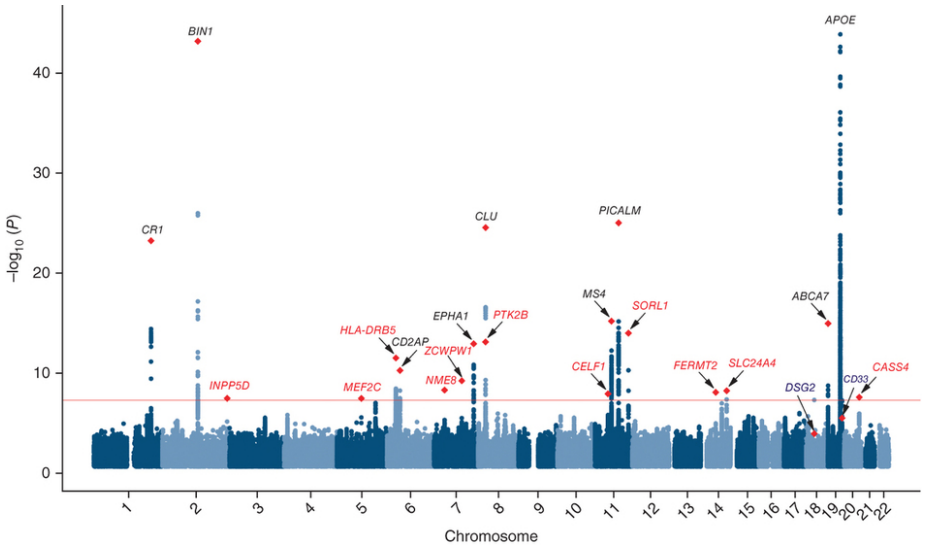
CS 68: Bioinformatics

Prof. Sara Mathieson

Spring 2018

Swarthmore College



Outline: April 25

- Lab 6 notes
- Finish Genome-Wide Association Studies (GWAS)
- Begin: machine learning for biology

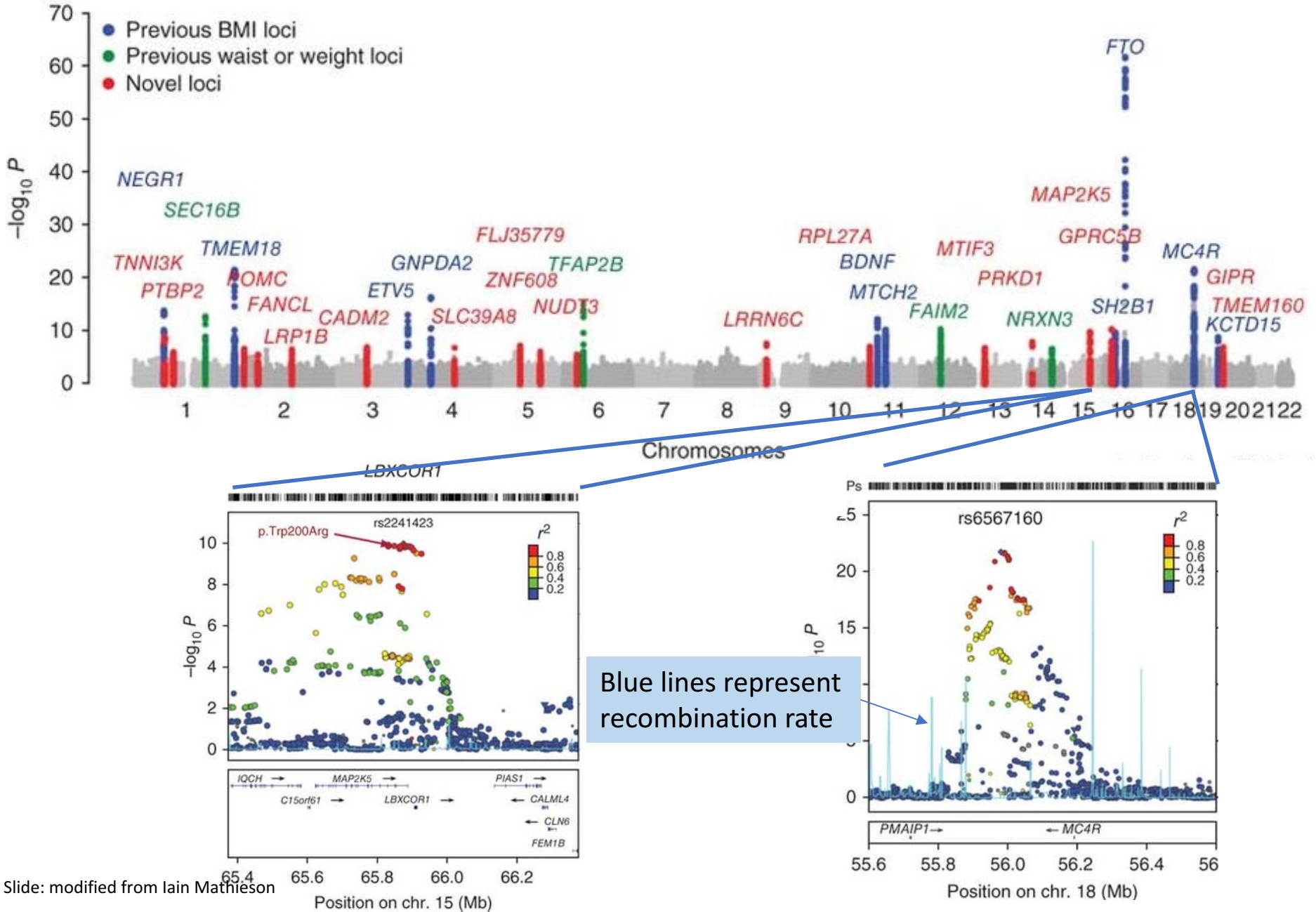
Notes:

- Hand back project proposals today
- Office hours TODAY 1-3pm
- Midterm 2 in-lab on Thursday (make/bring cheat-sheet)

Lab 6 Notes

- n = number of samples/sequences
- m = number of sites
- Runtime of naïve algorithm: $O(nm^2)$
  - Need to consider all pairs of sites => $O(m^2)$
  - Containment/disjoint linear in n by using a dictionary
- Runtime of Gusfield's algorithm: $O(nm)$
  - Each step (radix sort, transform rows, build trie) considers each entry in the matrix (n x m)
- Naïve is NOT exponentially faster than Gusfield! It is **quadratic** in m
- Recombination is the reason we don't expect a perfect phylogeny when considering many sites for samples from the same species

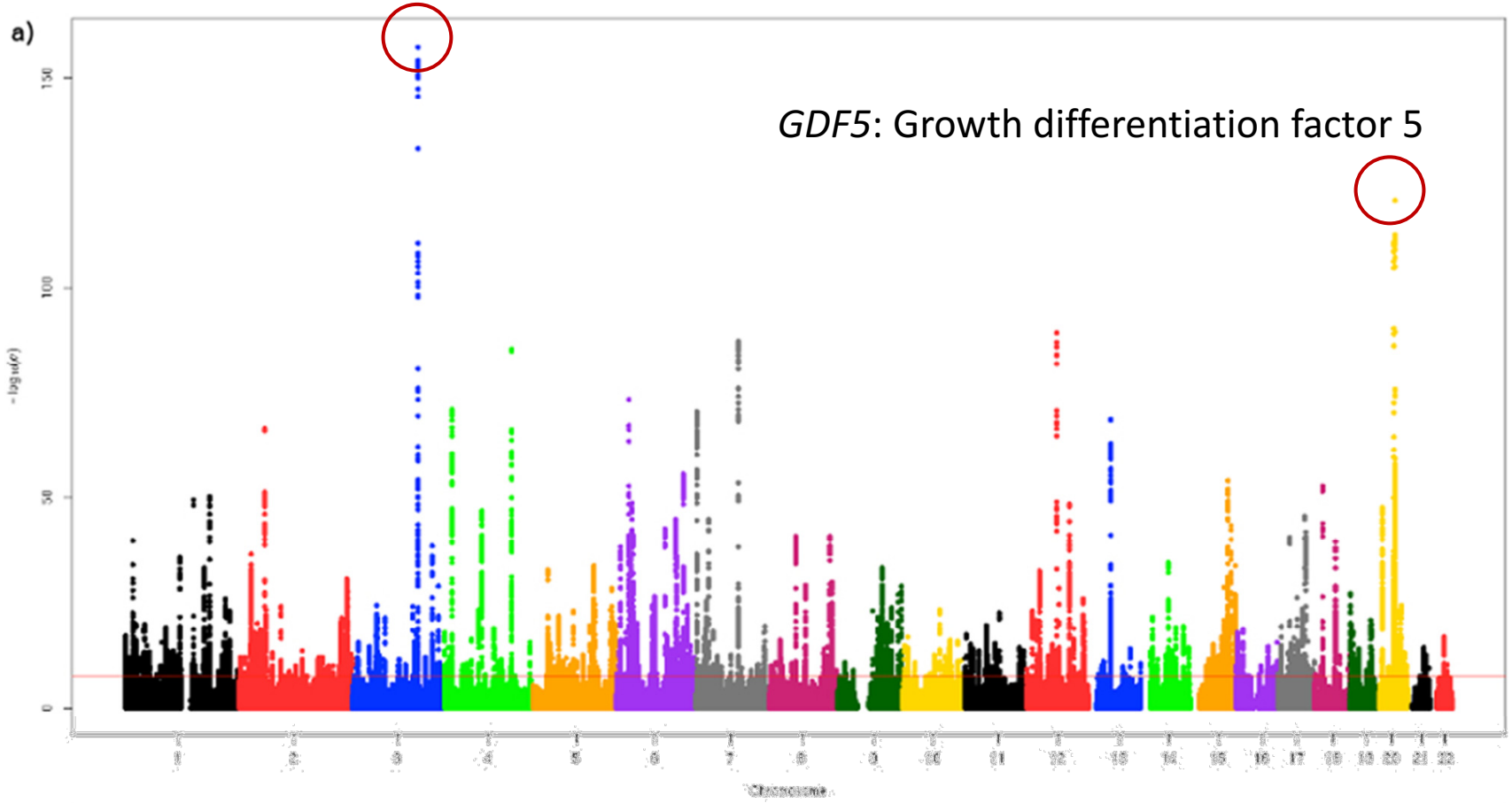
Fine-Mapping



Slide: modified from Iain Mathieson

Height GWAS

ZBTB38: Zinc Finger And BTB Domain Containing 38

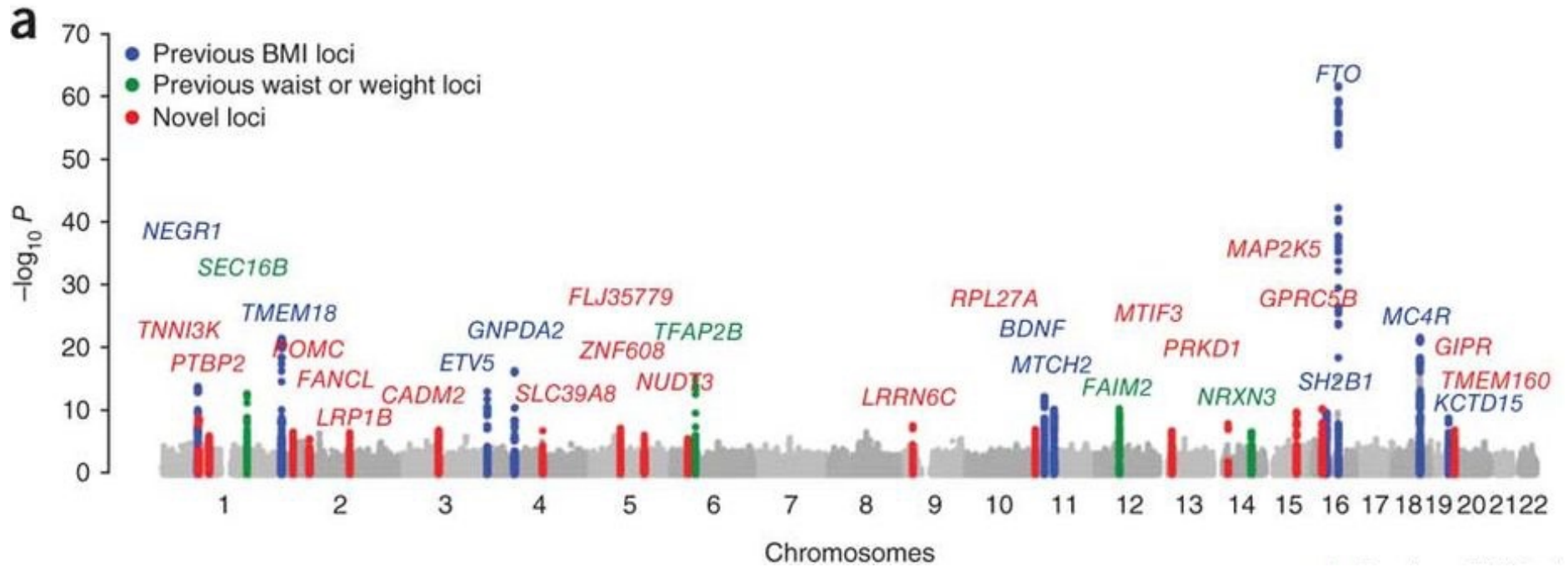


GDF5: Growth differentiation factor 5

697 independent SNPs significantly associated with height – Wood et al. 2014
Together explain about 15% of the phenotypic variance

Slide: modified from Iain Mathieson

BMI GWAS

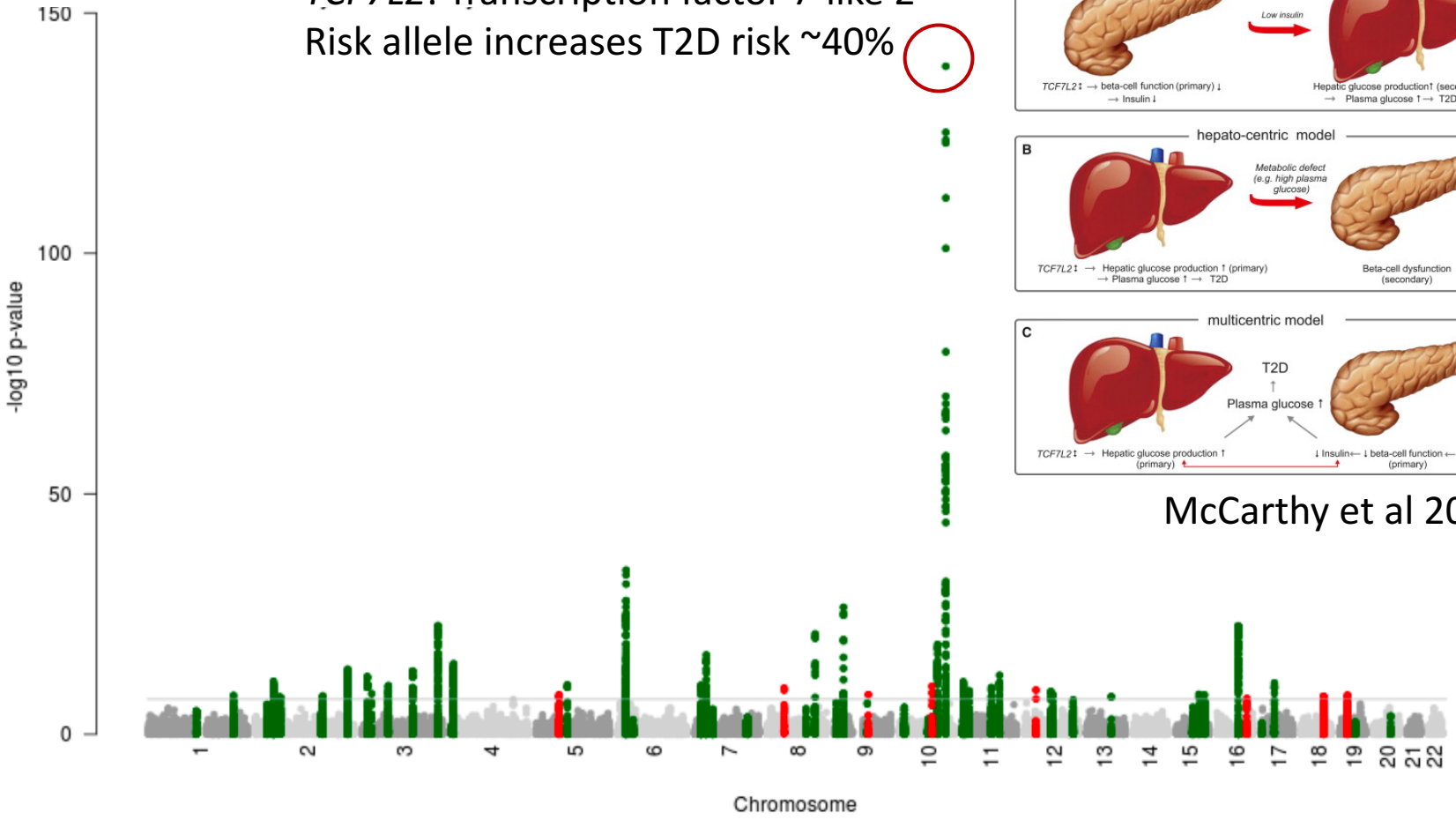


32 independent SNPs explain 1.45% of the variance in BMI – Speliotes et al. 2010

Slide: modified from Iain Mathieson

Type 2 Diabetes GWAS

TCF7L2: Transcription factor 7-like 2
Risk allele increases T2D risk ~40%

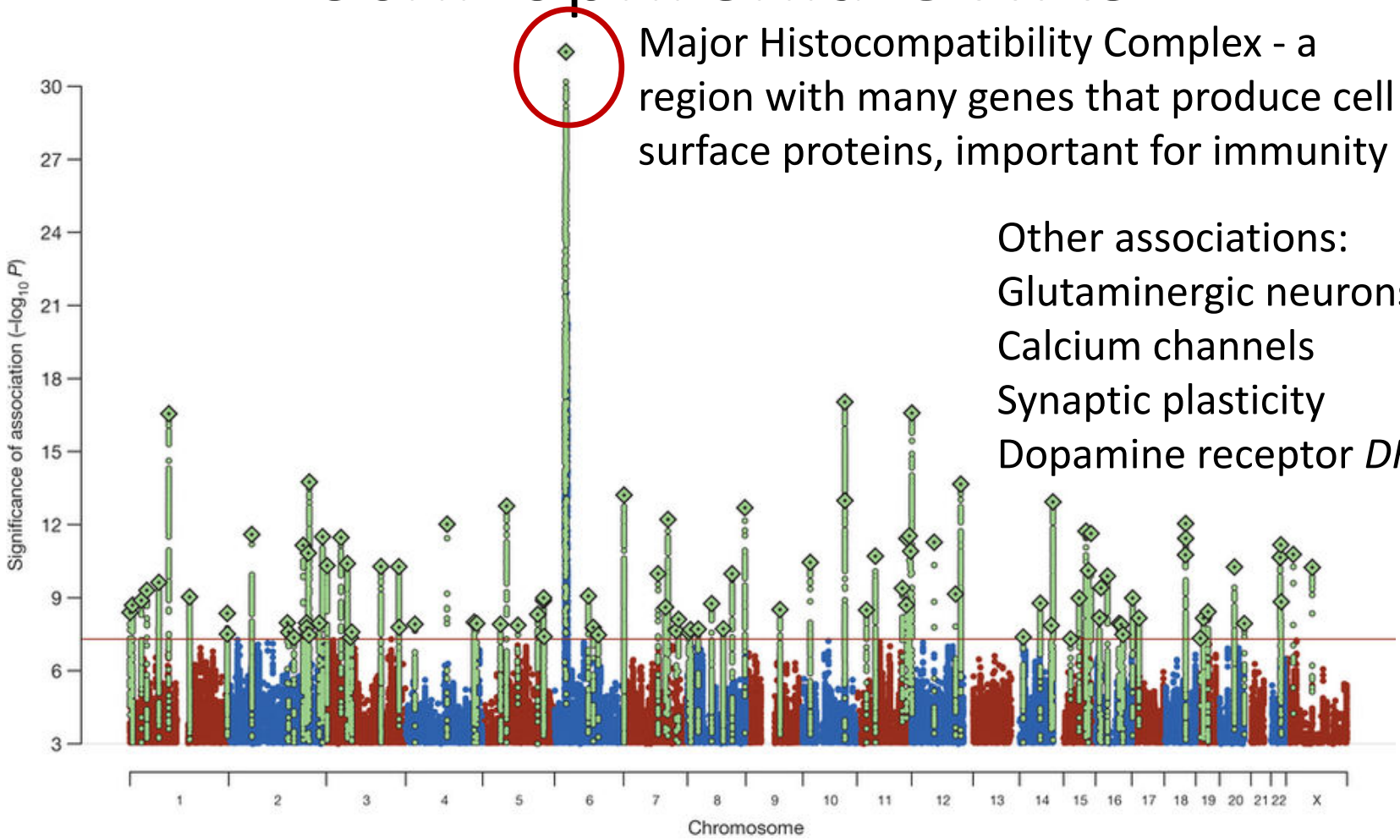


McCarthy et al 2013

63 independent loci explain 5.7% of the variance – Morris et al. 2012

Slide: modified from Iain Mathieson

Schizophrenia GWAS



108 independent loci explain 3.4% of the variance – Ripke et al. 2014

Slide: modified from Iain Mathieson

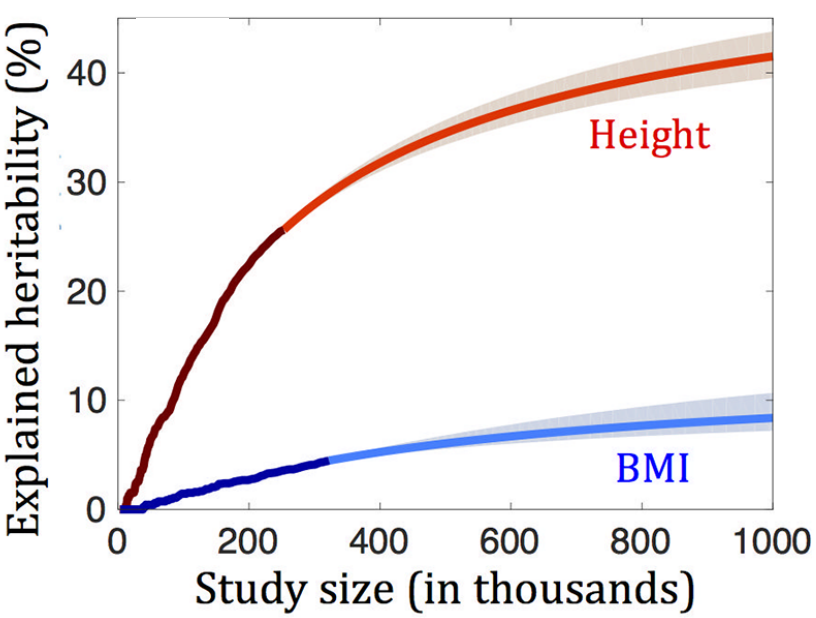
Missing Heritability?

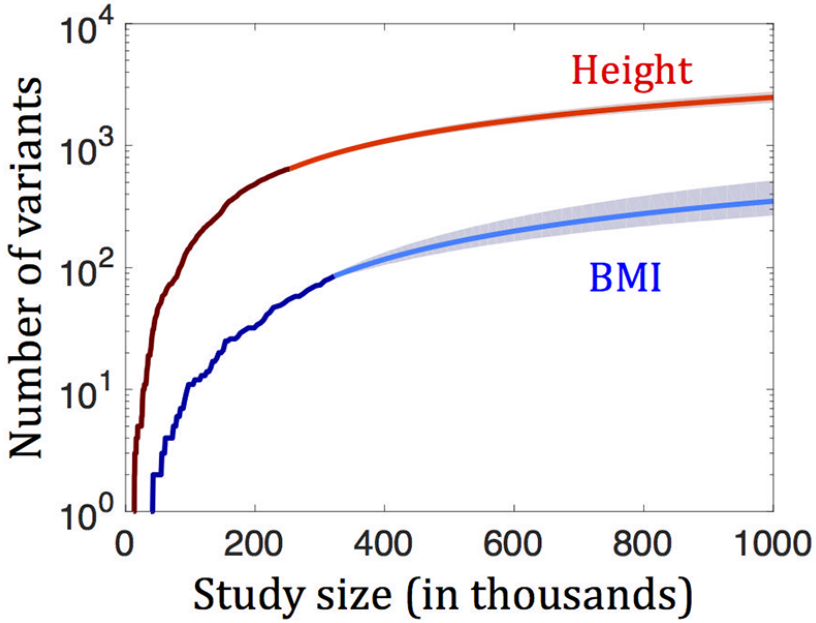
The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Nature 2008

The bigger the sample size, the more variants you find





Simons & Sella 2018

Slide: modified from Iain Mathieson

Missing Heritability?

"Missing heritability" is not really missing

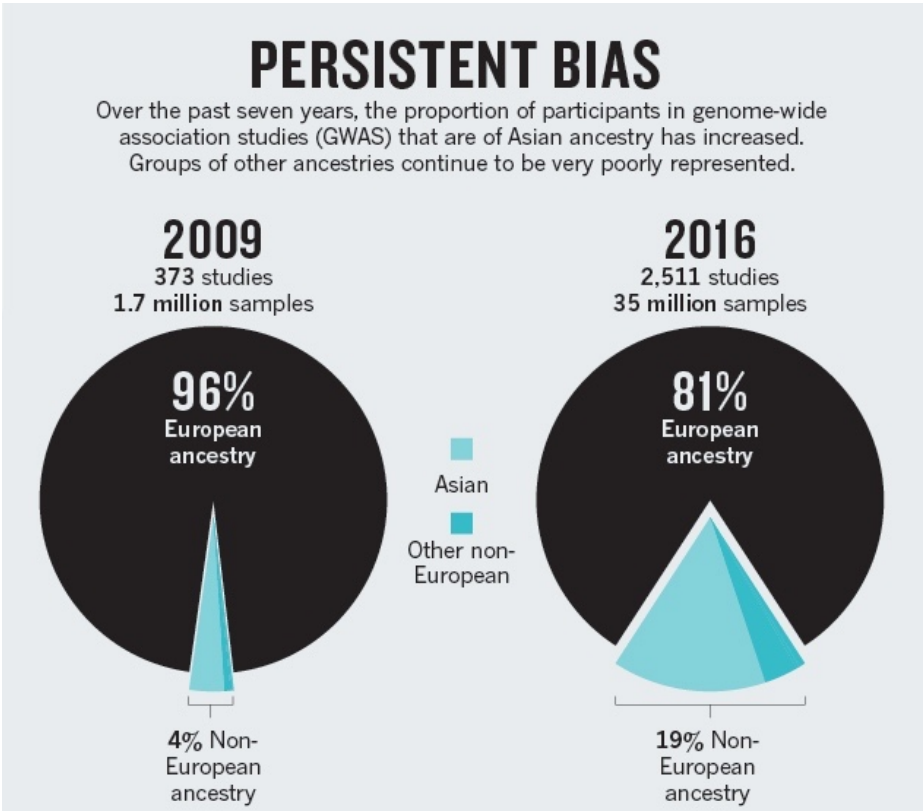
Mostly just hidden in very small effects that GWAS are not big enough to detect

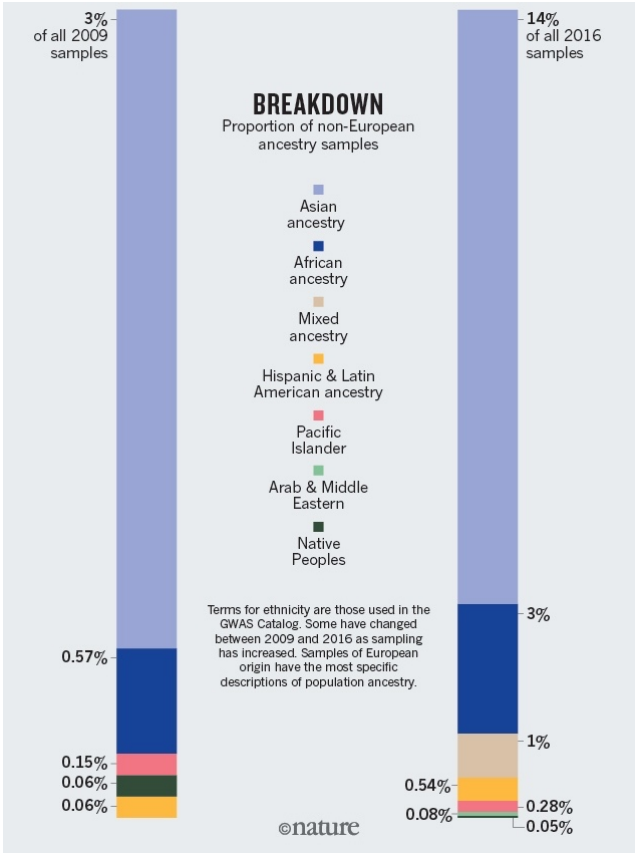
May be some hidden in epistatic effects or gene-environment interactions

Heritability estimates might be a bit too high

Slide: modified from Iain Mathieson

Almost all GWAS are carried out in European-Ancestry populations

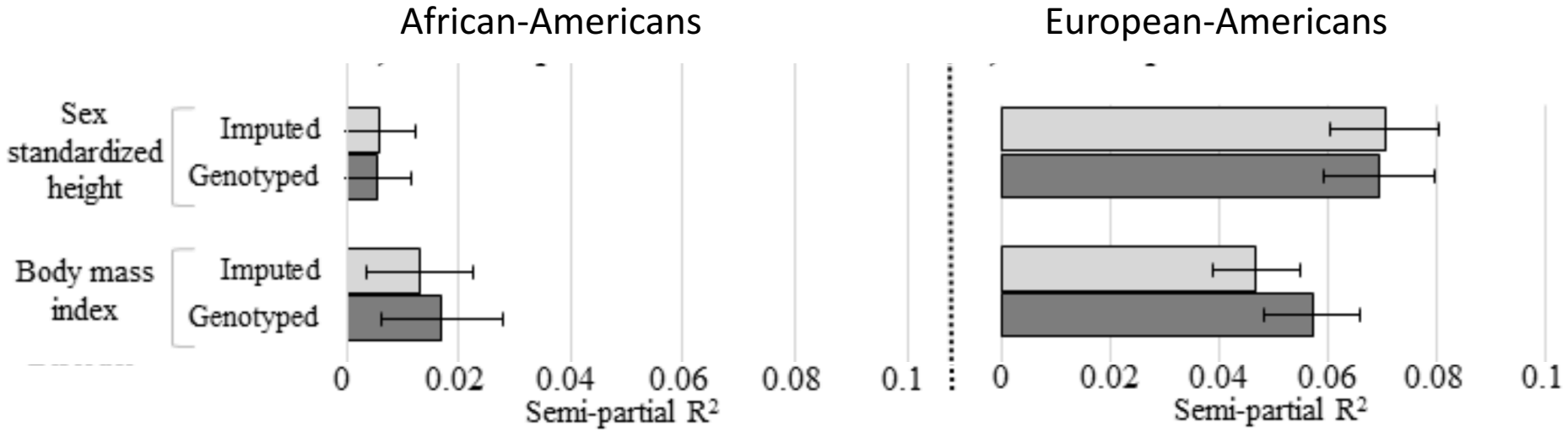




Popejoy & Fullerton 2016

Slide: modified from Iain Mathieson

European GWAS results do not translate to non-European ancestry populations



Ware et al 2018

Slide: modified from Iain Mathieson

How successful have GWAS been?

Twelve years.

Thousands of studies

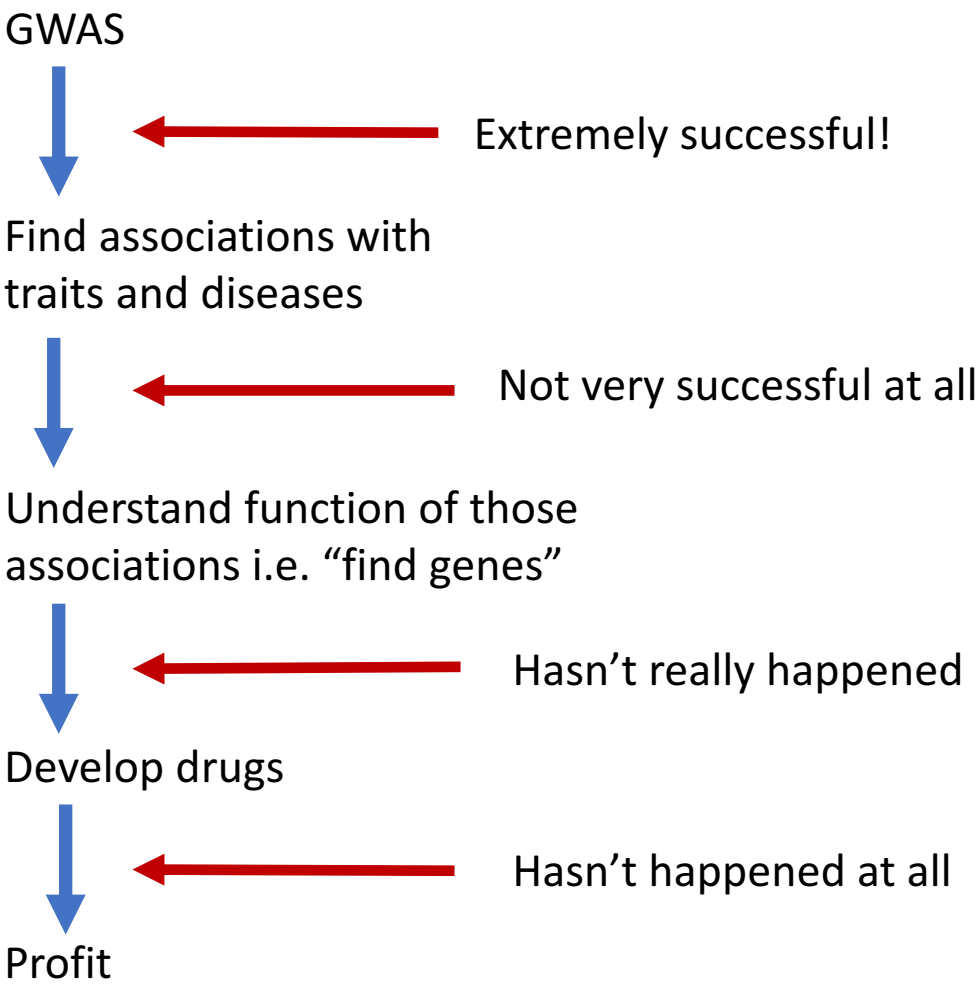
Tens of thousands of researchers

Tens of millions of patient-participants

Billions (?) of dollars

Slide: modified from Iain Mathieson

How successful have GWAS been?



New Opportunities

- Find connections between traits
- Predict genetic risk
- Understand complex trait evolution

Slide: modified from Iain Mathieson

Summary

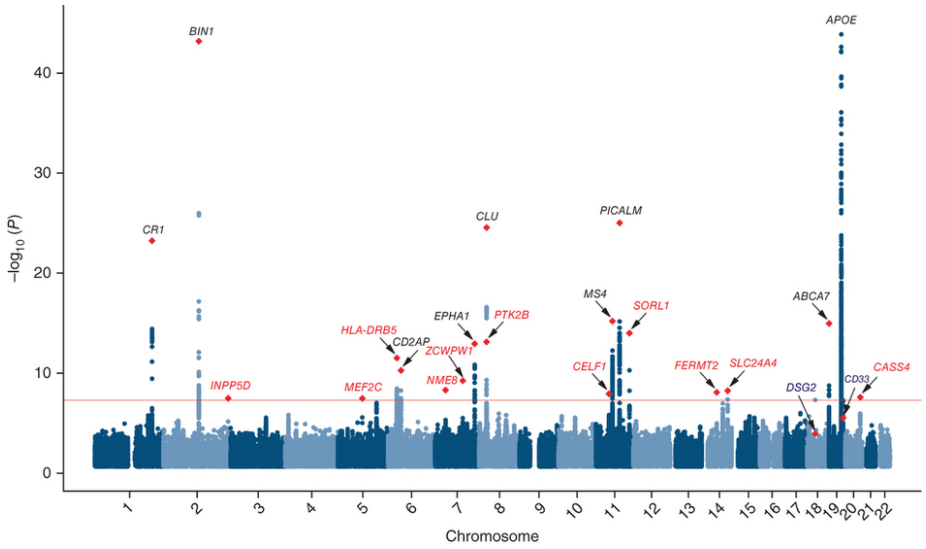
Genome-wide association studies:

Map common/low frequency variants associated with traits/disease

The bigger the sample size (more people) the smaller the effects you can detect

Do not tell us anything about function

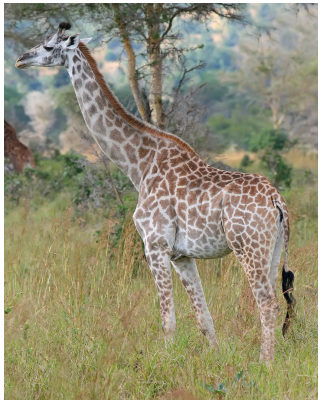
Need to be extremely careful about population structure and multiple testing



Slide: modified from Iain Mathieson

Machine Learning in Biology

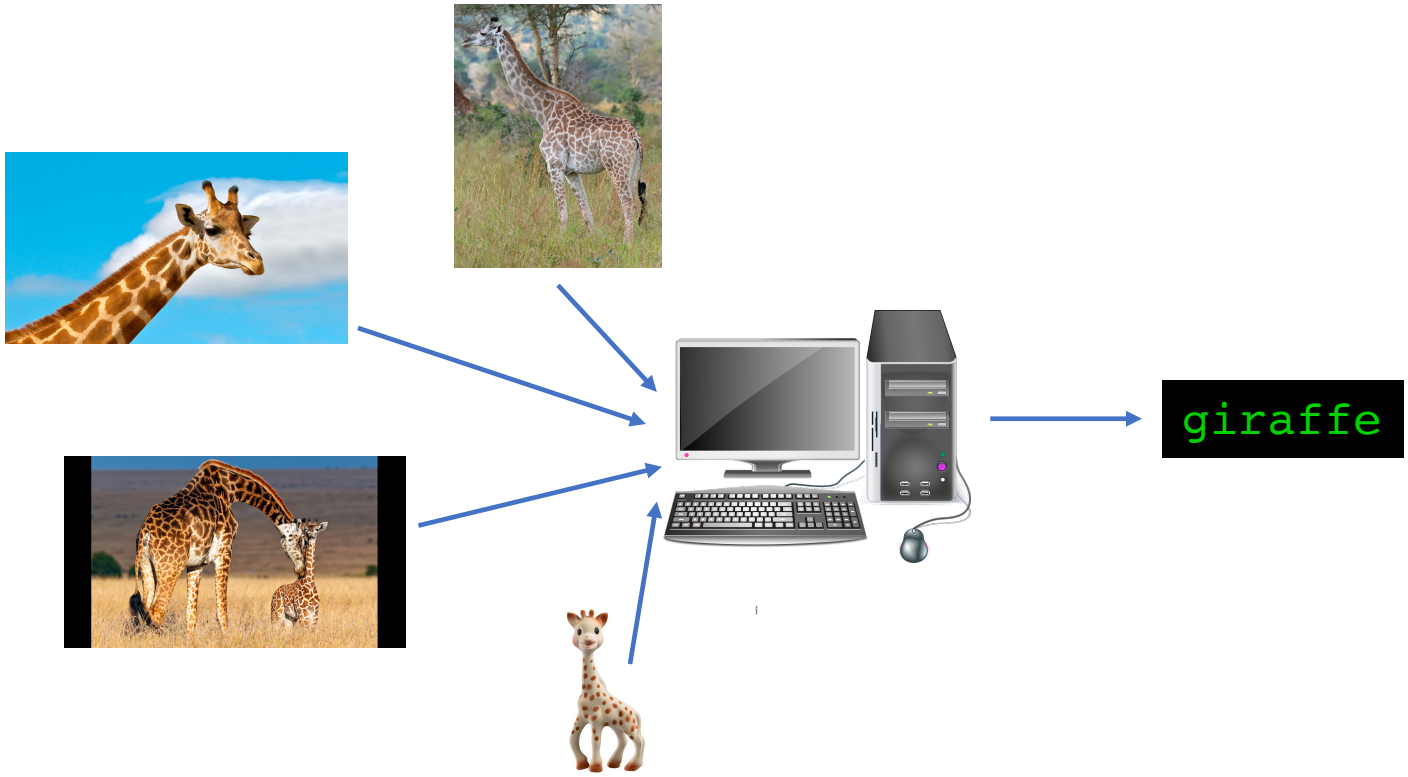
What is machine learning?



A child can see one giraffe and then be able to identify giraffes in many different contexts

Images: Wikipedia, San Diego Zoo, National Geographic, CNN.com

Can we train a computer to do the same thing?



Images: Wikipedia, San Diego Zoo, National Geographic, CNN.com

Can we train a computer to do the same thing?

How can we *distinguish* between similar objects?

Images: veriy.com

What is machine learning?

- One flavor of machine learning is *classification*

- Goal: separate examples into (many) different *classes*

`Example: bagel vs. dog`

Images: veriy.com

Why do we care?

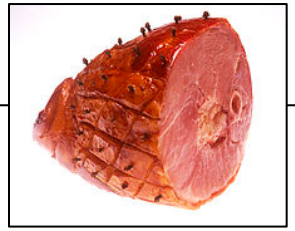
- Email filtering (spam vs. not-spam)

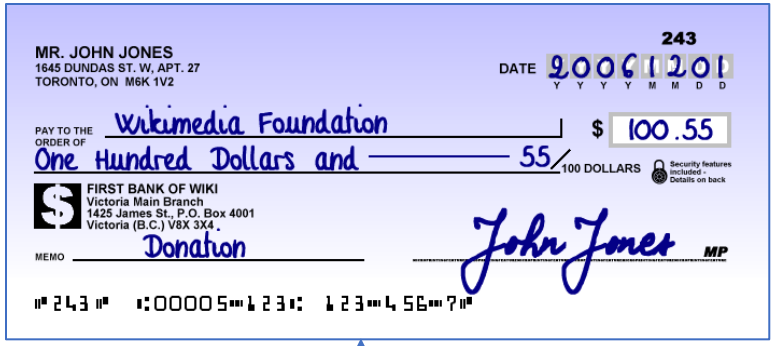
From: cheapsales@buystufffromme.com
To: ang@cs.stanford.edu
Subject: Buy now!

Deal of the week! Buy now!
Rolex w4tchs - $100
Medicine (any kind) - $50
Also low cost M0rgages
available.

From: Alfred Ng
To: ang@cs.stanford.edu
Subject: Christmas dates?

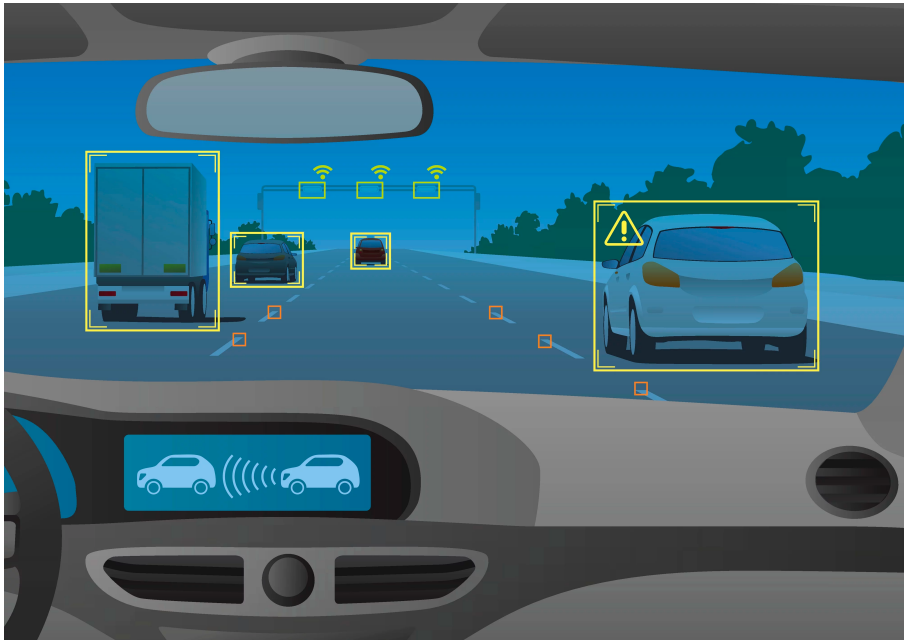
Hey Andrew,
Was talking to Mom about plans
for Xmas. When do you get off
work. Meet Dec 22?
Alf





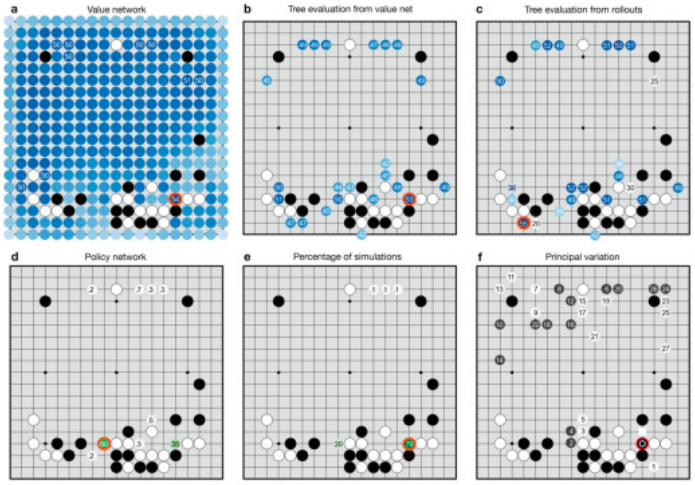
- Handwriting recognition (digits in a check)

Why do we care?



Self-driving cars are in our present and future

AlphaGo: plays humans never thought of

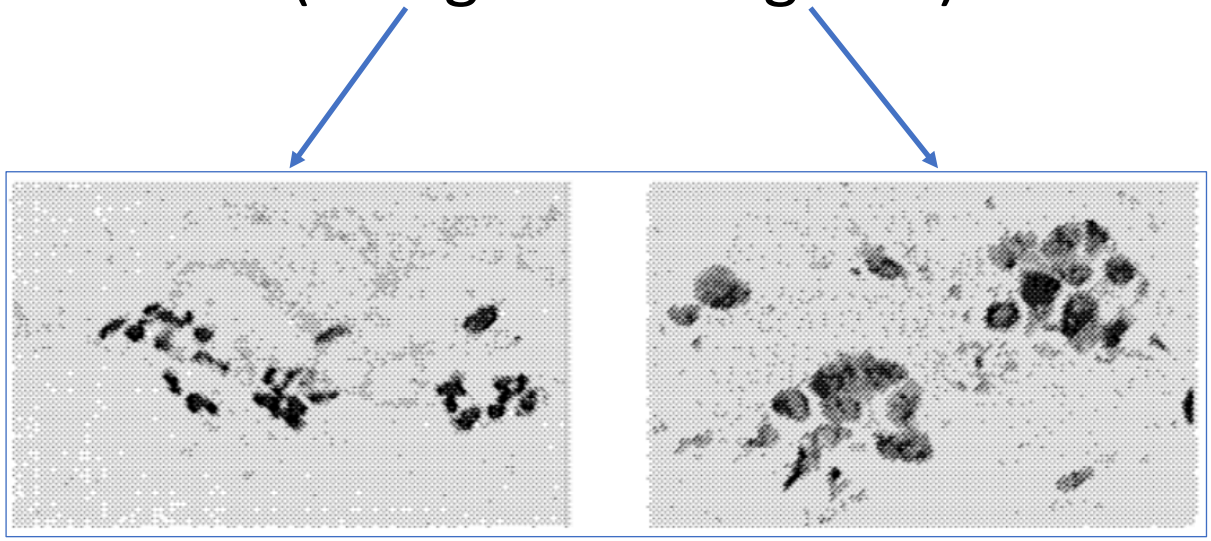


AlphaGo

Images: Scientific American

Why do we care?

- Tumor detection (benign vs. malignant)



"On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset"

ML and “Big Data”

- As datasets become larger and more complex, humans can no longer make sense of them without machines

- Machine learning is in all of our lives and understanding it will be increasingly valuable

Machine learning terminology

- *Training*: usually involves the program processing many *examples* (from different classes) where we know the "answer" or *label*, and learning how to separate them

- *Testing*: program classifies new examples

Machine learning terminology

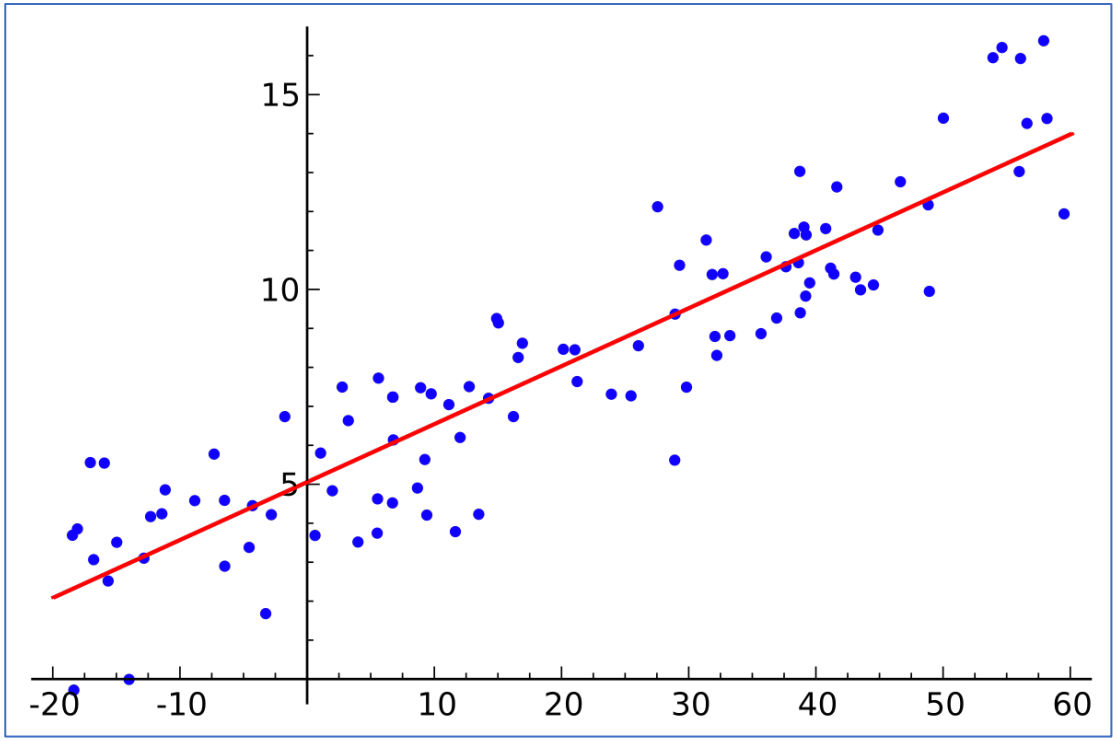
- *Supervised learning:* a human (usually) has hand-labeled the training examples, so it's easier for the computer to learn differences

- *Unsupervised learning:* data is unlabeled (no class information)

Machine Learning Methods

Regression

Training data: vectors ***x*** (independent variable) and ***y*** (dependent variable)

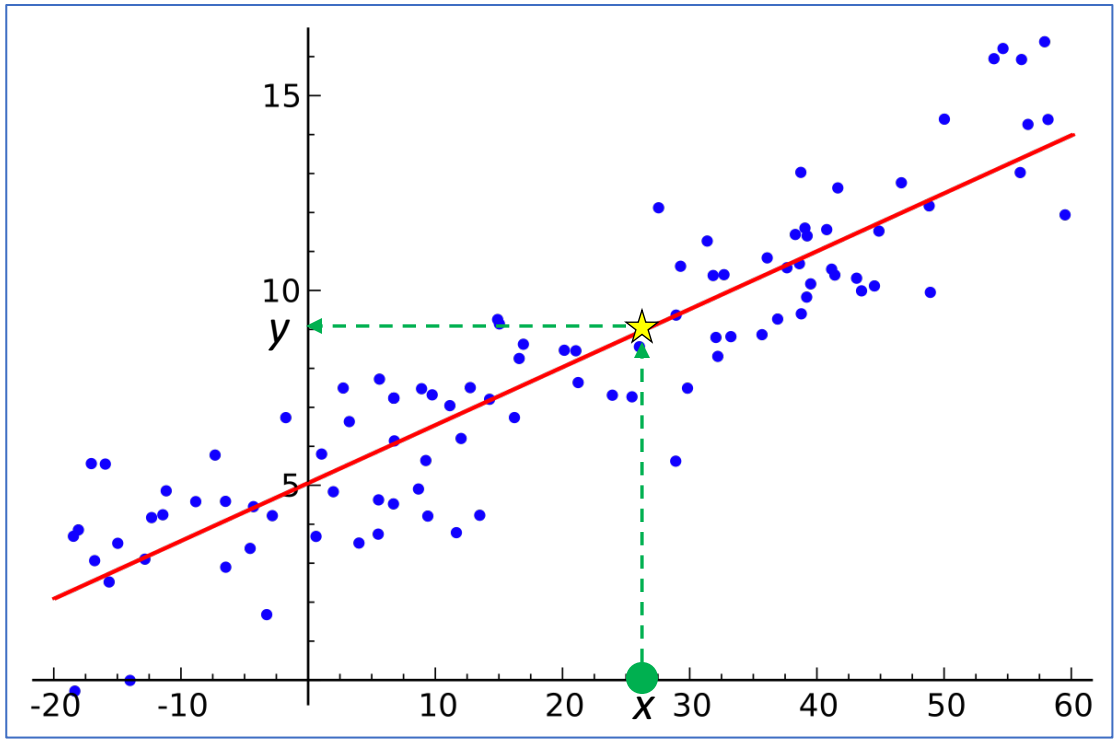


Images: Wikipedia

Regression

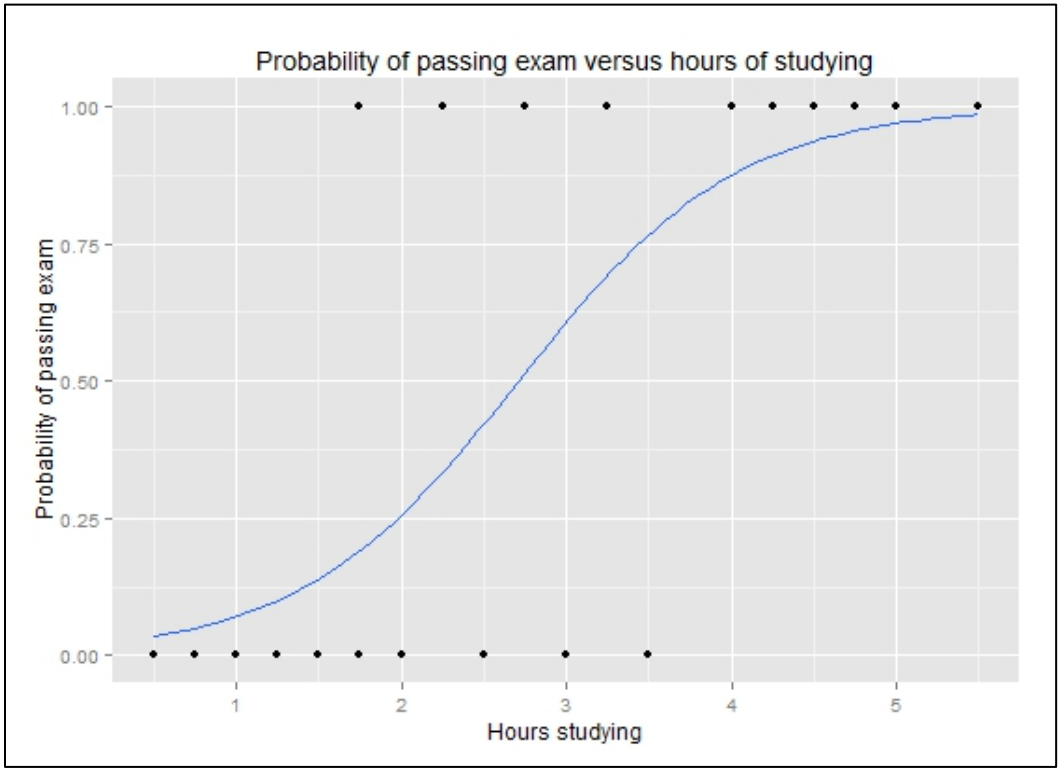
Training data: vectors ***x*** (independent variable) and ***y*** (dependent variable)

Testing goal: given a new *x* value, can we predict *y*?



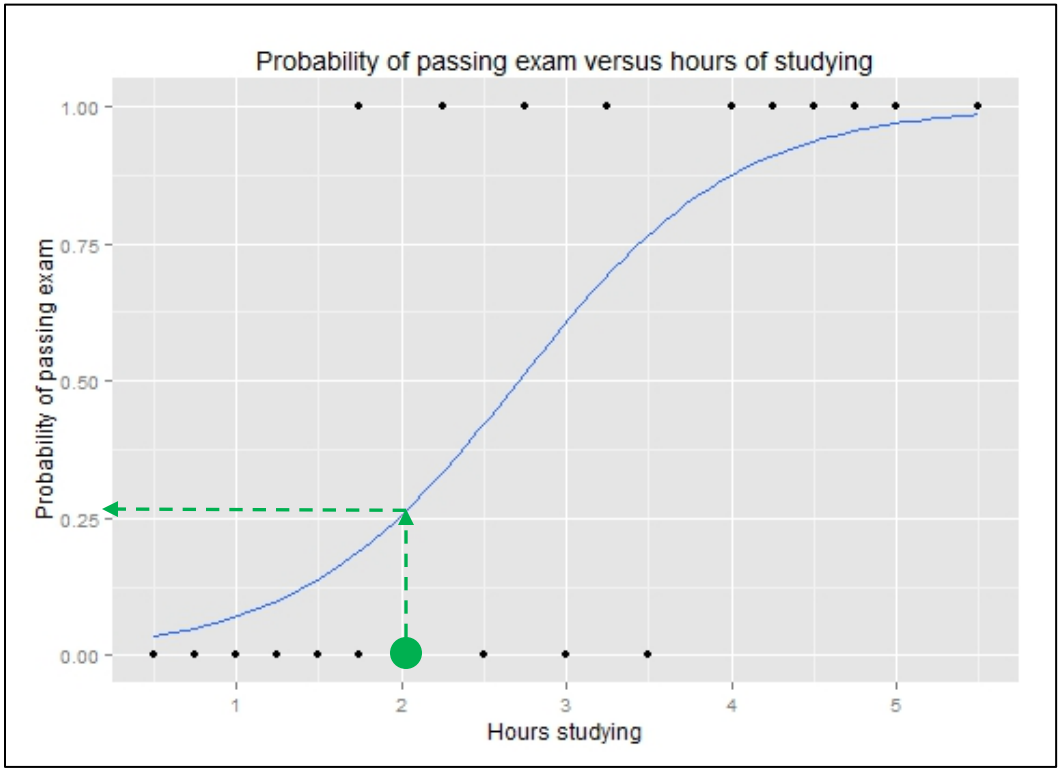
Images: Wikipedia

Logistic regression for classification



Images: Wikipedia

Logistic regression for classification

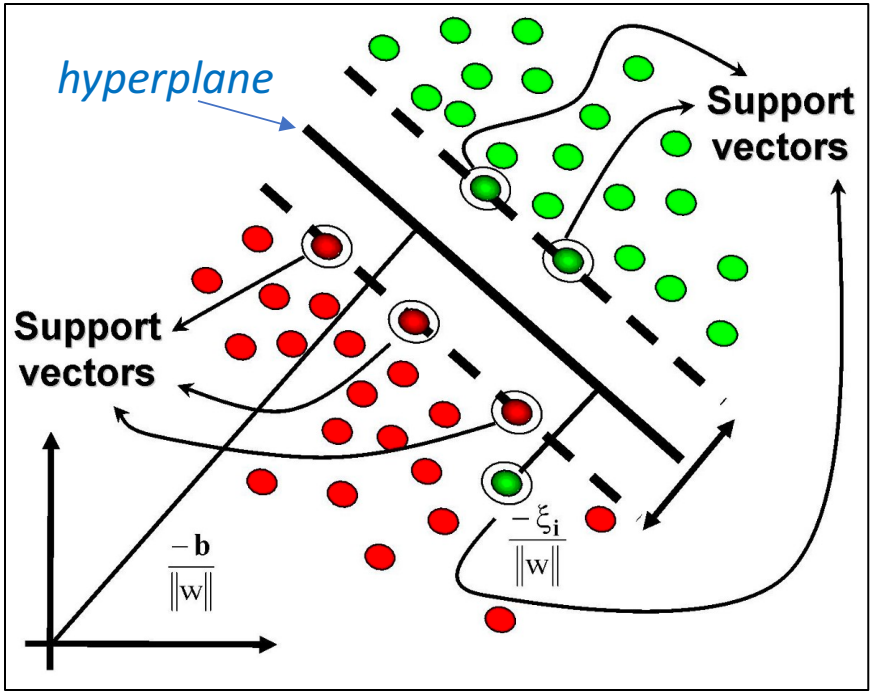


Images: Wikipedia

Support Vector Machines (SVM)

Idea: for 2 (or more classes), try to create the “best” boundary between them

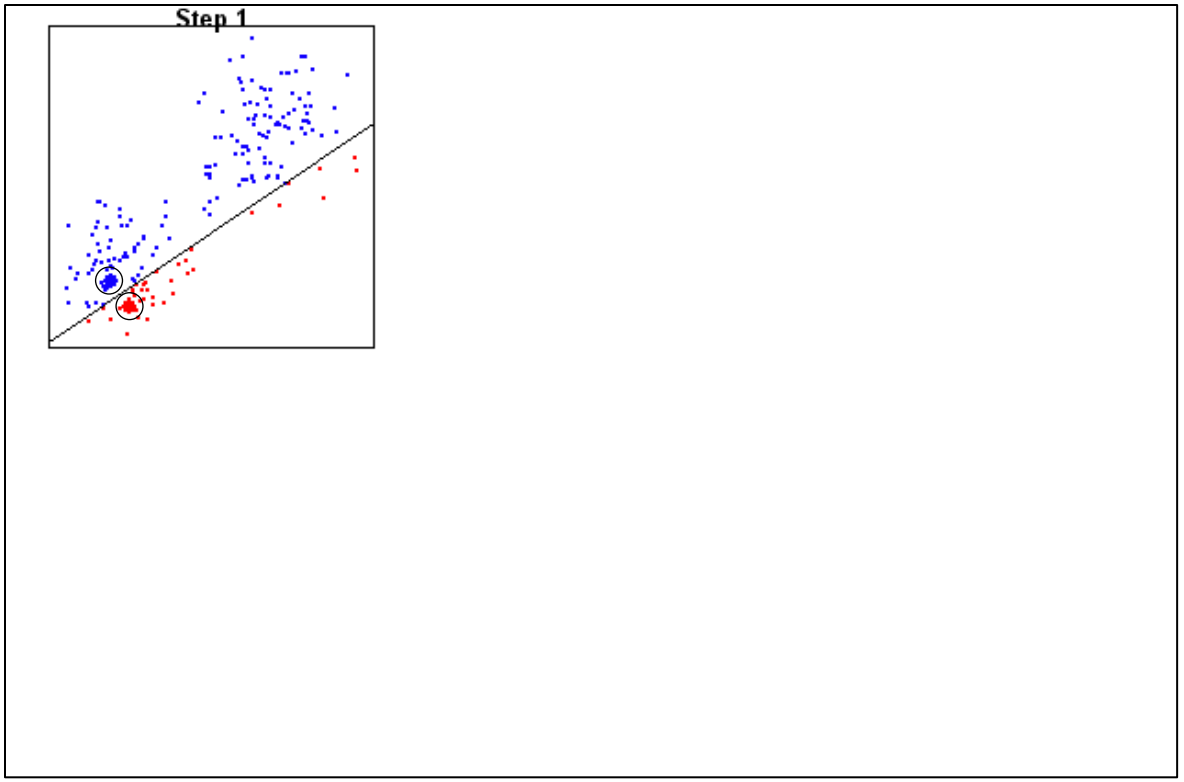
New examples can be classified based on which side of the *hyperplane* they fall on



Images: medium.com

Clustering (unsupervised learning)

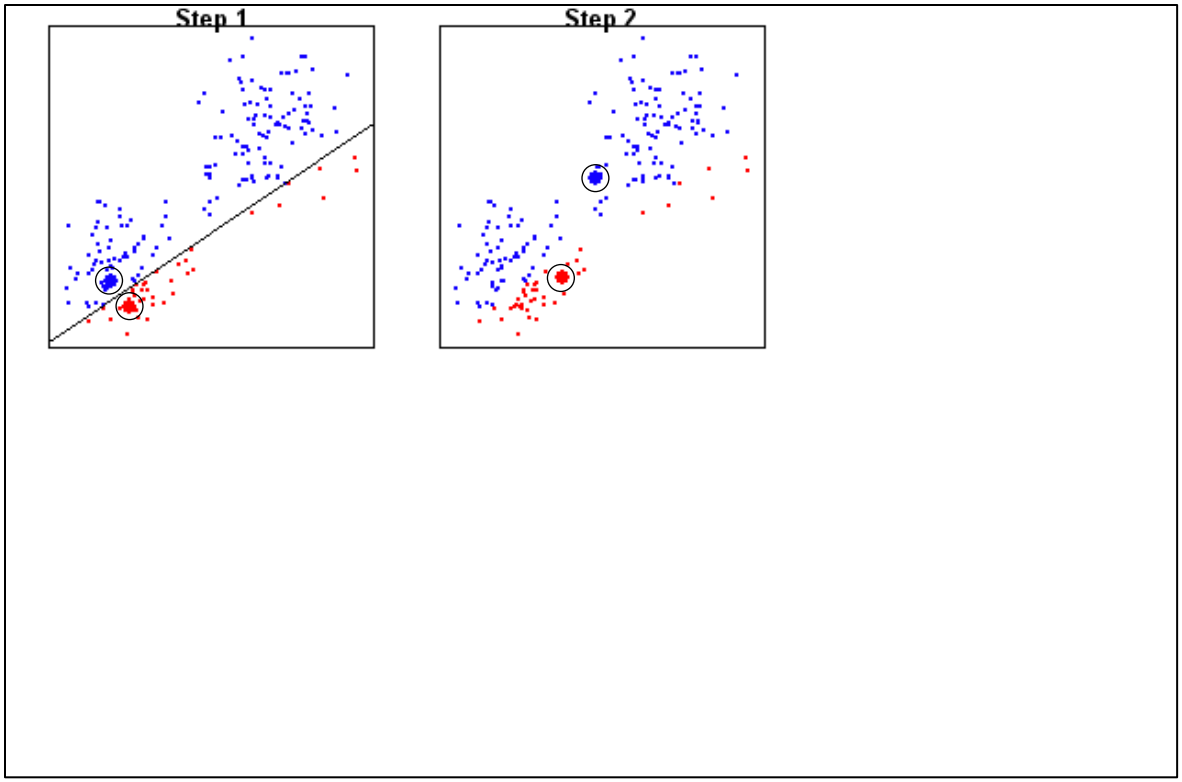
Choose two random data points to be the first means



Images: Polymatheia

Clustering (unsupervised learning)

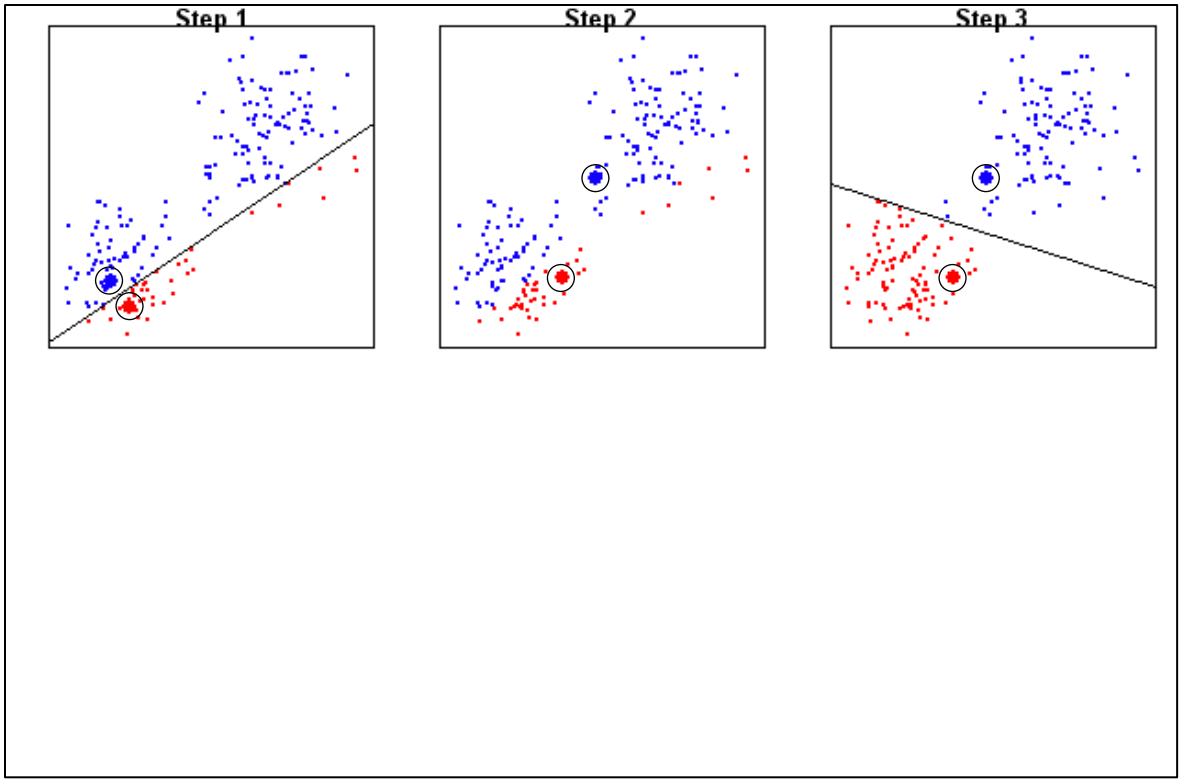
Color each point based on which mean is closest, then find means of resulting clusters



Images: Polymatheia

Clustering (unsupervised learning)

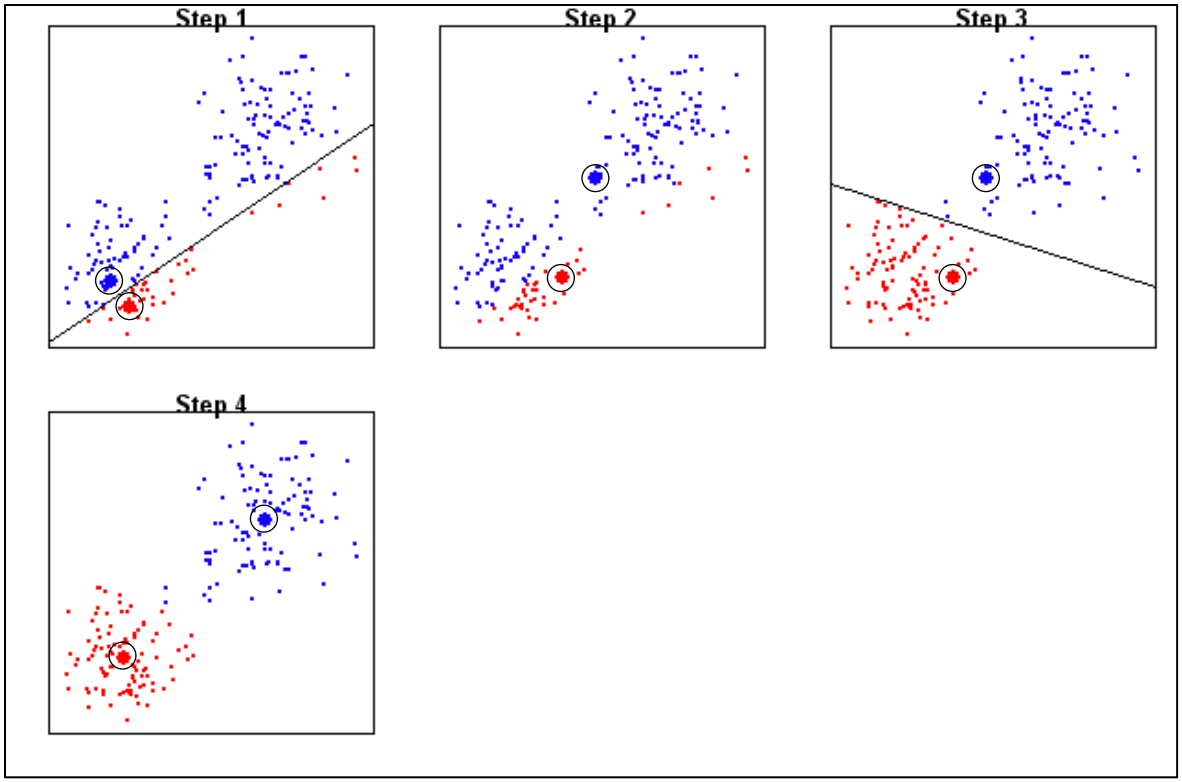
Repeat the process until the means are not changing



Images: Polymatheia

Clustering (unsupervised learning)

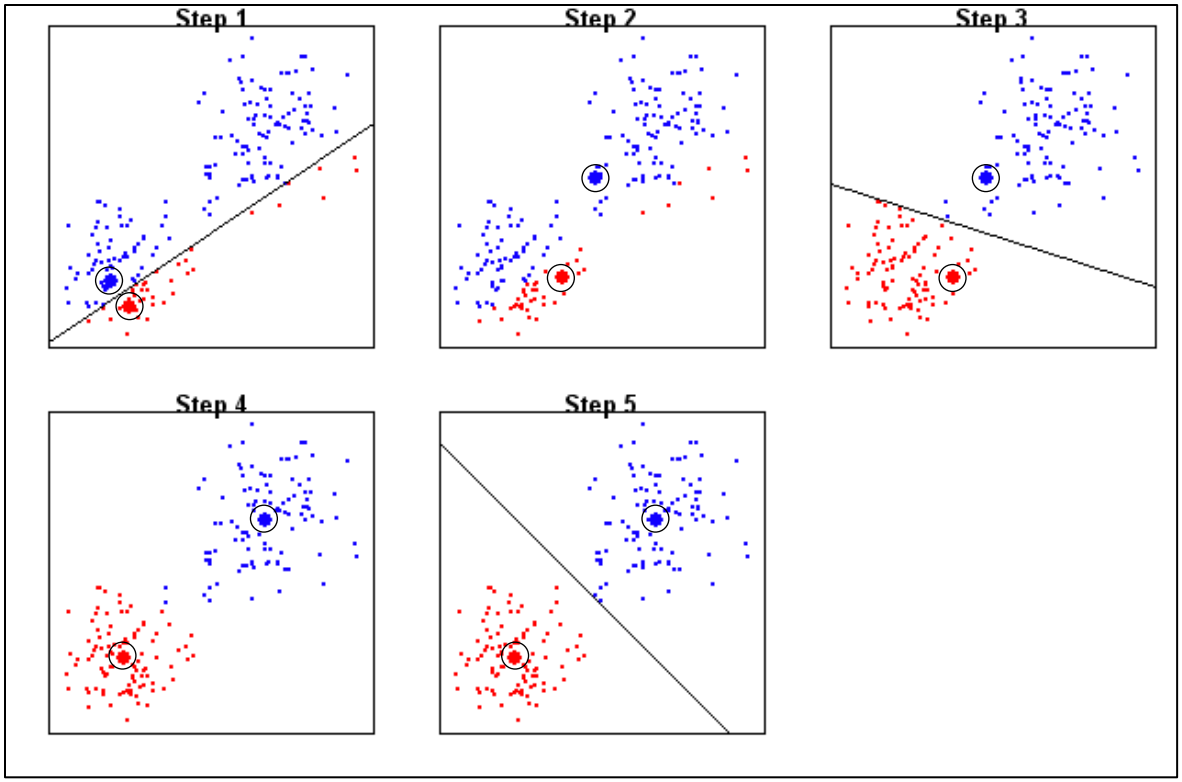
Repeat the process until the means are not changing



Images: Polymatheia

Clustering (unsupervised learning)

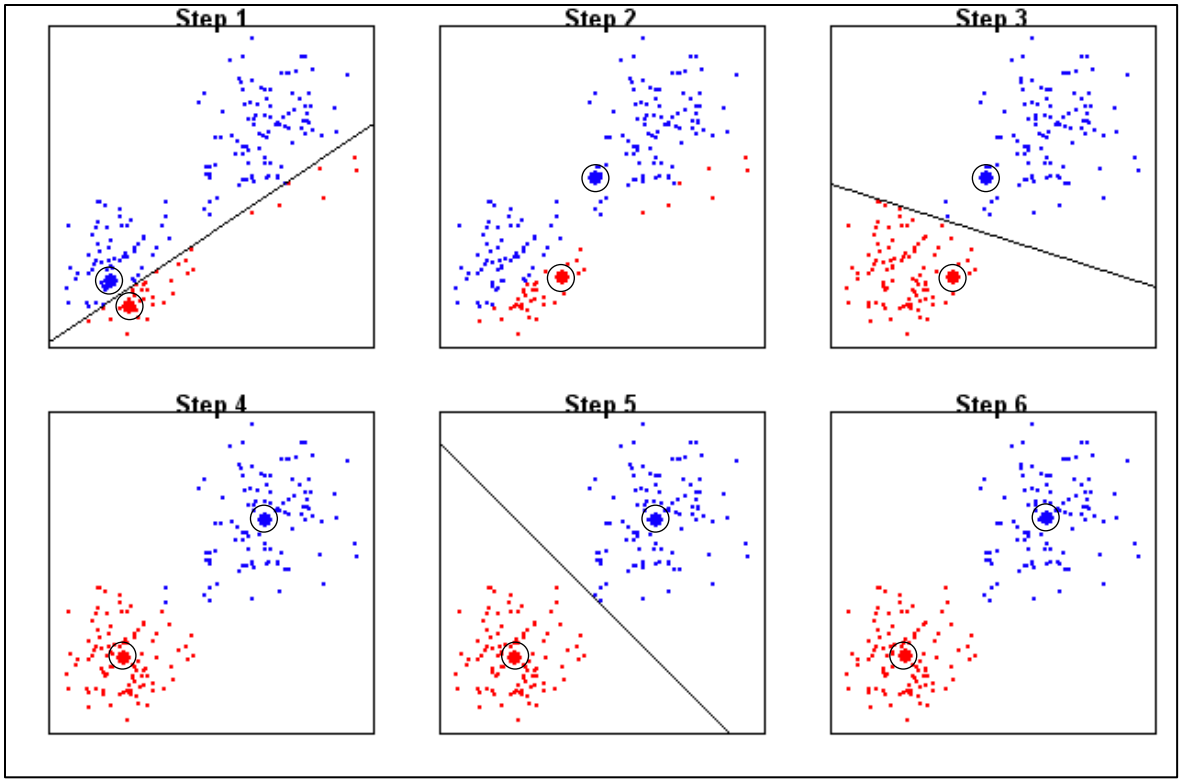
Repeat the process until the means are not changing



Images: Polymatheia

Clustering (unsupervised learning)

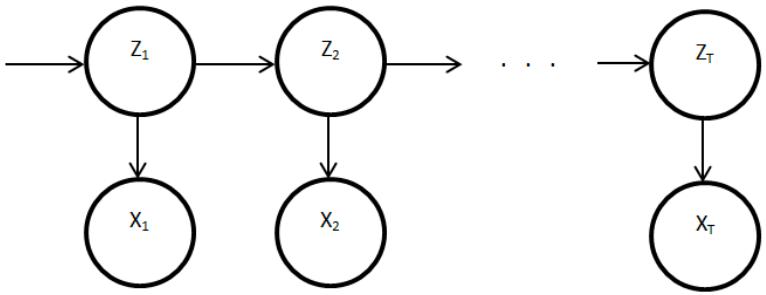
Repeat the process until the means are not changing



Images: Polymatheia

HMMs form a class of machine learning methods too

- Can be *supervised* (i.e. we know the hidden state sequence for some examples, use that to infer transition/emission probabilities)
- Then estimate hidden state sequence for new unlabeled data

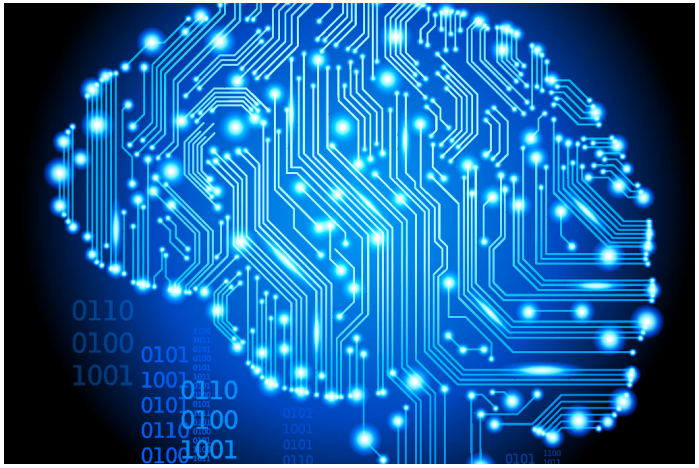


- Can be *unsupervised* (i.e. we don't know the hidden state sequence and we want to learn/predict this latent variable)

Images: Cross Validated

Recent trends in ML

- Inspired by how neurons are connected in our brains, "*deep learning*" has recently become successful in many fields



Images: QuantStart

Deep learning for images

X_1

X_2

X_3

X_4

X_5

X_6

input data

?

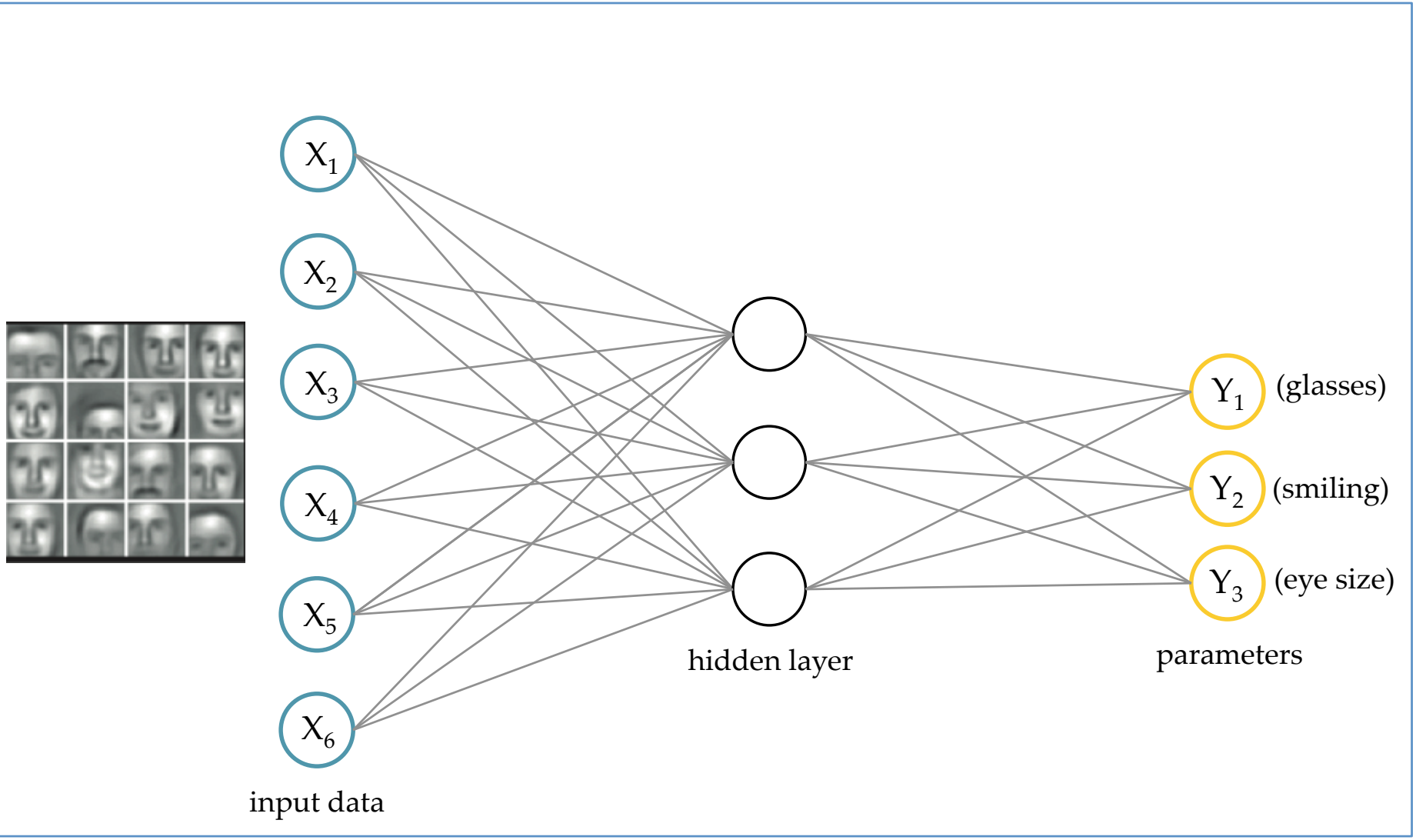
Y_1 (glasses)

Y_2 (smiling)

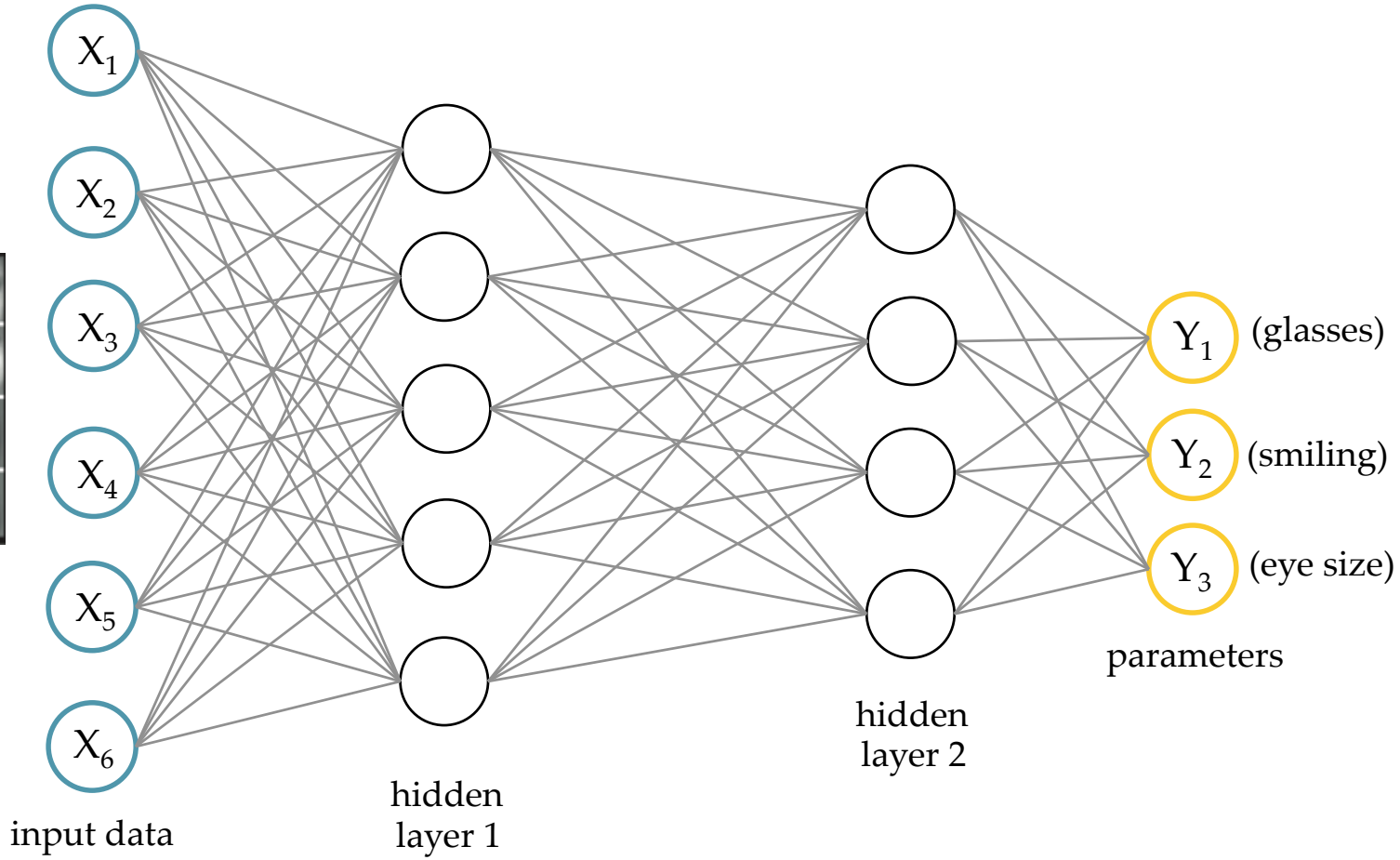
Y_3 (eye size)

parameters

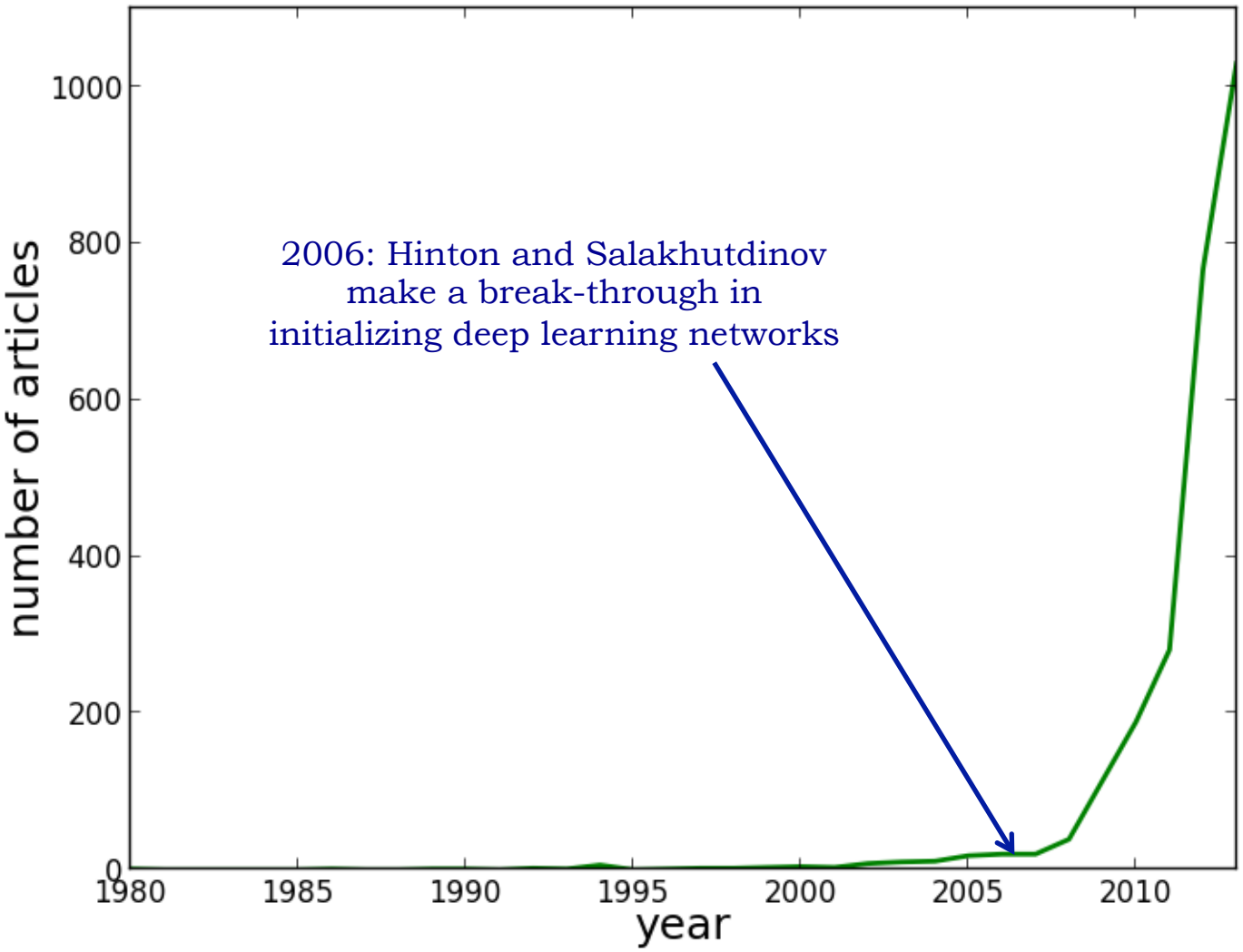
Classical neural network



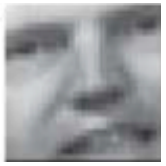
Deep network



Number of articles about deep learning over time



Break-through: unsupervised learning, autoencoder



x_1

x_2

x_3

x_4

x_5

x_6

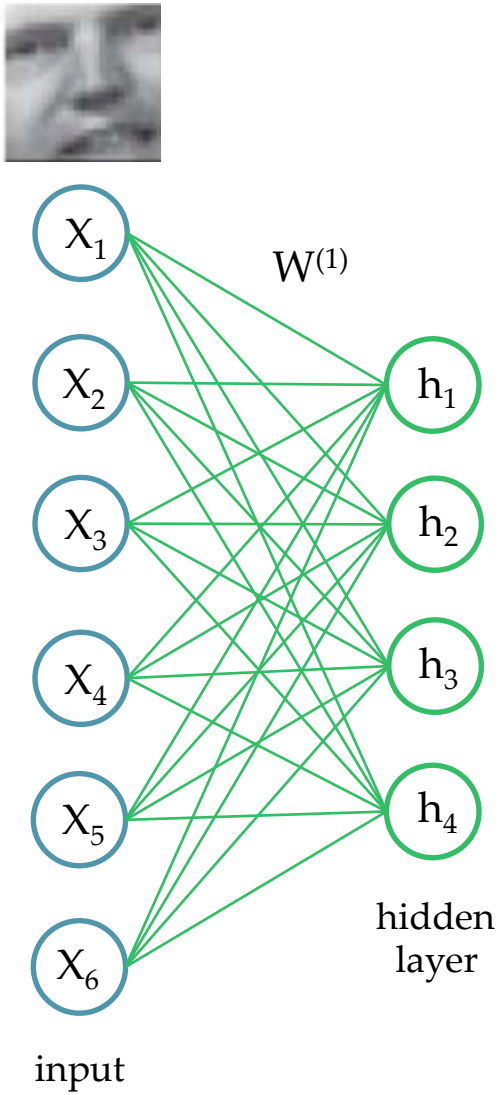
input

Break-through: unsupervised learning, autoencoder

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

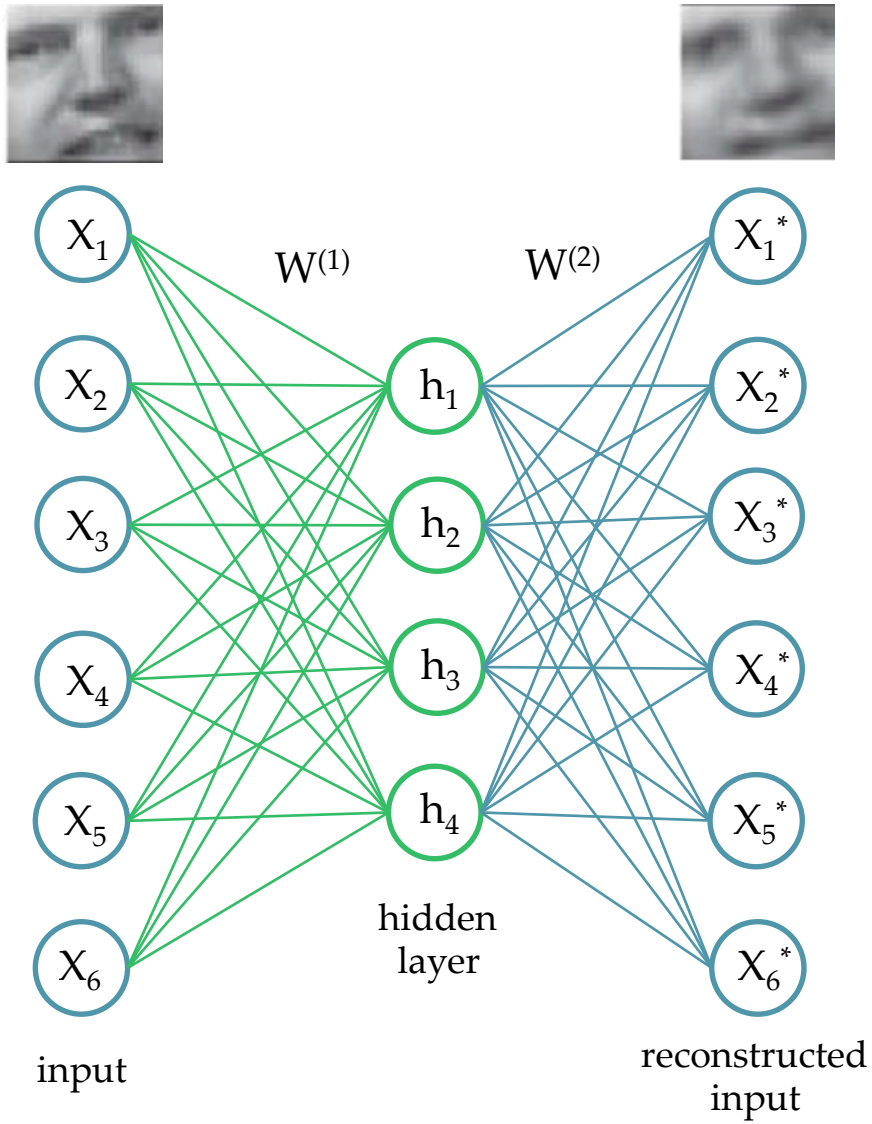
1. Project data into a lower dimension:

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

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

2. From reduced features, reconstruct:

$$x_i^* = \sigma\left(W_i^{(2)} \cdot h\right)$$



Break-through: unsupervised learning, autoencoder

1. Project data into a lower dimension:

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

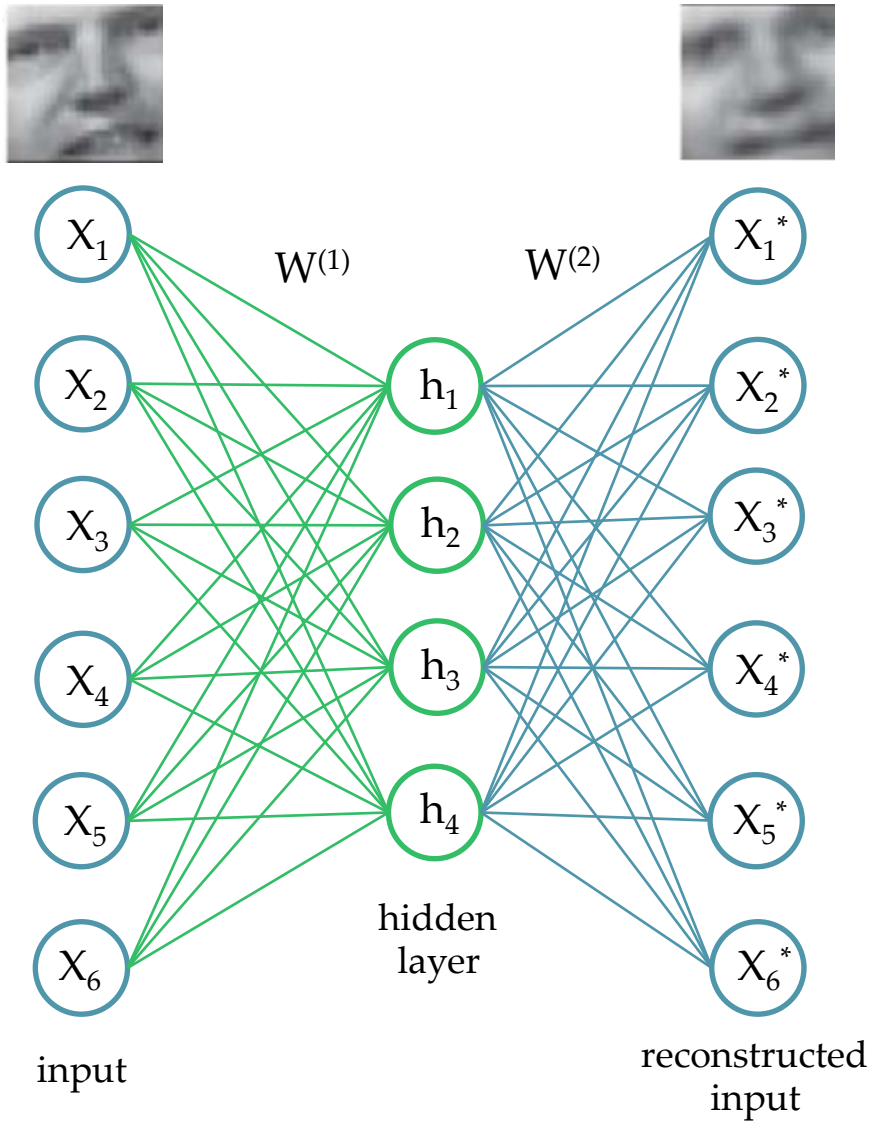
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

2. From reduced features, reconstruct:

$$x_i^* = \sigma\left(W_i^{(2)} \cdot h\right)$$

3. Minimize objective function:

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



PCA vs. Autoencoder

Original image

PCA vs. Autoencoder

Original image

PCA reconstruction

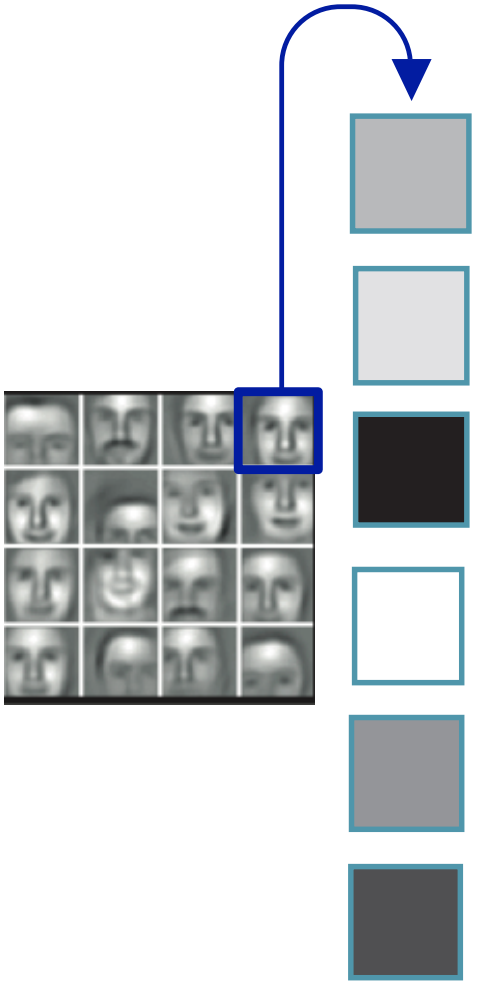
PCA vs. Autoencoder

Original image

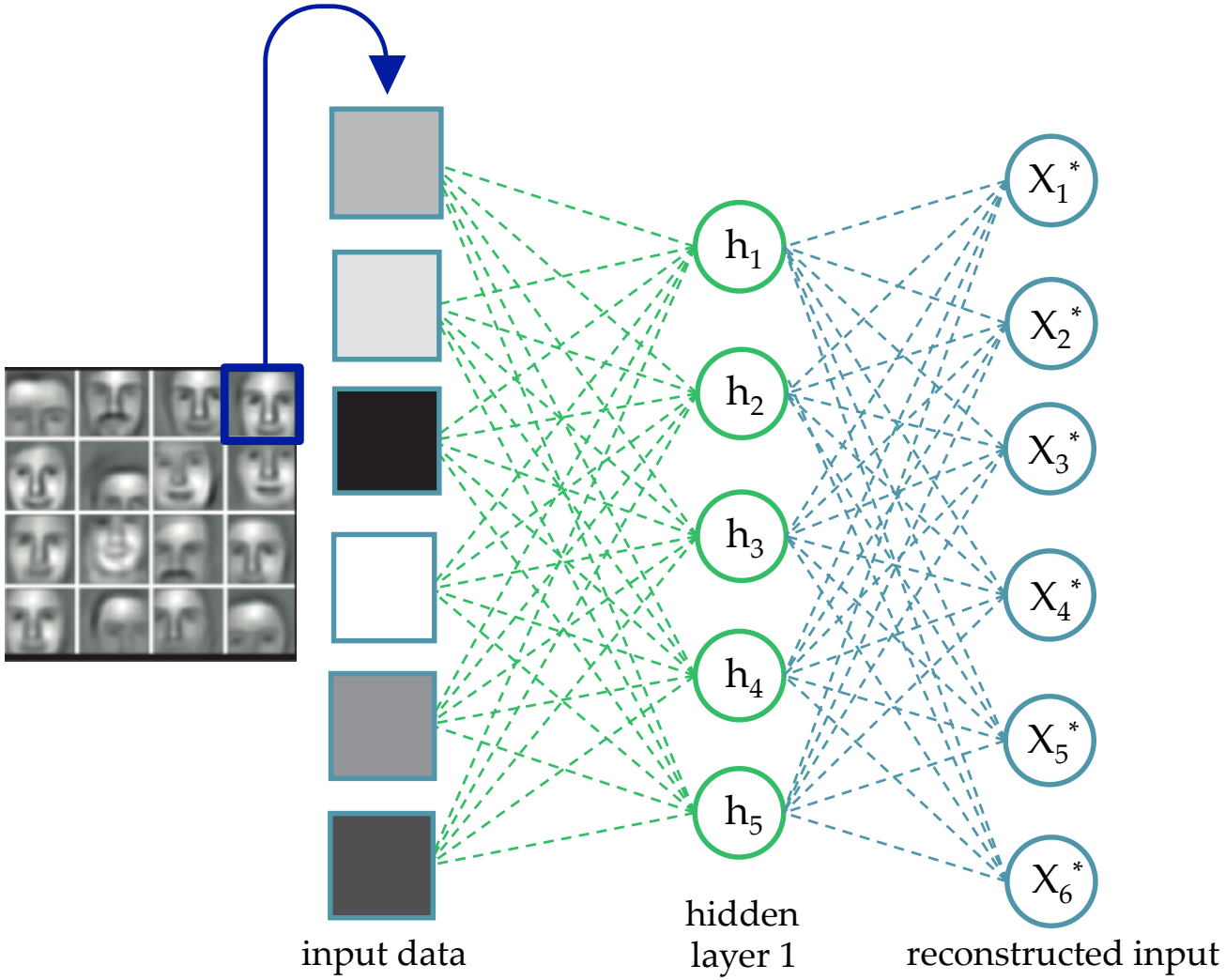
PCA reconstruction

Autoencoder reconstruction

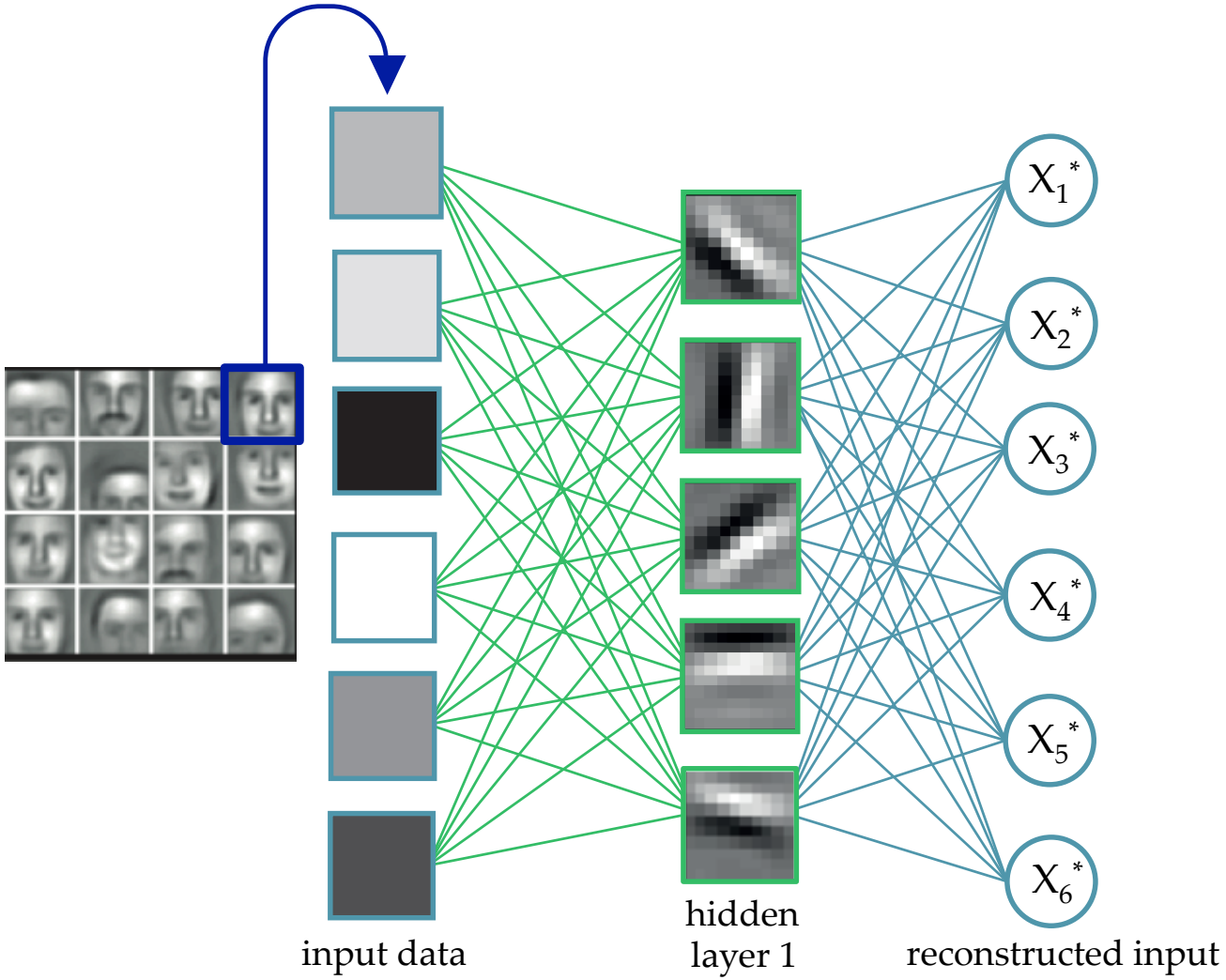
Transform the input data



Feature learning for hidden layer 1



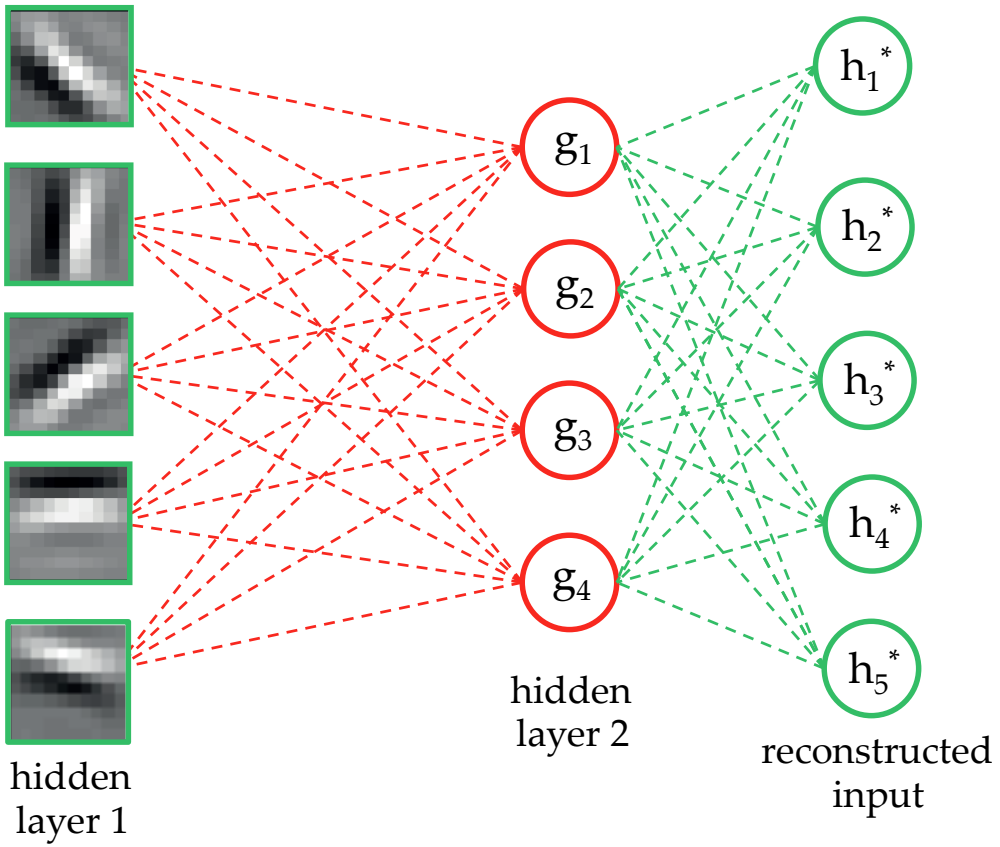
Feature learning for hidden layer 1



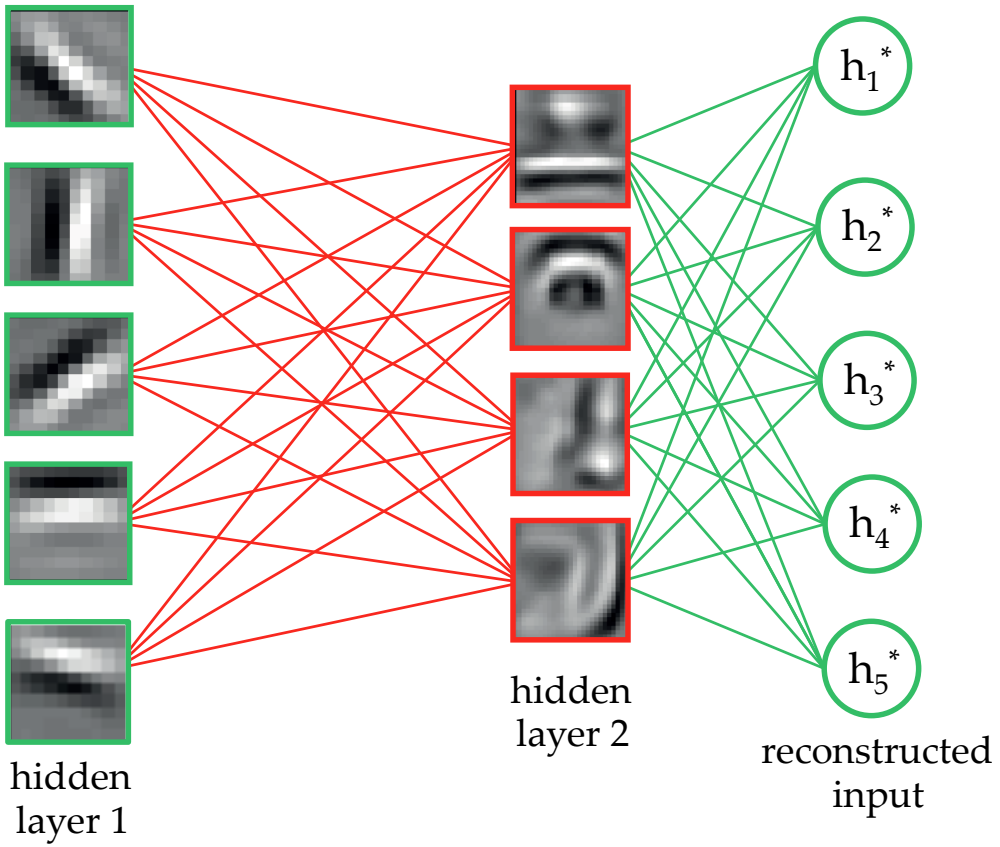
Low-level features become the new data

hidden
layer 1

Feature learning for hidden layer 2



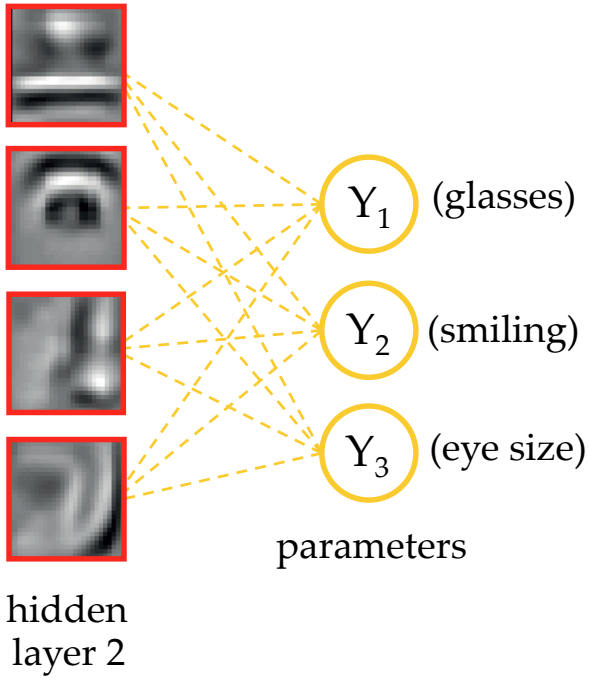
Feature learning for hidden layer 2



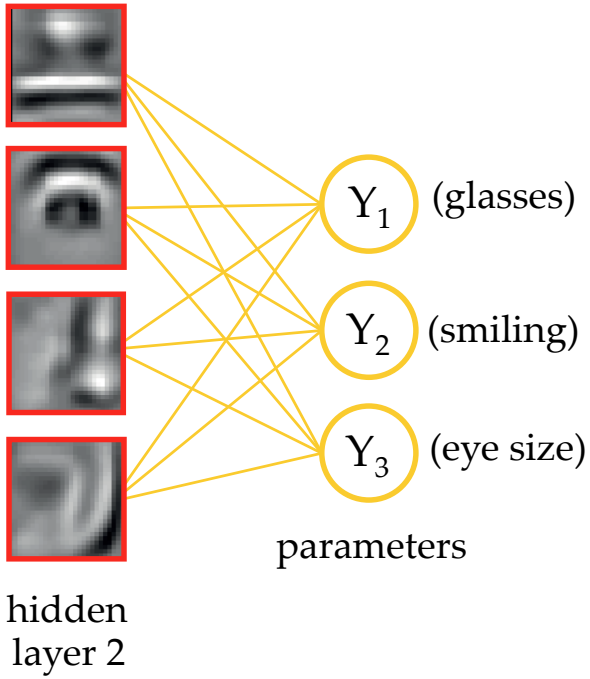
High-level features become the new data

hidden
layer 2

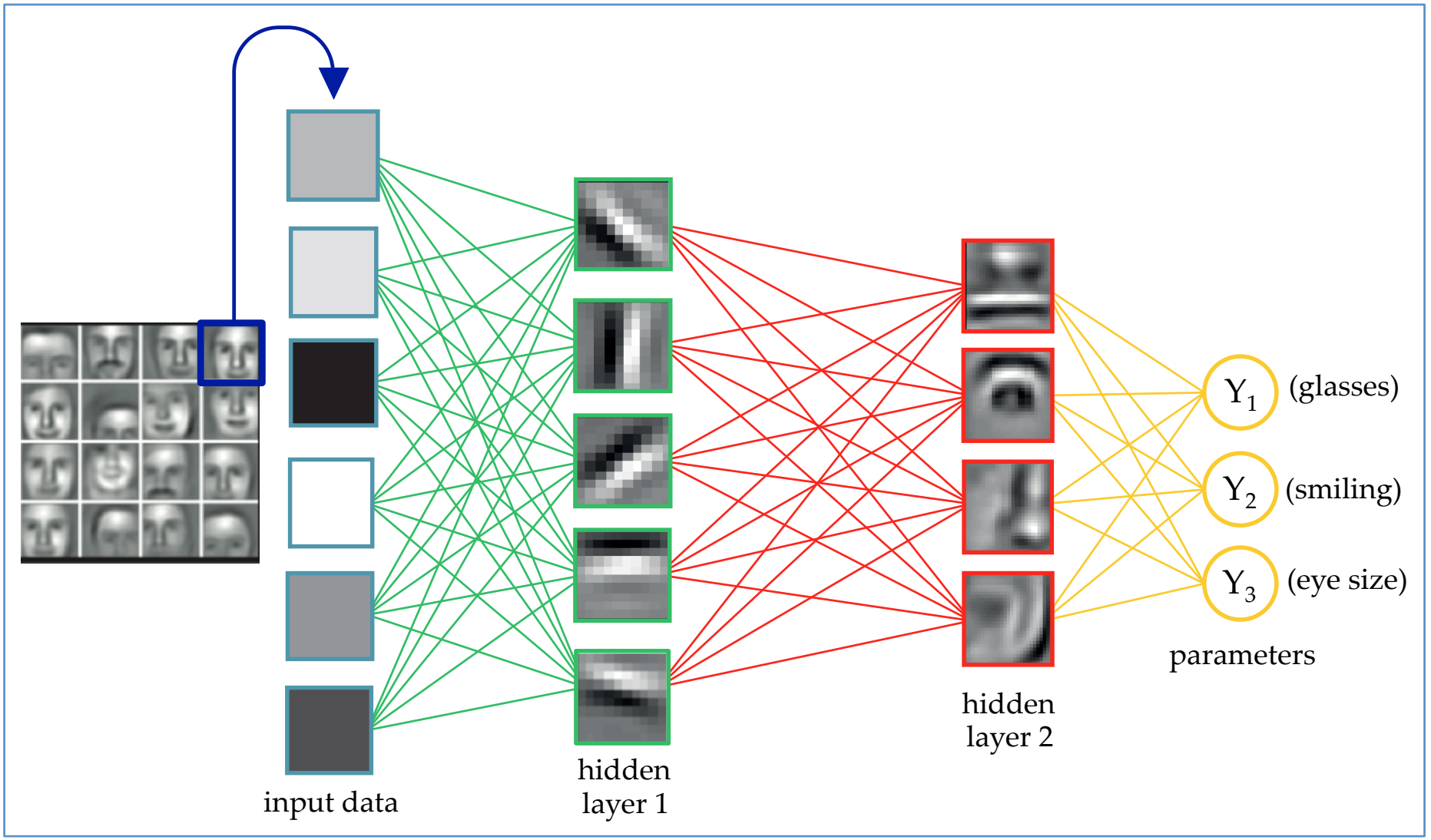
Last layer: supervised learning



Last layer: supervised learning

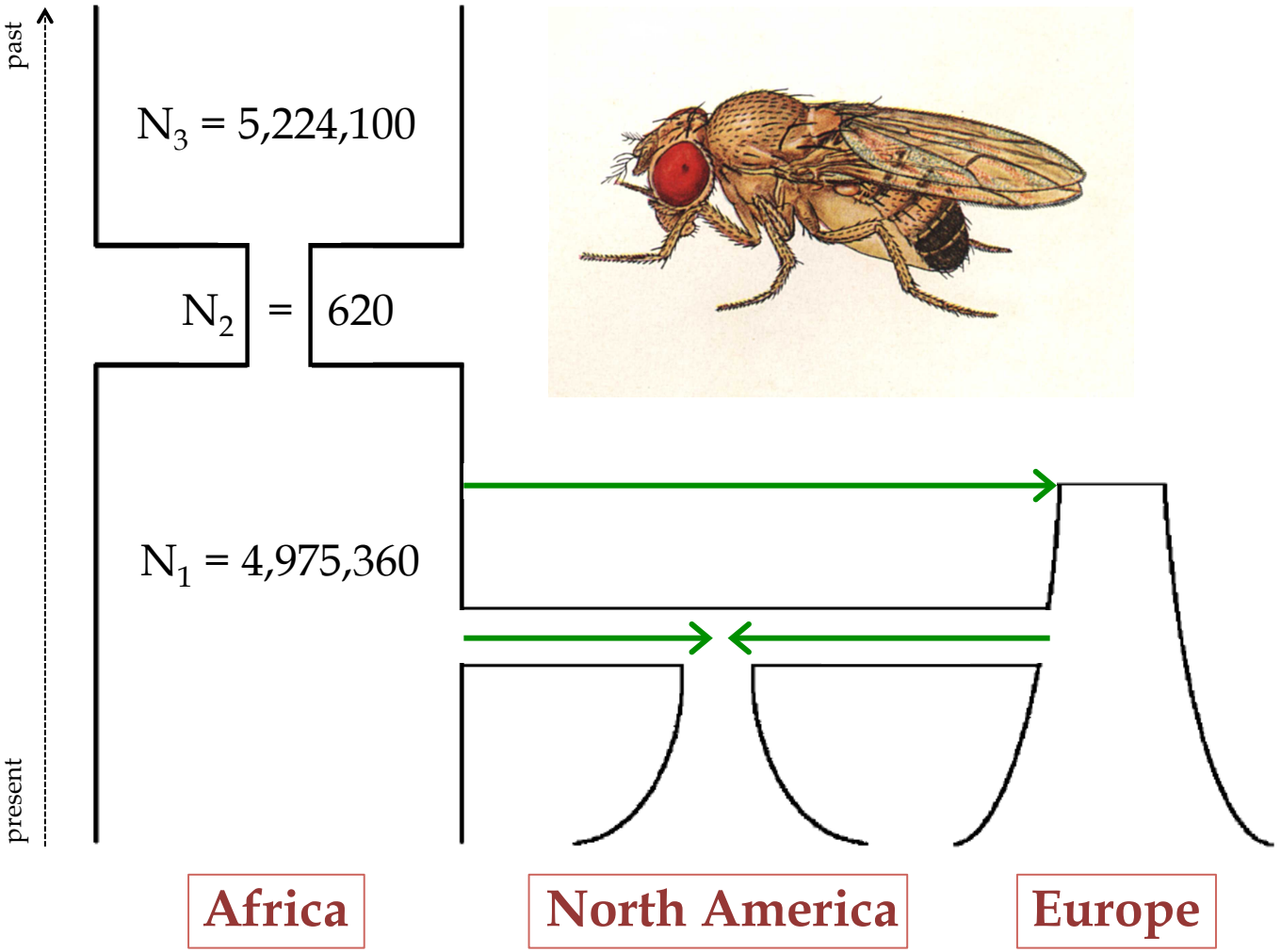


"Fine-tune" the entire deep network



Application of deep learning to population genetics

Motivation: demographic history of *Drosophila*

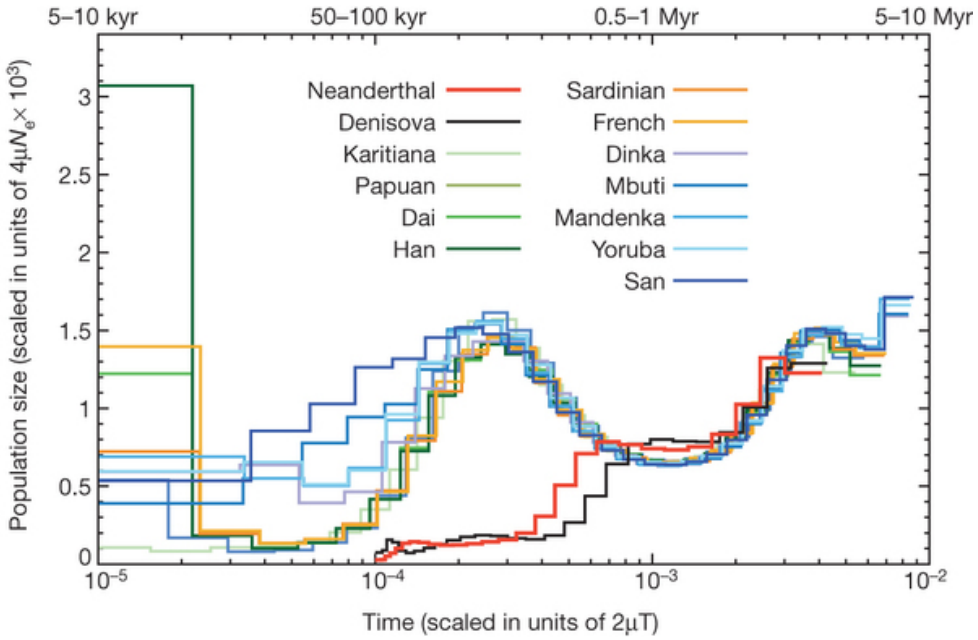


Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population

Pablo Duchen, Daniel Živković, Stephan Hutter, Wolfgang Stephan and Stefan Laurent

GENETICS *January 1, 2013 vol. 193 no. 1 291-301; https://doi.org/10.1534/genetics.112.145912*

Main goal: population sizes and natural selection

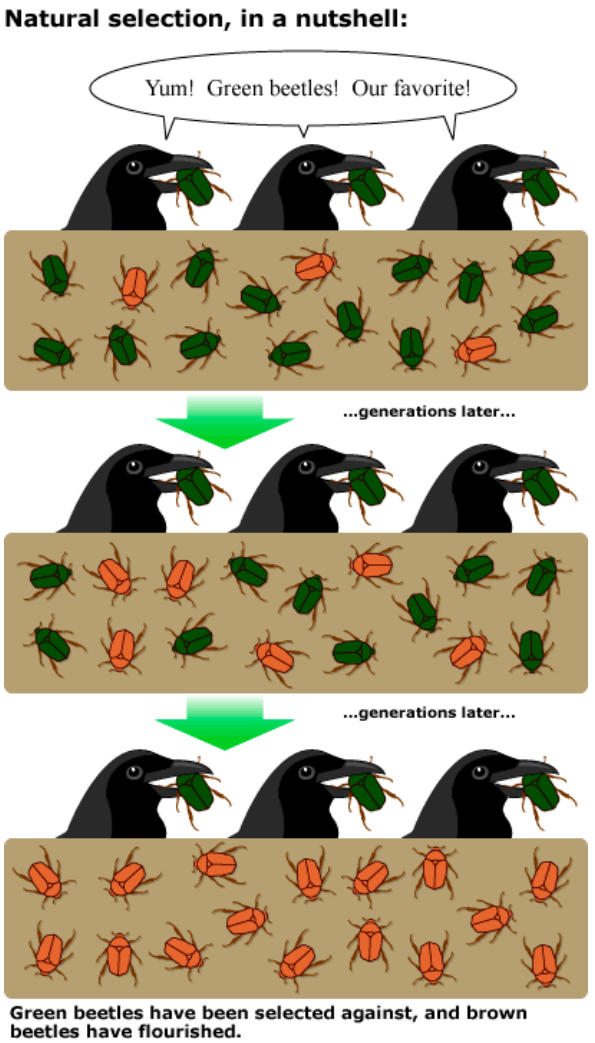


The complete genome sequence of a Neanderthal from the Altai Mountains, *Nature* (2013)

INVITED REVIEW

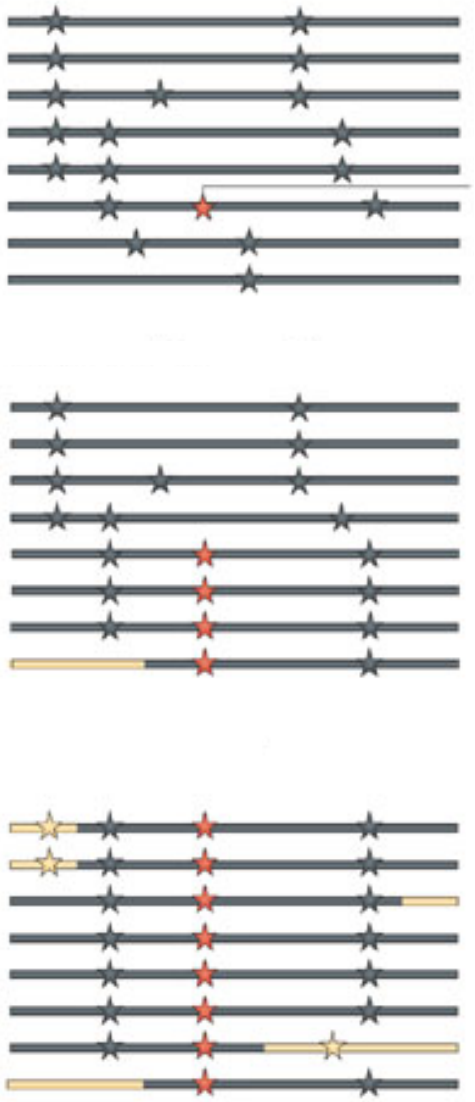
Joint analysis of demography and selection in population genetics: where do we go and where could we go?

JUNRUI LI,*† HAIPENG LI,* MATTIAS JAKOBSSON,‡ SEN LI,‡ PER SJÖDIN‡ and MARTIN LASCOUX*§

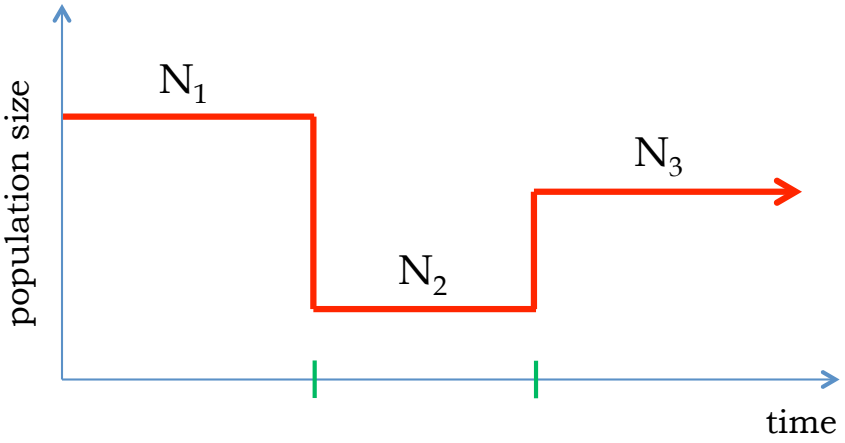


University of California Museum of Paleontology's "Understanding Evolution"

Selective sweeps can cause a loss of diversity



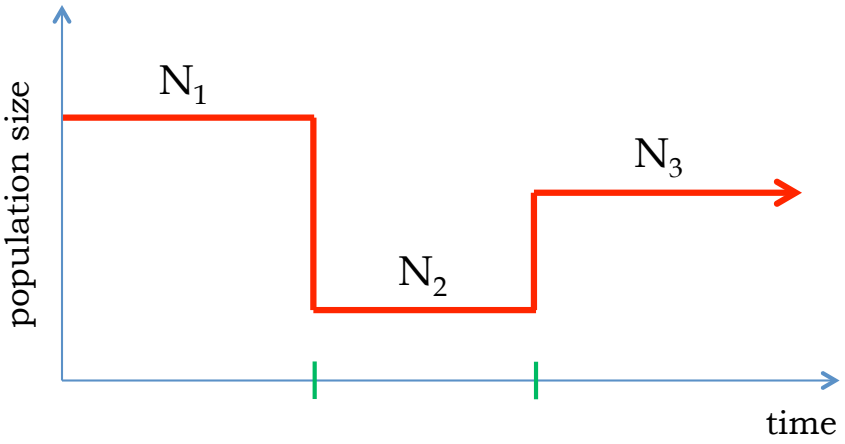
Training data: simulated datasets



400,000 datasets:

- 2,500 bottlenecks
- 160 regions/genome

Training data: simulated datasets

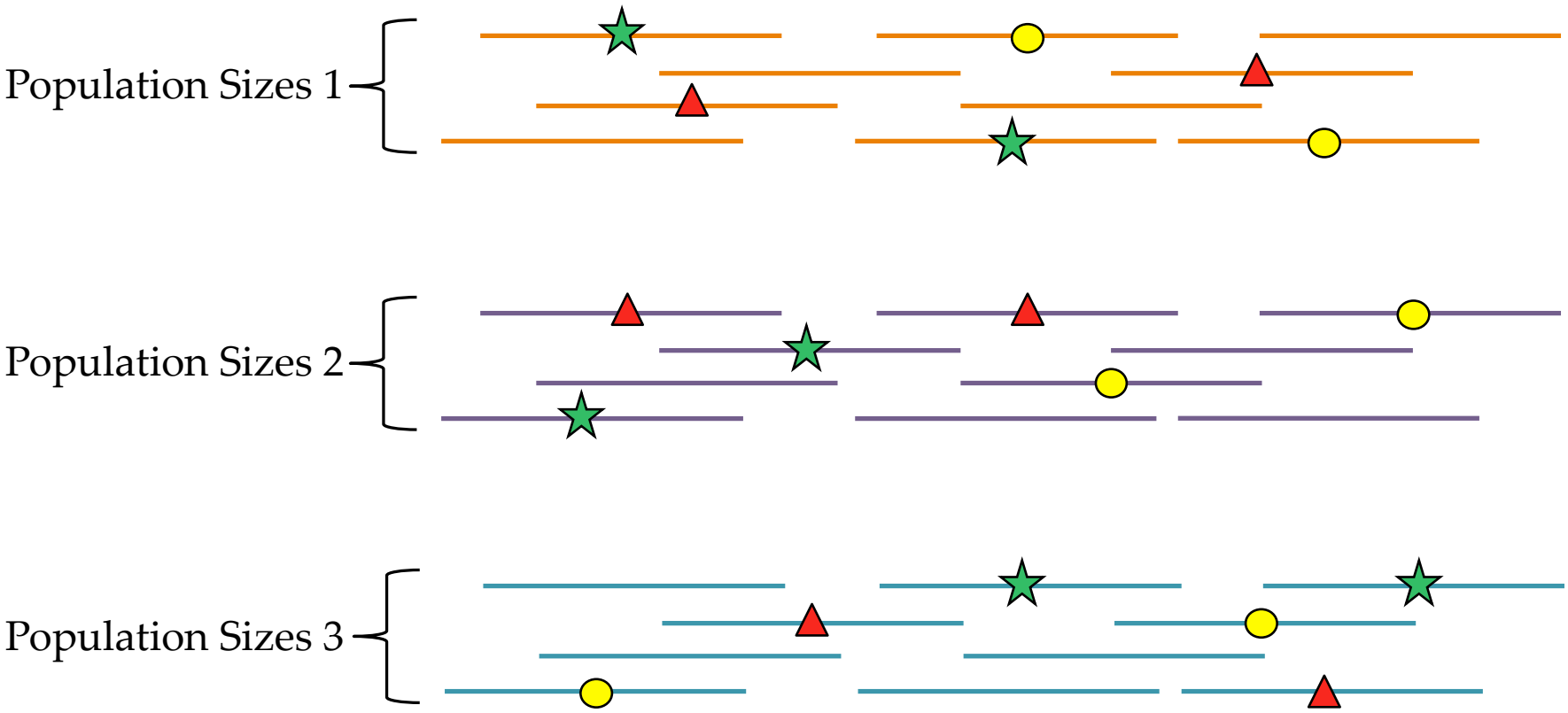


400,000 datasets:

- 2,500 bottlenecks
- 160 regions/genome

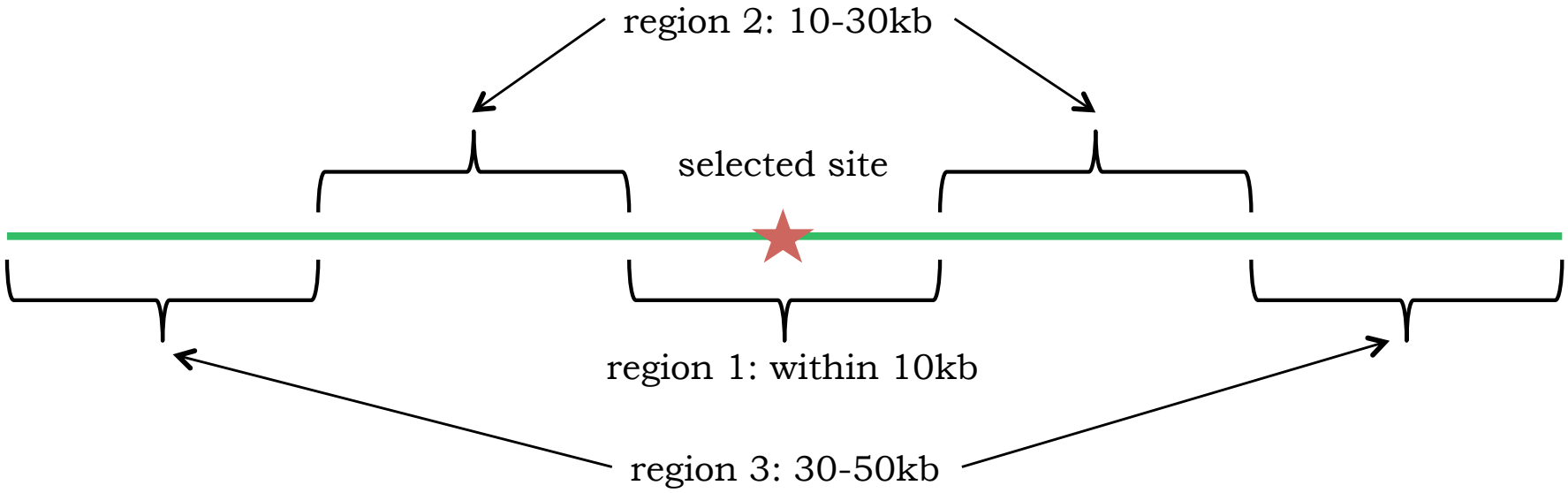
1. baseline effective population size: $N_e = 100,000$
2. $n = 100$ individuals
3. $L = 100,000$ bases per region
4. 75% of data for training and 25% for testing

Selection: four different classes



$\Rightarrow$ 4 selection classes

Compute statistics around selected site

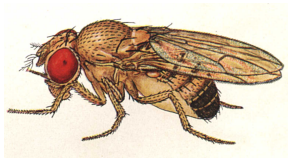


Summary statistics as features

- 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