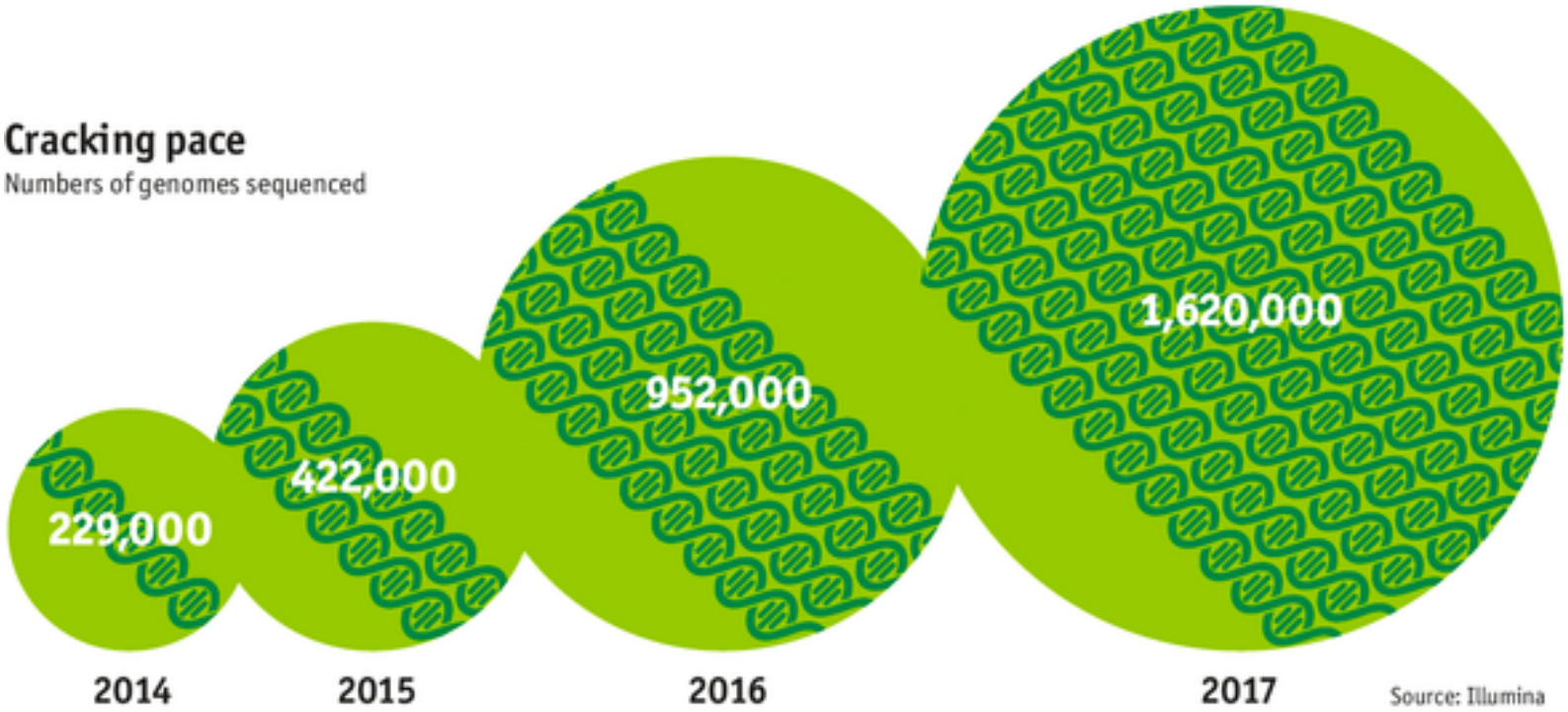
A Deep Learning Algorithm for Population Genetics

Sara Mathieson

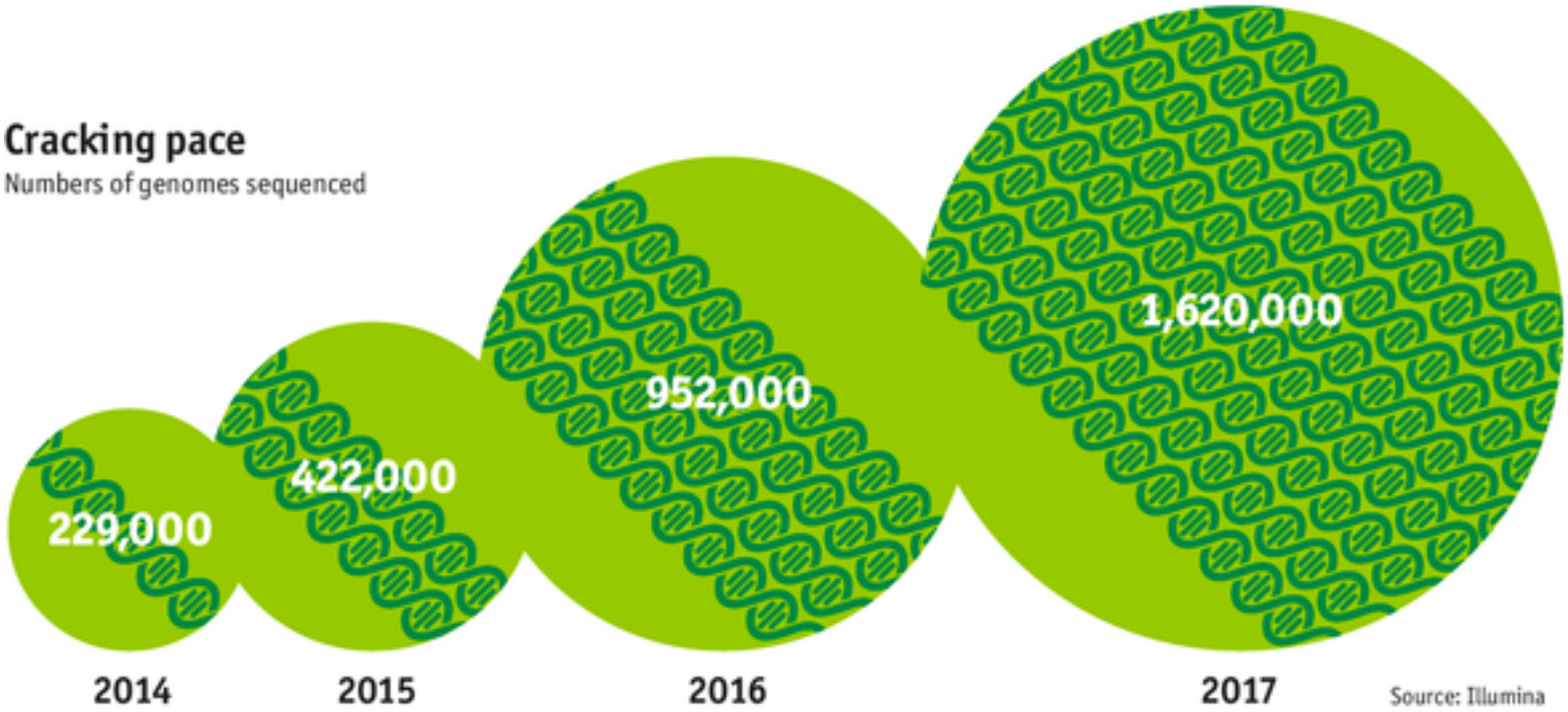
CSC 390
October 20, 2016

Number of human genomes sequenced



Economist, "The World in 2015"

Number of human genomes sequenced



Economist, "The World in 2015"

UK 10K

verinata Health®

10,000 species with genetic data

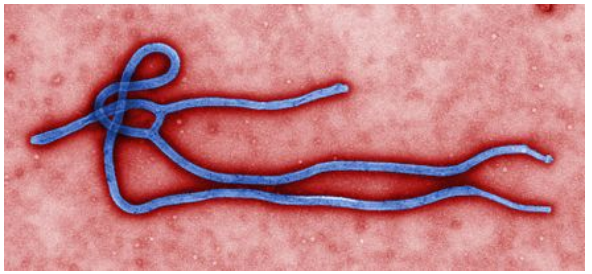
Chimp

Melitaea cinxia

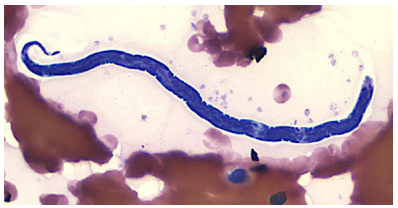
Buffalo

Images: wikipedia

Chinese liver worm

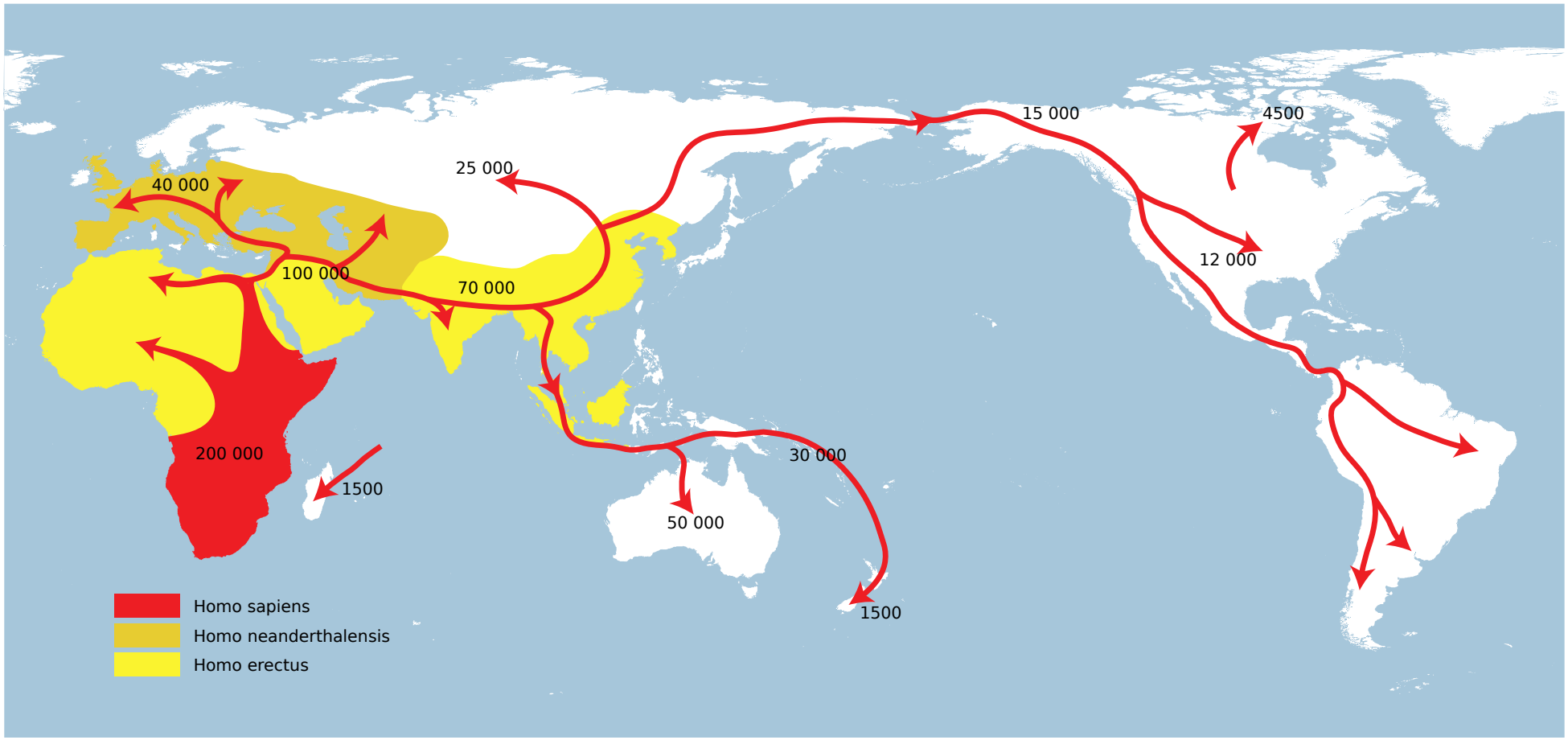


Ebola



Loa loa (eye worm)

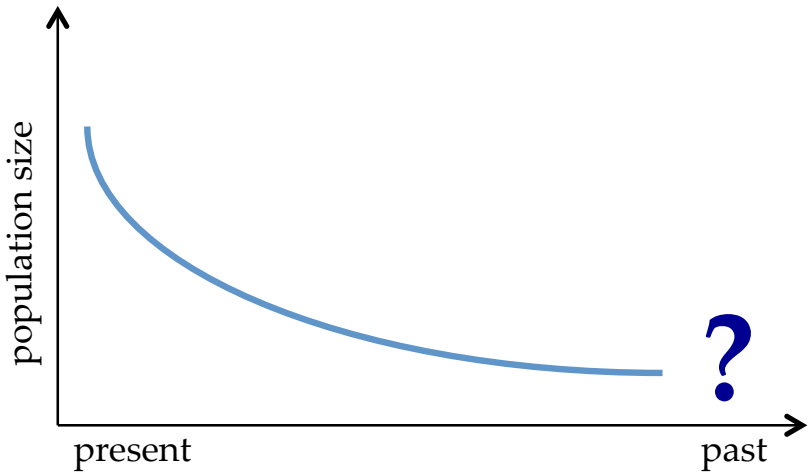
Modern DNA is a window into ancient history

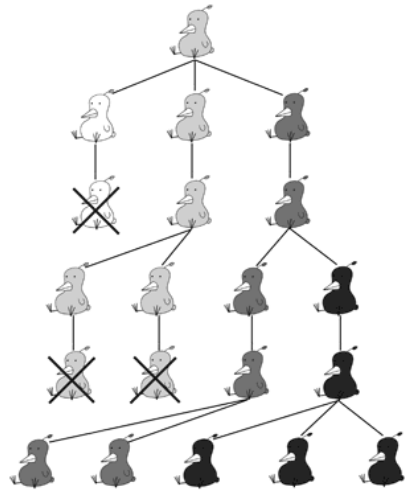


NordNordWest, wikipedia

Main question

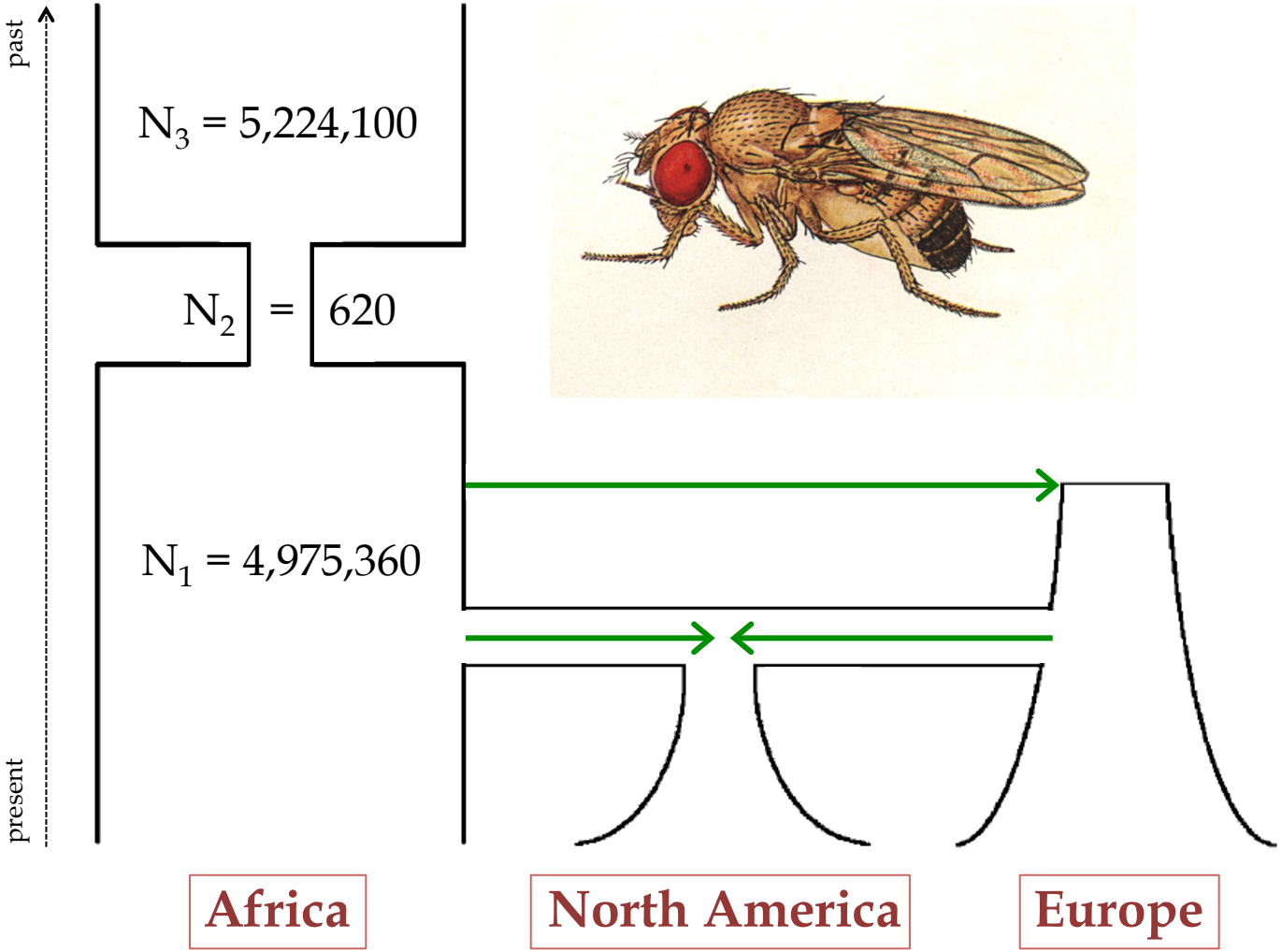
Population sizes + natural selection





evolutionexplained.blogspot.com

Motivation: demography and selection in *Drosophila*



"Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population."
Duchen, Živković, Hutter, Stephan, and Laurent. *Genetics* (2013).

Outline

- What is hard about joint inference? (*estimating multiple parameters at the same time*)

Outline

- What is hard about joint inference? (*estimating multiple parameters at the same time*)
- Why do we need a new method?

Outline

- What is hard about joint inference? (*estimating multiple parameters at the same time*)
- Why do we need a new method?
- Why deep learning?

Outline

- What is hard about joint inference? (*estimating multiple parameters at the same time*)
- Why do we need a new method?
- Why deep learning?

- A deep learning method for population genetics

Outline

- What is hard about joint inference? (*estimating multiple parameters at the same time*)
- Why do we need a new method?
- Why deep learning?

- A deep learning method for population genetics
- Applications, results, and future work

Motivation from Population Genetics

Sequence data as a matrix (X)

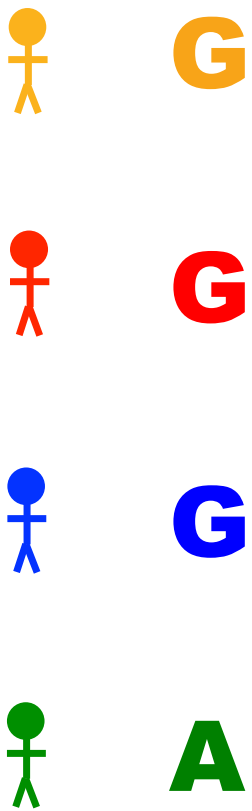

GCCTAGCTAGGTTACGTACG


GCCTAGCTAGGTTACGTACG

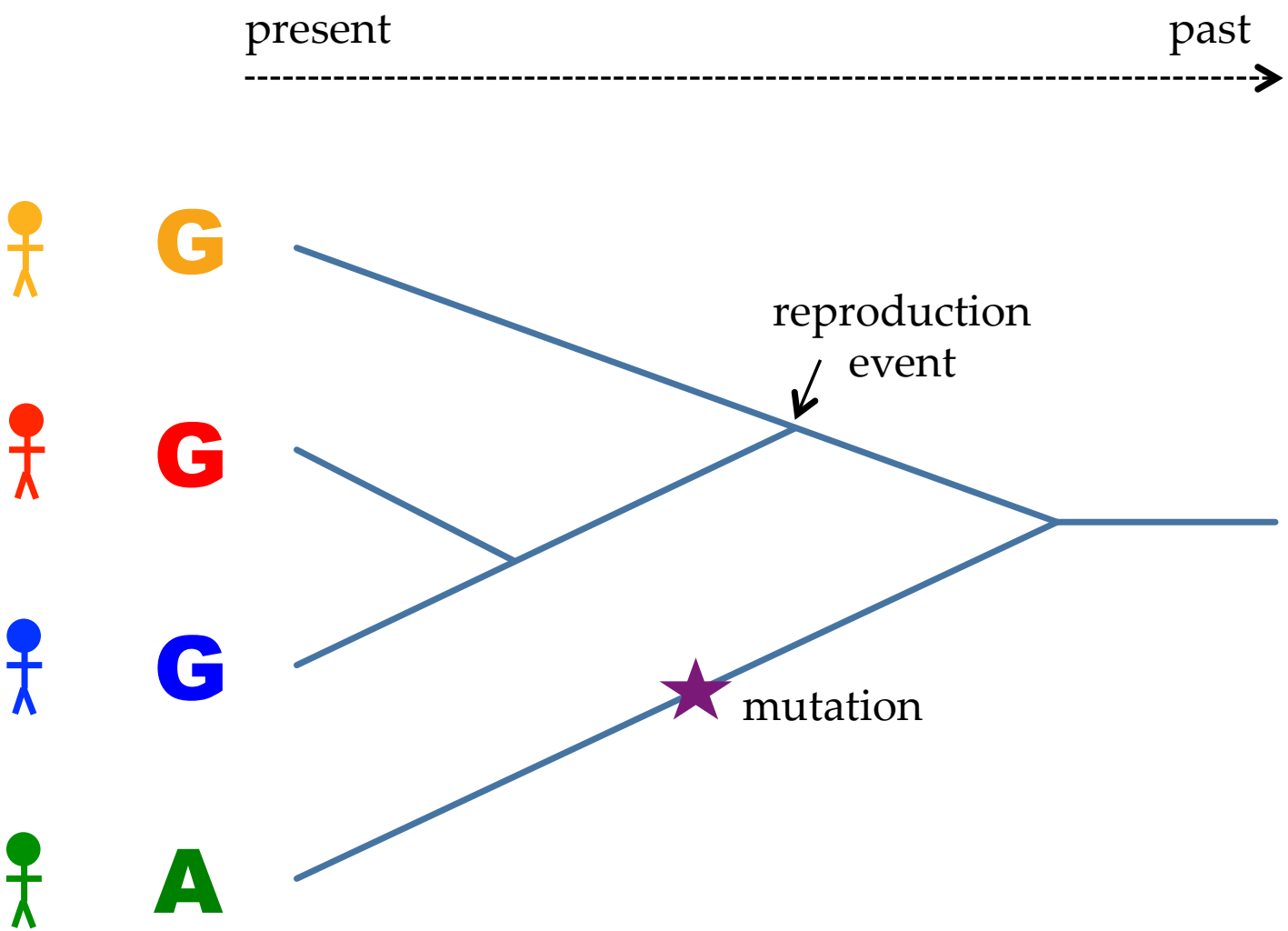

GCCTAGCTAGGTTACGTACG


ACCTAGCTAGGTTACGTACG

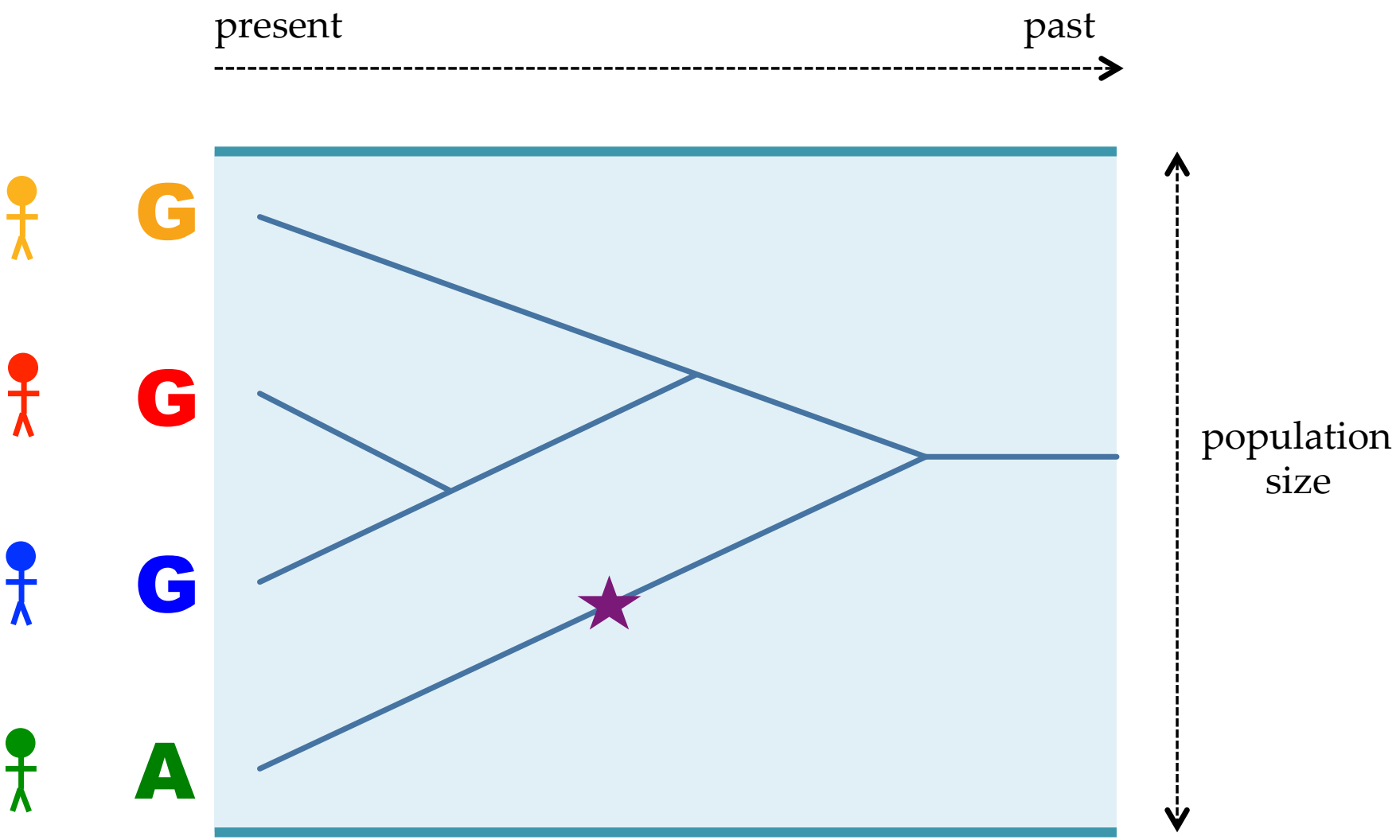
Sequence data at a single site



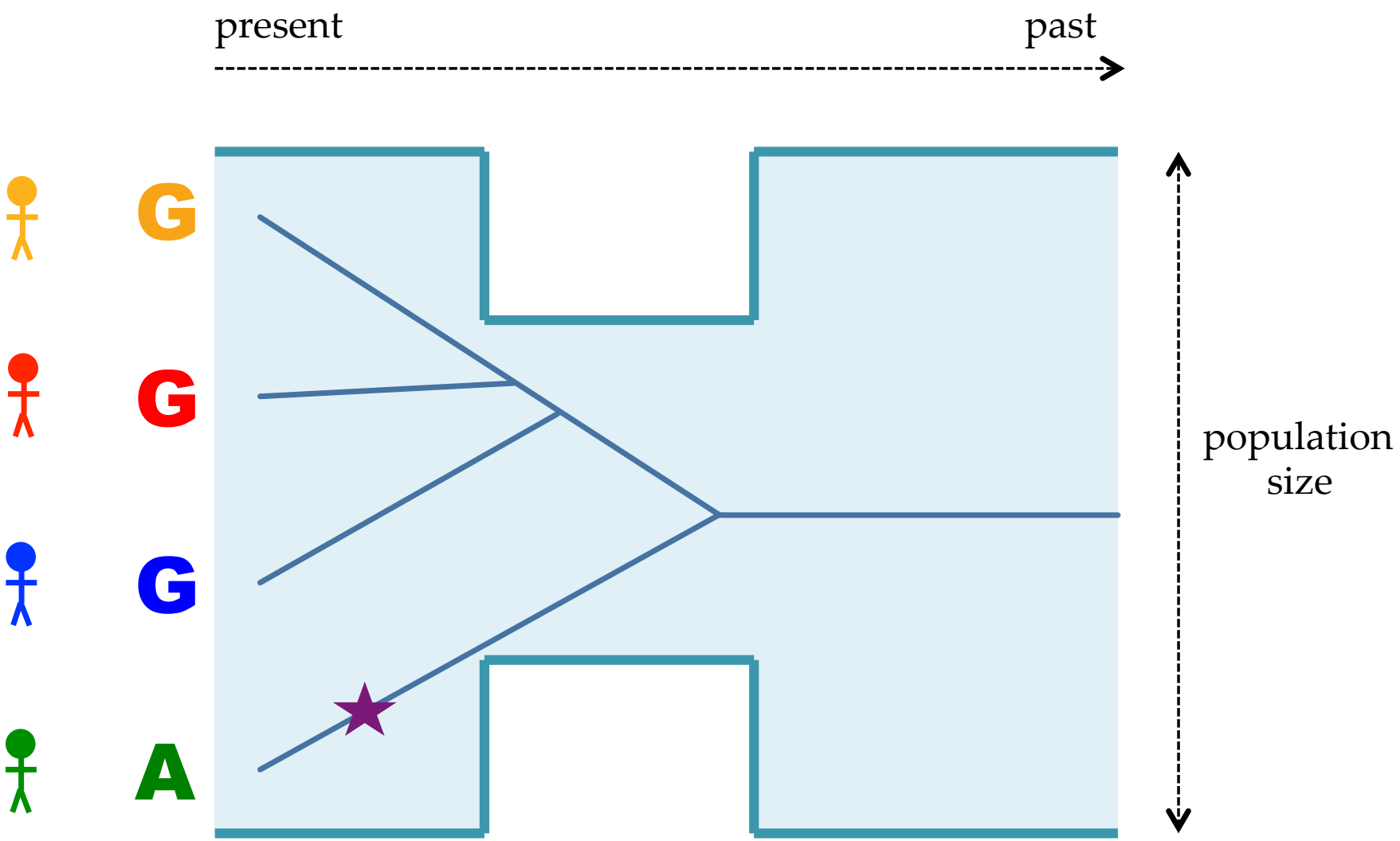
At a single site, everyone is related by a tree



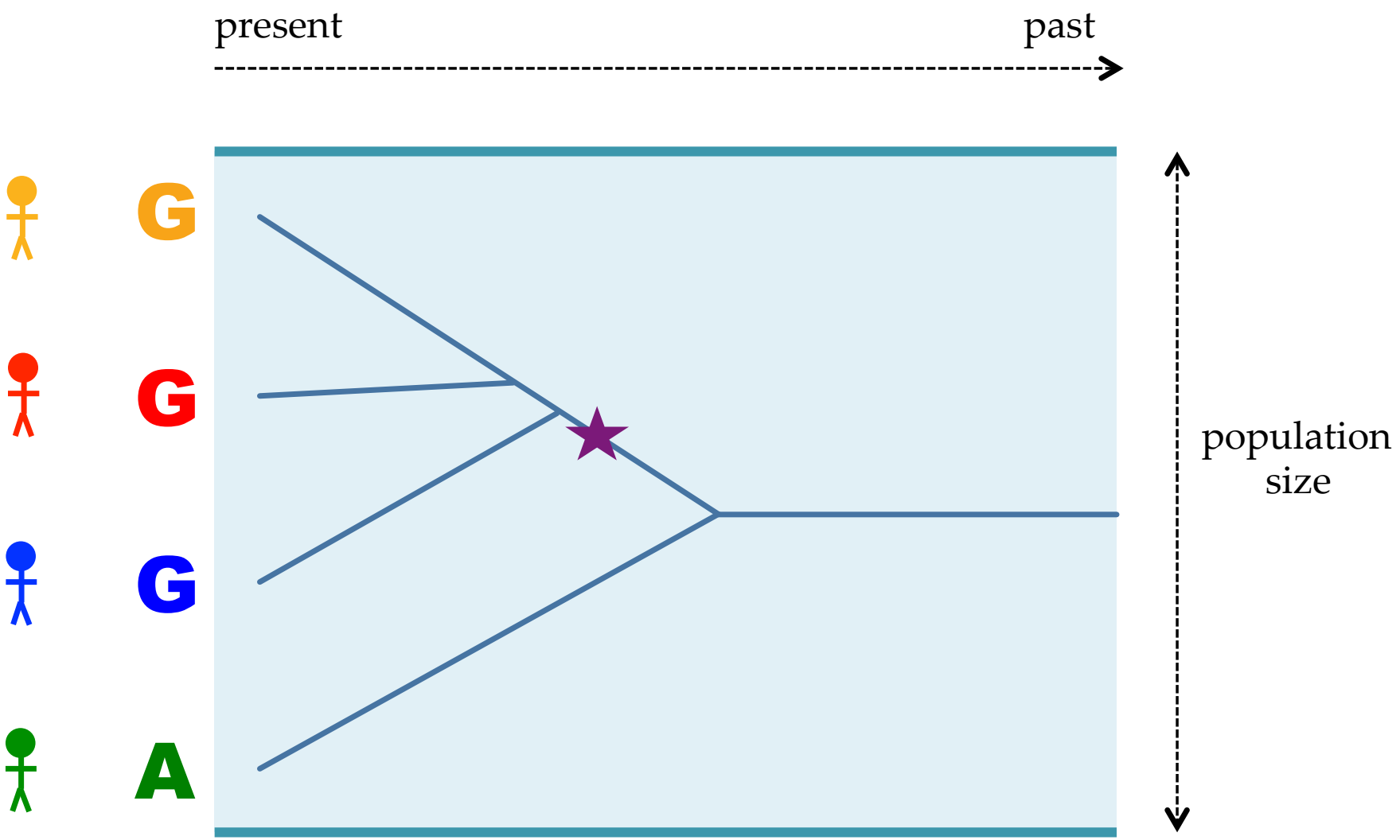
We call this tree the "genealogy"



Population bottleneck?



Or selection?



Likelihood-free inference

Big question: How can we infer population genetic parameters when we don't know the genealogy?

Likelihood-free inference

Big question: How can we infer population genetic parameters when we don't know the genealogy?

One answer: Approximate Bayesian Computation (ABC)

Likelihood-free inference

Big question: How can we infer population genetic parameters when we don't know the genealogy?

One answer: Approximate Bayesian Computation (ABC)

Likelihood-free inference

Big question: How can we infer population genetic parameters when we don't know the genealogy?

One answer: Approximate Bayesian Computation (ABC)

Likelihood-free inference

Big question: How can we infer population genetic parameters when we don't know the genealogy?

One answer: Approximate Bayesian Computation (ABC)

Main idea: try different combinations of ingredients until you get something close

Likelihood-free inference

Big question: How can we infer population genetic parameters when we don't know the genealogy?

One answer: Approximate Bayesian Computation (ABC)

Advantages of ABC:

- easy to use
- outputs a posterior

Likelihood-free inference

Big question: How can we infer population genetic parameters when we don't know the genealogy?

One answer: Approximate Bayesian Computation (ABC)

Advantages of ABC:

- easy to use
- outputs a posterior

Disadvantages of ABC:

- rejection method
- hard to interpret
- "curse of dimensionality"
- distance metric on statistics

Deep Learning Overview

Another answer: machine learning



X_1
X_2
X_3
X_4
X_5
X_6

input data

?

Y_1
Y_2
Y_3



parameters

Deep learning for images

X_1

X_2

X_3

X_4

X_5

X_6

input data

?

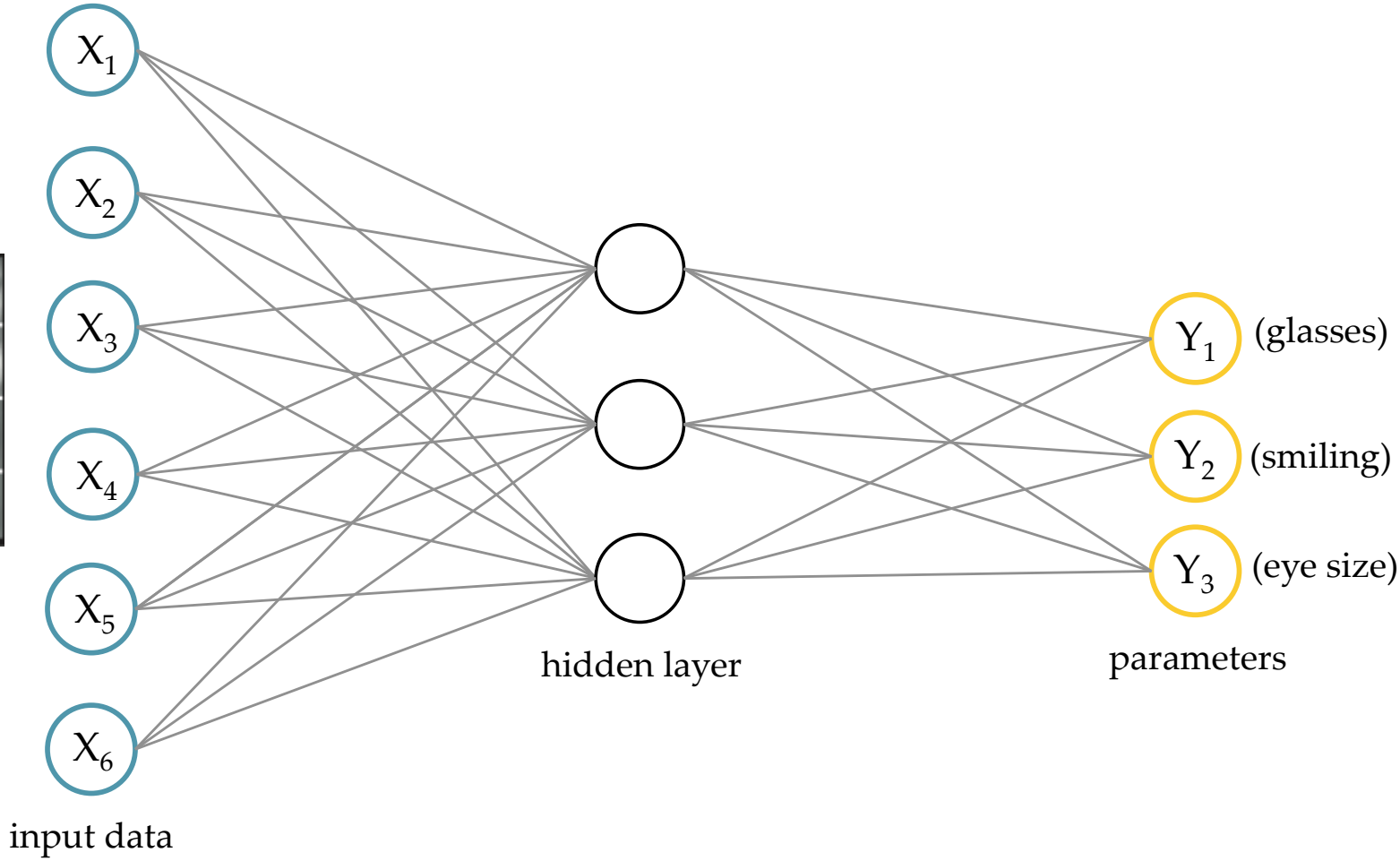
Y_1 (glasses)

Y_2 (smiling)

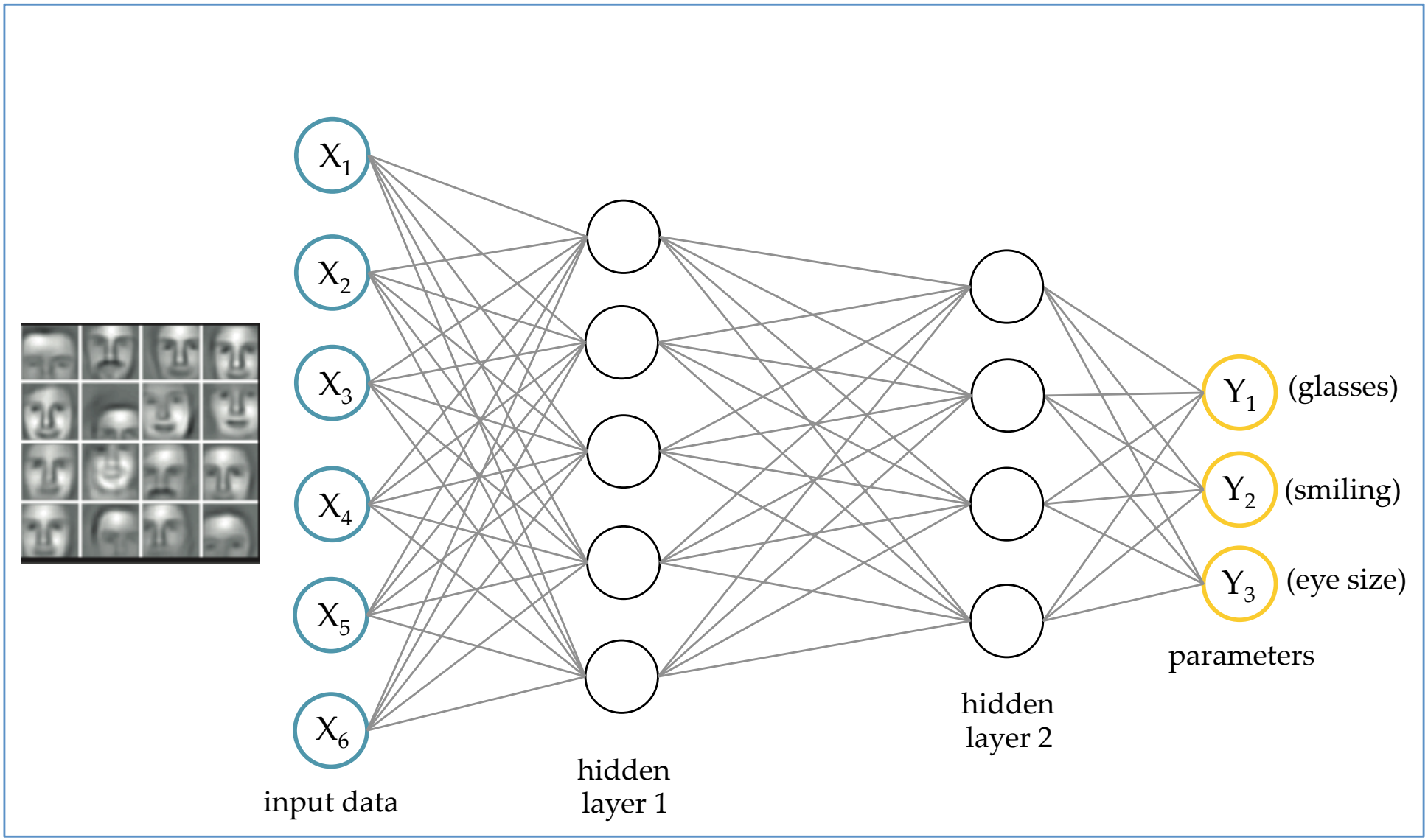
Y_3 (eye size)

parameters

Classical neural network

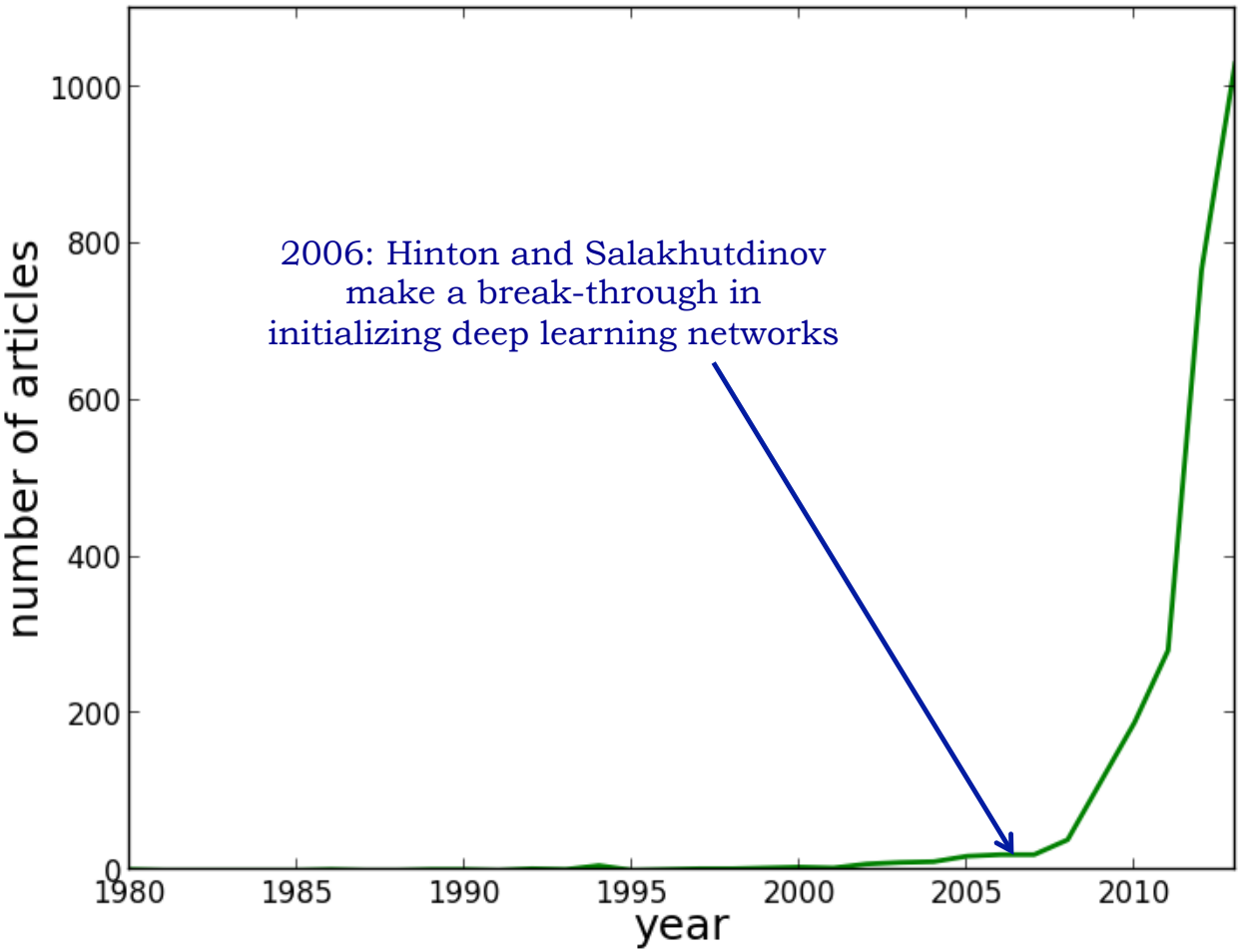


Deep network



Rise of deep learning

Number of papers that mention deep learning per year



Break-through: unsupervised learning, autoencoder

Goal: initialize the deep learning weights

input

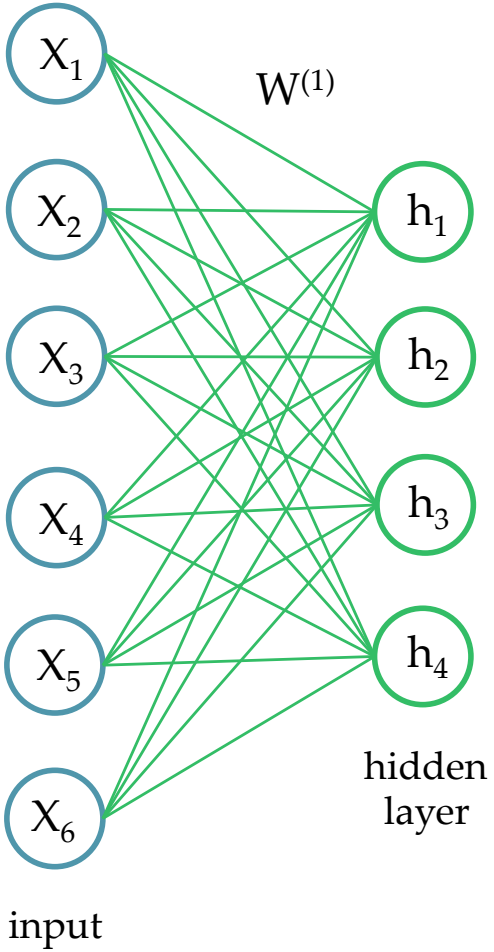
Break-through: unsupervised learning, autoencoder

Goal: initialize the deep learning weights

1. Project data into a lower dimension:

$$h_j = \sigma\left(W_j^{(1)} \cdot x\right)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



Break-through: unsupervised learning, autoencoder

Goal: initialize the deep learning weights

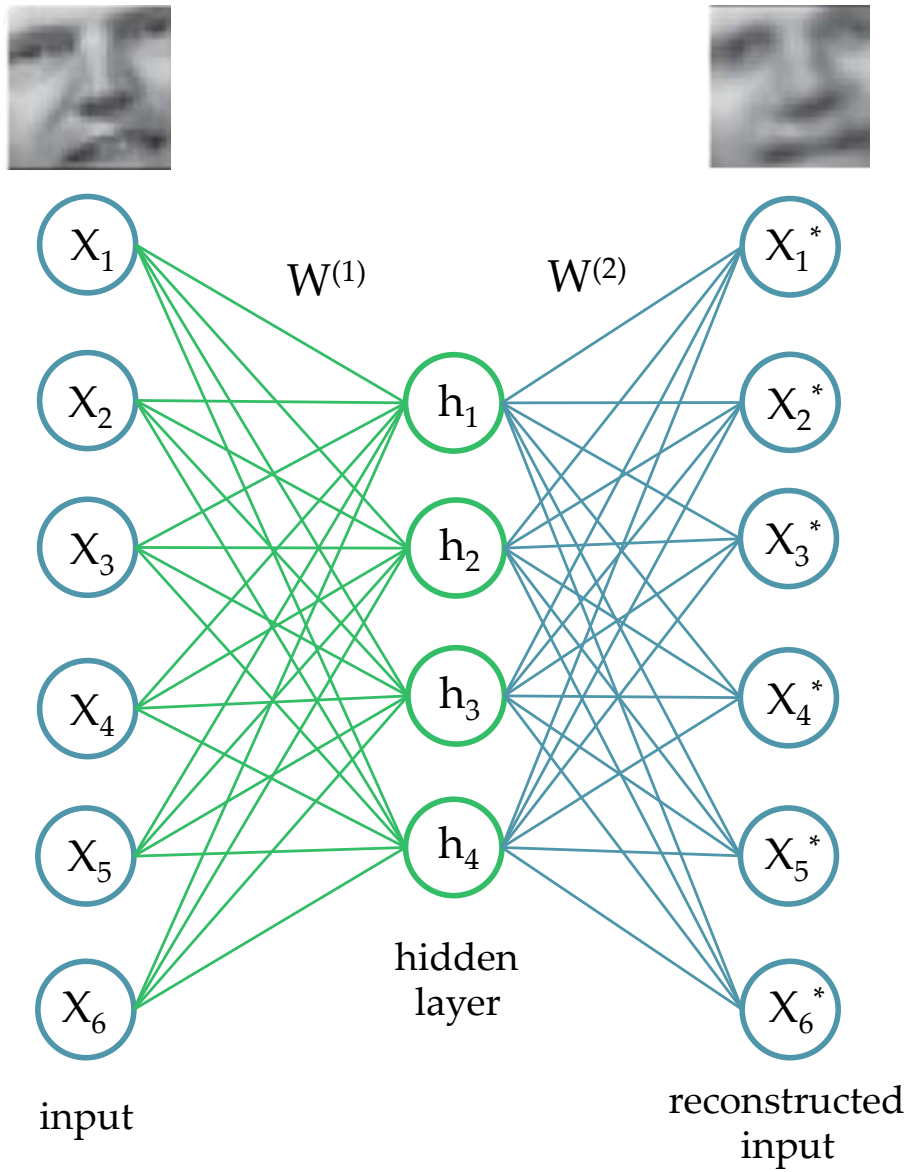
1. Project data into a lower dimension:

$$h_j = \sigma\left(W_j^{(1)} \cdot x\right)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

2. Minimize objective function:

$$J_x(W) = \frac{1}{2}\left|\left|x - x^*\right|\right|^2$$

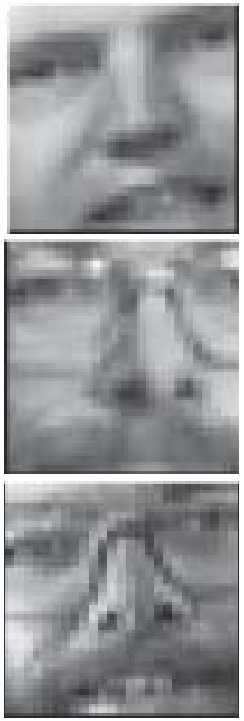


Break-through: unsupervised learning, autoencoder

original image

Break-through: unsupervised learning, autoencoder

original image



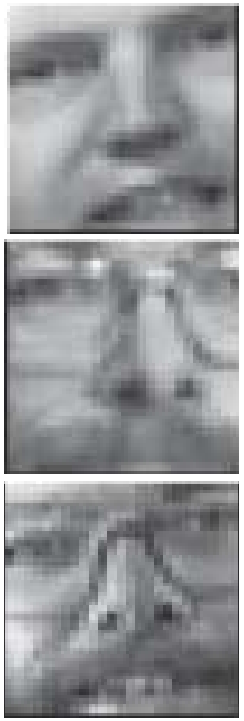
compression and feature reduction

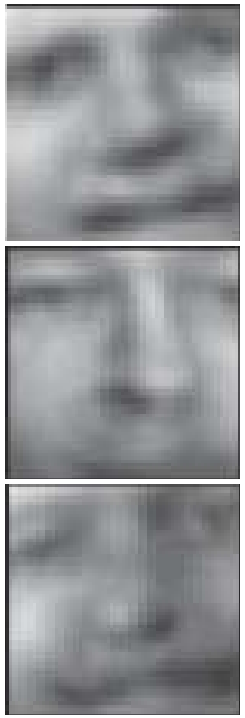
Break-through: unsupervised learning, autoencoder

original image

compression and feature reduction

reconstructed image





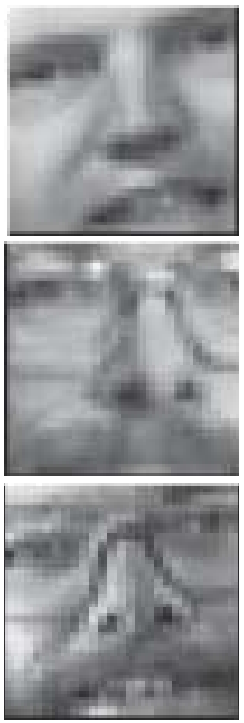
Break-through: unsupervised learning, autoencoder

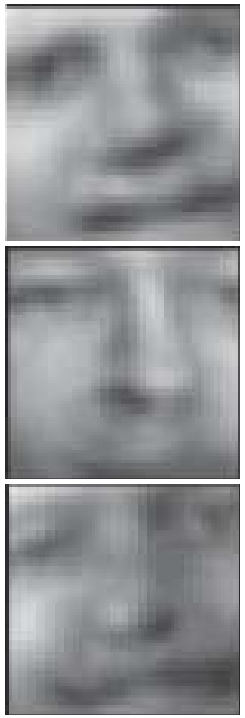
original image

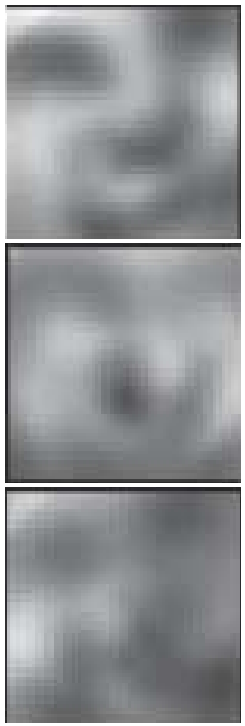
compression and feature reduction

reconstructed image

PCA

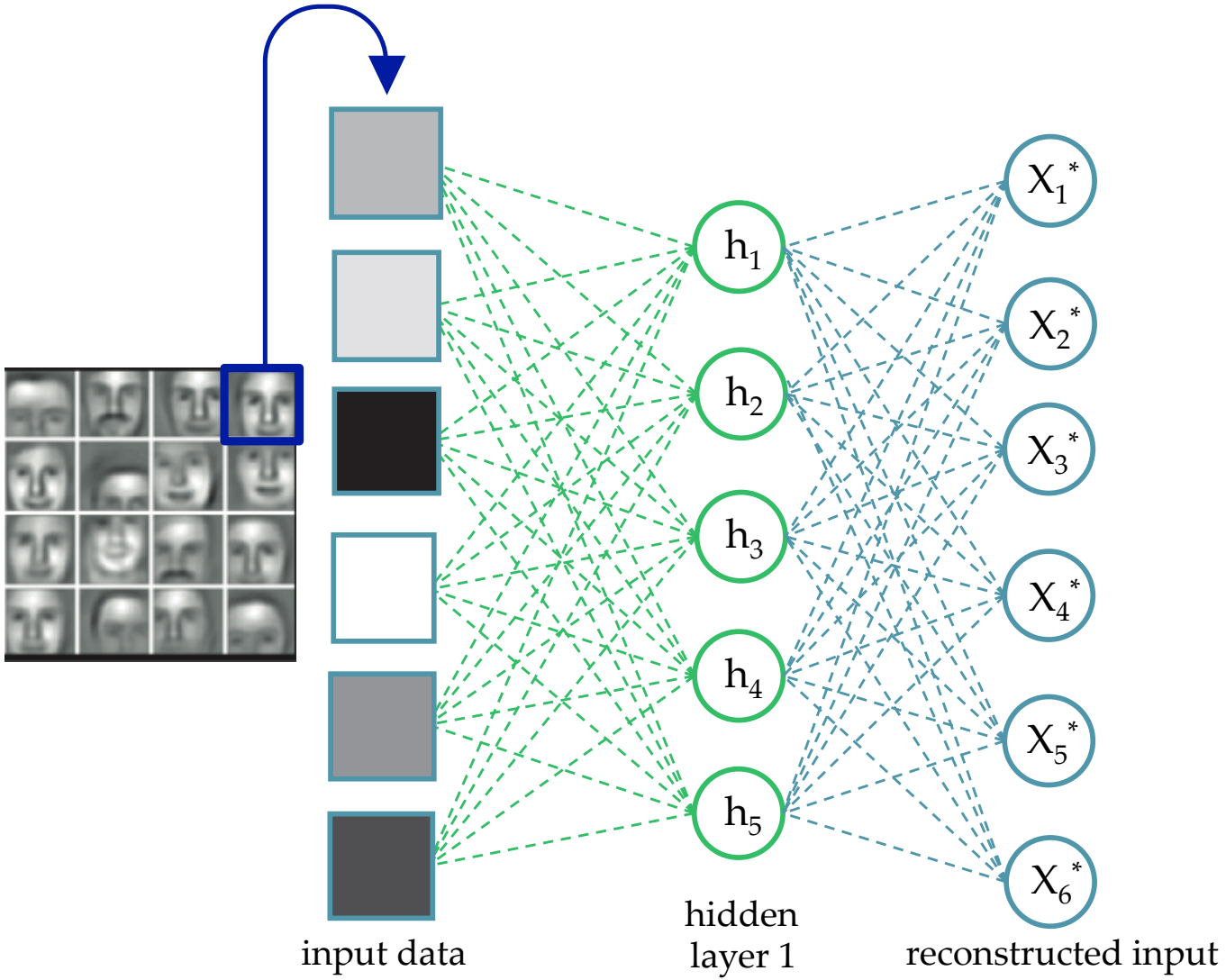




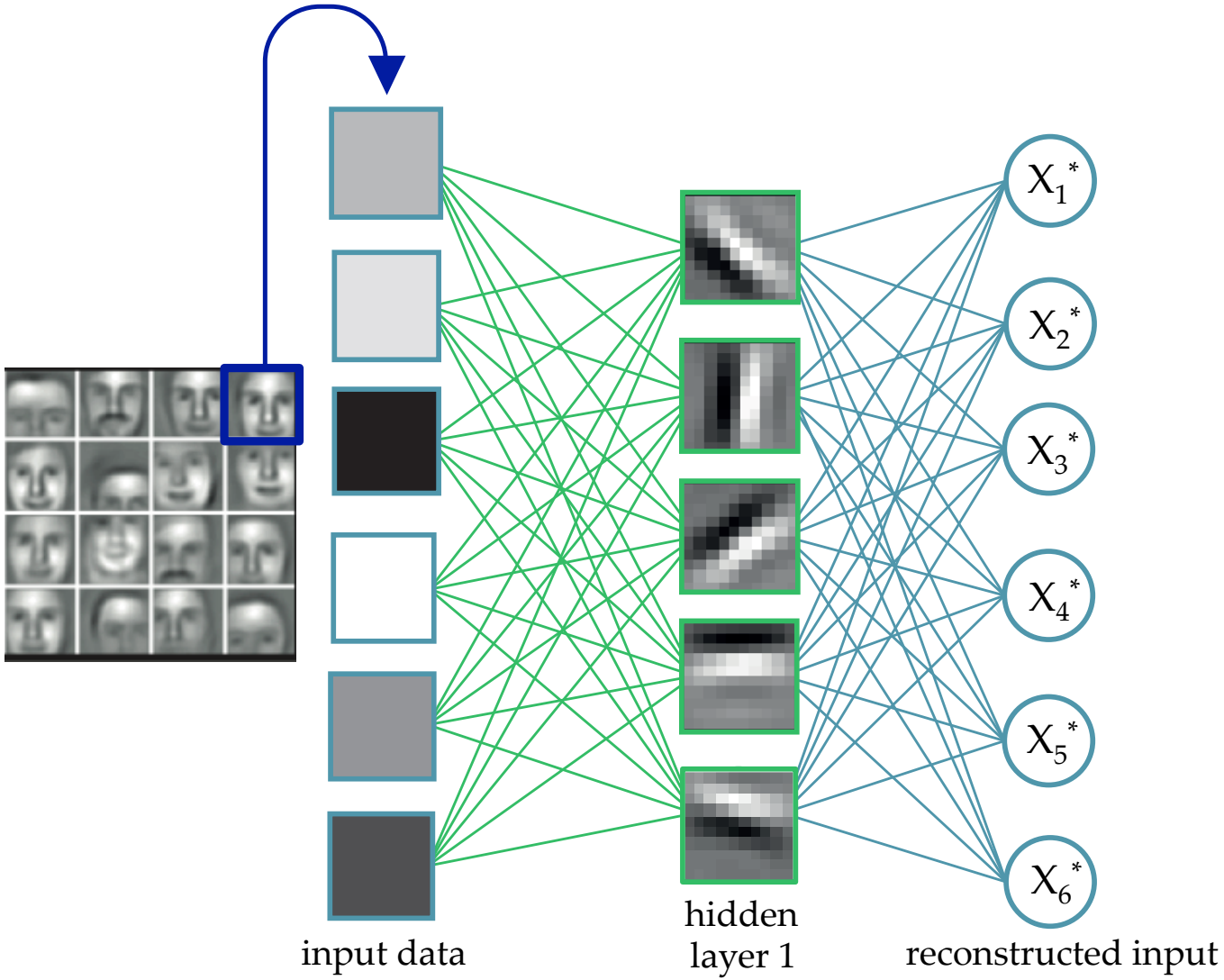


Transform the input data

Feature learning for hidden layer 1



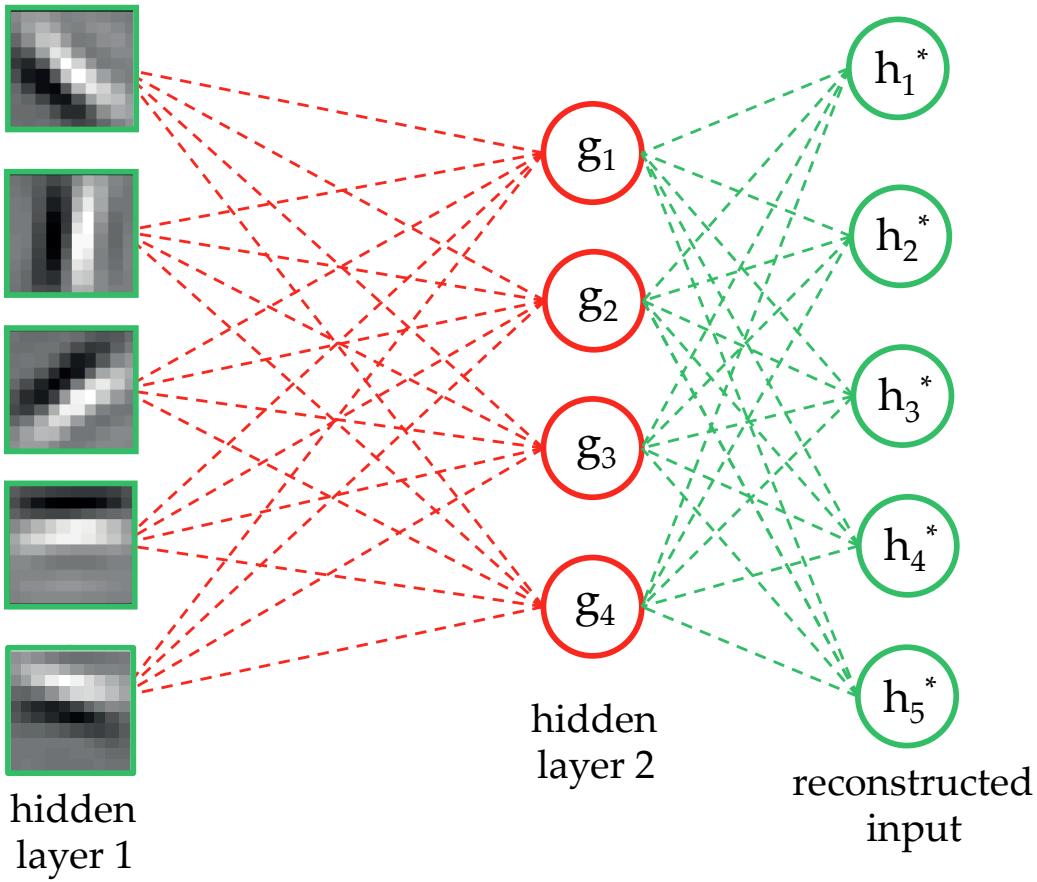
Feature learning for hidden layer 1



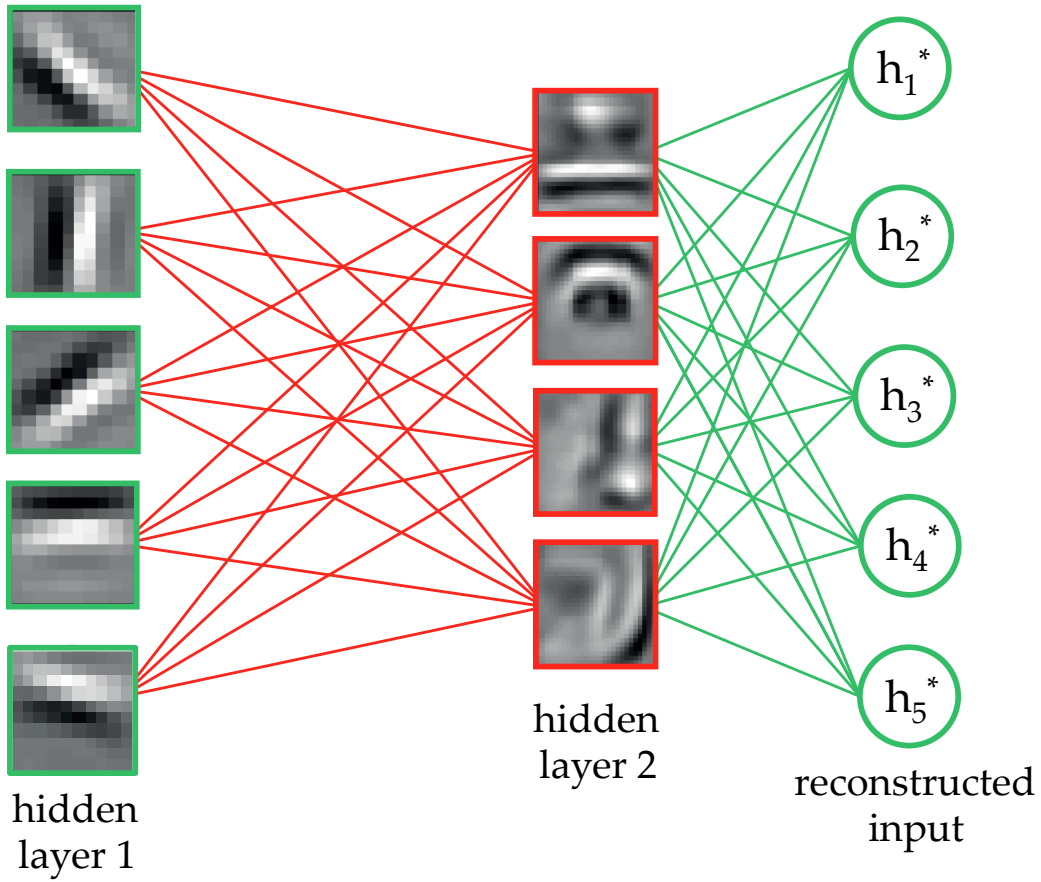
Low-level features become the new data

hidden
layer 1

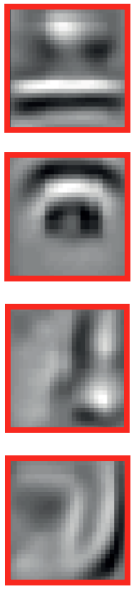
Feature learning for hidden layer 2



Feature learning for hidden layer 2

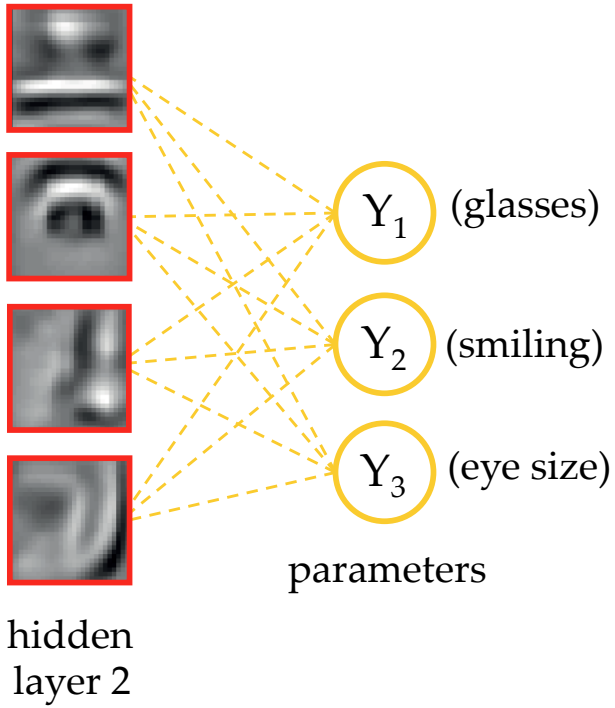


High-level features become the new data

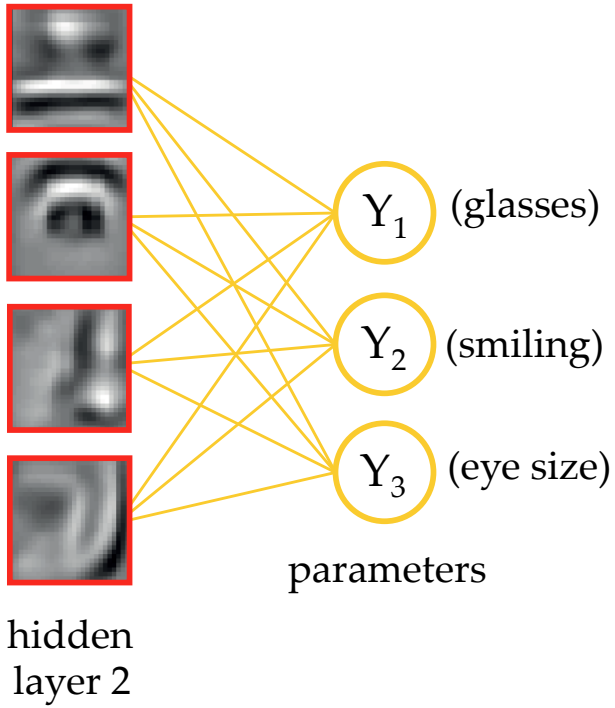


hidden
layer 2

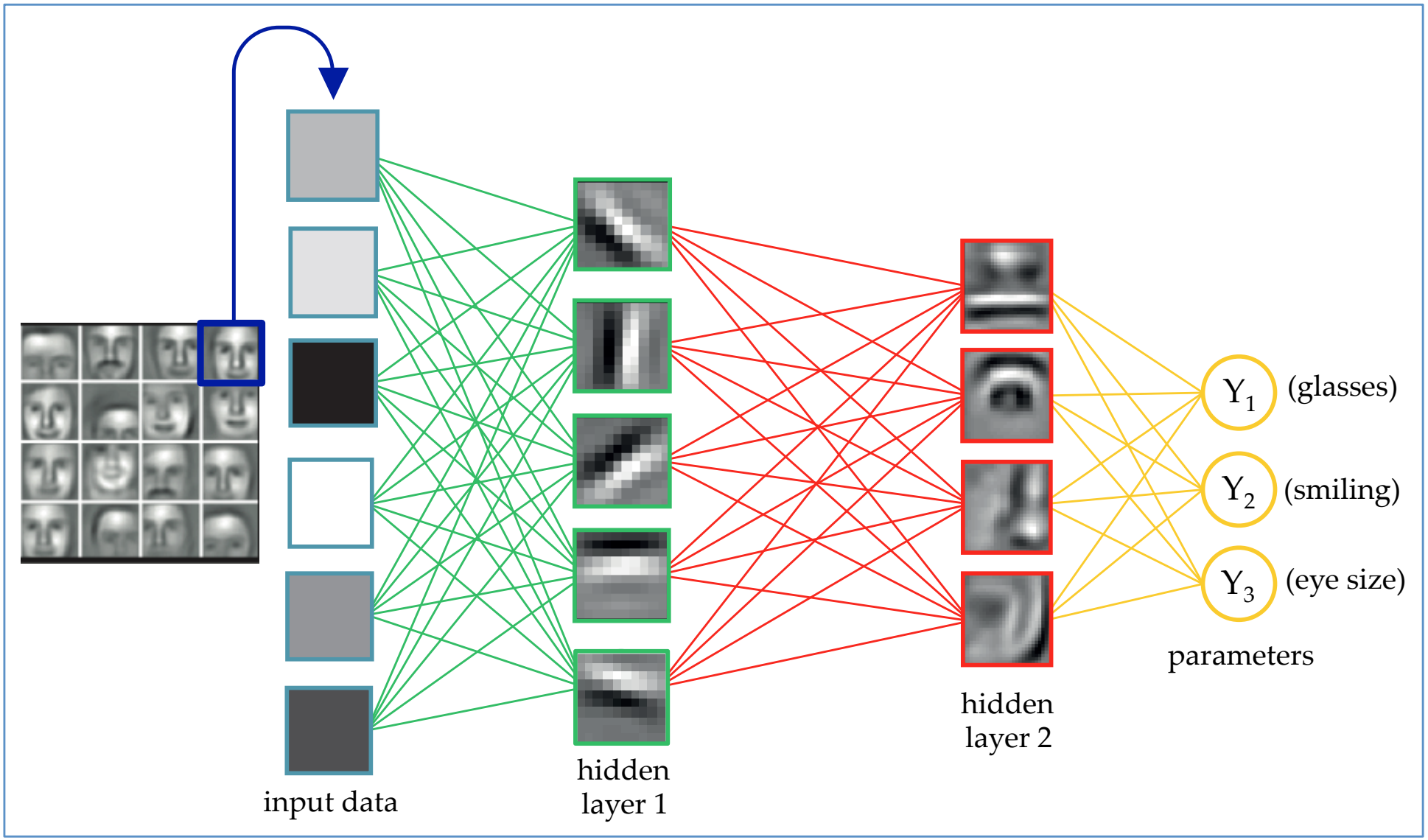
Last layer: supervised learning



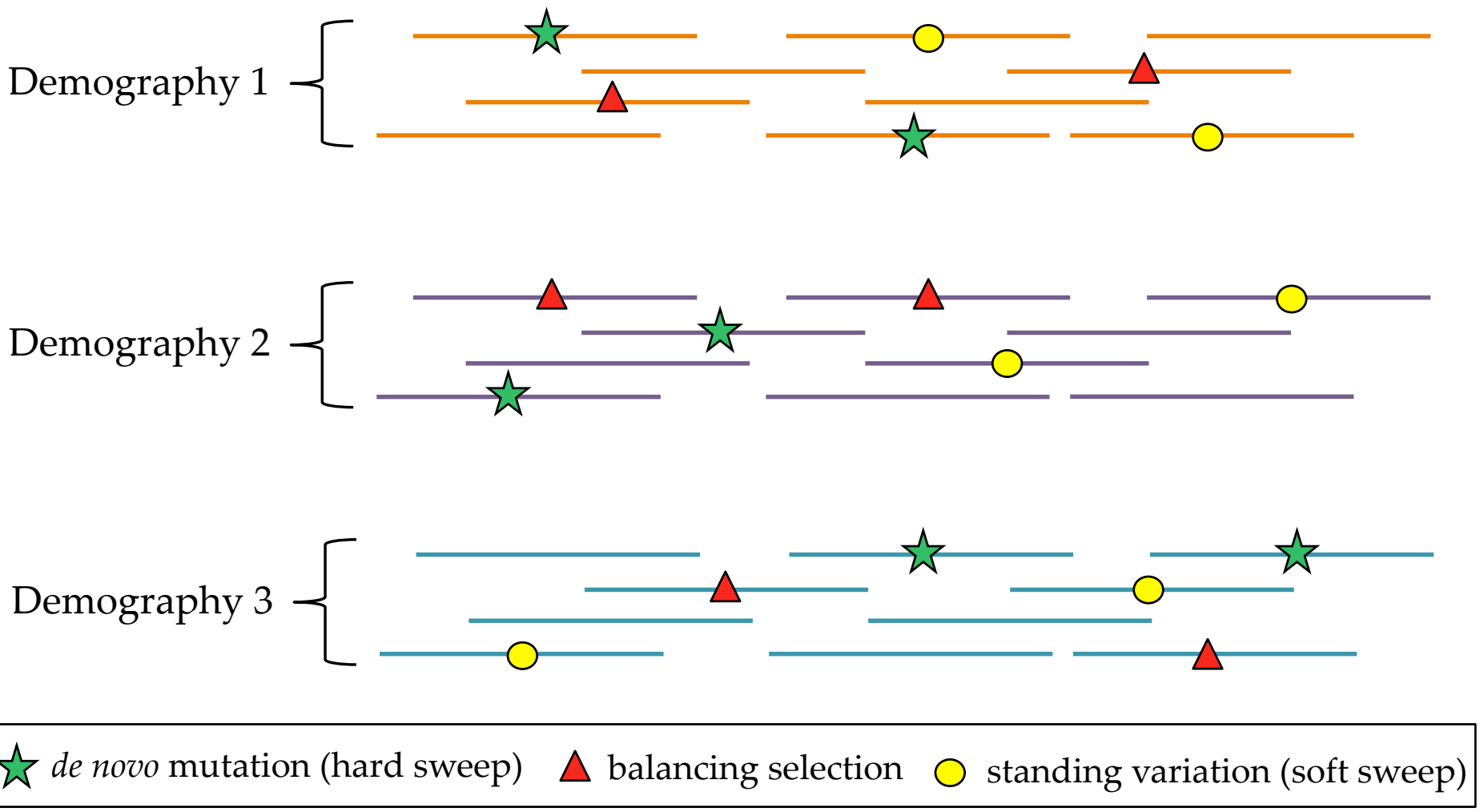
Last layer: supervised learning



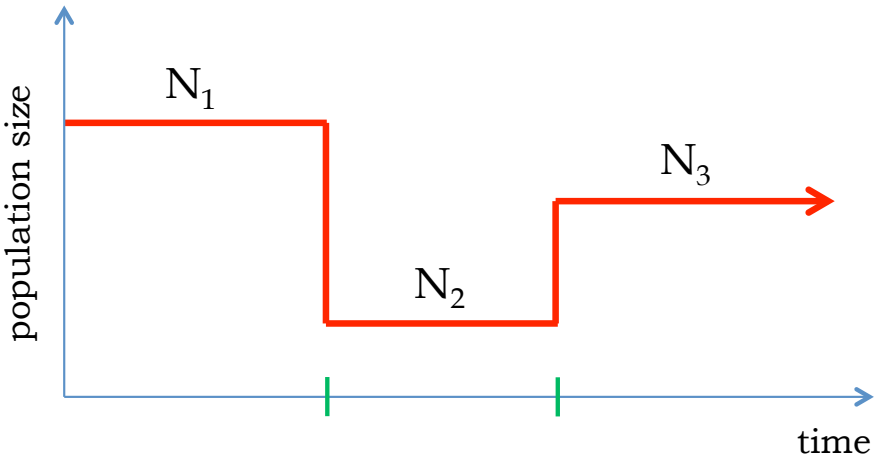
"Fine-tune" the entire deep network



Back to demography and selection: data design



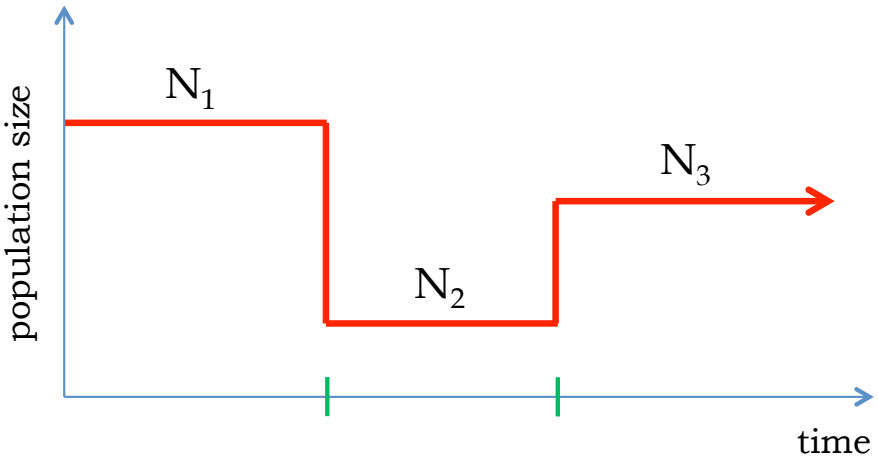
Simulated data details



400,000 datasets:

- 2,500 demographic models
- 160 regions/demography
- using `msms`

Simulated data details



400,000 datasets:

- 2,500 demographic models
- 160 regions/demography
- using `msms`

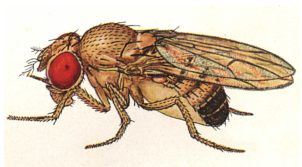
1. baseline effective population size: $N_e = 100,000$
2. $n = 100$ haplotypes
3. selection onset, selection strength, frequency of selected allele
4. 75% of data for training and 25% for testing

Summary statistics (features)

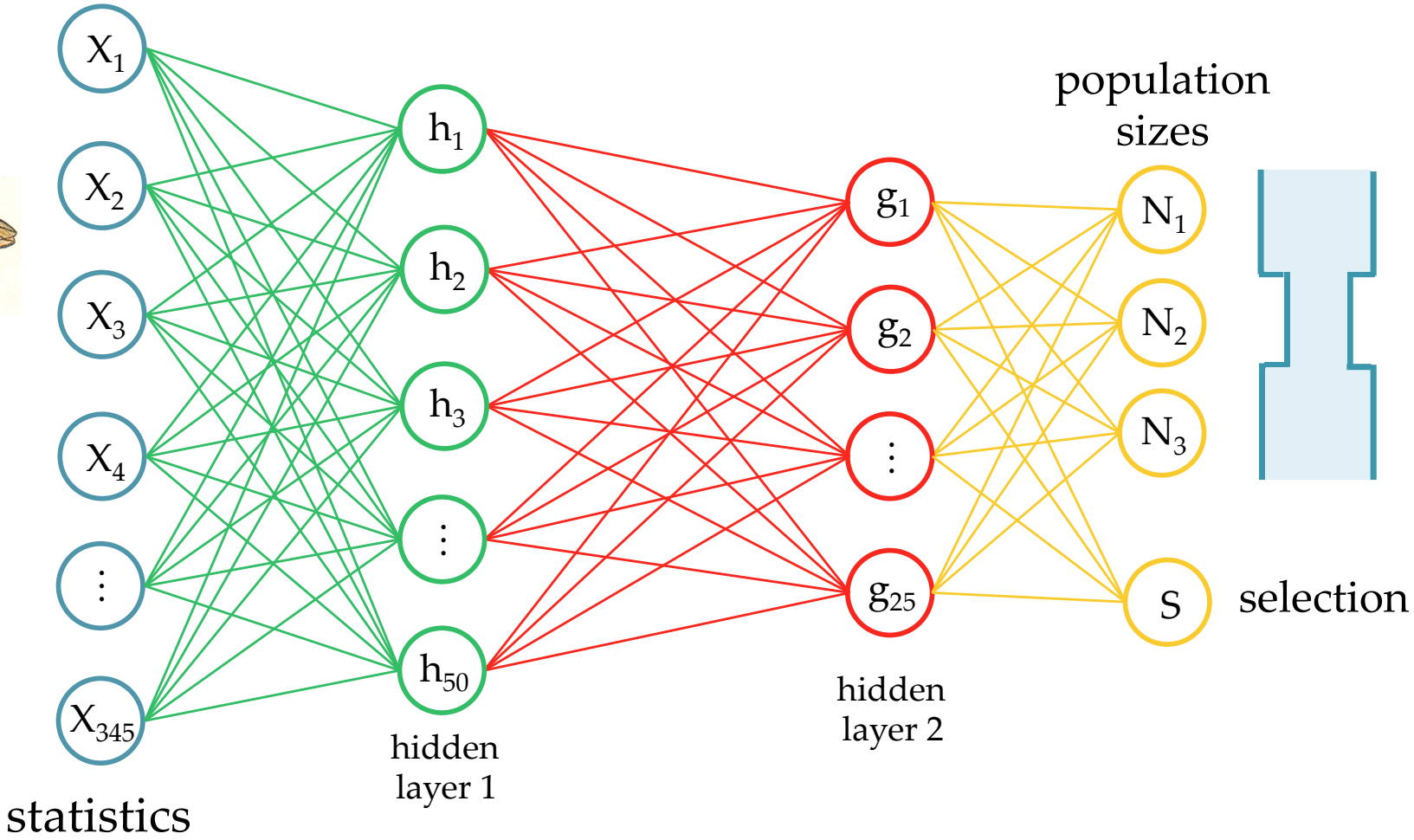
- Number of segregating sites **3 stats**
- Tajima's D **3 stats**
- Folded site frequency spectrum (SFS) **150 stats**
- Length distribution between segregating sites **48 stats**
- identity-by-state (IBS) tract length distribution **90 stats**
- Linkage disequilibrium (LD) distributions **48 stats**
- Haplotype frequency statistics **3 stats**

= 345 features total

A deep learning method for population genetics



GACTGGCTA
AGCTAGCTT
TAATCCGCA



Results

Confusion matrix for selection classes

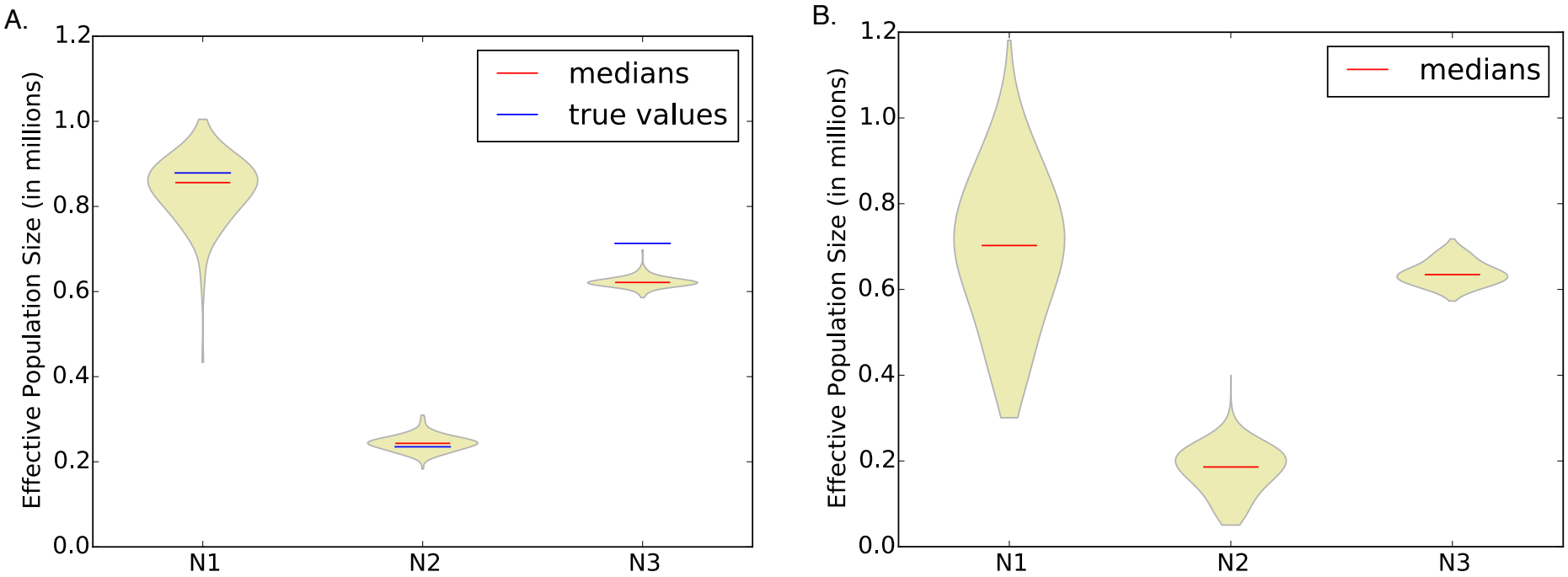
	Called Class			
True Class	Neutral	Hard Sweep	Soft Sweep	Balancing
Neutral	**0.9995**	0.0002	0.0003	0.0000
Hard Sweep	0.1434	**0.8333**	0.0032	0.0201
Soft Sweep	0.0096	0.0010	**0.9891**	0.0003
Balancing	0.0301	0.0356	0.0056	**0.9287**

Overall error: 6.2%

With and without unsupervised learning

		Called Class		
True Class	Neutral	Hard Sweep	Soft Sweep	Balancing
		Random Initialization		
Neutral	**1.000**	0.000	0.000	0.000
Hard Sweep	**0.978**	0.007	0.000	0.015
Soft Sweep	**1.000**	0.000	0.000	0.000
Balancing	**1.000**	0.000	0.000	0.000
		Autoencoder Initialization		
Neutral	**1.000**	0.000	0.000	0.000
Hard Sweep	0.145	**0.831**	0.004	0.021
Soft Sweep	0.011	0.001	**0.987**	0.000
Balancing	0.030	0.028	0.001	**0.941**

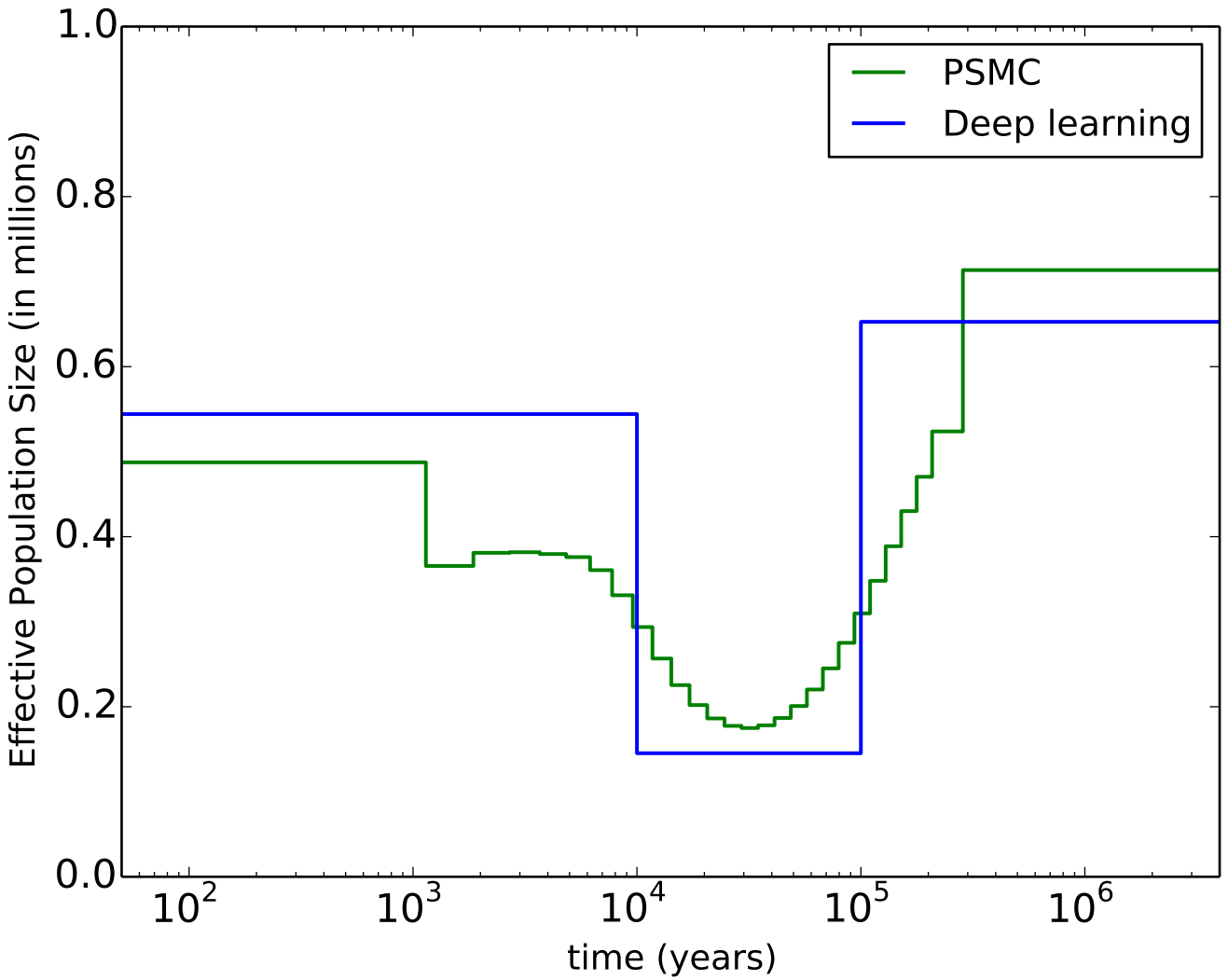
Population Size Accuracy



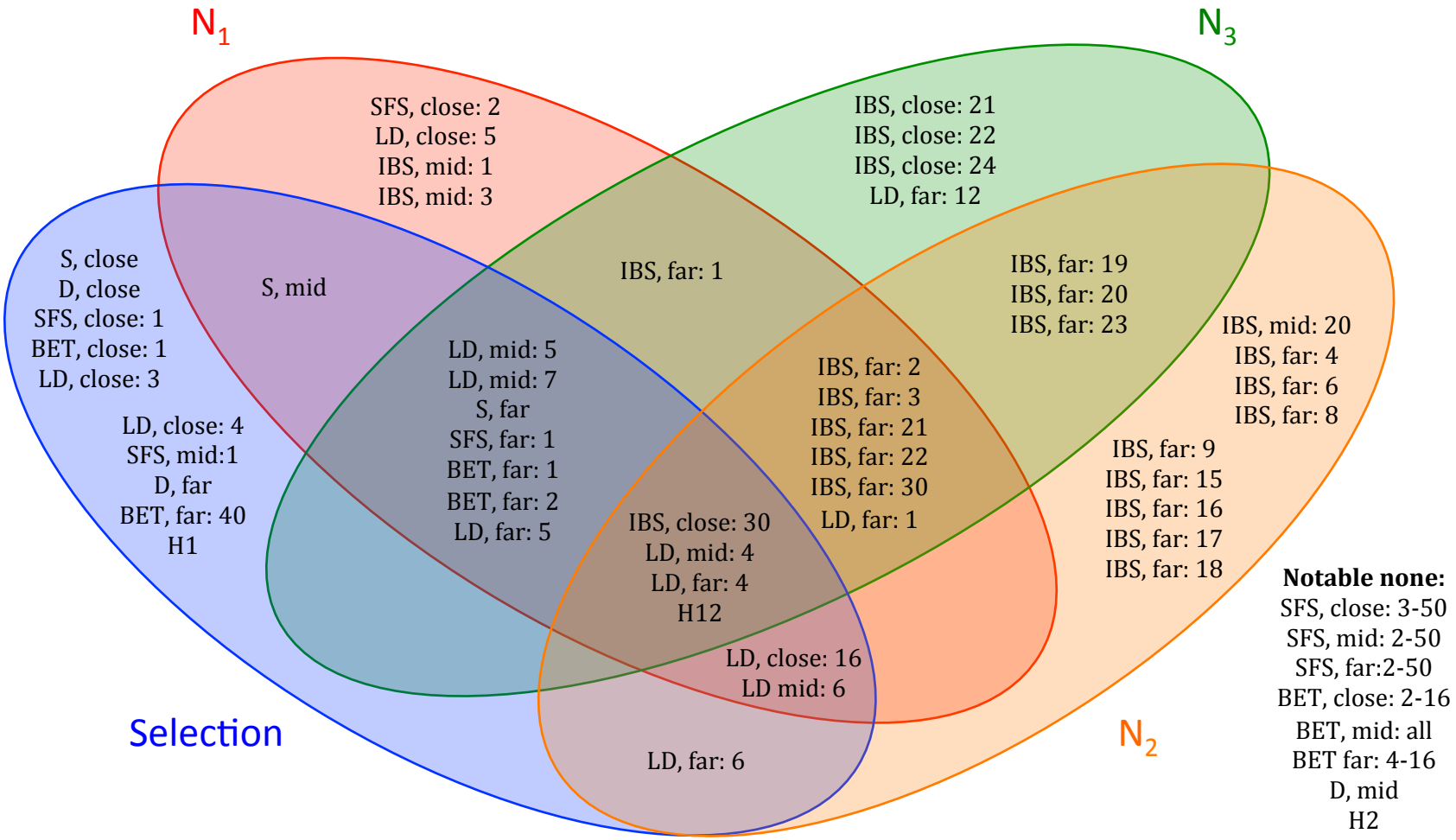
A. Population size results for an example simulated dataset.

B. Population size results for the real *Drosophila* data.

African *Drosophila*: history



Feature selection: "best" statistics



Conclusions and Future Work

Future Directions

Population sizes

- Recent population size changes
- Polar bears and other non-model species

Future Directions

Population sizes

- Recent population size changes
- Polar bears and other non-model species

Machine learning dual goals

- Incorporate biological models into deep learning
- Summary statistic free inference?

Thank you!

Collaborators

- Chloe Lee '17
- Artemis Metaxa-Kakavouli '19
- Silvana Saca '18
- Alice Yang '17
- Eline Lorenzen
- Yun S. Song

Funding