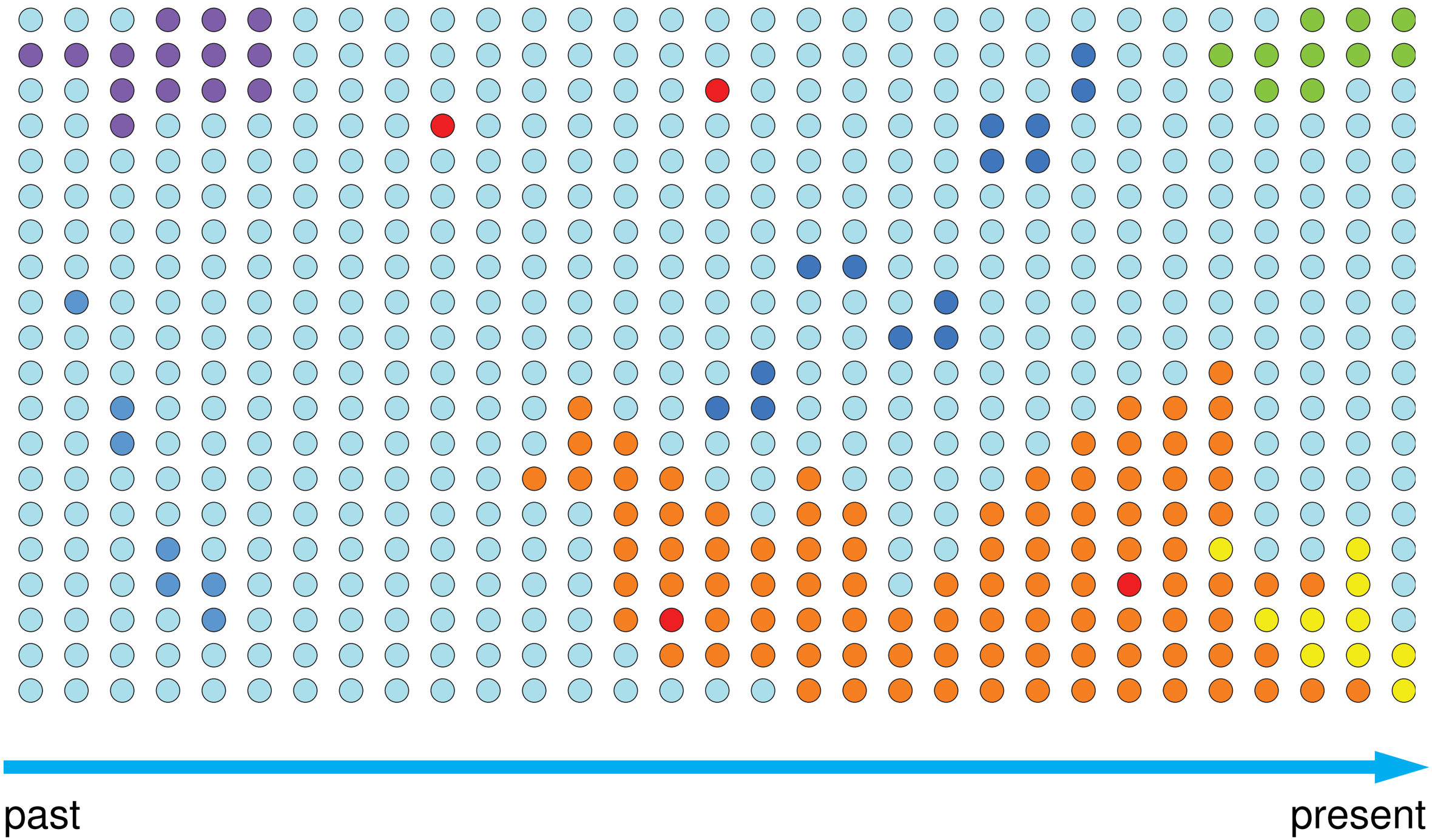


Population models

Fisher, Wright

Population models

Fisher, Wright



Population models

Wright

$$N_t = f(X, \mu, N_{t-1})$$

$$\theta = g(X)$$

$$F = h(\theta) = h(g(X))$$

$$F = h(\theta) \approx \frac{1+\theta}{1+2\theta}$$

$$F = s(X) \approx 1 - \frac{H_{\text{obs}}}{H_{\text{exp}}}$$

where $\theta = 4N\mu$ and F is the probability that 2 sampled allele from X are identical by descent. H is the expected or observed heterozygosity.

F-statistic is still the most commonly used method to infer population genetic parameters from allele frequency data.

F-statistic inference

Data	1	2
Aa	24	1
Aa	30	13
aa	46	86
	100	100

$$F = \frac{1+\theta}{1+2\theta}$$

$$F = 1 - \frac{H}{2pq}$$

$$p_1 = \frac{24 + \frac{1}{2}30}{100} \approx 0.39 \qquad q_1 = 1 - p_1 = 0.61$$

$$p_2 = \frac{1}{100} + \frac{1}{2}\frac{13}{100} = 0.075 \qquad q_2 = 0.925$$

$$F_1 = 1 - \frac{30/100}{2 \cdot 0.61 \cdot 0.39} \cong 1 - 0.48 = 0.52$$

$$\theta_1: \; 0.52 = \frac{1+\theta}{1+2\theta}$$

$$-1 + 0.52(1+2\theta) = \cancel{1+}\theta$$

$$-1 + 0.52 + 1.04\theta = \theta \qquad | -\theta$$

$$-0.48 + 0.04\theta = 0$$

$$\theta_1 = \frac{0.48}{0.04} = 12$$

$$F_2 = 1 - \frac{13/100}{2 \cdot 0.925 \cdot 0.075} \cong 1 - \frac{0.06}{0.0925} \cong 1 - \frac{1}{11} \cong 0.9$$

$$\theta_2: \; 0.9 = \frac{1+\theta}{1+2\theta} \rightarrow -1 + 0.9 + 1.8\theta = \theta$$

$$-0.1 + 1.8\theta = \theta$$

$$\theta_2 = \frac{0.1}{1.8} = \frac{1}{18}$$

$$\boxed{\theta_1 \gg \theta_2}$$

FE127-07

F-statistic inference

Example

Data	1	2
AH	24	1
Aa	30	13
aa	46	86
	100	100

$$F = \frac{1+\theta}{1+2\theta}$$

$$F = 1 - \frac{H}{2pq}$$

$$p_1 = \frac{24 + \frac{1}{2}30}{100} \approx 0.39 \qquad q_1 = 1 - p_1 = 0.61$$

$$p_2 = \frac{1}{100} + \frac{1}{2}\frac{13}{100} = 0.075 \qquad q_2 = 0.925$$

$$F_1 = 1 - \frac{30/100}{2 \cdot 0.61 \cdot 0.39} \cong 1 - 0.48 = 0.52$$

$$\theta_1: \; 0.52 = \frac{1+\theta}{1+2\theta}$$

$$-1 + 0.52(1+2\theta) = \cancel{1+}\theta$$

$$-1 + 0.52 + 1.04\theta = \theta$$

$$-0.48 + 0.04\theta = 0$$

$$\theta_1 = \frac{0.48}{0.04} = 12$$

$$F_2 = 1 - \frac{13/100}{2 \cdot 0.925 \cdot 0.075} \cong 1 - \frac{0.06}{0.0925}$$

$$\cong 1 - \frac{1}{11} \cong 0.9$$

$$\theta_2: \; 0.9 = \frac{1+\theta}{1+2\theta} \rightarrow -1 + 0.9 + 1.8\theta = \theta$$

$$-0.1 + 1.8\theta = \theta$$

$$\theta_2 = \frac{0.1}{1.8} = \frac{1}{18}$$

$$\boxed{\theta_1 \gg \theta_2}$$

FE127-07

F-statistic inference

Example

Data	1	2
AH	24	1
Aa	30	13
aa	46	86
	100	100

$$F = \frac{1+\theta}{1+2\theta}$$

$$F = 1 - \frac{H}{2pq}$$

$$p_1 = \frac{24 + \frac{1}{2}30}{100} \approx 0.39 \qquad q_1 = 1 - p_1 = 0.61$$

$$p_2 = \frac{1}{100} + \frac{1}{2}\frac{13}{100} = 0.075 \qquad q_2 = 0.925$$

$$F_1 = 1 - \frac{30/100}{2 \cdot 0.61 \cdot 0.39} \cong 1 - 0.48 = 0.52$$

$$\theta_1: \; 0.52 = \frac{1+\theta}{1+2\theta}$$

$$-1 + 0.52(1+2\theta) = \cancel{1+}\theta$$

$$-1 + 0.52 + 1.04\theta = \theta$$

$$-0.48 + 0.04\theta = 0$$

$$\Rightarrow \theta_1 = \frac{0.48}{0.04} = 12$$

$$F_2 = 1 - \frac{13/100}{2 \cdot 0.925 \cdot 0.075} \cong 1 - \frac{0.06}{0.0925}$$

$$\cong 1 - \frac{1}{11} \cong 0.9$$

$$\theta_2: \; 0.9 = \frac{1+\theta}{1+2\theta} \rightarrow -1 + 0.9 + 1.8\theta = \theta$$

$$-0.1 + 1.8\theta = \theta$$

$$\Rightarrow \theta_2 = \frac{0.1}{1.8} = \frac{1}{18}$$

$$\boxed{\theta_1 \gg \theta_2}$$

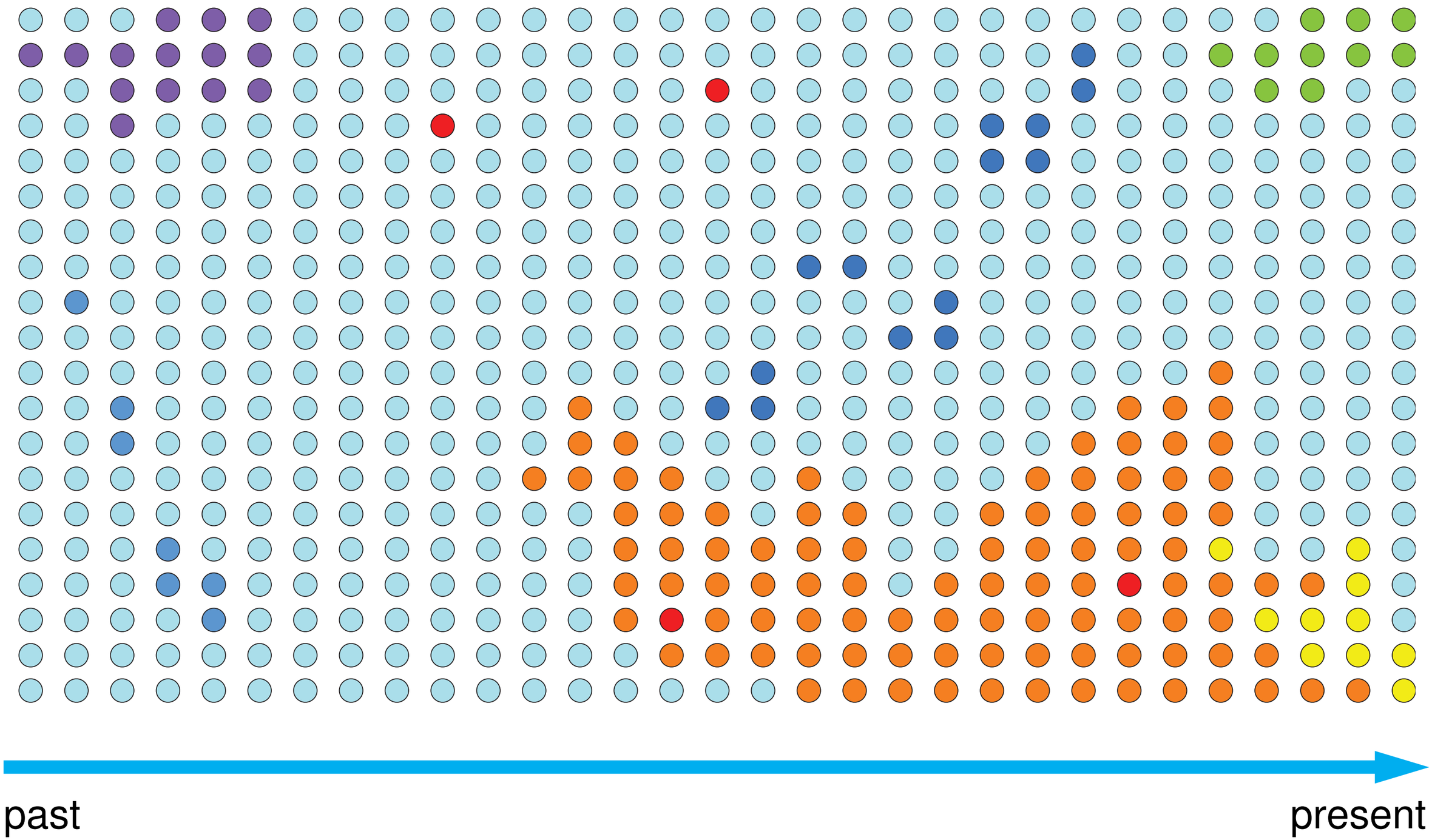
FE127-07

Computation device: IBM PC

1 generation ago

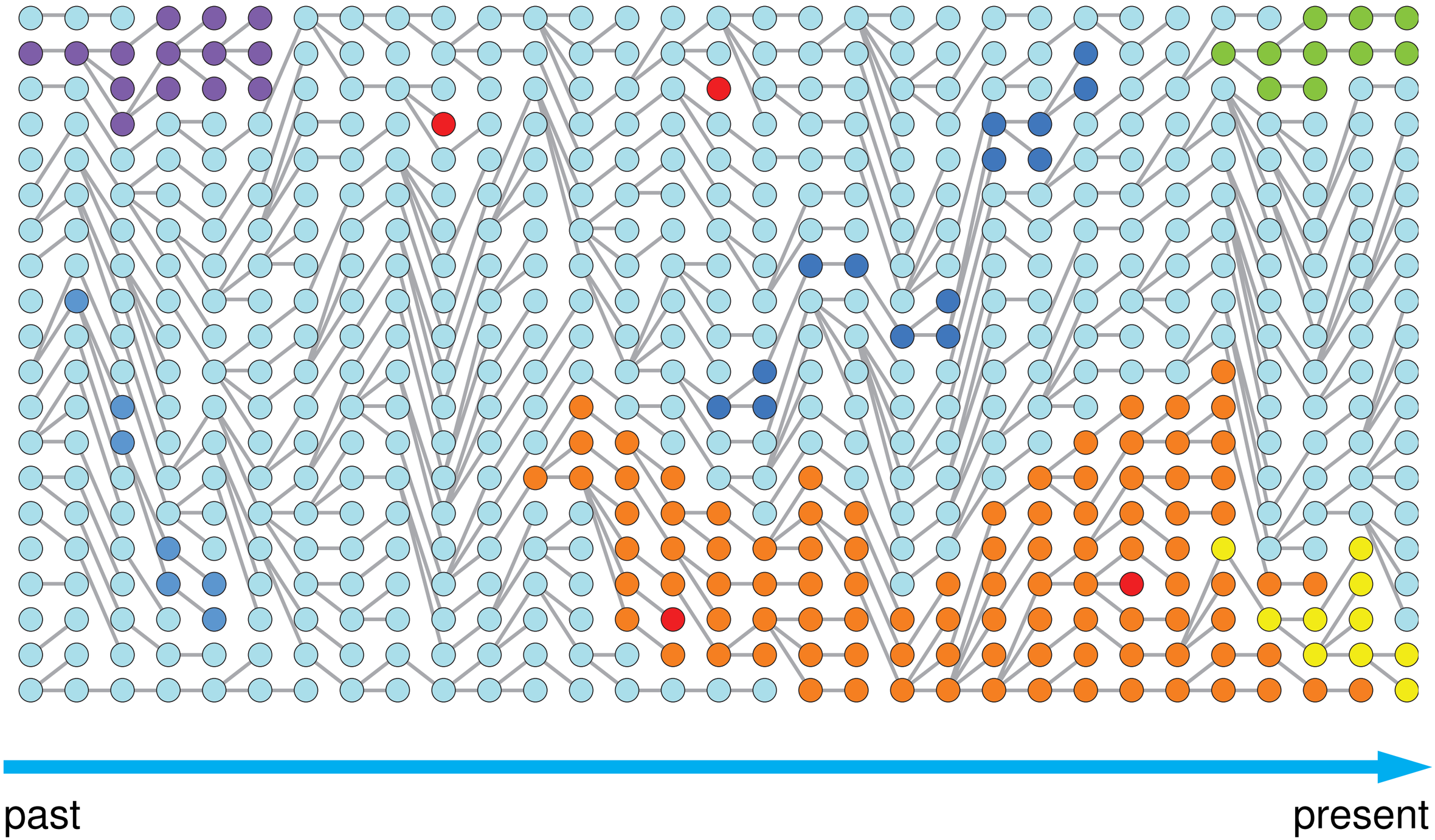
Population models

Fisher, Wright



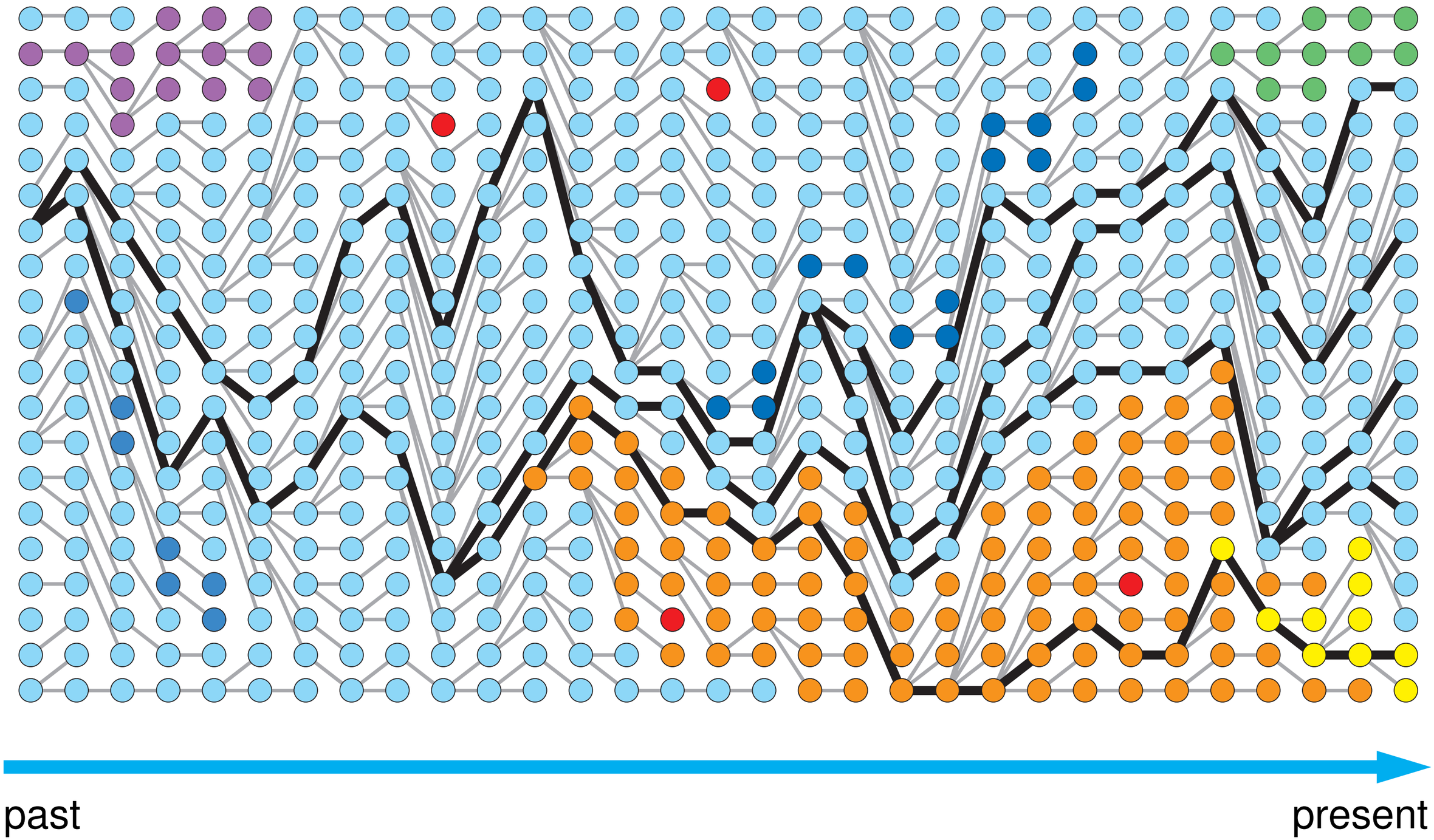
Population models

Fisher, Wright



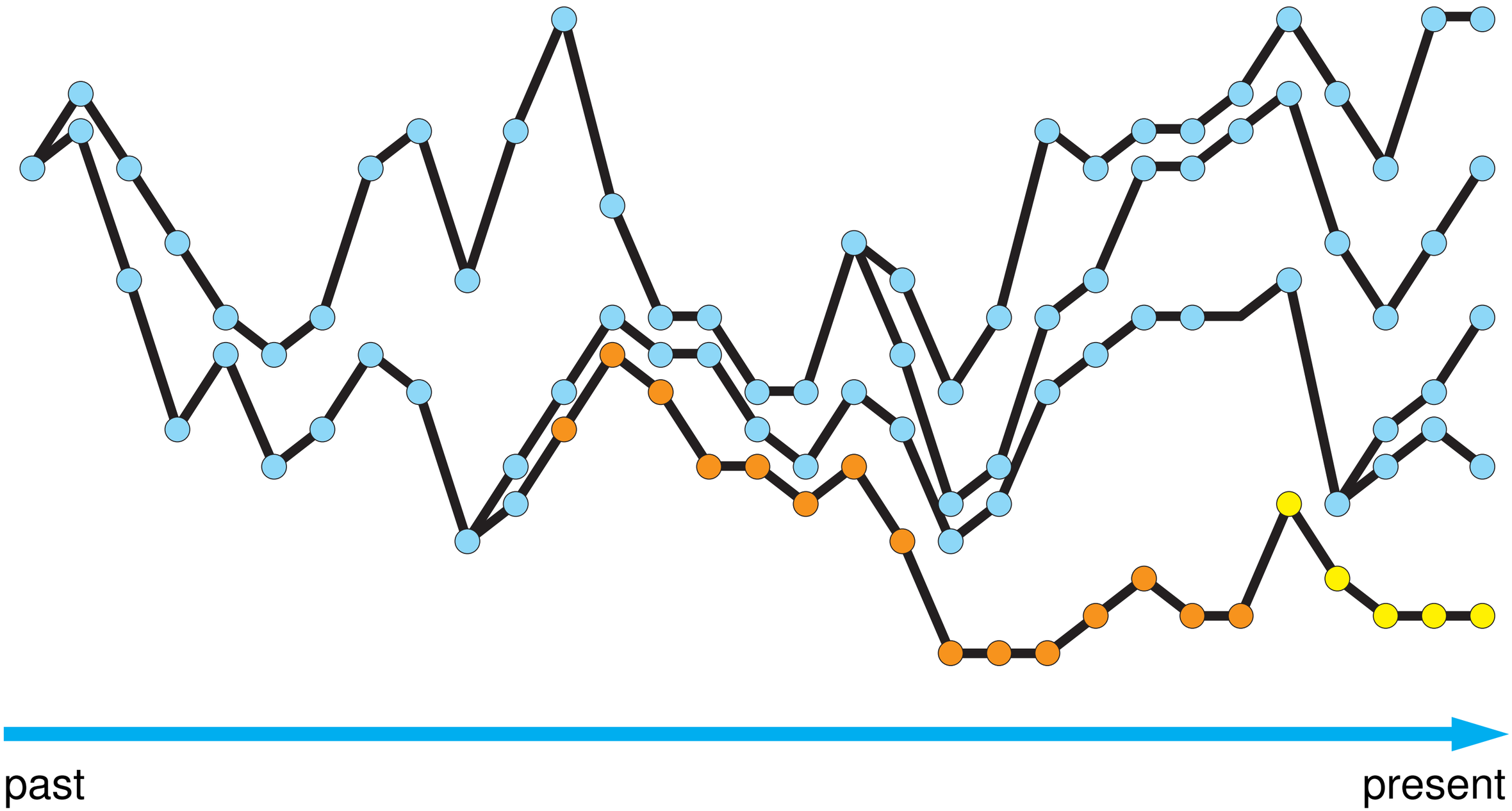
Population models

Fisher, Wright



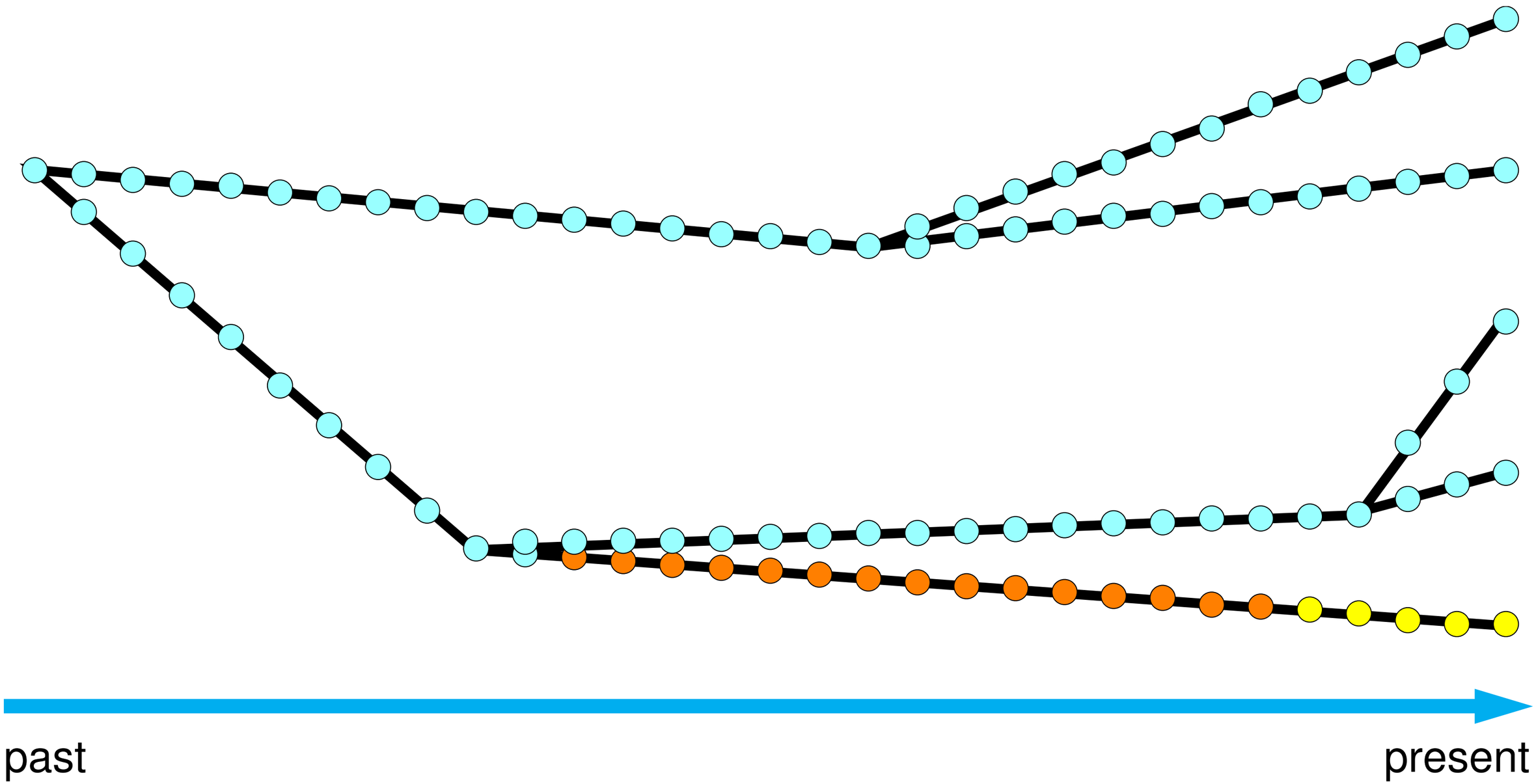
Population models

Fisher, Wright



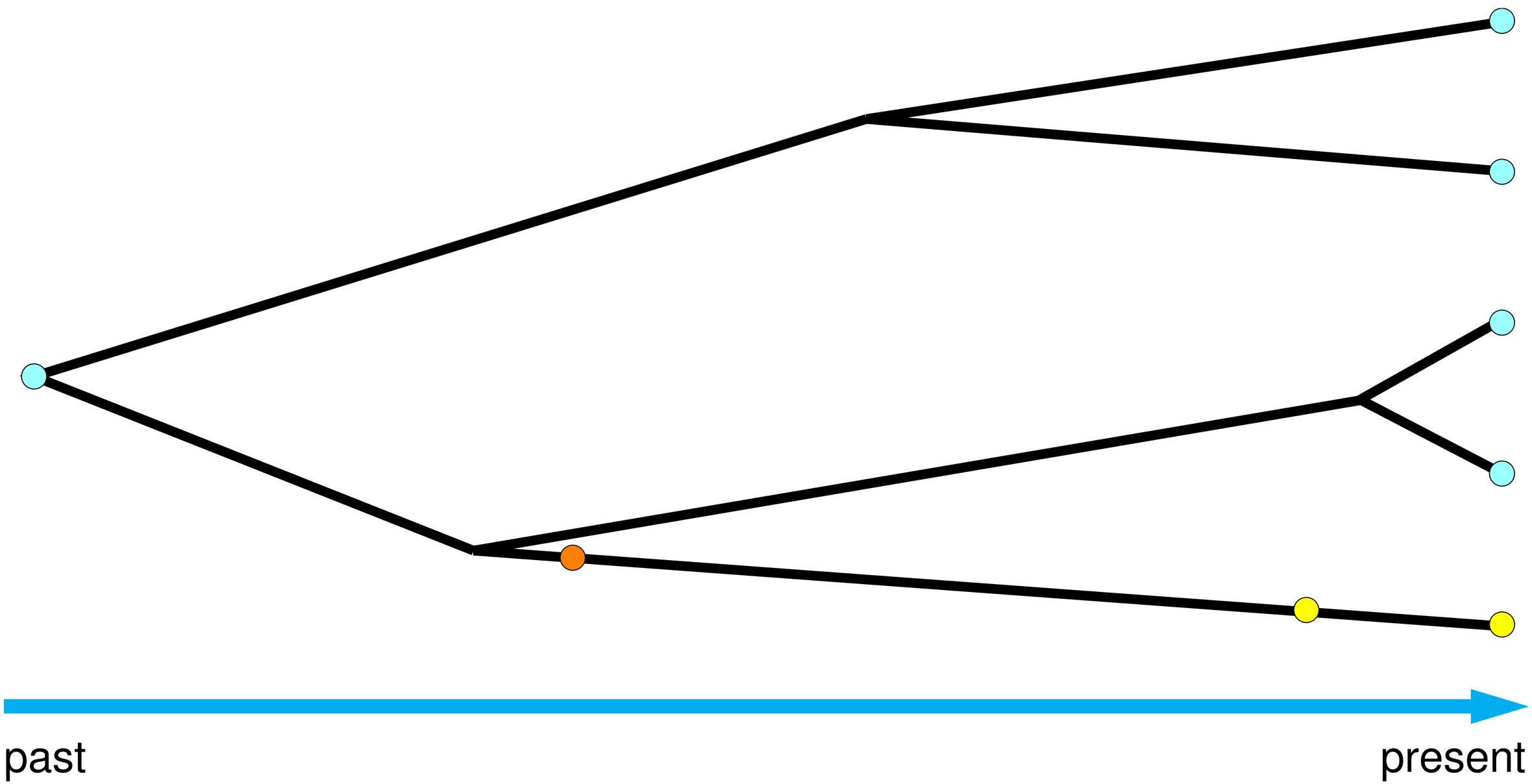
Population models

Fisher, Wright



Population models

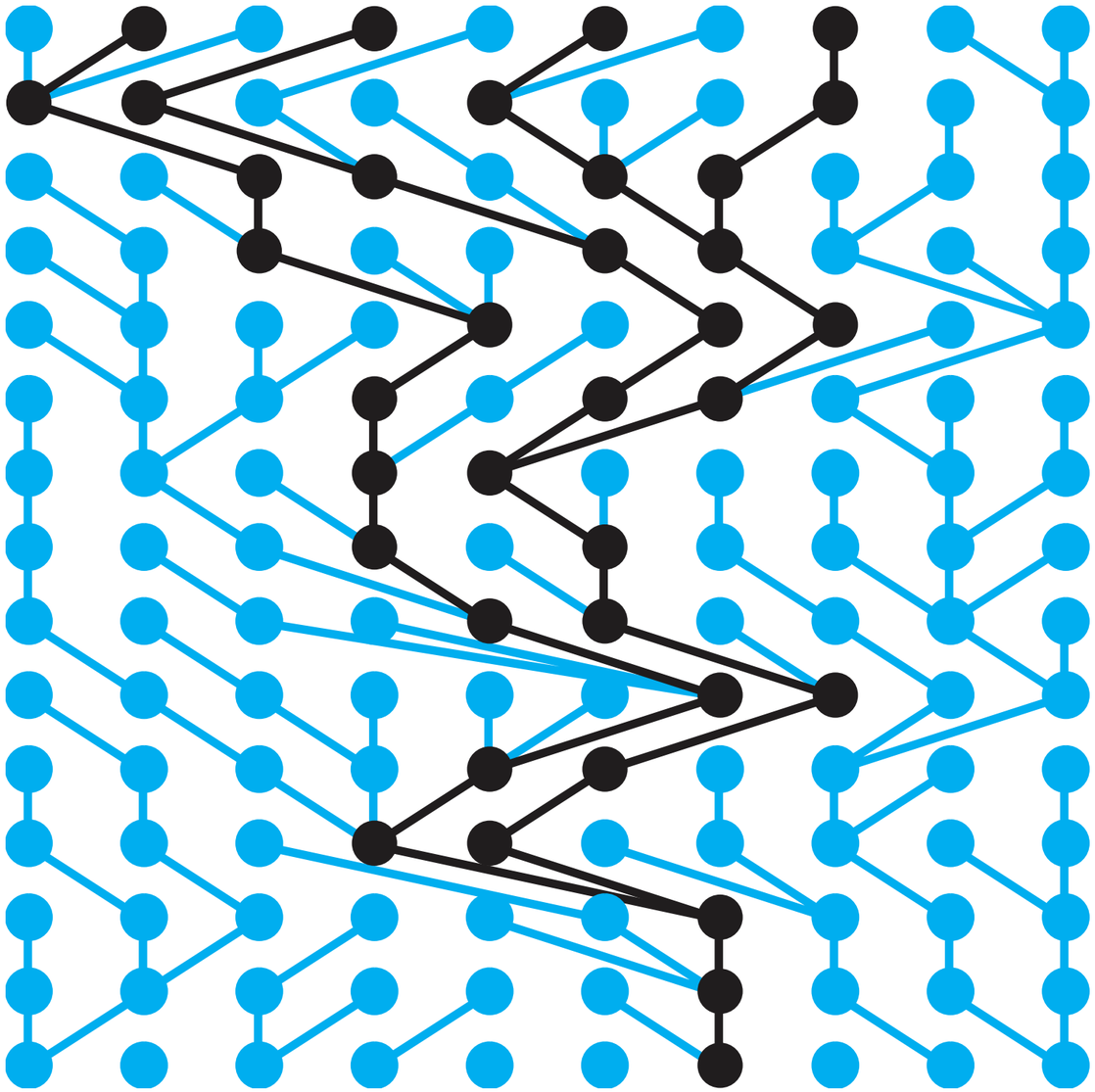
Fisher, Wright



Coalescence theory

Kingman

Present

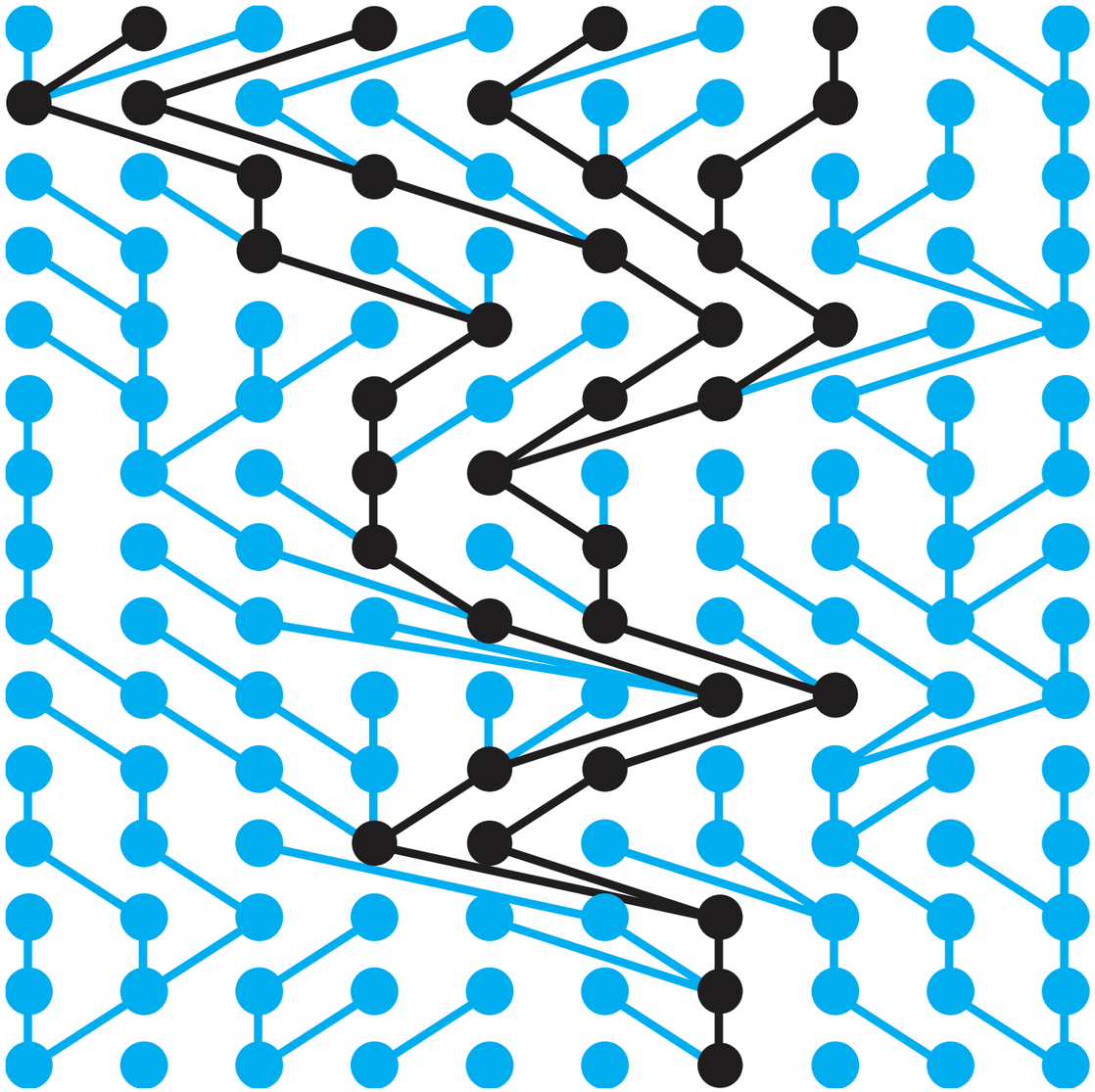


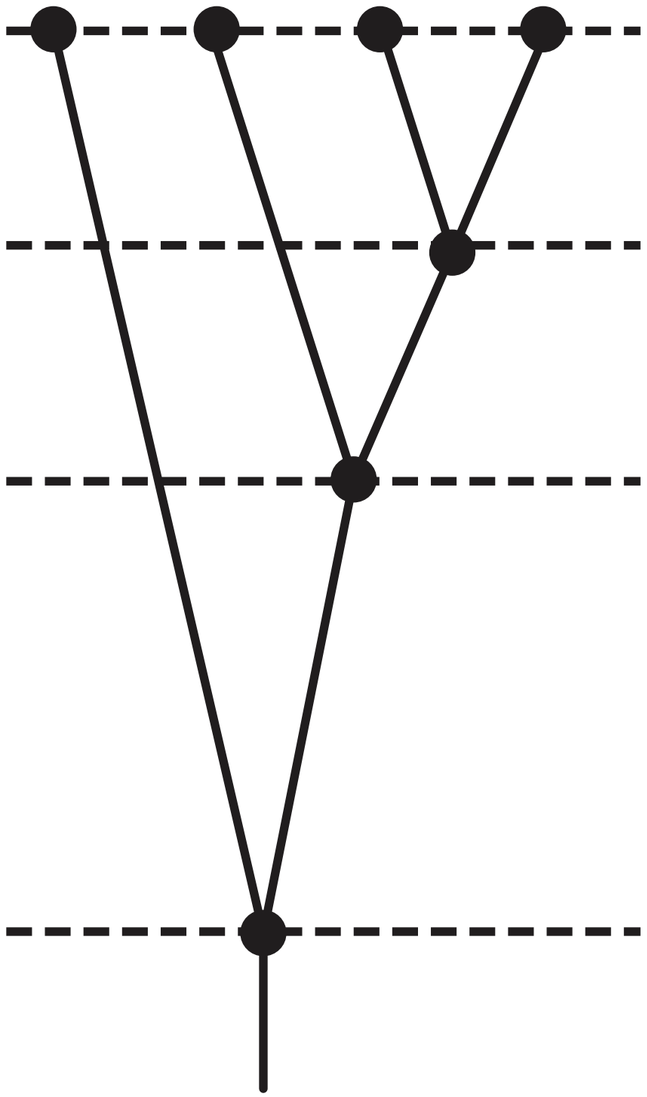
Past

Coalescence theory

Kingman

Present

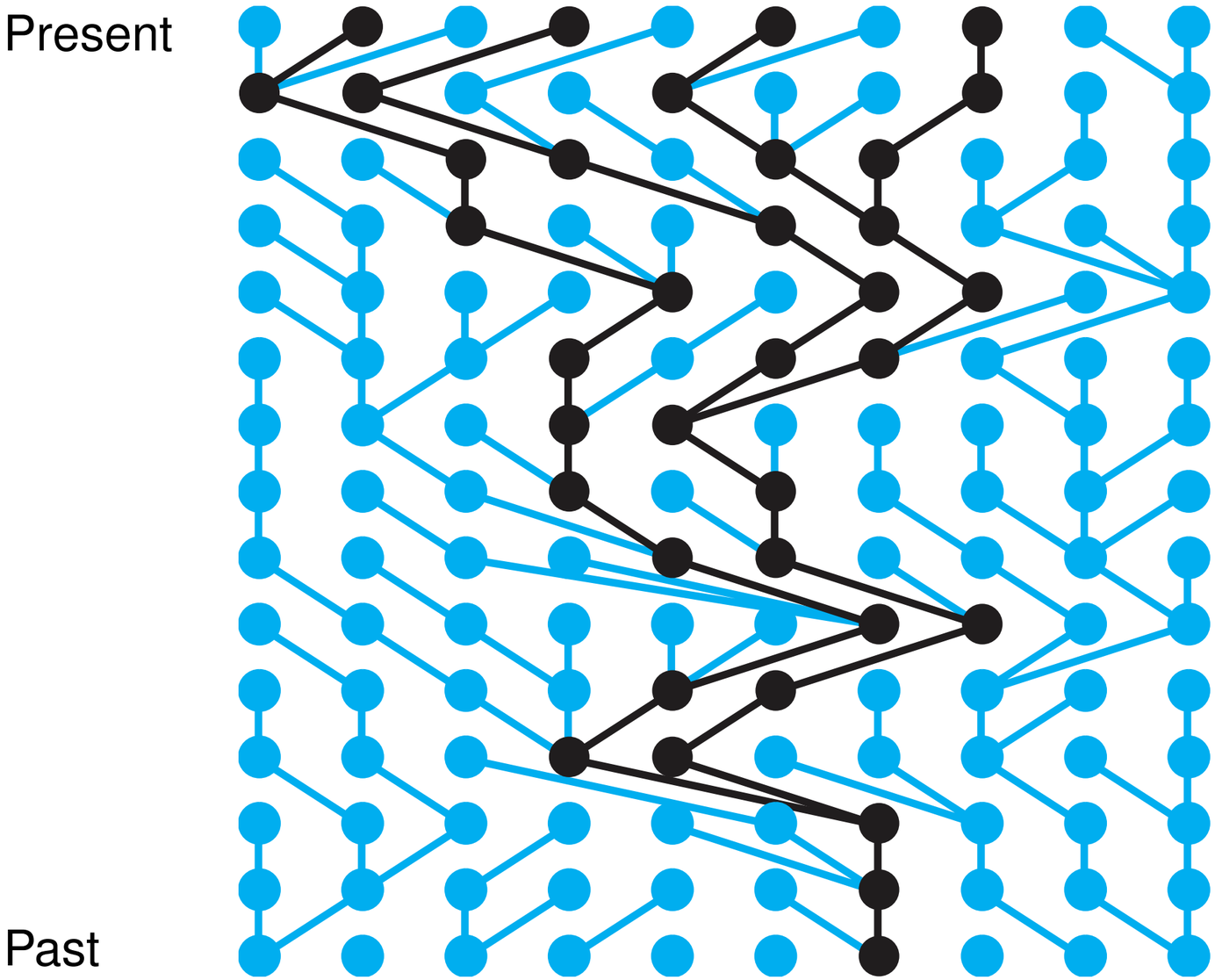


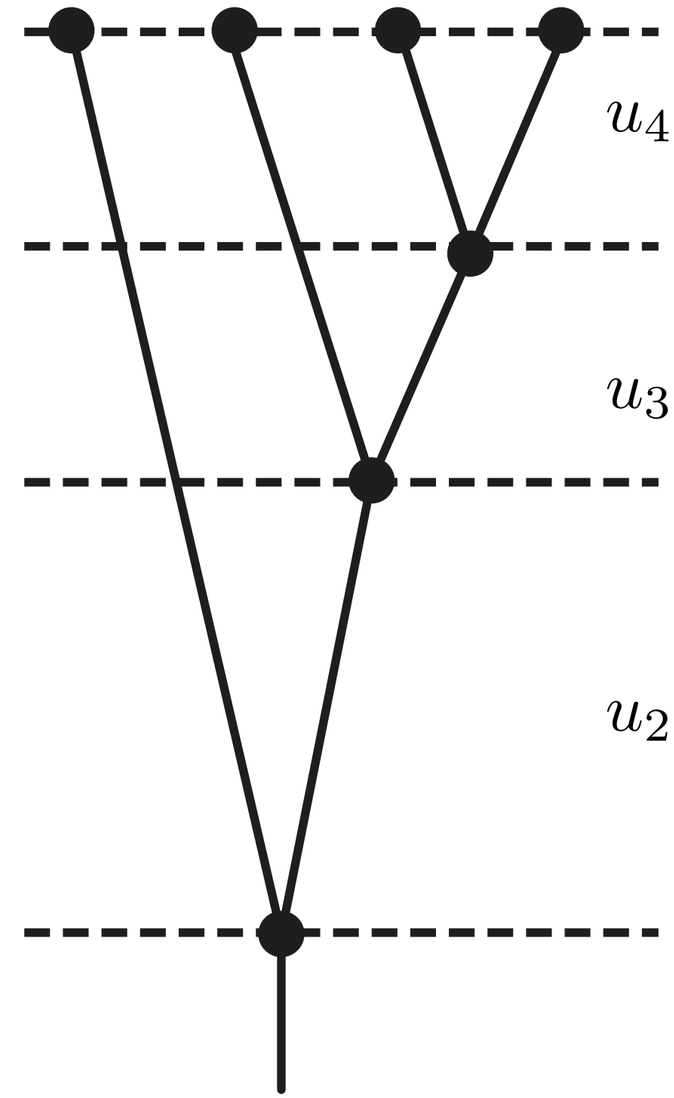


Past

Coalescence theory

Kingman



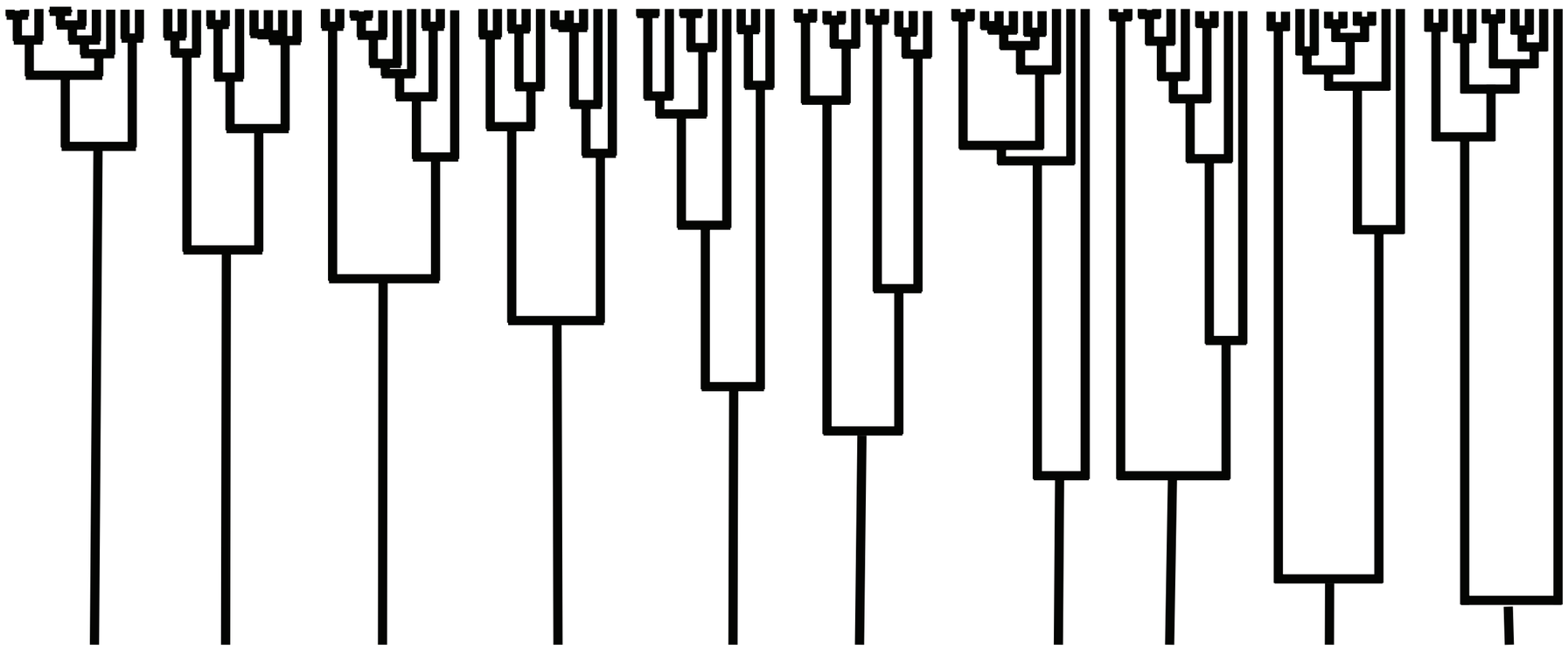


The distribution of time intervals u_k follows an exponential distribution with

$$\mathbb{E}(u_k) = \frac{\Theta}{k(k-1)}$$

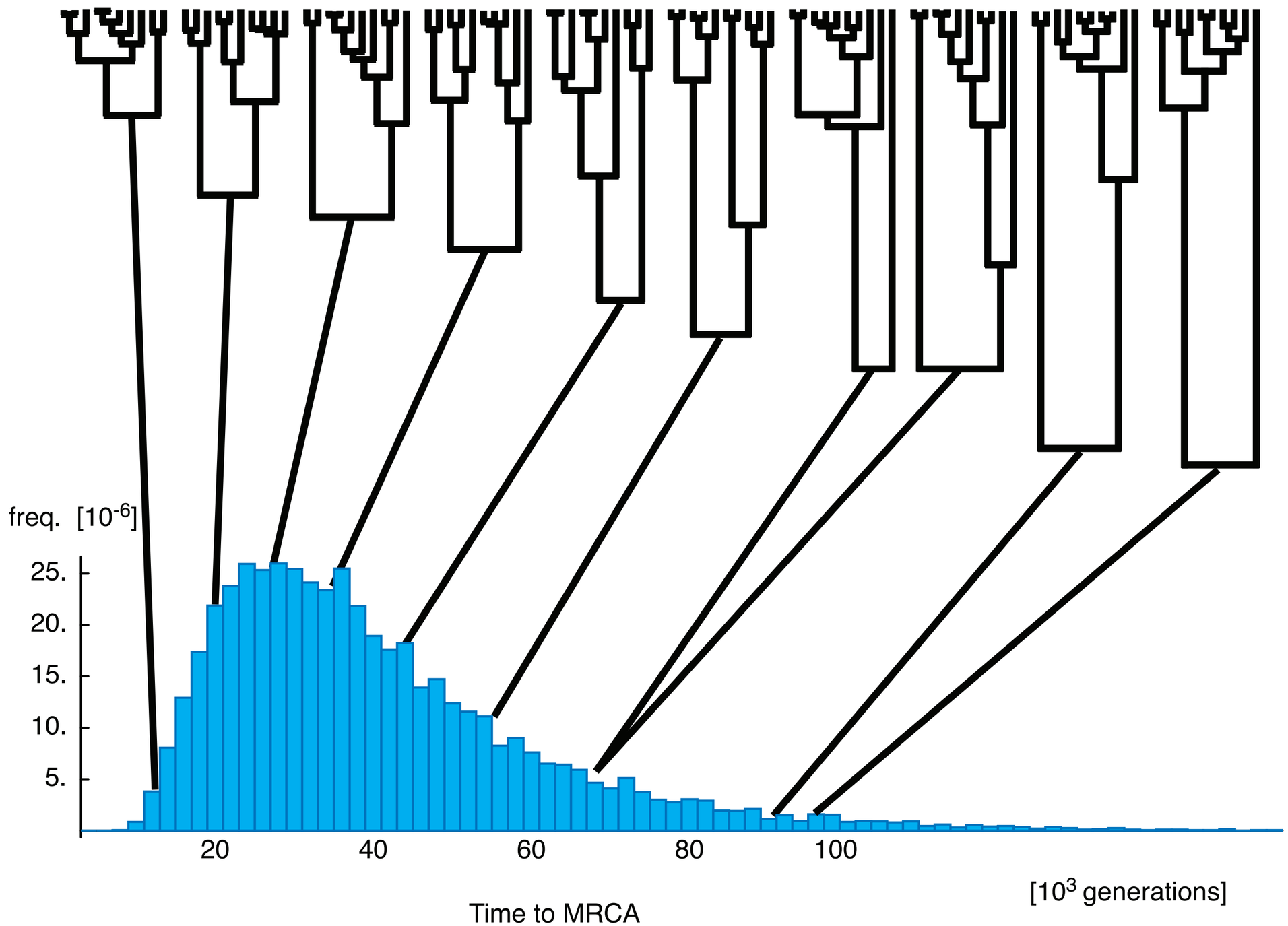
$$\mathrm{p(G \mid \Theta, n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k} \frac{\mathrm{k(k-1)}}{\Theta}\right) \frac{2}{\Theta}$$

Variability of the coalescent process



All genealogies were simulated with the same population size $N_e = 10,000$

Variability of the coalescent process



MRCA = most recent common ancestor (last node in the genealogy)

Coalescence applied

For a likelihood estimate we want to calculate the probability of the data given the model parameters $\mathrm{Prob}(\mathrm{X}|\text{model})$.

Coalescent	to describe the population genetic processes.
Mutation model	to describe the change of genetic material over time.

We find the maximum likelihood estimate by maximizing the likelihood function with respect to the parameters of interest.

$$L(\hat{\Theta}) = \max_{\Theta} L(\Theta)$$

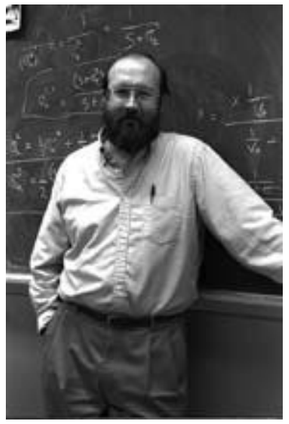
Maximum likelihood inference

$$p(X|\boldsymbol{\Theta}) = \int_G \mathrm{p(G|\boldsymbol{\Theta})p(X|G)dG}$$

$p(G|\boldsymbol{\Theta})$

The probability of a genealogy given parameters

$\mathrm{p(X|G)}$



The probability of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.

Problem

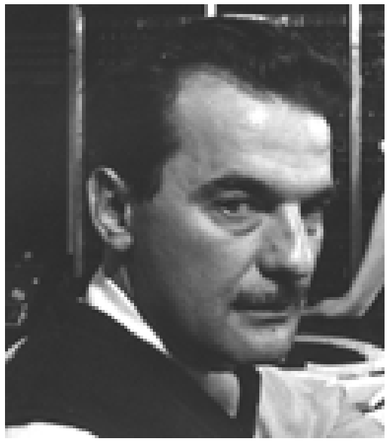
Does not fit on envelope

Tips	Labeled histories
3	3
4	18
5	180
6	2700
7	56700
8	1587600
9	57153600
10	2571912000
15	6958057668962400000
20	564480989588730591336960000000
30	436846661310306951246468019862076389144064000000000000
40	302733382994800735654630336455145720004293943205386250170788872192000000000000
50	3.28632×10^{112}
100	1.37416×10^{284}

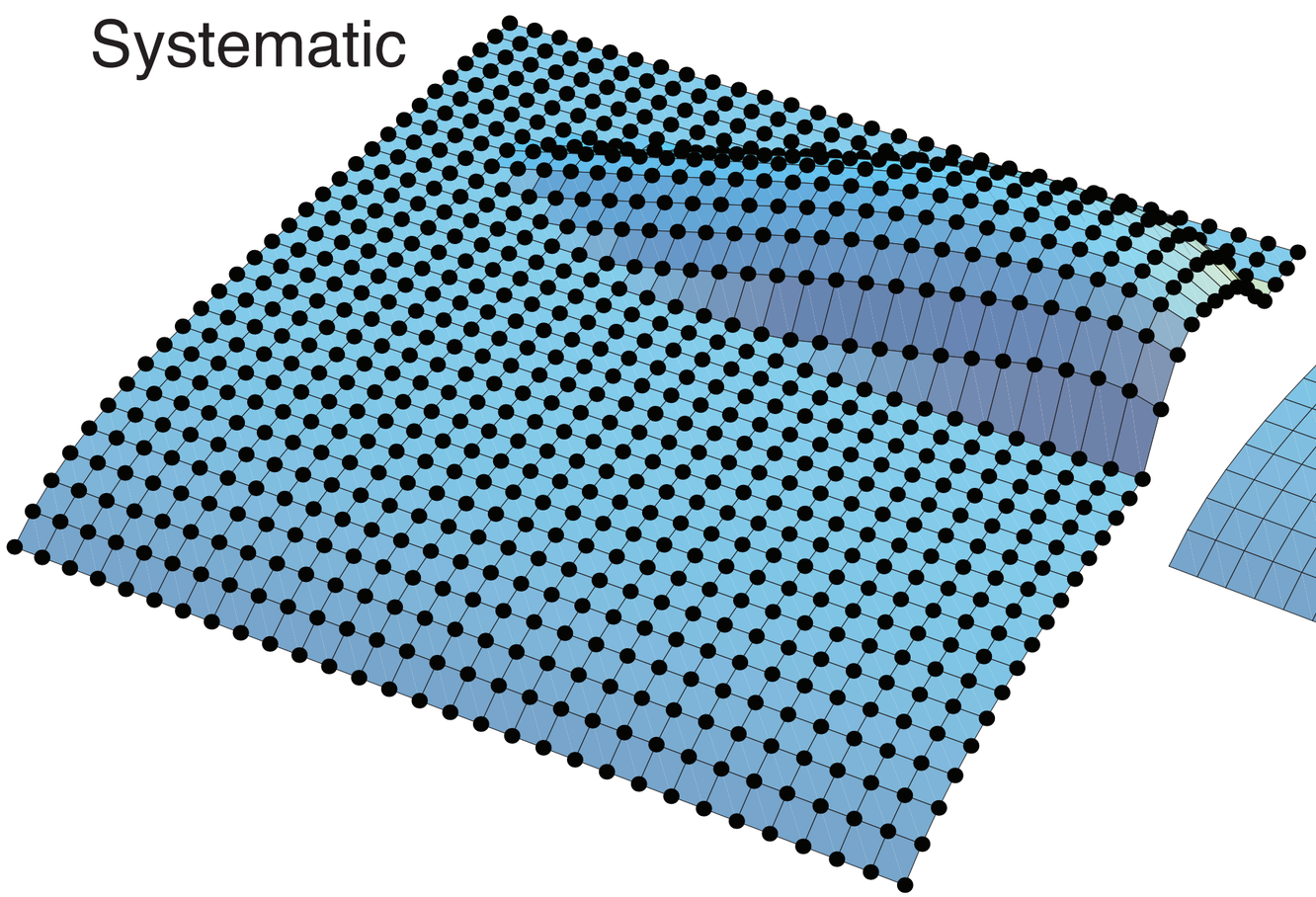
$$L(\Theta) = \int_G p(G|\Theta)\ \mathrm{Prob}(X|G)dG$$

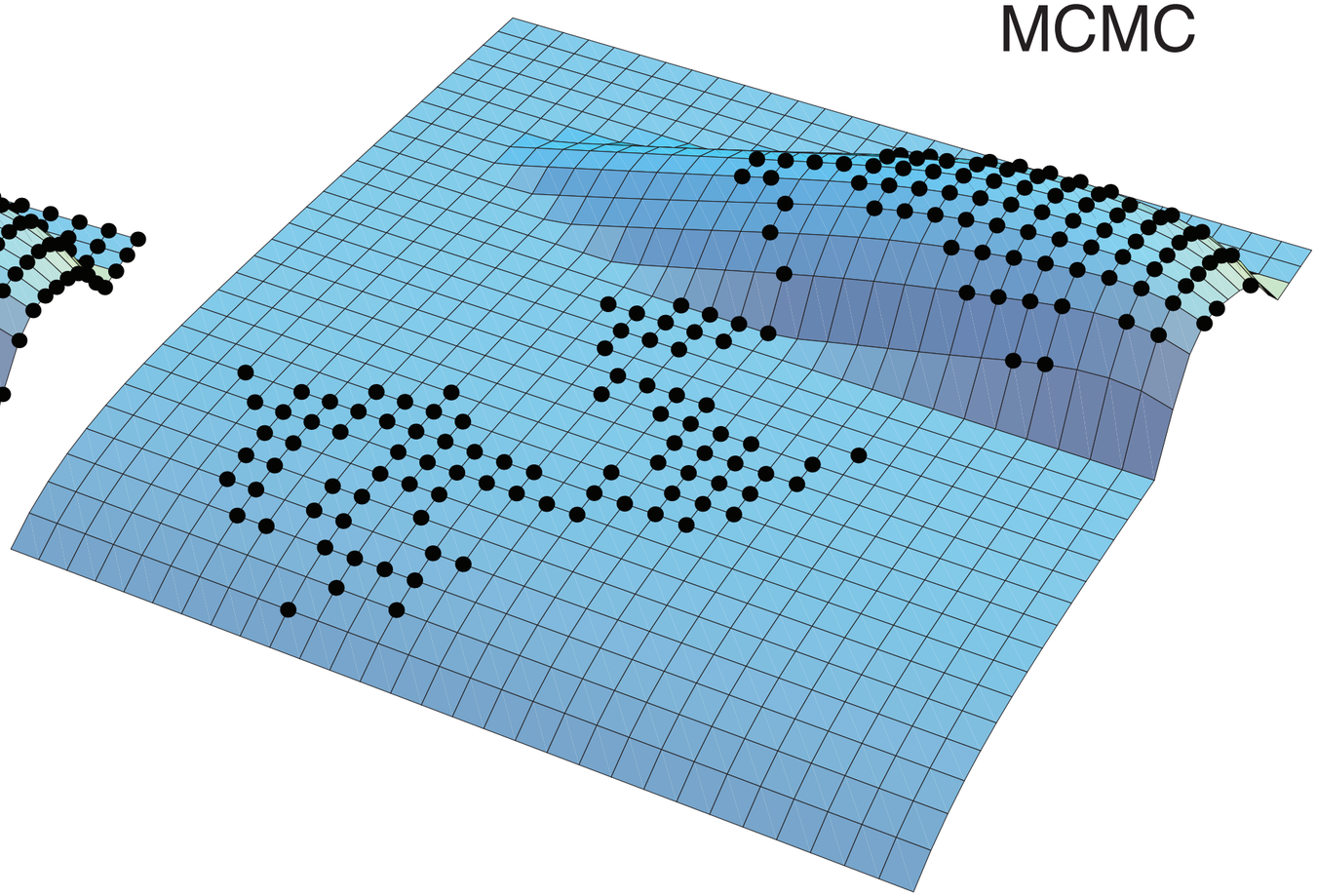
The number of possible genealogies is very large and for realistic data sets, programs need to use Markov chain Monte Carlo methods.

Metropolis-Hastings algorithm



$$\mathrm{L}(\boldsymbol{\Theta}) = \int_G \mathrm{p}(\mathrm{G}|\boldsymbol{\Theta})\ \mathrm{Prob}(\mathrm{X}|\mathrm{G})\mathrm{dG}$$





Computer

Computer

Current research problems in my group

Population genetics using non-Markovian waiting times.

Development of new landscape genetics algorithms

Development of new algorithms to infer recombination hotspots using Hawkes processes.

Gene flow and graph theory.

Unsolved problems

- What to do with large numbers of populations?
- How to incorporate the main population genetic forces into the analysis and still be available to run things? Main forces are recombination, selection, genetic drift, divergence, and mutation.

the end