Automatic evolutionary inference using Generative Adversarial Networks

Sara Mathieson

Assistant Professor

HAVERFORD COLLEGE
Department of Computer Science

NIH, NHGRI

Machine Learning in Genomics Workshop

April 13, 2021

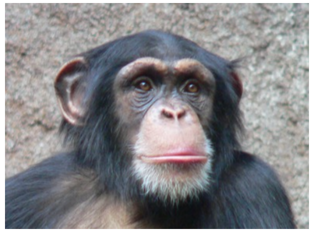
Central question in population genetics: data -> quantify evolution

INPUT

Sites or SNPs

samples/haplotypes

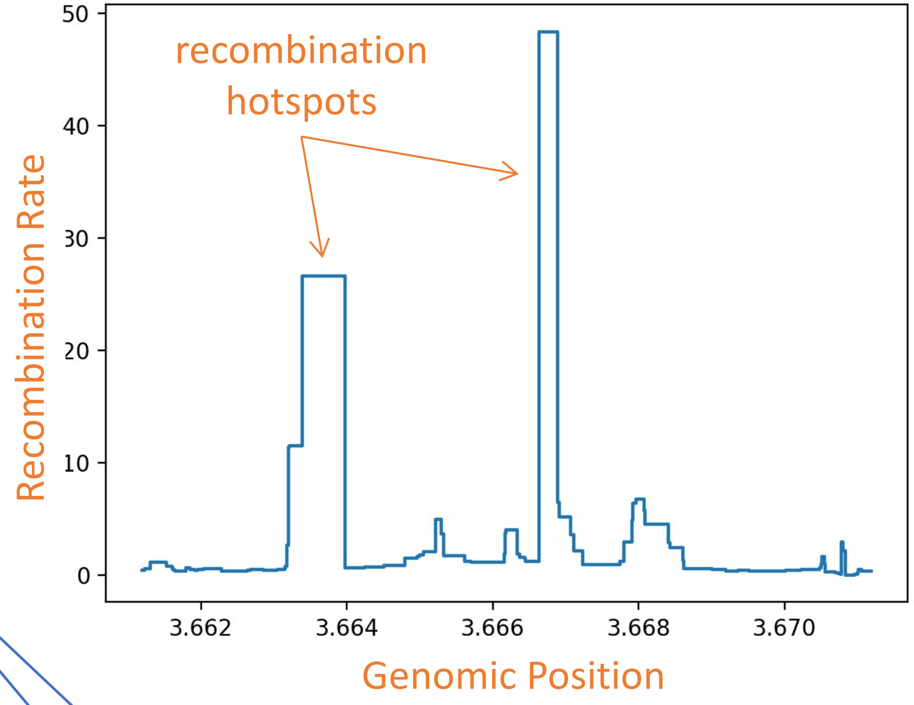
```
10000100010000001011000010100100101110000010
11000011101100111101000010111100011001000011
01001001010001000000000000001111011100001001
11011000001100100000000010001000110001110011
00101100100000000000011010111101001000111000
00100010100000000111000110000010101110010000
00000001010100110100101011001001000000010000
00000000101100010000000100000001000110110000
10101101001010010011000000011110011000000011
00001000111010000000000011101001010010100111
00000001110010000111000000111110000000110101
010001100001100000000001010000001011110000101
01000000101010000100001000010100000001001011
00000100000010101011000010110110100000100100
```



Images: wikipedia

?

OUTPUT



- Population size changes
- Natural selection
- Mutation rate variation
- Migration, admixture, introgression
- Heritable traits and diseases

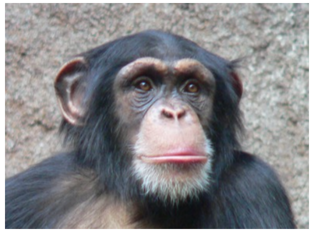
Central question in population genetics: data -> quantify evolution

INPUT

Sites or SNPs

samples/haplotypes

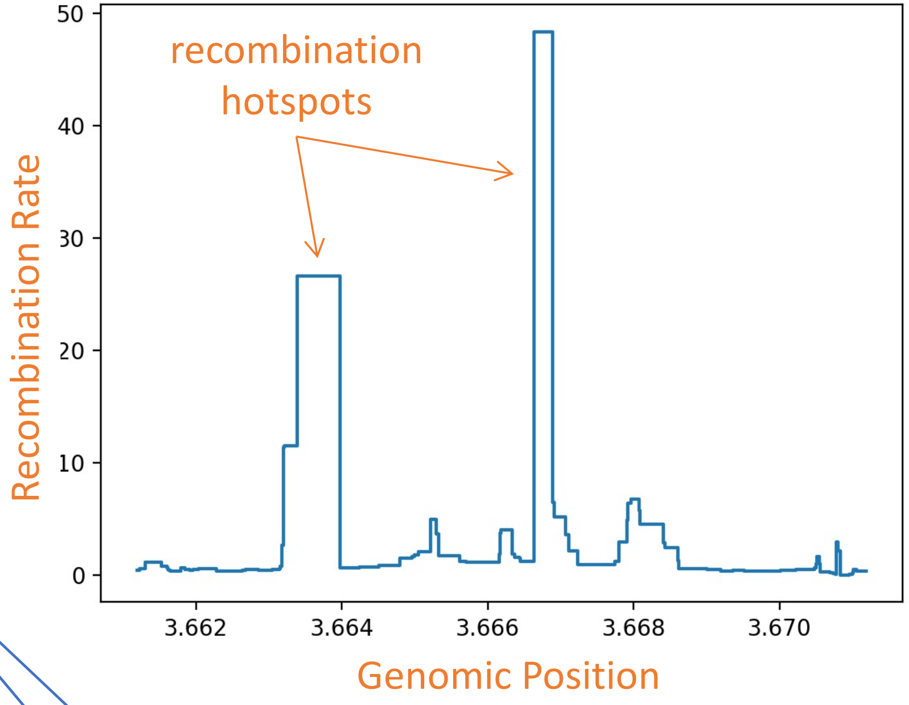
```
10000100010000001011000010100100101110000010
11000011101100111101000010111100011001000011
01001001010001000000000000001111011100001001
11011000001100100000000100010001100011100011
00101100100000000000011010111101001000111000
00100010100000000111000110000010101110010000
00000001010100110100101011001001000000010000
00000000101100010000000100000001000110110000
10101101001010010011000000011110011000000011
00001000111010000000000011101001010010100111
00000001110010000111000000111110000000110101
01000110000110000000001010000001011110000101
01000000101010000100001000010100000001001011
00000100000010101011000010110110100000100100
```



Images: wikipedia

Fast?
Flexible?
Machine learning?

OUTPUT



- Population size changes
- Natural selection
- Mutation rate variation
- Migration, admixture, introgression
- Heritable traits and diseases

Outline

- Shift to machine learning in population genetics
- Shift away from summary statistics to “raw” data
- GANs and adversarial training
  - pg−gan algorithm for creating realistic simulated data
- Results on human data from Africa, Europe, and East Asia

Outline

- Shift to machine learning in population genetics

- Shift away from summary statistics to “raw” data

- GANs and adversarial training
  - pg-gan algorithm for creating realistic simulated data

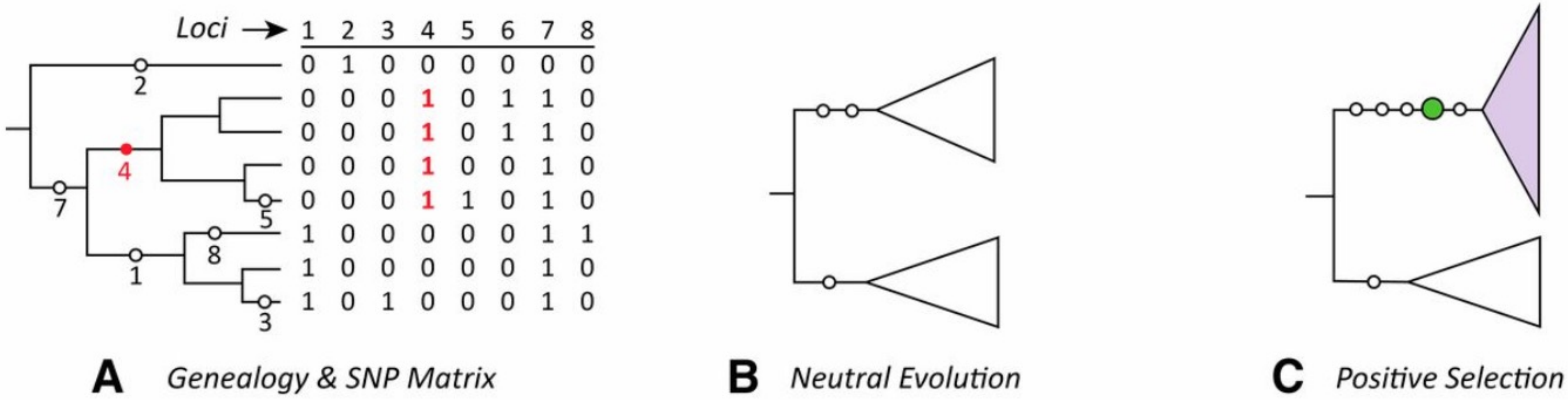
- Results on human data from Africa, Europe, and East Asia

2013: Using machine learning to infer selection

Learning Natural Selection from the Site Frequency Spectrum

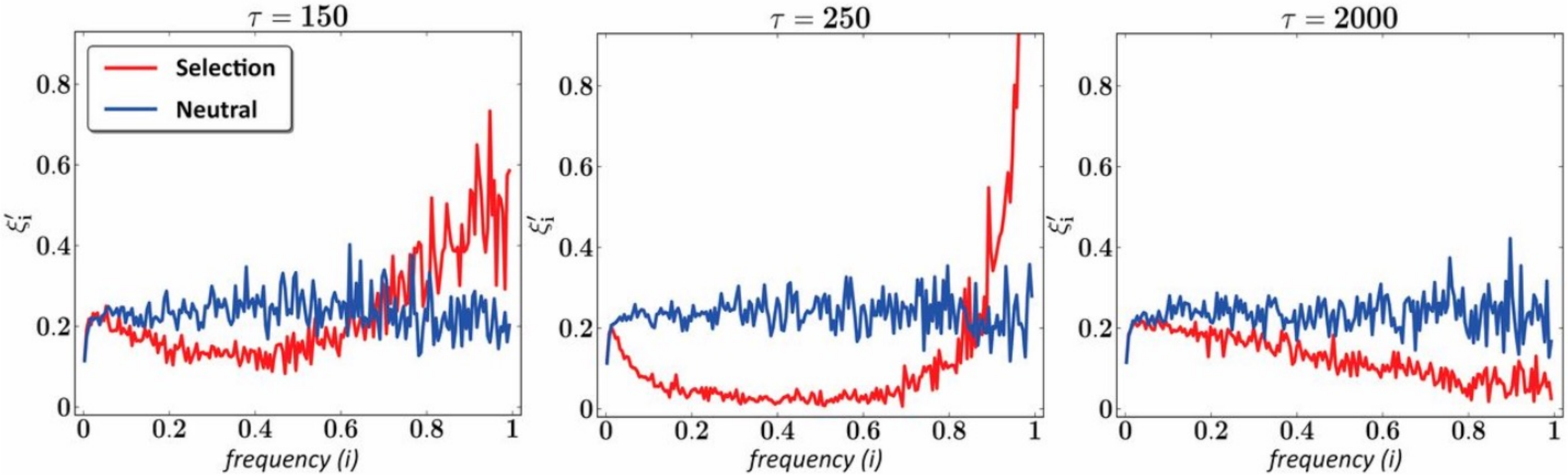
Roy Ronen, Nitin Udpa, Eran Halperin and Vineet Bafna

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

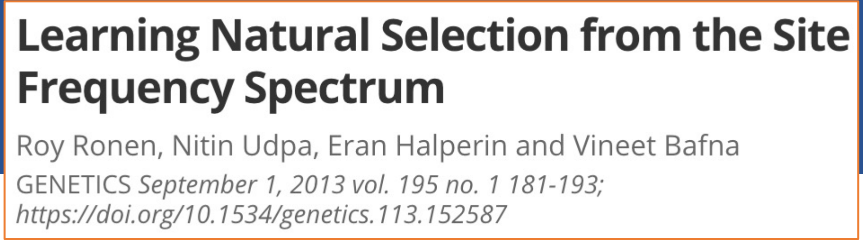


1	2	3	4	5	6	7	8
0	1	0	0	0	0	0	0
0	0	0	1	0	1	1	0
0	0	0	1	0	1	1	0
0	0	0	1	0	0	1	0
0	0	0	1	1	0	1	0
1	0	0	0	0	0	1	1
1	0	0	0	0	0	1	0
1	0	1	0	0	0	1	0

A *Genealogy & SNP Matrix* **B** *Neutral Evolution* **C** *Positive Selection*



2013: Using machine learning to infer selection

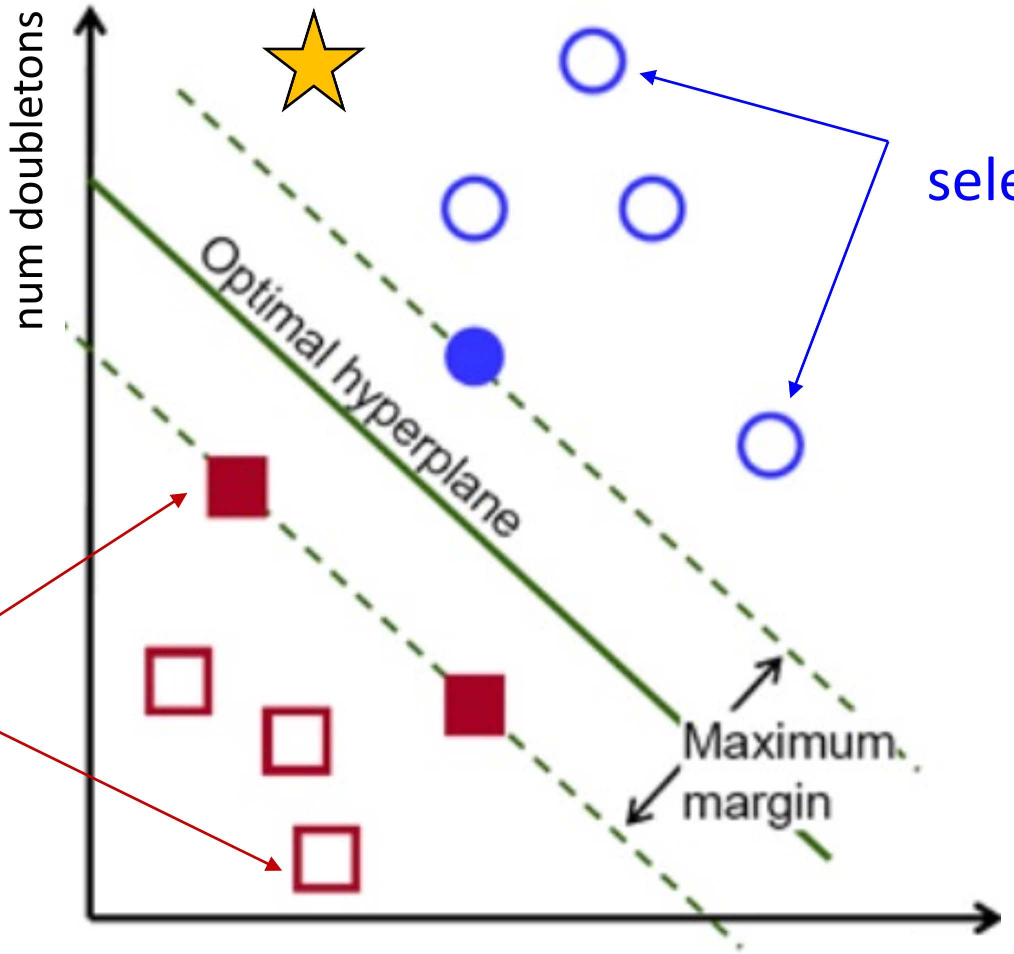


Learning Natural Selection from the Site Frequency Spectrum

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 machines (SVM)



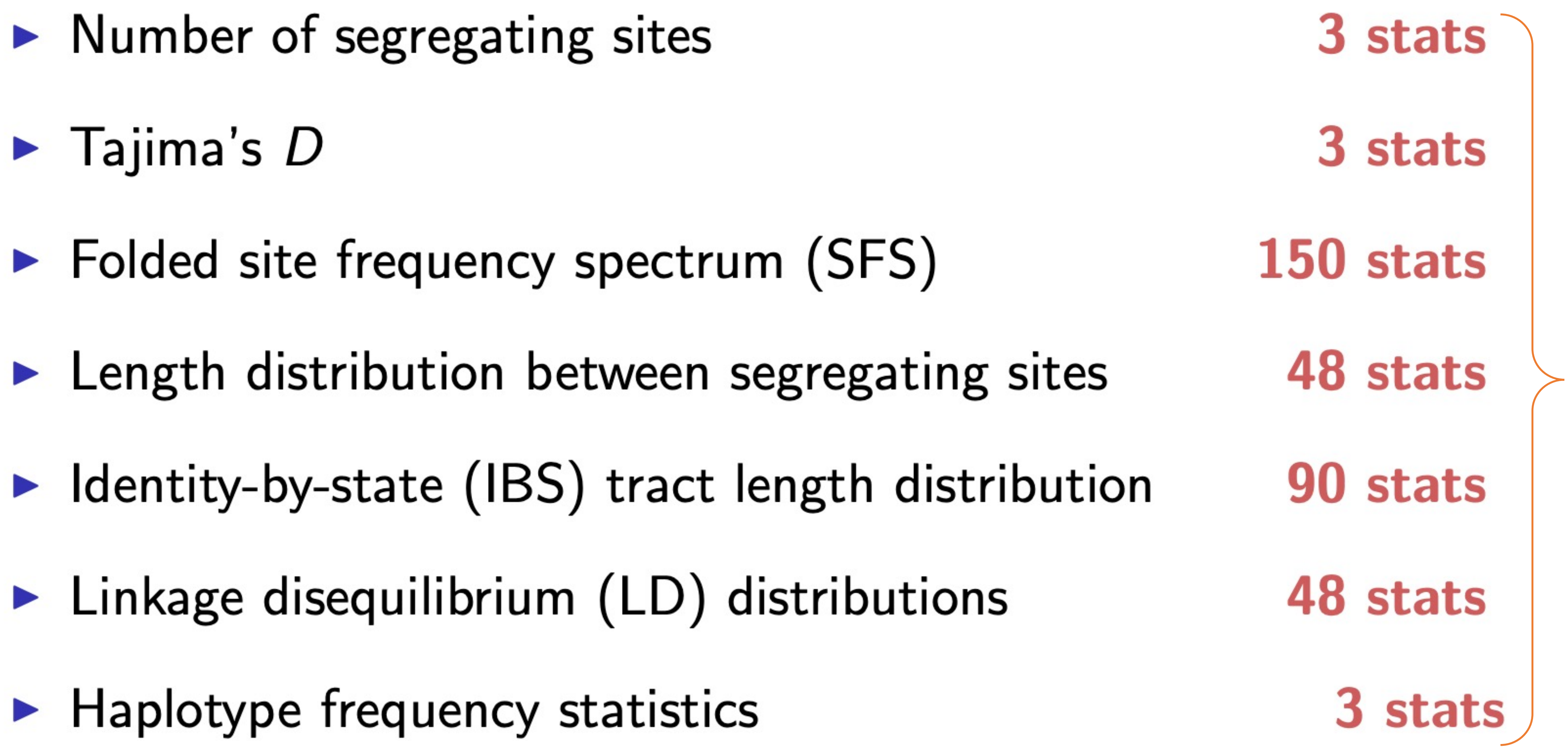
region of real data

Regions under selection (simulated)

Neutral regions (simulated)

Image from: “Towards Data Science”

Which summary statistics to use?

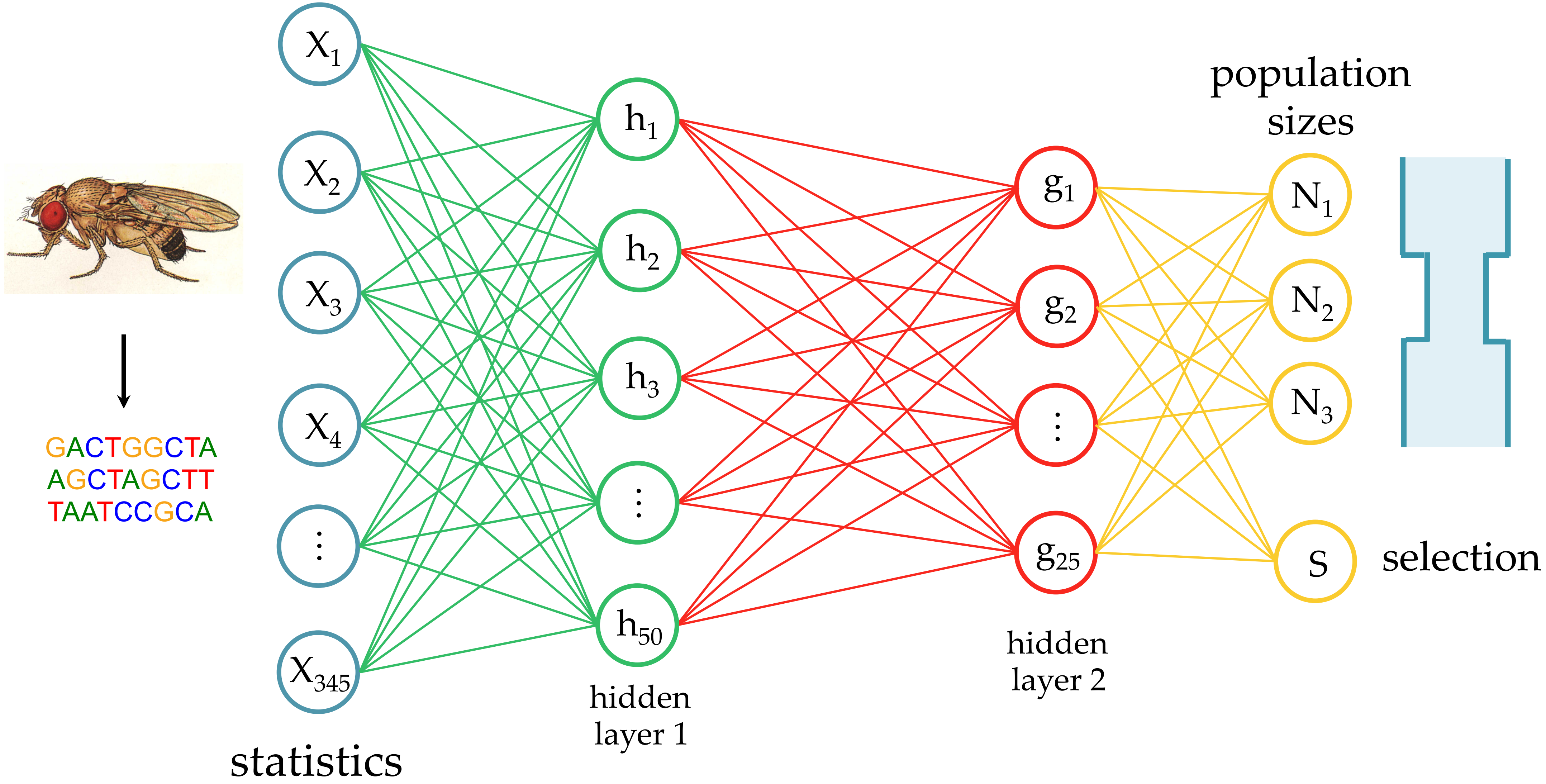


Number of segregating sites	3 stats
Tajima's *D*	3 stats
Folded site frequency spectrum (SFS)	150 stats
Length distribution between segregating sites	48 stats
Identity-by-state (IBS) tract length distribution	90 stats
Linkage disequilibrium (LD) distributions	48 stats
Haplotype frequency statistics	3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

345 total

Example summary statistics from "Deep learning for population genetic inference", Sheehan and Song, 2016

2016: deep learning with summary statistics



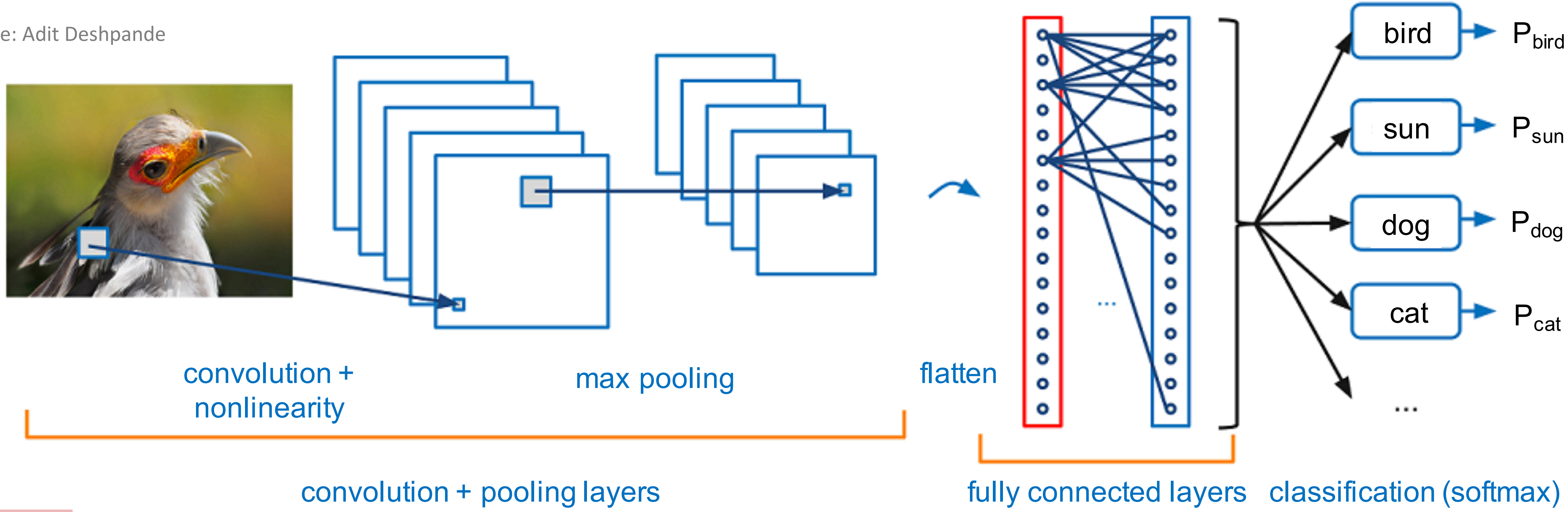
“Deep learning for population genetic inference”, Sheehan and Song, *PLOS Comp Bio*, 2016

Outline

- Shift to machine learning in population genetics
- Shift away from summary statistics to “raw” data
- GANs and adversarial training
  - pg-gan algorithm for creating realistic simulated data
- Results on human data from Africa, Europe, and East Asia

Can we do better? Convolutional neural networks (CNNs)

Figure: Adit Deshpande



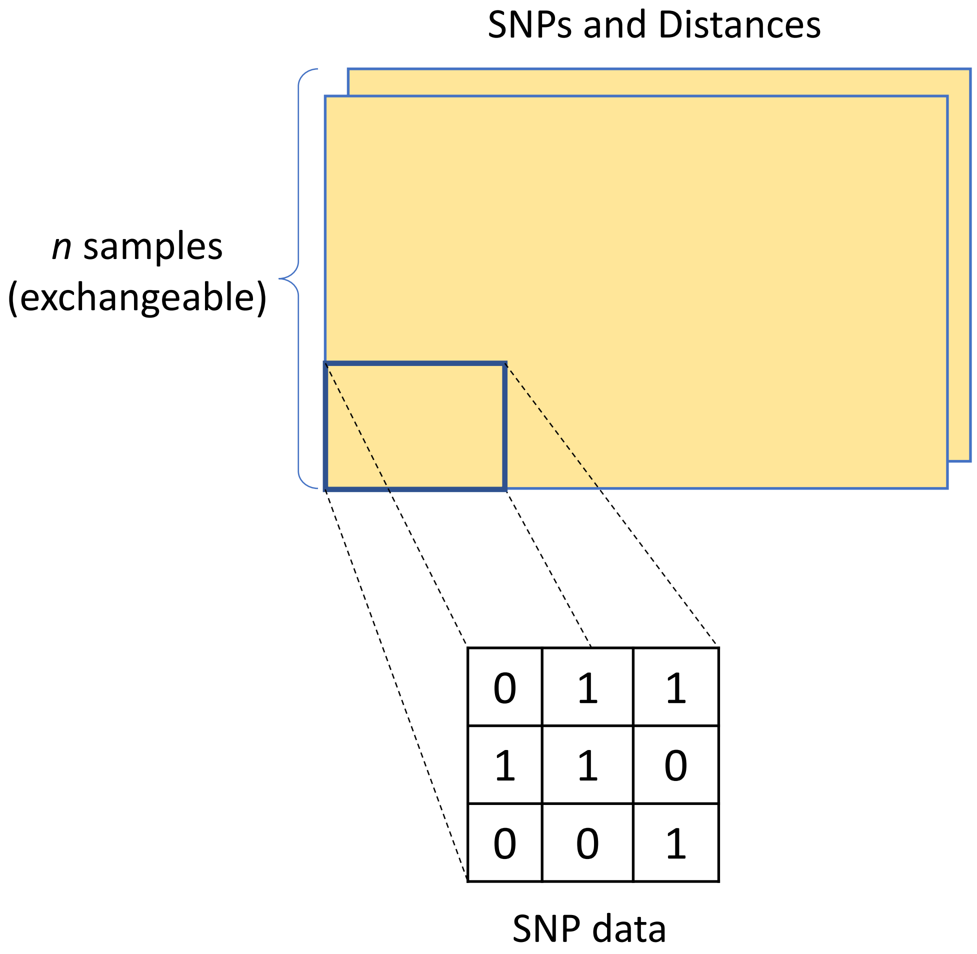
Issues

1. Image CNNs are optimized for different local features
2. For unstructured populations, sample (row) order doesn't matter

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

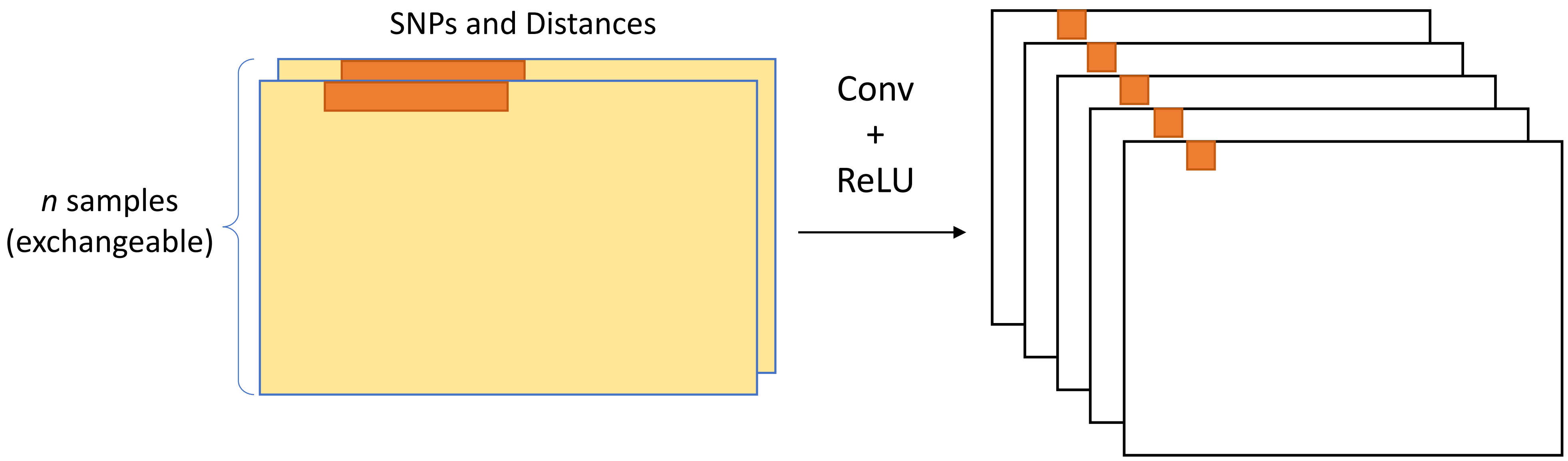
2018: CNN for "raw" population genetic data



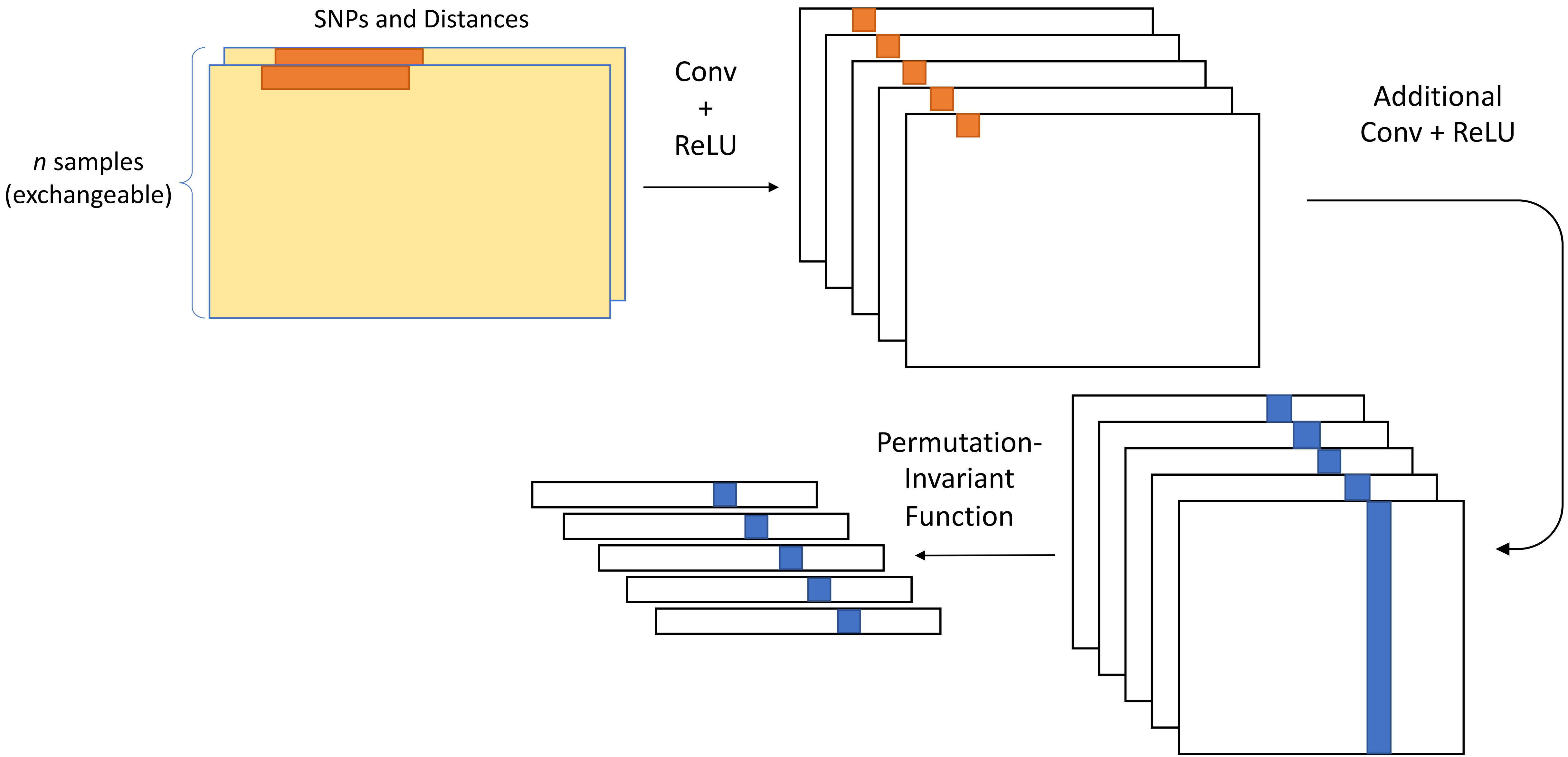
0	1	1
1	1	0
0	0	1

SNP data

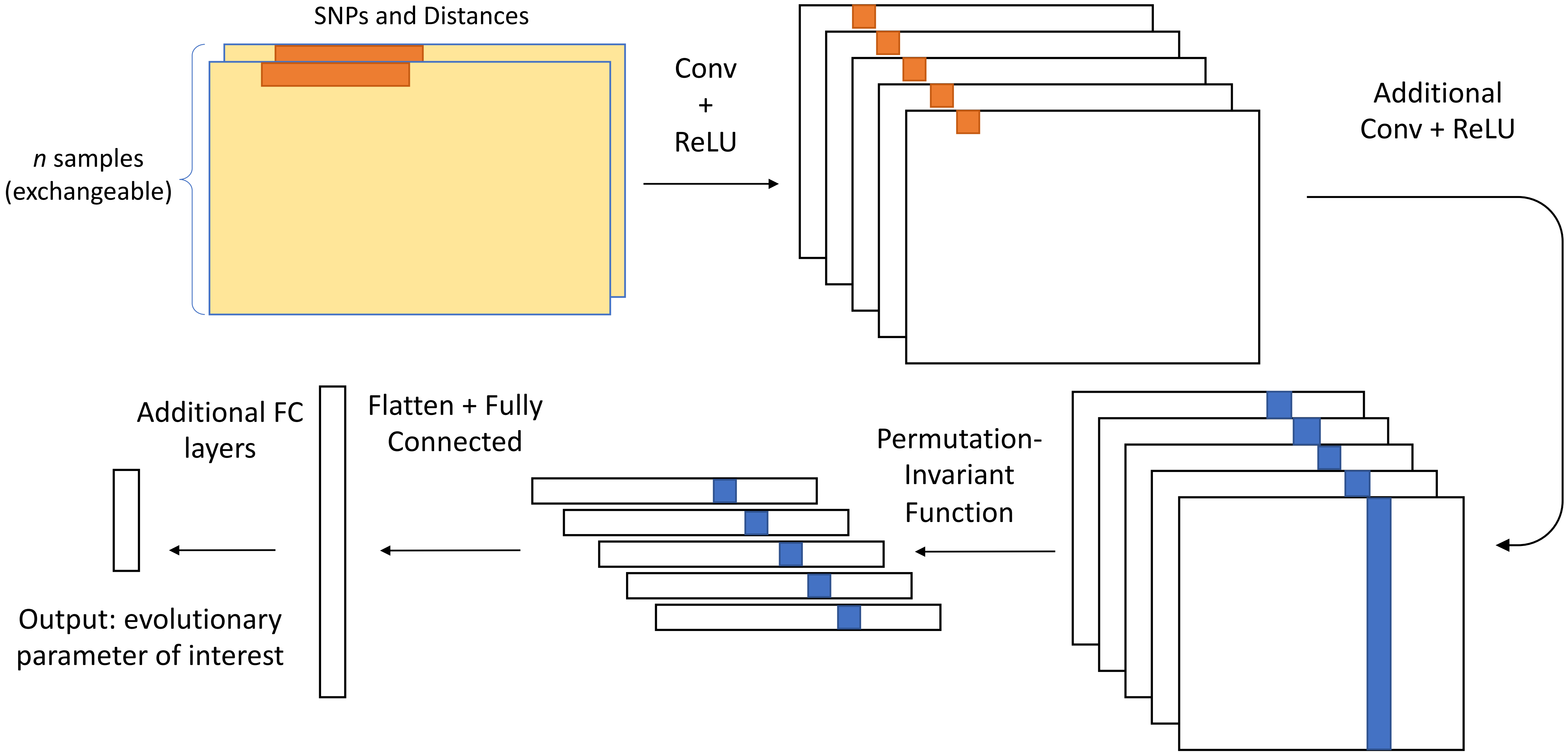
2018: CNN for “raw” population genetic data



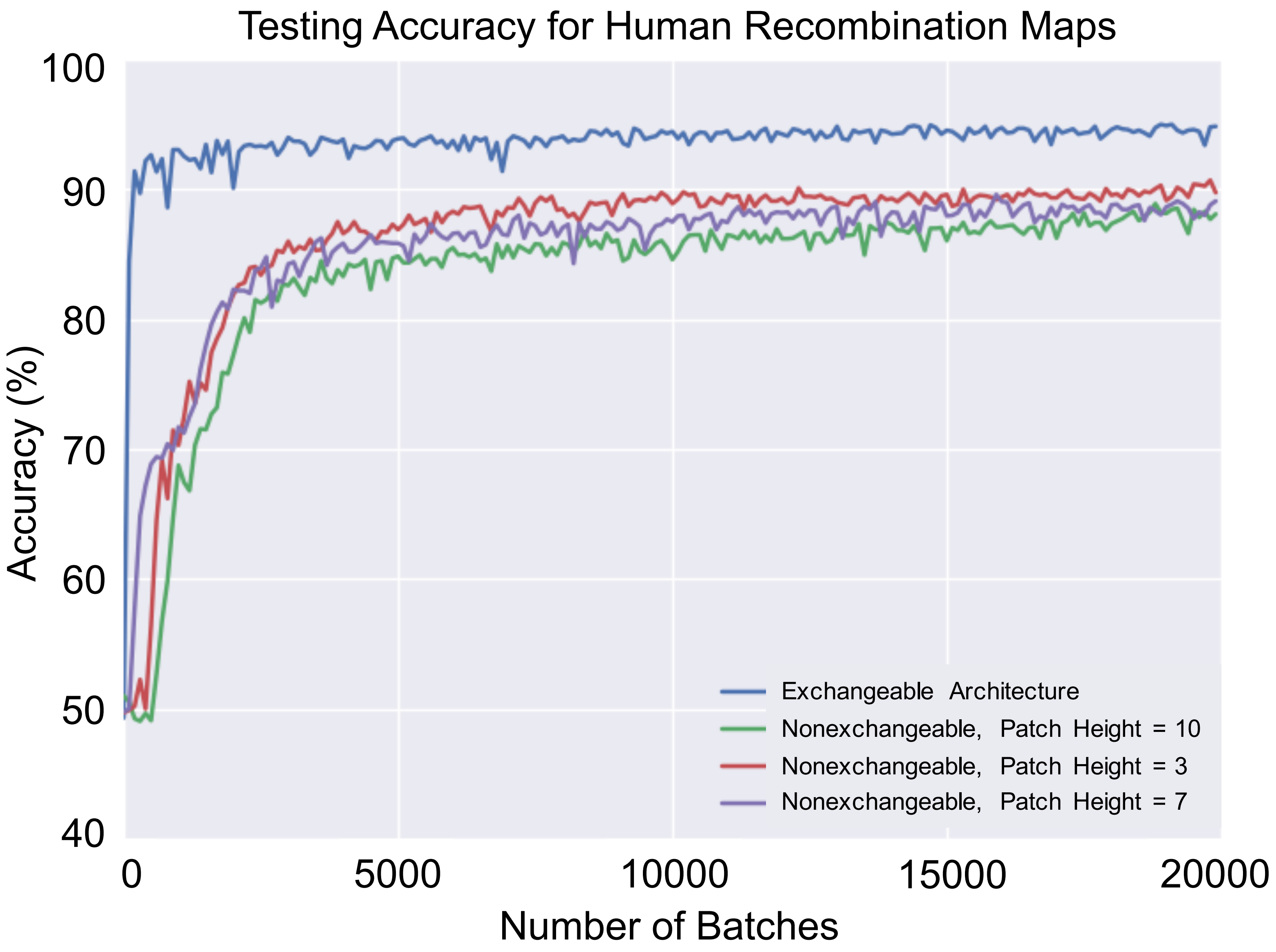
2018: CNN for “raw” population genetic data



2018: CNN for "raw" population genetic data


SNPs and Distances
n samples (exchangeable)
Conv + ReLU
Additional Conv + ReLU
Permutation-Invariant Function
Flatten + Fully Connected
Additional FC layers
Output: evolutionary parameter of interest

Impact of permutation-invariant architecture (recombination hotspots)



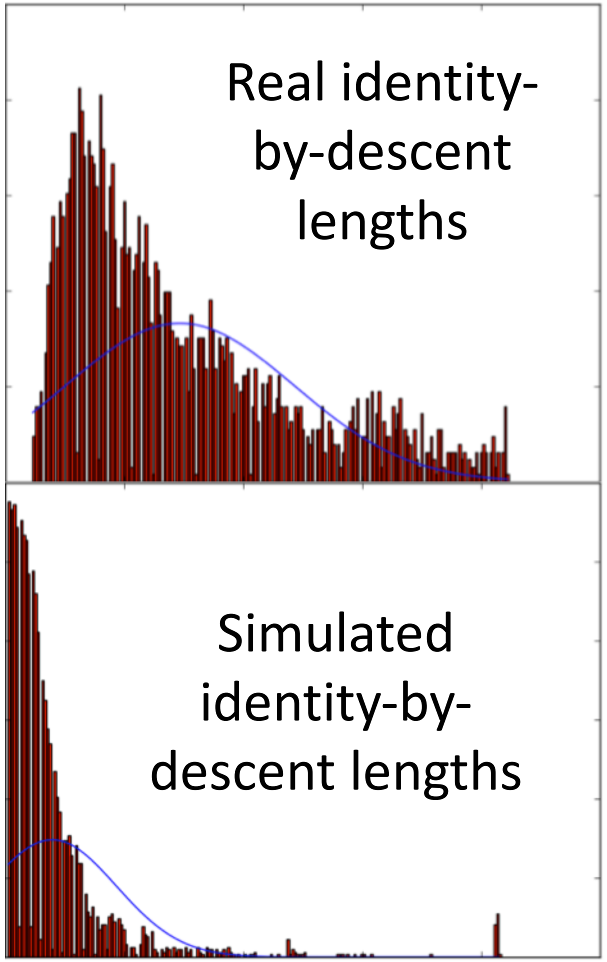
Outline

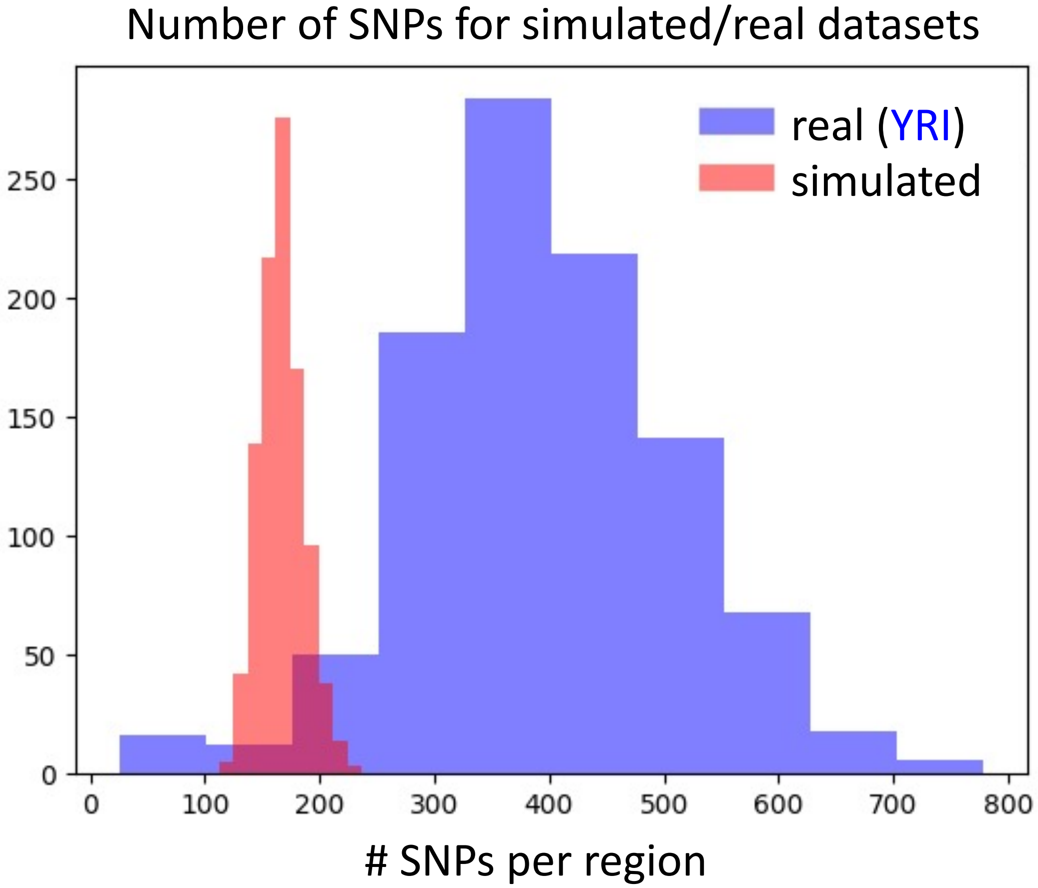
- Shift to machine learning in population genetics
- Shift away from summary statistics to “raw” data
- GANs and adversarial training
  - pg-gan algorithm for creating realistic simulated data
- Results on human data from Africa, Europe, and East Asia

Even using good simulation programs, it is difficult to match real data

High-quality simulated data is crucial!

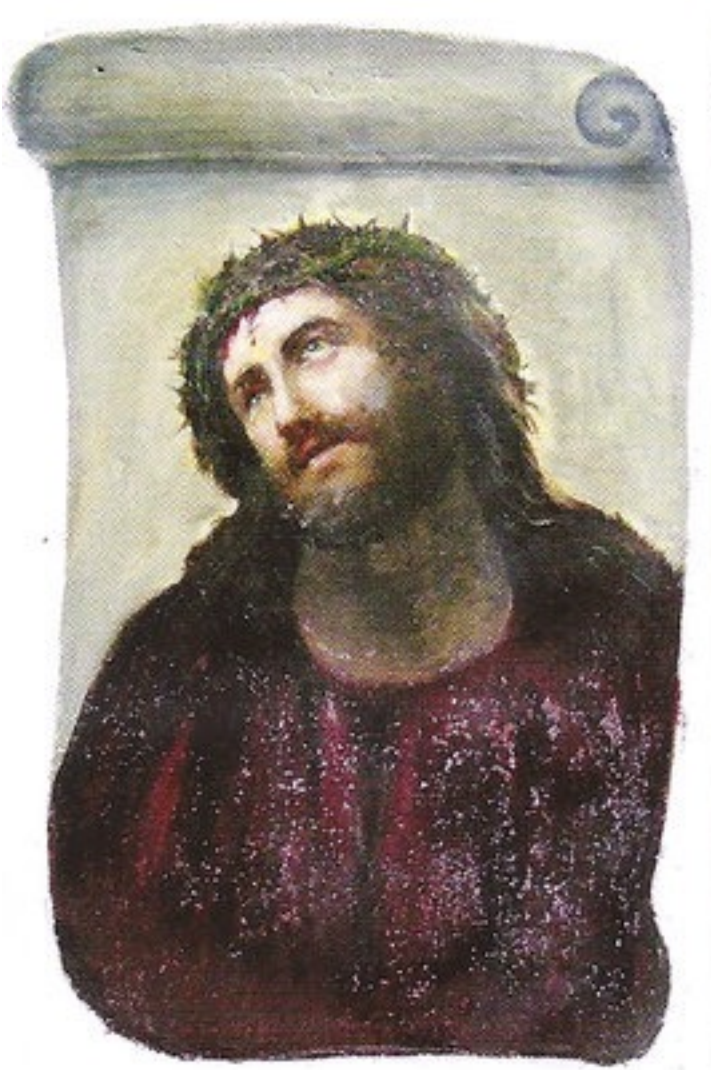
- Develop intuition
- Validate methods
- Provide training data for machine learning methods
- Popular simulators: SLiM, msprime



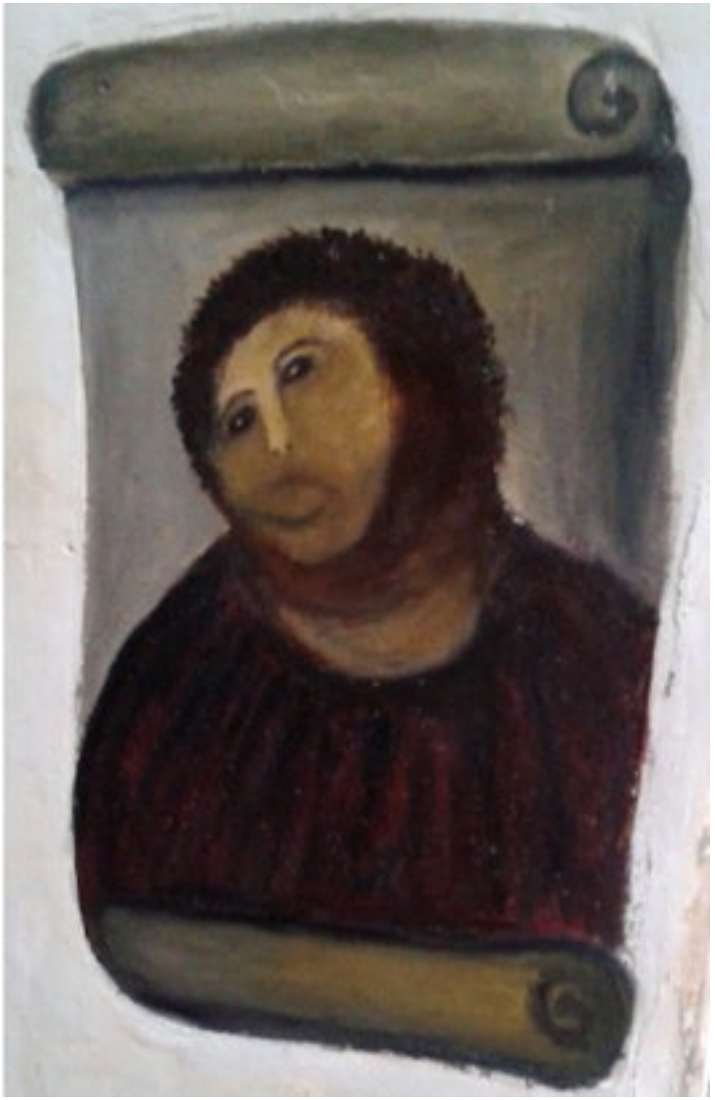

Number of SNPs for simulated/real datasets

YRI: Yoruba in Ibadan, Nigeria

Idea behind GANs (Generative Adversarial Networks)

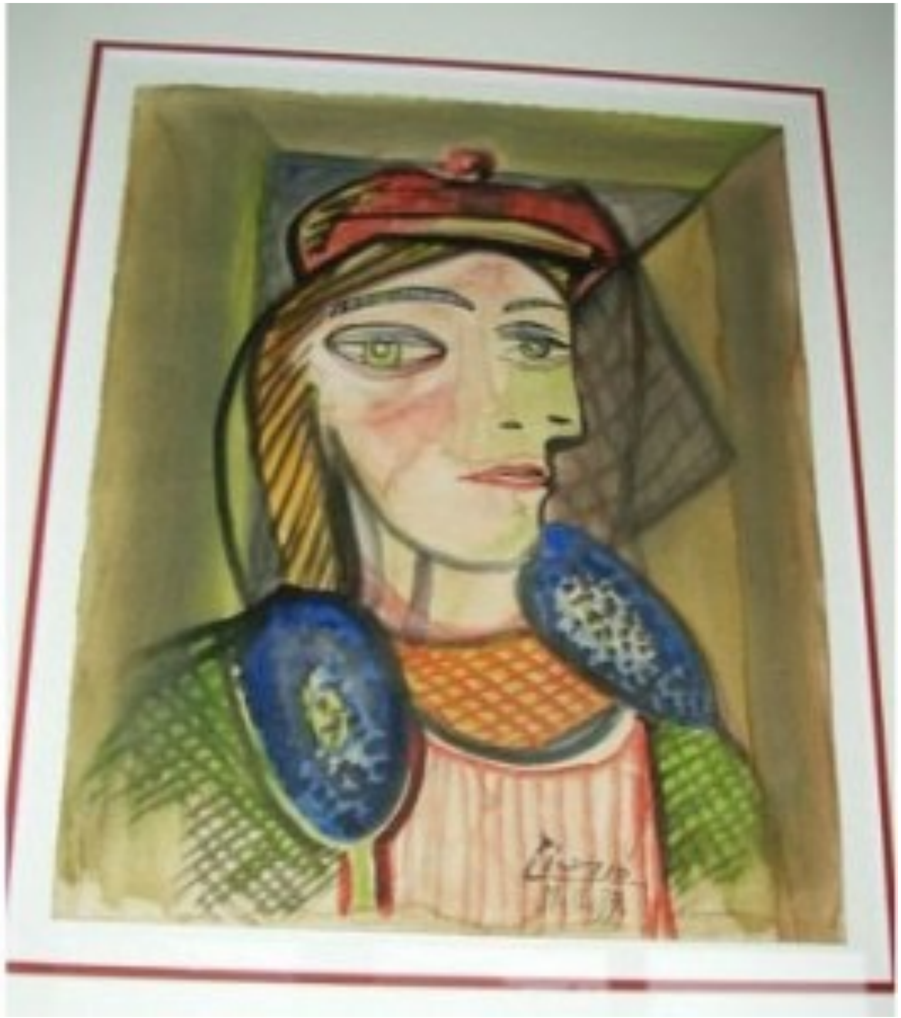


Which is “real” and which is “fake”?

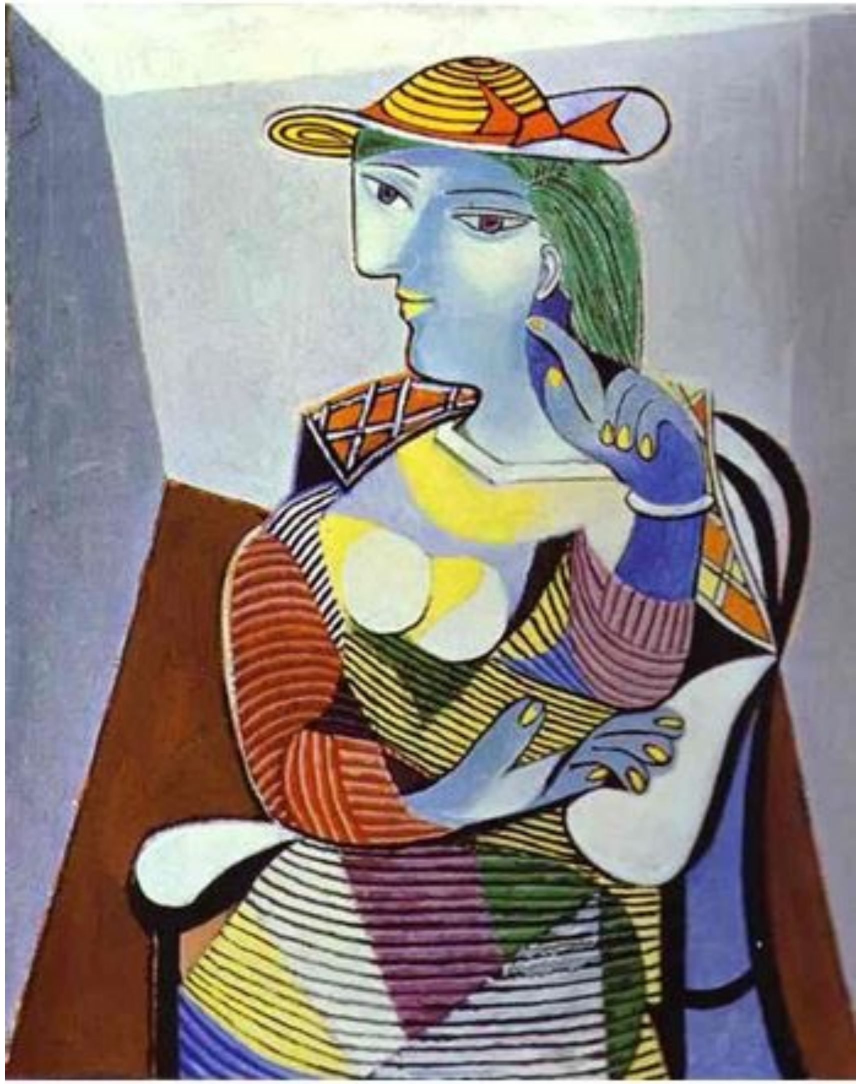


Centre de Estudios Borjanos/AFP/Getty Images

Idea behind GANs (Generative Adversarial Networks)

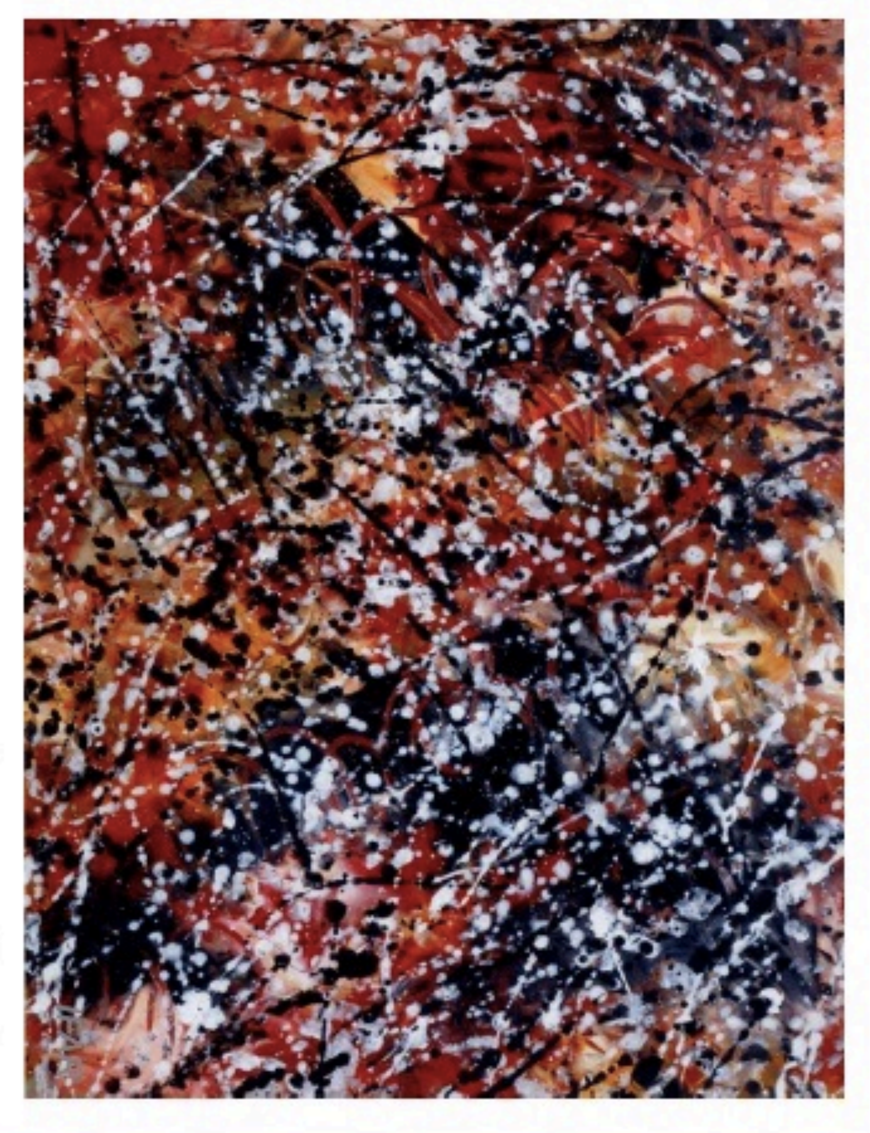


Which is "real" and which is "fake"?



https://webartacademy.com/fake-picasso

Idea behind GANs (Generative Adversarial Networks)



Which is “real” and which is “fake”?

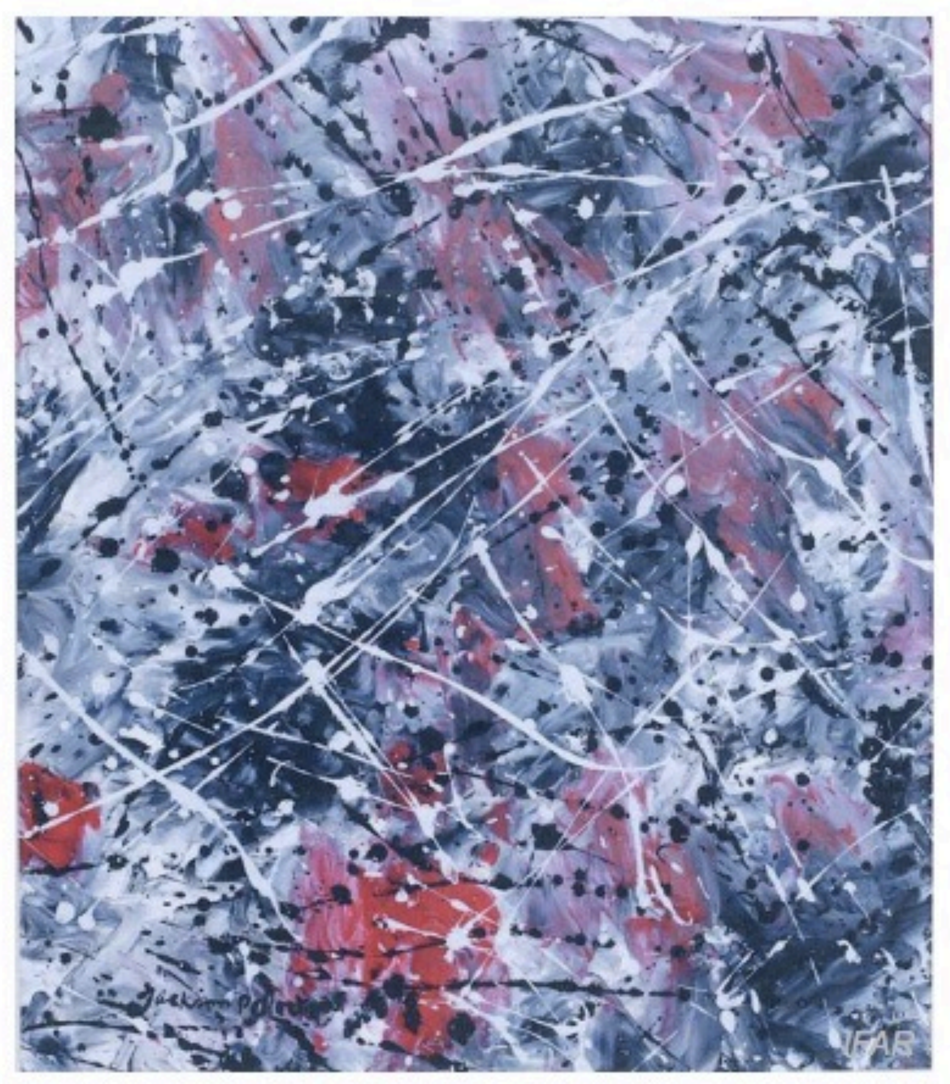
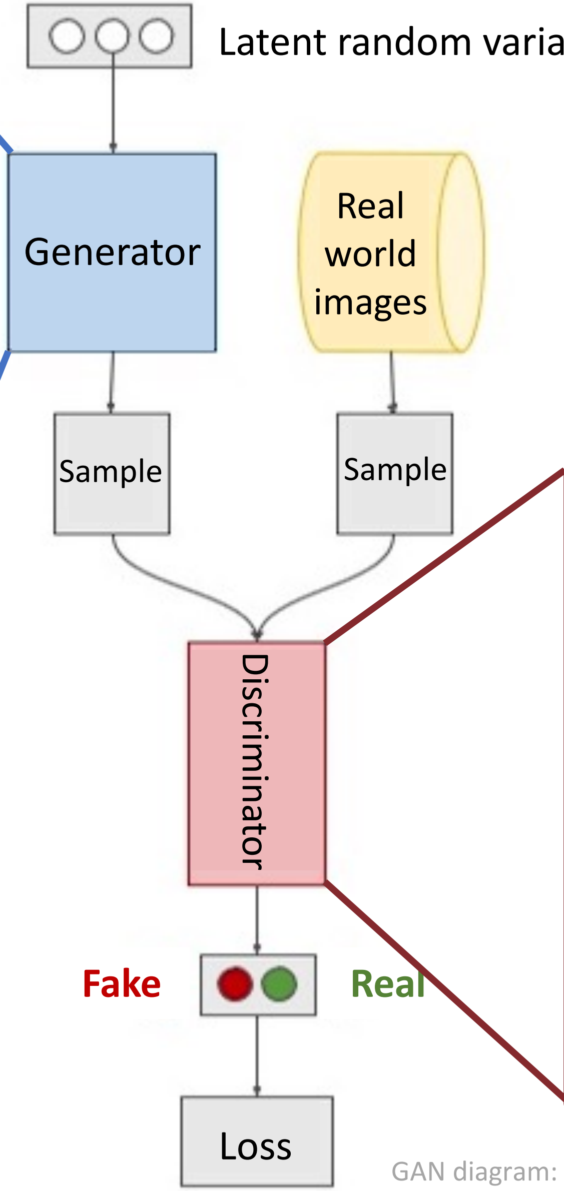


Photo: Courtesy International Foundation for Art Research (IFAR).

Idea behind GANs (Generative Adversarial Networks)

People Images / Getty Images

Generator ("forger") tries to create realistic artwork



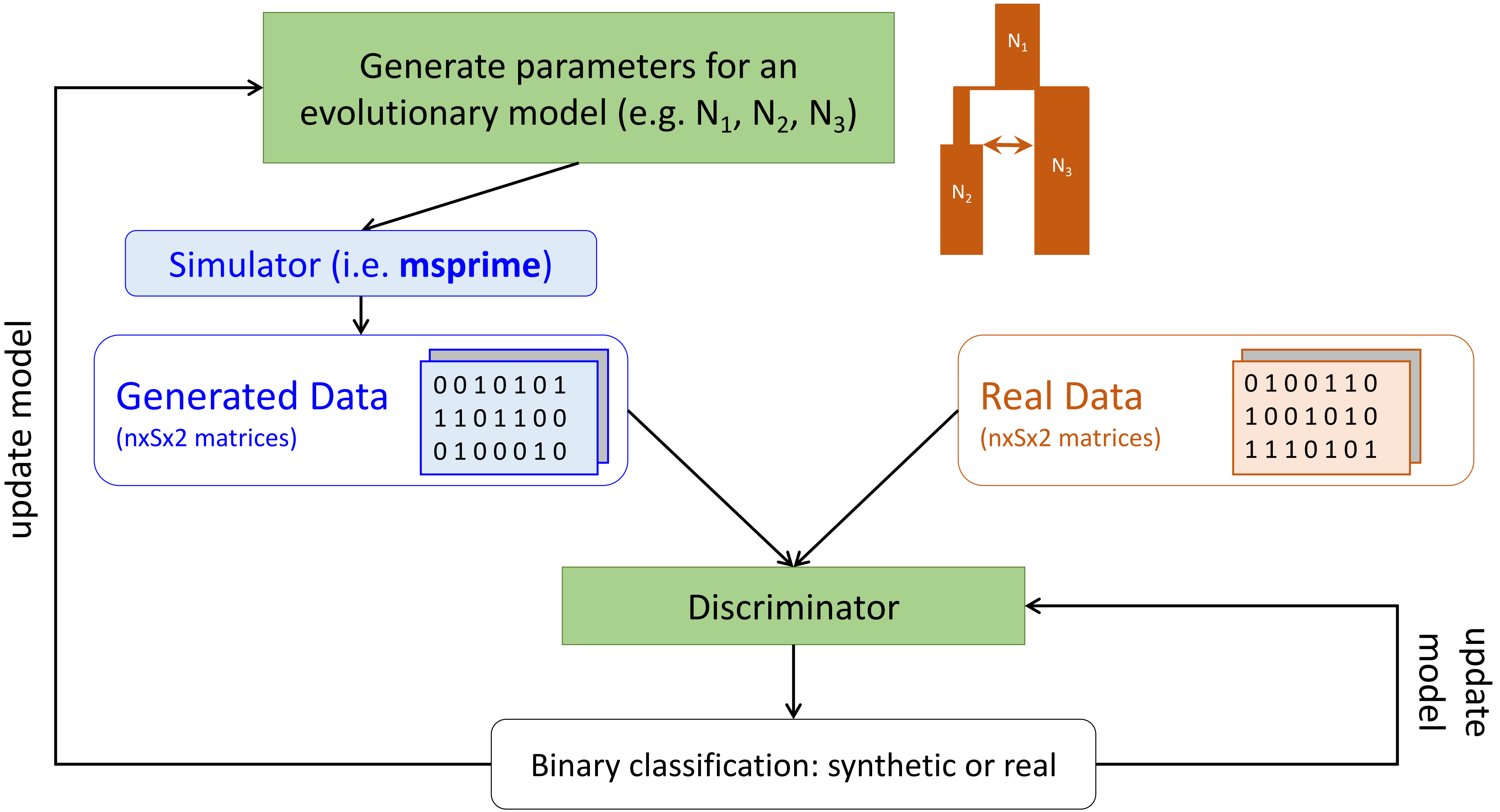
GAN diagram: Adapted from Kevin McGuinness

Discriminator ("art critic") tries to identify real vs. fake

(Source: Lost in the Louvre)

pg–gan algorithm overview

Wang, Wang, Kourakos, Hoang, Lee, Mathieson, Mathieson. "Automatic Inference of Demographic Parameters Using GANs". in press at *MER*, 2021 https://github.com/mathiesonlab/pg-gan



Generate parameters for an evolutionary model (e.g. N_1, N_2, N_3)

Simulator (i.e. **msprime**)

Generated Data

(nxSx2 matrices)

0 0 1 0 1 0 1
1 1 0 1 1 0 0
0 1 0 0 0 1 0

Real Data

(nxSx2 matrices)

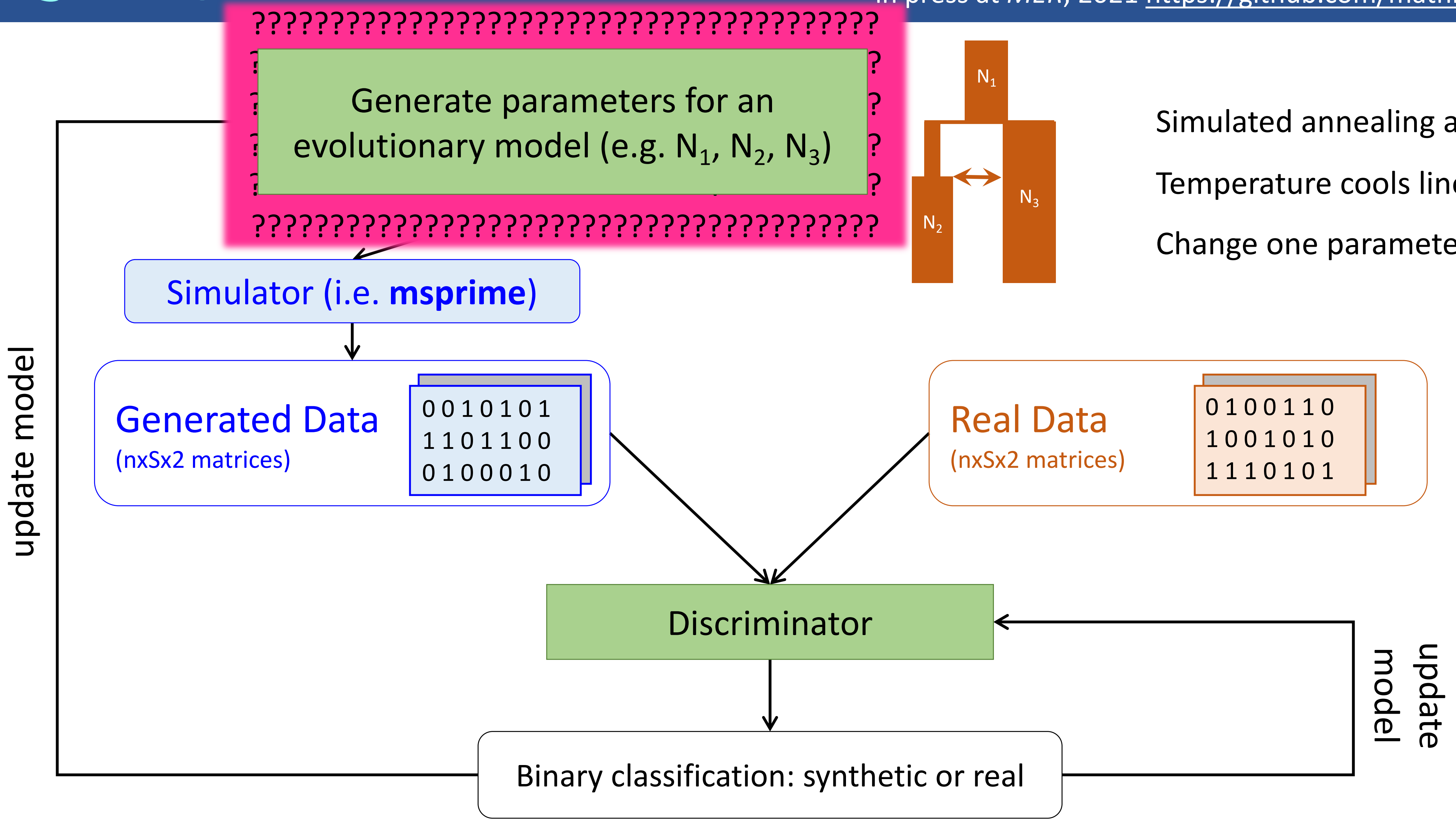
0 1 0 0 1 1 0
1 0 0 1 0 1 0
1 1 1 0 1 0 1

Discriminator

Binary classification: synthetic or real

pg-gan algorithm overview

Wang, Wang, Kourakos, Hoang, Lee, Mathieson, Mathieson. “Automatic Inference of Demographic Parameters Using GANs”. in press at *MER*, 2021 https://github.com/mathiesonlab/pg-gan



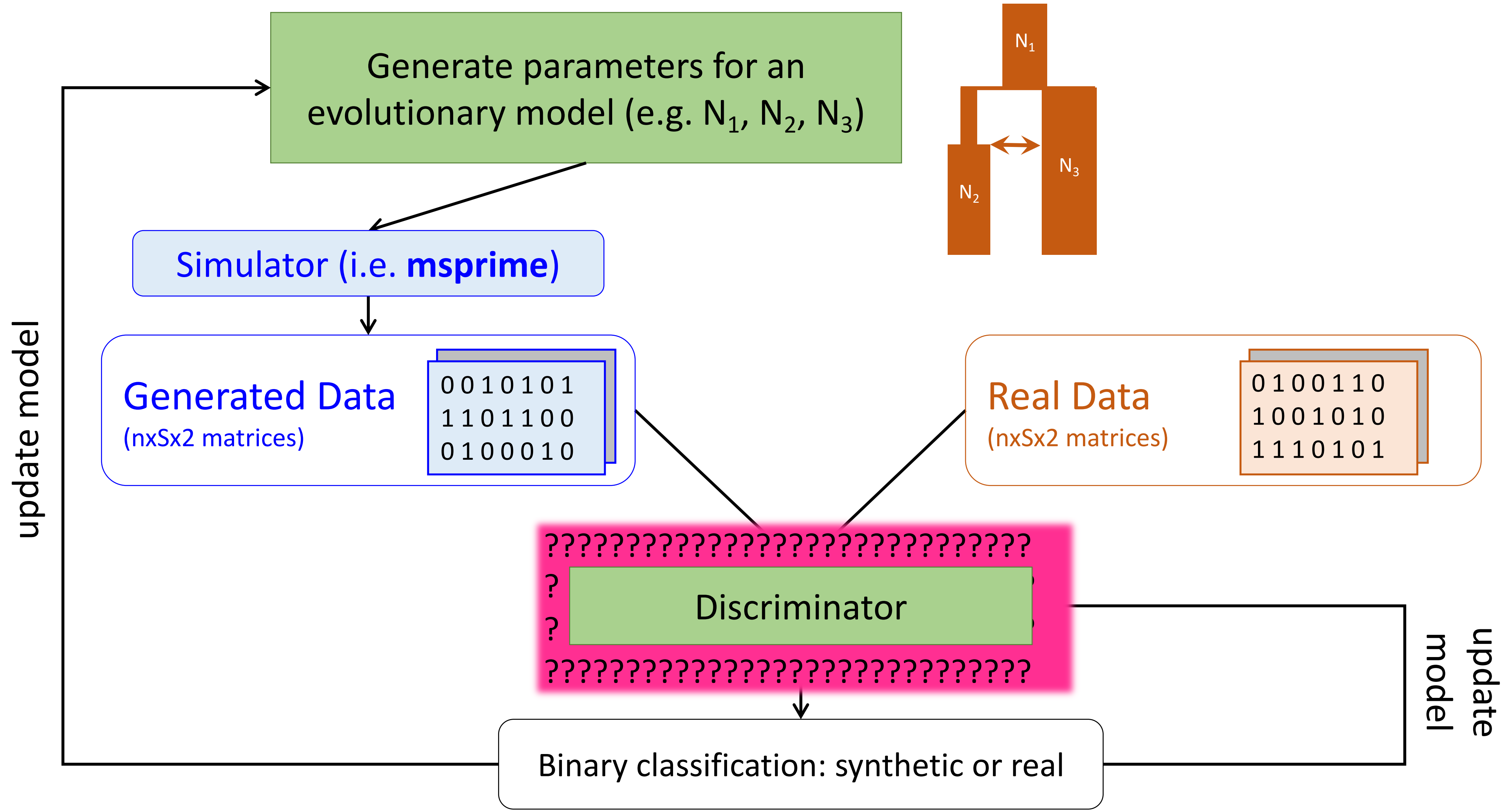
Simulated annealing algorithm

Temperature cools linearly

Change one parameter each iteration

pg-gan algorithm overview

Wang, Wang, Kourakos, Hoang, Lee, Mathieson, Mathieson. "Automatic Inference of Demographic Parameters Using GANs". in press at *MER*, 2021 https://github.com/mathiesonlab/pg-gan



Generate parameters for an evolutionary model (e.g. N_1, N_2, N_3)

Simulator (i.e. **msprime**)

Generated Data

(nxSx2 matrices)

0010101
1101100
0100010

Real Data

(nxSx2 matrices)

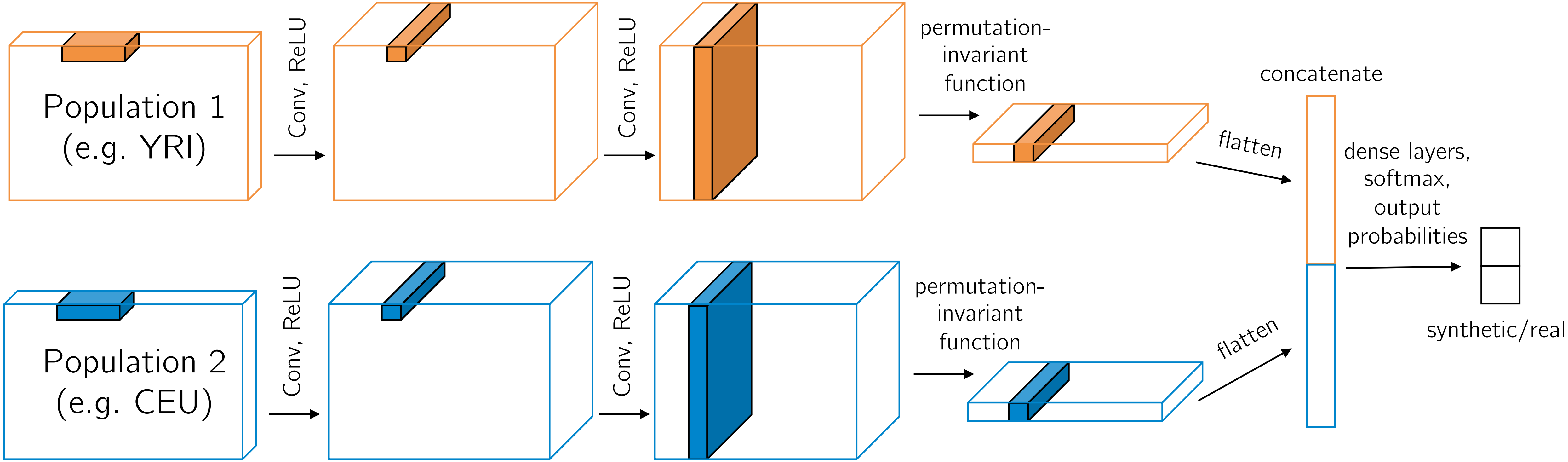
0100110
1001010
1110101

??????????????????????????????
?
?
??????????????????????????????

Discriminator

Binary classification: synthetic or real

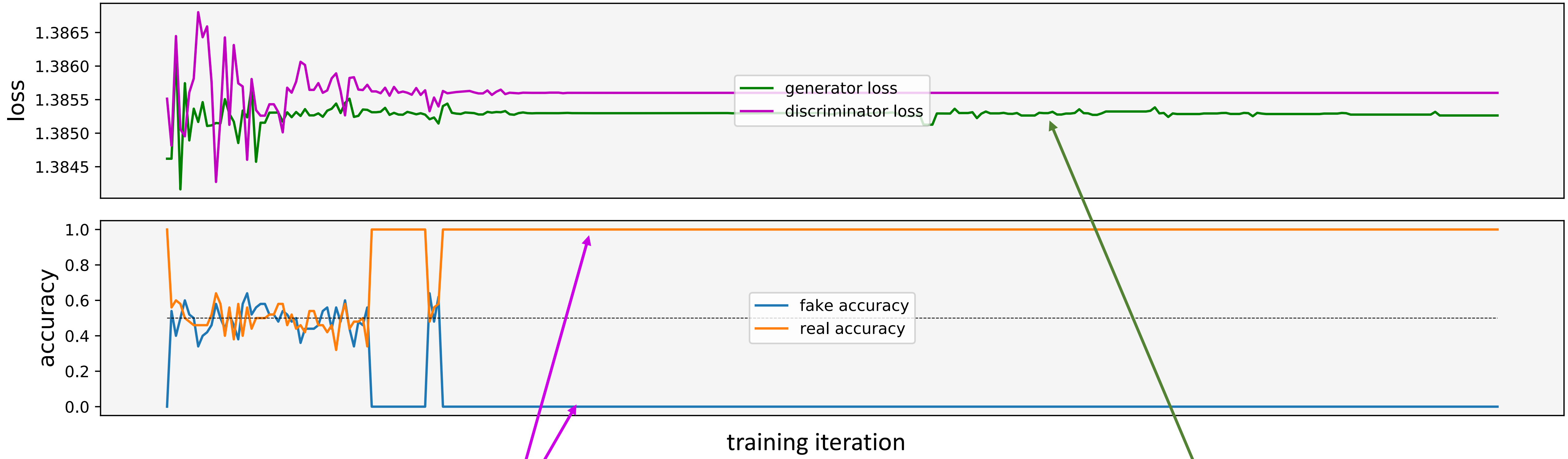
pg-gan discriminator architecture: extend to multiple populations



YRI: Yoruba in Ibadan, Nigeria

CEU: Utah residents with Northern and Western European ancestry

Example of failed GAN training



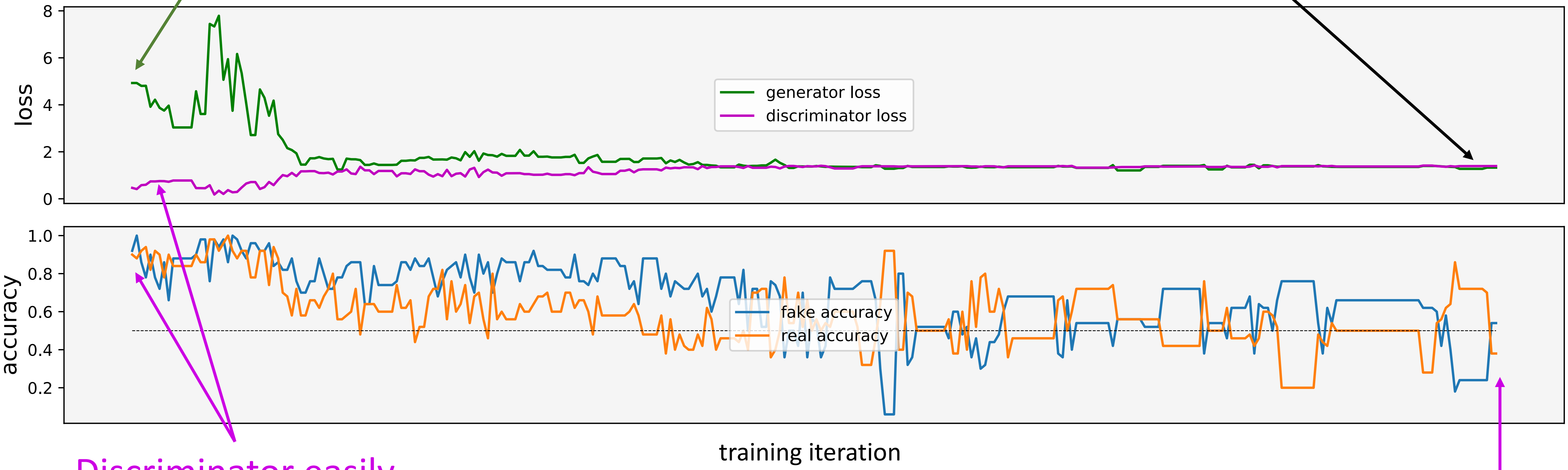
Discriminator classifies everything as real

Generator cannot learn and reduce loss

Example of successful GAN training

Generator not fooling discriminator

Generator and discriminator are balanced



Discriminator easily able to tell training from simulated

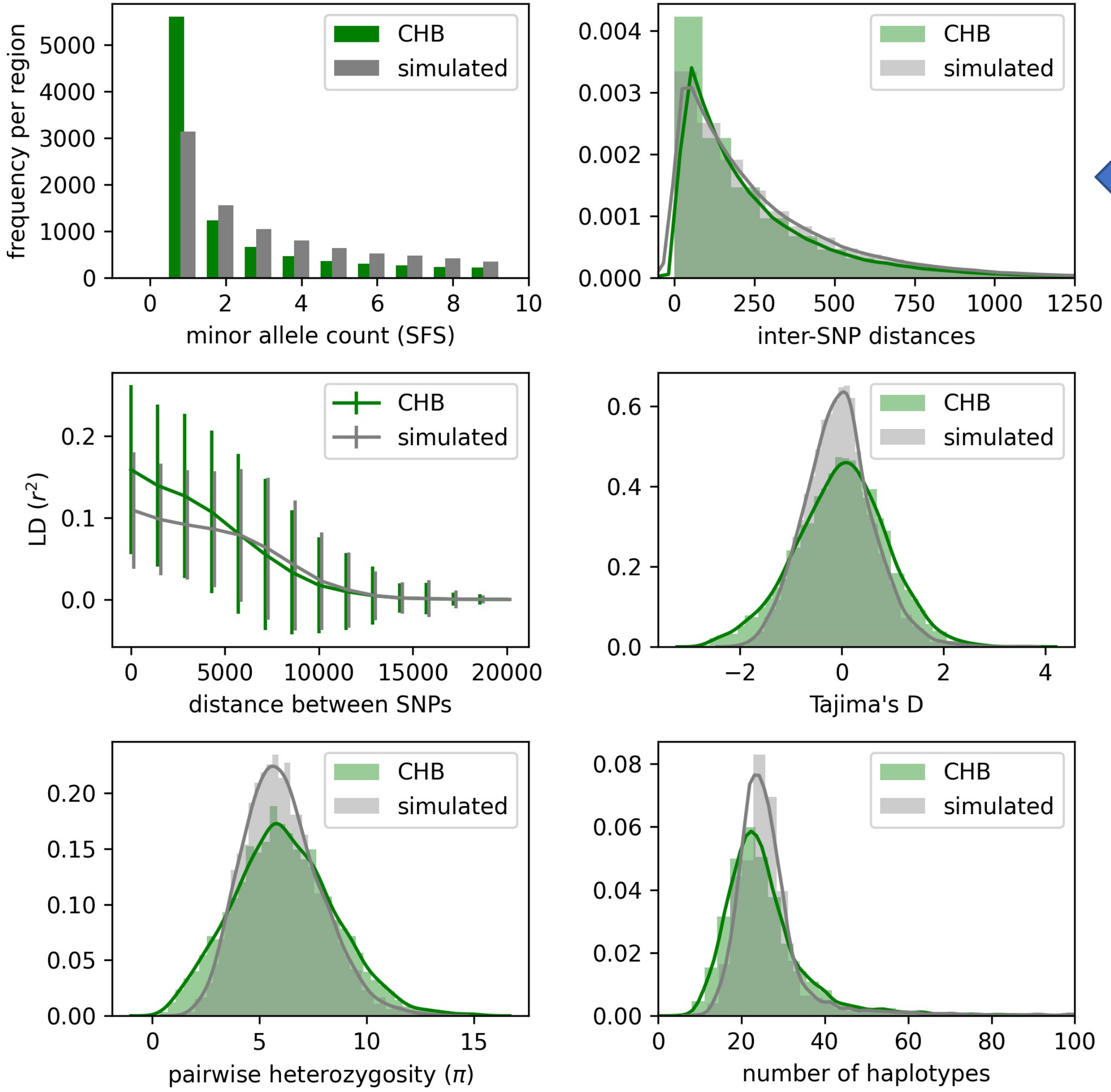
Discriminator is often confused

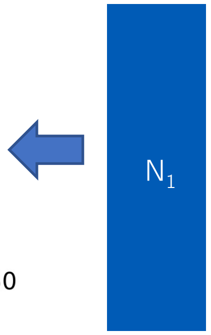
Outline

- Shift to machine learning in population genetics
- Shift away from summary statistics to “raw” data
- GANs and adversarial training
  - pg−gan algorithm for creating realistic simulated data
- Results on human data from Africa, Europe, and East Asia

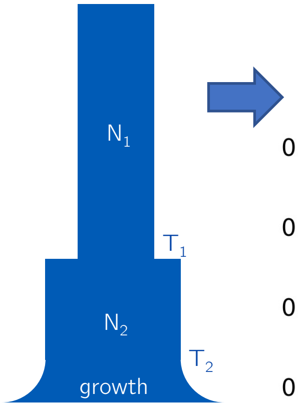
CHB: 1-param model

CHB: 5:param model

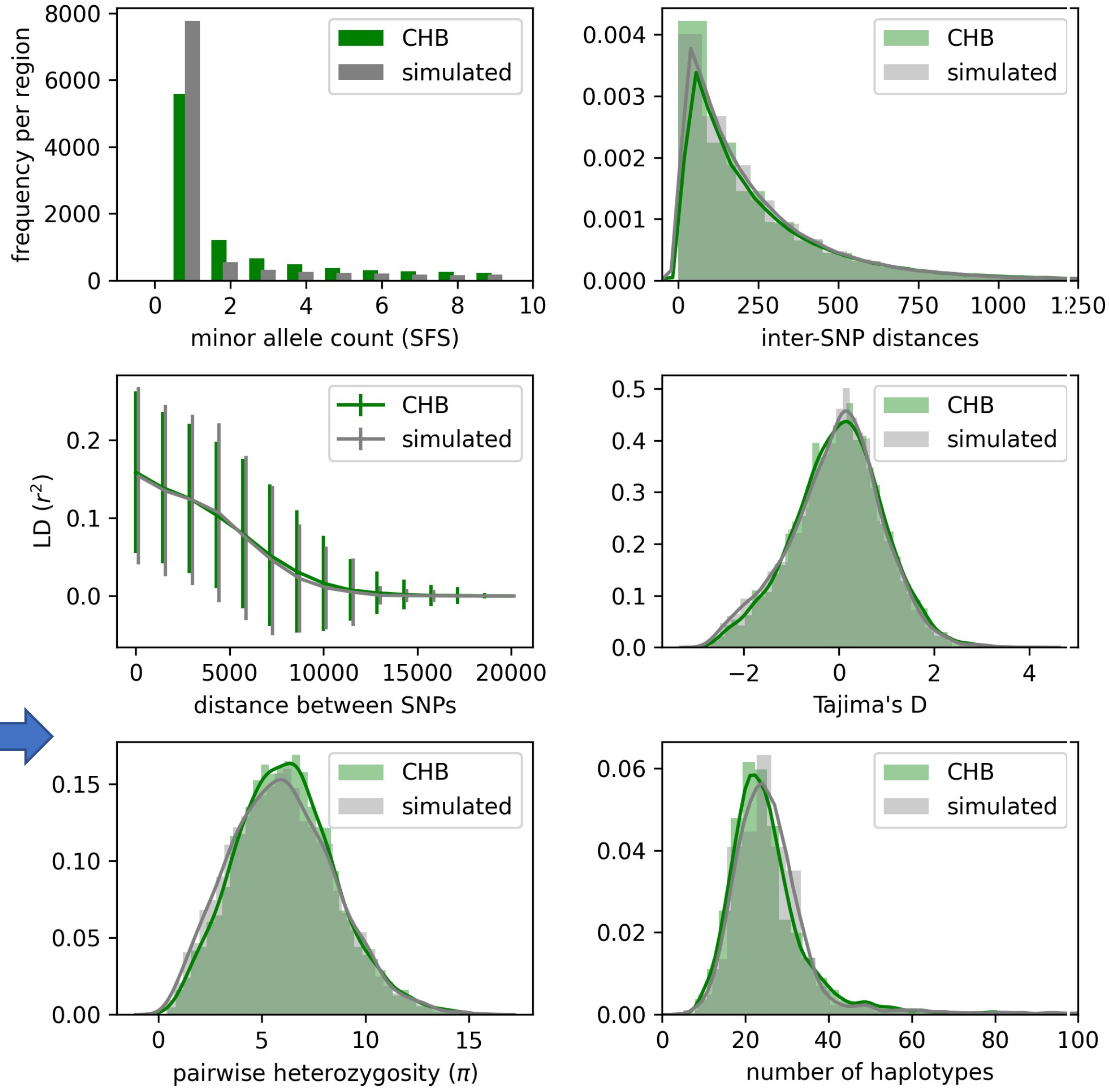




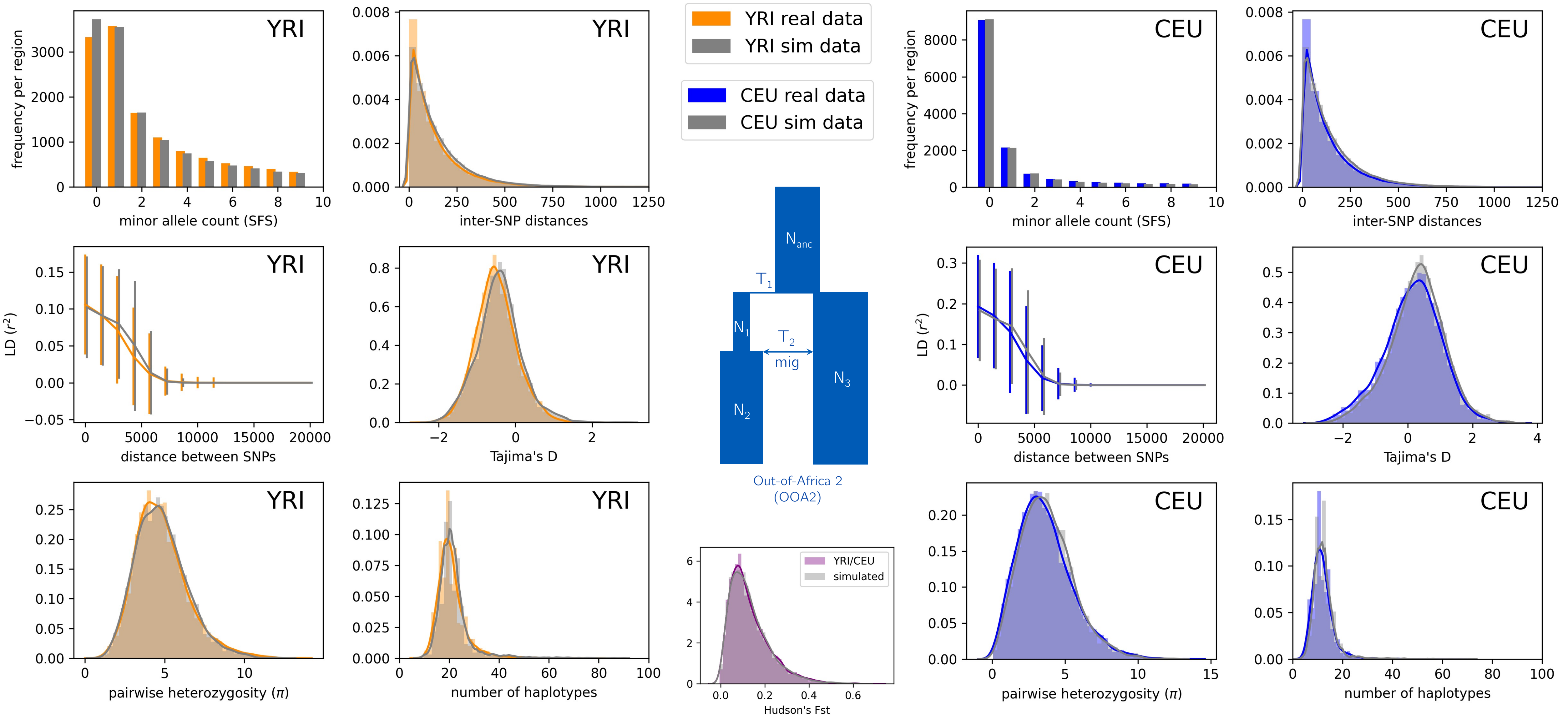
Constant



Single Population, growth (EXP)

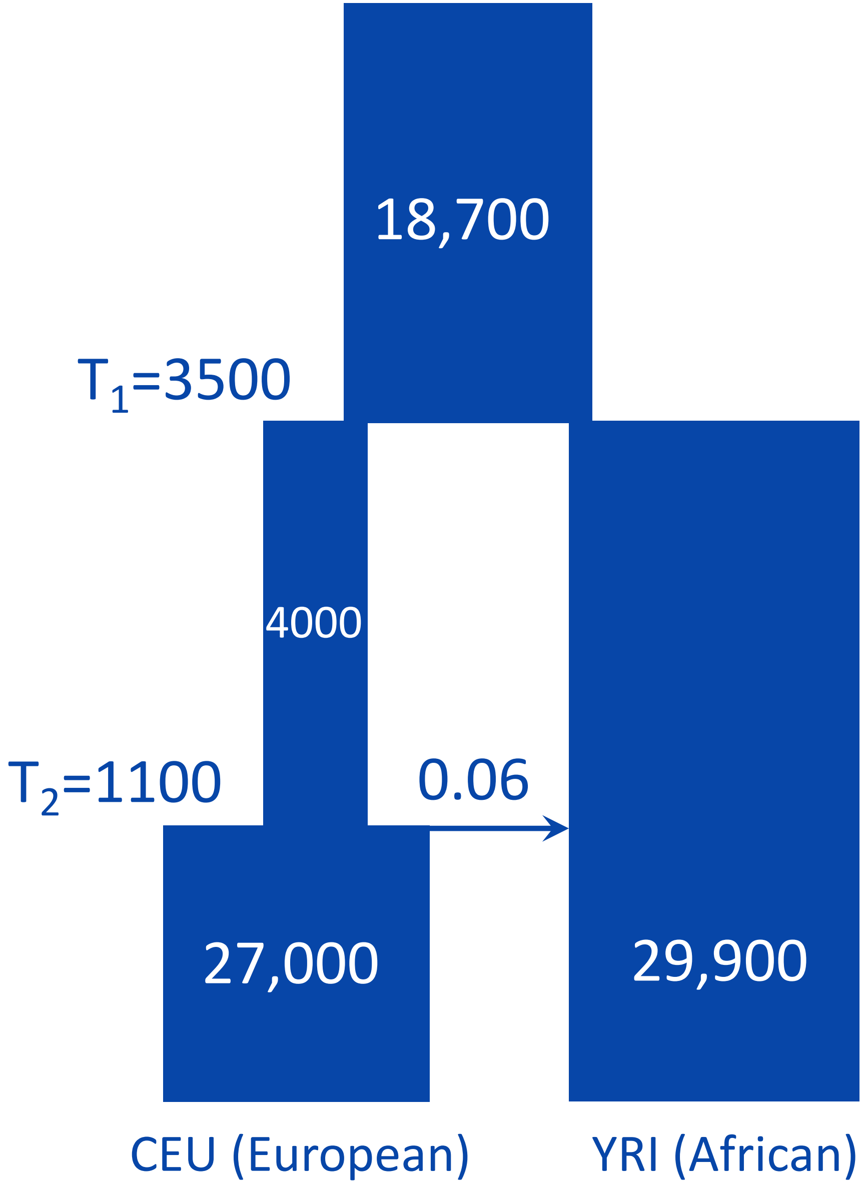


Simulated data under our GAN-inferred model matches real data



YRI/CEU split inference

- Time measured in generations
- Out-of-African bottleneck apparent



Conclusion for Machine Learning in Population Genetics

Future directions for `pg-gan`

- Apply to understudied populations
- Overcome data imbalance

Where are we going?

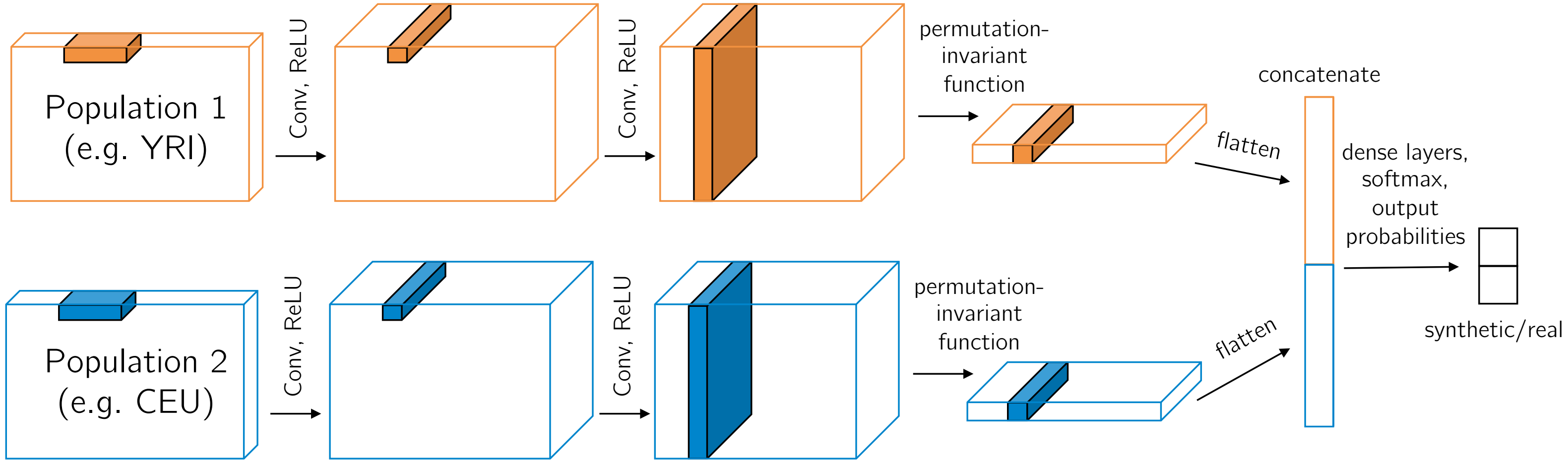
- Keep the data in mind
- ML methods need to be more interpretable
- Combine ML with evolutionary modeling
- Unsupervised learning

Thank you!

- Jeffrey Chan
- Nhung Hoang
- Paul Jenkins
- Michael Kourakos
- Hunter Lee
- Iain Mathieson
- Valerio Perrone
- Yun S. Song
- Jeffrey Spence
- Zhanpeng Wang
- Jiaping Wang

HAVERFORD COLLEGE

NIH



END