Evomics Machine Learning

Andrew Kern
University of Oregon

Leo Breiman's Two Cultures

the logic of data analysis

L. Breiman, Statistical Science (2001)

Leo Breiman's Two Cultures

Data Modeling Culture

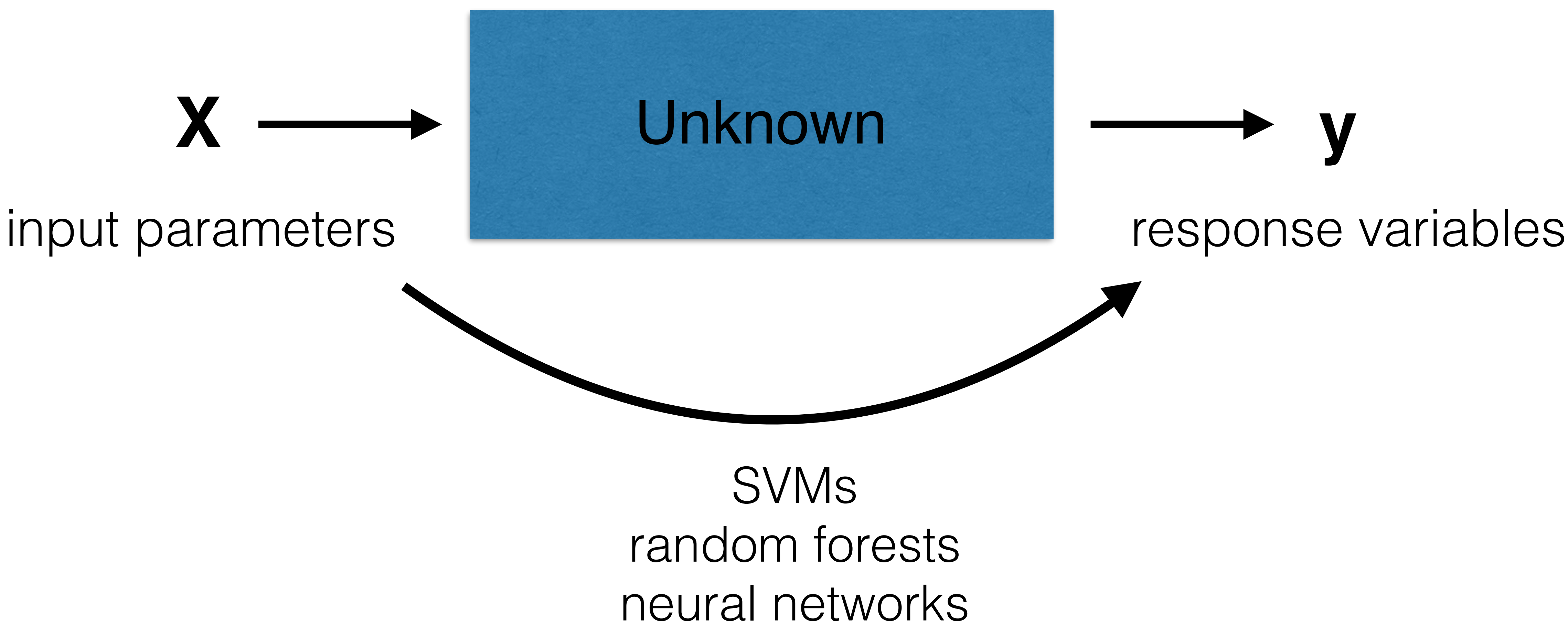
e.g. linear regression

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

98% of Statistics

Leo Breiman's Two Cultures

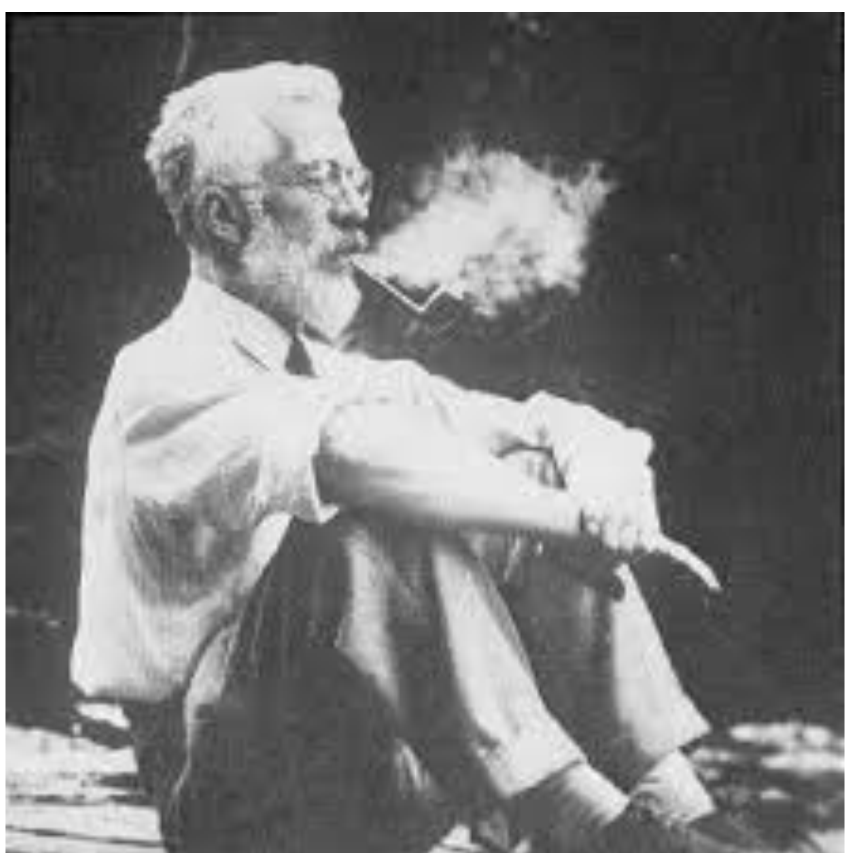
Algorithmic Modeling Culture
(machine learning)

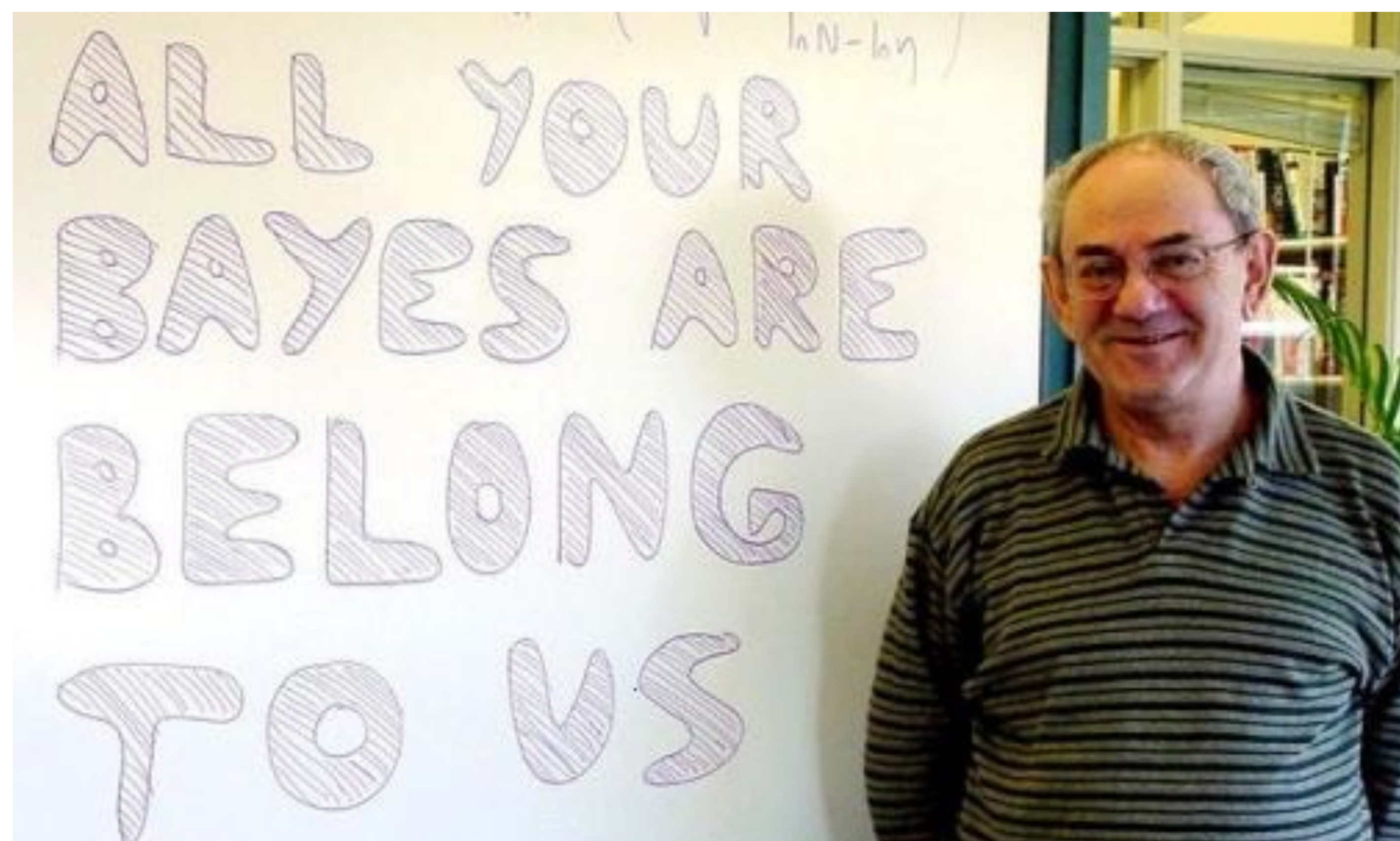


Ignore probabilistic generative model f(x)-> y

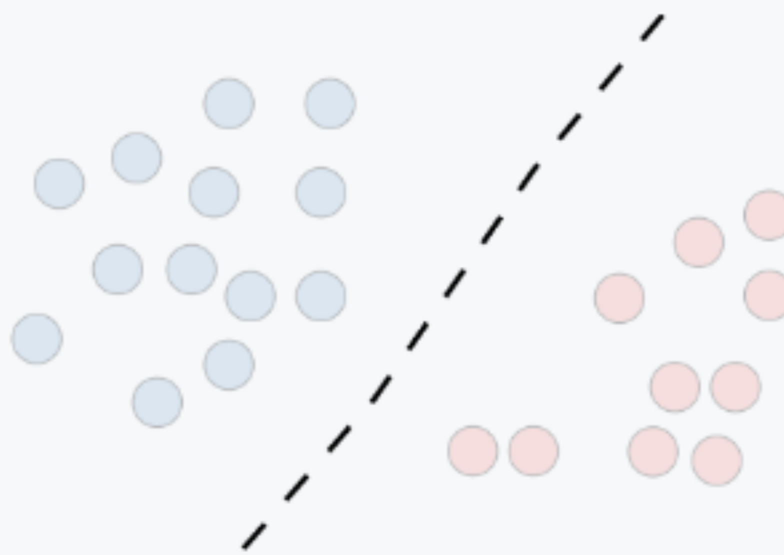
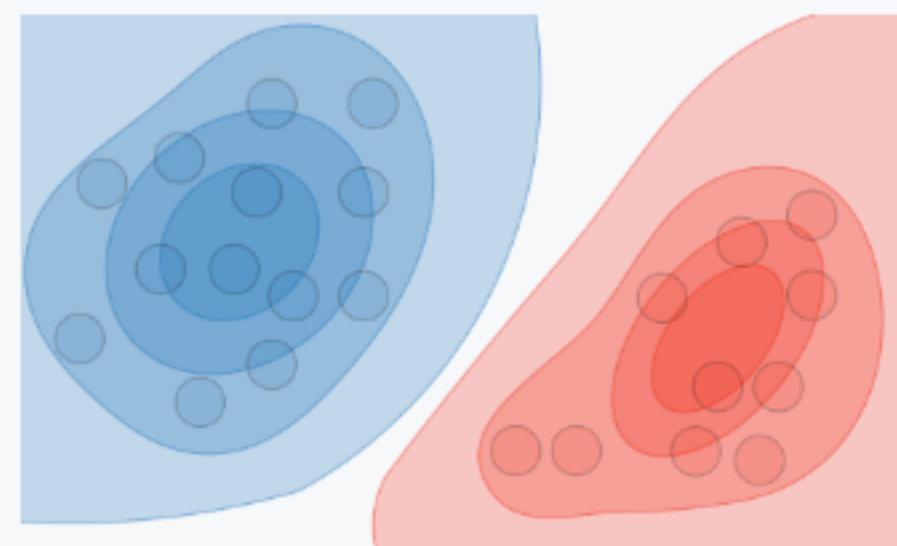
Machine Learning!

These guys don't
have generative model

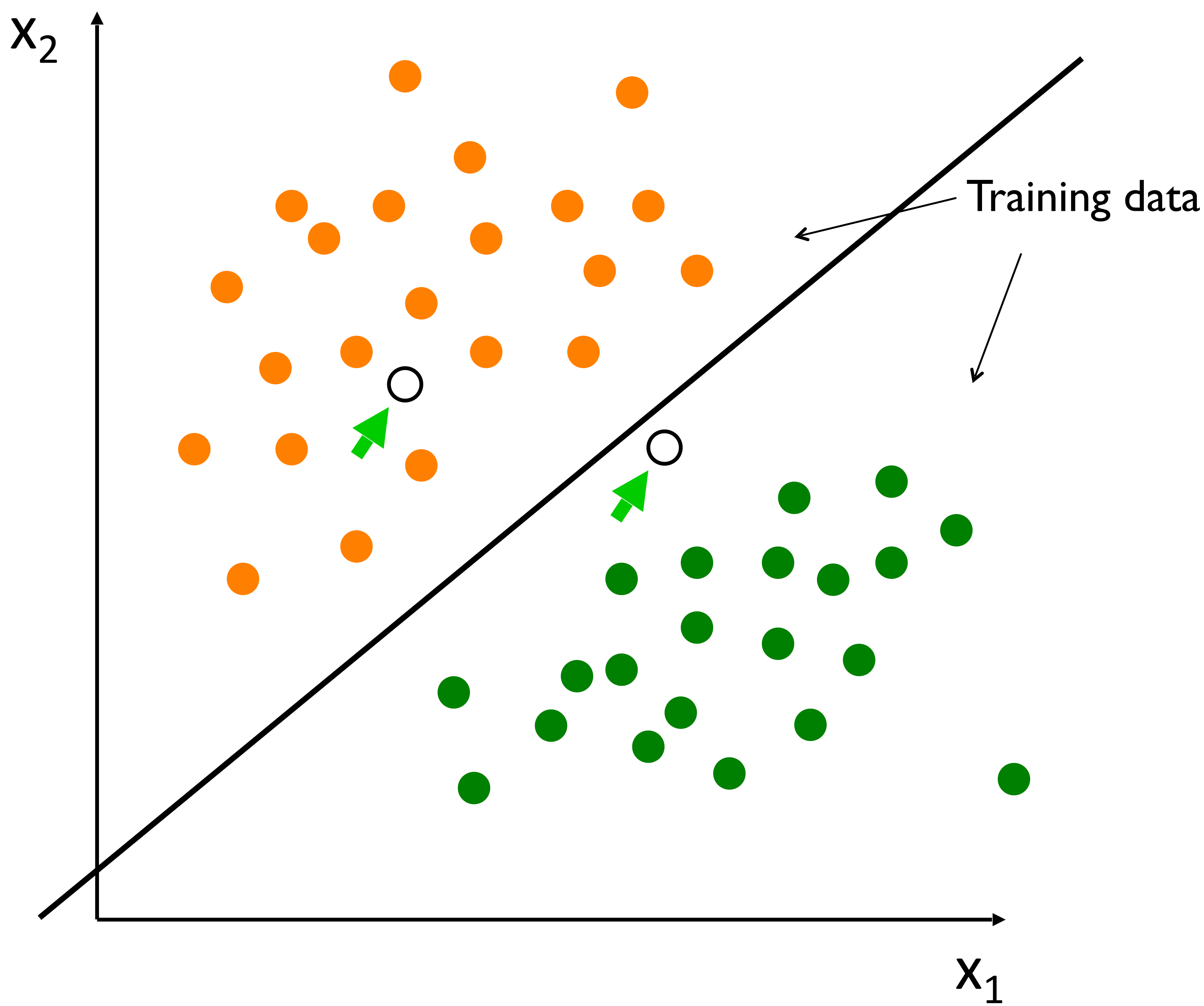




Discriminitive vs Generative Models

	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:

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

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:

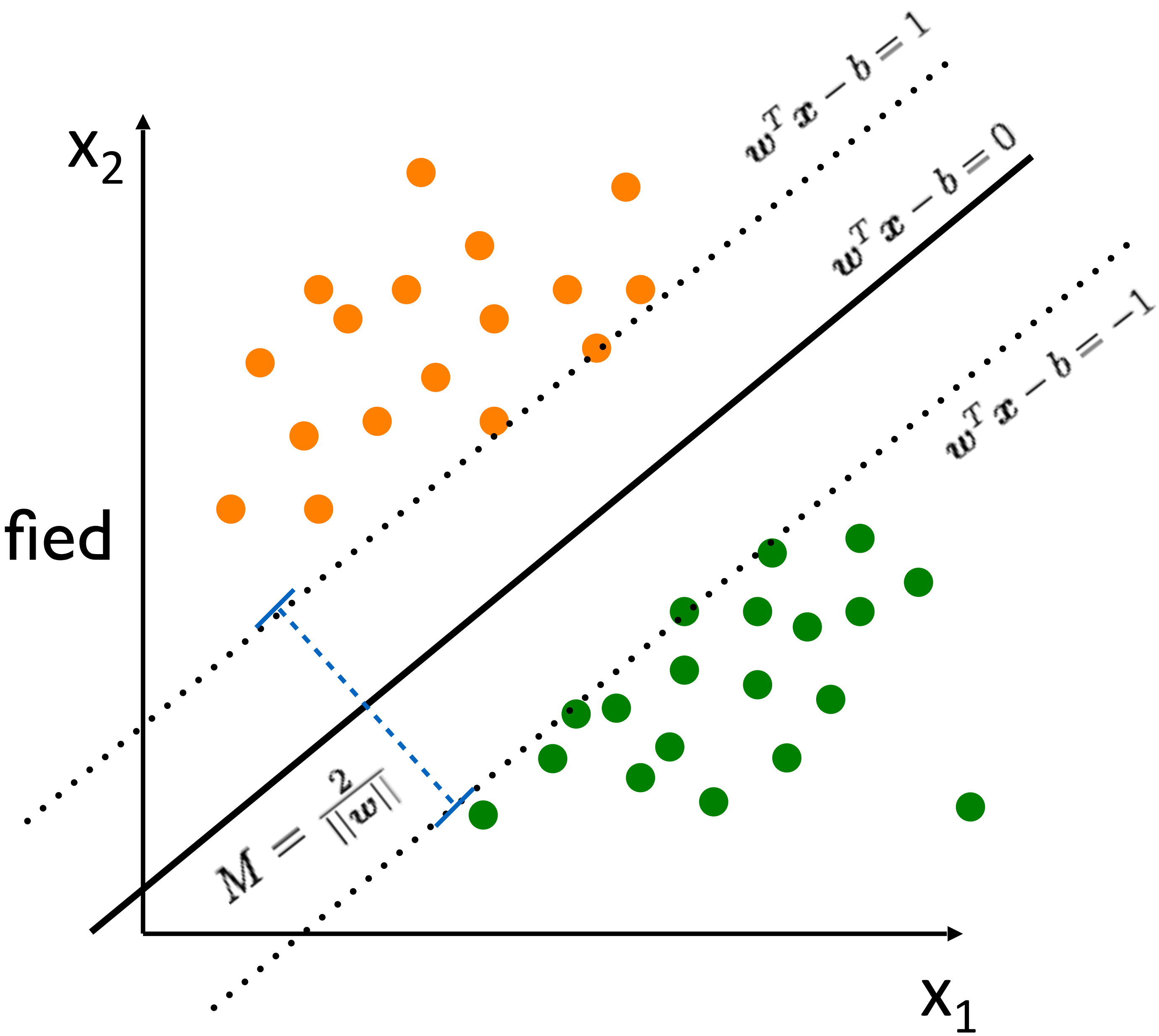
$$g(\boldsymbol{x}) = sign(\boldsymbol{w}^T\boldsymbol{x} - b)$$

simplest form = "Hard Margin"

i.e. all training points correctly classified

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

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



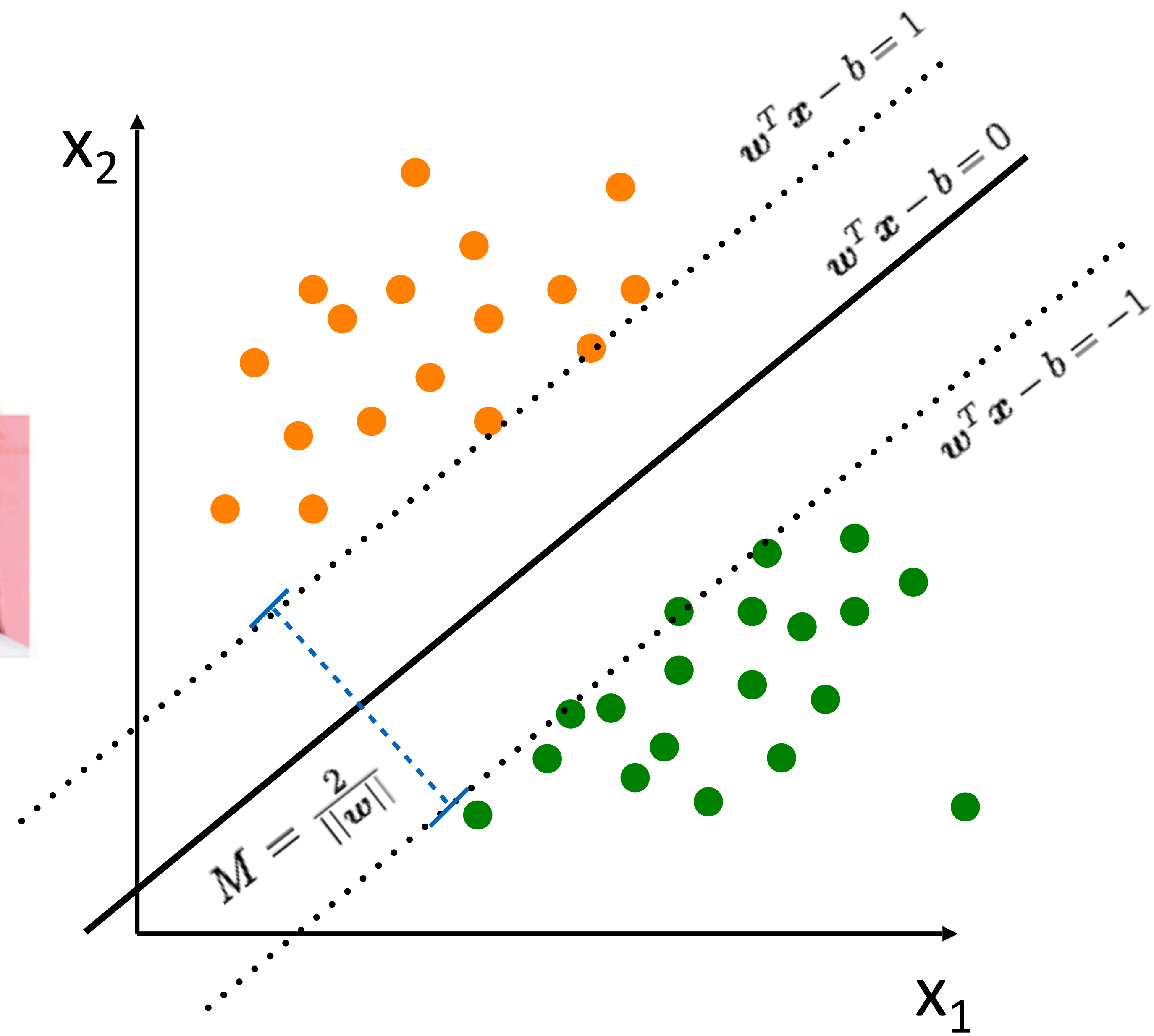
LOTS of variations on this e.g. soft margins, kernel trick for non-linear

Supervised Machine Learning

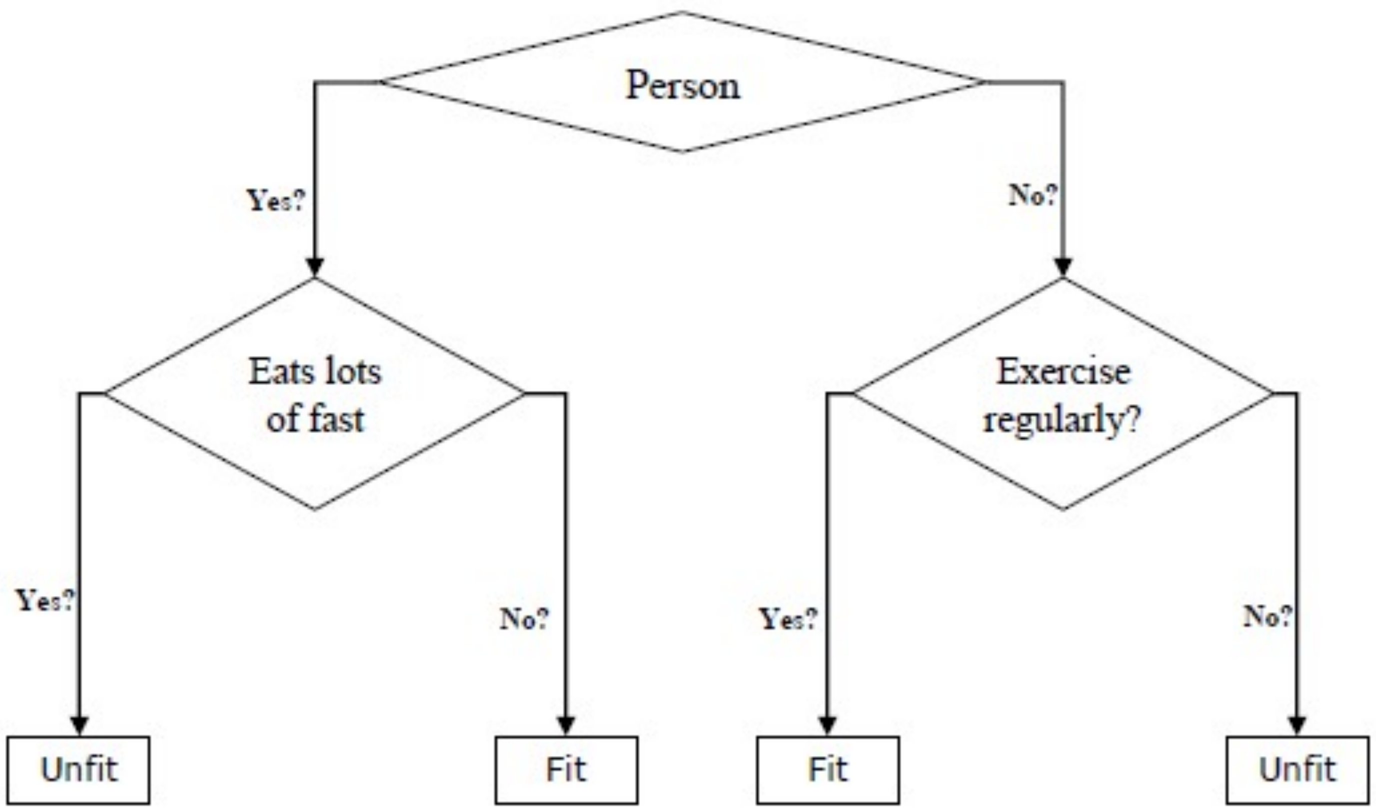
Support Vector Machines

Image recognition via SVM

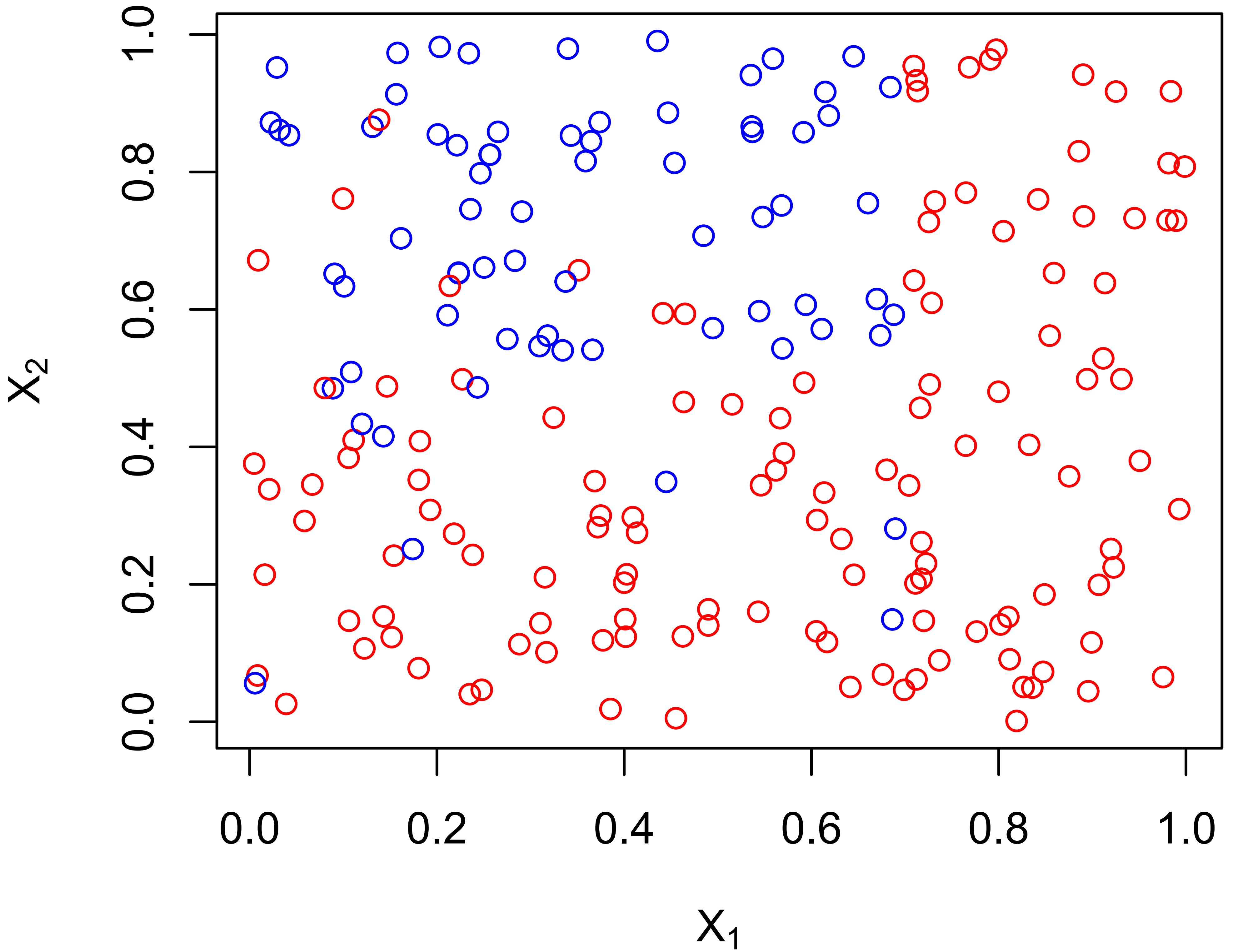
Happy Sad Surprised Angry



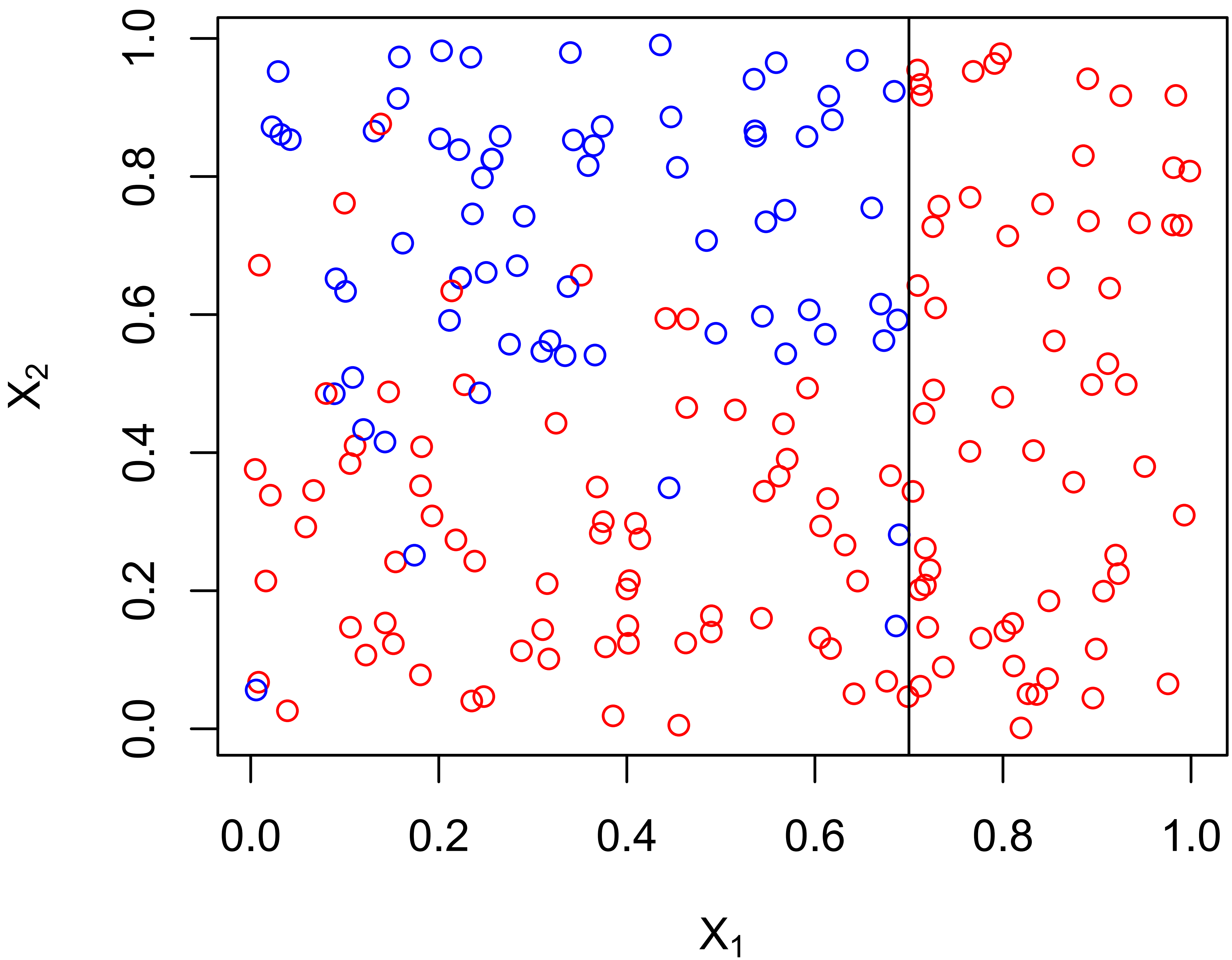
Decision Trees and Random Forests



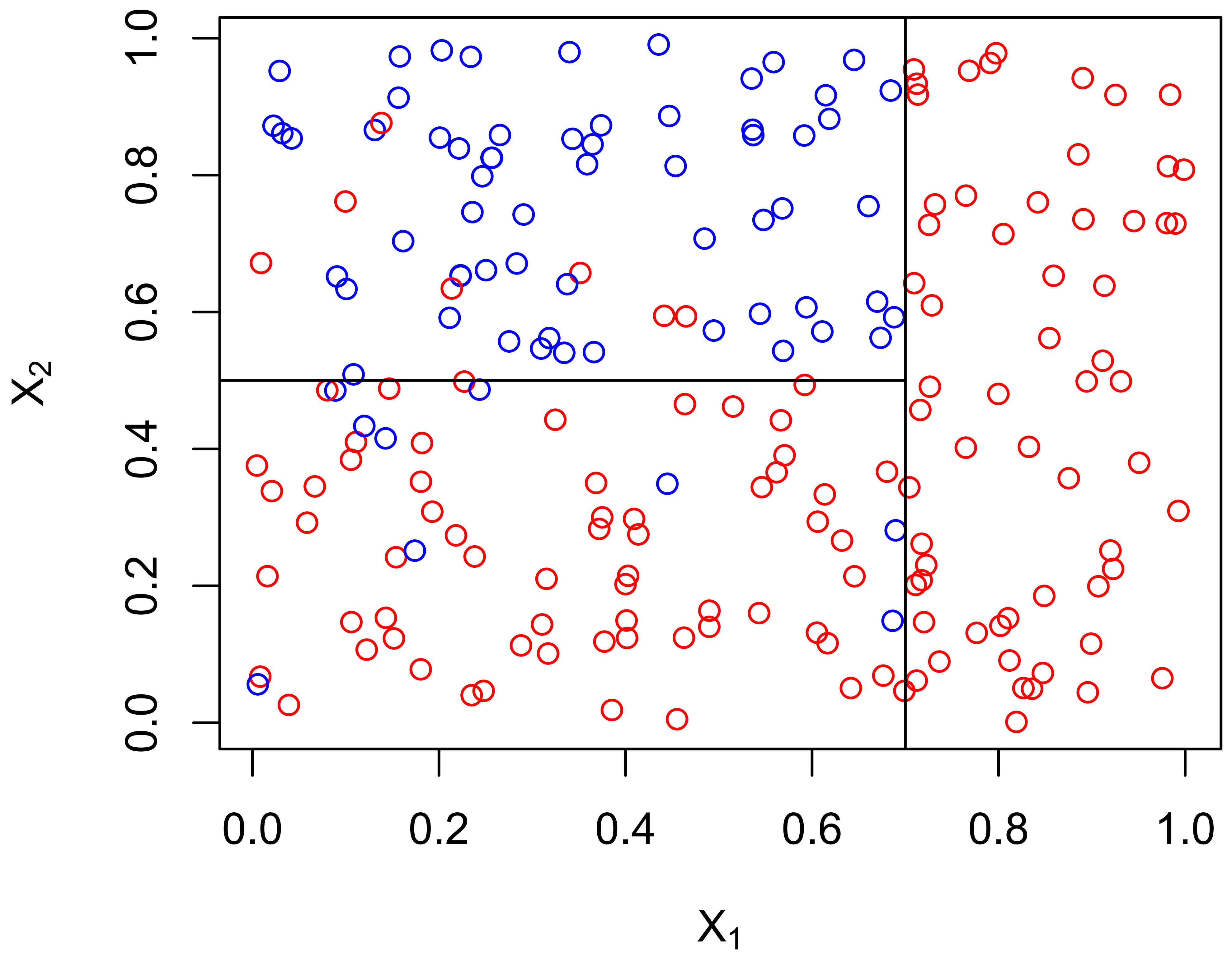
From decision trees to extra trees



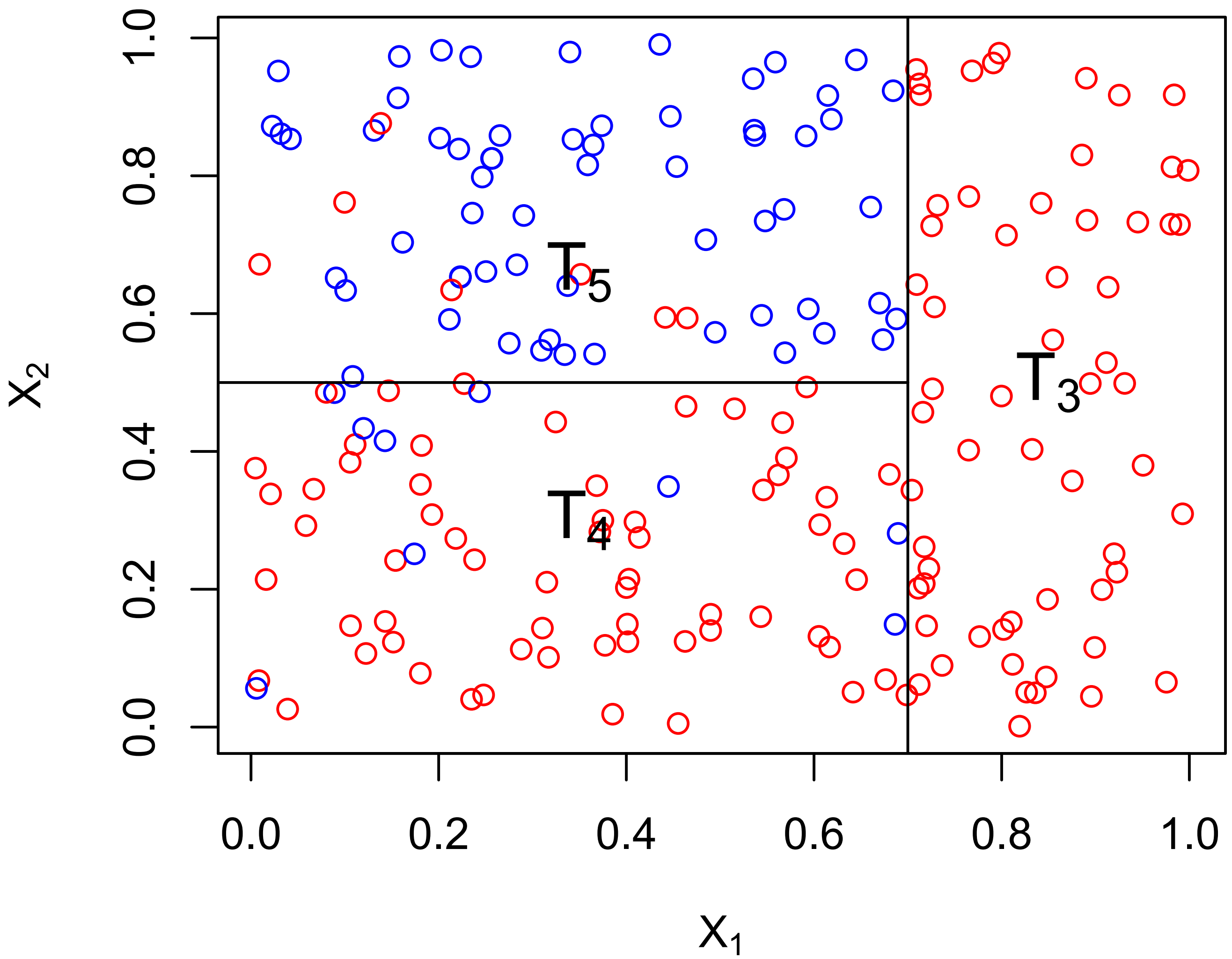
From decision trees to extra trees

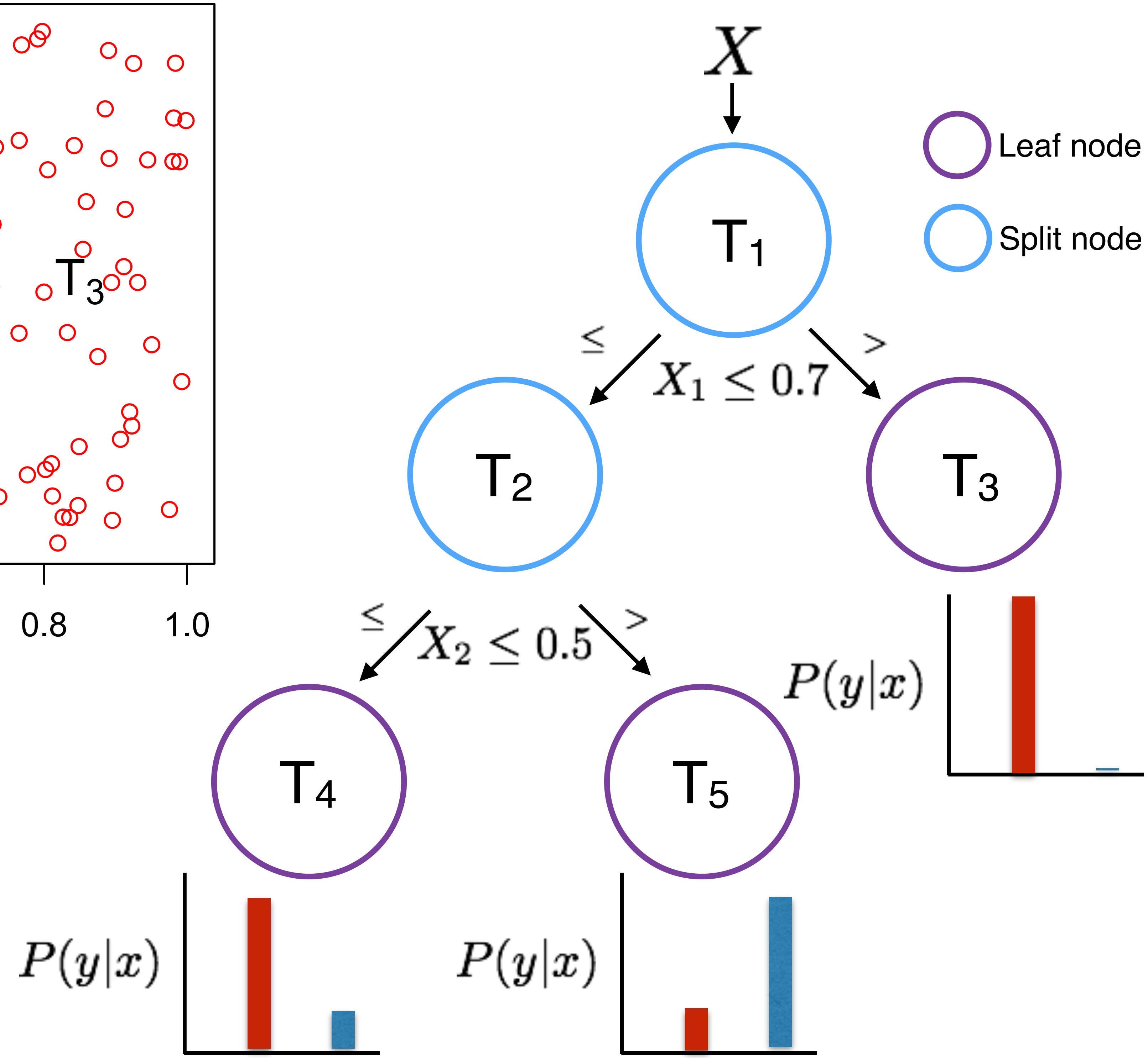


From decision trees to extra trees



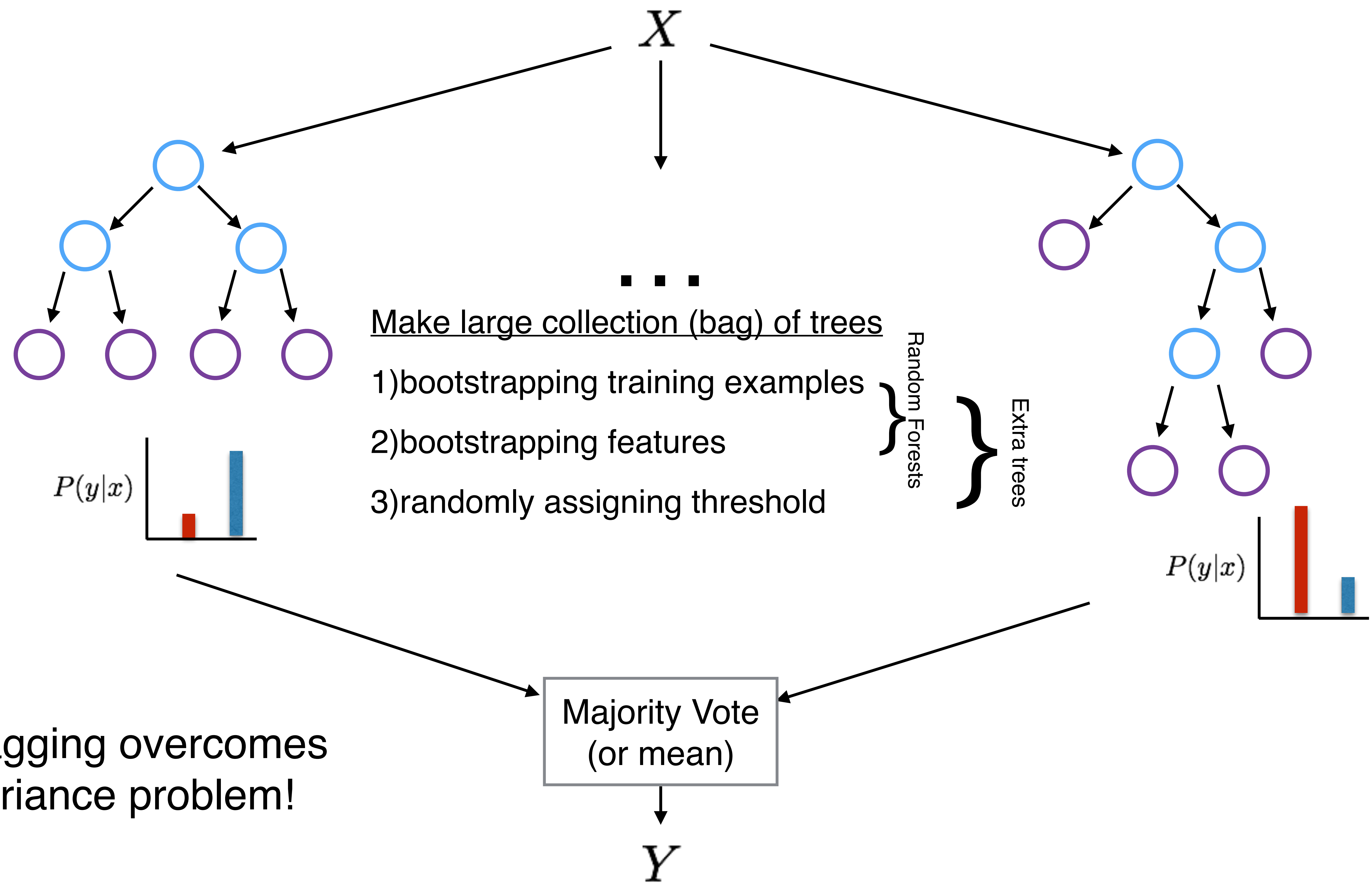
From decision trees to extra trees





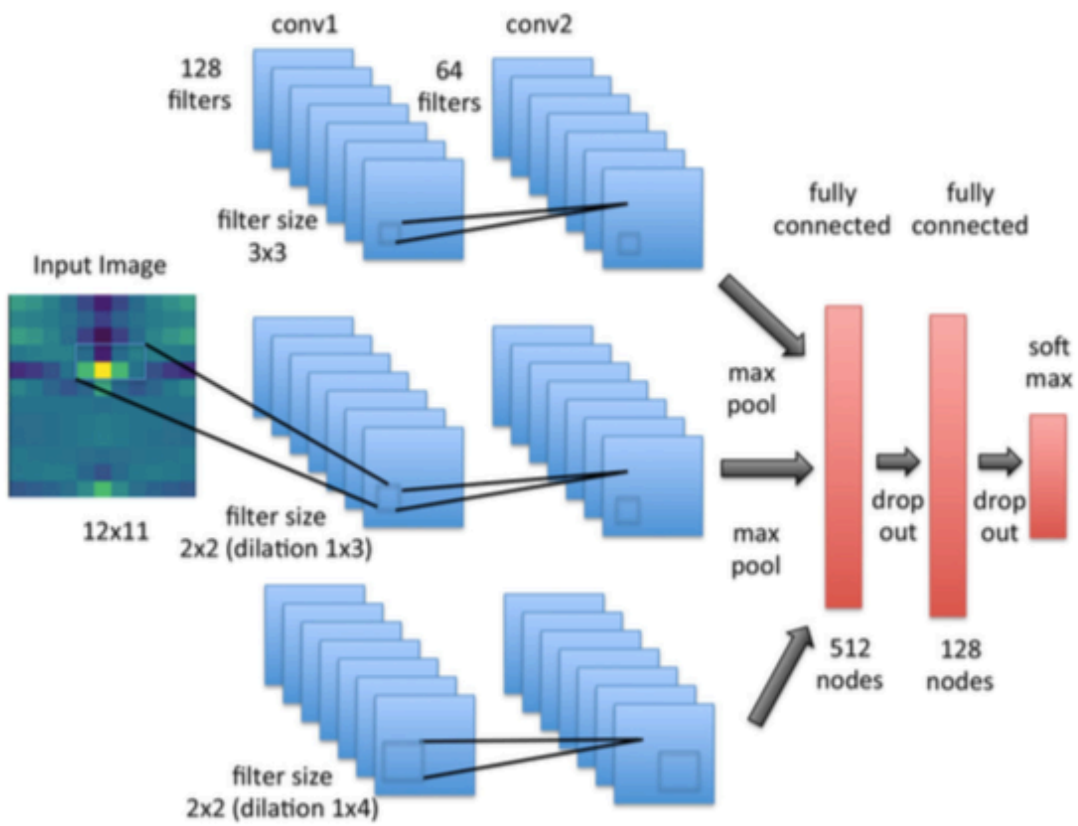
decision trees have low bias but suffer from high variance

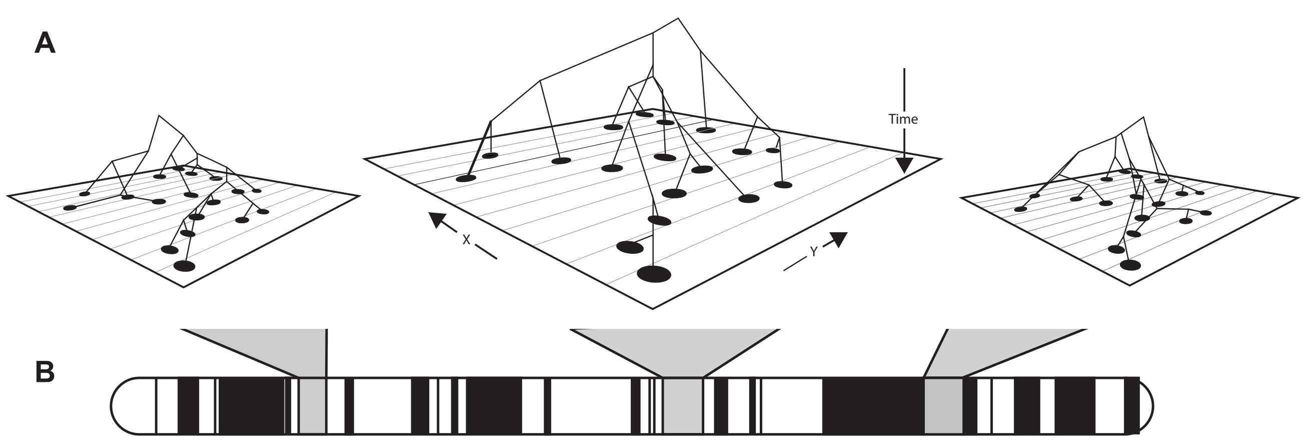
From decision trees to extra trees



bagging overcomes
variance problem!

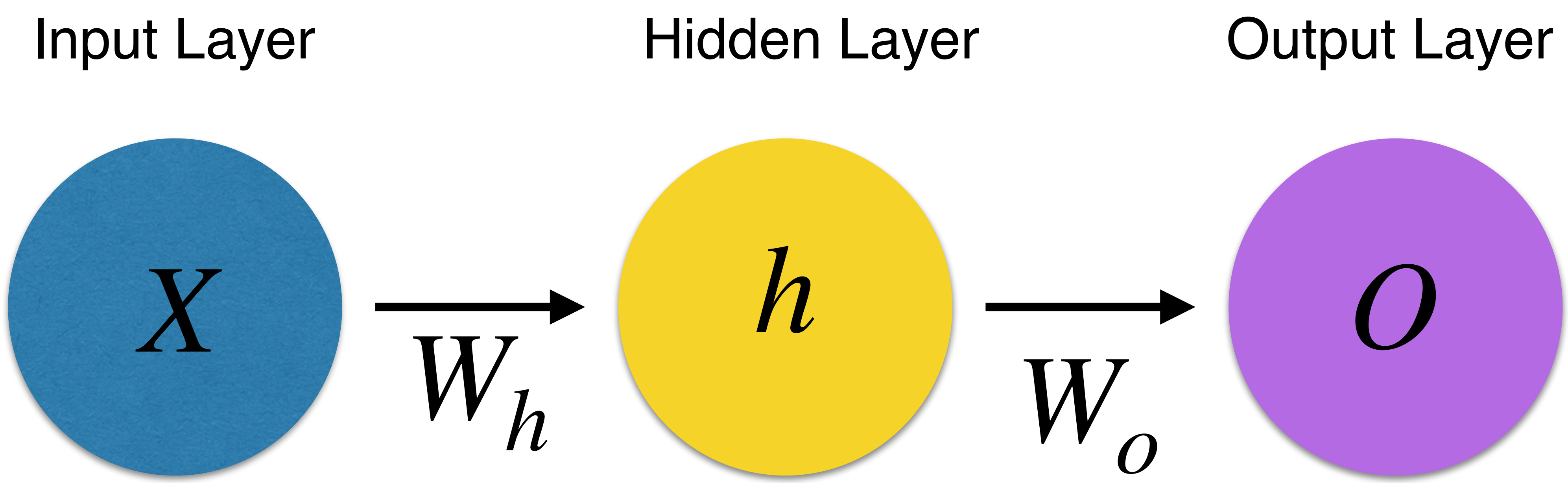
Some of Our Research





Computational Evolutionary Genetics

toy neural network



Feed-forward

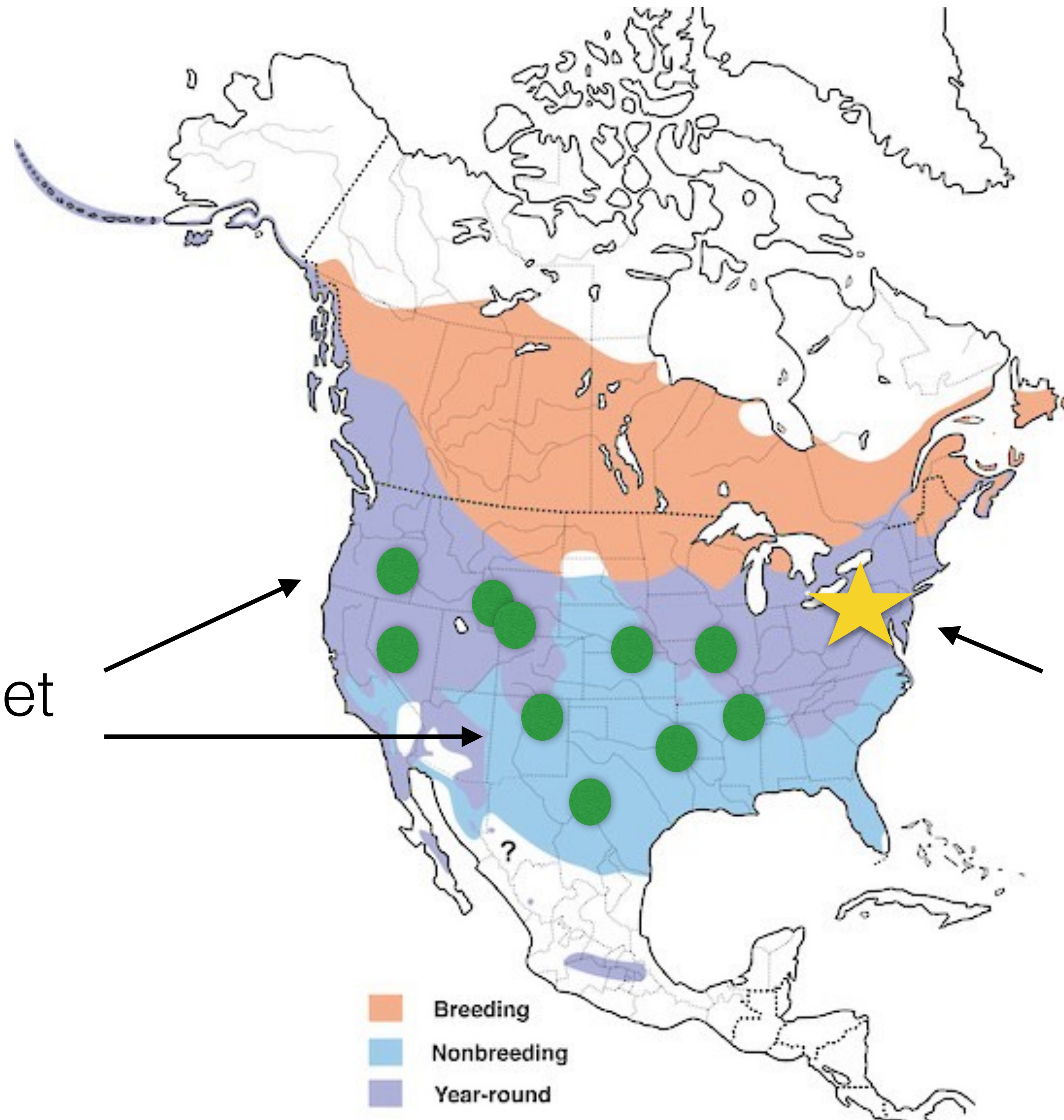
$$O = f(f(X \cdot W_h + b_h) \cdot W_o + b_o)$$

think of it like stacked linear regressions

Organisms live in space

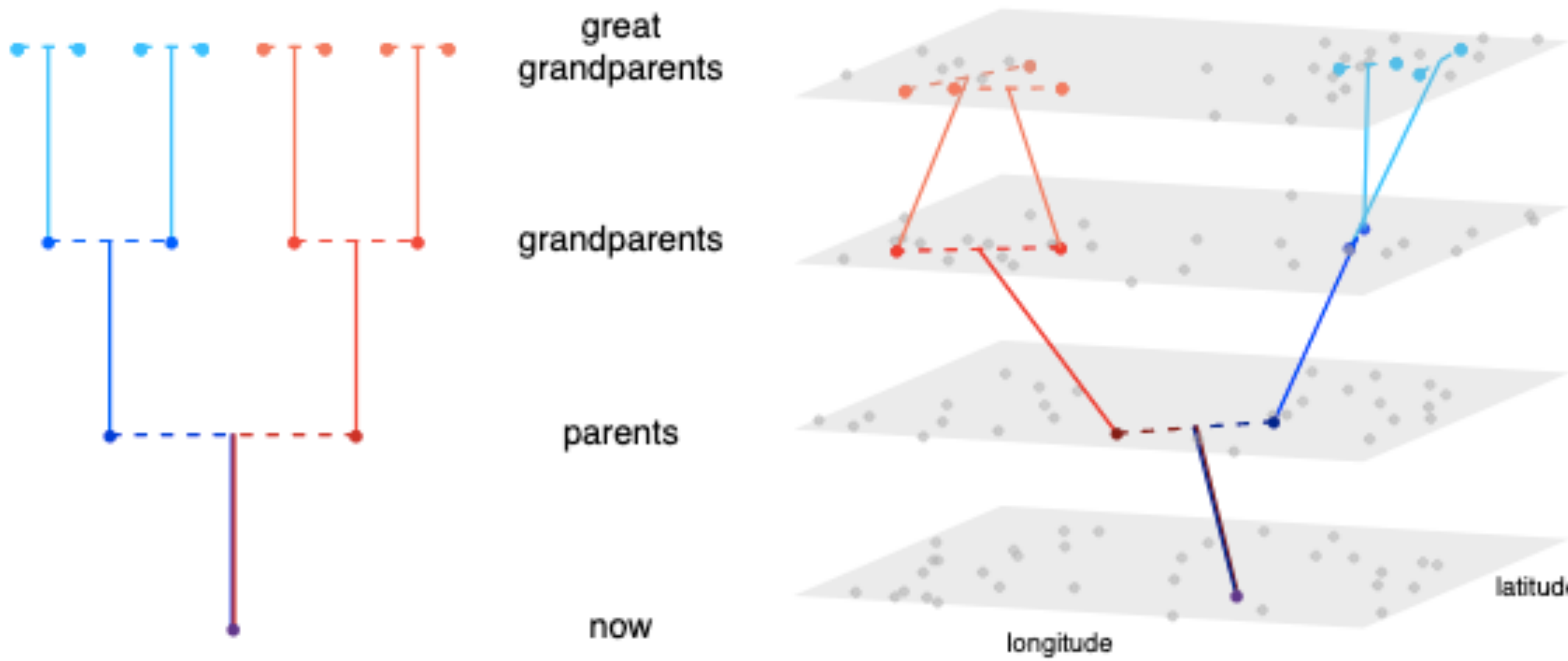
Song sparrow

Training set



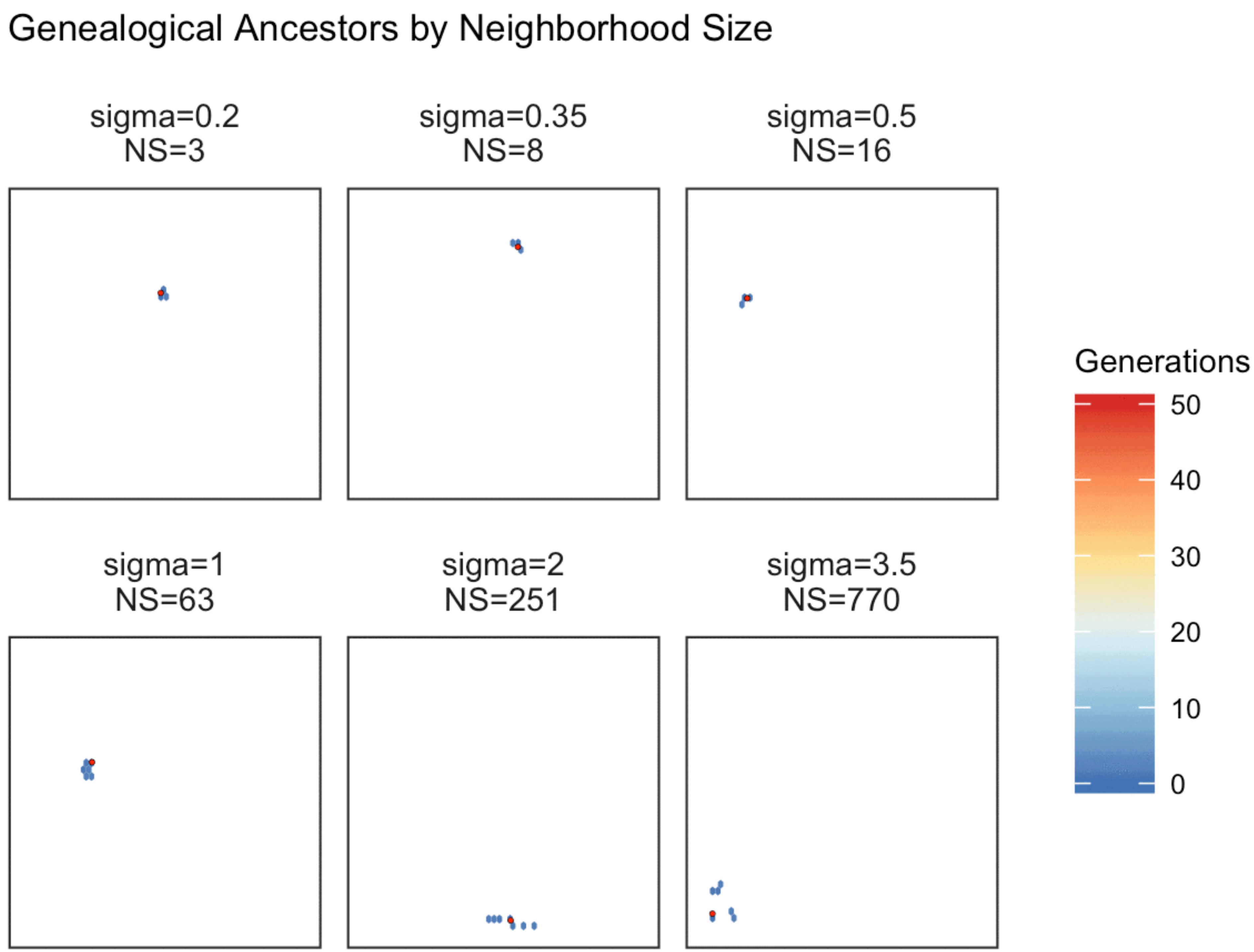
Predict locations of new genomes?

Space is the Place

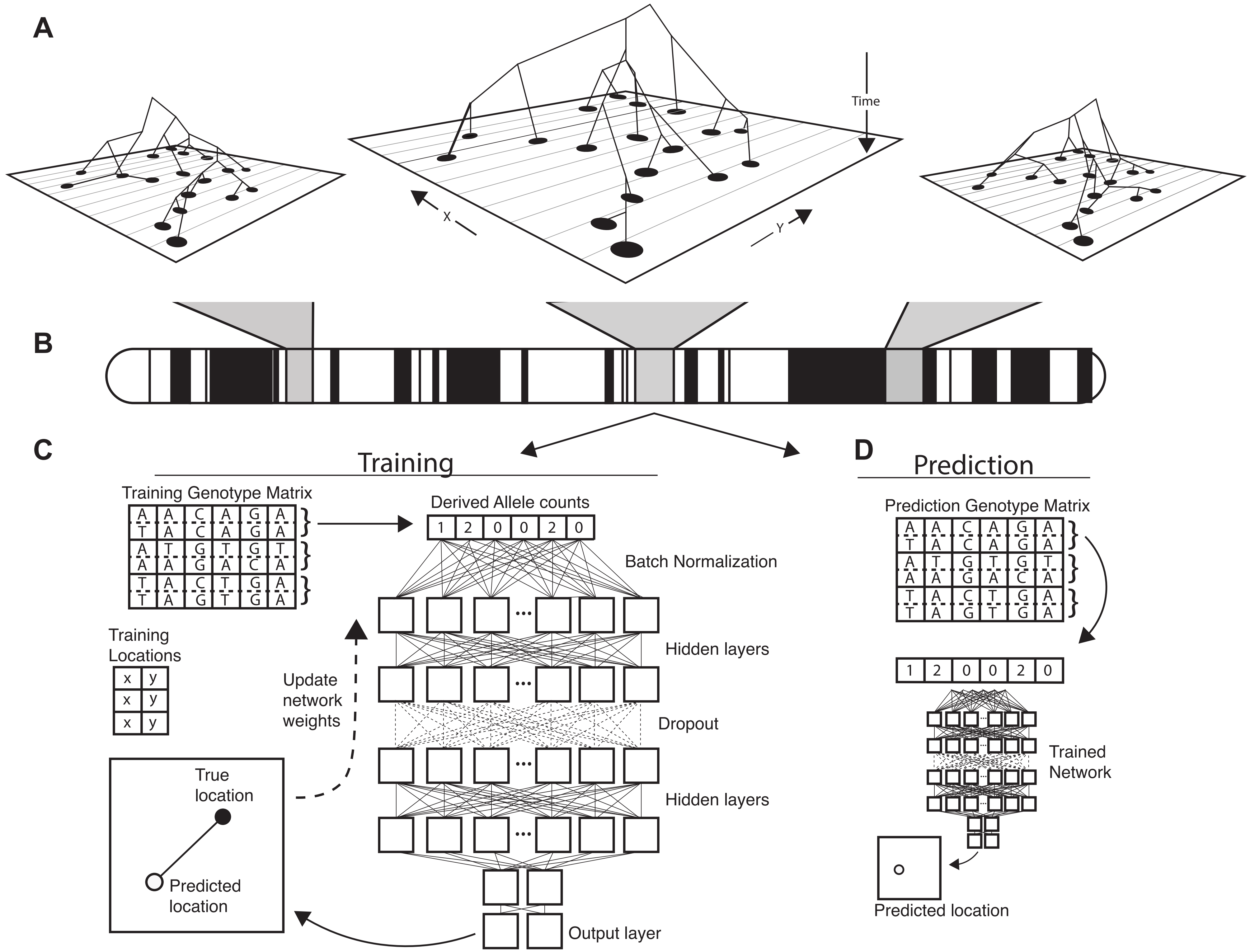


From Bradburd and Ralph (2019)

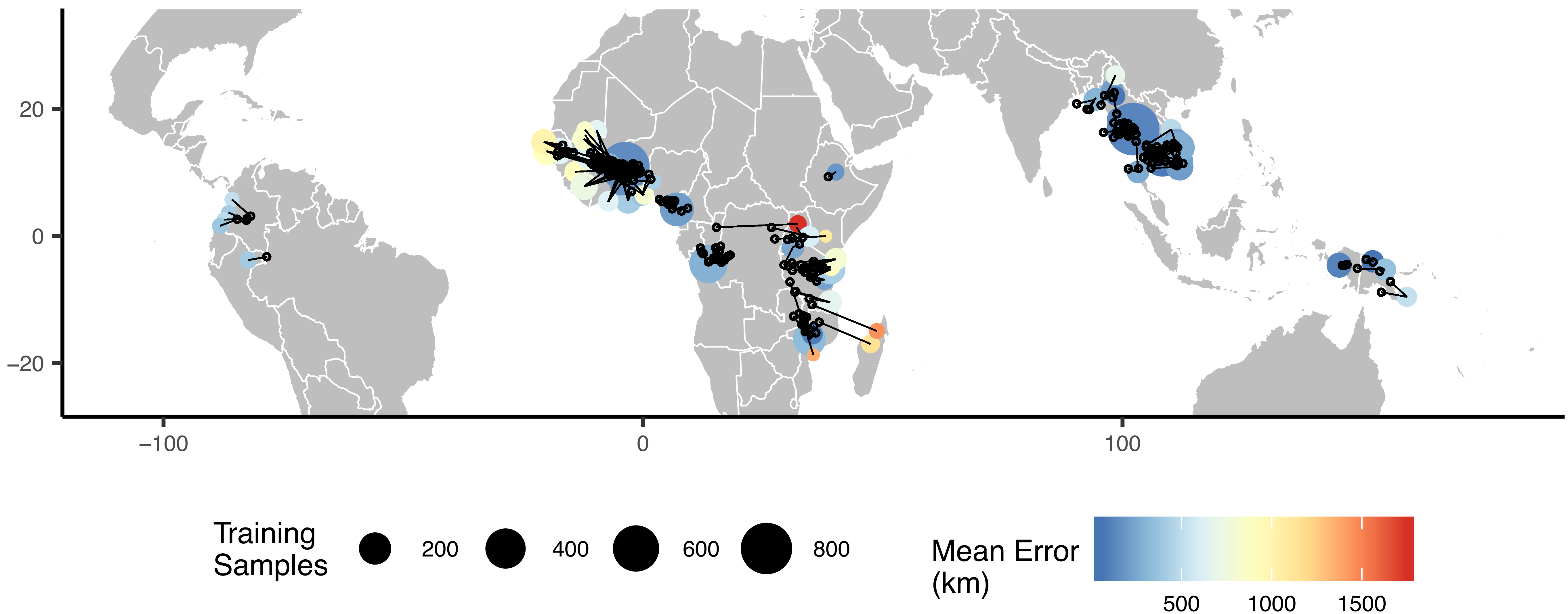
Space is the Place



Locator — (deep) learning space

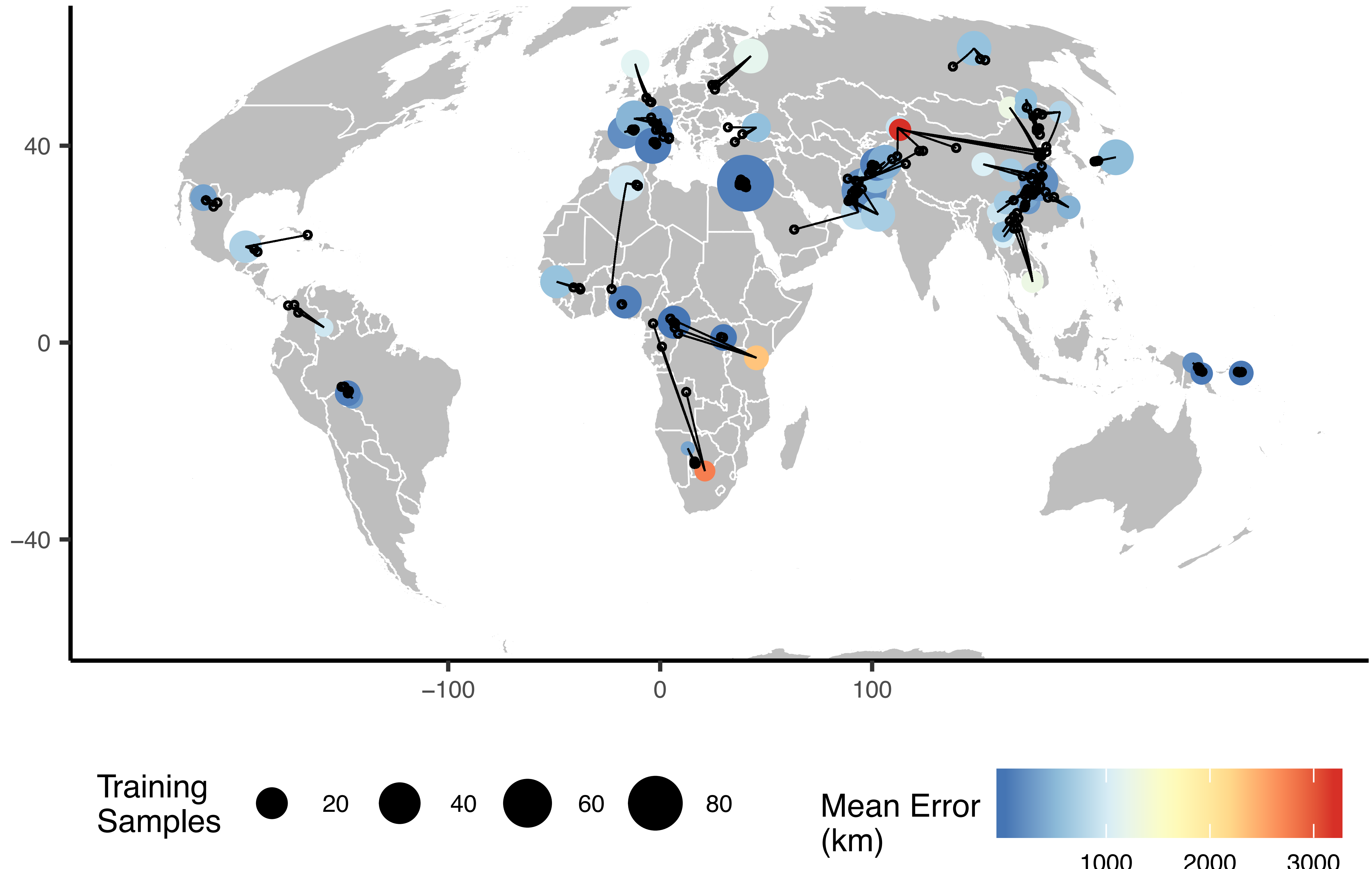


Locator— (deep) learning space



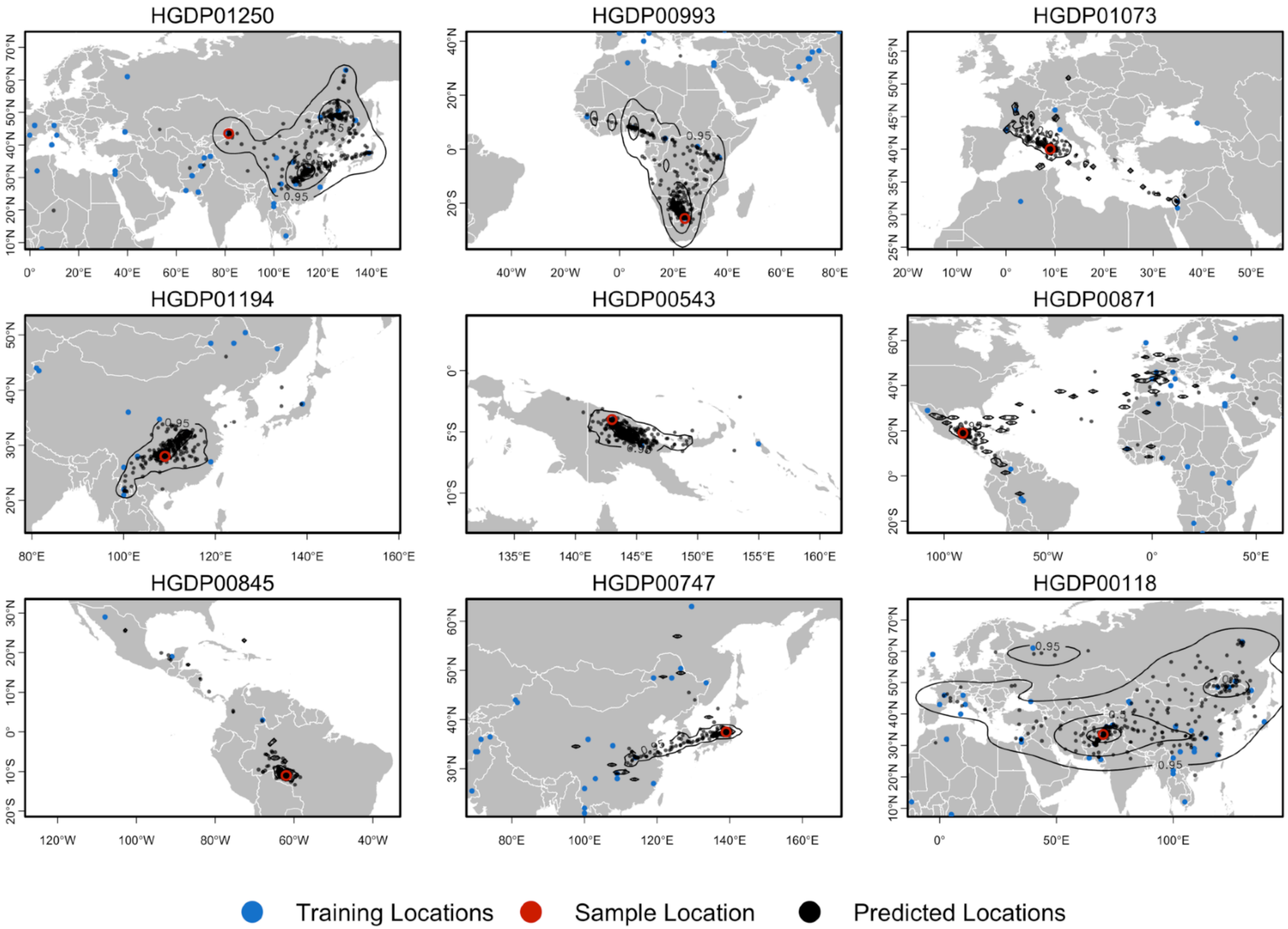
Plasmodium falciparum- Pf7K dataset
median error = 16.9 km

Locator— (deep) learning space



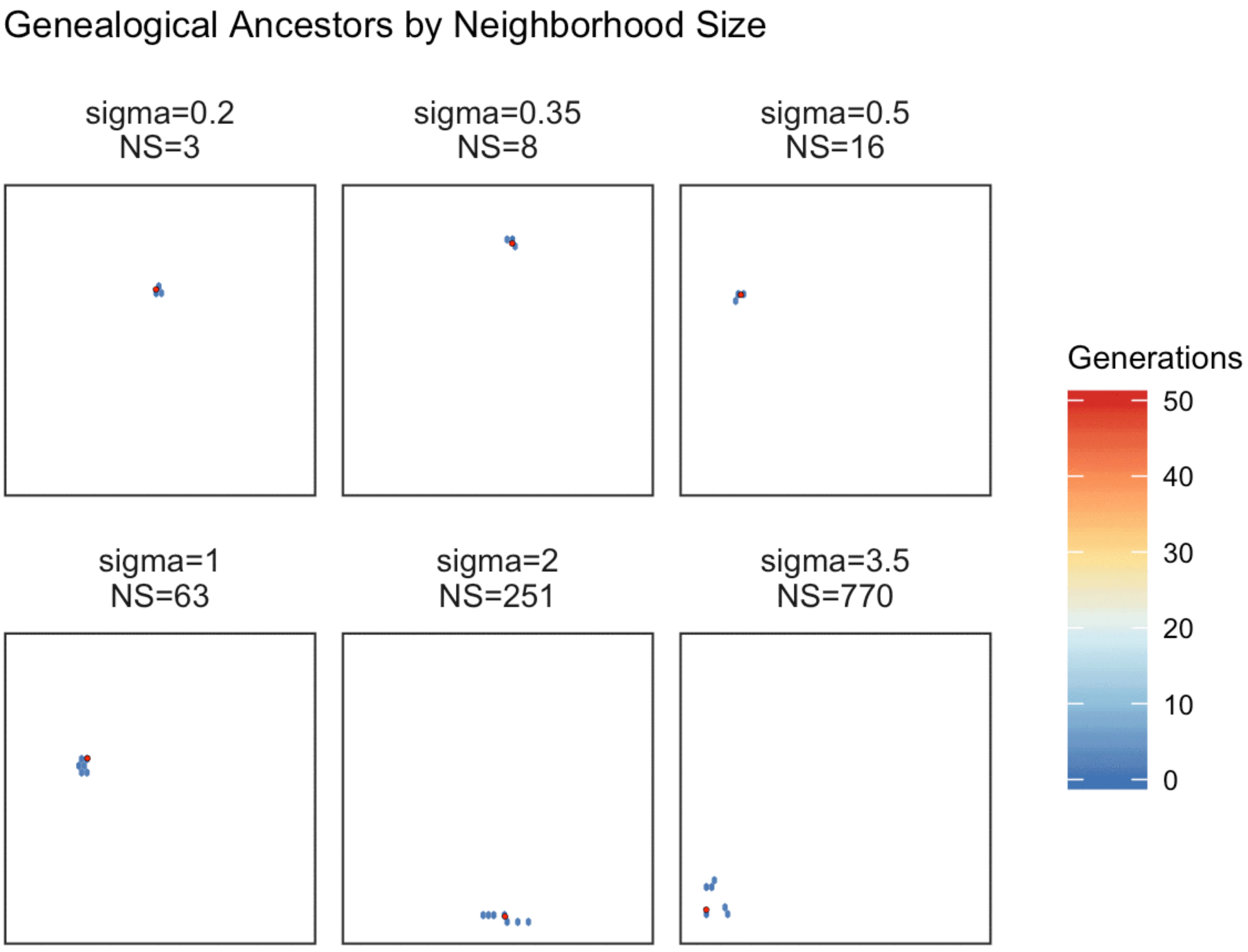
Humans - HGDP
median error = 85km

Locator— (deep) learning space



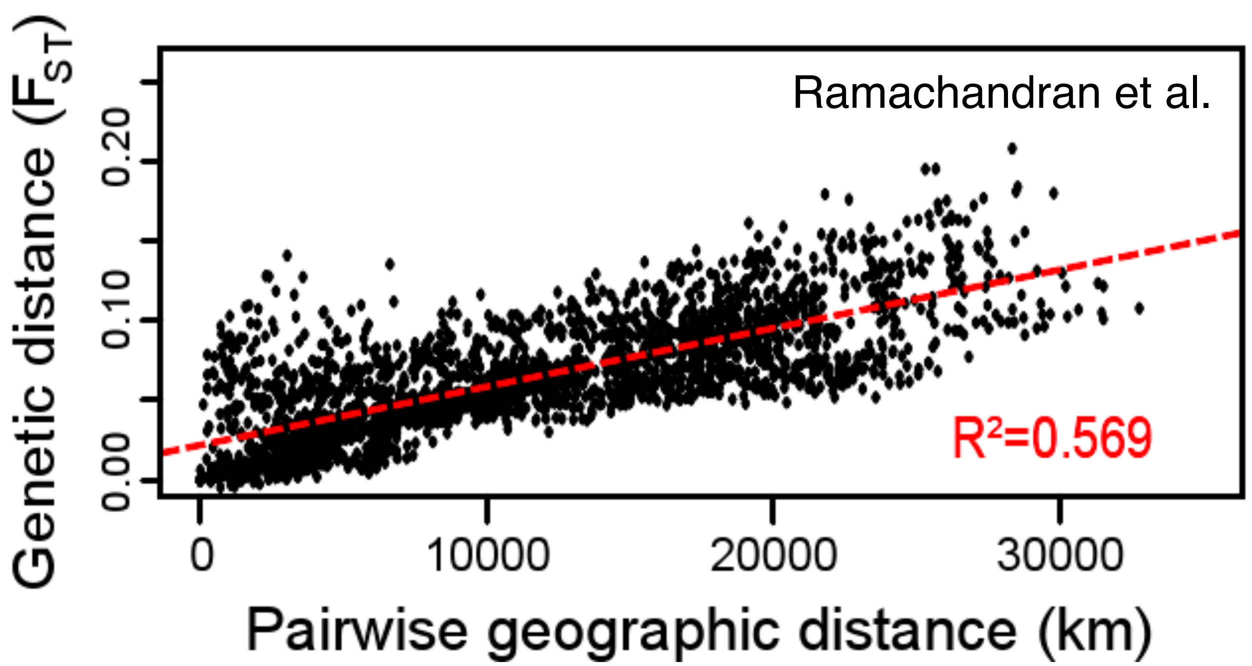
Humans - HGDP

Estimating σ



Estimating σ

Isolation by distance



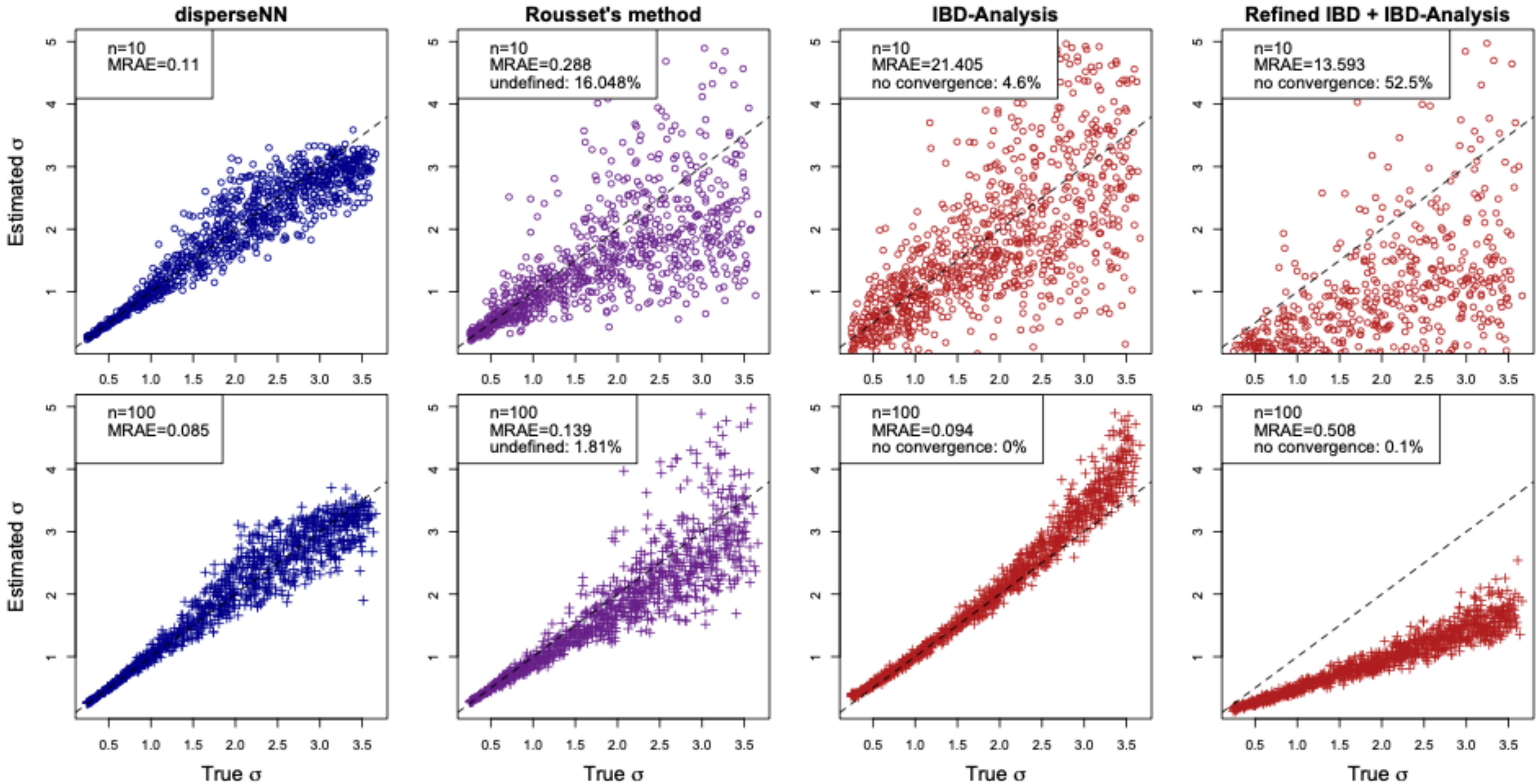
Classic result

Rousset's method— fit this regression, slope is approx $\dfrac{1}{4N\pi\sigma^2}$

BUT

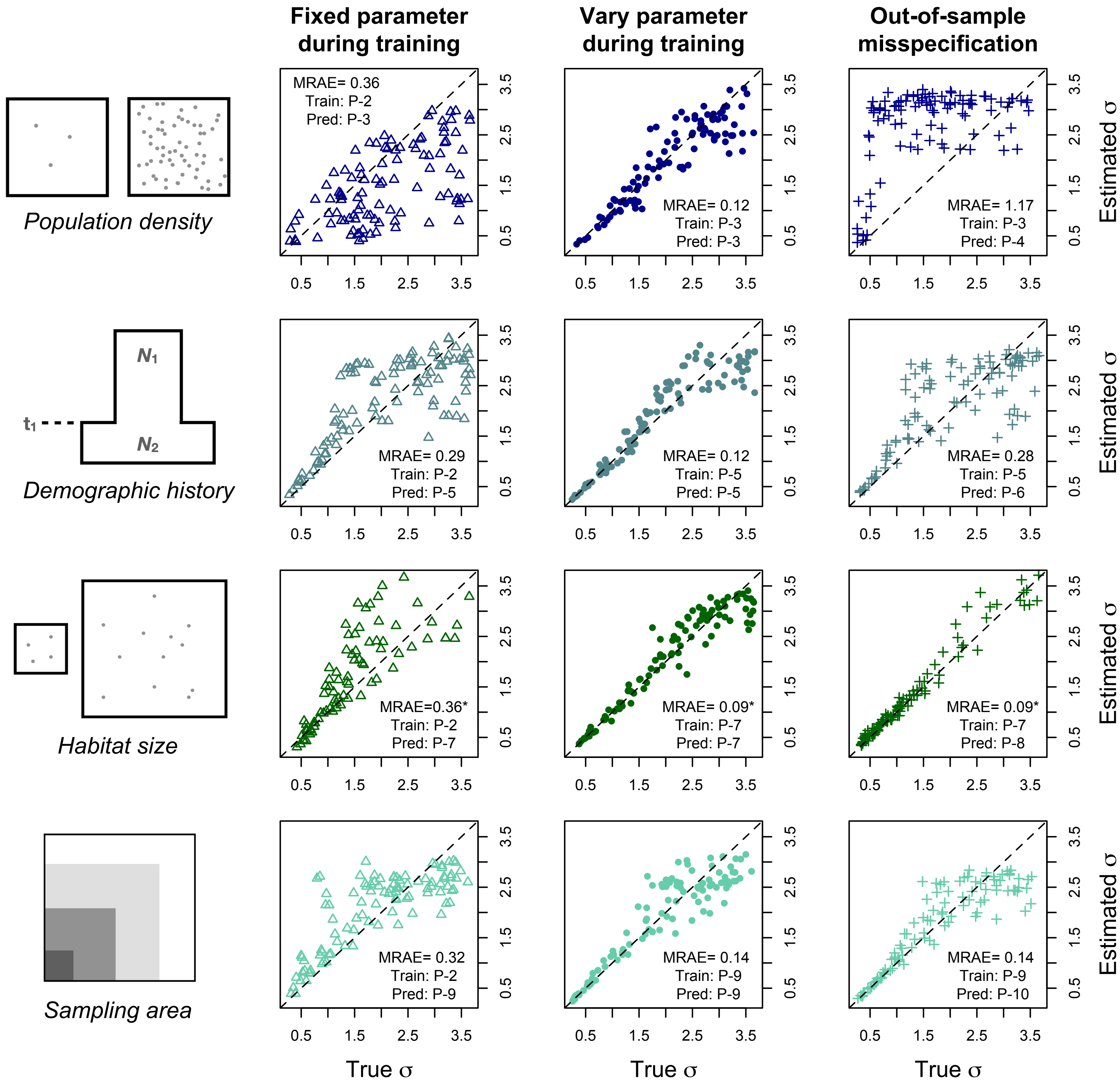
need to know local N!

disperseNN



disperseNN works really well, particularly at small sample size
(assume perfect knowledge of N or perfect IBD tract for other me)

disperseNN



disperseNN sensitive to misspecification but can train our way out of it (mostly)

disperseNN

Species	Common name	Region	σ (km)	95% CI (km)	Previous (km)	N_{loc}	n	S (km)	M. dist.
Zosterops borbonicus	Réunion grey white-eye	Réunion	4.97	(1.76, 13.83)	NA	295	41	62	4.59
Peromyscus leucopus	white-footed mouse	New York	0.77	(0.32, 1.67)	0.03-0.11	-231	12	38	8.15
Anopheles gambiae	African malaria mosquito	Cameroon	10.29	(2.00, 48.03)	0.04-0.5	52	29	278	9.62
Bombus bifarius	two-form bumble bee	Washington	14.75	(5.60, 37.28)	1.2-5	1,147	14	273	10.47
Bombus vosnesenskii	yellow-faced bumble bee	California	7.70	(1.21, 38.11)	1.2-5	3,944	18	169	11.83
Hippoglossus hippoglossus	Atlantic halibut	Canada	4.29	(0.71, 33.85)	NA	-5,546	11	193	14.59
Crassostrea virginica	eastern oyster	Canada	1.52	(0.72, 4.31)	21.9	1,435	13	187	19.69
Canis lupus	grey wolf	N. America	15.68	(2.36, 107.3)	98-147	35	13	721	25.42
Helianthus petiolaris	prairie sunflower	Kansas	1.00	(0.39, 3.52)	0.156	9	11	204	45.28
Zosterops olivaceus	Réunion olive white-eye	Réunion	1.05	(0.27, 4.36)	NA	2,392	10	50	45.97
Helianthus argophyllus	silverleaf sunflower	Texas	1.04	(0.38, 4.08)	0.156	57	30	307	86.49
Arabidopsis thaliana	thale cress	Spain	1.36	(0.28, 5.05)	0.001	35	35	80	198.25
Arabidopsis thaliana	thale cress	Sweden	0.44	(0.20, 0.93)	0.001	84	84	325	428.17

Empirical estimates from diverse set of organisms

Dispersal inference from population genetic variation using a convolutional neural network

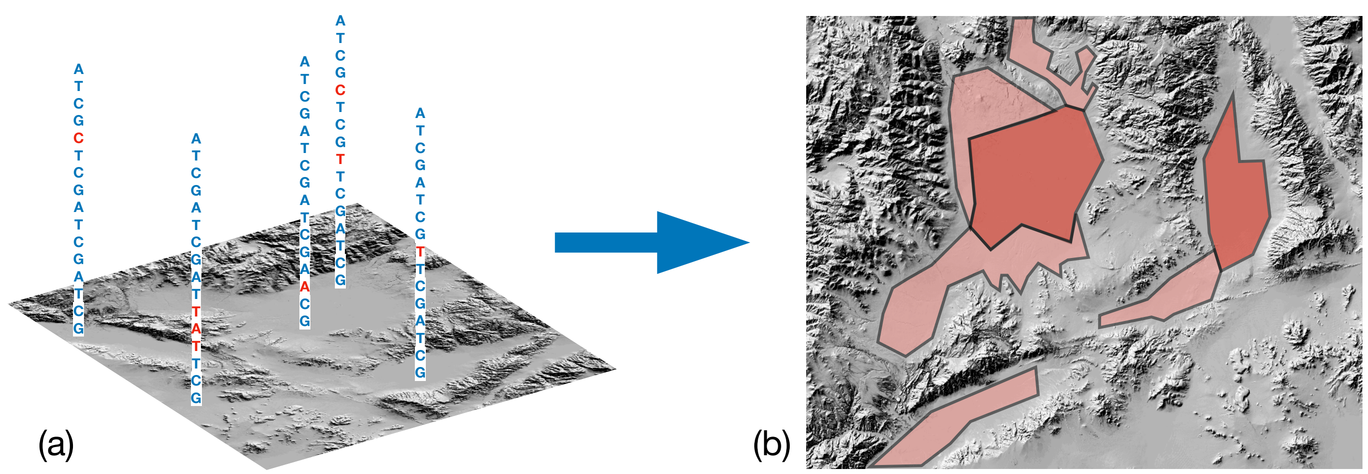
Chris C. R. Smith, Silas Tittes, Peter L. Ralph, Andrew D. Kern

doi: https://doi.org/10.1101/2022.08.25.505329

disperseNN

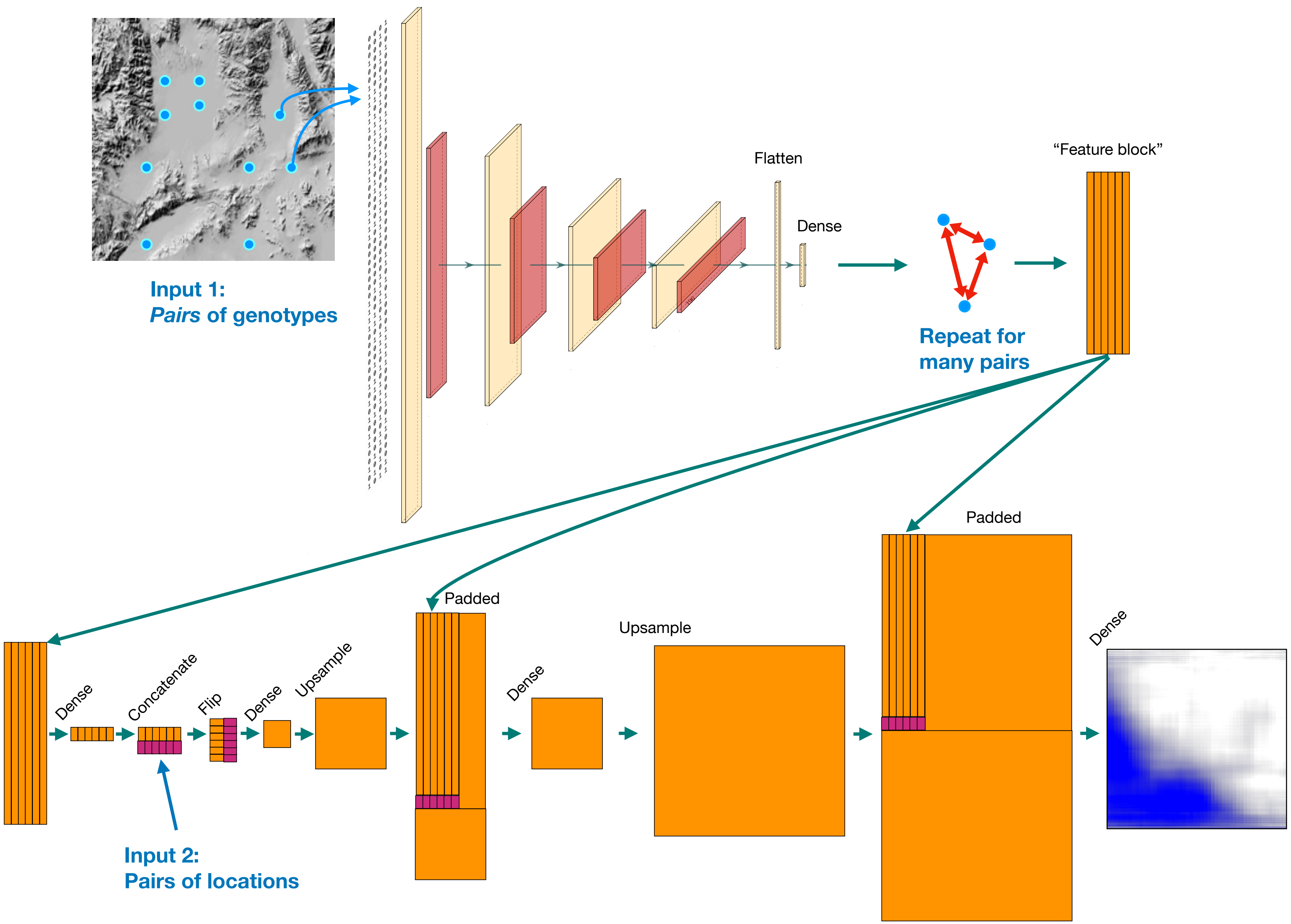
But dispersal need not be homogenous across space!

disperseNN



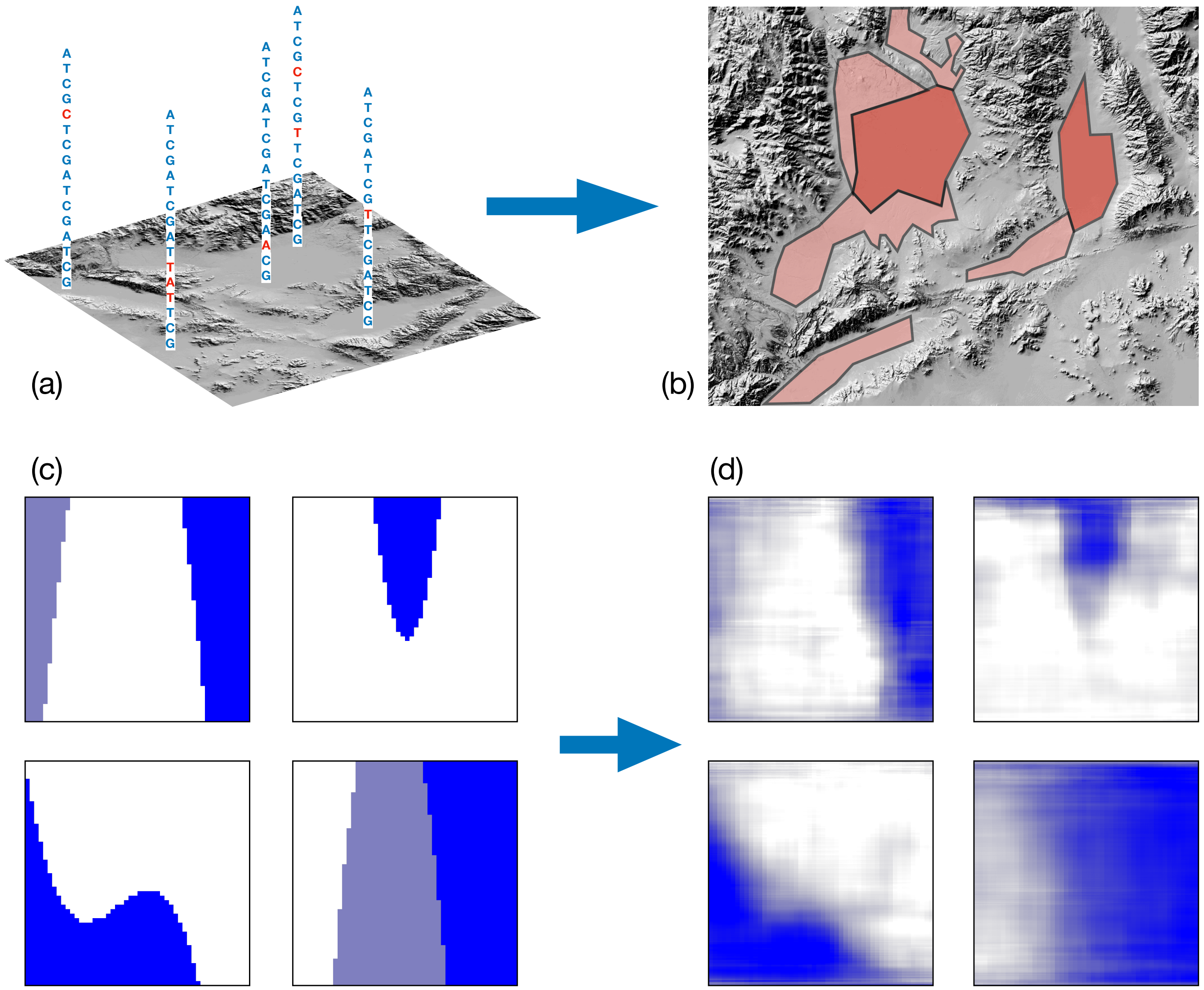
Predicting maps of dispersal with an segmentation network

disperseNN



Predicting maps of dispersal with an segmentation network

disperseNN



Predicting maps of dispersal with an segmentation network

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

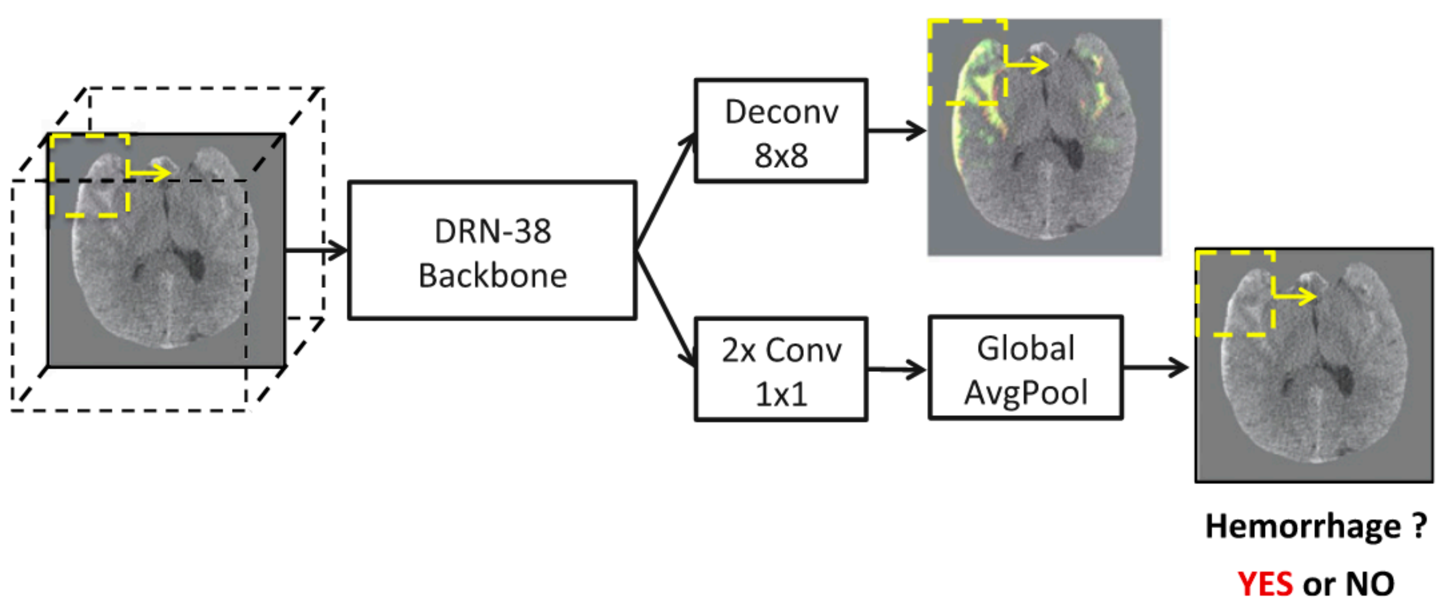
Artificial intelligence will radically disrupt the world of work, but the right policy choice
contract.

The promise of machine learning

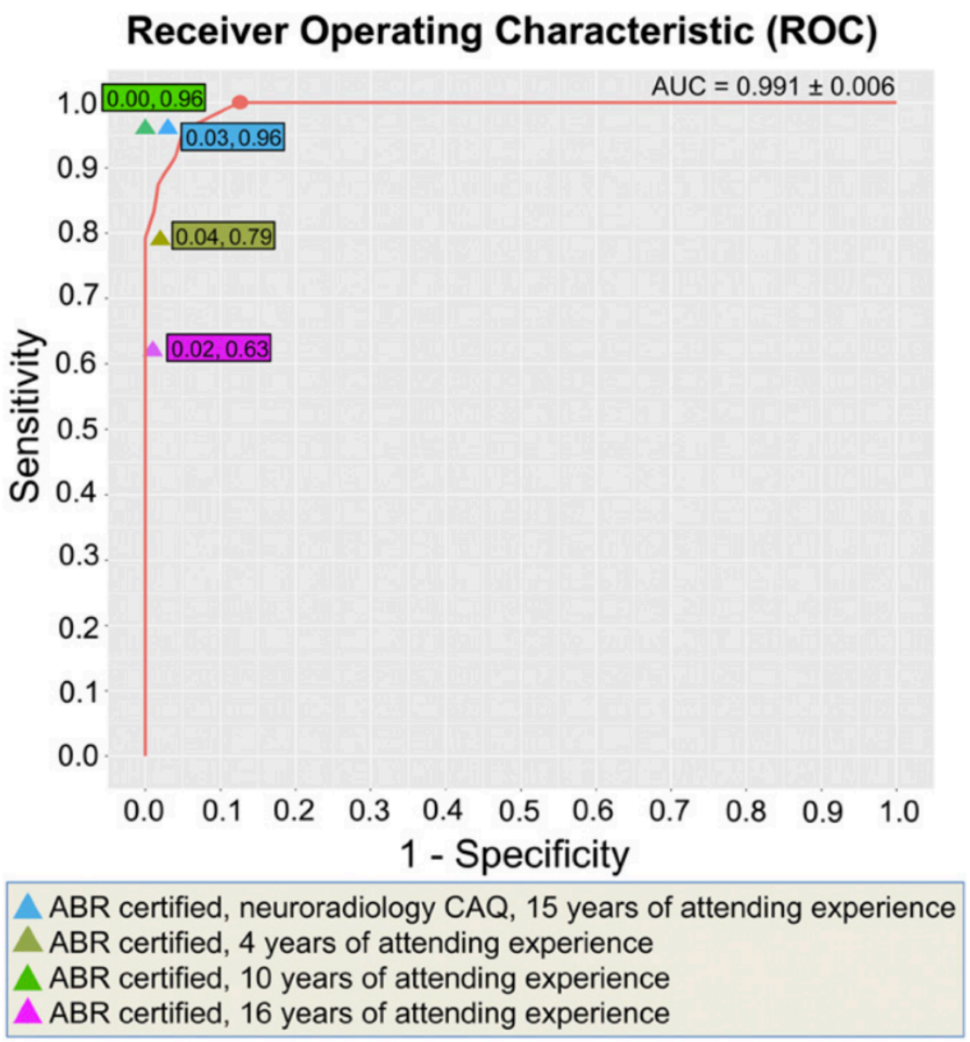
Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning

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



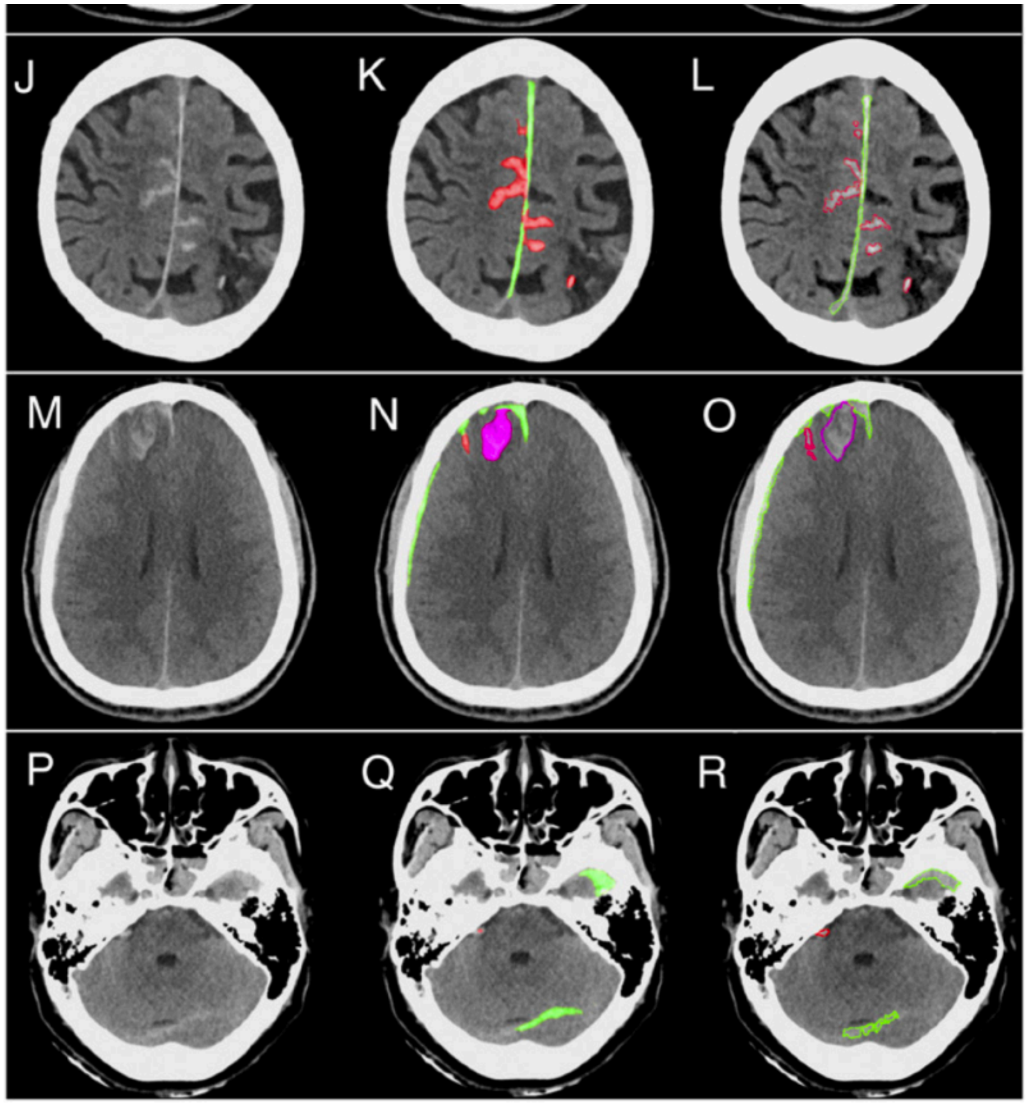


Fig. 5. Examples of multiclass segmentation by the algorithm and by an expert. (*A–C*) Small left holohemispheric subdural hematoma (SDH, green) and adjacent contusion (purple). (*D–F*) Small right frontal and posterior parafalcine SDH and anterior interhemispheric fissure SAH (red). (*G–I*) Small bilateral 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

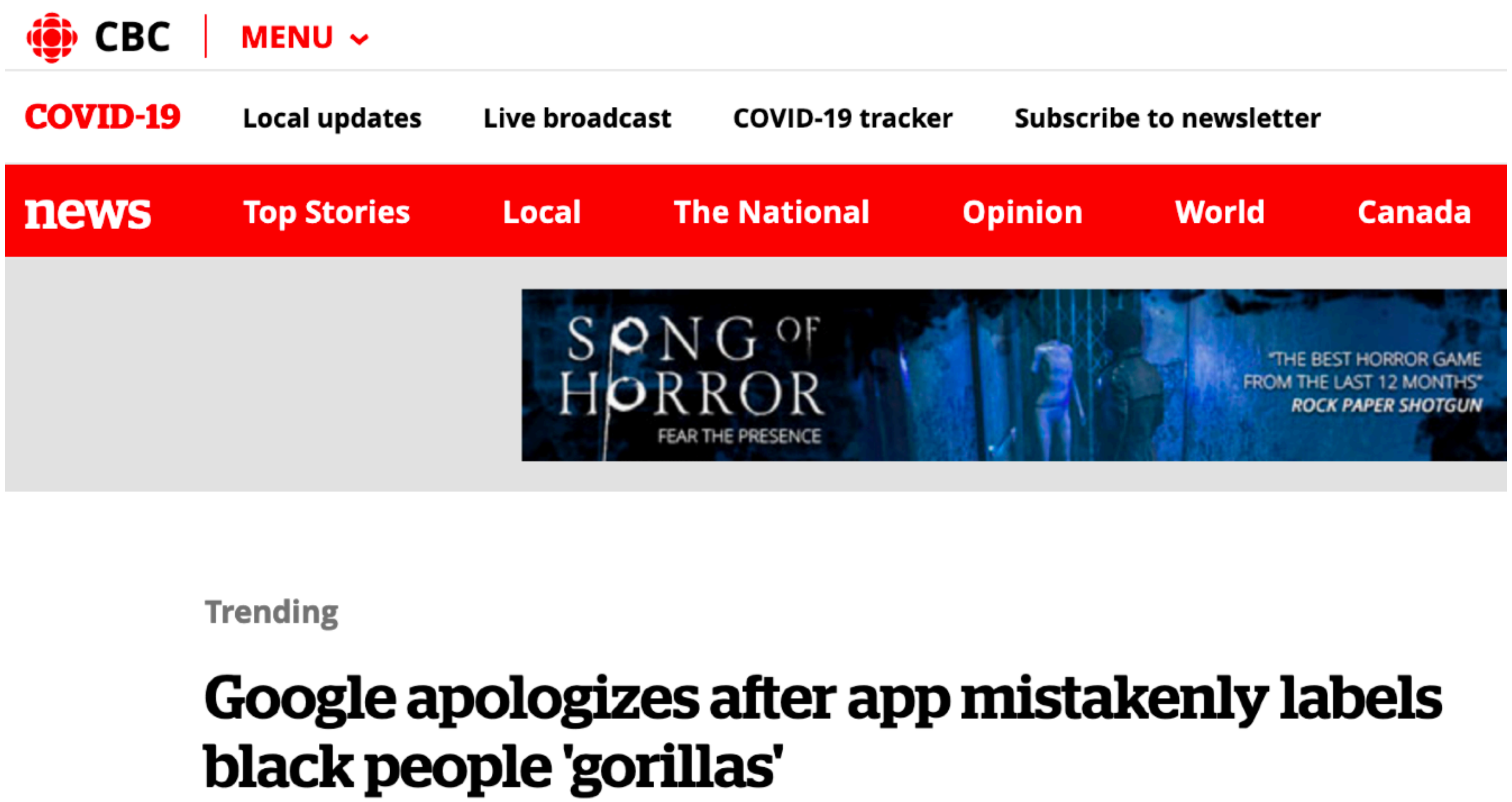
Facial Recognition Is Accurate, if You're a White Guy

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



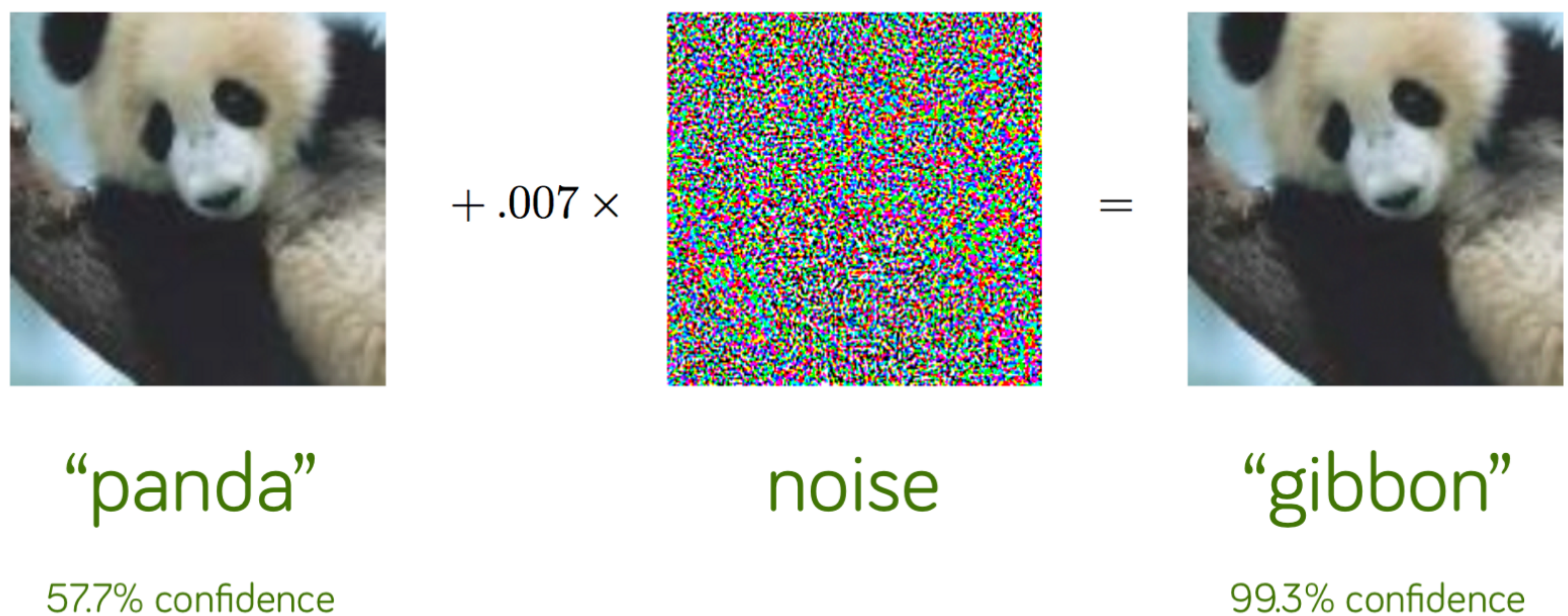
2. Biases in the training set

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

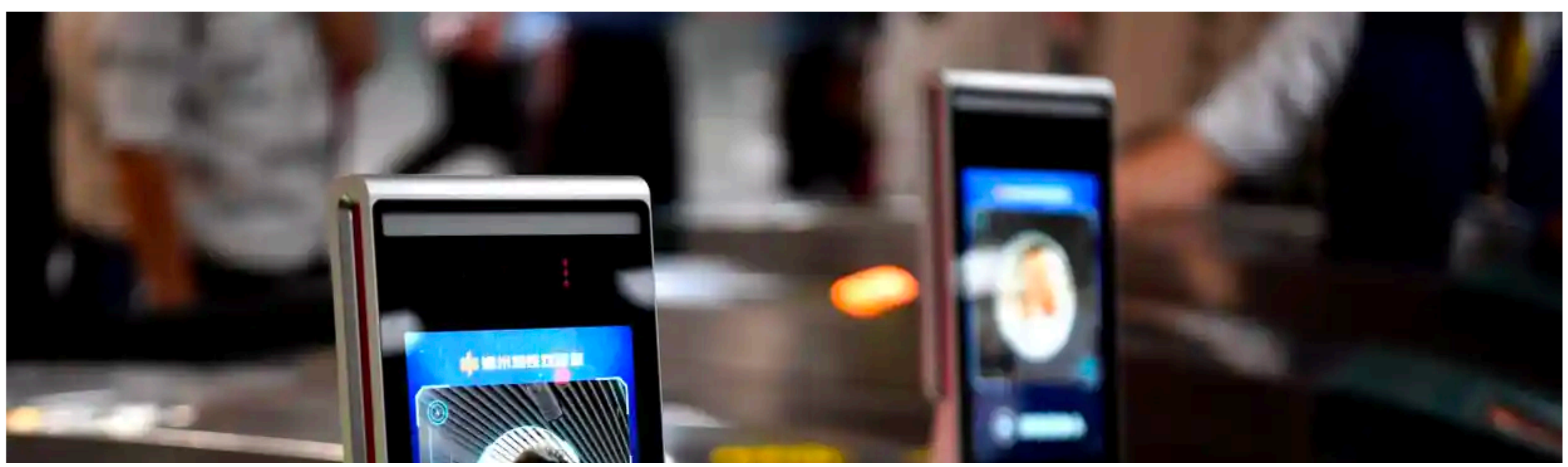


4. Fragile classifiers

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

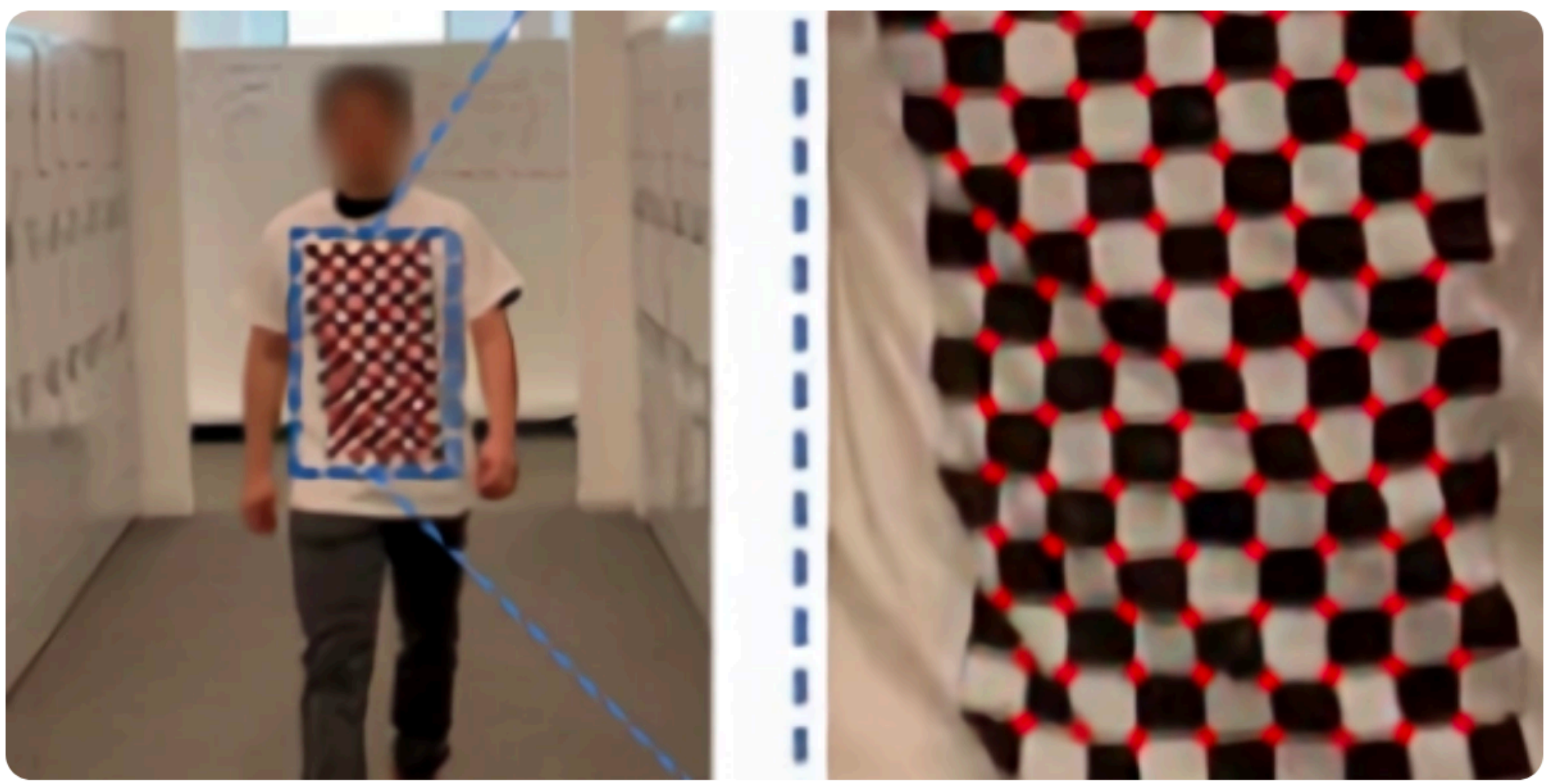


huge potential societal impacts

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



VB TRANSF
The AI even
business lea
Hosted Onli
July 15 - 17
Learn More

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

Generative 'AI'

DALL·E

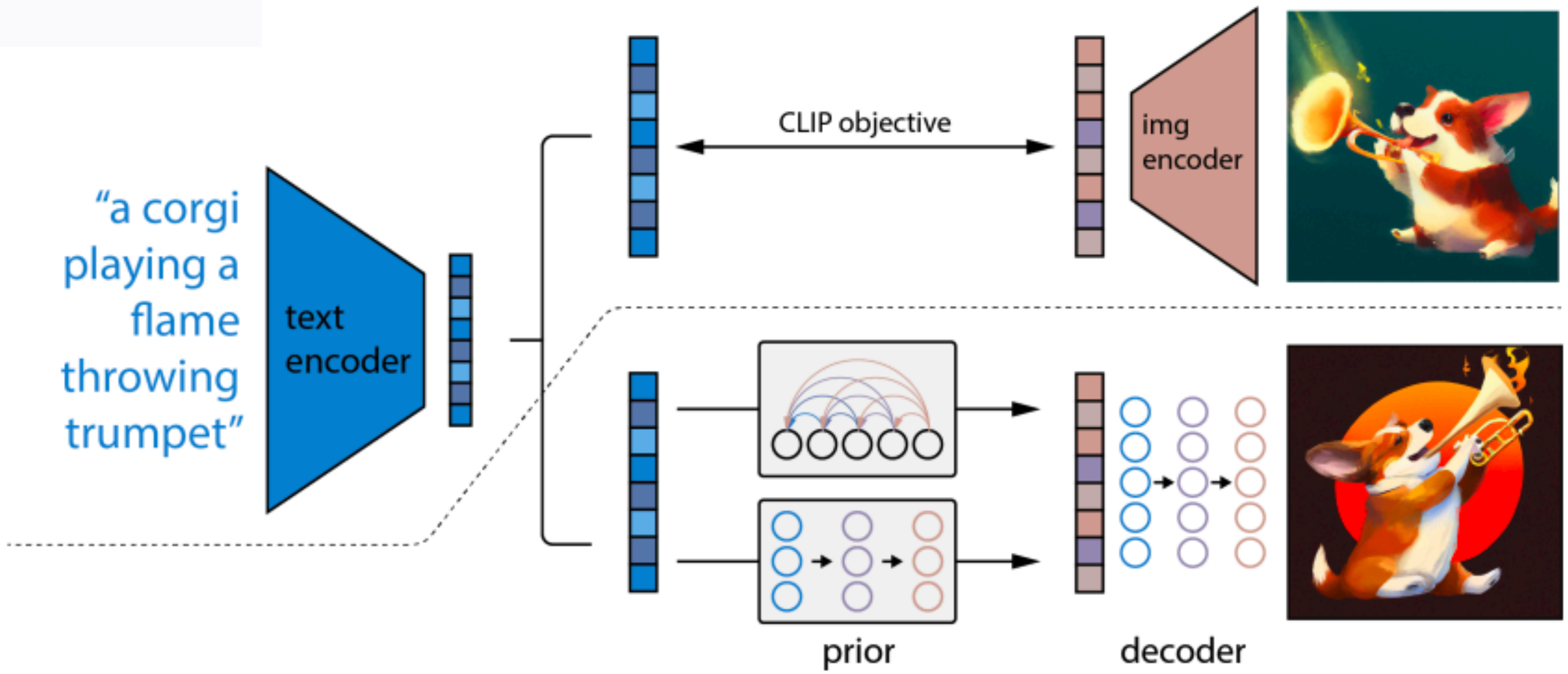
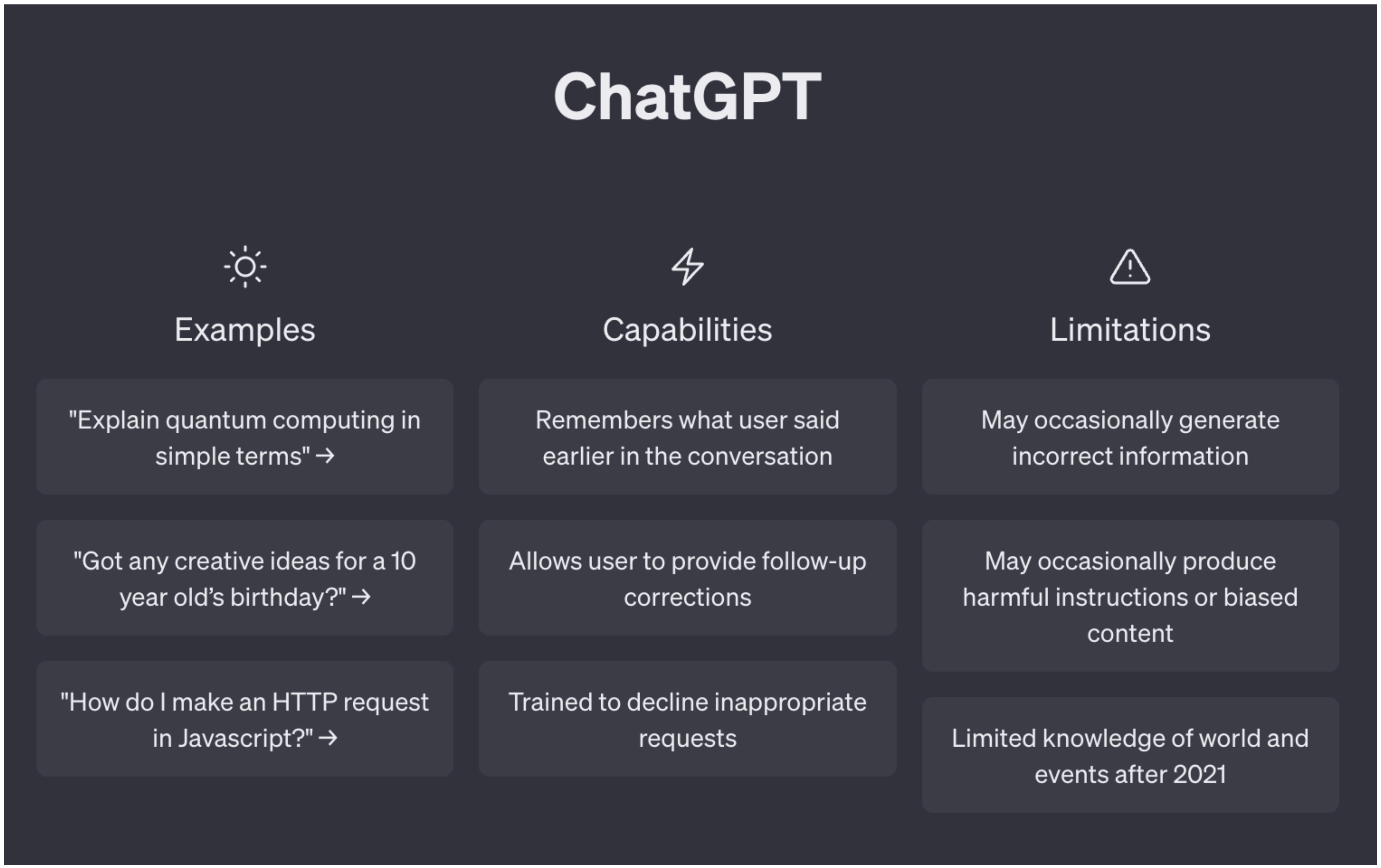
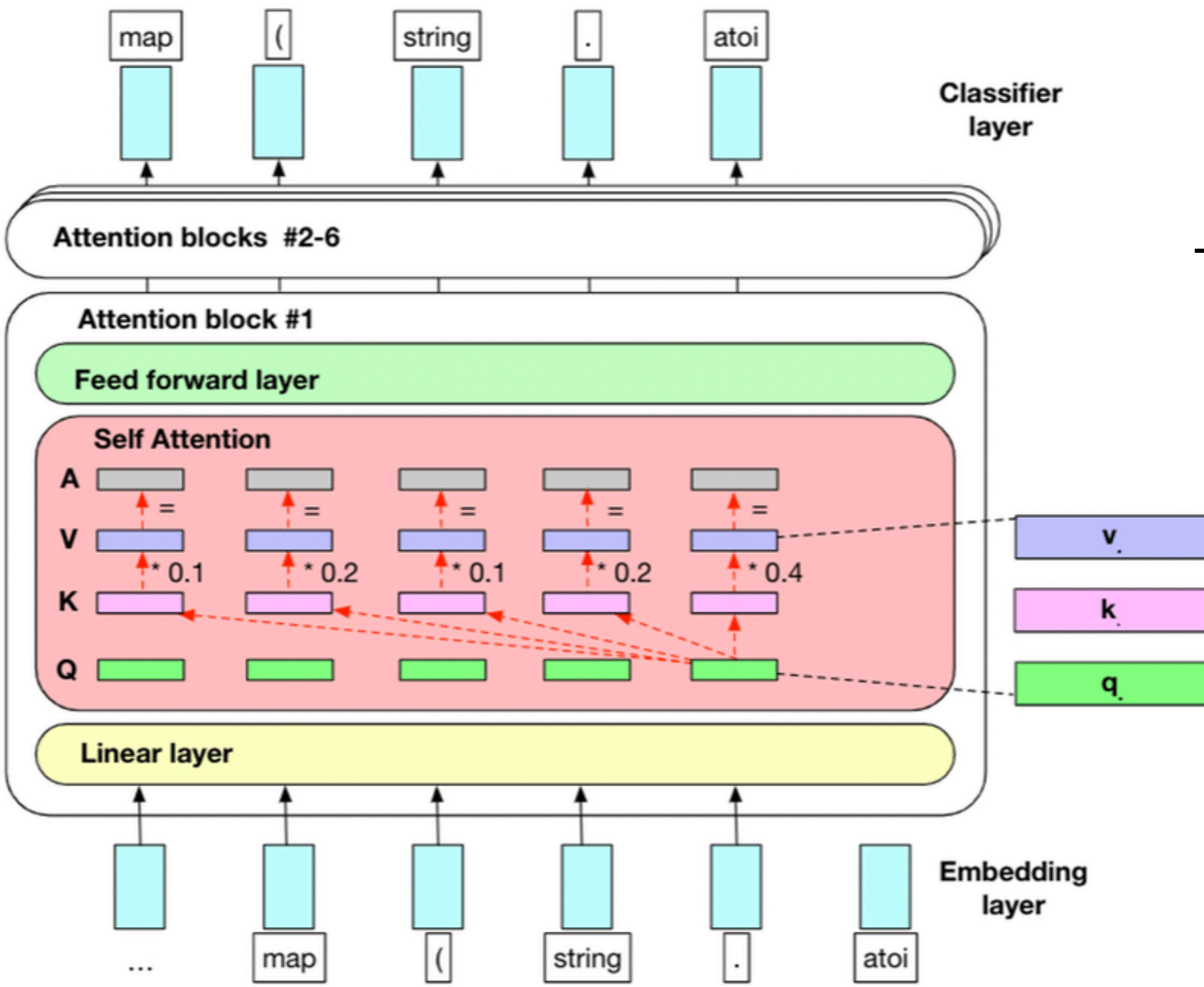


Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

Generative ‘AI’



Generative ‘AI’



Large Language Model - LLM

Transformer architecture shown

Generative 'AI'

ChatGPT