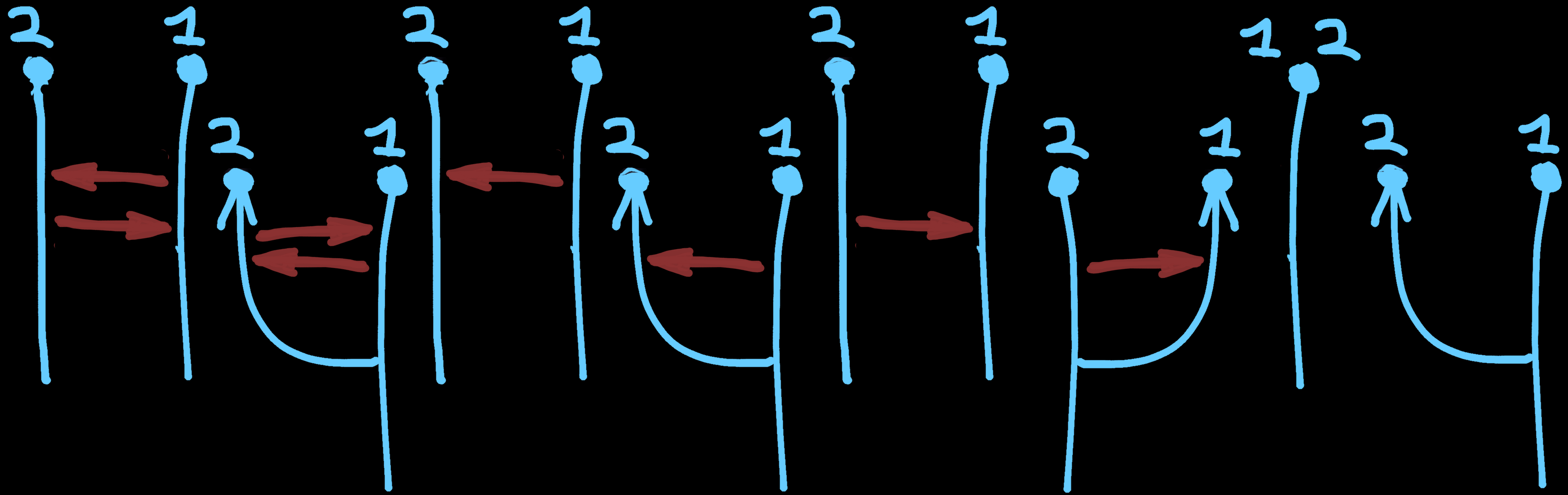
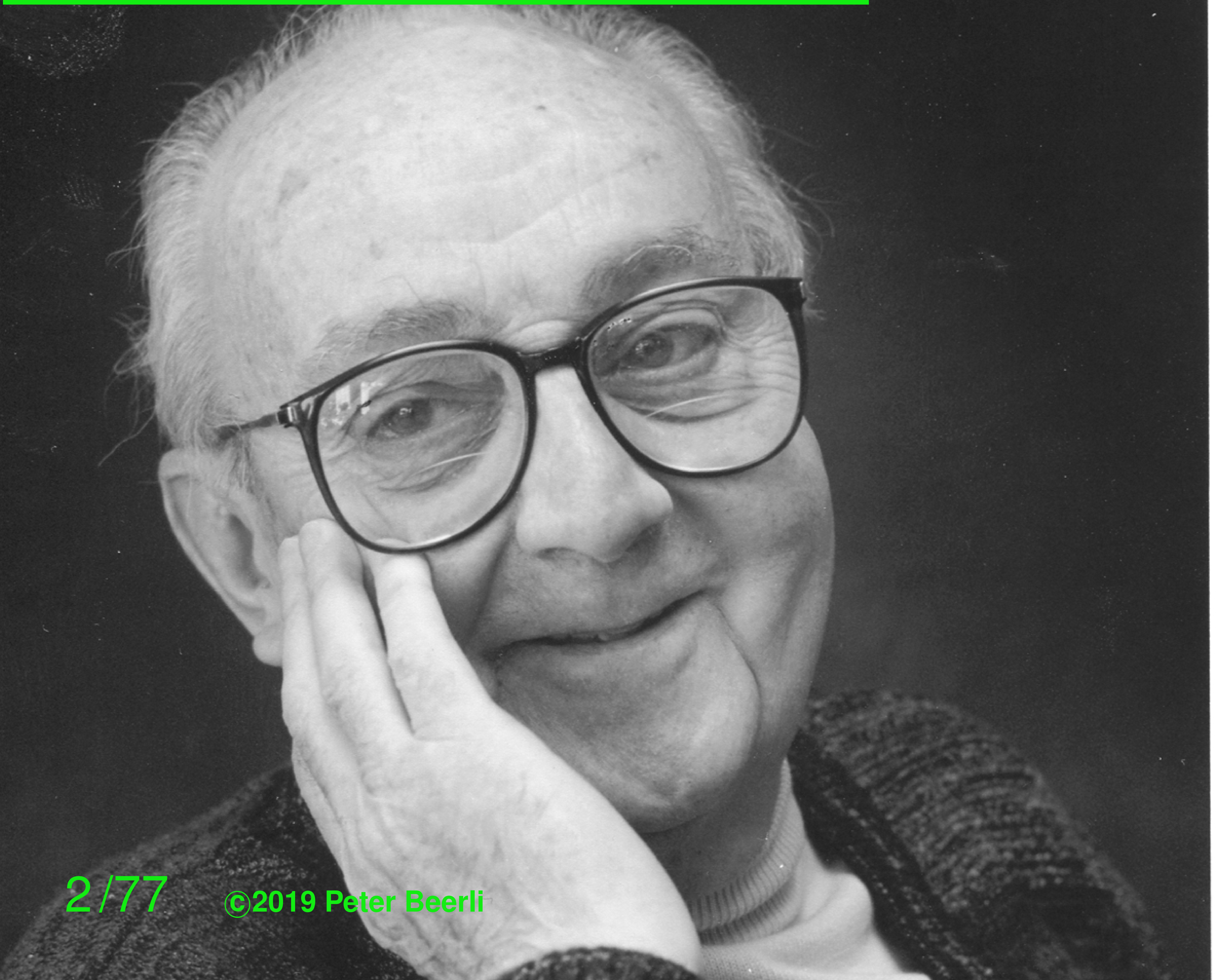
All models are good, but only some are useful



Peter Beerli Scientific Computing, Florida State University Twitter:@peterbeerli

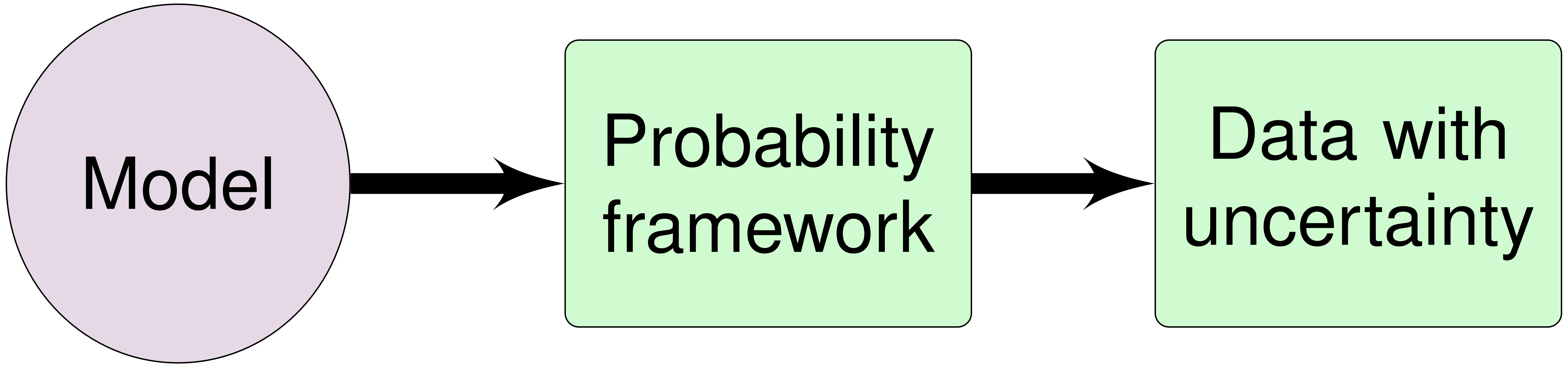
On models



Essentially, all models are wrong, but some are useful.

Box, George E. P.; Norman R. Draper (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley.

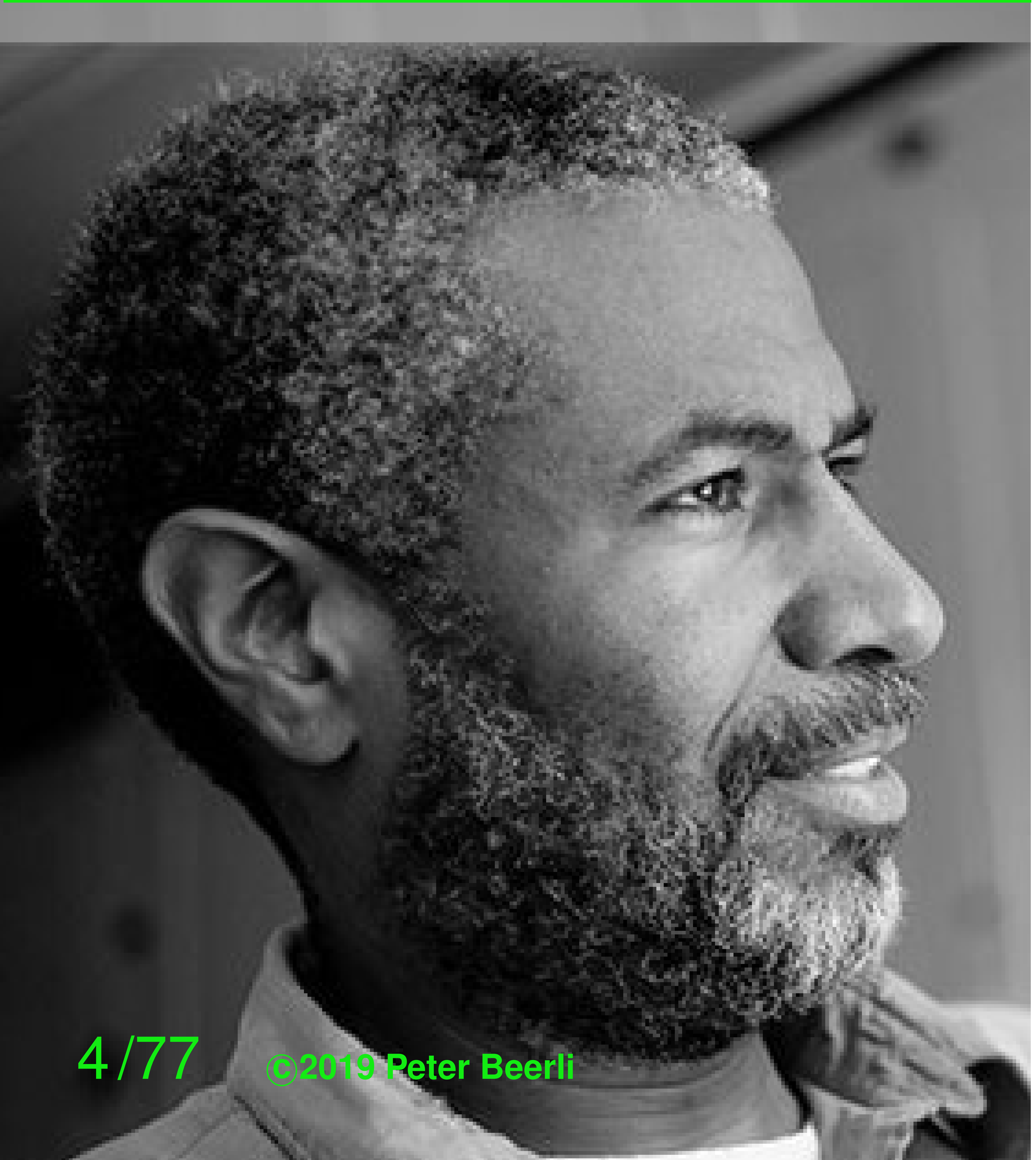
On models and data

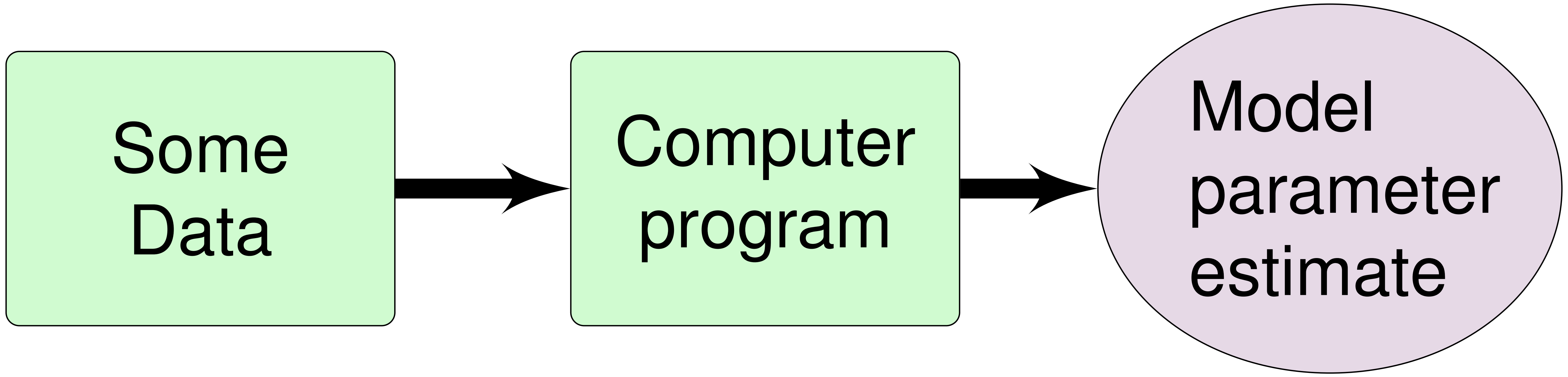


Theoretical Biologist: John Wakeley

©2019 Peter Beerli

On data and models




Some Data

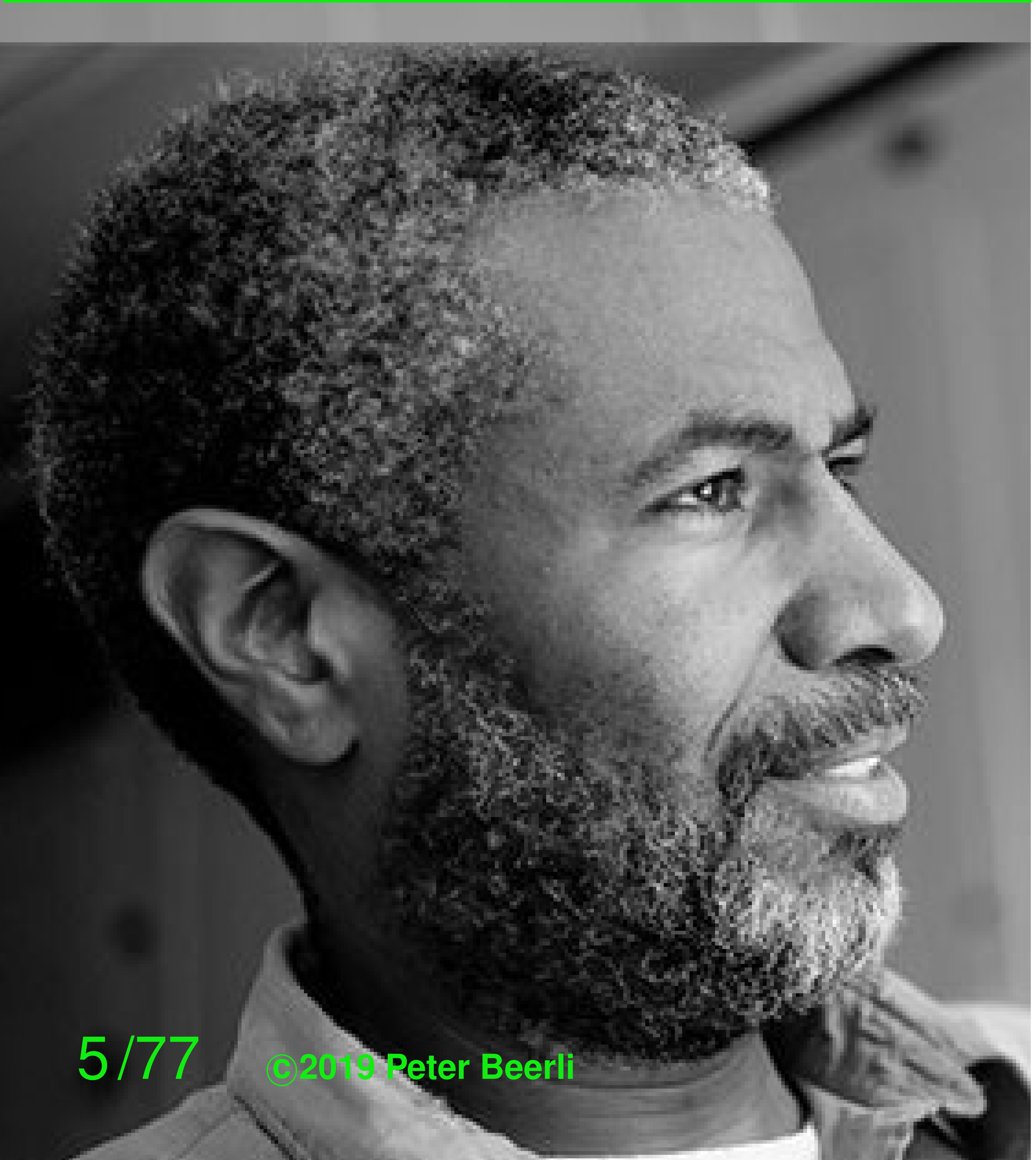
Computer program

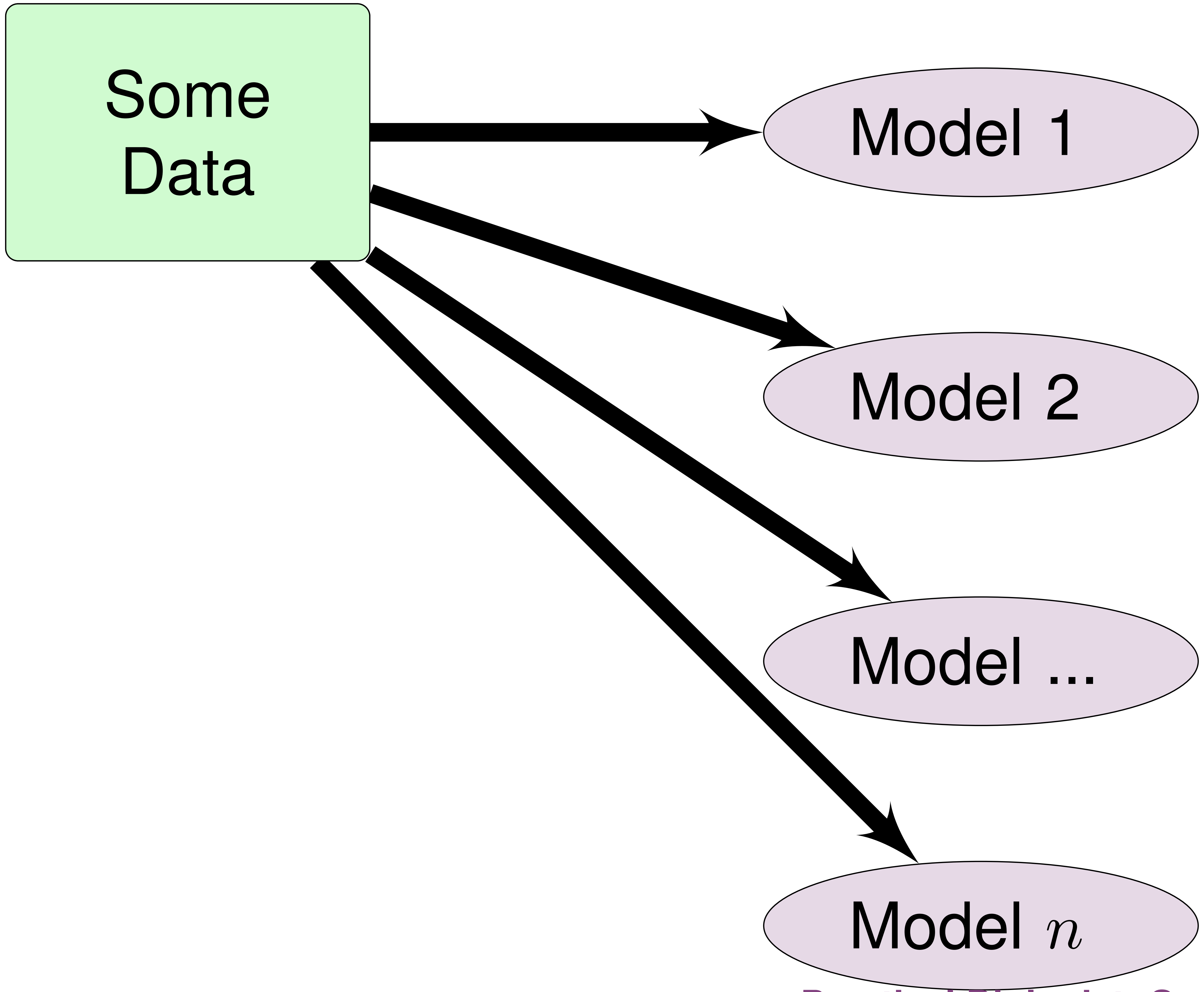
Model parameter estimate

©2019 Peter Beerli

Practical Biologist: Scott Edwards

On models and data

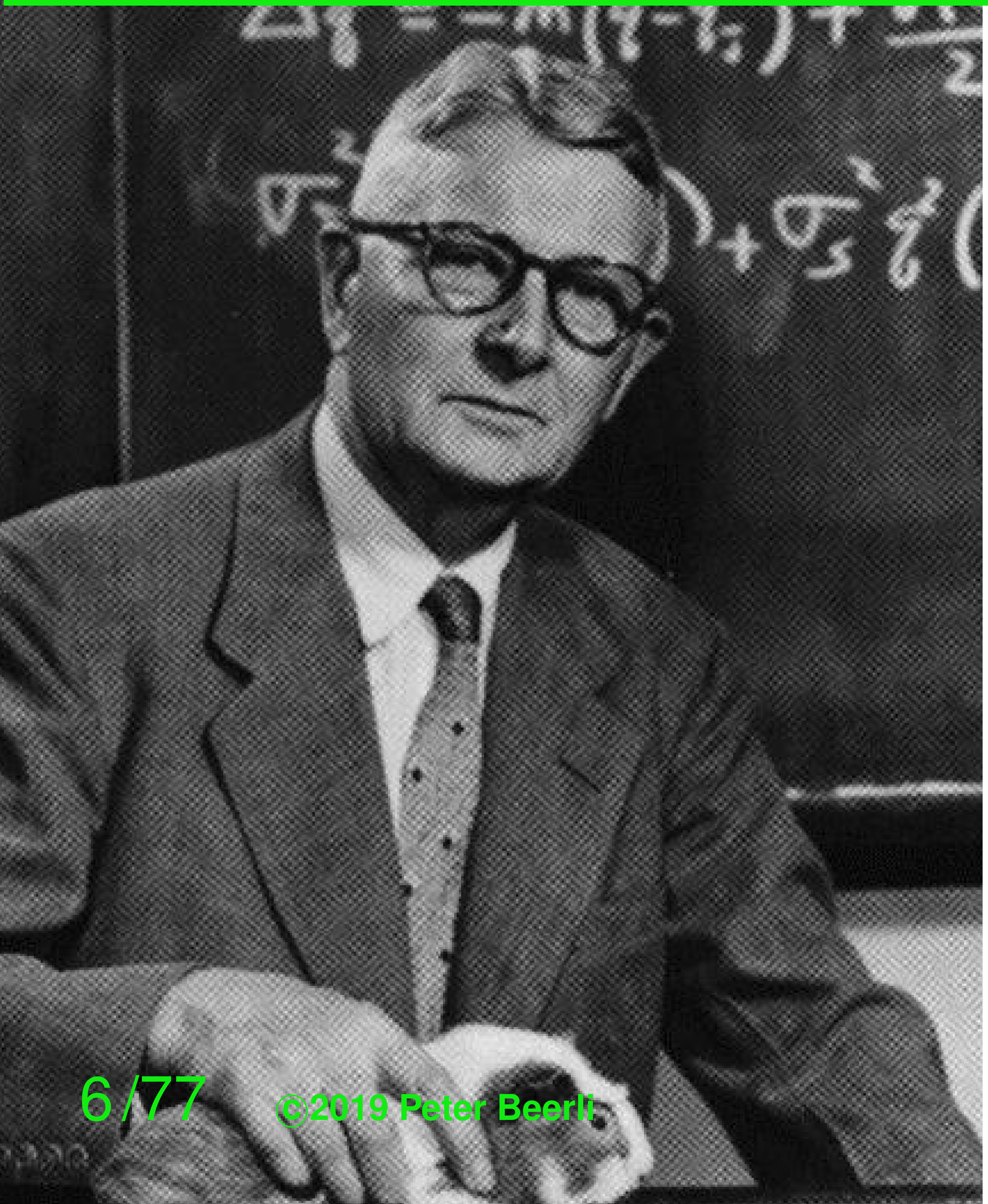


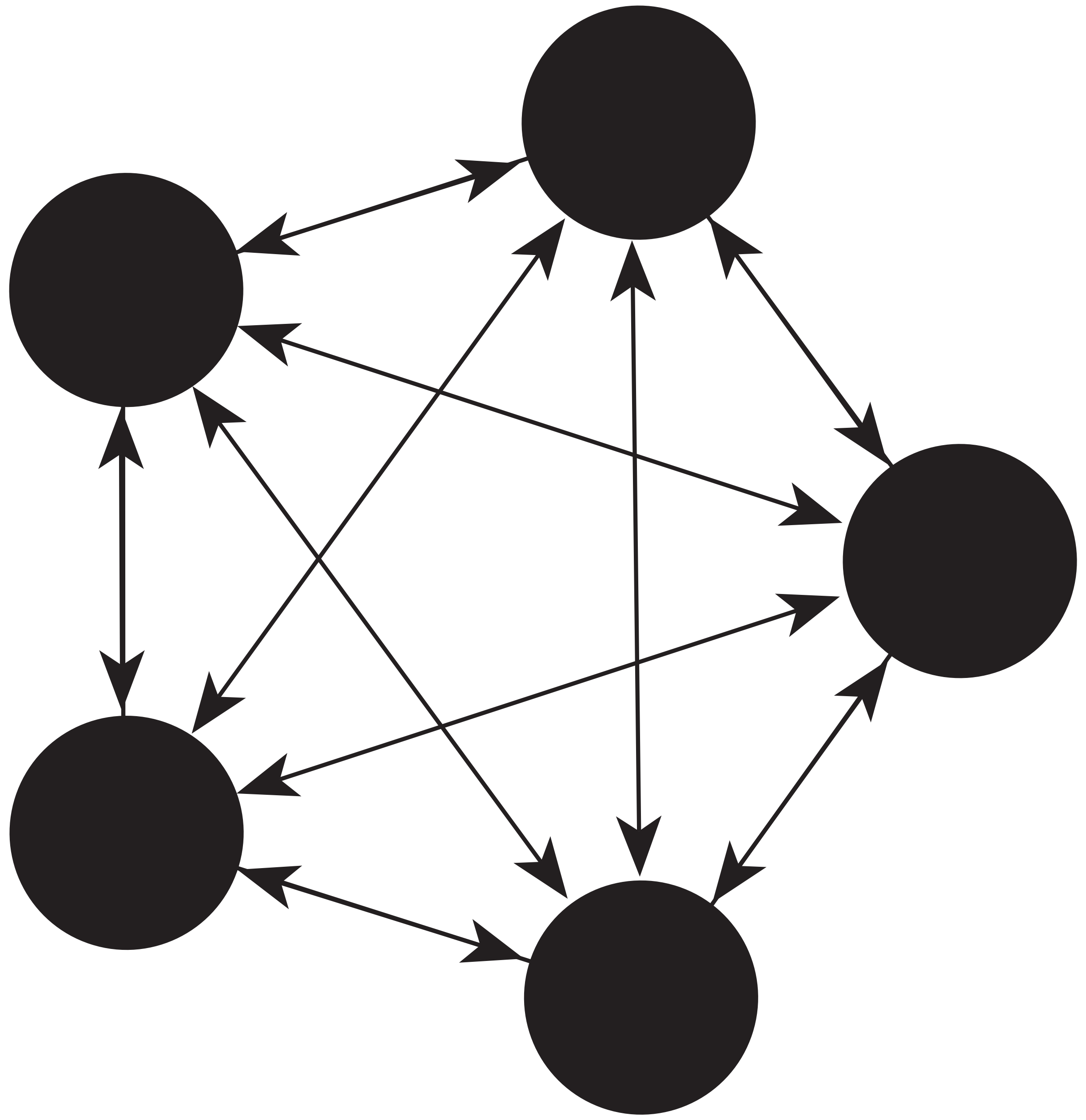


Practical Biologist: Scott Edwards

©2019 Peter Beerli

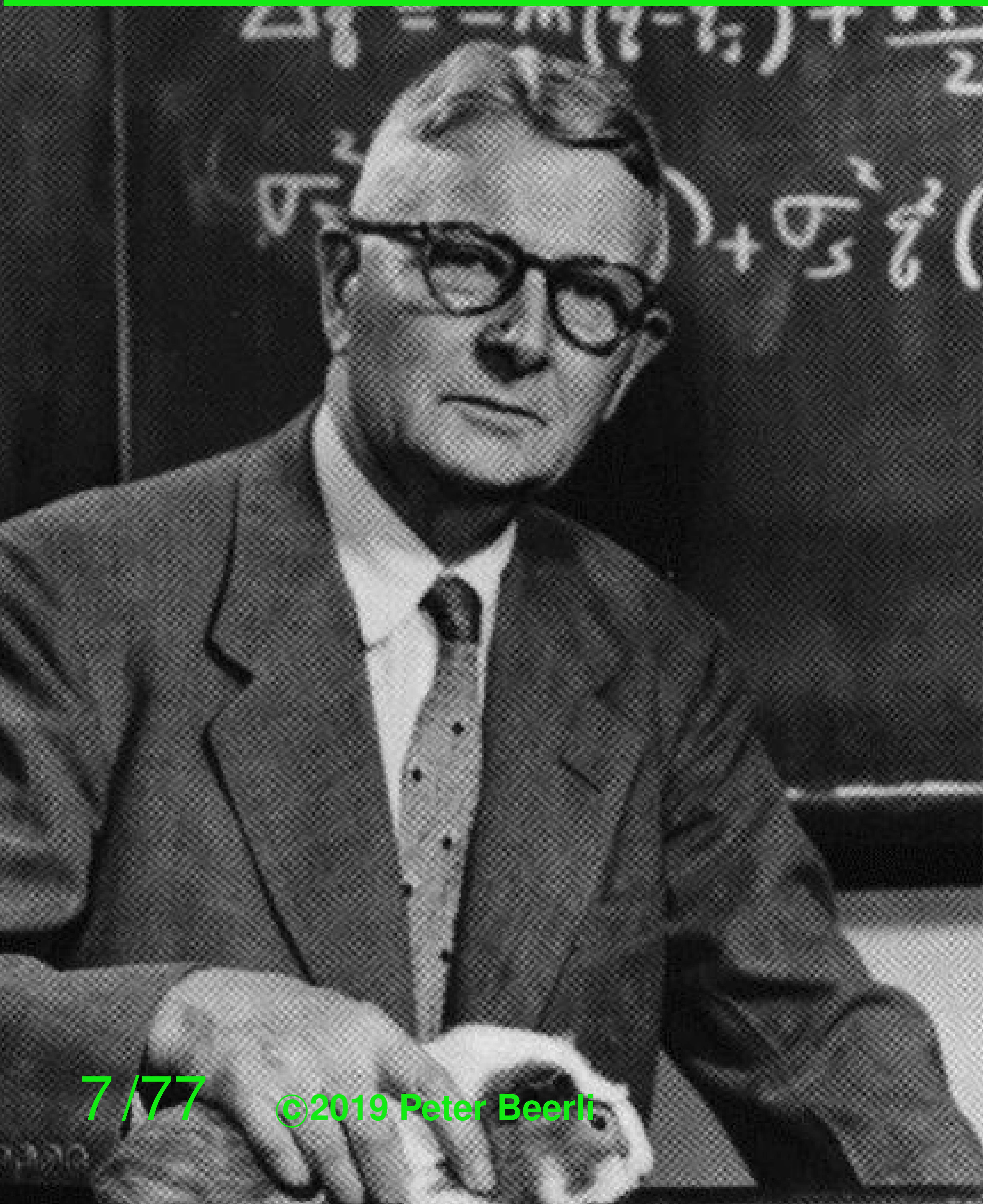
Population genetics models

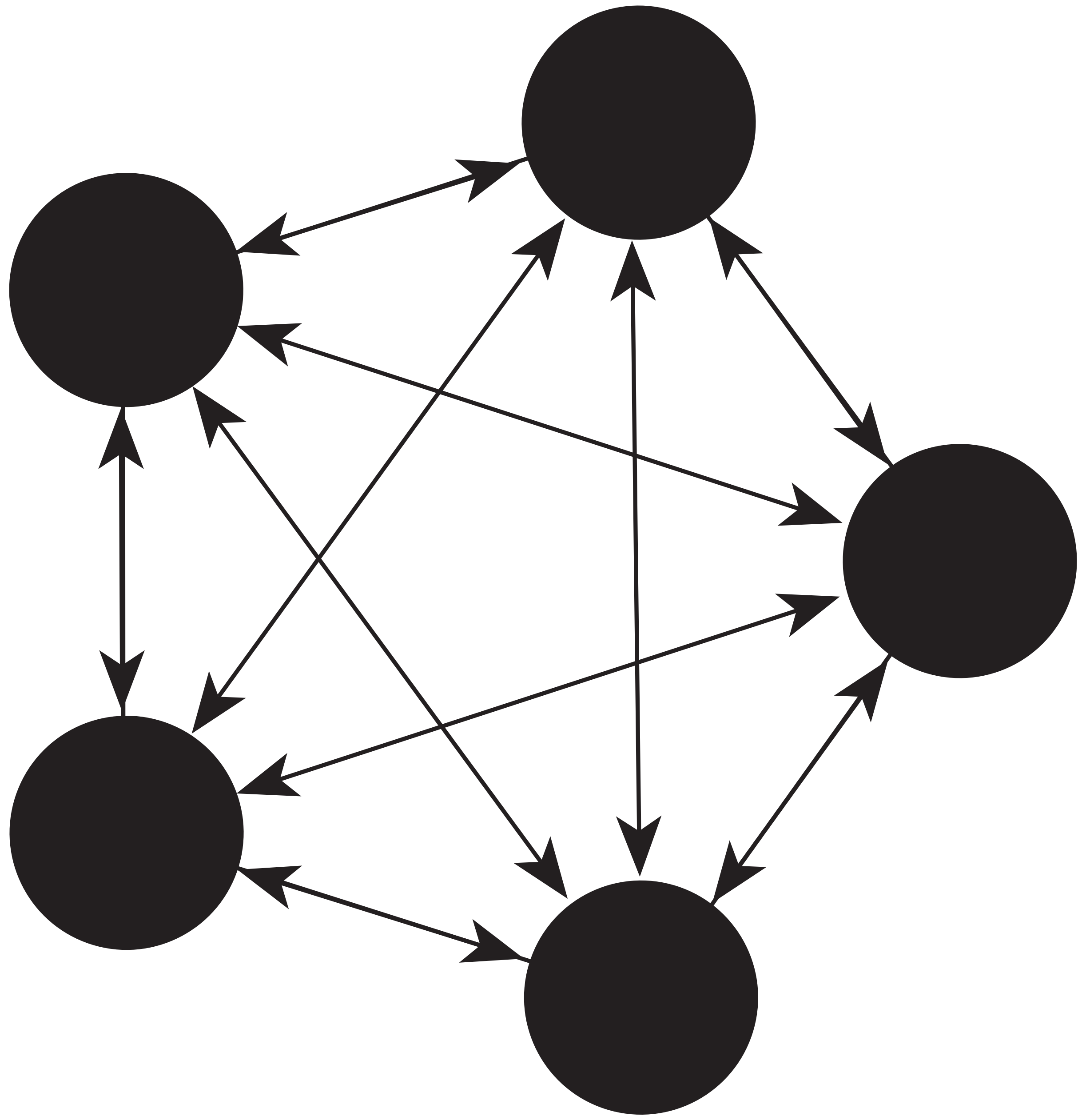




Theoretical Biologist: Sewall Wright

Population genetics models





$$F_{ST} = \frac{\sigma^2(p)}{p(1-p)} \simeq \frac{H_T - \overline{H}_S}{H_T}$$

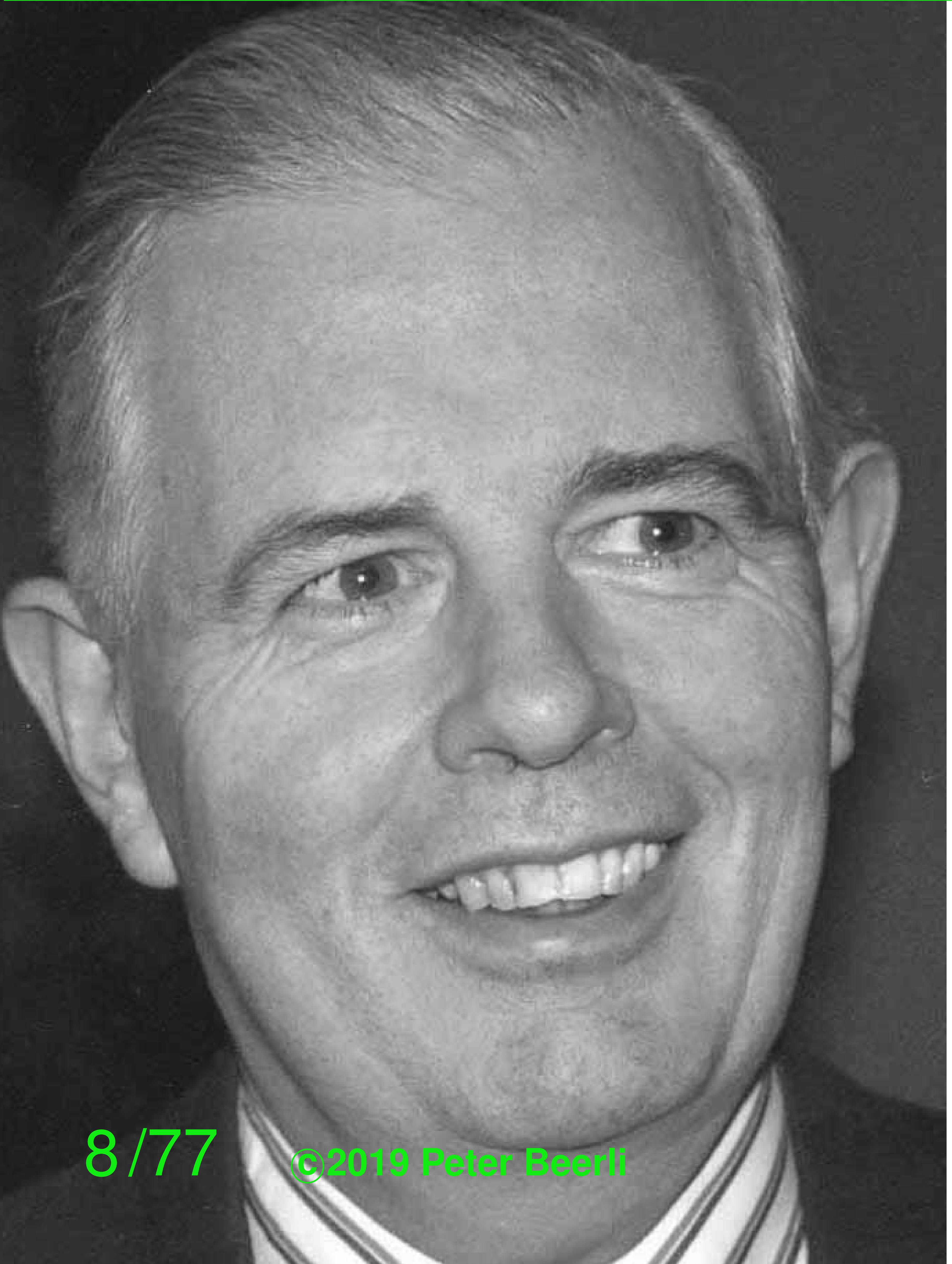
$$F_{ST} \approx \frac{1}{4Nm+1}$$

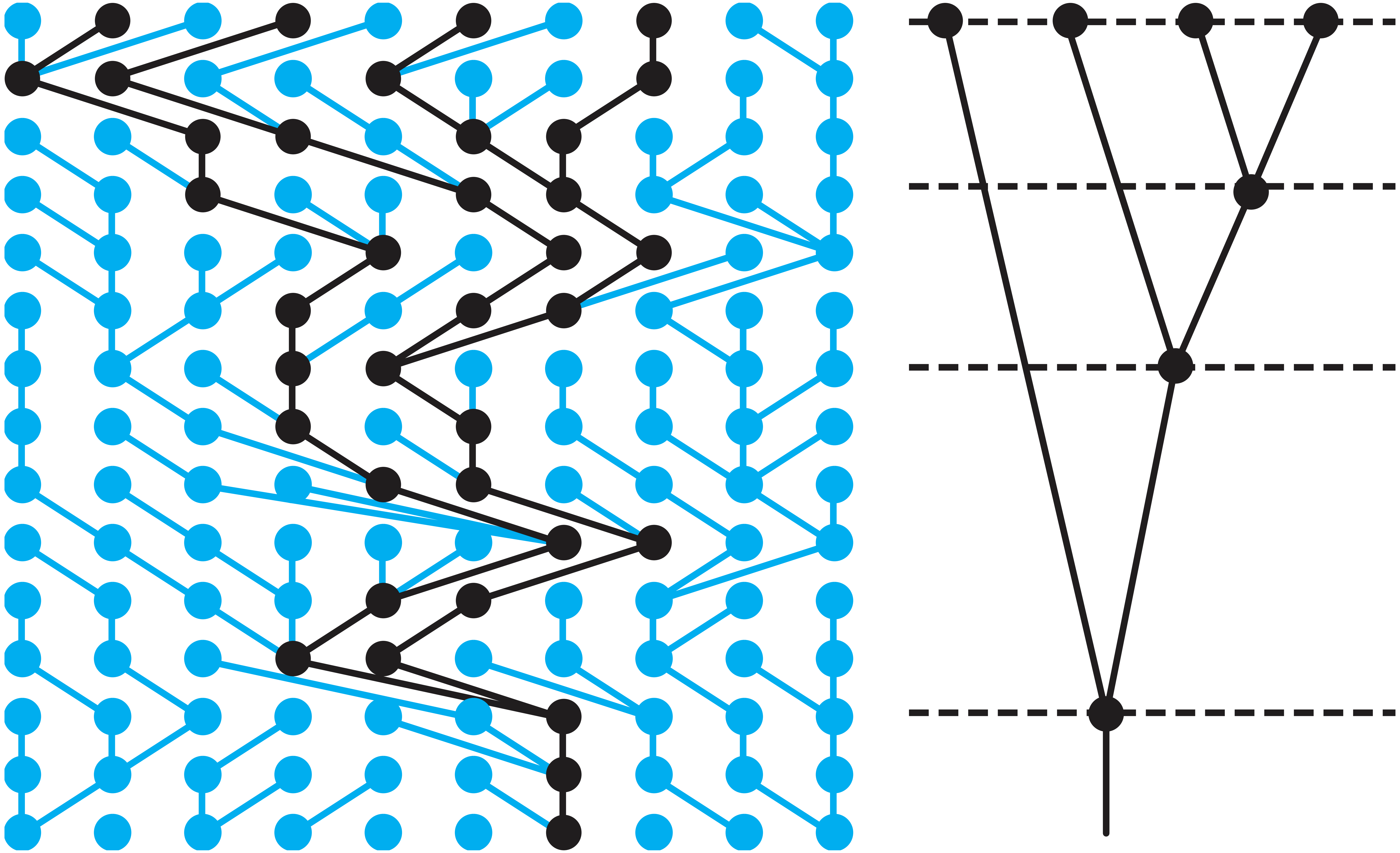
$$Nm \approx \frac{1}{4}\left(\frac{1}{F_{ST}} - 1\right)$$

Theoretical Biologist: Sewall Wright

©2019 Peter Beerli

Population genetics models

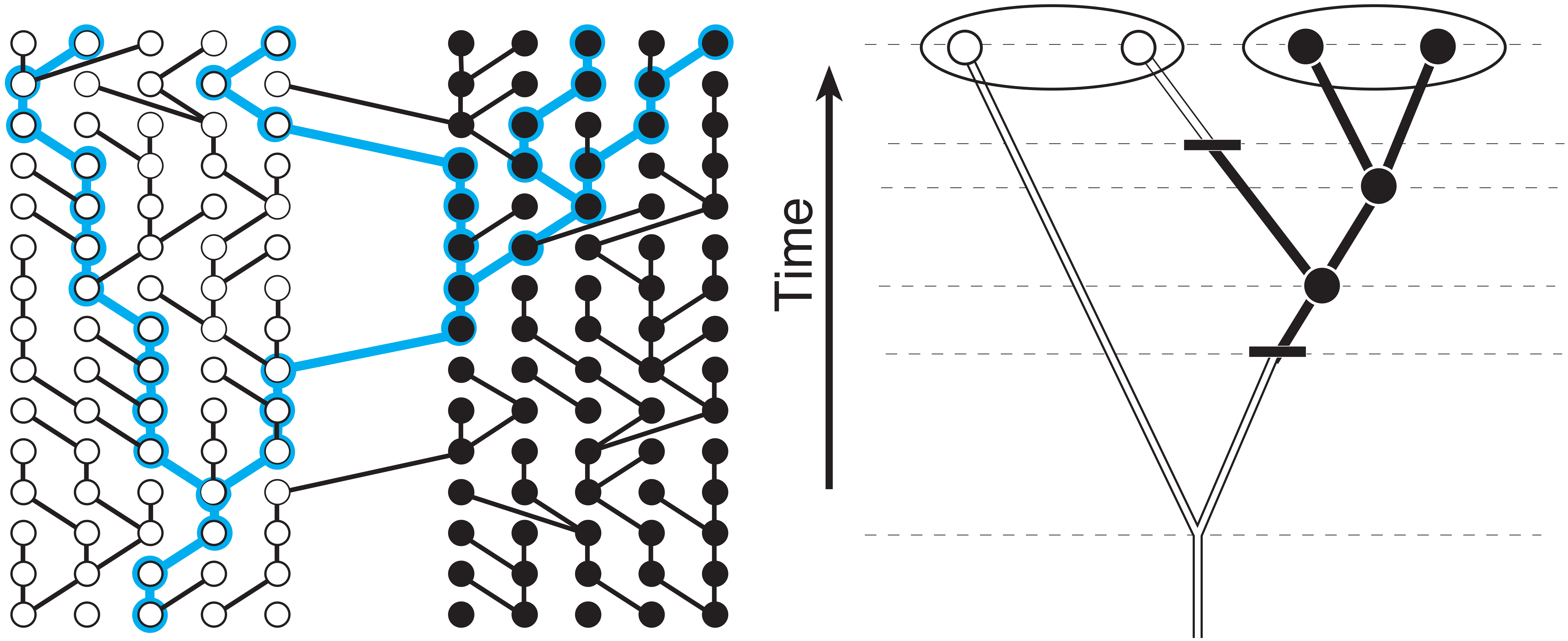




Probabilist: John F C Kingman

©2019 Peter Beerli

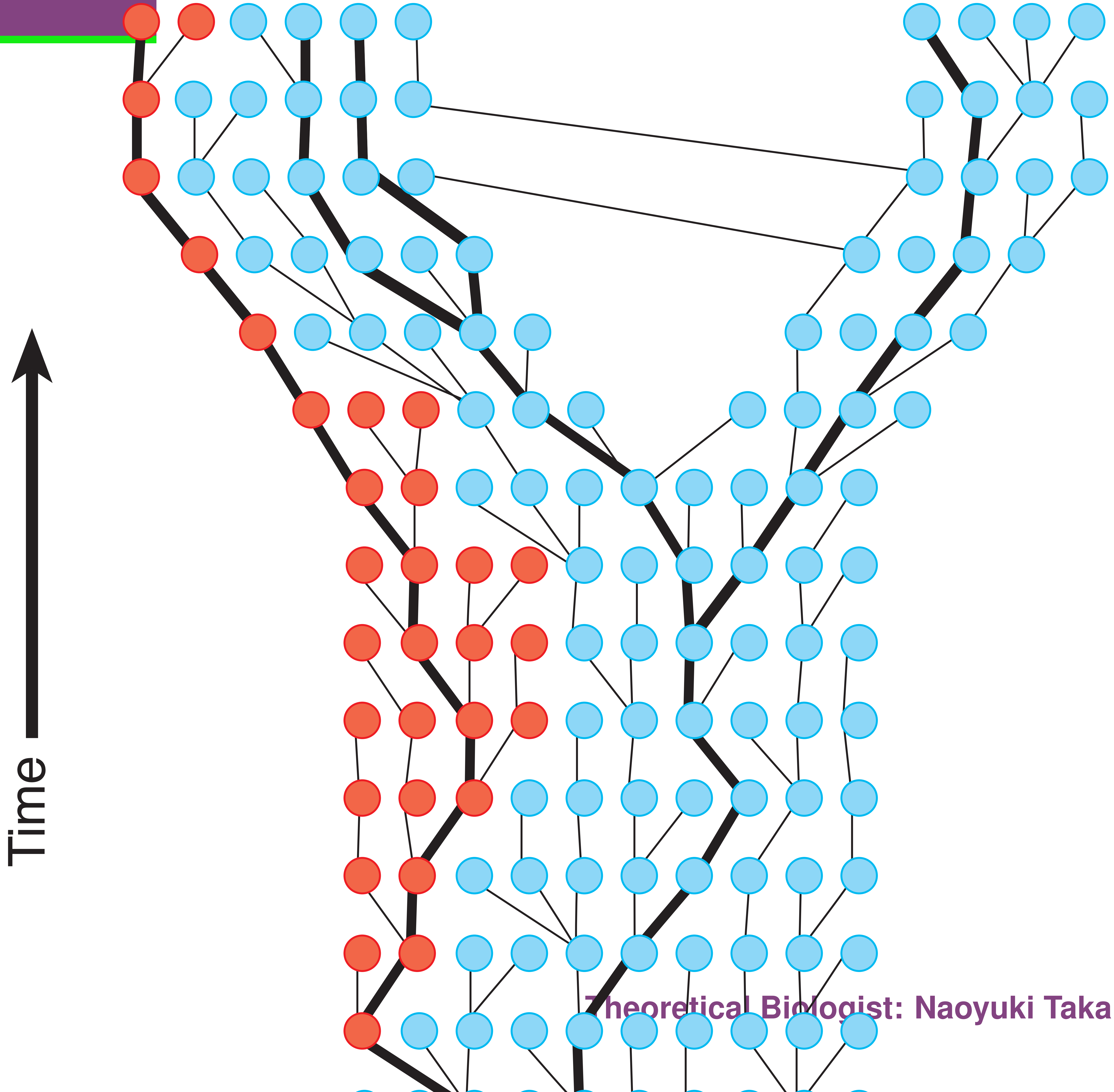
Population genetics models



Theoretical Biologist: Dick Hudson

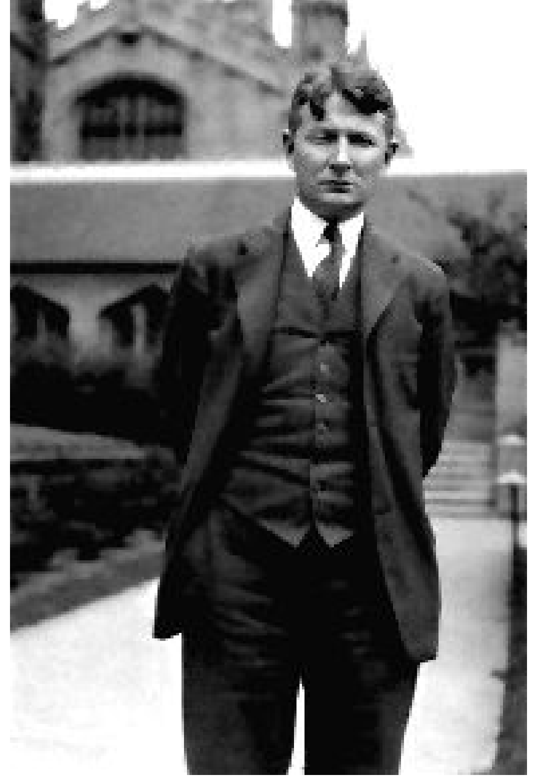
©2019 Peter Beerli

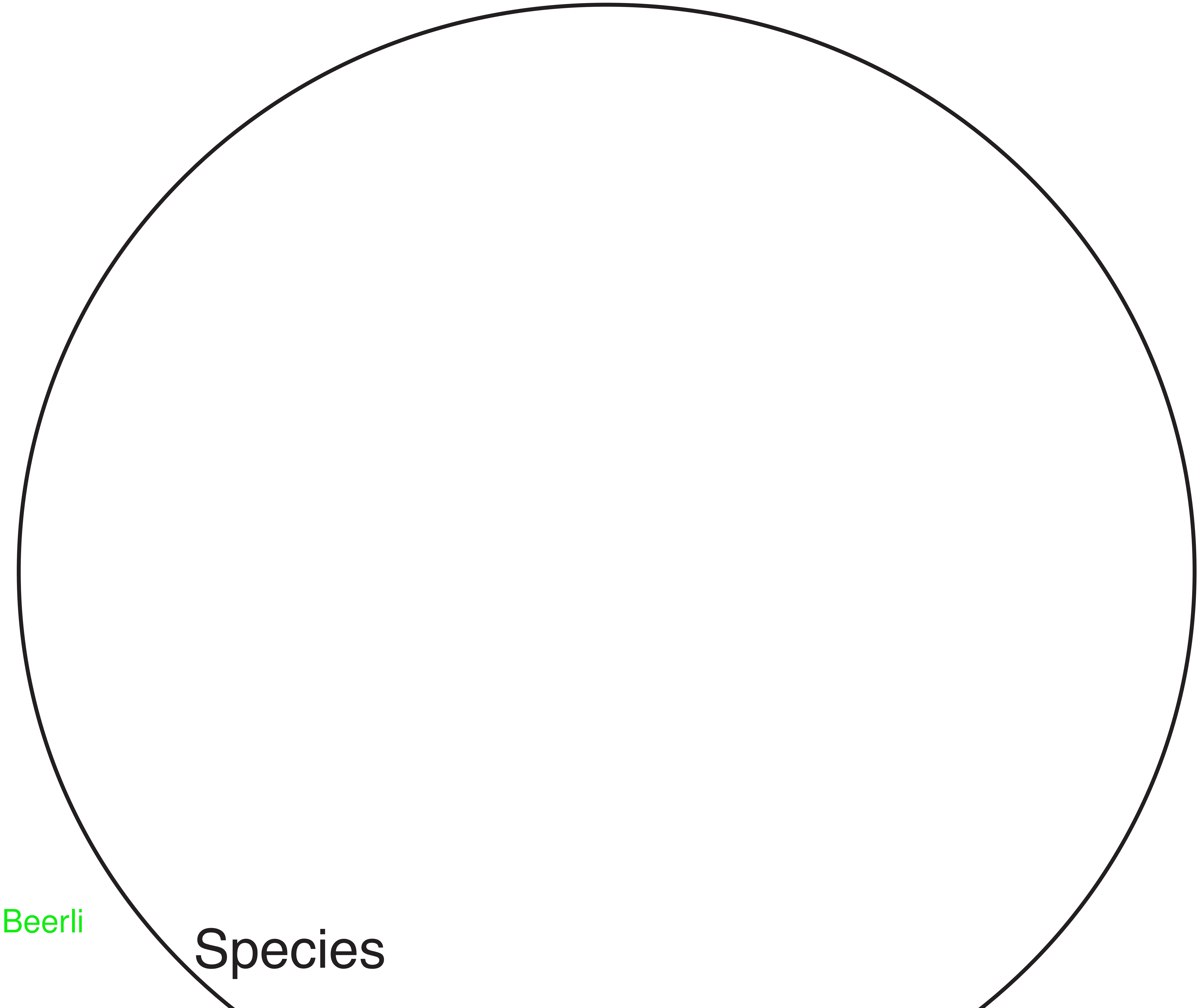
Population genetics models



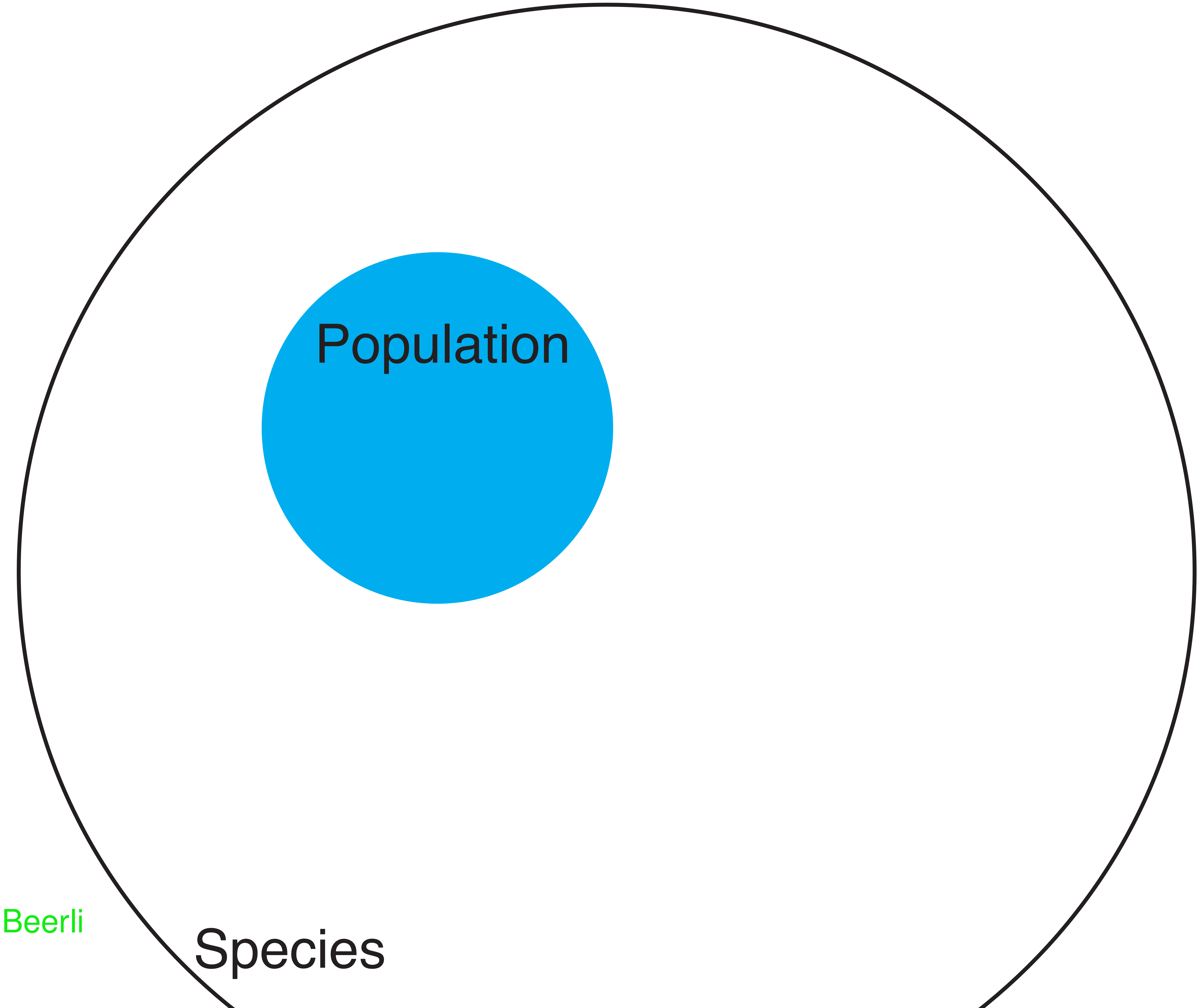
Theoretical Biologist: Naoyuki Takahata

Population models



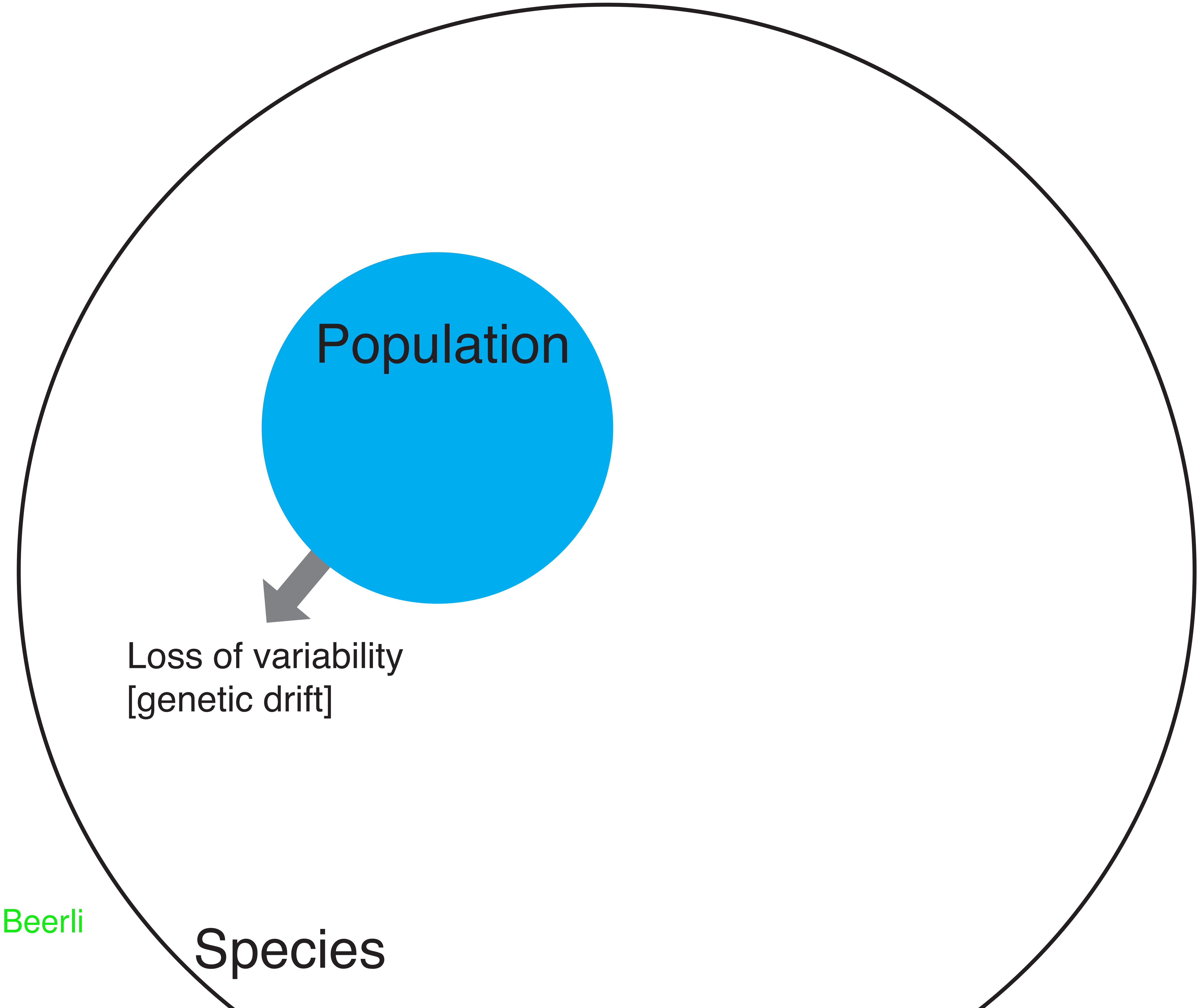


Population models



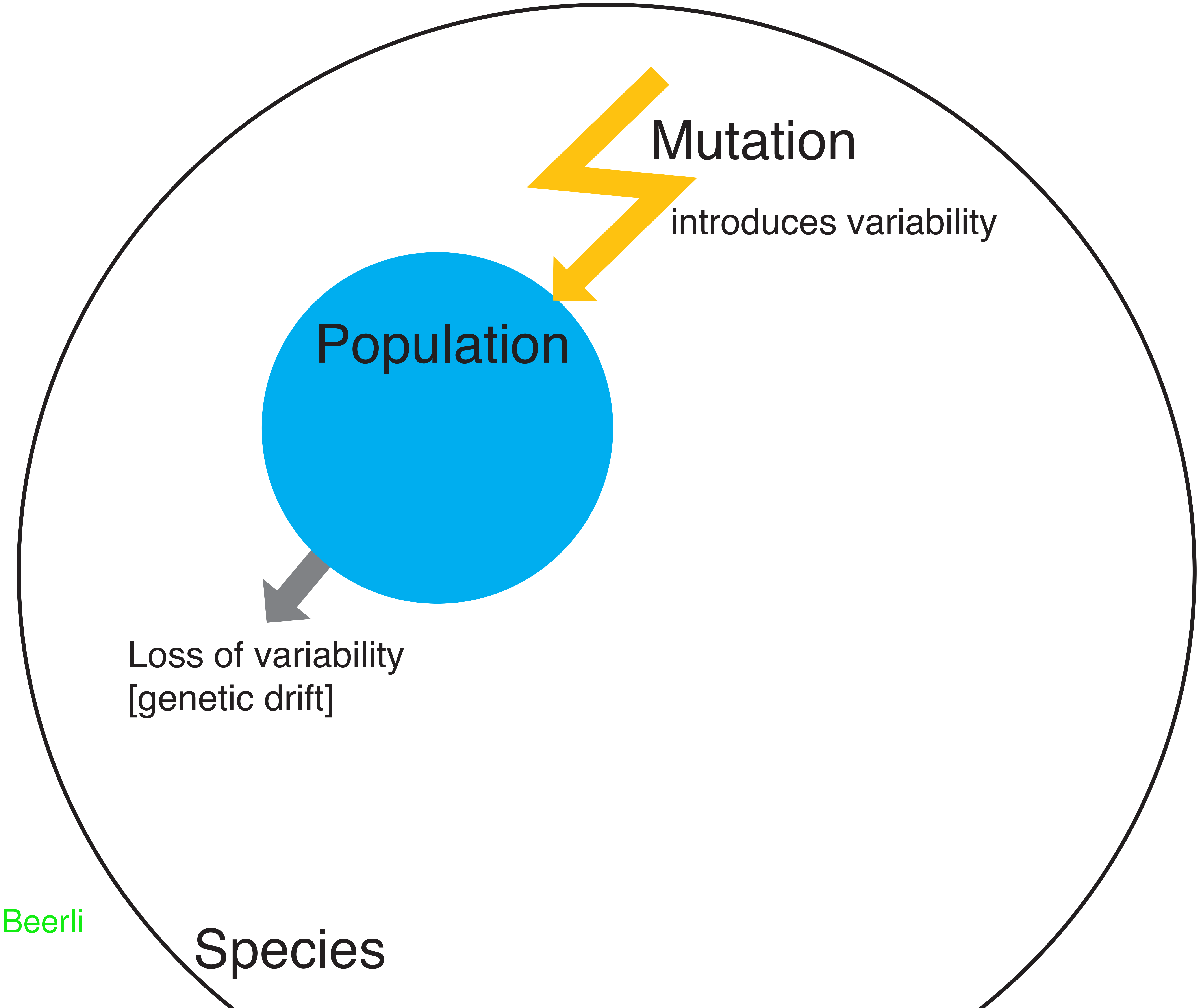
©2019 Peter Beerli

Population models



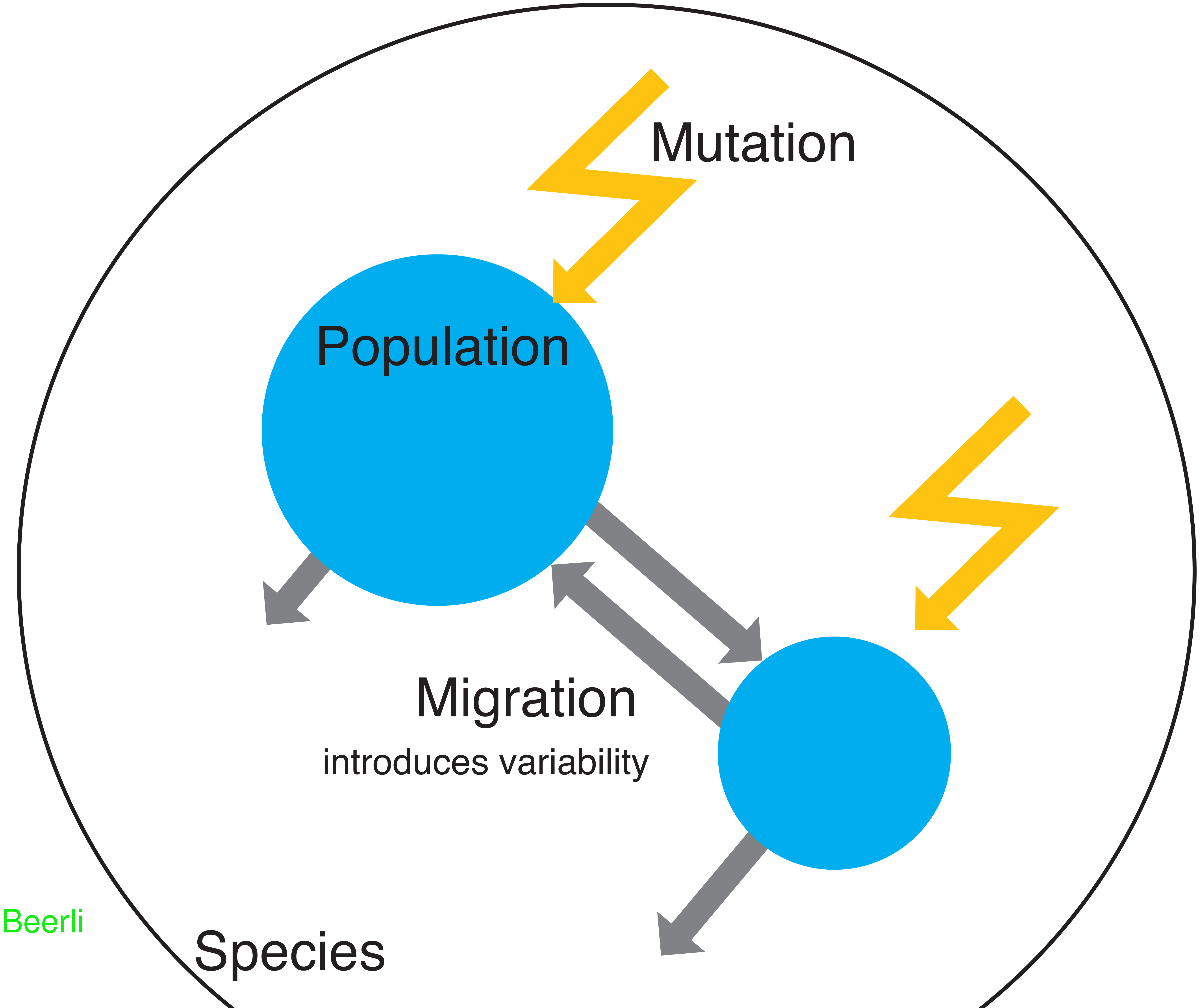
©2019 Peter Beerli

Population models



©2019 Peter Beerli

Population models



©2019 Peter Beerli

Population models

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

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

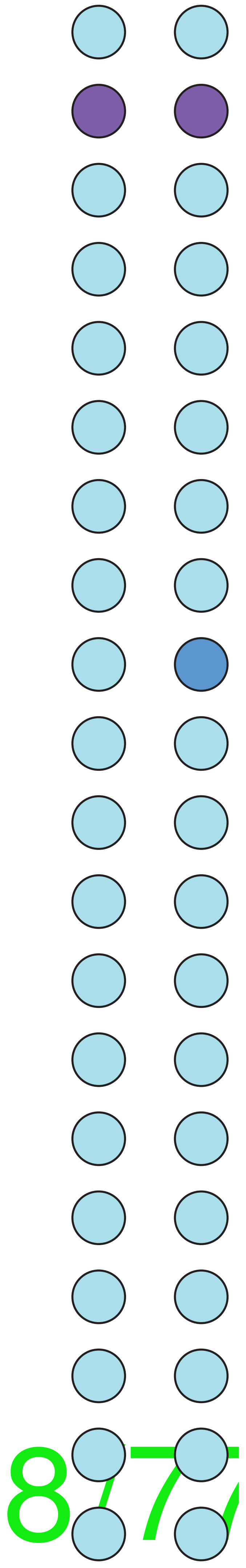
Simply looking only at a single population this is

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

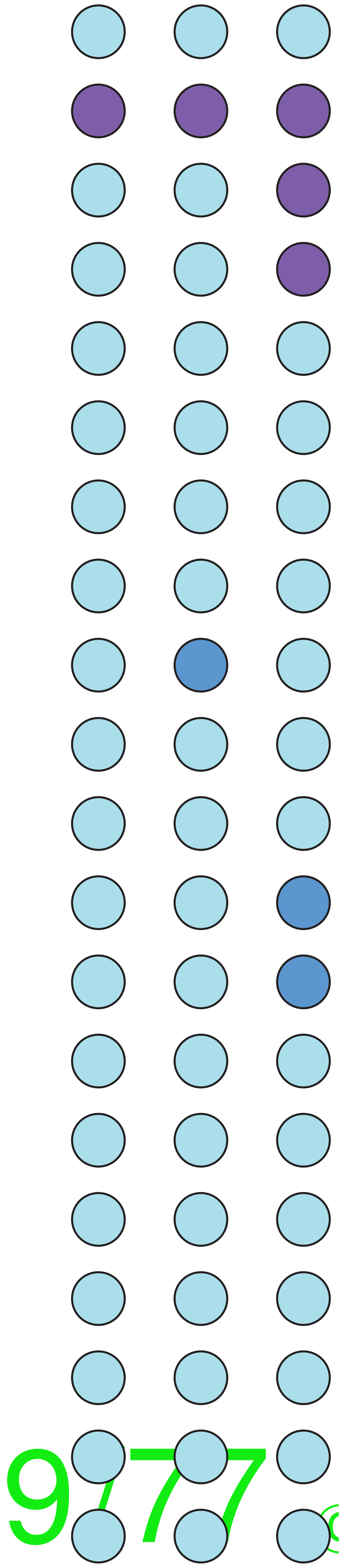
©2019 Peter Beerli

Population models

Population models

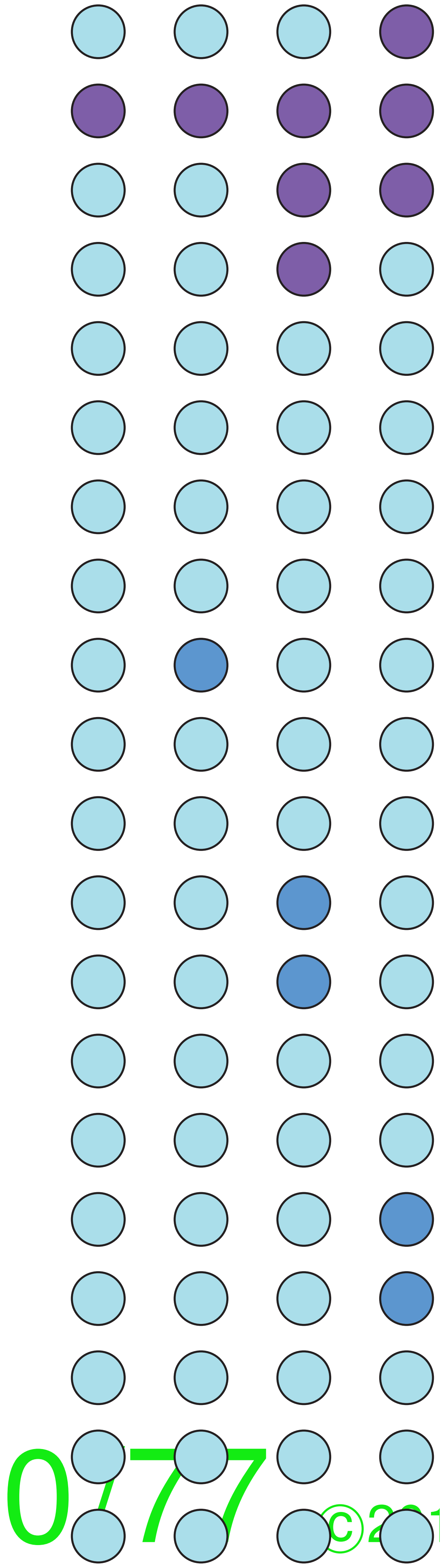


Population models

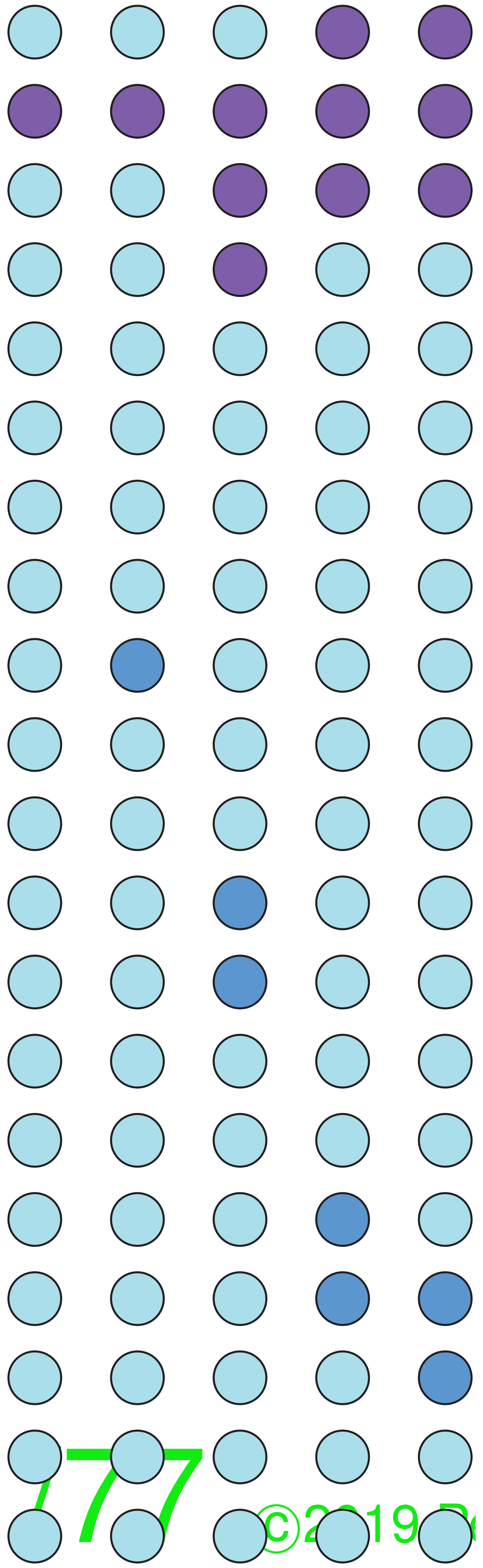


©

Population models

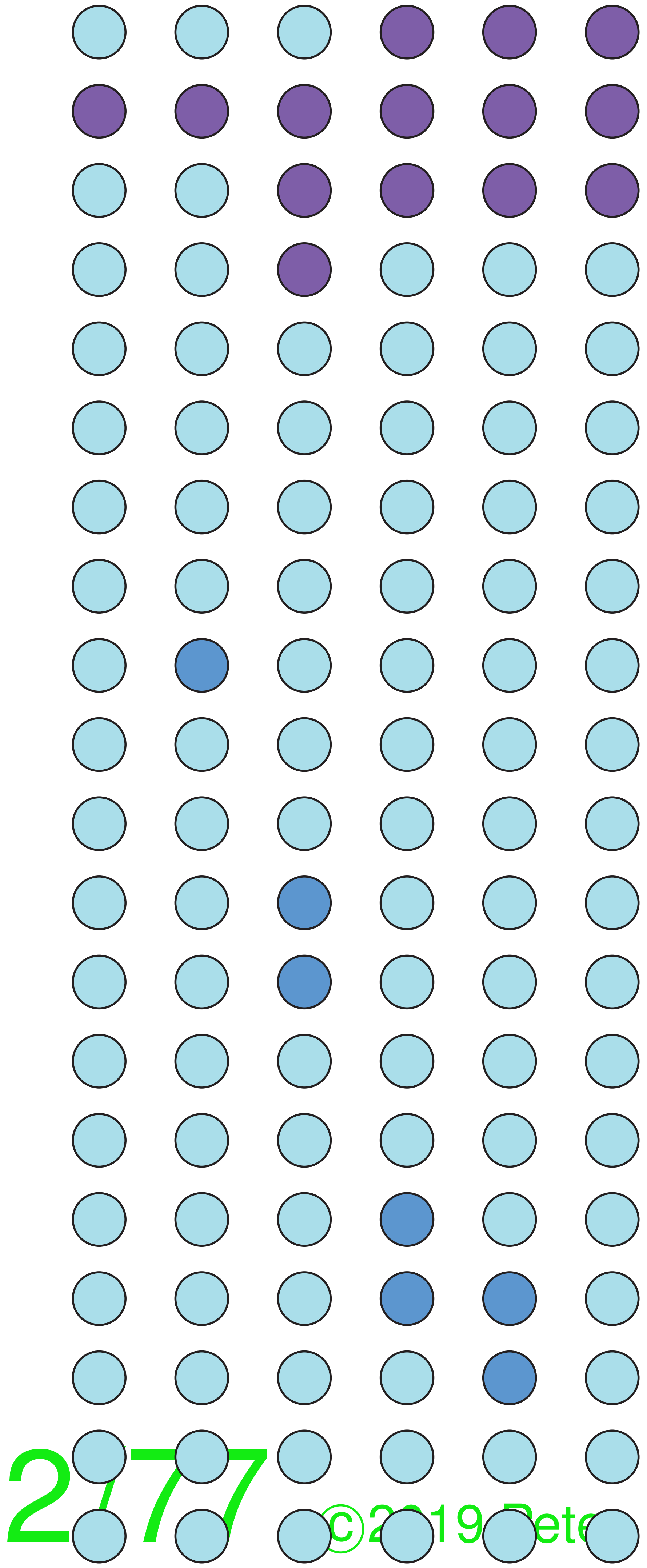


Population models



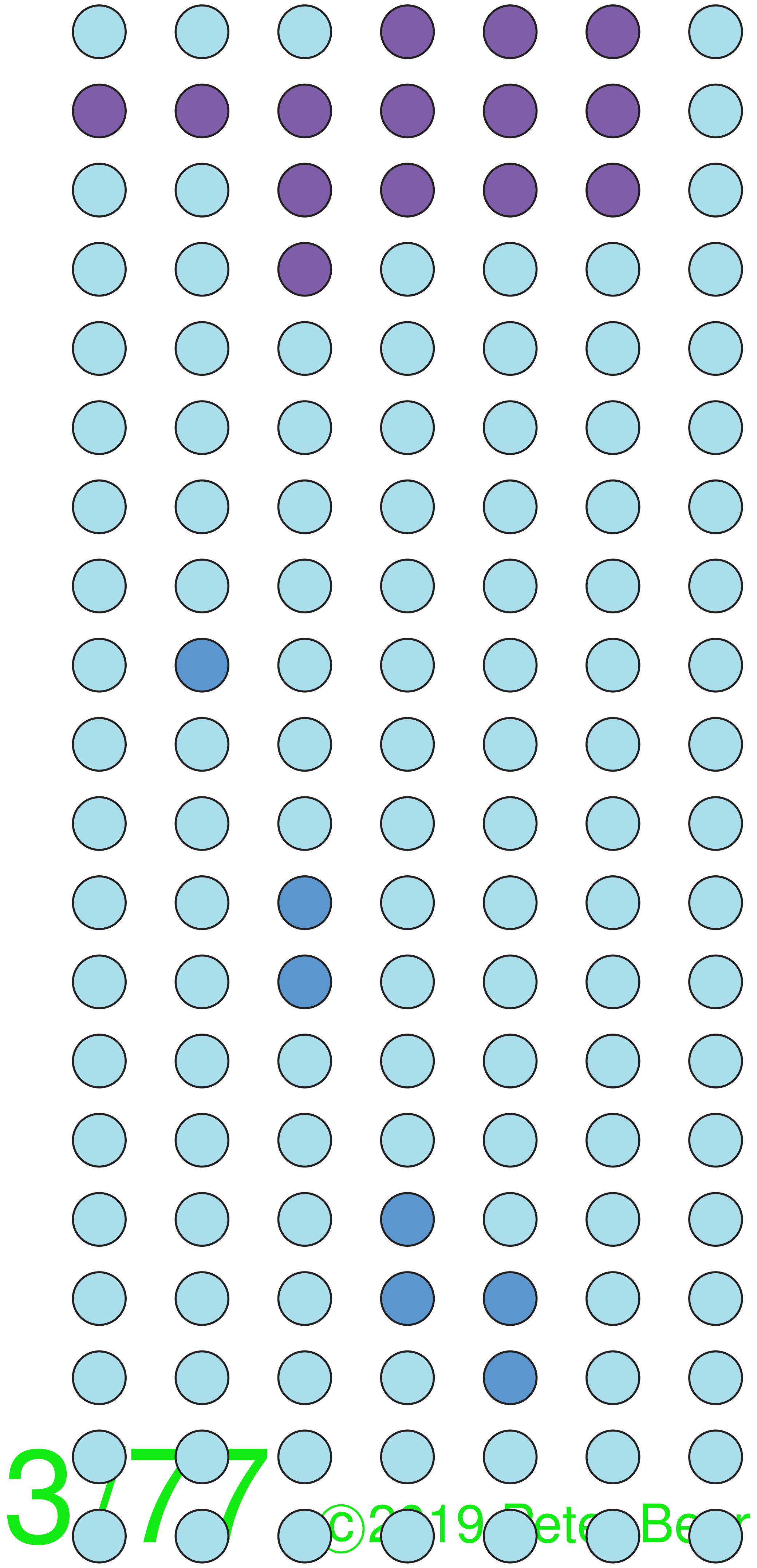
©2019 Re

Population models



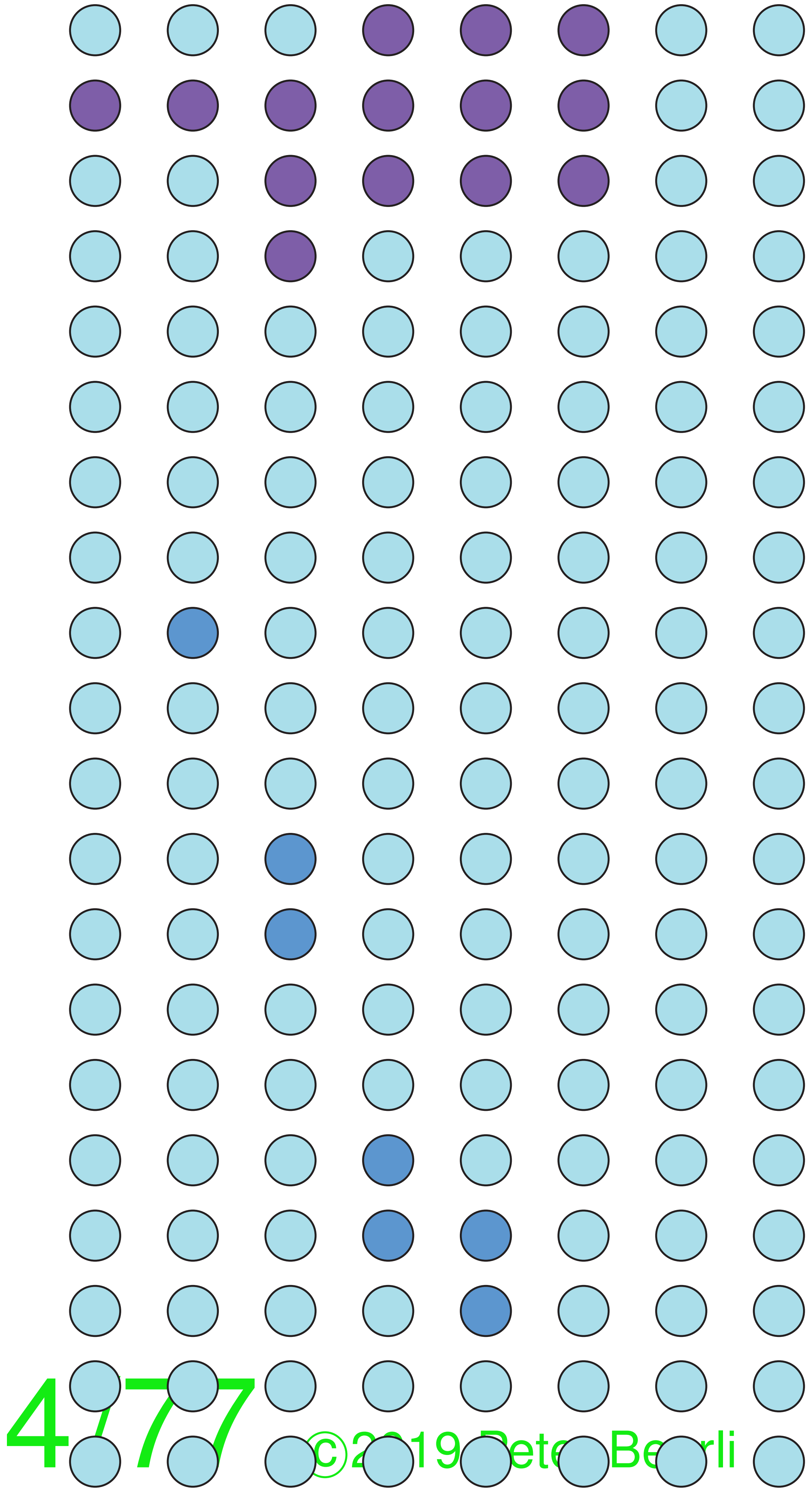
©2019 Pete

Population models



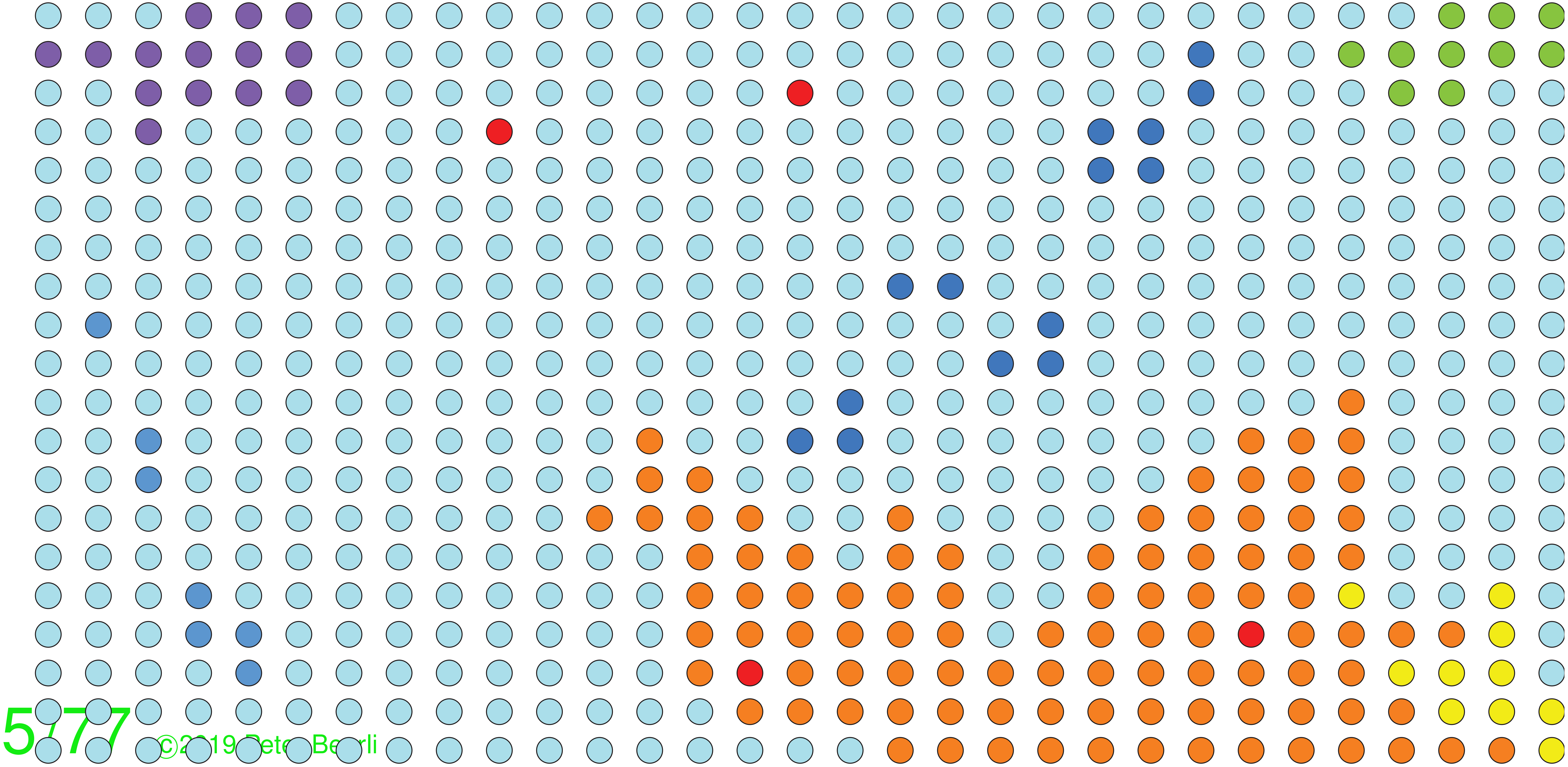
©2019 Peter Beerli

Population models



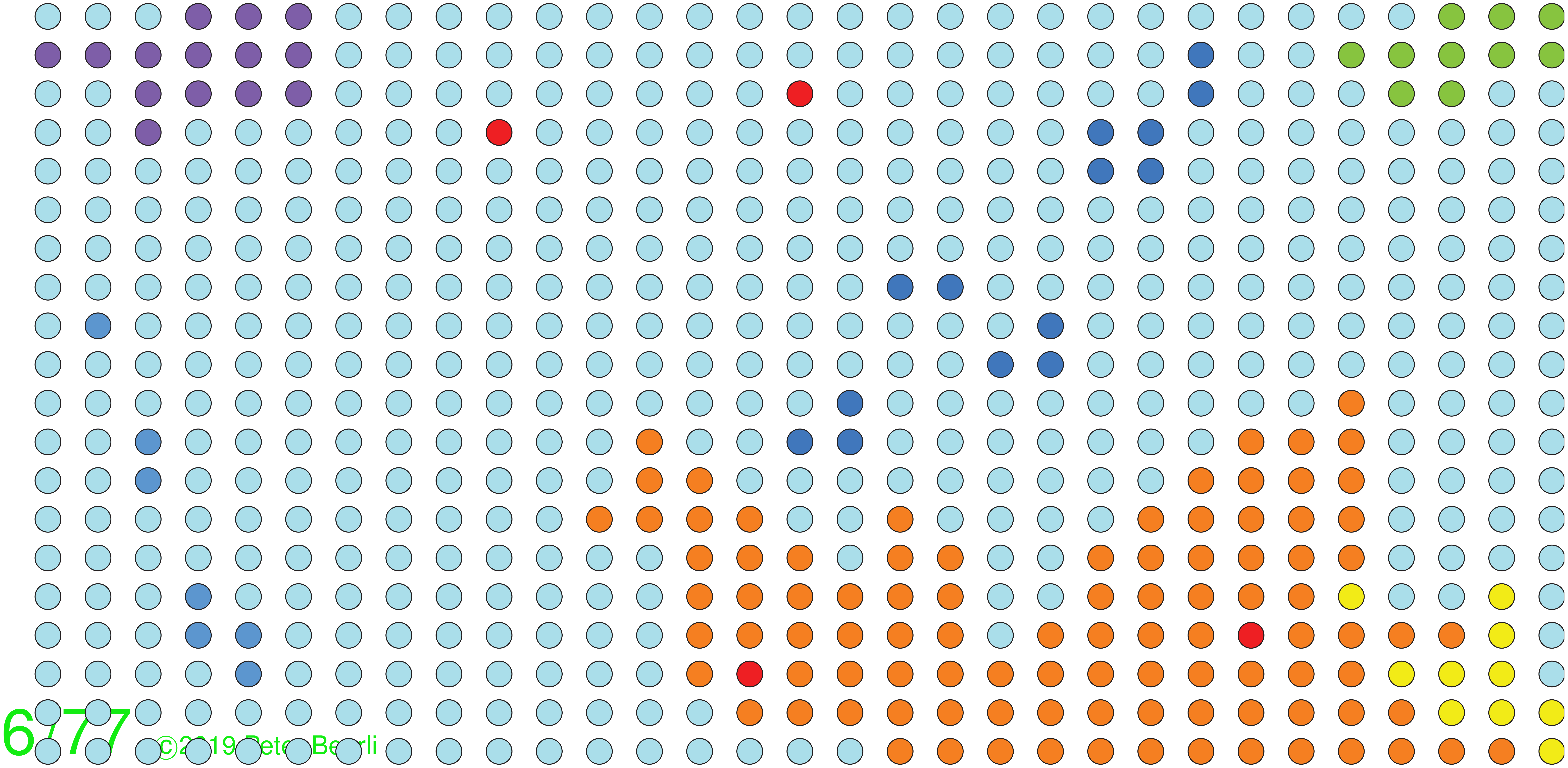
©2019 Peter Beerli

Population models



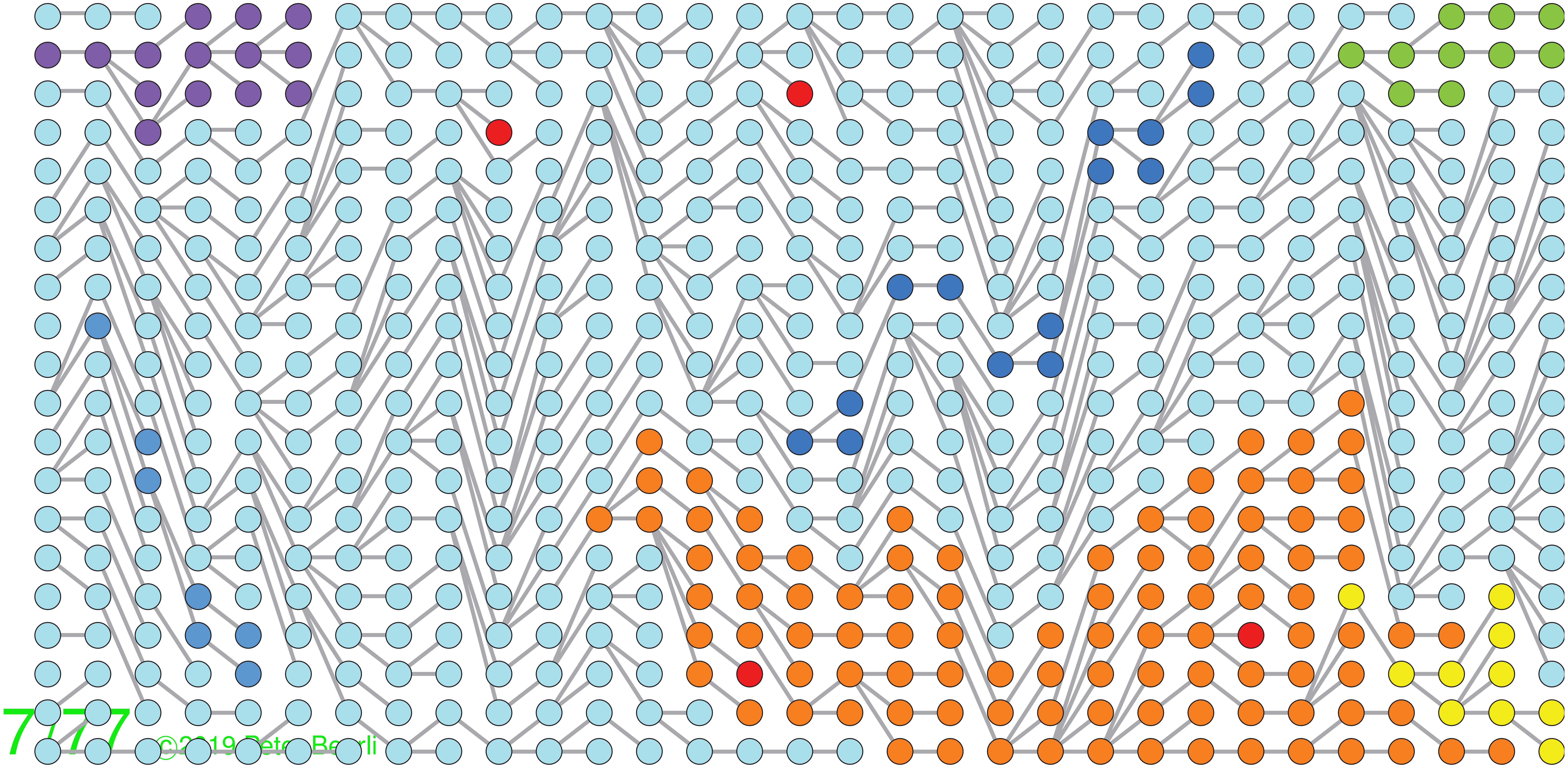
©2019 Peter Beerli

Population models



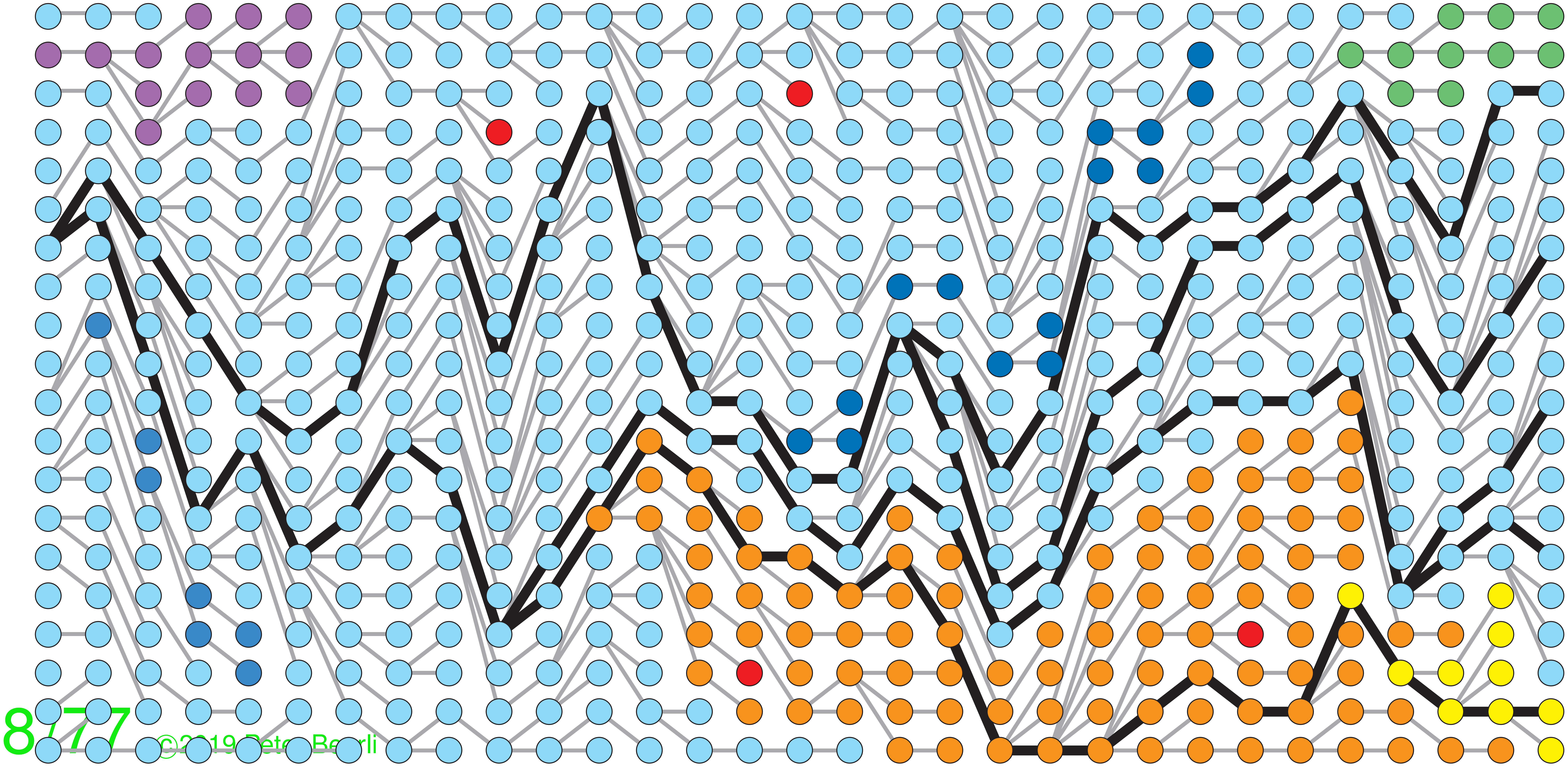
©2019 Peter Beerli

Population models



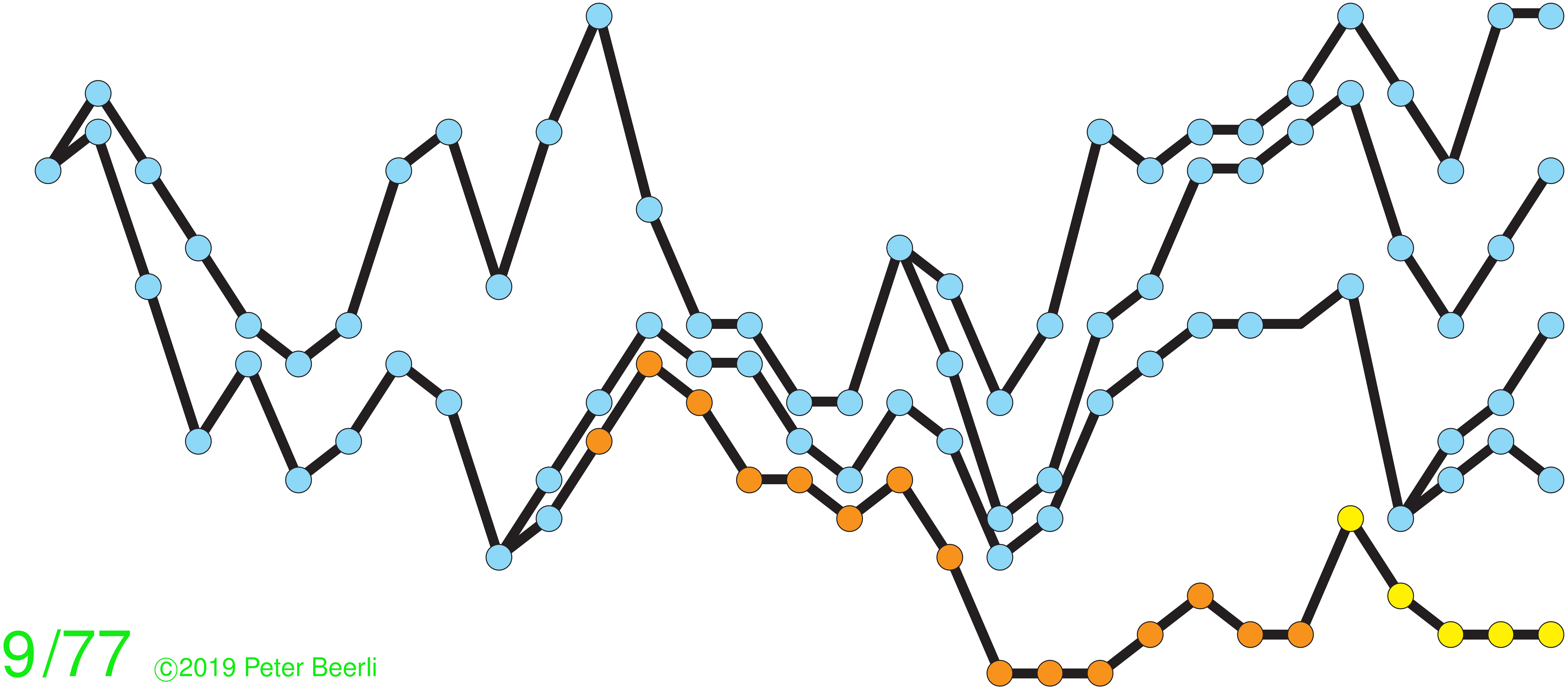
©2019 Peter Beerli

Population models



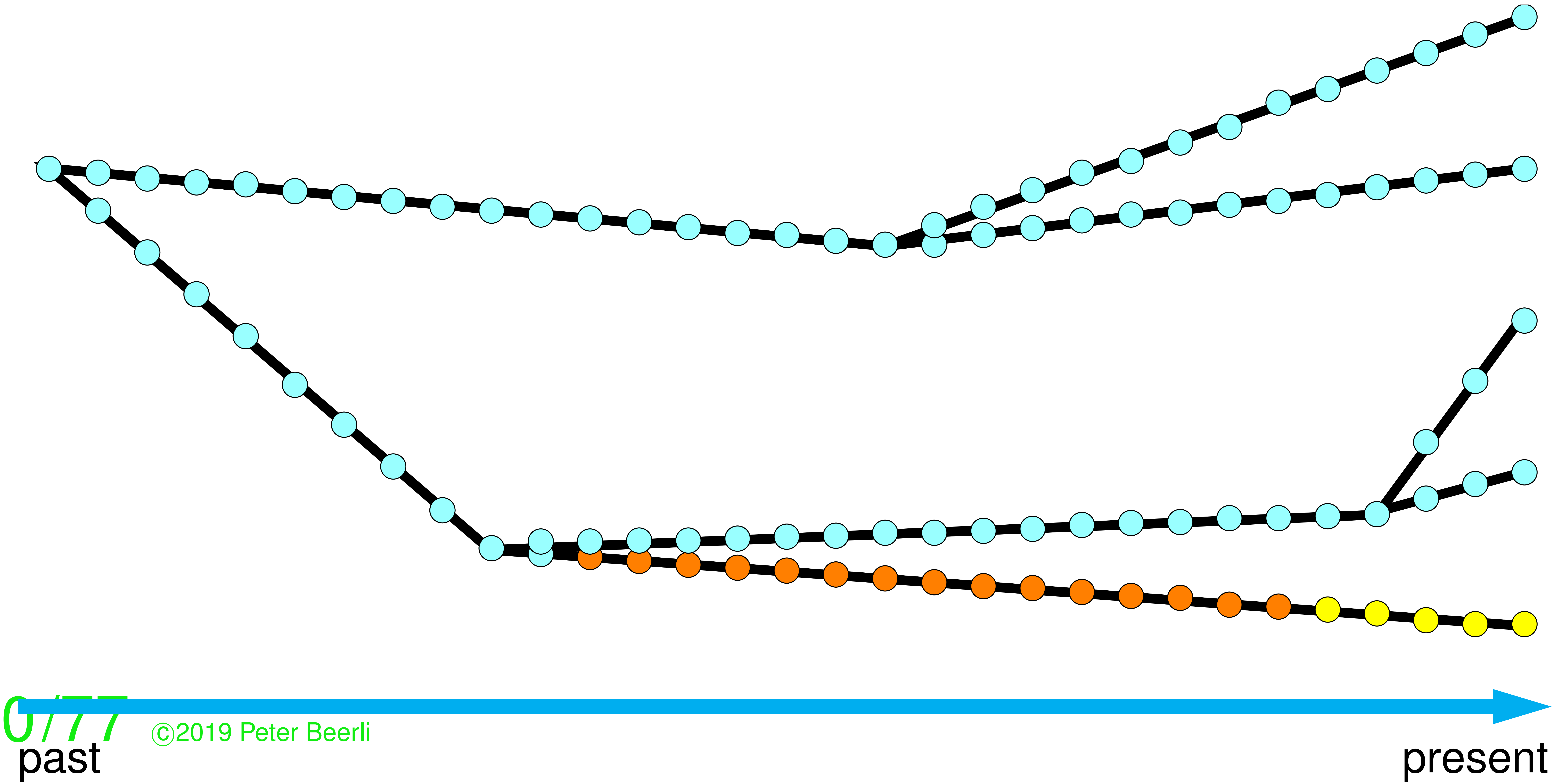
©2019 Peter Beerli

Population models



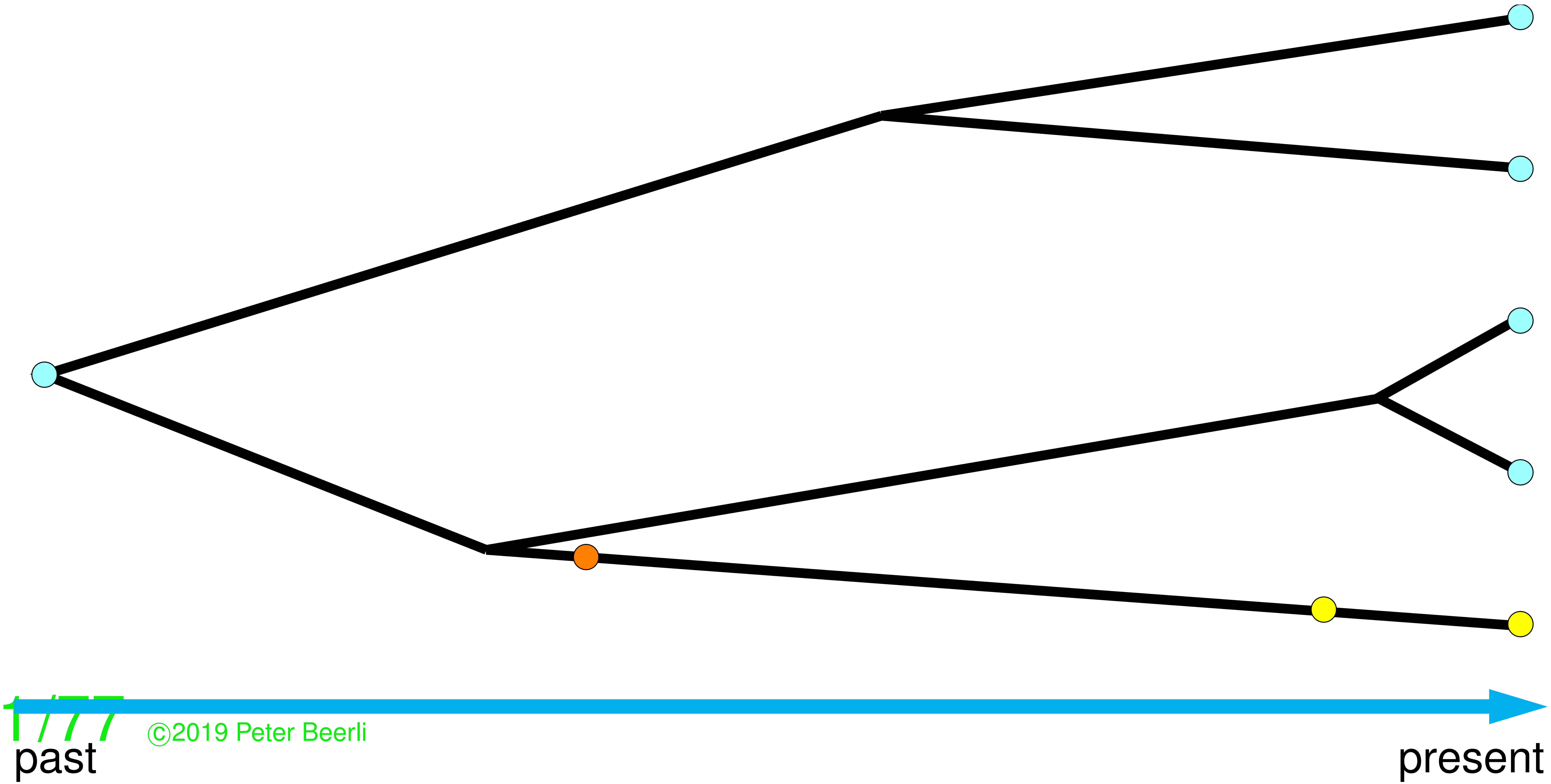
©2019 Peter Beerli

Population models



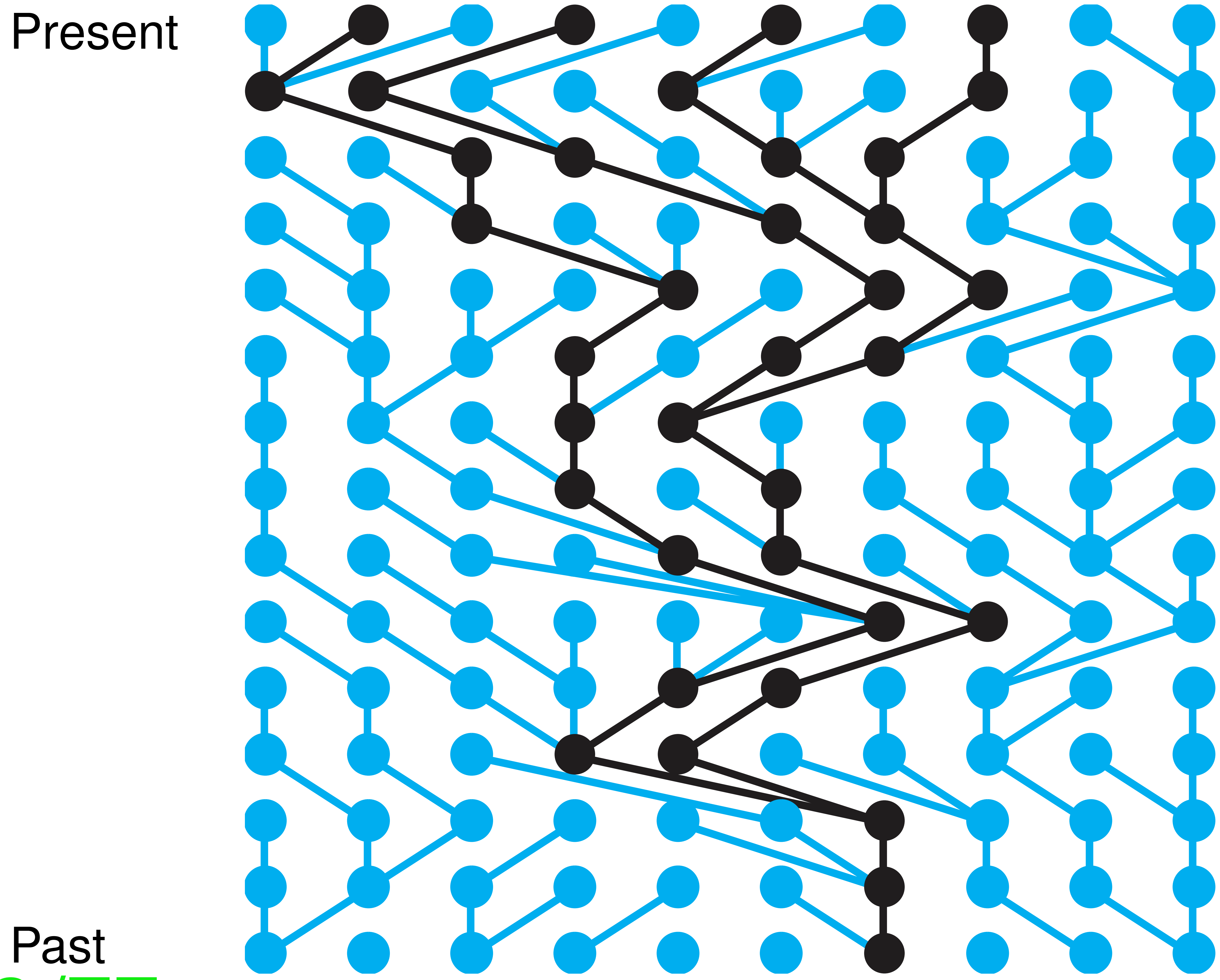
©2019 Peter Beerli

Population models



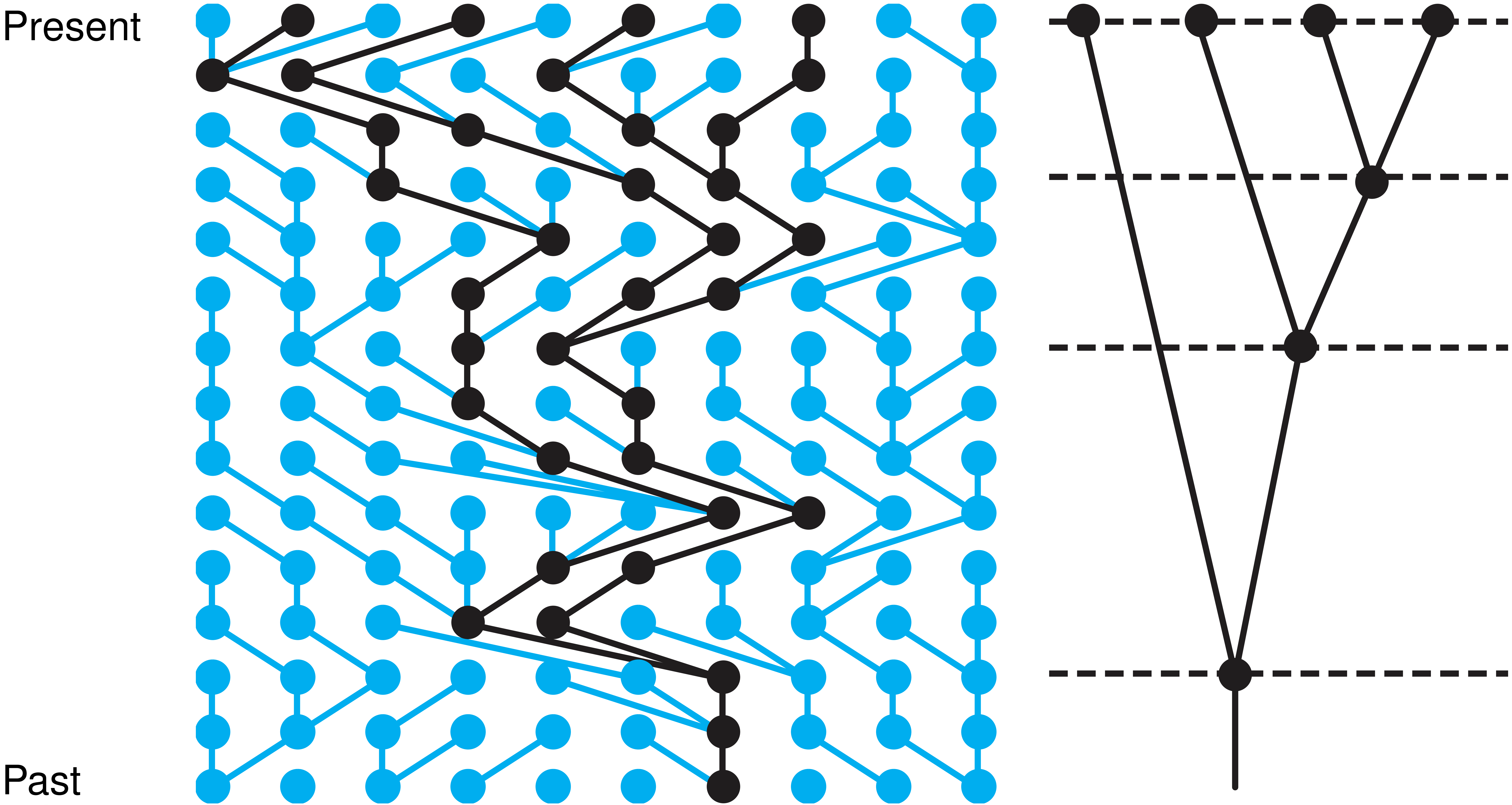
©2019 Peter Beerli

Coalescence theory



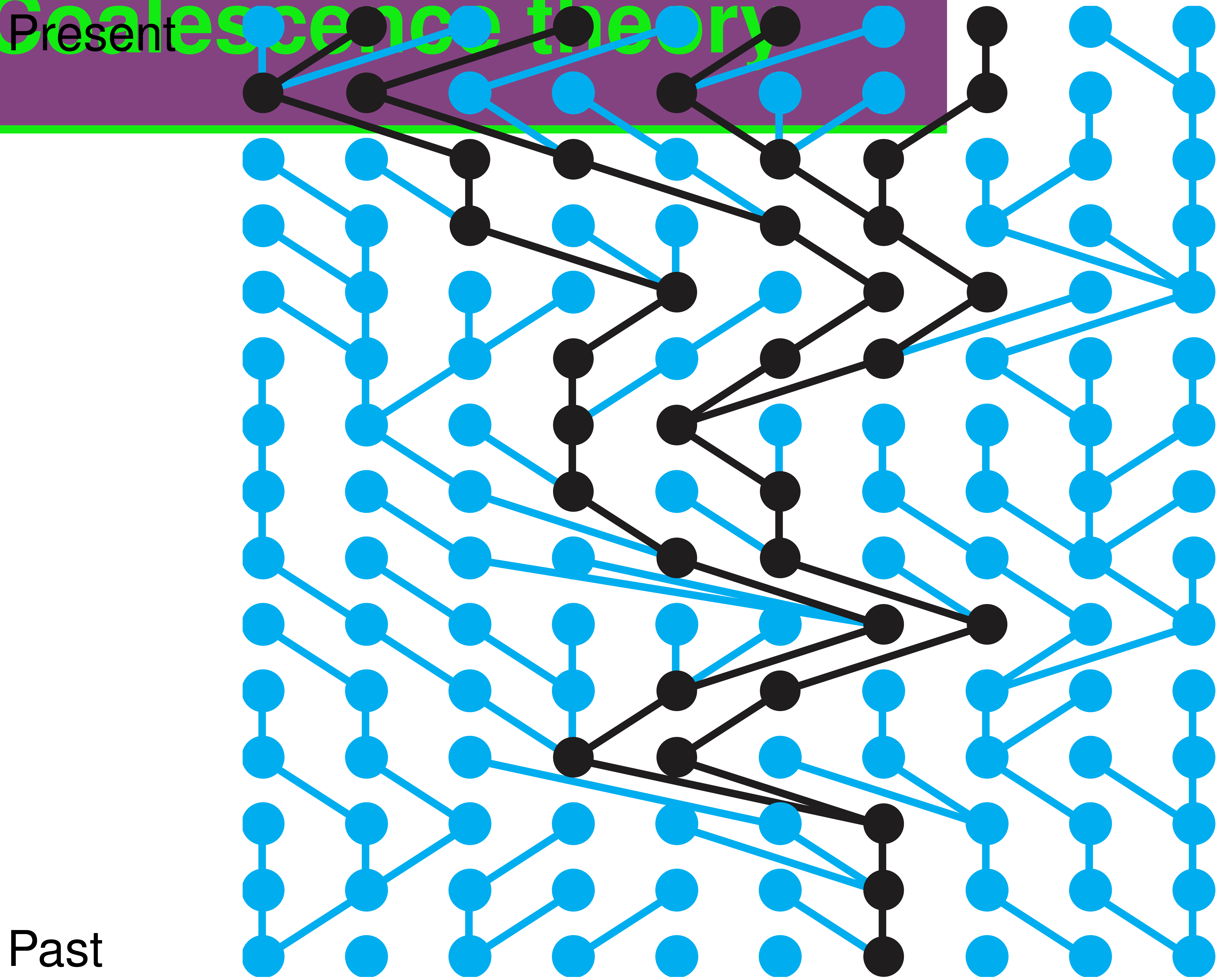
©2019 Peter Beerli

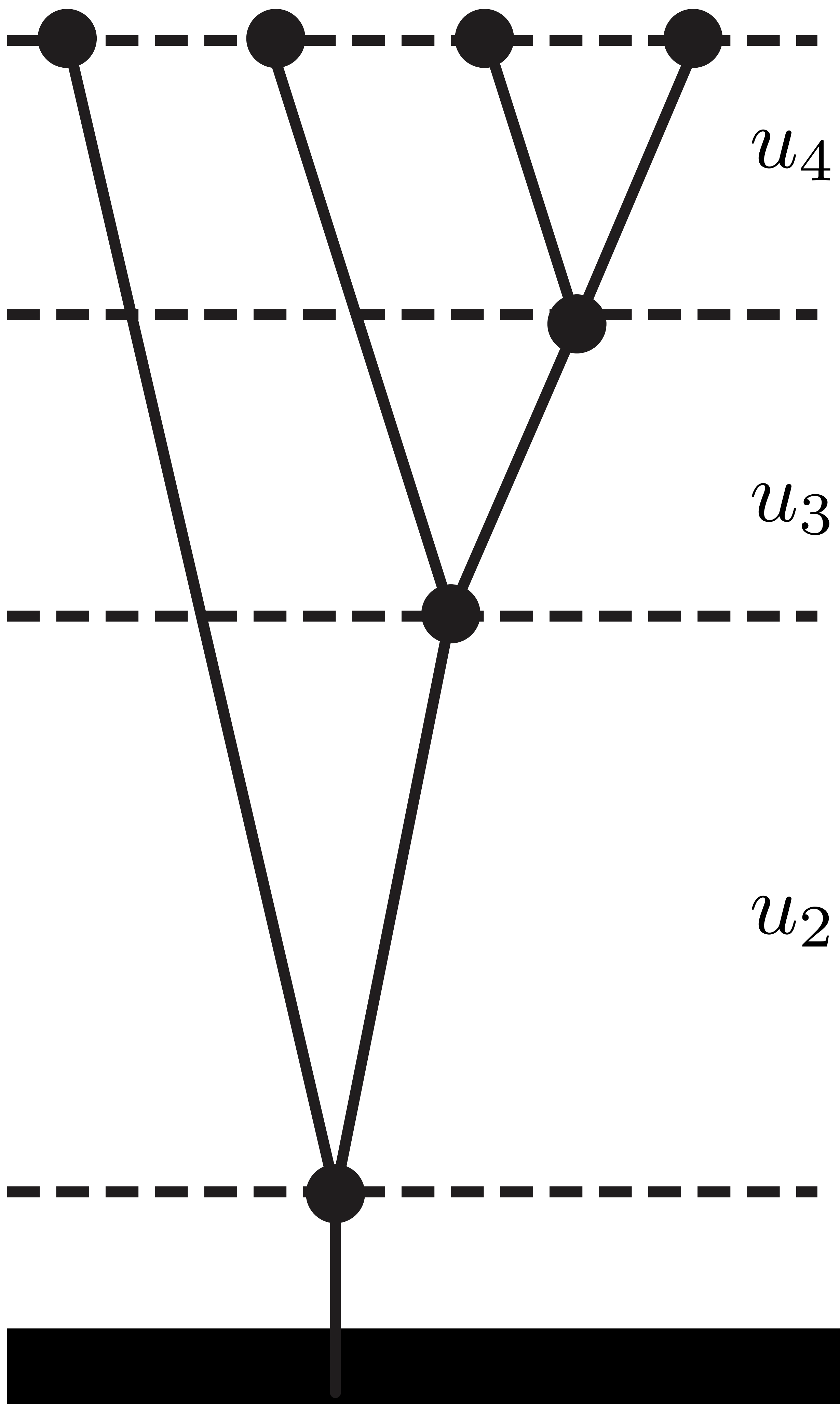
Coalescence theory



©2019 Peter Beerli

Coalescence theory



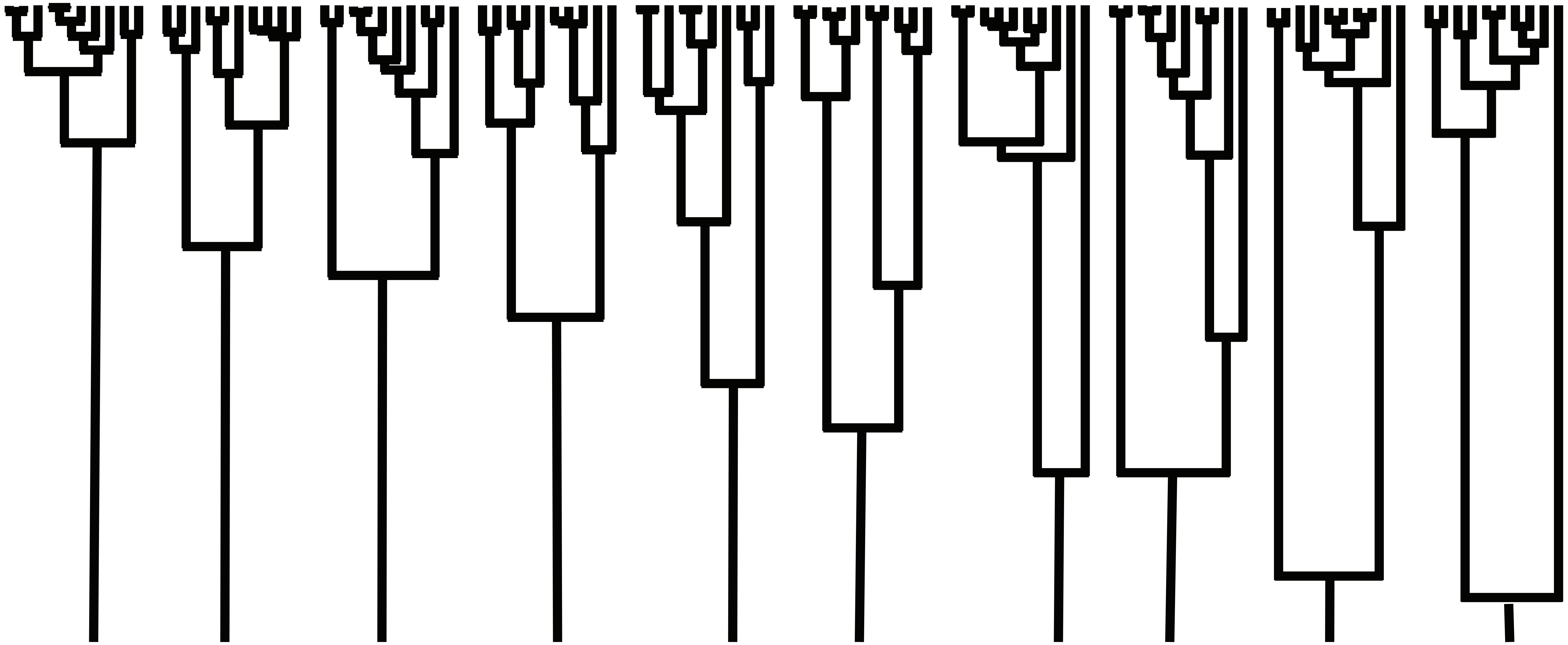


The time intervals u_k follows an exponential distribution with $\mathbb{E}(u_k) = \frac{\Theta}{k(k-1)}$

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

©2019 Peter Beerli

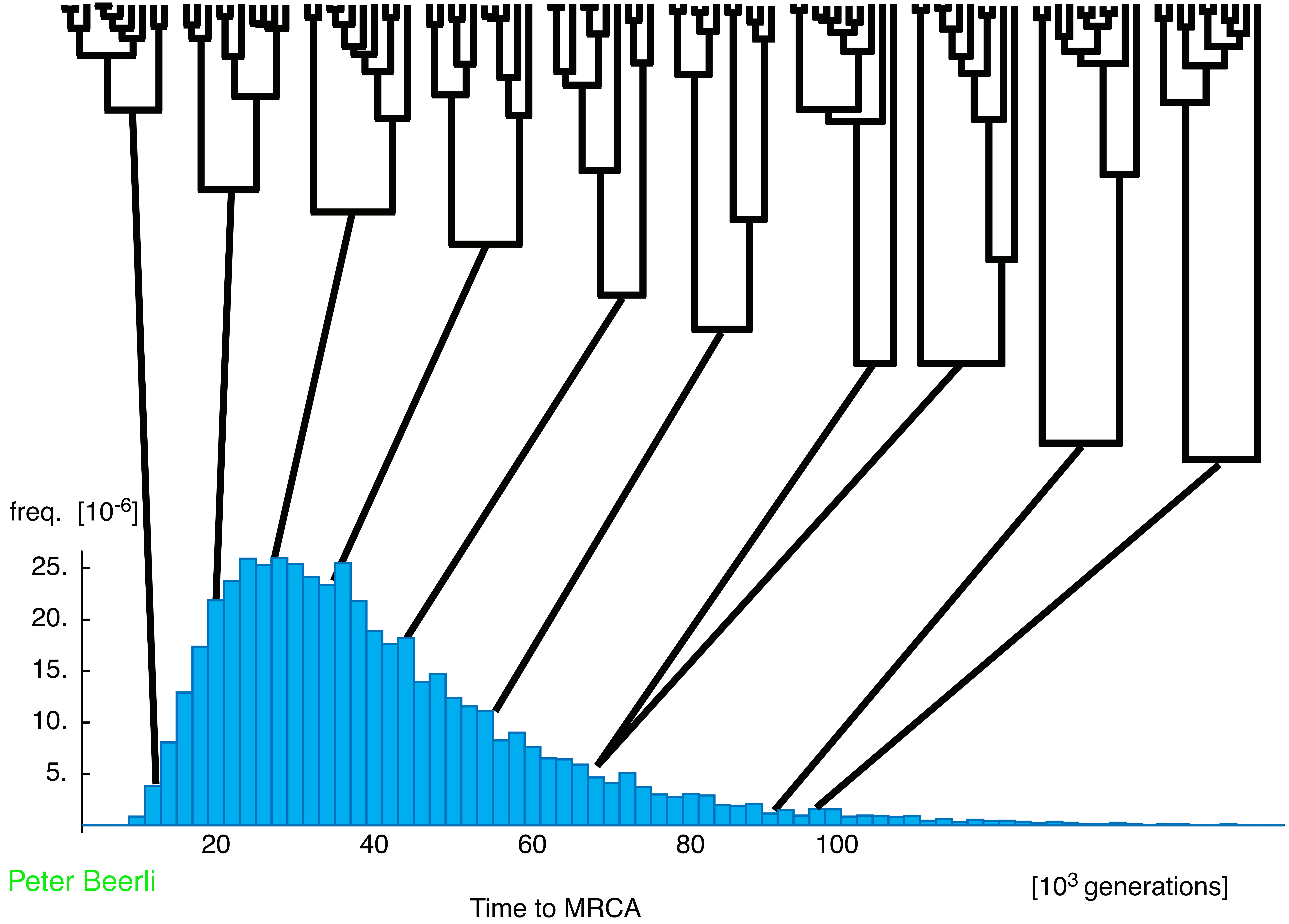
Variability of the coalescent process



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

©2019 Peter Beerli

Variability of the coalescent process



ⓒ2019 Peter Beerli

Population Parameter Inference

©2019 Peter Beerli

Wikimedia: Neon sign at Autonomy in Cambridge UK

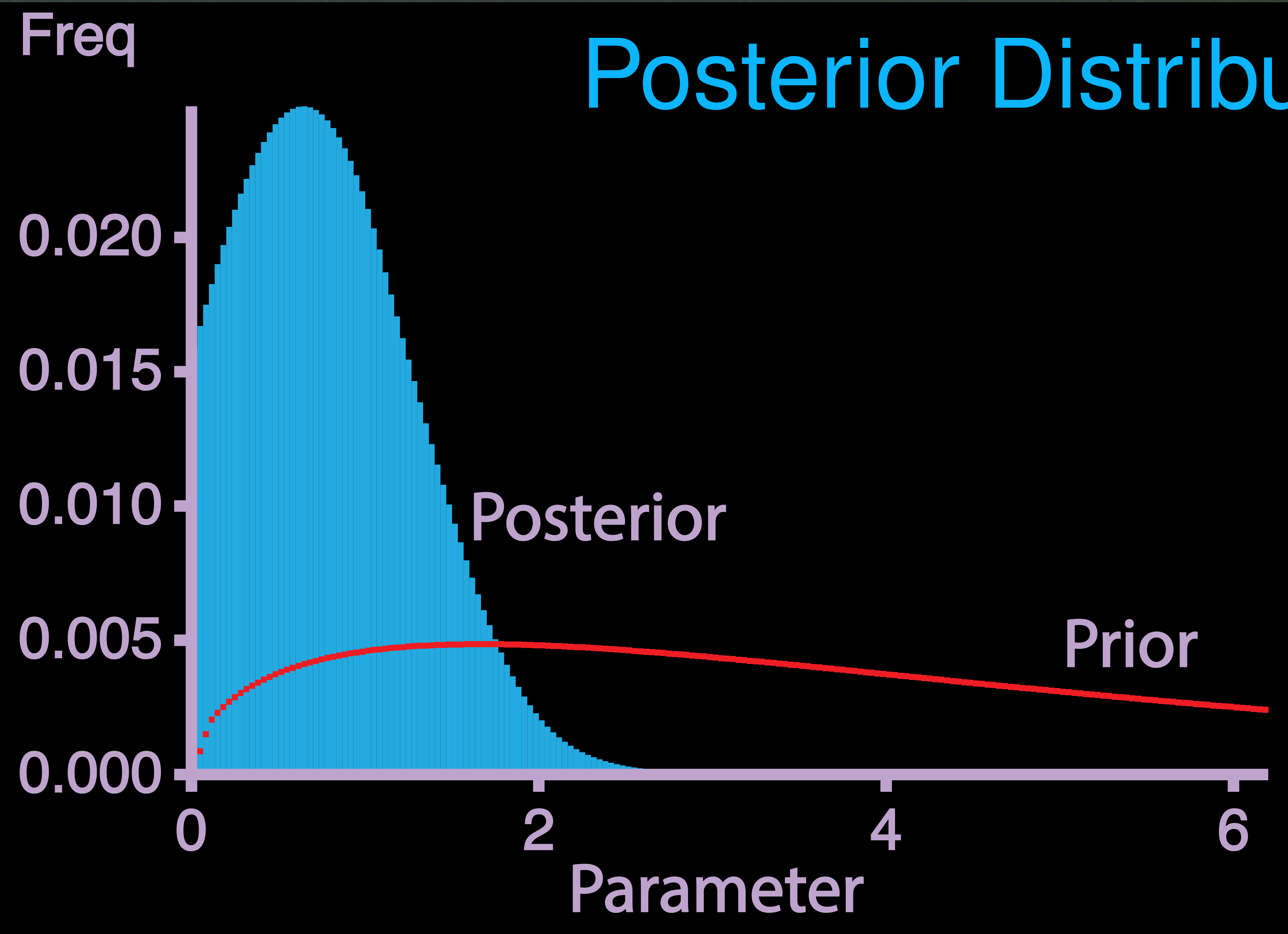
Population Parameter Inference

Genetic Data

Mutation model

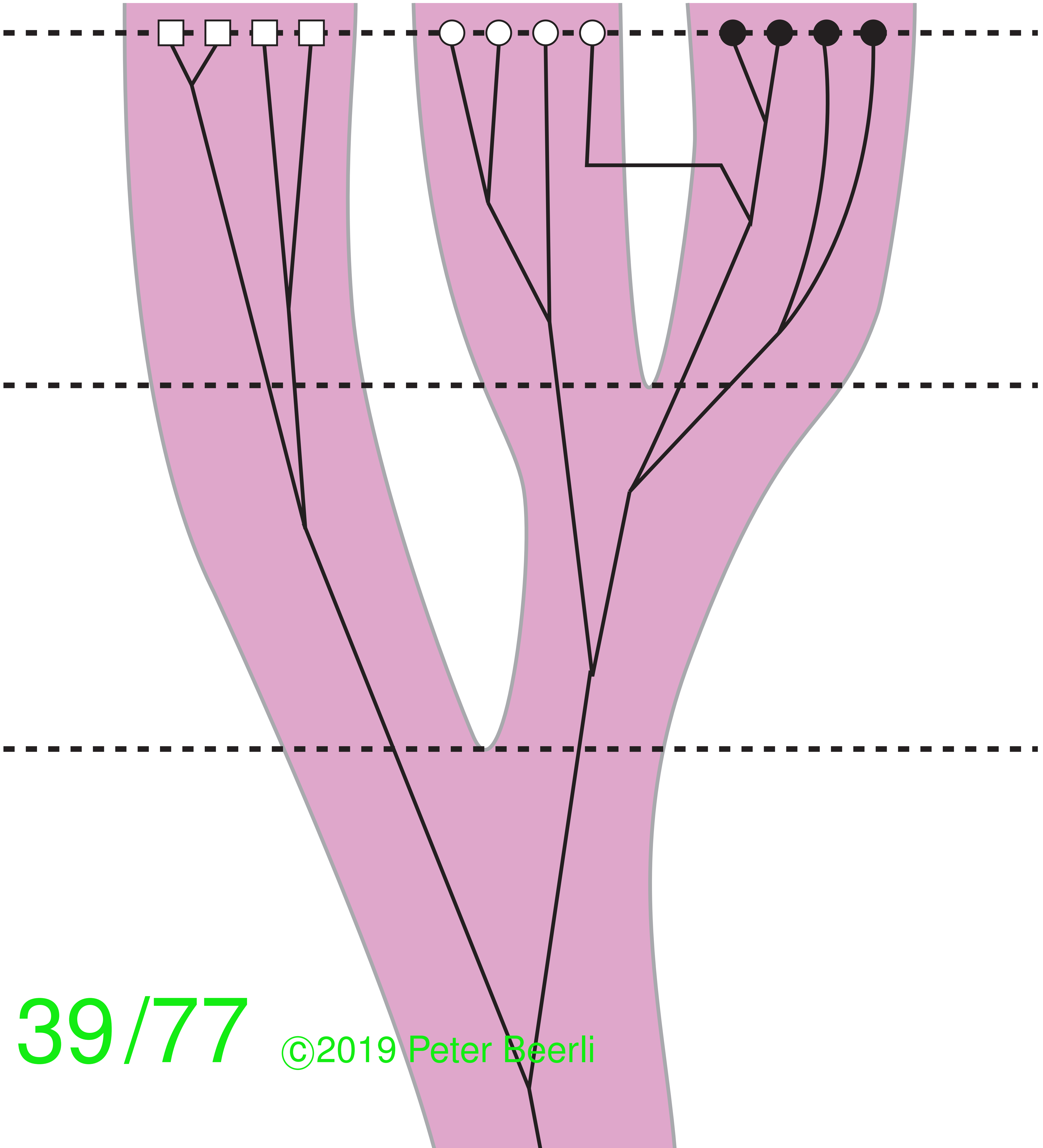
Population model

Posterior Distribution



©2019 Peter Beerli

Population model



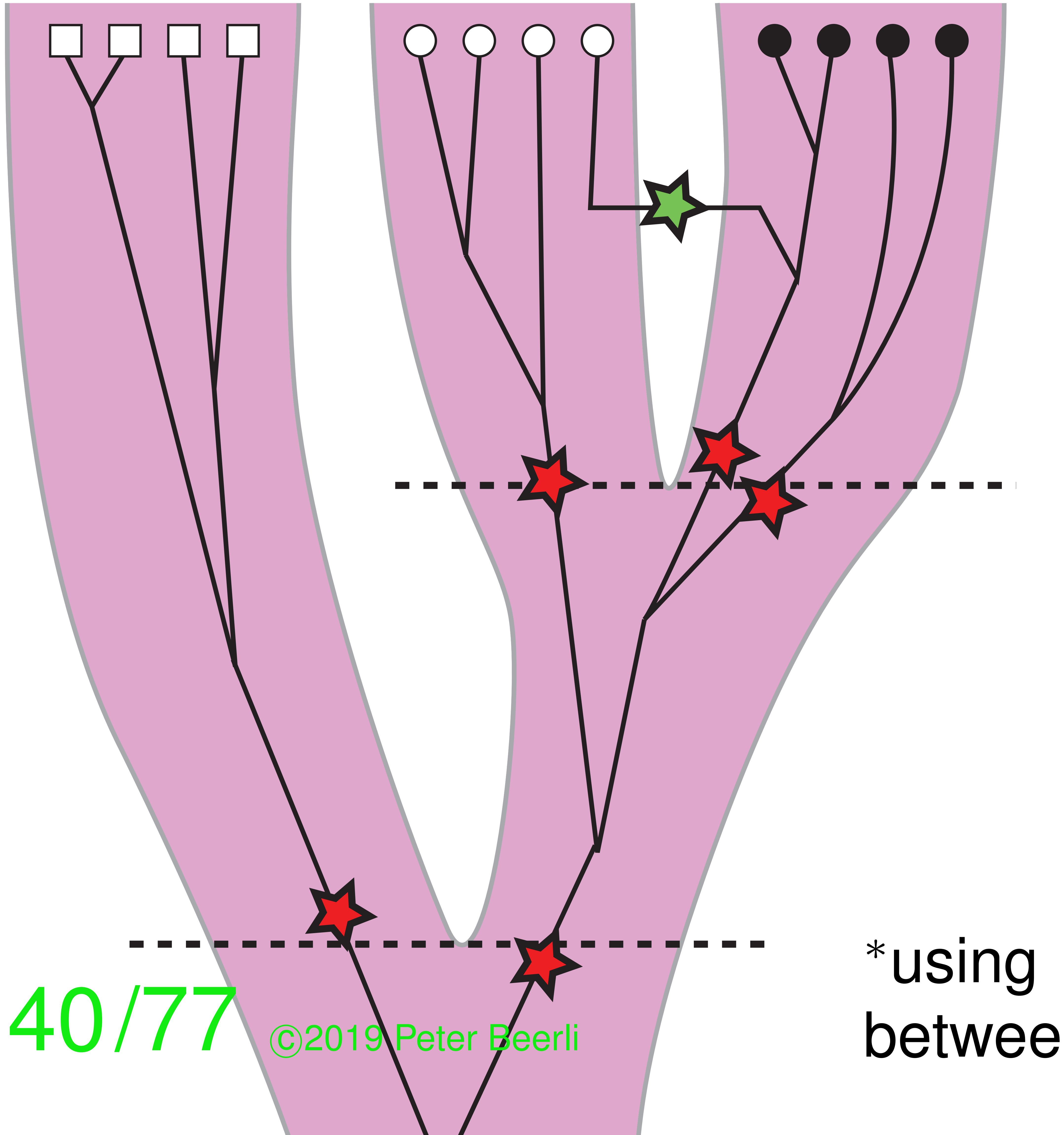
The relationship among individuals can be expressed, looking backward in time, by a waiting process where random lineages

- coalesce
- migrate between populations
- split off an ancestral population

©2019 Peter Beerli

Population genetics

Each of these processes can be expressed as a waiting time process with rate λ for N populations and k_j lineages in population j:



$$\lambda_{\text{two lineages coalesce}} = \sum_{j=1}^{N} \frac{k_j(k_j-1)}{4N}$$

$$\lambda_{\text{lineages migrate}} = \sum_{j=1}^{N} \sum_{i=1,i\neq j}^{N} k_j m_{ij}$$

$$\lambda_{\text{lineages split off}^*} = \frac{k\sqrt{\frac{2}{\pi}}e^{\frac{(t-\mu)^2}{2\sigma^2}}}{\sigma\left(1-\text{erf}\left(\frac{t-\mu}{\sqrt{2}\sigma}\right)\right)}$$

[*]using a Normal distribution to model the splitting time between two populations.

©2019 Peter Beerli

Combining the parts

$$P(\boldsymbol{\Theta}|\mathbf{D_1},\mathbf{D_2},...,\mu) = \frac{P(\boldsymbol{\Theta})P(\mathbf{D_1},\mathbf{D_2},...|\boldsymbol{\Theta})}{P(\mathbf{D_1},\mathbf{D_2},...)} = \frac{P(\boldsymbol{\Theta})\int_G P(G|\boldsymbol{\Theta})\prod_i^{n\mathrm{Loci}} P(\mathbf{D_i}|G,\mu)dG}{\int_\Theta P(\boldsymbol{\Theta})\int_G P(G|\boldsymbol{\Theta})\prod_i^{n\mathrm{Loci}} P(\mathbf{D_i}|G,\mu)dGd\Theta}$$

$$P(G|\boldsymbol{\Theta}) = \prod_{i=1}^{K} \lambda_x \exp(-t_i[\lambda_{\text{coalescence}} + \lambda_{\text{migration}} + \lambda_{\text{splitting}}])$$

$\boldsymbol{\Theta}$	vector of parameters for population size, migration and splitting parameters.
$\mathbf{D_1},\mathbf{D_2},...$	independent genetic sequence data,
μ	mutation model,
G	nuisance genealogies that we integrate out (we are interested in the parameters not the trees).
x	the particular event on the genealogy
K	number of total events on the genealogy

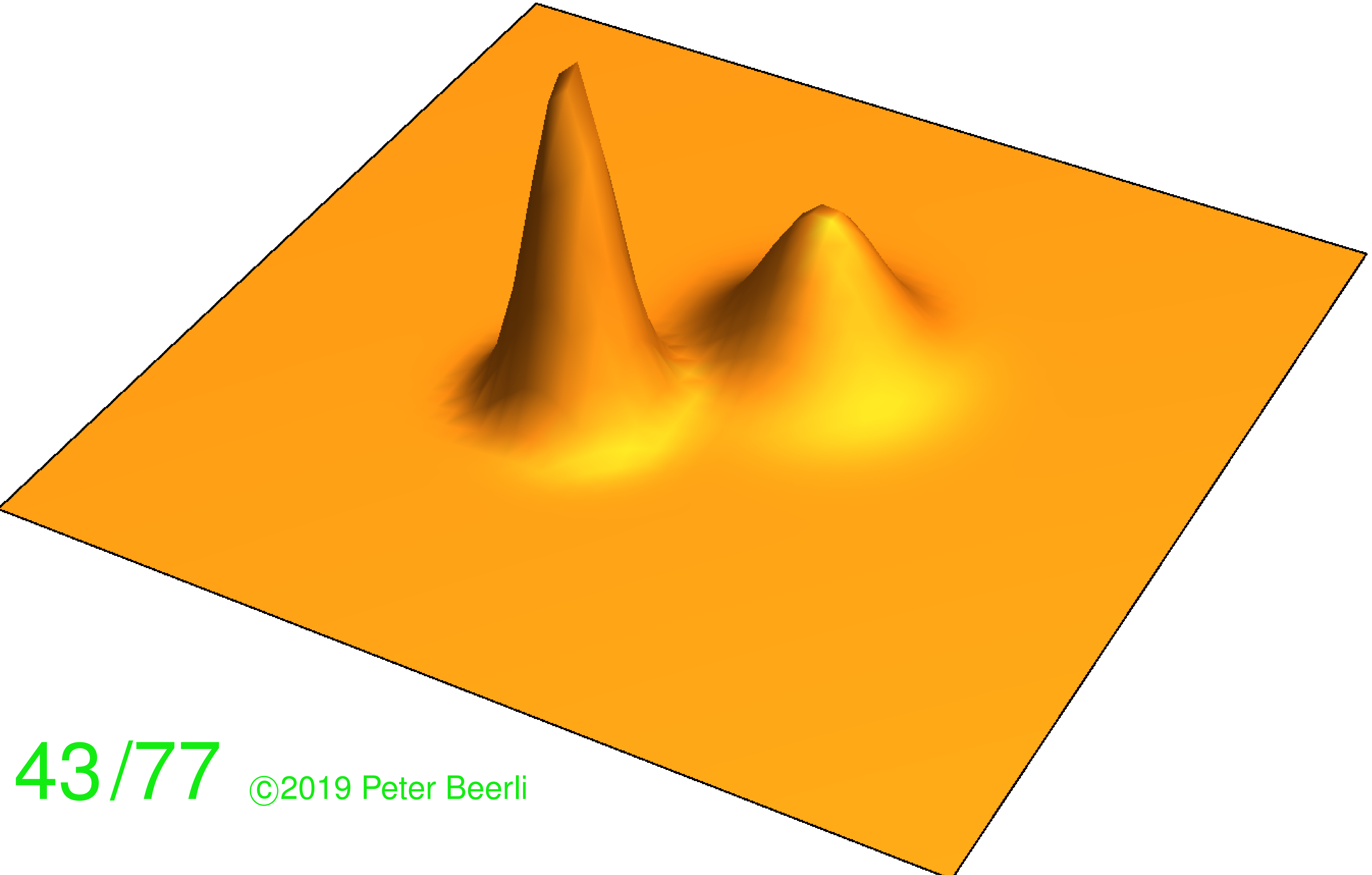
Finally....

$$p(D|\mathbf{\Theta}) = \int_G p(G|\mathbf{\Theta})p(D|G)dG$$

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

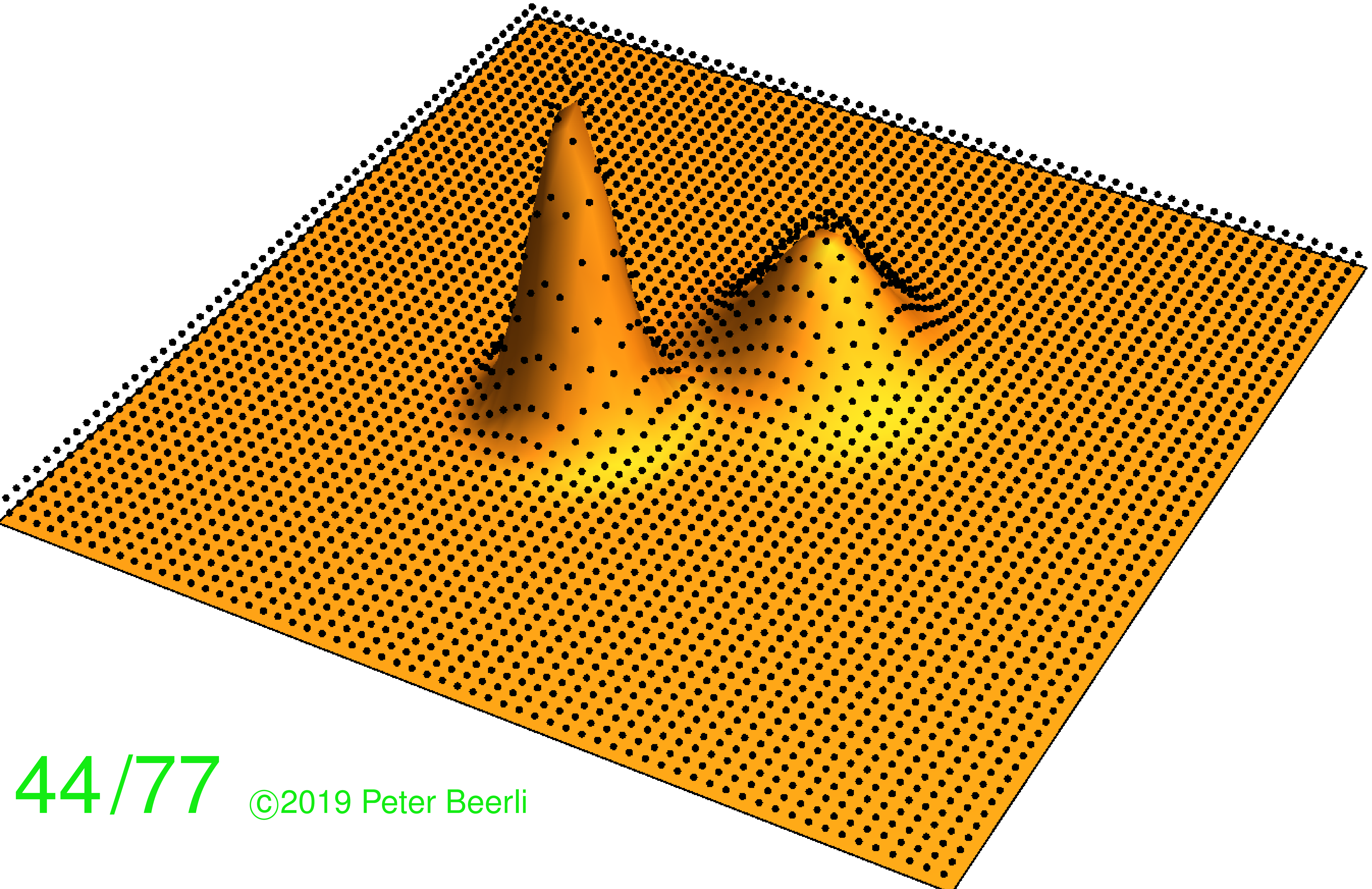
©2019 Peter Beerli

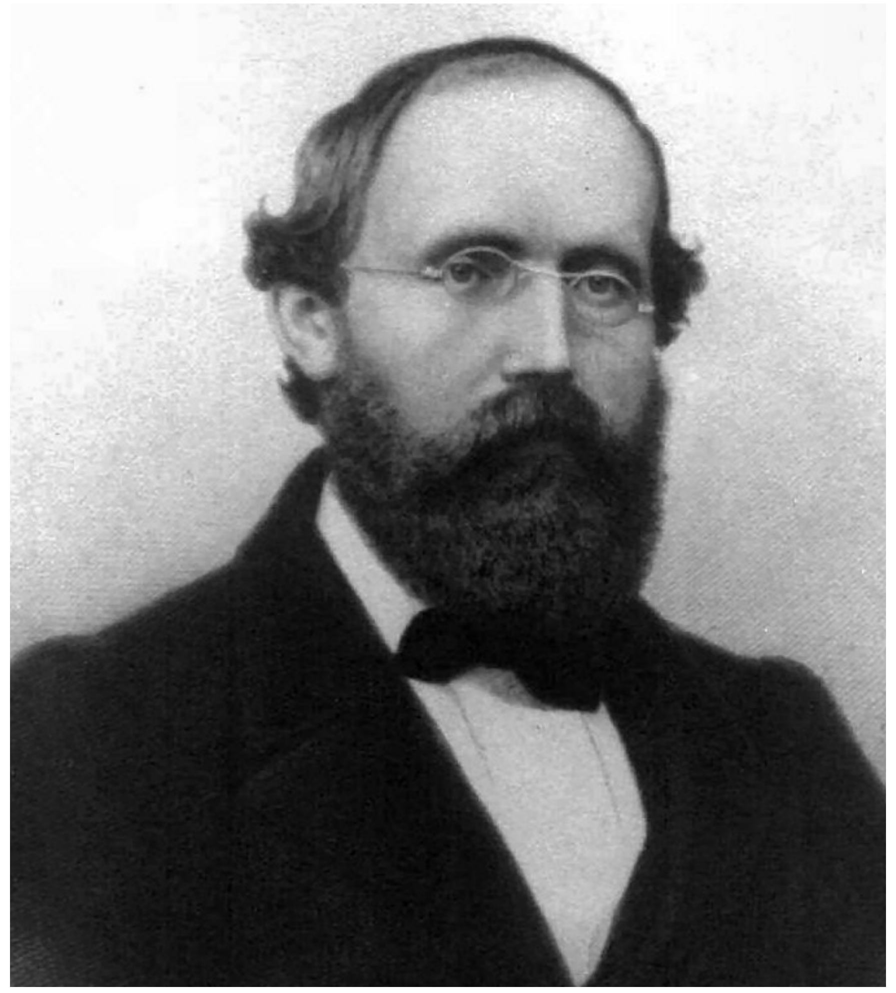
Naive integration approach



©2019 Peter Beerli

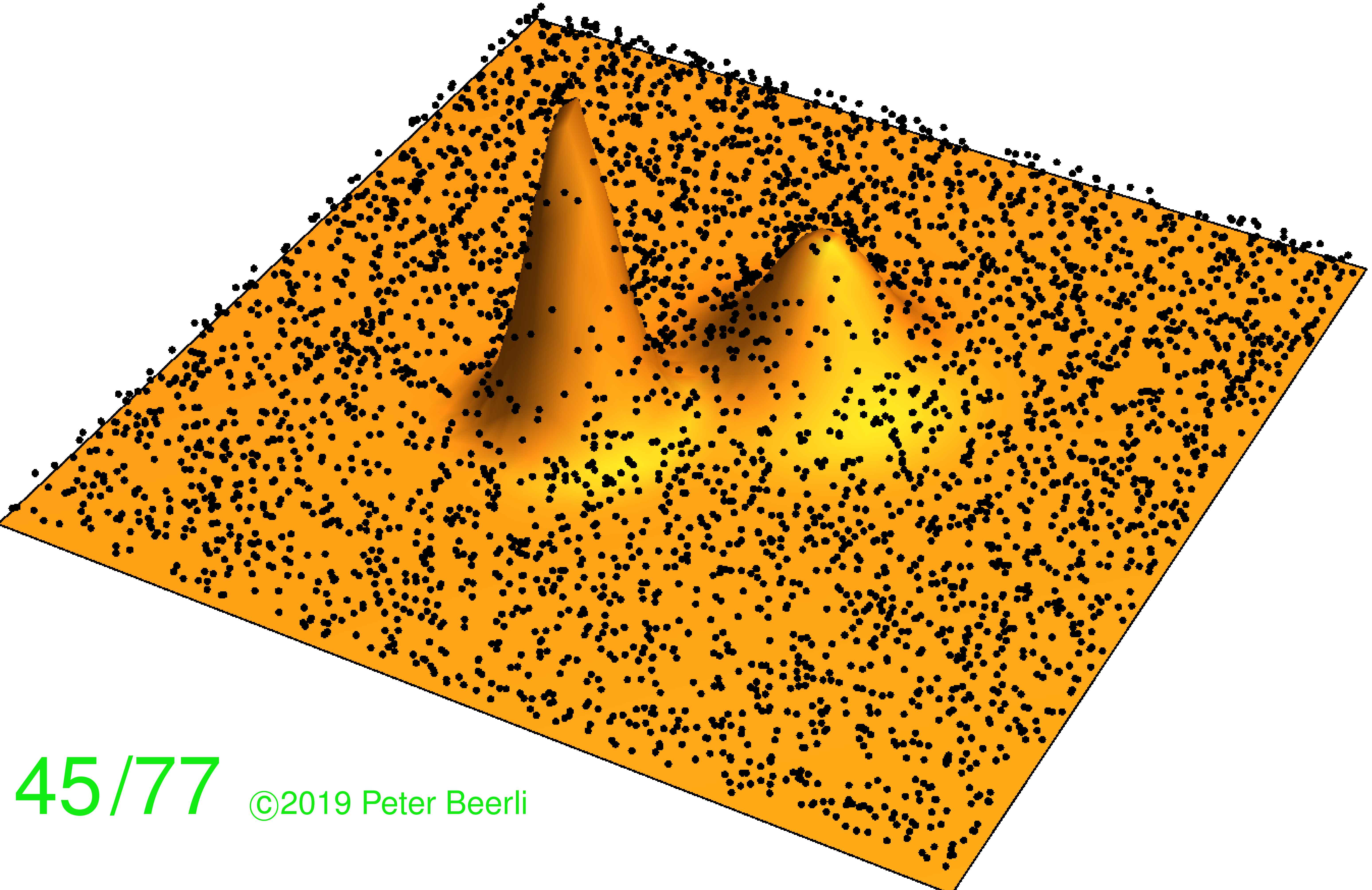
Naive integration approach



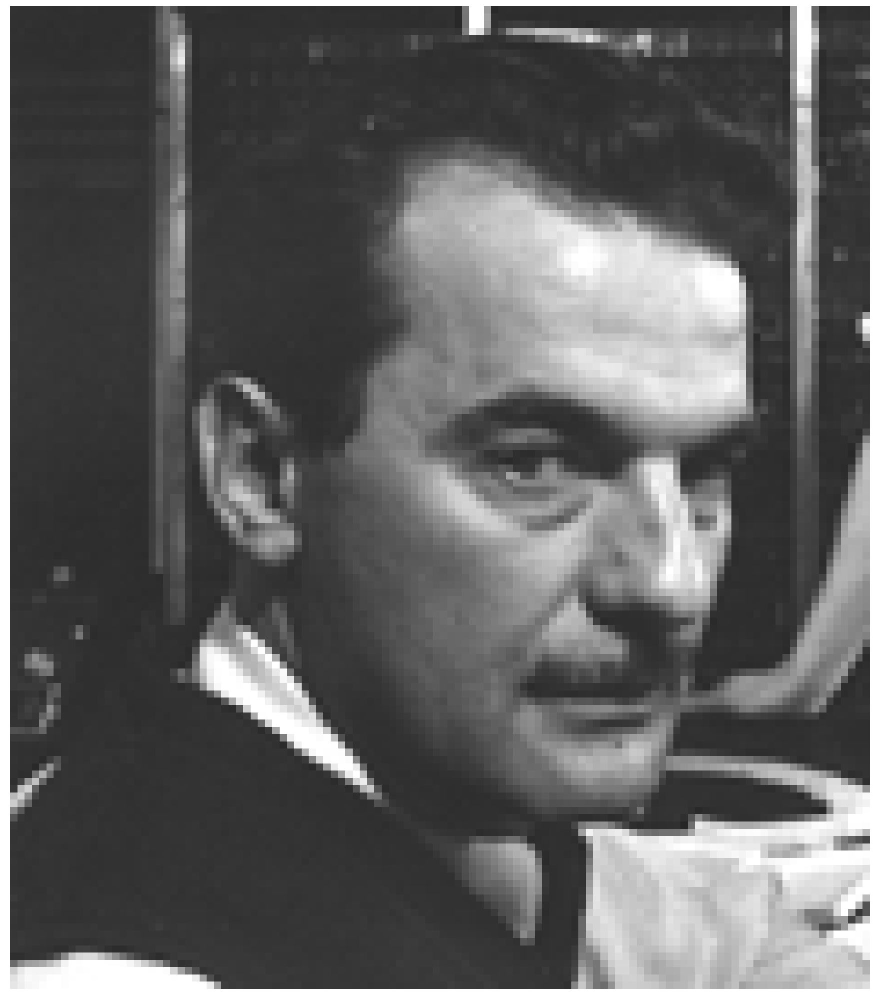


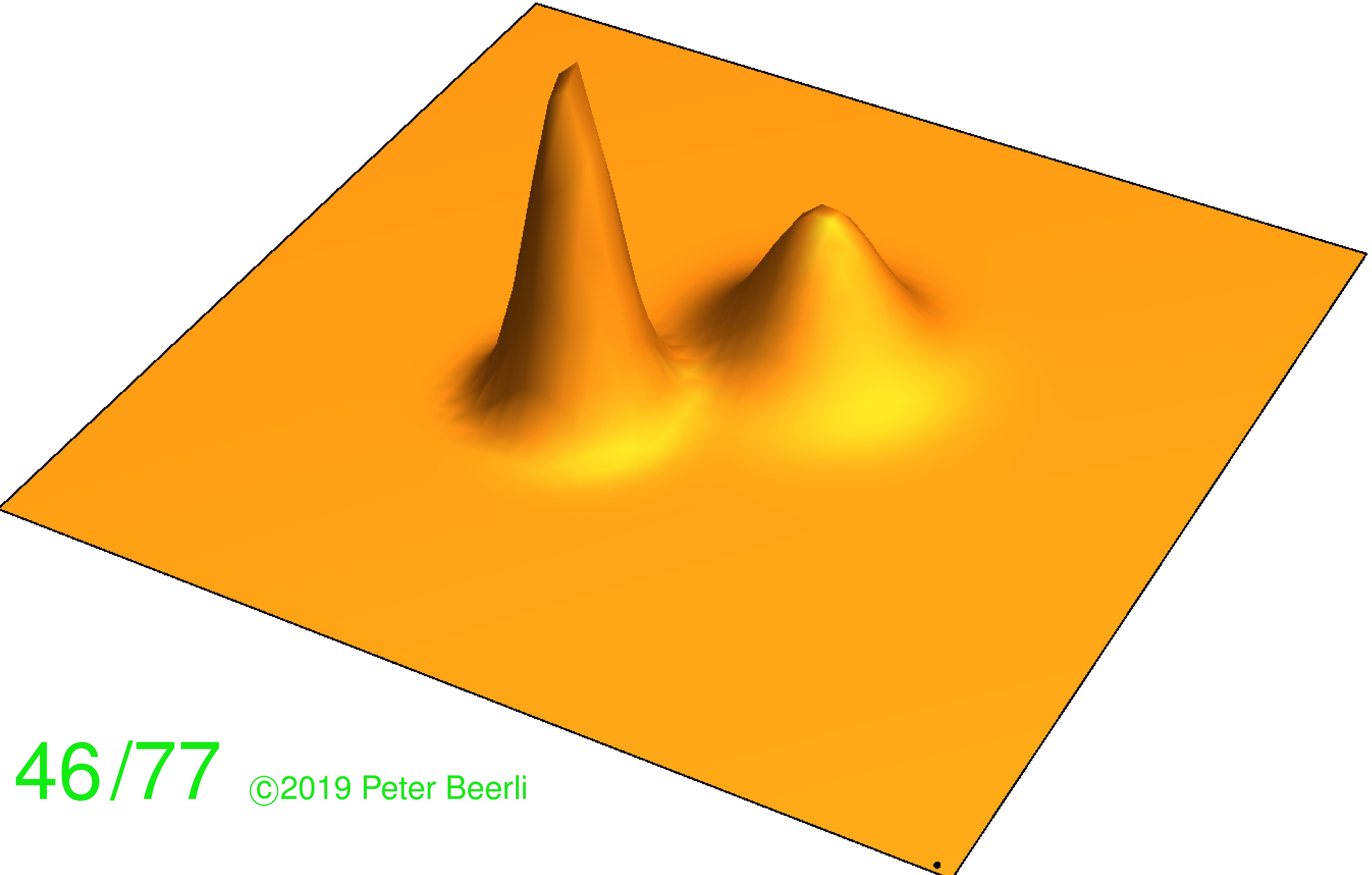
©2019 Peter Beerli

Another naive integration approach

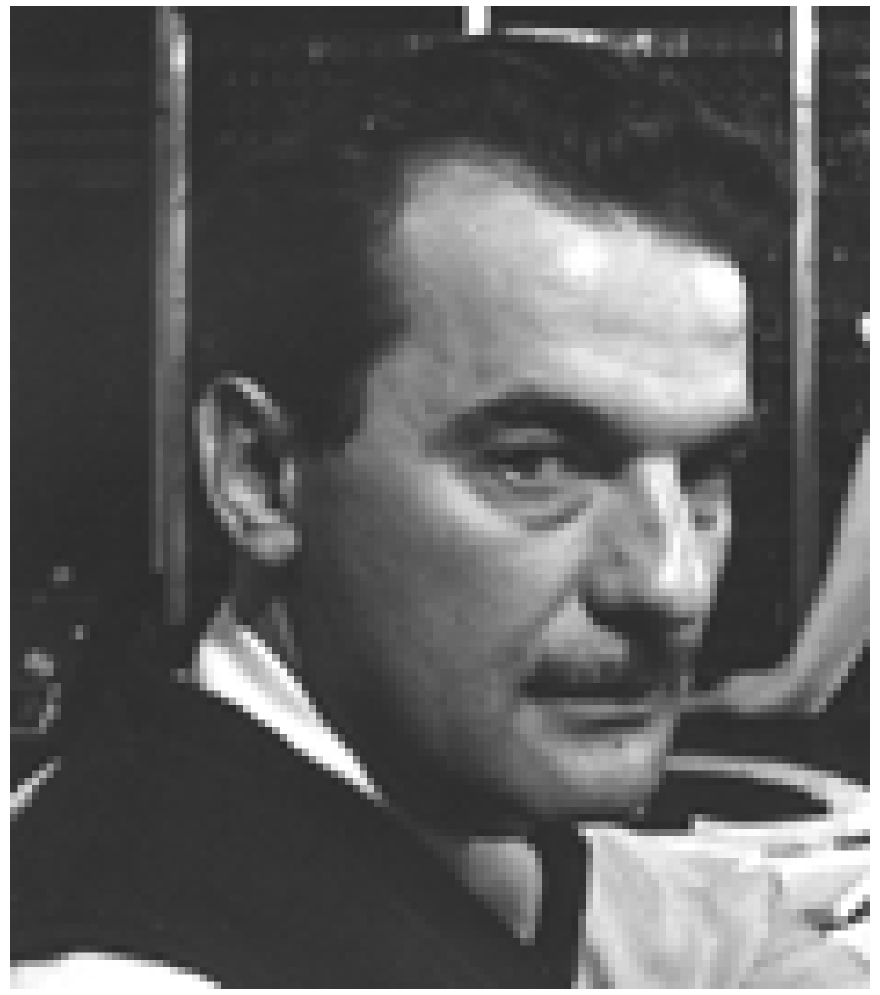


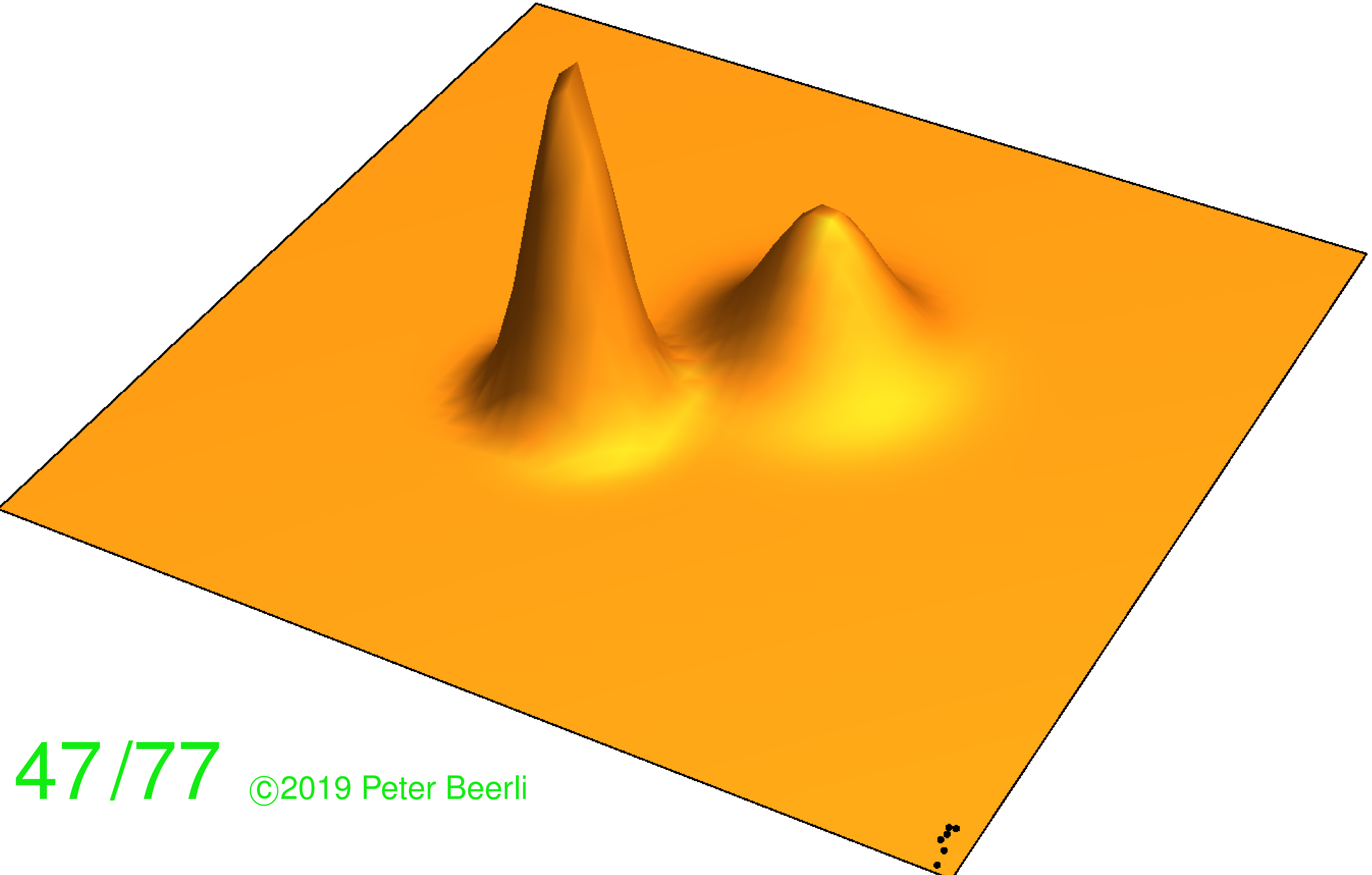
Metropolis-Hastings algorithm



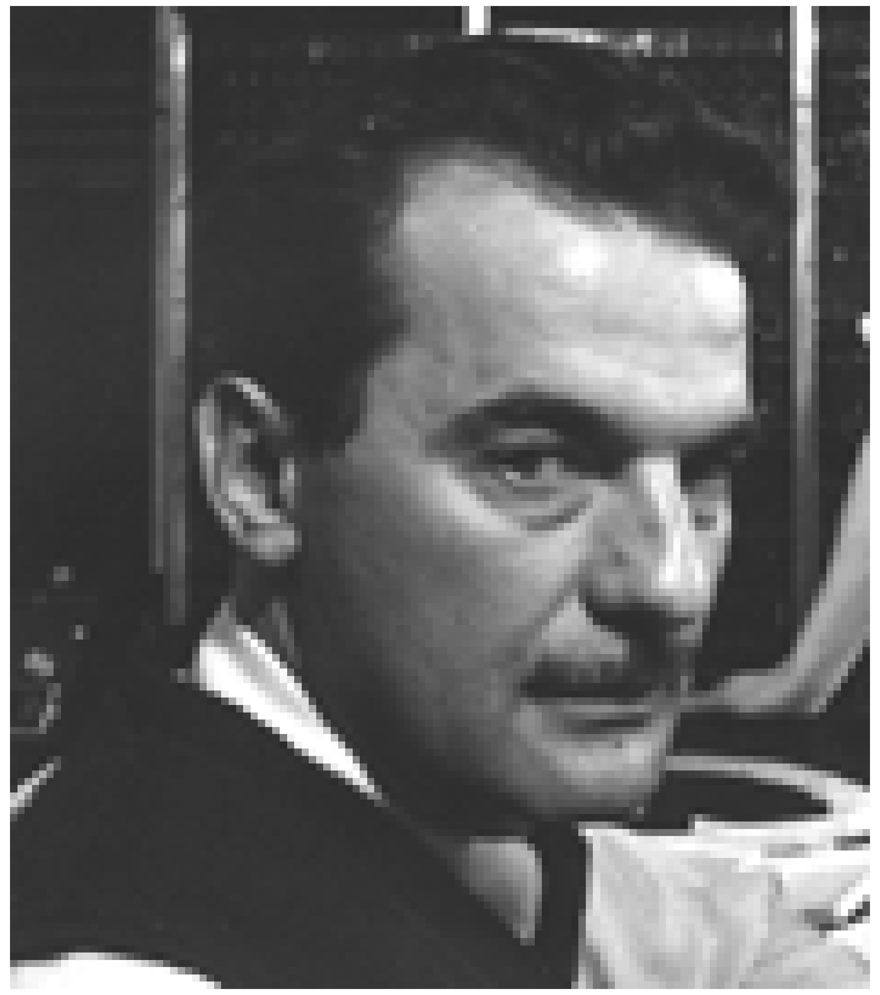


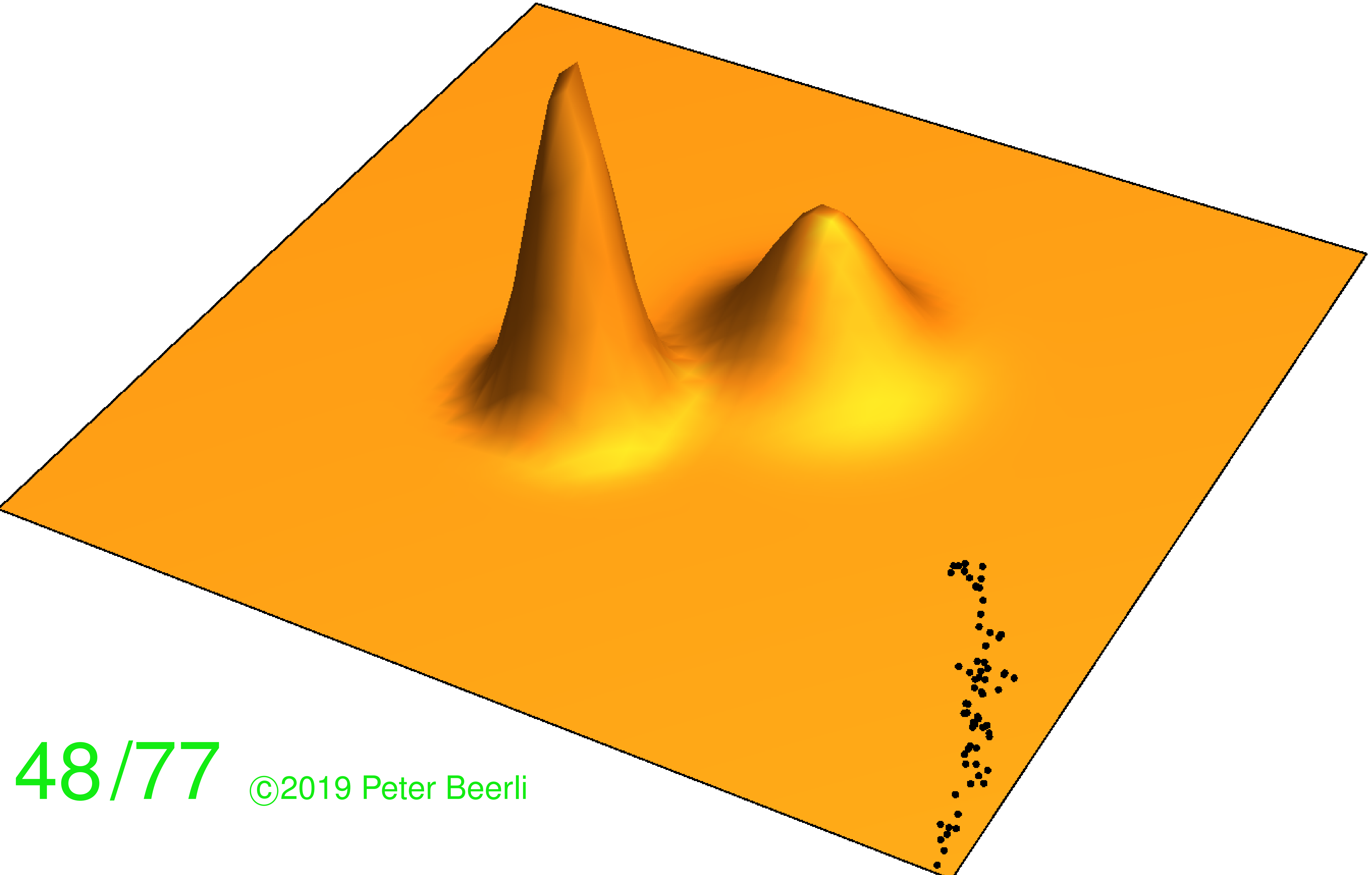
Metropolis-Hastings algorithm





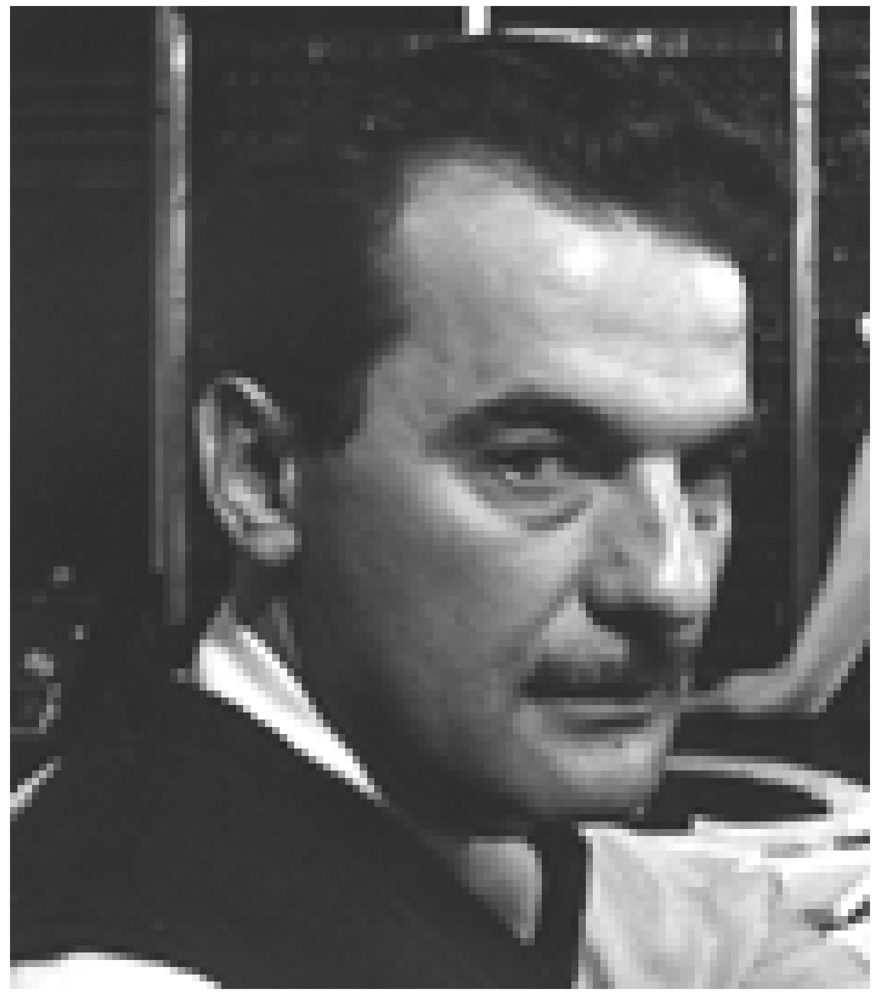
Metropolis-Hastings algorithm

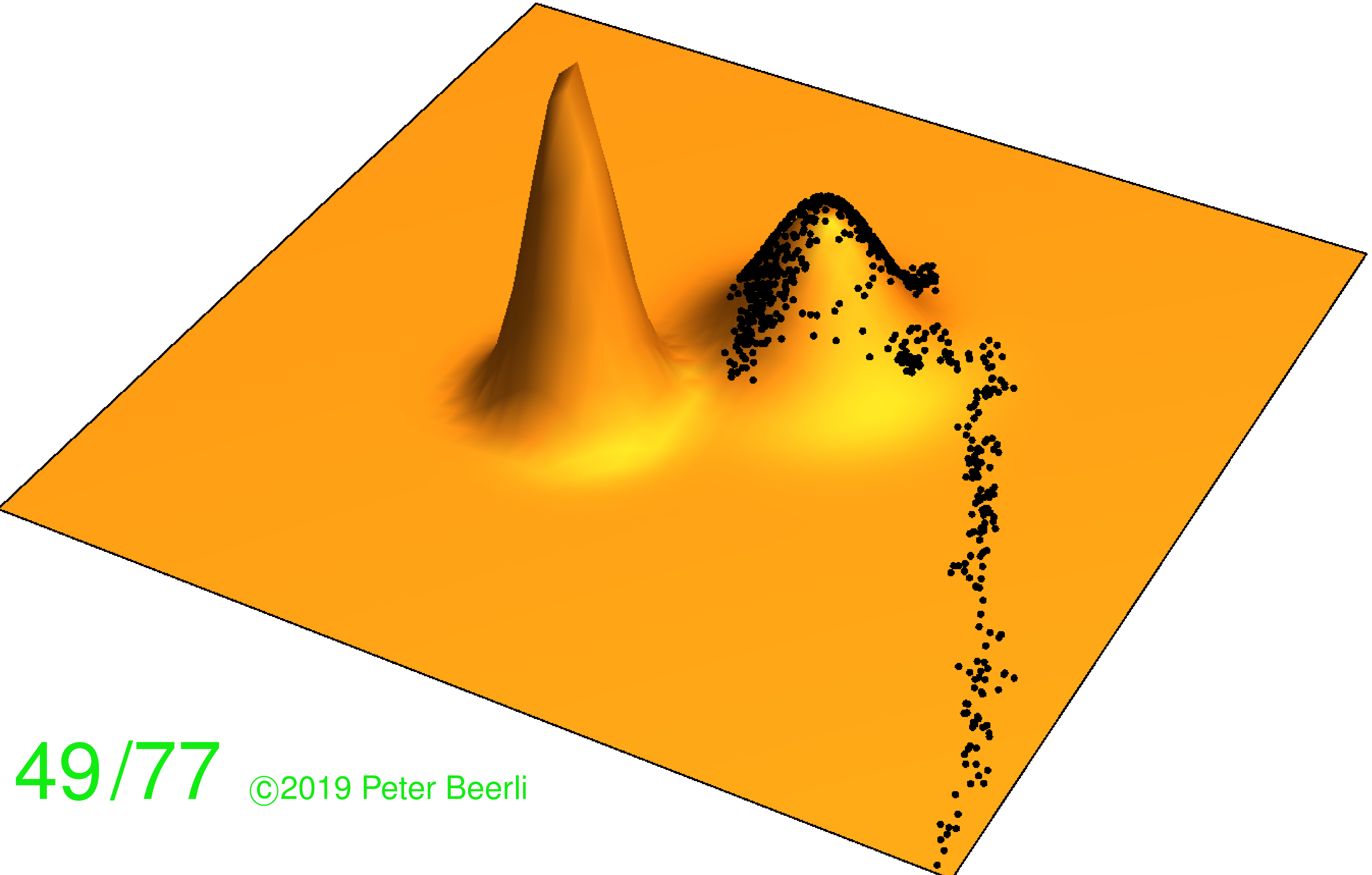




©2019 Peter Beerli

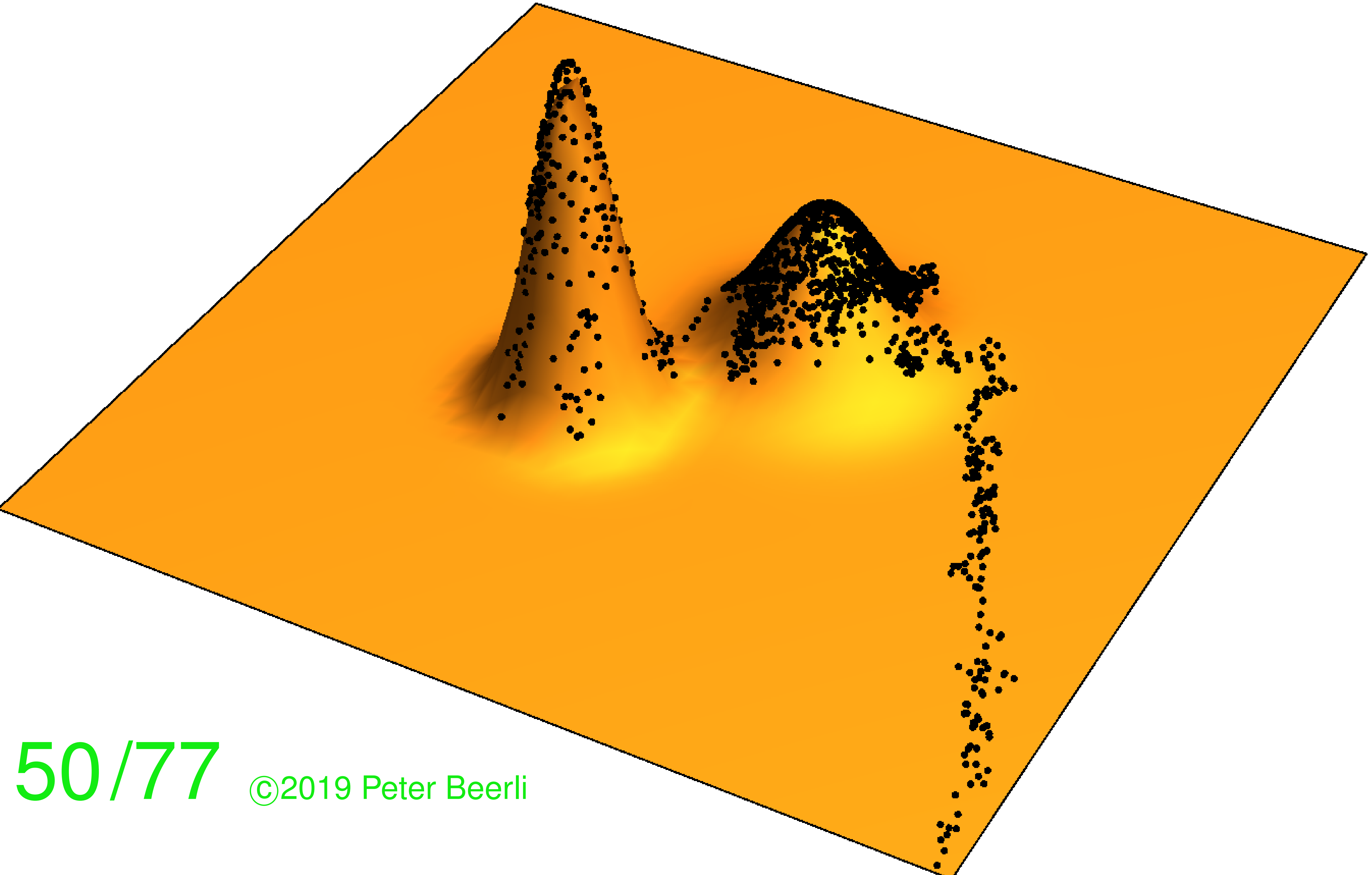
Metropolis-Hastings algorithm

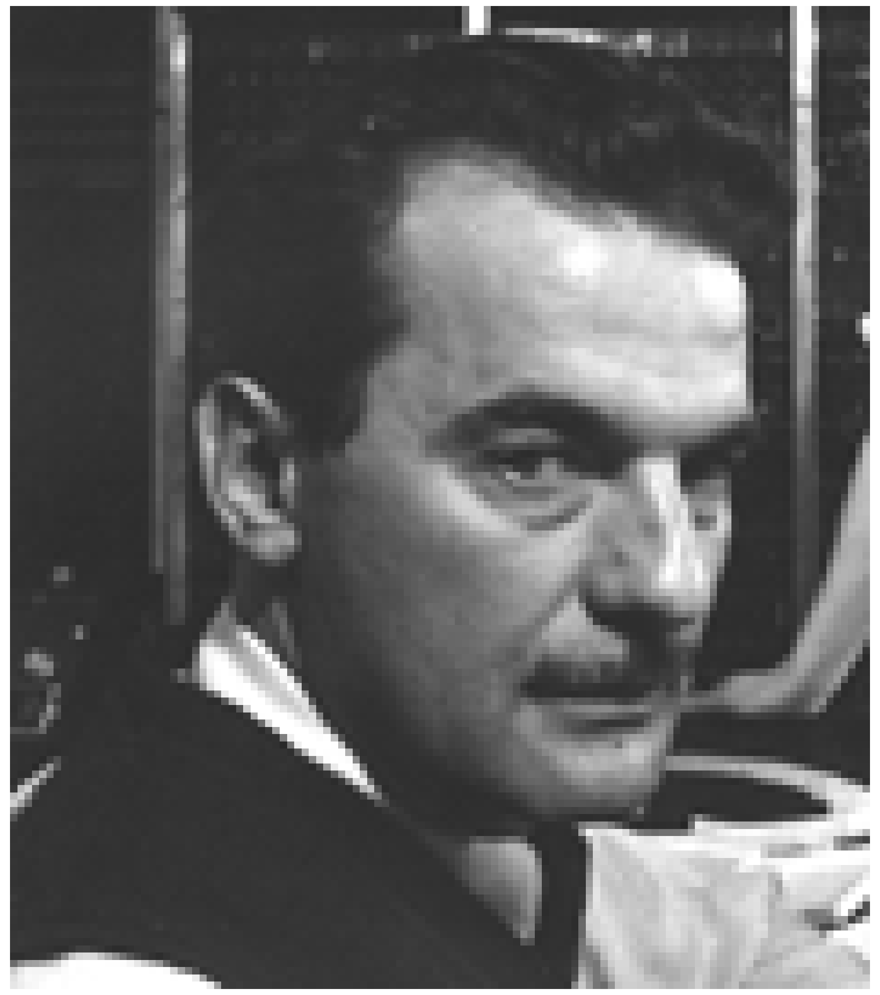




©2019 Peter Beerli

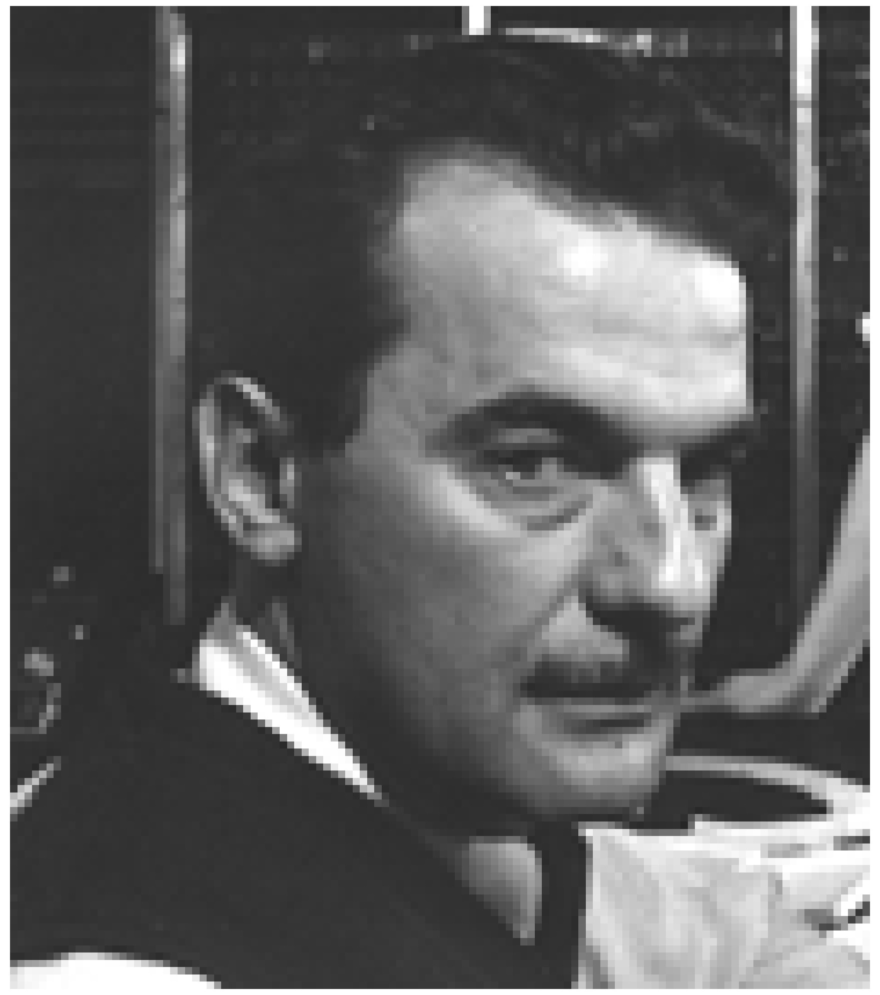
Metropolis-Hastings algorithm

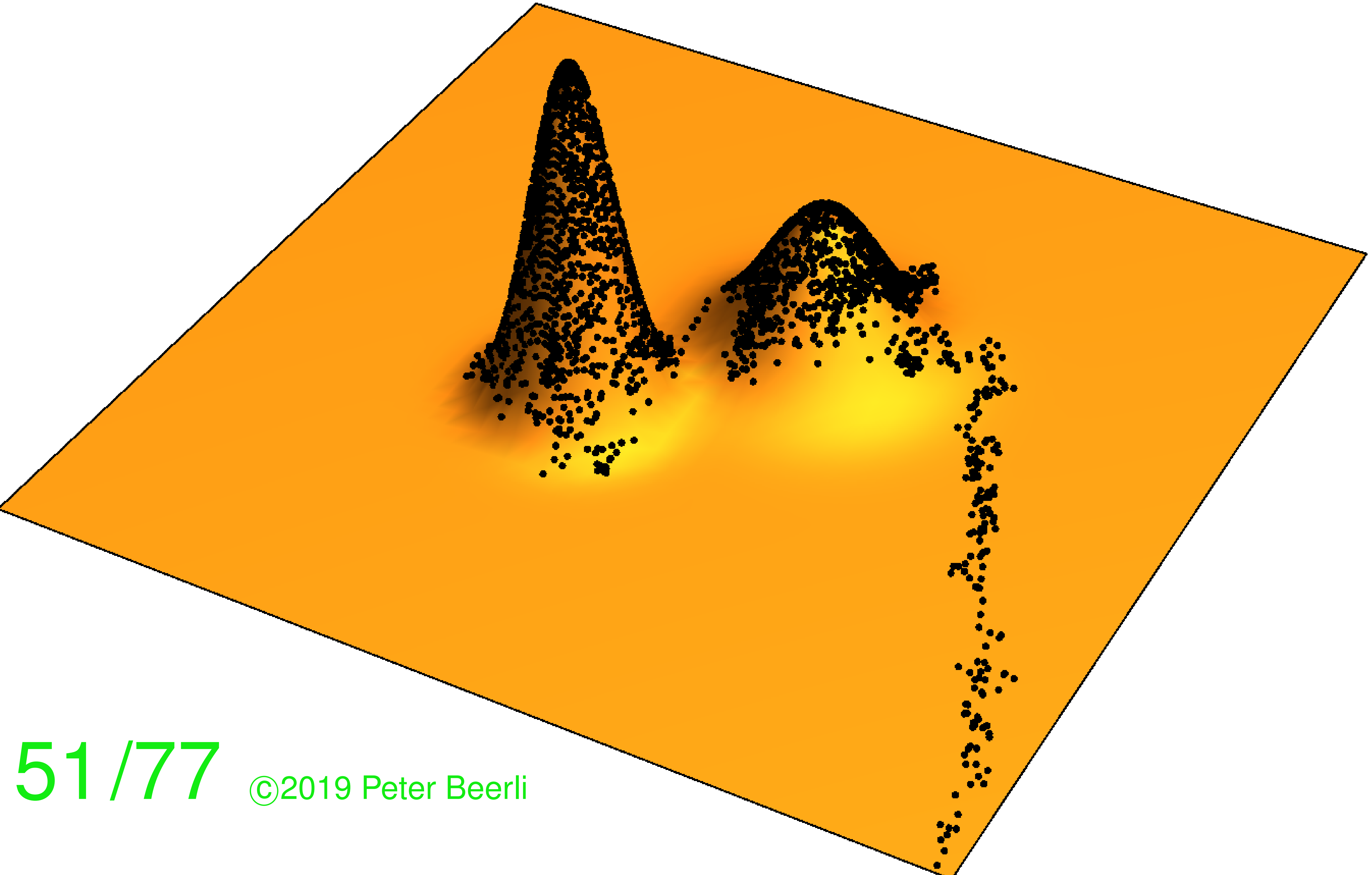




©2019 Peter Beerli

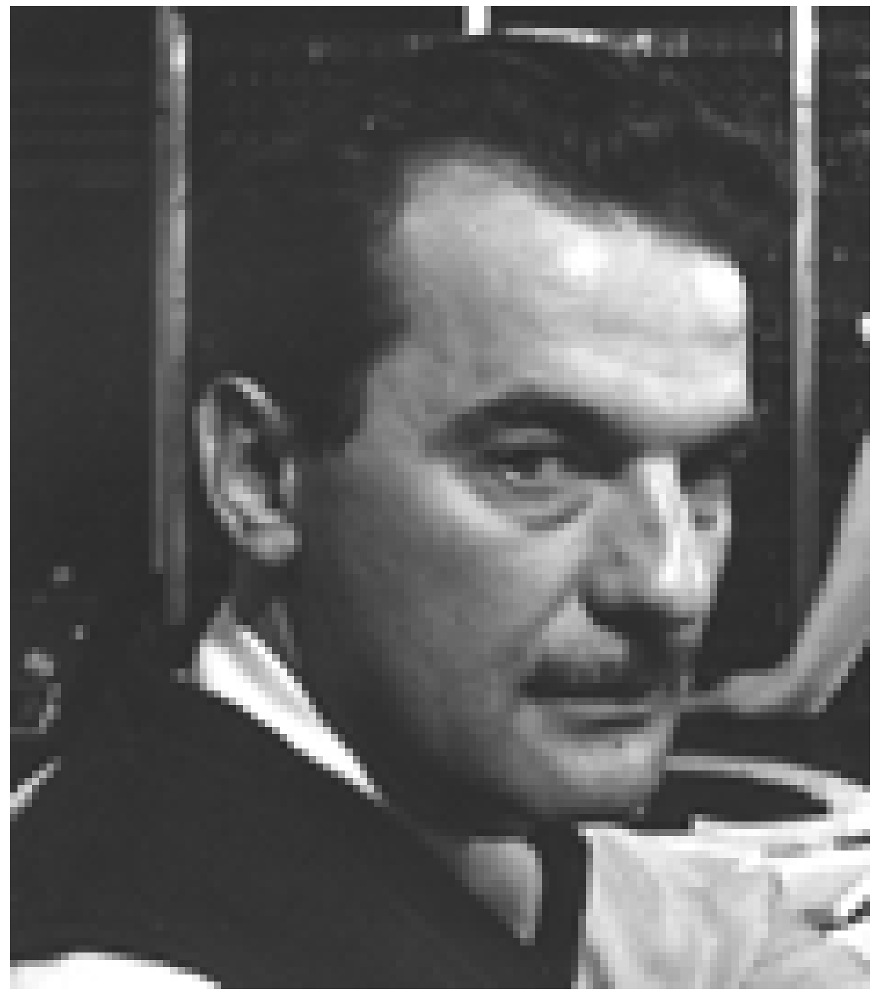
Metropolis-Hastings algorithm

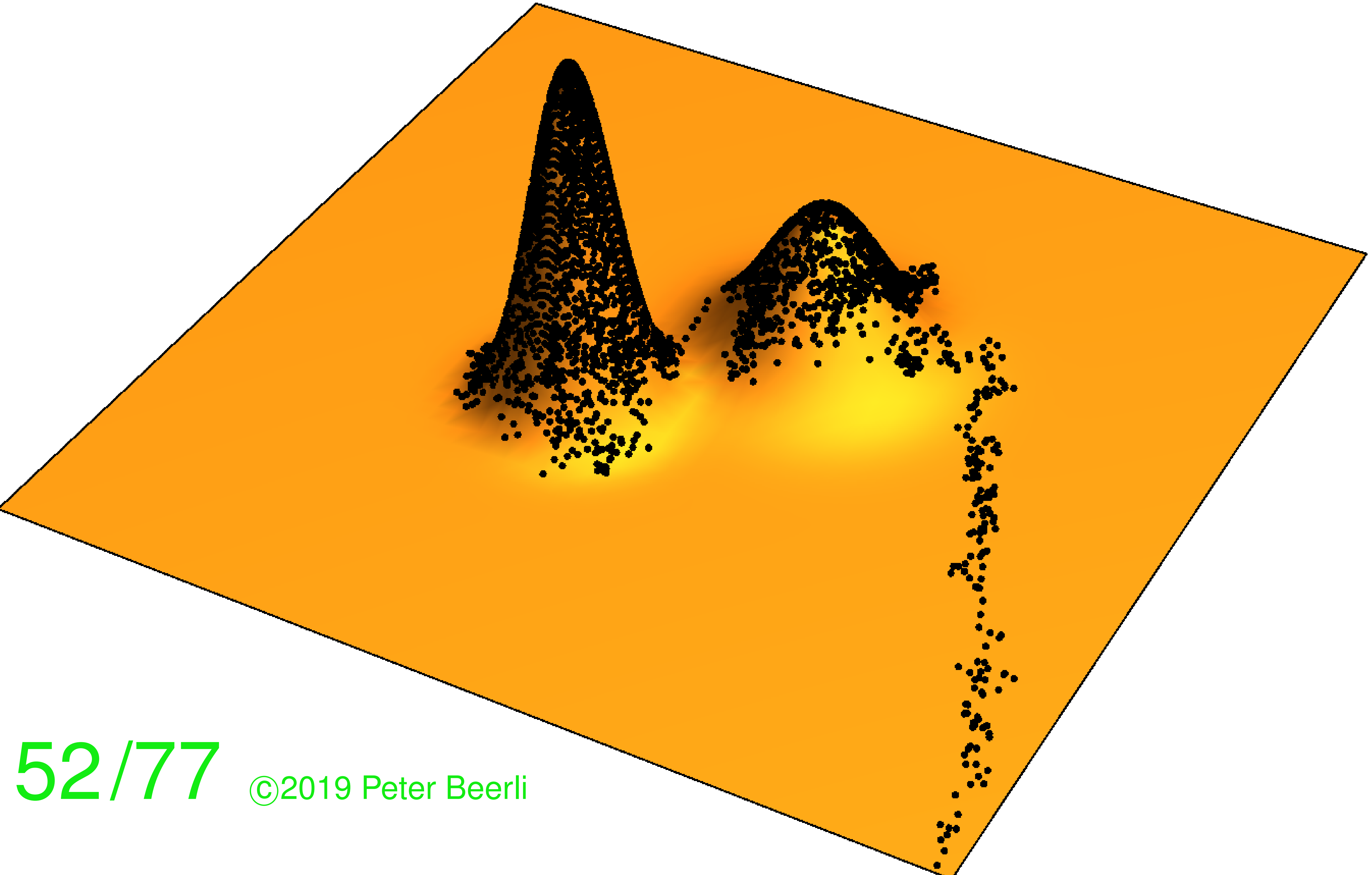




©2019 Peter Beerli

Metropolis-Hastings algorithm





©2019 Peter Beerli

So many models – so little time

©2019 Peter Beerli

Photo CC Wikimedia Wolfgang Sauber

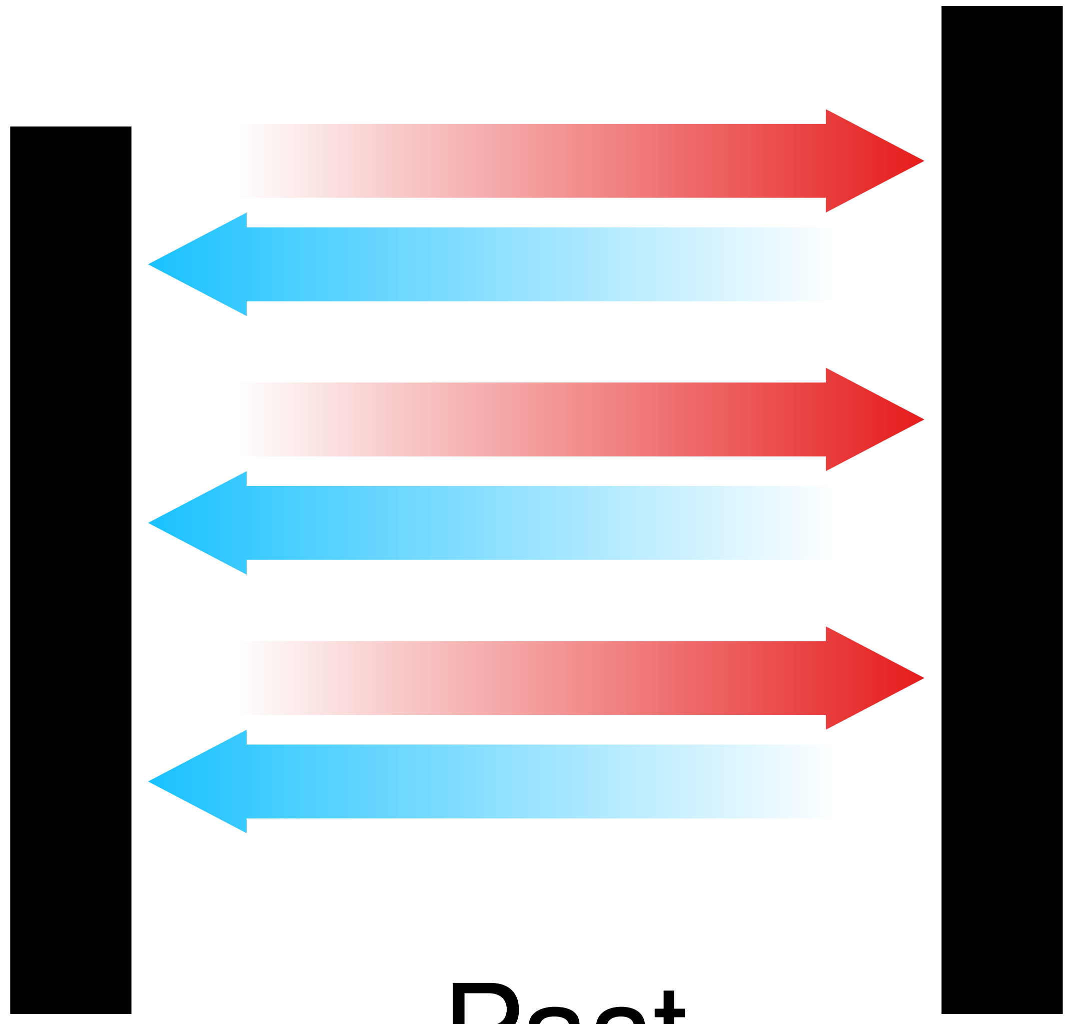
Gene flow

Neanderthal

'Modern' human

-30,000 years

Present



Past

©2019 Peter Beerli

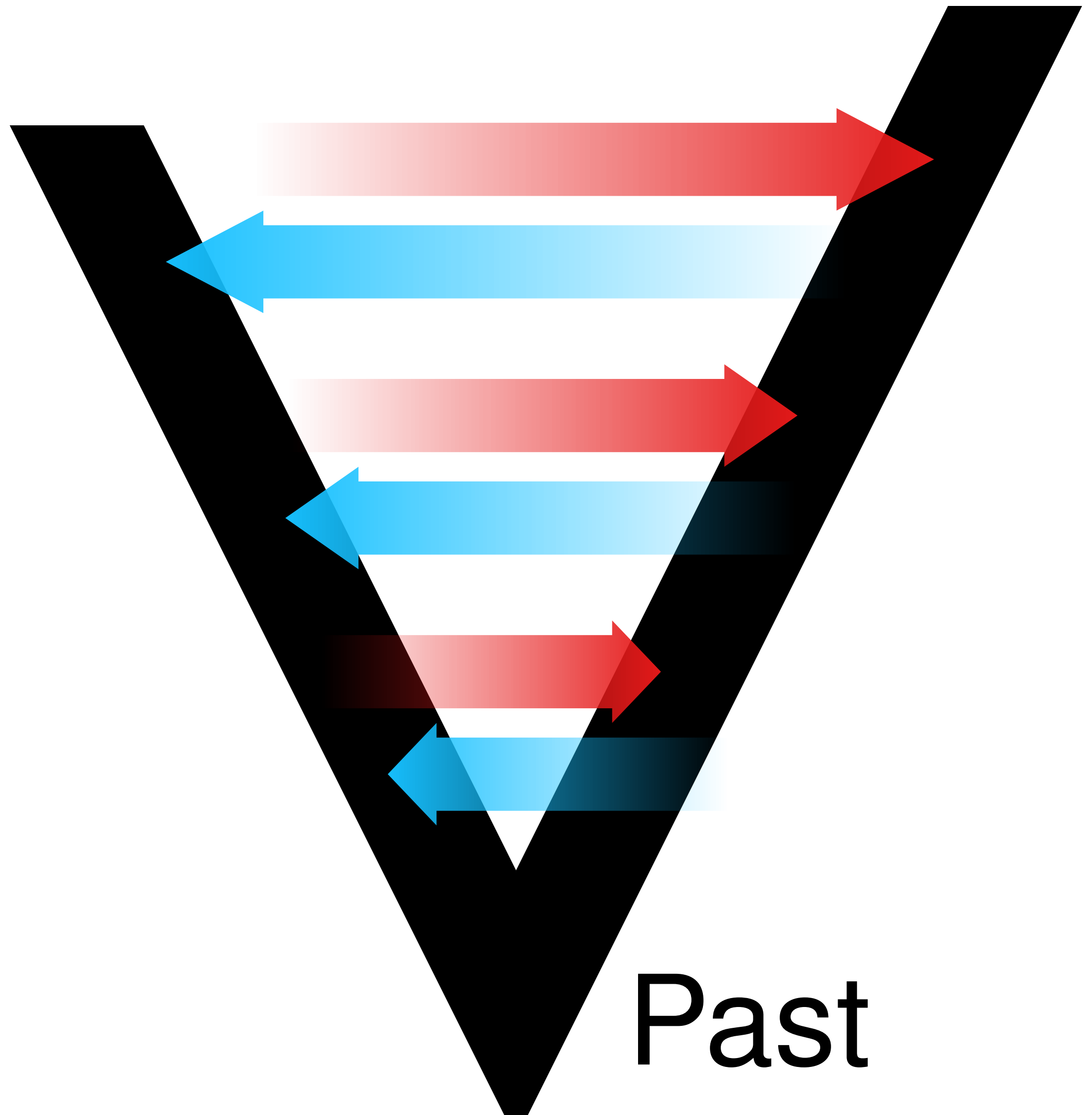
Data: Lego pictures from the net

Divergence and Gene flow

Neanderthal

'Modern' human

-30,000 years



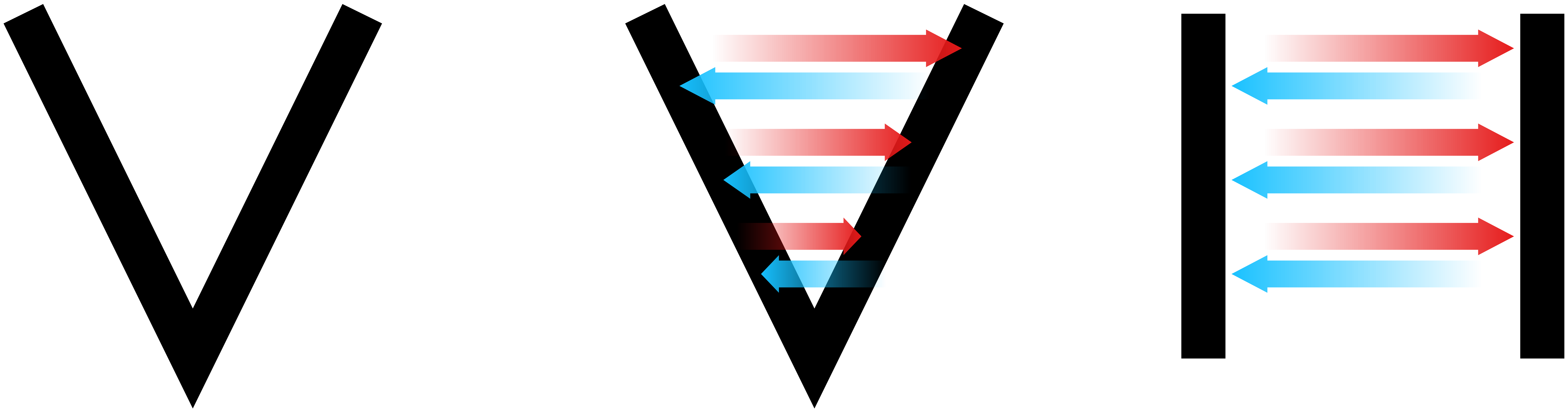
Present

Past

©2019 Peter Beerli

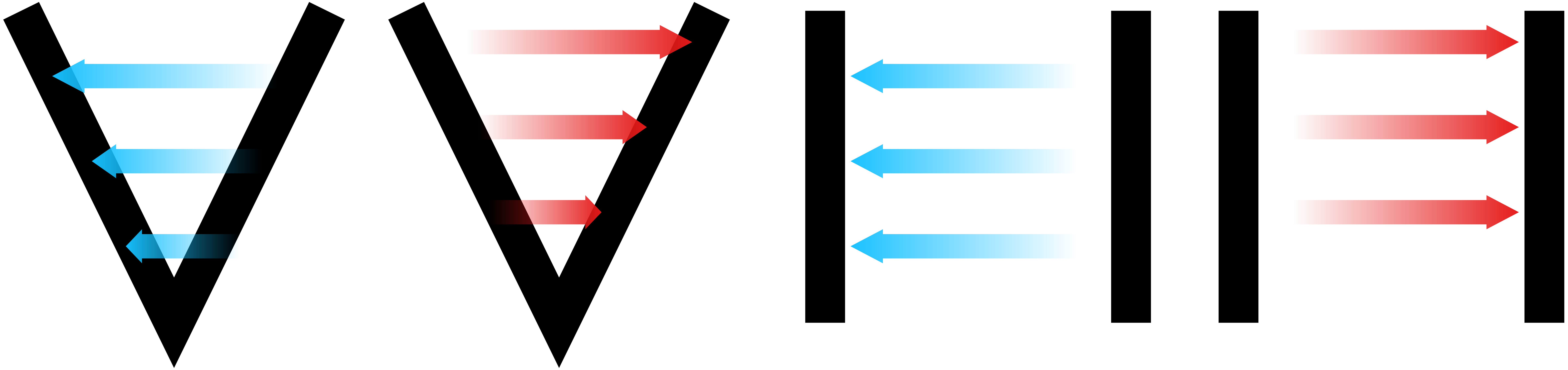
Data: Lego pictures from the net

Different structural models



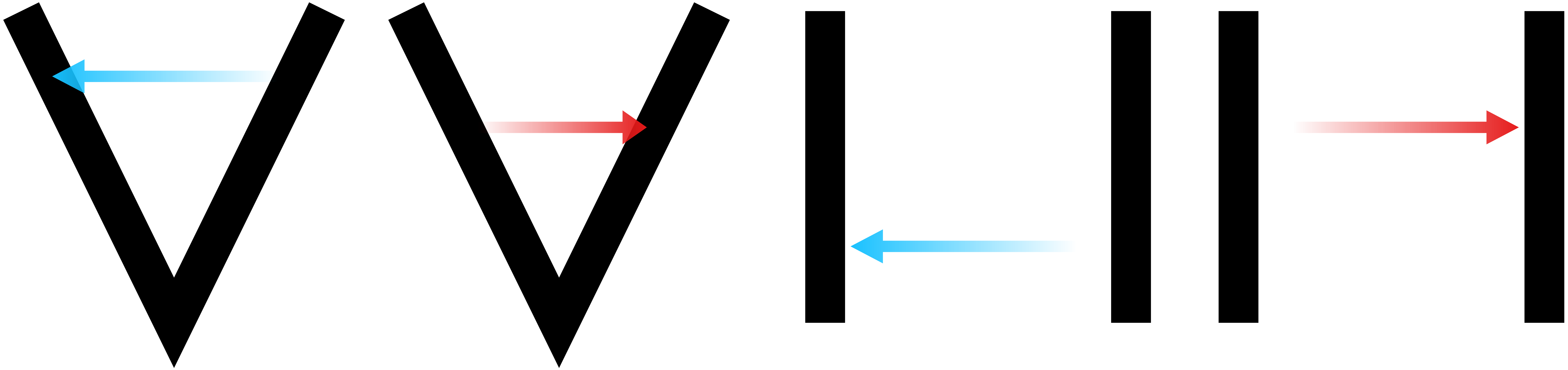
©2019 Peter Beerli

More different structural models



©2019 Peter Beerli

Even more different structural models



Model comparison

With a criterium such as likelihood we can compare nested models. Commonly we use a likelihood ratio test (LRT) or Akaike's information criterion (AIC) to establish whether phylogenetic trees are statistically different or mutation models have an effect on the outcome, etc.

Kass and Raftery (1995) popularized the Bayes Factor as a Bayesian alternative to the LRT.

Bayesian inference

©2019 Peter Beerli

Bayes factor

Theoretically, we can calculate the posterior probability density of the model

$$\mathrm{p}(\mathrm{M}_1|\mathrm{X}) = \frac{\mathrm{p}(\mathrm{M}_1)\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X})}$$

©2019 Peter Beerli

Bayes factor

Theoretically, we can calculate the posterior probability density of the model 1 and model 2

$$\mathrm{p}(\mathrm{M}_1|\mathrm{X}) = \frac{\mathrm{p}(\mathrm{M}_1)\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X})}$$

$$\mathrm{p}(\mathrm{M}_2|\mathrm{X}) = \frac{\mathrm{p}(\mathrm{M}_2)\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X})}$$

Bayes factor

Theoretically, we can calculate the posterior probability density of the model 1 and model 2

$$\frac{\mathrm{p}(\mathrm{M}_1|\mathrm{X})}{\mathrm{p}(\mathrm{M}_2|\mathrm{X})} = \frac{\frac{\mathrm{p}(\mathrm{M}_1)\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X})}}{\frac{\mathrm{p}(\mathrm{M}_2)\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X})}}$$

Bayes factor

We could look at the posterior odds ratio or equivalently the Bayes factors.

$$\frac{\mathrm{p}(\mathrm{M}_1|\mathrm{X})}{\mathrm{p}(\mathrm{M}_2|\mathrm{X})} = \frac{\mathrm{p}(\mathrm{M}_1)}{\mathrm{p}(\mathrm{M}_2)} \times \frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)}$$

$$\mathrm{BF} = \frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)} \qquad \mathrm{LBF} = 2\ln \mathrm{BF} = 2\ln\left(\frac{\mathrm{p}(\mathrm{X}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{X}|\mathrm{M}_2)}\right)$$

©2019 Peter Beerli

Bayes factor

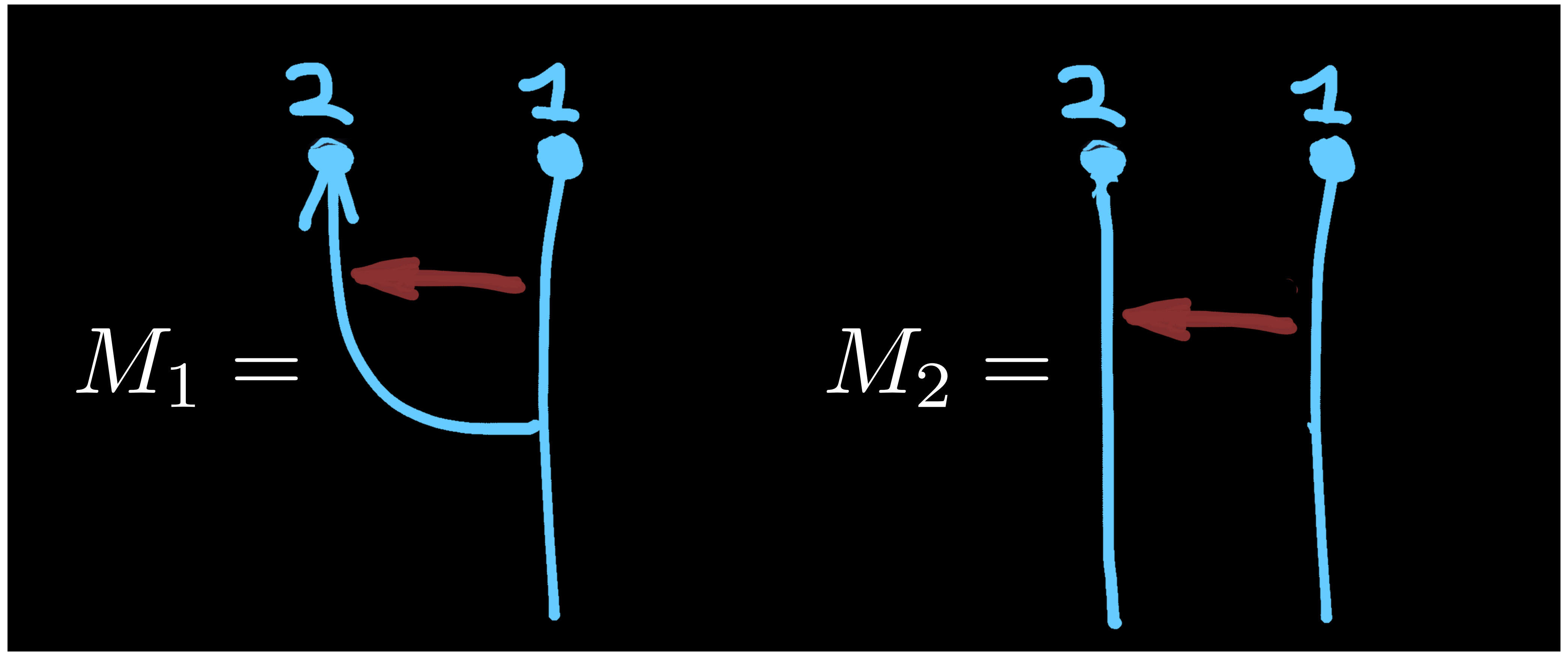
$$\mathrm{BF} = \frac{\mathrm{p(X|M_1)}}{\mathrm{p(X|M_2)}} \qquad \mathrm{LBF} = 2 \ln \mathrm{BF} = 2 \ln \left(\frac{\mathrm{p(X|M_1)}}{\mathrm{p(X|M_2)}} \right)$$

The magnitude of BF gives us evidence against or for hypothesis M_2

$$\mathrm{LBF} = 2 \ln \mathrm{BF} = z \quad \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

©2019 Peter Beerli

Bayes factor example



$$\mathrm{LBF} = 2\ln\mathrm{BF} = 2\ln\left(\frac{\mathrm{p(X|M_1)}}{\mathrm{p(X|M_2)}}\right) = 2(-9638.69) - (-9641.01) = 4.64$$

The magnitude of BF gives us evidence against or for hypothesis M_2

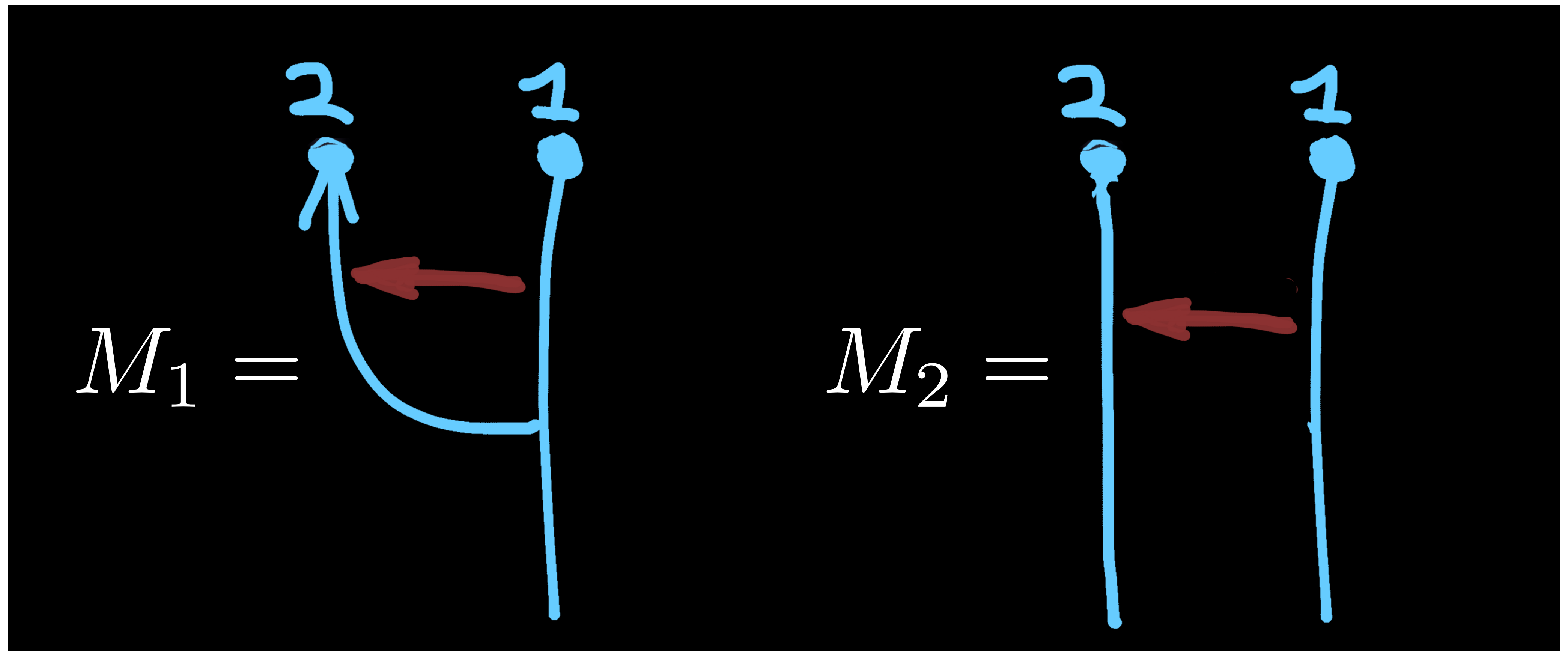
$$\mathrm{LBF} = 2\ln\mathrm{BF} = z \quad \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

©2019 Peter Beerli

Posterior model probability

(example continued)

Instead of calculating the Bayes factor we could use the probability of all tested models M_i and use them as weights (cf. Burnham and Anderson,1998)



$$p_i^* = \frac{\mathrm{p(X|M_i)}}{\sum_j \mathrm{p(X|M_j)}}, \qquad \sum_i p_i^* = 1, \qquad \ell_1 = -9638.61, \quad \ell_2 = -9641.01$$

$$p_1^* = \frac{\exp(\ell_1)}{\exp(\ell_1) + \exp(\ell_2)} = 0.911$$

$$p_2^* = \frac{\exp(\ell_2)}{\exp(\ell_1) + \exp(\ell_2)} = 0.089$$

©2019 Peter Beerli

Marginal likelihood

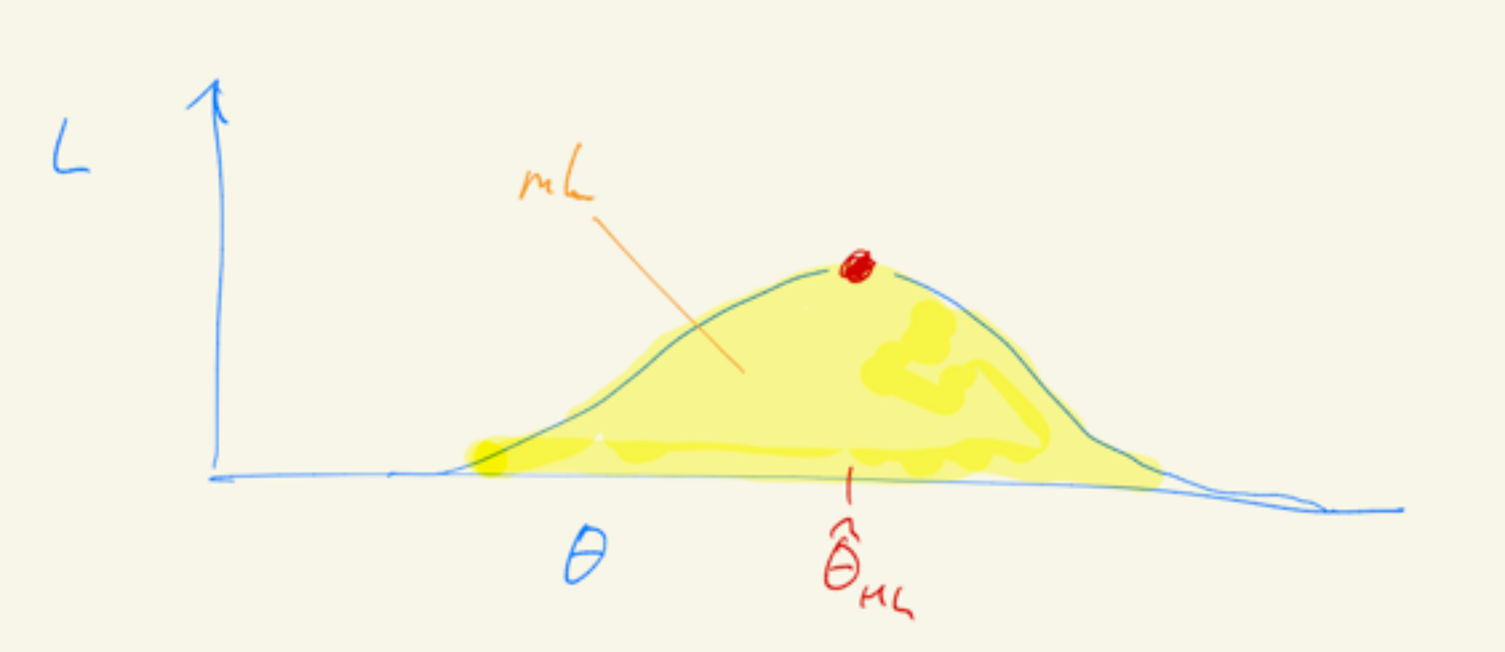
Typically, it is rather difficult to calculate the marginal likelihoods with good accuracy, because most often we only approximate the posterior distribution using Markov chain Monte Carlo (MCMC).
In MCMC we need to know only differences and therefore we typically do not need to calculate the denominator to calculate the Posterior distribution $\mathrm{p}(\Theta|\mathrm{X})$:

$$\mathrm{p}(\Theta|\mathrm{X},\mathrm{M}) = \frac{\mathrm{p}(\Theta)\mathrm{p}(\mathrm{X}|\Theta)}{\mathrm{p}(\mathrm{X}|\mathrm{M})} = \frac{\mathrm{p}(\Theta)\mathrm{p}(\mathrm{X}|\Theta)}{\int_\Theta \mathrm{p}(\Theta)\mathrm{p}(\mathrm{X}|\Theta)\mathrm{d}\Theta}$$

where $\mathrm{p}(\mathrm{X}|\mathrm{M})$ is the marginal likelihood, which we need for our model selection!

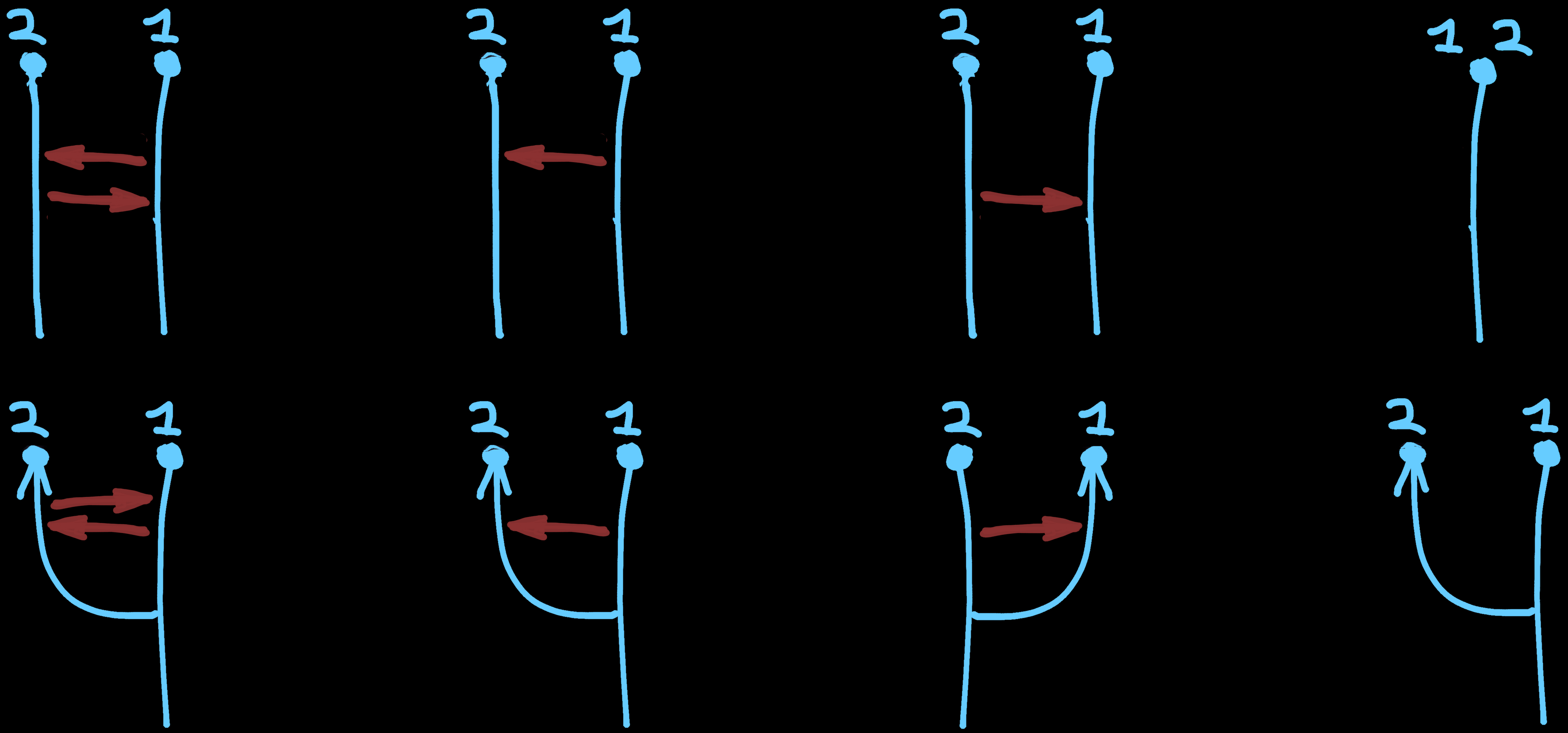
Estimation of the marginal likelihood

- Harmonic mean estimator [Kass and Raftery 1995]: methods is easy and used in many programs, results are biased and overestimate the marginal likelihood, variance of estimates can be very large.
- Thermodynamic integration (Path sampling) [Gelman and Meng 1997, Lartillot et al. 2006]: method is tedious to compute because several MCMC chains are needed. Results are accurate and reproducible with small variance when MCMC runs were run long enough.
- Stepping stone approach (Xie et al. 2011)



©2019 Peter Beerli

Population models

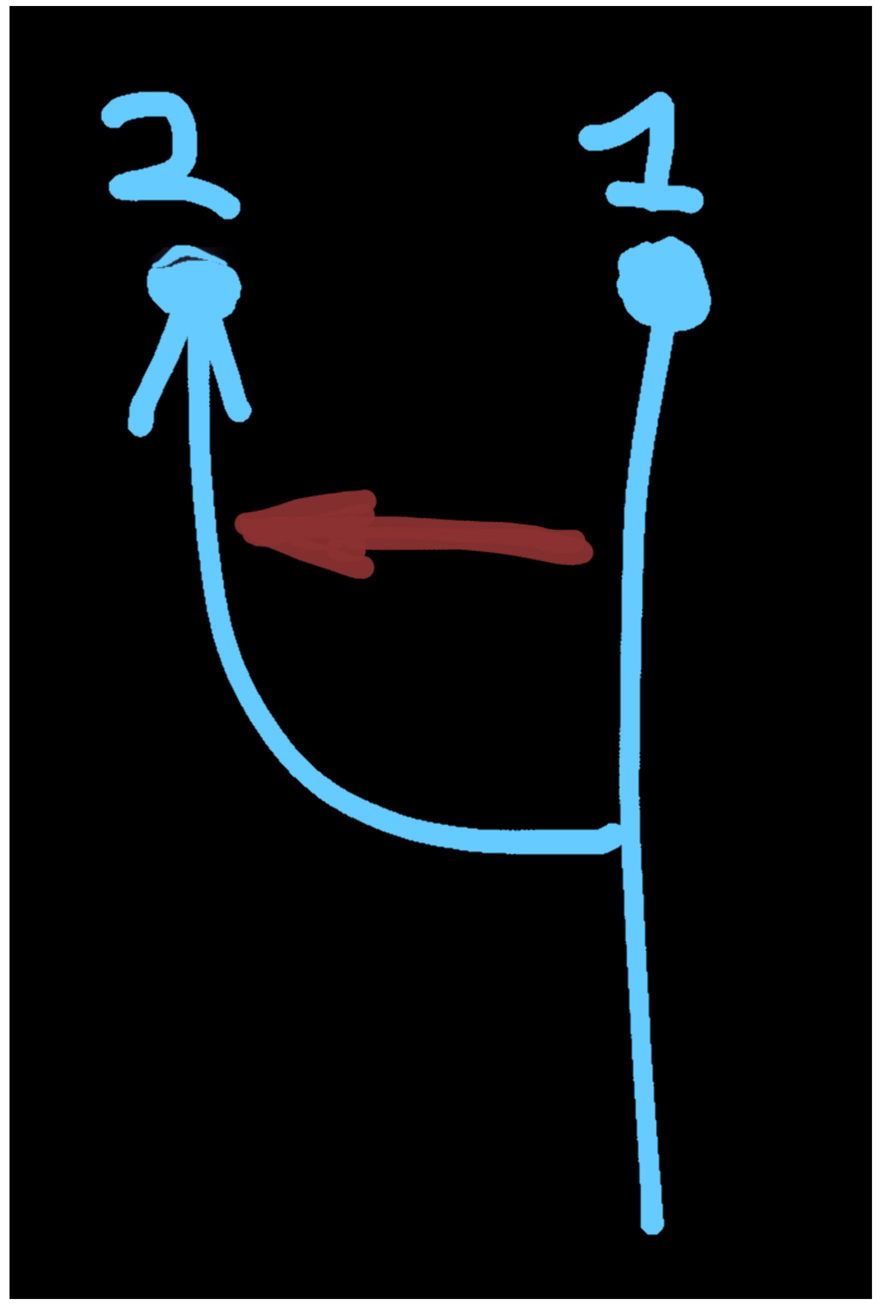


©2019 Peter Beerli

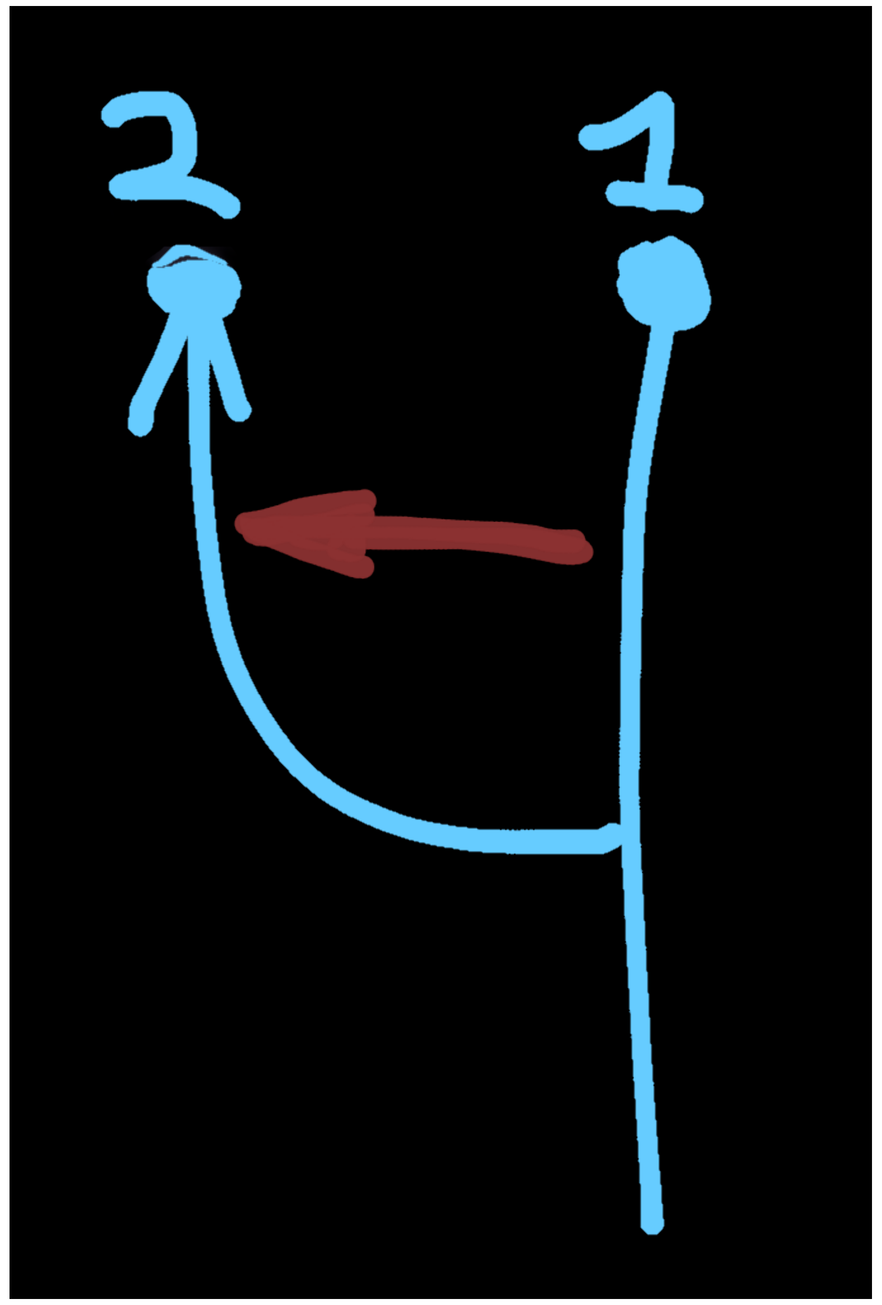
Simulated data

Two loci simulated from model x0Dx:

```
Model              Log(mL)    LBF*   Model-probability
-------------------------------------------------------
1: xxxx:           -9662.42   -23.73        0.0000
2: xDxx:           -9661.98   -23.29        0.0000
3: xxDx:           -9661.52   -22.83        0.0000
4: xd0x:           -9656.51   -17.82        0.0000
5: xD0x:           -9649.33   -10.64        0.0000
6: xx0x:           -9648.93   -10.24        0.0000
7: x0dx:           -9641.77    -3.08        0.0402
8: x0xx:           -9641.01    -2.32        0.0859
9: x0Dx:           -9638.69     0.00        0.8739
```



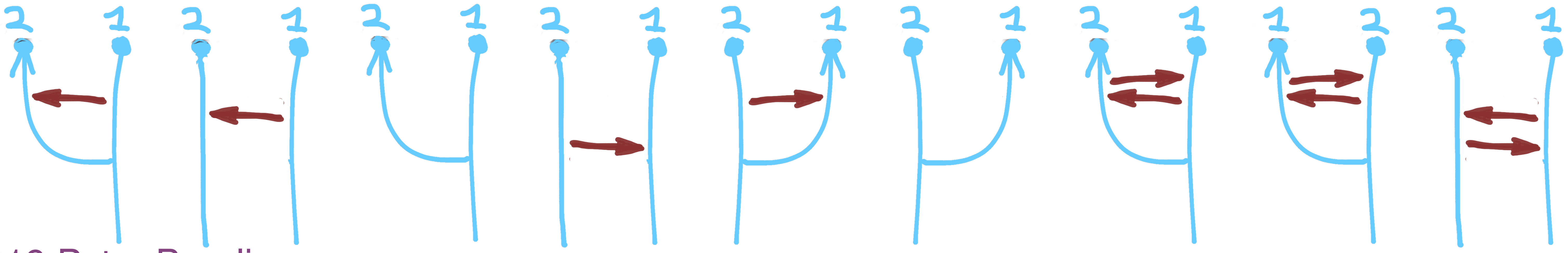
Simulated data



Two loci simulated from model `x0Dx`:

```
Model                  Log(mL)      LBF*    Model-probability
-------------------------------------------------------------
1: xxxx:              -9662.42    -23.73          0.0000
2: xDxx:              -9661.98    -23.29          0.0000
3: xxDx:              -9661.52    -22.83          0.0000
4: xd0x:              -9656.51    -17.82          0.0000
5: xD0x:              -9649.33    -10.64          0.0000
6: xx0x:              -9648.93    -10.24          0.0000
7: x0dx:              -9641.77     -3.08          0.0402
8: x0xx:              -9641.01     -2.32          0.0859
9: x0Dx:              -9638.69      0.00          0.8739
```

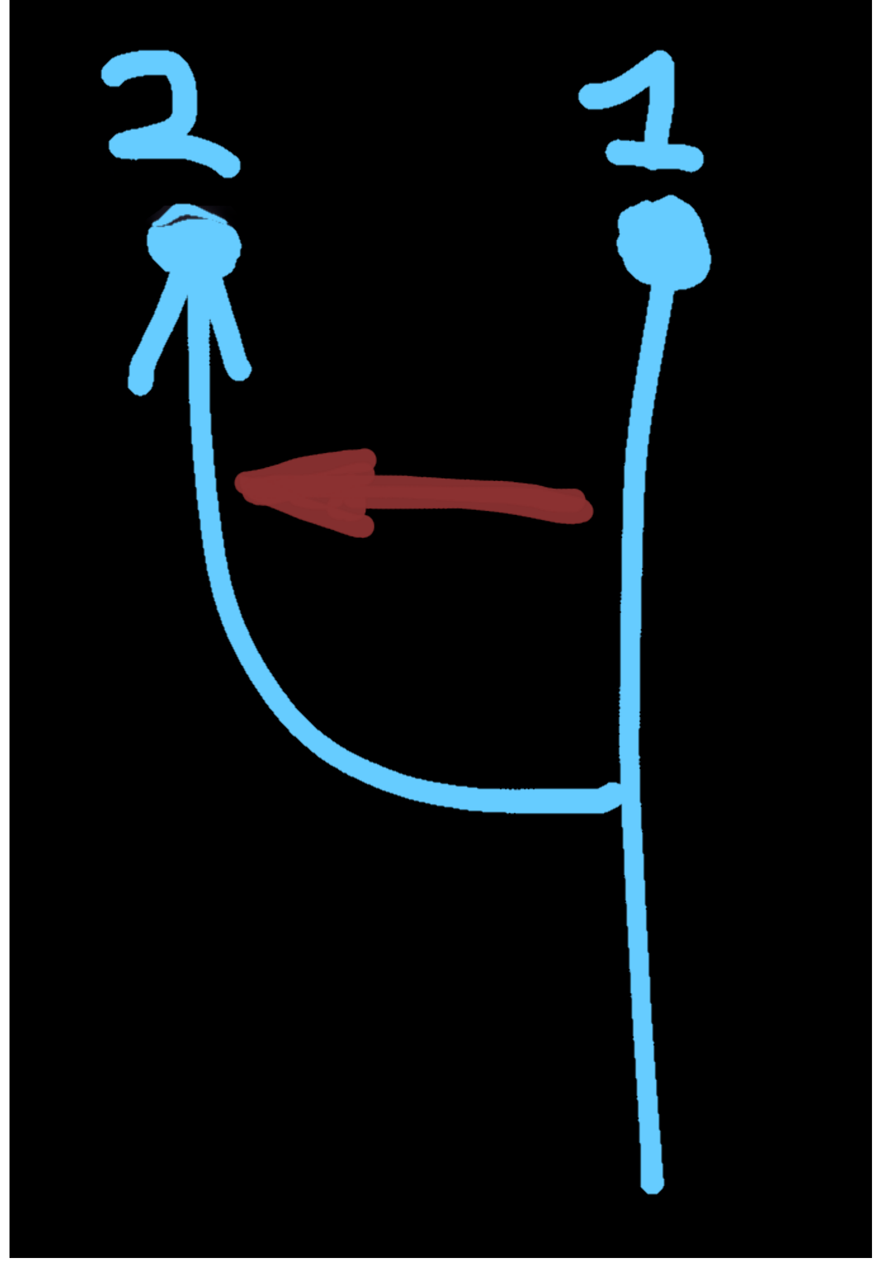
Best



Worst

©2019 Peter Beerli

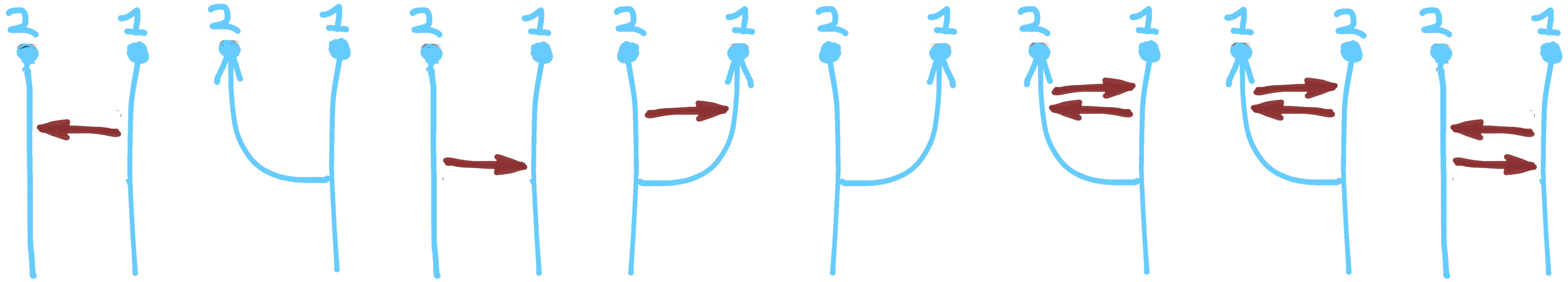
We did not include the correct model!



Two loci simulated from model x0Dx:

```
Model                  Log(mL)     LBF*   Model-probability
-------------------------------------------------------
1:xxxx:               -9662.42   -21.41        0.0000
2:xBxx:               -9661.98   -20.97        0.0000
3:xxBx:               -9661.52   -20.51        0.0000
4:xd0x:               -9656.51   -15.50        0.0000
5:xB0x:               -9649.33    -8.32        0.0002
6:xx0x:               -9648.93    -7.92        0.0002
7:x0dx:               -9641.77    -0.76        0.3185
8:x0xx:               -9641.01     0.00        0.6811
```

Best



Worst

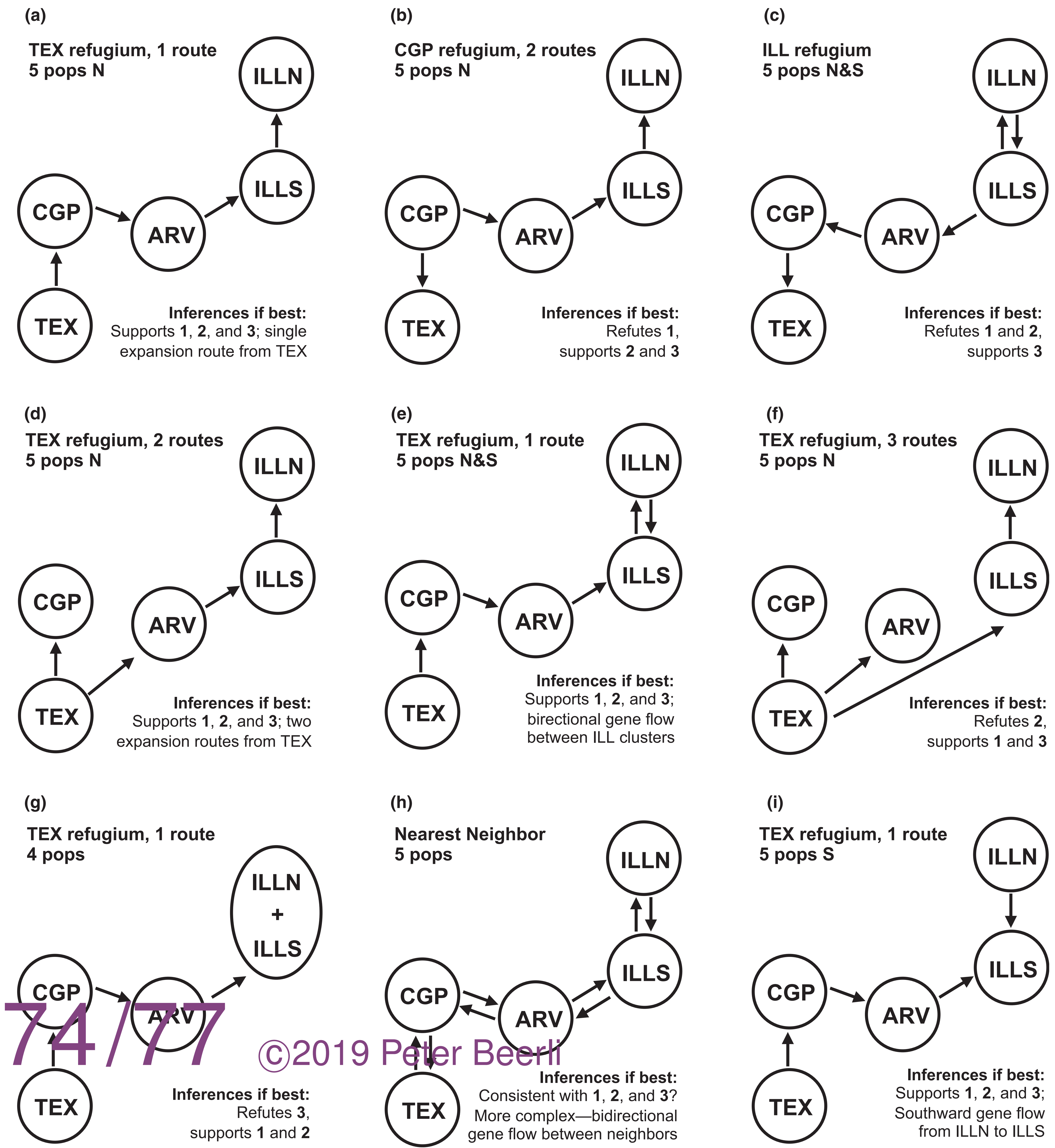
©2019 Peter Beerli

A real example

Frog picture: http://mdc.mo.gov/discover-nature/field-guide

Lisa N. Barrow, A. T. Bigelow, C. A. Phillips, and E. Moriarty Lemmon (2015) Phylogeographic inference using Bayesian model comparison across a fragmented chorus frog species complex. Molecular Ecology

3 There is detectable genetic structure within P. illinoensis consistent with the disjunct range.



(a) TEX refugium, 1 route
5 pops N

Inferences if best:
Supports 1, 2, and 3; single expansion route from TEX

(b) CGP refugium, 2 routes
5 pops N

Inferences if best:
Refutes 1,
supports 2 and 3

(c) ILL refugium
5 pops N&S

Inferences if best:
Refutes 1 and 2,
supports 3

(d) TEX refugium, 2 routes
5 pops N

Inferences if best:
Supports 1, 2, and 3; two expansion routes from TEX

(e) TEX refugium, 1 route
5 pops N&S

Inferences if best:
Supports 1, 2, and 3;
birectional gene flow between ILL clusters

(f) TEX refugium, 3 routes
5 pops N

Inferences if best:
Refutes 2,
supports 1 and 3

(g) TEX refugium, 1 route
4 pops

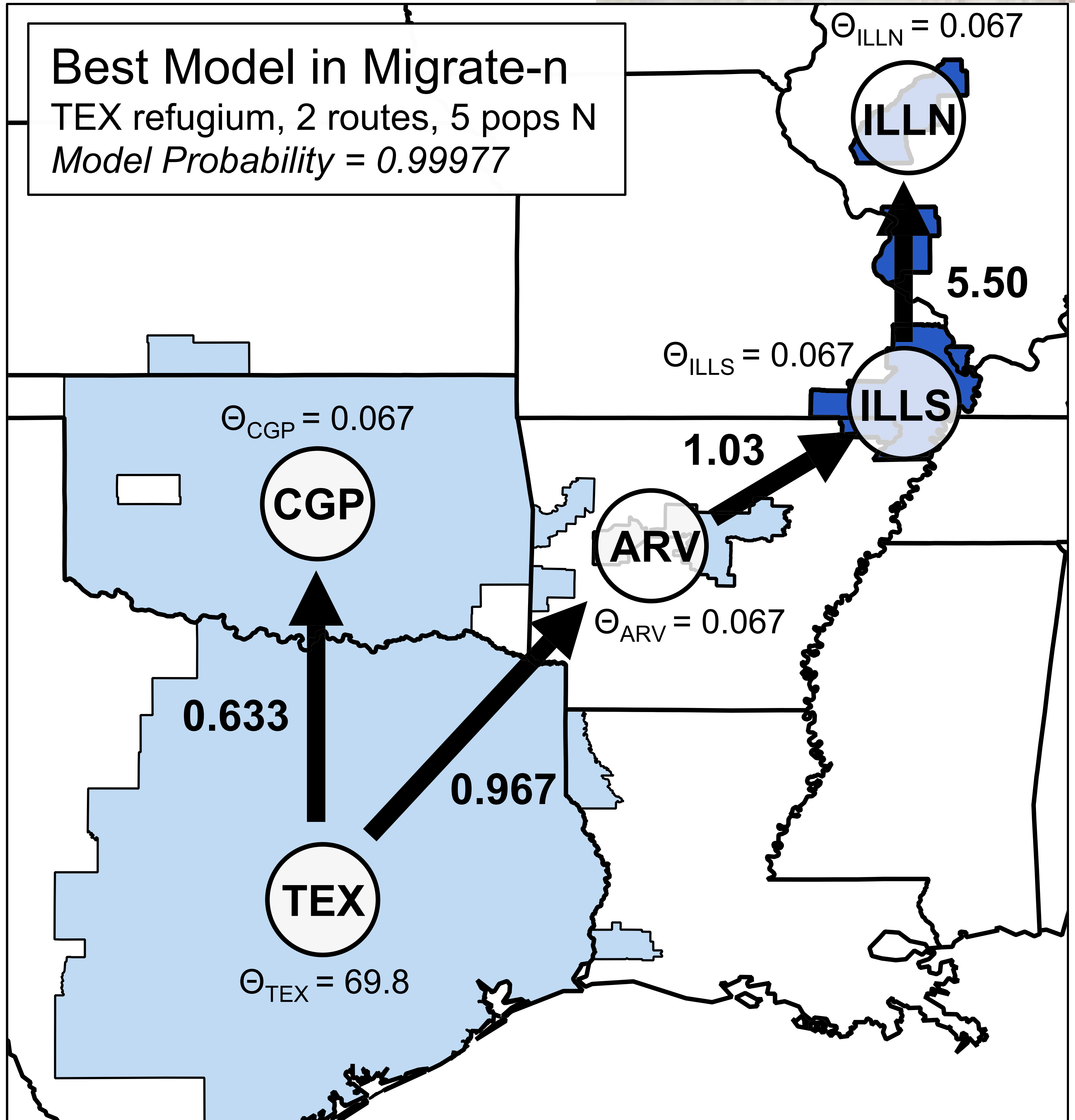
Inferences if best:
Refutes 3,
supports 1 and 2

(h) Nearest Neighbor
5 pops

Inferences if best:
Consistent with 1, 2, and 3?
More complex—bidirectional gene flow between neighbors

(i) TEX refugium, 1 route
5 pops S

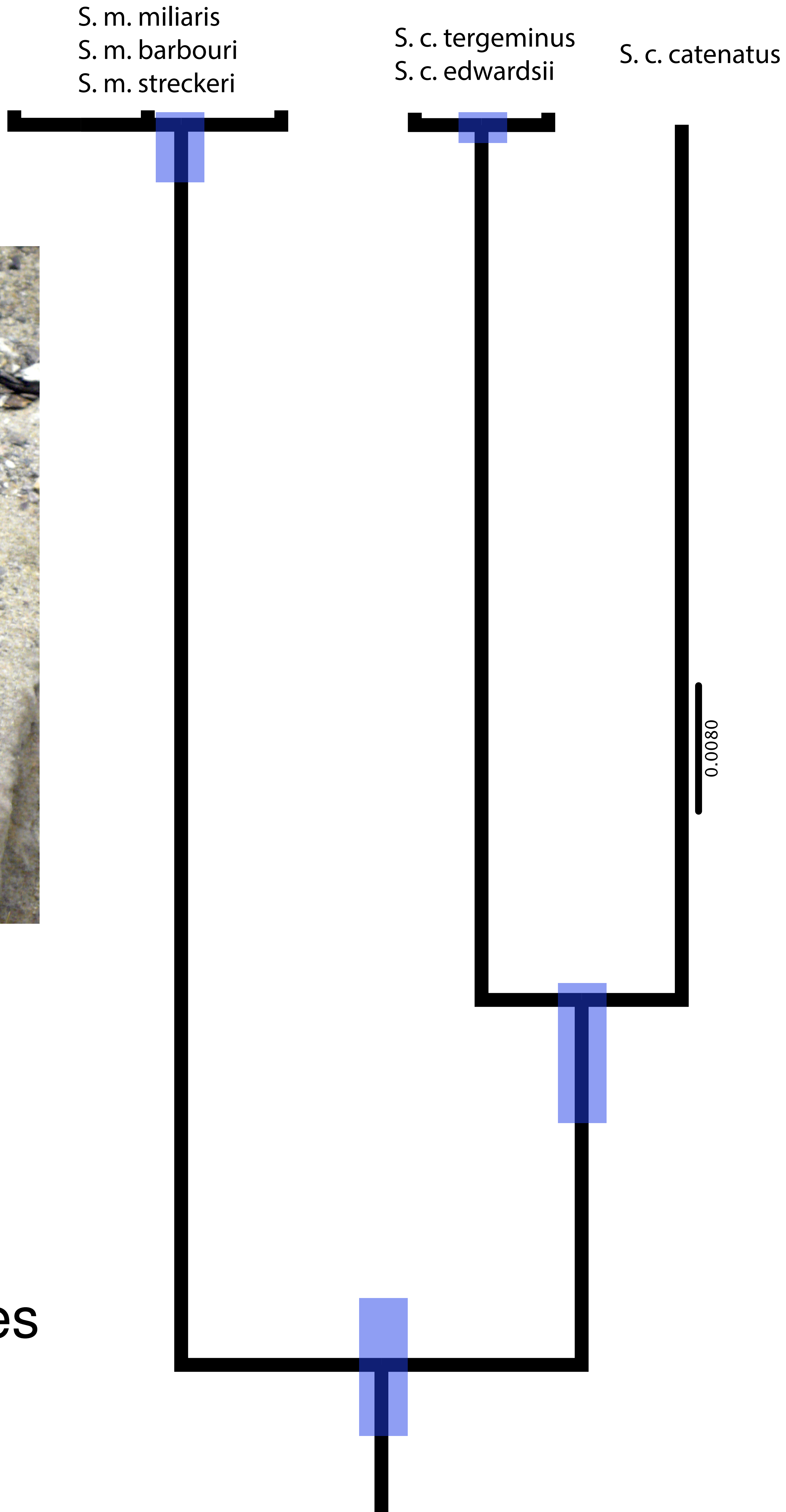
Inferences if best:
Supports 1, 2, and 3;
Southward gene flow from ILLN to ILLS



Best Model in Migrate-n
TEX refugium, 2 routes, 5 pops N
Model Probability = 0.99977

©2019 Peter Beerli

Population splitting



Model	Log(mL)	LBF	Model-probability
1: 3 species:	-15887.49	0.00	1.0000
2: 6 species:	-15961.95	-74.46	0.0000

Estimation of splitting dates of 6 subspecies of pygmy rattle snakes using MIGRATE (data from Kubatko et al. 2011)

Summary

- You may be surprised that your favored model does not win in a model comparison competition, but figuring out the model order leads oftentimes to new insights about the problem.
- Models by themselves are not true or wrong. BUT they may not fit your data well, OR they describe your data even when you “know” that the model is insufficient.

©2019 Peter Beerli

Thank you

Lucrezia Bieler

National Science Foundation

Michal Palzcewski,
Haleh Ashki,
Justin Bricker,
Somayeh Mashayekhi,
Kyle Shaw

http://popgen.sc.fsu.edu

©2019 Peter Beerli

Credit: ESO/C. Malin