Week 9

Machine Learning

Leo Breiman's Two Cultures

the logic of data analysis

L. Breiman, Statistical Science (2001)

Leo Breiman's Two Cultures

Data Modeling Culture

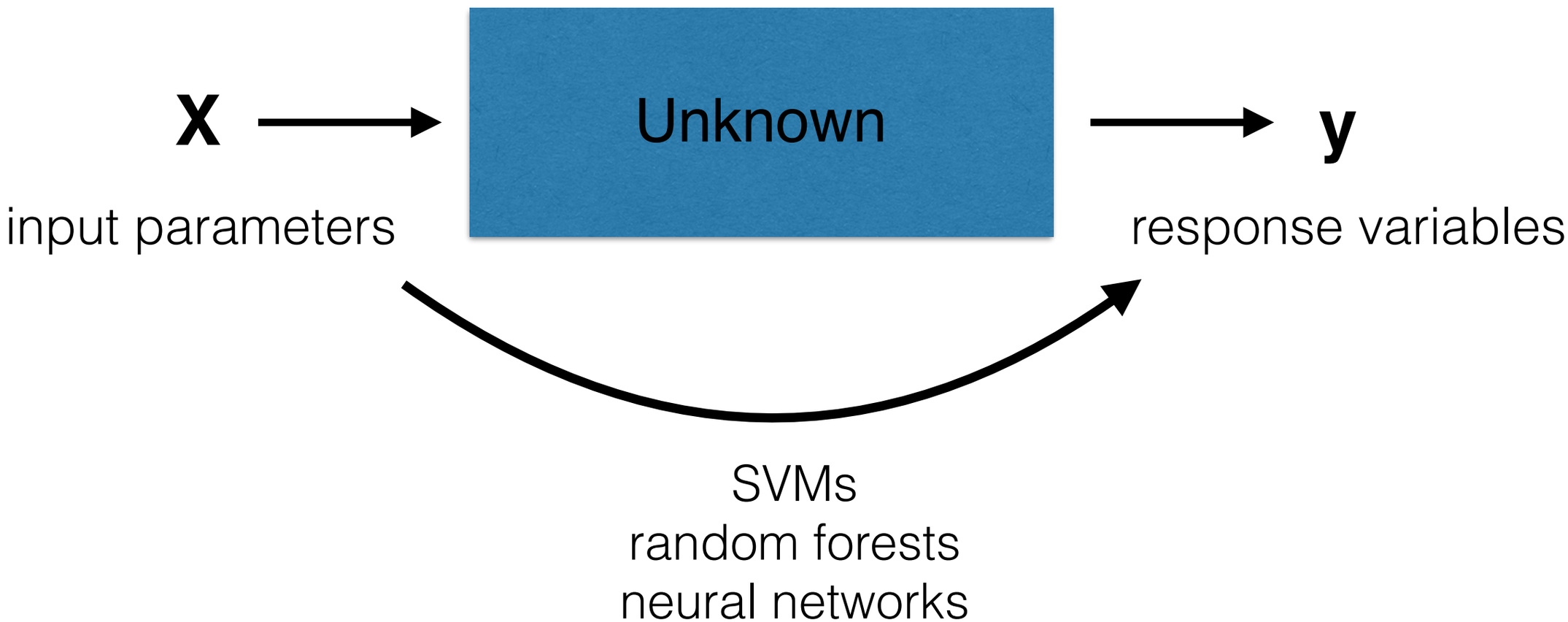
e.g. linear regression

Focus on stochastic model to explain how f(x)-> y

98% of Statistics

Leo Breiman's Two Cultures

Algorithmic Modeling Culture
(machine learning)



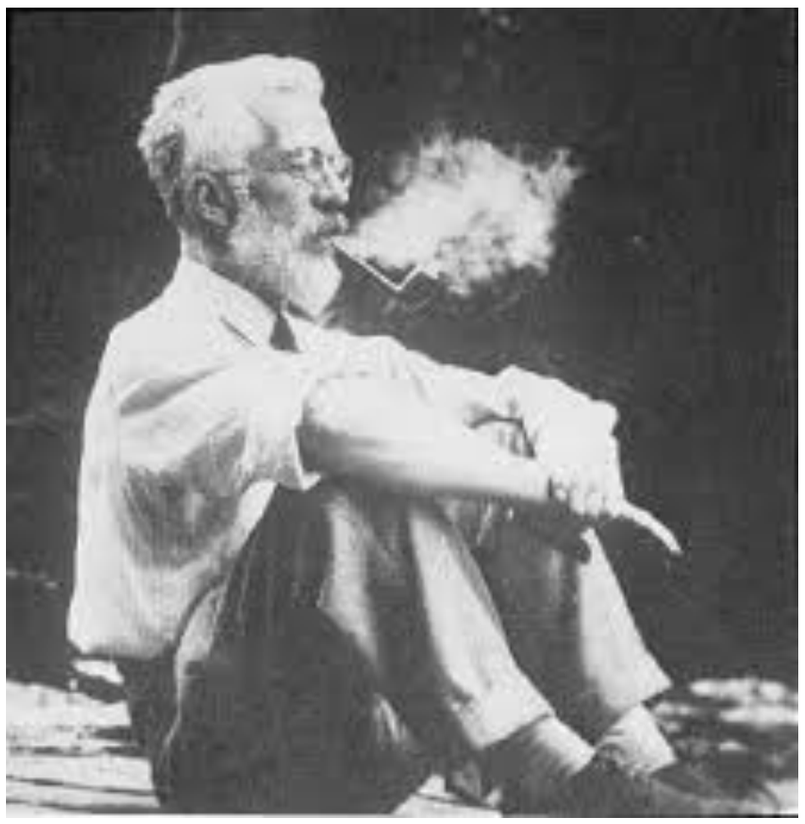
Ignore probabilistic generative model f(x)-> y

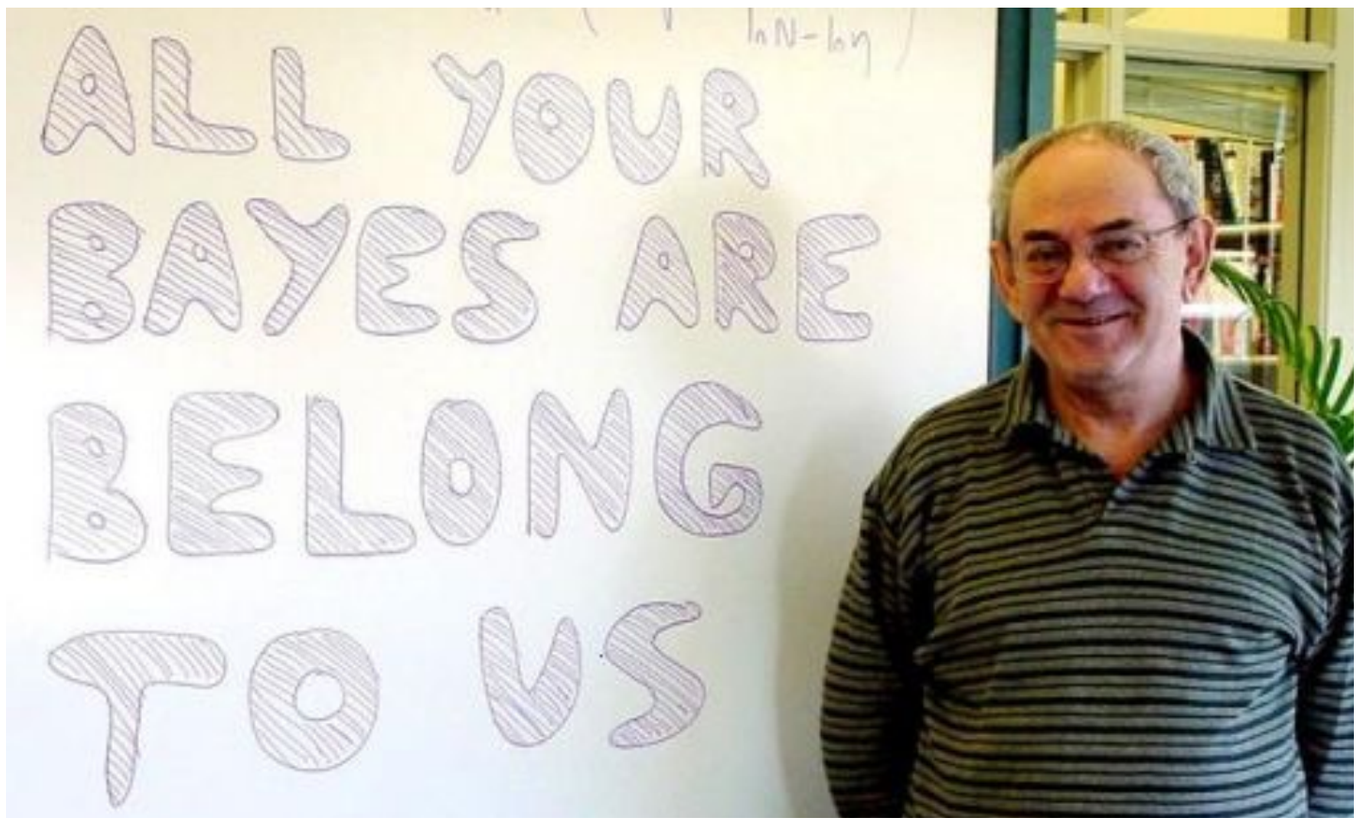
Machine Learning!

amazon

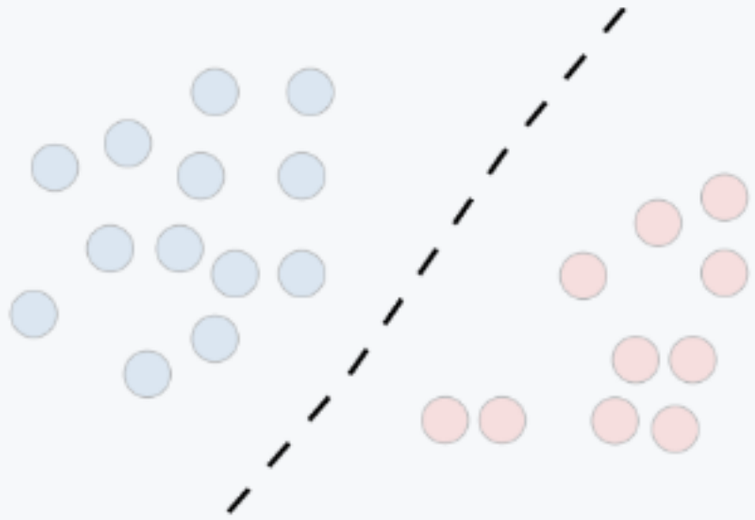
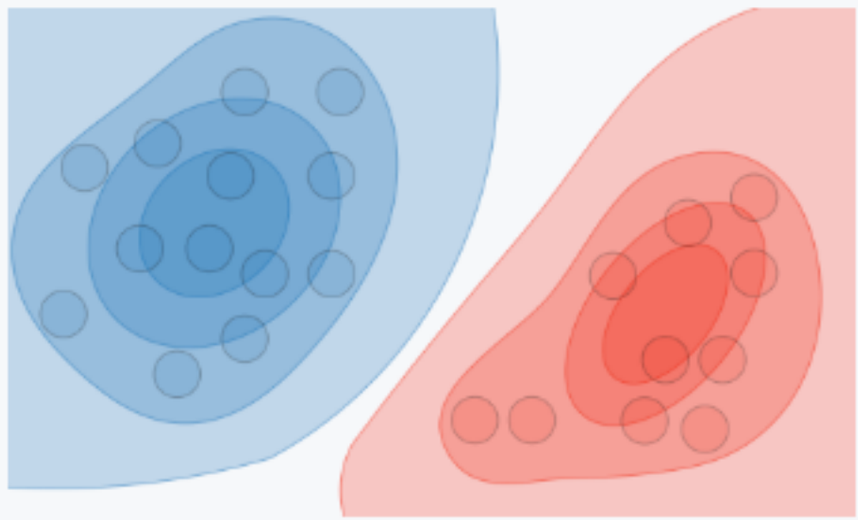
NETFLIX

These guys don't have generative model

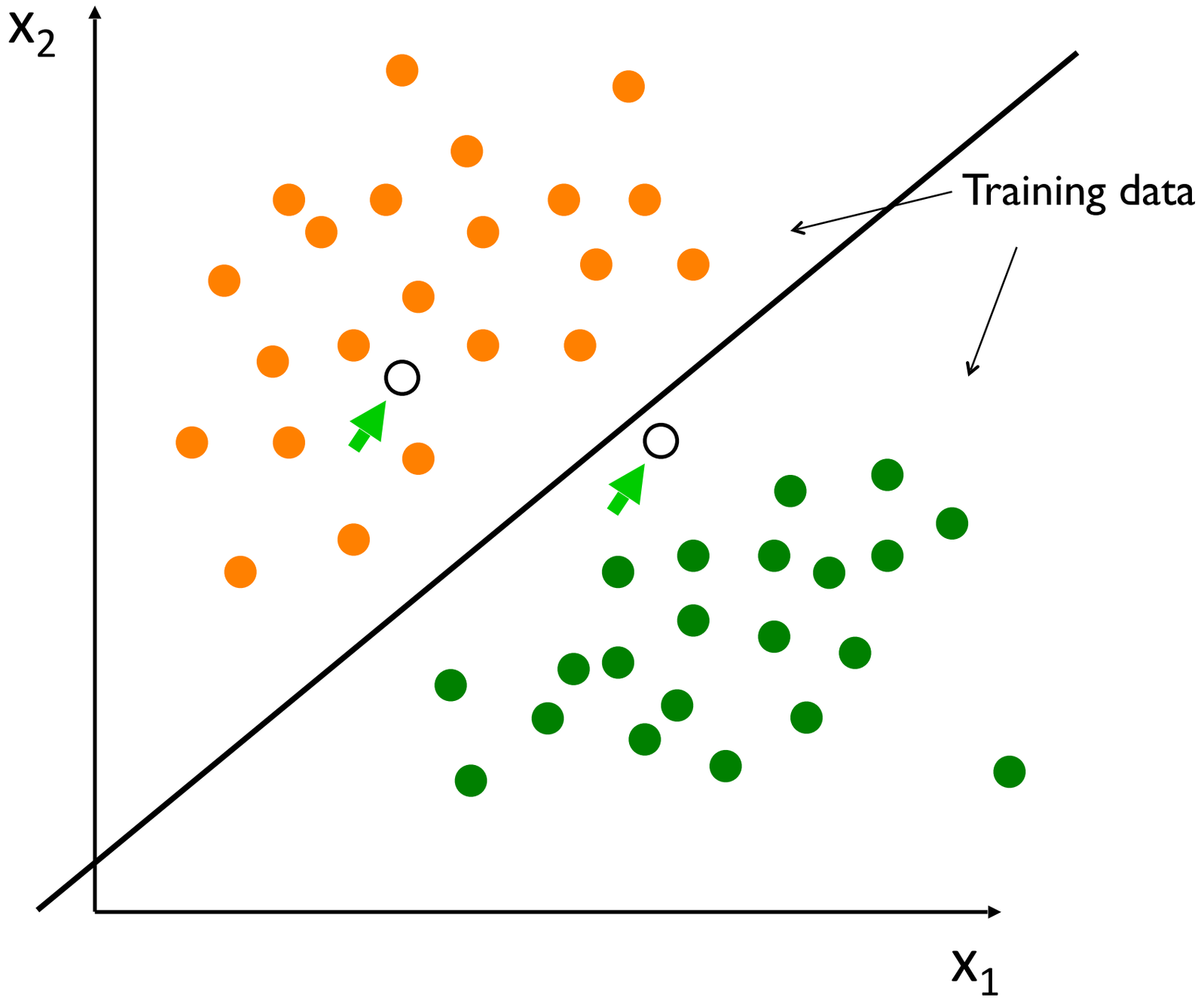




Discriminitive vs Generative Models

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

Supervised Machine Learning



We are using a Support Vector Machine (SVM)

Supervised Machine Learning

Given a set of *N* training (i.e. known, labelled) examples:

$$\{(x_1, y_1), ..., (x_N, y_N)\}$$

feature vector $\mathbb{R}^M$ class label $y \in \{-1, 1\}$

we define a learning function:

$$g : X \to Y \quad \text{e.g.} \quad g(x) = P(y|x)$$

and a loss function:

$$L : g(x) \times Y \to \mathbb{R}^{\geq 0} \text{ e.g. } L(g(x), y) = \mathbb{1}(g(x) \neq y)$$

then simply minimize a chosen risk function:

$$R(g) = \frac{1}{N} \sum_i L(y_i, g(x_i))$$

Supervised Machine Learning

Support Vector Machines

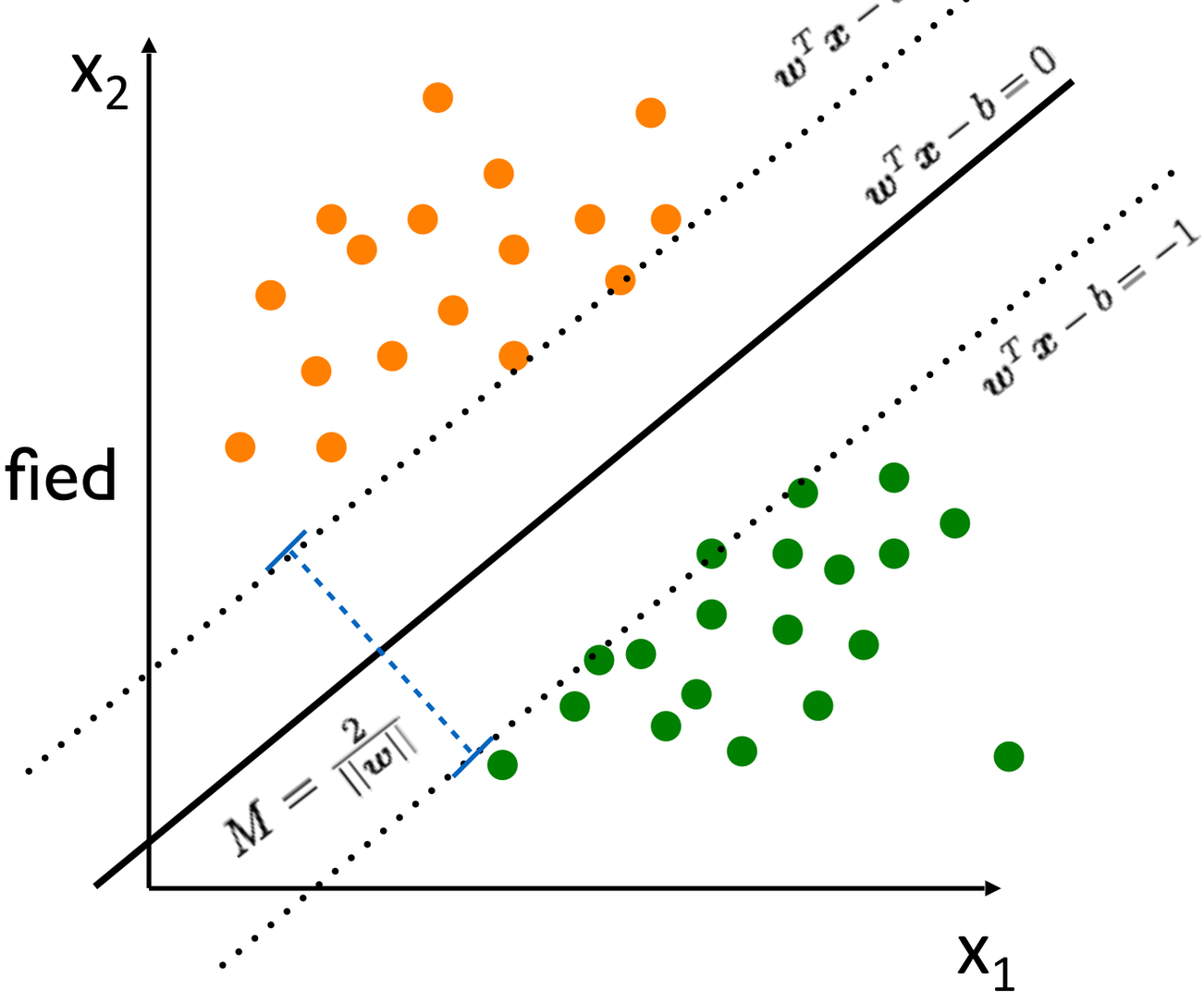
general learning function:

$$g(\boldsymbol{x}) = sign(\boldsymbol{w}^T\boldsymbol{x} - b)$$

simplest form = "Hard Margin"
i.e. all training points correctly classified

minimize $||\boldsymbol{w}||$ subject to,

$$y_i(\boldsymbol{w}^T\boldsymbol{x_i} - b) \geq 1$$



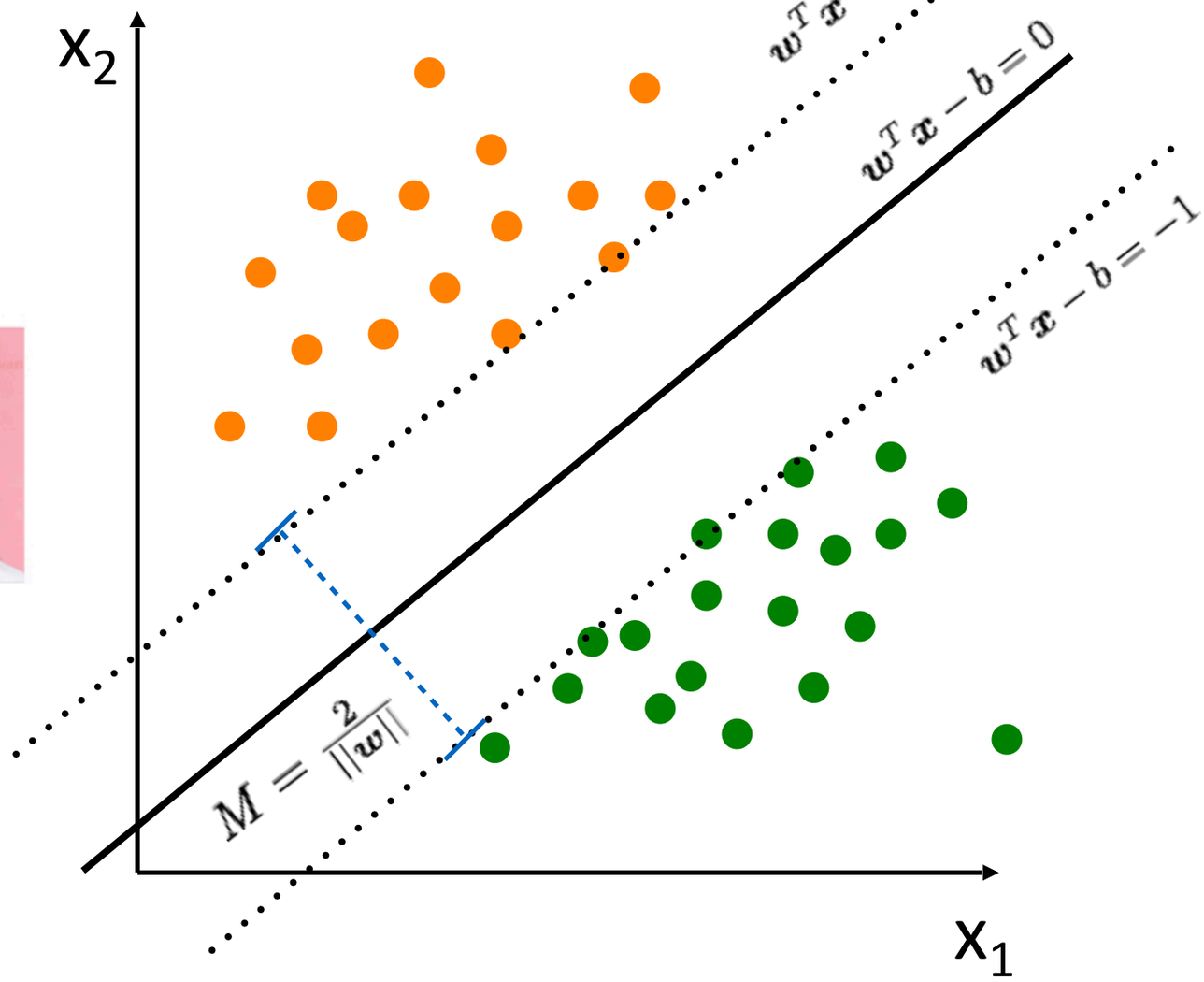
LOTS of variations on this e.g. soft margins, kernel trick for non-linear

Supervised Machine Learning

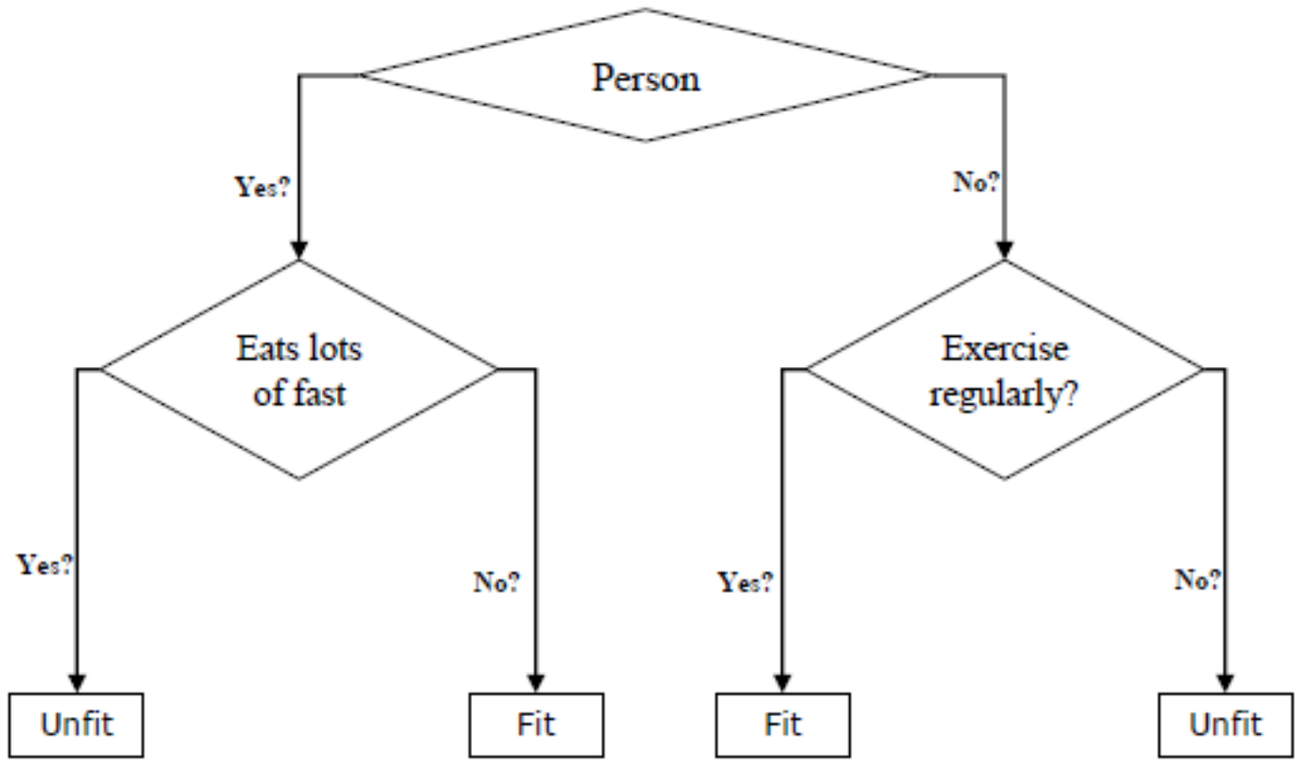
Support Vector Machines

Image recognition via SVM

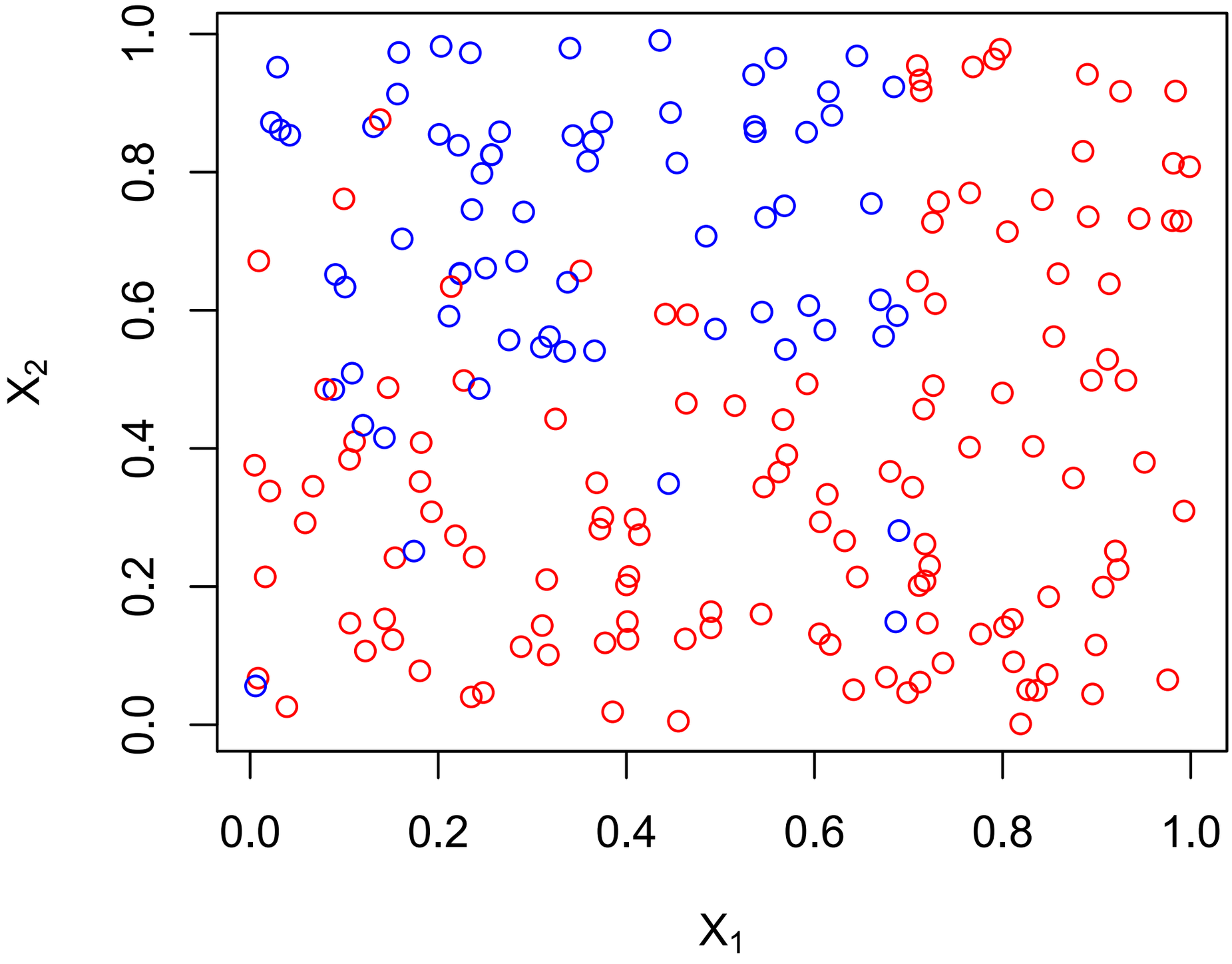
Happy Sad Surprised Angry



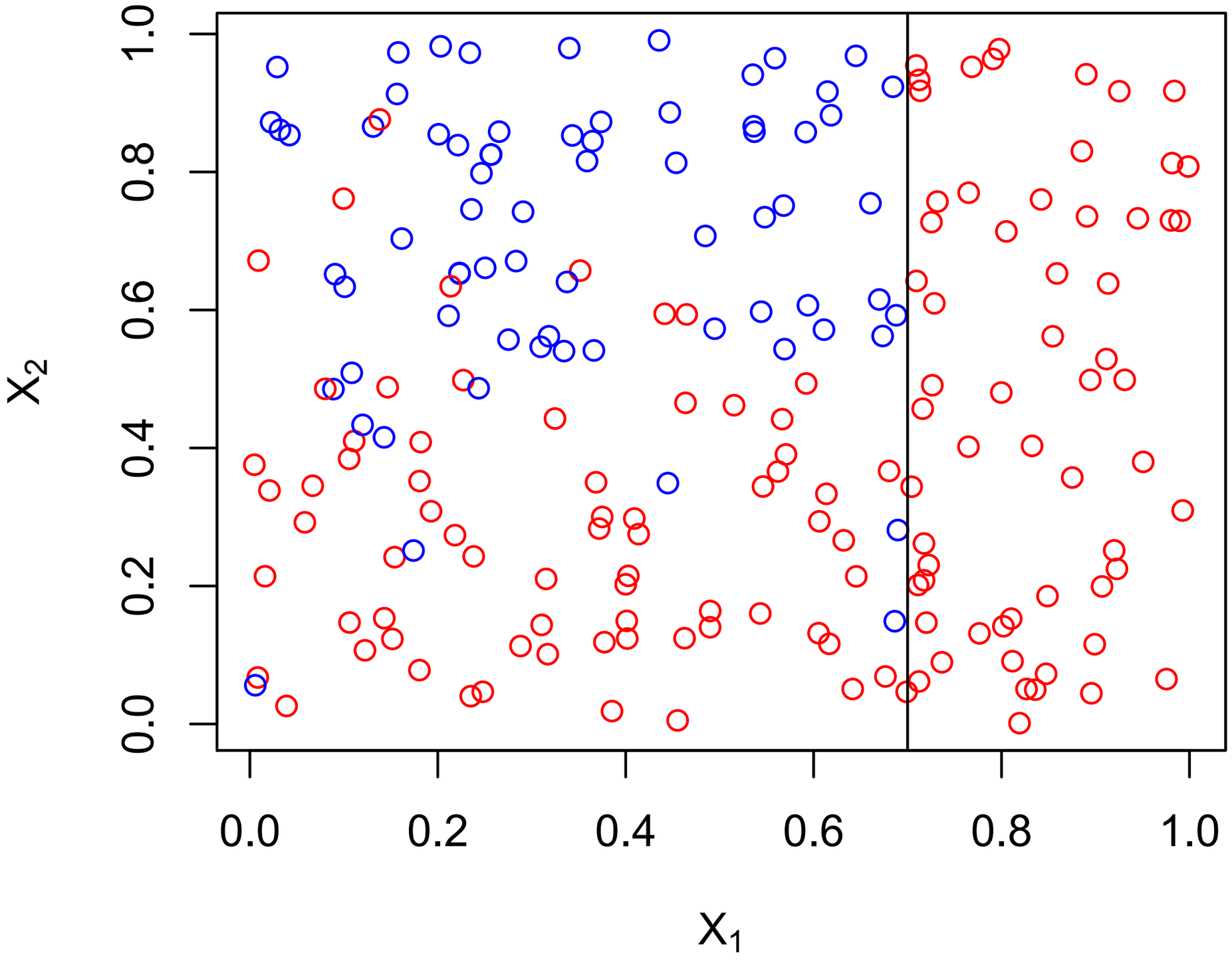
Decision Trees and Random Forests



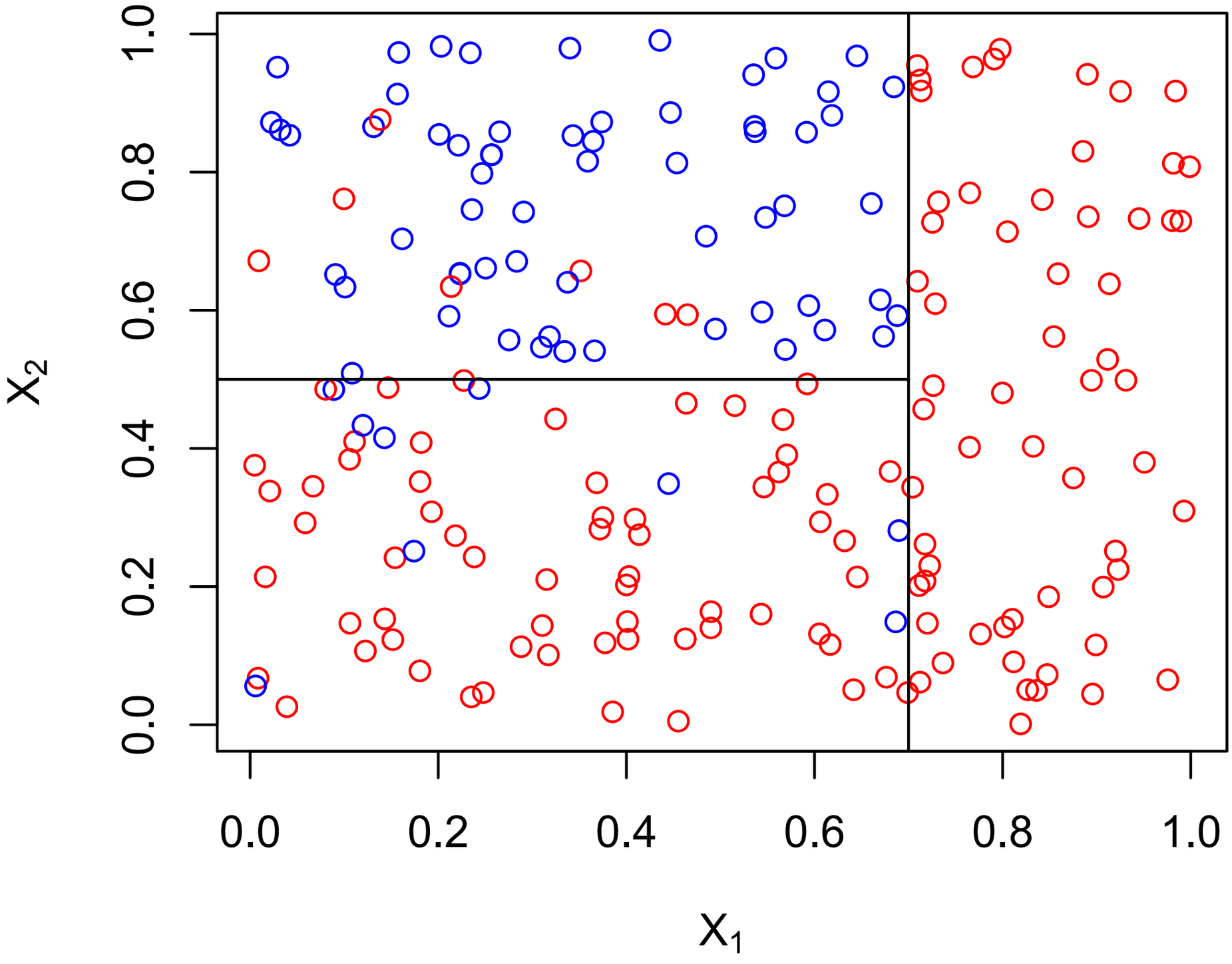
From decision trees to extra trees



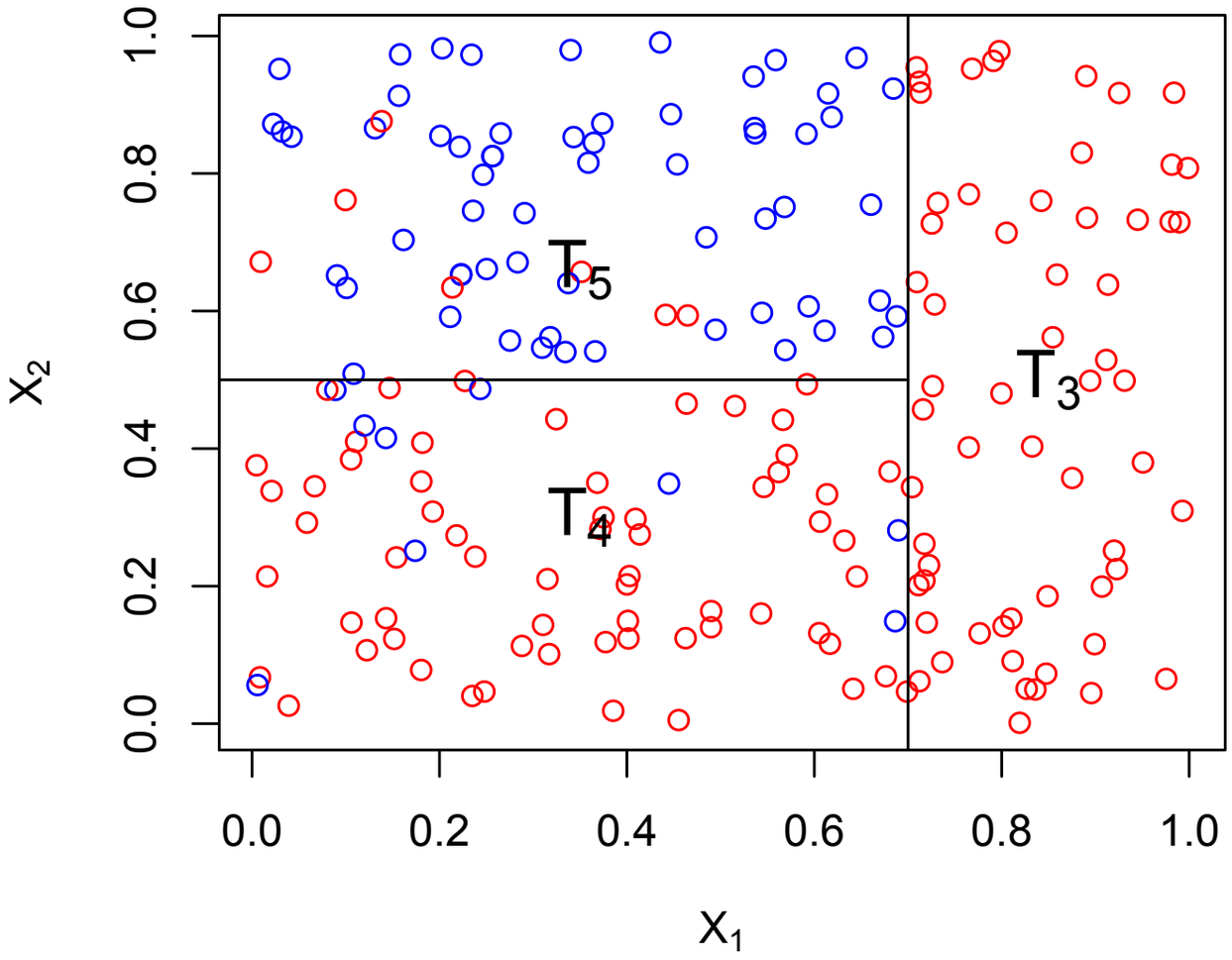
From decision trees to extra trees

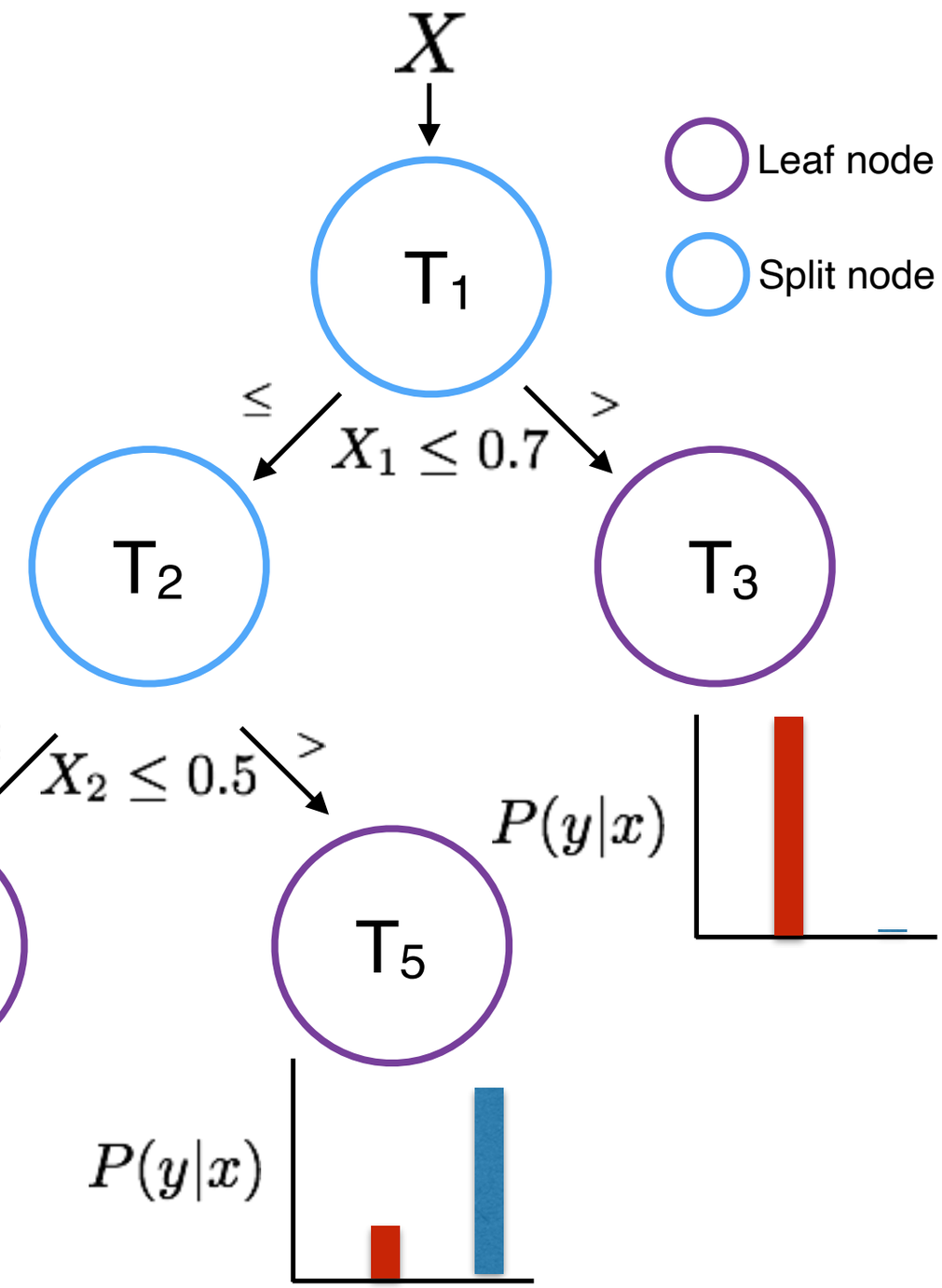


From decision trees to extra trees



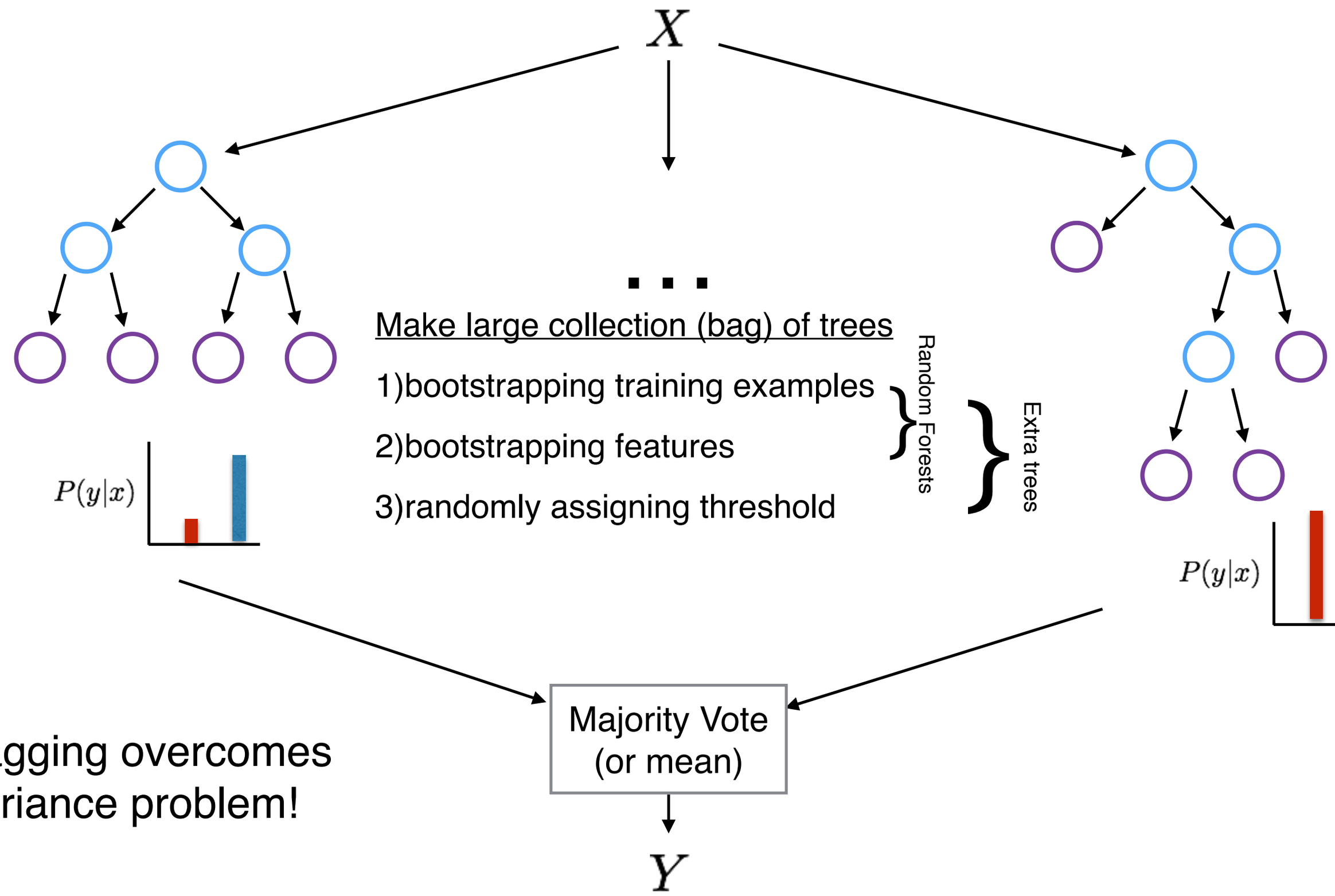
From decision trees to extra trees





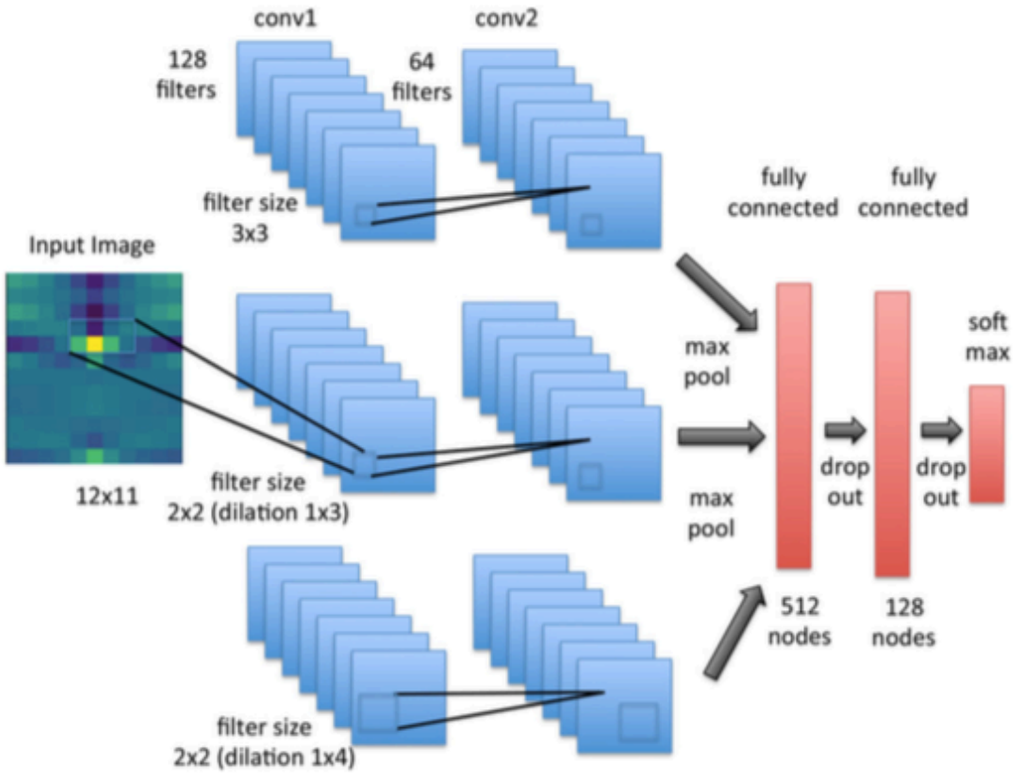
decision trees have low bias but suffer from high variance

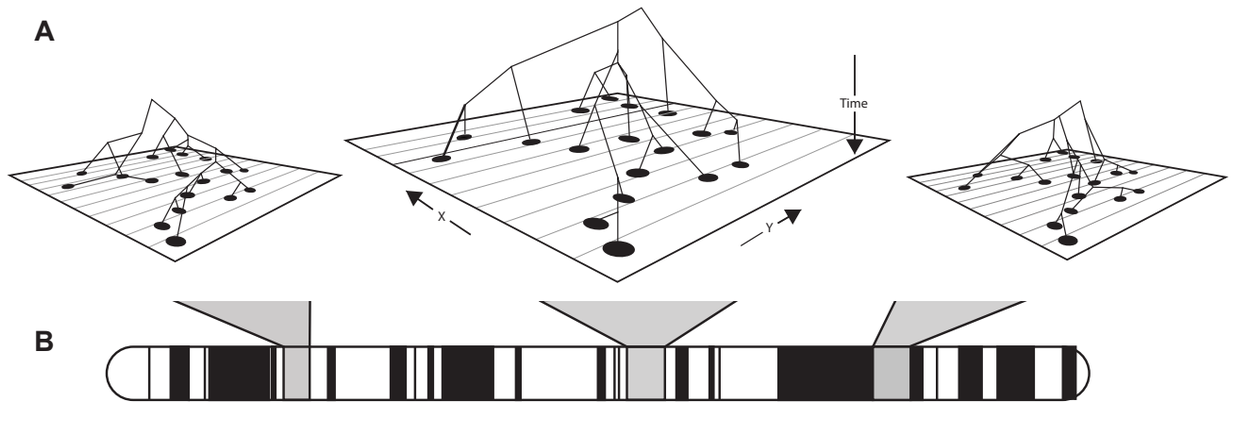
From decision trees to extra trees



bagging overcomes variance problem!

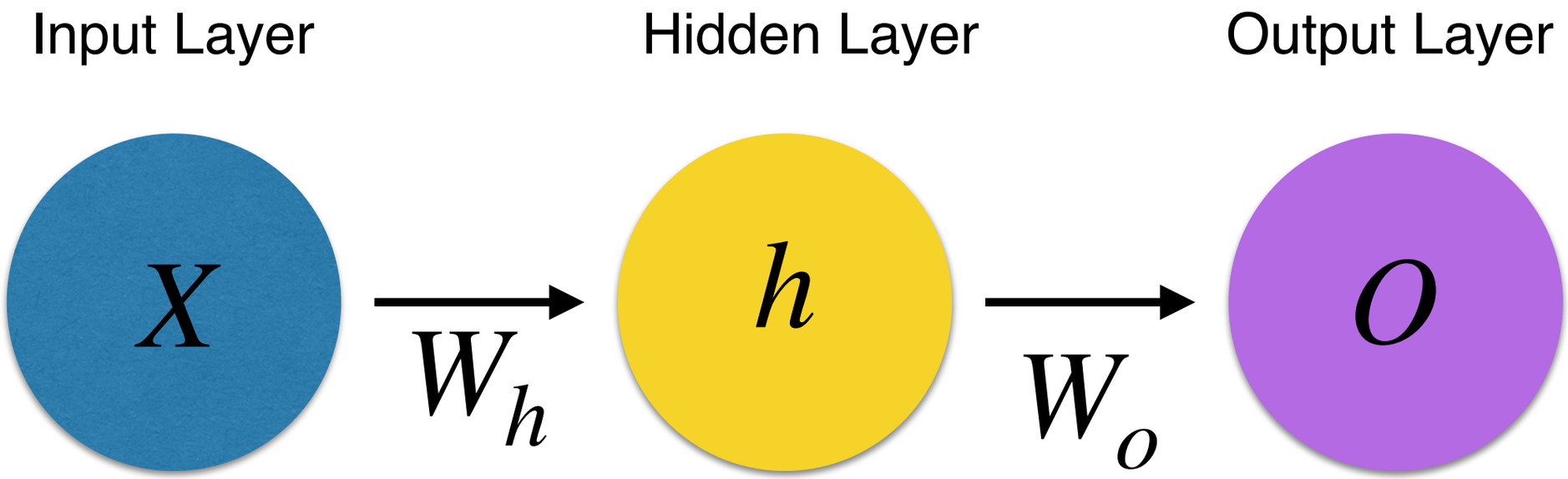
Some of Our Research





Computational Evolutionary Genetics

toy neural network



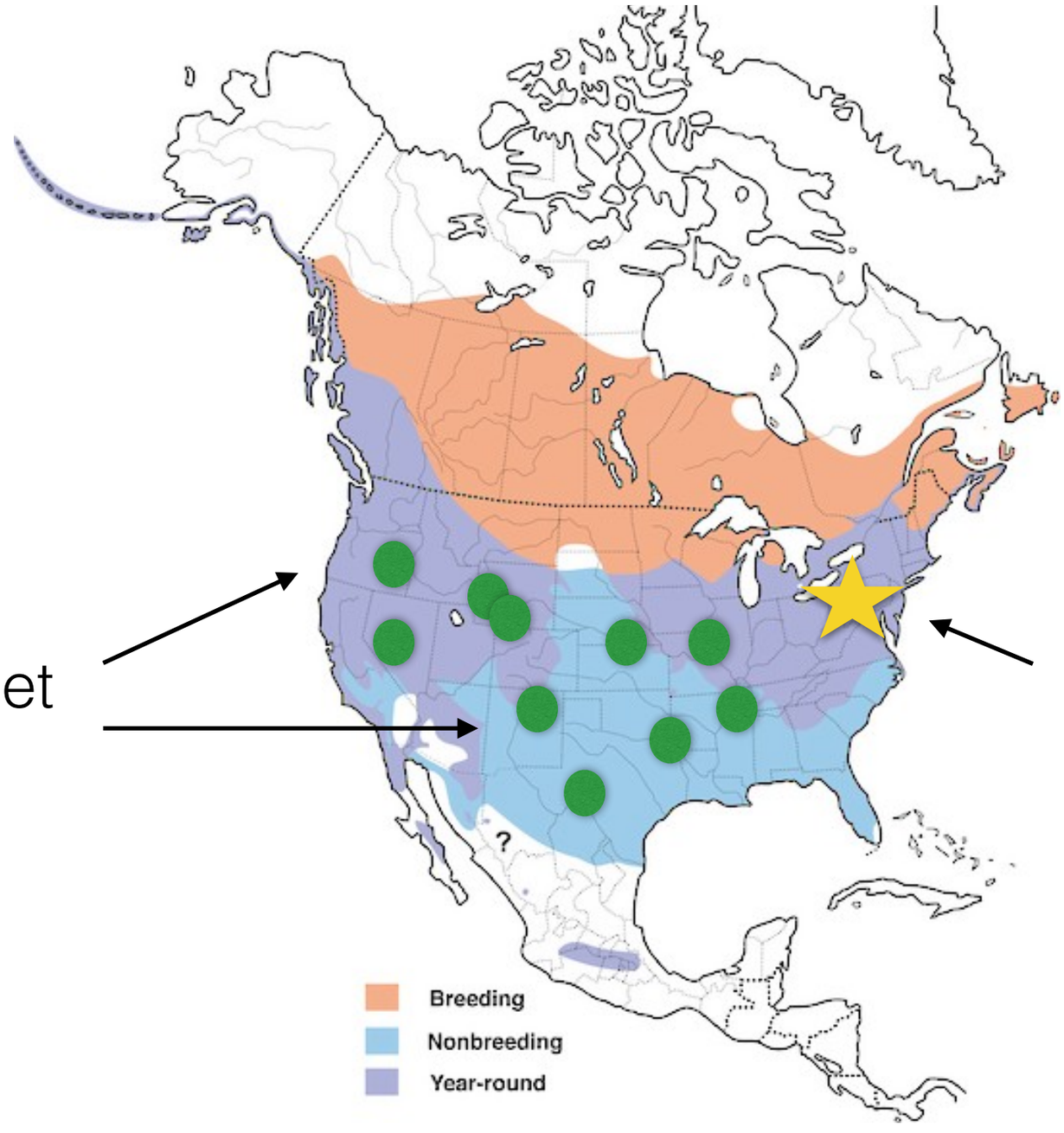
Feed-forward

$$O = f(f(X \cdot W_h + b_h) \cdot W_o + b_o)$$

think of it like stacked linear regressions

Organisms live in space

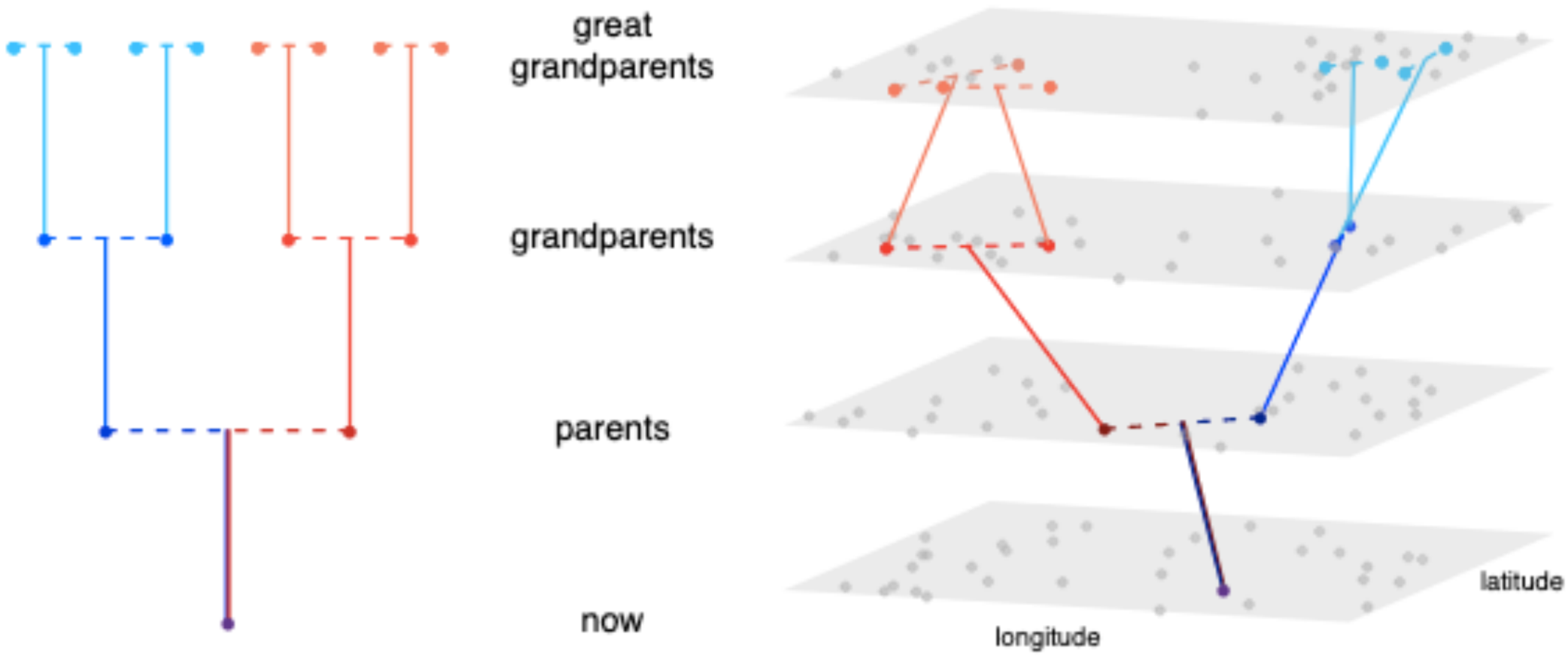
Song sparrow



Training set

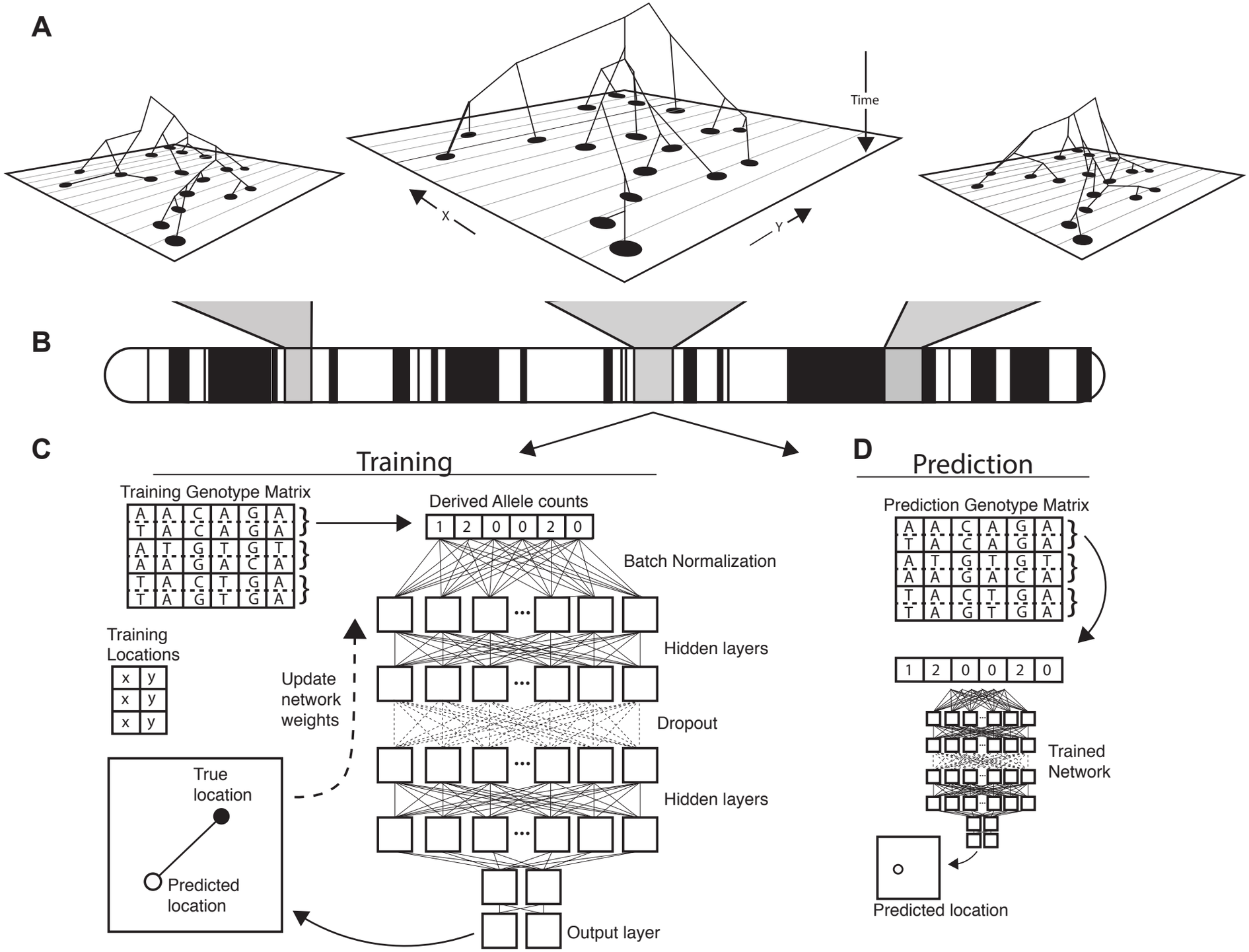
Predict locations of new genomes?

Space is the Place

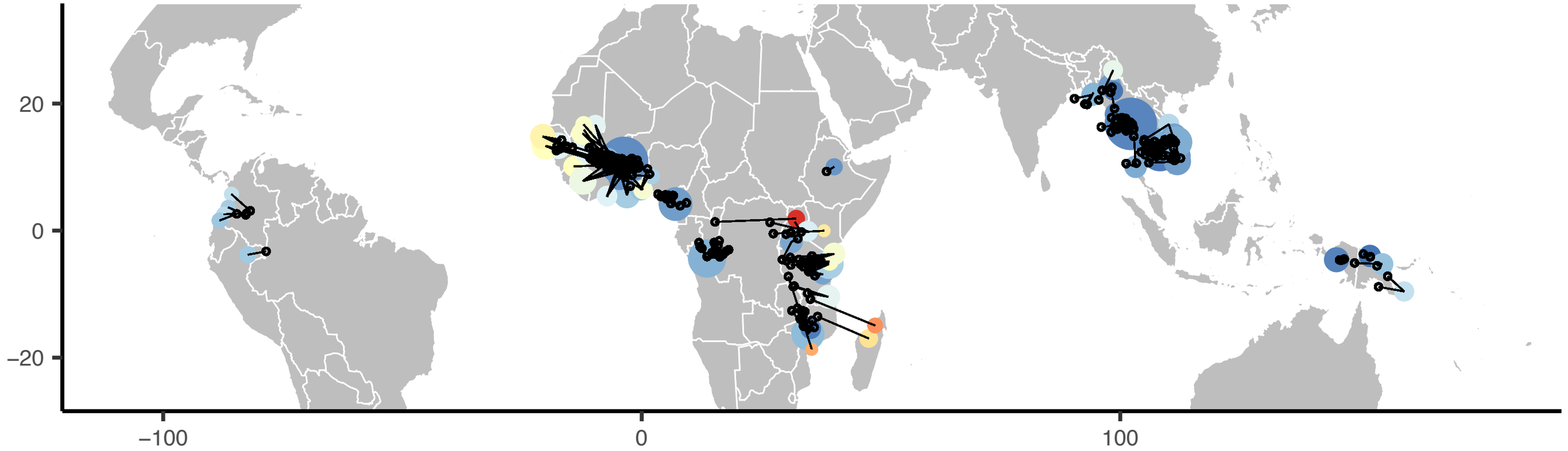


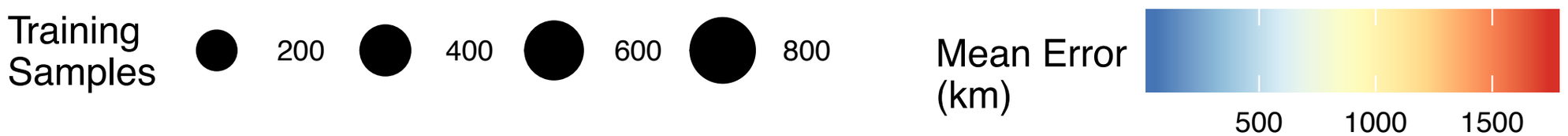
From Bradburd and Ralph (2019)

Locator— (deep) learning space



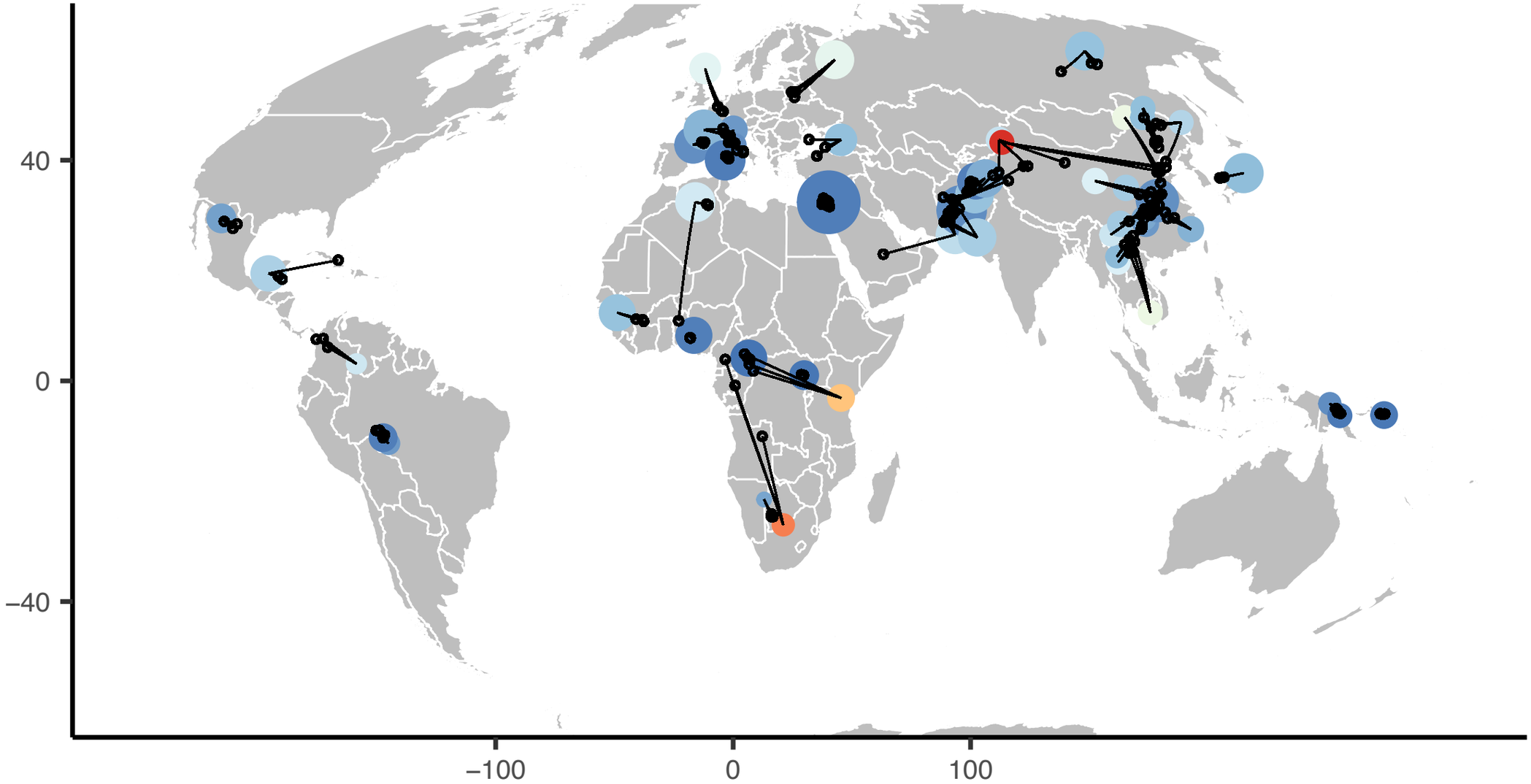
Locator— (deep) learning space





Plasmodium falciparum- Pf7K dataset
median error = 16.9 km

Locator— (deep) learning space



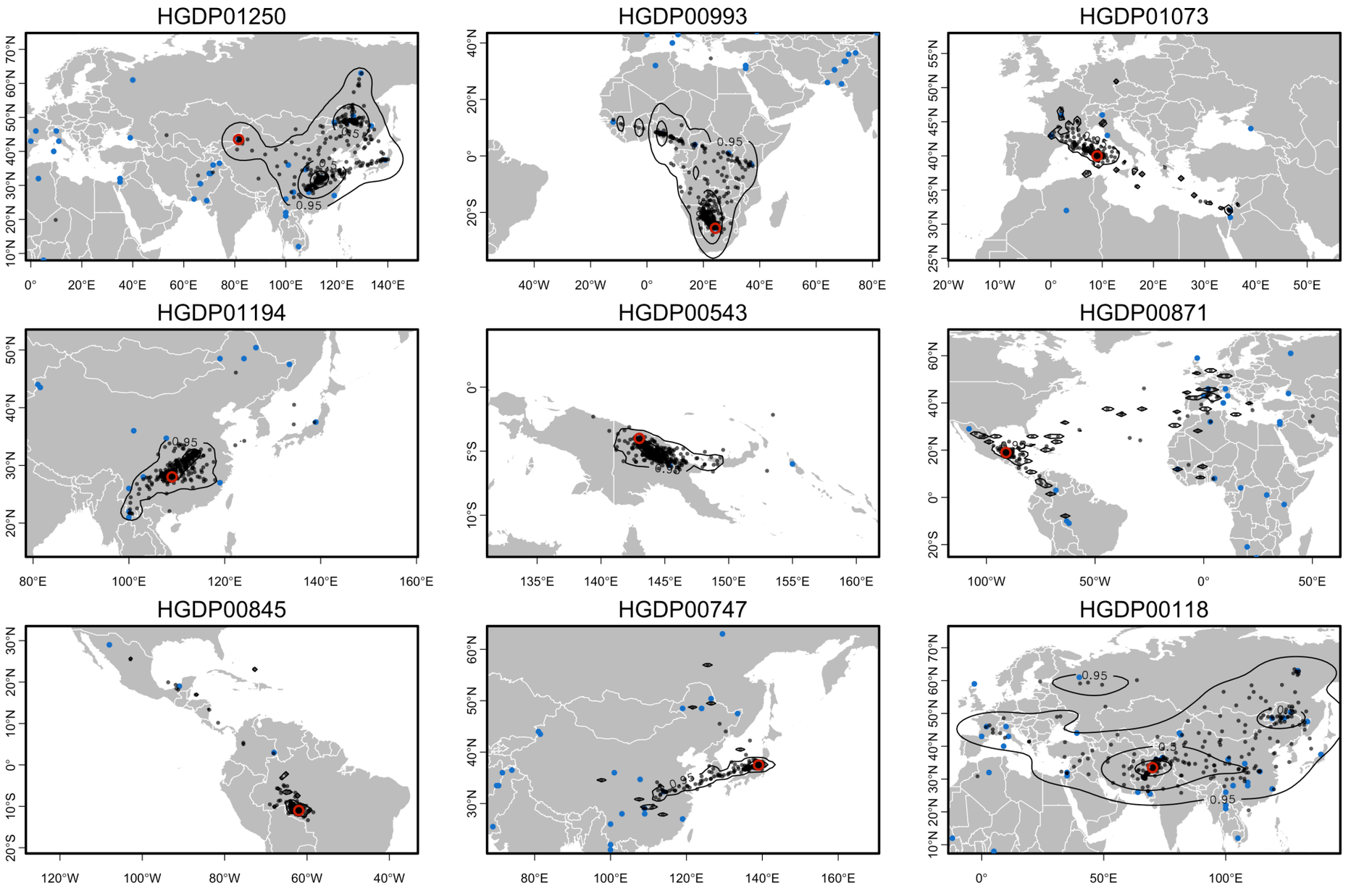
Training Samples 20 40 60 80

Mean Error (km) 1000 2000 3000

Humans - HGDP

median error = 85km

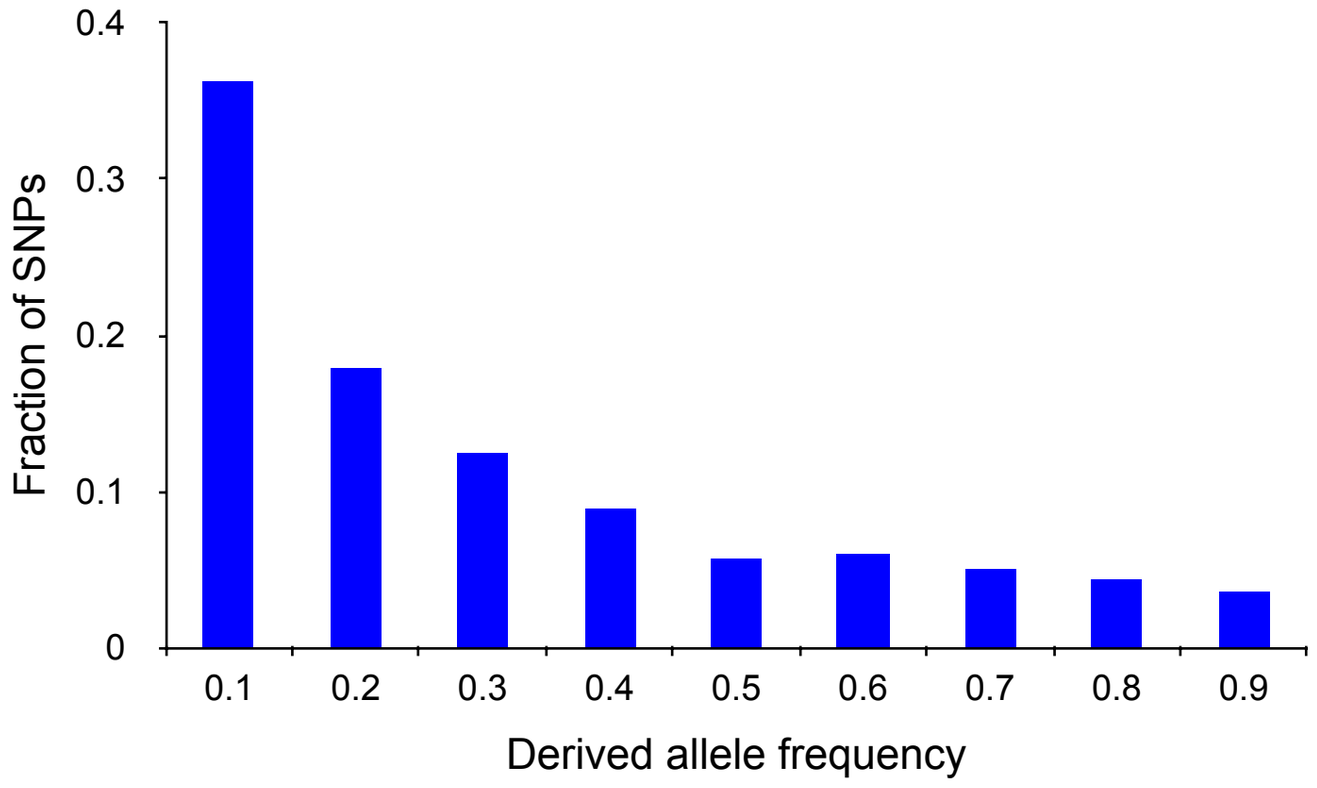
Locator— (deep) learning space



Humans - HGDP

Population Genomics Scan for Conserved Elements

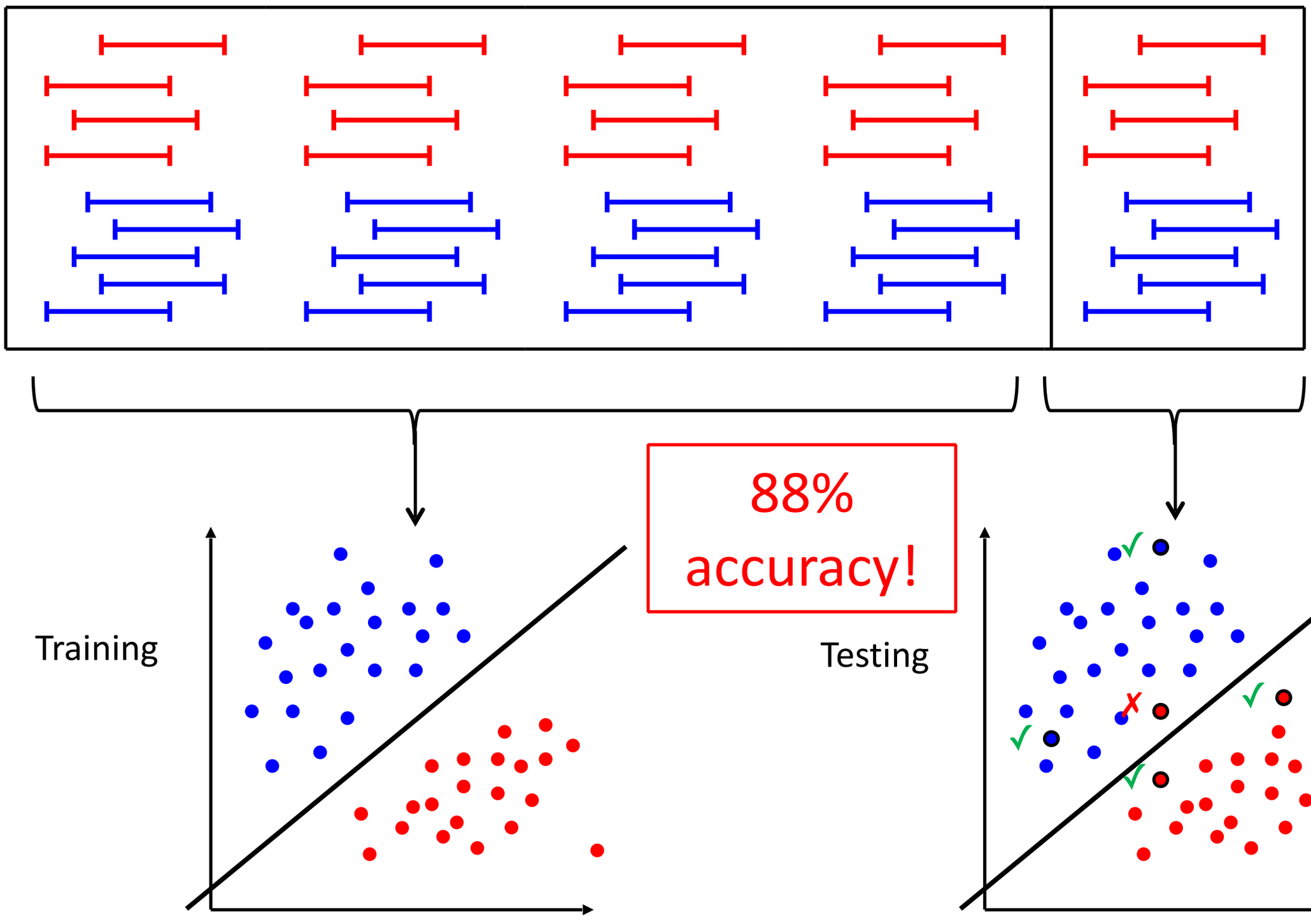
Training data



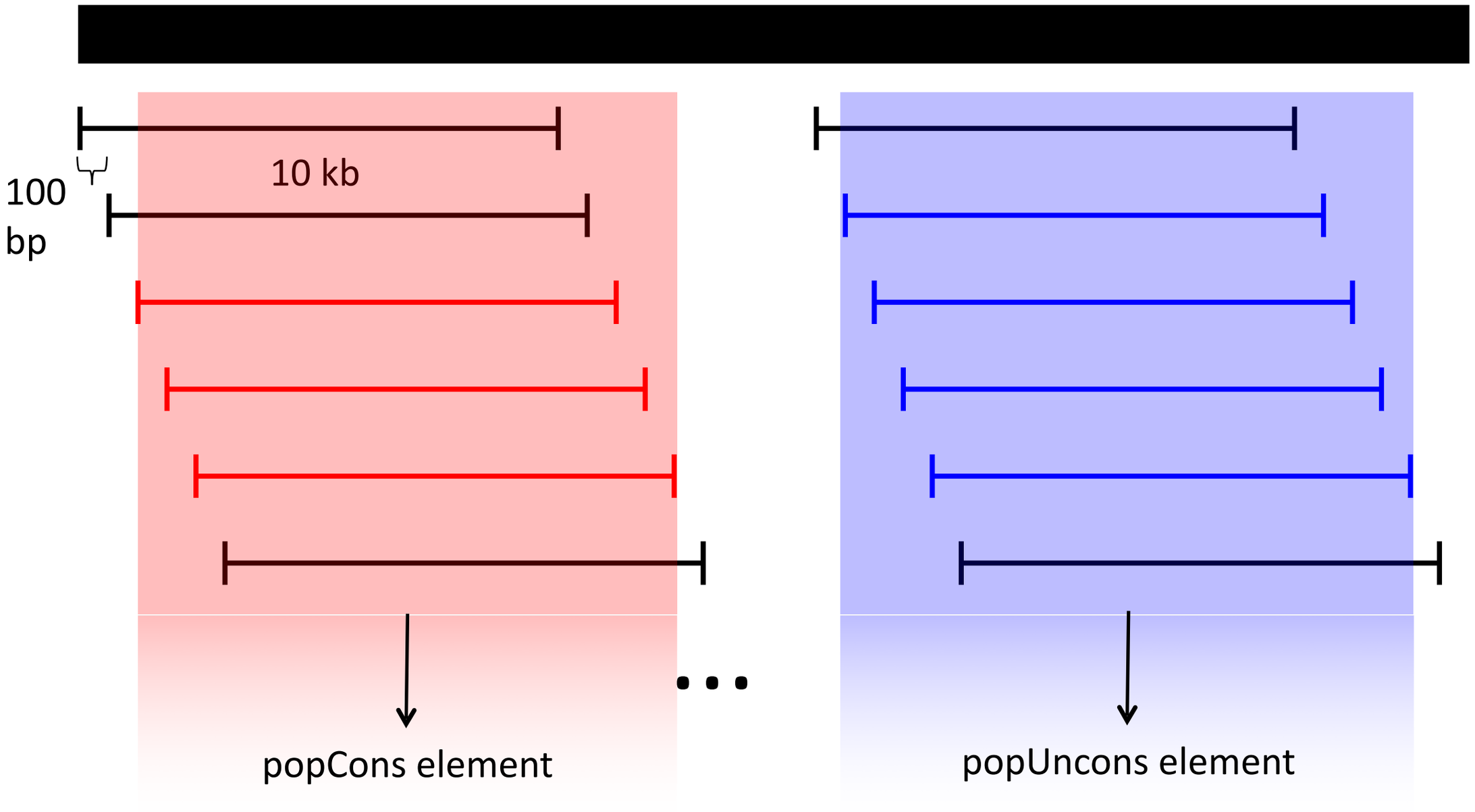
Feature vector is the SFS coarsened to 1000 bins

High accuracy on real data

cross validation



Classifying the human genome

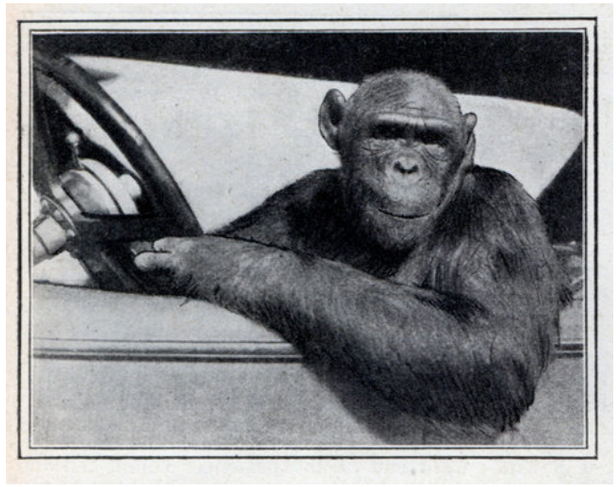


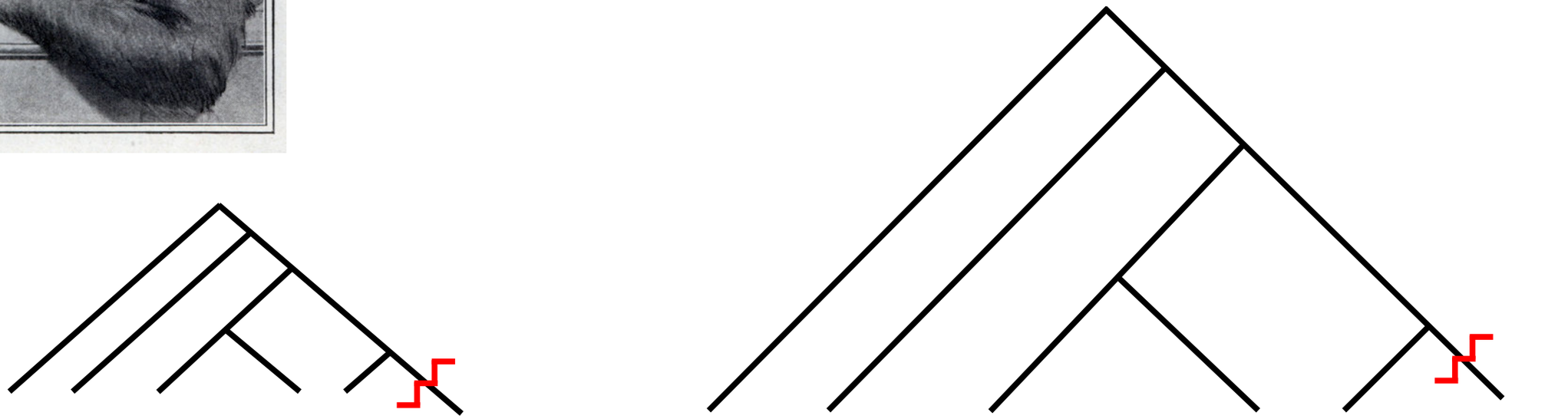
PopCons elements are functional

Enriched for:

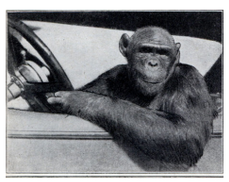
- Exons (P<0.001)
- TFBS (P<0.001)
- ORegAnno elements (P<0.001)
- miRNAs/snoRNAs (P<0.001)
- OMIM genes (P<0.001)
- Somatic mutations in cancer (P<0.001)
- Genetic Association Database genes (P<0.001)
- More . . .

Combine with Phylogenetic Info to get Gain-of-function and Loss-of-Function Predictions

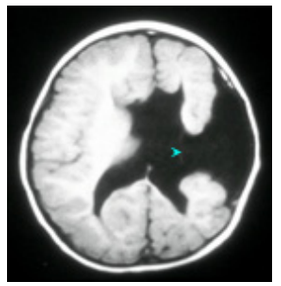




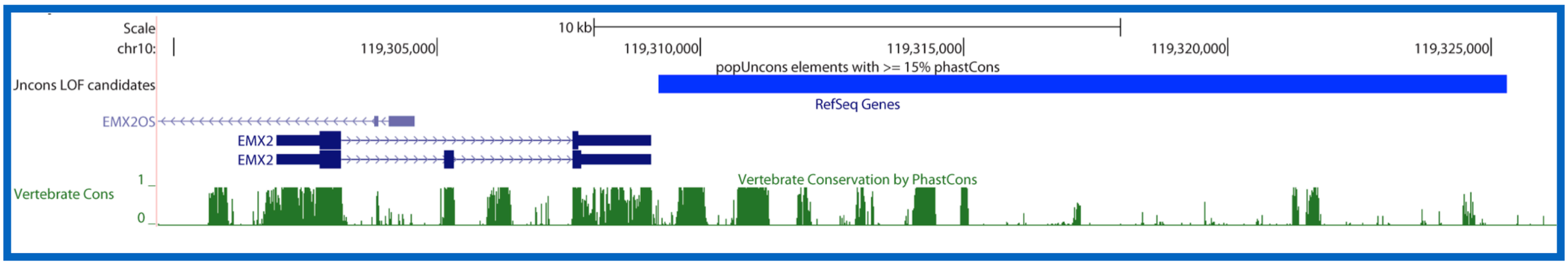
Recapitulates known examples of human-specific LOF (e.g. *MYH16* (Stedman et al. 2004))

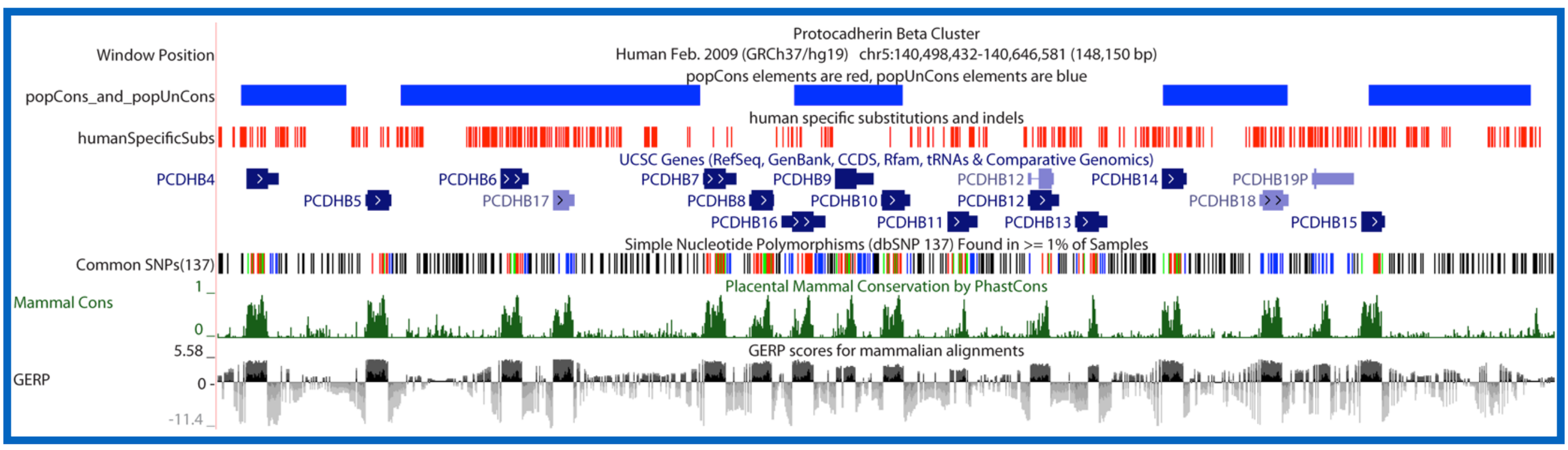


Loss-of-function examples



– *EMX2*: homeodomain tx factor gene. Required for cerebral cortex development





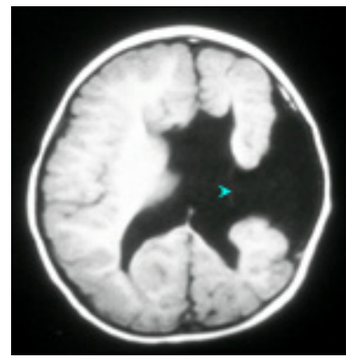
– *PCDHBs*: homeodomain tx factor gene. Cell-adhesion molecules of developing nervous system. Misexpression in Down, Rhett, and Fragile X

Loss-of-function candidates

Overwhelming associated with genes expressed in brain

Significant LOF enrichment near genes expressed in:

- Thalamus
- hypothalamus
- midbrain
- olfactory lobe
- dorsal root ganglion



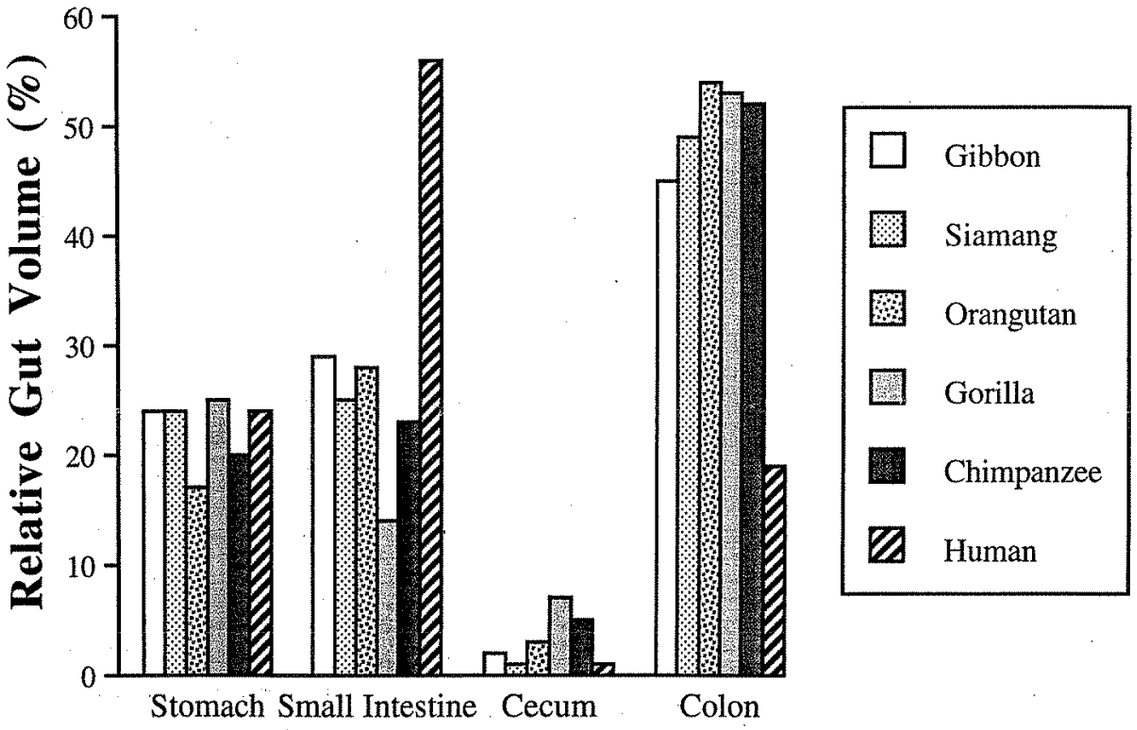
LOF associate genes associated with:

- Zinc finger domains
- C2H2-type domains
- DNA-binding domains

} Regulatory!

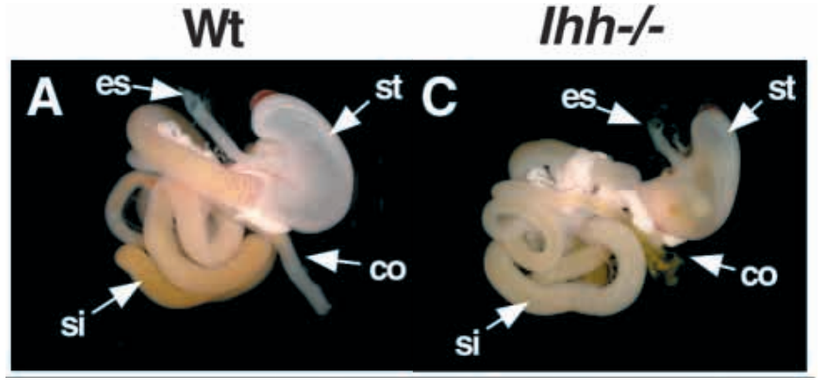
Schrider and Kern (2015)

Loss of constraint at Ihh

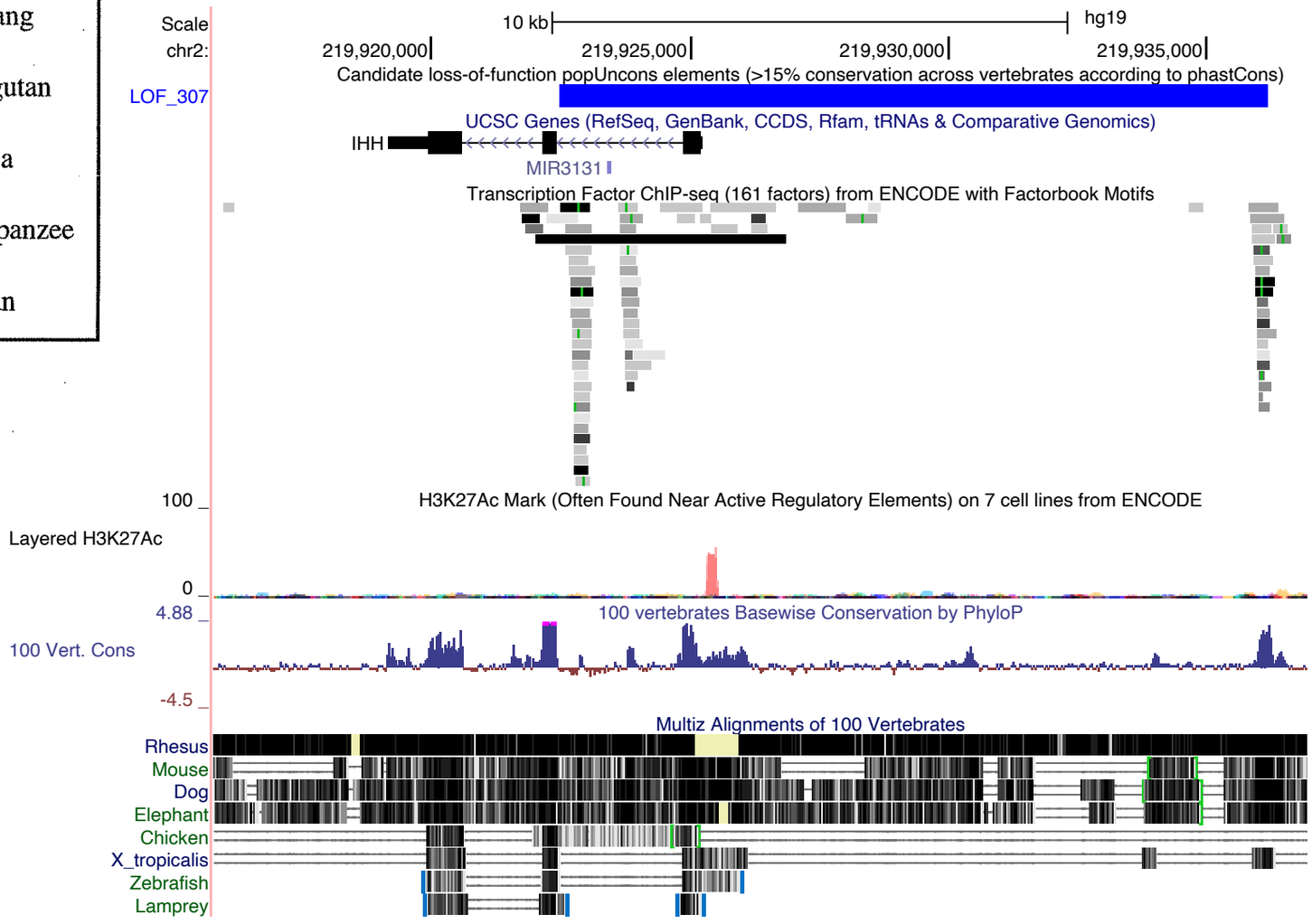


Gut Part By Species

From Milton (1999)



From Ramalho-Santos et al. (2000)



Ihh mutations known to lead to gut developmental identity transformations!
Currently collaborating on this element

The promise of machine learning

The New York Times

How Artificial Intelligence Could Transform Medicine

In “Deep Medicine,” Dr. Eric Topol looks at the ways that A.I. could improve health care, and where it might stumble.

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Life

LIFE & ARTS | IDEAS | THE SATURDAY ESSAY

The Human Promise of the AI Revolution

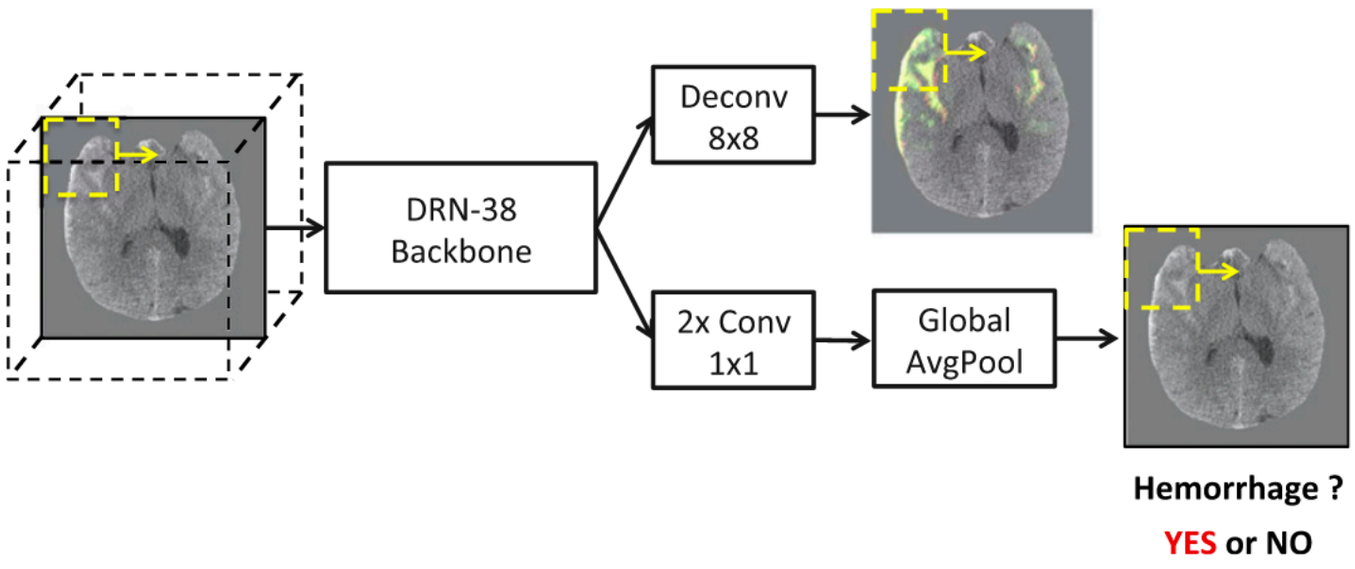
Artificial intelligence will radically disrupt the world of work, but the right policy choices contract.

The promise of machine learning

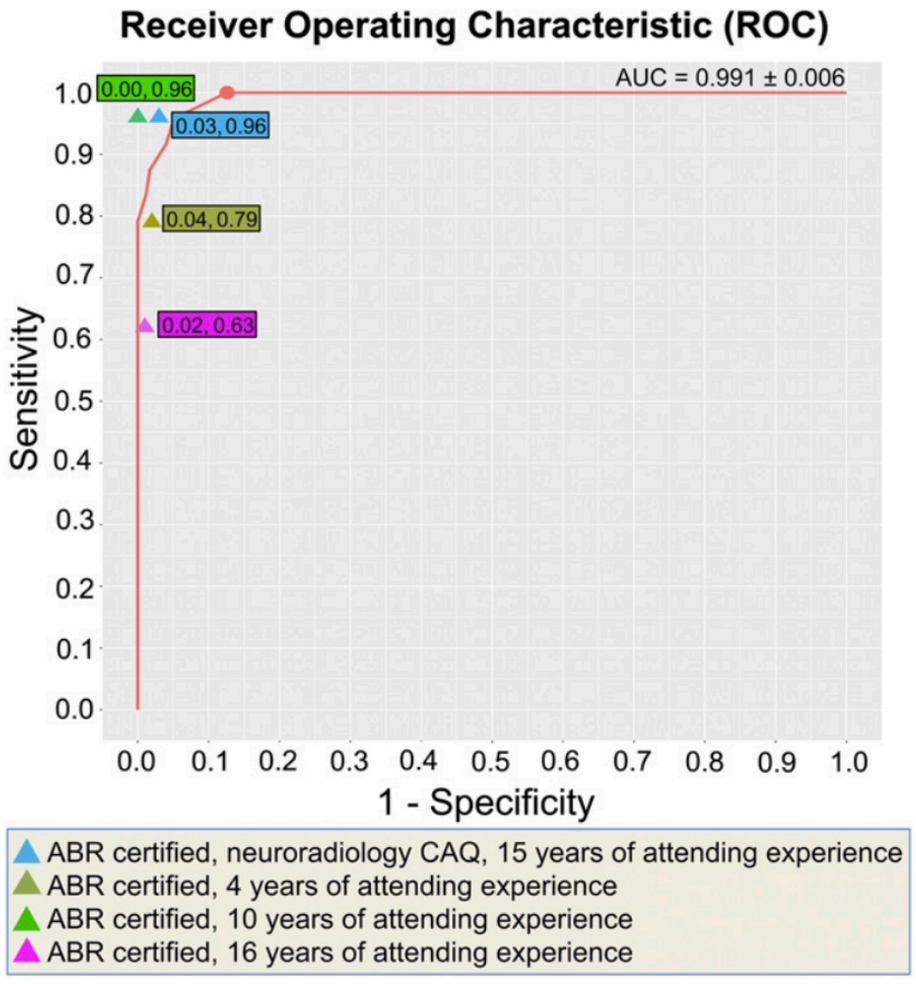
Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning

Weicheng Kuo[a], Christian Häne[a], Pratik Mukherjee[b], Jitendra Malik[a,1], and Esther L. Yuh[b,1]

[a]Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720; and [b]Department of Radiology and Biomedical Imaging,



The promise of machine learning



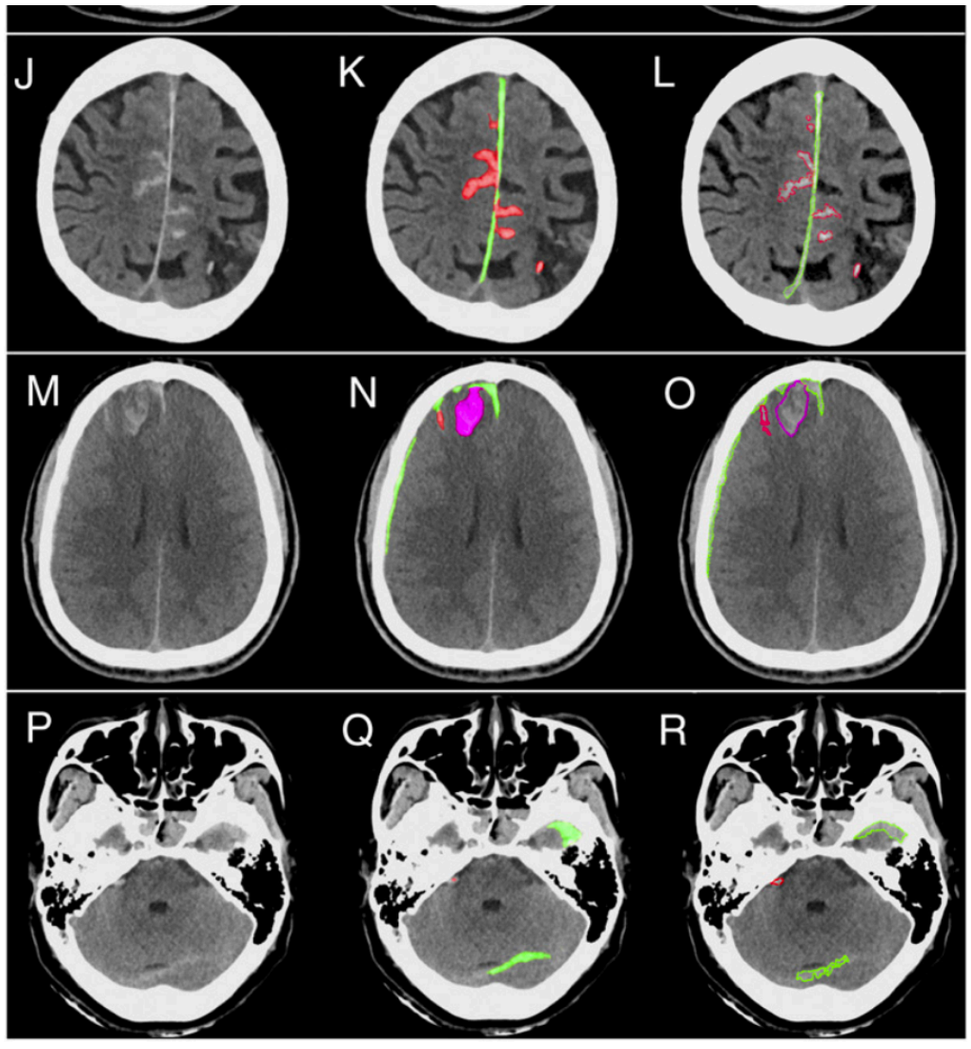


Fig. 5. Examples of multiclass segmentation by the algorithm and by an expert. (*A–C*) Small left holohemispheric subdural hematoma (SDH, green) and adjacent contusion (purple). (*D–F*) Small right frontal and posterior parafalcine SDH and anterior interhemispheric fissure SAH (red). (*G–I*) Small bilateral tentorial and left frontotemporal SDH (green) and subjacent contusions (purple) and SAH (red), in addition to shear injury in the left cerebral peduncle (purple). (*J–L*) Small parafalcine SDH (green) with surrounding SAH (red). (*M–O*) Several small right frontal areas of SDH (green) with subjacent contusion (purple) and SAH (red). (*P–R*) Small left tentorial and left anterior temporal SDH (green) and right cerebellopontine angle SAH (red). (*A*, *D*, *G*, *J*, *M*, and *P*) Original images. (*B*, *E*, *H*, *K*, *N*, and *Q*) Algorithmic delineation of hemorrhage with pixel-level probabilities >0.5 colored in red (SAH), green (SDH), and contusion/shear injury (purple). (*C*, *F*, *I*, *L*, *O*, and *R*) Neuroradiologist segmentation of hemorrhage.

Common pitfalls of machine learning

1. Not enough data

Common pitfalls of machine learning

Facial Recognition Is Accurate, if You're a White Guy

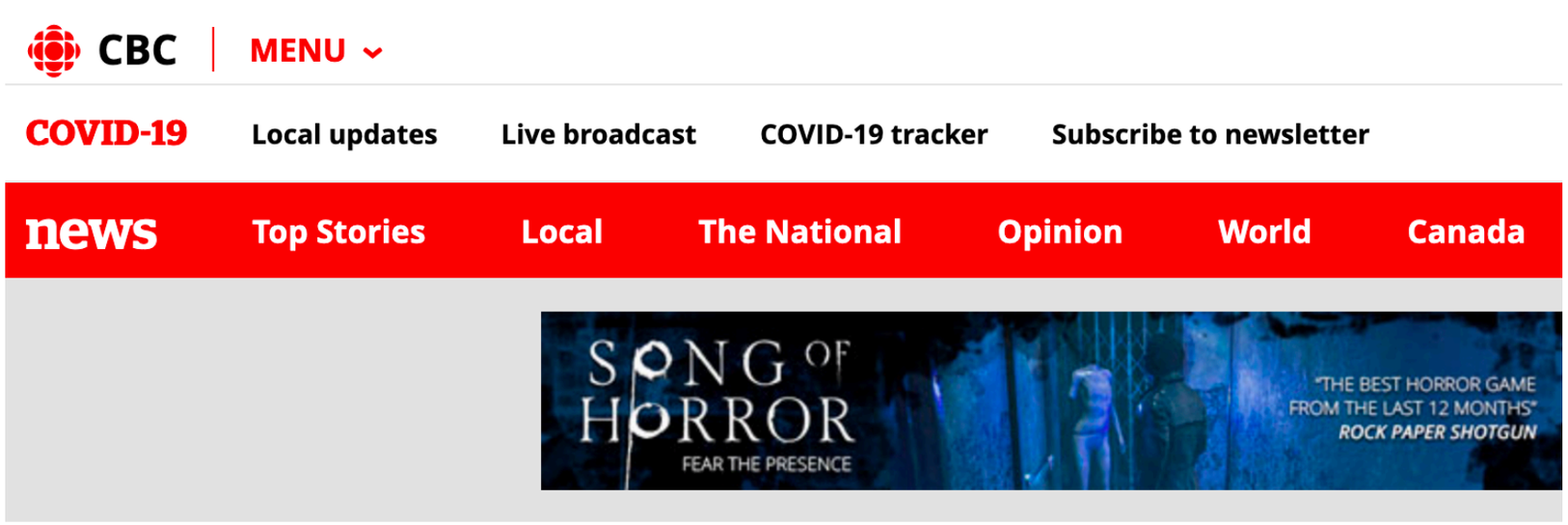
Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

---

2. Biases in the training set

Common pitfalls of machine learning



2. Biases in the training set

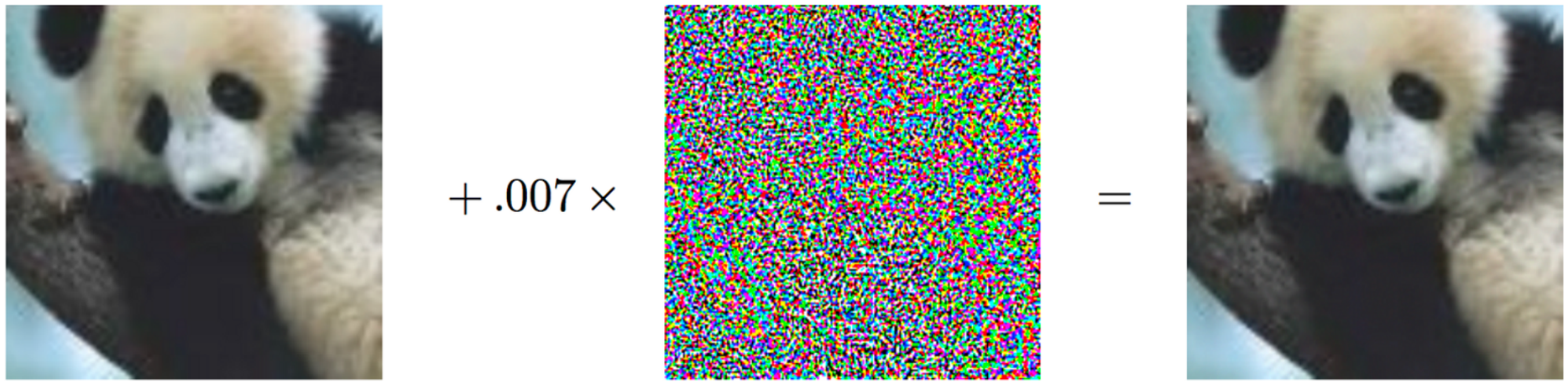
Common pitfalls of machine learning

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

---

3. Out of sample prediction doesn’t work well

Common pitfalls of machine learning



“panda”

57.7% confidence

noise

“gibbon”

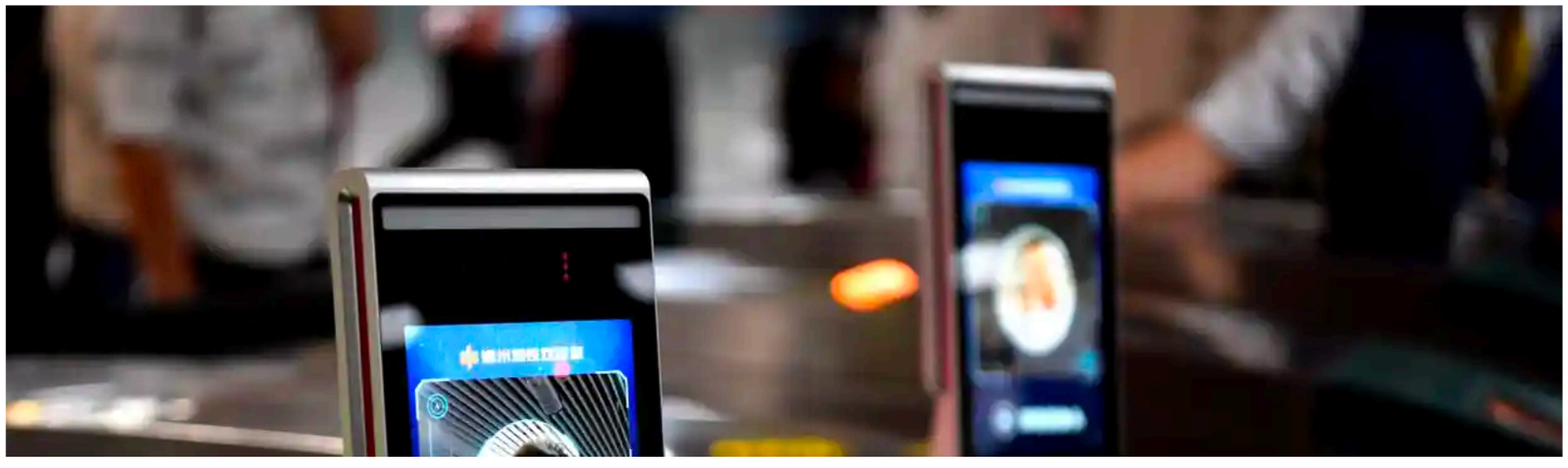
99.3% confidence

4. Fragile classifiers

Maybe not all good?

China brings in mandatory facial recognition for mobile phone users

Ministry claims change will 'protect the legitimate rights and interest of citizens in cyberspace' but critics say it's dystopian

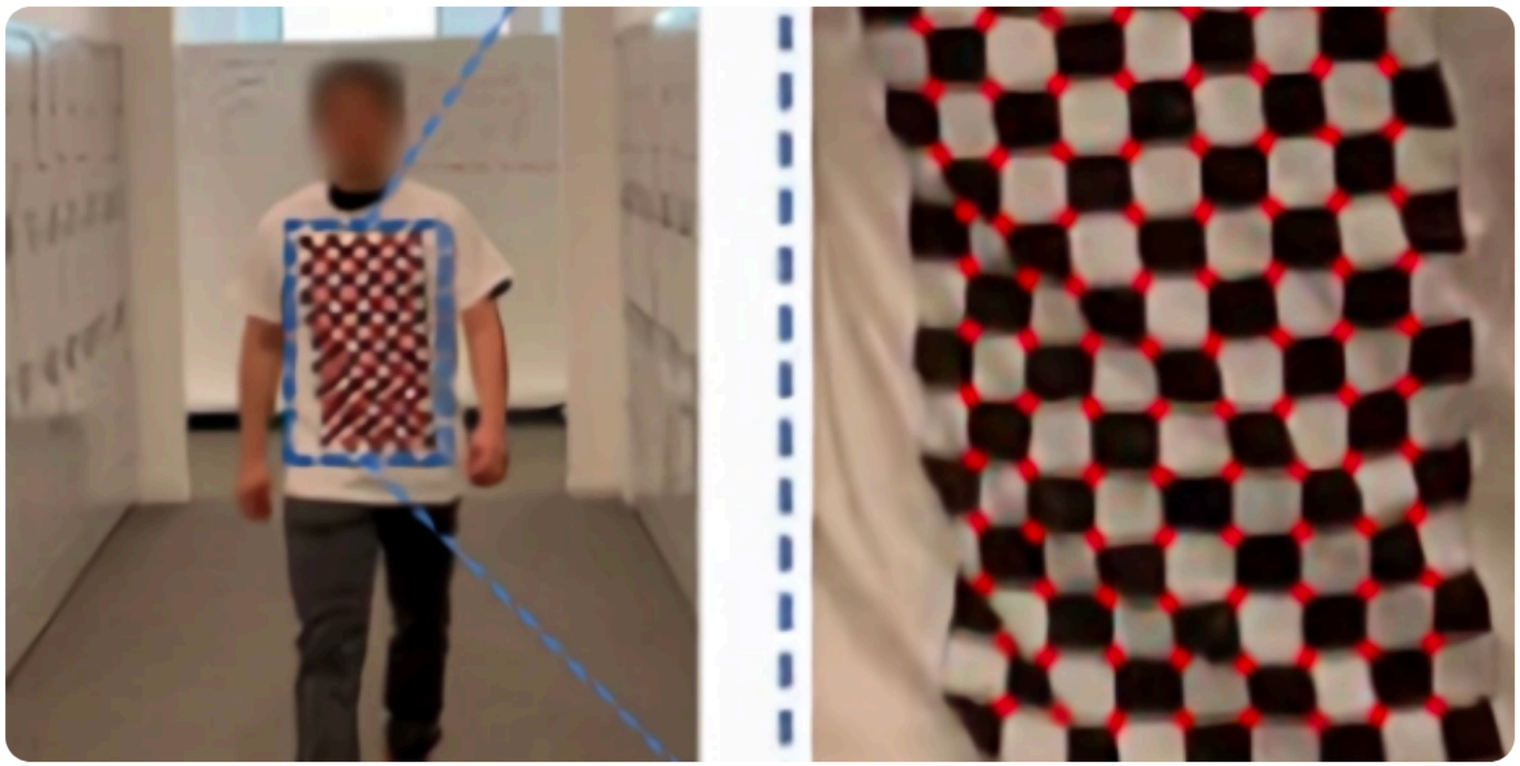


huge potential societal impacts

Maybe not all good?

Researchers foil people-detecting AI with an ‘adversarial’ T-shirt

KYLE WIGGERS @KYLE_L_WIGGERS OCTOBER 29, 2019 7:59 AM



huge potential societal impacts

Maybe not all good?

Search

Cart (0) Check Out

The patterns on the goods in this shop are designed to trigger Automated License Plate Readers, injecting junk data in to the systems used by the State and its contractors to monitor and track civilians and their locations.

Home

All Items

Shirts

Hoodies & Jackets

Skirts & Dresses

Backpacks

European Union

Graphic T-Shirts &

Featured collection

huge potential societal impacts